Quadrics Interconnect

Duncan Roweth

SC2001 Denver

Contents

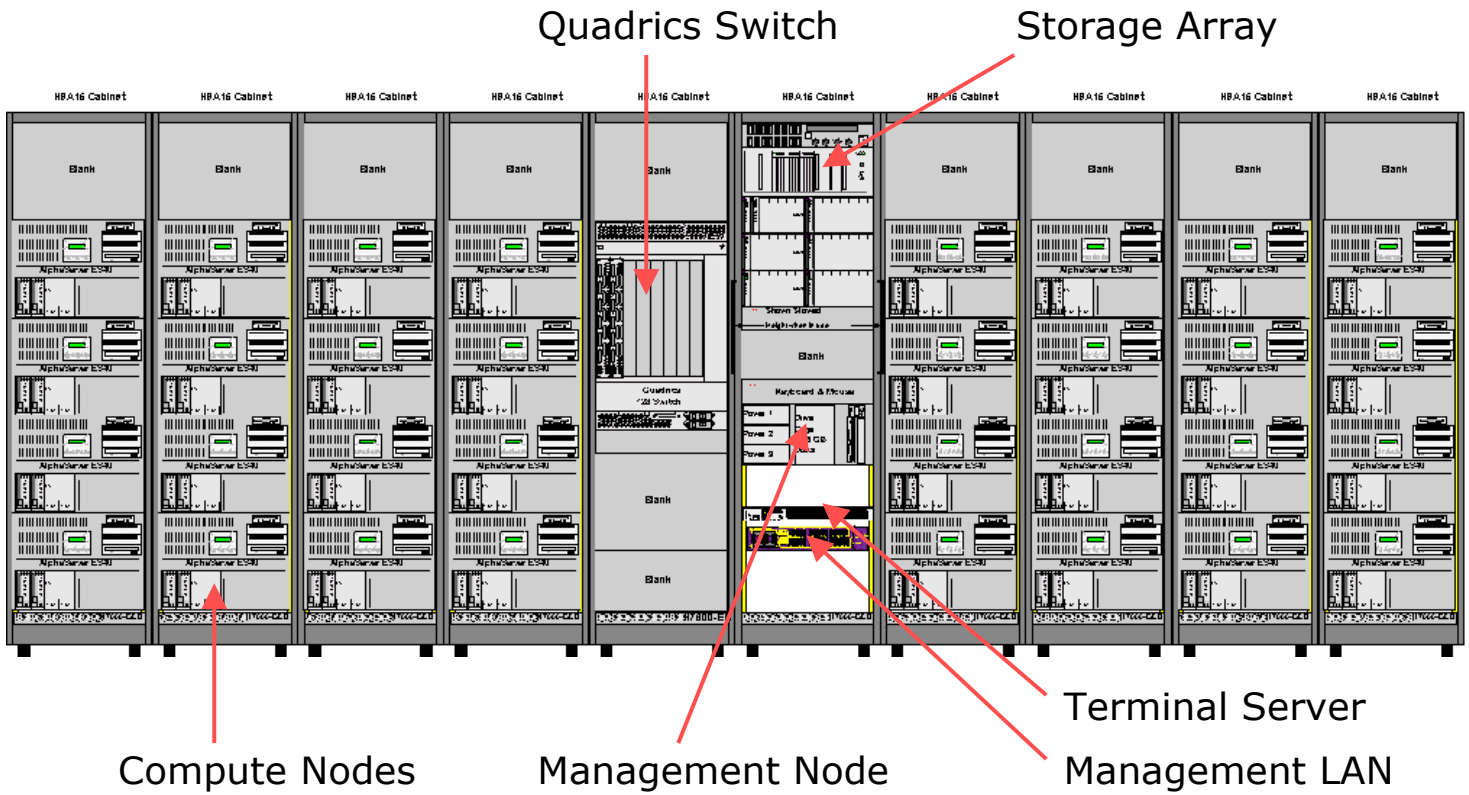
- QsNet Architecture
- Implementation
- Performance
- Software
- Roadmap

QsNet Architecture

- Adapter(s) in every node
- Multi-stage (fat-tree) network
- Model put, get, wait, queue
- Virtual & secure
- Fault tolerant

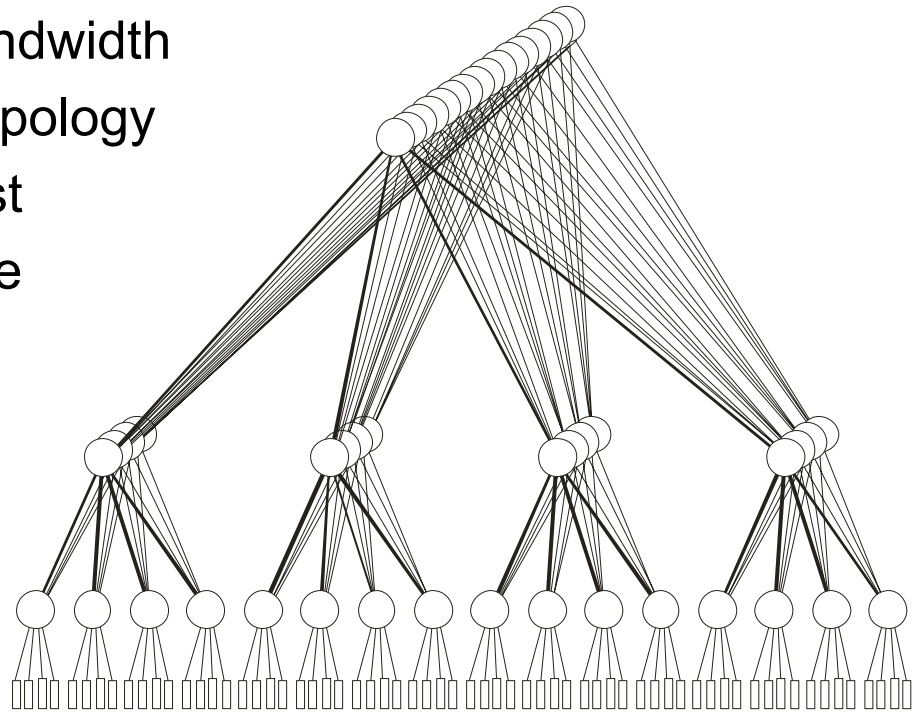
QsNet and Quadrics software provides the core components for building a high performance system from standard components.

