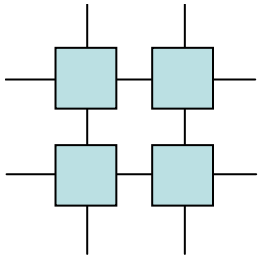
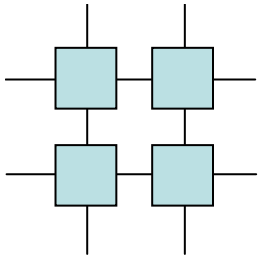
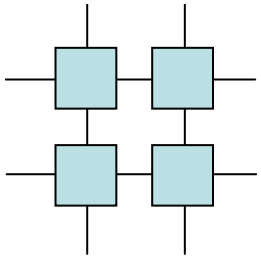


Interconnect Technologies for Clusters



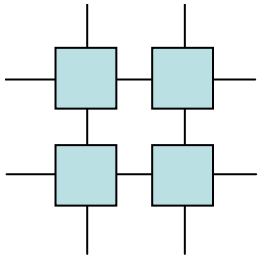
Interconnect approaches

- WAN
  - ’infinite distance’
- LAN
  - Few kilometers
- SAN
  - Few meters
- Backplane
  - Not scalable



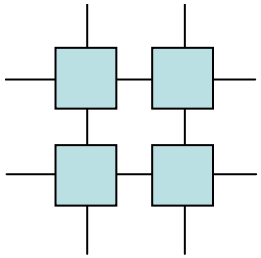
Physical Cluster Interconnects

- FastEther
- Gigabit EtherNet
- 10 Gigabit EtherNet
- ATM
- cLan
- Myrinet
- Memory Channel
- SCI
- Atoll
- ServerNet



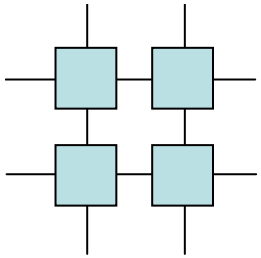
Switch technologies

- Switch design
  - Fully interconnected
  - Omega
- Package handling
  - Store and forward
  - Cut-through routing (worm-hole routing)



Implications of switch technologies

- Switch design
  - Affects the constant associated with routing
- Package handling
  - Affects the overall routing latency in a major may



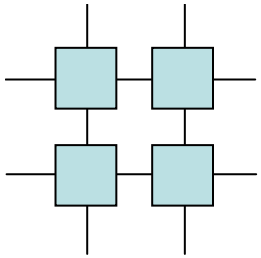
Store-and-fwd vs. Worm-hole one step

- $T(\nu)$ =Overhead+Channel+Time Routing Delay
- Cut through:

$$12us + \frac{128B}{100\frac{b}{us}} + 0.1us = 22.34us$$

- Store ’n fw:

$$12us + \frac{128B}{100\frac{b}{us}} + 0.1us + \frac{128B}{100\frac{b}{us}} = 32.58us$$



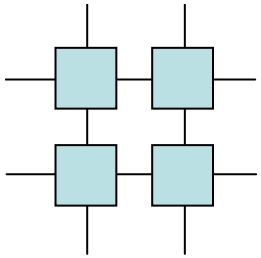
Store-and-fwd vs. Worm-hole ten steps

- $T(\nu) =$Overhead+Channel+Time Routing Delay
- Cut through:

$$12us + \frac{128B}{100\frac{b}{us}} + 10 \cdot 0.1us = 23.24us$$

- Store ’n fw:

$$12us + 10 \cdot \left(\frac{128B}{100\frac{b}{us}} + 0.1us \right) + \frac{128B}{100\frac{b}{us}} = 125.64us$$



FastEther

- 100 Mbit/sec

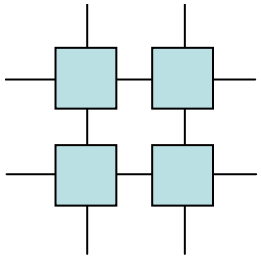
+ Generally supported

+ Extremely cheap

- Limited bandwidth

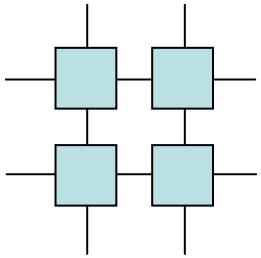
- Not really that standard

- Not all implementations support zero-copy protocols



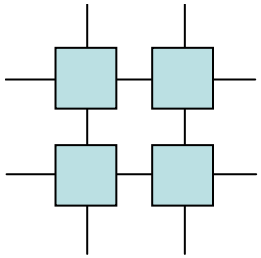
Gigabit EtherNet

- Ethernet is hype-only at this stage
- Bandwidth really is 1Gb/sec
- Latency is only slightly improved
  - Down to 20us from 22us in 100Mb
- Current standard
  - But NICs are as different as with FE



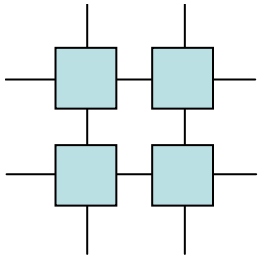
10 Gigabit EtherNet

- Not specified yet
  - But most producers already run demonstration setup
  - Several sells 10Gb
- Target applications not really defined
  - But clusters are not the most likely customers
  - Perhaps as backbone for large clusters
- Optical interconnects only
  - Copper currently being proposed



