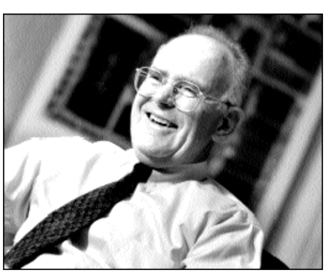
17th High-Performance Computing Symposium
1st OSCAR Symposium
May 11-14, 2003
Sherbrooke Delta Hotel
Québec, CANADA

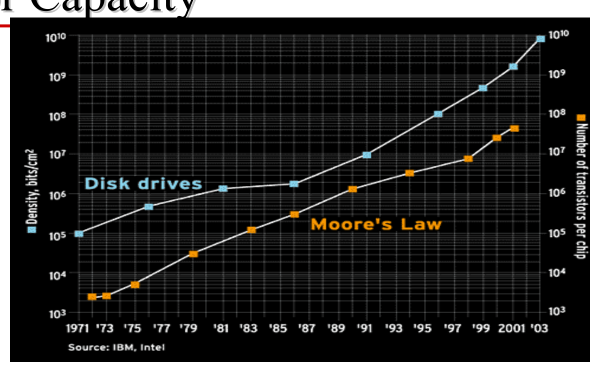
HPCS 2003 Sherbrooke, Québec CANADA

High Performance Computing, Computational Grid, and Numerical Libraries

Jack Dongarra
Innovative Computing Lab
University of Tennessee
and
Computer Science and Math Div
Oak Ridge National Lab
http://www.cs.utk.edu/~dongarra/

Technology Trends: Microprocessor Capacity



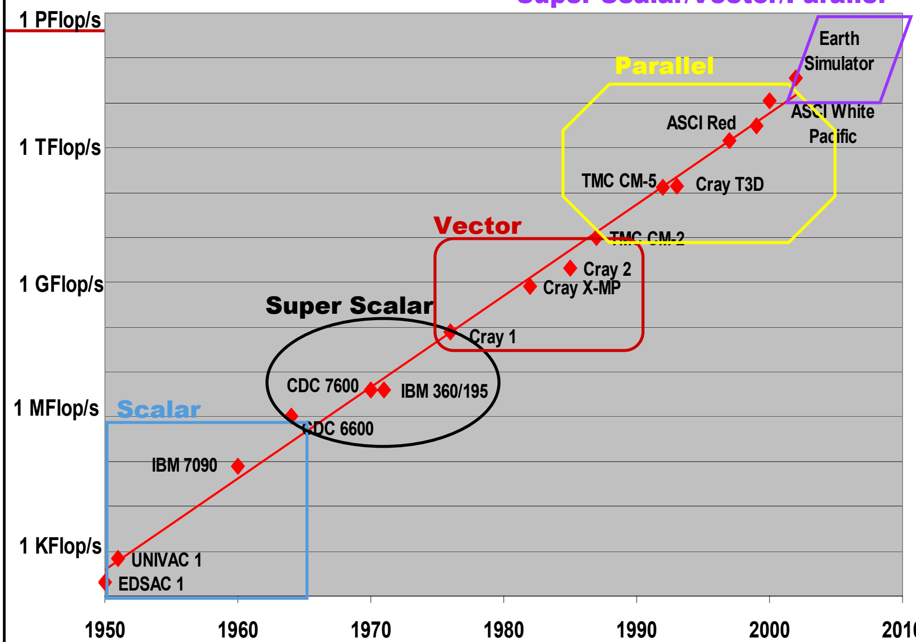


Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

2X transistors/Chip Every 1.5 years
Called "Moore's Law"

Microprocessors have become smaller, denser, and more powerful.
Not just processors, bandwidth, storage, etc.
2X memory and processor speed and ½ size, cost, & power every 18 months.

