Future Scientific Infrastructure

Ian Foster

Argonne National Laboratory

University of Chicago

Globus Alliance

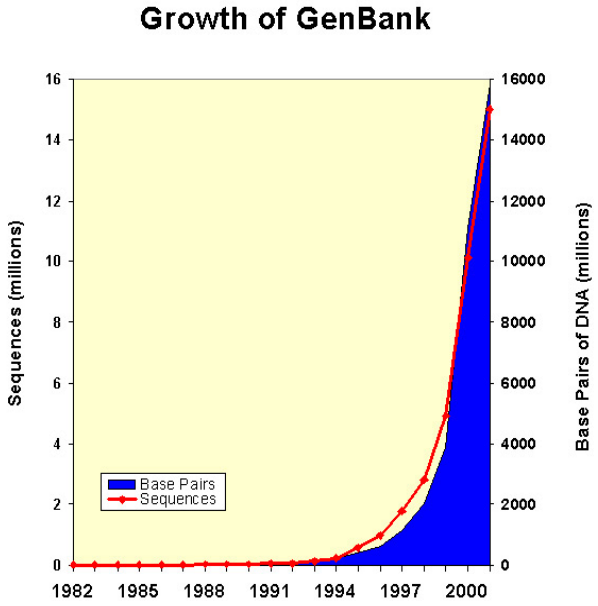
www.mcs.anl.gov/~foster

Overview

- Evolution of scientific infrastructure
  - Technology drivers & new requirements
- Building cyberinfrastructure
  - Standards, software, and facilities
  - Examples
- Futures
  - Virtual data
  - Cyberinfrastructure for life sciences
  - Challenges

Technology Drivers

- Internet revolution: 100M+ hosts
  - Collaboration & sharing the norm
- Universal Moore's law: x10^3/10 yrs
  - Sensors as well as computers
- Petascale data tsunami
  - Gating step is analysis
- & our old infrastructure?

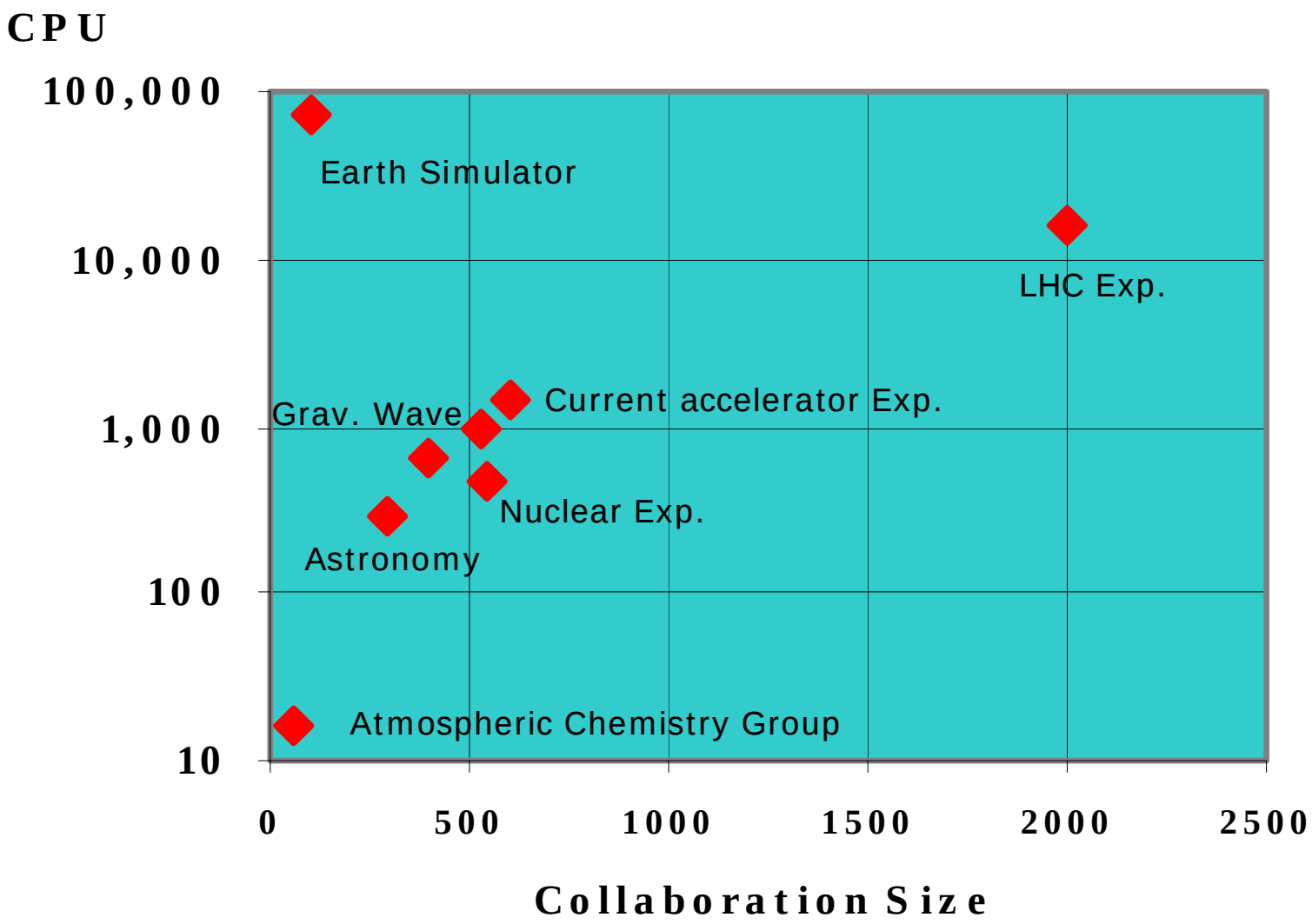


114 genomes
735 in progress

If We Can't Do it All in our Lab …

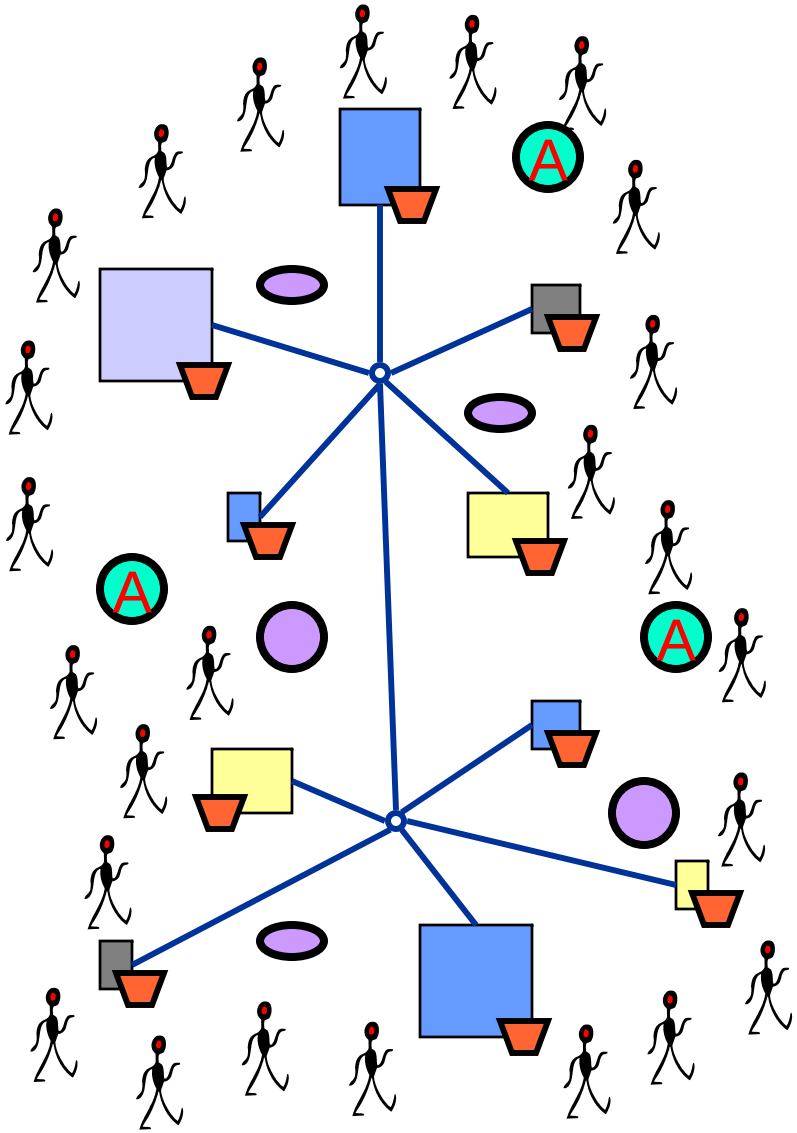
- Then we need to be able to reach out into the network to obtain needed resources
  - Clusters, archives, instrumentation
- And integrate those resources & services into our data processing workflows
  - Operating at scale (not Web browsing)
- Which means we need
  - Community facilities
  - Standard means of accessing those facilities
  - Means of coordinating across facilities
  - New methodologies and tools

Scale Metrics: Participants, Data, Tasks, Performance, Interactions, …



Cyberinfrastructure—AKA Grid (Metaphorical *and* Go-Fast)

- Resources
  - Computing, storage, data
- Services
  - Authentication, discovery, …
- Connectivity
  - Reduce tyranny of distance
- Technologies
  - Build services & applications
- Communities
  - Operational procedures, …



