ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

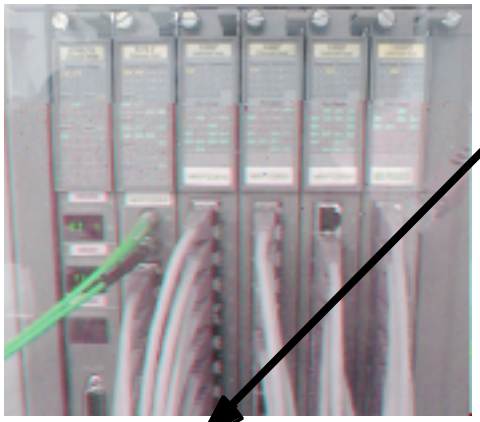
Reflections on Interconnects in PC Clusters

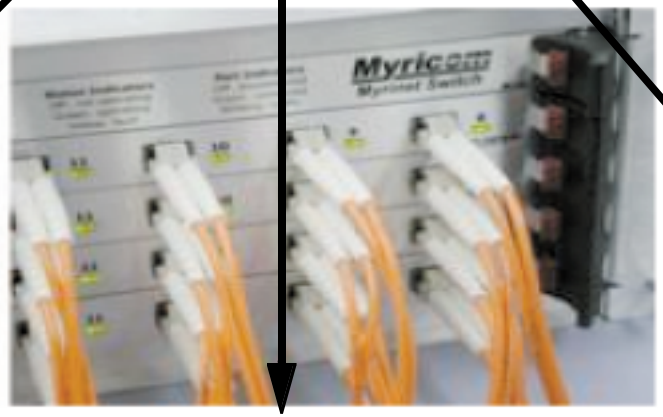
Thomas M. Stricker
Laboratory for Computer Systems
ETHZ - Swiss Institute of Technology
CH-8092 Zürich, Switzerland

Cluster 2003 Panel Discussion
December 3 - Hong-Kong, SAR, China

Classes of Cluster Interconnects

(according study in CAC03)





Beowulfs

$200 per node

UTP Ethernet

inexpensive

Switches

PC Clusters

$1000 per node

Myrinet, SCI

(Switches/Rings)

Infiniband

(Switches)

µP based MPP

$5000 per node

Cray T3E++

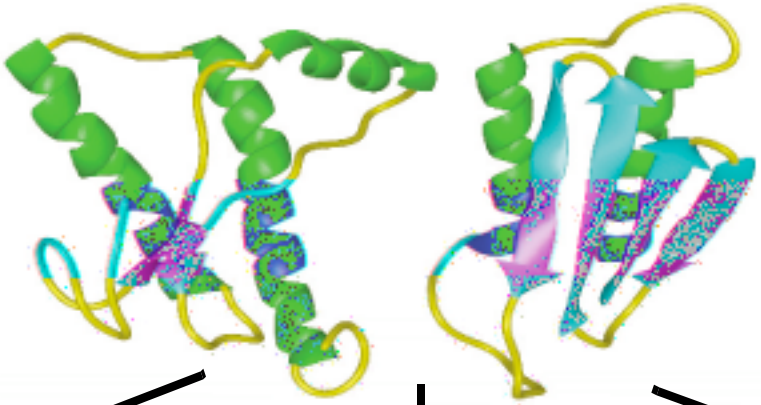
(3D,4D meshes)

Quadrics

(Fat Trees)

Effect on Applications

(e.g. Molecular Dynamics)



Beowulfs

simple, straight forward force calculation in 3D domain

inaccurate

PC Clusters

improved force calculations i.e. part.mesh Ewald in FFT domain

more accurate

µP based MPP

?

even better results for sci.

What would be a cluster with a very good interconnect today?

1993 Cray T3D	10 years	2003 Cluster MPP
Alpha CPU 21064 (150 MHz)	13.5x	AMD Opteron (2GHz)
64MB RAM per node	32x	2GB RAM per node
estim. 28 SPEC2000fp (based on a SPEC95 result)	50x	1400 SPECfp2000
Network Block transfers 1000Mb/s=125MB/s	25x	Network Block transfers 25Gb/s=3.125MB/s
Remote loads/stores (64b wrd) 300Mb/s≈35MB/s	25x	Remote loads/stores 7.5Gb/s≈933MB/s
Message Latency 1.5µs	1/25x	Message Latency 60ns
Barrier Synchronization 3.0µs @ 512 nodes	1/25x	Barrier Synchronization 120ns @ 512 nodes

How could we make that happen ? (Technology)

Bandwidth for contiguous Block Transfers

- **not unrealistic**... Infiniband 4x,8x will get there, Myrinet too!

Bandwidth for Remote Stores/Pipelined Remote Loads

- **unrealistic...** with any PCI, PCI-X I/O bus in the way.
- **realistic** with direct incorporation of communication operations into CPU and instruction set.