Moore's Law

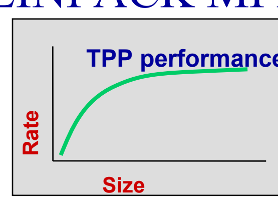


TOP500 supercomputer

H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

  Ax=b, dense problem

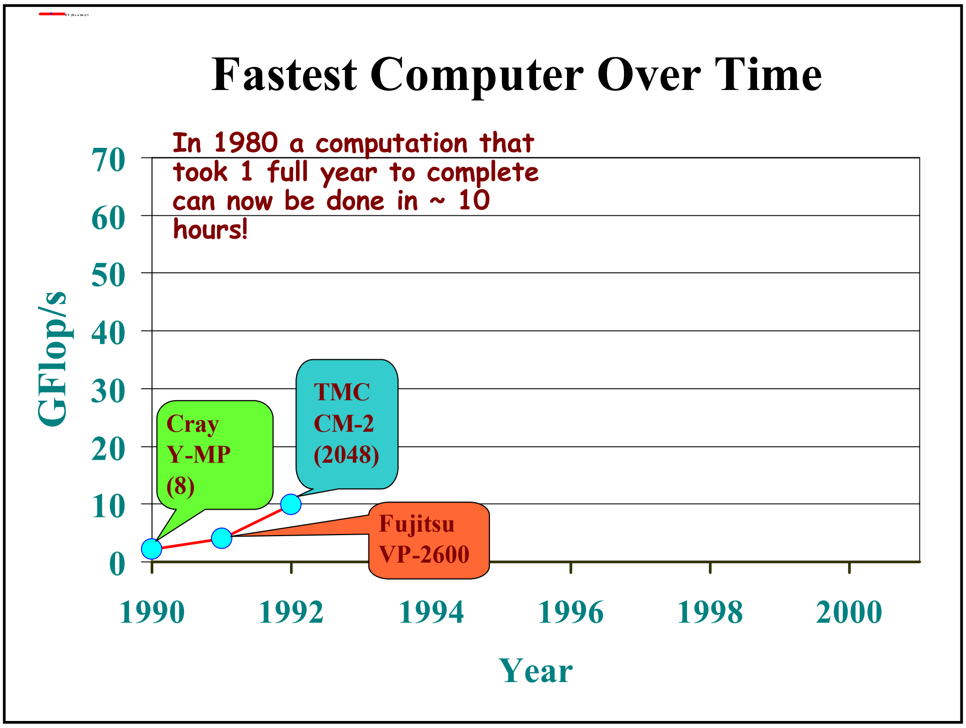


- Updated twice a year

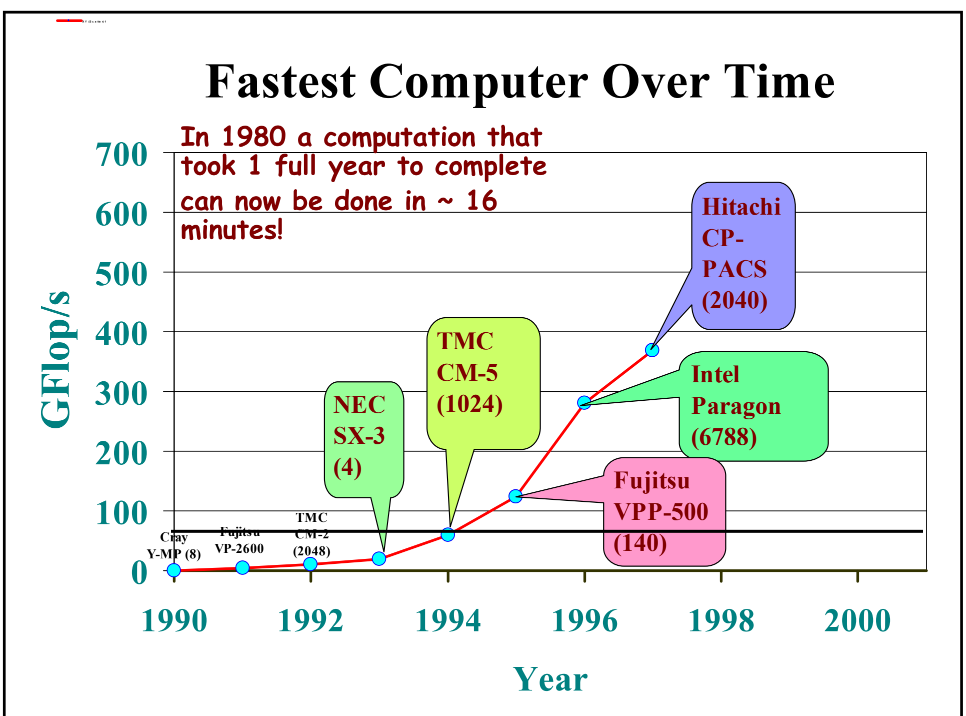
  SC'xy in the States in November

  Meeting in Mannheim, Germany in June

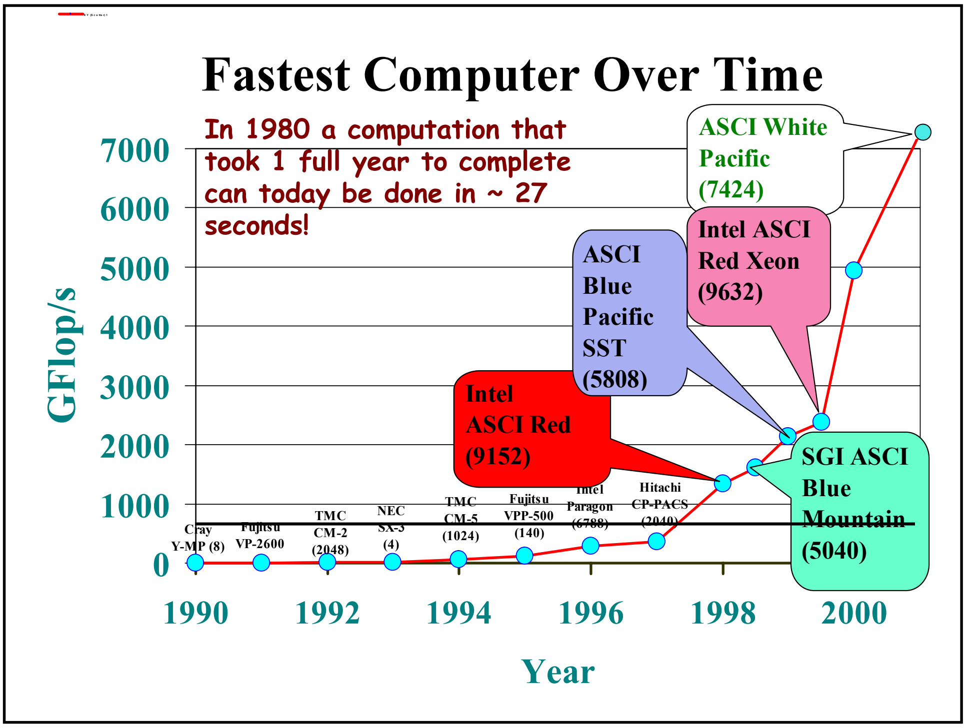
- All data available from **www.top500.org**

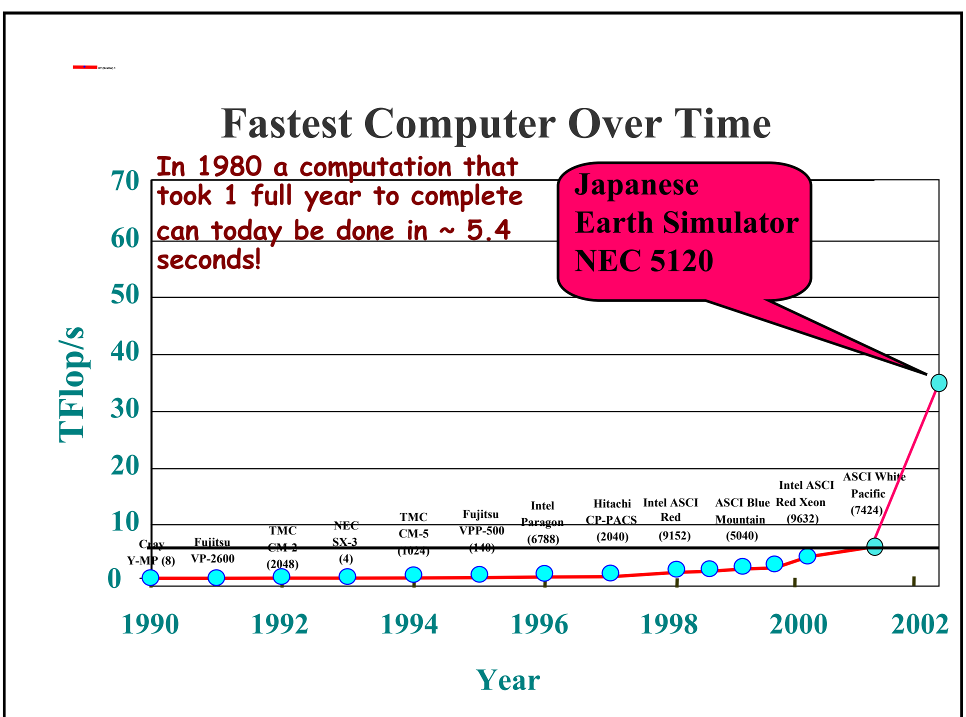


Fastest Computer Over Time
In 1980 a computation that took 1 full year to complete can now be done in ~ 10 hours!
GFlop/s
70
60
50
40
30
20
10
0
Cray Y-MP (8)
TMC CM-2 (2048)
Fujitsu VP-2600
1990
1992
1994
1996
1998
2000
Year



