InfiniBand™ Architecture Specification Volume 2

Release 1.2

October, 2004
Final

Copyright © 1999, 2001, 2002, 2003, 2004 by InfiniBand℠ Trade Association.
All rights reserved.
All trademarks and brands are the property of their respective owners.

Table 1 Revision History

Revision	Release Date	
1.0	10/24/2000	Released version
1.0.a	06/19/2001	Release 1.0 augmented with errata material. Updates only correct errors - no additional features have been added.
1.1	11/06/2002	Release 1.0.a augmented with additional features.
1.2	October, 2004	Release 1.1 augmented with additional features, including Enhanced Signaling.

LEGAL DISCLAIMER

"THIS VERSION OF THE IBTA SPECIFICATION IS PROVIDED "AS IS" AND WITHOUT ANY WARRANTY OF ANY KIND, INCLUDING, WITHOUT LIMITATION, ANY EXPRESS OR IMPLIED WARRANTY OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

IN NO EVENT SHALL IBTA OR ANY MEMBER OF IBTA BE LIABLE FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, PUNITIVE, OR CONSEQUENTIAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES."

TABLE OF CONTENTS

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

Draft

IBTA

LIST OF FIGURES

BETA

DRAFT

IBTA

BETA

DRAFT

LIST OF TABLES

IBTA

IBTA

BETA

DRAFT

IBTA

CHAPTER 1: INTRODUCTION

This document comprises the specification for the physical aspects of InfiniBand Architecture and addresses the following areas:

- Overview of the Physical Layer and the aspects that make it up
- Signal Definitions that comprise the defined connection interfaces
- Link/Phy layer that defines the encoding and lane striping functions
- High Speed Electrical Signaling defines the parameters for the basic link signals
- Copper Cable Interconnects defines the wire cable interface
- Fiber Optic Interconnects defines the fiber attachment interface
- Mechanical Form Factors defines the standard physical structures
- Backplane Connector defines the mechanical and electrical characteristics of the connector used on the standard form factors
- Low Speed Electrical Signaling defines the parameters pertaining to the power and management signals of the backplane connector
- Power / Hot Plug defines the power delivery, control and consumption requirements for the defined form factors
- System Management
- OS Power Management

1.1 DOCUMENT CONVENTIONS

The following conventions are used in this specification.

1.1.1 NORMATIVE TERMS

- Shall

  The use of the word **shall** indicates a mandatory requirement that must be implemented to claim compliance to this specification**.**

- Shall not

  The use of the word **shall not** indicates a mandatory requirement to not implement a given aspect in order to claim compliance to this specification**.**

1.1.2 INFORMATIVE TERMS

- Should

  The use of the word **should** indicates flexibility of choice in an implementation with a strongly preferred preference**.**

- May

  The use of the word **may** indicates flexibility of choice in an implementation with no implied preference**.**

1.1.3 COMPLIANCE NOTATION

This specification has statements of compliance within the body of the document. These are identified using two notational conventions:

1) **Mandatory compliance**: statements in the form of "**CXX-YY:** statement text" where 'XX' is the chapter number and 'YY' is the individual statement number. These statements are required to be satisfied depending on the compliance level being claimed.
2) **Optional compliance**: statements in the form of "**oXX-ZZ:** statement text" where 'XX' is the chapter number and 'ZZ' is the individual statement number. These statements are required to be satisfied if a given optional feature is implemented within a compliance level being claimed.

The 'ZZ' numbers for Optional compliance statements increment independently from the 'YY' numbers used for Mandatory compliance statements.

Chapter 15: Volume 2 Compliance Summary defines the compliance levels for Volume 2. Within that chapter, a summary listing of all the Mandatory compliance and Optional compliance statements that apply for that level are shown.

1.1.4 ARCHITECTURE NOTES

Architecture Note

This information appears in-line to clarify architected items.

1.1.5 IMPLEMENTATION NOTES

Implementation Note

This information appears in-line and describes hints on implementations.

1.1.6 RECOMMENDATION

Recommendation to Someone

This information appears in-line and describes a recommended approach of a feature.

1.2 REFERENCES AND RELATED DOCUMENTS

The following are documents to which this specification refers.

[1] ANSI/TIA/EIA-492AAAA-A-97 - Detail Specification for 62.5-um Core Diameter/125-um Cladding Diameter Class 1a Graded-Index Multimode Optical Fibers Jan. 1, 1998

[2] ANSI/TIA/EIA-492AAAB-98 - Detailed Specification for 50-um Core Diameter/125-um Cladding Diameter Class 1a Graded-Index Multimode Optical Fibers Nov. 16, 1998

[3] ANSI/TIA/EIA-492CAAA-98 – Detail Specification for Class 1Va Dispersion-Unshifted Single-Mode Optical Fibers May 1, 1998

[4] ANSI/TIA/EIA-604-10: FOCIS 10 – Fiber Optic Intermateability Standard Type LC, 1999

[5] Center for Devices & Radiological Health

[6] Fibre Channel - Methodologies for Jitter Specification revision 10

[7] Fibre Channel Physical Interface Standard revision 8

[8] Fibre Channel Physical Interface Standard revision 8, Annex A

[9] FOTP-107 (EIA/TIA-107A) – Return Loss for Fiber Optic Components: 1st Ed. Feb. 1989, 2nd Ed. Mar. 1999

[10] IEC 1754-7-3 - Push/Pull MPO Adapter Interface Standard

[11] IEC 1754-7-4 - Push/Pull MPO Female Plug Connector Interface

[12] IEC 1754-7-5 – Push/Pull MPO Male Plug Connector Interface

[13] IEC 60793-2 Type A1a Dec. 1998

[14] IEC 60793-2 Type A1b Dec. 1998

[15] IEC 60793-2 Type B1 Dec. 1998

[16] IEC 60825-1 - Radiation Safety of Laser Products - Equipment Classification, Requirements and User's Guide, 1st Ed. Nov. 1993, Amended Sep. 1997

[17] IEC 60825-1 Class I, IIIA, IIIB – Nov. 1993

[18] IEEE 802.3z Gigabit Ethernet Standard

[19] ITU-T G.957 – Optical Interfaces for Equipment and systems Relating to the Synchronous Digital Hierarchy June, 1999

[20] TIA2.2.1 working specification TIA/EIA/455-204-FOTP204 Measurement Method for Multimode Fiber Bandwidth – to be published

[21] InfiniBand Architecture Specification, Volume 1

InfiniBand℠ Trade Association, http://www.infinibandta.org

[22] ISO 8601 Date/Time Format, http://www.iso.ch/cate/d15903.html

[23] OFSTP-14 (TIA/EIA 526-14) - Optical power loss measurements of installed multimode fiber cable plant, under ballot

[24] System Management Bus Specification, Revision 1.1, December 11, 1998. Copyright(c)1996, 1997, 1998, Benchmarq Microelectronics Inc., Duracell Inc., Energizer Power Systems, Intel Corporation, Linear Technology Corporation, Maxim Integrated Products, Mitsubishi Electric Corporation, National Semiconductor Corporation, Toshiba Battery Co., Varta Batterie AG.

[25] The I2C-Bus Specification, Version 2.0, December 1998, Philips Semiconductors

[26] IEEE Standard 181-2003

1.3 ACKNOWLEDGMENTS

This specification represents the collaboration of a number contributing companies. Special thanks to those who contributed.

1.3.1 STEERING COMMITTEE

The following individuals served as directors of the InfiniBand℠ Trade Association through the Steering Committee during the creation of Release 1.0 and 1.1 or creation of Release 1.2 of this specification:

1.3.1.1 CO-CHAIRS - DIRECTORS

Tom Bradicich	Tom Macdonald

1.3.1.2 MEMBERS

Jacqueline Balfour	Ken Jansen	Jim Pinkerton
Kevin Deierling	Michael Krause	Martin Whittaker
Balint Fleisher	Todd Matters	Bob Zak
Dr. Alfred Hartmann	Ed Miller	
David Heisey	John Pescatore	

1.3.2 TECHNICAL WORK GROUP

The following individuals served as members of the Technical Work Group (TWG) with responsibility for the oversight of the technical production during the creation of Releases 1.0 and 1.1 or creation of Release 1.2 of this specification:

1.3.2.1 CO-CHAIRS

Dwight Barron	Jeff Hilland	David Wooten
Paul Grun	Irving Robinson	

1.3.2.2 MEMBERS

Dr. Alan Benner	Dr. Al Hartmann	Dr. Greg Pfister
Mark Bradley	Michael Krause	Greg Still
Wolfgang Cristl	Bill Lynn	Ken Ward
Diego Crupnicoff	Ed Miller	

1.3.3 VOLUME WORKING GROUP

This volume is authored by the membership the Electrical/Mechanical Workgroup (EWG) of the InfiniBand Trade Association. The other working groups that are part of the Trade Association are listed in *InfiniBand Architecture Specification, Volume 1*. The EWG is indebted for their efforts in consultation and review.

1.3.3.1 CO-CHAIRS

The following individuals served as co-chairs of the EWG during the creation of Releases 1.0 and 1.1 of this specification:

Mike Chastain	David Moss

The following individual served as chair of the EWG during the creation of Release 1.2 of this specification:

Moises Cases

1.3.3.2 VOLUME EDITOR

The following individual served as editor of Release 1.0 and 1.1 of this volume during its creation:

Greg Still

The following individual served as editor of Release 1.2 of this volume during its creation:

Alan Benner

1.3.3.3 SUBTEAM LEADERS

The following individuals served as subteam leaders and specific chapter editors of Release 1.0 and 1.1 of this volume during its creation:

Bill Bunton	Dominic Goodwill	Trevor Williams
Del Cecchi	John Leder	
Jay Diepenbrock	Siamak Tavallaei	

The following individuals served as subteam leaders and specific chapter editors of Release 1.2 of this volume during its creation:

Alan Benner	Jay Diepenbrock
Del Cecchi	Trevor Williams

1.3.3.4 INDIVIDUAL CONTRIBUTORS

The following individuals provided active contributions to Release 1.0 and 1.1 of this volume during its creation:

Andrew Alduino	Art Kimmel	Harry Rogers
Paul Artman	Don Lentz	Art Rousmaniere
Brian Beaman	Dennis Miller	Richard Schumacher
Dave Brown	J. P. Miller	Kim Sides
Ray Clemo	Mike Miller	Tom Slaight
Steve Contreras	Mark Myers	Bill Stanley
Casimer DeCusatis	Istvan Novak	Dan Stigliani
Hal Dozier	Shlomo Novotny	Farrukh Syed
Pat Egan	Tom Osten	Pat Thaler
Stillman Gates	Ali Oztaskin	Chris Trudeau
Ali Ghiasi	Ken Privitt	Jim Warren
Lowell Good	Kevin Przybylski	
Ken Gross	Eddie Reid	

The following individuals provided active contributions to Release 1.2 of this volume during its creation:

Moises Cases
John Calvin
Del Cecchi
William Cerreta
Hillel Chapman
Ian Colloff
Casimer M. DeCusatis
Christy Devonport
Joseph (Jay) Diepenbrock
Ronen Eckhouse
Shahar Eytan
Andy Kayner
Tadashi Kumamoto
Todd Leonard
Jacques Longueville
Gregory Mann
Padraig McDaid
J. P. Miller
Keith Murr
Dale Murray
Jay Neer
John Petrilla
Greg Pfister
Daniel Reed
Eddie Reid
Michael Rost
Peter Smith
Mike Sorna
Greg Still
Antonius Susanto
Trevor Williams

1.4 DISCLAIMER

Like any document, this specification is subject to errata for correctness, clarity, and enhancements. The InfiniBand Trade Association hosts a web site at http://www.infinibandta.org. Please visit this site to check for errata and updates to this specification.

CHAPTER 2: GLOSSARY

3U A chassis size defined by the IEEE 1101.10 specification.

6U A chassis size defined by the IEEE 1101.10 specification.

_L The _L suffix to a signal name indicates a low true signal.

Address Handle An object that contains the information necessary to transmit messages to a remote port over Unreliable Datagram service.

Actively Managed Chassis An **Actively Managed Chassis** provides an InfiniBand™ specified GUID and physical Slot Information to every IB Module on the Module's unique IB-ML. In addition, it provides a Chassis Management Entity on at least one IB Module's IB-ML. In an actively managed chassis, the Slot Information for every Module provides information identifying all of the Slot(s) that have access to the CME.

Adapter Also called I/O adapter. Please see Host Channel Adapter or Target Channel Adapter. A form of TCA which conforms to the InfiniBand™ Architecture form factor definition. The term adapter by itself is overloaded due to its general use in the computer industry and should not be used by itself. Typically the term adapter pertains to the channel adapter but is also used in the context of an IO adapter.

Anycast An identifier for a set of interfaces (typically belonging to different nodes). A packet sent to an anycast address is delivered to one of the interfaces identified by that address (the "nearest" one, according to the routing protocols' measure of distance).

Attribute The collection of management data carried in a Management Datagram.

Aux Power The 5V auxiliary power input to an IB Module that is provided on the VA_In pin. This is the standby power source used when Bulk Power is off.

Average Optical Power The optical power measured using an average reading power meter when transmitting a specified code sequence as defined in the test procedure.

B_Key Please see Baseboard Management Key.

Backplane Physical PCB into which an IB Module plugs, this may be into a Server, Switch, I/O Chassis, etc.

Backplane Connector — The connector located in a Server, Switch, I/O Chassis, etc. that mates with an IB Modules Edge Connector.

Base LID — The numerically lowest Local Identifier that refers to a Port.

Baseboard Managed Unit — Any Unit which provides InfiniBand™ specification defined information about itself by a Baseboard method MAD operation through the InfiniBand™ link.

Baseboard Management Key — A construct that is contained in IBA management datagrams to authenticate that the sender is allowed to perform the requested operation.

Baseboard Management — Management of physical hardware (environmentals, Vital Product Data, LEDs, etc.). Considered synonymous with System Management as opposed to “Subnet Management”.

Baseboard Management Proxy — A Unit which provides the means to translate InfiniBand™ Baseboard Management Datagram (MAD) requests and responses to IB-ML or other transports on behalf of one or more other entities.

The proxy must be addressable by the InfiniBand™ subnet to allow operations to be targeted to it (aka it has a LID)

Baud — The signaling speed on a lane in transitions per second.

Beacon Sequence — The periodic transmission of the Training Sequence 1 Control Ordered-set. The Beacon sequence is used to wake or begin Link Training with the device on the opposite end of a Link that may be in a power managed state or has been just attached to the link.

BER — Bit error rate.

BTH — Base Transport Header.

Board — Physical, pluggable entity that is defined by the InfiniBand™ Architecture.

Bulk Power — The 12V main power input to an IB Module that is provided on the **VB_In** pins.

Burst-BER — Number of bit errors measured within a sliding window.

Byte Striping — The byte stream representing the packet is sent by distributing the bytes sequentially across the available Physical Lanes, one byte per Lane, in order. The first byte goes in lane 0, the second in lane 1, the third in lane 2 and so on. When the last physical lane is reached, the process starts again with lane 0.

CA — Please see Channel Adapter.

Carrier Module — The carrier module, designed specifically for each size of IB Module defines the thermal, EMI, ESD and vibration/shock interface to a chassis.

CDR — Clock Data Recovery unit

Center Wavelength — The nominal value of the central wavelength of the operating, modulated laser. This is the wavelength (see FOTP-127) where the effective optical power resides.

Channel — The association of two queue pairs for communication.

Channel Adapter — Device that terminates a link and executes transport-level functions. One of Host Channel Adapter or Target Channel Adapter.

Chassis — The collection of IB Modules, and their associated power and cooling resources housed within a single mechanical package. This may be a Server, Switch, I/O Chassis, etc.

Chassis GUID — 8 bytes of Globally Unique ID for every Chassis

Chassis Management Entity — The Chassis Management Entity may or may not include a processor. In it's simplest form the CME is merely an IB-ML de/MUX. The CME may provide proxy access to the IB Modules IB-ML.

Clock Compensation — The SKIP ordered-set is used for clock compensation. A device may add or drop a SKP symbol from the Control Ordered-set to prevent it's link input/output buffer from over-running or under-running.

CME — Please see Chassis Management Entity.

COM — The Comma symbol is transmitted to identify the start of a Training Sequence 1, Training Sequence 2 or SKIP ordered-set.

Compliance Channel — A worst-case connection from driver to receiver. The compliance channel and the transmitter define the minimum acceptable inputs which a receiver shall be capable of receiving at a specified bit error rate.

Compliant Channel — A compliant channel is any channel which provides a signal at the receiver which is better than the Compliance Channel.

Control Ordered-set — Control Ordered-sets are used for Link Training and Clock Compensation. The first symbol of all ordered-sets is the COM symbol, additional symbols are unique to the set type.

Cover — The protective cover which mates with the carrier module defining the outer surface of an IB Module.

CQE

CRC Please see Cyclic Redundancy Check.

Cyclic Redundancy Check A number derived from, and stored or transmitted with, a block of data in order to detect corruption. By recalculating the CRC and comparing it to the value originally transmitted, the receiver can detect some types of transmission errors.

Data Payload The data, not including any control or header information, carried in one packet.

DETH Datagram Extended Transport Header.

DFE Decision Feedback Equalizer.

DGID Destination Globally Unique Identifier.

DCD Duty cycle distortion.

DDJ Data Dependant Jitter.

de/MUX Multiplexer / Demultiplexer

DLID Destination Local Identifier

Disparity The difference between the number of ones and number of zeros transmitted in a Physical Lane. The running disparity is a binary parameter with a value of positive or negative.

Dispersion A term used to denote pulse broadening and distortion. For fiber optic links, the two general categories of dispersion are modal dispersion, due to the difference in the propagation velocity of the propagation modes in a multimode fiber, and chromatic dispersion, due to the difference in propagation of the various spectral components of the optical source.

DJ Deterministic Jitter

EBP Please see End of Bad Packet Delimiter.

Edge Connector The connector on an IB Module that mates with a Backplane Connector.

EGP Please see End of Good Packet Delimiter.

EMC Electro-Magnetic Compatibility.

EMC Gasket The name of the gasket used between IB Modules to shield for electro-magnetic emissions.

IBTA

End of Good Packet Delimiter

The End of Good Packet Delimiter (EGP) symbol is used to mark the end of each packet as it is transmitted by the originating node.

End of Bad Packet Delimiter

The End of Bad Packet Delimiter (EBP) symbol is used to mark the end of a bad packet forwarded by a switch or router node.

Endnode

An endnode is any node that contains a Channel Adapter and thus it has multiple queue pairs and is permitted to establish connections, end to end context, and generate messages. Also referred to as Host Channel Adapter or Target Channel Adapter, two specific types of endnodes.

Endport

A Port which can be a destination of LID-routed communication within the same Subnet as the sender. All Channel Adapter ports on the subnet are endports of that subnet, as is Port 0 of each Switch in the subnet. Switch ports other than Port 0 may not be endports. When *port* is used without qualification, it may be assumed to mean *endport* whenever the context indicates that it is a destination of communication.

ESD

Electro-Static Discharge.

Error Event

A single root cause can result in multiple error events.

Error Propagation

A special character in the end of packet delimiter is used to delimit packets in which a transmission error has already been detected and reported.

Even Alignment

A special character in the code set used on the link to establish a given disparity.

Extinction Ratio

The ratio (in dB) of the average optical energy in a logic one level to the average optical energy in a logic zero level measured under modulated conditions at the specified baud rate.

Fabric

The collection of Links, Switches, and Routers that connects a set of Channel Adapters.

Fall Time

The time interval for the falling edge of a pulse to transition from its 80% amplitude level to its 20% amplitude level.

FFE

Feedforward Equalizer.

Fiber Optic Adapter

A device into which two optical connectors plug, joining two optical segments.

Fiber Optic Cable

A jacketed optical fiber or fibers.

Fiber Optic Segment

An unbroken length of optical fiber with an optical connector on each end. The fiber may contain splices. A fiber optic segment shall not con-

tain a fiber optic adapter.

Fiber Optic Test Procedure — EIA/TIA standards developed and published by the Electronic Industries Association (EIA) and Telecommunications Industry Association (TIA) under the EIA-RS-455 series of standards. Please see FOTP.

Form Factor — The definition of physical packaging that specifies mechanical limits, connectors, and placement of connectors and connector plates.

FOTP — Please see Fiber Optic Test Procedure.

Fully Managed Repeater — A Repeater which provides VPD information about itself by a GetInfo MAD operation targeted to it and allows IB-ML access to defined information. **A repeater cannot fulfill this definition.**

Fully Managed TCA — A TCA which provides Module Information about itself when a GetInfo MAD operation is targeted to it and allows IB-ML access to defined Module information.

Fully Managed Unit — Any Unit which is both a Baseboard Managed Unit and an IB-ML Managed Unit.

Gb/s — Giga-bits per second (10^9 bits per second)

GB/s — Giga-bytes per second (10^9 bytes per second)

General Services Interface — An interface providing management services (e.g., connection, performance, diagnostics) other than subnet management. QP1 is reserved for the GSI, which may redirect requests to other QPs.

GID — Please see Globally Unique Identifier.

Globally Unique Identifier — A software-readable number that uniquely identifies a device or component.

GMP — General Management Packet.

Graceful Hot Removal — The removal of an IB Module that has first been placed in a quiescent state. V_{Bulk} may or may not be on.

GSI — Please see General Services Interface.

GT/s — Giga-transfers per second (10^9 transfers per second).

GUID — Please see Globally Unique Identifier.

HCA — Please see Host Channel Adapter.

DRAFT

Host — One or more Host Channel Adapters governed by a single memory/CPU complex.

Host Channel Adapter — A Channel Adapter that supports the Verbs interface.

Host Connector — The connector interface associated with the mating of a pluggable device to a printed circuit board (PCB).

Host Node — A host node is a type of endnode consisting of one or more host channel adapters governed by a single memory/CPU complex. A host node supports one or more software processes, which use the subnet for communication between peer processes (IPC) and I/O services. A host may support processes that provide an I/O service (e.g., console service). Such a process is considered an I/O controller and the host enumerates that service the same as an I/O unit enumerates I/O controllers.

Hot Add — The insertion of an IB module into a backplane that has both V_{Bulk} and V_{Aux} present. The IB module powers up and initiates a training sequence.

IBA — InfiniBand™ Architecture.

IB-ML — InfiniBand™ Management Link.

IB-ML Managed Unit — Any Unit which provides InfiniBand™ specification defined information about itself through accesses from the IB-ML.

IB-ML Management Proxy — A Unit which serves as a means to translate IB-ML operations (requests and responses) to the InfiniBand™ Subnet link(s) on behalf of one or more other entities.

IB-ML Master — A device on the IB-ML that initiates operations and provides the clock for the transfer to or from an IB-ML Slave device.

IB-ML Slave — A device on the IB-ML that responds to an operation that is addressed to it by an IB-ML Master. The clock used for the transfer is provided by the IIB-ML Master.

IB Board — The PCB assembly inside an IB Module.

IB Module — A unit that conforms to any of the form factors defined in Volume 2 of the InfiniBand™ specification. The IB module minimally consists of the following: at least one IB Board, a Carrier Module, and a Cover. There are two defined module heights and two defined module widths. Double width modules occupy two IB chassis slots.

Standard Module

Standard Double Wide Module

Tall Module

Tall Double Wide Module

IBT InfiniBand™ Technology

IHV Independent Hardware Vendor

Idle Data Data symbols transmitted to fill idle time on a link. These symbols are not part of a packet and do not have delimiters symbols to mark their start and end.

In-band Management Refers to the monitoring and control of InfiniBand™ components using the IB Subnet Link.

Independent Hardware Vendor Any vendor providing hardware. Used synonymously at times with Hardware Vendor.

Inter IB-ML Management Proxy A unit which serves to translate IB-ML operations from one unit to another unit's IB-ML. (This is currently not specified in the InfiniBand™ Architecture but is included for completeness).

Intersymbol Interference The effect on a sequence of symbols in which the symbols are distorted by transmission through a limited bandwidth medium to the extent that adjacent symbols begin to interfere with each other. Also referred to as ISI.

Invalid Key A Key is invalid if it numerically different from the correct Key value associated with an IBA-defined resource. Please see Key.

I/O Input/Output.

I/O Adapter An I/O controller and TCA, usually implemented as an IB Module, that provides access to one or more I/O devices possible attached via a secondary peripheral bus of network.

I/O Chassis The collection of Slots and their associated power and cooling resources housed within a single mechanical package.

I/O Controller One of the two architectural divisions of an I/O Unit. An I/O controller (IOC) provides I/O services, while a Target Channel Adapter provides transport services.

I/O Hierarchy An I/O unit contains one or more I/O controllers**.** Each I/O controller provides access to one or more I/O devices or I/O ports. This is roughly equivalent to a PCI card that might contain up to 8 I/O functions.

On the host side, the equivalent of a controller is a process. The part of the I/O process that accesses an I/O controller is referred to as the I/O driver. An I/O driver controls one or more I/O controllers, thus there is an instance of an I/O device for each I/O controller.

I/O Node

A type of endnode that provides I/O functions.

I/O Plate

Physical end of an IB Module where the I/O media cables are mounted for exit.

I/O Unit

An I/O unit (IOU) provides I/O service(s). An I/O unit consists of one or more I/O Controllers attached to the fabric through a single Target Channel Adapter.

IOC

Please see I/O Controller.

IOU

Please see I/O Unit.

ISI

Please see Intersymbol Interference.

Jitter

Deviations from the ideal timing of an event which occur at frequencies >=10kHz. Jitter is customarily subdivided into deterministic and random components.

Jitter, Deterministic

Timing distortions caused by normal circuit effects in the transmission system. Deterministic jitter is often subdivided into duty cycle distortion (DCD) caused by propagation differences between the two transitions of a signal and data dependent jitter (DDJ) caused by the interaction of the limited bandwidth of the transmission system components and the symbol sequence.

Jitter, Random

Jitter due to thermal noise which may be modeled as a Gaussian process. The peak-to-peak value of RJ is of a probabilistic nature and thus any specified value yields an associated BER.

Key

A construct used to limit access to one or more resources, similar to a password. Security is provided by (1) limiting the ability to generate keys and check keys to well-trusted levels of function in the network, (2) making the keys large numbers, and managing how they are changed, so that each key should be unique within the system at once, and (3) the detection of invalid key usage is raised to high levels of software where the source of the invalid key can be identified and dealt with. The following keys are defined by the InfiniBand™ Architecture:

Baseboard Management Key.

Management Key.

Queue Key.

Partition Key.

L_Key Please see Local Route Header.

Lane Please see Physical Lane

Lane Identifier The second symbol of a training sequence that identifies the Physical Lane(s) of a 1x, 4x or 12x Link. It is used to recognize physical lane reversal and to determine link width, etc.

LED Light Emitting Diode.

'Let-Through' Failure A Failure mechanism of non-isolated voltage converters which allows the primary input power of the converter, through a DC path, onto the output.

LID Please see Local Identifier.

Link A full duplex transmission path between any two network fabric elements, such as Channel Adapters or Switches.

Link Heartbeat Ordered-Set The sixteen symbol ordered-set composed of a COM symbol, Lane Identifier, three D1.2 data symbols, an OpCode symbol, a reserved symbol, a PortNum symbol (only used for switch ports), and an eight-symbol Globally Unique ID (GUID), used for ensuring link aliveness and determining link round-trip latency.

Link Training Link Training is the process of establishing link synchronization between two Link Endpoints. The Link Training State Machine controls the transition between the Link Down and Link UP state. This process includes but is not limited to:

1) Bit synchronization
2) Symbol synchronization
3) Width and Speed negotiation
4) Physical Lane ordering
5) Physical Lane polarity
6) Training Sequence handshake
7) Error recovery

Local Identifier An address assigned to a port by the Subnet Manager, unique within the subnet, used for directing packets within the subnet. The Source and Destination LIDs are present in the Local Route Header.

Local Route Header Routing header present in all InfiniBand™ Architecture packets, used for routing through switches within a subnet.

Logic Ground — The voltage to which the logic signals (high speed and low speed) are referenced. This is synonymous with Signal Ground.

Longitudinal Airflow — As it pertains to the IB Module, longitudinal airflow is defined as airflow across the module predominantly in the direction normal to the back-plane and connector housing.

LRH — Please see Local Route Header.

M_Key — Please see Management Key.

MAD — Please see Management Datagram.

Managed Unit — A Unit which provides Module Information about itself to an external source.

Management Datagram — Refers to the contents of an Unreliable Datagram packet used for communication among the HCAs, Switches, Routers, and TCAs to manage the network. InfiniBand™ Architecture describes the format of a number of these management commands.

Management Key — A construct that is contained in IBA Management Datagrams to authenticate the sender by the receiver.

Management Proxy — A Unit which serves as a means to allow for system management operations to get from one “band” to the other (i.e. an InfiniBand™ Link to IB-ML or IB-ML to an InfiniBand™ link)

MB/s — Mega-bytes per second (10^6 bytes per second)

Message — A transfer of information between two or more Channel Adapters that consists of one or more packets.

MN — Please see Modal Noise.

Modal Noise — Noise in a laser based optical communication system caused by the incomplete collection of the spatially correlated interference pattern.

Mode-Partition Noise — Noise in a laser based optical communication system caused by the changing distribution of laser energy partitioning itself among the laser modes (or lines) on successive pulses in the data stream. The effect is a different center wavelength for the successive pulses resulting in arrival time jitter attributable to chromatic dispersion in the fiber.

Mode-Partition Noise k Factor — Empirically derived factor linking mode-partition noise to system penalty.

Modifiers — In a verb definition, the list of input and output objects that specify how,

	and on what, the verb is to be executed.
Module Information	Information provided to Management Software about an IB Module.
MPN	Please see Mode-Partition Noise.
MPN-k	Please see Mode-Partition Noise k Factor.
Numerical Aperture	The sine of the radiation or acceptance half angle of an optical fiber, multiplied by the refractive index of the material in contact with the exit or entrance face.
O/E	Optical/Electrical.
OFSTP	Please see Optical Fiber System Test Practice.
OMA	Please see Optical Modulation Amplitude.
ORL	Please see Optical Return Loss.
Operating System Vendor	The software manufacturer of the operating system that is running on the node under discussion.
Optical Cable Plant	All passive communications elements (e.g., optical fiber, connectors, splices, etc.) between a transmitter and a receiver.
Optical Connector	An optical connector connects the optical media to the receptacle of an optical transmitter, to the receptacle of an optical receiver, or to a fiber optic adapter.
Optical Connector Loss	The optical power lost between two optical connectors mated in a fiber optic adapter.
Optical/Electrical Converter	Also referred to as "O/E Converter". A device that converts electrical signals on a board to/from IB-compliant optical signals. It optionally contains CDR and/or de/MUX functionality, but it does not contain data buffers and is not protocol aware. The electrical signals into / out of an Optical/Electrical Converter may be vendor-specific.
Optical Eye Opening	For a fiber optic link, the time interval across the eye, measured at the 50% normalized eye amplitude which is error free to the specified BER.
Optical Fiber	Any filament or fiber, made of dielectric material, that guides light.
Optical Fiber System Test Practice	Standards developed and published by the EIA/TIA under the EIA/TIA-526 series of standards. This term is also referred to as OFSTP.

Optical Modulation Amplitude — The absolute difference between the optical power of a logic one level and the optical power of a logic zero level. This term is abbreviated OMA.

Optical Passive Loss — The insertion loss resulting from connections (adapters or splices), and attenuation attributable to the fiber cable plant.

Optical Receiver — The part of an Optical/Electrical Converter which receives an optical signal.

Optical Receiver Bandwidth — High frequency 3dB roll-off frequency of the optical receiver.

Optical Receiver Overload — The maximum acceptable value of the received average optical power at the receiver input to achieve the specified BER.

Optical Receiver Sensitivity — The minimum acceptable value of average received signal, at the defined optical test point, to achieve the specified BER. It takes into account power penalties caused by use of a transmitter with a worst-case output. In the case of an optical path it does not include power penalties associated with Dispersion, Jitter, effects related to the modal structure of the source or reflections from the optical path. These effects are specified separately in the allocation of maximum optical path penalty.

Optical Receptacle — The part of the Optical Transmitter or Receiver into which an optical connector plugs.

Optical Return Loss — The ratio (expressed in units of dB) of optical power incident upon a component port or an assembly to the optical power reflected by that component when that component or assembly is introduced into a link or system. This term is abbreviated ORL.

Optical Stressed Receiver Sensitivity — Minimum receiver sensitivity required in order to perform PLL locking on the data. This measure takes a transmitter with worst-case output, and adds the channel loss and the EYE losses. The stressed EYE is used principally used to know how the jitter affect the system and how low the low pass filter has to be in the receiver.

Optical System Penalty — An optical link penalty to account for those effects other than optical passive loss.

Optical Transceiver — A device that converts IB-compliant electrical signals on a board to/from IB-compliant optical signals. An optical transceiver may be 1x, 4x or 12x wide. In general, an Optical Transceiver will contain one or more Optical/Electrical Converters, CDR's, and data buffers.

Optical Transmitter — The part of an Optical/Electrical Converter which transmits an optical signal.

Ordered Set — See Control Ordered-set.

Out-of-band Management — Management messages which traverse a transport other than the InfiniBand™ fabric.

P_Key — Please see Partition Key.

Packet — The indivisible unit of IBA data transfer and routing, consisting of one or more headers, a Packet Payload, and one or two CQEs.

Packet Data — Data symbols transmitted as part of a Data Packet Payload or a Link Packet Payload. Packet data is delimited by start symbols SDP and SLP and terminated by end symbols EGP and EBP.

Packet Payload — The portion of a Packet between (not including) any Transport header(s) and the CRCs at the end of each packet. The packet payload contains up to 4096 bytes.

PAD — The PAD symbol is transmitted only on a 12x link to maintain packet framing alignment.

Partition Key — A value carried in packets and stored in Channel Adapters that is used to determine membership in a partition.

Partition Manager — The entity that manages partition keys and membership.

Passively Managed Chassis — A **Passively Manage**d chassis provides an InfiniBand™ specified GUID and physical Slot Information to every IB Module on the Modules unique IB-ML.

PCB — Printed Circuit Board

PD — Photodiode, used for converting optical signals to electrical at a receiver.

Permanent Errors — A permanent error is surfaced when a continuous and irreversible fault is present within an IBA component.

Physical Lane — A set of one transmit and one receive differential pairs. A 1x, 4x and 12x link is composed of one, four and twelve physical lanes respectively.

Pluggable — A description for a transceiver module which may be unplugged for repair or replacement.

PM — Please see Partition Manager.

PN — Processor Node

Polling — A port state where the transmitter is generating a Beacon Sequence and the receiver is waiting to respond to a Beacon Sequence.

IBTA

Port	Location on a Channel Adapter or Switch to which a link connects. There may be multiple ports on a single Channel Adapter, each with different context information that must be maintained. Switches/switch elements contain more than one port by definition.
Port Type 1	An electrical interface type that utilizes the parameters of the original InfiniBand specification (1.0, 1.0a).
Port Type 2	An electrical interface type that utilizes more stringent parameters than Port Type 1. Port Type 2 is currently associated with pluggable devices.
Power Management	The ability for an operating system to control the power consumption of InfiniBand™ Architecture compliant devices (Endnode devices and Switches).
Power Subsystem Management	The ability to monitor and control the power supplies of a system or chassis.
Processor Node	One or more general-purpose processors running under a single cache coherent memory model that uses a HCA to connect a processor bus to one or more Ports. Additionally, a PN is capable of performing initialization and configuration.
Proxy Managed Repeater	A Repeater which must have a proxy in-place to respond to GetInfo MAD operations but allows IB-ML access to defined VPD information. All repeater boards fulfill this definition.
Proxy Managed TCA	A TCA which must have a proxy in-place to respond to GetInfo MAD operations but allows IB-ML access to defined VPD information. This type of TCA in not defined by the InfiniBand™ specification.
Proxy Managed Unit	Any Unit which must have a proxy in-place because Baseboard information is not accessible through its native InfiniBand™ fabric link(s).
PSE	Protocol State Engine (Phy, Link, Frame)
QP	Please see Queue Pair.
Queue Pair	Consists of a Send and a Receive Work Queue. Send and receive queues are always created as a pair and remain that way throughout their lifetime. A Queue Pair is identified by its Queue Pair Number.
Queue Pair Number	Identifies a specific Queue Pair within a Channel Adapter.
RDETH	Reliable Datagram Extended Transport Header.
Receive Queue	One of the two queues associated with a Queue Pair. The receive queue contains Work Queue Elements that describe where to place incoming

data.

Relative intensity noise — Laser noise in dB/Hz with 12 dB optical return loss, with respect to the optical modulation amplitude.

Retiming Repeater — A device which recovers and retransmits data, using a local oscillator to eliminate jitter transfer, and hence perform a jitter reset. (SKIP ordered-set dependent)

RIN12-OMA — Please see Relative intensity noise.

Rise Time — The time interval for the rising edge of a pulse to transition from its 20% amplitude level to its 80% amplitude level.

RJ — Please see Jitter, Random.

RMS — Root Mean Square.

RNR Nak — Receiver Not Ready. A response signifying that the receiver is not currently able to accept the request, but may be able to do so in the future.

Router — A device that transports packets between IBA subnets.

Run Length — Maximum number of consecutive identical bits in the transmitted signal e.g., the pattern 0011111010 has a run length of five (5).

Running Disparity — A binary parameter indicating the cumulative disparity (positive or negative) of all previously issued bits.

SA — Please see Subnet Administration.

SDP — Please see Start of Data Packet Delimiter.

Send Queue — One of the two queues of a Queue Pair. The Send queue contains WQEs that describe the data to be transmitted.

Server

1) The passive entity in a connection establishment exchange.
2) An entity (e.g., a process) that provides services in response to requests from clients.
3) The class of computers that emphasize I/O connectivity and centralized data storage capacity to support the needs of other, typically remote, client computers.

SGID — Source Globally Unique Identifier.

Signal Ground — Please see Logic Ground.

Skew	The timing difference between Physical Lanes as measured using the zero crossings of the differential voltage of the COMmas present in the training sequences TS1 and TS2 or in the SKIP ordered-set.
SKP	The SKP symbol is transmitted as part of a SKIP ordered-set.
SKIP ordered-set	The SKIP ordered-set consists of a COM symbol followed by three SKP symbols when initially transmitted. It may consist of a COM symbol followed by one to five SKP symbols when received.
Sleeping	A port state where the transmitter is quiescent and the receiver is waiting to respond to a Beacon Sequence.
SLID	Source Local Identifier
Slot	An InfiniBand™ specified volumetric envelope with a specified backplane connector into which one of the defined IB Modules plug.
Slot ID	A Slot Designation provided in Slot Information
Slot Information	Information provided by the Chassis about individual Slots
SLP	Please see Start of Link Packet Delimiter.
SM	Please see Subnet Manager.
SMA	Please see Subnet Management Agent.
SMA GUID	The Globally Unique Identifier common to all resources within the scope of a single Subnet Management Agent.
SMD	See Subnet Management Data.
SMP	Please see Subnet Management Packet.
Solicited Event	A facility by which a message sender may cause an event to be generated at the recipient when the message is received.
Spectral Width	The weighted root mean square width of the optical spectrum (see FOTP-127).
Start of Data Packet Delimiter	The Start of Data Packet Delimiter symbol is transmitted to identify the start of a end-to-end data packet.
Start of Link Packet Delimiter	The Start of Link Packet Delimiter symbol is transmitted to identify the start of a link control packet.
Stressed Receiver	In a Stressed Receiver test the worst case transmitter eye opening is

applied to the optical receiver under test. The eye closure typically is generated with a combination of coaxial cable and/or Bessel 4th order filter driving a laser diode directly.

Stress Receiver ISI Penalty — Vertical eye closure penalty due to ISI.

Stressed Receiver DCD component of DJ — Horizontal eye closure caused by DCD measured at average optical power.

Subnet — A set of InfiniBand™ Architecture Ports, and associated links, that have a common Subnet ID and are managed by a common Subnet Manager. Subnets may be connected to each other through routers.

Subnet Administration — The architectural construct that implements the interface for querying and manipulating subnet management data.

Subnet Manager — One of several entities involved in the configuration and control of the subnet.

> **Active Subnet Manager:** Any subnet manager currently exercising control over all or part of the subnet. An active subnet manager may be the master subnet manager, or an alternate subnet manager acting on the behalf of the master.This is sometimes referred to as the formal subnet manager.
>
> **Alternate Subnet Manager:** Any subnet manager that is acting on the behalf of the master subnet manager, but is not the master subnet manager.
>
> **Master Subnet Manager:** The subnet manager that is authoritative, that has the reference configuration information for the subnet.
>
> **Standby Subnet Manager:** A subnet manager that is currently quiescent, and not in the role of a master SM, by agency of the master SM. Standby Isms are dormant managers.

Subnet Management Agent — An entity present in all IBA Channel Adapters and Switches that processes Subnet Management Packets from Subnet Manager(s).

Subnet Management Data — Vital Product Data required by the Subnet Manager.

Subnet Management Packet — The subclass of Management Datagrams used to manage the subnet. SMPs travel exclusively over Virtual Lane 15 and are addressed exclusively to Queue Pair Number 0.

Surprise Hot Removal — The removal of an IB Module from a backplane that has both V_{Bulk} and V_{Aux} present without first being placed in a quiescent state.

Switch — A device that routes packets from one link to another of the same Subnet, using the Destination Local Identifier field in the Local Route Header.

Switch Management Port

A virtual port by which a Switch may be managed.

Symbol Time

The transmit time for 1 symbol. With 8b/10b encoding, a symbol is 10 bits long; thus, a symbol time is 10*UI (e.g., 4 ns for 2.5Gb/s signaling, and 1 ns for 10Gb/s signaling).

TCA

Please see Target Channel Adapter.

Target Channel Adapter

A Channel Adapter typically used to support I/O devices. TCAs are not required to support the Verbs interface. See also I/O Unit.

Training Sequence 1

The sixteen symbol ordered-set which consists of a COM symbol, Lane Identifier and fourteen D10.2 data symbols.

Training Sequence 2

The sixteen symbol ordered-set which consists of a COM symbol, Lane Identifier and fourteen D5.2 data symbols.

Training Sequence 3

The sixteen symbol ordered-set which consists of a COM symbol, Lane Identifier, six D13.2 data symbols, a bit map of the active speeds, a bit map requesting transmitter driver de-emphasis and/or link heartbeat enabling, a symbol describing the driver de-emphasis setting which should be used, and five reserved symbols. TS3 is used for negotiating link configuration between two peer ports on a link.

Training Sequence for Test

The sixteen symbol ordered-set which is used by test equipment to place a port link/phy state machine into one of several states used for testing transmitter and receiver compliance with physical layer specifications.

Transparent Retiming Repeater

A device that recovers and retransmits data, to eliminate jitter transfer, and hence perform a jitter reset. (SKP symbols are not added or deleted to SKIP ordered-set)

Transport Service Type

Describes the reliability, sequencing, message size, and operation types that will be used between the communicating Channel Adapters.

Transport service types that use the IBA transport and that pertain to Volume 2 are:

- Unreliable Datagram

See Volume 1 for other Transport Service Types.

Transverse Airflow

As it pertains to the IB Module, transverse airflow is defined as airflow that might predominantly enter the module parallel to the backplane. In a vertical module orientation this direction would be an upper or lower entrance point. Transverse airflow may be directed longitudinally within the module, but the predominant entrance and/or exitpoint for the airflow is in a direction parallel to the backplane.

TS1 Please see Training Sequence 1.

TS2 Please see Training Sequence 2.

TS3 Please see Training Sequence 3.

TS-T Please see Training Sequence for Test.

UI Please see Unit Interval.

Unit One or more sets of processes and/or functions attached to the fabric by one or more channel adapters. Please see Host and I/O Unit.

Unit Interval The time interval equivalent to one bit on the high speed serial link at the active transmission speed.

Unmanaged Chassis An Unmanaged chassis does not implement any IB-specified GUID or Slot Information and leaves the Module IB-ML unconnected.

Unreliable Datagram A Transport Service Type in which a Queue Pair may transmit and receive single-packet messages to/from any other QP. Ordering and delivery are not guaranteed, and delivered packets may be dropped by the receiver.

VA_In The Auxiliary Voltage Input pin (5V nominal) that is defined on the IB Module edge connector and backplane connector.

VB_In The Bulk Voltage Input pins (12V nominal) that are defined on the IB Module edge connector and backplane connector.

Verbs An abstract description of the functionality of a Host Channel Adapter. An operating system may expose some or all of the verb functionality through its programming interface.

Virtual Lane A method of providing independent data streams on the same physical link.

Virtual Memory The address space available to a process running in a system with a memory management unit (MMU). The virtual address space is usually divided into pages, each consisting of 2**N bytes. The bottom N address bits (the offset within a page) are left unchanged, indicating the offset within a page, and the upper bits give a (virtual) page number that is mapped by the MMU to a physical page address. This is recombined with the offset to give the address of a location in physical memory

Vital Product Data Device-specific data to support management functions.

VL Please see Virtual Lane.

VPD — Please see Vital Product Data.

Wake Request Event — Events that can be produced by an IB Module that desires to return from a Sleeping or Polling state to an operation state.

Wander — Deviations at <10kHz from the ideal timing of an event.

Work Queue Element — The Host Channel Adapter's internal representation of a Work Request. The consumer does not have direct access to Work Queue Elements.

Work Queue Pair — See Queue Pair.

Work Request — The means by which a consumer requests the creation of a Work Queue Element.

Workstation, or Client Computer — The class of computers that emphasize numerical and/or graphic performance and provide an interface to a human being.

WRE — Please see Wake Request Event

IBTA

CHAPTER 3: PHYSICAL LAYER OVERVIEW

3.1 INTRODUCTION

This volume defines the low level physical interface protocols, electrical and mechanical specifications for developing applications based on the InfiniBand Architecture. The InfiniBand Architecture supports a range of applications from the backplane interconnect of a single host, to a complex system area network consisting of multiple independent and clustered hosts and I/O components.

In keeping with the layered nature of the InfiniBand Architecture, Figure 1 depicts the structure within the Physical Layer itself.

IBTA

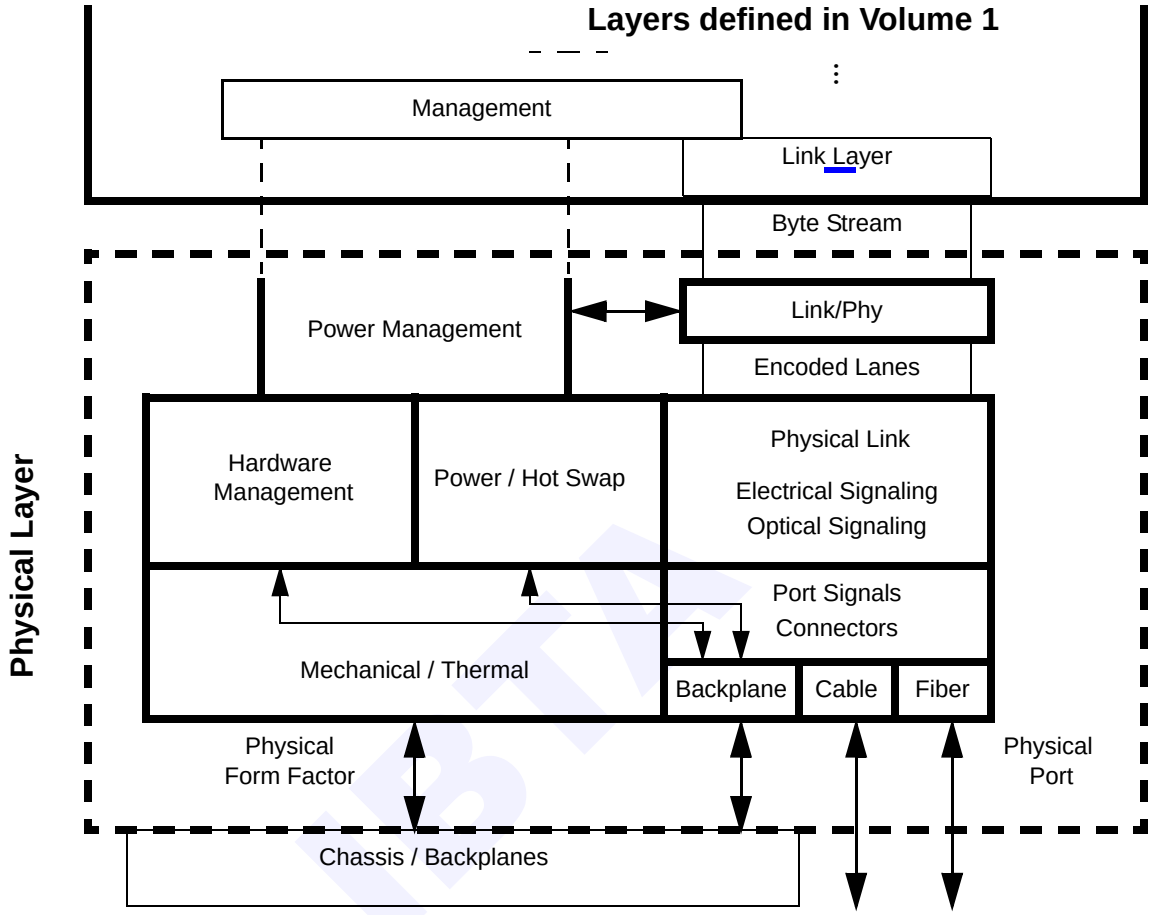


Figure 1 Physical Layer Structure

The basic interconnect of the InfiniBand Architecture is a link (or "physical link"), which is a full duplex transmission path between any two fabrics elements. A fabric is a collection of links, switches, repeaters and routers that connects a set of end nodes. A link physically terminates at a port.

The physical attach point to a port is either:

1) A Cable Connector, which is defined for use for copper cables.
2) A Fiber Connector, which is defined for use for optical cables.
3) A Backplane Connector, which is defined for accepting a specified form factor that houses a function (Channel Adapter, Switch, etc.).

For the form factors, elements of Power and Hardware Management are also specified.

The remaining chapters in this volume provide the detailed specifications to be met in order for devices to properly function on an InfiniBand fabric.

3.2 PHYSICAL PORT

A physical "Port" is a set of signals as seen on a connector interface identified by this specification. The physical ports defined are:

- Backplane Port
- Cable Port
- Fiber Optic Port

Some physical ports contain all signals (e.g. backplane) while others contain a subset (e.g. cable and optics).

A physical port consists of up to four groups of signals which serve different purposes. These groups are:

- Signaling Group
- Hardware Management Group
- Bulk Power Group
- Auxiliary Power Group

Figure 2 depicts a Backplane Port and the signal groupings containing a single physical link made of a number of physical lanes dependent on the link width. Similarly, Figure 3 and Figure 4 depict a Cable Port and a Fiber Optic Port, respectively.

See Chapter 4: Port Signal Definitions for further details on these port types.

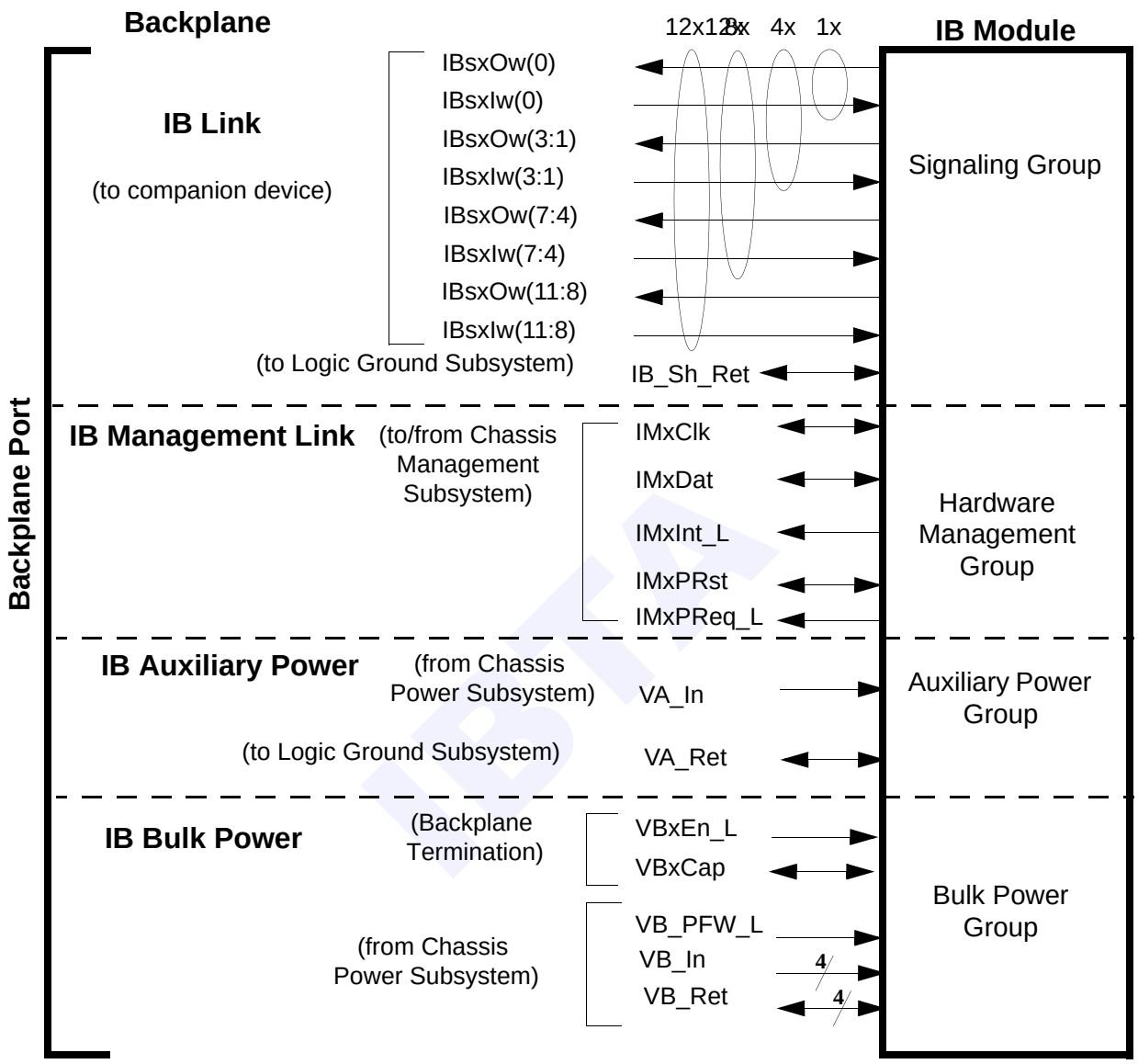


Figure 2 Backplane Port - Single Physical Link

IBTA

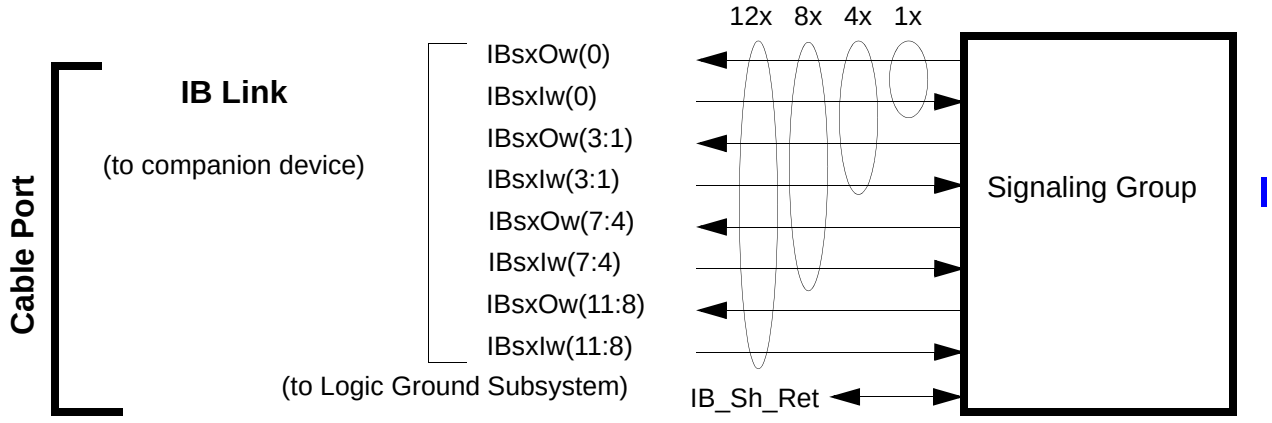


Figure 3 Cable Port

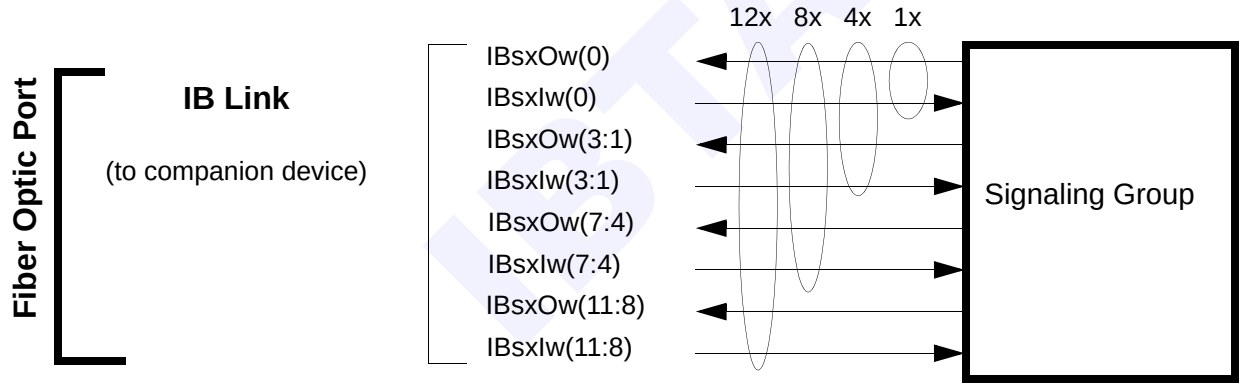


Figure 4 Fiber Optic Port

3.2.1 MULTI-PORTING

This version of the InfiniBand specification supports the utilization of the the backplane connector on the defined form factors in a "multi-ported" manner. As the physical connector, as defined in Chapter 10: Backplane Connector Specification on page 409, supports up to 12 physical lanes, it is possible to implement modules with multiple IBA links (i.e. ports) using only a single connector. Since some of the form factors defined in Chapter 9: Mechanical Specification on page 355 allow for the presence of multiple physical connectors, it is additionally possible to have multiple links on each of the physical connectors present.

Figure 5 depicts a Backplane Port containing multiple independent physical links in a single physical connector; this is also referred to in this specification as being "multi-ported".

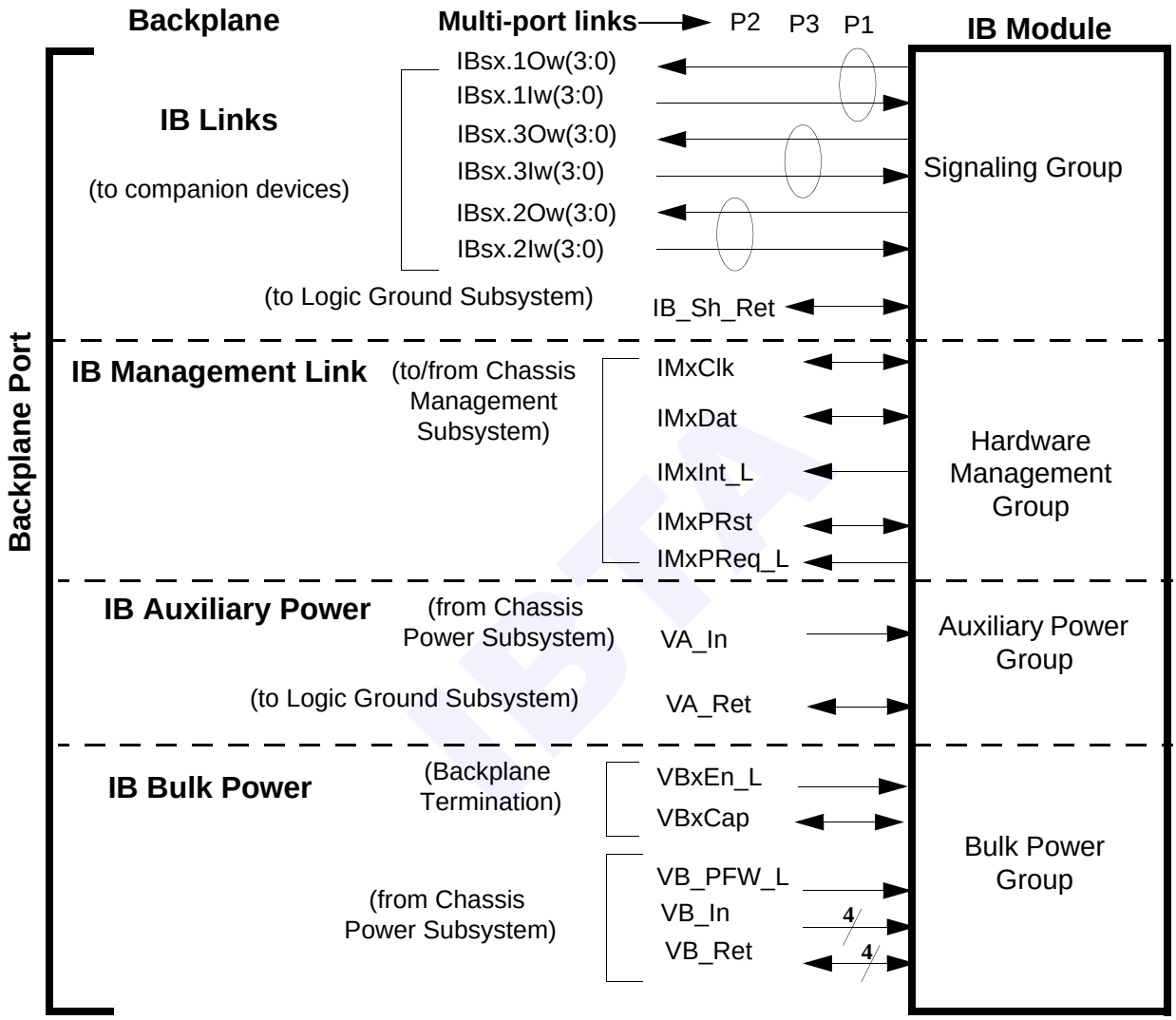


Figure 5 Backplane Port - Multiple Physical Links

A similar configuration of multiple 4x physical links in a 12x connector applies for cables also, as described in Section 7.7.10, "12x to 3-4x Copper Cables," on page 259.

3.2.2 ACTIVE CABLES

This release of this specification allows provision of power to cable connectors, to allow incorporation of active components in the cables, as described in Section 7.8, "Active Cables," on page 264.

3.3 LINK ELECTRICAL SIGNALING

This specification defines the characteristics required to communicate between the output of a port of one InfiniBand node and the input of a port of another InfiniBand node using copper printed wiring and optional cabling. The only signaling rate for encoded data on the media for prior releases is 2.5 Gbits/sec which results in a data rate that can be considered to be 250 MBytes/second per physical lane, at the “SDR” (Single Data Rate) speed. Additional enhanced signaling rates defined in this release are 5.0 Gbits/sec for the “DDR” (Double Data Rate) speed, and 10.0 Gbits/sec for the “QDR” (Quad Data Rate) speed. The connections are point to point and signaling is full duplex, unidirectional.

See Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s for details of SDR (2.5Gb/s), DDR (5.0Gb/s) and QDR (10.0 Gb/s) operation. See Chapter 7: Copper Cable for details of the copper cable and connectors used for electrical links.

3.4 LINK OPTICAL SIGNALING

This specification defines the characteristics required to communicate between the output of a port of one InfiniBand node and the input of a port of another InfiniBand node using optical fiber. The only signaling rate for encoded data on the media for prior releases was 2.5 Gbits/sec which results in a data rate that can be considered to be 250 MBytes/second per physical lane, at the “SDR” (Single Data Rate) speed. Additional enhanced signaling rates described in this release are 5.0 Gbits/sec for the “DDR” (Double Data Rate) speed, and 10.0 Gbits/sec for the “QDR” (Quad Data Rate) speed. The connections are point to point and full duplex, unidirectional.

See Chapter 8: Fiber Attachment - 2.5 Gb/s, 5.0 Gb/s, & 10 Gb/s for specification details of SDR (2.5 Gb/s), DDR (5.0 Gb/s) and QDR (10 Gb/s) operation over optical fiber.

3.5 LINK PHYSICAL LAYER

The Link Physical layer provides the interface between the packet byte stream of Link Layer defined in *InfiniBand Architecture Specification, Volume 1* and the serial bit stream(s) of the physical media. The packet byte stream will be byte striped across the available physical lanes. The byte stream on each physical lane is encoded using the industry standard 8B/10B coding. In addition to encode and decode, the link physical layer includes link training and initialization logic, clock tolerance compensation logic, receive error detecting logic, and link heartbeat logic. This layer is described in Chapter 5: Link/Phy Interface.

3.5.1 SPEED NEGOTIATION

The previous releases of the InfiniBand Architecture supported only 2.5Gb/s signaling. However, provisions were made for potential frequency (speed) increases in a future version of this specification. This release specifies negotiation for the use of SDR (2.5Gb/s), DDR (5.0 Gb/s) and QDR (10.0Gb/s) speeds.

The link speed negotiation mechanism is part of the Link Initialization and Training process. This process provides for two nodes to determine the speed of the interface that will allow for operation at the maximum frequency architecturally supported that is achievable based on endnode capability and interconnect signal integrity.

3.5.2 WIDTH NEGOTIATION

The link width negotiation mechanism is also part of the Link Initialization and Training. This process provides for two nodes to determine the width of the interface that will allow for the maximum bandwidth that is achievable based on endnode capability and interconnect.

This specification defines three interface widths as follows:

1) 1x
   - Electrical: 2 differential pair, 1 per direction for a total of 4 wires
   - Optical: 1 transmit/1 receive per direction for a total of 2 fibers
2) 4x
   - Electrical: 8 differential pair, 4 per direction for a total of 16 wires
   - Optical: 4 transmit / 4 receive per direction for a total of 8 fibers
3) 8x
   - Electrical: 16 differential pair, 8 per direction for a total of 32 wires
   - Optical: 8 transmit / 8 receive per direction for a total of 16 fibers
4) 12x
   - Electrical: 24 differential pair, 12 per direction, for a total of 48 wires
   - Optical: 12 transmit / 12 receive per direction for a total of 24 fibers

3.6 MODULE MECHANICAL

An InfiniBand Module conforms to any of the defined form factors in Chapter 9: Mechanical Specification. A module consists of the following:

a) The module carrier - the basic metal structure

IBTA

b) The module ejector & latch - a handle and latch

c) The module cover - easily removable cover to protect the board(s)

d) The board(s) which implement the module's functionality

This version of the specification defines two module heights (tall and standard) and two module widths. Double width modules occupy two IB defined chassis slots.

As double wide standard and both forms of tall modules (single and double wide) allow for the presence of more than one physical connector, a nomenclature is needed to refer to the connector positions.

Every module contains a connector positioned for the primary slot - primary being defined as the left-most slot the module covers. As shown in Chapter 9: Mechanical Specification, a tall module allow for two connector locations: one in the lower portion and one in the upper portion; lower and upper in this context is when the module is in the vertical orientation. For consistency between standard and tall module designations, the lower (for tall) or only (for standard) location is termed "Primary" and designated "C" by convention. The upper location (for tall) is designated "O" (for optional).

Architectural Note

The most obvious designation for "Primary" would be "P". However, "P" is used in this specification to designate "Ports". The designation of "C" indicates "Primary Connector".

For the case of a standard module, the only connector present is designated C1.

For the case of standard wide modules, there may be a second connector positioned for the adjacent slot - adjacent being defined as the slot to the right of the left-most slot the module covers. The two connectors, both in the "C" location, are designated C1 and C2.

For the case of a tall module, as noted previously, there is a primary connector designated C1 and may optionally have a connector in the upper location designated O1.

For the case of a tall wide module, in addition to the low there may optionally be connectors in the upper locations for both the primary and adjacent positions. The two upper connectors, both in the upper "O" locations are designated O1 and O2.

Within any of the physical connector locations, up to three (3) link ports is defined. The pin designation for these ports is defined in Chapter 10: Backplane Connector Specification.

Figure 6 depicts these designations.

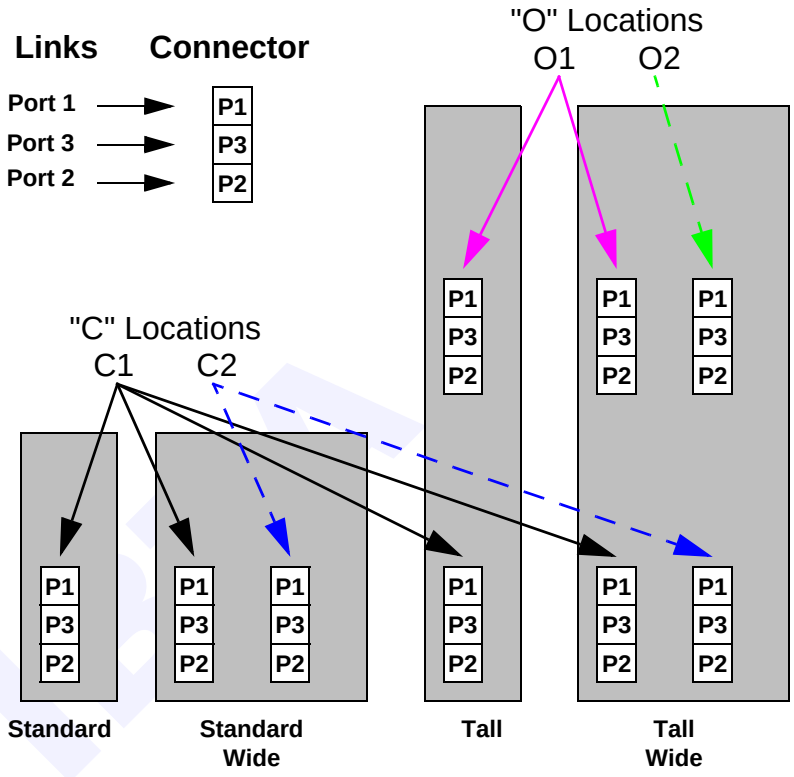


Figure 6 Module Connector Location Designation

3.7 CHASSIS SLOT MECHANICAL

A slot is the volumetric envelope with a specified backplane connector into which one of the defined InfiniBand Modules plug. This specification provides for the enablement of 3U and 6U chassis implementations with modules oriented vertically or horizontally.

Chapter 9: Mechanical Specification provides details for system and chassis designers to implement slots which will accommodate the module form factor options. Only single width slots are physically defined; double width modules occupy two single width slots.

As only single width slots are defined, the connector location designation only indicate whether a connector is in the lower position (for standard and

tall) designated "C" or in the upper position (for tall only) designated "O". The additional designation is for the slot number.

As for the module, any of the physical connector locations present may contain up to three (3) link ports. The pin designation for these ports is defined in Chapter 10: Backplane Connector Specification.

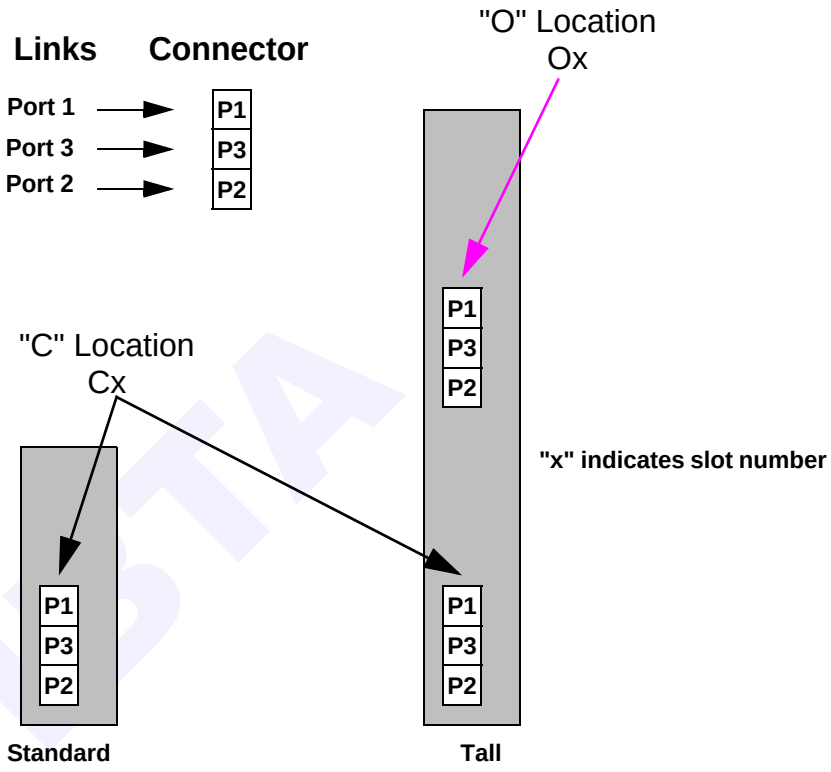


Figure 7 Chassis Connector Location Designation

3.8 POWER

InfiniBand modules are provided two forms of power:

- Bulk Power to perform the major intended functions of the module. This is supplied at nominally 12V and must be converted on the module by DC-DC converter(s) to produce the voltages required by the module's electronics.
- Auxiliary Power to perform management functions to the module even when the Bulk Power is not available to the module. This is supplied at nominally 5V and may be regulated as necessary by the module.

The combination of connector contact staggering and on-module power control circuitry allow for InfiniBand module to be hot plugged, both inserted and removed, without physical damage. Additionally, features are specified to allow for a "Graceful Removal" operation whereby software is notified of a pending removal such that appropriate actions can take place so as to not disrupt operations. LED indicators on the module indicate when an appropriate state has been achieved so that operationally safe removal can be performed.

Please see Chapter 12: Power / Hot Plug for details on power functionality and Chapter 11: Low Speed Electrical Signaling for details on the electrical parameters of the Bulk and Auxiliary Power groups.

3.9 HARDWARE MANAGEMENT

Hardware Management describes the functions that manage, control, and monitor physical components of InfiniBand modules and the Chassis in which they reside. Additionally, xCAs and switches that are packaged in some form factor other than those defined by this specification but optionally provide the hardware management functionality defined are described. The functions defined include:

1) Communication mechanisms for optionally present Baseboard Manager software running on one or more nodes attached to the InfiniBand fabric,

2) Communication mechanisms for an optionally present Chassis Management Entity (CME) which is local managing the physical elements of a chassis,

3) Graceful Hot Removal mechanisms,

4) Standard visual indicators (LEDs) to assist the user in Hot Add and Hot Removal operations,

5) Module Vital Product Data (VPD) accessible to both a Baseboard Manager and a CME,

6) Access to module optional environmental variables.

Baseboard Manager software access these facilities using defined datagrams of the Baseboard class on the InfiniBand fabric. The Chassis Management Entity, typically with firmware, access these facilities through the InfiniBand Management Link (IB-ML) interface defined on the standard InfiniBand backplane connector to a module.

Please see Chapter 13: Hardware Management for details on this functionality.

DRAFT

3.10 OPERATING SYSTEM (OS) POWER MANAGEMENT

Operating System Power Management defines a set of states and facilities that allow for an operating system (or a prescribed agent) to control the power consumption of InfiniBand modules and switches that provide this support. Power management of devices or media not directly attached to the InfiniBand fabric as an addressable node are outside the scope this specification.

The defined states are controlled from the InfiniBand fabric connection of the module. This includes the ability to "power down" to a state only drawing auxiliary power and "power up" from activity on the InfiniBand link. Further, in the presence of a fully operational link, states are defined to allow modules varied levels of power consumption.

Please see Chapter 14: OS Power Management for details on this functionality.

IBTA

CHAPTER 4: PORT SIGNAL DEFINITIONS

4.1 SIGNAL NAMING CONVENTIONS

The symbol substitution and type notation used for the signal naming within this specification are defined in Table 2 and Table 3.

Table 2 Symbol Substitution

Symbol Convention	Substitution
w in Signal Name	Differential pair distinction - IB Signaling Group Replaced by 'p' for positive rail of differential pair Replaced by 'n' for negative rail of differential pair
x in Signal Name	Replaced by respective board or backplane Port x may range from 1 to n where n is the number of connectors on a board edge or backplane slot
s in Signal Name	For Boards, replaced by 't' for I/O Plate connection, or 'b' for backplane connection. For Backplanes, replaced by slot number. Slots are numbered from 1 to n from left to right, or bottom to top for horizontal mounting

Table 3 Type Notation

Signal Type	Definition
InDiff	Differential input signal - IB signaling levels
OutDiff	Differential output signal - IB signaling levels
ShRet	Shield Return - Shield Ground for both the IB board and the IB backplane
Rx	Optical Receiver Input
Tx	Optical Transmitter Output
PwrU	Pull Up derived from 12V Bulk power input (**VB_In**)
PwrD	Pull Down to Bulk power return (**VB_Ret**)
APwrU	Pull Up to Auxilliary power input (**VA_In**)
APwrD	Pull Down to Auxiliary power return (**VA_Ret**)
InOut	Open Drain - Wired OR Input/Output pull-up referenced to Auxiliary power (**VA_In**)
In	Input - logic is powered from Auxiliary power (**VA_In**)

Table 3 Type Notation

Signal Type	Definition
Out	Output - derived from Auxiliary power (**VA_In**)
PwrIn	12 Volt Input from backplane/system power supply
PwrRet	12 Volt Return to backplane/system power supply
APwrIn	5 Volt Input from backplane/system power supply
APwrRet	5 Volt Return to backplane/system power supply

Signal names ending with "_**L**" indicates the "asserted" or "true" condition is a low voltage.

4.2 PORT SIGNAL SUMMARY

This section summarizes the signals that comprise the physical ports defined by this specification.

IBTA

4.2.1 BACKPLANE PORT

Table 4 summarizes the signals of backplane port.

Table 4 Backplane Port Signal Summary

Interface				Signal Name	Signal Type	Mate Order[a]	Break Order[b]	Number of Contacts	Description
Signaling Group - High Speed Differential									
12x	8x	4x	1x	**IBsxIw(0)**	InDiff	High Speed	High Speed	2 1 pair	IB Symbol Input Differential Signaling
				IBsxOw(0)	OutDiff	High Speed	High Speed	2 1 pair	IB Symbol Output Differential Signaling
				IBsxIw(3:1)	InDiff	High Speed	High Speed	6 3 pair	IB Symbol Input Differential Signaling
				IBsxOw(3:1)	OutDiff	High Speed	High Speed	6 3 pair	IB Symbol Output Differential Signaling
				IBsxIw(7:4)	InDiff	High Speed	High Speed	8 4 pair	IB Symbol Input Differential Signaling
				IBsxOw(7:4)	OutDiff	High Speed	High Speed	8 4 pair	IB Symbol Output Differential Signaling
				IBsxIw(11:8)	InDiff	High Speed	High Speed	8 4 pair	IB Symbol Input Differential Signaling
				IBsxOw(11:8)	OutDiff	High Speed	High Speed	8 4 pair	IB Symbol Output Differential Signaling
				IB_Sh_Ret	ShRet	High Speed	High Speed	24	Shield Return tied to Logic Ground
Hardware Management Group									
				IMxClk	InOut	3	2	1	IB Management Link Clock
				IMxDat	InOut	3	2	1	IB Management Link Data
				IMxInt_L	Out	3	2	1	IB Management Link Interrupt
				IMxPRst	InOut	4	1	1	Presence/Reset
				IMxPReq_L	Out	3	2	1	Power Request

Table 4 Backplane Port Signal Summary

Interface	Signal Name	Signal Type	Mate Order[a]	Break Order[b]	Number of Contacts	Description
Bulk Power Group						
	VB_In[c]	PwrIn	2	3	4	12 Volt Bulk Power Input
	VB_Ret	PwrRet	1	4	4	12 Volt Bulk Power Return
	VBxEn_L	PwrU / PwrD	4	1	1	Bulk Power Enable Not Asserted - Disable Asserted - Enable
	VBxCap	APwrU / APwrD	3	2	1	Bulk Power Capability Not Asserted - 25W; Asserted - 50W
	VBxPFW_L	In	3	2	1	Power Fail Warning
Auxilliary Power Group						
	VA_In[c]	APwrIn	2	3	1	5 Volt Auxiliary Power Input
	VA_Ret[c]	APwrRet	1	4	1	Logic Ground

a. See the *InfiniBand Architecture Specification, Volume 2*, Section 12.6.1 on page 471 for detailed description
b. See the *InfiniBand Architecture Specification, Volume 2*, Section 12.6.2 on page 473 and Section 12.6.3 on page 475 for detailed description
c. This signal may or may not be specific to a port. If it is not specific to a port (i.e, it is common among multiple ports), the "_" is used in the signal name; if it is specific to a port, then the "_" shown should be replaced with the port number in the same manner as "x" is used for other signals as described in Table 2 Symbol Substitution on page 69. For documentation clarity, the "_" is used through the remainder of this specification.

As shown in Figure 2 Backplane Port - Single Physical Link on page 59, the Signaling Group signals make up the IB Link for the width supported by a backplane port. The ports are physically located on the defined module form factors and the backplanes into which they plug as specified in Chapter 9: Mechanical Specification. Some of the module form factors defined allow for multiple backplane ports to be present. However, one is always required.

C4-1: In order to establish an IB Link, all modules **shall** have at least one backplane port containing the Signaling Group signals defined in Table 4 located at Primary Port (1) as defined by Figure 154 Module Bulk Power Ports (Logical) on page 460 and Section 12.5, "Chassis Power Rules," on page 470.

C4-2: In order to establish an IB Link, all chassis that accept InfiniBand modules **shall** have at least a backplane port containing the Signaling Group signals defined in Table 4 located at Primary Port (1) as defined by Figure 154 Module Bulk Power Ports (Logical) on page 460 and Section 12.5, "Chassis Power Rules," on page 470.

Primary Port (1) requirements for the Hardware Management Group are found in Chapter 13: Hardware Management. Requirements for the Bulk Power Group and Auxiliary Power Group are found in Chapter 12: Power / Hot Plug.

The backplane connector is specified in Chapter 10: Backplane Connector Specification.

4.2.2 CABLE PORT

Table 5 summarizes the signals of a cable port. Cable port requirements are found in Chapter 7: Copper Cable.

Table 5 Cable Port Signal Summary

Interface				Signal Name	Signal Type	Number of Pins	Description
Signaling Group - High Speed Differential							
12x	8x	4x	1x	**IBsxIw(0)**	InDiff	2 1 pair	IB Symbol Input Differential Signaling
				IBsxOw(0)	OutDiff	2 1 pair	IB Symbol Output Differential Signaling
				IBsxIw(3:1)	InDiff	6 3 pair	IB Symbol Input Differential Signaling
				IBsxOw(3:1)	OutDiff	6 3 pair	IB Symbol Output Differential Signaling
				IBsxIw(7:4)	InDiff	8 4 pair	IB Symbol Input Differential Signaling
				IBsxOw(7:4)	OutDiff	8 4 pair	IB Symbol Output Differential Signaling
				IBsxIw(11:8)	InDiff	8 4 pair	IB Symbol Input Differential Signaling
				IBsxOw(11:8)	OutDiff	8 4 pair	IB Symbol Output Differential Signaling
				IB_Sh_Ret	ShRet		Inner Shield Return tied to Logic Ground

4.2.3 FIBER OPTIC PORT

Table 6 summarizes the signals of a fiber optic port. Fiber optic port requirements are found in Chapter 8: Fiber Attachment - 2.5 Gb/s, 5.0 Gb/s, & 10 Gb/s.

Table 6 Fiber Optic Port Signal Summary

Interface				Signal Name[a]	Signal Type	Number of Fibers	Description
Signaling Group - High Speed Optical							
12x	8x	4x	1x	**IBsxIp(0)**	Rx	1	IB Symbol Input
				IBsxOp(0)	Tx	1	IB Symbol Output
				IBsxIp(3:1)	Rx	3	IB Symbol Input
				IBsxOp(3:1)	Tx	3	IB Symbol Output
				IBsxIp(7:4)	Rx	4	IB Symbol Input
				IBsxOp(7:4)	Tx	4	IB Symbol Output
				IBsxIp(11:8)	Rx	4	IB Symbol Input
				IBsxOp(11:8)	Tx	4	IB Symbol Output

a. As each lane in made of a single fiber per direction versus a differential pair per direction for electrical interfaces, the “w” portion of the signal name only will have one polarity present. By convention, the “p” (positive) polarity is used.

IBTA

4.2.4 ACTIVE CABLE PORT

Table 7 summarizes the signals of an active cable port. Active Cable port requirements are found in Section 7.8, "Active Cables," on page 264.

Table 7 Active Cable Port Signal Summary

Interface				Signal Name	Signal Type	Number of Pins	Description
Signaling Group - High Speed Differential							
12x	8x	4x	1x	**IBsxIw(0)**	InDiff	2 1 pair	IB Symbol Input Differential Signaling
				IBsxOw(0)	OutDiff	2 1 pair	IB Symbol Output Differential Signaling
				IBsxIw(3:1)	InDiff	6 3 pair	IB Symbol Input Differential Signaling
				IBsxOw(3:1)	OutDiff	6 3 pair	IB Symbol Output Differential Signaling
				IBsxIw(7:4)	InDiff	8 4 pair	IB Symbol Input Differential Signaling
				IBsxOw(7:4)	OutDiff	8 4 pair	IB Symbol Output Differential Signaling
				IBsxIw(11:8)	InDiff	8 4 pair	IB Symbol Input Differential Signaling
				IBsxOw(11:8)	OutDiff	8 4 pair	IB Symbol Output Differential Signaling
				IB_Sh_Ret	ShRet		Inner Shield Return tied to Logic Ground
Active Cable Power Group							
12x	8x	4x		**Sense-3.3V**	In	1	Detection of Active Cable operating at 3.3V
12x	8x	4x		**Sense-12V**	In	1	Detection of Active Cable operating at 12V
12x	8x			**Vcc**	PwrIn	1	3.3Volt or 12Volt Bulk Power Input (for 4x). Return is through IB_Sh_Ret Logic Ground
12x	8x			**Vcc**	PwrIn	2	3.3V or 12V Bulk Power Input (for 8x & 12x). Return is through IB_Sh_Ret Logic Ground

4.3 SIGNALING GROUP

This group is comprised of the high speed signals that make up the InfiniBand link. The link consists of two uni-directional interfaces, one operating as an input (or receiver) and one operating as an output (or driver). These are **IBsxIw** and **IBsxOw**, respectively.

4.3.1 HIGH SPEED ELECTRICAL

Based on the width of interface supported, the number of signals within the group varies: 1x consists of 4 conductors (1 differential pair in, 1 differential pair out); 4x consists of 16 conductors (4 differential pair in, 4 differential pair out); 8x consists of 32 conductors (8 differential pair in, 8 differential pair out); and 12x consists of 48 conductors (12 differential pair in, 12 differential pair out).

These signals conform to the electrical signaling as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s.

4.3.2 HIGH SPEED OPTICAL

Based on the width of interface supported, the number of signals within the group varies: 1x consists of 2 fibers (1 in, 1 out); 4x consists of 8 fibers (4 in, 4 out); 8x consists of 16 fibers (8 in, 8 out); and 12x consists of 24 fibers (12 in, 12 out).

These signals conform to the electrical signaling as defined in Chapter 8: Fiber Attachment - 2.5 Gb/s, 5.0 Gb/s, & 10 Gb/s.

4.4 HARDWARE MANAGEMENT GROUP

The Hardware Management Group provides a serial management interface which supports communication between the module and an optional intelligent environmental controller (hereafter called the Chassis Management Element or "CME") associated with the IB backplane. It also includes additional signals to allow for low level interaction between the chassis backplanes and the module.

The InfiniBand Management Link, also known as IB-ML, is made up by the **IMxDat** and **IMxClk** signals and allow access to a number of architected facilities that provide identification, status, and control of hardware management features.

The **IMxInt_L** signal originates from the module and allows the module to provide an interrupt notification of a condition that requires backplane attention. IB-ML operations are used in response to this signal to determine the event type.

The **IMxPRst** signal is a bidirectional signal that allows the backplane to detect module presence or force module reset

The **IMxPReq_L** signal originates from the module and allows the module to request that bulk power be supplied.

These signals conform to the electrical signaling as defined in Chapter 11: Low Speed Electrical Signaling.

4.5 BULK POWER GROUP

The Bulk Power Group provides the majority of the IB module's operational power. In particular, the IB Signaling Group and the IB module application (i.e. Fibre Channel, Gigabit Ethernet, etc.) are powered only by the Bulk Power Group. Power is delivered at nominally 12V.

The **VB_In** and **VB_Ret** connections drive an on-board power sequencer that controls power to the module's local DC-DC power converter(s). The IB standard requires an on module power sequencer capable of:

- Sensing full insertion of the module
- Checking power required by a board and the power provided by the backplane
- Enabling board power-up on board/backplane power compatibility
- Power-up override by a Chassis Management Element (CME) associated with a managed backplane (See Chapter 13: Hardware Management for information on Chassis Management Elements as it pertains to this specification.)

These capabilities are provided by the **VBxEn_L** and **VBxCap** pins.

The **VBxPFW_L** signal originates from the system or chassis power supply and provides an indication that the module's bulk power input may be about to go out of specified tolerance.

These signals conform to the electrical signaling as defined in Chapter 11: Low Speed Electrical Signaling. Further details of the usage of the Bulk Power group can be found in Chapter 12: Power / Hot Plug of this specification.

4.6 AUXILIARY POWER GROUP

The Auxiliary Power Group provides a low power, yet always available supply from which certain management functions can be available in the absence of Bulk Power. Functions include module information access through IB-ML (See Chapter 13: Hardware Management), beacon sequence detection for in-band Power Management (See Chapter 14: OS Power Management, Chapter 5: Link/Phy Interface), and detection or generation of Wake Request Events (See Chapter 14: OS Power Management).

Power is delivered at nominally 5V.

The **VA_In** connection provides the power. The return for this supply is **VA_Ret** which may be tied to logic ground.

These signals conform to the electrical signaling as defined in Chapter 11: Low Speed Electrical Signaling. Further details of the usage of the Auxiliary Power group can be found in Chapter 12: Power / Hot Plug of this specification.

4.7 ACTIVE CABLE POWER GROUP

The Active Cable Power Group provides power for active cables, which are specified in Section 7.8, "Active Cables," on page 264.

Two **Sense** pins are used to distinguish Active Cables operating at either 3.3V or 12V from passive (non-powered) cables, which use the same physical connector. The **VCC** pins provide power for active components in the cable, with return through signal ground. Power is delivered nominally at either 12V or 3.3V, as determined by needs of the active cable assembly.

IBTA

CHAPTER 5: LINK/PHY INTERFACE

5.1 INTRODUCTION

The Link Physical layer provides an interface between the packet byte stream of upper layers and the serial bit stream(s) of the physical media. The physical media may be implemented as 1, 4, 8, or 12 physical lanes. The packet byte stream will be byte striped across the available physical lanes. (See Figure 8) The byte stream on each physical lane is encoded using the industry standard 8B/10B coding. In addition to encode and de-code, the link physical layer includes link training and initialization logic, clock tolerance compensation logic, and receive error detecting logic.

The Transmit Data Flow is responsible for:

- Insertion of control sequence information
- 8B/10B encoding

The Receive Data Flow is responsible for:

- Deletion of control sequence information
- Error detection and handling
- 8B/10B decoding

Figure 8 shows a block diagram of the link physical layer.

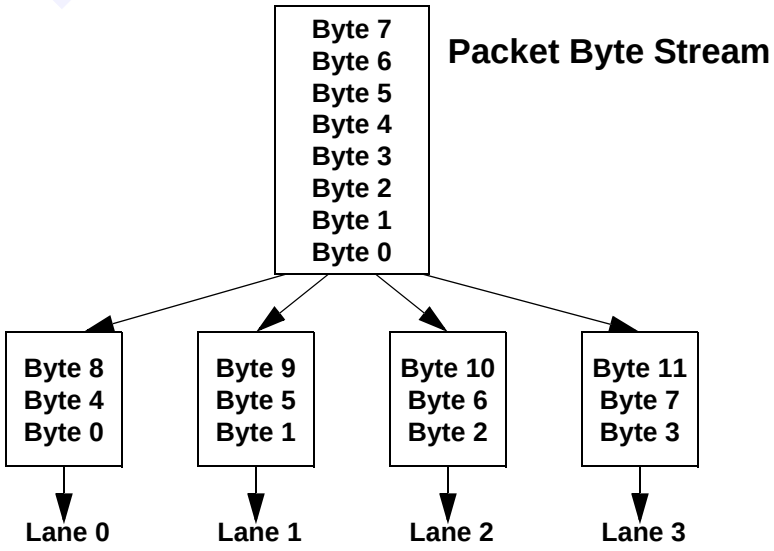


Figure 8 Example 4x Byte Striping Diagram

IBTA

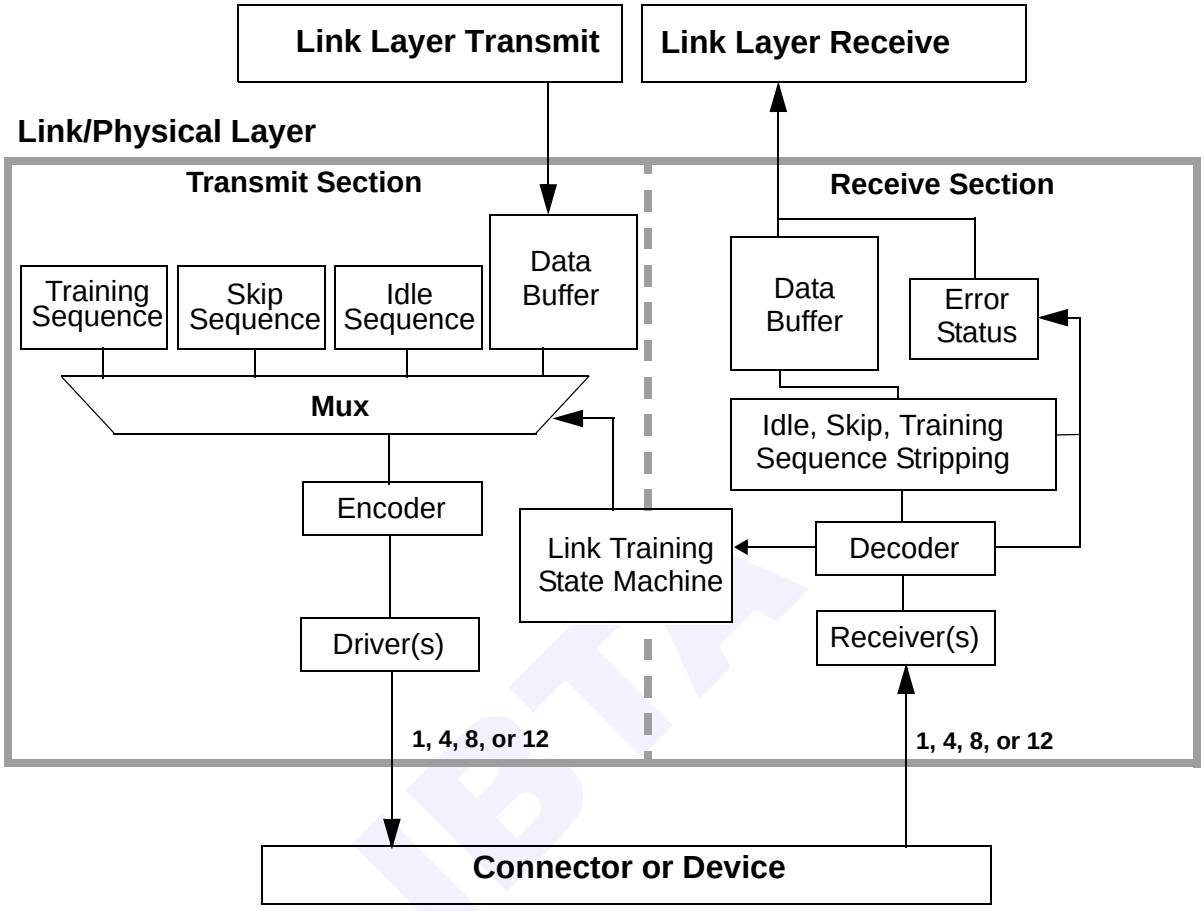


Figure 9 Link/Physical Interface Block Diagram

5.2 SYMBOL ENCODING (8B/10B CODING)

The InfiniBand physical lane encoding uses the industry standard 8B/10B code which is used by Fibre Channel, Gigabit Ethernet (IEEE 802.3z), FICON, and ServerNet. The 8B/10B code provides DC balance, limited run lengths, byte (symbol) synchronization, and the ability to distinguish between data characters and control characters.

C5-1: All ports **shall** use the industry standard 8B/10B code as defined in Section 5.2, “Symbol Encoding (8B/10B coding),” on page 80.

5.2.1 NOTATION CONVENTIONS

The 8B/10B transmission code uses letter notation for describing the bits of an un-encoded information byte and a single control variable. Each bit of the un-encoded information byte contains either a binary zero or a bi-

nary one. A control variable, Z, has either the value D or the value K. When the control variable associated with an un-encoded information byte contains the value D, the associated encoded code-group is referred to as a data code-group. When the control variable associated with an un-encoded information byte contains the value K, the associated encoded code-group is referred to as a special code-group.

The bit notation of A,B,C,D,E,F,G,H for an un-encoded information byte is used in the description of the 8B/10B transmission code. The bits A,B,C,D,E,F,G,H are translated to bits a,b,c,d,e,i,f,g,h,j of 10-bit transmission code-groups. Figure 10 illustrates the byte and bit nomenclature as a byte stream is encoded (decoded) and serialized (de-serialized). 8B/10B code-group bit assignments are illustrated in Table 8 and Table 9. Each valid code-group has been given a name using the following convention:

1) /Dx.y/ for the 256 valid data code-groups, and
2) /Kx.y/ for the special control code-groups where x is the decimal value of bits EDCBA, and y is the decimal value of bits HGF.

Examples of this are D10.2 or K28.5.

IBTA

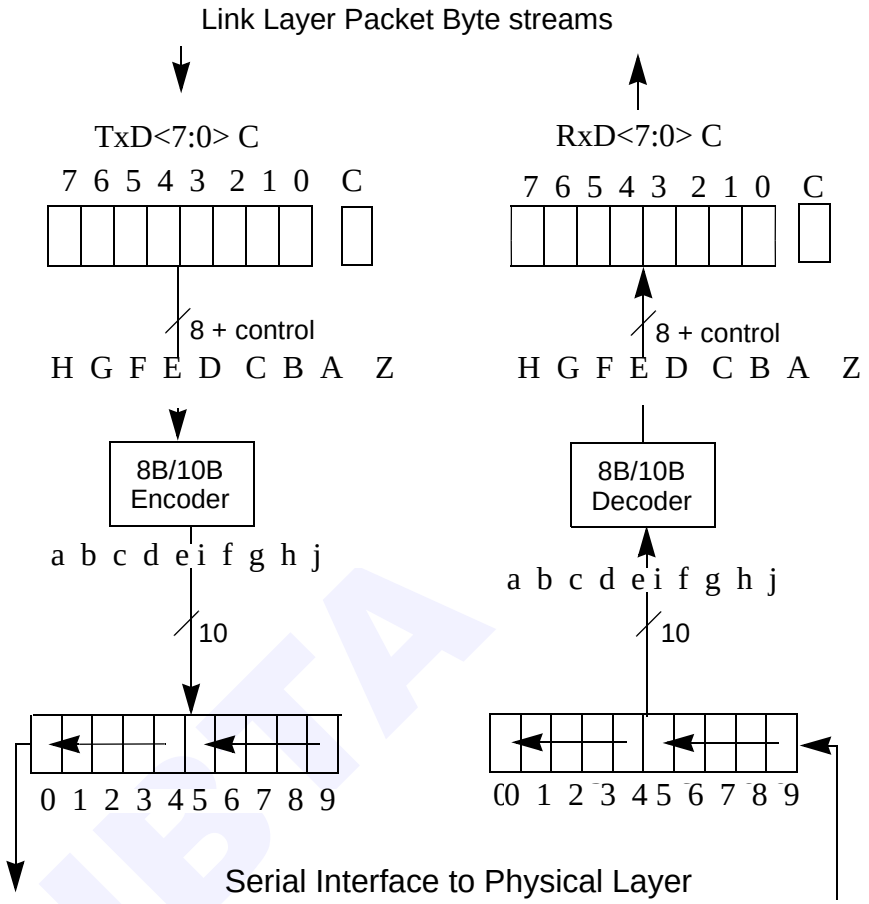


Figure 10 Transmit & Receive Data Ordering

5.2.2 VALID AND INVALID CODE-GROUPS

Table 8 defines the valid data code-groups (D code-groups) of the 8B/10B transmission code. Table 9 defines the valid special code-groups (K code-groups) of the code. The tables are used both for generating valid code-groups (encoding) and for checking the validity of received code-groups (decoding).

In the tables, each byte entry has two columns that represent two code-groups which are not necessarily different. The two columns correspond to the valid code-group based on the current value of the running disparity (Current RD - or Current RD +). Running disparity is a binary parameter with either the value negative (-) or the value positive (+). See Section 5.2.3, “Running disparity rules,” on page 91 for the definition of and rules for disparity.

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D0.0	00	000 00000	100111 0100	011000 1011
D1.0	01	000 00001	011101 0100	100010 1011
D2.0	02	000 00010	101101 0100	010010 1011
D3.0	03	000 00011	110001 1011	110001 0100
D4.0	04	000 00100	110101 0100	001010 1011
D5.0	05	000 00101	101001 1011	101001 0100
D6.0	06	000 00110	011001 1011	011001 0100
D7.0	07	000 00111	111000 1011	000111 0100
D8.0	08	000 01000	111001 0100	000110 1011
D9.0	09	000 01001	100101 1011	100101 0100
D10.0	0A	000 01010	010101 1011	010101 0100
D11.0	0B	000 01011	110100 1011	110100 0100
D12.0	0C	000 01100	001101 1011	001101 0100
D13.0	0D	000 01101	101100 1011	101100 0100
D14.0	0E	000 01110	011100 1011	011100 0100
D15.0	0F	000 01111	010111 0100	101000 1011
D16.0	10	000 10000	011011 0100	100100 1011
D17.0	11	000 10001	100011 1011	100011 0100
D18.0	12	000 10010	010011 1011	010011 0100
D19.0	13	000 10011	110010 1011	110010 0100
D20.0	14	000 10100	001011 1011	001011 0100
D21.0	15	000 10101	101010 1011	101010 0100
D22.0	16	000 10110	011010 1011	011010 0100
D23.0	17	000 10111	111010 0100	000101 1011
D24.0	18	000 11000	110011 0100	001100 1011
D25.0	19	000 11001	100110 1011	100110 0100
D26.0	1A	000 11010	010110 1011	010110 0100
D27.0	1B	000 11011	110110 0100	001001 1011
D28.0	1C	000 11100	001110 1011	001110 0100
D29.0	1D	000 11101	101110 0100	010001 1011
D30.0	1E	000 11110	011110 0100	100001 1011

IBTA

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D31.0	1F	000 11111	101011 0100	010100 1011
D0.1	20	001 00000	100111 1001	011000 1001
D1.1	21	001 00001	011101 1001	100010 1001
D2.1	22	001 00010	101101 1001	010010 1001
D3.1	23	001 00011	110001 1001	110001 1001
D4.1	24	001 00100	110101 1001	001010 1001
D5.1	25	001 00101	101001 1001	101001 1001
D6.1	26	001 00110	011001 1001	011001 1001
D7.1	27	001 00111	111000 1001	000111 1001
D8.1	28	001 01000	111001 1001	000110 1001
D9.1	29	001 01001	100101 1001	100101 1001
D10.1	2A	001 01010	010101 1001	010101 1001
D11.1	2B	001 01011	110100 1001	110100 1001
D12.1	2C	001 01100	001101 1001	001101 1001
D13.1	2D	001 01101	101100 1001	101100 1001
D14.1	2E	001 01110	011100 1001	011100 1001
D15.1	2F	001 01111	010111 1001	101000 1001
D16.1	30	001 10000	011011 1001	100100 1001
D17.1	31	001 10001	100011 1001	100011 1001
D18.1	32	001 10010	010011 1001	010011 1001
D19.1	33	001 10011	110010 1001	110010 1001
D20.1	34	001 10100	001011 1001	001011 1001
D21.1	35	001 10101	101010 1001	101010 1001
D22.1	36	001 10110	011010 1001	011010 1001
D23.1	37	001 10111	111010 1001	000101 1001
D24.1	38	001 11000	110011 1001	001100 1001
D25.1	39	001 11001	100110 1001	100110 1001
D26.1	3A	001 11010	010110 1001	010110 1001
D27.1	3B	001 11011	110110 1001	001001 1001
D28.1	3C	001 11100	001110 1001	001110 1001
D29.1	3D	001 11101	101110 1001	010001 1001
D30.1	3E	001 11110	011110 1001	100001 1001
D31.1	3F	001 11111	101011 1001	010100 1001

IBTA

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D0.2	40	010 00000	100111 0101	011000 0101
D1.2	41	010 00001	011101 0101	100010 0101
D2.2	42	010 00010	101101 0101	010010 0101
D3.2	43	010 00011	110001 0101	110001 0101
D4.2	44	010 00100	110101 0101	001010 0101
D5.2	45	010 00101	101001 0101	101001 0101
D6.2	46	010 00110	011001 0101	011001 0101
D7.2	47	010 00111	111000 0101	000111 0101
D8.2	48	010 01000	111001 0101	000110 0101
D9.2	49	010 01001	100101 0101	100101 0101
D10.2	4A	010 01010	010101 0101	010101 0101
D11.2	4B	010 01011	110100 0101	110100 0101
D12.2	4C	010 01100	001101 0101	001101 0101
D13.2	4D	010 01101	101100 0101	101100 0101
D14.2	4E	010 01110	011100 0101	011100 0101
D15.2	4F	010 01111	010111 0101	101000 0101
D16.2	50	010 10000	011011 0101	100100 0101
D17.2	51	010 10001	100011 0101	100011 0101
D18.2	52	010 10010	010011 0101	010011 0101
D19.2	53	010 10011	110010 0101	110010 0101
D20.2	54	010 10100	001011 0101	001011 0101
D21.2	55	010 10101	101010 0101	101010 0101
D22.2	56	010 10110	011010 0101	011010 0101
D23.2	57	010 10111	111010 0101	000101 0101
D24.2	58	010 11000	110011 0101	001100 0101
D25.2	59	010 11001	100110 0101	100110 0101
D26.2	5A	010 11010	010110 0101	010110 0101
D27.2	5B	010 11011	110110 0101	001001 0101
D28.2	5C	010 11100	001110 0101	001110 0101
D29.2	5D	010 11101	101110 0101	010001 0101
D30.2	5E	010 11110	011110 0101	100001 0101
D31.2	5F	010 11111	101011 0101	010100 0101
D0.3	60	011 00000	100111 0011	011000 1100

IBTA

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D1.3	61	011 00001	011101 0011	100010 1100
D2.3	62	011 00010	101101 0011	010010 1100
D3.3	63	011 00011	110001 1100	110001 0011
D4.3	64	011 00100	110101 0011	001010 1100
D5.3	65	011 00101	101001 1100	101001 0011
D6.3	66	011 00110	011001 1100	011001 0011
D7.3	67	011 00111	111000 1100	000111 0011
D8.3	68	011 01000	111001 0011	000110 1100
D9.3	69	011 01001	100101 1100	100101 0011
D10.3	6A	011 01010	010101 1100	010101 0011
D11.3	6B	011 01011	110100 1100	110100 0011
D12.3	6C	011 01100	001101 1100	001101 0011
D13.3	6D	011 01101	101100 1100	101100 0011
D14.3	6E	011 01110	011100 1100	011100 0011
D15.3	6F	011 01111	010111 0011	101000 1100
D16.3	70	011 10000	011011 0011	100100 1100
D17.3	71	011 10001	100011 1100	100011 0011
D18.3	72	011 10010	010011 1100	010011 0011
D19.3	73	011 10011	110010 1100	110010 0011
D20.3	74	011 10100	001011 1100	001011 0011
D21.3	75	011 10101	101010 1100	101010 0011
D22.3	76	011 10110	011010 1100	011010 0011
D23.3	77	011 10111	111010 0011	000101 1100
D24.3	78	011 11000	110011 0011	001100 1100
D25.3	79	011 11001	100110 1100	100110 0011
D26.3	7A	011 11010	010110 1100	010110 0011
D27.3	7B	011 11011	110110 0011	001001 1100
D28.3	7C	011 11100	001110 1100	001110 0011
D29.3	7D	011 11101	101110 0011	010001 1100
D30.3	7E	011 11110	011110 0011	100001 1100
D31.3	7F	011 11111	101011 0011	010100 1100
D0.4	80	100 00000	100111 0010	011000 1101
D1.4	81	100 00001	011101 0010	100010 1101

IBTA

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D2.4	82	100 00010	101101 0010	010010 1101
D3.4	83	100 00011	110001 1101	110001 0010
D4.4	84	100 00100	110101 0010	001010 1101
D5.4	85	100 00101	101001 1101	101001 0010
D6.4	86	100 00110	011001 1101	011001 0010
D7.4	87	100 00111	111000 1101	000111 0010
D8.4	88	100 01000	111001 0010	000110 1101
D9.4	89	100 01001	100101 1101	100101 0010
D10.4	8A	100 01010	010101 1101	010101 0010
D11.4	8B	100 01011	110100 1101	110100 0010
D12.4	8C	100 01100	001101 1101	001101 0010
D13.4	8D	100 01101	101100 1101	101100 0010
D14.4	8E	100 01110	011100 1101	011100 0010
D15.4	8F	100 01111	010111 0010	101000 1101
D16.4	90	100 10000	011011 0010	100100 1101
D17.4	91	100 10001	100011 1101	100011 0010
D18.4	92	100 10010	010011 1101	010011 0010
D19.4	93	100 10011	110010 1101	110010 0010
D20.4	94	100 10100	001011 1101	001011 0010
D21.4	95	100 10101	101010 1101	101010 0010
D22.4	96	100 10110	011010 1101	011010 0010
D23.4	97	100 10111	111010 0010	000101 1101
D24.4	98	100 11000	110011 0010	001100 1101
D25.4	99	100 11001	100110 1101	100110 0010
D26.4	9A	100 11010	010110 1101	010110 0010
D27.4	9B	100 11011	110110 0010	001001 1101
D28.4	9C	100 11100	001110 1101	001110 0010
D29.4	9D	100 11101	101110 0010	010001 1101
D30.4	9E	100 11110	011110 0010	100001 1101
D31.4	9F	100 11111	101011 0010	010100 1101
D0.5	A0	101 00000	100111 1010	011000 1010
D1.5	A1	101 00001	011101 1010	100010 1010
D2.5	A2	101 00010	101101 1010	010010 1010

IBTA

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D3.5	A3	101 00011	110001 1010	110001 1010
D4.5	A4	101 00100	110101 1010	001010 1010
D5.5	A5	101 00101	101001 1010	101001 1010
D6.5	A6	101 00110	011001 1010	011001 1010
D7.5	A7	101 00111	111000 1010	000111 1010
D8.5	A8	101 01000	111001 1010	000110 1010
D9.5	A9	101 01001	100101 1010	100101 1010
D10.5	AA	101 01010	010101 1010	010101 1010
D11.5	AB	101 01011	110100 1010	110100 1010
D12.5	AC	101 01100	001101 1010	001101 1010
D13.5	AD	101 01101	101100 1010	101100 1010
D14.5	AE	101 01110	011100 1010	011100 1010
D15.5	AF	101 01111	010111 1010	101000 1010
D16.5	B0	101 10000	011011 1010	100100 1010
D17.5	B1	101 10001	100011 1010	100011 1010
D18.5	B2	101 10010	010011 1010	010011 1010
D19.5	B3	101 10011	110010 1010	110010 1010
D20.5	B4	101 10100	001011 1010	001011 1010
D21.5	B5	101 10101	101010 1010	101010 1010
D22.5	B6	101 10110	011010 1010	011010 1010
D23.5	B7	101 10111	111010 1010	000101 1010
D24.5	B8	101 11000	110011 1010	001100 1010
D25.5	B9	101 11001	100110 1010	100110 1010
D26.5	BA	101 11010	010110 1010	010110 1010
D27.5	BB	101 11011	110110 1010	001001 1010
D28.5	BC	101 11100	001110 1010	001110 1010
D29.5	BD	101 11101	101110 1010	010001 1010
D30.5	BE	101 11110	011110 1010	100001 1010
D31.5	BF	101 11111	101011 1010	010100 1010
D0.6	C0	110 00000	100111 0110	011000 0110
D1.6	C1	110 00001	011101 0110	100010 0110
D2.6	C2	110 00010	101101 0110	010010 0110
D3.6	C3	110 00011	110001 0110	110001 0110

IBTA

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D4.6	C4	110 00100	110101 0110	001010 0110
D5.6	C5	110 00101	101001 0110	101001 0110
D6.6	C6	110 00110	011001 0110	011001 0110
D7.6	C7	110 00111	111000 0110	000111 0110
D8.6	C8	110 01000	111001 0110	000110 0110
D9.6	C9	110 01001	100101 0110	100101 0110
D10.6	CA	110 01010	010101 0110	010101 0110
D11.6	CB	110 01011	110100 0110	110100 0110
D12.6	CC	110 01100	001101 0110	001101 0110
D13.6	CD	110 01101	101100 0110	101100 0110
D14.6	CE	110 01110	011100 0110	011100 0110
D15.6	CF	110 01111	010111 0110	101000 0110
D16.6	D0	110 10000	011011 0110	100100 0110
D17.6	D1	110 10001	100011 0110	100011 0110
D18.6	D2	110 10010	010011 0110	010011 0110
D19.6	D3	110 10011	110010 0110	110010 0110
D20.6	D4	110 10100	001011 0110	001011 0110
D21.6	D5	110 10101	101010 0110	101010 0110
D22.6	D6	110 10110	011010 0110	011010 0110
D23.6	D7	110 10111	111010 0110	000101 0110
D24.6	D8	110 11000	110011 0110	001100 0110
D25.6	D9	110 11001	100110 0110	100110 0110
D26.6	DA	110 11010	010110 0110	010110 0110
D27.6	DB	110 11011	110110 0110	001001 0110
D28.6	DC	110 11100	001110 0110	001110 0110
D29.6	DD	110 11101	101110 0110	010001 0110
D30.6	DE	110 11110	011110 0110	100001 0110
D31.6	DF	110 11111	101011 0110	010100 0110
D0.7	E0	111 00000	100111 0001	011000 1110
D1.7	E1	111 00001	011101 0001	100010 1110
D2.7	E2	111 00010	101101 0001	010010 1110
D3.7	E3	111 00011	110001 1110	110001 0001
D4.7	E4	111 00100	110101 0001	001010 1110

IBTA

Table 8 Valid Data Code Groups

Data Byte Name	Data Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
D5.7	E5	111 00101	101001 1110	101001 0001
D6.7	E6	111 00110	011001 1110	011001 0001
D7.7	E7	111 00111	111000 1110	000111 0001
D8.7	E8	111 01000	111001 0001	000110 1110
D9.7	E9	111 01001	100101 1110	100101 0001
D10.7	EA	111 01010	010101 1110	010101 0001
D11.7	EB	111 01011	110100 1110	110100 1000
D12.7	EC	111 01100	001101 1110	001101 0001
D13.7	ED	111 01101	101100 1110	101100 1000
D14.7	EE	111 01110	011100 1110	011100 1000
D15.7	EF	111 01111	010111 0001	101000 1110
D16.7	F0	111 10000	011011 0001	100100 1110
D17.7	F1	111 10001	100011 0111	100011 0001
D18.7	F2	111 10010	010011 0111	010011 0001
D19.7	F3	111 10011	110010 1110	110010 0001
D20.7	F4	111 10100	001011 0111	001011 0001
D21.7	F5	111 10101	101010 1110	101010 0001
D22.7	F6	111 10110	011010 1110	011010 0001
D23.7	F7	111 10111	111010 0001	000101 1110
D24.7	F8	111 11000	110011 0001	001100 1110
D25.7	F9	111 11001	100110 1110	100110 0001
D26.7	FA	111 11010	010110 1110	010110 0001
D27.7	FB	111 11011	110110 0001	001001 1110
D28.7	FC	111 11100	001110 1110	001110 0001
D29.7	FD	111 11101	101110 0001	010001 1110
D30.7	FE	111 11110	011110 0001	100001 1110
D31.7	FF	111 11111	101011 0001	010100 1110

IBTA

Table 9 Valid Special Code Groups

Data Byte Name	Special Byte Value	Bits HGF EDCBA	Current RD - abcdei fghj	Current RD + abcdei fghj
K28.0	1C	000 11100	001111 0100	110000 1011
K28.1	3C	001 11100	001111 1001	110000 0110
K28.2	5C	010 11100	001111 0101	110000 1010
K28.3	7C	011 11100	001111 0011	110000 1100
K28.4	9C	100 11100	001111 0010	110000 1101
K28.5	BC	101 11100	001111 1010	110000 0101
K28.6	DC	110 11100	001111 0110	110000 1001
K28.7	FC	111 11100	001111 1000	110000 0111
K23.7	F7	111 10111	111010 1000	000101 0111
K27.7	FB	111 11011	110110 1000	001001 0111
K29.7	FD	111 11101	101110 1000	010001 0111
K30.7	FE	111 11110	011110 1000	100001 0111

Note: Underlined special code-groups contain the comma bit patterns 0011111 or 1100000. This pattern is not found in any valid Data code group or combination of Data code-groups, and is used for symbol synchronization.

Note: Codes that are not contained in the above two tables are considered invalid and **shall** generate a code violation error.

5.2.3 RUNNING DISPARITY RULES

Running disparity **shall** be calculated using the following rules. Calculations on code groups that have been transmitted will yield a transmitter's running disparity. Similarly, calculations on code groups that have been received will yield a receiver's running disparity.

Running disparity for a code-group is calculated on the basis of sub-blocks, where the first six bits (abcdei) (Table 10 and Table 12) form one six-bit sub-block, and the second four bits (fghj) (Table 11 and Table 13) form the other four-bit sub-block. Running disparity at the beginning of the six-bit sub-block is the running disparity at the end of the previous code-group. Running disparity at the beginning of the four-bit sub-block is the running disparity at the end of the six-bit sub-block. Running disparity at the end of the code-group is the running disparity at the end of the four-bit sub-block.

Running disparity for the sub-blocks is calculated as follows:

1) Running disparity at the end of any sub-block is positive if the sub-block contains more ones than zeros. It is also positive at the end of the six-bit sub-block if the six-bit sub-block is 000111. It is likewise positive at the end of the four-bit sub-block if the four-bit sub-block is 0011;

2) Running disparity at the end of any sub-block is negative if the sub-block contains more zeros than ones. It is also negative at the end of the six-bit sub-block if the six-bit sub-block is 111000. It is also negative at the end of the four-bit sub-block if the four-bit sub-block is 1100;

3) Otherwise, running disparity at the end of the sub-block is the same as at the beginning of the sub-block.

NOTE-All sub-blocks with equal numbers of zeros and ones are disparity neutral. In order to limit the run length of 0's or 1's between sub-blocks, the 8B/10B transmission code rules specify that sub-blocks encoded as 000111 or 0011 are generated only when the running disparity at the beginning of the sub-block is positive; thus, running disparity at the end of these sub-blocks is also positive. Likewise, sub-blocks containing 111000 or 1100 are generated only when the running disparity at the beginning of the sub-block is negative; thus, running disparity at the end of these sub-blocks is also negative.

Table 10 5B/6B Coding for Data Characters

Data Byte Name	unencoded Bits EDCBA	Current RD - abcdei	Current RD + abcdei
D0	00000	100111	011000
D1	00001	011101	100010
D2	00010	101101	010010
D3	00011	110001	110001
D4	00100	110101	001010
D5	00101	101001	101001
D6	00110	011001	011001
D7	00111	111000	000111
D8	01000	111001	000110
D9	01001	100101	100101
D10	01010	010101	010101
D11	01011	110100	110100
D12	01100	001101	001101
D13	01101	101100	101100

Table 10 5B/6B Coding for Data Characters

Data Byte Name	unencoded Bits EDCBA	Current RD - abcdei	Current RD + abcdei
D14	01110	011100	011100
D15	01111	010111	101000
D16	10000	011011	100100
D17	10001	100011	100011
D18	10010	010011	010011
D19	10011	110010	110010
D20	10100	001011	001011
D21	10101	101010	101010
D22	10110	011010	011010
D23	10111	111010	000101
D24	11000	110011	001100
D25	11001	100110	100110
D26	11010	010110	010110
D27	11011	110110	001001
D28	11100	001110	001110
D29	11101	101110	010001
D30	11110	011110	100001
D31	11111	101011	010100

Table 11 3B/4B Coding for Data Characters

Data Byte Name	Unencoded Bits HGF	Current RD - fghj	Current RD + fghj
--.0	000	1011	0100
--.1	001	1001	1001
--.2	010	0101	0101
--.3	011	1100	0011
--.4	100	1101	0010
--.5	101	1010	1010
--.6	110	0110	0110
--.7	111	1110/0111	0001/1000

IBTA

Table 12 5B/6B Coding for Special Characters

Data Byte Name	Unencoded Bits EDCBA	Current RD - abcdei	Current RD + abcdei
K28	11100	001111	110000
K23	10111	111010	000101
K27	11011	110110	001001
K29	11101	101110	010001
K30	11110	011110	100001

Table 13 3B/4B Coding for Special Characters

Data Byte Name	Unencoded Bits HGF	Current RD - fghj	Current RD + fghj
--.0	000	1011	0100
--.1	001	0110	1001
--.2	010	1010	0101
--.3	011	1100	0011
--.4	100	1101	0010
--.5	101	0101	1010
--.6	110	1001	0110
--.7	111	0111	1000

5.2.4 GENERATING CODE-GROUPS

The byte to be encoded and current value of the transmitter's running disparity **shall** be used to select the code-group from its Table 8 or Table 9. For each code-group transmitted, a new value of the running disparity is calculated. This new value is used as the transmitter's current running disparity for the next byte to be encoded and transmitted.

5.2.5 CHECKING THE VALIDITY OF RECEIVED CODE-GROUPS

The following rules **shall** be used to determine the validity of received code groups:

1) The columns in Table 8 and Table 9 corresponding to the current value of the receiver's running disparity **shall** be searched for the received code-group.

2) If the received code-group is found in the proper column according to the current running disparity, then the code-group **shall** be con-

sidered valid and the associated data byte determined (decoded) for data code-groups.

3) If the received code-group is not found in that column, then the code-group **shall** be considered invalid.

4) Independent of the code-group's validity, the received code-group **shall** be used to calculate a new value of running disparity. The new value **shall** be used as the receiver's current running disparity for the next received code-group.

Detection of an invalid code-group does not necessarily indicate that the code-group in which the invalid code-group was detected is in error. Invalid code-groups may result from a prior error which altered the running disparity of the bit stream but which did not result in a detectable error at the code-group in which the error occurred.

5.3 CONTROL SYMBOLS AND ORDERED-SETS

The InfiniBand link uses the Control Symbols and Ordered-Sets of control and data symbols to implement:

1) packet delimiters,
2) ordered-set delimiters,
3) packet padding;
4) clock tolerance compensation.

C5-2: This compliance statement is obsolete and has been replaced by C5-2.2.1:.

C5-2.2.1: All ports **shall** use the control symbols and the SKIP, TS1, and TS2 ordered-sets specified in Section 5.3, "Control Symbols and Ordered-Sets," on page 95 for link/phy control and communication.

o5-2.2.1: All ports claiming compliance with InfiniBand Rel. 1.2 Enhanced Signaling **shall** also use the TS3, HRTBT, and TS-T ordered-sets specified in Section 5.3, "Control Symbols and Ordered-Sets," on page 95 for link/phy control and communication.

IBTA

5.3.1 CONTROL SYMBOLS

The IB control symbols have been chosen from the 8B/10B special code-groups and **shall** be used as defined in Table 14 below. The control code-groups are non-data symbols that are uniquely identifiable. Of the twelve available special code-groups, seven are used as control symbols, one is provided for vendor specific use, and four are reserved.

Table 14 Link Control Symbols

Symbol	Encoding	Description
COM	K28.5	Comma, character boundary alignment symbol.
SDP	K27.7	Start of Data Packet Delimiter
SLP	K28.2	Start of Link Packet Delimiter
EGP	K29.7	End of Good Packet Delimiter
EBP	K30.7	End of Bad Packet Delimiter
PAD	K23.7	Packet padding symbol
SKP	K28.0	Skip symbol
	K28.1 K28.7	Reserved control symbols. These symbols have “comma” characteristics.
	K28.3 K28.4	Reserved control symbols.
	K28.6	Vendor specific control symbol.

5.3.1.1 COMMA CONTROL SYMBOL (COM)

The comma control symbol (K28.5) is used by the physical lane receiver logic to identify symbol boundaries. Comma symbols are required to synchronize the receive logic when the links are being trained. The comma symbol is also used as the start of ordered-set delimiter.

5.3.1.2 START OF DATA PACKET DELIMITER (SDP)

The Start of Data Packet Delimiter symbol (K27.7) is transmitted to identify the start of a data packet. Packet formatting rules specify which physical lanes may be used by the “SDP” control symbol. (See Section 5.5 on page 111)

5.3.1.3 START OF LINK PACKET DELIMITER (SLP)

The Start of Link Packet Delimiter symbol (K28.2) is transmitted to identify the start of a link control packet. Packet formatting rules specify which physical lanes may be used by the “SLP” control symbol. (See Section 5.5 on page 111)

5.3.1.4 END OF GOOD PACKET DELIMITER (EGP)

The End of Good Packet Delimiter symbol (K29.7) is used to mark the end of each packet as it is transmitted by the originating port. Packet length

rules restrict which physical lanes may be used to transmit the “EGP” control symbol. (See Section 5.5 on page 111)

5.3.1.5 END OF BAD PACKET DELIMITER (EBP)

The End of Bad Packet Delimiter symbol (K30.7) is used to mark the end of a bad packet forwarded by a switch or router node. When an error (e.g.: decode error, CRC error, etc.) is detected in a data packet it is marked bad by replacing the original “EGP” symbol with a “EBP” packet symbol. Receiving end nodes are required to recognize either EGP or EBP as the end of packet delimiter. Any data packet terminated with “EBP” symbol must be treated as if it had a CRC error.

5.3.1.6 PADDING SYMBOL (PAD)

The Padding symbol (K23.7) is used on the 8x and 12x physical link to align the physical lanes. Alignment is required at the end of any packet that does not end (EGP or EBP) in physical lane eleven (11) on a 12x physical link, or physical lane seven (7) on an 8x link. (See Section 5.5.5 on page 115 and Section 5.5.6 on page 117) Pad symbols are also used by the retiming repeaters to forward error conditions (See Section 5.10 on page 161).

5.3.1.7 SKIP SYMBOL (SKP)

The “Skip” symbol (K28.0) is used as part of the SKIP ordered-set which is used for clock tolerance compensation. (See Section 5.9 on page 160)

5.3.1.8 RESERVED CONTROL SYMBOLS

Four of the special code-groups listed in Table 14 are reserved for future use by the IB standard. Special symbols K28.1, K28.3, K28.4, and K28.7 are reserved. For this version of the IB standard, the following rules apply to reserved control symbols:

1) The use of these control symbols **may** be defined in a future revision of this specification.
2) Devices based on this version of the specification **shall** not transmit these control symbols.
3) When a device based on this version of the specification receives a reserved control symbol a symbol coding error **shall** be reported. (See Section 5.4.2 on page 107)

5.3.1.9 VENDOR SPECIFIC CONTROL SYMBOL

The IB standard reserves a special control symbol (K28.6) for vendor specific use. The function and use of this symbol is vendor-defined and interoperability between vendors is not guaranteed. The vendor negotiation process for the use of this symbol is not defined by this specification.

The following rules apply to the use of the vendor-specific control symbols:

1) The use of the vendor-specific control symbols **may** be defined for vendor unique function.
2) Devices supporting the vendor-specific control symbol **shall not** transmit the symbol until both ends have negotiated its use.
3) When an unsupported or un-negotiated vendor-specific control symbol is detected, a coding error **shall** be reported. (see Section 5.4.2 on page 107)

5.3.2 CONTROL ORDERED-SETS

In addition to the individual control symbols described above, IB defines control ordered-sets. The ordered-sets are used for link training and clock tolerance compensation. The first symbol of all ordered-sets **shall** be the “COM" symbol, with additional symbols unique to the set type. When ordered-sets are transmitted, the ordered-set **shall** be transmitted on all physical lanes.

The defined ordered-sets are illustrated in Figure 11 below.

Ordered Set Formats - per Lane

Byte	SKIP	TS1	TS2	TS3	HRTBT	TS-T
0	COM	COM	COM	COM	COM	COM
1	SKP	LANE ID	LANE ID	LANE ID	LANE ID	Reserved
2	SKP	D10.2 (4Ah)	D5.2 (45h)	D13.2 (4Dh)	D1.2 (41h)	D17.2 (51h)
3	SKP	D10.2 (4Ah)	D5.2 (45h)	D13.2 (4Dh)	D1.2 (41h)	D17.2 (51h)
4		D10.2 (4Ah)	D5.2 (45h)	D13.2 (4Dh)	D1.2 (41h)	D17.2 (51h)
5		D10.2 (4Ah)	D5.2 (45h)	D13.2 (4Dh)	OpCode	D17.2 (51h)
6		D10.2 (4Ah)	D5.2 (45h)	D13.2 (4Dh)	Reserved	D17.2 (51h)
7		D10.2 (4Ah)	D5.2 (45h)	D13.2 (4Dh)	PortNum	D17.2 (51h)
8		D10.2 (4Ah)	D5.2 (45h)	SpeedActive	GUID[63-56]	Speeds
9		D10.2 (4Ah)	D5.2 (45h)	HBR/ADD	GUID[55-48]	Test Opcode
10		D10.2 (4Ah)	D5.2 (45h)	DDSV/DDS	GUID[47-40]	Reserved
11		D10.2 (4Ah)	D5.2 (45h)	Reserved	GUID[39-32]	Reserved
12		D10.2 (4Ah)	D5.2 (45h)	Reserved	GUID[31-24]	TxCfg[15-8]
13		D10.2 (4Ah)	D5.2 (45h)	Reserved	GUID[23-16]	TxCfg[76-0]
14		D10.2 (4Ah)	D5.2 (45h)	Reserved	GUID[15- 8]	RxCfg[15-8]
15		D10.2 (4Ah)	D5.2 (45h)	Reserved	GUID[7- 0]	RxCfg[76-0]

Figure 11 Ordered-Sets

5.3.2.1 SKIP ORDERED-SET (SKIP)

When transmitted the Skip sequence (SKIP) is a four symbol ordered-set comprised of a comma (COM) and three consecutive “Skip” (SKP) sym-

bols. A SKIP ordered-set may be inserted or repeated by a retiming repeater in the link. (See Section 5.10 on page 161).

5.3.2.2 TRAINING SEQUENCE ONE ORDERED-SET (TS1)

Link Training Sequence One (TS1) is a sixteen symbol ordered-set composed of a comma (COM), a lane identifier data symbol, and fourteen data symbols unique to training sequence one. The lane identifiers used for TS1 and TS2 **shall** use the definitions found in Table 15 below.

Table 15 Lane Identifiers

Lane Identifier	Hex number	8B/10B Encoding	Description
0	00	D00.0	Physical lane 0 used by 1x, 4x 8x, and 12x links
1	01	D01.0	Physical lane 1 used by 4x 8x, and 12x links
2	02	D02.0	Physical lane 2 used by 4x 8x, and 12x links
3	04	D04.0	Physical lane 3 used by 4x 8x, and 12x links
4	08	D08.0	Physical lane 4 used by 8x and 12x links
5	0F	D15.0	Physical lane 5 used by 8x and 12x links
6	10	D16.0	Physical lane 6 used by 8x and 12x links
7	17	D23.0	Physical lane 7 used by 8x and 12x links
8	18	D24.0	Physical lane 8 used by 12x links
9	1B	D27.0	Physical lane 9 used by 12x links
10	1D	D29.0	Physical lane 10 used by 12x links
11	1E	D30.0	Physical lane 11 used by 12x links

The TS1 unique data symbol is D10.2 (or 4Ah), and the 10-bit encoded value is a toggling pattern (010101 0101) for both the positive and negative running disparity.

5.3.2.3 TRAINING SEQUENCE TWO ORDERED-SET (TS2)

Link Training Sequence Two (TS2) is a sixteen symbol ordered-set composed of a comma (COM), a lane identifier data symbol, and fourteen data symbols unique to training sequence two. The lane Identifiers used by TS2 are the same as for TS1 and are defined in Table 15 above.

The TS2 unique data symbol is D5.2 (or 45h), and the 10-bit encoded value is the same pattern (101001 0101) for both the positive and negative running disparity.

5.3.2.4 TRAINING SEQUENCE THREE ORDERED-SET (TS3)

Link Training Sequence Three (TS3) is a sixteen symbol ordered-set composed of a comma (COM), a lane identifier data symbol, six data symbols unique to training sequence three, a bit map of the active speeds, a bit

map requesting transmitter de-emphasis and/or link heartbeat enabling, and a byte describing the transmitter de-emphasis setting which should be used, followed by five reserved bytes. The lane Identifiers used by TS3 are defined in Table 15. The format of the TS3 ordered-set is shown in Figure 12 on page 100 below. The use of the TS3 ordered-set is described in Section 5.6.4.6, "Configuration States - Enhanced Signaling," on page 133.

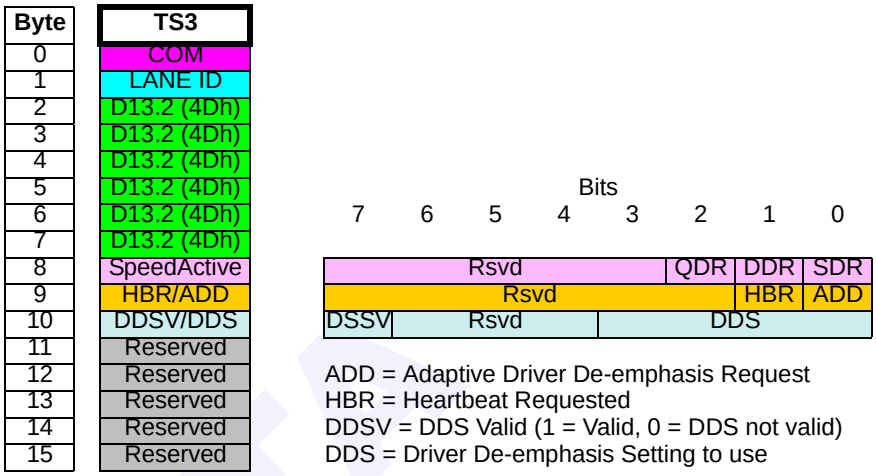


Figure 12 TS3 Ordered-Set: Detailed Format

Symbol 8, the ***SpeedActive*** link speed identifier, is used to identify and advertise the speeds at which the link/phy is enabled to operate, which the link medium supports. The enumerated values used for advertising enabled link speeds are the same as described for the ***SM.PortInfo(LinkSpeedEnabled)*** field.

Symbol 9 indicates whether the sender is requesting adaptive driver de-emphasis (ADD), and whether heartbeat is enabled and requested (HBR).

Symbol 10 indicates whether the transmitter of the TS3 is requesting a driver de-emphasis setting different from the default setting on its peer transmitter, and a 4-bit identifier for the driver de-emphasis setting requested. The transmitter driver may therefore have 17 de-emphasis settings - a default de-emphasis setting (DDSV = 0), and up to 16 adaptive driver de-emphasis settings (DDSV=1, DDS specifies a driver de-emphasis setting number). Driver de-emphasis is described in Section 6.3.2, "Equalization," on page 143.

Symbols 11 to 15 are Reserved. They are transmitted as D00.0, and valid Dxx.y symbols are ignored at the receiver.

The TS3 unique data symbol is D13.2 (or 4Dh), and the 10-bit encoded value is the same pattern (101001 0101) for both the positive and negative running disparity.

5.3.2.5 LINK HEARTBEAT ORDERED-SET (HRTBT)

Link Heartbeat ordered-set (HRTBT) is a sixteen symbol ordered-set composed of a common (COM), a lane identifier data symbol, three data symbols unique to the Link Heartbeat ordered-set, a 1-symbol OpCode, a 1-symbol Reserved field, a 1-symbol PortNum value (only used for switch ports), and an eight-symbol Globally Unique ID (GUID). The format of the HRTBT ordered-set is shown in Figure 13 on page 101 below. The use of the HRTBT ordered-set is described in Section 5.11.1, “Operation of Link Heartbeats,” on page 167

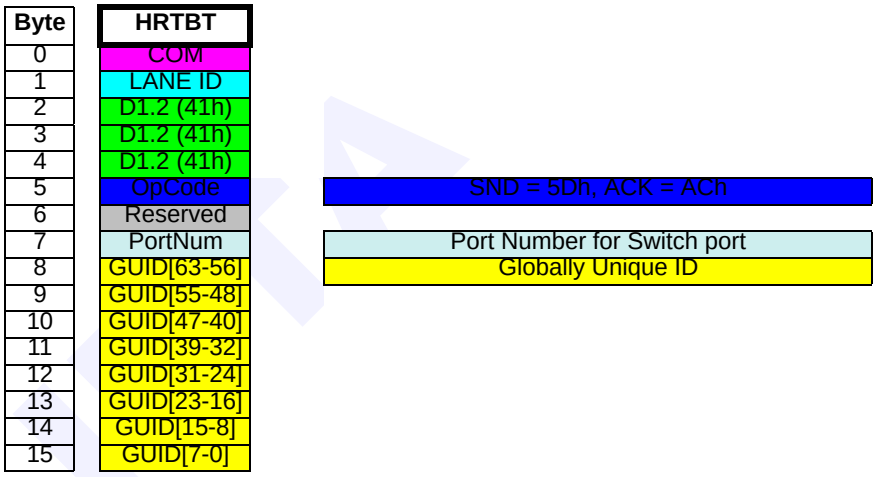


Figure 13 HRTBT Ordered-Set: Detailed Format

Symbol 5 is an OpCode, which signifies whether the Heartbeat is a Send Heartbeat (SND, OpCode = 5Dh), or an acknowledgement (ACK, OpCode = ACh).

Symbol 7 and Symbols 8-15 indicates the port number and the GUID or the port that the SND heartbeat came from, and which port the ACK is going back to, and Symbols

The Link Heartbeat unique data symbol is D1.2 (or 41h), and the 10-bit encoded value is either 011101 0101 for negative running disparity, or 100010 0101 for positive running disparity.

5.3.2.6 TRAINING SEQUENCE FOR TEST ORDERED-SET (TS-T)

The Training Sequence for Test is a sixteen symbol ordered-set composed of a comma (COM), a reserved symbol replacing the usual Lan ID

symbol, six data symbols unique to Training Sequence for Test, a bit map of the speed at which test phase is to occur (SDR, DDR, or QDR), a symbol indicating which of several defined test modes to use, and 2 16-bit transmitter configuration and receiver configuration fields, which define, in a vendor-dependent way, how the transmitter and receiver are configured during the test. The lane Identifiers used by TS-T are defined in Table 15. The format of the TS-T ordered-set is shown in Figure 14 below. The use of the TS-T ordered-set is described in Section 5.12, "Physical Layer Compliance Testing," on page 169.

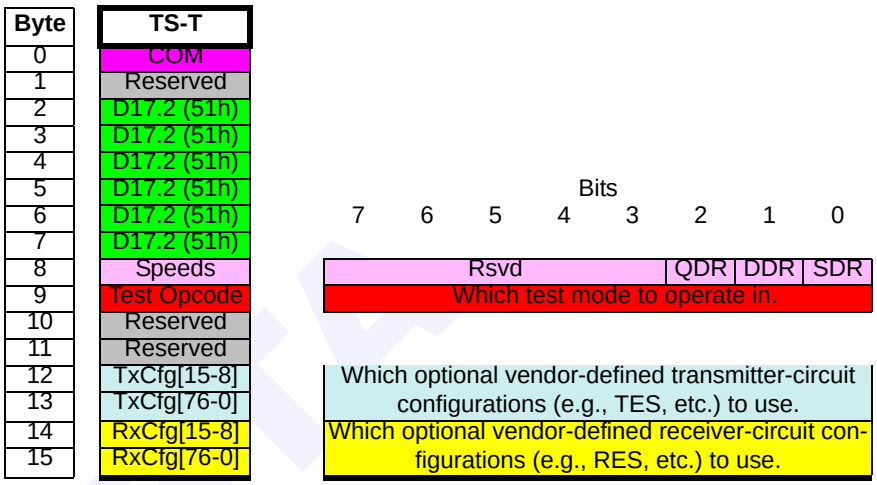


Figure 14 TS-T Ordered-Set: Detailed Format

The Training Sequence for Test (TS-T) is only generated by test equipment. There is no need for an InfiniBand device to be able to generate this ordered-set - only the need to recognize it and behave appropriately when 8 contiguous and valid copies of TS-T arrive at the receiver port on one or more lanes.

Symbol 8 of a TS-T, the link speed identifier, is used to identify the speed at which the test equipment requests the IB device to operate. At least one bit of this bit map must be asserted to 1. If multiple bits are asserted, the test will be conducted at the highest ***LinkSpeedEnabled*** speed. This allows testing at the highest enabled speed by simply asserting all bits in this symbol.

Symbol 9 of a TS-T allows the test equipment to determine which testing mode the port will be placed in. The following values are defined.

0: SKIP-less Idle Data

Each transmitter lane transmits a pseudo-random sequence of data symbols, generated by the 11th order LFSR = $X^{11} + X^9 + 1$ with no insertion of SKIP ordered-sets.

1: SKIP-less back-to-back TS1s

Each transmitter lane transmits an unbroken string of TS1 ordered-sets, with no insertion of SKIP ordered-sets.

2: Receiver test.

On each lane, the transmitter sends an indication of the validity of the data received on the corresponding receiver lane.

VALID DATA: For each received symbol on a lane which decodes to a valid 8b/10b code point with good running disparity, the corresponding transmitter lane transmits a D10.2 (010101 0101) character.

LOGICAL ERROR: For each received symbol on a lane which decodes with a logical error (e.g., bit error, or running disparity error), the corresponding transmitter lane transmits a K28.5 D00.0 pair of symbols.

LOSS OF SIGNAL: For each received symbol on a lane which indicates an inadequate signal (e.g., inadequate signal swing, all 0s, all 1s, or noise), the corresponding transmitter lane transmits a K28.5 D01.0 pair of symbols.

3-255: Optional Vendor-dependent opcodes to allow other testing modes.

Symbol 10 and Symbol 11 of a TS-T are reserved. They are transmitted as D00.0 and valid Dxx.y symbols are ignored at the receiver.

Symbols 12 and 13 allow the test equipment to optionally configure the transmitter in one of 65,536 vendor-dependent states, and symbols 14 and 15 allow the test equipment to configure the receiver in one of 65,536 vendor-dependent states. The values 0000h are identified for “normal” or “default” operation, so that test equipment can be expected to get good and correct operation when these fields are set to 0. Values other than 0 are optional and vendor-dependent. Typically only a very small subset of these states will be valid and will provide specific unique behaviors.

The TS-T unique data symbol is D17.2 (or 51h), and the 10-bit encoded value is the pattern (100011 0101) for both the positive and negative running disparity.

5.3.3 PACKET LOGICAL INTERFACES

This section describes the logical interfaces of the Link/Phy layer. This section is not intended to describe an actual implementation of a device but rather to explain the logical interface between protocol layers. This in-

terface may or may not be implemented as an internal or external device interface.

The logical interface section defines three logical interfaces:

1) Data/Status interface to the upper layer, Link to Link/Phy,
2) Interface to the lower layer, Link/Phy to Physical;
3) Control and Status interface to the Link/Phy layer.

5.3.3.1 LOGICAL LINK TO LINK/PHY DATA INTERFACE

The logical Link to Link/Phy data interface is defined in the *InfiniBand Architecture Specification, Volume 1*, Chapter 6: Physical Layer Interface.

5.3.3.2 LOGICAL LINK/PHY TO PHYSICAL INTERFACE

The Link/Phy to Physical interface consists of the following logical signals:

1) PhyTxData[11:0] **-** Transmit bit stream(s) 1x, 4x, 8x or 12x.
2) PhyRxData[11:0] **-** Receive bit stream(s) 1x, 4x, 8x or 12x.

5.3.3.3 LINK/PHY CONTROL AND STATUS INTERFACE

The Link/Phy control and status interface consists of the following logical signals:

1) PowerOnReset - Port Power on Reset input,
2) LinkPhyReset - Link Physical Reset control input,
3) LinkPhyRecover - Link Physical Recovery control input;
4) LinkPhyStat - Link Physical State (up or down) status output.

5.4 MANAGEMENT DATAGRAM CONTROL AND STATUS INTERFACE

The Management Datagram Control and Status interface is subdivided into two sub-sections: Control Inputs, and Status Outputs.

> **Implementation Note**
>
> Designers **should** not rely on the absence or characteristics of any features or commands marked “reserved” or “undefined”. The InfiniBand™ Trade Association reserves these for future definition.

C5-3: This compliance statement is obsolete and has been replaced by C5-3.2.1:.

C5-3.2.1: All ports **shall** implement the control and status management interface defined by Section 5.4, “Management Datagram Control and

Status Interface," on page 104, excluding DDR/QDR interoperability, 8x interoperability, PhyTest compliance testing, and LinkRoundTripLatency.

o5-3.2.1: All ports claiming compliance with InfiniBand Rel. 1.2 Enhanced Signaling **shall** implement the control and status management interface defined by Section 5.4, "Management Datagram Control and Status Interface," on page 104, including DDR/QDR interoperability, 8x interoperability, PhyTest compliance testing, and LinkRoundTripLatency.

5.4.1 CONTROL INPUTS (MAD SET)

Specific implementations **may** provide control information via proprietary mechanisms.

Multiple commands may be sent in the same ***SM.PortInfo(component)*** Management Datagram. Simultaneous commands, one of which changes the state of the Port Training State machine, **shall** set the other associated state variable(s) before the port training state change occurs.

The Control Input interface consists of the following Management Datagram to variable or logical signal mappings:

1) A Management Datagram ***SM.PortInfo(PortPhysicalState)*** set **shall** cause the Port Training State machine to transition state based on the following enumerated values:

   0: No State Change (NOP)

   1: Sleeping

   2: Polling

   3: Disabled

   4-15: Reserved (Ignored)

   Refer to Section 5.4.2 for fields returned by the MAD get operation. The default value following power on **shall** be set to ***Polling.***

2) A Management Datagram ***SM.PortInfo(LinkDownDefaultState)*** set **shall** set the logical signal LinkDownDefaultState based on the following enumerated values:

   0: No State Change (NOP)

   1: Sleeping

   2: Polling

   3-15: Reserved (Ignored)

   Refer to Section 5.4.2 for fields returned by the MAD get operation. The default value following power on **shall** be set to ***Polling.***

3) A Management Datagram **SM.PortInfo(LinkWidthEnabled)** set **shall** set the **LinkWidthEnabled** variable. No action **shall** be taken upon the variable until the Port Training State machine transitions to the **Configuration** state. The port **shall** only attempt to configure the link to width(s) based on the following enumerated values:

   0: No State Change (NOP)

   1: 1x

   2: 4x

   3: 1x or 4x

   4: 8x

   5: 1x or 8x

   6: 4x or 8x

   7: 1x or 4x or 8x

   8: 12x

   9: 1x or 12x

   10: 4x or 12x

   11: 1x, 4x or 12x

   12:8x or 12x

   13: 1x or 8x or 12x

   14: 4x or 8x or 12x

   15: 1x or 4x or 8x or 12x

   16-254: Reserved (Ignored)

   255: Set to ***LinkWidthSupported*** value

   Refer to Section 5.4.2 for fields returned by the MAD get operation. The default value following power on **shall** be set to ***LinkWidthSupported.***

4) A Management Datagram **SM.PortInfo(LinkSpeedEnabled)** set **shall** set the **LinkSpeedEnabled** variable. No action **shall** be taken upon the variable until the Port Training State machine transitions to the **Configuration** state. The port **shall** only attempt to configure the link to a speed based on the following enumerated values:

   0: No State Change (NOP)

   1: 2.5 Gb/s

   3: 2.5 or 5.0 Gb/s (SDR or DDR)

   5: 2.5 or 10.0 Gb/s (SDR or QDR)

   7: 2.5 or 5.0 or 10.0 Gb/s (SDR or DDR or QDR)

IBTA

2, 4, 6, 8-14: Reserved (ignored)

15: Set to ***LinkSpeedSupported*** value

Note that the SDR speed (2.5 Gb/s) must always be enabled, since link initialization occurs at this speed. Refer to Section 5.4.2 for fields returned by the MAD get operation. The default value following power on **shall** be set to ***LinkSpeedSupported.***

Implementation Note

MAD responders are required to respond to a **SM.PortInfo Set()** command with a **GetResp()** MAD. (*InfiniBand Architecture Specification, Volume 1*, Section 13.4.6.1) The **SM.PortInfo(PortPhysicalState)** set will cause the port training state machine to down the port when implementing the command. When this MAD is received on the port affected by the port state change command, the **GetResp()** MAD may not be transmitted before the port responds to the command. When a command is issued in this manner, the requestor should not expect a **GetResp()** MAD. Command execution can be verified by observing the state change at the other end of the link.

Implementation Note

Subnet Manager software designers should be aware that when a powered port is commanded to the Disabled State, the port will no longer respond to received packets or beacons. The port can only be re-enabled by using an alternate path, a second port on the same node, or some out-of-band connection. If there is no alternate path, the port must be reset to recover normal operation. For example, a single port IB device without an out-of-band management path would have to be physically reset to exit the Disable State.

5.4.2 STATUS OUTPUTS (MAD GET)

Specific implementations **may** provide status via proprietary mechanisms.

The Status Output interface consists of the following Management Datagram to variable or logical signal mappings:

1) A Management Datagram ***SM.PortInfo(PortPhysicalState)*** get **shall** return the Port Training State machines current state as one of the following enumerated values:

   1: Sleeping

2: Polling

3: Disabled

4: Configuration

5: LinkUp

6: Recovery

7: Phy Test

0, 8-15: Reserved

2) A Management Datagram ***SM.PortInfo(LinkDownDefaultState)*** get **shall** return the LinkDownDefaultState current value based on the following enumerated values:

1: Sleeping

2: Polling

0, 3-15: Reserved

3) A Management Datagram ***SM.PortInfo(LinkWidthEnabled)*** get **shall** return the previously written ***LinkWidthEnabled*** value based on the following enumerated values:

1: 1x

2: 4x

3: 1x or 4x

4: 8x

5: 1x or 8x

6: 4x or 8x

7: 1x or 4x or 8x

8: 12x

9: 1x or 12x

10: 4x or 12x

11: 1x, 4x or 12x

12: 8x or 12x

13: 1x or 8x or 12x

14: 4x or 8x or 12x

15: 1x or 4x or 8x or 12x

0, 16-255: Reserved

4) A Management Datagram ***SM.PortInfo(LinkSpeedEnabled)*** get **shall** return the previously written ***LinkSpeedEnabled*** value based on the following enumerated values:

IBTA

1: 2.5 Gb/s

3: 2.5 or 5.0 Gb/s (SDR or DDR)

5: 2.5 or 10.0 Gb/s (SDR or QDR)

7: 2.5 or 5.0 or 10.0 Gb/s (SDR or DDR or QDR)

0, 2, 4, 8-15: Reserved

5) A Management Datagram ***SM.PortInfo(LinkWidthSupported)*** get **shall** return the supported width(s) based on the following enumerated values:

1: 1x

3: 1x or 4x

7: 1x, 4x or 8x

11: 1x, 4x, or 12x (not valid for products supporting Rel. 1.2 Enhanced Signaling)

15: 1x, 4x, 8x or 12x

0, 2, 4-6, 8-10,12-14, 16-255: Reserved

6) A Management Datagram ***SM.PortInfo(LinkWidthActive)*** get **shall** return the Port Training State machines currently configured width based on the following enumerated values:

1: 1x

2: 4x

4: 8x

8: 12x

0, 3, 5-7, 9-255: Reserved

7) A Management Datagram ***SM.PortInfo(LinkSpeedSupported)*** get **shall** return the supported speeds based on the following enumerated values:

1: 2.5 Gb/s

3: 2.5 or 5.0 Gb/s (SDR or DDR)

5: 2.50 or 10.0 Gb/s (SDR or QDR)

7: 2.5 or 5.0 or 10.0 Gb/s (SDR or DDR or QDR)

0, 2, 4, 6, 8-15: Reserved

8) A Management Datagram ***SM.PortInfo(LinkSpeedActive)*** get **shall** return the Port Training State machines currently configured speed based on the following enumerated values:

1: 2.5 Gb/s

2: 5.0 Gb/s (DDR)

IBTA

4: 10 Gb/s (QDR)

0, 3, 5-15: Reserved

9) A Management Datagram ***SM.PortInfo(LinkRoundTripLatency)*** get **shall** return the minimum measured round trip link latency from the port to the connected port on the opposite side of the link. The LinkRoundTripLatency is a 32-bit value representing the minimum measured time for a bit in Link Heartbeat ordered-set to traverse the link in both directions, and can take on the following values:

   0xFFFF_FFFF: Link round-trip latency not yet measured, following reset on entry to LinkDownDefaultState.

   Others (0x0000_0000 to 0xFFFF_FFFE): Link round-trip latency between when a SND HEARTBEAT is transmitted and it's corresponding ACK HEARTBEAT is returned, reported in 4 nanosecond intervals.

   Measurement of LinkRoundTripLatency using HRTBT ordered-sets is described in Section 5.11.1, "Operation of Link Heartbeats," on page 167.

5.4.3 PORT PERFORMANCE COUNTERS

Each port implements the following performance counters. These counter are accessed using the Performance Management command defined in *InfiniBand Architecture Specification, Volume 1,* Section 15.2. The Link Physical Performance Counters **shall** implement both the get and set performance management methods. These counters do not rollover but **shall** stop at there maximum count. At set, operation is required to re-enable error counting. Specific implementations **may** provide performance information via proprietary mechanisms. The Performance Management interface **shall** consist of the following performance counters:

1) A Management Datagram ***Perf.PortCounters(SymbolErrorCounter)*** read **shall** return the current value of the 16-bit counter ***SymbolErrorCounter***. This counter is incremented each time an error is detected on one or more lanes. (See Section 5.7.2, "Minor Link Physical Errors Events," on page 157)

2) A Management Datagram ***Perf.PortCounters(LinkErrorRecoveryCounter)*** read **shall** return the current value of the 8-bit counter ***LinkErrorRecoveryCounter***. This counter is incremented each time the Port Training State machine successfully completes the link error recovery process. (See Section 5.6.4.8, "Link Error Recovery States," on page 144)

3) A Management Datagram ***Perf.PortCounters(LinkDownedCounter)*** read **shall** return the current value of the 8-bit counter ***LinkDownedCounter***. This counter is incremented each time the

Port Training State machine fails the link error recovery process and downs the link. (See Section 5.6.4.8, "Link Error Recovery States," on page 144). This counter is also incremented each time a Link Heartbeat error returns the Link Training State Machine to the Link-DownDefaultState. (See Section 5.11.2, "Heartbeat Error Handling," on page 168).

5.5 PACKET FORMATS FOR SINGLE AND MULTI LANE SUPPORT

This section describes the distribution and translation of the Link Layer packet byte stream to the physical lanes.

In addition to the control ordered-sets defined earlier, data symbols are used as the payload of link and data packets and as link idle data. Packet data symbols are framed by start of packet symbols SDP and SLP and end of packet symbols EGP and EBP. Idle data symbols are not part of a link or data packet and are not framed by packet delimiters symbols.

5.5.1 LINK PACKET ORDERING

The upper layers of the Protocol provide a stream of packets. The packet in the stream are composed of:

1) A Local Routing Header and other optional headers.

2) A packet payload, type dependent.

3) A Invariant CRC, type dependent.

4) And a Variant CRC.

The combined length of packet header(s), payload, and Invariant CRC are a multiple of four bytes. Two bytes of a Variant CRC plus packet start delimiter and packet end delimiter ensure that all packet are a multiple of four bytes in length. See *InfiniBand Architecture Specification, Volume 1*, Chapters 6 & 7 for complete details of internal packet formatting. The Link/Physical layer forwards these packets in the order received. As required, the Link/Physical layer will insert SKIP ordered-sets between packets. (See Section 5.9.2) An example of packet ordering on a 4x link is depicted below in Figure 15. When there is no packet or SKIP ordered-sets set to transmit the Link/Physical layer will fill the link Idle with a pseudo-random sequence of data symbols (idle data).

C5-4: All ports **shall** implement link packet ordering as defined by Section 5.5.1, "Link Packet Ordering," on page 111.

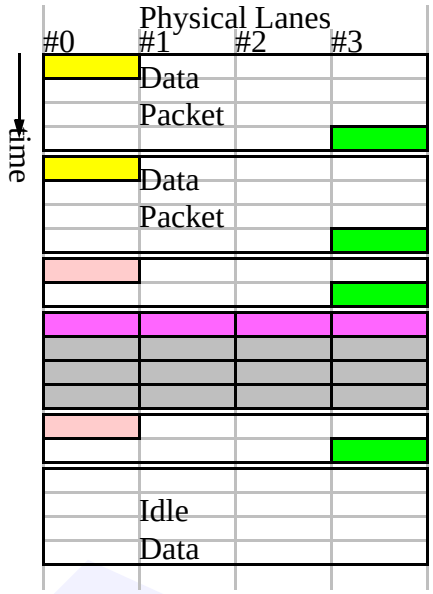


Figure 15 Packet Ordering Example

5.5.1.1 PACKET ORDERING RULES

1) Packets contain un-interruptible packet data content and delimiters. Ordered-sets and control symbols **shall not** be inserted into a packet's data content except as described in rule 4 below.

2) Packets received from the upper layers **shall** be transmitted in the order received.

3) Scheduled SKIP ordered-sets **shall** only be inserted between packets for clock tolerance compensation.

4) To initiate the error recovery process, TS1 ordered-sets **shall** be inserted at any symbol boundary, possibly interrupting the current packet.

5) When the link is idle (no packets or control ordered-sets to transmit), a pseudo-random sequence of data symbols (idle data) **shall** be transmitted on all lanes.

6) The link idle pseudo-random data sequence **shall** be generated by the 11th order LFSR = $X^{11} + X^{9} + 1$.

7) Each lane **may** start the link idle pseudo-random data pattern at an arbitrary valid value. No lane-to-lane dependence is specified for the link idle pseudo-random data pattern.

8) Idle data **shall** be terminated at any time there is a packet or control ordered-set to transmit.

5.5.2 PACKET FORMATS

Packets including start and end delimiters are formed by the upper layers of the protocol. This stream of packets is striped across the available physical lane(s) (1x, 4x, 8x, or 12x).

C5-5: All ports **shall** implement packet formatting as defined by Section 5.5.2, "Packet Formats," on page 113.

5.5.2.1 PACKET FORMATTING RULES

1) Total length of data packets including packet delimiter symbols **shall** be integer multiples of four symbols.
2) Data packets **shall** have SDP symbol as the first symbol of the packet.
3) The link packets **shall** be eight symbols long including the packet delimiter symbols.
4) Link packets **shall** have a SLP symbol as the first symbol of the packet.
5) When transmitted, all packets **shall** be terminated by an EGP or EBP symbol.
6) There are no per-lane even or odd alignment restrictions. Packets are not required to start with even or odd alignments. Comma symbols do not force even alignment.
7) The starting running disparity of packet delimiters and ordered-sets is not specified; a packet or ordered-set may start with positive or negative disparity. The disparity of all symbols **shall** comply with 8B/10B encoding rules. (See Section 5.2.3)

5.5.3 1X PACKET FORMAT

The 1x link is composed of a single physical lane. The combined packet symbol stream is serialized into a single stream of symbols. The SKIP ordered-set is inserted between packets as needed for clock tolerance compensation. The combined symbol stream (packets, SKIP ordered-sets, and idle data) is encoded using the 8B/10B code defined in Section 5.2 above. A 1x symbol stream containing data packets, link packet, idle data, and SKIP ordered-sets is illustrated in Figure 16 below.

C5-6: All 1x ports and 4x and 12x ports when configured as a 1x port **shall** implement packet formatting as defined by Section 5.5.3, "1x Packet Format," on page 113.

o5-6.2.1: All 8x ports when configured as a 1x port **shall** implement packet formatting as defined by Section 5.5.3, "1x Packet Format," on page 113.

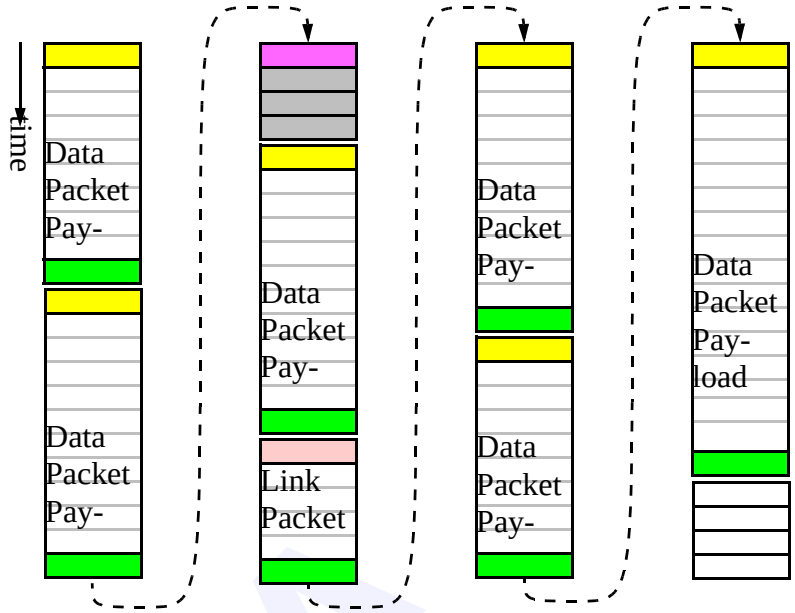


Figure 16 1x Packet Formats

5.5.3.1 1X LINK FORMATTING RULES

1) The start of packet delimiters (SDP & SLP) **shall** be transmitted in lane zero only.

2) The end of packet delimiters (EGP & EBP) **shall** be transmitted in lane zero only.

5.5.4 4X PACKET FORMAT

The 4x link is composed of four physical lanes (0 through 3). The combined data and link packet symbol stream (control and data) are byte striped across physical lanes 0 through 3. In the 4x configuration, the start packet delimiters will always be transmitted on physical lane 0 and the end packet delimiters will always be transmitted on physical lane 3. When SKIP ordered-sets are needed for clock tolerance compensation, they are inserted between packets simultaneously on all physical lanes. The four symbol streams composed of packets, SKIP ordered-sets, and idle data are individually encoded using the 8B/10B code defined in Section 5.2 above.The 4x symbol streams containing data packets, link packets, and SKIP ordered-sets is illustrated in Figure 17 below.

C5-7: All 4x ports and 12x ports when configured as a 4x port **shall** implement packet formatting as defined by Section 5.5.4, “4x Packet Format,” on page 114.

o5-7.2.1: All 8x ports when configured as a 4x port **shall** implement packet formatting as defined by Section 5.5.4, "4x Packet Format," on page 114.

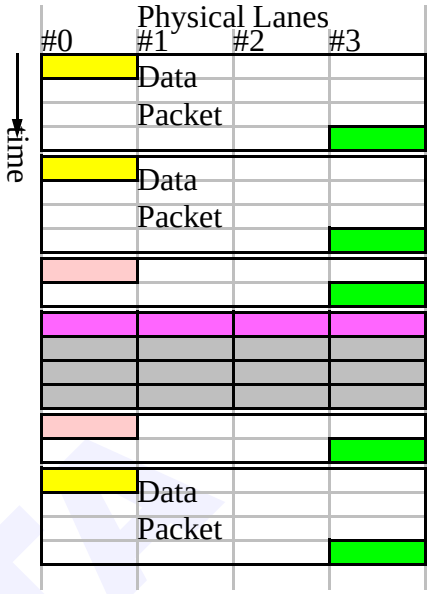


Figure 17 4x Packet Formats

5.5.4.1 4X PACKET FORMATTING RULES

1) The start of packet delimiters (SDP & SLP) **shall** be transmitted in lane zero (0) only.
2) The end of packet delimiters (EGP & EBP) **shall** be transmitted in lane three (3) only.
3) SKIP ordered-sets **shall** be transmitted on all (4) lanes simultaneously.
4) When the idle data is placed on the link, the idle data pattern **shall** be inserted on all four(4) lanes simultaneously.
5) The pseudo-random idle data for each lane **should** start a different point in the sequence.

5.5.5 8X PACKET FORMAT

The 8x link is composed of eight physical lanes (0 through 7). The combined data and link packet symbol stream (control and data) are byte striped across physical lanes 0 through 7. In the 8x configuration the start packet delimiters will always be transmitted on physical lane 0, and the end packet delimiters will always be transmitted on physical lane 3 or 7. The link will be padded (PAD control symbols) as necessary to maintain the start of packet alignment rule. When SKIP ordered-sets are needed

for clock tolerance compensation, they are inserted between packets simultaneously on all physical lanes. The eight symbol streams composed of packets, SKIP ordered-sets, and idle data are individually encoded using the 8B/10B code defined in Section 5.2 above.The 8x symbol streams containing data packets, link packets, and SKIP ordered-sets is illustrated in Figure 18 below.

o5-7.2.1: All 8x ports and 12x ports when configured as a 8x port **shall** implement packet formatting as defined by Section 5.5.5, "8x Packet Format," on page 115.

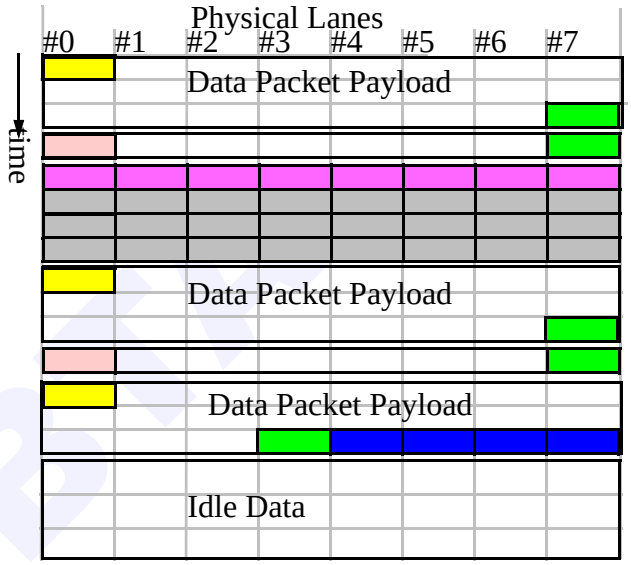


Figure 18 8x Packet Formats

5.5.5.1 8x PACKET FORMATTING RULES

1) The start of packet delimiters (SDP & SLP) **shall** be transmitted in lane zero (0) only.
2) The end of packet delimiters (EGP & EBP) **shall** be transmitted in lane three or seven (3 or 7) only.
3) When the currently transmitted packet does not end on lane 7, four PAD control symbols **shall** be inserted to align the link for the next transmit operation.
4) SKIP ordered-sets **shall** be transmitted on all eight (8) lanes simultaneously.
5) When the idle data is placed on the link, the idle data pattern **shall** be inserted on all eight (8) lanes simultaneously.

6) The pseudo-random idle data for each lane **should** start at a different point in the sequence.

5.5.6 12X PACKET FORMAT

The 12x link is composed of twelve physical lanes (0 through 11). The combined data and link packet symbol stream (control and data) are byte striped across physical lanes 0 through 11. In the 12x configuration the start packet delimiters will always be transmitted on physical lane 0, and the end packet delimiters will always be transmitted on physical lane 3, 7, or 11. The link will be padded (PAD control symbols) as necessary to maintain the start of packet alignment rule. When SKIP ordered-sets are needed for clock tolerance compensation, the link is padded as necessary to allow all twelve SKIP ordered-sets to start on the same clock. SKIPS are inserted between packets simultaneously on all physical lanes. The twelve symbol streams, composed of packets, SKIP ordered-sets, and idle data are individually encoded using the 8B/10B code defined in Section 5.2. The 12x symbol streams containing data packets, link packet, idle data and SKIP ordered-sets is illustrated in Figure 19 below.

C5-8: All 12x ports **shall** implement packet formatting as defined by Section 5.5.6, "12x Packet Format," on page 117.

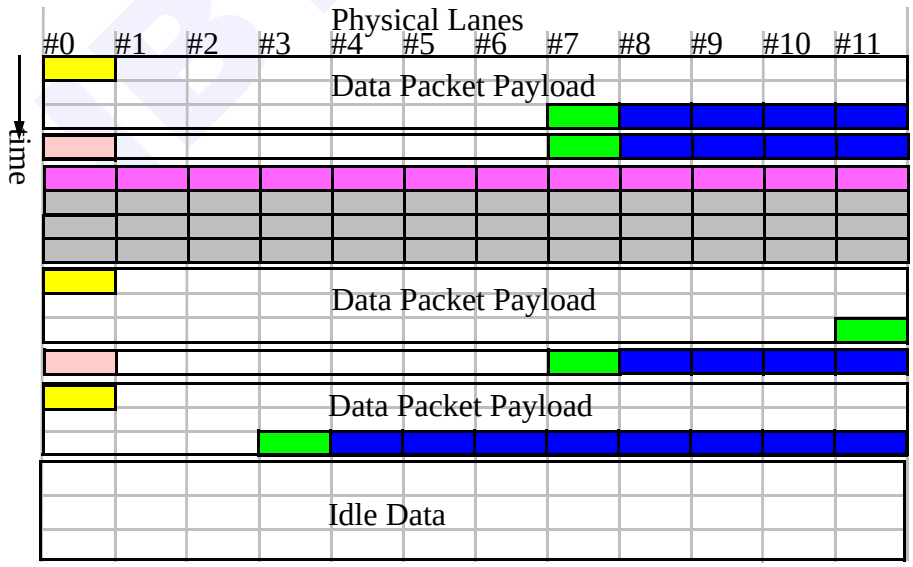


Figure 19 12x Packet Formats

5.5.6.1 12X PACKET FORMATTING RULES

1) The start of packet delimiters (SDP & SLP) **shall** be transmitted in lane zero (0) only.

2) The end of packet delimiters (EGP & EBP) **shall** be transmitted in lane three, seven, or eleven (3, 7, or 11) only.
3) When the currently transmitted packet does not end on lane 11, four or eight PAD control symbols **shall** be inserted to align the link for the next transmit operation.
4) SKIP ordered-sets **shall** be transmitted on all twelve(12) lanes simultaneously.
5) When the idle data is placed on the link, the idle data pattern **shall** be inserted on all twelve(12) lanes simultaneously.
6) The pseudo-random idle data for each lane **should** start at a different point in the sequence.

5.6 LINK INITIALIZATION AND TRAINING

This section defines the link/physical control process that configures and initializes a link for normal operation. A fully-powered port implements this process by accomplishing the following:

1) Waking a remote port on auxiliary power.
2) Configuring and initializing the link.
3) Supporting normal packet transfers
4) Recovering from transient link errors.

Waking a remote port is a special case operation with the local port operating fully-powered and the remote port operating on auxiliary power. The local port polls (See Section 5.6.4.2, “Polling States,” on page 127) the remote port. The remote port detects the presence of this signal and initiates fully-powered operation. In this special case the signal detected by the end operating on auxiliary power is referred to as a beacon. Its use is described further in *InfiniBand Architecture Specification, Volume 2*, Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s and in *InfiniBand Architecture Specification, Volume 2*, Chapter 14: OS Power Management.

During the training process, the ports at each end of the link learn each other’s capabilities and configure the following parameters:

1) Link width (1x, 4x, 8x or 12x).
2) Link speed (SDR-2.5 Gb/s, DDR-5.0 Gb/s, or QDR-10.0 Gb/s).
3) Optional correction of lane reversal.
4) Optional correction of inverted received serial data (crossed differential signals).

C5-9: This compliance statement is obsolete and has been replaced by C5-9.2.1:.

C5-9.2.1: All ports **shall** implement link initialization and training as defined by Section 5.6, "Link Initialization and Training," on page 118. Ports are not required to implement the lane reversal and serial data inversion options. Implementations that do not support Enhanced Signaling must use the Link Training State Machine described in Figure 21: Link Training State Machine - Legacy, and **shall not** implementation any Enhanced Signaling functions related to TS3 ordered-sets, Heartbeat ordered-sets, LinkRoundTripLatency, TS-T ordered-sets, or PhyTest compliance testing.

o5-9.2.1: All ports that claim compliance with the Rel. 1.2 Enhanced Signaling **shall** implement link initialization and training as defined by Section 5.6, "Link Initialization and Training," on page 118 and **shall** implement all functions related to TS3 ordered-sets, heartbeat ordered-sets, LinkRoundTripLatency measurement, TS-T ordered-sets, or PhyTest compliance testing,

o5-1: This compliance statement is obsolete and has been replaced by o5-9.2.1:.

o5-9.2.1: Ports that implement the serial data inversion option **shall** implement the following:
Section 5.6.7.3, "RxCMD = EnConfig," on page 152 rules #2.

o5-2: This compliance statement is obsolete and has been replaced by o5-9.2.1:.

o5-9.2.1: All 4x, 8x and 12x ports that implement the lane reversal option **shall** implement the following:
Section 5.6.4.6.3, "Config.WaitRmt State," on page 139 rule #7 and
Section 5.6.4.6.4, "Config.TxRevLanes State," on page 140 all rules and
Section 5.6.7.3, "RxCMD = EnConfig," on page 152 rule #4.

5.6.1 LINK DE-SKEW AND TRAINING SEQUENCE

The terms link de-skew and training sequence are used throughout this section. Both are briefly explained below. Additionally, the characteristics of the training sequence ordered-sets (TS1, TS2, and TS3) are summarized herein.

Link de-skew: A multi-lane link (4x, 8x and 12x) may have many sources of lane to lane skew. These sources include but are not limited to chip I/O drivers and receivers, printed wiring boards, electrical and optical cables, serialization and de-serialization logic, and retiming repeaters. Although symbols are transmitted simultaneously on all lanes, they cannot be expected to arrive at the receiver without lane-to-lane skew. The lane-to-lane skew may include components which are less than a bit time, bit time units (400ps at SDR, 200ps and 100ps at DDR and QDR, respectively),

or full symbol time units (4ns at SDR, 2ns and 1ns at DDR and QDR, respectively) of skew caused by the retiming repeaters' insert/delete operations. A link may have up to two retiming repeaters. Each repeater operates independently on multiple lanes. Because of this independent operation, they may insert or delete a skip symbol independently on each lane. Refer to Section 5.10 on page 161 for a complete description of retiming repeaters. The receiving node is required to remove this lane-to-lane skew in order to receive and process data on all lanes simultaneously. This process is called "link de-skew". Receivers use TS1, TS2, and TS3 ordered-sets to perform link de-skew functions.

TS1 and TS2: The training sequences TS1 and TS2 provide three fundamental types of information:

1) **Lane-to-lane skew:** Information is provided by the unique structure and length of the of the two training sequences and the fact that they are transmitted without lane-to-lane skew. Both training sequences' (TS1 and TS2) sixteen symbol length (comma and 15 data symbols) allows unambiguous de-skew of up to seven symbol times of skew.
2) **Lane Identification:** The second symbol of the training sequence contains the lane number information. The lane number uniquely identifies each of the twelve possible lanes.
3) **Lane serial data polarity:** The TS1 and TS2 data symbols contain serial data polarity information. When the bit stream on a lane is inverted, the comma and lane number symbols swap running disparity but still decode to the same value. The TS1 data symbol changes from a D10.2 (4Ah) to a D21.5 (B5h), and the TS2 data symbol changes from a D5.2 (45h) to a D26.5 (BAh) providing a clear indication of lane serial data polarity inversion.

TS3: The training sequence TS3 provides the same three fundamental types of information provided by TS1 and TS2. In addition, it provides the following types of information:

1) **Enabled and Supported Link Speeds:** Information is provided by a Link/Physical layer about the speed or speeds that the layer supports and is enabled to operate at. This information is provided in the Symbol 8, the SpeedActive symbol, of the sixteen symbols of the TS3 training sequences. During link training, each Link/Physical layer transmits a series of TS3 training sequences to the Link/Physical layer on the other side of its associated link, as part of the link speed auto-negotiation procedure. A link always operates at the highest speed that is supported and enabled on both ends and supported by the physical channel.
2) **Link Heartbeat Enabled:** The Link Heartbeat function is required at DDR and QDR speeds, and expected at SDR speed. A port may disable use of the Link Heartbeat function, e.g., for interoperation

with legacy devices at SDR speed. Link Heartbeat will only be put into effect when both ports on a negotiation exchange support and enable it.

3) **Adaptive Driver De-emphasis Request:** Allows a port to request handshake for a simple adaptive driver de-emphasis setting for equalization, with either a default driver de-emphasis setting or one of 16 adaptive driver de-emphasis settings negotiated between transmitter and receiver.

4) **Driver De-emphasis Setting request**: Allows a receiver to transmit to its peer transmitter the requested driver de-emphasis setting out of the 17 possible settings.

5.6.2 LINK INITIALIZATION AND TRAINING OPTIONS

Two independent optional features are defined as part of link initialization and training. Both are intended to allow on chip logic to correct non-optimum pin assignments which may occur when connecting an IBA link chip-to-chip or chip-to-connector. These options can eliminate the two common printed wiring board (PWB) layout issues of crossed differential pair and bus bow ties. Implementing these optional features allows the PWB layout to focus on signal integrity with simplified connections.

5.6.2.1 INVERTED SERIAL DATA CORRECTION (OPTIONAL)

The inverted serial data correction feature allows the receiver logic to correct a receive differential pair that is crossed in the PWB layout. Such a receiver will need to test the polarity of the received training sequences and correct inverted data as part of link configuration. This option does not provide the capability to correct the polarity of transmit data.

To be inter-operable with ports that do not implement inverted serial data correction, the polarity at the transmitting connector must be as specified in the InfiniBand *Architecture Specification,* Chapter 7: Copper Cable and Chapter 10: Backplane Connector Specification.

5.6.2.2 LANE REVERSAL CORRECTION (OPTIONAL)

The lane reversal feature allows a PWB layout to connect 4x, 8x or 12x ports in reversed lane number order. The receiver logic uses the lane number symbol in the training sequences (TS1 or TS2) to detect and correct the reversed lane connections in the PWB layout. When the remote port is incapable of correcting receive lane reversal (link width is less than the local port or the port does not implement this option) the local port will reverse its transmit lane as part of its configuration process. (see Section 5.6.4.6, “Configuration States - Enhanced Signaling,” on page 133)

5.6.3 INTERACTIONS WITH OTHER ENTITIES

Link training is an involved process that requires interactions with other entities. This section briefly describes these interfaces and refers the reader to other sections for more detailed information. Figure 20 illustrates all entities that interact with the link training state machine in the link initialization and configuration process.

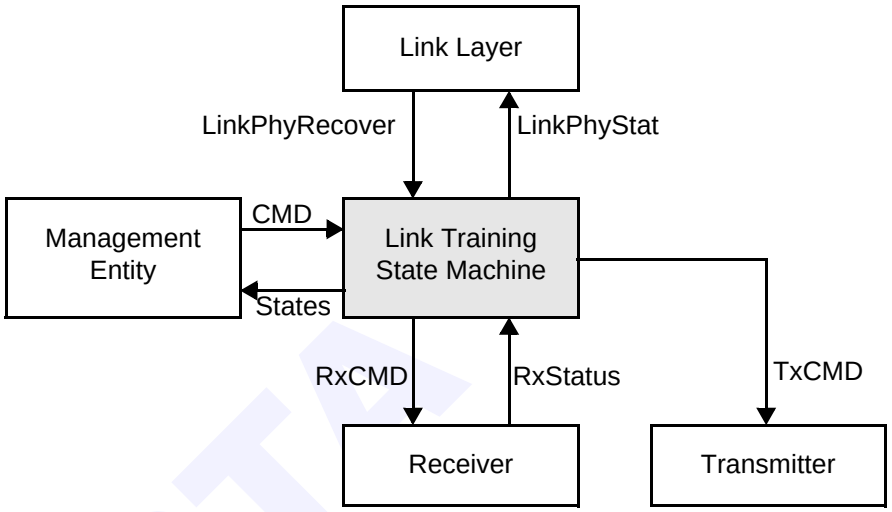


Figure 20 Link Training State Machine Interactions

The interface between the link layer and the link/phy layer is comprised of the signals LinkPhyRecover and LinkPhyStat. These logical signals are described in Section 5.3.3, "Packet Logical Interfaces," on page 103.

The link training state machine and the unit's management entity communicate over a set of logical signals. Although this logical interface is not defined completely in this specification, functions that need to be carried out over this interface are defined. The management messages related to the link training state machine operations are described in Section 5.4, "Management Datagram Control and Status Interface," on page 104.

The link training state machine also interfaces to the transmitter and receiver logic to coordinate the link initialization and configuration operations. The interface to the transmitter is in the form of commands (TxCMD) for the transmitter to perform. Similarly, the receiver interface has a set of commands (RxCMD) or configurations and a set of receiver status (RxStatus) values.

These commands and status conditions are listed in Table 16 below and are described in detail in Section 5.6.6 and Section 5.6.7. The receiver status conditions (RxStatus) are not mutually exclusive conditions.

Table 16 Transmitter and Receiver Interface

Transmitter		Receiver	
TxCMD	**Speed**	**RxCMD**	**Status (RxStatus)**
Disable		Disable	
SendTS1, SendTS2, SendTS3	SPEED = MinEnabledSpeed, SPEED = MaxBothActive	WaitTS1,WaitTS2, WaitTS3	RcvdTS1, RcvdTS2, RcvdTS3
RevLanes		EnConfig, EnDeSkew	RxTrained
SendIdle		WaitIdle	RcvdIdle
Enable		Enable	RxMajorError, RxHeartbeatError
		WaitTS-T	RcvdTS-T

5.6.4 LINK TRAINING STATE MACHINE

Link initialization, configuration, and link error recovery operations are performed by the link training state machine. This state machine is described in this section in the form of hierarchical states. As depicted in Figure 22, the link training state machine has seven primary states. Some of these states are super states composed of two or more states.

The link training state machine has the following primary states:

1) **Disabled**: Port drives its outputs to quiescent levels and does not respond to received data.
2) **Sleeping**: In this super state, the port drives its outputs to quiescent levels and responds to received training sequences.
3) **Polling**: In this super state, the port transmits training sequences and responds to received training sequences. This is the default state following power on.
4) **Configuration**: A transient super state with both the transmitter and receiver active. The port is attempting to configure and transition to the LinkUp state.
5) **LinkUp**: This is the normal link operation state. The port is available to transfer packets.
6) **Recovery**: The recovery super state attempts to re-synchronize the link and return it to normal operation. This state is entered when a port experiences loss of link synchronization, a major error, or when a link layer error triggers error recovery.
7) **Phy Test**: The Physical Layer test state allows simplified testing to determine compliance of physical layer transmitter driver and receiver circuitry with specified requirements. This state is defined for

Enhanced Signaling**,** and is not implemented in products not supporting Enhanced Signaling.

These primary states directly map to the enumerated port states defined in Section 5.4, "Management Datagram Control and Status Interface," on page 104.

The status of the link training state machine is also reported to the link layer in a more abstract fashion. The "LinkPhyStat" logical signal over this interface provides the values "Up" and "Down". This logical signal is given the "Up" value when this state machine is in the LinkUp state. It is given the "Down" value in all other states.

The initial or post-reset state of the link training state machine is **Polling**. This insures that a newly connected and powered on port will be recognized by any remote port. Management commands can force the link training state machine to the following states: **Disabled**, **Polling**, or **Sleeping**. Management commands also set a default state for link down. All other state transitions are under the control of the link training state machine. Test equipment can control the transition in and out of the **Phy Test** state by injecting patterns in the port's receiver.

Figure 21 shows the Link Training State Machine for products not supporting Enhanced Signaling. Figure 22 shows the Link Training State Machine for products that support Enhanced Signaling, which adds the **Phy Test** state, and additional substates within the **Configuration** super state.

In Figure 22, all state transition events are based on management commands, delay events, or receiver status. These receiver status conditions are described later in Section 5.6.7 on page 151.

Operations in certain states involve delays or time-out periods. These events are denoted the state diagrams as state transitions labeled with DelayTimeOut (time). These time-out periods start as the state is entered and cleared when a transition to a different state is taken. Time-out periods cannot be re-started within a state. (See Section 5.6.5 on page 147)

Note: This is the legacy state diagram, for products that do not support Enhanced Signaling. It has been modified for Enhanced Signaling as shown in the diagram in Figure 22.

