ROYAL IRISH ACADEMY
Acadamh Ríoga na hÉireann
Ireland's Academy for the Sciences & Humanities

Supercomputers and Clusters and Grid, Oh My!

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory
and
SFI Walton Visitor
University College Dublin

Apologies to Frank Baum…

Dorothy: "Do you suppose we'll meet any wild animals?"

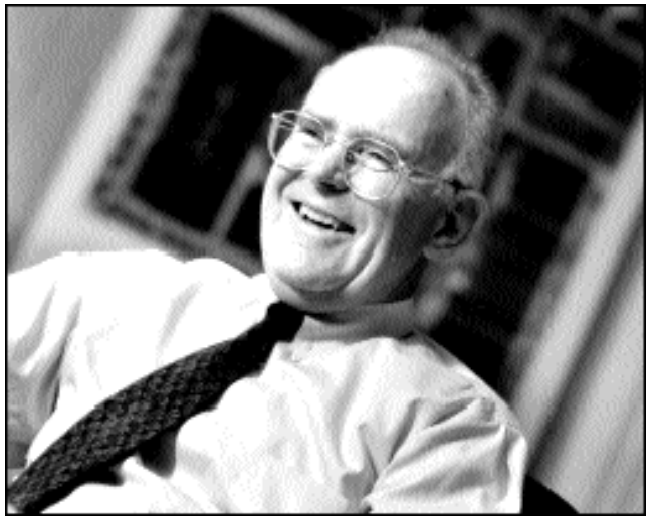
Tinman: "We might."

Scarecrow: "Animals that ... that eat straw?"

Tinman: "Some. But mostly lions, and tigers, and bears."

All: Supercomputers and clusters and grids, oh my!
Supercomputers and clusters and grids, oh my!

Technology Trends: Microprocessor Capacity



Gordon Moore (co-founder of Intel) Electronics Magazine, 1965

Number of devices/chip doubles every 18 months

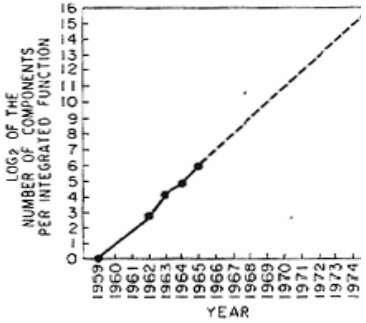
2X transistors/Chip Every 1.5 years

Called "Moore's Law"

The experts look ahead

Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip



By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new areas.

Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today.

But the biggest potential lies in the production of large systems. In telephone communications, integrated circuits in digital filters will separate channels on multiplex equipment. Integrated circuits will also switch telephone circuits and perform data processing.

Computers will be more powerful, and will be organized in completely different ways. For example, memories built of integrated electronics may be distributed throughout the machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

Present and future

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as well as any additional ones that result in electronics functions supplied to the user as irreducible units. These technologies were first investigated in the late 1950's. The object was to miniaturize electronics equipment to include increasingly complex electronic functions in limited space with minimum weight. Several approaches evolved, including microassembly techniques for individual components, thin-film structures and semiconductor integrated circuits.

Each approach evolved rapidly and converged so that each borrowed techniques from another. Many researchers believe the way of the future to be a combination of the various approaches.

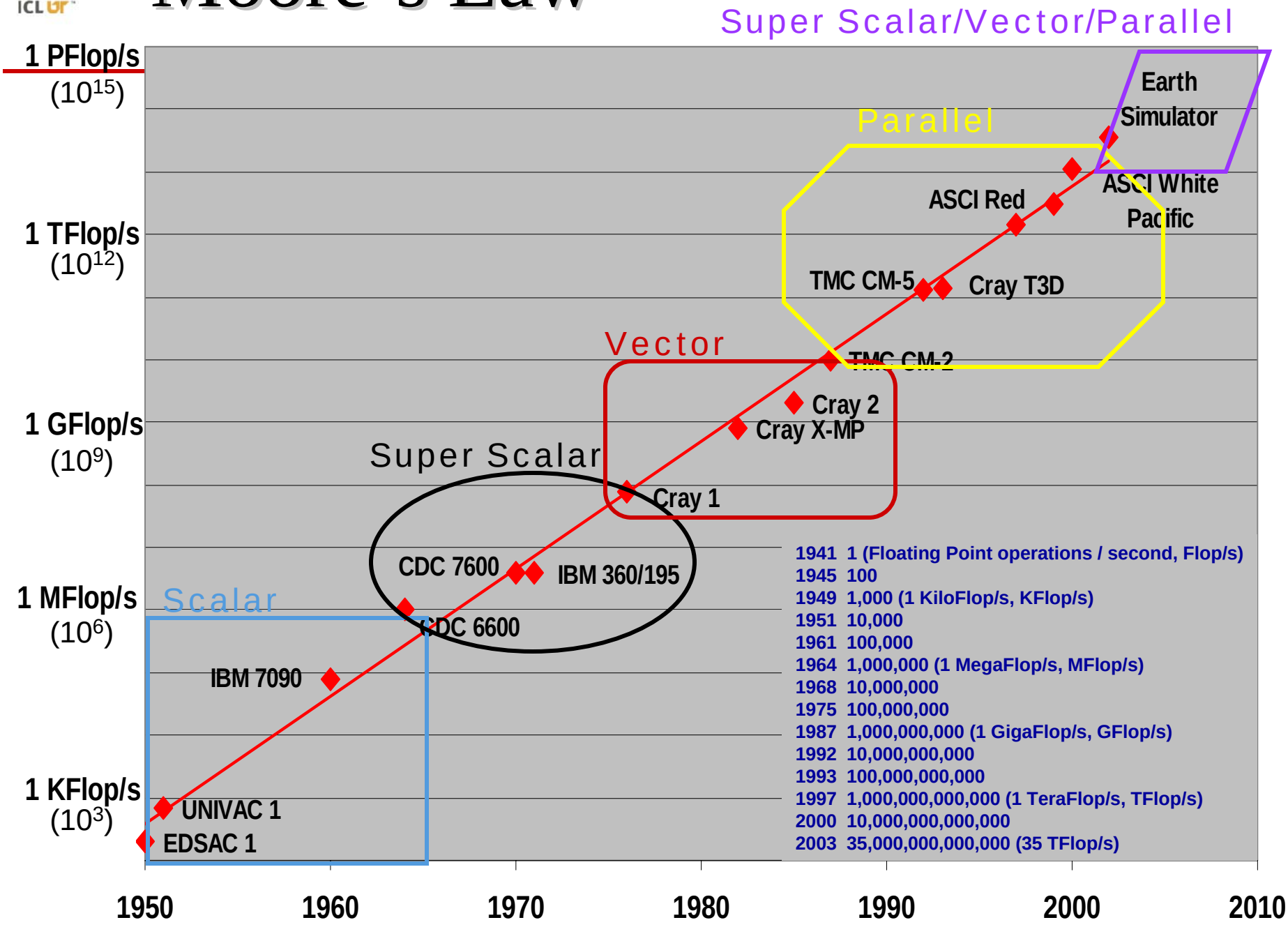
The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to an active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of active semiconductor devices to the passive film arrays.

Both approaches have worked well and are being used

The author

Dr. Gordon E. Moore is one of the new breed of electronic engineers, schooled in the physical sciences rather than in electronics. He earned a B.S. degree in chemistry from the University of California and a Ph.D. degree in physical chemistry from the California Institute of Technology. He was one of the founders of Fairchild Semiconductor and has been

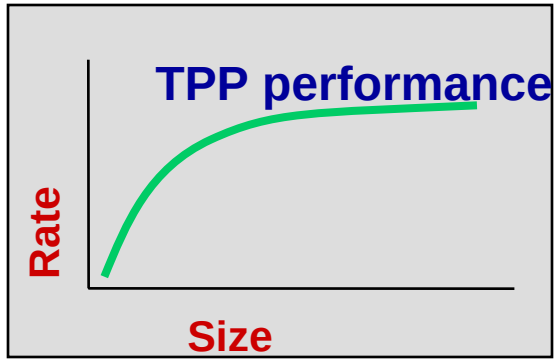
Moore's Law


Super Scalar/Vector/Parallel
1 PFlop/s
(10^15)
1 TFlop/s
(10^12)
1 GFlop/s
(10^9)
1 MFlop/s
(10^6)
1 KFlop/s
(10^3)
Parallel
Vector
Super Scalar
Scalar
Earth Simulator
ASCI White Pacific
ASCI Red
TMC CM-5
Cray T3D
TMC CM-2
Cray 2
Cray X-MP
Cray 1
CDC 7600
IBM 360/195
CDC 6600
IBM 7090
UNIVAC 1
EDSAC 1
1941 1 (Floating Point operations / second, Flop/s)
1945 100
1949 1,000 (1 KiloFlop/s, KFlop/s)
1951 10,000
1961 100,000
1964 1,000,000 (1 MegaFlop/s, MFlop/s)
1968 10,000,000
1975 100,000,000
1987 1,000,000,000 (1 GigaFlop/s, GFlop/s)
1992 10,000,000,000
1993 100,000,000,000
1997 1,000,000,000,000 (1 TeraFlop/s, TFlop/s)
2000 10,000,000,000,000
2003 35,000,000,000,000 (35 TFlop/s)
1950
1960
1970
1980
1990
2000
2010

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

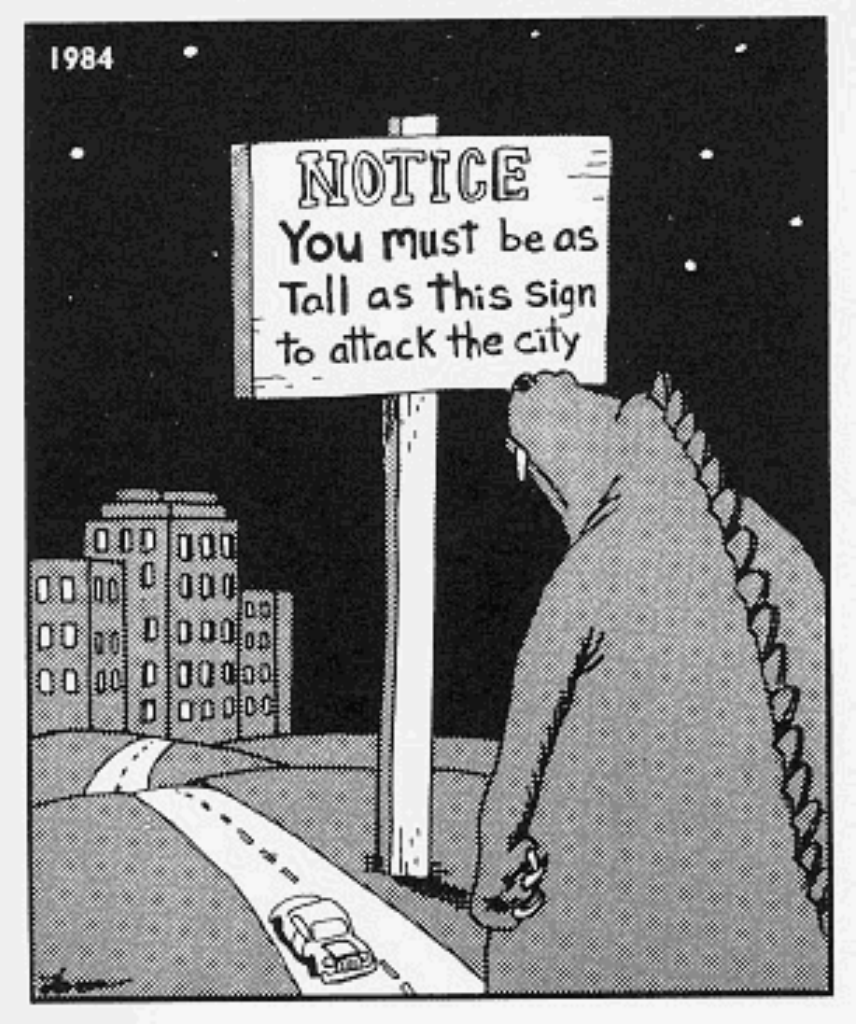
  Meeting in Mannheim, Germany in June

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

What is a Supercomputer?

