

High Perf. I/Cs

- Myrinet
- Quadrics
- Ethernet vs. Ethernet

Myrinet

Origins:

- based on MPP architecture – multicomputer parallel network
 - packet-based
 - regular topology – i.e., can figure out address mapping
 - cut-through routing
 - flow control on every link – asynchronous signals
 - low error rate assumption
- Caltech Mosaic multicomputer

Myrinet Design

- point-to-point, host NIs and switches
- flow control w/ stop and go signals
 - stop and go at threshold values, not at min/max to allow for some slack
- it's own frame format
 - header, payload (arbitrary length!), CRC at the end, on entire packet, interpacket gap...
- blocking-cut-through routing

Myrinet NICs

- LaNai chips
 - SRAM, bus (originally proprietary E-Bus, LaNai-X with PCI-X), DMA engine, packet engine
 - DMA engine can compute checksum on data transferred on bus
 - programmable processor – Myrinet Control Program (in memory write protected by LaNai processor).

Host interface

- command and acknowledgment queues
- scatter/gather DMA
- 1 vs. zero-copy data transfers
 - user space to NIC memory, checksum by DMA engine
 - packet in multiple dma transfers...
- interrupt enable/disable on rcv.
- automatically select address mappers in network
 - Self configuring/self healing
- IP multicast support