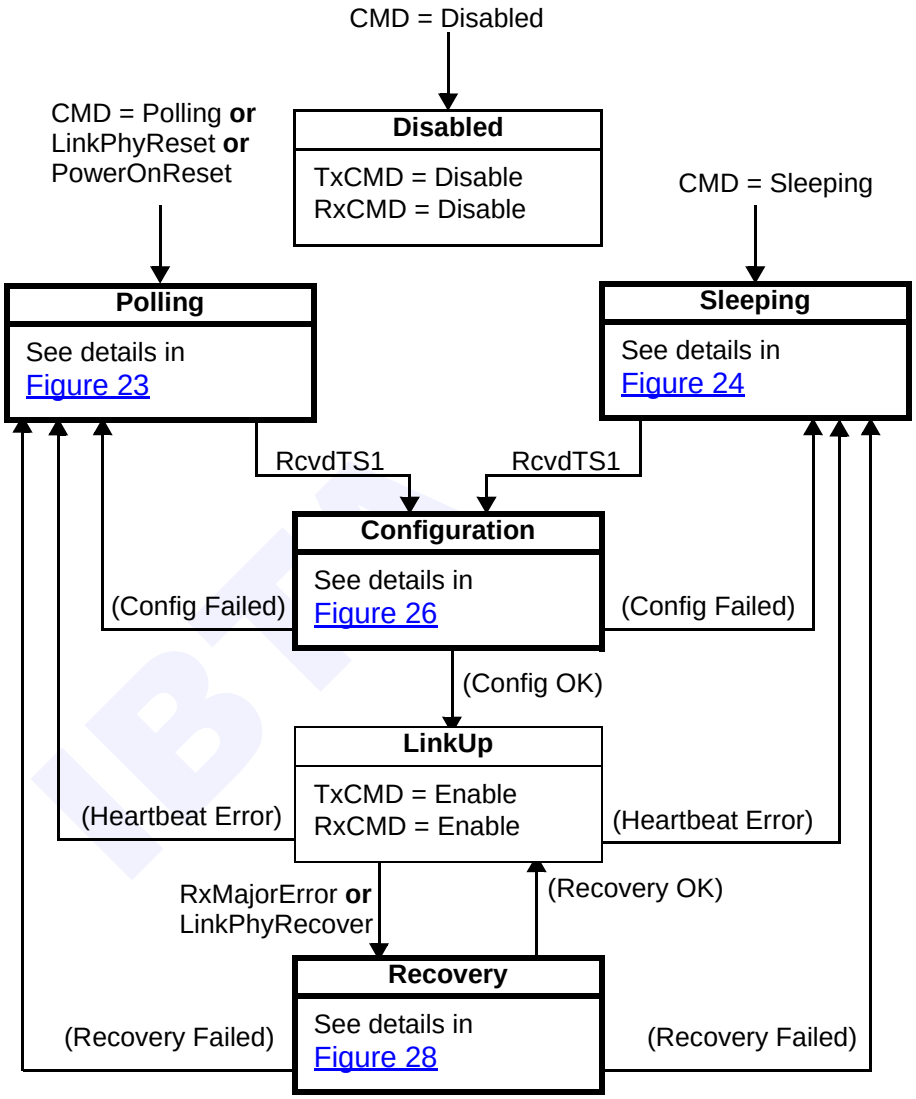


Figure 21 Link Training State Machine - Legacy

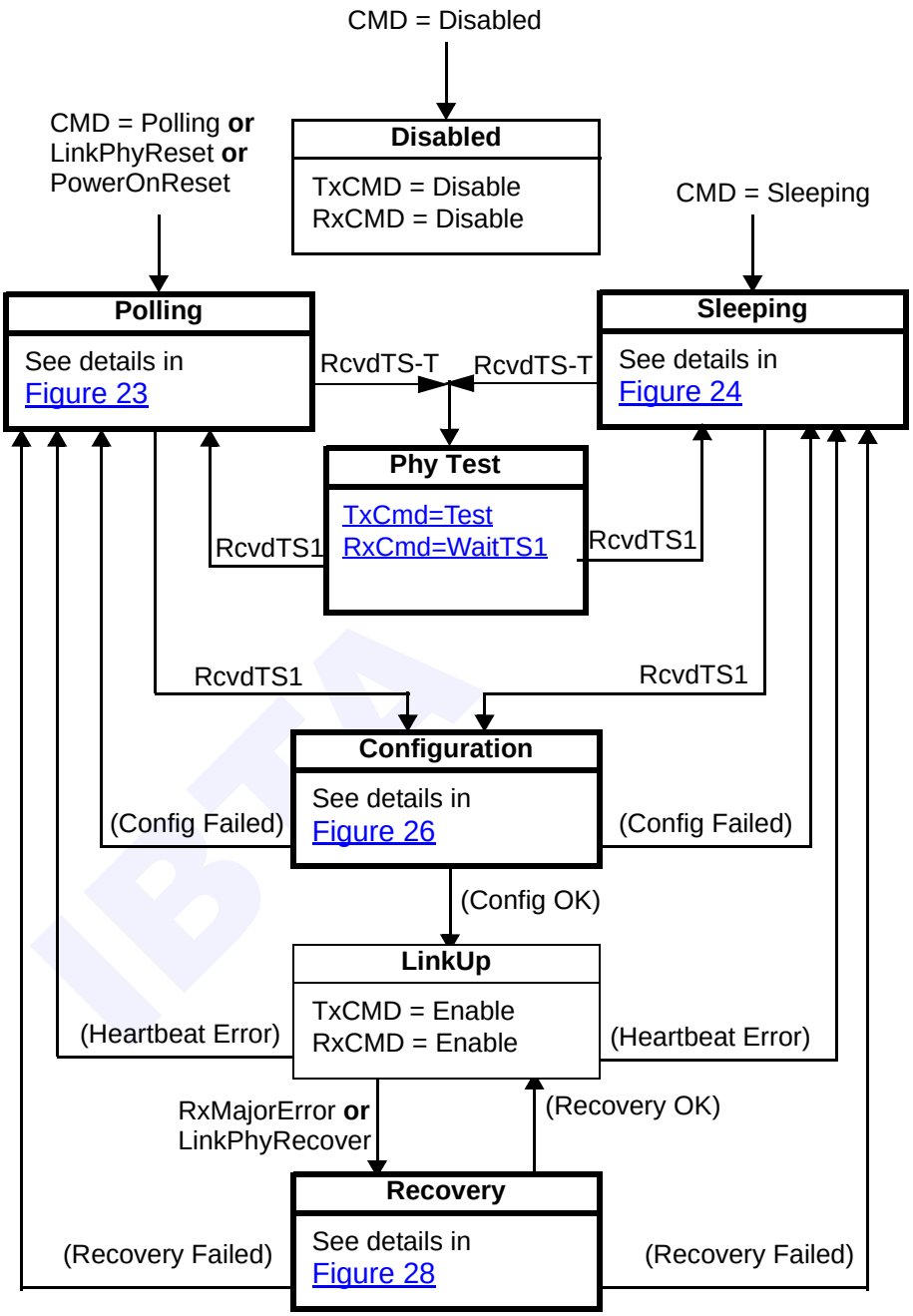


Figure 22 Link Training State Machine - Enhanced Signaling

5.6.4.1 DISABLED STATE

As the state name implies, there is no physical layer activity in this state, and both the transmitter and receiver are disabled. The transmitter outputs are forced to a quiescent condition, and receiver inputs are ignored. Entrance to and exit from the Disabled state are controlled by manage-

ment commands. In the Port Disabled State the following rules and commands **shall** be active:

1) TxCMD = Disable
2) RxCMD = Disable
3) LinkPhyStat = Down
4) ***SymbolErrorCounter*** is inhibited

5.6.4.2 POLLING STATES

The Polling States define the polling process used to initiate link configuration. The link training state machine cycles between the two states. These states are transmitting repeated TS1(s) for 2ms in Polling.Active followed by 100ms of a quiescent output in Polling.Quiet. In both states, the port's receiver is configured to detect TS1 ordered-sets. When TS1 is received, the link training state machine transfers control to the Configuration Super State. The states of the Polling Super State are expanded in the state graph shown in Figure 23.

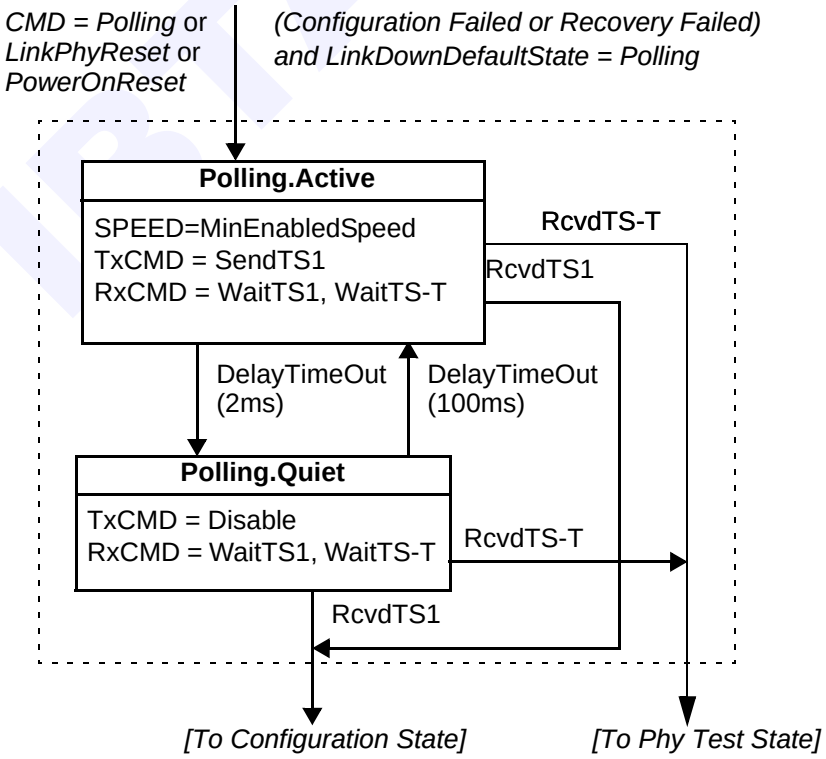


Figure 23 Polling Super State (Expanded)

5.6.4.2.1 POLLING.ACTIVE STATE

The Polling.Active state sends a stream of TS1s and at the same time enables detection of TS1s on all receiver lanes. The transmitter speed is set to the minimum enabled speed. In the Polling.Active state the following rules and commands **shall** be active:

1) SPEED = MinEnabledSpeed
2) TxCMD = SendTS1
3) RxCMD = WaitTS1, WaitTS-T
4) LinkPhyStat = Down
5) ***SymbolErrorCounter*** is inhibited.
6) ***LinkSpeedActive = LinkSpeedEnabled***
7) TS1 and SKIP ordered-sets are transmitted on all lanes.
8) If RxStatus = RcvdTS1, the next state is Configuration.
9) If RxStatus = RcvdTS-T, the next state is Phy Test
10) Else if DelayTimeOut(2ms), the next state is Polling.Quiet

5.6.4.2.2 POLLING.QUIET STATE

The Polling.Quiet state inhibits the transmitter and continues to enable detection of TS1 on all receiver lanes. When TS1 is detected, configuration can be started. In the Polling.Quiet state, the following rules and commands **shall** be active:

1) TxCMD = Disable
2) RxCMD = WaitTS1
3) LinkPhyStat = Down
4) ***SymbolErrorCounter*** is inhibited
5) If RxStatus = RcvdTS1, the next state is Configuration
6) If RxStatus = RcvdTS-T, the next state is Phy Test
7) Else If DelayTimeOut(100ms), the next state is Polling.Active

5.6.4.3 SLEEPING STATES

The Sleeping states are used to deactivate a link without powering off the port. When a port enters the sleeping state, it first disables its transmitter and receiver and then delays, allowing all link activity to cease. Following the delay, the receiver is enabled to detect TS1 ordered-sets. When a TS1 is detected, port configuration is started. The states of the Sleeping Super State are expanded in the state graph shown in Figure 24.

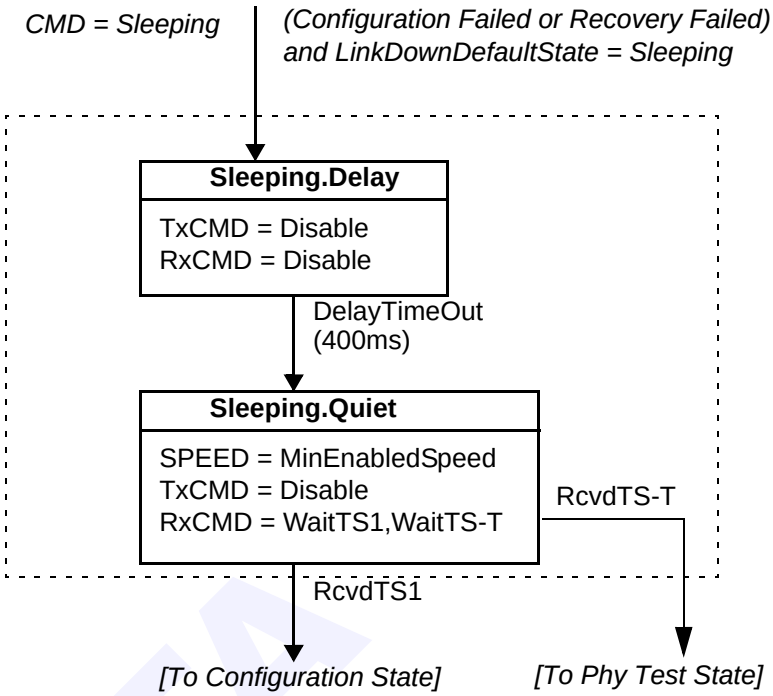


Figure 24 Sleeping Super State (Expanded)

5.6.4.3.1 SLEEPING.DELAY STATE

In the Sleeping Delay (Sleeping.Delay) state the transmitter is quiescent, and the receiver is disabled. The next state following the delay is Sleeping.Quiet. In the Sleeping.Delay state the following rules and commands **shall** be active:

1) TxCMD = Disable
2) RxCMD = Disable
3) LinkPhyStat = Down
4) ***SymbolErrorCounter*** is inhibited
5) If DelayTimeOut(400ms), the next state is Sleeping.Quiet

5.6.4.3.2 SLEEPING.QUIET STATE

The Sleeping.Quiet state inhibits the transmitter and at the same time enables detection of TS1 on all receiver lanes. The transmitter speed is set to the minimum enabled speed for when transmission resumes. When TS1 is detected, configuration can be started. In the Sleeping.Quiet state, the following rules and commands **shall** be active:

1) SPEED = MinEnabledSpeed

2) TxCMD = Disable
3) RxCMD = WaitTS1, WaitTS-T
4) LinkPhyStat = Down
5) ***SymbolErrorCounter*** is inhibited
6) ***LinkSpeedActive = LinkSpeedEnabled***
7) If RxStatus = RcvdTS1, the next state is Configuration
8) If RxStatus = RcvdTS-T, the next state is Phy Test

5.6.4.4 PHY TEST STATE

The Phy Test state allows the transmitter and receiver circuitry to be tested by external test equipment for compliance with transmitter and receiver specifications. This state is not used for normal operations. Operation of this state is described further in Section 5.12, "Physical Layer Compliance Testing," on page 169.Entry to this state occurs on receipt of a TS-T (Training Sequence for Test) ordered-set. In the Phy Test state, the following rules and commands **shall** be active:

1) SPEED = Value determined in TS-T ordered-set
2) RxCMD = WaitTS1
3) Transmitted signals will follow rules defined in Section 5.12, "Physical Layer Compliance Testing," on page 169.
4) If RxStatus = RcvdTS1, the next state is the LinkDownDefaultState (Polling or Sleeping)

5.6.4.5 CONFIGURATION STATES - LEGACY OPERATION

The Configuration Super State controls the configuration and training of the link. A low level protocol using TS1, TS2, and Idle Data is used to communicate the state of the two ports on the link.

The first step is to send a extended stream of TS1 ordered-sets indicating that the port has started the configuration process. This allows time for the physical connection to stabilize (debounce delay).

Following the delay, the second step of configuration begins by enabling the receivers to auto-configure and de-skew. The remote port may have completed receiver training and be transmitting TS2 ordered-sets. The receiver can auto-configure and de-skew while receiving TS1 or TS2 ordered-sets. The transmitter continues to send a stream of TS1 ordered-sets and waits for the receiver to complete configuration and de-skew (training).

The third step starts when the receiver has completed training, and the transmitter begins to send a stream of TS2 ordered-sets indicating that the local receiver is trained. A delay is started, and the port waits for the

remote port to indicate that its receiver has completed training. While waiting the received training sequences are checked for lane order errors.

At the end of the delay, if the remote port has not indicated that it is trained, a port implementing the optional lane reversal will reverse its transmit lanes, start a short delay, and continue to wait for the remote port to report that its receiver has completed training. While waiting the received training sequences are checked for lane order errors. If both ports are not trained at the end of the delay(s), configuration has failed. If the port completes training before the end of the delay(s), the final step of configuration starts when the port is both sending and receiving TS2 ordered-sets.

The final step of link configuration transmits the idle data stream and waits to receive idle data. When the port is both sending idle data and receiving idle data, configuration is complete, and the link is up. While waiting for idle data the receivers continue to verify TS2 lane order. Receiving TS1 in this state will force the restart of link configuration.

The states of Configuration Super State are expanded in the state graph shown in Figure 25.

IBTA

Note: This is the legacy state diagram, for products that do not support Enhanced Signaling. It has been modified for Enhanced Signaling as shown in the diagram in Figure 26.

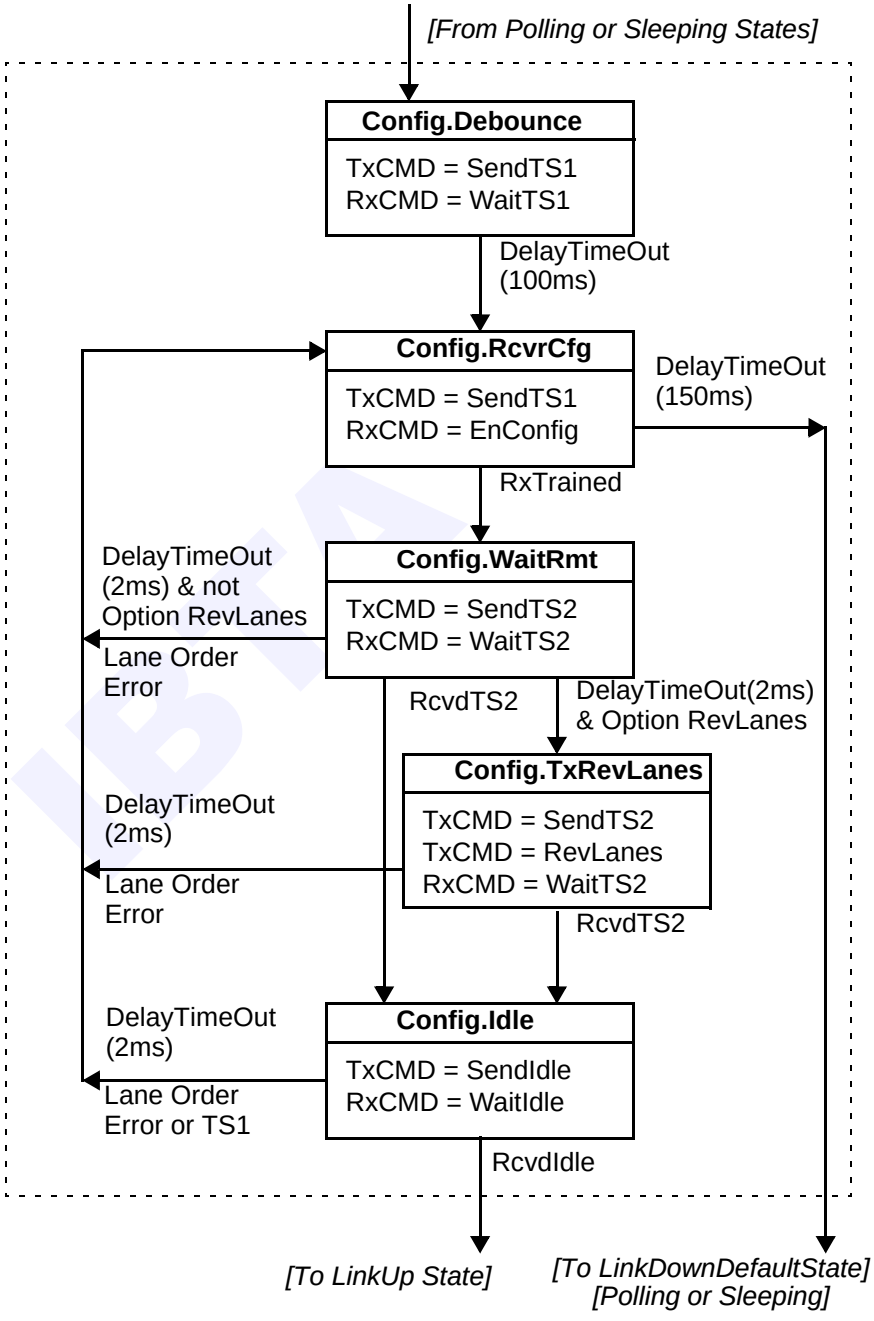


Figure 25 Configuration Super State (Expanded) - Legacy

5.6.4.6 CONFIGURATION STATES - ENHANCED SIGNALING

In this release of this specification, additional functions are added to the Configuration states to support enhanced signaling capabilities of the links, including higher speed operation at DDR (5.0 Gb/s) and QDR (10.0 Gb/s) bit rates, Link Heartbeat function (described in Section 5.11, "Link Heartbeat," on page 166), and support for transmitter and receiver equalization to better enable operation at higher bit rates over a variety of media. Several additional states and an additional training sequence (TS3) are added for the purpose of negotiating common use of these enhanced capabilities between the two peer ports on a link.

The first three steps of link initialization are the same as for legacy devices: a TS1 exchange assures that both transmitters are active, a TS2 exchange assures that both receivers are trained, and lane reversal assures correct polarity at each end of the link.

Following these steps, both ports transmit a TS3 training-set, indicating which enhanced capabilities they are enabled to support (DDR/QDR, Equalization, and Link Heartbeat). Each transmitter compares its own capabilities with the capabilities indicated in the TS3 received from the other end of the link. When both sides have sent and received TS3s, indicating agreement on common enhanced capabilities, they act to enable those capabilities. These negotiations occur at the lowest enabled speed, which will generally be the SDR speed.

The TS3 exchange for negotiating link speed operates by having each port send a bit map of the active speeds, in the format used in the ***SM.PortInfo(LinkSpeedEnabled)***, and ***SM.PortInfo(LinkSpeedSupported)*** MADs. The first time the ports traverse the states negotiating enhanced capabilities, they send TS3 ordered-sets indicating the link speeds enabled by the SM (***LinkSpeedEnabled***). If, however, a particular speed is attempted, and found to be unsupportable (for example, link attenuation is too high for intact data transmission), the bit corresponding to this speed is cleared to 0 in the ***LinkSpeedActive*** field, and in the following TS3 ordered-sets. A logical AND operation on the transmitted and received fields is used to determine the speed(s) supported on both ports, and further operation is attempted at the highest speed that is active on both.

A link pair will go through TS3 exchange at least twice, to ensure that the supported capabilities on both ends of the link are consistent, and supported by the link media. Links will always be ***Active*** at the highest speed which is ***Enabled*** in both ports driving the link, and supported by the link transmission medium. During the negotiation phase of the configuration super state, multiple speeds bits may be set to 1 in the ***LinkSpeedActive*** bit map exchanged in TS3 ordered-sets, but once consistent and correct operation has been established, only one ***LinkSpeedActive*** bit will be set

to 1. Links always operate at the same speed in both directions, and a port always configures its transmitter and receiver to operate at the same speed.

The TS3 exchange is also used to establish whether adaptive driver de-emphasis equalization training is requested. If either side requested adaptive equalization, both sides will participate. For equalization training, each transmitter sends idle data at the maximum speed commonly active in both ports (typically DDR or QDR, but possibly SDR as well), for a 100 millisecond period. The 100 ms. period is split into a 36 ms period where data is sent with a default transmitter equalization setting, for receivers which don't support adaptive equalization, and 16 separate adaptive transmitter equalization setting slots, of 4ms duration each. On each 4 ms. slot, the transmitter sends data using one of a set of vendor-dependent pre-emphasis settings. The receiver, on each slot, attempts to receive intact data, optionally adjusting a receiver equalization setting. The receiver determines which is the first slot that provides a signal good enough to be equalized at the receiver.

Implementation Note

On definition of the terms ***Enabled***, ***Supported***, and ***Active***, in the context of describing link widths and speeds:
The set of ***Enabled*** widths or speeds is determined by the subnet manager, and will be equal to the set of or a subset of the widths or speeds ***Supported*** by the design and implementation of the port. The ***Active*** width or speed is determined by a combination of the implemented capabilities in the two ports of a link, and by the link used to connect them. For example, a pair of ports sharing a link may both ***Support*** and be ***Enabled*** for SDR/DDR/QDR operation (7), but the cable over which they are connected may only allow intact data transmission at SDR speed. In this case, the port Link/Phy logic would set the ***LinkSpeed-Active*** value to 1 on both ports. The actual operating link speed or link width, as determined by the Link/Phy logic during the Configuration super state, is reflected in the ***Active*** fields.

Implementation Note:

Equalization is used to compensate for signal distortion (attenuation or dispersion) that occurs in the medium between transmitter and receiver. The distortion being compensated for may take a variety of forms, including attenuation of high-frequency components, frequency-dependent crosstalk, or frequency-dependent reflections. Signal-processing circuitry in the transmitter, or the receiver, or (typically) both, can equalize out some of the effect of some distortions. This equalization may, for example, take the form of emphasizing high frequencies (fast toggles) or de-emphasizing low frequencies (strings of 1s or 0s), or of de-emphasizing high frequencies, in cases where frequency-dependent crosstalk is the major distortion. The equalization circuitry should therefore be adaptive to the characteristics of the particular link medium.

Achieving ideal adaptive equalization over the wide variety of copper and optical media of the various distances used for InfiniBand™ links is very complex problem. The algorithm described here (16 different adaptive driver de-emphasis settings, which the receiver selects from, after attempting adaptive receiver equalization on each) provides a mechanism for both transmitter and receiver equalization with moderate complexity. The algorithm also allows interoperability with ports that don't implement adaptive equalization.

Since the maximum physical length of an InfiniBand cable this release using long wavelength optical fiber links is 10 km, with speed-of-light round-trip latency on the order of 100 microseconds (0.1 ms), the two ends of a link may be un-synchronized with each other by this much. This synchronization period determines, along with the (-1us/ +126us) clock tolerance and 100 ppm variation in clock frequency described in Section 5.6.5, "State Machine Delays and Timeouts," on page 147, the window of proper transmitter-equalized data within the 4ms slot period, and must be taken into account at the receiver when determining transmitter equalization setting slots.

Following attempted transmission at the maximum speed that is active on both ports, including transmitter and receiver adaptive equalization, the ports then transition through an WaitCfgMaxBothActive state, to ensure that both ports have completed attempting operation at the maximum active speed, and then do another round of TS3 exchange. This round allows both ports to indicate whether they were able to receive intact data at the attempted speed, and if so, to allow each port's receiver to indicate to its peer port's transmitter which of the 16 4 ms slots was the first that had a transmitter equalization setting with good enough signal for receiver

data recovery. The use of the first good slot allows the transmitter/receiver pair to optimize the link's overall power usage.

The TS3 exchange also establishes the use of Link Heartbeat. Link Heartbeat is used (a) to determine that both receivers on a link are receiving data from their peer transmitters rather than from crosstalk, and (b) to determine the link round-trip latency for data to traverse the link in both directions. Link Heartbeat is required at DDR and QDR speeds, due to the small received signal swings and lack of a signal detect capability at the higher bit rates. Link Heartbeat is also useful at SDR speed, and is required for devices supporting Enhanced Signaling, even if they only support SDR. A bit in the exchanged TS3 indicates that Link Heartbeat is requested. If both ports request the use of Link Heartbeat, it is used when the link transitions to the LinkUp state. Further details on Link Heartbeat may be found in Section 5.11, "Link Heartbeat," on page 166.

Once both ports on a link have exchanged consistent TS3s, and assured correct operation at the highest commonly-***Active*** speed, they will transition to the Config.WaitCfgEnhanced state, to indicate completion of configuration for enhanced operation, and from there to Config.LinkUp.

On connection with a port that does not support Enhanced Signaling, or during legacy operation, the port state machine would traverse the states directly from top to bottom, at SDR rate, bypassing the Config.Test, Config.WaitRmtTest, and Config.WaitCfgEnhanced states.

The states of Configuration Super State for ports supporting Chapter 5: operation are expanded in the state diagram shown in Figure 26.

A simplified diagram of the State graph is shown in Figure 27, along with a typical control flow through the states, showing negotiation from SDR up to higher speed operation.

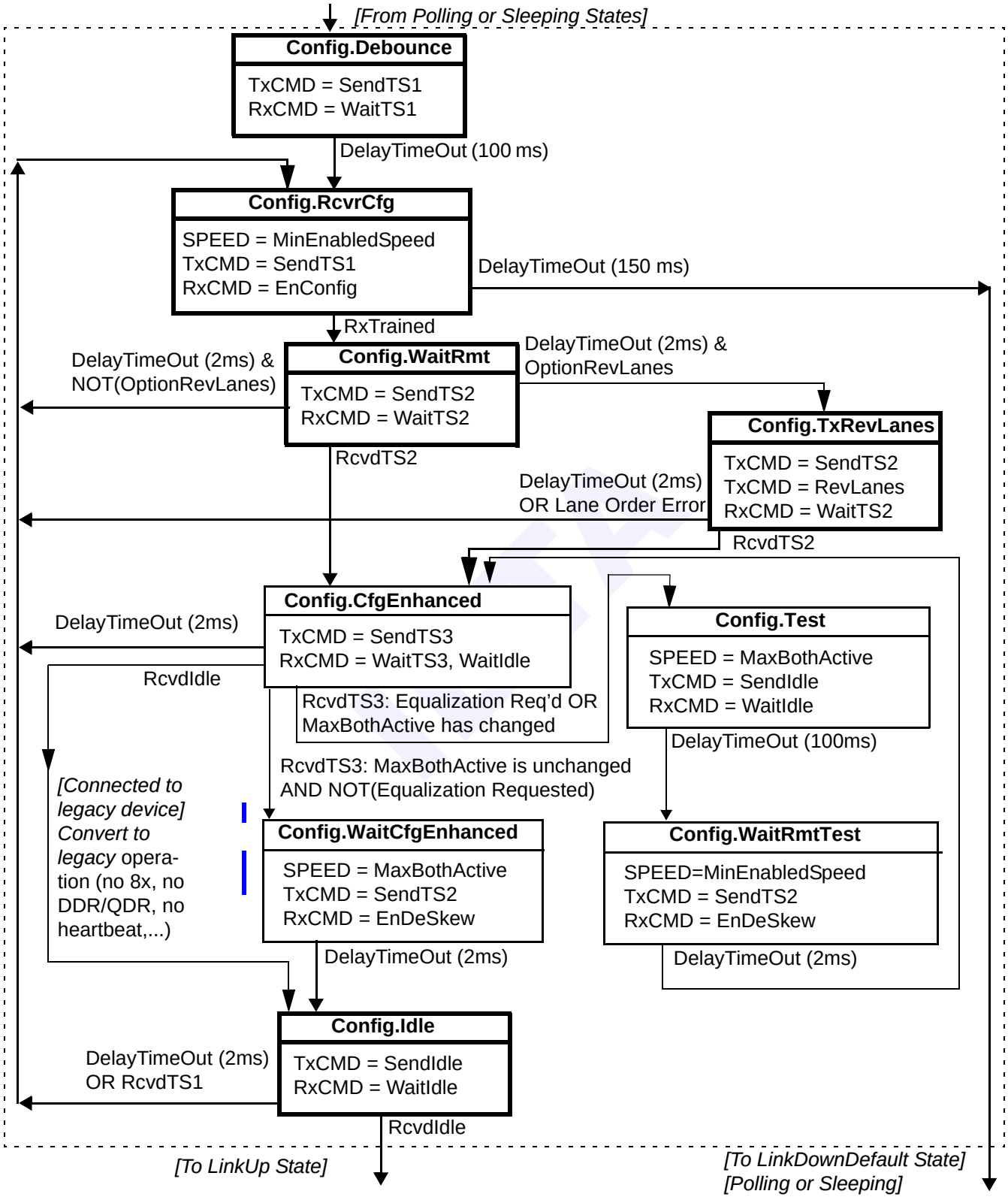


Figure 26 Configuration Super State (Expanded) - Enhanced Signaling

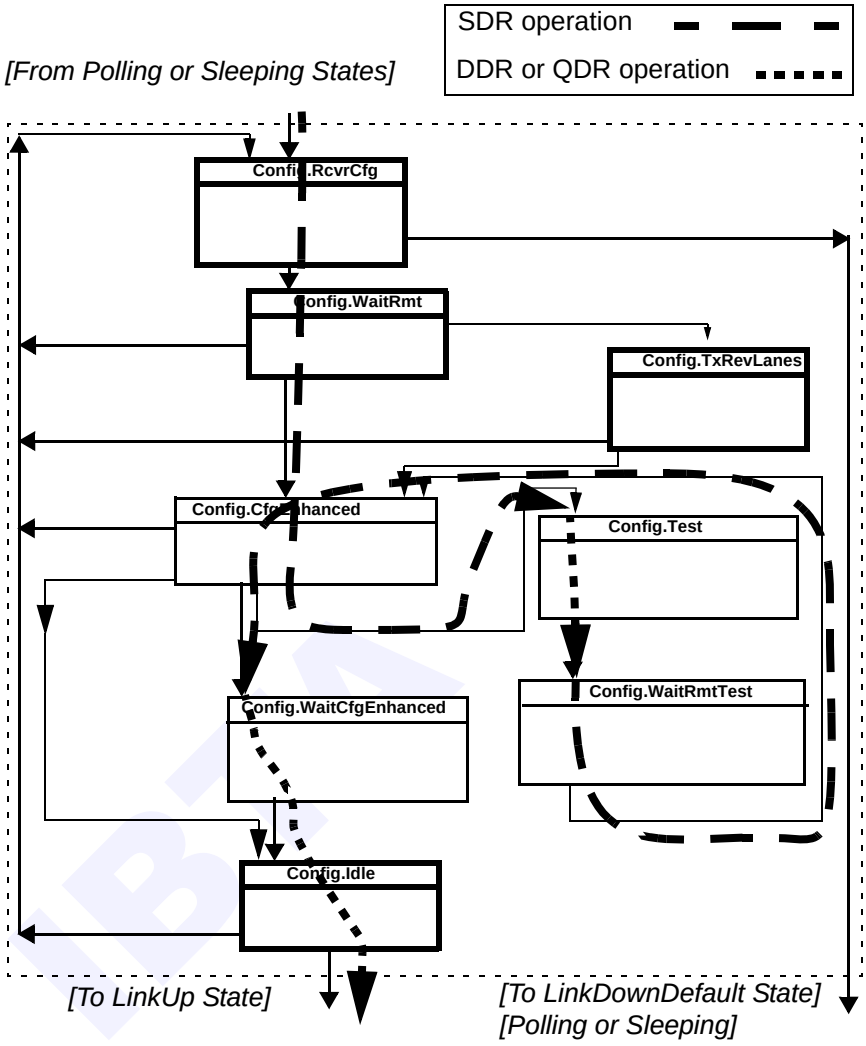


Figure 27 Typical Control Flow through Configuration Super State - Enhanced Signaling

5.6.4.6.1 CONFIG.DEBOUNCE STATE

In the debounce state (Config.Debounce), the transmitter sends a series of TS1 ordered-sets on all lanes. The receiver status is not checked in this state. In receivers with adaptive receiver equalization, the 100 ms duration of this state may optionally be used for determining receiver equalization settings. Operation in Config.Debounce is always at SDR speed. In the Config.Debounce state, the following rules and commands **shall** be active:

1) TxCMD = SendTS1
2) RxCMD = WaitTS1
3) LinkPhyStat = Down

4) TS1 and SKIP ordered-sets are transmitted on all lanes
5) ***SymbolErrorCounter*** is inhibited
6) If DelayTimeOut(100ms), next state is Config.RcvrCfg.

5.6.4.6.2 CONFIG.RCVRCFG STATE

In the Receiver Configure (Config.RcvrCfg) state, the transmitter sends a stream of TS1 ordered-sets on all lanes. The receiver uses received TS1 or TS2 ordered-sets to identify the width of the remote port and to optionally correct lane reversal and inverted lane data. When the receiver has completed configuration and training, it reports Receiver Trained status. The transmitter speed is set to the minimum enabled speed. In the Config.RcvrCfg state, the following rules and commands **shall** be active:

1) SPEED = MinEnabledSpeed
2) TxCMD = SendTS1
3) RxCMD = EnConfig
4) LinkPhyStat = Down
5) TS1 and SKIP ordered-sets are transmitted on all lanes
6) ***SymbolErrorCounter*** is enabled
7) ***LinkSpeedActive = LinkSpeedEnabled***
8) If the RxStatus = RxTrained, the next state is Config.WaitRmt.
9) Else if DelayTimeOut (150 ms) the next state is the LinkDownDefaultState (Polling or Sleeping).

5.6.4.6.3 CONFIG.WAITRMT STATE

In the Wait Remote (Config.WaitRmt) state, the transmitter sends a series of TS2 ordered-sets on all lanes. The receiver monitors the configured physical lanes waiting for a transition from TS1 to TS2 ordered-sets indicating the remote port has completed its configuration. If the lane reversal option is supported and the delay has timed out, the next state will reverse the port's transmit lanes. If the lane reversal option is not supported and the delay has timed out, the next state will be retry receiver configuration. In the Config.WaitRmt state the following rules and commands **shall** be active:

1) TxCMD = SendTS2 (minimum of 16 TS2s)
2) RxCMD = WaitTS2
3) LinkPhyStat = Down
4) TS2 and SKIP ordered-sets are transmitted on all lanes
5) ***SymbolErrorCounter*** is enabled
6) If received "lane order error" the next state is Config.RcvrCfg.

7) Else If the RxStatus = RcvdTS2, the next state is Config.CfgEnhanced for ports implementing Rel. 1.2 Enhanced Signaling capability. For ports implementing only legacy functionality, the next state is Config.Idle.

8) Else If "Lane Reversal Option" and DelayTimeOut(2ms), the next state is Config.TxRevLanes.

9) Else if not "Lane Reversal Option" and DelayTimeOut(2ms), the next state is Config.RcvrCfg.

5.6.4.6.4 CONFIG.TXREVLANES STATE

In the reverse transmit lanes (Config.TxRevLanes) state, the transmitter reverses its lanes and continues to sends a series of TS2 ordered-sets on all lanes. The receiver monitors the configured physical lanes waiting for a transition from TS1 to TS2 ordered-sets, indicating the remote port has completed its configuration. In the Config.TxRevLanes state, the following rules and commands **shall** be active:

1) TxCMD = SendTS2 (minimum of 16 TS2s)

2) TxCMD = RevLanes

3) RxCMD = WaitTS2

4) LinkPhyStat = Down

5) TS2 and SKIP ordered-sets are transmitted on all lanes

6) ***SymbolErrorCounter*** is enabled

7) If received "lane order error" the next state is Config.RcvrCfg

8) Else If the RxStatus = RcvdTS2, the next state is Config.CfgEnhanced.

9) Else If DelayTimeOut(2ms), the next state is Config.RcvrCfg.

5.6.4.6.5 CONFIG.CFGENHANCED STATE - RELEASE 1.2 ENHANCED SIGNALING

In the Configure Enhanced Capabilities state (Config.CfgEnhanced) the transmitter sends TS3 (Training Sequence 3) ordered-sets to negotiate with the link peer port what enhanced capabilities will be used. As described in Section 5.3.2.4, "Training Sequence Three Ordered-Set (TS3)," on page 99, a TS3 ordered-set includes fields which allow the transmitter to indicate

- which link speeds may be attempted (in a form consistent with the ***SM.PortInfo(LinkSpeedEnabled)*** and ***SM.PortInfo(LinkSpeedSupported)*** fields described in Section 5.4.2, "Status Outputs (MAD get)," on page 107),
- whether or not adaptive transmitter pre-emphasis is requested, and

- whether or not the LINK HEARTBEAT functionality, required for DDR and QDR operation as described in Section 5.11, “Link Heartbeat,” on page 166, is enabled, and
- which transmitter equalization setting should be used, if equalization training has already occurred in the Config.Equalization state.

In the Config.CfgEnhanced state the following rules and commands **shall** be active:

1) TxCMD = SendTS3
2) RxCMD = WaitTS3
3) LinkPhyStat = Down
4) TS3 and SKIP ordered-sets are transmitted on all lanes
5) SymbolErrorCounter is inhibited
6) If RxStatus = RcvdTS3, and the exchanged TS3 indicate both (a) a request for equalization on either port, OR (b) MaxBothActive changed from previous visit (i.e., first time on state since Config.RcvrCfg, or equalization failed on one of the peers), the next state is Config.Test. Also, ***LinkSpeedActive*** is set equal to the AND of the current ***LinkSpeedActive*** and the SpeedActive field of the received TS3 to determine the speed(s) active on both ports.
7) Else if RxStatus = RcvdTS3, and either no further equalization is requested AND MaxBothActive has not changed from previous visit, and ***LinkSpeedActive*** is set to the highest bit set on both peers, the next state is Config.WaitCfgEnhanced.
8) Else If RxStatus = RcvdIdle (indicating the port is connected to a legacy device port), the next state is Config.Idle.
   - For legacy compatibility, an 8x device should assert RxStatus = RcvdIdle if it receives Idles on only lanes 0...3, as described in Section 5.6.7.7, “RxCMD = WaitIdle,” on page 154.
   - After following this transition, since the port is connected to a legacy device, all Rel. 1.2 Enhanced Signaling capabilities (DDR/QDR, heartbeat, etc.) are inactivated. Furthermore, an 8x device reverting to legacy operation **shall** operate with 4x or 1x link width, with only lanes 0-3 or lane 0 configured.
9) Else if DelayTimeOut (2 ms), the next state is Config.RcvrCfg.

5.6.4.6.6 CONFIG.TEST STATE - RELEASE 1.2 ENHANCED SIGNALING

In the Configure Test state (Config.Test) the transmitter will send idle data at the maximum speed indicated in the ***LinkSpeedActive*** bit map, for a 100 millisecond period. The 100 ms. period is split into a 36 ms period of default transmitter equalization time, and 16 separate adaptive transmitter equalization setting slots of 4ms duration each. On each 4 ms slot, the transmitter sends one of a vendor-dependent set of adaptive equalization

settings. The receiver, on each slot, attempts to receive intact data (valid and supported 8b/10b code groups, with correct running disparity), while optionally adjusting receiver equalization settings.

The receiver equalization techniques and settings used are vendor-dependent, don't require transmitter/receiver handshaking across a link. Therefore, the receiver equalization is unspecified here.

Once the 100 ms period of the Config.Test state is over, the receiver decides which of the 16 adaptive transmitter equalization settings was the first usable one. This request is indicated in the TS3 ordered-sets transmitted in the Config.CfgEnhanced state following this Config.Test state.

Transmitter and receiver adaptive equalization capabilities are optional, and may not be implemented in some devices. In this case, the non-implementing device will return a TS3 with DDSV and the DDS bits all cleared to 0, and the attached device will use the default equalization setting of the first 36 ms.

In the Config.Test state the following rules and commands **shall** be active:

1) SPEED = MaxBothActive
2) TxCMD = SendIdle
3) RxCMD = WaitIdle
4) Idle and SKIP ordered-sets are transmitted on configured lanes
5) LinkPhyStat = Down
6) ***SymbolErrorCounter*** is enabled
7) If DelayTimeOut(100 ms), next state is Config.WaitRmtTest.
   - If the receiver was unable to recover data from the received signal without incurring any minor link errors, the bit indicating the currently-active speed, MaxBothEnabled, should be cleared to 0 in ***LinkSpeedsActive.*** This modified Speeds mask, indicating that the current speed is not supported, is reflected in further transmitted TS3 ordered-sets.
   - A receiver **MAY** use SKIP ordered-sets to ensure that the requirement of total link skew of 6 symbol times (60 UI) or less has been met. Please see Table 28, Table 41, Table 60, Table 61, and Table 62. If the total link skew is measured to be more than 6 symbol times, the bit indicating the currently-active speed, MaxBothEnabled, should be cleared to 0 in LinkSpeedsActive. This modified Speeds mask, indicating that the current speed is not supported, is reflected in further transmitted TS3 ordered-sets.

5.6.4.6.7 CONFIG.WAITRMTTEST STATE - RELEASE 1.2 ENHANCED SIGNALING

In the Configure Wait for Remote Test state (Config.WaitRmtTest) the transmitter sends a series of TS2 ordered-sets at the MinEnabledSpeed transmitter speed to indicate to the link peer that it has completed attempting operation at the maximum speed commonly active in both ports, with attempted equalization training. In the Config.WaitRmtTest state the following rules and commands **shall** be active:

1) SPEED = MinEnabledSpeed
2) TxCMD = SendTS2
3) RxCMD = EnDeskew
4) LinkPhyStat = Down
5) TS2 and SKIP ordered-sets are transmitted on all lanes at the currently-active transmitter speed.
6) ***SymbolErrorCounter*** is inhibited
7) If DelayTimeOut (2 ms), the next state is Config.CfgEnhanced

5.6.4.6.8 CONFIG.WAITCFGENHANCED STATE - RELEASE 1.2 ENHANCED SIGNALING

In the Configure Wait Remote Configure Enhanced state (Config.WaitCfgEnhanced) the transmitter sends a series of TS2 ordered-sets at the currently-active transmitter speed to indicate to the link peer that it has (a) completed TS3 exchange for negotiating enhanced capabilities, and (b) committed to operation at the speed negotiated in the Config.CfgEnhanced TS3 exchange and attempted in a previous Config.Test state. In the Config.WaitCfgEnhanced state the following rules and commands **shall** be active:

1) SPEED = MaxBothActive
2) TxCMD = SendTS2
3) RxCMD = EnDeSkew
4) LinkPhyStat = Down
5) TS2 and SKIP ordered-sets are transmitted on all lanes at the currently-active transmitter speed.
6) ***SymbolErrorCounter*** is inhibited
7) If DelayTimeOut (2 ms), the next state is Config.Idle.

5.6.4.6.9 CONFIG.IDLE STATE

In the configuration idle (Config.Idle) state, both the local and remote ports have completed configuration and training. The port now transmits link idle data at the configured active speed. The receiver monitors the configured physical lanes, waiting for a transition from TS2 ordered-sets to Idle Data. When Link Idle Data is received the link is up. In the Config.Idle state, the following rules and commands **shall** be active:

1) TxCMD = SendIdle
2) RxCMD = WaitIdle
3) LinkPhyStat = Down
4) Idle data and SKIP ordered-sets are transmitted on configured lanes only.
5) ***SymbolErrorCounter*** is enabled
6) If RxStatus = RcvdTS1 or If DelayTimeOut(2ms), the next state is Config.RcvrCfg.

5.6.4.7 LINKUP STATE

The LinkUp state is the normal operational state of the port. Packets provided by the Link Layer are transmitted, and received packets are forwarded to the Link Layer. This is the only state in which the Link Physical logical status signal LinkPhyStat is assigned the value “Up”, indicating that the link is available to transfer packets. In the LinkUp state, Physical Link error handling is enabled. The Physical Link Error Handling is defined in Section 5.7, “Link Physical Error Handling,” on page 156. In the LinkUP state the following rules and commands **shall** be active:

1) TxCMD = Enable
2) RxCMD = Enable
3) LinkPhyStat = Up
4) The ***SymbolErrorCounter*** is enabled
5) If RxStatus = RxMajorError or LinkPhyRecover, the next state is Recovery.
6) If RxStatus = RxHeartbeatError,
   a) The ***LinkDownedCounter*** is incremented
   a) The ***LinkRoundTripLatency*** value is reset to 0xFFFF FFFF
   b) The next state is the LinkDownDefaultState (Polling or Sleeping)

5.6.4.8 LINK ERROR RECOVERY STATES

The Link Error Recovery Super State controls port error recovery. The recovery process is triggered when an error is detected by the Link Layer or Link/Physical Layer. The Physical Layer error handling logic monitors minor errors and, when a rate threshold is reached or when a major error is detected, triggers recovery. (See Section 5.7 on page 156)

The recovery process starts by sending a stream of TS1 ordered-sets to trigger error recovery at the remote port. When TS1 or TS2 is received by the local receiver, it will use the current configuration (width, lane reversal, serial data inversion, and speed) to retrain.

The second step of recovery starts when the receiver has completed retraining. The port starts sending a stream of TS2 ordered-sets. The port then waits for the remote port to complete retraining. When the port is both sending and receiving TS2, both receivers have been retrained, and the second step of recovery is complete.

In the final step of link recovery, the port transmits an idle data stream and waits for idle data. When the port is both sending and receiving idle data, recovery is complete, and the link is up.

If the recovery process fails at any step, the Link/Physical Layer state machine returns to its link down default state.

The states of Recovery Super State are expanded in the state graph shown in Figure 28.

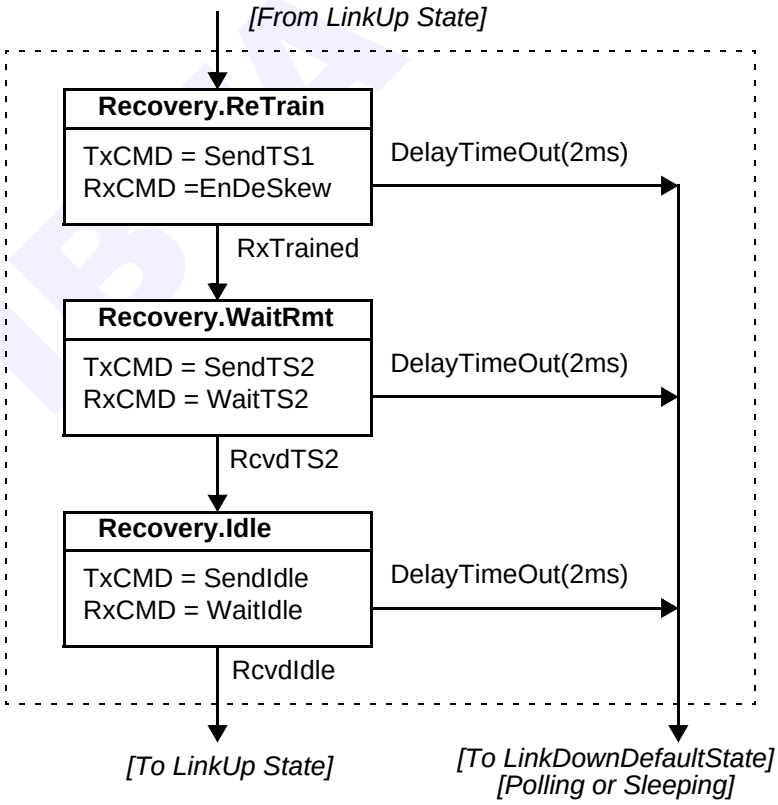


Figure 28 Recovery Super State (Expanded)

5.6.4.8.1 RECOVERY.RETRAIN STATE

In the Recovery Retrain (Recovery.Retrain) state, the transmitter is commanded to send a series of TS1 ordered-sets using the currently configured lanes. The receiver is enabled to reacquire symbol sync and then to

de-skew the lanes. In the Recovery.Retrain state the following rules and commands **shall** be active:

1) TxCMD = SendTS1 (minimum of 16 TS1s)
2) RxCMD = EnDeSkew
3) LinkPhyStat = Down
4) TS1 and SKIP ordered-sets are transmitted on configured lanes only.
5) The port ***SymbolErrorCounter*** is enabled.
6) If the RxStatus = RxTrained, the next state is Recovery.WaitRmt.
7) Else if DelayTimeOut(2ms),
   a) The ***LinkDownedCounter*** is incremented
   b) The ***LinkRoundTripLatency*** value is reset to 0xFFFF FFFF
   c) The next state is the LinkDownDefaultState (Polling or Sleeping)

5.6.4.8.2 RECOVERY.WAITRMT STATE

In the Recovery Wait Remote state (Recovery.WaitRmt), the transmitter is commanded to send a series of TS2 ordered-sets on the configured lanes. The receiver monitors the receive streams on the configured lanes. In the Recovery.WaitRmt state the following rules and commands **shall** be active:

1) TxCMD = SendTS2 (minimum of 16 TS2s)
2) RxCMD = WaitTS2
3) LinkPhyStat = Down
4) TS2 and SKIP ordered-sets are transmitted on configured lanes only
5) The port ***SymbolErrorCounter*** is enabled.
6) If the RxStatus = RcvdTS2, the next state is Recovery.Idle.
7) Else if DelayTimeOut(2ms)
   a) The ***LinkDownedCounter*** is incremented
   b) The ***LinkRoundTripLatency*** value is reset to 0xFFFF FFFF
   c) The next state is the LinkDownDefaultState (Polling or Sleeping)

5.6.4.8.3 RECOVERY.IDLE STATE

In the Recovery Idle state (Recovery.Idle), both ports have retrained, and the transmitter is commanded to send on the configured lanes. The receiver monitors the receive streams on the configured lanes waiting for idle data. In the Recovery.Idle state, the following rules and commands **shall** be active:

1) TxCMD = SendIdle (minimum of 16 symbol times)

2) RxCMD = WaitIdle
3) LinkPhyStat = Down
4) Idle Data and SKIP ordered-sets are transmitted on configured lanes only.
5) The port ***SymbolErrorCounter*** is enabled.
6) If RxStatus = RcvdIdle,
   a) The ***LinkErrorRecoveryCounter*** is incremented
   b) The next state is LinkUp.
7) Else If DelayTimeOut(2ms),
   a) The ***LinkDownedCounter*** is incremented
   b) The ***LinkRoundTripLatency*** value is reset to 0xFFFF FFFF
   c) The next state is the LinkDownDefaultState (Polling or Sleeping)

5.6.5 STATE MACHINE DELAYS AND TIMEOUTS

Operations in certain states involve delays or time-out periods. These time-out periods start as the state is entered and are cleared when a transition to a different state is taken.

C5-10: All ports **shall** use the delays and tolerances defined in Section 5.6.5, “State Machine Delays and Timeouts,” on page 147.

The delay timeouts used by the link training state machine **shall** have the following tolerances:

1) DelayTimeOut (2ms): 2ms (-1us/ +126us)
2) DelayTimeOut (100ms): 100ms (-1us/ +126us)
3) DelayTimeOut (150ms): 150ms (-1us/ +126us)
4) DelayTimeOut (400ms): 400ms (-1us/ +126us)

Note: The delay tolerances specified above do not include plus/minus 100ppm allocated to the clock oscillator. The two tolerances are cumulative (delay tolerance + oscillator tolerance).

5.6.6 TRANSMITTER INTERFACE AND BEHAVIOR

The link training state machine controls the transmitter behavior using a set of transmitter commands (TxCMD). The transmitter reports its status as TxStatus. The expected behavior of each command is specified in the following paragraphs.

C5-11: All ports **shall** implement the transmitter behaviors defined in Section 5.6.6, “Transmitter Interface and Behavior,” on page 147 but are not required to implement the optional lane reversal function.

o5-3: All ports that implement the optional lane reversal function **shall** implement the behavior as defined in Section 5.6.6.6, "TxCMD = RevLanes," on page 150.

5.6.6.1 TxCMD = DISABLE

When transmit command is "disabled", the transmitter output signals on all physical lanes **shall** be driven to a quiescent condition.

5.6.6.2 TxCMD = SENDTS1

The SendTS1 command instructs the transmitter to send a sequence of training sequence one ordered-sets (TS1). When commanded to send TS1 ordered-sets, the transmitter **shall** implement the following rules:

1) The first TS1 ordered-set **shall** be transmitted at the next available ordered-set or symbol boundary. The packet currently being transmitted **may** be terminated abnormally to send TS1 ordered-sets that trigger error recovery.
2) The first TS1 ordered-set transmitted **should** force the current running disparity in all lanes to the same value (positive or negative)
3) The SKIP ordered-sets have the highest transmit priority. When scheduled, a SKIP ordered-set **shall** be transmitted at the next ordered-set boundary
4) Complete TS1 ordered-sets **shall** be transmitted back-to-back as long as the SendTS1 command is active.
5) At a minimum, sixteen (16) TS1 ordered-sets **shall** be transmitted.
6) In the Polling and Configuration Super States:
   a) The SKIP ordered-sets **shall** be transmitted on all enabled physical lanes.
   b) The TS1 ordered-sets **shall** be transmitted on all enabled physical lanes.
7) In the Recovery Super State:
   a) The SKIP ordered-sets **shall** be transmitted only on configured physical lanes.
   b) The TS1 ordered-sets **shall** be transmitted only on configured physical lanes.
   c) Lanes not included in the configured link width (unused) **shall** be forced to a quiescent condition.

5.6.6.3 TxCMD = SENDTS2

The SendTS2 command instructs the transmitter to send a sequence of training sequence two ordered-sets (TS2). When commanded to send TS2 ordered-sets, the transmitter will implement the following rules.

1) The first TS2 ordered-set **shall** be transmitted at the next available ordered-set boundary.
2) Complete TS2 ordered-sets **shall** be transmitted back-to-back as long as the SendTS2 command is active.
3) At a minimum, sixteen (16) TS2 ordered-sets **shall** be transmitted.
4) The SKIP ordered-sets have the highest transmit priority. When scheduled, a SKIP ordered-set **shall** be transmitted at the next packet or ordered-set boundary.
5) In the Config.WaitRmt and Config.TxRevLane States:
   a) The SKIP ordered-sets **shall** be transmitted on all enabled physical lanes.
   b) The TS2 ordered-sets **shall** be transmitted on all enabled physical lanes.
6) In the Recovery.WaitRmt, the Config.WaitRmtTest, and the Config.WaitCfgEnhanced State:
   a) The SKIP ordered-sets **shall** be transmitted only on configured physical lanes.
   b) The TS2 ordered-sets **shall** be transmitted only on configured physical lanes.
   c) Lanes not included in the configured link width (unused) **shall** be forced to a quiescent condition.

5.6.6.4 TxCMD = SendTS3

The SendTS3 command instructs the transmitter to send a sequence of training sequence three ordered sets (TS3). When commanded to send TS3 ordered-sets, the transmitter will implement the following rules:

1) The first TS3 ordered set **shall** be transmitted at the next available ordered-set boundary.
2) Complete TS3 ordered sets **shall** be transmitted back-to-back as long as the SendTS3 command is active.
3) At a minimum, sixteen (16) TS3 ordered sets **shall** be transmitted.
4) The SKIP ordered sets have the highest transmit priority. When scheduled a SKIP ordered set **shall** be transmitted at the next packet or ordered set boundary.
5) The SKIP ordered-sets **shall** be transmitted on all enabled physical lanes.
6) The TS3 ordered-sets **shall** be transmitted on all enabled physical lanes.

IBTA

5.6.6.5 TxCMD = SendIdle

The SendIdle command instructs the transmitter to send the link idle data sequence. When commanded to send idle data, the transmitter **shall** implement the following rules.

1) The SKIP ordered-sets have the highest transmit priority. When scheduled, a SKIP ordered-set **shall** be transmitted as soon as possible by interrupting link idle data sequence transmission.
2) The SKIP ordered-sets **shall** be transmitted only on configured physical lanes.
3) The link idle data sequence **shall** be transmitted only on configured physical lanes.
4) Lanes not included in the configured link width (unused) **shall** be forced to a quiescent condition.
5) At a minimum, sixteen (16) symbols times of link idle data **shall** be transmitted on each physical lane.

5.6.6.6 TxCMD = RevLanes

The Reverse Lanes command is an optional modifier to the SendTS2 command. The RevLanes command instructs the transmitter to reverse the order of its transmit lanes. When commanded to reverse lanes, the transmitter **shall** implement the following rules.

1) The operation of the SendTS2 command **shall** not be interrupted.
2) The lane swap **shall not** cause the remote receiver to detect an error.
3) Four wide (4x) port **shall** reverse lanes causing the following lane swaps: 0 to 3, 1 to 2, 2 to 1, and 3 to 0.
4) Eight wide (8x) port **shall** reverse lanes causing the following lane swaps: 0 to 7, 1 to 6, 2 to 5, 3 to 4, 4 to 3, 5 to 2, 6 to 1, and 7 to 0.
5) Twelve wide (12x) port **shall** reverse lanes causing the following lane swaps: 0 to 11, 1 to 10, 2 to 9, 3 to 8, 4 to 7, 5 to 6, 6 to 5, 7 to 4, 8 to 3, 9 to 2, 10 to 1, and 11 to 0.

5.6.6.7 SPEED = MaxBothActive

This command configures the transmitter to transmit data at the maximum rate commonly enabled on both ports of the link that is also supported by the link medium.

5.6.6.8 SPEED = MinEnabledSpeed

This command configures the transmitter to transmit data at the minimum speed enabled on the port, generally the SDR rate.

5.6.6.9 TxCMD = Enable

This command enables the normal operation condition of the transmitter. When enabled, the transmitter will implement the following rules.

1) The transmission of SKIP ordered-sets, link heartbeats, packets, and link idle data **shall** be restricted to the configured link width.
2) Physical lanes not included in the configured link width (unused) **shall** be forced to a quiescent condition.
3) SKIP ordered-sets have the highest transmit priority. When scheduled, a SKIP ordered-set **shall** be transmitted as soon as possible at the next packet boundary, at the next ordered-set boundary, or by interrupting link idle transmission.
4) On ports that implement Rel. 1.2 Enhanced Signaling capability, Link Heartbeat Ordered-Sets have the second highest transmit priority. When a Link Heartbeat ordered-set is available for transmission, it **shall** be transmitted as soon as possible at the next packet boundary, at the next ordered-set boundary, or by interrupting link idle transmission.
5) Packets have the next highest transmit priority. When a packet is available for transmission, it **shall** be transmitted as soon as possible at the next packet boundary, at the next ordered-set boundary, or by interrupting link idle transmission.
6) When there are no SKIP ordered-sets, link heartbeats, or packets to transmit, the transmitter **shall** transmit the link idle data pattern.

5.6.7 Receiver Interface and Behavior

The link training state machine controls the receiver behavior using a set of receiver commands (RxCMD). The receiver reports its status as RxStatus. The expected behavior of the receiver in response to each command is described in the following paragraphs. Receiver status conditions are defined as part of receiver command definition.

C5-12: All ports **shall** implement the Receiver behaviors defined in Section 5.6.7, “Receiver Interface and Behavior,” on page 151 but are not required to implement the optional lane reversal function or correction of inverted serial data.

o5-4: All ports that implement the optional correction of inverted serial data **shall** implement the behavior as defined in Section 5.6.7.3, “RxCMD = EnConfig,” on page 152 rule 2.

o5-5: All ports that implement the optional lane reversal function **shall** implement the behavior as defined in Section 5.6.7.3, “RxCMD = EnConfig,” on page 152 rule 4.

5.6.7.1 RxCMD = Disable

When the receiver is disabled, the receiver **shall** implement the following rules:

1) The receiver **shall not** forward packets to the upper layer protocols.
2) The receiver **shall not** update the port error counters. (See Section 5.4.3 on page 110)

5.6.7.2 RxCMD = WaitTS1

This command instructs the receiver to look for TS1 on all physical lanes independently. When commanded to WaitTS1, the receiver **shall** implement the following rules.

1) Symbol synchronization **shall** be enabled on all lanes.
2) The receiver status **shall** be RcvdTS1 when at least one complete and error free TS1 or TS2 ordered-set is detected in one or more physical lanes.
3) The training sequence ordered-set lane number **shall not** be checked.

5.6.7.3 RxCMD = EnConfig

When the receiver is commanded to enable auto-configure, the receiver will attempt to configure the port as restricted by port capability and by capabilities enabled by management commands (See Section 5.4.1 on page 105). When commanded to enable auto-configure, the receiver will implement the following rules:

1) The receiver **shall** only attempt to configure the link to speeds enabled by the LinkSpeedEnabled variable. (Auto-configuration of link speed is supported under Rel. 1.2 Enhanced Signaling, but not supported by legacy devices).
2) The receiver **may** optionally correct inverted receiver data.
3) The receiver **shall** verify proper lane polarity using the training sequence data (symbols 3 through 16 of both TS1 and TS2) in the received training sequence. If the lane polarity is not correct (or cannot be corrected), the receiver **shall not** report RxTrained.
4) The receiver **may** optionally correct reversed lanes.
5) The receiver **shall** verify proper lane order using the lane number symbol (the second symbol in both TS1 and TS2) in the received training sequence. If proper lane ordering is not present (or cannot be corrected), the port **shall not** report RxTrained.
6) The receiver **shall** only attempt to configure the link to widths enabled by the LinkWidthEnabled variable.

a) When the 12x width is enabled, the receiver **shall** attempt symbol synchronization on twelve physical lanes (11-0) and **shall** verify that all twelve lanes are receiving TS1 or TS2 ordered-sets.

b) When the 8x width is enabled, the receiver **shall** attempt symbol synchronization on eight physical lanes (7-0) and **shall** verify that all eight lanes are receiving TS1 or TS2 ordered-sets.

c) When the 4x width is enabled, the receiver **shall** attempt symbol synchronization on four physical lanes (3-0) and **shall** verify that all four lanes are receiving TS1 or TS2 ordered-sets.

d) When the 1x width is enabled, the receiver **shall** attempt symbol synchronization on physical lane (0) and **shall** verify that it is receiving TS1 or TS2 ordered-sets.

7) The receiver **shall** use widest enabled and verified link width for completion of link training and report that width as the configured link width.

8) The receiver **shall** use the sixteen symbol long TS1 or TS2 as reference for link de-skew operations.

9) The receiver **shall** be capable of de-skewing a minimum of six symbol times of total link skew.

10) After successful link de-skew, the receiver **shall** receive eight (8) consecutive error free TS1 or TS2 (all TS1, all TS2 or TS1 followed by TS2) ordered-sets simultaneously on all configured lanes before reporting RxTrained status.

5.6.7.4 RxCMD = EnDeSkew

When the receiver is commanded to enable de-skew, it will implement the following rules:

1) The receiver **shall** use the sixteen symbol long TS1 or TS2 as reference for link de-skew operations.

2) The receiver **shall** be capable of de-skewing a minimum of six symbol times of total link skew.

3) After successful link de-skew, the receiver **shall** receive eight (8) consecutive error free TS1 or TS2 (all TS1, all TS2 or TS1 followed by TS2) ordered-sets simultaneously on all configured lanes before reporting RxTrained status.

5.6.7.5 RxCMD = WaitTS2

This command instructs the receiver to confirm configuration of the remote port. When the remote port configures its receiver, it starts to transmit TS2 ordered-sets. When commanded to wait for TS2, the receiver **shall** implement the following rules:

1) The receiver **shall** ignore data on the lanes that are not included in the configured link width (unused).
2) The receiver **shall** ignore properly formatted TS1 and SKIP ordered-sets.
3) The receiver **shall** receive eight (8) consecutive error free and properly lane ordered TS2 or TS3 ordered-sets simultaneously on all configured lanes before reporting RcvdTS2 status.

5.6.7.6 RxCMD = WaitTS3

The WaitTS3 command instructs the receiver to determine the advertised speed of the remote port. When the far port has finished sending and receiving idle data it starts to transmit TS3 ordered-sets. When commanded to wait for TS3, the receiver shall implement the following rules:

1) The receiver **shall** ignore data on lanes that are not included in the configured link width (unused).
2) The receiver **shall** receive eight (8) consecutive error free TS3 ordered sets simultaneously on all configured lanes before reporting RcvdTS3 status.
3) The receiver **shall** ignore error free idle data and SKIP ordered-sets
4) The Receiver **shall** compare the link speed indicator from the received TS3 ordered-sets with the link speed indicator from the links transmitted TS3 ordered-sets. The highest common speed will be saved and passed to the transmitter for use in the transmission of further data.

5.6.7.7 RxCMD = WaitIdle

This command instructs the receiver to confirm reception of the final handshake necessary to complete the link training process. When the far port confirms proper configuration of the near port, it starts to transmit the link idle data pattern. When commanded to wait for idle data, the receiver **shall** implement the following rules:

1) The receiver **shall** ignore properly formatted TS2 and SKIP ordered-sets.
2) The receiver **shall** assert RcvdIdle after reception of at least eight (8) error free symbol times of link idle data sequence on all configured lanes (i.e. eight (8) per lane).

5.6.7.8 RxCMD = Enable

This command enables the normal receiver operation. The receiver will implement the following rules:

1) Received packets **shall** be transferred to the upper layers of the protocol (link layer).

2) Link error detection **shall** be enabled. (See Section 5.7 on page 156)
3) If Rel. 1.2 Enhanced Signaling capabilities are implemented,
   a) detection of Heartbeat SND Ordered Sets **shall** be enabled, and
   b) detection of Heartbeat Errors **shall** be enabled.

5.6.7.9 RxCMD = WaitTS-T

This command enables the receiver to detect the TS-T ordered set. The receiver will implement the following rules:

1) The receiver status **shall** be RcvdTS-T when at least one complete and error free TS-T ordered-set is detected in one or more physical lanes.

5.6.7.10 RxStatus = RcvdTS1

The received TS1 status is valid at any time TS1 is received. The receiver status RcvdTS1 is defined in Section 5.6.7.2, "RxCMD = WaitTS1," on page 152.

5.6.7.11 RxStatus = RxTrained

The receiver trained status is valid only when the RxCMD is EnConfig or EnDeSkew. The receiver status RxTrained is defined in Section 5.6.7.4, "RxCMD = EnDeSkew," on page 153.

5.6.7.12 RxStatus = RcvdTS2

The received TS2 status is valid at any time TS2 is received. The receiver status RcvdTS2 is defined in Section 5.6.7.5, "RxCMD = WaitTS2," on page 153.

5.6.7.13 RxStatus = RcvdTS3

The received TS3 status is valid only when the RxCMD is WaitTS3. The receiver status RcvdTS3 is defined in Section 5.6.7.6, "RxCMD = WaitTS3," on page 154.

5.6.7.14 RxStatus = RcvdTS-T

The received TS-T status is valid only when the RxCMD is WaitTS-T. The receiver status RcvdTS-T is defined in Section 5.6.7.9, "RxCMD = WaitTS-T," on page 155.

5.6.7.15 RxStatus = RcvdIdle

The received idle status is valid when the RxCMD is WaitIdle and in states where a received Idle causes a state transition. The receiver status RcvdIdle is defined in Section 5.6.7.7, "RxCMD = WaitIdle," on page 154.

5.6.7.16 RxStatus = RxMajorError

Receiver Major Error status can be asserted by multiple sources. The receiver status RxMajorError is defined in Section 5.7.4, "Major Link Physical Errors Events," on page 158.

5.6.7.17 RxStatus = RxHeartbeatError

Receiver Heartbeat Error status can be asserted by multiple sources. The receiver status RxMajorError is defined in Section 5.11.2, "Heartbeat Error Handling," on page 168.

5.7 Link Physical Error Handling

This section describes link error detection and link error recovery implemented in the Link/Physical layer.

The Link/Physical layer does not interpret packet payloads, packet framing, or CRCs. These errors are detected at the Link Layer Protocol described in Volume 1.

C5-13: All ports **shall** implement link physical error handling as defined in Section 5.7, "Link Physical Error Handling," on page 156.

5.7.1 Link Physical Errors Events

Link/Physical errors stem from two fundamental sources: link bit errors and protocol violations. Bit errors may appear as 8B/10B coding violations, running disparity violations, or incorrect but valid 8B/10B code groups. Bit errors which result in incorrect but valid code groups may be detected as protocol errors or as CRC errors when checked by the upper layers of the protocol. Burst errors may severely corrupt multiple code groups on one or more lanes. These major error events may result in multiple coding violations, protocol errors, loss of lane to lane de-skew, or loss of symbol synchronization. Protocol errors may be the result of simple bit errors, or they may be the result of some other event. Receiver-detected errors are handled in several basic ways:

1) Minor error events which do not significantly impact link/physical layer processing **shall** be marked and forwarded to the upper layers.
   - When minor error events occur simultaneously on multiple lanes, they **shall** be treated as a single minor error event.
2) Major error events **shall** result in a link error event which triggers link/physical error recovery.

The error threshold logic described in Section 5.7.3 on page 157 monitors minor error events and may trigger a major event if the rate is too high.

5.7.2 MINOR LINK PHYSICAL ERRORS EVENTS

Minor error events are input to the error threshold logic described in Section 5.7.3 on page 157 and to the ***SymbolErrorCounter*** described in Section 5.4.2. The following errors are counted as minor error events:

1) Invalid 8B/10B codes groups and running disparity errors **shall** be counted as minor error events. (See Section 5.2).

2) Unsupported or disabled valid code groups **shall** be counted as minor error events. (See Section 5.3.1.8 and Section 5.3.1.9).

3) Start or End Packet Delimiter in the wrong lane of a multi-lane link **shall** be counted as a minor error event. (See Section 5.5.4, Section 5.5.5, and Section 5.5.6).

4) Any control symbol within the boundaries of a packet **shall** be counted as a minor error event.

5) PAD symbols on a non-12x and non-8x link **shall** be counted as minor error events. (See Section 5.5.5 and Section 5.5.6).

6) On an 8x or 12x link, PAD symbols not preceded by a End of Packet Delimiter **shall** be counted as minor error events. PAD symbols in the wrong lane **shall** also be counted as minor error events. (See Section 5.5.5, Section 5.5.6 and Section 5.10.3)

.

5.7.3 LINK PHYSICAL ERROR THRESHOLD ALGORITHM

To detect an excessive number of minor errors (see Section 5.7.2 above), an error threshold algorithm is implemented. The threshold **shall** be set to detect an error rate of four (4) or more minor errors within sixteen (16) symbol times. Both the “leaky bucket algorithm” and “sliding window algorithm” implementations are sufficient. The error threshold function will implement the following rules:

1) The error threshold **shall** be enabled only when the link training state machine is in the LinkUp state.

2) The error threshold **shall** be disabled and cleared when the link training state machine is not in the LinkUp state.

The following implementation notes describe leaky bucket and sliding window error threshold algorithms.

Implementation Note:

The error threshold counter which employs the “leaky bucket algorithm” **should** implement with the following rules:

1) When a minor error is detected, the error threshold counter is incremented.
2) The error threshold counter is decremented every sixteen (16) symbol times.
3) The decrement event may be a free-running symbol time counter that is not synchronized to any specific link state or error condition.

An error threshold event will be reported when the error threshold counter is value is four (4) or greater.

Implementation Note:

The error threshold counter which employs the “sliding window algorithm” **should** implement with the following rules:

1) Logic tracks the minor error status of the most recent 16 symbol times.
2) An error threshold event will be reported when four (4) or more minor errors are present in the in the most recent 16 symbol time of history

5.7.4 MAJOR LINK PHYSICAL ERRORS EVENTS

Major error events trigger the link error recovery process of the Link Training State Machine. These error events are not counted directly. However, successful error recovery attempts are counted by the ***LinkErrorRecoveryCounter*** and failed error recovery attempts are counted by the ***LinkDownedCounter***. Both counters are defined in Section 5.4.3, “Port Performance Counters,” on page 110. The following errors events will start the link error recovery process:

1) Training Sequence one or two (TS1 or TS2) received in the linkup state **shall** trigger link error recovery. (TS1 indicates that the remote port has initiated the link error recovery process)
2) The assertion of LinkPhyRecover from the Link Layer **shall** trigger link error recovery.

3) Loss of symbol(s) caused by elastic buffer overflow or underflow **shall** trigger link error recovery.
4) The absence of 4 expected SKIP ordered-sets **should** trigger link error recovery.
5) Clear loss of lane-to-lane de-skew **should** trigger link error recovery.
6) A minor error threshold event **should** trigger link error recovery.

5.7.5 HEARTBEAT ERROR

A Heartbeat Error error event triggers a transition to the default state in the Link Training State Machine. These error events are counted by the ***LinkDownedCounter***. The following error event is a Heartbeat Error event:

- 4 HEARTBEAT transmission periods (8 ms) elapse without receiving valid HEARTBEAT ACK ordered-sets simultaneously on all active lanes.

The characteristics of a valid HEARTBEAT ordered-set are described in Section 5.3.2.5, “Link Heartbeat Ordered-Set (HRTBT),” on page 101. A valid Link Heartbeat ordered-set has the following characteristics:

- Error-free reception of 8b/10b symbols on all configured lanes.
- Lane IDs match the lanes on which they were received.
- Opcode of either 5Dh (SND) or ACh (ACK).
- Received GUID on a SND HRTBT that doesn’t match the receive port’s own GUID. Matching SND GUID is interpreted as transmitter/receiver crosstalk on SND transmission from a port transmitter on the same device.
- Received GUID and PortNum on an ACK HRTBT that matches the receive port’s own GUID and PortNum. Mismatched ACK GUID and Portnum is interpreted as transmitter/receiver crosstalk on ACK transmission.

5.8 INTERNAL SERIAL LOOPBACK

The Internal Serial Loopback is optional and is not required on all ports. Internal serial loopback allows a port to receive its transmitted data stream without the need for external support. The internal serial loopback function is intended for self-test and fault isolation use only.

o5-6: All ports that implement internal serial loopback option **shall** implement it as defined in Section 5.8, “Internal Serial Loopback,” on page 159.

The following rules apply to the internal serial loopback:

1) When disabled, the internal serial loopback **shall** have no effect on port operation.

2) When enabled, internal serial loopback **shall** connect the transmit serial data output to the receive serial data input as close to the chip I/O cells as is practical.

3) When internal serial loopback is enabled, the receive data from the receiver input pins **shall** be blocked.

4) When internal serial loopback is enabled, the chip's transmitter output pins **shall** be forced to a quiescent condition.

5.9 CLOCK TOLERANCE COMPENSATION

Each end of a 1x link (or each physical lane on 4x, 8x and 12x links) utilizes a transmitter and a receiver. The transmitter logic operates using a tightly controlled reference clock called the "transmit clock". Likewise, the receiver logic operates using a clock recovered from the incoming bit stream called the "receive clock". Once the link is trained, the recovered receive clock operates at the same frequency as the transmit clock at the other end of the link.

The UI_D parameter in *InfiniBand Architecture Specification, Volume 2,* Table 17 Driver Characteristics for 2.5 Gb/s on page 185 specifies the transmit clock accuracy as +/- 100 ppm (parts per million). The worst-case frequency difference between the transmit and receive clocks of a link occurs when one of the transmitters is at the +100 ppm and the other one is at the -100 ppm tolerance, resulting in a 200 ppm difference. In other words, the transmit and receive clocks can shift by as much as one clock period every 5000 clocks.

A common design practice is to clock most of the receive path logic in the transmit clock domain. This is accomplished by using an "elastic buffer" in the very early stages of the receive path. The elastic buffer compensates for the differences between the transmit and receive clock domains by dropping or inserting symbols. The input side of the elastic buffer operates in the receive clock domain, and the output side operates in the transmit clock domain.

To perform the compensation function, the elastic buffer needs to identify safe zones in the incoming symbol sequence to insert or drop a symbol. This zone is provided by the "SKIP ordered-set".

C5-14: All ports **shall** implement clock tolerance compensation as defined in Section 5.9, "Clock Tolerance Compensation," on page 160.

5.9.1 TRANSMITTER "SKIP" REQUIREMENTS

The transmitters are required to transmit SKIP ordered-sets periodically, complying with the following rules:

1) The SKIP ordered-set **shall** be scheduled for insertion at least once every 4608 symbol times.

2) The SKIP ordered-set **shall** be scheduled for insertion at most once every 4352 symbol times.

3) A scheduled SKIP ordered-set **shall** be inserted at the next packet or ordered-set boundary.

5.9.2 RECEIVER "SKIP" REQUIREMENTS

The receivers are required to receive and process SKIP ordered-sets periodically, complying with the following rules:

1) The receivers **shall** recognize received SKIP ordered-sets that are comprised of one (1) comma (COM) symbol followed by one (1) to five (5) skip (SKP) symbols.

2) The receivers **shall** be tolerant to receive and process SKIP ordered-sets at an average rate of once in every 4352 to 4608 symbol times.

3) The receivers **shall** be tolerant to receive and process SKIP ordered-sets separated from each other at least 128 symbol times -- measured as the distance between the leading comma (COM) symbols.

4) The receivers **shall** be tolerant to receive and process SKIP ordered-sets separated from each other at most 8832 symbol times -- measured as the distance between the leading comma (COM) symbols.

5.10 RETIMING REPEATERS

The InfiniBand™ Architecture allows for the use of "retiming repeaters" to recover from potentially weakened signal strength and built-up jitter between two end nodes of a link.

There are two types of retiming repeater: "SKIP ordered-set dependent" retiming repeaters, which use the SKIP ordered-set to compensate for frequency difference, and "transparent" retiming repeaters, which do not depend on or use the SKIP ordered-set, but operate at a single common frequency for both transmission and reception.

This section describes, in general, the operation of SKIP ordered-set dependent retiming repeaters, except where transparent retiming repeaters are specifically mentioned. Both types of retiming repeater reset jitter and meet the signaling requirement specified in (*InfiniBand Architecture Specification, Volume 2,* Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s).

Implementation Note:

The major difference between the two types of retiming repeaters is in whether the retiming repeater uses an internal clock, or uses a clock recovered from the received data stream, for transmitting the repeated data. A SKIP ordered-set dependent retiming repeater will typically have its own internal clock, with the usual +/- 100 ppm frequency tolerance, and may need to insert or remove SKIP symbols to compensate for frequency differences with the originating port. A transparent retiming repeater will typically transmit the repeated data using a clock recovered from the received data.

Transparent retiming repeaters will typically be more generally useful, since they are not dependent on the specific InfiniBand link/phy protocol's use of the SKIP ordered-set.

C5-15: All retiming repeaters **shall** implement functions as defined in 5.10: Retiming Repeaters.

The following general rules apply to retiming repeaters:

1) Retiming repeaters **shall** reset the jitter budget.
2) Not more than two SKIP ordered set-dependent retiming repeaters **shall** be allowed between two protocol-aware ports.
3) Retiming repeaters **may** join dissimilar physical media such as copper-to-fiber optic links.
4) Retiming repeaters are not in-band-addressable devices. Hence they cannot be managed through in-band management messages.

IBTA

5.10.1 RETIMING REPEATER FUNCTIONS

Figure 29 below depicts a block diagram of a conceptual retiming repeater. This figure is provided as a visual aid to help explain the fundamental functions of retiming repeaters.

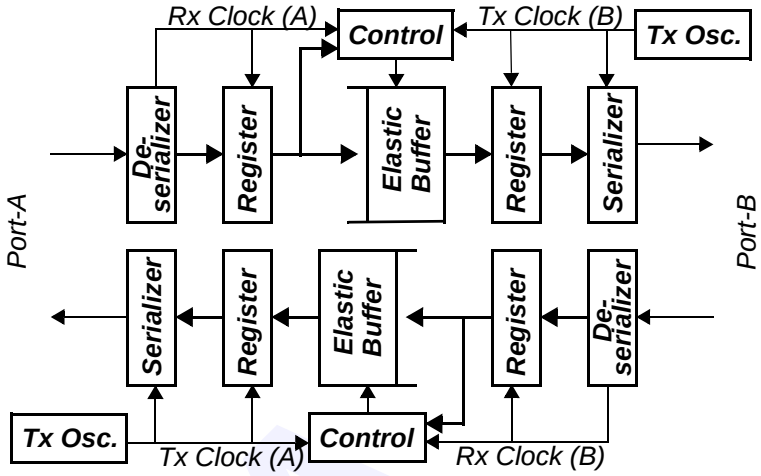


Figure 29 A Conceptual Retiming Repeater Block Diagram

Retiming repeaters are fairly simple devices. A 1x retiming repeater consists of two ports. It simply transmits every symbol received on one port to the other port. The receiver circuitry on each port operates from its own recovered receive clock (Rx Clock) domain. However, the transmit circuitry operates on a locally generated transmit clock (Tx Clock) domain, which may have a different frequency than the frequency of the incoming data. In order to compensate for the differences in the receive and transmit clock domains, elastic buffers are used in each direction. Transparent retiming repeaters transmit the data using the clock recovered from the received data, and have less requirement for elastic buffers.

Retiming repeaters are not link-protocol-aware devices. In other words, they do not recognize link and data packets. However, SKIP ordered-set dependent retiming repeaters do recognize two kinds of ordered-sets in the incoming symbol stream:

- SKIP ordered-set
- TS1 and TS2 Training Sequence ordered-sets

SKIP ordered-sets are both recognized in the incoming symbol stream and used in the elasticity operations performed in each direction. Elasticity operations performed by the retiming repeaters are specified in Section 5.10.2 later in this chapter.

During Link Training, repeated training sequence ordered-sets are transmitted between two end points. Repeaters use the periodic comma (COM) symbols in this stream to detect symbol boundary misalignment.

5.10.2 CLOCK TOLERANCE COMPENSATION

SKIP ordered-set dependent retiming repeaters perform clock tolerance compensation during SKIP sequences. The SKIP sequence -- as observed by repeaters -- is comprised of one comma (COM) symbol followed by two to four skip (SKP) symbols. Repeaters are allowed to insert or delete one skip symbol in this ordered-set in order to compensate for the differences between the receive and transmit clock frequencies.

SKIP ordered-set dependent retiming repeaters **shall** use the following rules to insert or delete a skip symbol in a SKIP ordered-set:

1) When necessary, the retiming repeaters **shall** insert a skip (SKP) symbol after the comma (COM) symbol within the current SKIP ordered-set.
2) When necessary, the retiming repeaters **shall** delete any skip (SKP) symbol after the comma (COM) symbol in the current SKIP ordered-set.

5.10.3 ERROR HANDLING CONSIDERATIONS

Retiming repeaters **shall not** check for code violations (i.e. decode and disparity errors). Symbols received on one port with or without transmission errors are simply transmitted by the other port.

The elastic buffers are not expected to underflow or overflow. However, when these conditions are detected, these rules apply:

1) When an underflow condition is detected, the retiming repeaters **shall** insert a pad (PAD) symbol.
2) When an overflow condition is detected, the retiming repeaters **shall** replace two arbitrary but consecutive symbols with a pad (PAD) symbol.

5.10.4 SYMBOL BOUNDARY ALIGNMENT

When protocol-aware nodes detect link errors (including loss-of-symbol synchronization), they initiate error recovery (see Section 5.7 on page 156) by sending training sequences that contain the commas needed for symbol synchronization. Retiming repeaters cannot initiate training sequences on their own. Instead, they detect periodic unaligned comma (COM) symbols within this sequence to determine the loss-of-symbol boundary alignment.

The following rules define the symbol boundary alignment process used by retiming repeaters:

1) Periodic comma symbols (COM) in TS1, TS2 or SKIP ordered-sets **shall** be used to acquire symbol boundary synchronization.
2) SKIP ordered-set dependent retiming repeaters **shall** reacquire symbol boundary synchronization when three consecutive misaligned comma symbols are detected. A misaligned comma symbol is the 0011111 or 1100000 symbol pattern which does not begin at the current symbol boundary.
   - SKIP ordered-set dependent retiming repeaters **may** use the 7-bit comma bit pattern 1100000 or 0011111 to acquire alignment, rather than the full 10-bit comma symbol pattern.
3) Retiming repeaters **shall** disable their symbol boundary synchronization circuitry when five consecutive instances of aligned comma symbols (COM) are detected.

5.10.5 MULTI-LANE REPEATER CONSIDERATIONS

Multi-lane retiming repeaters are used by multi-lane (4x, 8x and 12x) links. Like protocol aware ports, retiming repeaters are required to use a common clock source for all lanes in the link. The logic within a retiming repeater is not required to synchronize lane to lane operation. Multi-lane retiming repeaters can be shared by multiple links that use a subset of the retiming repeater's lanes. For example, a 4x retiming repeater can be used by up to 4 1x links. The 12x retiming repeater can be used by multiple combinations of 1x, 4x, and 8x links.

The following rules apply to operation of multi-lane retiming repeaters:

1) For each direction all lanes of multi-lane retiming repeaters **shall** have a common transmit clock source.
2) The lanes in opposed directions **may** have independent transmit clock sources.
3) Each lane of multi-lane retiming repeaters **shall** operate independently of the other lanes.
4) Multi-lane retiming repeaters **may** implement lane to lane de-skew.

5.10.6 POWER STATE CONSIDERATIONS

The retiming repeaters are not protocol-aware devices. Hence they cannot be managed through in-band management packets. However, they are expected to be managed by either:

1) the Management Entity of the chassis of which they are part, or
2) the Management Entity of the InfiniBand module of which they are part.

The retiming repeaters that are managed using one of these schemes are also said to be "proxy-managed."

During normal operation, a neighboring node may transition to $\mathbf{X_{Standby}}$ or $\mathbf{X_{Sleep}}$ state. In these states, the node drives the link to "quiescent" state. Similarly, when the neighboring node transitions to the $\mathbf{X_{Polling}}$ state, the link will be driven to its "quiescent" state periodically. Retiming repeaters are expected to detect the link going to "quiescent" state at its receiver on either side and to propagate this link "quiescent" state on the transmitter of the other side.

When the neighboring node transitions to $\mathbf{X_{On}}$ state, the link goes from "quiescent" to "active" state. Similarly, when the neighboring node transitions to the $\mathbf{X_{Polling}}$ state, the link will transition from "quiescent" to "active" state periodically. The retiming repeaters are expected to detect this link state change on its receiver at either side and to propagate it to the transmitter on the other side.

Proxy managed retiming repeaters **shall** comply with power states and behavior defined in *InfiniBand Architecture Specification, Volume 2,* Section 14.3, "Port Power Management States," on page 641. Additionally, operation of the retiming repeaters when propagating "quiescent" and "active" link activity are governed by the following rules:

1) Retiming repeaters **shall** detect "quiescent" to "active" transition on one side and propagate that state to the other side within 100 microseconds.
2) Retiming repeaters **shall** detect "active" to "quiescent" transition on one side and propagate that state to the other side within 100 microseconds.

5.11 LINK HEARTBEAT

At the DDR and QDR link bit rates described for Rel. 1.2 Enhanced Signaling, the lower peak-to-peak received voltage and the consequent lack of a signal detect at the receiver results in higher vulnerability to crosstalk in the receiver logic. It is possible that if a link is disconnected while a port is in the LinkUp state a receiver may interpret crosstalk from its own transmitter as valid data and incorrectly maintain LinkUp status. The link heartbeat function ensures that received data actually originates from the opposite end of a link, rather than from crosstalk from the port's own transmitter.

Link Heartbeats are also used to determine link round-trip latency. Since InfiniBand™ links use a variety of copper and optical physical layer transmission media, a link may be between a few inches and many kilometers long, causing a wide range of link latency values. Prior releases of the InfiniBand™ specification allowed no straightforward way for the link layer logic or for a subnet manager to determine the round-trip latency across a link. Link Heartbeats allow the determination of link round-trip latency,

for helping to improve link management and bandwidth allocation algorithms, which is useful for SDR as well as for DDR & QDR link rates.

Implementation Note:

Knowledge of link lengths can be useful for link management functions, e.g., to tell whether optical transceivers or retiming repeaters are likely to be present. A 1 km link, for example, may be assumed to have an optical transceiver inserted, even though, since they are simple devices, and are not in-band-addressable, this information can't be determined using in-band management messages.

Knowledge of link lengths is also useful for congestion control and bandwidth maximization mechanisms, since a long link with high link latency may be throttled in bandwidth due to credit starvation if it is not allocated extra buffering and extra credits. A switch chip, for example, with a combination of long and short links on different ports and with a flexible buffering allocation capability, can move buffer space and credit allocation from a short link to a long link, maximizing usable bandwidth on both.

C5-15.2.1: All ports claiming compliance with InfiniBand Rel. 1.2 Enhanced Signaling **shall** implement link heartbeat functionality as defined by Section 5.11, "Link Heartbeat," on page 166, at all link speeds.

5.11.1 OPERATION OF LINK HEARTBEATS

Each port capable of Rel. 1.2 Enhanced Signaling operation will create regularly transmitted HEARTBEAT ordered-sets when connected to another capable port, at any link speed. A HRTBT ordered-set, described in Section 5.3.2.5, "Link Heartbeat Ordered-Set (HRTBT)," on page 101 contains an OpCode identifying it as a SND HEARTBEAT or an ACK HEARTBEAT, along with information identifying its source.

A port schedules for transmission a Link Heartbeat with OpCode=5Dh, indicating SND, once every 100 milliseconds while the port is in the LinkUp state. The SND HEARTBEAT contains the sending port's base GUID and, if applicable, the sending port's switch port number. In the SND HEARTBEAT, the OpCode byte, is set to the value D29.2 (5Dh, encoded as 101110 0101 or 010001 0101).

Upon reception of a valid SND HEARTBEAT the port schedules an ACK (Acknowledge) HEARTBEAT for transmission at the next possible transmission time. The ACK heartbeat is sent with the GUID and switch port

number contained in the SND heartbeat, i.e., these values are reflected back in the ACK HEARTBEAT ordered-set. In the ACK HEARTBEAT, the OpCode byte, the sixth byte in the ordered-set, is set to the value D12.5 (ACh, encoded as 001101 1010 or 001101 1010).

Upon reception of a valid ACK (acknowledge) HEARTBEAT, the port finds the difference in time between its original SND transmission and the ACK reception. This difference correlates to the link round-trip latency, which correlates to link length.

Both SND HEARTBEAT and ACK HEARTBEAT ordered-sets are transmitted on every active lane simultaneously (like SKP ordered-sets).

SND HEARTBEAT ordered-sets are scheduled once every 100 ms in LinkUp State, to match the minimum DelayTimeOut period. Transmitting this often incurs negligible bandwidth loss and provides prompt notification of changes in link integrity.

Transmission of Link Heartbeat ordered-sets interspersed with other packets and ordered-sets uses the priority scheme described in Section 5.6.6.9, “TxCMD = Enable,” on page 151.

5.11.2 HEARTBEAT ERROR HANDLING

Heartbeat errors are detected under the conditions described in Section 5.7.5, “Heartbeat Error,” on page 159.

Heartbeat errors indicate that a link has been disconnected, and should result in the Link Training State Machine returning to LinkDownDefaultState (Polling or Sleeping) with the ***LinkDownedCounter*** being incremented.

5.11.3 HEARTBEAT LATENCY CALCULATION AND REPORTING

Round-trip latency is measured as the time between the end of Heartbeat SND transmission and the end of reception of the corresponding Heartbeat ACK. The longest possible round-trip latency, over a 10 km. single-mode optical link, with roughly 5 ns/meter propagation speed, will be roughly 100 microseconds, and the shortest possible round-trip latency will be less than 100 nanoseconds.

Implementation Note:

Since the fundamental granularity is limited by clock cycle and HEARTBEAT ordered-set lengths to at least 4 nanoseconds, and the maximum round-trip link latency is roughly 100,000 nanoseconds across a 10 km link, a 2-byte counter (e.g., counting to 64K cycles of 4ns. each) should be sufficient. A 3-byte counter, with 4 ns. granularity, would allow accurate round-trip latency measurement of a link roughly 6,700 km. long.

Use of the Heartbeat mechanism allows for a PortInfo value, ***SM.PortInfo(LinkRoundTripLatency)***, which is accessible through the subnet management MAD Get mechanism, as described in Section 5.4.2, "Status Outputs (MAD get)," on page 107. This is a 32 bit value representing the round-trip latency of the link, measured in 4 nanosecond intervals. This value is reset to 0xFFFF_FFFF upon entry to LinkDownDefaultState. When a SND heartbeat is transmitted, the current time (most likely via a free-running counter) is latched. Upon reception of an ACK heartbeat, the current time is compared against the latched time, and the difference, converted to nanoseconds, is reported as the LinkRoundTripLatency.

Due to delays in scheduling, an ACK may be sent significantly later (by as much as roughly 8 microsec.) than the time that the SND was received, resulting in an erroneously high link latency. Therefore, only the lowest value seen across a set of link latency measurements will indicate the most representative value, and only the lowest measured value should be reported as ***SM.PortInfo(LinkRoundTripLatency)***.

The precision of reporting is dependent on how well the link round trip latency has been measured, but is expected to be measured to precision of better than 100 nanoseconds. When a port is brought to the Sleeping, Polling, or Disabled State, the ***LinkRoundTripLatency*** is re-set to 0xFFFF FFFF.

5.12 PHYSICAL LAYER COMPLIANCE TESTING

The following facilities and procedures are intended to simplify physical layer device characterization and compliance testing to specifications described in Section 6:, "High Speed Electrical Signalling - 2.5Gb/s, 5.0 Gb/s, & 10 Gb/s," on page 139, while minimizing the amount of circuity and complexity required in the InfiniBand devices.

C5-15.2.2: All ports claiming compliance with InfiniBand Rel. 1.2 Enhanced Signaling **shall** implement physical layer compliance testing as

defined by Section 5.12, “Physical Layer Compliance Testing,” on page 169, at all link speeds.

5.12.1 COMPLIANCE TESTING OVERVIEW

Device testing is accomplished by attaching the device to test equipment over a link, with the receiver attached to a signal generator such as an arbitrary pattern generator, and the transmitter attached to a signal monitor such as an oscilloscope. Testing at the physical layer does not require intervention by a Subnet Manager. New facilities are included in the port to simplify compliance testing for the high-speed devices, including modification to the Link Training State Machine from the structure shown in Figure 21 to the structure shown in Figure 22, and definition of an ordered-set, Training Sequence for Test (TS-T), which is used exclusively for compliance testing. Multiple testing modes are possible, as determined by an Opcode in the TS-T ordered-set.

When the port is placed into the Phy Test state by reception of a TS-T ordered-set from attached test equipment on any lane, it disables all link-level protocol other than 8b/10b code/decode and TS1 detection, and only enables the physical layer circuitry, including clock and data recovery, link transmitter, driver de-emphasis, and receiver equalization circuitry.

Transmitter testing is accomplished using the TS-T to command the transmitter to transmit particular test patterns, and monitoring the characteristics (jitter, amplitude, rise/fall times, etc.) of the transmitted patterns.

For receiver testing, the test equipment uses the TS-T to command the transmitter to send particular symbols depending on the validity (signal strength, 8b/10b decode correctness as to bit errors, disparity, etc.) of the received data, and then feeding the receiver 8b/10b coded data with various characteristics (signal strength, deterministic jitter, pattern-dependent jitter, etc.) and monitoring the symbols the transmitter sends.

Since there is no link-level logic enabled during this procedure (aside from optional vendor-dependent capabilities which are beyond the scope of this specification), no link-level functions (flow control, CRC, packet framing, etc.) are tested with these procedures. These facilities are intended to test compliance with transmitter and receiver specifications described in Section 6:, “High Speed Electrical Signalling - 2.5Gb/s, 5.0 Gb/s, & 10 Gb/s,” on page 139.

5.12.2 COMPLIANCE TESTING FACILITIES

Facilities for physical layer compliance testing include changes to the Link Training State Machine (LTSM), addition of a new Training Sequence for Test (TS-T) ordered-set, and capability for the link to generate patterns specifically used for compliance testing.

5.12.2.1 LINK TRAINING STATE MACHINE MODIFICATIONS FOR COMPLIANCE TESTING

The differences from the legacy Link Training State Machine include the following.

1) A new state is defined, “Phy Test”, which allows physical layer compliance testing.
2) A new training sequence, Training Sequence-Test (TS-T) is added, which allows transitions from the LinkDownDefault State (either Polling or Sleeping) to the Phy Test state.
3) Two new transition arcs are added, from Polling and from Sleeping, to the Phy Test state, on reception of TS-T.
4) Two new transition arcs are added, from the Phy Test state to the LinkDownDefault state (either Polling or Sleeping), on RcvdTS1. These transitions allow exit from Testing mode, without a power-off of the device. Testing procedures must ensure that no test pattern generates 8 contiguous TS1 patterns before testing is completed.
5) Within the Phy Test state, there are three (3) different required modes of operation, as described in Section 5.12.2.2, “Use of the Training Sequence for Test Ordered-Set (TS-T),” on page 171, including (a) a “SKIP-less Idles” mode, of transmitting Idle data without SKIP ordered sets for clock compensation, (b) a “SKIP-less TS1” mode, of transmitting back-to-back TS1 without SKIPs, and (c) a “Receiver Test” mode, of transmitting simple patterns indicating whether the receiver is detecting good data, or is detecting logical bit errors, or is detecting a loss of signal.
6) The Phy Test state may optionally also include other test modes, and the TS-T ordered-set allows the transmitter and receiver to optionally be placed in multiple configurations, for optional compliance testing beyond the required modes described above.

The general operation of physical layer compliance testing is described below.

5.12.2.2 USE OF THE TRAINING SEQUENCE FOR TEST ORDERED-SET (TS-T)

The Training Sequence for Test (TS-T) is only generated by test equipment. There is no need for an InfiniBand device to be able to generate this ordered-set - only the need to recognize it and behave appropriately when 8 contiguous and valid copies of TS-T arrive at the receiver port on one or more lanes.

A TS-T may be received on only a single lane, or on multiple lanes. If a multi-lane TS-T ordered-set with different values on different lanes in symbols 8, 9 and 12-15, the port may use any of the received TS-T values.

Symbol 1 of a TS-T (the Lane ID in other ordered-sets) is reserved, so that the test equipment may insert a TS-T on any lane. Symbols 2-7 contain the TS-T unique data symbol, D17.2 (or 51h), whose 10-bit encoded value is the pattern (100011 0101) for both the positive and negative running disparity. Symbols 2-7 may be used by the port to determine the polarity of symbols 8-15, since the test equipment may insert a TS-T of either polarity.

Symbol 8 of a TS-T, the link speed identifier, is used to identify the speed at which the test equipment requests the IB device to operate. At least one bit of this bit map must be asserted to 1. If multiple bits are asserted, the test will be conducted at the highest ***LinkSpeedEnabled*** speed. This allows testing at the highest enabled speed by simply asserting all bits in this symbol.

Symbol 9 of a TS-T allows the test equipment to determine which testing mode the port will be placed in. The following values are defined.

0: SKIP-less Idle Data

Each transmitter lane transmits a pseudo-random sequence of data symbols, generated by the 11th order LFSR = $X^{11} + X^{9} + 1$ with no insertion of SKIP ordered-sets.

1: SKIP-less back-to-back TS1s

Each transmitter lane transmits an unbroken string of TS1 ordered-sets, with no insertion of SKIP ordered-sets.

2: Receiver test.

On each lane, the transmitter sends an indication of the validity of the data received on the corresponding receiver lane.

VALID DATA: For each received symbol on a lane which decodes to a valid 8b/10b code point with good running disparity, the corresponding transmitter lane transmits a D10.2 (010101 0101) character.

LOGICAL ERROR: For each received symbol on a lane which decodes with a logical error (e.g., bit error, or running disparity error), the corresponding transmitter lane transmits a K28.5 D00.0 pair of symbols.

LOSS OF SIGNAL: For each received symbol on a lane which indicates an inadequate signal (e.g., inadequate signal swing, all 0s, all 1s, or noise), the corresponding transmitter lane transmits a K28.5 D01.0 pair of symbols.

3-255: Optional Vendor-dependent opcodes to allow other testing modes.

Symbols 12 and 13 allow the test equipment to optionally configure the transmitter in one of 65,536 vendor-dependent states, and symbols 14

and 15 allow the test equipment to configure the receiver in one of 65,536 vendor-dependent states. The values 0000h are identified for “normal” or “default” operation, so that test equipment can be expected to get good and correct operation when these fields are set to 0. Values other than 0 are optional and vendor-dependent. Typically only a very small subset of these states will be valid and will provide specific unique behaviors.

Implementation Note:

One particular use of this transmitter and receiver configuration facility may be for measuring transmitter equalization. As described above in Section Chapter 5:, this annex allows transmitters and receivers to negotiate for use of up to 16 different transmitter equalization setting (TES) configurations which may be adapted to the characteristics of the link. Testing of this transmitter equalization facility over various physical links will be an important function. This facility in the TS-T ordered-set can allow the test equipment to configure the transmitter to use a particular TES configuration as it’s transmitting the data, in order to see how the different TES settings operate over various media.

Similarly, the facility allows the test equipment to test out various configurations of the receiver circuitry, such as using different receiver equalization setting (RES) configurations.

5.12.3 EXAMPLE COMPLIANCE TEST PROCEDURE

Efficient and fast compliance testing requires a simple and efficient procedure to measure transmitter and receiver characteristics. The proposed procedure described below allows simple testing of the most critical elements of the transmitter and receiver circuitry. More detailed procedures will be used in practice, as this procedure is intended to illustrate use of the facilities described above.

1) The Device Under Test (DUT) is powered up, and goes into Polling state, transmitting repeated TS1(s) for 2ms in Polling.Active followed by 100ms of a quiescent output in Polling.Quiet.

2) The Test Equipment (TE) (which may be a combination of a sophisticated arbitrary pattern generator and oscilloscope, or a simpler and more application-specific box with FPGAs, simple microprocessors and SERDES circuits) is attached to the DUT by a cable.

3) The TE transmits a TS-T to the DUT, with flags set to test a particular device capability (e.g., TS-T with fields set to Speed=QDR, Opcode=0, TxCfg=0000h and RxCfg=0000h).

4) The DUT detects the TS-T, transitions to the Phy Test state, and uses the values in the TS-T fields to determine the specific test mode.

   - If the Opcode in the TS-T was set to 0, indicating “SKIP-less Idle Data,” the DUT transmits Idle data. The TE measures parameters such as Rise/Fall Time, Driver Jitter, Transmitter Peak-Peak voltage, etc., using the wide-spectrum of data contained in the Idle data pattern.

   - If the Opcode in the TS-T was set to 1, indicating “SKIP-less back-to-back TS1s”, the DUT transmits TS1 patterns, allowing the TE to make similar measurements on a much shorter data pattern with a combination of high-frequency (1010...) and low-transition-density (K28.5) elements.

   - If the Opcode in the TS-T was set to 2, indicating “Receiver Test”, the DUT transmits either D10.2, K28.5 D0.0, or K28.5 D01.0 patterns, depending on the validity of the received signal. The TE can test the sensitivity of the receiver by varying the signal given to the receiver and monitoring the transmitted pattern. Similarly, the TE can test the receiver’s sensitivity to input jitter the same way, to do a full four-corners test of receiver sensitivity to signal strength and received jitter.

5) The TE can move the device out of Phy Test state and back to the LinkDownDefaultState (Polling or Sleeping) at any time by sending a stream of 8 contiguous TS1 ordered-sets on a receiver lane.

6) Once the device is back in the Polling state, the TE can test the transmitter under different conditions by sending in a new TS-T ordered-set (e.g., with speed set to DDR instead of QDR, or with TxCfg set to 1 instead of 0).

7) At finish of testing, the TE would send the DUT a series of 8 TS1 ordered-sets, placing it back in Polling mode, and then the cable can be removed. The device is then ready to be connected to other IB devices.

There is no cross-dependency between different lanes during the Phy Test state, other from TS1 detection and the requirement that all lanes operate at the same speed. All lanes of a port move in and out of the Phy Test state together.This allows the test procedure described here to be implemented either with serial test equipment, or with more complex parallel test equipment running multiple lanes at once.

As described above, the format of the TS-T ordered-set and the function of the Phy Test state allows various other capabilities to be defined and used in a vendor-dependent manner, limited by the device and the testing equipment capabilities.

CHAPTER 6: HIGH SPEED ELECTRICAL SIGNALING - 2.5, 5.0, & 10.0 GB/S

6.1 INTRODUCTION

This section describes the signaling that allows for InfiniBand™ link operation at 2.5Gbits/s (SDR), 5.0 Gbits/s (DDR) and 10 Gbits/s (QDR). Unless otherwise noted, the specifications contained herein apply to all signals of each of the interface width definitions.

The signaling rates are for encoded data on the media, and correspond to 10 bits/Byte. The data rate can also be considered to be 250 MBytes/second/direction for SDR, 500 MB/s for DDR and 1GB/s for QDR for each of the 1, 4, 8, or 12 physical lanes.

6.1.1 BACKGROUND AND REFERENCE MATERIAL

While this specification endeavors to be complete, there is a great deal of background material which is helpful in understanding and correctly implementing this interface that is not included.

Much useful information can be found in books such as “High-Speed Digital Design, A Handbook of Black Magic” by Howard Johnson and Martin Graham and “Digital Systems Engineering” by William J. Dally and John W. Poulton. The web site http://www.t11.org has a great deal of useful information as does the site for 10Gig Ethernet.

The draft 4 Common Electrical Interface (CEI) Interoperability agreement for 6+ and 11+ Gbps I/O from the Optical Internetworking Forum (OIF) was consulted in the preparation of this chapter. At this writing it is document OIF2003.104.5 and is available to members of OIF.

Additional information is available from the IEEE (I&M) subcommittee on Pulse measurement Techniques (SCOPT). Methods of performing pulse amplitude and parametric measurements should be conducted in accordance to IEEE Std.181-2003, and this will be referenced explicitly throughout this chapter.

6.1.2 OVERVIEW AND SUMMARY OF CHANGES FOR THIS SECTION

1) Transmit and Channel parameters for DDR and QDR
2) Increased total loss and noise budget to 20 dB from 15 dB (DDR/QDR only)
3) Required Compliance and Characterization functions
4) Receive equalization and pre-emphasis for DDR/QDR

5) Transmitter provides training pattern for receive equalizer
6) Heartbeat and latency ordered set for connection verification and latency determination
7) Channel definition by isolated pulse response and S-parameters
8) Common mode changes
9) Power is provided to enable “Active Cable” connector receptacles.

Note: Some enhanced functions are defined in other sections of this annex.

6.2 SIGNAL SPECIFICATION

6.2.1 BACKGROUND

This chapter of the InfiniBand™ specification defines characteristics required to communicate between an InfiniBand output and an InfiniBand input using copper printed wiring on a printed circuit board and optional cabling at signaling rates of 2.5, 5.0, or 10 Gbits/second.

This release of the specification extends and enhances but does not supersede prior releases. Products built using this release will be interoperable with legacy products at SDR rate (2.5 Gbits/second).

Connections are point to point and signaling is unidirectional. By definition, a physical lane comprises a differential pair or fiber in each direction, i.e. a transmit signal and a receive signal at each end.

The characteristic impedance of the cables and printed wiring is nominally 100 ohms differential. The single ended impedance is more variable, depending on the amount of coupling and the package design. The details of properly designing the packaging and interconnect of an InfiniBand link are beyond the scope of this specification and are the responsibility of the designer.

The definition of the DDR and QDR signaling includes a definition of a “Compliance Channel” which represents a worst case connection from driver to receiver. SDR, by contrast, uses a loss budget approach. The compliance channel and the transmitter define the minimum acceptable inputs which the receiver shall be capable of receiving at the specified bit error rate.

A compliant channel is any channel which provides a signal at the receiver which is better than the compliance channel.

An important requirement of InfiniBand is that devices which may be produced by different manufacturers **shall** be inter-operable and hot-pluggable.

DDR devices **shall** be capable of operating in compliance with SDR. QDR devices **shall** be capable of operating in compliance with SDR and **may** optionally operate in compliance with DDR. Restrictions on configurations or the use of DC blocking capacitors may be necessary to meet this requirement with respect to prior release backplane ports. Details of the speed negotiation and configuration are specified in Section 5.6, "Link Initialization and Training," on page 118.

Pluggable devices shall meet the same requirements as other InfiniBand ports.

Note: The differential amplitude represents the value of the voltage between the true and complement signals. This may be expressed as RMS, peak, or peak-peak. The Peak-Peak value is twice the Peak value. This document uses the IEEE Std.181-2003 terminology in which the value commonly referred to as "peak-peak" is rather called "unsigned amplitude".

6.2.2 COMPLIANCE

C6-1: This compliance statement is obsolete and has been replaced by C6-1.2.1:.

C6-1.1.1: This compliance statement is obsolete and has been replaced by C6-1.2.2:.

C6-1.2.1: Any device claiming InfiniBand compliance at the slot interface, or copper cable interface **shall** comply with the requirements of Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s for Port Type 1 when operating at 2.5 Gb/s (SDR).

C6-1.2.2: Any 1x pluggable device claiming InfiniBand compliance at the socket interface **shall** comply with the requirements of Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s for Port Type 2 when operating at 2.5 Gb/s (SDR).

o6-1.2.1: Any device claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance **may** support data rates higher than SDR. A port may support a non-contiguous set of link speeds, e.g., SDR and QDR without DDR.

o6-1.2.2: Any port claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance at the slot interface, or copper cable interface **shall** comply with the requirements of Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s while operating at 5.0 Gb/s (DDR).

o6-1.2.3: Any port claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance at the slot interface, or copper cable interface **shall** comply with

BETA

the requirements of Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s while operating at 10 Gb/s (QDR).

6.3 GENERAL REQUIREMENTS

6.3.1 ESD

C6-2: This compliance statement is obsolete and has been replaced by C6-2.2.1:.

C6-2.1.1: This compliance statement is obsolete and has been replaced by C6-2.2.2:.

C6-2.2.1: InfiniBand signal and power pins **shall** withstand 2000 V of ESD using the human body model and 500 V using the charged device model without damage. Class 2 per JEDEC JESD22-A114-B.

C6-2.2.2: InfiniBand pins **shall** withstand 2000 V of ESD using the human body model and 500 V using the charged device model, Class 2 per JEDEC JESD22-A114-B, with power applied, without damage or non-recoverable error including but not limited to latchup. A recoverable error is one that does not require reset or replacement of the device.

6.3.2 HOT INSERTION AND REMOVAL

C6-3: InfiniBand Devices, including modules and chassis, **shall not** be damaged by unexpected removal or insertion. Removal may occur while the link is operating without damage to either port on the link. Insertion or removal may occur with power on or power off.

6.3.3 ENVIRONMENT

C6-4: This compliance statement is obsolete and has been replaced by C6-4.2.1:.

C6-4.2.1: InfiniBand links **shall** meet all specifications in this chapter when operating in an InfiniBand chassis or other package within the environment as defined in Section 9.5.2, “Cooling Requirements,” on page 398. The IB power supply is 12 Volts as defined in Chapter 12: Power / Hot Plug.

Architecture Note

InfiniBand does not specify tolerances or values for the local regulated power supply because the local power supply is regulated locally from the bulk power supply. Only bulk power and aux power are supplied at the IB backplane connector. Regulated Power does not appear at any IB interface.

6.3.4 HIGH SPEED DIFFERENTIAL SHIELD RETURNS

C6-5: The **IB_Sh_Ret** connections are specified on the cable connector to support the isolation of the high speed differential inputs and outputs. These shield returns **shall** be connected to logic ground on the module.

Implementation Note

The primary purpose of these connections is to provide for isolation of the differential signals from each other. These shields also help to insure that the desired impedance of the link is maintained.

C6-6: The bulk shield is used for EMI control and **shall** be connected to chassis ground as defined in Section 7.9.4, “Cable Shield Connections,” on page 192.

6.4 DIFFERENTIAL DRIVER OUTPUTS

The SDR driver characteristics are specified so as to guarantee the specified differential signal characteristics at the receiver as described in Section 6.5, “Differential Receiver Inputs,” on page 191 within defined InfiniBand topologies. Some driver parameters are not specified, but can be derived from other specified parameters.

SDR, DDR and QDR Driver characteristics **shall** enable the receiver to achieve the specified BER when connected by a compliant channel. Values for driver characteristics are provided but may not be sufficient to guarantee proper operation. In case of conflict, driver characteristics shall be determined from the requirement to achieve the specified BER with a Compliant Channel.

Transmitters are specified at the board side of the backplane or cable connector, TP1 or TP5 in Figure 34. Values at the SERDES (ASIC) pin are informative.

C6-7: This compliance statement is obsolete and has been replaced by C6-7.2.1:.

C6-7.1.1: This compliance statement is obsolete and has been replaced by C6-7.2.2:.

C6-7.2.1: All output ports **shall** comply with the parameters and notes of Table 17 Driver Characteristics for 2.5 Gb/s using appropriate parameters as noted while operating at SDR. The parameters are defined in terms of values at IB port pins.

In addition to the parameters defined for "normal" IB signaling from device to device defined as Port Type 1, additional more restrictive parameters are defined as Port Type 2 for 1x Pluggable Devices such as Optical Transceivers in the Small Form Factor Pluggable (SFP) form factor in order to be compatible with the MSA for that package.

C6-7.2.2: 1x Pluggable ports **shall** meet the jitter J_{T2} and J_{D2} defined for Port Type 2 in addition to the other parameters of Table 17 Driver Characteristics for 2.5 Gb/s.

o6-7.2.1: Any output port claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance **shall** comply with the parameters and notes of Table 18 Driver Characteristics for 5.0 Gb/s while operating at DDR.

o6-7.2.2: Any output port claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance **shall** comply with the parameters and notes of Table 19 Driver Characteristics for 10 Gb/s while operating at QDR.

6.4.1 DC BLOCKING

o6-1: This compliance statement is obsolete and has been replaced by o6-7.2.1:

o6-7.2.1: The use of DC blocking capacitors is **optional** for backplane connections at SDR rate.

Explanation: The specified levels and termination allow the direct connection of driver to receiver if desired in environments with a common signal ground reference. Direct connection will minimize cost and maximize signal integrity. At higher speeds, DC blocking capacitors may be required.

C6-8: Cable ports **shall** incorporate DC blocking capacitors located at the receiver.

o6-2: Optionally, additional DC blocking capacitors may be located at the transmitter.

o6-8.2.1: DC blocking capacitors **may** be located at any IB pin (or pin pair).

Any loss or jitter caused by the addition of capacitors must be accounted for as part of the allocation for the printed wiring board on which the capacitors are mounted.

C6-8.2.1: DC blocking capacitors **shall not** be mounted inside the cable assembly.

Capacitor value selection is left to the designer. Capacitors must be large enough to provide a -3dB point below 100 MHz. Impedance and voltage values and measurements below 100 MHz will be affected by the presence of blocking capacitors. Appropriate allowances shall be made when evaluating compliance.

Capacitor DC working voltage is left to the designer. Since the differences in common mode voltage are limited by the power supply voltage and the grounding of the shields, a value of 10 volts should be more than adequate. ESD resistance should be taken into account, since the Cable and Backplane connector pins also need to meet the ESD specifications.

Layout of the mounting pads and vias for the blocking capacitor, and selection of the capacitor value and model must be carefully done in order to minimize impedance discontinuities. At higher speeds, blind or back drilled vias may be required.

6.4.2 EQUALIZATION

The frequency dependent attenuation of the interconnection media degrades the signal and thus produces Inter-Symbol Interference or Data Dependent Jitter which is a component of the Deterministic Jitter. The effects of high frequency attenuation can be reduced by techniques such as:

- Pre-distortion or Pre-emphasis/de-emphasis of the signal produced at the driver.
- Addition of a passive high pass filter network which has a frequency response complementary to that of the interconnect between driver and receiver. This is sometimes called Passive Equalization.
- Adaptive equalization using techniques such as partial response or DFE may be implemented at the receiver.

The method of equalization selected is dependent on the frequency of operation and the generation (version) of the ports.

Prior releases of the InfiniBand specification described fixed pre-emphasis or filtering rather than adaptive equalization when operating at 2.5 Gbits/second (SDR) for the first generation backplane interface. A fixed Total Jitter of 0.30 UI is allocated to the interconnect between the transmitter ASIC and the Receiver ASIC. Prior release SDR cables incorporated equalization filters as necessary to limit total jitter.

Ports supporting Rel. 1.2 Enhanced Signaling use a combination of fixed or adaptive transmit equalization and variable receive equalization. The equalization when Enhanced Signaling ports are inter-operating with legacy ports shall be transmit equalization for the ES transmitter and pre-emphasis or filtering for the legacy transmitter. This may require a new hybrid cable at longer lengths, or an adapter to provide the necessary filtering.

C6-9: This compliance statement is obsolete and has been replaced by C6-9.2.1:.

C6-9.2.1: InfiniBand connection equalization **shall** conform to the following restrictions.

- Backplane connections **shall** use either a passive equalization network, driver pre-emphasis, or some combination of the two to compensate for the maximum loss as specified in Table 23. This equalization or pre-emphasis **shall** be provided on the InfiniBand module which contains the transmitter. Backplane connections which exceed the specified loss **shall** provide additional equalization on the backplane. Note that this pre-emphasis/de-emphasis may cause the subsequent bits in a run to have an amplitude lower than the specified minimum. Figure 30: Waveform with De-emphasis for Backplane and Cable shows examples.
- Equalization filters **shall not** be provided on Infiniband module cable ports. Equalization for copper cables is provided within the cable assembly (see Section 7.7.3, “Equalization,” on page 247). Pre-emphasis/de-emphasis may be provided. Note that this pre-emphasis/de-emphasis may cause the subsequent bits in a run to have an amplitude lower than the specified minimum, as shown in Figure 30 Waveform with De-emphasis for Backplane and Cable on page 184.
- Port Type 2 outputs **shall not** implement pre-emphasis except to a limited extent such that the pre-emphasis **shall not** cause the amplitude for any bit to violate the maximum or minimum amplitude specifications in Table 17.

IBTA

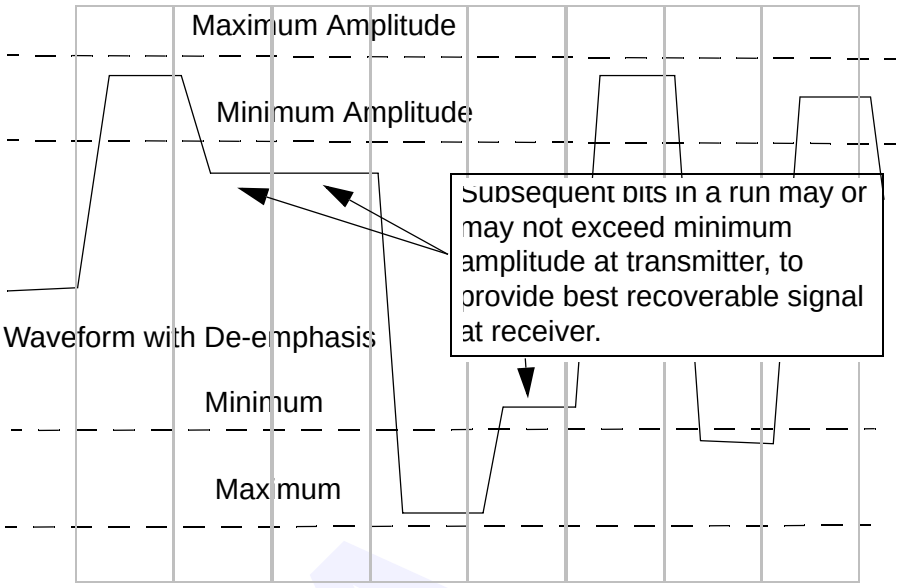
C6-9.2.2: InfiniBand Rel. 1.2 Enhanced Signaling connection equalization **shall** conform to the following restrictions for ports when operating at 2.5 Gb/s (SDR)

- Backplane connections **shall** use driver de-emphasis to compensate for the maximum loss as specified in Table 23.
- Cable ports **shall** use the specified driver de-emphasis to compensate for the maximum cable loss allowed. This is necessary for compatibility with legacy Receivers.
- Receive equalization **may** be provided however no training pattern is available other than the SDR initialization sequence as defined in Chapter 5: Link/Phy Interface.
- Equalization for copper cables **may** be provided external to the transmitter for those connections with legacy transmitters.

C6-9.2.3: InfiniBand Rel. 1.2 Enhanced Signaling connection equalization **shall** conform to the following restrictions for ports which may operate at 5.0 Gb/s (DDR) or 10 Gb/s (QDR)

- The transmitter **shall** provide the specified de-emphasis or equalization.
- The transmitter **shall** transmit a training pattern for the receive equalizer as defined in Chapter 5: Link/Phy Interface.
- Filters **shall not** be used in the cable or backplane. The allowed interconnect characteristics are defined in Section 6.8, “Compliance Channel - DDR and QDR,” on page 203
- Receivers **shall** incorporate equalization since the allowed de-emphasis will not be sufficient to guarantee an open eye under all circumstances

Figure 30 Waveform with De-emphasis for Backplane and Cable



IBTA

6.4.3 DIFFERENTIAL OUTPUT CHARACTERISTICS

Table 17 Driver Characteristics for 2.5 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Notes
V_{CM}	Output Common Mode Voltage (backplane connections)	1.2	0.30	V	(Vhigh+Vlow)/2 The common mode is undefined if DC blocking capacitors are used.
Vdiff (serdes)	Differential output Voltage At Transmit SerDes pins (Informative)	1.6	1.0	V	Differential unsigned waveform amplitude into 100 ohm differential load. Note 13,14,15.
Vdiffc	Differential output At cable connector pin (Normative)	1.6	0.89	V	Differential unsigned waveform amplitude into 100 ohm differential load. Note 13,14,15
Vdiffb	Differential output At Backplane connector pin (Normative)	1.6	0.80	V	Differential unsigned waveform amplitude into 100 ohm differential load Note 13,14,15
Vdisable	Disabled Mode output (Note 5)	1.6	0	V	The output in disabled or quiescent state may be zero volts differential.
Vstandby	Standby Mode output (Note 5)	1.6	0	V	
t_{DRF}	Driver Transition Time		30	ps	at 20-80% at the connector pins into 100 ohm load
I_{ACCM}	AC Common Mode current (Note 3, 6) (RMS Voltage)	5		uA	Determined by EMI restrictions and shielding effectiveness. 30 MHz to 6.25 GHz
V_{ACCM}	AC Common Mode Voltage (Note 3, 6, 7)	25		mV	RMS
S_{DD}	Differential Output return loss (Note 1)		10	dB	differential mode Note 11.
S_{SE}	Single Ended Output Return Loss (Note 1)		8	dB	Single ended, either output, current -15 to 15 ma, both pins driven. Note 11
Z_{SEDC}	Single Ended Output Impedance--Low Frequency Note 10, Note 12	10,000	30	Ω	Single ended, either output, current -15 to 15 ma, both pins driven Note 11

IBTA

Table 17 Driver Characteristics for 2.5 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Notes
S_{DC}	Common Mode to Differential Reflection	-20		dB	Reflected common mode that appears as differential mode. Note 11
I_{DShort}	Short Circuit Current	100	-100	mA	To any voltage between 1.6 and ground, power on or off.
S_{DBtB}	Skew	500		ps	Between any two physical lanes within a single transmitter.
J_D	Deterministic Jitter (Note 2, 8)	.17		UI	Without pre-emphasis
J_T	Total Jitter (Note 8)	.35		UI	At +/- 7s (10^{-12} BER)
J_{D2}	Deterministic Jitter (Note 8, Note 9)	0.09		UI	Without pre-emphasis Port Type 2
J_{T2}	Total Jitter (Note 8, Note 9)	0.24		UI	At +/- 7σ Port Type 2
UI_D	Unit Interval	400	400	ps	Plus/Minus 100 ppm measured over a minimum of 10,000 UIs.

1. At the InfiniBand connector pins from 100 MHz to 1.875 GHz.
2. A transmitter which implements pre emphasis shall meet the receiver mask with the minimum and maximum allowable interconnect configuration.
3. The AC common mode voltage is limited in order to control radiated EMI. Some causes of AC common mode are skew between the true and complement outputs of the differential driver, mismatched levels between the true and complement outputs of the differential driver, and power supply or ground noise such as that caused by switching activity being conducted through the driver.
4. The output level is increased above the minimum required by the specified backplane attenuation to allow for cable attenuation as well as provide headroom for noise. Amplitude is measured for the first bit in a run if a Backplane port, for all bits of a Cable Port.
5. Section 5:, “Link/Phy Interface,” on page 43 defines states in which the transmitter is quiescent. In these states driving node is inactive and there are no transitions on the link. Driver outputs shall be static in these states.
6. Common mode current is measured on an IB cable assembly using a detector bandwidth of 120 kHz below 1 GHz and 1MHz above 1 GHz. All physical lanes shall be transmitting the idle pseudo-random character sequence.
7. Since cable shield effectiveness is in excess of 40 dB and common mode impedance is approximately 50 ohms, driver common mode voltage of 25 mV is specified
8. Jitter is measured as defined in IBTA CIWG Test Specification.
9. Port Type 2 - Obsolete
10. DC to 100 MHz
11. Driver impedances selected to adequately absorb reflections and other noise Return Loss measured with respect to 100 ohms, 100 MHz to 3.75 GHz
12. In the absence of DC blocking capacitors. Undefined if DC blocking capacitors present.
13. Refer to IEEE Std.181-2003 for definitions and procedures around un-signed

IBTA

waveform amplitude measurements. (amplitude unsigned.)
14.Waveform amplitude measurements shall be conducted only on consecutive transition bits, over a 2UI Epoch (IEEE Std. 181-2003) unit interval, to a minimum population of 10E4 UI's. This test should apply to any unsigned consecutive transition bit in the test pattern.
15: Top and Base level calculations should be determined based on the histogram mean technique as described in IEEE Std. 181--2003 (Determining State Levels from the Histogram).

Table 18 Driver Characteristics for 5.0 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Notes
V_{CM}	Common Mode Voltage (Backplane connections) Note 12	1.2	0.75	V	(Vhigh+Vlow)/2 The common mode is undefined if DC blocking capacitors are used.
Vdiff	Differential output (Note 4) (serdes pins, informative)	1.6	0.8	V	Differential unsigned waveform amplitude into 100 ohm differential load Note 13,14,15
Vdiffc	Differential output (Note 4) (TP6, normative)	1.6	0.65	V	Differential unsigned waveform amplitude into 100 ohm differential load Note 13,14,15
Vdiff	Differential output (Note 4) (TP1, normative)	1.6	0.60	V	Differential unsigned waveform amplitude into 100 ohm differential load Note 13,14,15
Vdisable	Disabled Mode output (Note 5)	1.6	0	V	The output in disabled or quiescent state may be zero volts differential. DC value
Vstandby	Standby Mode output (Note 5)	1.6	0	V	DC value
t_{DRF}	Driver Transition Time		30	ps	At 20-80% at the package pins into 100 ohm load
I_{ACCM}	AC Common Mode current (Note 3, Note 6) (RMS Voltage)	5		uA	Determined by EMI restrictions and shielding effectiveness. 30 MHz to 6.25 GHz
V_{ACCM}	AC Common Mode Voltage (Note 3, 6, 7)	25		mV	RMS
S_{DD}	Differential Output Return Loss (Note 1)		10	dB	Differential mode Note 11.

Table 18 Driver Characteristics for 5.0 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Notes
S_{CC}	Single Ended Return Loss (Note 1)		8	dB	Single ended, either output, current -15 to 15 ma, both pins driven. Note 11
Z_{SEDC}	Single Ended Output Impedance--Low Frequency (Note 10, Note 12)	10,000	30	Ω	Single ended, either output, current -15 to 15 ma, both pins driven Note 11
S_{DC}	Common Mode to Differential Reflection	-20		dB	Reflected common mode that appears as differential mode. Note 11
I_{DShort}	Short Circuit Current	100	-100	mA	To any voltage between 1.6 and ground, power on or off.
S_{DBtB}	Skew	500		ps	between any two physical lanes within a single transmitter.
J_{D1}	Deterministic Jitter (Note 2, Note 8)	0.15		UI	Without pre-emphasis
J_{T1}	Total Jitter (Note 8)	0.30		UI	At +/- 7σ
UI_D	Unit Interval	200	200	ps	Plus/Minus 100 ppm measured over a minimum of 10,000 UIs.

1. At the InfiniBand connector pins from 100 MHz to 3.75 GHz.
2. A transmitter which implements pre emphasis **shall** meet the receiver mask with the minimum and maximum allowable interconnect configuration.
3. The AC common mode voltage is limited in order to control radiated EMI. Some causes of AC common mode are skew between the true and complement outputs of the differential driver, mismatched levels between the true and complement outputs of the differential driver, and power supply or ground noise such as that caused by switching activity being conducted through the driver.
4. The output level is increased above the minimum required by the specified backplane attenuation to allow for cable attenuation as well as provide headroom for noise. Amplitude is measured for the first bit in a run.
5. Section 5:. “Link/Phy Interface,” on page 43 defines states in which the transmitter is quiescent. In these states driving node is inactive and there are no transitions on the link. Driver outputs shall be static in these states.
6. Common mode current is measured on an IB cable assembly using a detector bandwidth of 120 kHz below 1 GHz and 1MHz above 1 GHz. All physical lanes shall be transmitting the idle pseudo-random character sequence.
7. Since cable shield effectiveness is in excess of 40 dB and common mode impedance is approximately 50 ohms, driver common mode voltage of 25 mV is specified
8. Jitter is measured as defined in IBTA CIWG Test Specification.
9. N/A - obsolete.
10. DC to 100 MHz

IBTA

11. Driver impedances selected to adequately absorb reflections and other noise Return loss with respect to 100 ohms 100 MHz to 6.25 GHz
12. In the absence of DC blocking capacitors. Undefined if DC blocking capacitors present.
13. Refer to IEEE Std.181-2003 for definitions and procedures around un-signed waveform amplitude measurements. (amplitude unsigned.)
14.Waveform amplitude measurements shall be conducted only on consecutive transition bits, over a 2UI Epoch (IEEE Std. 181-2003) unit interval, to a minimum population of 10E4 UI's. This test should apply to any unsigned consecutive transition bit in the test pattern.
15: Top and Base level calculations should be determined based on the histogram mean technique as described in IEEE Std. 181--2003 (Determining State Levels from the Histogram).

Table 19 Driver Characteristics for 10 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Notes
V_{CM}	Common Mode Voltage Backplane connection Note 12	1.2	0.75	V	(Vhigh+Vlow)/2 The common mode is undefined if DC blocking capacitors are used.
V_{diff}	Differential output (Note 4) (Serdes pins, informative)	1.6	0.6	V	Differential unsigned waveform amplitude into 100 ohm differential load Note 13,14,15
V_{diffc}	Differential output (Note 4) (TP6, normative)	1.6	0.50	V	Differential unsigned waveform amplitude into 100 ohm differential load Note 13,14,15
V_{diffb}	Differential output (Note 4) (TP1, normative)	1.6	0.45	V	Differential unsigned waveform amplitude into 100 ohm differential load Note 13,14,15
$V_{disable}$	Disabled Mode output (Note 5)	1.6	0	V	The output in disabled or quiescent state may be zero volts differential.DC
$V_{standby}$	Standby Mode output (Note 5)	1.6	0	V	DC
t_{DRF}	Driver Transition Time		30	ps	at 20-80% at the connector pins into 100 ohm load
I_{ACCM}	AC Common Mode current (Note 3, Note 6) (RMS Voltage)	5		uA	Determined by EMI restrictions and shielding effectiveness. 30 MHz to 6.25 GHz
V_{ACCM}	AC Common Mode Voltage (Note 3, Note 6, Note 7)	25		mV	RMS

IBTA

Table 19 Driver Characteristics for 10 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Notes
S_{DD}	Differential Output Return Loss (Note 1)	8		dB	Differential mode Note 11.
S_{CC}	Single Ended Output Return Loss (Note 1)	8		dB	Single ended, either output, current -15 to 15 ma, both outputs driven. Note 11
Z_{SEDC}	Single Ended Output Impedance--Low Frequency (Note 10, Note 12)	10,000	30	Ω	Single ended, either output, current -15 to 15 ma, both pins driven Note 11
S_{DC}	Common Mode to Differential Reflection	-20		dB	Reflected common mode that appears as differential mode. Note 11
I_{DShort}	Short Circuit Current	100	-100	mA	To any voltage between 1.6 and ground, power on or off.
S_{DBtB}	Skew	500		ps	between any two physical lanes within a single transmitter.
J_{D1}	Deterministic Jitter (Note 2, Note 8)	.15		UI	Without pre-emphasis
J_{T1}	Total Jitter (Note 8)	.30		UI	At +/- 7σ
UI_D	Unit Interval	100	100	ps	Plus/Minus 100 ppm measured over 10,000 UI.

1. At the InfiniBand connector pins from 100 MHz to 3.75 GHz.
2. A transmitter which implements pre emphasis shall meet the receiver mask with the minimum and maximum allowable interconnect configuration.
3. The AC common mode voltage is limited in order to control radiated EMI. Some causes of AC common mode are skew between the true and complement outputs of the differential driver, mismatched levels between the true and complement outputs of the differential driver, and power supply or ground noise such as that caused by switching activity being conducted through the driver.
4. Amplitude is measured for the first bit in a run if a Backplane port, for all bits of a Cable Port.
5. Section 5:, “Link/Phy Interface,” on page 43 defines states in which the transmitter is quiescent. In these states driving node is inactive and there are no transitions on the link. Driver outputs shall be static in these states.
6. Common mode current is measured on an IB cable assembly using a detector bandwidth of 120 kHz below 1 GHz and 1MHz above 1 GHz. All physical lanes shall be transmitting the idle pseudo-random character sequence.
7. Since cable shield effectiveness is in excess of 40 dB and common mode impedance is approximately 50 ohms, driver common mode voltage of 25 mV is specified
8. Jitter is measured as defined in IBTA CIWG Test Specification.

IB-A

9. N/A - obsolete
10. DC to 100 MHz
11. Driver impedances selected to adequately absorb reflections and other noise. Return Loss with respect to 100 ohms 100 MHz to 12.5 GHz.
12. In the absence of DC blocking capacitors. Undefined if DC blocking capacitors present.
13. Refer to IEEE Std.181-2003 for definitions and procedures around un-signed waveform amplitude measurements. (amplitude unsigned.)
14.Waveform amplitude measurements shall be conducted only on consecutive transition bits, over a 2UI Epoch (IEEE Std. 181-2003) unit interval, to a minimum population of 10E4 UI's. This test should apply to any unsigned consecutive transition bit in the test pattern.
15: Top and Base level calculations should be determined based on the histogram mean technique as described in IEEE Std. 181--2003 (Determining State Levels from the Histogram).

6.4.4 CABLE DRIVING

C6-10: This compliance statement is obsolete and has been removed.

6.5 DIFFERENTIAL RECEIVER INPUTS

C6-11: This compliance statement is obsolete and has been replaced by C6-11.2.1:.

C6-11.2.1: All input ports **shall** comply with the parameters and notes of Table 20 Receiver Characteristics for 2.5 Gb/s while operating at SDR. Parameters apply to the pin type as noted. The parameters are defined in terms of values at ASIC pins. They may be measured at accessible test points, with the values adjusted appropriately as defined in Section 6.7, "Compliance Points," on page 201. A BER of 10^{-12} shall be achieved when connected to the worst case transmitter through any compliant channel.

o6-11.2.1: All input ports claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance **shall** comply with the parameters and notes of Table 21 Receiver Characteristics for 5.0 Gb/s while operating at DDR. A BER of 10^{-12} shall be achieved when connected to the worst case transmitter through any compliant channel.

o6-11.2.2: All input ports claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance **shall** comply with the parameters and notes of Table 22 Receiver Characteristics for 10 Gb/s while operating at QDR. A BER of 10^{-12} shall be achieved when connected to the worst case transmitter through any compliant channel.

6.5.1 DIFFERENTIAL INPUT CHARACTERISTICS

Table 20 Receiver Characteristics for 2.5 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Conditions or Remarks
BER	Bit Error Ratio	10^{-12}			With minimum input
J_{DR}	Deterministic Jitter at Receiver	.47		UI	Port Type 1
J_{TR}	Total Jitter at Receiver	.65		UI	at 10^{-12} BER
Z_{RTerm}	Termination	62.5	40	Ω	To V_{tt} (differential impedance is double) See Figure 32.
V_{tt}	Termination Voltage Note 3, Note 6	1.0	0.5	V	See Figure 32
S_{DC}	Mode Conversion return loss		30	dB	The plus and minus rails of the signal are each terminated to V_{tt} and must be matched to avoid common mode to differential conversion. Note 4. Measured at TP1, TP6.
Z_{Vtt}	Vtt Impedance	30	0	Ω	Note 4.
Z_{Vtt}	Vtt Impedance Note 6	10,000	0	Ω	Note 5 Use of large values can reduce excess power caused by Vcm conflicts between driver and receiver.
L_{DR}	Differential Return Loss		10	dB	Note 4
L_{CMR}	Common Mode Return Loss		6	dB	Note 4
V_{RSense}	Input Sensitivity		175	mV	Minimum differential unsigned waveform amplitude. Note 1, 7, 8 See Figure 31 on page 193
V_{RSD}	Signal Threshold		85	mV	Minimum differential unsigned waveform amplitude. Note 2. See Figure 31 on page 193
Vrmax	Maximum Input Voltage	1.6		V	Maximum differential unsigned waveform amplitude.
V_{RCM}	Common Mode Voltage Note 6	1.25	0.25	V	(Vhigh+Vlow)/2 Note 3.
I_{ROff}	Off Current	50	-50	mA	Max. current into a pin with power off. Max voltage 1.6

IBTA

Table 20 Receiver Characteristics for 2.5 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Conditions or Remarks
V_{RHP}	Hot Plug Voltage (Voltage applied with power off or on)	1.6	-0.5	V	Applied without damage to any IB connector signal pin.
t_{REye}	Eye width		140	ps	Note 9 See Figure 31 on page 193.
S_{RBtB}	Total Skew	24		ns	Across receive physical lanes on a port. (see Section 5.6.7.4, "RxCMD = EnDeSkew," on page 153 Item 2)

1. Signals meeting the InfiniBand input specification shall be received with a maximum bit error rate of $1*10^{-12}$.
2. Signals having an un-signed differential amplitude less than VRSD shall be ignored. Signals having an un-signed differential amplitude greater than VRSD and less than VRSense may be ignored.
3. Unless DC blocking capacitors are present between the termination and the InfiniBand module pins.
4. Over a frequency range of 100 MHz to 1.875 GHz.
5. Frequency from DC to 100 MHz.
6. In the absence of DC blocking capacitors. Undefined if DC blocking capacitors present.
7. Refer to IEEE Std.181-2003 (Amplitude, Waveform, Unsigned) for definitions and measurement procedures.
8.Waveform amplitude measurement's shall be conducted only on consecutive transition bits, over a 2UI Epoch (IEEE Std. 181-2003), over a minimum population of 10E3 UI's. This test should apply to any unsigned consecutive transition bit in the test pattern.
9: Top and Base level calculation should be determined based on the histogram mean technique as described in IEEE Std. 181--2003 (Determining State Levels from the Histogram)

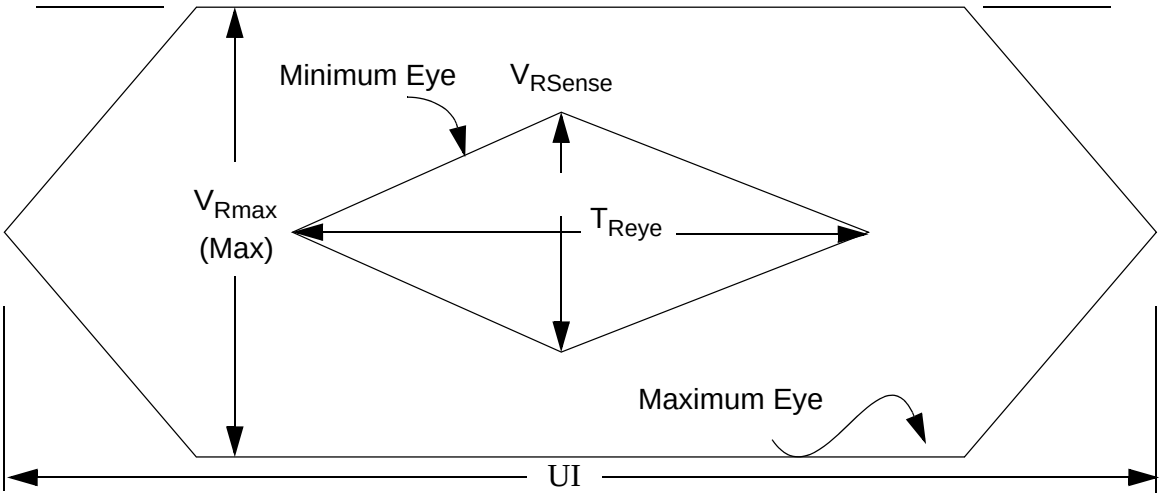


Figure 31 Eye Opening at receiver (Differential)

Table 21 Receiver Characteristics for 5.0 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Conditions or Remarks
BER	Bit Error Ratio	10^{-12}			Driven by compliant transmitter through compliance channel
S_{DD}	Differential return loss		8	dB	Reference to 100 ohms
S_{CC}	Common Mode Return Loss		6	dB	reference to 30 ohms. Note 4
V_{tt}	Termination Voltage	1.2	1.0	V	See Figure 32 Note 3, Note 6 May need to change
S_{DC}	Mode Conversion at input	-20		dB	The plus and minus rails of the signal are each terminated to V_{tt} and must be matched to avoid common mode to differential conversion. Note 4 Measured at IB connector pins.
Z_{Vtt}	Vtt Impedance		0	Ω	Note 5, Note 6 Use of larger values can reduce excess power caused by Vcm conflicts between driver and receiver.
V_{RSense}	Input Sensitivity (informative, at serdes input)		80	mV	Minimum differential unsigned waveform amplitude. Note 1,7,8 See Figure 31 on page 193
V_{CSense}	Input Sensitivity at cable connector (informative)		90	mV	Minimum differential unsigned waveform amplitude. Note 1,7,8,10 See Figure 31 on page 193
V_{BSense}	Input Sensitivity at Backplane connector (informative)		95	mV	Minimum differential unsigned waveform amplitude. Note 1,7,8,10 See Figure 31 on page 193
Vrmax	Maximum Input Voltage	1.6		V	Maximum differential unsigned waveform amplitude.

IBTA

Table 21 Receiver Characteristics for 5.0 Gb/s

V_{RCM}	Common Mode Voltage Note 6	0.70	1.2	V	(Vhigh+Vlow)/2 Note 3
I_{ROff}	Off Current	50	-50	mA	Max. current into a pin with power off. Max 1.6 V
V_{RHP}	Hot Plug Voltage (Voltage applied with power off or on)	1.6	-0.5	V	Applied without damage to any IB connector signal pin.
t_{REye}	Eye width	????		ps	Note 9 See Figure 31 on page 193.
S_{RBtB}	Total Skew	12		ns	Across receive physical lanes on a port. (see Section 5.6.7.4, "RxCMD = EnDeSkew," on page 153 Item 2)

1. Signals meeting the InfiniBand input specification shall be received with a maximum bit error rate of $1*10^{-12}$.
2. Signals amplitude informative only. DDR signal amplitude determined by transmit and channel
3. Unless DC blocking capacitors are present between the termination and the InfiniBand module pins.
4. Over a frequency range of 100 MHz to 3.75 GHz.
5. Frequency from DC to 100 MHz.
6. In the absence of DC blocking capacitors. Undefined if DC blocking capacitors present.
7. Refer to IEEE Std.181-2003 (Amplitude, Waveform, Unsigned) for definitions and measurement procedures.
8.Waveform amplitude measurement's shall be conducted only on consecutive transition bits, over a 2UI Epoch (IEEE Std. 181-2003), over a minimum population of 1000 UI's. This test should apply to any unsigned consecutive transition bit in the test pattern.
9: Top and Base level calculation should be determined based on the histogram mean technique as described in IEEE Std. 181--2003 (Determining State Levels from the Histogram mean technique)
10. Does not supersede Compliant channel requirement

Table 22 Receiver Characteristics for 10 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Conditions or Remarks
BER	Bit Error Ratio	10^{-12}			Driven by compliant transmitter through compliance channel
V_{tt}	Termination Voltage Note 3, Note 6	1.2	0.9	V	See Figure 32
S_{DD}	Differential Return Loss		8	dB	Note 4
S_{CC}	Common Mode Return Loss		6	dB	Note 4
S_{CD}	Mode Conversion, Common mode to differential	-20		dB	The plus and minus rails of the signal are each terminated to V_{tt} and must be matched to avoid common mode to differential conversion. Note 4 Measured at IB connector pins.
Z_{Vtt}	Vtt Impedance Note 6	10,000	0	Ω	Note 5 Use of large values can reduce excess power caused by Vcm conflicts between driver and receiver.
V_{RSense}	Input Sensitivity (informative, at serdes pins)		60	mV	Minimum differential unsigned waveform amplitude. Note 1,7,8 See Figure 31 on page 193
V_{CSense}	Input Sensitivity at cable connector.		70	mV	Minimum differential unsigned waveform amplitude. Note 1,7,8,10 See Figure 31 on page 193
V_{BSense}	Input Sensitivity at Backplane connector		60	mV	Minimum differential unsigned waveform amplitude. Note 1,7,8,10 See Figure 31 on page 193
Vrmax	Maximum Input Voltage	1.6		V	Maximum differential unsigned waveform amplitude.

IBTA

Table 22 Receiver Characteristics for 10 Gb/s

Symbol	Parameter	Maximum	Minimum	Units	Conditions or Remarks
V_{RCM}	Common Mode Voltage Note 6	1.25	.8	V	(Vhigh+Vlow)/2 Note 3.
I_{ROff}	Off Current	50	-50	mA	Max. current into a pin with power off.
V_{RHP}	Hot Plug Voltage (Voltage applied with power off or on)	1.6	-0.5	V	Applied without damage to any IB connector signal pin.
t_{REye}	Eye width	????		ps	Note 9 See Figure 31 on page 193.
S_{RBtB}	Total Skew	6		ns	Across receive physical lanes on a port. (see Section 5.6.7.4, "RxCMD = EnDeSkew," on page 153 Item 2)

1. Signals meeting the InfiniBand input specification shall be received with a maximum bit error rate of $1*10^{-12}$.
2. Signals VRSD for information only. Input signal determined by Transmitter and compliance channel.
3. Unless DC blocking capacitors are present between the termination and the InfiniBand module pins.
4. Over a frequency range of 100 MHz to 7.5 GHz.
5. Frequency from DC to 100 MHz.
6. In the absence of DC blocking capacitors. Undefined if DC blocking capacitors present.
7. Refer to IEEE Std.181-2003 (Amplitude, Waveform, Unsigned) for definitions and measurement procedures.
8.Waveform amplitude measurement's shall be conducted only on consecutive transition bits, over a 2UI Epoch (IEEE Std. 181-2003), over a minimum population of 1000 UI's. This test should apply to any unsigned consecutive transition bit in the test pattern.
9: Top and Base level calculation should be determined based on the histogram mean technique as described in IEEE Std. 181--2003 (Determining State Levels from the Histogram)
10. Does not supersede Compliant channel requirement.

6.5.2 TERMINATION

Figure 32 shows the concept of termination and signal definition. Differential and common mode (odd and even mode) termination is important to dampen noise and reflections in both modes.

Figure 32 Termination and Signaling

6.5.3 BEACON SIGNALING

C6-12: This compliance statement is obsolete and has been replaced by C6-13.2.1:.

C6-13: This compliance statement is obsolete and has been replaced by C6-13.2.2:.

C6-13.2.1: The beaconing sequence is described in detail in Section 5:, "Link/Phy Interface," on page 79. Devices which power down in $\mathbf{X_{Sleep}}$ state **shall** detect the beaconing sequence while operating on Auxiliary power. All devices **shall** detect the beaconing sequence. The beaconing sequence is transmitted at SDR only.

C6-13.2.2: The beacon detection shall meet the following requirements: The beaconing detector **shall** detect a minimum unsigned waveform amplitude of 175 mV p-p as a valid signal present. Unsigned waveform amplitudes less than 85 mV p-p **shall** be considered absent. Signals with transitions only at less than 10 MHz **shall** be considered absent. The beaconing sequence is transmitted at the in band signaling rate of 2.5 Gbits/second with an active period of 2 ms and a quiescent period of 100 ms (See Section 5.6.4.2, "Polling States," on page 127).

Note: Beaconing occurs at 2.5 Gb/s. All links begin the initialization process at 2.5 Gb/s. Continued link connection is verified by periodic transmission and reception of the "Link Heartbeat" ordered-set.

6.5.4 SIGNALING CAUTIONS

C6-14: This compliance statement is obsolete and has been replaced by C6-14.2.1:.

C6-14.2.1: Note that there are many circumstances in InfiniBand in which no differential voltage will be applied at the IB receiver. Examples include sleeping links, disconnected cables, and partially used ports resulting from width negotiation. Therefore means **shall** be provided to disable or de-gate the data coming from such a receiver to assure that false error signals are not generated. This gating **shall** be maintained across speed changes as appropriate. Note that the heartbeat will be absent in a disconnected link.

6.5.5 REPEATER

C6-15: This compliance statement is obsolete and has been replaced by C6-15.2.1:.

C6-15.2.1: A repeater which retimes by recovering the clock and filtering it is allowed. The transmit port of any SDR retimer **shall** meet the specifications of Table 17 on page 185. The receive port of any SDR retimer **shall** meet the specifications of Table 20 on page 192. A repeater need be neither protocol aware nor capable of any error detection. Outputs corresponding to quiescent inputs **shall** be quiescent. Repeaters shall be capable of handling SDR signaling in addition to any other signaling rate with which compliance is claimed. Non SDR repeaters may need to be protocol aware to a limited degree in order to participate in the speed negotiation and equalization or have capabilities to handle SDR data with DDR or QDR clocking.

6.6 LOSS MODEL FOR 2.5 GB/S SIGNALING

Note that Section 6.6, "Loss Model for 2.5 Gb/s signaling," on page 199 applies only to SDR 2.5 Gbit/s signaling.

6.6.1 BACKGROUND AND REASONING

Allowable interconnect is specified in terms of attenuation at various package boundaries and at two or more frequencies. This bounds the acceptable configurations but does not guarantee operation. Both the driver and the interconnect are responsible for producing the specified signal at the receiver pins.

6.6.2 LOSS VALUES

C6-16: This compliance statement is obsolete and has been replaced by C6-16.2.1:.

C6-16.2.1: Table 23 defines the maximum values of the loss assumed in IB interconnect topologies. The following sections define each of the

parameters. IB components **shall** limit loss at each package level to provide the eye openings and amplitudes at the Compliance Test Points as shown in Table 24 2.5 Gb/s Signal Test Points on page 202, when operating at 2.5 Gb/s (SDR).

Table 23 SDR Loss Parameter Maximum Values

Loss Parameter	Symbol	dB @ 2.5Gb/s, 1.25 GHz	dB at 625 MHz
Total Loss	L_T	15 dB	8 dB
Adapter Board[1]	L_A	2.5 dB	1.8 dB
Cable Assembly[2]	L_C	See Chapter 7: Copper Cable	
I/O Plate	L_S	1 dB	0.6 dB
Active Backplane	L_{BA}	9.5 dB	6.5 dB
Passive Backplane	L_{BP}	7 dB	5 dB
Crosstalk	L_{XT}	3dB	N/A

[1] includes a loss based on 1 pf via at the connector. Implementations must account for larger vias, if used.
[2] Cable attenuation is measured from the connecting vias on the IB board. Allowable values are defined in Chapter 7: Copper Cable.
3. Loss at 625 MHz specified for predictable equalization. Attenuation is assumed to be

6.6.2.1 TOTAL LOSS (LT)

Total loss is measured between the pins of communicating IC packages or between the receiving IC and the passive equalizer at the transmitter.

6.6.2.2 ADAPTER BOARD LOSS (LA)

The maximum loss from the IC package to the board connector is 2.5 db at 1250 MHz, including the connector.

6.6.2.3 CABLE ASSEMBLY LOSS (LC)

The maximum loss in a cable assembly, including the connector loss, is 10 db (value given here for reference only) as defined in Section 7:, "Copper Cable," on page 212.

6.6.2.4 I/O PLATE LOSS (LS)

The maximum loss from the IC package to the I/O plate connector is 1 dB not including the connector.

6.6.2.5 PLUGGABLE DEVICE LOSS

The maximum loss from the IC package to the Pluggable Device Connector is 1 dB.

6.6.2.6 CROSSTALK

80 mV referenced to the receiver of the loss budget is reserved for worst case crosstalk and other noise sources. Crosstalk is specified for the Backplane Connector in *InfiniBand Architecture Specification,* *Table 108, "Connector electrical performance requirements," on page 433*, and for the cable connectors in *InfiniBand Architecture Specification,* Table 31, "Cable connector electrical performance requirements," on page 214.

6.7 COMPLIANCE POINTS

Since the ASIC or Serdes pins are not accessible, other points within the IB defined topologies are defined and the signal levels listed in Table 24, "2.5 Gb/s Signal Test Points," on page 202 below. These signal levels and eye openings are derived from the already defined driver, receiver, and interconnect characteristics. All values are for an equivalent 100 ohm differential impedance load located at the test point. Correction or de-embedding must be performed to derive actual measured data, taking into account the configuration of the test setup.

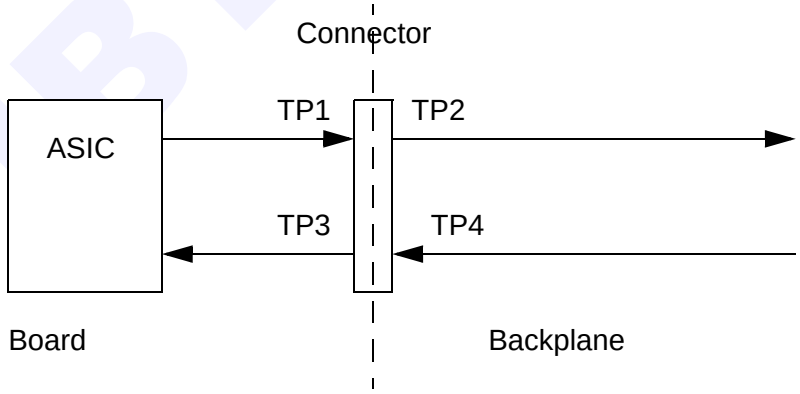


Figure 33 Board/Backplane test points

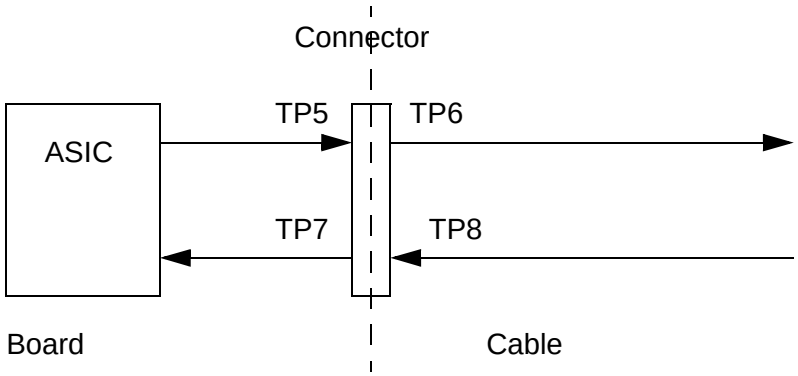


Figure 34 Board/Cable test points

Table 24 2.5 Gb/s Signal Test Points

Test Point	Description	Unsigned Waveform Amplitude Notes 1,2	Eye Width
TP1	Transmitted signal at board side of back-plane connector	.8	250 ps
TP2	Transmitted signal at backplane side of backplane connector	.75	240 ps
TP3	Received signal at board side of connector	.316	150 ps
TP4	Received signal at backplane side of con-nector	.335	160 ps
TP5	Transmitted signal at board side of cable connector	.89	250 ps
TP6	Transmitted signal at cable side of cable connector	.82	240 ps
TP7	Received signal at board side of cable con-nector	.282	150 ps
TP8	Received signal at cable side of cable con-nector	.30	160 ps
TP9	Transmitted signal at board side of plugga-ble interface socket	.89	296 ps
TP10	Received signal at board side of pluggable interface socket	.282	126 ps

Note: All measurements for only LUT active.

1. Refer to IEEE Std.181-2003 (Amplitude, Waveform, Unsigned) for definitions and

procedures around un-signed waveform amplitude measurements.
2. Waveform amplitude measurements shall be conducted only on consecutive transition bits, over a 2UI Epoch (IEEE Std. 181-2003) unit interval, over a minimum population of 10E4 UI's.

C6-16.1.1: This compliance statement is obsolete and has been replaced by C6-16.2.2:.

C6-16.1.2: This compliance statement is obsolete and has been replaced by C6-16.2.3:.

C6-16.1.3: This compliance statement is obsolete and has been replaced by C6-16.2.4:.

C6-16.2.2: SDR Backplane Connections shall comply with the amplitude and eye opening at TP1 and TP3. TP2 and TP4 are for reference.

C6-16.2.3: SDR Cable Connection shall comply with the amplitude and eye opening at TP5 and TP7. TP6 and TP8 are for reference.

C6-16.2.4: 1x Pluggable and 4x PluggableDevice Ports shall comply with the amplitude and eye opening at TP9 and TP10. Measurement may require de-embedding.

6.8 COMPLIANCE CHANNEL - DDR AND QDR

The signal characteristics seen by an InfiniBand DDR or QDR receiver are defined by the combination of the Transmitter characteristics as defined in the tables and any Compliant Channel as defined in this section. It is intended that the channel will be capable of being equalized by the receiver using equalizers such as a 5-tap DFE.

6.8.1 DDR (5.0 GBIT/SEC) COMPLIANT CHANNEL

A Compliant Channel is defined in terms of response to an isolated bit having an amplitude of 1.0 volt, measured at 1.0 UI intervals. Table 25 Compliant Channel Impulse response on page 204 defines the values. For test purposes, bits of alternating polarity should be used, separated by adequate time.

o6-16.2.1: Any receiver claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance **shall** achieve a BER of 10^{-12} when connected to an InfiniBand compliant Transmitter through any Compliant Channel when operating at 5.0 Gb/s (DDR).

6.8.2 QDR (10.0 GBIT/SEC) COMPLIANT CHANNEL

A Compliant Channel is defined in terms of response to an isolated bit having an amplitude of 1 volt, measured at 1.0 UI intervals as listed in

Table 25 Compliant Channel Impulse response on page 204. For test purposes, bits of alternating polarity should be used, separated by adequate time.

o6-16.2.1: Any receiver claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance **shall** achieve a BER of 10^{-12} when connected to an InfiniBand compliant Transmitter through any Compliant Channel when operating at 10.0 Gb/s (QDR).

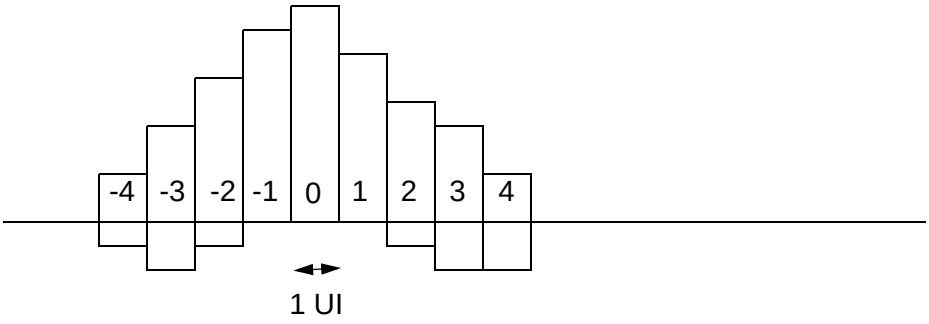
Table 25 Compliant Channel Impulse response

Interval	DDR Amplitude	QDR Amplitude
-5	.01	.01
-4	.01	.01
-3	.01	.01
-2	.01	.01
-1	.03	.03
0	.180	.180
1	.06	.06
2	.04	.04
3	.03	.02
4	.01	.01
5	.01	.01

IBTA

6.8.3 EXAMPLE COMPLIANT CHANNEL DEFINITION

Figure 35 Example Channel Definition



6.8.4 COMPLIANT CHANNEL S-PARAMETERS

Parameters are for maximum loss at Bit rate/2, and must be scaled for shorter configurations having less loss at Bitrate/2.

Table 26 S Parameters DDR (dB)

Frequency	SDD11	SDD21 (min)	SDD21 (max)	SDC11	SDC21
100 MHz	-10	0	-1	-20	-20
200 MHz	-10	-2	-1	-20	-20
625 MHz	-10	-7	-4	-20	-20
1250 MHz	-10	-10	-8	-18	-18
1875 MHz	-10	-12.5	-11	-15	-15
2500 MHz	-10	-15	-15	-10 dB	-10 dB

Table 27 S parameters QDR (dB)

Frequency	SDD11	SDD21 (min)	SDD21 (max)	SDC11	SDC21
100 MHz	-10		0	-20 dB	-20 dB
200 MHz	-10	-2	-1	-20	-20
625 MHz	-10	-3	-2	-20	-20
1250 MHz	-10	-5	-4	-20	-20

Table 27 S parameters QDR (dB)

Frequency	SDD11	SDD21 (min)	SDD21 (max)	SDC11	SDC21
1875 MHz	-10	-7	-6	-18	-18
2500 MHz	-10	-10	-8	-15	-15
3750 MHz	-10	-12.5	-11	-12	-12
5000 MHz	-10	-15 dB	-15	-10 dB	-10dB

6.9 BIT TO BIT SKEW

6.9.1 SKEW VALUES

C6-17: Table 28 defines the allowable bit to bit skew across the physical lanes. All IB ports **shall** limit skew to the values in Table 28 Bit to Bit Skew Parameter Maximum Values

Table 28 Bit to Bit Skew Parameter Maximum Values

Skew Parameter	Symbol	Value
Total Skew[a]	S_{RBtB}	60 UI
Driving Transceiver[b]	S_{DBtB}	500 ps
Backplane	S_{BPBtB}	500 ps
Adapter Board	S_{ABtB}	500 ps
Cable Assembly[c]	S_{CBtB}	See Chapter 7: and Chapter 8:
Transceiver to I/O Plate	S_{SBtB}	500 ps
Retiming Repeater (2 max per link) [d]	S_{RTR}	10 UI each, 20 UI total See Section 5.10, "Retiming Repeaters," on page 125

a. Defined in Table 20 Receiver Characteristics for 2.5 Gb/s on page 192 and included for reference.
b. Defined in Table 17 Driver Characteristics for 2.5 Gb/s on page 185 and included for reference.
c. Defined in Chapter 7: Copper Cable.
d. Retiming repeaters may independently insert or delete skip symbols on a per lane basis. Repeaters which do not utilize skips are not counted in this limit.

6.9.2 SKEW DEFINITION

Skew at any point is measured using the zero crossings of the differential voltage of the commas present in the training sequences TS1 and TS2 or in the Skip ordered-set as defined in Section 5.3.2, "Control Ordered-

sets," on page 98. Each of these sequences transmits a rank of commas on all physical lanes simultaneously.

6.10 ELECTRICAL TOPOLOGIES

This section outlines the electrical topologies accommodated within the InfiniBand specification.

Integrated Circuits (IC) that are depicted represent elements that generate and accept a specified InfiniBand interface. These ICs may be any one of the following types of functions:

- Host Channel Adapter (HCA)
- Target Channel Adapter (TCA)
- Switch
- Repeater (non-addressable retime and redrive component)
- Electrical/Optical Translation function
- Pluggable Device

6.10.1 ON-BOARD

This configuration is used for interconnection of IC on a common PCB.

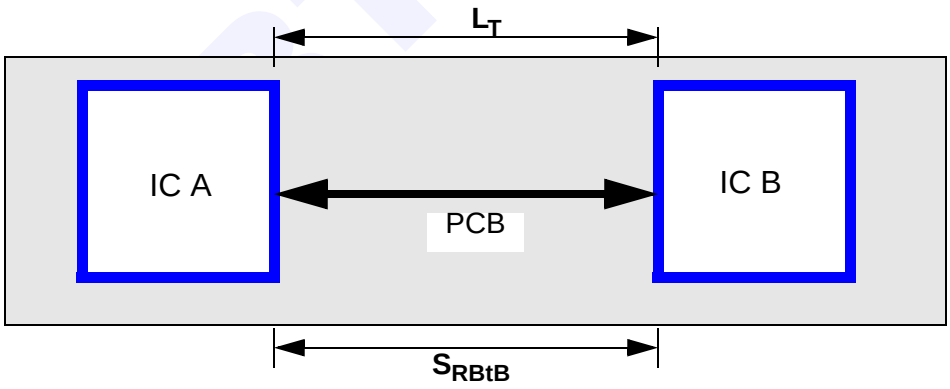


Figure 36 On-board Topology

6.10.2 SLOT ON ACTIVE BACKPLANE

This is the typical chipset bridge or switch chip on a backplane that interfaces to an adapter board plugged into the backplane. This topology is typical of PCI configurations.

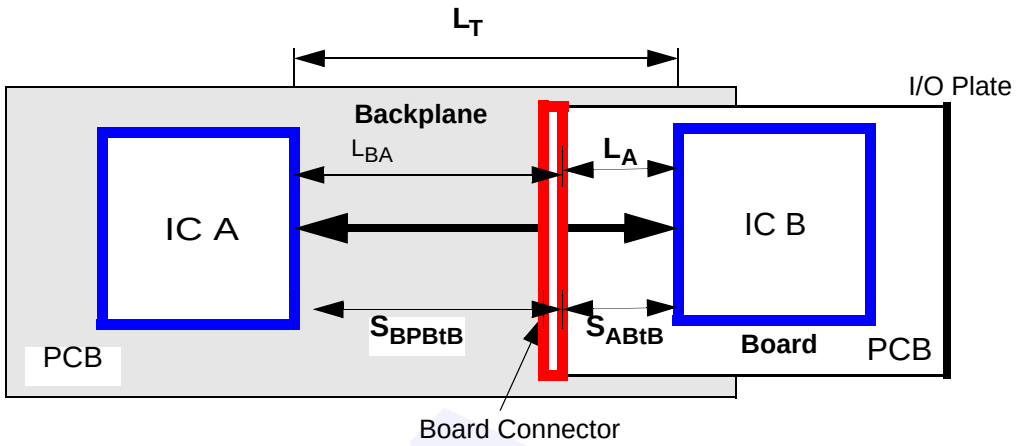


Figure 37 Slot on Active Backplane Topology

6.10.3 SLOT TO SLOT VIA BACKPLANE

This topology allows for one pluggable board (for example, a switch) to interface to another pluggable board (for example, a TCA) through a passive backplane.

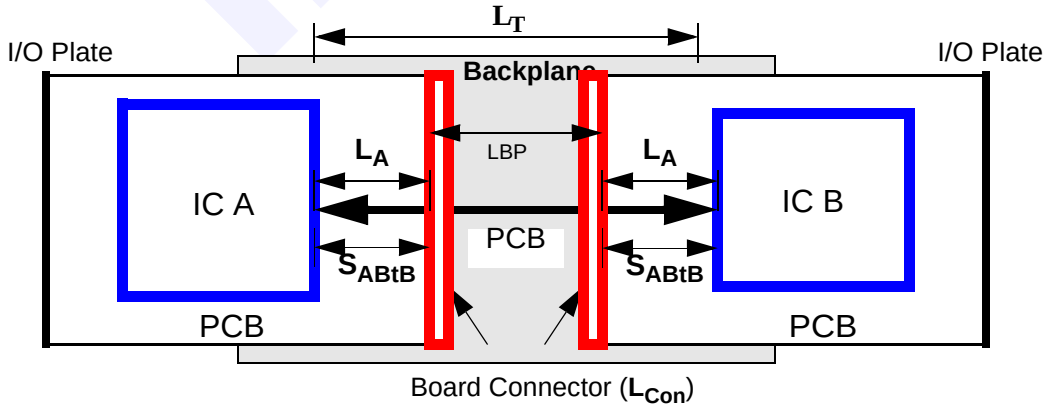


Figure 38 Slot to Slot on Passive Backplane Topology

6.10.4 COPPER INTERCONNECTS

These interfaces would be used for driving remote TCA or HCA through copper cables either from I/O Plate to I/O Plate, in the case of a switch to a switch or adapter, or from a backplane mounted switch to another backplane mounted switch. A cable will be one of 1x, 4x, 8x or 12x. A typical

application of this topology might be a stand-alone storage box connecting to another such box or a processor.

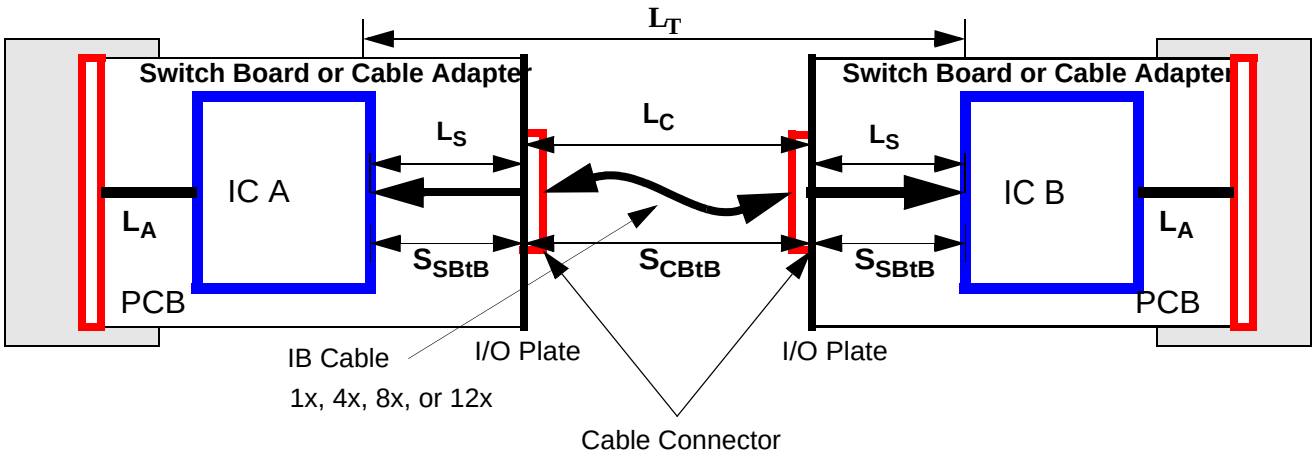


Figure 39 I/O Plate to I/O Plate via Cable Topology

Backplanes (or “System Boards”) may be connected together using cable.

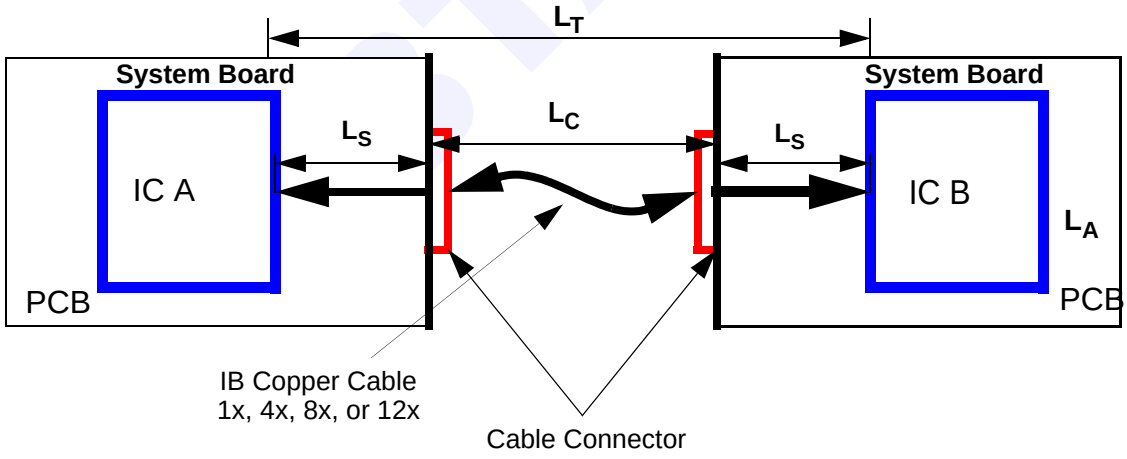


Figure 40 System Board to System Board via Cable Topology

6.10.5 OPTICAL FIBER

This topology may be used for driving remote IB ports at distances of up to several hundred meters (or several km using the long haul fiber and laser transmitters) through optical interfaces. Each end of the link may be mounted on pluggable boards with each housing optical transceiver components as shown in Figure 41 or be directly mounted on a system board as shown in Figure 42. This release of the InfiniBand specification specifies 1x and 4x short and long haul links, and 1x, 4x, 8x and 12x short haul

optical links, at several speeds. Backplanes may be connected together using optical links. The transceivers depicted translate the Infiniband™ copper interface into an optical interface to drive the optical componentry.

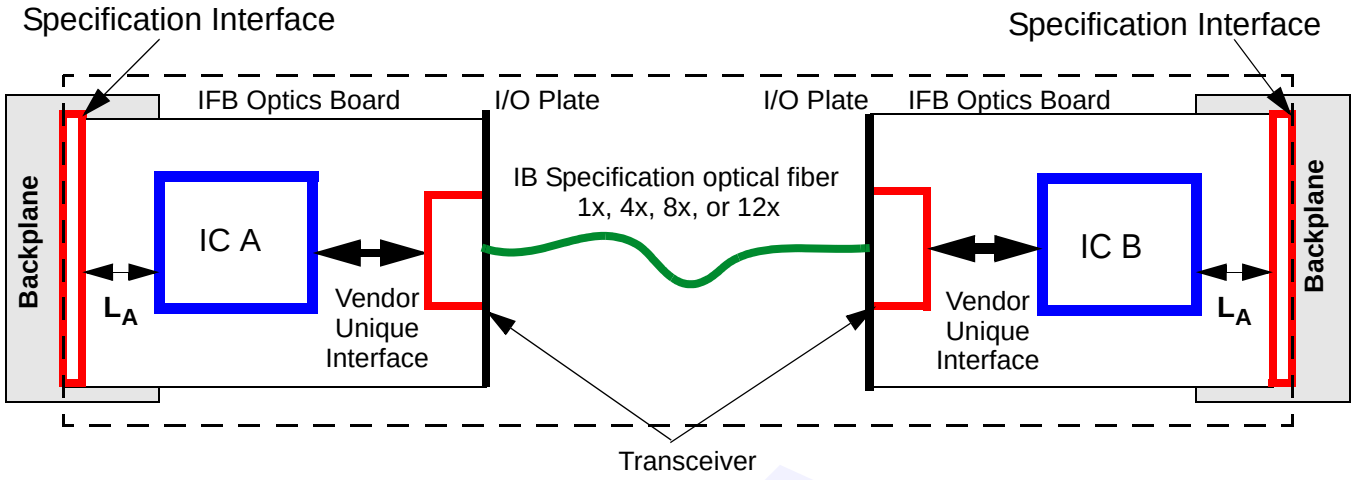


Figure 41 Optical Fiber Interconnect Topology

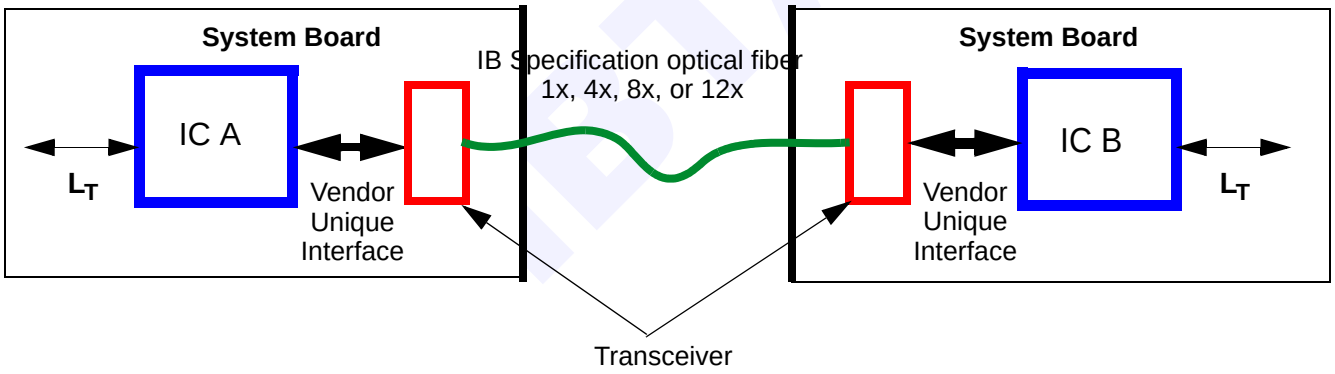


Figure 42 System Board Optical Fiber Interconnect Topology

6.10.6 PLUGGABLE

InfiniBand devices may be connected together by pluggable devices, as shown in Figure 43. ICs depicted translate the Infiniband™ copper inter-

face into an InfiniBand Pluggable Device interface to drive the interconnect componentry.

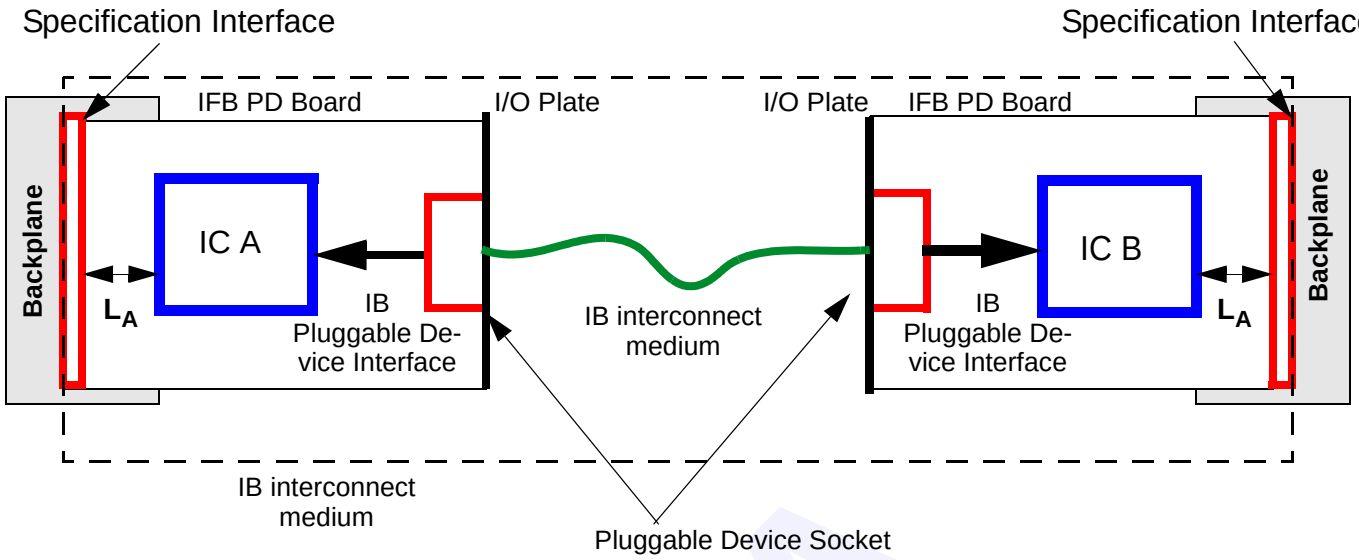


Figure 43 Pluggable Device Topology

6.11 PHYSICAL LAYER TEST AND CHARACTERIZATION FACILITIES

Test, characterization, and compliance verification of an IB port requires additional states, modes, or accessible signals that are not required for functional operation. These facilities, as specified in the Link/Phy section of this annex, are recommended but not required for legacy ports. These facilities are required for devices claiming compliance with Rel. 1.2 Enhanced Signaling.

The test and characterization modes are specified in the Link/Phy section of this annex, Section 5.12, “Physical Layer Compliance Testing,” on page 169.

o6-17.2.1: Any device claiming InfiniBand Rel. 1.2 Enhanced Signaling compliance at the slot interface, or copper cable interface **shall** comply with the requirements of Section 6.11 on page 211 regarding physical layer test and characterization facilities.

CHAPTER 7: COPPER CABLE

7.1 INTRODUCTION

This chapter defines InfiniBand cables and the connectors used to attach InfiniBand cables to InfiniBand modules. InfiniBand cable connectors **shall** incorporate features shown in the appropriate sections below that specify the connector to board interfaces.

It is the responsibility of the cable and connector designers or suppliers to perform the indicated tests and supply the data to potential customer companies to indicate compliance. It is recommended that the appropriate test groups specified in EIA-364.1000.01 be used for qualification testing, using the following conditions:

- 50 mating cycles preconditioning
- Un-mated exposure, option 2 mixed flowing gas exposure
- Five year product life
- Field operating temperature range up to 60 degrees C.

7.2 CABLE AND BOARD CONNECTORS - COMMON REQUIREMENTS

The requirements stated in this section apply to all connectors used to attach InfiniBand cables to InfiniBand modules. The following individual sections that define the specific connectors to be used for a given link width may also define additional requirements for that specific connector.

Use of the keying definitions included herein is optional; however, utilization of these definitions is highly recommended to prevent plugging of InfiniBand cables into ports using incompatible interfaces such as Fibre Channel.

7.2.1 PHYSICAL AND MECHANICAL PERFORMANCE REQUIREMENTS

C7-1: Connectors to be used on and for connection to InfiniBand modules **shall** meet or exceed the physical and mechanical performance requirements listed in Table 29.

It is also recommended that connector interfaces meet the parameters defined in Table 30. Unless otherwise noted, successful completion of a given test is indicated by an acceptable Low Level Contact Resistance measurement as indicated in EIA-364.1000.01 or other equivalent test sequence.

Note that all drawing dimensions in the following sections are in millimeters (mm).

Table 29 Cable connector physical requirements

	Parameter		Minimum	Maximum	Units	Conditions/Comments
N	Durability		250		mating cycles	Without physical damage or exceeding low level contact resistance when mated; no more than 1% of contacts with exposed base metal
F_i	Insertion force	1x		30	N	
		4x		56		
		12x		73		
F_w	Withdrawal force	1x		30	N	
		4x		49		
		12x		59		
F_r	Retention force		75		N	Load pull, per EIA 364-38A
F_{ls}	Side load capability		75		N	No damage to cable or board, no opens detected with LLCR test; force applied to the cable in a plane parallel to the I/O plate at a distance of 90 mm, in the direction of the smaller dimension of the receptacle
F_{ll}	Longitudinal load capability		100		N	No damage to cable or board, no opens detected with LLCR test; force applied to the cable in a plane parallel to the I/O plate at a distance of 90 mm, in the direction of the larger dimension of the receptacle
F_{rc}	Housing contact retention force		5		N	
t_{pm}	Contact finish – option 1		0.76 Au over 1.27 Nickel		μm	
t_{pm}	Contact finish – option 2		0.51 PdNi with Au flash over 1.27 Nickel		μm	Min. 75% Pd in PdNi alloy

IBTA

Table 30 Recommended cable connector physical requirements

Symbol	Parameter	Minimum	Maximum	Units	Conditions/Comment
F_n	Contact normal force	100		cN	per contact beam
S_{hcc}	Contact Hertz stress	170		kpsi	per contact beam
D_{wc}	Contact wipe	1.0		mm	

7.2.2 ELECTRICAL PERFORMANCE REQUIREMENTS

C7-2: Connectors to be used on and for connection to InfiniBand modules **shall** meet or exceed the electrical performance requirements for port type 1, listed in Table 31.

Table 31 Cable connector electrical performance requirements

Symbol	Parameter	Minimum	Maximum	Units	Conditions/Comment
LLCR	Low level contact resistance - initial		80	mΩ	through testing per EIA 364-23
ΔLLCR	Low level contact resistance - change		20	mΩ	through testing per EIA 364-23, as a result of any test group step
I_{max}	Current rating	0.5		A	per EIA-364-70 or IEC 512-5-1 Test 5a, at 30° C. temperature rise above ambient
Z_{dco}(peak)	Differential Impedance (peak)	90	110	Ω	Mated cable and board connector, average value measured over the propagation delay of the connector at 100 ps rise time (at the connector) per EIA 364-108. Includes connector, cable to connector interface, and board termination pads and vias but not equalizer.
Z_{dco}(nom)	Differential Impedance (nominal)	95	105	Ω	
L_{co}	Insertion loss		1.0	dB	Mated cable assembly and board connector, at frequencies up to 1.25 GHz, per EIA 364-101
S_{cop}	Within pair skew		5	ps	per EIA 364-103; by design, measurement not required
J_{co}	Jitter		10	ps	per EIA 364-107 with Fibre Channel CJTPAT stimulus until 10000 hits in max. 20 mV high jitter box or equivalent, with equipment and fixture contribution de-embedded; by design, measurement not required.
$NEXT_c$	Near end crosstalk		4	%	Mated cable and board connector, measured differentially with all adjacent neighbor pairs driven at 100 ps transition time, per EIA 364-90 and Section 6.6.2.6. A lower value will result in additional design margin. See Note 2.

IBTA

The signal eye pattern at the cable output must conform to the electrical requirements as listed in Table 31 and shown in Figure 60 in all cases.

Notes

1. All rise or transition times referenced in this chapter of the specification are assumed to be measured from the 20% to 80% levels of the waveform base state to top state amplitudes.

2. Crosstalk is calculated by dividing the larger of the amplitudes of the unsigned positive or negative differential crosstalk noise waveforms by the unsigned amplitude of the differential aggressor waveform, the result being multiplied by 100 to obtain percent. In the case of measurement by superposition, the noise voltage is the sum of the maximum noise amplitude values induced by the individual aggressor pair sources.

7.2.3 ENVIRONMENTAL PERFORMANCE REQUIREMENTS

C7-3: Connectors to be used on and for connection to InfiniBand modules **shall** meet or exceed the environmental performance requirements of EIA-364.1000.01, including exposure to Mixed Flowing Gas consistent with the required product life as defined in Section 7.1.

Unless otherwise noted, successful completion of a given test is indicated by an acceptable Low Level Contact Resistance measurement upon completion of the test, as defined above.

7.2.4 PORT LABELING

It is recommended that I/O ports be labeled to avoid confusion between InfiniBand ports and connections and those used for other incompatible interfaces. The recommended labeling is defined in Section 9.6.

7.3 1X INTERFACE BOARD CONNECTOR

This section defines the connector for the 1x cable interface on InfiniBand boards. The 1x interface is the narrowest InfiniBand link, providing for simultaneous transmit and receive of one bit of differential data.

7.3.1 DESCRIPTION

The connector used on InfiniBand boards for 1x cable ports is similar to that used for Fibre Channel. The receptacle is a surface-mounted design with seven contacts; four of the pins are signal contacts, and three con-

tacts are connected to signal ground. The connector is shielded to minimize EMI radiation. A suitable receptacle, referred to as HSSDC2 InfiniBand, is available from Tyco Electronics and others, and is shown in Figure 44. Detailed drawings of mating interface dimensions are shown in Figure 45.

C7-4: All 1x InfiniBand cable plugs **shall** be intermateable with this connector.