- A supercomputer is a hardware and software system that provides close to the maximum performance that can currently be achieved.

- Over the last 10 years the range for the Top500 has increased greater than Moore's Law
- 1993:
  - #1 = 59.7 GFlop/s
  - #500 = 422 MFlop/s
- 2003:
  - #1 = 35.8 TFlop/s
  - #500 = 403 GFlop/s

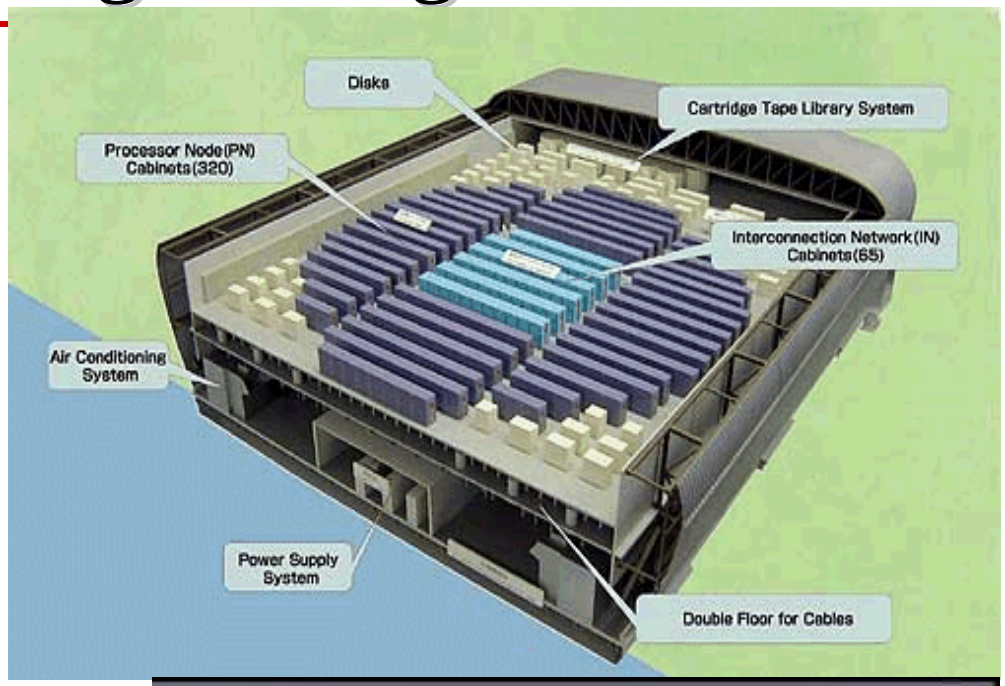


Why do we need them? Computational fluid dynamics, protein folding, climate modeling, national security, in particular for cryptanalysis and for simulating nuclear weapons to name a few.

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)
- ~ 1/2 Billion € for machine, software, & 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

- From the Top500 (November 2003)
  - Performance of ESC > Σ Next Top 3 Computers

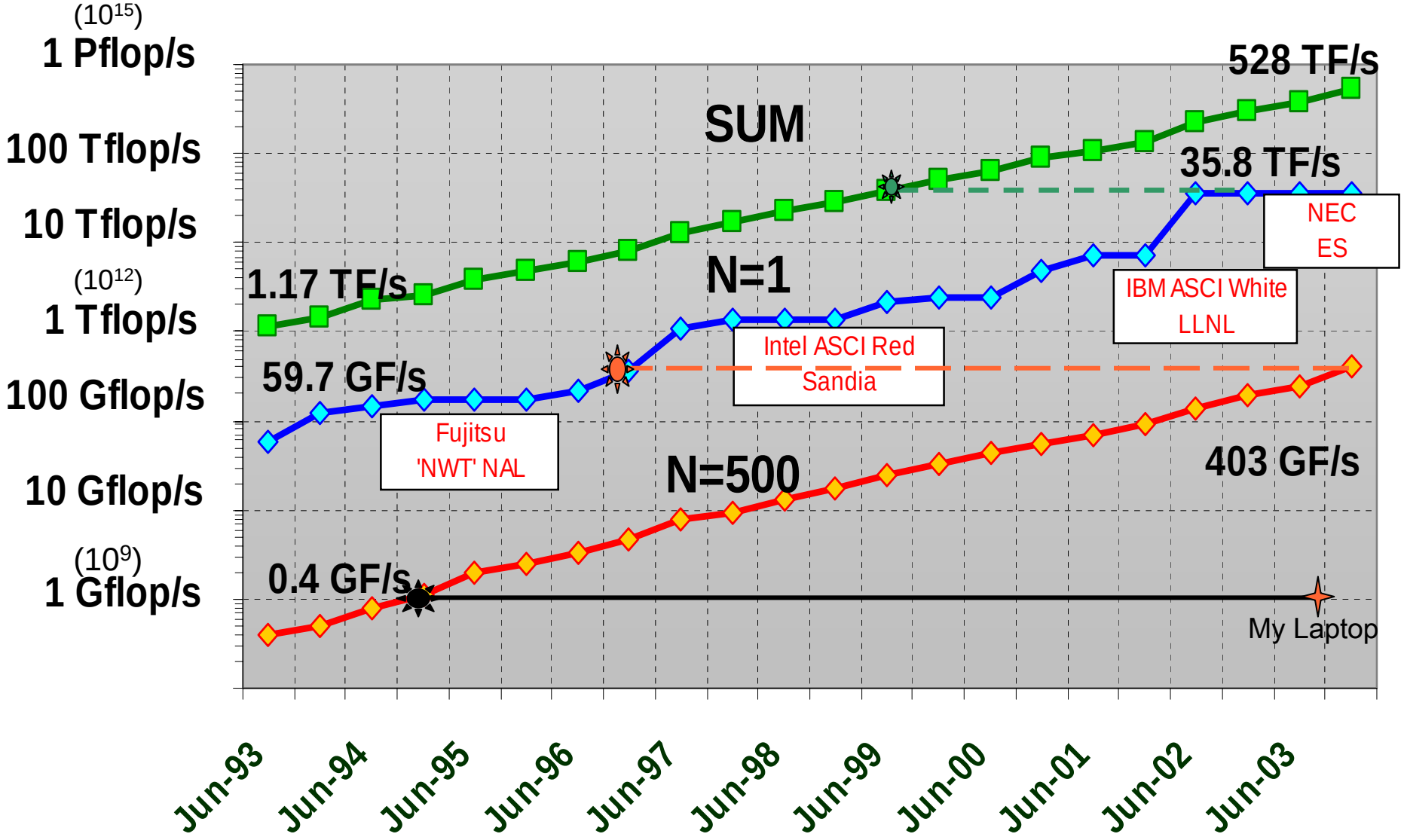


November 2003

	Manufacturer	Computer	Rmax Tflop/s	Installation Site	Year	# Proc	Rpeak Tflop/s
1	NEC	Earth-Simulator	35.8	Earth Simulator Center Yokohama	2002	5120	40.90
2	Hewlett-Packard	ASCI Q - AlphaServer SC ES45/1.25 GHz	13.9	Los Alamos National Laboratory Los Alamos	2002	8192	20.48
3	Self	Apple G5 Power PC w/Infiniband 4X	10.3	Virginia Tech Blacksburg, VA	2003	2200	17.60
4	Dell	PowerEdge 1750 P4 Xeon 3.6 Ghz w/Myrinet	9.82	University of Illinois U/C Urbana/Champaign	2003	2500	15.30
5	Hewlett-Packard	rx2600 Itanium2 1 GHz Cluster – w/Quadrics	8.63	Pacific Northwest National Laboratory Richland	2003	1936	11.62
6	Linux NetworX	Opteron 2 GHz, w/Myrinet	8.05	Lawrence Livermore National Laboratory Livermore	2003	2816	11.26
7	Linux NetworX	MCR Linux Cluster Xeon 2.4 GHz – w/Quadrics	7.63	Lawrence Livermore National Laboratory Livermore	2002	2304	11.06
8	IBM	ASCI White, Sp Power3 375 MHz	7.30	Lawrence Livermore National Laboratory Livermore	2000	8192	12.29
9	IBM	SP Power3 375 MHz 16 way	7.30	NERSC/LBNL Berkeley	2002	6656	9.984
10	IBM	xSeries Cluster Xeon 2.4 GHz – w/Quadrics	6.59	Lawrence Livermore National Laboratory Livermore	2003	1920	9.216

50% of top500 performance in top 9 machines; 131 system > 1 TFlop/s; 210 machines are clusters, 1 IE Vodophone

