



AMES LABORATORY

# Cluster Interconnect Overview

**Brett M. Bode**

**Scalable Computing Laboratory**

[www.scl.ameslab.gov](http://www.scl.ameslab.gov)

# Motivation



AMES LABORATORY

In the past 10 years the speed of the interconnects commonly found in clusters has improved by a factor of 1000!

While 1-3 years ago the PCs commonly available were unable to make full use of the available bandwidth, today's systems are demonstrating some impressive performance.

There are more good choices for interconnects today than ever.

# The Past



AMES LABORATORY

- 10 Mbps Ethernet
- ATM
- Fast Ethernet
- Others

# ALICE - 1997



AMES LABORATORY

- 64 dual 200 MHz Pentium Pro CPUs
  - Fast Ethernet
  - 256MB RAM per node
  - 2GB disk per node
  - 8 additional nodes dedicated to development
- Large Fast Ethernet based switch
- Theoretically capable of 100Mbits/sec (12 Mbytes/sec) full duplex between any pair of nodes (most systems can utilize 90% of that peak)
- Switch has a 55 Gbit/sec backplane and can take modules with Fast Ethernet, Gigabit Ethernet, or several other technologies.
- Servers connected via Gigabit Ethernet



# The Present



AMES LABORATORY

- Gigabit Ethernet
- Myrinet
- SCI - Dolphin Wulfskit
- Quadrics
- Infiniband
- 10 Gigabit Ethernet



# Do we need faster networks?

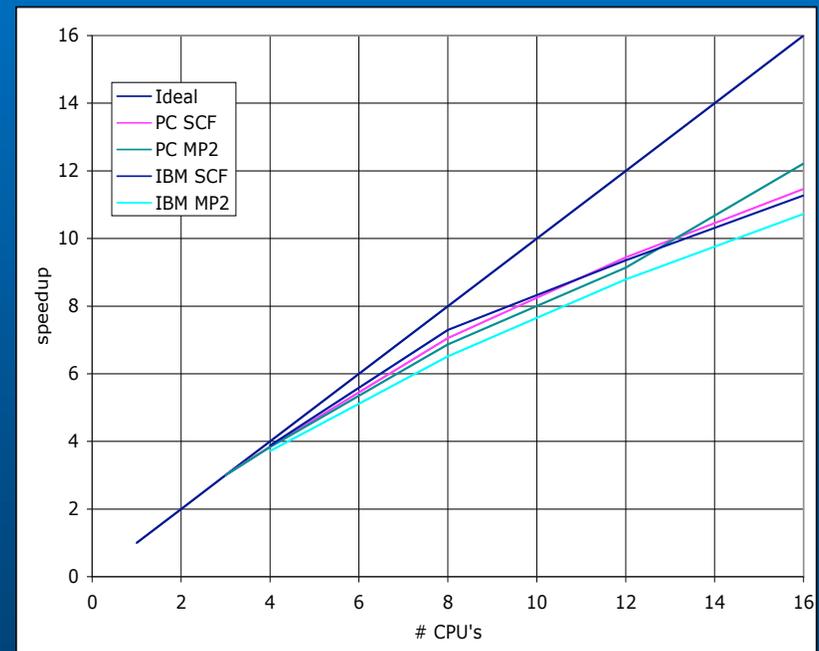


AMES LABORATORY

In general the compute/communication ratio in a cluster needs to remain fairly constant.

So as the computational power increases the network speed must also be increased.

This graph illustrates that balance working. It compares single CPU PII nodes with Fast Ethernet, versus dual CPU IBM workstations using Gigabit. (roughly factor of 6 in CPU, 10 in network)

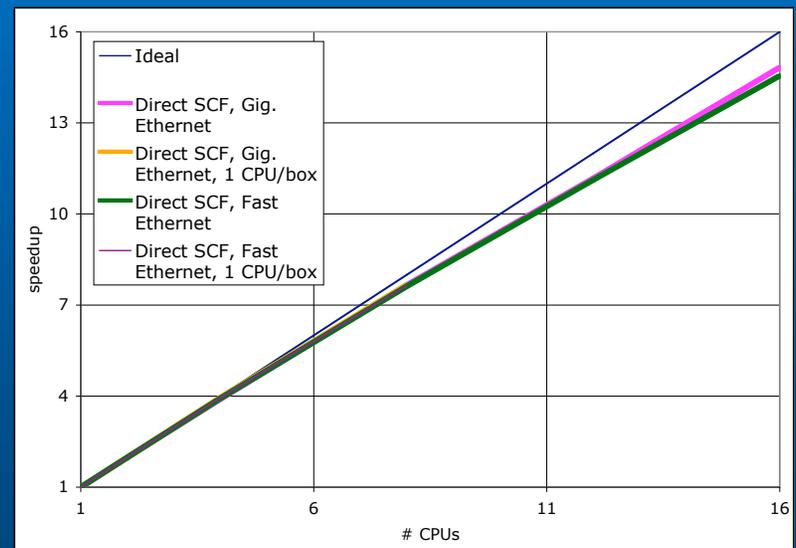


# Lightweight Communications



AMES LABORATORY

- Some applications, like the GAMESS SCF test in this example have very low communications requirements and scale very well on almost any interconnect.
- The test compares running a calculation over Fast Ethernet and Gigabit Ethernet and with 1 or 2 CPUs active per node.
- Here the network makes virtually no difference!