System Overview



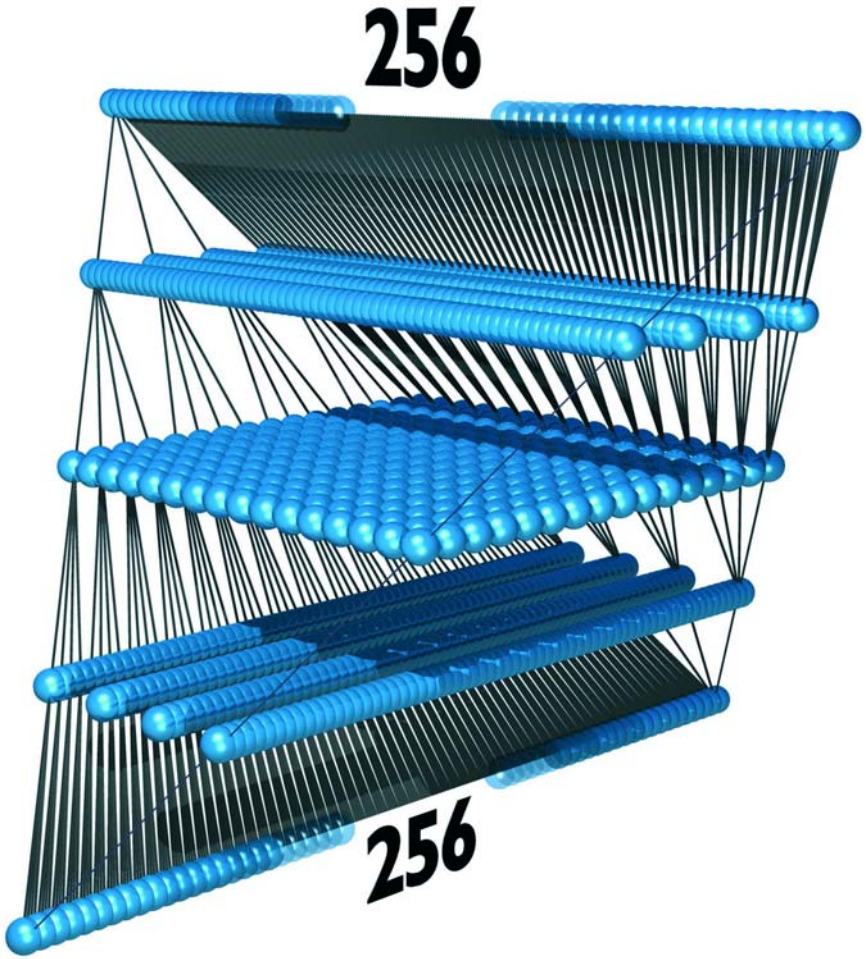
Multi-Stage Switches

- Full bi-sectional bandwidth
- General purpose topology
- Hardware broadcast
- Locality of reference
- Logarithmic cost
- Multiple routes



