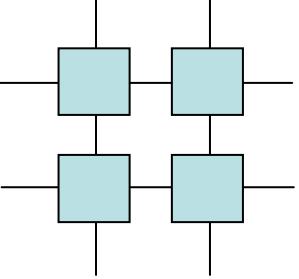
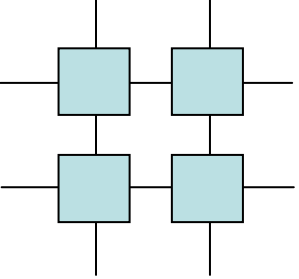


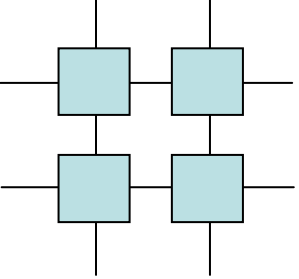
Cluster Computing

Cluster Architectures



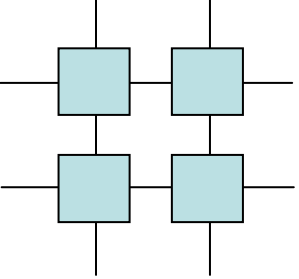
Overview

- The Problem
- The Solution
- The Anatomy of a Cluster
- The New Problem
- A big cluster example



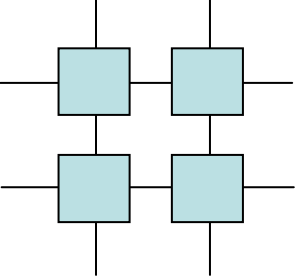
The Problem Applications

- Many fields have come to depend on processing power for progress:
  - Medicine / Biochemistry (molecular level simulations)
  - Weather forecasting (ocean current simulation)
  - Engineering problems (car crash simulation etc.)
  - Genetics Research (human genome project)
  - Physics (Quantum simulations)



The Hardware Problem

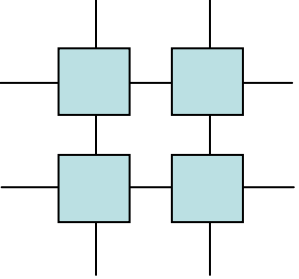
- The previous problems can only be handled by supercomputers
- Supercomputers are expensive, even when measuring $/Mflops
- Supercomputers are complex to build
- Few Supercomputers are build, which in turn makes them more expensive



Cluster Computing

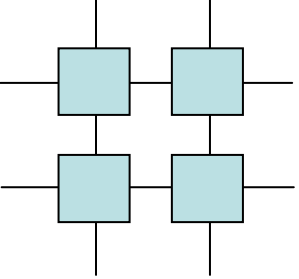
The Alternative

- Workstations are cheap, also when measuring $/Mflops
- Workstations are easy to build and readily available
- Workstations are sold in the millions, which makes them even cheaper
- Workstations are too slow



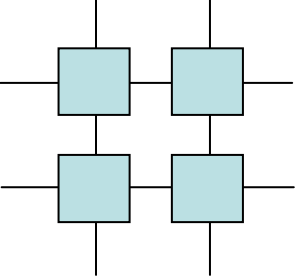
The Solution

- Workstations may be interconnected to function as a supercomputer
  - Cheap
  - In theory a set of workstations are powerful, e.g. N workstations may solve a problem in 1/N time
  - In practice things are not so simple



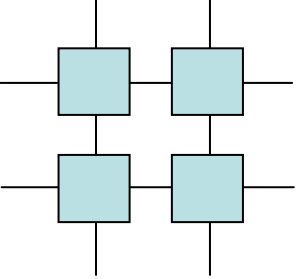
The Anatomy of a Cluster

- The field is new enough that there is not consensus on what a cluster is, check the debate on: http://www.eg.bucknell.edu/~hyde/tfcc/vol1no1-dialog.html
- On the abstract plane a cluster is a set of interconnected computers



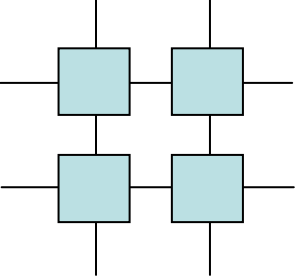
The Parallelization Problem

- If one man can dig a 10 by one by one ditch in ten hours, then two men can do so in five hours
  - Can 10 men dig the ditch in one hour?
  - What about a one by one by 10 hole?



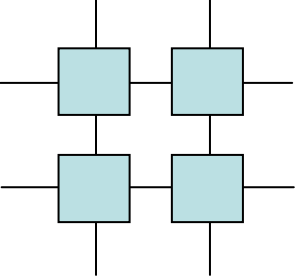
Programming the Cluster

- Even if we can parallelize the problem, how can we execute it on a cluster?
  - Using message exchange
  - Pretending we have shared memory



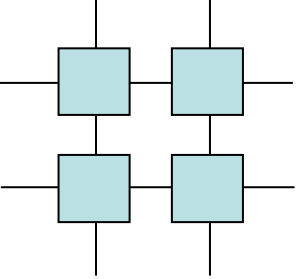
The New Problems

- An Cray X1 has a message latency of less than 2 microseconds, 1Gb/sec TCP is well over 65 microseconds
- Commercial supercomputers comes with optimized libraries - cluster architectures has none
  - Well – this is slowly changing



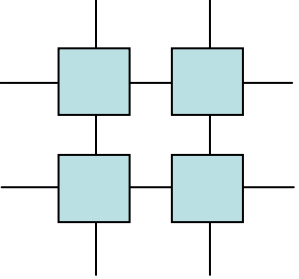
(what used to be) Denmark's fastest Supercomputer

Background, Architecture and Use



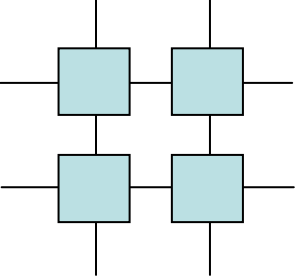
Next generation supercomputers

- Clusters of PC’s
- Emulating
  - SMP or
  - MPP machines
- Connected through standard Ethernet or custom cluster-interconnects



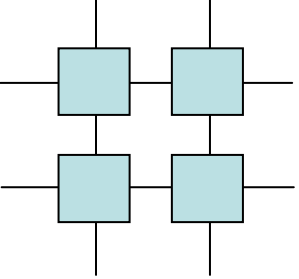
The advantages of cluster computers

- Commercial Of The Shelf (COTS)
- Drip model
  Supercomputer
  - ⇒ Workstation
  - ⇒ PC
- Easily adjusts to user needs



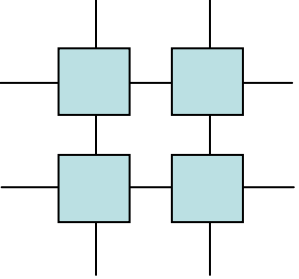
Cluster Machines

- ✚ Extremely cheap
- ✚ May grow infinitely large
- ✚ If one processor fails then the rest survives
- ▬ Quite hard to program



