Sparse Days and Grid Computing at St. Girons

Journées sur l'algèbre linéaire creuse et le calcul sur la grille.
Presentation succinte format ps format pdf

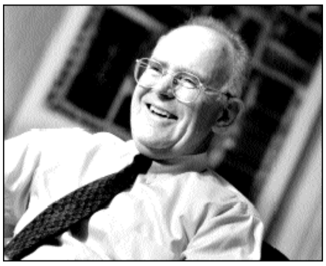
A four-day meeting on sparse matrix matters at St Girons in Ariege at foothills of Pyrenees, on June 10-13 2003.

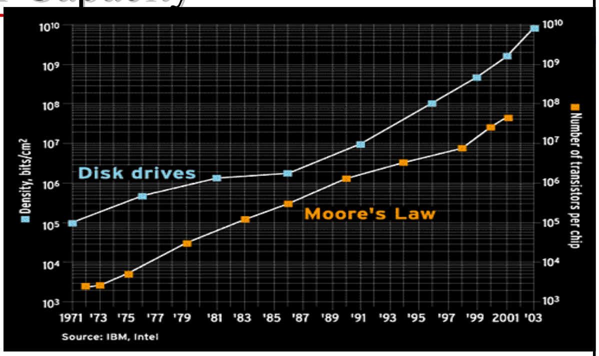
High Performance Computing and the Computational Grid

Jack Dongarra
University of Tennessee
and
Oak Ridge National Lab

June 12, 2003 1

Technology Trends: Microprocessor Capacity





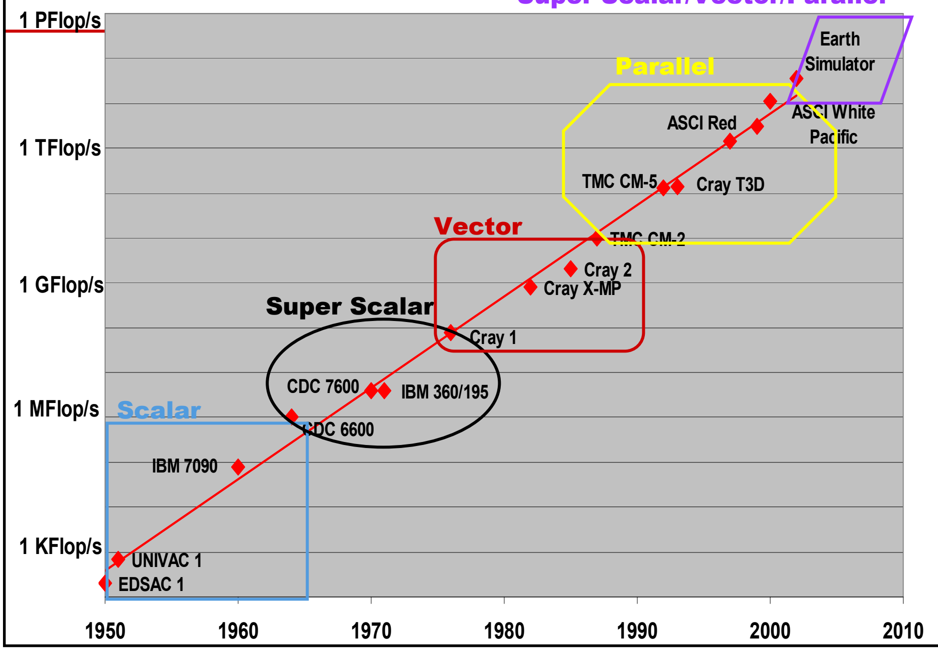
Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

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.

2

Moore's Law



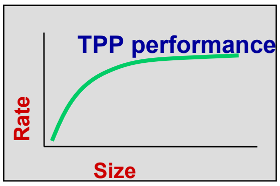
3

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



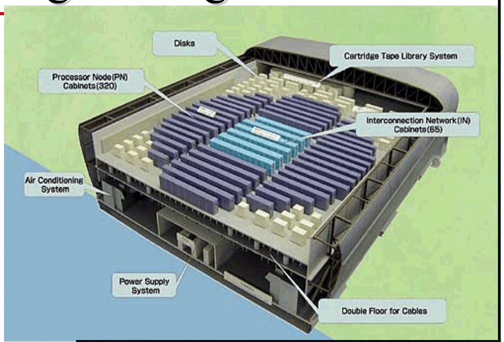
- Updated twice a year
  SC'xy in the States in November
  Meeting in Mannheim, Germany in June
- All data available from **www.top500.org**

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

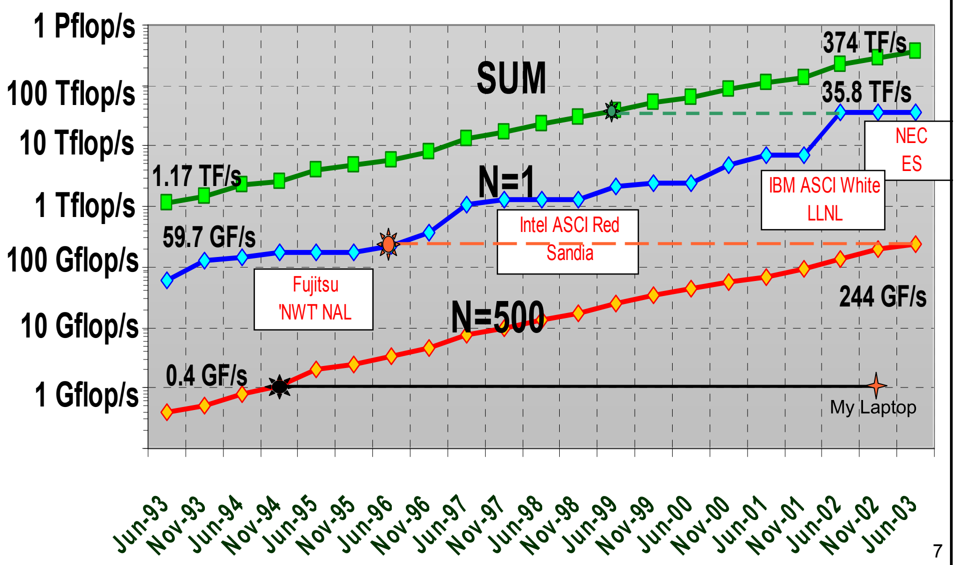
- From the Top500 (June 2003)
  - Performance of ESC ≈ Σ Next Top 4 Computers
  - ~ 10% of performance of all the Top500 machines