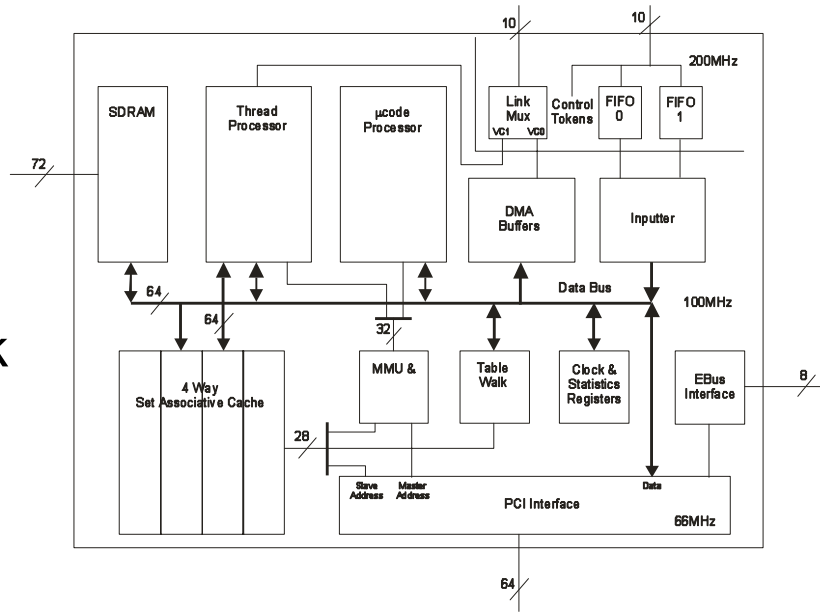
Adapter per node

1024-way Network

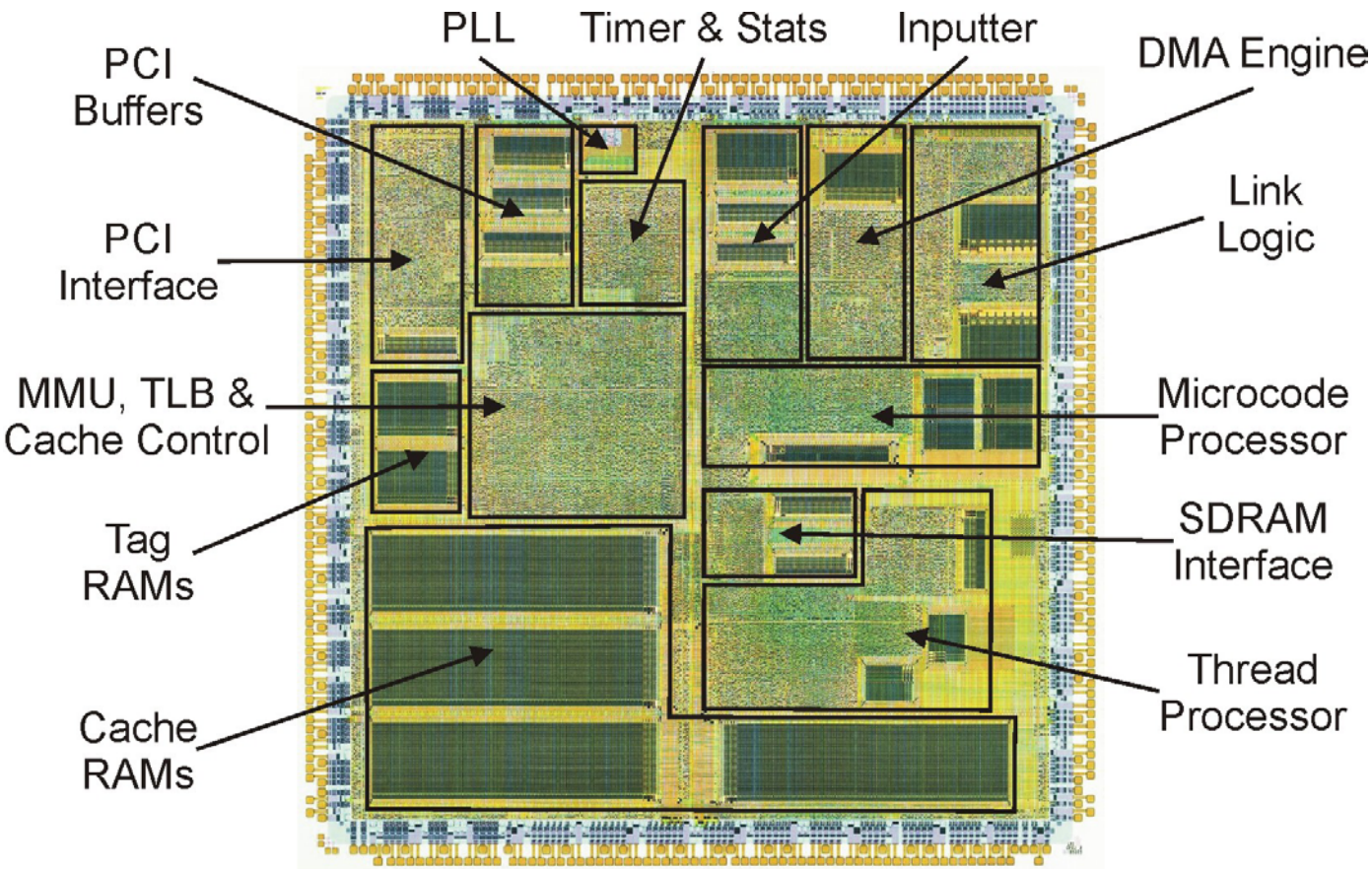


Elan3 Adapter

- DMA engine
- Thread processor
- On-chip cache
- MMU with hardware table walk and TLB
- 400Mhz byte wide LVDS link
- 64 bit 66MHz PCI interface
- 64Mbytes local ECC SDRAM

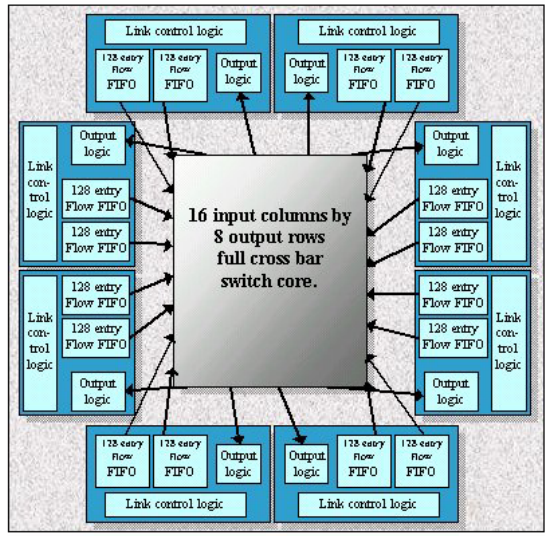


Elan3 Adapter

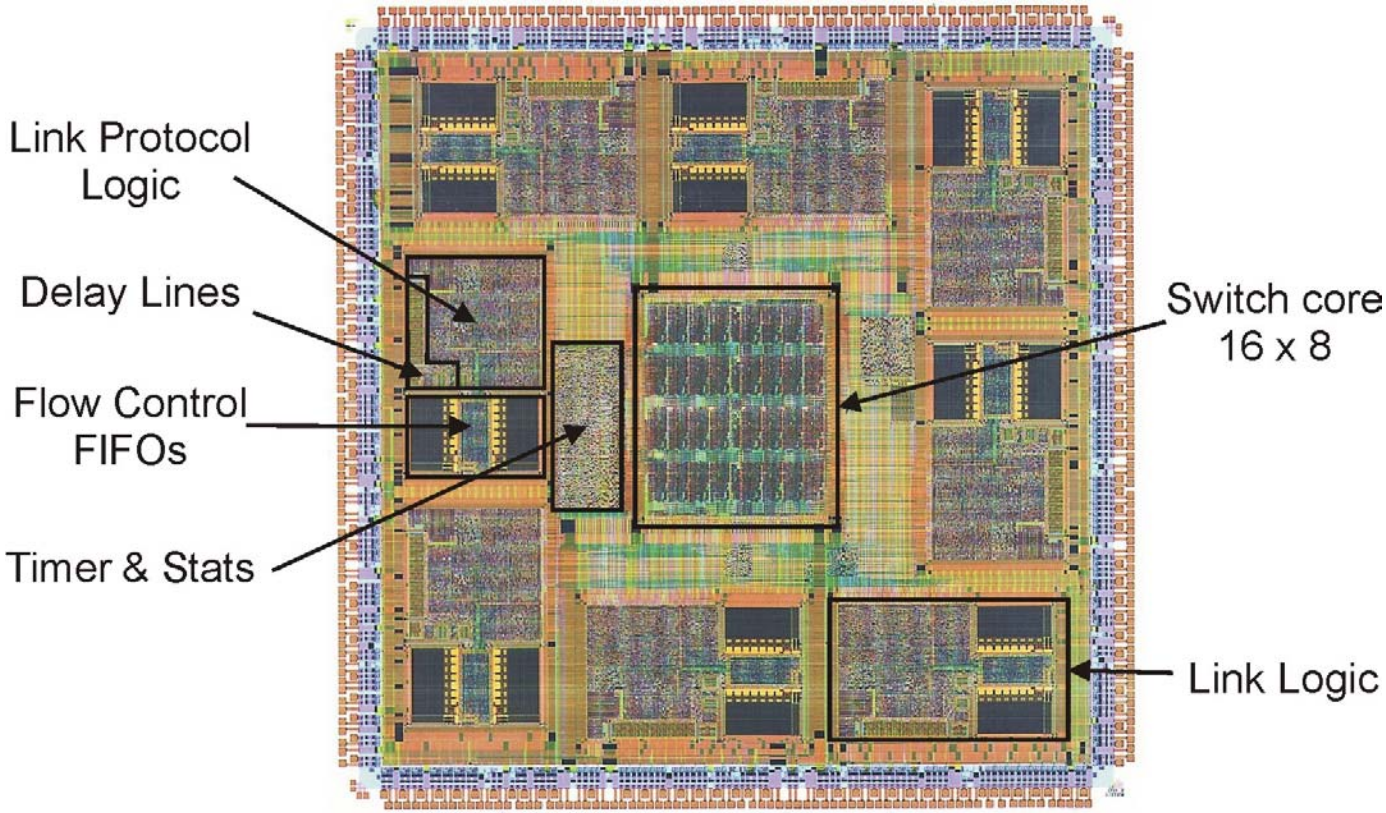


Elite3 Switch

- Byte wide bi-directional links
  - 400 MB/sec, 340 after protocol.
- 2 virtual channels per link
- Broadcast & network conditionals
- Adaptive worm-hole routing
- CRC on every packet
- 35ns per hop



