

Future Trends in High-Performance Interconnects for Parallel Computers?

System builder's view of the future...

Mark Seager
Asst. Dept. Head for Advanced Technology
Integrated Computing and Communications
Lawrence Livermore National Laboratory

Presented to Hot Interconnects Panel
Stanford University
August 25, 2004

UCRL-PRES-

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.



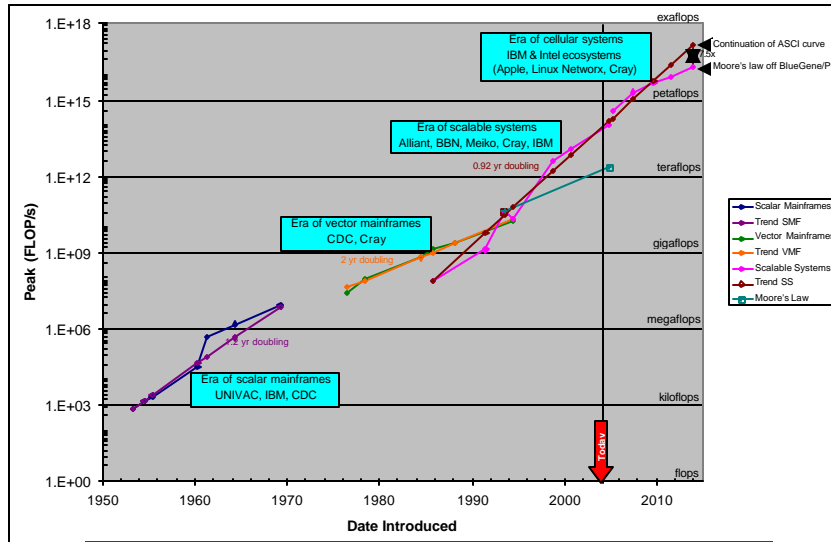
Major Points



- ◆ Future Trends in High-Performance Interconnects for Parallel Computing are:
 - Scalability of Systems
 - Much larger networks
 - Scalability of Applications
 - Much more intelligence in adapter
 - Convergence of Networks
 - Multiple functions for network



Looking into the future we see platforms that will be 2PF/s in 2008 and 20-140 PF/s in 10 years



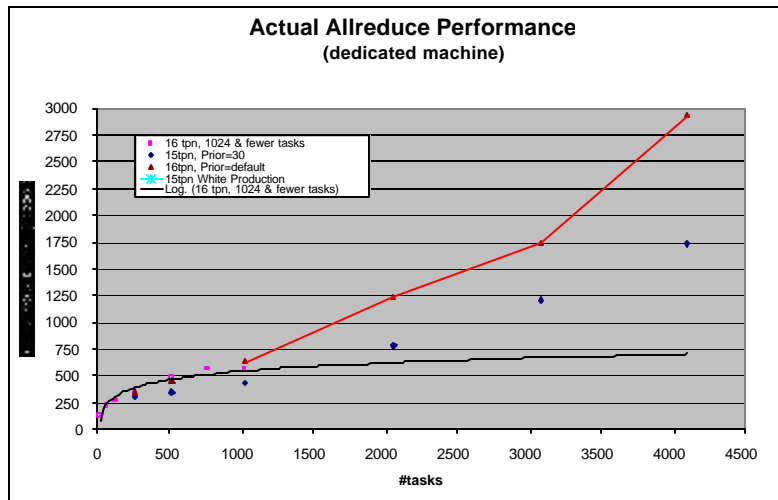
The 10 year forecast calls for cellular systems with between 2^{16} (65,536) and 2^{19} (524,288) CPUs. Let's get scalable!