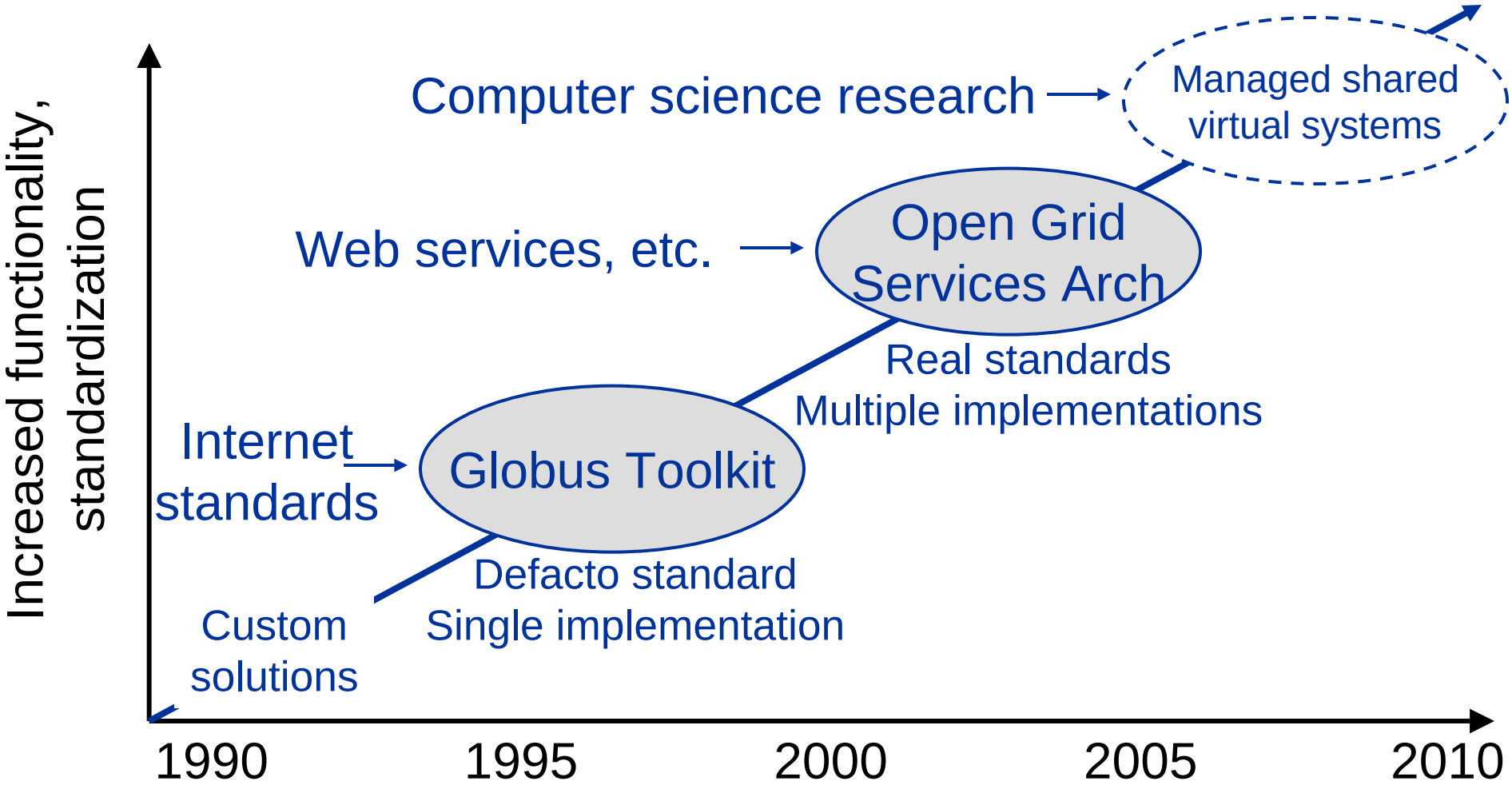
Overview

- Evolution of scientific infrastructure
  - Technology drivers & new requirements
- **Building cyberinfrastructure**
  - **Standards, software, and facilities**
  - **Examples**
- Futures
  - Virtual data
  - Cyberinfrastructure for life sciences
  - Challenges

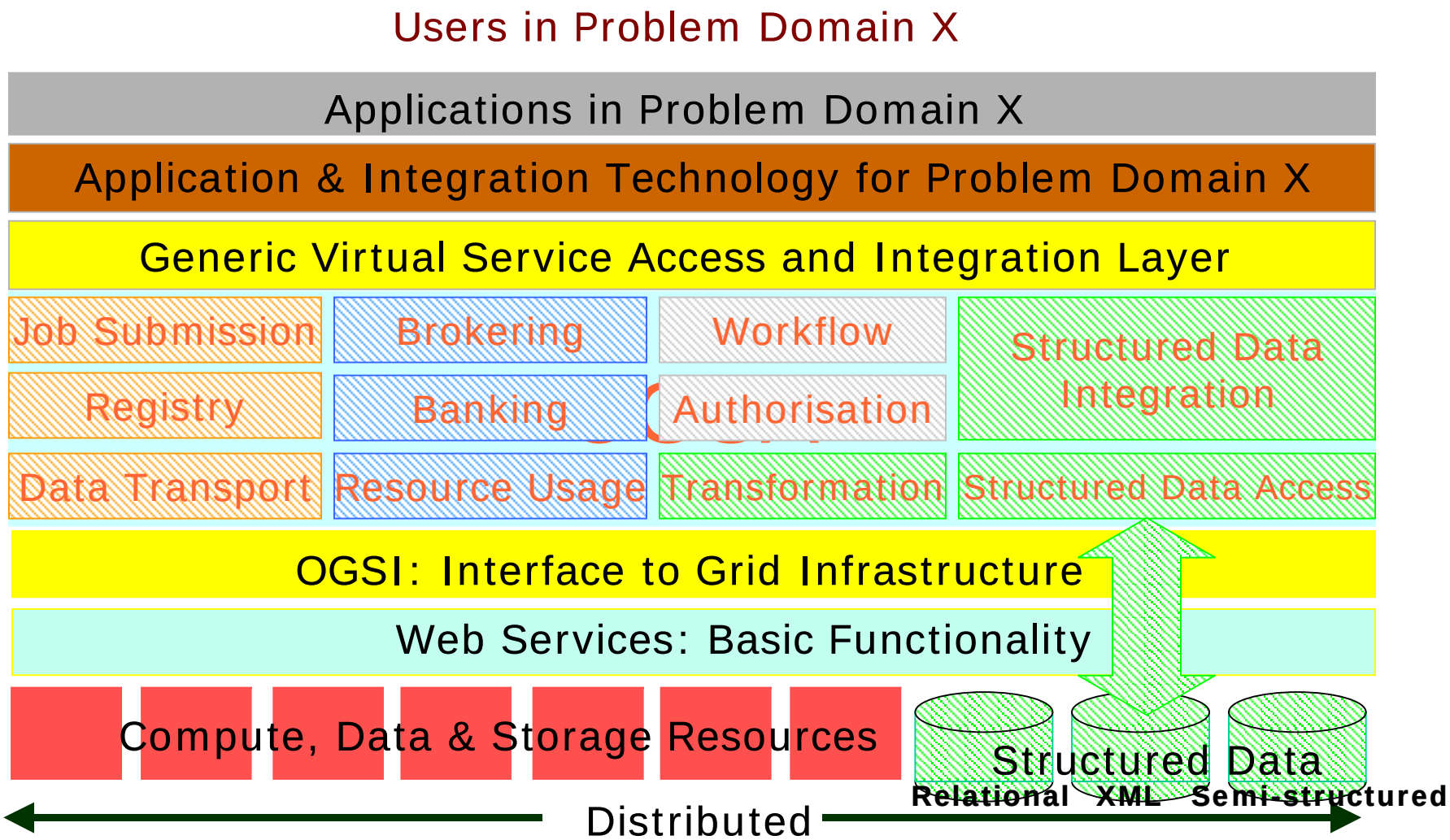
Building Cyberinfrastructure

- Open **standards**: absolutely vital
  - Define how we describe, discover, access, secure, monitor, & manage services
  - E.g., Open Grid Services Architecture
- Open **facilities**: enabled by standards
  - Standards-based access, community services (authentication, registry), operational support
  - E.g., TeraGrid, Grid3, NEESgrid, ESG, DOE Science Grid, LHC Computing Grid, NASA IPG
- Open **software**: accelerate standards adoption
  - Accelerates creation of facilities & applications
  - Globus Toolkit, NSF Middleware Initiative, etc.