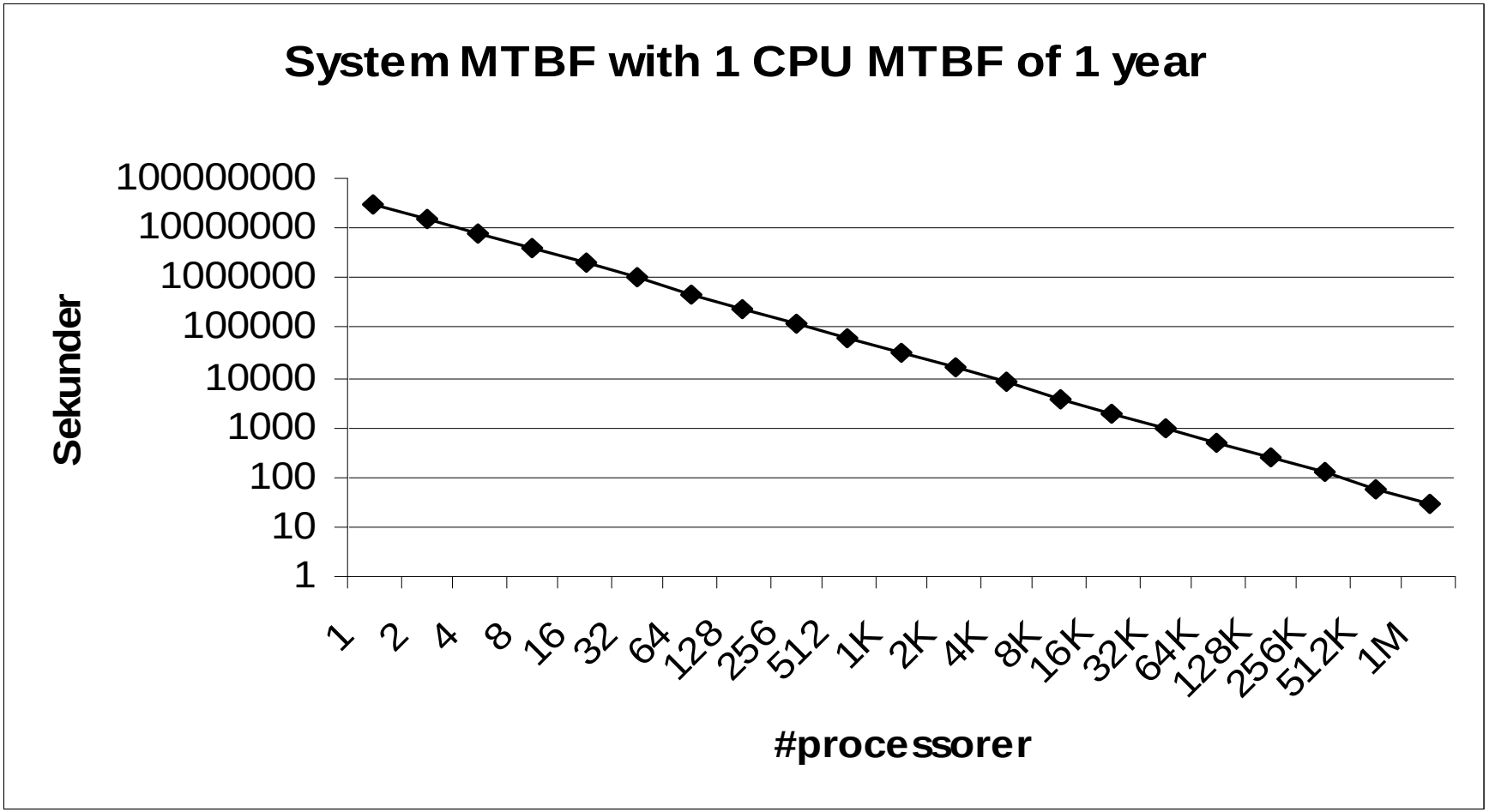
Why worry about errors?

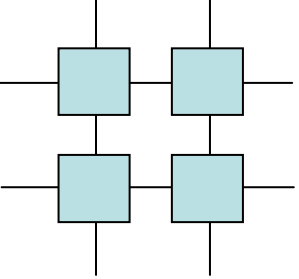
- Because the mean time between failure (MTBF) grows linearly with the number of CPUs
- Assuming one failure per CPU per year
  - With 1000 CPUs we should experience a failure every 9 hours



Cluster Computing

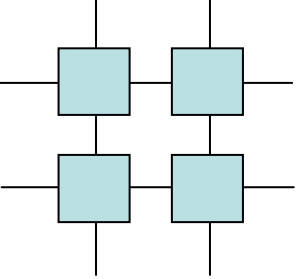
Why worry about errors?





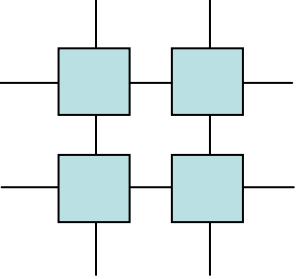
Important Decisions

- Which network to use?
  - Latency
  - Bandwidth
  - Price
- Which CPU architecture to use?
  - Performance (FP)
  - Price
- Which node architecture to use?
  - Performance: local and remote communication
  - Price



