



Grid Computing

GS1640

Raúl Soto

Joel Maysonet

Lucette Sánchez

Fernando González

What is Grid Computing?

- A **Grid Computing system** is a collection of distributed computing resource available over a local or wide area network, that appears to an end user or application as **one large virtual** computing system
- Is an approach to distributed computing that spans not only locations but also **organizations, machine architectures,** and **software boundaries**

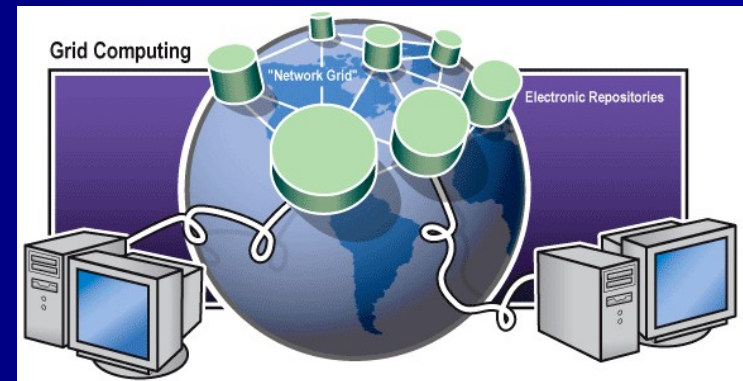


Characteristics

- Allows the integrated, collaborative use of computers, networks, databases, and scientific instruments owned and managed by multiple organizations
- Facilitates the solution of **computational problems**
 - large-scale
 - complex
 - multi-institutional
 - multidisciplinary
 - Large data storage and/or computational requirements
- Grid computing started out as the simultaneous application of the resources of many networked computers to a **single** (usually scientific) problem.
 - E.G. : *SETI @ Home, Human Proteome project, Anthrax research, Smallpox project, Cancer research project, etc.*

Characteristics

- For many years, computational grids have been used to solve large-scale problems in **science** and **engineering**
- **Currently** used in the following fields :
 - Medical research : protein folding, cancer drug development
 - Astronomy : SETI data analysis
 - Mathematical / Statistical problems
 - Climate models
- Grid computing is beginning to enter the **commercial** world
 - Financial analysis
 - Forecasting
 - Enterprise Grids



Characteristics

- Involves sharing of heterogeneous resources:
 - Hardware platforms
 - Hardware / software architectures
 - Computer languages
 - Different places
 - Different administrative domains

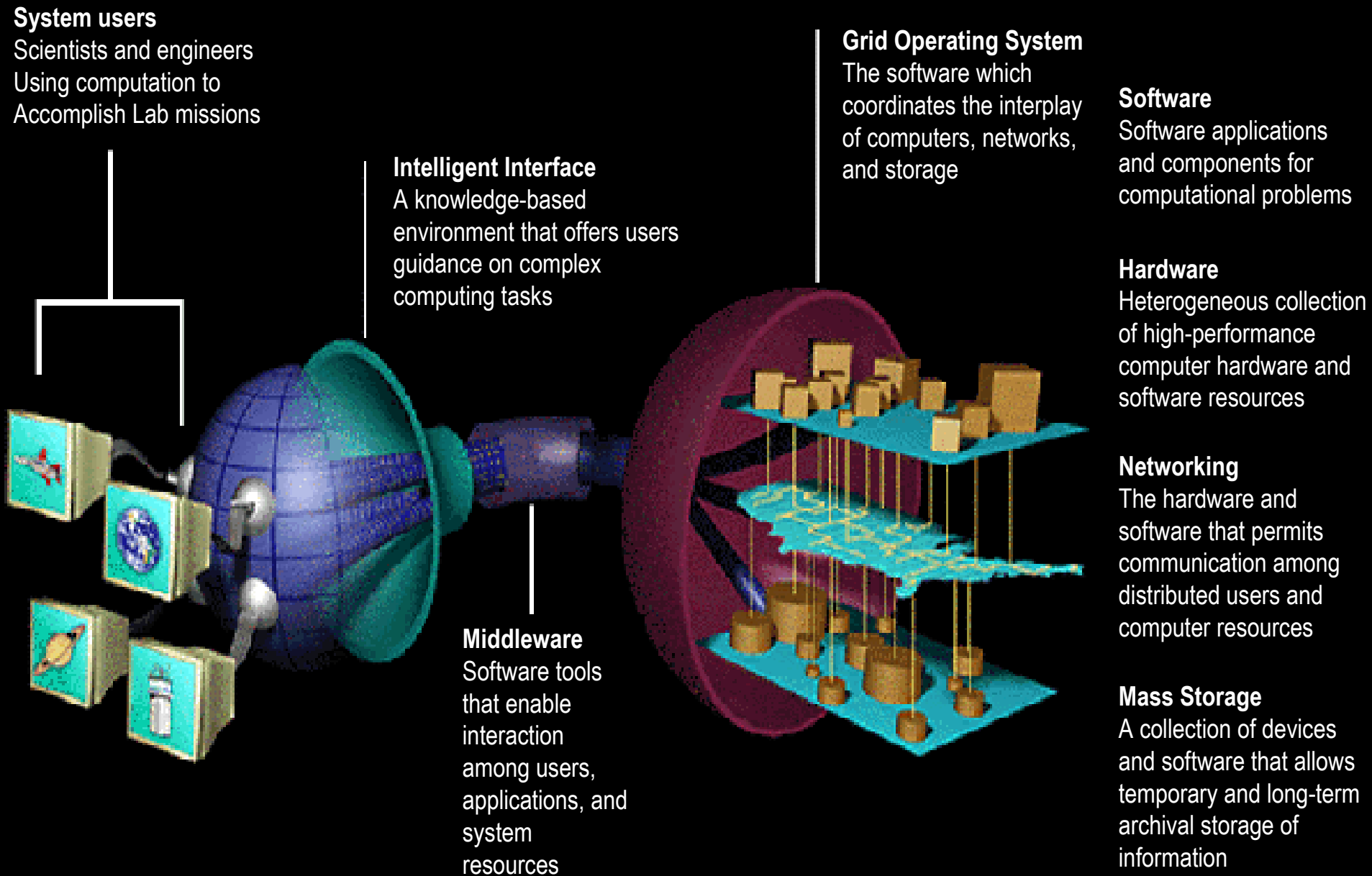


Main Classifications



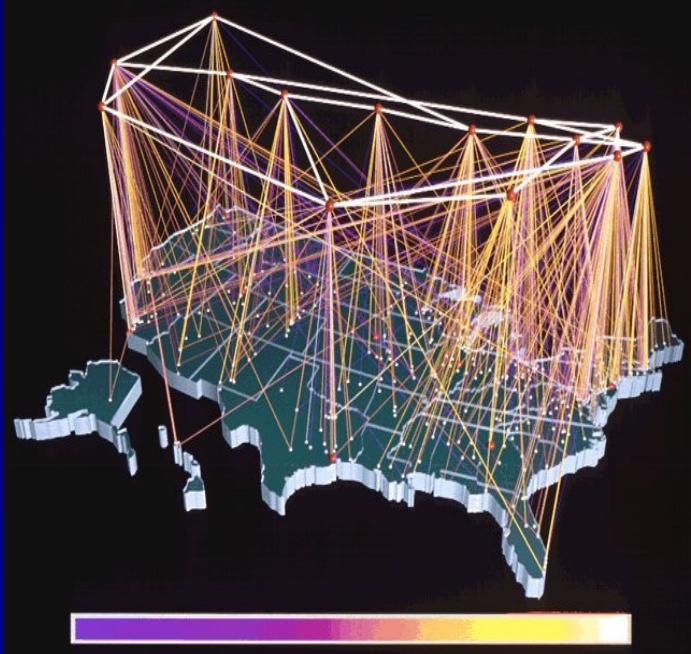
- Grid Types:
 - **Computational Grids** : computers set aside resources allocated to number-crunch data or provide coverage for CPU-intensive workloads
 - **Data Grids** : share data resources and storage capacity, unified interface for all data repositories in an organization, through which data can be queried, managed, and secured
 - **Scavenging Grids** : used to locate and exploit machine cycles on idle servers and desktops for use in resource-intensive tasks
- Internal vs External Grids
 - **External grids** : usually geographically-distributed, non-profit research efforts
 - **Internal grids** : large commercial enterprise with complex problems who aim to fully exploit their unused internal computing power

How does it work?