Elite3 Switch



MultiRail Systems

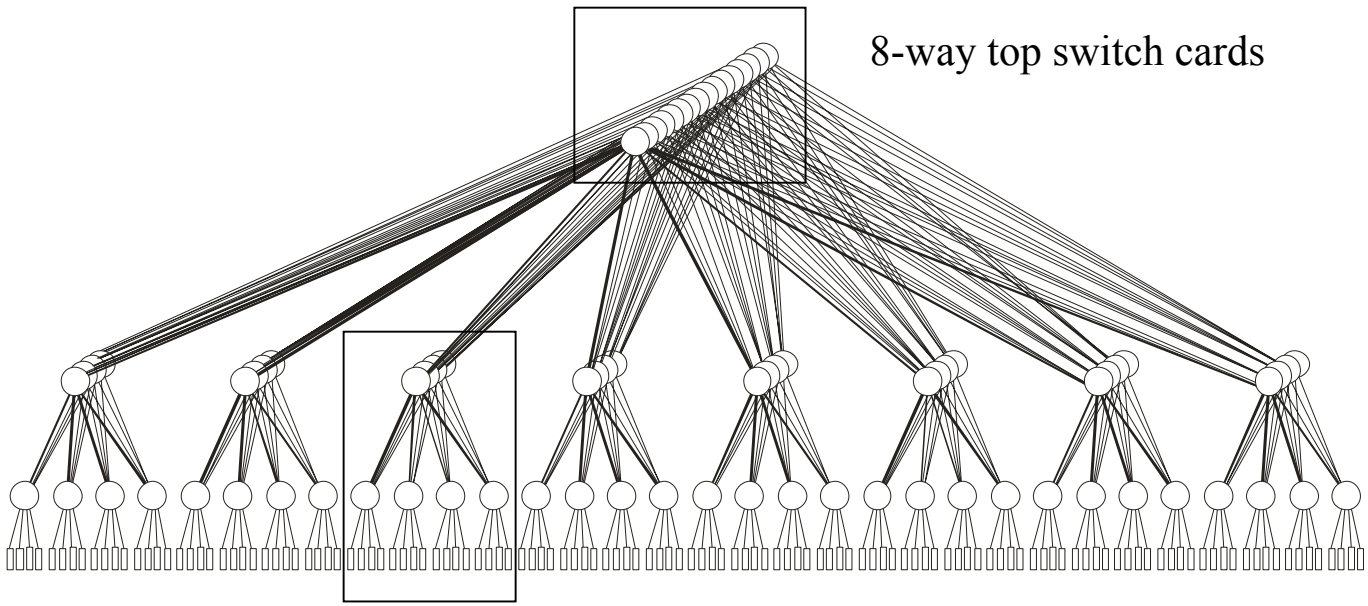
- Scaling comms performance with CPUs
  - Shared memory within the node
  - Quadrics data network between nodes.
- Replicate the data network
- Adapters connected to separate rails
  - 2 adapters per ES40 / ES45 in SC40 / SC45

QsNet Packaging

- Network Options
  - 16-128 nodes in units of 16
  - 64-1024 nodes in units of 64
- Switch Chassis
  - 42.5 Gbytes/sec bi-sectional bandwidth
  - 35 - 175 ns unblocked latency
  - Temperature, fan and PSU monitoring
  - Hot plug 1 from 2 48V PSUs
  - Live insertion
  - Embedded controller