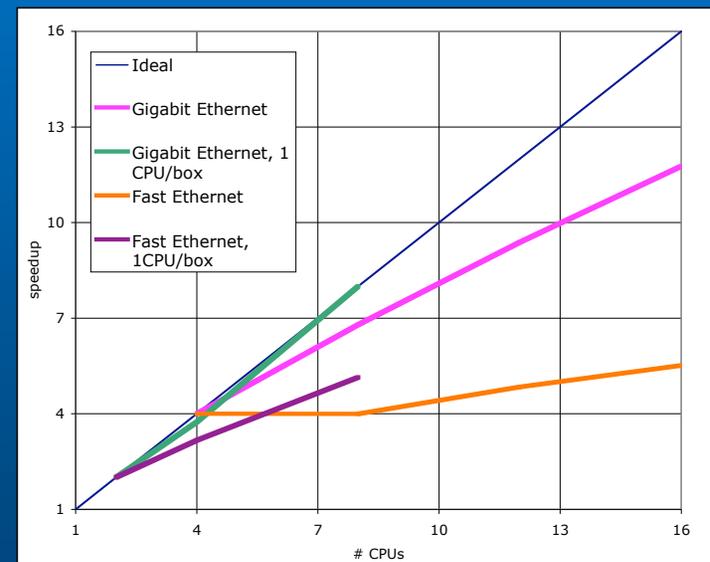


# Communication Intensive Applications



AMES LABORATORY

- In this test the GAMESS parallel MP2 code is used. It uses a pseudo global shared memory approach leading to a high communication requirement.
- Here Fast Ethernet is NOT good enough.
- Gigabit Ethernet works better, but is still not fully adequate for two CPUs.



# Ethernet



AMES LABORATORY

Ethernet has almost always been the defacto baseline network on clusters. Most clusters use Ethernet for at least administrative traffic, and any use it for the primary high-speed interconnect.

The principle advantage of Ethernet is its ubiquity. This leads to low prices and widespread driver availability.

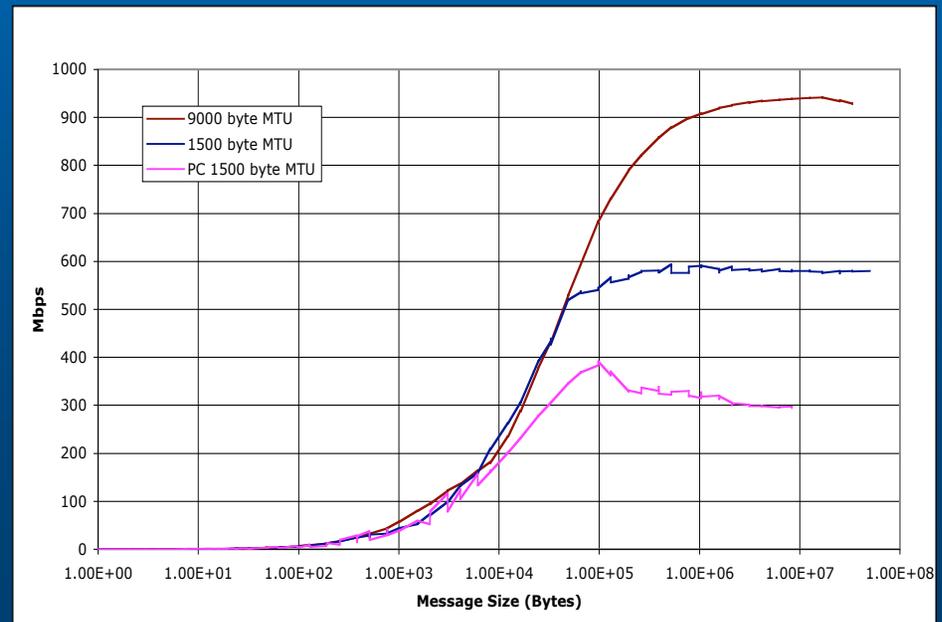
The principle disadvantage is that its switched architecture relies on intelligent switches that can be expensive to produce. These switches are designed for commodity networks leading to unneeded features (ex. Layer 3 and higher routing) and are difficult to aggregate together.

# Gigabit Ethernet

10

AMES LABORATORY

- Early Gigabit Ethernet boards obtained fairly meager performance (often 300-400 Mbps or below), at 10 times the cost of Fast Ethernet.
- One solution was to extend the Ethernet specification for MTU size to 9000 bytes.
  - Doubles the performance in some cases (up to 940Mbps)
  - Limits the choice of NICs and switches (Alteon, 3Com, SysKonnnect, Extreme, Force10...)
  - Have tended to be more expensive (\$300-\$800/NIC, \$500 or more / switch port)