Aug 24, 2004

Hot Interconnects Panel — page 3



Large scale applications highly dependent on scalable MPI operations

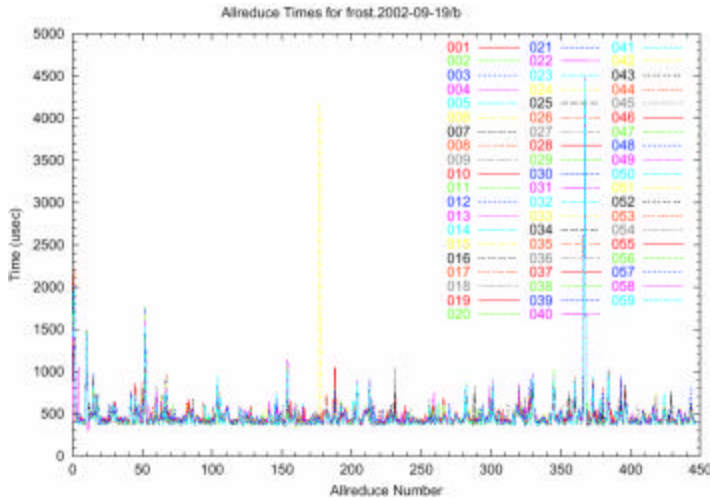


Aug 24, 2004

Hot Interconnects Panel — page 4



Cause is small, random perturbations in runtime due to interference with system activities

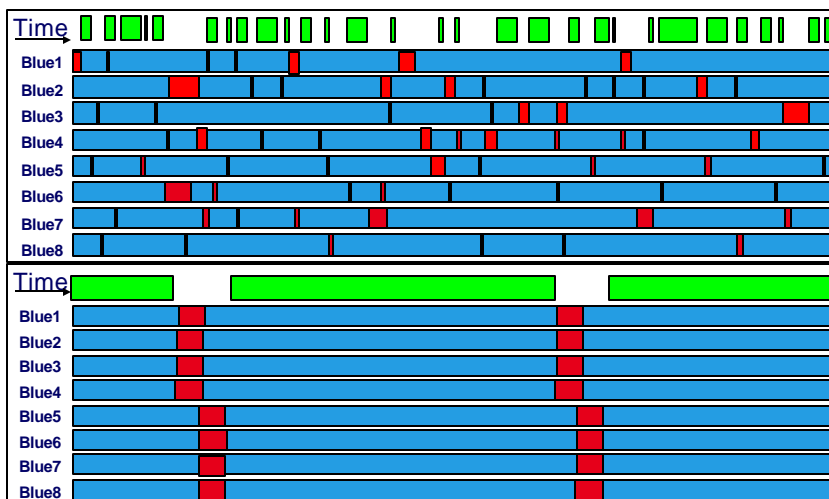


Aug 24, 2004

Hot Interconnects Panel — page 5



One possible solution is to synchronize OS schedules



Another solution is to simplify node OS and runtime to eliminate noise

Aug 24, 2004

Hot Interconnects Panel — page 6



Best solution for general purpose cluster computing is to put key MPI functions into adapter



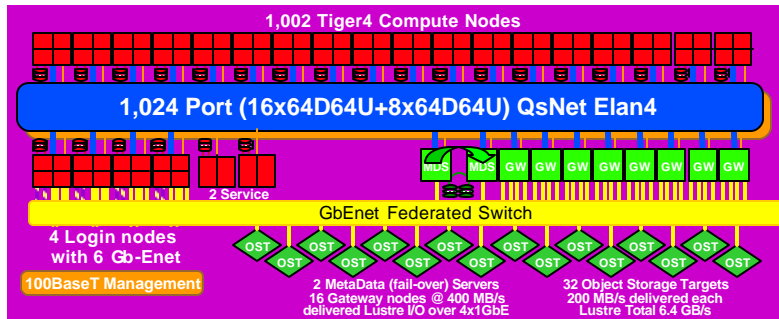
- ◆ MPI_ALLREDUCE
 - Min, Max – FP can be done as 64b integer operation
 - FP sum of small vector needs 64b FP ADD
 - software emulation good enough?
- ◆ MPI_BROADCAST
- ◆ MPI_BARRIER
- ◆ MPI_ALLGATHER, MPI_ALLSCATER
 - Launch and wait
- ◆ Fast implementation must allow for application reproducibility → reproducible ordering of FP operations

Aug 24, 2004

Hot Interconnects Panel — page 7



Larger Linux clusters require network convergence!



