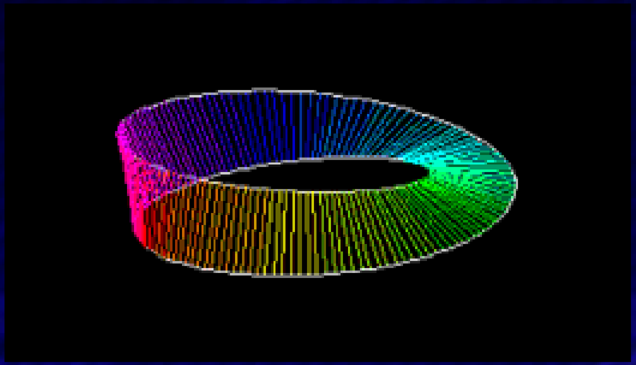
Beowulf Clusters

Pick Three

What is a Beowulf?

A cluster of M2COTS PCs connected by a low cost LAN running an Open Source OS and executing parallel applications

—Tom Sterling, 1999



What Is a Beowulf?

Collection of new or used computer hardware that is connected in parallel

First developed at GSFC in 1995

Beowulf is…

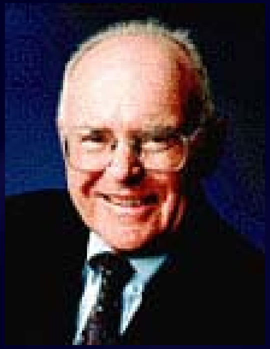
Offers near supercomputer speed on some complex algorithms

Second Generation developed at GSFC in 1997

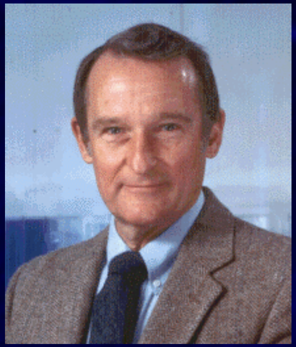
The Law Makers

Grace Hopper

Gene Amdahl



Gordon Moore



Seymour Cray

Marc Snir

Dr. Zaius

Definition: Good

- High-speed Processor
- Lots of RAM
- Fast network
- Amdahl's "Law" - one instruction per second requires one byte of memory and one bit per second of I/O
- *Ex. ASCI White (QCD / CFD)*

