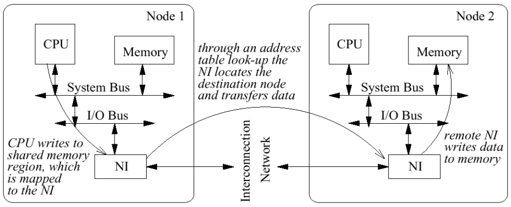
System Area Networks (SANs)

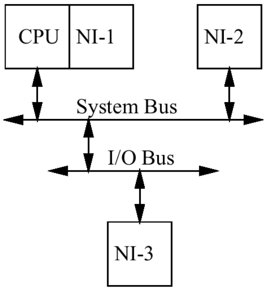
- Hardware
  - Nodes: Network Interface Card (NIC) on I/O bus (PCI. PCI-X) or maybe on motherboard
  - Components
    - Hardware to interface with physical layer of network (copper or fiber)
    - Hardware to interface with I/O bus
  - Transmission rate limited by speed of I/O bus and network
    - Currently more by I/O bus

Data Transfer Process


Node 1
CPU
Memory
System Bus
I/O Bus
NI
CPU writes to shared memory region, which is mapped to the NI
through an address table look-up the NI locates the destination node and transfers data
Interconnection Network
Node 2
CPU
Memory
System Bus
I/O Bus
NI
remote NI writes data to memory

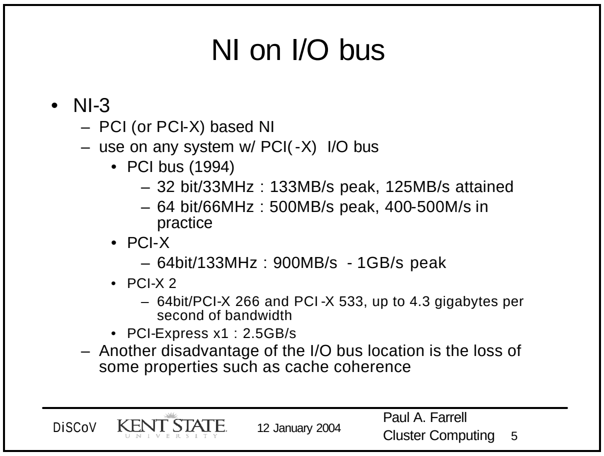
Network Interface Location

- NI location
  - Critical to performance and usability
  - NI1
    - transputer, most implemented at the prototype phase
  - NI2
    - best place for NI, but proprietary system buses
  - NI3
    - most common today, no way to support cache coherence



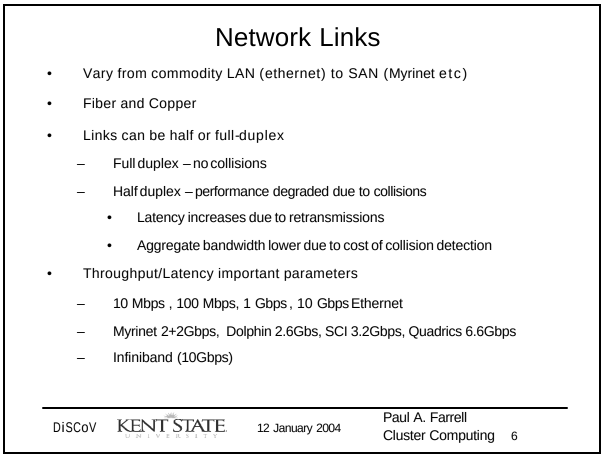
General Architecture (III)

- NI-1
  - instruction set (special communication registers)
  - Transputer from INMOS
  - iWrap, related systolic architecture
  - not successful (too small market)
- NI-2
  - ideal (case of high performance bus)
  - system bus based NI
  - poll on cache-coherent NI registers
  - DMA can read/write from/to main memory using burst cycle
  - NI implementation only



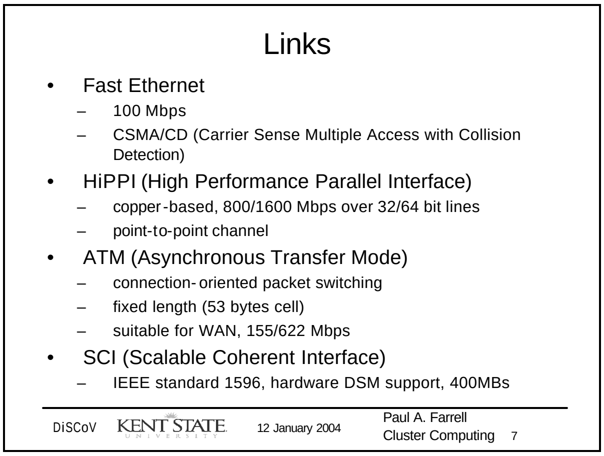
NI on I/O bus

- NI-3
  - PCI (or PCI-X) based NI
  - use on any system w/ PCI(-X) I/O bus
    - PCI bus (1994)
      - 32 bit/33MHz : 133MB/s peak, 125MB/s attained
      - 64 bit/66MHz : 500MB/s peak, 400-500M/s in practice
    - PCI-X
      - 64bit/133MHz : 900MB/s - 1GB/s peak
    - PCI-X 2
      - 64bit/PCI-X 266 and PCI -X 533, up to 4.3 gigabytes per second of bandwidth
    - PCI-Express x1 : 2.5GB/s
  - Another disadvantage of the I/O bus location is the loss of some properties such as cache coherence



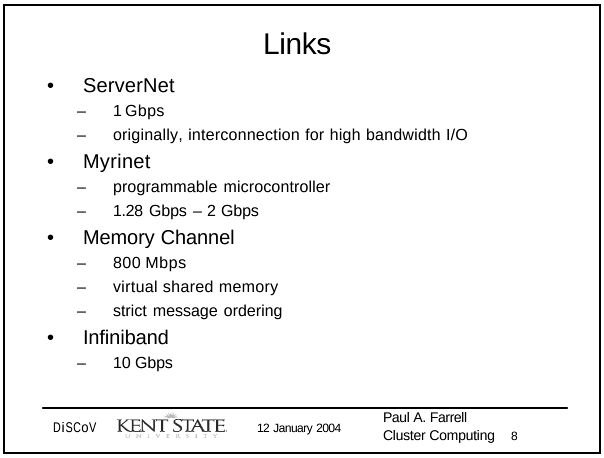
Network Links

- Vary from commodity LAN (ethernet) to SAN (Myrinet etc)
- Fiber and Copper
- Links can be half or full-duplex
  - Full duplex – no collisions
  - Half duplex – performance degraded due to collisions
    - Latency increases due to retransmissions
    - Aggregate bandwidth lower due to cost of collision detection
- Throughput/Latency important parameters
  - 10 Mbps , 100 Mbps, 1 Gbps , 10 Gbps Ethernet
  - Myrinet 2+2Gbps, Dolphin 2.6Gbs, SCI 3.2Gbps, Quadrics 6.6Gbps
  - Infiniband (10Gbps)