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

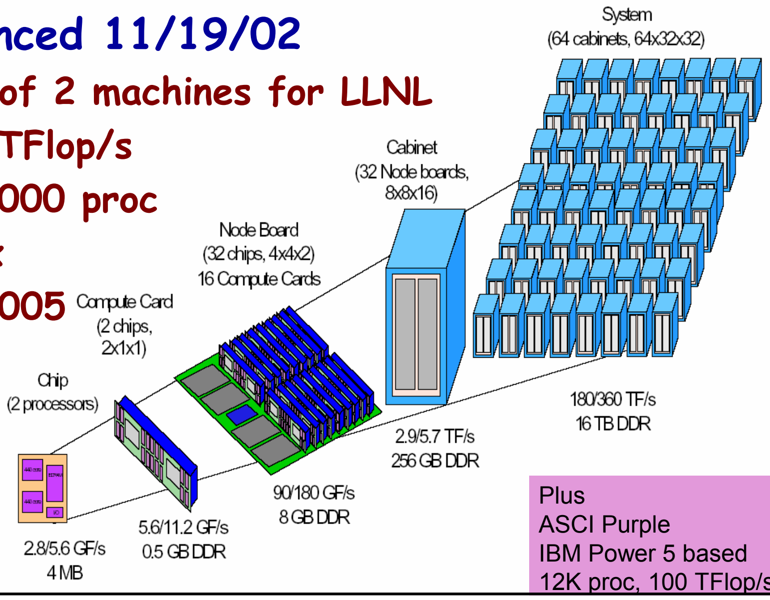
TOP500 – Performance - June 2003



7

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



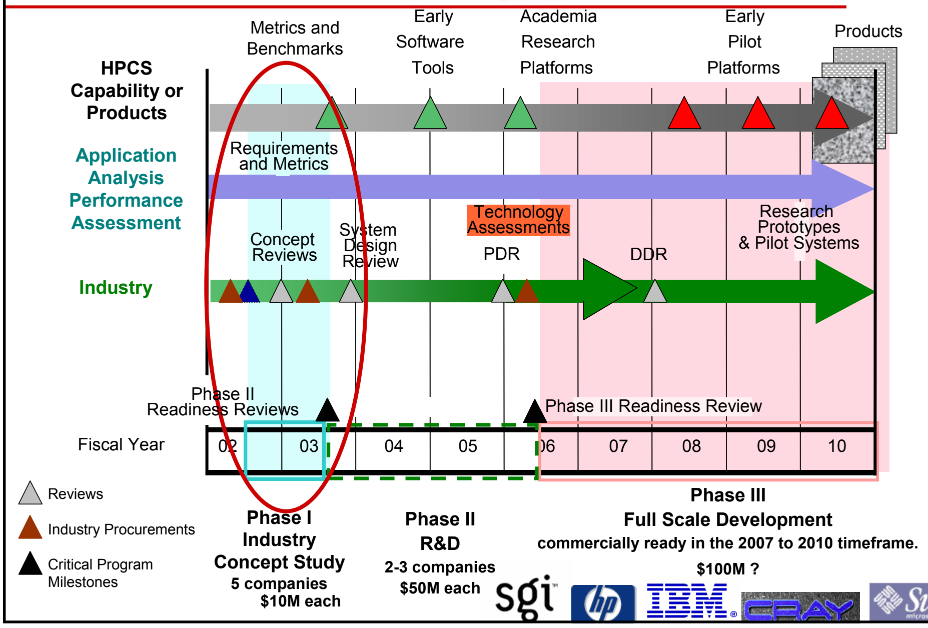
Plus
ASCI Purple
IBM Power 5 based
12K proc, 100 TFlop/s