The Emergence of Open Grid Standards

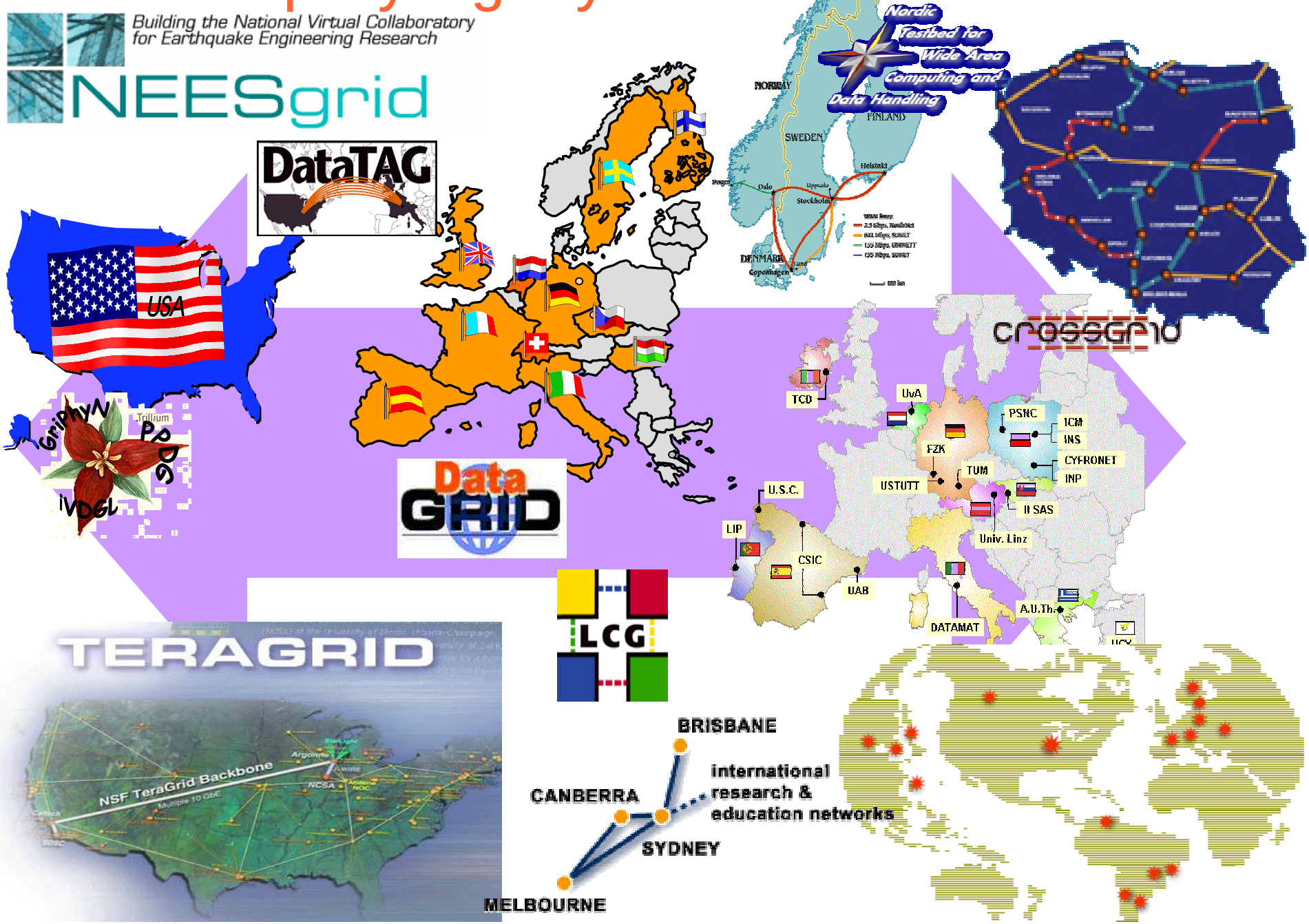


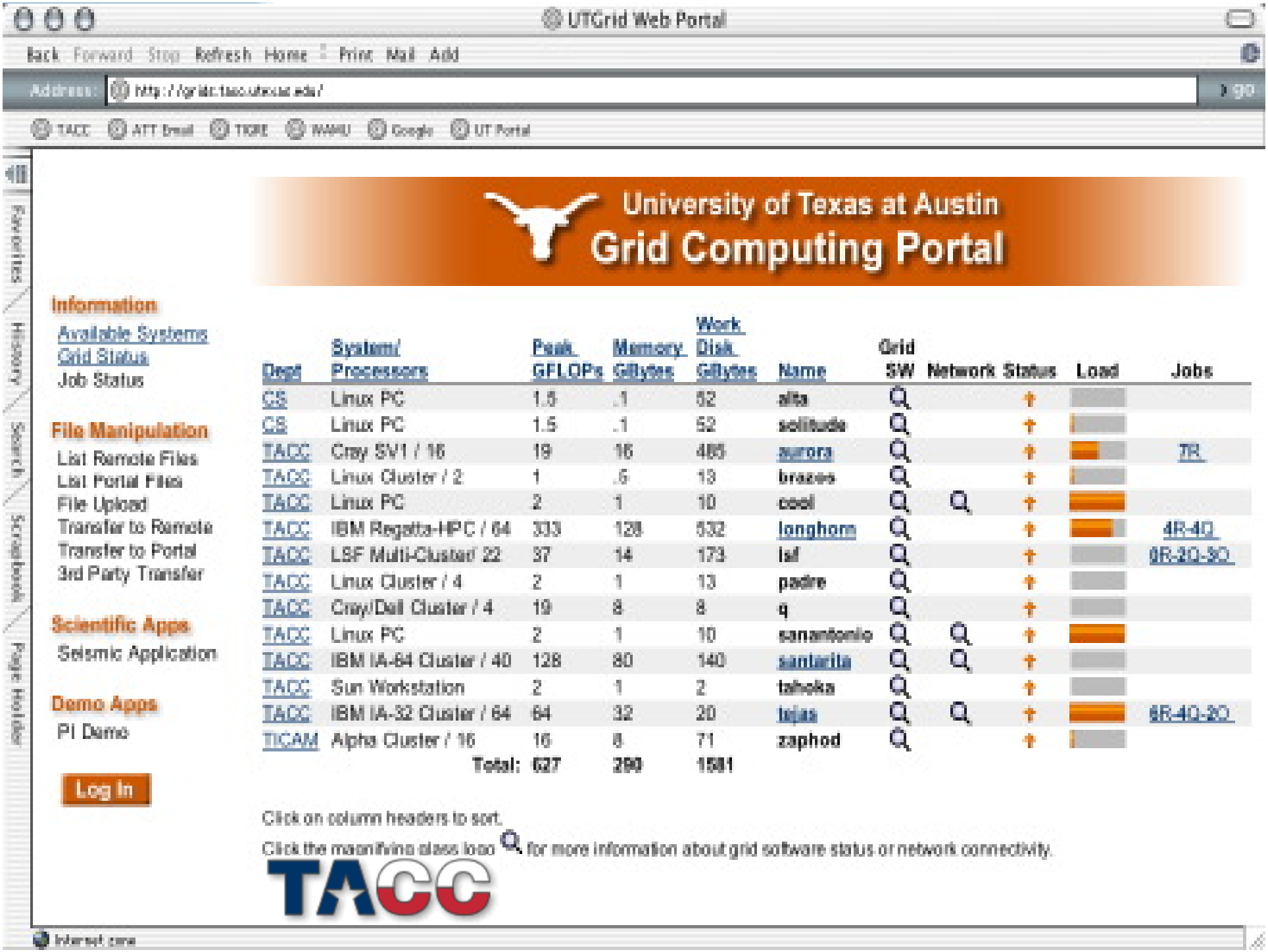
Open Grid Services Architecture (GGF, OASIS, W3C)



Virtual Integration Architecture

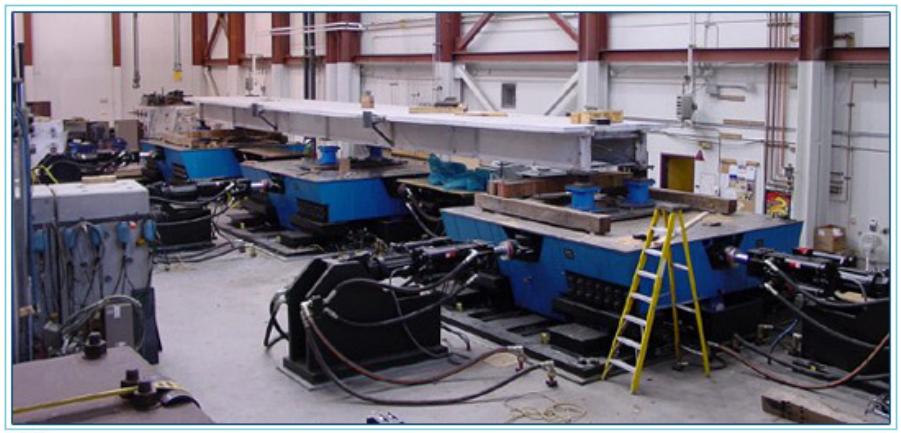
Deploying Cyberinfrastructure




UTGrid Web Portal
Back Forward Stop Refresh Home Print Mail Add
Address: http://grid.tacc.utexas.edu/
go
TACC ATT Email TIGRE WAMU Google UT Portal
Favorites History Search Scrapbook Page Holder
University of Texas at Austin
Grid Computing Portal
Information
Available Systems
Grid Status
Job Status
File Manipulation
List Remote Files
List Portal Files
File Upload
Transfer to Remote
Transfer to Portal
3rd Party Transfer
Scientific Apps
Seismic Application
Demo Apps
PI Demo
Log In
Dept | System/ Processors | Peak GFLOPs | Memory GBytes | Work Disk GBytes | Name | Grid SW | Network | Status | Load | Jobs
CS | Linux PC | 1.5 | .1 | 52 | alta
CS | Linux PC | 1.5 | .1 | 52 | solitude
TACC | Cray SV1 / 16 | 19 | 16 | 485 | aurora | 7R
TACC | Linux Cluster / 2 | 1 | .5 | 13 | brazos
TACC | Linux PC | 2 | 1 | 10 | cool
TACC | IBM Regatta-HPC / 64 | 333 | 128 | 532 | longhorn | 4R-4Q
TACC | LSF Multi-Cluster/ 22 | 37 | 14 | 173 | lsf | 0R-2Q-3O
TACC | Linux Cluster / 4 | 2 | 1 | 13 | padre
TACC | Cray/Dell Cluster / 4 | 19 | 8 | 8 | q
TACC | Linux PC | 2 | 1 | 10 | sanantonio
TACC | IBM IA-64 Cluster / 40 | 128 | 80 | 140 | santarita
TACC | Sun Workstation | 2 | 1 | 2 | tahoka
TACC | IBM IA-32 Cluster / 64 | 64 | 32 | 20 | tejas | 6R-4Q-2O
TICAM | Alpha Cluster / 16 | 16 | 8 | 71 | zaphod
Total: | 627 | 290 | 1581
Click on column headers to sort.
Click the magnifying glass logo for more information about grid software status or network connectivity.
TACC
Internet zone