Fastest Computer Over Time
In 1980 a computation that took 1 full year to complete can now be done in ~ 16 minutes!
GFlop/s
700
600
500
400
300
200
100
0
Hitachi CP-PACS (2040)
Intel Paragon (6788)
TMC CM-5 (1024)
NEC SX-3 (4)
Fujitsu VPP-500 (140)
Cray Y-MP (8)
Fujitsu VP-2600
TMC CM-2 (2048)
1990
1992
1994
1996
1998
2000
Year


Fastest Computer Over Time
In 1980 a computation that took 1 full year to complete can today be done in ~ 27 seconds!
GFlop/s
7000
6000
5000
4000
3000
2000
1000
0
1990
1992
1994
1996
1998
2000
Year
Cray Y-MP (8)
Fujitsu VP-2600
TMC CM-2 (2048)
NEC SX-3 (4)
TMC CM-5 (1024)
Fujitsu VPP-500 (140)
Intel Paragon (6788)
Hitachi CP-PACS (2040)
Intel ASCI Red (9152)
ASCI Blue Pacific SST (5808)
Intel ASCI Red Xeon (9632)
ASCI White Pacific (7424)
SGI ASCI Blue Mountain (5040)

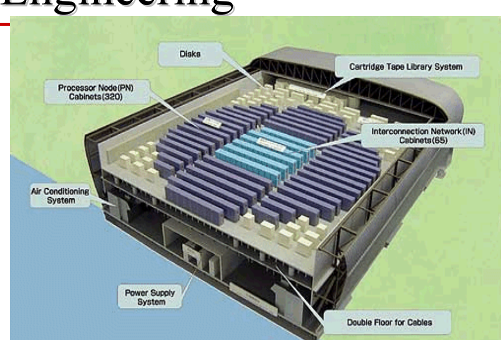

Fastest Computer Over Time
In 1980 a computation that took 1 full year to complete can today be done in ~ 5.4 seconds!
Japanese Earth Simulator NEC 5120
TFlop/s
70
60
50
40
30
20
10
0
1990
1992
1994
1996
1998
2000
2002
Year
Cray Y-MP (8)
Fujitsu VP-2600
TMC CM-2 (2048)
NEC SX-3 (4)
TMC CM-5 (1024)
Fujitsu VPP-500 (140)
Intel Paragon (6788)
Hitachi CP-PACS (2040)
Intel ASCI Red (9152)
ASCI Blue Mountain (5040)
Intel ASCI Red Xeon (9632)
ASCI White Pacific (7424)

Machines at the Top of the List

Year	Computer	Measured Gflop/s	Factor Δ from Pervious Year	Theoretical Peak Gflop/s	Factor Δ from Pervious Year	Number of Processors	Efficiency
2002	Earth Simulator Computer, NEC	35860	5.0	40960	3.7	5120	88%
2001	ASCI White-Pacific, IBM SP Power 3	7226	1.5	11136	1.0	7424	65%
2000	ASCI White-Pacific, IBM SP Power 3	4938	2.1	11136	3.5	7424	44%
1999	ASCI Red Intel Pentium II Xeon core	2379	1.1	3207	0.8	9632	74%
1998	ASCI Blue-Pacific SST, IBM SP 604E	2144	1.6	3868	2.1	5808	55%
1997	Intel ASCI Option Red (200 MHz Pentium Pro)	1338	3.6	1830	3.0	9152	73%
1996	Hitachi CP-PACS	368.2	1.3	614	1.8	2048	60%
1995	Intel Paragon XP/S MP	281.1	1	338	1.0	6768	83%
1994	Intel Paragon XP/S MP	281.1	2.3	338	1.4	6768	83%
1993	Fujitsu NWT	124.5		236		140	53%

A Tour de Force in Engineering

- Homogeneous, Centralized, Proprietary, Expensive!
- Target Application: CFD-Weather, Climate, Earthquakes
- 640 NEC SX/6 Nodes (mod)
  - 5120 CPUs which have vector ops
  - Each CPU 8 Gflop/s Peak
- 40 TFlop/s (peak)
- $250-$500 million for things in building
- Footprint of 4 tennis courts
- 7 MWatts
  - Say 10 cent/KWhr - $16.8K/day = $6M/year!
- Expect to be on top of Top500 until 60-100 TFlop ASCI machine arrives
- For the Top500 (November 2002)
  - Performance of ESC ≈ Σ Next Top 7 Computers
  - Σ of DOE computers (DP&OS) = 49 TFlop/s

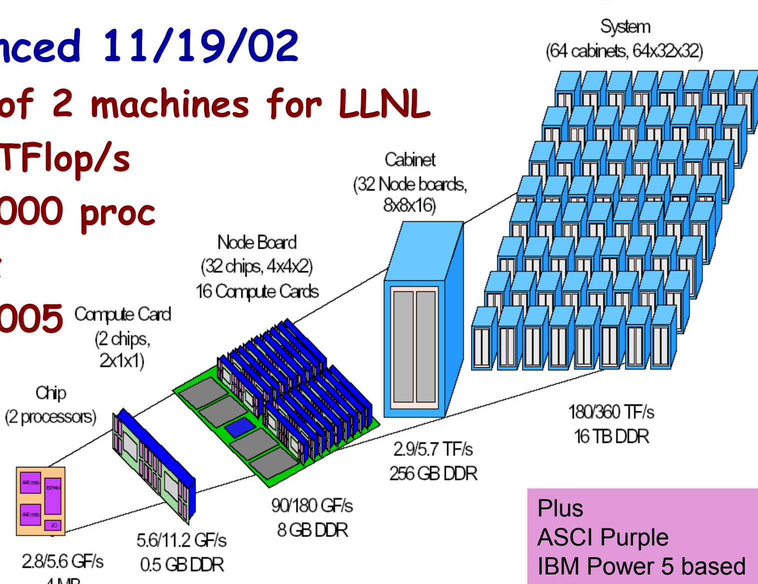


20th List: The TOP10

Rank	Manufacturer	Computer	R_{max} [TF/s]	Installation Site	Country	Year	Area of Installation	# Proc
1	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	Research	5120
2	HP	ASCI Q, AlphaServer SC	7.73	Los Alamos National Laboratory	USA	2002	Research	4096
2	HP	ASCI Q, AlphaServer SC	7.73	Los Alamos National Laboratory	USA	2002	Research	4096
4	IBM	ASCI White SP Power3	7.23	Lawrence Livermore National Laboratory	USA	2000	Research	8192
5	Linux NetworX	MCR Cluster	5.69	Lawrence Livermore National Laboratory	USA	2002	Research	8192
6	HP	AlphaServer SC ES45 1 GHz	4.46	Pittsburgh Supercomputing Center	USA	2001	Academic	3016
7	HP	AlphaServer SC ES45 1 GHz	3.98	Commissariat a l'Energie Atomique (CEA)	France	2001	Research	2560
8	HPTi	Xeon Cluster - Myrinet2000	3.34	Forecast Systems Laboratory - NOAA	USA	2002	Research	1536
9	IBM	pSeries 690 Turbo	3.16	HPCx	UK	2002	Academic	1280
10	IBM	pSeries 690 Turbo	3.16	NCAR (National Center for Atmospheric Research)	USA	2002	Research	1216