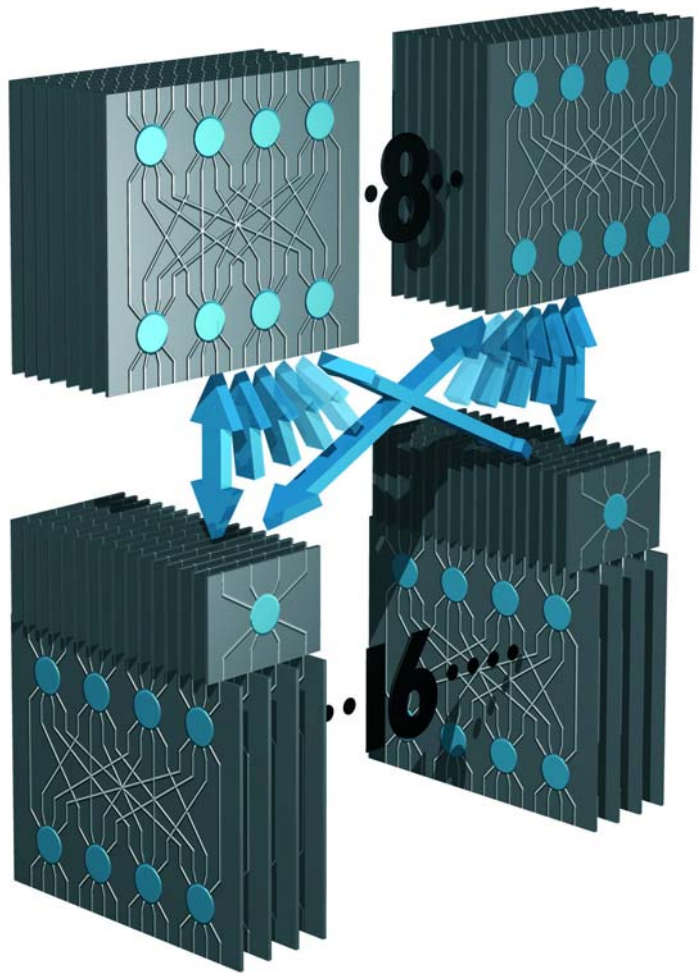
128-Way Switch Chassis

- 1-8 ×16 way switch cards
- 16 × 8 top switch cards



Federated Networks

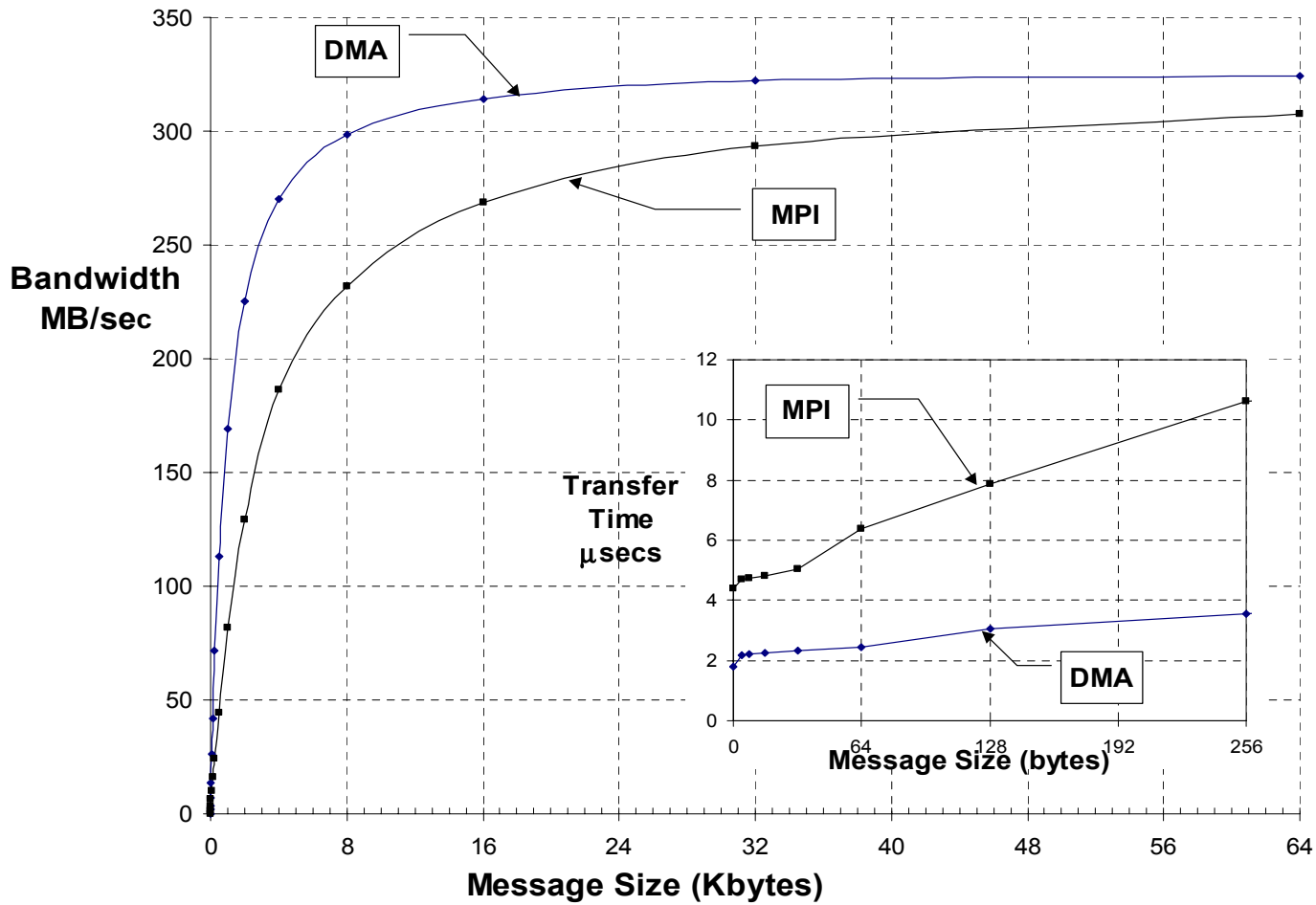
- Top switches
  - Multiple ×4, ×8, ×16 top switches connecting the node switches together
- Node switches
  - 64 links from nodes
  - 64 links to the top switches
- Common components



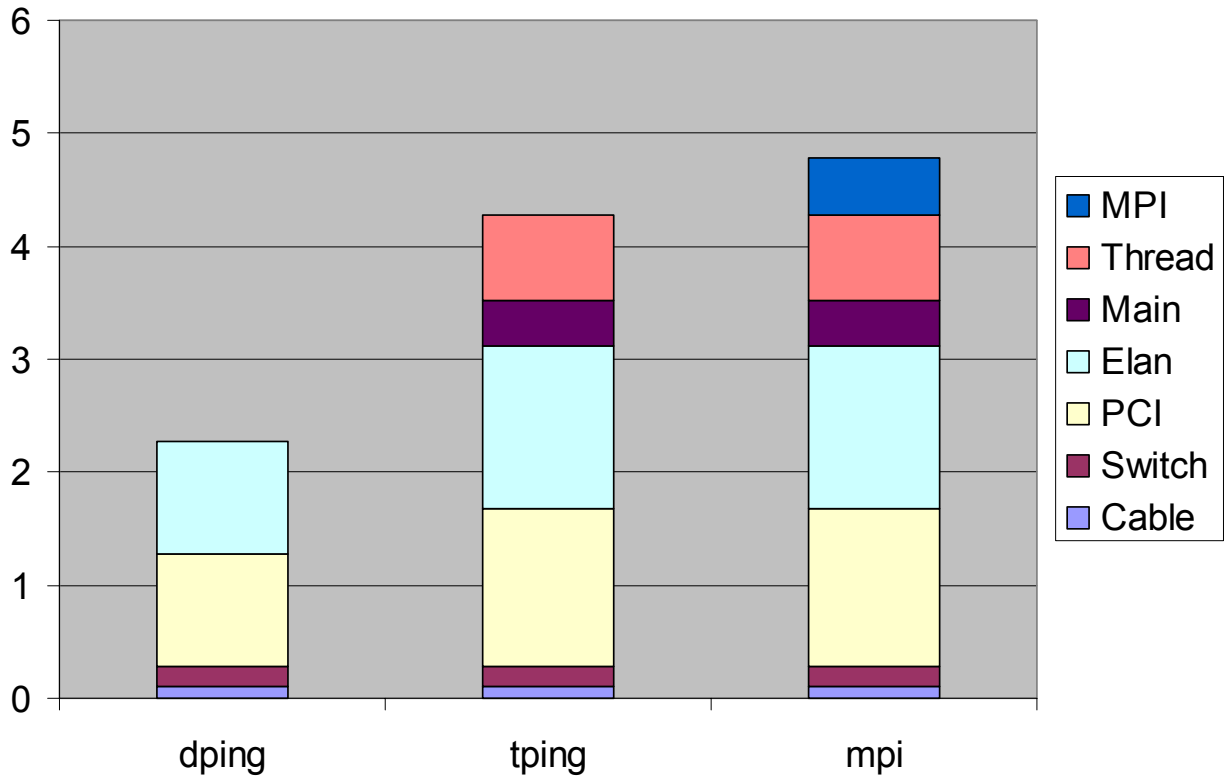
Elan-3 Performance Overview

- Latency
  - MPI send / receive 4 - 5 µsec
  - Shmem put / get 2.5 µsec
  - Pipelined put / get ≤ 1 µsec
- Bandwidth
  - MPI & Shmem 325 MB/sec
- Depends on PCI implementation
  - AlphaServer ES40 2.25 µsec 210 MB/sec
  - AlphaServer ES45 2.0 µsec 305 MB/sec

