Quadrics QsNet^{II}:

A network for Supercomputing Applications

David Addison, Jon Beecroft, David Hewson,
Moray McLaren (Quadrics Ltd.), Fabrizio Petrini (LANL)

You've bought your super computer

You've constructed your building for it!

So what do you want from the network?

Any process to any other process communication in *zero* time.

That's it! Simple

Can't have that so you need the next best thing ...

A Quadrics Supercomputer Network

- Ultra low user process to user process latency
- Highest possible (affordable?) compute communications ratio
- Seamless scaling to many 1000s of nodes
- High availability
- Reliable data transfer
- Mixed system and multiple user traffic on one network

Supercomputers

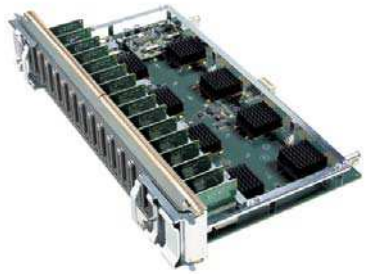
- **Los Alamos National Laboratory – ASCI Q 13880 Gflops**
- **Lawrence Livermore National Laboratory - MCR 7634 Gflops**
- **Lawrence Livermore National Laboratory - ALC 6586 Gflops**
- **Pacific Northwest National Laboratory - 4881 Gflops**
- **Pittsburgh Supercomputer Centre - Le Mieux 4463 Gflops**
- **CEA - Tera -3680 Gflops**