Response to the Earth Simulator: IBM Blue Gene/L and ASCI Purple

- **Announced 11/19/02**
  - One of 2 machines for LLNL
  - 360 TFlop/s
  - 130,000 proc
  - Linux
  - FY 2005



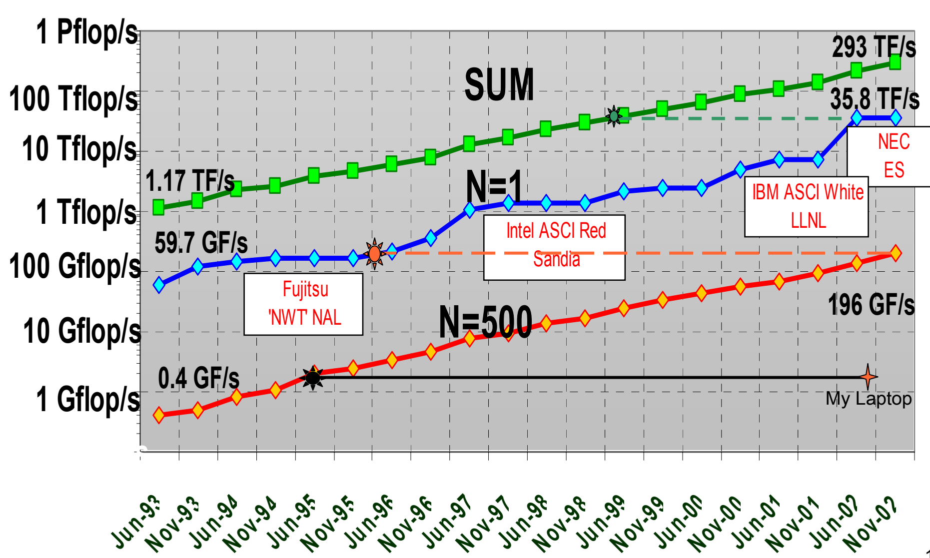
Plus
ASCI Purple
IBM Power 5 based
12K proc, 100 TFlop/s

DOE ASCI
Red Storm Sandia National Lab

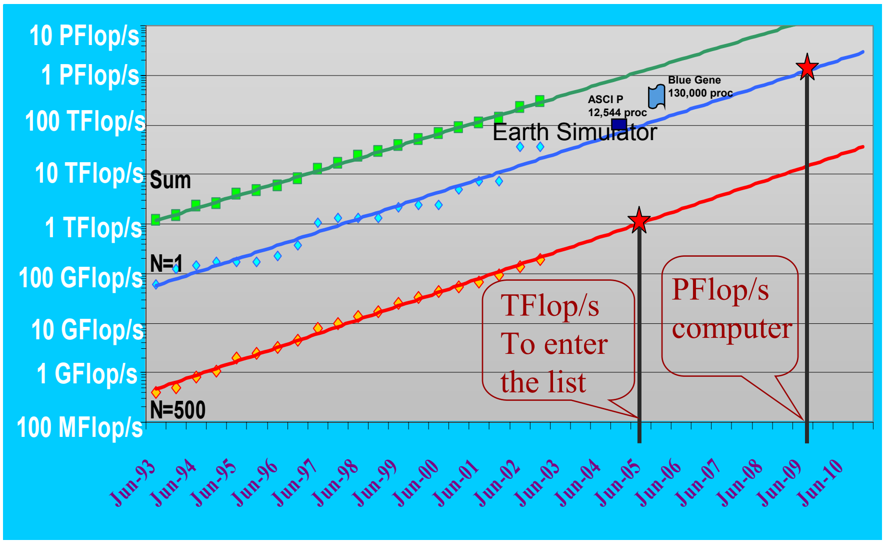
- 10,368 compute processors, 108 cabinets
  - AMD Opteron @ 2.0 GHz
  - Cray integrator and providing the interconnect
- Fully connected high performance 3-D mesh interconnect.
  - Topology - 27 X 16 X 24
- Peak of ~ 40 TF
  - Expected MP-Linpack >20 TF
- Aggregate system memory bandwidth - ~55 TB/s
- MPI Latency - 2 ms neighbor, 5 ms across machine
- Bi-Section bandwidth ~2.3 TB/s
- Link bandwidth ~3.0 GB/s in each direction