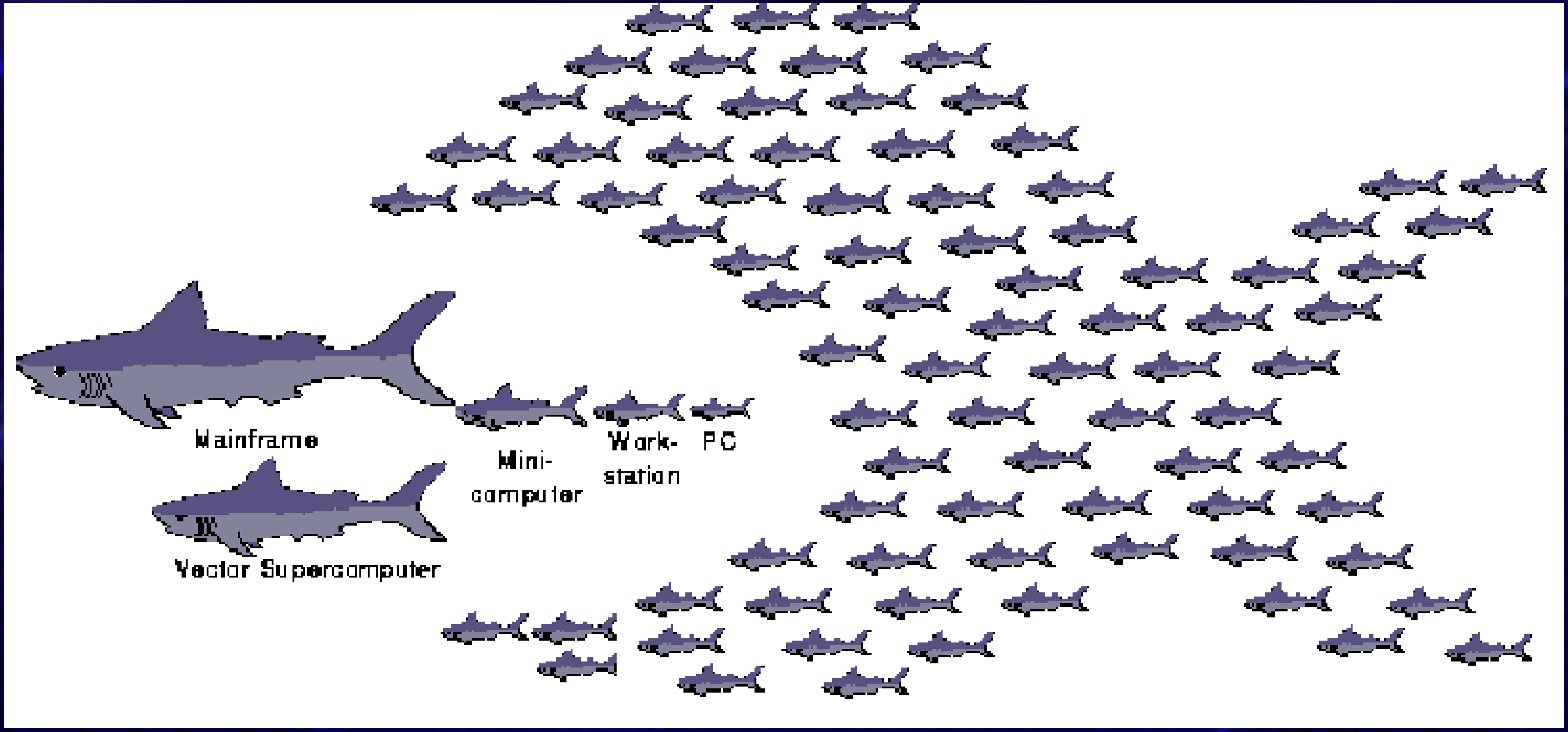
Definition: Cheap

- Stay one generation behind
- Last year’s network
- What’s on the table?
- Buy in quantity
- Getting something for nothing
- *Ex. Stone Soupercomputer*

Definition: Fast (acquisition)

- Build or buy?
- Many vendors in this space
- But not the “big guys”
- Arithmetic labor equation from node design/buildout
- Off-the-shelf clusters
- *Ex. ibm.com/servers/eserver/clusters*

Why do Clusters Blow Doors?



SPOF vs. distribution

- 1500 Mhz
- 99.999% uptime

- 3 x 500 Mhz
- 99.9% uptime

Chance of a single failure = MTBF / n

Chance of catastrophic (SPOF) failure = MTBF / n in the monolithic case

Amdahl’s Law…

If N is the number of processors, s is the amount of time spent (by a serial processor) on serial parts of a program and p is the amount of time spent (by a serial processor) on parts of the program that can be done in parallel, then Amdahl's law says that speedup is given by

Speedup = (s + p) / (s + p / N) = 1 / (s + p / N),

where we have set total time s + p = 1 for algebraic simplicity. For N = 1024, this is an unforgiving steep function of s near s = 0

…is overgeneralized

- Amdahl’s law is comprehensive, it is concise, it is mathematically elegant, and also demonstrably wrong. Modern clusters routinely violate Amdahl’s Law .
- A more realistic law is more like $S(N) \sim S_{AMDAHL}(N) \ / \ [1 + f_{COMM} \times R_{P/C}]^*$ Where f_{COMM} is the fraction of work devoted to communications and $R_{P/C}$ is the ratio of processor speed to communications speed.