IBTA

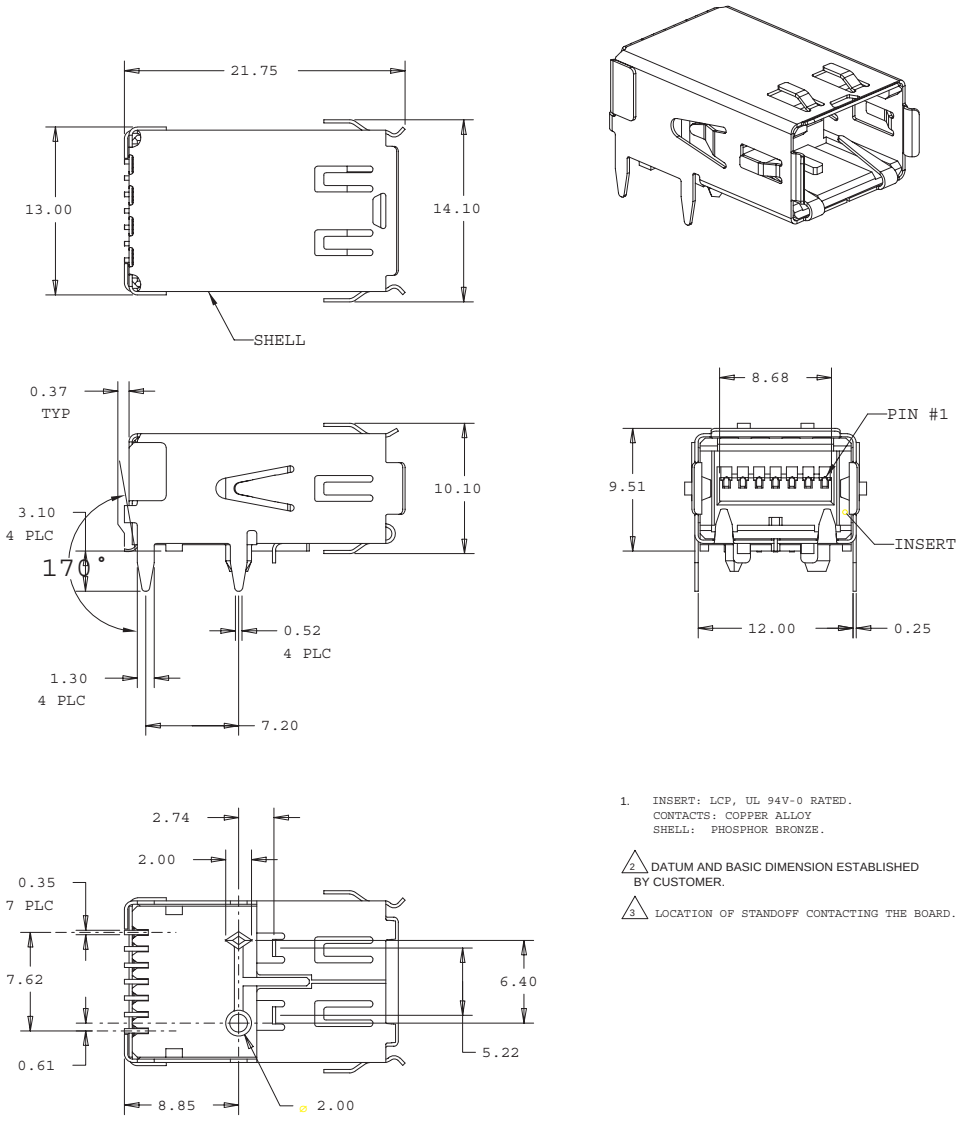


Figure 44 1x cable board connector

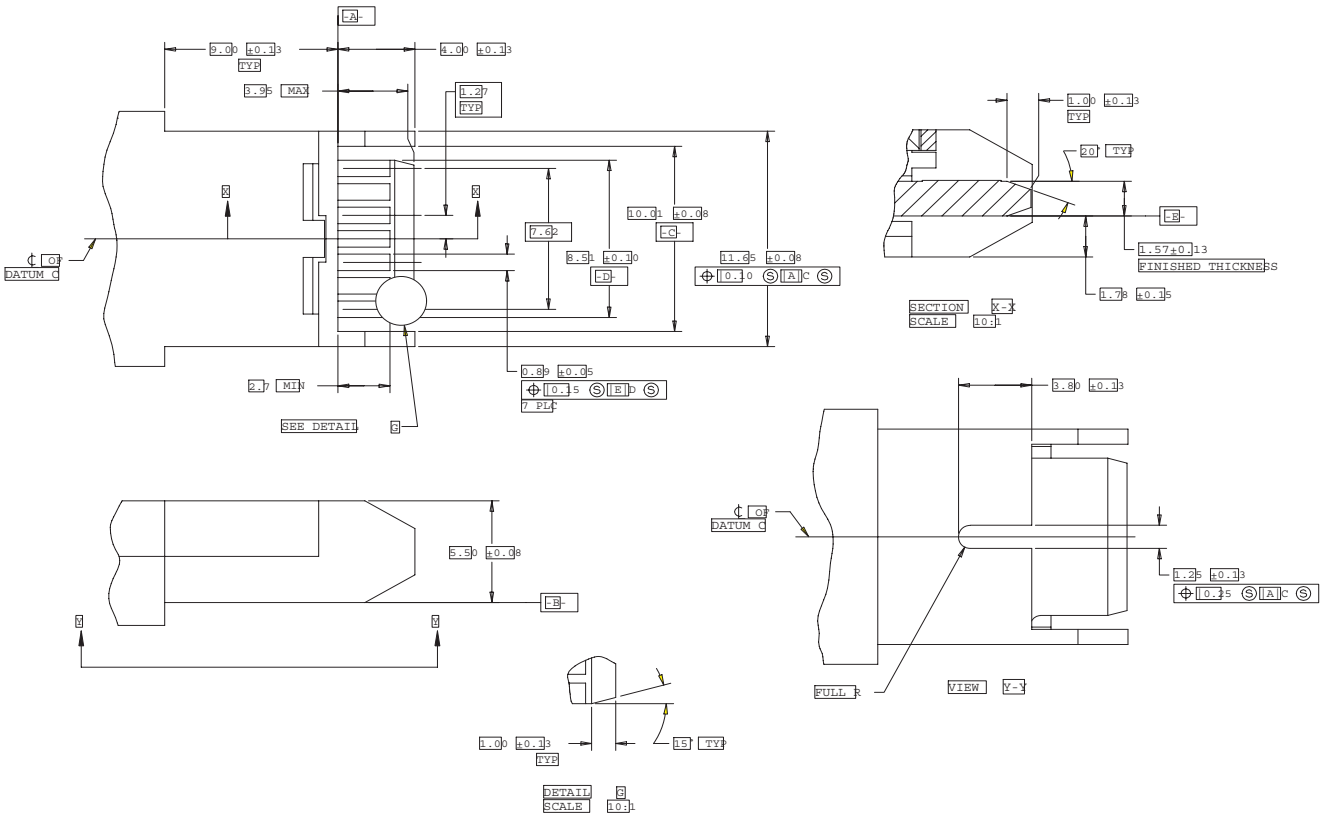


Figure 45 1x connector interface details

7.3.2 FOOTPRINT AND CONTACT PATTERN

The board footprint shown in Figure 46 should be used for the 1x cable connector shown in Figure 44. The connector is surface-mounted, so there are no contacts on the secondary side of the board.

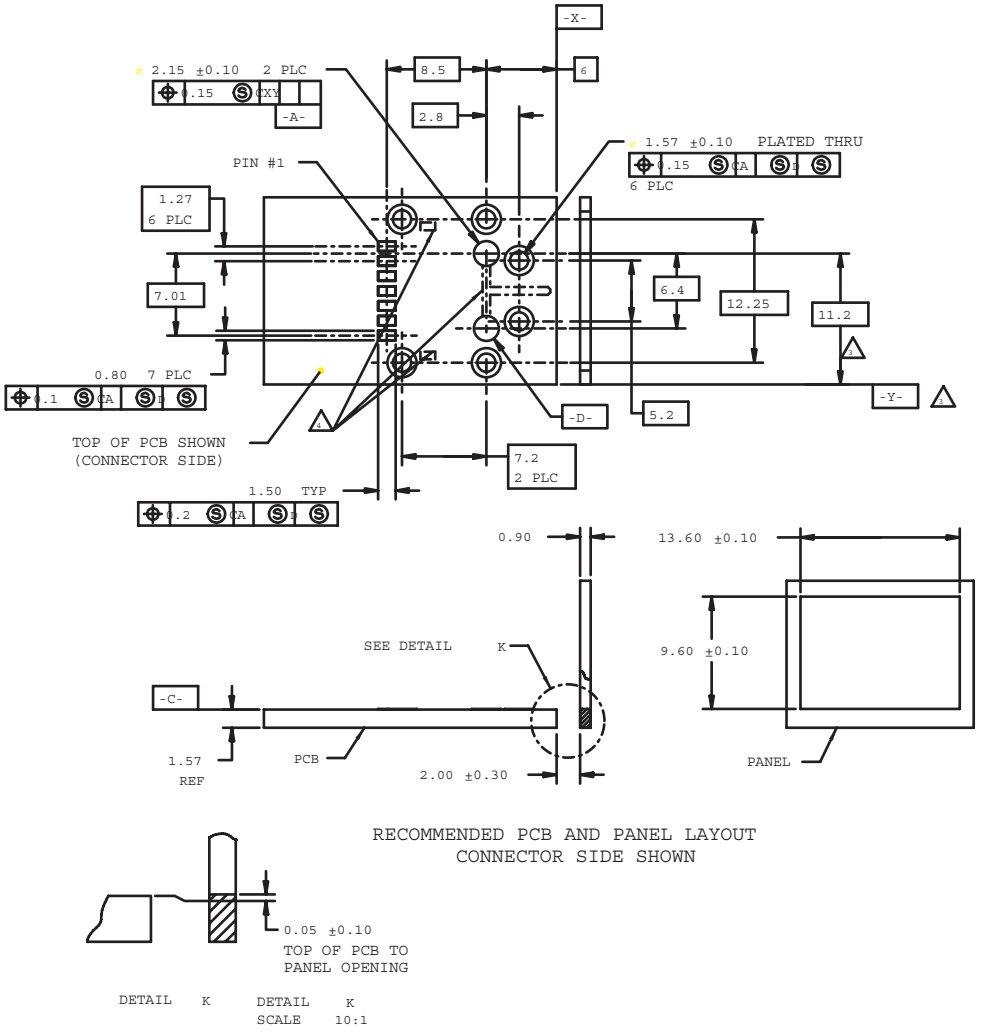


Figure 46 1x cable board connector footprint (top view) and panel layout

7.3.3 PIN ASSIGNMENT

C7-5: The pin assignment listed in Table 32 **shall** be used for the board connector for InfiniBand 1x cables.

The character 'x' in the signal symbol is the port number, as defined in Section 4.1, "Signal Naming Conventions," on page 69.

Table 32 1x cable board connector signal assignment

Pin Number	Signal
1	Signal Ground
2	IBtxIp(0)
3	IBtxIn(0)
4	Signal Ground
5	IBtxOn(0)
6	IBtxOp(0)
7	Signal Ground
Housing	Chassis Ground

C7-6: Signal Ground **shall** be connected to **IB_Sh_Ret** on the module.

C7-7: Signal Ground **shall not** be connected to Chassis Ground in the connector.

C7-7.1.1: A continuous ground path from the cable's inner shield(s) through the connector to the board signal ground **shall** be provided to insure low jitter, low crosstalk and EMI containment.

A cable constructed with only a bulk shield is not likely to meet the electrical requirements of Section 6.3.4.

7.3.4 KEYING

The 1x board connector accommodates a bottom-side (opposite the latch side) key with width of 1.0 mm, which is centered horizontally in the connector body, as shown in Figure 45.

7.4 4X INTERFACE BOARD CONNECTOR

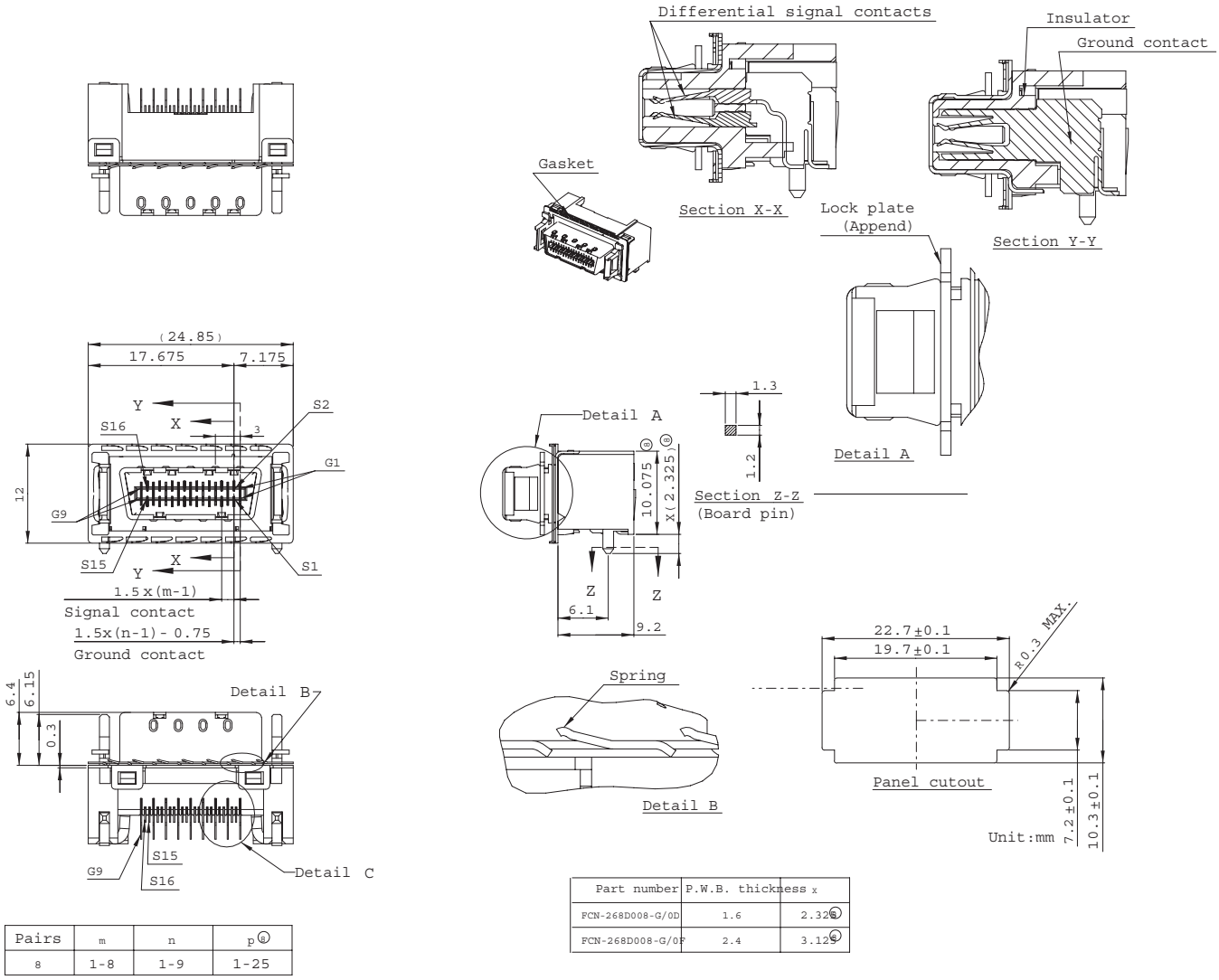
This section defines the connector for the 4x cable interface on InfiniBand boards. The 4x interface provides for simultaneous transmit and receive of four bits of differential data.

7.4.1 DESCRIPTION

This section defines the connector for the 4x cable interface on InfiniBand boards. The connector uses pairs of contacts separated by Ground contacts to reduce near end crosstalk (NEXT). The connector is surface-mounted, and all signal pins are used. A suitable receptacle, referred to as an eight pair MICRO GIGACN™ [1] receptacle, is available from Fujitsu

Components Ltd. and others, and is shown in Figure 47. Detailed drawings of mating interface dimensions are shown in Figure 48.

C7-8: All 4x InfiniBand cable plugs **shall** be intermateable with this connector.



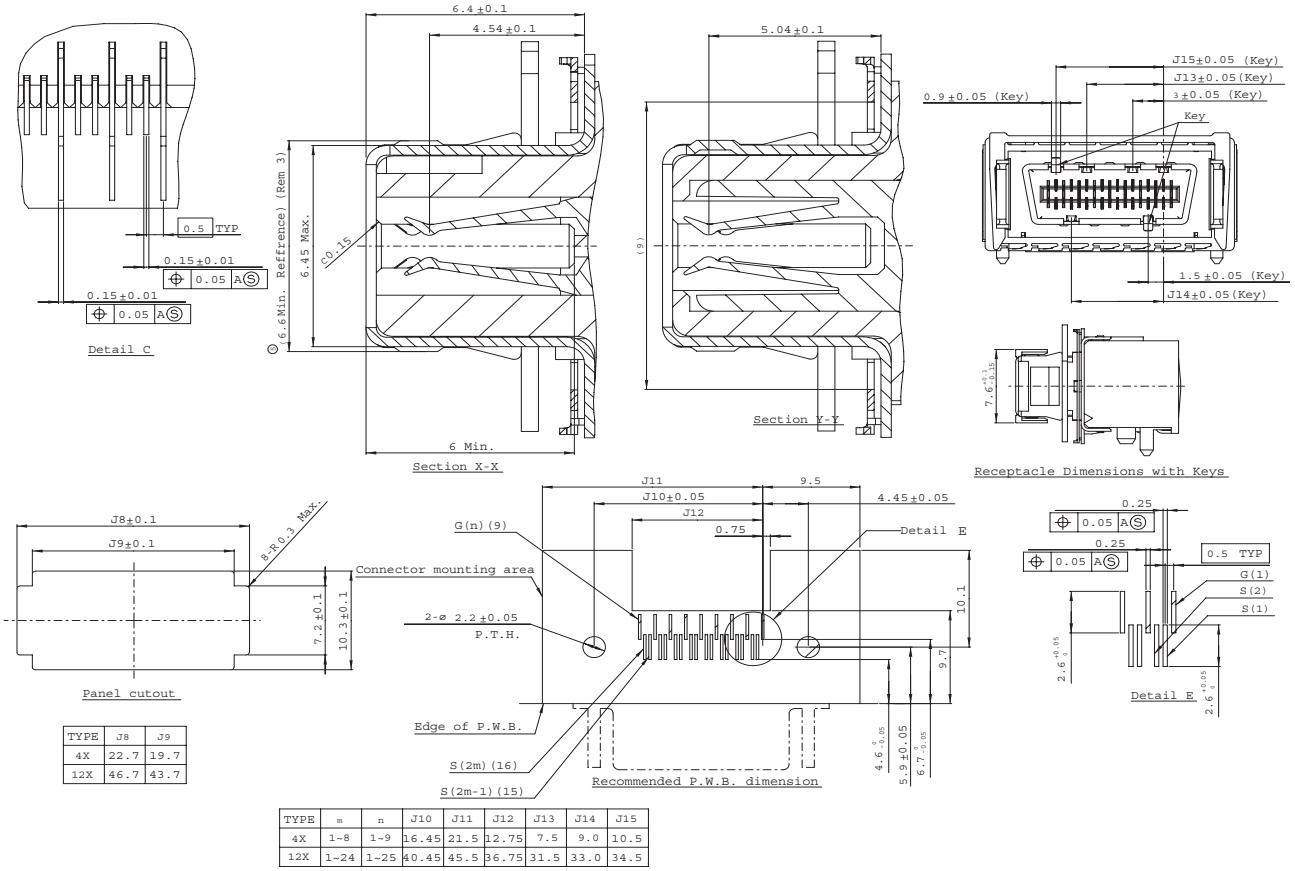
Pairs	m	n	p ②
8	1-8	1-9	1-25

Part number	P.W.B. thickness x	
FCN-268D008-G/0D	1.6	2.325 ②
FCN-268D008-G/0F	2.4	3.125 ②

Rem1) Unless otherwise specified,dimensions are in mm and tolerances are ±0.5mm.
Rem2) Coplanarity are 0.1mm

Figure 47 4x cable board connector

1. MICRO GIGACN is a trademark of Fujitsu Components Limited



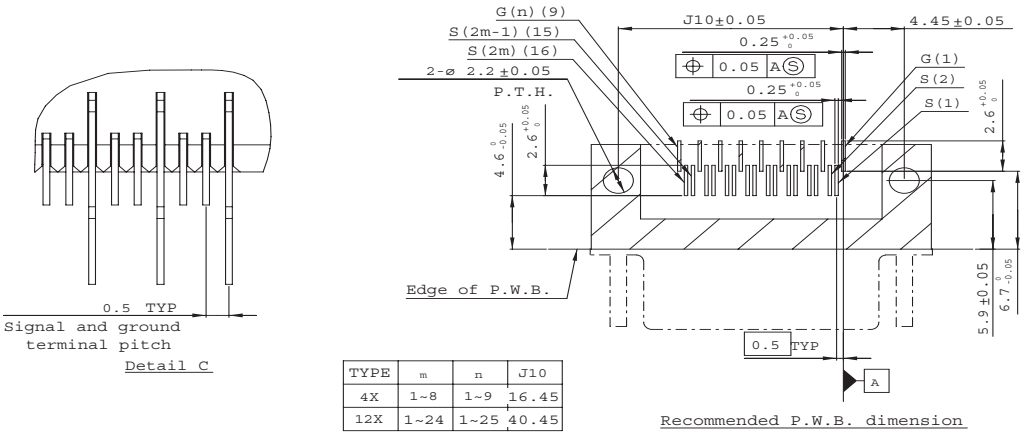
TYPE	J8	J9
4X	22.7	19.7
12X	46.7	43.7

TYPE	m	n	J10	J11	J12	J13	J14	J15
4X	1~8	1~9	16.45	21.5	12.75	7.5	9.0	10.5
12X	1~24	1~25	40.45	45.5	36.75	31.5	33.0	34.5

Figure 48 4x/12x cable board connector interface details

7.4.2 FOOTPRINT AND CONTACT PATTERN

The board footprint shown in Figure 49 should be used for the 4x cable connector shown in Figure 47. Connector pin numbers are indicated for information. The connector is surface-mounted, so there are no contacts on the secondary side of the board.



TYPE	m	n	J10
4X	1~8	1~9	16.45
12X	1~24	1~25	40.45

Figure 49 4x cable board connector footprint, top view

7.4.3 PIN ASSIGNMENT

C7-9: The pin assignment listed in Table 33 **shall** be used for the board connector for InfiniBand 4x cables.

The character 'x' in the signal symbol is the port number, as defined in Section 4.1, "Signal Naming Conventions," on page 69.

Table 33 4x board connector signal assignment

Pin Number	Signal
G1-G9	Signal Ground
S1	IBtxIp(0)
S2	IBtxIn(0)
S3	IBtxIp(1)

Table 33 4x board connector signal assignment

Pin Number	Signal
S4	IBtxIn(1)
S5	IBtxIp(2)
S6	IBtxIn(2)
S7	IBtxIp(3)
S8	IBtxIn(3)
S9	IBtxOn(3)
S10	IBtxOp(3)
S11	IBtxOn(2)
S12	IBtxOp(2)
S13	IBtxOn(1)
S14	IBtxOp(1)
S15	IBtxOn(0)
S16	IBtxOp(0)
Housing	Chassis Ground

C7-10: Signal Ground **shall** be connected to **IB_Sh_Ret** on the module.

C7-11: Signal Ground **shall not** be connected to Chassis Ground in the connector. See Section 9.5.4.

C7-11.1.1: A continuous ground path from the cable's inner shield(s) through the connector to the board signal ground **shall** be provided to insure low jitter, low crosstalk and EMI containment.

A cable constructed with only a bulk shield is not likely to meet the electrical requirements of Section 6.3.4.

7.4.4 KEYING

It is recommended that the 4x board and copper cable connectors as used for InfiniBand use keys in positions 4 and 5, shown in Figure 52 as viewed from the outside of the chassis. This is to prevent misplugging with other interfaces that might have chosen to use the same physical connector.

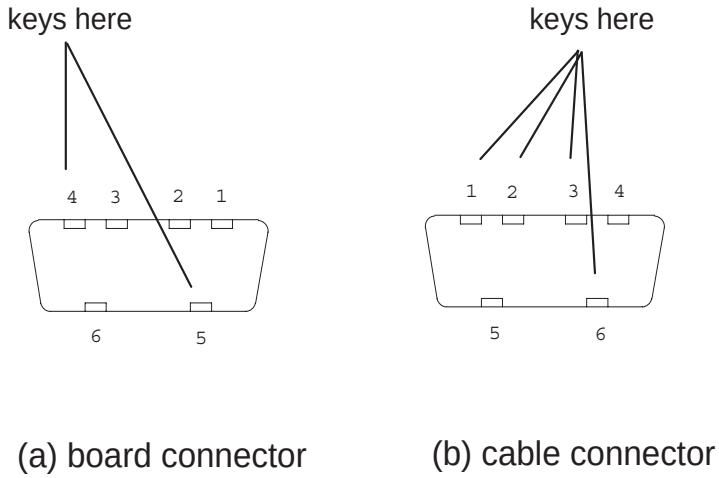


Figure 50 4x board and cable connector keying

7.5 12X AND 8X INTERFACE BOARD CONNECTOR

This section defines the connector for the 12x and 8x cable interface on InfiniBand boards. The 12x interface is the widest InfiniBand link, providing for simultaneous transmit and receive of twelve bits of encoded differential data.

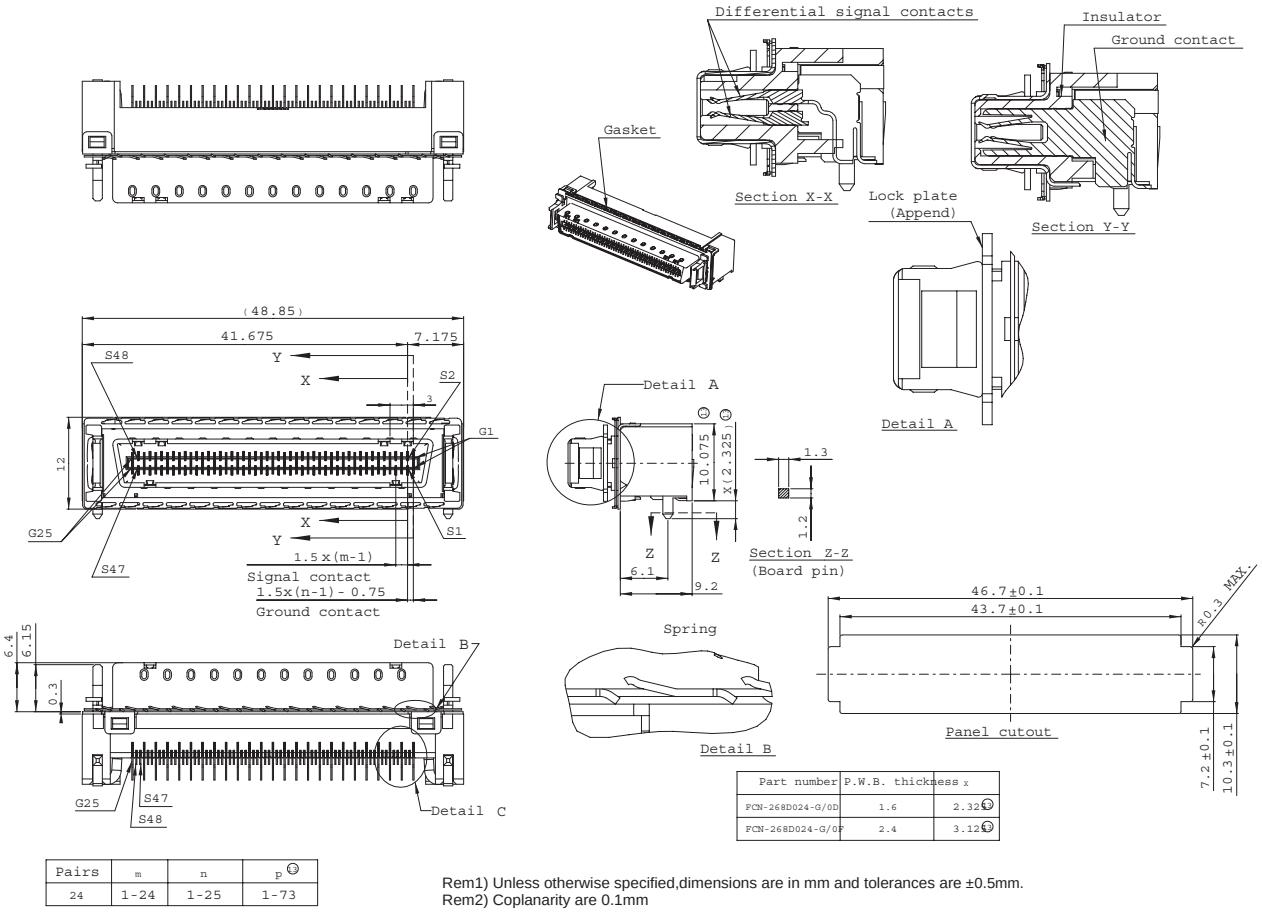
7.5.1 DESCRIPTION

This connector uses pairs of contacts separated by Ground contacts to reduce near end crosstalk (NEXT). The connector is a surface-mounted design similar to that of the 4x connector, and all signal pins are used. A suitable receptacle, referred to as a 24 pair MICRO GIGACN receptacle, is available from Fujitsu Components Ltd. and others, and is shown in Figure 51. Detailed drawings of mating interface dimensions are shown in Figure 48.

The 8x cable interface uses the same board connector as the 12x interface described in Section 7.5 on page 225, but with a subset of the pins used.

C7-12: All 12x InfiniBand cable plugs **shall** be intermateable with this connector.

C7-12.2.1: All 8x InfiniBand cable plugs **shall** be intermateable with this connector.



Part number	P.W.B. thickness x
FCN-268D024-G/0D	1.6
FCN-268D024-G/0F	2.4

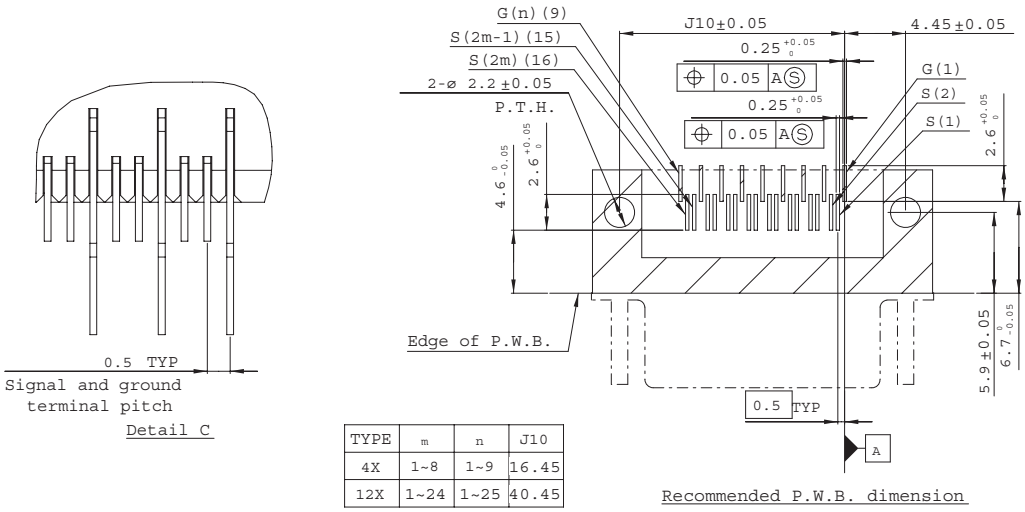
Pairs	m	n	p
24	1-24	1-25	1-73

Rem1) Unless otherwise specified,dimensions are in mm and tolerances are ±0.5mm.
Rem2) Coplanarity are 0.1mm

Figure 51 12x cable board connector

7.5.2 FOOTPRINT AND CONTACT PATTERN

The board footprint shown in Figure 52 should be used for the 12x cable board connector shown in Figure 51. Connector pin numbers are indicated for information. The connector is surface-mounted, so there are no contacts on the secondary side of the board.



TYPE	m	n	J10
4X	1~8	1~9	16.45
12X	1~24	1~25	40.45

Figure 52 12x cable board connector footprint, top view

7.5.3 PIN ASSIGNMENT FOR 12X OPERATION

C7-13: The pin assignment listed in Table 34 **shall** be used for the board connector for InfiniBand 12x cables.

The character 'x' in the signal symbol is the port number, as defined in Section 4.1, "Signal Naming Conventions," on page 69.

Table 34 12x board connector signal assignment

Pin Number	Signal
G1-G25	Signal Ground
S1	IBtxIp(0)
S2	IBtxIn(0)
S3	IBtxIp(1)
S4	IBtxIn(1)
S5	IBtxIp(2)
S6	IBtxIn(2)

Table 34 12x board connector signal assignment

Pin Number	Signal
S7	IBtxIp(3)
S8	IBtxIn(3)
S9	IBtxIp(4)
S10	IBtxIn(4)
S11	IBtxIp(5)
S12	IBtxIn(5)
S13	IBtxIp(6)
S14	IBtxIn(6)
S15	IBtxIp(7)
S16	IBtxIn(7)
S17	IBtxIp(8)
S18	IBtxIn(8)
S19	IBtxIp(9)
S20	IBtxIn(9)
S21	IBtxIp(10)
S22	IBtxIn(10)
S23	IBtxIp(11)
S24	IBtxIn(11)
S25	IBtxOn(11)
S26	IBtxOp(11)
S27	IBtxOn(10)
S28	IBtxOp(10)
S29	IBtxOn(9)
S30	IBtxOp(9)
S31	IBtxOn(8)
S32	IBtxOp(8)
S33	IBtxOn(7)
S34	IBtxOp(7)
S35	IBtxOn(6)
S36	IBtxOp(6)

IBTA

Table 34 12x board connector signal assignment

Pin Number	Signal
S37	IBtxOn(5)
S38	IBtxOp(5)
S39	IBtxOn(4)
S40	IBtxOp(4)
S41	IBtxOn(3)
S42	IBtxOp(3)
S43	IBtxOn(2)
S44	IBtxOp(2)
S45	IBtxOn(1)
S46	IBtxOp(1)
S47	IBtxOn(0)
S48	IBtxOp(0)
Housing	Chassis Ground

C7-14: Signal Ground **shall** be connected to **IB_Sh_Ret** on the module.

C7-15: Signal Ground **shall not** be connected to Chassis Ground in the connector.

C7-15.1.1: A continuous ground path from the cable's inner shield(s) through the connector to the board signal ground **shall** be provided to insure low jitter, low crosstalk and EMI containment.

A cable constructed with only a bulk shield is not likely to meet the electrical requirements of Section 6.3.4.

7.5.4 PIN ASSIGNMENT FOR 8X OPERATION

o7-15.2.1: The pin assignment listed in Table 34 **shall** be used for the board connector for InfiniBand 8x cables.

The character 'x' in the signal symbol is the port number, as defined in Section 4.1, "Signal Naming Conventions," on page 69.

Table 35 8x board connector signal assignment

Pin Number	Signal
G1-G25	Signal Ground
S1	IBtxIp(0)
S2	IBtxIn(0)
S3	IBtxIp(1)
S4	IBtxIn(1)
S5	IBtxIp(2)
S6	IBtxIn(2)
S7	IBtxIp(3)
S8	IBtxIn(3)
S9	IBtxIp(4)
S10	IBtxIn(4)
S11	IBtxIp(5)
S12	IBtxIn(5)
S13	IBtxIp(6)
S14	IBtxIn(6)
S15	IBtxIp(7)
S16	IBtxIn(7)
S17	reserved
S18	reserved
S19	reserved
S20	reserved
S21	reserved
S22	reserved
S23	reserved
S24	reserved
S25	reserved
S26	reserved
S27	reserved
S28	reserved

IBTA

Table 35 8x board connector signal assignment

Pin Number	Signal
S29	reserved
S30	reserved
S31	reserved
S32	reserved
S33	IBtxOn(7)
S34	IBtxOp(7)
S35	IBtxOn(6)
S36	IBtxOp(6)
S37	IBtxOn(5)
S38	IBtxOp(5)
S39	IBtxOn(4)
S40	IBtxOp(4)
S41	IBtxOn(3)
S42	IBtxOp(3)
S43	IBtxOn(2)
S44	IBtxOp(2)
S45	IBtxOn(1)
S46	IBtxOp(1)
S47	IBtxOn(0)
S48	IBtxOp(0)
Housing	Chassis Ground

7.5.5 KEYING

It is recommended that the 12x board and copper cable connectors as used for InfiniBand use keys in positions 4 and 5, shown in Figure 53 as viewed from the outside of the chassis. This is to prevent misplugging with other interfaces that might have chosen to use the same physical connector.

It is recommended that the 8x board and copper cable connectors as used for InfiniBand use the same keying as that for a 12x cable.

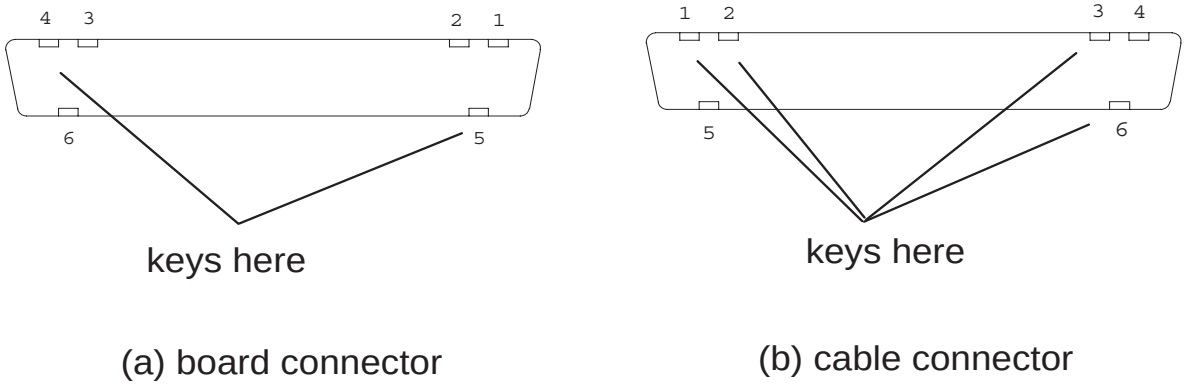


Figure 53 12x board and cable connector keying

7.6 PLUGGABLE INTERFACE CONNECTORS

Unlike the 1x, 4x, 8x, and 12x connectors described in the preceding sections, a pluggable interface port is designed to be used with either copper or optical cables. Due to the need to account for the jitter introduced in the assembly between the board and the cable, a new port type, referred to as port type 2, is defined for these devices. The unique electrical requirements for port type 2 are listed in Section 6.4. In the case of the copper pluggable devices, the transceiver may be part of the cable assembly that is mated with the port connector on the board. Only those cable types listed in Table 36 **shall** be supported.

The active devices contain circuitry to compensate for the jitter introduced in the cable assembly. The active optical device external interface is defined in Section 8.10. The passive copper device is permanently attached to the cable, so the port type 2 interface to the board is the only separable interface.

C7-15.1.2: Passive copper pluggable cables **shall** meet the requirements listed in Section 6.4.

Table 36 1x Pluggable interface supported device types

Pluggable device type	Board port type	External port type	Cable connector
Active copper	2	1	1x Copper (HSSDC2)

Table 36 1x Pluggable interface supported device types

Pluggable device type	Board port type	External port type	Cable connector
Active optical	2	**See Section 8.10**	**See Section 8.10**
Passive copper (captive)	2	None (see text)	None (see text)

7.6.1 1X PLUGGABLE INTERFACE

7.6.1.1 DESCRIPTION

The interface connector used on InfiniBand boards for 1x pluggable copper and optical ports is the same one defined for Small Form Factor Pluggables, as specified in SFF-8074 for Fibre Channel applications. The receptacle is a surface-mounted design with twenty contacts; eleven of the pins are signal contacts, two are power contacts, and seven contacts are connected to signal Ground. A separate metal cage is utilized on the board to minimize EMI radiation. The connector and cage are shown in Figure 54. The example shown in the figure is for a device in which the cable is separable from the transceiver, as in the case of the active copper device.

Note

Not all devices that will physically fit in the pluggable interface port are supported by this specification. See Table 36 for the supported plug-gable device types.

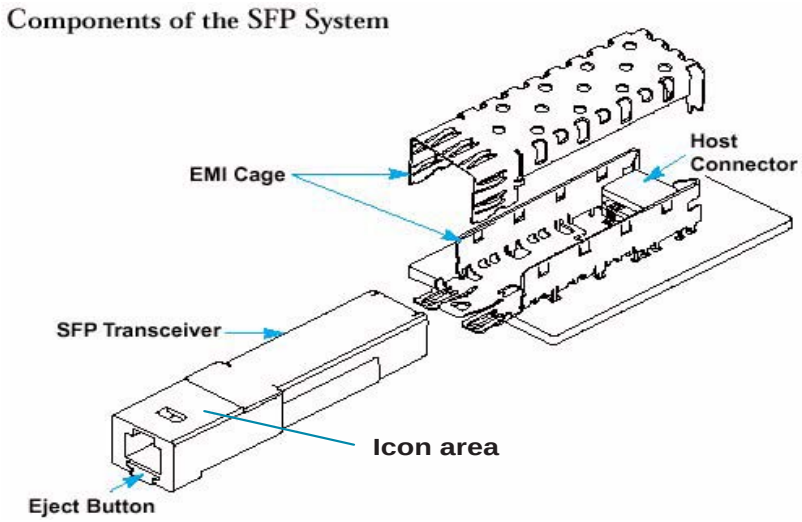


Figure 54 1x Pluggable interface and transceiver assembly

7.6.1.2 PHYSICAL REQUIREMENTS

The physical requirements for the pluggable interface are not common to those listed in Table 29.

C7-15.1.3: Connectors to be used for pluggable ports on InfiniBand modules **shall** meet or exceed the physical and mechanical performance requirements listed in Table 37, which are common to those in SFF-8074.

Table 37 Pluggable interface physical requirements

Symbol	Parameter	Minimum	Maximum	Units	Conditions/Comments
F_i	Insertion force	0	40	N	
F_w	Withdrawal force	0	11.5	N	
F_{rx}	Transceiver retention force	90	170	N	no damage to transceiver below 90 N
Frcl	Cage retention (latch strength)	180		N	no damage to latch below 180 N
F_{ck}	Cage kickout spring force	11.5	22	N	
N_{hc}	Durability, host connector and cage	100		cycles	

Table 37 Pluggable interface physical requirements

Symbol	Parameter	Minimum	Maximum	Units	Conditions/Comments
N_x	Durability, transceiver	50		cycles	

7.6.1.3 ELECTRICAL REQUIREMENTS

C7-15.1.4: The pluggable interface **shall** meet the additional electrical performance requirements listed in Section 6.4 for port type 2, as specified at the board contacts of the host connector.

7.6.1.4 FOOTPRINT AND CONTACT PATTERN

The board footprint and contact pattern for the pluggable interface is shown in Figure 55. The I/O plate panel cutout is shown in Figure 56.

IBTA

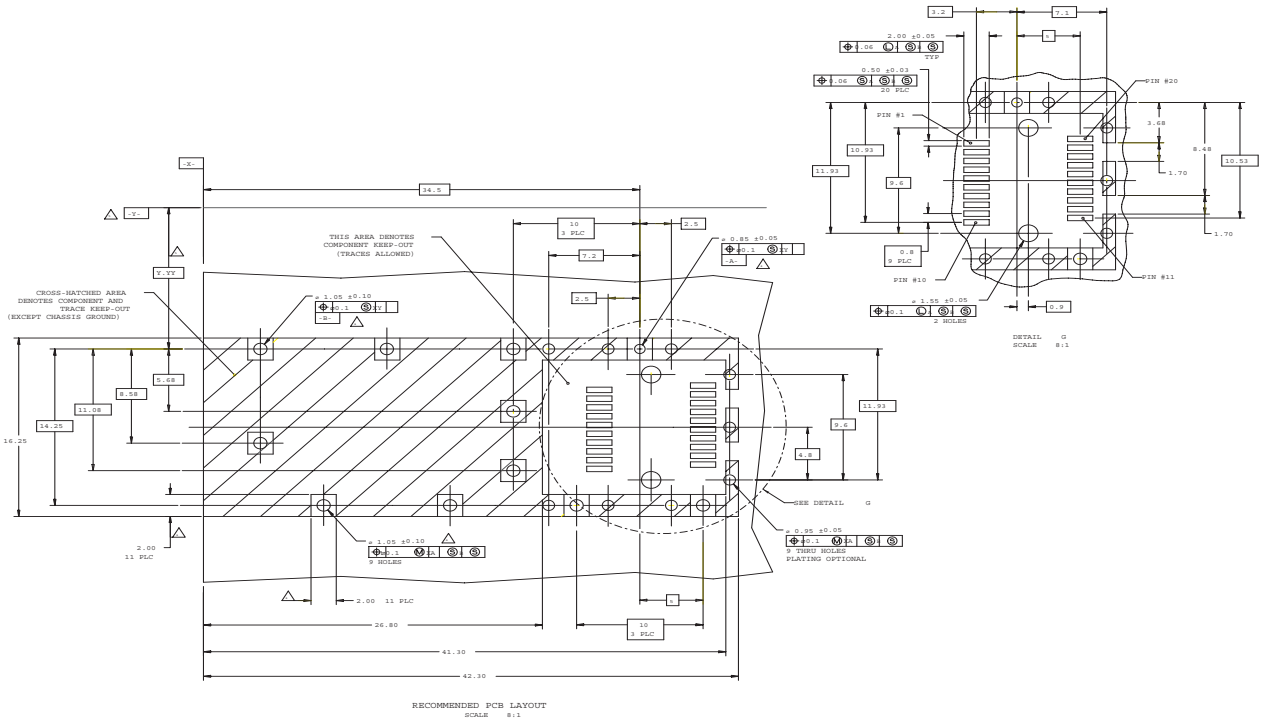


Figure 55 Pluggable interface board connector footprint

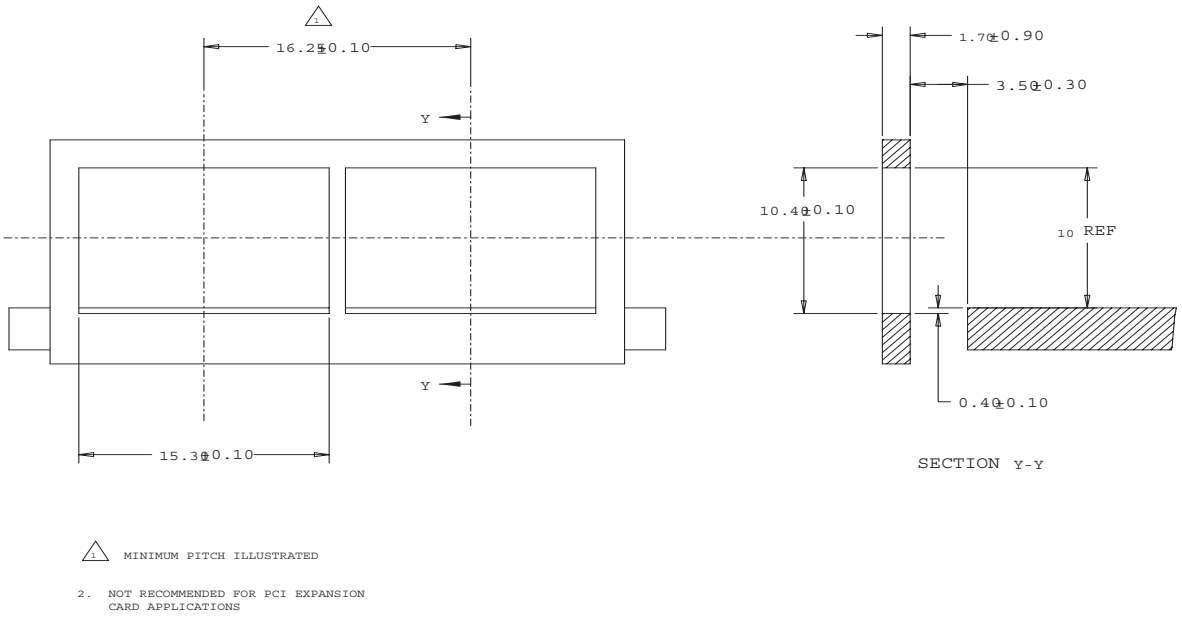


Figure 56 Pluggable interface I/O plate cutout

7.6.1.5 PIN ASSIGNMENTS

C7-15.1.5: The pin assignment shown in Table 38 **shall** be used for the 1x pluggable interface.

The reader is referred to SFF-8074 for the detailed descriptions of the pin functions.

Table 38 Pluggable interface pin assignment

Pin number	Signal	Description
1	VeeT	Transmit Ground
2	TxFault	Transmit Fault indication
3	TxDisable	Transmitter Disable (active high)
4	MOD-DEF(2)	Module Definition 2, 2-wire serial interface
5	MOD-DEF(1)	Module Definition 1, 2-wire serial interface

Table 38 Pluggable interface pin assignment

Pin number	Signal	Description
6	MOD-DEF(0)	Module Definition 0, internally Grounded in module
7	Rate Select	Selects receiver bandwidth; low or open = reduced, high = full
8	LOS	Loss of Signal (active high)
9	VeeR	Receive Ground
10	VeeR	Receive Ground
11	VeeR	Receive Ground
12	IBtxIn(0)	- Received data
13	IBtxIp(0)	+ Received data
14	VeeR	Receive Ground
15	VccR	Receiver power, 3.3 V ±5%
16	VccT	Transmitter power, 3.3 V ±5%
17	VeeT	Transmit Ground
18	IBtxOp(0)	+ Transmit data
19	IBtxOn(0)	- Transmit data
20	VeeT	Transmit Ground

7.6.2 4X PLUGGABLE INTERFACE

This section specifies the 4x Single Data Rate (SDR) (2.5 GT/s) pluggable interface. The 4x Pluggable Interface is designed to allow a copper cable or optical transceiver to be attached to an InfiniBand HCA, TCA or Switch. The 4x Pluggable Interface supports Active Copper, Passive Copper, and Optical (4x-SX and 4x-LX) type devices.

7.6.2.1 DESCRIPTION

The 4x Pluggable is based in part on the XPAK MSA. The mechanical specifications are identical to XPAK, while the High Speed electrical and optical interface specifications are based on the InfiniBand specifications listed in the *InfiniBand Architecture Specification, Volume 2*, Chapters 6 and 8.

7.6.2.2 PHYSICAL REQUIREMENTS

The physical dimensions of the 4x Pluggable transceiver are shown in Figure 57 for reference. The reader is referred to the official XPAK MSA documentation for specific dimensions and function.

C7-15.2.1: 4X pluggable interface **shall** meet or exceed the physical, electrical and mechanical performance specifications of the official XPAK MSA documentation.

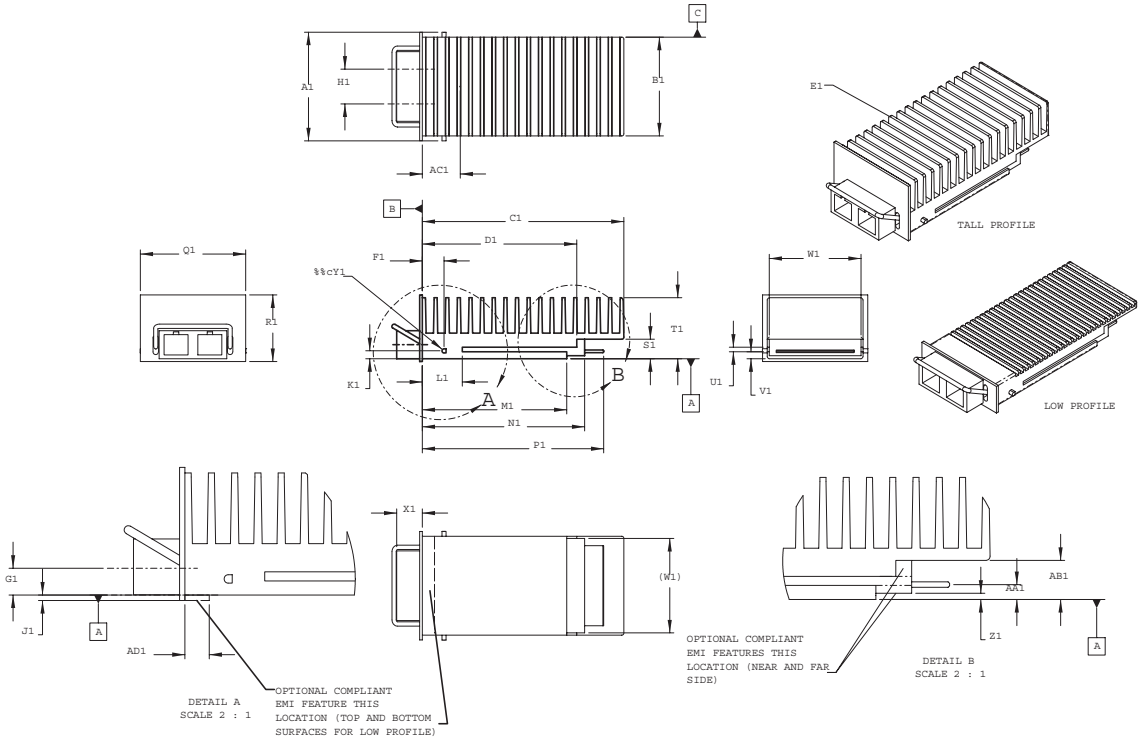


Figure 57 XPAK transceiver physical outline

7.6.2.3 FOOTPRINT AND CONTACT PATTERN

The host connector footprint used on the board **shall** be as shown in Figure 58.

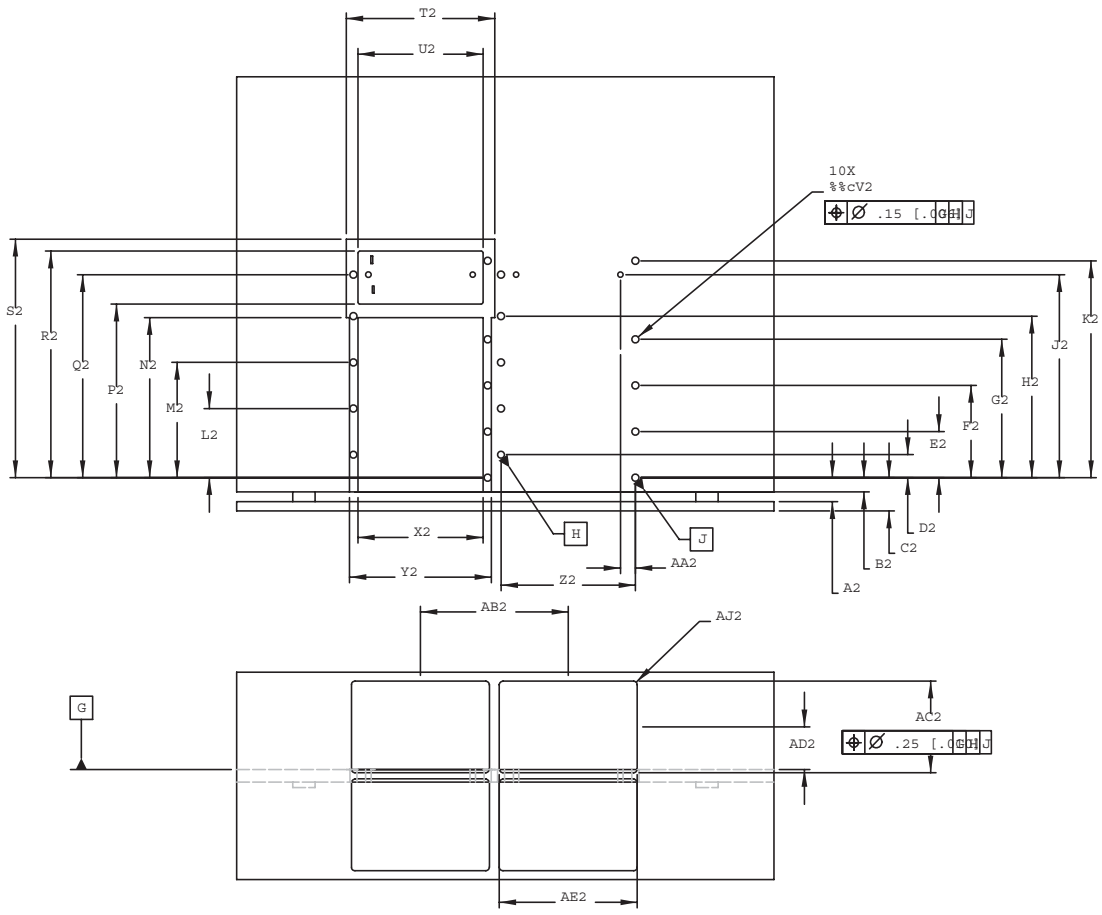


Figure 58 4x Pluggable host connector footprint

7.6.2.4 PIN ASSIGNMENTS

C7-15.2.2: The pin assignment shown in Table 39 **shall** be used for the for the 4x Pluggable interface.

Table 39 Pin Assignment for 4x Pluggable Interface

Pin No	Name	Pin No	Name
70	GND	1	GND
69	GND	2	GND
68	REF CLK-	3	GND
67	REF CLK+	4	5.0V
66	GND	5	3.3V
65	TX LANE3-	6	3.3V
64	TX LANE3+	7	APS
63	GND	8	APS
62	TX LANE2-	9	LASI
61	TX LANE2+	10	RESET
60	GND	11	VEND SPECIFIC
59	TX LANE1-	12	TX ON/OFF
58	TX LANE1+	13	TX LOCK ER#
57	GND	14	MOD DETECT
56	TX LANE0-	15	VEND SPECIFIC
55	TX LANE0+	16	VEND SPECIFIC
54	GND	17	MGMT I/F
53	GND	18	MGMT CLK
52	GND	19	MGMT Mode
51	RX LANE3-	20	PRTAD3
50	RX LANE3+	21	PRTAD2
49	GND	22	PRTAD1
48	RX LANE2-	23	PRTAD0
47	RX LANE2+	24	MODE XAUI/SFI4_P2
46	GND	25	APS SET
45	RX LANE1-	26	RX LOCK ER#
44	RX LANE1+	27	APS SENSE
43	GND	28	APS
42	RX LANE0-	29	APS
41	RX LANE0+	30	3.3V
40	GND	31	3.3V
39	SRCCLK-	32	5.0V
38	SRCCLK+	33	GND
37	GND	34	GND
36	GND	35	GND

7.6.2.5 REGISTER SET

The register map for InfiniBand is identical to XPAK 2.31 and XENPAK 3.0, with the following exceptions:

- The XPAK register OUI, in register 32818 (decimal) 0x8032 (hexadecimal) is 0x000ACB (hexadecimal) (this is different from XENPAK which is 0x0008BE).
- The XPAK register package type mask, in register 32768 (decimal), 0x8012 (hexadecimal) is 00000100 (binary), 4 (decimal), 0x04 (hexadecimal) (this is different from XENPAK which is 0x01)

In addition, the following extensions to the XENPAK MSA NVM register set is defined for InfiniBand™ 4x pluggable devices that conform to the XPAK MSA form factor.

Table 40 InfiniBand register additions for 4x Pluggable

Field	Dec	Hex	NVRAM Byte	Size	Name	Description	Register value, hex	Interpretation
Basic	32787	8013	12	1	Connector	Optical connector	40	MPO/MTP
Basic	32791	8017	16	1	Protocol	Protocol type	20	IB 1x
Basic	32791	8017	16	1	Protocol	Protocol type	40	IB 4x
Basic	32801	8021	26	1	Standards Compliance Code	IB Optical Distance/10GFC Code Byte 3	10	IB LX
Basic	32801	8021	26	1	Standards Compliance Code	IB Optical Distance/10GFC Code Byte 3	20	IB SX
Basic	32801	8021	26	1	Standards Compliance Code	IB Copper type/10GFC Code Byte 3	40	IB active copper
Basic	32801	8021	26	1	Standards Compliance Code	IB Copper type/10GFC Code Byte 3	80	IB passive copper

7.6.2.6 HIGH SPEED SIGNALING

7.6.2.6.1 GENERAL REQUIREMENTS

C7-15.2.3: The high-speed signaling interface to the 4x Pluggable **shall** comply with the requirements listed in the *InfiniBand Architecture Specification, Volume 2*, Chapter 6:

7.6.2.6.2 ELECTRICAL SPECIFICATIONS

The electrical interface, including pinout, power supplies and management functionality, is specified by XENPAK MSA v3.0 with the exceptions listed below. The pinout for SFI4-P2 implementation is identical to XENPAK MSA v3.0, and is shown for reference only.

7.6.2.6.3 DIFFERENTIAL DRIVER OUTPUTS

C7-15.2.4: All SDR output ports for the 4x Pluggable interface **shall** comply with the parameters and notes of Table 17 Driver Characteristics for 2.5 Gb/s. Parameters apply to the pin type as noted. The parameters are defined in terms of values at ASIC pins. They may be measured at accessible test points, with the values adjusted appropriately as defined in Section 6.7, "Compliance Points," on page 201.

7.6.2.6.4 DIFFERENTIAL RECEIVER INPUTS

C7-15.2.5: All SDR input ports for the 4x Pluggable interface **shall** comply with the parameters and notes of Table 20 Receiver Characteristics for 2.5 Gb/s. Parameters apply to the pin type as noted. The parameters are defined in terms of values at ASIC pins. They may be measured at accessible test points, with the values adjusted appropriately as defined in Section 6.7, "Compliance Points," on page 201.

7.6.2.6.5 LOSS MODEL

C7-15.2.6: The maximum loss from the IC package to the Pluggable Device Connector is 1 dB.

7.6.2.6.6 COMPLIANCE POINTS

C7-15.2.7: 1x Pluggable Device Ports shall comply with the amplitude and eye opening at TP9 and TP10 as described in Table 24 2.5 Gb/s Signal Test Points. Measurement may require de-embedding.

7.6.3 4X OPTICAL PLUGGABLE DEVICES

C7-15.2.8: All 4x SDR Optical Pluggable devices **shall** comply with requirements in its respective 4x optical distance classification (SX or LX) in the *InfiniBand Architecture Specification, Volume 2*, Chapter 8: Fiber Attachment - 2.5 Gb/s, 5.0 Gb/s, & 10 Gb/s.

7.7 CABLES

This section defines the characteristics of InfiniBand cables. It is the intent of this specification to allow for and encourage innovation in the market. Therefore, the emphasis is on interoperability and minimum requirements needed to insure functionality without restricting the means of implementation. Specific bulk wire conductor size, insulation materials, etc. are explicitly not specified herein, allowing suppliers to determine the optimum design for a given application.

7.7.1 PHYSICAL REQUIREMENTS

It is recommended that cable assemblies to be used for InfiniBand products have a minimum bend radius of no more than 100 mm (4 inches), as shown in Figure 59. The bend radius is the radius to which the cable can

BETA

be bent while continuing to meet the electrical characteristics as defined in Table 41.

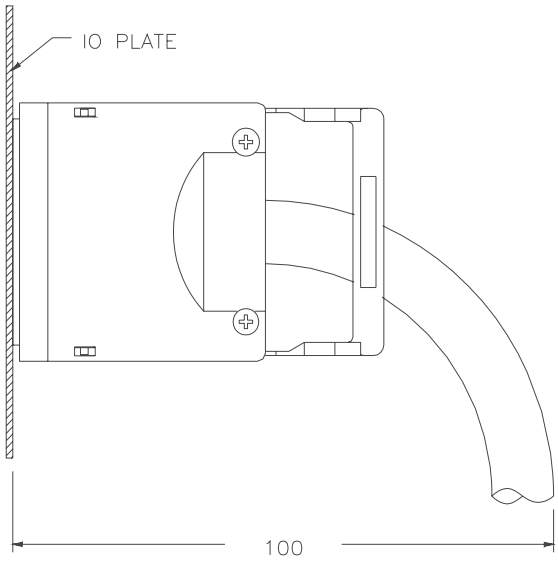


Figure 59 Cable assembly bend radius

7.7.2 ELECTRICAL REQUIREMENTS

C7-16: Cable assemblies to be used for InfiniBand **shall** meet the electrical requirements listed in Table 41 for all link widths for port type 1.

The sections of this specification that deal with the specific link widths may have additional requirements that **shall** also be met for that specific interface.

Notes

1. Although the cable differential impedance specification listed in Table 41 covers a relatively wide range, the board connector and wiring are specified at 100 Ohms nominal. The use of cables with impedance significantly different from 100 Ohms (due to higher wire impedance or equalizer circuitry, or both) will result in losses due to reflections at the impedance mismatches. These losses may be significant unless there is sufficient attenuation in the cable to reduce them to an acceptable level.

2. It is left to the cable designer to evaluate the trade-off between reflection loss and length in the cable design to insure reliable operation. Cables less than some minimum length may not operate reliably outside the range of 90-110 Ohms differential impedance, due to insufficient attenuation of these signal reflections. The signal eye pattern at the cable output must conform to the electrical requirements as listed in Table 41 and shown in Figure 60 in all cases.

IBTA

Table 41 Cable Assembly Electrical Requirements

Symbol	Parameter	Minimum	Maximum	Units	Conditions/Comments
Z_{dca}(nom)	Differential Impedance, nominal value	95	155	Ω	per EIA 364-108, measured over the length of each signal pair, from the unequalized end (if equalizer is used)
Z_{dca}(peak)	Differential Impedance, deviation from nominal value		5	Ω	measured over the length of each signal pair
L_{ca}	Insertion loss		10	dB	mated cable assembly and board connector, at frequencies up to 1.25 GHz, per EIA 364-101
S_{cal}	Pair to pair skew		500	ps	see Table 28 Bit to Bit Skew Parameter Maximum Values
J_{ca}	Jitter		0.25	UI	per EIA 364-107 and IEEE 191-2003, with 1 V differential Fibre Channel CJTPAT stimulus until 10000 hits in max. 20 mV high jitter box or equivalent, with equipment and fixture contribution de-embedded; worst case pair, with three adjacent pairs on one side of that pair (if they exist) to be driven at the same end of the cable as the measurement is performed, by an asynchronous source with 1.5 V unsigned differential amplitude and transition time of 100 ps or less at the board pins. The far end of each adjacent pair should be terminated in 50 Ohms to Ground. A PRBS generator or other source may be used for driving the adjacent pairs, at a minimum of 1.3 Gb/s. See Figure 60.
V_{Cout}	Eye height (voltage)	196		mV	minimum unsigned differential amplitude for 1 V differential Fibre Channel CJTPAT stimulus until 10000 hits in max. 20 mV high jitter box or equivalent, measured at board connector pins under the same conditions as J_{ca} (see above).See Figure 60. For information, 316 mV if measured without crosstalk.
T_{Reye}	Eye width (time)	0.75		UI	minimum time opening for 1 V differential Fibre Channel CJTPAT stimulus until 10000 hits in max. 20 mV high jitter box or equivalent, measured at board connector pins under the same conditions as J_{ca} (see above). See Figure 60.

The signal eye pattern at the cable output must conform to the electrical requirements as listed in Table 41 and shown in Figure 60 in all cases.

IBTA

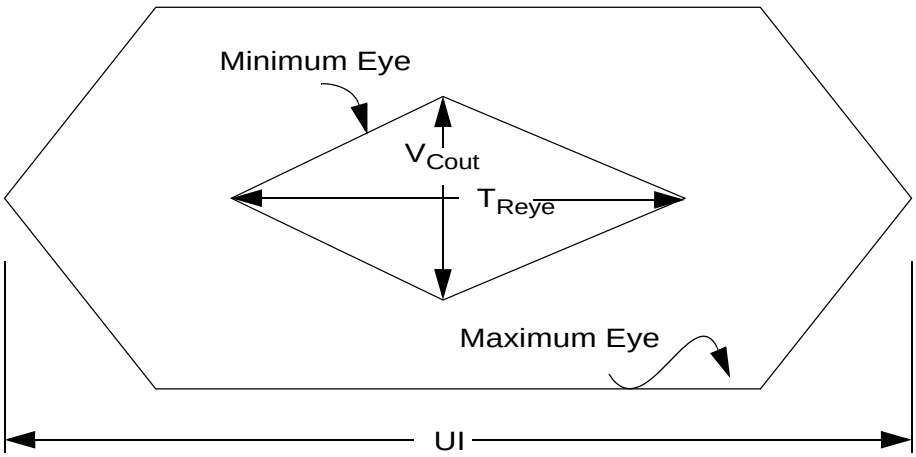


Figure 60 Eye Opening at receiving board connector pins (differential)

DRAFT

Note

It is recommended that within pair skew be no more than 120 ps total for any given pair in a cable assembly. Within pair skew may cause excessive jitter or induce sufficient shield current to exceed EMI radiation limits. It is expected that shield currents over 1 μA may be sufficient to cause radiation problems.

Within pair skew is a parameter that is easily measured at the cable assembly level, and is often useful if the eye parameters (opening and jitter) do not meet specification. The eye parameters are the critical test for functionality, hence within pair skew is primarily included here as guidance for diagnostic purposes.

7.7.3 EQUALIZATION

Equalization in the cable, either fabricated as part of the bulk wire or in the form of discrete components in the cable assembly (e. g., on a printed circuit board inside the cable connector backshell), is permitted. However, the cable assembly **shall** be required to meet the electrical requirements defined in Table 41 and Figure 60.

Note

Signal reflections can occur when using equalized cable assemblies, due to the presence of equalizer components which add in series with the cable characteristic impedance and may nearly double the apparent impedance seen by the signal. Normally these reflections are harmless due to cable attenuation and the use of source and/or receiver matching. Implementers should be aware of the potential for reflections if less than ideal impedance matches exist at the various interfaces in the path. The signal eye pattern at the cable output must conform to the electrical requirements as listed in Table 41 and in Figure 60 in all cases.

7.7.4 CABLE SHIELD CONNECTIONS

7.7.4.1 INNER (SIGNAL PAIR OR QUAD) SHIELD

C7-17: The **IB_Sh_Ret** signals **shall** be connected to the cable inner shield(s) in the cable connector.

C7-18: A continuous ground path from the cable's inner shield(s) through the connector to the board signal ground **shall** be provided to insure low jitter, low crosstalk and EMI containment.

Note

The primary purpose of these connections is to provide for isolation of the differential signals from each other. These shields also help to insure that the desired impedance of the link is maintained.

A cable connector or receptacle constructed with only a bulk shield is not likely to meet the electrical requirements of Section 6.3.4.

7.7.4.2 OUTER (BULK) SHIELD

C7-19: The cable bulk shield (shield used over the outside of the collection of conductors) used in InfiniBand cables **shall** be directly connected to chassis ground at both ends.

This specification does not permit AC coupling of the bulk shield.

7.7.5 ACTIVE CABLES

This release of this specification describes Active Cables in Section 7.8, “Active Cables,” on page 264.

7.7.6 PORT LABELING

It is recommended that I/O cables be labeled to avoid confusion between InfiniBand ports and connections and those used for other incompatible interfaces. The recommended labeling is defined in Section 9.6.

7.7.7 1X INTERFACE CABLE

7.7.7.1 CABLE CONNECTOR DESCRIPTION

The cable plug to be used on InfiniBand 1x cables is a derivative of the plug used for the Fibre Channel standard. The design uses seven in-line contacts, and the backshell incorporates a top side latch.

C7-19.1.1: A metal backshell or other means which fully shields the connector **shall** be bonded to the cable bulk shield through a continuous 360 degree contact to minimize EMI (Electromagnetic Interference).

Two pairs of signals are used, one each for transmit and receive. The signal pairs are isolated by the use of a signal Ground pin between them to minimize near end crosstalk (NEXT). A suitable plug, referred to as HSSDC2 InfiniBand, is available from Tyco Electronics and others, and is shown in Figure 61. Detailed drawings of mating interface dimensions are shown in Figure 48.

C7-20: 1x Board receptacles used on InfiniBand modules **shall** be intermateable with this connector.

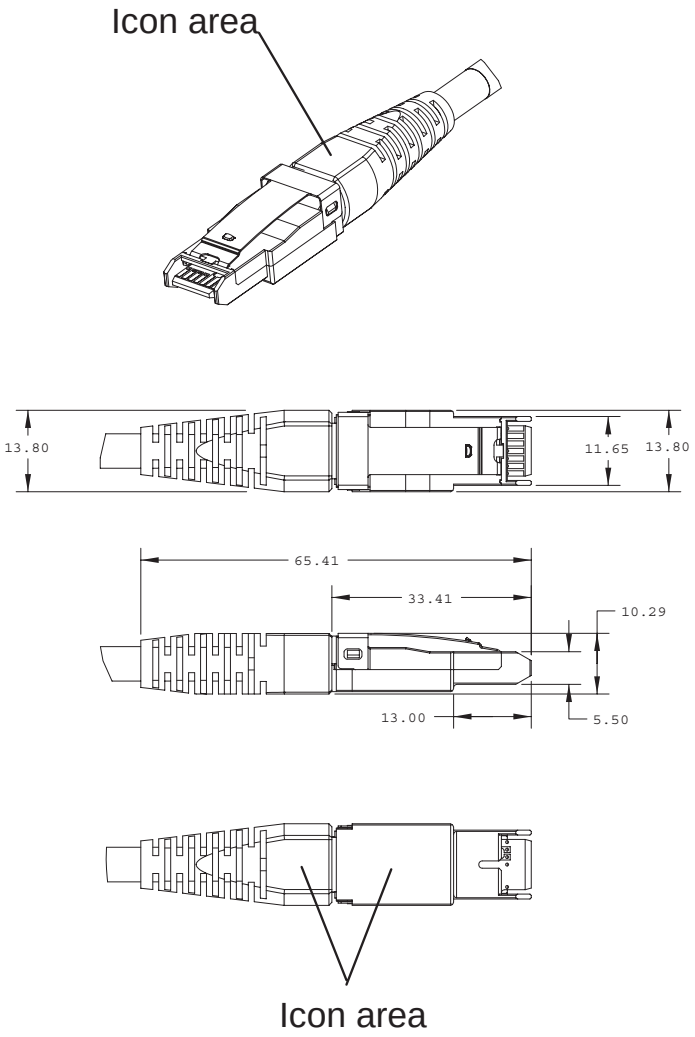


Figure 61 1x cable plug

7.7.7.2 PIN ASSIGNMENT

C7-21: The pin assignment listed in Table 42 **shall** be used for InfiniBand 1x cables.

There are no unused pins in the cable plug.

Note

Cable signal nets **shall** be connected from each *transmit* signal pair on the source port to the appropriate *receive* signal pair. For example, a cable from port y to port z **shall** be wired to connect IBtyOp(0) (positive output signal from port y, bit 0), to IBtzIp(0) (positive input to port z, bit 0) and IBtyOn(0) to IBtzIn(0), as shown in Table 42.

Table 42 1x cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
1	Signal Ground	7	Signal Ground
2	IBtxIp(0)	6	IBtxOp(0)
3	IBtxIn(0)	5	IBtxOn(0)
4	Signal Ground	4	Signal Ground
5	IBtxOn(0)	3	IBtxIn(0)
6	IBtxOp(0)	2	IBtxIp(0)
7	Signal Ground	1	Signal Ground
Housing	Chassis Ground	Housing	Chassis Ground

C7-22: Signal Ground **shall not** be connected to Chassis Ground in the cable or cable connector.

7.7.7.3 KEYING

The 1x cable plug includes a slot to accept a 1.0 mm bottom-side (opposite the latch side) key corresponding to the key in the receptacle, which is centered horizontally in the connector body.

7.7.8 4X CABLE

7.7.8.1 CABLE CONNECTOR DESCRIPTION

The cable plug to be used on InfiniBand 4x cables uses pairs of contacts interspersed with Ground contacts for crosstalk reduction. Eight pairs of signals are used, four each for transmit and receive. A suitable plug, referred to as an eight pair MICRO GIGACN plug, is available from Fujitsu

Components Component, Ltd. and others, and is shown in Figure 62. Detailed drawings of mating interface dimensions are shown in Figure 48.

C7-23: 4x Board receptacles used on InfiniBand modules **shall** be intermateable with this connector.

C7-23.1.1: A metal backshell or other means which fully shields the connector **shall** be bonded to the cable bulk shield through a continuous 360 degree contact to minimize EMI (Electromagnetic Interference).

Side latches are used in conjunction with receptacle features for retention, and are released by pulling on the "lanyard" handle shown in the drawing.

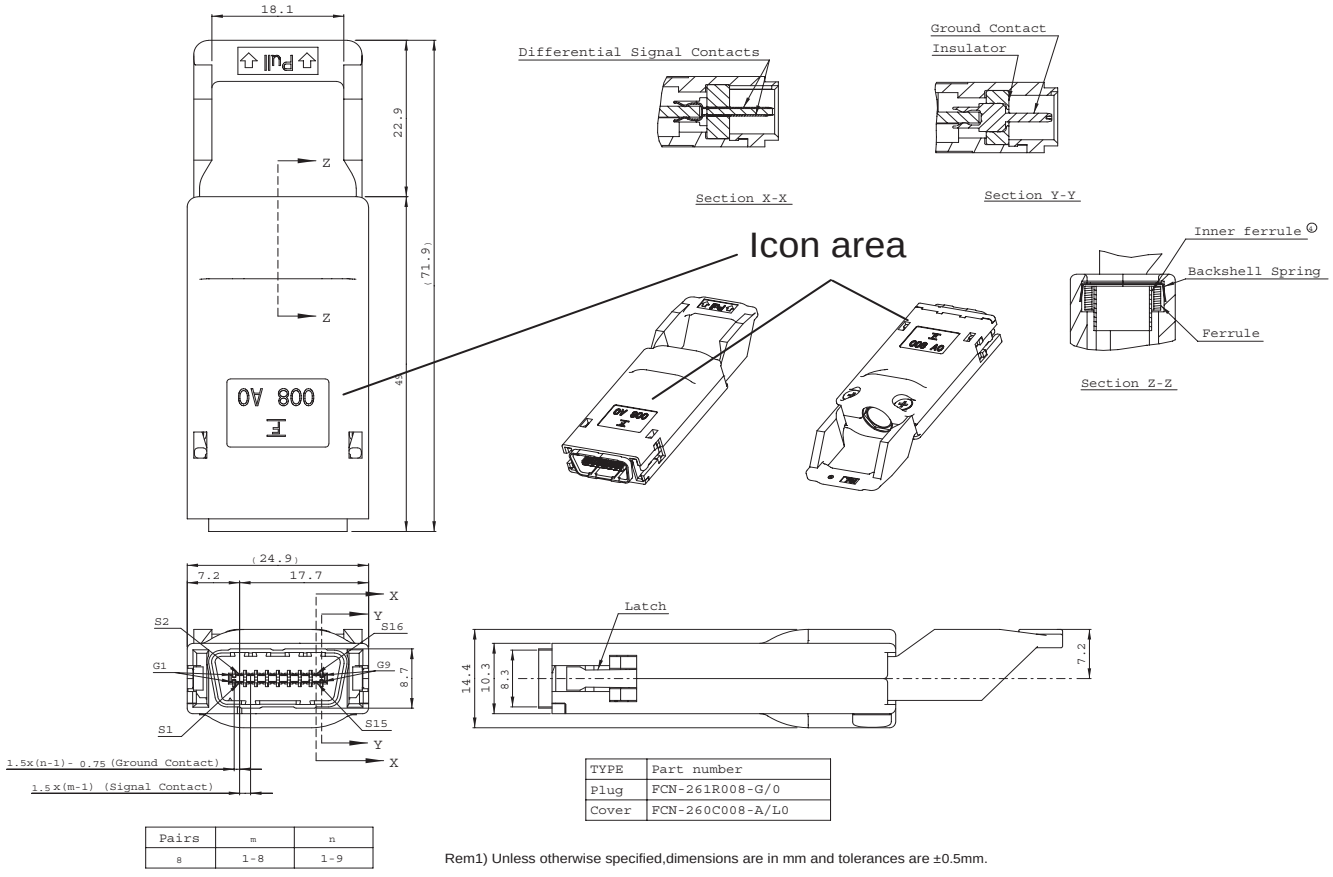
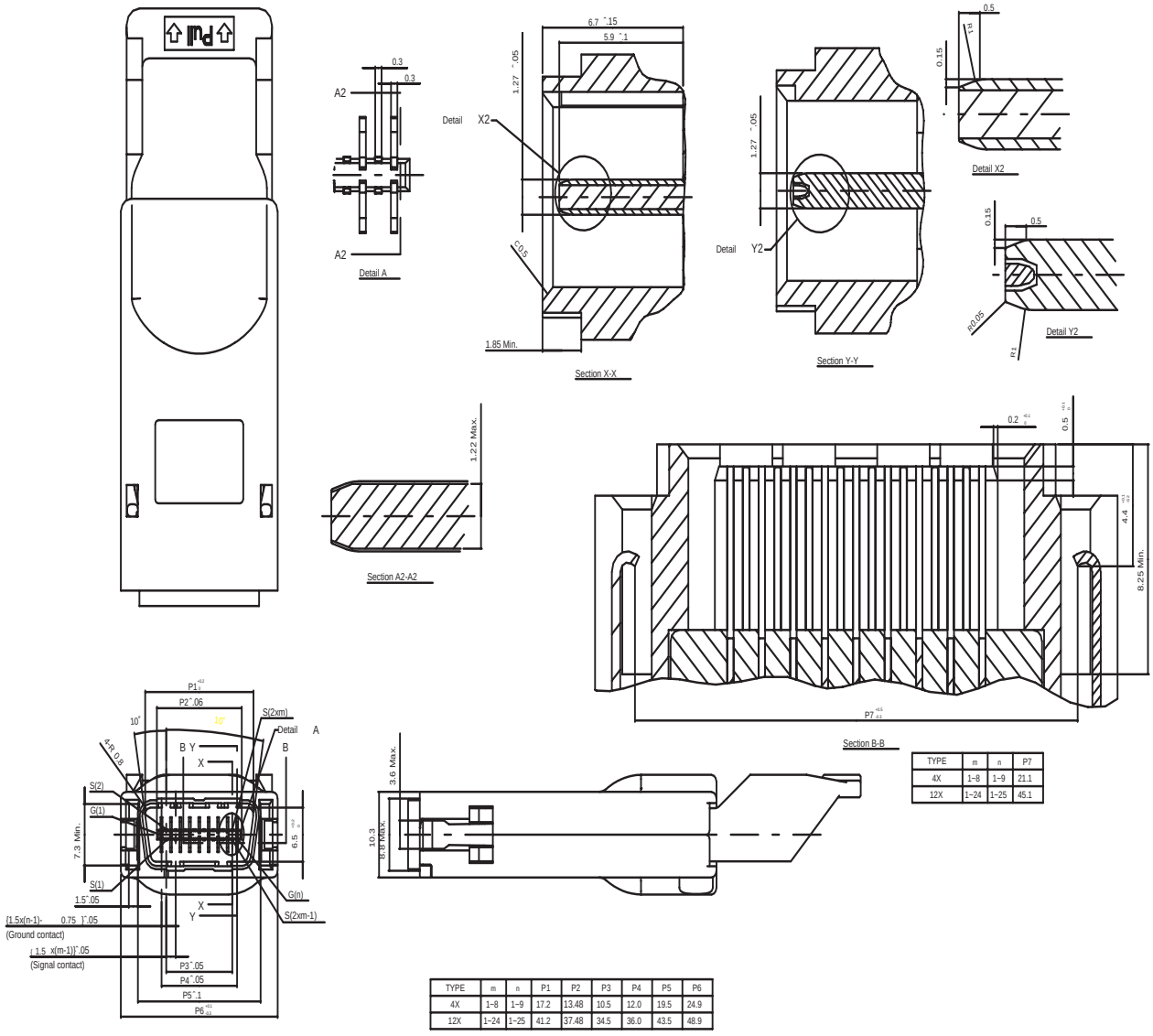


Figure 62 4x cable plug



TYPE	m	n	P1	P2	P3	P4	P5	P6
4X	1–8	1–9	17.2	13.48	10.5	12.0	19.5	24.9
12X	1–24	1–25	41.2	37.48	34.5	36.0	43.5	48.9

TYPE	m	n	P7
4X	1–8	1–9	21.1
12X	1–24	1–25	45.1

Figure 63 4x/12x cable plug interface details

7.7.8.2 PIN ASSIGNMENT

C7-24: The pin assignment listed in Table 43 **shall** be used for InfiniBand 4x cables.

There are no unused pins in the cable plug.

Note

Cable signal nets **shall** be connected from each *transmit* signal pair on the source port to the appropriate *receive* signal pair on the desired destination port. For example, a cable from port y to port z should be wired to connect IBtyOp(0) (positive output signal from port y, bit 0), to IBtzIp(0) (positive input to port z, bit 0) and IBtyOn(0) to IBtzIn(0), as shown in Table 43.

Table 43 4x cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
G1-G9	Signal Ground	G1-G9	Signal Ground
S1	IBtxIp(0)	S16	IBtxOp(0)
S2	IBtxIn(0)	S15	IBtxOn(0)
S3	IBtxIp(1)	S14	IBtxOp(1)
S4	IBtxIn(1)	S13	IBtxOn(1)
S5	IBtxIp(2)	S12	IBtxOp(2)
S6	IBtxIn(2)	S11	IBtxOn(2)
S7	IBtxIp(3)	S10	IBtxOp(3)
S8	IBtxIn(3)	S9	IBtxOn(3)
S9	IBtxOn(3)	S8	IBtxIn(3)
S10	IBtxOp(3)	S7	IBtxIp(3)
S11	IBtxOn(2)	S6	IBtxIn(2)
S12	IBtxOp(2)	S5	IBtxIp(2)
S13	IBtxOn(1)	S4	IBtxIn(1)
S14	IBtxOp(1)	S3	IBtxIp(1)
S15	IBtxOn(0)	S2	IBtxIn(0)
S16	IBtxOp(0)	S1	IBtxIp(0)
Housing	Chassis Ground	Housing	Chassis Ground

C7-25: Signal Ground **shall not** be connected to Chassis Ground in the cable or cable connector.

7.7.8.3 KEYING

It is recommended that the 4x copper cable connectors as used for InfiniBand use keys in positions 4 and 5, shown in Figure 52.

7.7.9 12X CABLE

7.7.9.1 CONNECTOR DESCRIPTION

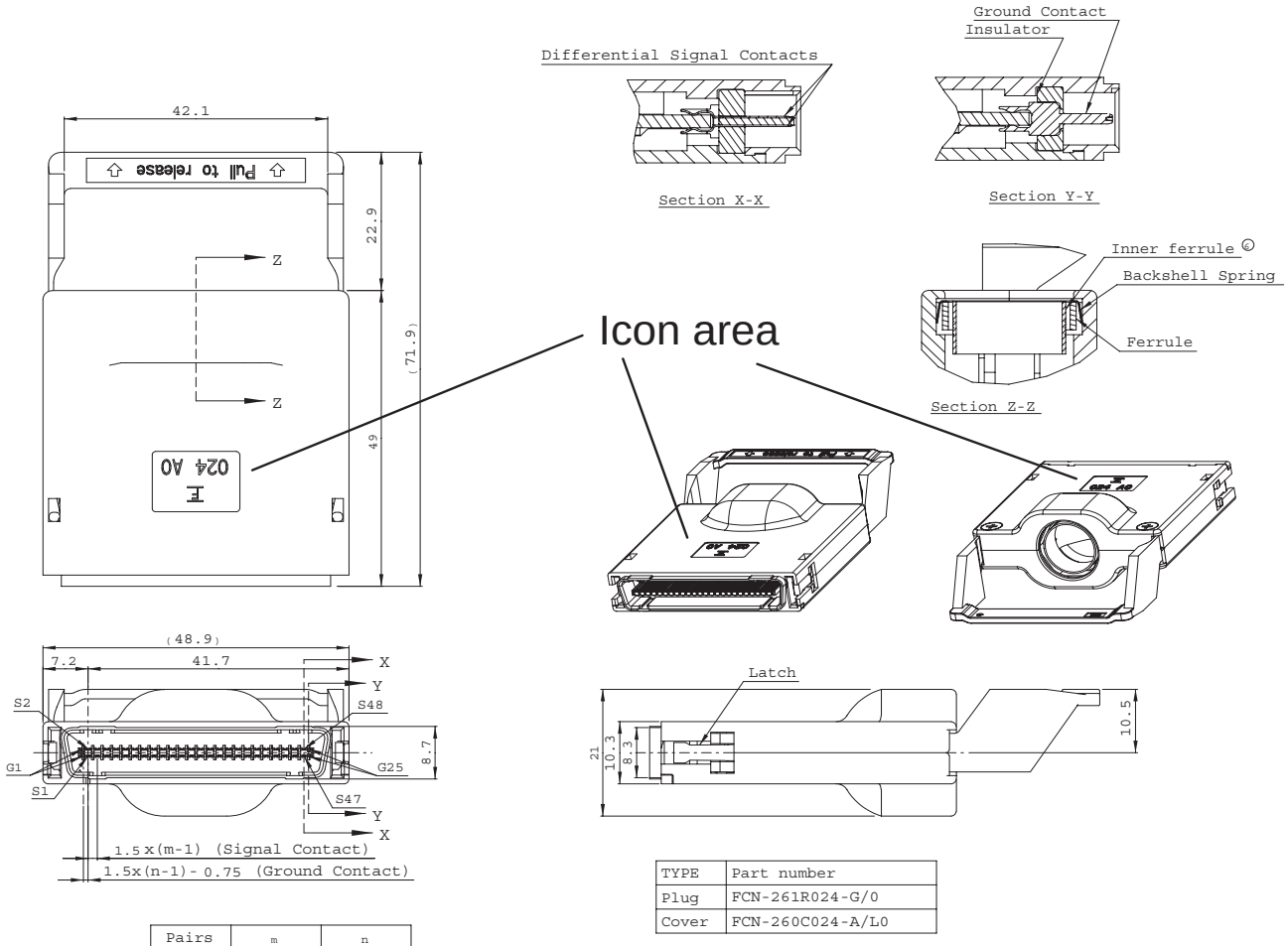
This cable plug uses a similar design to that of the 4x cable plug. The plug uses pairs of signal contacts interspersed with Ground contacts for crosstalk reduction. Side latches are used in conjunction with receptacle features for retention, and are released by pulling on the "lanyard" handle shown in the drawing. A suitable plug, referred to as a 24 pair MICRO GIGACN plug, is available from Fujitsu Components Ltd. and others, and is shown in Figure 64. Detailed drawings of mating interface dimensions are shown in Figure 63.

The cable plug may be designed using a dual bulk wire exit to accommodate the large bulk wire diameter while reducing bend radius. In that case, each of the two bulk cables would contain twelve signal pairs of wire.

C7-26: 12x Board receptacles used on InfiniBand modules **shall** be intermateable with this connector.

C7-26.1.1: A metal backshell or other means which fully shields the connector **shall** be bonded to the cable bulk shield through a continuous 360 degree contact to minimize EMI (Electromagnetic Interference).

Twenty-four pairs of signals are used, twelve each for transmit and receive.



Pairs	m	n
24	1-24	1-25

TYPE	Part number
Plug	FCN-261R024-G/0
Cover	FCN-260C024-A/L0

Rem1) Unless otherwise specified,dimensions are in mm and tolerances are ±0.5mm.

Figure 64 12x cable plug

7.7.9.2 PIN ASSIGNMENT

C7-27: The pin assignment listed in Table 44 **shall** be used for InfiniBand 12x cables.

There are no unused pins in the cable plug.

Note

Cable signal nets **shall** be connected from each *transmit* signal pair on the source port to the appropriate *receive* signal pair on the desired destination port. For example, a cable from port y to port z should be wired to connect IBtyOp(0) (positive output signal from port y, bit 0), to IBtzIp(0) (positive input to port z, bit 0) and IBtyOn(0) to IBtzIn(0), as shown in Table 44.

Table 44 12x cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
G1-G25	Signal Ground	G1-G25	Signal Ground
S1	IBtxIp(0)	S48	IBtxOp(0)
S2	IBtxIn(0)	S47	IBtxOn(0)
S3	IBtxIp(1)	S46	IBtxOp(1)
S4	IBtxIn(1)	S45	IBtxOn(1)
S5	IBtxIp(2)	S44	IBtxOp(2)
S6	IBtxIn(2)	S43	IBtxOn(2)
S7	IBtxIp(3)	S42	IBtxOp(3)
S8	IBtxIn(3)	S41	IBtxOn(3)
S9	IBtxIp(4)	S40	IBtxOp(4)
S10	IBtxIn(4)	S39	IBtxOn(4)
S11	IBtxIp(5)	S38	IBtxOp(5)
S12	IBtxIn(5)	S37	IBtxOn(5)
S13	IBtxIp(6)	S36	IBtxOp(6)
S14	IBtxIn(6)	S35	IBtxOn(6)
S15	IBtxIp(7)	S34	IBtxOp(7)
S16	IBtxIn(7)	S33	IBtxOn(7)
S17	IBtxIp(8)	S32	IBtxOp(8)
S18	IBtxIn(8)	S31	IBtxOn(8)
S19	IBtxIp(9)	S30	IBtxOp(9)

Table 44 12x cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
S20	IBtxIn(9)	S29	IBtxOn(9)
S21	IBtxIp(10)	S28	IBtxOp(10)
S22	IBtxIn(10)	S27	IBtxOn(10)
S23	IBtxIp(11)	S26	IBtxOp(11)
S24	IBtxIn(11)	S25	IBtxOn(11)
S25	IBtxOn(11)	S24	IBtxIn(11)
S26	IBtxOp(11)	S23	IBtxIp(11)
S27	IBtxOn(10)	S22	IBtxIn(10)
S28	IBtxOp(10)	S21	IBtxIp(10)
S29	IBtxOn(9)	S20	IBtxIn(9)
S30	IBtxOp(9)	S19	IBtxIp(9)
S31	IBtxOn(8)	S18	IBtxIn(8)
S32	IBtxOp(8)	S17	IBtxIp(8)
S33	IBtxOn(7)	S16	IBtxIn(7)
S34	IBtxOp(7)	S15	IBtxIp(7)
S35	IBtxOn(6)	S14	IBtxIn(6)
S36	IBtxOp(6)	S13	IBtxIp(6)
S37	IBtxOn(5)	S12	IBtxIn(5)
S38	IBtxOp(5)	S11	IBtxIp(5)
S39	IBtxOn(4)	S10	IBtxIn(4)
S40	IBtxOp(4)	S9	IBtxIp(4)
S41	IBtxOn(3)	S8	IBtxIn(3)
S42	IBtxOp(3)	S7	IBtxIp(3)
S43	IBtxOn(2)	S6	IBtxIn(2)
S44	IBtxOp(2)	S5	IBtxIp(2)
S45	IBtxOn(1)	S4	IBtxIn(1)
S46	IBtxOp(1)	S3	IBtxIp(1)
S47	IBtxOn(0)	S2	IBtxIn(0)
S48	IBtxOp(0)	S1	IBtxIp(0)

IBTA

Table 44 12x cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
Housing	Chassis Ground	Housing	Chassis Ground

C7-28: Signal Ground **shall not** be connected to Chassis Ground in the cable or cable connector. See Section 9.5.4.

7.7.9.3 KEYING

It is recommended that the 12x copper cable connectors as used for InfiniBand use keys as shown in Figure 53.

7.7.10 12X TO 3-4X COPPER CABLES

The 12x to 3-4x Copper Cables are used for connecting to devices which may configurably operate with a single 12x port, or with three separate 4x ports, using the same pins. The 12x to 3-4x copper cable provides an interface to a 12x interface board connector, operating as either a single 12x port or three 4x ports. The opposite side of the cable provides three separate 4x cable connectors.

7.7.10.1 12X INTERFACE BOARD CONNECTOR

This section defines the connector for the 12x cable interface on InfiniBand boards. The 12x interface is the widest InfiniBand link, providing for simultaneous transmit and receive of twelve bits of encoded differential data.

7.7.10.2 PIN ASSIGNMENTS

C7-28.2.1: The pin assignment listed in Table 45 **shall** be used for the board connector for InfiniBand 12x to 3-4x cables.

The character 'x' in the signal symbol is the port number, as defined in Chapter 4: Port Signal Definitions.

Table 45 12x board connector signal assignment for 12x to 3-4x cables

Pin Number	Signal (single 12x port)	Signal (three 4x ports)
G1-G25	Signal Ground	Signal Ground
S1	IBtxIp(0)	IBtx.1Ip(0)
S2	IBtxIn(0)	IBtx.1In(0)
S3	IBtxIp(1)	IBtx.1Ip(1)

Table 45 12x board connector signal assignment for 12x to 3-4x cables

Pin Number	Signal (single 12x port)	Signal (three 4x ports)
S4	IBtxIn(1)	IBtx.1In(1)
S5	IBtxIp(2)	IBtx.1Ip(2)
S6	IBtxIn(2)	IBtx.1In(2)
S7	IBtxIp(3)	IBtx.1Ip(3)
S8	IBtxIn(3)	IBtx.1In(3)
S9	IBtxIp(4)	IBtx.2Ip(0)
S10	IBtxIn(4)	IBtx.2In(0)
S11	IBtxIp(5)	IBtx.2Ip(1)
S12	IBtxIn(5)	IBtx.2In(1)
S13	IBtxIp(6)	IBtx.2Ip(2)
S14	IBtxIn(6)	IBtx.2In(2)
S15	IBtxIp(7)	IBtx.2Ip(3)
S16	IBtxIn(7)	IBtx.2In(3)
S17	IBtxIp(8)	IBtx.3Ip(0)
S18	IBtxIn(8)	IBtx.3In(0)
S19	IBtxIp(9)	IBtx.3Ip(1)
S20	IBtxIn(9)	IBtx.3In(1)
S21	IBtxIp(10)	IBtx.3Ip(2)
S22	IBtxIn(10)	IBtx.3In(2)
S23	IBtxIp(11)	IBtx.3Ip(3)
S24	IBtxIn(11)	IBtx.3In(3)
S25	IBtxOn(11)	IBtx.3On(3)
S26	IBtxOp(11)	IBtx.3Op(3)
S27	IBtxOn(10)	IBtx.3On(2)
S28	IBtxOp(10)	IBtx.3Op(2)
S29	IBtxOn(9)	IBtx.3On(1)
S30	IBtxOp(9)	IBtx.3Op(1)
S31	IBtxOn(8)	IBtx.3On(0)
S32	IBtxOp(8)	IBtx.3Op(0)
S33	IBtxOn(7)	IBtx.2On(3)

Table 45 12x board connector signal assignment for 12x to 3-4x cables

Pin Number	Signal (single 12x port)	Signal (three 4x ports)
S34	IBtxOp(7)	IBtx.2Op(3)
S35	IBtxOn(6)	IBtx.2On(2)
S36	IBtxOp(6)	IBtx.2Op(2)
S37	IBtxOn(5)	IBtx.2On(1)
S38	IBtxOp(5)	IBtx.2Op(1)
S39	IBtxOn(4)	IBtx.2On(0)
S40	IBtxOp(4)	IBtx.2Op(0)
S41	IBtxOn(3)	IBtx.1On(3)
S42	IBtxOp(3)	IBtx.1Op(3)
S43	IBtxOn(2)	IBtx.1On(2)
S44	IBtxOp(2)	IBtx.1Op(2)
S45	IBtxOn(1)	IBtx.1On(1)
S46	IBtxOp(1)	IBtx.1Op(1)
S47	IBtxOn(0)	IBtx.1On(0)
S48	IBtxOp(0)	IBtx.1Op(0)
Housing	Chassis Ground	Chassis Ground

7.7.10.3 12X TO 3-4X CABLE

This cable provides a means to connect a 12x board connector configured as three 4x ports to three separate 4x devices.

Notes

1. When using a 12x to 3-4x cable, ports 2 and 3 will not width negotiate since there is no lane 0 on those ports.

2. Port 3 (plug 4) of a 12x to 3-4x active cable plugged into an 8x board receptacle will not function.

Twenty-four pairs of signals are used, twelve each for transmit and receive. Eight pairs are used for each 4x port.

C7-28.2.2: A metal backshell or other means which fully shields the connector **shall** be bonded to the cable bulk shield through a continuous 360 degree contact to minimize EMI (Electromagnetic Interference).

7.7.10.4 12x CABLE CONNECTOR DESCRIPTION

The 12x cable plug uses the same connector as the 12x copper cable described in the *InfiniBand Architecture Specification, Volume 2*, Chapter 7: Copper Cable.

C7-28.2.3: 12x Board receptacles used on InfiniBand modules **shall** be intermateable with the cable connector used on the 12x end of the 12x to 3-4x cable.

The 4x cable plug uses the same connector as the 4x copper cable described in the *InfiniBand Architecture Specification, Volume 2*, Chapter 7: Copper Cable.

C7-28.2.4: 4x Board receptacles used on InfiniBand modules **shall** be intermateable with the cable connector used on the 4x end of the 12x to 3-4x cable.

7.7.10.5 PIN ASSIGNMENT

C7-28.2.5: The pin assignment listed in Table 46 **shall** be used for 12x to 3-4x cables.

> **Note**
>
> Cable signal nets **shall** be connected from each *transmit* signal pair on the source port to the appropriate *receive* signal pair on the desired destination port. For example, a cable from port y to port z should be wired to connect IBtyOp(0) (positive output signal from port y, bit 0), to IBtzIp(0) (positive input to port z, bit 0) and IBtyOn(0) to IBtzIn(0), as shown in Table 46.

Table 46 12x to 3-4x cable connector signal assignment

Plug 1		Plug number		
Pin number	Signal		Pin number	Signal
G1-G25	Signal Ground	2	G1-G9	Signal Ground
		3	G1-G9	Signal Ground
		4	G1-G9	Signal Ground

Table 46 12x to 3-4x cable connector signal assignment

Plug 1		Plug number		
Pin number	Signal		Pin number	Signal
S1	IBtx.1Ip(0)	2	S16	IBtx.1Op(0)
S2	IBtx.1In(0)	2	S15	IBtx.1On(0)
S3	IBtx.1Ip(1)	2	S14	IBtx.1Op(1)
S4	IBtx.1In(1)	2	S13	IBtx.1On(1)
S5	IBtx.1Ip(2)	2	S12	IBtx.1Op(2)
S6	IBtx.1In(2)	2	S11	IBtx.1On(2)
S7	IBtx.1Ip(3)	2	S10	IBtx.1Op(3)
S8	IBtx.1In(3)	2	S9	IBtx.1On(3)
S9	IBtx.2Ip(0)	3	S16	IBtx.2Op(0)
S10	IBtx.2In(0)	3	S15	IBtx.2On(0)
S11	IBtx.2Ip(1)	3	S14	IBtx.2Op(1)
S12	IBtx.2In(1)	3	S13	IBtx.2On(1)
S13	IBtx.2Ip(2)	3	S12	IBtx.2Op(2)
S14	IBtx.2In(2)	3	S11	IBtx.2On(2)
S15	IBtx.2Ip(3)	3	S10	IBtx.2Op(3)
S16	IBtx.2In(3)	3	S9	IBtx.2On(3)
S17	IBtx.3Ip(0)	4	S16	IBtx.3Op(0)
S18	IBtx.3In(0)	4	S15	IBtx.3On(0)
S19	IBtx.3Ip(1)	4	S14	IBtx.3Op(1)
S20	IBtx.3In(1)	4	S13	IBtx.3On(1)
S21	IBtx.3Ip(2)	4	S12	IBtx.3Op(2)
S22	IBtx.3In(2)	4	S11	IBtx.3On(2)
S23	IBtx.3Ip(3)	4	S10	IBtx.3Op(3)
S24	IBtx.3In(3)	4	S9	IBtx.3On(3)
S25	IBtx.3On(3)	4	S8	IBtx.3In(3)
S26	IBtx.3Op(3)	4	S7	IBtx.3Ip(3)
S27	IBtx.3On(2)	4	S6	IBtx.3In(2)
S28	IBtx.3Op(2)	4	S5	IBtx.3Ip(2)
S29	IBtx.3On(1)	4	S4	IBtx.3In(1)

Table 46 12x to 3-4x cable connector signal assignment

Plug 1		Plug number		
Pin number	Signal		Pin number	Signal
S30	IBtx.3Op(1)	4	S3	IBtx.3Ip(1)
S31	IBtx.3On(0)	4	S2	IBtx.3In(0)
S32	IBtx.3Op(0)	4	S1	IBtx.3Ip(0)
S33	IBtx.2On(3)	3	S8	IBtx.2In(3)
S34	IBtx.2Op(3)	3	S7	IBtx.2Ip(3)
S35	IBtx.2On(2)	3	S6	IBtx.2In(2)
S36	IBtx.2Op(2)	3	S5	IBtx.2Ip(2)
S37	IBtx.2On(1)	3	S4	IBtx.2In(1)
S38	IBtx.2Op(1)	3	S3	IBtx.2Ip(1)
S39	IBtx.2On(0)	3	S2	IBtx.2In(0)
S40	IBtx.2Op(0)	3	S1	IBtx.2Ip(0)
S41	IBtx.1On(3)	2	S8	IBtx.1In(3)
S42	IBtx.1Op(3)	2	S7	IBtx.1Ip(3)
S43	IBtx.1On(2)	2	S6	IBtx.1In(2)
S44	IBtx.1Op(2)	2	S5	IBtx.1Ip(2)
S45	IBtx.1On(1)	2	S4	IBtx.1In(1)
S46	IBtx.1Op(1)	2	S3	IBtx.1Ip(1)
S47	IBtx.1On(0)	2	S2	IBtx.1In(0)
S48	IBtx.1Op(0)	2	S1	IBtx.1Ip(0)
Housing	Chassis Ground	2, 3, 4	Housing	Chassis Ground

7.8 ACTIVE CABLES

It is desirable to provide for the development of cable assemblies or other devices which plug into cable receptacles which incorporate active circuitry. This would include but isn't limited to cables with built-in repeaters and copper cable substitutes which incorporate optical transceivers.

7.8.1 ACTIVE CABLE INTERFACE, 4X AND 12X WIDTHS

The active cable interface provides for simultaneous transmit and receive of four or twelve bits of encoded differential data with power available on the board connector for fiber or copper transponder devices that are external to the board. It uses the same board connector as the respective

copper cable interface, but with a different pin assignment for the signal ground pins.

7.8.1.1 4X ACTIVE CABLE INTERFACE BOARD CONNECTOR

This section defines the connector for the 4x active cable interface on InfiniBand boards.

C7-28.2.6: All 4x InfiniBand active cable plugs **shall** be intermateable with this board connector.

7.8.1.1.1 PIN ASSIGNMENTS

C7-28.2.7: The pin assignment listed in Table 47 shall be used for the board connector for InfiniBand 4x active cables. Usage requirements including current limitations on the power pins are described in Section 7.8.3, "Active Cable Power," on page 280. Power return is by way of the Signal Ground pins.

The 12 V or 3.3 V Sense signal is used to enable the respective voltage on the Vcc power supply pin used to provide power to the active components in the cable. Both signals are considered active if their voltage level is between 0.9 V and 2.4 V DC.

C7-28.2.8: Both 12 V and 3.3 V Sense signals **shall not** be active simultaneously.

The character 'x' in the signal symbol is the port number, as defined in Section 4.1, "Signal Naming Conventions," on page 69.

Table 47 4x active board connector signal assignment

Pin Number	Signal
G1	Sense-12V
G2-G6, G9	Signal Ground
S1	IBtxIp(0)
S2	IBtxIn(0)
S3	IBtxIp(1)
S4	IBtxIn(1)
S5	IBtxIp(2)
S6	IBtxIn(2)
S7	IBtxIp(3)

Table 47 4x active board connector signal assignment

Pin Number	Signal
S8	IBtxIn(3)
S9	IBtxOn(3)
S10	IBtxOp(3)
S11	IBtxOn(2)
S12	IBtxOp(2)
G7	Sense-3.3V
S13	IBtxOn(1)
S14	IBtxOp(1)
G8	Vcc
S15	IBtxOn(0)
S16	IBtxOp(0)
Housing	Chassis Ground

7.8.1.1.2 KEYING

It is recommended that the 4x active board and copper cable connectors use the same keying as that for a 4x cable as defined in the *InfiniBand Architecture Specification, Volume 2*, Section 7.4, "4x Interface board connector," on page 220.

7.8.1.2 8X ACTIVE CABLE INTERFACE BOARD CONNECTOR

This section defines the connector for the 8x active cable interface on InfiniBand boards. The active cable interface provides for simultaneous transmit and receive of four bits of encoded differential data with power available on the board connector for fiber or copper transponder devices. It uses the same board connector as the 12x copper cable interface, but with a different pin assignment for the signal ground pins.

C7-28.2.9: All 8x InfiniBand active cable plugs **shall** be intermateable with this board connector.

7.8.1.2.1 PIN ASSIGNMENTS

C7-28.2.10: The pin assignment listed in Table 48 **shall** be used for the board connector for active InfiniBand 8x cables. Usage requirements including current limitations on the power pins are described in Section 7.8.3, "Active Cable Power," on page 280. Power return is by way of the Signal Ground pins.

The 12 V or 3.3 V Sense signal is used to enable the respective voltage on the Vcc power supply pin used to provide power to the active components in the cable. Both signals are considered active if their voltage level is between 0.9 V and 2.4 V DC.

C7-28.2.11: Both 12 V and 3.3 V Sense signals **shall not** be active simultaneously.

The character 'x' in the signal symbol is the port number, as defined in Section 4.1, "Signal Naming Conventions," on page 69.

Table 48 8x active board connector signal assignment

Pin Number	Signal
G1	Sense-12V
G2, G4-G10, G12-G22, G25	Signal Ground
S1	IBtxIp(0)
S2	IBtxIn(0)
S3	IBtxIp(1)
S4	IBtxIn(1)
G3	Vcc
S5	IBtxIp(2)
S6	IBtxIn(2)
S7	IBtxIp(3)
S8	IBtxIn(3)
S9	IBtxIp(4)
S10	IBtxIn(4)
S11	IBtxIp(5)
S12	IBtxIn(5)
S13	IBtxIp(6)
S14	IBtxIn(6)
S15	IBtxIp(7)
S16	IBtxIn(7)
S17	reserved
S18	reserved

IBTA

Table 48 8x active board connector signal assignment

Pin Number	Signal
G11	Vcc
S19	reserved
S20	reserved
S21	reserved
S22	reserved
S23	reserved
S24	reserved
S25	reserved
S26	reserved
S27	reserved
S28	reserved
S29	reserved
S30	reserved
S31	reserved
S32	reserved
S33	IBtxOn(7)
S34	IBtxOp(7)
S35	IBtxOn(6)
S36	IBtxOp(6)
S37	IBtxOn(5)
S38	IBtxOp(5)
S39	IBtxOn(4)
S40	IBtxOp(4)
S41	IBtxOn(3)
S42	IBtxOp(3)
S43	IBtxOn(2)
S44	IBtxOp(2)
G23	Sense-3.3V
S45	IBtxOn(1)
S46	IBtxOp(1)

IBTA

Table 48 8x active board connector signal assignment

Pin Number	Signal
G24	Vcc
S47	IBtxOn(0)
S48	IBtxOp(0)
Housing	Chassis Ground

7.8.1.2.2 KEYING

It is recommended that the 8x active board and copper cable connectors use the same keying as that for a 12x cable as defined in the *InfiniBand Architecture Specification, Volume 2*, Chapter 7: Copper Cable.

7.8.1.3 12X ACTIVE CABLE INTERFACE BOARD CONNECTOR

This section defines the connector for the 12x active cable interface on InfiniBand boards. The active cable interface provides for simultaneous transmit and receive of four bits of encoded differential data with power available on the board connector for fiber or copper transponder devices. It uses the same board connector as the 12x copper cable interface, but with a different pin assignment for the signal ground pins.

C7-28.2.12: All 12x InfiniBand active cable plugs **shall** be intermateable with this board connector.

7.8.1.3.1 PIN ASSIGNMENTS

C7-28.2.13: The pin assignment listed in Table 49 **shall** be used for the board connector for InfiniBand 12x active cables. Usage requirements including current limitations on the power pins are described in Section 7.8.3, "Active Cable Power," on page 280. Power return is by way of the Signal Ground pins.

The 12 V or 3.3 V Sense signal is used to enable the respective voltage on the Vcc power supply pin used to provide power to the active components in the cable. Both signals are considered active if their voltage level is between 0.9 V and 2.4 V DC.

C7-28.2.14: Both 12 V and 3.3 V Sense signals **shall not** be active simultaneously.

The character 'x' in the signal symbol is the port number, as defined in Section 4.1, "Signal Naming Conventions," on page 69.

Table 49 12x active board connector signal assignment

Pin Number	Signal
G1	Sense-12V
G2, G4-G10, G12-G22, G25	Signal Ground
S1	IBtxIp(0)
S2	IBtxIn(0)
S3	IBtxIp(1)
S4	IBtxIn(1)
G3	Vcc
S5	IBtxIp(2)
S6	IBtxIn(2)
S7	IBtxIp(3)
S8	IBtxIn(3)
S9	IBtxIp(4)
S10	IBtxIn(4)
S11	IBtxIp(5)
S12	IBtxIn(5)
S13	IBtxIp(6)
S14	IBtxIn(6)
S15	IBtxIp(7)
S16	IBtxIn(7)
S17	IBtxIp(8)
S18	IBtxIn(8)
G11	Vcc
S19	IBtxIp(9)
S20	IBtxIn(9)
S21	IBtxIp(10)
S22	IBtxIn(10)
S23	IBtxIp(11)
S24	IBtxIn(11)
S25	IBtxOn(11)

IBTA

Table 49 12x active board connector signal assignment

Pin Number	Signal
S26	IBtxOp(11)
S27	IBtxOn(10)
S28	IBtxOp(10)
S29	IBtxOn(9)
S30	IBtxOp(9)
S31	IBtxOn(8)
S32	IBtxOp(8)
S33	IBtxOn(7)
S34	IBtxOp(7)
S35	IBtxOn(6)
S36	IBtxOp(6)
S37	IBtxOn(5)
S38	IBtxOp(5)
S39	IBtxOn(4)
S40	IBtxOp(4)
S41	IBtxOn(3)
S42	IBtxOp(3)
S43	IBtxOn(2)
S44	IBtxOp(2)
G23	Sense-3.3V
S45	IBtxOn(1)
S46	IBtxOp(1)
G24	Vcc
S47	IBtxOn(0)
S48	IBtxOp(0)
Housing	Chassis Ground

7.8.1.3.2 KEYING

It is recommended that the 12x active board and copper cable connectors use the same keying as that for a 12x cable as defined in the *InfiniBand Architecture Specification, Volume 2*, Section 7.5.

7.8.2 ACTIVE CABLES, 4X, 8X, AND 12X WIDTHS

7.8.2.1 4X ACTIVE CABLE

> **Note**
>
> A 4x active cable used with InfiniBand version 1 (unpowered) board connector pinout will not function.

7.8.2.1.1 4X ACTIVE CABLE CONNECTOR DESCRIPTION

This cable plug uses the same connector as the 4x copper cable described in the *InfiniBand Architecture Specification, Volume 2*, Chapter 7: Copper Cable.

C7-28.2.15: 4x board receptacles used on InfiniBand modules **shall** be intermateable with this active cable connector.

C7-28.2.16: A metal backshell or other means which fully shields the connector **shall** be bonded to the cable bulk shield through a continuous 360 degree contact to minimize EMI (Electromagnetic Interference).

Eight pairs of signals are used, four each for transmit and receive.

7.8.2.1.2 4X ACTIVE CABLE PIN ASSIGNMENT

C7-28.2.17: The pin assignment listed in Table 44 **shall** be used for active InfiniBand 4x cables.

C7-28.2.18: V_{cc} shall not be wired through the cable to the other end.

The 12 V or 3.3 V Sense signal is used to enable the respective voltage on the Vcc power supply pin used to provide power to the active components in the cable. Both signals are considered active if their voltage level is between 0.9 V and 2.4 V DC.

C7-28.1.1: Both 12 V and 3.3 V Sense signals shall not be active simultaneously.

> **Note**
>
> Cable signal nets **shall** be connected from each *transmit* signal pair on the source port to the appropriate *receive* signal pair on the desired destination port. For example, a cable from port y to port z should be wired to connect IBtyOp(0) (positive output signal from port y, bit 0), to IBtzIp(0) (positive input to port z, bit 0) and IBtyOn(0) to IBtzIn(0), as shown in Table 50.

Table 50 4x active cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
G1	Sense-12V	G9	Signal Ground
G4-G6	Signal Ground	G4-G6	Signal Ground
S1	IBtxIp(0)	S16	IBtxOp(0)
S2	IBtxIn(0)	S15	IBtxOn(0)
G2	Signal Ground	G8	Vcc
S3	IBtxIp(1)	S14	IBtxOp(1)
S4	IBtxIn(1)	S13	IBtxOn(1)
G3	Signal Ground	G7	Sense-3.3V
S5	IBtxIp(2)	S12	IBtxOp(2)
S6	IBtxIn(2)	S11	IBtxOn(2)
S7	IBtxIp(3)	S10	IBtxOp(3)
S8	IBtxIn(3)	S9	IBtxOn(3)
S9	IBtxOn(3)	S8	IBtxIn(3)
S10	IBtxOp(3)	S7	IBtxIp(3)
S11	IBtxOn(2)	S6	IBtxIn(2)
S12	IBtxOp(2)	S5	IBtxIp(2)
G7	Sense-3.3V	G3	Signal Ground
S13	IBtxOn(1)	S4	IBtxIn(1)
S14	IBtxOp(1)	S3	IBtxIp(1)
G8	Vcc	G2	Signal Ground
S15	IBtxOn(0)	S2	IBtxIn(0)
S16	IBtxOp(0)	S1	IBtxIp(0)
G9	Signal Ground	G1	Sense-12V
Housing	Chassis Ground	Housing	Chassis Ground

IBTA

7.8.2.2 8X ACTIVE CABLE

> **Note**
>
> A 8x active cable used with InfiniBand version 1 (unpowered) board connector pinout will not function.

7.8.2.2.1 8X ACTIVE CABLE CONNECTOR DESCRIPTION

This cable plug uses the same connector as the 12x copper cable described in the *InfiniBand Architecture Specification, Volume 2*, Chapter 7: Copper Cable.

C7-28.2.19: 8x Board receptacles used on InfiniBand modules **shall** be intermateable with this active cable connector.

C7-28.2.20: A metal backshell or other means which fully shields the connector **shall** be bonded to the cable bulk shield through a continuous 360 degree contact to minimize EMI (Electromagnetic Interference).

Sixteen pairs of signals are used, eight each for transmit and receive.

7.8.2.2.2 8X ACTIVE CABLE PIN ASSIGNMENT

C7-28.2.21: The pin assignment listed in Table 44 **shall** be used for active InfiniBand 8x cables.

C7-28.2.22: V_{cc} shall not be wired through the cable to the other end.

The 12 V or 3.3 V Sense signal is used to enable the respective voltage on the Vcc power supply pin used to provide power to the active components in the cable. Both signals are considered active if their voltage level is between 0.9 V and 2.4 V DC.

C7-28.1.2: Both 12 V and 3.3 V Sense signals shall not be active simultaneously.

> **Note**
>
> Cable signal nets **shall** be connected from each *transmit* signal pair on the source port to the appropriate *receive* signal pair on the desired destination port. For example, a cable from port y to port z should be wired to connect IBtyOp(0) (positive output signal from port y, bit 0), to IBtzIp(0) (positive input to port z, bit 0) and IBtyOn(0) to IBtzIn(0), as shown in Table 44.

Table 51 8x active cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
G1	Sense-12V	G25	Signal Ground
G4-10, G12-22	Signal Ground	G4-10, G12-22	Signal Ground
S1	IBtxIp(0)	S48	IBtxOp(0)
S2	IBtxIn(0)	S47	IBtxOn(0)
G2	Signal Ground	G24	Vcc
S3	IBtxIp(1)	S46	IBtxOp(1)
S4	IBtxIn(1)	S45	IBtxOn(1)
G3	Vcc	G23	Sense-3.3V
S5	IBtxIp(2)	S44	IBtxOp(2)
S6	IBtxIn(2)	S43	IBtxOn(2)
S7	IBtxIp(3)	S42	IBtxOp(3)
S8	IBtxIn(3)	S41	IBtxOn(3)
S9	IBtxIp(4)	S40	IBtxOp(4)
S10	IBtxIn(4)	S39	IBtxOn(4)
S11	IBtxIp(5)	S38	IBtxOp(5)
S12	IBtxIn(5)	S37	IBtxOn(5)
S13	IBtxIp(6)	S36	IBtxOp(6)
S14	IBtxIn(6)	S35	IBtxOn(6)
S15	IBtxIp(7)	S34	IBtxOp(7)
S16	IBtxIn(7)	S33	IBtxOn(7)
S17	reserved	S32	reserved
S18	reserved	S31	reserved
S19	reserved	S30	reserved
S20	reserved	S29	reserved
G11	Vcc	G16	Signal Ground
S21	reserved	S28	reserved
S22	reserved	S27	reserved

IBTA

Table 51 8x active cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
S23	reserved	S26	reserved
S24	reserved	S25	reserved
S25	reserved	S24	reserved
S26	reserved	S23	reserved
S27	reserved	S22	reserved
S28	reserved	S21	reserved
S29	reserved	S20	reserved
S30	reserved	S19	reserved
G16	Signal Ground	G11	Vcc
S31	reserved	S18	reserved
S32	reserved	S17	reserved
S33	IBtxOn(7)	S16	IBtxIn(7)
S34	IBtxOp(7)	S15	IBtxIp(7)
S35	IBtxOn(6)	S14	IBtxIn(6)
S36	IBtxOp(6)	S13	IBtxIp(6)
S37	IBtxOn(5)	S12	IBtxIn(5)
S38	IBtxOp(5)	S11	IBtxIp(5)
S39	IBtxOn(4)	S10	IBtxIn(4)
S40	IBtxOp(4)	S9	IBtxIp(4)
S41	IBtxOn(3)	S8	IBtxIn(3)
S42	IBtxOp(3)	S7	IBtxIp(3)
S43	IBtxOn(2)	S6	IBtxIn(2)
S44	IBtxOp(2)	S5	IBtxIp(2)
G23	Sense-3.3V	G3	Vcc
S45	IBtxOn(1)	S4	IBtxIn(1)
S46	IBtxOp(1)	S3	IBtxIp(1)
G24	Vcc	G2	Signal Ground
S47	IBtxOn(0)	S2	IBtxIn(0)
S48	IBtxOp(0)	S1	IBtxIp(0)

IBTA

Table 51 8x active cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
G25	Signal Ground	G1	Sense-12V
Housing	Chassis Ground	Housing	Chassis Ground

7.8.2.3 12X ACTIVE CABLE

Note

A 12x active cable used with InfiniBand version 1 (unpowered) board connector pinout will not function.

7.8.2.3.1 12X ACTIVE CABLE CONNECTOR DESCRIPTION

This cable plug uses the same connector as the 12x copper cable described in the *InfiniBand Architecture Specification, Volume 2*, Chapter 7: Copper Cable.

C7-28.2.23: 12x board receptacles used on InfiniBand modules **shall** be intermateable with this active cable connector.

C7-28.2.24: A metal backshell or other means which fully shields the connector **shall** be bonded to the cable bulk shield through a continuous 360 degree contact to minimize EMI (Electromagnetic Interference).

Twenty-four pairs of signals are used, twelve each for transmit and receive.

7.8.2.3.2 12X ACTIVE CABLE PIN ASSIGNMENT

C7-28.2.25: The pin assignment listed in Table 52 **shall** be used for InfiniBand 12x active cables.

C7-28.2.26: V_{cc} shall not be wired through the cable to the other end.

The 12 V or 3.3 V Sense signal is used to enable the respective voltage on the Vcc power supply pin used to provide power to the active components in the cable. Both signals are considered active if their voltage level is between 0.9 V and 2.4 V DC.

C7-28.1.3: Both 12 V and 3.3 V Sense signals shall not be active simultaneously.

Note

Cable signal nets **shall** be connected from each *transmit* signal pair on the source port to the appropriate *receive* signal pair on the desired destination port. For example, a cable from port y to port z should be wired to connect IBtyOp(0) (positive output signal from port y, bit 0), to IBtzIp(0) (positive input to port z, bit 0) and IBtyOn(0) to IBtzIn(0), as shown in Table 44.

Table 52 12x active cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
G1	Sense-12V	G25	Signal Ground
G4-10, G12-22	Signal Ground	G4-10, G12-22	Signal Ground
S1	IBtxIp(0)	S48	IBtxOp(0)
S2	IBtxIn(0)	S47	IBtxOn(0)
G2	Signal Ground	G24	Vcc
S3	IBtxIp(1)	S46	IBtxOp(1)
S4	IBtxIn(1)	S45	IBtxOn(1)
G3	Vcc	G23	Sense-3.3V
S5	IBtxIp(2)	S44	IBtxOp(2)
S6	IBtxIn(2)	S43	IBtxOn(2)
S7	IBtxIp(3)	S42	IBtxOp(3)
S8	IBtxIn(3)	S41	IBtxOn(3)
S9	IBtxIp(4)	S40	IBtxOp(4)
S10	IBtxIn(4)	S39	IBtxOn(4)
S11	IBtxIp(5)	S38	IBtxOp(5)
S12	IBtxIn(5)	S37	IBtxOn(5)
S13	IBtxIp(6)	S36	IBtxOp(6)
S14	IBtxIn(6)	S35	IBtxOn(6)
S15	IBtxIp(7)	S34	IBtxOp(7)

Table 52 12x active cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
S16	IBtxIn(7)	S33	IBtxOn(7)
S17	IBtxIp(8)	S32	IBtxOp(8)
S18	IBtxIn(8)	S31	IBtxOn(8)
S19	IBtxIp(9)	S30	IBtxOp(9)
S20	IBtxIn(9)	S29	IBtxOn(9)
G11	Vcc	G16	Signal Ground
S21	IBtxIp(10)	S28	IBtxOp(10)
S22	IBtxIn(10)	S27	IBtxOn(10)
S23	IBtxIp(11)	S26	IBtxOp(11)
S24	IBtxIn(11)	S25	IBtxOn(11)
S25	IBtxOn(11)	S24	IBtxIn(11)
S26	IBtxOp(11)	S23	IBtxIp(11)
S27	IBtxOn(10)	S22	IBtxIn(10)
S28	IBtxOp(10)	S21	IBtxIp(10)
S29	IBtxOn(9)	S20	IBtxIn(9)
S30	IBtxOp(9)	S19	IBtxIp(9)
G16	Signal Ground	G11	Vcc
S31	IBtxOn(8)	S18	IBtxIn(8)
S32	IBtxOp(8)	S17	IBtxIp(8)
S33	IBtxOn(7)	S16	IBtxIn(7)
S34	IBtxOp(7)	S15	IBtxIp(7)
S35	IBtxOn(6)	S14	IBtxIn(6)
S36	IBtxOp(6)	S13	IBtxIp(6)
S37	IBtxOn(5)	S12	IBtxIn(5)
S38	IBtxOp(5)	S11	IBtxIp(5)
S39	IBtxOn(4)	S10	IBtxIn(4)
S40	IBtxOp(4)	S9	IBtxIp(4)
S41	IBtxOn(3)	S8	IBtxIn(3)
S42	IBtxOp(3)	S7	IBtxIp(3)

DRAFT

Table 52 12x active cable connector signal assignment

Plug 1		Plug 2	
Pin number	Signal	Pin number	Signal
S43	IBtxOn(2)	S6	IBtxIn(2)
S44	IBtxOp(2)	S5	IBtxIp(2)
G23	Sense-3.3V	G3	Vcc
S45	IBtxOn(1)	S4	IBtxIn(1)
S46	IBtxOp(1)	S3	IBtxIp(1)
G24	Vcc	G2	Signal Ground
S47	IBtxOn(0)	S2	IBtxIn(0)
S48	IBtxOp(0)	S1	IBtxIp(0)
G25	Signal Ground	G1	Sense-12V
Housing	Chassis Ground	Housing	Chassis Ground

7.8.3 ACTIVE CABLE POWER

7.8.3.1 INTRODUCTION

This section defines the power use and control of the power provided for active cables as defined in the section of this specification on active cables. This power is provided to enable the construction of active devices such as repeaters and electro-optical converters which plug into InfiniBand version 2 Cable receptacles.

Since the active cable board connectors are intermateable with legacy InfiniBand cables, the associated power supplies must tolerate having their outputs shorted to ground for an indefinite period. Sense pins are supplied to assist in managing and switching the power supply output.

7.8.3.2 LOCAL DC-DC POWER CONVERTER(S)

The IB active cable will generate the design specific voltages from the supplied active cable voltage. These converters will be located in the Active Cable assembly.

7.8.3.3 POWER

The current and voltage specifications are intended to ensure support of active cable assemblies with power consumption at 12V of up to 4 W for a 4x interface and 10W for a 12x interface, and power consumption at 3.3V of up to 1.56 W for a 4x interface and 5.1W for a 12x interface. They are also intended to assure that current per ACP (Vcc) power pin is limited, to control connector heating.

7.8.3.4 VOLTAGE

C7-28.2.27: The Active Cable Power 12V (ACP12) voltage shall be a minimum of 10 volts at maximum current and 14 volts maximum at any current from zero to maximum, measured at the Vcc pins at the cable receptacle.

C7-28.2.28: The Active Cable Power 3.3V (ACP3) voltage shall be a minimum of 3.13 volts at maximum current and 3.47 volts maximum at any current from zero to maximum, measured at the Vcc pins at the cable receptacle.

7.8.3.5 CURRENT

C7-28.2.29: The Active Cable Power 12V (ACP12) interface shall be capable of supplying a minimum of 400 mA at 10V for a 4x interface and 1000 mA at 10V for a 12x interface.

C7-28.2.30: Each end of an active cable assembly using a 12V nominal power supply shall consume a maximum of 400 mA at 10V for a 4x interface, and 1000 mA at 10V for a 12x interface.

C7-28.2.31: The Active Cable Power 3.3V (ACP3) shall be capable of supplying a minimum of 500 mA at 3.13V for a 4x interface and 1.5 A at 3.13V for a 12x interface.

C7-28.2.32: Each end of an active cable assembly using a 3.3V nominal power supply shall consume a maximum of 500 mA at 3.13V for a 4x interface and 1.5 A at 3.13V for a 12x interface. Each end of an active cable assembly using a 3.3V nominal power supply shall consume a maximum of 500 mA per ACP (Vcc) power pin in the receptacle at the maximum voltage of 3.47V for the Active Cable Power 3.3V (ACP3) interface.

C7-28.2.33: An active cable assembly shall consume a maximum of 500 mA per ACP (Vcc) power pin in the receptacle at the maximum voltages of 14V for Active Cable Power 12V (ACP12) interface and 3.47V for the Active Cable Power 3.3V (ACP3) interface.

Note that the number of pins varies with the width of the port as defined in the section of this specification on active cables. Note also that, in this version of the specification, 8x links use the 12x interface, and are subject to 12x power specifications.

7.8.3.6 SENSE

C7-28.2.34: The active cable assembly shall connect a 5K +/- 5% resistor from SENSE-3.3V or SENSE-12V to ground to enable the active cable power.

C7-28.2.35: The ACP circuitry shall only enable power to the receptacle when the presence of the SENSE resistor is detected. If SENSE is connected to ground, or is open (no cable), ACP shall be disabled.

7.8.3.7 HOT PLUGGING

C7-28.2.36: The active cable power shall detect the sense pin value and provide full current availability within 50 milliseconds.

7.8.3.8 SHORT CIRCUIT PROTECTION

C7-28.2.37: The active cable power supply shall protect itself against indefinite connection to ground and shall limit short circuit current to less than 50 mA when the equivalent load resistance is less than 1 ohm.

7.8.3.9 LOAD IMPEDANCE

C7-28.2.38: The load shall present a maximum steady state current draw of 500 mA in parallel with a maximum capacitance of 500 microfarads on any Active Cable Power pin. The load shall limit the current drawn to 500 mA from any power pin.

7.8.3.10 AUXILIARY POWER

C7-28.2.39: Active Cable Power shall not be active when Bulk Power is not available.

7.8.3.11 BYPASSING

C7-28.2.40: All active cable power pins, including sense pins shall provide a low impedance to ground at frequencies from 100 MHz to 0.75 times the maximum bit rate supported on the port. This bypassing shall be provided on both the cable assembly and the receptacle.

IBTA

CHAPTER 8: FIBER ATTACHMENT - 2.5 GB/S, 5.0 GB/S, & 10 GB/S

8.1 INTRODUCTION

This chapter describes the InfiniBand Fiber Attachment for link operation at 2.5 Gb/s (SDR), 5.0 Gb/s (DDR), and 10Gb/s (QDR). This provides effective link bandwidths ranging from 5.0 Gb/s to 120 Gb/s. The Fiber Attachment is a media-level, point-to-point, duplex fiber optic interconnect.

A class of very short reach (VSR) fiber optic interconnect options is defined, and is referred to as -SX. A class of longer reach fiber optic interconnect options is defined, and is referred to as -LX. Specifications are provided for each of the interface widths (1x, 4x, 8x, 12x) at one or more distance classes. Certain optical specifications and fiber connector specifications differ between distance classes, and differ between interface width options within a distance class. Except where noted, the specifications contained herein apply to all distance classes for each of the interface width options.

8.2 SCOPE

This chapter defines the following attributes for the 2.5, 5.0 & 10.0 GTransfers/second Fiber Attachment:

- Optical transmission scheme
- Optical Transmitter mask compliance
- Jitter compliance
- Eye safety
- Optical link budget and distance
- Optical Receptacle and Optical Connector
- Optical Cable Plant
- Bulk and Aux power connections.
- Link encoding
- Pluggable Devices

An InfiniBand Optical Transceiver **may** be permanently attached with other electronic components on a printed circuit board, or **may** be fabricated as a component that plugs-in through the connector housing. Implementation notes are included at the end of this chapter describing recommended arrangements to interface an InfiniBand Optical Transceiver to other components on an InfiniBand printed circuit board.

IBTA

8.3 FIBER ATTACHMENT TECHNOLOGY OPTIONS

Fiber Optic Attachment technology options allowed in this version of the specification are listed in Table 53, Table 54, Table 55. Detailed specifications for these options are provided in the remainder of this chapter.

Table 53 Fiber Optic Attachment Option for SDR (2.5 Gb/s)

	Very short reach (VSR)	Longer reach
1x Wide		
Designation	**IB-1x-SX**	**IB-1x-LX**
Wavelength	850nm	1300nm
Connector	dual-LC	dual-LC
Worst-case operating range	2m - 250m using 50/125µm 500MHz.km fiber 2m - 500m using 50/125µm 2000MHz.km fiber 2m - 125m using 62.5/125µm 200MHz.km fiber	2m-10km with single mode fiber
4x Wide		
Designation	**IB-4x-SX**	**IB-4x-LX**
Wavelength	850nm	1300nm
Connector	single MPO	dual-SC
Worst-case operating range	2m - 125m using 50/125µm 500MHz.km fiber 2m - 200m using 50/125µm 2000MHz.km fiber 2m - 75m using 62.5/125µm 200MHz.km fiber	2m-10km with single mode fiber
8x & 12x Wide		
Designation	**IB-8x-SX, IB-12x-SX**	see Note 1
Wavelength	850nm	
Connector	dual MPO	
Worst-case operating range	2m - 125m using 50/125µm 500MHz.km fiber 2m - 200m using 50/125µm 2000MHz.km fiber 2m - 75m using 62.5/125 µm 200MHz.km fiber	

1: 8x wide LX and 12x wide LX links are not defined in this version of the specification.

IBTA

Table 54 Fiber Optic Attachment Options for DDR (5.0 Gb/s)

	Very short reach (VSR)	Longer reach
1x Wide		
Designation	**IB-1x-DDR-SX**	**IB-1x-DDR-LX**
Wavelength	850nm	1300nm
Connector	dual-LC	dual-LC
Worst-case operating range	2m-125m using 50/125µm 500MHz.km fiber 2m-200m using 50/125µm 2000MHz.km fiber 2m-65m using 62.5/125µm 200MHz.km fiber	2m-10km with single mode fiber
4x Wide		
Designation	**IB-4x-DDR-SX**	**IB-DDR-4x-LX**
Wavelength	850nm	see Note 1
Connector	single MPO	
Worst-case operating range	2m-75m using 50/125µm 500MHz.km fiber 2m-150m using 50/125µm 2000MHz.km fiber 2m-50m using 62.5/125µm 200MHz.km fiber	
8x & 12x Wide		
Designation	**IB-8x-DDR-SX, IB-12x-DDR-SX**	**IB-12x-DDR-LX**
Wavelength	850nm	see Note 1
Connector	dual MPO	
Worst-case operating range	2m-75m using 50/125µm 500MHz.km fiber 2m-150m using 50/125µm 2000MHz.km fiber 2m-50m using 62.5/125µm 200MHz.km fiber	

1: 4x, 8x and 12x wide DDR LX links are not defined in this specification.

IBTA

Table 55 Fiber Optic Attachment Options for QDR (10.0 Gb/s)

	Very short reach (VSR)	Longer reach
1x Wide Designation Wavelength Connector Worst-case operating range	 **IB-1x-QDR-SX** 850nm dual-LC 2m-82m using 50/125µm 500MHz.km fiber 2m-300m using 50/125µm 2000MHz.km fiber 2m-33m using 62.5/125µm 200MHz.km fiber	 **IB-1x-QDR-LX** 1300nm dual-LC 2m-10km with single mode fiber
4x Wide Designation Wavelength Connector	 **IB-4x-QDR-SX** see Note 1	 **IB-4x-QDR-LX** see Note 1
8x & 12x Wide Designation Wavelength Connector	 **IB-8x-QDR-SX, IB-12x-QDR-SX** see Note 1	 **IB-8x-QDR-LX, IB-12x-QDR-LX** see Note 1

1: 4x, 8x, and 12x wide QDR SX and LX links are not defined in this specification.

8.4 FIBER ATTACHMENT OVERVIEW

This section provides an overview of the structure, concepts, and mechanisms of the InfiniBand 1x, 4x, 8x, and 12x fiber optic links operating at SDR, DDR, and QDR speeds.

8.4.1 FIBER OPTIC SYSTEM OVERVIEW

A fiber optic link in general conveys InfiniBand signals between two InfiniBand Boards. A fiber optic link used as an external loopback on a single InfiniBand Board is also possible. A fiber optic link consists of a Fiber Optic Cable connected by means of an Optical Connector at each end to a pair of Optical Transceivers. Each Optical Transceiver typically resides on an InfiniBand Board. The Fiber Optic Cable consists of one or more segments of optical fiber, joined together using Fiber Optic Adapters. The Optical Transceiver consists of the following elements:

1) an Optical Transmitter section which converts intermediate electrical signals to InfiniBand-compliant optical signals,

2) an Optical Receiver section which converts InfiniBand-compliant optical signals into intermediate electrical signals,
3) an Optical Receptacle (which **shall** be single or dual as specified in subsequent sections) into which the Optical Connector plugs,
4) a Signal Conditioner section to convert the vendor-specific intermediate electrical signals into InfiniBand-compliant electrical signals which are specified in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s,
5) In the case of 4x LX - a serializer in the transmitter and a de-serializer in the receiver,
6) a power management section.

The InfiniBand-compliant optical signal parameters are specified in Section 8.5. Optical Receptacles and Connectors are specified in Section 8.6. Fiber Optic cables are specified in Section 8.7. Signal Conditioners are specified in Section 8.8.

The Optical Transceiver implements power management functionality (AUX power, beaconing and remote wake-up) as specified in Chapter 14: OS Power Management. Aux power is specified in Section 8.9.

Architecture Note

The intermediate electrical interface within the Optical Transceiver is intended to be vendor-specific and is not specified in this document, although recommended implementations are outlined in Section 8.8. The concept of intermediate electrical signals is intended to facilitate the use of existing optoelectronic components in an InfiniBand architecture with the addition of only a new electrical interface chipset, referred to as a "Signal Conditioner". The Signal Conditioner will typically contain retiming and control functionality to convert the intermediate electrical signals to/from IB-compliant electrical signals. The Signal Conditioner **may** be physically integrated with the optoelectronic components into a solderable or pluggable physical assembly for attachment to an IB board. The intermediate electrical signals are not intended to be connected through the boundaries of the Optical Transceiver.

Implementations are also envisaged wherein an Optical Transceiver converts directly from IB-compliant optical signals to IB-compliant electrical signals.

8.4.2 1X SYSTEM OVERVIEW - SDR, DDR & QDR

C8-1: All 1x Optical Cables shall comply with the requirements in Section 8.4.2 for respective distance classifications (SX or LX).

1x-SX and 1x-LX links differ only in optical specifications (Section 8.5.8), fiber specification (Section 8.7.1) and color coding (Section 8.6.1.4). This overview section therefore applies to both 1x-SX and 1x-LX links, except where stated.

A 1x optical link carries a duplex 1x link as shown in Figure 65. IB optical signals are generated in the 1x-SX Optical Transmitter using one laser and are detected in the 1x-SX Optical Receiver using one photodetector.

Each segment of the 1x Fiber Optic Cable **shall** have one fiber for each direction. The 1x-SX fiber **shall** be multimode, and the 1x-LX fiber **shall** be single-mode. The 1x Optical Connector **shall** be a dual LC-type connector.

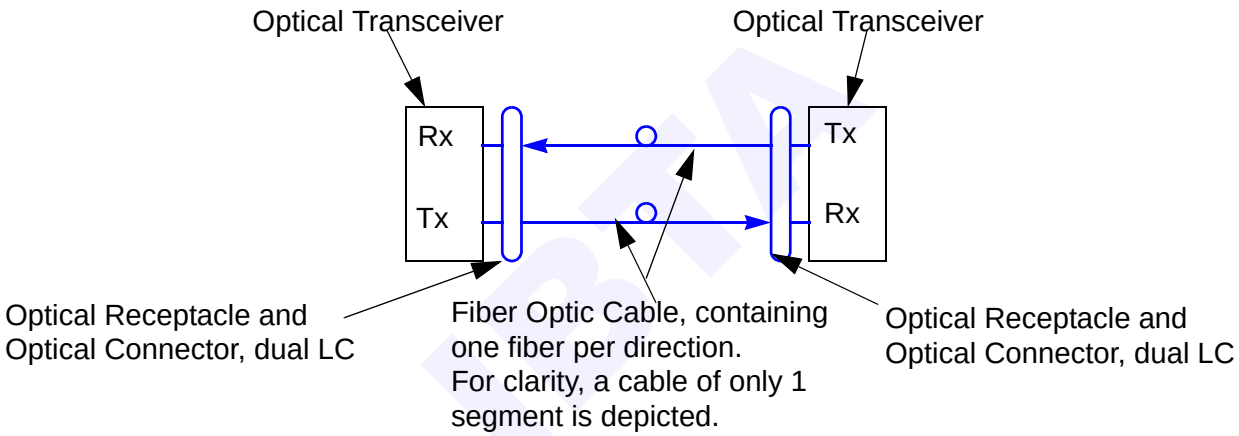


Figure 65 1x Optical Link Overview

8.4.3 4X SYSTEM OVERVIEW

C8-2: This compliance statement is obsolete and has been replaced by C8-2.1.1: and C8-2.1.2:

8.4.3.1 4X-SX OVERVIEW - SDR, DDR, & QDR

C8-2.1.1: All 4x-SX Fiber Optic Cables shall meet Section 8.4.3.1.

A 4x-SX optical link carries a duplex 4x link as shown in Figure 66. Typically the fibers are in a ribbon format. IB optical signals are typically generated in the 4x-SX Optical Transmitter using four lasers and are typically detected in the 4x-SX Optical Receiver using four photodetectors.

Each segment of the 4x SX Fiber Optic Cable **shall** have four multimode fibers for each direction. The 4x-SX Optical Connector **shall** be a single MPO connector. To simplify manufacturing, fibers **may** be present in the

four central positions of the connector, but if present these optional fibers **shall not** be used for IB signals.

4x optical cable links for SDR, DDR, and QDR differ only in supported data rate.

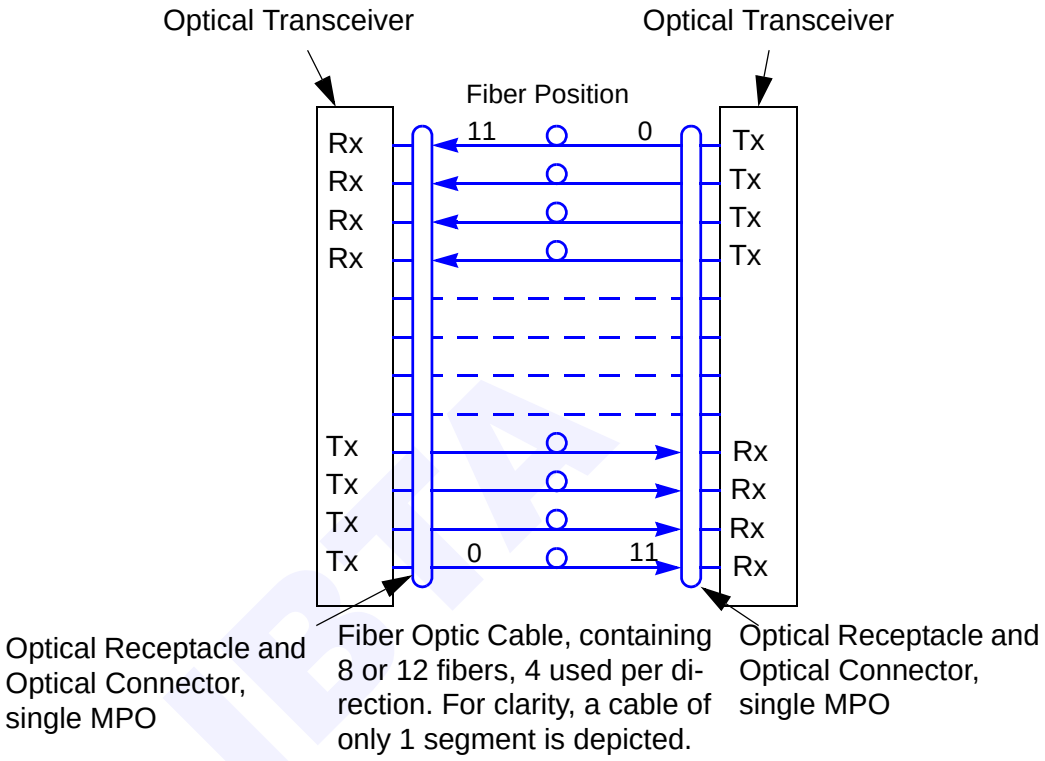


Figure 66 4x-SX Optical Link Overview

8.4.3.2 4x-LX OVERVIEW

Architecture Note

The IB-4x-LX link is somewhat unique, in comparison to the other optical links. The IB-4x-LX operates at a bit rate of 10.0 Gb/s, over a single fiber per direction, and is therefore similar technically similar to an IB-1x-QDR-LX link, but is not identical. It is, rather, a special case of the long-distance SDR 4x link, with unique data serialization, as described in Section 8.4.3.3 on page 290.

C8-2.1.2: All 4x-LX Fiber Optic Cables shall meet Section 8.4.3.2

A 4x-LX optical link carries a duplex 4x link as shown in Figure 67. IB optical signals are generated in the 4x-LX Optical Transmitter using one

laser and are detected in the 4x-LX Optical Receiver using one photodetector.

Each segment of the 4x-LX Fiber Optic Cable **shall** have one fiber for each direction. The 4x-LX fiber **shall** be single-mode. The 4x LX Optical Connector **shall** be a dual SC-type connector.

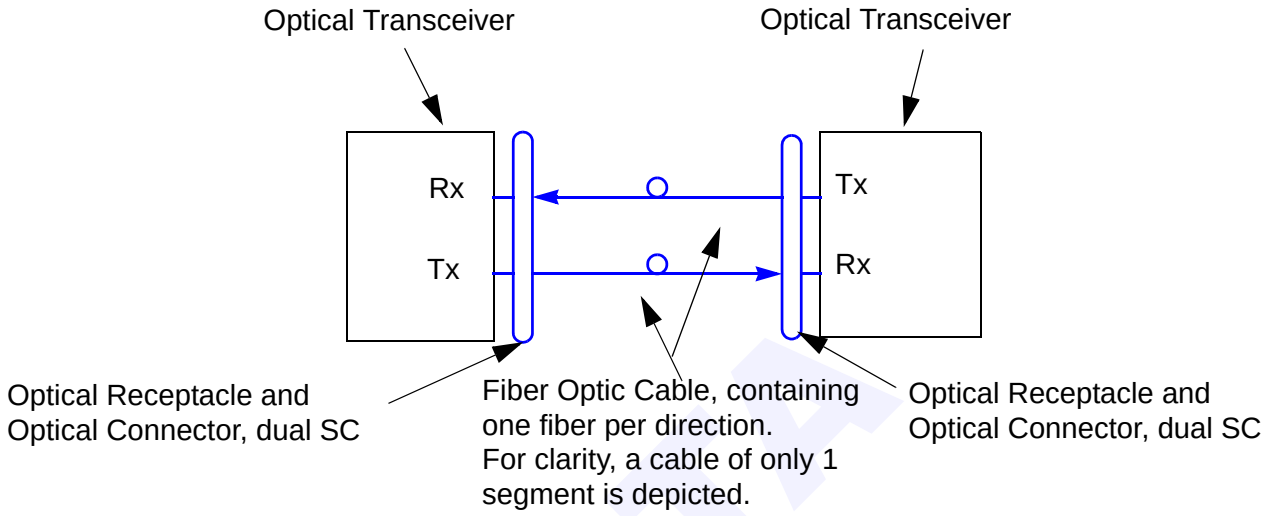


Figure 67 4x-LX Optical Link Overview

8.4.3.3 4X-LX SERIALIZATION

C8-2.1.3: All 4x-LX Optical transceivers **shall** comply with 8.4.3.3

The 4x-LX optical link is based on the 10Gb/s Ethernet specifications for 10GBase-LR link. The biggest optical difference is the link speed. 10Gb/s Ethernet serial runs at 10.3125 GBd while the 4x-LX link runs at 10.0 GBd. Both 10GBase-LR and 4x-LX transmitters accept an 8b/10b encoded byte-striped stream across 4 physical lanes. The 4 lanes for 10Gb/s Ethernet run at 3.125 GBd. To serialize the 4 lanes 10GbE removes the 8b/10b encoding from the bytes on the 4 lanes and then recodes using 64b/66b onto one lane. To keep the optical speed for 4x LX similar to 10GBase-LR, 4x LX optical transceivers will directly serialize the 4 lanes to achieve 10.0 Gbd.

A 4x-LX Optical Transmitter **shall** accept an 8b/10b encoded byte striped stream across the 4 physical lanes as described in Chapter 5: Link/Phy Interface. The output of the Optical transmitter **shall** serialize the traffic in the following manner:

The first byte on the cable **shall** be byte 0 from lane 0, followed by byte 1 from lane 1, then byte 2 from lane 2, then byte 3 from lane 3, then byte 4 from lane 0,… and so on. Each 10 bit byte **shall** be serialized in bit order starting with bit 0 and ending with bit 9. Refer to Figure 68 for a diagram detailing the transmitter serialization scheme.

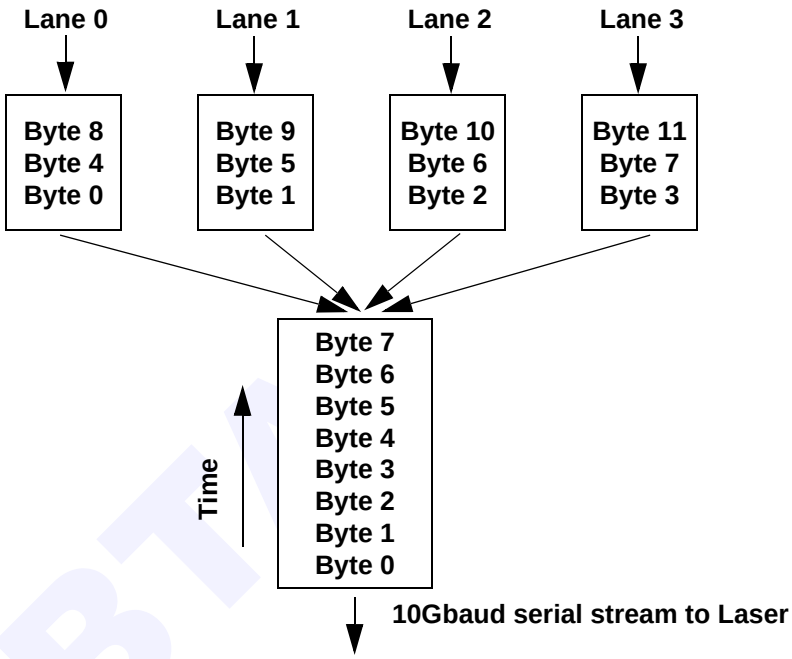


Figure 68 4x LX Transmitter Serialization

A 4x-LX Optical Receiver **shall** accept a 8b/10b encoded byte serialized stream as described above from a 4x-LX Transmitter. The output of the Optical Receiver **shall** be a byte-striped stream across 4 lanes as defined in Chapter 5: Link/Phy Interface. Refer to Figure 69 for a diagram detailing de-serialization.

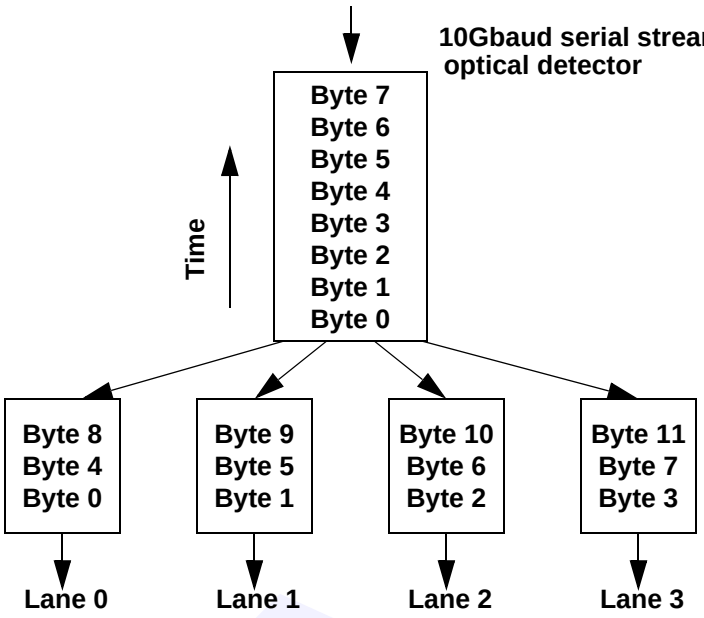


Figure 69 4x LX receiver de-serialization

Architecture Note

The transmitter function to serialize 4 lanes of byte striped traffic is fairly straight forward. The receiver is a little more complicated. One method for correctly de-serializing the symbol-multiplexed stream is to use the comma's to align on bytes. Then use the lane identifier information in TS1s and TS2s to rotate the incoming serial stream to align with the correct physical lane.

8.4.4 8X-SX OVERVIEW - SDR & DDR

C8-2.1.4: All 8x-SX Fiber Optic Cables shall meet Section 8.4.4

A 8x-SX optical link carries a duplex 8x link as shown in Figure 70. Typically the fibers are in a ribbon format. IB optical signals are typically generated in the 8x-SX Optical Transmitter using eight lasers and are typically detected in the 8x-SX Optical Receiver using eight photodetectors. To simplify manufacturing, fibers **may** be present in the four extra positions of the connector, but if present these optional fibers **shall** not be used for IB signals.

Each segment of the 8x SX Fiber Optic Cable **shall** have eight multimode fibers for each direction. The 8x-SX Optical Connector **shall** be a dual MPO connector

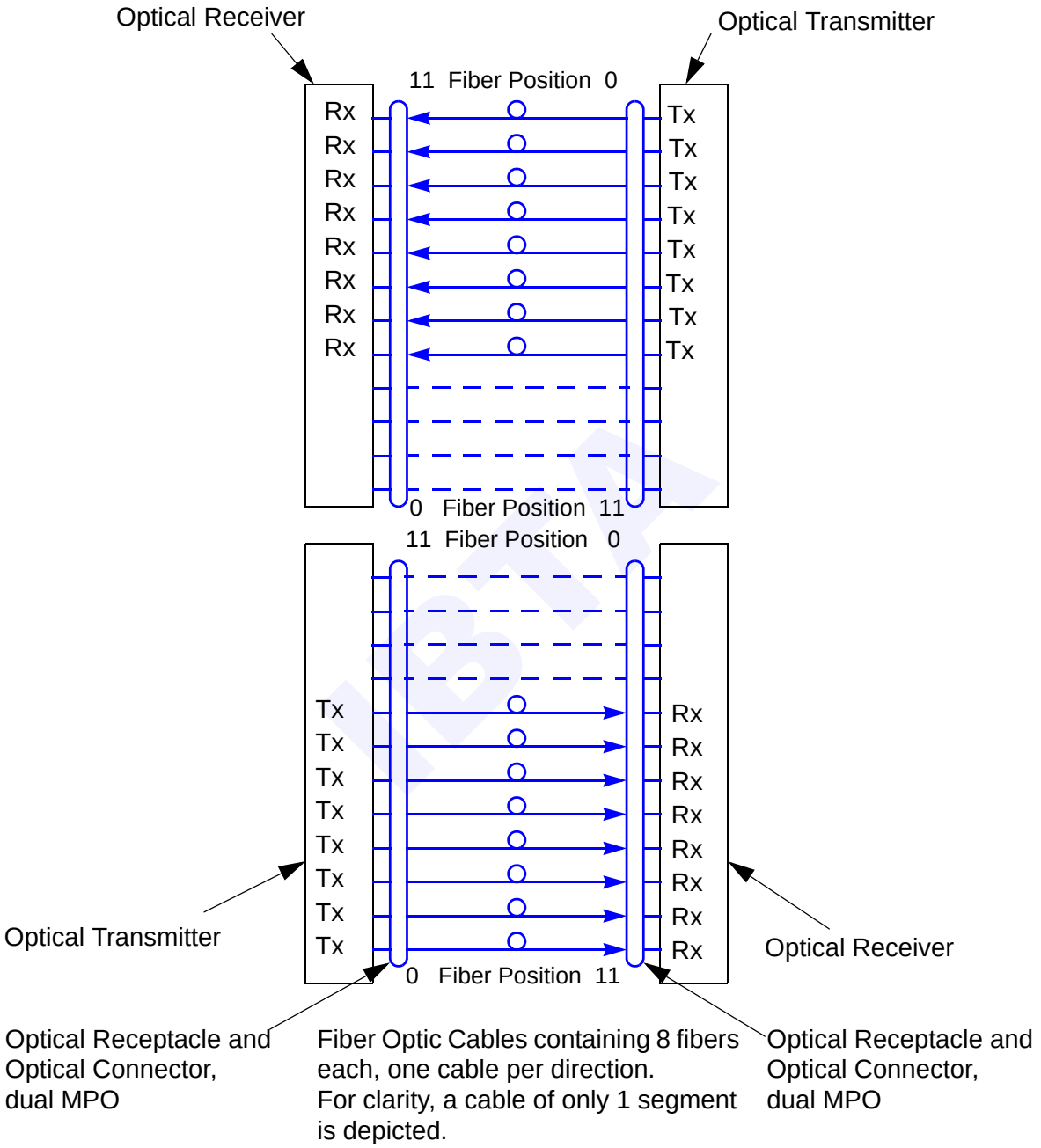


Figure 70 8x-SX Optical Link Overview

8.4.5 12X SYSTEM OVERVIEW - SDR, DDR, & QDR

C8-3: All 12x-SX Fiber Optic Cables shall meet Section 8.4.5.

A 12x-SX optical link carries a duplex 12x link as shown in Figure 71. Typically the fibers are in a ribbon format. IB optical signals are typically generated in the 12x-SX Optical Transmitter using twelve lasers and are typically detected in the 12x-SX Optical Receiver using twelve photodetectors.

Each segment of the 12x SX Fiber Optic Cable **shall** have twelve multimode fibers for each direction. The 12x-SX Optical Connector **shall** be a dual MPO connector.

IBTA

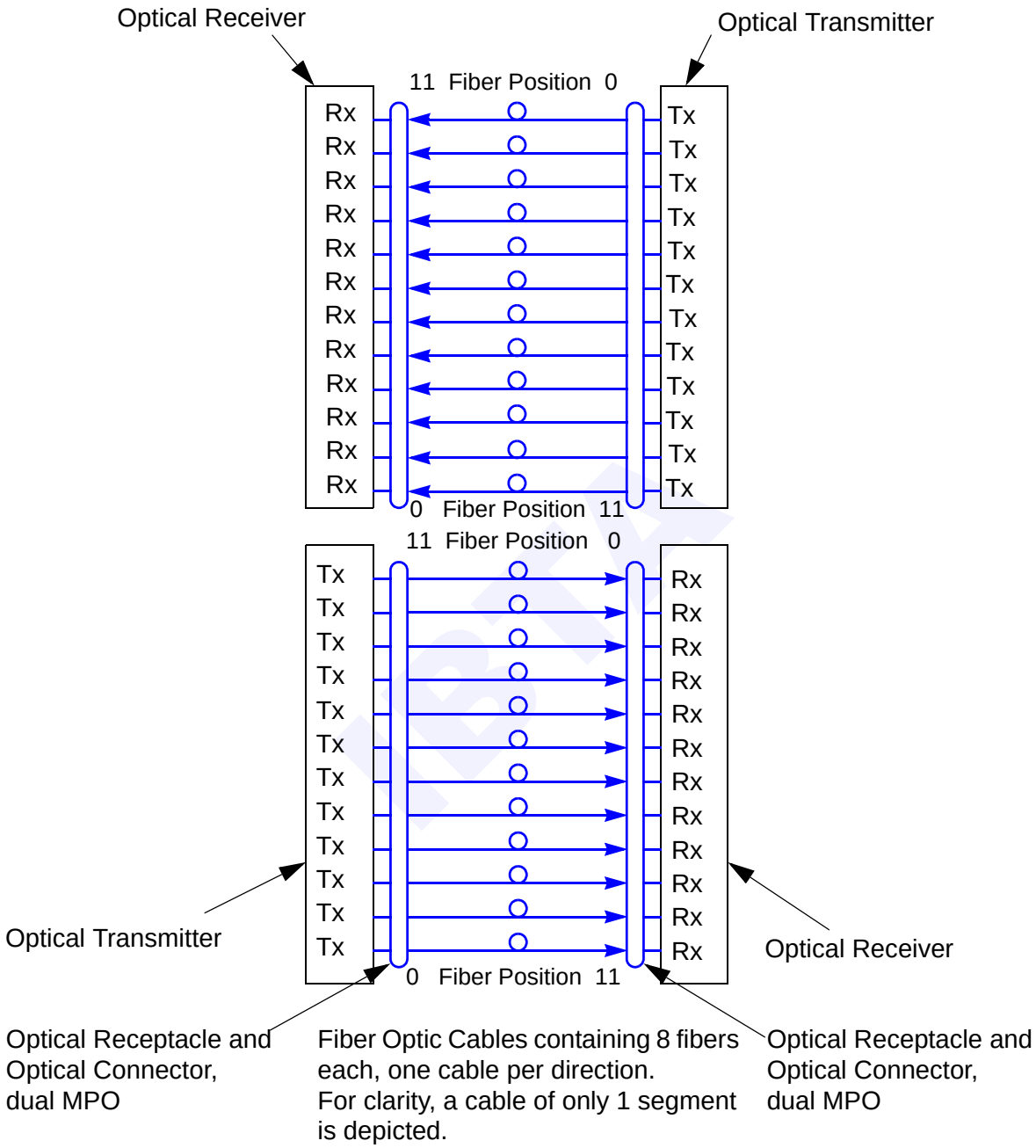


Figure 71 12x-SX Optical Link Overview

8.5 OPTICAL SPECIFICATIONS

This section defines the characteristics of InfiniBand-compliant optical signals. The Optical Transmitter parameters are specified immediately in-

side the optical fiber adjacent to the Optical Transmitter (TP2 of Figure 74). The Optical Receiver parameters are specified immediately inside the optical fiber adjacent to the Optical Receiver (TP3 of Figure 74). The corresponding fiber optic cable plant specifications are described in Section 8.7.

C8-4: The BER of each lane **shall not** exceed 10^{-12} under any conditions. The optical specifications in this section support meeting this requirement. In particular the Optical Receiver of each lane is expected to operate at a BER of $\leq 10^{-12}$ over lifetime, temperature, and operating range when driven through a cable plant as specified in Section 8.7 by an Optical Transmitter data stream compliant with the eye mask, jitter and optical parameter specifications.

Architecture Note

The maximum and minimum of the allowed range of average transmitter power coupled into the fiber are worst-case values to ensure that the specified minimum optical modulation amplitude can reliably be launched despite power supply variations, manufacturing variances, drift due to temperature variations, and aging effects.

The minimum received optical modulation amplitude and maximum average received power together define the input power range which is expected to achieve the required BER. The values specified take into account power penalties caused by the use of an Optical Transmitter and Fiber Optic Cable with a worst-case combination of spectral characteristics, optical modulation amplitude, maximum average power, and pulse shape characteristics.

The Gigabit Ethernet optical link model method of IEEE 802.3z was used to estimate the link performance. This model presently uses the parameters extinction ratio and average optical power. Transmitter and Receiver Optical Modulation Amplitudes were calculated from these data. DDR and QDR performance were estimated using 10GBE. This version uses OMA directly.

C8-5: This compliance statement is obsolete and has been replaced by C8-6.1.1: and C8-6.2.1:

C8-6: All Optical Ports **shall** meet the signal grounding requirements in Chapter 9: Mechanical Specification.

8.5.1 QUIESCENT CONDITION

C8-6.1.1: All Optical Ports **shall** meet the Quiescent conditions and Optical signal Polarity specified in Section 8.5.1, and Section 8.5.2

If an electrical input to an Optical Transmitter is quiescent, then the optical power of that lane **shall not** be modulated. If there are transitions at the electrical input to the Optical Transmitter, then the optical power **shall** be modulated according to the pattern of the transitions.

All parameters for each lane **shall** be met whether the other lanes are individually active or quiescent.

Implementation Note

Chapter 5: Link/Phy Interface includes a description of a beaconing sequence consisting of bursts of transitions separated by quiescent periods. It is recommended that during the quiescent periods DC current continue to flow through the lasers to bias the lasers at least to threshold and ideally to average optical power level. This will ensure that the optical signals meet specification as quickly as possible when the next beaconing burst starts.

8.5.2 OPTICAL SIGNAL POLARITY

C8-6.1.2: All Optical Ports **shall** meet the Optical Signal Polarity specified in Section 8.5.2.

A logic Zero high-speed signaling level at the electrical input to an Optical Transmitter **shall** generate a low level of optical power on the fiber. A logic One high-speed signaling level at the electrical input to an Optical Transmitter **shall** generate a high level of optical power on the fiber.

8.5.3 OPTICAL TRANSMITTER MASK COMPLIANCE FOR LINKS OPERATING AT 2.5 GB/S & 5.0 GB/S

C8-6.2.1: All 1x SX, 1x LX, 4x SX, and 12x SX Optical Ports **shall** meet the Optical Transmitter Mask specified in Section 8.5.3 while operating at SDR speed.

C8-6.2.2: All 1x SX, 1x LX, 4x SX, and 12x SX Optical ports **shall** meet the Optical Transmitter Mask specified in Section 8.5.3 while operating at DDR speed. For DDR, the O/E converter 3 dB frequency response **shall** be scaled up by a factor of 2 to 3.75 GHz.

8.5.3.1 EYE MASK SPECIFICATION

The optical transmitter pulse shape characteristics are specified in the form of a compliance mask on the eye diagram of Figure 72. This trans-

mitter compliance mask is used to verify the overall response of the Optical Transmitter for rise time, fall time, pulse overshoot, pulse undershoot, and ringing. Compliance with this optical mask is a very good indicator that deterministic effects are within generally acceptable limits, but it does not guarantee compliance with IB jitter specifications.

For uniform measurements, the Optical Transmitter eye **shall** be measured using an O/E converter with a equivalent fourth-order Bessel-Thompson response given by:

$$H_P = \frac{105}{105 + 105y + 45y^2 + 10y^3 + y^4}$$

where

$$y = 2.114p \qquad p = \frac{j\omega}{\omega_r} \qquad \omega_r = 2\pi f_r \qquad f_r = 1.875GHz$$

The O/E converter filter response is based on that described in ITU-T G.957, which provides a physical implementation. The specified O/E converter is only intended to provide uniform measurement and does not represent the noise response of an IB Optical Receiver. An actual SDR (**2.5 GB/S**) IB Optical Receiver has maximum 3dB bandwidth of 2.8 GHz, as specified in Table 65, Table 68, Table 77 and Table 83. The IB Optical Receiver maximum 3dB bandwidth does not apply to DDR (**5.0 GB/S**) Receivers.

The reference O/E converter **shall** have 3 dB frequency response of 1.88 GHz. The equivalent response of the 4th order (or higher) Bessel-Thompson reference O/E converter **shall** meet the values listed in Table 56 with tolerance **not** exceeding the values listed in Table 57.

Table 56 Equivalent Response of Reference O/E Converter

f/f_0	f/f_r	Attenuation (dB)	Distortion (UI)
0.15	0.2	0.1	0
0.3	0.4	0.4	0
0.45	0.6	1.0	0
0.6	0.8	1.9	0.002
0.75	1.0	3.0	0.008
0.9	1.2	4.5	0.025
1.0	1.33	5.7	0.044
1.05	1.4	6.4	0.055
1.2	1.6	8.5	0.10
1.35	1.8	10.9	0.14
1.5	2.0	13.4	0.19
2.0	2.67	21.5	0.30

Table 57 Attenuation Tolerance of Reference O/E Converter

Reference Frequency f/f_r	Attenuation Tolerance Δa (dB)
0.1 - 1.00 1.00 … 2.00	-0.0... + 0.5 +0.5 … +3.0
NOTE – Intermediate values of Δa **shall** be linearly interpolated on a logarithmic frequency scale.	

The Optical Transmitter compliance mask used for compliance testing **shall** be as in Figure 72. In this figure the amplitude has been normalized such that an amplitude of 0.0 represents logic ZERO and an amplitude of 1.0 represents logic ONE.

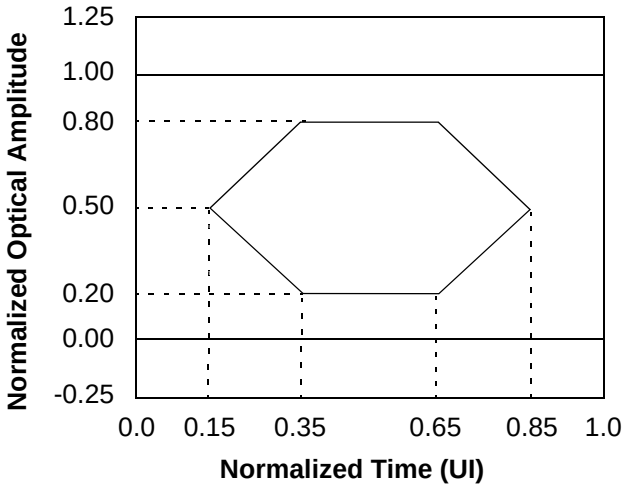


Figure 72 Normalized Optical Transmitter Compliance Mask

8.5.3.2 RISE/FALL TIME MEASUREMENT

Optical rise and fall time specifications are based on unfiltered waveforms. Some lasers have ringing or overshoot, which can reduce the accuracy of 20%-80% rise and fall time measurements. Therefore the 4th-order Bessel-Thompson filter defined in Section 8.5.3.1 is a convenient filter for measurement of the rise and fall time. Since the limited response of the 4th-order Bessel-Thompson filter will adversely impact the measured response, the following equation should be used to remove the filter response from the rise and fall times:

$$T_{rise} = \sqrt{(T_{riseMeasured})^2 - (T_{riseFilter})^2}$$

$$T_{fall} = \sqrt{(T_{fallMeasured})^2 - (T_{fallFilter})^2}$$

The filter 3 dB bandwidth used in the measurement **may** be different than the specified reference filter, but any filter used in the measurement **shall** be a fourth order Bessel-Thompson filter.

8.5.3.3 RMS RISE/FALL TIME

Optical rise time and fall time will not be identical in a typical implementation. Optical link models such as the IEEE 802.3z Gigabit Ethernet optical link model are generally defined using the larger of rise time and fall time. This provides an overly pessimistic analysis. Therefore IB optical specifications are defined using T_{rfRMS}, which is the RMS mean of rise time and fall time as defined below:

$$T_{rfRMS} = \sqrt{\frac{(T_{rise})^2 + (T_{fall})^2}{2}}$$

8.5.3.4 OPTICAL MODULATION AMPLITUDE

Optical Modulation Amplitude (OMA) is defined as the absolute difference between the optical power of a logic ONE level and the optical power of a logic ZERO level. *OMA* is related to Extinction Ratio (*ER* measured in dB) and Average Optical Power (P_{ave} measured in dBm) by the equation:

$$OMA = 2 \times 10^{P_{ave}/10} \times \frac{(1 - 10^{-ER/10})}{(1 + 10^{-ER/10})}$$

8.5.4 OPTICAL TRANSMITTER MASK COMPLIANCE FOR LINKS OPERATING AT 10.0 GB/S

8.5.4.1 1X QDR OPTICAL TRANSMITTER MASK COMPLIANCE - SX & LX

C8-6.2.3: The 1x SX and 1x LX Optical ports **shall** meet the Optical Transmitter Mask requirements in Section 8.5.4.1 while operating at QDR speed (**10.0 GB/S**).

The 1x SX Optical Port shall meet the Optical Transmitter Mask as defined for 10GBASE-SR in **ANSI/IEEE P802.3ae-2002® 10Gb/s Ethernet**. 1x SX QDR **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed shall be 10.0 GBaud ±100ppm

The 1x LX QDR Optical Port shall meet the Optical Transmitter Mask as defined for 10GBASE-LR in **ANSI/IEEE P802.3ae-2002® 10Gb/s Ethernet**. 1x LX QDR **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed shall be 10.0 GBaud ±100ppm

8.5.4.2 4X LX OPTICAL TRANSMITTER MASK COMPLIANCE

C8-6.1.3: All 4x-LX Optical Ports **shall** meet the Optical Transmitter Mask specified in Section 8.5.4.2.

Footnote 73. The 4x LX Optical Port **shall** meet the Optical Transmitter Mask as defined for 10GBASE-LR in **ANSI/IEEE P802.3ae-2002® 10Gb/s Ethernet**. 4x LX **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed shall be 10.0 GBaud ±100ppm

8.5.5 OPTICAL JITTER SPECIFICATION FOR LINKS OPERATING AT 2.5 GB/S AND 5.0 GB/S

C8-7: This compliance statement is obsolete and has been replaced by C8-7.2.1:

8.5.5.1 1X SX, 1X LX, 4X SX, AND 12X SX OPTICAL JITTER SPECIFICATIONS FOR 2.5 GB/S AND 5.0 GB/S

C8-7.1.1: This compliance statement is obsolete and has been replaced by C8-7.2.1:

C8-7.1.2: This compliance statement is obsolete and has been replaced by C8-7.2.4:

C8-7.2.1: All 1x-SX,1x-LX, 4x-SX, and 12x-SX Optical Ports shall comply with the Jitter requirements in Section 8.5.5.1 for respective Link widths/distances as defined in Table 58 while operating at SDR speed.

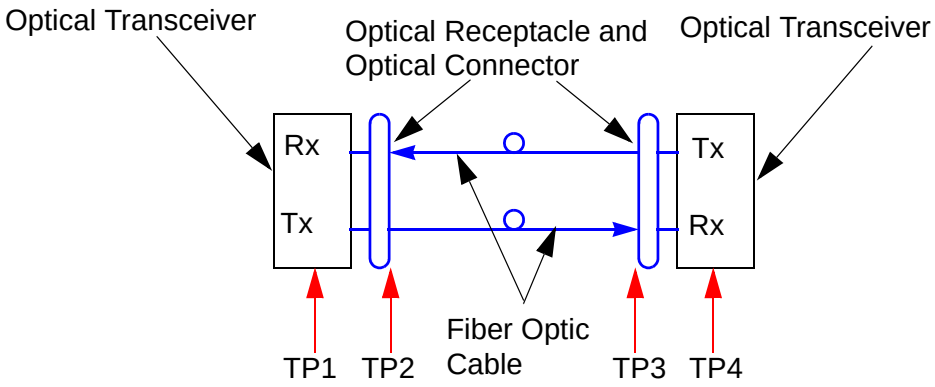
C8-7.2.2: All 1x-SX,1x-LX, 4x-SX, and 12x-SX Optical Ports shall comply with the Jitter requirements in Section 8.5.5.1 for respective Link widths/distances as defined in Table 59 while operating at DDR speed.

The IB jitter specification is based on the same methodology as the Fibre Channel - Methodologies for Jitter Specification revision 10.0 and the IEEE 802.3z Gigabit Ethernet standards. Figure 74 shows jitter compliance test points TP1, TP2, TP3 and TP4 for an IB-1x optical link.

TP2 is located immediately inside the output end of a test fiber of 2m length plugged into the Optical Transmitter and wrapped 10 turns around a 25.4mm-diameter mandrel, and it is the test point at which all Optical Transmitter optical parameters are defined. TP3 is located immediately inside the optical fiber adjacent to the Optical Receiver, and is the location at which all Optical Receiver optical parameters are defined. TP1 is located at the vendor-specific intermediate electrical signals within the Optical Transmitter. TP4 is located at the vendor-specific intermediate electrical signals within the Optical Receiver. The physical existence of TP1 and TP4 is optional.

Test points for 4x-SX, 8x-SX and 12x-SX links are defined similarly to those for 1x.

RFA



For clarity, only the jitter compliance test points for the left-to-right portion of the link are shown

Figure 74 Jitter Compliance Test Points

When operating at SDR speed, IB optical links **shall** not exceed the deterministic and total jitter values listed in Table 58 at TP2 and TP3 when the input jitter to the Optical Transceiver is compliant with the IB electrical specification. Typical jitter values at TP1 and TP4 are listed in Table 58 for reference. The total jitter of an optical component **shall** be measured at BER of 10^{-12} with a test sequence of K28.5+, K28.5- characters. An IB optical port must provide BER of $\leq 10^{-12}$ under worst case data patterns. Suitable test methods are defined in Fibre Channel - Methodologies for Jitter Specification revision 10.0. The jitter specification **shall** be met over the entire range of compliant optical parameters and compliant Fiber Optic Cables.

The total jitter listed for TP4 in Table 58 does not include a sinusoidal jitter (SJ) component.

Table 58 Maximum Jitter of Optical Links for SDR

InfiniBand Link	Compliance Point	Deterministic Jitter		Total Jitter	
		UI	ps	UI	ps
1x-SX, 1x-LX	TP1 (input test jitter)	0.10	40	0.25	100
	TP2	0.23	92	0.46	184
	TP3	0.30	120	0.54	216
	TP4	0.40	160	0.70	280
4x-SX, 12x-SX	TP1 (input test jitter)	0.10	40	0.25	100
	TP2	0.25	100	0.48	192
	TP3	0.30	120	0.53	212
	TP4	0.40	160	0.70	280

When operating at DDR speed, IB optical links **shall** not exceed the deterministic and total jitter values listed in Table 59 at TP2 and TP3 when the input jitter to the Optical Transceiver is compliant with the IB electrical specification.

Table 59 Maximum Jitter of Optical Links for DDR

InfiniBand Link	Compliance Point	Deterministic Jitter		Total Jitter	
		UI	ps	UI	ps
1x-SX, 1x-LX	TP1 (input test jitter)	0.14	28	0.26	52
	TP2	0.26	52	0.452	90.4
	TP3	0.265	53	0.58	116.1
	TP4	0.365	72.6	0.756	151.1
4x-SX, 8x-SX, 12x-SX	TP1 (input test jitter)	0.10	20	0.25	50
	TP2	0.22	44	0.443	89
	TP3	0.22	44.6	0.538	108
	TP4	0.32	64	0.757	151

For all link types, the Signal Conditioner or other component connected at TP4 **shall** tolerate total jitter of 0.80 UI for SDR (**2.5 GB/S**) and 0.085 for DDR (**5.0 GB/S**), which includes 0.10 UI of sinusoidal jitter (SJ) over a swept frequency from 1.5 MHz to1250 MHz as defined in Figure 75.

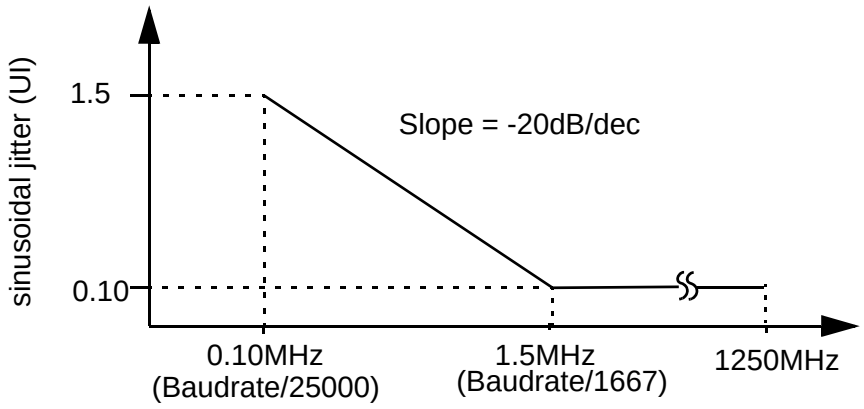


Figure 75 Jitter Tolerance Mask

Architecture Note - Jitter of Optical Links versus Jitter of High Speed Electrical Signaling

The attention of System Designers is drawn to the difference between the jitter specified at TP1 and TP4 in Table 58 compared to the jitter specifications of Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s.

In particular, a design which just meets the *informative* jitter specification of TP1 and TP4 will fail the *normative* jitter specification of Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s.

Implementation suggestions to deal with this issue are provided herein in Section 8.8.2.

Architecture Note - Jitter

Table 58 and Table 59 specify total jitter (TJ) at BER of 10^{-12} and specifies deterministic jitter (DJ). Random jitter (RJ) can be calculated from TJ and DJ:

RJ = TJ - DJ, where RJ is 14σ for BER of 10^{-12}.

DJ's of successive physical components add linearly. RJ's add in quadrature. TJ values can only be added by breaking them down into DJ and RJ components.

Jitter values are specified at each test point TP1, TP2, TP3, TP4. To determine the amount of deterministic jitter and random jitter that an Optical Transmitter under test adds from TP1 to TP2, the following analysis applies:

$$DJ(Transmitter) = DJ2 - DJ1$$

$$RJ(Transmitter) = \sqrt{(TJ2 - DJ2)^2 - (TJ1 - DJ1)^2}$$

where

DJ1 = DJ at TP1

DJ2 = DJ at TP2

TJ1 = TJ at TP1

TJ2 = TJ at TP2

A similar analysis using TP3 and TP4 applies to testing an Optical Receiver.

8.5.6 OPTICAL JITTER SPECIFICATIONS FOR LINKS OPERATING AT 10.0 GB/S

8.5.6.1 1X SX QDR AND 1X LX QDR OPTICAL JITTER SPECIFICATIONS

C8-7.2.3: All 1x-QDR-SX and 1x-QDR-LX Optical Ports shall comply with the Jitter requirements in Section 8.5.6.2 while operating at QDR speed.

The 1x SX QDR Optical Port shall meet the Optical Jitter specifications as defined for 10GBASE-SR in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. 4x SX QDR **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed shall be 10.0 GBaud ±100ppm

The 1x LX QDR and Optical Port shall meet the Optical Jitter specifications as defined for 10GBASE-LR in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. 4x LX QDR **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed shall be 10.0 GBaud ±100ppm

8.5.6.2 4X LX OPTICAL JITTER SPECIFICATIONS

C8-7.2.4: All 4x-LX Optical Ports shall comply with the Jitter requirements in Section 8.5.6.2.

The 4x LX Optical Port shall meet the Optical Jitter specifications as defined for 10GBASE-LR in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. 4x LX **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed shall be 10.0 GBaud ±100ppm

8.5.7 BIT TO BIT SKEW

C8-8: All IB optical ports **shall not** exceed the Maximum Skew values allowed across all physical lanes as defined in Table 60 while operating at SDR speed.

C8-8.2.1: All InfiniBand optical ports **shall not** exceed the Maximum Skew values allowed across all physical Lanes as defined in Table 61 while operating at DDR speed.

C8-8.2.2: All InfiniBand optical ports **shall not** exceed the Maximum Skew values allowed across all physical Lanes as defined in Table 62 while operating at QDR speed.

Table 60 defines the allowable bit to bit skew across the physical lanes for optical components. All IB optical ports **shall** limit skew to the maximum values defined in Table 60 when operating at SDR speed. All IB optical ports **shall** limit skew to the maximum values defined in Table 61 when operating at DDR speed. All IB optical ports **shall** limit skew to the maximum values defined in Table 62 when operating at QDR speed.

IBTA

Table 60 Maximum Optical Bit to Bit Skew Values for 2.5 Gb/s

Skew Parameter	Maximum Value
Optical Cable Assembly[a]	3.0ns
Transmitter[b]	500ps
Receiver[c]	500ps

a. An optical cable assembly **shall** include the optical cable and appropriate optical connectors at each end of the cable.
b. Between any two physical lanes within a transmitter.
c. Between any two physical lanes within a receiver.

Table 61 Maximum Optical Bit to Bit Skew Values for 5.0 Gb/s

Skew Parameter	Maximum Value
Optical Cable Assembly[a]	1.5ns
Transmitter[b]	250ps
Receiver[c]	250ps

a. An optical cable assembly **shall** include the optical cable and appropriate optical connectors at each end of the cable.
b. Between any two physical lanes within a transmitter.
c. Between any two physical lanes within a receiver.

Table 62 Maximum Optical Bit to Bit Skew Values for 10.0 Gb/s

Skew Parameter	Maximum Value
Optical Cable Assembly[a]	0.75ns
Transmitter[b]	125ps
Receiver[c]	125ps

a. An optical cable assembly **shall** include the optical cable and appropriate optical connectors at each end of the cable.
b. Between any two physical lanes within a transmitter.
c. Between any two physical lanes within a receiver.

8.5.8 1X SDR LINKS - AT 2.5 GB/S

A 1x-SX link operates in the 850nm wavelength band using multimode (MM) fiber. The optical parameters have been selected so as to allow typical 1x-SX Optical Transceivers to use a GaAs Vertical Cavity Surface Emitting Laser (VCSEL) and a photodetector, trans-impedance pre-amplifier and limiting post-amplifier. Three fiber types are specified for 1x-SX:

i) 1x-SX/50 link: 500MHz.km 50μm / 125μm MM fiber
ii) 1x-SX/50 link: 2000MHz.km 50μm / 125μm MM fiber

iii) 1x-SX/62 link: 200MHz.km 62.5µm / 125µm MM fiber

A 1x-LX link operates in the 1300nm wavelength band using singlemode (SM) fiber. The optical parameters have been selected to allow a typical 1x-LX Optical Transceiver to use an uncooled InP-based Fabry-Perot (FP) laser, Distributed Feedback (DFB) laser or VCSEL laser, and an In-GaAs photodetector, trans-impedance pre-amplifier and limiting post-amplifier. A trade-off curve is provided in Figure 76 between Center Wavelength and RMS Spectral Width to allow designers to select appropriate laser technology while still ensuring link operation. Only one fiber type is specified for 1x-LX: single-mode non-dispersion shifted.

8.5.8.1 1X EYE SAFETY

C8-9: The optical power coupled into the fiber for 1x-SX and 1x-LX **shall** be limited to a maximum value with Class I laser safety operation in accordance with CDRH and IEC 60825-1 Amendment 2 Radiation Safety of Laser Products: Equipment Classification, Requirements and User Guide.

8.5.8.2 1X-SX OPTICAL PARAMETERS

C8-10: This compliance statement is obsolete and has been replaced by C8-10.2.1:.

C8-10.2.1: All 1x-SX Optical Ports shall comply with the Transmitter and Receiver requirements in Section 8.5.8.2.

Table 63 gives the link budgets for 1x-SX fiber optic fiber links running at 2.5 GTransfers/second. Fiber plant specifications are described in Section 8.7.

Optical Passive Loss is the loss resulting from connections between Fiber Optic Segments (adapters or splices), and attenuation attributable to the fiber cable plant. Optical System Penalty includes all other link penalties other than Optical Passive Loss.

Table 63 Link Parameters - 1x-SX

Parameter	IB-1x-SX/50		IB-1x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Baud Rate		2500		2500	Mb/s	
Rate tolerance		±100		±100	ppm	

Table 63 Link Parameters - 1x-SX (Continued)

Parameter	IB-1x-SX/50		IB-1x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Optical Passive Loss		2.44		2.0	dB	
Optical System Penalty		3.56		4.0	dB	
Total link power budget	6		6		dB	
Worst case operating range		2 - 250[a] 2-500[b]		2 - 125	m	c
Fiber mode-field (core) diameter		50		62.5	µm	

a. 1x-SX/50 link: 500MHz.km 50µm / 125µm MM fiber
b. 1x-SX/50 link: 2000MHz.km 50µm / 125µm MM fiber
c. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 1x-SX link **shall** meet the parameters specified in Table 64 at TP2.

If any high-speed electrical input signals are less than the V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s then the optical modulation amplitude **shall not** exceed 2.0 µW.

