High-Performance Computing and Distributed Systems

Some Observations from TeraGrid

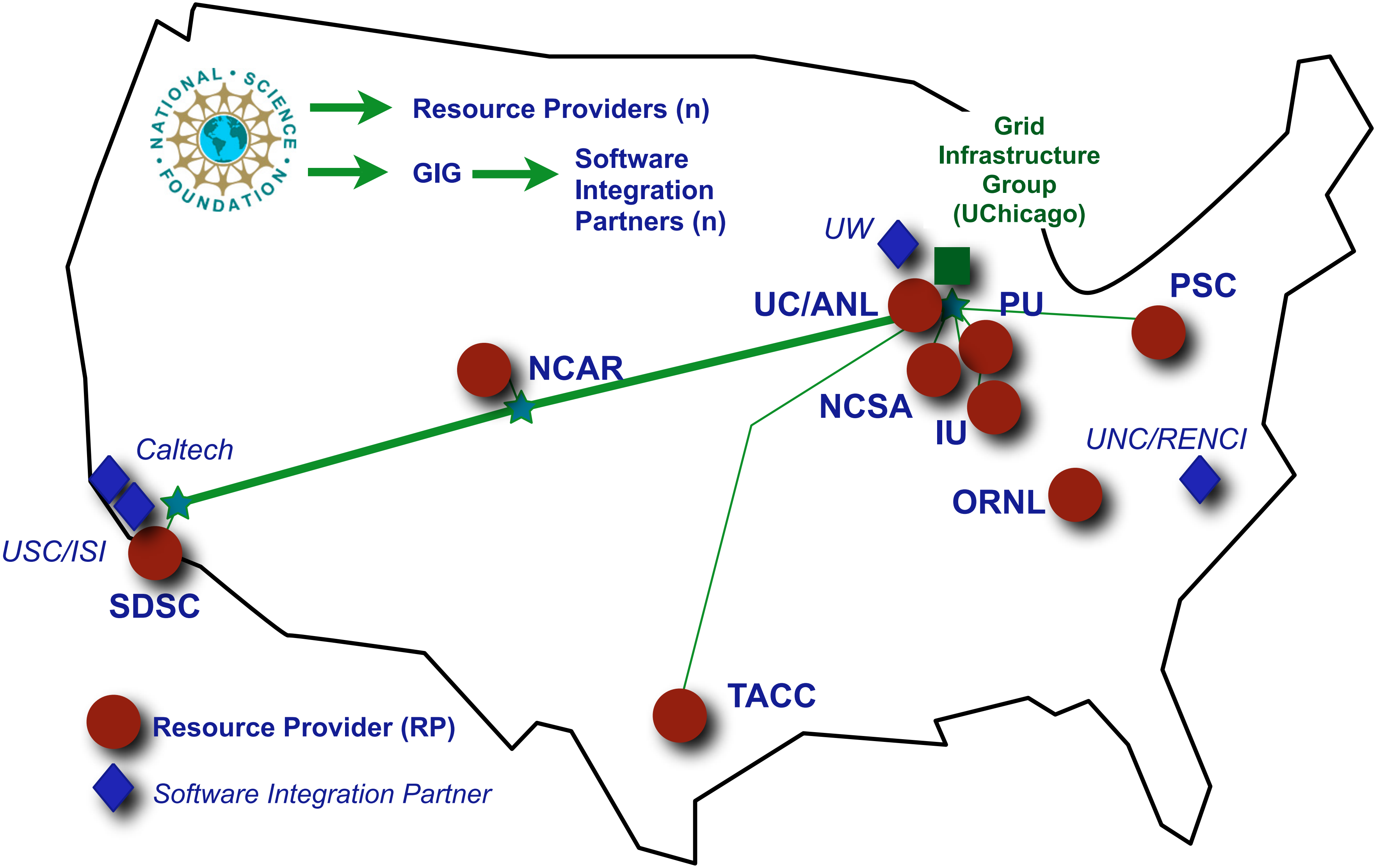
Charlie Catlett
CIO, Argonne National Laboratory
Chairman, TeraGrid Forum
Senior Fellow, Computation Institute
The University of Chicago and Argonne National Laboratory

