COMPUTER & COMPUTATIONAL SCIENCES

Battle of the Network Stars Panel
Supercomputing 2003

Fabrizio Petrini

fabrizio@lanl.gov

Computer and Computational Sciences Division
Los Alamos National Laboratory

NNSA

1943 - 2003 Los Alamos NATIONAL LABORATORY

Ideas That Change the World

Why is "your" solution the better one?

- There is no “better” solution in absolute terms
- It simply depends on what you are doing
- Quadrics QsNet: outstanding network design, fast, easy to program: it provides what you need, but **nothing else.**
- Target: high-end performance computing (top20) with standard I/O interfaces (PCI, PCI-X, PCI-Express)
- By far, the best “research tool”

Quadrics QsNETII: main features

- Low latency High Bandwidth: 2-3μs latency, 900 MB/s bandwidth (sensitive to the I/O bus performance)
- Scalability to thousands of nodes
- 64 bit virtual addressing
- Reliable transmission protocol
- Commodity interface (PCI-X)
- Network interface is highly programmable at user-level

More information on QsNETII

- Available at the following URL

http://www.c3.lanl.gov/~fabrizio/quadrics.html

Latency and Bandwidth enough to characterize a network?

CCS-3

- No, they are not enough
- Trend: latency and bandwidth tend to be similar for all commodity network
- Latency limited by the speed of the light, we are only a factor of 5 from that (MPI latency delivered by Elan4 is 1.4 micros + PCI-X latency, practical lower bound 300 ns)
- Bandwidth is “technologically free”: expected two orders of magnitude improvement by the end of the decade, > 100GB/sec
- System software is the key
- Reliability, possibly in HW
- Support for global coordination/collective primitives in HW: large clusters will be more common
- Scalability of a larger machine, rather that point-to-point

Will the “status quo” continue?

CCS-3

- Why not? For sure in the next 5 years
- High end of commodity: QsNet, IBA, Myrinet, will be very competitive
- SANs with GigE (which is more expensive of the previous 3)

With 10-Gigabit Ethernet processors on the horizon, will RDMA/TCP/10GigE be sufficient in matching the performance of Quadrics, InfiniBand,Myrinet

- The point to point performance is only one part of the story, in particular in a large cluster
- It will be close, but always lagging behind.
- Heavyweight HW and system SW
- Performance should be measured at MPI level,
- Raw performance misleading
- Where is scalable system SW?

Underlying architecture?

- PCI-X now,
- PCI Express in a year or two
- HyperT has its own niche

Moore's law and supernetworking

CCS-3

- Bandwidth is free from a technological point of view
- Expected increase of a factor of 500 by the end of the decade

Inefficient MPI??

- Not sure it is the right question: it is too efficient/too low level
- > 2 billions of dollars in MPI SW in the TriLabs
- Growing at > 250 million a year
- How much is your network?

Infiniband

CCS-3

- Promising technology
- My bet: ready for prime time in 2005
- Will it replace QSW and Myri? If the DARPA HPCS projects are successful, they will all be out of business (at the least for high end supercomputers)