Grid Computing vs the Internet

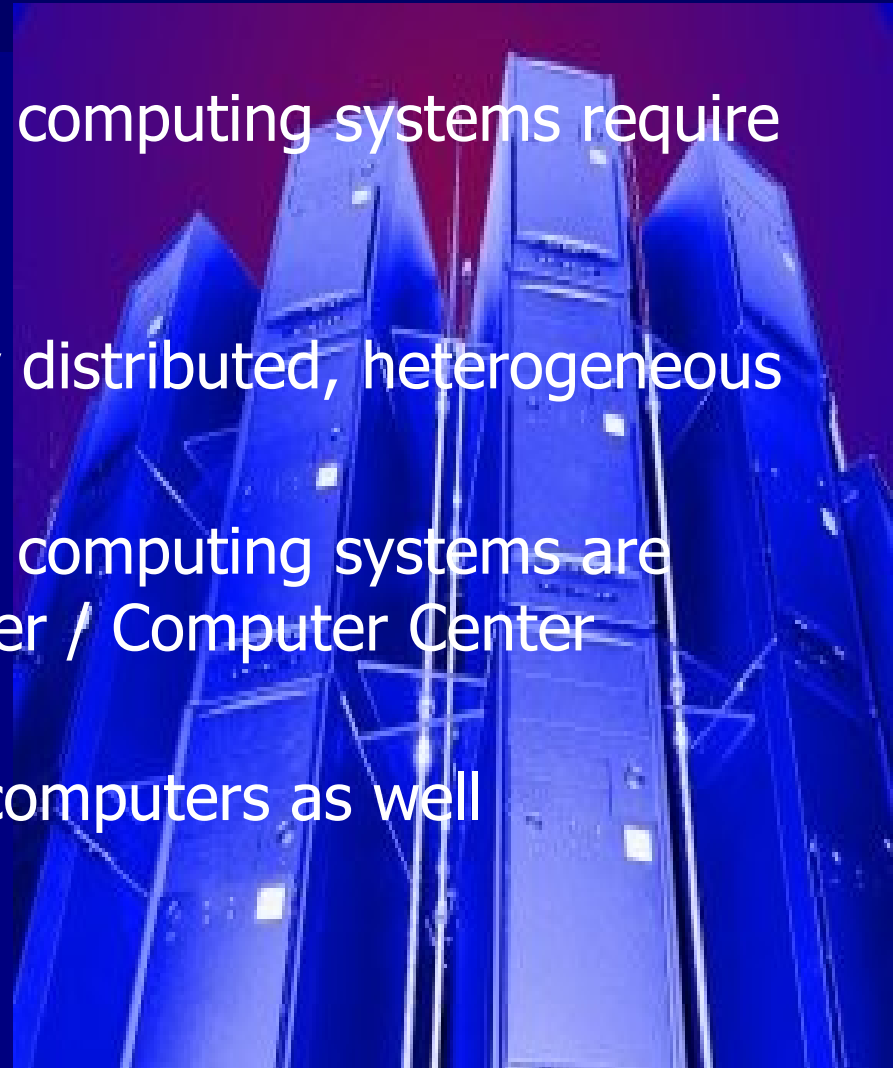
- The Internet is about getting computers to **talk** together
- Grid computing is about getting computers to **work** together



- The Internet is a network of **communication**
- Grid computing is a network of **computation**: provides tools and protocols for resource sharing of a variety of IT resources

Grid Computing vs. Clusters / Distributed Computing

- Clusters and Distributed computing systems require
 - physical proximity
 - Operational homogeneity
- Grids are geographically distributed, heterogeneous
- Clusters and Distributed computing systems are based on the Data Center / Computer Center computers
- Grids include end-user computers as well

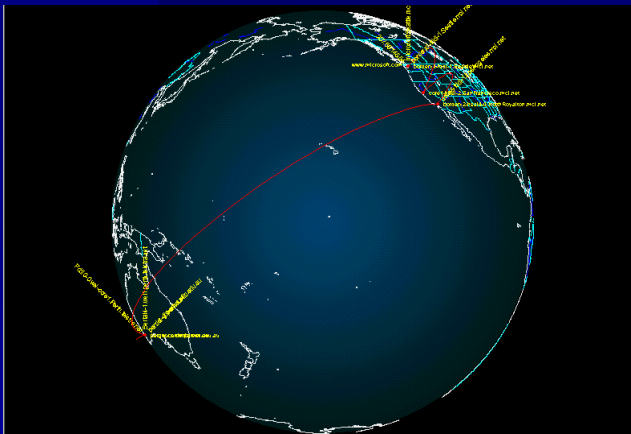
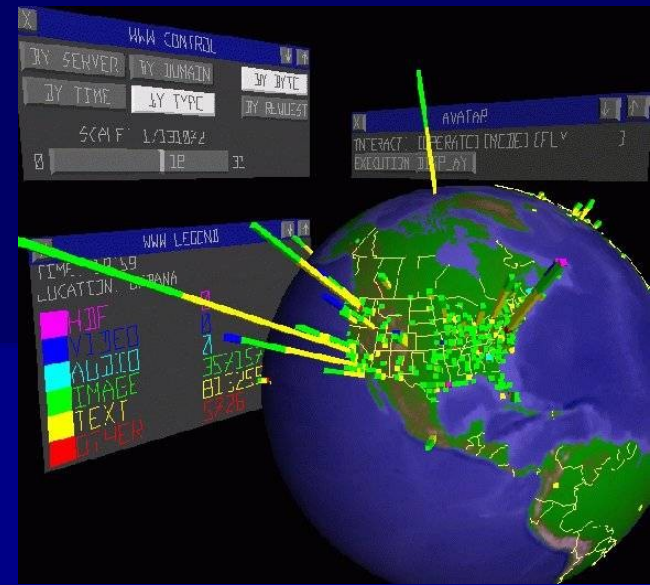


Analogy: Electric Power Grid

- Electric Power:
 - Household electrical devices simply plug to an electric outlet
 - Use **only** the electric power you need
 - Pay **only** for the electric power you used
- Computers – typical :
 - You pay for a computer with certain computing power (CPU flops)
 - If you use less computing power than what your computer provides, **you still pay for all** of it
 - If you use more computing power than what your computer provides, **you have to buy a better computer**
- Computers – Grid :
 - Multiple computers (**including yours**) share computing resources and make up a virtual computer
 - You only use the computing power you need

Grid Computing

- Computing power cost
 - 1980: \$100,000 per megaflop
 - 2000: \$1 per megaflop
- Computing power evolution
 - 1986: US National Science Foundation resources were five (5) **Cray XM-P supercomputers**
 - 2000: that's the equivalent computing power of ONE (1) **Nintendo 64 console**



Applications

- Distributed supercomputing / computational science
- High – capacity / throughput computing: large-scale simulation, chip design, and parameter studies
- Content sharing: for example, sharing digital content among peers
- Remote software access / renting

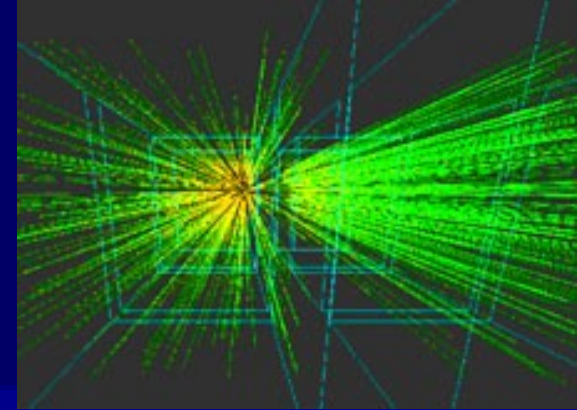


Applications

- Data-intensive computing: drug design, particle physics, stock prediction
- On-demand, real-time computing: medical instrumentation, mission critical initiatives
- Collaborative computing (e-science, e-engineering), collaborative design, data exploration



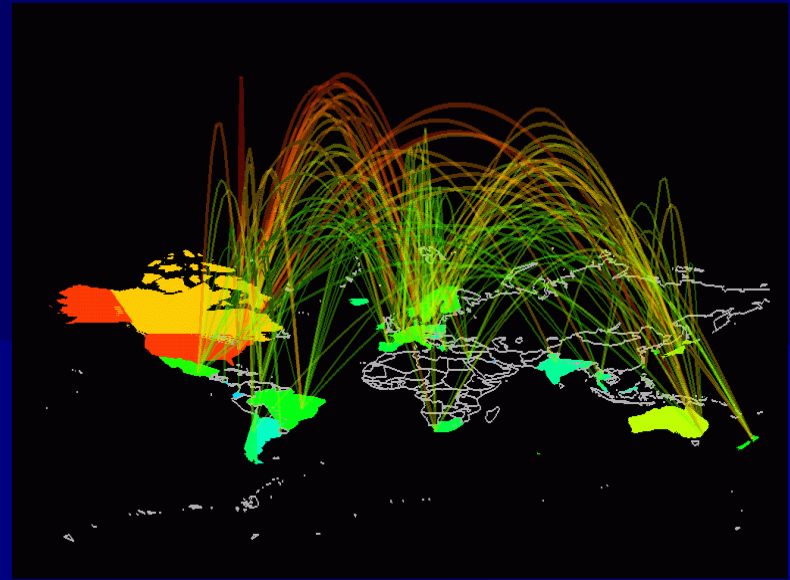
Benefits



- Permits sharing of resources throughout an organization, or among organizations
- Make effective use of **underused** computing resources
- Provide access to remote databases and software
- Reduce significantly the number of servers needed (25-75%)
- Allow **on-demand aggregation of resources** at multiple sites
- Reduce execution time for large-scale data processing applications