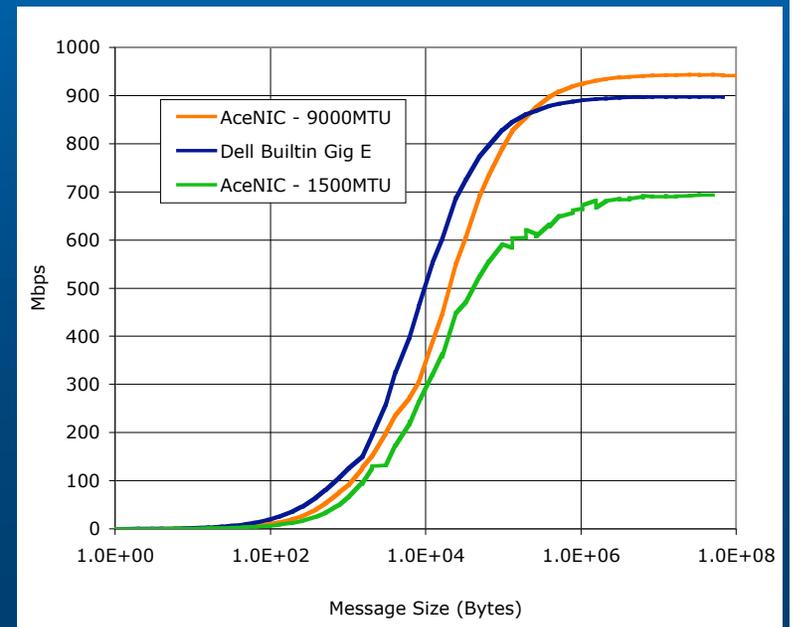


# Gigabit Ethernet



AMES LABORATORY

- A new generation of inexpensive (or even integrated) NICs offers very good performance for very low cost (< \$100/NIC). Often these use the Broadcom chipset.
- Similarly 4-24 port 1000BaseT switches are available at <\$100/port.
- Driven in part by the need to provide 10Gb Ethernet switches large chassis Gigabit switches have dropped in price dramatically in the last couple of years. Now mid-range switches run \$200-\$300/port and the largest switches run around \$667/port for up to 480 ports.



# Gigabit Ethernet



AMES LABORATORY

Thus Gigabit Ethernet provides several advantages:

- Works on almost any platform
- Wide choice of vendors
- Widest application portability
- Lowest cost solution:
  - \$100-\$200/port up to 64 ports
  - \$250-\$350/port up to 128 ports
  - under \$750/port up to 480 ports



# Gigabit Ethernet



AMES LABORATORY

Gigabit Ethernet also has several disadvantages:

- Lowest bandwidth and highest latency in this survey.
- Difficult/impossible to maintain full bi-section bandwidth with more than one switch

# Myrinet



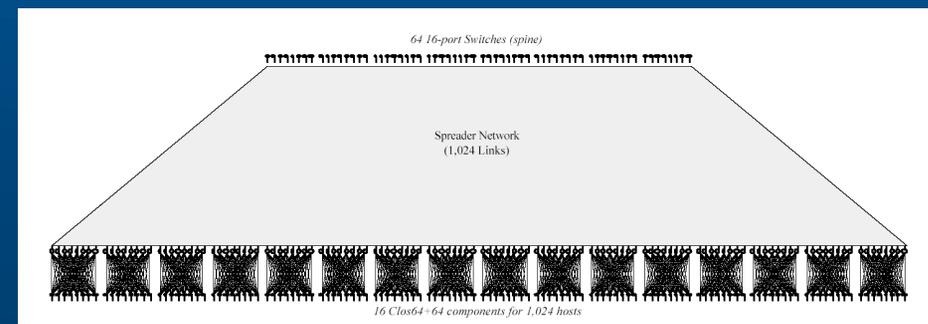
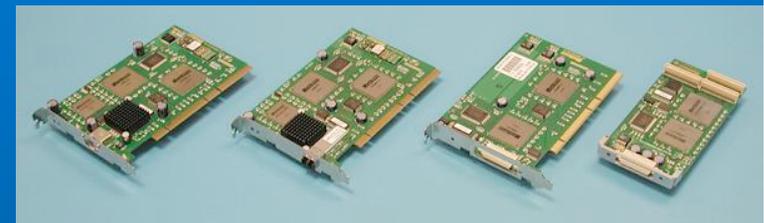
AMES LABORATORY

Myrinet was one of the first interconnects designed specifically for use in a cluster.

As such the design allows for good scalability and moderate cost.

Myrinet is a source routed, switched network with the switching elements laid out in a fat tree.

This leads to relatively inexpensive switches.

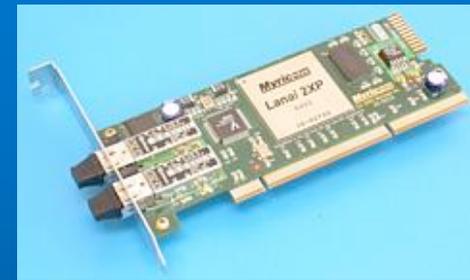


# Myrinet

15

AMES LABORATORY

- Early versions of Myrinet using copper cabling had a few hardware (mostly cabling) and software issues that affected stability
- About three years ago new software and the change to all fiber interfaces resulted in a dramatic improvement in reliability.
- Runs on a wide range of platforms.
- Scales to a large number of nodes via switch aggregation in Fat Tree arrangement.



