DEPARTMENT of MATHEMATICS

UMIST

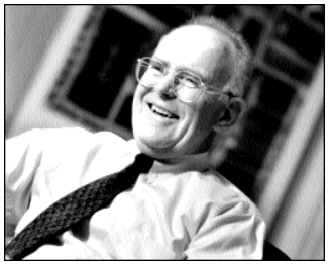
New Frontiers in Computational Mathematics

Saturday January 10 - Sunday January 11, 2004
Chancellors Hotel and Conference Centre, University of Manchester

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) Electronics Magazine, 1965

Number of devices/chip doubles every 12 months (later revised to 18 months)

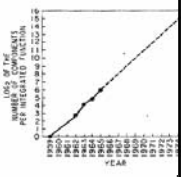
2X transistors/Chip Every 1.5 years

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.

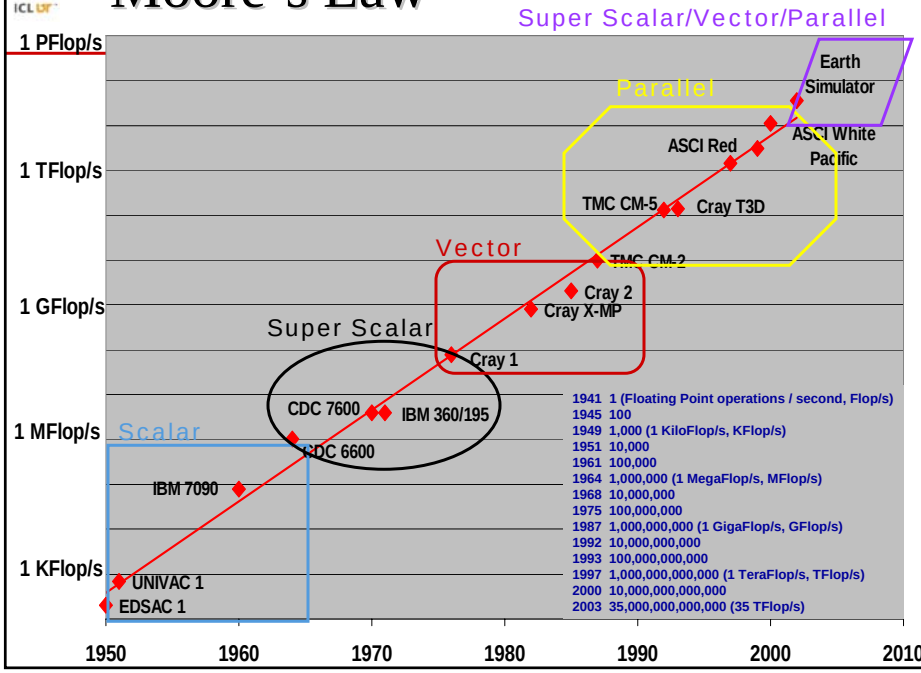
The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to an active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of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



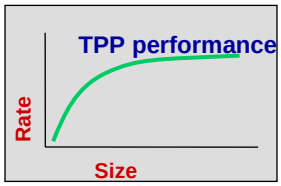
3

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*

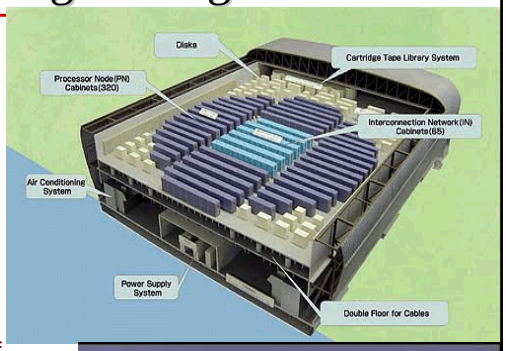


- Updated twice a year
  SC'xy in the States in November
  Meeting in Mannheim, Germany in June
- All data available from **www.top500.org**