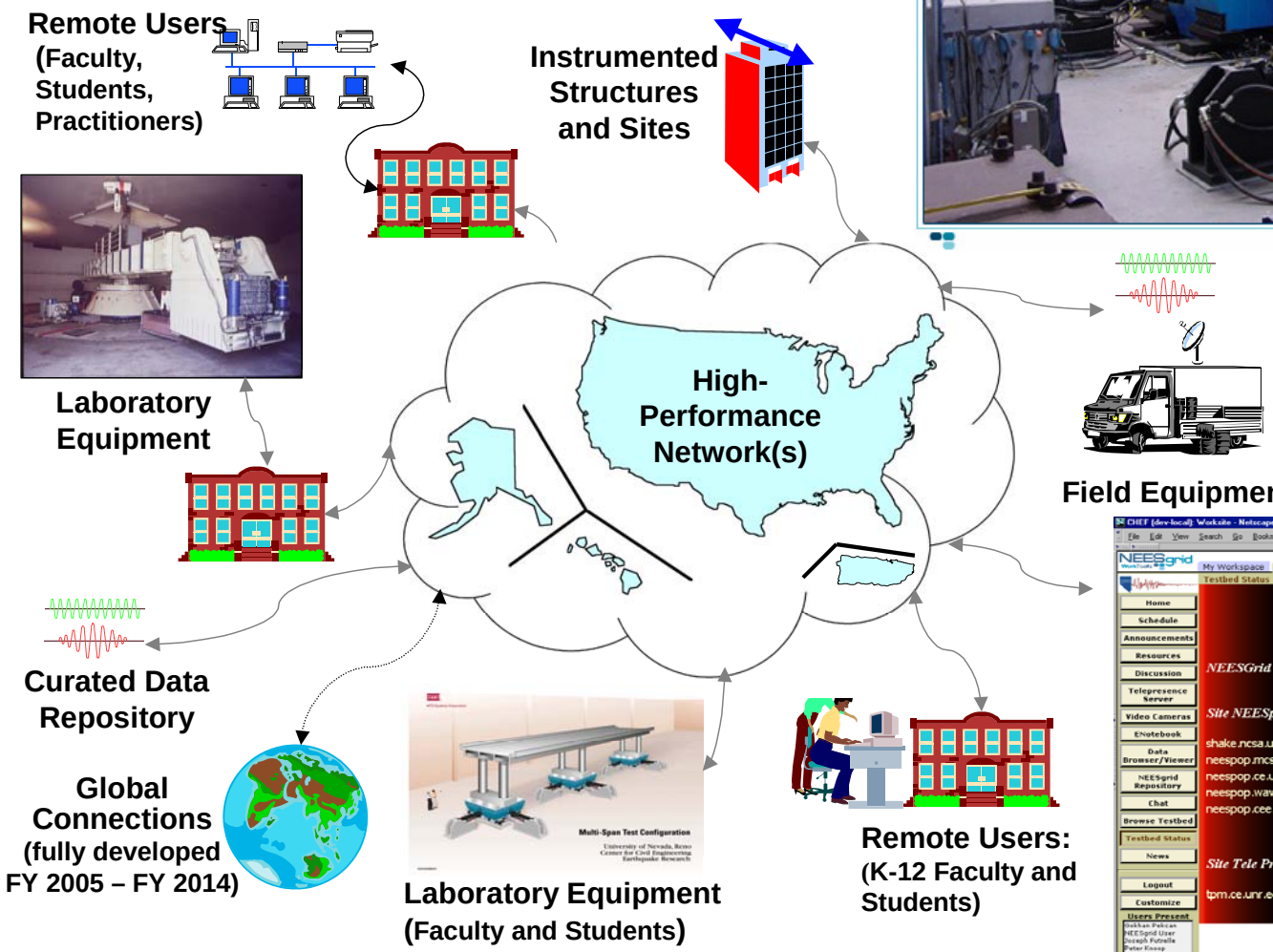
the globus alliance
www.globus.org

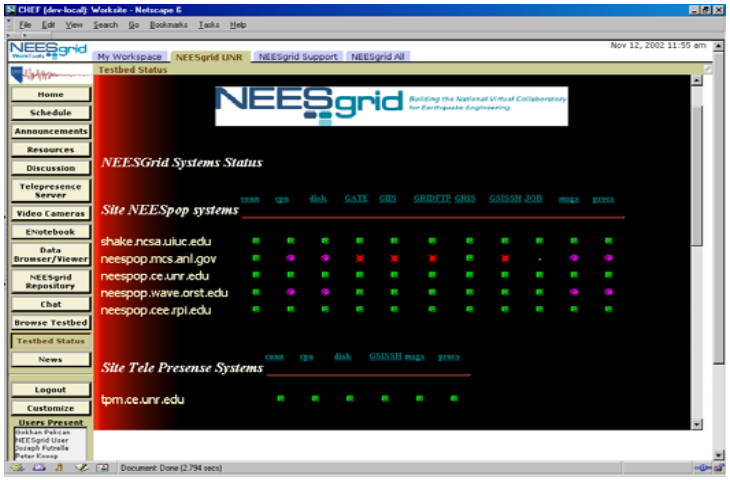
NEESgrid
Earthquake Engineering Collaboratory



