

Overview of Interconnects

Myrinet and Quadrics

Leading Modern Interconnects

Presentation Outline

- General Concepts of Interconnects
- Myrinet
 - Components
 - Communication features
 - Latest Products
- Quadrics
 - Components
 - Communication features
 - Performance
 - Latest Release
- Our Research

Interconnects

- Shared-medium Interconnects
 - LAN (Ethernet)
- Router-based Interconnects
 - Intel Paragon, Cray T3D, Cray T3E
- Switch-based Interconnects
 - Myrinet, Quadrics, InfiniBand

Switch-Based Interconnects

- Link Fiber or Cables
- Network Interfaces
- Switches: Crossbar Switches
- Interconnection of Switches

Interconnects Issues

- Communication Features
 - Basic issues:
 - bit encoding
 - framing
 - switching/routing
 - flow control (deadlock)
 - error-control (reliability)
 - Advanced Issues:
 - Memory management
 - Message passing semantics
 - offloading of the protocol processing
 - Multiple rails, message striping
- Performance and Scalability

Basic Switching Unit (Crossbar Switch)

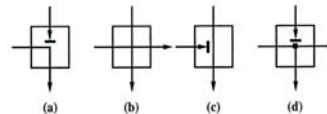


Figure 1.12: States of a switch point in a crossbar network

Switching Technology

- Circuit Switching
- Packet Switching
- Virtual Cut-through
- Wormhole Switching

Basic Switching

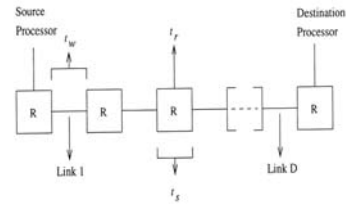


Figure 2.5: View of the network path for computing the no load latency

Circuit Switching

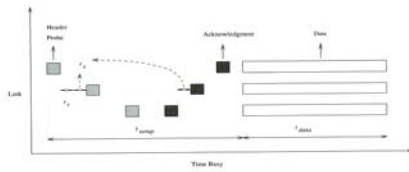


Figure 2.6: Time space diagram of a circuit switched message

$$t_{circuit} = t_{setup} + t_{data} \text{ where}$$

$$t_{setup} = D * (t_r + 2 * (t_s + t_w))$$

$$t_{data} = \frac{1}{B} * \left\lceil \frac{L}{W} \right\rceil$$

Packet Switching

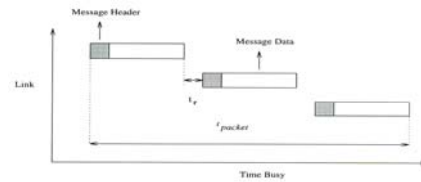


Figure 2.8: Time space diagram of a packet switched message

$$t_{packet} = D * \left(t_r + (t_s + t_w) * \left\lceil \frac{L+W}{W} \right\rceil \right)$$

Virtual Cut-Through Switching

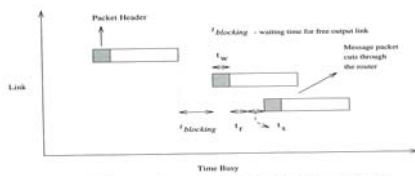


Figure 2.10: Time space diagram of a virtual cut-through switched message

$$t_{vct} = D * (t_r + t_s + t_w) + \max(t_s, t_w) * \left\lceil \frac{L}{W} \right\rceil$$

Wormhole Switching

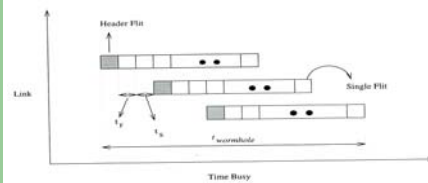


Figure 2.11: Time space diagram of a wormhole switched message

$$t_{wormhole} = D * (t_r + t_s + t_w) + \max(t_s, t_w) * \left\lceil \frac{L}{W} \right\rceil$$

Blocking in Wormhole Network

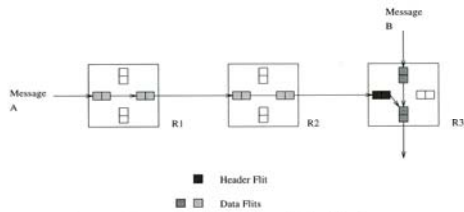


Figure 2.12: An example of a blocked wormhole switched message

Virtual Channels

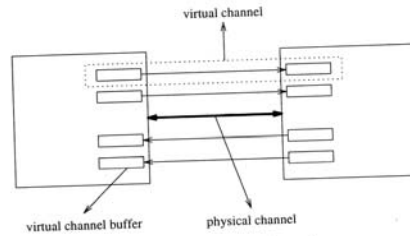


Figure 2.17: Virtual channels

Presentation Outline

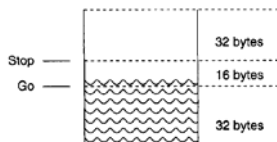
- General Concepts of Interconnects
- Myrinet
 - Components
 - Communication features
 - Performance
 - Latest Release
- Quadrics
 - Components
 - Communication features
 - Performance
 - Latest Release
- Our Research