4

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

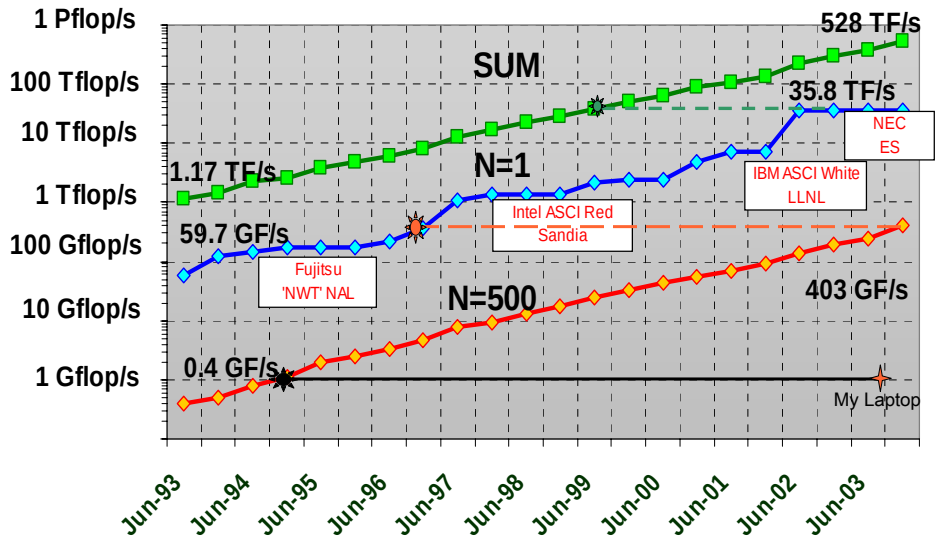


November 2003

	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	Self	Apple G5 Power PC w/Infiniband 4X	10280	Virginia Tech Blacksburg, VA	2003	2200	17600
4	Dell	PowerEdge 1750 P4 Xeon 3.6 Ghz w/Myrinet	9819	University of Illinois U/C Urbana/Champaign	2003	2500	15300
5	Hewlett-Packard	rx2600 Itanium2 1 GHz Cluster - w/Quadrics	8633	Pacific Northwest National Laboratory Richland	2003	1936	11616
6	Linux NetworX	Opteron 2 GHz, w/Myrinet	8051	Lawrence Livermore National Laboratory Livermore	2003	2816	11264
7	Linux NetworX	MCR Linux Cluster Xeon 2.4 GHz - w/Quadrics	7634	Lawrence Livermore National Laboratory Livermore	2002	2304	11060
8	IBM	ASCI White, Sp Power3 375 MHz	7304	Lawrence Livermore National Laboratory Livermore	2000	8192	12288
9	IBM	SP Power3 375 MHz 16 way	7304	NERSC/LBNL Berkeley	2002	6656	9984
10	IBM	xSeries Cluster Xeon 2.4 GHz - w/Quadrics	6586	Lawrence Livermore National Laboratory Livermore	2003	1920	9216

50% of top500 performance in top 9 machines; 131 system > 1 TFlop/s; 210 machines are clusters; 33 in UK 6

TOP500 – Performance - Nov 2003



7

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
    - #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

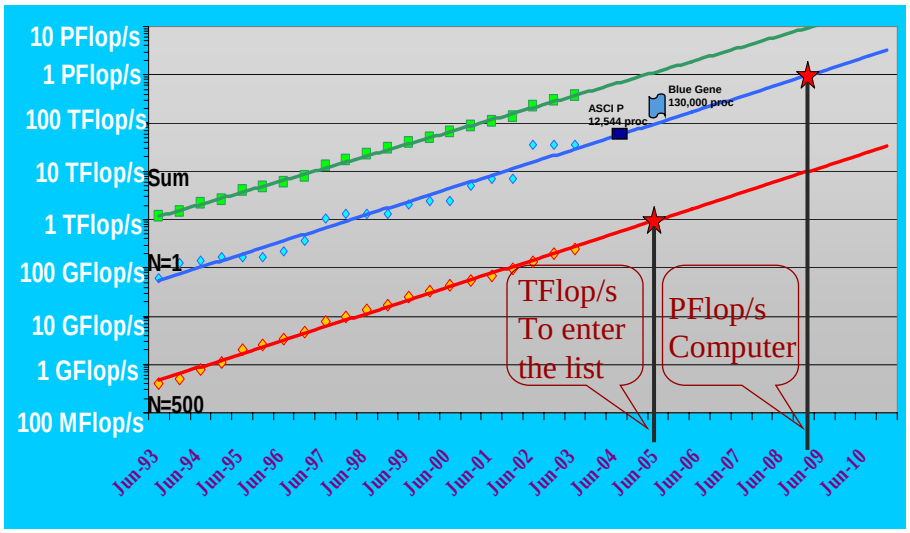
9

Top 5 Machines for the Linpack Benchmark

	Computer (Full Precision)	Number of Procs	*Achieved* GFlop/s	*T Peak* GFlop/s	*Efficiency*
1	Earth Simulator	5120	35860	40960	87.5%
2	LANL ASCI Q AlphaServer EV-68 (1.25 GHz w/Quadrics)	8160	13880	20480	67.7%
3	VT Apple G5 dual IBM Power PC (2 GHz, 970s, w/Infiniband 4X)	2112	10280	16896	60.9%
4	UIUC Dell Xeon Pentium 4 (3.06 Ghz w/Myrinet)	2500	9820	15300	64.2%
5	PNNL HP RX2600 Itanium 2 (1.5GHz w/Quadrics)	1936	8633	11616	74.3%