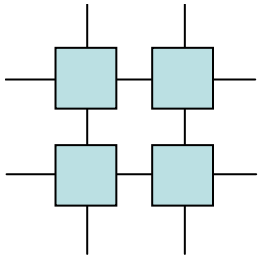
ATM

- Used to be the holy grail in cluster computing
- Turns out to be poorly suited for clusters
  - High price
  - Tiny packages
  - Designed for throughput not reliability



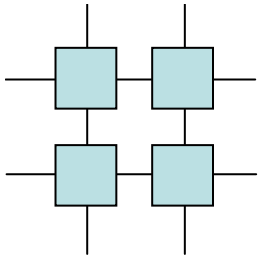
cLAN

- Virtual Interface Architecture
- API standard not HW standard
- 1.2 Gbit/sec



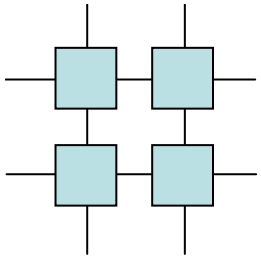
Myrinet

- Long time ’defacto-standard’
- LAN and SAN architectures
- Switch-based
- Extremely programmable



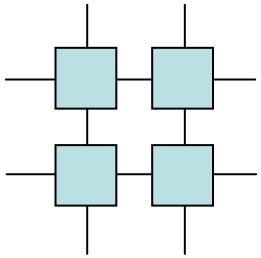
Myrinet

- Very high bandwidth
  - 0.64Gb + 0.64 Gb in gen 1 (1994)
  - 1.28Gb + 1.28 Gb in gen 2 (1997)
  - 2.0Gb + 2.0 Gb in gen 3 (2000)
- 18 bit parallel wires
- Error-rate at 1bit per 24 hours
- Very limited physical distance



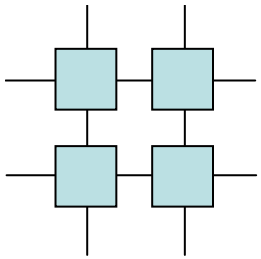
Myrinet Interface

- Hosts a fast RISC processor
  - 132 MHz in newest version
- Large memory onboard
  - 2,4 or 8MB in newest version
- Memory is used as both send and recieve buffers and run at CPU speed
  - 7.5ns in newest version