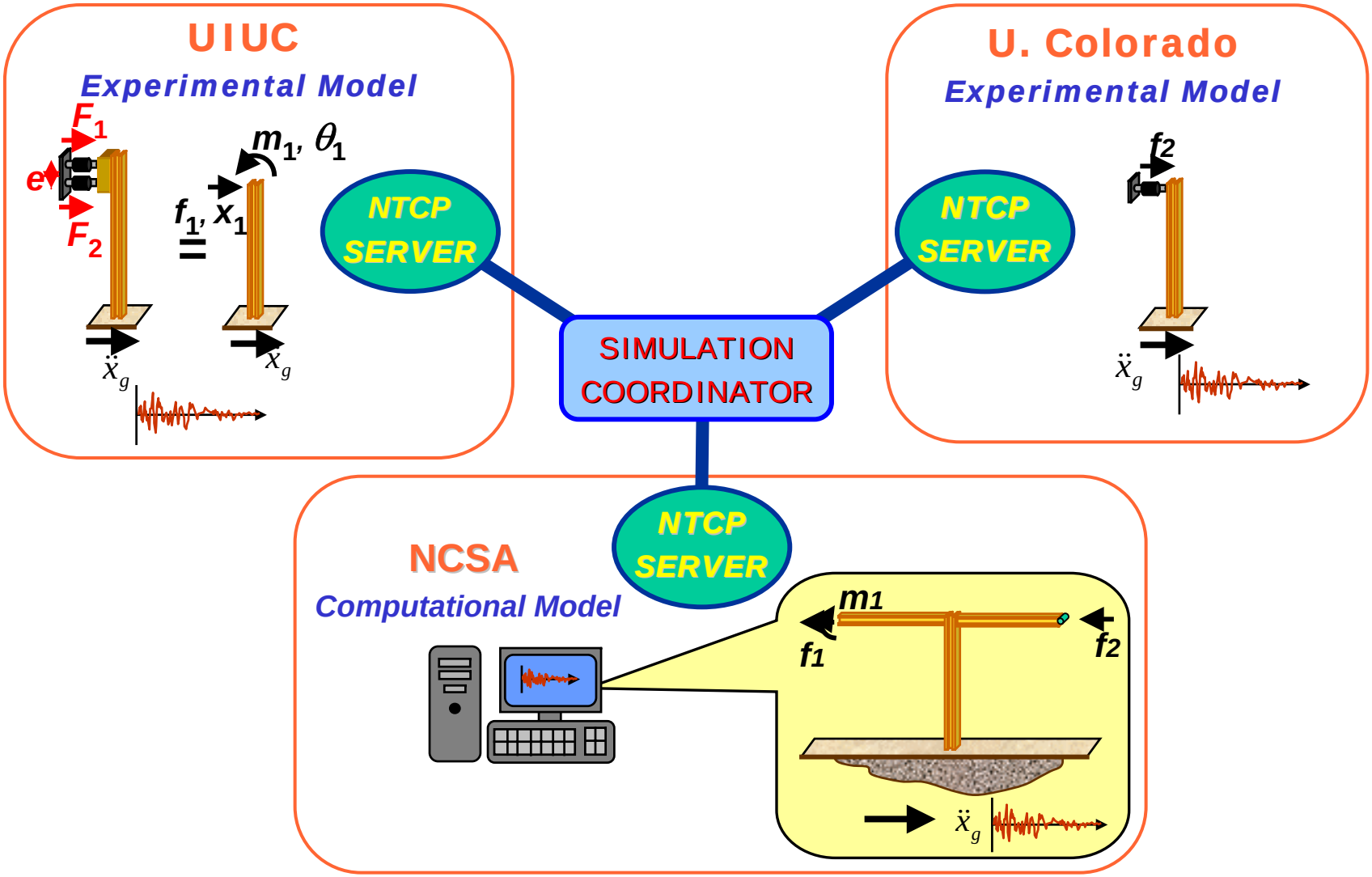
U.Nevada Reno

www.neesgrid.org



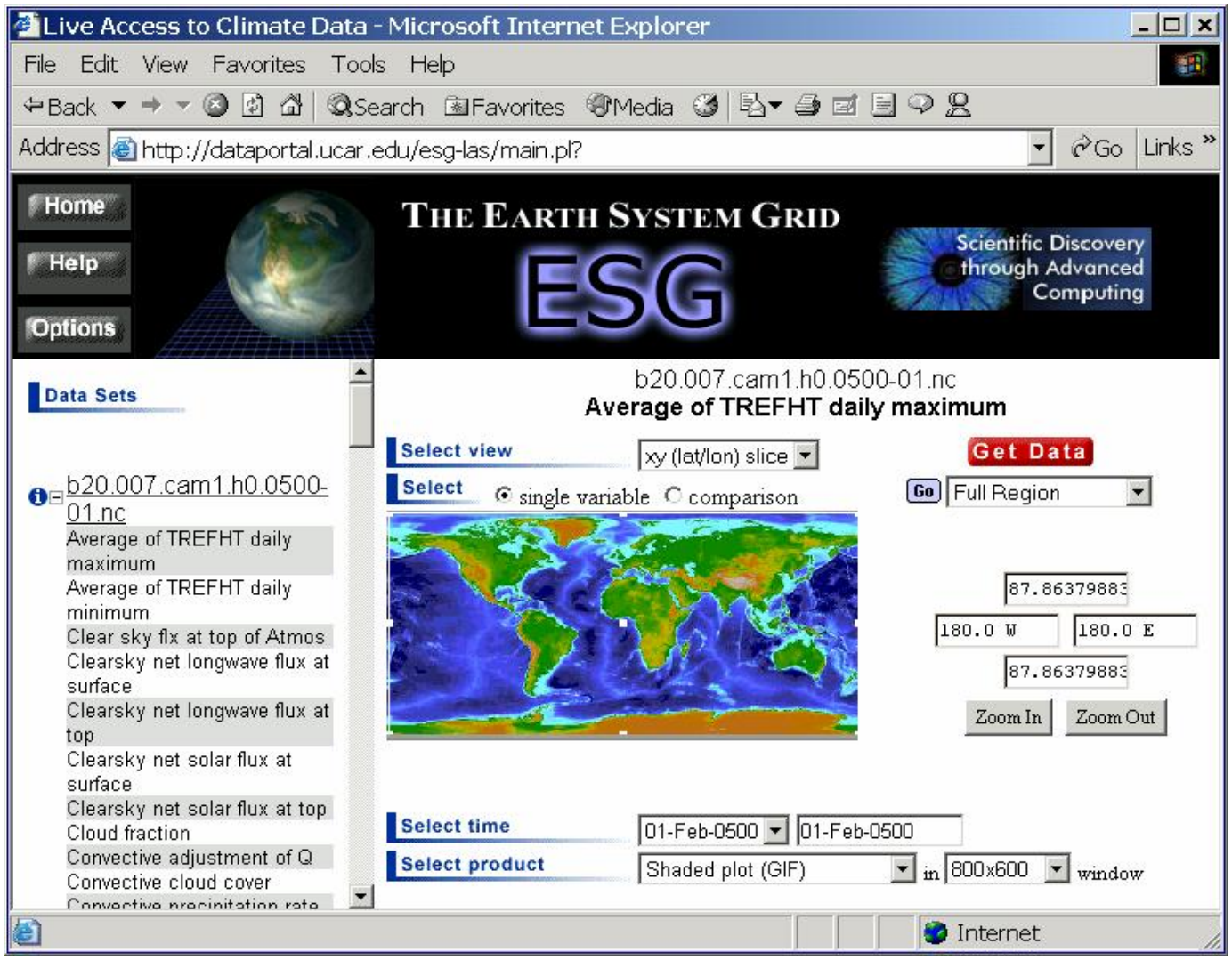


MOST: Multi-site Online Simulation Test

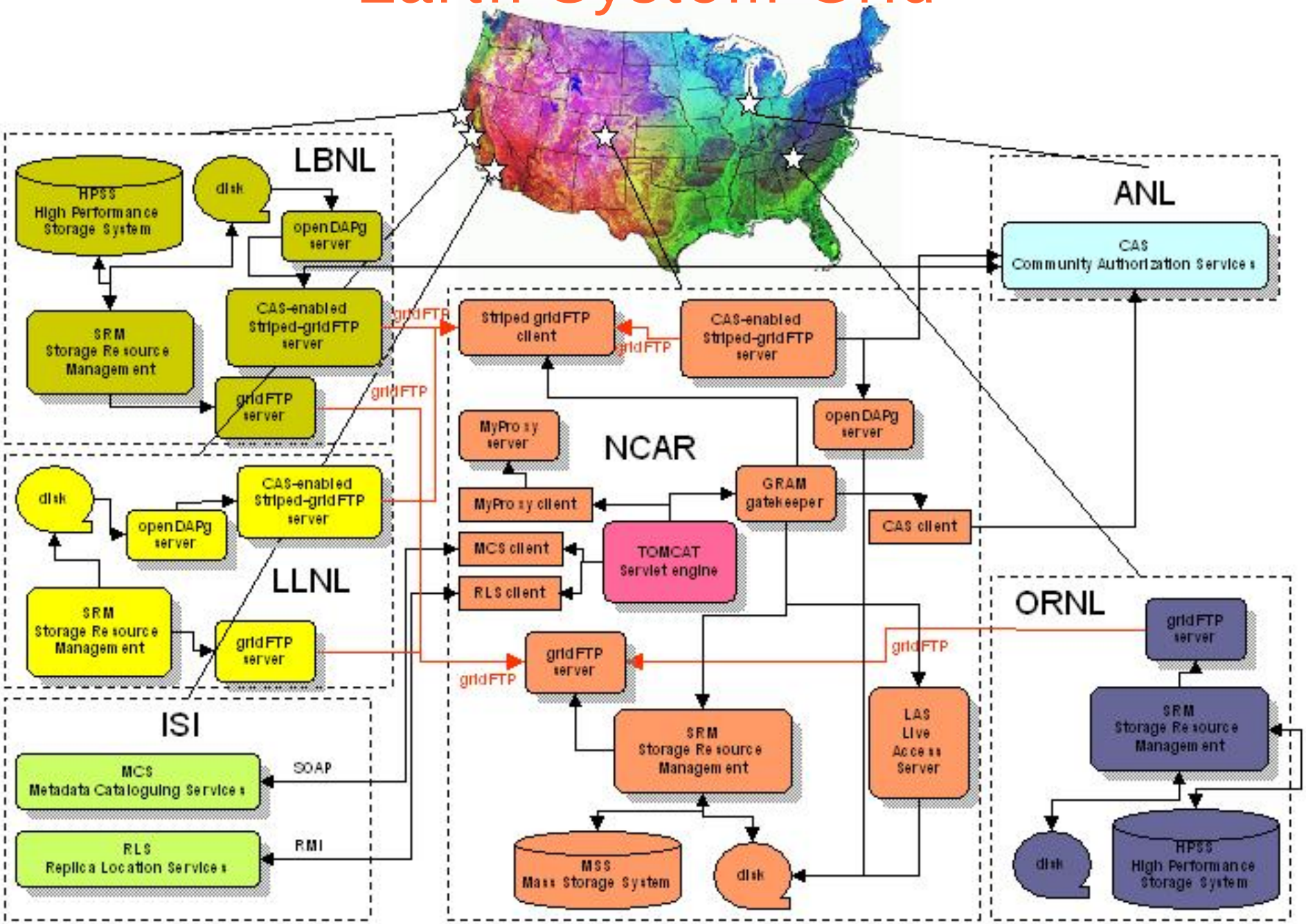


Earth System Grid (ESG)

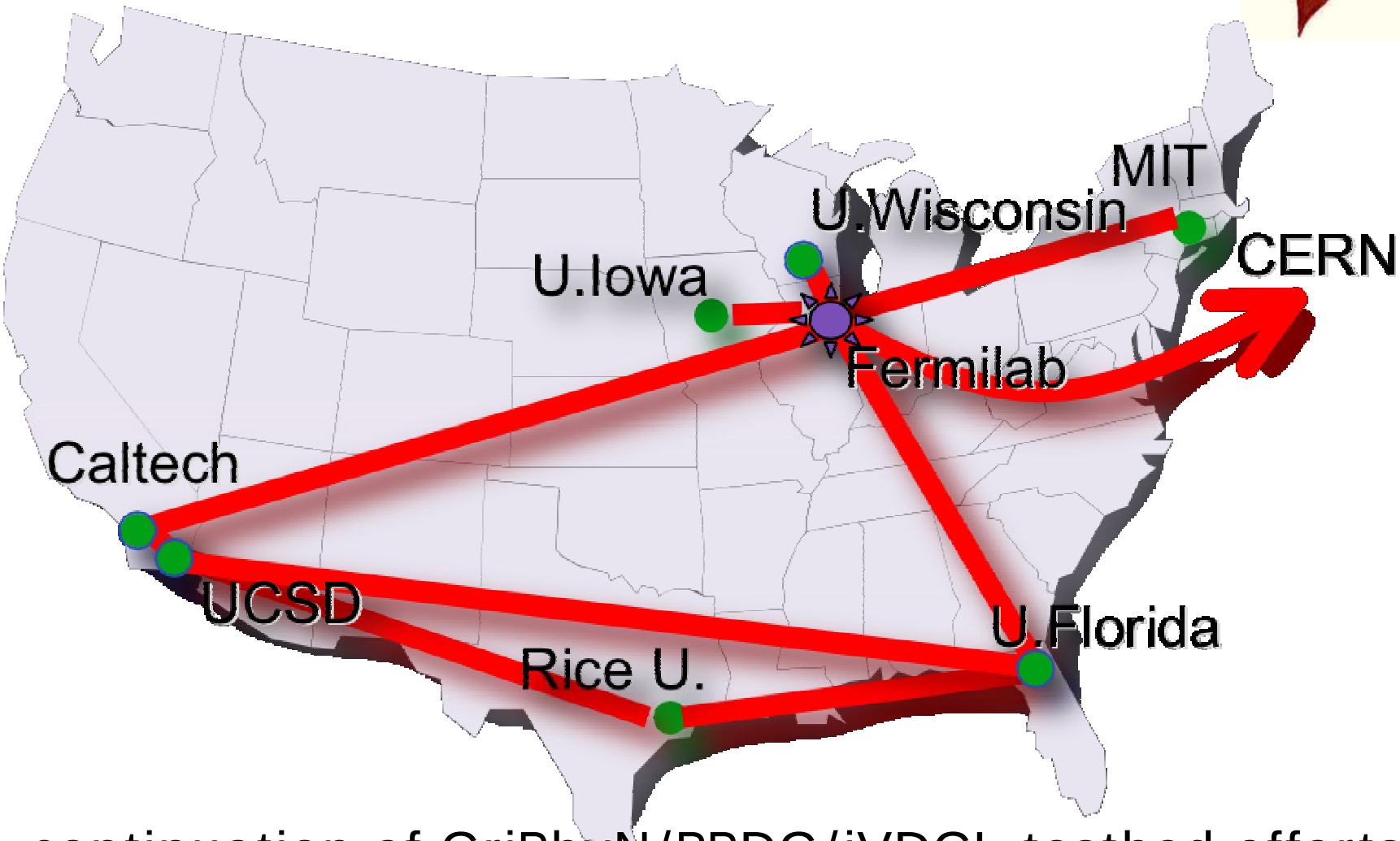
Goal: address technical obstacles to the sharing & analysis of high-volume data from advanced earth system models