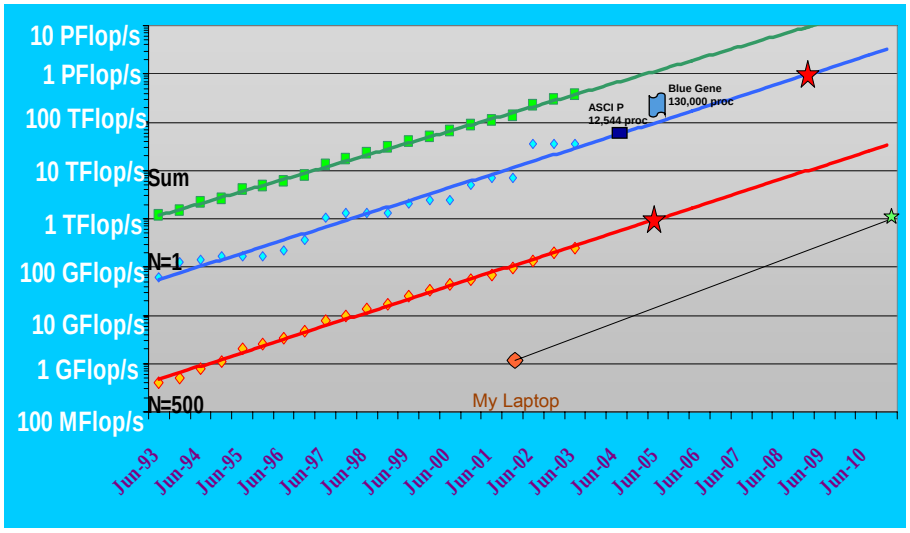
10

Performance Extrapolation



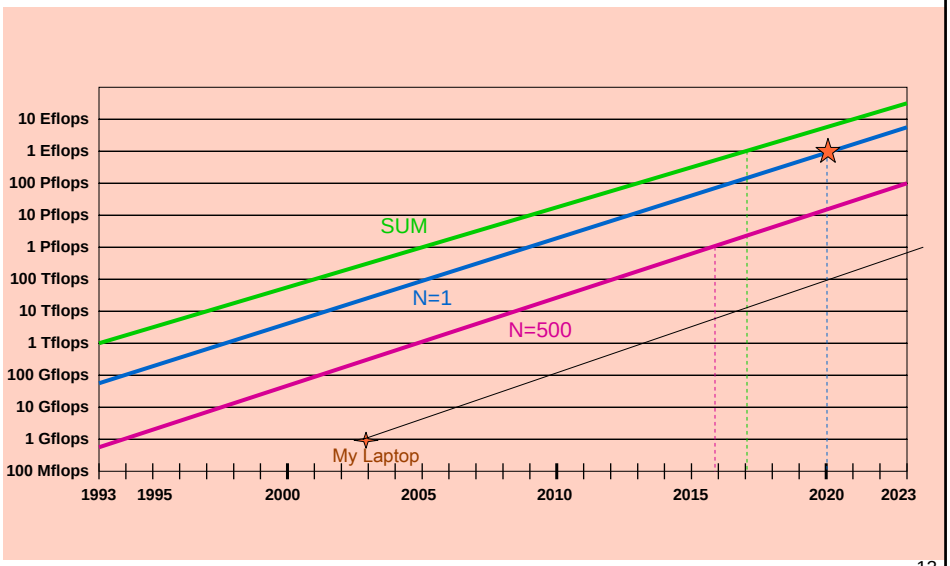
11

Performance Extrapolation



12

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

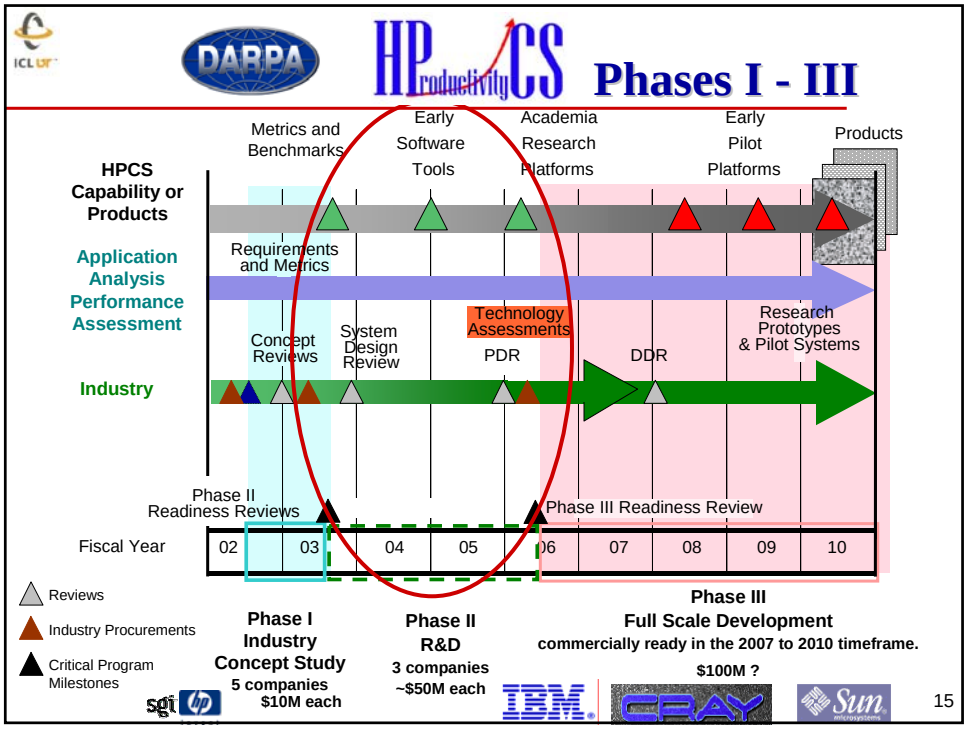


13

Selected System Characteristics

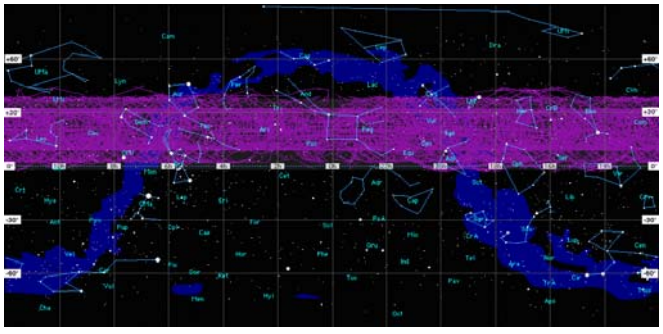
		Earth Simulator (NEC)	Cray X1 (Cray)	ASCI Q (HP ES45)	MCR (Dual Xeon)	VT Big Mac (Dual IBM PPC)
Year of Introduction		2002	2003	2003	2002	2003
Node Architecture		Vector SMP	Vector SMP	Alpha micro SMP	Xeon micro SMP	IBM 970 PPC SMP
System Topology		NEC single-stage Crossbar	2D Torus Interconnect	Quadrics QsNet Fat-tree	Quadrics QsNet Fat-tree	Infinaband Fat-tree
Number of Nodes		640	32	2048	1152	1100
Processors	- per node	8	4	4	2	2
	- system total	5120	128	8192	2304	2200
Processor Speed		500 MHz	800 MHz	1.25 GHz	2.4 GHz	2 GHz
Peak Speed	- per processor	8 Gflops	12.8 Gflops	2.5 Gflops	4.8 Gflops	8 Gflops
	- per node	64 Gflops	51.2 Gflops	10 Gflops	9.6 Gflops	16 Gflops
	- system total	40 Tflops	1.6 Tflops	30 Tflops	10.8 Tflops	17.6 Tflops
Memory	- per node	16 GB	8-64 GB	16 GB	16 GB	4 GB
	- per processor	2 GB	2-16 GB	4 GB	2 GB	2 GB
	- system total	10.24 TB		48 TB	4.6 TB	4.4 TB
Memory Bandwidth (peak)						
	- L1 Cache	N/A	76.8 GB/s	76.8 GB/s	76.8 GB/s	64 GB/s
	- L2 Cache	N/A		76.8 GB/s	76.8 GB/s	64 GB/s
	Main (per proc)	32 GB/s	34.1 GB/s	3.2 GB/s (400 MHz bus)	4.3 GB/s (533 MHz bus)	6.4 GB/s
Inter-node MPI						
	- Latency	8.6 µsec	8.6 µsec	5 µsec	4.75 µsec	9.5 µsec
	- Bandwidth	11.8 GB/s	11.9 GB/s	300 MB/s	315 MB/s	844 MB/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

