Science Gateways

and their tremendous potential for science and engineering

VO Management in Production Grids Workshop

June 24, 2008

Nancy Wilkins-Diehr

TeraGrid Area Director for Science Gateways

wilkinsn@sdsc.edu

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

TeraGrid

Phenomenal Impact of the Internet on Worldwide Communication and Information Retrieval

Only 16 years since the release of Mosaic!

- Implications on the conduct of science are still evolving
  - 1980's, Early gateways, National Center for Biotechnology Information BLAST server, search results sent by email, still a working portal today
  - 1992 Mosaic web browser developed
  - 1995 "International Protein Data Bank Enhanced by Computer Browser"
  - 2004 TeraGrid project director Rick Stevens recognized growth in scientific portal development and proposed the Science Gateway Program
- Simultaneous explosion of digital information
  - Emerging analysis needs in a variety of scientific areas
    - Bioinformatics
  - Sensors, telescopes, satellites, digital images and video
  - #1 machine on Top500 today is as powerful as *all combined* entries on the first list in 1993

Rapid Advances in Web Usability

- First generation
  - Static Web pages
- Second generation
  - Dynamic, database interfaces, cgi
  - Lacked the ease of use of desktop applications
- Third generation
  - True networked and internetworked applications that enable dynamic two-way, even multi-way, communication and collaboration on the Web.
  - Remarkable new uses of the Web in the organizational workplace and on the Internet

Source: Screen Porch White Paper, The University of Western Ontario (1996)

PEW / INTERNET

PEW INTERNET & AMERICAN LIFE PROJECT

The convenience of getting scientific material on the web opens doors to better attitudes and understanding of science.

November 20, 2006

John B. Horrigan, Associate Director

The Internet as a Resource for News and Information about Science: Summary of Findings at a Glance
40 million Americans rely on the internet as their primary source for news and information about science.
For home broadband users, the internet and television are equally popular as sources for science news – and the internet leads the way for young broadband users.
The internet is the source to which people would turn first if they need information on a specific scientific topic.
The internet is a research tool for 87% of online users. That translates to 128 million adults.
Consumers of online science information are fact-checkers of scientific claims. Sometimes they use the internet for this, other times they use offline sources.
Convenience plays a large role in drawing people to the internet for science information.
Happenstance also plays a role in users' experience with online science resources. Two-thirds of internet users say they have come upon news and information about science when they went online for another reason.
Those who seek out science news or information on the internet are more likely than others to believe that scientific pursuits have a positive impact on society.
Internet users who have sought science information online are more likely to report that they have higher levels of understanding of science.
Between 40% and 50% of internet users say they get information about a specific topic using the internet or through email.
Search engines are far and away the most popular source for beginning science research among users who say they would turn first to the internet to get more information about a specific topic.
Half of all internet users have been to a website which specializes in scientific content.
Fully 59% of Americans have been to a science museum in the past year.
Science websites and science museums may serve effectively as portals to one another.

SDSC

SAN DIEGO SUPERCOMPUTER CENTER

http://www.pewinternet.org/pdfs/PIP_Exploratorium_Science.pdf

TeraGrid

NSF (my sponsor) has long recognized the importance of science and technology interactions

- Interdisciplinary programs did much to facilitate application-technology integration and develop standard tools
  - 1997 PACI Program
    - "Shotgun marriages" between technologists and application scientists
      - A few groups served as path finders and benefited tremendously
      - NPACI neuroscience thrust in 1997 leads to Telescience portal and BIRN in 2001
  - Information Technology Research (ITR)
  - NSF Middleware Initiative (NMI)
    - Plug and play tools so more groups can benefit

NSF Continues Its Leadership Today

What Will Lead to Transformative Science?

- "Virtual environments have the potential to enhance collaboration, education, and experimentation in ways that we are just beginning to explore."
- "In every discipline, we need new techniques that can help scientists and engineers uncover fresh knowledge from vast amounts of data generated by sensors, telescopes, satellites, or even the media and the Internet."

NSF & Congress

Testimony

Dr. Arden L. Bement, Jr.
Director
National Science Foundation

Testimony
Before the Senate Commerce, Science & Transportation
Subcommittee on Technology, Innovation, and Competitiveness

April 19, 2007

Gateways are a terrific example of interfaces that can support transformative science

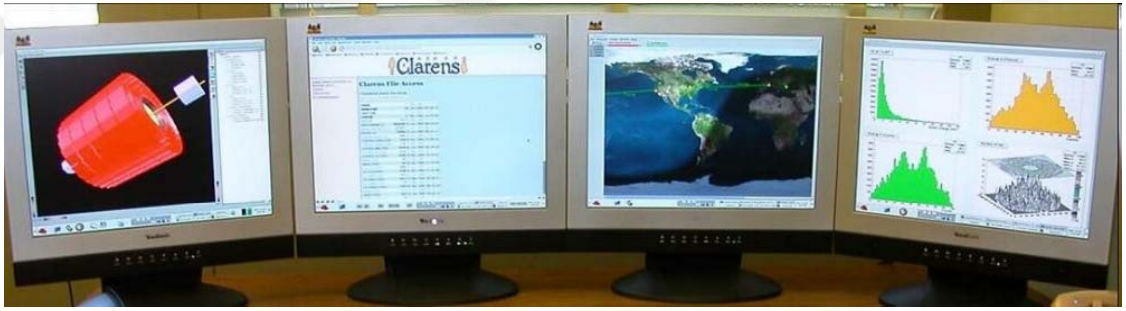
National Science Foundation
WHERE DISCOVERIES BEGIN

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

TeraGrid

Science Gateways are a Natural Extension of Internet Developments

- 3 common types of gateway
  - Web portal with users in front and services in back
  - Client server model where application programs running on users' machines (i.e. workstations and desktops) and accesses services
  - Bridges across multiple grids, allowing communities to utilize both community developed grids and shared grids
- Continued rapid changes ahead, must be adaptable, gateways can provide some nimbleness



Gateway Idea Resonates with Scientists

- Capabilities provided by the Web are easy to envision because we use them in every day life
- Researchers can imagine scientific capabilities provided through a familiar interface