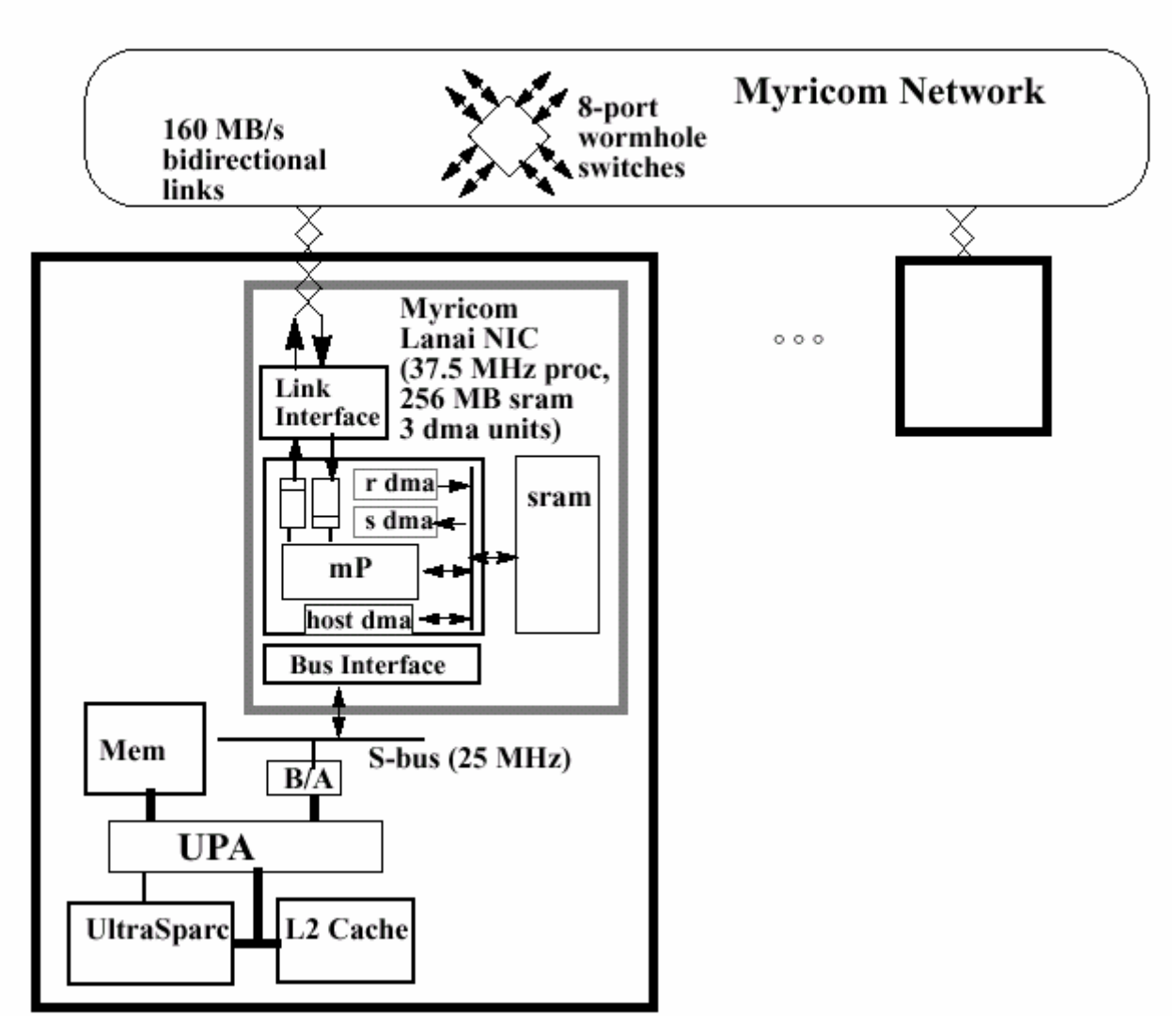
Myrinet-switch

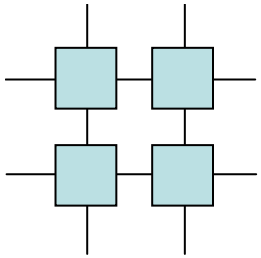
- Worm-hole routed
  - 5 ns route time
- Process to process
  - 9us (133 MHz LANai)
  - 7us (200 MHz LANai)



Cluster Computing

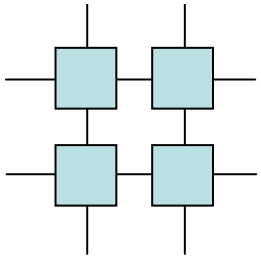
Myrinet





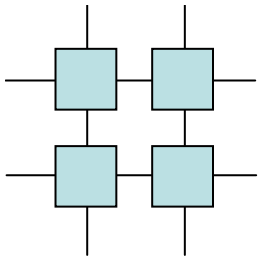
Myrinet Prices

- PCI/SAN interface
  - $995, $1295, $1595
- SAN Switch
  - 8 port $3250
  - 16 port $4825
- 10 ft. cable $190