14



SETI@home: Global Distributed Computing

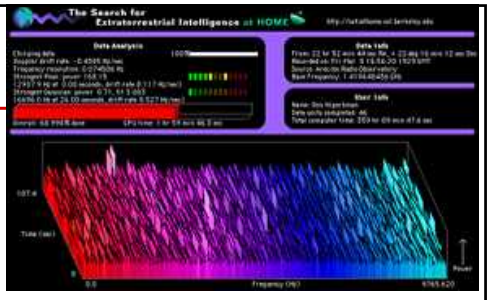
- Running on 500,000 PCs, ~1300 CPU Years per Day
  - 1.3M CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope



16

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.

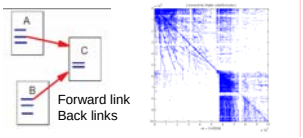


- **Largest distributed computation project in existence**
  - Averaging 55 Tflop/s
  - 1368 users

17

Google

- Google query attributes
  - 150M queries/day (2000/second)
  - 100 countries
  - 3.3B 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

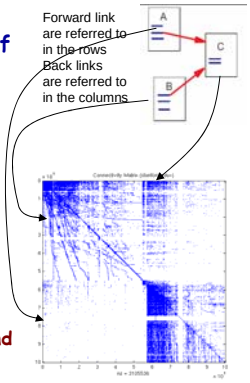
18

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
  - Sparse 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



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

When a query, set of words, comes in:

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

Source: Monika Henzinger, Google & Cleve Moler

19

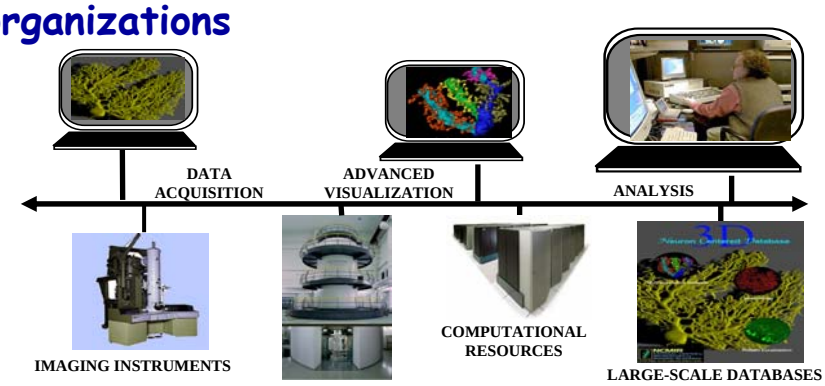
Science and Technology

- Today, large science projects are conducted by global teams using sophisticated combinations of
  - *People*
  - *Computers*
  - *Networks*
  - *Viz*
  - *data storage*
  - *remote instruments*
  - *other resources*
- *Information Infrastructure provides a way to integrate resources to support modern applications*