Table 64 Optical Transmitter Parameters - 1x-SX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	830	860	nm	
RMS spectral width		0.85	nm	
Average launched power		-4.0	dBm	a
Optical Modulation Amplitude (OMA)	0.196		mW	b
RMS mean of 20% - 80% Rise/Fall time		150	ps	c
RIN_{12} (OMA)		-117	dB/Hz	

a. Average launched power, max. is the lesser of the eye safety limit or Average receiver power, max.
b. Optical modulation amplitude values are peak-to-peak.
c. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 1x-SX link **shall** meet the parameters specified in Table 65 at TP3. Conformance testing for a stressed receiver at TP3 **shall** follow the methods of Annex A of Fibre Channel Physical Interface revision 8.0.

If the average received optical power on any Lane is less than -30 dBm then the corresponding high-speed electrical signaling output **shall** be squelched to less than V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s. The electrical outputs of every lane with a optical power greater than -20dBm **shall not** be squelched.

Table 65 Optical Receiver Parameters - 1x-SX

Parameter	Minimum	Maximum	Units	Note
Average received power		-1.5	dBm	
Optical modulation amplitude	0.050		mW	a
Return loss of receiver	12		dB	
Stressed receiver sensitivity (OMA)	0.102		mW	a
Stressed receiver ISI test	2.0		dB	
Stressed receiver DCD component of DJ (at Tx)	40		ps	
Receiver electrical 3 dB upper cutoff frequency		2.8	GHz	
Receiver electrical 10 dB upper cutoff frequency		6.0	GHz	

a. Optical modulation amplitude values are peak-to-peak.

8.5.8.3 1X-LX OPTICAL PARAMETERS

C8-11: This compliance statement is obsolete and has been replaced by C8-11.2.1:.

C8-11.2.1: All 1x-LX Optical Ports shall comply with the Transmitter and Receiver requirements in Section 8.5.8.3.

Table 66 gives the link budgets for 1x-LX single-mode fiber optic links running at 2.5GTransfers/second. Fiber plant specifications are described in Section 8.7.

Table 66 Link Parameters - 1x-LX

Parameter	Minimum	Maximum	Units	Note
Baud Rate		2500	Mb/s	
Rate tolerance		±100	ppm	
Optical Passive Loss		6.64	dB	
Optical System Penalty		2.36	dB	
Total link power budget	9.0		dB	
Worst case operating range		2 - 10,000	m	a
Fiber mode-field (core) diameter		9	µm	

a. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 1x-LX link **shall** meet the parameters specified in Table 67 at TP2.

If any high-speed electrical input signals are less than the V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s then the optical modulation amplitude **shall not** exceed 2.0 µW.

In addition, the RMS spectral width of a 1x-LX Optical Transmitter **shall** lie on or below the curve shown in Figure 76 as a function of Optical Transmitter Center Wavelength. The curve was calculated using the worst-case fiber at a given Optical Transmitter Center Wavelength for all operating conditions. Specifically, the zero-dispersion wavelength of the fiber was chosen to be 1324nm for Optical Transmitter Center Wavelengths below approximately 1312nm, and the zero-dispersion wavelength of the fiber

was chosen to be 1300nm for Optical Transmitter Center Wavelengths above approximately 1312nm.

Table 67 Optical Transmitter Parameters - 1x-LX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	1270	1360	nm	
RMS spectral width			nm	a
Average launched power		-3.0	dBm	b
Optical Modulation Amplitude (OMA)	0.186		mW	c
RMS mean of 20% - 80% Rise/Fall time		150	ps	d
RIN_{12} (OMA)		-120	dB/Hz	

a. See text and Figure 76.
b. Average launched power, max. is the lesser of the eye safety limit or Average receiver power, max.
c. Optical modulation amplitude values are peak-to-peak.
d. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 1x-LX link **shall** meet the parameters specified in Table 68 at TP3.

If the average received optical power on any Lane is less than -30 dBm then the corresponding high-speed electrical signaling output **shall** be squelched to less than V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s. The electrical outputs of every Lane with a optical power greater than -20dBm **shall not** be squelched.

Table 68 Optical Receiver Parameters - 1x-LX

Parameter	Minimum	Maximum	Units	Note
Average received power		-1.5	dBm	
Optical modulation amplitude	0.0234		mW	a
Return loss of receiver	20		dB	

Table 68 Optical Receiver Parameters - 1x-LX

Parameter	Minimum	Maximum	Units	Note
Stressed receiver sensitivity (OMA)	0.0365		mW	a
Stressed receiver ISI test	0.58		dB	
Stressed receiver DCD component of DJ (at TX)	40		ps	
Receiver electrical 3 dB upper cutoff frequency		2.8	GHz	
Receiver electrical 10 dB upper cutoff frequency		6.0	GHz	

a. Optical modulation amplitude values are peak-to-peak.

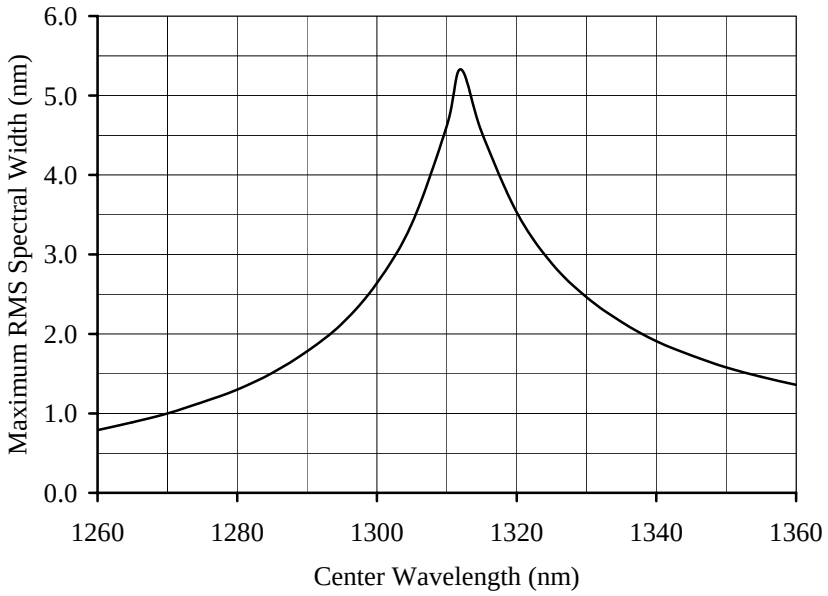


Figure 76 1x-LX Trade-off between RMS Spectral Width and Center Wavelength

8.5.9 1X DDR LINKS - AT 5.0 GB/S

A 1x-DDR-SX link uses similar transmitter, receiver, and fiber technology as 1x-SX links described in Section 8.5.8, and differ in bit rate, and in the parameters described below. A 1x-DDR-LX link also uses similar technology as 1x-LX links described in Section 8.5.8.

8.5.9.1 EYE SAFETY

C8-11.2.2: The optical power coupled into the fiber for 1x-DDR-SX and 1x-DDR-LX **shall** be limited to a maximum value with Class I laser safety operation in accordance with CDRH and IEC 60825-1 Amendment 2 Radiation Safety of Laser Products: Equipment Classification, Requirements and User Guide.

8.5.9.2 1X-DDR-SX OPTICAL PARAMETERS

C8-11.2.3: All 1x-DDR-SX Optical Ports shall comply with the Transmitter and Receiver requirements in Section 8.5.9.2.

Table 69 gives the link budgets for 1x-DDR-SX fiber optic fiber links running at 5.0 GTransfers/second. Fiber plant specifications are described in Section 8.7.

Optical Passive Loss is the loss resulting from connections between Fiber Optic Segments (adapters or splices), and attenuation attributable to the fiber cable plant. Optical System Penalty includes all other link penalties other than Optical Passive Loss.

Table 69 Link Parameters - 1x-DDR-SX

Parameter	IB-1x-SX/50		IB-1x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Baud Rate		5000		5000	Mb/s	
Rate tolerance		±100		±100	ppm	
Optical Passive Loss		1.97		1.76	dB	
Optical System Penalty		3.6		4.4	dB	
Total link power budget	7.93		7.93		dB	
Worst case operating range		2 - 125[a] 2-200[b]		2 - 65	m	[c]
Fiber mode-field (core) diameter		50		62.5	µm	

a. 1x-SX/50 link: 500MHz.km 50mm / 125mm MM fiber
b. 1x-SX/50 link: 2000MHz.km 50mm / 125mm MM fiber
c. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 1x-DDR-SX link **shall** meet the parameters specified in Table 70 at TP2.

Table 70 Optical Transmitter Parameters - 1x-DDR-SX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	830	860	nm	
RMS spectral width		0.85	nm	
Average launched power		-2.0	dBm	a
Optical Modulation Amplitude (OMA)	0.247		mW	b
RMS mean of 20% - 80% Rise/Fall time		75	ps	c
RIN_{12} (OMA)		-123	dB/Hz	

a. Average launched power, max. is the lesser of the eye safety limit or Average receiver power, max.
b. Optical modulation amplitude values are peak-to-peak.
c. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 1x-DDR-SX link **shall** meet the parameters specified in Table 71 at TP3. Conformance testing for a stressed receiver at TP3 **shall** follow the methods of Annex A of Fibre Channel Physical Interface revision 8.0.

Table 71 Optical Receiver Parameters - 1x-DDR-SX

Parameter	Minimum	Maximum	Units	Note
Average received power	0.0		dBm	
Optical modulation amplitude (Informative)		0.040 Informative	mW	a
Return loss of receiver	12		dB	
Stressed receiver sensitivity (OMA)	0.074		mW	**a**, b
Stressed receiver ISI test		2.48	dB	
Stressed receiver DCD component of DJ (at Tx)	17.0		ps	

a. Optical modulation amplitude values are peak-to-peak.

b. Stressed receiver entries for Sensitivity and ISI differ for each fiber type. The values here apply to 50u, 500 MHz-km fiber.

IBTA

8.5.9.3 1X-DDR-LX OPTICAL PARAMETERS

C8-11.2.4: All 1x-DDR-LX Optical Ports **shall** comply with the Transmitter and Receiver requirements in Section 8.5.9.3.

Table 72 gives the link budgets for 1x-LX single-mode fiber optic links running at 2.5GTransfers/second. Fiber plant specifications are described in Section 8.7.

Table 72 Link Parameters - 1x-DDR-LX

Parameter	Minimum	Maximum	Units	Note
Baud Rate		5000	Mb/s	
Rate tolerance		±100	ppm	
Optical Passive Loss		5.66	dB	
Optical System Penalty		2.4	dB	
Total link power budget	9.8		dB	
Worst case operating range		2 - 10,000	m	a
Fiber mode-field (core) diameter		9	µm	

a. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 1x-DDR-LX link **shall** meet the parameters specified in Table 73 at TP2.

In addition, the RMS spectral width of a 1x-DDR-LX Optical Transmitter **shall** lie on or below the curve shown in Figure 77 as a function of Optical Transmitter Center Wavelength and minimum Transmit OMA. The curve was calculated using the worst-case fiber at a given Optical Transmitter Center Wavelength for all operating conditions. Specifically, the zero-dispersion wavelength of the fiber was chosen to be 1324nm for Optical Transmitter Center Wavelengths below approximately 1312nm, and the zero-dispersion wavelength of the fiber was chosen to be 1300nm for Optical Transmitter Center Wavelengths above approximately 1312nm.

Table 73 Optical Transmitter Parameters - 1x-DDR-LX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	1270	1360	nm	
RMS spectral width		0.47	nm	a
Average launched power		-1.0	dBm	b
Optical Modulation Amplitude (OMA)	0.29		mW	c
RMS mean of 20% - 80% Rise/Fall time		75	ps	d
RIN_{12} (OMA)		-123	dB/Hz	

a. See text and Figure 77.
b. Average launched power, max. is the lesser of the eye safety limit or Average receiver power, max.
c. Optical modulation amplitude values are peak-to-peak.
d. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 1x-DDR-LX link **shall** meet the parameters specified in Table 74 at TP3.

Table 74 Optical Receiver Parameters - 1x-DDR-LX

Parameter	Minimum	Maximum	Units	Note
Average received power	-1.0		dBm	
Optical modulation amplitude (Informative)		0.029 Informative	mW	a
Return loss of receiver	12		dB	
Stressed receiver sensitivity (OMA)	0.030		mW	a
Stressed receiver ISI test	1.45		dB	
Stressed receiver DCD component of DJ (at TX)	16.3		ps	
Receiver electrical 3 dB upper cutoff frequency		5.3	GHz	b
Receiver electrical 10 dB upper cutoff frequency			GHz	2

a. Optical modulation amplitude values are peak-to-peak.
b. These rows may be deleted since receivers may operate over all three signal rates?

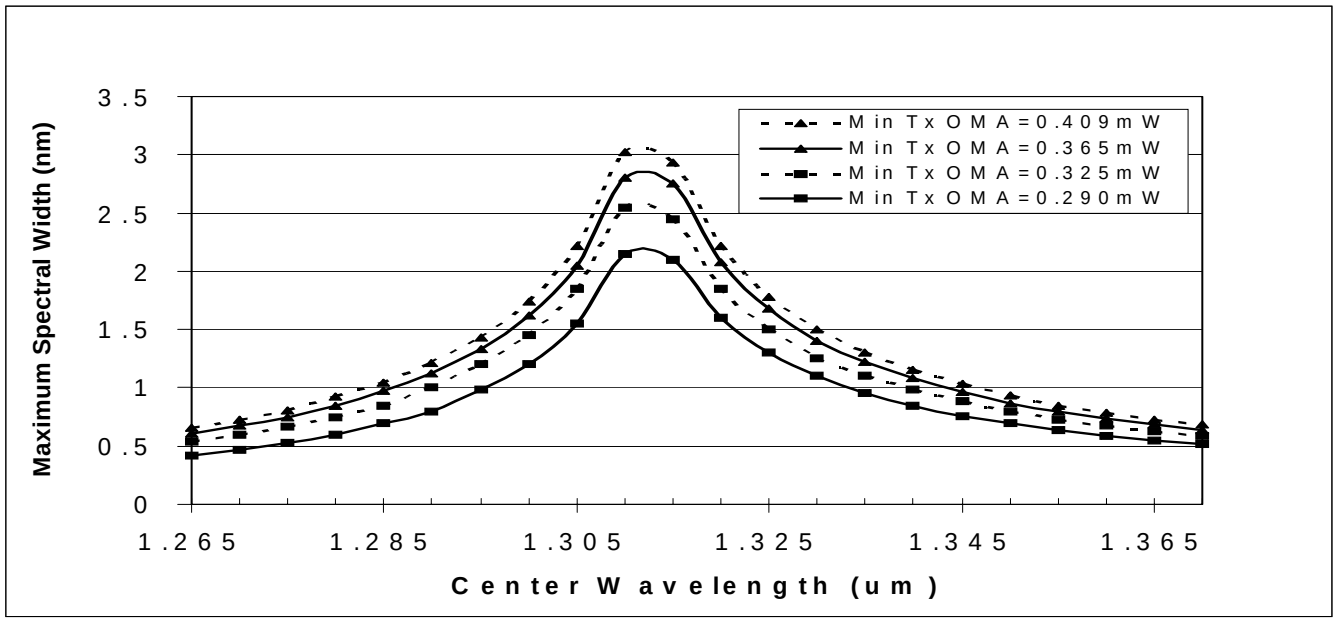


Figure 77 1x-DDR-LX Trade-off between OMA, RMS Spectral Width, and Center Wavelength

8.5.10 4X SDR LINKS - AT 2.5 GB/S

C8-12: This compliance statement is obsolete and has been replaced by C8-12.2.1:.

C8-12.1.1: This compliance statement is obsolete and has been replaced by C8-12.2.2:.

C8-12.2.1: All 4x-SX Optical Ports **shall** comply with the Eye Safety, Transmitter, and Receiver requirements in Section 8.5.10.1 and Section 8.5.10.2.

C8-12.2.2: All 4x-LX Optical Ports **shall** comply with the Eye Safety, Transmitter, and Receiver requirements in Section 8.5.10.3 and Section 8.5.10.4

A 4x-SX link operates in the 850nm wavelength band using multimode (MM) fiber. The optical parameters have been selected so as to allow typical 4x-SX Optical Transceivers to use an array of GaAs VCSELs and an

array of photodetectors, trans-impedance pre-amplifiers and limiting post-amplifiers. Three fiber types are specified for 4x-SX:

i) 4x-SX/50: 500MHz.km 50µm / 125µm MM fiber
ii) 4x-SX/50: 2000MHz.km 50µm / 125µm MM fiber
iii) 4x-SX/62: 200MHz.km 62.5µm / 125µm MM fiber

8.5.10.1 4X-SX EYE SAFETY

The optical power emitted from a 4x-SX port shall be limited to a maximum value for Class 1M laser safety in accordance with IEC/EN 60825-1 Amendment 2 Safety of Laser Products, part 1: Equipment classification, requirements, and user's guide and FDA/CDRH 21 CFR 1040.10

For systems that are required to meet existing IEC Class I laser safety regulations, Open Fiber Control (OFC) or other similar vendor-specific protocols **should** be implemented to meet the CDRH and IEC requirement. Any such implementation **shall** be transparent to all InfiniBand layers.

Recommendation to System Designer

Without any form of open fiber control, the power levels specified for the 4x-SX Optical Transmitter will meet the new relaxed specification for IEC Class 1M.

8.5.10.2 4X-SX OPTICAL PARAMETERS

Table 75 gives the link budgets for 4x-SX multimode fiber optic links running at 2.5 GTransfers/second. Fiber plant specifications are described in Section 8.7.

Table 75 Link Parameters - 4x-SX

Parameter	IB-4x-SX/50		IB-4x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Baud Rate		2500		2500	Mb/s	
Rate tolerance		±100		±100	ppm	

Table 75 Link Parameters - 4x-SX (Continued)

Parameter	IB-4x-SX/50		IB-4x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Optical Passive Loss		1.9		1.8	dB	
Optical System Penalty		2.9		3.0	dB	
Total link power budget	4.8		4.8		dB	
Worst case operating range		2 - 125[a] 2 - 200[b]		2 - 75	m	c
Fiber mode-field (core) diameter		50		62.5	µm	

a. 4x-SX/50: 500MHz.km 50µm / 125µm MM fiber
b. 4x-SX/50: 2000MHz.km 50µm / 125µm MM fiber
c. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 4x-SX link **shall** meet the parameters specified in Table 76 at TP2.

If any high-speed electrical input signals are less than the V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s then the optical modulation amplitude **shall not** exceed 5.0 µW.

Table 76 Optical Transmitter Parameters - 4x-SX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	830	860	nm	
RMS spectral width		0.85	nm	
Average launched power		-2.5	dBm	a
Optical Modulation Amplitude (OMA)	0.150		mW	b
RMS mean of 20% - 80% Rise/Fall time		150	ps	c
RIN_{12} (OMA)		-117	dB/Hz	

a. Average launched power per fiber, max. is the lesser of the eye safety limit or Average receiver power, max.
b. Optical modulation amplitude values are peak-to-peak.
c. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 4x-SX link **shall** meet the parameters specified in Table 77 at TP3.

If the average received optical power on any Lane is less than -26 dBm then the corresponding high-speed electrical signaling output **shall** be squelched to less than V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s. The electrical outputs of every lane with a optical power greater than -18 dBm **shall not** be squelched.

Table 77 Optical Receiver Parameters - 4x-SX

Parameter	Minimum	Maximum	Units	Note
Average received power		-1.5	dBm	
Optical modulation amplitude	0.050		mW	a
Return loss of receiver	12		dB	
Stressed receiver sensitivity (OMA)	0.085		mW	a
Stressed receiver ISI test	0.90		dB	
Stressed receiver DCD component of DJ (at TX)	60		ps	
Receiver electrical 3 dB upper cutoff frequency		2.8	GHz	
Receiver electrical 10 dB upper cutoff frequency		6.0	GHz	

a. Optical modulation amplitude values are peak-to-peak.

8.5.10.3 4x-LX EYE SAFETY

The 4x LX Optical Port **shall** meet the Eye safety as defined for 10GBASE-LR in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**.

8.5.10.4 4x-LX OPTICAL PARAMETERS

The 4x-LX Optical Port **shall** meet the Optical Parameters as defined for 10GBASE-LR in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. 4x LX **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed for 4x-LX shall be 10.0 GBaud ±100ppm

8.5.11 4X-DDR-SX LINK

C8-12.2.3: All 4x-DDR-SX Optical Ports **shall** comply with the Eye Safety, Transmitter, and Receiver requirements in Section 8.5.11.1 and Section 8.5.11.2.

A 4x-DDR-SX link operates in the 850nm wavelength band using multi-mode (MM) fiber. The optical parameters have been selected so as to allow typical 4x-DDR-SX Optical Transceivers to use an array of GaAs VCSELs and an array of photodetectors, trans-impedance pre-amplifiers and limiting post-amplifiers. Three fiber types are specified for 4x-DDR-SX:

i) 4x-SX/50: 500MHz.km 50µm / 125µm MM fiber
ii) 4x-SX/50: 2000MHz.km 50µm / 125µm MM fiber
iii) 4x-SX/62: 200MHz.km 62.5µm / 125µm MM fiber

8.5.11.1 4X-DDR-SX EYE SAFETY

The optical power emitted from a 4x-DDR-SX port **shall** be limited to a maximum value for Class 1M laser safety in accordance with IEC/EN 60825-1 Amendment 2 Safety of Laser Products, part 1: Equipment classification, requirements, and user's guide and FDA/CDRH 21 CFR 1040.10

For systems that are required to meet existing IEC Class I laser safety regulations, Open Fiber Control (OFC) or other similar vendor-specific protocols **should** be implemented to meet the CDRH and IEC requirement. Any such implementation **shall** be transparent to all InfiniBand layers.

Recommendation to System Designer

Without any form of open fiber control, the power levels specified for the 4x-DDR-SX Optical Transmitter will meet the new relaxed specification for IEC Class 1M.

8.5.11.2 4X-DDR-SX OPTICAL PARAMETERS

Table 78 gives the link budgets for 4x-DDR-SX multimode fiber optic links running at 5.0 GTransfers/second. Fiber plant specifications are described in Section 8.7.

Table 78 Link Parameters - 4x-DDR-SX

Parameter	IB-4x-SX/50		IB-4x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Baud Rate		5000		5000	Mb/s	
Rate tolerance		±100		±100	ppm	
Optical Passive Loss		1.78[a] 2.06[b]		1.70	dB	
Optical System Penalty		2.5[c] 2.3[d]		3.2	dB	
Total link power budget	6.25		6.25		dB	
Worst case operating range		2 - 75[e] 2 - 150[f]		2 - 50	m	g
Fiber mode-field (core) diameter		50		62.5	µm	

a. 4x-SX/50: 500MHz.km 50µm / 125µm MM fiber
b. 4x-SX/50: 2000MHz.km 50µm / 125µm MM fiber
c. 4x-SX/50: 500MHz.km 50µm / 125µm MM fiber
d. 4x-SX/50: 2000MHz.km 50µm / 125µm MM fiber
e. 4x-SX/50: 500MHz.km 50µm / 125µm MM fiber
f. 4x-SX/50: 2000MHz.km 50µm / 125µm MM fiber
g. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 4x-DDR-SX link **shall** meet the parameters specified in Table 79 at TP2.

Table 79 Optical Transmitter Parameters - 4x-DDR-SX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	830	860	nm	
RMS spectral width		0.85	nm	
Average launched power		-2.0	dBm	a

BTA

Table 79 Optical Transmitter Parameters - 4x-DDR-SX

Parameter	Minimum	Maximum	Units	Note
Optical Modulation Amplitude (OMA)	0.224		mW	b
RMS mean of 20% - 80% Rise/Fall time		75	ps	c
RIN_{12} (OMA)		-122	dB/Hz	

a. Average launched power per fiber, max. is the lesser of the eye safety limit or Average receiver power, max.
b. Optical modulation amplitude values are peak-to-peak.
c. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 4x-DDR-SX link **shall** meet the parameters specified in Table 80 at TP3.

Table 80 Optical Receiver Parameters - 4x-DDR-SX

Parameter	Minimum	Maximum	Units	Note
Average received power	-1.5		dBm	
Optical modulation amplitude (Informative)		0.053 Informative	mW	a
Return loss of receiver	12		dB	
Stressed receiver sensitivity (OMA)	0.078		mW	b
Stressed receiver ISI test	1.50		dB	
Stressed receiver DCD component of DJ (at TX)	16.3		ps	

a. Optical modulation amplitude values are peak-to-peak.
b. Stressed Rx entries for Sensitivity and ISI differ for each fiber type. The entries shown are for 50u, 500MHzkm fiber

8.5.12 8X-SX LINKS - AT 2.5 GB/S

C8-12.2.4: All 8x SX Optical Ports **shall** meet all the optical specifications for 12x-SX in Section 8.5.13 with the exception that 8x SX has only eight receive and transmit lanes.

8.5.13 12X-SX LINKS - AT 2.5 GB/S

C8-13: This compliance statement is obsolete and has been replaced by C8-13.2.1:.

C8-13.2.1: All 12x-SX Optical Ports shall comply with the Eye Safety, Transmitter, and Receiver requirements in Section 8.5.13.

A 12x-SX link operates in the 850nm wavelength band using multimode (MM) fiber. The optical parameters have been selected so as to allow typical 12x-SX Optical Transceivers to use an array of GaAs VCSELs and an array of photodetectors, trans-impedance pre-amplifiers and limiting post-amplifiers. Two fiber types are specified for 12x-SX:

i) 12x-SX/50 link: 500MHz.km 50µm / 125µm MM fiber
ii) 12x-SX/50 link: 2000MHz.km 50µm / 125µm MM fiber
iii) 12x-SX/62 link: 200MHz.km 62.5µm / 125µm MM fiber

8.5.13.1 12-SX EYE SAFETY

The optical power coupled into the fiber for IB-12x-SX **shall** be limited to a maximum value with Class IIIB laser safety operation in accordance with CDRH and EN 60825-1 Amendment 2 Radiation Safety of Laser Products: Equipment Classification, Requirements and User Guide.

For systems that are required to meet existing IEC Class I laser safety regulations, Open Fiber Control (OFC) or other similar vendor-specific protocols **should** be implemented to meet the CDRH and IEC requirement. Any such implementation **shall** be transparent to all InfiniBand layers.

Recommendation to System Designer

Without any form of open fiber control, the power levels specified for the 12x-SX Optical Transmitter will meet the proposed new relaxed specification for IEC Class 1M, which is currently in the process of ratifications. However, the InfiniBand Trade Association cannot guarantee when or if the new Class 1M power level will be ratified. Vendors **should** consider this when designing InfiniBand systems.

8.5.13.2 12X-SX OPTICAL PARAMETERS

Table 81 gives the link budgets for 12x-SX and 8x-SX multimode fiber optic links running at 2.5 GTransfers/second. Fiber plant specifications are described in Section 8.7.

Table 81 Link Parameters - 12x-SX and 8x-SX

Parameter	IB-12x-SX/50		IB-12x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Baud Rate		2500		2500	Mb/s	
Rate tolerance		±100		±100	ppm	
Optical Passive Loss		1.9		1.8	dB	
Optical System Penalty		2.9		3.0	dB	
Total link power budget	4.8		4.8		dB	
Worst case operating range		125[a] 200[b]		75	m	c
Fiber mode-field (core) diameter		50		62.5	µm	

a. 12x-SX/50 link: 500MHz.km 50µm / 125µm MM fiber
b. 12x-SX/50 link: 2000MHz.km 50µm / 125µm MM fiber
c. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 12x-SX or 8x-SX link **shall** meet the parameters specified in Table 82 at TP2.

If any high-speed electrical input signals are less than the V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s then the optical modulation amplitude **shall not** exceed 5.0 µW.

Table 82 Optical Transmitter Parameters - 12x-SX and 8x-SX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	830	860	nm	
RMS spectral width		0.85	nm	
Average launched power		-4	dBm	a
Optical modulation amplitude	0.150		mW	b
RMS mean of 20% - 80% Rise/Fall time		150	ps	c
RIN_{12} (OMA)		-117	dB/Hz	

a. Average launched power per fiber, max. is the lesser of the eye safety limit or Average receiver power, max.
b. Optical modulation amplitude values are peak-to-peak.

c. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 12x-SX link **shall** meet the parameters specified in Table 83 at TP3.

If the average received optical power on any Lane is less than -26 dBm then the corresponding high-speed electrical signaling output **shall** be squelched to less than V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s. The electrical outputs of every lane with a optical power greater than -18 dBm **shall not** be squelched.

Table 83 Optical Receiver Parameters - 12x-SX

Parameter	Minimum	Maximum	Units	Note
Average received power		-1.5	dBm	
Optical Modulation Amplitude (OMA)	0.050		mW	a
Return loss of receiver	12		dB	
Stressed receiver sensitivity (OMA)	0.085		mW	**a**
Stressed receiver ISI test	0.90		dB	
Stressed receiver DCD component of DJ (at TX)	60		ps	
Receiver electrical 3 dB upper cutoff frequency		2.8	GHz	
Receiver electrical 10 dB upper cutoff frequency		6.0	GHz	

a. Optical modulation amplitude values are peak-to-peak.

8.5.14 8x-DDR-SX AND 12x-DDR-SX LINKS

C8-13.2.2: All 8x-SX-DDR and 12x-SX-DDR Optical Ports shall comply with the Eye Safety, Transmitter, and Receiver requirements in Section 8.5.14

A 8x-DDR-SX and 12x-DDR-SX links operates in the 850nm wavelength band using multimode (MM) fiber. The optical parameters have been selected so as to allow typical 8x-DDR-SX and 12x-DDR-SX Optical Transceivers to use an array of GaAs VCSELs and an array of photodetectors,

trans-impedance pre-amplifiers and limiting post-amplifiers. Three fiber types are specified for 8x-DDR-SX and 12x-DDR-SX:

i) 12x-SX/50 link: 500MHz.km 50µm / 125µm MM fiber
ii) 12x-SX/50 link: 2000MHz.km 50µm / 125µm MM fiber
iii) 12x-SX/62 link: 200MHz.km 62.5µm / 125µm MM fiber

8.5.14.1 8x-DDR-SX and 12x-DDR-SX Eye Safety

The optical power coupled into the fiber for 8x-DDR-SX and 12x-DDR-SX **shall** be limited to a maximum value with Class 1M laser safety operation in accordance with CDRH and EN 60825-1 Radiation Safety of Laser Products: Equipment Classification, Requirements and User Guide.

For systems that are required to meet existing IEC Class I laser safety regulations, Open Fiber Control (OFC) or other similar vendor-specific protocols **should** be implemented to meet the CDRH and IEC requirement. Any such implementation **shall** be transparent to all InfiniBand layers.

8.5.14.2 8x-DDR-SX and 12x-DDR-SX Optical Parameters

Table 84 gives the link budgets for 8x-DDR-SX and 12x-DDR-SX multimode fiber optic links running at 5.0 GTransfers/second. Fiber plant specifications are described in Section 8.7.

Table 84 Link Parameters - 8x-DDR-SX and 12x-DDR-SX

Parameter	IB-12x-SX/50		IB-12x-SX/62		Units	Note
	Minimum	Maximum	Minimum	Maximum		
Baud Rate		5000		5000	Mb/s	
Rate tolerance		±100		±100	ppm	
Optical Passive Loss		1.78[a] 2.06[b]		1.68	dB	
Optical System Penalty		2.6[c] 2.5[d]		3.1	dB	
Total link power budget	6.25		6.25		dB	
Worst case operating range		2-75[e] 2-150[f]		2-45	m	g
Fiber mode-field (core) diameter		50		62.5	µm	

a. 12x-SX/50 link: 500MHz.km 50µm / 125µm MM fiber

b. 12x-SX/50 link: 2000MHz.km 50µm / 125µm MM fiber
c. 12x-SX/50 link: 500MHz.km 50µm / 125µm MM fiber
d. 12x-SX/50 link: 2000MHz.km 50µm / 125µm MM fiber
e. 12x-SX/50 link: 500MHz.km 50µm / 125µm MM fiber
f. 12x-SX/50 link: 2000MHz.km 50µm / 125µm MM fiber
g. Longer operating distance than the range specified here can be achieved using transmitters, receivers and/or cables meeting specification but performing better than worst-case.

An Optical Transmitter for a 8x-DDR-SX and 12x-DDR-SX link **shall** meet the parameters specified in Table 85 at TP2.

Table 85 Optical Transmitter Parameters - 8x-DDR-SX and 12x-DDR-SX

Parameter	Minimum	Maximum	Units	Note
Type		Laser		
Center Wavelength	830	860	nm	
RMS spectral width		0.85	nm	
Average launched power		-2.0	dBm	a
Optical modulation amplitude	0.224		mW	b
RMS mean of 20% - 80% Rise/Fall time		75	ps	c
RIN_{12} (OMA)		-122	dB/Hz	

a. Average launched power per fiber, max. is the lesser of the eye safety limit or Average receiver power, max.
b. Optical modulation amplitude values are peak-to-peak.
c. Optical rise and fall time specifications are based on unfiltered waveforms. See Section 8.5.3.2.

An Optical Receiver for a 12x-DDR-SX link **shall** meet the parameters specified in Table 83 at TP3.

Table 86 Optical Receiver Parameters - 12x-DDR-SX and 8x-DDR-SX

Parameter	Minimum	Maximum	Units	Note
Average received power	-1.5		dBm	
Optical Modulation Amplitude (OMA) (Informative)		0.053 Informative	mW	a
Return loss of receiver	12		dB	
Stressed receiver sensitivity (OMA)	0.078		mW	**a**, b
Stressed receiver ISI test	1.50		dB	2
Stressed receiver DCD component of DJ (at TX)	16.3		ps	

a. Optical modulation amplitude values are peak-to-peak.
b. Stressed Rx entries for Sensitivity and ISI differ for each fiber type. The entries shown are for 50u, 500MHzkm fiber

8.5.15 1X QDR LINKS

C8-13.2.3: All 1x-QDR-SX Optical Ports **shall** comply with the Eye Safety, Transmitter, and Receiver requirements in Section 8.5.15.1 and Section 8.5.15.2

C8-13.2.4: All 1x-QDR-LX Optical Ports **shall** comply with the Eye Safety, Transmitter, and Receiver requirements in Section 8.5.15.1 and Section 8.5.15.3

A 1x-QDR-SX link operates in the 850nm wavelength band using multi-mode (MM) fiber. The optical parameters have been selected so as to allow typical 1x-SX Optical Transceivers to use a GaAs Vertical Cavity Surface Emitting Laser (VCSEL) and a photodetector, trans-impedance pre-amplifier and limiting post-amplifier. Three fiber types are specified for 1x-QDR-SX:

i) 1x-SX/50 link: 500MHz.km 50μm / 125μm MM fiber
ii) 1x-SX/50 link: 2000MHz.km 50μm / 125μm MM fiber
iii) 1x-SX/62 link: 200MHz.km 62.5μm / 125μm MM fiber

A 1x-QDR-LX link operates in the 1300nm wavelength band using sin-glemode (SM) fiber. The optical parameters have been selected to allow a typical 1x-LX Optical Transceiver to use an uncooled InP-based Fabry-Perot (FP) laser, Distributed Feedback (DFB) laser or VCSEL laser, and an InGaAs photodetector, trans-impedance pre-amplifier and limiting post-amplifier. A trade-off curve is provided in Figure 65 between Center

Wavelength and RMS Spectral Width to allow designers to select appropriate laser technology while still ensuring link operation. Only one fiber type is specified for 1x-LX: single-mode non-dispersion shifted.

8.5.15.1 1X QDR EYE SAFETY

C8-13.2.5: The optical power coupled into the fiber for 1x-QDR-SX and 1x-QDR-LX **shall** be limited to a maximum value with Class I laser safety operation in accordance with CDRH and IEC 60825-1 Radiation Safety of Laser Products: Equipment Classification, Requirements and User Guide.

8.5.15.2 1X-QDR-SX OPTICAL PARAMETERS

The 1x QDR SX Optical Port shall meet the Optical Parameters as defined for 10GBASE-SR in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. 1x QDR SX **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed for 1x-QDR-SX shall be 10.0 GBaud ±100ppm

8.5.15.3 1X-QDR-LX OPTICAL PARAMETERS

The 1x LX QDR Optical Port shall meet the Optical Parameters as defined for 10GBASE-LR in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. 1x LX QDR **shall** use the same transmit and receive optical interface and budget specifications and measurements, with the following notes and exceptions:

- The nominal signaling speed for 1x-QDR-LX shall be 10.0 GBaud ±100ppm

8.6 OPTICAL RECEPTACLE AND CONNECTOR

The primary function of the optical fiber connector specification is to define mechanical alignment of the optical fibers to the optical port of an Optical Transceiver.

The objective of this section is to specify the optical interface sufficiently to ensure the following:

a) Intermateability
b) Mechanical/Optical Performance
c) Maximum Supplier Flexibility

In this section, only the dimensions necessary to specify the transmitter-receiver center-center distance of the 12x-SX connector are provided. All other dimensions are included by reference to other standards.

C8-14: All Optical Transceivers **shall** present the Optical Receptacle specified in Optical Receptacle and Connector (Section 8.6) through the system bulkhead. Typically, this bulkhead will be a connector housing.

Optical Receptacles, Optical Connectors, and Fibre Optic Adapters **shall** be as specified in this section.

The InfiniBand connector for 1x, 1x DDR and 1x QDR links **shall** be a duplex LC connector, for 4x SX and 4x SX DDR links **shall** be an MPO connector, for 4x LX links **shall** be a dual-SC connector and for 12x, 8x SX DDR and 12x SX DDR links **shall** be a dual-MPO connector.

8.6.1 1X CONNECTOR - LC

The connectors, adapters and receptacles described here **shall** be fully duplex, creating fully bi-directional optical connection with one mating action. Figure 78 and Figure 79 show outline drawings of the Duplex LC Optical Connector and Duplex LC Fiber Optic Adapter respectively.

The 1x Optical Connector defined by this specification **shall** conform to ANSI/TIA/EIA 604-10 (FOCIS 10), Fiber Optic Connector Intermateability Standard, Type “LC”.

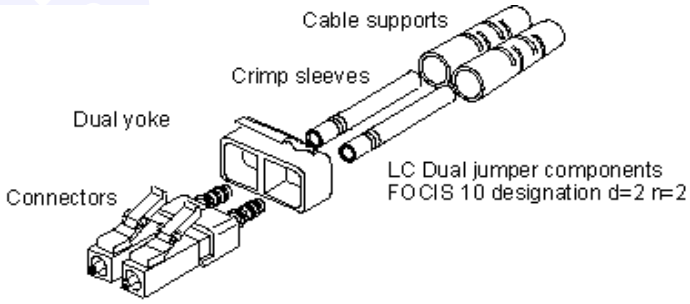
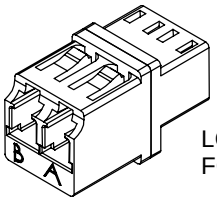


Figure 78 Dual LC Plug



LC Dual Adapter
FOCIS 10 designation n=2 m=2 (shown)

Figure 79 Dual LC Fiber Optic Adapter

8.6.1.1 1X FIBER OPTIC CONNECTOR

C8-15: All 1x Optical Cable Connectors shall comply with connector specifications of Section 8.6.1.1.

The InfiniBand connector for 1x links **shall** be a duplex LC connector.

The 1x Optical Connector on each end of the Fiber Optic Cable **shall** conform to:

i) ANSI/TIA/EIA 604-10 (FOCIS 10), Fiber Optic Connector Intermateability Standard, Type "LC", and

ii) Fibre Channel Physical Interface standard (FC-PI), revision 8.0

The implementation of connectors or adapters compliant with the FOCIS 10 standard **shall not** preclude the intermateability of connectors, adapters or receptacles compliant with the FOCIS 10A standard.

8.6.1.2 1X FIBER OPTIC RECEPTACLE

C8-16: All 1x Optical Port Receptacles shall comply with receptacle and fiber orientation specifications of Section 8.6.1.2 and Section 8.6.1.3.

The 1x (SX & LX) Optical Transceiver Port **shall** be a Type “LC” receptacle as defined by ANSI/TIA/EIA 604-10 (FOCIS 10), Fiber Optic Connector Intermateability Standard.

It **may** contain resilient sleeves to optically align the connector plug ferrules. The positioning of the ferrule endfaces to optimize optical coupling can be accomplished in a variety of ways not described here.

The implementation of connectors, adapters or receptacles compliant with the FOCIS 10 standard **shall not** preclude the intermateability of connectors, adapters or receptacles compliant with the FOCIS 10A standard.

8.6.1.3 1X FIBER OPTIC RECEPTACLE ORIENTATION

1x-SX and 1x-LX FIber Optic Transceivers **shall** follow the Transmit/Receive convention detailed in Figure 80. If the Optical Connector is orientated such that the keying features on the LC housings are at the top, then looking into the Fiber Optic Receptacle the fiber on the left **shall** be used for optical transmit and the fiber on the right **shall** be used for optical receive.

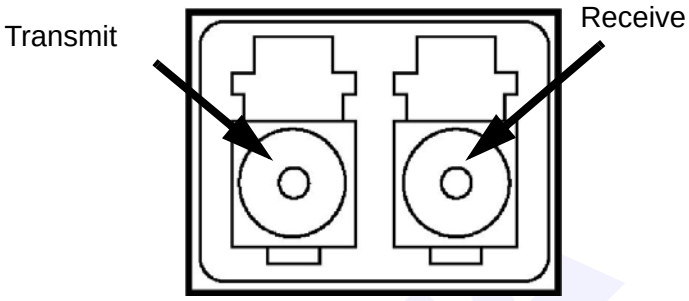


Figure 80 1x-SX and 1x-LX Optical Receptacle orientation looking into the Optical Transceiver

8.6.1.4 1X COLOR

C8-17: The 1x-SX multimode Optical Connector, Adapter and Receptacle, or a visible portion of these, **shall** be beige in color.

C8-18: The 1x-LX singlemode Optical Connector, Adapter and Receptacle, or a visible portion of these, **shall** be blue in color.

8.6.2 4X-SX CONNECTOR - SINGLE MPO

The connectors, adapters and receptacles described here **shall** be fully duplex, creating fully bi-directional optical connection with one mating action.

8.6.2.1 4X-SX FIBER OPTICAL CONNECTOR

C8-19: All 4x SX Optical Cable Connectors shall comply with fiber optical connector specifications of Section 8.6.2.1.

The 4x-SX Optical Connector on each end of the Fiber Optic Cable **shall** consist of a female MPO plug. The female MPO plug **shall** conform to IEC 1754-7-4, Push/Pull MPO Female Plug Connector Interface, and **shall** contain a female MT ferrule.

A female MT ferrule is similar to a male MT ferrule, except that the female ferrule does not have alignment pins. Instead, alignment holes are provided, which accept the alignment pins of the corresponding male MT ferrule. In this manner, precision alignment is achieved between the ferrule in an Optical Receptacle and the ferrule in an Optical Connector.

Figure 81 shows an outline drawing of a typical MPO connector with push-pull coupling mechanism. The MPO housing surrounds a rectangular male MT ferrule. The male MT ferrule has positions for 12 fibers. The male MT ferrule is typically 6.4 mm by 2.5 mm and contains two precision alignment pins of 0.7 mm diameter.

8.6.2.2 4X-SX FIBER OPTICAL RECEPTACLE

C8-20: All 4x SX Optical Port Receptacles shall comply with receptacle and fiber optical orientation specifications of Section 8.6.2.2 and Section 8.6.2.3

The 4x-SX Optical Transceiver Port **shall** have a 4x-SX Optical Receptacle which **shall** be a male MPO receptacle. The male MPO receptacle **shall** have two fixed pins conforming to IEC 1754-7-5, and **shall** conform to IEC 1754-7-3, Push/Pull MPO Adapter Interface standard.

Note: In the U.S.A., the MPO connector is also known as the MTP connector.

8.6.2.3 4X-SX FIBER OPTIC RECEPTACLE ORIENTATION

4x-SX Optical Transceivers **shall** follow the Transmit/Receive convention detailed in Figure 82. If the Optical Connector is orientated such that the keying feature on the MPO housing is at the top, then looking into the Fiber Optic Receptacle fibers are numbered left-to-right as 0-through-11. The 4 fiber positions on the left (fibers 0, 1, 2, 3) **shall** be used for optical transmit and the 4 fiber positions on the right (fibers 8, 9, 10, 11) **shall** be used for optical receive. Fibers 0, 1, 2, 3 **shall** carry transmit Lanes 0, 1, 2, 3 respectively. Fibers 8, 9, 10, 11 **shall** carry receive Lanes 3, 2, 1, 0 respectively.

The central four fibers (fibers 4, 5, 6, 7) **may** be physically present. If one or more of the central four fibers is present, then it **shall not** be used to carry IB signals.

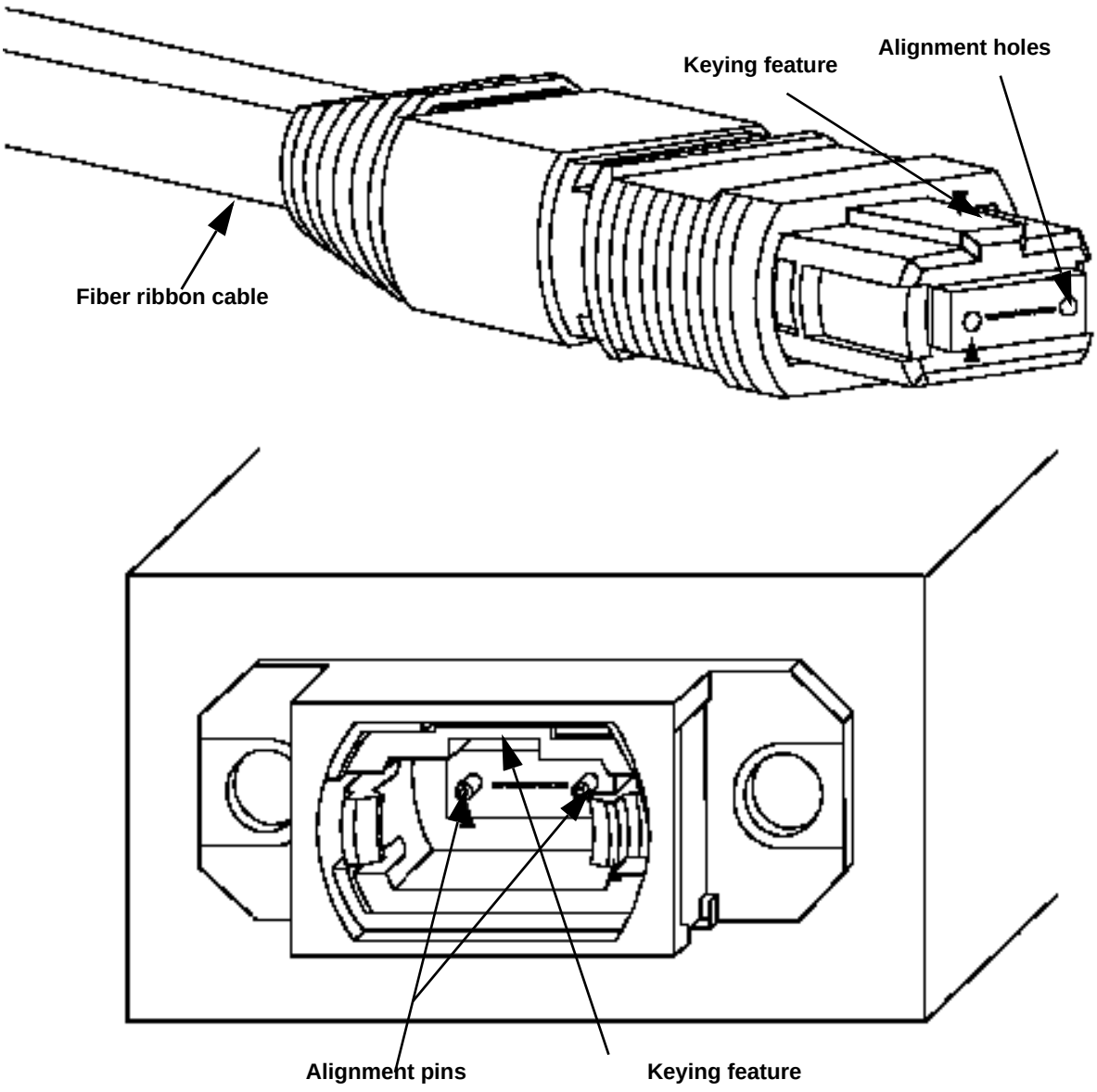


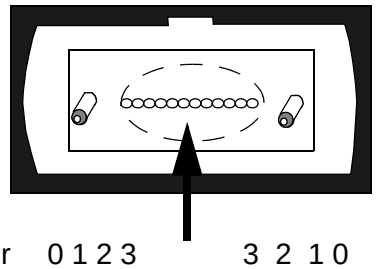
Figure 81 MPO Plug and Receptacle

8.6.3 4X LX CONNECTOR - SC

The connectors, adapters and receptacles described here **shall** be fully duplex, creating fully bi-directional optical connection with one mating action.

8.6.3.1 4X LX FIBER OPTIC CONNECTOR

C8-20.1.1: All 1x Optical Cable Connectors shall comply with connector specifications of Section 8.6.3.1.



Transmit Lane number	0 1 2 3	3 2 1 0 Receive Lane Number
Fiber number	0 1 2 3 4 5 6 7 8 91011	
Used (o) or not (x)	o o o o x x x x o o o o	

Figure 82 4x-SX Optical Receptacle orientation looking into the Transceiver

The InfiniBand connector for 4x LX links **shall** be a duplex SC connector.

The 4x LX Optical Connector on each end of the Fiber Optic Cable **shall** conform to the requirements of IEC 61754-4.

Only the Floating Duplex style Connector Plug **shall** be used. Rigid SC Duplex connector **shall not** be used.

8.6.3.2 4X LX FIBER OPTIC RECEPTACLE

C8-20.1.2: All 1x Optical Port Receptacles shall comply with receptacle and fiber orientation specifications of Section 8.6.3.2 and Section 8.6.3.3.

The 4x LX Optical Transceiver Port **shall** be a Type “SC” receptacle as defined by IEC 61754-4-5.

8.6.3.3 4X LX FIBER OPTIC RECEPTACLE ORIENTATION

4x LX Fiber Optic Transceivers **shall** follow the Transmit/Receive convention detailed in Figure 83. If the Optical Connector is orientated such that the keying features on the SC housings are at the Bottom, then looking into the Fiber Optic Receptacle the fiber on the right **shall** be used for optical transmit and the fiber on the left **shall** be used for optical receive.

8.6.4 8X-SX OPTICAL RECEPTACLE AND CONNECTOR

C8-20.2.1: All 8x SX Optical Cable Connectors and 8x Optical Port Receptacles shall comply with the 12x-SX connector and receptacle specifications in Section 8.6.4 and Section 8.6.5 with the exception of the following exception for Fiber Optic Receptacle Orientation:

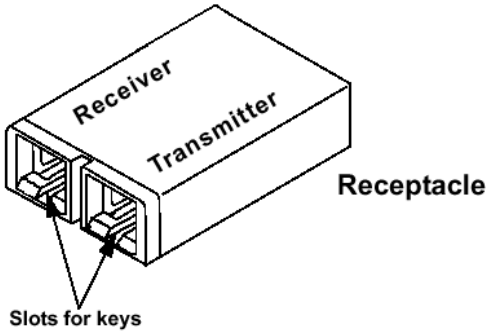


Figure 83 4x-LX Optical Receptacle orientation

8x-SX Optical Transceivers **shall** follow the transmit/receive convention detailed in Figure 84. If the Optical Connector is orientated such that the keying features on the MPO housings are at the top, then looking into the Optical Receptacle the fibers are numbered left-to-right as 0-through-11 for the left-hand section and 0-through-11 for the right hand section. The 8 fiber positions on the left **shall** be used for optical transmit and the 8 fiber positions on the right **shall** be used for optical receive. On the left, fibers 0, 1, 2, 3, 4, 5, 6, 7 **shall** carry transmit Lanes 0, 1, 2, 3, 4, 5, 6, 7 respectively. On the right, fibers 4, 5, 6, 7, 8, 9, 10, 11 **shall** carry receive Lanes 7, 6, 5, 4, 3, 2, 1, 0 respectively. On the left fibers 8, 9, 10, 11 **may** be present but **shall not** be used to carry IB signals. On the right fibers 0, 1, 2, 3 **may** be present but **shall not** be used to carry IB signals.

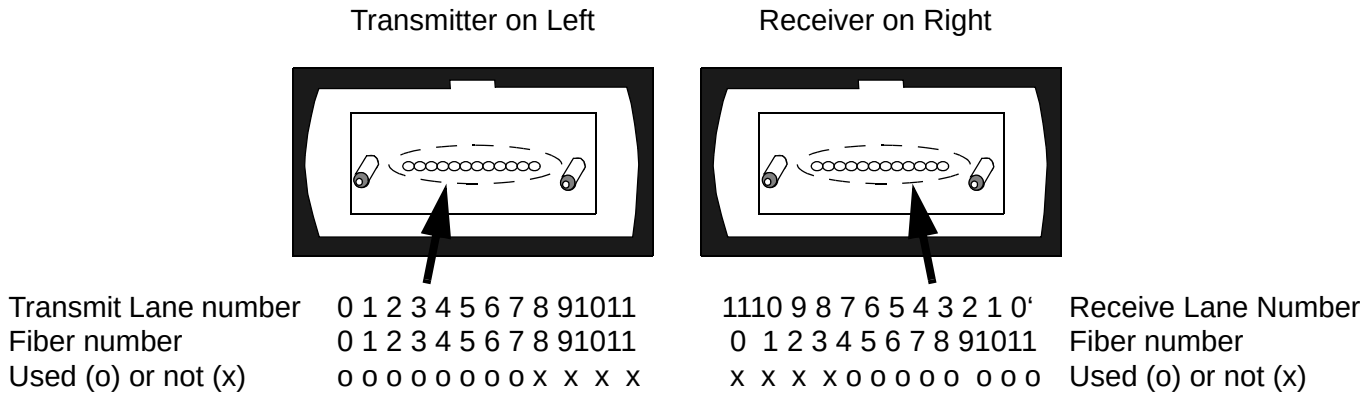


Figure 84 8x-SX Optical Receptacle orientation looking into the Transceiver

8.6.5 12X-SX CONNECTOR - DUAL MPO

The connectors, adapters and receptacles described here **shall** be fully duplex, creating fully bi-directional optical connection.

8.6.5.1 12X-SX FIBER OPTIC CONNECTOR

C8-21: All 12x-SX Optical Cable Connectors shall comply with connector specifications of Section 8.6.5.1.

The 12x-SX Optical Connector on each end of the Fiber Optic Cable **shall** consist of a double female MPO plug. Each female plug **shall** conform to IEC 1754-7-4, Push/Pull MPO Female Plug Connector Interface, and **shall** contain a female MT ferrule.

8.6.5.2 12X-SX FIBER OPTIC RECEPTACLE

C8-22: This compliance statement is obsolete and has been replaced by C8-22.2.1:

C8-22.2.1: All 12x-SX Optical Port Receptacles shall comply with receptacle and fiber optical orientation specifications of Section 8.6.5.2 and Section 8.6.2.3

The 12x-SX Optical Transceiver Port **shall** have a 12x-SX Optical Receptacle which **shall** consist of a double male MPO receptacle. Each male MPO receptacle **shall** be as described in Section 8.6.2, **shall** have two fixed pins conforming to IEC 1754-7-5, and **shall** conform to IEC 1754-7-3, Push/Pull MPO Adapter Interface standard.

A 12x Optical Transceiver **may** consist of physically separate Transmit and Receive Modules or a single physical Transceiver.

If a 12x solution is based on separate transmit and receive modules, then these modules **shall** support a centerline to centerline receptacle spacing of 20.0 mm +/-0.5 mm as defined in Figure 85.

If a 12x solution is based on a single transceiver module, then these modules **shall** support a centerline to centerline receptacle spacing of 16.0 mm +/- 0.1mm as defined in Figure 85.

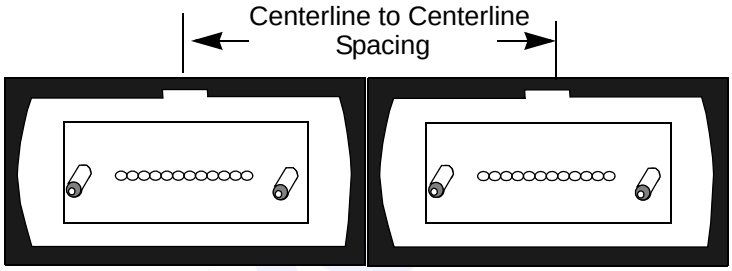


Figure 85 Double MPO Optical Receptacle Configuration

Implementation Note

To simplify cable management, in typical installations a duplex Fiber Optic Cable will be mated to a 12x-SX Optical Transceiver in a single mating action. This reduces the probability of mis-configuration. A plastic clip would serve to join two individual MPO housings together to form a double-MPO connector. Alternatively, a common housing could be moulded to house two sets of MPO actuation mechanisms plus two MT ferrules that float independently. In all cases, the two MT ferrules would float independently of each other to some extent.

The Optical Transceiver within the IB module could be constructed either as one transceiver component. Alternatively it could be constructed as separate Optical Transmitter and Optical Receiver components mounted to the board independently, but mutually aligned to within the total travel of the two MT ferrules.

8.6.5.3 12X-SX FIBER OPTIC RECEPTACLE ORIENTATION

12x-SX Optical Transceivers **shall** follow the transmit/receive convention detailed in Figure 86. If the Optical Connector is orientated such that the

keying features on the MPO housings are at the top, then looking into the Optical Receptacle the fibers are numbered left-to-right as 0-through-11 for the left-hand section and 0-through-11 for the right hand section. The 12 fiber positions on the left **shall** be used for optical transmit and the 12 fiber positions on the right **shall** be used for optical receive. On the left, fibers 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 **shall** carry transmit Lanes 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 respectively. On the right, fibers 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 **shall** carry receive Lanes 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0 respectively.

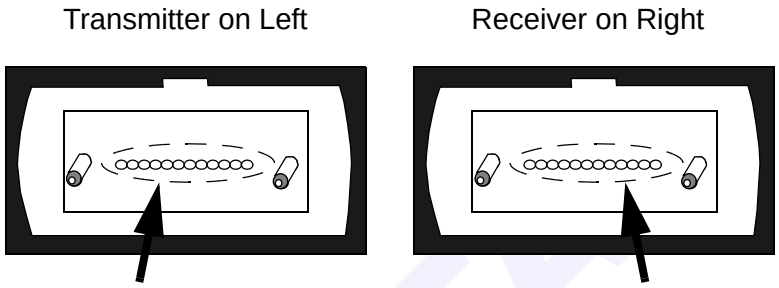


Figure 86 12x-SX and 8x-SX Optical Receptacle orientation looking into the Transceiver

8.7 FIBER OPTIC CABLE PLANT SPECIFICATIONS

8.7.1 OPTICAL FIBER SPECIFICATION

C8-23: This compliance statement is obsolete and has been replaced by C8-23.2.1:.

C8-23.2.1: All Optical Cables shall comply with the fiber cable requirements in Section 8.7.1 and Section 8.7.2 for respective Fiber Modes (Multimode or Single mode) as defined in Table 87.

C8-23.2.2: All 8x-SX Optical Ports **shall** meet all the optical cable plant specifications for 12x-SX in Section 8.7 with the exception that 8x-SX has only eight receive and transmit lanes.

Fiber Optic Cables for links in the -SX class **shall** use either 50/125 µm optical fiber or 62.5/125 µm multimode optical fiber compliant respectively with the "MMF 50/125 um" or "MMF 62.5/125 µm" specification in this section.

Fiber Optic Cables for links in the -LX class **shall** use non-dispersion-shifted single mode fiber compliant with the "SMF" specification in this section.

SMF **shall** conform to TIA/EIA-492CAAA-98 “Dispersion-Unshifted Single-Mode Optical Fibers”.

MMF 500 MHz.km 50/125 µm **shall** conform to TIA/EIA-492AAAB-98 "Detail Specification for 50-µm Core Diameter/125-µm Cladding Diameter Class Ia Graded-Index Multimode Optical Fibers" or IEC 60793-2 Type A1a.

MMF 2000 MHz.km 50/125 µm **shall** conform to TIA/EIA-492AAAC. TIA/EIA-492AAAC is presently in ballot.

MMF 62.5/125 µm **shall** conform to TIA/EIA-492AAAA-A-97 "Detail Specification for 62.5-µm Core Diameter/125-µm Cladding Diameter Class Ia Graded-Index Multimode Optical Fibers" or IEC 60793-2 Type A1b.

Recommendation to Optical System Designer

For new -SX installations, MMF 50/125 µm fiber **should** be used because of the increased range of 1x-SX/50 links compared to 1x-SX/62 links.

4x-SX, 8x-SX and 12x-SX Fiber Optic Cables **may** use ribbonized fibers.

SMF, MMF 50/125µm and MMF 62.5/125 µm fibers **shall** also conform to the respective columns of Table 87.

8.7.2 MODAL BANDWIDTH

The Modal Bandwidth for the 500 MHz.km and 200 MHz.km cables with Overfilled Launch specified in Table 87 is the worst case modal bandwidth, measured according to the methods of TIA2.2.1 working specification TIA/EIA/455-204-FOTP204 Measurement method for Multimode Fiber Bandwidth. The 2000MHz.km cable is measured using TIA/EIA-492AAAC. Worst case modal bandwidth is defined as the lowest bandwidth that can occur in a fiber under reasonable launch conditions.

In practice, worst case modal bandwidth is used to account for differences in multimode fiber bandwidth that can occur under restricted launch conditions relative to bandwidth observed using an overfilled launch condition. The Optical System Penalty limits presented in Section 8.5.8, Section 8.5.10 and Section 8.5.13 represent conservative calculations

based on an overfilled launch model. In practice, better link performance is typically expected.

Table 87 Optical Fiber Specifications

Parameter	SMF (9 μm)	MMF 50/125 μm	MMF 62.5/125 μm	Units
Nominal Fiber Specification Wavelength	1310	850	850	nm
Fiber Cable Attenuation (Max)	0.5	3.5	3.5	dB/km
Modal Bandwidth with Overfilled Launch (Min)	not applicable	500[a] and 2000[b]	200	MHz.km
Zero Dispersion Wavelength λ_0	$1300 \le \lambda_0 \le 1320$	$1295 \le \lambda_0 \le 1320$	$1320 \le \lambda_0 \le 1365$	nm
Zero Dispersion Slope S_0 (Max)	0.093	0.11 for $1300 \le \lambda_0 \le 1320$ and $0.001(\lambda_0 - 1190)$ for $1295 \le \lambda_0 \le 1300$	0.11 for $1320 \le \lambda_0 \le 1348$ and $0.001(1458 - \lambda_0)$ for $1348 \le \lambda_0 \le 1365$	ps/nm^2.km

a. Overfilled launch bandwidth per IEC 60793-1-41 or TIA/EIA 455-204.
b. Effective modal bandwidth for fiber meeting TIA /EIA-492AAAC

8.7.3 OPTICAL PASSIVE LOSS OF FIBER OPTIC CABLE

C8-24: This compliance statement is obsolete and has been replaced by C8-24.1.1:

C8-24.1.1: All optical cables shall comply with the passive loss specifications of Section 8.7.3

IBTA

Optical Passive Loss **shall** be not exceed than the values specified in Table 63, Table 66, Table 75 and Table 81 for 1x-SX, 1x-LX, 4x-SX and 12x-SX links respectively. Optical Passive Loss **shall not** exceed the values specified in Table 69, Table 72, Table 78, Table 84, and Table 84 for 1x-DDR SX, 1x-DDR LX, 4x-DDR SX, 8x-DDR-SX and 12x-DDR-SX links respectively. The loss of the fiber plant **shall** be verified by the methods of OFSTP-14A.The Optical Passive Loss of a Fiber Optic Cable is the sum of attenuation losses due to the fiber, Fiber Optic Adapters and splices.

Connection Insertion Loss for 4x-LX **shall not** exceed the values specified for 10GBASE-L optical cables in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. The loss of the fiber plant **shall** be verified by the methods defined in **ANSI/IEEE P802.3ae-2002 10Gb/s Ethernet**. The Connection Insertion Loss for 4x-LX of a Fiber Optic Cable is the sum of attenuation losses due to the fiber, Fiber Optic Adapters and splices.

A Fiber Optic Cable **may** contain one or more Fiber Optic Adapters and/or splices, provided that the total Optical Passive Loss conforms to the optical budget of this specification. In calculating worst case operating range values for Table 63, Table 66, Table 69, Table 72, Table 75, Table 78 and Table 81, a total loss budget of 1.5dB was assigned to Fiber Optic Adapters and splices. For 4x-LX, 10G Ethernet assumed a budget of 2.0dB for Fiber Optic Adapters and splices.

8.7.4 FIBER OPTIC ADAPTERS AND SPLICES

8.7.4.1 1X (SX & LX) FIBER OPTIC ADAPTERS AND SPLICES

C8-25: All 1x (SX & LX) Optical Cable adaptors and splices **shall** comply with Section 8.7.4.1

If the 1x Fiber Optic Cable consists of more than one Fiber Optic Segment, then the Fiber Optic Adapter used to join the Fiber Optic Segments **shall** conform to ANSI/TIA/EIA 604-10 (FOCIS 10), Fiber Optic Connector Intermateability Standard, Type "LC".

Fiber Optic Adapters and splices for 1x-SX (multimode) links **shall** have a return loss of 20dB minimum as measured by the methods of FOTP-107 or equivalent. Fiber Optic Adapters and splices for 1x-LX (single-mode) links **shall** have a return loss of 26dB minimum as measured by the methods of FOTP-107 or equivalent.

8.7.4.2 4X-SX FIBER OPTIC ADAPTERS AND SPLICES

If the 4x -SX Fiber Optic Cable consists of more than one Fiber Optic Segment, then Optical Adapters are used to join the Fiber Optic Segments. These Optical Adapters are not detailed in this specification.

C8-26: Fiber Optic Adapters and splices for 4x-SX links **shall** have a return loss of 20dB minimum as measured by the methods of FOTP-107 or equivalent.

8.7.4.3 4X-LX FIBER OPTIC ADAPTERS AND SPLICES

If the 4x -LX Fiber Optic Cable consists of more than one Fiber Optic Segment, then Optical Adapters are used to join the Fiber Optic Segments. These Optical Adapters are not detailed in this specification.

C8-26.1.1: Fiber Optic Adapters and splices for 4x-LX links **shall** have a return loss of 26dB minimum as measured by the methods of FOTP-107 or equivalent.

8.7.4.4 12X-SX FIBER OPTIC ADAPTERS AND SPLICES

If the 12x -SX Fiber Optic Cable consists of more than one Fiber Optic Segment, then Optical Adapters are used to join the Fiber Optic Segments. These Optical Adapters are not detailed in this specification.

C8-27: Fiber Optic Adapters and splices for 12-SX links **shall** have a return loss of 20dB minimum as measured by the methods of FOTP-107 or equivalent.

8.8 SIGNAL CONDITIONER IN OPTICAL TRANSCEIVER

C8-28: All Optical Ports shall comply with the Signal Conditioner requirements in Section 8.8

C8-28.2.1: All 8x SX Signal conditioners **shall** meet all the signal conditioner specifications for 12x-SX in Section 8.8 with the exception that 8x SX has only eight receive and transmit lanes.

8.8.1 MOTIVATION FOR SIGNAL CONDITIONER

At this time, optical functionality is relatively expensive compared to electronic functionality. The overall system cost is generally minimized by using optical components (lasers, Fiber Optic Cables and photoreceivers) that are as low performance as can be tolerated, while including relatively sophisticated signal conditioning and control functionality in the electrical domain in the Optical Transmitter and Optical Receiver. Hence, optical links in general, including those specified for InfiniBand, are generally designed to operate at higher jitter than the electrical interconnects found on printed circuit boards

The optical jitter specification of Section 8.5.5 is relaxed from the jitter specification for high-speed electrical signaling specified in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s. However, the elec-

trical inputs (Optical Transmitter side) and electrical outputs (Optical Receiver side) of an Optical Transceiver **shall** meet the high-speed signaling specification of Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s.

These high speed electrical signaling jitter limits should be compared to the optical jitter specification of Table 58 for the different links widths and distance classes.

If an Optical Transceiver does not perform signal conditioning and an IB Electrical Transceiver driver output is connected directly through a short board trace to the intermediate electrical interface TP1 in an Optical Transceiver, then the recommended input jitter specification will be exceeded at TP1. The use of appropriate pre-emphasis can reduce deterministic jitter (DJ), making it more likely that the system will meet the mandatory jitter compliance at TP2 and TP33.

There is a related problem on the Optical Receiver side, assuming that the jitter at the intermediate electrical interface TP4 meets the recommended typical value. The typical TP4 jitter exceeds the specified input jitter tolerance of an IB Electrical Transceiver. Hence the system may not work correctly.

8.8.2 SIGNAL CONDITIONER IMPLEMENTATION

This section describes possible implementations of the Optical Transceiver to include the signal conditioning functionality.

Figure 87, Figure 88, Figure 89, Figure 90 and Figure 91 show typical IB Optical Transceiver implementations for 1x, 1x-DDR-SX, 1x-QDR-SX, 4x-SX, 4x-DDR-SX, 4x-LX, 8x-DDR-SX, 12x-SX, and 12x-DDR-SX respectively. In each case an optional Signal Conditioner is shown. This Signal Conditioner **shall** be present in an Optical Transceiver if the optoelectronic components do not directly provide IB-compliant high-speed electrical signals.

The function of the Signal Conditioner is to convert between IB-compliant high-speed electrical signaling and the intermediate electrical interface of the optoelectronic components. The Signal Conditioner **should** be implemented as a repeater, IB-compliant Retiming Repeater, or an adaptive equalizer.

Implementation Note

The system designer may choose to integrate the Signal Conditioner with the optoelectronic components, or may choose to install it separately on the Board.

Retiming functionality generally dissipates considerable power. For thermal considerations it may be advantageous to physically separate the signal conditioner from the optoelectronic components. This is especially true for 4x-SX and 12-SX links. In this case the intermediate electrical interface requires careful design to ensure signal integrity.

Implementation Note

In some implementations it may be expedient for an Optical Transceiver not to present an IB-compliant electrical interface. For instance, the Signal Conditioner functionality could be collapsed within the SerDes block of a Switch chip to reduce system power and cost and to improve functional density. In this case the "Optical Transceiver" no longer presents an IB-compliant high-speed electrical signaling interface. Nonetheless the TP2 and TP3 optical specifications can still be met. Hence the overall IB Module and the optical ports may still be able to claim IB-compliance, since the non-compliant electrical interface is hidden within the IB Module.

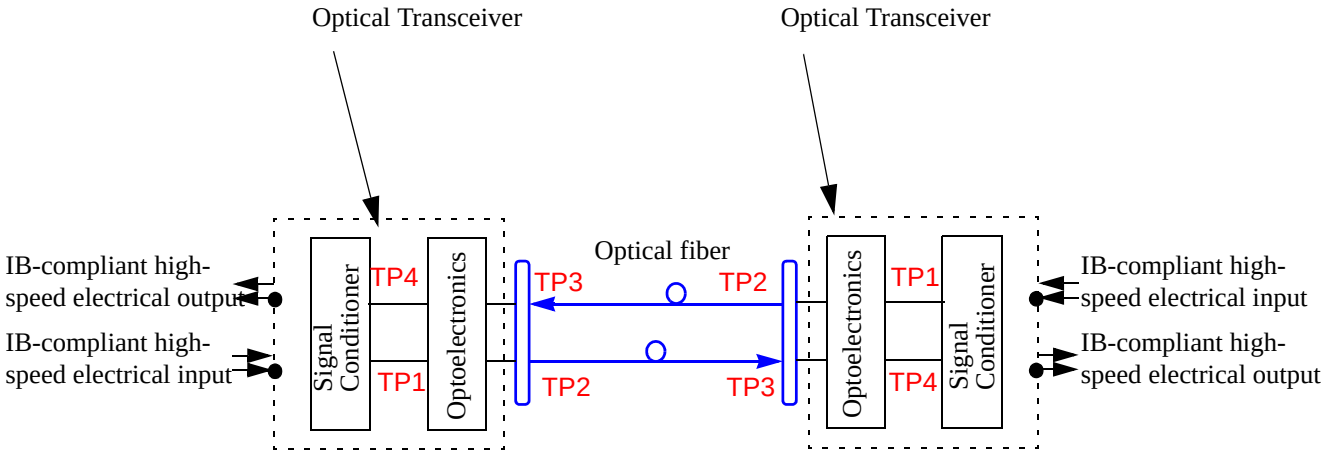


Figure 87 Recommended 1x, 1x DDR, and 1x QDR Optical Link Implementation

IBTA

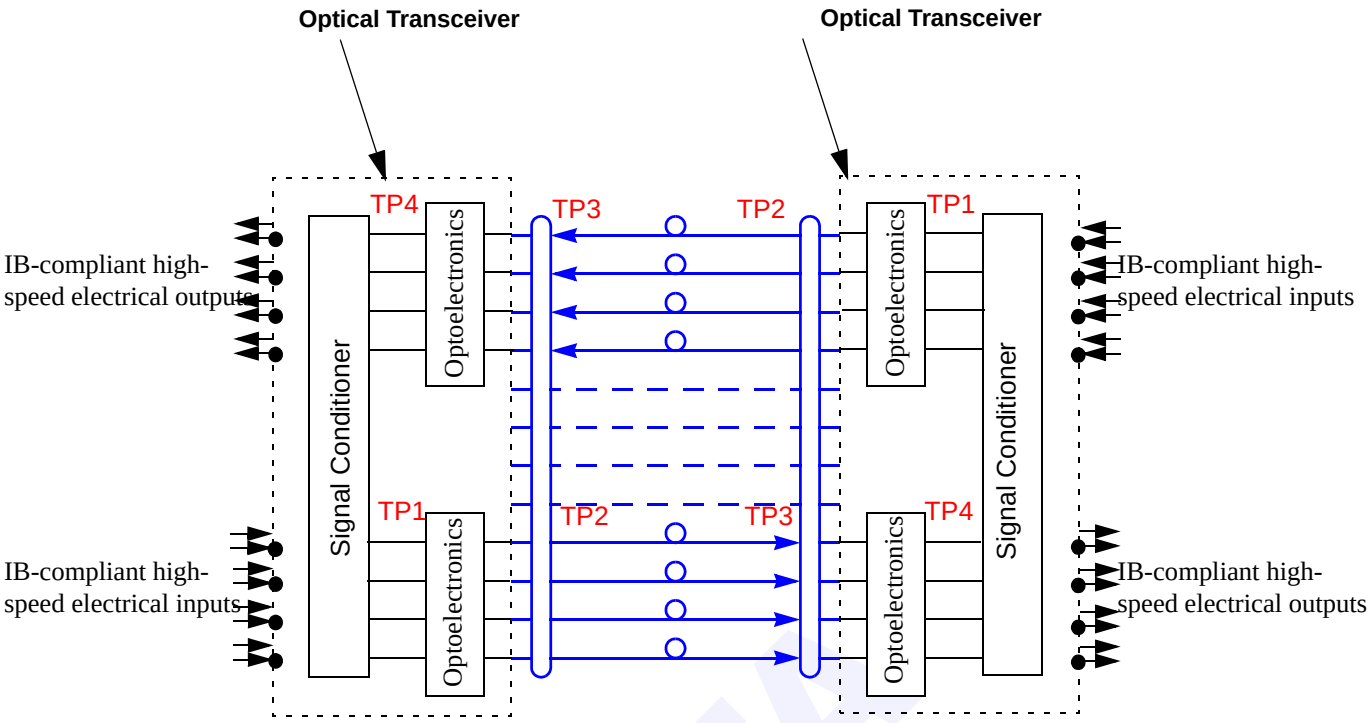


Figure 88 Recommended 4x-SX & 4x-DDR-SX Optical Link Implementation

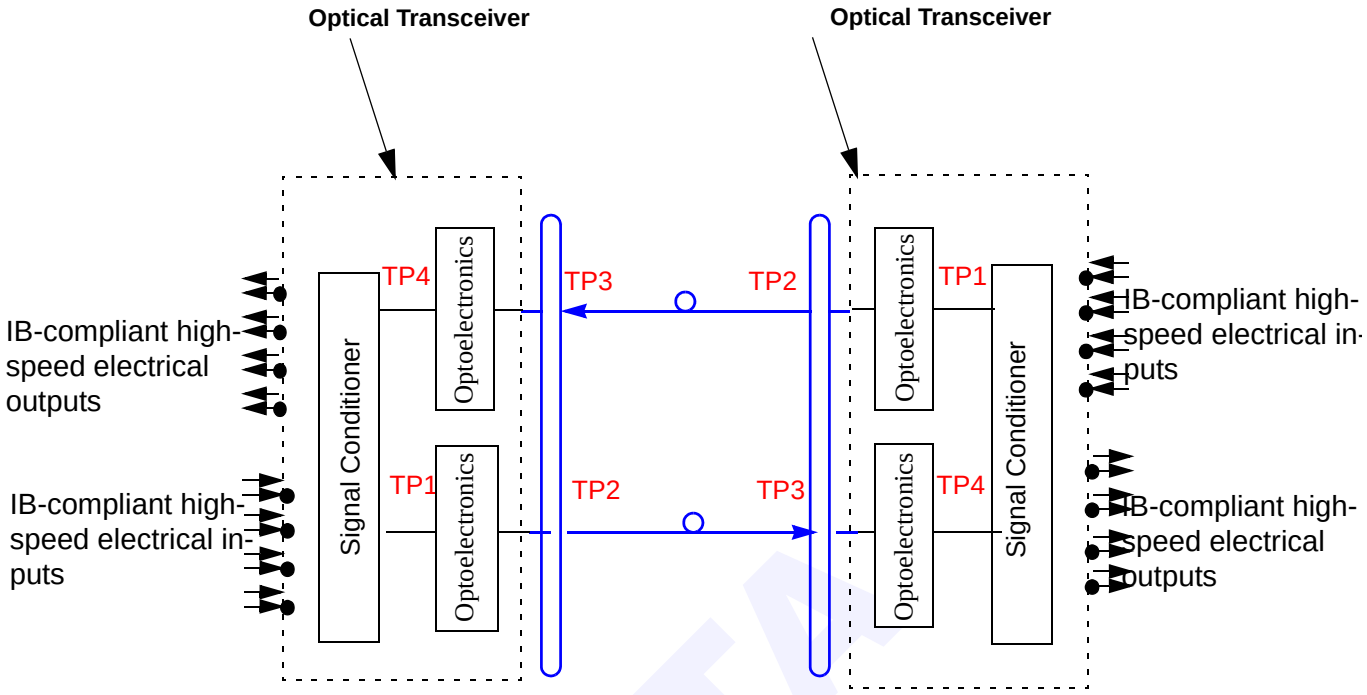


Figure 89 Recommended 4x-LX Optical Link Implementation

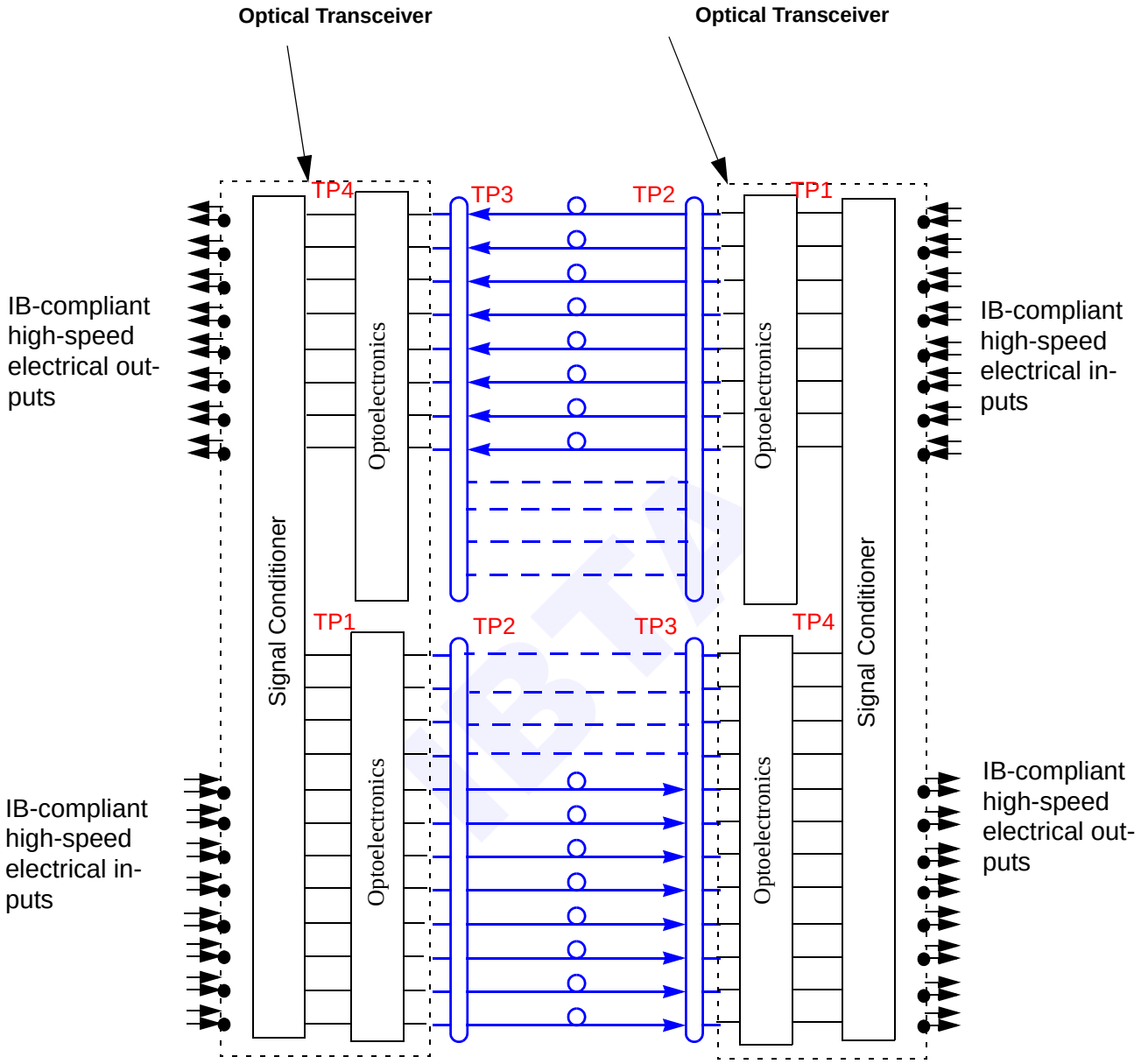


Figure 90 Recommended 8x-SX and 8x-DDR-SX Optical Link Implementation

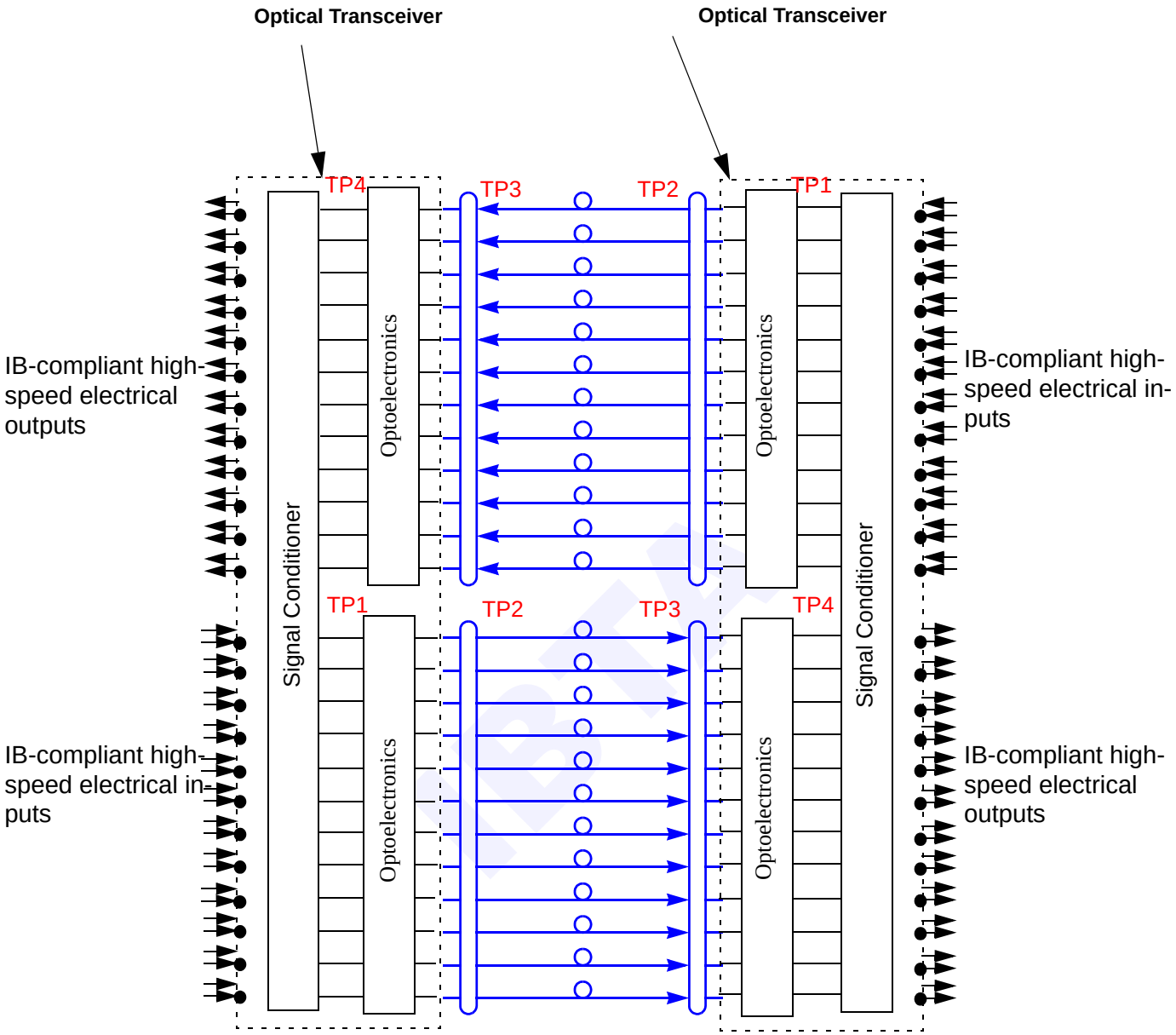


Figure 91 Recommended 12x-SX & 12x-DDR-SX Optical Link Implementation

8.9 AUX POWER

C8-29: All Optical Ports shall comply with Aux power behavior in Section 8.9

C8-29.2.1: All 8x-SX Optical Ports **shall** meet all the Aux power specifications for 12x-SX in Section 8.9 with the exception that 8x SX has only eight receive and transmit lanes.

8.9.1 BEHAVIOR IN AUX POWER MODE

To facilitate compliance with the overall Auxiliary power budget, an Optical Transceiver **should not** draw more than 150mW in Aux power mode. This limit depends on how much other functionality the IB Module is required to provide in Aux power mode. More than 150mW of Aux power may be available to an Optical Transceiver in certain implementations.

The high-speed electrical outputs of an Optical Receiver operating under Aux power **shall** be squelched to less than V_{RSD} (Signal Threshold) as defined in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s.

8.9.2 BEACONING AND WAKE-UP

An Optical Receiver operating only under Aux power **shall** detect the presence of a beaconing sequence (refer to the appropriate receiver specifications in Section 8.5 for the minimum valid optical levels) on the Fiber Optic Cable and trigger the appropriate response. See Chapter 14: OS Power Management and Chapter 5: Link/Phy Interface.

8.10 OPTICAL PLUGGABLE DEVICES

8.10.1 1X OPTICAL PLUGGABLE DEVICES

C8-29.1.1: All 1x SX and 1x LX Optical Pluggable Devices **shall** comply with Section 8.10.1

Any 1x Optical Pluggable device shall comply with electrical and mechanical 1x pluggable requirements in Chapter 6: High Speed Electrical Signaling - 2.5, 5.0, & 10.0 Gb/s and Chapter 7: Copper Cable, and the respective 1x optical distance classifications (SX or LX) in Section 8.4.2, Section 8.5, Section 8.6.1, Section 8.7, and Section 8.9

CHAPTER 9: MECHANICAL SPECIFICATION

9.1 MECHANICAL OVERVIEW

This section describes form factors, chassis slot details, and environmental considerations necessary to implement InfiniBand (IB) systems and board modules.

This specification provides for a 3U chassis implementation with modules oriented vertically as shown in Figure 92, as well as a 6U chassis, and double width modules for both a 3U and 6U chassis. Chassis designs which orient the modules horizontally are also supported.

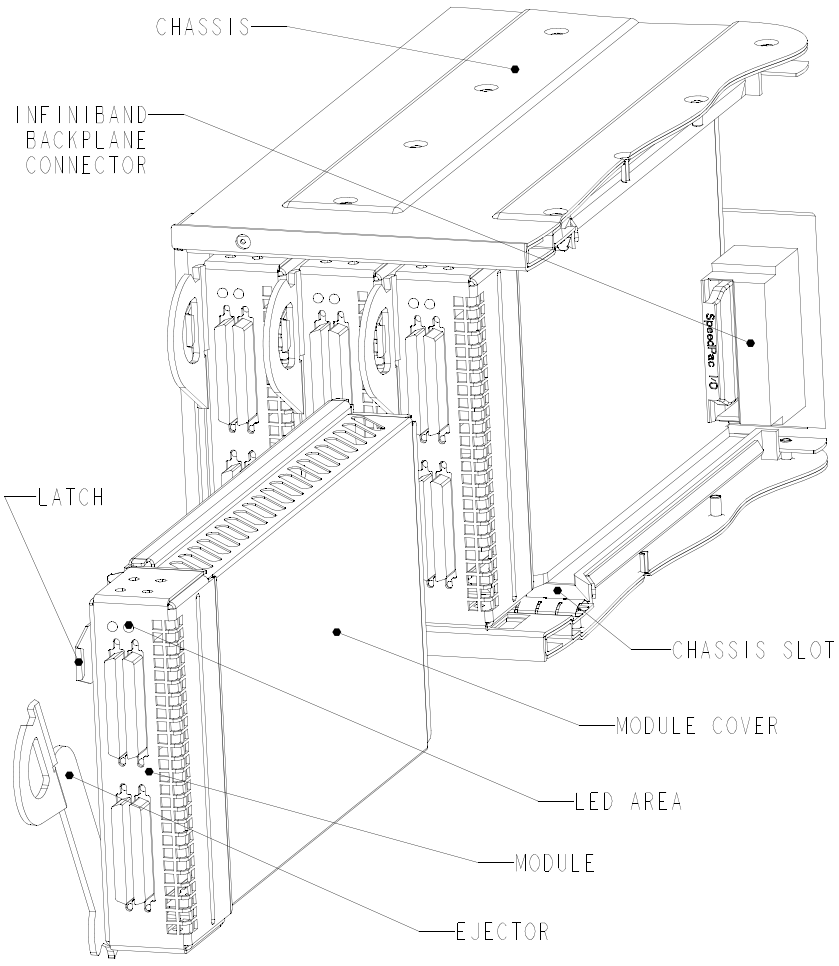


Figure 92 Standard Module, 3U Chassis Slots

An InfiniBand mechanical system consists of InfiniBand Modules which are inserted into InfiniBand Slots, where a slot is defined to be a chassis location capable of accepting a single width module. To be explicit, there is no definition for a "double width slot". Wide modules, by definition, occupy two adjacent regular width slots. This specification does not define a particular chassis or backplane design, but it does define all relevant dimensions and interfaces for a single slot, as well as the spacing between adjacent slots. The designer can step-and-repeat this slot definition as required to build any size system.

9.1.1 DIMENSIONS AND TOLERANCES

C9-1: All dimensions included in this chapter are in millimeters and **shall** have a tolerance of +/- 0.25 mm unless otherwise specified.

> **Note to Designers**
>
> Some assembly tolerances in Figure 97 on page 366, Figure 101 on page 370, Figure 105 on page 374, or Figure 109 on page 378 may require assembly fixturing. Specific examples include dimensions between paddle guard and ejector surfaces.

9.1.2 MODULE DESCRIPTION

Most of the mechanical definition for InfiniBand is centered around the module. The module, shown in Figure 93 on page 358, consists of the following parts:

a) The module carrier - a basic metal structure which provides:
   - A metal plate which aligns and supports the circuit board(s) and provides EMI suppression between adjacent boards
   - Metal guides which mate with the surfaces of the slot for alignment, guidance, thermal management, and additional EMI suppression
   - A connector housing which provides for external connector mounting, four sided EMI sealing, ESD protection, and thermal management

b) The module ejector(s) & latch(es) - a handle and latch which provide:
   - Controlled insertion and ejection of the module for hot add/remove
   - Retention of the module during operation or shipping

c) The module cover - easily removable cover to protect the board
   - Provides ESD protection for the board and components
   - Prevents damage during module handling

- Protects the user from high temperature components
- May optionally provide air dams / deflectors to improve the thermal characteristics of the module
- Provides area for additional transverse airflow

d) The board(s) which implement the module's functionality

IBTA

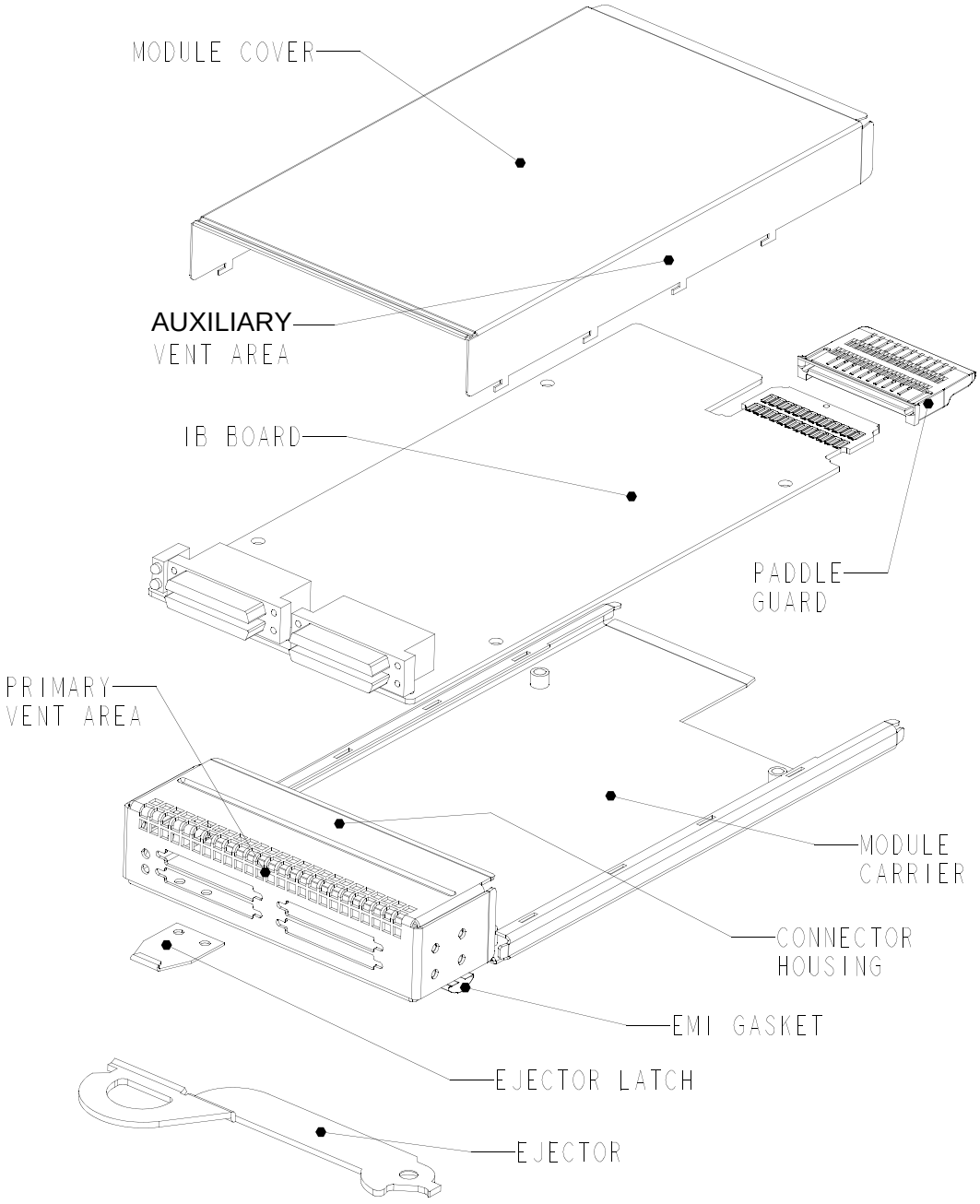


Figure 93 Exploded View of InfiniBand Standard Module

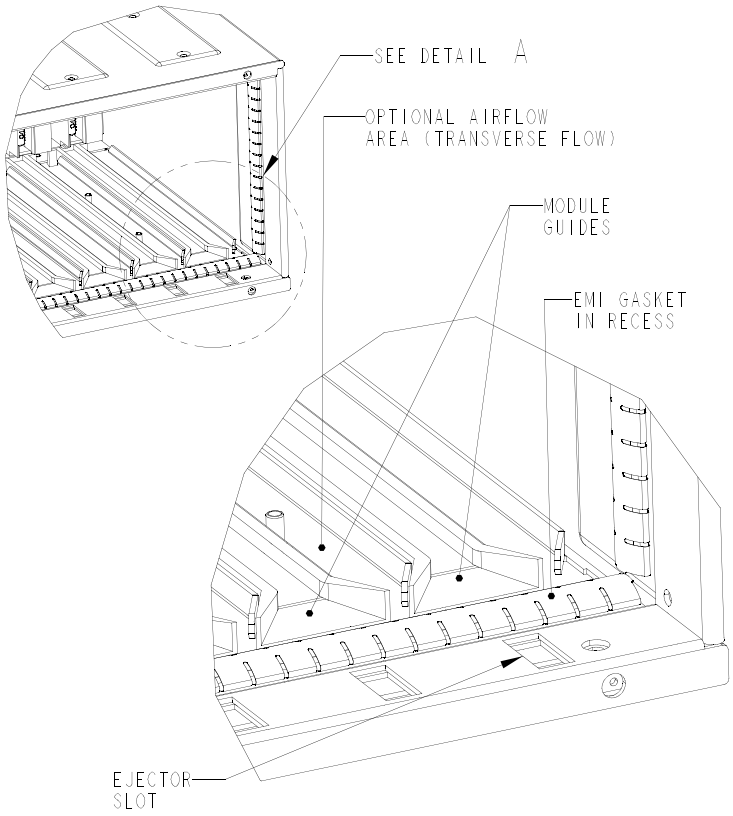


Figure 94 Chassis Slot Detail

9.1.3 SLOT DESCRIPTION

The remaining mechanical definition for an InfiniBand system is centered around the slot which receives the module. Figures shown are of a design example, not of specified geometry except where explicitly dimensioned. Minimal dimensions are specified to allow product differentiation. The slot, shown in Figure 92 on page 355 with additional detail in Figure 94, consists of the following key features:

a) EMI/ESD contacts at the top and bottom of the chassis slot to complete the four sided EMI seal

b) Mating module guides which may be isolated, directly connected, or resistively connected to the chassis slot as required by the system designer to complete the EMI management solution

c) Features which mate with the module's ejector(s) outside of the module/slot EMI enclosure

d) The backplane and backplane connector

e) An optional area in the module guide area to provide additional airflow in a predominantly transverse direction.

Note to Chassis Designers

The right most slot in a chassis must provide the EMI gasket for the right side of the last module. Thereafter, the modules seal against one another or against the left wall of the chassis.

9.1.3.1 SLOT DESIGNATIONS

Viewed from the external surface of the carrier connector housing, slot numbering convention is from left to right or bottom to top depending upon the orientation of InfiniBand Modules. System designers should label chassis slots in a manner identical to the slot designation reported through the system management structure to facilitate module hot add/remove.

This release of the specification supports more than one port per slot connector. This highlights the need to provide the user a consistent means to identify chassis and modules that support greater that one port per slot connector. The following table, Table 88, outlines this naming convention for chassis slot labeling.

Table 88 Chassis Slot Configurations and Labeling

Label[a]	Descriptions
CxP0	Primary slot connector with no ports (Power and or Management only)
CxP1	Primary slot connector with port 1 populated
CxP2	Primary slot connector with ports 1 and 2 populated
CxP3	Primary slot connector with ports 1, 2, and 3
OxP0	Optional slot connector with no ports (Power and or Management only)
OxP1	Optional slot connector with port 1 populated
OxP2	Optional slot connector with ports 1 and 2 populated
OxP3	Optional slot connector with ports 1, 2 and 3 populated

a. C = Primary Slot Connector, O = Optional Slot Connector, x = Slot Number. Port position information can be found in Table 100 Backplane connector board and backplane contact assignments for three 4x ports - primary side on page 421 and Table 101 Backplane connector board and backplane contact assignments for three 1x ports - primary side on page 422.

Example slot labeling:

- Chassis Slot 4 with 2 ports on the primary connector would be labeled as "C4P2".
- Tall Slot 6 with optional connector with two ports on the primary connector and one port on the optional connector would be "C6P2-O6P1".

9.1.3.2 MODULE PORT DESIGNATIONS

The addition of support for more than one port per slot / module connector in this version of the specification creates the need to provide a consistent means for user to match module and slot that support these additional ports. The following labeling convention is defined for modules and should be placed clearly on the module cover.

Table 89 Module Port Configurations and Labels

Label[a]	Description
C1P1	Primary backplane connector with port 1 populated
C1P2	Primary backplane connector with ports 1 and 2 populated
C1P3	Primary backplane connector with ports 1, 2 and 3 populated
C2P0	Second primary backplane connector with no port (This is for wide modules that use the second primary port for power or management or both)
C2P1	Second primary backplane connector with port 1 populated
C2P2	Second primary backplane connector with ports 1 and 2 populated
C2P3	Second primary backplane connector with ports 1, 2 and 3 populated
O1P0	Optional backplane connector with no port (This is for Tall modules that use the optional port for power or management or both)
O1P1	Optional backplane connector with port 1populated (Tall Modules)
O1P2	Optional backplane connector with ports 1 and 2 populated
O1P3	Optional backplane connector with ports 1, 2 and 3 populated
O2P0	Second optional backplane connector with no ports (Tall Wide Modules)
O2P1	Second optional backplane connector with port 1 populated
O2P2	Second optional backplane connector with ports 1 and 2 populated
O2P3	Second optional backplane connector with ports 1, 2 and 3 populated

a. C = Primary Slot Connector, O = Optional Slot Connector in Tall Modules
Ports location information can be found in Table 100 Backplane connector board and backplane contact assignments for three 4x ports - primary side on page 421 and Table 101 Backplane connector board and backplane contact assignments for three 1x ports - primary side on page 422.

Example module labeling:

- Standard module with 2 ports labeled as "C1P2"
- Standard Wide Module with 2 port one on each connector = "C1P1-C2P1".
- Standard Wide module with 4 port two on each connector = "C1P2-C2P2".
- Tall module with 2 ports on primary connector = "C1P2"
- Tall Wide module with 12 ports 3 on each connector = "C1P3-C2P3-O1P3-O2P3"

o9-1.1.1: Chassis that support greater than one InfiniBand port per backplane connector **shall** populate the ports in the connector in the defined configurations in Table 88 Chassis Slot Configurations and Labeling on page 360 using the connector pin assignments defined for 4x ports in Table 101 Backplane connector board and backplane contact assignments for three 1x ports - primary side on page 422 or for 1x ports in Table 102 Backplane connector board and backplane contact assignments - secondary side on page 423. Each slot **shall** be clearly labeled with the appropriate label from Table 88 Chassis Slot Configurations and Labeling on page 360.

o9-1.1.2: Modules that support greater than one InfiniBand port on each backplane connector **shall** populate the ports in the backplane connectors in the defined configurations in Table 89 Module Port Configurations and Labels on page 361 using the connector pin assignments defined for 4x ports in Table 101 Backplane connector board and backplane contact assignments for three 1x ports - primary side on page 422 or for 1x ports in Table 102 Backplane connector board and backplane contact assignments - secondary side on page 423. Each module **shall** be clearly labeled with the appropriate labels from Table 89 Module Port Configurations and Labels on page 361.

9.1.4 DATUM PLANE DEFINITION

The remaining sections of the mechanical specification, as well as the backplane connector specification, are based on detailed dimensioned drawings which all use a common set of reference Datum planes. These Datum planes are defined in Table 90 which is referenced to a vertical module orientation.

Table 90 Datum Planes

Datum Name	Definition
A	Horizontal plane, perpendicular to both the board and backplane, which passes through the center of the InfiniBand backplane connector

Table 90 Datum Planes

Datum Name	Definition
B	Vertical plane, coincident with the primary (connector) surface of the InfiniBand backplane
C	Vertical plane, coincident with the primary component surface of the InfiniBand board
D	Vertical plane, parallel to the backplane and coincident with the inside surface of the carrier's connector housing
E	Vertical plane, parallel to Datum C and offset from Datum C by 7mm; coincident with the left side of the carrier plate.

9.2 MODULE DESCRIPTION

The InfiniBand Module completely encloses the board. It minimally consists of a board, carrier, cover, standard ejector, and latch.

IBTA

C9-2: Dimensions and tolerances for all modules **shall** conform to Figure 95.

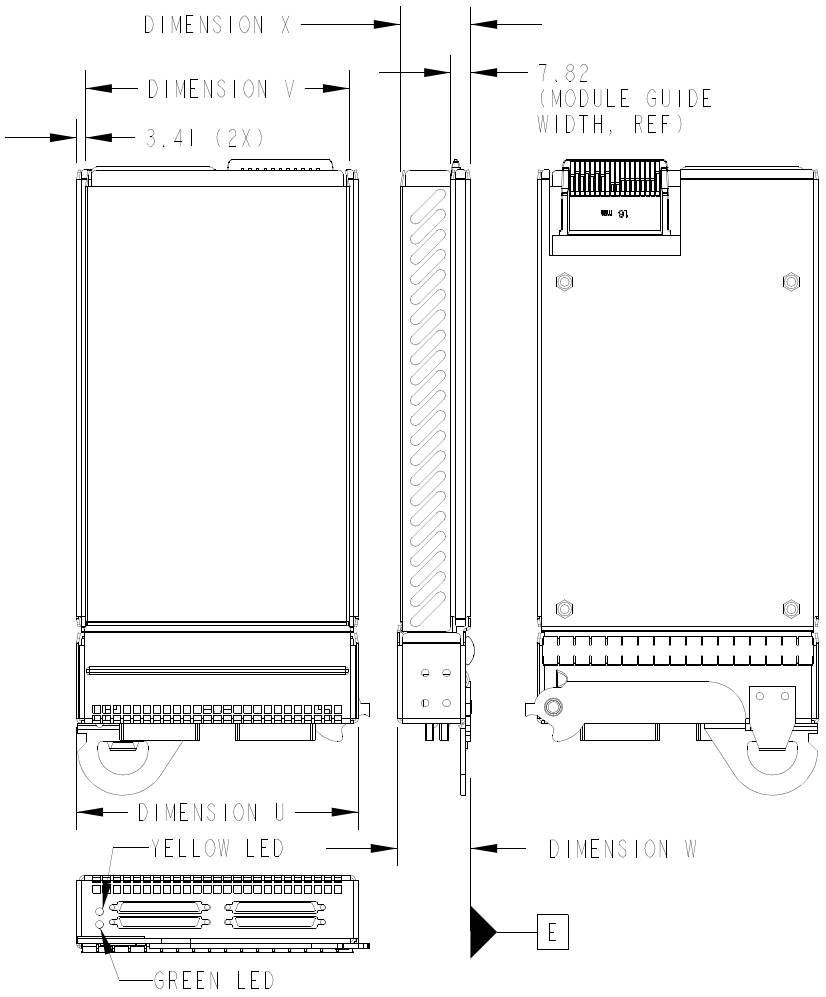


Figure 95 General Module Assembly Dimensions (Parametric)

Table 91 Parametric Dimensions for Figure 95

Module Size	Dimension U	Dimension V	Dimension W	Dimension X
Standard	112.00	"U" - 7mm	29.0	"W" - 1mm
Standard Wide	112.00	"U" - 7mm	59.0	"W" - 1mm
Tall	245.35	"U" - 7mm	29.0	"W" - 1mm
Tall Wide	245.35	"U" - 7mm	59.0	"W" - 1mm

9.2.1 STANDARD SIZE MODULE

C9-3: Standard modules **shall** conform to the specific dimensions shown in Figure 97 on page 366 through Figure 99 on page 368

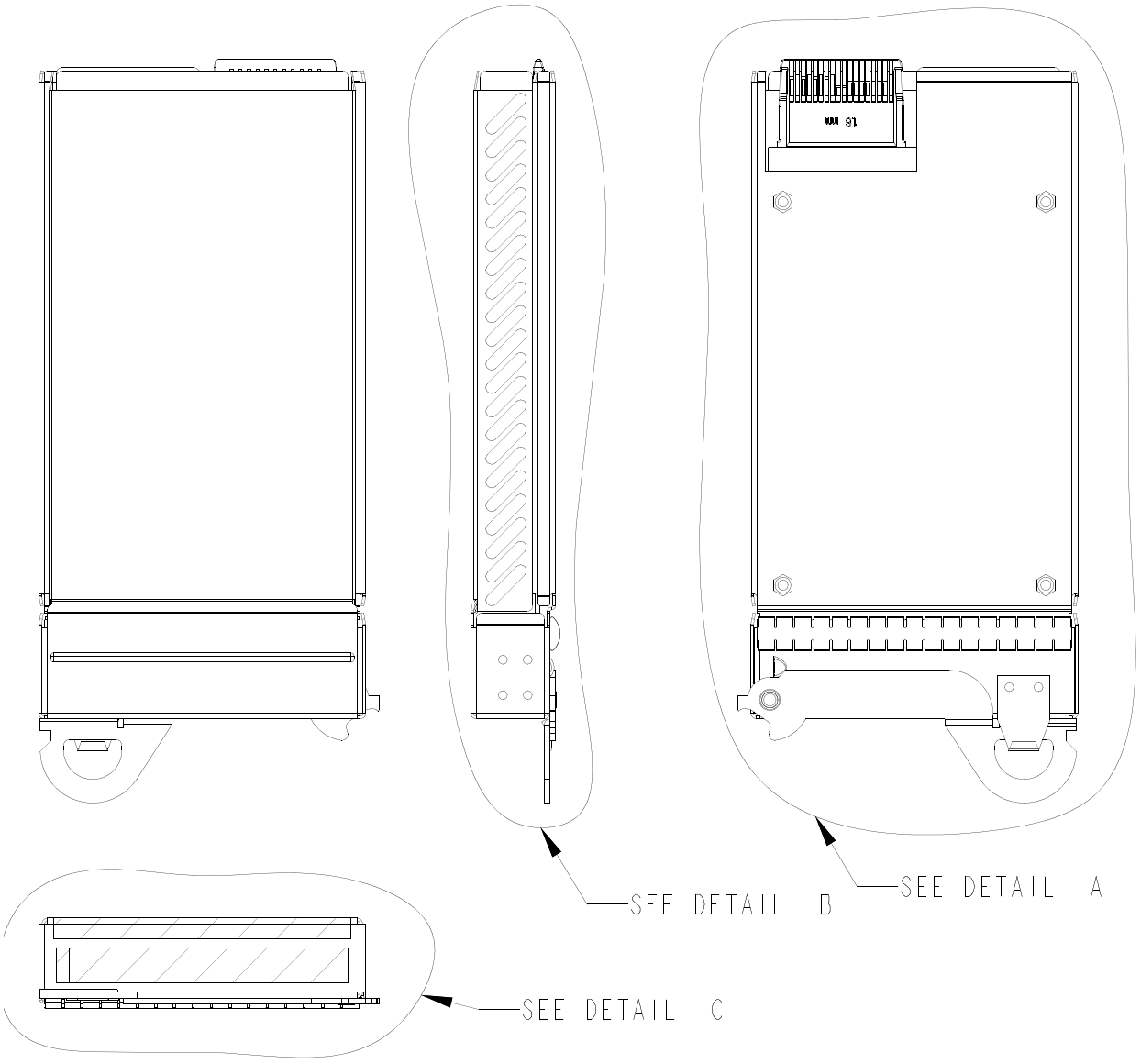


Figure 96 Standard Module Critical Dimensions

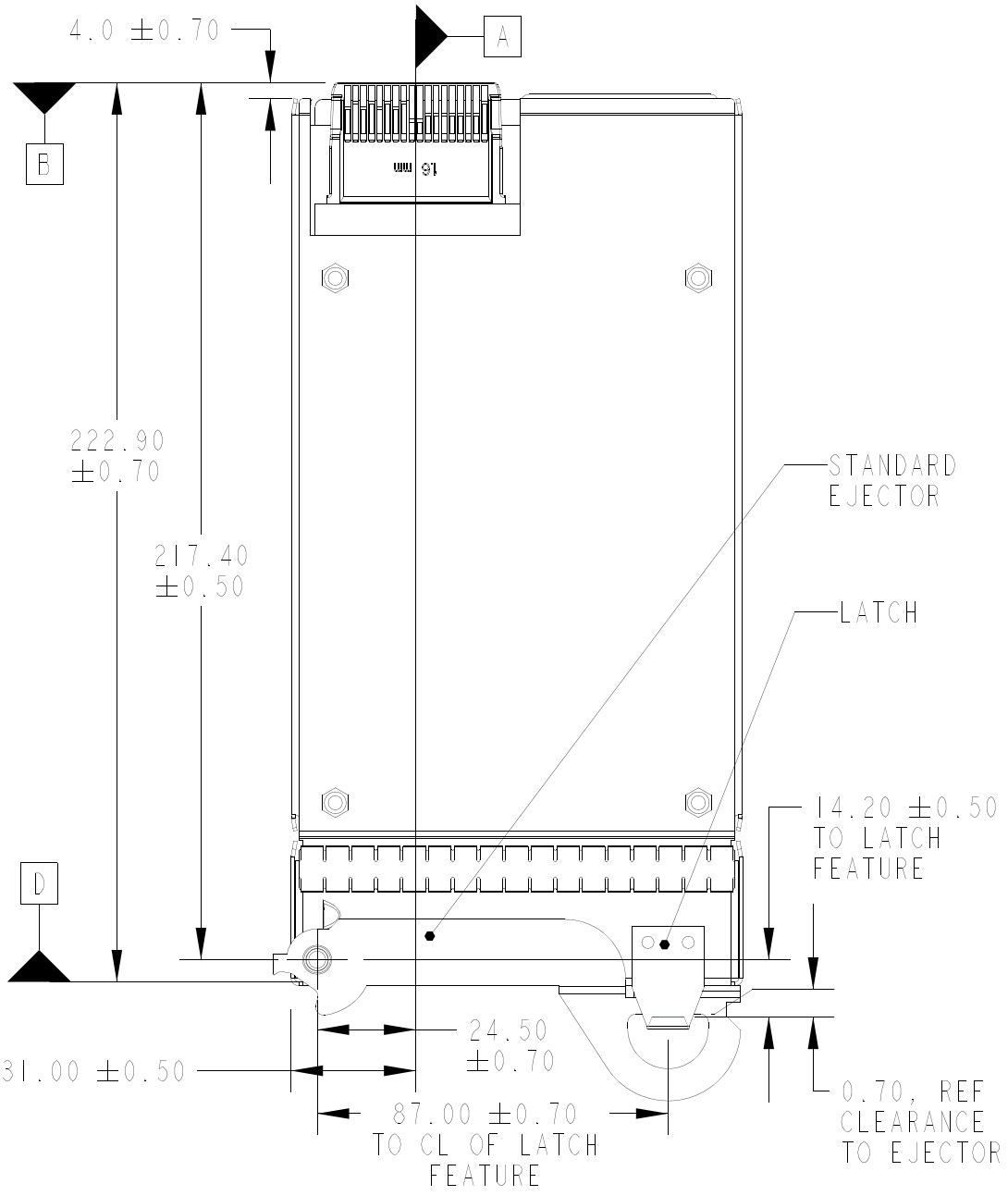


Figure 97 Standard Critical Dimensions - Detail A

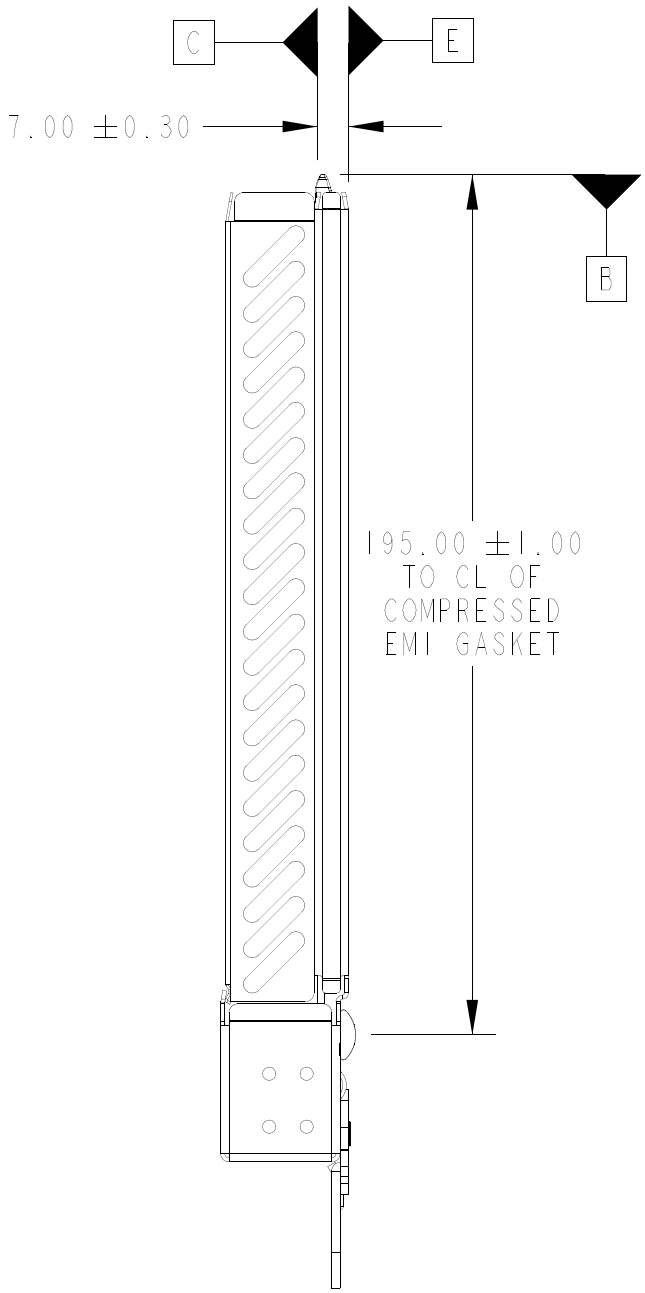


Figure 98 Standard Critical Dimensions - Detail B

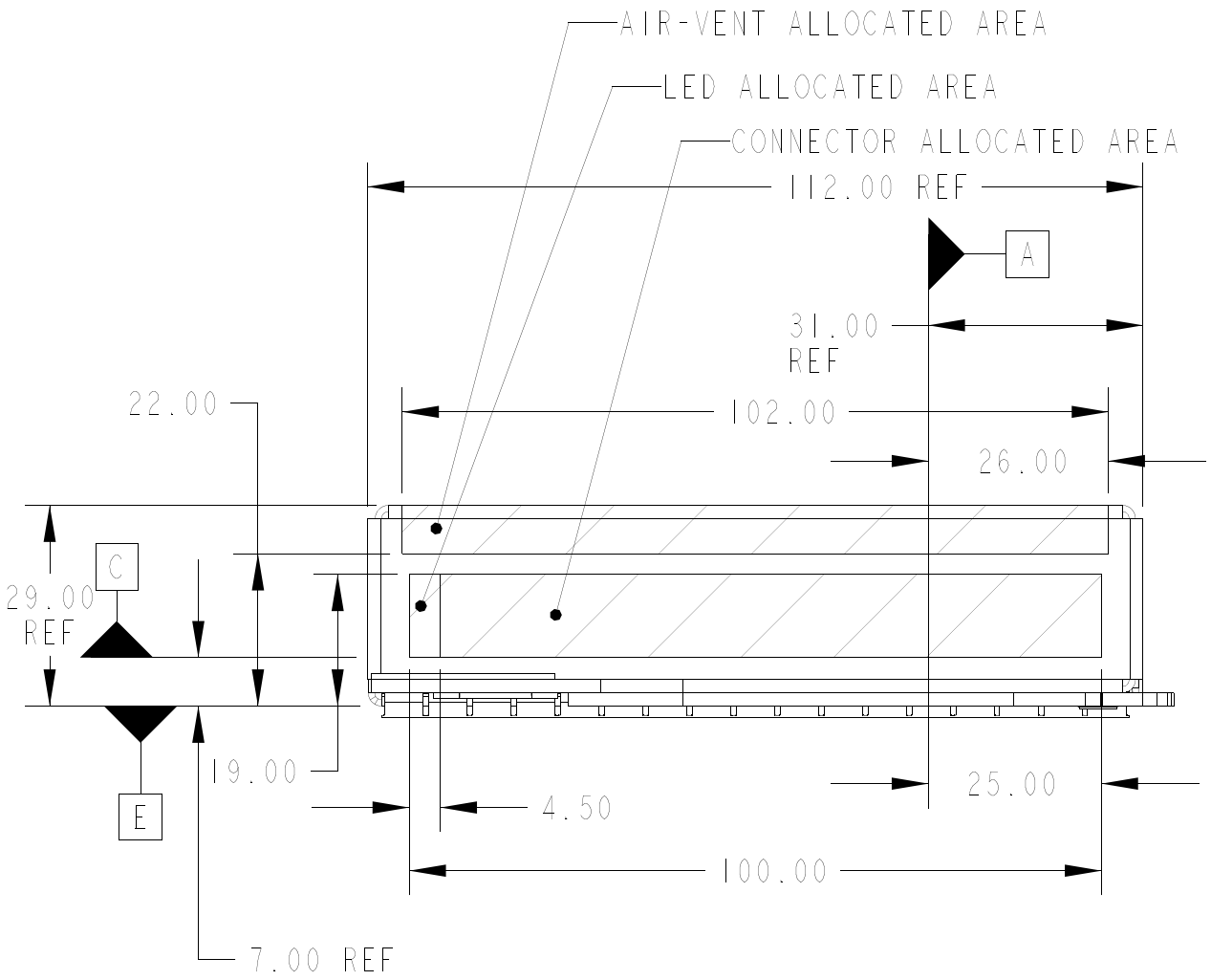


Figure 99 Standard Critical Dimensions - Detail C

9.2.2 TALL SIZE MODULE

C9-4: Tall modules **shall** conform to the specific dimensions shown in Figure 101 on page 370 through Figure 103 on page 372.

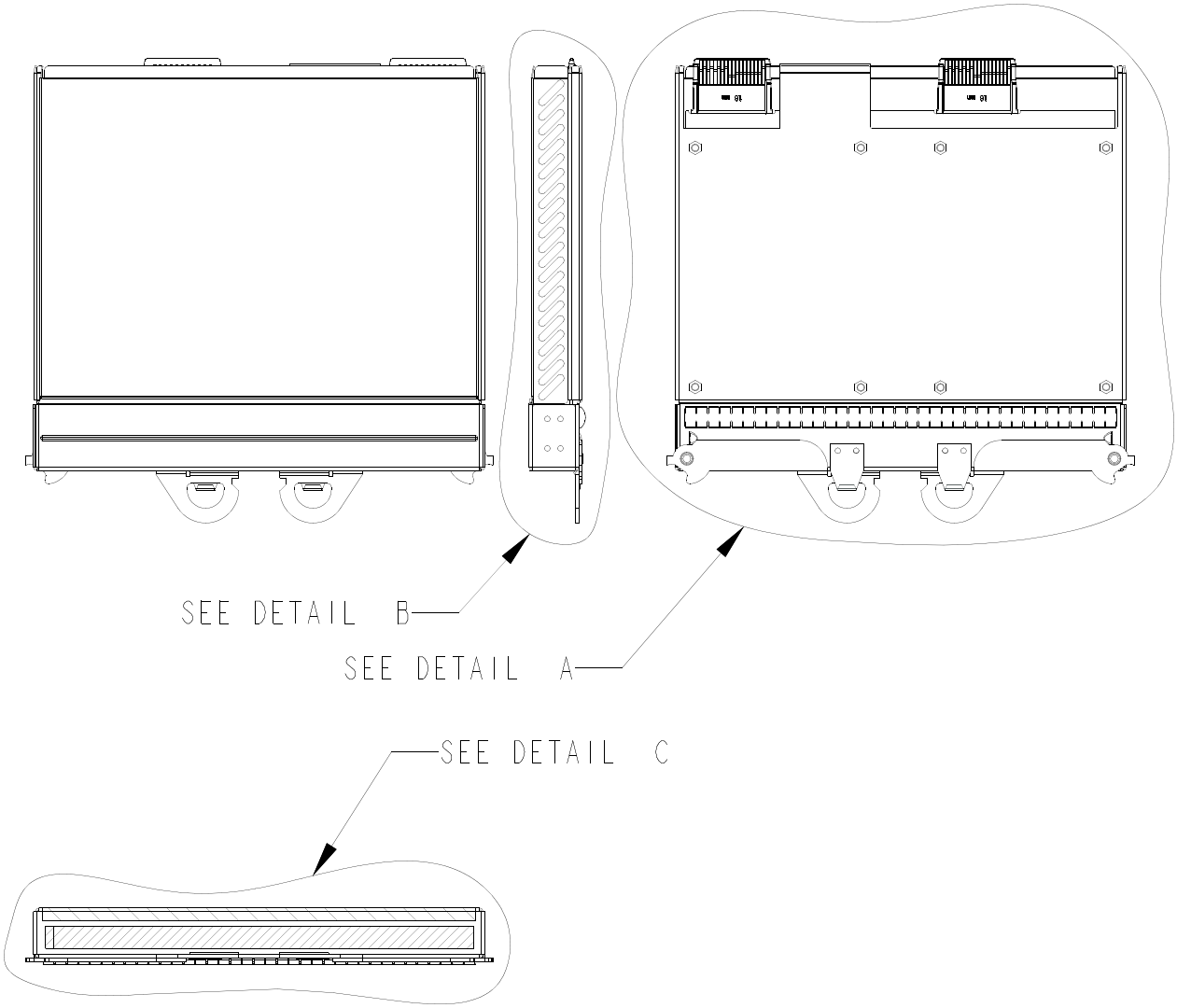


Figure 100 Tall Module Critical Dimensions

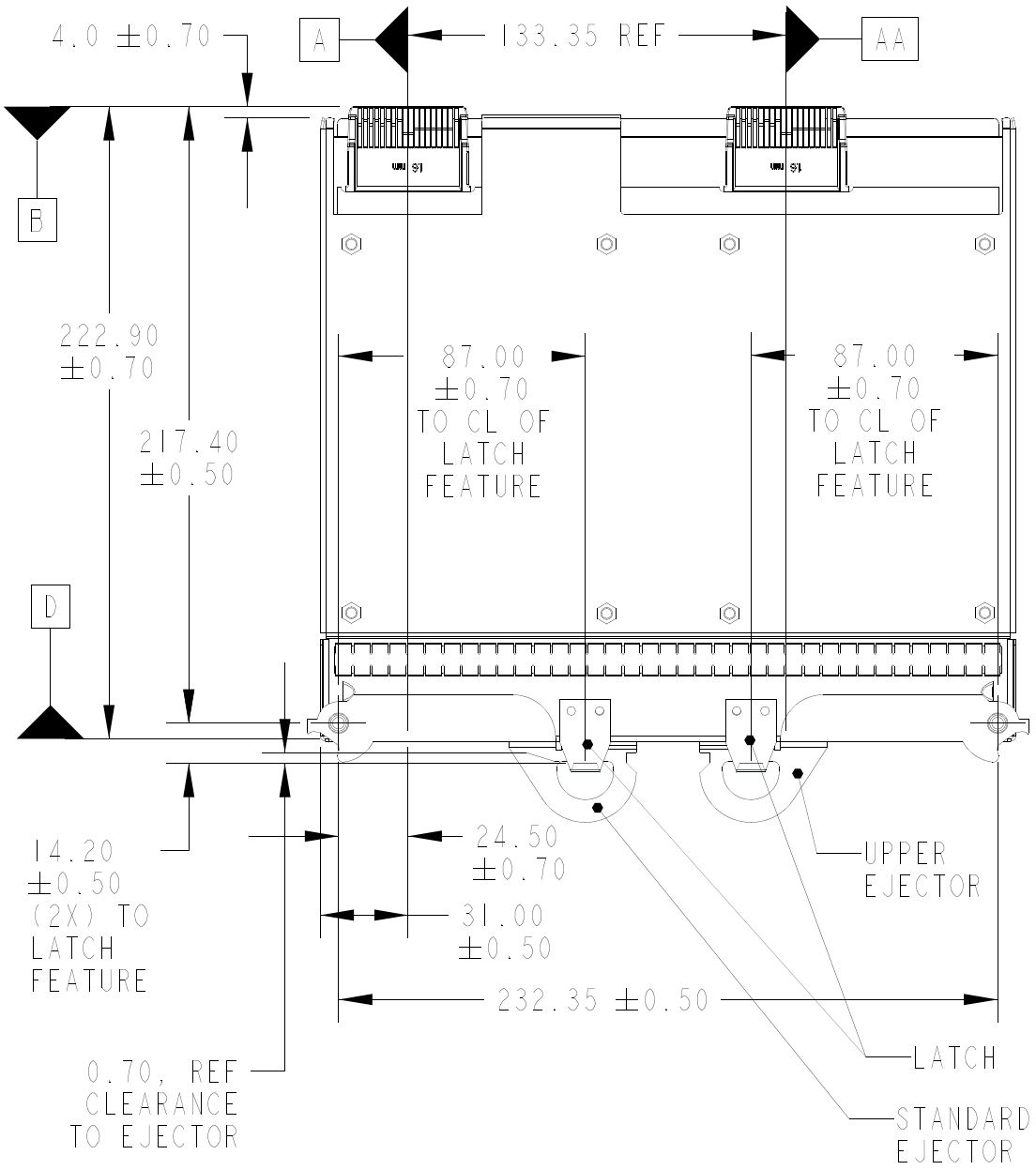


Figure 101 Tall Critical Dimensions - Detail A

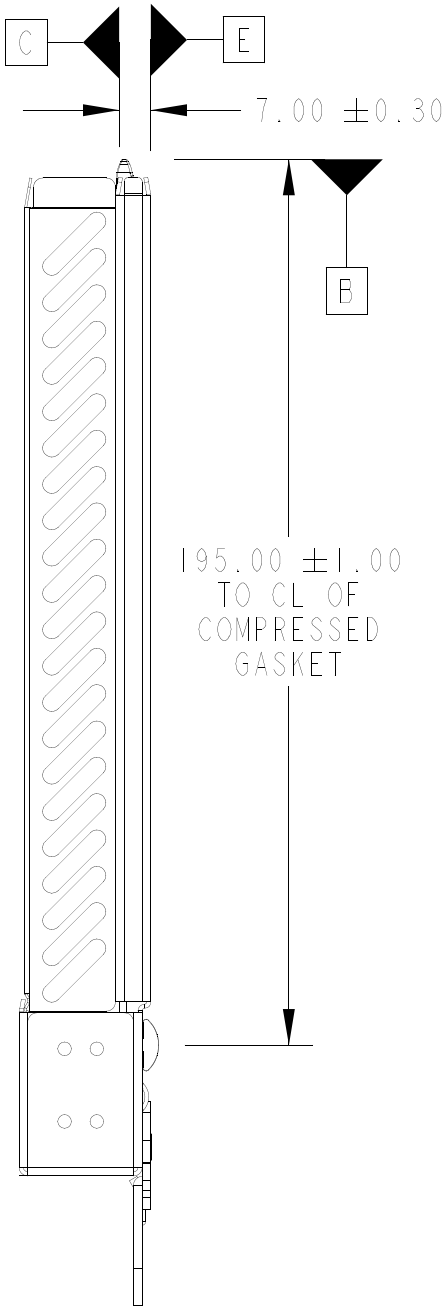


Figure 102 Tall Critical Dimensions - Detail B

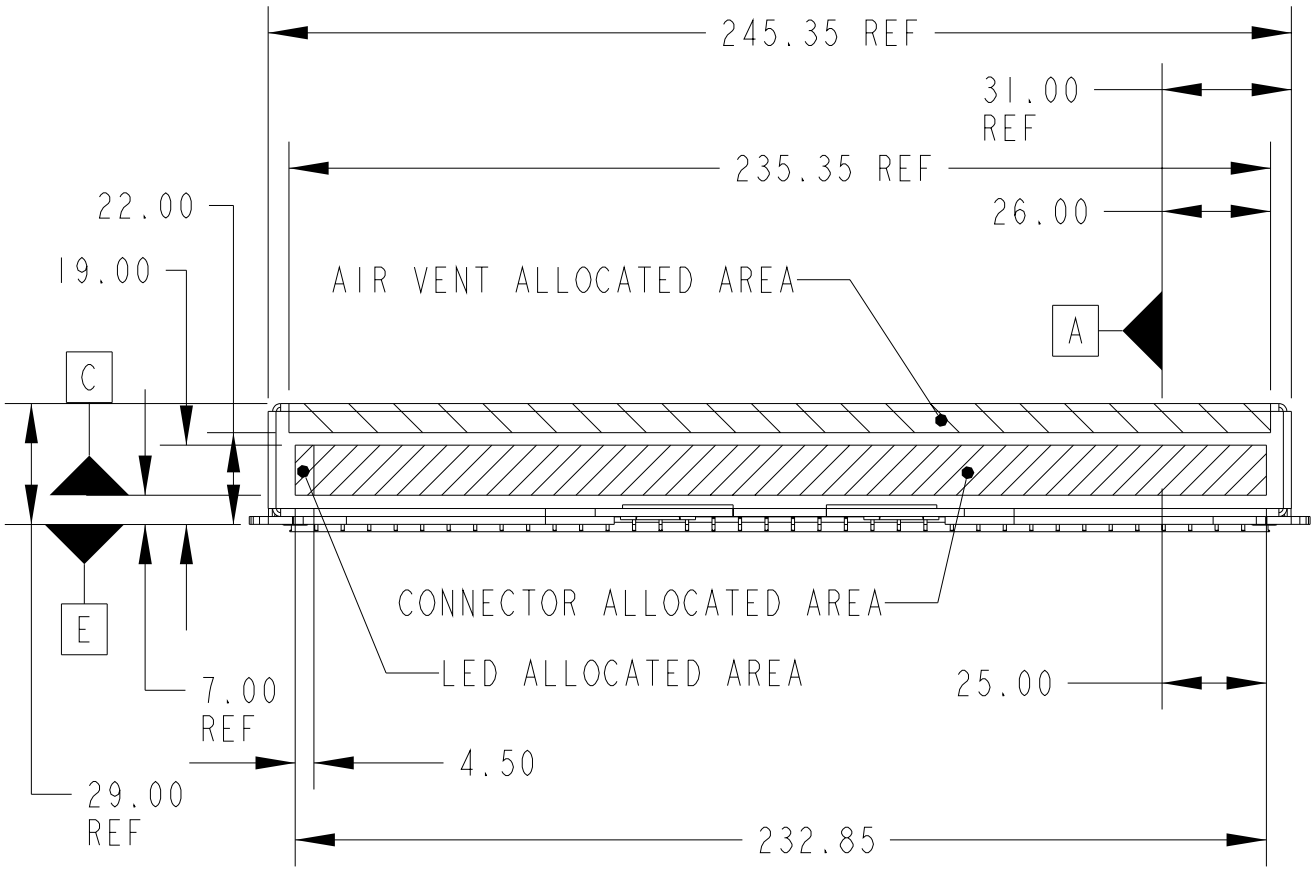


Figure 103 Tall Critical Dimensions - Detail C

9.2.3 STANDARD WIDE MODULE

C9-5: Standard Wide Modules **shall** conform to the specific dimensions shown in Figure 105 on page 374 through Figure 107 on page 376.

The upper two hatched areas in Figure 107 on page 376 correspond to the connector and venting area for the second optional board. Absent of

the second board, this entire space is free for venting provided the module still complies with Figure 123 on page 401.

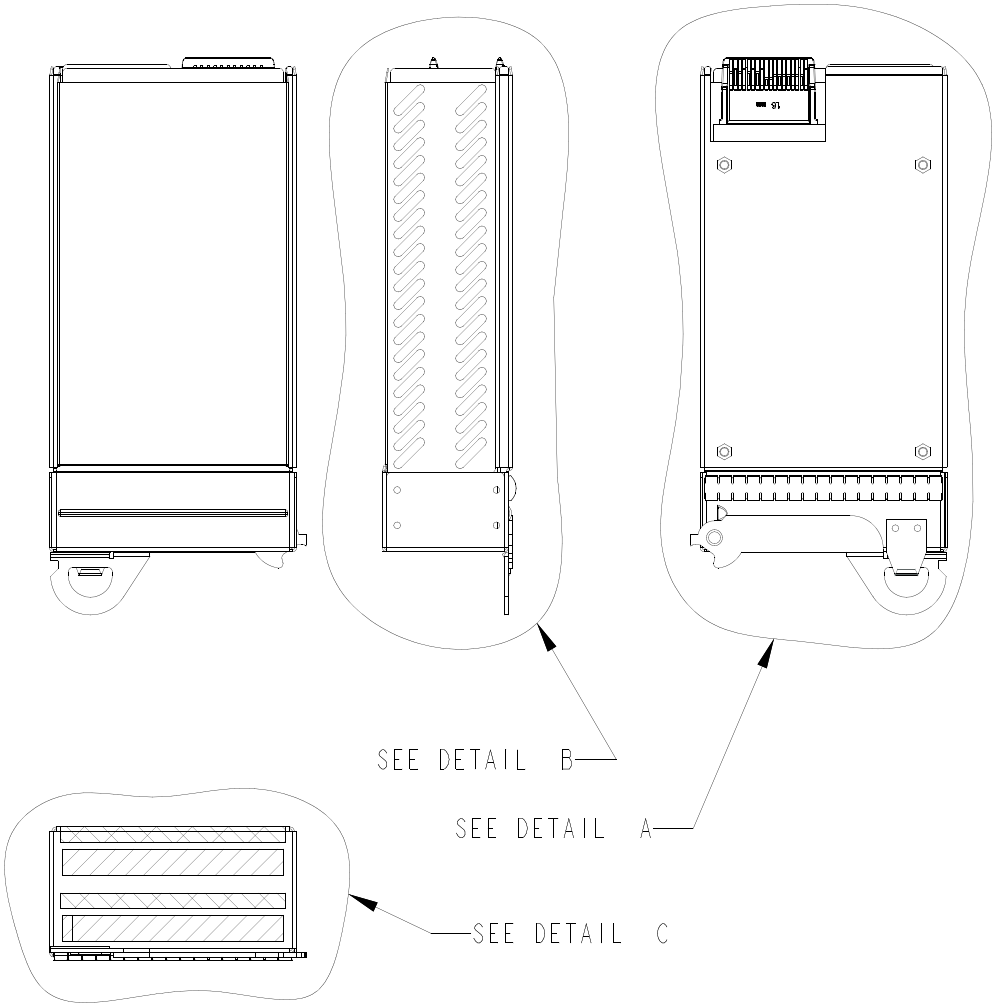


Figure 104 Wide Module Critical Dimensions

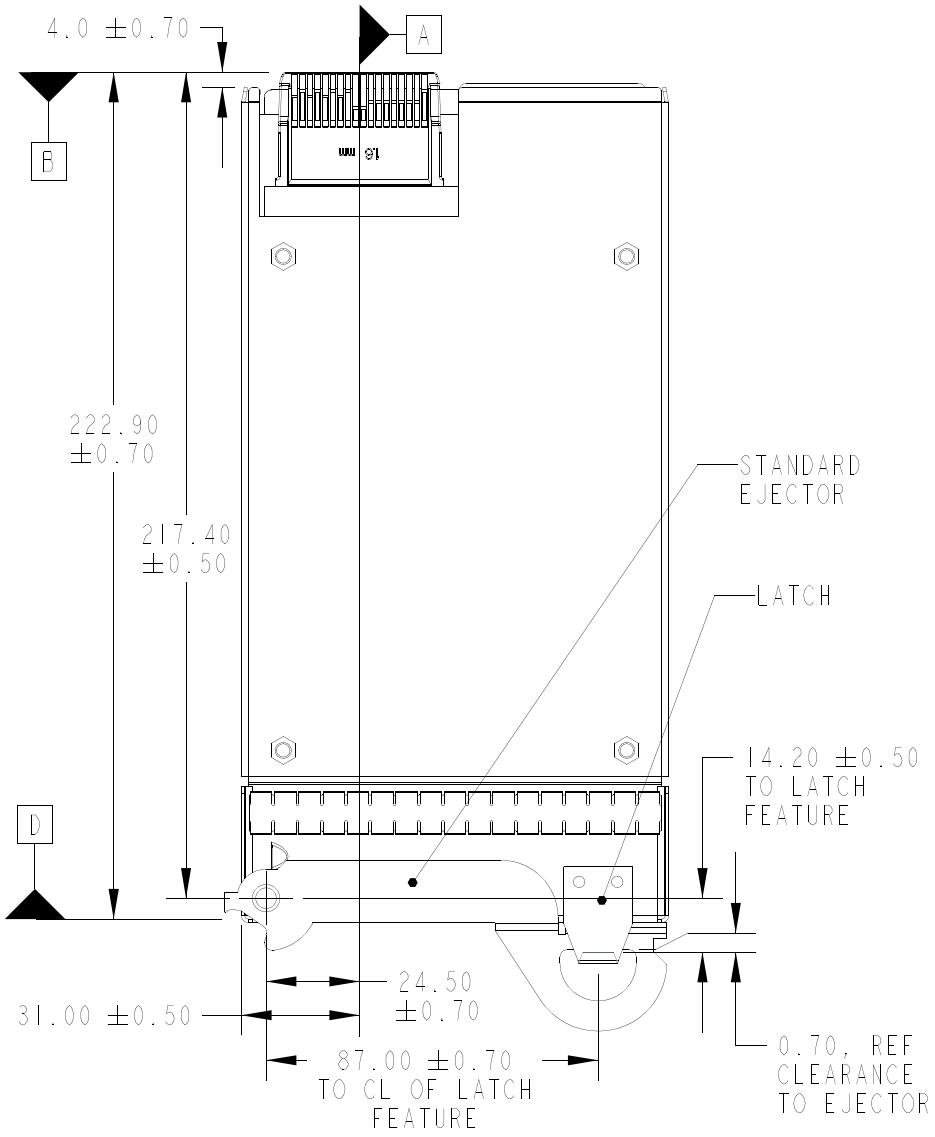


Figure 105 Wide Critical Dimensions - Detail A

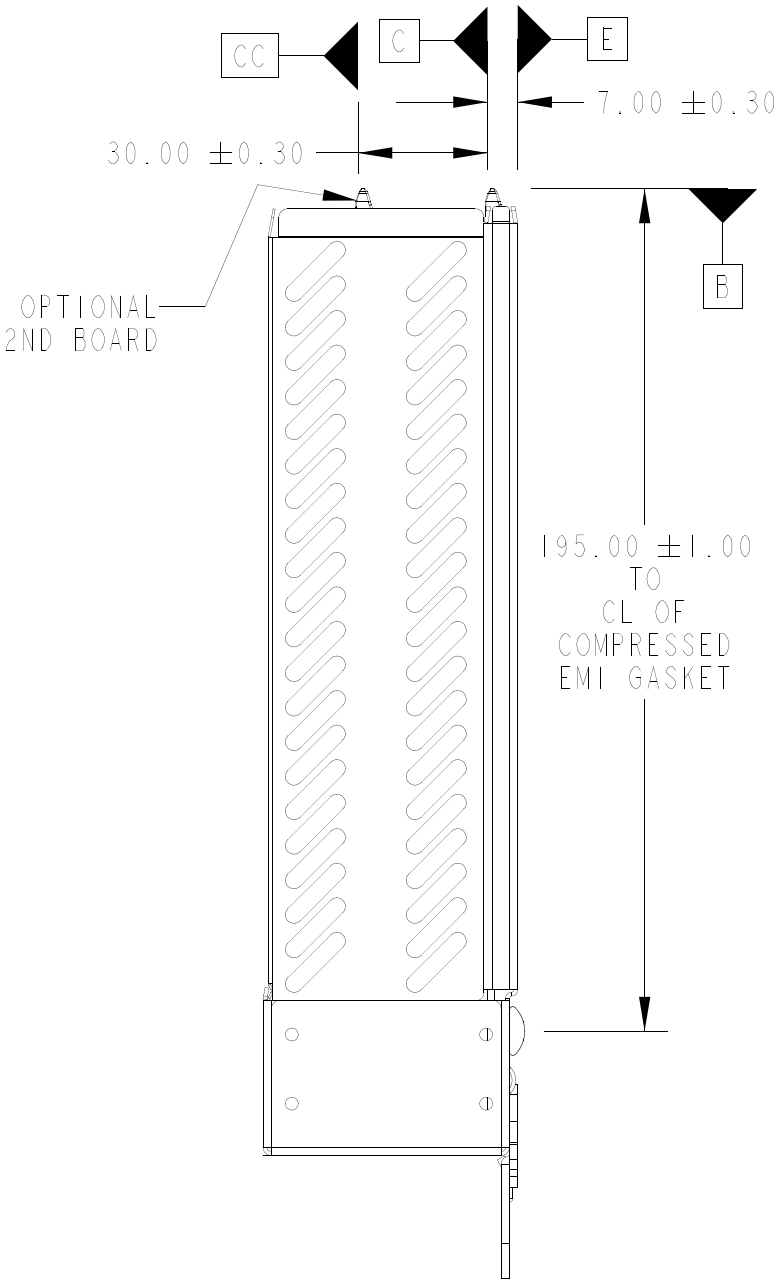


Figure 106 Wide Critical Dimensions - Detail B

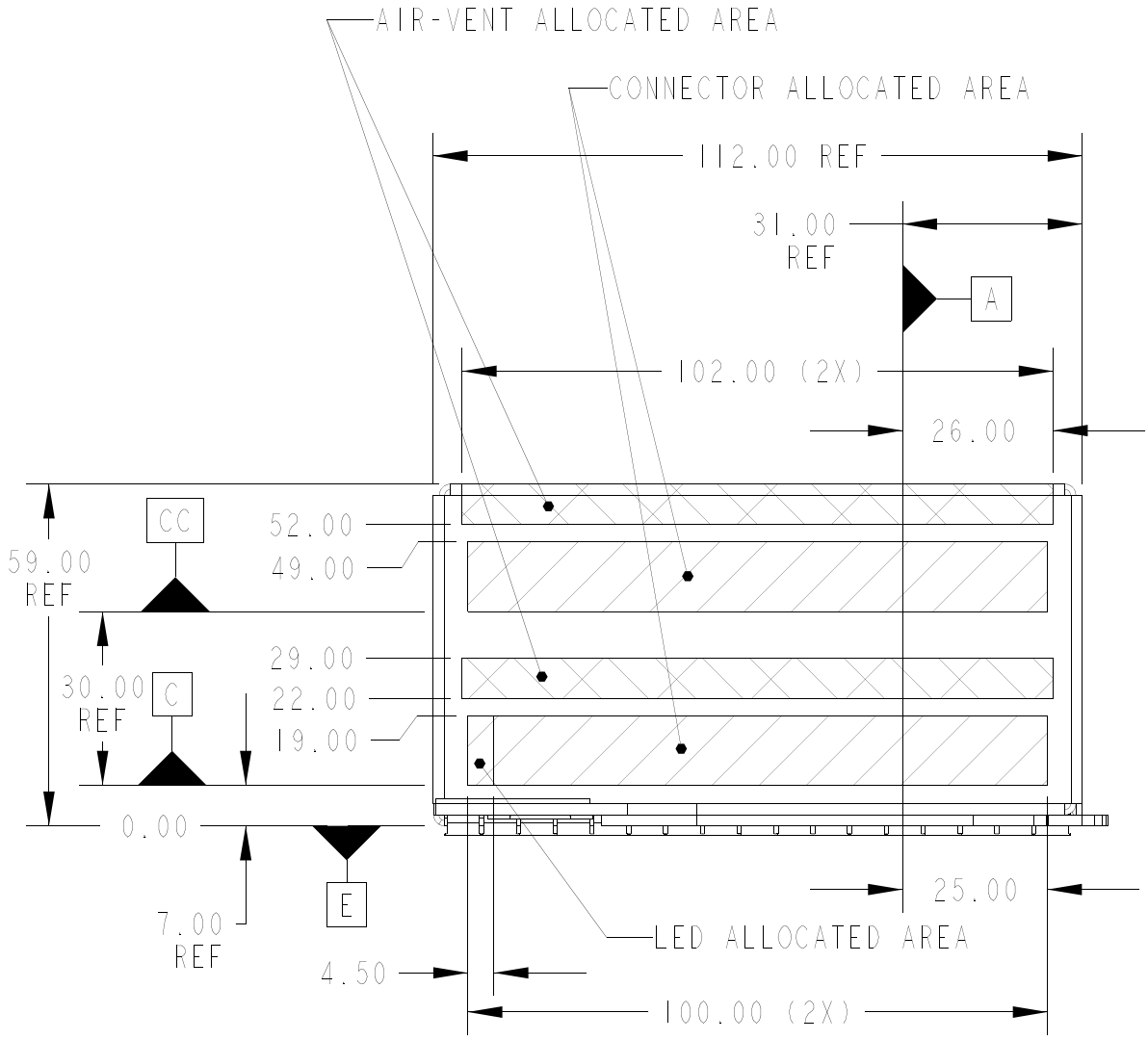


Figure 107 Wide Critical Dimensions - Detail C

9.2.4 TALL WIDE MODULE

C9-6: Tall Wide Modules **shall** conform to the specific dimensions shown in Figure 109 on page 378 through Figure 111 on page 380.

The upper two hatched areas in Figure 111 on page 380 correspond to the connector and venting area for the second optional board. Absent of the

second board, this entire space is free for venting provided the module still complies with the pressure/flow characteristics specified in Figure 123 on page 401.

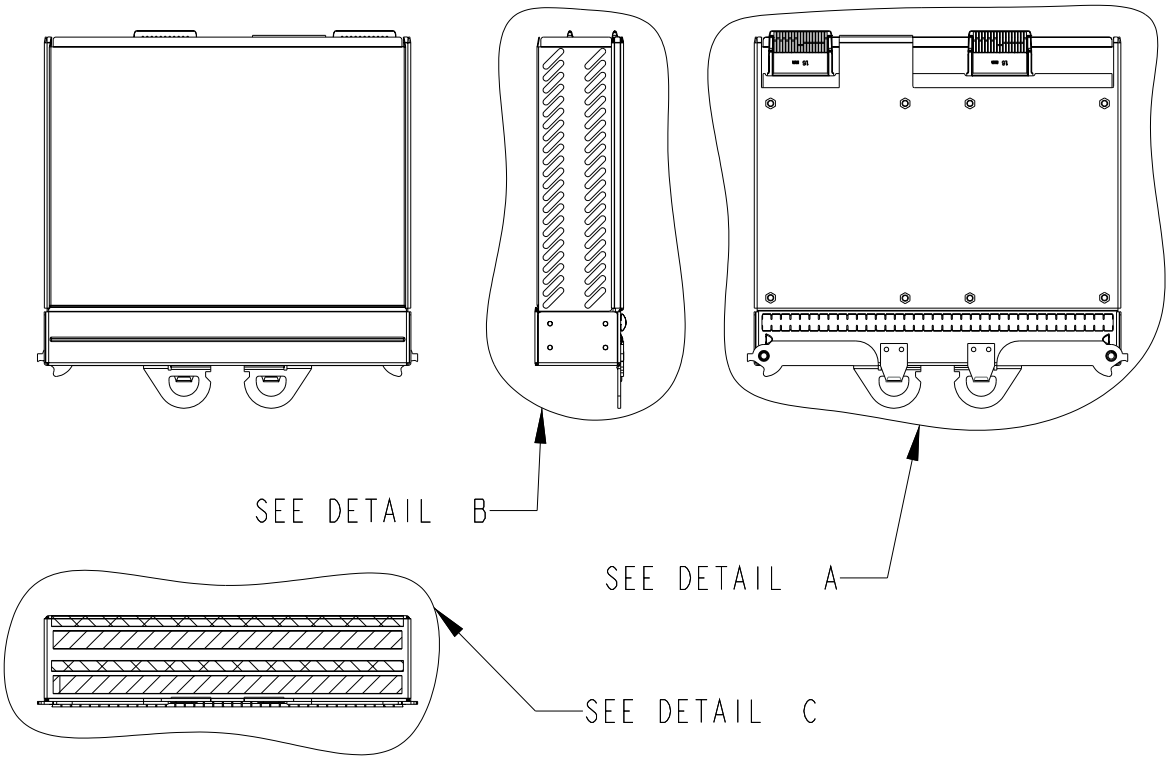


Figure 108 Tall Wide Module Critical Dimensions

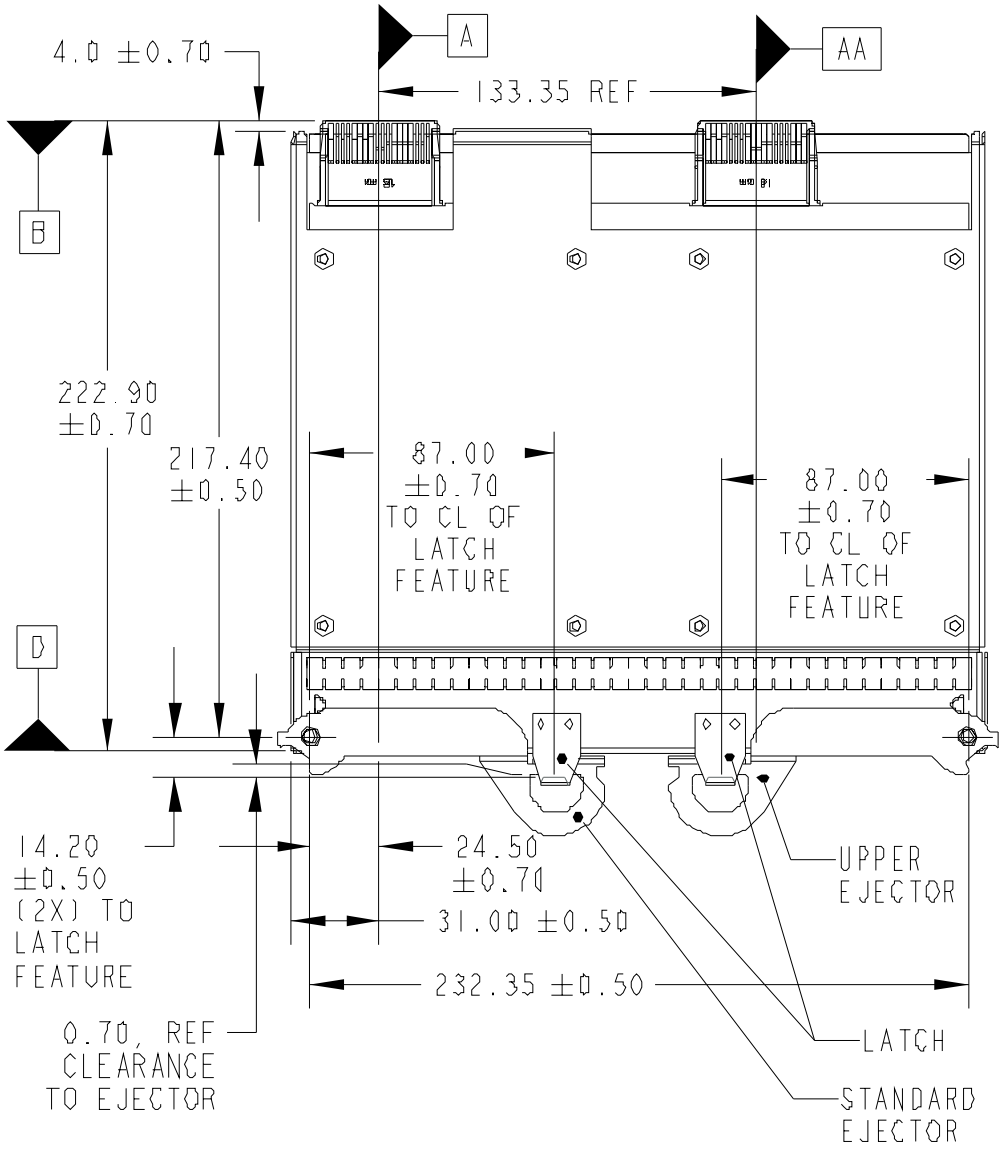


Figure 109 Tall Wide Critical Dimensions - Detail A

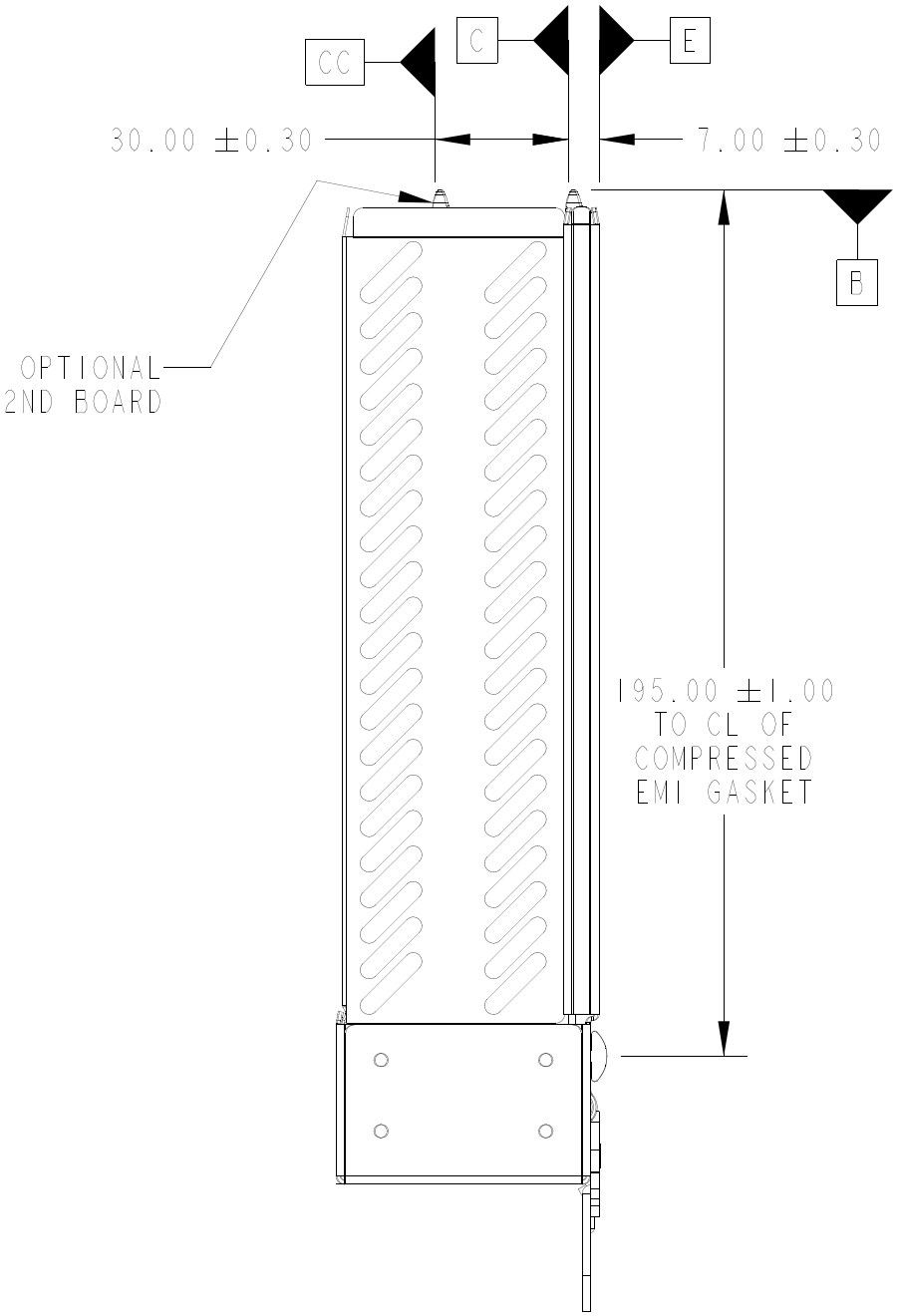


Figure 110 Tall Wide Critical Dimensions - Detail B

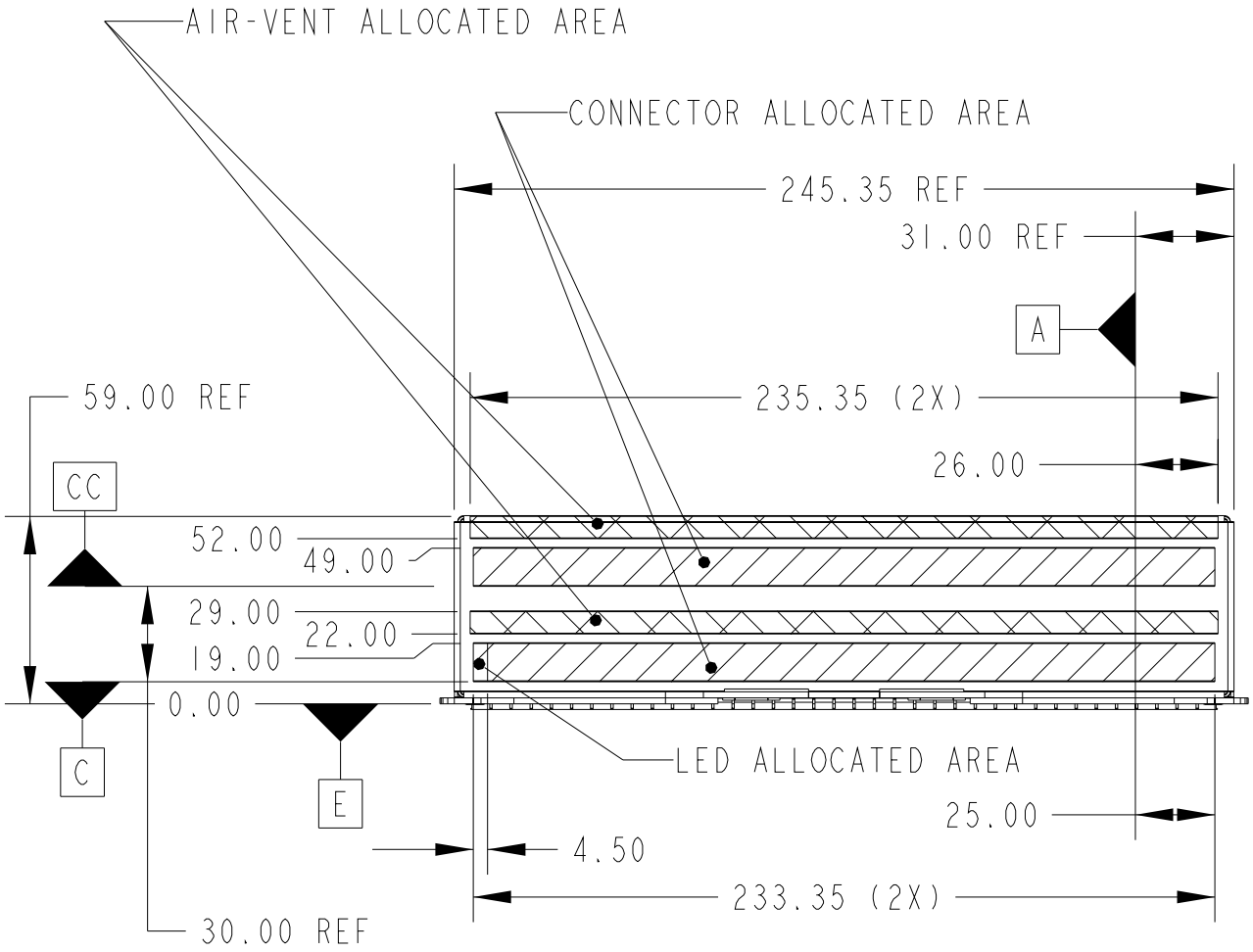


Figure 111 Tall Wide Critical Dimensions - Detail C

9.2.5 MODULE LEAD-IN

C9-7: In an effort to have a consistent "feel" during module insertion/extraction, all modules **shall** have the lead in features specified in Figure 112.

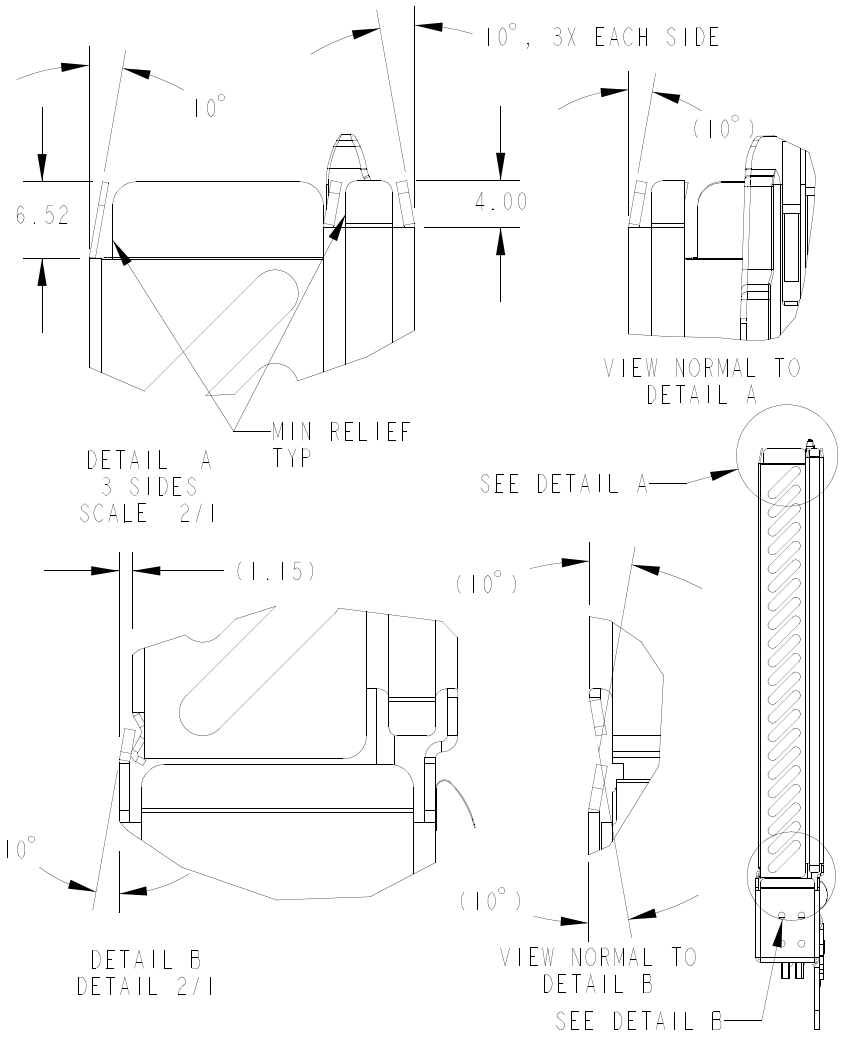


Figure 112 Module Lead-in Features, Standard Module

9.2.6 MODULE MATERIALS

C9-8: Module materials (carrier, ejector, latch) **shall** be galvanically compatible to zinc finishes and any EMI gasket material chosen. All module parts **shall** be free of burrs and sharp edges.

9.2.7 FILLER MODULES

C9-9: Each unoccupied IB Slot **shall** contain a filler module to maintain the EMI shielding of the faraday cage and specified airflow characteristics. The filler module **shall** operate within the minimum and maximum airflows established for normal modules (see Table 94 on page 399).

A filler module **may** consist of, but is not limited to, an empty module carrier without a cover and with the Standard Module ejector and latch.

9.2.8 MODULE INDICATORS

C9-10: All requirements detailed in Section 9.2.8 **shall** be followed regarding module LEDs.

Each InfiniBand Module **shall** have two indicator LEDs, one green and one yellow. These indicators **shall** be positioned at the upper portion of all modules (vertical module orientation) within the specified connector/LED area shown in Figure 99 on page 368, Figure 103 on page 372, Figure 107 on page 376, or Figure 111 on page 380.

All LEDs **shall** meet the following wavelength and luminescence requirements:

9.2.8.1 WAVELENGTH

Table 92 Module Indicator Specification

LED	Wavelength
Green (Blue-Green)	555 - 565 nm
Yellow (Amber)	582 - 592 nm

9.2.8.2 LUMINESCENCE

Module Indicators **shall** have a minimum luminescence of 80 candellas per square meter over an area of at least 1.5 mm by 1.5 mm.

9.2.9 MODULE COMPONENT HEIGHT

Component height limitations are governed by the module design and depend on material thicknesses of the board, carrier and cover. The location

of the primary side of the board in relation to the carrier is defined in Section 9.2.12 and can also be seen in the figures earlier in this section.

Note to Board Designers

There is decreased secondary side component height in the connector end of the module due to the recess for the ejector

9.2.10 MODULE EJECTOR AND LATCH DETAILS

Figure 113 through Figure 118 on page 390 show ejector and latch details and their relationships to respective modules and the chassis slots. Ejector may be removable, this will allow Ejectors to meet system vendor's requirement.

C9-11: All InfiniBand™ Modules **shall** have Standard ejectors.

Standard ejectors should comply with all dimensions specified in Figure 115 on page 387

C9-11.a1: No feature of the Ejector when assembled to the Carrier shall extend beyond datum "E".

C9-11.a2: When Ejector is latched no feature of the Ejector **shall** extend over the connector housing to an extent that it will interfere with the mated connectors for the module it is attached to. For a universal handle (works with all connectors that fit within the connector allocated area) this dimension **shall not** extend more than 2.8mm from Datum E over the carrier connector housing.

C9-11.a3: No feature of the Ejector, when in the latched (closed) position, **shall** protrude more than 35mm in front of the Ejector pivot centerline.

C9-11.a4: Ejector **shall** have no sharp edges. Ejector edges **shall** comply with UL1439 as a minimum condition for safe handling.

C9-11.a5: To prevent the Ejector from being damaged in normal handling outside of the chassis the Ejector "Latch" **shall** require a minimum of 1.4 Newtons (5 ounces) of force to release the Ejector from the "closed" position.

C9-11.a6: When the ejector is latched the Ejector latch mechanism **shall** prevent the Module from being inadvertently released from its slot when a force of 111 Newtons (25 pounds) is applied to the face of the Module in the direction of removal of the Module.

C9-11.a7: The Carrier **shall** include a hard stop feature that prevents the Ejector from extending beyond 95 degrees from the latched position.

C9-12: All Tall form factors (including Tall Wide) modules **shall** have Upper ejectors.

Upper ejectors should comply with all dimensions specified in Figure 116 on page 388

C9-13: Each Ejector **shall** have a latch mechanism.

Latch mechanisms for all All InfiniBand™ Modules should comply with all dimensions specified in Figure 117 on page 389

Figure 118 on page 390 specifically deals with module assembly details and shows how the camming mechanism works during insertion and extraction. The camming motion itself allows a more controlled insertion with decreased potential for backplane connector damage.

Note to Designers

Ejector mechanism must be designed to handle insertion/extraction forces equivalent to two backplane connector's insertion force with a reasonable margin of safety to also include frictional forces from guides and EMI gaskets. The Ejector should be strong enough to withstand a module insertion force of 40 lb.

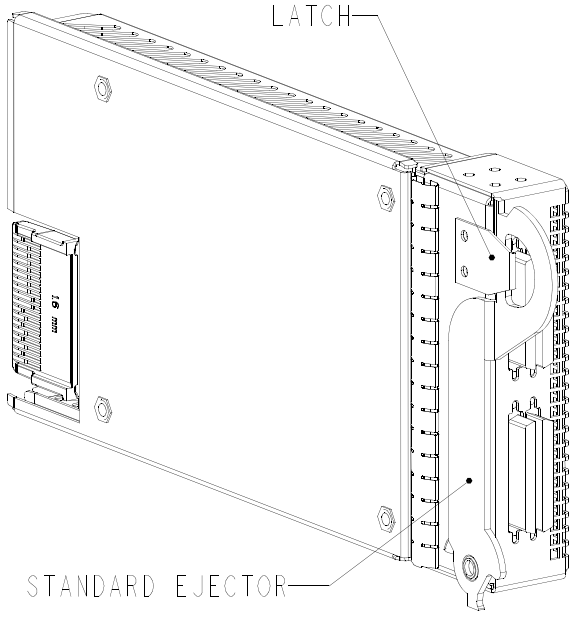


Figure 113 Standard Module With Ejector and Latch

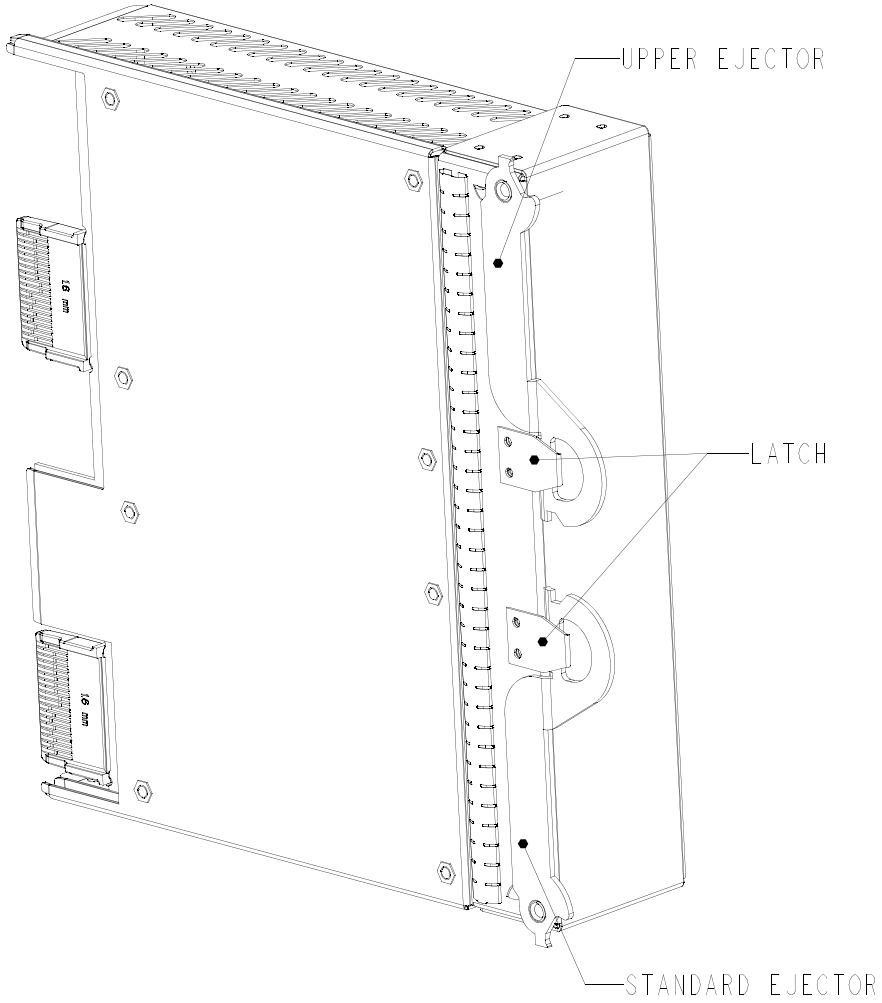


Figure 114 Tall Wide Module With Ejectors and Latches

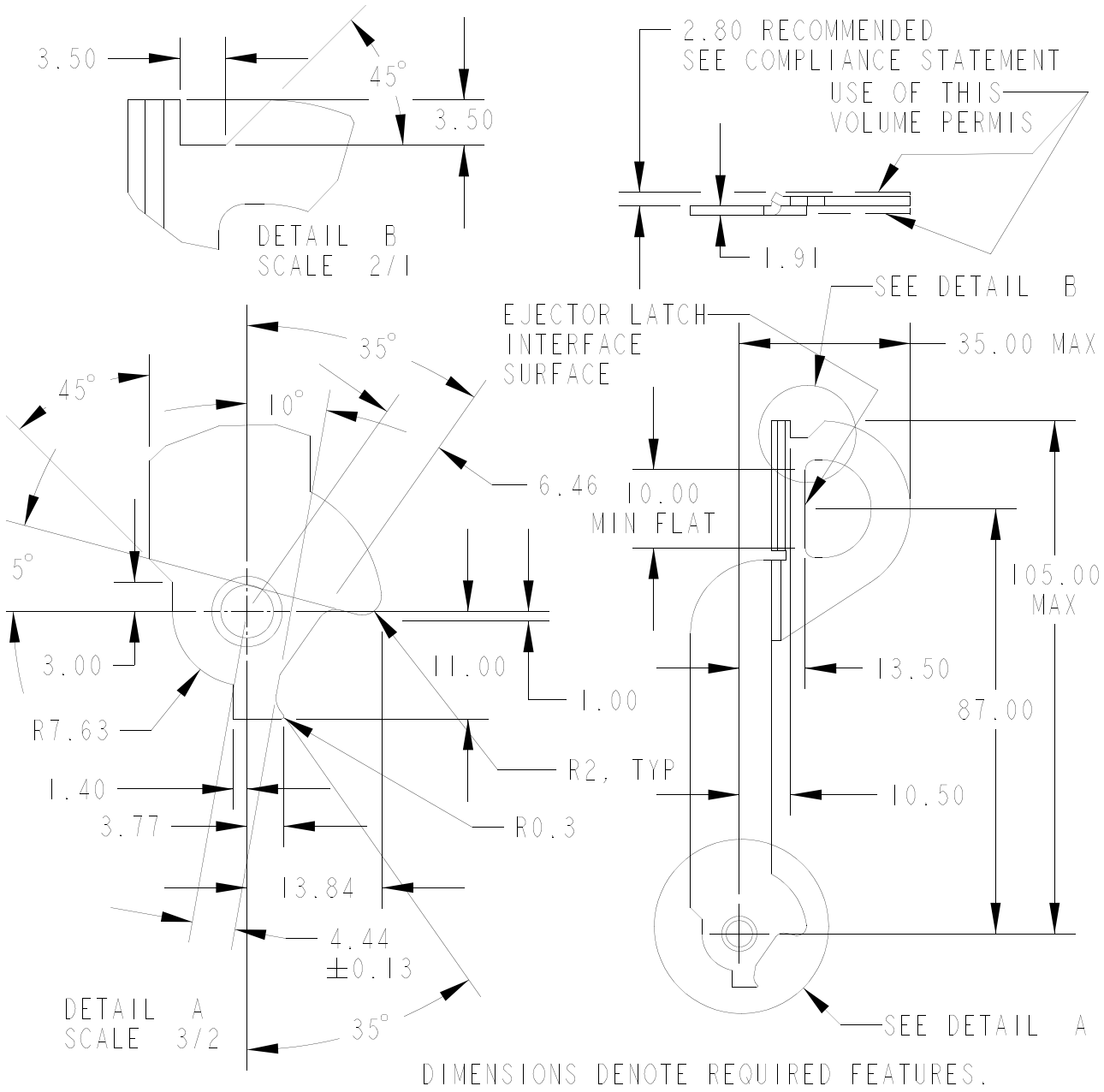


Figure 115 Standard Ejector Dimensions

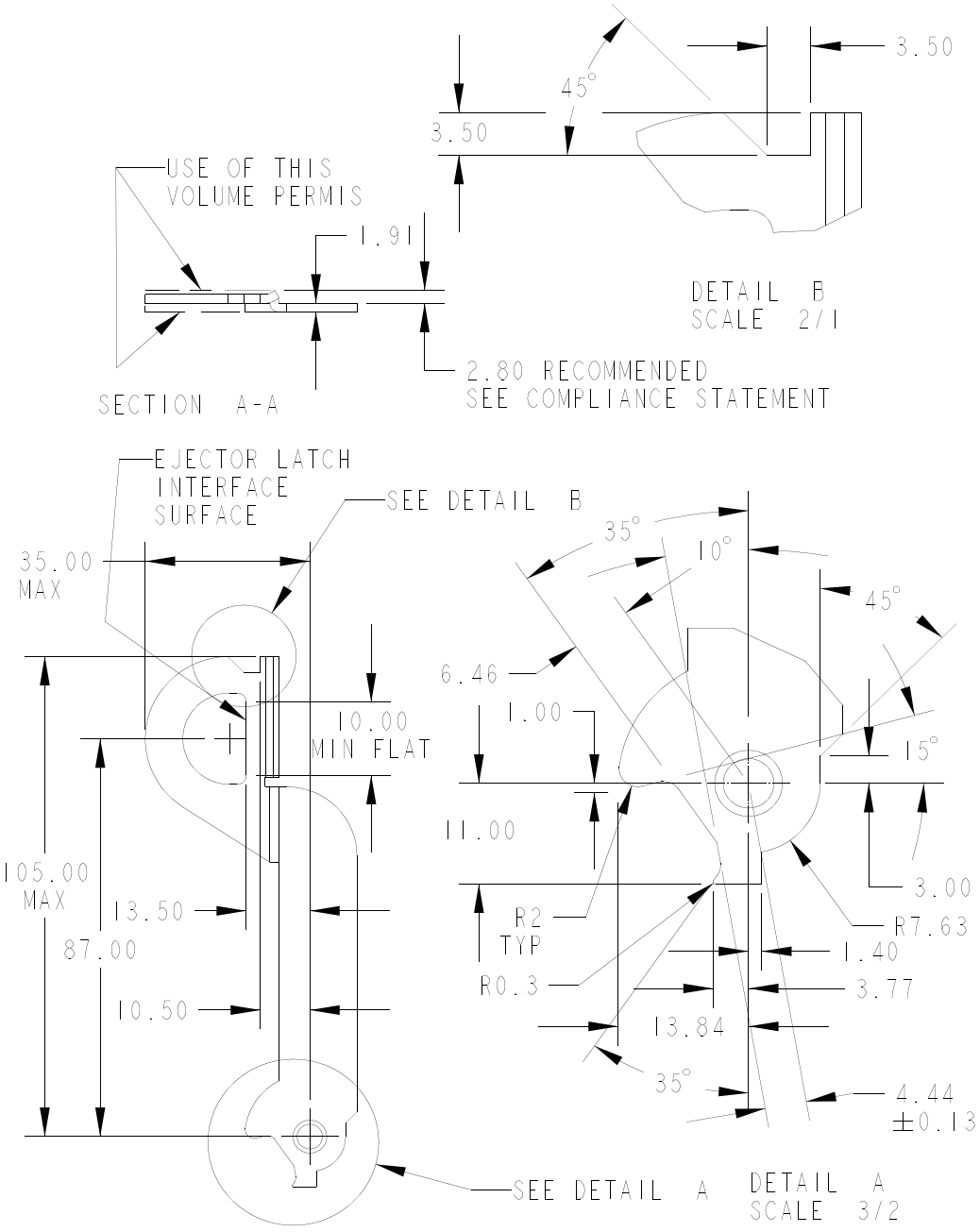


Figure 116 Upper Ejector Dimensions

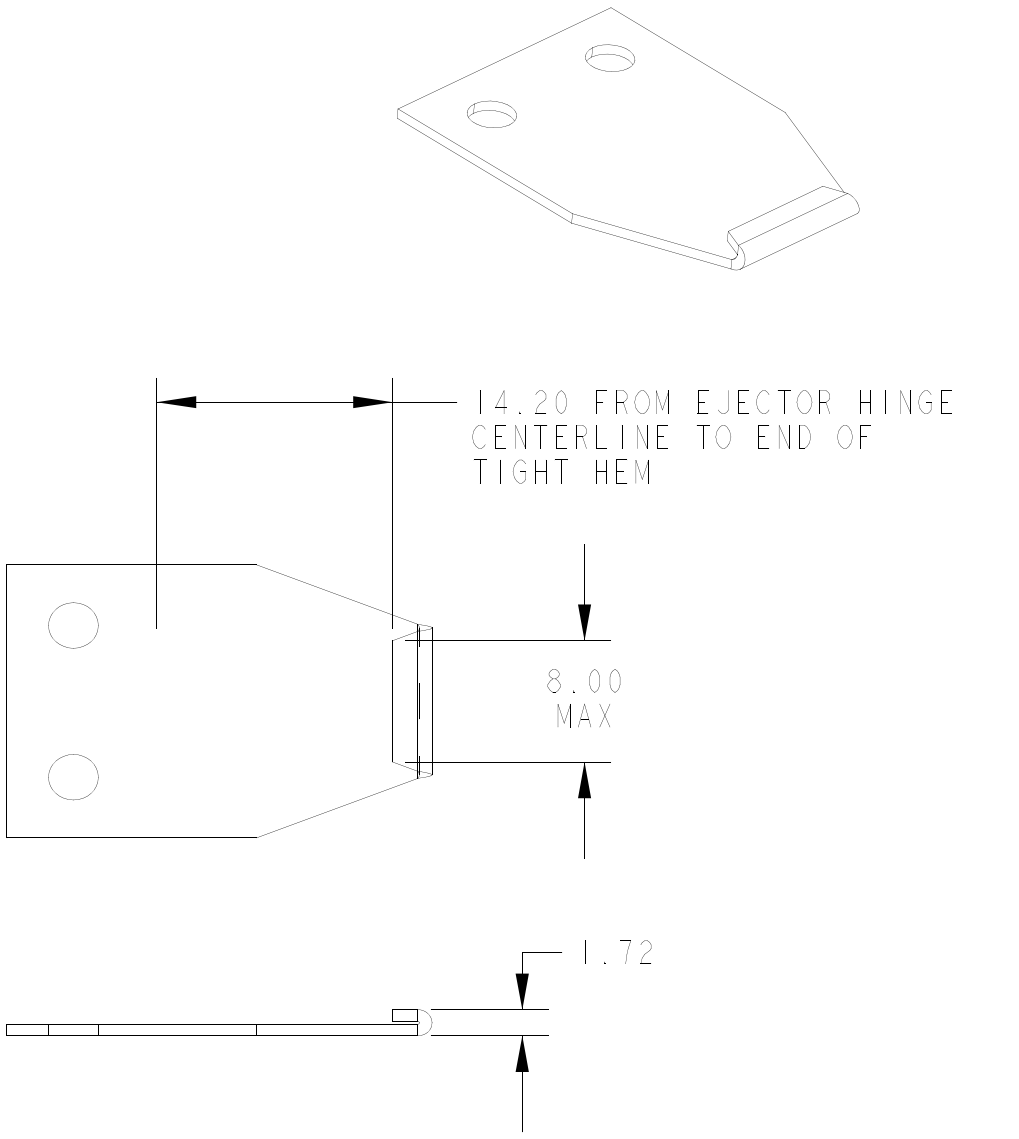


Figure 117 Latch Dimensions

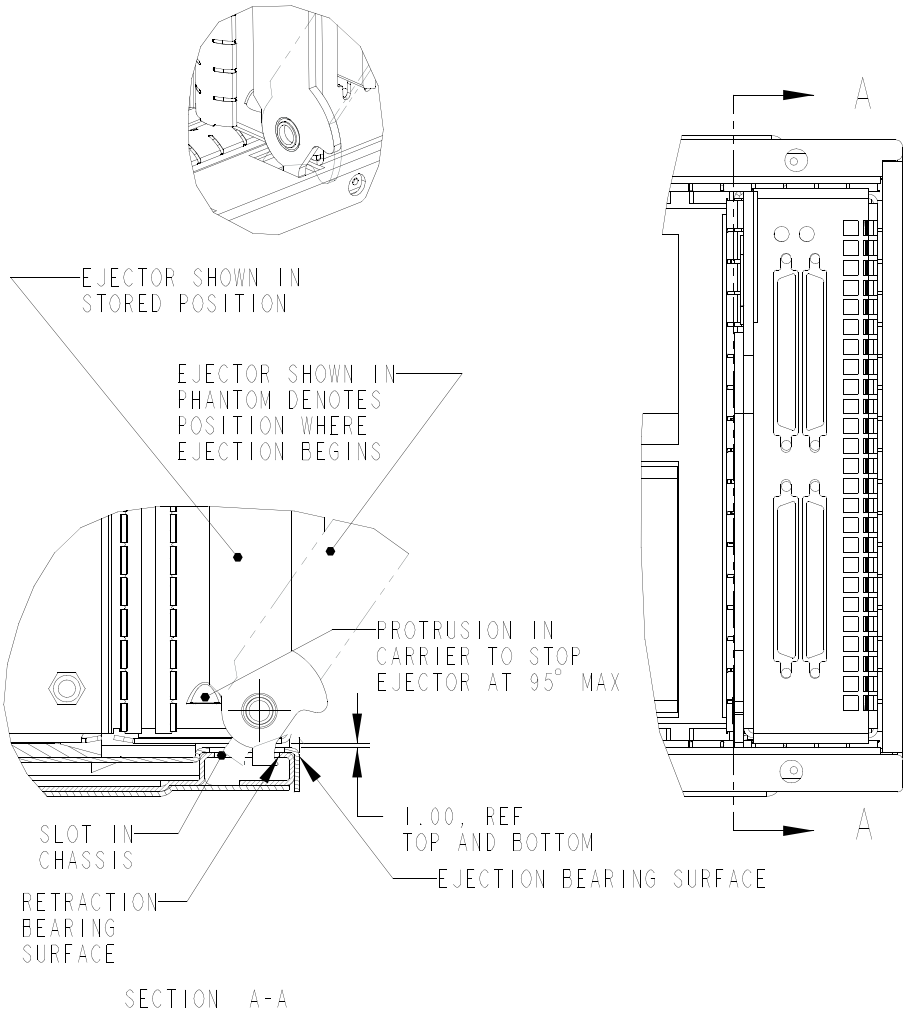


Figure 118 Chassis slot/Ejector Interface

9.2.11 MODULE COVER DETAILS

C9-14: InfiniBand Modules **shall** have a cover over the module component side. Module covers **shall** comply with the dimensions specified in Figure 95 on page 364: specifically cover width, parametric dimension "V" and cover height, parametric dimension "X".