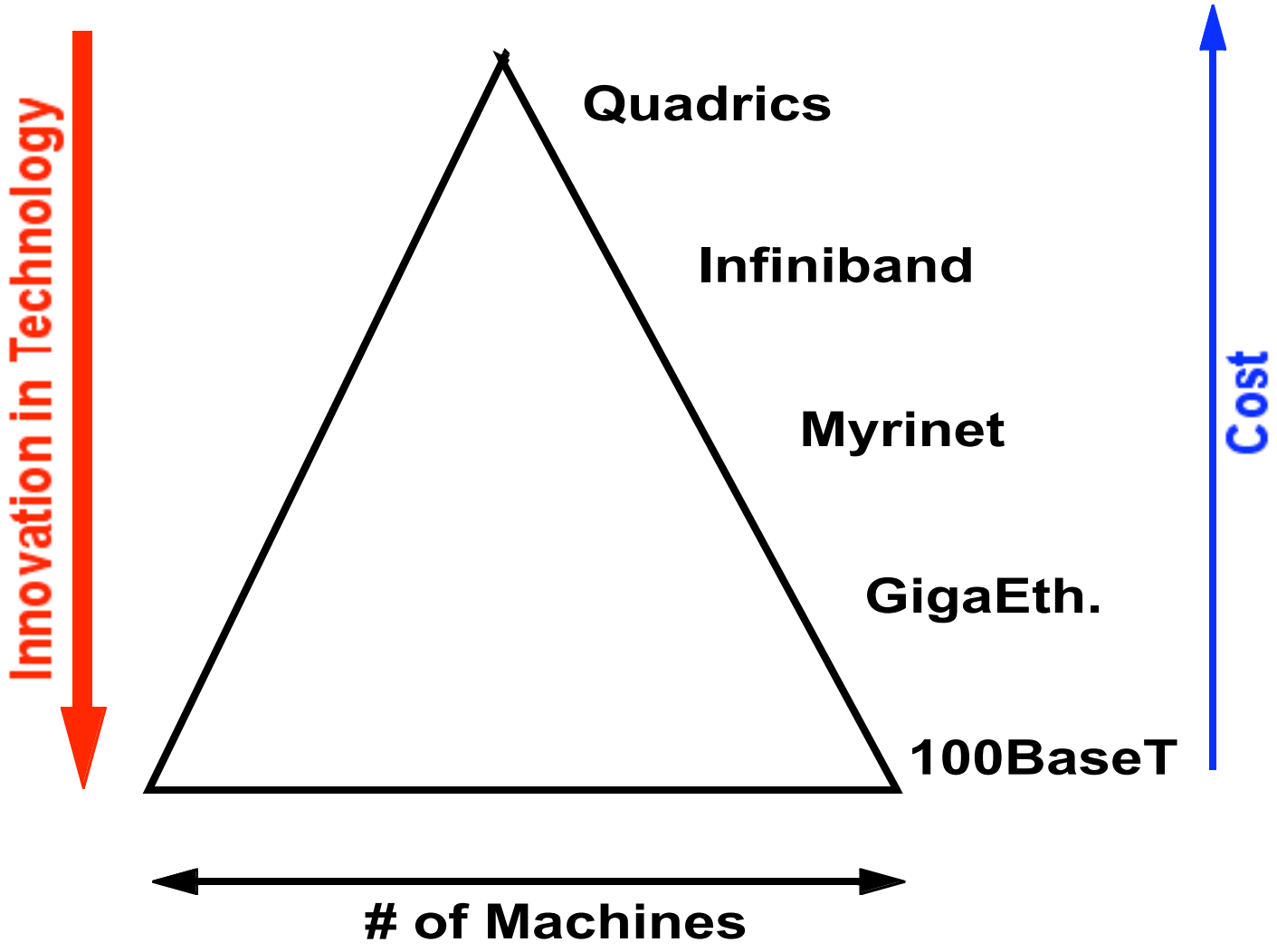
Latency for a Single Message Transfer

- **unrealistic**, electricity/light makes only 0.15-0.2 m/ns, there is maybe room for an **5x improvement** with perfect switches!

Latency for a Barrier Synchronization of a full machine

- there is room for a **maybe 10x improvement** with dedicated sync. network that switches faster.

Affordability of High Speed Networks



How could we make that happen ? (Economics)

Problem of the purchasing decision for clusters

- Mostly a deal between a computational science researcher and a funding agency... none of which have some in depth knowledge in cluster architectures.
- Unlike in building construction - architects are rarely consulted.

The cluster business lacks the application specialists that supercomputer manufacturers provided 10 yrs. ago.

- Often there is little understanding what a better network could do to improve computational science in an application field.
- Often all a researcher can get is the bare bone Beowulf with minimal networking.

Conclusions

Microprocessor technology needs to **integrate computation** and **communication**.

Interaction between **cluster architects** and **applications specialists** is needed to create a **market for better cluster interconnects**.

A sole, narrow focus of PC clusters to **best cost/performance** remains **questionable in the long run**, because:

- the definition of measurable performance is often unrelated to an appropriate definition of utility within an application field.
- after 50 years of computing **it is still not clear** if better computers **enable** new computational applications or if new applications **lead to** the construction of better computers.