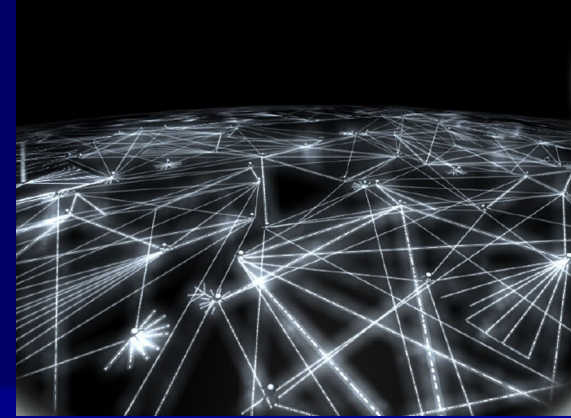
Benefits

- Provide load sharing across a set of platforms
- Provide **fault tolerance**
- Take advantage of time-zone and random diversity (in peak hours, users can access resources in off-peak zones)
- Provide flexibility to meet unforeseen emergency demands: can rent external resources for a required period instead of buying additional capacity
- Virtual data centers

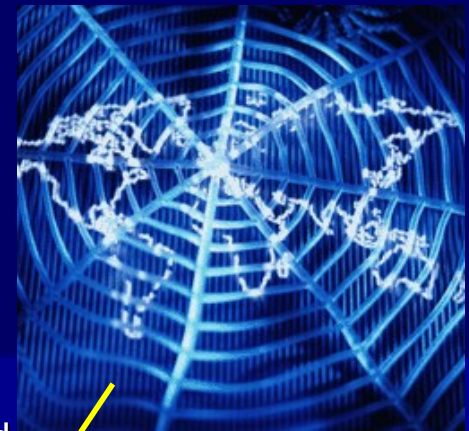


Disadvantages

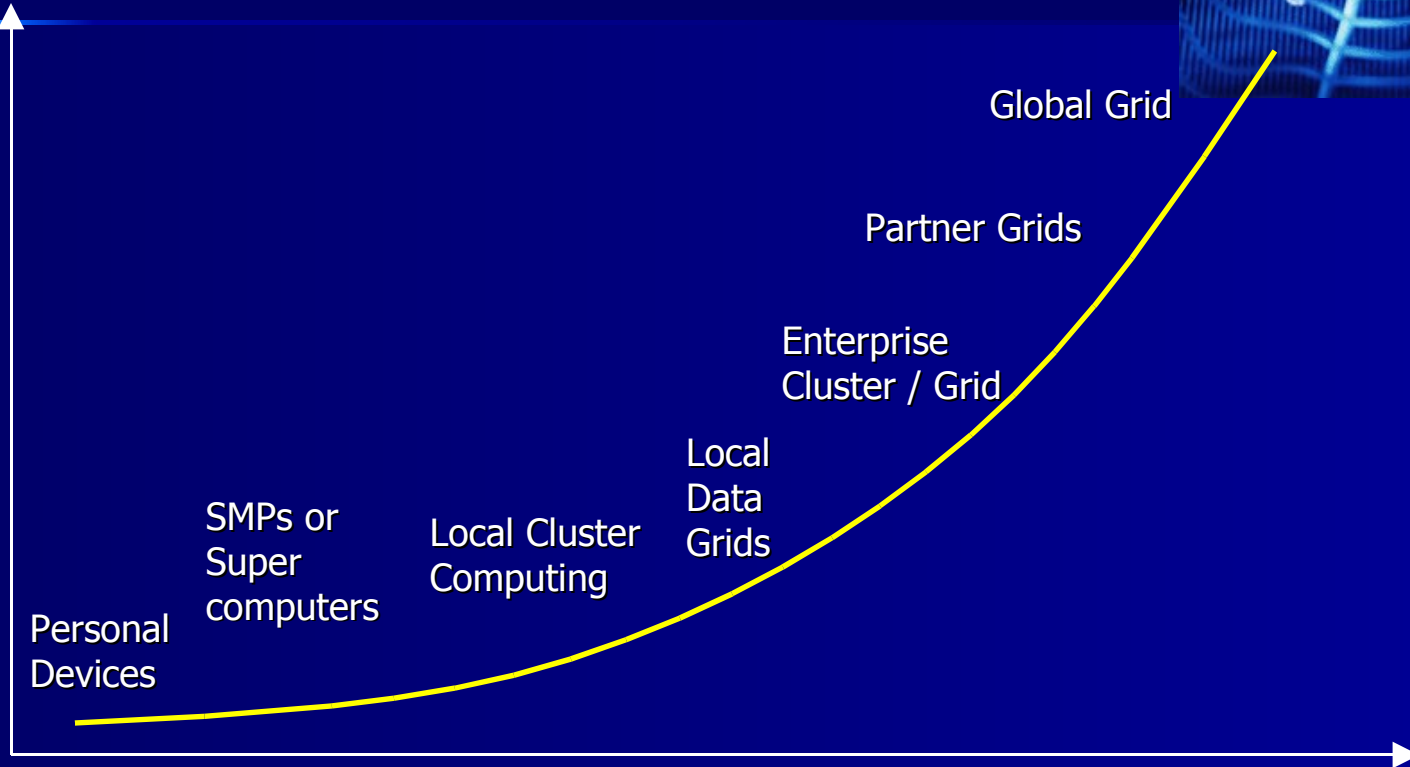
- Proprietary approaches
 - Leading vendors (HP, Sun, IBM, MS, Oracle, etc.) have proprietary, **incompatible** approaches,
 - which **defeats the purpose** of Grid computing
- Business case not always easy to sell to upper management
 - Need to present a business case based on economics, efficiency, not technical details
- Vendors need to show how their software effectively manages a grid environment
- **Security**
 - Confidentiality, Integrity, Access to resources, data
- Performance monitoring
 - Zoning mechanisms to ensure applications competing for resources do not affect each other



Evolution of Grids



Performance
and QoS



Early Stage
1990s

Second Stage
Early
2000s Mid
2000s

Third Stage
Late 2000s

Enterprise Grids Example : AstraZeneca PLC Grids

■ Data Grid

- Connects R&D databases from sites in UK, Sweden, and USA
- Significant savings in finding information
- Efficiency gains due to shortening the time R&D or design staff needs to find information
- Large investment in broadband links to connect data centers in different countries



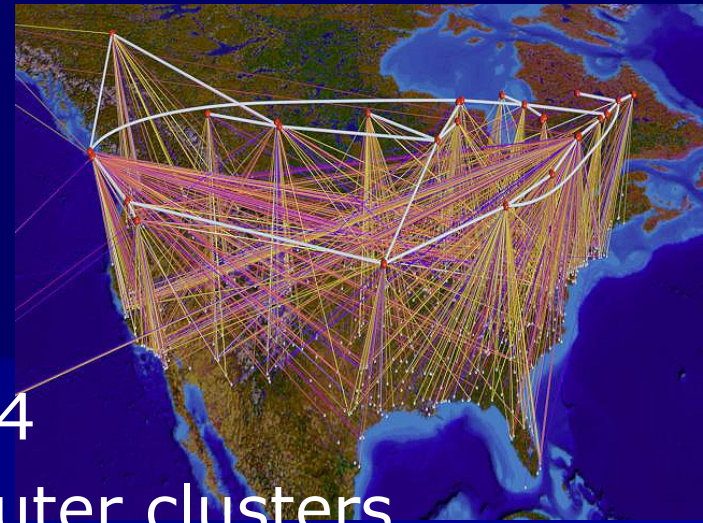
■ Enterprise Grid

- Processing power shared between R&D sites in UK, Sweden, and USA
- Efficiency due to processing power sharing, plus access to data
- Savings on R&D time, time to market
- Permits more efficient collaboration between sites
- Significant investment in security, and in high-performance broadband links

Some Examples of Grids

Grid Name	Sponsor	Purpose
BlueGrid	IBM	IBM computation R&D
DISCOM	Sandia National Labs	Defense research
DOE Science Grid	DOE Office of Science	Scientific research
European Union DataGrid	European Union	Scientific research
EuroGrid GRIP	European Union	Computation R&D
Globus Project	DARPA, NASA, Msoft, others	Grid tech research
GridLab	European Union	Grid tech research
Grid Research Integration	National Science Foundation	Grid middleware developmt
Intern. Data Grid Lab	National Science Foundation	International large scale grid tech research
Information Power Grid	NASA	Aerospace research
Earthquake Eng. Simulations	National Science Foundation	Earthquake engineering
Particle Physics Data Grid	DOE Science	High-energy physics research
TeraGrid	National Science Foundation	Link major US universities
UK Grid Support Center	UK eScience	Grid projects in UK

TeraGrid









(www.teragrid.com)

- Completed in September 2004
- Massively parallel supercomputer clusters
- 40 teraflops of computing power
- 2 petabytes of rotating storage
- Connected network of US supercomputing centers (currently 8, and growing)
- Each of the four original sites operates a Linux cluster, interconnected by means of a 10-30 Gigabit/sec dedicated optical network

TERAGRID

CALTECH: Data Collection Analysis

-  0.4 TF Intel IA-64
-  80 TB Online Disk
-  IA-32 Datawulf
-  Sun Storage Server