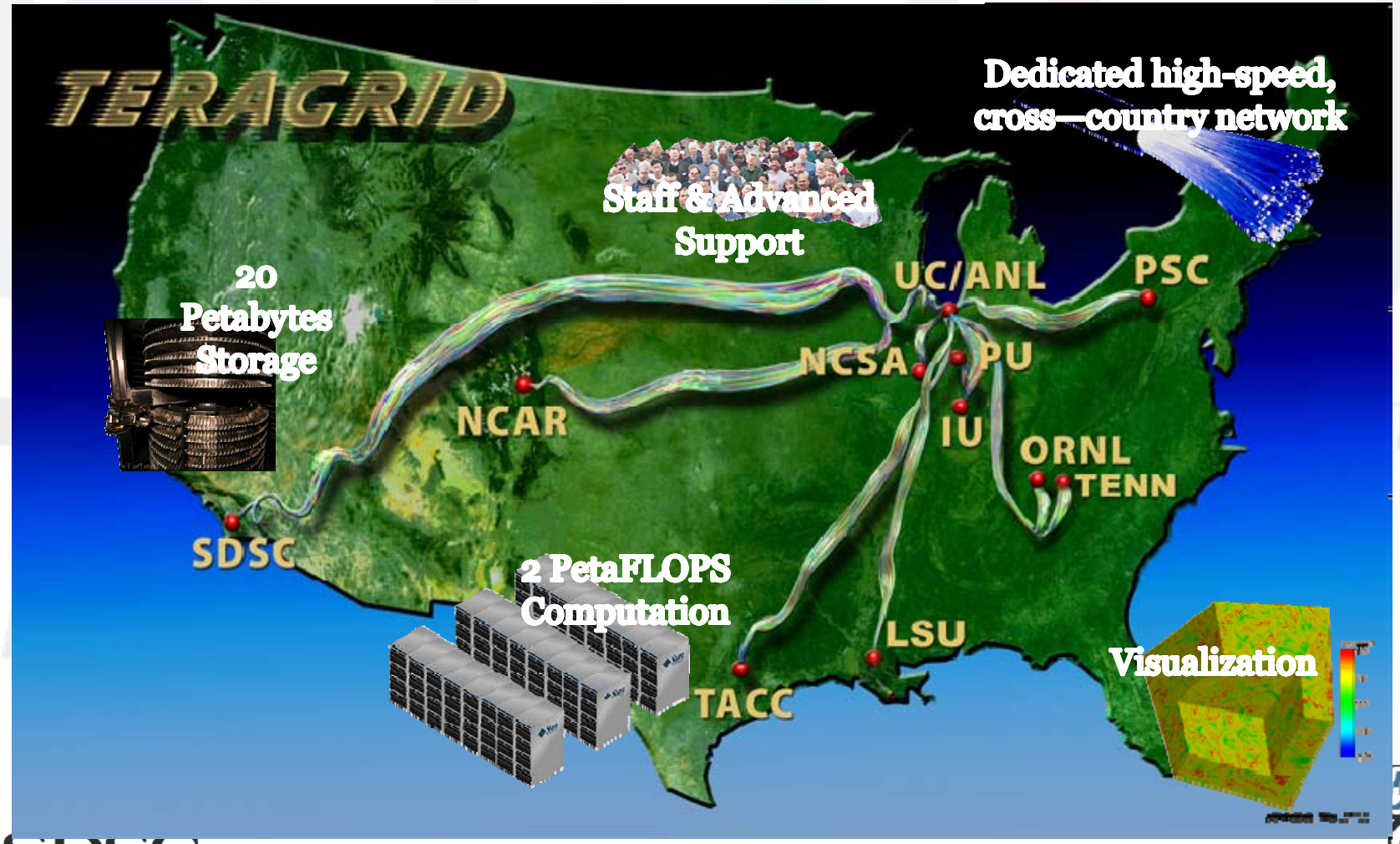
- Groups resonate with the fact that gateways are designed by communities and provide interfaces understood by those communities
  - But also provide access to greater capabilities on the back end without the user needing to understand the details of those capabilities
  - Scientists know they can undertake more complex analyses and that's all they want to focus on
- But this seamless access doesn't come for free. It all hinges on very capable developers

Tremendous Opportunities Using the Largest Shared Resources - Challenges too!

- What's different when the resource doesn't belong just to me?
  - Resource discovery
  - Accounting
  - Security
  - Proposal-based requests for resources (peer-reviewed access)
    - Code scaling and performance numbers
    - Justification of resources
    - Gateway citations
- Tremendous benefits at the high end, but even more work for the developers
- Potential impact on science is huge
  - Small number of developers can impact thousands of scientists
  - But need a way to train and fund those developers and provide them with appropriate tools

What is the TeraGrid?

A unique combination of fundamental CI components



Opportunities and Challenges as a Virtual Organization (TeraGrid)

- Full vision of cyberinfrastructure
  - Data, compute, visualization, workflows
  - But need to do a better job of representing the capabilities to researchers
  - Creating prototypes for others to follow
  - Never underestimate the value in keeping things SIMPLE
- Work with top notch people regardless of location
  - Better for end users
    - Single request process for all types of resources
    - Single place for documentation
- But must work harder
  - To sustain momentum in projects
    - Set a few high-level goals
    - Clear management structure
      - Individual responsibility
      - Project accountability
  - To provide clarity for users

Evolution of the Gateway Program

- 2004 "TeraGrid Science Gateway" term originates
  - We will help them build gateway portals that leverage TeraGrid capabilities and provide web-based interfaces to community tools
- 2005 Initial Gateway requirements analysis team
  - Areas of identified commonality include:
    - Web services, auditing, community accounts, flexible allocations, scheduling, outreach
  - Needs of command-line supercomputing users fairly well defined
    - ssh to tg-login
    - Data transfer
    - Software
      - MPI, math libraries, domain software
    - Compilers
    - Batch queue submission
    - Help desk
  - **Need to address Gateway developer needs just as efficiently**

