QsNet II and beyond

Moray McLaren

Networks for Supercomputers

QsNet characteristics

- Ultra low user process to user process latency
- Maximum available bandwidth on standard buses
- Seamless scaling to many 1000s of nodes
- High availability
- Reliable data transfer
- Mixed system and multiple user traffic on one network

QsNet unique features

- Operates on pageable virtual memory
  - No page lock down requirement
- Reliable hardware broadcast
  - Optimised global operations such a barrier synchronisation
- User programmable IO processor
  - Easy to implement multiple protocols
  - Minimize main processor interrupts

QsNet II Components

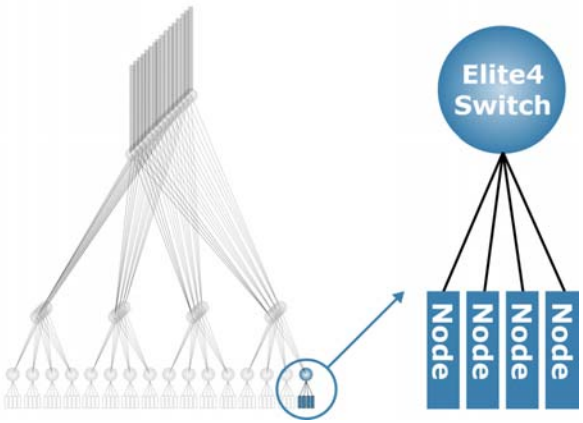
- Elan 4 network interface card

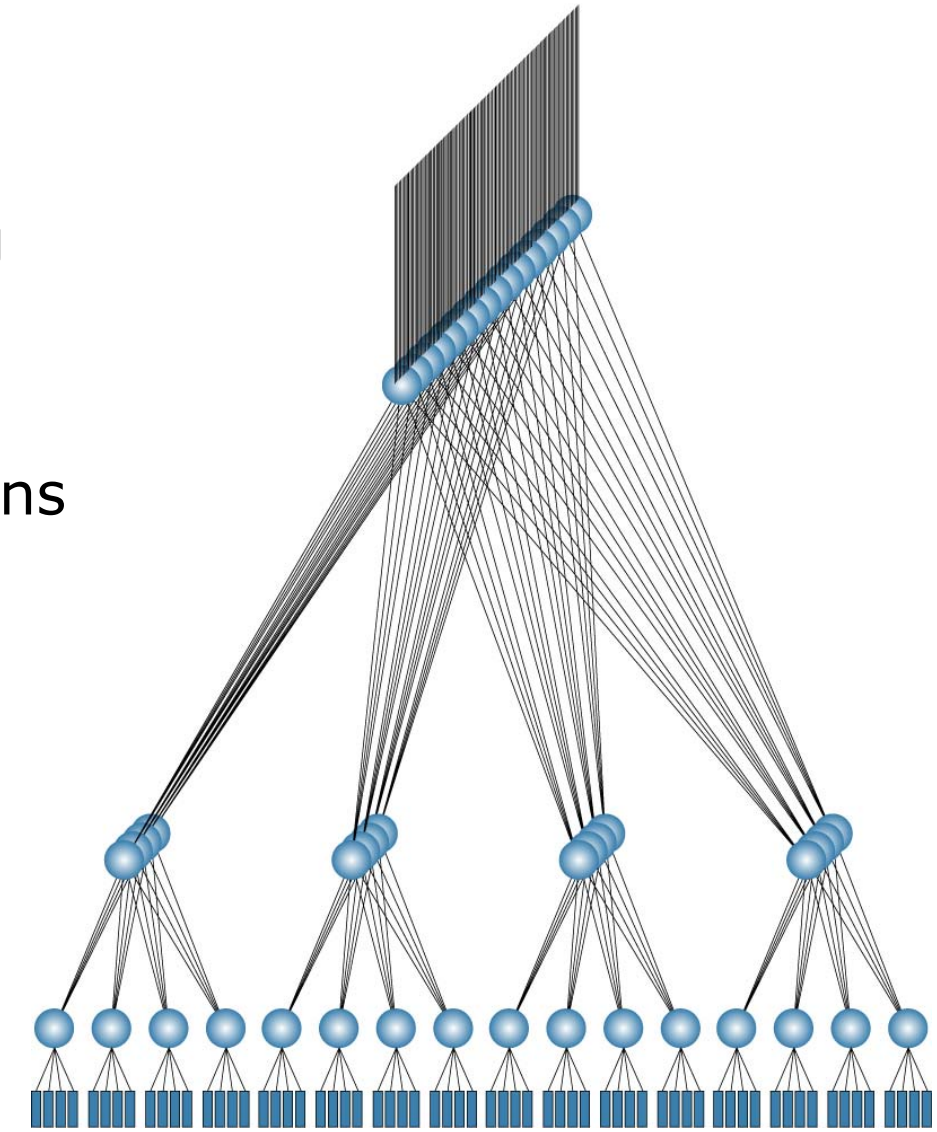
Elite 4 switch component

QsNet II Switch

Fat Tree Topology

- Benefits
  - Linear bandwidth scaling
  - Fault tolerance structure
  - Uniform connectivity
  - Supports global operations
  - Simple adaptive routing