-  Clusters
-  Storage Server
-  Online Distributed Storage
-  Visualization Cluster
-  Shared Memory
-  Large-Scale Routers

ANL: Visualization

-  1.25 TF Intel IA-64
-  20 TB Online Disk
-  96 Visualization Nodes

LA Hub

40 GB/sec Extensible Backplane Network

Chicago Hub




SDSC: Data-Intensive

-  4 TF Intel IA-64
-  1.1 TF IBM Power4
-  500 TB Online Disk
-  Sun Disk Server
-  IBM Database Server

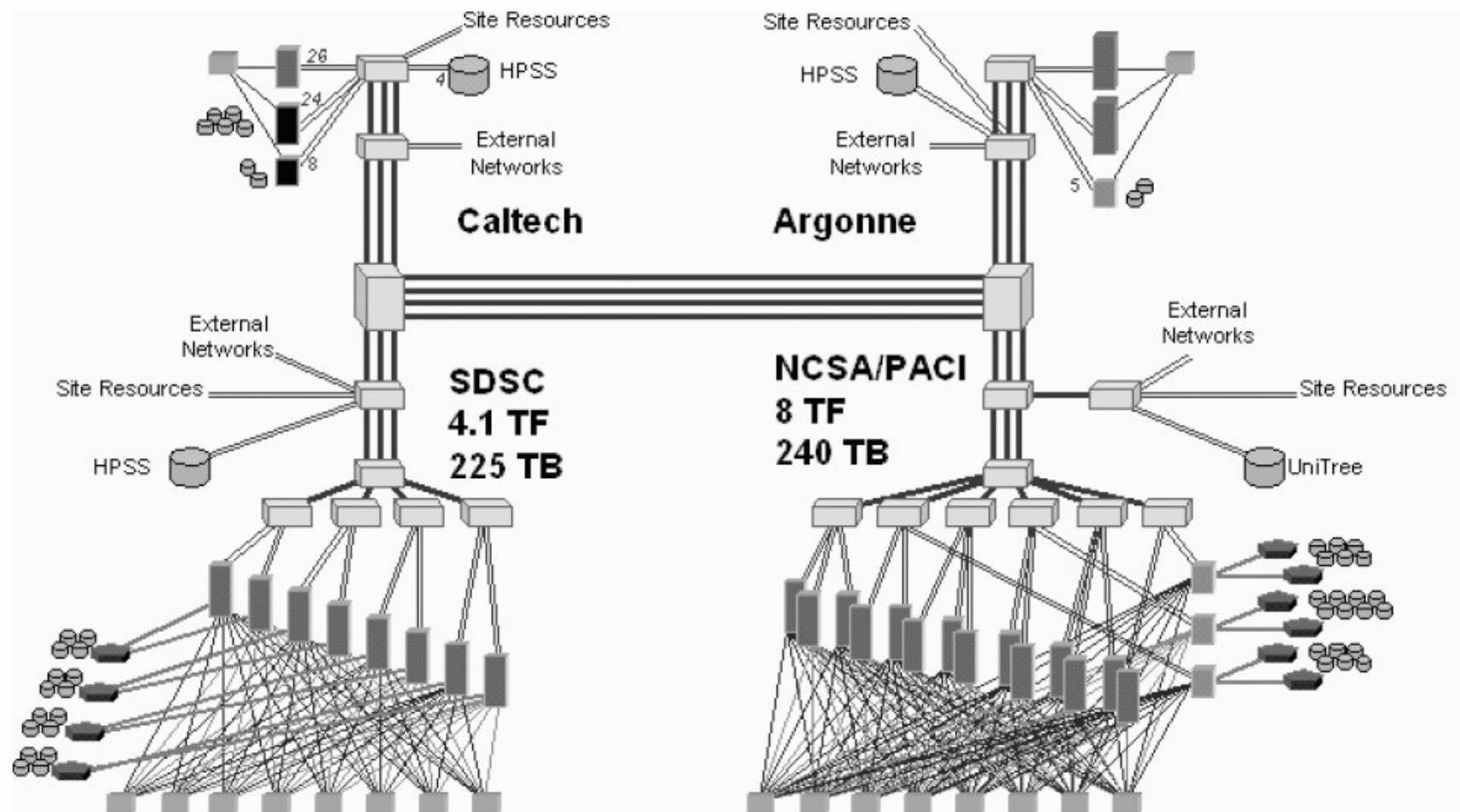
NCSA: Compute-Intensive

-  10 TF Intel IA-64
(128 Large-Memory Nodes)
-  230 TB Online Disk

PSC: Heterogeneity

-  6.3 TF Compaq EV7
-  221 TB Online Disk
-  Disk Cache Server

TeraGrid

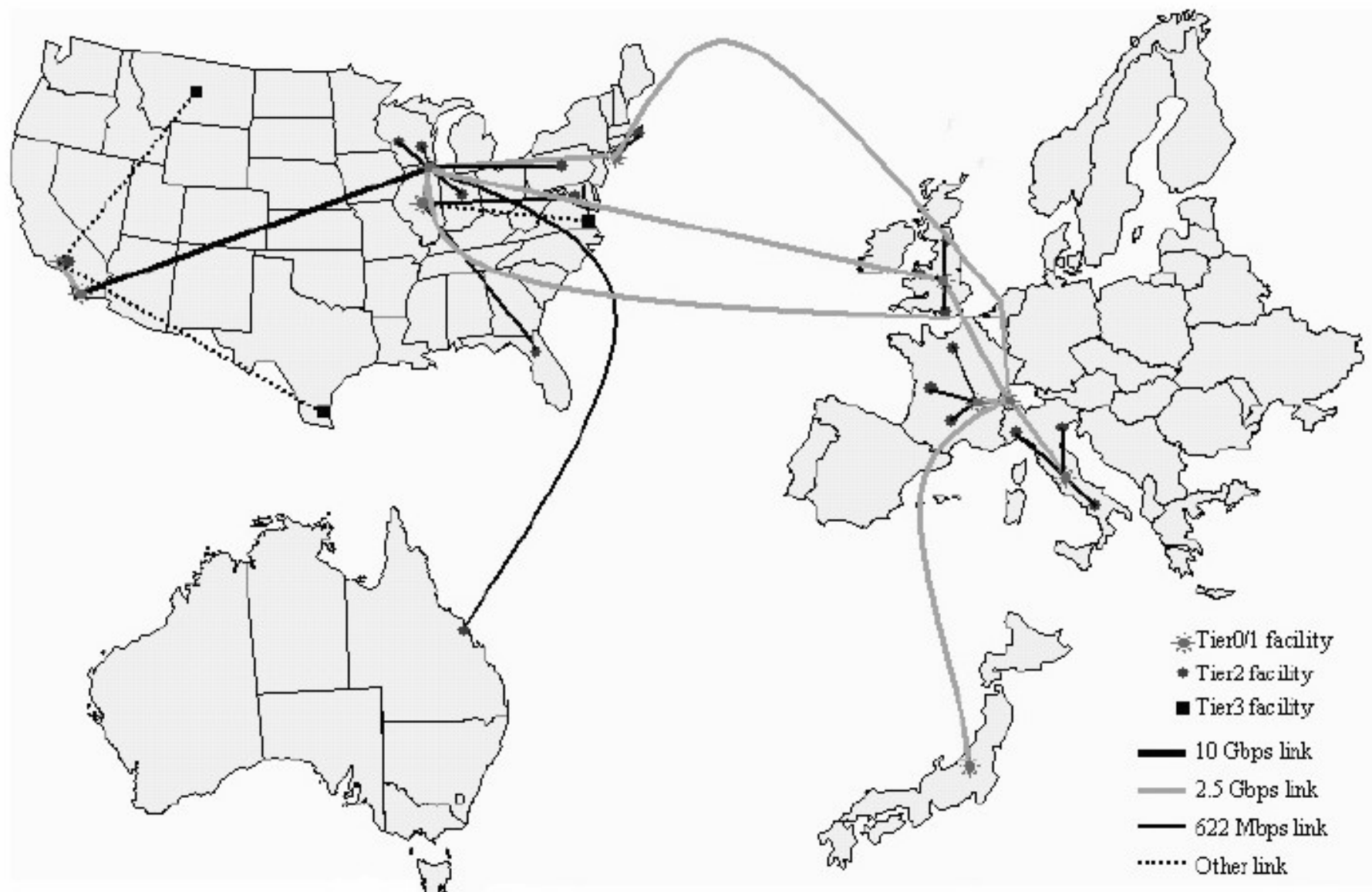


TeraGrid connections



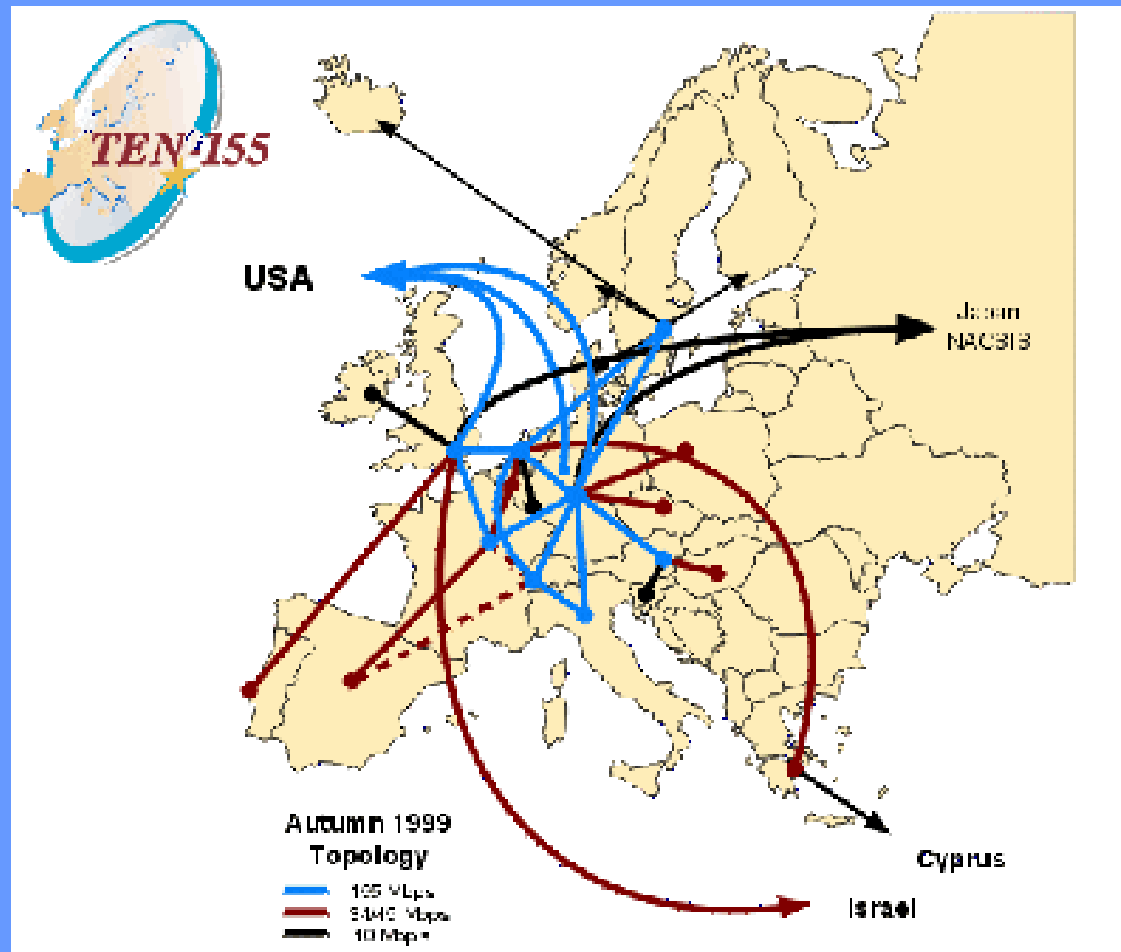
Other Grid Projects

iVDGL



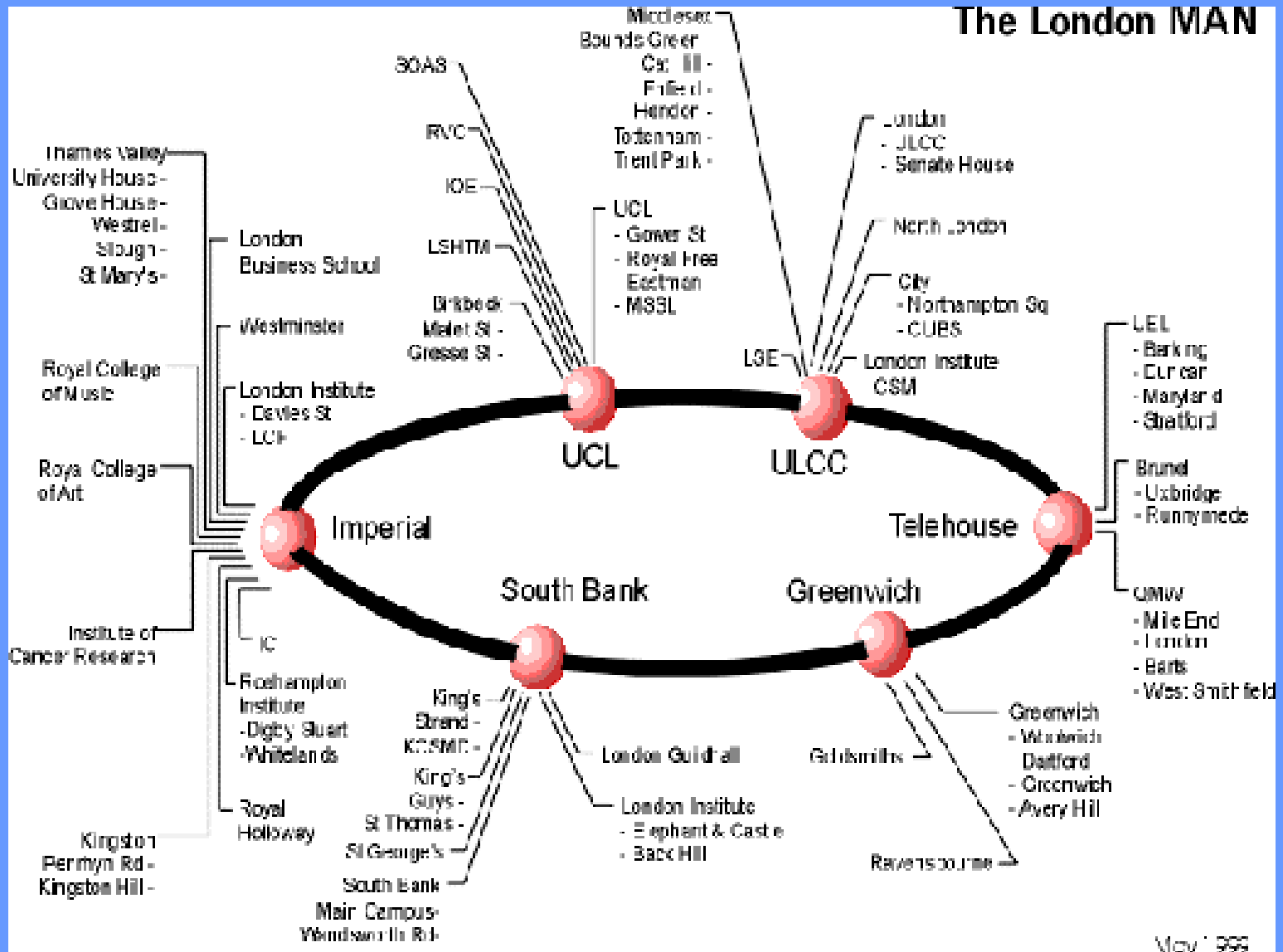
Other Grid Projects

European Research Network



Other Grid Projects

London Metropolitan Area Network



View: 333



Pauá Grid Project

HP Brazil Sponsored Grid



Introduction

- *Pauá* means "everything" in ancient native Brazilian tribes language (*Tupi*)
- A federation made by many research centers that have relationship with HP Brazil
- Each research center contributes with computational resources and specific grid technology expertise



Motivation

- Creating a geographically distributed Grid in Brazil
- Fostering existing grid knowledge in current HP Brazil projects
- Attracting projects from other knowledge areas that need high performance computing

Solution

- MyGrid/OurGrid based solution
- Sharing dedicated/non-dedicated resources
- *Pauá* federation regulates grid policies