The cover is used to prevent accidental contact of conductive components or solder leads with adjacent modules. It also prevents human contact with hot components, and provides ESD protection. The cover **may** also be used to control/channel airflow.

The design of the cover is not further specified, but it is recommended that all covers be easily removable, without tools, to aid the module manufacturing process.

Note to Designers

It is anticipated that most covers will be made from plastic with appropriate ESD treatment. If used, metal covers must make reliable electrical contact with the carrier to prevent EMI problems. An example of a module cover design can be found in appendix Section 1.1, "Module Design Examples," on page 729.

9.2.12 BOARD MOUNTING

C9-15: The PCB **shall** be secured to the module carrier. Regardless of the PCB thickness, the board primary side **shall** be located 7 +/- 0.3 mm from the outside of the carrier, datum E (see Figure 98 on page 367, Figure 102 on page 371, Figure 106 on page 375, or Figure 110 on page 379 for standard, tall, wide, or tall wide modules respectively).

This dimension will insure the PCB will properly mate with the backplane connector. Recommended board mounting points can be found in Figure 119 and Figure 120. These are recommendations **only** to facilitate some standardization of carrier design. These hole sizes, recommended in Figure 119 and Figure 120 are based on M3.5 size fasteners. The board designer is responsible for providing adequate mounting points to meet the EMI requirements of Section 9.5.4.

C9-16: Datums A and B are defined by the board connector position within the module. All Standard Modules **shall** comply with the dimensions that define these datums as shown in Figure 97 on page 366

C9-17: Datums A, AA, and B are defined by the board connector position within the module. All Tall Modules **shall** comply with the dimensions that define these datums as shown in Figure 101 on page 370

C9-18: Datums A and B are defined by the board connector position within the module. All Wide Modules **shall** comply with the dimensions that define these datums as shown in Figure 105 on page 374

C9-19: Datums A, AA, and B are defined by the board connector position within the module. All Tall Wide Modules **shall** comply with the dimensions that define these datums as shown in Figure 109 on page 378

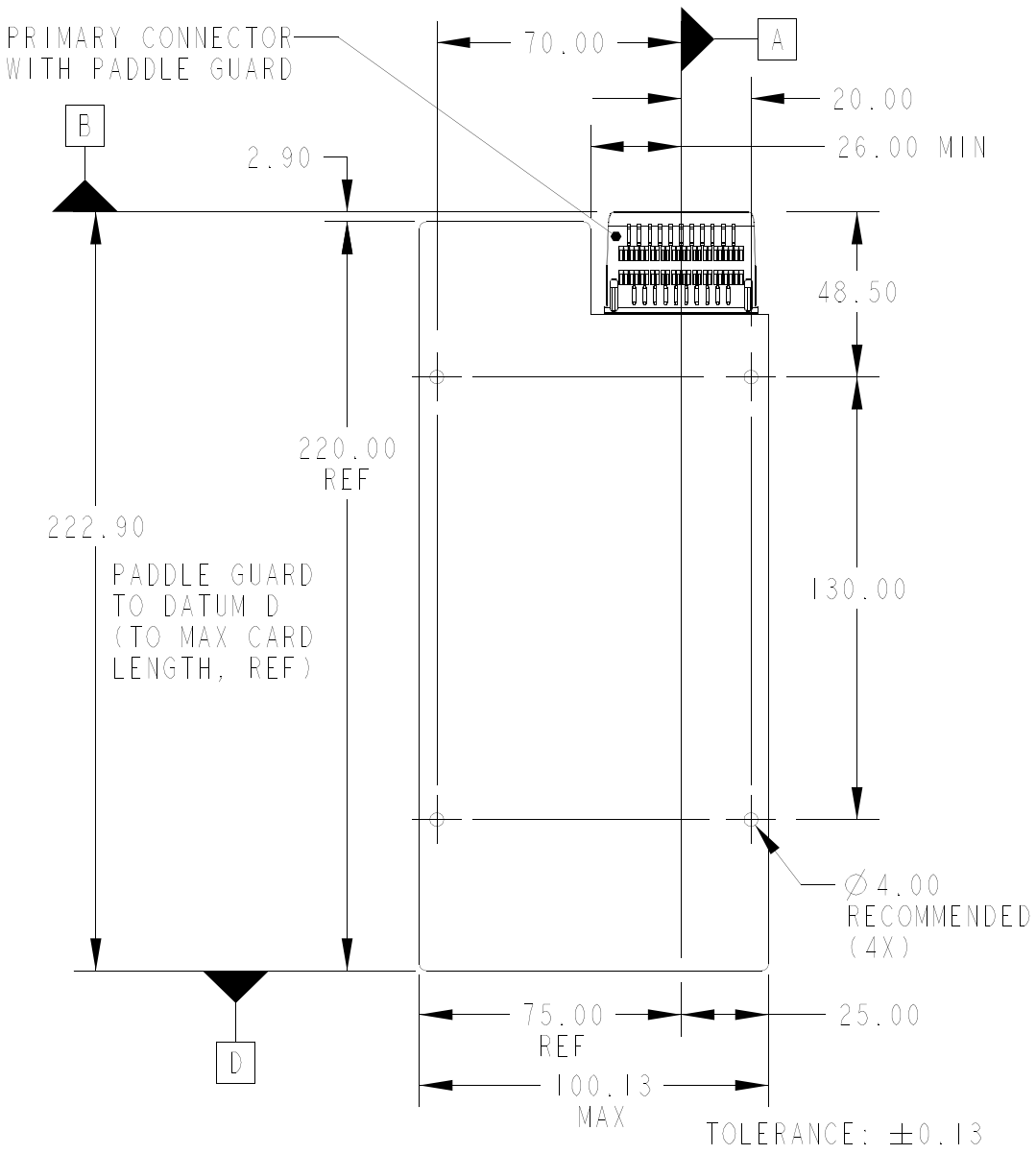
9.3 INFINIBAND BOARD PHYSICAL DESCRIPTION

9.3.1 BOARD DIMENSIONS

The dimensions included in this section describe the maximum allowable board dimensions within a module. The goal of this specification is to define the interface dimensions loosely to enable internal flexibility for the module designer.

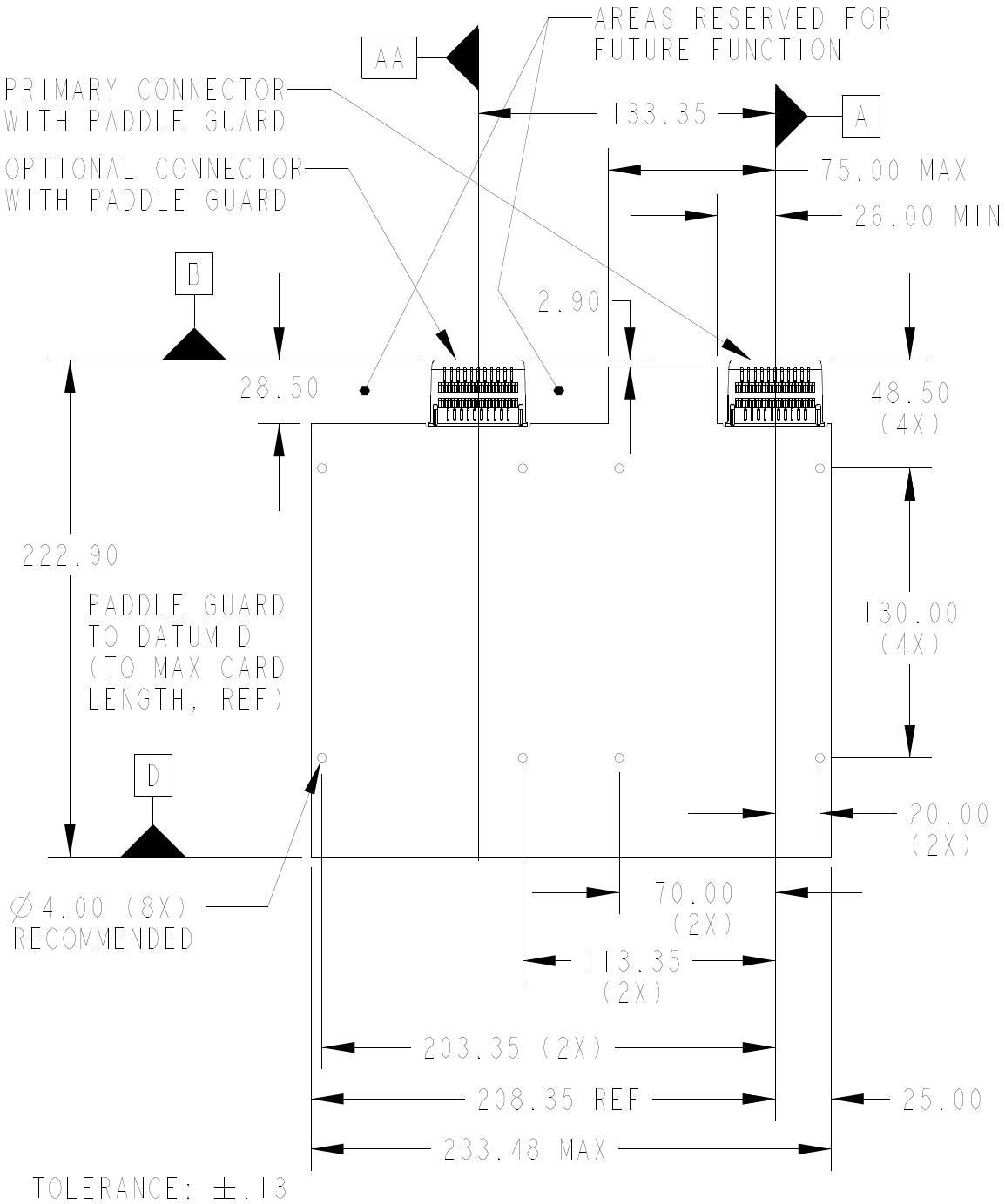
Figure 119 illustrates the board dimensions and connector location for a standard height module. Figure 120 illustrates the board dimensions and connector locations for a tall height module.

IBTA



Shown from Primary Side

Figure 119 Board Dimensions for Standard Module



Shown from Primary Side

Figure 120 Board Dimensions for Tall Module

9.3.2 PADDLE GUARD

A paddle guard is a requirement for every IB board edge connector. The paddle guard helps to separate connector contacts and it adapts different thickness boards to the connector. A detailed description of the paddle guard is given in Section 10.3.2, “Paddle Guard,” on page 414.

9.3.3 BOARD THICKNESS

C9-20: The thickness of InfiniBand printed boards is left to the module designer and **shall** range from 1.44 to 2.64 mm thick.

Although any thickness of board is allowable within this stated range, early adopters of InfiniBand™ **should** design to nominal thickness of either 1.6 or 2.4mm to take advantage of pre-tooled paddle guards. Regardless of board thickness, the mounting **shall** comply with the 7mm dimension referencing the board primary side to the module external surface (see Figure 98 on page 367, Figure 102 on page 371, Figure 106 on page 375, or Figure 110 on page 379).

9.3.4 INFINIBAND BOARD-EDGE CONNECTOR

Refer to IEC 61076 for detailed specification of the IB connector, board paddle routing, artwork, paddle guard details, and connector environmental information. For InfiniBand specific details for this connector, refer to Chapter 10: Backplane Connector Specification.

9.4 CHASSIS SLOT DESCRIPTION

Most of the mechanical definition for InfiniBand is centered around the module. InfiniBand does not define an entire "chassis"; it defines a minimal set of requirements for a "chassis slot".

9.4.1 BACKPLANE REQUIREMENTS

C9-21: Connector spacing between adjacent InfiniBand Slots **shall** be 30 +/- 0.10 mm (see Figure 121).

C9-22: Connector spacing between two ports in a tall slot **shall** be 133.35 +/-0.20 mm (see Figure 121).

Recommendation to Chassis/Backplane Designers

The backplane will likely be in the direct path of airflow for the module. Backplane height should be minimized to maximize module cooling.

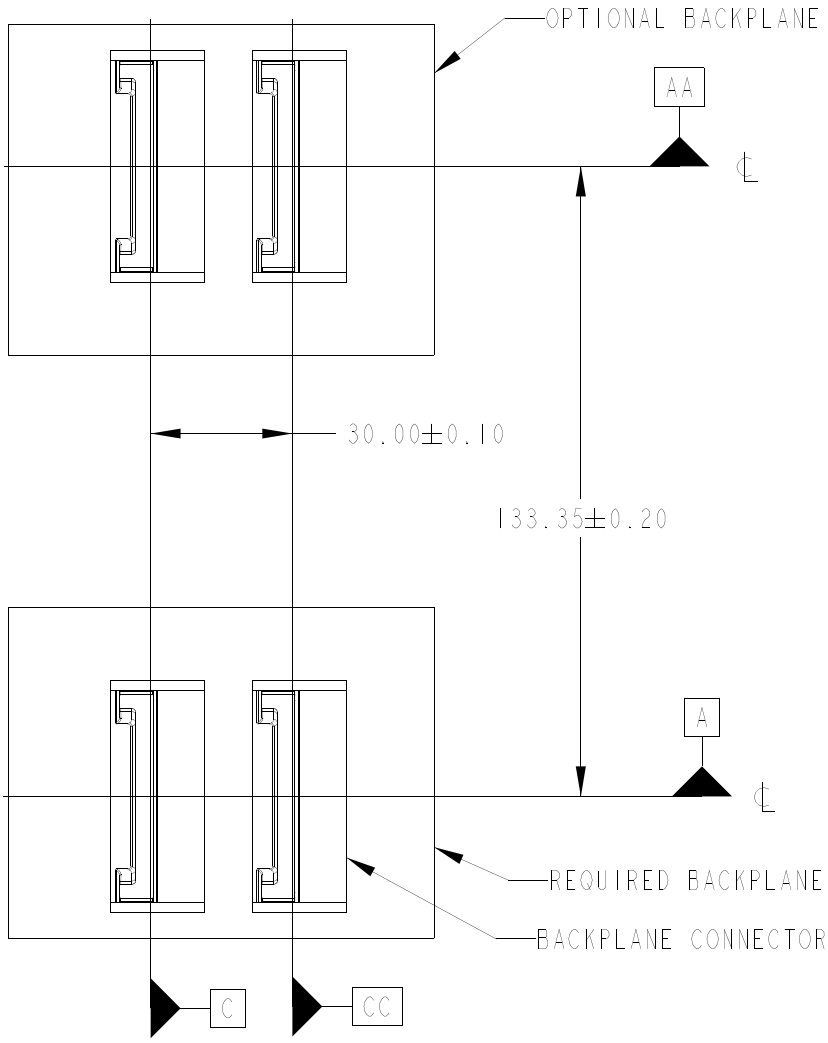


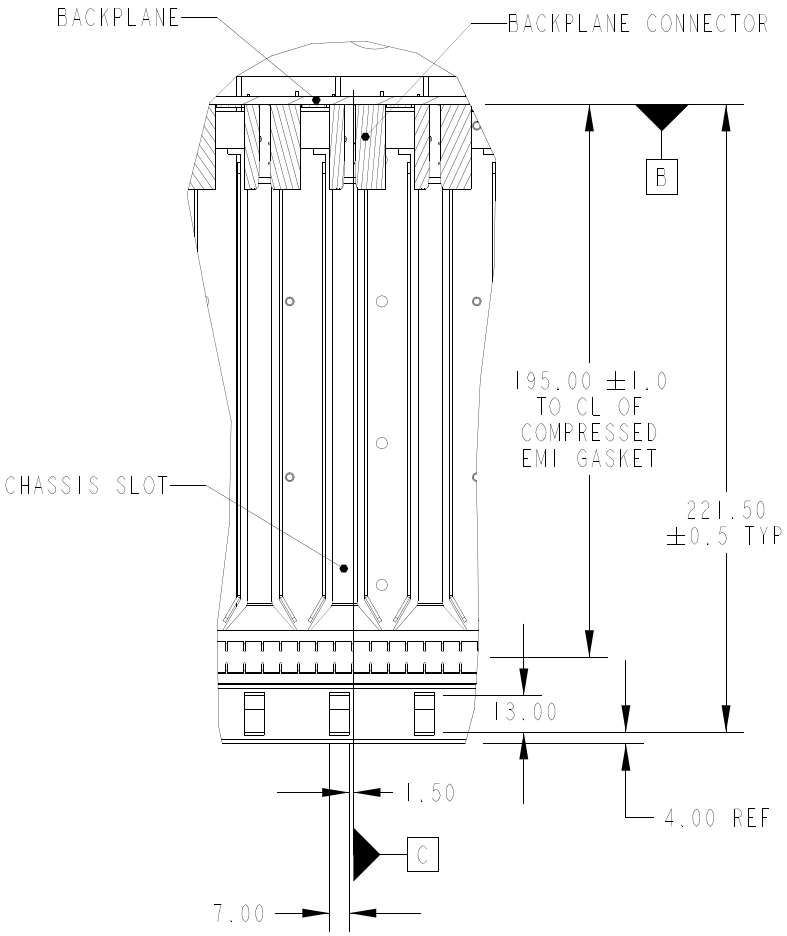
Figure 121 Backplane Connector Spacing

9.4.2 BACKPLANE CONNECTOR

Refer to Chapter 10: Backplane Connector Specification for connector details

9.4.3 CHASSIS SLOT DETAIL

C9-23: Chassis slots **shall** comply with all dimensions and tolerances specified in Figure 122.



Shown Through Connector- Datum A

Figure 122 Chassis Slot Dimensions

9.5 DESIGN CONSIDERATIONS

This specification is intended to provide a consistent and repeatable method for ensuring a minimum level of integrity and environmental performance. This section details a standard test process that all chassis,

modules, and back-planes specified in this standard can use to be compliant.

It is the system designer's responsibility to evaluate the relationship between the modules, chassis, and the system environment.

9.5.1 COOLING/THERMAL

The design objective is to supply the necessary airflow at the required temperature to insure that component temperatures are met. The chassis will provide sufficient cooling to meet all specifications when cooled by the fan(s) in the system enclosure. Fans are required to provide sufficient airflow to cool the modules under all conditions. Filler modules (or equivalent) will be required to prevent excessive airflow through spaces not occupied by modules.

9.5.2 COOLING REQUIREMENTS

The InfiniBand Module will be used in various ambient environments. To insure that the module meets temperature requirements for these various environments, two thermal profiles have been developed. Allowable inlet temperatures to the module for these thermal profiles are given in Table 93. Type "A" equipment refers to general purpose computing equipment and has been elevated under the assumption that the module receives pre-heated air. Type "B" refers to Network Equipment Building Systems (NEBS), Level 3.

Table 93 Operational Inlet Temperature Specification

Thermal Profile	Module Inlet Temperature (C)
A	0 - 45
B	0 - 55

To insure that components in the InfiniBand Module receive adequate cooling to meet reliability and temperature goals, minimum airflow rates per module have been specified. Table 94 lists minimum airflow requirements per module type. It is the system engineer's responsibility to meet the minimum flow rate requirements. To insure adequate airflow it is important to insure that the combined module and system resistance curve intersect the fan curve(s) at a point above the minimum required flow rate

(see Appendix, Section 1.2, "Computational Fluid Dynamics (CFD) Results," on page 736 for details).

Table 94 Minimum Airflow Rates Per Module

Module Type	Minimum Airflow, Type A cubic feet/minute (cfm)	Minimum Airflow, Type B cubic feet/minute (cfm)
Standard	6	9
Wide Standard	12	18
Tall	12	18
Wide Tall	24	36

C9-24: A chassis slot **shall** deliver the appropriate volumetric flow rate as specified in Table 94 within the temperature range specified in Table 93 on page 398.

9.5.3 MODULE THERMAL DESIGN

The design of the module vents is critical to insure proper airflow. If the module design has too much resistance, components in the module may not receive adequate cooling air. On the other hand if the resistance of the module is too low, airflow may bypass other modules with higher resistances, reducing cooling airflow to these modules and therefore raising component temperatures. To insure cooling requirements can be met and to reduce the impact of bypass, Figure 123 on page 401 has been constructed giving minimum and maximum allowable resistances per module.

In addition, Table 95 on page 401 has been constructed to give estimates of minimum and maximum vent areas per module to meet Figure 123. See Appendix, Section 1.3, "Module Pressure Drop Testing," on page 739 for details associated with pressure drop testing of modules.

C9-25: Longitudinal airflow is the presumed norm for most general purpose computing equipment since it suggests the most dense packaging. In the most dense system implementations where a vertical module occupies all but a fraction of a chassis' height, all of the module's cooling air is presumed to pass through the connector housing vents travelling in a longitudinal direction through the module. The increased airflow requirements of systems conforming to Table 94, type B, may not be achievable in a purely longitudinal cooling scenario. Additional chassis height may be required to open up plenums above and/or below the module and to potentially add additional chassis venting in parallel with the connector

housing venting. InfiniBand™ modules are required to have a minimum of open vent area along the top and bottom cover surfaces to facilitate additional airflow that would tend to flow in a more transverse direction. The decision is left to the system designer on whether or not to selectively open corresponding areas in the chassis to take advantage of transverse airflow. Regardless of the intended system implementation (longitudinal vs. transverse), all modules **shall** fall within the minimum and maximum resistance requirements for modules with longitudinal airflow shown in Figure 123 on page 401.

Note that the transverse airflow is the primary method for cooling in some market segments for example the telephone industry. Products targeted at these markets should insure adequate transverse venting to support the environment requirements. See annex A figure-143 for an example of a cover venting.

The curves in Figure 123 represent flow dependent resistance through a module. The pressure axis is constant regardless of module size. The flow rate axis has been differentiated for the Standard, Tall, Wide, and the Tall Wide size.

The connector housing areas specified for connectors may also be used for cooling provided that EMI attenuation is met.

Note to designers

The minimum flowrates described in Table 94 were analytically determined to produce similar component temperatures at the respective inlet temperatures (Table 93 on page 398). In other words, a module designed for Type A environments should achieve similar component temperatures in Type B environments given the Type B enclosure provides the specified airflow

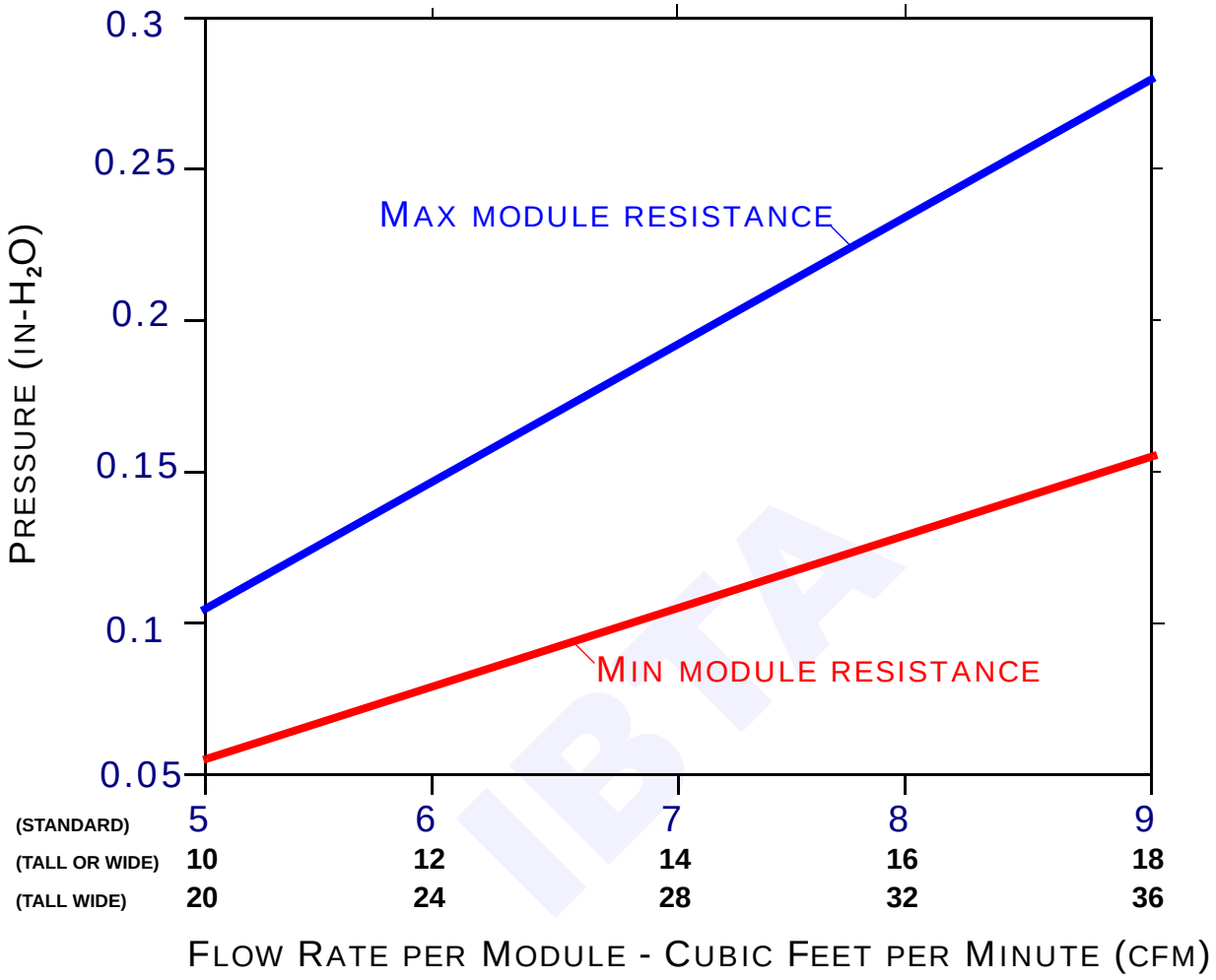


Figure 123 Module Resistances for Longitudinal Airflow

Table 95 Estimated Vent Area to Meet Specified Module Resistances

Module Type	Minimum Open Vent Area (mm^2)	Maximum Open Vent Area (mm^2)
Standard	500	750
Standard Wide	1000	1500
Tall	1000	1500
Tall Wide	2000	3000

9.5.4 EMI & ESD REQUIREMENTS

The implementation of the connection between Signal ground and Chassis ground is the responsibility of the system integrator. That connection should be made at a point that minimizes Electromagnetic Interference (EMI).

C9-26: Although not recommended, it is permissible to connect signal ground and chassis ground together on an InfiniBand™ module. If such connection(s) must be made, the InfiniBand™ module **shall** be fully EMC tested and qualified with this connection in place, and the module **shall** continue to meet the ESD specifications of sections 9.5.4.4.1 or 9.5.4.4.2 of the InfiniBand™ Mechanical specification, as well as all applicable FCC and EC EMC regulatory requirements. Since a logic to frame ground connection has a great potential for causing EMC problems if done improperly, modules making this connection **shall** be EMC tested with all supported I/O cables attached while exercising all I/O interfaces using typical customer data. Note that, in this instance, idle data alone are not sufficient to test the module I/O interfaces, as the InfiniBand™ idle data pattern has been defined to minimize EMI. Typical customer data **shall** be used when verifying EMC compliance.

C9-27: The module **shall** make a near continuous (no gap greater than 4 mm) EMI seal with the chassis and surrounding modules around its perimeter.

This is necessary in order to meet the stringent requirements of Figure 124.

C9-28: The module **shall** provide an EMI gasket located per Figure 98 on page 367, Figure 102 on page 371, Figure 106 on page 375, or Figure 110 on page 379.

C9-29: The chassis **shall** provide an EMI gasket on the top and bottom surfaces of the slot that mate with the module per Figure 122 on page 397.

C9-30: The gasket material **shall** be used with a nominal compression of up to 50% assuming it is compressed in a nominal gap of 1 mm. The gasket material **shall** be such that sliding contact causes no degradation. The gasket material **shall** have a finish that is galvanically compatible with zinc.

9.5.4.1 EMI

The IB chassis and modules are of closed construction and are designed to maintain an EMI environment compatible with worldwide regulatory requirements for various classes of equipment. The modules and chassis, with conductive gaskets at the seams form a Faraday cage enclosure. The modules themselves will be an integral part of the chassis EMI containment strategy.

C9-31: Attenuation levels of Figure 124 **shall** be met by **all** module carriers regardless of the relative level of internal module emissions.

The intent of Figure 124 is to provide a known level of containment at the carrier connector housing as an integral part of the system EMI enclosure. The curve specifies containment provided by the carrier as tested in a defined carrier attenuation reference chassis and in no way specifies the completed module EMI containment characteristics nor provides any assurance that the module will pass industry standards. Refer to normative appendix Section 1.5, "Carrier Attenuation Test Procedure (Normative)," on page 749 for EMI test procedures and appendix A1.6 Fabrication and Assembly of the EMI Reference Chassis on page 754 for the definition of the carrier attenuation reference chassis.

There is a carrier guide surface provided to allow a continuous contact between module top and bottom and chassis (conductive, resistive, or iso-

lated). This continuous contact along the length of the module is intended to limit interference between modules.

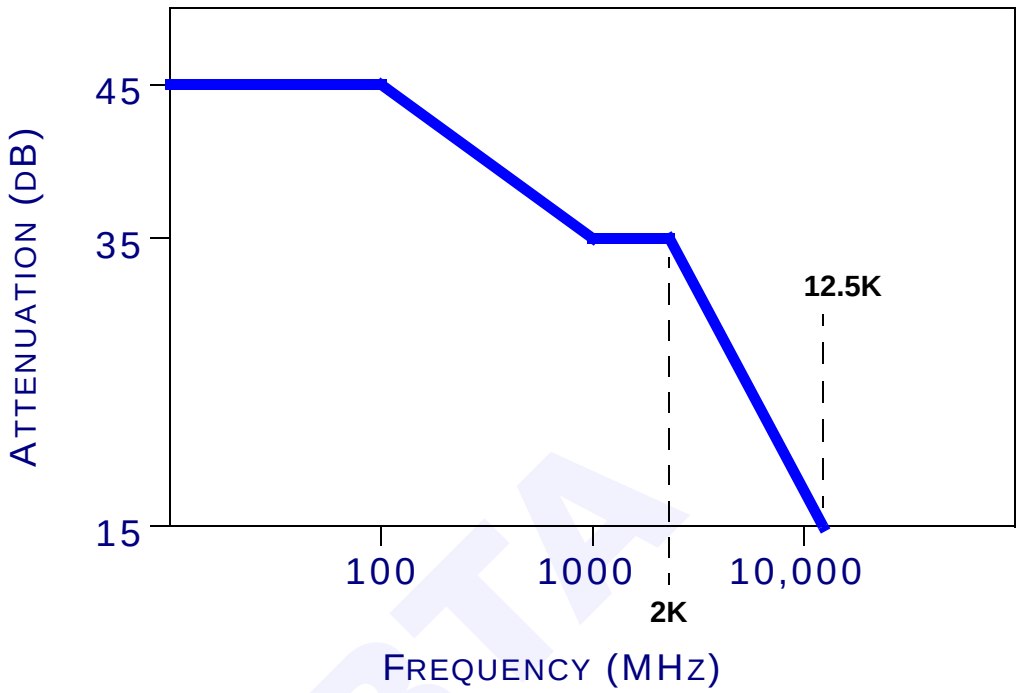


Figure 124 Required Carrier Attenuation

9.5.4.2 IB MODULE AIR VENT DESIGN FOR EMI

In order to meet the specification for shielding effectiveness given in Figure 124, the number, size, and spacing of the vent holes and the thickness of the connector housing must be controlled. To meet this specification, alternative vent designs have been evaluated using Finite Difference Time Domain (FDTD) modeling of an electric field source near the module connector housing. These results indicate that the required shielding effectiveness may be obtained by the designing the holes per Table 96.

Table 96 IB Module Vent Design for EMI Shielding

Hole Size (LxW, mm)	Conn. Housing Thickness, mm	Hole Spacing, Center-Center, mm
2 x 2	0.5	3

Table 96 IB Module Vent Design for EMI Shielding

Hole Size (LxW, mm)	Conn. Housing Thickness, mm	Hole Spacing, Center-Center, mm
2 x 2	1	3
4 x 4	2	5
6 x 6	4	7
8 x 8	8	9

These vent dimension design points will permit the maximization of the air vent area (assumes no connectors) while still allowing an IB Module to meet the required EMI shielding effectiveness. If more shielding is necessary and less open area is needed for airflow, the diameter and number of holes may be reduced to the level required to provide the minimum open vent area as specified in Table 95 on page 401.

There are three levels of EMI containment that the module vendor must create; the first level is base carrier containment per the curve specified in fig which may be met simply by adhering to the guidelines in this section. The second level is an extension of the carrier mechanical containment which may be needed to compensate for holes created by various connectors. The first and potentially the second levels are required to meet the InfiniBand™ specification. The third level, which is not addressed by this specification, is additional suppression at the board/component level which may be required for the product to pass agency tests.

9.5.4.3 IB EXTERNAL CONNECTOR EMI REQUIREMENTS

In addition to controlling the vent holes, connector openings in the connector housing must also be controlled. The maximum single opening allowed for an unshielded connector is a 10mm waveguide.

Note to designers

Waveguide thickness for one or more unshielded openings should conform to Table 97. All other unused or unshielded connector openings should be capped or provided with a second internal layer of shielding.

Table 97 Connector Housing Design Requirements for Normally Unshielded Connectors (i.e. Optical, Unshielded Twisted Pair, etc.)

Connector Hole Size, mm	Number of Holes	Conn. Housing Thickness, mm
10 x 10	1	4
10 x 10	2	10
10 x 10	4	10
10 x 10	8	15

Refer to Appendix, Section 1.4, "Shielding Effectiveness," on page 743 for the shielding attenuation provided by the waveguides shown in Table 97.

To avoid the use of waveguides for optical and other usually unshielded connectors, the connector itself must be shielded and its maximum aperture size limited to 4 mm.

9.5.4.3.1 SHIELDED CONNECTORS

D Shell or other shielded external connectors used on InfiniBand Modules must also be grounded through a full 360 degree contact around the interface between the connector and connector housing.

9.5.4.4 ESD

C9-32: All modules **shall** meet the appropriate ESD requirements described in Section 9.5.4.4

The ESD performance of the module **shall** meet the requirements of EN55024 following the procedures specified in IEC 61000-4-2. This standard requires that the InfiniBand Module provides sufficient shielding and grounding to meet ESD discharges of the appropriate category:

9.5.4.4.1 GENERAL PURPOSE COMPUTING EQUIPMENT

Level 2 test of 4 kV contact and 8 kV indirect both polarities

9.5.4.4.2 NETWORK EQUIPMENT BUILDING SYSTEMS (NEBS)

Level 4 test of 8 kV contact and 15 kV indirect both polarities

IBTA

Note to Designers

Meeting this requirement may require the use of light pipes or other techniques to recess the LEDs behind the connector housing surface.

9.6 CHASSIS AND CABLE ICONS

InfiniBand has leveraged the use of industry connectors for both copper and optical cables solutions. Because of this it is important to clearly identify InfiniBand cables and chassis connectors to avoid confusion and interoperability issues in multi-system environments. This highlights the need for a cable / connector icon.

9.6.1 ICONS

The following icons have been chosen to identify InfiniBand connectors for both cables and chassis. There is one icon for each cable width identified below. The use of the icon is optional but highly recommended.

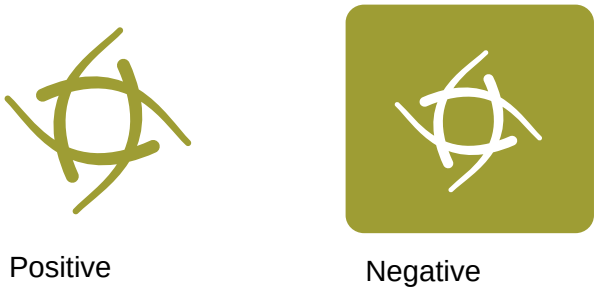


Figure 125 1x Icon

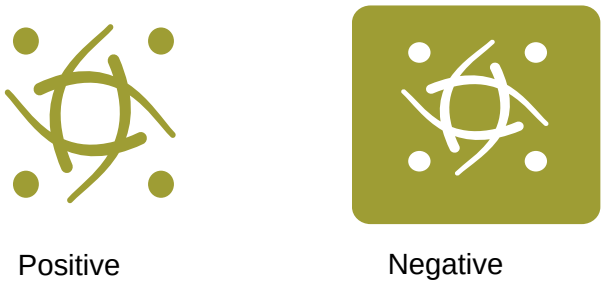


Figure 126 4x Icon

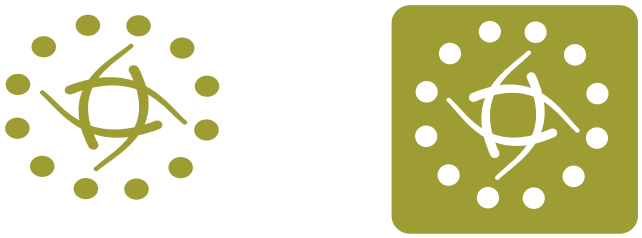


Figure 127 12x Icon

o9-32.1.1: If an icon is associated with an InfiniBand port, it shall be a scaled version of the icon in Figure 125, Figure 126, or Figure 127 that represents the interface width of the associated port. The icon shall not be less than 5mm height and 5mm in width.

The physical placement (top, bottom, side) and means (label, stamping or molding) is left to the implementation.

o9-32.1.2: If an InfiniBand icon is used to identify a copper or optical port and color coding is used to further enhance identification, Pantone 399 C (Green) shall be used.

Recommendation: If the InfiniBand icon is not used for port association and a color scheme is used., it is recommended that Pantone 399 C (Green) be used.

o9-32.1.3: If an icon is associated with an InfiniBand copper cable or optical fiber, it shall be a scaled version of the icon in Figure 125, Figure 126, or Figure 127 that represents the interface width of the associated cable or fiber. The size and contrast of the icon shall allow it to be visible from a distance of 0.5m under standard office lighting (500 lux).

Recommendation: If an icon is affixed, stamped or molded on an Infini-Band copper cable, it is recommended that it be placed in the space noted as "Icon Area" in Figure 61: 1x cable plug, Figure 62: 4x cable plug or Figure 64: 12x cable plug for copper cables or Figure 54 on page 234 of chapter 7 for optical pluggables.

CHAPTER 10: BACKPLANE CONNECTOR SPECIFICATION

10.1 INTRODUCTION

The connectors used to attach InfiniBand boards to InfiniBand backplanes are defined in this section. These connectors are a one-piece design mounted on the backplane, into which the InfiniBand board is inserted. One example of a suitable connector is specified in a draft New Work Proposal to the International Electrotechnical Commission (IEC). The detailed connector specification is included in final draft International Standard specification IEC 61076-4-115.

Two connectors are defined, one type for use with 1x and 4x boards, and a second type for 12x boards. InfiniBand™ boards and backplanes **shall** incorporate features shown in the appropriate sections below that specify the connector to board and connector to backplane interfaces. The 1x/4x and 12x connectors are externally identical, but the backplane footprint for the 1x/4x connector may have a reduced footprint from that used for the 12x connector if desired, due to the smaller number of contacts utilized. The board footprint for the connector contacts is the same in all cases.

The board edge "paddle" which mates with the connector is covered by a "paddle guard," a plastic shroud that fits over the paddle to protect the contacts during insertion and withdrawal of the board from the backplane connector. The paddle guard also serves to initiate closure of the connector housing when the board is inserted.

It is the responsibility of the connector supplier to perform the indicated tests on any backplane connectors to be used for InfiniBand boards and supply the data to potential customer companies to indicate compliance. It is recommended that the appropriate test groups specified in EIA-364.1000.01 for one piece connectors be used for qualification testing, using the following conditions:

- 50 mating cycles preconditioning
- Unmated exposure
- Five year product life
- Field operating temperature range up to 60 degrees C.

10.2 CONNECTOR DESCRIPTION

The InfiniBand backplane connector is a low insertion force connector with two sets of contacts. One set of contacts is used on the primary side of the InfiniBand board for high-speed differential pair signals and their

corresponding grounds. A second set of contacts is used on the secondary side of the board for low-speed signals, power, and ground. The 12x connector contains 24 pairs of high-speed contacts (48 pins) and 18 low speed/power contacts. The 1x/4x connector contains eight pairs of high-speed contacts and the same number (18) low speed/power contacts.

Closure of the mechanism that engages the high-speed contacts is achieved by an internal mechanism which is actuated by outline features on the paddle guard. It is thus vitally important that the paddle guard outline not deviate from that described in the sections below. Sequencing of low speed/power contacts for hot insertion and withdrawal of boards is accomplished through the use of contact staggering, controlled by openings in the paddle guard, as described in Section 10.3.2.

Figure 128 shows a diagram of the backplane connector mounted on a backplane and an exploded board assembly with paddle guard to be inserted into the connector.

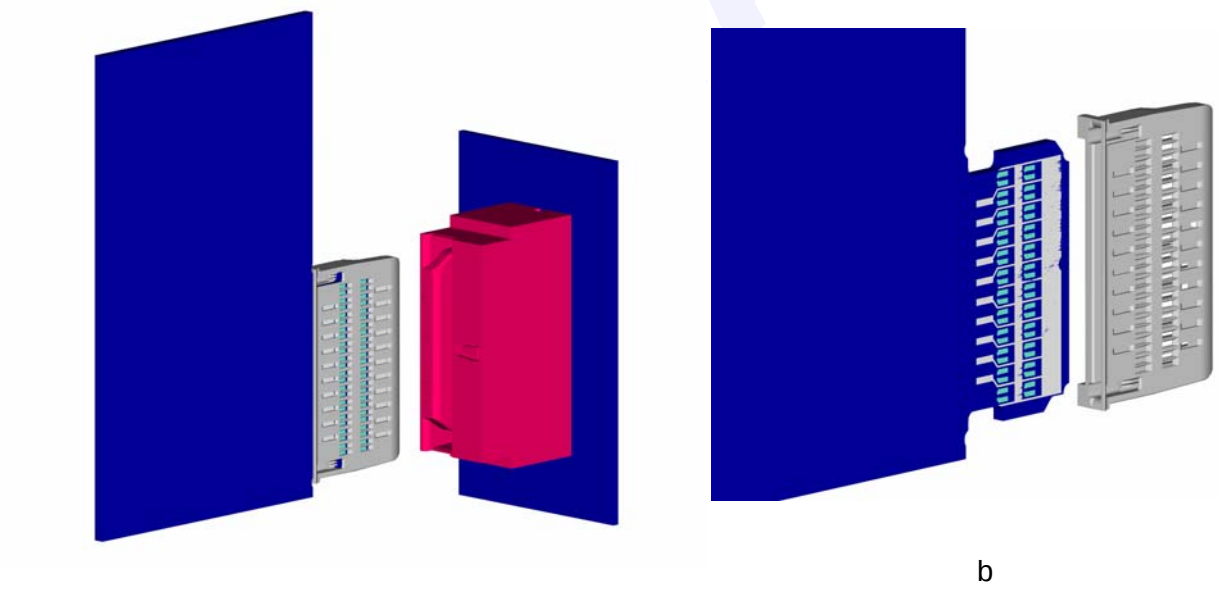


Figure 128 Backplane connector and board paddle (a), paddle guard assembly (b)

The mechanical, electrical, and environmental performance requirements for InfiniBand backplane connectors are defined in Section 10.5, Section 10.6, and Section 10.7 respectively.

10.3 CONNECTOR TO BOARD PHYSICAL REQUIREMENTS

Note that all drawing dimensions in the following sections are in millimeters (mm), and are in accordance with ASME Y14.5M-1994. The connector drawings contained in this document are intended for reference purposes only. The reader is encouraged to consult IEC specification 61076 and the respective supplier's drawings for detailed design information.

10.3.1 BOARD CONTACT PATTERNS

C10-1: The contact patterns shown in Figure 129 **shall** be used on the primary side of InfiniBand boards for high speed contacts to the backplane connector.

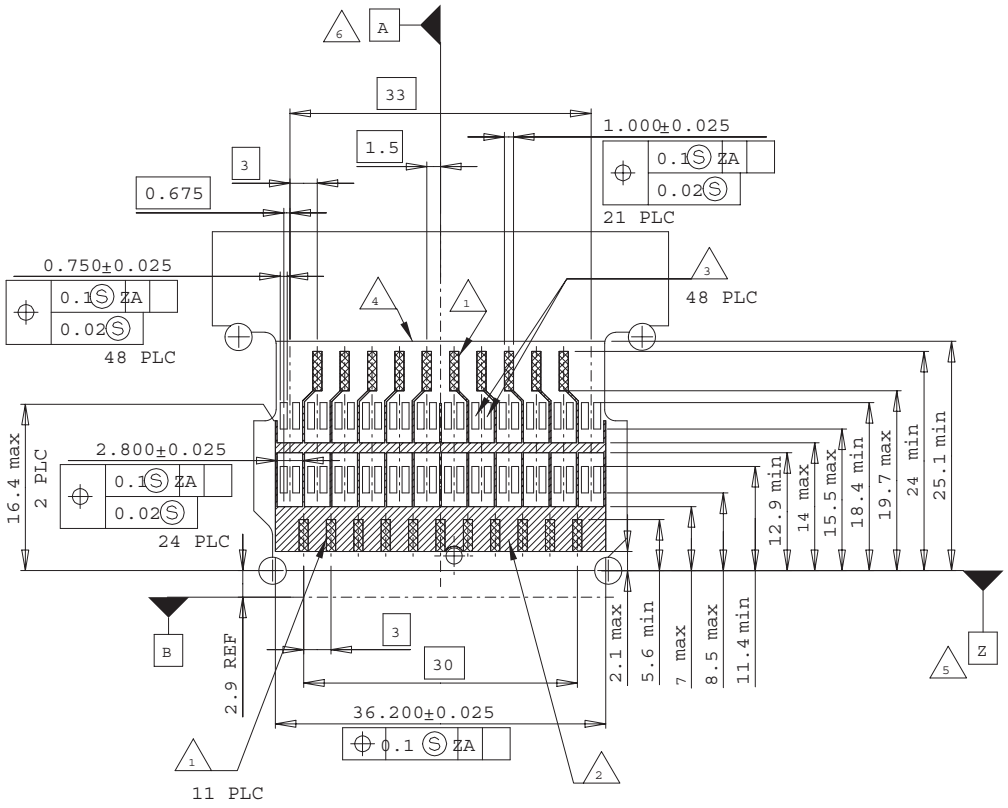
C10-2: The contact pattern shown in Figure 130 **shall** be used on the secondary side of InfiniBand boards for low speed and power contacts.

C10-3: The board paddle design shown in Figure 131 **shall** be used on InfiniBand boards to insure interoperability.

As mentioned in the previous sections, the sequencing of signals for hot insertion and withdrawal is controlled by the openings in the paddle guard, which are defined in the following section.

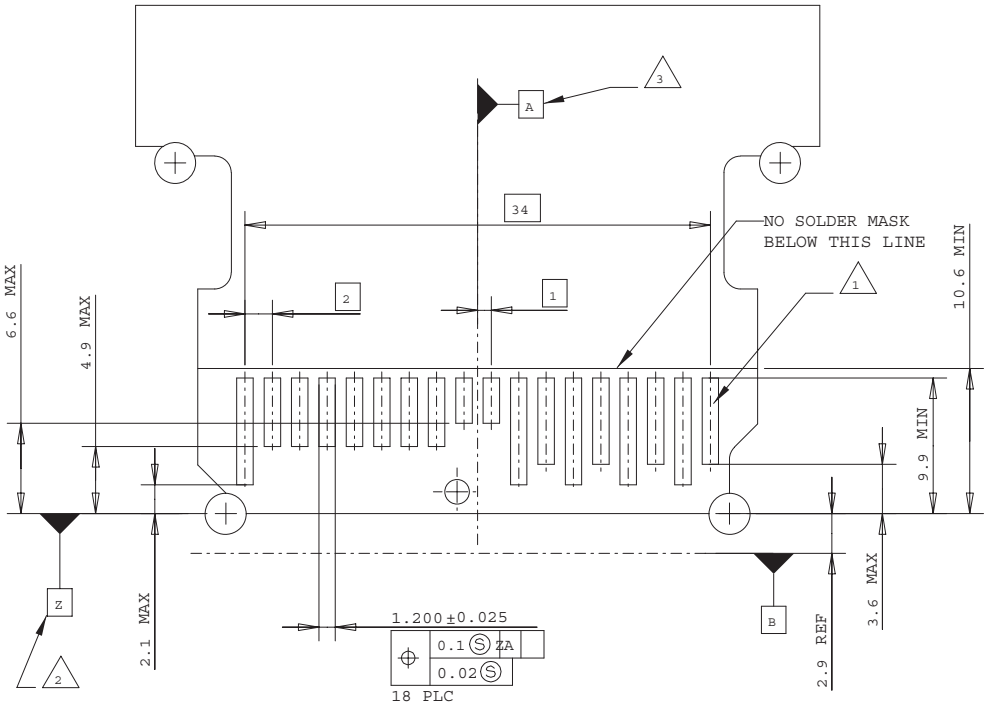
IBTA

The requirements for contacts used on InfiniBand boards are defined in Section 10.5, Section 10.6, and Section 10.7.



1. PAD FOR GROUND FRAME, NO VIA HOLES IN THIS AREA.
2. PREFERRED GROUND AREA.
3. PAD FOR SIGNAL CONTACTS.
4. NO SOLDERMASK BELOW THIS LINE.
5. ALL DIMENSIONS FROM REFERENCE Z = CENTER OF 2 DRILLED HOLES ON BOTTOM OF PCB.
6. CENTER OF 4 DRILLED HOLES WITH ø3.

Figure 129 Board contact pattern - primary side of board



1 PAD FOR LOW SPEED/POWER CONTACTS, NO VIA HOLES IN THIS AREA.

2 ALL DIMENSIONS FROM REFERENCE Z = CENTER OF THE 2 DRILLED HOLES ON BOTTOM OF PCB.

3 CENTER OF 4 DRILLED HOLES WITH ⌀3.

Figure 130 Board contact pattern - secondary side of board

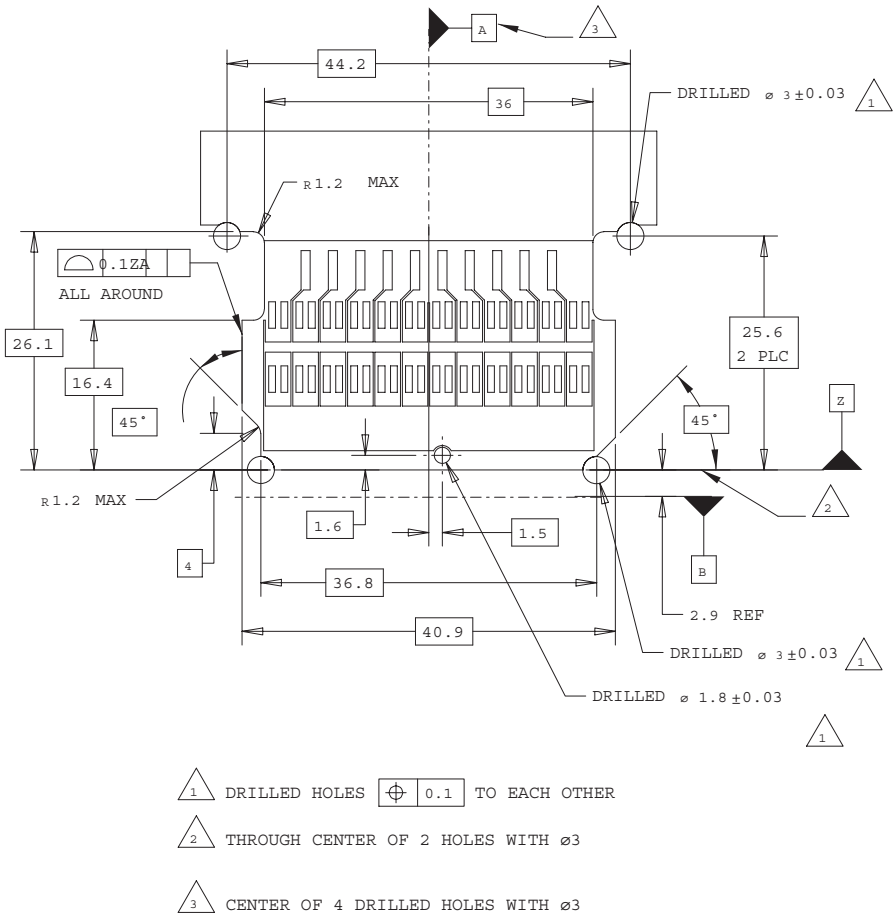


Figure 131 Board paddle outline - primary side of board

10.3.2 PADDLE GUARD

C10-4: InfiniBand boards **shall** utilize a paddle guard, a plastic shroud that fits over the board connector tab to protect the contacts during insertion and withdrawal of the board from the backplane connector. The paddle guard also actuates the closure of the high speed contacts and serves as a pin alignment and pin sequencing mechanism.

C10-5: The mechanical design shown in Figure 132 through Figure 135 **shall** be used for the paddle guard used on InfiniBand boards.

The paddle guard is designed to touch the InfiniBand backplane (as indicated by Datum B in Figure 132) when the module is fully inserted. The paddle guard offsets the IB board itself from the backplane; the exact relationships between the connector, board, and backplane are defined beginning in Section 9.3.

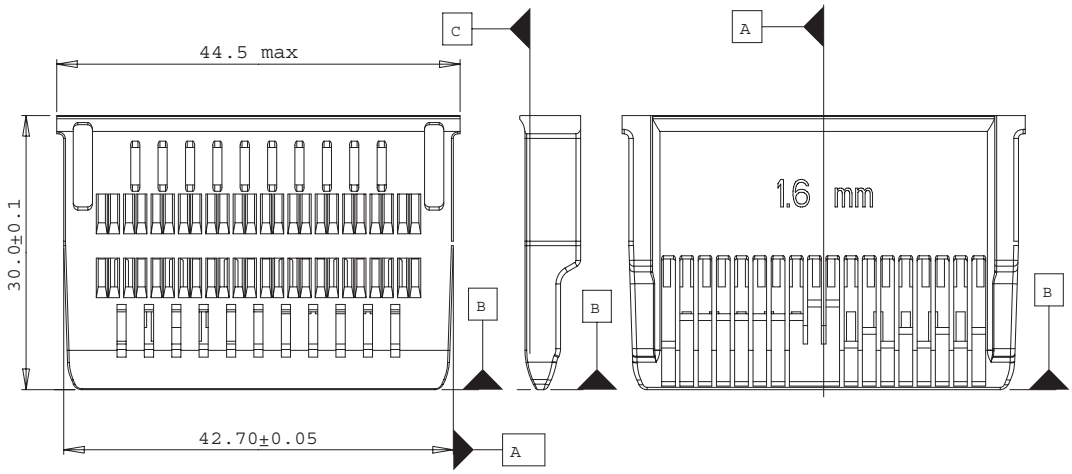


Figure 132 Paddle guard, board primary, side and secondary views

C10-6: The paddle guard contact window design shown in Figure 133 **shall** be used for the high-speed contacts on the primary side of InfiniBand boards.

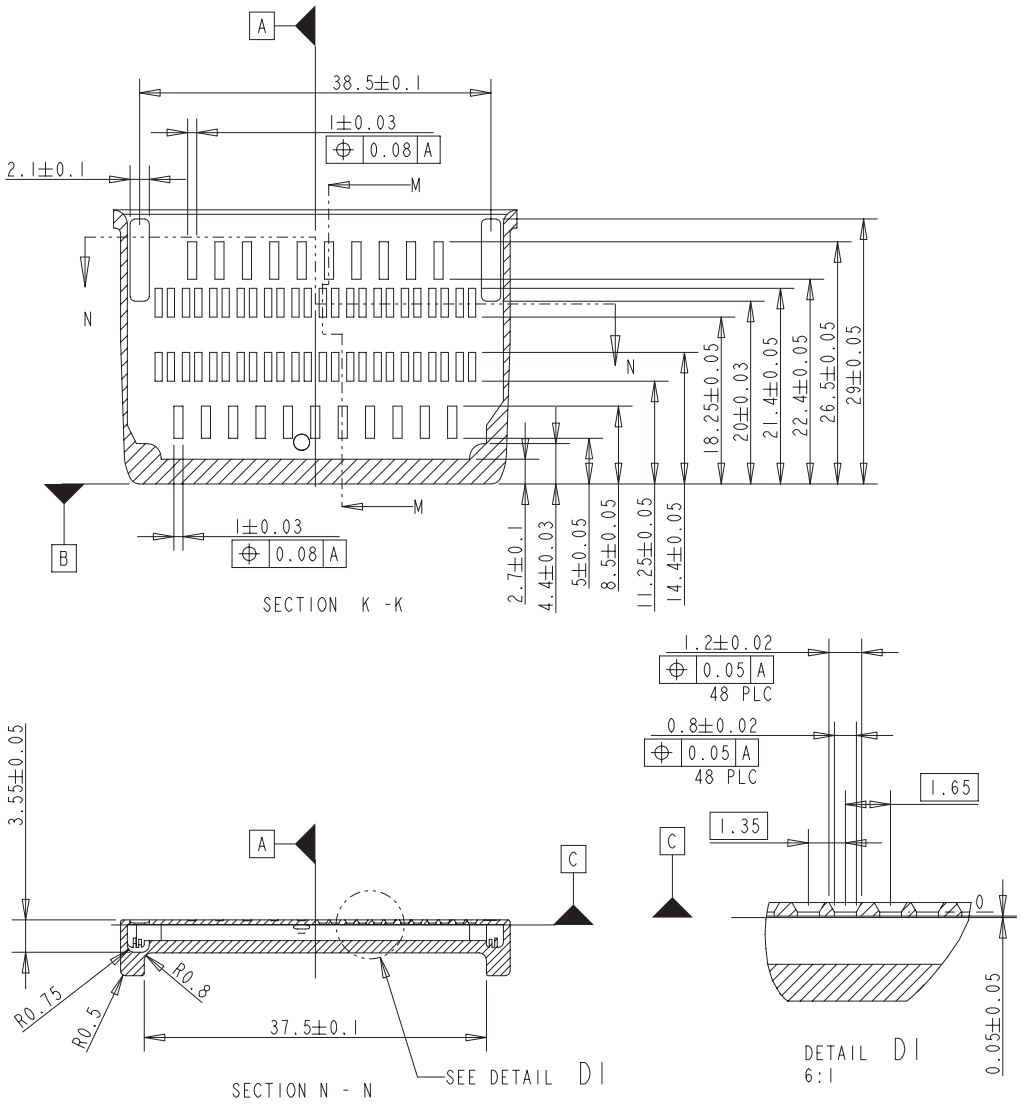


Figure 133 Paddle guard contact window detail, board primary side

C10-7: The paddle guard contact window design shown in Figure 134 **shall** be for low-speed/power contacts on the secondary side of InfiniBand boards.

This pattern provides for four levels of low-speed signal pin make/break timing. The use of these levels is defined in Chapter 12: Power / Hot Plug. Pin sequencing is controlled by the design of the openings in the paddle guard over the board contacts.

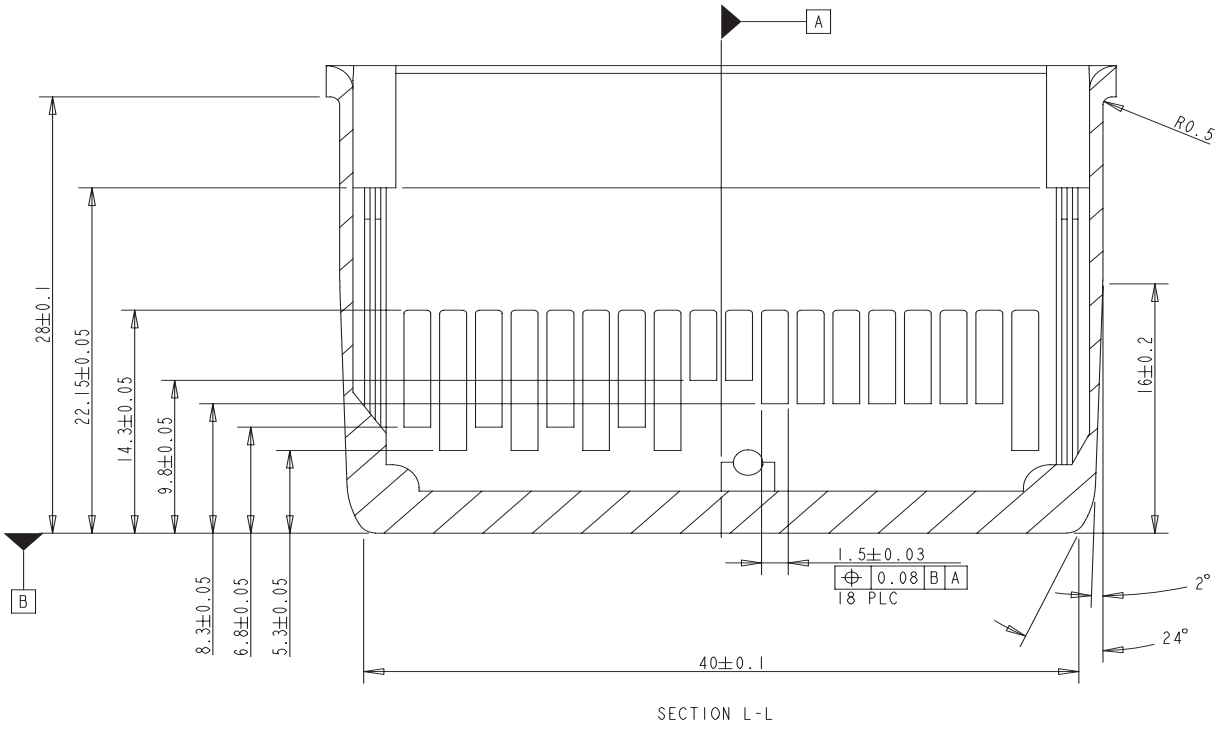


Figure 134 Paddle guard contact window detail, board secondary side

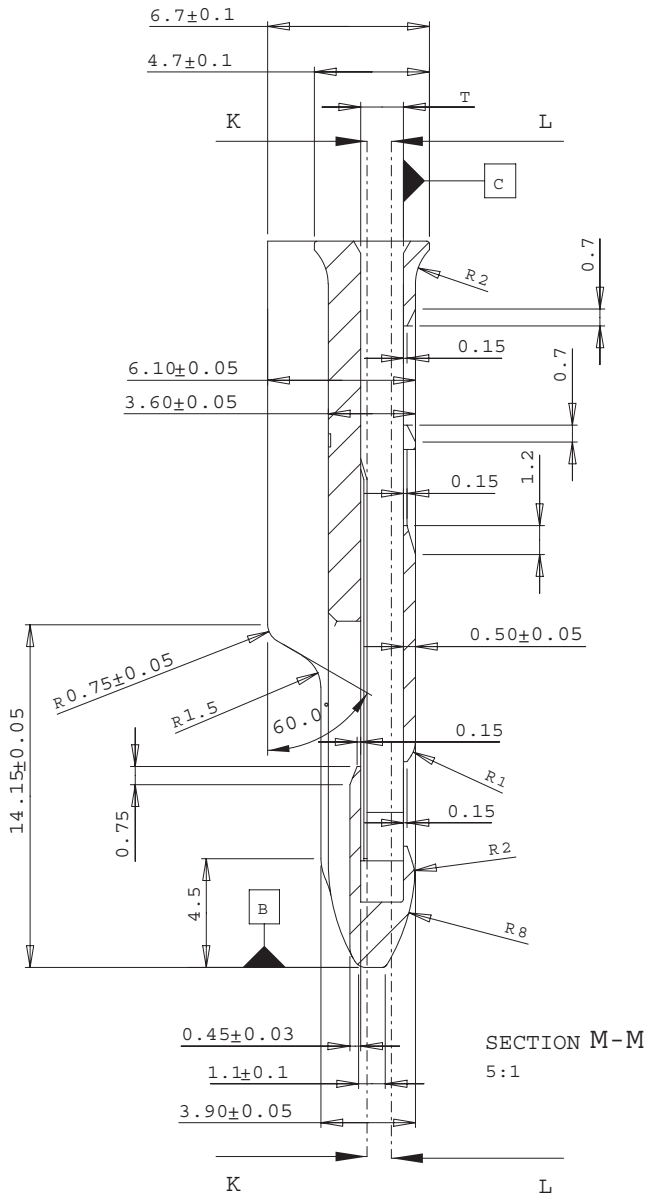


Figure 135 Paddle guard, cross section view

The values for dimension T in Figure 135 are listed in Table 98.

Table 98 Paddle guard slot width, T

Board thickness, mm	Slot width T, min.	Units
1.6 +/- 10%	1.76	mm
2.4 +/- 10%	2.64	mm

10.3.3 BACKPLANE CONNECTOR SIGNAL ASSIGNMENTS

The InfiniBand signals are defined in Chapter 4: Port Signal Definitions.

C10-8: InfiniBand boards and backplanes **shall** use the pin assignments as shown in Table 99 for high-speed signals, and those in Table 102 for low-speed and power signals.

o10-8.1.1: Boards and backplanes supporting multiple 4x ports shall use the high-speed signal pin assignments shown in Table 100.

o10-8.1.2: Boards and backplanes supporting multiple 1x ports shall use the high-speed signal pin assignments shown inTable 101.

IBTA

Table 99 Backplane connector board and backplane contact assignments for single port - primary side

Interface				Row a		Row b	
				Contact	Signal Name	Contact	Signal Name
IB Signaling Group - High Speed Differential Inputs and Outputs							
12x	8x	4x	1x				
T/R	T/R	T/R	T/R	ax01	IBbxIn(0)	bx01	IBbxOn(0)
				ay01	IBbxIp(0)	by01	IBbxOp(0)
				ax02	IBbxIn(1)	bx02	IBbxOn(1)
				ay02	IBbxIp(1)	by02	IBbxOp(1)
				ax03	IBbxIn(2)	bx03	IBbxOn(2)
				ay03	IBbxIp(2)	by03	IBbxOp(2)
				ax04	IBbxIn(3)	bx04	IBbxOn(3)
				ay04	IBbxIp(3)	by04	IBbxOp(3)
				ax05	IBbxIn(4)	bx05	IBbxOn(4)
				ay05	IBbxIp(4)	by05	IBbxOp(4)
				ax06	IBbxIn(5)	bx06	IBbxOn(5)
				ay06	IBbxIp(5)	by06	IBbxOp(5)
				ax07	IBbxIn(6)	bx07	IBbxOn(6)
				ay07	IBbxIp(6)	by07	IBbxOp(6)
				ax08	IBbxIn(7)	bx08	IBbxOn(7)
				ay08	IBbxIp(7)	by08	IBbxOp(7)
				ax09	IBbxIn(8)	bx09	IBbxOn(8)
				ay09	IBbxIp(8)	by09	IBbxOp(8)
				ax10	IBbxIn(9)	bx10	IBbxOn(9)
				ay10	IBbxIp(9)	by10	IBbxOp(9)
				ax11	IBbxIn(10)	bx11	IBbxOn(10)
				ay11	IBbxIp(10)	by11	IBbxOp(10)
				ax12	IBbxIn(11)	bx12	IBbxOn(11)
				ay12	IBbxIp(11)	by12	IBbxOp(11)
				sa01 - sa11 & sb02 - sb11	IB_Sh_Ret - high speed shield; multiple redundant contacts		

Table 100 Backplane connector board and backplane contact assignments for three 4x ports - primary side

Interface	Row a		Row b	
	Contact	Signal Name	Contact	Signal Name
IB Signaling Group - High Speed Differential Inputs and Outputs				
T/R, port 1	ax01	IBbx.1In(0)	bx01	IBbx.1On(0)
	ay01	IBbx.1Ip(0)	by01	IBbx.1Op(0)
	ax02	IBbx.1In(1)	bx02	IBbx.1On(1)
	ay02	IBbx.1Ip(1)	by02	IBbx.1Op(1)
	ax03	IBbx.1In(2)	bx03	IBbx.1On(2)
	ay03	IBbx.1Ip(2)	by03	IBbx.1Op(2)
	ax04	IBbx.1In(3)	bx04	IBbx.1On(3)
	ay04	IBbx.1Ip(3)	by04	IBbx.1Op(3)
T/R, port 3	ax05	IBbx.3In(0)	bx05	IBbx.3On(0)
	ay05	IBbx.3Ip(0)	by05	IBbx.3Op(0)
	ax06	IBbx.3In(1)	bx06	IBbx.3On(1)
	ay06	IBbx.3Ip(1)	by06	IBbx.3Op(1)
	ax07	IBbx.3In(2)	bx07	IBbx.3On(2)
	ay07	IBbx.3Ip(2)	by07	IBbx.3Op(2)
	ax08	IBbx.3In(3)	bx08	IBbx.3On(3)
	ay08	IBbx.3Ip(3)	by08	IBbx.3Op(3)
T/R, port 2	ax09	IBbx.2In(0)	bx09	IBbx.2On(0)
	ay09	IBbx.2Ip(0)	by09	IBbx.2Op(0)
	ax10	IBbx.2In(1)	bx10	IBbx.2On(1)
	ay10	IBbx.2Ip(1)	by10	IBbx.2Op(1)
	ax11	IBbx.2In(2)	bx11	IBbx.2On(2)
	ay11	IBbx.2Ip(2)	by11	IBbx.2Op(2)
	ax12	IBbx.2In(3)	bx12	IBbx.2On(3)
	ay12	IBbx.2Ip(3)	by12	IBbx.2Op(3)
	sa01 - sa11 & sb02 - sb11	IB_Sh_Ret - high speed shield; multiple redundant contacts		

Table 101 Backplane connector board and backplane contact assignments for three 1x ports - primary side

Interface	Row a		Row b	
	Contact	Signal Name	Contact	Signal Name
IB Signaling Group - High Speed Differential Inputs and Outputs				
T/R, port 1	ax01	IBbx.1In(0)	bx01	IBbx.1On(0)
	ay01	IBbx.1Ip(0)	by01	IBbx.1Op(0)
	ax02		bx02	
	ay02		by02	
	ax03		bx03	
	ay03		by03	
	ax04		bx04	
	ay04		by04	
T/R, port 3	ax05	IBbx.3In(0)	bx05	IBbx.3On(0)
	ay05	IBbx.3Ip(0)	by05	IBbx.3Op(0)
	ax06		bx06	
	ay06		by06	
	ax07		bx07	
	ay07		by07	
	ax08		bx08	
	ay08		by08	
T/R, port 2	ax09	IBbx.2In(0)	bx09	IBbx.2On(0)
	ay09	IBbx.2Ip(0)	by09	IBbx.2Op(0)
	ax10		bx10	
	ay10		by10	
	ax11		bx11	
	ay11		by11	
	ax12		bx12	
	ay12		by12	
	sa01 - sa11 & sb02 - sb11	IB_Sh_Ret - high speed shield; multiple redundant contacts		

The physical locations of the signals on the backplane are defined in Section 10.4.

Table 102 Backplane connector board and backplane contact assignments - secondary side

Contact	Signal Name
z01	VA_Ret
z02	VA_In
z03	IMxDat
z04	IMxClk
z05	IMxPReq_L
z06	IMxInt_L
z07	VBxPFW_L
z08	VBxCap
z09	IMxPRst
z10	VBxEn_L
z11	VB_Ret(0)
z12	VB_In(0)
z13	VB_Ret(1)
z14	VB_In(1)
z15	VB_Ret(2)
z16	VB_In(2)
z17	VB_Ret(3)
z18	VB_In(3)

C10-9: Signals **shall** utilize physical contact locations on the paddle and paddle guard as shown in Figure 136.

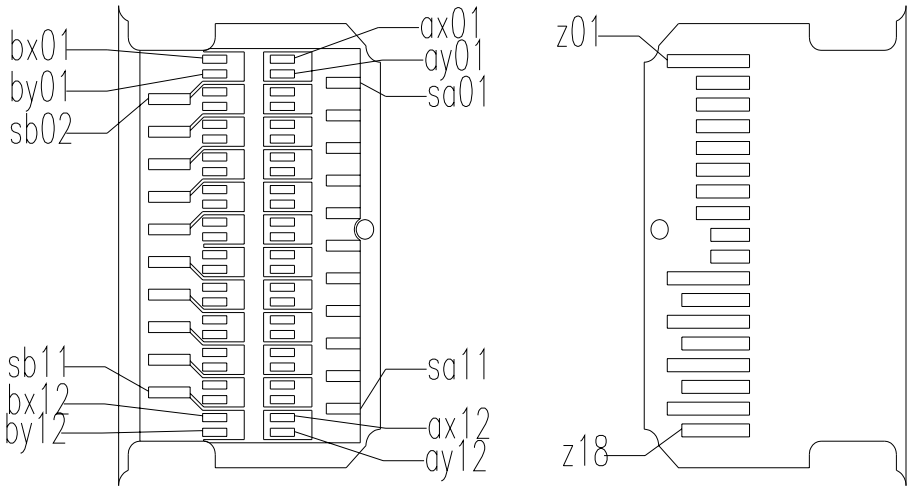


Figure 136 Board contact assignment, primary and secondary sides

10.3.4 BACKPLANE WIRING

C10-9.1.1: Backplane wiring shall be designed with 100+/- 10% Ohms differential impedance.

10.3.5 BOARD WIRING

C10-9.1.2: Board wiring shall be designed with 100+/- 10% Ohms differential impedance.

10.4 CONNECTOR TO BACKPLANE PHYSICAL REQUIREMENTS (MOUNTING SPECIFIC)

10.4.1 BACKPLANE INTERFACE – INITIAL COMPRESSION / PRESS-FIT MOUNTING

The initial backplane connector is based on compression mounting for the high-speed differential contacts and conventional press-fit mounting for the low-speed and power contacts. Press-fit posts are also used as additional mounting points on the primary board side.

While this mounting technique is initially believed to offer the best combination of high speed signal integrity and manufacturability, this standard acknowledges that other mounting techniques may evolve over time and these designs will require additional specification requirements. Therefore, sections are reserved and will be added as required in future revisions of this standard.

C10-10: Backplane connectors used for connection to InfiniBand boards **shall** be intermateable with the board paddle and paddle guard described in Sections 10.3.1 and 10.3.2.

Figure 137 shows a drawing one example of a backplane connector housing of this type.

C10-11: The external dimensions of backplane connectors used for connection to InfiniBand boards **shall** not exceed the outline dimensions shown in Figure 137.

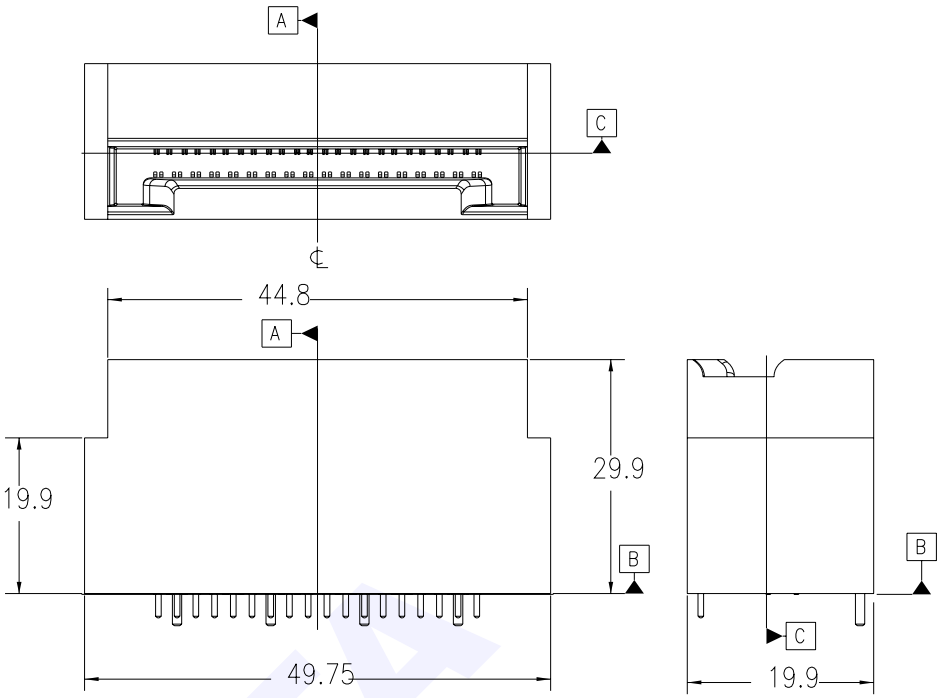


Figure 137 Compression/press-fit backplane connector housing

Figure 138 indicates the resting location of the contacts on the board with respect to the connector housing. The high speed contacts are on the left, and the low speed/power on the right.

IBTA

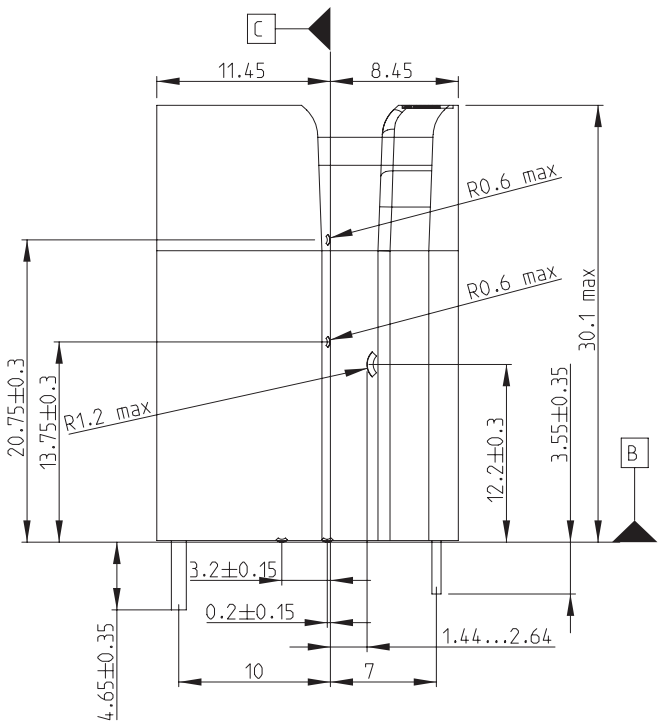


Figure 138 Contact resting locations on board

10.4.1.1 BACKPLANE INTERFACE REQUIREMENTS

This section defines the backplane interface for high-speed compression / low-speed press-fit mounted backplane connectors.

The contact footprint shown in Figure 139 should be used for a 12x implementation using a compression/press-fit backplane connector.

The signal contact assignment shown in Figure 140 should be used on 12x backplanes 12x using a compression/press-fit backplane connector.

Note

Should the designer choose to do so, the 1x/4x connector footprint may be implemented omitting signal contacts ax5-ax12, bx5-bx12, and s5-12, as shown in Figure 141. However, all remaining contacts must be included as shown, and the 12x contact pattern shown in Figure 129 and Figure 130 must still be included on the module board.

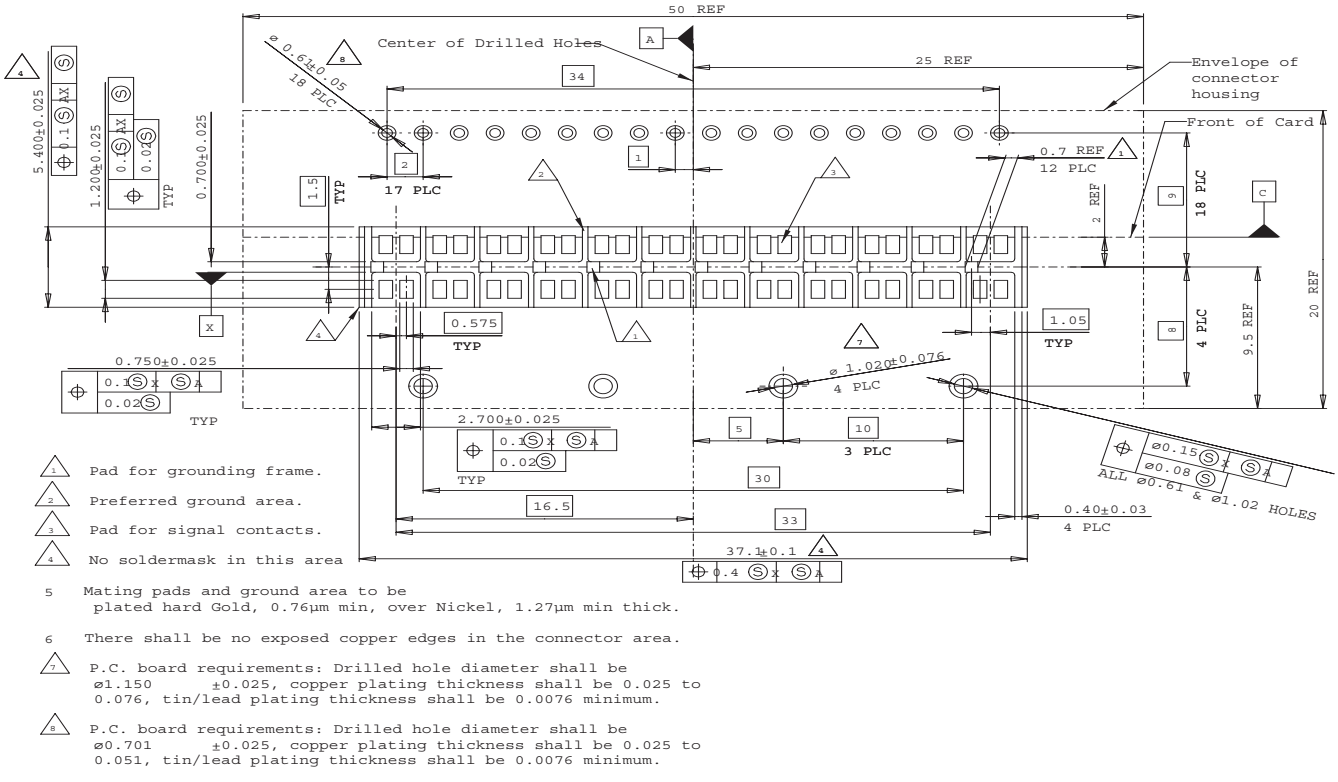


Figure 139 Backplane connector footprint for 12x port

Recommended backplane hole sizes for compression/press-fit connectors are listed in Table 103.

Table 103 Recommended hole dimensions for compression/press-fit backplane connectors

Hole type	Drilled hole size	Finished hole size	Units
Low speed/power	0.701±0.025	0.61±0.05	mm
Retention pins (on high speed side of connector)	1.150±0.025	1.02±0.076	mm

Notes

1. In the event that the press-fit holes are solder plated, the finished hole dimension is the diameter after solder plating.

2. Nickel plating **shall** not be used in press-fit contact holes.

The requirements for contacts used on InfiniBand backplanes are defined in Section 10.4.1.2.

IBTA

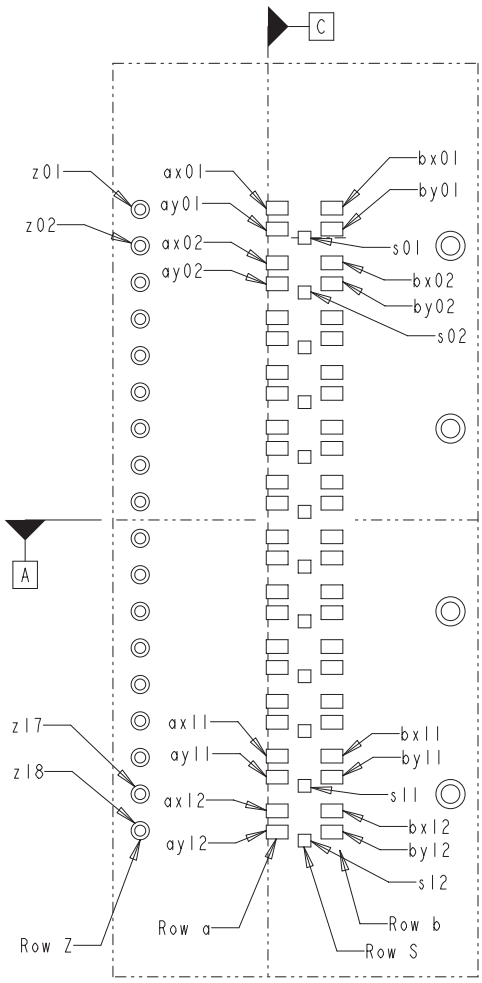


Figure 140 Backplane signal contact assignment, 12x port

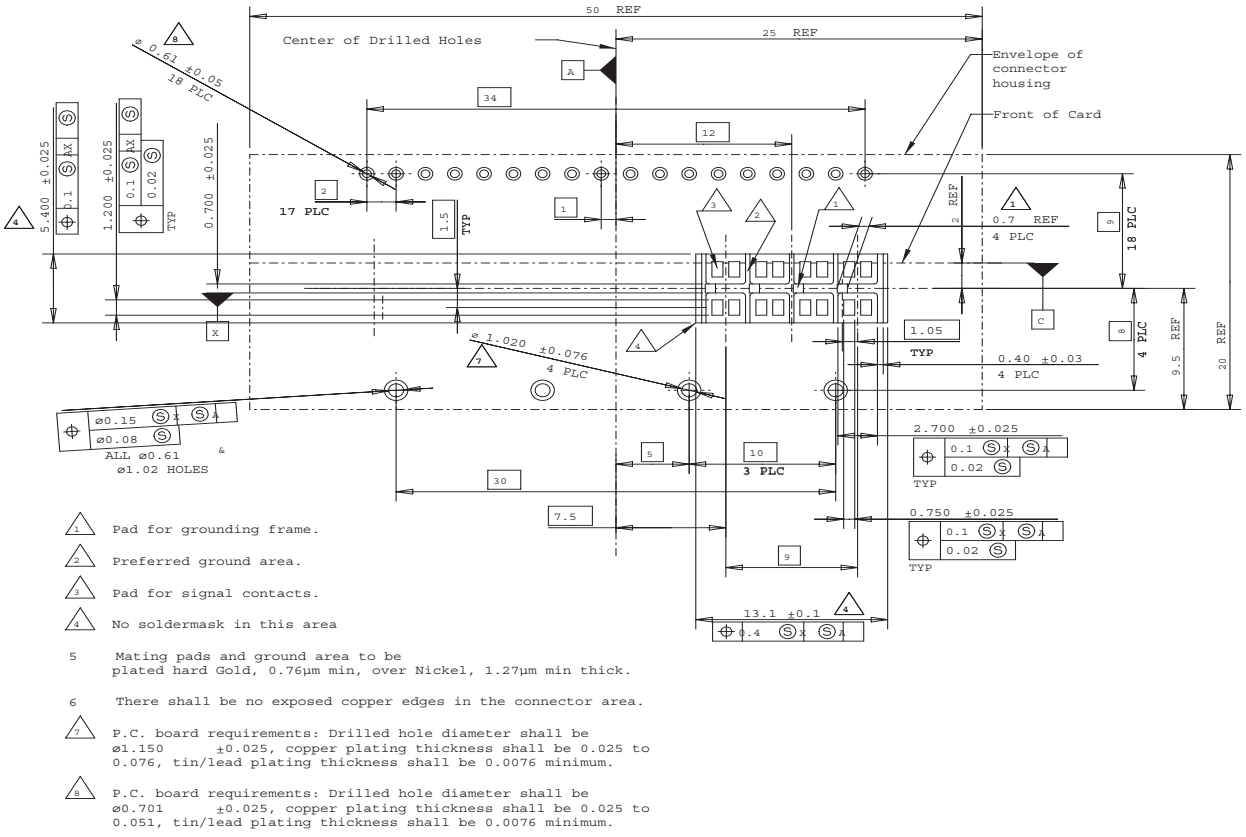


Figure 141 Backplane connector footprint for 1x/4x port

10.4.1.2 BACKPLANE REQUIREMENTS

C10-12: The Backplane **shall** meet the specifications as defined in Table 104.

It is also recommended that a compression / press-fit style backplane connector meet the requirements in Table 105.

Table 104 Backplane requirements - compression / press-fit

Symbol	Parameter	Minimum	Maximum	Units	Comment
t_{bp}	Backplane thickness	1.44	not limited by connector design	mm	1.6 mm nominal
W_{bp}	Backplane coplanarity/ warpage		0.2	mm / 100 mm	
t_{pm}	Contact pad finish, high speed contacts	0.76 min. Au over 1.27 min. Ni		µm	Electroplated. Nickel plating **shall** not be used in press-fit contact holes
	Contact surface finish	TBD	N/A		

Recommendation

To insure reliable connection, backplane contacts should be free of any residues from the board manufacturing processes (such as solder flux, contact masking tape adhesive, etc.) as well as subsequent handling.

Table 105 Recommended connector to backplane mechanical parameters

Symbol	Parameter	Applicable contacts	Minimum	Maximum	Units	Conditions/Comment
F_{ncp}	Contact normal force	high speed	100		cN	per conductor
S_{hcp}	Contact Hertz stress	all	170		kpsi	per contact beam
D_{wp}	Contact Wipe	high speed	0.13		mm	backplane interface only

10.4.2 BACKPLANE INTERFACE – SURFACE SOLDER MOUNTING

The definition of this interface may be present in future versions of this specification.

10.4.3 BACKPLANE INTERFACE – PRESS-FIT MOUNTING

The definition of this interface may be present in future versions of this specification.

10.5 MECHANICAL PERFORMANCE REQUIREMENTS

C10-13: Connectors to be used for connection to InfiniBand boards and contacts on those boards **shall** meet or exceed the physical and mechanical performance requirements listed in Table 106.

Table 106 Backplane connector to board mechanical performance requirements

Symbol	Parameter	Applicable Contacts	Minimum	Maximum	Units	Conditions/Comment
N	Durability	All	250		mating cycles	With paddle guard installed; no more than 1% of contacts with exposed base metal. The specified low level contact resistance **shall** not be exceeded.
F_i	Insertion force	1x/4x		30	N	
		12x		75		
F_w	Withdrawal force	1x/4x		30	N	
		12x		75		
F_{rc}	Housing contact retention force	All	5		N	Backplane interface only
d_{1l}	Distance between contact mating on board insertion/withdrawal	low speed	4.5		mm	First to last low speed contacts
t_{pm}	Contact finish in contact area option 1		0.76 Au over 1.27 Ni		µm	
t_{pm}	Contact finish in contact area option 2		0.51 PdNi with Au flash over 1.27 Ni		µm	Min. 75% Pd in PdNi alloy
F_{rpg}	Paddle guard retention force		100		N	to board paddle

It is further recommended that the backplane connector contact interface to the board meet the mechanical parameters specified in Table 107.

Note

Inserting a paddle guard alone into a backplane connector without a board may cause the paddle guard to be irretrievably retained in the connector, and may cause damage to the connector.

IBTA

Table 107 Recommended connector to board mechanical parameters

Symbol	Parameter	Applicable contacts	Minimum	Maximum	Units	Conditions/Comment
F_{ncb}	Contact Normal Force	low speed	100		cN	per conductor
		high speed	100		cN	per conductor
S_{hcb}	Contact Hertz stress	all	170		kpsi	per contact beam
D_{wb}	Contact wipe	low speed	1.5		mm	board interface only
		high speed	1.5		mm	board interface only

10.6 ELECTRICAL PERFORMANCE REQUIREMENTS

C10-14: Backplane connectors to be used for connection to InfiniBand boards **shall** meet or exceed the electrical performance requirements listed in Table 108.

Table 108 Connector electrical performance requirements

Symbol	Parameter	Applicable contacts	Minimum	Maximum	Units	Conditions/Comment
LLCR	Low level contact resistance - initial	low speed/ power		15	mΩ	through testing per EIA 364-23
		high speed		30	mΩ	
ΔLLCR	Low level contact resistance - change	low speed/ power		10	mΩ	through testing per EIA 364-23, as a result of any test group step
		high speed		20	mΩ	
I_{max}	Current rating	low speed/ power	2.5		A	per EIA-364-70 or IEC 512-5-1 Test 5a, at 30° C. temperature rise above ambient with all contacts energized
		high speed	0.5		A	per conductor pair, per EIA-364-70 or IEC 512-5-1 Test 5a, at 30° C. temperature rise above ambient
L_{ls}	Inductance	low speed/ power		30	nH	measured as loop inductance of two adjacent pins per EIA 364-69 at 1 MHz

Table 108 Connector electrical performance requirements

Z_{dco} (peak)	Differential Impedance (peak)	high speed	90	110	Ω	Average value measured over the propagation delay of the connector at 100 ps rise time (at the connector), per draft EIA 364-108; includes connector, board and backplane pads, and vias
Z_{dco} (nom)	Differential Impedance (average)	high speed	95	105		
L_{co}	Insertion loss	high speed		0.75	dB	frequencies up to 1.25 GHz, per EIA 364-101
S_{cop}	Within pair skew	high speed		5	ps	by design; measurement not required.
J_{co}	Jitter	high speed		10	ps	by design; measurement not required.
$NEXT_c$	Near end crosstalk	high speed		3	%	measured differentially with all adjacent neighbor pairs driven at 100 ps rise time, per EIA 364-90

10.7 ENVIRONMENTAL PERFORMANCE REQUIREMENTS

C10-15: Connectors to be used for connection to InfiniBand boards **shall** meet or exceed the environmental performance requirements of EIA-364.1000.01, including exposure to benign dust and Mixed Flowing Gas consistent with the required product life as defined in Section 10.1.

Unless otherwise noted, successful completion of a given test is indicated by an acceptable Low Level Contact Resistance measurement upon completion of the test, as defined above.

CHAPTER 11: LOW SPEED ELECTRICAL SIGNALING

11.1 INTRODUCTION

This chapter describes the electrical signaling used on the low speed contacts of the InfiniBand backplane connector. Specifically, these are those signals within the Bulk Power, Auxiliary Power and System Management Groups.

11.2 GENERAL REQUIREMENTS FOR ALL GROUPS

11.2.1 ESD

C11-1: InfiniBand signal and power contacts **shall** withstand ESD voltages as defined in Section 6.3.1, "ESD," on page 178.

11.3 BULK POWER GROUP

This section provides the electrical specifications for the Bulk Power Group Pins and contains the pertinent parameters needed by the board power design. The parameters reference the topologies in Section 12.1.3 on page 457.

11.3.1 VB_IN

The bulk power input is provided on the **VB_In** pins and provides for either 25W or 50W of power capability.

C11-2: The backplane **shall** supply 25W of power capability on the **VB_In** pins.

o11-1: The backplane **may** optionally provide an additional 25W on the **VB_In** pins for a total of 50W capability.

C11-3: If the backplane only provides 25W of power to a port, it **shall not** assert the **VBxCap** signal for that port.

o11-2: If a module has multiple power groups available, the **VB_In** pins **shall not** be connected between power groups (ports). The outputs of the local DC-DC converter(s) **may** be connected together.

The electrical parameters for these pins are shown in Table 109 on page 436.

11.3.2 VB_RET

The **VB_Ret** connections are specified on the connector to support the isolated configuration. These Power returns may be connected locally to

logic ground on the IB Module to support Class I, or left isolated on the IB Module to be connected at a common point in the system to support Class II. The IB Modules power class must be reported in the modules *ModuleInfo*. The detailed description of the *ModuleInfo* information is located in Section 13.3.2.10.

11.3.3 BULK POWER PARAMETERS

C11-4: The Bulk Power pins, **VB_In** and **VB_Ret**, **shall** conform to the values defined in Table 109.

The nominal bulk voltage is 12V.

Note: All capacitance values specified in Table 109 include the tolerances of the components implemented.

Table 109 Bulk Power Parameters

Symbol	Parameter	Minimum	Maximum	Units	Comment
Board					
V_{Bulk}	Operational Voltage - Static	10	14	V	Measured at the bottom of the backplane connector.
	Operational Voltage - Transient	--	16[a]	V	Maximum voltage that can be dealt with for 1 ms and still operate. Measured at the bottom of the backplane connector.
I_{BAvg}	Average Current	--	2.5	A	The average current shall not exceed the values shown within any 1 sec. sliding window under all conditions.
$I_{BRefRip}$	Reflected Ripple Current (Peak-Peak)	--	100	mA	The ripple that the board may produce into a resistive load equivalent to the module's maximum load. This is measured from 5Hz to 20MHz.
I_{BMax}	Peak Current	--	2.8	A	The current that the board **shall** provide limiting for applicable during power-on sequences and during normal operation
$I_{BLeakage}$	Leakage Current	--	20	mA	The maximum current that a module may draw when **VBxEN_L** is de-asserted. (Power is disabled)

Table 109 Bulk Power Parameters

Symbol	Parameter	Minimum	Maximum	Units	Comment
$didt_{BMax}$	Current Transient	--	40	A/ms	The current transient that the board **shall** provide limiting for applicable during power-on sequences and during normal operation. $\mathbf{I_{BMax}}$ must not be violated.
t_{BCT}	Current Transient Time	--	28	ms	I_{BMax} / $didt_{BMax}$
C_{BIn}	Initial Hot Plug Capacitance	--	5000	pf	Capacitance at hot plug connector mating.
C_{BBulk}	Input Capacitance	--	500	µf	Refer to Figure 152 on page 458 and Figure 153 on page 459.
t_{BOn}	Turn-on Time	--	500	ms	This is measured from the AND of **VBxPFW_L** being asserted inactive AND **VBxEn_L** being asserted active AND Local Power Enable being asserted active to when the converter is enabled and within regulation.
t_{BOff}	Turn-off Time	--	1	ms	This is measured from the deassertion of **VBxEn_L** deassertion time is the hot removal case.

a. These are the tolerance requirements, across a 20 MHz bandwidth, at the backplane side of the InfiniBand backplane connector. **VB_In** must return to within the static voltage specification (V_{Bulk}) within 1 ms after a transient event.

11.3.3.1 MEASUREMENTS

C11-5: Bulk voltages designated in Table 109 **shall** be measured at the backplane side of the InfiniBand backplane connector.

11.3.4 VBXEN_L

The **VBxEn_L** pin provides for strong sensing that the IB Module is fully seated prior to power converter enabling. This signal is assigned on the shortest pin level of the connector such that it is the "last to mate" / "first to break".

The signal is terminated as shown in Figure 142 on page 438. The backplane provides for the overriding "Asserted" level by drawing current through the **VBxEn_L** pin. The maximum current required to drive the **VBxEn_L** signal asserted is $\mathbf{I_{VBxEN\_ol}}$. The voltage level required to assert the **VBxEn_L** signal is implementation dependent. The module pro-

vides for the default "Not Asserted" level. The "Not Asserted" voltage level is implementation dependant, the circuit shown in Figure 142, has a "virtual"[2] pull-up R_{PU} to **VB_In** as an indication to use the bulk voltage rail to drive the level. The maximum voltage level on **VBxEn_L** that may applied to the backplane is V_{VBxEN_Max}. Please see Table 110 on page 443 for the values of V_{VBxEN_Max} and I_{VBxEN_ol}.

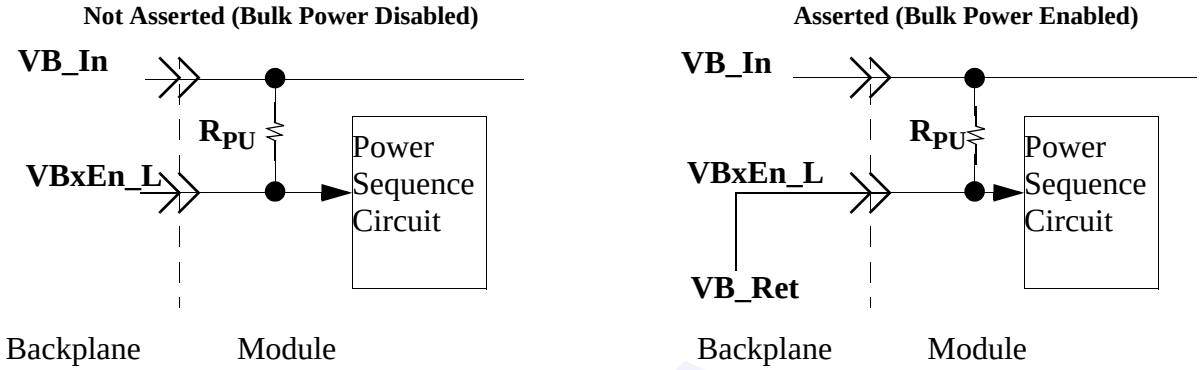


Figure 142 VBxEn_L Termination

11.3.4.1 GENERAL RULES

C11-6: The **VBxEn_L** signal **shall** conform to the DC parameters defined in Table 110 on page 443.

11.3.4.2 MODULE RULES

C11-7: The module **shall** provide the **VBxEn_L** signal for each power port.

Also see Section 12.4, "Module Power Rules," on page 467 for additional rules pertaining to this signal on the module.

11.3.4.3 CHASSIS RULES

o11-3: A "Managed Chassis" **may** optionally control the **VBxEn_L** signal to selectively enable or disable the power sequence circuit of the module.

Also see Section 12.5, "Chassis Power Rules," on page 470 for additional rules pertaining to this signal on the chassis.

2. "Virtual", in this context, means it is to behave from the outside as a pull up to something like 5 Volts but it is to get its source of power from the **VB_In** for a Class II module. A Class I module may use either **VB_In** or **VA_In**.

11.3.5 VBxCAP

The Bulk Capability (**VBxCap**) signal is bidirectional and indicates the power capability provided by the chassis and the power required by the IB Module.

For modules requiring between 25 and 50W, a backplane must be capable of delivering this power on the **VB_In** pins. IB Modules (single wide high or double wide standard) monitor the level of the **VBxCap** pin to determine if the backplane in which the module is plugged is capable of delivering this power. A low level (not asserted) indicates that the port is only capable of 25W; a high level (asserted) indicates that 50W is available. The "level" is referenced to **VA_Ret.** Refer to Figure 143 on page 439 and Figure 144 on page 440 for termination structures.

A 50W IB Module supplies a pull-down R_{PD} on the board to provide an initial 25W indication to the IB Management Link Controller. The signal can be 1) driven high by a 50W backplane with a pull-up R_{PU} to indicate a 50W capability or 2) be grounded by a 25W backplane. In the second situation, the backplane does not provide the power required by the IB module. The module must therefore either a) not power on or b) recognize that it is plugged into a backplane that will only deliver 25W and take appropriate actions to limit its power accordingly.

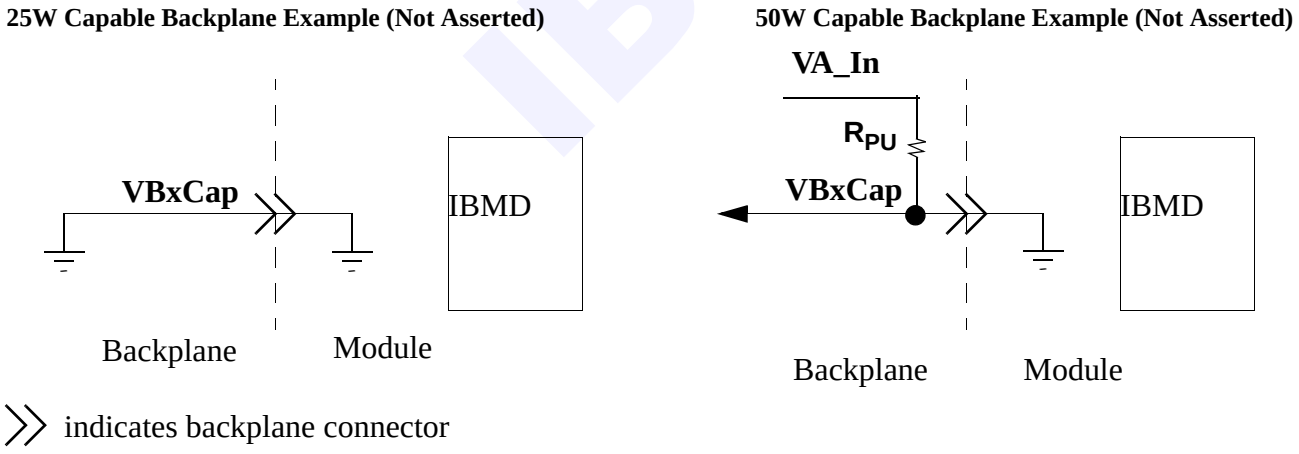


Figure 143 VBxCap Termination - 25W IB Module

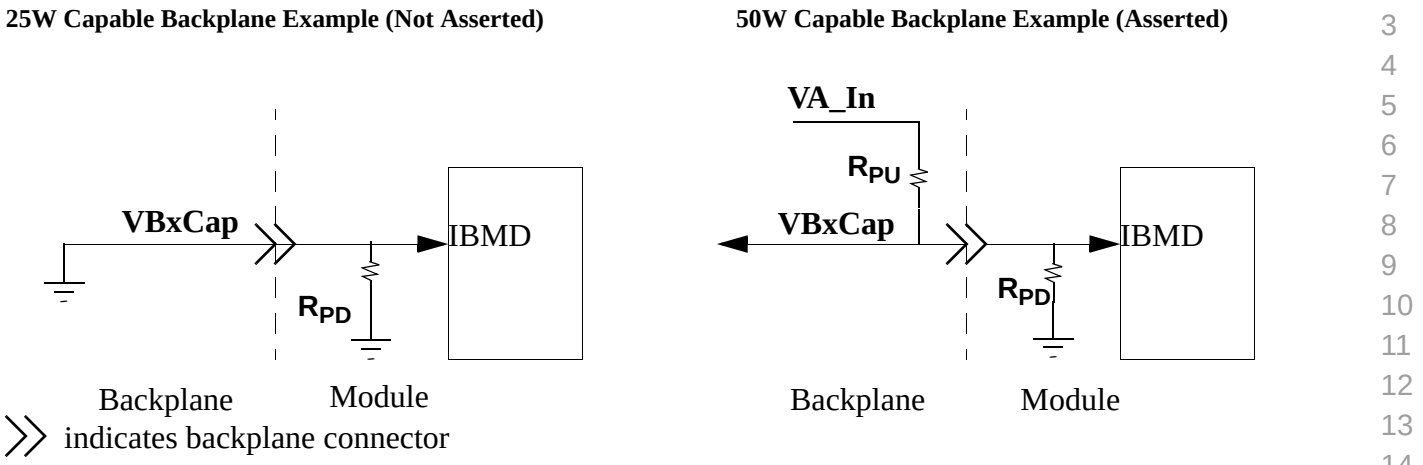


Figure 144 VBxCap Termination - 50W IB Module

11.3.5.1 GENERAL RULES

C11-8: The **VBxCap** signal **shall** conform to the DC parameters defined in Table 110 on page 443.

The parameters V_{VBxCap_oh} and V_{VBxCap_ol} define the required voltage levels for the InfiniBand Management Device (IBMD) and backplane. The parameter $I_{VBxCap\_Leakage}$ defines the maximum leakage current generated by the IBMD or backplane inputs. The parameters R_{PU} and R_{PD} define the allowed values for the pull-up and pull-down devices respectively. The values for these parameters are defined in Table 110 on page 443.

11.3.5.2 MODULE RULES

C11-9: A power port on a module requiring 25W or less **shall** directly connect **VBxCap** to **VA_Ret** to communicate to a 50W capable backplane that the port will consume no more than 25W.

C11-10: A power port on a module requiring between 25W and 50W **shall** provide a pull-down R_{PD} as indicated in Figure 144 on page 440 and of a value as specified in Table 110 on page 443.

Implementation Note

Typically available 10% resisters of 2.2KΩ for R_{PU} and 3.0KΩ for R_{PD} are recommended for use on the **VBxCap** signal.

o11-4: If a power port on a module requires between 25W and 50W, the module **shall** monitor the level of the **VBxCap** pin to determine if the backplane in which the module is plugged is capable of delivering this power.

If the port detects a voltage less than or equal to V_{VBxCap_ol} on **VBxCap**, the module **shall** either a) not draw power from the port or b) recognize that it is plugged into a backplane that will only deliver 25W and take appropriate actions to limit its power draw accordingly.

11.3.5.3 CHASSIS RULES

C11-11: A power port on a chassis that delivers less than 50W **shall** directly connect **VBxCap** to **VA_Ret**.

C11-12: A power port on a chassis that can deliver at least 50W **shall** provide R_{PU} pull-up to **VA_In**.

11.3.6 VBxPFW_L

The Bulk Power Fail Warning (**VBxPFW_L**) signal originates from the system or chassis power supply and serves two functions.

1) When it transitions from deasserted (high) to asserted (low), this alerts the IB Module that bulk power may be about to go out of specified tolerance, and is only guaranteed to remain within tolerance for $t_{VBINVALID}$. When an adapter sees this transition, it may begin whatever house keeping functions it deems necessary to prepare for power loss.
2) When it transitions from asserted to deasserted, this alerts the IB Module card that bulk power has been within specified tolerance for at least $t_{VBVALID}$. Modules should not begin initialization until this signal has been sampled deasserted.

11.3.6.1 GENERAL RULES

C11-13: The **VBxPFW_L** signal **shall** conform to the DC parameters defined in Table 110 on page 443 and the AC parameters Table 111 on page 443.

11.3.6.2 MODULE RULES

o11-5: IB Modules **may** optionally utilize **VBxPFW_L** functionality.

> **Recommendation to System Designer**
>
> Buffering of this signal on a slot basis prevents glitches on **VBxPFW_L** from affecting adjacent slots.

11.3.6.3 CHASSIS RULES

C11-14: The Chassis **shall** provide the defined **VBxPFW_L** function for every port.

C11-15: Once the **VBxPFW_L** signal transitions to asserted, it **shall** remain asserted for a minimum of t_{PW_low} even if the bulk power did not go out of specification.

C11-16: Once the **VBxPFW_L** signal transitions to deasserted, it **shall** remain deasserted for a minimum of t_{PW_high} even if the bulk power immediately indicates a failure is imminent.

11.3.7 VBx* SIGNAL PARAMETERS

This section defines the electrical parameters for the control signals within the Bulk Power group.

If parameters apply to all signals beginning with "**VBx**", they are referred to as "**VBx***"; the "*" indicates a generic placeholder. If parameters apply to specific signals, these are named explicitly.

IBTA

11.3.7.1 VBx* DC SPECIFICATIONS

Table 110 VBx* DC Specifications

Symbol	Parameter	Minimum	Maximum	Units	Comments
V_{VBxEn_Max}	**VBxEn_L** maximum voltage		5.5	V	See Figure 146 on page 446.
I_{VBxEn_ol}	**VBxEn_L** output low	4		mA	See Figure 146 on page 446.
I_{VBxEn_il}	**VBxEn_L** input low		4	mA	See Figure 146 on page 446.
V_{VBxCap_oh}	**VBxCap** output high - module	2.1	3.63	V	See Figure 143 on page 439.
	VBxCap output high - backplane	2.1	5.50	V	See Figure 143 on page 439.
V_{VBxCap_ol}	**VBxCap** output low	-0.5	0.8	V	See Figure 143 on page 439.
R_{PU}[a]	Pull-up Termination	1.870	2.500	KΩ	
R_{PD}[a]	Pull-down Termination	2.625	3.630	KΩ	
$I_{VBxCap\_Leakage}$	**VBxCap** leakage current		+/- 100	µA	See Figure 143 on page 439.
$V_{VBxPFW\_L\_il}$	**VBxPFW_L** Input Low Voltage	-0.5	.8V	V	See Figure 145 on page 445.
$V_{VBxPFW\_L\_ih}$	**VBxPFW_L** Input High Voltage	2.1	3.63	V	See Figure 145 on page 445.

a. The values specified for this parameter do not include all tolerances of implemented components.

11.3.7.2 VBxPFW_L AC SPECIFICATIONS

Table 111 VBxPFW_L AC Timing Specifications

Symbol	Parameter	Minimum	Maximum	Units	Comments
$t_{VBINVALID}$	**VBxPFW_L** Asserted To **VB_In** Invalid	2		ms	See Figure 145 on page 445

Table 111 VBxPFW_L AC Timing Specifications

Symbol	Parameter	Minimum	Maximum	Units	Comments
$t_{VBVALID}$	**VB_In** Valid to **VBxPFW_L** Deasserted	100		ms	See Figure 145 on page 445
t_{PW_low}	Signal Pulse Width Low	1		ms	See Figure 145 on page 445
t_{PW_high}	Signal Pulse Width High	10		µs	See Figure 145 on page 445
t_{Fall}	Fall Time		300	ns	See Figure 145 on page 445
t_{Rise}	Rise Time		1000	ns	See Figure 145 on page 445

IBTA

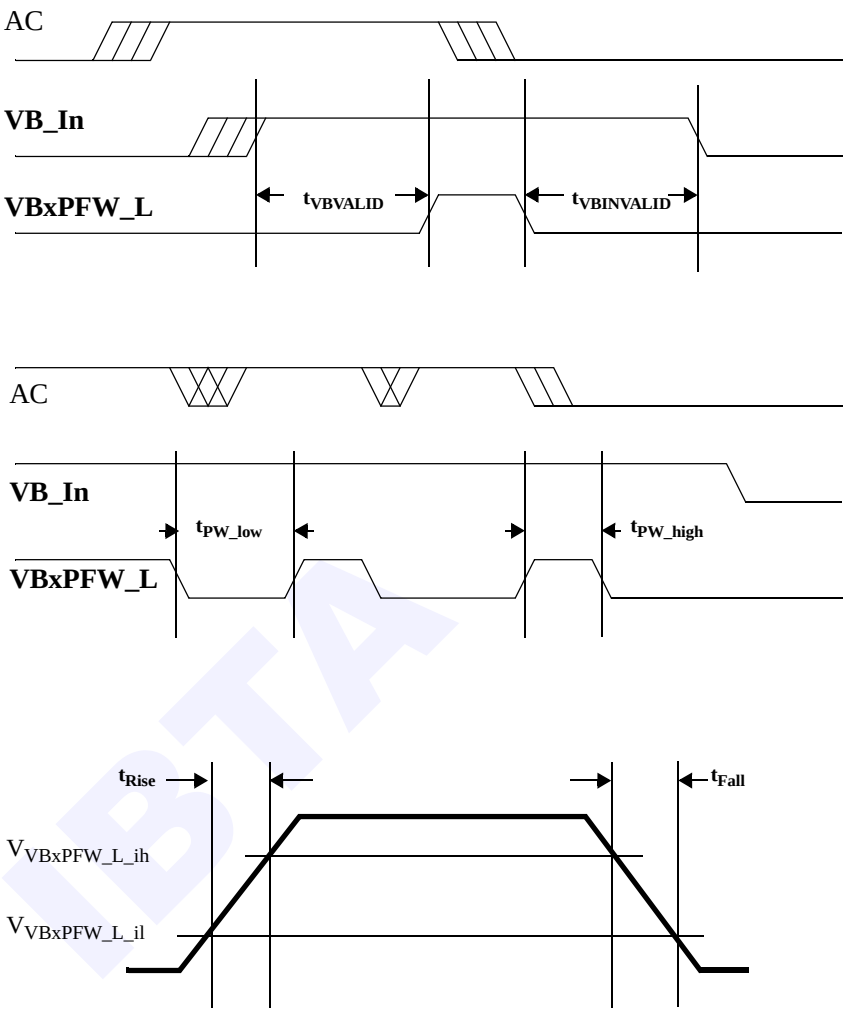


Figure 145 VBxPFW_L Timing Measurements

IBTA

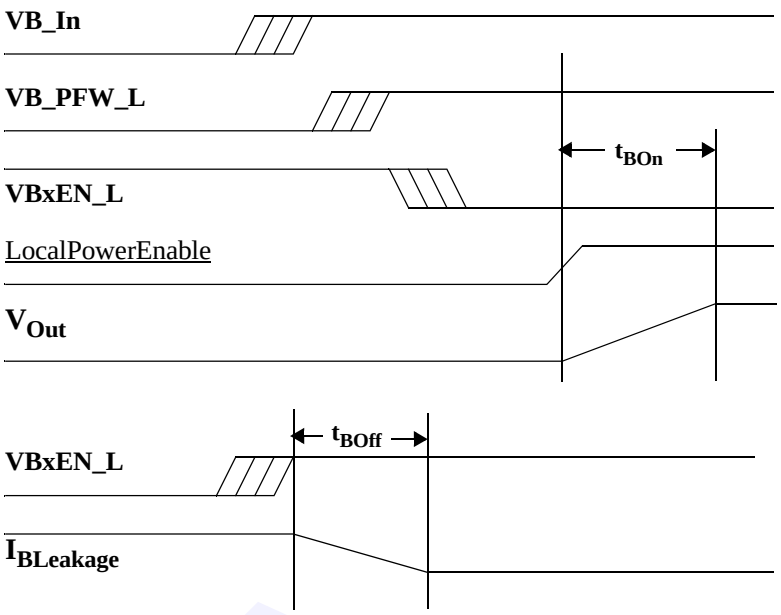


Figure 146 VBxEn_L Timing Measurements

11.4 AUXILIARY POWER GROUP

This section provides the electrical specification for the Bulk Power Group Pins and contains the pertinent parameters needed by the board power design. The parameters reference the topologies in Section 12.1.3 on page 457.

11.4.1 VA_IN

Auxiliary power input is provided on the **VA_In** pin and provides for up to 2.4W based on form factor as defined in Table 116 on page 461.

11.4.2 VA_RET

The **VA_Ret** connection provides the return path for the Auxiliary power functions and is connected to logic ground on the board.

11.4.3 AUXILIARY POWER PARAMETERS

C11-17: The Auxiliary Power pins, **VA_In** and **VA_Ret**, **shall** conform to the values defined in Table 112.

The nominal auxiliary voltage is 5V.

Note: All capacitance values specified in Table 112 include the tolerances of the components implemented.

Table 112 Auxiliary Power Parameters

Symbol	Parameter	Minimum	Maximum	Units	Comment
Board					
V_{Aux}	Operational Voltage - Static	4.60	5.50	V	Measured at the bottom backplane connector.
I_{AAvg}	Average Current	--	.260	A	The average current **shall not** exceed the values shown within any 1 sec. sliding window under all conditions.
I_{AMax}	Peak Current	--	.390	A	The current that the board **shall** provide limiting for applicable during power-on sequences and during normal operation
$didt_{AMax}$	Current Transient	--	0.1	A/ms	The current transient that the board **shall** provide limiting for. This is applicable during power-on sequences and during normal operation
LS					
C_{AIn}	Initial Hot Plug Capacitance	--	1000	pf	Capacitance at hot plug connector mating.
C_{ABulk}	Input Capacitance	--	100	µf	
t_{AOn}	Turn-on Time	--	2	s	This is measured from the presence assertion of the **IMxPRst** pin. See Figure 147 on page 448.

11.4.3.1 MEASUREMENTS

C11-18: Auxiliary voltages designated in Table 112 **shall** be measured at the backplane side of the InfiniBand backplane connector.

IBTA

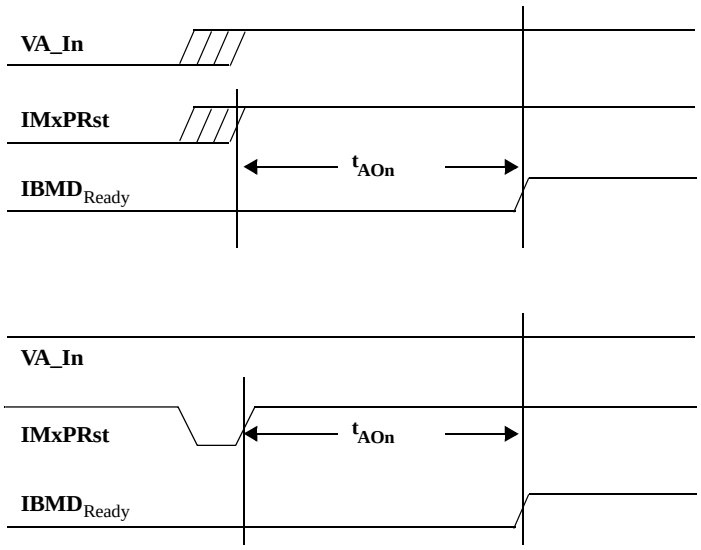


Figure 147 VA_In Timing Measurements

11.5 SYSTEM MANAGEMENT GROUP

The system management group consists of the InfiniBand Management Link IB-ML (**IMxClk** and **IMxDat**), the management interrupt (**IMxInt_L)**, the module presence and reset (**IMxPRst)**), and the power request (**IMxPReq_L)** signals.

11.5.1 IMXCLK, IMXDAT

The **IMxClk** and **IMxDat** signals provide a serial management interface which supports communication between an IB module and a chassis.

11.5.1.1 GENERAL RULES

C11-19: The **IMxDat** and **IMxClk** signals **shall** conform to the DC parameters defined in Table 113 on page 453 and the AC parameters Table 114 on page 455.

C11-20: IB-ML devices on modules that are powered down **shall** provide a mechanism to isolate the device from the management link. The load presented by the isolated devices shall be less than $I_{Isolate}$.

11.5.1.2 MODULE RULES

C11-21: A module **shall not** present a capacitive load on the IB-ML greater than C_{IBML_Module}. The IB Module **shall not** present a current leakage load on the IB-ML greater than I_{IBML_Module}.

C11-22: A module **shall** provide pull-ups to 3.3V on the **IMxClk** and **IMxDat** signals. The pull-up value **shall** be designed for the maximum load on the module and chassis.

Please see Annex A2: IB-ML Design Guidelines on page 767 for design guidelines in selecting an appropriate value.

11.5.1.3 CHASSIS RULES

C11-23: The chassis **shall not** present a capacitive load on the IB-ML greater than $C_{IBML\_Chassis}$. The chassis **shall not** present a current leakage load on the IB-ML greater than $I_{IBML\_Chassis}$.

C11-24: The chassis **shall** provide pull-ups to 3.3V on the **IMxClk** and **IMxDat** signals. The pull-up value **shall** be designed for the maximum load on the module and chassis.

Please see Annex A2: IB-ML Design Guidelines on page 767 for design guidelines in selecting an appropriate value.

11.5.2 IMXINT_L

The Management Interrupt (**IMxInt_L**) signal originates from the IB Module to indicate that an attention condition has arisen to a present CME. See Section 13.3.2.8, "IMxInt_L," on page 515 for an operational description.

C11-25: A Module **shall** provide the **IMxInt_L** signal on all ports with an implemented IB-ML.

11.5.2.1 GENERAL RULES

C11-26: The **IMxInt_L** signal **shall** conform to the DC parameters defined in Table 113 on page 453 and the AC parameters defined in Table 115 on page 456.

11.5.3 IMXPRST

The **IMxPRst** signal is bidirectional and provides the chassis with the ability to both detect the physical presence of an IB Module in a slot and to perform a reset of the IB module, if necessary.

11.5.3.1 PRESENCE DETECT

The Presence function is achieved through the R_{PRup} and R_{PRdn} resistors. The values chosen would have R_{PRdn} being weaker than R_{PRup} so that an inserted module would have **IMxPRst** being pulled to an asserted high level. An empty slot would have R_{PRdn} pulling **IMxPRst** to **VA_Ret** which represents the deasserted level. A chassis would monitor the level (as defined by Vil and Vih in Table 113 IMx* DC Specifications on page 453) to determine module presence.

Implementation Note

Typically available 10% resisters of 2.2KΩ for R_{PRup} and 3.0KΩ for R_{PRdn} are recommended for use on the **IMxPRst** signal.

Upon auxiliary power being available to a module, an amount of time is necessary to allow the functions implemented on auxiliary power to achieve the reset condition. Thus, upon detection of a module in a slot, the chassis needs to wait before attempting operations on IB-ML to that slot (See t_{AOn} in Table 112 on page 447).

11.5.3.2 RESET

The Reset function is provided to allow a chassis to effect a slot reset, including the functions implemented on the Auxiliary power domain, without power cycling the **VA_In** inputs. This reset is to place the module into the state that would have been achieved by a cold power on of both auxiliary and primary power.

Referring to Figure 148 IMxPRst Termination on page 451, the reset indication is achieved by the module monitoring the level of the **IMxPRst** signal for a low level. Under normal operations, the **IMxPRst** is held at a high level by the strong R_{PRup} resistor. The chassis can implement an ac-

IBTA

tive override mechanism to pull the **IMxPRst** signal low to achieve the reset.

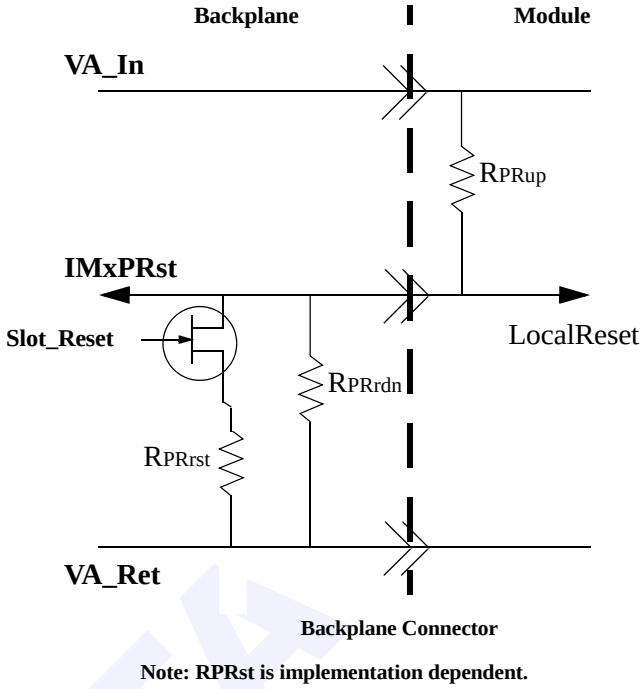


Figure 148 IMxPRst Termination

11.5.3.3 GENERAL RULES

C11-27: The **IMxPRst** signal **shall** conform to the DC parameters defined in Table 113 on page 453 and the AC parameters defined in Table 115 on page 456.

11.5.3.4 MODULE RULES

C11-28: A module **shall** implement the **IMxPRst** signal on all implemented backplane ports.

> **Architectural Note**
>
> This provides a means for the chassis to determine that the module is fully inserted.

C11-29: The module **shall** provide for an initial reset condition upon **VA_In** reaching $\mathbf{V_{Aux}}$(min) or other means.

C11-30: Upon release of the reset condition, the module **shall** clear the IB-ML agent state machines, return the MME facilities their "default" values, and perform any module specific functions that would take place within *ModulePowerInfo.InitTime*. (See Table 156 on page 622).

11.5.3.5 CHASSIS RULES

o11-6: The chassis **may** monitor the level of **IMxPRst** to determine module presence from a backplane port.

o11-7: If the chassis uses IB-ML, the chassis **shall** wait for the time specified by $\mathbf{t_{AOn}}$ (See Table 112 on page 447) before attempting operations on IB-ML to that slot.

o11-8: The chassis **may** perform a reset through an active override mechanism to pull the **IMxPRst** signal low to achieve the reset. If the override mechanism is implemented, the **IMxPRst** signal **shall** be deasserted a minimum of $\mathbf{t_{Reset}}$ as defined in Table 115.)

o11-9: The chassis **may** perform a reset prior to module operation.

11.5.4 IMXPREQ_L

The **IMxPReq_L** pin signals the chassis that the IB Module is requesting **VB_In** to be supplied.

This signal is controlled by the Baseboard Manager and WakeRequestEvent.

11.5.4.1 GENERAL RULES

C11-31: The **IMxPReq_L** signal **shall** conform to the DC parameters defined in Table 113 on page 453 and the AC parameters defined in Table 115 on page 456.

11.5.4.2 MODULE RULES

o11-10: If the module supports power management as specified in Chapter 14: OS Power Management, the module **shall** implement the **IMxPReq_L** signal on each backplane port.

11.5.4.3 CHASSIS RULES

o11-11: The chassis **may** support the reenablement of power subsystems with the **IMxPReq_L** signal as outlined in Chapter 14: OS Power Management.

11.5.5 IMX* SIGNAL PARAMETERS

This section defines the electrical parameters for the System Management group.

If parameters apply to all signals beginning with "**IMx**", they are referred to as "**IMx***"; the "*" indicates a generic placeholder. If parameters apply to specific signals, these are named explicitly.

11.5.5.1 IMx* DC SPECIFICATIONS

Table 113 IMx* DC Specifications

Symbol	Parameter	Minimum	Maximum	Units	Comments
Vil	Input Low Voltage	-0.5	.8V	V	
Vih	Input High Voltage	2.1	3.63	V	
Vol	Output Low Voltage		0.4	V	at Ipullup Max.
Voh	Output High Voltage	2.2	3.63	V	
Ipullup	Pullup Current		4.0[a]	mA	
C_{IBML_Module}	Module Segment Capacitance		200	pF	
$C_{IBML\_Chassis}$	Chassis Segment Capacitance		100	pF	
I_{IBML_Module}	Module Segment Leakage Current		+/- 80	µA	
$I_{IBML\_Chassis}$	Chassis Segment Leakage Current		+/- 40	µA	
$I_{Isloate}$	Isolation Current		+/- 10	µA	
R_{PRup}[b]	Pull-up Termination	1.870	2.500	KΩ	
R_{PRdn}[b]	Pull-down Termination	2.625	3.630	KΩ	
V_{PU}	IBML Termination Voltage	3.036	3.63	V	Nominal 3.3V

a. The maximum value includes both the current through the pull-up resistor and current from all bus agents. In this context the parameter Ipullup is equivalent to Iol. The minimum value is implementation dependant
b. The values specified for this parameter do not include all tolerances of implemented components.

11.5.5.2 IMxDAT, IMxCLK AC SPECIFICATIONS

Figure 149 shows the electrical timing parameters for the **IMxDat** and **IMxClk** signals. The specification values for these parameters are found in Table 114.

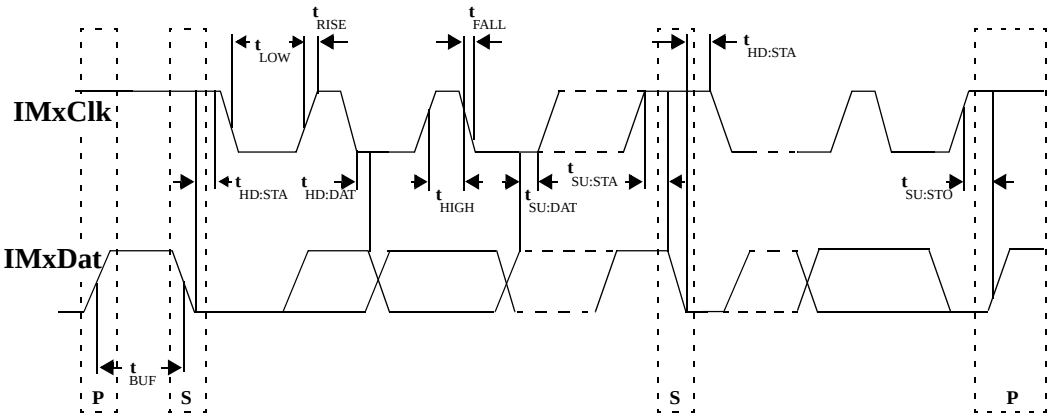


Figure 149 IMxDat, IMxClk Timing Measurements

IBTA

Table 114 IMxDat, IMxClk AC Timing Specifications

Symbol	Parameter	Minimum	Maximum	Units	Comments
f_{IBML}	IB-ML Operating Frequency		100	KHz	
$IBML_{TP}$	IB-ML Throughput	10		KHz	
t_{FALL}	Clock/Data Fall Time		300	ns	See note [a]
t_{RISE}	Clock/Data Rise Time		1000	ns	See note [a]
t_{OF}	Output fall time from V_{IH}min to V_{IL}max with a bus capacitance from 10 pF to 400 pF		250	ns	See note [b]
t_{BUF}	Bus free time between Stop and Start Condition	4.7		µs	See Figure 149 on page 454.
$t_{HD:STA}$	Hold time after (Repeated) Start Condition. After this period, the first clock is generated.	4.0		µs	See Figure 149 on page 454
$t_{SU:STA}$	Repeated Start Condition setup time	4.7		µs	See Figure 149 on page 454
$t_{SU:STO}$	Stop Condition setup time	4.0		µs	See Figure 149 on page 454
$t_{HD:DAT}$	Data hold time	300		ns	See Figure 149 on page 454
$t_{SU:DAT}$	Data setup time	250		ns	See Figure 149 on page 454
t_{LOW}	Clock low period	4.7		µs	
t_{HIGH}	Clock high period	4.0	50	µs	See note [c]

a. Rise and fall time are defined as follows: T_{RISE} = (V_{IL}MAX - 0.15) to (V_{IH}MIN + 0.15), t_{FALL} = 0.9V_{PU}NOM to (V_{IL}MAX - 0.15)
b. The maximum t_{FALL} for the **IMxClk** and **IMxDat** lines specified is longer than the specified maximum t_{OF} for the output stages. This allows series protection resistors (Rs) to be connected between the **IMxClk**/**IMxDat** pins and the **IMxClk**/**IMxDat** lines without exceeding the maximum specified t_{FALL}.
c. t_{HIGH} MAX provides a simple guaranteed method for devices to detect bus idle conditions.

11.5.5.3 IMxINT_L, IMxPRST, IMxPREQ_L AC SPECIFICATIONS

Table 115 IMxInt_L, IMxPRst, IMxPReq_L AC Timing Specifications

Symbol	Parameter	Minimum	Maximum	Units	Comments
$t_{IMxFALL}$	Fall Time		300	ns	See note [a]
$t_{IMxRISE}$	Rise Time		1000	ns	See note [a]
t_{Reset}	IMxPRst pulse width	1		ms	

a. Rise and fall time for the **IMxPRst** signal is defined as follows: $t_{RISE} = V_{IL}MAX$ to $V_{IH}MIN$, $t_{FALL} = V_{IH}MIN$ to $V_{IL}MAX$. Rise and fall time for the **IMxInt and IMxPReq_L** signals are defined as follows: $t_{RISE} = V_{OL}MAX$ to $V_{OH}MIN$, $t_{FALL} = V_{OH}MIN$ to $V_{IL}MAX$.

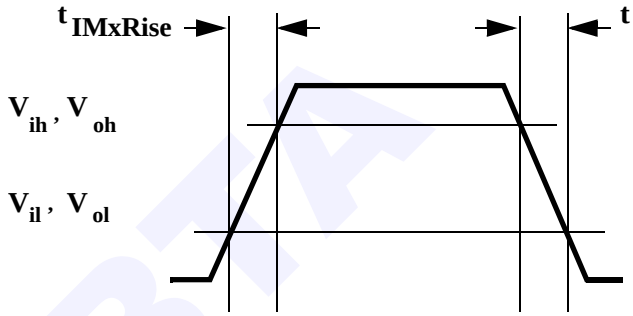


Figure 150 IMxPRst, IMxInt, IMxPReq_L Timing Measurements

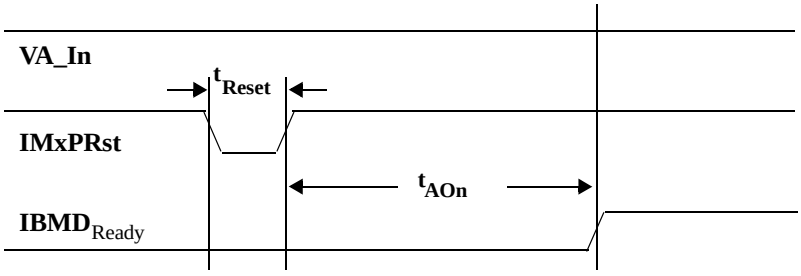


Figure 151 IMxPRst Timing Measurements

CHAPTER 12: POWER / HOT PLUG

12.1 INTRODUCTION

This chapter defines the power use and control of the module form factors defined by this specification. Where necessary, backplane requirements are stated to produce module to chassis interoperability as intended by the architecture.

The power control of non-IB devices that are sub-ordinate to xCAs (e.g. I/O controllers) are outside of the scope of this specification.

Due to the point-to-point nature of the InfiniBand link, the issues with data signal isolation that exist with multi-dropped bus topologies are not present. However, power needs to be controlled so as to not damage a module due to high in-rush currents.

12.1.1 LOCAL DC-DC POWER CONVERTER(S)

The DC-DC power converter(s) on the IB adapter module supply local power to on-board devices at their required voltage(s). This allows the IB standard to be independent of the unique voltages required by any particular adapter.

12.1.2 POWER CLASSES

Two Power Classes are defined for InfiniBand modules:

- Class I implementations utilize converters that may or may not isolate **VB_Ret** from logic ground on the module.
- Class II implementations ONLY utilize converters that isolate **VB_Ret** from logic ground on the module.

  See Section 12.4, “Module Power Rules,” on page 467 for the parametric definition of isolation for Class II.

The local DC-DC power converter(s) incorporated on an IB adapter may be implemented under either of the two power classes.

12.1.3 POWER TOPOLOGIES

InfiniBand adapters that draw from bulk power have a representative topology as depicted in Figure 152 Class I Power Topology on page 458 and Figure 153 Class II Power Topology on page 459.

IBTA

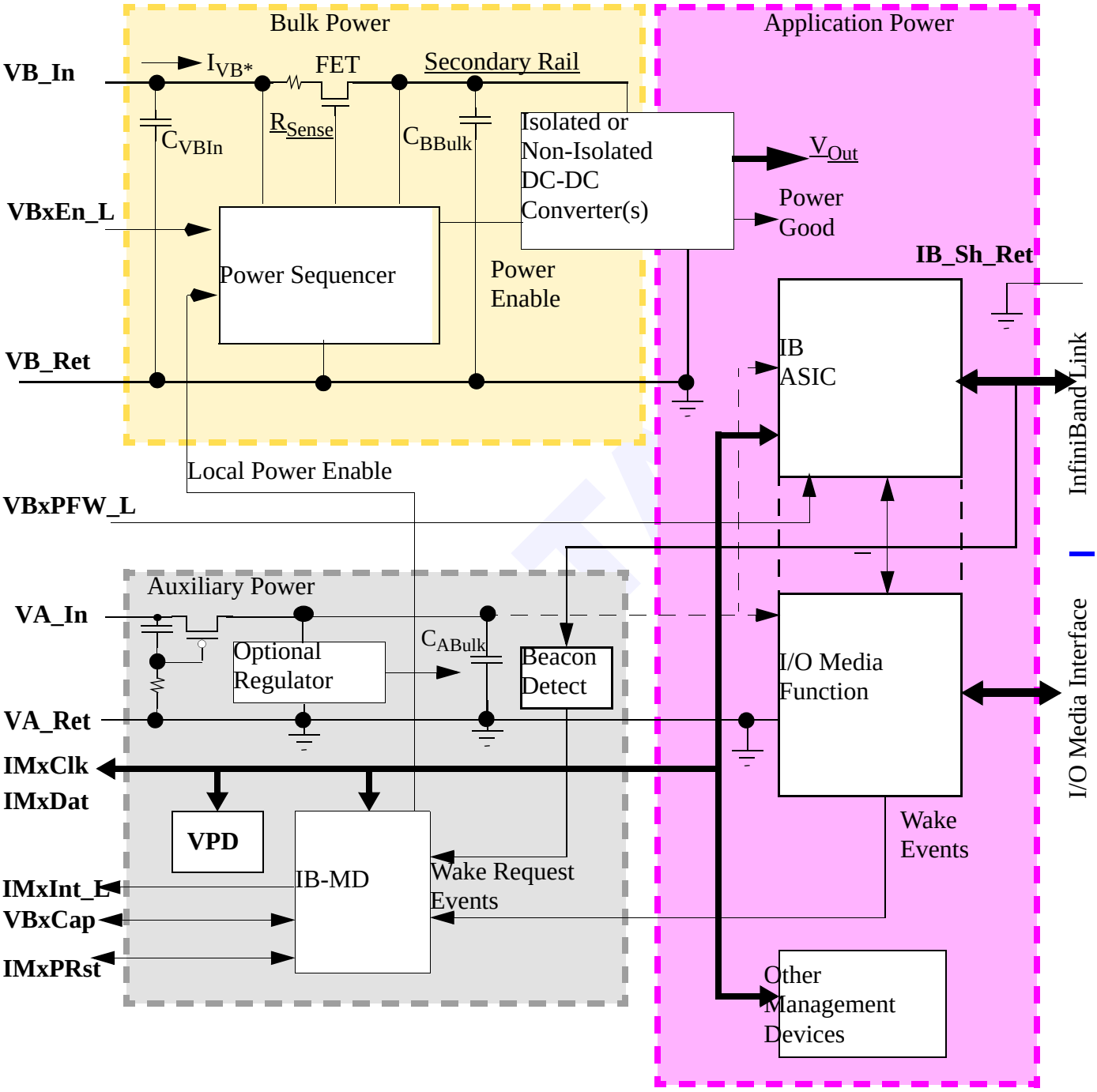
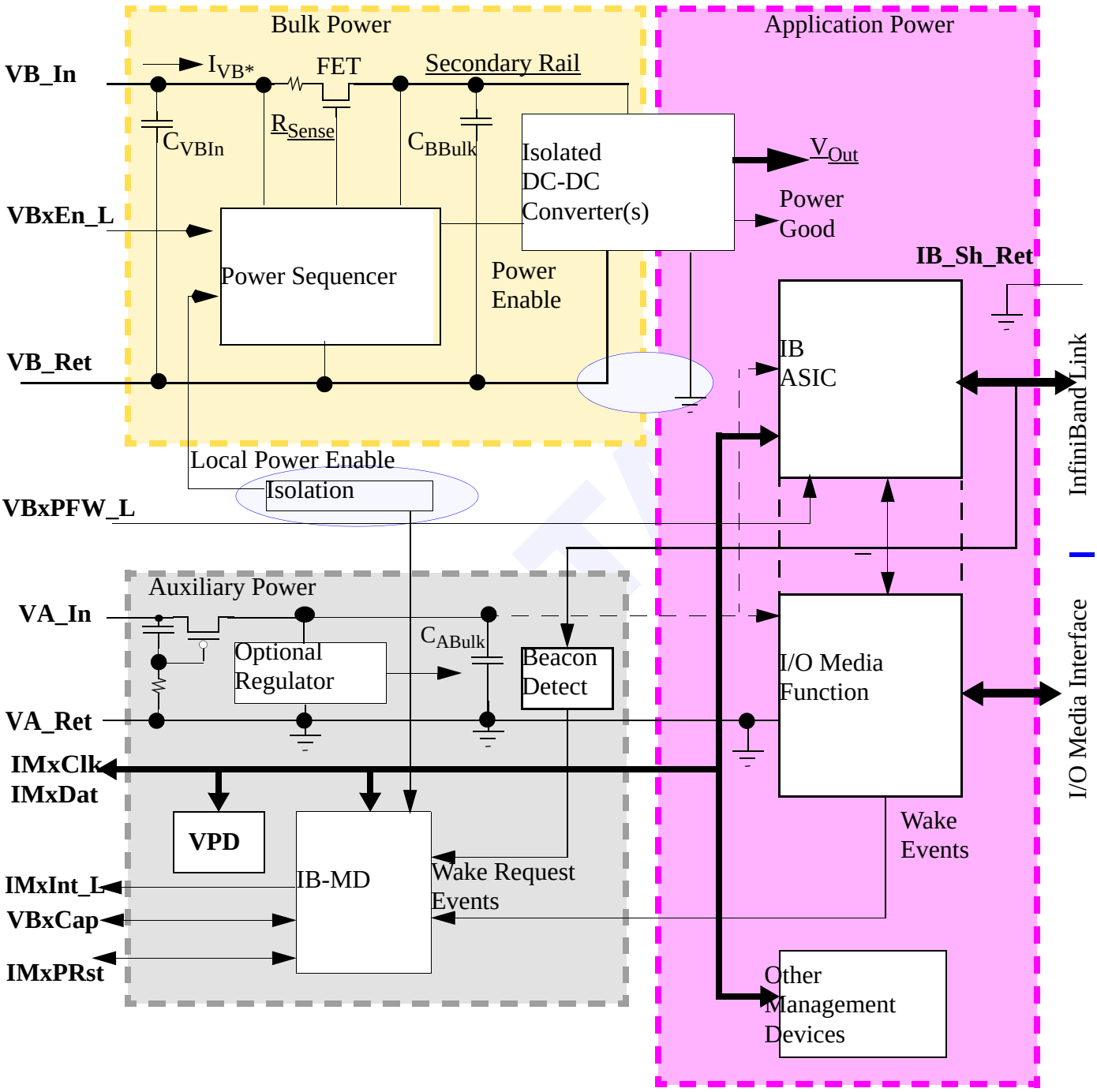


Figure 152 Class I Power Topology



Note: The differences between the Class I and Class II topologies are highlighted with

Figure 153 Class II Power Topology

12.1.4 POWER PORTS

Power is supplied from the backplane within a chassis to the module through the "Bulk Power Group" and "Auxiliary Power Group" contacts. These groups collectively define a "Power Port". Figure 154: Module Bulk Power Ports (Logical) depicts the relationship of Power Port to physical modules.

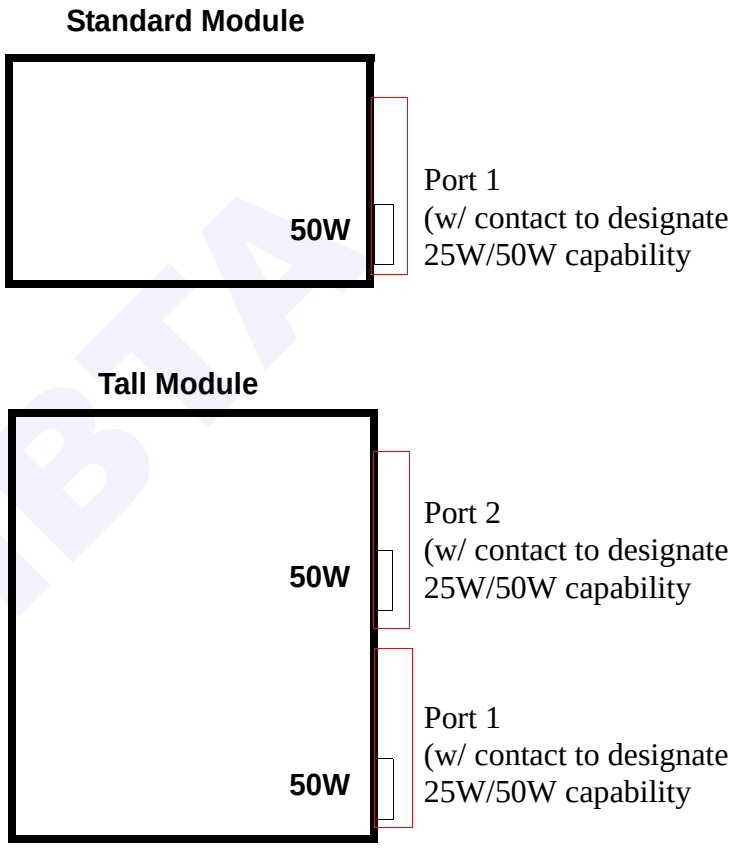


Figure 154 Module Bulk Power Ports (Logical)

12.2 MODULE POWER PORT AND CONSUMPTION RULES

C12-1: All modules **shall** limit their power consumption to the values defined in Table 116 for their respective form factor.

C12-2: All modules **shall** have at least one power port located at Primary Port 1 as defined by Figure 154 and Section 12.5, "Chassis Power Rules," on page 470.

o12-1: Tall or Wide modules **may** have additional power ports as defined in Table 116 for the respective form factor.

Table 116 Module Form Factor Power Summary

Module Form Factor	Power Ports	Maximum Bulk Power Consumed	Maximum Auxiliary Power Consumed	Maximum Thermal Dissipation[ab]
Standard, Single Wide	1	25W	1.2W	25W
Standard, Double Wide	1or 2	50W	1.2W	50W
Tall, Single Wide	1 or 2	50W	2.4W	50W
Tall, Double Wide	1, 2, 3, or 4	100W	2.4W	100W

a. Dissipation is integrated over a 5 second interval
b. The module must ensure that the dissipated power does not exceed the Dissipation value for the form factor regardless of input power sources.

12.3 MODULE / CHASSIS COMPATIBILITY

This section defines the compatibility of the defined chassis and module types.

In this section, the terms "Redundancy" or "Redundant" refer to the design technique where power is drawn from multiple power ports, if available, in such a manner that if a power port fails, the full module power may be drawn through other power ports. Conversely, "Not Redundant" indicates that additional power ports are not available to take on the full load; this could be due to the lack of additional physical power ports (i.e. a standard module) or by design (i.e. a 100W double wide tall module).

12.3.1 STANDARD MODULE, STANDARD CHASSIS

The compatibility of Standard InfiniBand modules within a Standard Chassis from a power view is defined in Table 117.

Table 117 Standard Module, Standard Chassis Slot Power

Standard Module Type				Slot[a]	Backplane Slot Connections[b]	
Width[c]	Power[d]	Boards	Redundancy[e]		Power Port 1, 25W	Power Port 1, 50W
S	25	1	N	P	**25W**	**25W**
D	25	1	N	P	**25W**	**25W**
				A	**No Connection**	**No Connection**
D	25	2	Y	P	**25W**	**25W**
				A	**25W(R)**	**25W(R)**
D	50	1	N	P	**Not Supported**	**50W**
				A	**No Connection**	**No Connection**
D	50	2	N	P	**Not Supported**	**50W**
				A[f]	**No Connection for Power**	**No Connection for Power**
D	50	2	N	P	**25W**	**25W**
				A	**25W**	**25W**
D	50	2	Y	P	**Not Supported**	**50W**
				A	**Not Supported**	**50W(R)**

a. "P" - Primary slot; "A" - Adjacent slot to the right as viewed from the I/O Plate. Adjacent is applicable for Double Wide modules only.
b. A port must be enabled with **VBxEn_L** being asserted (low). "25W" at a Power Port indicates that **VBxCap** is driven low by the backplane; "50W" at a Power Port indicates that **VBxCap** is pulled high by the backplane.
c. "S" - Single wide; "D" - Double wide.
d. Power value stated is the maximum allowed.
e. Indicates whether the particular module supports redundant power ports. "N" - No; "Y" - Yes. If "Y", one or more power ports are designated as "redundant" with the form " xxW(R)" where "xx" is the maximum power drawn.
f. While this module type has 2 boards and may contact the backplane for attachment of links, it is not contacting the power contacts to avoid the need for multiple converters. Rather, the module is pulling all of its power from the Primary slot.

Recommendation to Module Designer:

While it is possible for a Standard Double Wide module to draw 50W from two adjacent 25W ports, it is recommended that the module draw its power from only the primary port. (See recommendation below.)

Recommendation to System Designer

The chassis may provide 50W to each port unconditionally.

If the power delivery requirement for a slot is exceeded (25W for this case), the chassis may disable power to the adjacent port to the right. This is because only a double wide standard module can draw more than 25W (via **VBxCap**) and it inherently occupies the adjacent slot to the right.

12.3.2 TALL MODULE, TALL CHASSIS

The compatibility of Tall InfiniBand modules within a Tall Chassis from a power view is defined in Table 118.

IBTA

Table 118 Tall Module, Tall Chassis Slot Power

Tall Module Type				Slot[a]	Backplane Slot Connections[b]			
Width[c]	Power[d]	Boards[e]	Redundancy[f]		Power Port 1, 25W	Power Port 1, 50W	Power Port 2, 25W	Power Port 2, 50W
S	25	1	N	P	**25W**	**25W**	**No Connection**	**No Connection**
S	25	1	Y	P	**25W**	**25W**	**25W(R)**	**25W(R)**
S	50	1	N	P	**25W (Requires Port 2)**	**50W**	**25W**	Don't Care
S	50	1	Y	P	**Not Supported**	**50W**	**Not Supported**	**50W(R)**
D	50	1	N	P	**25W (Requires Port 2)**	**50W**	**25W**	Don't Care
				A	**No Connection**	**No Connection**	**No Connection**	**No Connection**
D	50	2	N	P	**25W**	**50W**	**25W**	Don't Care
				A[g]	**No Connection for Power**	**No Connection for Power**	**No Connection for Power**	**No Connection for Power**
D	50	2	N	P	**25W**	**25W**	**No Connection**	Don't Care
				A	**25W**	**25W**	**No Connection**	**No Connection**
D	50	2	Y	P	**25W**	**50W**	**25W(R)**	Don't Care
				A	**25W**	**50W(R)**	**25W (R)**	Don't Care
D	100	1	N	P	**Not Supported**	**50W**	**Not Supported**	**50W**
				A	**No Connection**	**No Connection**	**No Connection**	**No Connection**
D	100	2	N	P	**25W**	**50W**	**25W**	Don't Care
				A	**25W**	**50W**	**25W**	**No Connection**
D	100	2	Y	P	**Not Supported**	**50W**	**Not Supported**	**50W(R)**
				A	**Not Supported**	**50W**	**Not Supported**	**50W(R)**

a. "P" - Primary slot; "A" - Adjacent slot to the right as viewed from the I/O Plate. Adjacent is applicable for Double Wide modules only.
b. A port must be enabled with **VBxEn_L** being asserted (low). "25W" at a Power Port indicates that **VBxCap** is driven low by the backplane; "50W" at a Power Port indicates that **VBxCap** is pulled high by the backplane.
c. "S" - Single wide; "D" - Double wide.
d. Power value stated is the maximum allowed.
e. Indicates the number of PCBs implemented within the form factor that connect to the backplane.

f. Indicates whether the particular module supports redundant power ports. "N" - No; "Y" - Yes. If "Y", one or more power ports are designated as "redundant" with the form " xxW(R)" where "xx" is the maximum power drawn.
g. While this module type has 2 boards and may contact the backplane for attachment of links, it is not contacting the power contacts to avoid the need for multiple converters. Rather, the module is pulling all of its power from the Primary slot.

Recommendation to Module Designer:

While it is possible for a Tall Double Wide module to draw 100W from two adjacent 50W ports, it is recommended that the module draw its power from only the primary slot's Port 1 and Port 2. (See recommendation below.)

Recommendation to System Designer

The chassis may provide 50W to each port unconditionally.

Alternatively, if the power delivery required for a slot is exceeded (50W for this case of a tall module), the chassis may disable power to the right adjacent slot's Port 1 and Port 2. This is because only a double wide tall module can draw more than 50W (via **VBxCap**(s)) and it inherently occupies the adjacent slot to the right.

12.3.3 STANDARD CHASSIS, TALL CHASSIS

The compatibility of Standard InfiniBand modules within a Tall Chassis from a power view is defined in Table 119.

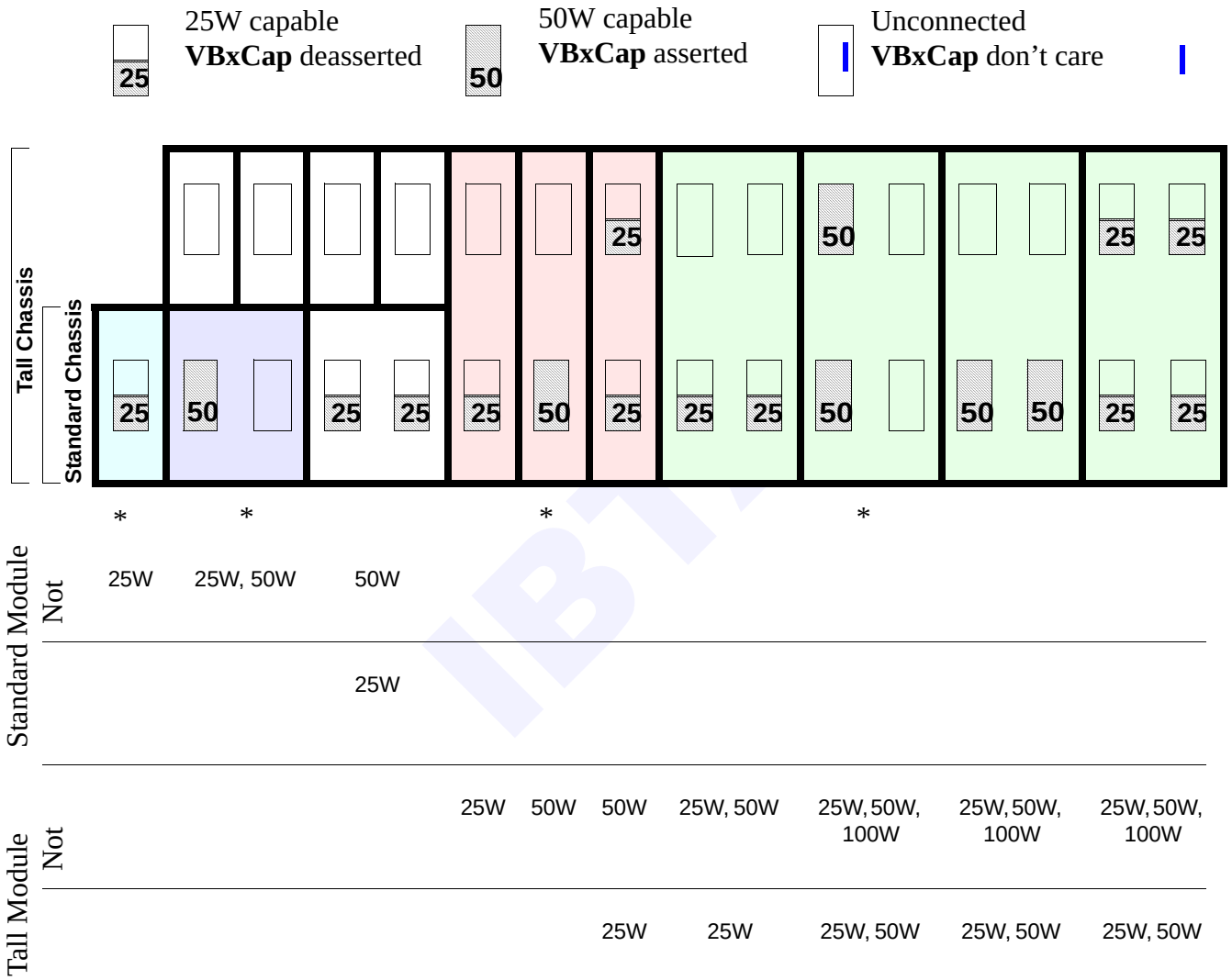
Table 119 Standard Module, Tall Chassis Slot Power

Standard Module Type				Slot[a]	Backplane Slot Connections[b]			
Width[c]	Power[d]	Boards[e]	Redundancy[f]		Power Port 1, 25W	Power Port 1, 50W	Power Port 2, 25W	Power Port 2, 50W
S	25	1	N	P	**25W**	**25W**	**No Connection**	**No Connection**
D	25	1	N	P	**25W**	**25W**	**No Connection**	**No Connection**
				A	**No Connection**	**No Connection**	**No Connection**	**No Connection**
D	25	2	Y	P	**25W**	**25W**	**No Connection**	**No Connection**
				A	**25W(R)**	**25W(R)**	**No Connection**	**No Connection**
D	50	1	N	P	**Not Supported**	**50W**	**No Connection**	**No Connection**
				A	**No Connection**	**No Connection**	**No Connection**	**No Connection**
D	50	2	N	P	**Not Supported**	**50W**	**No Connection**	**No Connection**
				A[g]	**No Connection for Power**	**No Connection for Power**	**No Connection**	**No Connection**
D	50	2	N	P	**25W**	**25W**	**No Connection**	**No Connection**
				A	**25W**	**25W**	**No Connection**	**No Connection**
D	50	2	Y	P	**Not Supported**	**50W**	**No Connection**	**No Connection**
				A	**Not Supported**	**50W(R)**	**No Connection**	**No Connection**

a. "P" - Primary slot; "A" - Adjacent slot to the right as viewed from the I/O Plate. Adjacent is applicable for Double Wide modules only.
b. A port must be enabled with **VBxEn_L** being asserted (low). "25W" at a Power Port indicates that **VBxCap** is driven low by the backplane; "50W" at a Power Port indicates that **VBxCap** is pulled high by the backplane.
c. "S" - Single wide; "D" - Double wide.
d. Power value stated is the maximum allowed.
e. Indicates the number of PCBs implemented within the form factor that connect to the backplane.
f. Indicates whether the particular module supports redundant power ports. "N" - No; "Y" - Yes. If "Y", one or more power ports are designated as "redundant" with the form " xxW(R)" where "xx" is the maximum power drawn.
g. While this module type has 2 boards and may contact the backplane for attachment of links, it is not contacting the power contacts to avoid the need for multiple converters. Rather, the module is pulling all of its power from the Primary slot.

Implementation Note

The following shows some example subset backplane power ports configurations that are allowable. A * indicates the configurations that are likely to be predominent.



12.4 MODULE POWER RULES

12.4.1 GENERAL

This section defines power rules that apply to all modules defined by this specification.

C12-3: A module that draws power from **VB_In shall** supply a bulk DC-DC converter.

C12-4: The module **shall** disable its port-specific DC-DC converter(s) upon the sensing of a deasserted level on **VBxEn_L** on that port.

C12-5: The module **shall not** draw more than $\mathbf{I_{BLeakage}}$ current while **VBxEn_L** is deasserted. Upon **VBxEn_L** going from asserted to deasserted, the module **shall** reduce its current draw to $\mathbf{I_{BLeakage}}$ or less within the time specified by $\mathbf{t_{BOff}}$.

Note: Chassis are not required to provide $I_{BLeakage}$ current while **VBxEn_L** is deasserted.

C12-6: The module power converter **shall** provide input fault current limiting[3] and set its threshold at less than 200% of $\mathbf{I_{BMax}}$. The current limiting **shall** be activated within 20µs of crossing the current limit threshold.

Implementation Note

The current limit implementation is intended to be located in the hot-plug circuit, where the Hot-plug FET is turned off or controlled for a over-current fault. Thus, protection for over current faults or shorts on the regulator's output, internal to the regulator and of the Secondary Rail (after the hot plug FET) are covered. Examples of acceptable over current protection methods are:

- Fold Back - Over current limit is reached and FET is controlled to fold back the allowed current
- Hiccup - A failure mode where the power sequencer supplies power to the voltage converter, current limiting occurs, and power is removed. This sequence can continuously repeat.
- Latch off - Over-current cause hot plug FET to turn off, and an on/off command is required to reset

C12-7: The module **shall** set *InterruptClearStatus.Power_Converter_Fault* upon the detected violation of the $\mathbf{I_{BMax}}$ parameter or any condition that causes the DC-DC converter to be off.

C12-8: The module **shall** provide over voltage protection sufficient to not produce a voltage greater than $\mathbf{V_{RHP}}$ defined in Table 20 Receiver Char-

3. The faults for which the fault current limiting is intended are shorts to ground on the Secondary Rail, within the DC-DC converter, or for V_{Out}. (Refer to Figure 152 on page 458 for terminology references). Shorting faults across any active current limiting components (series FETs for example) are not considered covered by this mechanism.

acteristics for 2.5 Gb/s on page 192 on the **IBsxIw** or **IBsxOw** signals, even in the presence of "let-through" converter fault.

> **Implementation Note**
>
> Many ASIC driver and receiver designs intrinsically supply clamping circuitry (typically diodes) to prevent signal levels from exceeding the input power supply specifications of that ASIC. Many of these clamping circuits can allow fault generated over-voltage conditions to pass on to the external (off-ASIC) portion of the driver or receiver and cause a violation of far end parameters (V_{RHP} specifically). Designers should take this possibility into account when selecting components and designing over-voltage protection.

C12-9: The module **shall** set the *ModulePowerInfo.PowerClass* field to reflect the implemented Power Class.

o12-2: If the *ModulePowerInfo.PowerClass* field indicates Class II, the module **shall** provide a minimum isolation of 2.0 MΩ DC resistance @ 500V of isolation between **VB_Ret** and **Varlet** and a minimum isolation of 2.0 MΩ DC resistance @ 500V between **VB_Ret** and **IB_Sh_Ret**.

> **Note to Module Implementers**
>
> A Class II chassis is not required to permit Class I adapters to power on. A Class I chassis must permit Class II adapters to power on if their power requirements do not exceed the chassis/slot limit.

C12-10: A module **shall** support "Hot Add" functionality as defined in Section 12.6.1 on page 471.

C12-11: A module **shall** support "Surprise Hot Removal" functionality as defined in Section 12.6.2 on page 473.

C12-12: A module **shall** draw auxiliary power only from the **VA_In** contact of Primary Port 1.

C12-13: The module **shall** set the *ModulePowerInfo.RedundantPower* field to reflect implemented redundant power port connections.

12.4.2 MULTIPLE POWER PORTS

If a module contains more than one power port, the rules of this subsection apply.

o12-3: If a module contains more than one power port, it **shall not** connect the **VB_In** signals from multiple power ports together.

Implementation Note

The above non-bridging requirement implies that when power multiple connections are made (using multiple power ports), a DC-DC converter per port is required.

12.4.3 TALL DOUBLE WIDE

o12-4: A Tall Double Wide module **may** consume up to 2.4W only from **VA_In** from Primary Port 1.

12.5 CHASSIS POWER RULES

This section states the chassis power requirements for supporting the modules defined by this specification.

For vertical module orientations, the "primary" slot is defined to be the left-most slot of an adjacent pair as viewed from the I/O plate. For horizontal module orientations, the "primary" slot is defined to be the lower-most slot of an adjacent pair as viewed from the I/O plate. Refer to Section 9.1.3.1, “Slot Designations,” on page 360.

12.5.1 GENERAL

This section defines the power rules for a chassis that accepts a module defined by this specification.

C12-14: An Unmanaged standard height chassis **shall** be capable of supplying a minimum of 25W on **VB_In** at any existent port.

o12-5: An Actively Managed standard chassis **may** choose to manage the power distribution according to *ModulePowerInfo* rather than providing 25W to all existent ports simultaneously.

C12-15: The chassis **shall** provide for the assertion of the **VBxEn_L** signal per power port for which it intends to supply power.

C12-16: The chassis power supply(s) **shall** be capable of supplying a minimum of 1.2W on **VA_In** at every slot simultaneously.

o12-6: If a chassis accepts Tall modules, it **shall** provide distribution and any protection mechanisms that allow a minimum of 2.4W to be delivered to **VA_In** at any Port 1.

Implementation Note

The module does not inherently provide 240VA protection. This is left to the system implementation.

Implementation Note

High Availability chassis will need to account for a dead short between any contacts on a module prior protection circuitry.

12.6 HOT PLUG

12.6.1 HOT ADD

"Hot Add" is defined as the insertion of an InfiniBand module into a backplane that has both V_{Bulk} and V_{Aux} present. The module powers up and initiates a training sequence.

The InfiniBand backplane connector defines four (4) mating contact levels and that are referred to as Level 1, Level 2, Level 3 and Level 4. See Chapter 10: Backplane Connector Specification. The contact sequencing that takes place upon module insertion, referred to a "mate order", is de-

picted in Figure 155: Module Insertion Contact Sequence and then further described.

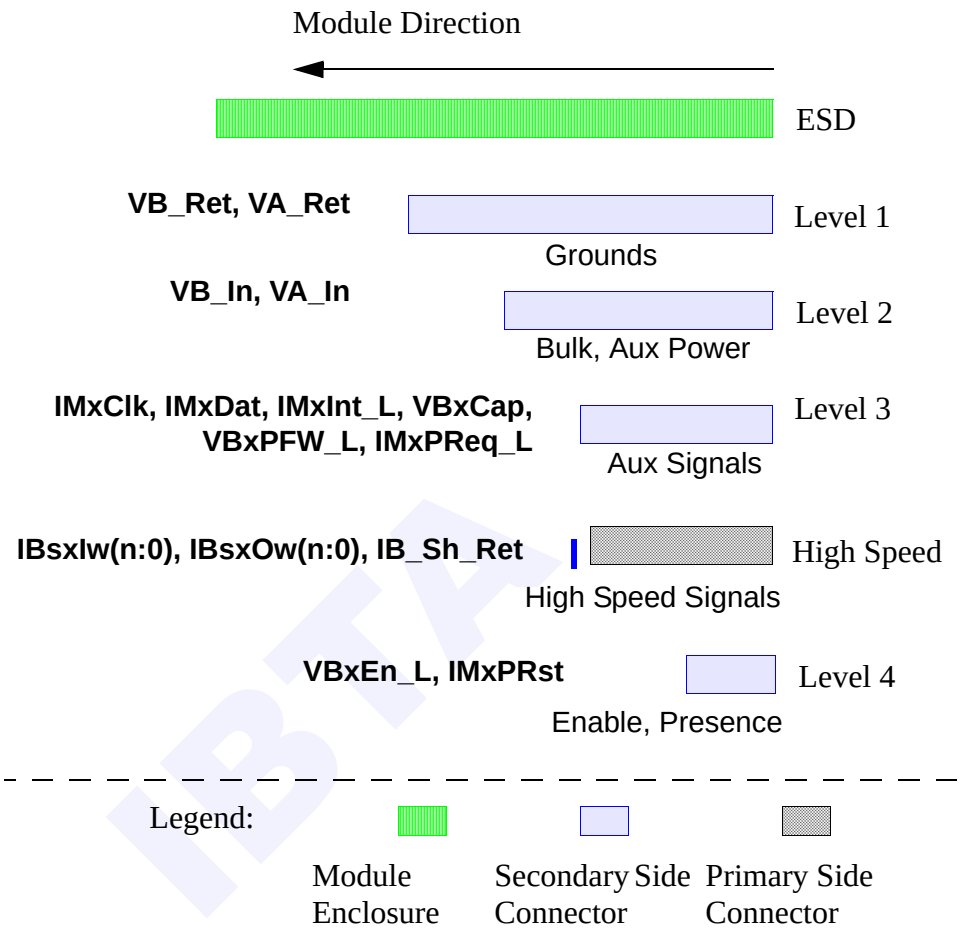


Figure 155 Module Insertion Contact Sequence

1) ESD**:** Upon insertion of the module in the chassis slot, the EMI springs provide a path for ESD from the module to the chassis.
2) Level 1 Mate: **VB_Ret**, **VA_Ret** and **IB_Sh_Ret** mate to provide current return contacts and establish reference voltage levels.
3) Level 2 Mate: Power contacts **VB_In** and **VA_In** mate.

   The bulk domain does not begin ramping due to **VBxEn_L** being pulled high (deasserted) through pull-up termination. See Figure 142 VBxEn_L Termination on page 438 for details.

   The auxiliary domain begins ramping based on input capacitance $\mathbf{C_{ABulk}}$.
4) Level 3 Mate: The **IMxDat**, **IMxClk**, **IMxInt_L**, **IMxPReq_L**, **VBxPFW_L**, **VBxCap** contacts mate to allow for IB-ML functionality.

High Speed Mate: The high speed link interface contacts, **IBsxIw** and **IBsxOw**, mate.

The timing relationship between Level 3 Mate and High Speed Mate is connector dependent and not specified. However, for the operability of Hot Add, this relationship is unimportant. What is important is the High Speed Mate to Level 4 Mate; this is strictly controlled by the connector (See Chapter 10: Backplane Connector Specification on page 409).

5) Level 4 Mate: The **VBxEn_L** contact mates to allow for bulk power enablement. **IMxPRst** mates to signal the backplane that a connection is now present on this IB port.

   For a Unmanaged Chassis, the **VBxEn_L must** be asserted to allow the on-board sequencer to be enabled.

   For a Managed Chassis, the **VBxEn_L may** be deasserted by the backplane to prevent the enablement of the sequencer. A CME, using the componentry implemented on auxiliary power, **may** read VPD elements to make whatever determination appropriate before asserting **VBxEn_L**. The sequencer, seeing that **VB_In** has reached $\mathbf{V_{Bulk}}$(Min), and having been control enabled (**VBxEn_L**, Local Power Enable), will turn on the input to the DC-DC converter. This allows for the module logic to be powered from the output of the converter.

12.6.2 SURPRISE HOT REMOVAL

"Surprise Hot Removal" is defined as the removal of an InfiniBand module from a backplane that has both $\mathbf{V_{Bulk}}$ and $\mathbf{V_{Aux}}$ present without first being placed in a quiescent state.

The InfiniBand backplane connector defines four (4) mating contact levels and are referred to as Level 1, Level 2, Level 3 and Level 4. See Chapter 10: Backplane Connector Specification.

The contact sequencing that takes place upon module removal, referred to a "break order", is depicted in Figure 156: Module Removal Contact Sequence and then further described.

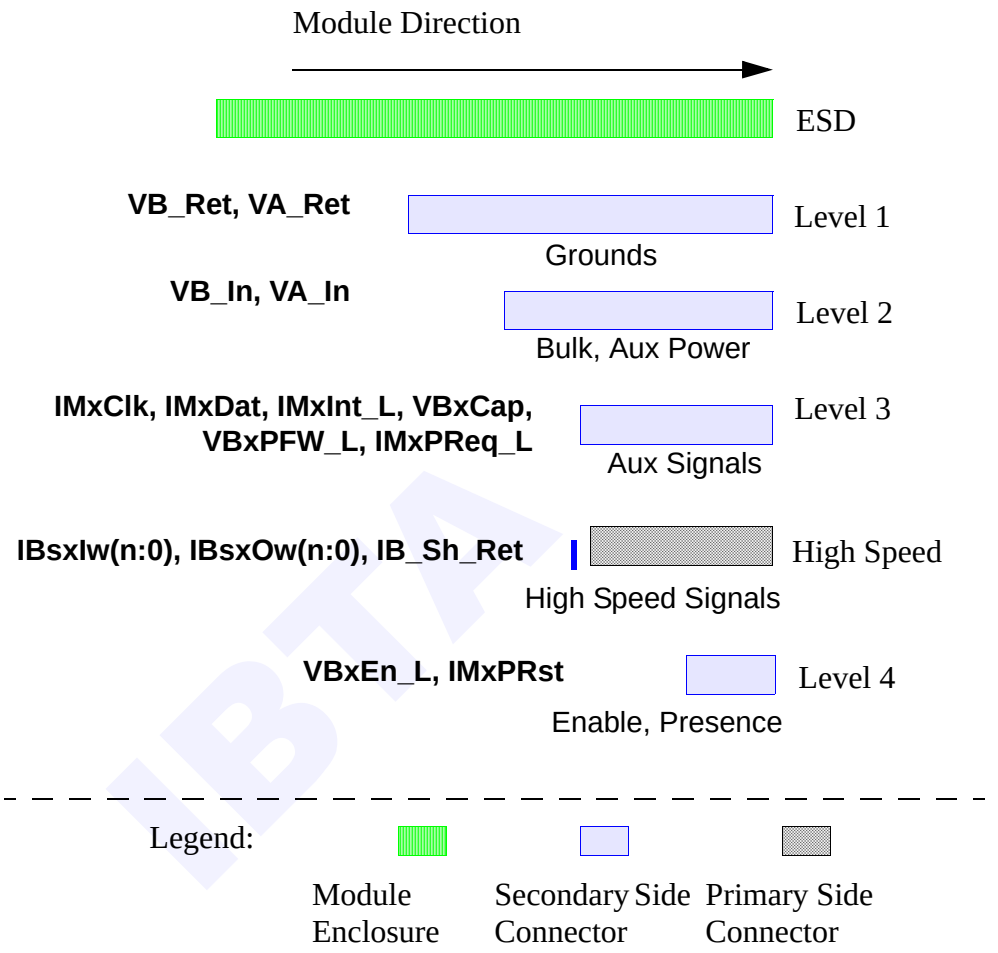


Figure 156 Module Removal Contact Sequence

1) Level 4 Break: The Bulk Power control signal, **VBxEn_L,** is the first to break from a power view. The module **must** initiate converter shutdown and have the current draw below $I_{BLeakage}$ within t_{BOff}.

2) Level 3 Break: The **IMxDat**, **IMxClk**, **IMxInt_L**, **IMxPReq_L**, **VBxPFW_L**, **VBxCap** contacts break to disconnect IB-ML.

   High Speed Break: The **IBsxIw** and **IBsxOw** contacts disengage. Power to the logic function driving the **IBsxOw** signals must be removed due to **VBxEn_L** being deasserted.

   The timing relationship between Level 3 Break and High Speed Break is connector dependent and unspecified. However, for the operability of Surprise Hot Removal, this relationship is unimportant. What is im-

portant is the Level 4 Break to High Speed Break timing relationship which is strictly controlled by the connector.

3) Level 2 Break: Power contacts **VB_In** and **VA_In** break. The current being below $\mathbf{I_{BLeakage}}$ on **VB_In** prevents contact damage upon the break. **VA_In** can break at full current ($\mathbf{I_{AMax}}$) without contact damage.

4) Level 1 Break: Return contacts **VB_Ret**, **VA_Ret** and **IB_Sh_Ret** break.

5) ESD: This allows for residual charge to be dissipated.

12.6.3 GRACEFUL HOT REMOVAL

"Graceful Hot Removal" is the removal of an adapter module that has first been placed in a quiescent state. $\mathbf{V_{Bulk}}$ may or may not be on.

For the case that $\mathbf{V_{Bulk}}$ is still on, the Graceful Hot Removal is dependent on the Surprise Hot Removal functionality described in Section 12.6.2 and required in Section 12.4 on page 467.

The features required for the achievement of the quiescent state are described in Section 13.3.2.4, "Graceful Hot Removal," on page 503.

12.7 POWER MANAGEMENT WAKE-UP

12.7.1 BEACON SEQUENCE - CHASSIS POWER ON

This sequence assumes that the module has been inserted into a chassis and that the chassis is in a "Full On" condition. (both Auxiliary and Bulk Power are available to the slot). Additionally, it is assumed that the module was powered down by setting the PowerDown bit, which is derived from the ***ModulePMControl.PMState*** field; this would deassert Local Power Enable to the sequencer.

1) A beaconing sequence is detected.

2) PowerDown is reset. This causes Local Power Enable to assert assuming the absence of other environmental faults.

   Additionally, the **IMxPReq_L** is asserted to an existent CME or power subsystem control element.

3) The sequencer is enabled (**VBxEn_L** is asserted due to assumed proper mating with backplane and **VB_In** being at assumed appropriate level).

4) The sequencer, seeing that **VB_In** has reached $\mathbf{V_{Bulk}}$(Min), and having been control enabled (**VBxEn_L**, Local Power Enable), will turn on the input to the DC-DC converter. This allows for the module logic to be powered from the output of the converter.

12.7.2 BEACON SEQUENCE - CHASSIS POWER OFF

This sequence assumes that the module has been inserted into a chassis and that the chassis is in a "Off but Power Available" condition. (Auxiliary Power is on, Bulk Power is off to the slot). Additionally, it is assumed that the module was powered down by setting a PowerDown bit, which is derived from the ***ModulePMControl.PMState*** field; this would deassert Local Power Enable to the sequencer.

1) A beaconing sequence is detected under Auxiliary Power.
2) PowerDown is reset. This causes Local Power Enable to assert assuming the absence of other environmental faults.

   Additionally, the **IMxPReq_L** is asserted to the backplane.

   The sequencer is enabled (**VBxEn_L** is asserted due to assumed proper mating with backplane). However, **VB_In** is off due to the chassis being "Off but Power Available". The sequencer will monitor the availability of voltage on **VB_In** to enable the converter(s).
3) The chassis, being in the "Off but Power Available" condition, sees the assertion of **IMxPReq_L**. This indication allows the chassis to enable the Bulk Power.

   The means by which the chassis was put into the "Off but Power Available" state are not specified; it is outside the scope of the InfiniBand specification. It is likely that a chassis which implements this type of functionality will have a CME of some form.
4) **VB_In** ramps up through the chassis power-on. The sequencer, seeing that **VB_In** has reached $\mathbf{V_{Bulk}}$(Min), and having been control enabled (**VBxEn_L**, Local Power Enable), will turn on the input to the DC-DC converter. This allows for the module logic to be powered from the output of the converter.

CHAPTER 13: HARDWARE MANAGEMENT

13.1 INTRODUCTION

This chapter describes the functions that manage, control, and monitor physical components of InfiniBand Modules and the Chassis in which they reside. This chapter addresses hardware management for InfiniBand topologies mentioned in Section 6.8, "Compliance Channel - DDR and QDR," on page 203. The xCAs and Switches that are packaged in a form factor other than those defined by this specification, which optionally provide the defined hardware management functionality, are also described.

Additionally, this chapter defines required and optional features of the InfiniBand IB-Modules, the InfiniBand Chassis, the InfiniBand Target Channel Adapters, the InfiniBand Switches, and the InfiniBand Host Channel Adapters to support the following:

1) Communication with the Baseboard Manager
2) In-Band and Out-of-Band facilities for Graceful Hot Removal
3) Standard visual indicators to assist the user in Hot Add and Hot Removal of IB-Modules
4) Access to IB-Module Vital Product Data (VPD)
5) Access to IB-Module optional environmental variables
6) Access to IB-Module's LEDs
7) Remote control of a Chassis power state
8) Provide In-band facilities to manage Chassis and IB-Modules on the InfiniBand Fabric

13.1.1 NOMENCLATURE

For consistency with *InfiniBand Architecture Specification, Volume 1*. transfers on IB-ML (defined in Section 13.3.2.2, "IB-ML," on page 495) are done most significant byte, most significant bit first unless otherwise specified. Multi-field structures, as shown in tabular form in this chapter, are delivered from the top of the table to the bottom.

13.2 OVERVIEW

This section briefly describes IB-Module Interfaces, Hardware Management Methods, Managed Units, Classes of Management Transactions, IB-Module Management, Non-Module Device Management, and Chassis Management.

IBTA

13.2.1 MANAGEMENT MODEL

Figure 157 shows the generic Baseboard Management model for an InfiniBand Module as described in *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”, Section “Baseboard Management”.

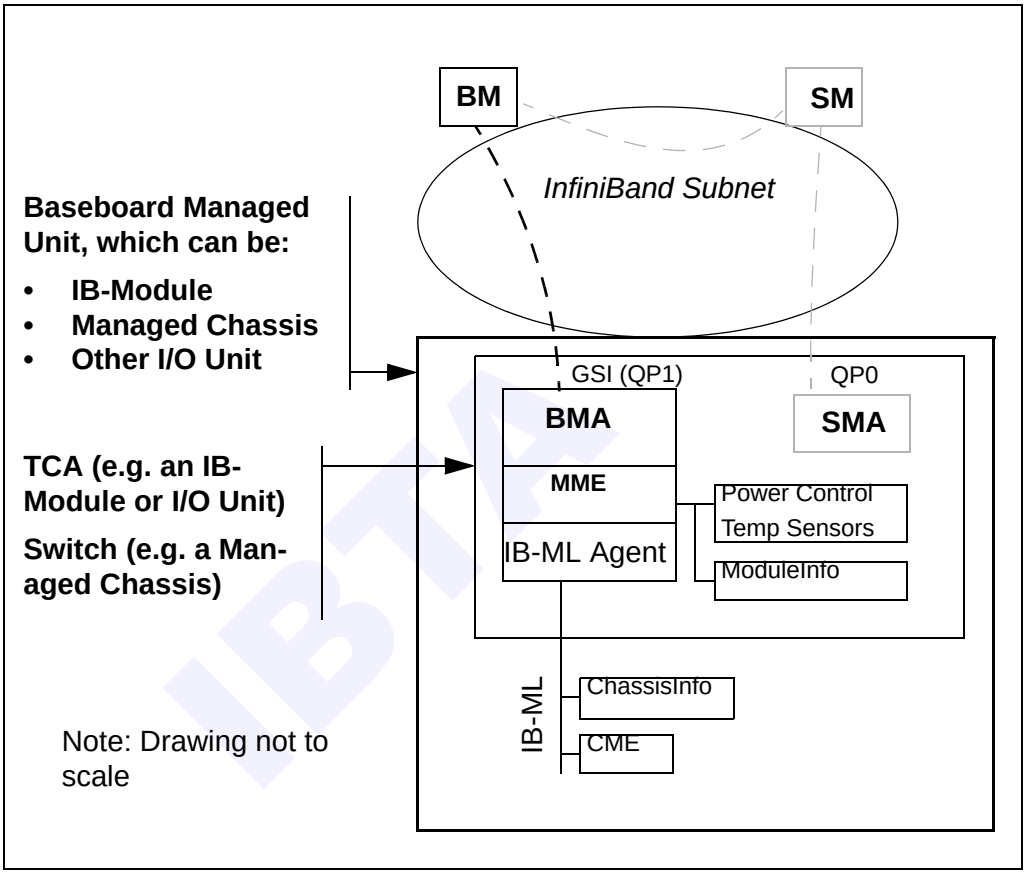


Figure 157 Baseboard Management Architecture

Figure 158 shows a generic management model for an InfiniBand Module.

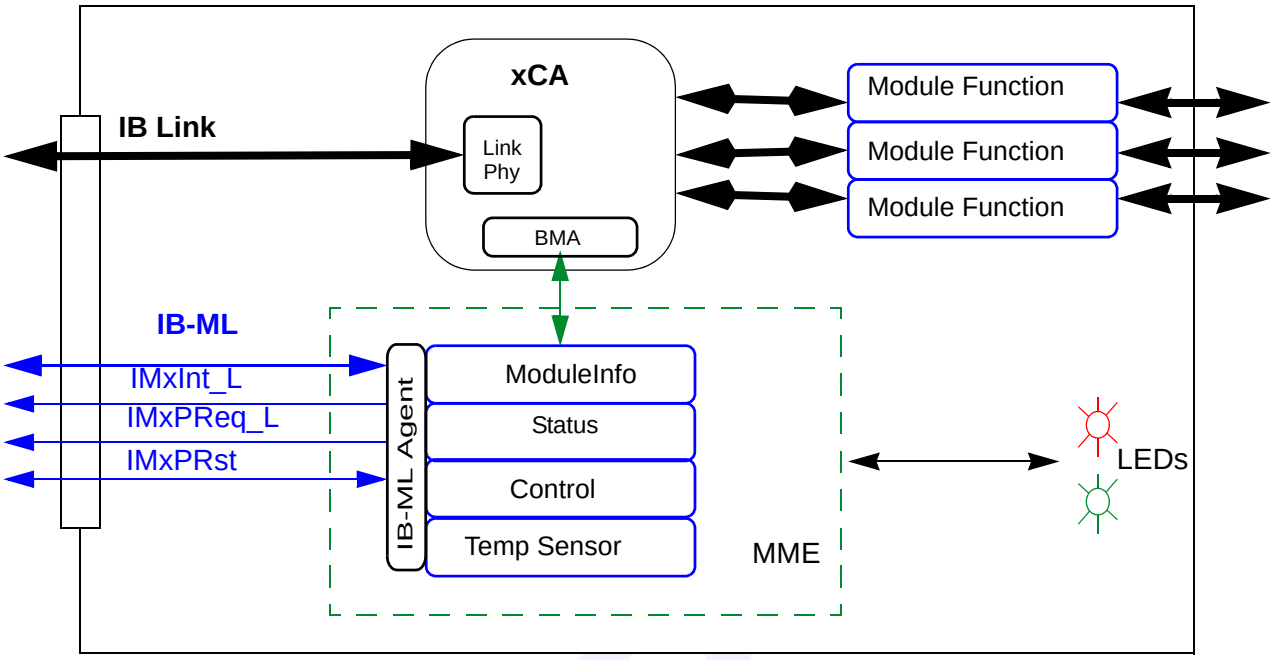


Figure 158 Module Management Model

An InfiniBand Module includes at least one InfiniBand Link, a Baseboard Management Agent (BMA: See *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”, Section “Baseboard Management”), one IB-ML Agent (an interface to the InfiniBand Management Link (IB-ML)), one Module Management Entity (MME), and the applications (functions) the Module performs. The Module, through the logical MME, includes (or emulates) a SEEPROM as a non-volatile storage to hold ModuleInfo (including Vital Product Data, VPD), InfiniBand-required LEDs, local Power Regulation, and other optional management devices. The IB-ML interface is the means by which an optional Chassis Management Entity (CME) may appropriately monitor and control certain aspects of a Module.