# Myrinet



AMES LABORATORY

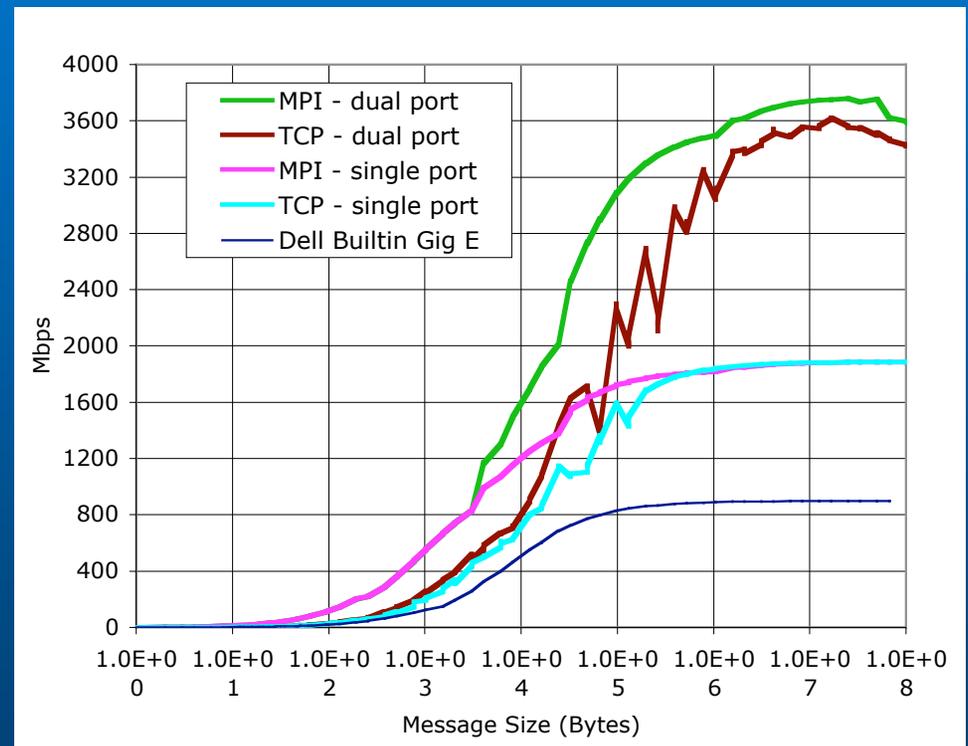
Performance is quite good with the limit of the 2Gbps carrier.

Latency is low 6-7  $\mu$ sec.

Peak bandwidth 1880+Mbps.

Performance in IP mode is also quite good, 1880+Mbps and 27-30  $\mu$ sec latency.

Dual port NICs provide almost double the performance of the single port NIC: 3750Mbps for MPI and 3500Mbps with TCP/IP



# Myrinet



AMES LABORATORY

The cost of Myrinet has dropped steadily over the past couple of years. NICs are down to \$595 for the single and \$995 for the dual.

This leads to:

~\$850/port up to 128 ports

Up to \$1737/port up to 1024 ports

The dual port NIC nearly doubles the cost adding at least \$800/Node.

Thus the single port solution is quite competitive, but the dual port solution is much less so...

# Myrinet - future



AMES LABORATORY

Myricom has a large installed base and relatively good products.

Higher link speeds will be needed!

The software stack still needs improvement.

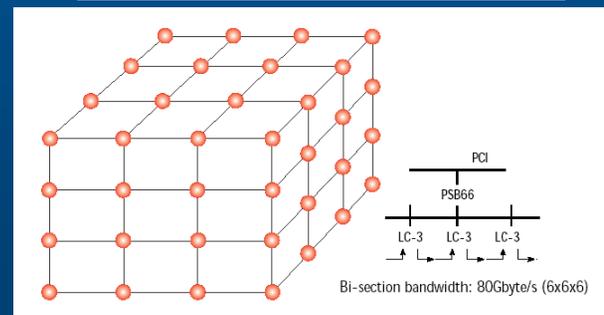
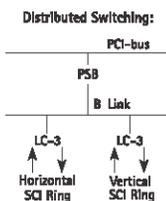
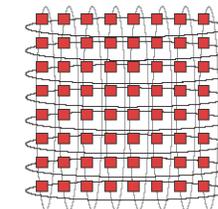
Myrinet needs better hardware and software diagnostics!

# SCI: Dolphin Wulfskit

19

AMES LABORATORY

- SCI offers 2 or three 500MByte per second external links
- Nodes are configured in a 2D or 3D wrapped mesh without switches.
- Pass through messages are handled entirely on the NIC at a very low additional latency.
- Dolphin provides an Open Source driver (including a TCP socket implementation). A 3rd party Open Source MPI is also available.
- Scali provides a commercial driver and MPI.