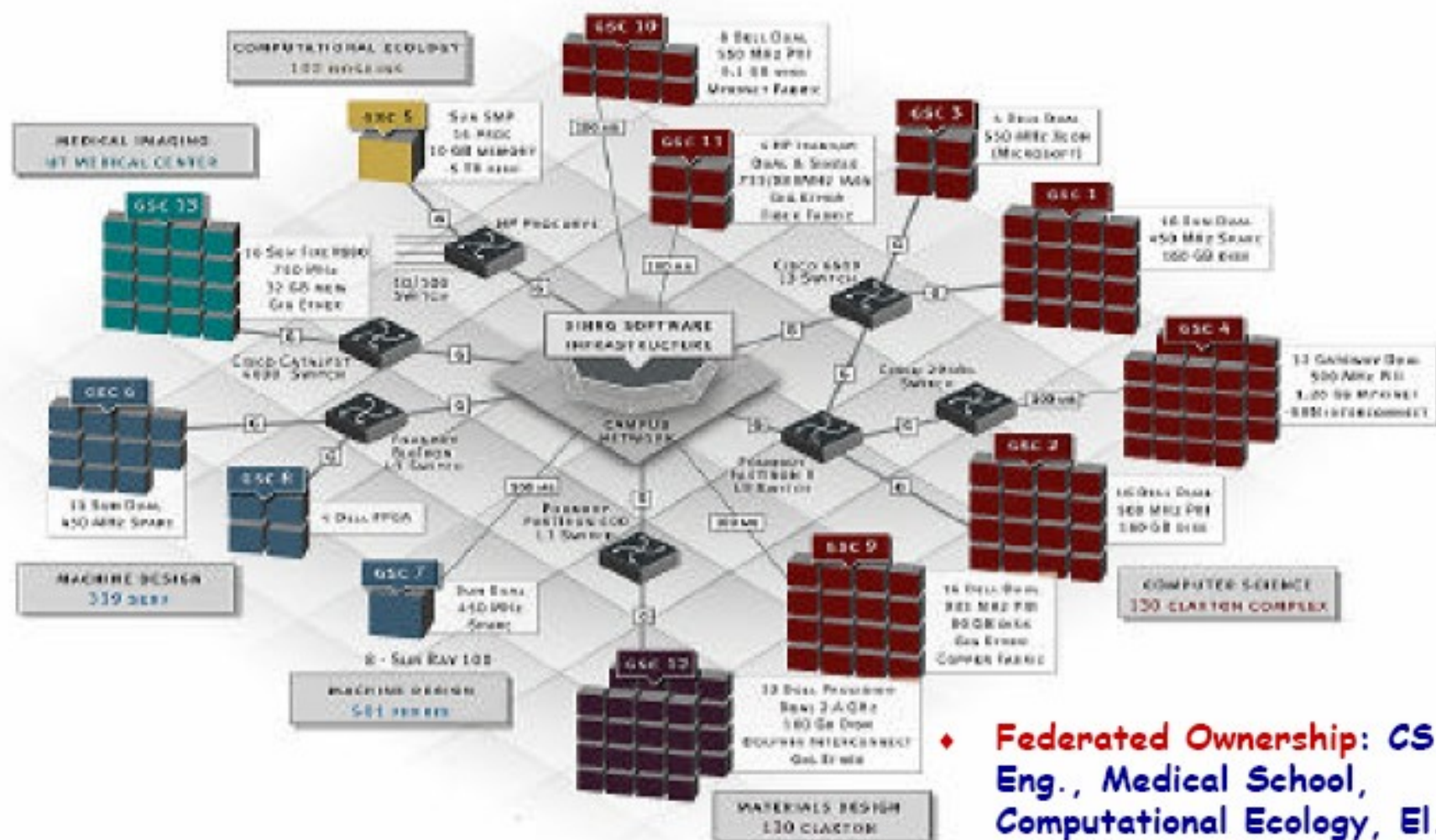
Status

- A Grid covering the south, south-east and north-east of Brazil
- Total nodes: ~200
- Coverage: ~1,900 mi.

Pauá Grid Project - Brasil

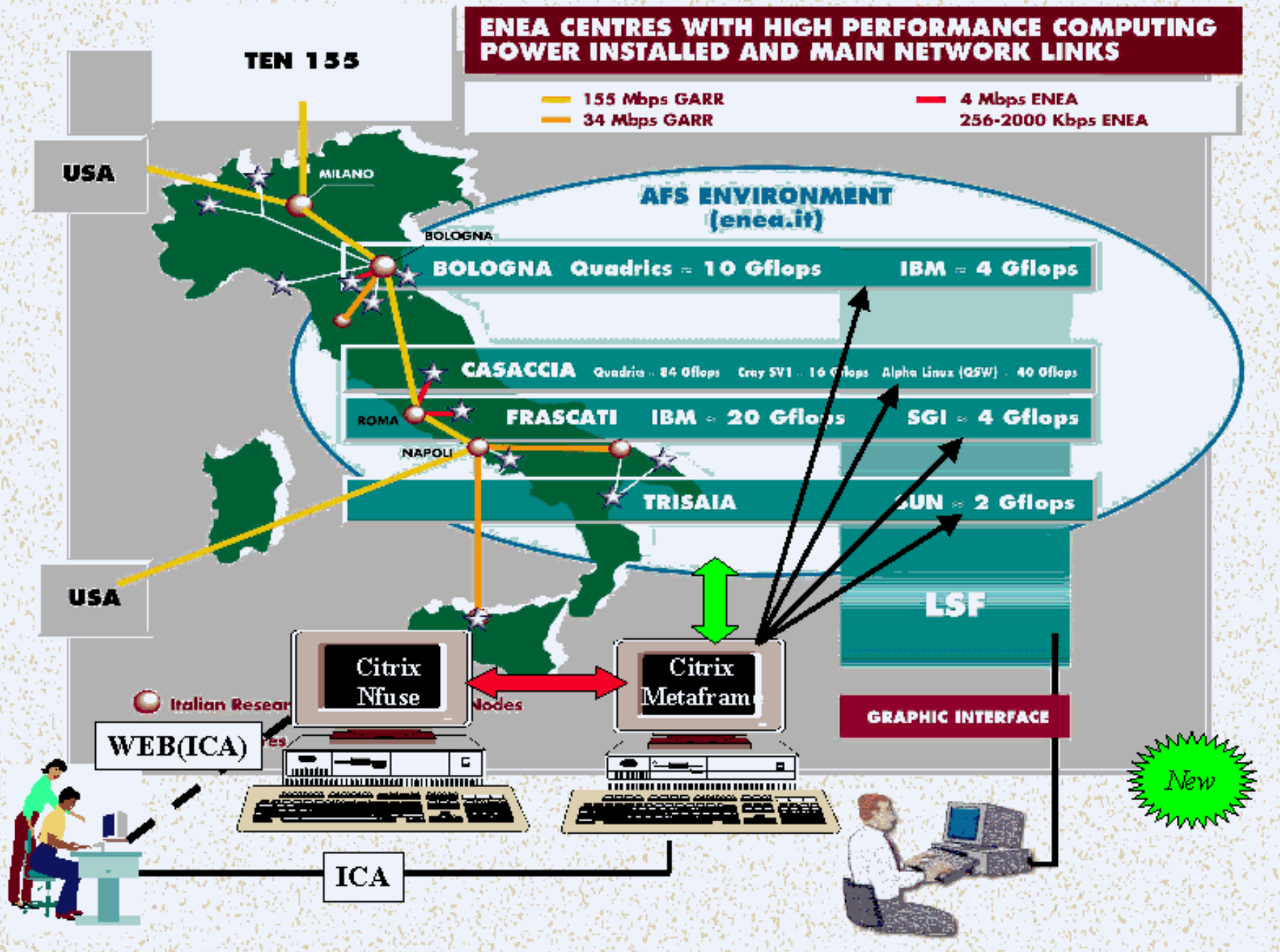
- Partnership with HP
- Links various academic sites across the nation
- 7 sites, 200 nodes, 1900 miles
- Bioinformatics, data mining, security applications
- Attracts projects from other knowledge areas that need high performance computing

University of Tennessee Deployment: Scalable Inter-campus Research Grid: SInRG



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

Other Grid Projects



Grid Academic Projects

Many Active Projects

- ◆ Akenti
- ◆ AppLeS
- ◆ Arcade
- ◆ CIF
- ◆ Condor
- ◆ CUMULVUS
- ◆ EveryWare
- ◆ Globus
- ◆ Habanero
- ◆ Harness
- ◆ IceT
- ◆ IPG NAS - NASA
- ◆ JINI
- ◆ Llava
- ◆ Legion
- ◆ NCSA Workbench Project
- ◆ NEOS
- ◆ NetSolve
- ◆ NINF
- ◆ Ninja
- ◆ PAWS
- ◆ PARDIS
- ◆ POEMS
- ◆ Sweb
- ◆ Teraweb
- ◆ UNICORE
- ◆ WebFlow

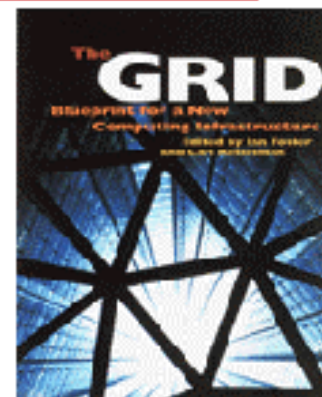
Grid Academic Projects

For More Information ...

