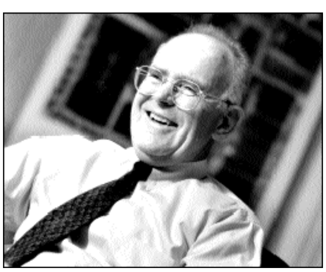
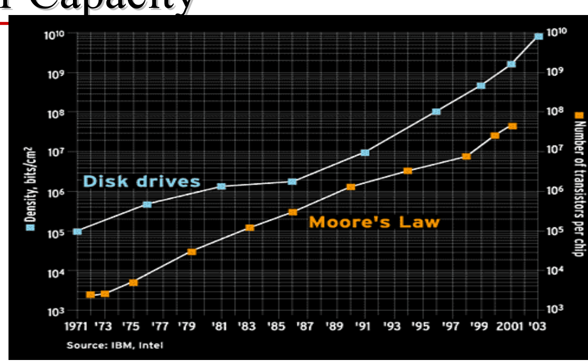
Trends in High Performance Computing and the Grid

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory

Technology Trends: Microprocessor Capacity



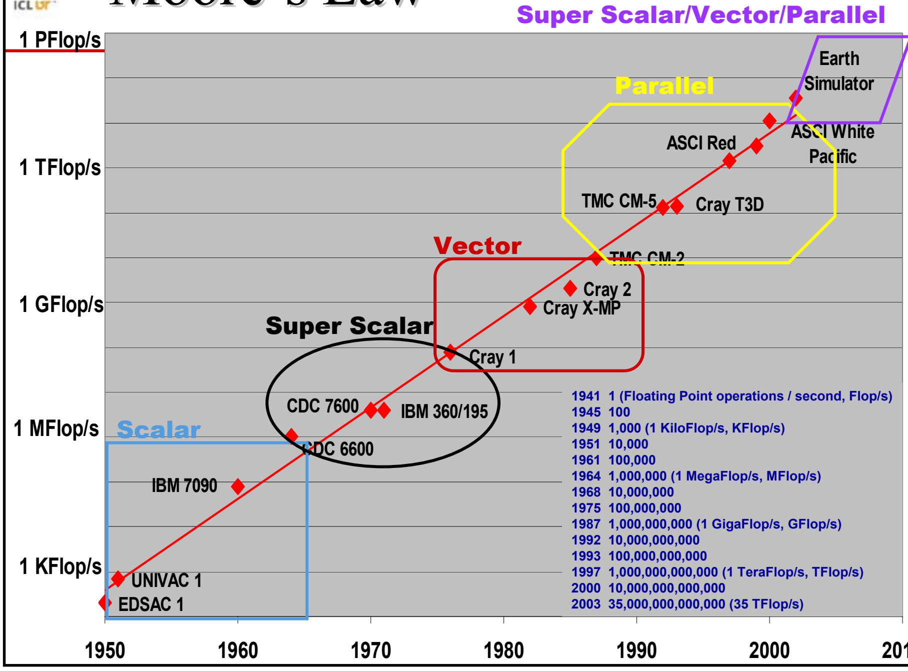


Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

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

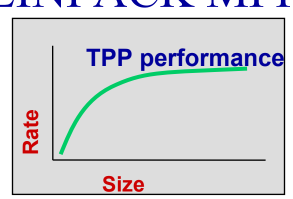


TOP500 superCOMPUTER

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

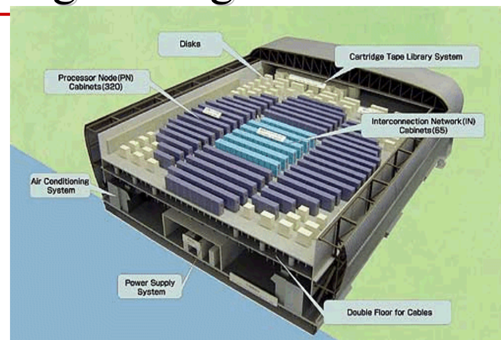
  SC'xy in the States in November

  Meeting in Mannheim, Germany in June
- All data available from **www.top500.org**

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 & 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

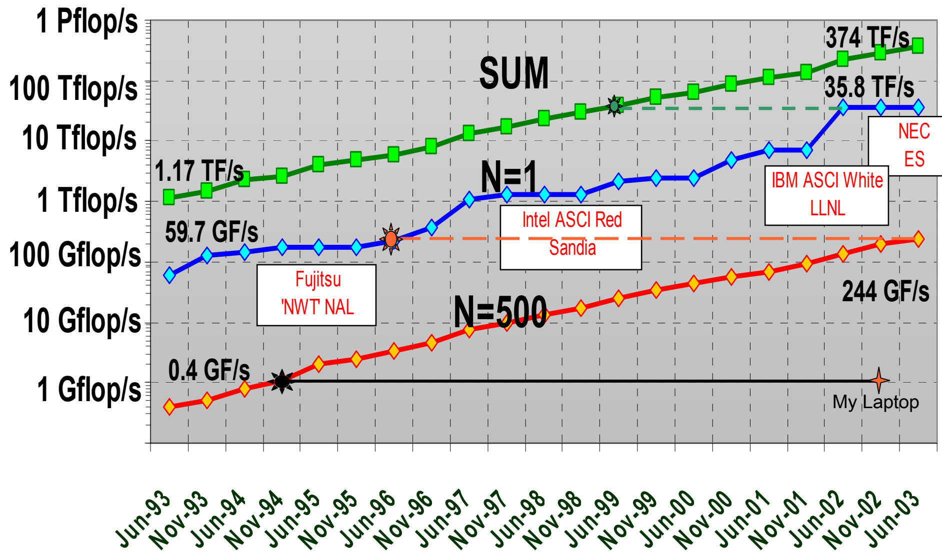
- From the Top500 (June 2003)
  - Performance of ESC ≈ Σ Next Top 4 Computers
  - ~ 10% of performance of all the Top500 machines