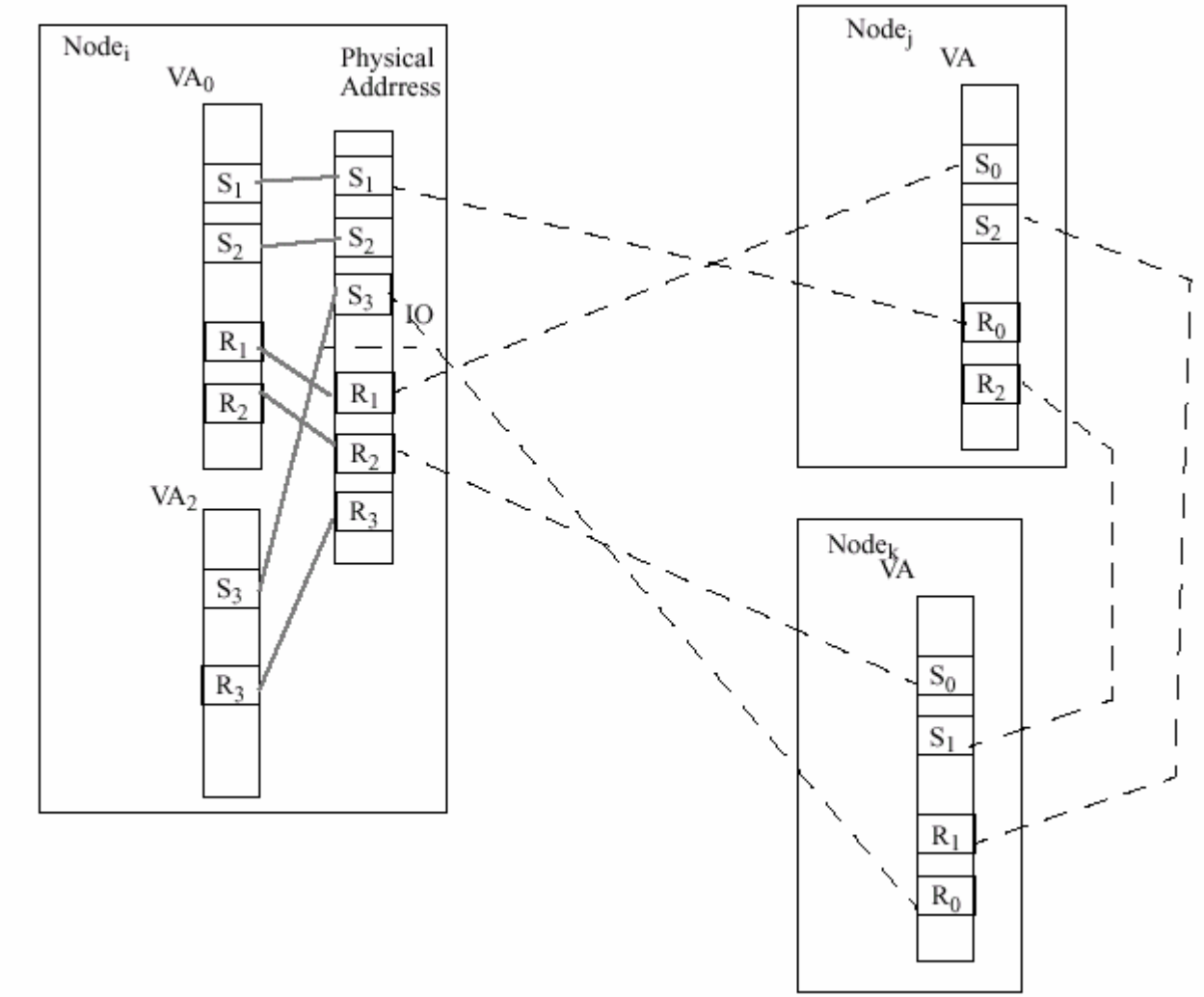
Memory Channel

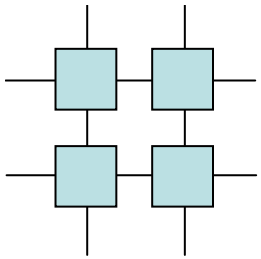
- Digital Equipment Corporation product
- Raw performance:
  - Latency 2.9 us
  - Bandwidth 64 MB/s
- MPI performance
  - Latency 7 us
  - Bandwidth 61 MB/s