Images Courtesy of LANL, LLNL,PNNL, PSC, CEA

QsNet II Components

- Elan 4 network interface card

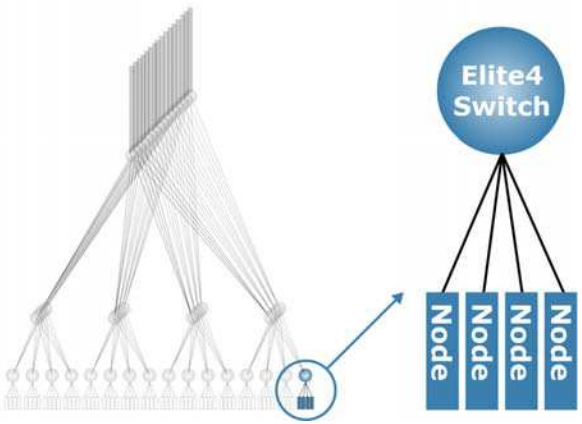
- Elite 4 switch component

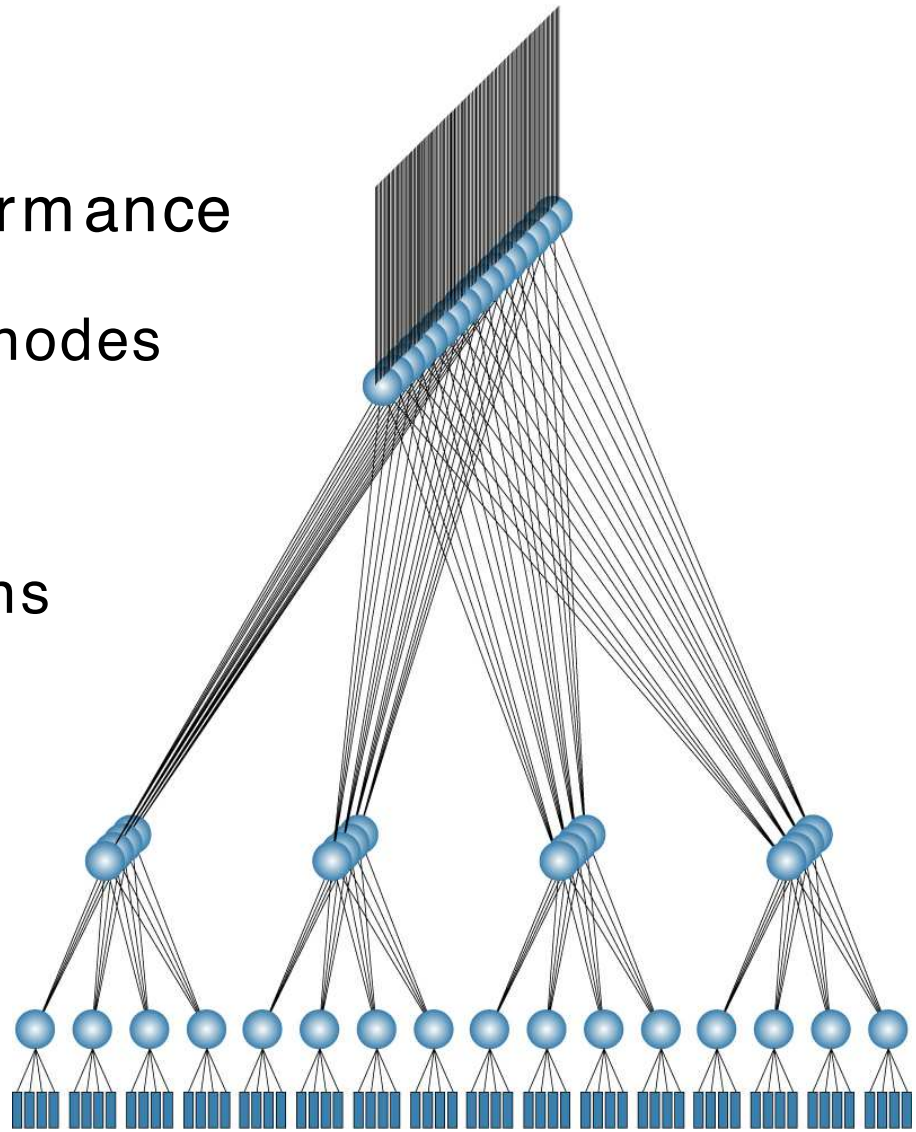


- QsNet II Switch

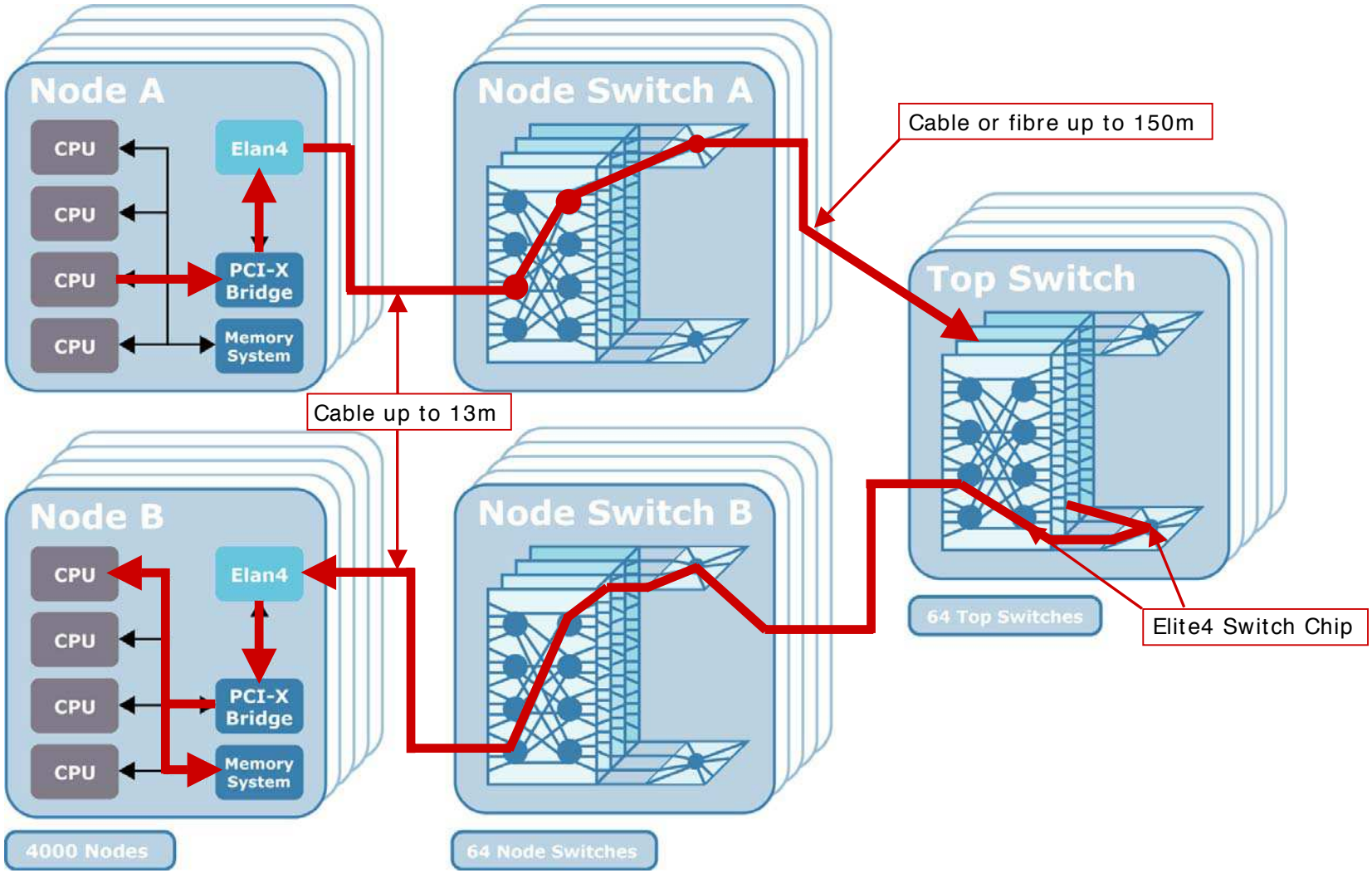
Topology

- Fat trees give great performance
  - Scales to the number of nodes
  - Fault tolerance structure
  - Uniform connectivity
  - Supports global operations