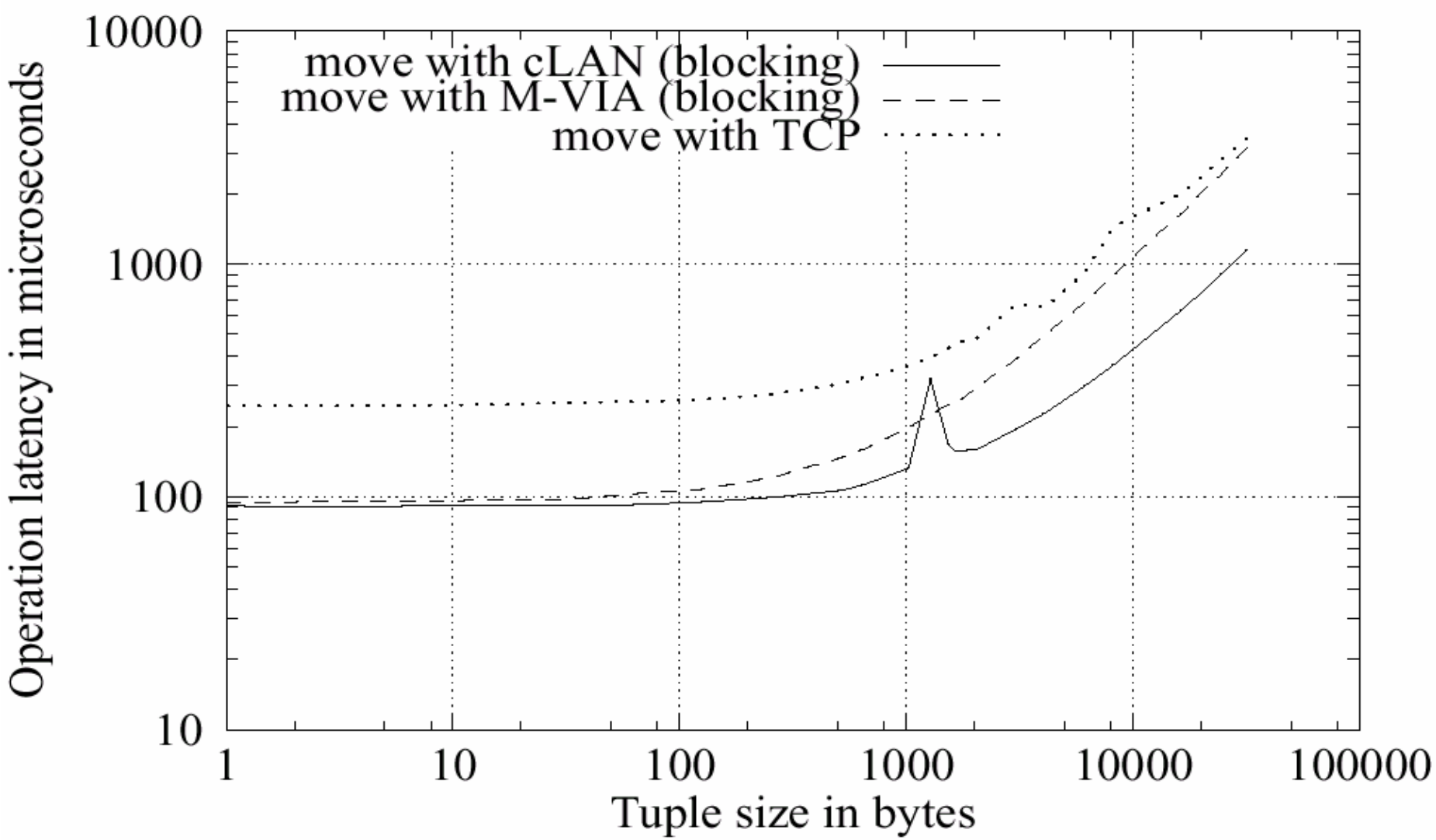
Cluster Networks

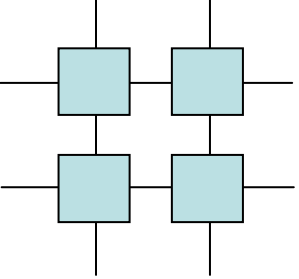
• FastEther	$ 50 per node
• VIA (cLan, etc...)	$1200 per node
• Myrinet	$2000 per node
• SCI	$2500 per node
• Quadrics	$4000 per node



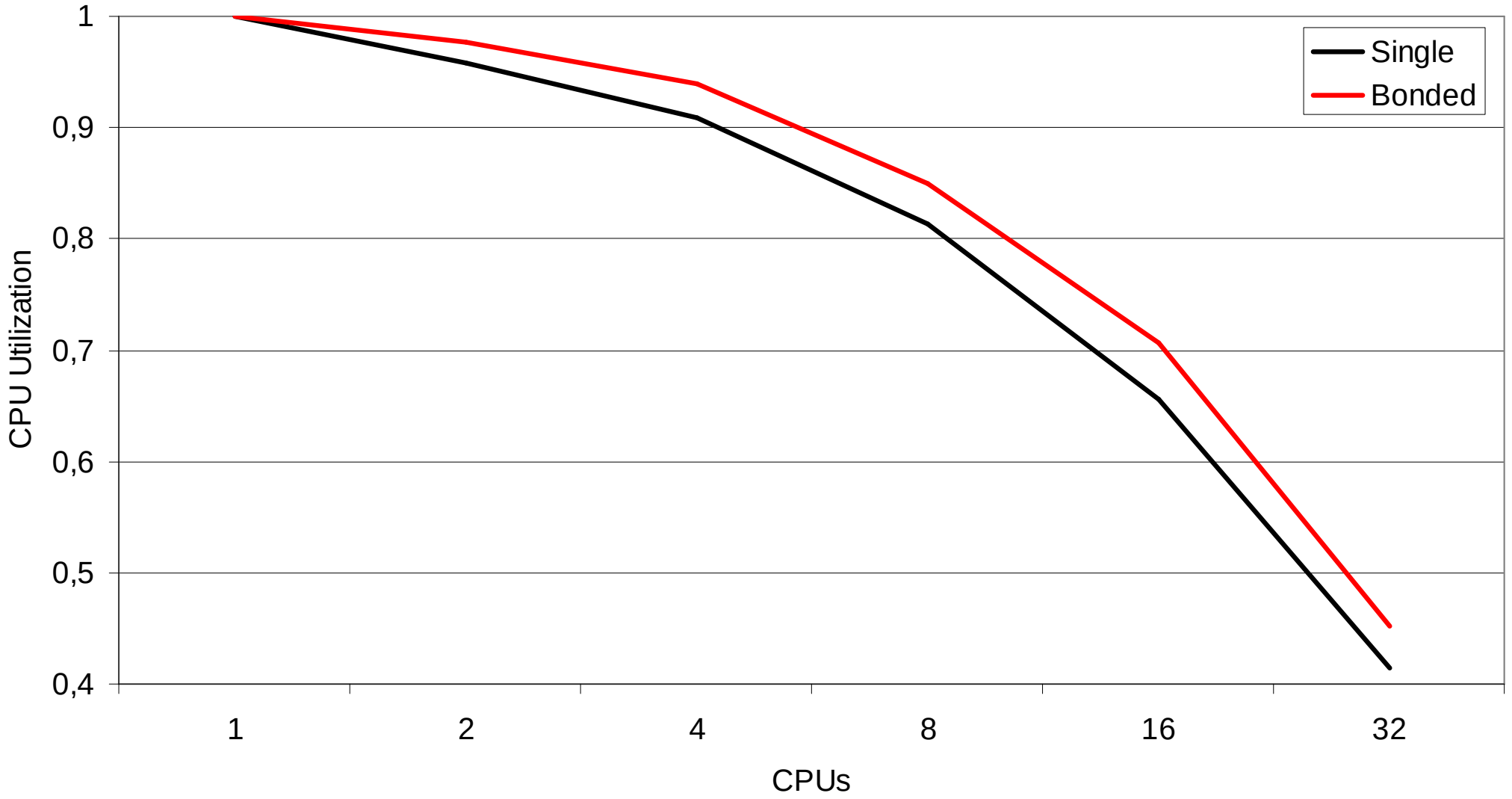
Cluster Computing

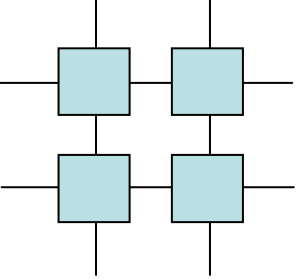
Elimination of TCP





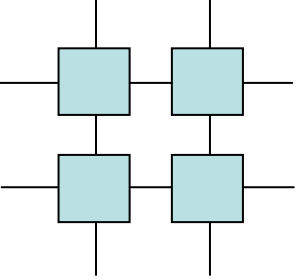
Gaussian Elimination Using one and two NICs





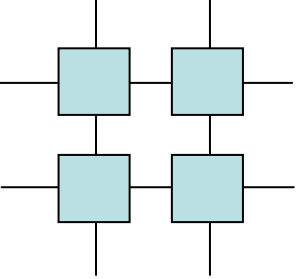
Which CPU?