Cluster Computing

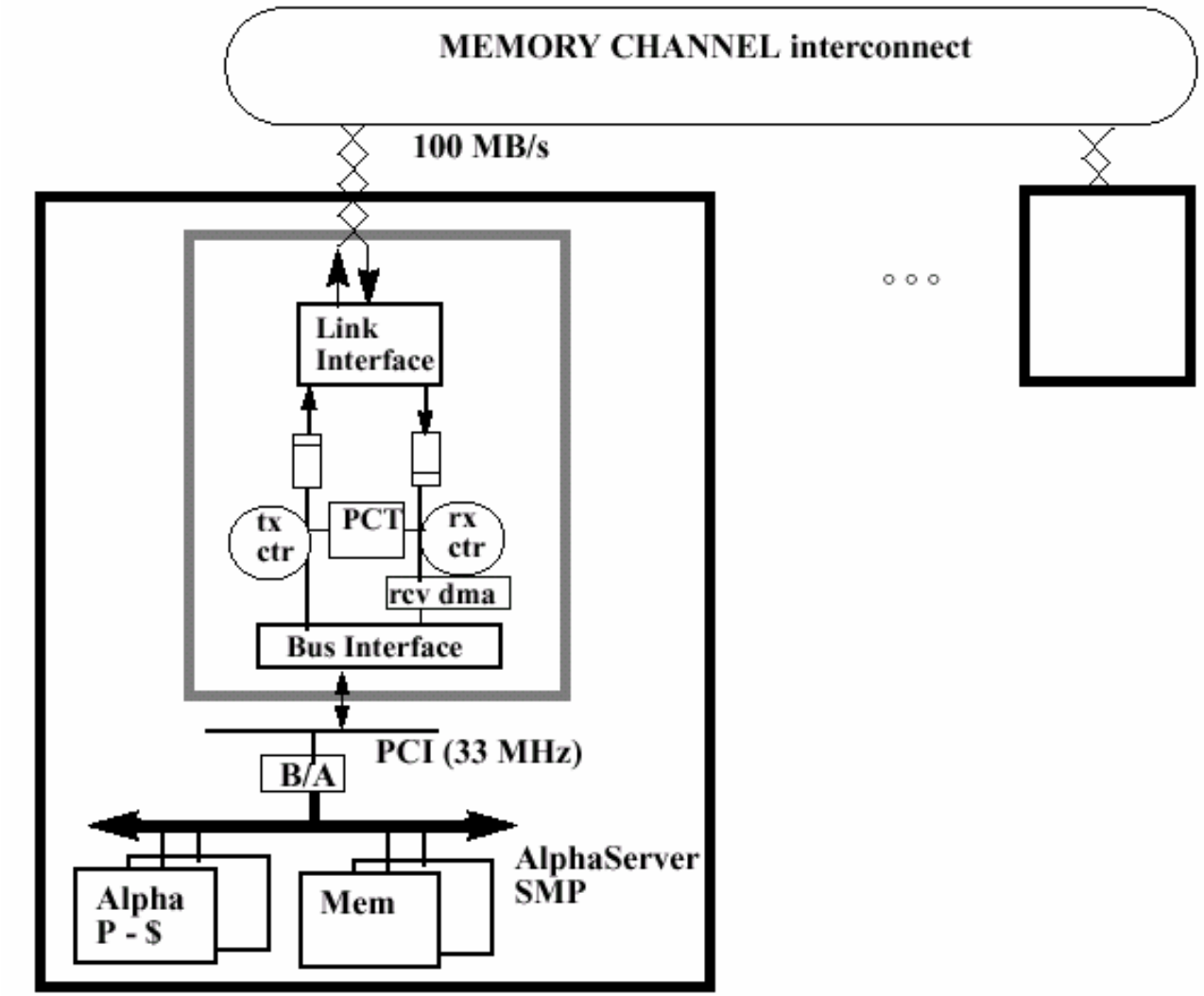
Memory Channel

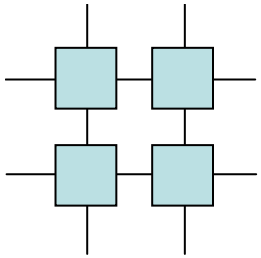




Cluster Computing

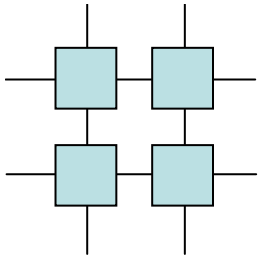
Memory Channel



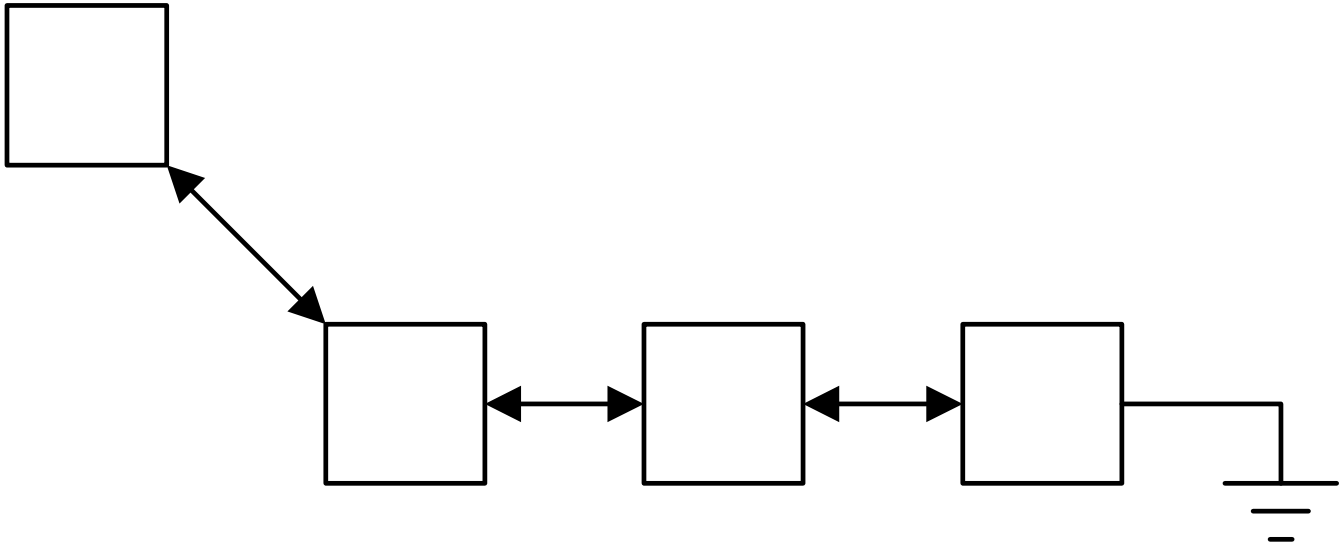


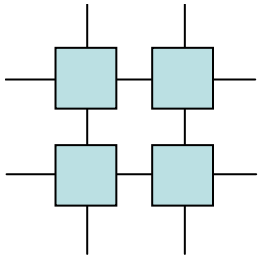
SCI

- Scalable Coherent Interface
- IEEE standard
- Not widely implemented
- Coherency protocol is very complex
  - 29 stable states
  - An enourmous amount of transient states



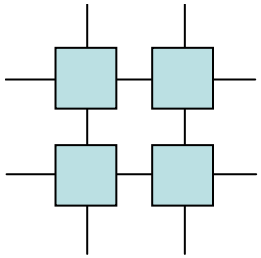
SCI





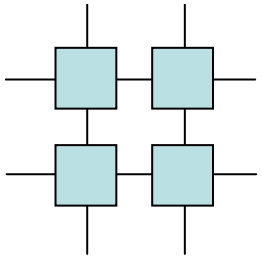
SCI Coherency

- States
  - *Hom*e: no remote cache in the system contains a copy of the block
  - *Fres*h: one or more remote caches may have a read-only copy, and the copy in memory is valid.
  - *Gon*e: another remote cache contains a writeable copy. There is no valid copy on the local node.



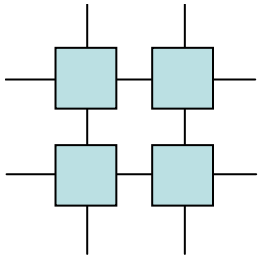
SCI Coherency

- State is named by two components
  - ONLY
  - HEAD
  - TAIL
  - MID
  - Dirty: modified and writable
  - Clean: unmodified (same as memory) but writable
  - Fresh:data may be read, but not written until memory is informed
  - Copy: unmodified and readable



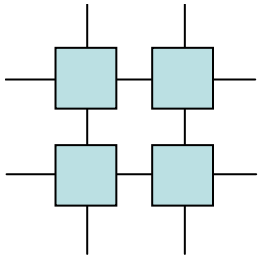
SCI Coherency

- *List construction*: adding a new node (sharer) to the head of a list
- *Rollout*: removing a node from a sharing list, which requires that a node communicate with its upstream and downstream neighbors informing them of their new neighbors so they can update their pointers
- *Purging (invalidation)*: the node at the head may purge or invalidate all other nodes, thus resulting in a single-element list. Only the head node can issue a purge.