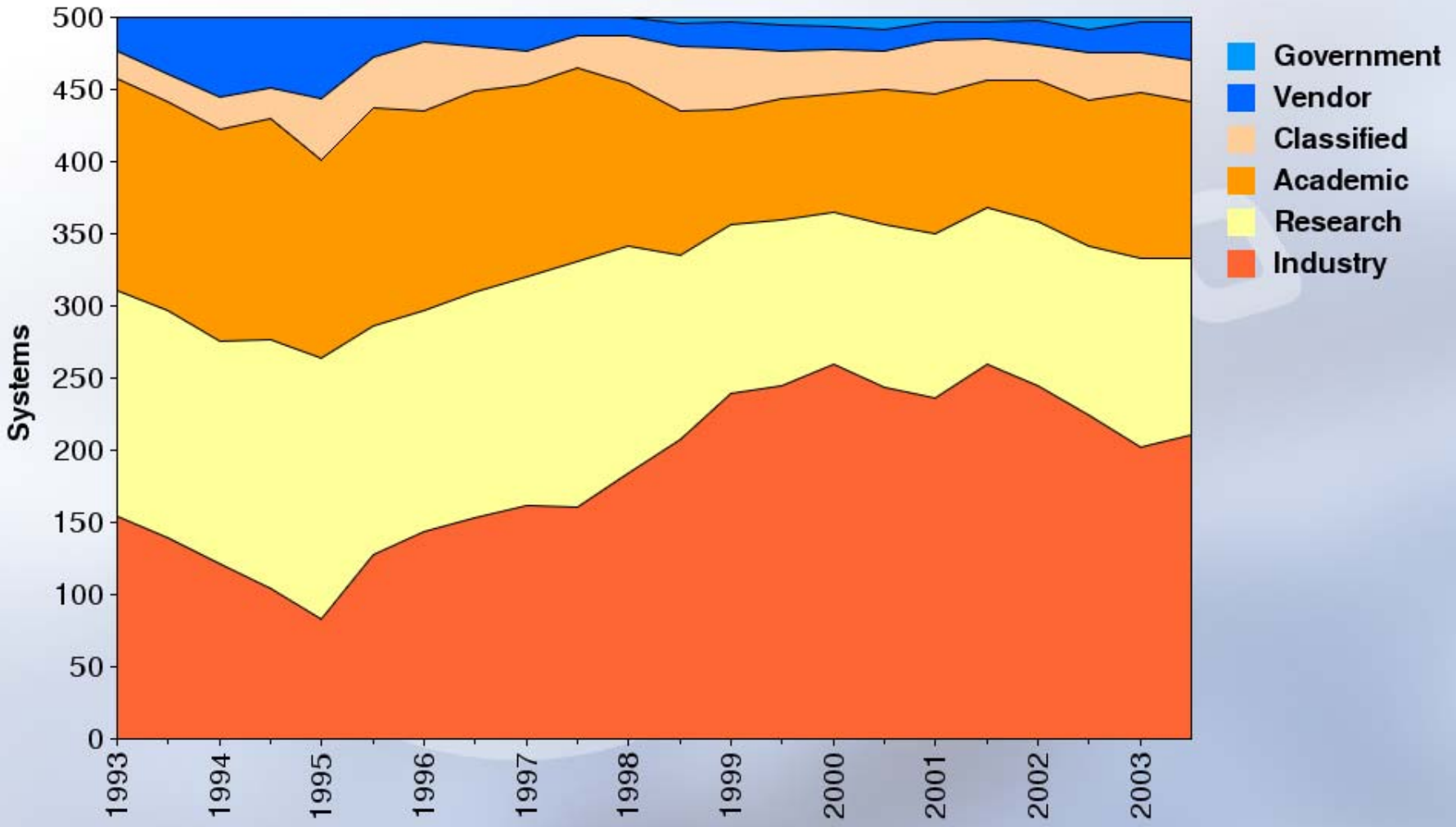
TOP500 – Performance - Nov 2003



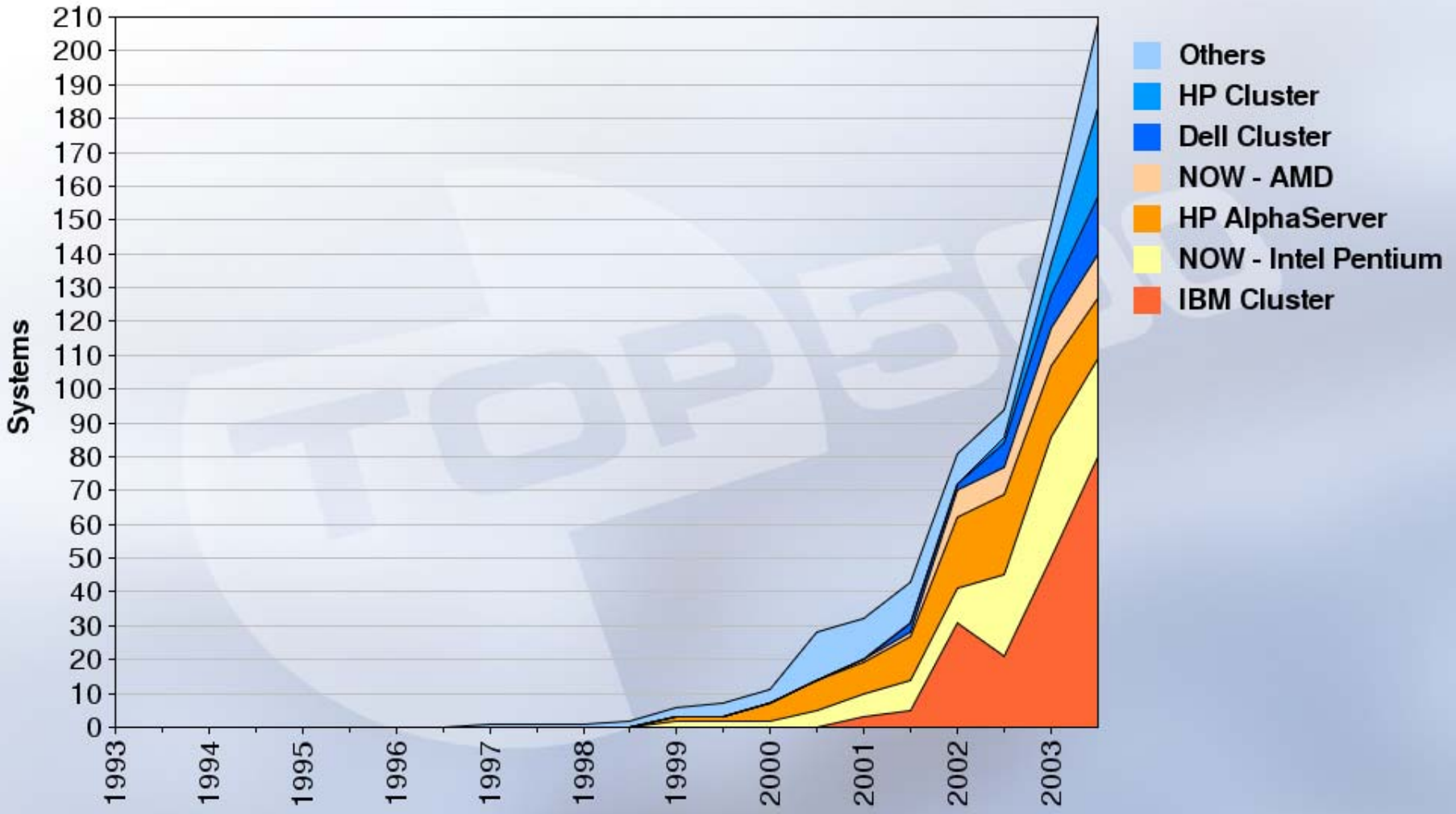
Virginia Tech "Big Mac" G5 Cluster

- **Apple G5 Cluster**
  - Dual 2.0 GHz IBM Power PC 970s
    - 16 Gflop/s per node
      - 2 CPUs * 2 fma units/cpu * 2 GHz * 2(mul-add)/cycle
  - 1100 Nodes or 2200 Processors
    - Theoretical peak 17.6 Tflop/s
  - Infiniband 4X primary fabric
    - Cisco Gigabit Ethernet secondary fabric
  - Linpack Benchmark using 2112 processors
  - Theoretical peak of 16.9 Tflop/s
  - Achieved 10.28 Tflop/s
    - Could be #3 on 11/03 Top500
  - Cost is $5.2 million which includes the system itself, memory, storage, and communication fabrics

Customer Segment / Systems

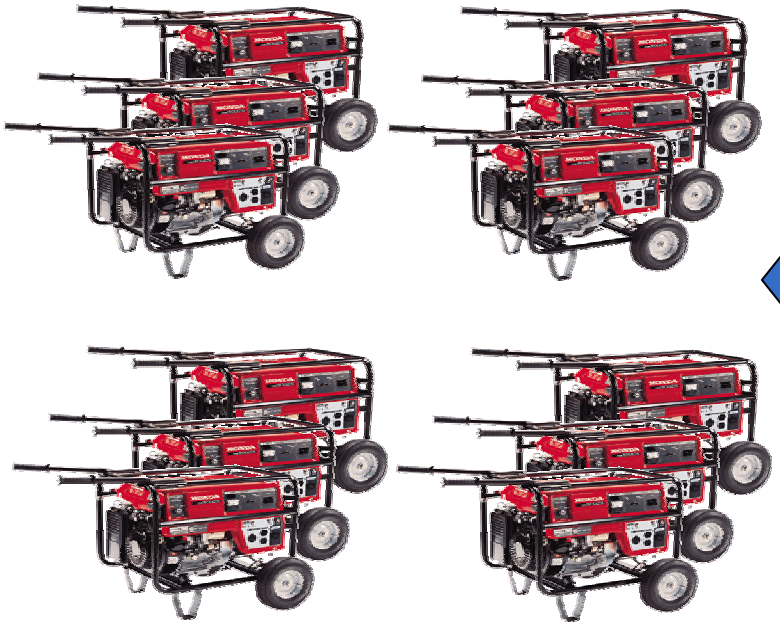


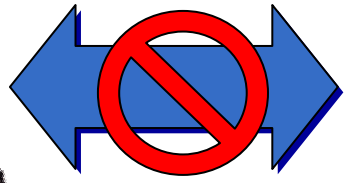
Clusters (NOW) / Systems



A Tool and A Market for Every Task

200K Honda units at 5 KW to equal a 1 GW nuclear plant



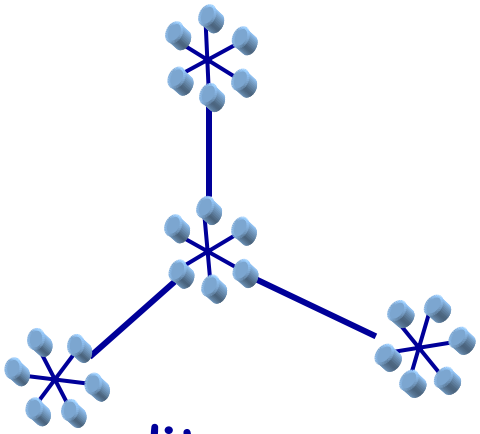


Capability

- **Each targets different applications**
  - understand application needs

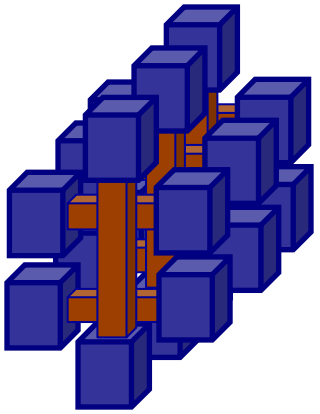
Taxonomy

Cluster Computing



- Commodity processors and switch
- Processors design point for web servers & home pc's
- Leverage millions of processors
- Price point appears attractive for scientific computing

Capability Computing



- Special purpose processors and interconnect
- High Bandwidth, low latency communication
- Designed for scientific computing
- Relatively few machines will be sold
- High price

High Bandwidth vs Commodity Systems

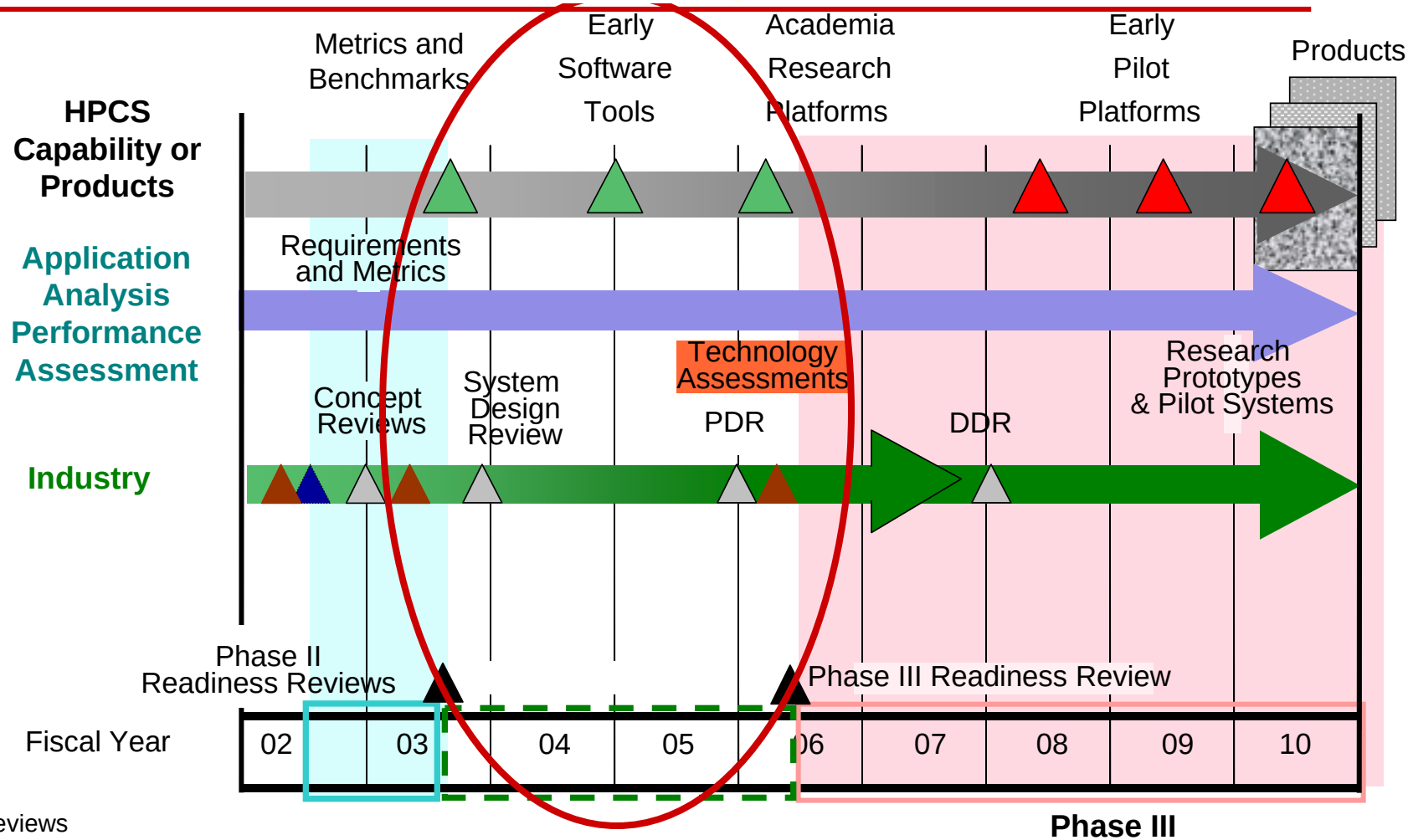
- High bandwidth systems have traditionally been vector computers
  - Designed for scientific problems
  - Capability computing
- Commodity systems are designed for web servers and the home PC market
  - Used for cluster based computers leveraging price point
- Scientific computing needs are different
  - Require a better balance between data movement and floating point operations. Results in greater efficiency.