Earth System Grid



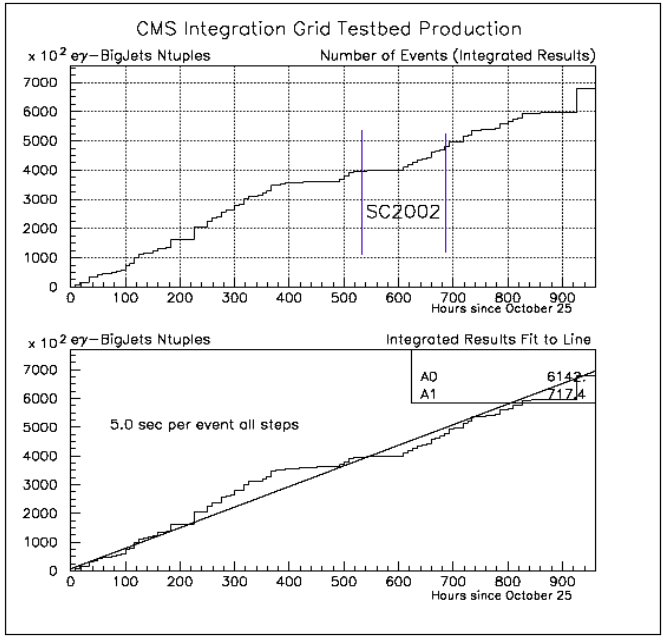
Grid3: Building A Community Science Grid

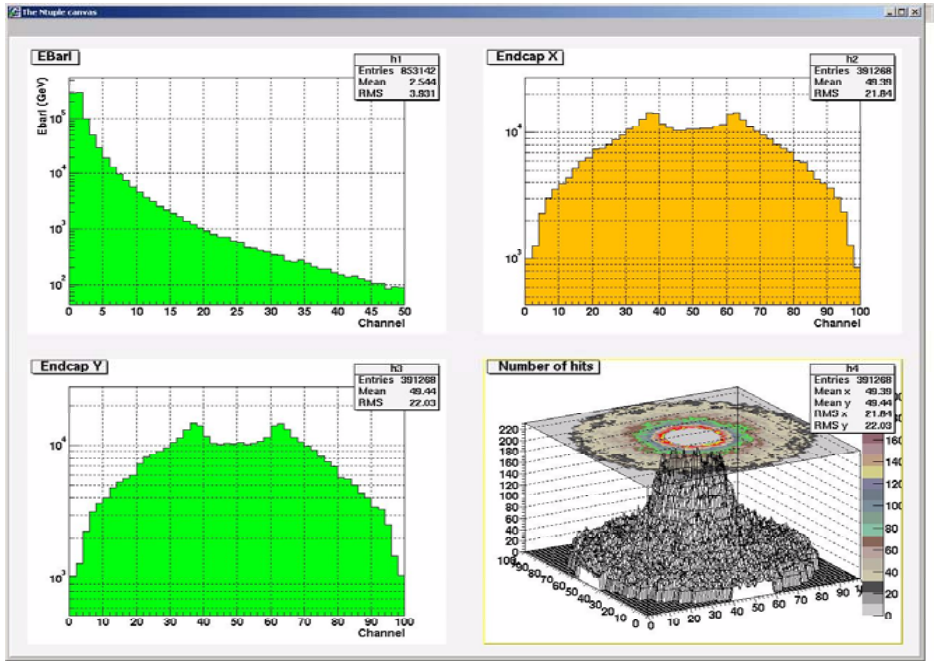


A continuation of GriPhyN/PPDG/iVDGL testbed efforts, focused on establishing a functional federated Grid

CMS Event Simulation Production

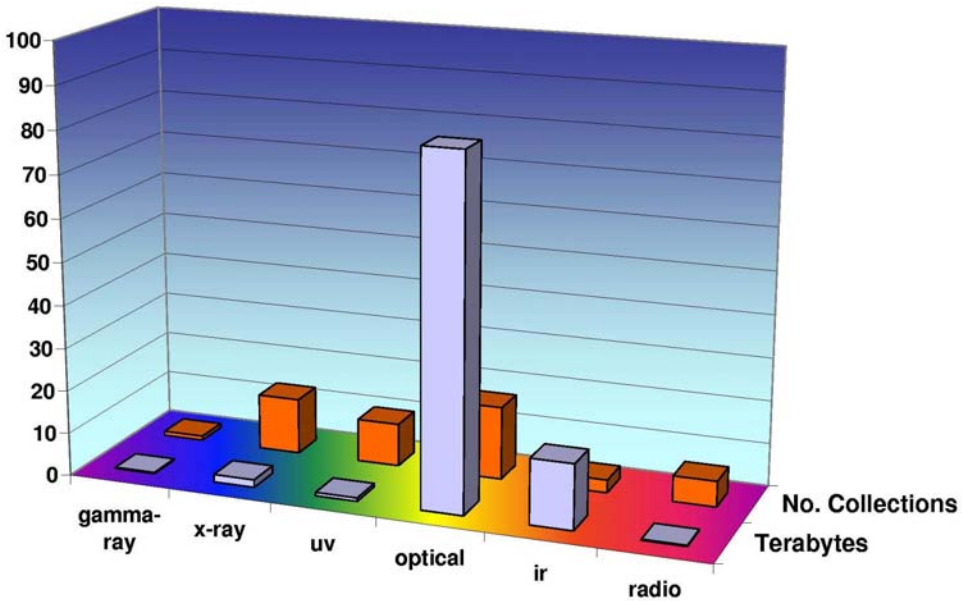
- Production run on the integration testbed
  - Simulate 1.5 million full CMS events for physics studies: ~500 sec per event on 850 MHz processor
  - 2 months continuous running across 5 testbed sites
  - Managed by a single person at the US-CMS Tier 1



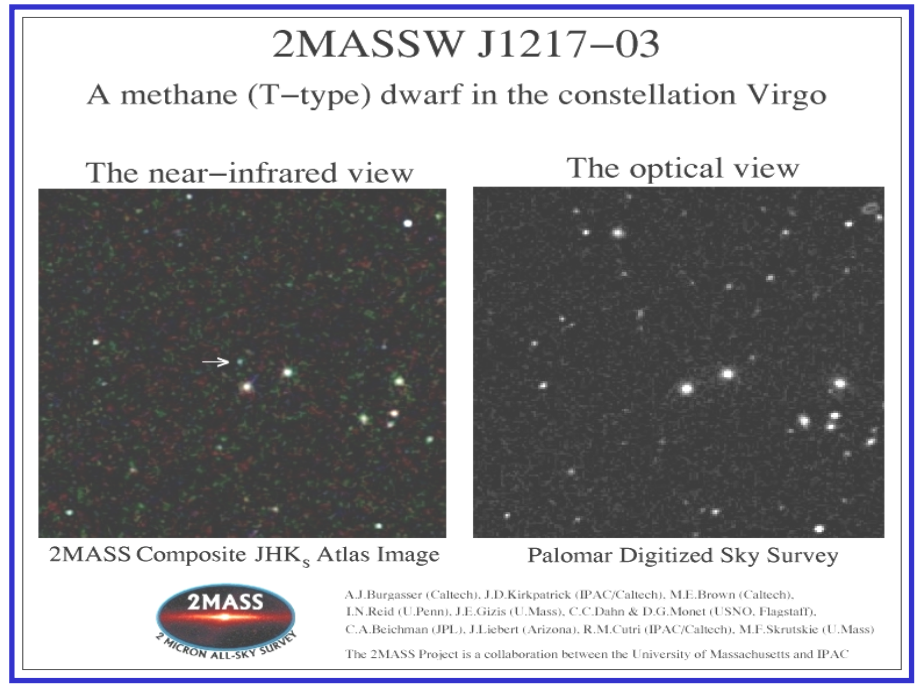


- EU DataGrid and LCG-1 operating at similar scales

Virtual Observatories



No. & sizes of data sets as of mid-2002, grouped by wavelength



THE NEW YORK TIMES NATIONAL TUESDAY, JUNE 1, 1999

Astronomers Detect New Category of Elusive 'Brown Dwarfs'