20

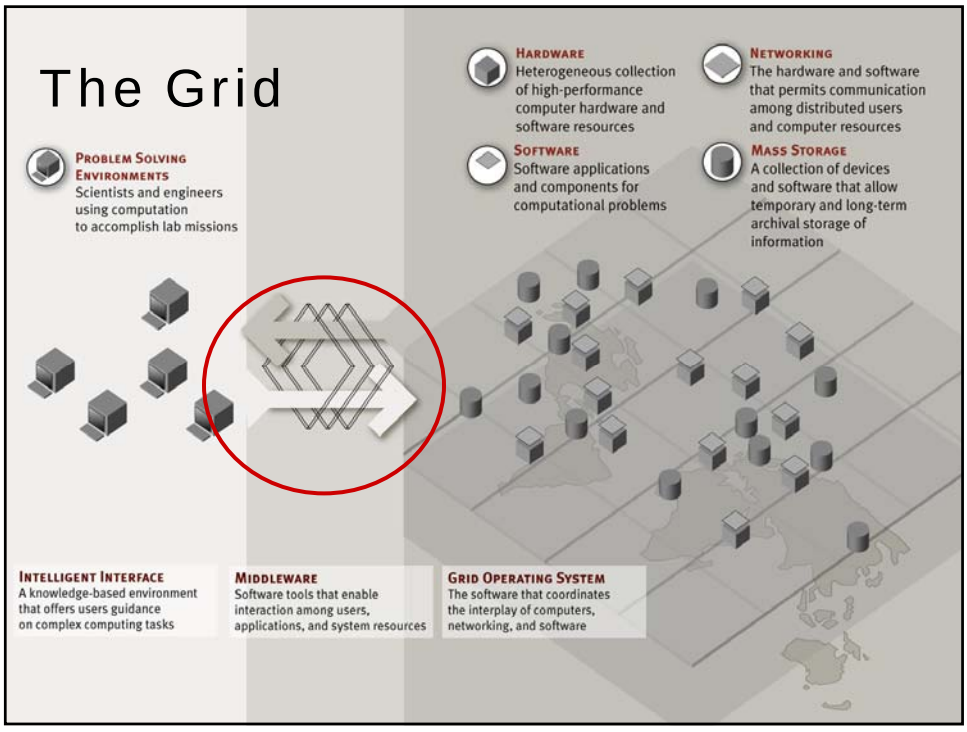
Grid Computing is About …

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



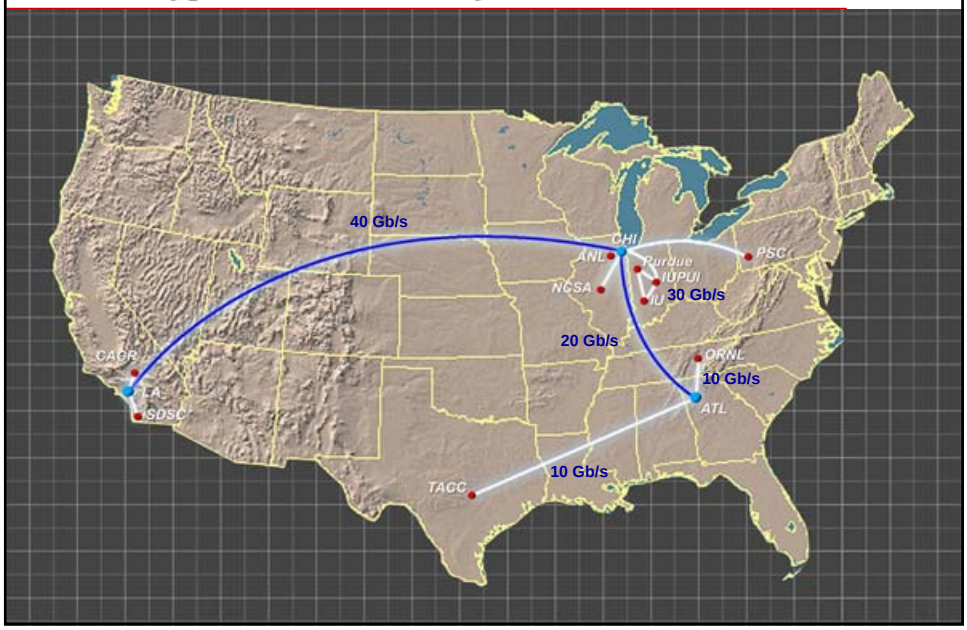
"Telescience Grid", Courtesy of Mark Ellisman