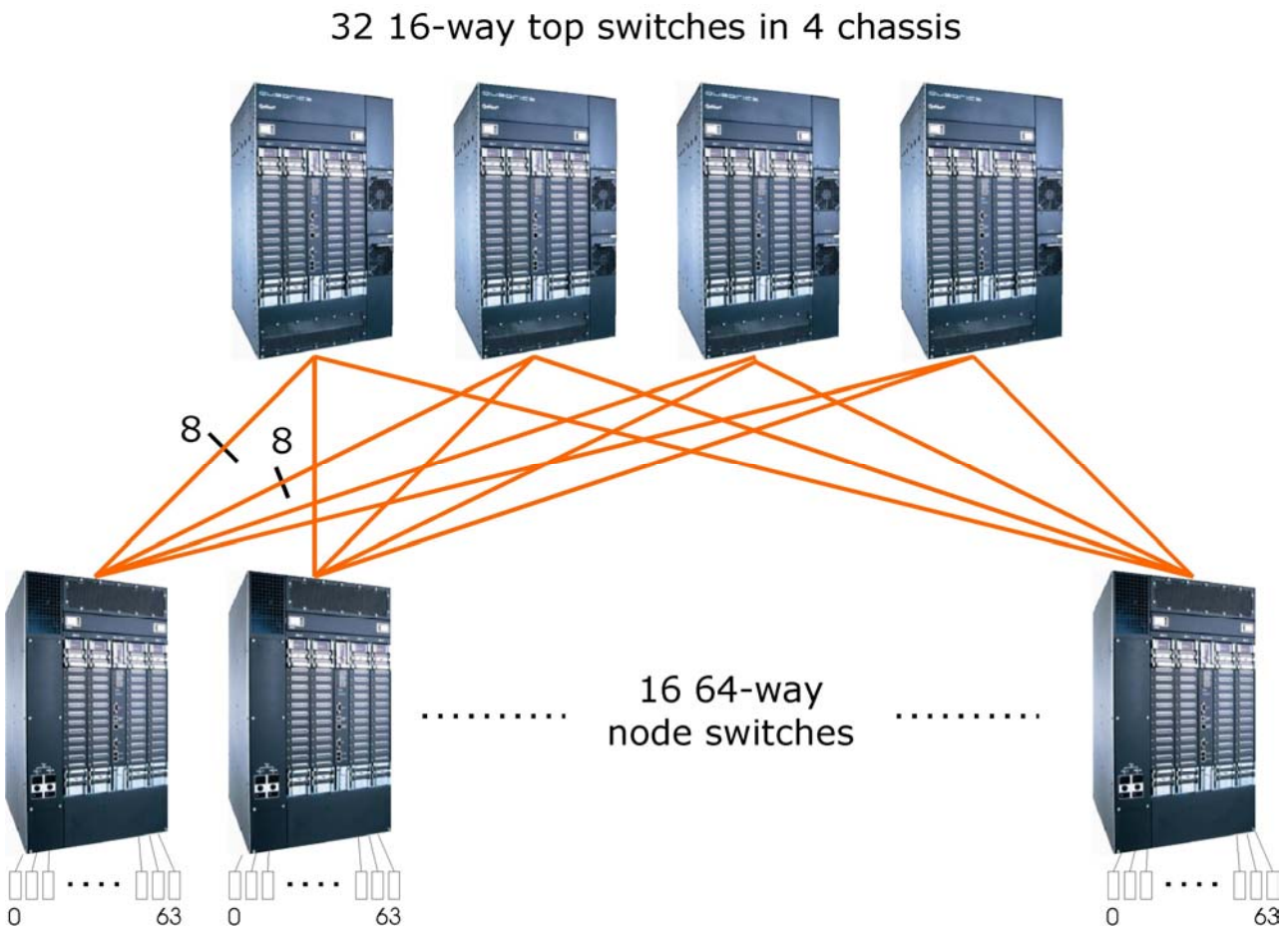
QsNet II Update

- Standalone Switches
- Fibre networks
- 2048-way switch
- Software developments

QsNet^{II} Product Development

- Introduction of low cost standalone 8/32/128-way switches

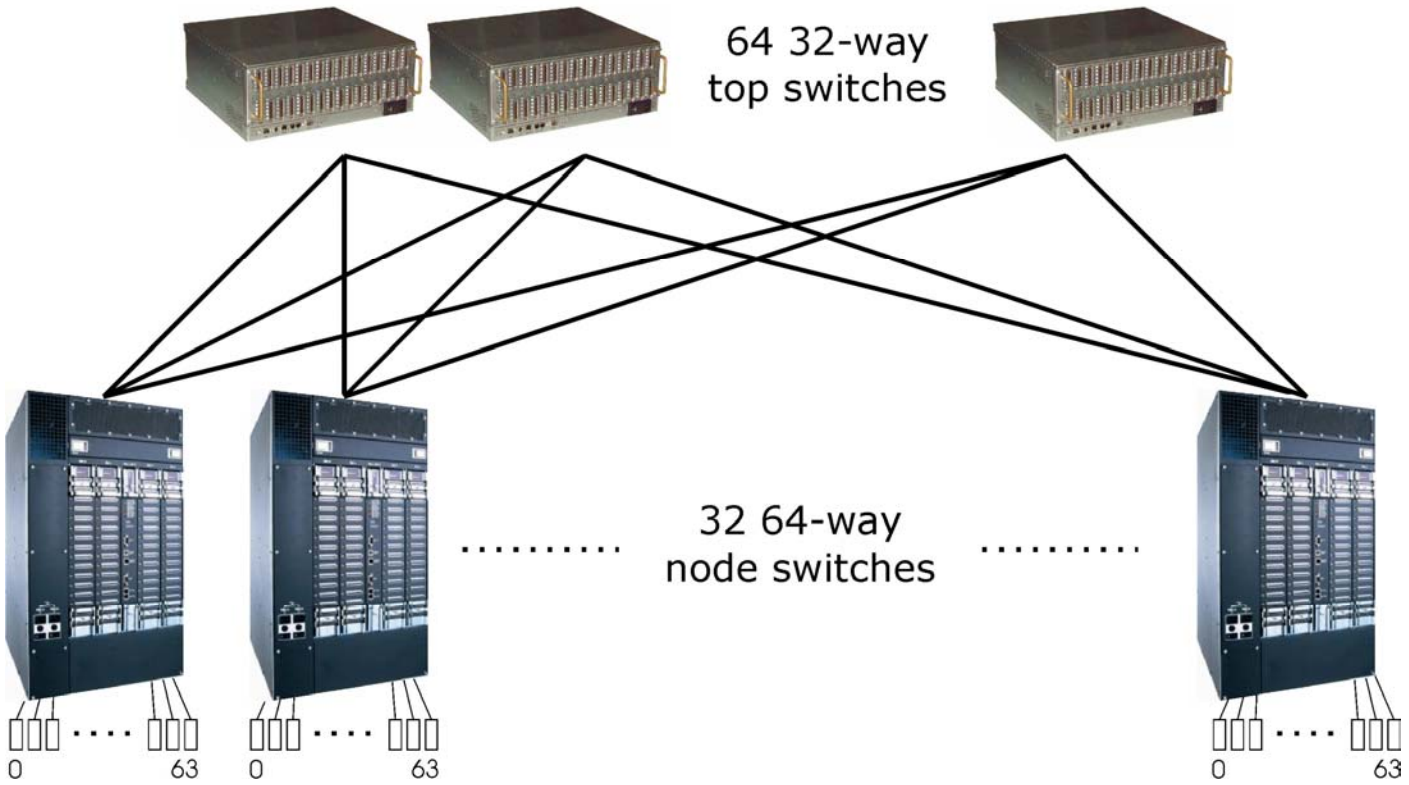
QsNet^{II} Development - Fibre Network



Roadmap for Fibre Parts

Dec 2003	Components in house
Sept 2004	Mechanical design of new faceplates
Nov 2004	Proto build of new cards for UL & EMC testing
Mar 2005	Approvals complete
Apr 2005	Production build