Myrinet Origin (www.myri.com)

- Mosaic
 - High data rates
 - Regular topology and scalability
 - Very low data rates
 - Cut-through routing
 - Flow-control at every link
- Atomic LAN
 - Achieve high data rates 10^{15} bits per second
 - Limitations:
 - asynchronous signaling
 - complex mapping
 - lack of DMA engine
 - multiple copies through TCP/IP stack

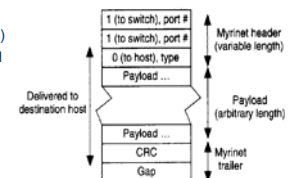
Myrinet Links

- Cable links
 - 18 twisted pairs, nine in each directions
 - Synchronous transmission, avoids asynchronous signaling
 - Maximal 25m cables
- Flow Control
 - Receiver Driver
 - Slack Buffer
 - Stop and Go signals



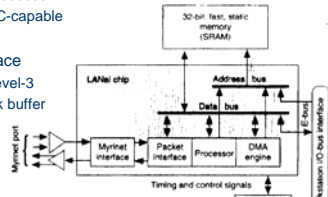
Myrinet Packets

- Packet Format
 - Header (up to 24 bytes)
 - Arbitrary length payload
 - CRC, error-control
 - Gap



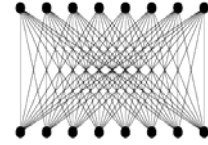
Myrinet Network Interface

- Host Interface
 - Programmable Processor
 - DMA engine, CRC-capable
 - Packet interface
- Optical-Fiber Interface
 - OSI level-2 and level-3
 - ~1500 bytes slack buffer



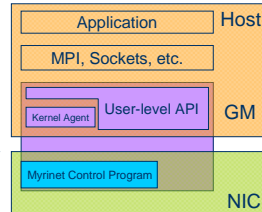
Myrinet Switch

- Basic Unit
 - Crossbar switch
 - Worm-hole routing
 - Easy network-mapping
 - 550ns switch latency
- Topology
 - Clos Network
 - Full bisectional bandwidth
 - Easy Connections into larger network



Software Stack

- MCP
 - running on the host interface
 - Perform continuous mapping, monitoring and route updating
 - IP Multicast capable
- GM
 - kernel module
 - user-level API
 - Provide interface between user processes and NIC
- Programming Libraries
 - MPI, sockets, etc.



Myrinet Products

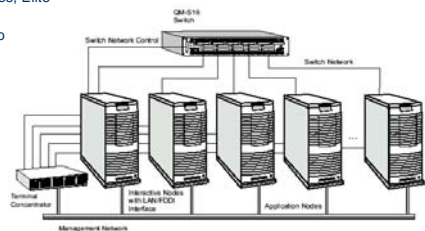
- Cutting-edge interconnect technology for many years (Many TOP500 systems during 1995-2002)
- High performance, low latency and highly reliable
- Self configurable and fault-tolerant
- Capable of being I/O Fabric
- Ideal for cluster-computing
- Recently moved to a dual strategy
 - Proprietary adapter
 - 10GbE adapter

Presentation Outline

- General Concepts of Interconnects
- Myrinet
 - Components
 - Communication features
 - Latest Products
- Quadrics
 - Components
 - Communication features
 - Performance
 - Latest Release
- Our Research

Quadrics Components

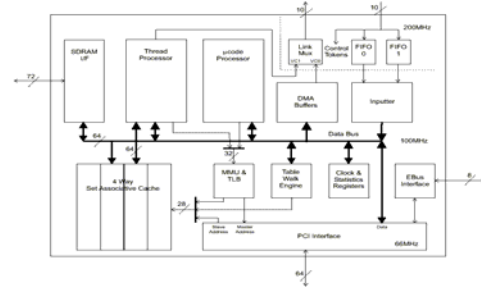
- Hardware Components
 - Network Interfaces, Elan 3
 - Switches, Elite
- Software
 - elan3lib
 - elanlib



Network Interface

- Link physical layer
 - Full duplex 10 bit, 400Mbaud Link
- Elan 3 (QM400) Network Adapter
 - 64 bit/66MHz PCI Bus
 - Programmable I/O processor
 - Support Multiple threads 100MHz
 - Integrated DMA engine
 - Automatic packetisation and scheduling
 - Dedicated input packet processing engine
 - 8KB on chip cache
 - 64MB SDRAM with MMU + TLB
 - Supported OS: Tru64 UNIX™ and Linux™
 - Communication Libraries
 - MPI, Shmem, kernel messaging & IP

NIC: Elan 3



Microcode Processor

- Control Processor for Elan 3
- Execute four threads
 - Command processing
 - Thread scheduling
 - Inputter thread
 - DMA thread

Thread Processor

- Basic Features
 - 100 MHz
 - 32 bit RISC
 - Extended instruction set
 - 4-stage pipeline
 - 32 registers
- Execute user threads
 - Provide NIC programmability