	Earth Simulator (NEC)	Cray X1 (Cray)	ASCI Q (HP ES45)	MCR (Dual Xeon)	VT Big Mac (Dual IBM PPC)
Year of Introduction	2002	2003	2003	2002	2003
Node Architecture	Vector	Vector	Alpha	Pentium	Power PC
Processor Cycle Time	500 MHz	800 MHz	1.25 GHz	2.4 GHz	2 GHz
Peak Speed per Processor	8 Gflop/s	12.8 Gflop/s	2.5 Gflop/s	4.8 Gflop/s	8 Gflop/s
Bytes/flop to main memory	4	3	1.28	0.9	0.8
Bytes/flop interconnect	1.5	1	0.12	0.07	0.11

Top 5 Machines for the Linpack Benchmark

	Computer (Full Precision)	Number of Procs	*Achieved* TFlop/s	*T Peak* TFlop/s	*Efficiency*
1	Earth Simulator NEC SX-6	5120	35.9	41.0	87.5%
2	LANL ASCI Q AlphaServer EV-68 (1.25 GHz w/Quadrics)	8160	13.9	20.5	67.7%
3	VT Apple G5 dual IBM Power PC (2 GHz, 970s, w/Infiniband 4X)	2112	10.3	16.9	60.9%
4	UIUC Dell Xeon Pentium 4 (3.06 Ghz w/Myrinet)	2500	9.8	15.3	64.1%
5	PNNL HP RX2600 Itanium 2 (1.5GHz w/Quadrics)	1936	8.6	11.6	74.1%

HPCS Phases I - III



Reviews

Industry Procurements

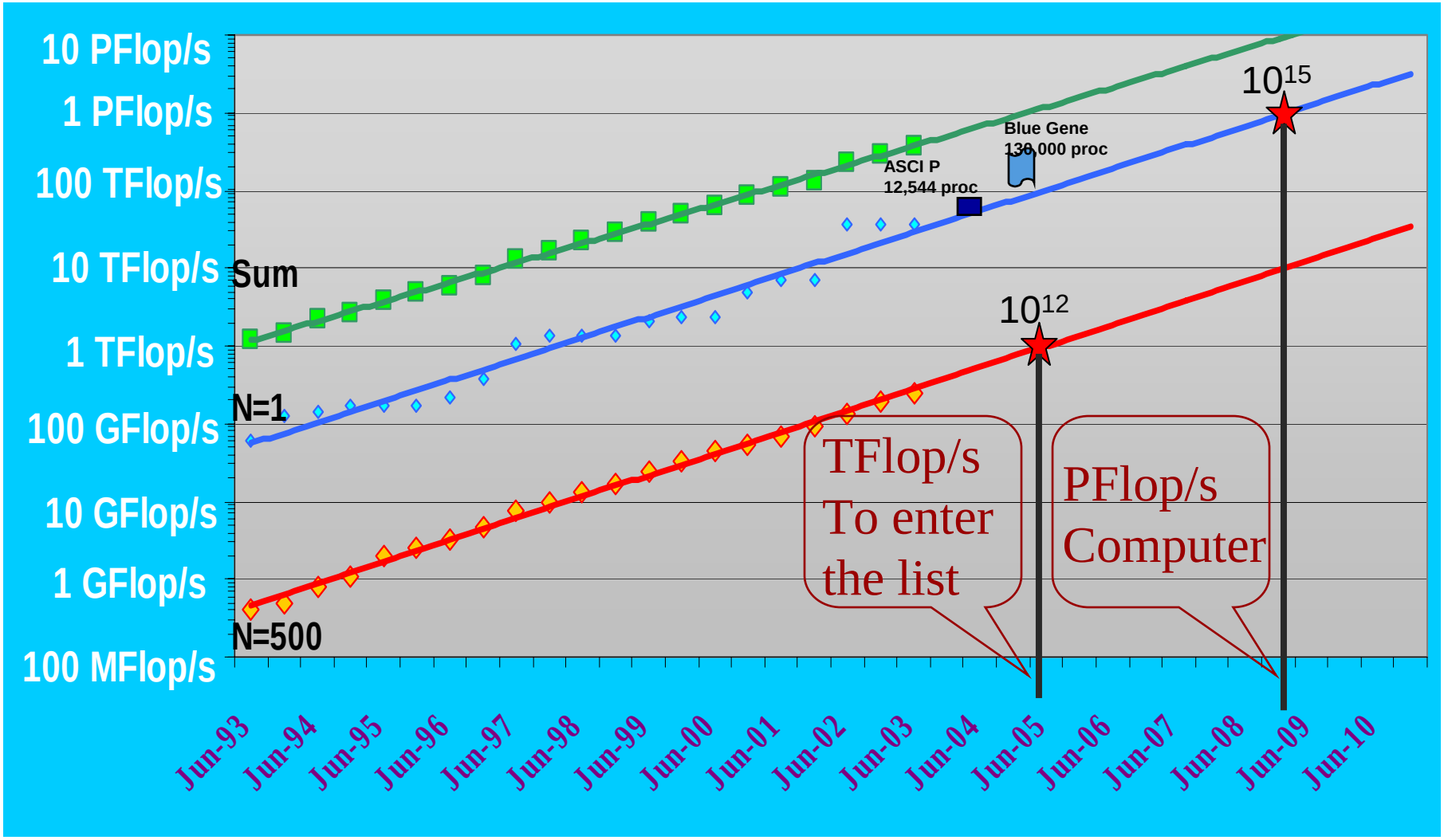
Critical Program Milestones

Phase I
Industry Concept Study
5 companies
$10M each

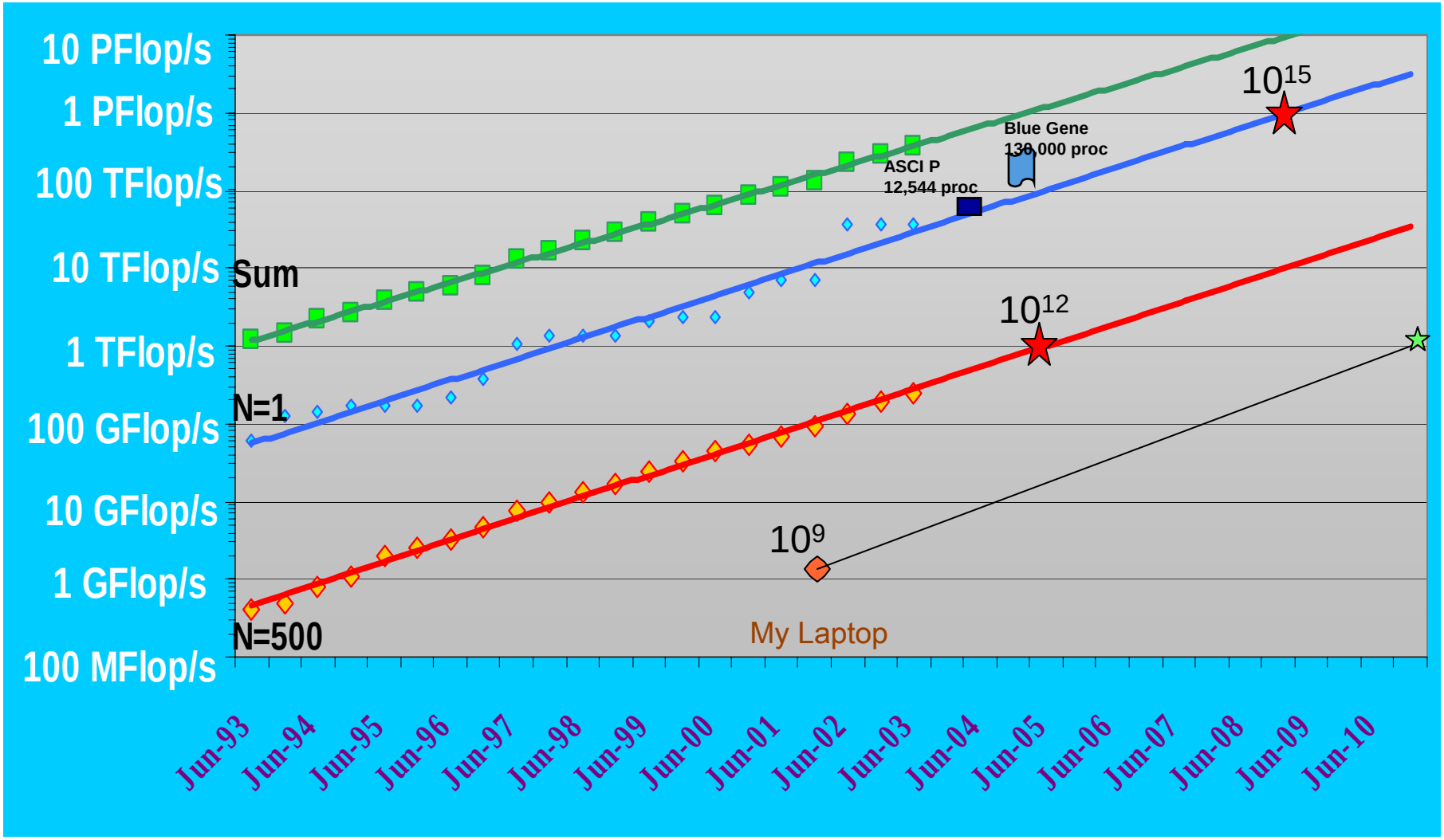
Phase II
R&D
3 companies
~$50M each

Phase III
Full Scale Development
commercially ready in the 2007 to 2010 timeframe.
$100M ?