21

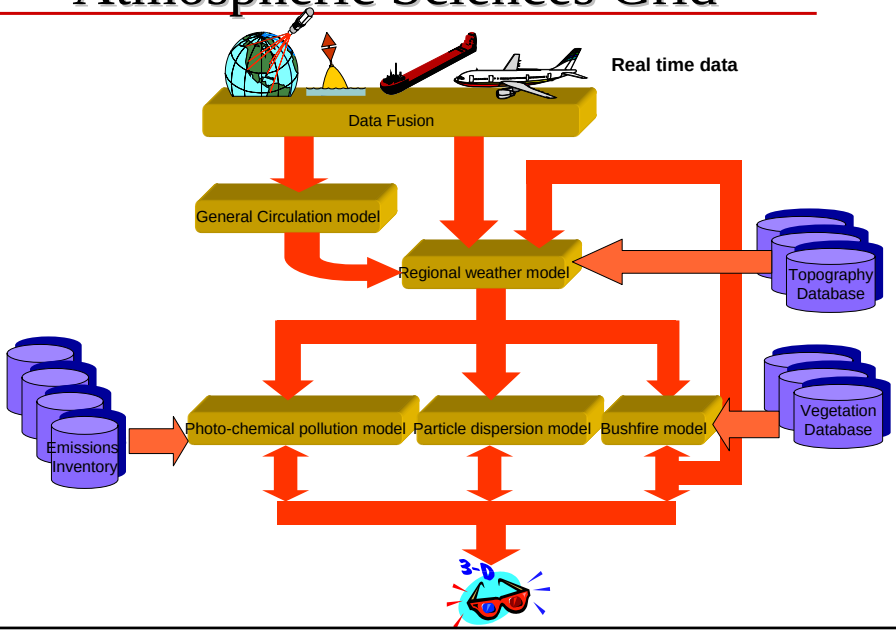


TeraGrid 2003

Prototype for a National Cyberinfrastructure

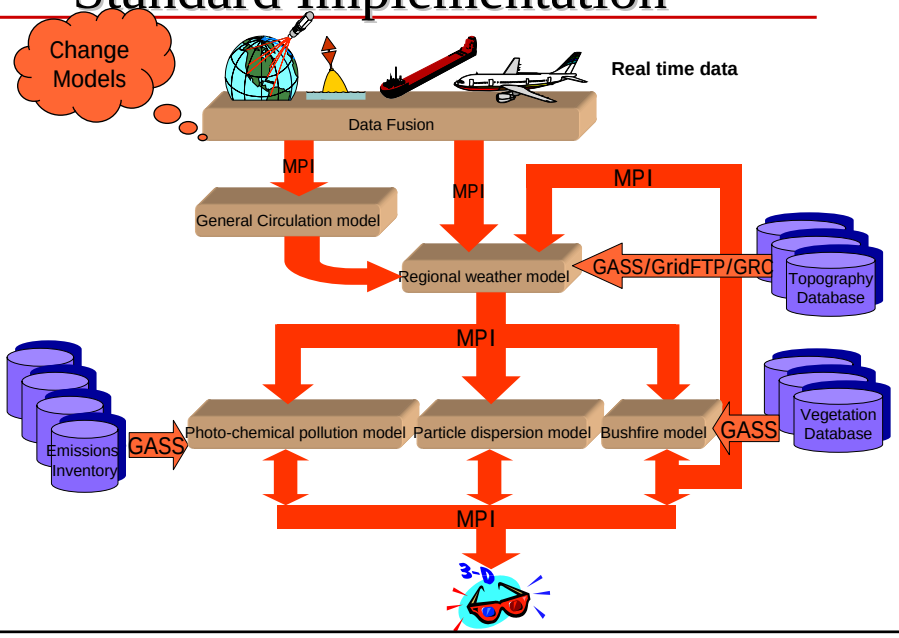


Atmospheric Sciences Grid



24

Standard Implementation



25

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

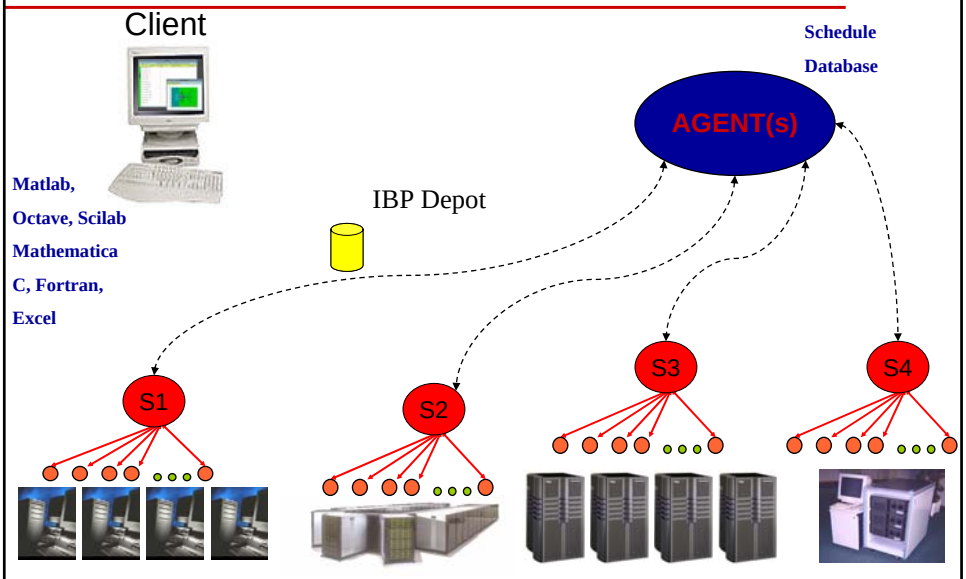
26

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.

27

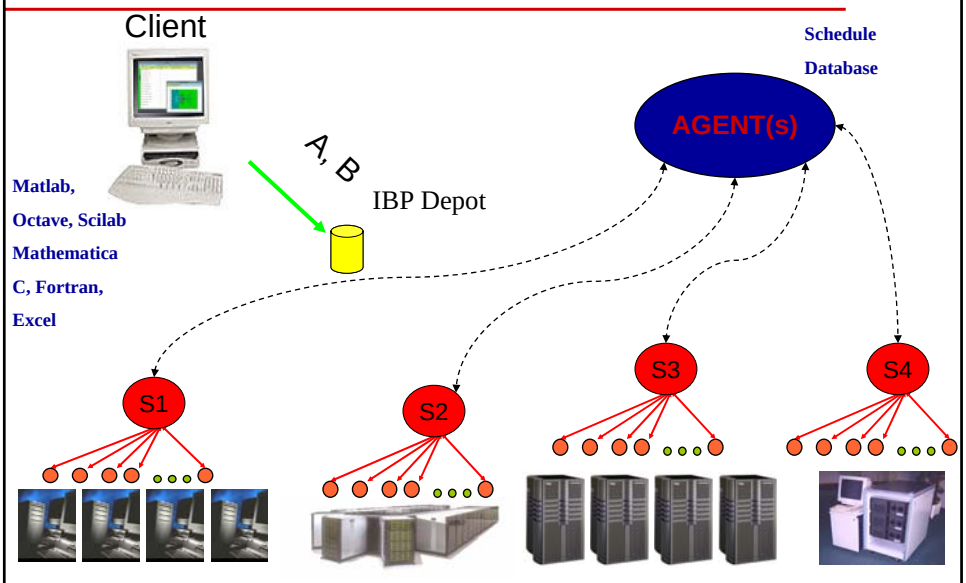
NetSolve: The Big Picture



No knowledge of the grid required, RPC like.