June 2003

	Manufacturer	Computer	Rmax	Installation Site	Year	# Proc	Rpeak
1	NEC	Earth-Simulator	35860	Earth Simulator Center Yokohama	2002	5120	40960
2	Hewlett-Packard	ASCI Q - AlphaServer SC ES45/1.25 GHz	13880	Los Alamos National Laboratory Los Alamos	2002	8192	20480
3	Linux NetworX Quadrics	MCR Linux Cluster Xeon 2.4 GHz - Quadrics	7634	Lawrence Livermore National Laboratory Livermore	2002	2304	11060
4	IBM	ASCI White, SP Power3 375 MHz	7304	Lawrence Livermore National Laboratory Livermore	2000	8192	12288
5	IBM	SP Power3 375 MHz 16 way	7304	NERSC/LBNL Berkeley	2002	6656	9984
6	IBM/Quadrics	xSeries Cluster Xeon 2.4 GHz - Quadrics	6586	Lawrence Livermore National Laboratory Livermore	2003	1920	9216
7	Fujitsu	PRIMEPOWER HPC2500 (1.3 GHz)	5406	National Aerospace Lab Tokyo	2002	2304	11980
8	Hewlett-Packard	rx2600 Itanium2 1 GHz Cluster - Quadrics	4881	Pacific Northwest National Laboratory Richland	2003	1540	6160
9	Hewlett-Packard	AlphaServer SC ES45/1 GHz	4463	Pittsburgh Supercomputing Center Pittsburgh	2001	3016	6032
10	Hewlett-Packard	AlphaServer SC ES45/1 GHz	3980	Commissariat a l'Energie Atomique (CEA) Bruyeres-le-Chatel	2001	2560	5120

TOP500 – Performance - June 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 9.555 Tflop/s
    - Could be #3 on 11/03 Top500
  - Cost is $5.2 million which includes the system itself, memory, storage, and communication fabrics

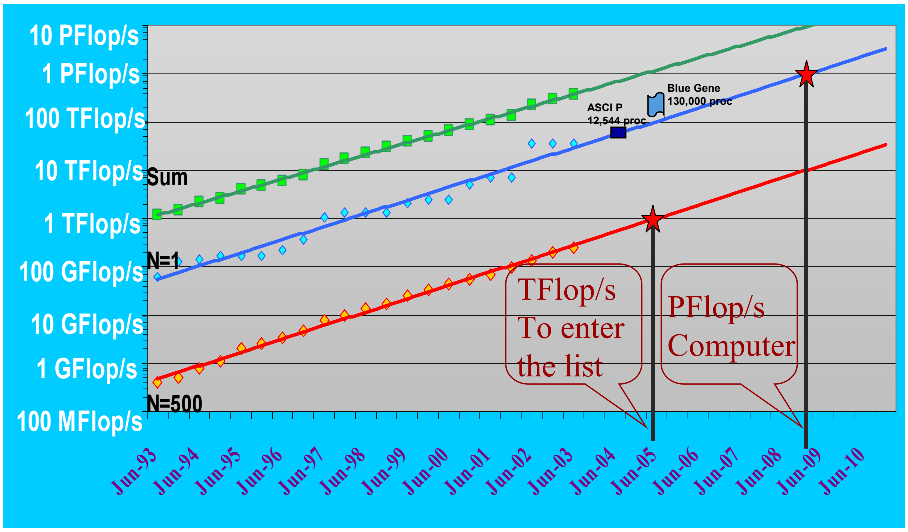
Detail on the Virginia Tech Machine

- Dual Power PC 970 2GHz
  - 4 GB DRAM.
  - 160 GB serial ATA mass storage.
  - 4.4 TB total main memory.
  - 176 TB total mass storage.
- Primary communications backplane based on infiniband technology.
  - Each node can communicate with the network at 20 Gb/s, full duplex, "ultra-low" latency.
  - Switch consists of 24 96-port switches in fat-tree topology.
- Secondary Communications Network:
  - Gigabit fast ethernet management backplane.
  - Based on 5 Cisco 4500 switches, each with 240 ports.
- Software:
  - Mac OSX.
  - MPIch-2
  - C, C++ compilers - IBM xlc and gcc 3.3
  - Fortran 95/90/77 Compilers - IBM xlf and NAGWare

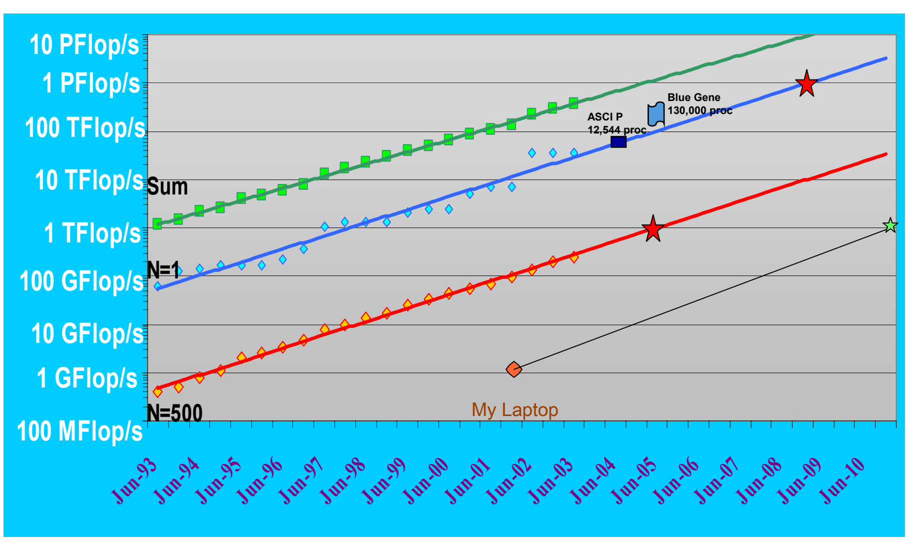
Top 5 Machines for the Linpack Benchmark

	Computer (Full Precision)	Number of Procs	R_{max} GFlop/s	R_{Peak} GFlop/s
1	Earth Simulator	5120	35860	40960
2	ASCI Q AlphaServer EV-68 (1.25 GHz w/Quadrics)	8160	13880	20480
3	Apple G5 dual IBM Power PC (2 GHz, 970s, w/Infiniband 4X)	2112	9555	16896
4	HP RX2600 Itanium 2 (1.5GHz w/Quadrics)	1936	8633	11616
5	Linux NetworX (2.4 GHz Pentium 4 Xeon w/Quadrics)	2304	7634	11059