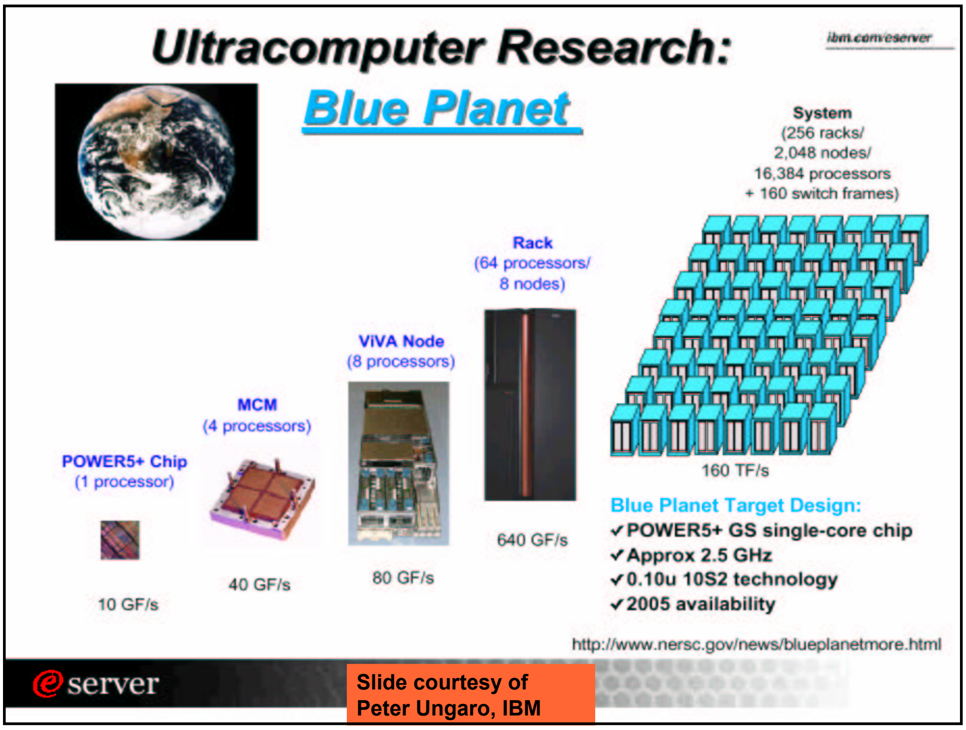
DOE ASCI
Red Storm Sandia National Lab

- 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

HPCS Phases I - III





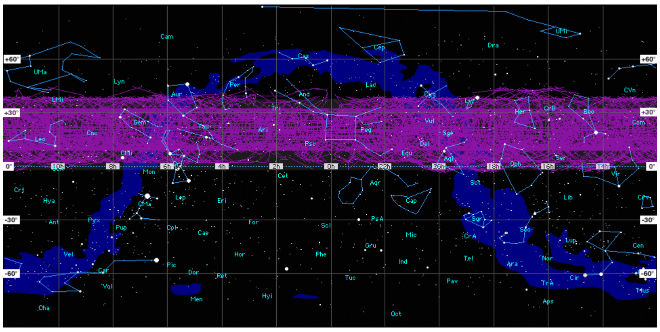
ORNL – September 2003

- 3.2 TFlops
- 256 processors
- 1 TB shared memory
- 32 TB of disk space
- 8 cabinets
- programming model
  - MPI, Co-Array Fortran, or Shmem

12

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



13

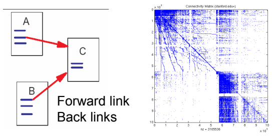
SETI@home

- 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.

14

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
  - $n=2.7\times10^9$ (see: Cleve's Corner)

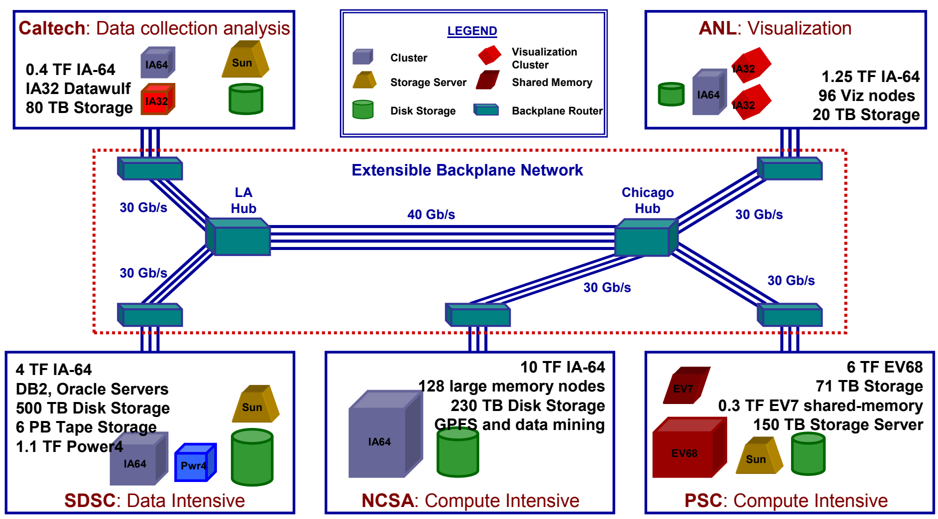


  - 1 if there's a hyperlink from page *i* to *j*
- Form a transition probability matrix of the Markov chain
  - Matrix is not sparse, but it is a rank one modification of a sparse matrix
- Largest eigenvalue is equal to one; want the corresponding eigenvector (the state vector of the Markov chain).
  - The elements of eigenvector are Google's PageRank (Larry Page).
- When you search: They have an inverted index of the web pages
  - Words and links that have those words
- Your query of words: find links then order lists of pages by their PageRank.

Source: Monika Henzinger, Google & Cleve Moler

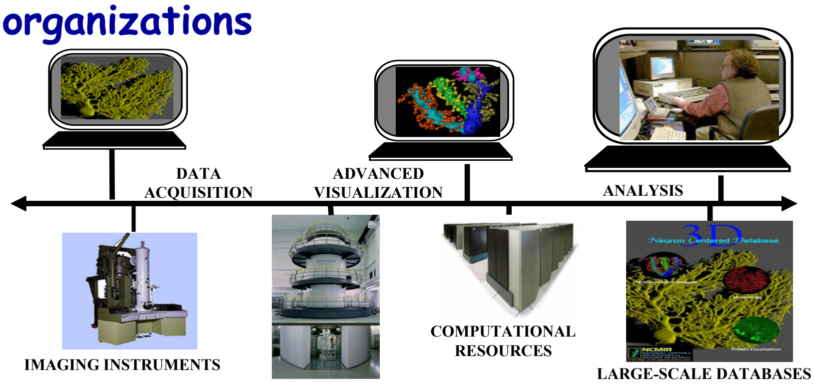
Extensible TeraGrid Facility (ETF)