Links

- Fast Ethernet
  - 100 Mbps
  - CSMA/CD (Carrier Sense Multiple Access with Collision Detection)
- HiPPI (High Performance Parallel Interface)
  - copper-based, 800/1600 Mbps over 32/64 bit lines
  - point-to-point channel
- ATM (Asynchronous Transfer Mode)
  - connection-oriented packet switching
  - fixed length (53 bytes cell)
  - suitable for WAN, 155/622 Mbps
- SCI (Scalable Coherent Interface)
  - IEEE standard 1596, hardware DSM support, 400MBs



Links

- ServerNet
  - 1 Gbps
  - originally, interconnection for high bandwidth I/O
- Myrinet
  - programmable microcontroller
  - 1.28 Gbps – 2 Gbps
- Memory Channel
  - 800 Mbps
  - virtual shared memory
  - strict message ordering
- Infiniband
  - 10 Gbps

Network Devices

- Hardware interconnecting links
- Main types : hubs, switches
  - Hubs
    - Only possible if link allows contention
    - Single broadcast domain, half-duplex links, inexpensive
    - Need collision/contention detection
    - In presence of contention throughput can drop to 35%
    - Common in 10/100 Mbps
    - Not suitable for clusters

Network Devices

- Switches
  - Predominant due to price drops/performance benefits
  - Switches build database mapping ethernet hardware address to port last seen on
  - Only first frame need be broadcast
  - Performance of switches
    - Backplane bandwidth e.g. 16 Gbps = 16 ports at 1 Gbps
    - Packets per second
    - Non-blocking
  - Small networks – one switch
  - Larger networks – require multiple switches
  - To reduce bottlenecks on inter-switch links, link aggregation or trunking can be used i.e use multiple links and treat as one

Hashing Problems in Trunked Links

- Hashing used to distribute traffic over links
- Sub-optimal in cluster due to:
  - Uniformity of hardware
  - Sequential IP and possibly NIC addresses
  - Round robin hashing : good traffic distribution but packet reordering causes problem for higher network layers
- Some switches e.g. Myricom use source routing
  - More scalable
  - Client need to maintain routes to all other clients
  - Leads to better overall performance

Aims

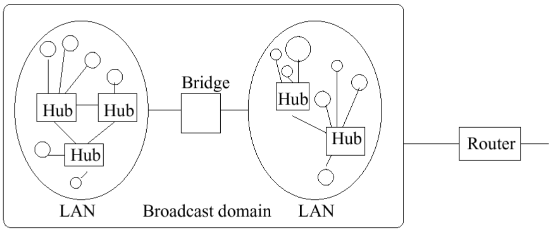
- Price vs. Performance
  - production volume, expensive physical layer, amount of storage
  - Fast Ethernet($50-100) vs. Myrinet or ServerNet ($1000 or more)
- Scalability
  - fixed topology vs. dynamic topology, shared media vs. private media
  - traditionally fixed network topology (mesh, hypercube)
  - **clusters are more dynamic**
  - network can tolerate the increased load and deliver nearly the same bandwidth latency
  - can afford larger number of nodes

Aims

- Reliability
  - CRC check level/provider, buffering storage for retransmission, protocol complexity
  - two classes of parallel computer
    - scientific and business computing
  - Networks can operate without software overhead
    - error freed physical layer
    - CRC can be computed by NIC itself
    - error signaling (interrupt or status registers)
    - NIC side buffer

Fast Ethernet (I)

- 100 Mbps over UTP or fiber-optic cable
- MAC protocol: CSMA/CD (Carrier Sense Multiple Access with Collision Detection)



Fast Ethernet (II)

- Interconnection devices
  - Repeater
    - restore data and collision signal
    - amplify and regenerate signals
  - Hub
    - central point
    - repeat and copy: All can see it
  - Bridge
    - link adjacent LANs: datalink layer
    - filtering
    - forward to other segment
  - Router
    - link adjacent LANs: network layer
    - shortest path

Gigabit Ethernet Fundamentals

- Gigabit Ethernet (802.3z) 1Gbps
  - Modified Fiber Channel physical layer
  - Actual Bit Rate 1,250,000,000 bits/second!
- Media: Fiber (multimode,singlemode), UTP Cat 5
- Uses the 802.3 Ethernet frame format.
- Full-duplex (point to point) mode and/or half-duplex (switched) mode.
- Half-duplex mode
  - Uses *enhanced* CSMA/CD access method
  - requires carrier extension to 512 byte time slot to preserve 200m collision domain
  - thus wastes bandwidth

Gigabit Ethernet Fundamentals Cont.

- Support Fiber and Copper media.
  - 25 meters short link copper.
  - 100 meters horizontal copper.
  - 500 meters multimode fiber.
  - 3000 meters single mode fiber.
- Paul A. Farrell, Hong Ong, Communication Performance over a Gigabit Network "http://discov.cs.kent.edu/publications/2000/ipccc2000.pdf"
- Ethernet – Next Generation.
  - 10 Gigabit Ethernet (802.3ae).
  - http://www.10gea.org/.

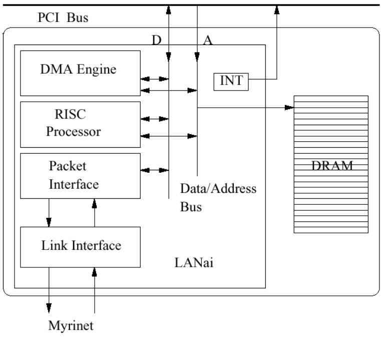
Myrinet (I)

- A SAN evolved from supercomputer technology
- A main product of Myricom (founded in 1994)
- Quite popular in the research community
  - all HW & SW specifications are open & public
- Based on 2 research projects
  - Mosaic by Caltech
    - a fine grain supercomputer, need a truly scalable interconnection network with lots of bandwidth
  - Atomic LAN by USC
    - based on Mosaic technology, a research prototype of Myrinet
- Speed: 1.28 Gbps
- Good price/performance ratio

Myrinet (II)

- Host interface
  - LANai chip
    - a custom VLSI chip, a programmable microcontroller
    - control the data transfer between the host & the network
  - SRAM memory
    - Message data must first be written to the NI SRAM, before it can be injected into the network
  - (+) the great flexibility of the HW due to a programmable microcontroller,
  - (-) but can also be a bottleneck with respect to performance since the LANai runs only at moderate frequencies