OGF20 May 2007

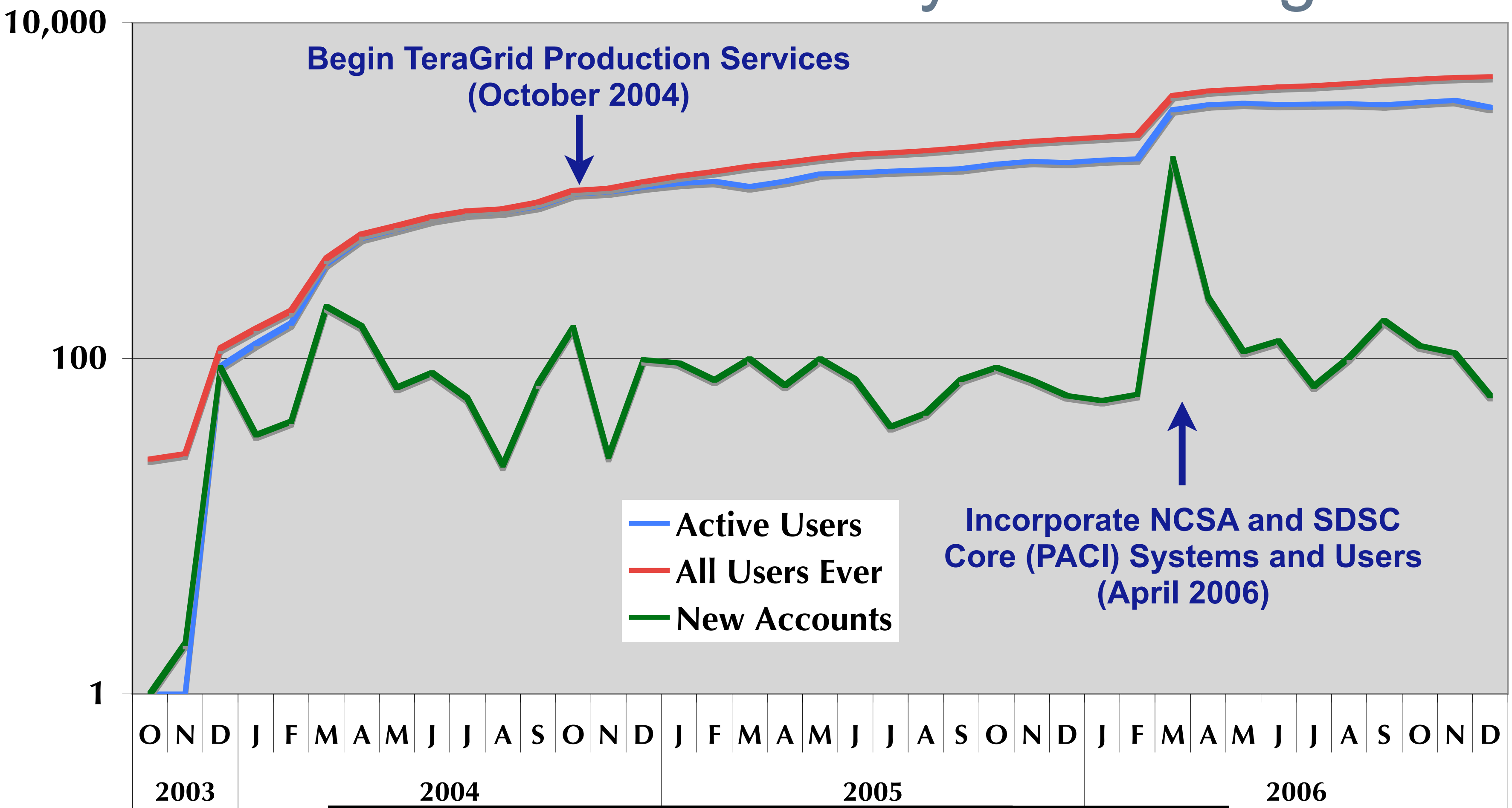
TeraGrid is supported by the National Science Foundation

Argonne National Laboratory is managed by
The University of Chicago for the U.S. Department of Energy

9 Resource Providers, One Facility


NATIONAL SCIENCE FOUNDATION
Resource Providers (n)
GIG
Software Integration Partners (n)
Grid Infrastructure Group (UChicago)
UW
UC/ANL
PU
PSC
NCAR
NCSA
IU
UNC/RENCI
ORNL
Caltech
USC/ISI
SDSC
TACC
● Resource Provider (RP)
◆ Software Integration Partner

HPC User Community is Growing



Begin TeraGrid Production Services (October 2004)

Incorporate NCSA and SDSC Core (PACI) Systems and Users (April 2006)