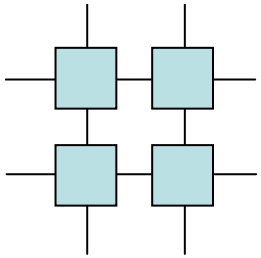
Atoll

- University research project
- Should be very fast and very cheap
- Keeps comming ’very soon now’
- I have stopped waiting



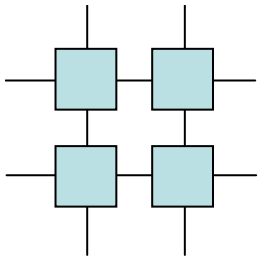
Atoll

- Grid architecture
- 250 MB/sec bidirectional links
  - 9 bit
  - 250MHz clock

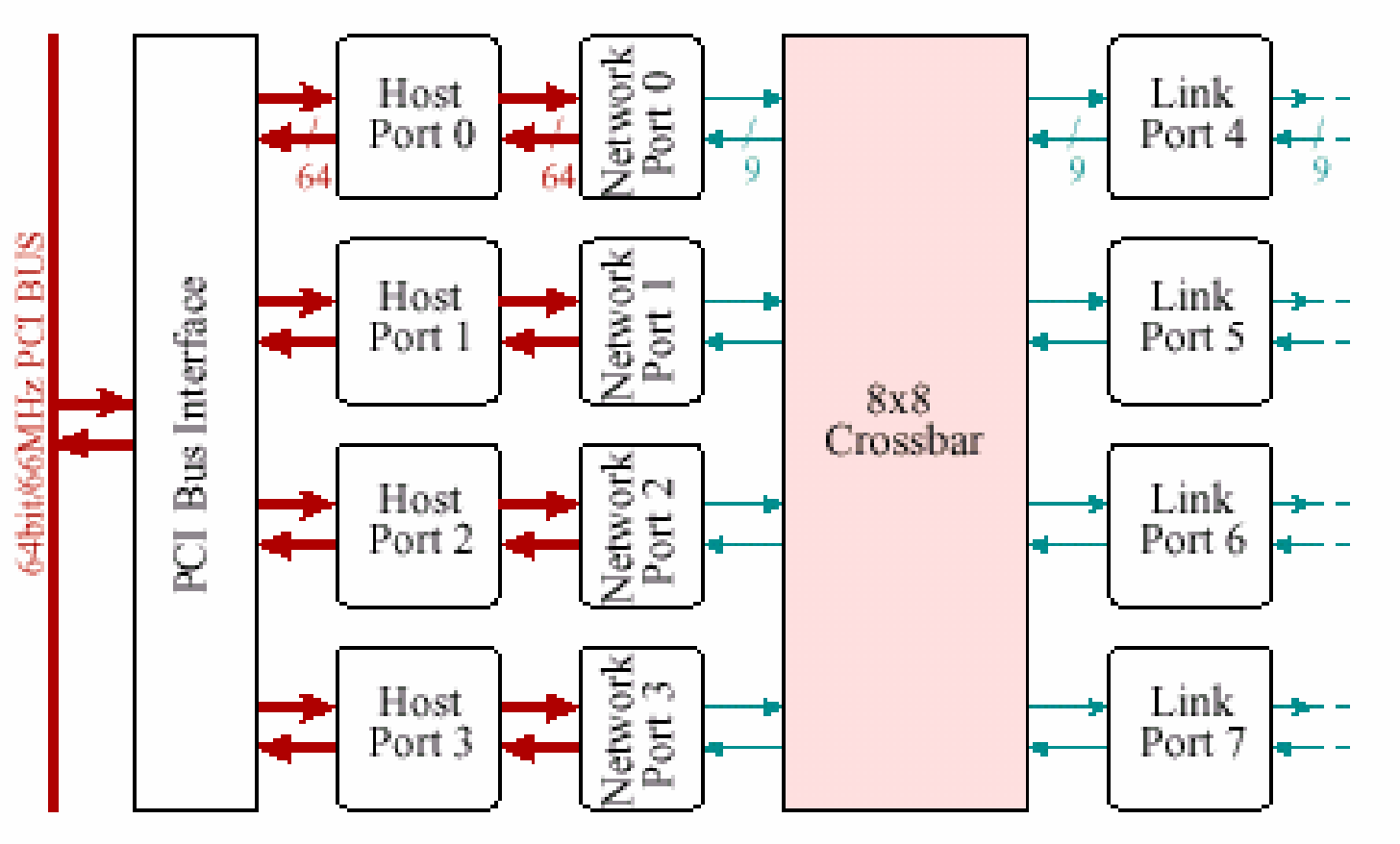


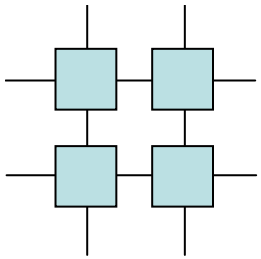
Atoll

msg size (byte)	n_{snd}	t_{atoll} (μs)	n_{rcv}	$T_{start}^{n+1} - T_{start}^{n}$ (μs)	latency (μs)	single NI BW (Mbyte/s)	acc. BW (Mbyte/s)
32	3	2.9	3	2.2	5.6	10.8	43.2
256	3	4.9	6	3.9	9.9	54.2	216.8
512	4	7.3	8	6.1	13.2	70.7	282.8
1024	6	12.9	14	10.6	21.8	82.7	330.8
4096	18	39.4	32	36.8	72.6	96.6	386.4