Proposed 2002, Becoming operational



Grid Computing is About …

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

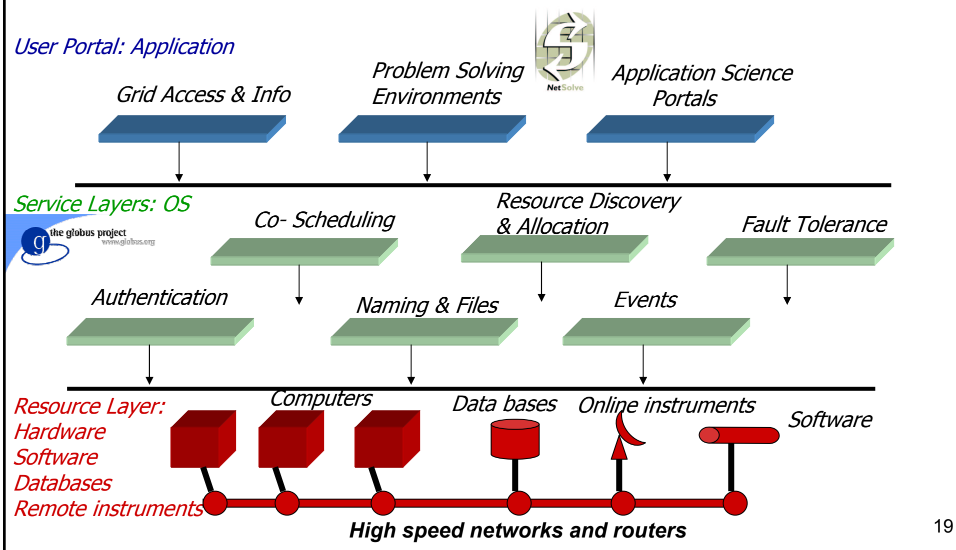


"Telescience Grid", Courtesy of Mark Ellisman

The Computing Continuum

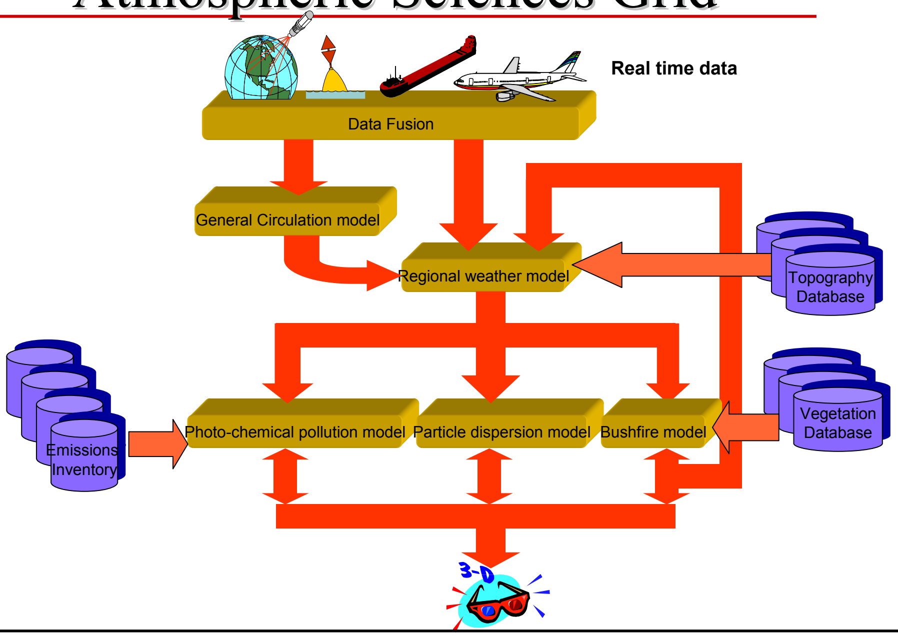
- Each strikes a different balance
  - computation/communication coupling
- Implications for execution efficiency
- *Applications for diverse needs*
  - *computing is only one part of the story!*