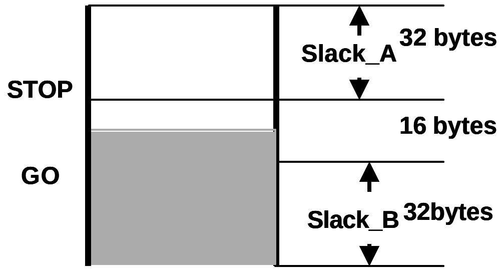
Myrinet Host Interface



Myrinet (III)

- Link and Packet Layer
  - similar to ServerNet
  - full duplex 9 bit parallel channel in one direction running at 80MHz
  - network offer 160Mbyte/s physical bandwidth over one channel
  - two different cable type (SAN, LAN)
    - 3m SAN link, 10m LAN link
  - variable length data format
  - route with wormhole switching
  - source path routing
  - consist of routing header
  - special control symbols (STOP, GO)

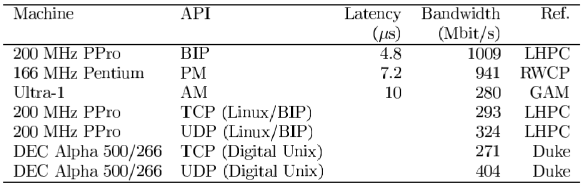
Flow Control (Slack Buffer Operation)



Myrinet (IV)

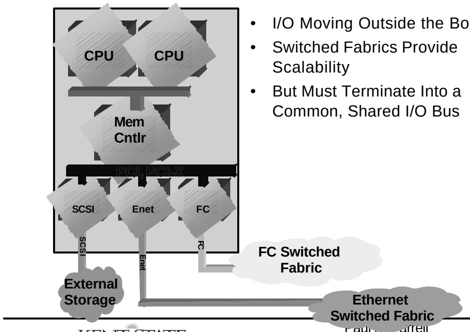
- Switches
  - 4, 8 and 16 ports, mixable SAN and LAN
  - any network topology
  - autodetect the absence of a link
  - starting up, host interface detect network topology automatically
- Error Handling
  - MTBF: million hours are reported
  - cable fault and node failure
    - alternative routing by LANai
  - prevent deadlock: time out generates a forward reset (FRES) signal

Performance of Message Layers over Myrinet



Machine	API	Latency	Bandwidth	Ref.
		(μs)	(Mbit/s)	
200 MHz PPro	BIP	4.8	1009	LHPC
166 MHz Pentium	PM	7.2	941	RWCP
Ultra-1	AM	10	280	GAM
200 MHz PPro	TCP (Linux/BIP)		293	LHPC
200 MHz PPro	UDP (Linux/BIP)		324	LHPC
DEC Alpha 500/266	TCP (Digital Unix)		271	Duke
DEC Alpha 500/266	UDP (Digital Unix)		404	Duke

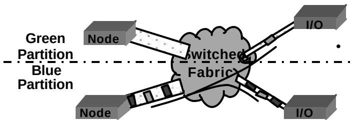
I/O Architecture Trends



- I/O Moving Outside the Box
- Switched Fabrics Provide Scalability
- But Must Terminate Into a Common, Shared I/O Bus

InfiniBand™ Protocol Features

- Flow & rate control
  - Static rate control to control sources and destinations of different speeds
  - Credit based link- level flow control for efficient congestion avoidance

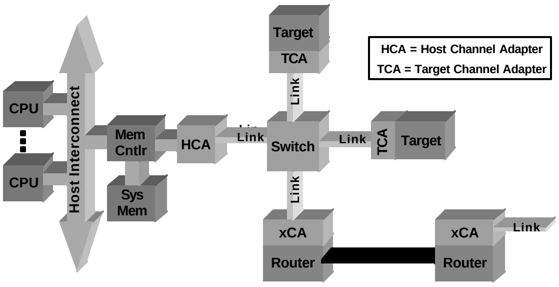


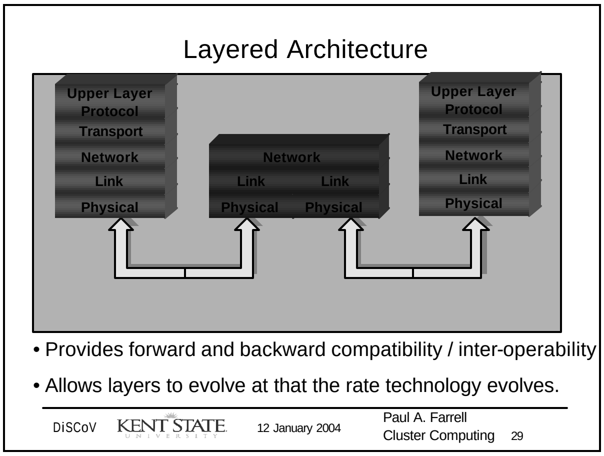
- Partitioning
  - For performance and functional isolation
  - Transparent I/O Sharing
- Multicasts
  - A single message is distributed by the fabric to multiple destinations
- Network topology
  - Subnet support
  - IPv6 GUIDs for unique end-point id
  - IPv6 addressing headers for global routing
  - Speed matched to the backbone
  - IP compatible fabric management

InfiniBand Components

- Link
- Switch
- Router
- Target Channel Adapter (TCA)
- Host Channel Adapter (HCA)

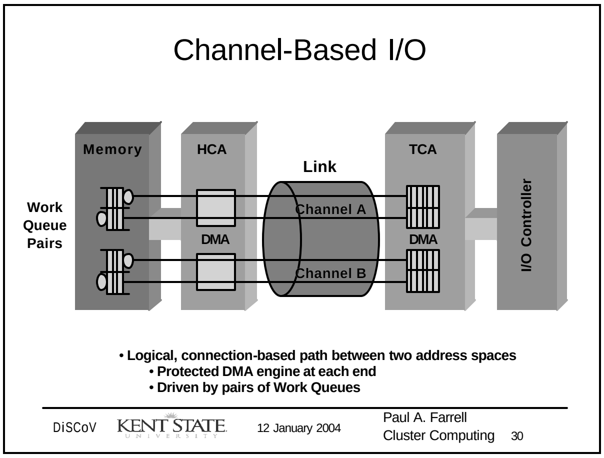
InfiniBand™ Architecture (IBA)





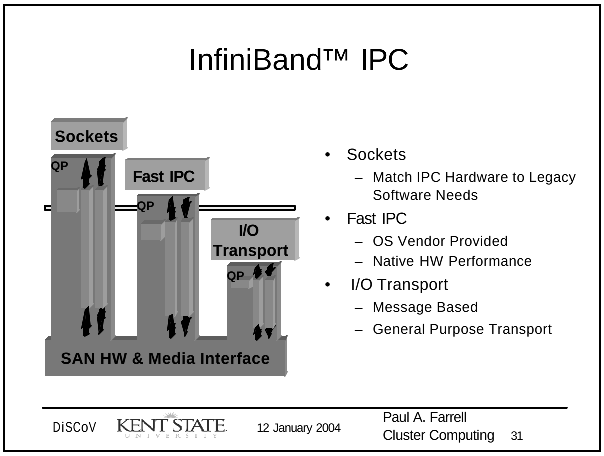
Layered Architecture

- Provides forward and backward compatibility / inter-operability
- Allows layers to evolve at that the rate technology evolves.