Applications

sockets, MPI, other

user-level API

kernel agent

GM or MX

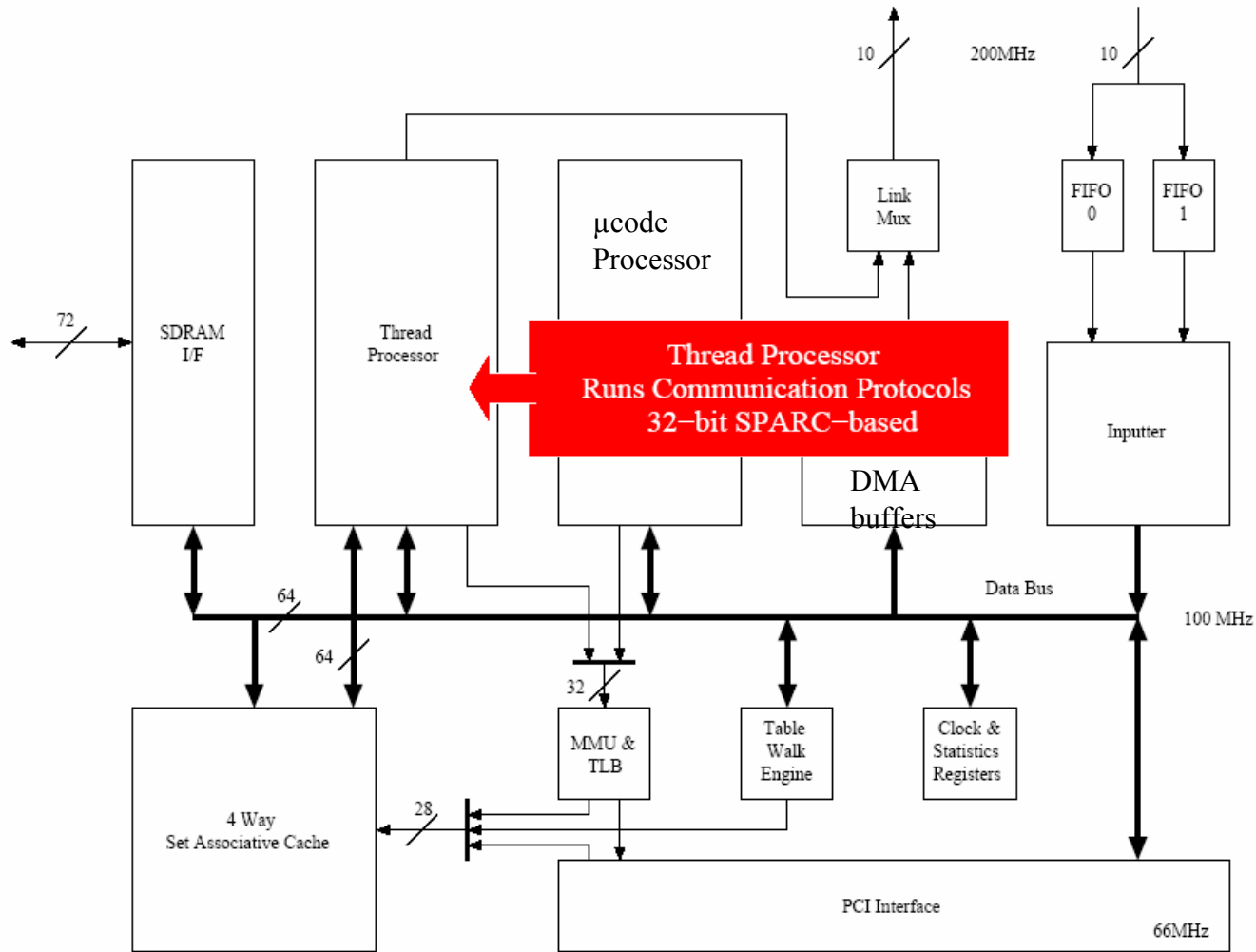
MCR

- third generation 2x2Gbps
- fourth 2x10Gbps, 10GbE compliant
- LaNai-X – multiple interconnects supported in firmware (Myrinet, GbE, Infiniband).
 - PCI-X, plus proprietary SAN interface...
- Current top - #1 1, 4800 processors, 9.6TB Mm

Quadrics Overview

- Quadrics Interconnection Network QsNet (QsNet II)
 - Integrated global shared memory
 - Programmable NIC
 - Hardware support for collectives, integrated fault detection and fault tolerance in network hw
 - Currently -- #5, 8704 processors, 26.1TB Mm
- Based on Elan NICs and Elite switches

Elan NICs



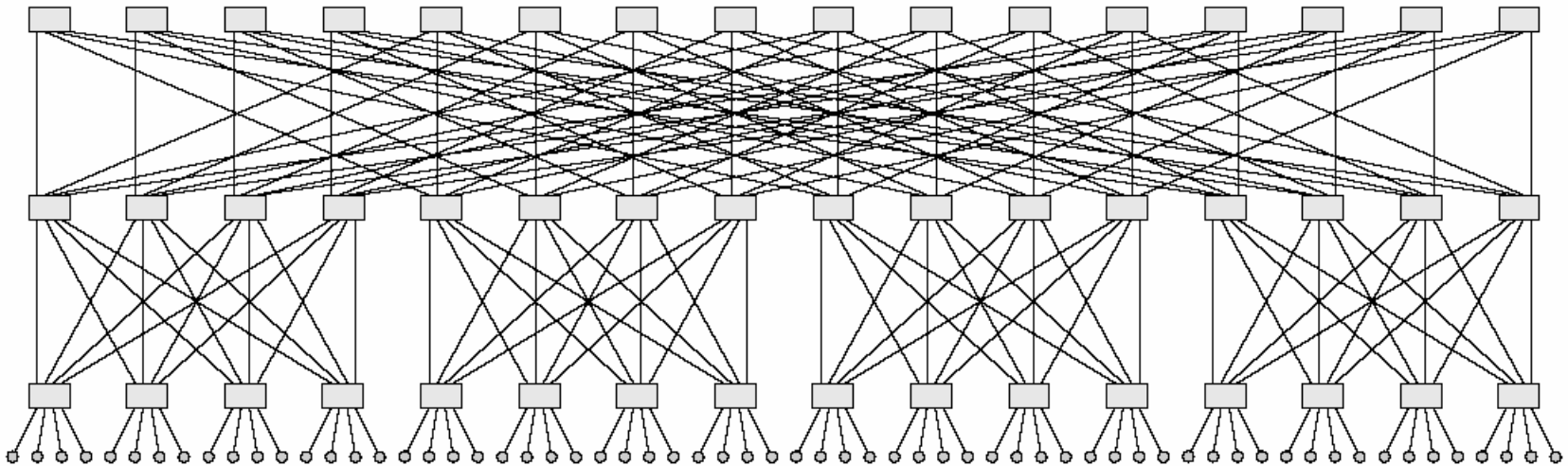
- μcode Proc threads: inputter, DMA thread, processor-scheduling thread, command-processor thread