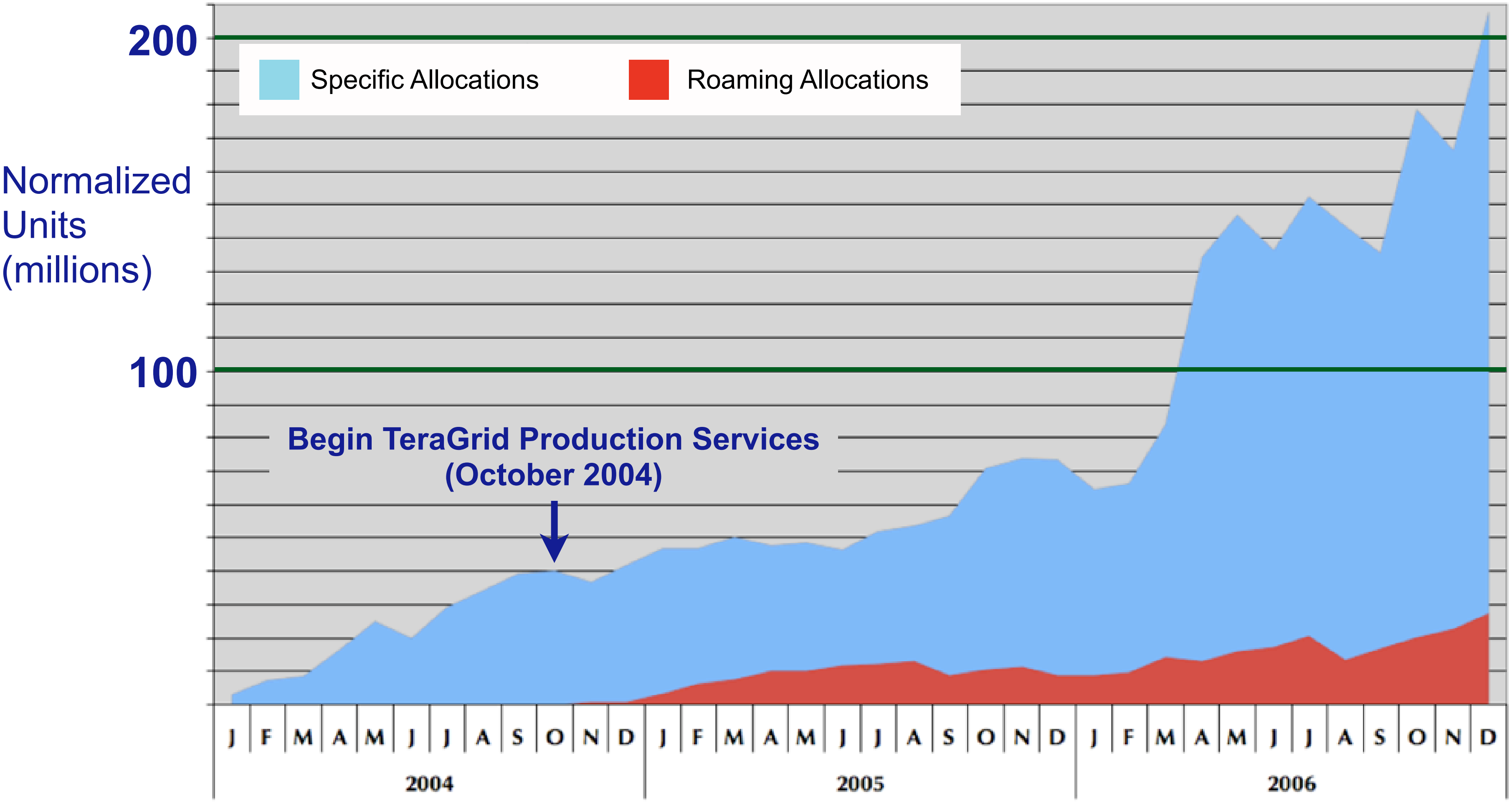
Decommissioning of systems typically causes slight reductions in active users. E.g. December 2006 is due to decommissioning of Lemieux (PSC).

	FY05	*FY06*
New User Accounts	948	2,692
Avg. New Users per Quarter	315	365*
Active Users	1,350	3,228
All Users Ever	**1,799**	**4,491**

(*FY06 new users/qtr excludes Mar/Apr 2006)

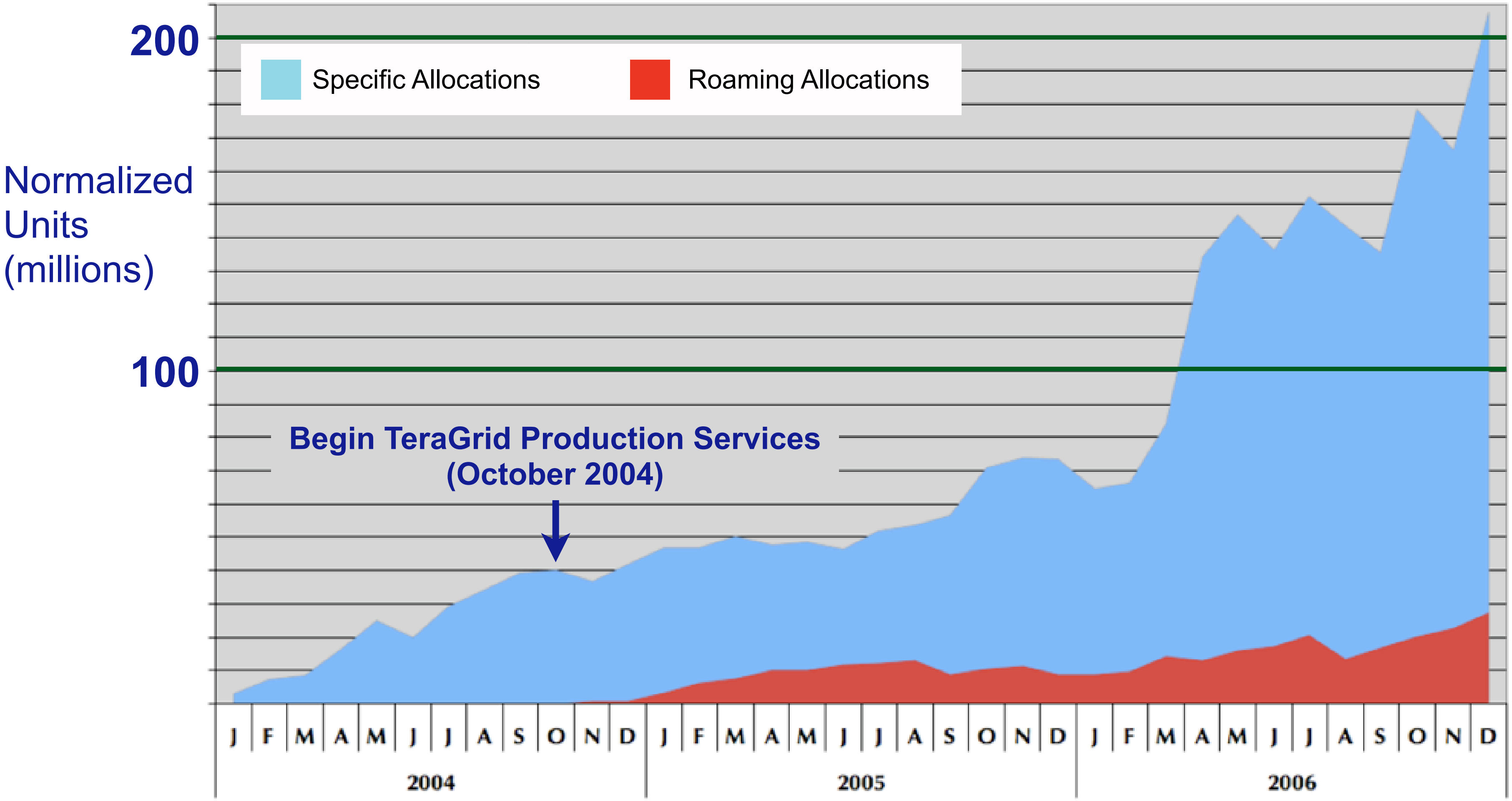
Does not include gateway users (expecting >10x)