Four System Networks Does Not Scale!

- Quadrics ELAN4 for MPI
- 1000 BaseT Ethernet for Lustre
- 100 BaseT Ethernet for Management
- Serial console → 100 BaseT Ethernet

System Parameters

- Quad 1.4 GHz Itanium2 Madison Tiger4 nodes with 8.0 GB DDR266 SDRAM
- <3 μs, 900 MB/s MPI latency and Bandwidth over QsNet Elan4
- Support 400 MB/s transfers to Archive over quad Jumbo Frame Gb-Enet and QSW links from each Login node

Thunder achieved 19.94 teraFLOP/s (87% of peak and 2 on TOP500) on April 3, 2004

Partners

- LLNL: system architecture, CHAOS distro, system admin
- CDC: integration and delivery
- Intel: nodes & processors
- Qudaries: ELAN4
- CFS: Lustre
- DDN: SATA RAID

Aug 24, 2004

Hot Interconnects Panel — page 8



QsNet Elan4 modified fat-tree interconnect for a 1,024 node, 23 TF/s cluster



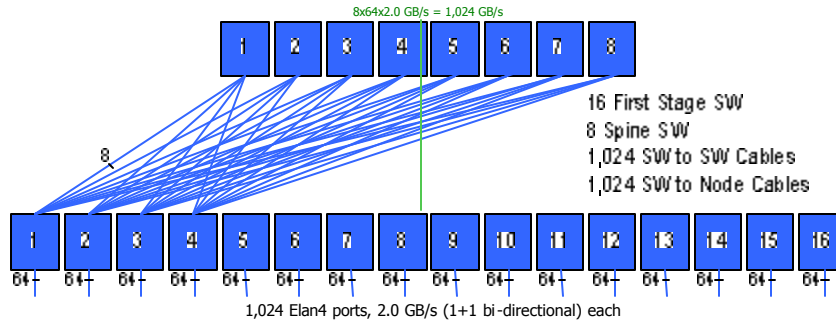
$F = 1,024 \times 22.4 \text{ GF/s} = 22.94 \text{ TF/s}$

Node B:F = 2.0 GB/s / 22.94 GF/s = 0.087

System B:F = 1.0 TB/s / 22.94 TF/s = 0.0446

Node : System = 2.0

Requires 24 QsNet Elan4 128-way switches & 2,000 Cables



This scales up to 64x64 or 4,096 ports

Aug 24, 2004

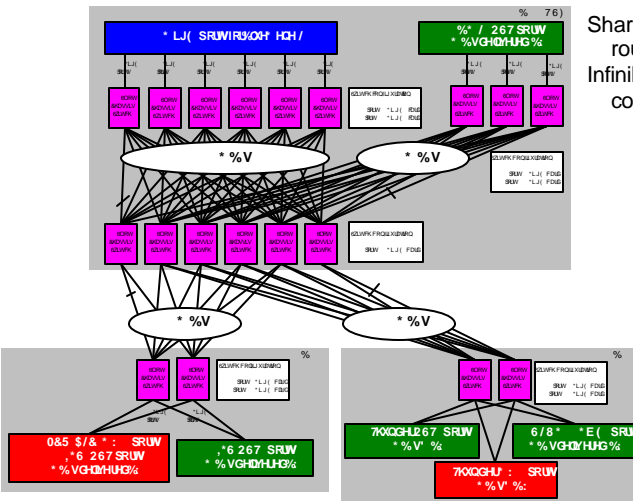
Hot Interconnects Panel — page 9



Federated 1 Gb Ethernet switching infrastructure requires distributed network between multiple building



2 &) HGUDMG (WHCHZLAK
9ULMQ S8LO



Shared file system require shared routable networking.
Infiniband is not going to solve the complexity problem.

Aug 24, 2004

Hot Interconnects Panel — page 10