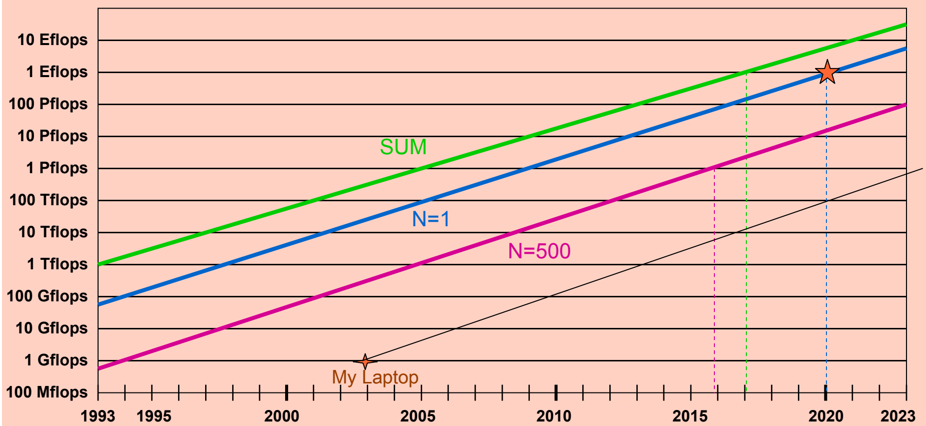
Performance Extrapolation



Performance Extrapolation

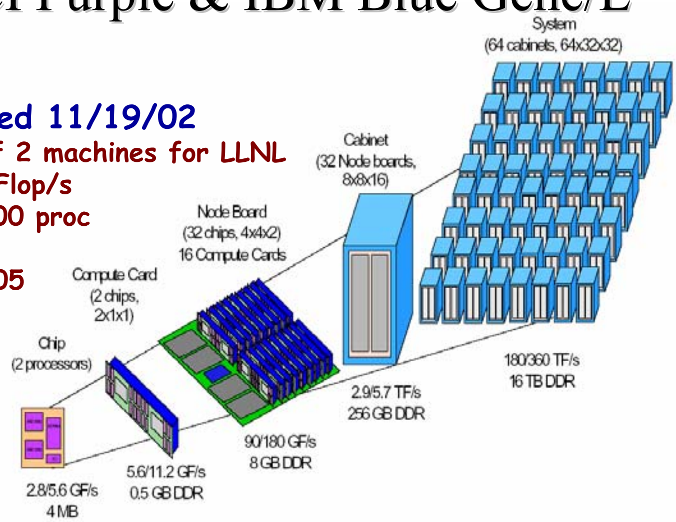


To Exaflop/s (10^{18} and Beyond)



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**



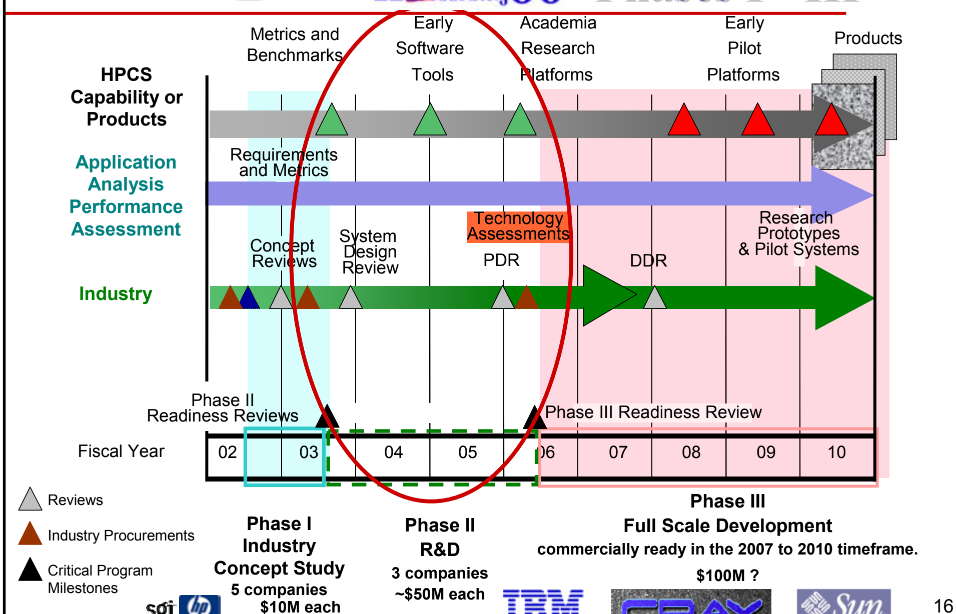
- **Preliminary machine**
  - **IBM Research BlueGene/L**
    - PowerPC 440, 500MHz w/custom proc/interconnect
    - 512 Nodes (1024 processors)
    - 1.435 Tflop/s (2.05 Tflop/s Peak)

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

Selected System Characteristics

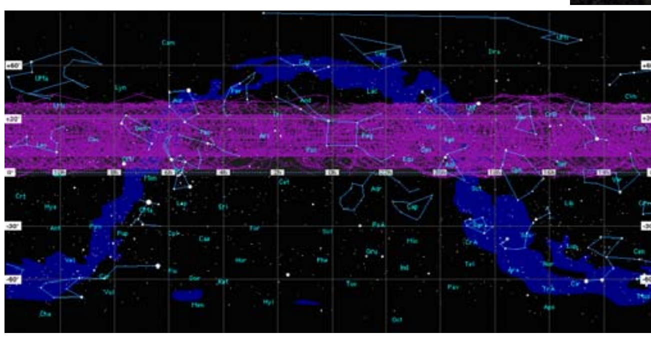
		Earth Simulator (NEC)	Cray X1 (Cray)	ASCI Q (HP ES45)	MCR (Dual Xeon)
Year of Introduction		2002	2003	2003	2002
Node Architecture		Vector SMP	Vector SMP	Alpha micro SMP	Xeon micro SMP
System Topology		NEC single-stage Crossbar	2D Torus Interconnect	Quadrics QsNet Fat-tree	Quadrics QsNet Fat-tree
Number of Nodes		640	32	2048	1152
Processors	- per node	8	4	4	2
	- system total	5120	128	8192	2304
Processor Speed		500 MHz	800 MHz	1.25 GHz	2.4 GHz
Peak Speed	- per processor	8 Gflops	12.8 Gflops	2.5 Gflops	4.8 Gflops
	- per node	64 Gflops	51.2 Gflops	10 Gflops	9.6 Gflops
	- system total	40 Tflops	1.6 Tflops	30 Tflops	10.8 Tflops
Memory	- per node	16 GB	8-64 GB	16 GB	16 GB
	- per processor	2 GB	2-16 GB	4 GB	2 GB
	- system total	10.24 TB		48 TB	4.6 TB
Memory Bandwidth (peak)					
	- L1 Cache	N/A	76.8 GB/s	20 GB/s	20 GB/s
	- L2 Cache	N/A		13 GB/s	1.5 GB/s
	Main (per proc)	32 GB/s	34.1 GB/s	2 GB/s	2 GB/s
Inter-node MPI					
	- Latency	8.6 μsec	8.6 μsec	5 μsec	4.75 μsec
	- Bandwidth	11.8 GB/s	11.9 GB/s	300 MB/s	315 MB/s
Bytes/flop to main memory		4	3	0.8	0.4
Bytes/flop interconnect		1.5	1	0.12	0.07