Usage is also Growing....



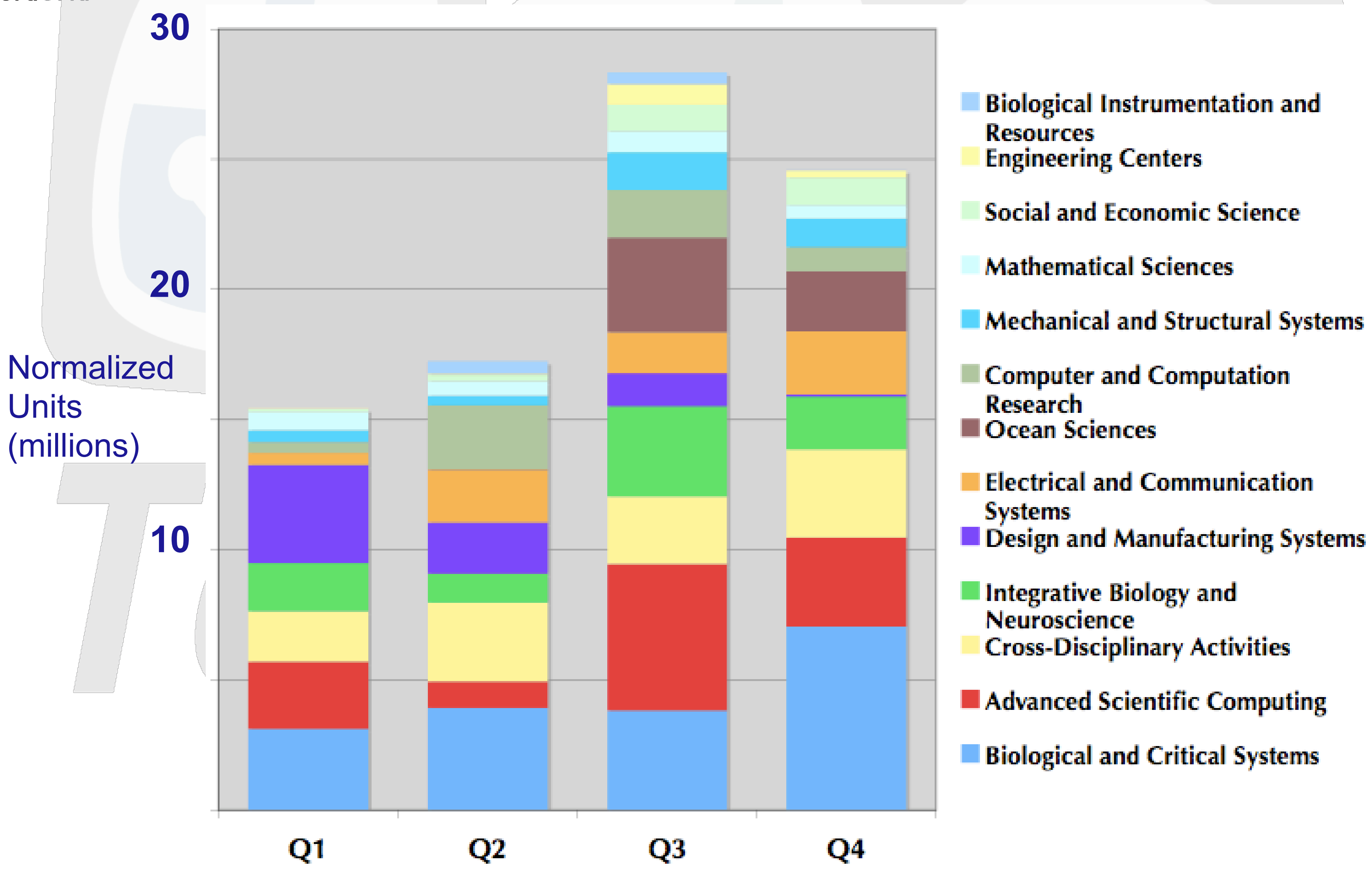
Begin TeraGrid Production Services (October 2004)

Usage is also Growing....