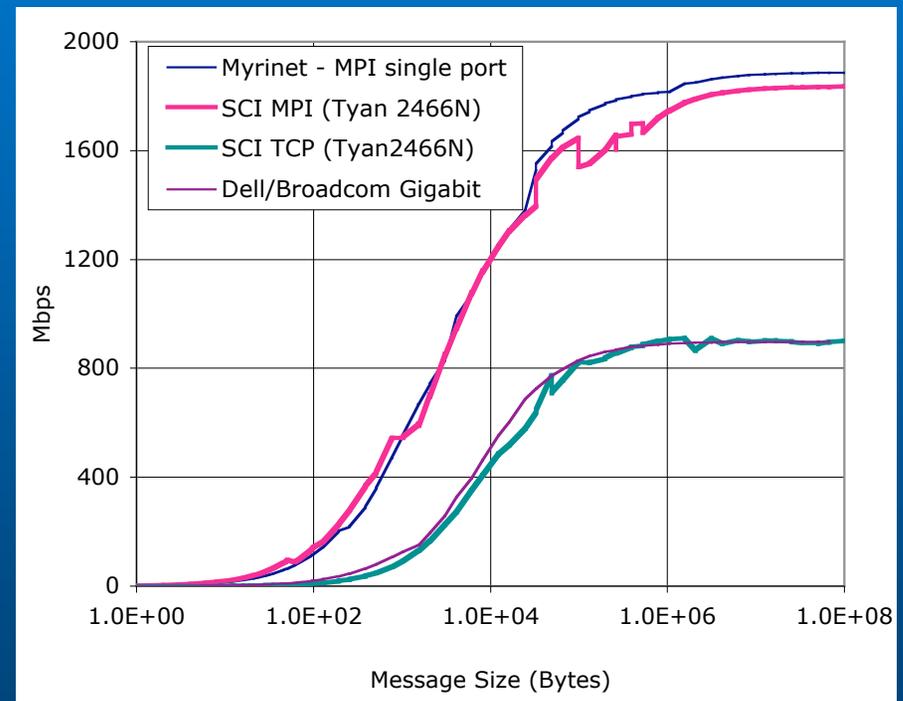


# SCI: Dolphin Wulfkit



AMES LABORATORY

- Performance is quite good with very low latency ( $\sim 4 \mu\text{sec}$ ) and high peak bandwidth (1800 Mbps).
- Performance in the 1KB-100KB range is particularly impressive.
- IP mode is much less impressive.
- The mesh topology is particularly well suited to applications that divide work up on a rectangular grid. If these apps are mapped to the physical grid layout the scalability would be very good.



# SCI: Dolphin Wulfskit



AMES LABORATORY

- Scalability is limited by aggregate bandwidth on the link, a mesh dimension of 8-10 is a reasonable limit.
- Cable lengths are very short making setup a challenge.
- Per port costs are fixed at ~\$1095 for 2D, ~\$1595 for 3D.
- 3rd party MPI and driver software from Scali is recommended, but is not Open Source and adds up to \$378/node.

# Quadrics - QsNetII



AMES LABORATORY

- The original QSNET has been used extensively on Compaq/HP AlphaServer SC systems.
- New QSNET II has much higher performance and lower cost.
- Now focused more on high end x86 and IA64 clusters.
- Switched architecture.
- Hardware is capable of 2-5  $\mu$ sec latencies
- Hardware link bandwidth is 900 MB/sec
- Configurations up to 4096 nodes with switches in a fat tree arrangement.



# Quadrics Software



AMES LABORATORY

## Good Points

Most of the software stack is Open Source (driver/MPI).

Good support for variety of Linux kernel versions and distributions.

Most of the communications overhead is offloaded to the NIC.

The NIC/driver supports DMA directly to paged VM! Thus no memory locking is needed.

## Not so good points

A kernel patch is required (to support DMA to VM)

Software stack includes RMS which is a licensed product.

The TCP/IP mode did not build properly for us.

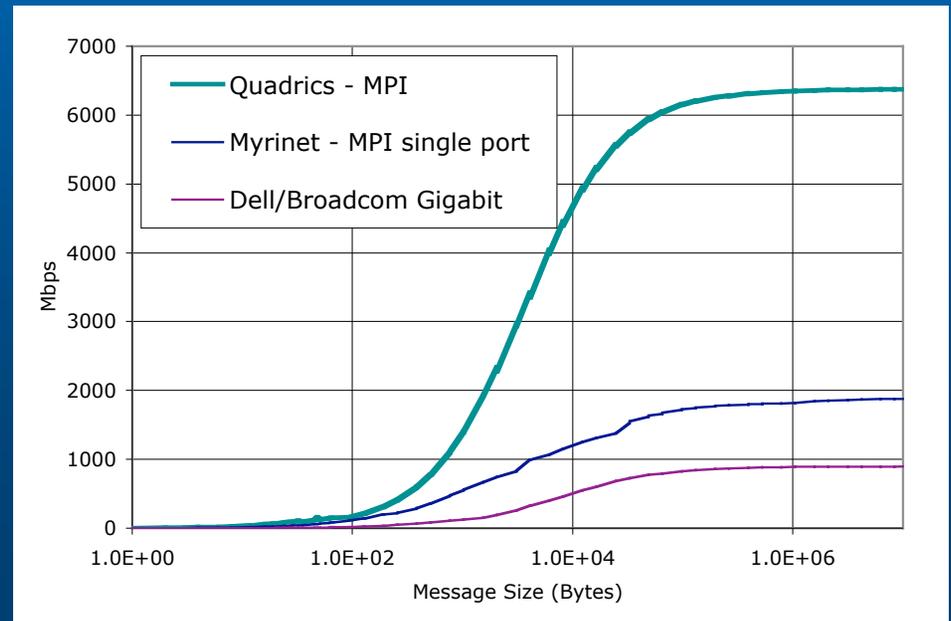
# Quadrics - QSNNet II



AMES LABORATORY

Achieved performance is excellent.

- 2-3  $\mu$ sec latencies
- Peak bandwidth of 6370 Mbps



# Quadrics performance

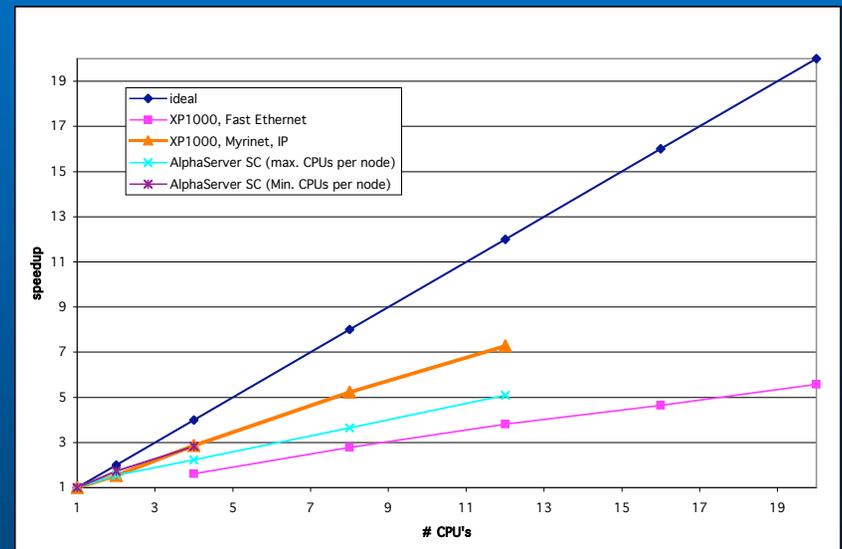


AMES LABORATORY

Application performance on the SC has been a bit mixed.