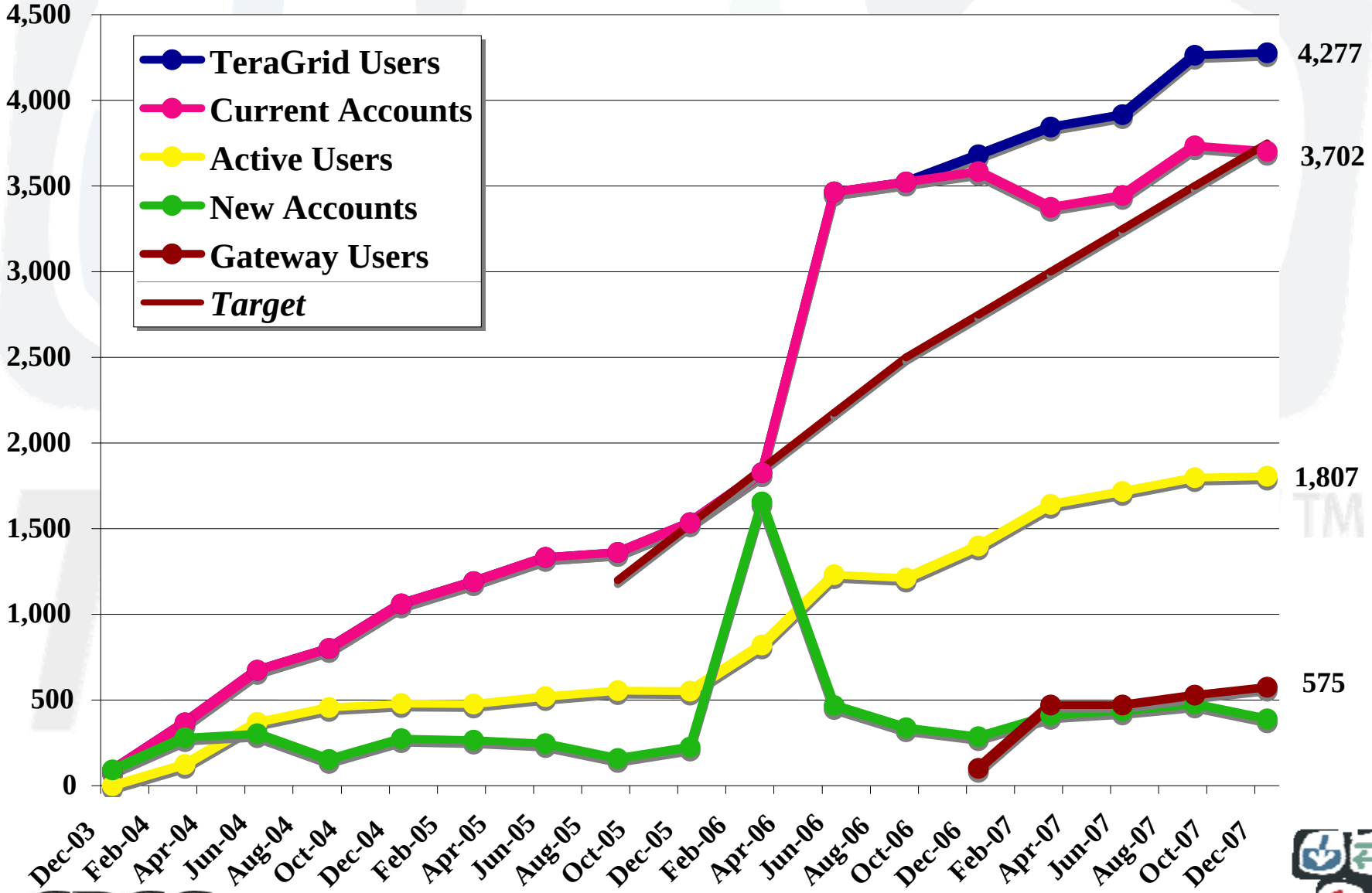
Ongoing Work to Meet Common Needs

- Web Services
  - GT4 deployment, identification of remaining capabilities
  - Information services, MDS
  - **Registry of Gateway services**
  - TG-specific "where can I run soonest" with QBETS – DONE
- Auditing
  - GRAM audit to retrieve usage information for individual compute jobs - DONE
  - **GridShib**
    - **Counting gateway users, individualized accounting, increased security**
- **Community Accounts**
  - **Policy finalized, security approaches being tested by RPs**
  - GridShib development, testing with gateways
- Resource requests - DONE
  - Collaboration with reviewers to develop guidelines for Gateway PIs
  - Adapt to usage uncertainties, ability to assess impact, Gateway management structure
- Scheduling
  - Metascheduling
  - On-demand via SPRUCE framework
- Outreach
  - Pathways project
    - Gateway use by educators
    - Training MSI students to build Gateways
- Documentation
  - Extensive wiki information transformed into navigable documentation
- Gateway Hosting
  - Available at IU through peer review
- Staff Support
  - Targeted support, general capabilities, production coordinator

Variety of Gateways Available Today

Title	Discipline
Open Science Grid (OSG)	Advanced Scientific Computing
Special PRiority and Urgent Computing Environment (SPRUCE)	Advanced Scientific Computing
Massive Pulsar Surveys using the Arecibo L-band Feed Array (ALFA)	Astronomical Sciences
National Virtual Observatory (NVO)	Astronomical Sciences
Linked Environments for Atmospheric Discovery (LEAD)	Atmospheric Sciences
Computational Chemistry Grid (GridChem)	Chemistry
Computational Science and Engineering Online (CSE-Online)	Chemistry
Network for Earthquake Engineering Simulation (NEES)	Earthquake Hazard Mitigation
GEON(GEOsciences Network) (GEON)	Earth Sciences
Network for Computational Nanotechnology and nanoHUB	Emerging Technologies Initiation
TeraGrid Geographic Information Science Gateway (GISolve)	Geography and Regional Science
CIG Science Gateway for the Geodynamics Community	Geophysics
QuakeSim (QuakeSim)	Geophysics
The Earth System Grid (ESG)	Global Atmospheric Research
National Biomedical Computation Resource (NBCR)	Integrative Biology and Neuroscience
Developing Social Informatics Data Grid (SIDGrid)	Language, Cognition, and Social Behavior
Neutron Science TeraGrid Gateway (NSTG)	Materials Research
Biology and Biomedicine Science Gateway	Molecular Biosciences
Open Life Sciences Gateway (OLSG)	Molecular Biosciences
The Telescience Project	Neuroscience Biology
Grid Analysis Environment (GAE)	Physics
SCEC Earthworks Project	Seismology
TeraGrid Visualization Gateway	Visualization, Graphics, and Image Processing