The Grid Architecture Picture



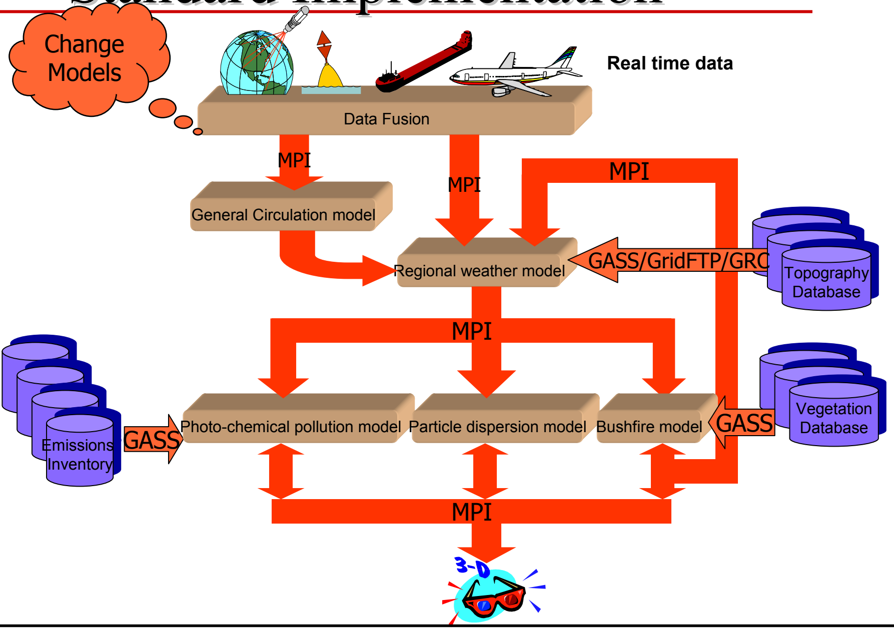
19

Atmospheric Sciences Grid



20

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

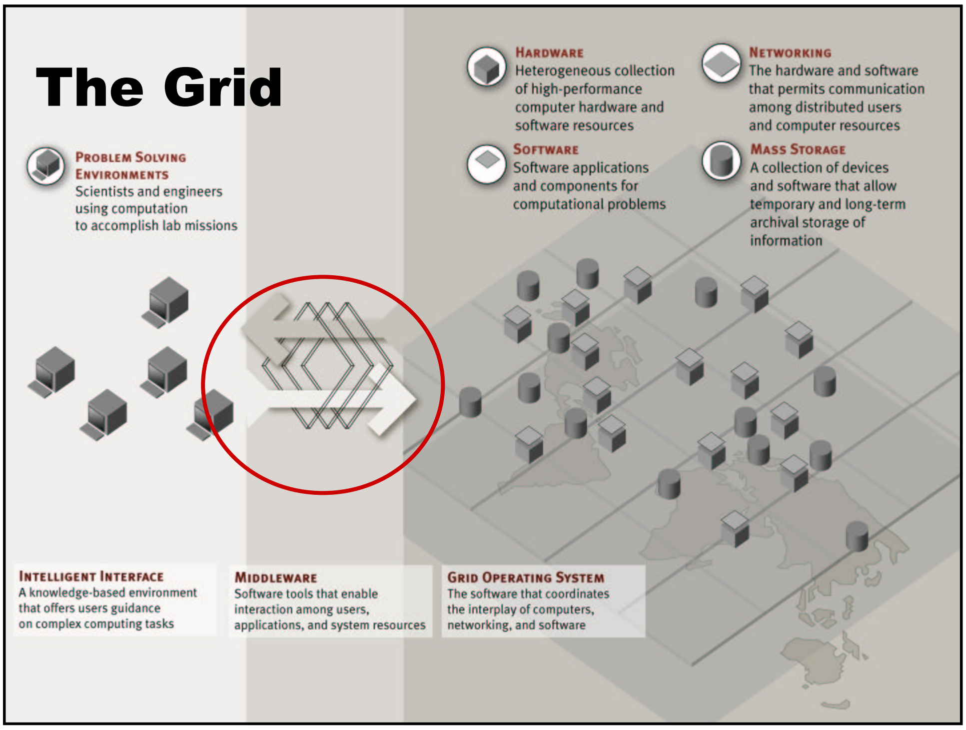
Some Grid Requirements – Systems/Deployment Perspective

- Identity & authentication
- Authorization & policy
- Resource discovery
- Resource characterization
- Resource allocation
- (Co-)reservation, workflow
- Distributed algorithms
- Remote data access
- High-speed data transfer
- Performance guarantees
- Monitoring
- Adaptation
- Intrusion detection
- Resource management
- Accounting & payment
- Fault management
- System evolution
- Etc.

Globus Grid Services

the globus project
www.globus.org

- The Globus toolkit provides a range of basic Grid services
  - Security, information, fault detection, communication, resource management, ...
- These services are simple and orthogonal
  - Can be used independently, mix and match
  - Programming model independent
- For each there are well-defined APIs
- Standards are used extensively
  - E.g., LDAP, GSS-API, X.509, ...
- You don't program in Globus, it's a set of tools like Unix

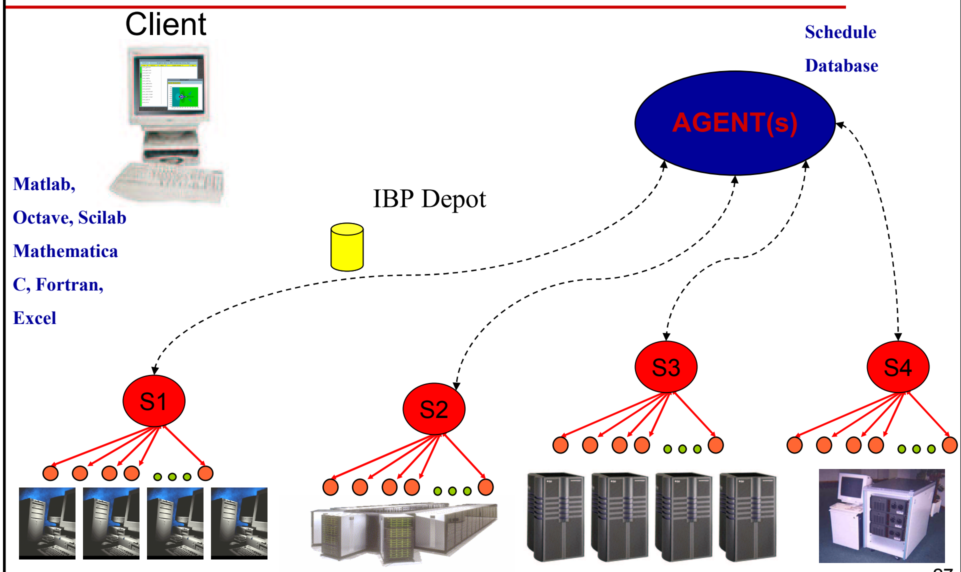


NetSolve Grid Enabled Server

- 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.

