

## Hot Chips Panel – August 2003

Moray McLaren  
*Quadrics Ltd.*

## EtherNet and EtherNot!

*Will standard interconnects solve all our problems?*

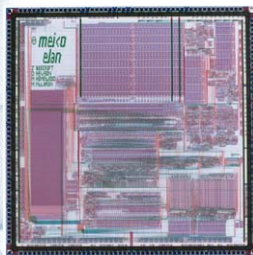
- Whatever the volume interconnect of the future is, it will be called Ethernet.
- Incorporate ideas from specialised low latency interconnects into Ethernet?
  - RDMA is a start
  - Common DDI with high performance NICs?
  - Price advantage not so clear for equivalent BW.
- Successful EtherNot technologies need clear performance advantages that deliver in applications.

## What's special about Supercomputing?

- Pushing the extremes of scale
  - Seamless switch scaling
  - Global operations
  - Fault tolerance
- They still count compute cycles
  - Compute communications ratio
  - Ultra low latency
  - Overhead

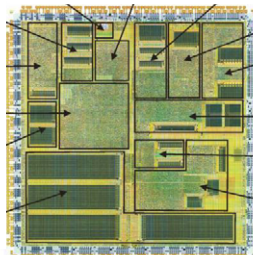
## Historical scaling...

Elan – 1990



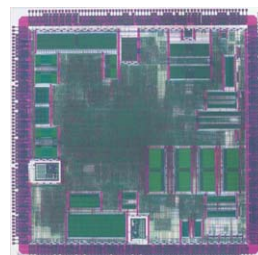
Put - 9 $\mu$ s  
MPI - 78 $\mu$ s  
44Mbytes/s

Elan 3 – 1998



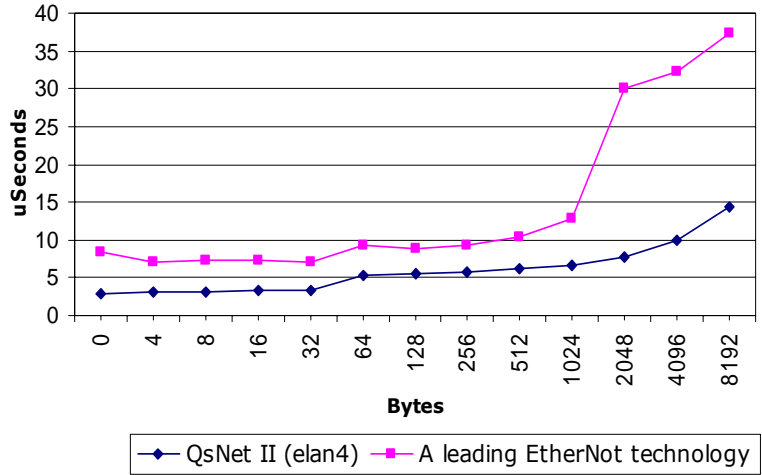
Put - 2 $\mu$ s  
MPI - 5 $\mu$ s  
320Mbytes/s

Elan 4 – 2003

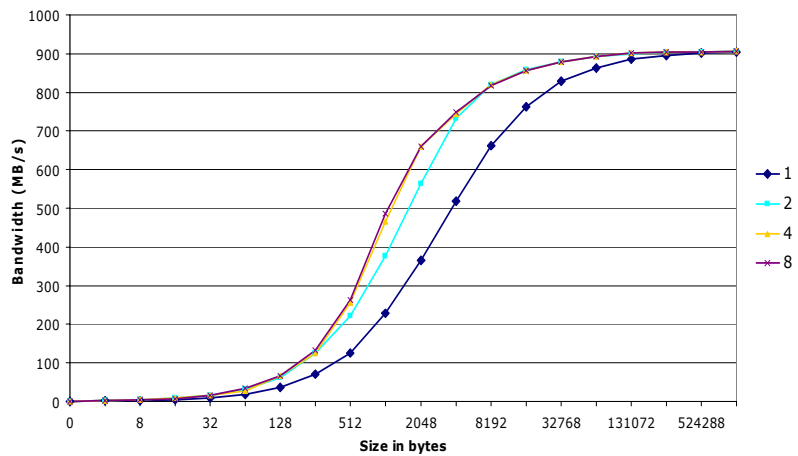


Put - 1.7 $\mu$ s  
MPI - 3 $\mu$ s  
900Mbytes/s

### MPI short message latency



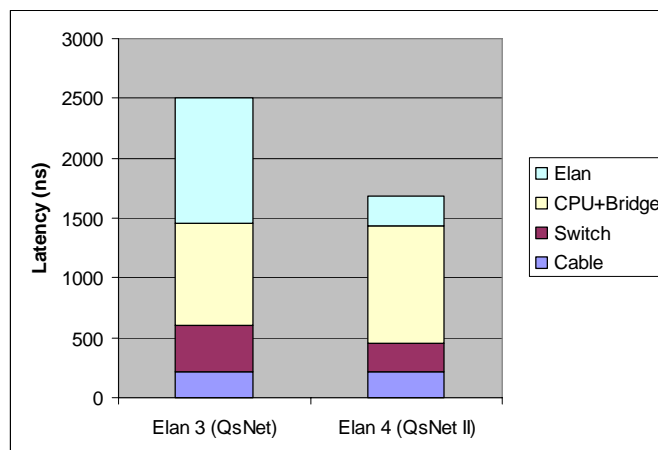
### MPI Bandwidth – Elan 4



## Measuring performance

- Latency measurement
  - Lowest latency implementation have highest CPU overhead
  - Throughput as important as basic latency
- Bandwidth measurement
  - Measure overhead as well as bandwidth
  - Alignment issues
- Network performance
  - Bisectonal bandwidth
  - Variation in latency across node space
  - How does it relate to the application performance?

## Elan-4 Latency Breakdown



## The way forward on latency

- **Basic hardware latency**
  - Many factors reaching practical limits.
  - Closer integration to CPU removes some delays
  - Pipelining to support multiple outstanding short messages
- **Real application latency**
  - MPI well understood
  - Lower level API need for compilers etc.
  - What's the API for kernel messaging?
  - Reliably, ordered, datagram.
  - Several alternates, Portals, Via constructs..

## Bandwidth going forward

- **Limited by where you can connect to**
  - Double and Quad clock PCI-X
  - PCI-Express
  - Direct connections.
- **Large scale multi rail systems with large SMPs**
  - NUMA challenges

## QsNet<sup>II</sup> Physical Link

- 1.333Ghz design speed
  - 4b5b coding for DC balance
  - ~900 Mbytes/s after protocol
- Copper
  - 10 bit lvds – total 40 wires
  - 10-12m range
- Optics
  - 12 bit parallel optical fiber
  - 100m



## Future link technologies

- Still copper on the backplane for cost and reliability
  - Careful design gets to up to 5Gbit/s per wire for moderate runs. More with clever equalisation.
  - Max length decreases as speed increase
  - Improved packing technology to reduce connection lengths, pack more into the copper zone.
- Rack to rack all fibre
  - Future generations of parallel fibre
  - 12 x 5 Gbits/s ~ 6Gbytes/s

## Optical switching ?

- Optical technology has been driven by telecomms requirements
  - Long haul not short haul
  - Circuit switched not packet switched
  - They're not buying anything!
- Combining logic, switching and buffering – easy for silicon
  - Silicon switches – a distributed arbiter which delivers data as a side effect.
- To best exploit optical switches need to consider radically different architectures.