Number of processors

Degree of parallelism

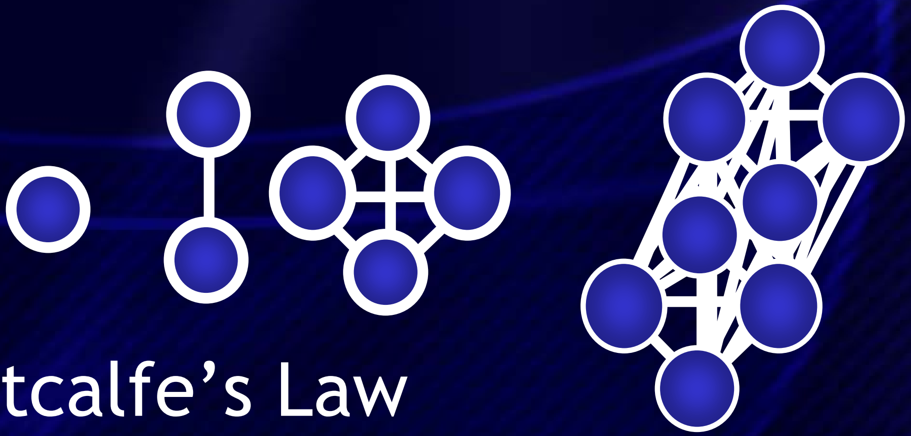
	2	4	8	32	64	256	512	1024
0.7	1.54	2.11	2.58	3.11	3.22	3.30	3.32	3.33
0.8	1.67	2.50	3.33	4.44	4.71	4.92	4.96	4.98
0.85	1.74	2.76	3.90	5.66	6.12	6.52	6.59	6.63
0.9	1.82	3.08	4.71	7.80	8.77	9.66	9.83	9.91
0.95	1.90	3.48	5.93	12.55	15.42	18.62	19.28	19.64
0.97	1.94	3.67	6.61	16.58	22.15	29.60	31.35	32.31
0.99	1.98	3.88	7.48	24.43	39.26	72.11	83.80	91.18
0.999	2.00	3.99	7.94	31.04	60.21	203.98	338.85	506.18

Expected Performance

- Relates to Amdahl's law
  - Adding more compute nodes may not be realized
  - Must write better parallelized code
  - If 8 processors are employed and code is only 90% parallelized net result is 4.71
    - 3.29 processors will not be fully utilized

General Cluster Advantages

- Beowulf clusters are virtually immune from the split brain effect, vaporlock, vicious cycles and the dreaded Christmas tree light syndrome.
- High-availability
- Fail-over
- Mission-critical
- Capitalizes on Metcalfe’s Law



“The usefulness, or utility, of a network equals the square of the number of users”.

Who and Where
A few archetypal examples:

- Warren / Salmon / Savarese (cnls.lanl.gov/avalon/)
- Don Becker (loki-www.lanl.gov/)
- Tom Sterling (www.cacr.caltech.edu/beowulf/)
- Hank Dietz (www.aggregate.org/KLAT2)
- Eadline, Lindahl, Lindheim

Why

- Traditional SC Apps (top500.org)
- Render Farms (POVPVM)
- Compute intensive operations
- Gordon Bell Prize
- Revitalize old PCs
- DBMS (Oracle Parallel Server)
- RC5, (Seti, Folding, Genome)@home

