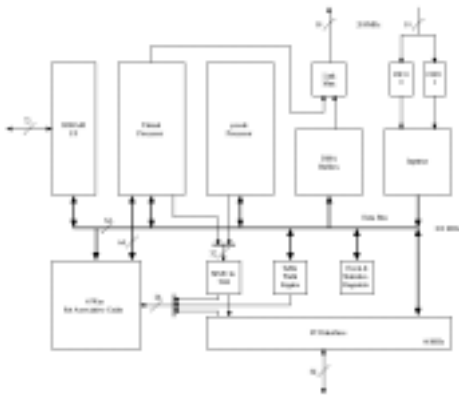


- Quadrics network
  - Fabrizio Petrini HotI papers (01 and 03)

## Quadrics Network

- Target
  - parallel supercomputers, very low latency
  - scalable in terms of bandwidth, #links with #nodes, cost of send with #receivers...
  - hardware/software support for collective communication
- Parallel applications
  - user-to-user level low latency
  - high aggregate bandwidth
  - collective communication/global operations barriers
    - broadcasts, hotspots, integer collectives...

## Elan3 NIC



## Elan3 Features

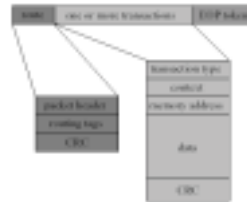
- ucode processor – dedicated threads:
  - inputter, DMA, processor-scheduling, command-processor thread
- thread processor – programmable, RISC proc, specialized instructions, efficient h/w supported context switch
- MMU translate VA to SDRAM physical or PCI physical;
- Routing tables for route translation based on virtual processor number

## 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



## Quadrics packets

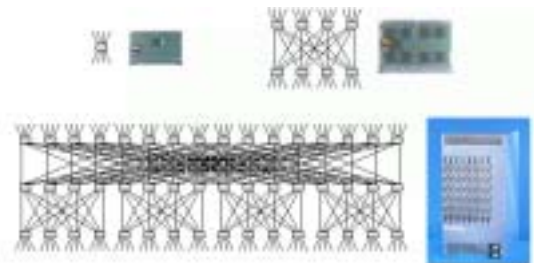


- router path in header -> source driven routing
- wormhole routing –
  - packet in buffers.
  - size 320B
  - flit-level flow control (16B)

## Elite Switch

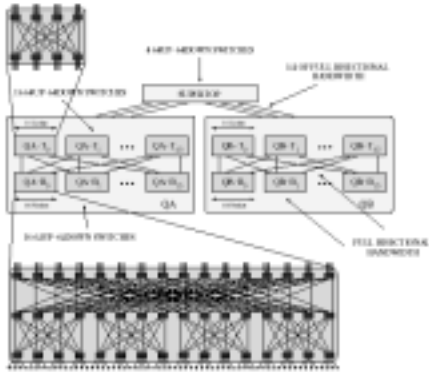
- basic unit
  - 8 bidirectional links with 2 virtual channels
  - broadcast to range of outputs
  - full automatic error detection / recovery
  - arbitration based on age of packet
  - two levels of priority
  - adaptive routing support
  - unblocked latency of ~20ns
  - traceroute transaction for interrogating the network
- fat-tree topology
  - scalable
  - fault-tolerance – multiple paths
  - broadcast/multicast operations

## Fat-tree topology

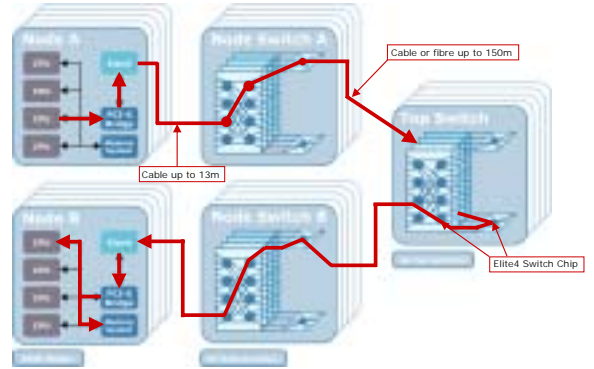


- single Elite
- 16 up/16 down – level 2 fat tree
- 64 up/64 down – level 3 fat tree