HPCS Phases I - III

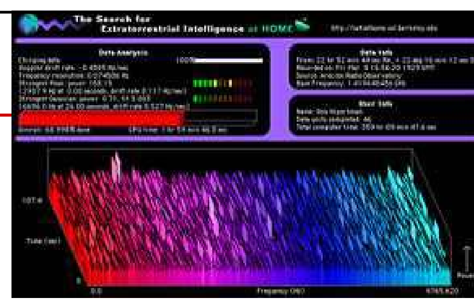


SETI@home: Global Distributed Computing

- 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

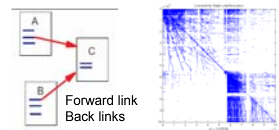


- Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
- When their computer is idle or being wasted this software will download ~ half a MB chunk of data for analysis. Performs about 3 Tflops for each client in 15 hours.
- The results of this analysis are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

- Largest distributed computation project in existence
  - Averaging 55 Tflop/s
  - 1368 users

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



Source: Monika Henzinger, Google & Cleve Moler

How Google Works; You have to think big

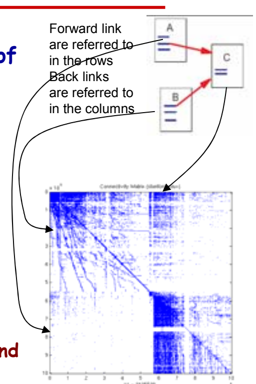
This is done "offline" …

Number of inlinks to a web page is a sign of the importance of the web page

- Generate an incidence matrix of links to and from web pages
  - For each web page there's a row/column
  - Matrix of order $3x10^9$
- Form a transition probability matrix of the Markov chain
  - Matrix is not sparse, but it is a rank one modification of a sparse matrix
- Compute the eigenvector corresponding to the largest eigenvalue, which is 1.
  - Solve $Ax = x$.
  - Use the power method? (x=initial guess; iterate $x \leftarrow Ax$;)
  - Each component of the vector x corresponds to a web page and represents the weight (importance) for that web page.
  - This is the basis for the "Page rank"
- Create an inverted index of the web;
  - word : web pages that contain that word

When a query, set of words, comes in:

- Go to the inverted index and get the corresponding web pages for the query
- Rank the resulting web pages by the "Page rank" and return pointers to those page in that order.



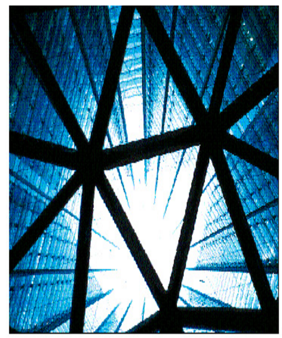
Eigenvalue problem
$n=3x10^9$
(see: MathWorks Cleve's Corner)

Source: Monika Henzinger, Google & Cleve Moler

Computational Science and the Grid

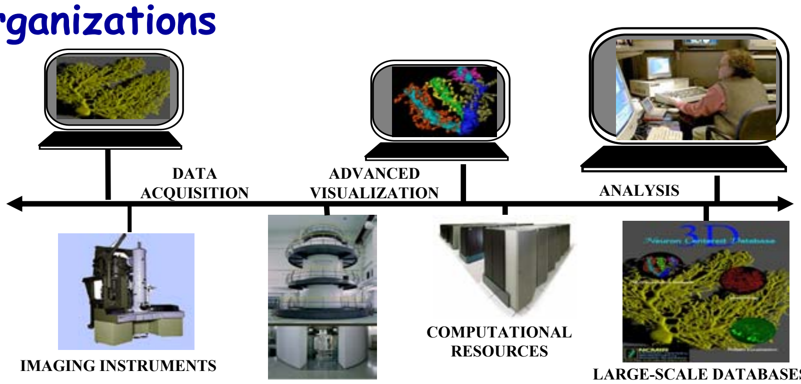
Today's computational and information infrastructure must address both science and technology trends

- Proliferation of resources
  - Everyone has computers
  - Multiple IP addresses per person
- Increased demand
  - Immense amounts of data
  - Applications increasingly
    - Multi-scale
    - Multi-disciplinary
    - Information-driven
- Coordination/collaboration is a default mode of interaction
  - The Internet
  - Globalization, virtualization
  - Open source movement