Other Processors

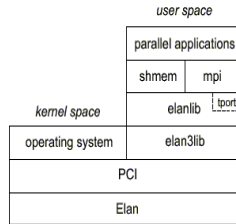
- Input Processor
 - Processing network packets,
 - Assemble data into transactions
 - Initiate the transactions for Microcode Processor
- DMA Processor
 - Service user RDMA read and write requests
 - Handle arbitrary source/destination buffer alignment
 - Support broadcast/flood and Queue DMAs

Memory Management

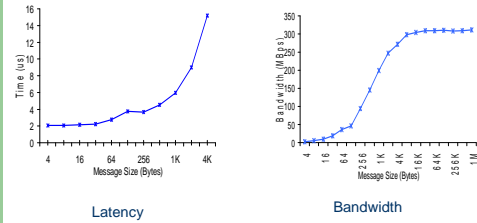
- 64 MB SDRAM
- 8K 4-way Set Associative Cache
- MMU
 - Address Elan or Main Memory
 - Synchronized with Main Memory
 - 16-entry TLB
 - Table Walk Engine

Communication Libraries

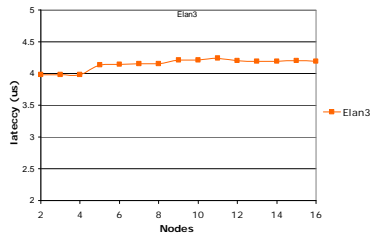
- elan3lib:
 - Basic Communications
 - Hardware-related
- elanlib:
 - Hardware Independent
 - Tagged Message Passing
 - Collective Communications
 - Broadcast, Barrier, Reduce



Performance (Elan-level)



Barrier with hw/bcast



Later Products

- QsNet-II (Elan 4 and Elite 4)
 - PCI-X
 - Link rate (1.333Gbaud)
 - 200MHz IO processor
 - MMU (128-entry TLB, 64-bit addressing)
 - MPI latency < 3μs
 - Bandwidth 900Mbytes/s
 - Max system size > 4K nodes
- Moved to 10GbE world

Presentation Outline

- General Concepts of Interconnects
- Myrinet
 - Components
 - Communication features
 - Latest Products
- Quadrics
 - Components
 - Communication features
 - Performance
 - Latest Products
- **Our Research**

Myrinet

- Active Network Interface Support
 - A. Gulati, D. K. Panda, P. Sadayappan, and P. Wyckoff, [NIC-based Rate Control for Proportional Bandwidth Allocation in Myrinet Clusters](#), ICPP '01
 - S. Senapathi, B. Chandrasekharan, D. Stredney, H.-W. Shen, and D. K. Panda, [QoS-aware Middleware for Cluster-based Servers to Support Interactive and Resource-Adaptive Applications](#), HPDC '03
 - D. Buntinas, D. K. Panda, J. Duato, and P. Sadayappan, [Broadcast/Multicast over Myrinet using NIC-Assisted Multidestination Messages](#), CANPC '03
 - D. Buntinas, D. K. Panda and P. Sadayappan, [Fast NIC-Based Barrier over Myrinet/GM](#), IPDPS '01.
 - D. Buntinas, D.K. Panda, and W. Gropp, [NIC-Based Atomic Operations on Myrinet/GM](#), SAN-1
 - D. Buntinas and D. K. Panda, [NIC-Based Reduction in Myrinet Clusters: Is It Beneficial?](#), SAN-2
 - W. Yu, D. Buntinas, and D. K. Panda, [High Performance and Reliable NIC-Based Multicast over Myrinet/GM-2](#), ICPP '03

Myrinet

- Application-Bypass Collectives
 - D. Buntinas, D. K. Panda, and R. Brightwell, Application-Bypass Broadcast in MPICH over GM, CCGrid '03.
 - A. Wagner, D. Buntinas, R. Brightwell, and D. K. Panda, Application-Bypass Reduction for Large-Scale Clusters, Cluster 2003
- Efficient Support to Programming Models
 - D. Buntinas, A. Saify, D. K. Panda, and J. Nieplocha, Optimizing Barrier and Lock Operations in ARMCI CAC '03
 - R. Noronha and D. K. Panda, Implementing TreadMarks over GM on Myrinet: Challenges, Design Experience, and Performance Evaluation, CAC '03
 - V. Tipparaju, M. Krishnan, J. Nieplocha, G. Santhanaraman, and D. K. Panda, Optimizing Mechanisms for Latency Tolerance in Remote Memory Access Communication, Cluster 2003

Quadrics

- Active Network Interface Support
 - A. Moody, J. Fernandez, F. Petrini, and D. K. Panda, Scalable NIC-based Reduction on Large-scale Clusters, (SC '03)
- Efficient Support to Programming Models
 - W. Yu, S. Sur, D. K. Panda, R. T. Aulwes, and R. Graham, High Performance Broadcast Support in LA-MPI over Quadrics, LACSI '03