26

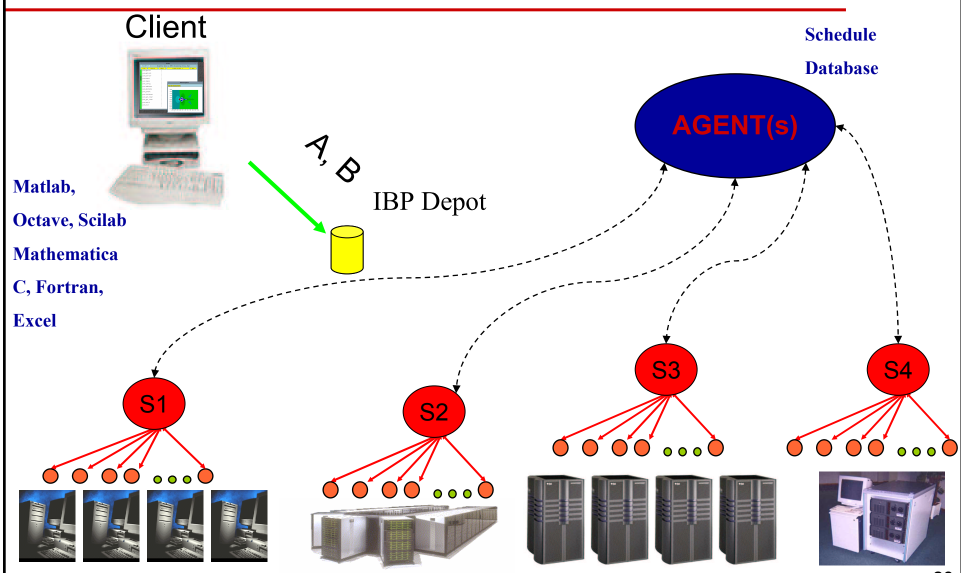
NetSolve: The Big Picture



No knowledge of the grid required, RPC like.