- MMU/TLB kept consistent with the host (no pinning/registering)
- Handshake in HW

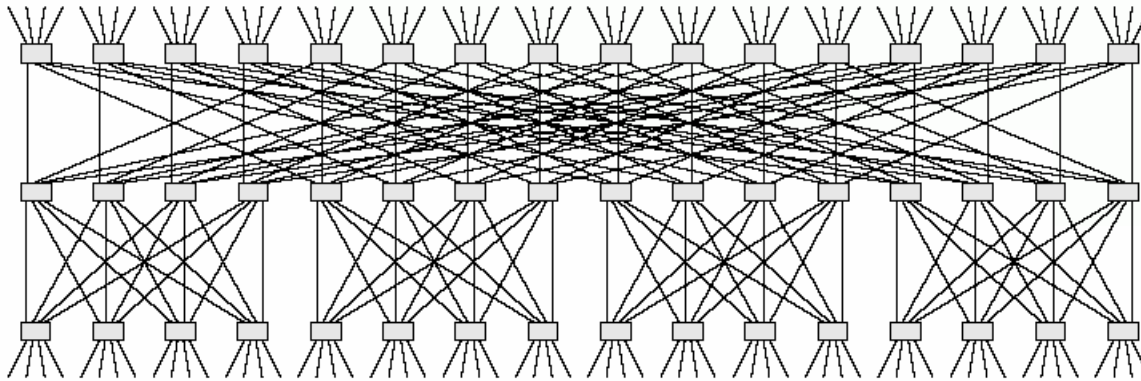
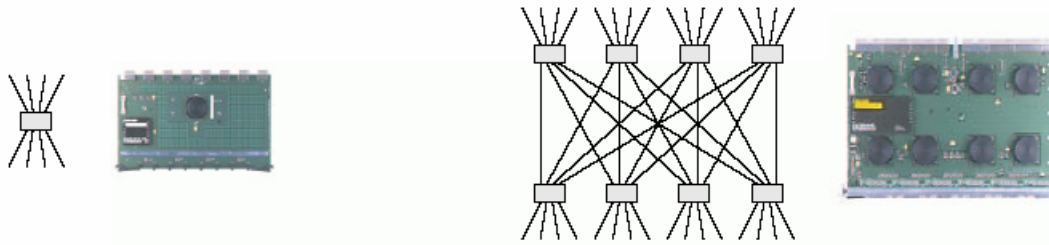
Elite switches

- 8 bidirectional links with 2 virtual channels in each direction
- An internal 16x8 full crossbar switch
 - 2 input ports per input link to deal with virtual channels
- 400 MB/s on each link direction
- 2 priority levels plus an aging mechanism
- Adaptive routing
- Hardware support for broadcast