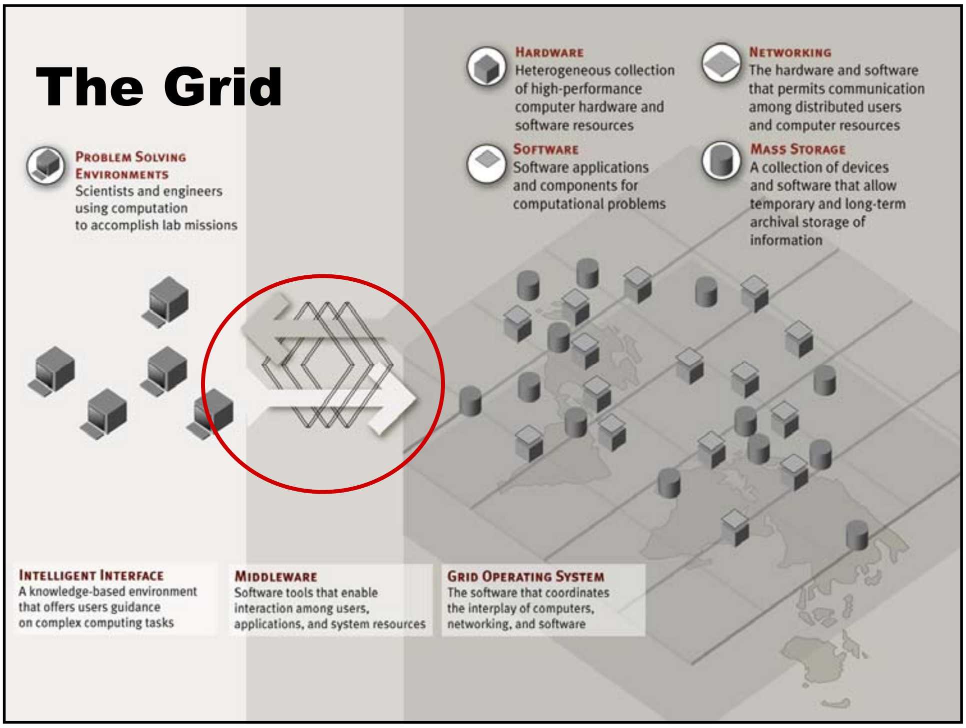


Grid Computing is About …

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



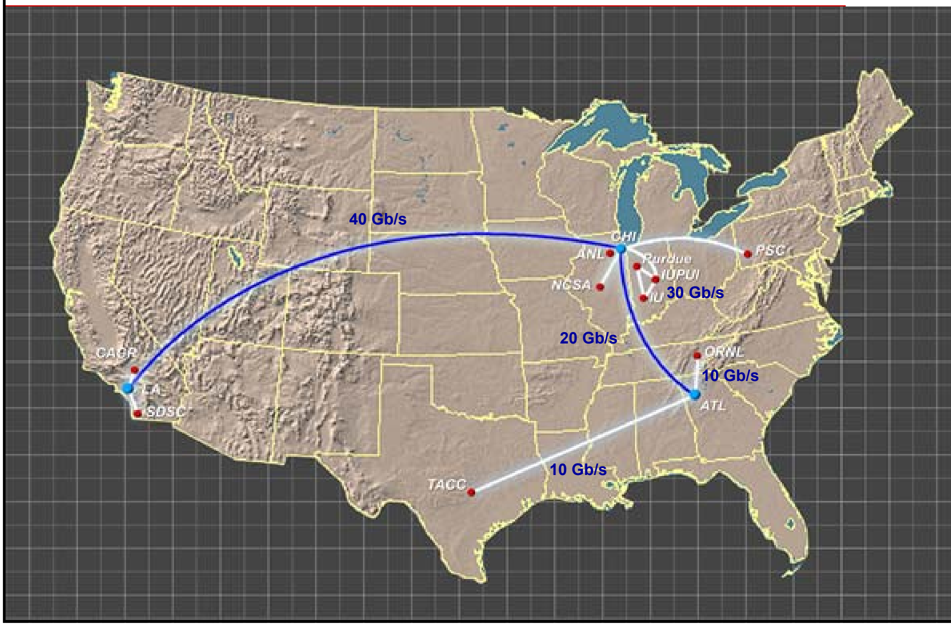
"Telescience Grid", Courtesy of Mark Ellisman



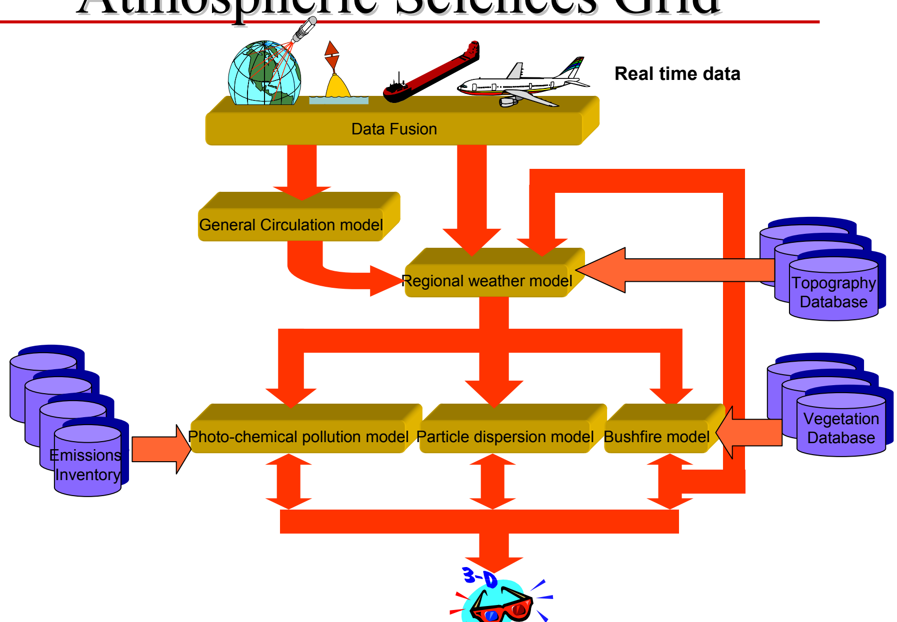
"Benefits of the Grid"

1. Higher Utilization of Distributed Resources
   - E.g., Supercomputing Center Grid, Nationwide Virtual Supercomputers
   - No increase in the overall "pie"
2. Higher Reliability and Upgradability of compute resources
   - Same objectve as the Internet (or, ARPANet)
3. "Collaboratory Science": tight collaboration of virtual organizations over the network
   - E.g., EU DataGrid w/3000 worldwide high-energy physicists
4. Tight Integration of Data, Sensors, Human Resources
   - VLBI (Astronomy) Project
5. Ultra-Scaling of Resources
   - Distributed placement of (otherwise oversized) resources
6. Exploitation of Idle Cycles/Storage on non-dedicated, commodity resources
   - Peer-to-Peer(P2P), Voluntary Computing

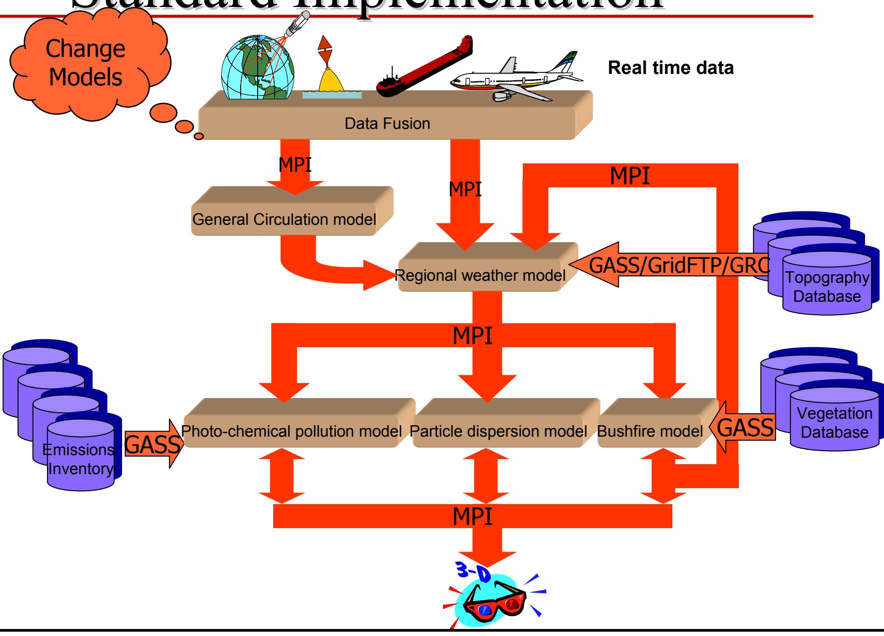
TeraGrid 2003
Prototype for a National Cyberinfrastructure