By JOHN NOBLE WILFORD

CHICAGO, May 31 — Ambitious

Apache Point, N.M. Dr. Michael Strauss and a graduate student,

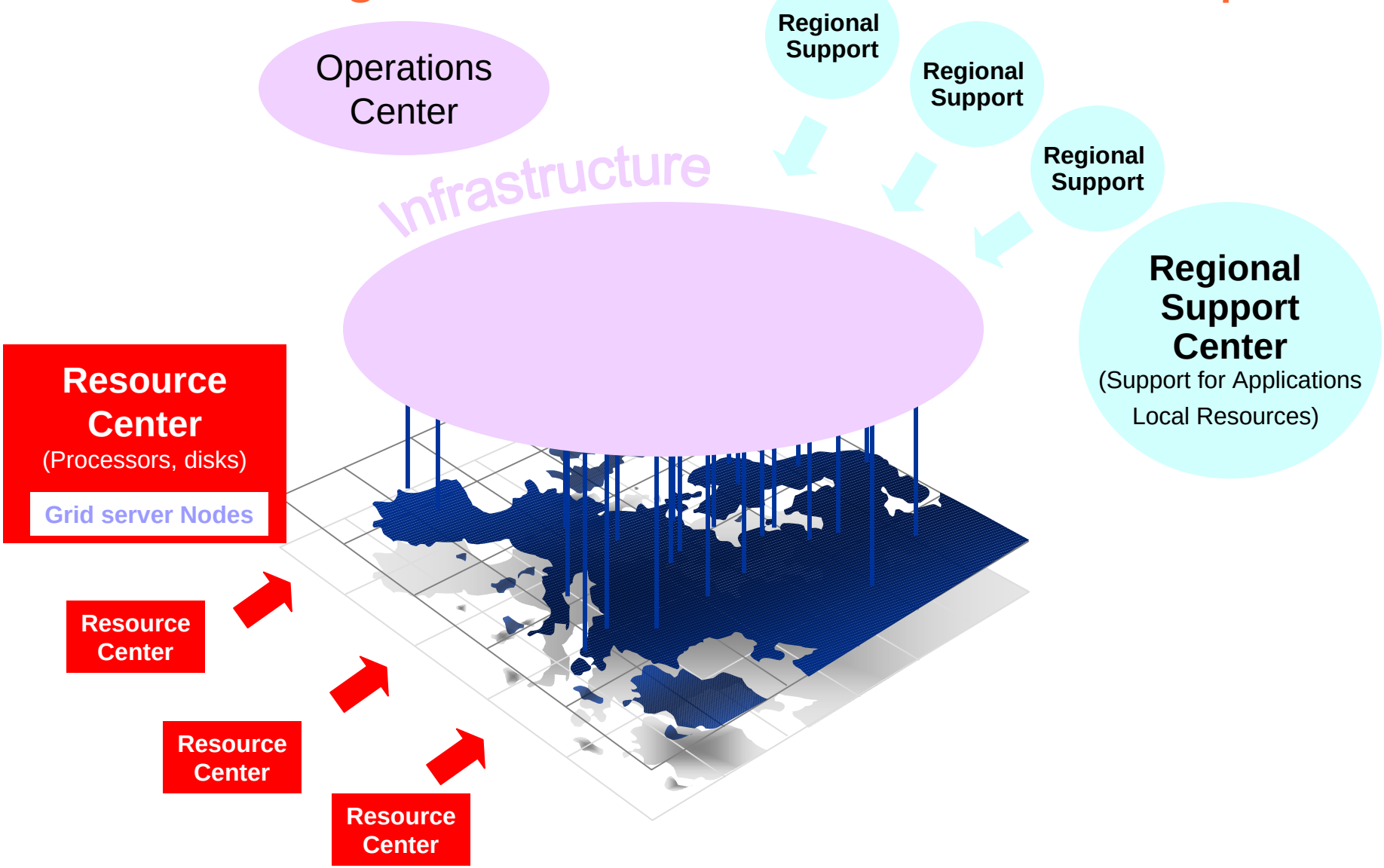
was a brown dwarf, but was not associated with a star companion.

have been possible in hotter, younger objects. An estimate of their mass

- 12 waveband coverage of large areas of the sky
- Total about 200 TB data
- Doubling every 12 months
- Largest catalogues near 1B objects

Data and images courtesy Alex Szalay, John Hopkins

EGEE:
Enabling Grids for E-Science in Europe



Overview

- Evolution of scientific infrastructure
  - Technology drivers & new requirements
- Building cyberinfrastructure
  - Standards, software, and facilities
  - Examples
- **Futures**
  - **Virtual data**
  - **Cyberinfrastructure for life sciences**
  - **Challenges**

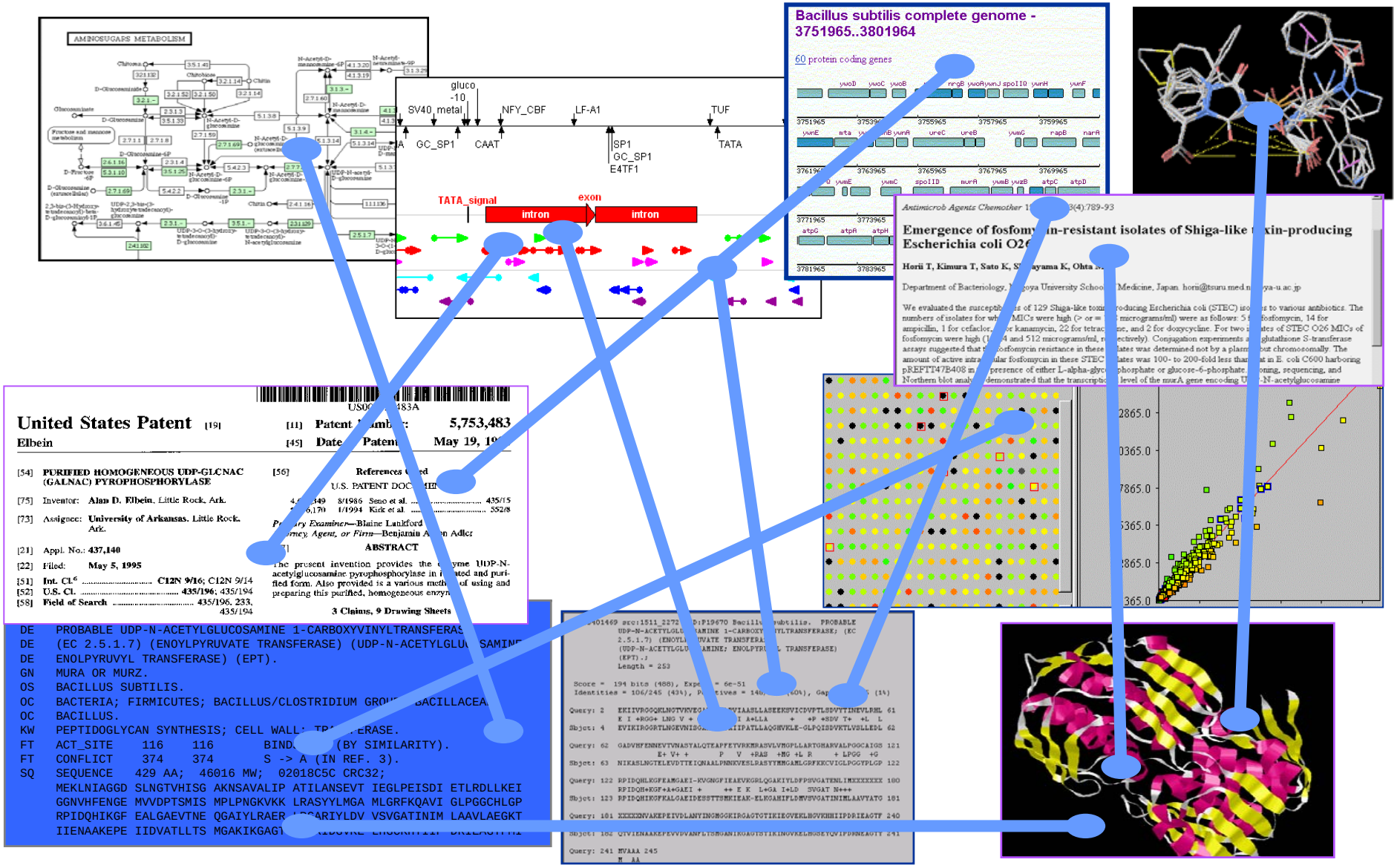
Science as Workflow: Integrating & Evolving Data and Computation

"*Science advances through tentative answers to a series of more & more subtle questions which reach deeper & deeper into the essence of natural phenomena*." (L. Pasteur)

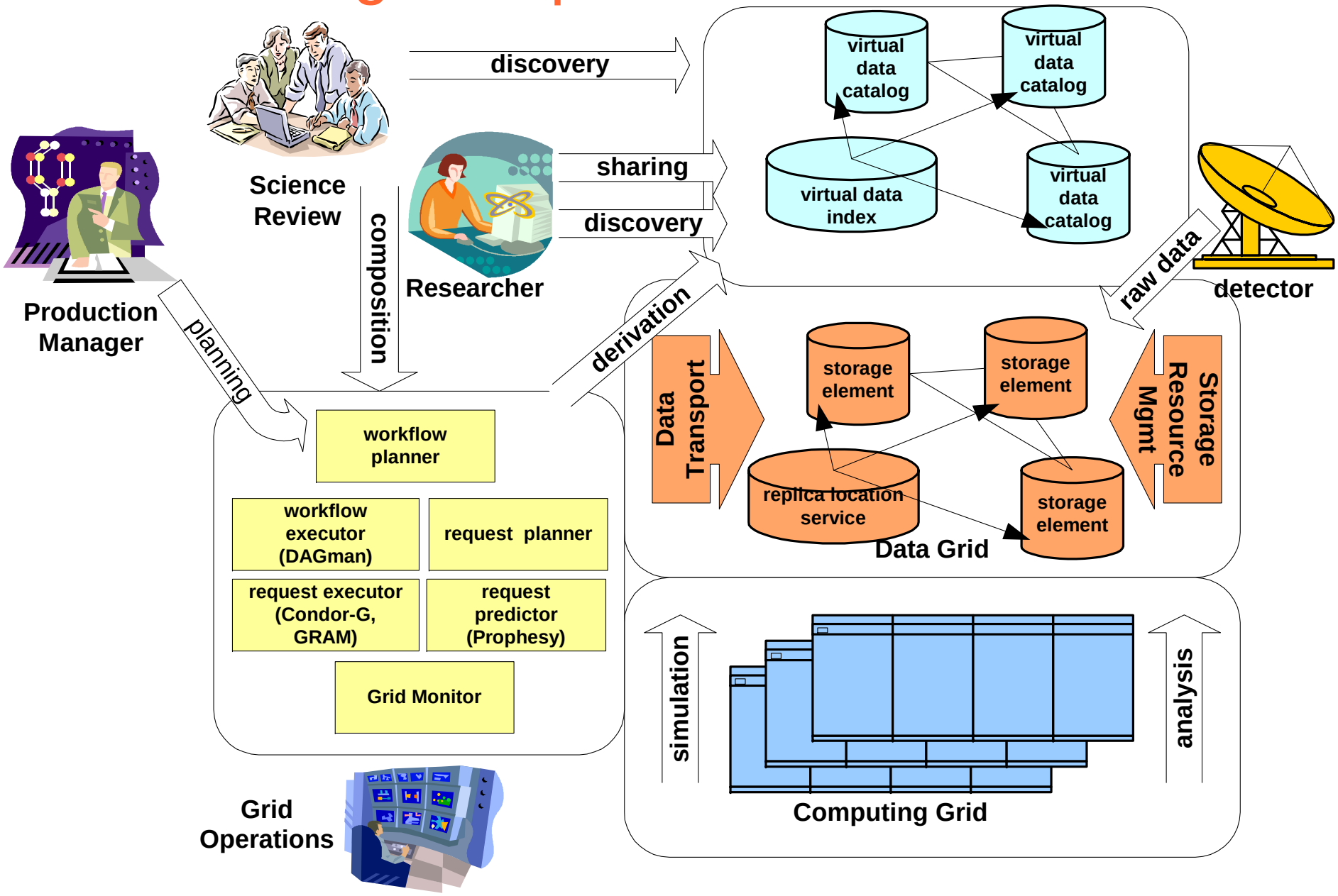
- In a networked petascale world, this means leveraging distributed data & computation to
  - Discover relevant data/hypotheses
  - Derive new tentative data/hypotheses from old
  - Publish new data/hypotheses for use by others
- In a collaborative framework in which many can contribute to, & draw from, the whole