Network topology: Fat-tree

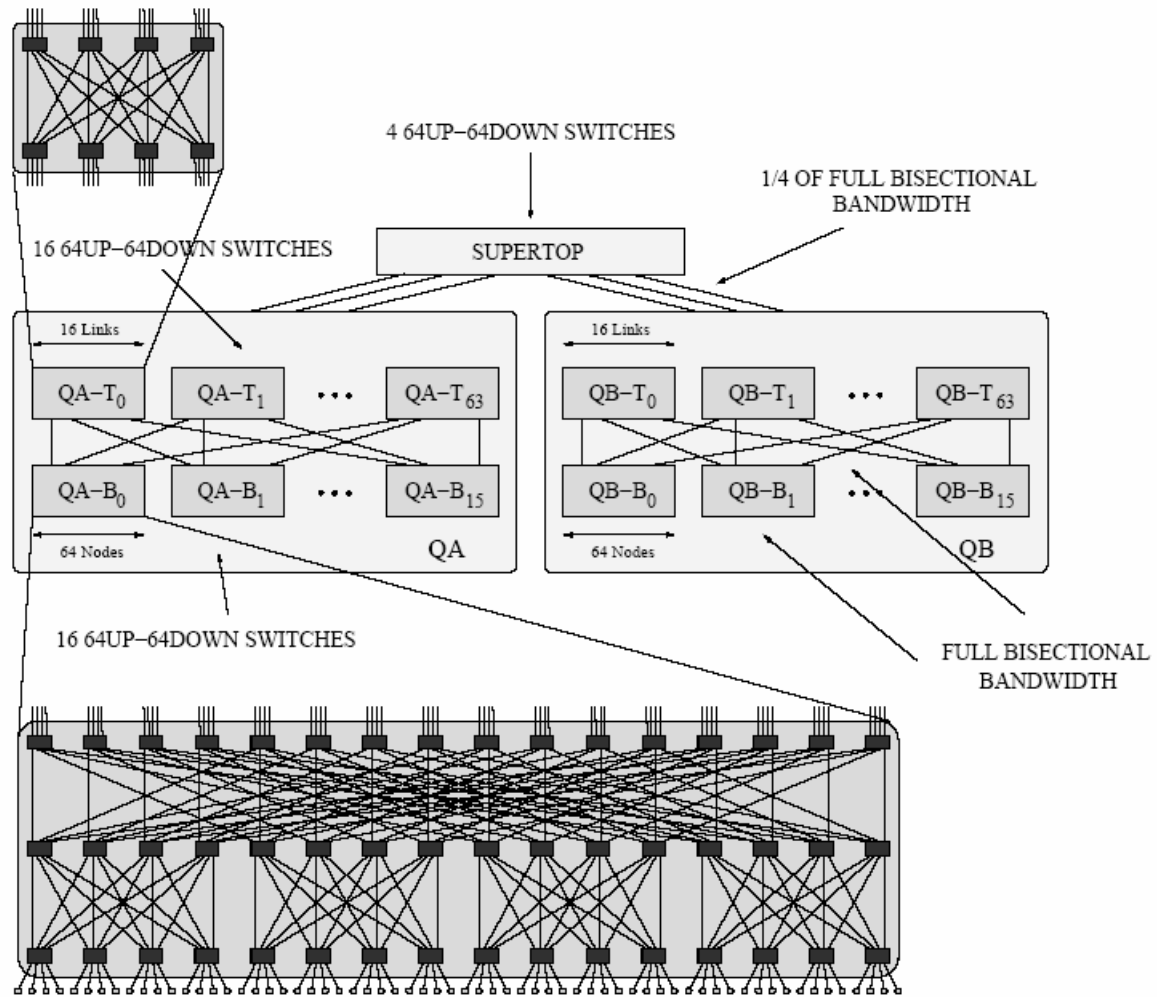


- Adaptive routing: choose lightest path through tree
- Path remains “reserved” for acknowledgement