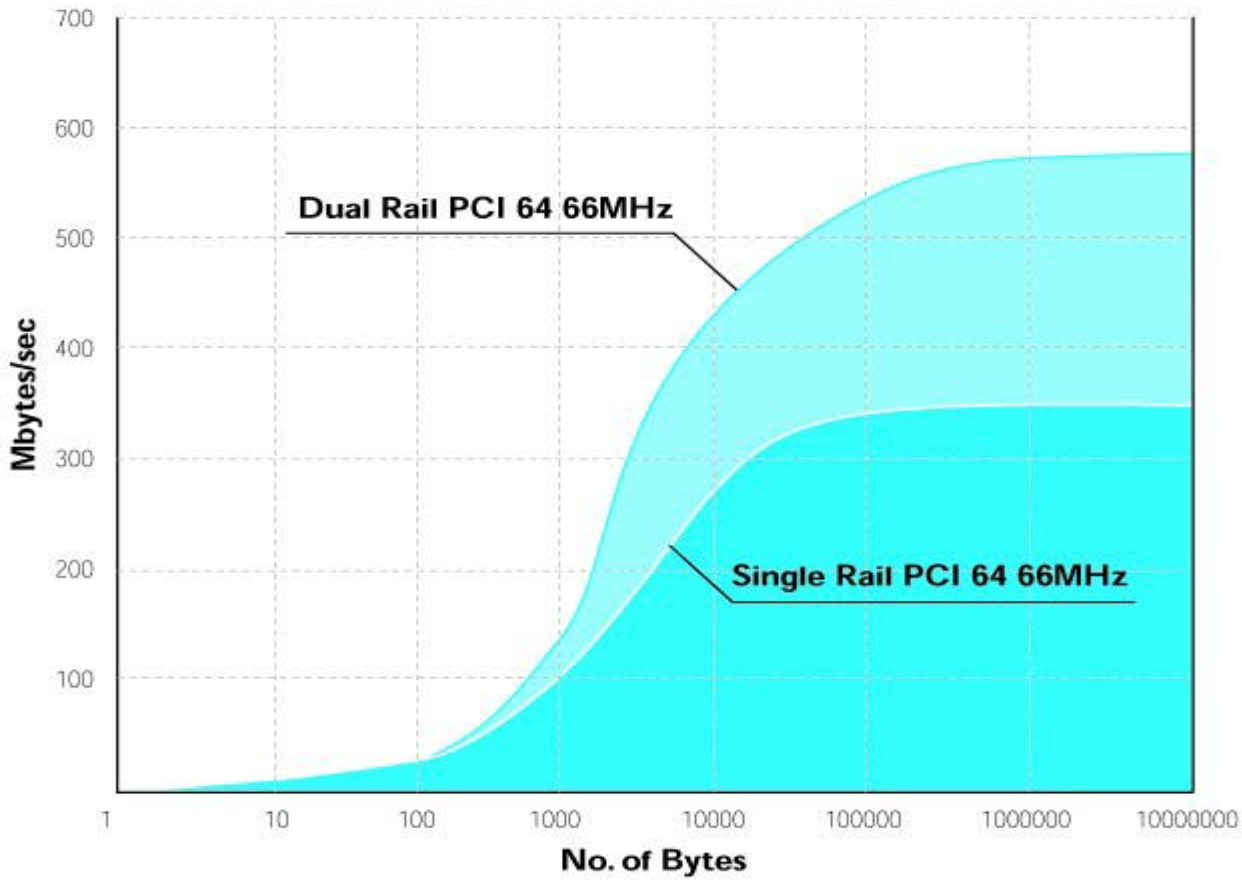
Latency and Bandwidth



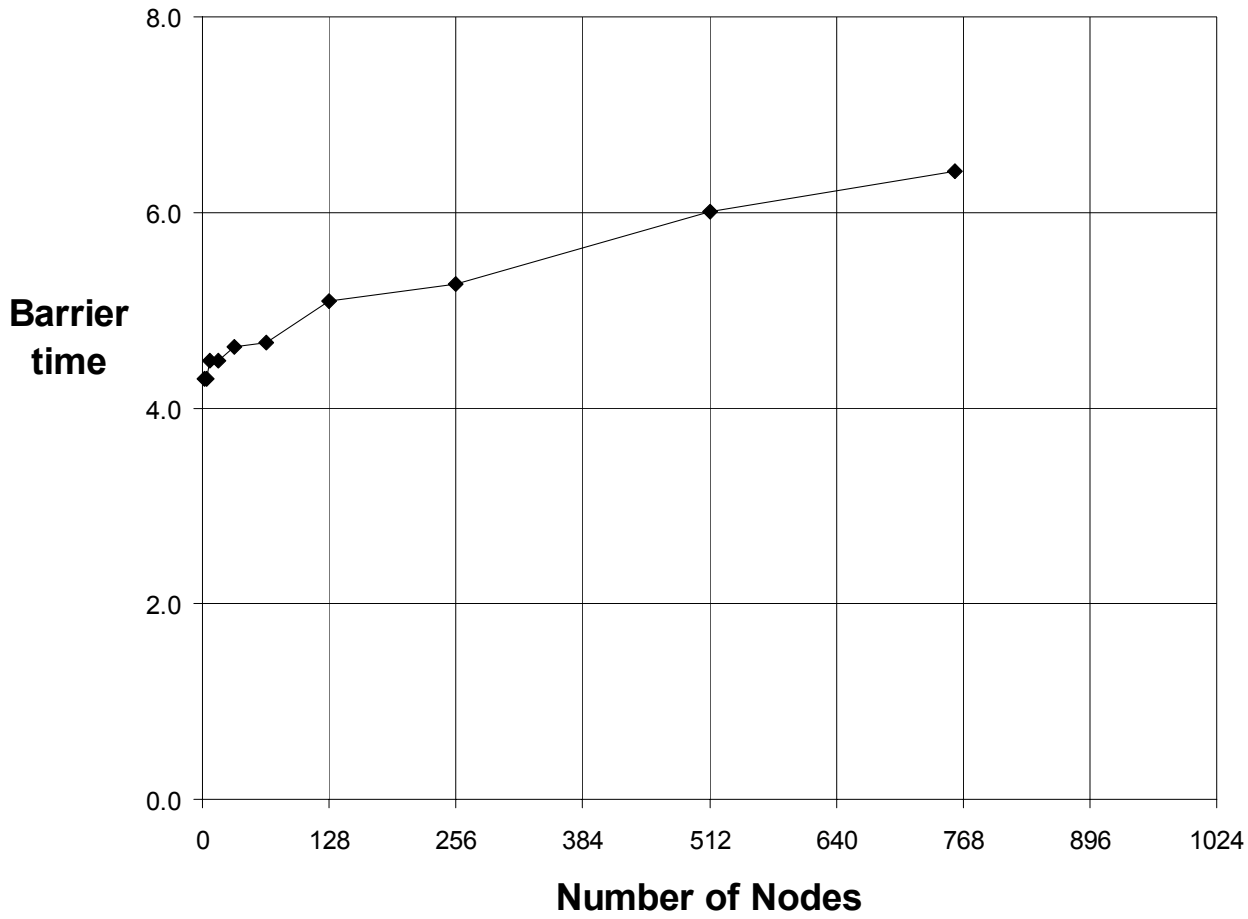
Elan-3 Latency Breakdown



MultiRail Exchange Performance



Barrier Scaling



Fault Tolerance

- Multiple routes – network failure
- Multiple rails – adapter / cable failure
- CRC on every packet
- Automatic retry / re-route on error
- Network errors collected by embedded controller and logged to database

Software Overview

Not just switches and cables – all the components you need to add to a standard node and OS.

- Device driver
- Optimised MPI/Shmem libraries
- Kernel services
- Resource management system
- Third-party development tools

Elan Driver

- Not used to communicate!
- Sets up the rights to communicate
  - Virtual address and process number
- Exception Handling
- Security
  - Kernel services and multiple users on the same hardware - Elan3 comms contexts.

Elan Libraries