Overall scaling is quite good as the number of nodes is increased.

But a closer look at the scaling within a box reveals a significant problem (probably a shared lock).



# Cost



AMES LABORATORY

Quadrics has always been a premium product. This has come down somewhat, though it remains higher in cost than other solutions.

Costs scale with port count:

\$1700-1900/port 8-128 ports

\$3300-3600/port 256-4096 ports

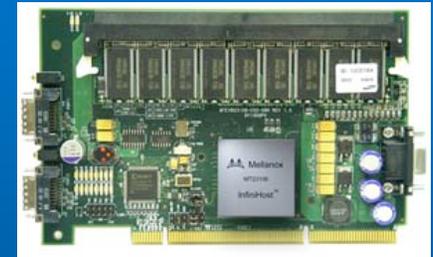
# Infiniband



AMES LABORATORY

Infiniband is a standard created by committee to solve all of our problems providing a solution for everything from a system bus to an external interconnect for storage and communications.

The basic link in Infiniband is 2.5Gbps, but the design includes multiples of the basic link. Currently most implementations use 4 links (aka 4x) providing 10Gbps. 12x links are also supported providing 30Gbps.



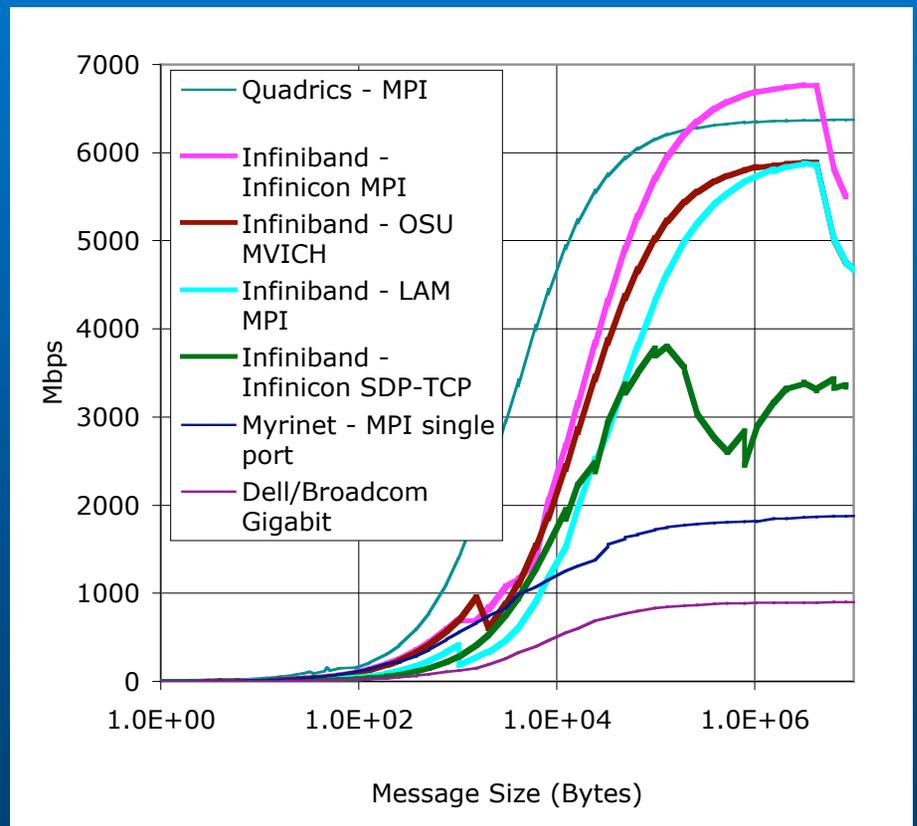
# Infiniband



Infiniband performance is quite encouraging.

Peak performance is in excess of 6750Mbps, though is quite dependant on various settings and the software implementation.

Current TCP/IP performance is limited to around 3500Mbps.



# Infiniband



AMES LABORATORY

Currently most of the Infiniband hardware is based on Silicon from Mellanox Corp.

However, there are quite a few vendors producing NICs and switches based on this hardware.

The software story is still very much in flux. We need a good Open Source stack!

Cost is quite reasonable ranging from \$1200-\$1600/port.

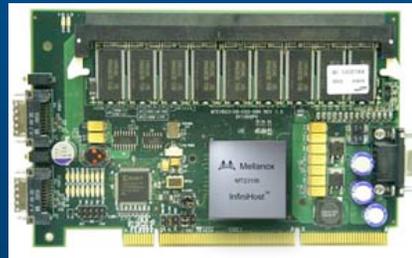
# The Future



AMES LABORATORY

- Infiniband

- Companies may be nearing a funding transition point.
- The software stack(s) need to improve!
- In order to really achieve its promise it must be tied in closer to the processor/memory bus (no PCI).
- IF Infiniband takes off as a standard storage, etc interconnect the price should plummet.



# 10 Gigabit Ethernet



AMES LABORATORY

10 Gb Ethernet has come a long way in the past two years!

Two companies offer NICs.