What a Gflops cost (when)
63K in ’96 (Loki), 28K in ‘97

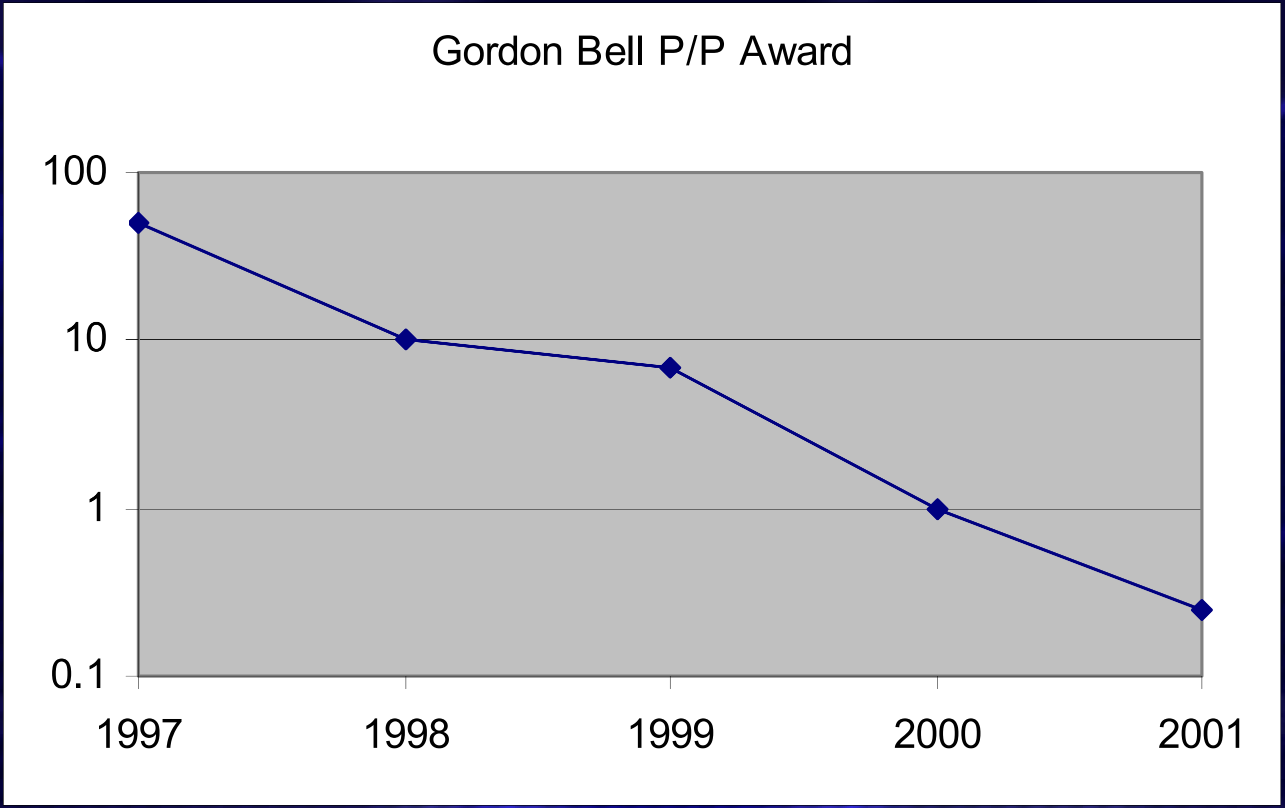
6 Pentium II's with 64MB RAM, 2GB disks	$12,000
6 SMC EtherPro 10/100 Network Cards	600
1 Bay 350T Fast Ethernet Switch	2,500
6 Category 5 UTP Cables	60
Linux CD and MPICH	40
Total, 1998	**$15,200**
Total, 1999	**$10,000**
Total, 2000	**$4,000**
Total, 2001	**$1,300**

What

Supercomputing Conference Award Winners

Team and Year	$ / MFLop
97 Becker (Loki)	50
98 Christ QCD	10
00 Bunyip	1
01 Hwang /Kim /Lee	.25

97 Becker (Loki)	$50
98 Christ QCD	$10
00 Bunyip / KLAT2	$ 1
01 Hwang /Kim /Lee	25¢

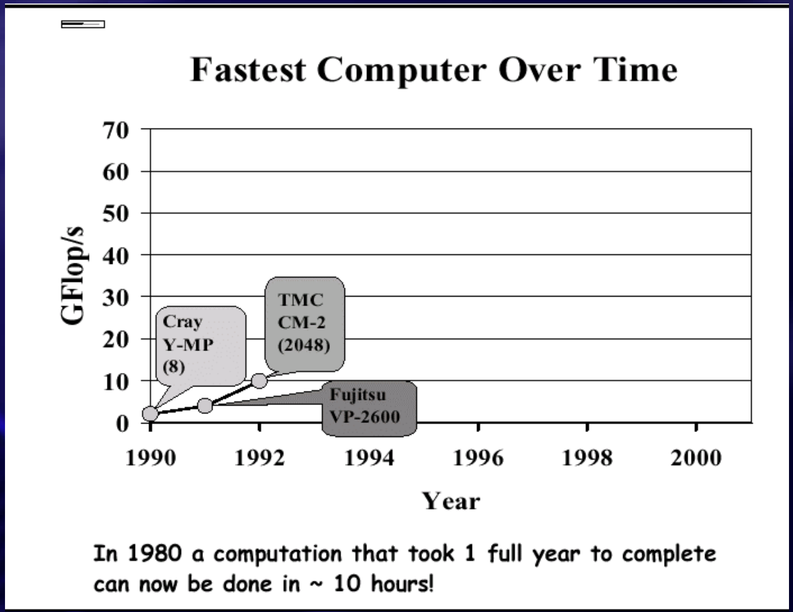


Team and Year	$ / MFLop
02 None (see www.sc-2002.org)	-
03 Phoenix (sc-conference.org/sc2003/)	1, Top 100