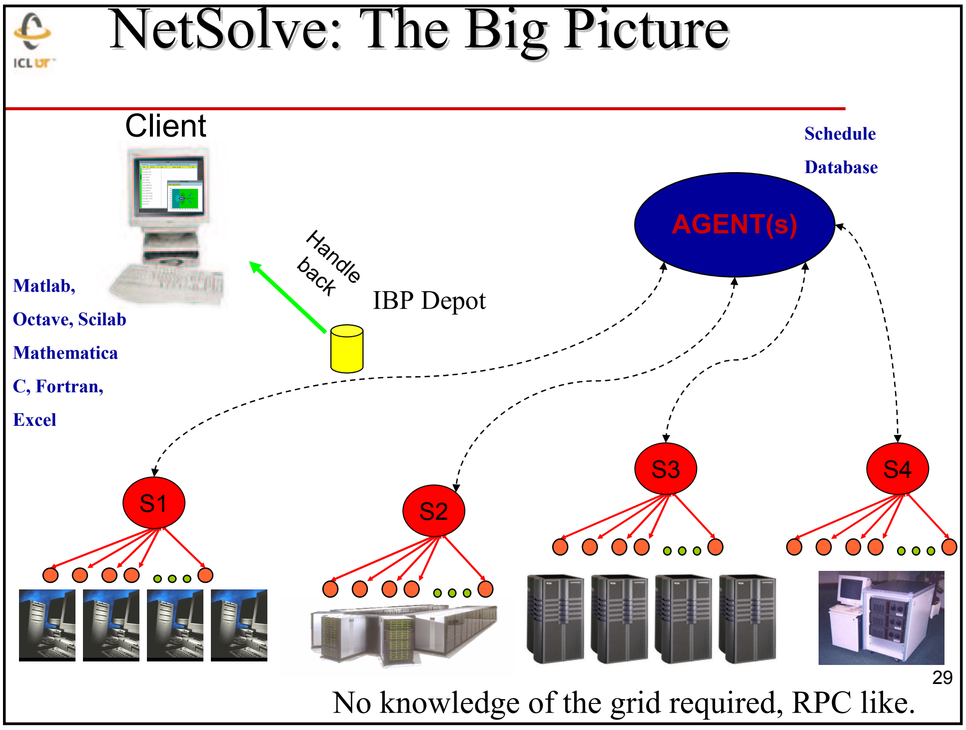
27

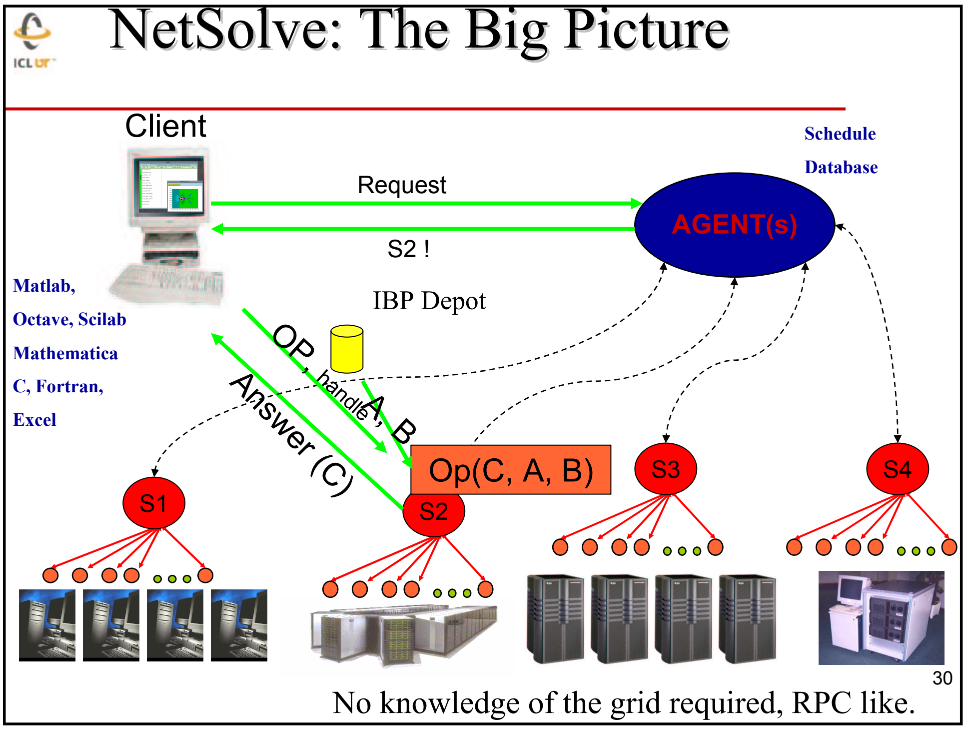
NetSolve: The Big Picture

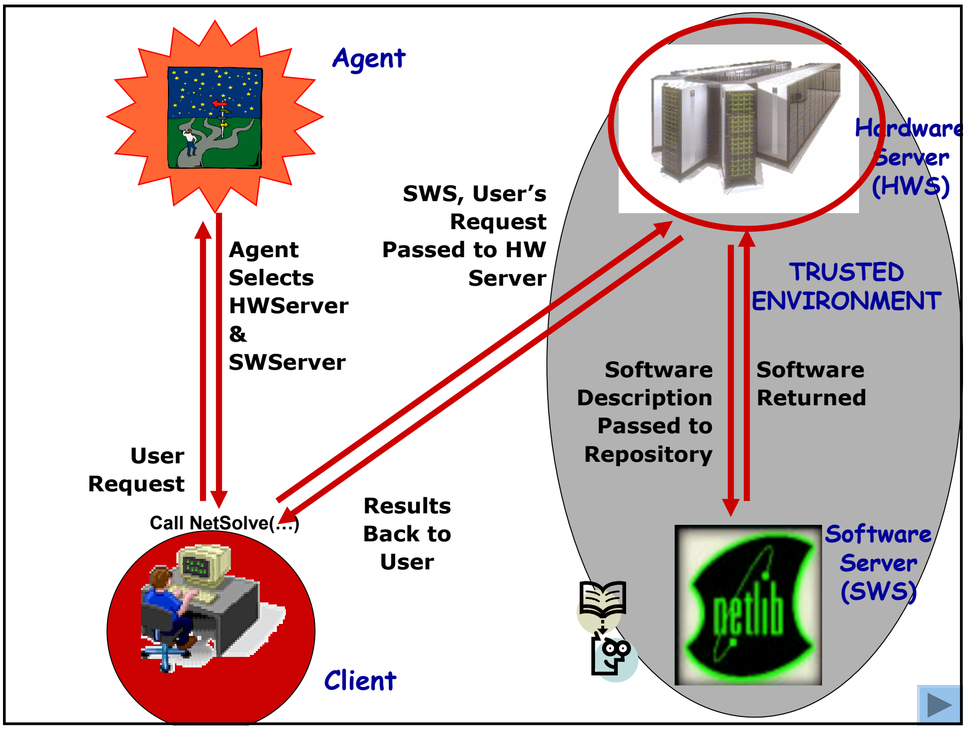


No knowledge of the grid required, RPC like.

28


NetSolve: The Big Picture
Client
Schedule
Database
AGENT(s)
Handle back
IBP Depot
Matlab,
Octave, Scilab
Mathematica
C, Fortran,
Excel
S1
S2
S3
S4
29
No knowledge of the grid required, RPC like.


NetSolve: The Big Picture
Client
Schedule
Database
Request
AGENT(s)
S2 !
IBP Depot
Matlab,
Octave, Scilab
Mathematica
C, Fortran,
Excel
OP, handle, A, B
Answer (C)
Op(C, A, B)
S1
S2
S3
S4
30
No knowledge of the grid required, RPC like.



NetSolve Agent

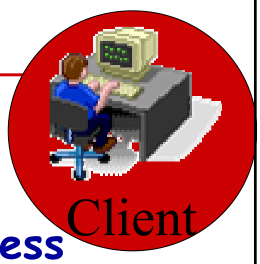
- Name server for the NetSolve system.
- Information Service
  - client users and administrators can query the hardware and software services available.
- Resource scheduler
  - maintains both static and dynamic information regarding the NetSolve server components to use for the allocation of resources

NetSolve Agent

- Resource Scheduling (cont'd):
  - CPU Performance (LINPACK).
  - Network bandwidth, latency.
  - Server workload.
  - Problem size/algorithm complexity.
  - Calculates a "Time to Compute." for each appropriate server.
  - Notifies client of most appropriate server.