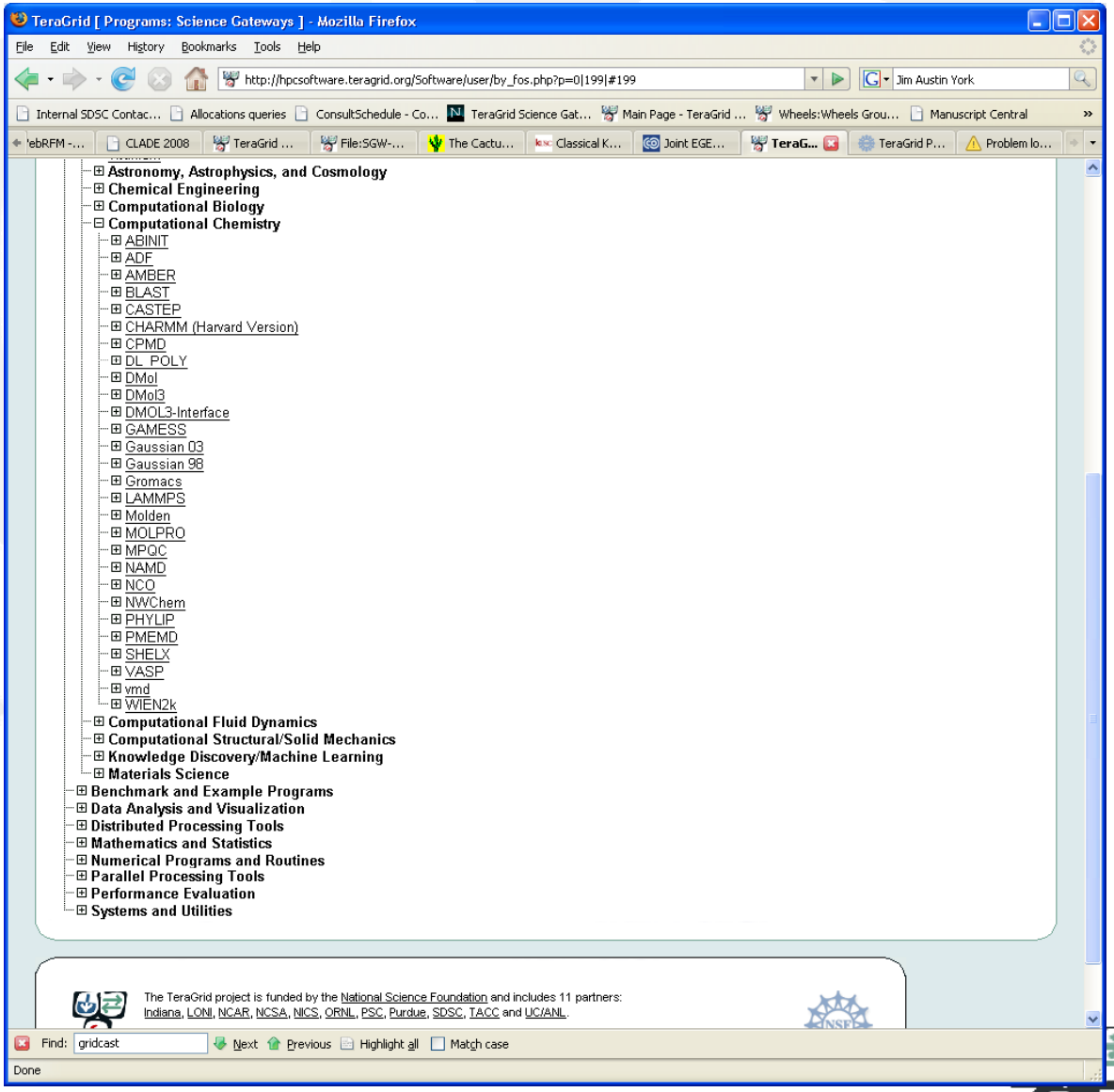
HPC User Community is Growing



Source: John Towns, TeraGrid Central Database

Software Discovery on TeraGrid Today

- Researchers view software catalog to see packages available on the command line
- Significant effort to discover software available through a Gateway

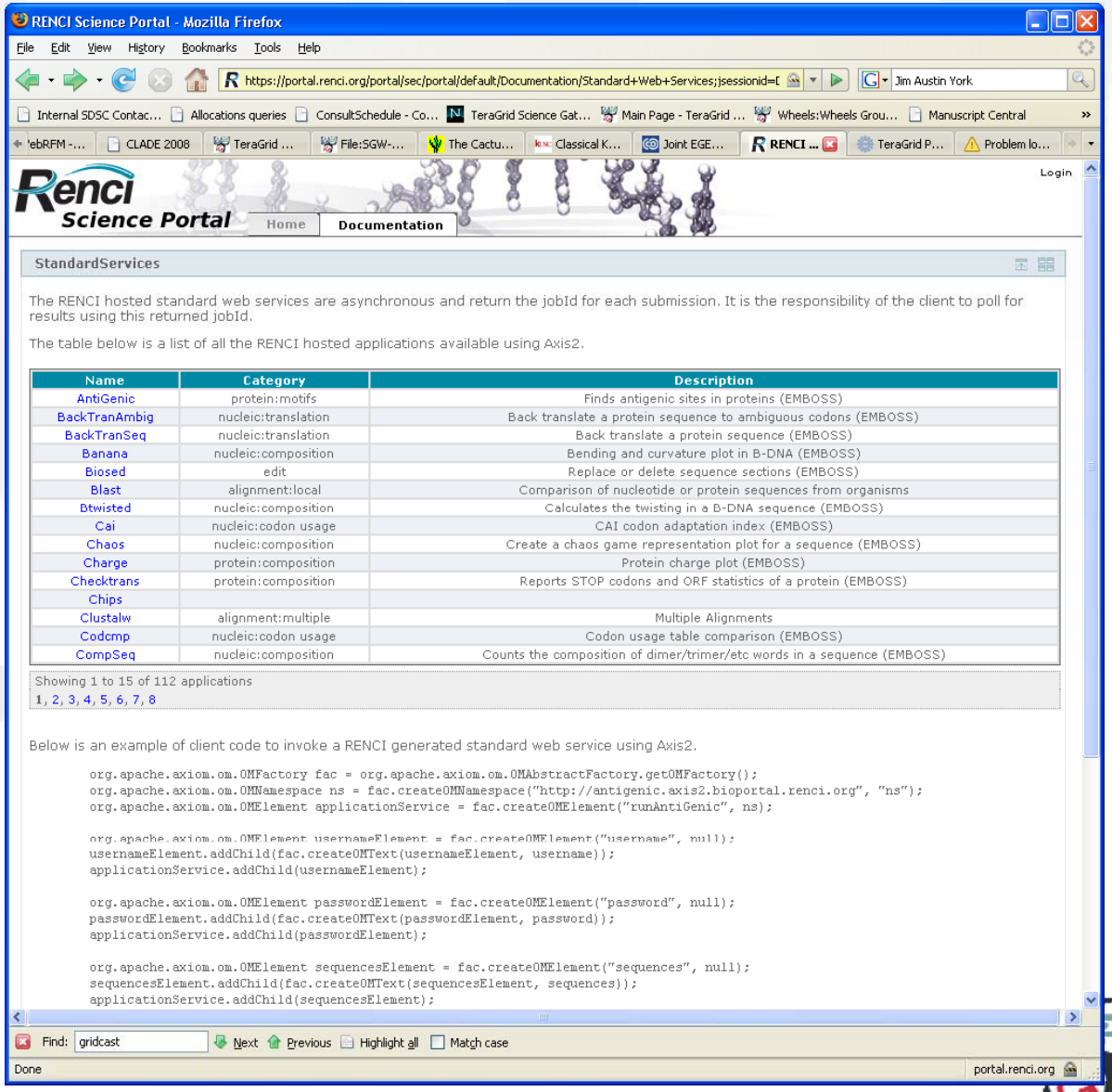


SDSC
SAN DIEGO SUPERCOMPUTER CENTER

TeraGrid

Software Discovery on TeraGrid Tomorrow

- Create registry of services offered by gateways for both end users and other developers to discover



SDSC
SAN DIEGO SUPERCOMPUTER CENTER

TeraGrid

Community Accounts
The Key to Gateways

- Early viewpoint
  - Scalability is great, we don't have to or want to know how many end users are using a community account through a gateway
- Current viewpoint
  - We want to understand the impact of gateways and that includes keeping records of how many unique users access TeraGrid through each gateway
    - Store this info in TeraGrid central database (TGCDB)
  - Additional security features are nice too
    - Blacklist individuals rather than pulling an entire community account
    - Emails associated with each job
    - IP filtering
  - TeraGrid could now do accounting for gateways rather than delegating

- So, how do we send attributes along with jobs associated with community accounts?

GridShib SAML Tools

Science Gateways extension

- Bind a SAML token to a proxy certificate signed by a gateway's community credential
  - *entityID*: a globally unique identifier for the SAML issuer (i.e., the gateway)
  - *name identifier*: a globally unique identifier for the authenticated user
  - *authentication statement*: a description of the act of authentication at the gateway
    - *authentication method*: an identifier specifying the method of authentication (e.g., password)
    - *authentication instant*: a date-time stamp indicating the exact time the authentication took place
    - *IP address*: the IP address of the user agent involved in the authentication event
  - *attribute statement*: a collection of user attributes
    - *isMemberOf attribute*: an attribute declaring the user's membership in a virtual organization (VO)
    - *mail attribute*: the user's e-mail address

But

- Will GridShib address all gateway architectures?
  - Non-Globus based? Any gateways not using certificates?
- One more hurdle developers must clear to use the TeraGrid
  - Currently many hoops, will this be the final straw?
    - Proposals to request time
    - Forms to be listed as a gateway
    - Forms to request a community software area
    - Forms to request community accounts
    - Must report quarterly on unique individuals using gateways
    - Risk and vulnerability assessment
    - Etc.