- P3
  - SPEC-2000: 454/292 kr. 5.200 per CPU; 1Ghz
    256KB cache, 512MB ram!
- P4
  - SPEC-2000: 515/543 kr. 7.000 per CPU; 1.5 GHz
    256KB cache, 1GB ram
- Athlon
  - SPEC-2000: 496/426 kr. 5000 per node; 1.4 GHz
    256 KB cache 1GB ram



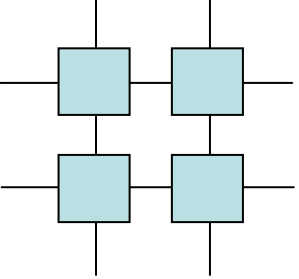
Which CPU?

- Itanium
  - SPEC-2000 370/711 kr. 50.000 per CPU; 733 MHz 2MB cache, 1GB ram
- Alpha
  - SPEC-2000 380/514 kr. 50.000 per CPU; 667 MHz 4MB cache 256 MB ram
- Power604e
  - SPEC-2000 248/330 kr. 80.000 per CPU; 375 MHz 8 MB cache, 512 MB ram

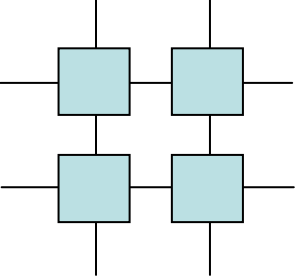


Why P4 (and not Athlon)

- Athlon had a 10% price performance advantage, but…
- Heat problems
  - We burn 95KW
- Because Athlon burns if it overheats
  - Well – it did in 2001 :)
- But P4 uses Thermal Throttling...