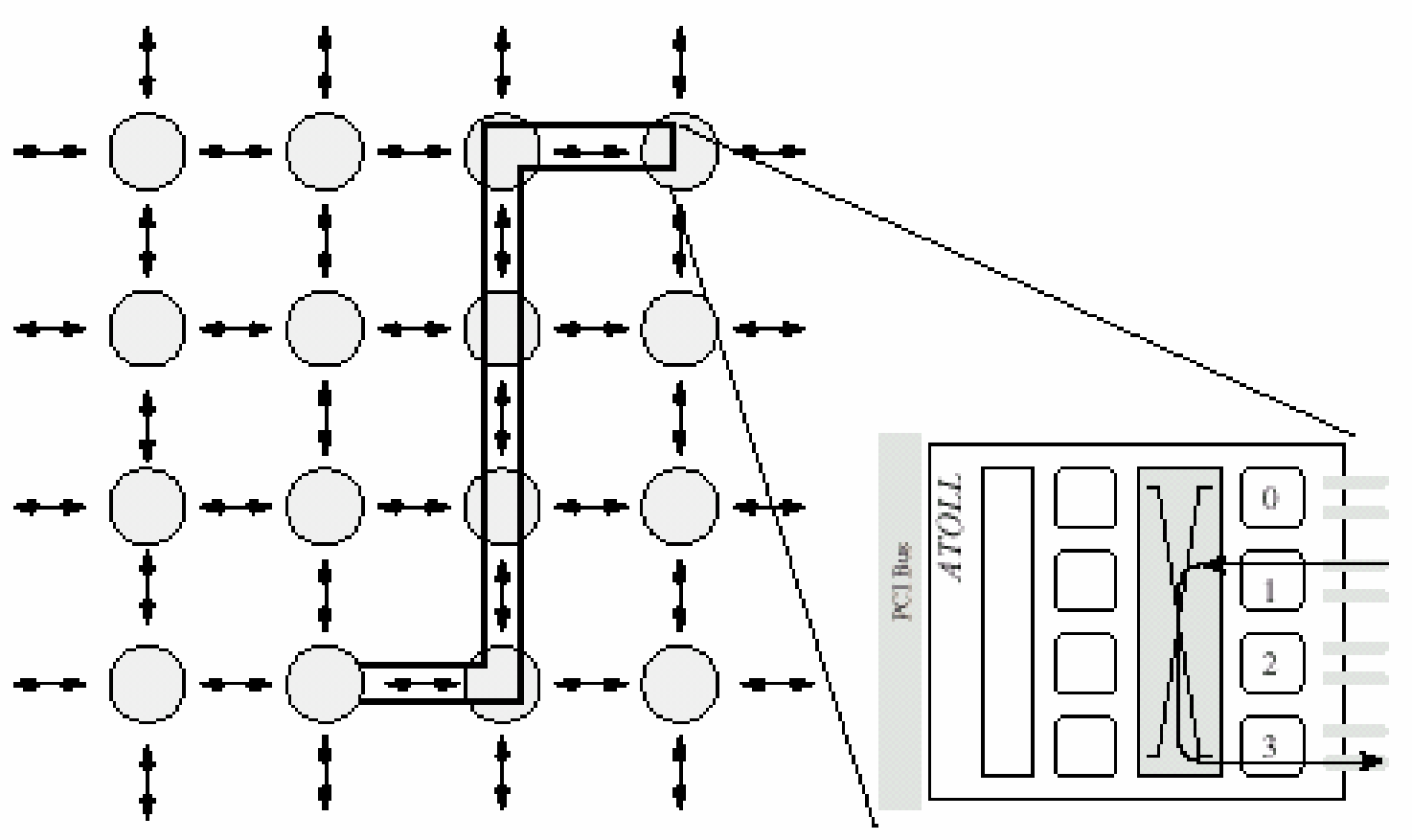

Cluster Computing

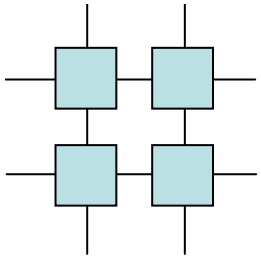
Atoll


64bit/66MHz PCI BUS
PCI Bus Interface
Host Port 0
Host Port 1
Host Port 2
Host Port 3
64
64
Network Port 0
Network Port 1
Network Port 2
Network Port 3
9
9
8x8 Crossbar
Link Port 4
Link Port 5
Link Port 6
Link Port 7
9