Performance Extrapolation

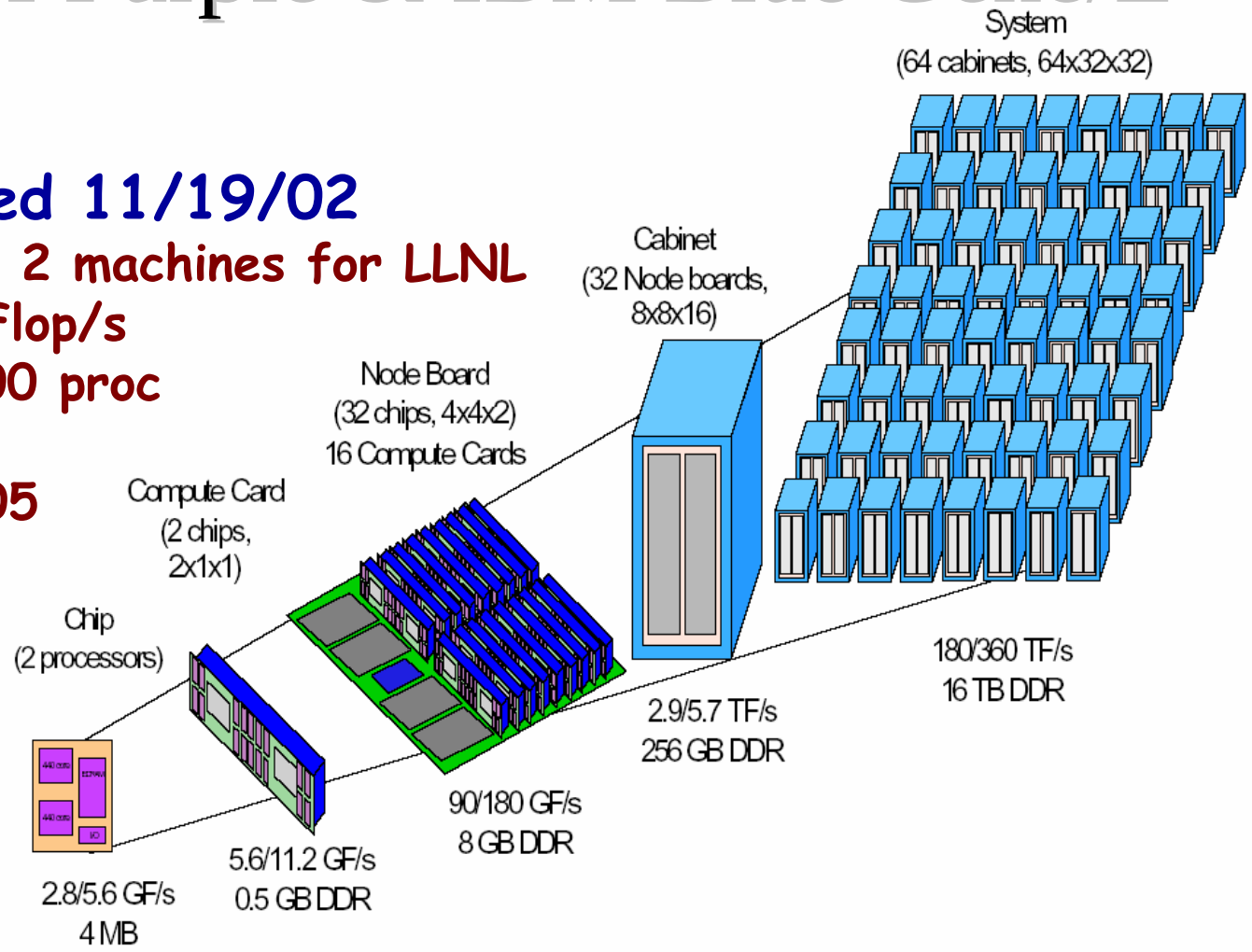


Performance Extrapolation



ASCI Purple & IBM Blue Gene/L

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

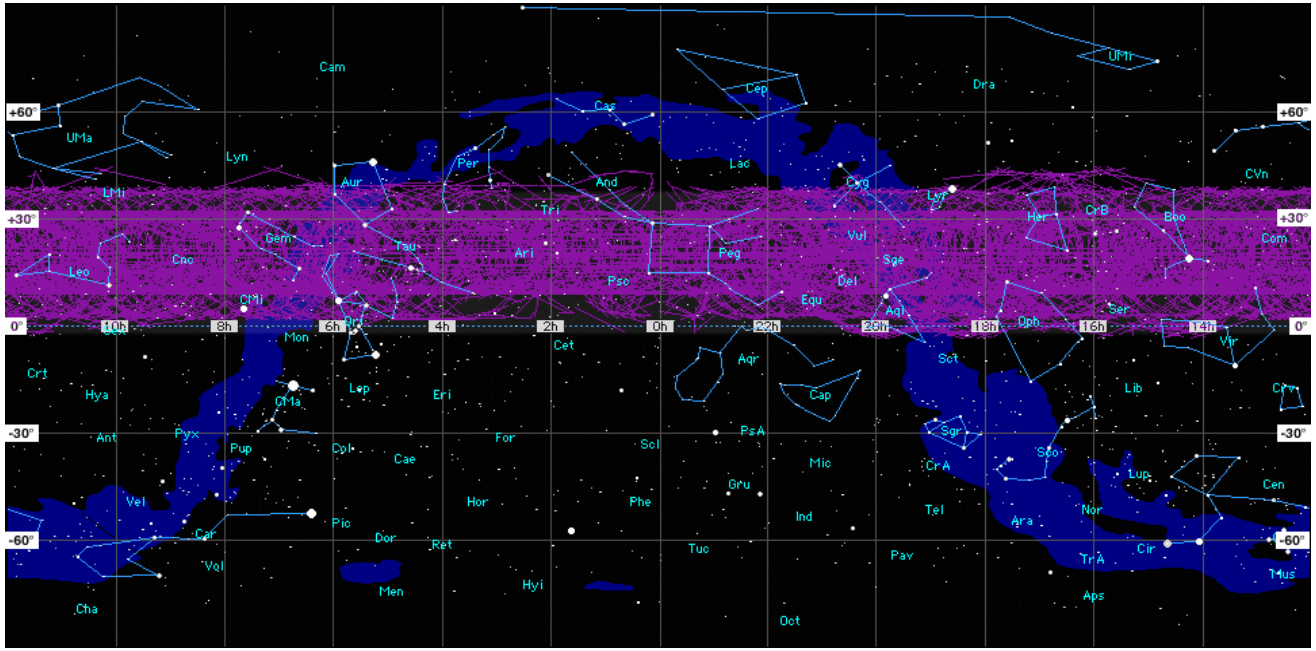


  - Preliminary machine
    - IBM Research BlueGene/L
      - PowerPC 440, 500MHz w/custom proc/interconnect
      - 512 Nodes (1024 processors)
      - 1.435 Tflop/s (2.05 Tflop/s Peak)

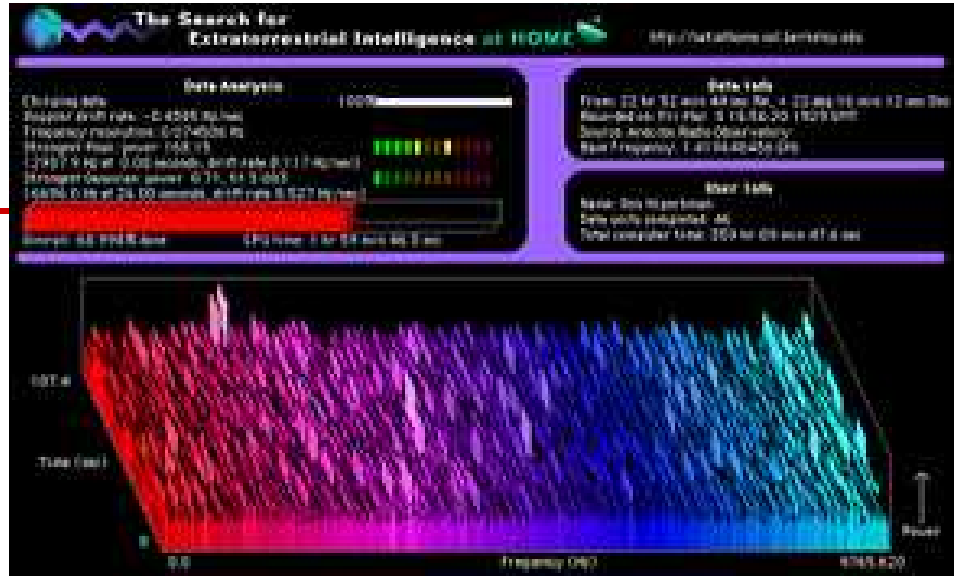
Plus
ASCI Purple
IBM Power 5 based
12K proc, 100 TFlop/s

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

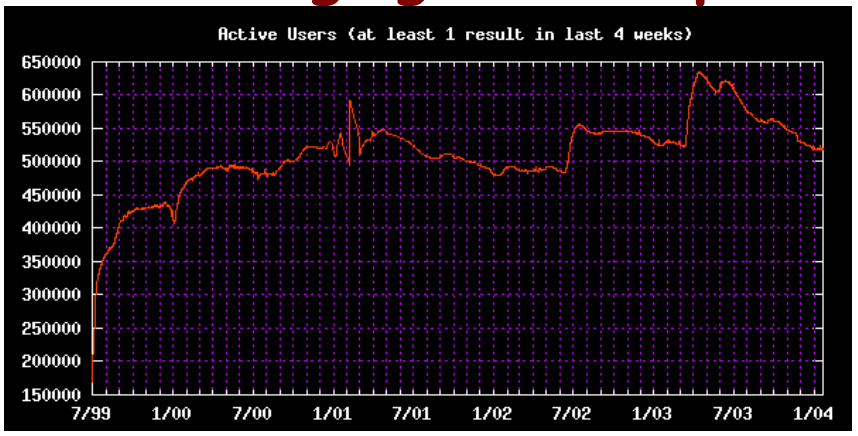


SETI@home



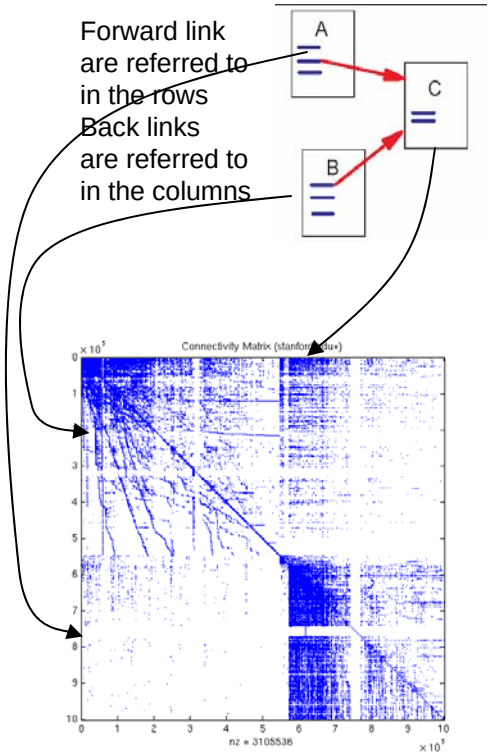
- Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
- When their computer is idle or being wasted this software will download ~ half a MB chunk of data for analysis. Performs about 3 Tflops for each client in 15 hours.
- The results of this analysis are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

- **Largest distributed computation project in existence**
  - **Averaging 72 Tflop/s**



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
    - problems are not reproducible



Eigenvalue problem
$n=3x10^9$
(see: MathWorks Cleve's Corner)