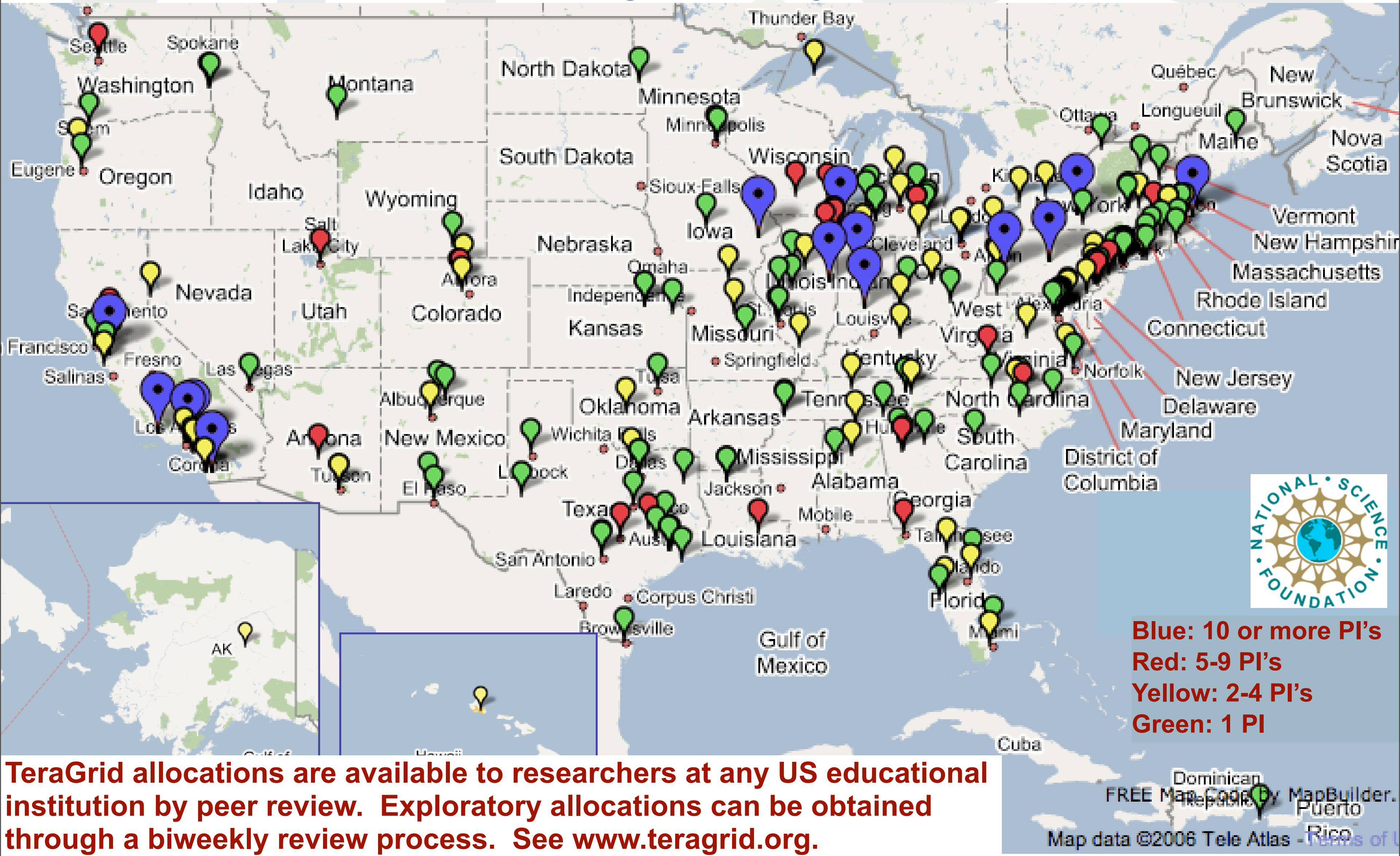


... and an increasing number of users prefer not to be tied to a specific machine.