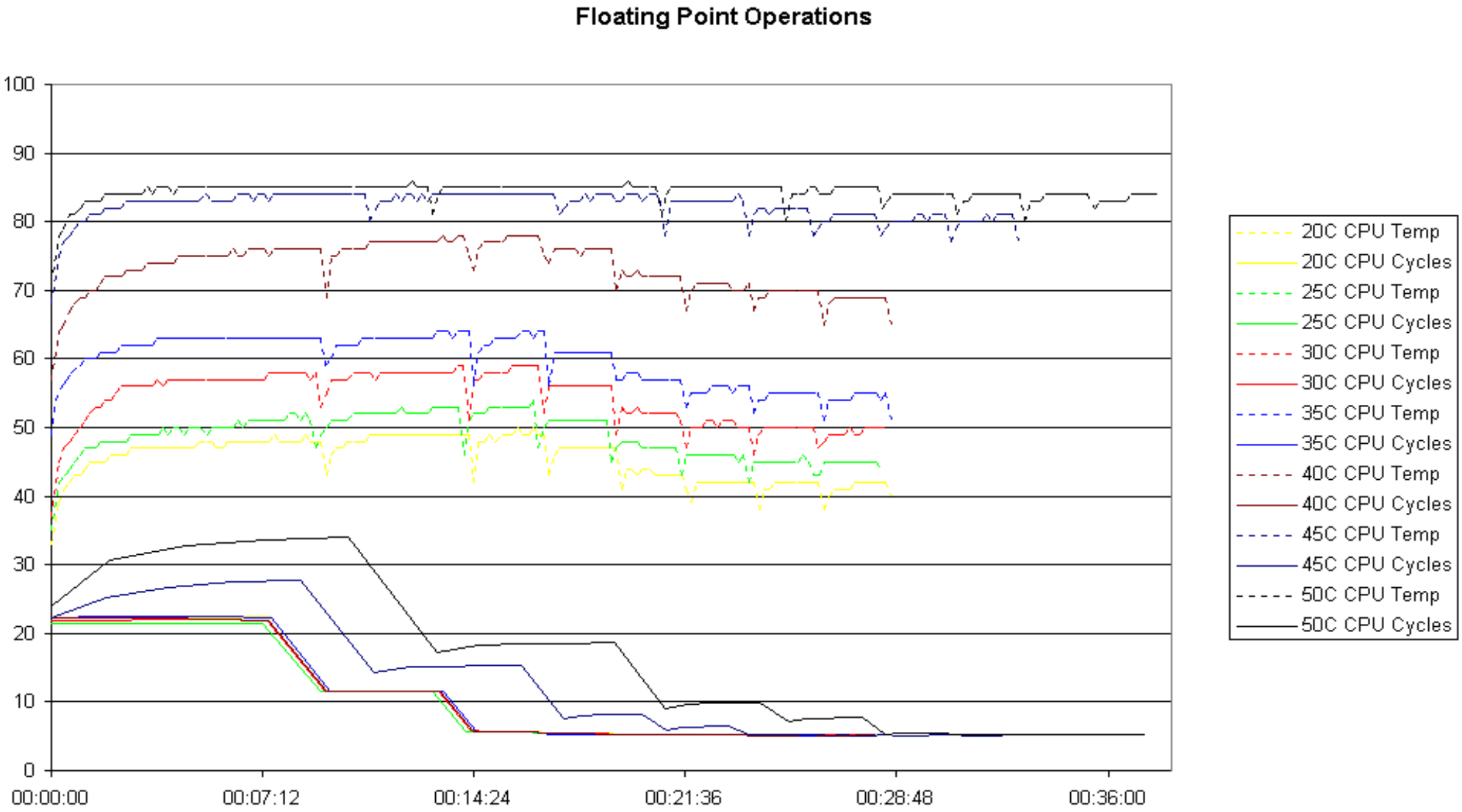


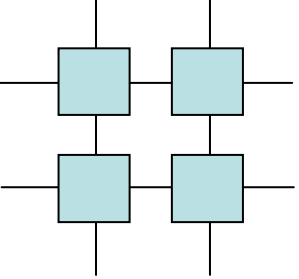
Thermal Throttling



Cluster Computing

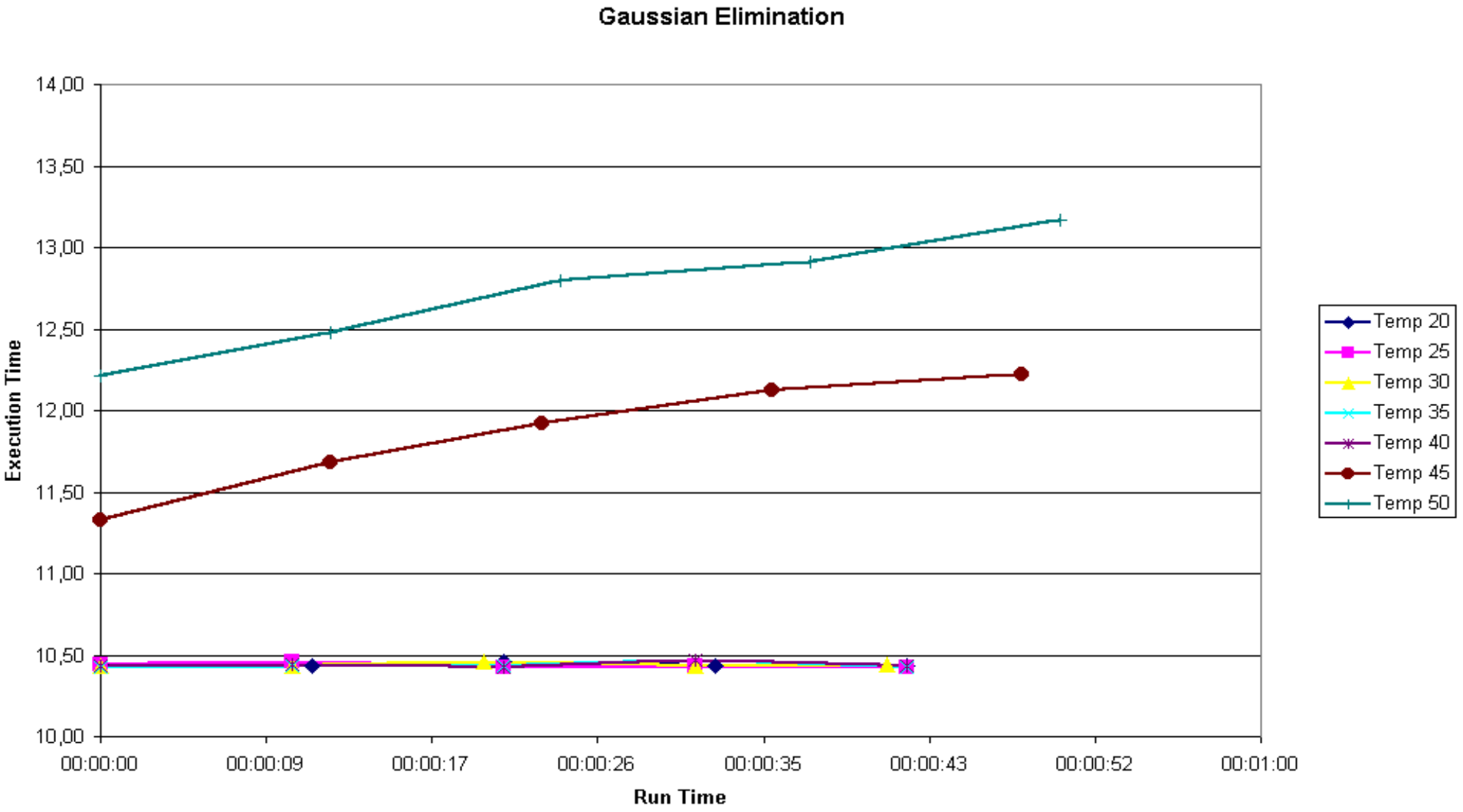
Thermal Throttling

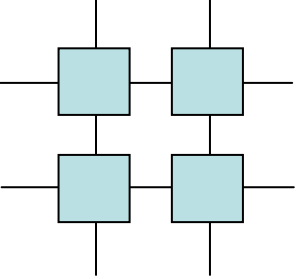




Cluster Computing

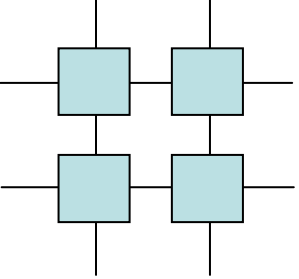
Thermal Throttling





Why uniprocessors

- Processor memory bandwidth is the most scarce resource in the system
  - Most users can’t code efficiently for large caches
- Interrupt latency is drastically increased in SMP mode