Atmospheric Sciences Grid



Standard Implementation



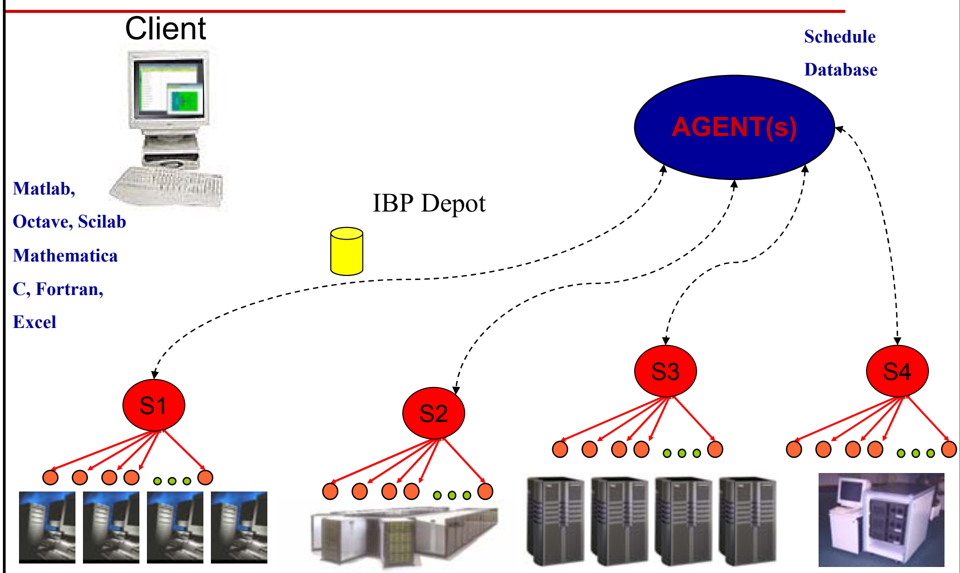
Some Grid Requirements – User Perspective

- Single sign-on: authentication to any Grid resources authenticates for all others
- Single compute space: one scheduler for all Grid resources
- Single data space: can address files and data from any Grid resources
- Single development environment: Grid tools and libraries that work on all grid resources

NetSolve Grid Enabled Server

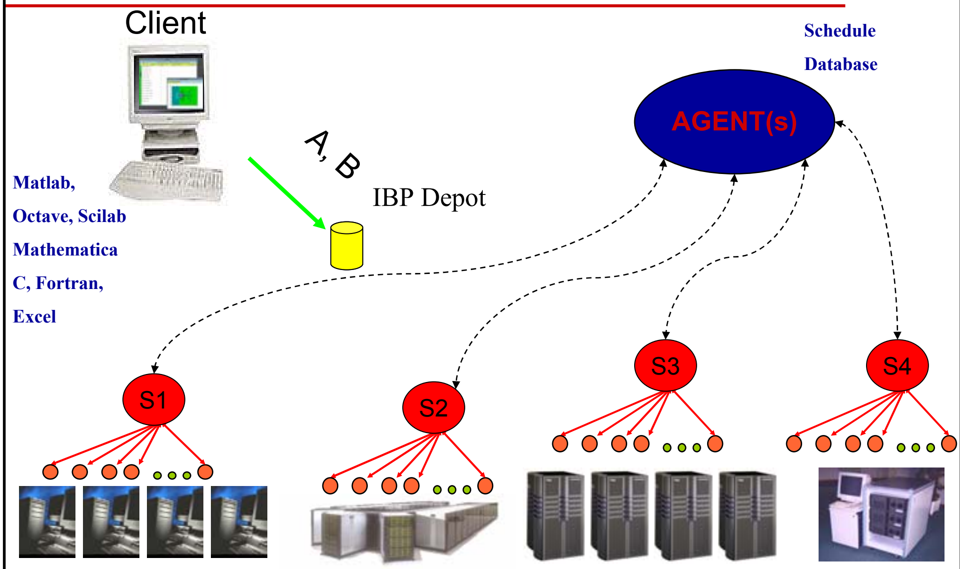
- NetSolve is an example of a Grid based hardware/software/data server.
- Based on a Remote Procedure Call model but with …
  - resource discovery, dynamic problem solving capabilities, load balancing, fault tolerance asynchronicity, security, …
- Easy-of-use paramount
- Its about providing transparent access to resources.

NetSolve: The Big Picture



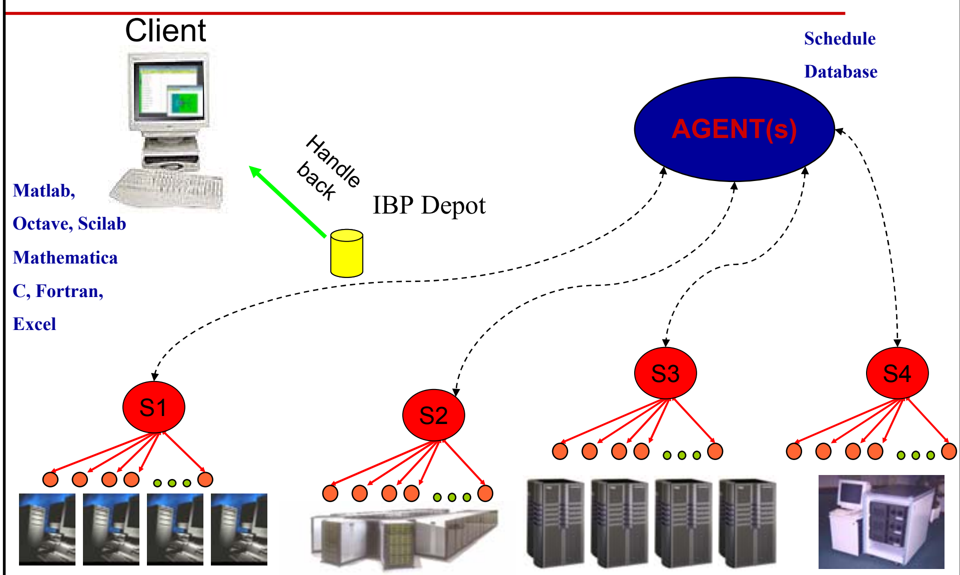
No knowledge of the grid required, RPC like.