Moore's Law:

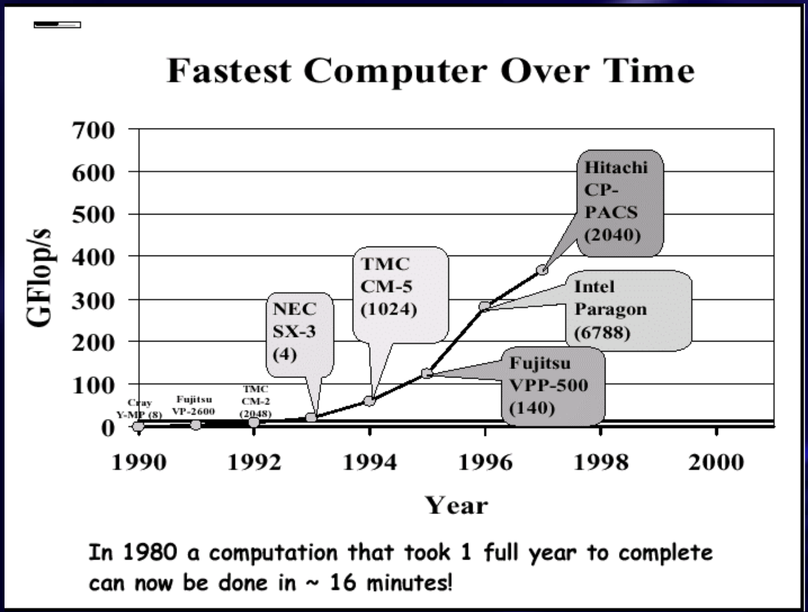
I'm not quite dead yet!

- New fab techniques
- 3-D Processor
- Infiniband
- Blue Gene (/L)
- Greater than 60% over last 5 years



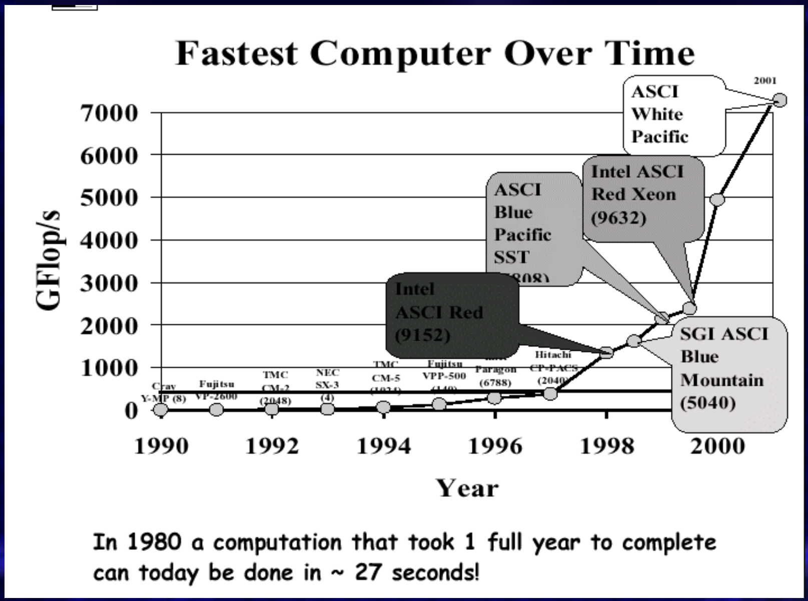
Fastest Computer Over Time

In 1980 a computation that took 1 full year to complete can now be done in ~ 10 hours!



Fastest Computer Over Time

In 1980 a computation that took 1 full year to complete can now be done in ~ 16 minutes!