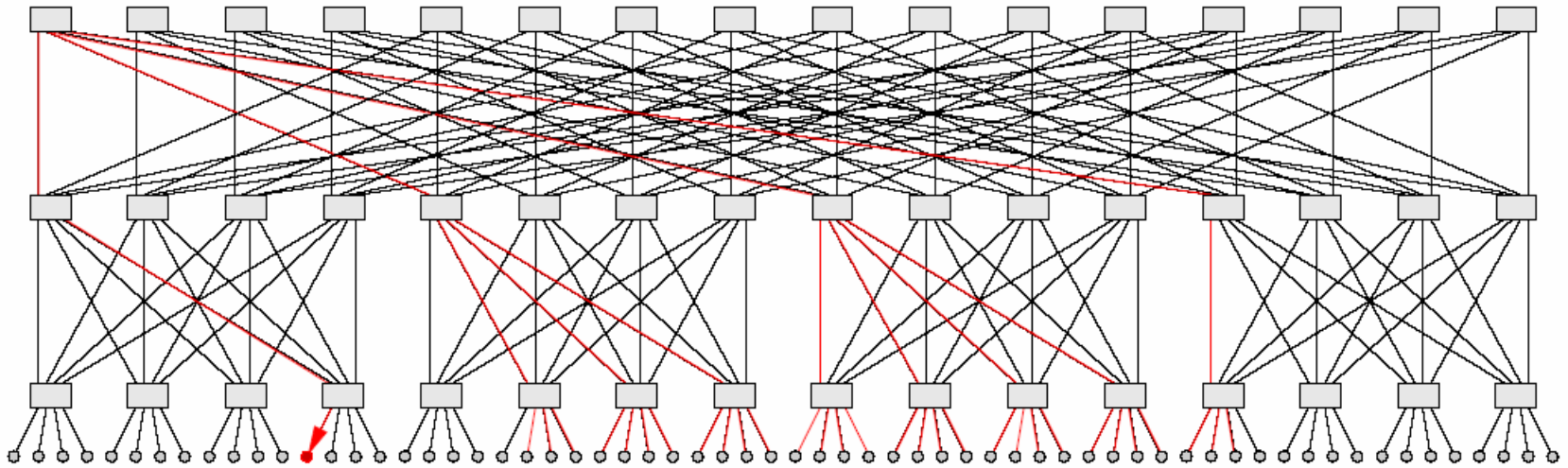


- Building blocks:
 - Single Elite (backplane)
 - 16 up/16 down (level 2 fat-tree)
 - 64 up/64 down (level 3 fat-tree)

rent



Hardware support for collectives



- E.g., multicast – provided that addresses are contiguous

War of the Interconnects

- Interconnects status
 - HotInterconnects 03 “War of the Interconnects”
 - Supercomputing 03 “Battle of the Network Stars”
 - HotInterconnects 04
 - ...
 - OSU benchmarks (everyone quotes them, just not all of them...)
- Today issue is Ethernet vs. Ethernet technologies
 - Myrinet, Quadrics have Ethernet compatible options
 - Long haul solution for wide/wider area connectivity
 - Ethernet technology with protocol offload in hardware

I/Cs on Top500 list (top500.org)

	11/03	06/04	11/04	05/06	06/06
GigE	109 (-)	163 (0)	176 (0)	212 (0)	256 (0)
Myrinet	73 (2)	186 (2)	193 (2)	69 (1)	56 (0)
Quadrics	26 (4)	23 (3)	20 (2)	13 (1)	14 (1)
Infiniband	3 (1)	10 (0)	10 (1)	17 (1)	37 (3)
clusters	208 (7)	290 (6)	294 (5)	304 (3)	364 (4)

HPLinpack benchmark -> Tflops

application mix evaluates compute power at nodes, also bandwidth (aggregate and point-to-point), and ping-pong and collective comm. latency

2 years old status...

	10GigE	Infiniband	Myrinet	SCI	Quadrics
Network Environment	Any: LAN, SAN, MAN, WAN	SAN	SAN	SAN	SAN
Scalability	IP-routed → "Infinite" # of nodes		Source-routed → ?		Source-routed → ?
Cost Per Port	\$\$\$\$ (2004: \$\$) (2005: \$)	\$\$	\$\$	\$\$\$	\$\$\$
Performance MPI-to-MPI	916 MB/s 10.0 us (at socket level)	843 MB/s 6.0 us	250 MB/s 6.3 us	230 MB/s 3.7 us	908 MB/s 2.4 us
Protocols	Native TCP/IP TCP offload RDMA over TCP/IP	RDMA	RDMA	RDMA (or RSM: Remote Shared Memory)	RDMA
Total Cost of Ownership	\$	\$\$\$	\$\$\$\$	\$\$\$	\$\$\$\$\$

SCI – scalable coherent interface

Both 10GigE and IB cost down

MN/IB latency - ~2us range, 10GigE 6+us range

Understanding performance factors

- Microbenchmarks – (uni-/bi-directional) latency/bandwidth; collective operations (barrier, allreduce, hotstop, multicast);
- Higher layers: sockets and MPI of main concern
- Actual applications
- Price/performance (\$)

Other possible metrics

- Also important
 - overheads, overlap of communication and computation, CPU loads, reuse of buffers, scalability, manageability, reliability, network topology, communication patterns, efficient collective communication, standards compliance, virtualization, portability...
 - what is your application's bottom line?
- Cost of ownership – not just \$ to set it up, but also to maintain and upgrade over time.