NetSolve: The Big Picture



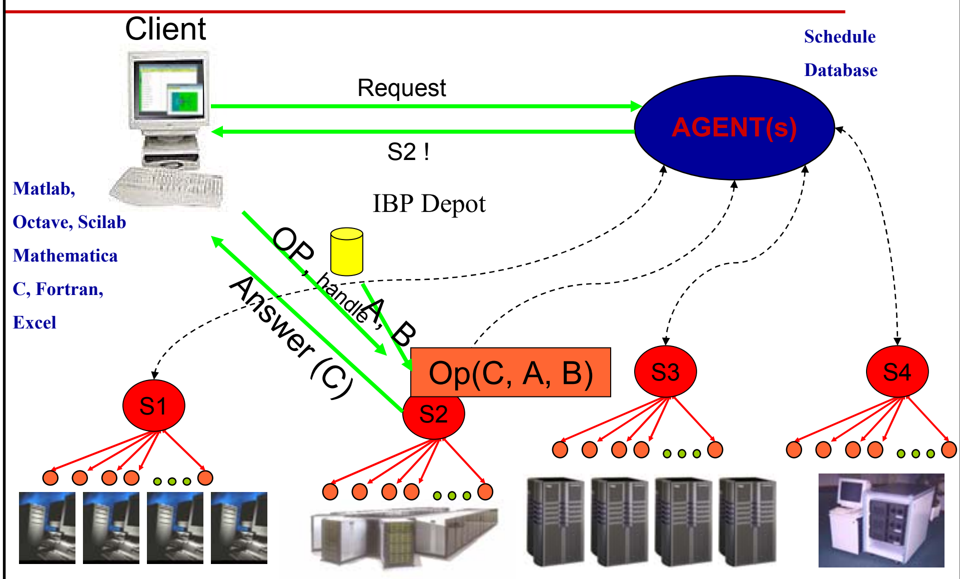
No knowledge of the grid required, RPC like.

NetSolve: The Big Picture

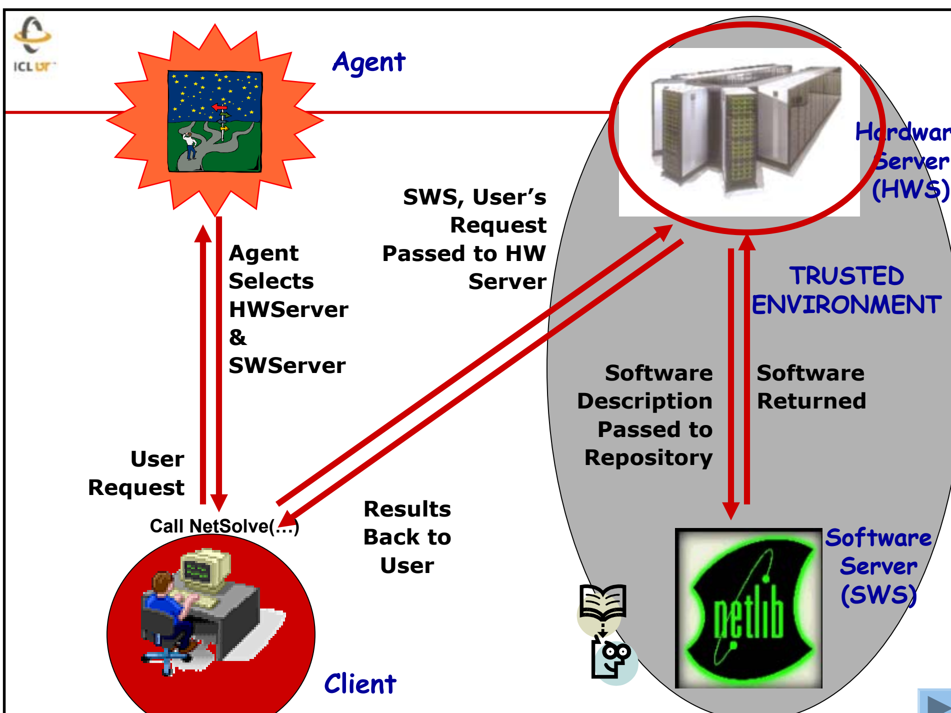


No knowledge of the grid required, RPC like.

NetSolve: The Big Picture

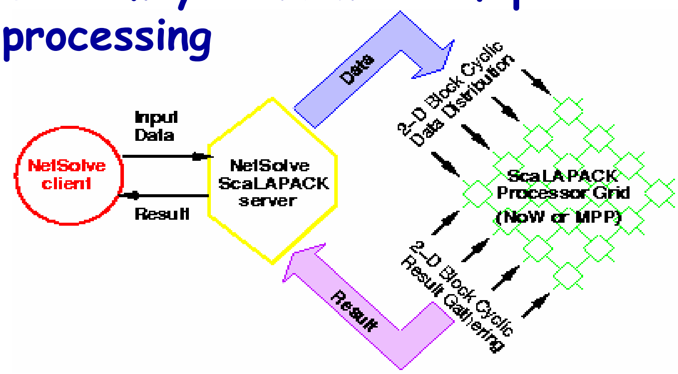


No knowledge of the grid required, RPC like.



Hiding the Parallel Processing

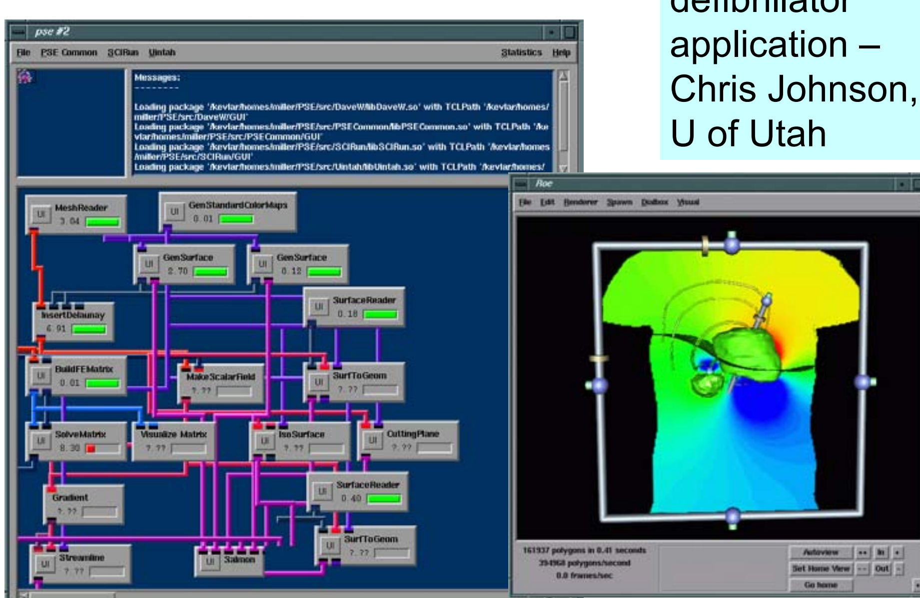
- User maybe unaware of parallel processing