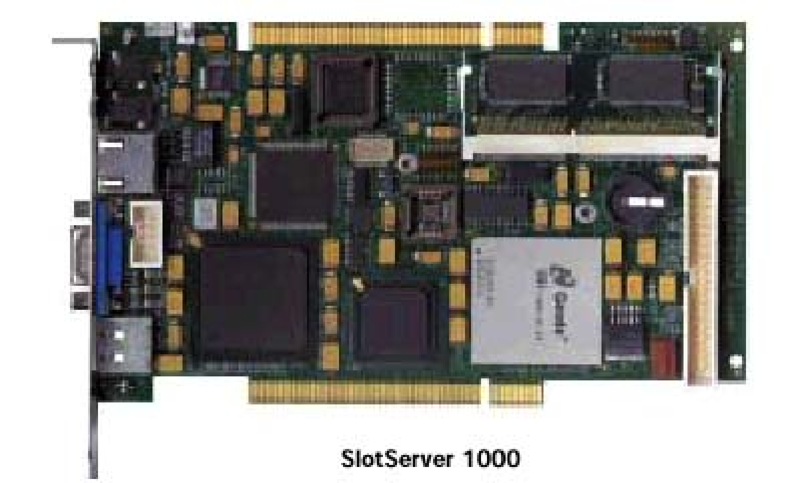


Fastest Computer Over Time

In 1980 a computation that took 1 full year to complete can today be done in ~ 27 seconds!

What about today?

Densely integrated HPC solutions



Single-board Computers

Blade solutions

Ripped from today’s headlines

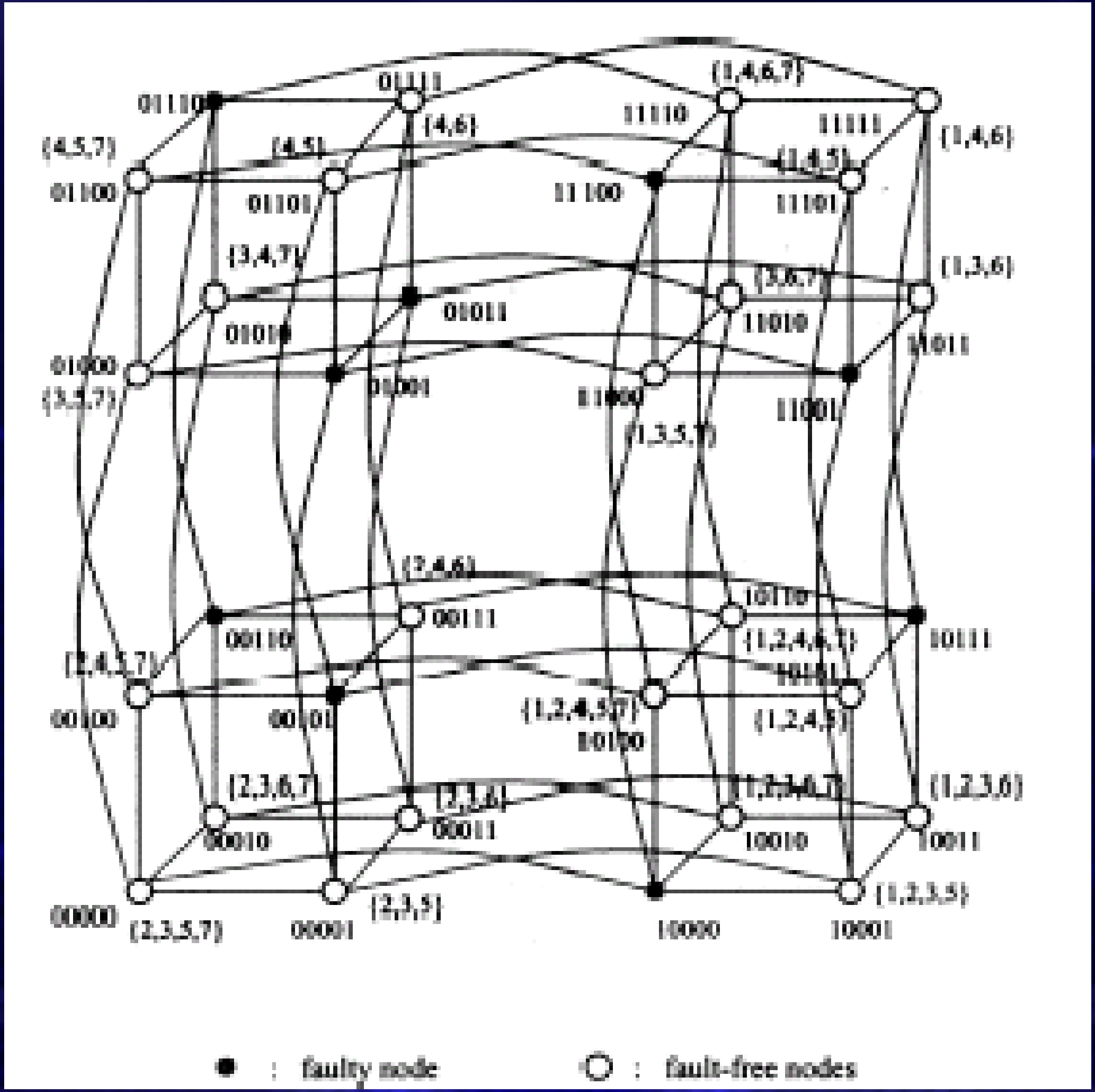
- “Off-the-shelf supercomputer”
  - MachineDesign.com
- Superspecialized SBCs
  - Clearspeed.com
- Exotic Topologies
  - MIT Roofnet