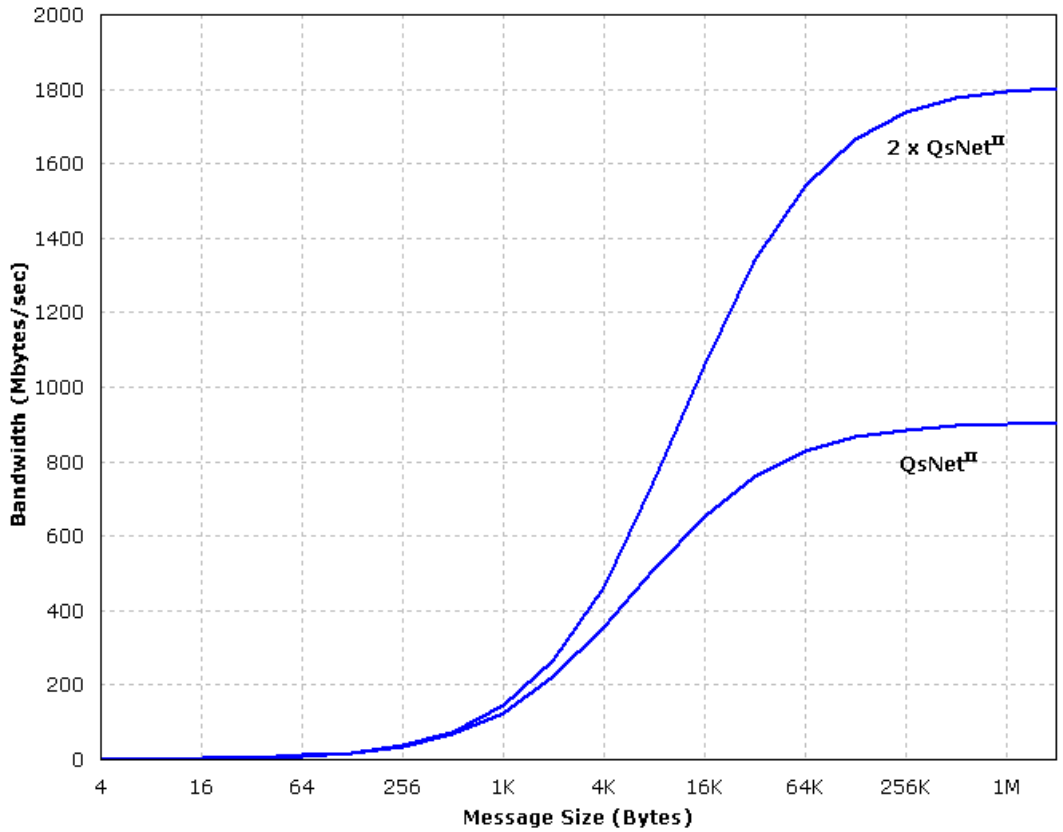
QsNet^{II} Development – 2K Port Network