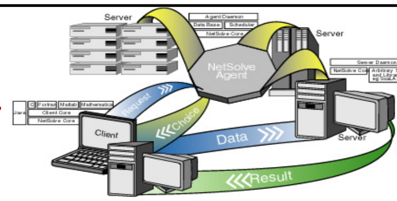
- NetSolve takes care of the starting the message passing system, data distribution, and returning the results. (Using LFC software)

Netsolve and SCIRun

SCIRun torso defibrillator application – Chris Johnson, U of Utah

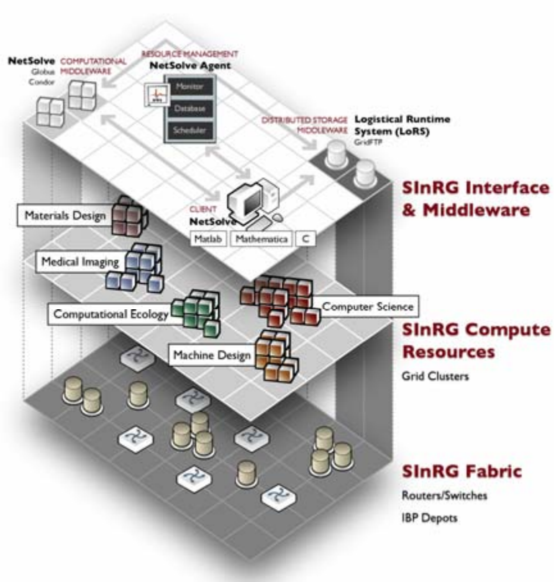


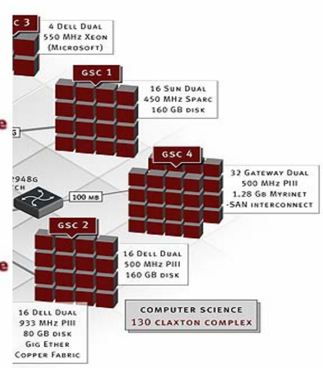
Basic Usage Scenarios



- Grid based numerical library routines
  - User doesn't have to have software library on their machine, LAPACK, SuperLU, ScaLAPACK, PETSc, AZTEC, ARPACK
- Task farming applications
  - "Pleasantly parallel" execution eg Parameter studies
- Remote application execution
  - Complete applications with user specifying input parameters and receiving output
- "Blue Collar" Grid Based Computing
  - Does not require deep knowledge of network programming
  - Level of expressiveness right for many users
  - User can set things up, no "su" required
  - In use today, up to 200 servers in 9 countries
- Can plug into Globus, Condor, NINF, …

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

New Features for NetSolve 2.0

New version available!

- New easy to use Interface Definition Language
  - Simplified PDF
- Dynamic servers
  - Add/delete problems without restarting servers
- New bindings for
  - GridRPC
  - Octave
  - Condor-G
- Separate hardware/software servers
- Support for Mac OS X & Windows 2K/XP
- Web based monitoring
- Allow user to specify server
- Allow user to abort execution

GridRPC - Introduction

- Attempting to provide:
  - Simple API upon which higher-level services could be implemented
  - Low burden on programmer attempting to transition code to the Grid
- Provide standardized, portable, and simple programming interface for Remote Procedure Call
- Attempt to unify client access to existing grid computing systems (such as NetSolve and Ninf-G)
- Working towards standardization through GGF WG
  - Initially standardize API; later deal with protocol
  - Standardize only minimal set of features; higher-level features can be built on top
  - Provide several reference implementations
    - Not attempting to dictate any implementation details

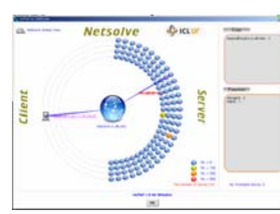
GridRPC - Features

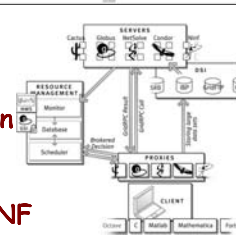
- Medium to coarse-grained calls (due to communication overhead)
- Asynchronous task-parallel programming
- Dynamic resource discovery and scheduling

- Jack Dongarra & Keith Seymour
  - University of Tennessee
- Hidemoto Nakada
  - National Institute of Advanced Industrial Science and Technology (AIST)
  - Tokyo Institute of Technology
- Satoshi Matsuoka
  - Tokyo Institute of Technology
  - National Institute of Informatics
- Craig Lee
  - The Aerospace Corporation
- Henri Casanova
  - San Diego Supercomputer Center
  - UCSD

NetSolve- Things Not Touched On

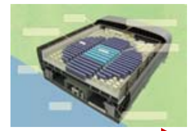
- Integration with other NMI tools
  - Globus, Condor, Network Weather Service
- Security
  - Using Kerberos V5 for authentication.
- Separate Server Characteristics
  - Hardware and Software servers
- Monitor NetSolve Network
  - Track and monitor usage
- Fault Tolerance
- Local / Global Configurations
- Dynamic Nature of Servers
- Automated Adaptive Algorithm Selection
  - Dynamic determine the best algorithm based on system status and nature of user problem
- NetSolve evolving into GridRPC
  - Being worked on under GGF with joint with NINF





The Computing Continuum

Loosely Coupled



Special Purpose "SETI / Google" | "Grids" | Clusters | Highly Parallel

Tightly Coupled

- **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

Collaborators / Support

- **TOP500**
  - H. Meuer, Mannheim U
  - H. Simon, NERSC
  - E. Strohmaier, NERSC
- **NetSolve**
  - Sudesh Agrawal, UTK
  - Henri Casanova, UCSD
  - Kiran Sagi, UTK
  - Keith Seymour, UTK
  - Sathish Vadhiyar, UTK