Integrating Across Different Types of Information

[source: GlaxoSmithKline via Carole Goble]



Virtual Data Research: Sharing Recipes As Well as Data



Cyberinfrastructure for Life Sciences

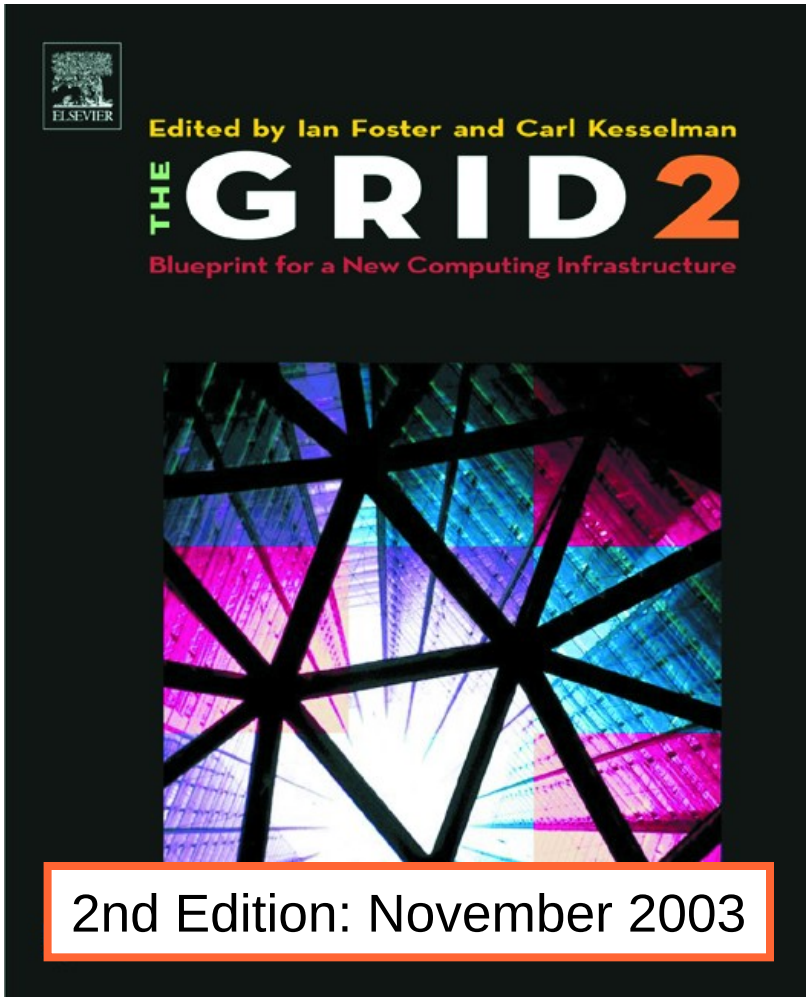
- Leverage emerging "eScience infrastructure"
  - Storage, data archives, networks, cluster
- Create an extensible service-rich environment
  - Workflows (standard protocols)
  - Experimental facilities: MRI, EM, arrays, …
  - Data update and access services
  - Computational services: BLAST, etc., etc.
  - Tools (informatics) suites
  - Ontologies & conceptual discovery frameworks
- Bio-applns-neutral: genomics, proteomics, …

Other Observations

- Building & applying cyberinstructure is hard
  - Need sustained effort and commitment
- Not clear that current funding models apply
  - Must support a critical mass of experts to build & operate infrastructure
  - Software is a vital part of the whole
- Life sciences need to obtain experience
  - Start multidisciplinary exploratory projects
- International cooperation is vital
  - Our partners “get it” and are hard at work

For More Information

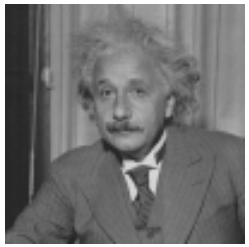
- The Globus Alliance
  - www.globus.org
- Global Grid Forum
  - www.ggf.org
- Background information
  - www.mcs.anl.gov/~foster
- GlobusWORLD 2004
  - www.globusworld.org
  - Jan 20–23, San Fran



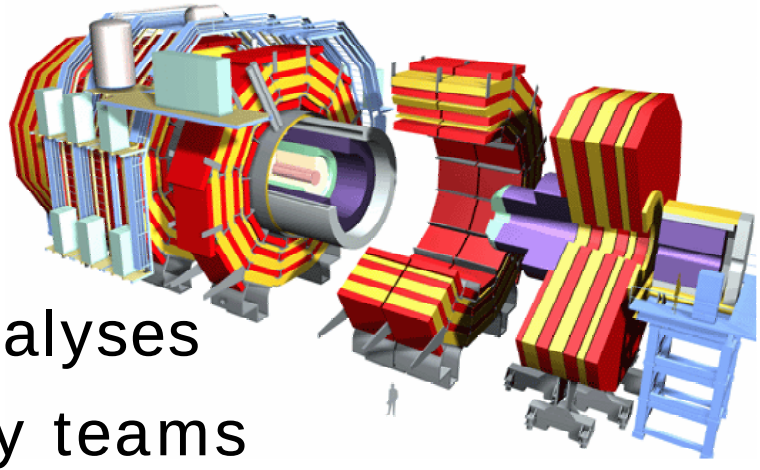
2nd Edition: November 2003

Evolution of Scientific Infrastructure

- Pre-electronic science
  - Theorize &/or experiment
  - “Computer” takes vacation
  - Individuals or small teams



- Post-electronic science
  - Specialized instruments
  - Very large databases
  - Computer simulations & analyses
  - Distributed multidisciplinary teams



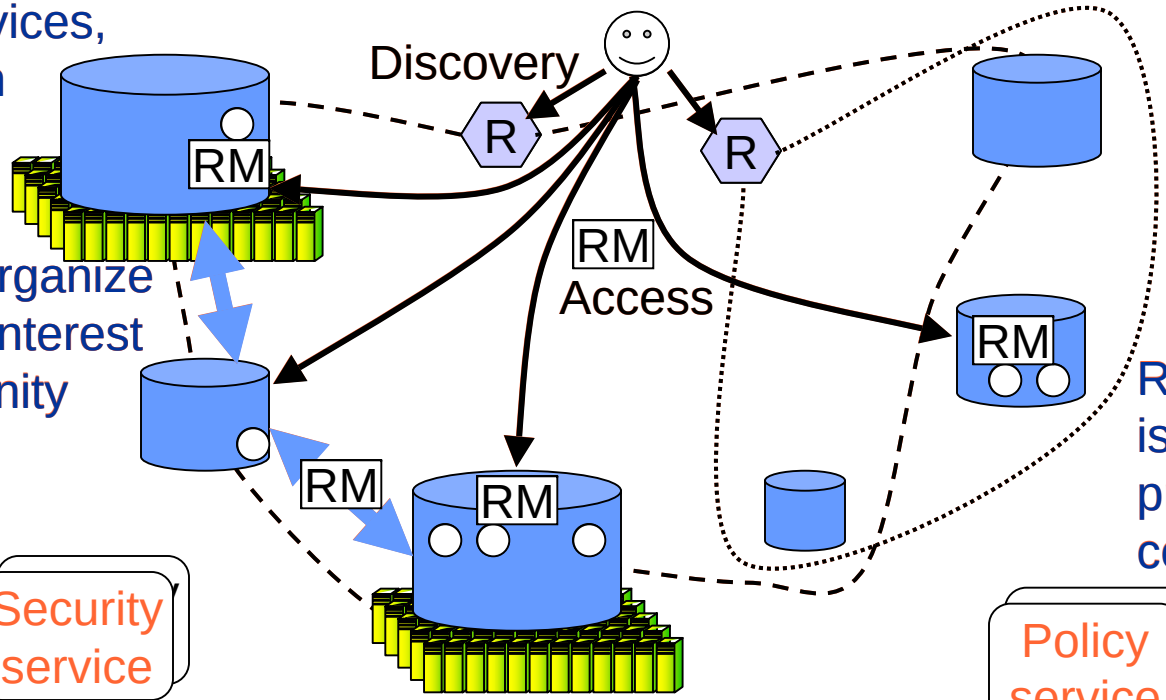
Technical Challenges

- Familiar issues: describe, name, discover, understand distributed data & other services
  - E.g., see Ken Buetow's talk
- But also new challenges relating to data volumes, resource counts, network speeds
  - Coordinate, manage, provision, & monitor of workflows & resources
  - Manage infinite demand for every resource, in a way that reflects scientific priorities
- And to dynamic, distributed communities
  - Establish, negotiate, manage, & evolve multi-organizational federations

Challenges of Scale

Many sources of data, services, computation

Security & policy must underlie access & management decisions



Registries organize services of interest to a community

Resource management is needed to ensure progress & arbitrate competing demands

Security service

Policy service

Data integration activities may require access to, & exploration of, data at many locations

Exploration & analysis may involve complex, multi-step workflows