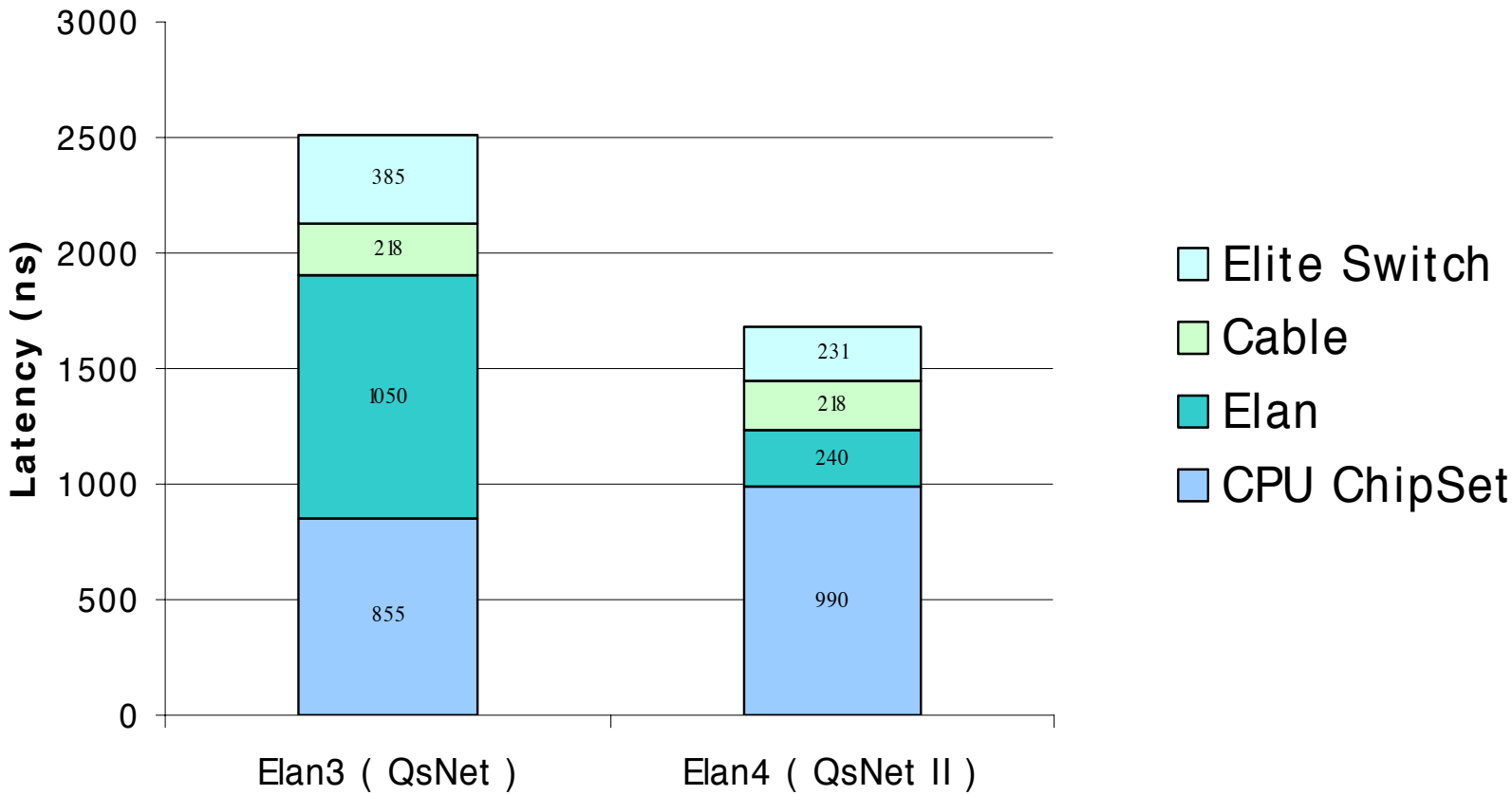




A Process Communication

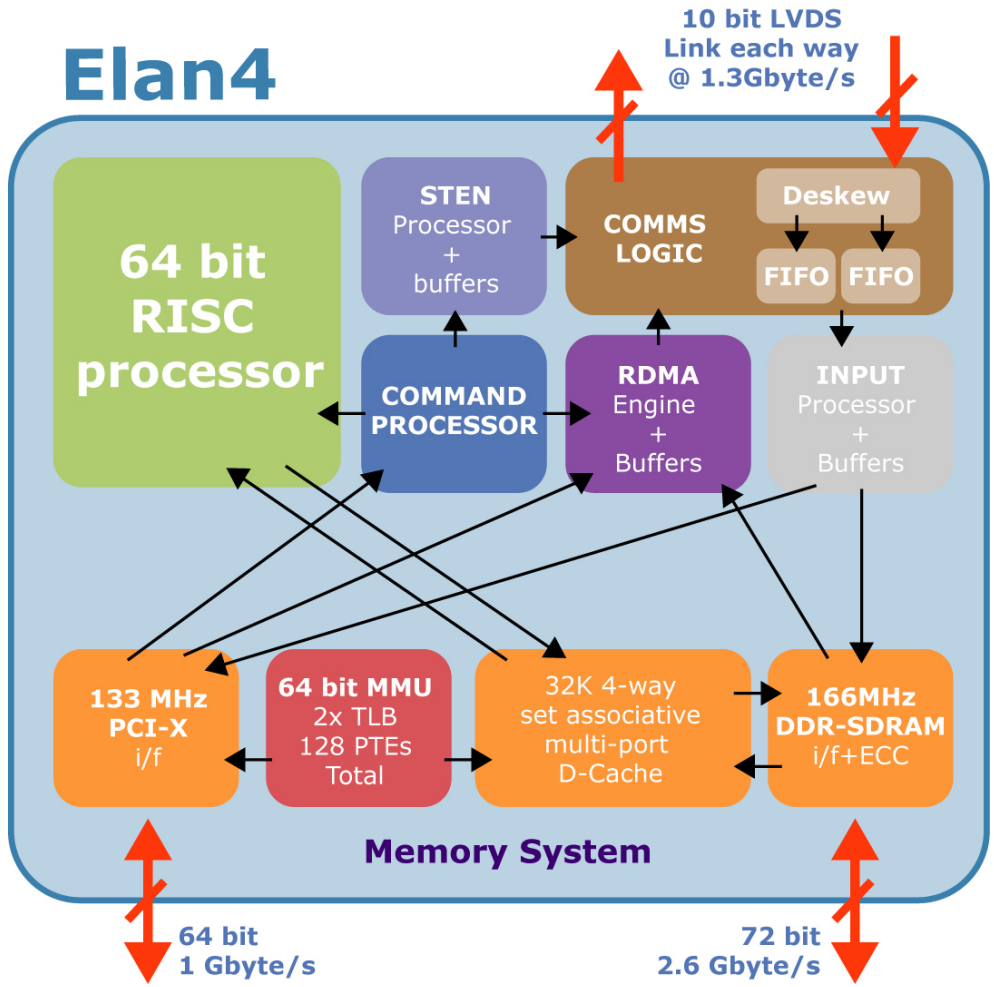


8 byte write latency on a 4000 node machine with 50m of cable



Elan 4 functional units

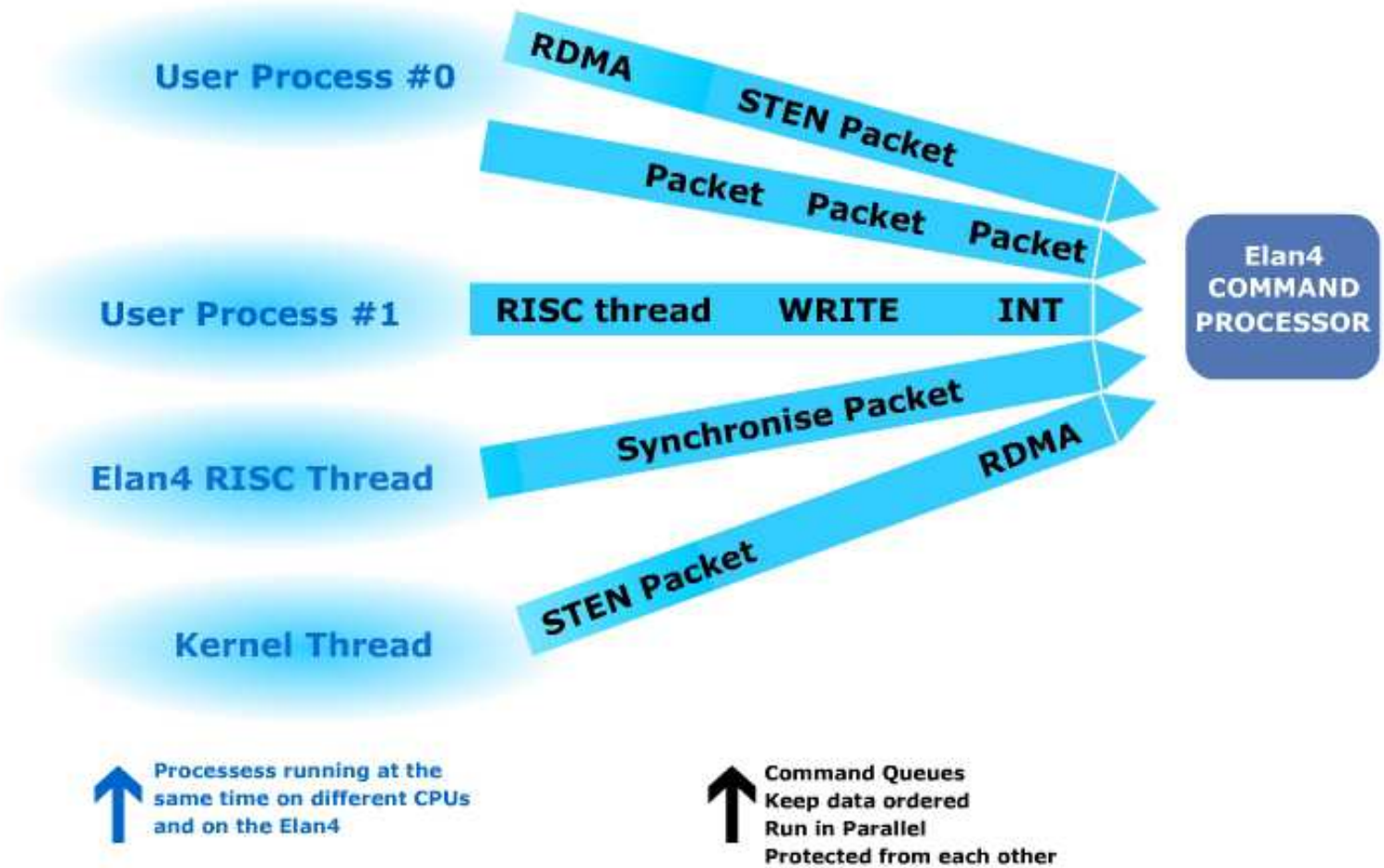
- 64-bit virtual addressing
- Short Transaction ENgine
- Pipelined (R)DMA engine
- 64-bit RISC processor
  - 16Kbyte On-chip I-cache
- Memory System
  - 32Kbyte On-chip D-cache, pipelined fills, multi-port.
  - 64-bit MMU 128-TLB entries, hash walk engine, mixed page sizes, 16 bit context.
  - 64-bit/133MHz PCI-X
  - 64Mbytes ECC DDR RAM
- Link. 2.6 Gbytes/sec total



Elan4 Command Queues

- Enables a user process to send packets into the network with very low latency
- Used to start all operations. (RDMA, STEN, RISC threads etc)
- Up to 8K command queues can be allocated
- Command queues can by used by many processes simultaneously from multiple CPUs