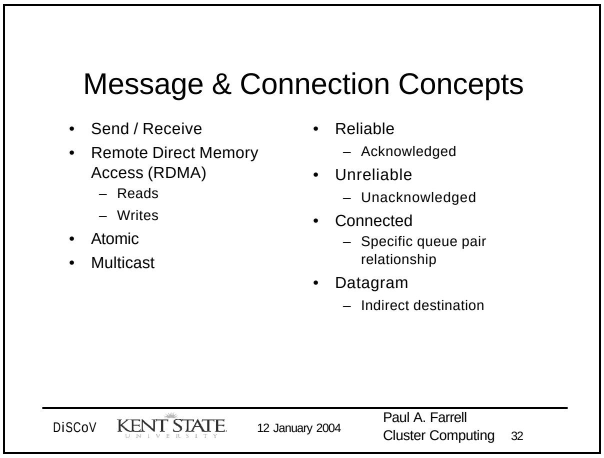
Channel-Based I/O

- Logical, connection-based path between two address spaces
  - Protected DMA engine at each end
  - Driven by pairs of Work Queues



InfiniBand™ IPC

- Sockets
  - Match IPC Hardware to Legacy Software Needs
- Fast IPC
  - OS Vendor Provided
  - Native HW Performance
- I/O Transport
  - Message Based
  - General Purpose Transport



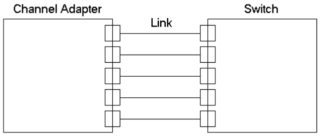
Message & Connection Concepts

- Send / Receive
- Remote Direct Memory Access (RDMA)
  - Reads
  - Writes
- Atomic
- Multicast
- Reliable
  - Acknowledged
- Unreliable
  - Unacknowledged
- Connected
  - Specific queue pair relationship
- Datagram
  - Indirect destination

InfiniBand Link

- Link
  - Full duplex
  - Point-to-point
  - 1-bit (1X) – 2.5Gbps
  - 4-bit (4x) – 10Gbps
  - 12-bit (12x) – 30Gbps

- Copper Cable, Optical Fiber, and printed circuit.
- Vendor Specific.
- May be used in Parallel.



InfiniBand Switch

- Switch
  - Local network routing
  - Implements up to 255 ports
  - Local Route Header
    - 16-bit Source Local ID
    - 16-bit Destination Local ID

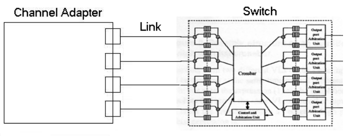
Port 0

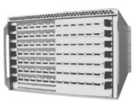
Forwarding table

Service Level to Virtual Lane mapping table

Non-multiplexed

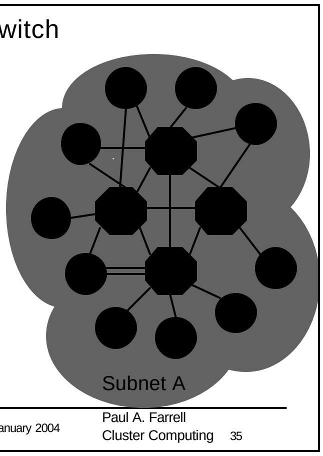
Full-multiplexed





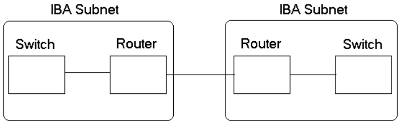
Switch

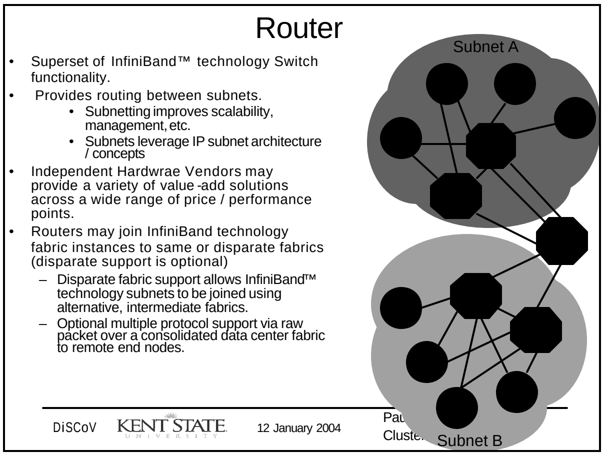
- Switches route packets only within a single subnet to reduce complexity and solution cost.
- Optional switch-based multicast for unreliable, datagram service



InfiniBand Router

- Router
  - Outside local network routing
  - Implements up to 255 ports
  - IBA, WAN, and MAN
  - Global Route Header
    - 128-bit Source Global ID
    - 128-bit Destination Global ID





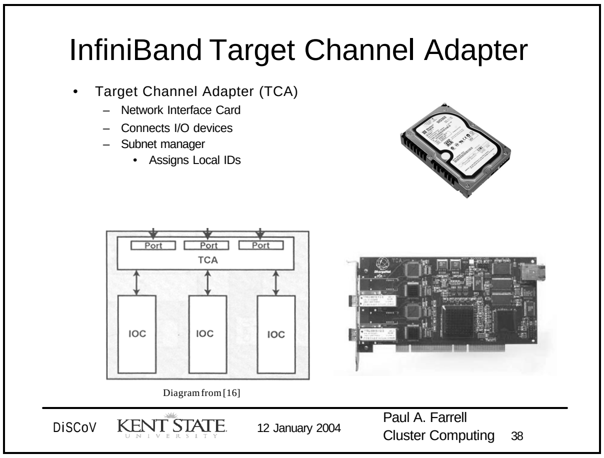
Router

- Superset of InfiniBand™ technology Switch functionality.
- Provides routing between subnets.
  - Subnetting improves scalability, management, etc.
  - Subnets leverage IP subnet architecture / concepts
- Independent Hardwrae Vendors may provide a variety of value -add solutions across a wide range of price / performance points.
- Routers may join InfiniBand technology fabric instances to same or disparate fabrics (disparate support is optional)
  - Disparate fabric support allows InfiniBand™ technology subnets to be joined using alternative, intermediate fabrics.
  - Optional multiple protocol support via raw packet over a consolidated data center fabric to remote end nodes.