A Second Community Account Challenge

- Variety of approaches to secure accounts amongst Resource Provider (RP) sites
  - Gateway Security Summit, January, 2008
    - 30 attendees
      - Security representatives from 8 TeraGrid RP sites, as well as the Ohio Supercomputer Center and OSG
      - Gateway developers representing 11 projects
      - TeraGrid staff working in accounting, documentation and attribute-based authentication
  - TeraGrid RPs to describe implementation approaches by March, 2008
    - Good news
      - Some sites have chosen to impose no additional restrictions on community accounts
    - Bad news
      - Those that are imposing restrictions are each doing it differently
        - » Good news within the bad news
          - » Job security for me as a negotiator to improve this situation!

- Site 1
  - Set up site-specific “acceptable use policy” with PI
  - Identify a minimal group of developers
  - Site tells PI that if any developer or community account is implicated in a compromise, then the entire group of accounts will be locked until the scope of the breach can be determined.
  - Each community account gets a project directory for community executables that can only be written by developers.
  - No shell access for community accounts
    - Developers can 'sudo su - {commacct}‘to mimic commandline behavior of community account
  - Community account can only run a series of specifically-approved paths
  - Community accounts to run any programs which they could modify
    - But developers can make modifications

- Site 2
  - Community Shell to allow pre-ws gram submission of reviewed job scripts
  - May add restrictions via the scheduler(such as maximum job length, etc.)
- Site 3
  - Community Shell to allow WS-GRAM job submission of approved job scripts to community accounts
  - No interactive shell access to community accounts.
  - Attribute-based auditing and authorization for community accounts.
- Site 4
  - Leverage features of the Globus grid toolkit to enable remote job submission within a service oriented architecture.
  - No shell logins for community accounts
- Need to make developers' lives easier or we will not have a successful Gateway program

The Community Shell - Commsh

- Mitigate this potential for abuse by placing restrictions on the applications that may be run by a community account
- Configuration file explicitly lists which applications (commands) the account can run
  - It can further specify the form of the parameters (arguments) passed to those applications.
- File will establish a directory for applications used by the community
  - Any applications within this single directory will be allowed
  - Applications outside of this directory will be restricted
  - Applications can then be added to or removed from this directroy as needed.
- Community account itself does not have access to modify the contents of this directory.
  - If the account can upload scripts into this applications directory, it can execute arbitrary code on the system and thereby circumvent this security precaution
- Recommended that a second, community administrator account be established for each community.
  - Used by the community administrators to manage which applications the community account can run.
  - Has its own set of credentials (password or private key) that are separate from the community account's.
    - The administrator account credentials should **not** be stored on the same server as the community account credentials, so that a compromise of the community gateway server will not compromise the administrative account.
- http://security.ncsa.uiuc.edu/research/commaccts/docs/howto.php
  - Downloads, installation instructions

Source: Kevin Price, developed at NCSA

Commsh implementation at PSC

- ***Securing a Community Account with commsh Step-by-Step:***
- *1. Installation of the commsh utility.*
- *2. Testing community shell (including Globus job submission to a community shell*
- *restricted account).*
- *3. Set community account's shell to /path/to/commsh/bin/commsh*
- 4. *Edit /path/to/commsh/etc/commsh.conf to include*
- *#Read the file 'commsh.rc' in the user's home directroy*
- ReadConfig ~/.commsh.rc
- # Allow all users except for 'root' to run commsh
- AllowUser **communityUser**
- DenyUser root
- *5. Retrieve **tested known-working job script from technical staff of Community and***
- *place it in communityUser's home directory*
- *6. Remove write access to the communityUser's home directory and all files*
- *included within.*
- *7. Edit ~/.commsh.rc to include*
- DirectAccess /complete/path/to/home/**communityUser/jobwrapper**

A Third Community Account Challenge

- Who is allowed access to supercomputers?
  - "To qualify for an award, the principal investigator (PI) must be a researcher or educator at a U.S. academic or non-profit research institution. A qualified advisor may apply for an allocation for his or her class, but a high school, undergraduate or graduate student may not be a PI. A postdoctoral researcher can also be a PI. (After receiving an award, PIs can request that students be granted accounts to use the PI's allocation.)"
    - **Foreign organizations** - NSF rarely provides support to foreign organizations. NSF will consider proposals for cooperative projects involving US and foreign organizations, provided support is requested only for the US portion of the collaborative effort.
    - **Other Federal agencies** - NSF does not normally support research or education activities by scientists, engineers or educators employed by Federal agencies or Federally Funded Research and Development Centers (FFRDCs).
- What about export control?
  - May 21, 2008 "**UT professor indicted by grand jury"**
- TeraGrid user responsibility form
- What about gateways?

How does Export Compliance relate to the TeraGrid? (i.e. Why do we care?)

Aside from the strict penalties for non-compliance, TeraGrid RPs need to consider:

- TeraGrid RPs who are Universities have varying policies on Research, but in general they have one mission - to *create and disseminate knowledge openly*.
- Websites of individual TeraGrid parent organizations contain information (FAQs, policy statements) related to export compliance. RPs must comply with policy of associated organization.
- There does not appear to be a cohesive TeraGrid-wide policy or procedure for dealing with Export Compliance issues.
- Over the past several years, there has been an increased concern over national security, particularly by Commerce and State. This has resulted in funding agencies trying to impose more restrictive measures on what was once considered fundamental research.

Source: David Moses, PSC

How should we Manage Controlled Information?

Some ideas and suggestions that may help assist us in implementing a compliance program:

- Dililgence – A good part of this battle is demonstrating diligence to the government agencies, especially in the case of an audit or other inquiry.
- The early project stage – Recent updates to the procedures at many Universities show that early evaluation of the project helps identify potential issues with Export Compliance.
- Query the researchers - Ask specific questions regarding the nature of the project. Although sometimes it is sensitive to ask citizenship, it is never a problem to ask if a user if their access to certain technical information or materials is in compliance with all U.S. Export regulations. If they are not sure, they can ask.
- During the project – Keeping track of the status of ongoing projects may aid in compliance.
- Organization – Elect a compliance representative from each of the TG member sites. That person would be responsible for keeping their respective site apprised of export issues as they may arise, changes in regulations, etc.
- Communication – As with most management issues such as this, it is always best to encourage as much communication as possible.