InfiniBand does not specify how to implement the required IB-ML and the IB-ML functional blocks. The Module may implement these functions using discrete components on a PCB, integrate them into the IB ASIC, or use a micro-controller; however, they must all be visible via the InfiniBand-specified methods on the IB-ML and the IB Link. Please see Table 120 Module Feature Requirements on page 494 for a list of the required and optional features and Table 126 Module Facility Requirements on page 520 for requirements for facilities supporting the features on Modules.

13.2.2 MANAGED UNITS

The Baseboard Manager desires to have knowledge of the following units:

- HCA
- IB Modules
- Non-Module xCA_ID
- Non-Module xCA_NID
- Non-Module Switch_NID
- Non-Module Switch_ID
- Switches
- Routers
- Repeater Modules
- Chassis into which the above are resident

A unit is considered “manageable” if it provides basic information about itself and where it is located. This information is called “Vital Product Data” (VPD). The InfiniBand specification defines the minimal information required in ModuleInfo and ChassisInfo for a unit to be considered “manageable”.

C13-1: For the purposes of Baseboard Management, a Router **shall** be considered equivalent to a Switch.

13.2.3 MANAGEMENT METHODS

- HCA
- IB Modules
- Non-Module xCA_ID
- Non-Module xCA_NID
- Non-Module Switch_NID
- Non-Module Switch_ID
- Switches
- Routers
- Repeater Modules
- Chassis into which the above are resident

A unit is considered “manageable” if it provides basic information about itself and where it is located. This information is called “Vital Product Data” (VPD). The InfiniBand specification defines the minimal information required in ModuleInfo and ChassisInfo for a unit to be considered “manageable”.

13.2.3.1 IN-BAND MANAGEMENT

In-band management refers to the monitoring and the control of InfiniBand components using messages transferred on their IB Fabric Link. After connecting to the fabric, IB components establish their physical link automatically and are able to receive Management Datagram (MAD) packets on QP0 and QP1.

In-band hardware management messages are Baseboard Management messages that traverse the InfiniBand Fabric. This class is defined in *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”. The Baseboard Management Software (referred to as the “Baseboard Manager”) uses these messages to gather health and inventory information about InfiniBand-attached devices and Chassis. The Baseboard Manager may also perform management control operations such as reset or power control on InfiniBand-attached devices and Chassis. Also see *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”, Section “Baseboard Management”.

InfiniBand Architecture Specification, Volume 1 defines HCAs, Switches, Routers, and TCAs as protocol-aware units and can therefore process In-band management messages; conversely, Repeaters are not protocol-aware and do not comprehend management messages.

All protocol-aware Modules that receive Baseboard class management messages on their General Services Interface (GSI) forward those messages onto the BMA implemented within the Module. As specified in *InfiniBand Architecture Specification, Volume 1*, the Baseboard class uses the General Service Interface (GSI) on QP1 or a redirected QP.

An agent across the IB Link may query *this* IB-Module's management information (ModuleInfo), or the Chassis management information (ChassisInfo) *via* this IB-Module.

13.2.3.1.1 MANAGEMENT BY PROXY

Non-protocol-aware units are not directly addressable on the InfiniBand fabric; a proxy is required to be manageable from the fabric. Chassis and Repeater Modules require proxies.

A proxy is a device that provides InfiniBand Baseboard Management access on behalf of an entity that does not accept InfiniBand Baseboard Management messages. For example, an xCA Module or a Switch may act as a proxy for accessing a Chassis Management Entity (CME) (See Section 13.6.3.2, “Ib2CME,” on page 574 and Section 13.5.2.7, “ChassisInfo Device,” on page 548) while a Chassis Management Entity (CME) or a Switch may act as a proxy for accessing a Repeater Module (See Section 13.5.2.6, “IB-ML Selector Proxy,” on page 548).

13.2.3.2 OUT-OF-BAND MANAGEMENT

Out-of-band management messages traverse a transport other than the InfiniBand fabric.

13.2.3.2.1 IB-ML

InfiniBand specification defines an interface called InfiniBand Management Link (IB-ML). IB-ML is a multi-drop, multi-master, two-wire serial bus which uses SMBus 1.1-based data transfer and arbitration protocols[24]. This interface connects IB-ML devices on IB-Modules with the Chassis. IB-ML does not electrically connect to more than one IB-Module or more than one Backplane connector. (See Table 113 IMx* DC Specifications on page 453 and Table 114 IMxDat, IMxClk AC Timing Specifications on page 455.)

IB-ML Out-of-Band messages, along with other side-band signals such as Power Request (**IMxPReq_L**), Presence Detect (**IMxPRst**), and Interrupt (**IMxInt_L**) allow communication between Chassis and Module entities. IB-ML is available even when the InfiniBand fabric is not operational and before a link is connected. For example, at power up, a Chassis backplane may use the out-of-band IB-ML to enumerate Modules and determine their power requirements prior to enabling their Bulk power.

IB-ML, primarily, provides an out-of-band access media for Modules; however, Switches, Routers, and other Channel Adapter applications may use IB-ML as well.

13.2.3.2.2 VIRTUAL IB-ML

A Channel Adapter or a Switch may provide access to an IB-ML via Baseboard Management messages received via the InfiniBand fabric. To software sending IB-ML access messages on InfiniBand, it is unimportant whether there is a physical IB-ML or not, as long as the messages are accepted and responded to correctly.

A device may respond to Baseboard Management messages for IB-ML accesses without implementing a physical IB-ML. This specification refers to such a device as having a Virtual IB-ML.

13.2.3.2.3 MULTI-PORTED IB-MODULES

Using IB-ML, a Managed InfiniBand Chassis provides some Slot-specific Information (SlotInfo) to every Backplane IB connector. Using a TCA's BMA and IB-ML, the Baseboard Manager accesses the SlotInfo provided by the Chassis. From a backplane's point of view, the SlotInfo for every slot resides at the same location. See Section 13.5.2.7.1, “SlotInfo,” on page 549.

A double-wide IB-Module connects to two adjacent backplane connectors and occupies two slots. The Chassis provides the same ChassisGUID to each slot but different SlotInfo to the two slots.

A Standard or Tall Double-wide IB-Module may connect one or more of its Backplane Connectors to multiple TCAs, to some ports of a multi-ported TCA, or to only one port of a TCA, Switch, Router, or Repeater. The IB-Module ModuleInfo describes the number of TCAs, Ports, and IB-MLs connected to its Backplane Connector. See Section 13.3.2.10, "ModuleInfo Device," on page 517.

As an example, Figure 159 depicts a two-ported IB-Module which implements two IB-MLs. It shares ModuleInfo and the MME Function Registers while it provides individual MME IbML2Ib Registers for the two IB-MLs. A four-ported IB-Module follows a similar concept.

IBTA

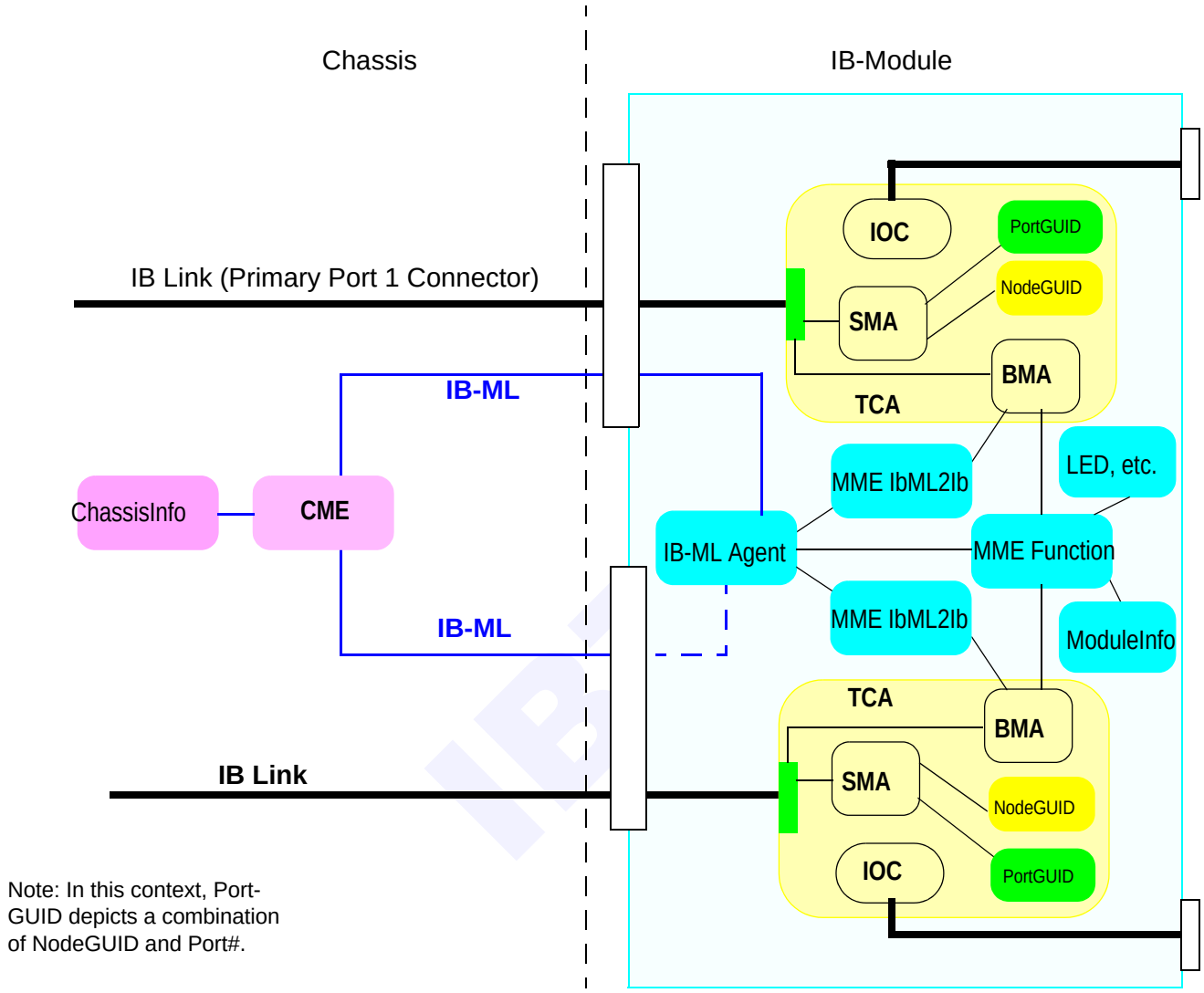


Figure 159 Multi-ported IB-Module (an example)

13.2.3.3 INTERRUPTING METHODS

Modules may detect conditions that need to inform interested entities of these conditions. InfiniBand specifies two (2) mechanisms to allow for the raising of such conditions: an in-band trap (***BMTrap***) sent over the IB fabric to the Baseboard Manager, and an out-of-band interrupt on a Back-plane Connector signal (**IMxInt_L**) to an optionally present CME. See Section 13.3.2.7, "BMTrap," on page 514 and Section 13.3.2.8, "IMxInt_L," on page 515 for feature descriptions.

13.2.3.4 INTER-METHOD OPERATIONS

Baseboard management messages may contain requests for an IB-ML access. The BMA maps these messages into a Write/Read IB-ML transaction. Some of these transactions will return data from IB-ML in the Baseboard Management response message sent back via IB. Other transactions may be split-transactions that cause an IB-ML device to later return a separate IB-ML response back to IB. See Section 13.6.1, “IB-to-MME (Baseboard MAD) Commands,” on page 549.

13.2.3.5 CLASSES OF MANAGEMENT TRANSACTIONS

InfiniBand Hardware Management recognizes the following classes of transactions:

1) In-band, directed at a Module in a Chassis
2) In-band, directed at a non-Module device supporting a BMA in a Chassis
3) In-band, directed at the CME (The CME has a direct IB connection.)
4) Out-of-band, from a CME, directed at a Module (The Chassis or CME accesses the Module via IB-ML.)
5) Out-of-band, from a CME, directed at a non-Module device supporting IB-ML
6) Out-of-band, from a Module, directed at the CME (An entity on the Module accesses the CME via IB-ML. Typically, this access will be via IB using an IB to IB-ML access.)
7) Out-of-band from a CME to IB (The CME has a physical or virtual IB-ML to IB connection).

13.2.4 MODULE MANAGEMENT

Functions such as IB Module VPD data are modeled as physical devices that are accessed via IB-ML. Common functions, such as VPD, are put at fixed “well-known” addresses on IB-ML.

IB-ML access from IB is accomplished via a single “IB to IB-ML” (IB2IBML) baseboard management message. Encapsulated within that message are parameters that can be used to direct a low-level transaction to a physical IB-ML. This is done to allow separation of the entity that performs IB to IB-ML operations from the functions on IB-ML, as follows.

The device that interprets the IB2IBML message can perform the request transaction without caring about what operation the transaction is performing on IB-ML. This makes it possible to create a device that does the IB to IB-ML operation independent from the functions being accessed on IB-ML. Such an implementation can allow new functions to be added to the IB-ML without impacting the IB to IB-ML device.

Alternatively, a device may implement IB to IB-ML with a virtual IB-ML that is either contained within the device, or is implemented using proprietary interfaces. For example, a channel adapter may be designed to provide VPD by implementing a proprietary non-volatile storage mechanism. As long as the IB baseboard management messages are responded to correctly, the implementation will be transparent to management software accessing the device via IB.

The IB-Module presents its Temperature Sensor and SEEPROM information to the CME at IB-specified addresses.

13.2.4.1 TCA MODULE

Figure 160 depicts a manageable single ported TCA Module. As described in Section 13.2.1, "Management Model," on page 478, in-band access to the Module Management Entity (MME) occurs through the InfiniBand link and the BMA. Out-of-band access to the MME occurs through the IB-ML Agent. Neither the logical partitioning of the TCA, BMA, MME and IB-ML Agent and IOCs nor the interfaces between any of them are specified.

The Module provides an interrupt line to the Chassis in the form of the Backplane Connector signal **IMxInt_L**. All sources of interrupts (power faults, thermal warnings, Hot Removal operations, etc.) on the Module make contribution to this interrupt. See Section 13.3.2.8, "IMxInt_L," on page 515 for a function description of this signal and Table 140 MME Function Registers on page 584 for the definition of the interrupt sources.

In response to Environmental events or Graceful Hot Removal Requests, TCA Modules may send traps to the Baseboard Manager using a Baseboard Management Trap (**BMTrap**). Additionally, a mechanism is specified to allow a CME to trigger these traps with an IB-ML command to a protocol-aware Module. See Section 13.3.2.7, "BMTrap," on page 514.

TCA Modules include the LEDs defined in Section 13.3.2.6, "Module Indicators (LEDs)," on page 508.

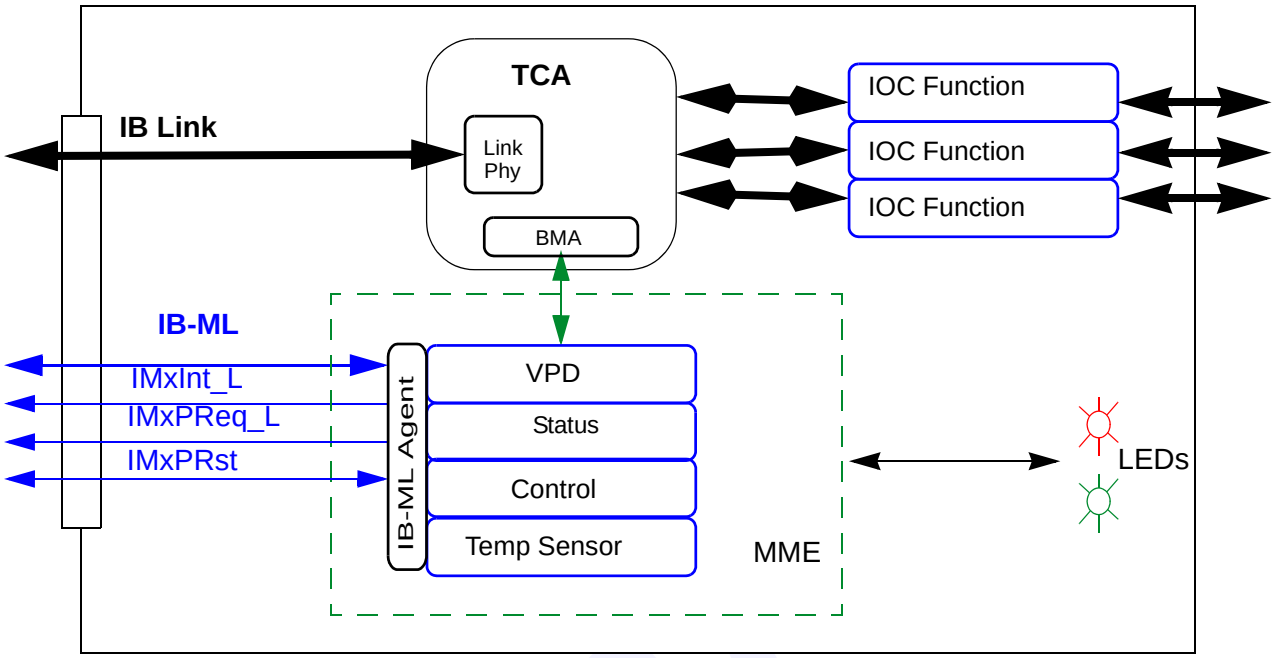


Figure 160 Manageable TCA Module

13.2.4.2 RETIMING REPEATER MODULE

A Repeater Module provides a means for retiming the InfiniBand Link but does not interpret the symbol stream other than as defined in Section 5.10, "Retiming Repeaters," on page 161.

Figure 161 depicts a manageable Repeater Module.

Since Repeaters are not protocol-aware, they are not addressable from the fabric. Thus, direct in-band access to the Module is not possible; a proxy must be provided. Out-of-band access to the MME is provided via the IB-ML through the IB-ML Agent.

A Repeater Module provides an interrupt line to the Chassis in the form of the Backplane Connector signal **IMxInt_L**. All sources of interrupts applicable to this type of Module (power faults, thermal warnings, etc.) make a contribution to this interrupt. See Section 13.3.2.8, "IMxInt_L," on page 515 for a function description of this signal and Table 140 MME Function Registers on page 584 for the definition of the interrupt sources.

As a repeater is not protocol-aware, **BMTraps** are not generated from a Repeater Module. The mechanism that allow a CME to trigger traps with an IB-ML command is not provided either. Rather, a proxy must be used on behalf of a Repeater Module.

Repeater Modules include the same LEDs as other Modules to provide a consistent user interface. See Section 13.3.2.6, “Module Indicators (LEDs),” on page 508.

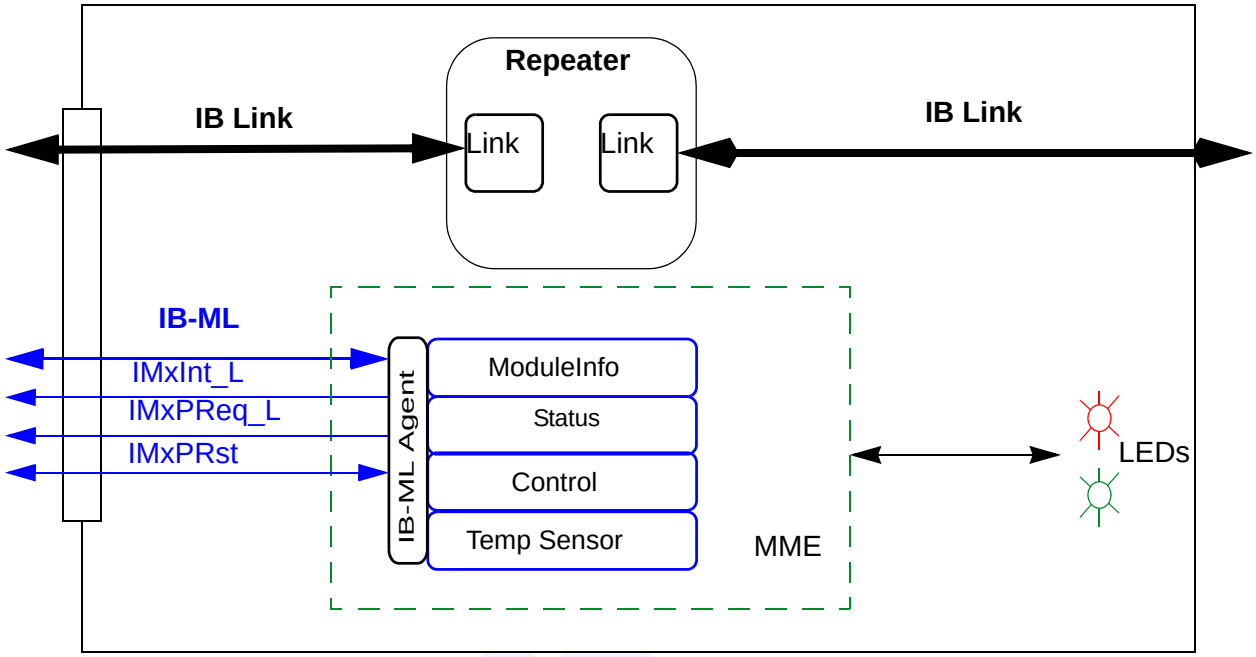


Figure 161 Manageable Retiming Repeater Module

13.2.5 NON-MODULE DEVICE MANAGEMENT

A protocol-aware device is considered “Non-Module” if it is not packaged in an InfiniBand module form factor.

A Non-Module Switch_ID is a Switch that is not on an InfiniBand Module and provides ChassisGUID. A Non-Module xCA_ID is an xCA that is not on an InfiniBand Module and provides ChassisGUID. A Non-Module Switch_NID is a Switch that is not on an InfiniBand Module and does not provide ChassisGUID. A Non-Module xCA_NID is an xCA that is not on an InfiniBand Module and does not provide ChassisGUID.

The Baseboard Manager may use ReadVPD to assess the physical location of a Non-Module device through ChassisInfo.

A Non-Module device only provides a subset of Baseboard information as specified in Section 13.3.2.10, “ModuleInfo Device,” on page 517 and Section 13.7.7, “ChassisInfo Record,” on page 612.

13.2.5.1 TCA

Non-Module TCAs perform InfiniBand operations on the link, but are not part of an InfiniBand defined Module form factor. They are typically mounted to be permanently associated with a given Switch port. These types of TCAs follow the same in-band programming model as the TCA Modules so that they may be managed in the same manner. Their out-of-band interface is not specified; if one exists, IB-ML may be used.

13.2.5.2 SWITCH

Non-Module Switches perform InfiniBand defined switch operations, but are not part of an InfiniBand defined Module form factor. Being a protocol-aware device, a Switch provides a BMA; Baseboard Management MADs may be directed to a Switch in-band. The out-of-band interface is not specified; however, IB-ML may be used.

A form of non-Module Switch is one that may be mounted within a Chassis and provide a number of ports to slots which accept InfiniBand Modules. Such a Switch may directly provide IB-ML proxy functions for the slots it supports or it may interface or emulate a CME. See Section 13.5.2.6, "IB-ML Selector Proxy," on page 548

13.2.6 CHASSIS MANAGEMENT

Within the scope of InfiniBand Hardware Management, a Chassis is the collection of InfiniBand Modules and their associated power and cooling resources housed within a single mechanical package.

A Chassis can be Unmanaged, Passively Managed, or Actively Managed.

13.2.6.1 ACTIVELY MANAGED CHASSIS

An ***Actively Managed*** Chassis provides an InfiniBand specified GUID and physical Slot Information to every InfiniBand Module on the Module's unique IB-ML. In addition, it provides a Chassis Management Entity (CME) on at least one InfiniBand Module's IB-ML. In an actively managed Chassis, the Slot Information (SlotInfo) for every Module provides a field that identifies all of the Slot(s) that may access the CME.

Figure 162 depicts a system with an actively managed Chassis. This system may be a server (including a CPU, memory complex, and an HCA) or a Box of Slots. Several IB-Modules plug into the Chassis backplane. The Switch may reside on an IB-Module or on the Chassis backplane.

A system with an actively managed Chassis includes a Chassis Management Entity (CME). The CME has direct connections via the IB Management Link (IB-ML) to at least one Module and optionally to all Modules and to the Switch. The CME and the Baseboard Manager use these links to

move management data and commands between the CME and IB-Modules and Switches.

The CME is a proxy for Chassis specific elements (power, thermal, security, etc.). A CME may optionally support a private management link to connect to another Chassis. This private link is outside of the scope of InfiniBand specification. The CME may provide a proxy for communication between Modules (See Section 13.5.2.6, "IB-ML Selector Proxy," on page 548).

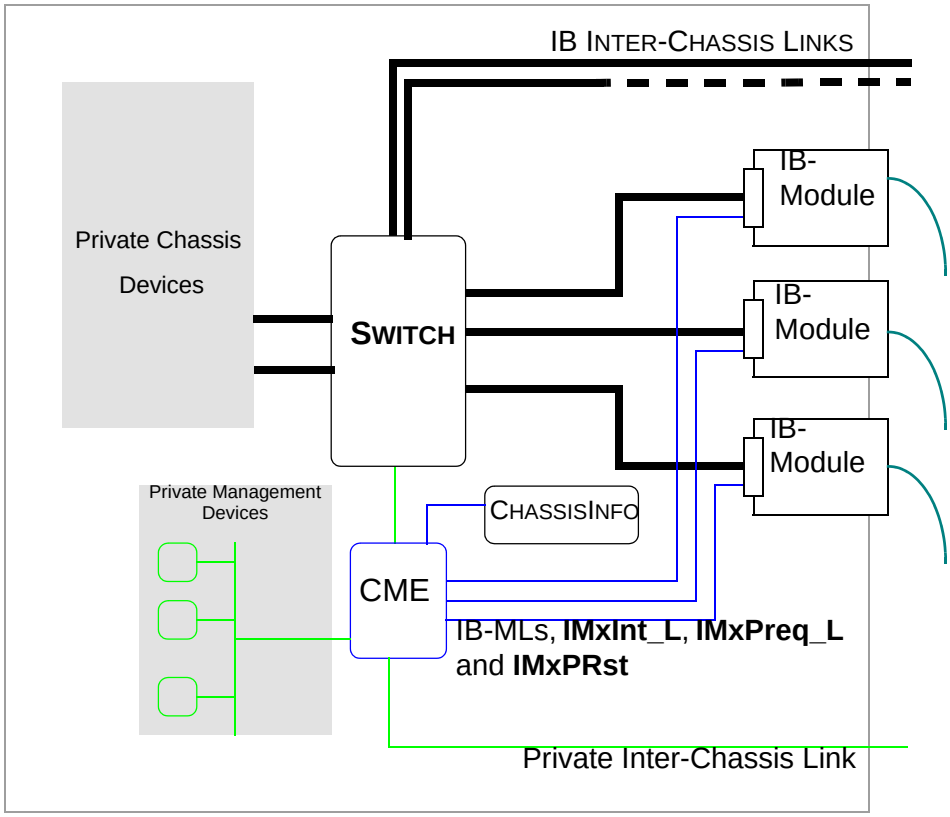


Figure 162 Actively Managed Chassis

The CME may maintain management information for the Chassis as a whole and may store this data in non-volatile storage located on a private management bus. The CME may use this private management bus for sensors and other Chassis management devices.

The CME may optionally connect to the Power Enable (**VBxEn_L**) signals on the InfiniBand backplane connectors (not shown in Figure 162). The Power Enable signals allow the CME to power-up or power-down IB-Mod-

ules. See Chapter 4: Port Signal Definitions for a description of the **VBxEn_L** signals.

IB-Modules contain their own management data. These data are available to the CME via IB-ML and to agents on the InfiniBand fabric via IB.

Table 129 Chassis Feature Requirements on page 544 identifies the optional and required features of an actively managed Chassis.

13.2.6.2 PASSIVELY MANAGED CHASSIS

A ***Passively Managed*** Chassis provides an InfiniBand specified ChassisGUID and physical Slot Information to every InfiniBand Module on the Module's unique IB-ML. See Section 13.2.6.4, "Globally Unique Identifiers," on page 492 and Section 13.5.2.7, "ChassisInfo Device," on page 548.

Figure 163 shows an example of a passively managed Chassis. The ChassisInfo blocks depict the ChassisGUID, the ChassisInfo, and the Slot-specific Information (SlotInfo). These ChassisInfo blocks are separate from their ModuleInfo counterparts on IB-Modules.

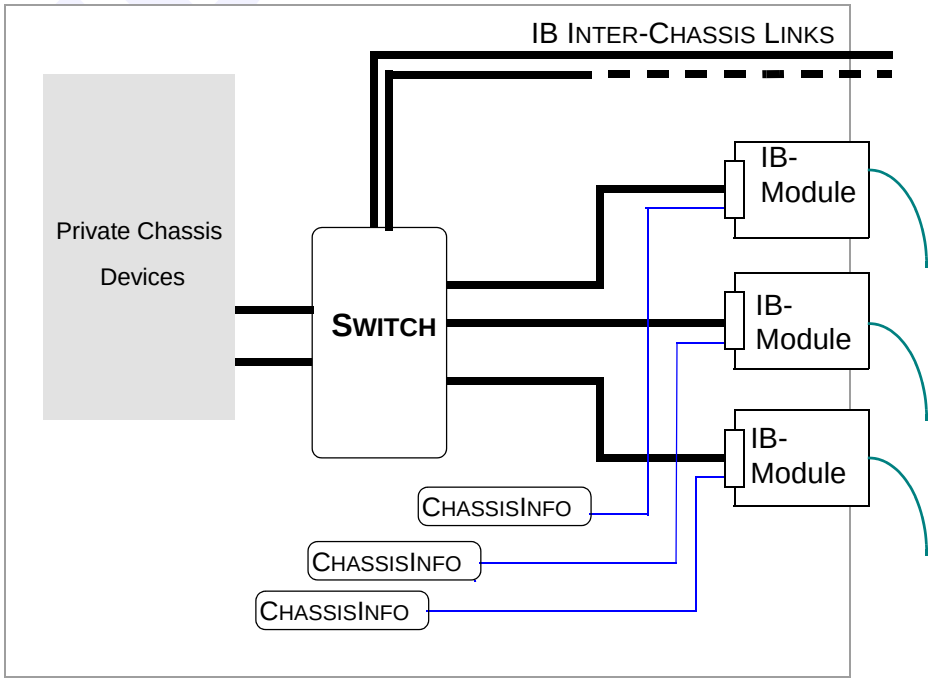


Figure 163 Passively Managed Chassis

Table 129 Chassis Feature Requirements on page 544 identifies the optional and required features of a passively managed Chassis.

13.2.6.3 UNMANAGED CHASSIS

An ***Unmanaged*** Chassis does not implement any IB-specified GUID or Slot Information and leaves Module IB-ML unconnected.

13.2.6.4 GLOBALLY UNIQUE IDENTIFIERS

For the purposes of Hardware Management, two EUI-64 compliant Globally Unique Identifiers (GUIDs) are defined for the physical elements of an IB-Module and a Chassis. These GUIDs allow a Baseboard Manager to create a logical to physical mapping of the elements that comprise an InfiniBand fabric. These identifiers are:

- **ModuleGUID**: associated with the physical entity on which an xCA, Switch or Repeater resides
- **ChassisGUID**: associated with the physical entity into which a Module is plugged

If an xCA or Switch is permanently affixed to a Chassis (i.e. Non-Module), the ModuleGUID and ChassisGUID may be the same.

When used in conjunction with the Globally Unique Identifiers defined in *InfiniBand Architecture Specification, Volume 1*, Chapter “IBA Addressing”, the mappings created with these GUIDs allow a Baseboard Manager to determine physical relationship of fabric elements for operations where actions to one element may affect other elements. Such operations include physical Module removal with multiple xCAs present, power management with multiple xCAs present, repeater management, and partitioning. Figure 164 depicts the logical view of these GUIDs for an

IB-Module and its associated Chassis. (In this context, PortGUID depicts a combination of NodeGUID and Port#).

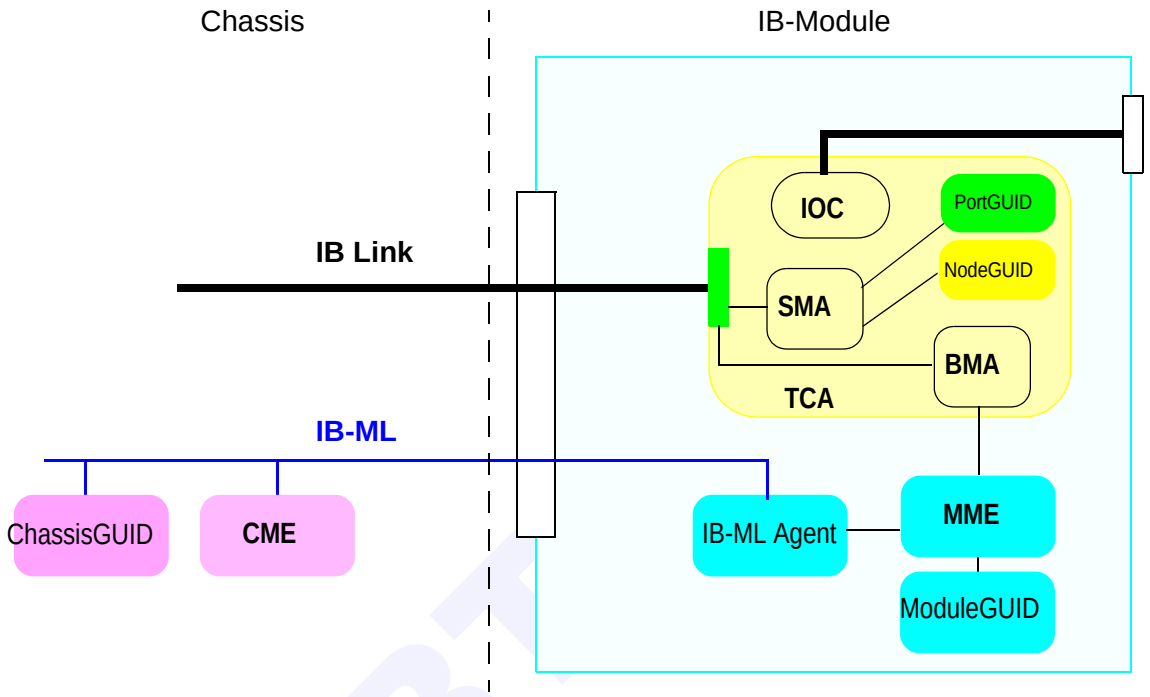


Figure 164 GUID Hierarchy

13.3 MODULE FEATURES

13.3.1 MODULE FEATURE REQUIREMENTS

This section defines the required and optional features for various types of InfiniBand devices and physical packages.

The column cells of Table 120 **shall** have the following definition:

- **Req** - A Module **shall** implement the facility labeled **Req** in the "Requirement" column.
- **Rec** - A Module **may** implement the feature labeled **Rec** in the "Requirement" column. It is recommended that the feature be provided. If the feature is provided, the Module **shall** follow the provisions of the feature description section.
- **Opt** - A Module **may** implement the features labeled **Opt** in the "Requirement" column. If the feature is provided, the Module **shall** follow the provisions of the feature description section.
- **N/A** - The feature is not applicable to the package being addressed.

C13-2: An InfiniBand defined Module containing an InfiniBand protocol-aware device **shall** provide those features and the dependent facilities labeled as "**Req**" in the "Protocol-aware Module" column of Table 120.

C13-3: An InfiniBand defined Module not containing an InfiniBand protocol-aware device **shall** provide those features and the dependent facilities labeled as "**Req**" in the "Non-Protocol-aware Module" column of Table 120.

C13-4: An InfiniBand Channel Adapter device mounted on a package other than an InfiniBand Module **shall** provide those features and the dependent facilities labeled as "**Req**" in the "Non-Module xCA" column of Table 120.

C13-5: An InfiniBand Switch device mounted on a package other than an InfiniBand Module **shall** provide those features and the dependent facilities labeled as "**Req**" in the "Non-Module Switch" column of Table 120.

Table 120 Module Feature Requirements

Feature	Protocol-aware Module	Non-Protocol-aware Module	Non-Module xCA	Non-Module Switch
BMA	**Req**	N/A	**Req**[a]	**Req**[a]
IB-ML (physical)	**Req**	**Req**	Opt	Opt
Module Management Entity (MME) (via IB)	**Req**	N/A	**Req**[a]	**Req**[a]
MME Function Registers	**Req**[b]	**Req**[b]	Opt	Opt
MME IbML2Ib Registers	**Req**[b]	N/A	Opt	Opt
MME Commands	Opt	N/A	Opt	Opt
Graceful Hot Removal	**Req**	**Req**[c]	Opt	Opt
P_Key, Q_Key, and B_Key	**Req**	N/A	**Req**[d]	**Req**[d]
Module Indicators (LEDs)	**Req**	**Req**	Opt	Opt
BMTrap	**Req**	N/A	Opt	Opt
IMxInt_L	**Req**	**Req**	Opt	Opt
Presence Detection	**Req**	**Req**	Opt	Opt
ModuleInfo Device	**Req**	**Req**	Opt [e f]	Opt [g f]
ChassisInfo Device	N/A	N/A	**Req** [i h]	**Req** [i j]

a. Required to support the "ReadVPD" and "WriteVPD" Baseboard MAD command, especially when packaged on a removable form factor. To provide a logical-to-physical association, an assembly **shall** provide ModuleGUID and ChassisGUID.
b. Access to this feature can be implemented either using a register interface, or using MME Commands.
c. Only the **CME_RTR** and **CME_CTR** facilities, accessed via IB-ML, are required. The **MME_CTR** facility, accessed via IB-ML, is optional. The **SW_RTR** and **SW_CTR** facilities are not applicable.
d. Required since BMA is required. See Section 13.3.2.10, "ModuleInfo Device," on page 517 for detailed field requirements.
e. Optional for non-Module xCA_ID only if ChassisInfo Device is present. Required for non-Module xCA_ID if ChassisInfo Device is not present.
f. Instead of implementing a ModuleInfo Device, the ModuleInfo may reside in the provided ChassisInfo Device.
g. Optional for non-Module Switch_ID only if chassisInfo Device is present. Required for non-Module Switch_ID if ChassisInfo Device is not present.
h. Required for non-Module xCA_ID if ModuleInfo Device is not present.
i. Via a connection (such as IB-ML) to the Chassis that this Unit attaches.
j. Required for non-Module Switch_ID if ModuleInfo Device is not present.

13.3.2 MODULE FEATURE DESCRIPTIONS

13.3.2.1 BMA

The Baseboard Management Agent (BMA) is a logical entity that accepts, responds to, and potentially generates Baseboard Management MADs as defined in *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "Baseboard Management" and interfaces with the logical MME (See Section 13.3.2.3, "Module Management Entity (MME)," on page 497) to perform the provided MAD commands.

13.3.2.2 IB-ML

IB-ML is a multi-drop, multi-master, two-wire serial bus which uses signaling and arbitration protocols similar to that of the SMBus 1.1 bus [24]. It primarily allows for access to defined facilities on the Module from the Chassis, but it also allows for certain defined operations to be sourced from the Module to the Chassis.

IB-ML is architecturally considered "point-point" in that the address space from one IB-ML port is not shared with the address space of any other port. This allows for the same device address assignments to be used on each Module without having to resort to dynamic address assignment techniques.

An IB-ML is made up of 2 Backplane Connector signals: **IMxClk** and **IMxDat**. See Section 11.5.1, "IMxClk, IMxDat," on page 448 and Annex A2: IB-ML Design Guidelines for the electrical and functional details of the IB-ML signals.

C13-6: An IB-Module **shall** implement an IB-ML on its Primary Backplane Connector (Port 1) (See Section 12.5, "Chassis Power Rules," on page 470 for the definition of the Primary connector.). It **may** implement

IB-MLs on other available Backplane Connectors; if so, it **shall** keep the multiple IB-MLs electrically isolated from each other.

Logical slave devices on the IB-ML allow access to the ModuleInfo (See Section 13.3.2.10, "ModuleInfo Device," on page 517) and the MME Registers (See Section 13.6.5, "MME Function Registers," on page 581).

C13-7: All Modules **shall** respond as IB-ML slaves for Read and Write access to the device addresses labeled as "**Req**" in the "Access" column of Table 121.

C13-8: A Module **shall not** respond as an IB-ML slave for Read and Write access to the device addresses labeled as Not Permitted "NP" in the "Access" column of Table 121.

o13-1: Modules **may** respond as IB-ML slaves for Read and Write access to the following device addresses labeled as "Opt" in the "Access" column of Table 121.

Table 121 Module IB-ML Slave Addresses

Function	Device Address[a]		Access	Reference
	Hex	Binary		
ModuleInfo Device	A0h	1010_000b	**Req**	Section 13.3.2.10 on page 517
Reserved for future IB definition	A2h	1010_001b	NP	
Reserved for future IB definition	A4h	1010_010b	NP	
Module Specific Use	A6h	1010_011b	Opt	
Module Specific Use	ACh	1010_110b	Opt	
Module Specific Use	AEh	1010_111b	Opt	
ChassisInfo Device	A8h	1010_100b	NP[b]	Section 13.5.2.7 on page 548
Chassis Specific Use	AAh	1010_101b	NP	
CME Slave	E8h	1110_100b	NP	Section 13.5.2.3.1 on page 547
MME Function	E0h	1110_000b	**Req**[c]	Section 13.6.5 on page 581
MME IbML2Ib	E2h	1110_001b	**Req**[c]	Section 13.6.5 on page 581
MME Command	E4h	1110_010b	Opt	Section 13.3.2.3.3 on page 498

a. This 7-bit Binary notation for IB-ML slave addresses leaves out the Write/Read direction bit which is encoded as the least significant bit on the IB-ML bus during the address phase. The accompanied Hex notation follows this Binary notation by left justifying the 7-bit IB-ML address to fit into an 8-bit wide byte and setting the list significant bit to 0.
b. "NP" for this Chassis device address means that the Module, **as a slave, shall not** (not permitted to) implement the resources associated with these addresses. It means that the slave device for this address is present outside of the Module. It **does not** mean that accesses to these addresses using the "ReadVPD"Baseboard MAD command to this address are denied if a physical or virtual IB-ML exists.
c. Access to this feature can be implemented either using a register interface, or using MME Commands. This table lists the required slave address to be used if the feature is implemented using the register interface.

Table 122 shows the only other slave addresses that InfiniBand Modules may use.

Table 122 Other Allowed Addresses for InfiniBand Modules

Slave Address	8-bit Hex notation	Function
0100 000 - 0100 011	40h-46h	Module-specific use. Typically used for 8574-8-bit and 8575-type 16-bit I^2C latches.
0101 001 - 0101 111	50h-5Eh	Module-specific use. LM78/ADM1024 and other "Heceta"-type monitoring devices.
1001 000 - 1001 111	90h-9Eh	Module-specific use. Typically used for LM75-style temperature sensors.
1011 110 - 1011 111	BCh-BEh	Module-specific use.
1100 001	C2h	SMBus Device Default. Module Devices that use this address must do so per the SMBus 2.0 specification.
1101 110 - 1101 111	DCh-DEh	Module-specific use.
1110 011	E6h	Module-specific use. Often used for PCA9544-style I^2C bus multiplexer.

13.3.2.3 MODULE MANAGEMENT ENTITY (MME)

The MME provides a logical view of the applicable specified Module facilities for access in-band via the BMA and Baseboard Management MADs and out-of-band via IB-ML.

C13-9: All Modules, at a minimum, **shall** implement a logical Module Management Entity that implements the required elements of this specification for Baseboard Management and IB-ML operations.

BETA

Module designs are free to choose any implementation mechanism; these may consist of, but are not limited to: discrete components, ASICs, micro-controllers, or Integration with xCA components.

Implementation Note

For example, the ModuleInfo portion of the MME may be implemented as a discrete SEEPROM device directly on IB-ML, be emulated as part of a micro-controller, or be included as part of the same chip containing the xCA attached to IB-ML.

13.3.2.3.1 MME FUNCTION REGISTERS

The MME Function Registers, described in Section 13.6.6, "MME Function Register Summary," on page 583, provide facilities for the CME to access Module functions via IB-ML and for the Baseboard Manager to access the Module functions via IB and via BMA. See also Figure 159 Multi-ported IB-Module (an example) on page 484.

The MME Function Registers can be implemented in two ways: As 'physical' registers that are written and read using IB-ML 'register' transactions via address E0h, or as 'logical' registers that are written and read using the WriteMMERegister and ReadMMERegister MME Commands using a 'split-transaction' via address E4h. In either case, Table 140 on page 584 is used to specify the register selector and data bytes for the MME Function Registers.

13.3.2.3.2 MME IBML2IB REGISTERS

The MME IbML2Ib Registers, described in Section 13.6.9, "MME IbML2Ib Register Descriptions," on page 588, provide facilities for the CME to send messages and traps to the Baseboard Management LID. See also Figure 159 Multi-ported IB-Module (an example) on page 484.

The MME IbML2Ib Registers can be implemented in two ways: As 'physical' SendBMTrap and SendIbML2BM registers that are written using IB-ML 'register' transactions via address E0h, or as split-transaction SendBMTrap and SendIbML2BM MME Commands via address E4h. Table 141 on page 588 specifies the data format for the physical 'E0h' register access, while Table 123 on page 499 provides the format for the MME command.

13.3.2.3.3 MME COMMANDS

MME commands are split-transaction IB-ML request and response transactions, as described in section 13.6.1.4 IB-ML Split-transaction Messaging on page 561. MME commands are part of the InfiniBand Hardware Management command set (Section 13.6.1.7, "IB-ML CMD Set Specific Fields," on page 564). The MME performs the requested operation, formats a response using the sequence number that was passed in the re-

quest. MME Commands are transferred on IB-ML via the MME Command address specified in Table 121 on page 496.

For the standard MME commands (commands other than OEM CMD messages), the Source Device ID in the request determines where the MME delivers the response to the commands. The response **shall** be sent as an IB-ML split-transaction response to the slave address specified by the Source Device ID in the request.

o13-2: If an entity implements an MME with the features described in Table 123, it shall follow the definitions of Section 13.3.2.3.3, "MME Commands," on page 498.

Table 123 defines the split-transaction commands for CMD Set = IB-ML sent to the MME Slave Address. See Section 13.6.1.7, "IB-ML CMD Set Specific Fields," on page 564 for CMD field placement in the split-transaction message.

Table 123 MME Commands

Command	CMD	Requirement	Request data	Response data
GetCMDSetVersion (13.3.2.3.4)	00h	**Req**	none	byte 1: completion status[a] byte 2: Version of MME command set (CMD Set = IB-ML) that this controller supports. BCD encoded. Initial value = 10h for CMD Set version 1.0
GetVendorID (13.3.2.3.5)	01h	Opt[b]	none	byte 1: completion status[a] bytes 2+ - A series of one or more 3-byte IDs for each vendor or organization that have defined controller-specific (OEM) commands implemented by this MME. FFFFFFh if no controller-specific commands. 3-byte ID uses same format as the vendor ID field in ChassisGUID.
Reserved	02h			
ReadVPD (13.3.2.3.6)	03h	Opt[c]	byte 1 - VPD Device Selector (address) byte 2 - VPD number of bytes to read (1-based. 0 count is allowed) bytes 3:4 - VPD offset to read (2 bytes)	byte 1: completion status[a] bytes 2+ - read data bytes (if any)

Table 123 MME Commands

Command	CMD	Requirement	Request data	Response data
WriteVPD (13.3.2.3.7)	04h	Opt[c]	byte 1 - VPD Device Selector (address) byte 2 - # bytes to write (2-bytes, 1-based. 0 count is allowed) byte 3:4 - VPD offset to read-write bytes 5+: - write data (N bytes)	byte 1: completion status[a.]
OEM (13.3.2.3.8)	05h	Opt	bytes 1-3: vendor ID bytes 4+: vendor-specific data	byte 1: completion status[a.] bytes 2-4: vendor ID bytes 5+: vendor-specific response data
Reserved	06h-0Fh			
WriteMMERegister (13.3.2.3.9)	10h	Opt[d]	byte 1 - register selector per Table 140 bytes 2+ - write data per Table 140	byte 1: completion status[a.]
ReadMMERegister (13.3.2.3.10)	11h	Opt[d]	byte 1 - register selector per Table 140	byte 1: completion status[a.] bytes 2+ - read data corresponding to bytes 1+ per Table 140
Reserved	12h-1Fh			
SendBMTrap (13.3.2.3.11)	20h	Opt[e]	byte 1 - BMTrapDataLength = count of following bytes, per Table 141 byte 2 - BMTrapType per Table 141 byte 3:5 - BMTrapTypeModifier varies based on trap type per Table 141 byte 6 + - BMTrapData per Table 141.	byte 1: completion status[a.] Includes command-specific status per Table 141
SendIbML2BM (13.3.2.3.12)	21h	Opt[e]	byte 1 - BMDataLength per Table 141 byte 2+ - BMData per Table 141	byte 1: completion status[a.] Includes command-specific status per Table 141

a. The completion status field uses the values defined in Table 135 Completion Status Values on page 565. Additional completion status values that pertain to a given command, if any, are enumerated in the Response Data for the command.
b. Required if Controller-specific commands are implemented by the MME. (If you implement any controller-specific commands, then you're required to implement the GetVendorID command.)
c. Required if the ModuleInfo Device or ChassisInfo Device features are required per Table 120 Module Feature Requirements on page 494, but it is not implemented via address A0h.
d. Required if the MME Function feature is required per Table 120, but it is not implemented via address E0h.
e. Required if the MME IbML2Ib feature is required per Table 120, but is not implemented via address E2h.

13.3.2.3.4 GETCMDSETVERSION

GETCMDSETVERSION is an IB-ML split-transaction command that returns the version of the IB-ML command set (CMD Set = IB-ML) implemented by the MME. The version information is stored in BCD (binary coded decimal) format, where the most-significant nibble holds the major version number and the least-significant nibble holds the minor version number. For example, a value of 09h corresponds to IB-ML command set version 0.9.

13.3.2.3.5 GETVENDORID

GetVendorID command is an IB-ML split-transaction command that returns the ID for the vendor that has defined controller-specific (CMD Set = OEM) messages for the MME. The vendor ID format is the same as that used in the ChassisGUID. If no CMD Set = OEM messages are defined, a value of 00_00_00h **shall** be returned in the response. Note that it is possible for more than one set of vendor-specific commands to be implemented on an MME.

13.3.2.3.6 READVPD

ReadVPD command is used to retrieve data from the logical VPD Device that is selected by the “Device Selector” field. If a physical ModuleInfo Device is not provided at address IB-ML address A0h, this command **shall** be used to provide a logical VPD Device that is accessed using a “Device Selector” field value of A0h. The amount of data that can be read is limited by the overall IB-ML command length limit of 36-bytes from Slave address through PEC, inclusive.

Architectural Note

The ***ReadVPD*** Number of Bytes to Read (NBR) field is defined to be 2 bytes so as to 1) allow for larger data abstraction implementations that may not use physical IB-ML devices that are limited to 256B of address ability and 2) to accommodate emerging physical devices that allow for 16KB of address ability.

13.3.2.3.7 WRITEVPD

WriteVPD command is used to write data to the logical VPD Device that is selected by the “Device Selector” field. If a physical ModuleInfo Device is not provided at address IB-ML address A0h, this command **shall** be used to provide a logical VPD Device that is accessed using a “Device Selector” field value of A0h. The amount of data that can be written is limited

by the overall IB-ML command length limit of 36-bytes from Slave address through PEC, inclusive.

Architectural Note

The ***WriteVPD*** Number of Bytes to Write (NBW) field is defined to be 2 bytes so as to 1) allow for larger data abstraction implementations that may not use physical IB-ML devices that are limited to 256B of address ability and 2) to accommodate emerging physical devices that allow for 16KB of addressablity.

13.3.2.3.8 OEM

This IB-ML split-transaction message allows a vendor or organization to define additional MME functions. The message includes a vendor-ID field that helps ensure that data parameter values specified by one vendor or organization do not interfere or need to be reconciled with the values selected by another vendor or organization. All other data fields are specified by the vendor or organization identified by the vendor-ID field.

The OEM MME command specified here is within the IB-ML Command Set (Table 123: MME Commands) and **shall** utilize the same message fields (Destination Device ID, Source Device ID, Rs, CMD Set, Length, etc.) as the other standard MME commands specified in this section.

13.3.2.3.9 WRITEMMEREGISTER

WriteMMERegister command provides the mechanism to write to a given logical MME Function Register. The “register selector” value and “write data” field contents are obtained from the “Register (selector)” and “Write Data” columns in Table 140 MME Function Registers on page 584, respectively.

13.3.2.3.10 READMMEREGISTER

ReadMMERegister command provides the mechanism to write to a given logical MME Function Register. The “register selector” value and “read data” field contents are obtained from the “Register (selector)” and “Read Data” columns in Table 140 MME Function Registers on page 584, respectively.

13.3.2.3.11 SENDBMTRAP

SendBMTrap command provides the mechanism to direct the MME to send a BMTrap method message to the Baseboard Management LID. The Module uses the components of the BM.ClassPortInfo and the “BMTrap Type”, “BMTrapTypeModifier” and “IBMTrapData” fields (See SendBMTrap in Table 123 on page 499) to form the datagram message as defined in InfiniBand Architecture Specification, Volume 1, Chapter “General Services”, Section “Baseboard Management”. The size of the BMTrapData field is limited by the requirement that MME commands be

limited to an overall length of 36-bytes from slave address through PEC, inclusive.

Architecture Note

This command is intended to be used by a CME to communicate with the Baseboard Manager and should be used in lieu of the *SendIBML2BM* command. A node running the Baseboard Manager can use the method in the Baseboard class MAD (BMTrap) to direct the MAD to the Baseboard Manager versus the node's BMA as a BMA is only to act upon BMSend methods with the R bit being 0 (requests). See Table 131 Common MAD Field usage for Baseboard Management on page 551.

13.3.2.3.12 SENDIBML2BM

SendIbML2BM command provides the mechanism to direct the MME to send a "BMSend" method message to the Baseboard Management LID. The Module uses the components of the BM.ClassPortInfo and the "BM-Data" fields (See SendIbML2BM in Table 123 on page 499) to form the datagram message as defined in InfiniBand Architecture Specification, Volume 1, Chapter "General Services", Section "Baseboard Management". The size of the BMData field is limited by the requirement that MME commands be limited to an overall length of 36-bytes from slave address through PEC.

This command is not recommended for use and is subject to removal in future versions of the specification. A CME should use the SendBMTrap command with "OEM CME" TrapType to communicate with the Baseboard Manager. TrapType of "Generic CME" is defined as a placeholder for potential future definition of IB2CME.

13.3.2.4 GRACEFUL HOT REMOVAL

"Graceful Hot Removal" is the removal of an IB Module that has first been placed in a quiescent state where no activity is present that would cause disruption upon link detachment. $\mathbf{V_{Bulk}}$ may or may not be on.

The quiescent state is under control of a central software agent running from across InfiniBand fabric that has access to the B_Key for a given IB Module and claims or otherwise arbitrates ownership of the Module.

In this description, the Baseboard Manager serves as the central software entity; however, depending on implementation, this central agent may be a real Baseboard Manager, be a component of some other software entity that acts as a Baseboard Manager, or be some other software entity with the Baseboard Manager acting as a proxy. Additionally, software implementations may allow for Claiming to be directly controlled by other, non-Baseboard Manager that has been given the B_Key for that IB Module.

The term “interested software entities” refers to those elements that are dependent on the presence of a given IB Module for operation.

The following facilities or bits are used for interaction with the Baseboard Manager to accomplish Graceful Hot Removal:

- **CME_RTR** (CME Request to Remove)
- **SW_RTR** (Software Request to Remove)
- **MME_RTR** (MME Request to Remove)
- **CME_CTR** (CME Clear to Remove)
- **SW_CTR** (Software Clear to Remove)
- **MME_CTR** (MME Clear to Remove)

An implementation may elect to make these bits accessible either by implementing the MME Function feature at address E0h, or by implementing the ReadMMERegister and WriteMMERegister MME commands. See Section 13.6.7, “MME Function Register Descriptions,” on page 586 for detailed bit locations.

Additionally, the ***BMTrap*** MAD mechanism is used to notify the Baseboard Manager that a CME or an MME Request to Remove is pending.

Operation

To allow for Graceful Hot Removal, it is required that software “Claim” the IB devices on the Module using the **SW_CTR** bit to indicate that software is presently using or preparing to use the Module.

At power-up, the **SW_CTR** is set to 1b. Once the **SW_CTR** bit is set to 0b, the LEDs on the IB Module indicate that it is not OK to Remove the Module (See Section 13.3.2.6, “Module Indicators (LEDs),” on page 508 for a full description).

Once an xCA is claimed, the general steps of Graceful Hot Removal are:

1) A software process (e.g. a console application), the MME on the Module, or a CME on the Chassis containing the IB device initiates a request for removal to the Baseboard Manager.
   a) For the case that a software process initiates the removal, the Baseboard Manager **shall** issue a ***BMSend.SetModuleState(SW_RTR)*** prior to collecting the necessary responses from the interested software entities. This allows for the IB Module to indicate the “transition” condition through the blinking of an LED.

b) For the case that the request is coming from a CME, the CME **shall** set **CME_RTR** and then forms and sends a "**Send BMTrap**" message on IB-ML to the IB Module. Based on the CME_RTR, the Module **shall** blink the Module Status LED.

Upon receipt of a **BMTrap** or equivalent polling means that indicates **CME_RTR** is 1b, the Baseboard Manager **shall** issue a **BMSend.SetModuleState(SW_RTR)**=1. The setting of the **SW_RTR** bit to 1b acts as an acknowledgement to the setting of **CME_RTR** through either polling or ***BMTrap*** and stops the assertion of the *CME_RTR_Trap_Timeout* facility.

Based on the setting of the **SW_RTR** bit to 1b, the Module **shall** assert the **IMxInt_L** interrupt to the CME if the CME has enabled the Removal Status Change interrupt. While awaiting **SW_RTR** bit to be set to 1b, the CME **shall** retry the ***BMTrap*** message once every $t_{BMTrapRetryPeriod}$ for $t_{BMTrapTimeout}$ Period of time. See Table 144: IB-ML Timeout Parameters.

2) The Baseboard Manager notifies all interested software entities of the removal request and determines when all such entities have released dependence on the xCA. This is termed "Claim Release".

3) For the case that not all interested software entities concur with the claim release, the Baseboard Manager **shall** issue a ***BMSend.SetModuleState(SW_RTR)=0*** to cancel the request.

   For the case that not all responses are received within a Baseboard Manager specific time-out period, the Baseboard Manager **may** issue a ***BMSend.SetModuleState(SW_RTR)=0*** to cancel the request. Failure to issue this operation will leave the Module in a transition state for an indefinite period of time; the LED state for this case indicates that the Module is not OK to Remove.

4) For the case that all interested software entities have released the IB devices on the Module, the Baseboard Manager issues a ***BMSend.SetModuleState(SW_CTR)=1***. This completes the software contribution to the Module Status LED indication for being OK to Remove.

5) If the CME had previously set *RemovalControl/Status.CME_CTR*=0, upon the CME detecting a state change of *InterruptSource.SW_CTR*, either through an interrupt or polling, CME **may** perform Chassis specific actions to prepare for removal. The CME **shall** then set *RemovalControl/Status.CME_CTR*=1 to contribute to the LED indication for being OK to Remove.

6) If the Module has implemented Module specific functions that require a contribution to the LED indication of being OK to Remove, it **should** set *RemovalControl/Status.MME_CTR=0.* Upon determining that these Module specific functions are in an appropriate state for re-

moval, it **shall** set *RemovalControl/Status.MME_CTR=1* to contribute to the LED indication for being OK to Remove.

See Section 13.4.2, "Removal Control/Status Facility Descriptions," on page 526 for a detailed description of each of the facilities used above.

C13-10: All InfiniBand Modules **shall** implement the MME and Module operations outlined in Section 13.3.2.4, "Graceful Hot Removal," on page 503.

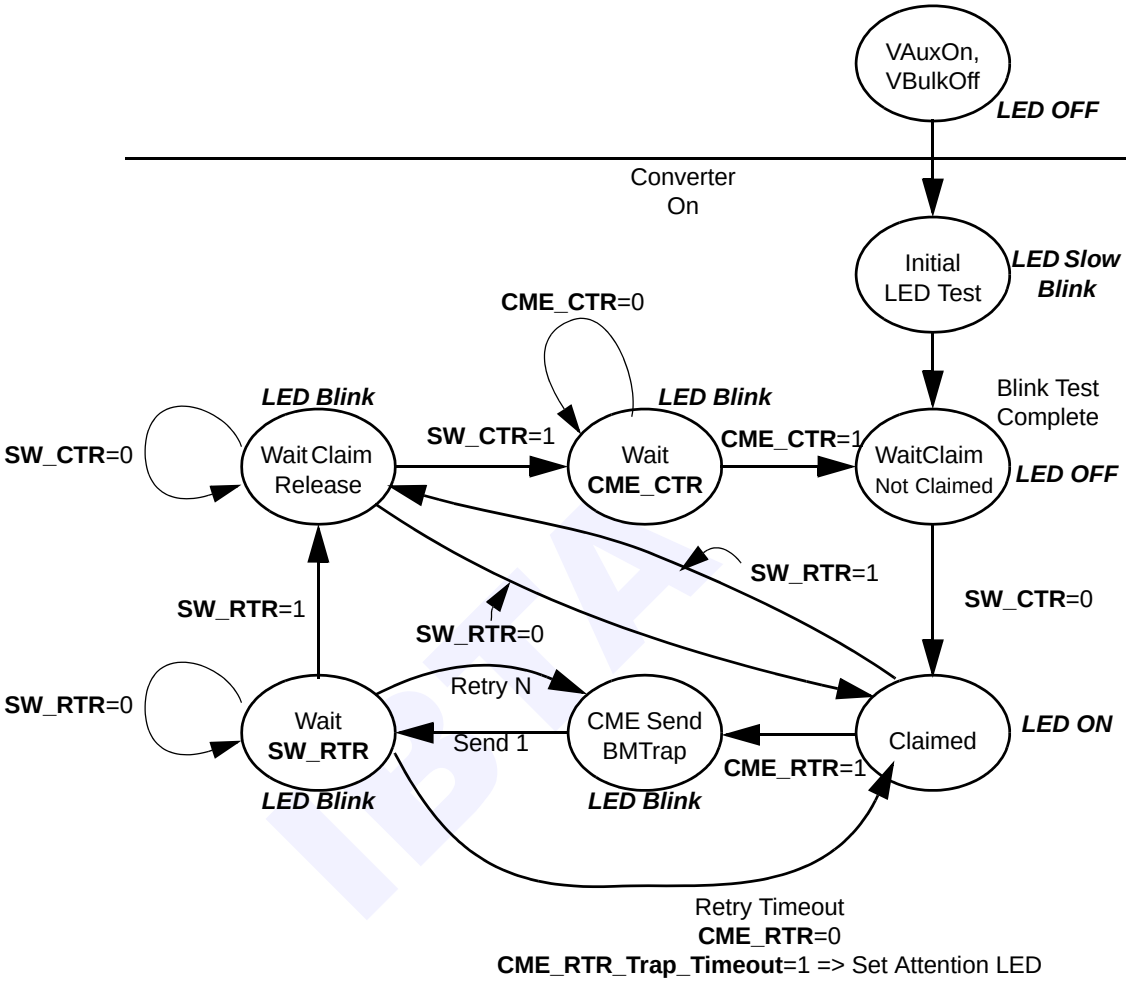
C13-11: The Baseboard Manager **shall** implement the Baseboard Management operations outlined in Section 13.3.2.4, "Graceful Hot Removal," on page 503.

o13-3: All actively managed InfiniBand Chassis that implements a Module Request-to-Remove function **shall** implement the CME operations outlined in Section 13.3.2.4, "Graceful Hot Removal," on page 503.

13.3.2.4.1 STATE DIAGRAM

The Graceful Hot Removal process is summarized in the state diagram of Figure 165 and generally depicts the behavior of the Module Status LED. (See Section 13.3.2.6.3, "Module Status (Green) LED," on page 510).

IBTA



Note:
The indicated LED is the Module Status LED.
The MME_RTR path is similar to that of CME_RTR

Figure 165 Graceful Hot Removal State Diagram

13.3.2.5 P_KEY, Q_KEY, AND B_KEY

The Baseboard Management MADs use the General Service Interface (GSI) on QP 1. P_Key and Q_Key are checked by QP1. P_Keys not in the node's partition table will cause the MAD to be discarded. GMPs destined for QP 1 whose Q_Key is not 0x80010000 are discarded. Baseboard Management class utilizes the GSI and may be forwarded across subnets. See *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "MAD Usage Model" for the details on P_Key and Q_Key used for GSIs.

For flexible Baseboard Management, xCAs reserve a P_Key for Baseboard Manager(s). If there is a P_Key available for the Baseboard Manager, devices from different data partitions can belong to one Baseboard Partition, and the Baseboard Manager can manage them regardless of data partition boundaries and therefore regardless of the Operating System or Software running on those data partitions.

The Baseboard Management Key (B_Key) provides a separate level of authentication that helps protect against receipt of bad management request messages. The Baseboard Manager includes the B_Key in the BM MAD to obtain authorization. The B_Key is used to authenticate a trusted source. Similar to the model used for the M_Key, this model assumes that the fabric has some level of physical security. While the B_Key is located in the header of the BM MAD, B_Key handling depends on whether BM MAD contains a *request* or a *response* message or is a BMTrap. See *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”, Section “B_Key Usage” for the details on B_Key.

13.3.2.6 MODULE INDICATORS (LEDS)

Two functionally independent single-colored LEDs provide visual indications to the user to show if the Module is Removable, in Removal Transition, Claimed by some software entity, or Requires Attention. An Identify Function is also provided.

C13-12: All IB-Modules **shall** implement two LEDs as defined by Section 13.3.2.6, “Module Indicators (LEDs),” on page 508 and its subsections.

o13-4: If a Non-Module InfiniBand assembly implements indicators to claim InfiniBand behavior, it **shall** implement two LEDs as defined by Section 13.3.2.6, “Module Indicators (LEDs),” on page 508 and its subsections.

C13-13: All InfiniBand Modules **shall** implement the MME and Module operations outlined in Section 13.3.2.6, “Module Indicators (LEDs),” on page 508.

o13-5: All actively managed InfiniBand Chassis that implements a Module Request-to-Remove function **shall** implement the CME operations outlined in Section 13.3.2.6, “Module Indicators (LEDs),” on page 508.

Table 124 summarizes the behavior of the two LEDs.

Table 124 Module Indicator LED Summary[a]

Green (Module Status)	Yellow or Amber (Attention)	Definition	Comment
OFF	**X**	Module is **Removable**	
Blink	**X**	Bulk power is stable Module's DC-DC converter is enabled. **Removal Transition** is in progress. Module is considered Claimed. Module is not Removable.	Blink Rate is defined in Table 125.During an IB Module Request-to-Remove negotiation, if the Module Status LED blinks and then goes OFF, the request for OK-to- Remove was successful. If the Module Status LED blinks and then returns to ON, the Request-to- Remove was rejected for some reason. For example the CME timed out waiting for the acknowledgement to CME_RTR BMTrap (**SW_RTR** stays 0b for "too long") or the system software could not relinquish the devices (**SW_CRT** stays 0b and **SW_RTR** goes back to 0b).
ON	**X**	Module is **Claimed** (normal operation or in a power managed state) Module is not removable.	Module Power Management State and the Module LEDs are independent of each other. In a Power Managed State, the Module Power may be off while the Module is still Claimed. In that case, the Module Status LED is accordingly ON.
X	**OFF**	No Attention or Identify to the Module is requested.	
X	**Blink**	**Module Identify** is active	Blink rate is defined in Table 125. The CME may invoke the Module Identify Function when Module Power is OFF. The CME may act as a proxy to invoke the Identify Function.
X	**ON**	The Module, the CME, or the S/W has requested **Attention** to the Module.	Module Power Management State and the Module LEDs are independent of each other. Due to an attention event, the Module Attention LED may be ON when the Module is in a Power Managed State or when $\mathbf{V_{Bulk}}$ is not available or when the Module's DC-DC converter is off. If the Module has power, the nature of the requested attention can be determined from status bits on the Module using the ***BMSend.GetModuleStatus*** command.
Slow Blink	**Slow Blink**	Bulk power is stable. Module's DC-DC converter is enabled. **LED Test** Module is not removable.	Slow Blink rate is defined in Table 125. After Module Insertion, the LEDs blink to provide an LED Test, and then go OFF.

a. An "X" in the Module Status and Attention LED columns of this table indicates ON, OFF, or Blink.

13.3.2.6.1 BLINK DEFINITIONS AND RATES

When used in this specification, the terms defined in Table 125 **shall** apply to the repetitive ON/OFF behavior of the LED indicators.

C13-14: An InfiniBand Module **shall** implement the LED Blink Rates as defined in Table 125.

.

Table 125 Blink Rate Definitions

Term	Frequency	Units	Period	Duty Cycle	Units	Comments
Blink	2	Hz	500ms	50 +/- 5	%	Nominally 250ms ON, 250ms OFF
Slow Blink	1/4	Hz	4s	50 +/- 10	%	Nominally 2s ON, 2s OFF

13.3.2.6.2 COLOR AND INTENSITY

The physical requirements of the LED indicators are specified in Section 9.2.8, “Module Indicators,” on page 382.

13.3.2.6.3 MODULE STATUS (GREEN) LED

The Module Status LED generally indicates the removal status of the Module and may take on three states: OFF, ON, or Blinking. For a detailed description of a Graceful Hot Removal, refer to section Section 13.3.2.4, “Graceful Hot Removal,” on page 503. It also participates in the LED Test functions. See Section 13.3.2.6.6, “LED Test,” on page 513 for a detailed description.

Operation

The Module MME causes the Module Status LED to go ON whenever the Module is “Claimed” by the Software. In this context, being “claimed by software” means that the unexpected removal of the Module might cause unexpected software behavior such as a loss of data or a software hang condition. The user is strongly encouraged not to remove the IB-Module when the Module Status LED is not OFF.

The Module MME causes the Module Status LED to start to Blink when an IB Software entity such as the Baseboard Manager sets the **SW_RTR** bit (Software Request-to-Remove) or when the CME sets the **CME_RTR**. (The CME initiates the **CME_RTR BMTrap**.)

The Module MME causes the Module Status LED to continue to Blink until the IB Software entity sets the **SW_CTR** (Software Clear-to-Remove) bit

to indicate that the removal request was granted, or until the IB Software removes the **SW_RTR** bit (SW denied the removal).

The CME uses **SW_RTR** as an acknowledgement that the Request-to-Remove BMTrap has been received. Since there is a small possibility that the Request-to-Remove BMTrap could get "lost" due to a data corruption or other error, the CME **shall** periodically re-send the Request-to-remove BMTrap while waiting for SW_RTR. (See SECTION 13.3.2.4, "GRACEFUL HOT REMOVAL," ON PAGE 503.) If the CME times out waiting for the **SW_RTR**, the CME **shall** clear the **CME_RTR** bit and **shall** set the Attention LED using the CME_RTR_Trap_Timeout status bit. While SW_RTR is 0b, in response to CME_RTR getting cleared, the Module MME **shall** cause the Module Status LED to go ON (back to the "SW claimed" state).

To indicate that it is "OK to Remove" the IB-Module, the Module MME **shall** cause the Module Status LED to go OFF when the **SW_CTR, CME_CTR,** and the **MME_CTR** are all asserted.

There might be several reasons for which a request for removal may be delayed or denied.

- The MME **may** hold **MME_CTR** negated while it is logging events or performing a shut-down operation such as flushing data buffers.
- The CME **may** hold **CME_CTR** while it performs Chassis specific operations.
- The Software **may** hold **SW_CTR** while it is in the process of bringing the Module into a quiescent state.

Control

C13-15: Using the following pseudo-code, the Module's MME **shall** implement the composite behavior of the Module Status LED with priority of facilities:

if **Initiate_LED_Test** is 1b
    the Module Status LED "slow blinks" as defined in Table 125;
else If **LED_Enable** is 0b,
    the Module Status LED is OFF;
else If **CME_Force_OKTR** is 1b,
    the Module Status LED is OFF;
else If **Module_OKTR** is 1b,
    the Module Status LED is OFF;
else If **SW_RTR**, **MME_RTR**, or **CME_RTR** is 1b, but (**SW_CTR**, **MME_CTR**, **CME_CTR**) is not 111b,
    the Module Status LED "blinks" as defined in Table 125;
else If **Module_OKTR** is 0b,
    the Module Status LED is ON;
else,

Error since all cases are supposed to have been covered by now!

Status

The status of the Module Status LED is provided through the Module_Status_LED_State facility (See Section 13.4.3.7 on page 532).

13.3.2.6.4 ATTENTION (AMBER) LED

The Attention LED actively indicates that an Attention condition has been raised for the Module. It also participates in the LED Test and Module Identify functions. See Section 13.3.2.6.6, "LED Test," on page 513 and Section 13.3.2.6.7, "Module Identify," on page 513 for detailed descriptions.

The Module MME keeps the Attention LED OFF when neither the CME, the Module, nor the IB Software has detected any problem with the Module to request attention.

The Module MME turns the Attention LED ON when the CME, the Module, or the IB Software is requesting attention to the IB-Module.

The Module MME causes the Attention LED to perform a "Blink" cycle as defined in Table 125 Blink Rate Definitions on page 510 when the Module Identify facility has been asserted. (See 13.3.2.6.7 on page 513). Upon the Module Identify being deasserted, the Attention LED returns to indicating whether an Attention condition exists or not.

Control

C13-16: Using the following pseudo-code, the Module's MME **shall** implement the composite behavior of the Attention LED with priority of facilities. The Attention LED **shall** reflect the Module's Attention state while $\mathbf{V_{Aux}}$ is available.

if **Initiate_LED_Test** is 1b
    the Attention LED "slow blinks" as defined in Table 125;
else If **LED_Enable** is 0b,
    the Attention LED is OFF;
else If **Module_Identify** is 1b,
    the Attention LED "blinks" as defined in Table 125;
else,
    the Attention LED is ON if any of the following is set:
        MME_Module_Attention
        CME_Module_Attention
        SW_Module_Attention
        CME_RTR_Trap_Timeout

13.3.2.6.5 STATUS

The status of the Attention LED is provided through the Attention_LED_State facility (See Section 13.4.3.6 on page 531).

13.3.2.6.6 LED TEST

LED Test is defined as a single “Slow Blink” cycle (as defined in Table 125 on page 510) of both indicators in synchronism after which they are returned to the state indicating the current Module condition. If both LEDs are ON when the test is to occur, they both turn OFF for 2 seconds before the “Slow Blink” cycle.

C13-17: All Modules **shall** perform an LED Test as indicated by the following:

a) A Module **shall** initiate an LED Test upon the Module DC-DC converter’s **Vout** reaching its valid voltage level.

b) A Module **shall** initiate an LED Test upon a write to the LEDTestStatus.Initiate_LED_Test register on IB-ML.

c) A Module **shall** initiate an LED Test upon receipt of ***BMSend.SetModuleAttention(***Initiate_LED_Test***)*** from an IB link.

Software Implementation Note

Before B_Keys are assigned, any IB Software may issue the ***BMSend.SetModuleAttention(assert Module Identify)*** command. This feature serves as a diagnostic tool. After B_Keys are assigned, the Node checks Baseboard MADs (including the SetModuleAttention Attribute) against B_Key to restrict access.

13.3.2.6.7 MODULE IDENTIFY

Module Identify is a modified extension to the LED Test operation whereby the Baseboard Manager or the CME may repeatedly Blink the Attention LED to Identify the Module.

The Baseboard Manager, using ***BMSend.SetModuleAttention(***Assert_Identify***)***, or the CME, using the LEDTestStatus.Assert_Identify IB-ML register, **may** periodically re-arm the Module Identify operation.

C13-18: If the Module Identify operation is rearmed prior to the operation being complete, the MME **shall** repeat the Module Identify operation.

Software Implementation Note

The CME or the Baseboard Manager **should** repeat the Module Identify command within the "IdentifyPeriod" interval (as defined in Table 128 on page 530) so that the viewer perceives the LED blinking continuously.

13.3.2.6.8 LED BEHAVIOR IN POWER MANAGEMENT STATES

The meaning of the Module Status and Attention LEDs do not change when a power management state is entered.

Architectural Note

By definition, if the Module is claimed by a software entity, it is not OK-to-Remove. If the Module is in a power managed state, the Module Status LED reflects whether it is "claimed" or not. If the Module is claimed, the Module Status LED is ON.

The Attention LED reflects whether the Module requires attention regardless of its power management state.

13.3.2.7 BMTRAP

The Baseboard Management Trap (BMTrap) is defined in *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "BM Methods" and the related sections supporting this method. It is a mechanism that a Module can use to send an unsolicited Unreliable Datagram to the Baseboard Manager in response to environmental events or Graceful Hot Removal Requests. Additionally, a Module may be used by a CME to inject a ***BMTrap*** as a part of the Graceful Hot Removal process (See Section 13.3.2.4, "Graceful Hot Removal," on page 503).

InfiniBand Architecture Specification, Volume 1 defines that each of the General Service classes have the attribute ***ClassPortInfo***. Information in the ***BM.ClassPortInfo*** allow for ***BMTrap*** to be routed through the fabric to the Baseboard Manager. Namely, the following components of this attribute are directly applicable; the other components apply as defined in *InfiniBand Architecture Specification, Volume 1*:

- **CapabilityMask:0** - Indicates whether BMTraps are supported
- **TrapGID** - GID to use in a GRH for global routing
- **TrapLID** - LID to use in the LRH for local routing
- **TrapQ_Key** - Q_Key associated with the QP through which the BMTrap is sent

The Module embeds the GID of the Baseboard Manager and sends a Baseboard Management Class MAD Trap to the Baseboard Manager. See Section 13.6.9.1, "SendBMTrap," on page 588 for the format used for BMTrap.

A Module may send a generic Baseboard Management Trap (***BMTrap***) over IB when it detects Module-specific events.

As Baseboard Management Traps use Unreliable Datagrams, Modules **should** attempt to send three (3) BMTraps for every event. Traps may be suppressed upon the receipt of a **BMTrapRepress** MAD as defined in *InfiniBand Architecture Specification, Volume 1*, Section 13.4.9.

The Baseboard Manager **should** read the Module Status to determine the source of the event. See Section 13.6.3.16, "GetModuleStatus," on page 579 for a description of this command.

13.3.2.8 IMxINT_L

IMxInt_L is a Backplane Connector signal that a Module asserts in response to a state change of any event that the CME has enabled on the Module.

An interrupt facility has the following four (4) defined associated functions:

1) Status - indicates that a state change has occurred since the last "Clear" operation
2) Clear - clears the Status function of the facility
3) Source - indicates whether the condition is presently active. This typically is the actual "interrupt generating device" implemented as part of the MME Function.
4) Control - indicates whether the status indication is allowed to assert the **IMxInt_L** signal

These facility functions are accessed by the CME through three (3) MME registers:

- *InterruptClearStatus (Section 13.6.7.1 on page 586)*
- *InterruptSource (Section 13.6.7.2 on page 586)*
- *InterruptControl (Section 13.6.7.3 on page 586)*

Every bit in the *InterruptSource* register has a corresponding bit in the *InterruptClearStatus* and a bit in the *InterruptControl* register. See the register bit definition in InterruptClearStatus on page 584, InterruptSource on page 584, and InterruptControl on page 585 for the individual specified interrupts. Requirements for the individual interrupt source facilities are

found in Section 13.4.1, "Facility Requirements," on page 519. More than one of these facilities are required for Module form factors.

C13-19: If a Module does not implement a particular bit in *InterruptSource*, *InterruptClearStatus*, and *InterruptControl* registers, it **shall** return 0b in response to a read of that bit. Writes to un-implemented bits **shall** not have any side effects.

C13-19.1.1: A Module provides the **IMxInt_L** signal on all ports with an implemented IB-ML. The **IMxInt_L shall** comply with the operations specified in **Section 13.3.2.8, "IMxInt_L," on page 515**.

C13-20: A Module **shall** provide the corresponding functions within the *InterruptClearStatus*, *InterruptSource*, and *InterruptControl* registers for those facilities which have IB-ML "Read" access through these registers and which are labeled as "**Req**" in the "Requirements" column of Table 126.

Operation

The CME sets the respective *InterruptControl* bits for which it desires to have contribution to the assertion of **IMxInt_L**. Additionally, the CME writes 1b to all locations in the *InterruptClearStatus* register to clear the status of a detected transition.

Upon one or more events occurring, the MME sets the corresponding *InterruptSource* register bits to reflect the present state of the interrupt generating events. A low-to-high or a high-to-low transition on an event sets its corresponding bit in the *InterruptClearStatus* register.

Once the CME detects that **IMxInt_L** is asserted, it queries the MME to discern which event happened. The CME reads the *InterruptSource* and *InterruptClearStatus* registers to get the current status of the sources and to know which event(s) has had a transition since the last *InterruptClearStatus* write (Clear).

The *InterruptClearStatus* is a "Write 0b to clear, Write 1b has no effect" register (See *13.6.7.1 on page 586*). Thus, using the data returned from the *InterruptClearStatus* command, the CME can write to *InterruptClearStatus* to clear the contributions of those facilities to **IMxInt_L**. The "Clear" operation to *InterruptClearStatus* only affects the *InterruptClearStatus* register-- not the *InterruptSource* register. The Clear Interrupt command does not affect the *InterruptSource* register. The *InterruptSource* register always reflects the present state of the interrupt generating event.

13.3.2.9 PRESENCE DETECTION

Module presence detection is provided by the **IMxPRst** signal on the Backplane Connector (See Section 11.5.3, "IMxPRst," on page 449).

A Module supports the **IMxPRst** signal on every Backplane Connector it implements.

13.3.2.10 MODULEINFO DEVICE

The ModuleInfo Device holds data that provide Vital Product Data (VPD) and other capability information necessary to allow the Hardware Management of an IB Module. The ModuleInfo Device consists of one base VPD device and zero or more extension VPD devices. The base VPD device always resides at the ModuleInfo Device address specified in Table 121: Module IB-ML Slave Addresses

See Section 13.7, "xInfo Format," on page 598 for the format and structure of records and fields in ModuleInfo.

13.3.2.10.1 SPECIFIED READABLE AREA

These areas are readable via IB and IB-ML.

ModuleInfo may be obtained using the Baseboard Manager (BM) and the General Signaling Interface (GSI) over QP1. It may also be obtained using IB-ML.

13.3.2.10.2 MODULEPOWERINFO RECORD

The ModulePowerInfo record returns information about the power consumption, startup characteristics, and power management capabilities of a Module. See Section 13.7.10, "ModulePowerInfo Record," on page 622 for a list of Module power management attributes. See Section 13.7, "xInfo Format," on page 598 for the format, requirements, and structure of records and fields in ModulePowerInfo.

13.3.2.10.3 IOCPMINFO

The IOCPMINFO record indicates the power management capabilities of the IOCs associated with an IOUnit. If a Module contains multiple IOUnits and several IOCs, the information about all such IOCs appear sequentially in the IOCPMInfo record. See Section 13.7, "xInfo Format," on page 598 for the format, requirements, and structure of records and fields in IOCPMINFO.

13.3.2.10.4 CHASSIS SPECIFIC NONVOLATILE WRITABLE AREA

C13-21: IB-Modules **shall** implement 64 bytes of non-volatile writable area in ModuleInfo out of which the first 32 bytes are reserved for the Chassis. The Module may implement this writable area in the base VPD

device or in an extension VPD device. Also See xInfo Format (Section 13.7 on page 598) for more information.

This area is readable and writable via IB-ML and IB.

The data in these specified writable areas **shall** follow the OEM Record format. See Table 159 OEM Record on page 628.)

13.3.2.10.5 BASEBOARD MANAGER SPECIFIC NONVOLATILE WRITABLE AREA

C13-22: IB-Modules **shall** implement at least 64 bytes of non-volatile writable area in ModuleInfo out of which the first 32 bytes are reserved for the Chassis and the second 32 bytes are reserved for the Baseboard Manager. It is recommended that Modules provide an additional 64 writable bytes for OEM-software. The Module may implement this writable area in the base VPD device or in an extension VPD device. Also See xInfo Format (Section 13.7 on page 598) for more information.

This area is readable and writable via IB-ML and IB.

The data in these specified writable areas **shall** follow the OEM Record format. See Table 159 OEM Record on page 628.)

13.3.2.11 MODULE PRIVATE AREAS

o13-6: For private use, IB-Modules **may** optionally implement read-only or writable non-volatile data areas using IB-ML devices that are mentioned in Module IB-ML Slave Addresses (Section Table 121 on page 496). Also See xInfo Format (Section 13.7 on page 598) for more information.

These devices are readable via IB-ML and IB.

13.4 MODULE FACILITIES

Module Facilities provide for status and logical control of the management features of an InfiniBand Module. As outlined in other sections of this chapter, in-band (Baseboard MADs) and out-of-band (IB-ML) access paths are architected for many features; however, in many cases, the access is not the same. For example, a given facility may be read accessible from both paths, but only one path is architecturally allowed to write it. This section details which paths are architected for each facility.

For the Baseboard MAD path, facilities are accessed using “commands” that are summarized in Table 136 Baseboard MAD Commands on page 570. For the IB-ML path, facilities are accessed using “registers” that are summarized in Table 140 MME Function Registers on page 584.

C13-23: For each of the Module facilities described in this section, a "default" state is indicated. Upon a power-on or reset, a Module **shall** force the facilities to the default state values unless otherwise noted.

See Section 13.4.1, "Facility Requirements," on page 519 for a list of requirements for all facilities by physical packaging and for a list of the applicable Baseboard commands and IB-ML registers that enable access to those facilities. Special considerations that apply to any of the access paths are noted in the individual facility description.

13.4.1 FACILITY REQUIREMENTS

13.4.1.1 MODULES

Table 126 summarizes the defined facilities for an InfiniBand Module.

C13-24: A Module containing an InfiniBand protocol-aware device **shall** provide those facilities labeled as "**Req**" in the "Requirements" column of Table 126.

C13-25: A Module containing an InfiniBand non-protocol-aware device **shall** provide those facilities labeled as "**Req**" in the "Requirements" column of Table 126. Only the "IB-ML" Access Path applies.

The "Requirement" column cells of the Table 126 **shall** have the following definition:

- **Req** - A Module **shall** implement the facility labeled **Req** in the "Requirement" column.
- **Opt** - The Module **may** implement the facility labeled **Opt** in the "Requirement" column. If the facility is provided, the Module **shall** implement it in accordance with the provisions of this section. If the facility is not provided, the Module **shall** provide the default values for read accesses that are specified for the facility bit(s); write accesses **shall not** cause updates to occur from the default value for the facility.

The "Access Path" columns of Table 126 **shall** have the following definitions:

- **Baseboard, Read** - The ability to read the current state of the named facility through the noted command and associated response as detailed in Table 136 Baseboard MAD Commands on page 570 under "Response Data".
- **Baseboard Access - Write** - The ability to update the state of the facility with the value in a noted received command as detailed in Table 136 Baseboard MAD Commands on page 570 under "Request Data".

- **IB-ML Access - Read** - The ability to read the current state of the named facility through the noted register as defined in Table 140 MME Function Registers on page 584 under “Read Data” by accessing the facility either through physical registers via address E0h, or via the WriteMMERegister MME Command
- **IB-ML Access - Write** - The ability to update the current state of the named facility through the noted register as defined in Table 140 MME Function Registers on page 584 under “Write Data” by accessing the facility either through physical registers via address E0h, or via the WriteMMERegister MME Command

The “Access Path” column cells of the Table 126 **shall** have the following definitions:

- **NP** - The Module **shall not** permit access to the facility with the path indicated by the column.
- **NS** - Access to the facility with the access path indicated by the column is not specified.
- **“cmd/reg”** - If the cell contains the name of a Baseboard command or an MME register as applicable to the access path, then the Module **shall** permit access to the facility with the path indicated by the column. Note that in the case that the access path is a register, the access may be implemented either by using a physical register via E0h or via the corresponding split-transaction Write/ReadMMERegister MME command via E4h.

Table 126 Module Facility Requirements

Facility	Requirement	Access Path			
		Baseboard[a]		IB-ML	
		Write	Read	Write	Read
Removal Control and Status					
Module_OKTR	**Req**	NP	GetModuleStatus	NP	RemovalControlStatus
Attention_Status_Change	**Req**	NS	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Removal_Status_Change	**Req**	NP	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
MME_RTR	**Opt**	NS	GetModuleStatus	NP	RemovalControlStatus
MME_CTR	**Opt**	NS	GetModuleStatus	NP	RemovalControlStatus
CME_Force_OKTR	**Req**	NS	GetModuleStatus	RemovalControlStatus	RemovalControlStatus

Table 126 Module Facility Requirements

Facility	Requirement	Access Path			
		Baseboard[a]		IB-ML	
		Write	Read	Write	Read
CME_RTR	Req	NP	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
CME_CTR	Req	NP	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
SW_CTR	Req	SetModuleState	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
SW_RTR	Req	SetModuleState	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
Indicator Control and Status					
Module_Identify	Req	NP	GetModuleStatus	NP	LEDTestStatus
Assert_Identify	Req	SetModuleState	NS	LEDTestStatus	NS
MME_Module_Attention	Opt	NS	GetModuleStatus	NP	LEDTestStatus
CME_Module_Attention	Req	NP	GetModuleStatus	LEDTestStatus	LEDTestStatus
SW_Module_Attention	Req	SetModuleAttention	GetModuleStatus	NP	LEDTestStatus
Attention_LED_State	Req	NP	GetModuleStatus	NP	LEDTestStatus
Module_Status_LED_State	Req	NP	GetModuleStatus	NP	LEDTestStatus
LED_Enable	Req	NP	GetModuleStatus	LEDTestStatus	LEDTestStatus
LED_Test	Req	NP	GetModuleStatus	NP	LEDTestStatus
Initiate_LED_Test	Req	SetModuleAttention	NS	LEDTestStatus	NP
Interrupt Control and Status					
CME_RTR_Trap_Timeout	Req	NP	GetModuleStatus	ModuleControlStatus	ModuleControlStatus
Power_Converter_Fault	Req	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Environmental_Fault	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Functional_Fault	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Hard_Fault	Opt	NP	NS	NP	LEDTestStatus
Degraded_Fault	Opt	NP	NS	NP	LEDTestStatus
Predictive_Fault	Opt	NS	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Cooling_Non_Crit	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource

Table 126 Module Facility Requirements

Facility	Requirement	Access Path			
		Baseboard[a]		IB-ML	
		Write	Read	Write	Read
Cooling_Crit	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Cooling_Non_Recoverable	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Voltage_Crit	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Voltage_Non_Recoverable	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Module_Power_On	Req	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Module Control and Status					
Pwr_Conv_Redundancy	Opt	NP	GetModuleStatus	NP	ModulePower
WakeOnIB	Opt	NP	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
WakeOnWRE	Opt	NP	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
IMxInt_L_State	Req	NP	NS	NP	ModuleControlStatus
WakeOnIB_Enable	Opt	NP	NS	ModuleControlStatus	ModuleControlStatus
IMxInt_L_Enable	Req	NP	NS	ModuleControlStatus	ModuleControlStatus
WRE_Enable	Opt	SetIOCPMControl	GetIOCPMControl	NP	ModuleControlStatus
WRE_Status	Opt	SetIOCPMControl	GetIOCPMControl	NP	ModuleControlStatus

a. The access path is only applicable for protocol-aware Modules.

13.4.1.2 NON-MODULE DEVICES

Table 127 summarizes the facilities defined for “Non-Module” devices.

C13-26: An InfiniBand protocol-aware device mounted on a package other than an InfiniBand Module and supporting the optional Baseboard Management Agent (BMA) **shall** provide those facilities labeled as “**Req**” in the “Requirement” column of Table 127.

The “Requirement” column cells of the Table 127 **shall** have the following definition:

- **Req** - A Non-Module **shall** implement the facility labeled **Req** in the “Requirement” column.
- **Opt** - A Non-Module **may** implement the facility labeled **Opt** in the “Requirement” column. If the facility is provided, the Module **shall** implement it in accordance with the provisions of this section. If the facility is not provided, the Module **shall** provide the default values for read accesses that are specified for the facility bit(s); write accesses **shall not** cause updates to occur from the default value for the facility.

The “Access Path” columns of Table 127 **shall** have the following definitions:

- **Baseboard, Read** - The ability to read the current state of the named facility through the noted command and associated response as detailed in Table 136 Baseboard MAD Commands on page 570 under “Response Data”.
- **Baseboard Access - Write** - The ability to update the state of the facility with the value in a received noted command as detailed in Table 136 Baseboard MAD Commands on page 570 under “Request Data”.
- **IB-ML Access - Read** - The ability to read the current state of the named facility through the noted register as defined in Table 140 MME Function Registers on page 584 under “Read Data”.
- **IB-ML Access - Write** - The ability to update the current state of the named facility through the noted register as defined in Table 140 MME Function Registers on page 584 under “Write Data”.

The “Access Path” column cells of the Table 127 **shall** have the following definitions:

- **NP** - The Module **shall not** permit access to the facility with the path indicated by the column.
- **NS** - Access to the facility with the access path indicated by the column is not specified.
- **“cmd/reg”** - If the cell contains the name of a Baseboard command, or an MME register as applicable to the access path, then the Module **shall** permit access to the facility with the path indicated by the column. Note that in the case that the access path is

a register, the access may be implemented either by using a physical register via E0h or via the corresponding split-transaction Write/ReadMMERegister MME command via E4h.

Table 127 Non-Module Device Facility Requirements

Facility	Requirement	Access Path			
		Baseboard		IB-ML	
		Write	Read	Write	Read
Removal Control and Status					
Module_OKTR	Opt	NP	GetModuleStatus	NP	RemovalControlStatus
Attention_Status_Change	Opt	NS	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Removal_Status_Change	Opt	NP	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
MME_RTR	Opt	NS	GetModuleStatus	NP	RemovalControlStatus
MME_CTR	Opt	NS	GetModuleStatus	NP	RemovalControlStatus
CME_Force_OKTR	Opt	NS	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
CME_RTR	Opt	NP	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
CME_CTR	Opt	NP	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
SW_CTR	Opt	SetModuleState	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
SW_RTR	Opt	SetModuleState	GetModuleStatus	RemovalControlStatus	RemovalControlStatus
Indicator Control and Status					
Module_Identify	Opt	NP	GetModuleStatus	NP	LEDTestStatus
Assert_Identify	Opt	SetModuleState	NS	LEDTestStatus	NS
MME_Module_Attention	Opt	NS	GetModuleStatus	NP	LEDTestStatus
CME_Module_Attention	Opt	NP	GetModuleStatus	LEDTestStatus	LEDTestStatus
SW_Module_Attention	Opt	SetModuleAttention	GetModuleStatus	NP	LEDTestStatus
Attention_LED_State	Opt	NP	GetModuleStatus	NP	LEDTestStatus
Module_Status_LED_State	Opt	NP	GetModuleStatus	NP	LEDTestStatus
LED_Enable	Opt	NP	GetModuleStatus	LEDTestStatus	LEDTestStatus
LED_Test	Opt	NP	GetModuleStatus	NP	LEDTestStatus
Initiate_LED_Test	Opt	SetModuleAttention	NS	LEDTestStatus	NP
Interrupt Control and Status					
CME_RTR_Trap_Timeout	Opt	NP	GetModuleStatus	ModuleControlStatus	ModuleControlStatus

Table 127 Non-Module Device Facility Requirements

Facility	Requirement	Access Path			
		Baseboard		IB-ML	
		Write	Read	Write	Read
Power_Converter_Fault	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Environmental_Fault	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Functional_Fault	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Predictive_Fault	Opt	NS	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Cooling_Non_Crit	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Cooling_Crit	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Cooling_Non_Recoverable	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Voltage_Crit	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Module_Power_On	Opt	NP	GetModuleStatus	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
Module Control and Status					
Pwr_Conv_Redundancy	Opt	NP	GetModuleStatus	NP	ModulePower
WakeOnIB	Opt	NP	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
WakeOnWRE	Opt	NP	NS	InterruptClearStatus, InterruptControl	InterruptClearStatus, InterruptControl, InterruptSource
IMxInt_L_State	Opt	NP	NS	NP	ModuleControlStatus
WakeOnIB_Enable	Opt	NP	NS	ModuleControlStatus	ModuleControlStatus
IMxInt_L_Enable	Opt	NP	NS	ModuleControlStatus	ModuleControlStatus
WRE_Enable	Opt	SetIOCPMControl	GetIOCPMControl	NP	ModuleControlStatus
WRE_Status	Opt	SetIOCPMControl	GetIOCPMControl	NP	ModuleControlStatus

IBTA

13.4.2 REMOVAL CONTROL/STATUS FACILITY DESCRIPTIONS

13.4.2.1 MODULE_OKTR

Module_OKTR (Module OK to Remove) is a bit facility that the Module controls to provide a summary of the defined “Clear to Remove” facilities.

The **Module_OKTR** facility has the following definition:

0b = Module is not OK to Remove
1b = Module is OK to Remove (default)

C13-27: The Module **shall** set the **Module_OKTR** bit to 1b if all of the following are set to 1b:

- **MME_CTR**
- **CME_CTR**
- **SW_CTR**

13.4.2.2 ATTENTION_STATUS_CHANGE

Attention_Status_Change is a bit facility that the Module controls to indicate that a change of state has occurred on any facility bits that contribute to the state of the Attention LED. This bit includes state status change for the following facilities:

- **MME_Module_Attention**
- **CME_Module_Attention**
- **SW_Module_Attention**

The following bits are included in the **Attention_Status_Change** facility:

- *InterruptSource.Attention_Status_Change*
- *InterruptControl.Attention_Status_Change*
- *InterruptClear/Status.Attention_Status_Change*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in 13.3.2.8 IMxInt_L on page 515.

The *InterruptSource.Attention_Status_Change* indicates whether any of the attention conditions are presently active.

0b = MME_Module_Attention, CME_Module_Attention, and SW_Module_Attention are all presently 0b (default).
1b = One or more of the MME_Module_Attention, CME_Module_Attention, or SW_Module_Attention bits are presently 1b.

The *InterruptStatus.Attention_Status_Change* bit indicates the current interrupt condition:

0b = Attention Status bits have not changed since last Interrupt Clear. (default)
1b = At least one Attention Status bit has changed states since last Interrupt Clear.

13.4.2.3 REMOVAL_STATUS_CHANGE

Removal_Status_Change is a bit facility that the Module controls to indicate that a change of state has occurred on any facility bits of the Module Removal Status register so as to contribute to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515. This bit includes state status change for the following facilities:

- **Module_OKTR**
- **MME_RTR**
- **MME_CTR**
- **CME_Force_OKTR**
- **CME_RTR**
- **CME_CTR**
- **SW_CTR**
- **SW_RTR**

The following bits are included in the **Removal_Status_Change** facility:

- *InterruptSource.Removal_Status_Change*
- *InterruptControl.Removal_Status_Change*
- *InterruptClear/Status.Removal_Status_Change*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The *InterruptSource.Removal_Status_Change* bit indicates the current interrupt condition:

0b = All Module Removal Status bits are presently 0b (default)
1b = One or more of the Module Removal Status bits are presently 1b.

The *InterruptStatus.Removal_Status_Change* bit indicates the current interrupt condition:

0b = The Module Removal Status bits have not changed since last interrupt clear (default)
1b = At least one bit in the Module Removal Status bits has changed states since last interrupt clear.

IBTA

13.4.2.4 MME_RTR

MME_RTR (MME Request to Remove) is a bit facility that the MME controls to indicate a request to the Baseboard Manager, other claiming software, and the CME to release the Module. See Section 13.3.2.4, "Graceful Hot Removal," on page 503.

The facility has the following definition:

0b = No Request has been Initiated, or Request has been removed (default)
1b = A Request is Pending.

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.2.5 MME_CTR

MME_CTR (MME Clear to Remove) is a bit facility that the MME controls to indicate that Module removal is allowed. See Section 13.3.2.4, "Graceful Hot Removal," on page 503 for a description on potential uses of this facility.

The **MME_CTR** facility has the following definition:

0b = Claimed
1b = Not Claimed (default)

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.2.6 CME_FORCE_OKTR

CME_Force_OKTR is a bit facility that the CME controls to indicate that the Module is OK to Remove independent of the states of **SW_CTR**, **MME_CTR**, and **CME_CTR**.

The **CME_Force_OKTR** facility has the following definition:

0b = Not forcing OKTR (default)
1b = Force OKTR

13.4.2.7 CME_RTR

CME_RTR (CME Request to Remove) is a bit facility that the CME controls to indicate that the CME is requesting the Baseboard Manager and other claiming software to release the IB Module containing this xCA in order to enable removal of the Module.

The **CME_RTR** facility has the following definition:

0b = Claimed

1b = Not Claimed (default)

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.2.8 CME_CTR

CME_CTR (CME Clear to Remove) is a bit facility that the CME controls to indicate that the CME has performed all Chassis specific operations that would allow the Module to be removed.

The **CME_CTR** facility has the following definition:

0b = Not Clear to Remove
1b = Clear to Remove (default)

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.2.9 SW_CTR

SW_CTR (Software Clear to Remove) is a bit facility that the Baseboard Manager controls to indicate that the Module is claimed such that it is not OK to Remove this Module. See Section 13.3.2.4, "Graceful Hot Removal," on page 503 for further description on the use of the this facility.

The **SW_CTR** facility has the following definition:

0b = Claimed
1b = Not Claimed (default)

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.2.10 SW_RTR

SW_RTR (Software Request to Remove) is a bit facility that the Baseboard Manager controls to indicate it is about to initiate the request actions with the interested software entities to determine when a quiescent state can be achieved. See Section 13.3.2.4, "Graceful Hot Removal," on page 503 for further description on the use of this facility.

The **SW_RTR** facility has the following definition:

0b = No Request has been Initiated, or Request has been removed (default)
1b = A Request is Pending.

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.3 INDICATOR CONTROL/STATUS FACILITY DESCRIPTIONS

13.4.3.1 MODULE_IDENTIFY

Module_Identify is a bit facility that provides status that the LED function defined in Section 13.3.2.6.7, "Module Identify," on page 513 was initiated by Baseboard Manager software or the CME.

The **Module_Identify** facility has the following definition:

0b = Not performing Identify (default)
1b = Performing Identify

13.4.3.2 ASSERT_IDENTIFY

Assert_Identify is a bit facility that the Baseboard Manager controls which initiates or rearms the LED function defined in Section 13.3.2.6.7, "Module Identify," on page 513.

The **Assert_Identify** facility has the following definition:

0b = Stop the Module Identify function. (default)
1b = Initiate the Module Identify on the Attention LED and repeat it for a period of time equal to IdentifyPeriod. See Table 128: Identify Function Period

BASEBOARD ACCESS - READ

Read operations to the **Assert_Identify** facility using Baseboard MADs are not specified. The Module_Identify facility provides the state of the Identify function. (See Section 13.4.3.1).

IB-ML ACCESS - READ

Read operations to the **Assert_Identify** facility using IB-ML are not specified. The Module_Identify facility provides the state of the Identify function. (See Section 13.4.3.1).

Table 128 Identify Function Period

Term	Period	Comments
IdentifyPeriod	8s +/- 25%	Once Assert_Identify facility been invoked, initiate the Module Identify on the Attention LED and repeat it for this amount of time.

13.4.3.3 MME_MODULE_ATTENTION

MME_Module_Attention is a bit facility that the MME controls to assert the Attention LED if the LEDs are enabled with **LED_Enable** (See Section 13.4.3.8 on page 532).

The **MME_Module_Attention** facility has the following definition:

0b = No Attention condition present (default)
1b = Attention condition present

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

Although MME Registers may not be available if a Module's DC-DC converter is OFF, if $\mathbf{V_{Aux}}$ is available, the Module's Attention LED maintains its state (ON or OFF). It reflects the state it had prior to Module's DC-DC going OFF. For example, the Attention LED may reflect a DC-DC converter failure.

13.4.3.4 CME_MODULE_ATTENTION

CME_Module_Attention is a bit facility that the CME controls to assert the Attention LED if the LEDs are enabled with **LED_Enable** (See Section 13.4.3.8 on page 532).

The **CME_Module_Attention** facility has the following definition:

0b = No Attention condition present (default)
1b = Attention condition present

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.3.5 SW_MODULE_ATTENTION

SW_Module_Attention is a bit facility that the Baseboard Manager controls to assert the Attention LED given LEDs are enabled with **LED_Enable** (See Section 13.4.3.8 on page 532).

The **SW_Module_Attention** facility has the following definition:

0b = No Attention condition present (default)
1b = Attention condition present

INTERRUPT EFFECTS

See 13.4.2.3 Removal_Status_Change on page 527 for implications this bit has on the **IMxInt_L** signal.

13.4.3.6 ATTENTION_LED_STATE

Attention_LED_State is a facility which reflects the current state of the Attention LED.

The **Attention_LED_State** facility has the following definition:

00b = OFF (Neither CME, MME, nor the Baseboard Manager has indicated an attention event and *CME_RTR_Trap_Timeout is 0b*.)

DRAFT

01b = ON (Either CME, MME, or the Baseboard Manager has indicated an attention event OR *CME_RTR_Trap_Timeout is 1b*.)
10b = Blink (indicates Module Identify)
11b = reserved

13.4.3.7 MODULE_STATUS_LED_STATE

Module_Status_LED_State is a facility which reflects the current state of the Module Status LED.

The **Module_Status_LED_State** facility has the following definition:

00b = OFF
01b = ON
10b = BLINK
11b = reserved

13.4.3.8 LED_ENABLE

LED_Enable is a bit facility indicating to the CME and to the Baseboard Manager whether the LEDs are enabled for operation. The CME has write access to this bit.

The **LED_Enable** facility has the following definition:

0b = LEDs are always OFF.
1b = LEDs follow the states detailed in Section 13.3.2.6, "Module Indicators (LEDs)," on page 508. (default)

13.4.3.9 LED_TEST

LED_Test is a bit facility indicating the state of the function defined in Section 13.3.2.6.6, "LED Test," on page 513 to the CME and to the Baseboard Manager.

The **LED_Test** facility has the following definition:

0b = LED Test function is not in progress.
1b = LED Test function is in progress.

13.4.3.10 INITIATE_LED_TEST

Initiate_LED_Test is a bit facility that either the Baseboard Manager or the CME controls to initiate or rearm the LED function defined in Section 13.3.2.6.6, "LED Test," on page 513.

The **Initiate_LED_Test** facility has the following definition:

0 = Write of 0b has no effect. Once the test starts, it goes to completion.
1 = Initiate the LED Test on both Green and Attention LEDs.

BASEBOARD ACCESS - READ

Read operations to the **Initiate_LED_Test** facility using Baseboard MADs are not specified. The LED_Test facility provides the state of the **LED Test** function. (See Section 13.4.3.9).

IB-ML ACCESS - READ

Read operations to the **Initiate_LED_Test** facility using IB-ML are not architected. The LED_Test facility provides the state of the **LED Test** function. (See Section 13.4.3.9).

13.4.4 INTERRUPT CONTROL/STATUS FACILITY DESCRIPTIONS

13.4.4.1 CME_RTR_TRAP_TIMEOUT

CME_RTR_Trap_Timeout is a bit facility that the CME controls to indicate that the CME has timed-out while waiting for the **SW_RTR** to be set in acknowledgement to a CME Request to Remove (**CME_RTR**) ***BMTrap***.

The CME **may** choose to over-ride the Baseboard Manager using the **CME_Force_OKTR** facility (See 13.4.2.6 CME_Force_OKTR on page 528).

Implementation Note

This time-out is an indication that the Baseboard Manager responsible for receiving the BMTrap may not be operating correctly. This time-out does not mean that the interested software entities have taken too long to respond with a **SW_CTR** because the Baseboard Manager is supposed to acknowledge the ***BMTrap*** before it queries the other interested software entities for their “vote”. See Section 13.3.2.4, “Graceful Hot Removal,” on page 503.

Although not directly related to Module OK To Remove, this time-out indicates that the **SW_RTR** and the **SW_CTR** is not set, and therefore, the Module is not OK To Remove from the Software’s point of view. See Section 13.3.2.4, “Graceful Hot Removal,” on page 503.

The **CME_RTR_Trap_Timeout** facility has the following definition:

0b = No time-out present (default)
1b = Time condition present

13.4.4.2 POWER_CONVERTER_FAULT

Power_Converter_Fault is a bit facility that the Module controls to indicate that the Module has turned off its main power converter(s) due to a power fault condition -- not in response to **VBxEn_L** negation, and not in response to an IB Power Down Request.

The following bits are included in the **Power_Converter_Fault** facility:

- *InterruptSource.Power_Converter_Fault*
- *InterruptControl.Power_Converter_Fault*
- *InterruptClear/Status.Power_Converter_Fault*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The **InterruptSource.Power_Converter_Fault** bit has the following definition:

0b = No fault condition exists (default)
1b = Module power converter(s) off due to a fault condition

BASEBOARD ACCESS - READ

Architectural Note

If the Module DC-DC converter is off, it is very likely that its IB port is non-functional and thus renders the Baseboard MAD access as moot. However, to facilitate redundant power designs or designs that may use alternate power sources, this access is architecturally provided.

13.4.4.3 ENVIRONMENTAL_FAULT

Environmental_Fault is a facility that the Module controls to indicate that an environmental condition other than voltage or cooling conditions has occurred.

The following bits are included in the **Environmental_Fault** facility:

- *InterruptSource.Environmental_Fault*
- *InterruptControl.Environmental_Fault*
- *InterruptClear/Status.Environmental_Fault*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The **InterruptSource.Environmental_Fault** bit has the following definition:

0b = Module is not in an environmental fault condition (default)
1b = Module has detected an environmental fault condition.

BASEBOARD ACCESS - READ

Architectural Note

If the Module DC-DC converter is off, it is very likely that its IB port is non-functional and thus renders the Baseboard MAD access as moot. However, to facilitate redundant power designs or designs that may use alternate power sources, this access is architecturally provided.

13.4.4.4 FUNCTIONAL_FAULT

Functional_Fault is a bit facility that the Module controls to indicate that a Module specific functional fault condition was detected. This bit includes state from the following facilities:

- **Hard_Fault**
- **Degraded_Fault**

The following bits are associated with the **Functional_Fault** facility:

- *InterruptSource.FUNCTIONAL_FAULT*
- *InterruptControl.FUNCTIONAL_FAULT*
- *InterruptClear/Status.FUNCTIONAL_FAULT*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The **InterruptSource.Functional_Fault** bit has the following definition:

0b = Module is not in a functional fault condition (default)
1b = Module has detected a functional fault condition.

13.4.4.5 HARD_FAULT

HARD_FAULT is a bit facility that the Module controls to indicate that a Module specific functional fault condition that renders the Module inoperable was detected.

The **HARD_FAULT** bit has the following definition:

0b = Module is not in a hard functional fault condition (default)
1b = Module has detected a hard functional fault condition.

13.4.4.6 DEGRADED_FAULT

Degraded_Fault is a bit facility that the Module controls to indicate that a Module specific functional fault condition was detected but the Module is capable to functioning with degrade capabilities.

The **Degraded_Fault** bit has the following definition:

0b = Module is not in a degraded functional fault condition (default)
1b = Module has detected a degraded functional fault condition.

Architectural Note

The **Degraded_Fault** bit is intended to indicate that degraded faults such as loss of cache memory due to excessive ECC error and loss of power converter redundancy had occurred. Other error examples are certainly possible.

13.4.4.7 PREDICTIVE_FAULT

Predictive_Fault is a bit facility that the Module controls to indicate that a non-critical functional event or condition has occurred that may eventually lead to a critical failure.

The following bits are associated with the **Predictive_Fault** facility:

- *InterruptSource.Predictive_Fault*
- *InterruptControl.Predictive_Fault*
- *InterruptClear/Status.Predictive_Fault*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, “IMxInt_L,” on page 515.

The **InterruptSource.Predictive_Fault** bit has the following definition:

0b = Module is functioning optimally. (default)
1b = Module is in a degraded state.

BASEBOARD ACCESS - WRITE

The setting of the **InterruptSource.Predictive_Fault** bit through InfiniBand link operations is Module specific.

13.4.4.8 COOLING_NON_CRIT

Cooling_Non_Crit is a bit facility that the Module controls to indicate a temperature sensor(s) on the Module has crossed its Module-specific warning temperature threshold but is still operating.

The following bits are associated with the **Cooling_Non_Crit** facility:

- *InterruptSource.Cooling_Non_Crit*
- *InterruptControl.Cooling_Non_Crit*
- *InterruptClear/Status.Cooling_Non_Crit*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The **InterruptSource.Cooling_Non_Crit** bit has the following definition:

0b = Module Temperature and Cooling Status are normal. (default)
1b = Module Temperature has exceeded its warning threshold or Module cooling apparatus is not running optimally

13.4.4.9 COOLING_CRIT

Cooling_Crit is a bit facility that the Module controls to indicate a temperature sensor(s) on the Module has crossed its Module-specific critical temperature threshold and that the Module has turned off its Power Converter(s).

The following bits are associated with the **Cooling_Crit** facility:

- *InterruptSource.Cooling_Crit*
- *InterruptControl.Cooling_Crit*
- *InterruptClear/Status.Cooling_Crit*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The **InterruptSource.Cooling_Crit** bit has the following definition:

0b = Module Temperature and Cooling Status are not critical. (default)
1b = Module Temperature has exceeded its critical threshold and Module Power Converter is off.

BASEBOARD ACCESS - READ

Architectural Note

As the Module DC-DC converter is off, it is very likely that its IB port is non-functional and thus renders the Baseboard MAD access as moot. However, to facilitate potential design options that may use alternate power sources, this access is architecturally provided.

13.4.4.10 COOLING_NON_RECOVERABLE

Cooling_Non_Recoverable is a bit facility that the Module controls to indicate that the temperature on the Module has gone over its Module-specific non-recoverable temperature threshold and that the Module has turned off its power converter(s). In addition, the Module is not reliable anymore because of a potential damage due to excessive over-heating. Once the Module reaches this temperature, the status indication of this fa-

cility will be permanently asserted (even through power sequences). (e.g. use a Fuse.)

The following bits are associated with the **Cooling_Non_Recoverable** facility:

- *InterruptSource.Cooling_Non_Recoverable*
- *InterruptControl.Cooling_Non_Recoverable*
- *InterruptClear/Status.Cooling_Non_Recoverable*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The **InterruptSource.Cooling_Non_Recoverable** bit has the following definition:

0b = Module Temperature and Cooling Status are not non-recoverable. (default)
1b = Module Temperature has exceeded its Non-recoverable threshold, Module Power Converter is off, and this Module is not reliable anymore.

BASEBOARD ACCESS - READ

Architectural Note

As the Module DC-DC converter is off, it is very likely that its IB port is non-functional and thus renders the Baseboard MAD access as moot.

13.4.4.11 VOLTAGE_CRIT

Voltage_Crit is a bit facility that the Module controls to indicate a voltage monitoring device on the Module has crossed its Module-specific critical voltage threshold and that the Module has turned off its power converter(s).

The following bits are associated with the **Voltage_Crit** facility:

- *InterruptSource.Voltage_Crit*
- *InterruptControl.Voltage_Crit*
- *InterruptClear/Status.Voltage_Crit*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The **InterruptSource.Voltage_Crit** bit has the following definition:

0b = Module Voltage is within range. (default)
1b = Module Voltage is out of range and the Module Power Converter is off.

BASEBOARD ACCESS - READ

Architectural Note

As the Module DC-DC converter is off, it is very likely that its IB port is non-functional and thus renders the Baseboard MAD access as moot. However, to facilitate potential design options that may use alternate power sources, this access is architecturally provided.

13.4.4.12 VOLTAGE_NON_RECOVERABLE

Voltage_Non_Recoverable is a bit facility that the Module controls to indicate that the voltage on the Module has gone over its Module-specific non-recoverable voltage threshold and that the Module has turned off its power converter(s). In addition, the Module is not reliable anymore because of a potential damage due to excessive voltage. Once the Module reaches this threshold, the status indication of this facility will be permanently asserted (even through power sequences). (e.g. use a Fuse.)

The following bits are associated with the **Voltage_Non_Recoverable** facility:

- *InterruptSource.Voltage_Non_Recoverable*
- *InterruptControl.Voltage_Non_Recoverable*
- *InterruptClear/Status.Voltage_Non_Recoverable*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, “IMxInt_L,” on page 515.

The **InterruptSource.Voltage_Non_Recoverable** bit has the following definition:

0b = Module Voltage is not non-recoverable. (default)
1b = Module Voltage has exceeded its non-recoverable threshold, Module Power Converter is off, and this Module is not reliable anymore.

DRAFT

BASEBOARD ACCESS - READ

Architectural Note

As the Module DC-DC converter is off, it is very likely that its IB port is non-functional and thus renders the Baseboard MAD access as moot.

13.4.4.13 MODULE_POWER_ON

Module_Power_On is a bit facility indicating whether one or more of the Module power converter(s) are enabled and providing output voltage and current within Module specific parameters.

The following bits are included in the **Module_Power_On** facility:

- *InterruptSource.Module_Power_On*
- *InterruptControl.Module_Power_On*
- *InterruptClear/Status.Module_Power_On*

These bits are associated as part of the control and status function that contributes to the backplane interrupt signal as defined in Section 13.3.2.8, "IMxInt_L," on page 515.

The *InterruptSource.Module_Power_On* bit has the following definition:

0b = All Module DC-DC Converters are Off
1b = At least one DC-DC Converter is On and providing Module specified power

Architecture Note

If only one power converter is present, the bit represents that state of that converter; if more than one converter is present in a redundant arrangement, this bit conveys if any of the converters are ON.

BASEBOARD ACCESS - READ

> **Architectural Note**
>
> If the Module DC-DC converter is off, it is very likely that its IB port is non-functional and thus renders the Baseboard MAD access as moot. However, to facilitate redundant power designs or designs that may use alternate power sources, this access is architecturally provided.

13.4.5 MODULE CONTROL/STATUS FACILITY DESCRIPTIONS

13.4.5.1 PWR_CONV_REDUNDANCY

Pwr_Conv_Redundancy is a facility controlled by the Module to indicate the redundancy state of its power converter function.

The **Pwr_Conv_Redundancy** facility has the following definition:

00b = non-redundant converter
01b = redundant converter is operating fully redundant
10b = redundant converter is operating with degraded redundancy
11b = redundant converter is operating non-redundant

13.4.5.2 WAKEONIB

WakeOnIB is a bit facility controlled by the Module to indicate that the last wake-up was caused by a WakeOnIB event. (Beacon detected.)

The **WakeOnIB** facility has the following definition:

0b = Last wake-up was not due to WakeOnIB (default)
1b = Last wake-up was due to WakeOnIB

13.4.5.3 WAKEONWRE

WakeOnWRE is a bit facility controlled by the Module to indicate that the last wake-up was caused by a Wake-Request-Event.

The **WakeOnWRE** facility has the following definition:

0b = Last wake-up was not due to WRE (default)
1b = Last wake-up was due to WRE

13.4.5.4 IMXINT_L_STATE

IMxInt_L_State is a bit facility controlled by the Module to indicate the current state of the **IMxInt_L** pin.

The **IMxInt_L_State** facility has the following definition:

0b = **IMxInt_L** signal is not asserted; it is at a high voltage.
1b = **IMxInt_L** signal is asserted; it is at a low voltage.

13.4.5.5 WAKEONIB_ENABLE

WakeOnIB_Enable is a bit facility controlled by the CME to control and indicate whether the power controller on the Module will allow wake-up based on a WakeOnIB event. (Beacon detected.)

The **WakeOnIB_Enable** facility has the following definition:

0b = wake-up due to WakeOnIB is Disabled
1b = wake-up due to WakeOnIB is Enabled (default)

Implementation Note

Use of this facility is not recommended and is subject to removal in future versions of the specification.

If it is used to disable the detection of beaconing events while the module is the $M_{Standby}$ state (which was established by in-band Power Management operations), the awakening anticipated by the in-band Power Manager through beaconing will not occur. Depending on the system design, this may result in service replacement of the module as being faulty that would otherwise not be warranted.

13.4.5.6 IMXINT_L_ENABLE

IMxInt_L_Enable is a bit facility controlled by the CME to control and indicate whether the Module will assert **IMxInt_L** based on *InterruptStatus* and *InterruptControl* registers.

The **IMxInt_L_Enable** facility has the following definition:

0b = Will not assert **IMxInt_L** (default)
1b = Will assert **IMxInt_L** based on *InterruptControl* and *InterruptStatus* registers

13.4.5.7 WRE_ENABLE

WRE_Enable (Wake Request Event Enable) is a bit facility controlling whether an IOC is enabled to generate a Wake Request Event defined in Section 14.2.5.1, “Wake Request Event,” on page 635.

This bit is “sticky” and is not to be affected by resets (power-on or **IMxPRst**) or other power state transitions; thus, the default state of this facility is undefined. Power Management software must explicitly establish the initial state through Baseboard commands. With the sticky nature defined, this facility is considered to be run on auxiliary power.

If a Module supports Power Management (*ModulePowerInfo.ModulePMCapability.IsPowerManagementSupported* = 1b) and additionally sup-

ports WRE generation as indicated in its *IOCPMInfo.IOCPmCapability* (any of *WREisIDozeSupported, WREisINapSupported*, *WREisISleep-Supported, WREisIStandbySupported* = 1b), it **shall** provide the **WRE_Enable** facility.

The **WRE_Enable** facility has the following definition:

0b = WREs Not Enabled

1b = WREs Enabled

13.4.5.8 WRE_STATUS

WRE_Status (Wake Request Event Status) is a bit facility indicating whether an IOC has generated a Wake Request Event defined in Section 14.2.5.1, "Wake Request Event," on page 635 independent of the setting of WRE_Enable (See Section 13.4.5.7).

This bit is "sticky" and is not to be affected by resets (power-on or **IMx-PRst**) or other power state transitions; thus, the default state of this facility is undefined. Power Management software must explicitly establish the initial state through Baseboard commands. With the sticky nature defined, this facility is considered to be run on auxiliary power.

If a Module supports Power Management (*ModulePowerInfo.ModulePM-Capability.IsPowerManagementSupported* = 1b) and additionally supports WRE generation as indicated in its *IOCPMInfo.IOCPmCapability* (any of *WREisIDozeSupported, WREisINapSupported*, *WREisISleep-Supported, WREisIStandbySupported* = 1b), it **shall** provide the **WRE_Status** facility.

The **WRE_Status** facility has the following definition:

0b = WRE not present

1b = WRE present

13.5 CHASSIS FEATURES

13.5.1 CHASSIS FEATURE REQUIREMENTS

This section defines the required and optional features for various types of Chassis which will accept InfiniBand Modules. Those indicated as required ensure interoperability between such a Chassis and Modules defined by this specification.

The column cells of Table 129 **shall** have the following definition:

- **Req** - A Chassis **shall** implement the facility labeled **Req** in the "Requirement" column.

- **Rec** - The Chassis **may** implement the features labeled **Rec** in the "Requirement" column. It is recommended that the feature be provided. If the feature is provided, the Chassis **shall** follow the provisions of the feature description section.
- **Opt** - The Chassis **may** implement the features labeled **Opt** in the "Requirement" column. If the feature is provided, the Chassis **shall** follow the provisions of the feature description section.
- **N/A** - The feature is not applicable to the Chassis being addressed.

C13-28: Removed since it is a null set.

C13-29: An InfiniBand defined Chassis not containing a Chassis Management Entity which accepts at least one InfiniBand defined Module and provides an IB-ML to it **shall** provide those features and the dependent facilities labeled as "**Req**" in the "Passively Managed" column of Table 129.

C13-30: An InfiniBand defined Chassis containing a Chassis Management Entity which accepts at least one InfiniBand defined Module and provides an IB-ML to it **shall** provide those features and the dependent facilities labeled as "**Req**" in the "Actively Managed" column of Table 129.

Table 129 Chassis Feature Requirements

Register	Unmanaged	Passively Managed	Actively Managed
IB-ML	N/A	**Req**	**Req**
IB-ML Slave Support	N/A	**Req**	**Req**
CME	N/A	N/A	**Req**
Module Power Control	N/A	N/A	Opt
Module Removal Control	N/A	N/A	Opt
IB-ML Selector Proxy	N/A	Opt	Opt
ChassisInfo Device	N/A	**Req**	**Req**

13.5.2 CHASSIS FEATURE DESCRIPTIONS

13.5.2.1 IB-ML

IB-ML is a multi-drop, multi-master, two-wire serial bus which uses signaling and arbitration protocols similar to that of SMBus 1.1 bus [24]. It primarily allows for access to defined facilities on the Module from the

Chassis, but it also allows for certain defined operations to be sourced from the Module to the Chassis.

Architectural Note

IB-ML is architecturally considered "point-point" in that the address space from one IB-ML port is not shared with the address space of any other port. This allows for the same device address assignments to be used on each Module without having to resort to dynamic address assignment techniques.

An IB-ML is made up of 2 Backplane Connector defined signals: **IMxClk** and **IMxDat**. See Section 11.5.1, "IMxClk, IMxDat," on page 448 and Annex A2: IB-ML Design Guidelines for the electrical details of the IB-ML signals.

C13-31: A managed Chassis **shall** implement an IB-ML on at least the Primary Backplane Connector (Port 1) of all Slots (See Section 12.5, "Chassis Power Rules," on page 470). It **may** implement IB-MLs on other available Backplane Connectors; if so, it **shall** keep the multiple IB-MLs electrically isolated from each other.

See Section 11.5, "System Management Group," on page 448 and Annex A2: IB-ML Design Guidelines for the details of the IB-ML signals.

13.5.2.2 IB-ML SLAVE SUPPORT

Logical slave devices on the IB-ML provide access to the ChassisInfo Device (See Section 13.5.2.7, "ChassisInfo Device," on page 548 and, optionally, to a CME (See Section 13.5.2.3, "CME," on page 546).

C13-32: A Passively Managed or an Actively Managed Chassis **shall** respond as IB-ML slaves for Read and Write access to the device addresses labeled as "**Req**" in the "Access" column of Table 130.

C13-33: A Passively Managed or an Actively Managed Chassis **shall not** respond as IB-ML slaves for Read and Write access to the device addresses labeled as Not Permitted "**NP**" in the "Access" column of Table 130.

A Passively Managed or an Actively Managed Chassis **may** respond as IB-ML slaves for Read and Write access to the following device addresses labeled as “Opt” in the “Access” column of Table 130.

Table 130 Chassis IB-ML Slave Addresses

Function	Device Address[a]		Access	Reference
	Hex	Binary		
ModuleInfo Device	A0h	1010_000b	NP	Section 13.3.2.10 on page 517
Reserved for future IB definition	A2h	1010_001b	NP	
Reserved for future IB definition	A4h	1010_010b	NP	
Module Specific Use	A6h	1010_011b	NP	
Module Specific Use	ACh	1010_110b	NP	
Module Specific Use	AEh	1010_111b	NP	
Module Specific Use	40h-46h 50h-5Eh 90h-9Eh BCh-BEh C2h DCh-DEh		NP	Section 13.3.2.2 on page 495
ChassisInfo Device	A8h	1010_100b	**Req**	Section 13.5.2.7 on page 548
Chassis Specific Use	AAh	1010_101b	Opt	
CME Slave	E8h	1110_100b	**Req**[b]	Section 13.5.2.3.1 on page 547
MME Function	E0h	1110_000b	NP	Section 13.6.5 on page 581
MME IbML2Ib	E2h	1110_001b	NP	Section 13.6.5 on page 581
MME Command	E4h	1110_010b	NP	Section 13.3.2.3.3 on page 498

a. This 7-bit Binary notation for IB-ML slave addresses leaves out the Write/Read direction bit which is encoded as the least significant bit on the IB-ML bus during the address phase. The accompanied Hex notation follows this Binary notation by left justifying the 7-bit IB-ML address to fit into an 8-bit wide byte and setting the list significant bit to 0.
b. Not required for Passively Managed Chassis.

13.5.2.3 CME

C13-34: A system with an actively managed Chassis **shall** include a Chassis Management Entity (CME).

The CME has point-to-point connections via the IB Management Link (IB-ML) to each Module it supports. The CME **may** optionally have access to Switches within the Chassis. The CME and the Baseboard Manager use these links to move management data and commands between the CME, Modules and Switches.

The CME **may** have access to all Backplane Connector signals such as **VBxEn_L, IMxPReq_L, IMxInt_L,** and **IMxPRst**.

C13-35: The CME in an Actively Managed Chassis **shall** implement the CMEInfo Record and its fields that are labeled "**Req**" in CMEInfo Record (Section Table 158 on page 627). The CME **shall** return data for these required fields in response to a ReadVPD CME command from the ports indicated by the CMEAccess field of the SlotInfo.

13.5.2.3.1 CME SLAVE

As an IB-ML slave, the CME responds to CME Slave address listed in Table 130 Chassis IB-ML Slave Addresses on page 546.

13.5.2.4 MODULE POWER CONTROL

The Chassis Management Entity (CME) **may** control a Module power via the **VBxEn_L** pin as part of Chassis specific power sequencing operations. See Section 11.3.4, "VBxEn_L," on page 437 and Section 13.6.12.7, "OutBandModuleCtl," on page 596)

The power control of non-IB devices (I/O controllers) sub-ordinate to TCAs are outside of the scope of InfiniBand specification.

13.5.2.5 MODULE REMOVAL CONTROL

An Actively Managed Chassis **may** choose to implement a Chassis-provided mechanism to hold a Module in place temporarily or permanently. Using the **CME_CTR**, the Chassis **may** block the assertion of the "OK To Remove" indication.

o13-7: If Module Removal Control feature is implemented, the CME **shall** set the **CME_CTR** facility in the MME to 1b when the Module is clear to remove from the Chassis point of view. The CME **may** set the **CME_CTR**

bit to 0b when the Chassis wishes to indicate that the Module is not OK to remove, whether for physical or logical reasons.

Implementation Note

For example, the CME and the Management Software can be enabled to control an Electro-Mechanical Lock which can hold a Module in place. The CME may release the lock in response to a locally generated Request to Remove (e.g. via a Chassis mounted Request to Remove push-button, Chassis front panel keypad, etc.) or remotely in response to **SW_CTR** being asserted from management software. Proprietary IB messages and out-of-band interfaces may also be used to control the lock.

If it is known that the Chassis implements a mechanical lock, and that the mechanical lock condition is the only one that would cause the Chassis to deassert **CME_CTR**, then the state of the **CME_CTR** bit could be used to infer the state of the lock."

13.5.2.6 IB-ML SELECTOR PROXY

A CME can act as a proxy to "route" IB-ML operations to a specified slot's IB-ML using a selector field in an applicable command. (Section 13.6.12, "CME Commands," on page 592). A proxy of this type is necessary for managing non-protocol-aware Modules (i.e. Repeater Modules).

ChassisInfo.CMEAccess provides an indication as to whether this feature is present (See Section 13.5.2.7, "ChassisInfo Device," on page 548).

13.5.2.7 CHASSISINFO DEVICE

The ChassisInfo Device holds data that provide Vital Product Data (VPD) and other capability information such as **ChassisGUID** that are necessary to allow for the Hardware Management of a Chassis as it pertains to InfiniBand. ChassisInfo includes readable and write-protect areas. The ChassisInfo Device consists of one base VPD device and zero or more extension VPD devices. The base VPD device always resides at the ChassisInfo Device address specified in Table 130: Chassis IB-ML Slave Addresses.

These areas are readable and writable via IB-ML and IB.

C13-36: For every IB-ML-connected Slot, Managed Chassis **shall** implement at least 64 bytes of non-volatile writable area in ChassisInfo out of which the first 32 bytes are reserved for the Module and the second 32 bytes are reserved for the Baseboard Manager. It is recommended that the Chassis implement an additional 64 non-volatile writable bytes for OEM-software. The Chassis **may** implement this writable area in the base VPD device or in an extension VPD device

C13-37: In a ChassisInfo Device, a Managed Chassis **shall** implement the ChassisInfo Record and its fields that are labeled "**Req**" in Table 153 ChassisInfo Record - Record Format Version=2 on page 613. The ChassisInfo Record shall immediately follow the Device Header.

13.5.2.7.1 SLOTINFO

C13-38: A Managed Chassis **shall** implement this field to provide slot specific information. This element consists of **SlotNumber** and **CMEAccess**. See ChassisInfo Record - Record Format Version=2 (Section Table 153 on page 613) for the details.

13.6 MANAGEMENT COMMANDS

This section provides an outline for a transport mechanism from the InfiniBand (IB) to the InfiniBand Management Link (IB-ML) and vice-versa.

13.6.1 IB-TO-MME (BASEBOARD MAD) COMMANDS

13.6.1.1 GENERIC MANAGEMENT REQUESTS AND RESPONSES

InfiniBand Architecture Specification, Volume 1, Chapter "General Services", Section "Baseboard Management", Subsection "MAD Fields" defines the primary fields that are used for a baseboard management request MAD (Management Datagram) packets transmitted and received via the InfiniBand media.

Baseboard Management requests and responses are delivered using the ***BMSend*** method described in *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "BM Methods". Similarly, baseboard management traps are delivered using the ***BMTrap*** method. BMTraps may be suppressed by the receipt of a MAD having the **BMTrapRepress** method (See *InfiniBand Architecture Specification, Volume 1*, Chapter "Management Model", Section "Traps").

Typically, remote software will issue a ***BMSend*** with an attribute value and data parameters that select a particular request (command) to be executed on the managed entity's MME. This will normally be followed by a matching ***BMSend*** containing a response with a *completion status* code that indicates whether the request was accepted. The response may also hold additional parameter data based on the type of request that was issued.

The ***BMSend*** method is also used for delivering requests to the CME via an IB -to- IB-ML message to the MME. The CME can then format a corresponding ***BMTrap*** message for the response and return it to the requester using an IB-ML -to- IB mechanism. (See Section 13.6.9.1, "SendBMTrap," on page 588)

Figure 166 depicts the structure of the a Baseboard Management MAD as defined in *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "Baseboard Management". The Common Header portion is defined in *InfiniBand Architecture Specification, Volume 1*, Chapter "Management Model", Section "Management Datagrams" and is repeated as Figure 167 for convenience[4].

Figure 166 Baseboard Management MAD Format

Offset	Byte 0	Byte 1	Byte 2	Byte 3
0	Common MAD Header			
...				
20				
24	B_Key			
28				
32	Reserved			
...				
60				
64	Baseboard Management Data			
...				
252				

Figure 167 Common MAD Header Format

bytes					
0	BaseVersion	MgmtClass	ClassVersion	R	Method
4	Status		ClassSpecific		
8	TransactionID				
12					
16	AttributeID		Reserved		
20	AttributeModifier				

4. *InfiniBand Architecture Specification, Volume 1* **shall** take precedence over Figure 167 should differences exist.

Table 131 outlines the baseboard management MAD usage of the common MAD fields. Refer to *InfiniBand Architecture Specification, Volume 1*, Chapter “Data Packet Format and Chapter “Management Model”, Section “Management Datagrams” for additional information on packet format and field offsets.

Table 131 Common MAD Field usage for Baseboard Management

Field Name	Length (bits)	Description
BaseVersion	8	Version of MAD base format. Set to 1.
MgmtClass	8	Designates Baseboard Management class. Value is defined in *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "Management Datagrams".
ClassVersion	8	Version of MAD class-specific format. Set to 1 for this version of the specification.
R	1	This field is used in combination with Method to differentiate “Send” methods from “Response” methods. 0 indicates a request, 1 a response. Baseboard Management Commands are only transported using “datagram” send methods ***BMSend*** and ***BMTrap***; these methods only produce "requests" as represented by this bit. Other methods supported by the Baseboard Management Class do produce "responses" (i.e. BMGet, BMSet, etc.). The Baseboard Management commands utilize the Attribute Modifier to communicate command level request / response indications.
Method	7	Baseboard management commands use the ***BMSend*** Method. Traps use the ***BMTrap*** Method. The other valid methods are defined in *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "Baseboard Management".
Status	16	This field is only valid for Methods with the R bit set. Thus, this is always 0000_0000h for ***BMSend*** and ***BMTrap*** Methods.
ClassSpecific	16	Reserved for this version of the specification.

Table 131 Common MAD Field usage for Baseboard Management

Field Name	Length (bits)	Description
TransactionID	64	For baseboard management commands from software, the Transaction ID is incremented indicate a new instance of a request at the IB packet level. For retried requests, the Transaction ID is not incremented, but instead Transaction ID from the initial request is repeated. The MME maintains its own Transaction ID counter for outgoing requests that it generates, including messages generated using the **IbML2BM** (IB-ML -to Baseboard Manager) functionality. This counter is also incremented for each new request, and repeated for retried requests. The MME is allowed to restart its outgoing Transaction ID value from 0 whenever the communication on the Module's InfiniBand interface is activated. It is recommended that the MME keep the outgoing Transaction ID on standby power if possible in order to reduce this occurrence. **BMSend** and **BMTrap** Methods are unacknowledged datagrams. The actual matching of Baseboard MAD requests and responses occurs at the Baseboard Management command level, using the BMSequenceNumber that is part of the Baseboard Management Data field in **BMSend** and **BMTrap** requests.
AttributeID	16	This field holds a value that identifies the baseboard management "command" to be performed. Table 136: Baseboard MAD Commands provides an overview list of the commands and their associated request and response parameters. The BM Attributes defined in *InfiniBand Architecture Specification, Volume 1,* Chapter "General Services", section "BM Attributes" provides the attribute ID values assigned to the individual commands.
Reserved	16	
AttributeModifier	32	[31:1] Reserved. 0000_0000_0000_000h. [0] Response/Request or BMTrap 0b = Attribute is for a Baseboard Management Command *request*. 1b = Attribute is for a Baseboard Management Command *response* or BMTrap.

Table 132 B_Key Field usage for Baseboard Management

Field Name	Length (bits)	Description
B_Key	64	A 64-bit (8-byte) key that is used to protect Baseboard Management operations. This specification allows more than one B_Key in a partition. The Baseboard Manager sets Nodes' B_Keys using BMSet(BKeyInfo). A Node **shall** include its B_Key in every Baseboard Management MAD that it emits. The Method and the **Attribute Modifier** determine how the B_Key is handled. If the message is a *request*, it **shall** only be accepted if its B_Key matches the B_Key set earlier by the Baseboard Manager via a BMSet(BKeyInfo). If the message is a *response* or a BMTrap, by default, the receiver **shall** ignore the B_Key. An implementation **may** include a vendor-specific option to check the B_Key in responses. This is a provision to allow an MME associated with a Baseboard Manager or a system software to check B_Keys on responses, if desired. See *InfiniBand Architecture Specification, Volume 1,* Chapter "General Services", section "B_Key General Use" and section "BKeyInfo". Also see Section 13.6.1.1, "Generic Management Requests and Responses," on page 549 for more information.

Table 133 Baseboard Management Data Field

Field Name	Length (bits)	Description
BMSequence	16	This value is used to track requests and responses at the Baseboard MAD Command level. The value is incremented (with wrap-around to 0000h at FFFFh) for new BMTraps and BMSend requests, and repeated for retried traps and requests. For responses, the value that was passed in the request is returned in the response. The MME maintains its own BMSequence number that is used to tag requests that it generates (typically as the result of generating an IB-ML -to- Baseboard Manager transaction).
BMSourceDevice	8	Identifies the source of the Baseboard Management Command request or response. The MME and CME values are based on the MME and CME 7-bit slave addresses in bits [7:1] of this field. By convention, bit 0 is always 0b. 00h = Source is unspecified. E0h = source is the MME. E8h = source is the CME. (Used when CME generates Baseboard Management request or responses. FEh = source is Baseboard Manager.
BMParmCount	8	Indicates the number of parameter bytes in the BMParameters field. 1-based. 0 means no additional command parameter data follows.
BMParameters	0 to 40*8	Baseboard MAD Command request and response parameters for the given Attribute ID. Refer to Table 136: Baseboard MAD Commands for the parameter definition and offsets of the bytes used within this field, and for command-specific completion status values. Refer to Table 135: Completion Status Values for completion status values that can be used with any Baseboard MAD Command response. Background: The 40 byte maximum specified is for a 32 byte data write using the command with the largest command header / trailer information per Table 136: Baseboard MAD Commands (WriteVPD). 40B = 1B (Device Selector) + 2B (# of bytes to write (NBW) + 2B (Offset) + 1B (CMD) + 1B (Length) + 32B (Data) + 1B (PEC)
Reserved	var	Fill to produce the required 192 byte MAD Data Field. Content is unspecified. Number of bits equal 1504 - BMParameters field size (in bits). Background: 1504 = 1536 (192B*8) - 32 (BMSequence (16)+BMSourceDevice(8)+BMParmCount(8))

13.6.1.2 DELIVERING REQUEST MESSAGES TO THE MME

The ***BMSend*** request message is used for both delivering requests to the MME from IB, and responses from the MME to IB. (The request/response formats are actually symmetric, so it would also be possible for the MME to generate a request to IB, and receive a response from IB).

As an example, Table 136 Baseboard MAD Commands on page 570 defines a *GetModuleStatus* attribute that allows Baseboard Manager software to get information such as whether an Attention indication is being asserted by the Module, or whether a fault condition exists. For this example, assume that software will be using the command to retrieve the present state of the Module LEDs. The following outlines the steps that are used for sending this command and getting the corresponding response.

C13-38.1.1: A Protocol Aware module **shall** perform the actions to be performed by either the MME or Node in Section 13.6.1.2.

1) Baseboard Management software formats up the *request* by filling in the appropriate fields of the *request* message:

   - The Baseboard Management Software **shall** populate the source and destination queue-pair addressing and Q_Key information for the channel that routes the *request* from the Baseboard Manager to the Module.
   - The Baseboard Management Software **shall** set the **Method** field to the value for ***BMSend*** (see *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”, Section “BM Methods”).
   - The IB packet communication function associated with the Baseboard Management Software **shall** set the **Transaction ID** field. This function **shall** increment the transaction ID field for each new *request*, and **shall** leave it the same for retried requests.
   - The Baseboard Management Software **shall** set the **Attribute ID** field to identify which baseboard management command it is performing.

     For this example, the **Attribute ID** field will be set to the value for the *GetModuleStatus* attribute (see *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”, Section “BM Attributes”). Table 136 Baseboard MAD Commands on page 570 lists the parameters and bit-fields associated with the baseboard management attributes. Table 136 is followed by command descriptions for each Attribute ID. The actual values for the Attribute IDs are specified in the “BM Attributes” section in the Management Chapter of the InfiniBand™ Architecture Specification, Volume 1.

- The Baseboard Management Software **shall** set the least significant bit of the **Attribute Modifier** field to 0b for a *request*.
- The Baseboard Management Software **shall** set the **B_Key** value to the required value obtained from the Configuration Manager (CFM) or other entity managing the **B_Keys**.
- The Baseboard Management Software **shall** set the **BMSequence** and **BMSourceDevice** sub-fields within the **Baseboard Management Data** field. (The Baseboard Management Software **shall** set the **BMSourceDevice** sub-field to FEh to indicate the message is from the Baseboard Manager.)
- The Baseboard Management Software **shall** set the BMParmCount and BMParm sub-fields within the **Baseboard Management Data** field as specified in Table 133: Baseboard Management Data Field and according to the Request Data column for the specific command in Table 136 Baseboard MAD Commands on page 570.

  In this example, **BMParmCount** is set to zero because no additional parameters are required for the *GetModuleStatus* command.

2) The remaining header fields and data integrity check fields **shall** be filled in according to the packet format specifications defined in *InfiniBand Architecture Specification, Volume 1*.
3) The message is delivered via the fabric to the managed IB Module.
4) The Node in the IB Module receives the packet, and passes it to the Baseboard Management Agent (BMA) function. The BMA function identifies Baseboard Management class messages. If the message is valid, the BMA passes the necessary message fields to the MME. The MME **shall** be delivered the **Attribute Modifier** and all InfiniBand packet fields that are necessary to format the *response* back to the requester. (The Transaction ID is not required to be passed to the MME.) The B_Key gets checked if bit 0 of the **Attribute Modifier** is 0b (a *request* message and not a BMTrap). See Table 131 Common MAD Field usage for Baseboard Management on page 551 and Table 132 B_Key Field usage for Baseboard Management on page 553.
5) The MME processes the *request* by examining the values of the **Attribute**, the **Attribute Modifier**, and baseboard management data fields. The MME may ignore unsolicited *responses*.
6) If the requested action is valid and accepted, the MME **shall** format a *response* message with a completion status of 00h = “OK”. Otherwise, the MME **shall** return a non-zero completion status as defined in Completion Status Values (Section Table 135 on page 565). The response will typically be delivered after the command is executed; however, in some cases, such as for SetModulePMControl

command, the execution of a command may prevent the response from being delivered. In these cases, the MME **shall** return the response before acting on the requested command.

7) The MME passes the following fields to allow the BMA to format a ***BMSend*** message with the following content:

   - The MME **shall** set the InfiniBand addressing information for the destination as defined in *InfiniBand Architecture Specification, Volume 1*.

     For this example, the addressing information is for the originator of the *request*.

   - The MME **shall** set the **Method** field to the value for ***BMSend*** (see *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "BM Methods").
   - The MME **shall** set the **R** bit to 0b.
   - The MME **shall** set the **Transaction ID** field to the value from MME's Transaction ID counter.
   - The MME **shall** set the **Attribute ID** field to the value for the specific command attribute to which it is responding.

     For this example: the value for the *GetModuleStatus* attribute (see *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "BM Attributes").
   - The MME **shall** set Bit 0 of the **Attribute Modifier** to1b to indicate a *response.* See Table 131 Common MAD Field usage for Baseboard Management on page 551.
   - The MME **shall** set the **BMSequence** sub-field within the **Baseboard Management Data** field to the **BMSequence** value that was passed in the *request*.
   - The MME **shall** set the **BMSourceDevice** sub-field to E0h to indicate the message is coming directly from the MME functionality.
   - The MME **shall** set the **BMParmCount** and **BMParameter** sub-fields in the Baseboard Management Data field as specified by the *Response Data* column for the *GetModuleStatus* command in Table 136 Baseboard MAD Commands on page 570.

     For this example, there are six bytes of parameters returned, including the completion status, so the BMParmCount field will be 06h and the BMParameter Field will hold the six response data bytes for the *GetModuleStatus* command. According to Table 136: Baseboard MAD Commands, byte six contains the desired LED state information.

8) The Node **shall** add its B_Key and complete the formatting of the message using the ***BMSend*** Method and deliver the formatted message to the IB fabric.

9) The response is then delivered to the originating software and the process is completed.

13.6.1.3 IB-TO-IB-ML ACCESS

C13-38.1.2: Protocol aware modules **shall** provide IB2IBML operations defined in 13.6.1.3 IB-to-IB-ML Access on page 558.

The Management Commands support "low-level" access to devices on an IB-ML. This is accomplished using the *Ib2IbML* command. The command can be used for the following three types of IB-ML transactions:

Atomic Write

Atomic write to an IB-ML device. This is a write of N-bytes of data to the IB-ML device at the specified slave (device) address, as a single transaction. No data is returned from the device. Devices such as SEEPROMs typically use this type of transaction when being written.

Atomic Read

Atomic Read from an IB-ML device. This is a read of M-bytes data from the IB-ML device at the specified slave (device) address, as a single transaction. In this transaction, no parametric data is written to device prior to the read. Because of this, this transaction is typically only used with devices such as simple latches that do not need an "index" value or other write to select the data to be returned.

Atomic Write-Read

Atomic Write-Read to an IB-ML device. This is a write of N-bytes of data followed by a read of M-bytes of data to/from the device at the specified slave (device) address. The write and read transactions are tied together with a "repeated start" to form a single, atomic transaction on the IB-ML. Devices such as SEEPROMs use this type of transaction for supporting random reads, where the write data are used to set an offset value into the SEEPROM from which the read data are to be returned.

The *Ib2IbML* command includes parameters for the number of bytes to write to and the number of bytes to read from the specified device. The three IB-ML transactions are selected according to what values are used for the write and read counts. If a non-zero write-count is used and the read count is set to zero an Atomic Write is performed. If the write count is set to zero and a non-zero number used for the read count an Atomic Read is performed. Lastly, if both values are non-zero an Atomic Write-Read transaction is performed.

Figure 168: IB-to-IB-ML Access illustrates how the MME takes the parameters passed in the *Ib2IbML* command, creates the transaction on the IB-ML, and generates the corresponding response. A random read of three bytes at offset 5Ah from a 24C02-style SEEPROM is shown as the example. This requires a write-read transaction where the offset to read is written to the device and then the data read out.

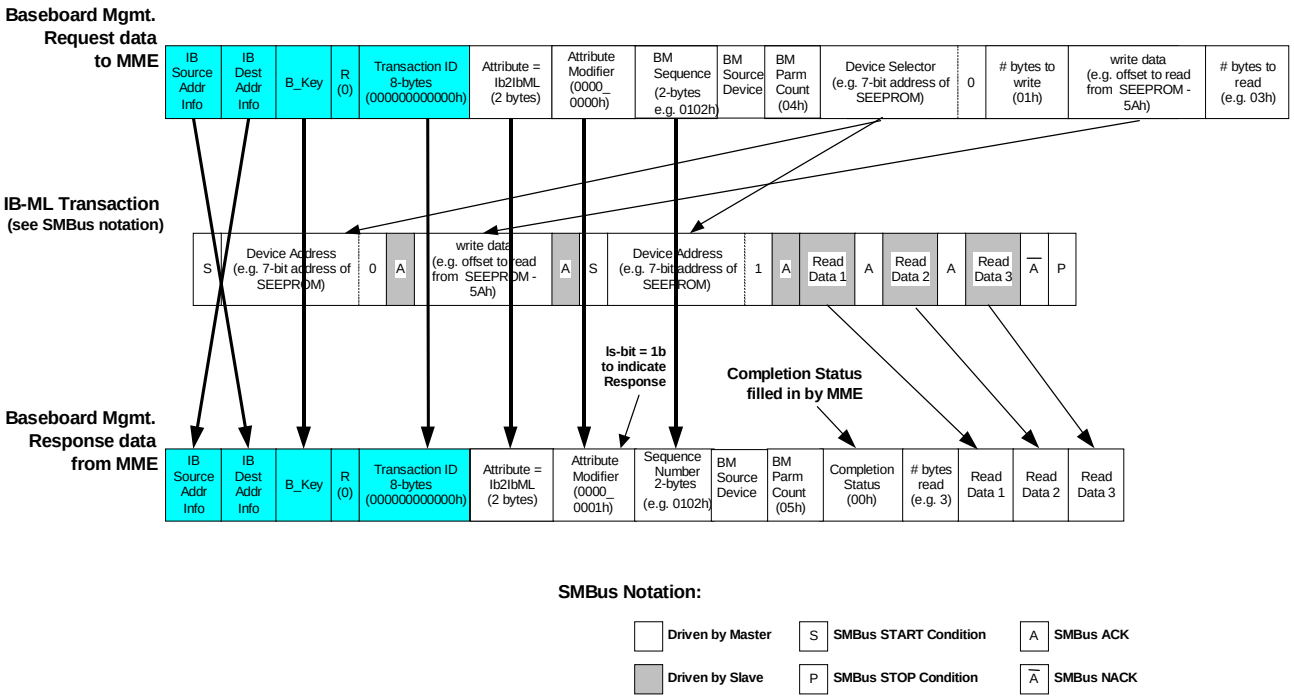


Figure 168 IB-to-IB-ML Access

The thick arrow lines identify field data that the MME essentially copies from the request to the response. The thin arrow lines identify data that are transferred to and from the IB-ML.

Transactions and arbitration on IB-ML **shall** follow the SMBus 1.1 specification.

A transaction is begun with a master (the IB-ML Agent in the MME in this case) arbitrating for the bus and then issuing a START condition followed by the 7-bit slave address of the device it wishes to access. The least significant bit of the slave address byte indicates whether the transfer is a write (0) or a read (1). All bytes transferred are followed by an ACK bit interval. For master writes to a slave, the slave device is responsible for ACK'ing any bytes written to it, including its slave address. If the slave NACKs one of the written bytes, the transaction is terminated by the master and the MME returns an error value for the completion status. Conversely for bytes read by a master from a slave, the master is responsible for ACK'ing the bytes, with the exception of the last read byte which the master NACKs.

The transaction is concluded by the master driving a STOP condition onto the bus. SMBus write and read transactions can be concatenated by is-

suing another START condition and slave address instead of a STOP. This is referred to as a “repeated START”. The IB-ML only specifies an Atomic Write-Read transaction, which is a single write transaction followed by a single read transaction concatenated to the write transaction by a repeated START.

Referring to Figure 168,the following steps outline how SEEPROMs (e.g. VPD Devices) are accessed on the IB-ML. Note that for simplicity, the handling of the ACK bit intervals is left out.

1) The BMA receives the message via the TCA and passes the request data to the MME (See Section 13.6.1.2, “Delivering Request Messages to the MME,” on page 555).
2) The MME interprets the **Attribute** field and finds an *Ib2IbML* request.
3) MME checks the write and read byte counts. If the number of bytes to write and the number of bytes to read are both zero, the request is invalid and the MME **shall** return an error value for the completion status as defined in Completion Status Values (Section Table 135 on page 565).
4) MME gets IB-ML Agent to arbitrate for the IB-ML.

   If the MME cannot win arbitration, times out waiting for a bus free condition, or a bus error occurs during the transaction, it **shall** return a response with an appropriate error value for the completion status as defined in Completion Status Values (Section Table 135 on page 565).

Implementation Note

The MME is not allowed to arbitrate until detecting a bus free condition. An MME implementation can elect to automatically retry arbitrating for the IB-ML on the next bus free condition instead of giving up on the first attempt. This type of implementation is recommended, since it reduces the possibility that software will need to retry the IB request.

5) MME formats the Write data for the IB-ML based on the number of bytes to write.

   If Write Count is not zero, the MME **shall** issue an IB-ML START condition, write the device slave address with LSB=Wr (0), and then write the indicated number of data bytes to IB-ML. If the Write Count is zero, the MME skips to the following step.

   In this example, just one data byte, 5Ah, is written.
6) If the Read Count is not zero, the MME **shall** issue an IB-ML START condition, write the device slave address with LSB=R (1), and clock

in data bytes based on read count. If the Read Count is zero, the MME **shall** skip the following step.

In this example, three bytes are read from the specified IB-ML device.

7) MME **shall** fill in appropriate completion status, address, and R fields as defined in Section 13.6.1.2, "Delivering Request Messages to the MME," on page 555) and forwards the response message data to the BMA for transmission to IB.

8) The BMA waits for an IB transmit opportunity and sends the response.

13.6.1.4 IB-ML SPLIT-TRANSACTION MESSAGING

The CME is accessed using a split-transaction messaging mechanism on IB-ML. The messaging is defined symmetrically; the CME can generate both requests and responses, and a party can send both requests and responses to the CME. An MME may also be accessed using a split-transaction messaging mechanism on IB-ML. See MME Commands (Section 13.3.2.3.3 on page 498).

Messages to the CME can be delivered from Baseboard Management MADs by using a ***BMSend*** method with the ***Ib2CME*** attribute. The *Ib2CME* command is similar to the ***Ib2IbML*** command, except that it only supports writing data to the CME on IB-ML. (The CME's device address is implicit in the command). The data contains the message for the CME.

This specification also defines how the MME would be accessed with the same messaging approach. In this case, the ***Ib2MME*** attribute is used to deliver messages to an MME.

When the ***Ib2CME*** attribute is used to send messages to the CME via the MME, the MME does not check the semantics of the CME message data, (although it can check that the message data meets the maximum IB-ML transaction specifications, see 13.6.1.11: Maximum Transaction / Message Lengths) it simply transfers the data to the IB-ML.

Because of this, and because messaging communication with the CME uses a split-transaction, the immediate response to the ***Ib2CME*** request only indicates that the message data was forwarded to IB-ML and the CME. (If the message data could not be delivered, perhaps because of an IB-ML bus error, or a busy CME, the MME will return an error value for the completion status.)

If the message data held a request for the CME, sometime later, the CME will asynchronously format and send a Baseboard Management MAD using the *SendBMTrap* message. See Section 13.6.9, "MME IbML2Ib Register Descriptions," on page 588.

Messages from IB to the MME are directly handled by the MME, however, the same split-transaction rules apply. The response to the ***Ib2MME*** command indicates that the request was simply transferred to the MME, *not* whether the MME has actually processed the request. A separate response message will be sent from the MME once it has processed the ***Ib2MME*** command.

13.6.1.5 IB-ML SPLIT-TRANSACTION MESSAGE FORMAT

Figure 170 shows the generic format used for requests and responses on IB-ML. For simplicity, the ACK interval is not shown. The most visible difference between requests and response formats is that responses include a completion status field and have the "Rs" bit set to 1. The individual fields are described below.

Request

S	Destination Device ID (e.g. CME address)	0	CMD Set	Length	Source Device ID (e.g. MME address)	Rs (0)	Seq	Data	PEC	P

Response

S	Destination Device ID (e.g. MME address)	0	CMD Set	Length	Source Device ID (e.g. CME address)	Rs (1)	Seq	Data	PEC	P

S = START condition, P = STOP condition. ACK bits not shown.

Figure 169 Generic IB-ML Split-Transaction Format

The positions of Seq and CMD Set fields in the above formats allow all IB-ML "split transaction" messages, regardless of CMD Set, to have a common header. The Seq field in the CME message serves a purpose similar to that for the Transaction ID field in the baseboard management MAD. The Seq field provides a way to differentiate new instances of a request from retried requests, and provide a mechanism for matching up requests with responses at the CME messaging level.

13.6.1.6 GENERIC IB-ML SPLIT-TRANSACTION MESSAGE PARAMETERS

C13-39: An entity that supports IB-ML Split-transaction messages to a CME or MME **shall** have the fields as indicated by Figure 170 and defined in this section.

Destination Device ID

7-bit ID of device/function on IB-ML that is to receive the message. This directly maps to an IB-ML slave address. Typically, this will be either the CME or MME address. The 7-bits of the Destination occupy the most significant bits of the first byte of the message. The least significant bit of the byte is always 0.

Source Device ID

7-bit ID of device/function on IB-ML that is sending the message. This will also typically be the slave address of the CME or MME. For request messages, this field is used to tell where a corresponding response message should be sent. For responses, this identifies the device that originated the response. The 7-bits of the Source Device ID occupy the most significant bits of the second byte of the message. The least significant bit of the byte is the Rs bit, described below.

Rs

1-bit field that differentiates request messages from response message. A 1 indicates a response, 0 indicates a request.

CMD Set

Identifies the command set from which the command is defined. Command sets are:

Table 134 IB-ML Command Sets

Command	Description
00h	IB-ML, InfiniBand Hardware Management.
01h:1Fh	reserved
20h	IPMI Hardware Management
21h	DMTF Pre-OS Working Group (ASF)
23h	OEM - controller specific. The commands are specified relative to the OEM that has specified the functionality of the device that receives the message. The device must respond to a command that retrieves the OEM ID from the device itself, or if not present, the OEM ID from the Chassis info is used.
24h	OEM/Std. Other Standard body. These commands are required to have an OEM ID as the first 3 bytes of the data field in the request, and as the first three bytes of data in the response. The OEM ID can specify a manufacturer or standardization body.
25h	PCI (placeholder)
26h	Compact PCI (placeholder)

Length

1-byte field indicating the number of bytes following the length byte, but excluding the PEC. A length of 1 indicates there is one byte following the length byte.

Seq#

Used to identify different instances of a particular command. Can be used by the requester to match up responses with particular requests. Can be used by destination device to discriminate retries from new requests. The value 00h is used to indicate a request that should never be interpreted as a retry. The Seq field is tracked per Source Device ID.

Data

Data parameters specific to given CMD Set/CMD, or OEM ID if CMD Set = OEM/Std.

PEC

Packet Error Code. This is an 8-bit CRC on all the preceding message bytes, including the initial slave address and read/write bit. The polynomial for the CRC is: $C(x) = x^8 + x^2 + x^1 + 1$ and **shall** be calculated in the order of the bits as received. The PEC algorithm starts off with an initial value of 00h.

o13-8: The PEC field is optional unless otherwise stated for a given command set. If it is present, it **shall** be the first byte after the last byte indicated by the length field.

13.6.1.7 IB-ML CMD SET SPECIFIC FIELDS

C13-40: An entity that supports CMD Set = IB-ML as defined in Table 134 **shall** provide the additional fields defined in Section 13.6.1.7.

CMD

Command. Identifies the requested function under the IB-ML command set. (See Table 140 MME Function Registers on page 584, Table 141 MME IbML2Ib Registers on page 588, Table 123 MME Commands on page 499, and Table 143 CME Commands on page 593). The CMD is the first data byte of both requests and responses.

Completion Status

This is the second byte of the data field of responses with CMD Set = IB-ML. The field indicates whether the request was successfully processed or not. Completion Status values are specified for the individual commands. In addition, there is a set of generic completion status values that can be used. See Table 135 Completion Status Values on page 565.

PEC

Packet Error Code. See Section 13.6.1.6, “Generic IB-ML Split-transaction Message Parameters,” on page 562.

C13-41: The PEC field **shall** be present for all CMD Set = IB-ML messages.

Request

S	Destination Device ID (e.g. CME address)	0	CMD Set = IB-ML	Length	Source Device ID (e.g. MME address)	Rs (0)	Seq	CMD	Data	PEC	P

Response

S	Destination Device ID (e.g. MME address)	0	CMD Set = IB-ML	Length	Source Device ID (e.g. CME address)	Rs (1)	Seq	CMD	Completion Status	Data	PEC	P

S = START condition, P = STOP condition. ACK bits not shown.

Figure 170 IB-ML Message Formats

13.6.1.7.1 GENERIC SPLIT-TRANSACTION COMPLETION STATUS VALUES

Table 135: Completion Status Values, lists completion status values that can be used with any CMD Set = IB-ML or CMD Set = OEM Controller-specific split-transaction response. In addition, Table 135 contains the list of common completion status values that can be returned in Baseboard MAD responses shown in Table 136 and the lists of common completion status values that a CME may return in response to CME commands in Table 143.

Table 135 Completion Status Values

Value	Definition
00h-3Fh	**generic completion status**
00h	OK
01h	Unspecified error
02h	CME busy
03h	MME busy
04h	command not supported
05h	illegal request parameter
06h-3Fh	reserved for generic completion status
40h-BFh	**command-specific completion status**
40h	Write protected
41h	NACK'd
42h	Bus error
43h	Busy (arb loss)
44h	Invalid VPD device selector
45h	Illegal offset
46h	Illegal byte count
C0h-FFh	Vendor-specific Error (defined by vendor identified by the Vendor ID field in the Module info)

13.6.1.8 IPMI/DMTF PRE-OS WORKING GROUP CMD SET FIELDS

o13-9: An entity that supports CMD Sets = IPMI and ASF as defined in Table 134 **shall** have the fields as indicated by Figure 171 on page 566 and defined in Section 13.6.1.8.

Note that the IB-ML defined Completion Status field is not a standard part of these response messages, although equivalent fields may be specified.

Data — Additional data parameters, as specified by the IPMI or DMTF Pre-OS Working Group specifications.

PEC — Packet Error Code.

o13-10: The PEC field **may** be present for all CMD Set = IPMI and ASF messages. If present, it **shall** be as specified in Section 13.6.1.6, "Generic IB-ML Split-transaction Message Parameters," on page 562.

Request

S	Destination Device ID (e.g. CME address)	0	CMD Set = IPMI or ASF	Length	Source Device ID (e.g. MME address)	Rs (0)	Seq	Data	PEC	P

Response

S	Destination Device ID (e.g. MME address)	0	CMD Set = IPMI or ASF	Length	Source Device ID (e.g. CME address)	Rs (1)	Seq	Data	PEC	P

Figure 171 IPMI/ASF Message Formats

13.6.1.9 OEM CONTROLLER-SPECIFIC CMD SET FIELDS

o13-11: An entity that supports CMD Sets = OEM as defined in Table 134 **shall** have the fields as indicated by Figure 172 OEM Controller-specific Message Formats on page 567 and defined in Section 13.6.1.9.

These commands are formatted like CMD Set = IB-ML messages, except the CMD Set = OEM. The meaning of the CMD field is specified by the OEM associated with the controller that accepts the request.

CMD — Command. Identifies the requested function under the OEM command set. The CMD values and semantics are specified by the vendor / organization identified by the ID returned by the GetVendorID CME Command. The CMD is the first data byte of both requests and responses.

Data — Additional data parameters, as specified by OEM.

Completion Status — This is the second byte of the data field of responses with CMD Set = OEM. Command-specific completion status codes are specified by the OEM. In addition, there is a set of generic completion status values that can be used. See Table 135 Completion Status Values on page 565.

PEC — Packet Error Code. See Section 13.6.1.6, "Generic IB-ML Split-transaction Message Parameters," on page 562

o13-12: The PEC field **shall** be present for all CMD Set = OEM messages.

Request

S	Destination Device ID (e.g. CME address)	0	CMD Set = OEM	Length	Source Device ID (e.g. MME address)	Rs (0)	Seq	CMD	Data	PEC	P

Response

S	Destination Device ID (e.g. MME address)	0	CMD Set = OEM	Length	Source Device ID (e.g. CME address)	Rs (1)	Seq	CMD	Completion Status	Data	PEC	P

Figure 172 OEM Controller-specific Message Formats

13.6.1.10 OEM/STD CMD SET SPECIFIC FIELDS

o13-13: An entity that supports CMD Sets = OEM/Std as defined in Table 134 **shall** have the fields as indicated by Figure 173 OEM/Std Message Formats on page 568 and defined in Section 13.6.1.10.

Note that the Completion Status field is not a standard part of these response messages.

OEM ID — Three byte OEM ID. Same format as OEM/Mfr. ID used in GUID. Fields within data portion of message are specified by given OEM/Mfr or standards body identified by the OEM ID.

Data — Additional data parameters, as specified by OEM/Mfr or standards body identified by the OEM ID.

PEC — Packet Error Code.

o13-14: The PEC field **may** be present for all CMD Set = OEM/Std messages. If present, it **shall** be as specified in Section 13.6.1.6, "Generic IB-ML Split-transaction Message Parameters," on page 562.

Request

S	Destination Device ID (e.g. CME address)	0	CMD Set = OEM/Std	Length	Source Device ID (e.g. MME address)	Rs (0)	Seq	OEM ID (3 bytes)	Data	PEC	P

Response

S	Destination Device ID (e.g. MME address)	0	CMD Set = OEM/Std	Length	Source Device ID (e.g. CME address)	Rs (1)	Seq	CMD	OEM ID (3 bytes)	Data	PEC	P

Figure 173 OEM/Std Message Formats

13.6.1.11 MAXIMUM TRANSACTION / MESSAGE LENGTHS

C13-42: An entity that implements IB-ML operations **shall** adhere to the requirements of Section 13.6.1.11

All IB-ML transactions are limited to an overall write transaction length of 36 bytes and an overall read transaction length of 36 bytes. That is to say, there are no more than 36 bytes from START condition to STOP or Repeated Start to STOP.

Architectural Note

Because there is varying overhead required for address bytes, length fields, CMD Set fields, PEC, etc., the maximum amount of data payload will vary. Factors such as proxy accesses and using IB-to-IBML commands also introduce variations in the number of data bytes that can be carried based on the access mechanism.

For the command in (CMD set == IBML), the 36-byte limit allows sufficient number of bytes of data field payload to be carried to the final targeted IB-ML device across all interfaces. The software does not need to adjust payload size based on the path (Ib2IBML, CME proxy, a peer on IBML) it takes to deliver the transaction.

1) All MME's **shall** accept at least 42-bytes of Baseboard Management Data from the BMA. Note that IB addressing fields, Transaction ID, etc. are not part of the Baseboard Management Data field.
2) All MME's **shall** accept a 36-byte (slave address through PEC) IB-ML write for accessing the IbML2Ib register. All other MME write transactions **shall** support at least the number bytes in registers specified in Table 140 MME Function Registers on page 584 and Table 141 MME IbML2Ib Registers on page 588.

3) An MME is allowed to accept write transactions greater than 36 bytes. Until the next STOP condition, an MME **shall** NACK every received byte of a write-transaction once the MME's input buffer is full.
4) All CME's **shall** accept at least 36-bytes written to it as a slave on IB-ML.
5) All MME's that implement MME Command **shall** accept at least 36-bytes written to it as a slave on IB-ML.
6) A CME is allowed to accept write transactions greater than 36 bytes. A CME **shall** NACK the first byte of a write-transaction that exceeds its buffer size. Until the next STOP condition, a CME **shall** NACK every received byte of a write-transaction once the CME's input buffer is full.
7) The writable IB-ML VPD devices **shall** accept write transactions of at least 8 data bytes. The writable IB-ML VPD devices can NACK the write transactions that cause writes to go beyond the device boundary.
8) IB-ML VPD devices **shall** accept read transactions of at least 36 bytes, excluding read transactions that would cause reads beyond the storage limit of the device.
9) All other "proprietary" or unspecified devices on IB-ML are allowed to define their own limits, but designers should be aware that the CME "proxy" access and IB-to-IBML transactions will limit the size of any IB-ML transactions that can be driven through a CME or MME.

Implementation Note

Applications should be aware that proxy operations, used to transfer data between interfaces, will introduce additional overhead that will reduce the data payload that can be transferred.

13.6.2 BASEBOARD MAD COMMANDS (BM MAD ATTRIBUTES)

Table 136 lists the named attributes and bitfields that constitute the "commands" used for managing the Module via the Baseboard Management MADs.

Table 136 Baseboard MAD Commands

Command (Description Section)	Request data	Response data	Facility Reference
Ib2IbML (13.6.3.1)	(3+NBW)x8 bits byte 1: device selector byte 2: # bytes to write (NBW) byte 3 to NBW+2: write data bytes (if NBW is 0, this byte is absent) byte NBW+3: # bytes to read (NBR)	(1+NBR)x8 bits byte 1: completion status[a] bytes 2+: read data bytes (if any)	
Ib2CME (13.6.3.2)	(1+NBW)x8 bits byte 1: # bytes to write byte 2+: write data bytes (see text)	(1+NBR)x8 bits byte 1: completion status[a] bytes 2+: read data bytes (if any)	
Ib2MME (13.6.3.3)	(1+NBW)x8 bits byte 1: #bytes to write byte 2+: write data bytes (see text)	(1+NBR)x8 bits byte 1: completion status[a] bytes 2+: read data bytes (if any)	
OEM (13.6.3.4)	3x8 + vendor-specific number of bits bytes 1:3: vendor ID bytes 4+: vendor-specific data	4x8 + vendor-specific number of bits byte 1: completion status[a] (vendor defined) bytes 2:4: vendor ID bytes 5+: vendor-specific response data	
WriteVPD (13.6.3.5)	(5+NBW)x8 bits byte 1: VPD device selector (address) bytes 2:3 # bytes to write (NBW: 2 bytes, 1-based. 0 allowed) bytes 4:5 VPD offset to write (2 bytes, 0-based) byte 6+: write data bytes (NBW bytes)	8 bits byte 1: completion status[a]	
ReadVPD (13.6.3.6)	5x8 bits byte 1: VPD device selector (address) bytes 2:3 VPD number of bytes to read (NBR: 2 bytes. 1-based. 0 allowed) bytes 4:5 VPD offset to read (2 bytes)	(1+NBR)x8 bits byte 1: completion status[a] bytes 2+: read data bytes (if any)	

Table 136 Baseboard MAD Commands

Command (Description Section)	Request data	Response data	Facility Reference
GetModuleStatus (13.6.3.16)	none	6x8 bits	
		byte 1: completion status[a]	
		byte 2: Removal status	
		7 - Module_OKTR	13.4.2.1
		6 - MME_RTR	13.4.2.4
		5 - MME_CTR	13.4.2.5
		4 - CME_Force_OKTR	13.4.2.6
		3 - CME_RTR	13.4.2.7
		2 - CME_CTR	13.4.2.8
		1 - SW_CTR	13.4.2.9
		0 - SW_RTR	13.4.2.10
		byte 3: Module Attention	
		7:4 - reserved	
		3 - Module_Identify	13.4.3.1
		2 - MME_Module_Attention	13.4.3.3
		1 - CME_Module_Attention	13.4.3.4
		0 - SW_Module_Attention	13.4.3.5
		byte 4: Attention source	
		7:6 - reserved	
		5 - CME_RTR_Trap_Timeout	13.4.4.1
		4 -Power_Converter_Fault	13.4.4.2
		3 -Environmental_Fault	13.4.4.3
		2 - Reserved	
		1 -Functional_Fault	13.4.4.4
		0- Predictive_Fault	13.4.4.7
		byte 5: Power/environmental status	
		7 - Cooling_Non_Crit	13.4.4.8
		6 - Cooling_Crit	13.4.4.9
		5 - Cooling_Non_Recoverable	13.4.4.10
		4 - Voltage_Crit	13.4.4.11
		3 - Voltage_Non_Recoverable	13.4.4.12
		2 - Module_Power_On	13.4.4.13
		1:0 - Pwr_Conv_Redundancy	13.4.5.1
		byte 6: LED Status	
		7 - reserved. Return as 0b	13.4.3.6
		6:5 - Attention_LED_State	
		4:3 - Module_Status_LED_State	13.4.3.7
		2 - Module_Identify	13.4.3.1
		1 - LED_Enable	13.4.3.8
		0 - LED_Test	13.4.3.9

IBTA

Table 136 Baseboard MAD Commands

Command (Description Section)	Request data	Response data	Facility Reference
ResetIbML (13.6.3.7)	none	8 bits byte 1: completion status[a] 01h = could not clear bus	
SetModuleState (13.6.3.14)	8 bits byte 1: Removal Control 7:2 - reserved 1 - SW_CTR 0 - SW_RTR	8 bits byte 1: completion status[a]	13.4.2.9 13.4.2.10
SetModuleAttention (13.6.3.15)	8 bits byte 1: Module Attention 7:4 - reserved 3 - Assert_Identify 2 - reserved 1 - Initiate_LED_Test 0 - SW_Module_Attention	8 bits byte 1: completion status[a]	13.4.3.1 13.4.3.10 13.4.3.5
Power Control [b]			
SetModulePMControl (13.6.3.8)	8 bits byte 1: control 00h = M_{On} 05h = $M_{Standby}$ Others = reserved	8 bits byte 1: completion status[c a]	14.4
GetModulePMControl (13.6.3.9)	none	2x8 bits byte 1: completion status[a] byte 2: Module Power State[d] 00h = M_{On} 05h = $M_{Standby}$ 06h = M_{NoWake} Others = reserved	14.4
SetUnitPMControl (13.6.3.10)	8 bits byte 1: control 00h =U_{On} 04h =U_{Sleep} 05h =$U_{Standby}$ Others = reserved	8 bits byte 1: completion status[a]	14.5

IBTA

Table 136 Baseboard MAD Commands

Command (Description Section)	Request data	Response data	Facility Reference
GetUnitPMControl (13.6.3.11)	none	2x8 bits byte 1: completion status[a] byte 2: Unit Power State 00h =U_{On} 04h =U_{Sleep} 05h =$U_{Standby}$ Others = reserved	14.5
SetIOCPMControl (13.6.3.12)	2x8 bits byte 1: IOC Selector byte 2: control 7 - WRE_Status 6 - WRE_Enable 5:4 - Reserved 3:0 - IOC Power States 0h = $I_{Operational}$ 1h = I_{Uninit} 2h = I_{Doze} 3h = I_{Nap} 4h = I_{Sleep} 5h = $I_{Standby}$ 6h-Fh = reserved	8 bits byte 1: completion status[a]	13.4.5.8 13.4.5.7 14.6
GetIOCPMControl (13.6.3.13)	8 bits byte 1: IOC Selector	2x8 bits byte 1: completion status[a] byte 2: control 7 - WRE_Status 6 - WRE_Enable 5:4 - Reserved 3:0 - IOC Power States 0h = $I_{Operational}$ 1h = I_{Uninit} 2h = I_{Doze} 3h = I_{Nap} 4h = I_{Sleep} 5h = $I_{Standby}$ 6h-Fh = reserved	13.4.5.8 13.4.5.7 14.6

a. The completion status field uses the values defined in Table 135. Additional completion status values that may pertain to a given command are enumerated.

IBTA

b. Power state value assignments used in this table are:

0h = Operational or On
1h = Uninit
2h = Doze
3h = Nap
4h = Sleep
5h = Standby
6h = NoWake

c. The Module **shall** return the *response* to the BM MAD before attempting to perform the requested command. See Section 13.6.1.2, "Delivering Request Messages to the MME," on page 555.

d. The Module is not expected to respond to BM MADs when it is not in M_{On} power state.

13.6.3 BASEBOARD MAD COMMAND DESCRIPTIONS

13.6.3.1 IB2IBML

The ***Ib2IbML*** command provides the facility for driving a low-level IB-ML transaction via a Baseboard Management MAD. There are three types of transaction supported: An Atomic Write, Atomic Read, and Atomic Write-Read. If the transaction is successful, the response to this command will return an "OK" completion status value and the requested IB-ML data (if any). If the MME cannot complete the transaction because of an IB-ML bus error, the corresponding non-zero completion status will be returned in the response. See Section 13.6.1, "IB-to-MME (Baseboard MAD) Commands," on page 549 for more information.

A *request* for an Atomic Write-Read operation specifies the Selector for the targeted Device, the number of data bytes to write (NBW: # bytes to write), NBW bytes of Write data, and the number of data bytes to Read (NBR: # bytes to read). It delivers (3+NBW)x8 bits.

The *response* to an Atomic Write-Read operation returns a *completion status* and NBR bytes of Read data. It returns (1+NBR)x8 bits.

13.6.3.2 IB2CME

The ***Ib2CME*** command provides the facility for delivering a split-transaction request or response message to the CME via IB-ML using a Baseboard Management MAD. The MME takes the parameters of this command and creates the corresponding CME IB-ML message. If MME is able to successfully transmit the message to the CME, the response will contain an "OK" completion status. If the MME cannot perform the transfer because of an IB-ML bus error, the corresponding non-zero completion status will be returned in the response. Note that the response to this command only indicates that the message was successfully transferred to the CME. It does not indicate whether the CME was able to process the message. See Section 13.6.1.4, "IB-ML Split-transaction Messaging," on page 561 for more information.

All protocol-aware Modules provides the Ib2CME command to their implemented physical IB-MLs. Non-Module forms of packaging **may** choose to abstract IB-ML and perform Ib2CME function. See Table 137 Baseboard MAD Command Requirements on page 580.

13.6.3.3 IB2MME

The ***Ib2MME*** command provides the facility for delivering a split-transaction request or response message to the MME using a baseboard management MAD. The transaction is formatted the same way as if it is being delivered to a device on IB-ML, although there is no requirement for it to be physically implemented that way. If MME accepts the message, the response will contain an "OK" completion status. Otherwise, the corresponding non-zero completion status will be returned. Note that the response to this only indicates that the message was successfully transferred to the MME. It does not indicate whether the MME has processed the message. See Section 13.6.1.4, "IB-ML Split-transaction Messaging," on page 561 for more information.

13.6.3.4 OEM

The ***OEM*** optional command provides a mechanism for sending baseboard management MADs to a vendor-specific function within an MME. The request parameters include the id of the vendor that has defined the remaining data parameters in the request. If this attribute is not supported, the MME **shall** provide a response with a completion status value of "Attribute not supported". If the command is supported, but not for the given vendor ID, an "Illegal request parameter" completion status **shall** be returned.

13.6.3.5 WRITEVPD

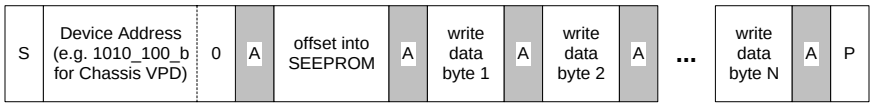
The ***WriteVPD*** command provides the mechanism for writing to different VPD devices, including the ModuleInfo and the ChassisInfo. The device selector parameter in the request holds the "slave address" of the VPD device to be accessed. For Chassis VPD devices, the MME takes the request parameters and formats a physical IB-ML transaction to the Chassis VPD using a 24C02 SEEPROM style access. Figure 174 shows the format of such a SEEPROM write transaction on IB-ML. An example of this type of access is also shown in Figure 168 IB-to-IB-ML Access on page 559.

As long as the baseboard management MAD is handled correctly, software does not care whether Module VPD is implemented as a physical SEEPROM on IB-ML, or not. Thus, Module VPD could be implemented as a physical SEEPROM device on the IB-ML, in which case the MME would format the same type of IB-ML transaction as used for Chassis VPD access. While a different implementation may implement the Module VPD using proprietary hardware, in which case the MME would be responsible

for translating the command and performing whatever steps necessary to perform an equivalent write to the Module VPD.

If the write transaction is performed successfully, The MME **shall** return a response with an “OK” completion status value. Otherwise, the MME **shall** return an appropriate non-zero completion status value.

IB-ML 'SEEPROM Write'



IB-ML 'SEEPROM Read'

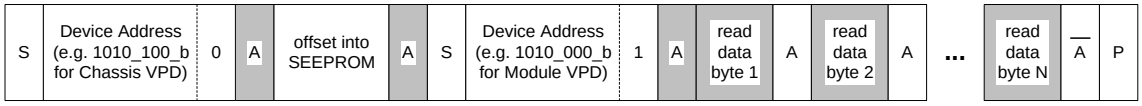


Figure 174 IB-ML SEEPROM Access Formats

Architectural Note

The ***WriteVPD*** Number of Bytes to Write (NBW) field is defined to be 2 bytes so as to 1) allow for larger data abstraction implementations that may not use physical IB-ML devices that are limited to 256B of address ability and 2) to accommodate emerging physical devices that allow for 16KB of addressablity.

13.6.3.6 READVPD

The ***ReadVPD*** command is similar to the ***WriteVPD*** command, except that if the access is successful the response will include the requested VPD data in addition to with the completion status value. See Figure 174 for the IB-ML format of a SEEPROM read.

Architectural Note

The ***ReadVPD*** Number of Bytes to Read (NBR) field is defined to be 2 bytes so as to 1) allow for larger data abstraction implementations that may not use physical IB-ML devices that are limited to 256B of address ability and 2) to accommodate emerging physical devices that allow for 16KB of addressablity.

13.6.3.7 RESETIBML

The ***ResetIbML*** command directs the MME to reset its IB-ML Agent function. If the command is successful, the response **shall** return an "OK" completion status value. Otherwise, the MME **shall** return an appropriate non-zero (error) completion status value.

Implementation Note

Optionally, the MME can attempt to restore a "hung" IB-ML bus state by clocking "1" data bits onto the bus until it sees the data line go high, at which time it can issue a STOP condition. For this operation to be successful, the hung bus condition must not render the IB-ML Agent's master function un-operational.

13.6.3.8 SETMODULEPMCONTROL

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=1b, the ***SETMODULEPMCONTROL*** command **shall** allow access to the current Power Management state for the Module as defined in Section 14.4, "Module Power States," on page 643.

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=0b, the ***SETMODULEPMCONTROL*** command **shall not** cause any updates, not produce any side effects, and return a non-zero completion status value in the response that accompanies this command.

Valid states are indicated through *ModuleInfo.PmCapablity* facility (see Section 13.3.2.10, "ModuleInfo Device," on page 517).

Issuing this command with an unsupported state value **shall not** result in any update to the state.

13.6.3.9 GETMODULEPMCONTROL

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=1b, the ***GETMODULEPMCONTROL*** command **shall** allow access to the current Power Management state for the Module as defined in Section 14.4, "Module Power States," on page 643.

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=0b, the ***GETMODULEPMCONTROL*** command **shall** always produce 00h as the state.

Valid states are indicated through *ModuleInfo.PmCapablity* facility (See Section 13.3.2.10, "ModuleInfo Device," on page 517).

13.6.3.10 SETUNITPMCONTROL

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=1b, the ***SETUNITPMCONTROL*** command **shall** allow access to the current I/O Unit Power Management state as defined in Section 14.5, "I/O Unit Power States," on page 646 for the addressed IOC.

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=0b, the ***SETUNITPMCONTROL*** command **shall not** cause any updates, not produce any side effects, and return a non-zero completion status value in the response that accompanies this command.

Valid states are indicated through *ModuleInfo.UnitPmInfo* facility (see Section 13.3.2.10, "ModuleInfo Device," on page 517).

Issuing this command with an unsupported state value **shall not** result in any update to the state.

13.6.3.11 GETUNITPMCONTROL

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=1b, the ***GETUNITPMCONTROL*** command **shall** allow access to the Power Management state for the I/O Unit as defined in Section 14.5, "I/O Unit Power States," on page 646.

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=0b, the ***GETUNITPMCONTROL*** command **shall** always produce 00h as the state.

Valid states are indicated through *ModuleInfo.UnitPmInfo* facility (see Section 13.3.2.10, "ModuleInfo Device," on page 517).

13.6.3.12 SETIOCPMCONTROL

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=1b, the ***SETIOCPMCONTROL*** command **shall** allow access to the current IOC Power Management state as defined in Section 14.6, "IOC Power States," on page 648 for the addressed IOC.

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=0b, the ***SETIOCPMCONTROL*** command **shall not** cause any updates, not produce any side effects, and return a non-zero completion status value in the response that accompanies this command.

Valid states are indicated through *ModuleInfo.IOCPmInfo* facility (see Section 13.3.2.10, "ModuleInfo Device," on page 517).

Issuing this command with an unsupported state value **shall not** result in any update to the state.

13.6.3.13 GETIOCPMCONTROL

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=1b, the ***GETIOCPMCONTROL*** command **shall** allow access to the Power Management state for the IOC as defined in Section 14.6, "IOC Power States," on page 648.

If *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported*=0b, the ***GETIOCPMCONTROL*** command **shall** always produce 00h as the state.

Valid states are indicated through *ModuleInfo.IOCPmInfo* facility (see Section 13.3.2.10, "ModuleInfo Device," on page 517).

13.6.3.14 SETMODULESTATE

The ***SETMODULESTATE*** command provides update access to the **SW_CTR** and **SW_RTR** facilities in support of Module Graceful Removal. See also Section 13.3.2.4, "Graceful Hot Removal," on page 503.

13.6.3.15 SETMODULEATTENTION

The ***SETMODULEATTENTION*** command provides control access to the Attention LED and to initiate the Identify function. See also Section 13.3.2.6.4, "Attention (Amber) LED," on page 512 and Section 13.3.2.6.7, "Module Identify," on page 513.

13.6.3.16 GETMODULESTATUS

The ***GETMODULESTATUS*** command provides access to the state of the facilities on a Module. The facilities accessible by this command are indicated by the Facility Reference column for GetModuleStatus in Table 136.

13.6.4 BASEBOARD MAD COMMAND REQUIREMENTS

This section defines the required and optional commands for various types of InfiniBand devices and physical packages.

The column cells of Table 137 **shall** have the following definition:

- **Req** - A Module **shall** implement the command labeled **Req** in the "Requirement" column.
- **Rec** - A Module **may** implement the command labeled **Rec** in the "Requirement" column. It is recommended that the command be provided. If the command is provided, the Module **shall** follow the provisions of the command description section. If the command is not provided, the Module **shall** silently drop the MAD containing the command.

IBTA

- **Opt** - The Module **may** implement the facility if so labeled. If the command is provided, the Module **shall** do so in accordance with the provisions of the command description section. If the command is not provided, the MAD containing the command is silently dropped.
- **N/A** - The command is not applicable to the package being addressed. The MAD containing the command is silently dropped.

C13-43: An InfiniBand defined Module containing an InfiniBand protocol-aware device **shall** provide those facilities labeled as "**Req**" in the "Protocol-aware Module" column of Table 137.

C13-44: An InfiniBand defined Module not containing an InfiniBand protocol-aware device **shall** provide those facilities labeled as "**Req**" in the "Non-Protocol-aware Module" column of Table 137.

C13-45: An InfiniBand Channel Adapter device mounted on a package other than an InfiniBand Module **shall** provide those facilities labeled as "**Req**" in the "Non-Module xCA" column of Table 137.