Subnet A

Subnet B

DiSCoV KENT STATE UNIVERSITY 12 January 2004 Pau
Cluste



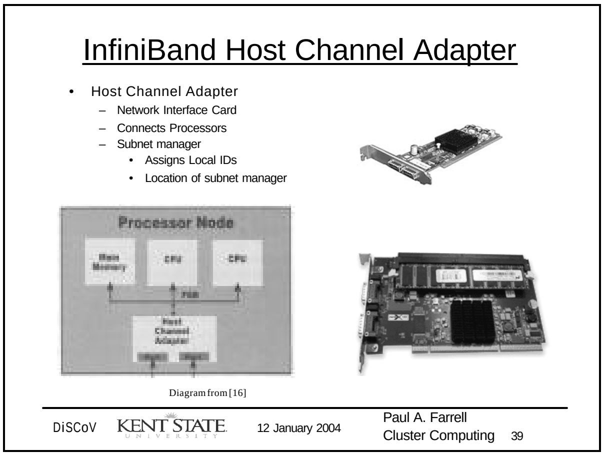
InfiniBand Target Channel Adapter

- Target Channel Adapter (TCA)
  - Network Interface Card
  - Connects I/O devices
  - Subnet manager
    - Assigns Local IDs

Port Port Port
TCA
IOC IOC IOC

Diagram from [16]

DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell
Cluster Computing 38



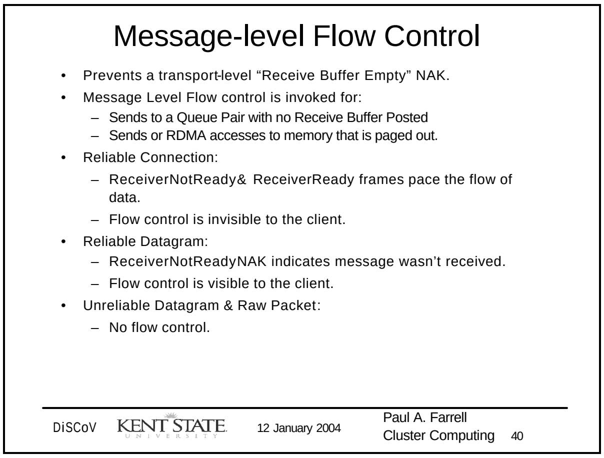
InfiniBand Host Channel Adapter

- Host Channel Adapter
  - Network Interface Card
  - Connects Processors
  - Subnet manager
    - Assigns Local IDs
    - Location of subnet manager

Processor Node
Main Memory
CPU
CPU
PSB
Host Channel Adapter

Diagram from [16]

DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell
Cluster Computing 39



Message-level Flow Control

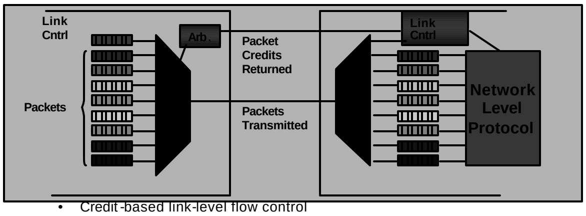
- Prevents a transport-level “Receive Buffer Empty” NAK.
- Message Level Flow control is invoked for:
  - Sends to a Queue Pair with no Receive Buffer Posted
  - Sends or RDMA accesses to memory that is paged out.
- Reliable Connection:
  - ReceiverNotReady& ReceiverReady frames pace the flow of data.
  - Flow control is invisible to the client.
- Reliable Datagram:
  - ReceiverNotReadyNAK indicates message wasn’t received.
  - Flow control is visible to the client.
- Unreliable Datagram & Raw Packet:
  - No flow control.

DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell
Cluster Computing 40

Link Characteristics: Addressing

- GUID (Globally Unique Identifier) - Each TCA, HCA, Switch, and Router has a single unique GUID.
- Local ID (LID) - Subnet-unique, 16-bit ID used to identify and route a packet to an endnode port within a single subnet.
- IP Address - Global-unique 128-bit IPv6 ID used to identify an endnode by applications and to route packets between subnets.
  - Each switch and router has one or more LIDs and IPv6 addresses that are used when it is the destination endnode for management messages.

Link Level Flow Control



- Credit-based link-level flow control
- Link Receivers grant packet receive buffer space credits per VL
- Multiple Virtual Lanes (VLs) on each Physical link provide:
  - Priority arbitration, with VLs assigned priority scheme.
  - Alleviation of head-of-line blocking.

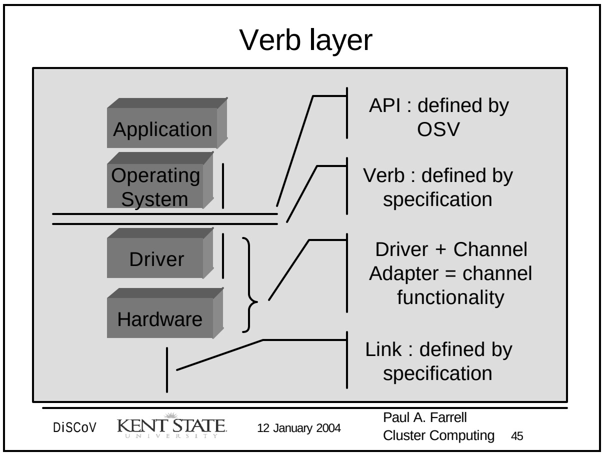
Link Characteristics

- 2.5 Gbaud signaling rate
  - Auto-negotiation for future higher speed signaling
- All links full-duplex
- 1, 4, and 12 wide link widths
- Common backplane connector(s)
  - With auto-negotiation to mutually acceptable width

Link Width	Signaling Rate	Bandwidth	
		Unidirectional	Bi-directional
1	2.5 Gbaud	250 MBytes/s	500 MBytes/s
4	2.5 Gbaud	1 GByte/s	2 GByte/s
12	2.5 Gbaud	3 GBytes/s	6 GBytes/s

Software Characteristics Overview

- Software interface to transport layer
  - Verbs specification
- Major specification topics
  - Partitioning
  - Work Request management
  - Memory management