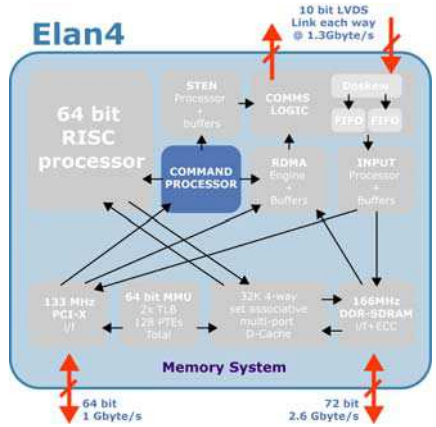
Command Programming Model

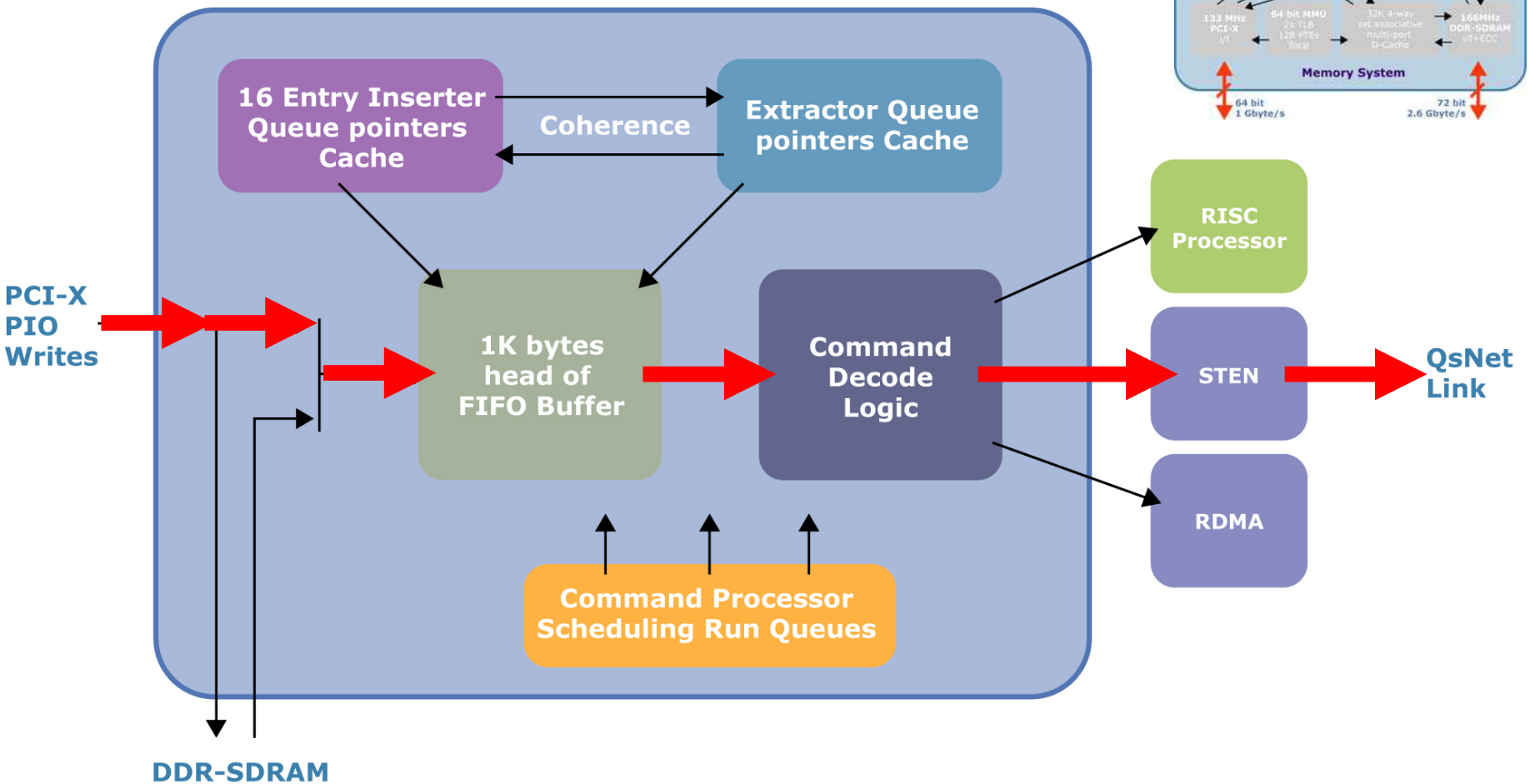


Elan4 Command Queues

- The Command Processor can execute commands directly as they are written from the PCI-X bus
- 80ns from PCI-X bus to network link
- Provides auto retry of STEN packets
- DDR-SDRAM used as backing store for Queues
- Copes with a main CPU process timeslice in a command stream and concurrent access by multiple CPUs
- Copes with occasional “out of order” PIO writes

Command Proc Implementation



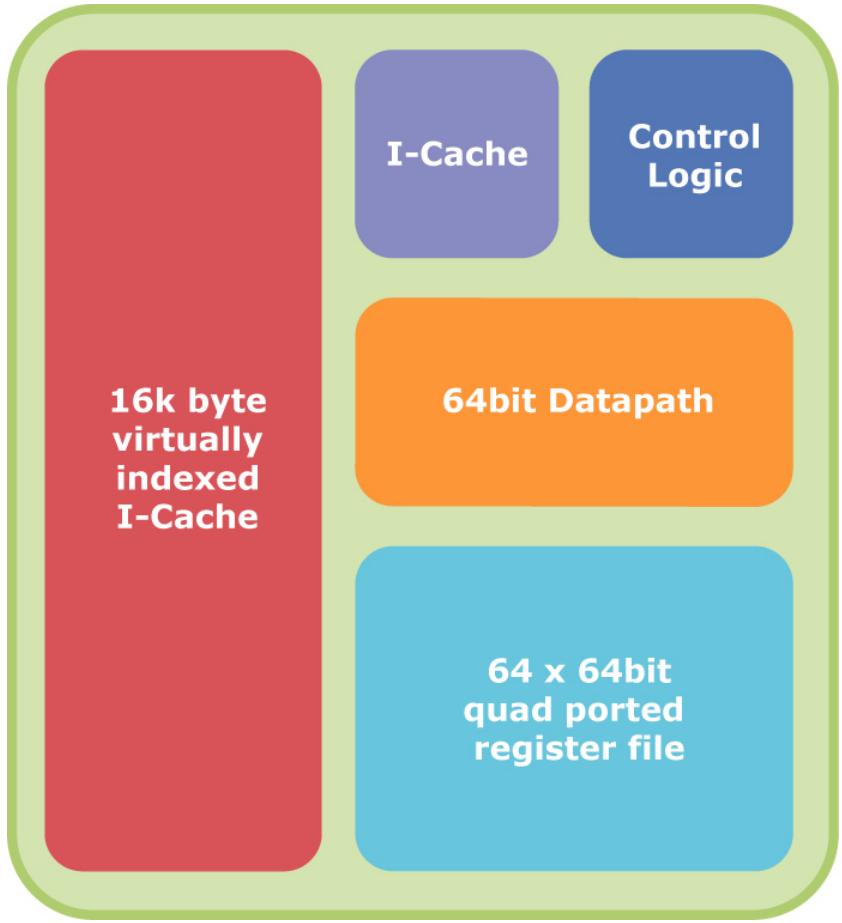


Elan4 RDMA processor

- Addresses are 64 bit
- Pipelined operation to hide large PCI-X read latency
- Processes two RDMA descriptors concurrently to achieve peak bandwidth with multiple small RDMAs
- Two run queues for high and low priority scheduling
- Timeslices between multiple RDMAs