33

NetSolve Client

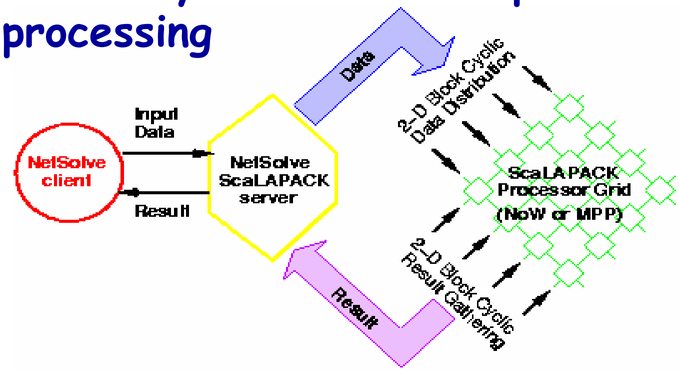


- Function Based Interface.
- Client program embeds call from NetSolve's API to access additional resources.
- Interface available to C, Fortran, Matlab, Octave, Mathematica, …
- Opaque networking interactions.
- NetSolve can be invoked using a variety of methods: blocking, non-blocking, task farms, …

34

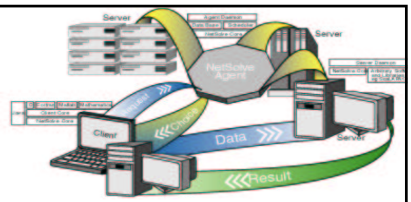
Hiding the Parallel Processing

- User maybe unaware of parallel processing



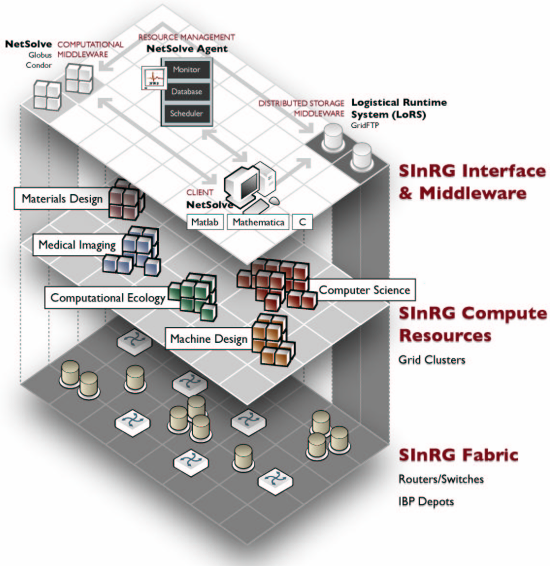
- NetSolve takes care of the starting the message passing system, data distribution, and returning the results.

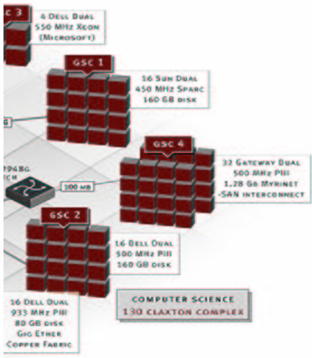
Basic Usage Scenarios



- 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



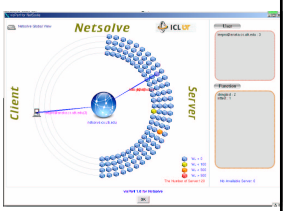


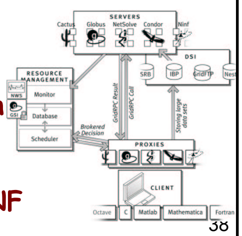
- **Federated Ownership**: CS, Chem Eng., Medical School, Computational Ecology, El. Eng.
- **Real** applications, middleware development, logistical networking

37

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



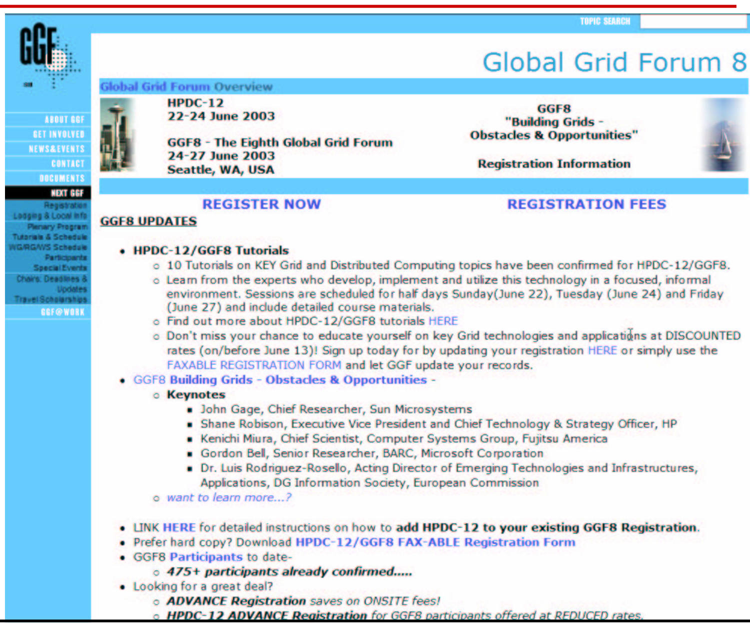


38

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

If You Want to Participate …



40

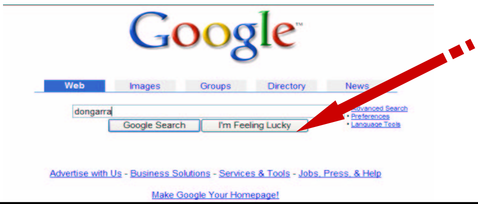
Futures for Numerical Algorithms and Software

- **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
- **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.**

41

Collaborators / Support

- **TOP500**
  - H. Mauer, Mannheim U
  - H. Simon, NERSC
  - E. Strohmaier, NERSC
- NetSolve
  - Sudesh Agrawal, UTK
  - Henri Casanova, UCSD
  - Keith Seymour, UTK
  - Sathish Vadhiyar, UTK
- **For more information…**



- Thanks

Next Generation Software

SciDAC
Scientific Discovery through Advanced Computing

Many opportunities within my group at Tennessee

42

43