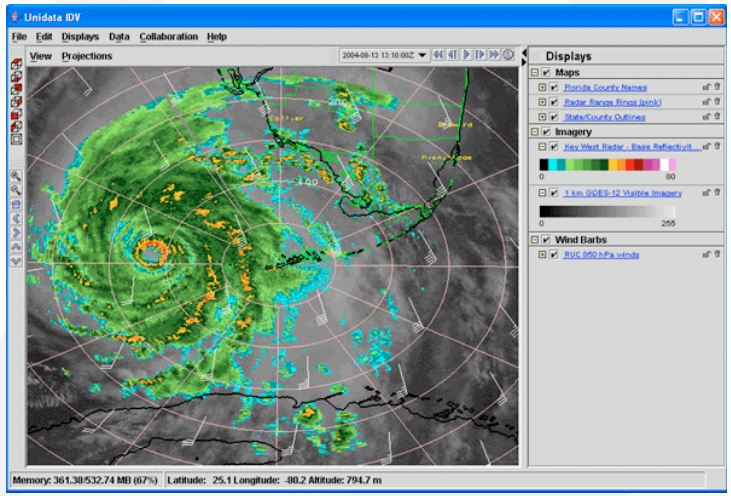
Source: David Moses, PSC

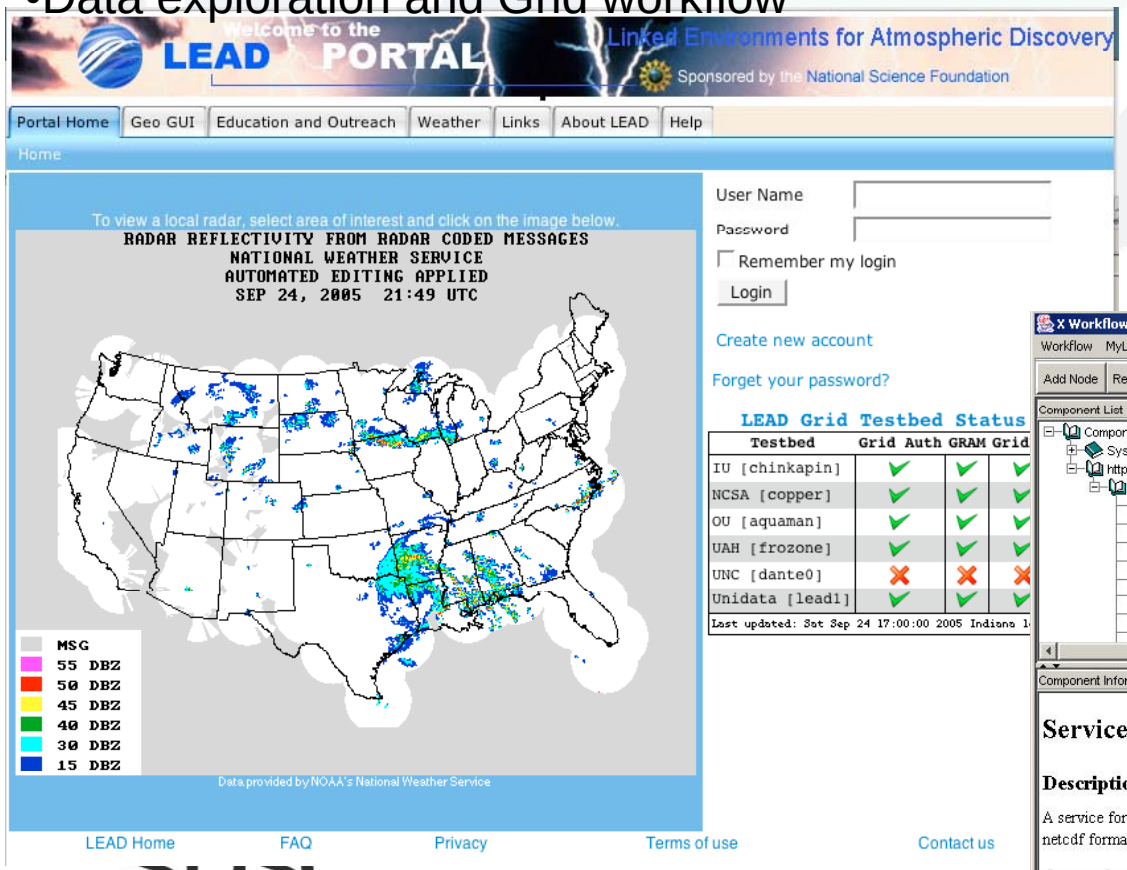
Highlights: NanoHub Explosive User Growth

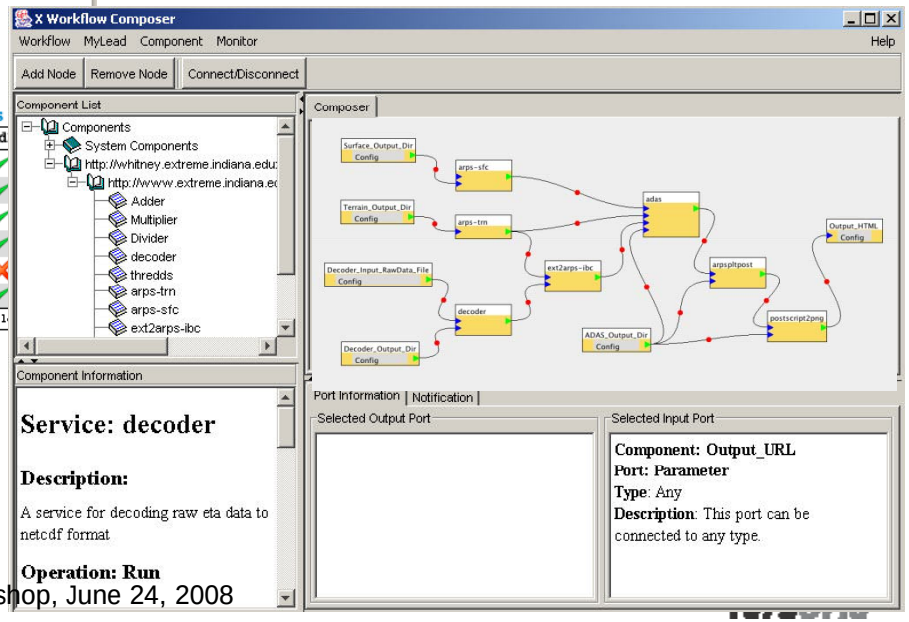
- In past 12 months
  - 68,975 users
    - 43% from U.S.
  - 25,187 course downloads
  - 8,287 podcast downloads
  - 371 online meetings
- Full featured gateway
  - Simulation tools, curricula, multimedia, user contributions, collaborations

Linked Environments for Atmospheric Discovery

- Providing tools that are needed to make accurate predictions of tornados and hurricanes
  - Meteorological data
  - Forecast models
  - Analysis and visualization tools
- Data exploration and Grid workflow