Many switches support 10Gb ports and the per port price is down to about \$10k per port.

However, switch density is fairly small (up to about 48 ports).

# 10 Gb Ethernet



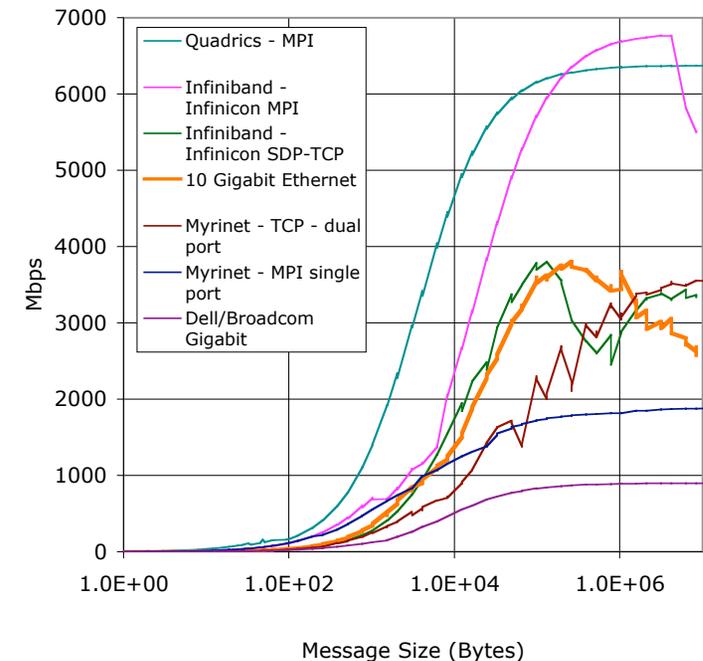
AMES LABORATORY

Performance remains a concern:

Current TCP/IP performance peaks around 3700 Mbps.

OS/bypass techniques are having some success at improving the peak bandwidth and latency.

This limit is probably related more to the overhead of TCP/IP than 10Gb Ethernet, indeed the limit is similar to that seen on other technologies.



# 10 Gb Ethernet



AMES LABORATORY

10Gb will be a player in the future, but for now it is limited to use for trunking between switches.

The price of the optics will have to drop a lot, or the 10Gb over copper standard will be needed to get the prices down to a reasonable level.

By the time this happens the CPU/bus performance will likely improve enough such that the overhead of TCP/IP will not be such a critical issue.

# GAMESS performance



AMES LABORATORY

<b>#node, 2Procs per</b>	<b>1</b>	<b>2</b>	<b>4</b>
Gigabit Ethernet	5026	1961	1529
Myrinet MPI	3277	1984	1559
Myrinet TCP	3477	1924	1279
Infiniband MPI	4868	2785	1319
Quadrics MPI		2593	1427
<b># node, 1 proc per</b>	<b>1</b>	<b>2</b>	<b>4</b>
Gigabit Ethernet	5823	3062	1852
Myrinet MPI	5651	3030	1743
Myrinet TCP	5843	3059	1801
Infiniband MPI	5743	3006	1727
Quadrics MPI		3052	1775

No big problems.

Why is Myrinet IP mode the fastest?

# GAMESS Performance



AMES LABORATORY

GAMESS uses a pseudo global shared memory model called DDI (similar to Global Arrays)

This leads to one compute and one data server on each processor.

Data servers spend most of their time waiting for messages.

This cause many MPI implementations to poll burning CPU time.

# The Future



- **PCI-Express**
  - **One of the most important near-term advances will be the arrival of PCI-Express.**
  - **Similar to Infiniband in that it provides a scalable bandwidth from a single 2.5Gpbs channel on up with options for 2, 4, 8 ,16 and 32 channel widths.**
  - **Most 10Gbps (or higher) interconnect vendors will likely release PCI- Express products quickly to get around the 8Gbps limit of PCI-X.**

# The Future



AMES LABORATORY

- **This will lead to new versions of existing products (ie Myrinet, SCI, etc) at 10Gbps. It will also allow better utilization of Infiniband and Quadrics.**
- **Enable even faster technologies such as 12x Infiniband.**

# Conclusions:

## So which one is the right one?



AMES LABORATORY

### Gigabit Ethernet

Obvious choice for small clusters due to its extremely low cost.

For latency tolerant applications it can scale easily up to 480 nodes.

On small systems bandwidth can be increased by using two NICs and two switches.

Only systems with very light communications needs should scale beyond 480 nodes since there will be bottlenecks in the network.

# Conclusions:

## So which one is the right one?



### Myrinet

Myrinet provides excellent performance, particularly at the IP level.

Myrinet offers a fairly competitive price when using the single port products.

Not clear the the dual-port offering provides more value than the competition.

# Conclusions: So which one is the right one?



AMES LABORATORY

## SCI

SCI offers very low latency and impressive small message performance.

It is showing its age though, a speed increase is needed to be competitive.

The open source software still needs more work.

# Conclusions:

## So which one is the right one?



AMES LABORATORY

### Quadrics

Quadrics offers leading edge performance and scales to large cluster sizes.

The price has dropped but remains above the competitors.

Still an excellent choice given the good software support and performance.

# Conclusions: So which one is the right one?



AMES LABORATORY

## Infiniband

Infiniband provides excellent price/performance.

Seems to scale reasonably well.

Still a bit unproven with some rough edges.