2004 in operation

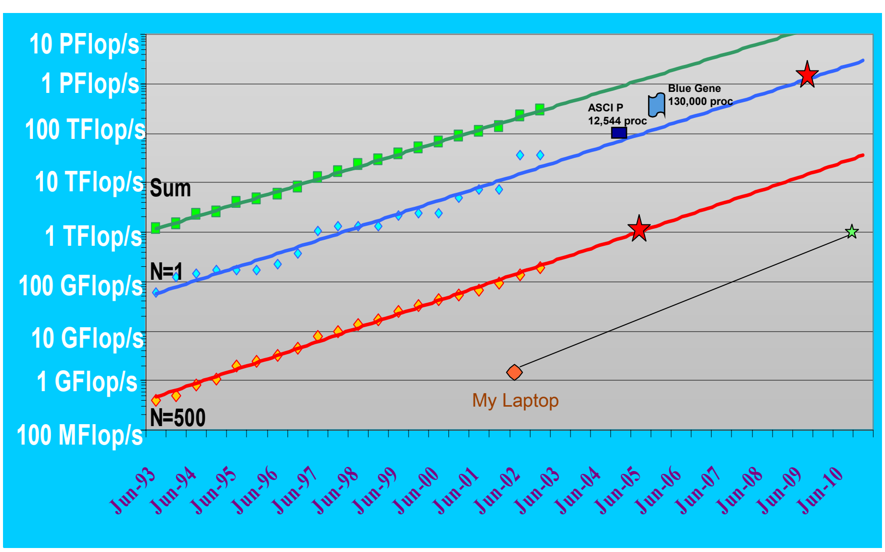
TOP500 - Performance



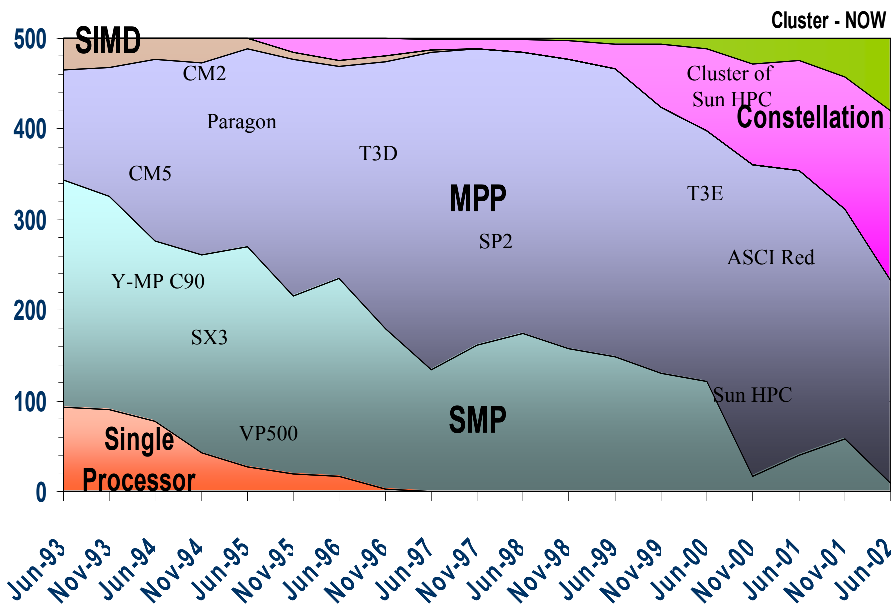
Performance Extrapolation



Performance Extrapolation



Architectures



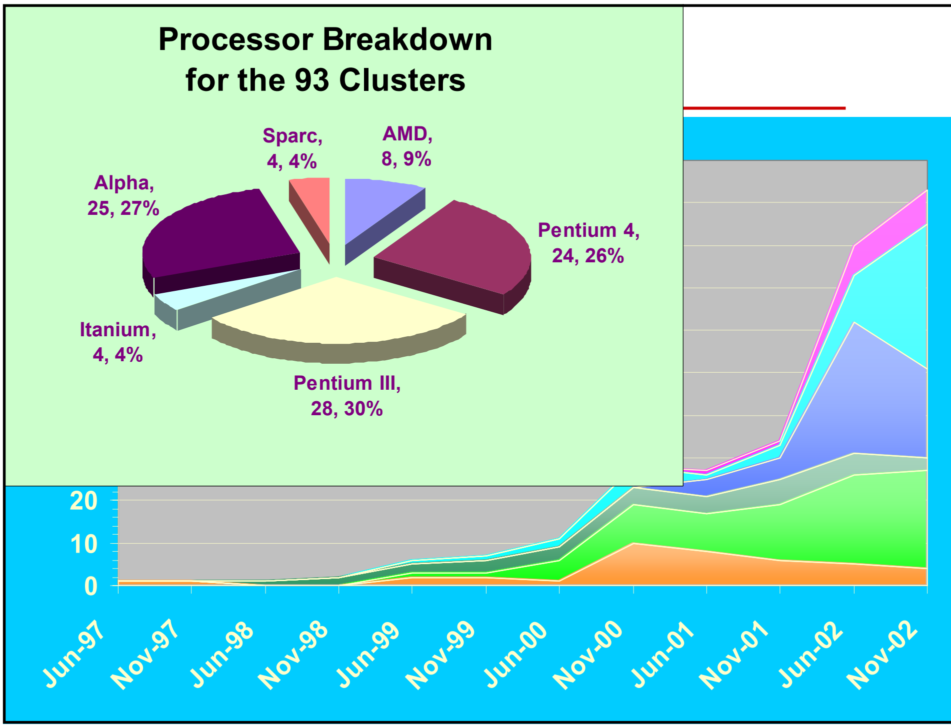
7

Constellation: # of p/n ≥ n

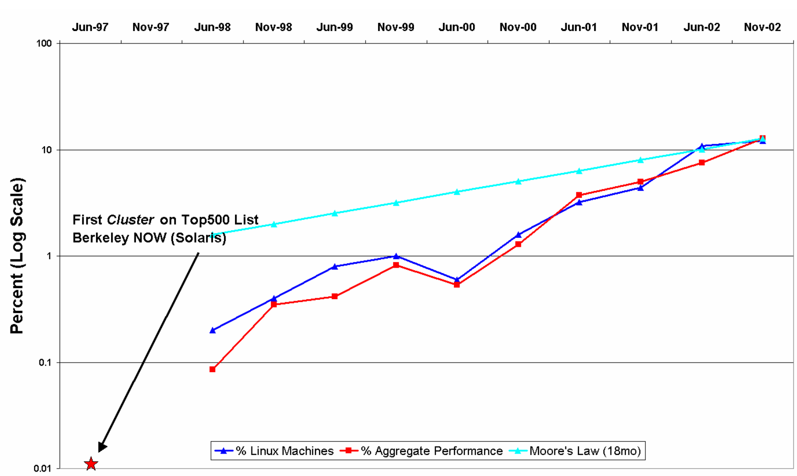
93 Clusters on the Top500

- A total of 56 Intel based and 8 AMD based PC clusters are in the TOP500.
  - 31 of these Intel based cluster are IBM Netfinity systems delivered by IBM.
- A substantial part of these are installed at industrial customers especially in the oil-industry.
  - Including 5 Sun and 5 Alpha based clusters and 21 HP AlphaServer.
- 15 of these clusters are labeled as 'Self-Made'.