Virginia Tech is building a supercomputer said to be one of the cheapest and most-powerful home-built machines in the world. It consists of 1,100 Apple G5 desktop computers tied together into a configuration that is reliable and much cheaper than huge mainframe supercomputing systems. Virginia Tech engineering professor Srinidhi Varadarajan is incorporating into the 1,100-node cluster a software package called Deja vu that he developed for stabilizing such systems. The University hopes the new supercomputer will bring it numerous big science research projects that it did not have resources to handle previously.

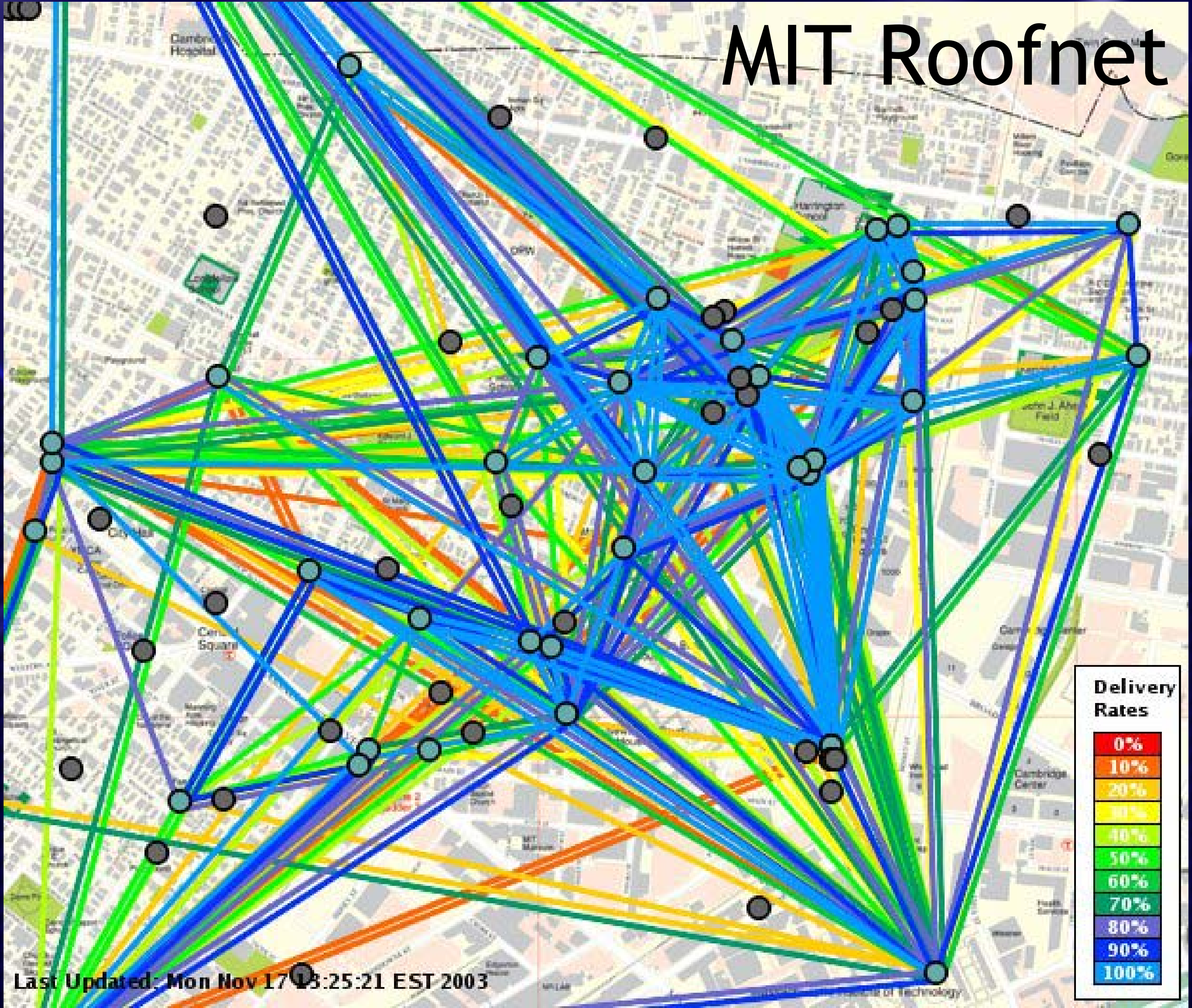
ClearSpeed

A small chip-design firm will unveil a new processor Tuesday it says will transform ordinary desktop PCs and laptops into supercomputers. The new chip is a parallel processor capable of performing 25 billion floating-point operations per second, or 25 gigaflops. According to the company, the chip has the potential to bring supercomputer performance to the desktop. An ordinary desktop PC outfitted with six PCI cards, each containing four of the chips, would perform at about 600 gigaflops (or more than half a teraflop). At this level of performance, the PC would qualify as one of the 500 most powerful supercomputers in the world. "That's a supercomputer on the desktop," said Simon McIntosh-Smith, ClearSpeed's director of architecture. The souped-up PC would cost about $25,000, ClearSpeed said. By comparison, most of the supercomputers on the Top 500 list are clusters of hundreds of processors and cost millions of dollars.

Hypercubes, wormholes



MIT Roofnet


Delivery Rates