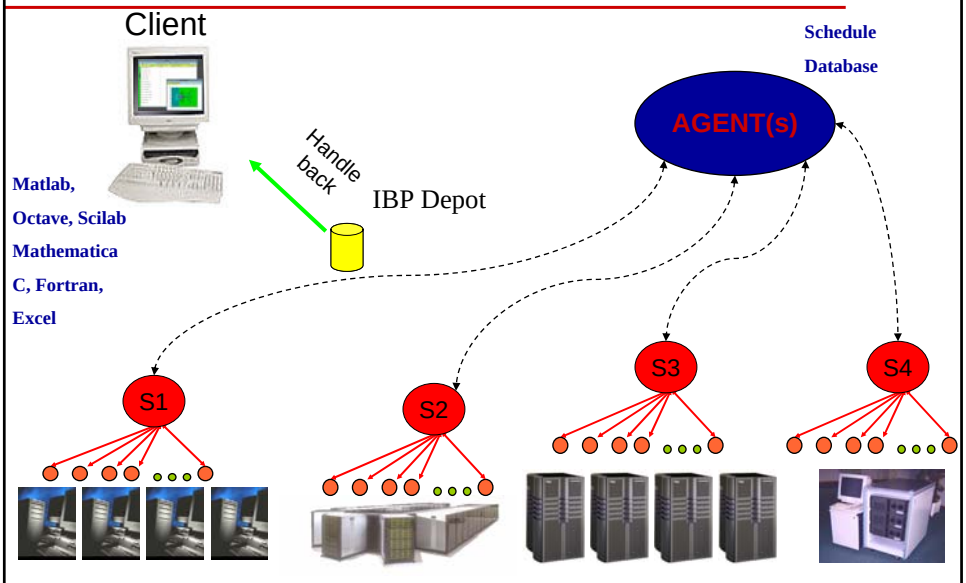
28

NetSolve: The Big Picture



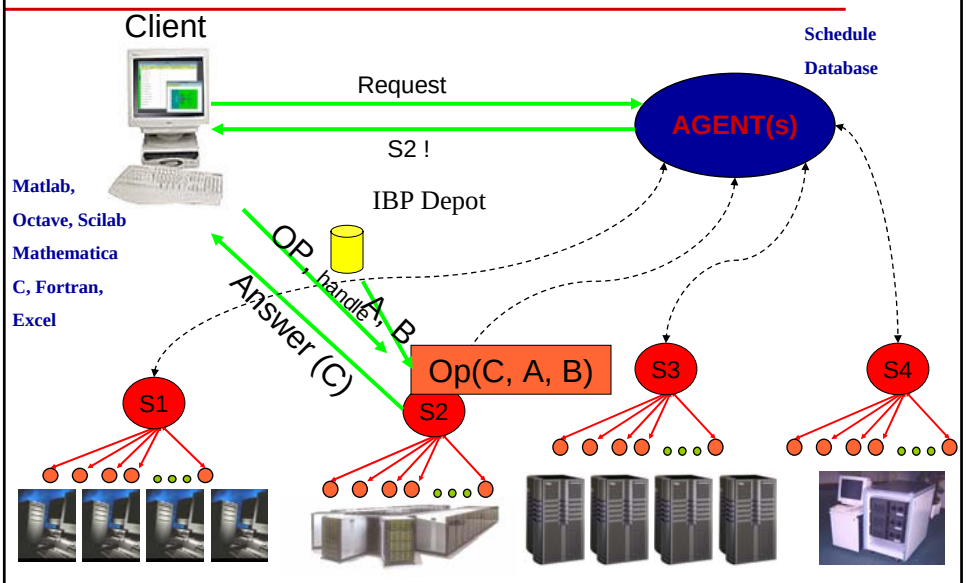
No knowledge of the grid required, RPC like. 29