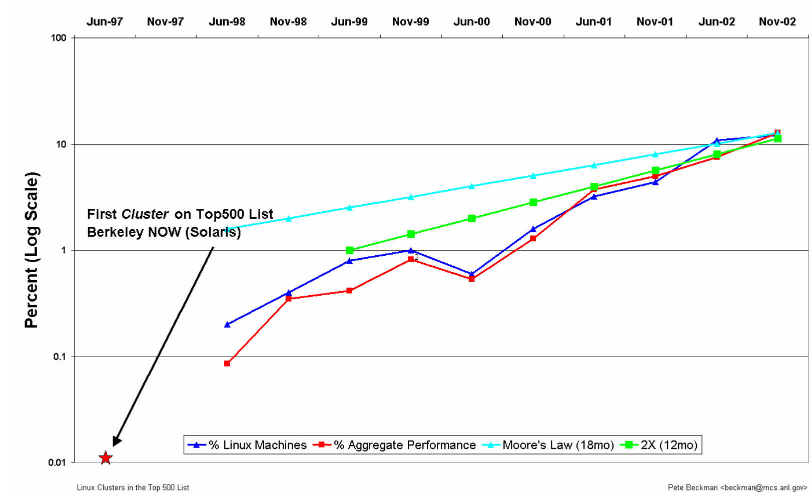
18



Linux: Plotting The Future

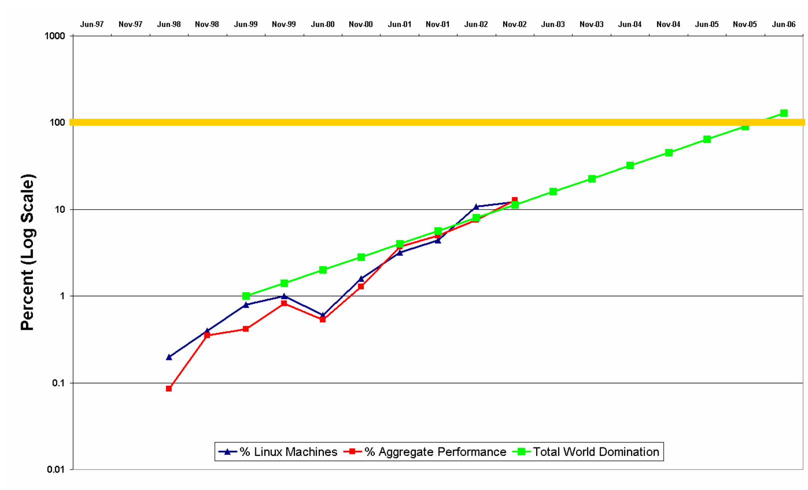


Linux: Plotting The Future

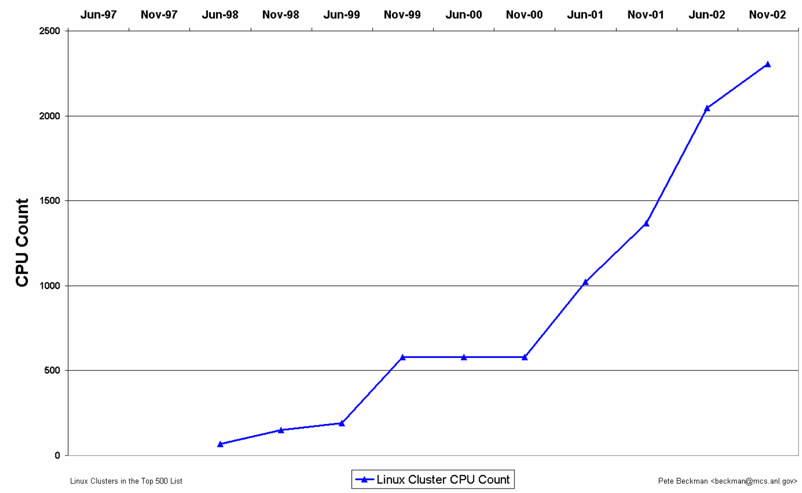


Predicting Future Market Share

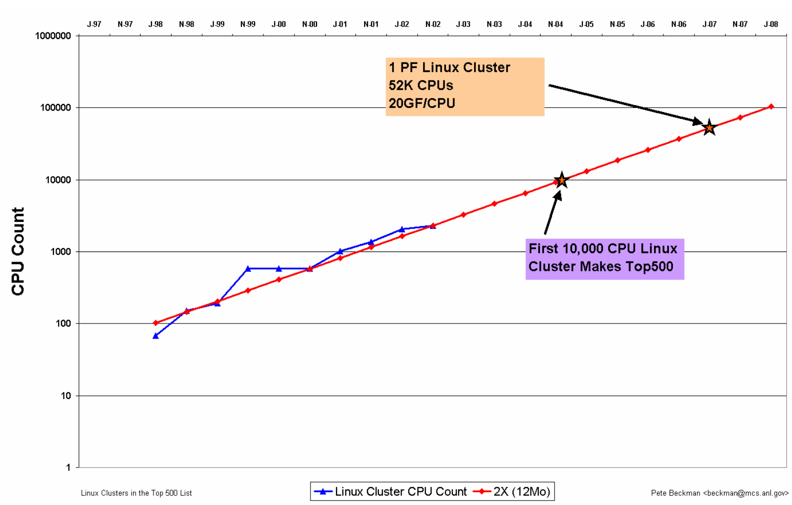
How Long Until Total World Domination?



How Large Can Linux Clusters Get?



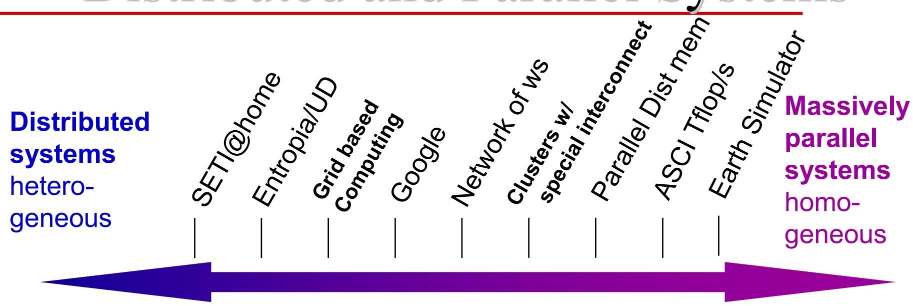
Linux Cluster Sizes: Plotting The Future



Observations

- The adoption rate of Linux HPC is phenomenal!
  - Linux in the Top500 is doubling every 12 months
  - Linux adoption is not driven by bottom feeders
    - *Adoption is actually faster at the ultra-scale!*
- The CPU counts for the largest Linux clusters are currently doubling every year
- Prediction: by 2005, we will have a 10,000 CPU Linux cluster
- Prediction: by 2005, most top-performing supercomputers will be running Linux
- Adoption rate driven largely by economics and human factors

Distributed and Parallel Systems

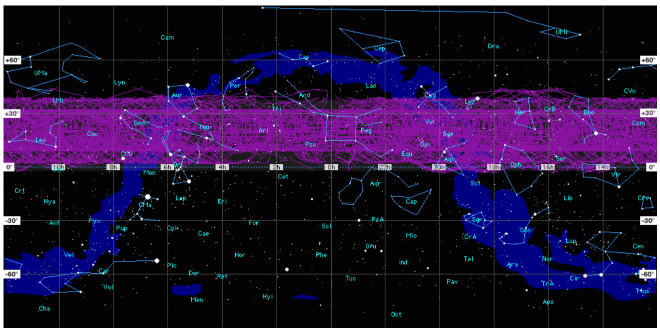


- Gather (unused) resources
- Steal cycles
- System SW manages resources
- System SW adds value
- 10% - 20% overhead is OK
- Resources drive applications
- Time to completion is not critical
- Time-shared
- SETI@home
  - ~ 500,000 machines
  - Averaging 55 Tflop/s