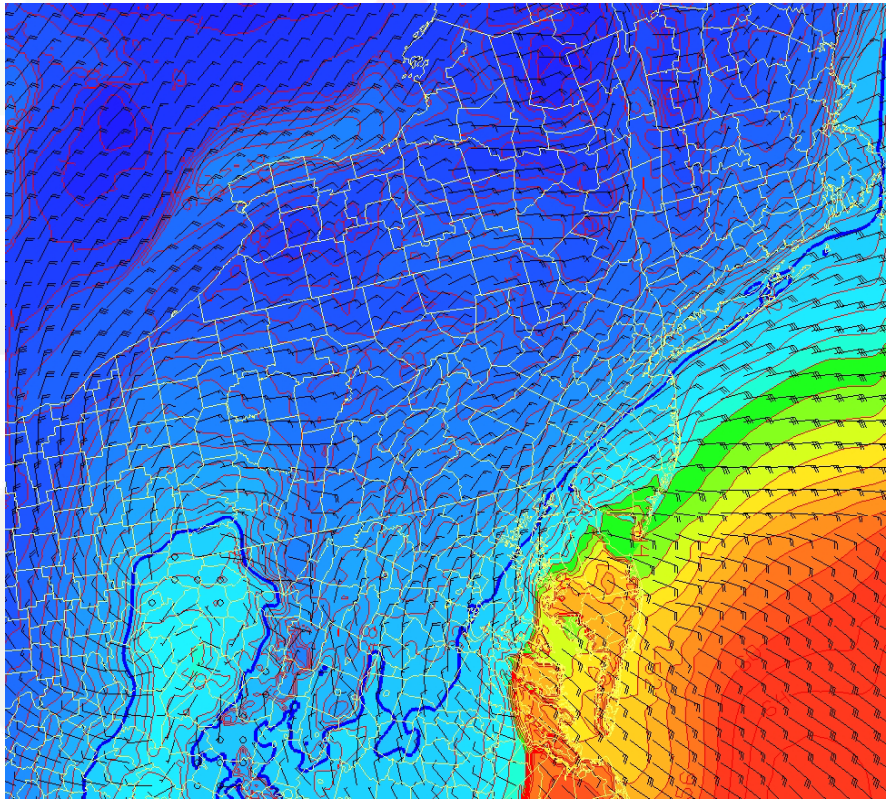




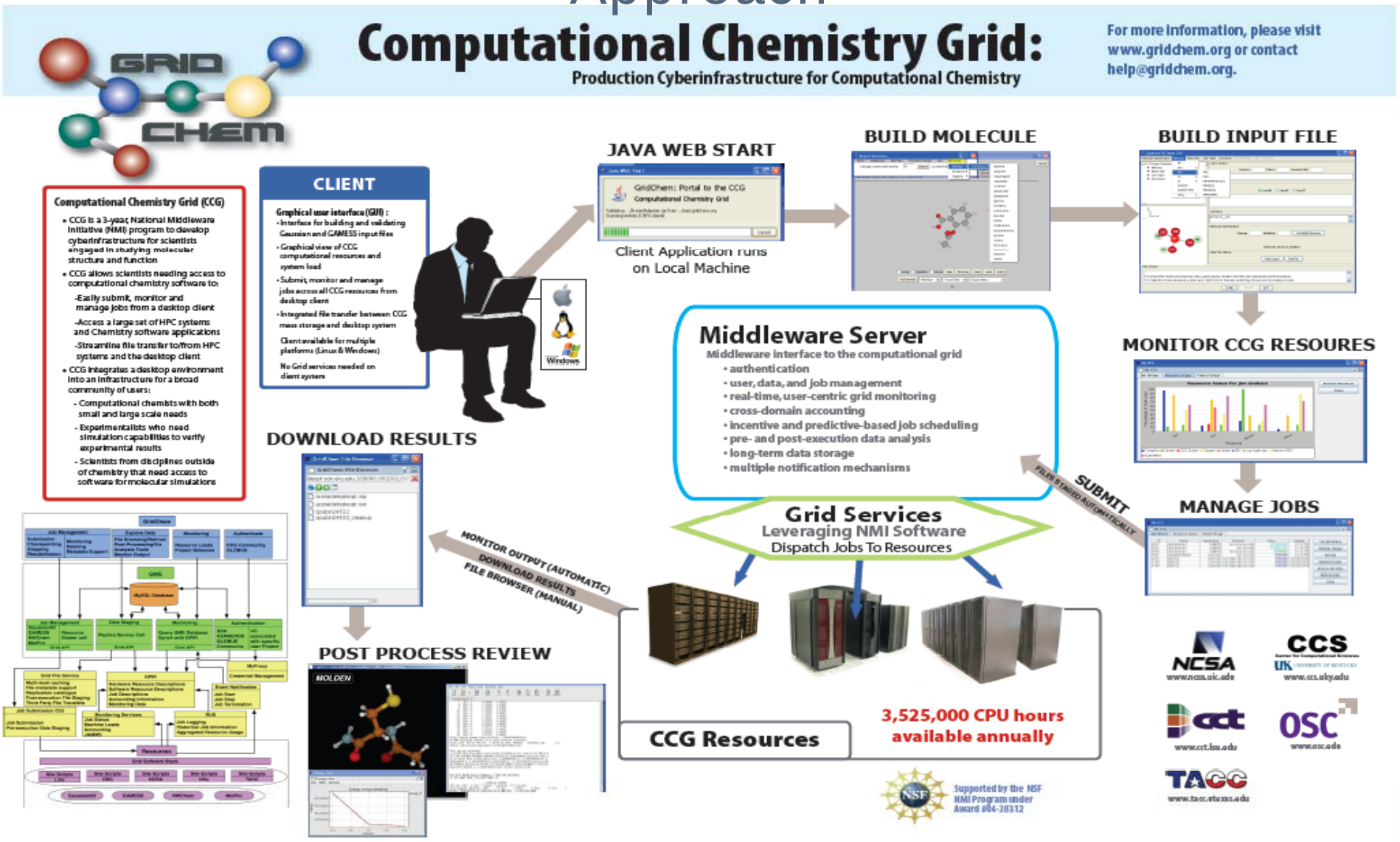


Highlights: LEAD Inspires Students
Advanced capabilities regardless of location

- A student gets excited about what he was able to do with LEAD
- "Dr. Sikora:Attached is a display of 2-m T and wind depicting the WRF's interpretation of the coastal front on 14 February 2007. It's interesting that I found an example using IDV that parallels our discussion of mesoscale boundaries in class. It illustrates very nicely the transition to a coastal low and the strong baroclinic zone with a location very similar to Markowski's depiction. I created this image in IDV after running a 5-km WRF run (initialized with NAM output) via the LEAD Portal. This simple 1-level plot is just a precursor of the many capabilities IDV will eventually offer to visualize high-res WRF output. Enjoy!"
- Eric (email, March 2007)



GridChem Employs a Client-Server Approach



GridChem Used for Production Science

- **Chemical Reactivity of the Biradicaloid (HO...ONO) Singlet States of Peroxynitrous Acid. The Oxidation of Hydrocarbons, Sulfides, and Selenides**. Bach, R. D et al. J. Am. Chem. Soc. 2005, 127, 3140-3155.

- **The "Somersault" Mechanism for the P-450 Hydroxylation of Hydrocarbons. The Intervention of Transient Inverted Metastable Hydroperoxides.** Bach, R. D.; Dmitrenko, O. J. Am. Chem. Soc. 2006, 128(5), 1474-1488.