# Conclusions: So which one is the right one?



AMES LABORATORY

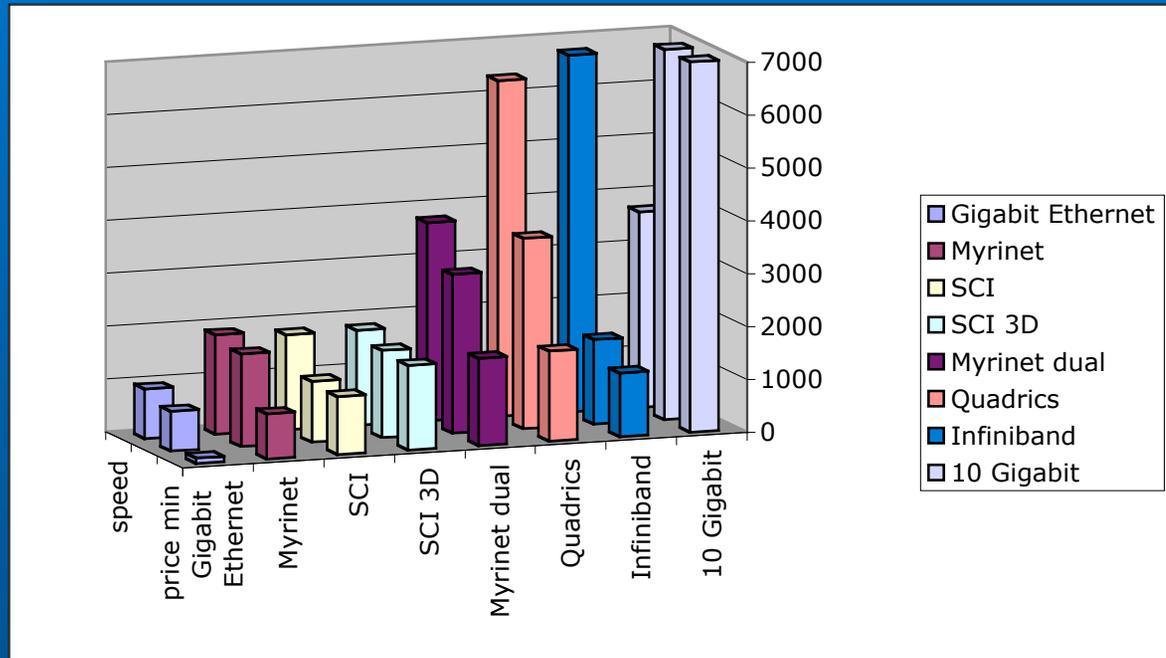
## 10 Gigabit Ethernet

Currently 10Gb is simply not competitive due to its high cost.

# Cost/performance



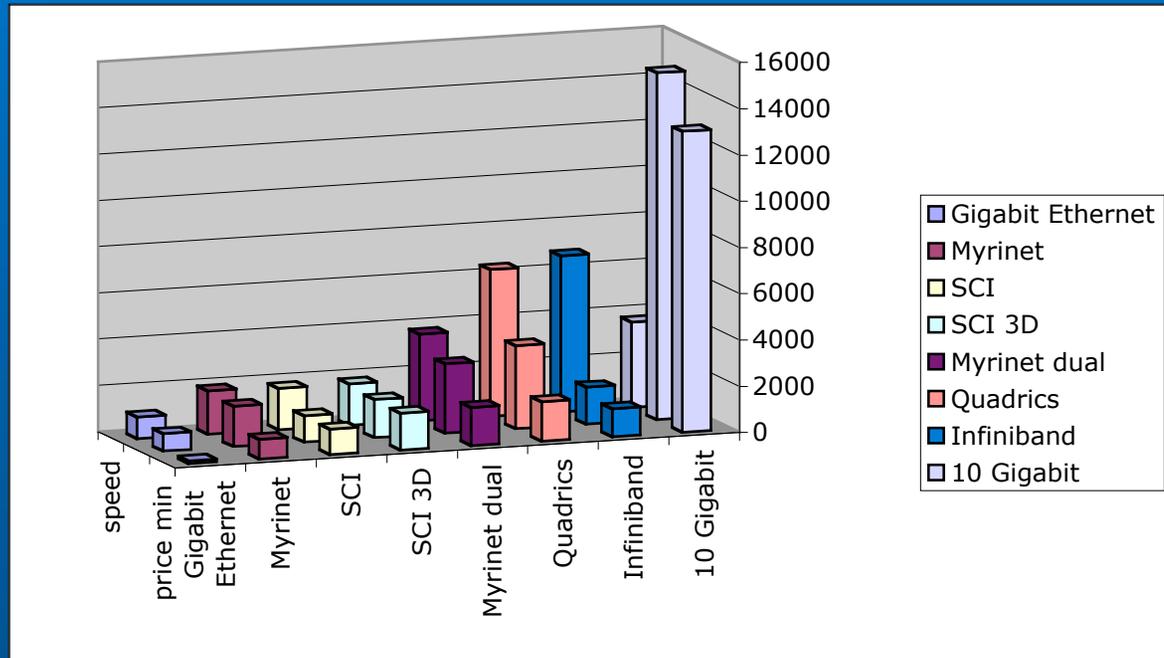
AMES LABORATORY



# Cost/performance



AMES LABORATORY



# Thank you



AMES LABORATORY

- **\$\$\$'s**
  - U.S. Department of Energy
  - NSF
  - IBM
- **People**
  - Jason Hill
  - David Schwenker

# Questions?



AMES LABORATORY