Verb layer

Application

Operating System

Driver

Hardware

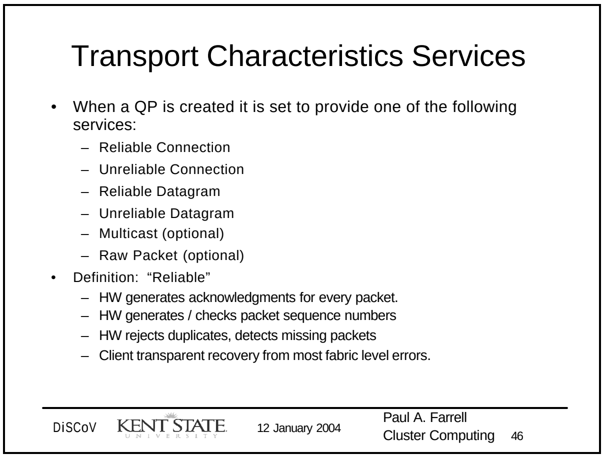
API : defined by OSV

Verb : defined by specification

Driver + Channel Adapter = channel functionality

Link : defined by specification

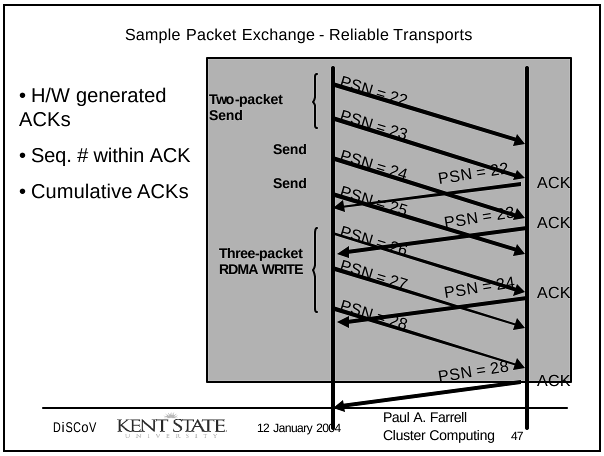
DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 45



Transport Characteristics Services

- When a QP is created it is set to provide one of the following services:
  - Reliable Connection
  - Unreliable Connection
  - Reliable Datagram
  - Unreliable Datagram
  - Multicast (optional)
  - Raw Packet (optional)
- Definition: “Reliable”
  - HW generates acknowledgments for every packet.
  - HW generates / checks packet sequence numbers
  - HW rejects duplicates, detects missing packets
  - Client transparent recovery from most fabric level errors.

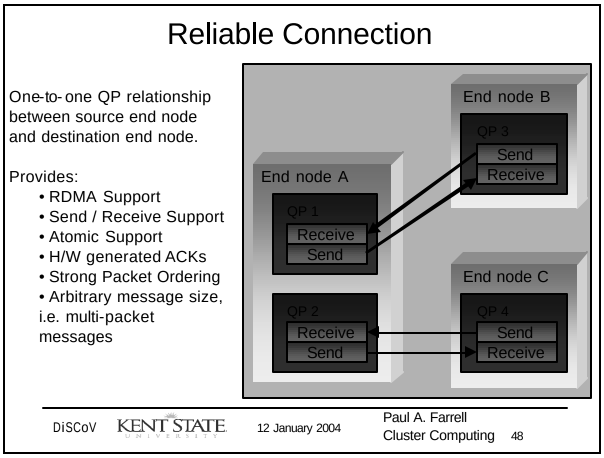
DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 46



Sample Packet Exchange - Reliable Transports

- H/W generated ACKs
- Seq. # within ACK
- Cumulative ACKs

DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 47

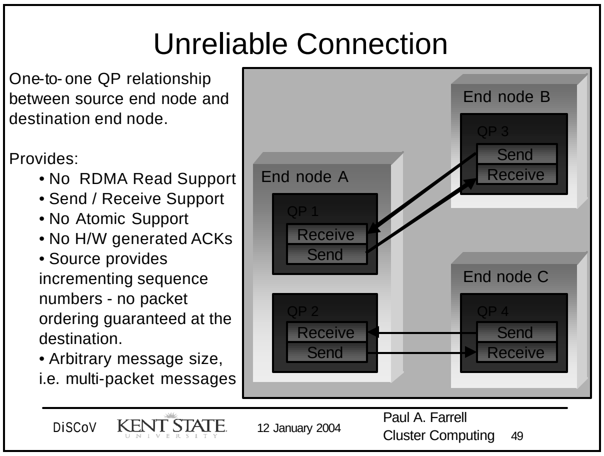


Reliable Connection

One-to-one QP relationship between source end node and destination end node.

Provides:
- RDMA Support
- Send / Receive Support
- Atomic Support
- H/W generated ACKs
- Strong Packet Ordering
- Arbitrary message size, i.e. multi-packet messages

DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 48

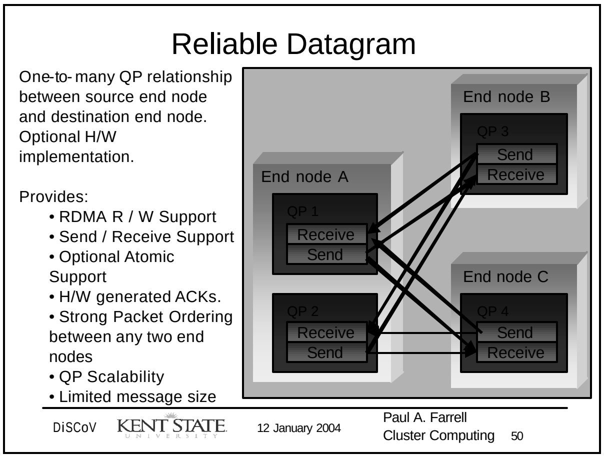


Unreliable Connection

One-to-one QP relationship between source end node and destination end node.

Provides:

- No RDMA Read Support
- Send / Receive Support
- No Atomic Support
- No H/W generated ACKs
- Source provides incrementing sequence numbers - no packet ordering guaranteed at the destination.
- Arbitrary message size, i.e. multi-packet messages

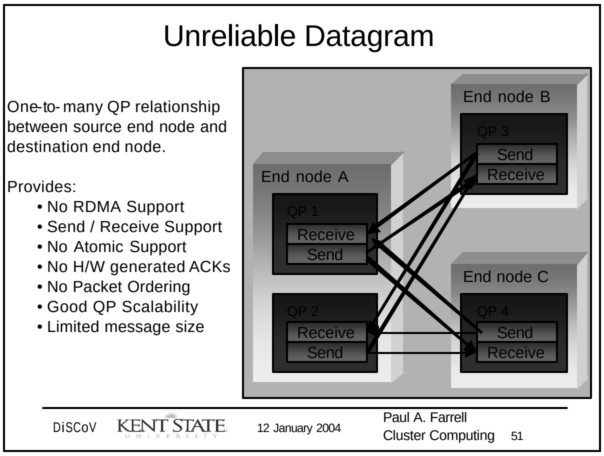


Reliable Datagram

One-to-many QP relationship between source end node and destination end node. Optional H/W implementation.