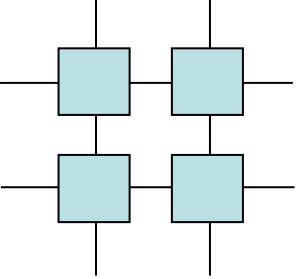


Cluster Computing

Elimination of TCP

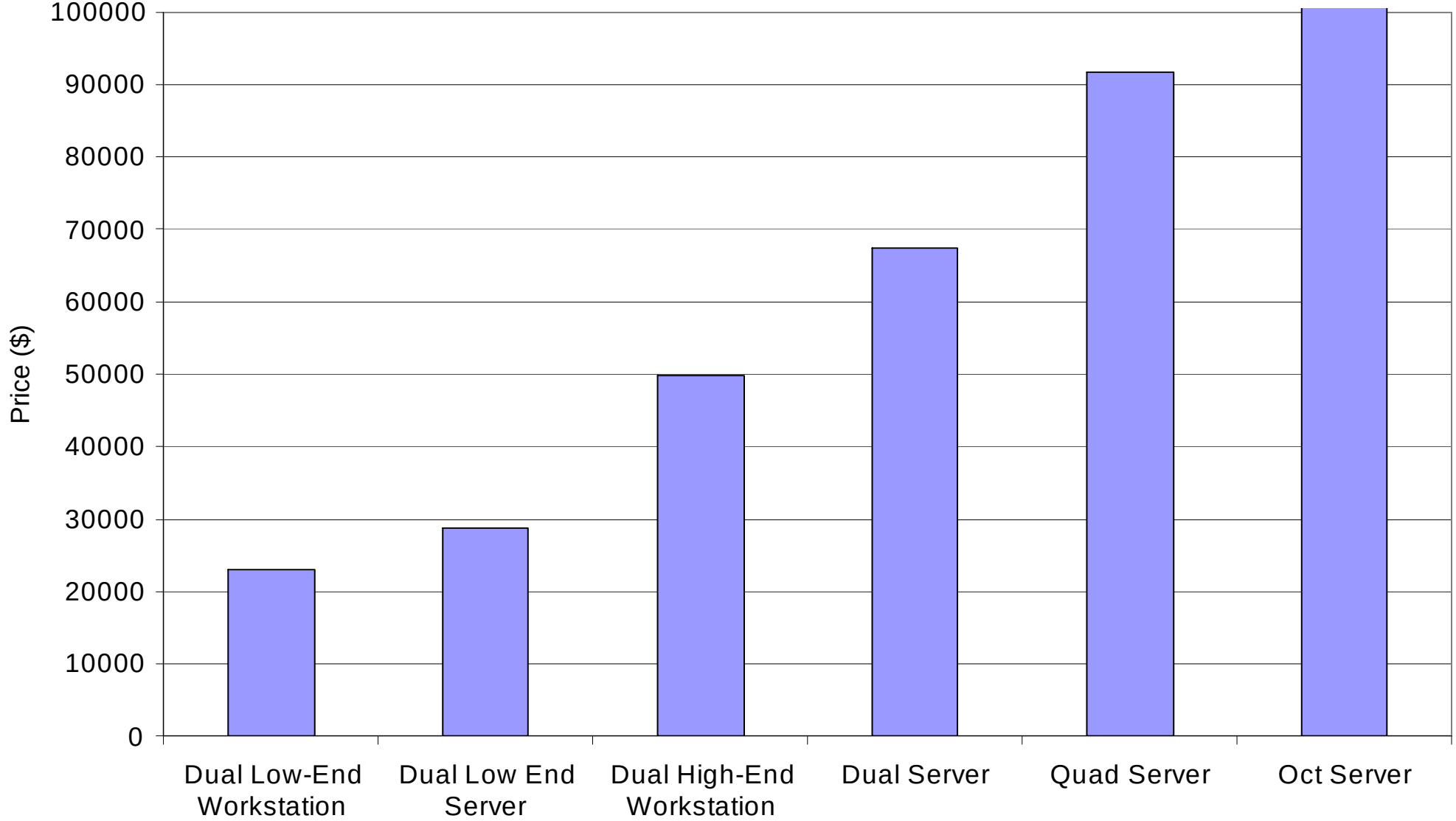
32 bytes payload

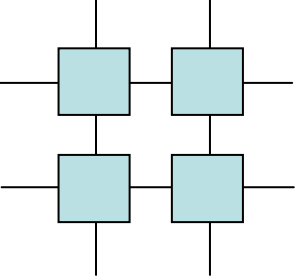
Communication mechanism	One-way latency SMP	One-way latency uniprocessor
TCP	246 us	206 us
UDP	193 us	156 us
PF_PACKET	165 us	126 us
UL-UL over proc	127 us	100 us
UL threads, loops in kernel	105 us	89 us
Kernel interrupt handler version	75 us	66 us
Same with HUB	63 us	54 us
Same with B2B cable	62 us	53 us



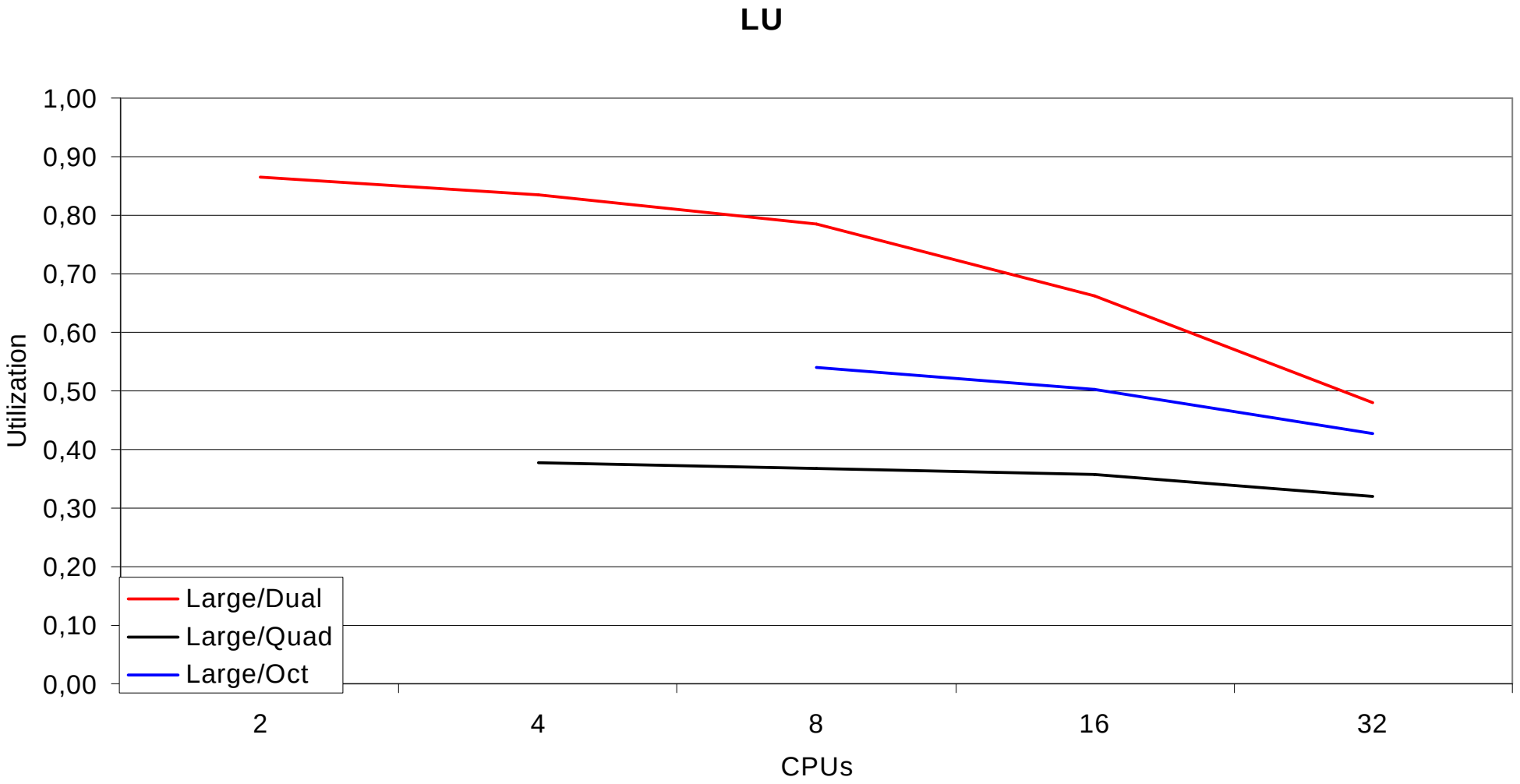
Cluster Computing

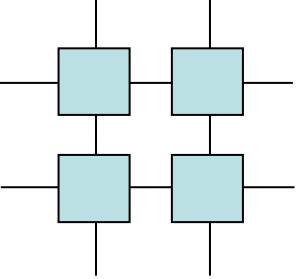
Single or SMP?