- **The Effect of Carbonyl Substitution on the Strain Energy of Small Ring Compounds and their Six-member Ring Reference Compounds** Bach, R. D.; Dmitrenko, O. J. Am. Chem. Soc. 2006,128(14), 4598.
- **Azide Reactions for Controlling Clean Silicon Surface Chemistry: Benzylazide on Si(100)-2 x 1**
  Semyon Bocharov et al..
  J. *Am. Chem. Soc.,* **128** (29), 9300 -9301, 2006
- **Chemistry of Diffusion Barrier Film Formation: Adsorption and Dissociation of Tetrakis(dimethylamino)titanium on Si(100)-2 × 1**
  Rodriguez-Reyes, J. C. F.; Teplyakov, A. V.
  J. Phys. Chem. C.; 2007; 111(12); 4800-4808.
- **Computational Studies of [2+2] and [4+2] Pericyclic Reactions between Phosphinoboranes and Alkenes. Steric and Electronic Effects in Identifying a Reactive Phosphinoborane that Should Avoid Dimerization** Thomas M. Gilbert* and Steven M. Bachrach *Organometallics,* 26 (10), 2672 -2678, 2007.

cancer Bioinformatics Grid

Addressing today's challenges in cancer research and treatment

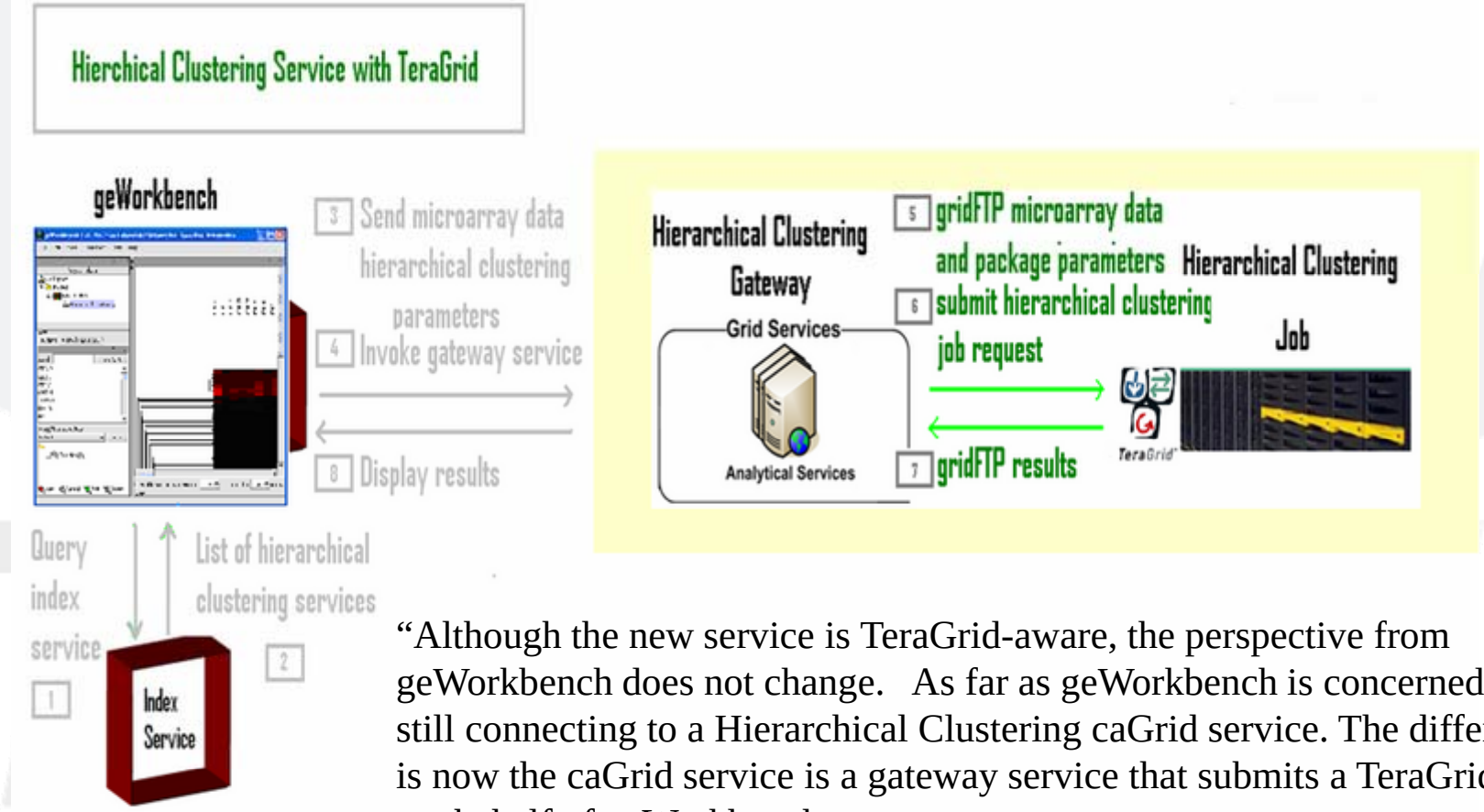
- The **mission** of caBIG™ is to develop a truly collaborative information network that accelerates the discovery of new approaches for the detection, diagnosis, treatment, and prevention of cancer, ultimately improving patient outcomes.
- The **goals** of caBIG™ are to:
- Connect scientists and practitioners through a shareable and interoperable infrastructure
- Develop standard rules and a common language to more easily share information
- Build or adapt tools for collecting, analyzing, integrating, and disseminating information associated with cancer research and care.

Source: cabig.cancer.gov

caBIG and TeraGrid

- caBIG conducted study of all Gateways
  - Pleased to discover that community accounts and web services will exactly meet their requirements
- TeraGrid resources incorporated into **geWorkbench**
  - an open source platform for integrated genomics used to
    - Load data from local or remote data sources.
    - Visualize gene expression and sequence data in a variety of ways.
    - Provide access to client- and server-side computational analysis tools such as t-test analysis, **hierarchical clustering**, self organizing maps, regulatory networks reconstruction, BLAST searches, pattern/motif discovery, etc.
      - Clustering is used to build groups of genes with related expression patterns which may contain functionally related proteins, such as enzymes for a specific pathway
    - Validate computational hypothesis through the integration of gene and pathway annotation information from curated sources as well as through Gene Ontology enrichment analysis.

geWorkbench Integrages TeraGrid Resources



"Although the new service is TeraGrid-aware, the perspective from geWorkbench does not change. As far as geWorkbench is concerned, it is still connecting to a Hierarchical Clustering caGrid service. The difference is now the caGrid service is a gateway service that submits a TeraGrid job on behalf of geWorkbench.
geWorkbench, however, does not notice this difference."

Source: http://wiki.c2b2.columbia.edu/informatics/index.php/GeWorkbench_Example

When is a gateway appropriate?

- Researchers using defined sets of tools in different ways
  - Same executables, different input
    - GridChem, CHARMM
  - Creating multi-scale workflows
  - Datasets
- Common data formats
  - National Virtual Observatory
  - Earth System Grid
  - Some groups have invested significant efforts here
    - caBIG, extensive discussions to develop common terminology and formats
    - BIRN, extensive data sharing agreements
- Difficult to access data/advanced workflows
  - Sensor/radar input
    - LEAD, GEON

Tremendous Potential for Gateways

- In only 16 years, the Web has fundamentally changed human communication
- Science Gateways can leverage this amazingly powerful tool to:
  - Transform the way scientists collaborate
  - Streamline conduct of science
  - Influence the public's perception of science
- Like e-commerce, Science Gateways need to build trust in the infrastructure, tools, and methods that they use
- Unlike the public or commercial arena, scientists will be vested in these gateways
  - Science Gateways will need to build trust in the organization behind them. Gateways need to have continuity
- High end resources can have a profound impact
- **The future is very exciting!**

Thank you for your attention

and for the trip to Boston

- wilkinsn@sdsc.edu
- www.teragrid.org

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

TeraGrid