◆ Recent book

"The Grid: Blueprint for a New Computing Infrastructure," I. Foster and C. Kesselman (Eds), Morgan-Kaufmann, 1999

<http://www.mkp.com/grids>



IPG NAS-NASA <http://nas.nasa.gov/~wej/home/IPG>

Globus <http://www.globus.org/>

Legion <http://www.cs.virginia.edu/~grimshaw/>

AppLeS <http://www-cse.ucsd.edu/groups/hpcl/apples>

NetSolve <http://www.cs.utk.edu/netsolve/>

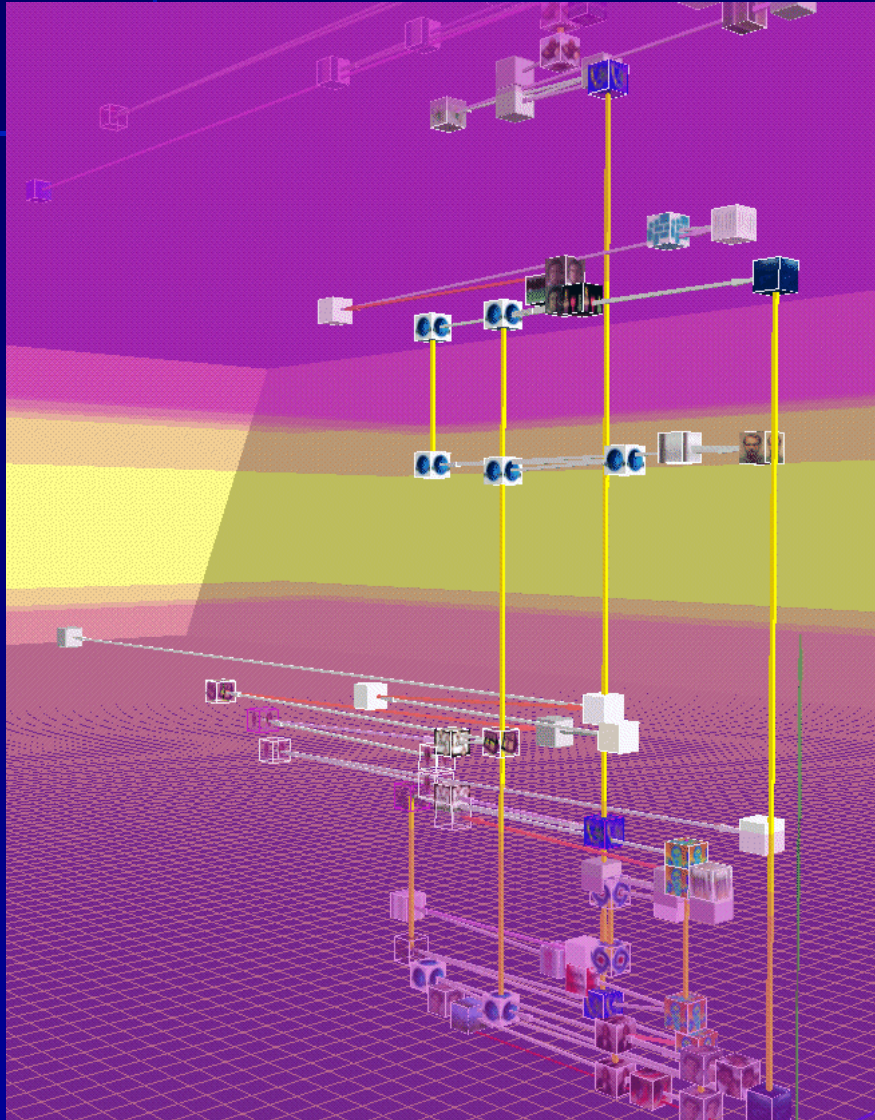
NINF <http://phase.etl.go.jp/ninf/>

Condor <http://www.cs.wisc.edu/condor/>

CUMULVS <http://www.epm.ornl.gov/cs/cumulvs.html>

WebFlow <http://www.npac.syr.edu/users/gcf/>

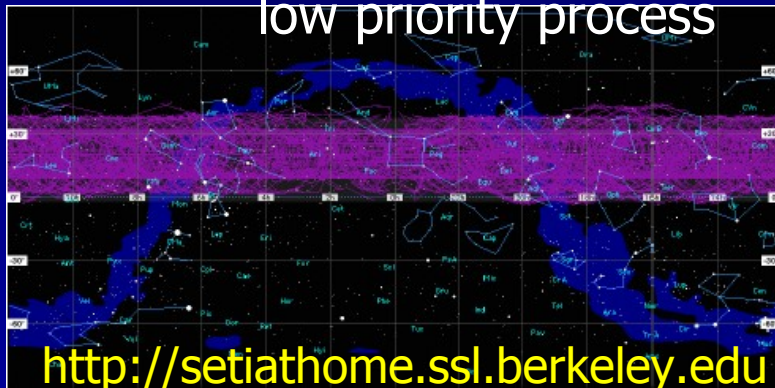
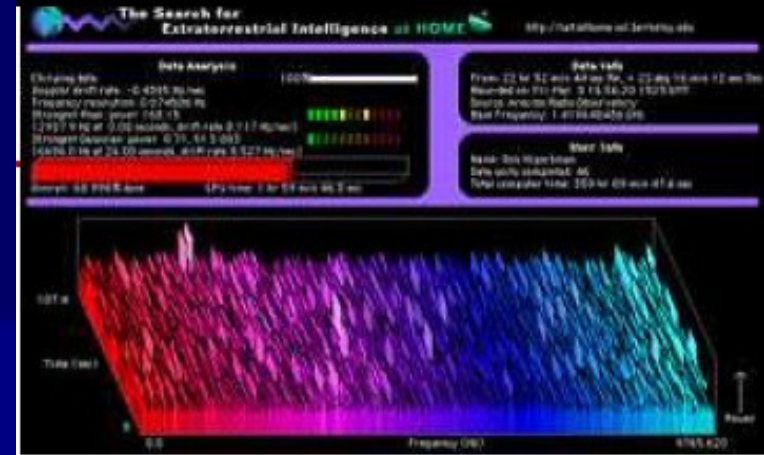
Grid Computing Applications



- SETI@home
- Human Proteome project
- Anthrax research project
- Smallpox project
- Cancer research project
- SciRun environment

SETI@home

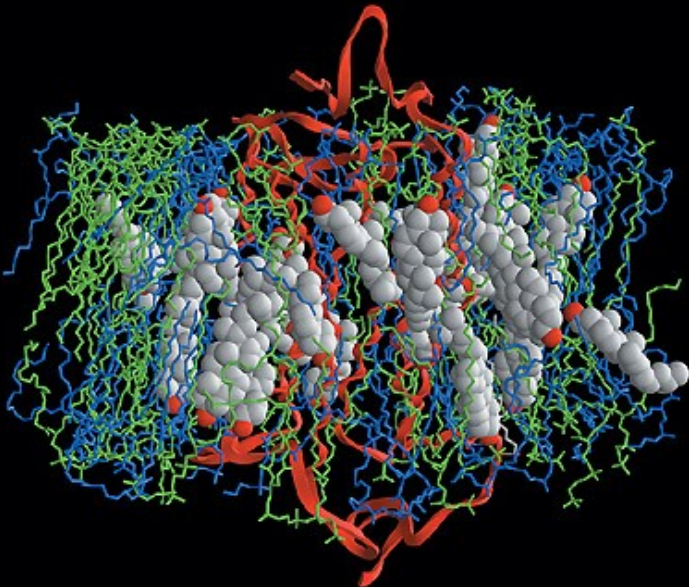
- LARGEST GRID COMPUTING PROJECT IN HISTORY
 - Runs on over 50,000 PCs
 - Generates approximately 1,000 CPU years / day
 - Approx 500,000 CPU years so far
 - Averages 40 Tflops/second
 - Distributes data sets from the Arecibo Radiotelescope
 - Performs sophisticated data and signal processing analysis
 - Can run as a **screen saver** or as a low priority process
- 
- 



Human Proteome Folding Project

- Find functions for all the proteins encoded in the Human Genome
- When human protein structures are known, scientists can use them to research disease treatments and cures
- Only a fraction of 30,000 human proteins have known structures and functions

<http://www.grid.org/projects/hpf/>



- Examining the entire human genome could require up to 1 million years of computational time on a Pentium IV.
- Using a commercial 1000-node cluster would require 50 years and, while faster, would still be impractical.
- Can run as a **screen saver** or as a low priority process

Score

Roetta
98.81
Mass
1839.34
Size
-0.83

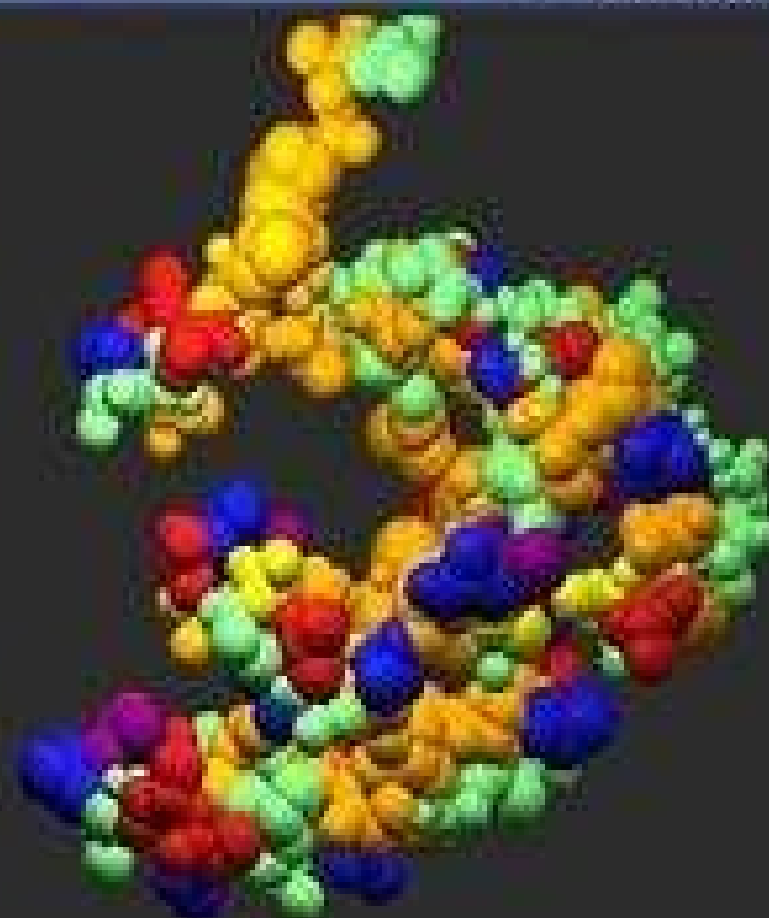
Electrostatics
-47.57
Mass
12.78
Size
-12.00

Water
-47.71
Mass
81.40
Size
-55.42

Current Progress



14.8%



Anthrax Research Project

- As of February 14, 2002, the **screening** phase of the Anthrax Research Project has been completed. In **4 weeks** it achieved what previously took **years**.
- This project's goal was to accelerate what is usually a time-consuming step in the lengthy drug discovery process.
- The project entailed presenting a key protein component of anthrax into the general rotation of the United Devices Member Community's current virtual screening project, which works with the Grid MP platform over the Internet.
- This allowed UD Members to lend their computers in the screening of **3.57 billion molecules** for suitability as a treatment for advanced-stage Anthrax.