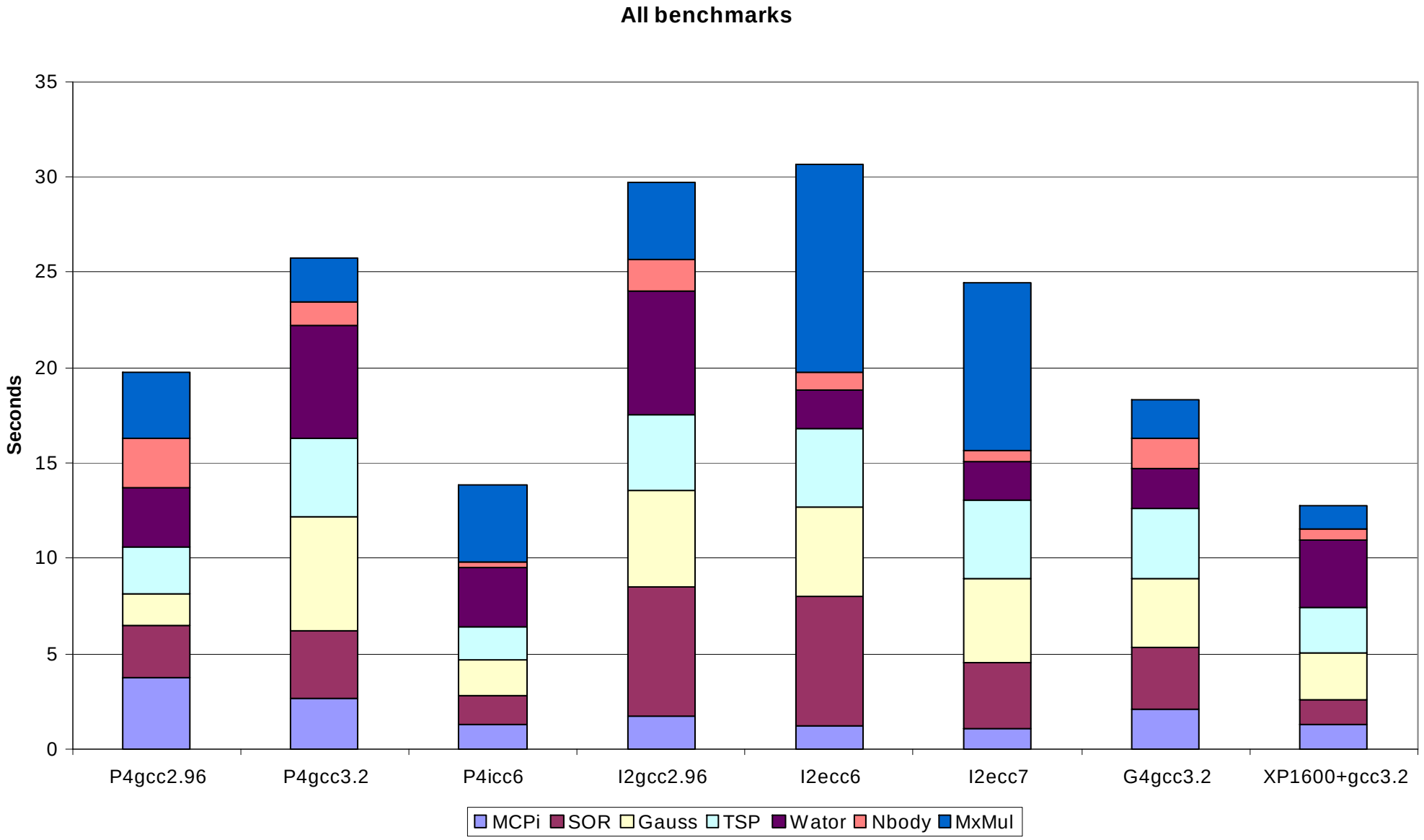


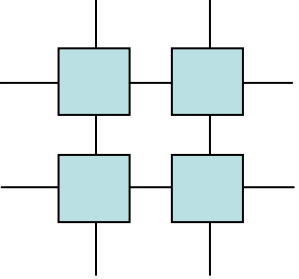
Single or SMP?




Cluster Computing

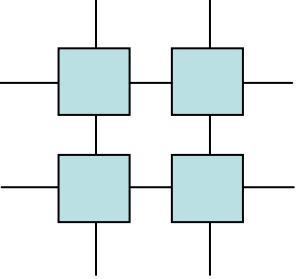
Compilers


All benchmarks
Seconds
35
30
25
20
15
10
5
0
P4gcc2.96
P4gcc3.2
P4icc6
I2gcc2.96
I2ecc6
I2ecc7
G4gcc3.2
XP1600+gcc3.2
MCPi
SOR
Gauss
TSP
Wator
Nbody
MxMul



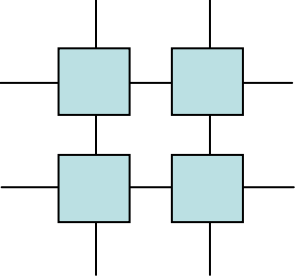
Implementation

- Use a brand name cluster solution
- Do it yourself
  - Lots of money to be saved here!