FY06 Quarterly Usage by Discipline



TeraGrid Projects by Institution

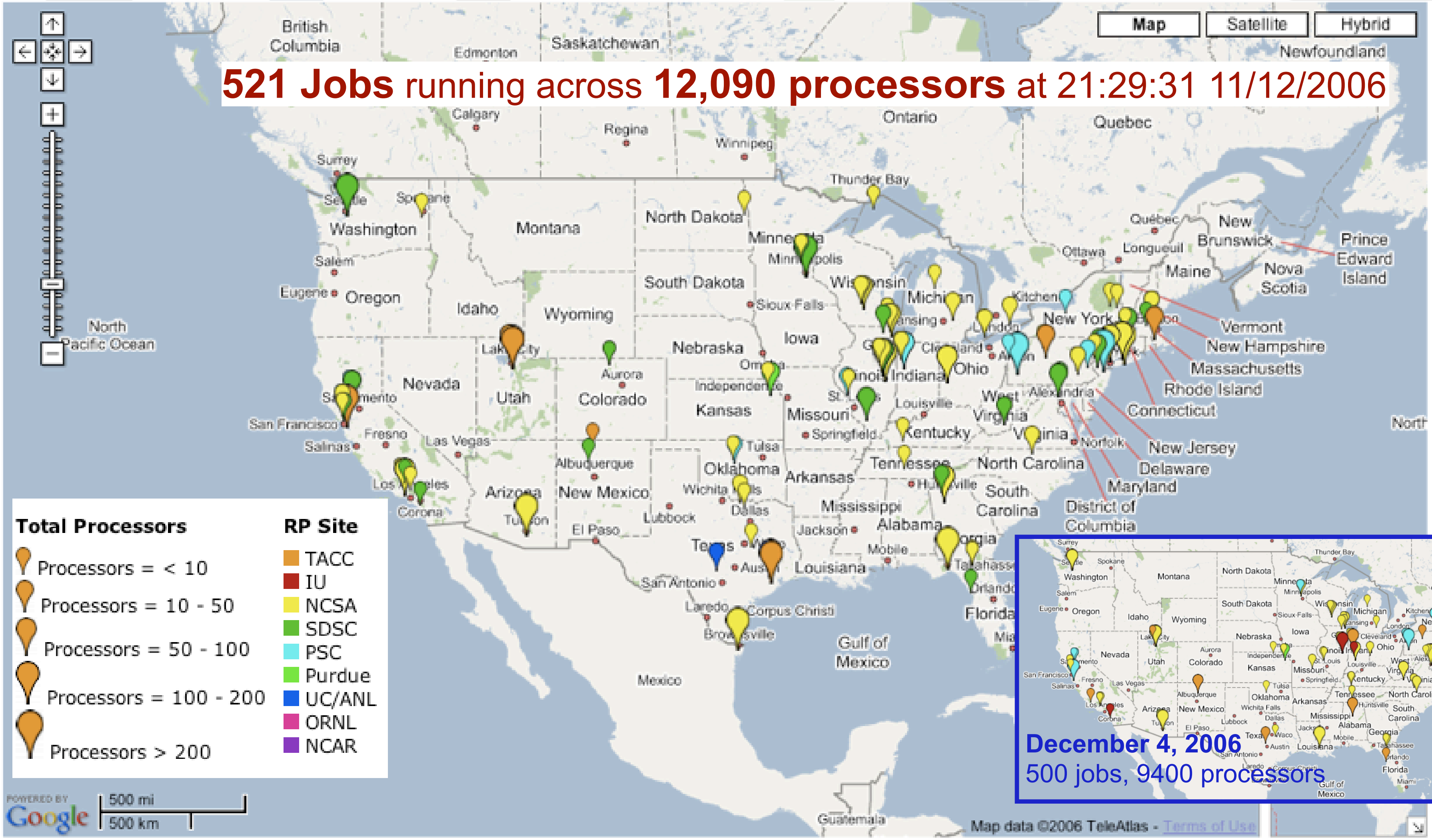


Blue: 10 or more PI's
Red: 5-9 PI's
Yellow: 2-4 PI's
Green: 1 PI

TeraGrid allocations are available to researchers at any US educational institution by peer review. Exploratory allocations can be obtained through a biweekly review process. See www.teragrid.org.

1000 projects (VOs), 4000 users

Real-Time Usage Mashup



521 Jobs running across **12,090 processors** at 21:29:31 11/12/2006

Total Processors

- Processors = < 10
- Processors = 10 - 50
- Processors = 50 - 100
- Processors = 100 - 200
- Processors > 200