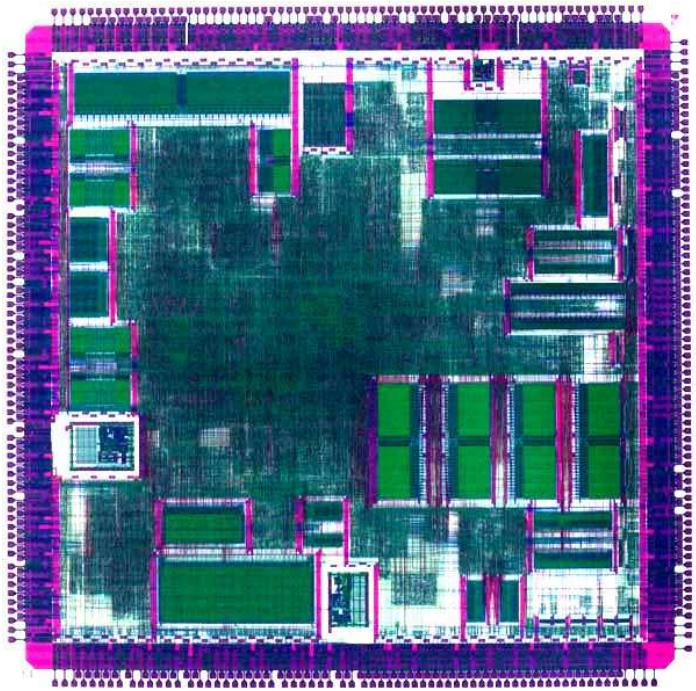
RISC processor

- 64 bit word
- Instruction set optimised for low latency scheduling
- 16 Kbyte I-Cache
- Registers loaded directly from the network
- Block load/store instructions for copy bandwidth



Elan4 implementation

- LSI G12 process
- 0.18um 4 + R layer metal process
- 7.5 mm x 7.5 mm
- ~800,000 gates
- ~283 signal pins
- 512 ball BGA
- 3 watts



QsNet^{II} Physical Link

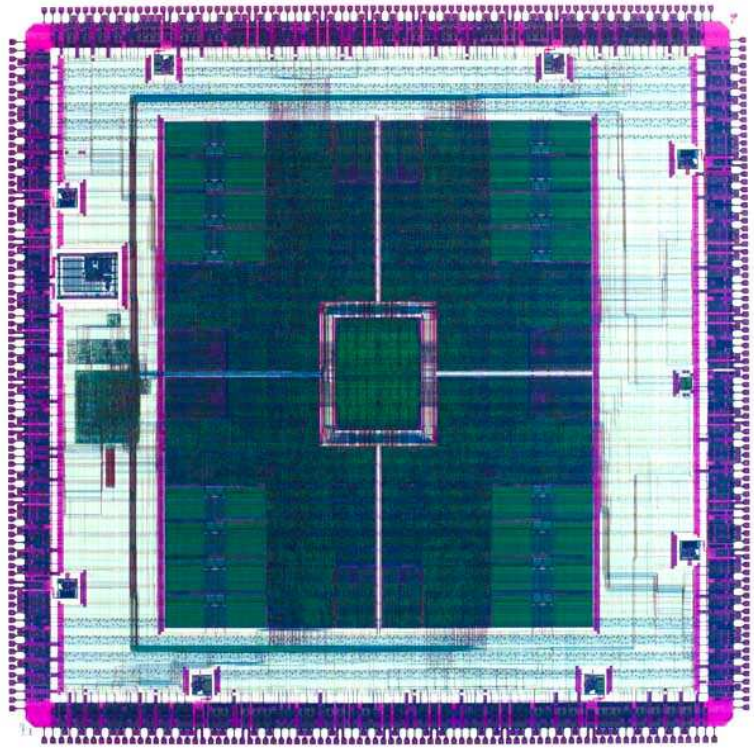
- 1.333Ghz design speed
  - 4b5b coding for DC balance on cables and fiber
  - ~920 Mbytes/s after protocol
  - Internal switch links deliver 1.18 Gbytes/s after protocol
  - 2 virtual channels
- Copper
  - 10 bit LVDS – total 40 wires
  - 12m range
- Optics
  - 12 bit parallel optical fiber
  - 150m range

QsNet^{II} Elite4 Switch Component

- 8 QsNet^{II} links × 2 virtual channels
- Broadcast to range of outputs
- Full automatic error detection / recovery
- Arbitration based on age of packet
- Two levels of priority
- Adaptive routing support
- Unblocked latency of ~20ns
- Traceroute transaction for interrogating the network

Elite 4 implementation

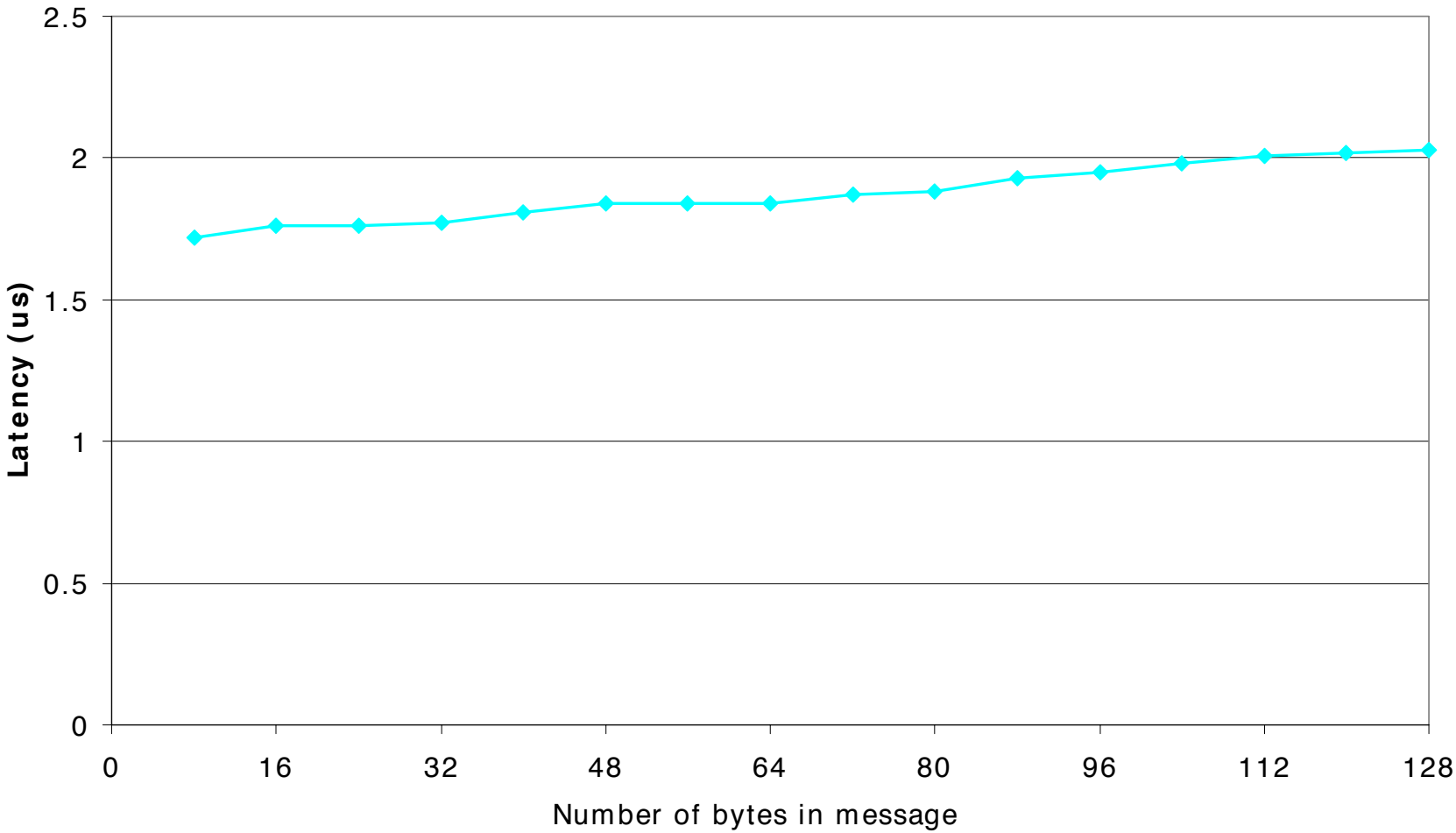
- LSI G12 process
- 0.18um 4 + 1 layer metal process.
- 8.67mm x 8.67mm
- ~ 1 million gates
- ~348 signal pins
- 608 ball BGA
- 6.5 watts