NetSolve: The Big Picture

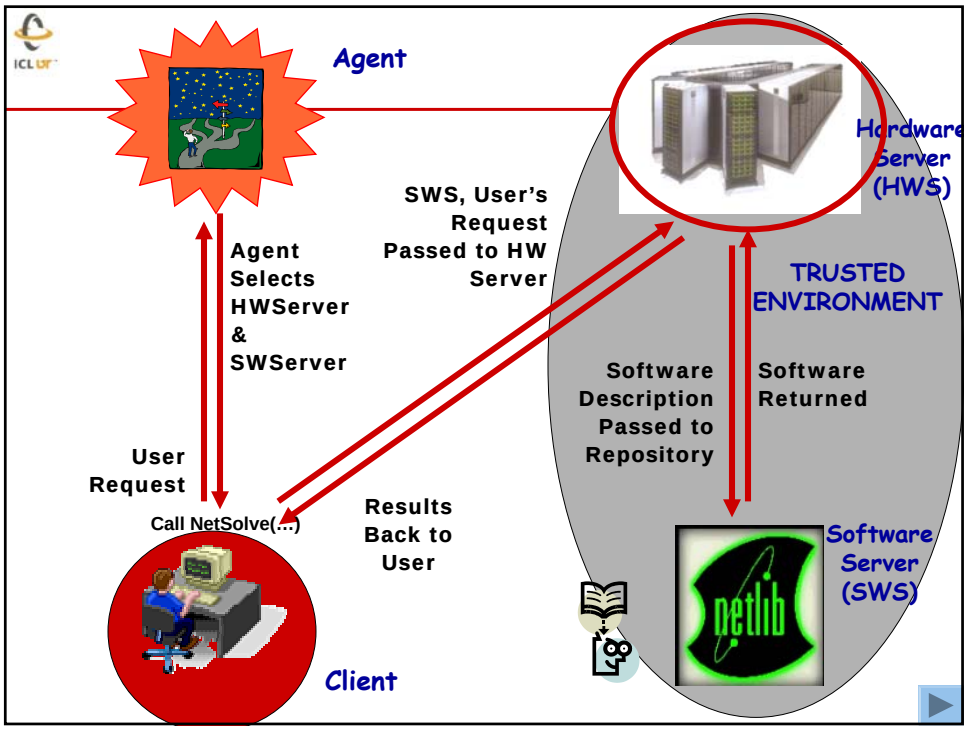


No knowledge of the grid required, RPC like. 30

NetSolve: The Big Picture

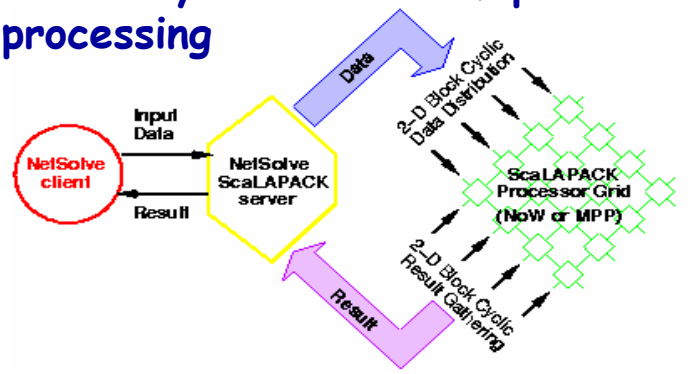


No knowledge of the grid required, RPC like. 31



Hiding the Parallel Processing

- User maybe unaware of parallel processing

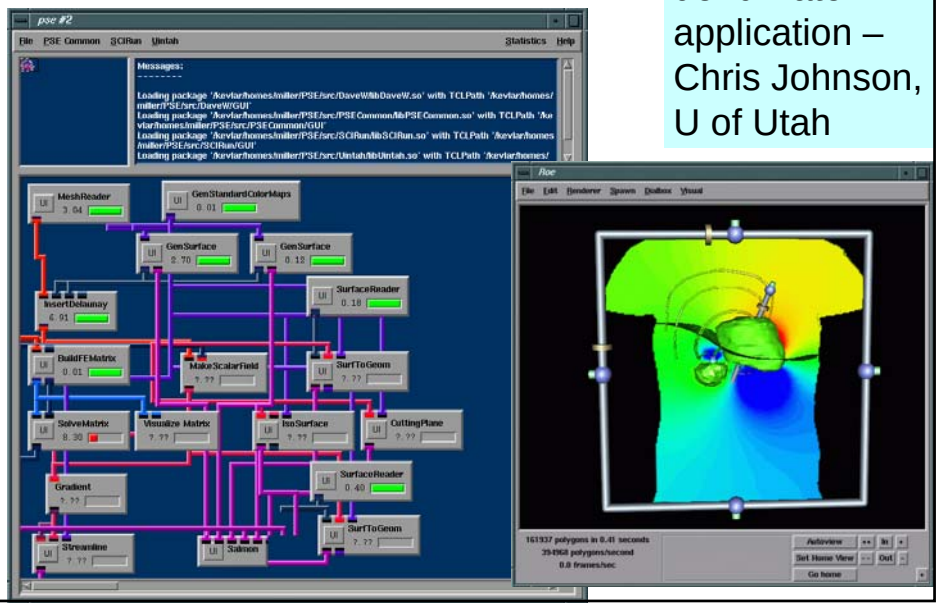


- NetSolve takes care of the starting the message passing system, data distribution, and returning the results. (Using LFC software)

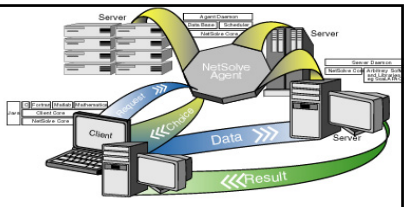
33

Netsolve and SCIRun

SCIRun torso defibrillator application – Chris Johnson, U of Utah



Basic Usage Scenarios

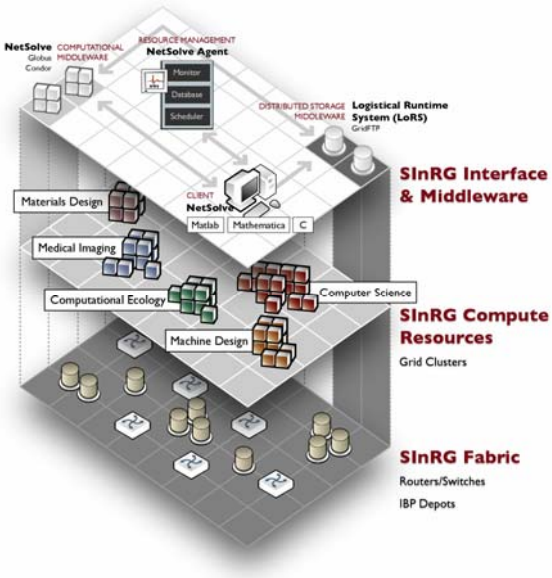


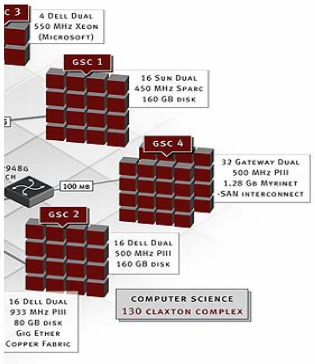
- Grid based numerical library routines
  - User doesn't have to have software library on their machine, LAPACK, SuperLU, ScaLAPACK, PETSc, AZTEC, ARPACK
- Task farming applications
  - "Pleasantly parallel" execution eg Parameter studies
- Remote application execution
  - Complete applications with 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, …

35

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

36

New Features for NetSolve 2.0

New version available!

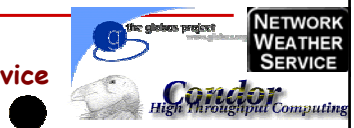
http://icl.cs.utk.edu/netsolve/

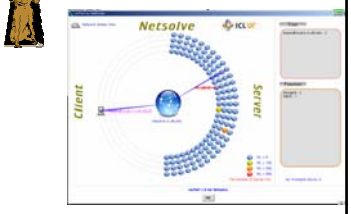
- New easy to use Interface Definition Language
  - Simplified PDF
- Dynamic servers
  - Add/delete problems without restarting servers
- New bindings for
  - GridRPC
  - Octave
  - Condor-G
- Separate hardware/software servers
- Support for Mac OS X & Windows 2K/XP
- Web based monitoring
- Allow user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

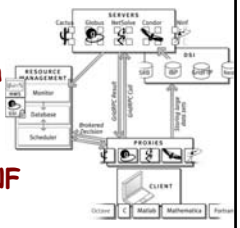
37

NetSolve- Things Not Touched On

- Integration with other NMI tools
  - Globus, Condor, Network Weather Service
- Security
  - Using Kerberos V5 for authentication.
- 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







38

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!*

39

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

40

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