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Last Updated: Mon Nov 17 13:25:21 EST 2003
Cambridge Hospital
Harrington School
City Hall
Central Square
John J. Ahern Field
Cambridge Center
Massachusetts Institute of Technology

TOP 5

SUPERCOMPUTER SITES (November 2003)

1

EARTH SIMULATOR

Earth Simulator Center
Yokohama
NEC
Rmax: 35.86 TFlops

2

ASCI Q

LANL
Los Alamos
HP Alphaserver SC
Rmax: 13.88 TFlops

3

Virginia Tech's X

Virginia Tech
Blacksburg, USA
Dual Apple G5/Mellanox
Rmax: 10.28 TFlops

4

Tungsten

NCSA
Urbana-Champaign, USA
Dell PowerEdge 1750/Myrinet
Rmax: 9.819 TFlops

5

MPP2

PNNL
Richland, USA
HP rx2600 Itanium2/Quadrics
Rmax: 8.633 TFlops

What’s on the table?

Academia

- Cheap student labor
- Cheap real estate
- Cheap electricity

The Real World

- ISP colocation expensive
- They meter everything
- $1000 / month for a 20 amp circuit

Still Important

- MPI, PVM
- Proprietary NICs (7 μs min)
- Interprocess commo; MOSIX, Enfuzion
- Interprocessor commo; SMP, RapidIO
- NICs; 100Mbps, 1 Gbps, 10Gbps (M-VIA, Gamma?)

Yesterday's news

- **Earth Simulator can perform 35.8 trillion operations a sec. and cost $700 million to build.**
- Japan's GDP was $3.73 trillion last year
- http://www.research.ibm.com/bluegene/
- Distros ; Callident, Scyld, Gentoo, Fedora?
- www.Windowsclusters.org
- Linux-ha.org, www.hp.connectthe.com/glinuxclus1/
- www.mersenne.org/prime.htm
- http://3.141592653589793238462643383279502884197169399375105820974944592.jp/

References

- lyre.mit.edu/~kkeville
- www.lcic.org
- www.extreme-linux.com
- *www.cacr.caltech.edu/cluster2001/program/talks/camp.pdf
- www.cacr.caltech.edu/cluster2001/program/talks/top500.pdf