RP Site

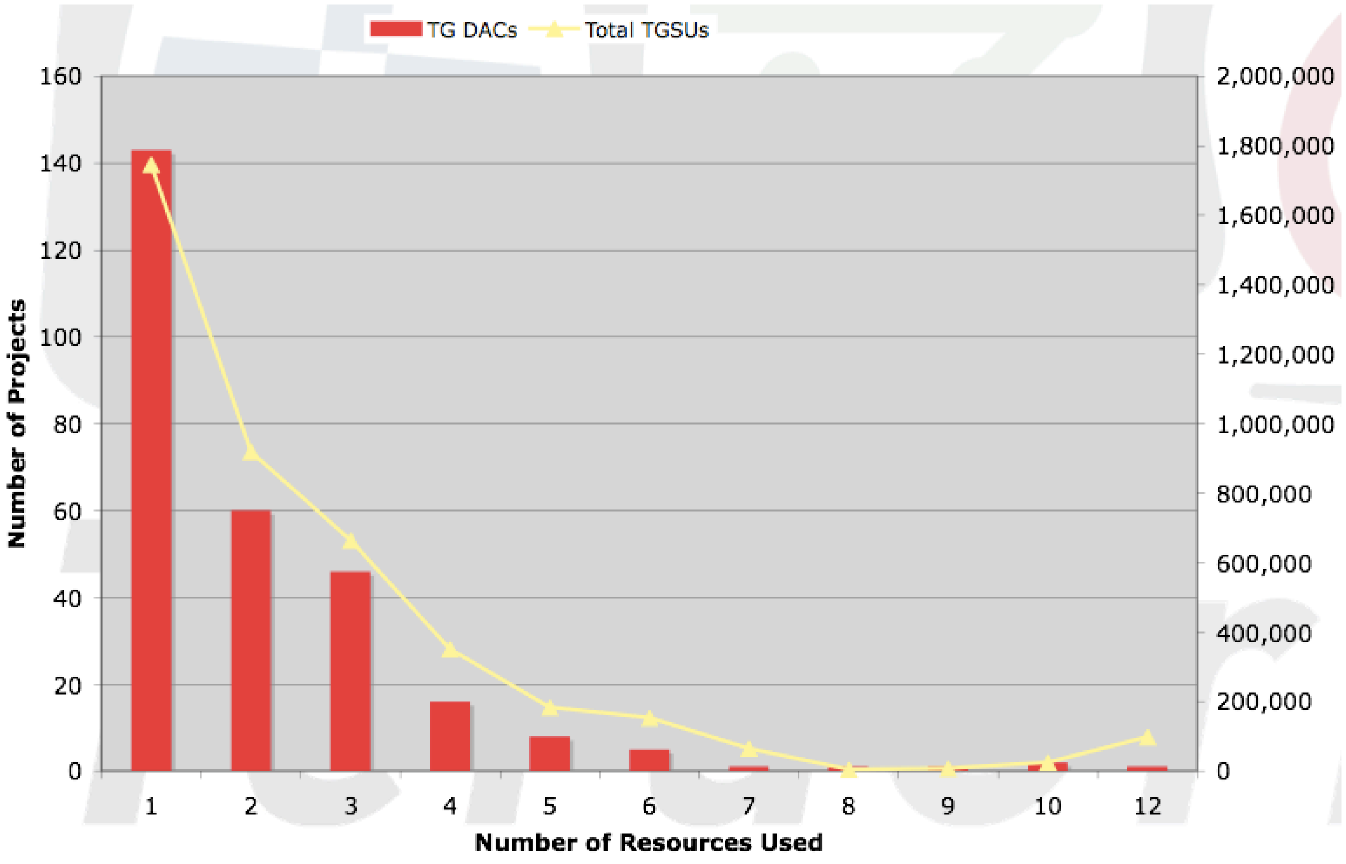
- 🟧 TACC
- 🟥 IU
- 🟨 NCSA
- 🟩 SDSC
- 🟦 PSC
- 🟩 Purdue
- 🟦 UC/ANL
- 🟪 ORNL
- 🟪 NCAR

December 4, 2006
500 jobs, 9400 processors

[Last Updated: Sun Nov 12 21:39:31 2006 CT]

The TeraGrid job map displays the current running jobs across TeraGrid. Each pin location denotes the location of the job owner, the color of the pin denotes the RP site of the job(s), the size of the pin denotes the total number of processors for the jobs. By clicking on the pin you can see the users job informormation - RP site, total number of jobs running, total number of processors - in addition to the user's location, department, and institution.

Is a coordinated user environment across many resources useful to new users?



Resources Used	Projects	Usage (SUs)
1	143	1,745,314
2	60	919,461
3	46	664,231
4	16	351,340
5	8	183,271
6	5	153,083
7	1	64,270
8	1	3,878
9	1	6,979
10	2	25,121
12	1	97,774
Total	284	4,214,722