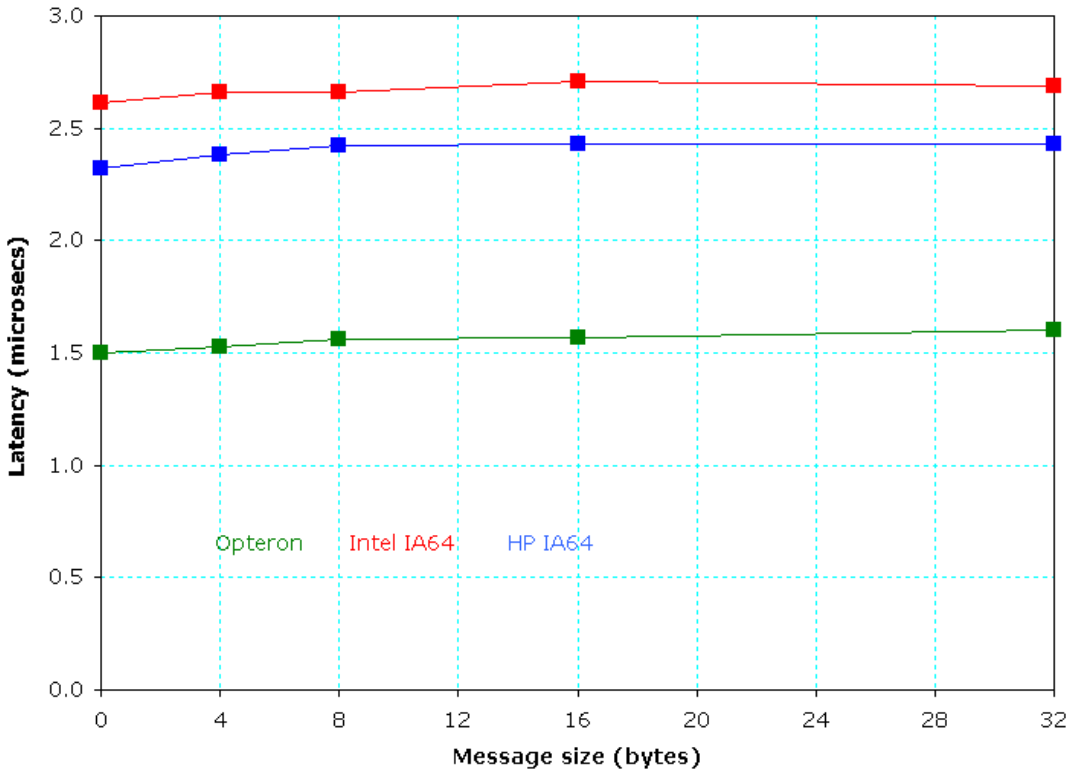
QsNet^{II} Performance Update

- MPI bandwidth & Latencies
- Event processor gather on a tree
- Alltoall optimisation
- Thread Processor Reduction
  - Ref Fabrizio Petrini & Adam Moody
- Pallas b_eff