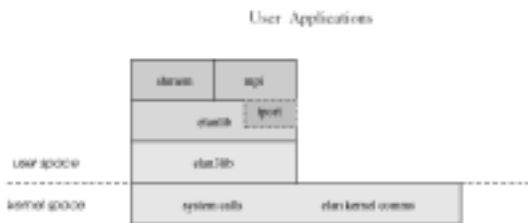
## Network topology



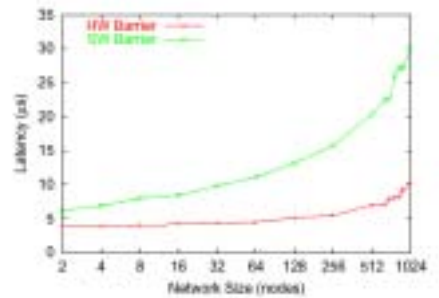
## A Process Communication

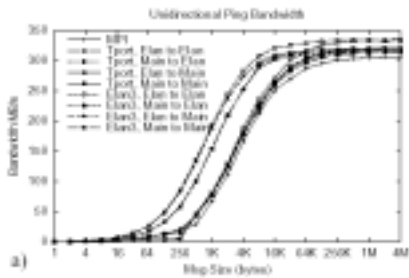


## software stack

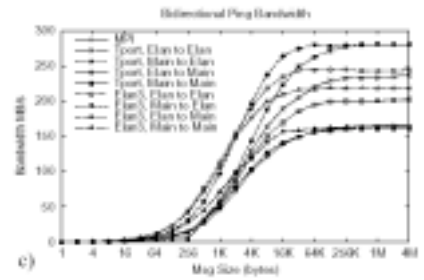


## barrier performance w/ Elan3

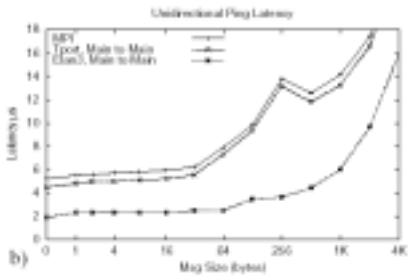




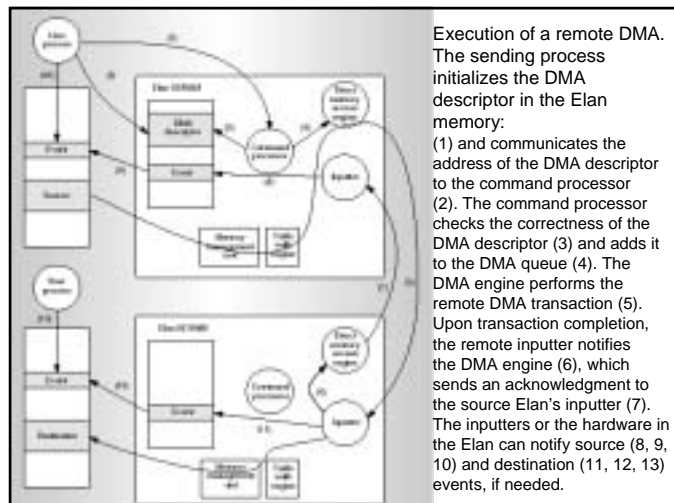
peak for elan – elan memory  
 main-elan < elan-main => PCI write < PCI read



bidir BW <math>\diamond</math> unidir BW -> bottlenecks in network and network I/F



message `envelopes' for smaller messages...



Execution of a remote DMA. The sending process initializes the DMA descriptor in the Elan memory: (1) and communicates the address of the DMA descriptor to the command processor (2). The command processor checks the correctness of the DMA descriptor (3) and adds it to the DMA queue (4). The DMA engine performs the remote DMA transaction (5). Upon transaction completion, the remote inputter notifies the source Elan's inputter (7). The inputters or the hardware in the Elan can notify source (8, 9, 10) and destination (11, 12, 13) events, if needed.