Source: Monika Henzinger, Google & Cleve Moler

Extreme Example: Sony PlayStation2

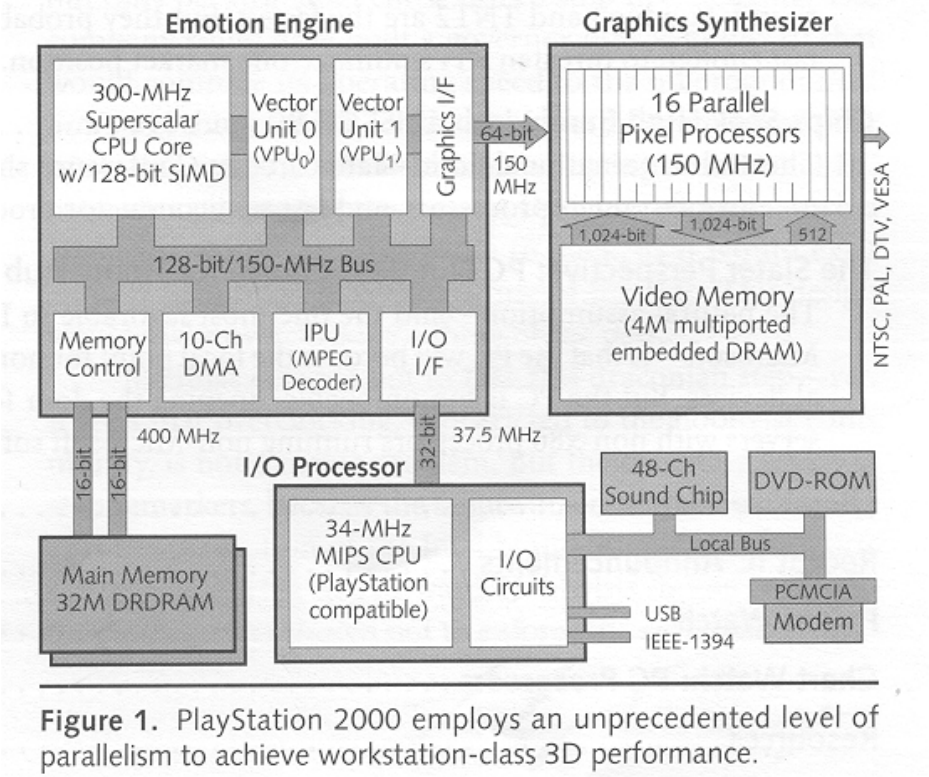
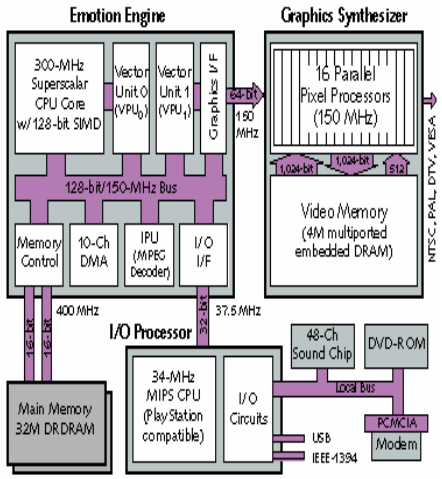


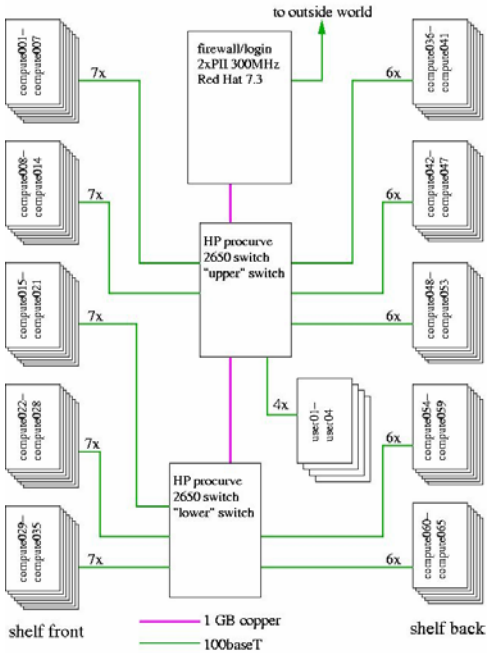
Figure 1. PlayStation 2000 employs an unprecedented level of parallelism to achieve workstation-class 3D performance.

- **Emotion Engine:**
- **6.2 Gflop/s, 75 million polygons per second (Microprocessor Report, 13:5)**
  - **Superscalar MIPS core + vector coprocessor + graphics/DRAM**
  - **About $200**

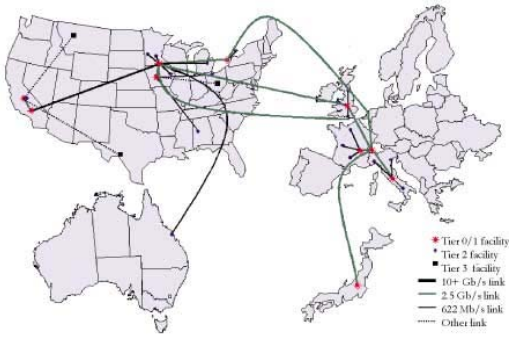
Computing On Toys

- **Sony PlayStation2**
  - 6.2 GF peak
  - 70M polygons/second
  - 10.5M transistors
  - superscalar RISC core
  - plus vector units, each:
    - 19 mul-adds & 1 divide
    - each 7 cycles
- ***$199 retail***
  - *loss leader for game sales*
- **100 unit cluster at U of I**
  - Linux software and vector unit use
    - over 0.5 TF peak
  - but hard to program & hard to extract performance …

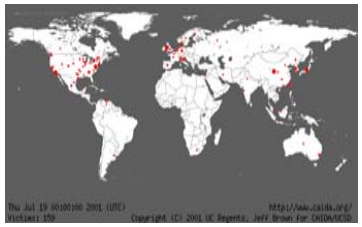


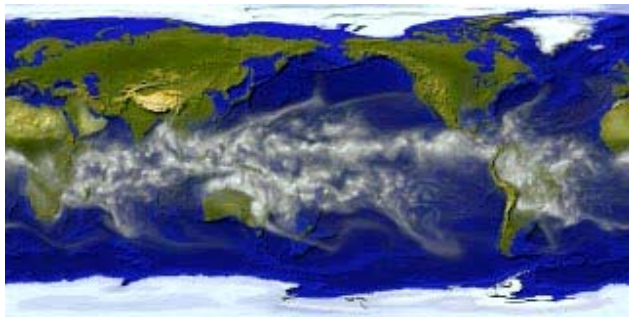


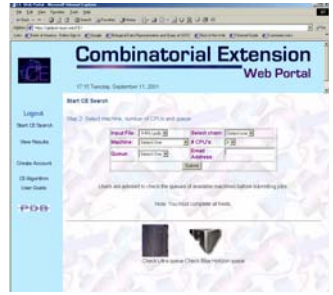
Science and Technology



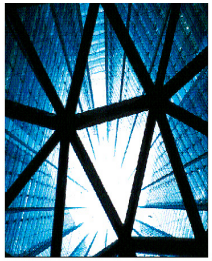
- Today, large science projects are conducted by global teams using sophisticated combinations of
  - *Computers*
  - *Networks*
  - *Visualization*
  - *Data storage*
  - *Remote instruments*
  - *People*
  - *Other resources*





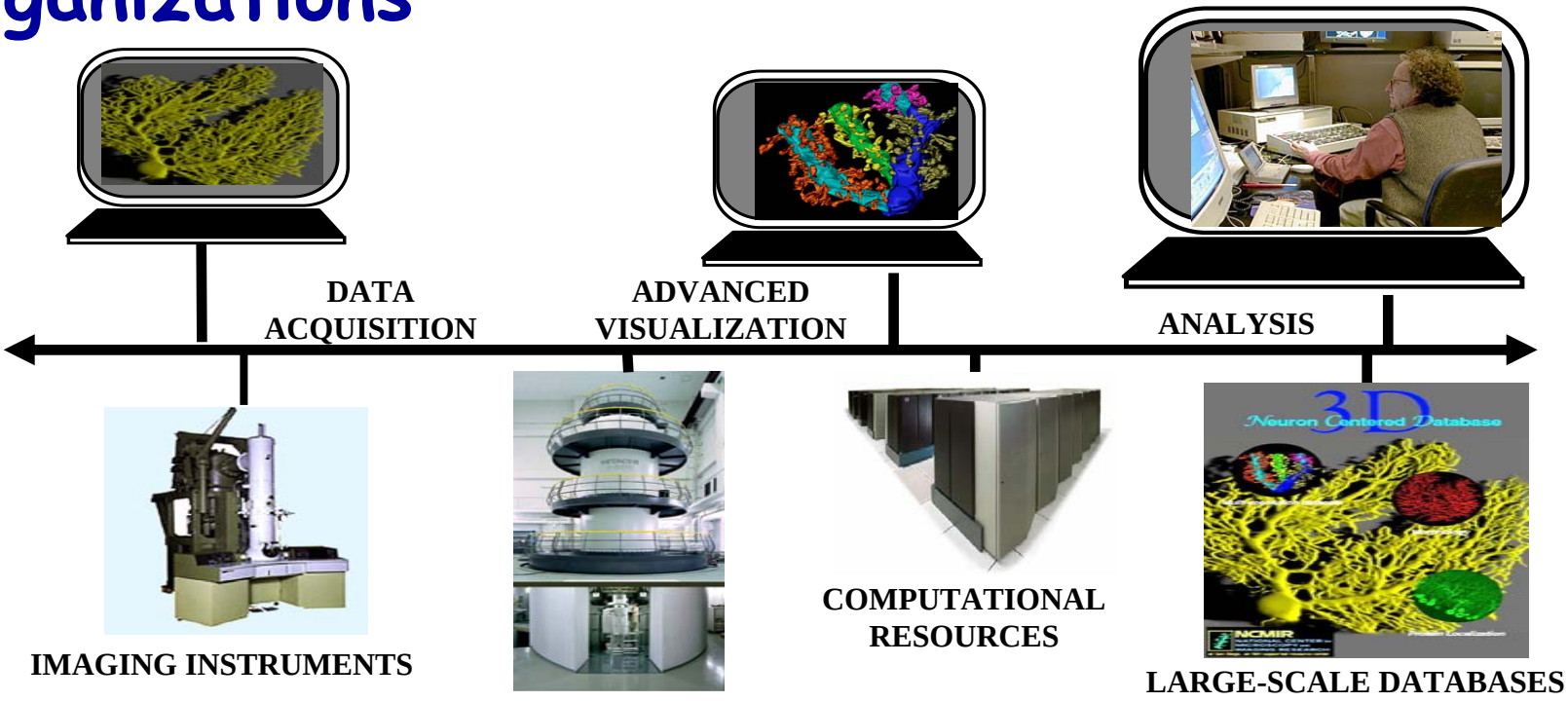


- *Information Infrastructure provides a way to integrate resources to support modern applications*



Grid Computing is About …

Resource sharing & coordinated problem solving in dynamic, multi-institutional virtual organizations



The most pressing scientific challenges require application solutions that are multidisciplinary and multi-scale.

The Grid

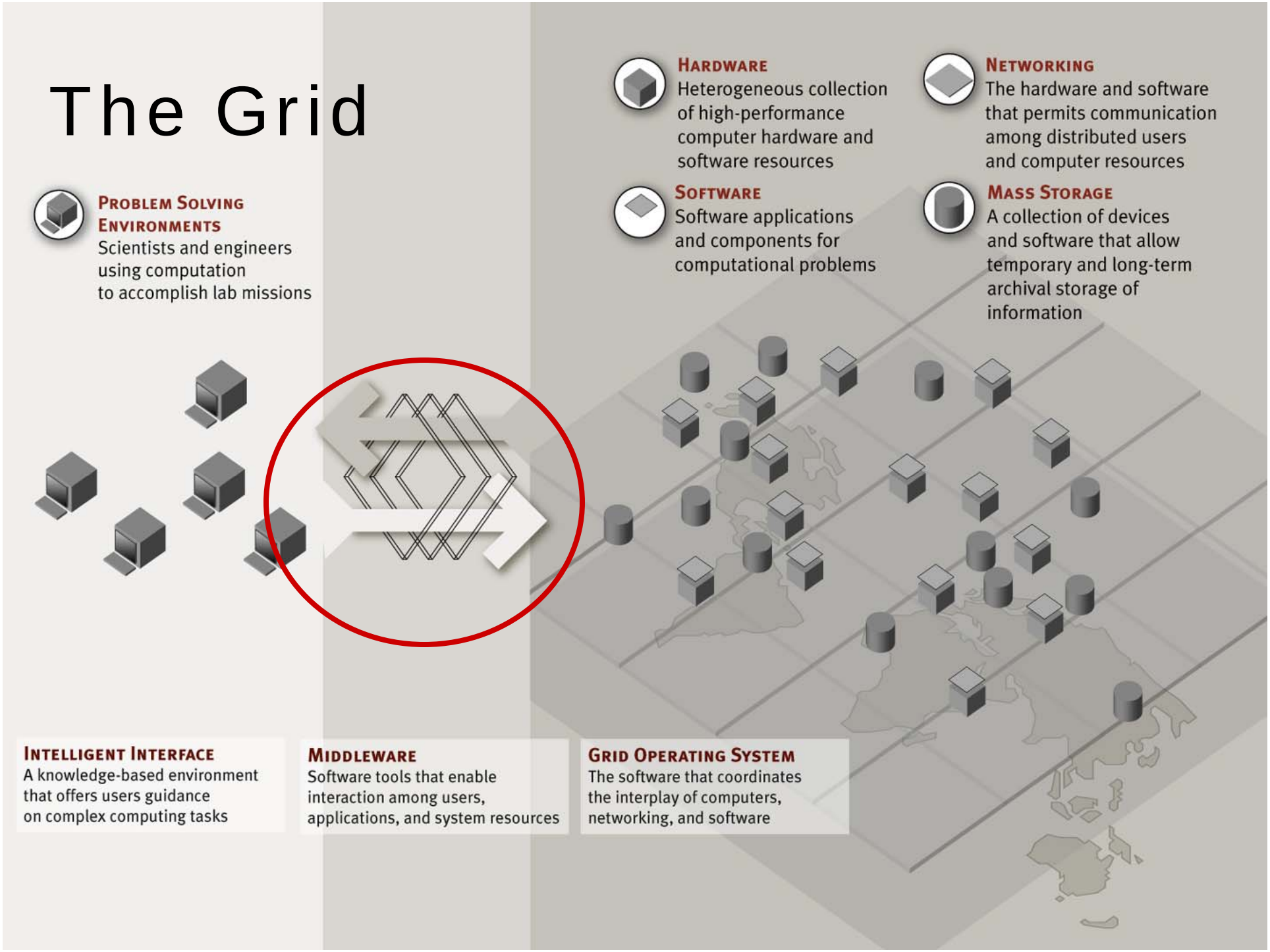
- Motivation: When communication is close to free we should not be restricted to local resources when solving problems.