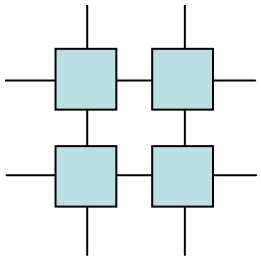
Atoll



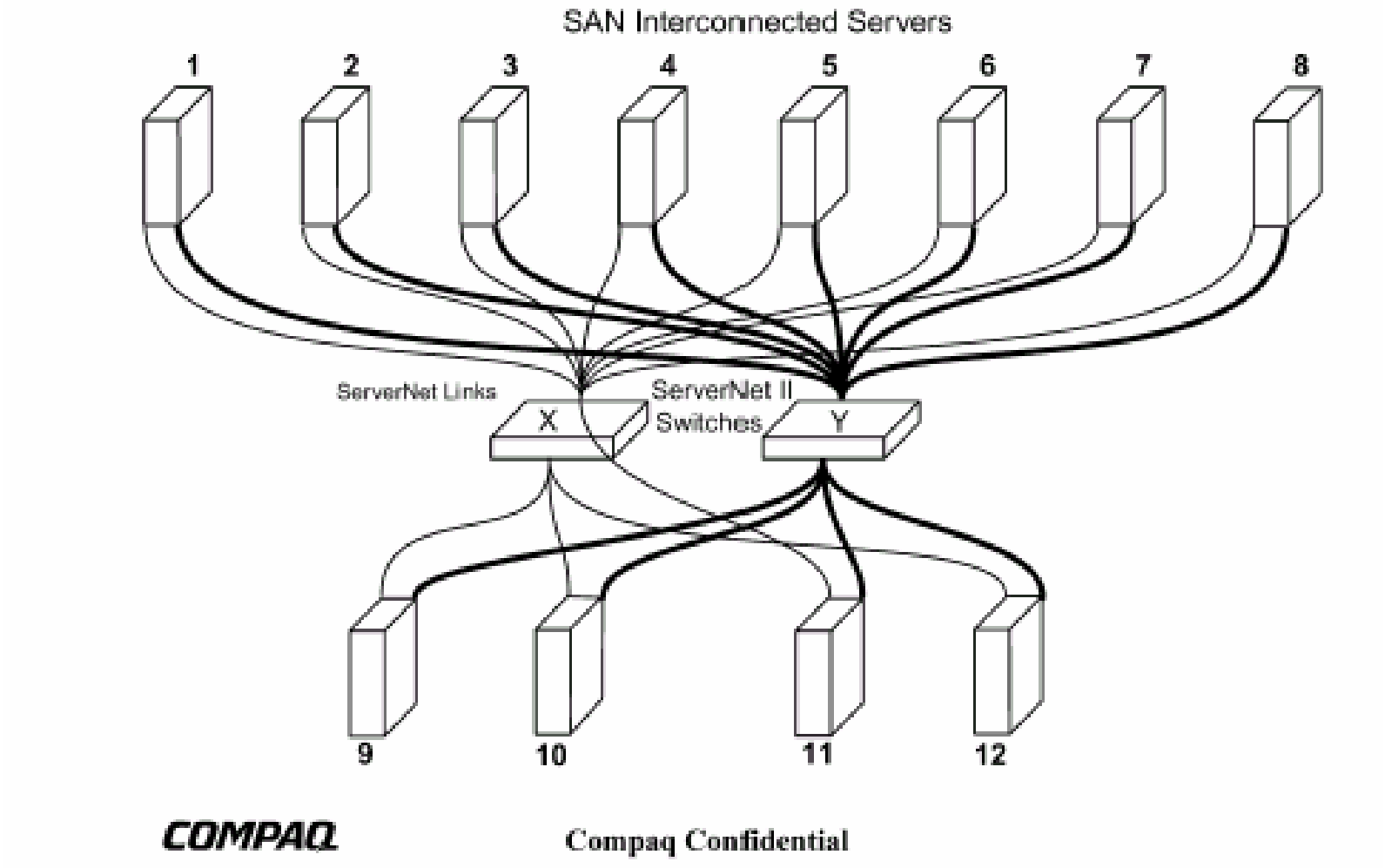


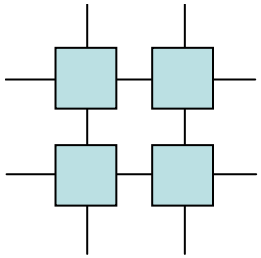
Servernet-II

- Supports 64-bit, 66-MHz PCI
- Bidirectional links
  - 1.25+1.25Gbit/sec
- VIA compatible



Servernet II



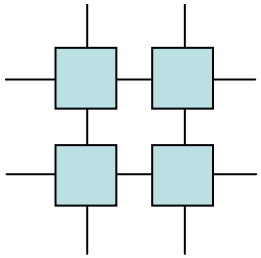


Servernet-II

ServerNet II Performance Summary*

Performance Test:	Measured
8 byte delivery latency with Send/Receive poll	12 µs
8 byte delivery latency with Send/Receive wait	32 µs
64 byte CPU cost with Send/Receive wait	27 µs
64 byte CPU cost with Lazy Send/Receive wait¶	29 µs
64K 1-way throughput RDMA writes (1VI-4 VIs)	92-132 MB/s
64K 1-way throughput RDMA reads (1 VI-4 VIs)	129-134 MB/s
64K 2-way throughput RDMA (reads-writes)	181-194 MB/s
64K RDMA throughput test CPU utilization	~0%

** This number was measured on 33MHz, 32-bit PCI with a 500MHz CPU.*



Infiniband

- New standard
  - Keeps coming any day now
- An extension of PCI-X
  - 1x = 2.5Gbps
  - 4x = 10Gbps – current standard
  - 12x = 30Gbps