Performance

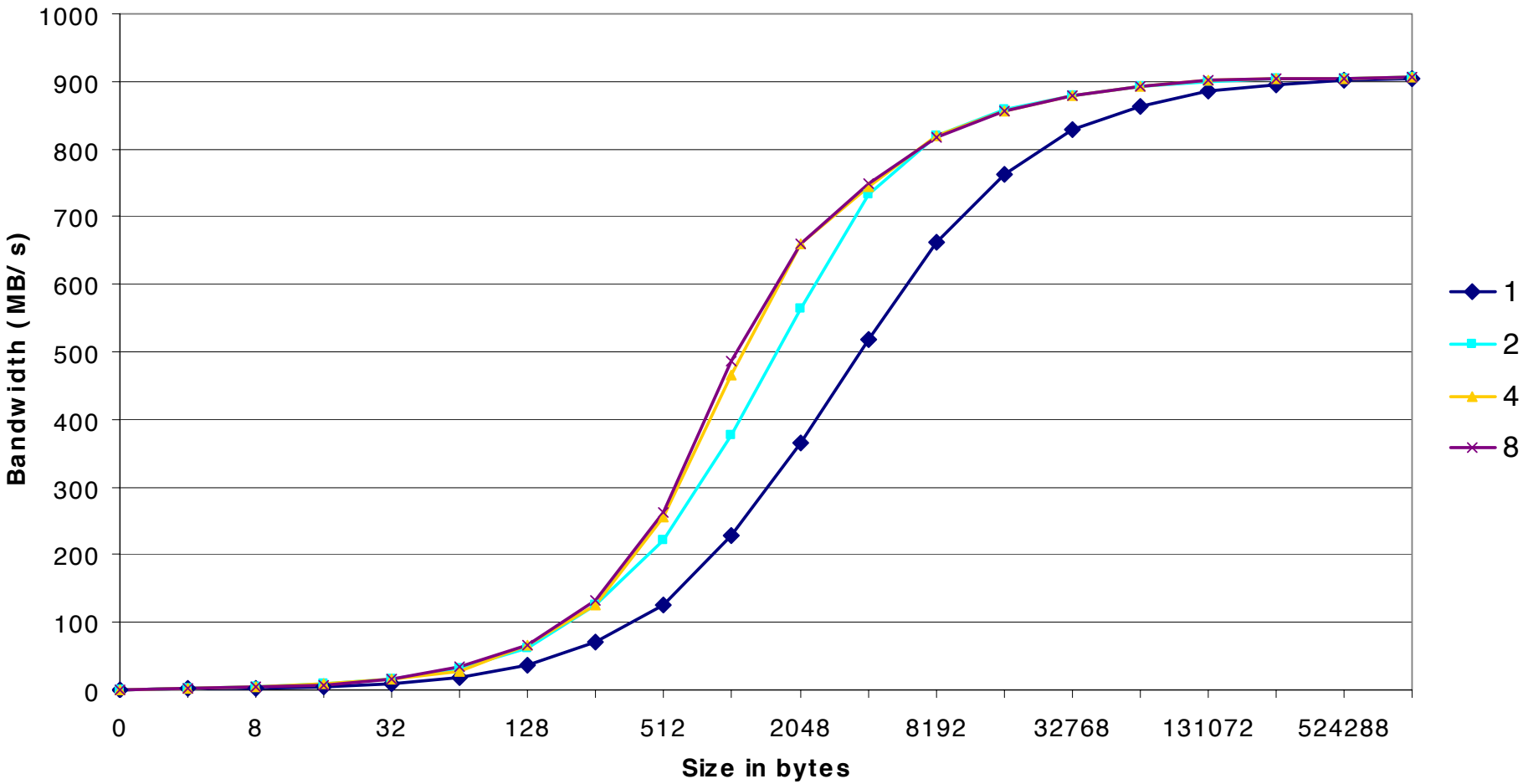
- Short message latency
- RDMA bandwidth
- Performance in applications
  - MPI message passing
  - Global operations