- Infrastructure that builds on the Internet and the Web
- Enable and exploit large scale sharing of resources
- Virtual organization
  - Loosely coordinated groups
- Provides for remote access of resources
  - Scalable
  - Secure
  - Reliable mechanisms for discovery and access

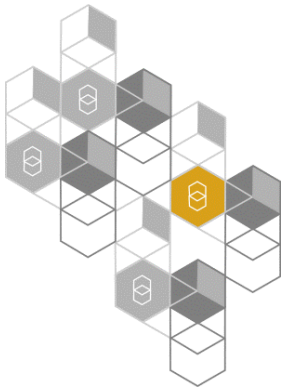
Grid Software Challenges

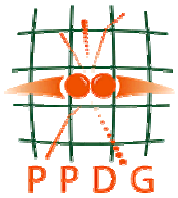
- Simplified programming
  - reduced complexity and coordination
- Accounting and resource economies
  - "non-traditional" resources and concurrency
    - shared resource costs and denial of service
  - negotiation and equilibration
    - exchange rates and sharing
- Scheduling and adaptation
  - performance, fault-tolerance, and access
    - networks, computing, storage, and sensors
- On-demand access
  - unique observational events and sensor fusion
    - "instant" access and nimble scheduling
- Managing bandwidth and latency
  - lambda dominance and exploitation

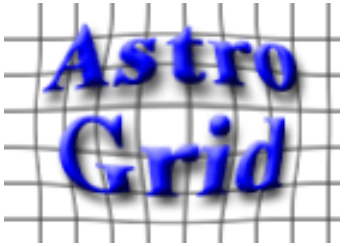
The Grid

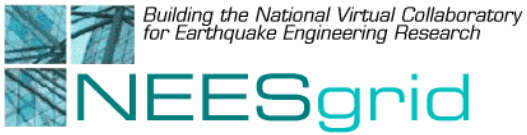

Problem Solving Environments
Scientists and engineers using computation to accomplish lab missions
Hardware
Heterogeneous collection of high-performance computer hardware and software resources
Networking
The hardware and software that permits communication among distributed users and computer resources
Software
Software applications and components for computational problems
Mass Storage
A collection of devices and software that allow temporary and long-term archival storage of information
Intelligent Interface
A knowledge-based environment that offers users guidance on complex computing tasks
Middleware
Software tools that enable interaction among users, applications, and system resources
Grid Operating System
The software that coordinates the interplay of computers, networking, and software

Science Grid Projects



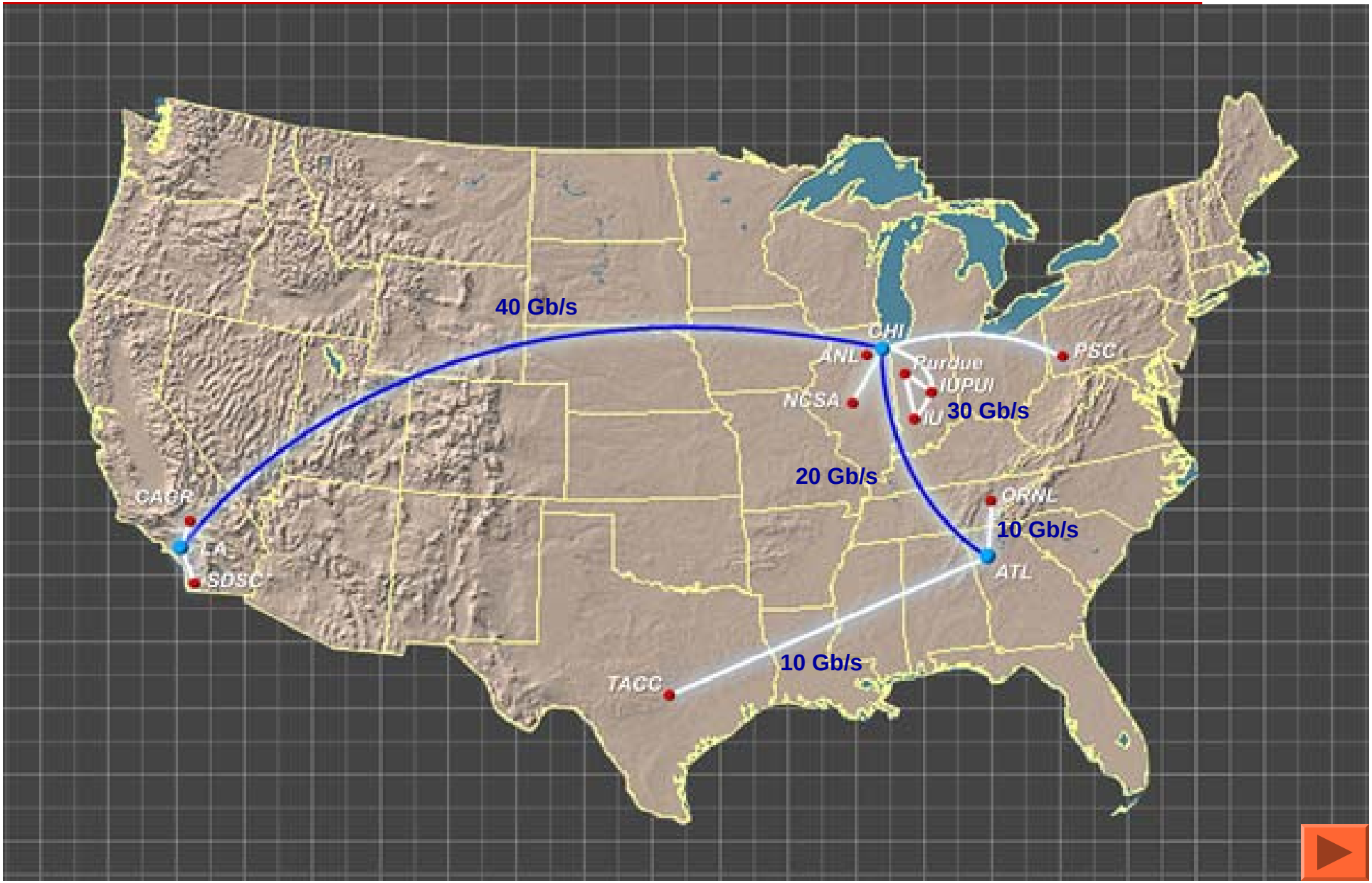




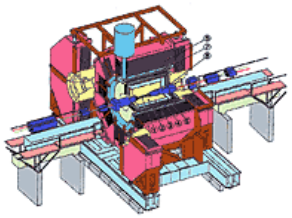


TeraGrid 2003

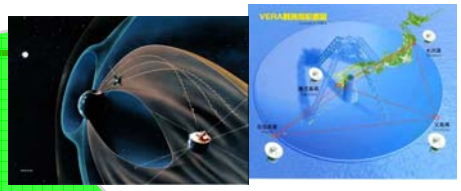
Prototype for a National Cyberinfrastructure