Provides:

- RDMA R / W Support
- Send / Receive Support
- Optional Atomic Support
- H/W generated ACKs.
- Strong Packet Ordering between any two end nodes
- QP Scalability
- Limited message size

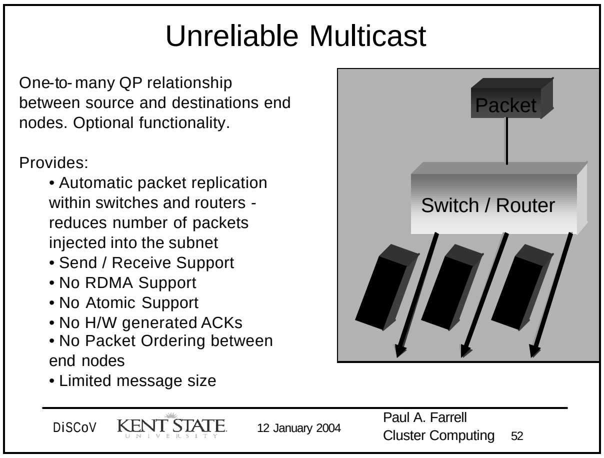


Unreliable Datagram

One-to-many QP relationship between source end node and destination end node.

Provides:

- No RDMA Support
- Send / Receive Support
- No Atomic Support
- No H/W generated ACKs
- No Packet Ordering
- Good QP Scalability
- Limited message size

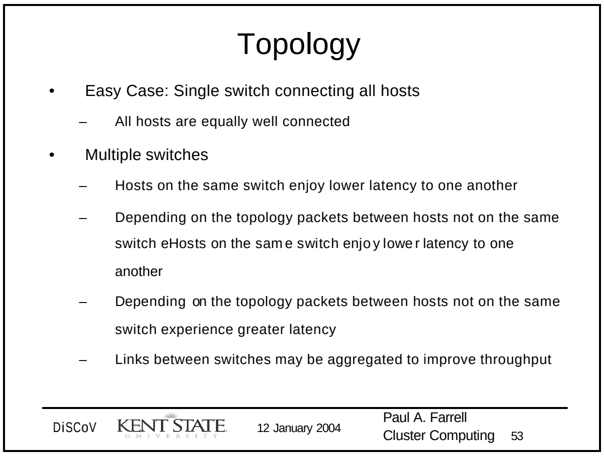


Unreliable Multicast

One-to-many QP relationship between source and destinations end nodes. Optional functionality.

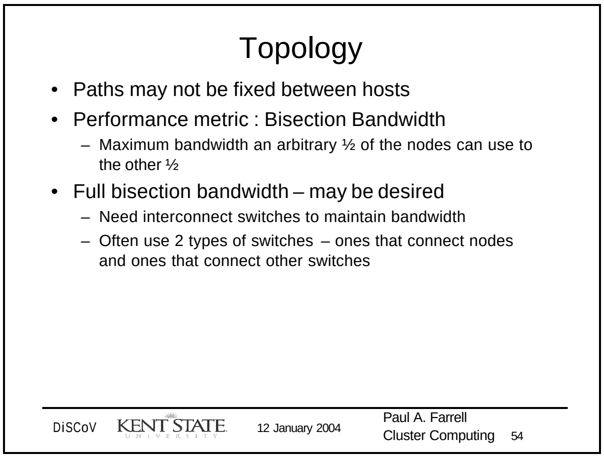
Provides:

- Automatic packet replication within switches and routers - reduces number of packets injected into the subnet
- Send / Receive Support
- No RDMA Support
- No Atomic Support
- No H/W generated ACKs
- No Packet Ordering between end nodes
- Limited message size



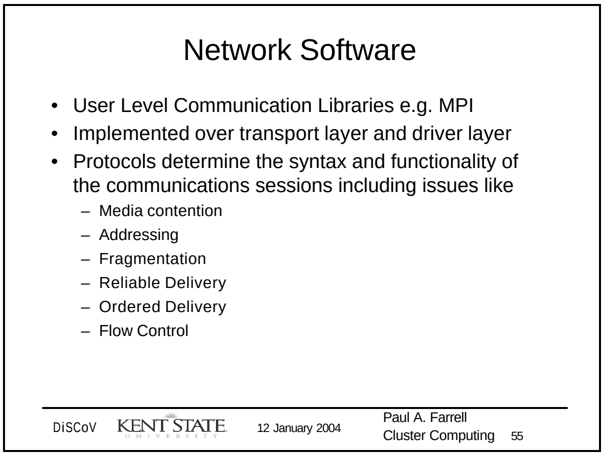
Topology

- Easy Case: Single switch connecting all hosts
  - All hosts are equally well connected
- Multiple switches
  - Hosts on the same switch enjoy lower latency to one another
  - Depending on the topology packets between hosts not on the same switch eHosts on the sam e switch enjo y lowe r latency to one another
  - Depending on the topology packets between hosts not on the same switch experience greater latency
  - Links between switches may be aggregated to improve throughput



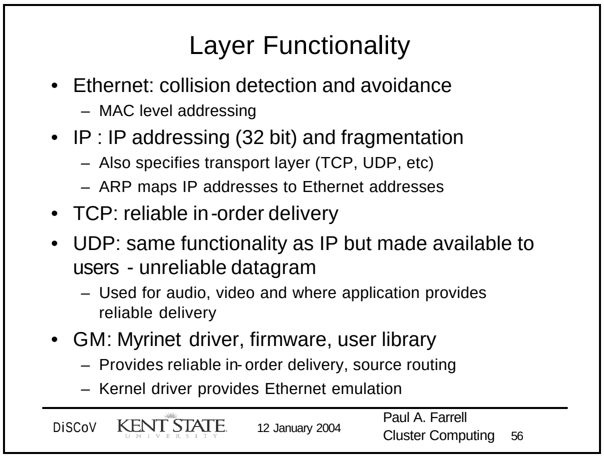
Topology

- Paths may not be fixed between hosts
- Performance metric : Bisection Bandwidth
  - Maximum bandwidth an arbitrary ½ of the nodes can use to the other ½
- Full bisection bandwidth – may be desired
  - Need interconnect switches to maintain bandwidth
  - Often use 2 types of switches – ones that connect nodes and ones that connect other switches



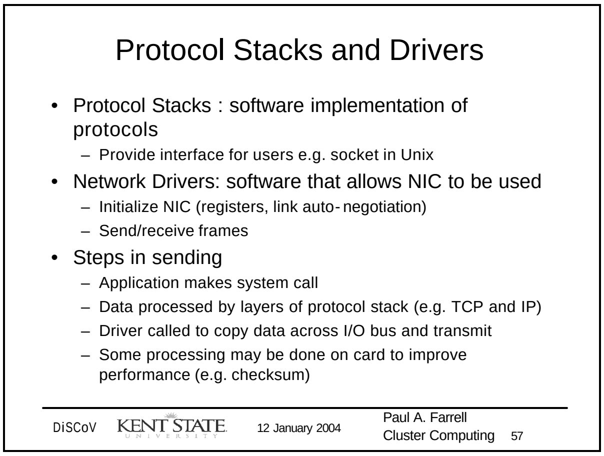
Network Software

- User Level Communication Libraries e.g. MPI
- Implemented over transport layer and driver layer
- Protocols determine the syntax and functionality of the communications sessions including issues like
  - Media contention
  - Addressing
  - Fragmentation
  - Reliable Delivery
  - Ordered Delivery
  - Flow Control