Short message latency on a 4000 node machine with 50m cable

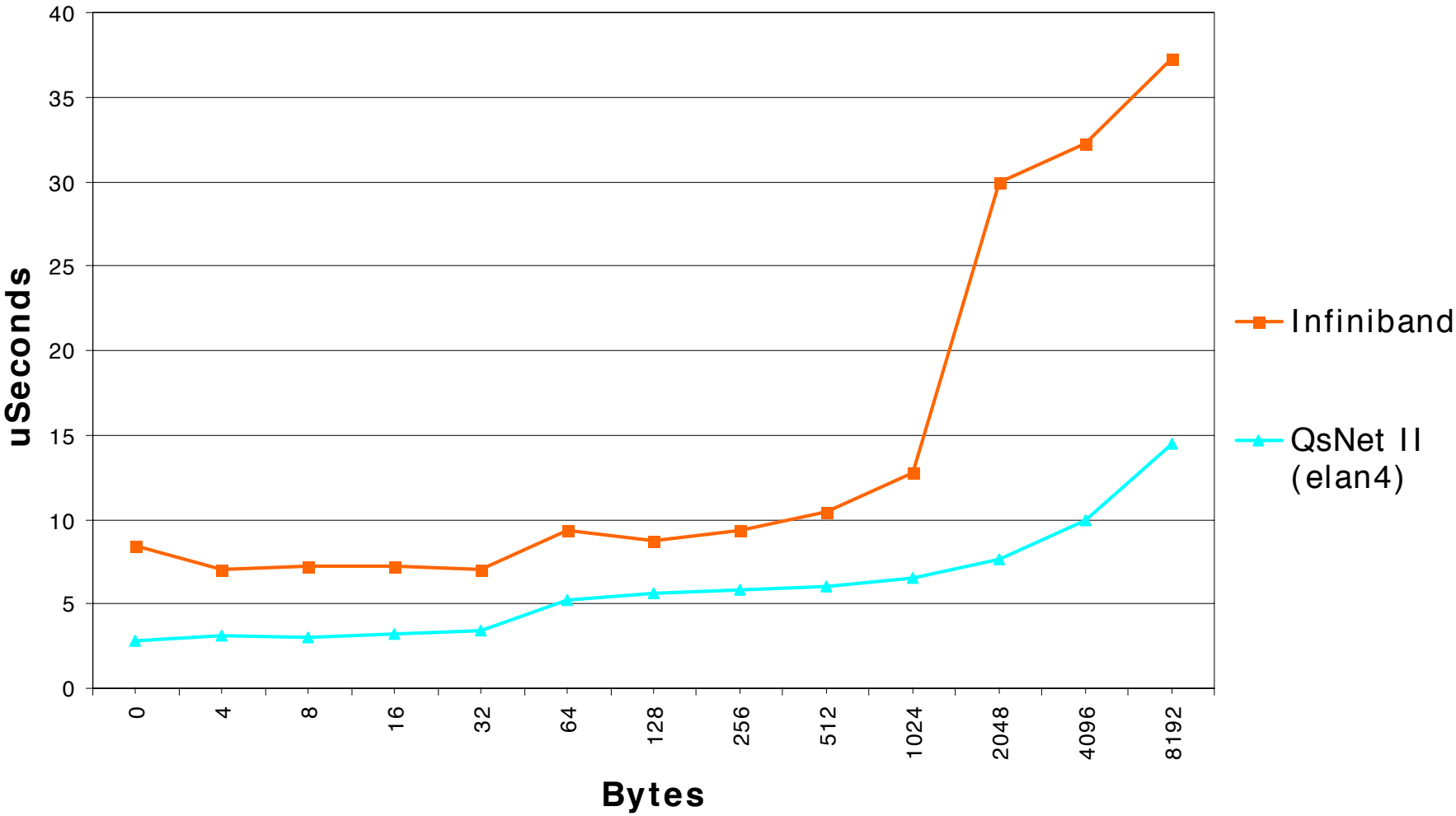


RDMA bandwidth

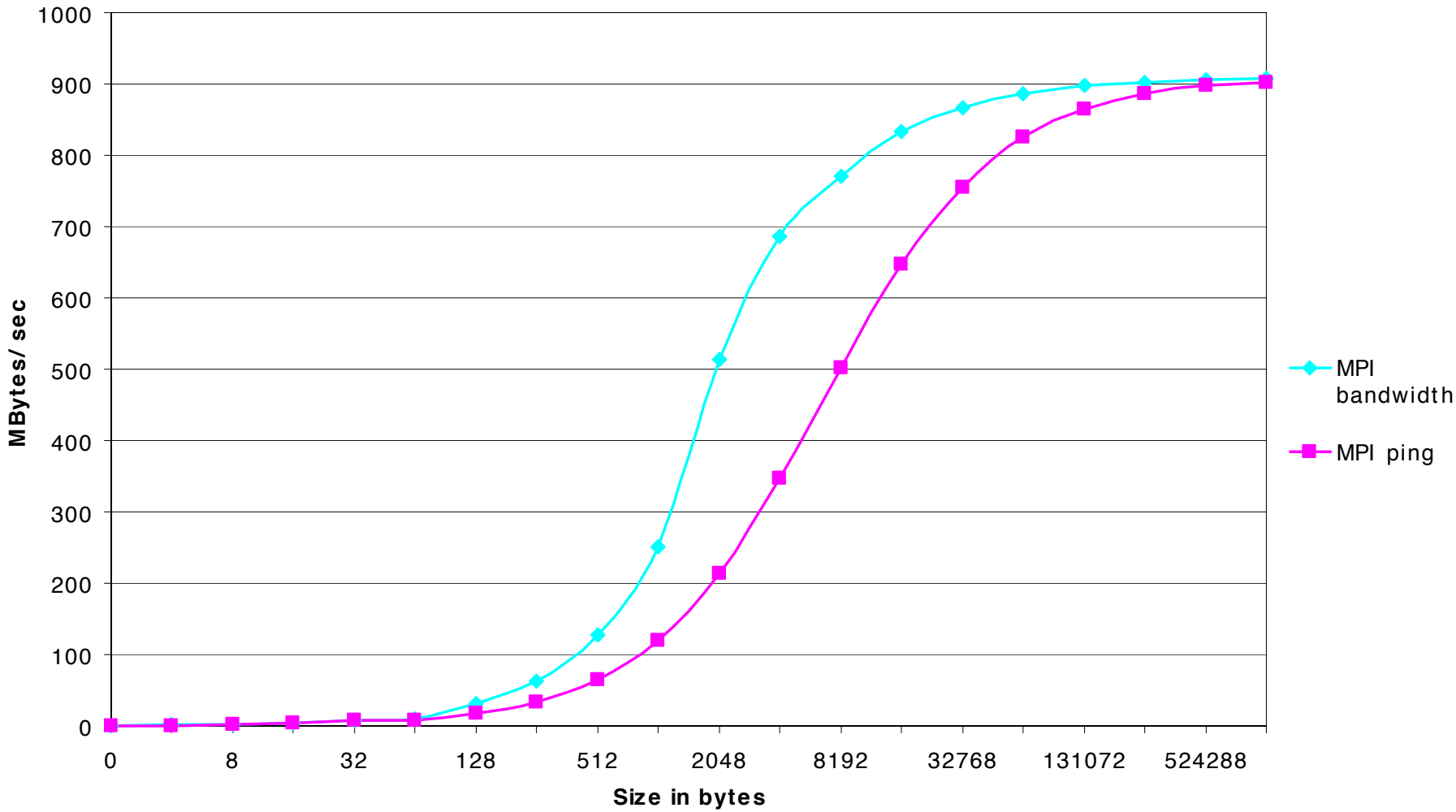
Batched RDMAs



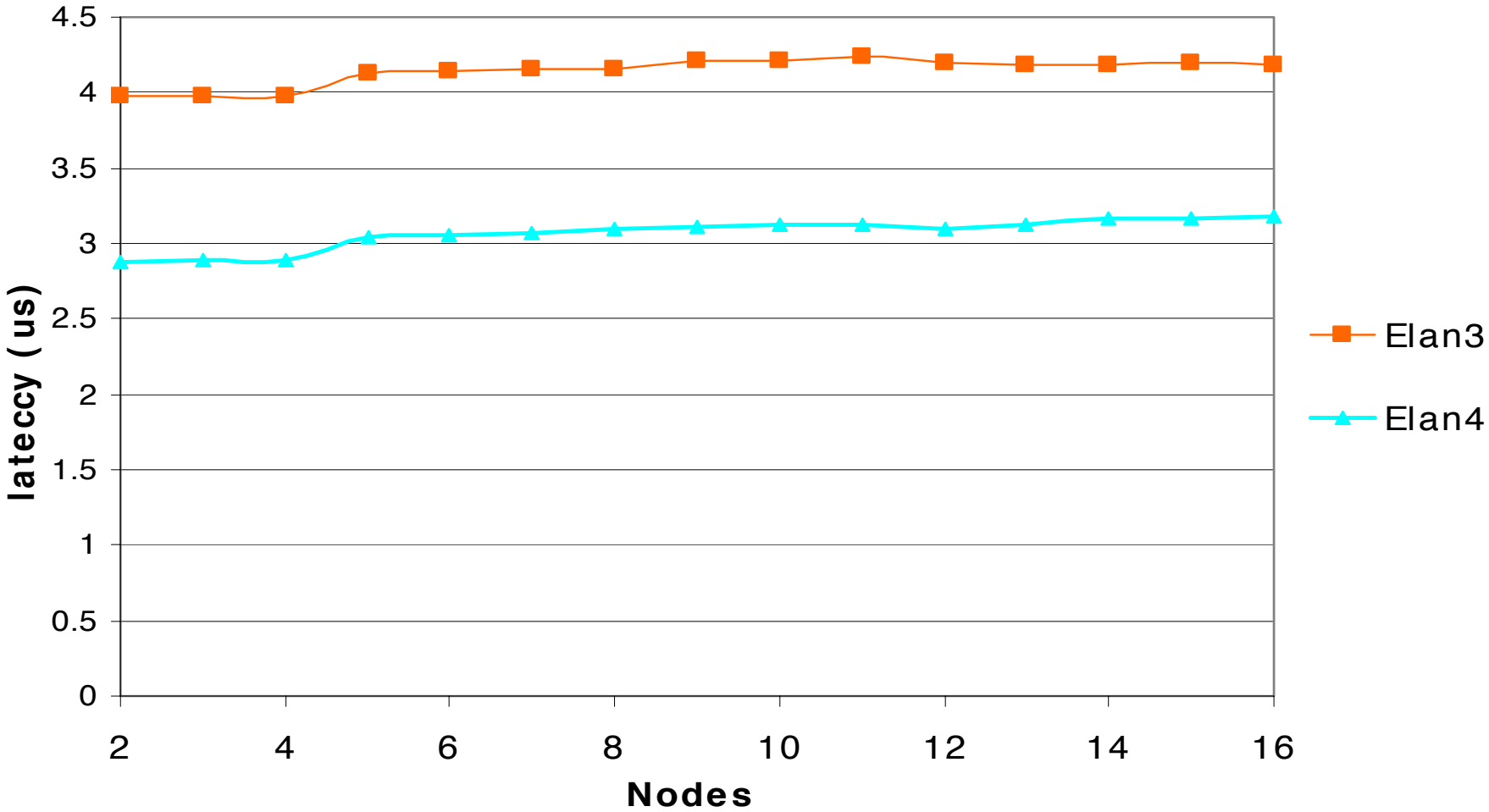
MPI short message latency



MPI Bandwidth - Elan4