C13-46: An InfiniBand Switch device mounted on a package other than an InfiniBand Module **shall** provide those facilities labeled as "**Req**" in the "Non-Module Switch" column of Table 137.

C13-47: An entity that implements Baseboard MAD Commands **shall** adhere to Section 13.6.1, "IB-to-MME (Baseboard MAD) Commands," on page 549, Section 13.6.2, "Baseboard MAD Commands (BM MAD Attributes)," on page 570, and Section 13.6.3, "Baseboard MAD Command Descriptions," on page 574.

C13-48: An entity that implements IB-ML operations **shall** adhere to the requirements of Section 13.6.1.11, "Maximum Transaction / Message Lengths," on page 568.

Table 137 Baseboard MAD Command Requirements

Register	Protocol-aware Module	Non-Protocol-aware Module	Non-Module xCA[a]	Non-Module Switch[a]
Ib2IbML	**Req**	N/A	Opt	Opt
Ib2CME	**Req**	N/A	Opt[b]	Rec[b]
Ib2MME	Opt	N/A	Opt	Opt
OEM	Opt	N/A	Opt	Opt
WriteVPD	**Req**	N/A	**Req**	**Req**
ReadVPD	**Req**	N/A	**Req**	**Req**

Table 137 Baseboard MAD Command Requirements

Register	Protocol-aware Module	Non-Protocol-aware Module	Non-Module xCA[a]	Non-Module Switch[a]
ResetIbML	**Req**	N/A	Opt	Opt
SetModulePMControl	Opt	N/A	Opt	Opt
GetModulePMControl	Opt	N/A	Opt	Opt
SetUnitPMControl	Opt	N/A	Opt	Opt
GetUnitPMControl	Opt	N/A	Opt	Opt
SetIOCPMControl	Opt	N/A	Opt	Opt
GetIOCPMControl	Opt	N/A	Opt	Opt
SetModuleState	**Req**	N/A	Opt	Opt
SetModuleAttention	**Req**	N/A	Opt	Opt
GetModuleStatus	**Req**	N/A	Opt	Opt

a. (See Section 13.3.2.1, "BMA," on page 495 and Table 120 Module Feature Requirements on page 494) All protocol-aware InfiniBand devices **shall** support ReadVPD and WriteVPD.
b. The IB2CME command is required to perform the IB-ML Selector Proxy function (See 13.5.2.6 on page 548).

13.6.5 MME FUNCTION REGISTERS

IB-ML access to the MME is accomplished by reading and writing to "registers" within a device address that operate on facilities on the Module. Another device address provides for IB-ML to IB messaging whereby message data are provided to the MME which, in turn, forwards correspondingly formatted data to the BMA for delivery to the InfiniBand media via the channel adapter.

Table 138 and Table 139 illustrate the format of the IB-ML atomic Write-Read transactions used to access MME Registers.

Table 138 MME Register Write from IB-ML[a]

IB-ML	data direction
START condition	
MME Slave Address | 0	into MME
Register Selector	into MME
# bytes to read from register = 00h	into MME
data to write to register (N bytes)	into MME
PEC	into MME

Table 138 MME Register Write from IB-ML[a]

IB-ML	data direction
START condition	
MME Slave Address | 1	into MME
Register Selector	out from MME
completion status: 00h = OK 01h = PEC error in write[b] 02h = illegal register selector 03h = illegal write length 04h = illegal read count	out from MME
PEC	out from MME
STOP condition	

a. The rows in this table correspond to successive operations on IB-ML.
b. The MME ACKs the PEC byte in the write-portion of the transaction, even if the PEC is in error. This is done in order to allow it to return an explicit completion status indicating the PEC error was detected.

Table 139 MME Register Read from IB-ML[a]

IB-ML	data direction
START condition	
MME Slave Address | 0	into MME
Register Selector	into MME
# bytes to read from register = N	into MME
PEC	into MME
START condition	
MME Slave Address | 1	into MME

Table 139 MME Register Read from IB-ML[a]

IB-ML	data direction
Register Selector	out from MME
completion status: 00h = OK 01h = PEC error in write[b] 02h = illegal register selector 03h = illegal write length 04h = illegal read count	out from MME
read data from register (N bytes)	out from MME
PEC	out from MME
STOP condition	

a. The rows in this table correspond to successive operations on IB-ML.
b. The MME ACKs the PEC byte in the write-portion of the transaction, even if the PEC is in error. This is done in order to allow it to return an explicit completion status indicating the PEC error was detected.

13.6.6 MME FUNCTION REGISTER SUMMARY

Table 140 lists the named registers, their selector indication, and the fields that are used for managing a Module using the IB-ML Device Address for "MME Function" defined in Table 121 Module IB-ML Slave Addresses on page 496.

Table 140 MME Function Registers

Register (Selector)	Write Data		Read Data	
	Description	Facility Refer	Description	Facility Refer
Interrupt-ClearStatus (00h)	Writing a "0" clears the corresponding interrupt capture flag. Writing a "1" causes no change. byte 1: 7:6 - reserved. Write as 11b. 5 - Attention_Status_Change 4 - Removal_Status_Change 3 - Power_Converter_Fault 2 - Environmental_Fault 1 - Functional_Fault 0 - Predictive_Fault byte 2: 7 - Cooling_Non_Crit 6 - Cooling_Crit 5 - Cooling_Non_Recoverable 4 - Voltage_Crit 3 - Voltage_Non_Recoverable 2 - Module_Power_On 1 - WakeOnIB 0 - WakeOnWRE	 13.4.2.2 13.4.2.3 13.4.4.2 13.4.4.3 13.4.4.4 13.4.4.7 13.4.4.8 13.4.4.9 13.4.4.10 13.4.4.11 13.4.4.12 13.4.4.13 13.4.5.2 13.4.5.3	"1" indicates which source has had a state change captured. If enabled, that change will be contributing to the interrupt signal. byte 1: 7:6 - reserved. Return as 00b. 5 - Attention_Status_Change 4 - Removal_Status_Change 3 - Power_Converter_Fault 2 - Environmental_Fault 1 - Functional_Fault 0 - Predictive_Fault byte 2: 7 - Cooling_Non_Crit 6 - Cooling_Crit 5 - Cooling_Non_Recoverable 4 - Voltage_Crit 3 - Voltage_Non_Recoverable 2 - Module_Power_On 1 - WakeOnIB 0 - WakeOnWRE	 13.4.2.2 13.4.2.3 13.4.4.2 13.4.4.3 13.4.4.4 13.4.4.7 13.4.4.8 13.4.4.9 13.4.4.10 13.4.4.11 13.4.4.12 13.4.4.13 13.4.5.2 13.4.5.3
InterruptSource (01h)	not writable		Presents the present state of the interrupt source. "1" indicates that the corresponding condition is presently asserted, "0" indicates it is presently deasserted. byte 1: 7:6 - reserved. Return as 00b. 5 - Attention_Status_Change 4 - Removal_Status_Change 3 - Power_Converter_Fault 2 - Environmental_Fault 1 - Functional_Fault 0 - Predictive_Fault byte 2: 7 - Cooling_Non_Crit 6 - Cooling_Crit 5 - Cooling_Non_Recoverable 4 - Voltage_Crit 3 - Voltage_Non_Recoverable 2 - Module_Power_On 1 - WakeOnIB 0 - WakeOnWRE	 13.4.2.2 13.4.2.3 13.4.4.2 13.4.4.3 13.4.4.4 13.4.4.7 13.4.4.8 13.4.4.9 13.4.4.10 13.4.4.11 13.4.4.12 13.4.4.13 13.4.5.3

Table 140 MME Function Registers

Register (Selector)	Write Data		Read Data	
	Description	Facility Refer	Description	Facility Refer
InterruptControl (02h)	Writing a "1" enables the corresponding interrupt capture flag to contribute to the interrupt signal. byte 1: 7:6 - reserved. Write as 11b. 5 - Attention_Status_Change 4 - Removal_Status_Change 3- Power_Converter_Fault 2 - Environmental_Fault 1 - Functional_Fault 0 - Predictive_Fault byte 2: 7 - Cooling_Non_Crit 6 - Cooling_Crit 5 - Cooling_Non_Recoverable 4 - Voltage_Crit 3 - Voltage_Non_Recoverable 2 - Module_Power_On 1 - WakeOnIB 0 - WakeOnWRE	 13.4.2.2 13.4.2.3 13.4.4.2 13.4.4.3 13.4.4.4 13.4.4.7 13.4.4.8 13.4.4.9 13.4.4.10 13.4.4.11 13.4.4.12 13.4.4.13 13.4.5.2 13.4.5.3	"1" indicates which interrupt capture flags are enabled. byte 1: 7:6 - reserved. Return as 00b. 5 - Attention_Status_Change 4 - Removal_Status_Change 3 - Power_Converter_Fault 2 - Environmental_Fault 1 - Functional_Fault 0 - Predictive_Fault byte 2: 7 - Cooling_Non_Crit 6 - Cooling_Crit 5 - Cooling_Non_Recoverable 4 - Voltage_Crit 3 - Voltage_Non_Recoverable 2 - Module_Power_On 1 - WakeOnIB 0 - WakeOnWRE	 13.4.2.2 13.4.2.3 13.4.4.2 13.4.4.3 13.4.4.4 13.4.4.7 13.4.4.8 13.4.4.9 13.4.4.10 13.4.4.11 13.4.4.12 13.4.4.13 13.4.5.2 13.4.5.3
LEDTestStatus (03h)	byte 1: 7:3 - reserved. Write as 000000b 2 - Assert_Identify 1 - LED_Enable 0 - Initiate_LED_Test byte 2: Module Attention control 7:2 - reserved 1 - CME_Module_Attention 0 - reserved	 13.4.3.2 13.4.3.8 13.4.3.10 13.4.3.4	byte 1: 7 - reserved. Return as 0b 6:5 - Attention_LED_State 4:3 - Module_Status_LED_State 2 - Module_Identify 1 - LED_Enable 0 - LED_Test byte 2: Module Attention status 7:6 - reserved 5 - Hard_Fault 4 - Degraded_Fault 3 - reserved 2 - MME_Module_Attention 1 - CME_Module_Attention 0 - SW_Module_Attention	 13.4.3.6 13.4.3.7 13.4.3.1 13.4.3.8 13.4.3.9 13.4.4.5 13.4.4.6 13.4.3.1 13.4.3.3 13.4.3.4 13.4.3.5
ModuleControl-Status (04h)	byte 1: 7 - CME_RTR_Trap_Timeout 6:3 - reserved. Write as 00000b 2 - WakeOnIB_Enable 1 - reserved 0 - IMxInt_L_Enable	 13.4.4.1 13.4.5.5 13.4.5.6	byte 1: 7 - CME_RTR_Trap_Timeout 6 - WakeOnIB 5 - WakeOnWRE 4 - IMxInt_L_State 3 - WRE_Enable 2 - WakeOnIB_Enable 1 - WRE_Status 0 - IMxInt_L_Enable	 13.4.4.1 13.4.5.2 13.4.5.3 13.4.5.4 13.4.5.7 13.4.5.5 13.4.5.8 13.4.5.6

Table 140 MME Function Registers

Register (Selector)	Write Data		Read Data	
	Description	Facility Refer	Description	Facility Refer
ModulePower (05h)	byte 1: 7:0 - reserved. Write as 00000000b byte 2: 7:0 - reserved. Write as 00000000b		byte 1: 7:3 - 00000b 2 - M_{On} 1 - $M_{Standby}$ 0 - M_{NoWake} byte 2: 7:2 - 000000b 1:0 - Pwr_Conv_Redundancy	 13.4.5.1
RemovalControl-Status (06h)	byte 1: Removal status 7:5 - reserved. Write as 000b 4 - CME_Force_OKTR 3 - CME_RTR 2 - CME_CTR 1:0 - reserved. Write as 000b	 13.4.2.6 13.4.2.7 13.4.2.8	byte 1: Removal status 7 - Module_OKTR 6 - MME_RTR 5 - MME_CTR 4 - CME_Force_OKTR 3 - CME_RTR 2 - CME_CTR 1 - SW_CTR 0 - SW_RTR	 13.4.2.1 13.4.2.4 13.4.2.5 13.4.2.6 13.4.2.7 13.4.2.8 13.4.2.9 13.4.2.10
Reserved for IB (7Fh:07h)				
OEM (FFh:80h)				

13.6.7 MME FUNCTION REGISTER DESCRIPTIONS

13.6.7.1 INTERRUPTCLEARSTATUS

The *INTERRUPTCLEARSTATUS* register provides a CME access to the facilities listed for InterruptClearStatus in Table 140 to perform the “Clear” function upon a “write” and provide state change “Status” upon a “read” as defined in Section 13.3.2.8, “IMxInt_L,” on page 515.

13.6.7.2 INTERRUPTSOURCE

The *INTERRUPTSOURCE* register provides a CME access to the facilities listed for InterruptSource in Table 140 to get current interrupt source status upon a “read” as defined in Section 13.3.2.8, “IMxInt_L,” on page 515.

13.6.7.3 INTERRUPTCONTROL

The *INTERRUPTCONTROL* register provides a CME access to the facilities listed for InterruptSource in Table 140 to control the **IMxInt_L** contribution enabling as defined in Section 13.3.2.8, “IMxInt_L,” on page 515.

13.6.7.4 LEDTESTSTATUS

The *LEDTESTSTATUS* register provides a CME access to the Module LEDs defined in Section 13.3.2.6, "Module Indicators (LEDs)," on page 508 using the facilities listed for LEDTestStatus in Table 140.

13.6.7.5 MODULECONTROLSTATUS

The *MODULECONTROLSTATUS* register provides a CME access to the state of the facilities listed for ModuleControlStatus in Table 140.

13.6.7.6 MODULEPOWER

The *MODULEPOWER* register provides a CME access to the current Module Power Management state defined in Section 14.4, "Module Power States," on page 643.

13.6.7.7 REMOVALCONTROLSTATUS

The *REMOVALCONTROLSTATUS* register provides a CME access to the facilities listed for RemovalControlStatus in Table 140 for use in supporting the functions described in Section 13.3.2.4, "Graceful Hot Removal," on page 503.

13.6.8 MME IbML2Ib REGISTER SUMMARY

Table 141 lists the named registers, their selector indication, and the fields that are used for managing a Module using the IB-ML Device Address for

IBTA

"MME IbML2Ib" defined in Table 121 Module IB-ML Slave Addresses on page 496.

Table 141 MME IbML2Ib Registers

Register (Selector)	Write Data		Read Data	
	Description	Facility Refer	Description	Facility Refer
SendBMTrap (00h)	byte 1: BMTrapDataLength - number of BMTrapData bytes to include in trap byte 2: BMTrapType 00h = Generic MME 01h = OEM MME 02h = CME_RTR 03h = WRE 04h = Generic CME 05h = OEM CME All others - reserved bytes 3:5 - BMTrapTypeModifier varies based on trap type: holds 3-byte OEM ID if Trap Type is OEM MME or OEM CME. holds SlotSelector, 00h, 00h if trap type is CME_RTR holds 00h,00h,00h (reserved) if Trap Type is Generic MME, WRE, or Generic CME byte 6 + - BMTrapData =data bytes for Trap Type of OEM MME or OEM CME		byte 1: completion status 01h = IB buffer not available (busy) 02h = IB media not ready 03h = Invalid length	
SendIbML2BM (02h)	byte 1: BMDataLength - number of data bytes to include in the Baseboard Management data field of the BM MAD starting after the BMSequence Field byte 2+ - BMData - data for BM MAD		byte 1: completion status 01h = IB buffer not available (busy) 02h = IB media not ready 03h = Invalid length	

13.6.9 MME IBML2IB REGISTER DESCRIPTIONS

13.6.9.1 SENDBMTRAP

The *SENDBMTRAP* register allows the CME to send a "BMTrap" method message to the Baseboard Management LID. The Module uses the components of the ***BM.ClassPortInfo*** and the "BMTrapType" and "BMTrapData" bytes (See SendBMTrap in Table 141) for the "Baseboard Management Data" field to form the datagram message as defined in *InfiniBand Architecture Specification, Volume 1*, Chapter "General Services", Section "Baseboard Management". The value indicated in the BMTrapDataLength byte (See SendBMTrap in Table 141) must not ex-

ceed the “Baseboard Management Data” field length specified in *InfiniBand Architecture Specification, Volume 1*.

Architecture Note

This register is intended to be used by a CME to communicate with the Baseboard Manager and should be used in lieu of the *SendIBML2BM* register. A node running the Baseboard Manager can use the method in the Baseboard class MAD (BMTrap) to direct the MAD to the Baseboard Manager versus the node’s BMA as a BMA is only to act upon BMSend methods with the R bit being 0 (requests). See Table 131 Common MAD Field usage for Baseboard Management on page 551.

13.6.9.2 SENDIBML2BM

The *SENDIBML2BM* register allows the CME to send a “BMSend” method message to the Baseboard Management LID. The Module uses the components of the ***BM.ClassPortInfo*** and the “IBData” bytes for the “Baseboard Management Data” field to form the datagram message as defined in *InfiniBand Architecture Specification, Volume 1*, Chapter “General Services”, Section “Baseboard Management”. The value indicated in the IBLength byte (See SendIbML2BM in Table 141) must not exceed the “Baseboard Management Data” field length specified in *InfiniBand Architecture Specification, Volume 1*.

This register is not recommended for use and is subject to removal in future versions of the specification. A CME should use the SendBMTrap register with "OEM CME" BMTrapType to communicate with the Baseboard Manager. BMTrapType of "Generic CME" is defined as a placeholder for potential future definition of IB2CME.

13.6.10 MME REGISTER REQUIREMENTS

This section defines the required and optional registers for various types of InfiniBand devices and physical packages.

The “Requirement” column cells of Table 142 **shall** have the following definition:

- **Req** - A Module **shall** implement the facility labeled **Req** in the “Requirement” column. An implementation may elect to make MME facilities available either as appropriate physical addresses (at E0h and E2h) or via MME Commands at address E4h. An implementation may select which registers it makes available via a given access mechanism. If MME Function Registers are not available at address E0h, then they all shall be available via MME Commands at address E4h. If MME IbML2Ib Registers are not available at address E2h, then they all shall be available via MME

commands at address E4h. An implementation is allowed to make registers available via both the physical and MME command interfaces.

- **Rec** - A Module **may** implement the command labeled **Rec** in the "Requirement" column. It is recommended that the command be provided. If the register is provided, the Module **shall** follow the provisions of the register description section. If the register is not provided, the Module **shall** provide the default values for read accesses that are specified for the facility bit(s); write accesses **shall not** cause updates to occur from the default value for the facilities within the register.
- **Opt** - The Module **may** implement the facility if so labeled in the "Requirement" column. If the register is provided, the Module **shall** do so in accordance with the provisions of the register description section. If the register is not provided, the Module **shall** provide the default values for read accesses that are specified for the facility bit(s); write accesses **shall not** cause updates to occur from the default value for the facilities within the register.
- **N/A** - The register is not applicable to the package being addressed. Reads from the register **shall** return all 0s for the bits designated; writes to the register **shall not** have any effect.

C13-49: This compliance statement is obsolete and has been replaced by C13-49.1.1:.

C13-49.1.1: An InfiniBand defined Module containing an InfiniBand protocol-aware device **shall** provide those facilities labeled as "**Req**" in the "Protocol-aware Module" column of Table 142.

C13-50: An InfiniBand defined Module not containing an InfiniBand protocol-aware device **shall** provide those facilities labeled as "**Req**" in the "Non-Protocol-aware Module" column of Table 142.

C13-51: An InfiniBand protocol-aware device mounted on a package other than an InfiniBand Module **shall** provide those facilities labeled as "**Req**" in the "Non-Module Device" column of Table 142.

o13-15: An InfiniBand defined Module that implements the facilities labeled as "**Rec**" or "**Opt**" in Table 142 **shall** follow the described specification in Section 13.6.7, "MME Function Register Descriptions," on

IBTA

page 586 and 13.6.11 MME Facilities in Power Management States on page 591.

Table 142 MME Function Register Requirements

Register	Protocol-aware Module	Non-Protocol-aware Module	Non-Module Device
InterruptClearStatus	**Req**	**Req**	Opt
InterruptSource	**Req**	**Req**	Opt
InterruptControl	**Req**	**Req**	Opt
LEDTestStatus	**Req**	**Req**	Opt
ModuleControlStatus	**Req**	**Req**	Opt
ModulePower	**Req**	**Req**	Opt
RemovalControlStatus	**Req**	**Req**	Opt
SendBMTrap	**Req**	N/A	Opt
SendIbML2BM	**Opt**	N/A	Opt

13.6.11 MME FACILITIES IN POWER MANAGEMENT STATES

o13-51.1.1: A Module that sets its *ModulePowerInfo.ModulePMCapability.IsPowerManagedSupported to 1b* **shall** provide the following facilities in power management states:

- If MME commands are implemented via IB-ML (at E4h), the GetCMDSetVersion command **shall** be available from IB-ML in all power-managed states.
- If the MME Function Registers are accessible using MME commands via IB-ML, the ReadMMERegister and WriteMMERegister commands **shall** be available from IB-ML in all power-managed states. In this case, the commands **shall** accept the selectors for the LEDTestStatus, ModulePower, and RemovalControlStatus registers.
- The module **shall** return an 'illegal request parameter' completion status if other IB-specified register selectors are used in the ReadMMERegister and WriteMMERegister commands, but the register is unavailable due to the power state.
- Similarly, if MME Function Registers are accessible via E0h, an 'illegal request parameter' completion status **shall** be returned for register selectors where the register is not available in a given power state.

- For both E0h and E4h accesses, if register access is provided in a given power state, all facilities for the selected register **shall** be functional as specified in Section 13.6.6: MME Function Register Summary.

13.6.12 CME COMMANDS

CME commands are split-transaction IB-ML request and response transactions, as described in section 13.6.1.4 IB-ML Split-transaction Messaging on page 561. CME commands are part of the InfiniBand Hardware Management command set (Section 13.6.1.7, "IB-ML CMD Set Specific Fields," on page 564). The CME performs the requested operation, formats a response using the sequence number that was passed in the request.

For the standard CME commands (commands other than OEM CMD messages), the Source Device ID in the request determines where the CME delivers the response to the commands. A request that comes from the MME's IbML2Ib function is assumed to have come from the Baseboard Manager. If the Source ID in the request matches the MME IbML2Ib slave address (Table 130: Chassis IB-ML Slave Addresses), the CME **shall** send the response back to the Baseboard Manager using the *SendIBML2BM* MME register or via the SendIbML2BM MME command, depending on which is available. Otherwise, the response **shall** be sent as an IB-ML split-transaction response to the slave address specified by the Source Device ID in the request.

OEM CME commands are allowed to include parameters that override returning the response to the Baseboard Manager and direct the CME to return the response via other mechanisms.

C13-52: Actively Managed Chassis **shall** implement a CME with the features described in Table 143.

C13-52.1.1: Actively Managed Chassis **shall** implement a CME with the features described in 13.2.6 Chassis Management on page 489.

Table 143 defines the split-transaction commands for CMD Set = IB-ML sent to the CME Slave Address. See Section 13.6.1.7, "IB-ML CMD Set

Specific Fields," on page 564 for CMD field placement in the split-transaction message.

Table 143 CME Commands

Command	CMD	Request data	Response data
GetCMDSetVersion (13.6.12.1)	00h	none	byte 1: completion status[a] byte 2: Version of CME command set (CMD Set = IB-ML) that this controller supports. BCD encoded. Initial value = 10h for CMD Set version 1.0
GetVendorID (13.6.12.2)	01h	none	byte 1: completion status[a] bytes 2+ - A series of one or more 3-byte IDs for each vendor or organization that have defined controller-specific (OEM) commands implemented by this CME. FFFFFFh if no controller-specific commands. 3-byte ID uses same format as the vendor ID field in ChassisGUID.
ProxyWriteRead (13.6.12.3)	02h	byte 1: SlotSelector[b], 1-based byte 2: device selector byte 3: # bytes to write byte 4: write data (if any) byte 5: # bytes to read	byte 1: completion status[a]
ReadVPD (13.6.12.4)	03h	byte 1: SlotSelector[b], 1-based 00h = No s. Device is directly associated with the CME. byte 2: VPD device selector (address) byte 3 VPD number of bytes to read (1 byte. 1-based. 0 count is allowed) bytes 4:5 VPD offset to read (2 bytes)	byte 1: completion status[a] byte 2+ - read data bytes (if any)
WriteVPD (13.6.12.5)	04h	byte 1: SlotSelector[b], 1-based. 00h =No proxy. Device is directly associated with the CME. byte 2: VPD device selector (address) byte 3 number of bytes to write (1-bytes, 1-based. 0 count is allowed) bytes 4:5 VPD offset to write (2-bytes, 0-based) byte 6+: data (N bytes)	byte 1: completion status[a]
OEM (13.6.12.6)	05h	bytes 1-3: vendor ID byte 4+: vendor-specific data	byte 1: completion status[a] bytes 2-4: vendor ID byte 5+: vendor-specific response data

Table 143 CME Commands

Command	CMD	Request data	Response data
OutBandModuleCtl (13.6.12.7)	06h	byte 1: SlotSelector[b], 1-based. ResetIbML, Reset, or VBxEn_L for the specified slot. byte 2: IB-ML Reset and Module Reset and Power Control bit 7:3 - reserved bit 3 - DisableBulkPower 0 = Do not Disable bulk power via **VBxEn_L** 1 = Disable bulk power via **VBxEn_L** bit 2 - reserved bit 1 - ResetModule 0 = Negate **IMxPRst** 1 = Reset Module via **IMxPRst** bit 0 - ResetIBML 0 = Do not reset IB-ML 1 = Reset IB-ML	byte 1: completion status[a] 51h = could not clear IB-ML 52h = Module Reset not supported 53h = Module Power Control not supported
ReadModuleCtlStatus (13.6.12.8)	07h	byte 1: SlotSelector[b], 1-based. Data associated with OutBandModuleCtl for specified slot. byte 2: IB-ML Reset, Module Reset, Module Present and Bulk Power Control Status bit 7:4 - reserved bit 3 - BulkPowerDisabled 0 = Bulk Power is not Disabled via **VBxEn_L**. 1 = Bulk Power is Disabled via **VBxEn_L** bit 2:1 - Module **IMxPRst** 00 = CME is not asserting **IMxPRst**. 01 = CME is asserting **IMxPRst** 10 = CME samples **IMxPRst** as Module not present. 11 = CME samples **IMxPRst** as Module present. bit 0 - ResetIBML 0 = IB-ML is not in reset. 1 = IB-ML is in reset.	byte 1: completion status[a]

a. The completion status field uses the values defined in Table 135. Additional completion status values that may pertain to a given command are enumerated.
b. A non-zero SlotSelector refers to the designating number of an InfiniBand Slot in the Chassis. SlotSelector 0 refers to the CME function which relates to the Chassis itself.

13.6.12.1 GETCMDSETVERSION

GETCMDSETVERSION is an IB-ML split-transaction command that returns the version of the IB-ML command set (CMD Set = IB-ML) implemented

by the CME. The version information is stored in BCD (binary coded decimal) format, where the most-significant nibble holds the major version number and the least-significant nibble holds the minor version number. For example, a value of 09h corresponds to IB-ML command set version 0.9.

13.6.12.2 GETVENDORID

GetVendorID is an IB-ML split-transaction command that returns the ID for the vendor that has defined controller-specific (CMD Set = OEM) messages for the CME. The vendor ID format is the same as that used in the ChassisGUID. If no CMD Set = OEM messages are defined, a value of 00_00_00h **shall** be returned in the response. Note that it is possible for more than one set of vendor-specific commands to be implemented on a CME.

13.6.12.3 PROXYWRITEREAD

ProxyWriteRead, an IB-ML split-transaction command, directs the CME to perform a Write-Read transaction to the IB-ML for the specified slot. The CME will perform the requested transaction.

13.6.12.4 READVPD

ReadVPD, an IB-ML split-transaction command, allows the CME to return VPD data from the IB-ML associated with the slot specified by the *slot selector* parameter. A slot selector value of 00h indicates a VPD device that is directly associated with the CME rather than a given slot. The ReadVPD command can be used for getting VPD information from a slot (IB-ML) other than the one used to deliver the command to the CME. This allows the CME to act as a "proxy" to access VPD from Modules, such as Repeater Modules, that lack direct support for providing VPD information via the ***ReadVPD*** Baseboard Management MAD.

A CME implementation that receives a ReadVPD CME command from one IB-ML to read VPD from a slot's IB-ML will typically use the parameters from the request to format an atomic Write-Read IB-ML transaction to read the VPD SEEPROM on the specified slot's IB-ML. If successful, the CME will then take the data read from this transaction and use it to format a ReadVPD response with a 00h "OK" completion status. If the CME encountered errors while attempting to access the IB-ML, it **shall** return an appropriate non-zero completion status.

13.6.12.5 WRITEVPD

WriteVPD, an IB-ML split-transaction command, allows the CME to write to the VPD device for the slot specified by the slot selector parameter. A slot selector value of 00h indicates a VPD device that is directly associated with the CME rather than a given slot. It can be for writing into writable VPD devices of Modules, such as Repeater Modules, that lack direct

support for providing VPD write access via the ***WriteVPD*** Baseboard Management MAD. The operation of the WriteVPD CME is similar to that of the ReadVPD command, except that data from the request is written to the specified device, and the response does not contain data other than the completion status value.

13.6.12.6 OEM

This IB-ML split-transaction message allows a vendor or organization to define additional CME functions. The message includes a vendor-ID field that helps ensure that data parameter values specified by one vendor or organization do not interfere or need to be reconciled with the values selected by another vendor or organization. All other data fields are specified by the vendor or organization identified by the vendor-ID field.

The OEM CME command specified here is within the IB-ML Command Set (Table 143: CME Commands) and **shall** utilize the same message fields (Destination Device ID, Source Device ID, Rs, CMD Set, Length, etc.) as the other standard CME commands specified in this section.

13.6.12.7 OUTBANDMODULECTL

The **OUTBANDMODULECTL** split-transaction command directs the CME to control a specific IB-ML, Module bulk power (via **VBxEn_L**), and Module reset (via **IMxPRst**).

The slot selector value selects the IB-ML and the Module for this function.

If the command is supported and successful, the response **shall** return an “OK” completion status value; otherwise, the CME **shall** return an appropriate non-zero (error) completion status value.

13.6.12.8 READMODULECTLSTATUS

The **READMODULECTLSTATUS** split-transaction command directs the CME to provide information about IB-ML, Module bulk power (via **VBxEn_L**), and Module reset and presence status (via **IMxPRst**) for a specific slot.

The slot selector value selects the IB-ML and the Module for this function.

If the command is supported, the response **shall** return an “OK” completion status value; otherwise, the CME **shall** return an appropriate non-zero (error) completion status value.

13.6.13 IB-ML TIMEOUT PARAMETERS

C13-53: All Modules **shall** conform to the parameters listed in Table 144 IB-ML Timeout Parameters on page 597.

o13-16: All non-Modules that implement IB-ML **shall** conform to the parameters listed in Table 144 IB-ML Timeout Parameters on page 597.

C13-54: All Managed Chassis **shall** conform to the parameters listed in Table 144 IB-ML Timeout Parameters on page 597.

13.6.13.1 IB-ML SLAVE TIMEOUT

A slave device always times out when any clock is held low longer than $t_{TIMEOUT}$ maximum. See Figure 175 on page 598.

13.6.13.2 IB-ML MASTER TIMEOUT

$t_{LOW:SEXT}$ is defined as the cumulative time a master device is allowed to extend its clock cycles within one byte in a message as measured from:

start to ack

ack to ack

ack to stop

See Figure 175 on page 598 for a timing diagram depicting this parameter.

Table 144 IB-ML Timeout Parameters

Symbol	Parameter	Minimum	Maximum	Units	Comments
$t_{TIMEOUT}$	Clock low time-out	25	35	ms	See note [a]
$t_{LOW:SEXT}$	Cumulative clock low extend time (slave device)		25	ms	See note [b]
$t_{LOW:MEXT}$	Cumulative clock low extend time (master device)		10	ms	See note [c]
$t_{BMTrapRetryPeriod}$	BMTrap Retry Period	200	300	ms	See note [d]
$t_{BMTrapTimeout}$	BMTrap Timeout Period	4	6	s	See note [d]
$t_{Module:CmdTOut}$	BM Command Timeout		20	ms	See note [e]
$t_{CME:NoPrxCmdTOut}$	CME Command Timeout while not using a Proxy Command		20	ms	See note [f]
$t_{CME:PrxCmdTOut}$	CME Command Timeout when using a Proxy Command		50	ms	See note [f]
$t_{CME:OEMCmdTOut}$	CME Command Timeout for an OEM Command	-	-		Not specified [f]
$t_{IbML2IbDelay}$	IB-ML to IB message delay		20	ms	See note [g]

a. Devices participating in a transfer will timeout when any clock low exceeds the value of $t_{TIMEOUT}$MIN). Devices that have detected a timeout condition must reset the communication no later than $t_{TIMEOUT}$MAX). The maximum value specified must be adhered to by both a master and a slave as it incorporates the cumulative stretch limit for both a master ($t_{LOW: MEXT}$) and a slave ($t_{LOW: SEXT}$).
b. $t_{LOW: SEXT}$ is the cumulative time a slave device is allowed to extend the clock cycles in one message from the initial start to the stop. If a slave device exceeds this time, it is expected to release both its clock and data lines and reset itself.
c. $t_{LOW: MEXT}$ is the cumulative time a master device is allowed to extend its clock cycles within each byte of a message as defined from start-to-ack, ack-to-ack, or ack-to-stop. A master device **shall not** violate $t_{LOW: MEXT}$ except while forcing a slave device timeout.
d. While awaiting **SW_RTR** bit to be set to 1b, the CME **shall** retry the ***BMTrap*** message once every $t_{BMTrapRetryPeriod}$ for $t_{BMTrapTimeout}$ Period of time.
e. From the time the IB device receives the entire BM MAD, the amount of time it takes for a Module to respond to a BM command before the Baseboard Manager will have to retry the message
f. Amount of time it takes for a CME to respond to a CME Command from the time the CME receives a CME Command
g. The amount of time it takes for a Module to respond to an IbML2Ib register access (IBML -to- IB delay) (form the time the Module receives the IbML2Ib message into the MME's register to the time the Module sends the IB message on the IB) See Section 13.6.8 on page 587.

Figure 175 shows the cycle-to-cycle timeout parameters that are mentioned in Table 144.

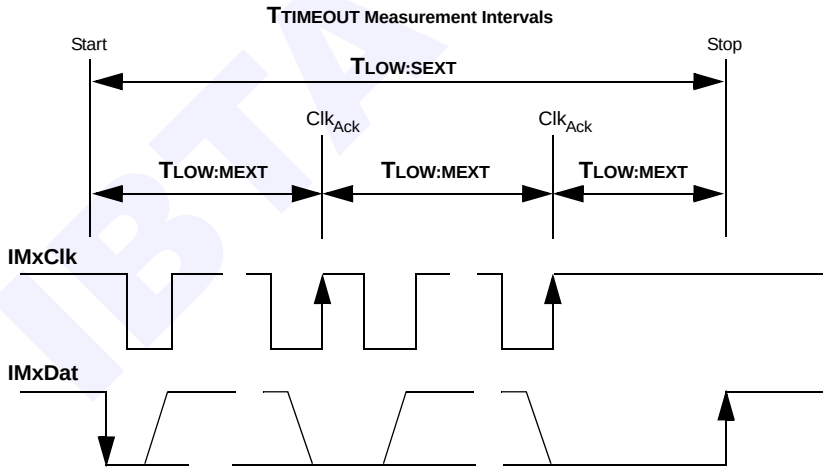


Figure 175 IMxDat, IMxClk $t_{TIMEOUT}$ Measurement Intervals

13.7 xINFO FORMAT

The following section describes the format of xInfo such as ChassisInfo, ModuleInfo, FRUInfo, PortConnectionInfo, ModulePowerInfo, CMEInfo, BuddyInfo, and OEM Record used in InfiniBand Chassis and Modules. InfiniBand devices on non InfiniBand-specified form factors may also adopt this format.

13.7.1 OVERVIEW

ModuleInfo and ChassisInfo include Vital Product Data (VPD). VPD, also referred to as Field Replaceable Unit information or FRU data, includes

information for identifying the type and version information of a replaceable unit or an entire system. This often includes elements such as serial numbers and part numbers.

In addition to VPD, ChassisInfo and ModuleInfo include elements that report about the configuration such as the number of slots a Chassis supports, whether a slot provides CME access, power management capabilities for a Module, GUIDs, etc.

The ChassisInfo and ModuleInfo data are stored as a series of typed records. Each record has a record header that indicates the type, version, and size of the data in that record. Following the header is a data area consisting of a number of fixed length and/or variable length fields.

A single device that emulates 24C02 SEEPROM style non-volatile memory may contain several xInfo Records; conversely, an xInfo may be large enough to require extension devices. See Section 13.7.3, “Device Header,” on page 603.

There are non-volatile writable regions in xInfo devices into which the CME, the Module, or the Baseboard Manager may write. The data in these specified writable areas follow the OEM Record format. See Table 159 OEM Record on page 628.) The Device Header of each device identifies the write-protected regions of that device.

IBTA

13.7.1.1 NOTATION

Table 145 identifies the notations used in the xInfo format tables and descriptions.

Table 145 Notations used in xInfo Field Descriptions

Abbreviation	Description
PA_Module	Protocol-aware InfiniBand Module
NPA_Module	non-Protocol-aware InfiniBand Module
N_ModuleXCA	Non-Module xCA. An xCA that is not on an InfiniBand Module
N_ModuleSw	Non-Module Switch. A Switch that is not on an InfiniBand Module
Req_UNA	Required Unspecified Not Allowed (if the Record is present, the field is required to be present and populated with valid data other than the “unspecified” value.)
Req_UA	Required Unspecified Allowed (if the Record is present, the field is required to be present and populated with valid data.)
Req	Required (if the Record is present, the field is required to be present and populated with valid data. The “unspecified” value is not defined for this field.)
Opt	Optional The presence of the field in the Record is optional. If present, the field **shall** be populated with valid data.

13.7.2 COMPLIANCE

C13-55: All xInfo devices that a Module implements **shall** be non-volatile memory devices and **shall** conform to the Header and Record definitions in this section. The Module **shall** provide IB-ML access to its xInfo devices when $\mathbf{V_{Aux}}$ is available on **VA_In**. (See Section 11.4, “Auxiliary Power Group,” on page 446.)

C13-56: All xInfo devices that a Chassis implements **shall** be non-volatile memory devices and **shall** conform to the Header and Record definitions in this section.

C13-57: All xInfo devices that a non-Module implements **shall** be non-volatile memory devices and **shall** conform to the Header and Record definitions in this section.

C13-58: This statement has been made obsolete by C13-58.1.1:.

C13-58.1.1: A Protocol-aware InfiniBand Module **shall** implement the Records and their fields that are labeled "**Req**", "**Req_UNA**", and "**Req_UA**" in the **PA_Module** column of the tables in Section 13.7.6, ModuleInfo Record, Section 13.7.8, FRUInfo Record, and Section 13.7.9, PortConnectionInfo Record, Section 13.7.14, BuddyInfo Record, Section 13.7.10, ModulePowerInfo Record, and Section 13.7.15, AssetTag Record.

C13-59: This statement has been made obsolete by C13-59.1.1:.

C13-59.1.1: .A non-Protocol-aware InfiniBand Module **shall** implement the Records and their fields that are labeled "**Req**", "**Req_UNA**", and "**Req_UA**" in the **NPA_Module** column of the tables in Section 13.7.6, ModuleInfo Record, Section 13.7.8, FRUInfo Record, Section 13.7.9, PortConnectionInfo Record, Section 13.7.14, BuddyInfo Record, Section 13.7.10, ModulePowerInfo Record, and Section 13.7.15, AssetTag Record.

o13-17: This statement has been made obsolete by o13-59.1.1:.

o13-59.1.1: A Module that optionally implements, ChassisInfo Record, CMEInfo Record, or OEM Record **shall** follow the format of the appropriate sections Section 13.7.7, ChassisInfo Record, Section 13.7.12, CMEInfo Record, and Section 13.7.13, OEM Record respectively.

o13-18: A non-Module xCA_NID or non-Module Switch_NID that optionally implements ModuleInfo Record, ChassisInfo Record, PortConnectionInfo Record, ModulePowerInfo Record, IOCPMInfo Record, CMEInfo Record, OEM Record, FRUInfo Record, BuddyInfo Record or AssetTag Record **shall** follow the format of the appropriate sections Section 13.7.6, ModuleInfo Record, Section 13.7.7, ChassisInfo Record, Section 13.7.9, PortConnectionInfo Record, Section 13.7.10, ModulePowerInfo Record, Section 13.7.11, IOCPMInfo Record, Section 13.7.12, CMEInfo Record, Section 13.7.13, OEM Record, Section 13.7.8, FRUInfo Record, Section 13.7.14, BuddyInfo Record, and Section 13.7.15, AssetTag Record respectively.

o13-19: A non-Module xCA_ID or non-Module Switch_ID that optionally implements PortConnectionInfo Record, ModulePowerInfo Record, IOCPMInfo Record, CMEInfo Record, OEM Record, FRUInfo Record, BuddyInfo Record or AssetTag Record **shall** follow the format of the appropriate sections Section 13.7.9, PortConnectionInfo Record, Section 13.7.10, ModulePowerInfo Record, Section 13.7.11, IOCPMInfo Record, Section 13.7.12, CMEInfo Record, Section 13.7.13, OEM Record, Section 13.7.8, FRUInfo Record, Section 13.7.14, BuddyInfo Record, and Section 13.7.15, AssetTag Record respectively.

C13-60: A non-Module Switch_ID **shall** implement the Records and their fields that are labeled "**Req**", "**Req_UNA**", and "**Req_UA**" in the **N_ModuleSw** column of the tables in Section 13.7.6, ModuleInfo Record and Section 13.7.7, ChassisInfo Record.

C13-61: A non-Module xCA_ID **shall** implement the Records and their fields that are labeled "**Req**", "**Req_UNA**", and "**Req_UA**" in the **N_ModuleXCA** column of the tables in Section 13.7.6, ModuleInfo Record, and Section 13.7.7, ChassisInfo Record.

C13-62: An Actively Managed Chassis **shall** implement the Section 13.7.7, ChassisInfo Record, Section 13.7.12, CMEInfo Record, and Section 13.7.15, AssetTag Record.

C13-63: A Passively Managed Chassis **shall** implement the Section 13.7.7, ChassisInfo Record.

o13-20: This statement has been made obsolete by o13-63.1.1:.

o13-63.1.1: An Actively Managed Chassis that optionally implements FRUInfo Record, PortConnectionInfo Record, BuddyInfo Record or OEM Record **shall** follow the format of the appropriate sections Section 13.7.8, FRUInfo Record, Section 13.7.9, PortConnectionInfo Record, Section 13.7.14, BuddyInfo Record and Section 13.7.13, OEM Record.

o13-21: This statement has been made obsolete by o13-63.1.1:.

o13-63.1.1: A Passively Managed Chassis that optionally implements FRUInfo Record, PortConnectionInfo Record, BuddyInfo Record, AssetTag Record, or OEM Record **shall** follow the format of the appropriate sections Section 13.7.8, FRUInfo Record, Section 13.7.9, PortConnectionInfo Record, Section 13.7.14, BuddyInfo Record and Section 13.7.13, OEM Record.

o13-22: A Protocol-aware Module, a Non-Module xCA, or a Non-Module Switch that implements Power Management as indicated by *ModulePowerInfo.ModulePMCapability.IsPowerManaged* = 1b, **shall** implement the Records and their fields that are labeled "**Req**", "**Req_UNA**", and "**Req_UA**" in the respective **PA_Module**, **N_ModuleXCA, and N_ModuleSw** columns of the tables in Section 13.7.11, IOCPMInfo Record.

C13-64: An entity that implements ModuleInfo Record, ChassisInfo Record, PortConnectionInfo Record, IOCPMInfo Record, CMEInfo Record, OEM Record, FRUInfo Record, BuddyInfo Record, or AssetTag Record **shall** provide the appropriate Record Header as defined in Section 13.7.4, "Record Header," on page 604. The xInfo device in which the Record resides, shall have a Device Header as defined in Section 13.7.3, "Device Header," on page 603. If an entity implements a

ChassisInfo Record, the ChassisInfo Record shall immediately follow the Device Header.

C13-65: An entity that implements ModuleInfo Record, ChassisInfo Record, PortConnectionInfo Record, ModulePowerInfo Record, IOCPMInfo Record, CMEInfo Record, OEM Record, FRUInfo Record, BuddyInfo Record or AssetTag Record **shall** follow the format of the appropriate sections Section 13.7.6, ModuleInfo Record, Section 13.7.7, ChassisInfo Record, Section 13.7.9, PortConnectionInfo Record, Section 13.7.10, ModulePowerInfo Record, Section 13.7.11, IOCPMInfo Record, Section 13.7.12, CMEInfo Record, Section 13.7.13, OEM Record, Section 13.7.8, FRUInfo Record, Section 13.7.14, BuddyInfo Record, and Section 13.7.15, AssetTag Record respectively.

13.7.3 DEVICE HEADER

The Device Header identifies the writable regions within a device. It also provides a version number indicating the format of the Device Header that matches this specification as well as the device address for the next xInfo device in a chain of xInfo devices. The Device Header is present at the beginning of all xInfo devices (VPD devices). xInfo Records immediately follow the Device Header.

xInfo Device Header (Section Table 146 on page 603) outlines the format of the xInfo Device Header.

Table 146 xInfo Device Header

Byte Offset	Field
0	VPD_FormatVersion in BCD. [7:4] - MSN = major revision [3:0] - LSN = minor revision 11h for this specification (format version 1.1)
1	[7:1] - ExtensionDeviceSlaveAddress 0000_000 = no extension device. [0] - reserved (set to 0b)
2:5	byte 2: LSB of offset to first byte of write-protected area byte 3: MSB of offset to first byte of write-protected area byte 4: LSB of offset to last byte of write-protected area byte 5: MSB of offset to last byte of write-protected area FFFF_0000h if no write-protected area.
6	Header Checksum

13.7.3.1 VPD_FORMATVERSION

The version information is stored in BCD (binary coded decimal) format, where the most-significant nibble holds the major version number and the least-significant nibble holds the minor version number.

13.7.3.2 EXTENSIONDEVICESLAVEADDRESS

ExtensionDeviceSlaveAddress is the address of the device on IB-ML that is the next extension to the device containing this device header. From the IB view, this address is the address that is used in a ReadVPD Baseboard Management command to the MME. From the IB-ML view, this address will either be a physical IB-ML device address accessed using an IB-ML SEEPROM write-read transaction or a logical IB-ML device accessed using a WriteVPD or ReadVPD command. Software should first try to use the same access method for the Extension device that was used to access the present device, and if that fails try the alternate implementation.

13.7.3.3 OFFSETFORMAT

OFFSETFORMAT defines the offset format to use in the subsequent bytes which are offsets to the first and the last bytes of a write-protected area.

If both start and end offset are the same value, it indicates that 1 byte is write protected. The value of FFFFh as the starting offset and the 0000h as the ending offset indicates that NO write-protected bytes are present.

13.7.4 RECORD HEADER

The Record Header contains offset information needed to find any additional records that will sequentially follow in a list manner. The Record Header has the following format.

Table 147 Record Header

Field Size (in Bytes)	Field
1	Record Type ID See Table 151 Record Type IDs on page 609

DRAFT

Table 147 Record Header

Field Size (in Bytes)	Field
1	Record Parameters [7] - LastRecord 0b = more records follow 1b = last record in this read-only or writable area. To find all Records, Software shall check both read-only and writable areas. (The device header identifies the read-only and writable areas.) Other Records may exist in the extension device(s). [6] - Length Multiplier 0b = record length given in bytes 1b = record length given in multiples of 8 bytes. This does not imply that the record data starts on an 8-byte boundary. The record data always immediately follows the record header. [5] - Read Only 0b = The record contents are not Write Protected 1b = S/W **shall not** attempt to write to any field in the record; it is recommended that the h/w device physically write-protect the record. [4] - reserved [3:0] - Record Format version (=1h for this specification, unless otherwise specified.)
1	Record Length, bits 7:0 (zero based)
1	Header Checksum

Any data associated with the Record will follow the last field of the header.

13.7.4.1 CHECKSUM

13.7.4.1.1 CHECKSUM CALCULATION

Record and field checksums are calculated as 256-([sum(bytes)+bytecount] mod 256). This is equivalent to taking an 8-bit sum of the bytes plus the bytecount and then performing an 8-bit 2's complement of the result.

Example:

Assume that we have three data bytes, AAh, 87h, and 32h. The bytecount is then 3. The 8-bit sum of the bytes plus the bytecount is AAh+87h+32h+03h = 66h. 66h is 01100110b. The two's complement of 01100110b is equal to 10011001b+1b = 10011010b or 9Ah.

13.7.4.1.2 VERIFYING THE CHECKSUM

If the checksum is added to the preceding data bytes plus the bytecount, the sum modulo 256 should be 00h.

Example:

Assume that we have three data bytes, AAh, 87h, and 32h. Per the preceding examples, the checksum is 9Ah. If we add AAh, 87h, 32h, 03h, and 9Ah we get 200h. 200h modulo 256 = 00h.

13.7.4.2 ENCODING/LENGTH BYTE

Records can be a combination of fixed-length and variable length fields. The Encoding/Length byte serves several purposes for variable length fields: It holds the length value for the field. It allows 'string' fields to be encoded in a way that is more efficient based on whether the string can be encoded as BCD (two characters per byte), ASCII (one character per byte), or UNICODE v3.0 (two bytes per character).

Architectural Note

The character encoding is done in a way that helps utility software present data, even if it doesn't know the semantics of the field. For example, it can display BCD, ASCII, or UNICODE strings without knowing their exact definition and can display binary encoded fields as 'hex'. Lastly, it provides a mechanism to tag OEM-defined fields.

The checksum is always the last byte in a Record. The Record Header specifies the length of the Record; therefore, the checksum is always at a known location regardless of the number of PADded bytes.

Table 148 presents the specification of the encoding/length byte.

Table 148 ENCODING/LENGTH BYTE

bits	Sub-field	Description
[7:6]	EncodingType	Identifies the type of encoding used for the subsequent data bytes. 00b - Binary or unspecified. Binary data bytes are stored least-significant byte first unless otherwise specified. Binary data fields are limited to 62 bytes, maximum. 01b - BCD plus (see below). BCD values are stored most-significant byte first. E.g. the BCD bytes for 1234h are stored as: byte 0 = 12h, byte 1=34h. This is done because it is expected that most BCD encoded fields will be treated as strings rather than converted to binary values. 10b - UNICODE v3.0 string, or OEM Field prefix. 11b - 8-bit ASCII + Latin 1. "ASCII+LATIN1" is derived from the first 256 characters of Unicode 3.0. The first 256 codes of Unicode follow ISO 646 (ASCII) and ISO 8859/1 (Latin 1). The Unicode "C0 Controls and Basic Latin" set defines the first 128 8-bit characters (00h-7Fh) and the "C1 Controls and Latin-1 Supplement" defines the second 128 (80h-FFh).

Table 148 ENCODING/LENGTH BYTE

bits	Sub-field	Description
[5:0]	CharacterCount	Number of characters (number of bytes for a binary-encoded field). A character count of 00_0000b has special meaning in combination with the Encoding Type field. In addition, a count of 11_1111b is reserved when used with Encoding Type = 00b (binary). See Table 149 Special Encoding/Length Field Values

13.7.4.2.1 SPECIAL ENCODING/LENGTH FIELD VALUES

The values in Table 149 for an Encoding/Length field have special meanings.

Table 149 Special Encoding/Length Field Values

value	name	meaning
00h	reserved	Reserved. This value **shall** not be used where an Encoding/Length byte is expected.
3Fh	reserved	Reserved. This value **shall** not be used where an Encoding/Length byte is expected.
40h	Null Field	Empty (null) field. This value is used to identify a field that is present but unpopulated. This value must be used for null fields even when the normal encoding for this field when populated is not ASCII.
80h	OEM Prefix	80h = OEM Prefix field. The field indicates the start of a set of vendor-specific fields within a standard InfiniBand record. All following fields in the record, up to the next OEM Prefix field if any, are assumed to be specified by the vendor or organization identified by the OEM ID. The value FFh, FFh, FFh for a EUI-64 Company ID is reserved for InfiniBand specification defined records. OEM Fields are only allowed to follow the InfiniBand specification defined fields in a record.
C0h	Last Field	Last variable length field in record. This value is used when a record has a variable number of Encoding/Length encoded fields and padding follows the last field. A record is allowed to end immediately with the checksum byte. I.e. a C0h does not need to follow the last field unless there are additional characters (PAD) between the last field and the checksum.

13.7.4.2.2 OEM PREFIX FIELD

The OEM Prefix field is specified as follows. Note that the length of the field is implied by having an Encoding/Length value of 80h and is fixed at five bytes.

Table 150 OEM Prefix Field

field size (in bytes)	field
1	**OEM Prefix** Encoding/Length = 80h (OEM ID Field)
1	[7:4] **ID Type** 1h = Company ID per EUI-64 / 48 (3 bytes) FFh, FFh, FFh = InfiniBand specified fields 2h = IANA Enterprise Number-based ID (3 bytes). FFh, FFh, FFh = reserved all others = reserved **[3:0] ID Length** - # of OEM ID data Bytes following
3	OEM ID data bytes least-significant byte first

13.7.4.3 USING OEM RECORDS AND DATA FIELDS

The VPD format supports both OEM-unique records, and OEM-unique data field extensions within InfiniBand specification defined ‘standard’ records. The following outlines the usage rules for OEM records and OEM fields within standard records.

1) OEM Records
   - With the exception of the ending checksum byte, the format of data fields within OEM-unique Records is not specified.
2) OEM Fields in Standard Records
   - One or more groups of OEM-unique data fields can be included in a standard record.
   - Each *group* of OEM-unique data fields for a given vendor or organization is required to be prefixed by a special ‘OEM ID Field’ that contains a 3-byte ID for the OEM.
   - Each individual field within the group is required to be formatted as a variable length field that starts with an Encoding /Length byte.
   - OEM-unique fields in a standard record can only follow the last standard field defined for the record. They are not allowed to be inserted between standard fields, even if the standard field starts with an Encoding/Length byte.

13.7.4.4 USE OF OPTIONAL FIELDS

Unless otherwise specified, Optional Fields must be present, but can be null or filled with the appropriate value indicating the field contents are ‘unspecified’. Software identifies fields based on their order in the record.

Since individual fields are not tagged according to type (to save space) all fields must be present to maintain ordering.

13.7.4.5 USE WITH NON-MODULE VPD DEVICES

All Record Types are allowed to be used on non-Module VPD devices, although some records only have meaning in the context of an InfiniBand Module. A non-Module device can optionally return records, such as ChassisInfo Records, at the ModuleInfo VPD device address. This implies that software should always check the ModuleInfo VPD device address first as the root device for VPD discovery. The ModuleInfo Record will indicate whether there is additional CME or ChassisInfo that can be obtained from other device addresses.

13.7.5 RECORD TYPES

Table 151 Record Type IDs on page 609 lists the different VPD record types and the associated Record Type ID values that are used in the Record Header.

Table 151 Record Type IDs

Record Type	Record Type ID
ModuleInfo	00h
ChassisInfo	01h
FRUInfo	02h
ModulePowerInfo	03h
PortConnectionInfo	04h
CMEInfo	05h
OEM	06h
BuddyInfo	07h
AssetTag	08h
IOCPMINFO	09h

13.7.6 MODULEINFO RECORD

The ModuleInfo Record is typically included on an InfiniBand Module to describe the Module class and the number of InfiniBand and IB-ML links

exiting the Module. Non-Modules can also utilize this record to describe their link and IB-ML connection support.

Table 152 ModuleInfo Record - Record Format Version=2

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
8	**ModuleGUID** - EUI-64 ID assigned for this Module	RO	**Req**	**Req**	**Req**	**Req**
1	[7:5] **IBModuleType** 000b = Module is not an InfiniBand Module 001b = Module is an InfiniBand Module All others = reserved [4:0] **ModuleClass** 0x00 - TCA 0x01 - HCA 0x02 - Switch 0x03 - Router 0x04 - 1x Repeater (port connection info record indicates whether repeater is Copper or Optical) 0x05 - 4x Repeater 0x06 - 12x Repeater All others - Reserved	RO	**Req**	**Req**	**Req**	**Req**
1	**NodeCount** - Number of InfiniBand Nodes on this Module with unique NodeGUIDs. A value of FFh represents more than FEh.	RO	**Req**	**Req**[a]	**Req**[a]	**Req**[a]
1	**PortCount** - Number of InfiniBand port on this Module with unique Port-GUIDs (i.e. NodeGUID plus Port Number) A value of FFh represents more than FEh.	RO	**Req**	**Req**	**Req**[a]	**Req**[a]
1	**BackplaneLinkCount** - Number of InfiniBand Links exiting this Module through its Backplane Connectors. Defines the number of **Backplane Connection** fields present in the associated **PortConnectionInfo** record. For modules that are not InfiniBand Modules, this count is the number of InfiniBand links that exit the module.	RO	**Req**	**Req**	**Req**[a]	**Req**[a]
1	**IBMLCount -** Number of IB-ML interfaces on this Module	RO	**Req**	**Req**	**Req**[a]	**Req**[a]
1	**BackplaneIBMLCount -**Number of IB-MLs exiting this Module through its Backplane Connectors. For modules that are not InfiniBand Modules, this count is the number of IB-MLs that exit the module.	RO	**Req**	**Req**	**Req**[a]	**Req**[a]

Table 152 ModuleInfo Record - Record Format Version=2

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
1 or 6 or 8	**ModuleSize -**Volumetric information for standard form factors in millimeters. For InfiniBand Modules, the origin is defined as the 0,0,0 datum used in the mechanical specifications for the Module. For non-standard modules, this is the dimension of the overall envelope of the module. Format (stored least significant byte first) Field size is 1-byte if 'null', 8-bytes if DimensionType = 1b, 6-bytes if DimensionType=0b. For DimensionType = 1b (standard from origin) 7-bytes: Six 8-bit bitfields to describe 6 offsets from the origin: [63:57] - Reserved. [56] - Extension (reserved. Write as 0b) [55:48] - Right. Positive X direction from origin in right-hand coordinate system. [47:40] - Left. Negative X direction from origin in right-hand coordinate system. [39:32] - Up. Positive Y direction from origin in right-hand [31:24] - Down. Negative Y direction from origin in right-hand coordinate system. [23:16] - Front. Distance from origin to front (faceplate) of Module. Positive Z direction from origin in right-hand coordinate system. [15:08] - Rear. Distance from origin to rear or backplane connection of Module. Negative Z direction from origin. For DimensionType = 0b (envelope) [47:44] - Reserved. [43:32] - Module Width in mm 12-bits. [31:20] - Module Height in mm 12-bits. [19:08] - Module Depth in mm 12-bits. For DimensionTypes 0b and 1b [07:02] - Reserved [1] - DimensionType 1b = offsets specified relative to standardized Module origin. 0b = size of envelope specified. [0] - FieldPopulated 1b = bits [63:0] are populated. 0b = field is unspecified: only bits [7:0] are present. (bits 7:1 should be 000_0000b)	RO	**Req_ UA**	**Req_ UA**	**Req_ UA**	**Req_ UA**

Table 152 ModuleInfo Record - Record Format Version=2

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
1	**FormFactor** 00h = unspecified 01h = non-removable (integrated, embedded, or non- user -service-able) 02h = Standard IB Module 03h = Standard Wide IB Module 04h = Tall IB Module 05h = Tall Wide IB Module 13h = PCI 14h = low-profile PCI 15h = Compact PCI 16h = VME 17h = Internal Mezzanine Board 18h = Card Edge Board 19h = Device Bay 1Ah = Other Removable All others reserved	RO	**Req_ UNA**	**Req_ UNA**	**Req_ UA**	**Req_ UA**
var	**OEM fields**. A group of one or more OEM fields must be preceded with an OEM Prefix field.	-	Opt	Opt	Opt	Opt
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	-	Opt	Opt	Opt	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	-	Opt	Opt	Opt	Opt
1	**Checksum**	-	**Req**	**Req**	**Req**	**Req**

a. Zero is a valid number for this count.

13.7.7 CHASSISINFO RECORD

This record returns information about an InfiniBand Chassis and slot-specific information for Chassis VPD accessed from an InfiniBand Mod-

ule or a CME.

Table 153 ChassisInfo Record - Record Format Version=2

field size (bytes)	field	Read/Write	Requirement
8	**ChassisGUID** - EUI-64 ID that identifies this Chassis.	RO	**Req**
1	**SlotCount** - Total number of IB Module slots available in this Chassis. (1-based)	RO	**Req**
var	**SlotNumbers** - HighestSlotNumber[a] / 3 (rounded up) 8-bit bitfields indicating which slot numbers are valid for the Chassis. This field allows slot numbers to be non-sequential. Slot numbers must be in the range 1-254. 0 is assigned to the CME and the Chassis containing the CME. [7] - extension. 1b = There is at least one other SlotNumbers field byte following this one. 0b = This is the last byte of the SlotNumbers field in this record. [6]-reserved [5:0] - Each bit-pair represents one of three corresponding slot numbers, with bits [1:0] of the first SlotNumbers field corresponding to slot number 1. 00b = Chassis does not implement this slot number 01b = Chassis provides one Backplane Connector to this Standard Slot 10b = Chassis provides one Backplane Connector to this Tall Slot 11b = Chassis provides two Backplane Connectors to this Tall Slot	RO	**Req**
var	**CMEAccess** - HighestSlotNumber[a] / 7 (rounded up) contiguous, 8-bit, bitfields indicating which InfiniBand Module slots in this Chassis provide CME Access via their Primary Port (1). If this record is for the CME itself (Slot number = 0) then this indicates the number of IB-MLs connected to the CME. [7] - extension. 1b = There is at least one other CMEAccess field byte following this one. 0b = This is the last byte of the CMEAccess field in this record. [6:0] - Each bit represents one of seven corresponding slot numbers, with bit 1 of the first CMEAccess field corresponding to the first implemented slot as defined in Section 9.1.3.1, "Slot Designations," on page 360. (bit 0 of the first CMEAccess field indicates whether a CME is present at all. 0b = no CME, 1b = CME exists.) 0b = slot does not provide CME access. 1b = slot provides CME access.	RO	**Req**
SlotInfo			
1	**SlotNumber**- Chassis-assigned slot number for the slot through which this ChassisInfo device is being accessed by the Module. 1-based. Slot numbers must be in the range 1-254. 0 is assigned to the CME and the Chassis containing the CME. FFh means 'unspecified' and is used to support non-Module applications of this record.	RO	**Req_UNA**

Table 153 ChassisInfo Record - Record Format Version=2

field size (bytes)	field	Read/Write	Requirement
1	**SlotDetails -** One byte field giving details on the slot through which this ChassisInfo device is being accessed [7:6] CMEAccess 00b = unspecified (used to support non-Module applications of this record) 01b = A CME is accessible via the Primary Port (1) for this slot. If this info is for a non-Module, this bit indicates that a CME is accessible by BM MADs to the port used to access this VPD Device. 10b = A CME is not accessible from this slot / port. 11b = reserved [5] ProxyAccess 0b = There is no CME proxy access to other slots via the slot or port used to access this VPD Device. 1b = This slot or port provides access to a CME that provides proxy access to other slots. [4] reserved. [3:2] LockDrivesCTR - Indicates whether the lock (if present) keeps CME_CTR deasserted (0) until the lock is released. 00b = unspecified (used to support non-Module applications of this record) 01b = CME uses only state of lock to drive CME_CTR. 10b = CME_CTR will not be asserted unless lock is released, but CME has other criteria that can also cause CME_CTR to be kept deasserted. 11b = reserved [1:0] MechanicalLockPresent - Indicates whether there is a mechanical lock associated with this slot. A mechanical lock physically prevents removal of the Module from the slot. The implementation of the lock can be a purely mechanical (e.g. a mechanical key lock) or electro-mechanical (e.g. a solenoid). 00b = unspecified (used to support non-Module applications of this record) 01b = mechanical lock present 10b = no mechanical lock on this slot 11b = reserved	RO	**Req_ UNA**
1	**NodeCount** - Number of InfiniBand Nodes on this Chassis with unique Node-GUIDs. A value of FFh represents more than FEh.	RO	**Req_ UNA**
var	**OEM fields**. A group of one or more OEM fields must be preceded with an OEM Prefix field.	-	Opt
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	-	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	-	Opt
1	**Checksum**	-	**Req**

a. Highest Slot Number is the highest one that the chassis intends to potentially assign.

13.7.8 FRUINFO RECORD

FRU stands for “Field Replaceable Unit”. The FRUInfo Record provides information that can potentially help a user identify and inventory the field replaceable units that make up a given Module, system, or sub-system. The record can be used for many types of InfiniBand specification defined and non-InfiniBand specification defined FRUs.

C13-66: All InfiniBand Modules **shall** implement a FRUInfo Record in an xInfo device. They **shall** also follow the FRUInfo Record, the Device Header, and the Record Header specifications.

IBTA

o13-23: All entities that implement a FRUInfo Record **shall** follow the FRUInfo Record, the Device Header, and the Record Header specifications.

Implementation Note

There can be more than one FRUInfo Record within a VPD device. A typical example of this is a low-cost system design where a single VPD device is used to hold both FRUInfo for the chassis as a replaceable unit, and FRUInfo for the overall product. Similarly, a system where there is a single VPD device on the main system board could hold three FRUInfo Records. One for the system as an overall product, one for the main system board itself as a replaceable unit, and one for the Chassis as a replaceable unit.

Table 154 FRUInfo Record

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
1	**FRUType** 00h = unspecified 01h = InfiniBand Module 02h = InfiniBand Module Backplane 03h = InfiniBand Switch module[a] 04h = Main Chassis. This is FRUInfo for the Chassis itself as a replaceable unit. 05h = Platform/System Product. This identifies a separate FRU record used to hold information about the overall platform (system) as a product. It is usually kept in the same VPD device as a Main Chassis FRU record. 06h = Product. This identifies a FRU that is a standalone product, but does not fit the definition of an InfiniBand Platform or other predefined FRU types. 07h = Board or Card. This FRU holds info about a board or card that is a sub-unit of a system, a Chassis, or a Module. 08h = Power Converter / Supply module[a] 09h = Other module[a] (assembly) 0Ah = Cooling module[a] (assembly) (e.g. FAN) 0Bh = Sub-Chassis 0Ch = Processor module[a] (or field replaceable chip) 0Dh = Memory module[a] 0Eh = Memory Card 0Fh-DFh = reserved for future InfiniBand specifications E0h-FFh = OEM specified per manufacturer / company identified by ManufacturerID field, below.	RO	**Req_ UNA**	**Req_ UNA**	**Req_ UA**	**Req_ UA**

Table 154 FRUInfo Record

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
1	**FRU_Handle** The FRU_Handle identifies Records that are associated with the same FRU. The scope of the value for the FRU_Handle is the xInfo device and its extension devices. Different FRUs shall be represented by different FRU_Handles.	RO	**Req**	**Req**	**Req**	**Req**
1	**FRUGUID** All FRUGUID data bytes are encoded as binary, least-significant byte first, unless otherwise specified. 00h = none specified (0 data bytes follow) 01h = EUI-64 (8 data bytes follow) 02h = EUI-48 (6 data bytes follow) 03h = SMBIOS/IPMI/OSF/ Wired for Management/Microsoft GUID/UUID (16 data bytes follow) 04h = SMBus 2.0 UDID (8 bytes) All others reserved	RO	**Req**[b]	**Req**[b]	Opt	Opt
var	FRUGUID data bytes.	RO	**Req**[b]	**Req**[b]	Opt	Opt
1	**SerialNumber** Encoding/Length[c] (Table 148: Encoding/Length Byte)	RO	**Req**	**Req**	**Req**	**Req**
var	SerialNumber data bytes The FRU manufacturer assigns the FRU serial number.	RO	**Req**	**Req**	Opt	Opt
1	**PartNumber** Encoding/Length (Table 148: Encoding/Length Byte)	RO	**Req**	**Req**	Opt	Opt
var	PartNumber data bytes The FRU manufacturer assigns the FRU part number.	RO	**Req**	**Req**	Opt	Opt
1	**Model** Encoding/Length (Table 148: Encoding/Length Byte)	RO	**Req**	**Req**	Opt	Opt
var	Model data bytes The FRU manufacturer assigns the FRU model name.	RO	**Req**	**Req**	Opt	Opt
1	**Version** Encoding/Length (Table 148: Encoding/Length Byte)	RO	**Req**	**Req**	Opt	Opt
var	Version data bytes The FRU manufacturer assigns the FRU version number.	RO	**Req**	**Req**	Opt	Opt
1	**ManufacturerName** Encoding/Length (Table 148: Encoding/Length Byte)	RO	**Req**	**Req**	Opt	Opt
var	ManufacturerName	RO	**Req**	**Req**	Opt	Opt
1	**ProductName** Encoding/Length (use null if FRU is not sold or intended to be viewed as a separate product.) (Table 148: Encoding/Length Byte)	RO	**Req**	**Req**	Opt	Opt
var	ProductName data bytes	RO	**Req**	**Req**	Opt	Opt

Table 154 FRUInfo Record

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
1	**ManufacturerID** All following ManufacturerID data bytes are encoded as binary, least significant byte first, unless otherwise specified. 00h = none specified (0 data bytes follow) 01h = Company ID per EUI-64 / 48 (3 bytes) 02h = IANA Enterprise Number-based ID (3 bytes)	RO	**Req_ UA**	**Req_ UA**	**Req_ UA**	**Req_ UA**
var	ManufacturerID data bytes	RO	Opt	Opt	Opt	Opt
1 or 4	**Mfg. Date / Time** Number of minutes from 0:00 hrs, GMT 1/1/2000. (0:00 hrs is midnight 12/31/1999). Field size = four bytes if populated, one byte if null. (The definitions of month and year in this field are based on ISO 8601 specification.) A value of 0000_0000h indicates un-specified. Format (stored least significant byte first): **[31:22]** - **Year** Given as number of years from GMT 1/1/2000. 1-based. 0 corresponds to the year 2000. (1023 years) **[21:18] - Month** 1-based. 1 = January. **[17:13] - Day** of the month. 1-based. **[12:08] - Hours** Number of hours from 0:00 hrs on day of manufacture 1-based. ('24 hr' format). 0 means time is between 0:00 hours and 1:00 hours. **[07]** - DateTimeDataPresent 0b = bits 8:31 not present. This field is null. Bits 7:1 are reserved and should be 0000000b. 1b = bits 8:31 present. Bits 31:1 hold time value. **[06]** -reserved. **[05:00] - Minutes** from 0:00 hrs on day of manufacture. 1-based. 0 means time is exactly on the hour.	RO	**Req_ UA**	**Req_ UA**	**Req_ UA**	**Req_ UA**
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	RO	Opt	Opt	Opt	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	RO	Opt	Opt	Opt	Opt
1	**Checksum**	RO	**Req**	**Req**	**Req**	**Req**

a. The lower case "m" in module means that the module (assembly) is not an IB-Module.
b. ModuleGUID is required in ModuleInfo Record. FRUGUID data bytes are not required unless the FRUGUID does not follow EUI-64. If not populated, set FRUGUID to 00 (none specified).
c. These areas are always encoded as ASCII to facilitate serial number comparison and tracking by management software applications.

13.7.9 PORTCONNECTIONINFO RECORD

The PortConnectionInfo variable-length record provides information about the entity to which the port connects A BuddyInfo record **must** be present in the device, including any extensions, to provide the NodeGUIDHandles necessary for this record.

Table 155 PortConnectionInfo Record - Record Format Version=2

field size (bytes)		field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
1		**ConnectionInfoCount** - Number of NodeGUIDHandle/PortNo/ConnectionType/BackplaneConnection field sets following. 1-based. It is recommended, but not required, that one NodeGUID/PORTNO/ConnectionType field sets be provided for each link present on the Module. A value of FFh represents more than FEh.	RO	**Req**	**Req**	**Req**	**Req**
ConnectionInfoCount x	1	**NodeGUIDHandle** - Handle of NodeGUID to which the related fields apply.	RO	**Req**	**Req**	**Req**	**Req**
	2	**PortNo** - Port Number The value that this port will return in **PortInfo:LocalPortNumber** (see *InfiniBand Architecture Specification, Volume 1*, Chapter "Subnet Management", Section "Subnet Management Class")	RO	**Req**	**Req**	**Req**	**Req**

IBTA

Table 155 PortConnectionInfo Record - Record Format Version=2

field size (bytes)		field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
ConnectionInfoCount x (structure continuation)	1	**ConnectionType** [7:6] - **InternalConnection**[a] 00b = unspecified. 01b = Connection terminates at another Node without going through a cable or a connector. 10b = Connection terminates at another Node after going through a cable or a connector. 11b = no internal connection [5:3] - **Media Class** 000b = unspecified 001b = InfiniBand specified Copper Cable 010b = InfiniBand specified Fiber-optic Cable 011b = PCB 100b = a Repeater on the FRU, then Copper Cable[b] 101b = a Repeater on the FRU, then Fiber-optic Cable 110b = a Repeater on the FRU, then a Node 111b = reserved [2:0] - **ConnectionClass** 000b = unspecified 001b = Non-specified Removable [Connection is via a non-InfiniBand specified connector type] 010b = Chassis side of a Backplane Slot Connector 011b = Module side of a Backplane Slot Connector 100b = Non-removable [Port is 'hard wired' to the Module] 111b = The information is not found in this field; other fields are fully descriptive. others = reserved Examples: `01_011_111` Non-removable (Two Nodes connected via PCB) `10_000_111` Non-specified Removable (proprietary connector for internal use) `11_000_111` Non-specified Removable (proprietary connector for use outside of the enclosure housing the FRU) `10_011_010` Backplane Slot Connector (Chassis HCA connected to a Slot) `10_011_011` Module Slot Connector (Module TCA connected to its Backplane Connector `10_001_111` Copper Cable (e.g. Switch connected to a Copper Cable) `10_010_111` Fiber Cable (e.g. HCA connected to a Fiber Cable) `10_101_111` Fiber Cable (e.g. HCA, connected to a Fiber Cable via a Repeater)	RO	**Req_ UNA**	**Req_ UNA**	**Req_ UA**	**Req_ UA**

DRAFT

Table 155 PortConnectionInfo Record - Record Format Version=2

field size (bytes)		field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
ConnectionInfoCount x (structure continuation)	1	**BackplaneConnection** - Indicates mapping to backplane designations based on value of **ConnectionType:ConnectionClass** If **ConnectionType:ConnectionClass** = 010b (Chassis side), the field encodes the designation defined in Section 9.1.3.1, "Slot Designations," on page 360 [7] - Reserved [6] - Connector Location[c] 0b = Primary 1b = Optional [5:3] - Reserved [2:0] - Physical Port Number within the connector (as defined in Section 10.3.3 on page 419) 0h = Reserved 1h = Port 1 2h = Port 2 3h = Port 3 All others reserved If **ConnectionType:ConnectionClass** = 011b (Module side), the field encodes the designation defined in Section 9.1.3.2, "Module Port Designations," on page 361 [7] - Reserved [6] - Connector Location[c] 0b = Primary 1b = Optional [5:4] - Connector Slot Number[c] within the slots the module occupies 0h = Reserved 1h = 1 - Primary 2h = 2 - Adjacent 3h = Reserved [3] - Reserved [2:0] - Physical Port Number within the connector (as defined in Section 10.3.3 on page 419) 0h = Reserved 1h = Port 1 2h = Port 2 3h = Port 3 All others reserved If **ConnectionType:ConnectionClass** = any others, field contents are undefined	RO	**Req**	**Req**	**Req**	**Req**

Table 155 PortConnectionInfo Record - Record Format Version=2

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
var	Additional NodeGUIDHandle/PortNo/ConnectionType/BackplaneConnection field sets, if any.	RO	**Req**[d]	Opt	Opt	**Req**[d]
var	**OEM** fields. A group of one or more OEM fields must be preceded with an OEM Prefix field.	-	Opt	Opt	Opt	Opt
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	-	Opt	Opt	Opt	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	-	Opt	Opt	Opt	Opt
1	Checksum	-	**Req**	**Req**	**Req**	**Req**

a. If a connection terminates within the enclosure housing the FRU, it is said to be internal.
b. The Link goes through a Repeater prior to leaving the FRU
c. See Section 3.2, "Physical Port," on page 58.
d. There **shall** be a PORTNO/Connection Type field pair for each IB connection.

13.7.10 MODULEPOWERINFO RECORD

The ModulePowerInfo record returns information about the power consumption, startup characteristics, and power management capabilities for a Module.

Table 156 ModulePowerInfo Record

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
3	**OperationalThermalPower** Maximum amount of power dissipated under normal operation including any vendor supplied exerciser over any sliding 60 second window. 0 = unspecified. [23:17] - Reserved [16: 0] - Power in mW. Range: 1 - 131071 mW	RO	**Req_UNA**	**Req_UNA**	**Req_UA**	**Req_UA**
2	**OperationalCurrent** Maximum amount of current drawn under normal operation including any vendor supplied exerciser across the range of VBulk. 0 = unspecified. [15:14] - Reserved [13: 0] - Current in mA. Range: 1-16383 mA	RO	**Req_UNA**	**Req_UNA**	**Req_UA**	**Req_UA**

Table 156 ModulePowerInfo Record

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
2	**IdleCurrent** Amount of current that a fully initialized device draws when waiting for functional requests. 0 = unspecified. [15:14] - Reserved [13: 0] - Current in mA. Range: 1-16383	RO	**Req_ UNA**	**Req_ UNA**	**Req_ UA**	**Req_ UA**
2	**InitCurrent** Amount of current drawn during Modules Built-In Self Test or self-initialization execution across the range of VBulk. 0 = unspecified. [15:14] - Reserved [13: 0] - Current in mA. Range: 1-16383	RO	**Req_ UNA**	**Req_ UNA**	**Req_ UA**	**Req_ UA**
2	**InitTime** The amount of time (in 10's of ms) from the end of Module Reset to when the Module is available for initialization by the Subnet Manager using SMPs and by other software. The InitTime includes any implemented BIST functionality. 0=unspecified.	RO	**Req_ UNA**	**Req_ UNA**	**Req_ UA**	**Req_ UA**
1	**ModulePMCapability**. [7:2] - Reserved [1] IsMStandbySupported 0 =: No 1 =: Yes [0] - IsPowerManagementSupported 0 = No 1 = Yes	RO	**Req**	**Req**	**Req**	**Req**
1	**UnitPMCapability** A multi-unit Module shares the same UnitPMInfo for all of its units [7:2] - Reserved [1] IsUSleepSupported 0 =: No 1 =: Yes [0] - IsUStandbySupported 0 = No 1 = Yes	RO	**Req**	**Req**	**Req**	**Req**

IBTA

Table 156 ModulePowerInfo Record

field size (bytes)	field	Read/Write	PA_Module	NPA_Module	N_ModuleXCA	N_ModuleSw
1	**ModulePowerParms** [7:6] = reserved [5:4] PowerClass 00b = unspecified 01b = Power Class I 10b = Power Class II 11b = reserved [3:2] = reserved [1:0] = RedundantPower Module incorporates redundant power converters. 00b = unspecified 01b = No redundancy 10b = Redundancy 11b = reserved	RO	**Req_ UNA**	**Req_ UNA**	**Req_ UA**	**Req_ UA**
var	**OEM fields**. A group of one or more OEM fields must be preceded with an OEM Prefix field.	-	Opt	Opt	Opt	Opt
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	-	Opt	Opt	Opt	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	-	Opt	Opt	Opt	Opt
1	**Checksum**	-	**Req**	**Req**	**Req**	**Req**

13.7.11 IOCPMINFO RECORD

The **IOCPMINFO** record indicates the power management capabilities of the IOCs associated with an IOUnit. If a Module contains multiple IOUnits

and several IOCs, the information about all such IOCs appear sequentially in the IOCPMInfo record.

Table 157 IOCPMInfo Record

field size (bytes)	field	Read/Write	PA_Module	N_ModuleXCA	N_ModuleSw
1	**IOC_Count** The number of nine-byte IOC capabilities (**IOCPMCapability** field through **IStandbyCurrent**) defined in this record.	RO	**Req**[a]	**Req**[a]	**Req**[a]
1	**IOCPMCapability** IOC Power Management Capability Mask [7] IsIDozeSupported 0: No 1: Yes [6] IsINapSupported 0: No 1: Yes [5] IsISleepSupported 0: No 1: Yes [4] IsIStandbySupported 0: No 1: Yes [3] WREisIDozeSupported 0: No 1: Yes - WRE can be generated by this node while in $\mathbf{I_{Doze}}$ State [2] WREisINapSupported 0: No 1: Yes - WRE can be generated by this node while in $\mathbf{I_{Nap}}$ State [1] WREisISleepSupported 0: No 1: Yes - WRE can be generated by this node while in $\mathbf{I_{Sleep}}$ State [0] WREisIStandbySupported 0: No 1: Yes - WRE can be generated by this node while in $\mathbf{I_{Standby}}$ State	RO	**Req**	**Req**	**Req**
2	**IDozeCurrent** The amount of total current drawn from the Bulk Power while in $\mathbf{I_{Doze}}$ PM state. Range: 0 - 16384 mA (0 - 16.384 A) **[15:14]** - Reserved **[13: 0]** - Current in mA	RO	**Req**	**Req**	**Req**

IBTA

Table 157 IOCPMInfo Record

field size (bytes)	field	Read/Write	PA_Module	N_ModuleXCA	N_ModuleSw
2	**INapCurrent** The amount of total current drawn from Bulk Power while in $\mathbf{I_{Nap}}$ PM state. Range: 0 - 16384 mA (0 - 16.384 A) **[15:14]** - Reserved **[13: 0]** - Current in mA	RO	**Req**	**Req**	**Req**
2	**ISleepCurrent** The amount of total current drawn from Bulk Power while in $\mathbf{I_{Sleep}}$ PM state. Range: 0 - 16384 mA (0 - 16.384 A) **[15:14]** - Reserved **[13: 0]** - Current in mA	RO	**Req**	**Req**	**Req**
2	**IStandbyCurrent** The amount of total current drawn from Auxiliary Power while in $\mathbf{I_{Standby}}$ PM state. Bulk Power drawn is defined to be 0A. Range: 0 - 16384 mA (0 - 16.384 A) **[15:14]** - Reserved **[13: 0]** - Current in mA	RO	**Req**	**Req**	**Req**
var	Additional nine-byte IOC capabilities (**IOCPMCapability** through **IStandbyCurrent**) field sets as indicated by IOC_Count.	-	Opt	Opt	Opt
var	**OEM fields**. A group of one or more OEM fields must be preceded with an OEM Prefix field.	-	Opt	Opt	Opt
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	-	Opt	Opt	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	-	Opt	Opt	Opt
1	Checksum	-	**Req**	**Req**	**Req**

a. There **shall** be one IOC capabilities for each IOC if one or more IOC's IsPowerManaged.

13.7.12 CMEINFO RECORD

The CMEInfo record contains additional information about the CME.

Table 158 CMEInfo Record

field size (bytes)	field	Read/Write	Requirement
1	**CMEGUID** All CMEGUID data bytes are encoded as binary, least-significant byte first, unless otherwise specified. 00h = none specified (0 data bytes follow) 01h = EUI-64 (8 data bytes follow) 02h = EUI-48 (6 data bytes follow) 03h = SMBIOS/IPMI/OSF/ Wired for Management/Microsoft GUID/UUID (16 data bytes follow) 04h = SMBus 2.0 UDID (8 bytes)	RO	**Req_UA**
var	CMEGUID data bytes. Field not present if CMEGUID field is set to 'none'.	RO	Opt
2	**CME Firmware Revision** in BCD Least Significant Number first. Format is xx.yy where the first byte holds BCD digits for the minor revision and the second byte holds two BCD digits for the major revision. 0000h = unspecified.	RO	**Req_UA**
var	**SlotNumbers** - One or more 8-bit bitfields indicating for which slots (identified by slot numbers) the CME provides IB-ML access. The field allows slot numbers to be non-sequential. Slot numbers must be in the range 1-254. 0 is assigned to the CME and its containing Chassis via CME. [7] - extension. 1b = There is at least one other SlotNumbers field byte following this one. 0b = This is the last byte of the SlotNumbers field in this record. [6:0] - Each bit represents one of seven corresponding slot numbers, with bit 1 of the first SlotNumbers field corresponding to slot number 1, and bit 0 of the second SlotNumbers field corresponding to slot number 7. (bit 0 of the first SlotNumbers field is reserved.)	RO	**Req_UA**
var	**OEM fields.** A group of one or more OEM fields must be preceded with an OEM Prefix field.	-	Opt
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	-	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	-	Opt
1	Checksum	-	**Req**

13.7.13 OEM RECORD

The OEM record contains OEM-specified data. The format of the data in the OEM Data portion of the field.

Table 159 OEM Record

field size (bytes)	field	Read/Write	Requirement
1	[7:4] **OEM ID Type** 1h = Company ID per EUI-64 / 48 (3 bytes) FFh, FFh, FFh = InfiniBand specified fields 2h = IANA Enterprise Number-based ID (3 bytes). FFh, FFh, FFh = reserved all others = reserved [3:0] **ID Length** - # of OEM ID data Bytes following	RO	**Req**
var	**OEM ID** data bytes	RO	**Req**
var	**OEM Data** Format per OEM identified by OEM ID Type and OEM ID data fields.	-	Opt
1	**Checksum**	-	**Req**

There are non-volatile writable regions in xInfo devices into which the CME, the Module, or the Baseboard Manager may write. The written data in specified writable areas **shall** follow the OEM Record format.

13.7.14 BUDDYINFO RECORD

The BuddyInfo record contains NodeGUIDs of InfiniBand Nodes and other entities that belong to the same enclosure.

There **shall** not be any connector or cable between any Node or other entity whose NodeGUID or GUID appear in a BuddyInfo Record. There **shall not** be any connector or cable between the Nodes or entities whose NodeGUID or GUID appear in a BuddyInfo Record and the Node reporting the BuddyInfo Record.

Table 160 BuddyInfo Record - Record Format Version=2

field size (bytes)		field	Read/Write	Requirement
1		**BuddyCount** [7:2] The number of InfiniBand nodes that belong to the same enclosure as the node reporting this record. 00000b = 'none' (No other node belongs to the same enclosure as this node.) 11111b = More than 31d [1:0] GUID encoding: All BuddyNodeGUID data bytes are encoded as binary, least-significant byte first, unless otherwise specified. 00b = EUI-64 (8 data bytes per GUID follow) 01b = EUI-48 (6 data bytes per GUID follow) 10b = SMBIOS/IPMI/OSF/ Wired for Management/Microsoft GUID/UUID (16 data bytes per GUID follow) 11b = SMBus 2.0 UDID (8 bytes per GUID follow)	R/W	**Req**
BuddyCount ×	var	**BuddyGUID** - data bytes for all the entities counted by BuddyCount. Field not present if **BuddyCount** field is set to 'none'. If GUID represents an IBA node, the value is the EUI-64 ID assigned for the node as reported in **NodeInfo:NodeGUID** to which the port is associated. (see *InfiniBand Architecture Specification, Volume 1*, Chapter "Subnet Management", Section "Subnet Management Class")	RO	**Req**
	1	**NodeGUIDHandle** - Unique value associated with the above NodeGUID. Serves as an index into **PortConnectionInfo** records.	RO	**Req**
var		**OEM fields.** A group of one or more OEM fields must be preceded with an OEM Prefix field.	R/W	Opt
1		**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	R/W	Opt
var		**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	R/W	Opt
1		Checksum	R/W	**Req**

13.7.15 ASSETTAG RECORD

As Table 161 shows, the AssetTag record is a writable Record which contains an AssetTag field.

Table 161 AssetTag Record

field size (bytes)	field	Read/Write	Requirement
1	FRU_Handle The FRU_Handle identifies Records that are associated with the same FRU. The scope of the value for the FRU_Handle is the xInfo device and its extension devices.	R/W	**Req**
1	**AssetTag** Encoding/Length (Table 148: Encoding/Length Byte)	R/W	**Req**
var	AssetTag data bytes	R/W	**Req**
var	**OEM fields.** A group of one or more OEM fields must be preceded with an OEM Prefix field.	R/W	Opt
1	**C0h** (Encoding/Length byte encoded to indicate no more info fields). [Only required when space exists between the end of the last field and the Checksum]	R/W	Opt
var	**PAD** [0 to N bytes]. An implementation is allowed to pad records with FFh values.	R/W	Opt
1	Checksum	R/W	**Req**

IBTA

CHAPTER 14: OS POWER MANAGEMENT

14.1 INTRODUCTION

"**Power Management**" is the set of state definitions and facilities that allow for an Operating System (or a prescribed agent) to control the power consumption of InfiniBand adapter module and, to a limited degree, switches.

Power Management of devices or media not directly attached to the InfiniBand fabric as an addressable node are outside the scope of this specification. (aka Fibre Channel or SCSI disks, Ethernet, etc).

14.2 OVERVIEW

14.2.1 POWER MANAGEMENT STATES

Power management policies rely on the existence of states which define the possible levels of power consumption. For InfiniBand, there are four (4) types of states defined:

- Port Power State (X)

  Applies to a given link in the fabric
- Module Power State (M)

  Applies to a module which may contain one or more I/O units
- Unit Power State (U)

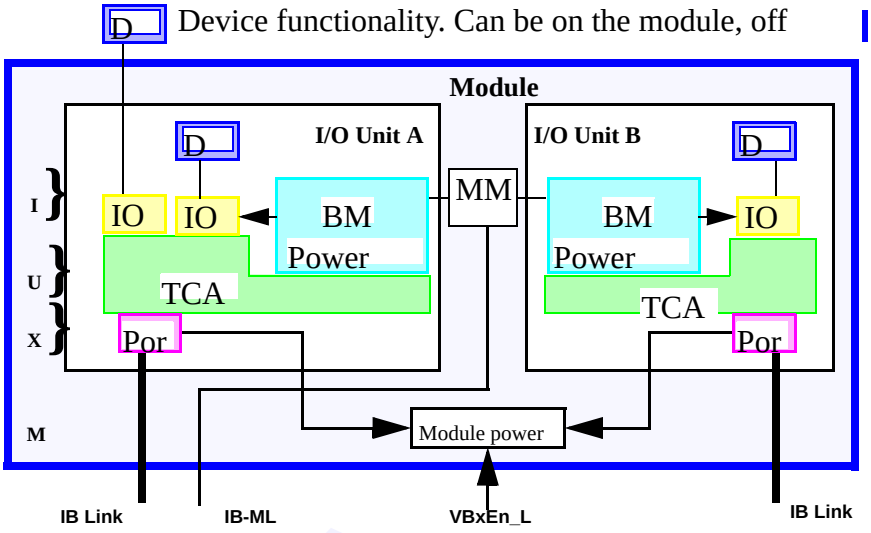
  Applies to the I/O unit at the end of a link, typically an xCA
- IOC Power State (I)

  Applies to the I/O Controller(s) (IOC) on an I/O Unit

14.2.1.1 META STATES

Some of the individual states within the above types are termed "meta" states because they are only discernible by implication. That is, they are derived states based on the independent states that make them up. There are no specific architectural facilities defining such states.

Figure 176 shows the applicability of these states to a generic InfiniBand module.



* Refer to Chapter 13: Hardware Management on IB-ML, IB-ML Agents and implications of multiple IB-ML interfaces

Figure 176 I/O Node Power Management Structure

Typically, operating system level software, based on policies it or the system owner establishes, controls the state entry and exit of a given unit.

In this chapter, the software component responsible for power management policy is referred to as the “Power Manager”. Through its policies, this software may place none, some or all of the units that it owns into varying degrees of power management states such to achieve the savings desired. It is outside the scope of the InfiniBand specification to define these policies; rather, InfiniBand defines mechanisms through which implemented policies may be realized. These policies may be based on system load, timed operations, or external events.

The majority of the facilities that control the Power Management states are architected as part of the Baseboard Management class; thus, the Power Manager must have software relationship with the Baseboard Manager for the fabric to be able to access these facilities. This relationship is not specified. However, possible implementations include, but are not limited to:

- having the Power Manager be a component of the Baseboard Manager,
- having the Power Manager be separate and utilize the Baseboard Manager as a proxy to the control facilities,

- having the Power Manager directly controlling the states through a sharing of B-Keys.

Additionally, the Power Manager would need a relationship with the Subnet Manager if it implements policies that utilize the X States.

Some of the states defined allow for the continued functionality of the InfiniBand links such that the Power Manager can, through the use of defined Baseboard management datagrams (MAD), restore a unit to an operational state. However, to achieve the highest degrees of power savings, states are defined that allow node implementations to remove bulk power from the unit and yet allow for it to be powered back on from the InfiniBand fabric. Additionally, means are defined to allow for an InfiniBand unit to indicate that it, having been previously configured to do so, desires to have the system be restored to the operating condition as defined by the policy. These are known as “Wake Request Events”.

14.2.2 WAKE REQUEST EVENTS

Wake Request Events (WREs) are those events that can be produced by an InfiniBand module that desires to return a “sleeping” or “low power” system to an operational state given the system had been previously enabled to do so. The actions that need to take place when a device recognizes a WRE are:

1) Provide an indication to the chassis power subsystem to affect a "power on" as needed
2) Provide an indication to the “Power Manager” that the WRE has occurred such that this agent can perform the remaining wake-up actions as set by system policy.

14.2.3 SIMPLE SYSTEM TOPOLOGY

Figure 177 depicts a basic two power domain topology for Power Management.

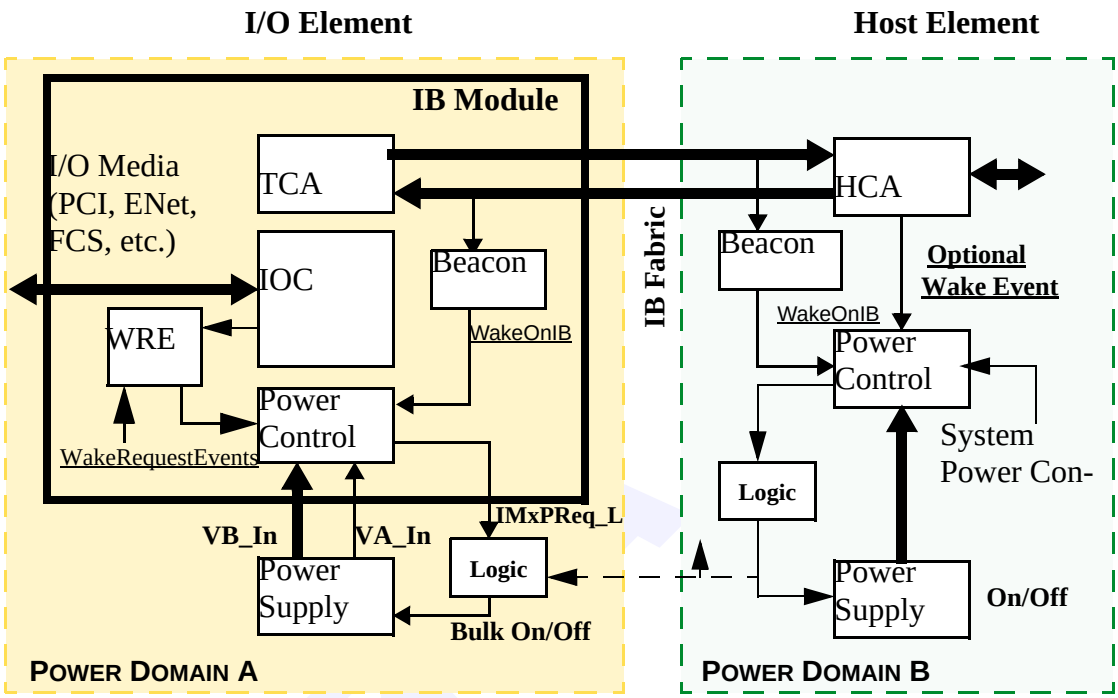


Figure 177 Power Management Topology

In the more generic case, one or more switch elements can be inserted between the I/O Unit and the Host whereby Power Management operations (state changes and wake-up "events") must be routable between these.

14.2.4 LOW POWER STATE ENTRY

Through the Baseboard class and the ***[Set/Get]ModulePMControl***, ***[Set/Get]UnitPMControl*** and ***[Set/Get]IOCPMControl*** attributes, the Power Manager software can control the transition to a low powered state. See Section 13.6.2, "Baseboard MAD Commands (BM MAD Attributes)," on page 570.

Hardware will, upon link training failure, enter either the $X_{Polling}$ or X_{Sleep} state based on link port attribute modifiers. See Section 5.6, "Link Initialization and Training," on page 118 for details.

14.2.5 LOW POWER STATE EXIT

A node may exit from a low power state due to one of two mechanisms:

- An event generated by the node. These are termed Wake Event Requests (WRE).
- Activity on the InfiniBand link indicating wake-up. This is termed WakeonIB.

Further, the waking due to the InfiniBand fabric may be due to one of two source actions:

- Power management or other software targeted at specific link or node using already powered portions of the fabric.
- The cascading of a WakeonIB conditions through the use of Beacon Sequences through a low powered fabric

14.2.5.1 WAKE REQUEST EVENT

IOCs may detect or otherwise generate Wake Request Events (WRE), if enabled. The module enables its power converter and asserts the **IMxPReq_L** signal on the backplane connector. This signal being asserted may cause the system power supply to transition from OFF to ON if the system power is enabled to do so. If the system power is already in the ON state, this assertion is ignored.

The enablement of module power, either due to the converter or system power or both, will cause the module to enter the Link Polling state whereby a training sequence (otherwise known as the “Beacon sequence”) is periodically transmitted on the InfiniBand link. Upon establishment of a trained link, further state transitions are under software control.

On the companion end (“Host Entity” in this topology), the presence of the beacon sequence on the IB link will generate a logical WakeOnIB signal. This signal, powered from $\mathbf{V_{Aux}}$, is used to transition the Host/Target system power supply for that domain to ON ($\mathbf{M_{On}}$ or equivalent). The WakeOnIB logical signal is typically ignored if the power supply is in the ON state.

The software interface defined below to transition to a low powered state may also be used to transition from a low powered state if it is accessible.

14.2.5.2 WAKE ON FABRIC

For the case that the Wake Event Request was generated on an HCA node or that software on a powered node desires to cause other parts of the fabric to return to operational conditions, the Wake on Fabric mechanism is used.

A Wake on Fabric operation is a two step process.

1) The Power Manager interfaces with the Subnet Manager to have a link taken from Link Sleeping ($\mathbf{X_{Sleep}}$) to Link Up ($\mathbf{X_{On}}$) state such to

reestablish the link to a low power endnode. The process of bringing the link to the Link Up state involves the use of the Beacon Sequence and will effectively produce a WakeOnIB logical signal that will restore power ($\mathbf{M_{On}}$) to the node if enabled to do so.

2) The Power Manager then controls the Unit and IOC states using ***[Get/Set]UnitPMControl*** and ***[Get/Set]IOCPMControl*** attributes, respectively.

14.2.6 POWER SUBSYSTEM NOTIFICATION OF WRE

Systems using InfiniBand may be built in one of two basic structures:

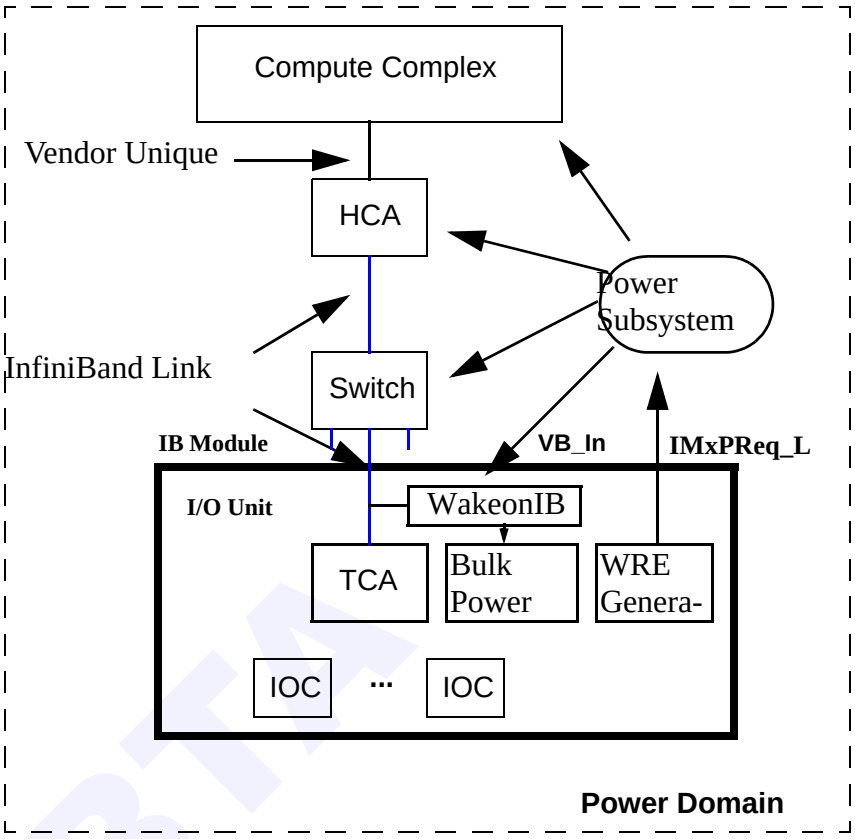
- **Single Domain** - where all nodes capable of generating a wake request event are in the same power domain as the node or nodes that will execute the policy based actions.
- **Multiple Domain** - where any nodes capable of generating a wake request event is in a different power domain as the node or nodes that will execute the policy based actions.

Each of these domains can be viewed as a “Local Domain” relative to the InfiniBand devices that reside in them. In the case of the “Single Domain” structure, the “Local Domain” is synonymous with “Single Domain”.

14.2.6.1 SINGLE DOMAIN / LOCAL DOMAIN

The Single Domain structure represents the single board system topology typical in many low-end systems (desktops, workstations, single box servers). In this case, the notification mechanism available from the WRE detecting adapter board can be used to wake-up the power subsystem of the entire domain.

The indication to the local power subsystem happens through the asser-



Note: For this example, the IB Module only contains one

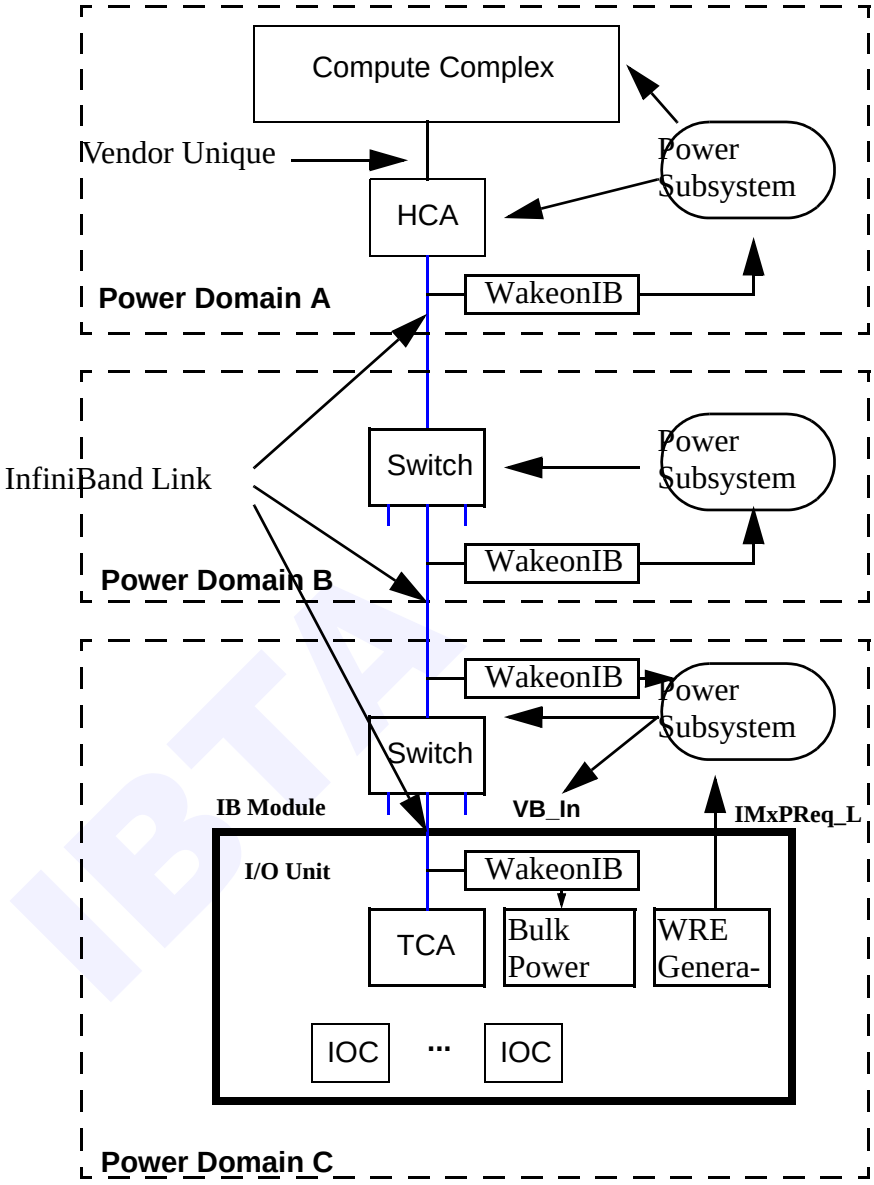
Figure 178 Single Power Domain Example

tion of **IMxPReq_L** pin upon the Wake Request Event occurrence. The chassis **should** interpret this line in the following manner:

- If the bulk power supplies are off (system is in “standby” or “auxiliary” power state) AND **IMxPReq_L** is asserted, then the bulk supplies should be enabled. If the bulk supplies are enabled AND the **IMxPReq_L** is asserted, the indication is ignored from a power viewpoint.

14.2.6.2 MULTIPLE DOMAINS

The Multiple Domain structure represents the modular system case where InfiniBand modules and their local switches may be on one domain, some number of other switches that make up the fabric can be in one or more domains, and the HCA nodes can be on any of these or still other domains. It is up to OS power management policy to decide the level of power savings and exercise control over some, none or all of a subnet. Figure 179 Multiple Power Domain Example on page 638 depicts this topology.



Note: For this example, the IB Module only contains one

Figure 179 Multiple Power Domain Example

The “wake-on” indication that spans power domains is produced on the InfiniBand link in the form of Training Sequence 1 as described in Chapter 5: Link/Phy Interface on page 79. For the purposes of Power Management, this is also referred to as the “Beacon Sequence”.

There are three possibilities for the state of a given local domain (i.e. that domain receiving an incoming beacon sequence on a port) that are addressed:

1) Bulk Power On, Auxiliary Power On

   Training Sequence 1 received on a port that is not in Disabled state simply begins the normal training process without any additional interaction with the power subsystem.

2) Bulk Power Off, Auxiliary Power On

   The detection of Training Sequence 1 must produce an indication to the local power subsystem similar to that as described in Section 14.2.6.1, “Single Domain / Local Domain,” on page 636. Once power is restored, further receipt of Training Sequence 1 begins the normal link training process to bring a link active.

   Note: this version of the specification does not define the means for restoring a switch from auxiliary power while retaining its context.

3) Bulk Power Off, Auxiliary Power Off

   As no power is available, the Training Sequence 1 can not be detected and is ignored.

Given power is restored to the domain, further operations flow as InfiniBand defined sequences or packets.

14.2.7 CONTEXTS

"Context" is information or states that pertain to a given aspect from which functions can resume operation. This section describes those contexts that are pertinent to Power Management.

14.2.7.1 FUNCTION CONTEXT

The Function Context is that information specific to the implemented device function that indicates the device’s operational state. Initialized Function Context is that state after the hardware device and any associated driver has initialized but operations have not commenced.

14.2.7.2 FABRIC CONTEXT

The Fabric Context is that information which allows for management operations to be received and transmitted on the InfiniBand fabric. This includes, but is not limited to, the following:

- M_Keys and associated control bits
- P_Keys and associated control bits
- B_Keys and associated control bits
- QP0 context

IBTA

- QP1 Baseboard Class

14.2.7.3 WRE CONTEXT

The WRE Context is the state information that allows for the enabling, detection, generation, and reporting of Wake Request Events.

Specifically, this includes Attribute ***IOCPMControl***, Components ***WREEnabled*** and ***WREStatus***.

14.2.7.4 SWITCH CONTEXT

The Switch Context is that information that allows for the distribution of a Beacon Sequence event.

If the switch has bulk power remaining to it, no explicit power management context is required. If a Beacon Sequence is received on a port and the port successfully trains, the Subnet Manager will find the state change of the port upon its next fabric sweep. Optionally, the switch may produce a ***SubnTrap(NewNodeonFabric)*** to the Subnet Manager. The Power Manager should subscribe for these events for nodes that it uses the $\mathbf{X_{Sleep}}$ state for as part of its policy.

Auxiliary powered switches are not architected in this version of the specification.

14.2.8 POWER MANAGER NOTIFICATION

System power management policy and system design affect the "entry" point of the WRE notification into the "in-band" portion of the system. This is basically enabling a given set of end nodes to wait on a WRE while also enabling a potentially different set of end nodes to get notified of the WRE. Given system policy and implemented features, this may involve the interaction with one or more power systems (due to the power domains in which the intervening switches reside) along the path from the "detecting" end node and the "notification" end node(s).

Ultimately, this notification needs to get to a Channel Adapter that, upon power reenablement, allows the Power Manager to gain knowledge that a WRE has occurred such that it can then perform policy based function restoration.

The mechanism for passing this notification from the "detecting" end node to the "notification" end node involves the following operational steps:

1) An Auxiliary Powered I/O Unit generates a WRE.
2) The local power domain is brought up using the mechanism described in Section 14.2.6.1, “Single Domain / Local Domain,” on page 636.

3) The I/O Unit sends Training Sequence 1 as a Beacon Sequence to its companion, presumably a Switch or HCA.

4) The Switch or HCA is either on Bulk Power or Auxiliary Power.

   If the Switch or HCA is on Bulk Power, then:

   - The "waking link" is trained;
   - If a Switch, the training of a previous Sleeping or Polling link has the Switch create ***SubnTrap(NewNodeonFabric)*** message to the Subnet Manager.
   - If an HCA, it provides implementation specific notification to interested entities in that node

   If the Switch or HCA is on Auxiliary Power, then

   - The HCA power subsystem is enabled as described in Section 14.2.6.2, “Multiple Domains,” on page 637.
   - A Switch on Auxiliary Power is not defined in this version of the specification.

14.3 PORT POWER MANAGEMENT STATES

The Port Power Management States are meta[5] states that combine the operational states of the link with the power available. They are defined to facilitate a bridge between power management terminology and operational link states.

Table 162 summarize these states; they are defined fully in the following sections.

5. Meta is defined in Section 14.2.1.1, “Meta States,” on page 631.

Table 162 Port Power Management States

X State	Port Power		Link States[a]	Power Consumption	Port Implementation	Near End Link Recovery Latency
	Recv	Xmit				
X_{On}	Bulk	Bulk	LinkUp, Configuration, Recovery	Full	Required[b]	None
$X_{Polling}$	Bulk	Bulk	Polling	Low	Required[b]	1 ms
X_{Sleep}	Bulk	Bulk	Sleeping	Low	Required[b]	1 ms
$X_{Disabled}$	Bulk	Bulk	Disabled	Low	Required[b]	N/A
$X_{Standby}$	Aux	Off	Beacon Enabled[c]	Lowest (while restorable)	Optional[d]	100 ms
X_{Off}	Off	Off	N/A	None	Given	Indefinite

a. The states shown are defined in Section 5.6.4, "Link Training State Machine," on page 123 except as noted.
b. The requirement for this state is defined in Section 5.6, "Link Initialization and Training," on page 118 and repeated here for a complete description of these meta states.
c. This link state is not defined in Section 5.6.4. Rather, this is the enablement of the Beacon Detection circuitry which may or may not be present in the same device that implements the other Link States.
d. Based on whether $M_{Standby}$ is supported (as indicated by *IsMStandbySupported* in *ModulePowerInfo* record.

14.3.1 X_{ON} LINK POWER STATE

This state indicates that the InfiniBand link is fully powered. Any link state, as reported by ***Subn.PortInfo(LinkState),*** other than Port Polling, Port Sleep, and Port Disabled defines the port as being in $\mathbf{X_{On}}$.

14.3.2 $X_{POLLING}$ LINK POWER STATE

This state is defined as having the port receiver look for the beaconing sequence from the companion end while periodically sending a beacon se-

quence on the port transmitter. The companion end may or may not be present.

Architecture Note

This state is considered a low power state for an uncontested port but was not designed for power management policy of a connected link. This is due to the fact that if the companion end is in a low power state (X_{Sleep} or $X_{Standby}$), the receipt of a beacon sequence during the poll will cause a transition to X_{On}. Thus, power management policy software would be advised to use X_{Sleep} for placing a powered port which is connected to a companion into a low power state.

14.3.3 X_{SLEEP} LINK POWER STATE

This state is defined as having the port receiver look for the beaconing sequence from the companion end while leaving the port transmitter static. The companion end may or may not be present.

Architecture Note

This state is useful when waiting for a hot add event or when one or both ends is in a power saving mode. Either end can initiate a wake-up event by sending the Beacon Sequence. In the case of an endnode, this would happen due to a WRE or initial power-up of the endnode. In the case of a switch, this could also be due to Power Management software desiring to reenable the link.

14.3.4 $X_{STANDBY}$ LINK POWER STATE

This state is defined as having the port receiver look for the beaconing sequence while operating on auxiliary power. The companion end may or may not be present.

14.3.5 X_{OFF} LINK POWER STATE

This state is defined as having neither Bulk nor Auxiliary Power available to the port and the port will not respond to the beaconing sequence.

14.4 MODULE POWER STATES

Module Power States control the local power converter enablement for InfiniBand defined pluggable modules. Other nodes on the fabric that are not packaged on these form factors may or may not utilize these states.

Table 163 provides a summary of the Module Power states that are fully defined in the following sections.

Table 163 Module Power States

Module Power State	Bulk Power Converter	Auxiliary Power	Wake Request Event (if supported)	Requirement
M_{On}	Enabled	Enabled	No effect	Required
$M_{Standby}$	Disabled	Enabled	Enabled	Optional
M_{NoWake}	Disabled	Enabled	Disabled	Meta
M_{Off}	Disabled	Disabled	Disabled	Meta

o14-1: All modules that implement Power Management as indicated by *ModulePowerInfo.ModulePMCapability(IsPowerManagementSupported)*=1 **shall** provide the "M" states indicated as “Required” in the column labeled “Requirement” in Table 163 and provide access and control of these states through the ***[Get/Set]ModulePMControl*** Attributes defined in Section 13.6.2, “Baseboard MAD Commands (BM MAD Attributes),” on page 570.

14.4.1 M_{ON}

The **M_{On}** state indicates that the power converters on the module that are required for operation are enabled.

14.4.2 $M_{STANDBY}$

$M_{Standby}$ indicates that the power converter(s) on the module are disabled but can be reenabled without Backplane intervention. Auxiliary power is available.

o14-2: All modules that implement the **$M_{Standby}$** state as indicated by *ModulePowerInfo.ModulePMCapability(IsMStandbySupported)*=1 **shall** provide the ability to detect a Beacon Sequence on the InfiniBand link. A delay of 10ms +100/-0% **shall** be provided between the entry into **$M_{Standby}$** and the enablement of the Beacon Detect circuitry to allow for the link quiescent condition to be achieved.

14.4.3 M_{NOWAKE}

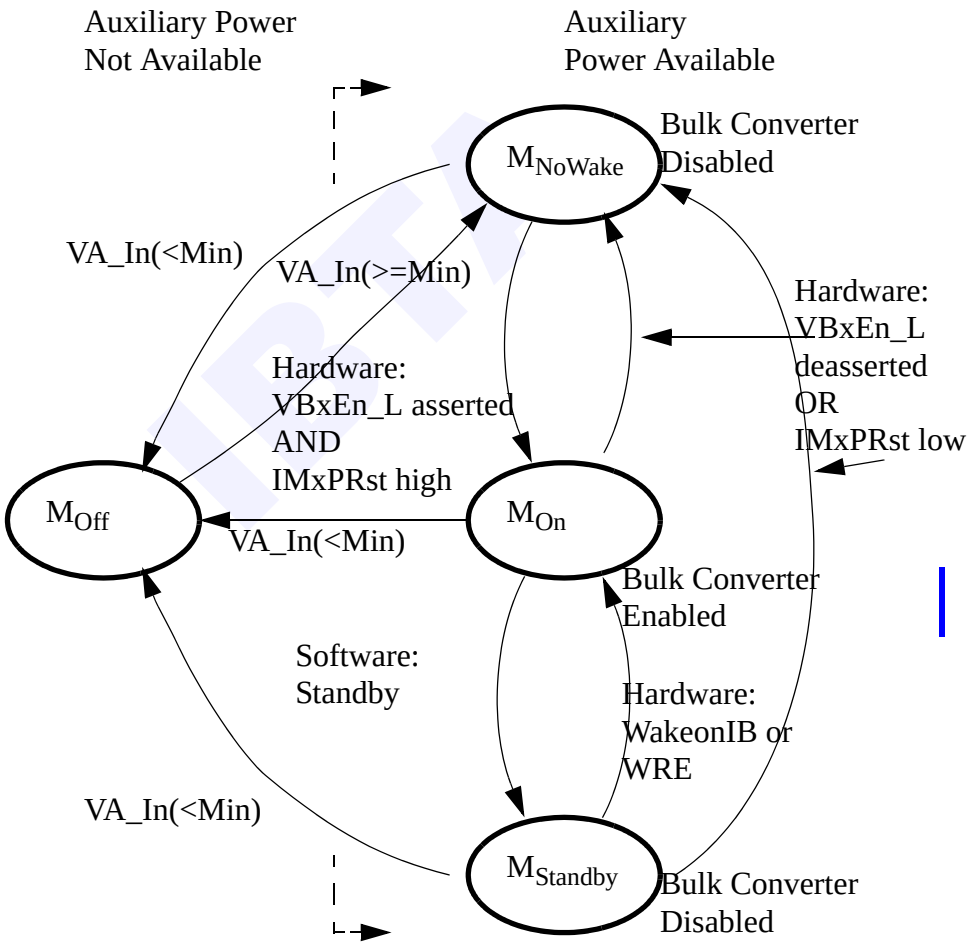
The **M_{NoWake}** state indicates that power converter(s) on the module are disabled by the backplane through **VBxEn_L** and can only be enabled by the backplane using **VBxEn_L**. Auxiliary power is available. Entry into and exit out of this state is exclusively under control of the backplane.

14.4.4 M_{OFF}

The **M_{Off}** state is a mechanical off state whereby either the module is not inserted in a slot or that the chassis in which the module is plugged has no auxiliary power available. In this state, the module does not receive Bulk power

14.4.5 MODULE POWER STATE TRANSITIONS

Figure 180 shows the valid transitions for the Module Power States.



Note: All software transitions shown are done through the ***BMSend.SetModulePMControl*** operation. The “value” shown is the mnemonic representation of a state value.

Figure 180 Module Power State Diagram

14.5 I/O UNIT POWER STATES

I/O Unit States are defined for endnodes that have a TCA and one or more IOCs having one or more power management functions implemented.

Table 164 provides a summary of the I/O Unit Power states that are defined fully in the following sections.

Table 164 I/O Unit Power States

Unit Power State	Module State	Wake Request (if supported)	Requirement
U_{On}	M_{On}	Don't Care	Required
U_{Sleep}	M_{On}	Don't Care	Optional
$U_{Standby}$	$M_{Standby}$	Enabled	Meta
U_{Off}	M_{Off}	Disabled	Meta

o14-3: All modules that implement Power Management as indicated by *ModulePowerInfo.ModulePMCapability(IsPowerManagementSupported)*=1 **shall** provide the "U" states indicated as "Required" in the column labeled "Requirement" in Table 164 and provide access and control of these states through the ***[Get/Set]UnitPMControl*** Attributes defined in Section 13.6.2, "Baseboard MAD Commands (BM MAD Attributes)," on page 570.

14.5.1 U_{ON}

The U_{On} state indicates that the I/O Unit is generally enabled for operations. Specific power management policy is further influenced through affecting the "*I*" states of underlying IOCs.

14.5.2 U_{SLEEP}

o14-4: All modules that implement the U_{Sleep} state as indicated by *ModulePowerInfo.UnitPMCapability(IsUSleepSupported)*=1 **shall** provide the functionality described in Section 14.5.2.

The U_{Sleep} state indicates that the I/O Unit is functionally disabled, but the TCA is capable of responding to the following if its Port is in X_{on}:

- All QP0 operations
- QP1 operations:
  - BMSend.GetUnitPMControl
  - BMSend.SetUnitPMControl

- BMSend.GetModulePMControl
- BMSend.SetModulePMControl

From the IB S/W point of view, the underlying IOCs are not accessible in this state and may be in I_{Off} state.

If a port is not in X_{on}, although not accessible, the underlying IOCs and I/O Unit may be in any state.

14.5.3 $U_{STANDBY}$

The $U_{Standby}$ state is a meta state indicating that the module on which this I/O Unit resides is in $M_{Standby}$.

From the software point of view, the underlying IOCs are not accessible in this state and may be in I_{Off} state.

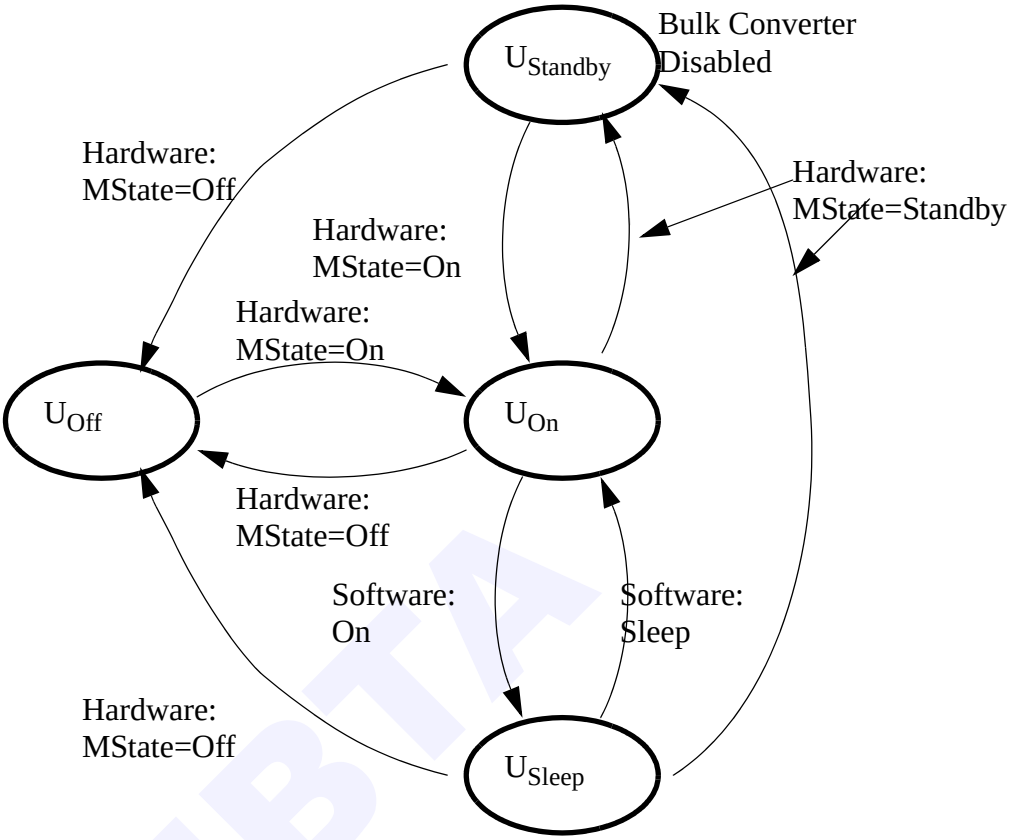
14.5.4 U_{OFF}

The U_{Off} state is a mechanical state of having both bulk and auxiliary power being unavailable. If the module on which this I/O Unit resides is in M_{Off}, the unit is considered to be in this state.

14.5.5 I/O UNIT POWER STATE TRANSITIONS

Figure 181 shows the valid transitions for the I/O Unit Power States.

IBTA



Note: All software transitions shown are done through the ***BMSend.SetUnitP-MControl*** operation. The "value" shown is the mnemonic representation of a state value.

Figure 181 I/O Unit Power State Diagram

14.6 IOC POWER STATES

IOC Power States are defined for IOCs that may have one or more power management functions implemented.

Table 165 provides a summary of the IOC Power states that are defined fully in the following sections.

Table 165 IOC Power States

IOC State	Operations to Unit	Operations from Unit	Context	Power	Latency to $I_{Operational}$	Requirement
$I_{Operational}$	All	All	Function	Full	None	Required
I_{Uninit}	Driver Specified	Driver Specified	None	Full	Drv Spec'd	Required
I_{Doze}	QP0, QP1 Base-board Class	WRE Trap	Function, WRE[a]	Low	us	Optional
I_{Nap}	QP0, QP1 Base-board Class	WRE Trap	Function, WRE[a]	Lower	ms	Optional
I_{Sleep}	QP, QP1 Base-board Class	WRE Trap	WRE[a]	Lower	s	Optional
$I_{Standby}$	Beacon Sequence	None	WRE[a]	<1.2W		Optional
I_{Off}	None	None	None	None	N/A	Meta

a. WRE Context is only maintained for those IOCs that support Wake Request Event generation.

o14-5: All modules that implement Power Management as indicated by *ModulePowerInfo.ModulePMCapability(IsPowerManagementSupported)*=1 **shall** provide the "I" states indicated as "Required" in the column labeled "Requirement" in Table 165 and provide access and control of these states through the ***[Get/Set]IOCPMControl*** Attributes defined in Section 13.6.2, "Baseboard MAD Commands (BM MAD Attributes)," on page 570.

14.6.1 $I_{OPERATIONAL}$

The $I_{Operational}$ state indicates that the IOC is fully operational.

14.6.2 I_{UNINIT}

The I_{Uninit} state is transitional and is not visible to Power Management software. It indicates that the IOC is capable of running at rated speed although it is not currently operational as the IOC has not been initialized by a driver.

14.6.3 I_{DOZE}

o14-6: All modules that implement the I_{Doze} state as indicated by *IOCPMInfo.IOCPMCapability(IsIDozeSupported)*=1 for an IOC **shall** provide the functionality described in Section 14.6.3 and provide the *IOCPMInfo.IDozeCurrent* field for that IOC.

I_{Doze} is a light sleep state where some functional activity may be occurring on the board (such as media network monitoring for special packets) but other activity is typically suppressed. This state requires the IOC to maintain implementation specific context to fully recover to its ***$I_{Operational}$*** condition upon the detection of a “Wake-on” event.

In the ***I_{Doze}*** state, the IOC **shall** only initiate link operations when a Wake Request Event has occurred whereby a ***BMTrap.WRE*** is sent to the Baseboard Manager.

14.6.4 I_{NAP}

o14-7: All modules that implement the **I_{Nap}** state as indicated by *IOCPMInfo.IOCPMCapability(IsINapSupported)*=1 for an IOC **shall** provide the functionality described in Section 14.6.4 and provide the *IOCPMInfo.INapCurrent* field for that IOC.

I_{Nap} is a long sleep state where an endnode is not intended to be used for some amount of time as set by system policy. This state requires the IOC to maintain implementation specific context to fully recover to its ***$I_{Operational}$*** condition upon the detection of a “Wake-on” event.

In the ***I_{Nap}***state, the IOC **shall** only initiate link operations when a Wake Request Event (WRE) has occurred whereby a *BMTrap.WRE* is sent to the Baseboard Manager.

14.6.5 I_{SLEEP}

o14-8: All modules that implement the **I_{Sleep}** state as indicated by *IOCPMInfo.IOCPMCapability(IsISleepSupported)*=1 for an IOC **shall** provide the functionality described in Section 14.6.3 and provide the *IOCPMInfo.ISleepCurrent* field for that IOC.

I_{Sleep} is a long sleep state where an endnode is not intended to be used for some amount of time as set by system policy. This state requires the IOC to only maintain WRE context to fully recover to its ***$I_{Operational}$*** condition upon the detection of a “Wake-on” event.

In the ***I_{Sleep}***state, the IOC **shall** only initiate link operations when a Wake Request Event has occurred whereby a ***BMTrap.WRE*** is sent to the Baseboard Manager.

14.6.6 $I_{STANDBY}$

o14-9: All modules that implement the **$I_{Standby}$** state as indicated by *IOCPMInfo.IOCPMCapability(IsIStandbySupported)*=1 for an IOC Management **shall** provide the functionality described in Section 14.6.3 and provide the *IOCPMInfo.IStandbyCurrent* field for that IOC.

$I_{Standby}$ is a truly dormant state whereby only auxiliary power is provided to the IOC. No function context is maintained in this state. If Wake Request Events are supported from ***$I_{Standby}$***, the associated WRE context **shall** be maintained. Full initialization by software will need to take place when exiting this state (to ***I_{Uninit}***).

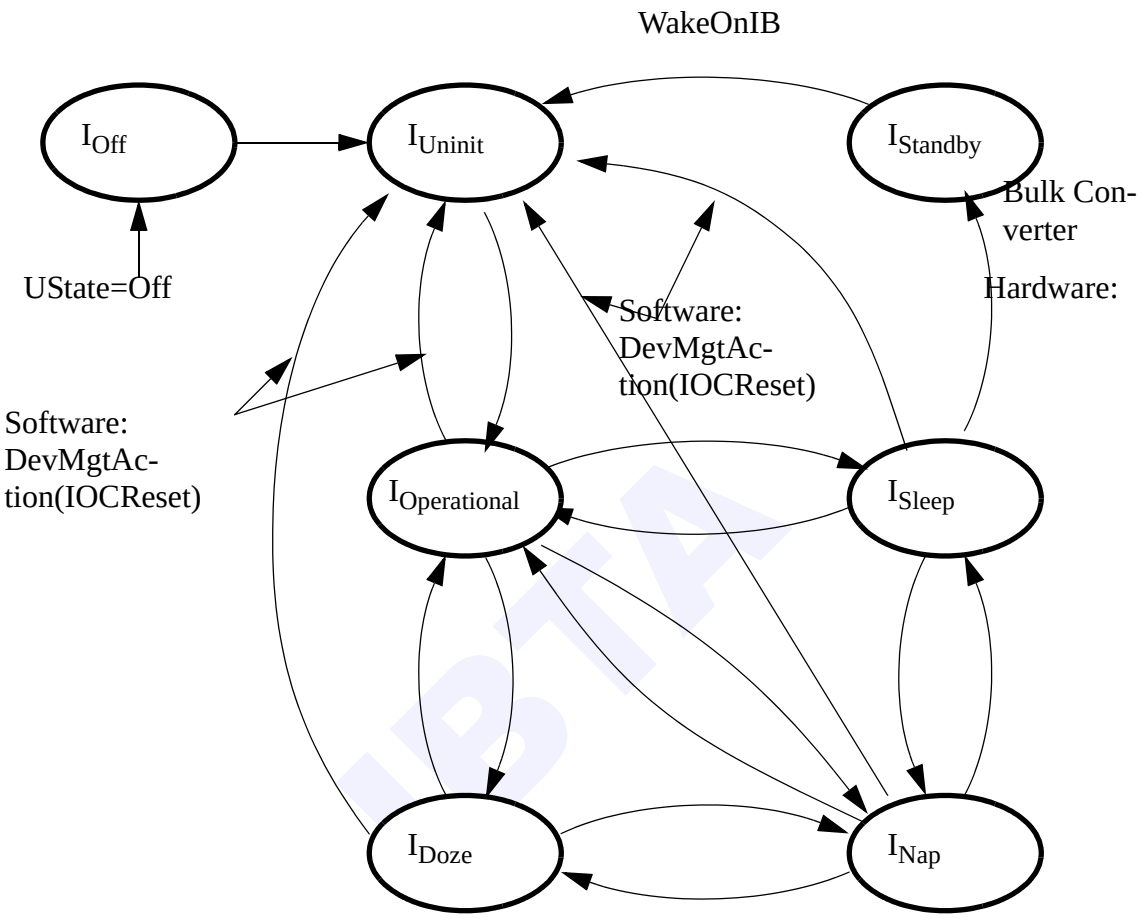
14.6.7 I_{OFF}

I_{Off} is the logical state of either mechanical off (the IOC having no bulk or auxiliary power available). No context is retained in this state.

IBTA

14.6.8 IOC POWER STATE TRANSITIONS

This section shows, via a state diagram, the valid transitions between each of the IOC Power states.



Note: All unlabeled transitions are affected by performing ***BMSend.SetIOCPMControl(IOC Power State)*** operations.

Figure 182 IOC Power Management State Diagram

14.7 STATE COMBINATIONS

This section shows the combinations of Transceiver Power states and Unit Power state and, where valid, the parameters that pertain.

14.7.1 $I_{OPERATIONAL}$

Table 166 $I_{Operational}$ Power State Combinations

Module State	Unit State	Xcvr State	IOC State	Operations to IOC	Operations from IOC	Context	Power
M_{On}	U_{On}	**X_{On}**	$I_{Operational}$	All	All	Function, Fabric	Full
< ---------All others---------->			$I_{Operational}$	Invalid			

14.7.2 I_{DOZE}

Table 167 I_{Doze} Power State Combinations

Module State	Unit State	Xcvr State	IOC State	Operations to IOC	Operations from IOC	Context	Power
M_{On}	U_{On}	**X_{On}**	I_{Doze}	BMSend[a], DevMgtAc-tion(IOCreset)	WRE[b]	Function, Fabric WRE[b]	<Full
< ---------All others---------->			I_{Doze}	Invalid			

a. Required to exit this state.
b. If supported in **I_{Doze}** mode

14.7.3 I_{NAP}

Table 168 I_{Nap} Power State Combinations

Module State	Unit State	Xcvr State	IOC State	Operations to IOC	Operations from IOC	Context	Power
M_{On}	U_{On}	**X_{On}**	I_{Nap}	BMSend[a], DevMgtAc-tion(IOCreset)	WRE[b]	Function, Fabric, WRE[b]	<Full
M_{On}	U_{On}	**X_{On}**	I_{Nap}	TS1[c]	WRE[b]	Function, WRE[b]	<Full
< ---------All others---------->			I_{Nap}	Invalid			

a. Required to exit this state.
b. If supported in **I_{Nap}** mode
c. Required to exit **X_{Sleep}** state

IBTA

14.7.4 I_{SLEEP}

Table 169 I_{Sleep} Power State Combinations

Module State	Unit State	Xcvr State	IOC State	Operations to IOC	Operations from IOC	Context	Power
M_{On}	U_{On}	**X_{On}**	I_{Sleep}	BMSend[a], DevMgtAction(IOCreset)	WRE[b]	Function, Fabric, WRE[b]	Full
M_{On}	U_{On}	**X_{On}**	I_{Sleep}	TS1[c]	WRE[b]	Function, WRE[b]	<Full
M_{On}	U_{Sleep}	**X_{Sleep}**	I_{Sleep}	TS1[c]	WRE[b]	WRE[b]	<Full
< ---------All others---------->			I_{Sleep}	Invalid			

a. Required to exit this state.
b. If supported in **I_{Sleep}** mode
c. Required to exit **X_{Sleep}** state

14.7.5 $I_{STANDBY}$

Table 170 $I_{Standby}$ Power State Combinations

Module State	Unit State	Xcvr State	IOC State	Operations to IOC	Operations from IOC	Context	Power
$M_{Standby}$	$U_{Standby}$	**$X_{Standby}$**	$I_{Standby}$	TS1[a]	**IMxPReq_L**[b]	WRE[b]	<1.2W
< ---------All others---------->			$I_{Standby}$	Invalid			

a. Required to exit this state from the fabric. A local Wake Event can also cause an exit if supported.
b. If WRE is supported in this mode

14.8 SWITCH POWER MANAGEMENT

The architecture provides for the ability to put individual ports into a low power state to allow system policy to reduce power dissipation if desired.

If the port is in a low power mode (**X_{Sleep}**), two ways are provided to exit this mode and have the port return to an active state:

1) under management software control
2) due to a hot plug or WRE event

Auxiliary powered switches are not architected in this version of the specification.

14.8.1 HOT PLUG EVENTS

For the case of hot plug events, upon the completion of a module power-on, a Training Sequence 1 is sent from the plugged module to the companion port. This will initiate the full training sequence described in Section 5.6, "Link Initialization and Training," on page 118.

14.8.2 WRE EVENTS

For WRE events, there are two cases: 1) the WRE endnode is local (in the same power domain as this switch) and 2) the WRE endnode is remote (in some other domain from the switch).

For the local domain case, the architecture only specifies the case where the switch has had bulk power maintained to it. Thus, the WRE event on the detecting endnode looks like a hot plug event (See Section 14.8.1).For the remote domain case, there are 2 possibilities: 1) the switch is currently powered and 2) the switch's power domain is currently not powered, other than auxiliary power, and needs to be enabled. For the case of the switch is powered and the port is in an $\mathbf{X_{Sleep}}$ state, the WRE is seen as a Beacon Sequence and looks like a hot plug event.

As previously stated, the case where the switch is on auxiliary power is not architected in this version of the specification.

IBTA

CHAPTER 15: VOLUME 2 COMPLIANCE SUMMARY

15.1 COMPLIANCE DEFINITION

This chapter specifies the Compliance Categories which are approved for labeling various products which contain InfiniBand content. This will allow vendors to label their products and claim InfiniBand compliance without creating confusion in the marketplace. This chapter addresses compliance to the feature set defined by Volume 2 of the InfiniBand Specification. Additional cross-references to the Compliance Summary for Volume 1 of the InfiniBand Specification will be added when necessary.

15.1.1 PRODUCT APPLICATION

Each product, which has InfiniBand content, **may** claim InfiniBand Compliance at one or more of the categories defined in the Compliance Summaries of the InfiniBand Specification. A product **shall not** simply claim "InfiniBand Compliant". Each claim of compliance **shall** be composed of two parts as follows:

Volume 2 Compliance Category: Volume 1 Compliance Category

The Volume 1 Compliance categories are defined in *InfiniBand Architecture Specification, Volume 1*, Chapt 20: Volume 1 Compliance Summary.

The valid Volume 2 Compliance categories are defined in Section 15.2

15.2 VALID VOLUME 2 COMPLIANCE CATEGORIES

Volume 2 Compliance Categories refer to the functionality of a physical entity which has one or more InfiniBand elements described in Volume 2. Table 171 summarizes all valid Volume 2 compliance categories.

Detailed compliance requirements are listed for each category in Section 15.3, "Compliance Requirements," on page 662.

Table 171 Valid Volume 2 Compliance Categories

Category	Product Type	Qualifier	Product Definition	Reference[a]
CU	C (Chassis)	U (Unmanaged)	Unmanaged IB Chassis	Section 15.3.1 on page 662
CA		A (Actively Managed)	Actively Managed Chassis	Section 15.3.2 on page 663
CP		P (Passively Managed	Passively Managed Chassis	Section 15.3.3 on page 665

Table 171 Valid Volume 2 Compliance Categories

Category	Product Type	Qualifier	Product Definition	Reference[a]
PS	P (Protocol aware module)	S (Standard)	Standard IB Protocol Aware Module	Section 15.3.4 on page 667
PW		W (standard Wide)	Wide IB Protocol Aware Module	Section 15.3.5 on page 671
PT		T (Tall)	Tall IB Protocol Aware Module	Section 15.3.6 on page 675
PTW		TW (Tall-Wide)	Tall-Wide Protocol Aware Module	Section 15.3.7 on page 679
RS	R (Repeater module)	S (Standard)	Standard IB Repeater Module	Section 15.3.8 on page 683
RW		W (standard, Wide)	Wide IB Repeater Module	Section 15.3.9 on page 686
RT		T (Tall)	Tall IB Repeater Module	Section 15.3.10 on page 689
RTW		TW (Tall, Wide)	Tall Wide Repeater Module	Section 15.3.11 on page 692
CC01	CC (Copper Cable)	01 (1x cable)	1x Copper cable	Section 15.3.12 on page 695
CC04		04 (4x cable)	4x Copper cable	Section 15.3.13 on page 696
CC08		08 (8x cable)	8x Copper cable	Section 15.3.14 on page 698
CC12		12 (12x cable)	12x Copper cable	Section 15.3.15 on page 699
CC12-4x		12x to 3-4x (cable)	12x to 3-4x Copper cable	Section 15.3.16 on page 700
CS01		01 (1x cable)	1x SFP Copper cable	Section 15.3.17 on page 701
FS01	FS (Fiber, Short haul)	01 (1x cable)	1x SX Fiber optic cable	Section 15.3.18 on page 702
FS04		04 (4x cable)	4x SX Fiber optic cable	Section 15.3.23 on page 704
FS08		08 (8x cable)	8x SX Fiber optic cable	Section 15.3.20 on page 703
FS12		12 (12x cable)	12x SX Fiber optic cable	Section 15.3.21 on page 703
FL01	FL (Fiber, Long haul)	01 (1x cable)	1x LX Fiber optic cable	Section 15.3.22 on page 703
FL04		04 (4x cable)	4x LX Fiber optic cable	Section 15.3.23 on page 704
FL08		08 (8x cable)	8x LX Fiber optic cable	See note a.
FL12		12 (12x cable)	12x LX Fiber optic cable	See note a.
PC01	PC (Port, Copper cable)	01 (1x cable port)	1x Copper cable port	Section 15.3.24 on page 704
PC04		04 (4x cable port)	4x Copper cable port	Section 15.3.25 on page 705
PC08		08 (8x cable port)	8x Copper cable port	See note a.
PC12		12 (12x cable port)	12x Copper cable port	Section 15.3.27 on page 709
PS01	PS (Port, Short haul fiber)	01 (1x cable port)	1x SX Fiber optic cable port	Section 15.3.28 on page 711
PS04		04 (4x cable port)	4x SX Fiber optic cable port	Section 15.3.29 on page 711
PS12		12 (12x cable port)	12x SX Fiber optic cable port	Section 15.3.30 on page 712

Table 171 Valid Volume 2 Compliance Categories

Category	Product Type	Qualifier	Product Definition	Reference[a]
PL01	PL (Port, Long haul fiber)	01 (1x cable port)	1x LX Fiber optic cable port	Section 15.3.31 on page 712
PL04		04 (4x cable port)	4x LX Fiber optic cable port	Section 15.3.32 on page 713
PL12		12 (12x cable port)	12x LX Fiber optic cable port	See note a.
PP01	PP (Port, Pluggable)	01 (1x)	1x Pluggable Port	Section 15.3.33 on page 714
PP04		04 (4x)	4x Pluggable Port: - R1.2	Section 15.3.34 on page 714
PP08		08 (8x)	8x Pluggable Port	See note a.
PP12		12 (12x)	12x Pluggable port	See note a.
AP04	AP (Active Port)	04 (4x)	4x Active Cable Port	Section 15.3.35 on page 714
AP12		12 (12x)	12x Active Cable Port	Section 15.3.36 on page 715
AC04	AC (Active Cable)	04 (4x)	4x Active Cable	Section 15.3.37 on page 715
AC12		12 (12x)	12x Active Cable	Section 15.3.38 on page 716
DC01	DC (Device Copper, Pluggable)	01 (1x)	1x Pluggable copper cable device	Section 15.3.39 on page 716
DC04		04 (4x)	4x Pluggable copper cable device - R1.2	Section 15.3.40 on page 717
DC08		08 (8x)	8x Pluggable copper cable device	See note a.
DC12		12 (12x)	12x Pluggable copper cable device	See note a.
DS01	DS (Device Short haul Fiber, Pluggable)	01 (1x)	1x Pluggable SX fiber optic device	Section 15.3.41 on page 718
DS04		04 (4x)	4x Pluggable SX fiber optic device	Section 15.3.42 on page 719
DS08		08 (8x)	8x Pluggable SX fiber optic device	See note a.
DS12		12 (12x)	12x Pluggable SX fiber optic device	See note a.
DL01	DL (Device Longhaul Fiber, Pluggable)	01 (1x)	1x Pluggable LX fiber optic device	Section 15.3.43 on page 721
DL04		04 (4x)	4x Pluggable LX fiber optic device	See note a.
DL12		12 (12x)	12x Pluggable LX fiber optic device	See note a.

Table 171 Valid Volume 2 Compliance Categories

Category	Product Type	Qualifier	Product Definition	Reference[a]
NMA_ID	NM (Not an IB defined Module or chassis slot form factor)	xCA (Functionality of an IB defined Adaptor; a TCA or HCA) _ID (Chassis GUID)	An IB TCA or HCA function whose physical implementation does not comply with the InfiniBand Volume 2 Specification and provides a Chassis GUID.	Section 15.3.44 on page 721
NMA_NID		xCA (Functionality of an IB defined Adaptor; a TCA or HCA) _NID (NoChassis GUID)	An IB TCA or HCA function whose physical implementation does not comply with the InfiniBand Volume 2 Specification and provides no Chassis GUID.	Section 15.3.45 on page 722
NMS_ID		S (Functionality of an IB defined Switch) _ID (Chassis GUID)	An IB Switch function whose physical implementation does not comply with the InfiniBand Volume 2 Specification and provides a Chassis GUID.	Section 15.3.46 on page 724
NMS_NID		S (Functionality of an IB defined Switch) _NID (No Chassis GUID)	An IB Switch function whose physical implementation does not comply with the InfiniBand Volume 2 Specification and provides no Chassis GUID.	Section 15.3.47 on page 725
NMR_ID		R (Functionality of an IB defined Router) _ID (Chassis GUID)	An IB Router function whose physical implementation does not comply with the InfiniBand Volume 2 Specification.	Section 15.3.48 on page 726
NMR_NID		R (Functionality of an IB defined Router) _NID (No Chassis GUID)	An IB Router function whose physical implementation does not comply with the InfiniBand Volume 2 Specification.	Section 15.3.49 on page 727
BM	BM	N/A	Baseboard Manager Software	Section 15.3.50 on page 728

a. These compliance categories are named but not defined in this level of documentation.

15.2.1 VOLUME 2 COMPLIANCE QUALIFIERS

Compliance Qualifiers indicate which compliance statements apply only if a product supports an optional feature or specified combination of optional features.

Some compliance statements may apply to multiple Compliance Categories, and thus appear in the Compliance Statement List under each applicable Category. Some of these “shared” compliance statements may include Qualifiers associated with functionality that is optional in some Categories and mandatory in others. In each Category where the functionality is mandatory, the associated Qualifier will be shown in ***bold italics*** for that Category’s “Valid Qualifier’s” entry in Table 171. (This release of the specification does not include any such qualifiers or valid qualifiers entries.)

15.2.1.1 CLAIMING SUPPORT FOR OPTIONAL FEATURES

C15-1: If a product claims to support a given optional feature, the product **must** comply with **all** compliance statements that apply to that optional feature.

For example, a Protocol Aware Module that claims to compliance with "Enhanced Signaling" must comply with all statements under the relevant Protocol Aware Statement List that apply, given the ES Qualifier.

A product **shall not** include in its list of supported *optional* features any features that are in fact *mandatory* for the Category the product claims compliance to. Qualifiers for these mandatory features will be shown in ***bold italics*** in Table 171, as necessary.

A product may claim support for multiple optional features, in which case the product must comply with all compliance statements that apply to the particular set of optional features claimed by the product, noting that some compliance statements apply only for specific combinations of qualifiers.

Table 172 lists and describes the Volume 2 Compliance Qualifiers that a product can claim compliance to. To abbreviate the optional support, one or more Qualifiers can be listed after the Category in braces. For example, a Standard IB Protocol Aware Module that supports Enhanced Signaling can be abbreviated with PS{ES}.

Table 172 Volume 2 Compliance Qualifiers

Qualifier	Description
ES	Enhanced Signaling (Release 1.2 extensions)

If an optional compliance statement does not contain a valid qualifier, refer to the text of the optional compliance statement to determine its applicability. The text of a compliance statement always takes precedence over the compliance qualifier.

15.2.1.2 COMPLIANCE STATEMENTS WITH MULTIPLE QUALIFIERS

Some compliance statements may contain combinations of Qualifiers, and apply only if the specified combination is true. For example, a compliance statement beginning with “RD and Atomics:” applies only if *both* RD and Atomics are supported. If a compliance statement begins with “RD or Atomics:”, the statement **shall** apply if *either* RD or Atomics is supported.

15.2.2 COMPLIANCE STATEMENT LISTS

Within each Compliance Category section is a list of the compliance statements that apply to that particular category. Here is a sample list entry:

15.2.2.1 HYPERTEXT LINKS

Online versions of this specification have hypertext links present before each of the lines in the Compliance Statement lists. These links are indicated by the "●" at the beginning of the line and will lead to the actual statement in the body of the specification that contains the details for each of the compliance entries.

Each Compliance Statement List entry also contains the page number for use with hard-copy versions of the specification.

15.2.2.2 COMPLIANCE STATEMENT LABELS

All formal compliance statements throughout the specification are labeled so they can be uniquely identified. Each label begins with either a "C" or an "o", indicating whether the compliance statement applies in all cases with respect to its category or whether the compliance statement is qualified with respect to optional features. The "o" is uncapitalized to make it more easily distinguishable from the "C" in Compliance Statement Lists.

The next portion of the label is the number of the chapter in which the formal compliance statement appears. The final portion of the label is a compliance statement number, which starts with "1" in each chapter. "C" and "o" compliance statements are numbered independently.

15.2.2.3 COMPLIANCE STATEMENT TITLES

Each line within a Compliance Statement List contains a brief title for the respective compliance statement. Because of the limited space and lack of context, each title is only intended to convey the topic of the compliance statement, and not necessarily convey its actual requirements.

Compliance statements that apply only to optional functionality is indicated by the presence of one or more Qualifiers at the beginning of the title, followed by a colon. For example, the above sample Compliance Statement Title contains the "ES" qualifier.

15.2.3 COMMON REQUIREMENTS

Some Compliance Categories share common requirements, such as those that apply to all ports. To avoid unnecessary duplication, certain common requirement sets have been collected and referenced by the appropriate Compliance Categories instead of replicating those lists of requirements under each separate Category.

15.3 COMPLIANCE REQUIREMENTS

15.3.1 CATEGORY "CU" - CHASSIS, UNMANAGED

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of “CU”, or “Chassis, Unmanaged”, a product **shall** meet all requirements specified below.

15.3.2 CATEGORY "CA" - CHASSIS, ACTIVELY MANAGED

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "CA", or "Chassis, Actively Managed", a product **shall** meet all requirements specified below.

15.3.3 CATEGORY "CP" - CHASSIS, PASSIVELY MANAGED

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "CP", or "Chassis, Passively Managed", a product **shall** meet all requirements specified below.

15.3.4 CATEGORY "PS" - PROTOCOL AWARE MODULE, STANDARD

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of “PS”, or “Protocol aware module, Standard”, a product **shall** meet all requirements specified below.

15.3.5 CATEGORY "PW" - PROTOCOL AWARE MODULE, WIDE

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PW", or "Protocol aware module, Wide", a product **shall** meet all requirements specified below.

15.3.6 CATEGORY "PT"- PROTOCOL AWARE MODULE, TALL

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PT", or "Protocol aware module, Tall", a product **shall** meet all requirements specified below.

15.3.7 CATEGORY "PTW" - PROTOCOL AWARE MODULE, TALL-WIDE

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PTW", or "Protocol aware module, Tall-Wide", a product **shall** meet all requirements specified below.

15.3.8 CATEGORY "RS" - REPEATER MODULE, STANDARD

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "RS", or "Repeater module, Standard", a product **shall** meet all requirements specified below.

15.3.9 CATEGORY "RW" - REPEATER MODULE, WIDE

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "RW", or "Repeater module, Wide", a product **shall** meet all requirements specified below.

15.3.10 CATEGORY "RT" - REPEATER MODULE, TALL

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "RT", or "Repeater module, Tall", a product **shall** meet all requirements specified below.

15.3.11 CATEGORY "RTW"- REPEATER MODULE, TALL-WIDE

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "RTW", or "Repeater module, Tall-Wide", a product **shall** meet all requirements specified below.

15.3.12 CATEGORY "CC01" - COPPER CABLE, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "CC01", or "Copper Cable, 1x", a product **shall** meet all requirements specified below.

15.3.13 CATEGORY "CC04" - COPPER CABLE, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "CC04", or "Copper Cable, 4x", a product **shall** meet all requirements specified below.

15.3.14 CATEGORY "CC08" - COPPER CABLE, 8X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "cc08", or "Fiber optic cable, Short haul, 8x", a product **shall** meet all requirements specified below.

15.3.15 CATEGORY "CC12" - COPPER CABLE, 12X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "CC12", or "Copper Cable, 12x", a product **shall** meet all requirements specified below.

15.3.16 CATEGORY "CC12-4X" - COPPER CABLE, 12X TO 3-4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "CC12-4x", or "Copper Cable, 12x to 3-4x", a product **shall** meet all requirements specified below.

15.3.17 CATEGORY "CS01" - COPPER SFP CABLE, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "CS01", or "Copper SFP Cable, 1x", a product **shall** meet all requirements specified below.

15.3.18 CATEGORY "FS01" - FIBER OPTIC CABLE, SHORT HAUL, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "FS01", or "Fiber optic cable, Short haul, 1x", a product **shall** meet all requirements specified below.

15.3.19 CATEGORY "FS04" - FIBER OPTIC CABLE, SHORT HAUL, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "FS04", or "Fiber optic cable, Short haul, 4x", a product **shall** meet all requirements specified below.

15.3.20 CATEGORY "FS08" - FIBER OPTIC CABLE, SHORT HAUL, 8X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "FS08", or "Fiber optic cable, Short haul, 8x", a product **shall** meet all requirements specified below.

15.3.21 CATEGORY "FS12" - FIBER OPTIC CABLE, SHORT HAUL, 12X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "FS12", or "Fiber optic cable, Short haul, 12x", a product **shall** meet all requirements specified below.

15.3.22 CATEGORY "FL01" - FIBER OPTIC CABLE, LONG HAUL, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "FL01", or "Fiber optic cable, Long haul, 1x", a product **shall** meet all requirements specified below.

15.3.23 CATEGORY "FL04" - FIBER OPTIC CABLE, LONG HAUL, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "FL04", or "Fiber optic cable, Long haul, 4x", a product **shall** meet all requirements specified below.

15.3.24 CATEGORY "PC01" - PORT, COPPER CABLE, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PC01", or "Port, Copper Cable, 1x", a product **shall** meet all requirements specified below.

15.3.25 CATEGORY "PC04" - PORT, COPPER CABLE, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PC04", or "Port, Copper Cable, 4x", a product **shall** meet all requirements specified below.

15.3.26 CATEGORY "PC08" - PORT, COPPER CABLE, 8X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PC08", or "Port, Copper Cable, 8x", a product **shall** meet all requirements specified below.

15.3.27 CATEGORY "PC12" - PORT, COPPER CABLE, 12X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PC12", or "Port, Copper Cable, 12x", a product **shall** meet all requirements specified below.

15.3.28 CATEGORY "PS01"- PORT, SHORT HAUL FIBER, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PS01", or "Port, Short haul fiber, 1x", a product **shall** meet all requirements specified below.

15.3.29 CATEGORY "PS04" - PORT, SHORT HAUL FIBER, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PS04", or "Port, Short haul fiber, 4x", a product **shall** meet all requirements specified below.

15.3.30 CATEGORY "PS12" - PORT, SHORT HAUL FIBER, 12X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PS12", or "Port, Short haul fiber, 12x", a product **shall** meet all requirements specified below.

15.3.31 CATEGORY "PL01" - PORT, LONG HAUL FIBER, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PL01", or "Port, Long haul fiber, 1x", a product **shall** meet all requirements specified below.

15.3.32 CATEGORY "PL04" - PORT, LONG HAUL FIBER, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PL04", or "Port, Long haul fiber, 4x", a product **shall** meet all requirements specified below.

15.3.33 CATEGORY "PP01" - PORT, PLUGGABLE, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PP01", or "Port, Pluggable, 1x", a product **shall** meet all requirements specified below.

15.3.34 CATEGORY "PP04" - PORT, PLUGGABLE, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "PP04", or "Port, Pluggable, 4x", a product **shall** meet all requirements specified below.

15.3.35 CATEGORY "AP04" - ACTIVE CABLE PORT, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "AP04", or "Active Cable Port, 4x", a product **shall** meet all requirements specified below.

- **C7-28.2.38:**Active port, load impedance . . . . . Page 282
- **C7-28.2.39:**Active port, power when Bulk Power not available . . . Page 282
- **C7-28.2.40:**Active port, bypassing at freq's over 100 MHz . . . . . Page 282
- **C15-1:** Optional features: compliance statement applicability . Page 660

15.3.36 CATEGORY "AP12" - ACTIVE CABLE PORT, 12X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "AP12", or "Active Cable Port, 12x", a product **shall** meet all requirements specified below.

- **C7-28.2.9:**12x active port, use by 8x active cable . . . . . Page 266
- **C7-28.2.11:**12x active port for 8x, no 12V & 3.3V simultaneous . . Page 267
- **C7-28.2.12:**12X active port, board connector . . . . . Page 269
- **C7-28.2.13:**12x active port, connector pin assignment . . . . . Page 269
- **C7-28.2.14:**12x active port, no 12V and 3.3V simultaneous . . . . . Page 269
- **C7-28.1.2:**12x active port for 8x, no 12 & 3.3V simultaneous . . . . Page 274
- **C7-28.1.3:**12x active port, no 12V and 3.3V simultaneous . . . . . Page 278
- **C7-28.2.27:**Active port, 12V voltage tolerance . . . . . Page 281
- **C7-28.2.28:**Active port, 3.3V voltage tolerance . . . . . Page 281
- **C7-28.2.29:**Minimum current supplied, 12V. . . . . Page 281
- **C7-28.2.31:**Minimum current supplied, 3.3V . . . . . Page 281
- **C7-28.2.34:**Active port, sense pin requirements . . . . . Page 281
- **C7-28.2.35:**Active port, power disable when sense grounded. . . . Page 282
- **C7-28.2.36:**Active port, hot plugging . . . . . Page 282
- **C7-28.2.37:**Active port, short circuit protection . . . . . Page 282
- **C7-28.2.38:**Active port, load impedance . . . . . Page 282
- **C7-28.2.39:**Active port, power when Bulk Power not available . . . Page 282
- **C7-28.2.40:**Active port, bypassing at freq's over 100 MHz . . . . . Page 282
- **C15-1:** Optional features: compliance statement applicability . Page 660

15.3.37 CATEGORY "AC04" - ACTIVE CABLE, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "AC04", or "Active Cable, 4x", a product **shall** meet all requirements specified below.

- **C7-28.2.8:**4x active cable, no 12V and 3.3V simultaneous . . . . . Page 265
- **C7-28.2.16:**4x active cable, metal backshell . . . . . Page 272
- **C7-28.2.17:**4x active cable, pin assignment . . . . . Page 272
- **C7-28.2.18:**4x active cable, Vcc not wired through . . . . . Page 272
- **C7-28.1.1:**4x active cable, no 12V and 3.3V simultaneous . . . . . Page 272
- **C7-28.2.27:**Active cable, 12V voltage tolerance . . . . . Page 281
- **C7-28.2.28:**Active cable, 3.3V voltage tolerance. . . . . Page 281
- **C7-28.2.30:**Maximum current draw at minimum voltage, 12V. . . . Page 281
- **C7-28.2.32:**Maximum current draw at minimum voltage, 3.3V . . . Page 281
- **C7-28.2.33:**Max current draw at max voltage, 3.3V & 12V . . . . . Page 281
- **C7-28.2.34:**Active cable, sense pin requirements . . . . . Page 281
- **C7-28.2.35:**Active cable, power disable when sense grounded . . Page 282
- **C7-28.2.36:**Active cable, hot plugging . . . . . Page 282
- **C7-28.2.37:**Active cable, short circuit protection . . . . . Page 282
- **C7-28.2.38:**Active cable, load impedance . . . . . Page 282
- **C7-28.2.39:**Active cable power when Bulk Power not available . . Page 282
- **C7-28.2.40:**Active cable, bypassing at freq's over 100 MHz . . . . . Page 282
- **C15-1:** Optional features: compliance statement applicability . Page 660

15.3.38 CATEGORY "AC12" - ACTIVE CABLE, 12X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "AC12", or "Active Cable, 12x", a product **shall** meet all requirements specified below.

15.3.39 CATEGORY "DC01" - DEVICE, COPPER SFP PLUGGABLE, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "DC01", or "Device, Copper SFP Pluggable, 1x", a product **shall** meet all requirements specified below.

15.3.40 CATEGORY "DC04" - DEVICE, COPPER SFP PLUGGABLE, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "DC04", or "Device, Copper SFP Pluggable, 4x", a product **shall** meet all requirements specified below.

15.3.41 CATEGORY "DS01" - DEVICE SHORT HAUL FIBER, PLUGGABLE, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "DS01", or "Device Short haul fiber, Pluggable, 1x", a product **shall** meet all requirements specified below.

- **C6-16.1.2:** Cable connection amplitude and eye ... - Obsolete . . . Page 203
- **C6-16.2.2:** Backplane connection amplitude and eye at SDR . . . . Page 203
- **C6-16.2.3:** Cable connection amplitude and eye at SDR . . . . . Page 203
- **o6-16.2.1:** ES: BER with compliant tx & channel, DDR . . . . . Page 203
- **o6-16.2.1:** ES: BER with compliant tx & channel at QDR . . . . . Page 204
- **C6-17:** Maximum skew specified across physical lanes . . . . . Page 206
- **C7-15.1.2:** 1x Pluggable devices, requirements . . . . . Page 232
- **C8-4:** BER requirements for all optical ports . . . . . Page 296
- **C8-5:** Optical Port, Tx Mask, Polarity, & Quie... - Obsolete . . Page 296
- **C8-6:** Optical Port, Signal grounding requirements . . . . . Page 296
- **C8-6.1.1:** Optical Port, Quiescent conditions &Optical Polarity . . Page 297
- **C8-6.2.1:** Optical Transmitter Mask, Nx-SX & 1x-LX . . . . . Page 297
- **C8-6.2.2:** Optical Transmitter Mask, Nx-SX & 1x-LX - at DDR . . . Page 297
- **C8-7:** Jitter requirements for all Optical Ports - Obsolete . . . . Page 302
- **C8-7.1.1:** Nx-SX & 1x-LX - Optical Jitter requi... - Obsolete . . . . Page 302
- **C8-7.1.1:** 1x-SX Optical Jitter Requirements - Obsolete . . . . . Page 302
- **C8-7.2.1:** Nx-SX & 1x-LX Optical Jitter requirements at SDR . . . Page 302
- **C8-7.2.1:** 1x-SX Optical Jitter Requirements at SDR . . . . . Page 302
- **C8-7.2.2:** Nx-SX & 1x-LX Optical Jitter requirements at DDR . . . Page 302
- **C8-7.2.2:** 1x-SX Optical Jitter Requirements at DDR . . . . . Page 302
- **C8-7.2.3:** Optical Port 1x Jitter at QDR . . . . . Page 306
- **C8-8:** Optical ports, skew while operating at SDR . . . . . Page 307
- **C8-8.2.1:** Optical ports, skew while operating at DDR . . . . . Page 307
- **C8-8.2.2:** Optical ports, skew while operating at QDR . . . . . Page 307
- **C8-9:** 1x-SX and 1x-LX optical port, eye safety requirements Page 309
- **C8-10.2.1:** 1x-SX Optical Transmitter and Receiver requirements. Page 309
- **C8-11.2.2:** 1x-SX and 1x-LX optical eye safety at DDR . . . . . Page 315
- **C8-11.2.3:** 1x-SX Optical Transmitter and Receiver at DDR . . . . . Page 315
- **C8-13.2.3:** 1x-SX Optical Eye Safety, Tx & Rx req's at QDR . . . . . Page 331
- **C8-13.2.5:** 1X Optical port, eye safety requirements at QDR . . . . . Page 332
- **C8-14:** Optical Receptacle requirements . . . . . Page 333
- **C8-16:** Optical Port receptacle and fiber orientation . . . . . Page 334
- **C8-28:** Optical Signal Conditioner requirements. . . . . Page 346
- **C8-29:** Optical Port Aux power behavior. . . . . Page 354
- **C8-29.1.1:** 1x Optical Pluggable Devices . . . . . Page 354
- **o9-32.1.3:** InfiniBand Icon Visibility . . . . . Page 408
- **C10-9.1.1:** backplane wiring impedance. . . . . Page 424
- **C15-1:** Optional features: compliance statement applicability . Page 660
- **C15-1:** Optional features: compliance statement applicability . Page 660
- **C15-1:** Optional features: compliance statement applicability . Page 660

15.3.42 CATEGORY "DS04" - DEVICE SHORT HAUL FIBER, PLUGGABLE, 4X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "DS04", or "Device Short haul fiber, Pluggable, 4x", a product **shall** meet all requirements specified below.

- **C6-1:** General coverage statement for ... - Obsolete . . . . . Page 177
- **C6-1.2.1:** General coverage statement for signaling, SDR. . . . . Page 177
- **o6-1.2.1:** ES: DDR operation is optional . . . . . Page 177
- **o6-1.2.1:** ES: QDR operation is optional . . . . . Page 177
- **o6-1.2.1:** ES: Non-contiguous LinkSpeed support . . . . . Page 177
- **o6-1.2.2:** ES: DDR speed operation. . . . . Page 177
- **o6-1.2.3:** ES: QDR speed operation. . . . . Page 177
- **C6-2:** ESD Requirements - Obsolete . . . . . Page 178
- **C6-2.1.1:** ESD Requirements, recoverability - Obsolete. . . . . Page 178
- **C6-2.2.1:** ESD Requirements . . . . . Page 178
- **C6-2.2.2:** ESD Requirements, recoverability . . . . . Page 178
- **C6-3:** Unexpected hot removal or insertion allowed . . . . . Page 178
- **C6-4:** Specifications apply within defined env. - Obsolete . . . Page 178

15.3.43 CATEGORY "DL01" - DEVICE LONG HAUL FIBER, PLUGGABLE, 1X

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "DL01", or "Device Long haul fiber, Pluggable, 1x", a product **shall** meet all requirements specified below.

- **C6-1.1.1:** Port Type 2 - Obsolete ..... Page 177
- **C6-1.2.2:** Port Type 2 ..... Page 177
- **C7-15.1.2:** 1x Pluggable devices, requirements ..... Page 232
- **C7-15.2.8:** 4X Optical Pluggable Devices, general ..... Page 243
- **C8-4:** BER requirements for all optical ports ..... Page 296
- **C8-5:** Optical Port, Tx Mask, Polarity, & Quie... - Obsolete ..... Page 296
- **C8-6:** Optical Port, Signal grounding requirements ..... Page 296
- **C8-6.1.1:** Optical Port, Quiescent conditions &Optical Polarity ..... Page 297
- **C8-6.2.1:** Optical Transmitter Mask, Nx-SX & 1x-LX ..... Page 297
- **C8-6.2.2:** Optical Transmitter Mask, Nx-SX & 1x-LX - at DDR ..... Page 297
- **C8-7:** Jitter requirements for all Optical Ports - Obsolete ..... Page 302
- **C8-7.1.1:** Nx-SX & 1x-LX - Optical Jitter requi... - Obsolete ..... Page 302
- **C8-7.1.1:** 1x-LX Optical Jitter Requirements - Obsolete ..... Page 302
- **C8-7.2.1:** Nx-SX & 1x-LX Optical Jitter requirements at SDR ..... Page 302
- **C8-7.2.1:** 1x-LX Optical Jitter Requirements at SDR ..... Page 302
- **C8-7.2.2:** Nx-SX & 1x-LX Optical Jitter requirements at DDR ..... Page 302
- **C8-7.2.2:** 1x-LX Optical Jitter Requirements at DDR ..... Page 302
- **C8-7.2.3:** Optical Port 1x Jitter at QDR ..... Page 306
- **C8-8:** Optical ports, skew while operating at SDR ..... Page 307
- **C8-8.2.1:** Optical ports, skew while operating at DDR ..... Page 307
- **C8-8.2.2:** Optical ports, skew while operating at QDR ..... Page 307
- **C8-9:** 1x-SX and 1x-LX optical port, eye safety requirements Page 309
- **C8-11:** 1x-LX Optical Transmitter and Receiver - Obsolete ..... Page 311
- **C8-11.2.1:** 1x-LX Optical Transmitter and Receiver Req'ments ..... Page 311
- **C8-11.2.2:** 1x-SX and 1x-LX optical eye safety at DDR ..... Page 315
- **C8-11.2.4:** 1x-LX Optical Transmitter and Receiver at DDR ..... Page 317
- **C8-13.2.4:** 1x-LX Optical Eye Safety, Tx & Rx req's at QDR ..... Page 331
- **C8-13.2.5:** 1X Optical port, eye safety requirements at QDR ..... Page 332
- **C8-14:** Optical Receptacle requirements ..... Page 333
- **C8-16:** Optical Port receptacle and fiber orientation ..... Page 334
- **C8-28:** Optical Signal Conditioner requirements ..... Page 346
- **C8-29:** Optical Port Aux power behavior ..... Page 354
- **C8-29.1.1:** 1x Optical Pluggable Devices ..... Page 354
- **o9-32.1.3:** InfiniBand Icon Visibility ..... Page 408
- **C15-1:** Optional features: compliance statement applicability ..... Page 660
- **C15-1:** Optional features: compliance statement applicability ..... Page 660

15.3.44 CATEGORY "NMA_ID" - NON-MODULE ADAPTER, ID

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "NMA_ID", or "Non-Module Adapter with Chassis GUID", a product **shall** meet all requirements specified below.

- **C5-1:** Symbol Encoding (8B/10B) ..... Page 80
- **C5-2:** Control Symbols and Ordered-sets - Obsolete ..... Page 95
- **C5-2.2.1:** Control Symbols & Ordered-sets ..... Page 95
- **o5-2.2.1:** ES: Control Symbols & Ordered-sets, enhanced ..... Page 95
- **C5-3:** Management Datagram Interface - Obsolete ..... Page 104
- **C5-3.2.1:** Management Datagram Interface ..... Page 104
- **o5-3.2.1:** ES: Management Datagram Interface, enhanced ..... Page 105
- **C5-4:** Packet Ordering ..... Page 111
- **C5-5:** Packet Formats ..... Page 113
- **C5-6:** 1x Packet Format ..... Page 113

15.3.45 CATEGORY "NMA_NID" - NON-MODULE ADAPTER, NO ID

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "NMA_NID", or "Non-Module Adapter without Chassis GUID", a product **shall** meet all requirements specified below.

15.3.46 CATEGORY "NMS_ID" - NON-MODULE SWITCH, ID

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "NMS_ID", or "Non-Module Switch with Chassis GUID", a product **shall** meet all requirements specified below.

- **o13-22:** xInfo requirements for Power Management . . . . . Page 602
- **C13-64:** xInfo Headers . . . . . Page 602
- **C13-65:** xInfo Format . . . . . Page 603
- **o13-23:** FRUInfo Record . . . . . Page 616
- **C15-1:** Optional features: compliance statement applicability . Page 660
- **C15-1:** Optional features: compliance statement applicability . Page 660

15.3.47 CATEGORY "NMS_NID" - NON-MODULE SWITCH, NO ID

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "NMS_NID", or "Non-Module Switch with Chassis GUID", a product **shall** meet all requirements specified below.

- **C5-1:** Symbol Encoding (8B/10B) . . . . . Page 80
- **C5-2:** Control Symbols and Ordered-sets - Obsolete . . . . . Page 95
- **C5-2.2.1:** Control Symbols & Ordered-sets . . . . . Page 95
- **o5-2.2.1:** ES: Control Symbols & Ordered-sets, enhanced . . . . . Page 95
- **C5-3:** Management Datagram Interface - Obsolete . . . . . Page 104
- **C5-3.2.1:** Management Datagram Interface . . . . . Page 104
- **o5-3.2.1:** ES: Management Datagram Interface, enhanced . . . . . Page 105
- **C5-4:** Packet Ordering . . . . . Page 111
- **C5-5:** Packet Formats . . . . . Page 113
- **C5-6:** 1x Packet Format . . . . . Page 113
- **o5-6.2.1:** ES: 1x Packet Format, 8x ports . . . . . Page 113
- **C5-7:** 4x Packet Format . . . . . Page 114
- **C5-7:** 4x Packet Format, 12x ports . . . . . Page 114
- **o5-7.2.1:** ES: 4x Packet Format, 8x ports . . . . . Page 115
- **o5-7.2.1:** ES: 8x Packet Format . . . . . Page 116
- **o5-7.2.1:** ES: 8x Packet Format, 12x ports . . . . . Page 116
- **C5-8:** 12x Packet Format . . . . . Page 117
- **C5-9:** Link Initialization and Training - Obsolete . . . . . Page 118
- **C5-9.2.1:** Link Initialization and Training . . . . . Page 119
- **o5-9.2.1:** ES: Link Initialization and Training, enhanced . . . . . Page 119
- **o5-1:** Serial Data Inversion - Obsolete . . . . . Page 119
- **o5-9.2.1:** Serial Data Inversion . . . . . Page 119
- **o5-2:** Lane Reversal - 4x - Obsolete . . . . . Page 119
- **o5-2:** Lane Reversal - 12x - Obsolete . . . . . Page 119
- **o5-9.2.1:** Lane Reversal - 4x . . . . . Page 119
- **o5-9.2.1:** Lane Reversal - 8x . . . . . Page 119
- **o5-9.2.1:** Lane Reversal - 12x . . . . . Page 119
- **C5-10:** State Machine Delays . . . . . Page 147
- **C5-11:** Transmitter Interface and Behavior . . . . . Page 147
- **o5-3:** Transmitter Lane Reversal - 4x . . . . . Page 148
- **o5-3:** Transmitter Lane Reversal - 8x . . . . . Page 148
- **o5-3:** Transmitter Lane Reversal - 12x . . . . . Page 148
- **C5-12:** Receiver Interface and Behavior . . . . . Page 151
- **o5-4:** Receiver Correction of Inverted Serial Data . . . . . Page 151
- **o5-5:** Receiver Lane Reversal - 4x . . . . . Page 151
- **o5-5:** Receiver Lane Reversal - 8x . . . . . Page 151
- **o5-5:** Receiver Lane Reversal - 12x . . . . . Page 151
- **C5-13:** Link Physical Error Handling . . . . . Page 156
- **o5-6:** Internal Serial Loopback . . . . . Page 159
- **C5-14:** Clock Tolerance Compensation . . . . . Page 160
- **C5-15.2.1:** ES: Link Heartbeat . . . . . Page 167
- **C5-15.2.2:** ES: Physical Layer Compliance Test (Phy Test) . . . . . Page 169
- **o6-17.2.1:** ES: Physical Test and Characterization Facilities . . . . . Page 211
- **o9-32.1.1:** InfiniBand port icon . . . . . Page 408
- **o9-32.1.2:** InfiniBand icon color . . . . . Page 408
- **C13-5:** Non-Module Switch HM Features . . . . . Page 494
- **o13-4:** InfiniBand LEDs . . . . . Page 508
- **C13-46:** Baseboard MAD commands . . . . . Page 580

15.3.48 CATEGORY "NMR_ID" - NON-MODULE ROUTER, ID

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "NMR_ID", or "Non-Module Router with Chassis GUID", a product **shall** meet all requirements specified below.

15.3.49 CATEGORY "NMR_NID" - NON-MODULE ROUTER, NID

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "NMR_NID", or "Non-Module Router without Chassis GUID", a product **shall** meet all requirements specified below.

15.3.50 CATEGORY "BM" - BASEBOARD MANAGER

In order to claim compliance to the InfiniBand Volume 2 specification at a compliance category of "BM", or "Baseboard Manager", a product **shall** meet all requirements specified below.

IBTA

ANNEX A1: MECHANICAL ANNEX

A1.1 MODULE DESIGN EXAMPLES

The design of the module cover is left extremely flexible. Figure 95 on page 364 describes required cover dimensions to ensure proper clearances with other modules and with chassis slot guides. Material and internal cover features are entirely flexible. It is recommended, however, that if a metal cover is used, careful attention be paid to the ground contact between the module and cover in order to avoid potential EMI problems. Figure 183 shows an example of a cover that might be used on an IB Module.

IBTA

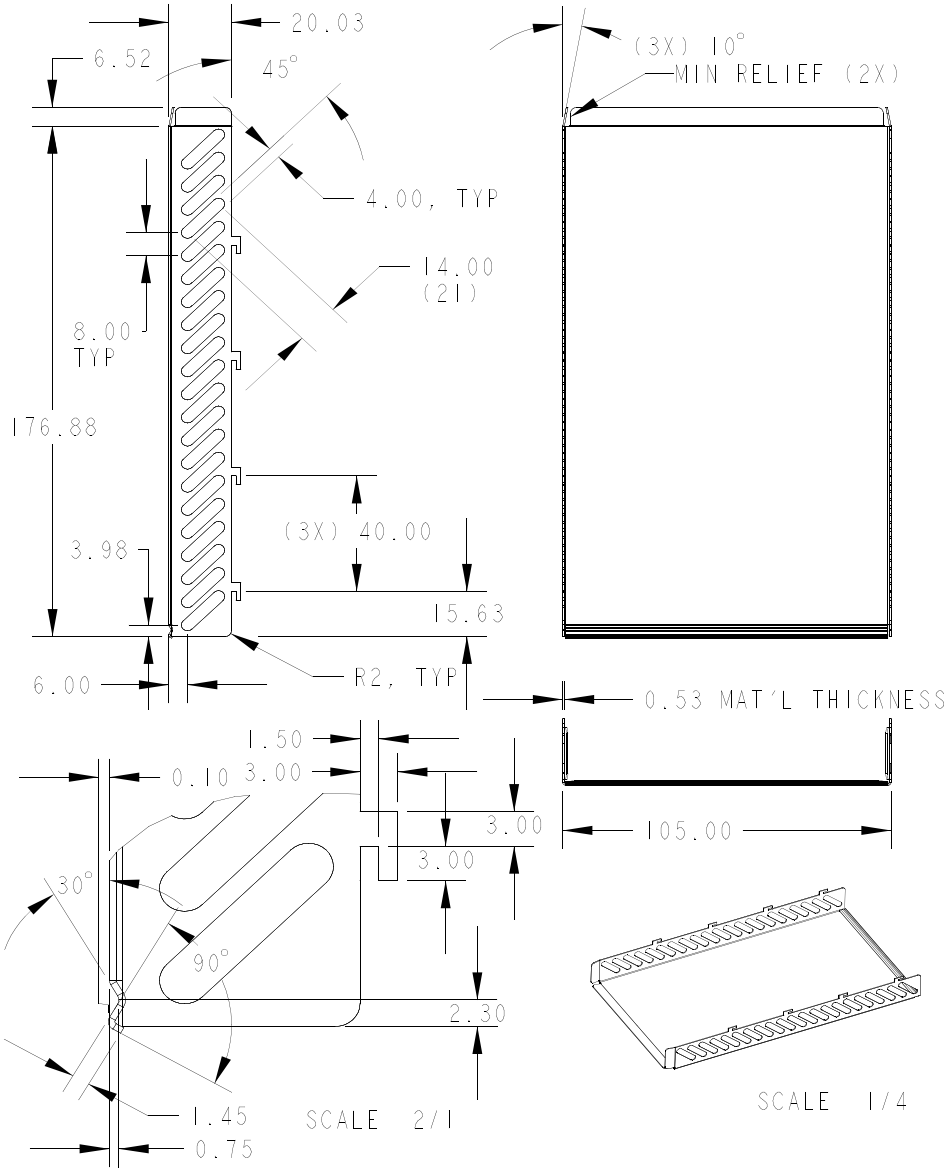


Figure 183 IB Cover Example

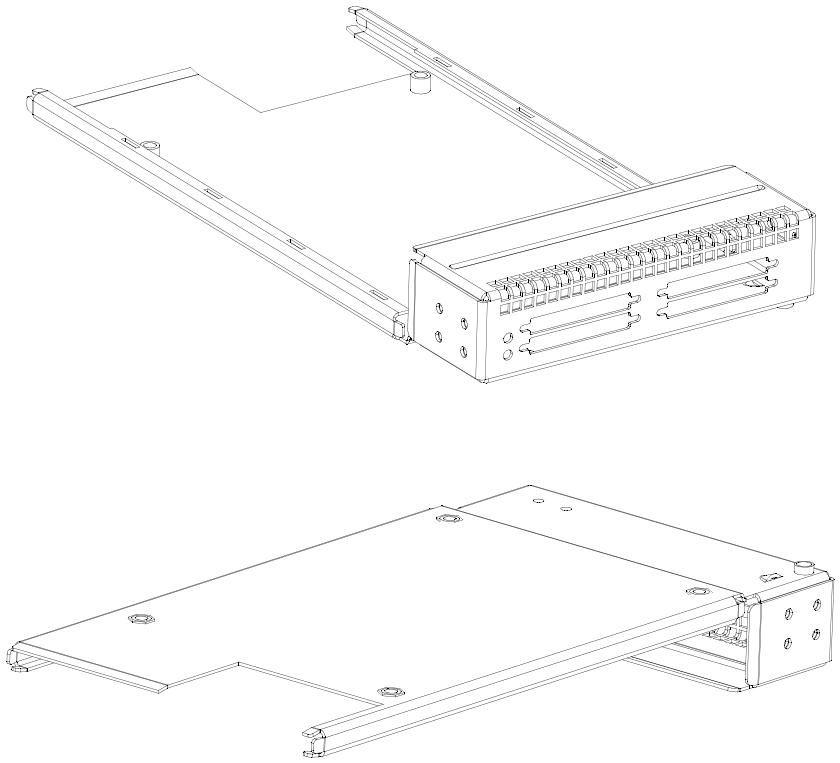


Figure 184 Example IB Module Carrier

The module carrier is also intentionally left somewhat flexible in design with mainly interface dimensions and material compatibility specified. The drawings provided in this annex are of an example design that has been detailed and prototyped and should provide an excellent starting point.

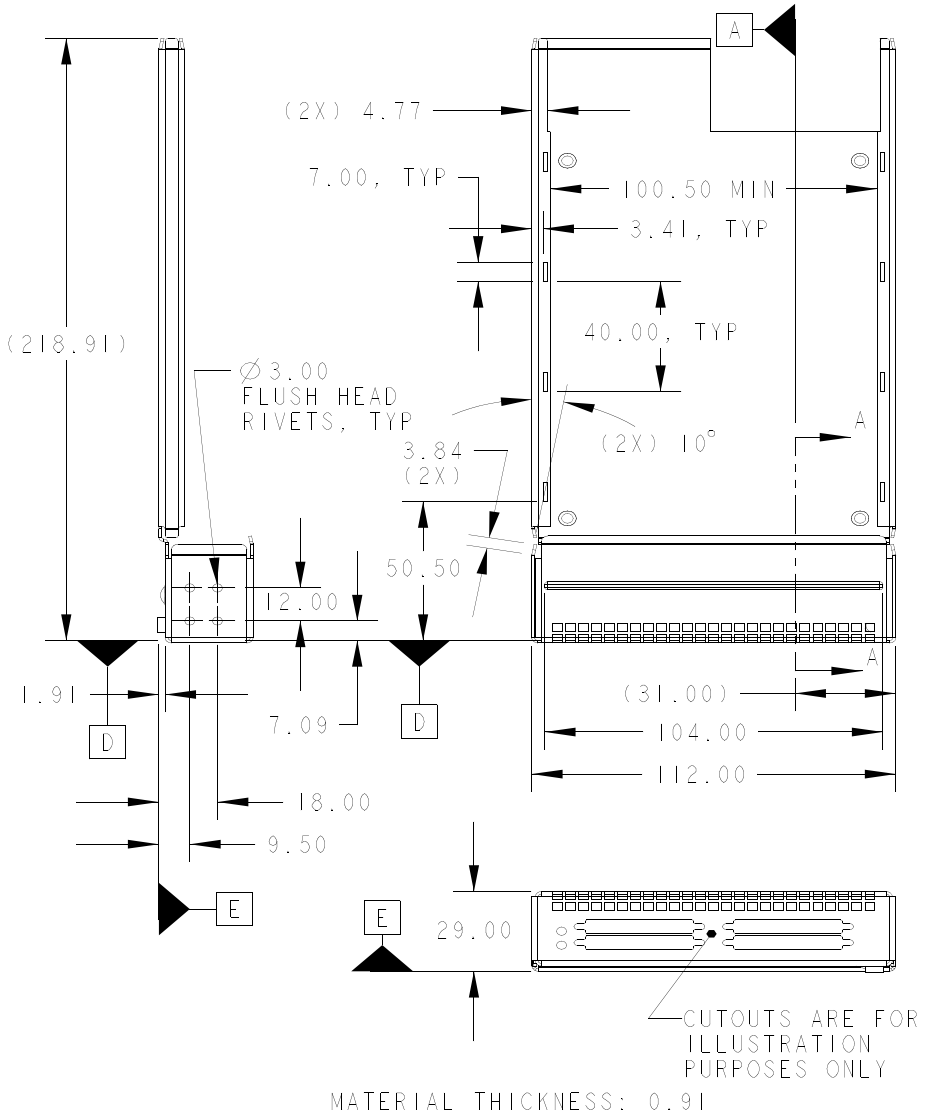


Figure 185 IB Carrier Example, sh.1

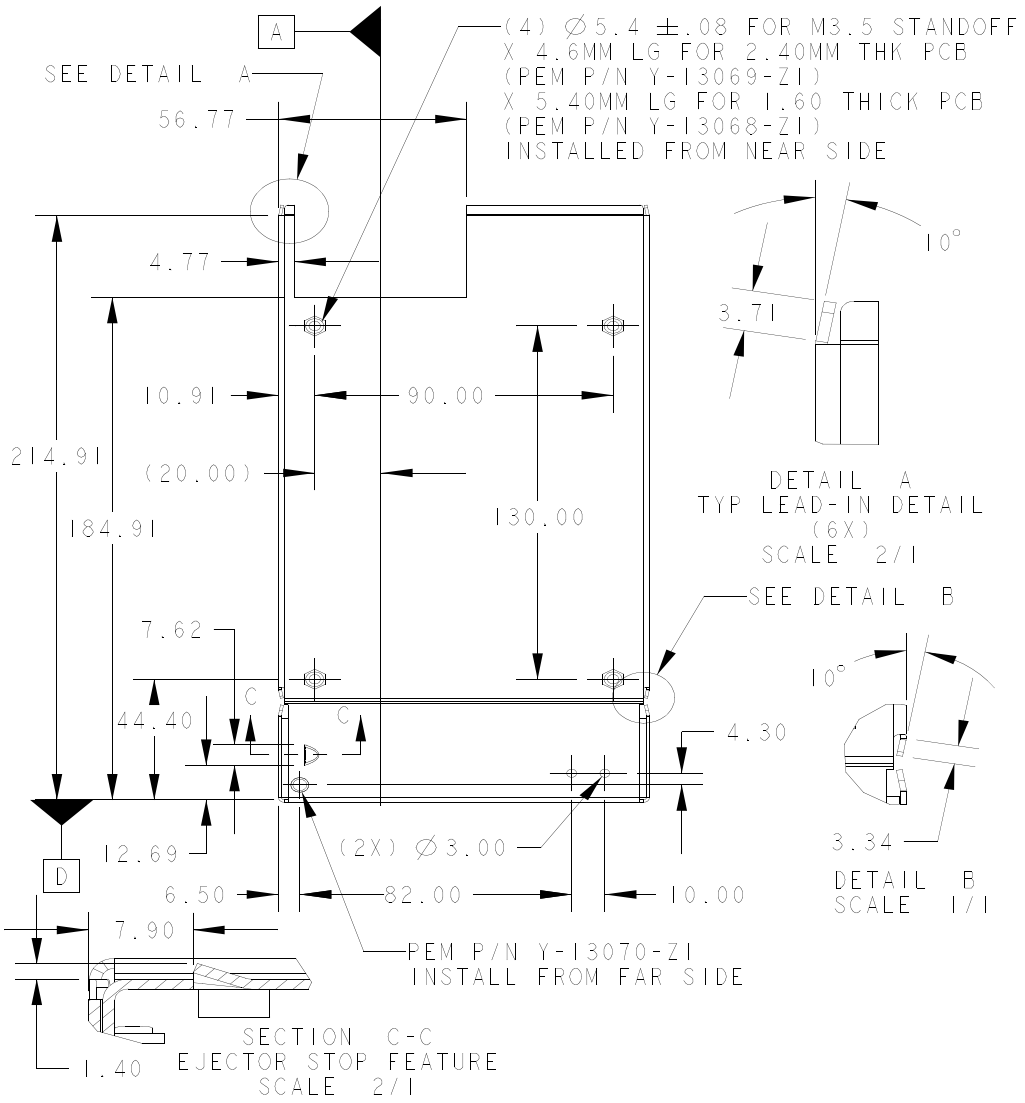


Figure 186 IB Carrier Example, sh.2

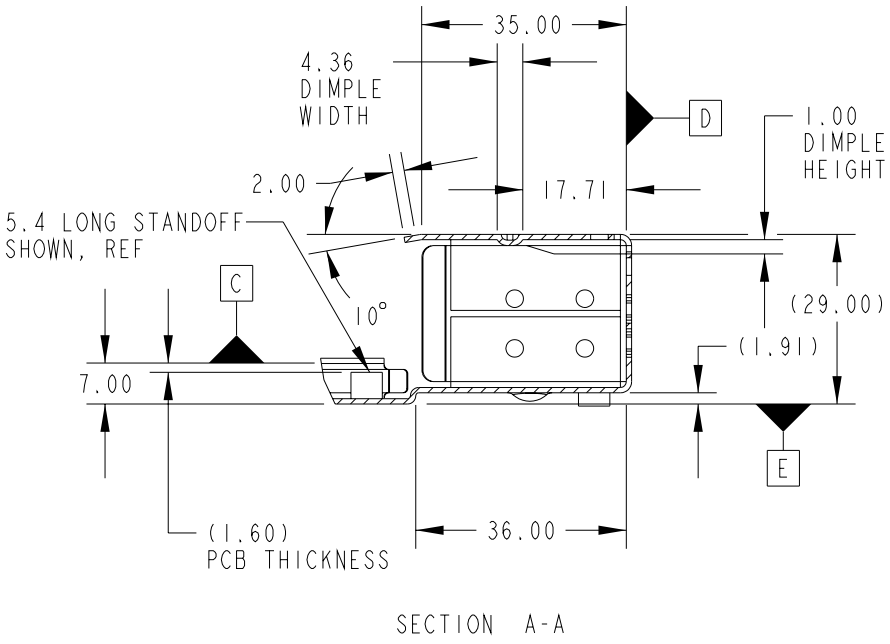


Figure 187 IB Carrier Example, sh.3

The example in Figure 187 implies a board thickness specific to this example only. Refer to Section 9.3.3, “Board Thickness,” on page 395 for complete information on board thickness.