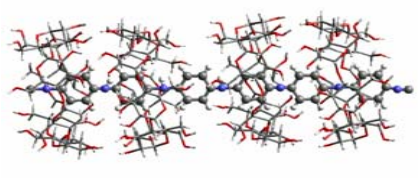
SuperSINET and Applications



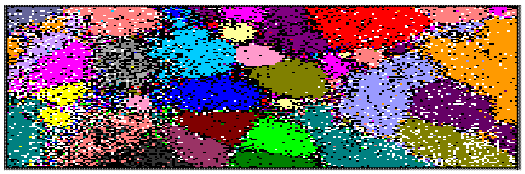
DataGRID for
High-energy Science



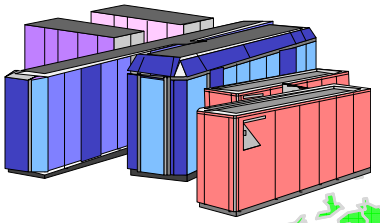
OC-48+ transmission
for Radio Telescope



Nano-Technology
For GRID Application



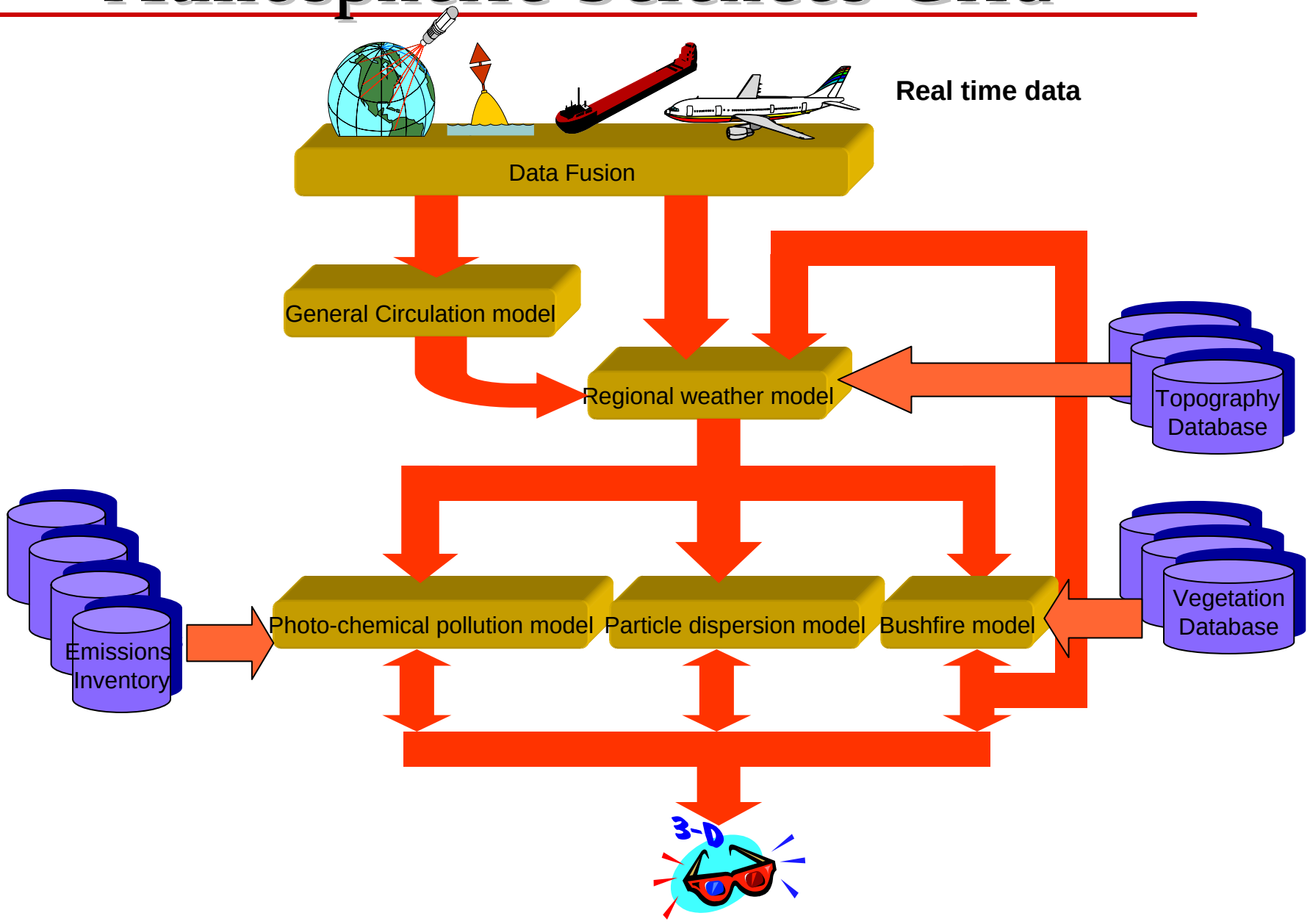
Bio-Informatics



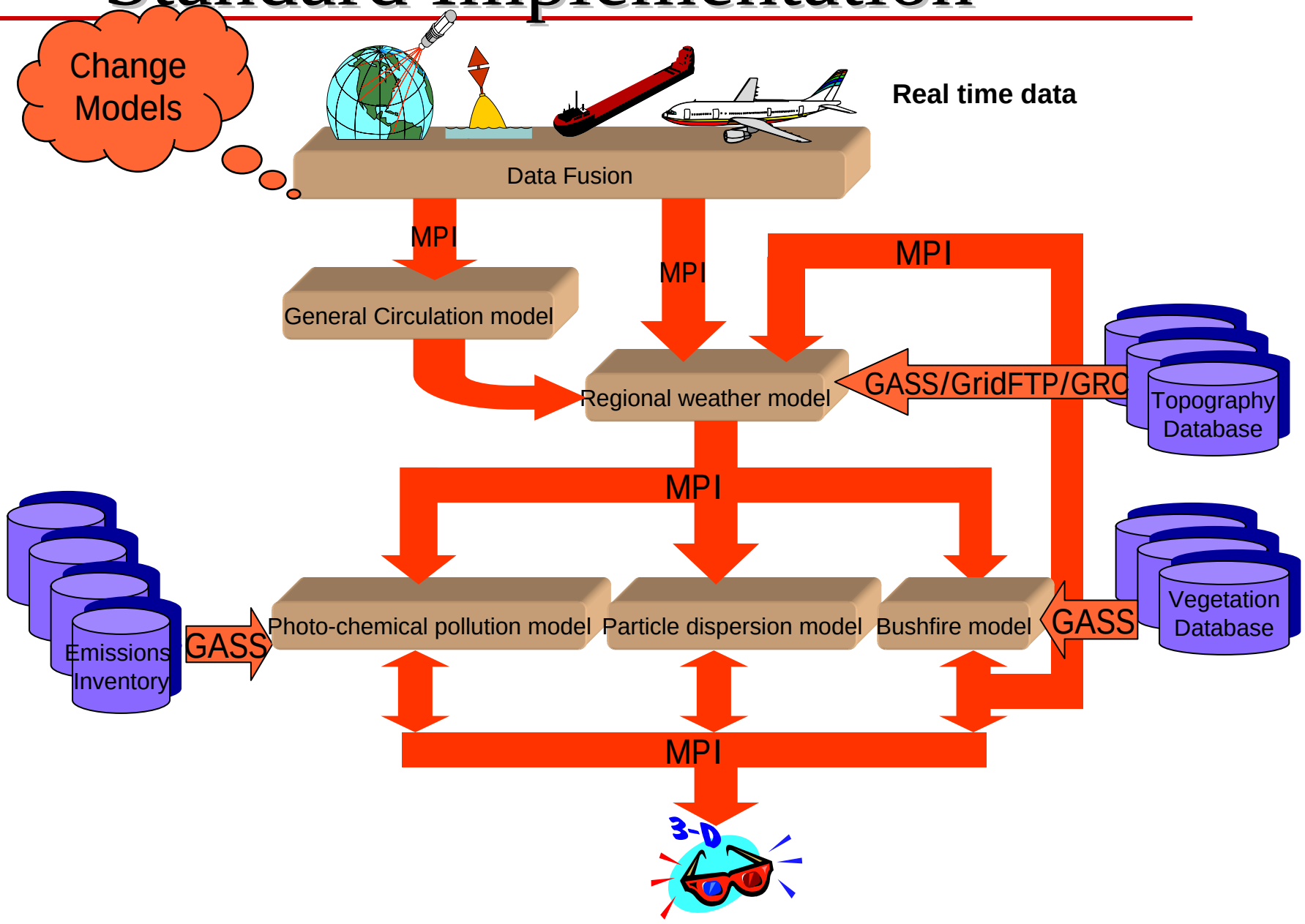
Computational
GRID and
NAREGI

Hokkaido U.
Tohoku U.
NIFS
Waseda U.
Kyoto U.
NAO
KEK
Osaka U.
Tsukuba U.
Operation (NII)
Doshidha U.
U. of Tokyo
Tokyo Institute of Tech.
Nagoya U.
Okazaki Research
Institutes
NII R&D
ISAS
NIG
Kyushu U.

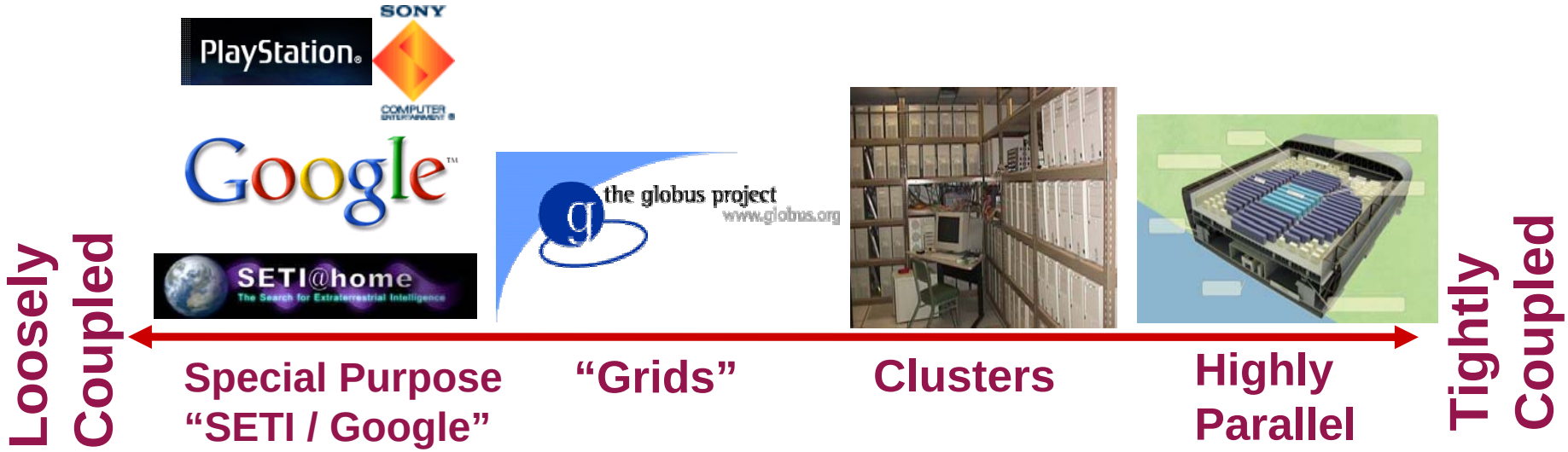
Atmospheric Sciences Grid



Standard Implementation



The Computing Continuum



- Each strikes a different balance
  - computation/communication coupling
- Implications for execution efficiency
- *Applications for diverse needs*
  - *computing is only one part of the story!*

Grids vs. Capability vs. Cluster Computing

- Not an "either/or" question
  - Each addresses different needs
  - Each 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 computations and teams
- Clusters
  - Low cost, group solution
  - Potential hidden costs
- Key is easy access to resources in a transparent way

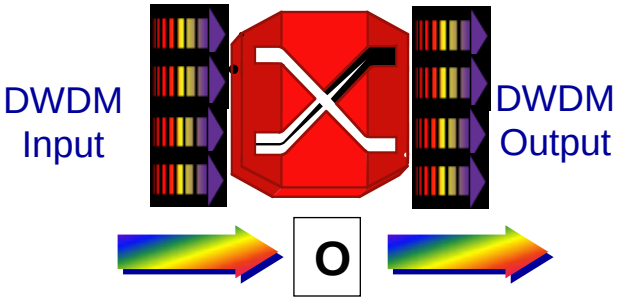
Real Crisis With HPC Is With The Software

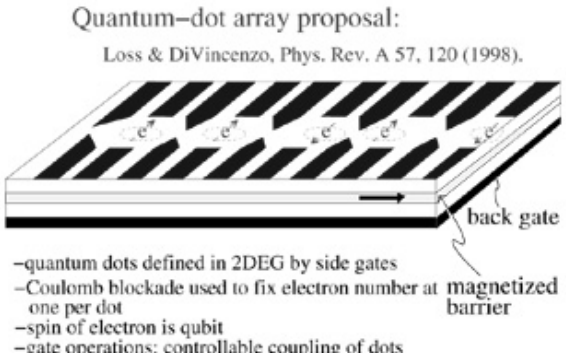
- It's time for a change
  - complexity is rising dramatically
    - highly parallel and distributed systems
      - From 10 to 100 to 1000 to 10000 to 100000 of processors!!
    - multidisciplinary applications
- Programming is stuck
  - arguably hasn't changed since the 60's
- A supercomputer application and software are usually much more long-lived than a hardware
  - Hardware life typically five years at most.
  - Fortran and C are the main programming models
- Software is a major cost component of modern technologies.
  - The tradition in HPC system procurement is to assume that the software is free.
- We don't have any great ideas about how to solve this problem.

Future Directions

- Silicon
  - escaping the von Neumann bottleneck
  - streaming, smart, dense packages, …
  - processing in memory (PIM)
- Optical computing
- Biological computing
- Quantum computing

Courage, Heart, and Brains

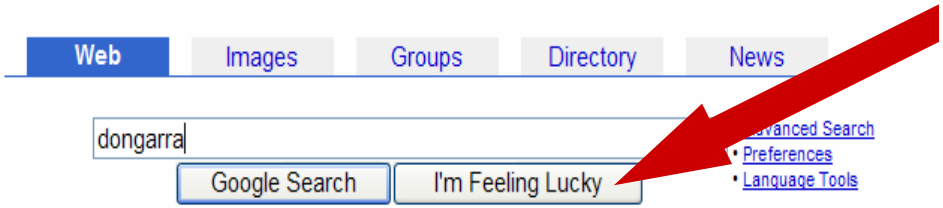




Collaborators / Support

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

An Coláiste Ollscoile, Baile Átha Cliath
University College Dublin



Advertise with Us - Business Solutions - Services & Tools - Jobs, Press, & Help

Make Google Your Homepage!

©2003 Google - Searching 3,083,324,652 web pages