Layer Functionality

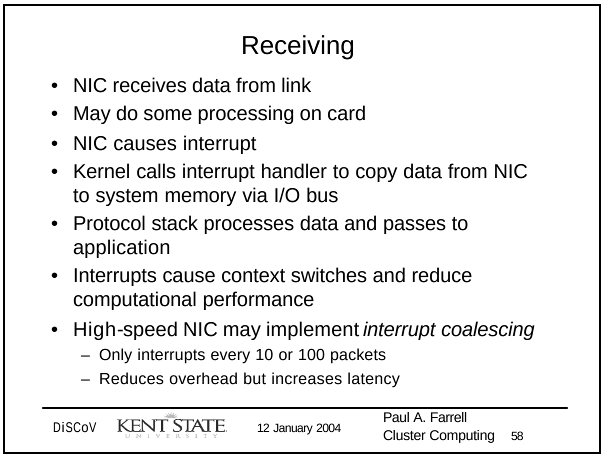
- Ethernet: collision detection and avoidance
  - MAC level addressing
- IP : IP addressing (32 bit) and fragmentation
  - Also specifies transport layer (TCP, UDP, etc)
  - ARP maps IP addresses to Ethernet addresses
- TCP: reliable in-order delivery
- UDP: same functionality as IP but made available to users - unreliable datagram
  - Used for audio, video and where application provides reliable delivery
- GM: Myrinet driver, firmware, user library
  - Provides reliable in-order delivery, source routing
  - Kernel driver provides Ethernet emulation



Protocol Stacks and Drivers

- Protocol Stacks : software implementation of protocols
  - Provide interface for users e.g. socket in Unix
- Network Drivers: software that allows NIC to be used
  - Initialize NIC (registers, link auto-negotiation)
  - Send/receive frames
- Steps in sending
  - Application makes system call
  - Data processed by layers of protocol stack (e.g. TCP and IP)
  - Driver called to copy data across I/O bus and transmit
  - Some processing may be done on card to improve performance (e.g. checksum)

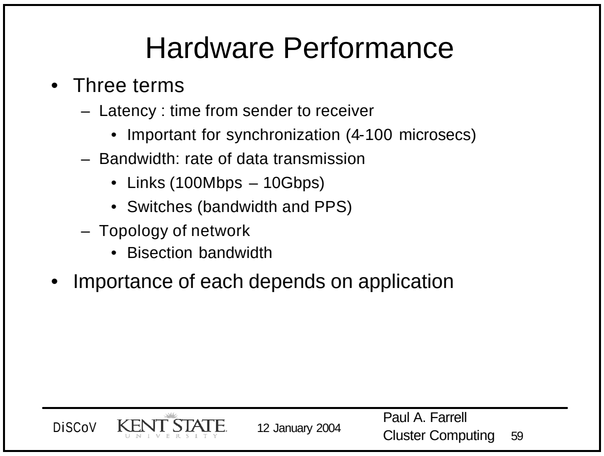
DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 57



Receiving

- NIC receives data from link
- May do some processing on card
- NIC causes interrupt
- Kernel calls interrupt handler to copy data from NIC to system memory via I/O bus
- Protocol stack processes data and passes to application
- Interrupts cause context switches and reduce computational performance
- High-speed NIC may implement *interrupt coalescing*
  - Only interrupts every 10 or 100 packets
  - Reduces overhead but increases latency

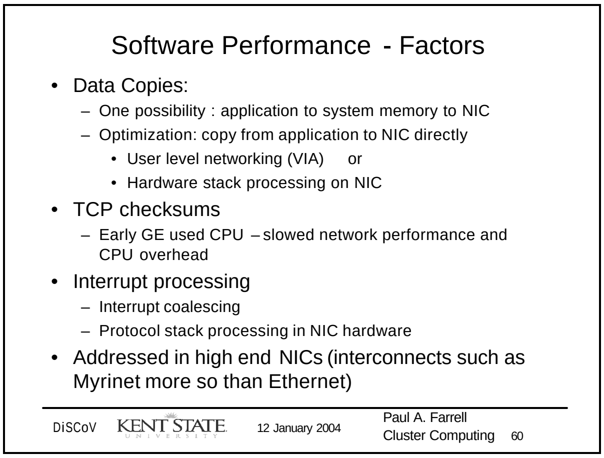
DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 58



Hardware Performance

- Three terms
  - Latency : time from sender to receiver
    - Important for synchronization (4-100 microsecs)
  - Bandwidth: rate of data transmission
    - Links (100Mbps – 10Gbps)
    - Switches (bandwidth and PPS)
  - Topology of network
    - Bisection bandwidth
- Importance of each depends on application

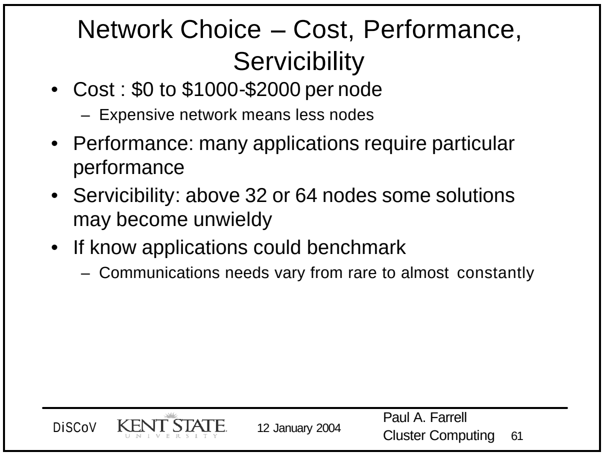
DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 59



Software Performance - Factors

- Data Copies:
  - One possibility : application to system memory to NIC
  - Optimization: copy from application to NIC directly
    - User level networking (VIA) or
    - Hardware stack processing on NIC
- TCP checksums
  - Early GE used CPU – slowed network performance and CPU overhead
- Interrupt processing
  - Interrupt coalescing
  - Protocol stack processing in NIC hardware
- Addressed in high end NICs (interconnects such as Myrinet more so than Ethernet)

DiSCoV KENT STATE UNIVERSITY 12 January 2004 Paul A. Farrell Cluster Computing 60



Network Choice – Cost, Performance, Servicibility

- Cost : $0 to $1000-$2000 per node
  - Expensive network means less nodes
- Performance: many applications require particular performance
- Servicibility: above 32 or 64 nodes some solutions may become unwieldy
- If know applications could benchmark
  - Communications needs vary from rare to almost constantly