Global operations - Hardware barrier scaling



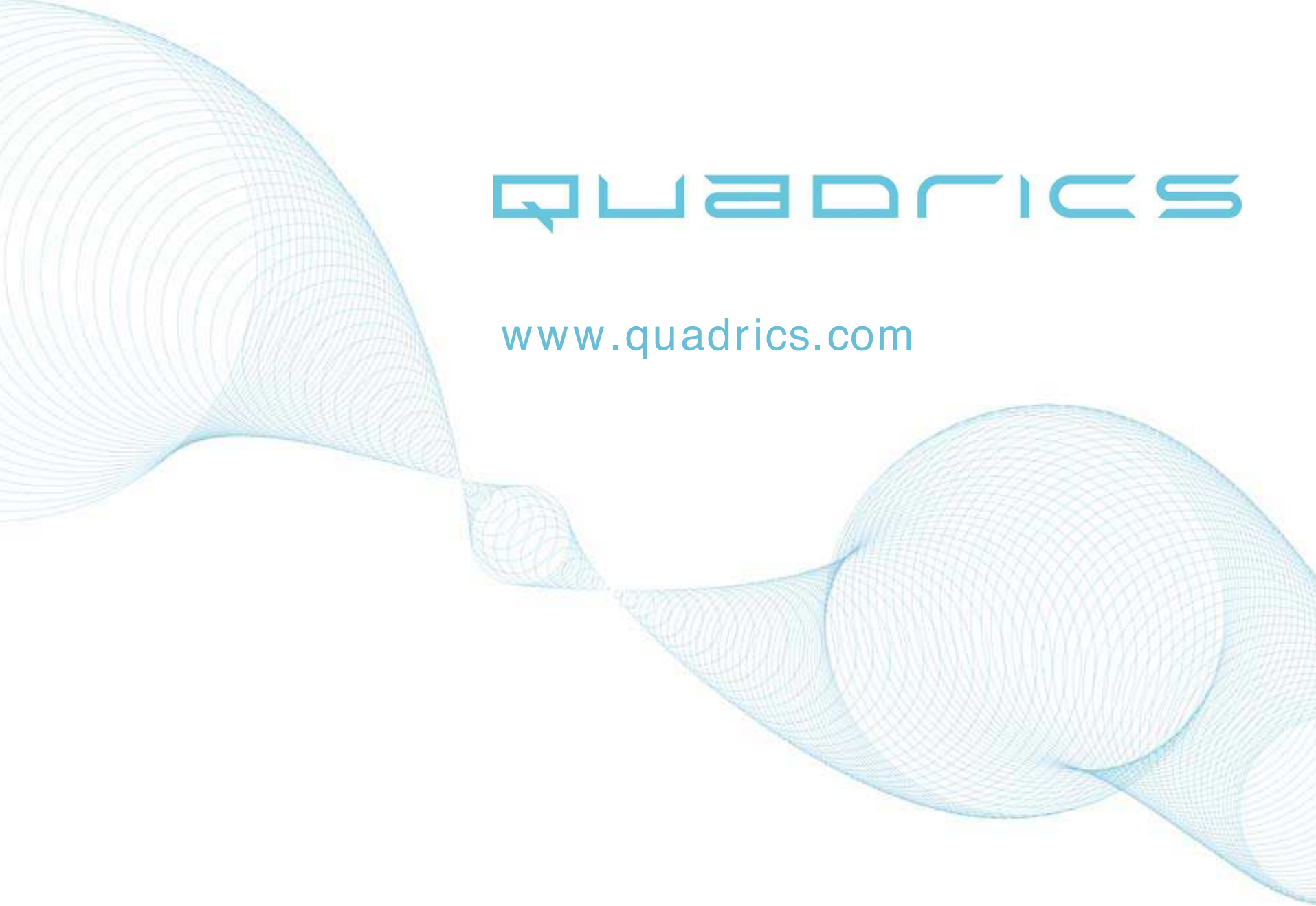
Acknowledgements

Los Alamos NATIONAL LABORATORY

Pacific Northwest National Laboratory
...delivering breakthrough science and technology

LSI LOGIC

hp invent


QUADRICS
www.quadrics.com