Anthrax Research Project

- Screening is **only one step** in a long drug discovery process that ultimately must move from the computational realm into the actual laboratory.
- The project used a **5-time redundancy** rate for each molecule to ensure a high level of accuracy and quality.
- Preliminary indications are that we have narrowed the original pool of 3.57 billion molecules down considerably, having identified over **300,000 crude unique hits** in the course of the project.
- This significantly reduces the next phase of the discovery process, in which the ranked hits will be **further refined and analyzed**, accelerating the overall time to availability of a treatment.

<http://www.grid.org/projects/anthrax/>

Cancer Research Project

- Processes molecular research being conducted by the Department of Chemistry at the University of Oxford in England and the National Foundation for Cancer Research.
- To participate, users download a very small, no cost, non-invasive software program that works like a screensaver: it runs when your computer isn't being used, and processes research until you need your machine. Your computer never leaves your desk, and the project never interrupts your usual PC use.
- The research centers on proteins that have been determined to be a possible target for cancer therapy. Through a process called "virtual screening", special analysis software will identify molecules that interact with these proteins, and will determine which of the molecular candidates has a high likelihood of being developed into a drug.

Cancer Research Project

- The process is similar to finding the right key to open a special lock — by looking at millions upon millions of molecular keys.
- It allows computers to screen molecules that may be developed into drugs to fight cancer. Each individual computer analyzes a few molecules and then sends the results back over the Internet for further research.
- This project is anticipated to be **the largest** computational chemistry project ever undertaken and represents a genuine hope to find a better way to fight cancer.
- The computational power to perform research of this scale is only available through the generosity of participants.

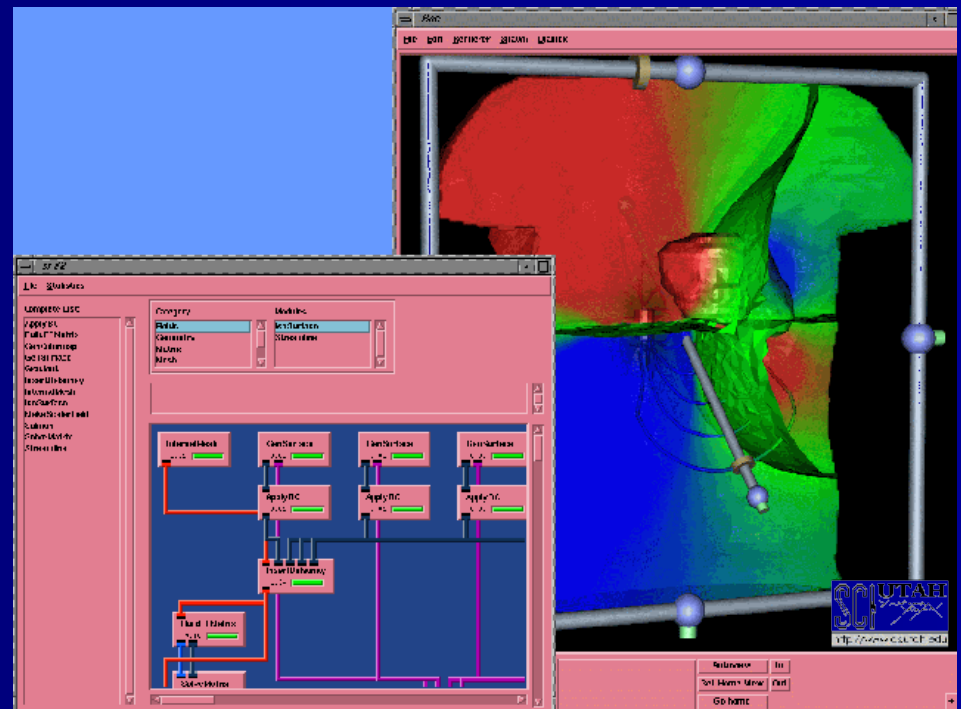
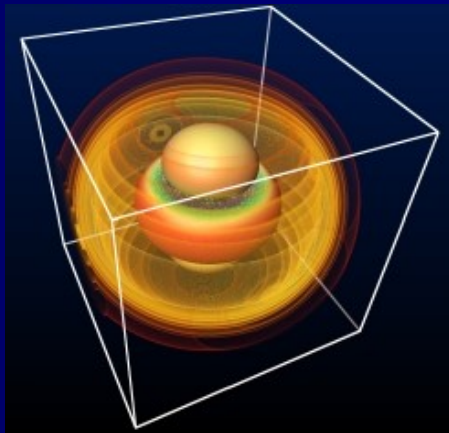
<http://www.grid.org/projects/cancer/>

Smallpox Research Project

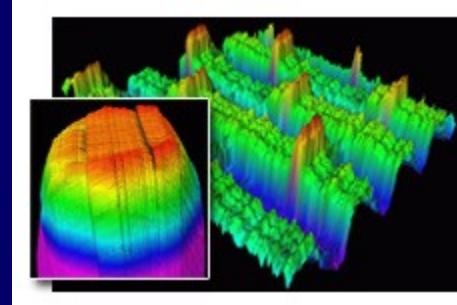
- Smallpox was eliminated from the world in 1977
- Stocks of the variola virus still exist, potential terrorist use
- Vaccination ended in 1972, so an outbreak would kill millions
- There is a possible molecular target whose blockade would prevent the ravages of an infection.
- The Smallpox Research Grid Project involves screening **35 million** potential drug molecules against several protein targets — one of the largest computational chemistry project ever undertaken.
- This will involve the use of the United Devices Grid MP Global, which we have successfully used in the past towards cancer and anthrax research.
- The project can harness millions of computers belonging to people in over two hundred countries, all of whom will benefit from protection against smallpox.
- Can run as a **screen saver** or as a low priority process

SCIRun

- Interactive Scientific Programming
- Interactively composes large-scale scientific computations through visual dataflow programming
- Integrated with visualization packages
- Supports interactive steering during the design, computation, and visualization phases of a simulation



Key Vendors



- Parabon Computation www.parabon.com
- DataSynapse www.datasynapse.com
- IBM Grid Computing www.ibm.com/grid
- Sun Microsystems Grid Computing www.sun.com/grid
- Oracle Corp. www.oracle.com/grid
- HP Grid Computing
www.hp.com/techservers/grid/index.html
- United Devices www.ud.com
- 1st Port for Grid Computing www.1stport.co.uk

References

- *Computational Intermediation and the Evolution of Computation as a Commodity*, Applied Economics, June 2004
www.business/duq.edu/faculty/davies/research/economicsofcomputation.pdf
- *The Grid: Blueprint for a New Computing Infrastructure*
www.mkp.com/grids
- *Grid Computing: Making the Global Infrastructure a Reality*
www.grid2002.org
- *The Grid : Core Technologies*
<http://coregridtechnologies.org>
- *CERN: The Grid Café – What is a Grid?*
<http://gridcafe.web.cern.ch/gridcafe/whatisagrid/whatis.html>
- *Grid Computing: A Brief Technology Analysis*
www.ctonet.org/documents/gridcomputing_analysis.pdf
- *IBM: What is Grid Computing*
http://www-1.ibm.com/grid/about_grid/what_is.shtml



References

- *IBM Grid Computing Benefits*
http://www-1.ibm.com/grid/about_grid/benefits.shtml
- *Sun Microsystems: What is Grid?*
<http://www.sun.com/executives/iforce/integratedsolutions/gridsolutions/index.html>
- *HP Grid Computing*
<http://www.hp.com/techservers/grid/index.html>
- *Oracle Grid Index Report*
<http://www.oracle.com/global/eu/pressroom/nagridreport.pdf>
- Juhasz, Zoltan, et al, *Distributed and Parallel Systems, Cluster and Grid Computing*, Springer Science and Business Media, 2005 [ISBN 0-387-23094-7]
- Minoli, Daniel, *A Networking Approach to Grid Computing*, John Wiley and Sons, 2005 [ISBN 0-471-68756-1]
- *Wikipedia: Grid Computing*
http://en.wikipedia.org/wiki/Grid_computing