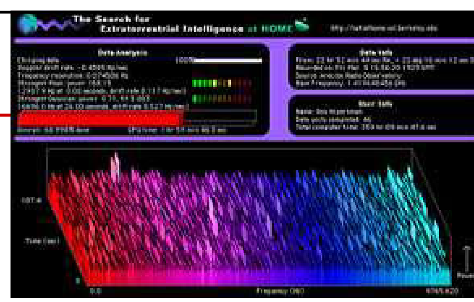
- Bounded set of resources
- Apps grow to consume all cycles
- Application manages resources
- System SW gets in the way
- 5% overhead is maximum
- Apps drive purchase of equipment
- Real-time constraints
- Space-shared
- Earth Simulator
  - 5000 processors
  - Averaging 35 Tflop/s

SETI@home: Global Distributed Computing

- Running on 500,000 PCs, ~1300 CPU Years per Day
  - 1.3M CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope

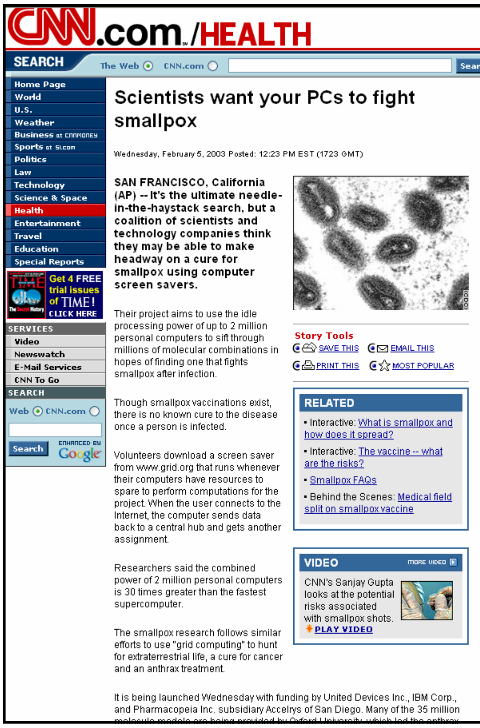


SETI@home



- Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
- When their computer is idle or being wasted this software will download ~ half a MB chunk of data for analysis. Performs about 3 Tflops for each client in 15 hours.
- The results of this analysis are sent back to the SETI team, combined with thousands of other participants.

- Largest distributed computation project in existence
  - Averaging 55 Tflop/s
- Today a number of companies trying this for profit.



CNN.com /HEALTH

SEARCH The Web CNN.com

Home Page
World
U.S.
Weather
Business at CNNMoney
Sports at SI.com
Politics
Law
Technology
Science & Space
Health
Entertainment
Travel
Education
Special Reports

Get 4 FREE trial issues of TIME! CLICK HERE

SERVICES
Video
Newswatch
E-Mail Services
CNN To Go

SEARCH
Web CNN.com
Search
Enhanced by Google

Scientists want your PCs to fight smallpox

Wednesday, February 5, 2003 Posted: 12:23 PM EST (1723 GMT)

SAN FRANCISCO, California (AP) -- It's the ultimate needle-in-the-haystack search, but a coalition of scientists and technology companies think they may be able to make headway on a cure for smallpox using computer screen savers.

Their project aims to use the idle processing power of up to 2 million personal computers to sift through millions of molecular combinations in hopes of finding one that fights smallpox after infection.

Though smallpox vaccinations exist, there is no known cure to the disease once a person is infected.

Volunteers download a screen saver from www.grid.org that runs whenever their computers have resources to spare to perform computations for the project. When the user connects to the Internet, the computer sends data back to a central hub and gets another assignment.

Researchers said the combined power of 2 million personal computers is 30 times greater than the fastest supercomputer.

The smallpox research follows similar efforts to use "grid computing" to hunt for extraterrestrial life, a cure for cancer and an anthrax treatment.

It is being launched Wednesday with funding by United Devices Inc., IBM Corp., and Pharmacopeia Inc. subsidiary Accelrys of San Diego. Many of the 35 million

(CDC)

Story Tools
SAVE THIS | EMAIL THIS
PRINT THIS | MOST POPULAR

RELATED
- Interactive: What is smallpox and how does it spread?
- Interactive: The vaccine -- what are the risks?
- Smallpox FAQs
- Behind the Scenes: Medical field split on smallpox vaccine

VIDEO — MORE VIDEO
CNN's Sanjay Gupta looks at the potential risks associated with smallpox shots.
PLAY VIDEO

Grid Computing - from ET to Smallpox

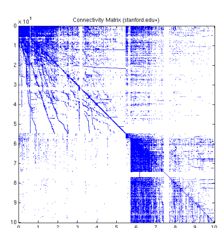
The project employs computational chemistry to analyze chemical interactions between a library of 35 million potential drug molecules and several protein targets on the smallpox virus in the search for an effective anti-viral drug to treat smallpox post-infection.

UNITED DEVICES™

ICL UT

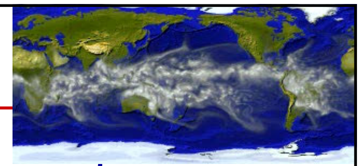
Google™

- **Google query attributes**
  - 150M queries/day (2000/second)
  - 100 countries
  - 3B documents in the index
- **Data centers**
  - 15,000 Linux systems in 6 data centers
    - 15 TFlop/s and 1000 TB total capability
    - 40-80 1U/2U servers/cabinet
    - 100 MB Ethernet switches/cabinet with gigabit Ethernet uplink
  - growth from 4,000 systems (June 2000)
    - 18M queries then
- **Performance and operation**
  - simple reissue of failed commands to new servers
  - no performance debugging