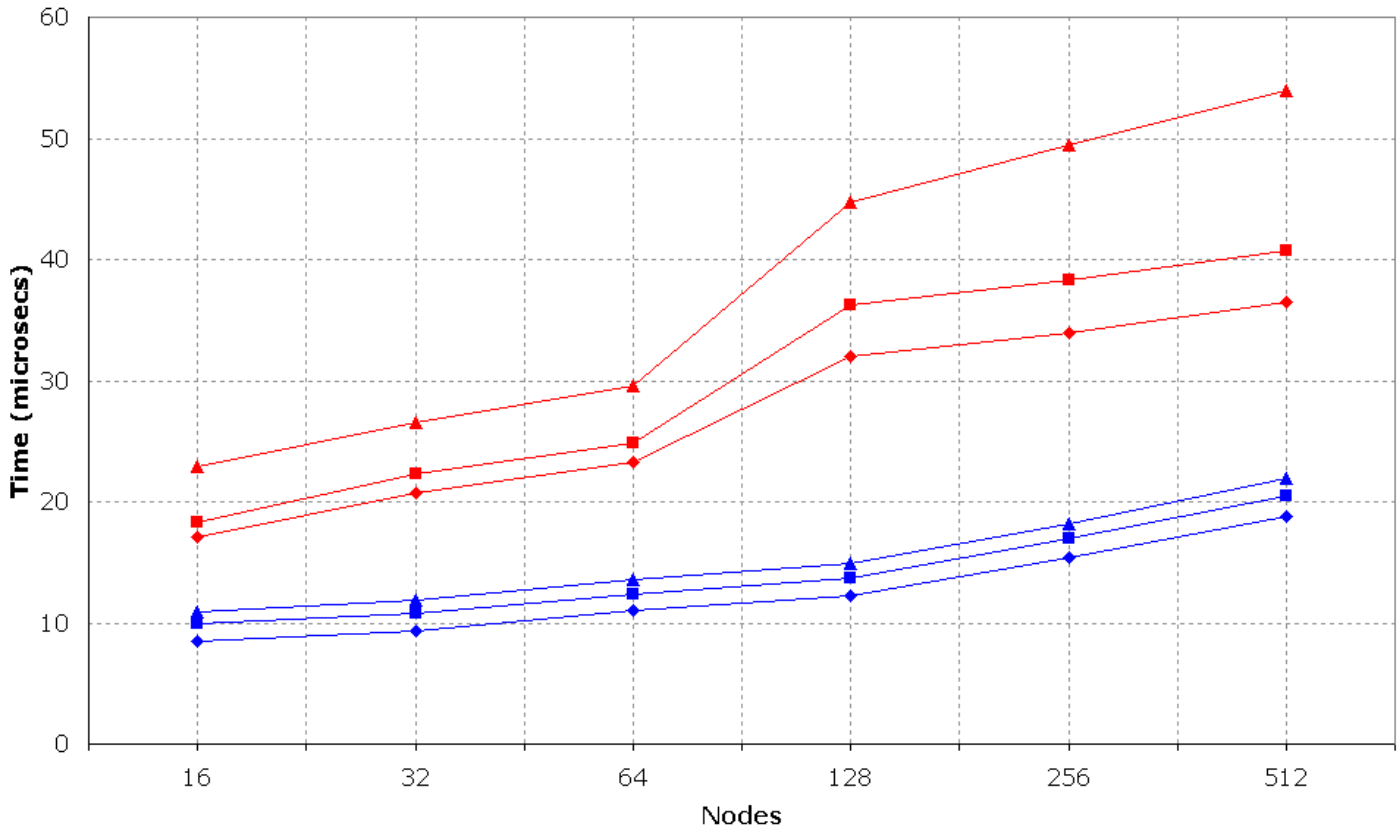
QsNet^{II} Performance – MPI Bandwidth



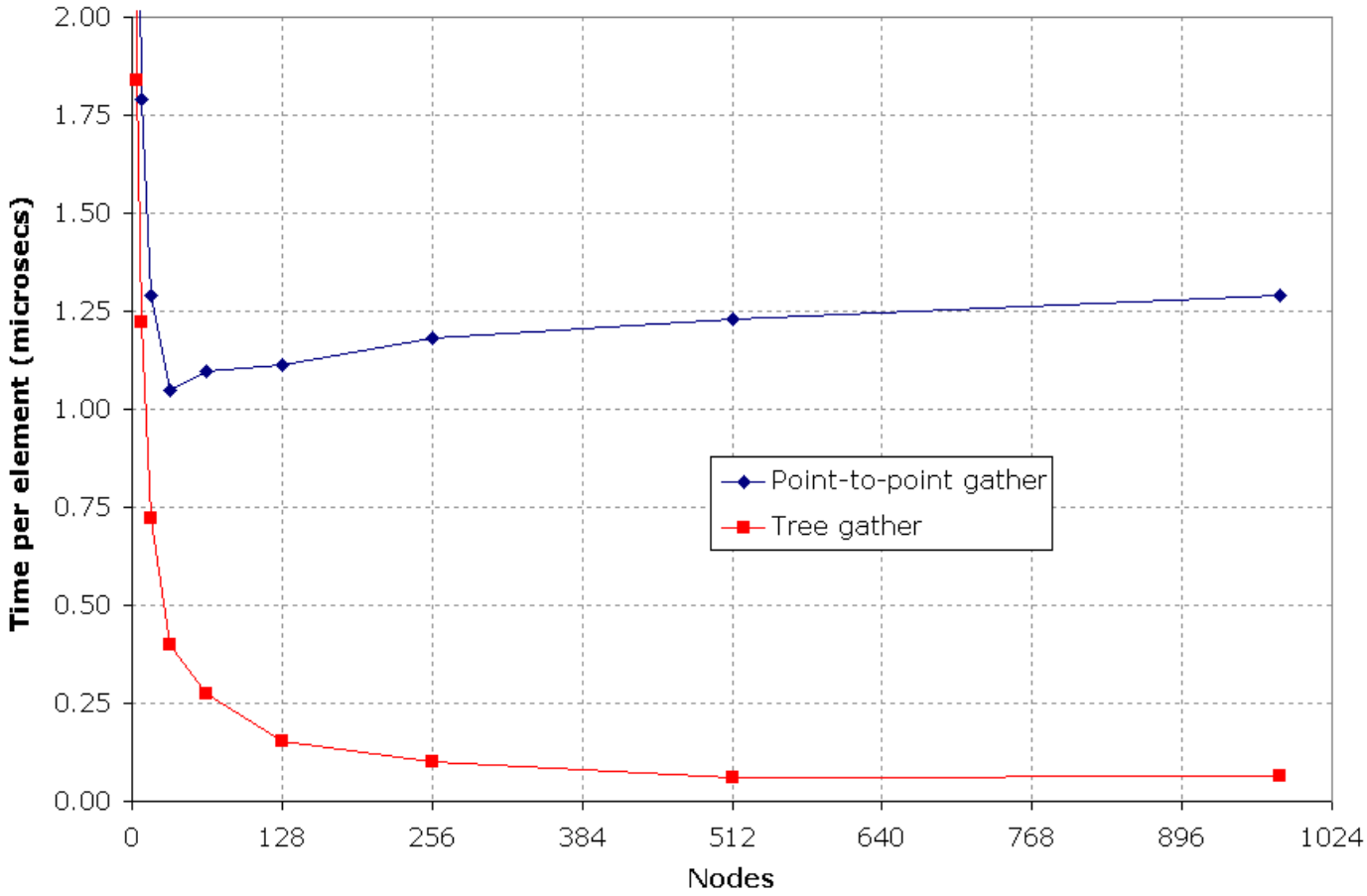
QsNet^{II} Performance – MPI Latency



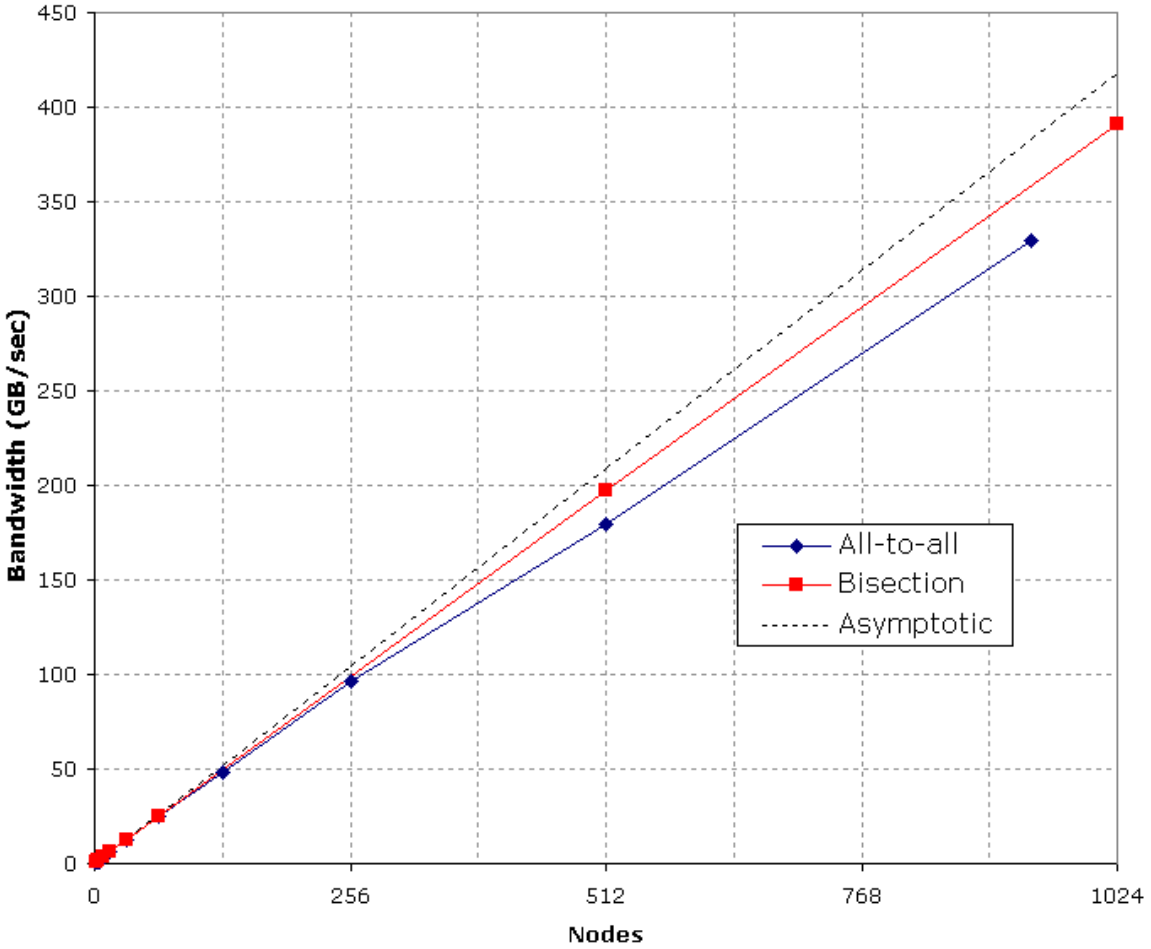
QsNet^{II} Performance – DSUM Reduction



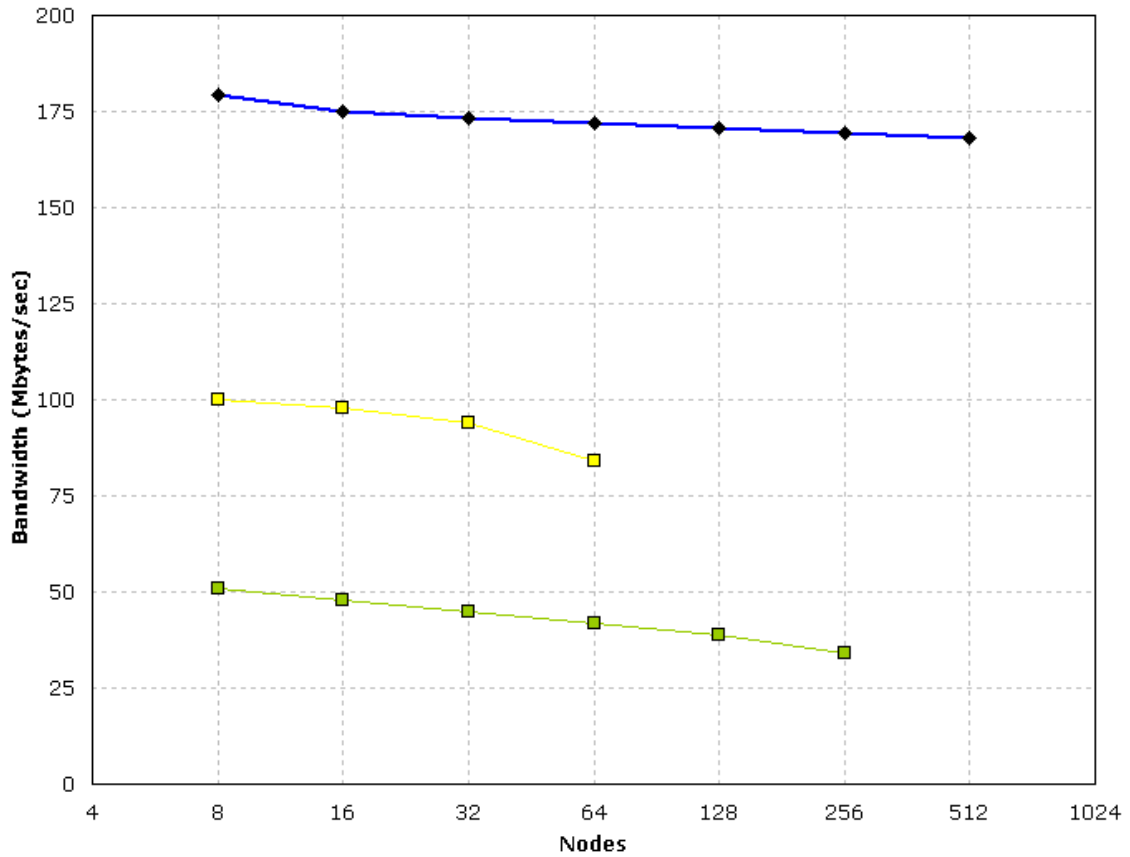
QsNet^{II} Performance – Event Gather



QsNet^{II} Performance – All to All



QsNet^{II} Performance – b_eff



QsNet^{II} Software Development

- Lustre support for multiple rails
  - Bandwidth
  - Transparent rail failover
- “patch free kernel”
  - Uses pin down cache
  - Single source tree

Next generation - QsNet III

- Performance
  - Need to offer the highest performance in our chosen application space
  - REAL application performance not just spec sheet.
- Standards
  - We need to be able to offer standards based solutions
- Re-use
  - Design for re-use of silicon IP, easily develop chip variants

Costs reduction - the real challenge

- Driven by continuing reduction in node costs
- NIC costs
  - Volume is everything. Supercomputing market not enough on it's own.
- Fabric costs
  - 40% ASIC, 30% cables 30% other stuff..

Elan 5 program

- Design objectives
  - Utilise new CPU interfaces – PCI Express
  - Provide range of bandwidth options
  - Maintain position as lowest latency interconnect
  - Generalise processor to support alternate protocols
  - Improve support for mixed system and user traffic