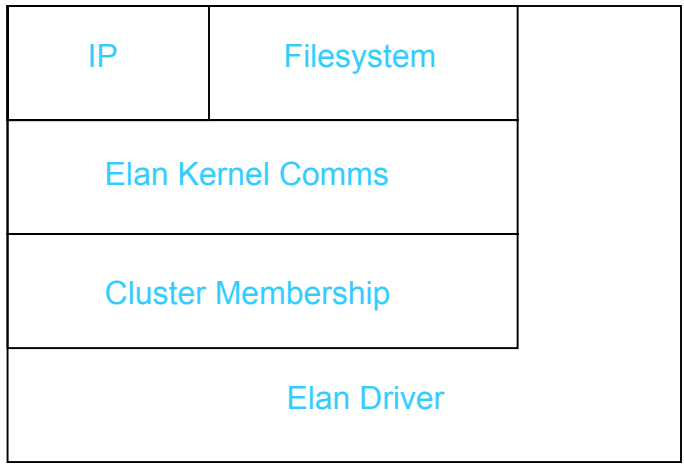
- Optimised MPI Implementation
  - Mpich 1.2 base
  - Point-to-point comms uses libelan
  - Local comms via shared memory
  - Optimised collectives
- Shmem Implementation
  - Wrapper around put / get
  - Local comms via block copy
  - Thread code for pipelined single word puts

Elan Libraries - MultiRail Support

- Allocation of communication to rails
  - Default policy for MPI / Shmem is to allocate a preferred rail per process
  - User over-ride
  - Full control in libelan

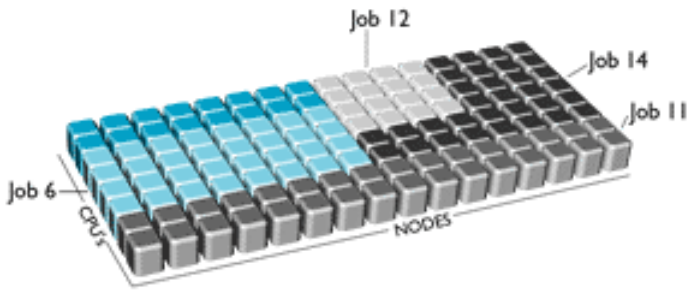
Kernel Services

- Cluster Membership
  - State of each node
- Kernel Comms
  - Message passing
  - RPC, put / get
- IP over Kernel Comms
- Filesystem over Kernel Comms
- Parallel Filesystem

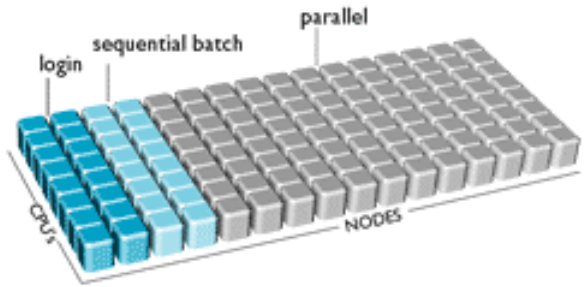


RMS

- Gang Scheduler
- Parallel runtime
- Performance and event monitoring
- Integration with LSF PBS, TotalView
- Management GUI