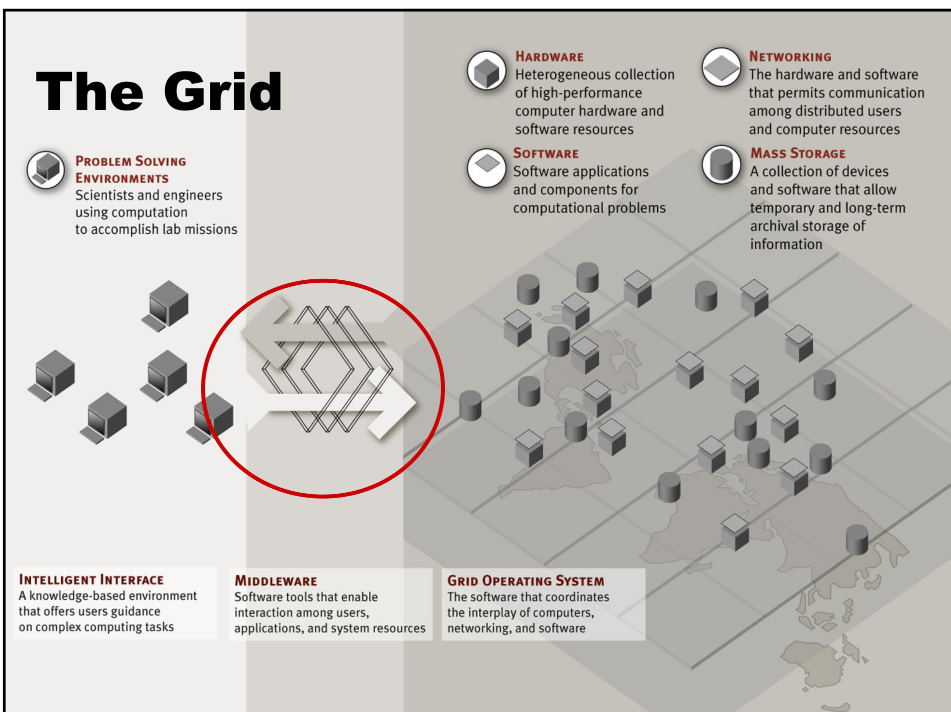
Source: Monika Henzinger, Google

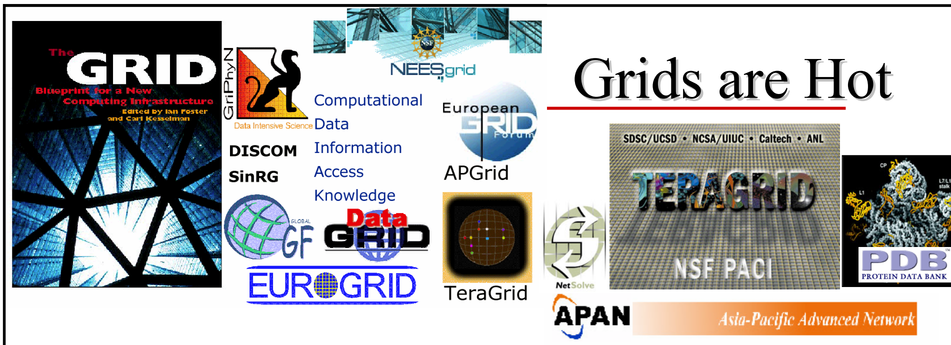
In the past: Isolation
Motivation for Grid Computing



- Today there is a complex interplay and increasing interdependence among the sciences.
- Many science and engineering problems require widely dispersed resources be operated as systems.
- What we do as collaborative infrastructure developers will have profound influence on the future of science.
- Networking, distributed computing, and parallel computation research have matured to make it possible for distributed systems to support high-performance applications, but...
  - Resources are dispersed
  - Connectivity is variable
  - Dedicated access may not be possible

Today: Collaboration



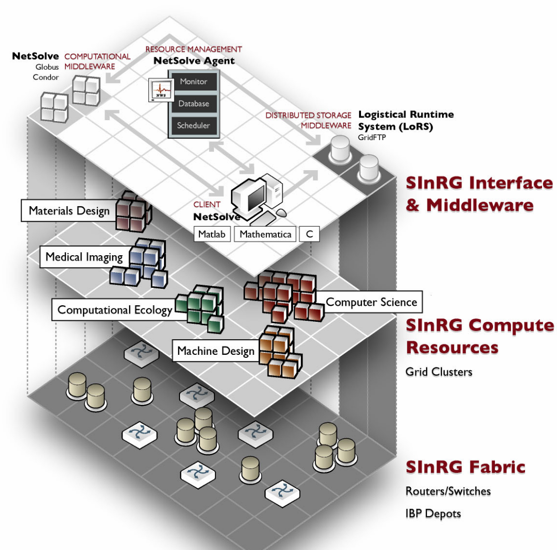


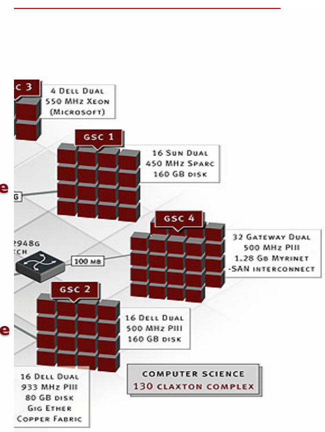
Grids are Hot

IPG NASA	http://nas.nasa.gov/~wej/home/IPG
Globus	http://www.globus.org/
Legion	http://www.cs.virgina.edu/~grimshaw/
AppLeS	http://www-cse.ucsd.edu/groups/hpcl/
NetSolve	http://www.cs.utk.edu/netsolve/
NINF	http://phase.etl.go.jp/ninf/
Condor	http://www.cs.wisc.edu/condor/
CUMULVS	http://www.epm.ornl.gov/cs/
WebFlow	http://www.npac.syr.edu/users/gcf/
NGC	http://www.nordicgrid.net

ICL UT

University of Tennessee Deployment: Scalable Intracampus Research Grid: SInRG





- Federated Ownership: CS, Chem Eng., Medical School, Computational Ecology, El. Eng.
- Real applications, middleware development, logistical networking

Grids vs. Capability Computing

- Not an "either/or" question
  - Each addresses different needs
  - Both are part of an integrated solution
- Grid strengths
  - Coupling necessarily distributed resources
    - instruments, software, hardware, archives, and people
  - Eliminating time and space barriers
    - remote resource access and capacity computing
  - Grids are not a cheap substitute for capability HPC
- Capability computing strengths
  - Supporting foundational computations
    - terascale and petascale "nation scale" problems
  - Engaging tightly coupled teams and computations

Futures for Numerical Algorithms and Software

- Numerical software will be adaptive, exploratory, and intelligent
- Determinism in numerical computing will be gone.
  - After all, its not reasonable to ask for exactness in numerical computations.
  - Auditability of the computation, reproducibility at a cost
- Fault Tolerance
  - Google claims 15K nodes, what do they do when one goes down?
  - We must do better than "restart ALL nodes from last chkpt"
- Importance of floating point arithmetic will be undiminished.
  - 16, 32, 64, 128 bits and beyond.
- Reproducibility, fault tolerance, and auditability
- Adaptivity is a key so applications can effectively use the resources.

Collaborators / Support

- TOP500
  - H. Mauer, Mannheim U
  - H. Simon, NERSC
  - E. Strohmaier, NERSC

- Thanks

Next Generation Software

SciDAC
Scientific Discovery through Advanced Computing