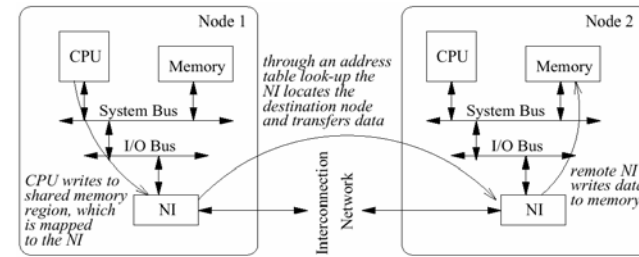


## System Area Networks (SANs)

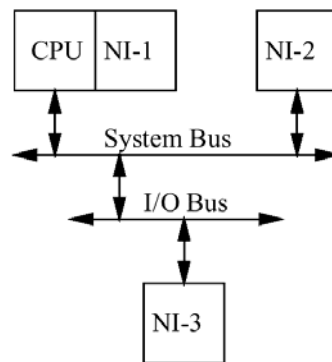
- Hardware
  - Nodes: Network Interface Card (NIC) on I/O bus (PCI, PCI-X, PCI-Express) or maybe on motherboard
  - Components
    - Hardware to interface with physical layer of network (copper or fiber)
    - Hardware to interface with I/O bus
  - Transmission rate limited by speed of I/O bus and network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

## Data Transfer Process



## 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
  - iWarp, 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 (PCI-X or PCI-Express) based NI
  - use on any system w/ PCI(-X etc) 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

- Ethernet -1975 Xerox /1980 standard
  - 10 Mbps
  - CSMA/CD (Carrier Sense Multiple Access with Collision Detection)
- HiPPI (High Performance Parallel Interface) - 1980
  - copper-based, 800/1600 Mbps over 32/64 bit lines
  - point-to-point channel
  - used a massive 50-pair cable with bulky connectors, and had limited cable lengths
- Fast Ethernet – 1995 standard IEEE 802.3
  - 100 Mbps
  - CSMA/CD (Carrier Sense Multiple Access with Collision Detection)

## Links

- ATM (Asynchronous Transfer Mode) – 1980s
  - connection-oriented packet switching
  - fixed length (53 bytes cell)
  - suitable for WAN, 155/622 Mbps
- Fibre Channel 1985/1994 standard
  - Designed as replacement for HiPPI
  - Also supported SCSI connections of disks
  - Primarily used for storage connection
  - 200Mbps – 2000Mbps
- SCI (Scalable Coherent Interface) - 1992
  - IEEE standard 1596, hardware DSM support, 400MBs

## Links

- ServerNet
  - 1 Gbps
  - originally, interconnection for high bandwidth I/O
- Myrinet - 1998
  - programmable microcontroller
  - 1.28 Gbps – 2 Gbps
- Memory Channel
  - 800 Mbps
  - virtual shared memory
  - strict message ordering
- Infiniband - 2000
  - 10 Gbps initially

## 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 free physical layer
    - CRC can be computed by NIC itself
    - error signaling (interrupt or status registers)
    - NIC side buffer

## Network Speeds

[http://en.wikipedia.org/wiki/List\\_of\\_device\\_bandwidths](http://en.wikipedia.org/wiki/List_of_device_bandwidths)

Architecture	Speed Mbps	Speed MB
LocalTalk	0.230 Mbit/s	0.0288 MB/s
Econet	0.800 Mbit/s	0.1 MB/s
PC-Network	2 Mbit/s	0.25 MB/s
ARCNET (Standard)	2.5 Mbit/s	0.3125 MB/s
Ethernet Experimental	3 Mbit/s	0.375 MB/s
Token Ring (Original)	4 Mbit/s	0.5 MB/s
Ethernet (10base-X)	10 Mbit/s	1.25 MB/s
Token Ring (Later)	16 Mbit/s	2 MB/s
ARCnet Plus	20 Mbit/s	2.5 MB/s
Token Ring IEEE 802.5t	100 Mbit/s	12.5 MB/s
Fast Ethernet (100base-X)	100 Mbit/s	12.5 MB/s
FDDI	100 Mbit/s	12.5 MB/s

## Network Speeds

Architecture	Speed Mbps	Speed MB
HIPPI	800 Mbit/s	100 MB/s
Token Ring IEEE 802.5v (no known implementations)	1,000 Mbit/s	125 MB/s
Gigabit Ethernet (1000base-X)	1,000 Mbit/s	125 MB/s
Myrinet 2000	2,000 Mbit/s	250 MB/s
Infiniband SDR 1X[17]	2,000 Mbit/s	250 MB/s
Quadrics QeNett	3,600 Mbit/s	450 MB/s
Infiniband DDR 1X[17]	4,000 Mbit/s	500 MB/s
Infiniband QDR 1X[17]	8,000 Mbit/s	1,000 MB/s
Infiniband SDR 4X[17]	8,000 Mbit/s	1,000 MB/s
Quadrics QeNettII	8,000 Mbit/s	1,000 MB/s

## Network Speeds

Architecture	Speed Mbps	Speed MB
10 gigabit Ethernet (10Gbase-X)	10,000 Mbit/s	1,250 MB/s
Myri 10G	10,000 Mbit/s	1,250 MB/s
Infiniband DDR 4X[17]	16,000 Mbit/s	2,000 MB/s
Scalable Coherent Interface (SCI) Dual Channel SCI, x8 PCIe	20,000 Mbit/s	2,500 MB/s
Infiniband SDR 12X[17]	24,000 Mbit/s	3,000 MB/s
Infiniband QDR 4X[17]	32,000 Mbit/s	4,000 MB/s
Infiniband DDR 12X[17]	48,000 Mbit/s	6,000 MB/s
Infiniband QDR 12X[17]	96,000 Mbit/s	12,000 MB/s
100 gigabit Ethernet (100Gbase-X)	100,000 Mbit/s	12,500 MB/s

## Network 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 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 half of the nodes can use to the other half
- 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)

## 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

## Hardware Performance

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

### 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 caused 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)

### 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 have know applications could benchmark
  - Communications needs vary from rare to almost constantly