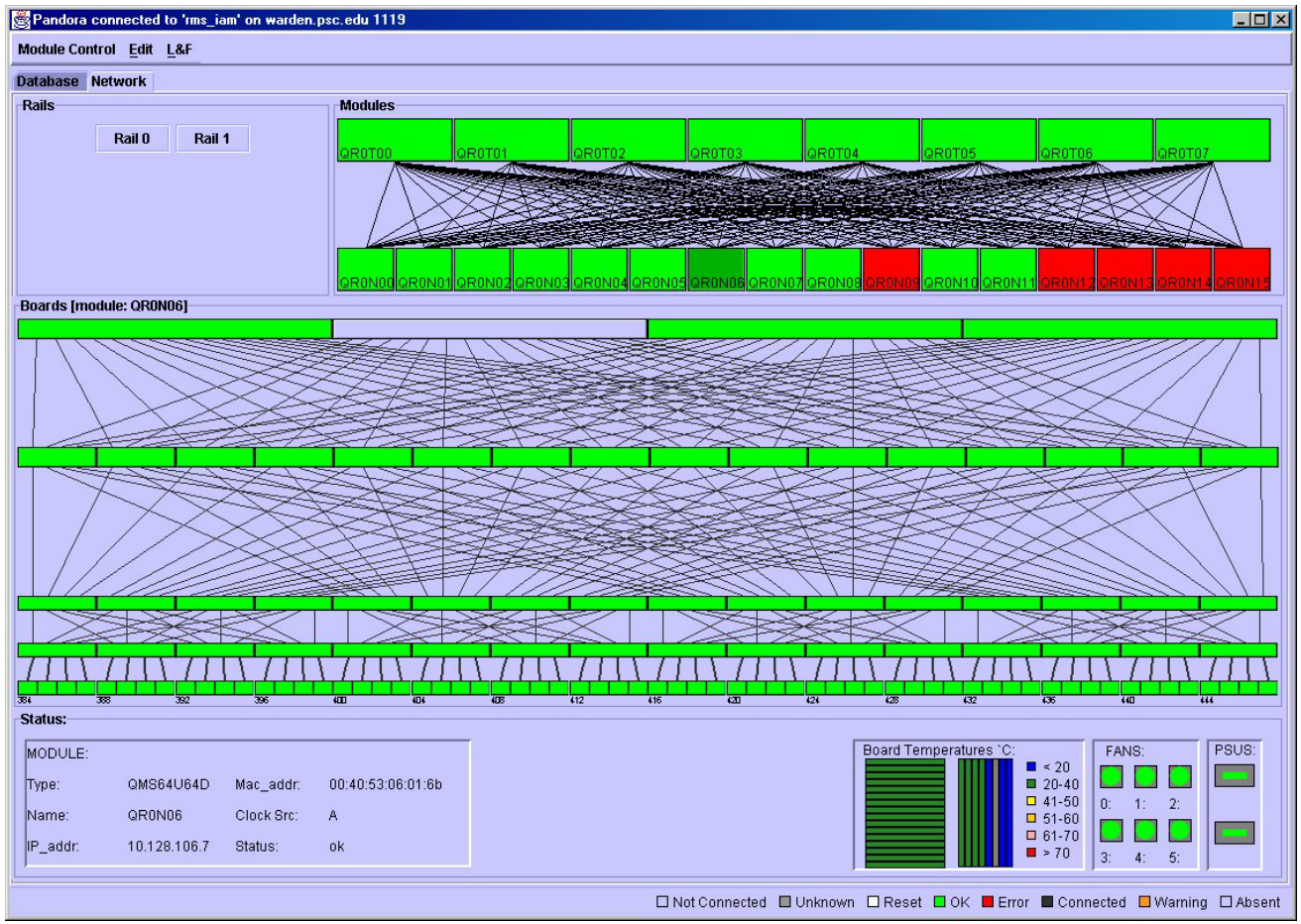


A typical configuration with login, sequential batch and parallel partitions



RMS Job Scheduling. Each job is allocated the optimal CPU's on a contiguous range of nodes

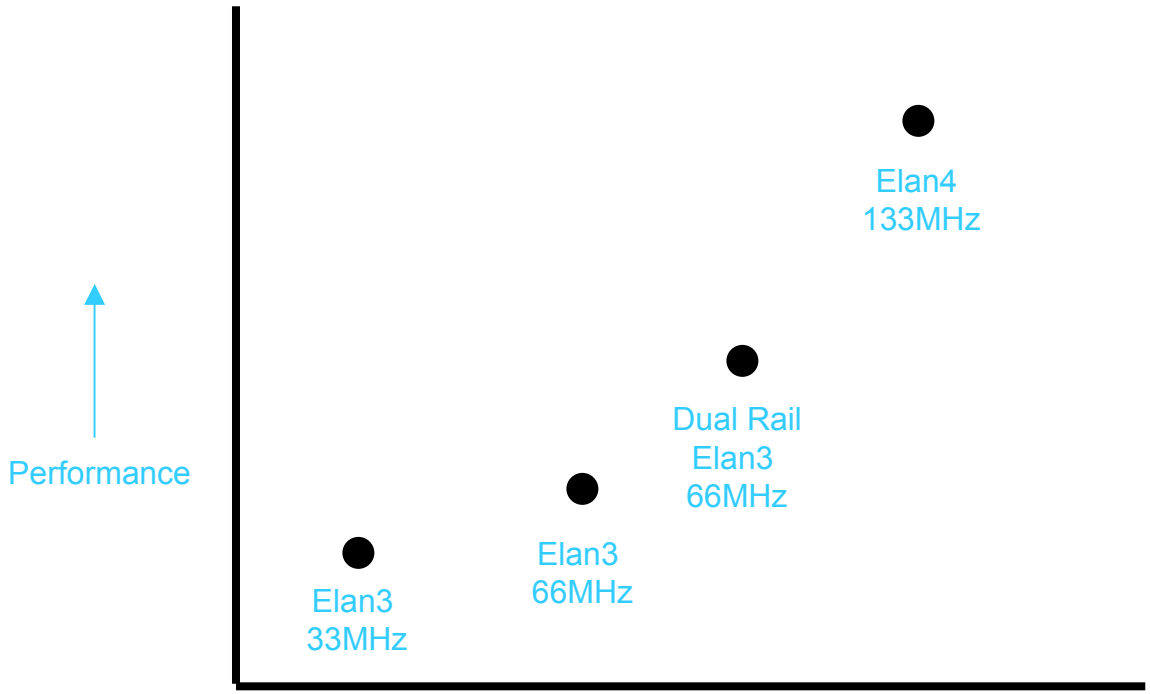
Pandora Network View



Largest System Installed

- AlphaServer SC at PSC
  - 760 ES45 nodes
  - 1024 port dual rail network

QsNet Roadmap



Next Generation Hardware

- No change to the overall architecture
  - Full 64-bit adapter
  - Hardware support for pipelined 64-bit writes
- PCI-X interface
- Faster network
  - Comfortably exceed PCI-X bandwidths
- Copper & parallel fibre links
  - Infiniband ×12 physical layer

Elan4 Compatibility

- No mixed Elan3 / Elan4 systems
- Backwards compatible to 66MHz PCI
- Libelan interface preserved

QUADRICS

www.quadrics.com