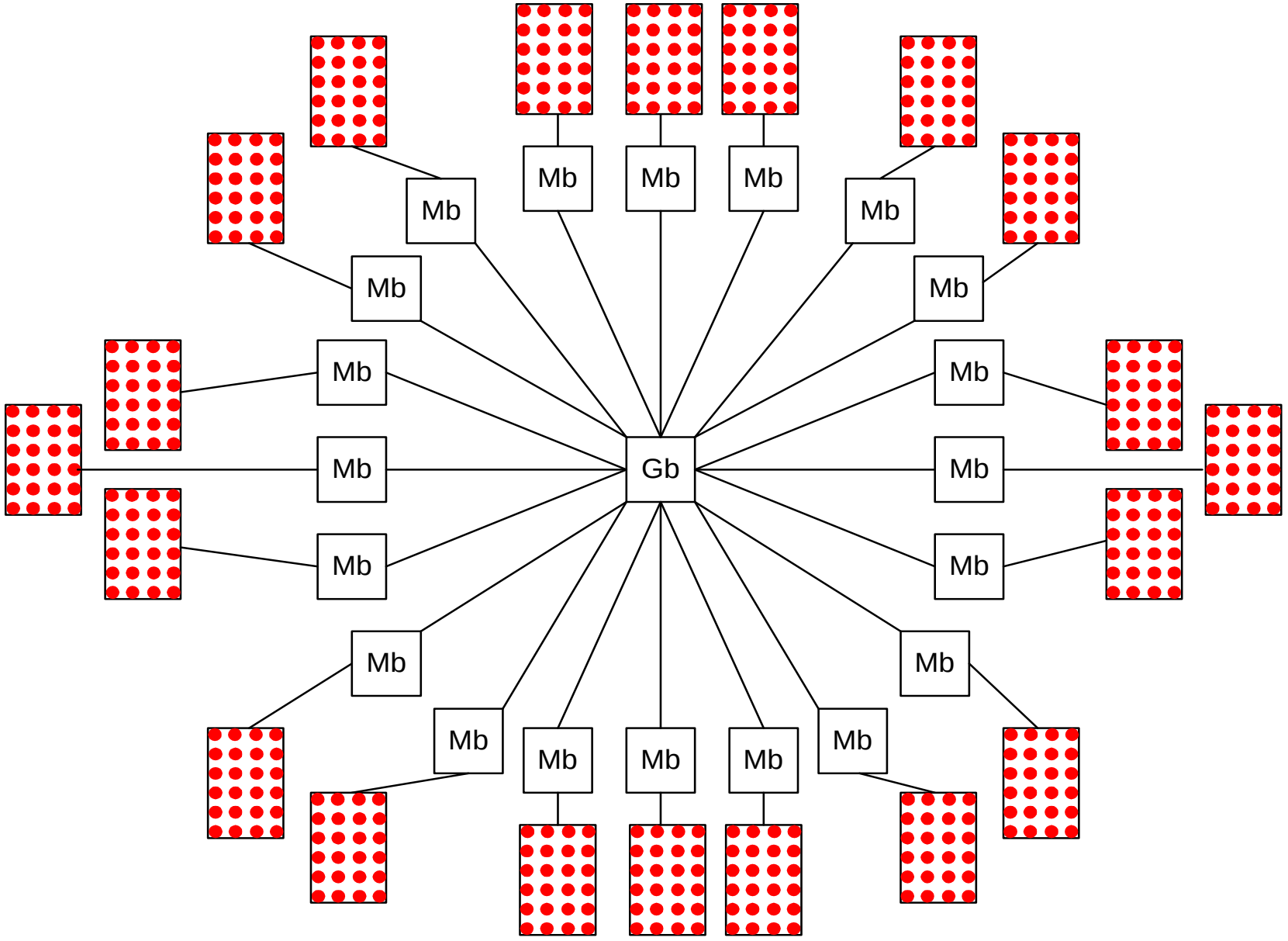


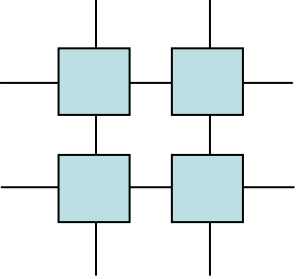
Our recipe

- One takes
  - 520 computers
  - 26 switches
  - 1.5 KM Cat-5e cable
  - 1200 TP plugs
  - 7 TP pliers
  - 7 students
  - 2 ks of beer and 35 pizzas

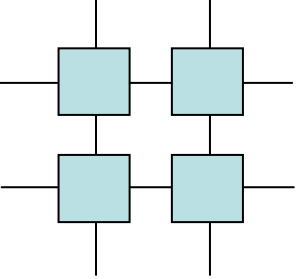


Architecture

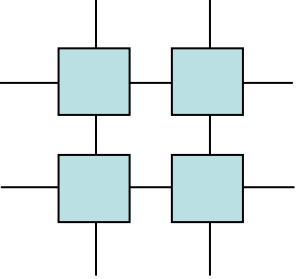




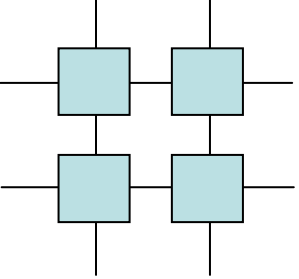
SDU Cluster



SDU Cluster

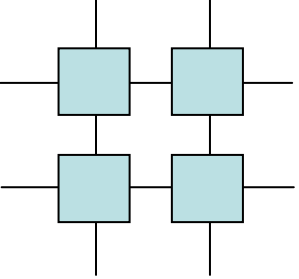


DTU Cluster



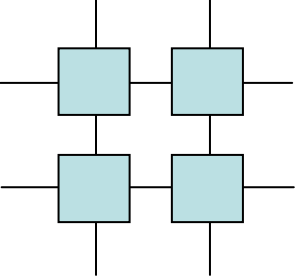
Cluster Software

- Installation programs
- Administration programs
- Programming



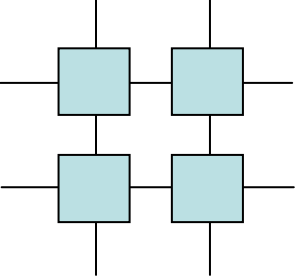
Installation Programs

- OSCAR
- Mandrake CLIC
- System Imager
- KA-BOOT
  - Very efficient
  - Thus our choice



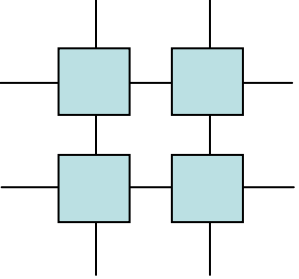
Administration programs

- Portable Batch System
  - OpenPBS
  - PBS-Pro
    - Commercial
    - But use UDP rather than TCP
- MAUI Scheduler
  - All the degrees of freedom one can ask for



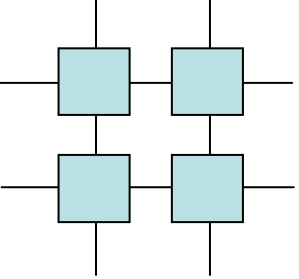
Cluster Programming

- Message Passing Interface
  - LAM MPI
  - MPICH
  - MESH-MPI
- Parallel Virtual Machine
  - PVM
- Distributed Shared Memory
  - Linda
  - PastSet/TMem



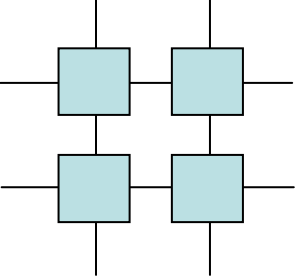
Unforeseen problems

- Air-condition
  - The air-condition had the reverse airflow from what we specified
- Power
  - Machines use far more power that specified
  - After a power failure power consumption approximates infinite...



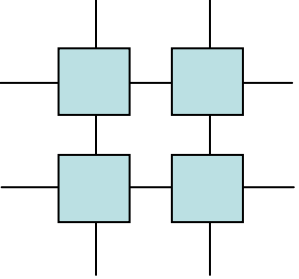
Unforeseen problems

- There is more to a hard drive than rotation speed and seek latency
  - One brand runs 10C hotter than the other
- When you order 4TB disk is comes configured for Windows as default...
- Large manufactures are far less professional at logistics than one would expect



Conclusion

- It’s a success
  - The users are very happy and the now 1430 CPU’s provide more than 80% of the available resources in Denmark
- A large production cluster is harder than an experimental department cluster



Conclusion

- But it’s still worth while
  - We provide three times more performance than if we bought a brand-name cluster
  - There are five times more CPUs than if we’d gone with cluster-interconnect

HORSESHOE

Scandinavia's Largest Cluster Supercomputer.
Funded by the Danish Center for Scientific Computing.

University of Southern Denmark

950 Processors
4.7 TFLOPS Peak Performance
950 GB Distributed Memory
100 TB On-line Disk Storage
10 GBIT/1 GBIT Network Infastructure
Powered by INTEL Pentium™ Processors
Operational Since 15/7 2002