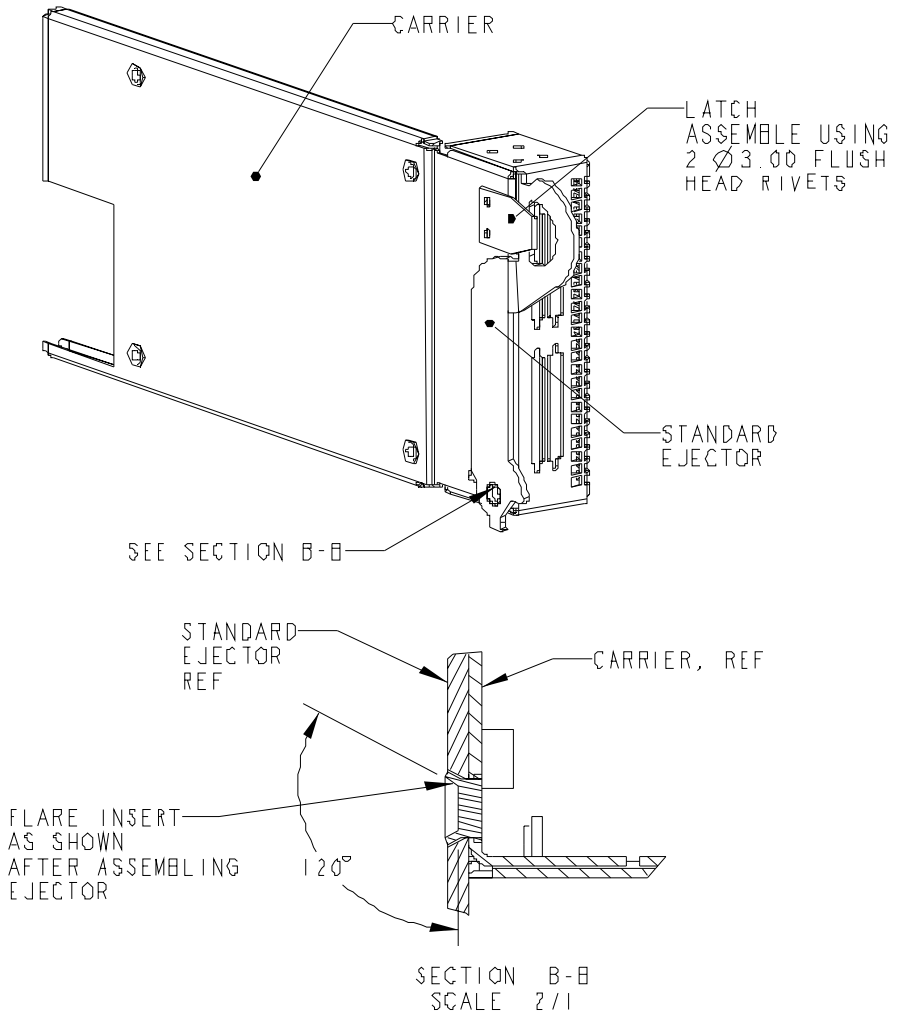


Figure 188 Example IB Carrier Assembly Details

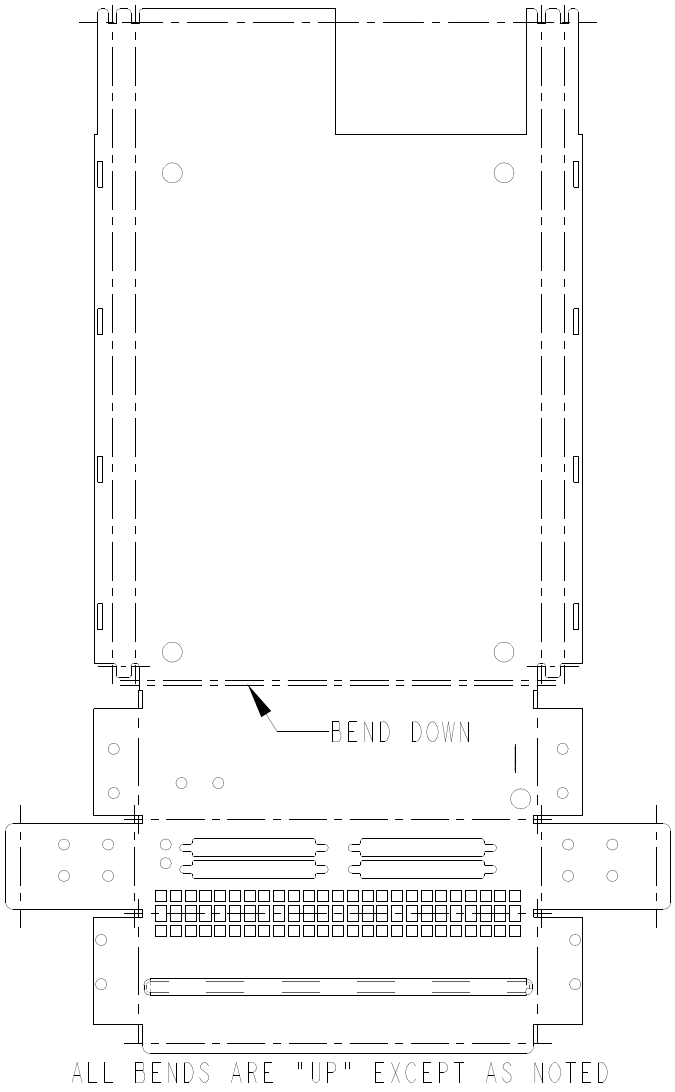


Figure 189 Example IB Carrier Flat Pattern

A1.2 COMPUTATIONAL FLUID DYNAMICS (CFD) RESULTS

The InfiniBand Module has been sized specifically to operate within component temperature limits in what is considered a severe system implementation. CFD models and empirical testing have been employed to determine module and system resistance's and flow rates. The models have also been used to study the effect of variables such as card height, slot spacing, vent size, etc. on module flow velocities and component tem-

peratures. These models may be a useful starting point for the engineer when designing a system to insure that component temperature goals are met, therefore a reference design has been included in this annex.

The reference design was constructed using Flotherm V2.2 and consists of an enclosure 560 mm in length, 133 mm tall, and 217 mm wide. This model represents the implementation of seven InfiniBand Modules in a system enclosure. The airflow is modeled back-to-front (airflow enters module near InfiniBand connector and exits at front vent), but the airflow is allowed to be in either direction (front-to-back or back-to-front) across the module. Directly behind the IB Module section is a fan/blower which splits the enclosure into a low-pressure (rear) compartment and a high-pressure (front) compartment. Typical components occupying the front compartment of a system would be hard drives, tape drives, power supplies, etc. These components may significantly affect the system by adding flow resistance and heat to the incoming air. Figure 190, shows an isometric view of the Flotherm model.

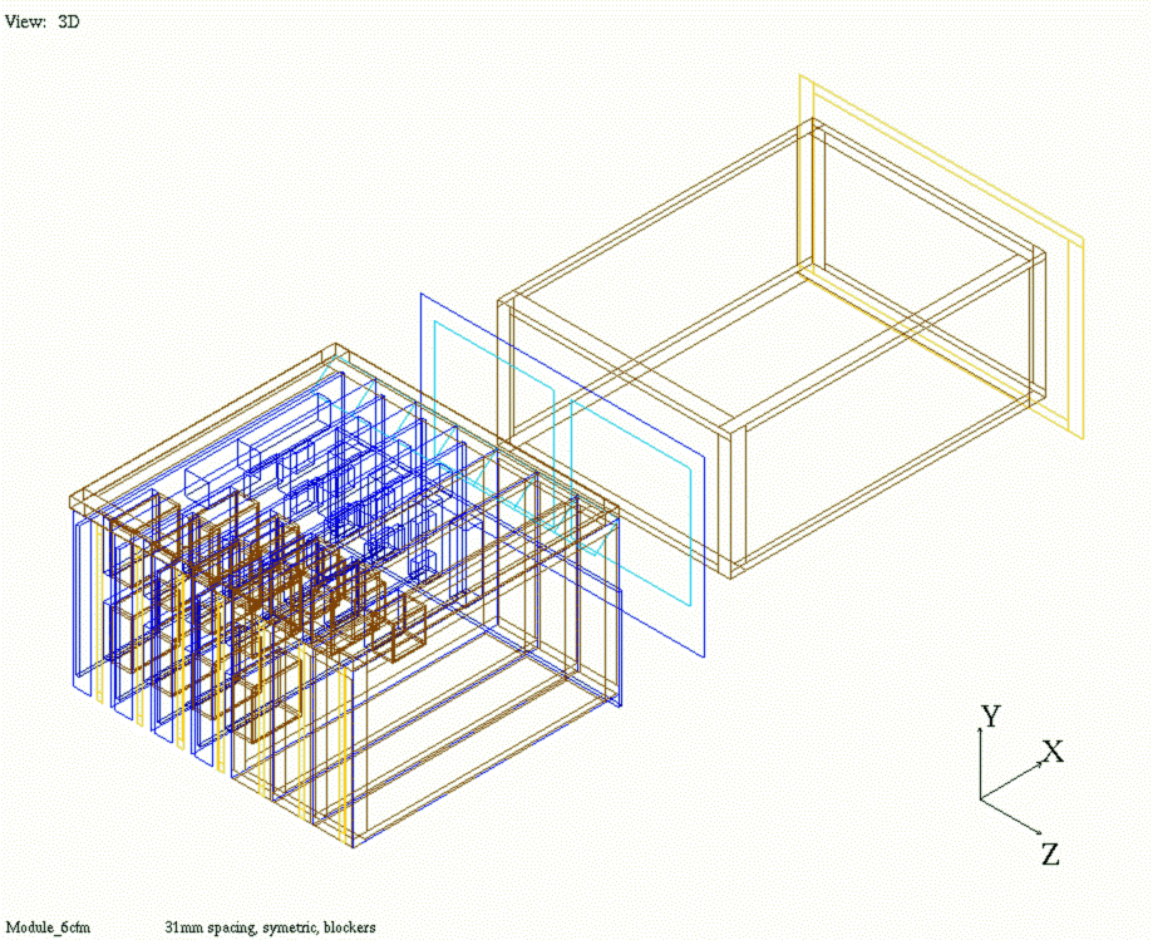


Figure 190 Isometric View of Flowtherm Model

This reference design assumes the majority of the front compartment to be occupied by a power supply with some area around its perimeter left to bleed ambient air in and around the supply. The power supply resistance was modeled using a Flotherm linear source term with a loss coefficient defined for each direction of the Cartesian axis system. In addition, 10 C of preheat was assumed across the power supply. The fans modeled were high performance 120 mm fans in series, with two pairs of fans per system. Four of the InfiniBand Modules were explicitly modeled with each component on the circuit board represented by a cuboid or volume source. The remaining modules were modeled with one volume source term each to reduce the size of the model. On the modules that were modeled explicitly, cuboids were used to model components with 2 watts or less heat dissipation, with the heat being dissipated conducted to the board and air. For components with more then 2 watts of heat dissipation, volume source terms were used to model the components with all the heat convected to the air (i.e. components were not coupled to the board). The total model size was about 250,000 nodes.

The goal of the engineer is to insure that component temperature goals are met. To this end, it is necessary to layout components on the board to insure that they receive adequate airflow. In addition, it may be necessary to install heat sinks on high heat density components. It also may be necessary to install a flow defector on the ESD cover to redirect airflow. A flow deflector will redirect airflow over board components, while reducing the shadowing effect caused by the backplane. Figure 191, shows velocity vectors generated by a CFD model across the InfiniBand Module. In this CFD model a flow deflector was used to redirect airflow towards the center of the board. From the figure it can be seen that the airflow is fairly uniform with only minimal recirculation near the backplane. This figure may be used as a reference for insuring adequate velocities across the module.

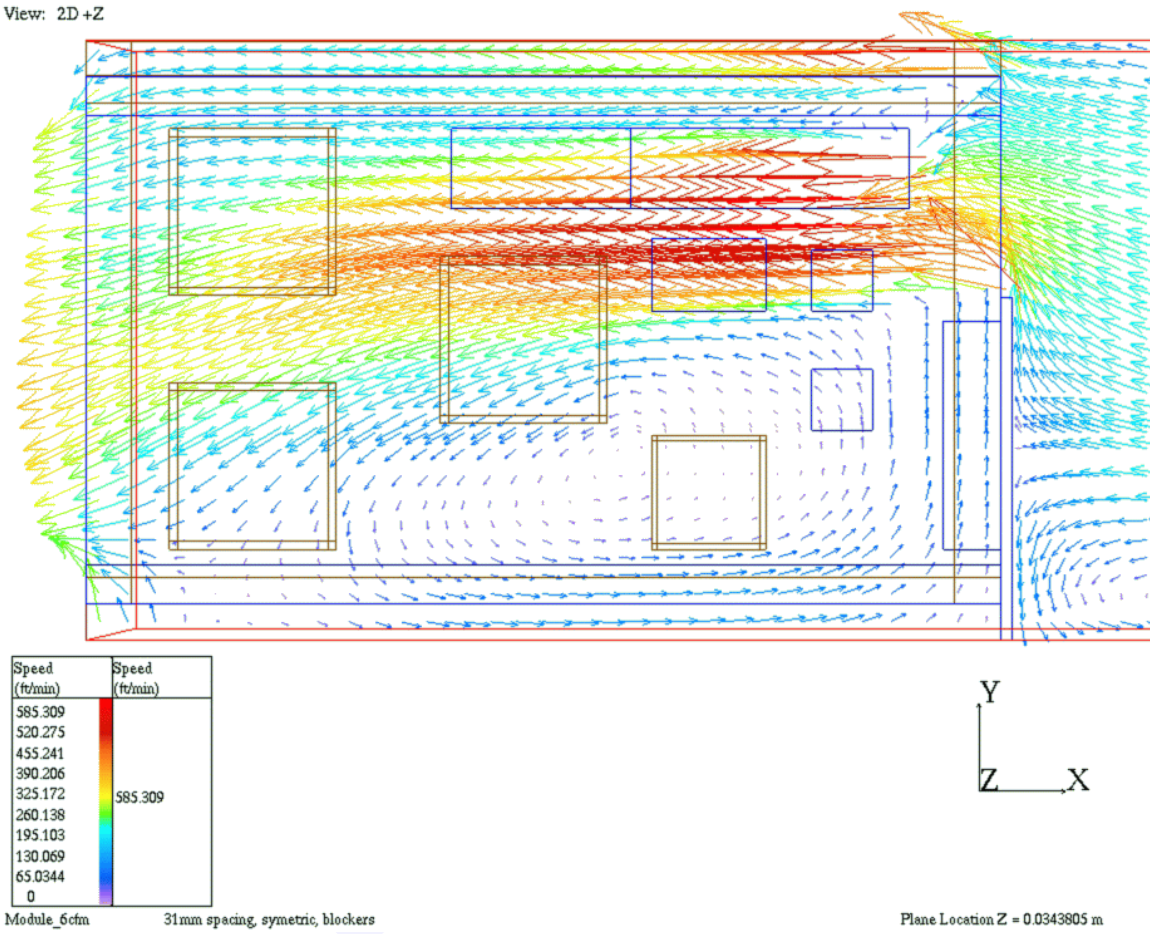


Figure 191 Velocity Vector Plot Through Module

A1.3 MODULE PRESSURE DROP TESTING

A1.3.1 GENERAL DESCRIPTION OF A MODULE PRESSURE DROP TEST FIXTURE

The module must have the airflow/pressure drop characteristics as shown in Figure 88. The module pressure drop is defined as the difference in static pressure from point A to point B of Figure 192 on page 741. Although systems may move air in either direction through the module, one direction has been chosen to characterize a module's pressure drop.

An airflow test fixture is required to test that the airflow and pressure of the module fall within the required limits. In general, such a fixture should deliver airflow at the backplane (point A) that is mostly unidirectional towards the module. This can be achieved by constructing a test fixture with a four-

walled chamber just upstream of the module that is sufficiently long (approximately 250mm minimum) to allow for the airflow to fully develop. Flow straighteners are recommended that can assist in straightening the airflow. The fan that supplies the air should be sufficiently upstream of the flow straighteners (approximately 250mm minimum) in order to reduce fan-induced swirling effects. In addition, the width of the four-walled chamber should be large enough (at least 200mm overall wider than the air inlet to the module in both axes) to minimize wall-induced turbulence effects. The fixture must measure the difference in air pressure at point A with respect to ambient air (i.e. outside the fixture) as a function of volumetric airflow.

A1.3.2 SPECIFIC EXAMPLE OF A MODULE PRESSURE DROP TEST FIXTURE

There are several instruments available to measure pressure drop and volumetric flow rate, such as solid state pressure transducers and manometers (to measure pressure), hot wire anamometers and thermistor sensors (to measure velocity, so multiples are required across a known cross-sectional area to determine volumetric airflow rates) and manometer-across-a-calibrated-nozzle (for measuring airflow rate). Figure 192 on page 741 is of a schematic of a pressure drop test fixture based upon the lowest-cost available technologies. The overall size of the test fixture is approximately 1000mm long by 240mm wide by 240mm tall and construction was based upon standard 6mm (¼ inch) plywood. The test fixture includes four main components: the airmover (fan), the airflow measuring device (in this case, a nozzle and manometer A), the static pressure measurement device (manometer B), and the module holder. A more detailed description follows.

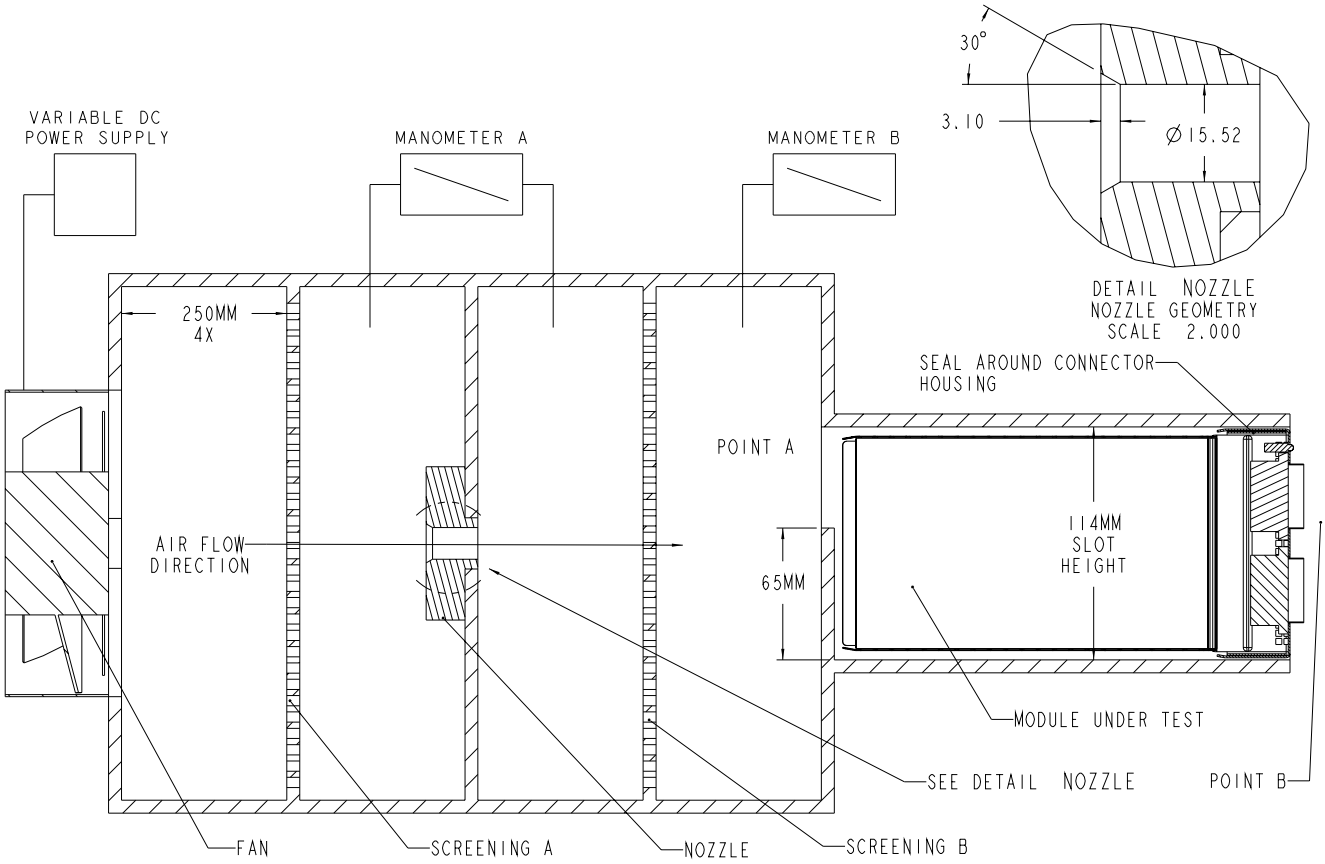


Figure 192 Schematic of pressure drop test fixture

The fan has sufficient capacity to supply enough volumetric airflow and static pressure to overcome the resistance in the fixture while supplying the required airflow and pressure to the module under test. A DC fan was selected so that the it's speed can be varied in order to test various air-flow-pressure drop points of the module. In this model, a standard 6-inch diameter by 2-inch deep fan was selected. This type of fan is readily available from most fan vendors. In the model, the fan was powered by a simple 12V (2A) power supply.

Two layers of standard window screen mesh (Screening A) with 1.60mm (.062 inch) square openings that are spaced 25 mm (1.00 in) apart are located approximately 250mm downstream of the fan in order to approximately straighten the airflow before the air reaches the nozzle. The volumetric airflow is measured by means of measuring the pressure drop across a calibrated nozzle using manometer A. The manometer that was selected is a simple, low-cost fluid/mechanical manometer (Dwyer model number 40-1) from Dwyer Instruments (Michigan City, Indiana, USA) that

is connected via 5.0mm (0.20 in) diameter plastic flexible tubing to barb fittings on the test fixture that are both upstream and downstream of the nozzle. By using the related calibration curve for the nozzle, volumetric airflow can be deduced.

Using a calibrated nozzle is critical to yielding accurate results for volumetric airflow. The nozzle for this model was designed and sized for a standard module, using the "minor headloss" equation found in any fluid mechanics textbook. It was machined from 25.4 mm (1.00 inch) thick aluminum and consisted of a 15.52 mm (0.611 inch) diameter hole and a 30 degree (per side) by 3.10 mm (0.122 inch) deep chamfered edge that is oriented on the upstream side of the nozzle (see detail in Figure 192 on page 741). The overall outer dimension of the nozzle is not critical. This particular nozzle was calibrated in an airflow chamber/wind tunnel (see Figure 193 on page 742). The volumetric airflow can be determined by measuring the pressure across the nozzle, then using Figure 193 on page 742 to determine the resultant volumetric airflow.

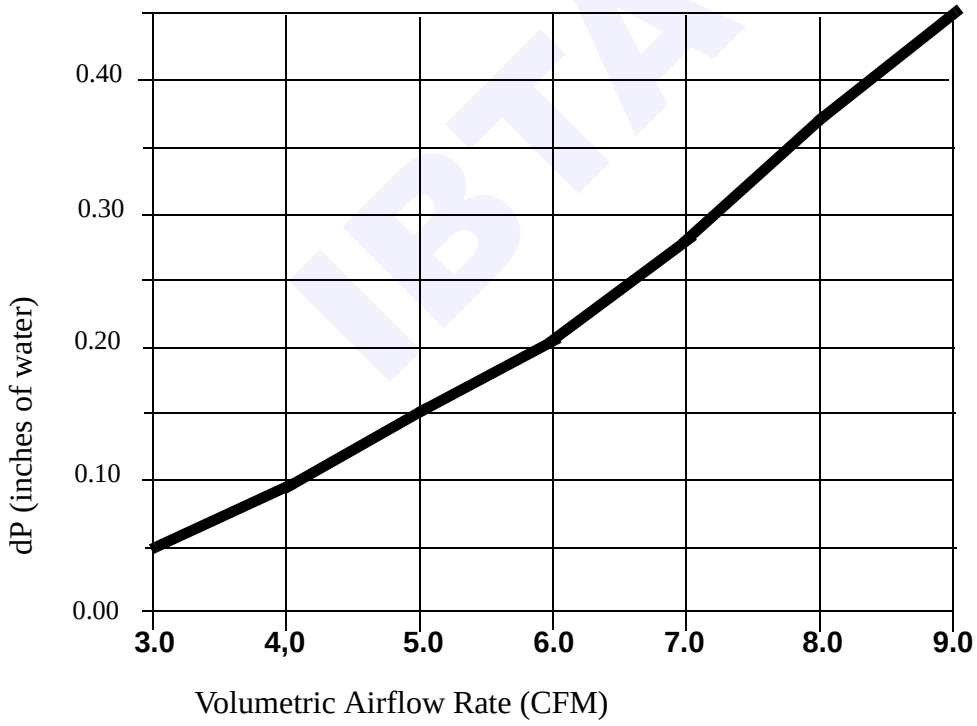


Figure 193 Nozzle calibration curve for standard module

A second dual-layer standard screen mesh (Screening B) is located approximately 250mm downstream of the nozzle in order to approximately straighten the airflow before the air reaches the module under test. The module is supported by walls that mimic the walls of the enclosure surrounding the module (including the backplane) as described in Figure 192

on page 741. Manometer B (the same type as used for manometer A) is used to measure the static pressure just upstream of the module under test and as described in Figure 192 on page 741.

The gap around the module between the connector housing and the inside of the slot should be sealed to prevent leakage and to better mimic EMI gaskets that will be in place on a final system.

A1.3.3 OPERATION OF THE PRESSURE DROP TEST

Fixture example:

To operate the test fixture, the operator sets the speed of the fan so that the volumetric airflow is set to a desired rate by reading manometer A, then the related static pressure is determined by reading manometer B.

A more specific description of operation of the test fixture is as follows:

1. Place the module in the slot of the test fixture.
2. Turn on power to the fan.
3. Select test criteria. It is suggested that the two airflow levels as shown in Table 94 on page 399 for each module size be tested to confirm that the related module airflow resistance falls between the minimum and maximum levels as described in Figure 123 on page 401. For example, for the Standard module, two different measurements would be made for airflow rates of 6 and 9 CFM.
4. From the horizontal axis of Figure 193 on page 742, select airflow level that is to be tested. Then determine the related differential pressure for the airflow level to be tested by reading the related pressure along the vertical axis.
5. Adjust the speed of the fan until manometer A reads the appropriate pressure from step 4. For example, to achieve 6 CFM of airflow, adjust the fan so that Manometer A reads 0.20 inches of water.
6. Read the measured pressure from manometer B to determine the related module resistance for that particular airflow rate.
7. Plot the airflow and pressure values onto Figure 123 on page 401. If the point falls between the minimum and maximum curves, then the module meets the specification.

A1.4 SHIELDING EFFECTIVENESS

The following paragraphs describe efforts in shielding performance simulations and modeling for InfiniBand™ Modules

In order to meet the InfiniBand™ shielding effectiveness specification shown in Figure 124 on page 404, a number of different hole patterns,

sizes and configurations were modeled using the Finite-Difference Time-Domain (FDTD) technique. This technique is particularly effective for this type of problem, since the thickness of the metal can be varied without the need to increase the model size (and hence its run time).

In Figure 194, Figure 195, Figure 196, and Figure 197, the shielding effectiveness of an InfiniBand™ module was modeled for various hole sizes and metal thickness to determine the required sizes to meet the specified shielding standard.

For figures Figure 198 on page 746, Figure 199, Figure 200, Figure 201, and Figure 202, another set of models was created for larger holes (10 x 10 mm), in 1, 2, 4, or 8 hole configurations to determine the effect of connector openings in the InfiniBand™ connector housing. Once again, various metal thickness were used to determine the required metal thickness to meet the specified shielding standard.

For all of the above simulations a wire source was placed 20 mm from the inside of the shield with holes, in order to represent a worst case situation of an EMI source close to the holes.

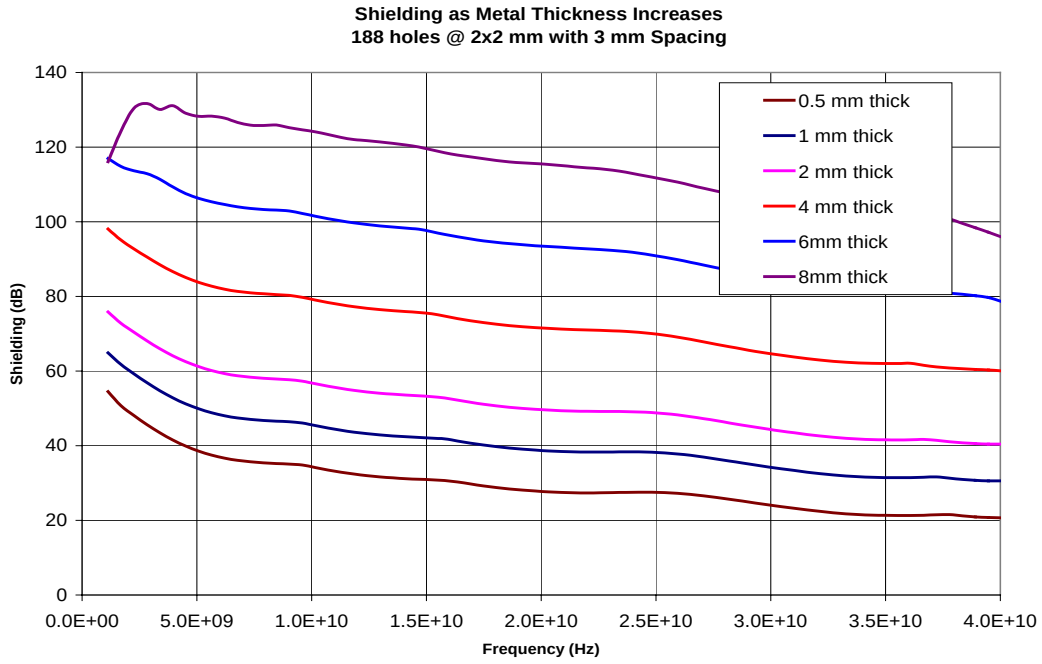


Figure 194 188 Holes, 2x2 mm Square 3mm Center-to-Center

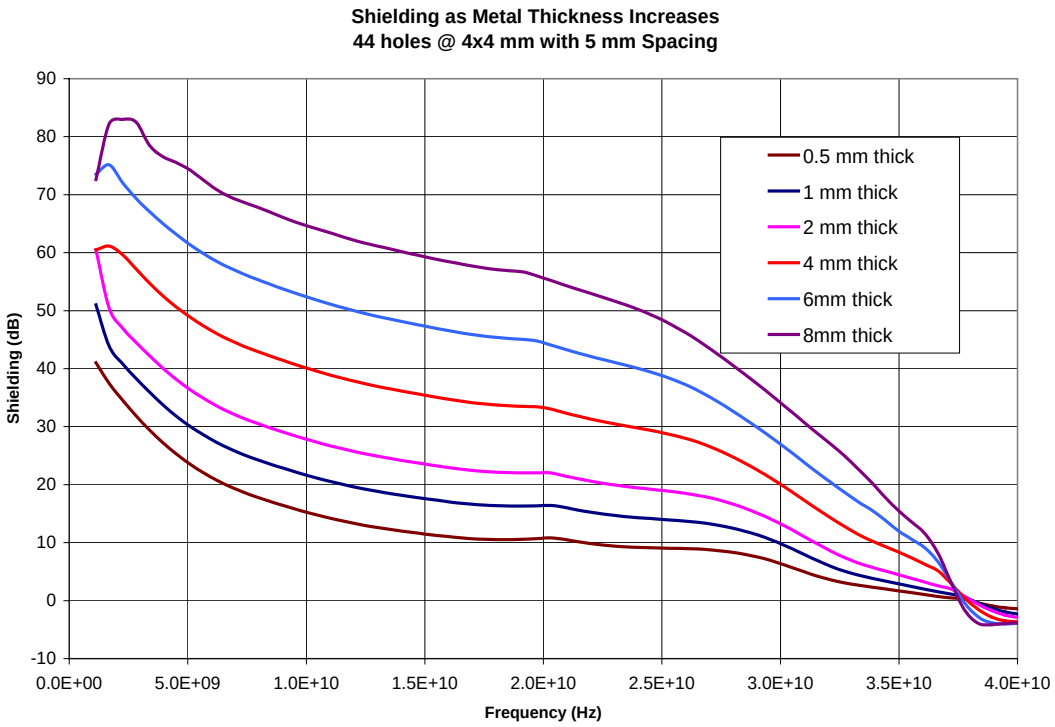


Figure 195 44 Holes, 4x4 mm Square 5mm Center-to-Center

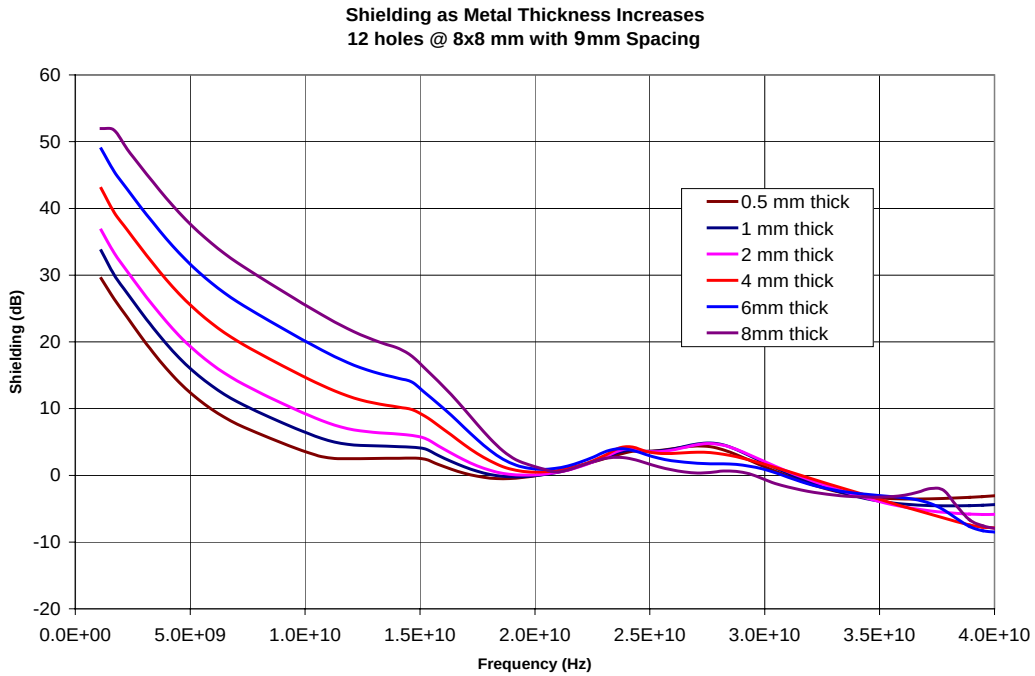


Figure 197 12 Holes, 8x8 mm Square 7mm Center-to-Center

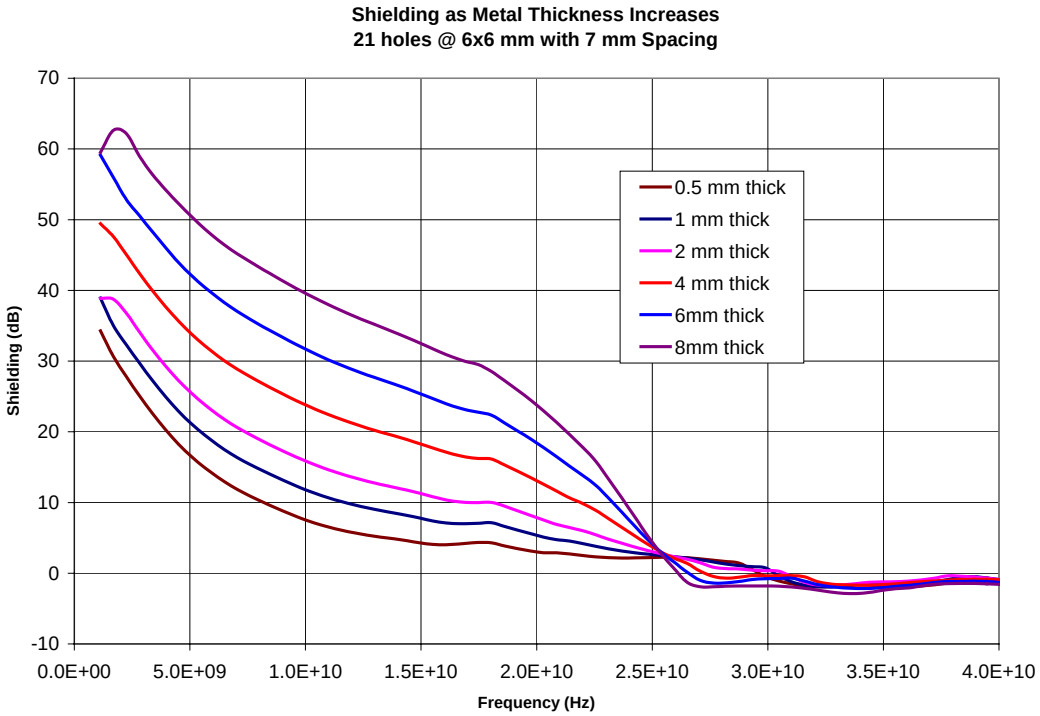


Figure 196 21 Holes, 6x6 mm Square 7mm Center-to-Center

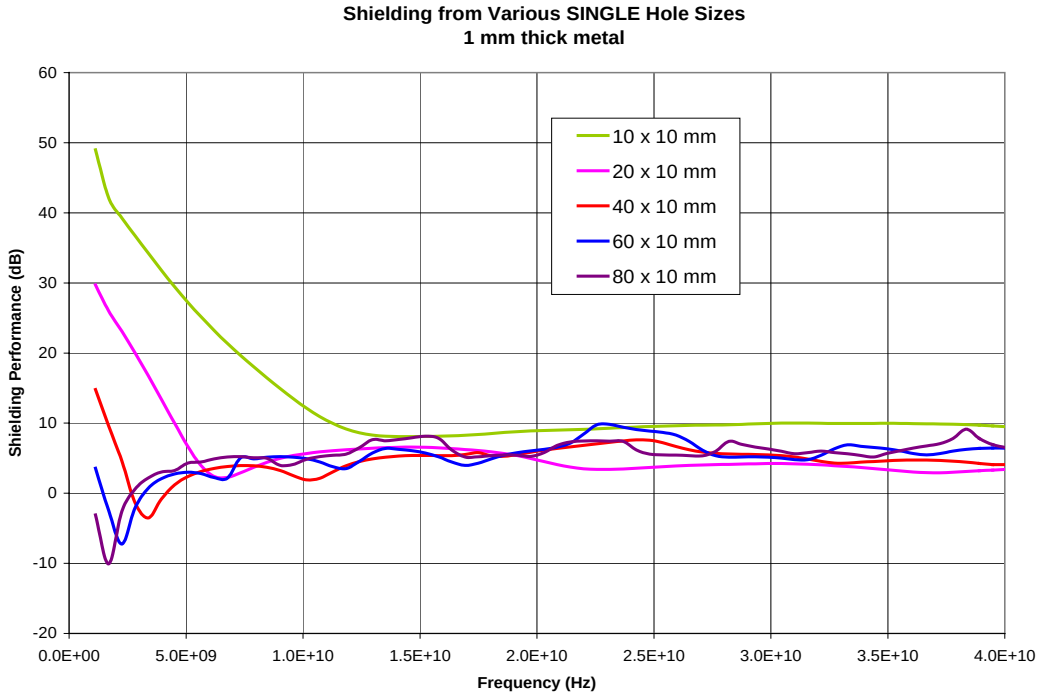


Figure 198 Shielding Performance: Single Hole, Various Sizes, 1mm Thick

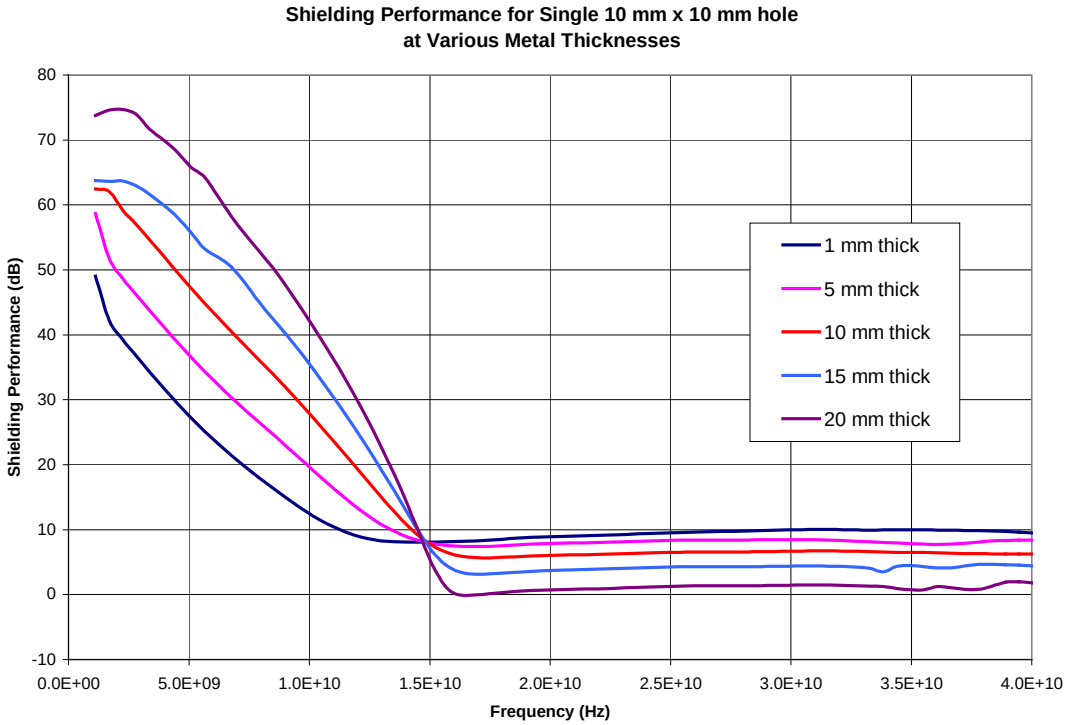


Figure 199 Shielding Performance: 10 x 10, Qty-1

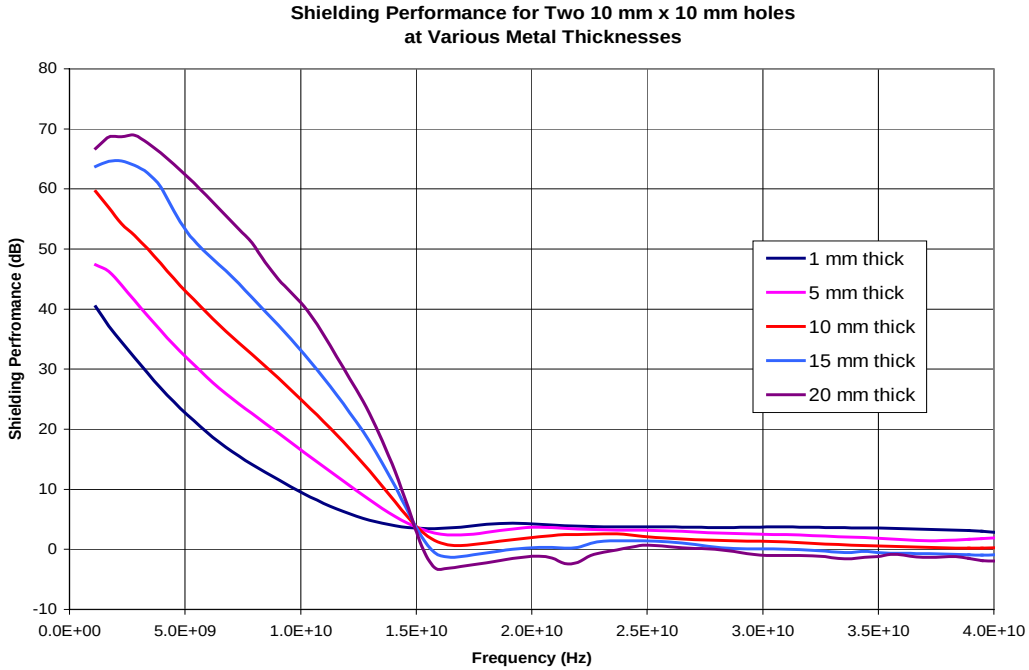


Figure 200 Shielding Performance: 10 x 10, Qty-2

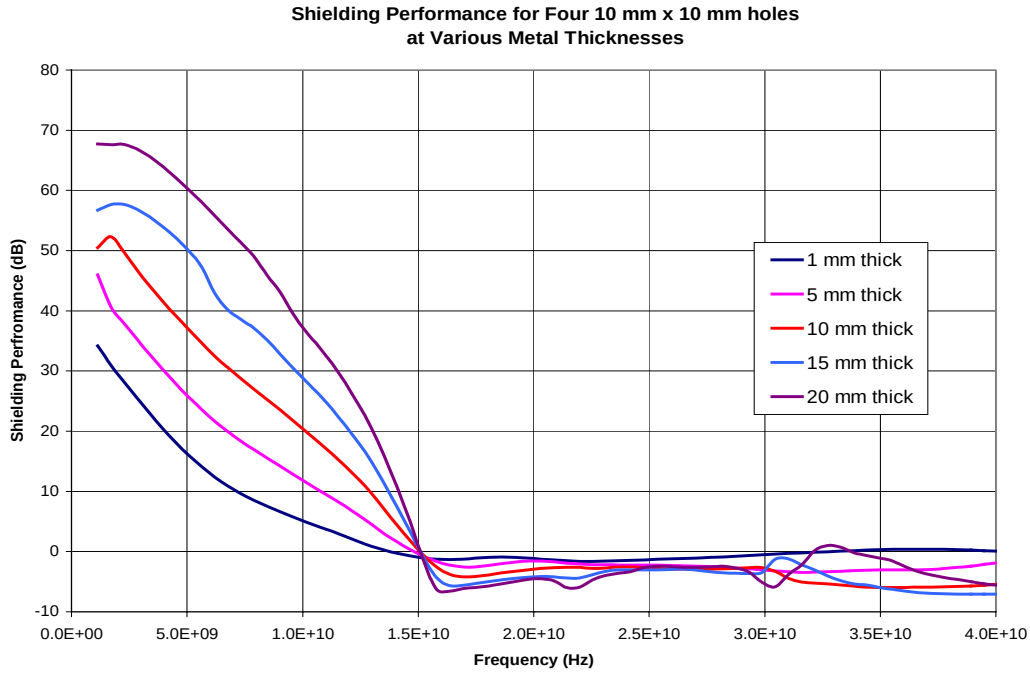


Figure 201 Shielding Performance: 10 x 10, Qty-4

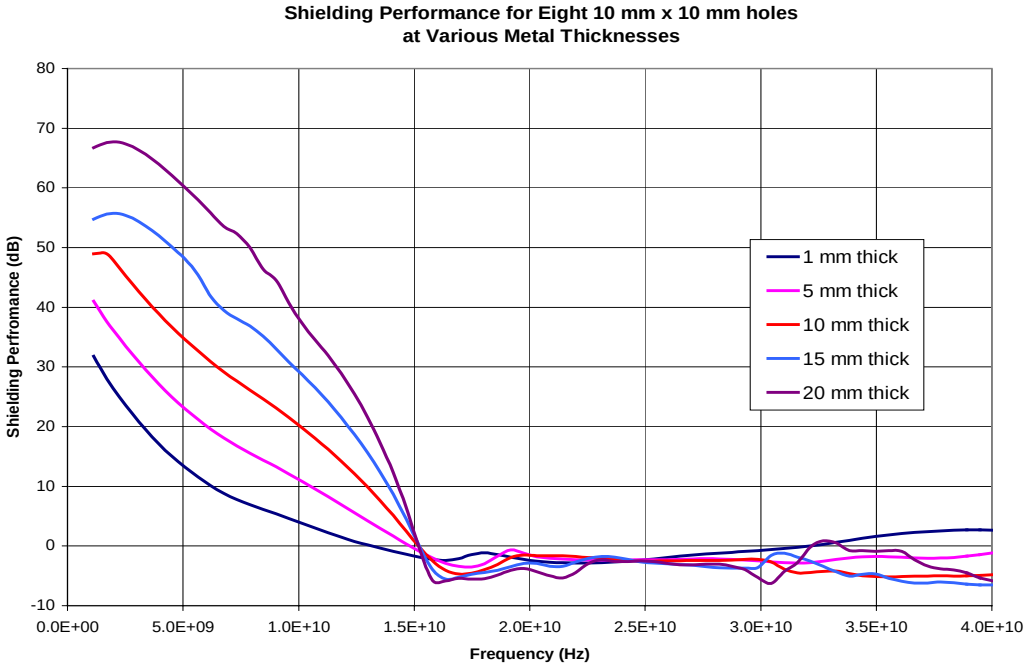


Figure 202 Shielding Performance, 10 x 10: Qty-8

A1.5 CARRIER ATTENUATION TEST PROCEDURE (NORMATIVE)

A1.5.1 INTRODUCTION

This section is normative. InfiniBand™ specifies very aggressive requirements for module carrier attenuation. With cooling vents included in the module carrier, the module itself becomes a strategic portion of a chassis' emi containment. Figure 124 on page 404 is the specified attenuation curve required for all module carriers. In order to verify these attenuation levels, an attenuation reference chassis has been developed. Details for this test vessel are included in Annex A1.6 on page 754. The intent of this procedure is to test the attenuation of the mechanics of the module carrier and not specifically the module's potential emissions. Each carrier must not only shield its own module emissions, but it also must attenuate potential emissions generated elsewhere in a real system. For these reasons, this test is set up to be run with non-operational module (or empty carrier) and a battery powered noise source capable of producing EMI spanning the band from 100 MHz to 12.5 GHz.

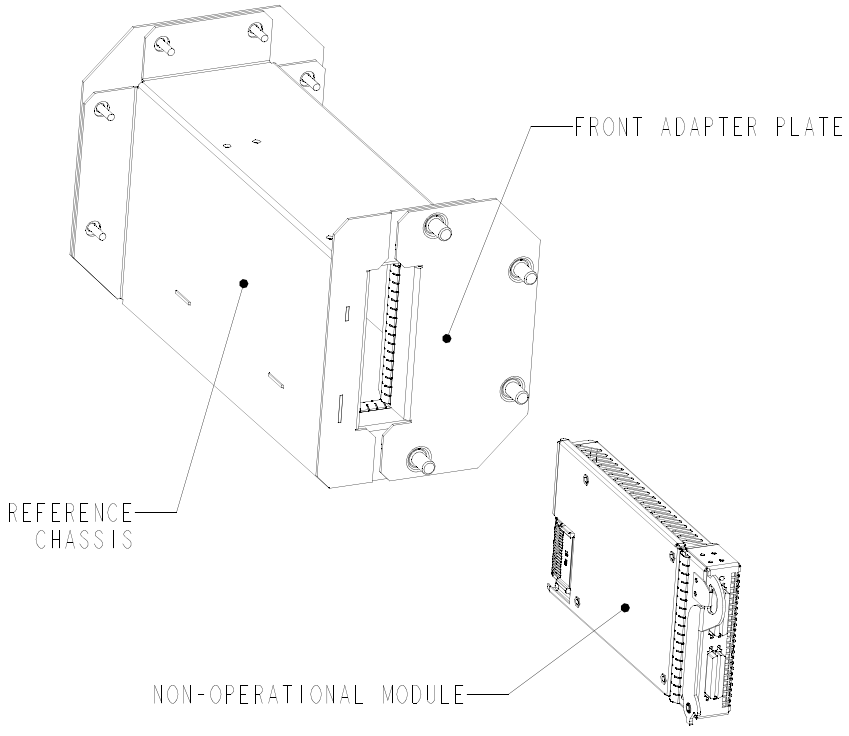


Figure 203 Attenuation Reference Chassis and Standard Module

For the purpose of this procedure the shielding effectiveness of the module shall be defined as the ratio of two values:

- Numerator: EMI measured with the noise source unobstructed (out in the open)
- Denominator: EMI measured with the noise source in the reference chassis and the module under test in place.

CA1-0.1.1: The shielding effectiveness **shall** be measured for both horizontal and vertical antenna polarities, at a source and antenna height of one meter, and a source to antenna distance of 2m.

CA1-0.1.2: The EMI measurements **shall** be made with the spectrum analyzer in maximum hold mode, while the unit under test is rotated through a 180 degree rotation. The test should be performed within a semi-anechoic chamber and **shall** be conducted over a ground plane.

A1.5.2 EQUIPMENT REQUIRED

- Semi-Anechoic Chamber with turntable
- HP 8566B Spectrum Analyzer or equivalent (measurement capability up to 12.5 GHz)
- EMCO 3148 Log Periodic Antenna or equivalent (measurement capability from.1 to 1 GHz)
- EMCO 3115 Horn Antenna or equivalent (measurement capability of 1 to12.5 GHz)
- Miteq NSP1800-N Preamplifier or equivalent (30 dB preamp, measurement capability.1 to 12.5 GHz)
- EMI Reference Chassis
- Battery Powered Noise Source(s) (EMI source for.1 to 12.5 GHz)

A1.5.3 MEASUREMENT PROCEDURE: 100 MHZ - 1.1 GHZ

1. Place the low frequency noise source on a wooden table 1m above the ground-plane of the turntable, set its antenna for vertical polarity.

2. Position the Log Periodic Antenna 2m away from the noise source, set its height to 1m, and rotate it for vertical polarization.

3. Connect the Preamplifier and Spectrum Analyzer to the Log Periodic Antenna. Set up the spectrum analyzer as follows:

Table 173 Spectrum Analyzer Settings, 100 MHz - 1.1 GHz

50dBuV	Reference Level
10 dB/div	Log Scale
0dB	RF attenuation
300 KHz	Video & IF(Resolution)BW
100 MHz	Start Frequency
1100 MHz	Stop Frequency
Channel A - Trace Clear/Write	Display Mode

4. Turn on the noise source.

5. Change the display mode for channel A of the Spectrum Analyzer from Clear/Write to Maximum Hold.

6. Rotate the turntable through 180 degrees, after rotation is complete set the display mode for channel A to view. Set the display mode for channel B of the spectrum analyzer to clear/write.

7. Connect the standard or wide front adapter plate as appropriate to the front of the reference-IB chassis as shown in Figure 203 on page 749.

8. Insert the noise source into the rear compartment of the reference chassis, making sure its antenna is vertically polarized, and seal it inside using the rear cover for the chassis.

9. Insert the module/carrier combination to be tested into the IB chassis slot.

10. Change the display mode for channel B of the Spectrum Analyzer from Clear/Write to Maximum Hold.

11. Rotate the turntable through 180 degrees, making sure that EMI from that module has been measured across the entire front of the IB chassis/module combination. After rotation is complete, set the display mode for channel B to view.

12. Set the spectrum analyzer to subtract trace b from trace a, the resulting trace will be the shielding effectiveness of the module and must comply with the shielding effectiveness specified in Figure 124 on page

404 of the specification. Print the results and maintain as a record to show compliance with the specification.

13. Take the noise source out of the reference chassis, and remove the reference chassis from the wooden table.

14. Set the noise on the table such that its antenna is in the horizontal polarity.

15. Rotate the Log Periodic antenna into the horizontal Polarity, and repeat steps 3 through 12, insuring that the antenna of the noise source is in the horizontal polarity when it is inserted into the reference chassis.

A1.5.4 MEASUREMENT PROCEDURE: 1 GHZ - 12.5 GHZ

1. Place the high frequency noise source on a wooden table 1m above the ground-plane of the turntable, set its antenna for vertical polarity.

2. Position the Horn Antenna 2m away from the noise source, set its height to 1m, and rotate it for vertical polarization.

3. Connect the Preamplifier and Spectrum Analyzer to the Horn Antenna. Set up the spectrum analyzer as follows:

Table 174 Spectrum Analyzer Settings, 1 GHz - 12.5 GHz

50dBuV	Reference Level
10 dB/div	Log Scale
0dB	RF attenuation
300 KHz	Video & IF(Resolution)BW
1.00 GHz	Start Frequency
12.50 GHz	Stop Frequency
Channel A - Trace Clear/Write	Display Mode

4. Turn on the noise source.

5. Change the display mode for channel A of the Spectrum Analyzer from Clear/Write to Maximum Hold.

6. Rotate the turntable through 180 degrees, after rotation is complete set the display mode for channel A to view. Set the display mode for channel B of the spectrum analyzer to clear/write.

7. Connect the standard or wide front adapter plate as appropriate to the front of the reference chassis as shown in Figure 203 on page 749.

8. Insert the noise source into the rear compartment of the reference chassis, making sure its antenna is vertically polarized.

9. Insert the module/carrier combination to be tested into the IB chassis slot.

10. Change the display mode for channel B of the Spectrum Analyzer from Clear/Write to Maximum Hold.

11. Rotate the turntable through 180 degrees, making sure that EMI from that module has been measured across the entire front of the IB chassis/module combination. After rotation is complete, set the display mode for channel B to view.

12. Set the spectrum analyzer to subtract trace b from trace a, the resulting trace will be the shielding effectiveness of the module and must comply with the shielding effectiveness specified in Figure 124 on page 404 of the specification. Print the results and maintain as a record to show compliance with the specification.

13. Take the noise source out of the reference chassis, and remove the reference chassis from the wooden table.

14. Set the noise on the table such that its antenna is in the horizontal polarity.

15. Rotate the Horn antenna into the horizontal polarity, and repeat steps 3 through 12, insuring that the antenna of the noise source is in the horizontal polarity when it is inserted into the reference chassis.

A1.5.5 TEST COMPLETION

CA1-0.1.3: After this procedure has been completed, four charts showing the shielding effectiveness of the module will have been generated. These charts fully document the shielding effectiveness of the module over the frequency range of 100 MHz-12.5 GHz for both horizontal and vertically polarized noise sources and **shall** meet or exceed the shielding effectiveness specified in Figure 124 of the specification. These charts **shall** be maintained as records to show that the module is in compliance with the IB mechanical specification.

IBTA

A1.6 FABRICATION AND ASSEMBLY OF THE EMI REFERENCE CHASSIS

A1.6.1 INTRODUCTION

Two reference chassis are detailed in this annex. The Standard Height Reference Chassis will support testing of Standard or Wide modules using a separate front cover assembly to adapt to standard or wide. The Tall Height Reference Chassis will support testing of Tall or Tall Wide modules.

Detailed fab parts are shown in the following figures. Some of the fab parts are parametrically represented where a parts dimensions deviate for standard versus wide or standard versus tall. In each of these cases, the standard size dimension is represented first followed by the tall (T) or wide (W) dimension in parentheses.

The chassis consists of a welded enclosure open in the rear and partially open in the front. The chassis is specified with internal removable module guides and specifically located EMI gasketing. The rear of the chassis is void for placement of the signal generator which can be accessed through the removable rear panel (see Figure 204 on page 756).

A1.6.2 PARTS LIST

Table 175 is a detailed parts list for each size of EMI reference chassis.

Table 175 Detailed Parts List for Standard and Tall EMI Reference Chassis

Standard Height Reference Chassis		Tall Height Reference Chassis	
QTY	Assy, Fab, or Purchase Part	QTY	Assy, Fab, or Purchase Part
1	Chassis Weldment, Standard	1	Chassis Weldment, Tall
1	Housing, Standard	1	Housing, Tall
1	Front, Standard	1	Front, Tall
1	Base	1	Base
2	Guide	2	Guide
1	Rear Cover Assy, Standard	1	Rear Cover Assy, Tall
1	Rear Cover, Standard	1	Rear Cover, Tall
8	SouthCo 84-20-080-30	8	SouthCo 84-20-080-30
*	EMI Gaskets Cut to Length *	*	EMI Gaskets Cut to Length *
1	Front Cover Assy, Standard	1	Front Cover Assy, Tall
1	Front Cover, Standard	1	Front Cover, Tall
4	SouthCo 84-20-080-30	4	SouthCo 84-20-080-30

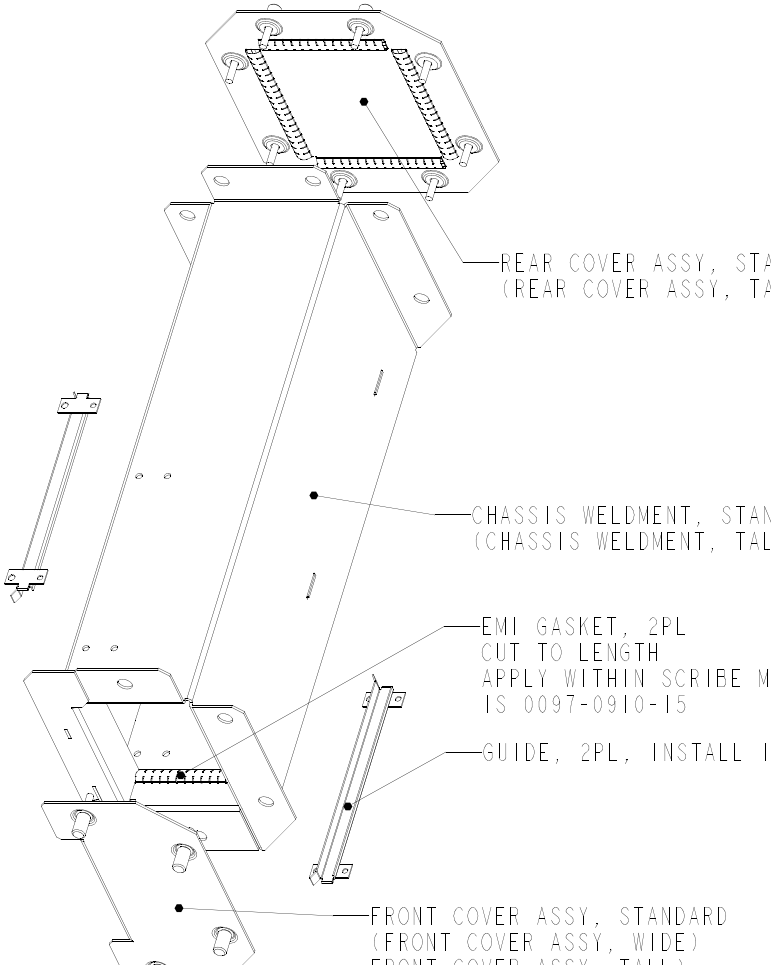
IBTA

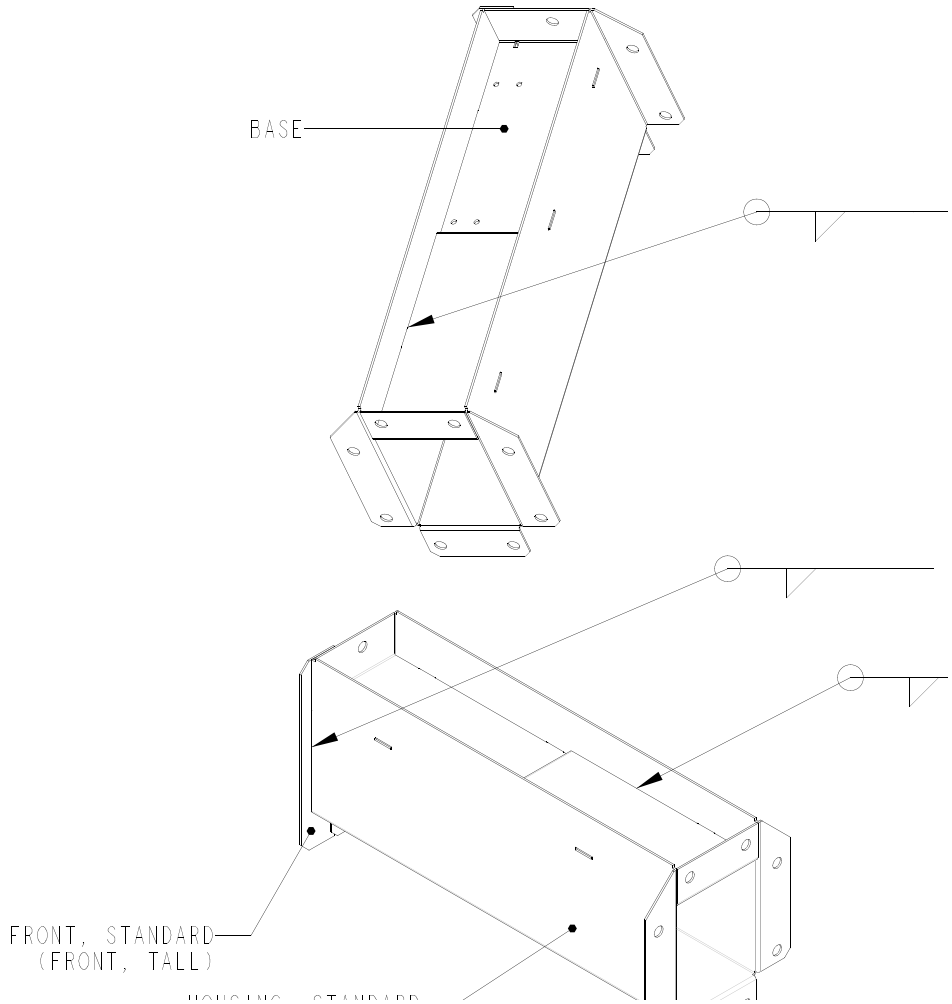
Table 175 Detailed Parts List for Standard and Tall EMI Reference Chassis

Standard Height Reference Chassis		Tall Height Reference Chassis	
QTY	Assy, Fab, or Purchase Part	QTY	Assy, Fab, or Purchase Part
*	EMI Gaskets Cut to Length *	*	EMI Gaskets Cut to Length *
1	Front Cover Assy, Wide	1	Front Cover Assy, Tall/Wide
1	Front Cover, Wide	1	Front Cover, Tall/Wide
4	SouthCo 84-20-080-30	4	SouthCo 84-20-080-30
*	EMI Gaskets Cut to Length *	*	EMI Gaskets Cut to Length *
3	* EMI Gasket (15" length) Instrument Specialties 0097-0910-15	3	* EMI Gasket (15" length) Instrument Specialties 0097-0910-15
1	Quarter Turn Application Tool-SouthCo 82-0-14719-11	1	Quarter Turn Application Tool-SouthCo 82-0-14719-11
1	Signal Generator, **TBD**	1	Signal Generator, **TBD**

IBTA

A1.6.3 ASSEMBLY/FABRICATION DRAWINGS




BASE
FRONT, STANDARD
(FRONT, TALL)

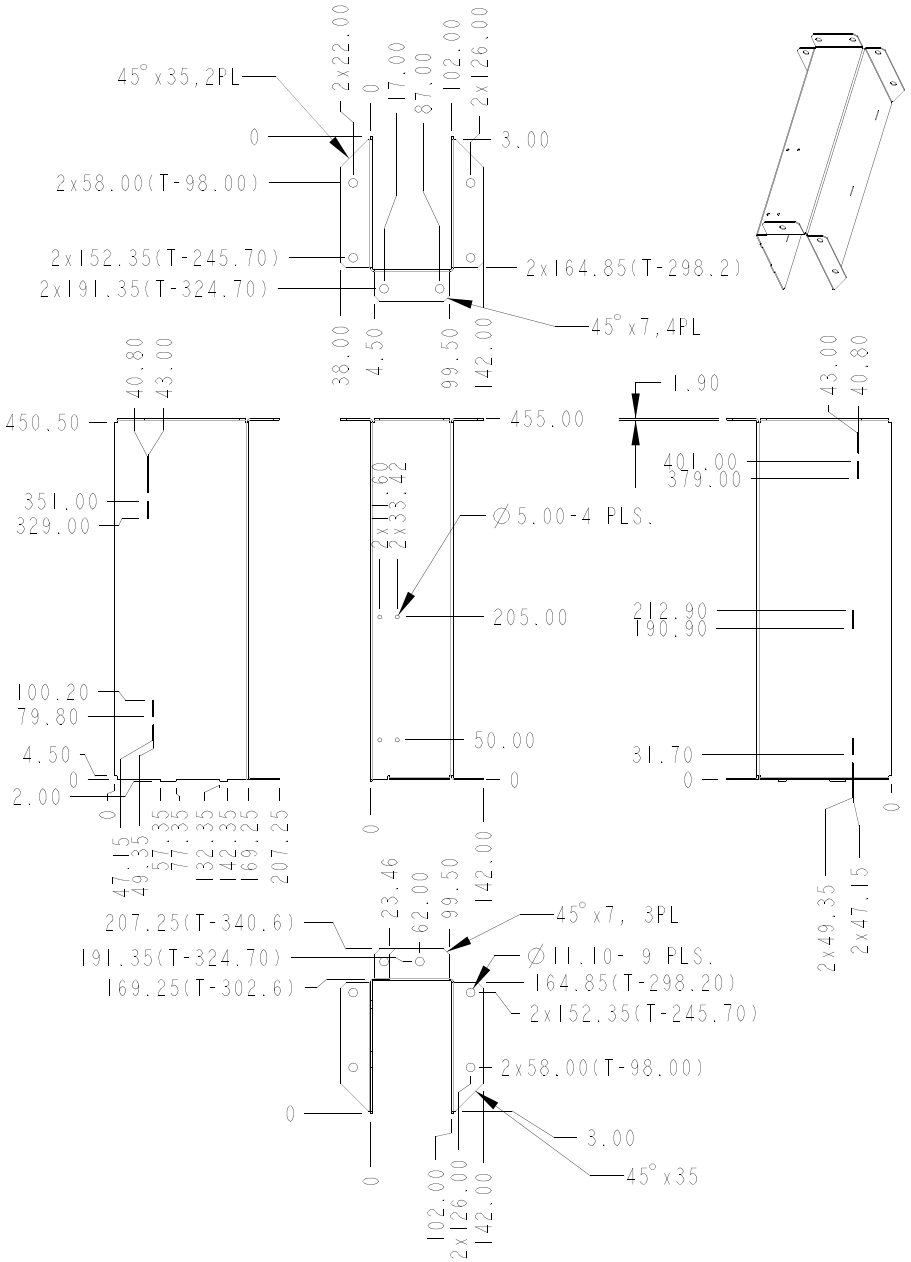


Figure 206 Housing - Standard; Tall (T)

*** Note:** Primary dimensions (shown first) are for the standard height enclosure. Dimensions in parentheses are for a tall enclosure, (T- XX.X)

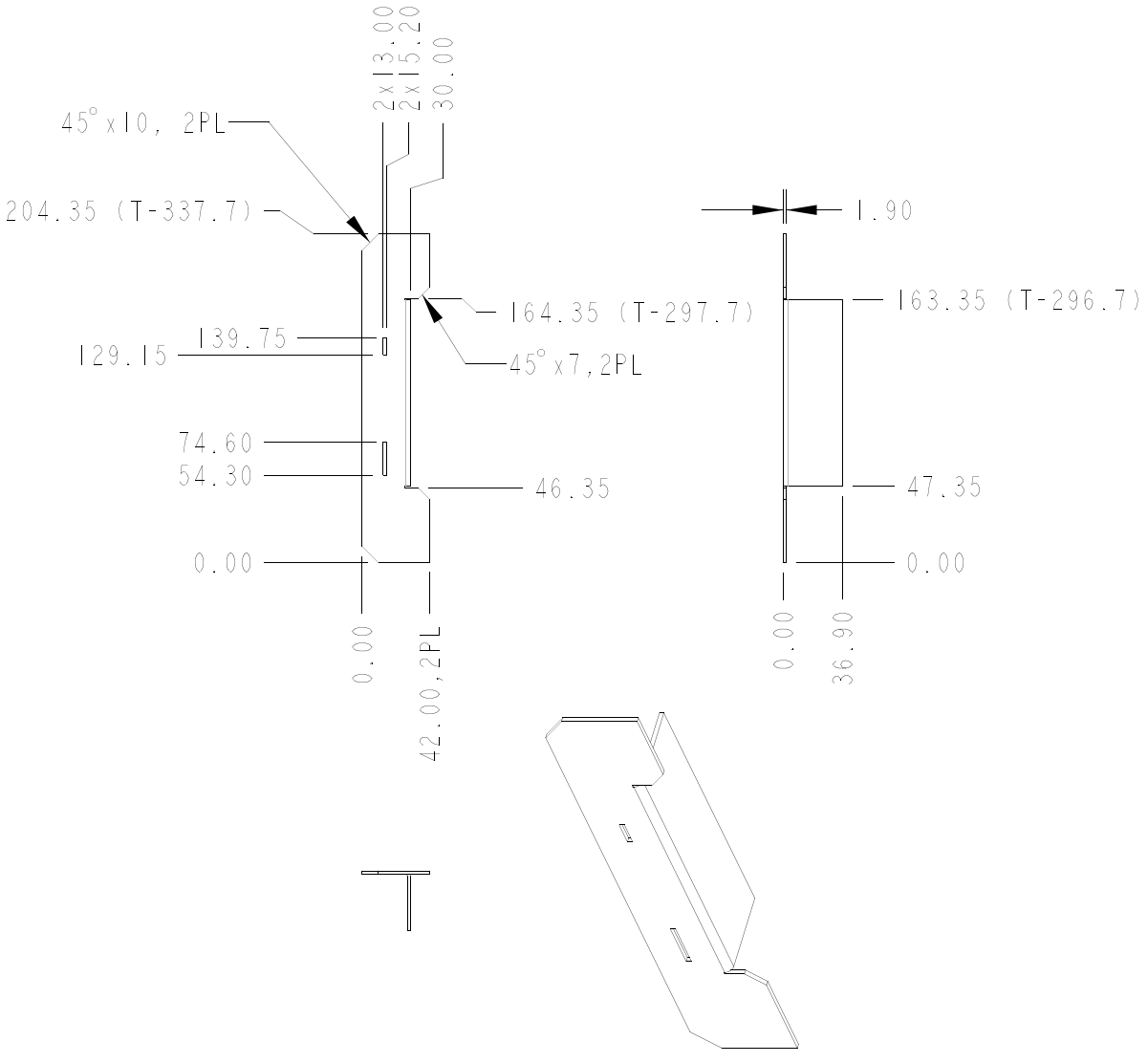


Figure 207 Front - Standard; Tall (T)

*** Note:** Primary dimensions (shown first) are for the standard height enclosure. Dimensions in parentheses are for a tall enclosure, (T- XX.X)

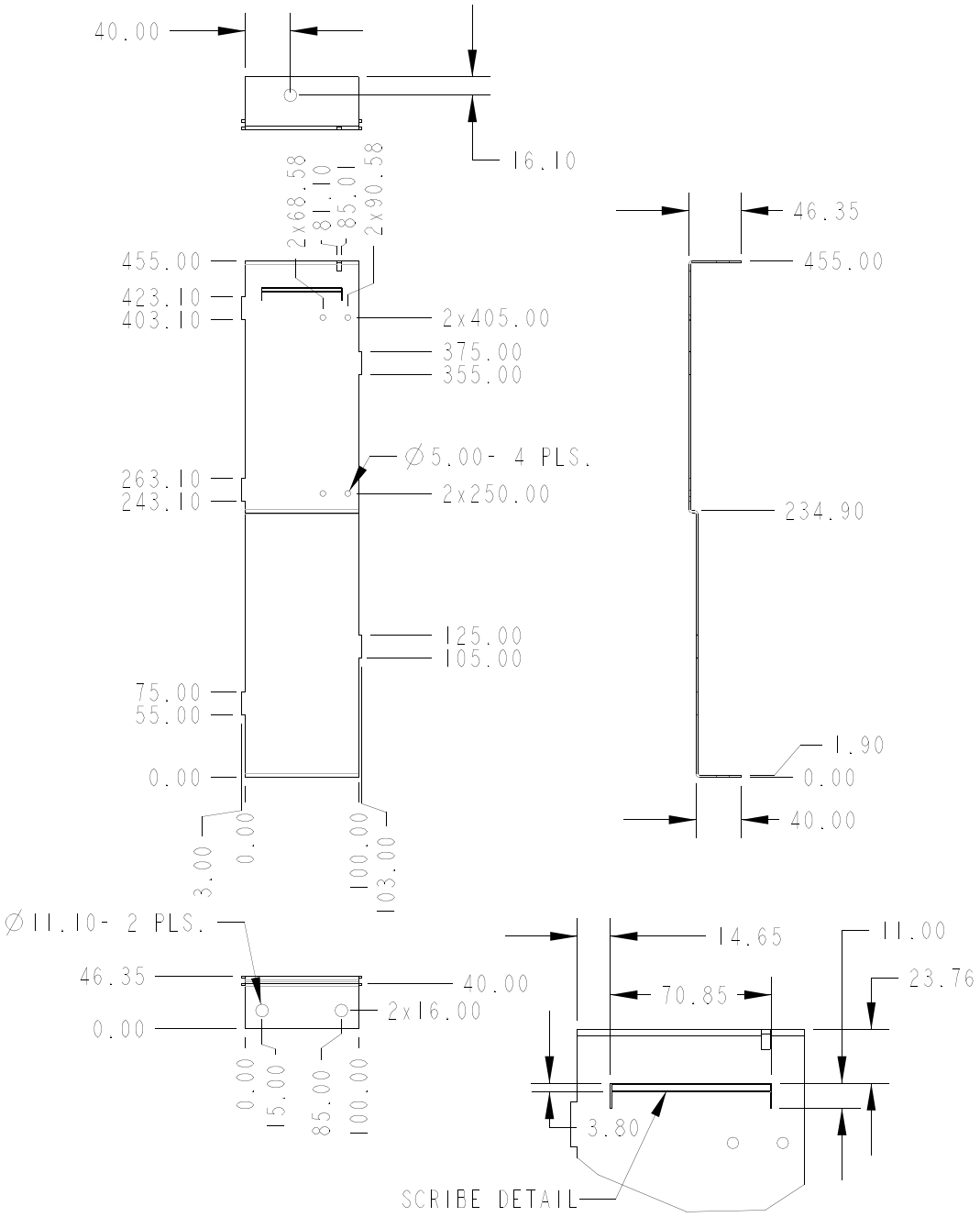


Figure 208 Base

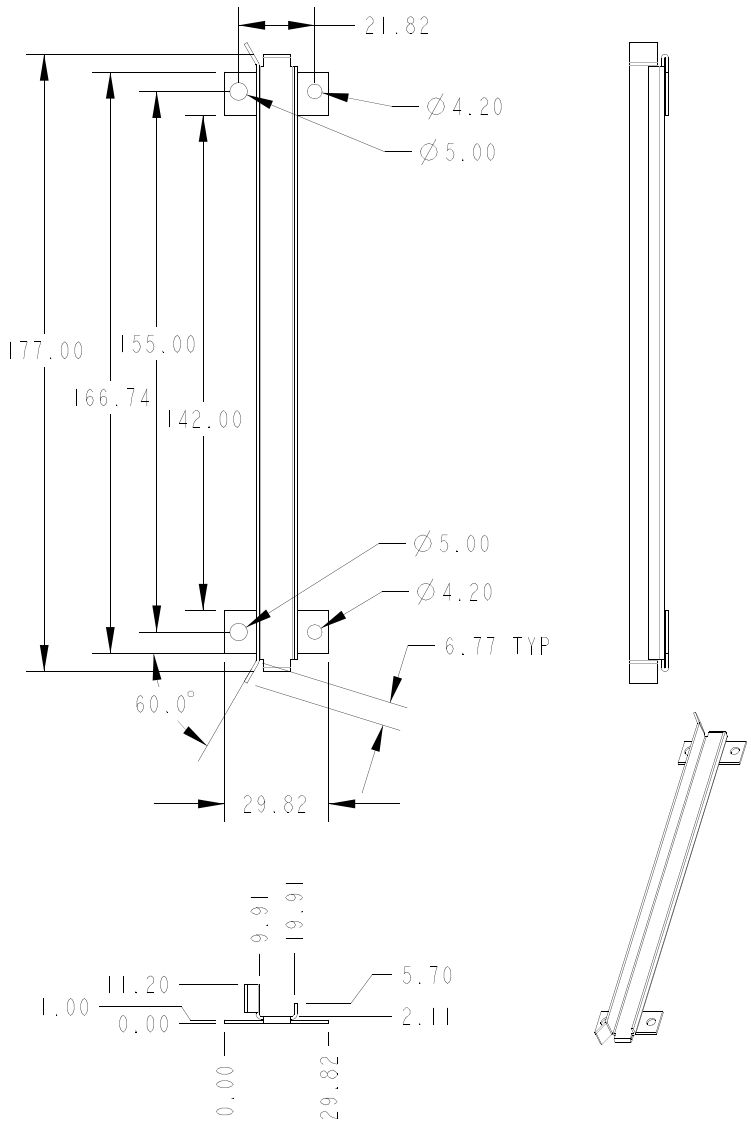


Figure 209 Guide

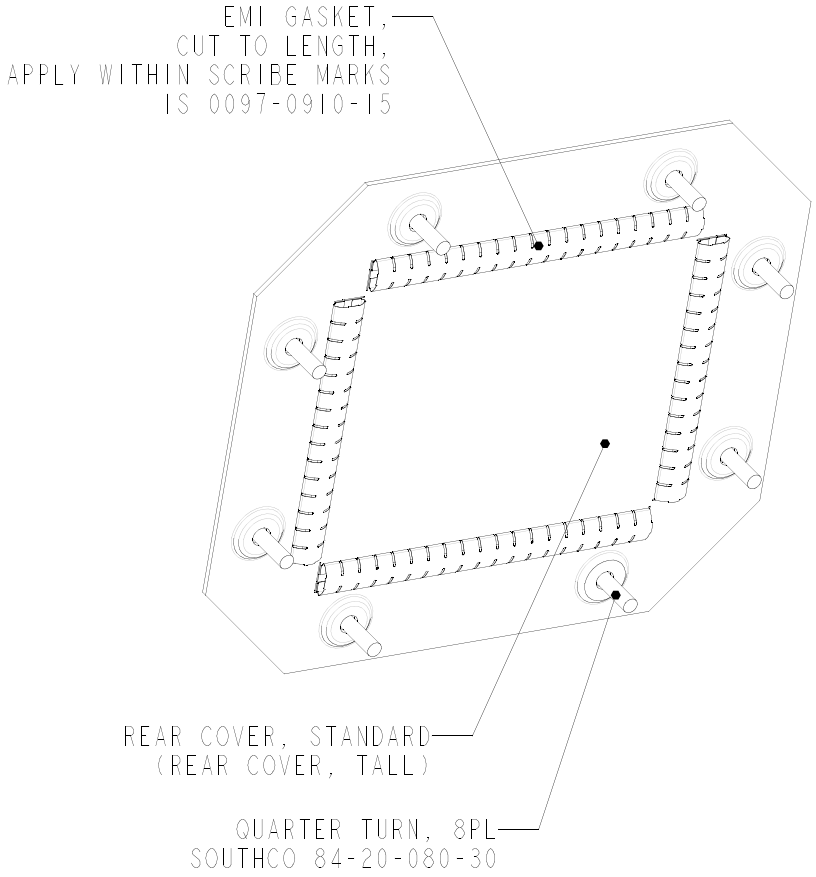


Figure 210 Rear Cover Assembly - Standard; Tall

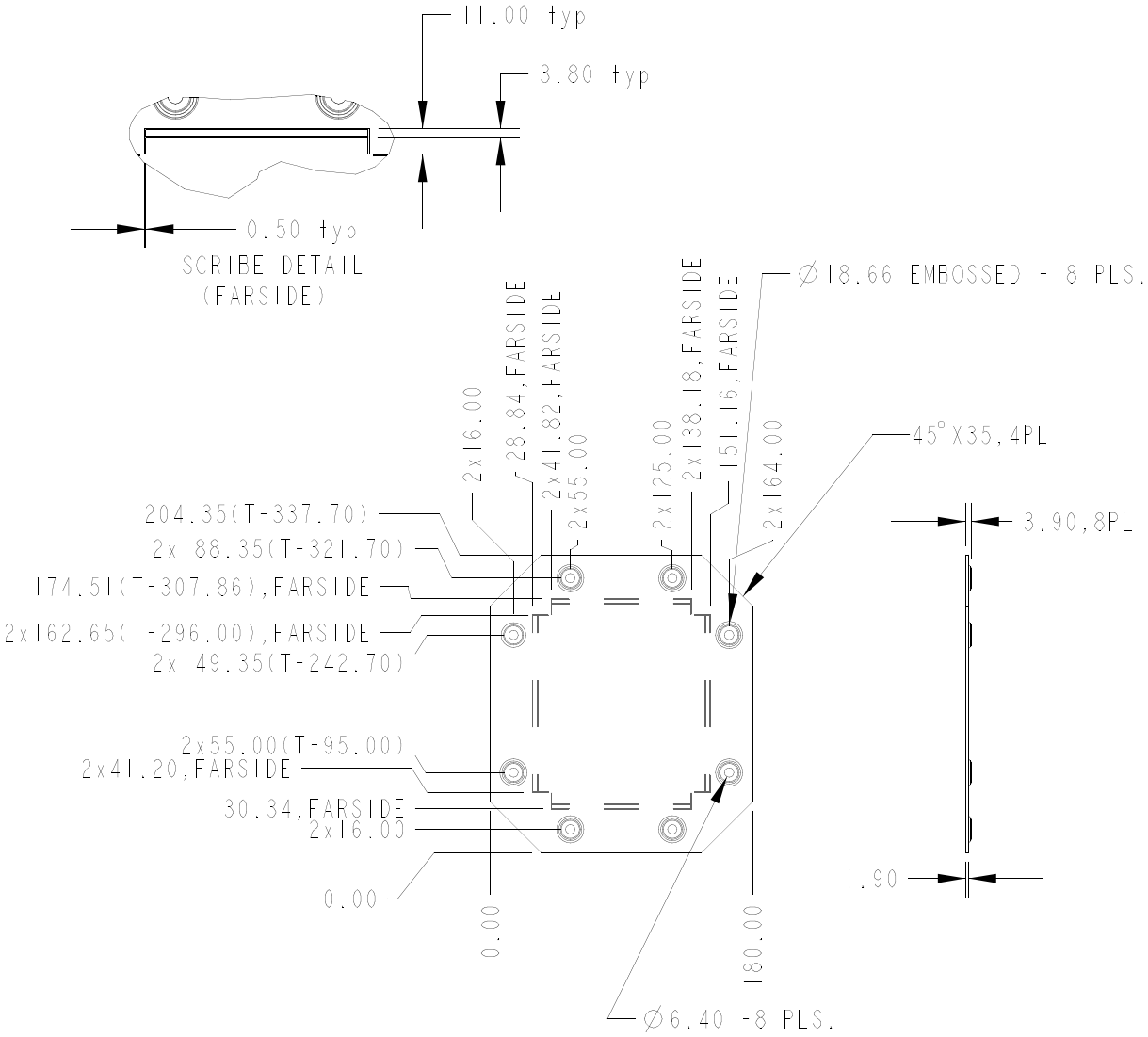


Figure 211 Rear Cover - Standard; Tall (T)

*** Note:** Primary dimensions (shown first) are for the standard height enclosure. Dimensions in parentheses are for a tall enclosure, (T- XX.X)

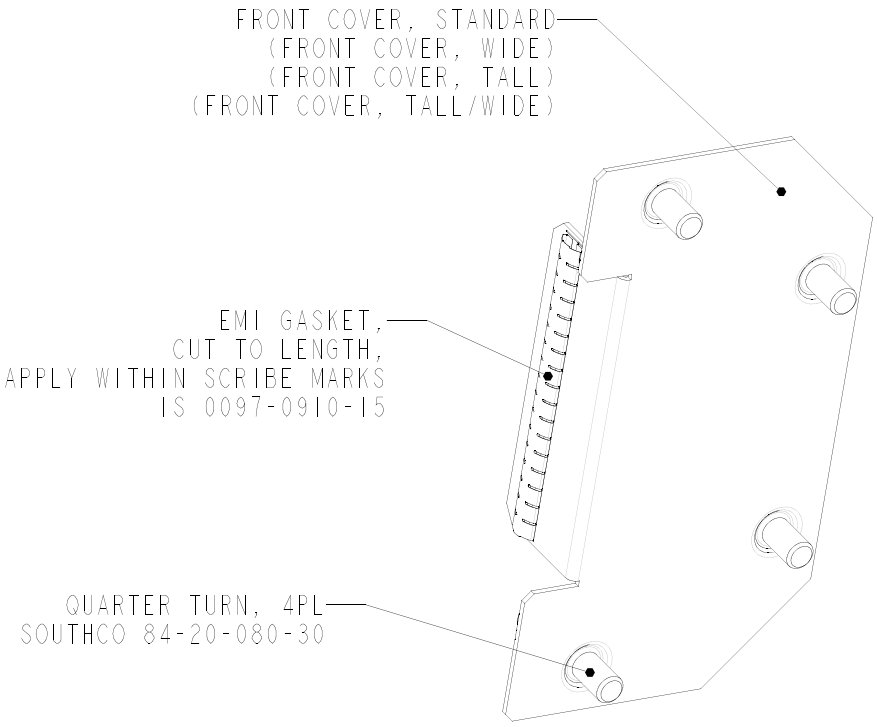


Figure 212 Front Cover Assembly - Standard; Wide; Tall; Tall/Wide

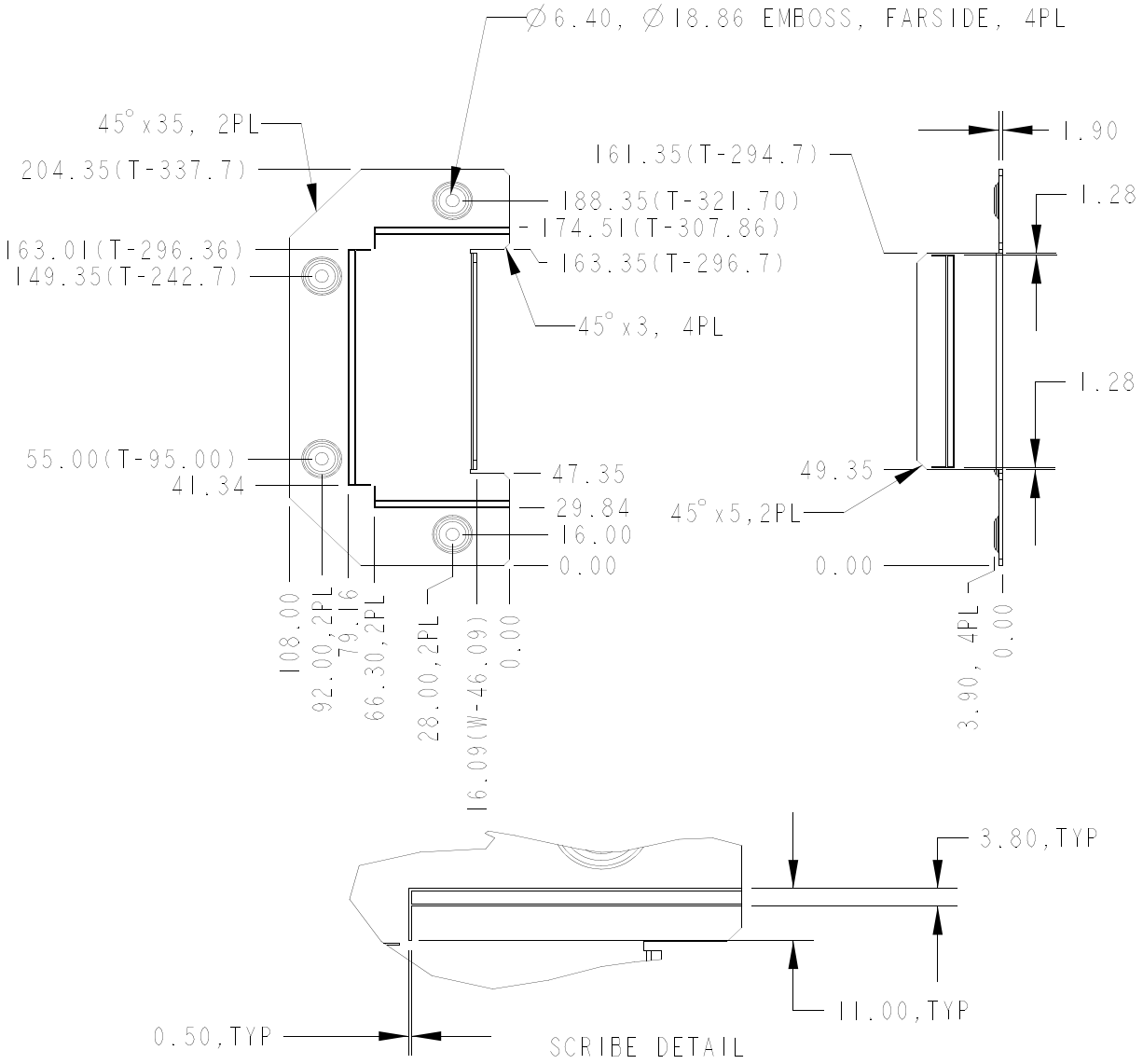


Figure 213 Front Cover - Standard; Tall; Wide; Tall/Wide

*** Note:** Primary dimensions (shown first) are for the standard height enclosure. Dimensions in parentheses are for either the tall or wide option, (T- XX.X or W - XX.X)

IBTA

ANNEX A2: IB-ML DESIGN GUIDELINES

A2.1 INTRODUCTION

The purpose of this annex is to aid in the design of the InfiniBand Management Link (IB-ML) using SMBus and/or I^2C agents. This annex covers design considerations of the SMBus and I^2C bus and also highlights the protocol, electrical, and timing differences between the SMBus and I^2C bus. Calculation of proper bus signal pullup resistor values is also covered. It is assumed that the reader is familiar with the concepts of the IB-ML, SMBus and I^2C bus.

A2.2 PROTOCOL

This section shows the protocol differences between the IB-ML, SMBus and the I^2C bus. The InfiniBand Management Link timing and multi-master arbitration protocol is based on the SMBus 1.1 specification[24]. It is recommended that interfaces be designed to also meet the I^2C timing specifications.

Although highly recommended, devices are not restricted to using the SMBus protocols. Devices must use protocols that use simple write-read transactions supported by the IB-to-IBML command. Devices are recommended to implement PEC (Packet Error Code) on their interfaces to help ensure data integrity.

One should consider the protocol and specific addresses supported by all agents on the bus to ensure proper bus operation. For a complete list and explanation of protocols supported by each type of bus, refer to Chapter 13: Hardware Management, the I^2C bus and the SMBus specifications. The I^2C Bus specification is available from Phillips Semiconductors [25]. Copies of the SMBus Specification Revision 1.1 are available from Intel Corporation and the Smart Battery Systems Implementers Forum [24].

A2.2.1 SMBUS AND I^2C BUS PROTOCOL DIFFERENCES

The following protocol differences exist between the SMBus and the I^2C bus:

1) The SMBus specifies device time-outs which can be used to signal a device error condition or that a device is not ready. Device time-out is accomplished by holding either SMBCLK or SMBDATA low for longer than $T_{TIMEOUT}$. It is recommended that devices automatically drop off the bus whenever $T_{TIMEOUT}$ is exceeded.

2) The SMBus allows a slave device to stretch the cumulative clock low time, in a single message, up to $T_{LOW:SEXT}$.

3) The SMBus allows a master device to stretch the commutative clock low time, in any single byte, up to $T_{LOW:MEXT}$.

4) The SMBus can specify the protocol an SMBus agent is allowed to use when acting as a slave when communicating with an SMBus host.

5) The I^2C bus protocol doesn't include the Quick Command, but it is valid and will not cause an error from I^2C devices. The Quick Command will look to I^2C devices as a command abort by the bus master.

6) I^2C does not specify a fixed host device address as does the SMBus.

7) The SMBus reserves the same bus protocol addresses as the I^2C bus, as well as additional SMBus-specific reserved addresses.

8) SMBus timing specifies that a master must hold data for 300 ns following the falling edge of clock, whereas I2C specifies a 0 ns data hold time. As a result, some SMBus slave devices may not work correctly if presented with the I^2C timing from a master. However, if a master implements SMBus timing for driving the bus, both I2C and SMBus slaves should work. Conversely, if a slave can work with I2C timing, it should work with both I^2C and SMBus master. Thus, it is recommended that devices provide the SMBus 300 ns data hold time when driving the bus as a master, and accept the 0 ns I^2C hold time when receiving data as a slave.

The SMBus exclusive reserved protocol addresses are listed in Table 176.

Table 176 SMBus Reserved Addresses

Address	SMBus Function
0001 000	SMBus Host (devices on IB-ML should avoid this address unless they implement SMBus Host functionality.)
0001 100	SMBus Alert Response (devices on IB-ML that use this address must do so per the SMBus specification)
1100 001	SMBus Device Default (devices on IB-ML that use this address must do so per the SMBus 2.0 specification)
0101 000	Reserved for ACCESS.BUS (allowed for device use on IB-ML)
0110 111	Reserved for ACCESS.BUS default (allowed for device use on IB-ML)
1001 0xx	Unrestricted Address (allowed for device use on IB-ML)

Besides the addresses listed as reserved by the SMBus, Table 177 provides a list of command addresses reserved for protocol use by both the

SMBus and the I²C bus. Table 177 provides a list of addresses used by some motherboards.

Table 177 SMBus and I²C Bus Reserved Addresses

Slave Address	Read/Write # Bit	SMBus and I²C Bus Function
0000 000	0	General call address
0000 000	1	START byte[a]
0000 001	X	CBUS address[b]
0000 010	X	Address reserved for different bus format[c]
0000 011	X	Reserved
0000 1xX	X	Reserved
1111 0XX	X	10-bit slave addressing[d]
1111 1xX	X	Reserved

a. No device is allowed to acknowledge at the reception of the START byte.
b. The CBUS address has been reserved to enable the inter-mixing of CBUS compatible and I²C-bus compatible devices in the same system. I²C-bus compatible devices are not allowed to respond on reception of this address.
c. The address reserved for a different bus format is included to enable I²C and other protocols to be mixed. Only I²C-bus compatible devices that can work with such formats and protocols are allowed to respond to this address.
d. Support for 10-bit slave addressing is not specified nor required for IB-ML.

Table 178 Commonly Used Motherboard Addresses

Slave Address	Read/Write # Bit	SMBus and I²C Bus Function
1010 XXX	X	Common 24C02-style SEEPROM address range. Also used for DIMM module SEEPROMs.
1001 110	X	I²C Mux address
1101 001	X	Clock generator address
0101 101	X	LM-78/ADM1024-style Monitoring ASIC addresses.

Table 179 on page 770 lists the recommended ranges of addresses usable by an InfiniBand™ module. This list is an extension of the list given in the Chapter 13: Hardware Management. Modules that are embedded in a chassis can use other addresses at the discretion of the system designer. Note that functional addresses for elements such as the ModuleInfo, ChassisInfo, MME Function registers, and MME IB2IBML registers should not be claimed even for an embedded module. In particular, the ModuleInfo address must be preserved, since software uses that

address as the root for discovering Module and Chassis capabilities via the IB-to-IBML Baseboard Management command functionality.

Implementation Note

IMPORTANT - Check your bus device addresses carefully.

All SMBus addresses (other than those listed in Table 176 and Table 177) are stated as being reserved for assignment by the SMBus address coordination committee of the Smart Battery Systems Implementer's Forum (SBS-IF). This is generally only applied to specifying addresses for certain devices that have standardized functional interfaces specified by the SBS-IF, such as battery charger ICs and battery monitors. It is common for device manufacturers to utilize these addresses without coordination with the SBS-IF, particularly on busses where the SMBus functionally addressed devices are not likely to be used, such as on the IB-ML between an InfiniBand™ chassis and InfiniBand™ module.

Philips Semiconductor (the I^2C address assignment agency) has created many addresses other than the common SMBus and I^2C protocol addresses for use by their specific devices. Designers may want to consult Philips and other vendor catalogs when choosing addresses in case there are components from those vendors the designer may want to use in the future.

Pay close attention to the specific addresses of the bus agents to avoid address conflicts. This especially applies when mixing SMBus devices and I^2C bus devices on the same physical bus.

Table 179 Allowed Addresses for InfiniBand™ Modules

Slave Address	8-bit Hex notation	Function
0100 000 - 0100 011	40h-46h	Module-specific use. Typically used for 8574-8-bit and 8575-type 16-bit I^2C latches.
0101 001 - 0101 111	50h-5Eh	Module-specific use. LM78/ADM1024 and other 'Heceta' -type monitoring devices.
1001 000 - 1001 111	90h-9Eh	Module-specific use. Typically used for LM75-style temperature sensors.
1010 000	A0h	ModuleInfo / Module VPD

Table 179 Allowed Addresses for InfiniBand™ Modules

Slave Address	8-bit Hex notation	Function
1010 011	A6h	Module-specific use. Typically used for additional SEEPROMs.
1010 110	ACh	Module-specific use. Typically used for additional SEEPROMs.
1010 111	AEh	Module-specific use. Typically used for additional SEEPROMs.
1011 110 - 1011 111	BCh-BEh	Module-specific use.
1100 001	C2h	SMBus Device Default. Module Devices that use this address must do so per the SMBus 2.0 specification, and not use assigned addresses that conflict with the InfiniBand specification.
1101 110 - 1101 111	DCh-DEh	Module-specific use.
1110 000	E0h	MME Function registers
1110 001	E2h	MME IBML2IB registers
1110 100	E4h	MME Command
1110 011	E6h	Module-specific use. Often used for PCA9544-style I^2C bus multiplexer.

A2.2.2 IB-ML PROTOCOL DIFFERENCES

The IB-ML supports the SMBus protocol. The differences from I^2C are the same as above. Specific addresses reserved for IB-ML devices and the Baseboard Management Agent are located in Table 121 Module IB-ML Slave Addresses on page 496 and Table 130 Chassis IB-ML Slave Addresses on page 546.

A2.3 ARCHITECTURAL DIFFERENCES

This section describes the architectural protocol differences between the IB-ML, SMBus and the I^2C bus.

A2.3.1 SMBUS AND I^2C BUS ARCHITECTURAL DIFFERENCES

The architectural differences between the SMBus and the I^2C bus are:

1) The SMBus may have an optional SMBUS# signal used in power-down mode. This signal is an output from the system management device and is used to signal the state of system suspend mode.

2) The SMBus may have an optional SMBALERT# signal used for a slave-only device to signal the bus master that it wishes to communicate.

3) Response to the SMBALERT# signal generates a 7 or 10-bit alert response using the defined alert response address.

A2.3.2 IB-ML ARCHITECTURAL DIFFERENCES

The architectural differences between the IB-ML and the SMBus/I²C bus are:

1) The IB-ML does not support the SMBus and I²C bus sideband signals described above.

2) The IB-ML is a point to point link between an IB Module and a Chassis. This eliminates address conflict issues between Modules, but support for multi-master arbitration is still required. This is because a CME and module can both attempt to initiate a transaction at the same time. CMEs, MMEs, and any other master on IB-ML must support multi-master arbitration.

3) IB-ML has its own electrical drive specifications. Refer to Table 180: Electrical Differences (IB-ML, SMBus and I2C Bus).

A2.4 ELECTRICAL AND TIMING

This section describes the IB-ML, SMBus and I²C electrical and timing differences. Since the IB-ML, SMBus and I²C bus are open-collector technologies, calculating the proper signal pullup resistor values will also be covered.

A2.4.1 IB-ML, SMBUS AND I²C BUS ELECTRICAL DIFFERENCES

The main difference between the IB-ML, SMBus and I²C Bus electrical characteristics is that the IB-ML/SMBus uses fixed levels and the I²C bus uses levels relative to the Vcc of the bus device. In addition, IB-ML specifies greater drive than the SMBus 1.1 specification does. Table 180 lists the electrical differences between the IB-ML, SMBus and the I²C bus.

Table 180 Electrical Differences (IB-ML, SMBus and I²C Bus)

Parameter	IB-ML		SMBus		I²C Bus	
	Min.	Max.	Min.	Max.	Min.	Max.
Vil	-0.5V	0.8V	-0.5V	0.6V	-0.5V	0.3*Vcc
Vih	2.1V	3.63V	1.4V	5.5V	0.7*Vcc	
Vol		0.4V @ 4.0mA		0.4V @Ipullup Min.	0	0.4V @ 3.0mA (0.6V @ 6.0mA fast)
Iil		+/- 10µA		+/- 10µA		+/- 10µA
Ipullup		4.0mA[a]	100µA[a]	350µA[a]		3.0mA (6.0mA fast)

a. The value includes both the current through the pullup resistor and current from all bus agents. In this context the parameter Ipullup is equivalent to Iol.

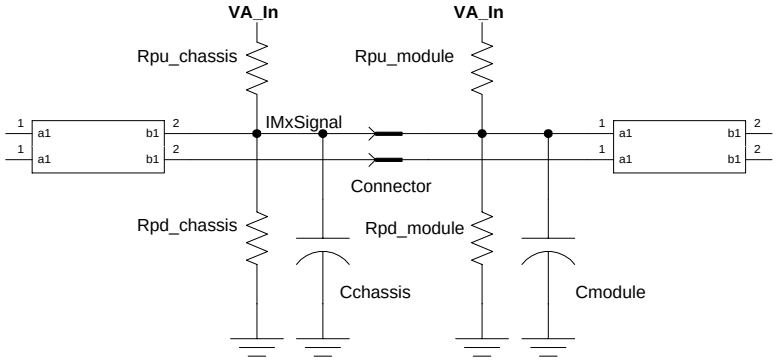
The value for Ipullup listed in the SMBus specification has been shown to be difficult to meet when larger numbers of agents are connected to the bus. For this reason, some SMBus devices may differ from the Ipullup parameter listed in Table 180. For example a Ipullup of 3.0mA Max will guarantee a Vol of 0.4V @ Ipullup of 3.0mA. This allows a larger number of devices to be connected to the bus without violating the operating levels.

A2.4.2 BUS AND AGENT VCC VOLTAGE LEVELS

The I²C bus specifies that device input levels are in most cases dependent on Vcc. Note the voltage to which Vcc of each I²C bus agent is connected and to what voltage the bus signal pullup resistors are connected. Mixing these voltages in the wrong way could cause improper input high levels.

The IB Modules auxiliary voltage is used as the supply for the termination of the IB-ML. A module may use a local 3.3V regulator from **VA_In** for the termination supply or use the VA_In supply with a resistive divider termination.

Figure 214 IB-ML Termination to VA_In



In some cases, bus agents may only have 3.3V tolerant input levels and therefore the bus signals must only be pulled up to a Vcc of 3.3V. Any other bus agents will have to be connected to the proper voltage levels so that their input levels are compliant with that of the bus. For a reference implementation to provide the necessary translation see Section A2.4.7.

A2.4.3 IB-ML, SMBUS AND I²C TIMING DIFFERENCES

The IB-ML/SMBus and I²C Bus timing differences are mostly concerned with time-out and clock stretching protocols supported by the IB-ML/SMBus. These timing differences are listed in Table 181. Note, the designer should still review the actual specifications to get a full comparison of the different bus timings.

Table 181 Timing Differences (IB-ML, SMBus and I²C Bus)

Parameter	IB-ML		SMBus		I²C Bus	
	Min.	Max.	Min.	Max.	Min.	Max.
Bus Frequency	10 KHz	100KHz	10 KHz	100KHz	0	100KHz (400KHz Fast)
Clock Low Time-out	25mS	35mS	25mS	35mS	N/A	N/A
Clock High period (THIGH)	4 us	50 uS	4 us	50 uS[a]	4 us	-
Clock Low Extend (Slave)		25mS		25mS	N/A	N/A
Clock Low Extend (Master)		10mS		10mS	N/A	N/A

a. This parameter provides a simple guaranteed method for masters to detect bus idle conditions. A master can assume that the bus is free if it detects that the clock and data signals have been high for greater than THIGH,MAX.

A2.4.4 ARBITRATION RETRIES

If software requests an IB-to-IBML operation, an MME can return an error completion status if it tries to arbitrate for the IB-ML and fails. Software would then need to retry the operation by sending another IB-to-IBML command. It is recommended that a IB-ML master implementation (MME or CME) automatically retry arbitrating for the bus at the next bus-free opportunity at least three times and no more than ten times.

A2.4.5 ARBITRATION DELAY

IB-ML, SMBus, and I^2C arbitration share the characteristic that the master that transmits the first 0' bit in its message will win arbitration. This means that devices with slave addresses with zeros in lower positions will always win arbitration over devices that have zeros in higher positions. For example, a transaction from the CME to the MME Function registers (1110 000) will always win arbitration over a simultaneous transaction from the MME to the CME (1110 100).

In general, delays and randomness in timing between implementations, and the unlikelihood of arbitration collisions usually prevent one device from locking out another. However, automated back-to-back transfers could cause one device to lock another off the bus until the transfers are completed. It is recommended that devices limit the number of back-to-back transactions to three, and then institute a delay of at least 100 us (a little over 1 byte interval) before again trying to arbitrate for the bus. A random factor in the delay interval will further help avoid the potential for 'lock out' occurrences. Devices that solely drive transactions in response to a software request (e.g. the MME's IBML-to-IB functionality) generally do not need to implement such limits in hardware, since system software can institute appropriate delays if necessary.

A2.4.6 BUS PULLUP RESISTOR CALCULATIONS

Due to the architecture of both the IB-ML, SMBus and the I^2C Bus, all agents on the bus must have open collector (drain) outputs. Because of the nature of these outputs, both the bus clock and data lines must be pulled up to some Vcc value. Both the SMBus and the I^2C bus specifications require only one pullup resistor per bus signal. The IB-ML specification require two pullup resistors one on the IB Module and one on the chassis. As discussed in Section A2.4.2, the voltage used for the pullup resistors on the bus signals depends on the nature of all of the bus agents.

A2.4.6.1 GENERAL BUS PULLUP RESISTOR CALCULATIONS

There are three cases that must be considered when calculating the pullup resistor values for either an SMBus or I^2C Bus:

- Vih level at rated current.

- Vil level at rated current.
- Rise and fall time conditions.

Use the above three cases when calculating the bus pullup values. The DC values of Vih and Vil will be used to define a solution space and the rise and fall time requirements will select a specific value in that solution space. We recommend this approach to provide a direct method of resistor value calculation that will satisfy all applicable conditions.

The steps to calculate the solution space are listed in the “Determine Maximum Pullup Resistor Value” and “Determine Minimum Pullup Resistor Value” procedures in the following sections.

A2.4.6.1.1 PROCEDURE 1. DETERMINE MAXIMUM PULLUP RESISTOR VALUE

To find the maximum pullup resistor value to use on clock and data lines, follow these steps:

1) Identify the minimum value of the *Vcc* rail *(Vccmin*) where the bus signal pullup resistors will be connected.
2) Identify the minimum *Vih* value for each bus agent when powered by the minimum *Vcc* value found in step 1. Use the highest of these *Vih* values *(Vihmm)* for the calculations. Choosing the highest minimum *Vih* value will guarantee all other *Vih* conditions are met.
3) Choose a desired noise margin *(NMmin*) for the logic high condition on the bus. This value is typically 0.1V - 0.2V above the maximum value chosen for *Vih*. It. is recommended that the first choice for the noise margin be 0.2V. This will be the minimum noise margin for the system when Vcc is at a minimum value. As the system *Vcc* increases to typical and maximum value, the noise margin will also increase.
4) Find the total maximum current *(Iihmax*) sunk by all agents for the input high condition.
5) Calculate the maximum value *(Rpmax*) for the bus pullup resistor from the following equation:

   Rpmax = (Vccmin - (Vihmm + NMmin))/ Iihmax **- Equation 1**

6) This result is the upper limit (maximum value) of the pullup resistor solution space.

A2.4.6.1.2 PROCEDURE 2. DETERMINE MINIMUM PULLUP RESISTOR VALUE

To find the minimum pullup resistor value to use on clock and data lines, follow these steps:

1) Identify the maximum value of the *Vcc* rail *(Vccmax*) where the bus pullup resistors will be connected.
2) Identify the lowest *Vol* value *(Volmin*) for all bus agents.

3) Identify the maximum current sunk *(Iolmax*) by this bus agent at the *Volmax* value.

4) Calculate the minimum value *(Rpmin*) for the bus pullup resistor from the following equation:

$$Rpmin = (Vccmax - Volmin) / Iolmax$$ **- Equation 2**

5) This result is the lower limit (minimum value) of the pullup resistor solution space.

Now we have a minimum and maximum resistance value that our pullup resistor can be while satisfying the DC specifications of our bus. This is our solution space for choosing a resistor value that will satisfy our rise and fall times.

A2.4.6.1.3 PROCEDURE 3. PERFORM BUS SIGNAL RISE AND FALL TIME CALCULATIONS

The value found from the rise and fall time calculations, together with the solution space defined from the bus DC bus values, will determine the proper value for the bus pullup resistors. The solution from the rise and fall time calculations should fall within the DC solution space to satisfy all bus requirements.

If a value of pullup resistor is found that satisfies our rise and fall time requirements but is not in our DC solution space, parameters in the calculations or the bus design itself may need to be modified.

A good rule of thumb for the pullup value is $Rp < Trise / 2 * Cbus$. Where *Trise* is the maximum allowable rise time minus some margin and *Cbus* is the total capacitance on the bus.

In this design guideline, a more precise calculation will be used to calculate the pullup resistor values. This calculation will also help verify the rule of thumb calculation. The equation that will be used to calculate signal rise time is one for charging capacitors.

The general form of this equation is:

$$Vc = Vcc + [Vo - Vcc] e^{-t/RC}$$ **- Equation 3**

where *R* is pullup Rp, *C* is the total bus capacitance, *t* is the rise time, *Vo* is the initial voltage on the bus capacitance, and *Vc* is the voltage across the bus capacitance at any given time (in this case Vil and Vih values).

Solving Equation 3 for the rise time t yields:

$$t = -RCln[(Vc - Vcc)/(Vo - Vcc)]$$ **- Equation 4**

A2.4.6.2 RISE TIME CALCULATION

Use the following steps to calculate the bus rise time:

1) The maximum rise time for both the SMBus and the I^2C Bus is defined to be 1.0µs.

2) Choose a rise time margin such that the desired rise time is less than the maximum allowed rise time of 1.0µs. Usually this is around 50 - 100nS.

3) Calculate the total capacitance of the bus. This includes the capacitance of the bus and of all bus agents.

4) For a rising edge, 0 to 1 transition, assume an initial starting voltage *(Vo*) for the logic 0 state. Assuming the bus has settled out to a ground state, this value would be approximately 0V. For worst case bus calculations, assume an initial voltage of 0V.

5) Determine the minimum value of *Vcc (Vccmin*) in the system where the pullup resistor will be connected.

6) Determine the maximum value for *Vih (Vihmm*) for the minimum value of *Vcc* found in step 5.

7) Solve Equation 5 for the pullup resistor *R* and substitute the known values for the bus parameters. This will yield an equation for the value of the pullup resistor needed.

$$Rp = -t / (C * ln [(Vihmm - Vccmin)/ (Vo - Vccmin)])$$ **- Equation 5**

If the value obtained from Equation 5 is not in the solution space found in the “General Bus Pullup Resistor Calculations” of Section A2.4.6.1 for the DC levels, one of two steps must be taken. If the value from the above rise time equation is greater than the maximum limit of the DC solution space found in Equation 4, the resistor value may be reduced to a value within the solution space.

Lowering the resistor value will satisfy the DC solution space and will not violate the rise time specification. A lower resistance value than the one found in Equation 5 will not violate the maximum rise time parameter but will in fact make the rise time less. This can be seen from Equation 4. Assuming no parameters change except for the resistance R, a lower resistance value will yield a lower rise time t.

If the result of Equation 5 is lower than the minimum value for the DC solution space found in Equation 2, either the factors influencing the minimum value for the DC solution space must be changed or the factors influencing the value from the rise time calculation must be changed. In the case of the minimum value in the DC solution space, this value must be lowered.

From Equation 2 we see that the resistance value may be lowered if the product of *Vccmax - Volmin* is decreased and/or *Iolmax* is increased. Since *Iolmax* and *Volmin* are a function of the bus agent in question, the only so-

lution without changing the driving bus agent is to select a lower *Vccmax* of the system.

In the case of the value for maximum rise time, this value must be raised. From Equation 5 we can see that in order to make the value for Rp greater we must increase the desired maximum rise time, decrease the bus capacitance, decrease the *Vihmax – Vccmin* product, and/or increase the *Vo – Vccmin* product.

Generally, the bus capacitance is a function of the bus agents and cannot be changed unless the number of bus agents is changed or the characteristics of the bus agents are changed. The products of *Vihmax – Vccmin* and *Vo – Vccmin* can be changed appropriately by decreasing the value of *Vo*. The fact that *Vihmax* is a function of *Vccmin* means changing one will change the other so that no net change will occur when the division of the numerator and denominator is performed. The most easily changed parameter is the desired rise time. If some margin has been added to the rise time, the amount of margin can be decreased so that the desired maximum rise time will increase.

The above cases are listed for completeness. In general, the resistor from the rise time calculation will either be in the DC solution space or will be greater than the DC solution space. In these cases little or no work will be needed to adjust for a resistor value that is acceptable for proper bus operation.

A2.4.6.3 FALL TIME CALCULATIONS

The calculations for fall time involve equations that require knowledge of the output driver, as well as several integrations over the active regions of the output driver. These calculations are beyond the scope of this document and need not be used here.

Instead, we can use the fact that the SMBus and I^2C Bus specifications require a maximum fall time of 300ns. We can also use the fact that these same specifications require that the output driver of a compliant agent must guarantee a maximum fall time of 250ns over a bus capacitance of 10 - 400pF with up to 3.0mA (6.0mA in the fast I^2C case) of sink current.

Implementation Note

Although the SMBus specifies a maximum fall time of 300ns, it makes no specification of the output driver guaranteed fall time. Only the I^2C bus specifies the 250ns guaranteed driver fall time, as mentioned above.

From the previous paragraph we can assume our bus fall times to be within the 300ns specification if three criteria are matched:

- All agents that are capable of driving the bus are I^2C compliant to guarantee falltimes of 250ns maximum.
- Our bus has a total capacitance of between 10 and 400pF.
- The total current that has to be sunk by any output driver in the low state is 3.0mA or less (6.0ma or less in the fast I^2C case).

For the total bus current that has to be sunk by any driver in the low state, you can use this equation:

I total = ((Vcc-Volmin)/Rp) + Iil (from all bus agents) **- Equation 6**

Note that the maximum pullup resistor current occurs at *Volmax* and *Vccmax*.

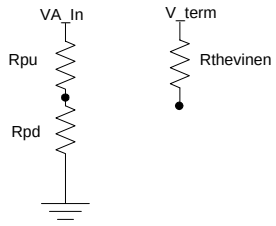
A2.4.6.4 IB-ML BUS PULLUP RESISTOR CALCULATIONS

The IB-ML termination target voltage is 3.3V nominal. Implementations may utilize a local 3.3V regulator (e.g. +-5%) from **VA_In** or may use a resistor divider from **VA_In**. In order to maintain the 3.63V VihMAX and include consideration of the tolerance of **VA_In** results in a termination voltage of:

VTermMAX: 3.63V (**VA_In** tolerance of +5%)

VTermMIN: 3.063V (**VA_In** tolerance of -8%).

Figure 215 Thevinen Equivalent Termination



If the IB-ML termination solution is determined for the Thevinen equivalent circuit, it can be implemented as the resistor divider shown above. Either solution is valid.

The following is a MathCAD analysis for the IB-ML resistor termination using **VA_In** as the termination voltage.

Given:

For the IMxClk, IMxDat and IMxPRst bi-directional signals:

1) The chassis and IBML input voltage requirements are:
   Vil = -0.5V Min and 0.8V Max
   Vih = 2.1V Min and 3.63V Max

2) The Vaux specification:
   VA_in = 5.5V Max and 4.6V Min

3) The Module IBML capacitance is 200pF Max

4) The Chassis IBML capacitance is 100pF Max

5) The Module leakage current is 80uA

6) The Chassis leakage is 40uA

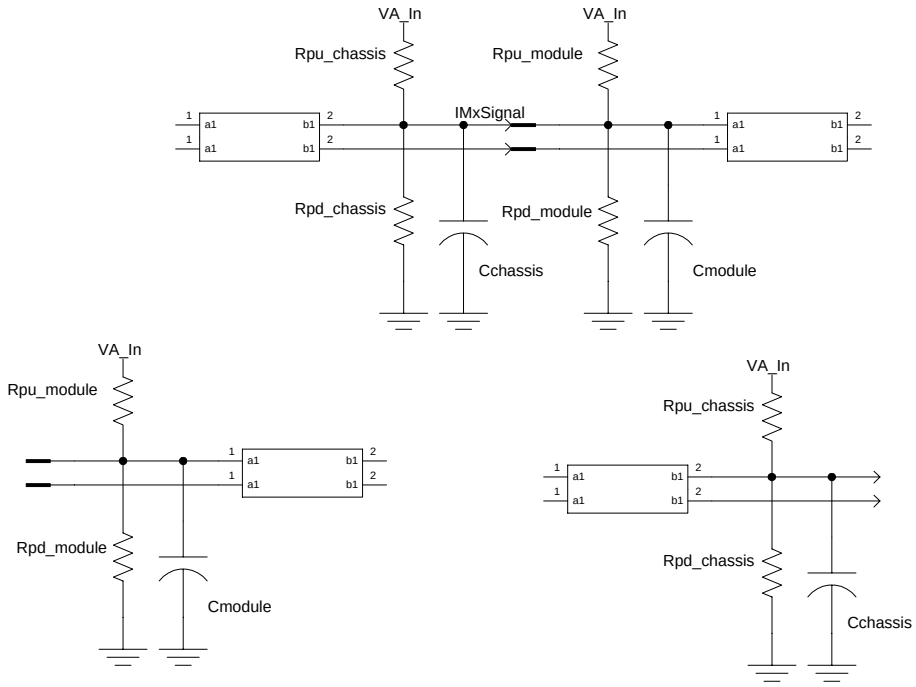
Requirement:

Rpu and Rpd must provide valid Vih/Vil levels and AC timing under all the following conditio

1) The resistor tolerance range
2) The VA_In range
3) The leakage range
4) The capacitance range

Three configurations are to be studied:

1) The IB Module alone
2) The Chassis alone
3) The IB Module and chassis together



Variables:

VA_In_max:= 5.5 VA_In_min:= 4.6 VA_In_nom:= 5.0

Vterm_max:= 3.63 Vterm_min:= 3.036 Vterm_nom:= 3.3

Vih_min:= 2.1 Vih_max:= 3.63

Vil_min:= –0.5 Vil_max:= 0.8

Vol_min:= 0 Vol_max:= 0.4

$Iol\_max := 4 \cdot 10^{-3}$

NM_min:= 0.2

$Ileakage\_max := 120 \cdot 10^{-6}$

Ileakage_min:= 0

$Ichassis\_max := 40 \cdot 10^{-6}$

$Imodule\_max := 80 \cdot 10^{-6}$

$Cchassis := 100 \cdot 10^{-12}$

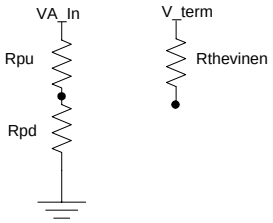
$Cmodule := 200 \cdot 10^{-12}$

Ctotal := Cchassis+ Cmodule

$Trise := 1000 \cdot 10^{-9}$

$Tmargin := 100 \cdot 10^{-9}$

T := Trise– Tmargin



1) Vterm_nom target is 3.3V

2) Vterm_max target is 3.63V which uses the VA_In tolerance of +5%

3) Vterm_min target is 3.036 which uses the VA_In tolerance of -8%

Step 1: Determine the Maximum Pullup Value based on DC parameters:

Case 1:The chassis alone

$$Rpumax\_chassis := \frac{[Vterm\_min - (Vih\_min + NM\_min)]}{Ichassis\_max} \qquad Rpumax\_chassis = 1.84 \times 10^{4}$$

Case 2: The module alone

$$Rpumax\_module := \frac{[Vterm\_min - (Vih\_min + NM\_min)]}{Imodule\_max} \qquad Rpumax\_module = 9.2 \times 10^{3}$$

Case 3: The module and chassis together

$$Rpumax := \frac{[Vterm\_min - (Vih\_min + NM\_min)]}{Ileakage\_max} \qquad Rpumax = 6.133 \times 10^{3}$$

Step 2: Determine the Minimum Pullup Value based on DC parameters:

Case 1, Case 2, Case 3: Are all the same

$$Rpumin := \frac{(Vterm\_max - Vol\_min)}{Iol\_max} \qquad Rpumin = 907.5$$

Step 3: Determine the Maximum Pullup Value based on the Rise Time requirements:

Case 1: The chassis alone

$$Rpu\_chassis := \frac{-T}{\left[Cchassis \cdot \left(\ln\left(\frac{Vih\_min - Vterm\_min}{Vol\_min - Vterm\_min}\right)\right)\right]} \qquad Rpu\_chassis = 7.649 \times 10^{3}$$

Case 2: The module alone

$$Rpu\_module := \frac{-T}{\left[Cmodule \cdot \left(\ln\left(\frac{Vih\_min - Vterm\_min}{Vol\_min - Vterm\_min}\right)\right)\right]} \qquad Rpu\_module = 3.824 \times 10^{3}$$

Case 3: The module and chassis together

$$Rpu := \frac{-T}{\left[Ctotal \cdot \left(\ln\left(\frac{Vih\_min - Vterm\_min}{Vol\_min - Vterm\_min}\right)\right)\right]} \qquad Rpu = 2.55 \times 10^{3}$$

Choosing a nominal termination value of 907.5 Ohms:

This is the minimum pullup value based on the DC requirements.
We will split the Thevinen equivalent resistor into pullup/pulldown pairs.

Rthevinen := 907.5

Given

$$Vterm\_nom = \frac{Rpd \cdot VA\_In\_nom}{Rpu + Rpd}$$

$$Rthevinen = \frac{Rpu \cdot Rpd}{Rpu + Rpd}$$

$$Find(Rpu, Rpd) \rightarrow \begin{pmatrix} 1375. \\ 2669.1176470588235294 \end{pmatrix}$$

Rpu := 1375. Rpd := 2669.1176470588235294

Splitting the resistors between the module and the chassis (using the Chassis/IB Mod capacitance ratio) while maintaining the equivalent parallel resistance gives:

$Rpu\_mod\_min := Rpu \cdot \frac{3}{2}$ $Rpu\_mod\_min = 2.063 \times 10^3$

$Rpu\_chas\_min := Rpu\_mod\_min \cdot 2$ $Rpu\_chas\_min = 4.125 \times 10^3$

$Rpd\_mod\_min := Rpd \cdot \frac{3}{2}$ $Rpd\_mod\_min = 4.004 \times 10^3$

$Rpd\_chas\_min := Rpd\_mod\_min \cdot 2$ $Rpd\_chas\_min = 8.007 \times 10^3$

Choosing a nominal termination value of 2250 Ohms:

This is the maximum pullup value based on the rise time requirements.
We will split the Thevinen equivalent resistor into pullup/pulldown pairs.

Rthevinen := 2250

Given

$$Vterm\_nom = \frac{Rpd \cdot VA\_In\_nom}{Rpu + Rpd}$$

$$Rthevinen = \frac{Rpu \cdot Rpd}{Rpu + Rpd}$$

$$Find(Rpu, Rpd) \rightarrow \begin{pmatrix} 3409.0909090909090909 \\ 6617.6470588235294118 \end{pmatrix}$$

Rpu := 3409.0909090909090909 Rpd := 6617.6470588235294118

Splitting the resistors between the module and the chassis (using the Chassis/IB Mod capacitance ratio) while maintaining the equivalent parallel resistance gives:

$Rpu\_mod\_max := Rpu \cdot \frac{3}{2}$ $Rpu\_mod\_max = 5.114 \times 10^3$

$Rpu\_chas\_max := Rpu\_mod\_max \cdot 2$ $Rpu\_chas\_max = 1.023 \times 10^4$

$Rpd\_mod\_max := Rpd \cdot \frac{3}{2}$ $Rpd\_mod\_max = 9.926 \times 10^3$

$Rpd\_chas\_max := Rpd\_mod\_max \cdot 2$ $Rpd\_chas\_max = 1.985 \times 10^4$

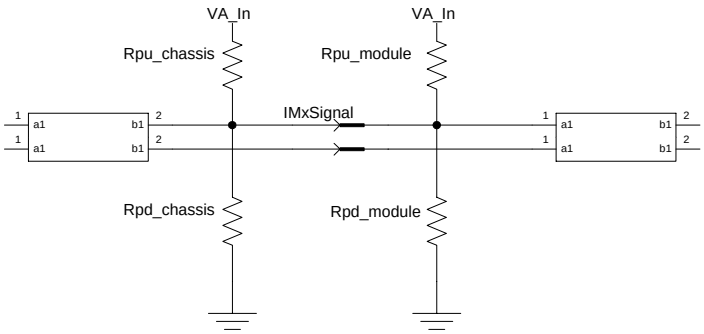
Summary: The following resistor value pairs will meet the DC voltage levels required over the VA_In range as well as the IB-ML AC timing requirements. The resistor selected should include the appropriate tolerances.

$Rpu\_mod\_max = 5.114 \times 10^3$

$Rpu\_mod\_min = 2.063 \times 10^3$

$Rpd\_mod\_max = 9.926 \times 10^3$

$Rpd\_mod\_min = 4.004 \times 10^3$



$Rpu\_chas\_max = 1.023 \times 10^4$

$Rpu\_chas\_min = 4.125 \times 10^3$

$Rpd\_chas\_max = 1.985 \times 10^4$

$Rpd\_chas\_min = 8.007 \times 10^3$

NOTE The pullup/pulldown resitors should not be selected independantly based on the range, but should be selected as a pair.

A2.4.6.5 IB-ML PULLUP RESISTOR CALCULATION EXAMPLE

The following is an example of a pullup resistor calculation for the IB-ML using I^2C slave agents and a single SMBus host.

This example assumes six I^2C compliant agents are on the IB Module, one I^2C compliant agent and one SMBus master controller on the chassis.

An example SMBus master controller spec:

- Vil = 0.6V max.
- Vih = 1.4V min.
- Iil, Iih = +/- 10µA
- Vol = 0.4V max. @ rated Iol
- Iol = 4.0mA max.

- Ci/o = 12pF max.

Implementation Note

The SMBus master controller inputs are assumed to only be 3.3V tolerant so they bus up to a 3.3V supply. In this case Vcc3 is specified at 3.3V + 5% -8% to reflect the tolerance on **VA_In**.

I^2C compliant bus agents specs.

- Vih = 0.7*Vcc with Vcc = 3.036 - 3.63V: Vih = 2.12V min., 2.54V max.
- Vil = 0.3*Vcc with Vcc = 3.036 - 3.63V: Vil = 0.91V min., 1.09V max.
- Iil, Iih = +/- 10µA
- Vol = 0.4V max. @ rated Iol
- Iol = 4.0mA max.
- Ci/o = 10pF max.

Implementation Note

All of the bus agents have the same specs when they are driving the bus low. This greatly simplifies our calculations since we only need to perform one set of calculations for these common output characteristics.

First we will calculate the pullup resistor solution space from the specified DC parameters for the various agents. In this example we will select a minimum noise margin *(NMmin*) of 0.2V. We see the largest minimum value for *Vih* on any bus agent is 2.1V at minimum *Vcc*. We also note that *Iihmax* for the seven agents receiving is 7*10µA.

From our equation for maximum pullup resistance:

- Rpmax = (Vccmin - (Vihmax + NMmin)) / Iihmax
- *Rpmax* = (3.036 - (2.1 + 0.2)) / $70x10^{-6}$
- *Rpmax* = 10.514 K

Next we need to calculate the pullup resistor minimum value. The minimum *Vol* value for any agent driving the bus is 0V at an *Iol* of 4.0mA Max. Max. *Vcc* has been determined to be 3.63V. From our equation for *Rpmin*:

- Rpmin = (Vccmax - Volmin) / Iolmax
- *Rpmin* = (3.63 - 0) / $4x10^{-3}$

- *Rpmin* = 907.5 K

We now have our solution space and may calculate our specific pullup value based on desired rise and fall times.

We have a maximum rise time for both the SMBus and I^2C bus of 1.0µs. We now choose a margin for rise time, in this case 100nS. Thus our maximum rise time is 900nS.

Note that we now have eight input loads on the bus at any one time during a logic high condition. Seven of these loads are the same, 10pF, and the master controller is 12pF. Thus, we have a total bus capacitance of 82pF.

Voltage conditions are *Vccmax* = 3.63V, *Vihmax+NM* = 2.7V, and our assumed worst case *Vol* is approximately 0V.

Substituting back into Equation 5 (our Rise Time Calculation), we may now calculate the pullup value based on the above conditions:

- Rp = -t / (C * ln[(Vihmax+NM - Vccmax) / (Vo - Vccmax)])
- $Rp = -900x10^{-9} / (82x10^{-12} * \ln[(2.7 - 3.036) / (0 - 3.036)])$
- *Rp* = 4.986 K

We see from our result that the resistor value falls in the pullup resistor solution space calculated from our DC parameters. We may choose this value or choose a different resistor value, as long as it is in the calculated solution space and is less than the value derived from the above rise-time calculation.

Our decision to choose a different resistor value than the one calculated may be the desire to use a more standard resistor value or the desire to use a value that is already used elsewhere in our design. This decision is left to the designer.

In this example, we will choose a more standard resistor value for our pullup. Let this resistor be 4.7K. As mentioned previously, we may adjust the pullup value to be less than the value derived from our rise-time calculation, without causing the bus to function improperly. We see from Equation 4 that if the resistance *Rp* is decreased, the rise time of the bus will decrease accordingly. This change will not violate the maximum rise time but will instead add more margin to our rise time.

We can verify this by calculating the rise time with our chosen value of pullup resistor from the equation:

- t = -R * C * ln[(Vohmax - Vccmax) / (Volmax - Vccmax)]
- $t = -4.7x10^{3} * 82x10^{-12} * \ln[(2.7 - 3.036) / (0 - 3.036)]$

- t = 848 ns

To verify that our choice does not violate the bus fall time we must check that we do not violate the specification for maximum current sunk by an output driver. We note that *Itotal* delivered to the output driver is:

- Itotal = ((Vccmax - Volmin) / Rp) + Iil (from all bus agents)
- *Itotal* = ((3.63 - 0) / 4.7×10^{3}) + 70×10^{-6}
- *Itotal* = 784µA

This is significantly less than the rated current of 3.0mA, so our value for *Rp* is satisfactory.

While this example shows that a stronger pullup is not required for a design that is not fully loaded, an IB Module or chassis should select a termination value that accounts for the maximum load on either end. The purpose of this example is to explain the concepts and the design concerns that should be addressed in using I^2C slave agents or SMBus devices in an IB-ML design.

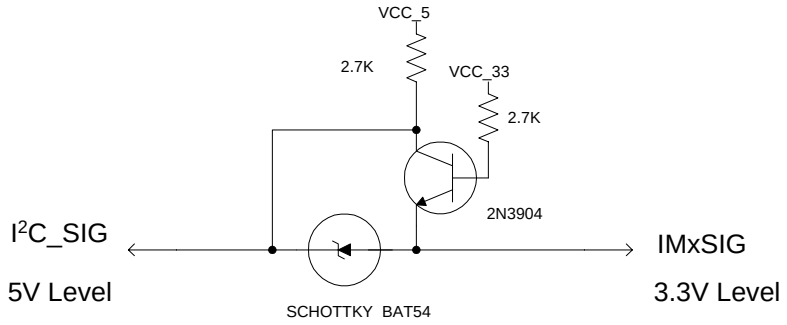
A2.4.7 EXAMPLE 5V TO 3.3V TRANSLATION CIRCUIT

The following is an example circuit to translate from 5V to 3.3V signaling.

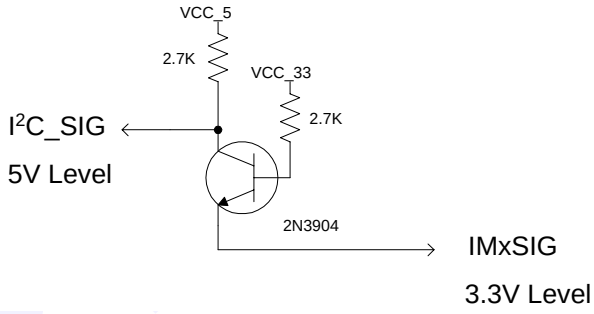
IBTA

Figure 216 IB-ML to 5V I²C Translation Circuit

Bi-directional 5V to 3.3V Level Shifter



Uni-directional 5V to 3.3V Level Shifter



A2.4.8 EXAMPLE ISOLATION LOGIC

The following is an example circuit for isolating a device from the IB-ML:

To be added in a future revision.

ANNEX A3: HARDWARE MANAGEMENT EXAMPLES

A3.1 INTRODUCTION

The purpose of this annex is to provide representative implementation examples of various Hardware Management feature described in Chapter 13: Hardware Management.

A3.2 MODULE AND CHASSIS MANAGEABILITY COMBINATIONS

This section outlines InfiniBand Module management and proxies in different manageable Chassis types.

A3.2.1 MODULES BY CHASSIS TYPE

Table 182 shows the manageability characteristics of modules in different Chassis types.

Table 182 Module Management by Chassis Type

InfiniBand Module	Passively Managed Chassis	Actively Managed Chassis	Unmanaged Chassis
xCA Module			
Fully Managed	In-band MADs targeted and responded Out-of-band access (chassis->Module) available, but not utilized	In-band MADs targeted and responded Out-of-band access (chassis->Module) supported	In-band MADs targeted and responded
Proxy Managed	Not Defined	Not Defined	Not Defined
Unmanaged	Not Defined	Not Defined	Not Defined
Repeater Module			
Fully Managed	Not Defined	Not Defined	Not Defined
Proxy Managed	In-band requires proxy to be manageable Out-of-band access (chassis->Module) available, but not utilized	In-band requires proxy to be manageable Out-of-band access (chassis->Module) supported	Not Defined
Unmanaged	Not Defined	Not Defined	Not Defined

15.3.50.1 PROXIES BY CHASSIS TYPES

Since TCAs, Switches, and HCAs are protocol-aware and addressable, they can act as proxies for IB to IB-ML transactions. Table 183 shows the possible protocol-aware proxies for the defined chassis types.

Table 183 Proxy Capability per Chassis Type

Protocol-aware Unit	Unmanaged Chassis	Passively Managed Chassis	Actively Managed Chassis	
			SEEPROM Interface	Split Transactions
Fully Managed TCA Module as a Proxy	Not Applicable No proxy necessary for chassis	Capable SEEPROM Write-Read support is required	Capable SEEPROM Write-Read support is required	Conditional IBML2Ib support is required; only a TCA supporting IBML2Ib can be a chassis proxy in a Chassis that uses slots as chassis portals.
Switch as a Proxy	Not Applicable No proxy necessary for chassis	Capable SEEPROM Write-Read support is required	Capable SEEPROM Write-Read support is required	Conditional IBML2Ib support is optional; only a Switches supporting IBML2Ib can be a chassis proxy in a Chassis that uses the Switch as a chassis portal.
Repeater Module	Not Capable (not addressable)	Not Capable (not addressable)	Not Capable (not addressable)	Not Capable (not addressable)
HCA Unit as a Proxy	Not Applicable No proxy necessary for chassis	Capable SEEPROM Write-Read support is required	Capable SEEPROM Write-Read support is required	Conditional As the HCA and the Chassis in which it resides are both choices of the system implementer, the HCA may be chosen as the proxy and thus would need IB-ML-IB support. Similarly, the system implementer may chose to use a switch or embedded TCA for the proxy.

A3.3 MANAGED CHASSIS USING SLOT SWITCH AND IB-ML SELECTOR PROXY

Figure 217 shows a Managed Chassis including an InfiniBand Switch. This Switch may be on the backplane or in a proprietary slot.

As Figure 217 shows, from the Baseboard Manager's and the module's

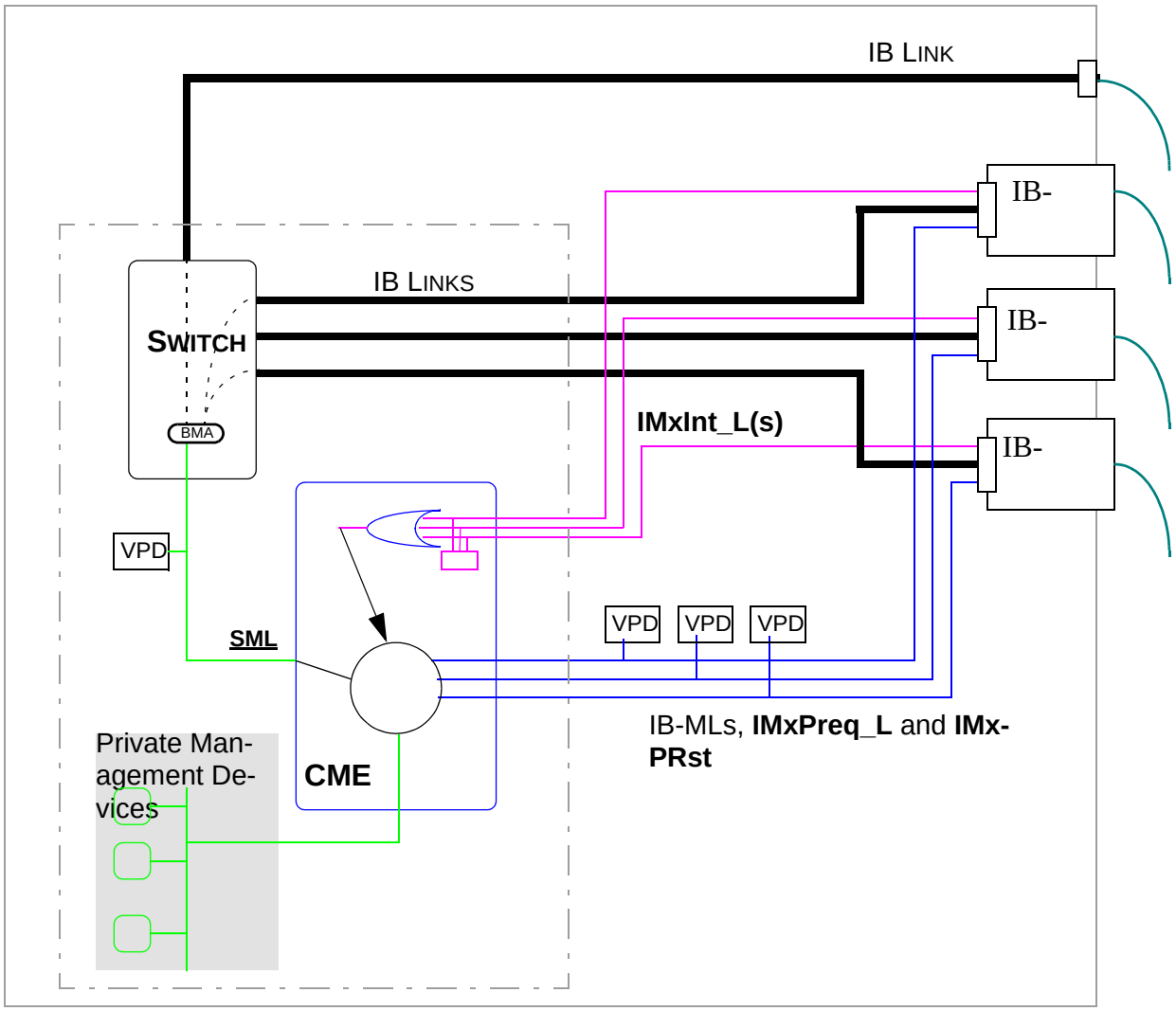


Figure 217 Example Managed Chassis with Switch

points of view, there exists one Chassis VPD per module. The chassis vendor may chose to combine these VPDs in one device.

Figure 217 shows a Chassis Management Entity (CME) which may or may not include a processor. The CME is merely an IB-ML De-Mux in its simplest form. The CME may provide a proxy for access to the modules' IB-MLs. See Section 13.5.2.6, "IB-ML Selector Proxy," on page 548. A proxy is necessary for managing Repeater Modules.

In-band access to the Switch VPD via Switch Management Link (SML). The SML may be implemented as an IB-ML. The Baseboard Manager

running on the Fabric may access the Switch VPD in-band via any port on the Switch-- including the ones connected to the Module Repeaters or TCAs.

The CME includes a Selector as a mechanism for selecting individual IB-MLs in its SML address space.

Once the Baseboard Manager detects the capabilities of the CME via the IB Switch VPD, it may select any Management Link using the Selector. The Switch VPD includes information about the number of IB-MLs connected to the CME.

Once selected, any module's IB-ML is available to the Baseboard Manager as though it was connected directly to the SML. The CME's Selector optionally appears in every IB-ML's address space so that the CME may act as a proxy for inter-IB-ML communication.

To manage Repeater Boards, the Baseboard Manager uses either the Switch or a TCA's access to CME. The CME has access to all modules' IB-MLs; otherwise, accessing Repeater modules' VPD is limited to the ones connected to the CME.

From the Switch VPD and the Fabric Manager, the Baseboard Manager knows through which port it is entering the Switch. If the module through which the Baseboard Manager is accessing the Switch has only a Repeater on it, the Baseboard Manager uses that Switch Port Number to select the IB-ML of the Repeater module.

An module may assert an interrupt. The CME may latch all of the modules' Interrupts in a register, may functionally OR them and may send an IB-ML style message to the Switch. The Switch, in turn, generates a Trap on the Fabric. The implementation of IB-ML to BMTrap is an optional Chassis feature.

TCAs are protocol-aware; they provide direct access to TCA Module's IB-ML. Although the IB-ML of a TCA Module is available via the TCA, the Baseboard Manager may choose to access its IB-ML in the same manner it accesses that of a Repeater module.

A3.4 xINFO EXAMPLE

The following example shows how some of the xInfo records could be populated using the topology shown in Figure 218 Topology for xInfo Example on page 796. All modules are assumed to be standard sized.

Given its illustrative nature, there is no consideration intended toward what is practically realizable for other considerations (i.e. building up 6 port switches with 3 port building blocks is not likely to be predominant).

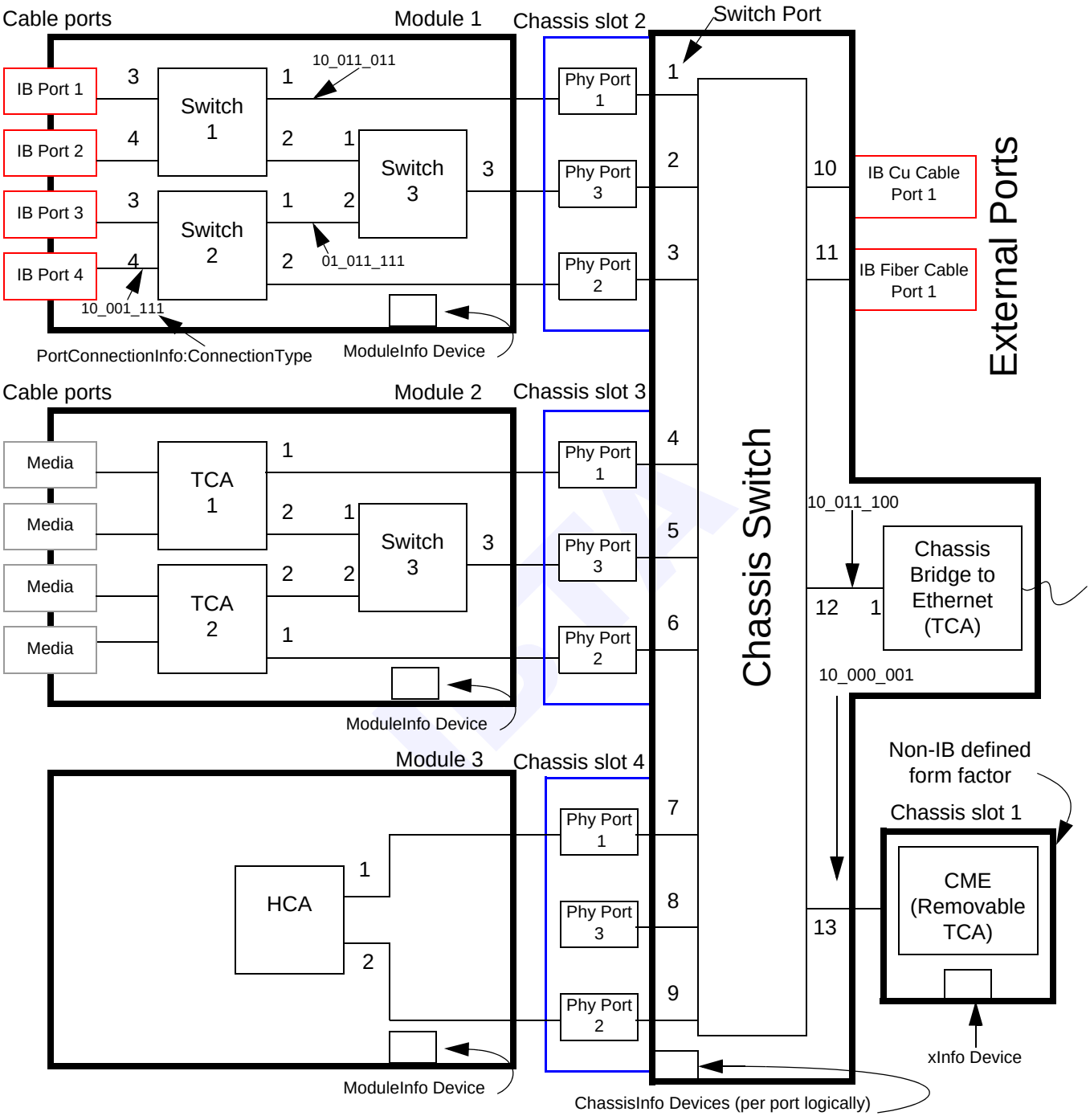


Figure 218 Topology for xInfo Example

Table 184 shows a potential layout of the records on Module 1 containing three switch nodes as shown in Figure 218 Topology for xInfo Example on page 796.

Table 184 Example Module 1 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
DeviceHeader					
0	0	1	VPD_FormatVersion	0x11	Version 1.1
1	1	1	ExtensionDeviceSlaveAddress	0x00	No extension device
2	2	1	LSB of offset to first byte of write-protected area	0x00	Begins at offset 0x0000
3	3	1	MSB of offset to first byte of write-protected area	0x00	
4	4	1	LSB of offset to last byte of write-protected area	0xFF	
5	5	1	MSB of offset to last byte of write-protected area	0x00	
6	6	1	Header Checksum (Device Header)	Calculated	
ModuleInfo					
7	7	1	RecordID	0x00	ModuleInfo
8	8	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
9	9	1	RecordLength	0x11	ModuleGUID through Checksum
10	A	1	Header Checksum	Calculated	
11	B	8	ModuleGUID	set by vendor	
19	13	1	IBModuleType, ModuleClass	xb001_00010	Module is an IB module, Class=Switch
20	14	1	NodeCount	0x03	Three (3) nodes present
21	15	1	LinkCount	0x0B	Needs definition work!!!
22	16	1	BackplaneLinkCount	0x03	Three links go into the backplane
23	17	1	IBMLCount	0x01	One(1) IB-ML per physical connector
24	18	1	BackplaneIBMLCount	0x01	One(1) IB-ML per physical connector
25	19	1	ModuleSize	0x00	Field unpopulated in this example
26	1A	1	FormFactor	0x02	Standard Module
27	1B	1	Checksum	Calculated	ModuleGUID through NodeGUIDHandle(3)
FRUInfo					

IBTA

Table 184 Example Module 1 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
28	1C	1	RecordID	0x02	FRUInfo
29	1D	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
30	1E	1	RecordLength	0x42	FRUType through Checksum
31	1F	1	Header Checksum	Calculated	
32	20	1	FRUType	0x01	InfiniBand Module
33	21	1	FRU_Handle	0x0A	Assigned by module vendor. 0x0A done for example.
34	22	1	FRUGUID	0x01	EUI-64 GUID format. 8 bytes to follow.
35	23	8	FRUGUID data bytes	set by vendor	
43	2B	1	Serial Number	0b11_001000	Encoding type=ASCII, Character Count=8
44	2C	8	Serial Number data bytes	0x49_42_41_2D_39_38_37_36	IBA-9876 (example) - Vendor assigned
52	34	1	Part Number	0b11_0000110	Encoding type=ASCII, Character Count=6
53	35	6	Part Number data bytes	0x31_32_33_34_30_39	123409 (example) - Vendor assigned
59	3B	1	Model Number	0b00_000000	Encoding type=Binary/unspecified, Character Count=0 (Model number not present in this example)
60	3C	1	Version Number	0b00_000001	Encoding type=Binary/unspecified, Character Count=1
61	3D	1	Version Number data bytes	0x02	2 (example) - Vendor assigned
62	3E	1	Manufacturer Name	0b11_001100	Encoding type=ASCII, Character Count=12
63	3F	12	Manufacturer Name data bytes	ASCII of vendor name	Vendor name
75	4B	1	Product Name	0b11_010000	Encoding type=ASCII, Character Count=16
76	4C	16	Product Name data bytes	ASCII of vendor product name	Vendor assigned name
92	5C	1	ManufacturerID	0x01	EUI-64 Company ID format, 3 bytes to follow
93	5D	3	ManufacturerID data bytes	vendor company ID	Vendor company ID
96	60	1	Mfg.Date/Time	0x00	Indicates date and time are not present
92	5C	1	Checksum	Calculated	FRUType through Mfg Date / Time
ModulePowerInfo					
93	5D	1	RecordID	0x03	ModulePowerInfo

Table 184 Example Module 1 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
94	5E	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
95	5F	1	RecordLength	0x10	FRUType through Checksum
96	60	1	Header Checksum	Calculated	
97	61	3	OperationalThermalPower	0x0061A8	Assume 25W
100	64	3	OperationalCurrent	0x007D0	Assume 2A
103	67	2	IdleCurrent	0x1F4	Assume.5A
105	69	2	InitCurrent	0x898	Assume 2.2A
107	6B	2	InitTime	0x320	Assume 8 secs (8 s * 1000ms/s / 10)
109	6D	1	ModulePMCapability	0b00000_0_0	Reserved, IsMStandbySupported=n, IsPowerManagmentSupported=n
110	6E	1	ModuleUnitCapability	0b00000_0_0	Reserved, IsUsleepSupported=n, IsUStandbySupported=n
111	6F	1	ModulePowerParms	0b00_01_00_01	Reserved, PowerClass=I, Reserved, RedundantPower=n
109	6D	1	Checksum	Calculated	OperationalThermalPower through ModulePowerParms
BuddyInfo					
110	6E	1	RecordID	0x07	BuddyInfo
111	6F	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
112	70	1	RecordLength	0x1D	BuddyCount through Checksum
113	71	1	Header Checksum	Calculated	
114	72	1	BuddyCount	0b000011_00	Nodes present=3, EUI-64 GUID format.
115	73	8	NodeGUID	set by vendor	Switch 1
123	7B	1	NodeGUIDHandle	0x01	Handles assigned beginning from 1 in this example
124	7C	8	NodeGUID	set by vendor	Switch 2
132	84	1	NodeGUIDHandle	0x02	Incremented
133	85	8	NodeGUID	set by vendor	Switch 3
141	8D	1	NodeGUIDHandle	0x03	Incremented
142	8E	1	Checksum	Calculated	BuddyCount through last NodeGUIDHandle
PortConnectionInfo					

IBTA

Table 184 Example Module 1 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
143	8F	1	RecordID	0x04	PortConnectionInfo
144	90	1	RecordParameters	0b1_0_1_0_0010	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
145	91	1	RecordLength	0x2E	ConnectionInfoCount through Checksum
146	92	1	Header Checksum	Calculated	
147	93	1	ConnectionInfoCount	0x0B	Number of ports connected
148	94	1	NodeGUIDHandle	0x01	Switch 1
149	95	1	PortNo	0x01	Port 1
150	96	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
151	97	1	BackplaneConnection	0b0_0_001_001	Reserved, Primary Slot, Primary Connector, Physical Port 1
152	98	1	NodeGUIDHandle	0x01	Switch 1
153	99	1	PortNo	0x02	Port 2
154	9A	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
155	9B	1	BackplaneConnection	0x00	Undefined for internal connections
156	9C	1	NodeGUIDHandle	0x01	Switch 1
157	9D	1	PortNo	0x03	Port 3
158	9E	1	ConnectionType	0b10_001_111	Connector, Copper Cable, Fully defined
159	9F	1	BackplaneConnection	0b0_0_001_001	Undefined for cable connections
160	A0	1	NodeGUIDHandle	0x01	Switch 1
161	A1	1	PortNo	0x04	Port 4
162	A2	1	ConnectionType	0b10_001_111	Connector, Copper Cable, Fully defined
163	A3	1	BackplaneConnection	0x00	Undefined for cable connections
164	A4	1	NodeGUIDHandle	0x02	Switch 2
165	A5	1	PortNo	0x01	Port 1
166	A6	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
167	A7	1	BackplaneConnection	0x00	Undefined for internal connections
168	A8	1	NodeGUIDHandle	0x02	Switch 2
169	A9	1	PortNo	0x02	Port 2

IBTA

Table 184 Example Module 1 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
170	AA	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
171	AB	1	BackplaneConnection	0b0_0_001_010	Reserved, Primary Slot, Primary Connector, Physical Port 2
172	AC	1	NodeGUIDHandle	0x02	Switch 2
173	AD	1	PortNo	0x03	Port 3
174	AE	1	ConnectionType	0b10_001_111	Connector, Copper Cable, Fully defined
175	AF	1	BackplaneConnection	0x00	Undefined for cable connections
176	B0	1	NodeGUIDHandle	0x02	Switch 2
177	B1	1	PortNo	0x04	Port 4
178	B2	1	ConnectionType	0b10_001_111	Connector, Copper Cable, Fully defined
179	B3	1	BackplaneConnection	0x00	Undefined for cable connections
180	B4	1	NodeGUIDHandle	0x03	Switch 3
181	B5	1	PortNo	0x01	Port 1
182	B6	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
183	B7	1	BackplaneConnection	0x00	Undefined for internal connections
184	B8	1	NodeGUIDHandle	0x03	Switch 3
185	B9	1	PortNo	0x02	Port 2
186	BA	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
187	BB	1	BackplaneConnection	0x00	Undefined for internal connections
188	BC	1	NodeGUIDHandle	0x03	Switch 3
189	BD	1	PortNo	0x03	Port 3
190	BE	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
191	BF	1	BackplaneConnection	0b0_0_001_011	Reserved, Primary Slot, Primary Connector, Physical Port 3
192	C0	1	No more fields indicator	0xC0	
193	C1	62	Padding	Filled with 0xFF	Pad to Writable Area
255	FF	1	Checksum	Calculated	ConnectionInfoCount through last Back-planeConnection
AssetTag					
256	100	1	RecordID	0x08	AssetTag

Table 184 Example Module 1 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
257	101	1	RecordParameters	0b1_0_0_0_0001	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=n; Reserved=0, RecordFormat=1
258	102	1	RecordLength	0x13	FRUType through Checksum
259	103	1	Header Checksum	Calculated	
260	104	1	FRU_Handle	0xA	Handle for FRU
261	105	1	AssetTag	0b11_001000	Encoding type=ASCII, Character Count=16
262	106	16	AssetTag data bytes	Value written	16 ASCII characters for this asset
278	116	1	Checksum	Calculated	FRU_Handle through Asset Tag data bytes

Table 185 shows a potential layout of the records on Module 2 containing two TCA nodes and a switch node as shown in Figure 218 Topology for xInfo Example on page 796. For classification purposes, this module is surfacing itself as a "TCA" even though it physically contains a switch element.

Table 185 Example Module 2 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
DeviceHeader					
0	0	1	VPD_FormatVersion	0x11	Version 1.1
1	1	1	ExtensionDeviceSlaveAddress	0x00	No extension device
2	2	1	LSB of offset to first byte of write-protected area	0x00	Begins at offset 0x0000
3	3	1	MSB of offset to first byte of write-protected area	0x00	
4	4	1	LSB of offset to last byte of write-protected area	0xDF	
5	5	1	MSB of offset to last byte of write-protected area	0x00	

Table 185 Example Module 2 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
6	6	1	Header Checksum (Device Header)	Calculated	
FRUInfo					
7	7	1	RecordID	0x02	FRUInfo
8	8	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
9	9	1	RecordLength	0x4F	FRUType through Checksum
10	A	1	Header Checksum	Calculated	
11	B	1	FRUType	0x01	InfiniBand Module
12	C	1	FRU_Handle	0x0A	Assigned by module vendor. 0x0A done for example.
13	D	1	FRUGUID	0x01	EUI-64 GUID format. 8 bytes to follow.
14	E	8	FRUGUID data bytes	set by vendor	
22	16	1	Serial Number	0b11_001000	Encoding type=ASCII, Character Count=8
23	17	8	Serial Number data bytes	set by vendor	
31	1F	1	Part Number	0b11_001000	Encoding type=ASCII, Character Count=8
32	20	8	Part Number data bytes	set by vendor	
40	28	1	Model Number	0b11_001000	Encoding type=ASCII, Character Count=8
41	29	8	Model Number data bytes	set by vendor	
49	31	1	Version Number data bytes	0x01	1 (example) - Vendor assigned
50	32	1	Manufacturer Name	0b11_010000	Encoding type=ASCII, Character Count=16
51	33	16	Manufacturer Name data bytes	set by vendor	Vendor name
67	43	1	Product Name	0b11_010000	Encoding type=ASCII, Character Count=16
68	44	16	Product Name data bytes	ASCII of vendor product name	Vendor assigned name
84	54	1	ManufacturerID	0x01	EUI-64 Company ID format, 3 bytes to follow
85	55	3	ManufacturerID data bytes	vendor company ID	Vendor company ID
88	58	1	Mfg.Date/Time	0x00	Indicates date and time are not present
84	54	1	Checksum	Calculated	FRUType through Mfg Date / Time
ModulePowerInfo					
85	55	1	RecordID	0x03	ModulePowerInfo

IBTA

Table 185 Example Module 2 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
86	56	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
87	57	1	RecordLength	0x10	FRUType through Checksum
88	58	1	Header Checksum	Calculated	
89	59	3	OperationalThermalPower	0x0061A8	Assume 25W
92	5C	3	OperationalCurrent	0x007D0	Assume 2A
95	5F	2	IdleCurrent	0x1F4	Assume .5A
97	61	2	InitCurrent	0x898	Assume 2.2A
99	63	2	InitTime	0x320	Assume 8 secs (8 s * 1000ms/s / 10)
101	65	1	ModulePMCapability	0b00000_0_0	Reserved, IsMStandbySupported=n, IsPowerManagmentSupported=n
102	66	1	ModuleUnitCapability	0b00000_0_0	Reserved, IsUsleepSupported=n, IsUStandbySupported=n
103	67	1	ModulePowerParms	0b00_01_00_01	Reserved, PowerClass=I, Reserved, RedundantPower=n
101	65	1	Checksum	Calculated	OperationalThermalPower through ModulePowerParms
ModuleInfo					
102	66	1	RecordID	0x00	ModuleInfo
103	67	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
104	68	1	RecordLength	0x2C	ModuleGUID through Checksum
105	69	1	Header Checksum	Calculated	
106	6A	8	ModuleGUID	set by vendor	
114	72	1	IBModuleType, ModuleClass	xb001_00000	Module is an IB module, Class=TCA
115	73	1	NodeCount	0x03	Three (3) nodes present
116	74	1	LinkCount	0x07	Needs definition work!!!
117	75	1	BackplaneLinkCount	0x03	Three links go into the backplane
118	76	1	IBMLCount	0x01	One(1) IB-ML per physical connector
119	77	1	BackplaneIBMLCount	0x01	One(1) IB-ML per physical connector
120	78	1	ModuleSize	0x00	Field unpopulated in this example
121	79	1	FormFactor	0x02	Standard Module

IBTA

Table 185 Example Module 2 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
122	7A	1	Checksum	Calculated	ModuleGUID through NodeGUIDHandle(3)
BuddyInfo					
123	7B	1	RecordID	0x07	BuddyInfo
124	7C	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
125	7D	1	RecordLength	0x2C	BuddyCount through Checksum
126	7E	1	Header Checksum	Calculated	
127	7F	1	BuddyCount	0b000011_00	Nodes present=3, EUI-64 GUID format.
128	80	8	NodeGUID	set by vendor	TCA 1
136	88	1	NodeGUIDHandle	0x80	Handles assigned beginning from 80h in this example
137	89	8	NodeGUID	set by vendor	TCA 2
145	91	1	NodeGUIDHandle	0x81	Incremented
146	92	8	NodeGUID	set by vendor	Switch 3
154	9A	1	NodeGUIDHandle	0x82	Incremented
155	9B	1	Checksum	Calculated	BuddyCount through last NodeGUIDHandle
PortConnectionInfo					
156	9C	1	RecordID	0x04	PortConnectionInfo
157	9D	1	RecordParameters	0b1_0_1_0_0010	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
158	9E	1	RecordLength	0x1E	ConnectionInfoCount through Checksum
159	9F	1	Header Checksum	Calculated	
160	A0	1	ConnectionInfoCount	0x07	Number of ports connected
161	A1	1	NodeGUIDHandle	0x80	TCA 1
162	A2	1	PortNo	0x01	Port 1
163	A3	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
164	A4	1	BackplaneConnection	0b0_0_001_001	Reserved, Primary Slot, Primary Connector, Physical Port 1
165	A5	1	NodeGUIDHandle	0x80	TCA 1
166	A6	1	PortNo	0x02	Port 2
167	A7	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable

Table 185 Example Module 2 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
168	A8	1	BackplaneConnection	0x00	Undefined for internal connections
169	A9	1	NodeGUIDHandle	0x81	TCA 2
170	AA	1	PortNo	0x01	Port 1
171	AB	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
172	AC	1	BackplaneConnection	0b0_0_001_010	Reserved, Primary Slot, Primary Connector, Physical Port 2
173	AD	1	NodeGUIDHandle	0x81	TCA 2
174	AE	1	PortNo	0x02	Port 2
175	AF	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
176	B0	1	BackplaneConnection	0x00	Undefined for internal connections
177	B1	1	NodeGUIDHandle	0x82	Switch 3
178	B2	1	PortNo	0x01	Port 1
179	B3	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
180	B4	1	BackplaneConnection	0x00	Undefined for internal connections
181	B5	1	NodeGUIDHandle	0x82	Switch 3
182	B6	1	PortNo	0x02	Port 2
183	B7	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
184	B8	1	BackplaneConnection	0x00	Undefined for internal connections
185	B9	1	NodeGUIDHandle	0x82	Switch 3
186	BA	1	PortNo	0x03	Port 3
187	BB	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
188	BC	1	BackplaneConnection	0b0_0_001_011	Reserved, Primary Slot, Primary Connector, Physical Port 3
189	BD	1	Checksum	Calculated	ConnectionInfoCount through last Back-planeConnection
AssetTag					
224	E0	1	RecordID	0x08	AssetTag
225	E1	1	RecordParameters	0b1_0_0_0_0001	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=n; Reserved=0, RecordFormat=1
226	E2	1	RecordLength	0x23	FRUType through Checksum

Table 185 Example Module 2 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
227	E3	1	Header Checksum	Calculated	
228	E4	1	FRU_Handle	0xA	Handle for FRU
229	E5	1	AssetTag	0b11_010000	Encoding type=ASCII, Character Count=32
230	E6	32	AssetTag data bytes	Value written	32 ASCII characters for this asset
262	106	1	Checksum	Calculated	FRU_Handle through Asset Tag data bytes

Table 186 shows a potential layout of the records on Module 3 containing a single HCA node as shown in Figure 218 Topology for xInfo Example on page 796.

Table 186 Example Module 3 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
DeviceHeader					
0	0	1	VPD_FormatVersion	0x11	Version 1.1
1	1	1	ExtensionDeviceSlaveAddress	0x00	No extension device
2	2	1	LSB of offset to first byte of write-protected area	0x00	Begins at offset 0x0000
3	3	1	MSB of offset to first byte of write-protected area	0x00	
4	4	1	LSB of offset to last byte of write-protected area	0xE5	End of PortConnectionInfo (for this example)
5	5	1	MSB of offset to last byte of write-protected area	0x00	
6	6	1	Header Checksum (Device Header)	Calculated	
ModuleInfo					
7	7	1	RecordID	0x00	ModuleInfo
8	8	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
9	9	1	RecordLength	0x1A	ModuleGUID through Checksum
10	A	1	Header Checksum	Calculated	

IBTA

Table 186 Example Module 3 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
11	B	8	ModuleGUID	set by vendor	
19	13	1	IBModuleType, ModuleClass	xb001_00001	Module is an IB module, Class=HCA
20	14	1	NodeCount	0x01	One (1) node present
21	15	1	LinkCount	0x02	Needs definition work!!!
22	16	1	BackplaneLinkCount	0x02	Three links go into the backplane
23	17	1	IBMLCount	0x01	One(1) IB-ML per physical connector
24	18	1	BackplaneIBMLCount	0x01	One(1) IB-ML per physical connector
25	19	1	ModuleSize	0x00	Field unpopulated in this example
26	1A	1	FormFactor	0x02	Standard Module
27	1B	1	Checksum	Calculated	ModuleGUID through NodeGUIDHandle(3)
FRUInfo					
28	1C	1	RecordID	0x02	FRUInfo
29	1D	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
30	1E	1	RecordLength	0x89	FRUType through Checksum
31	1F	1	Header Checksum	Calculated	
32	20	1	FRUType	0x01	InfiniBand Module
33	21	1	FRU_Handle	0xCD	Assigned by module vendor. 0xCD done for example.
34	22	1	FRUGUID	0x01	EUI-64 GUID format. 8 bytes to follow.
35	23	8	FRUGUID data bytes	set by vendor	
43	2B	1	Serial Number	0b10_001000	Encoding type=Unicode, Character Count=8 (note Unicode 3 have 2 byte characters)
44	2C	16	Serial Number data bytes	set by vendor	Vendor assigned
60	3C	1	Part Number	0b10_010000	Encoding type=Unicode, Character Count=16
61	3D	32	Part Number data bytes	set by vendor	Vendor assigned
93	5D	1	Model Number	0b10_010000	Encoding type=Unicode, Character Count=16
94	5E	32	Model Number data bytes	set by vendor	Vendor assigned
94	5E	1	Version Number	0b00_000010	Encoding type=Binary/unspecified, Character Count=2

IBTA

Table 186 Example Module 3 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
95	5F	2	Version Number data bytes	0x0005	5 (example) - Vendor assigned
97	61	1	Manufacturer Name	0b10_010000	Encoding type=Unicode, Character Count=16
98	62	32	Manufacturer Name data bytes	set by vendor	Vendor assigned
130	82	1	Product Name	0b10_010000	Encoding type=Unicode, Character Count=16
131	83	32	Product Name data bytes	set by vendor	Vendor assigned
163	A3	1	ManufacturerID	0x01	EUI-64 Company ID format, 3 bytes to follow
164	A4	3	ManufacturerID data bytes	vendor company ID	Vendor company ID
167	A7	1	Mfg.Date/Time	0x00	Indicates date and time are not present
168	A8	1	Checksum	Calculated	FRUType through Mfg Date / Time
ModulePowerInfo					
169	A9	1	RecordID	0x03	ModulePowerInfo
170	AA	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
171	AB	1	RecordLength	0x10	FRUType through Checksum
172	AC	1	Header Checksum	Calculated	
173	AD	3	OperationalThermalPower	0x0061A8	Assume 25W
176	B0	3	OperationalCurrent	0x007D0	Assume 2A
179	B3	2	IdleCurrent	0x1F4	Assume .5A
181	B5	2	InitCurrent	0x820	Assume 2.08A
183	B7	2	InitTime	0x1770	Assume 60 secs (60 s * 1000ms/s / 10)
185	B9	1	ModulePMCapability	0b00000_1_1	Reserved, IsMStandbySupported=y, IsPowerManagmentSupported=y
186	BA	1	ModuleUnitCapability	0b00000_0_0	Reserved, IsUsleepSupported=n, IsUStandbySupported=n (for this example, the implementation uses IOC functionality to perform power management)
187	BB	1	ModulePowerParms	0b00_10_00_01	Reserved, PowerClass=II, Reserved, RedundantPower=n
185	B9	1	Checksum	Calculated	OperationalThermalPower through ModulePowerParms
IOCPMInfo					

Table 186 Example Module 3 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
186	BA	1	RecordID	0x09	IOCPMInfo
187	BB	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
188	BC	1	RecordLength	0x1A	BuddyCount through Checksum
189	BD	1	Header Checksum	Calculated	
190	BE	1	IOC_Count	0x01	1 node present
191	BF	1	IOCPMCapabilty	0b1_1_1_1_1_1_1_1	Doze=y, Nap=y, Sleep=y, Standby=y, WREDoze=y, WRENap=y, WRESleep=y, WREStandby=n
192	C0	2	IDozeCurrent	0x05DC	Assume 1.5A
194	C2	2	INapCurrent	0x044C	Assume 1.1A
196	C4	2	ISleepCurrent	0x012C	Assume .3A
198	C6	2	IStandbyCurrent	0x00E6	Assume .23A (@ 5V = 1.15W)
200	C8	1	Checksum	Calculated	BuddyCount through last NodeGUIDHandle
BuddyInfo					
201	C9	1	RecordID	0x07	BuddyInfo
202	CA	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
203	CB	1	RecordLength	0x1A	BuddyCount through Checksum
204	CC	1	Header Checksum	Calculated	
205	CD	1	BuddyCount	0b000001_00	Nodes present=1, EUI-64 GUID format.
206	CE	8	NodeGUID	set by vendor	HCA
214	D6	1	NodeGUIDHandle	0x00	Handles assigned beginning from 00h in this example
215	D7	1	Checksum	Calculated	BuddyCount through last NodeGUIDHandle
PortConnectionInfo					
216	D8	1	RecordID	0x04	PortConnectionInfo
217	D9	1	RecordParameters	0b1_0_1_0_0010	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
218	DA	1	RecordLength	0x0A	ConnectionInfoCount through Checksum
219	DB	1	Header Checksum	Calculated	
220	DC	1	ConnectionInfoCount	0x02	Number of ports connected

IBTA

Table 186 Example Module 3 xInfo

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
221	DD	1	NodeGUIDHandle	0x00	HCA
222	DE	1	PortNo	0x01	Port 1
223	DF	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
224	E0	1	BackplaneConnection	0b0_0_001_001	Reserved, Primary Slot, Primary Connector, Physical Port 1
225	E1	1	NodeGUIDHandle	0x00	HCA
226	E2	1	PortNo	0x02	Port 2
227	E3	1	ConnectionType	0b10_011_011	Connector, PCB, Module side of Backplane connector
228	E4	1	BackplaneConnection	0b0_0_001_010	Reserved, Primary Slot, Primary Connector, Physical Port 2
229	E5	1	Checksum	Calculated	ConnectionInfoCount through last Back-planeConnection
AssetTag					
230	E6	1	RecordID	0x08	AssetTag
231	E7	1	RecordParameters	0b1_0_0_0_0001	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=n; Reserved=0, RecordFormat=1
232	E8	1	RecordLength	0x13	FRUType through Checksum
233	E9	1	Header Checksum	Calculated	
234	EA	1	FRU_Handle	0xCD	Handle for FRU
235	EB	1	AssetTag	0b11_001000	Encoding type=ASCII, Character Count=16
236	EC	16	AssetTag data bytes	Value written	16 ASCII characters for this asset
252	FC	1	Checksum	Calculated	FRU_Handle through Asset Tag data bytes

IBTA

Table 187 shows a potential layout of the records on the Chassis containing the switch and Ethernet TCA shown in Figure 218 Topology for xInfo Example on page 796.

Table 187 Example ChassisInfo Device

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
DeviceHeader					
0	0	1	VPD_FormatVersion	0x11	Version 1.1
1	1	1	ExtensionDeviceSlaveAddress	0x00	No extension device
2	2	1	LSB of offset to first byte of write-protected area	0x00	Begins at offset 0x0000
3	3	1	MSB of offset to first byte of write-protected area	0x00	
4	4	1	LSB of offset to last byte of write-protected area	0x80	
5	5	1	MSB of offset to last byte of write-protected area	0x00	
6	6	1	Header Checksum (Device Header)	Calculated	
ChassisInfo					
7	7	1	RecordID	0x01	ChassisInfo
8	8	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
9	9	1	RecordLength	0x1B	ChassisGUID through Checksum
10	A	1	Header Checksum	Calculated	
11	B	8	ChassisGUID	set by vendor	
19	13	1	SlotCount	0x03	Three (3) IB slots for this example, removable CME and any others are not included. Assume High Slot Number for the chassis is 8.
20	14	1	SlotNumbers	0b1_0_01_01_00	Last byte=n, reserved, Standard Slots for each of slots 3, 2, 1
21	15	1	SlotNumbers	0b1_0_00_00_01	Last byte=n, reserved, Standard Slots for each of slots 6, 5, 4
22	16	1	SlotNumbers	0b0_0_00_00_00	Last byte=y, reserved, Standard Slots for each of slots 9, 8, 7
23	17	1	CMEAccess	0b1_0001111	Last byte=n, CME access for each of slots 7, 6, 5, 4, 3, 2, 1

IBTA

Table 187 Example ChassisInfo Device

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
24	18	1	CMEAccess	0b0_0000000	Last byte=y, CME access for each of slots 14,13, 12, 11, 10 , 9 , 8
25	19	1	SlotNumber	0x00	Indicates this is the chassis containing the CME
26	1A	1	SlotDetails	0b01_1_0_00_00	CMEAccess=y, ProxyAccess=y, reserved, LockDriveCTR=unspecified, Mechanical-LockPresent=unspecified
27	1B	1	NodeCount	0x02	Nodes on this Chassis (2) - Switch and Ethernet device
28	1C	1	Checksum	Calculated	ChassisGUID through NodeGUIDHandle(2)
BuddyInfo					
29	1D	1	RecordID	0x07	BuddyInfo
30	1E	1	RecordParameters	0b0_0_1_0_0010	LastRecord=no, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
31	1F	1	RecordLength	0x24	BuddyCount through Checksum
32	20	1	Header Checksum	Calculated	
33	21	1	BuddyCount	0b000010_00	Nodes present=2, EUI-64 GUID format.
34	22	8	NodeGUID	set by vendor	Chassis Switch
42	2A	1	NodeGUIDHandle	0x00	Handles assigned beginning from 00h in this example
43	2B	8	NodeGUID	set by vendor	TCA - Chassis Ethernet
51	33	1	NodeGUIDHandle	0x01	Incremented
52	34	1	Checksum	Calculated	BuddyCount through last NodeGUIDHandle
PortConnectionInfo					
53	35	1	RecordID	0x04	PortConnectionInfo
54	36	1	RecordParameters	0b1_0_1_0_0010	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
55	37	1	RecordLength	0x3A	ConnectionInfoCount through Checksum
56	38	1	Header Checksum	Calculated	
57	39	1	ConnectionInfoCount	0x0E	Number of ports connected - 13 for switch, 1 for Ethernet device
58	3A	1	NodeGUIDHandle	0x00	Switch
59	3B	1	PortNo	0x01	Port 1

IBTA

Table 187 Example ChassisInfo Device

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
60	3C	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
61	3D	1	BackplaneConnection	0b0_0_000_001	Reserved, Connector Location=Primary, Reserved, Physical Port Number=1
62	3E	1	NodeGUIDHandle	0x00	Switch
63	3F	1	PortNo	0x02	Port 2
64	40	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
65	41	1	BackplaneConnection	0b0_0_000_011	Reserved, Connector Location=Primary, Reserved, Physical Port Number=3
66	42	1	NodeGUIDHandle	0x00	Switch
67	43	1	PortNo	0x03	Port 3
68	44	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
69	45	1	BackplaneConnection	0b0_0_000_010	Reserved, Connector Location=Primary, Reserved, Physical Port Number=2
70	46	1	NodeGUIDHandle	0x00	Switch
71	47	1	PortNo	0x04	Port 4
72	48	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
73	49	1	BackplaneConnection	0b0_0_000_001	Reserved, Connector Location=Primary, Reserved, Physical Port Number=1
74	4A	1	NodeGUIDHandle	0x00	Switch
75	4B	1	PortNo	0x05	Port 5
76	4C	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
77	4D	1	BackplaneConnection	0b0_0_000_011	Reserved, Connector Location=Primary, Reserved, Physical Port Number=3
78	4E	1	NodeGUIDHandle	0x00	Switch
79	4F	1	PortNo	0x06	Port 6
80	50	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
81	51	1	BackplaneConnection	0b0_0_000_010	Reserved, Connector Location=Primary, Reserved, Physical Port Number=2
82	52	1	NodeGUIDHandle	0x00	Switch

IBTA

Table 187 Example ChassisInfo Device

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
83	53	1	PortNo	0x07	Port 7
84	54	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
85	55	1	BackplaneConnection	0b0_0_000_001	Reserved, Connector Location=Primary, Reserved, Physical Port Number=1
86	56	1	NodeGUIDHandle	0x00	Switch
87	57	1	PortNo	0x08	Port 8
88	58	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
89	59	1	BackplaneConnection	0b0_0_000_011	Reserved, Connector Location=Primary, Reserved, Physical Port Number=3
90	5A	1	NodeGUIDHandle	0x00	Switch
91	5B	1	PortNo	0x09	Port 9
92	5C	1	ConnectionType	0b10_011_010	Connector, PCB, Chassis side of Backplane connector
93	5D	1	BackplaneConnection	0b0_0_000_010	Reserved, Connector Location=Primary, Reserved, Physical Port Number=2
94	5E	1	NodeGUIDHandle	0x00	Switch
95	5F	1	PortNo	0x0A	Port 10
96	60	1	ConnectionType	0b10_001_111	Connector, Copper Cable, Fully defined
97	61	1	BackplaneConnection	0x00	Undefined for cable connections
98	62	1	NodeGUIDHandle	0x00	Switch
99	63	1	PortNo	0x0B	Port 11
100	64	1	ConnectionType	0b10_010_111	Connector, Fiber Cable, Fully defined
101	65	1	BackplaneConnection	0x00	Undefined for fiber connections
102	66	1	NodeGUIDHandle	0x00	Switch
103	67	1	PortNo	0x0C	Port 12
104	68	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
105	69	1	BackplaneConnection	0x00	Undefined for internal connections
106	6A	1	NodeGUIDHandle	0x00	Switch
107	6B	1	PortNo	0x0D	Port 13
108	6C	1	ConnectionType	0b10_011_001	Connector, PCB, Non-IBA Removable

IBTA

Table 187 Example ChassisInfo Device

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
109	6D	1	BackplaneConnection	0x00	Undefined for internal connections
110	6E	1	NodeGUIDHandle	0x01	Ethernet TCA
111	6F	1	PortNo	0x01	Port 1
112	70	1	ConnectionType	0b01_011_100	No connector, PCB, Non-removable
113	71	1	BackplaneConnection	0x00	Undefined for internal connections
114	72	1	Checksum	Calculated	ConnectionInfoCount through last Back-planeConnection
AssetTag					
128	80	1	RecordID	0x08	AssetTag
129	81	1	RecordParameters	0b1_0_0_0_0001	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=n; Reserved=0, RecordFormat=1
130	82	1	RecordLength	0x13	FRUType through Checksum
131	83	1	Header Checksum	Calculated	
132	84	1	FRU_Handle	0xA	Handle for FRU
133	85	1	AssetTag	0b11_001000	Encoding type=ASCII, Character Count=16
134	86	16	AssetTag data bytes	Value written	16 ASCII characters for this asset
150	96	1	Checksum	Calculated	FRU_Handle through Asset Tag data bytes

Table 188 shows a potential layout of the records on the CME pluggable element shown in Figure 218 Topology for xInfo Example on page 796.. Note that this element is not an InfiniBand defined module form factor but it is a node on the InfiniBand fabric.

Table 188 Example CMEInfo Device

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
DeviceHeader					
0	0	1	VPD_FormatVersion	0x11	Version 1.1
1	1	1	ExtensionDeviceSlaveAddress	0x00	No extension device

Table 188 Example CMEInfo Device

Offset (dec)	Offset (hex)	Length (bytes)	Field	Value	Comment
2	2	1	LSB of offset to first byte of write-protected area	0xFF	Begins at offset 0x0000
3	3	1	MSB of offset to first byte of write-protected area	0xFF	
4	4	1	LSB of offset to last byte of write-protected area	0x00	No writable area
5	5	1	MSB of offset to last byte of write-protected area	0x00	
6	6	1	Header Checksum (Device Header)	Calculated	
CMEInfo					
7	7	1	RecordID	0x05	CMEInfo
8	8	1	RecordParameters	0b1_0_1_0_0010	LastRecord=yes, LengthMultiplier=bytes; ReadOnly=y; Reserved=0, RecordFormat=2
9	9	1	RecordLength	0xE	BuddyCount through Checksum
10	A	1	Header Checksum	Calculated	
11	B	1	CMEGUID	0x01	EUI-64 GUID format. 8 bytes to follow.
12	C	8	CMEGUID data bytes	set by vendor	
20	14	2	CMEFirmware Revision	0x0312	"12.3" set by vendor
22	16	1	SlotNumbers	0b1_0_01_01_00	Last byte=n, reserved, Standard Slots for each of slots 3, 2, 1
23	17	1	SlotNumbers	0b1_0_00_00_01	Last byte=n, reserved, Standard Slots for each of slots 6, 5, 4
24	18	1	SlotNumbers	0b0_0_00_00_00	Last byte=y, reserved, Standard Slots for each of slots 9, 8, 7
25	19	1	Checksum	Calculated	BuddyCount through last NodeGUIDHandle

IBTA

ANNEX A4: HARDWARE MANAGEMENT INDEX OF TERMS

IBTA

IBTA

IBTA

L

IBTA

M

IBTA

IBTA

IBTA

IBTA

IBTA

IBTA

End of Document

IBTA