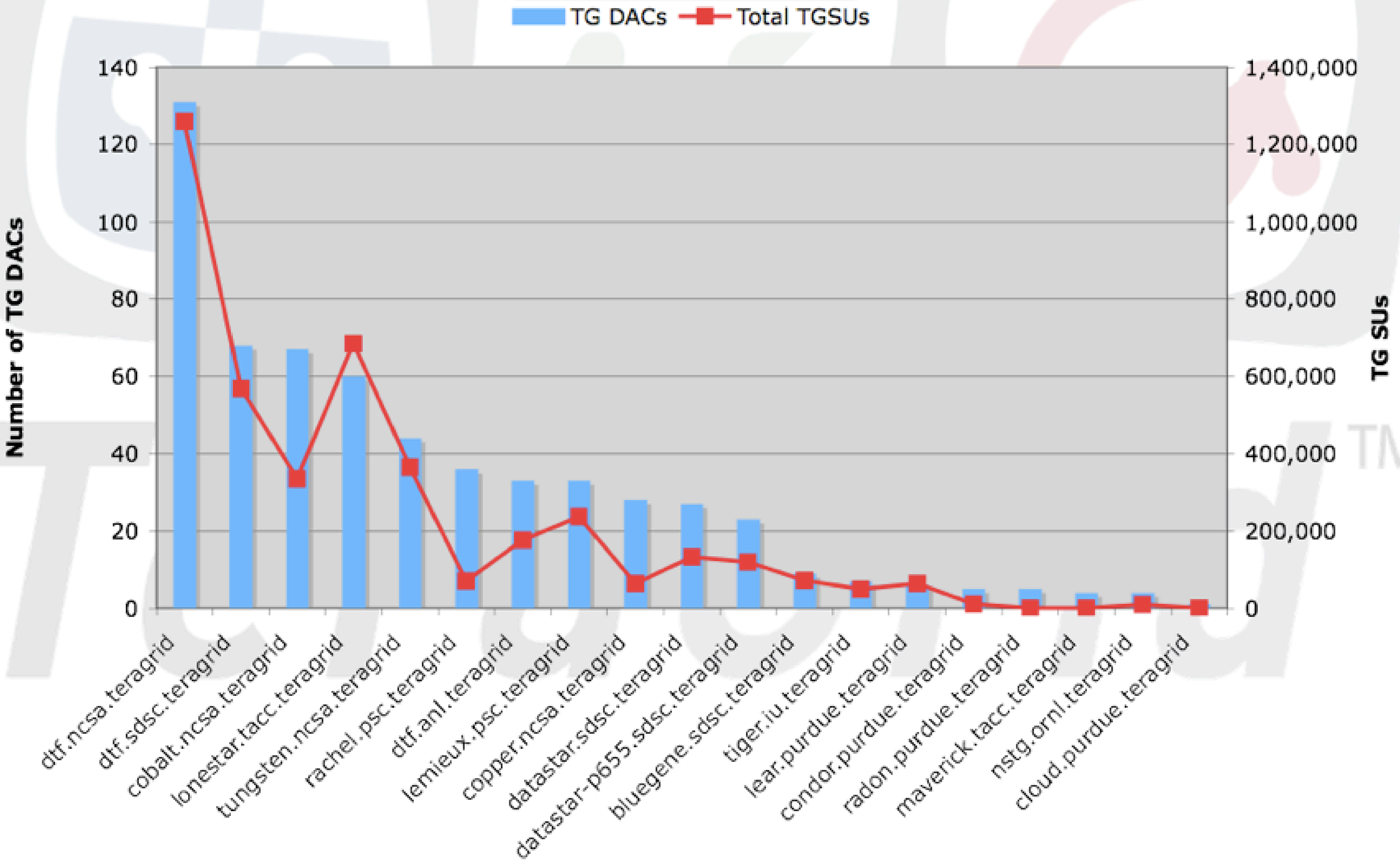
DAC - Development Allocations - new users with up to 30k hour allocations for exploring TeraGrid, porting codes, benchmaking. DAC allocations can be used on any TeraGrid resource. This chart shows the DAC awards sorted by the number of resources they have used (e.g. trying out various machines).

321 DACs used resources **EVER** !!

(only 37 before 2006)

Source: Dave Hart, SDSC

Do new users take advantage of freedom to choose from many resources?

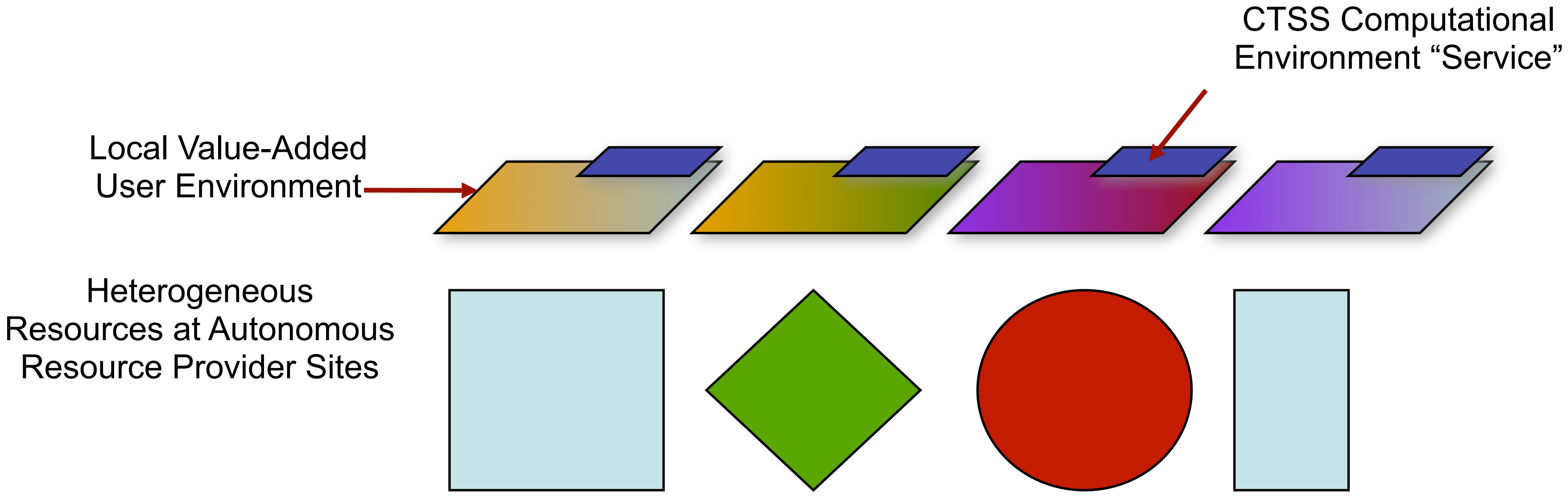


Source: Dave Hart, SDSC

There are many ways to provide an integrated, distributed facility...but supercomputers are heterogeneous and they are operated by autonomous (competing) organizations.

Heterogeneity and Autonomy must be leveraged - this requires agreement about services, central coordination, and ***local control***

Coordinated Facilities



CTSS Computational Environment “Service”

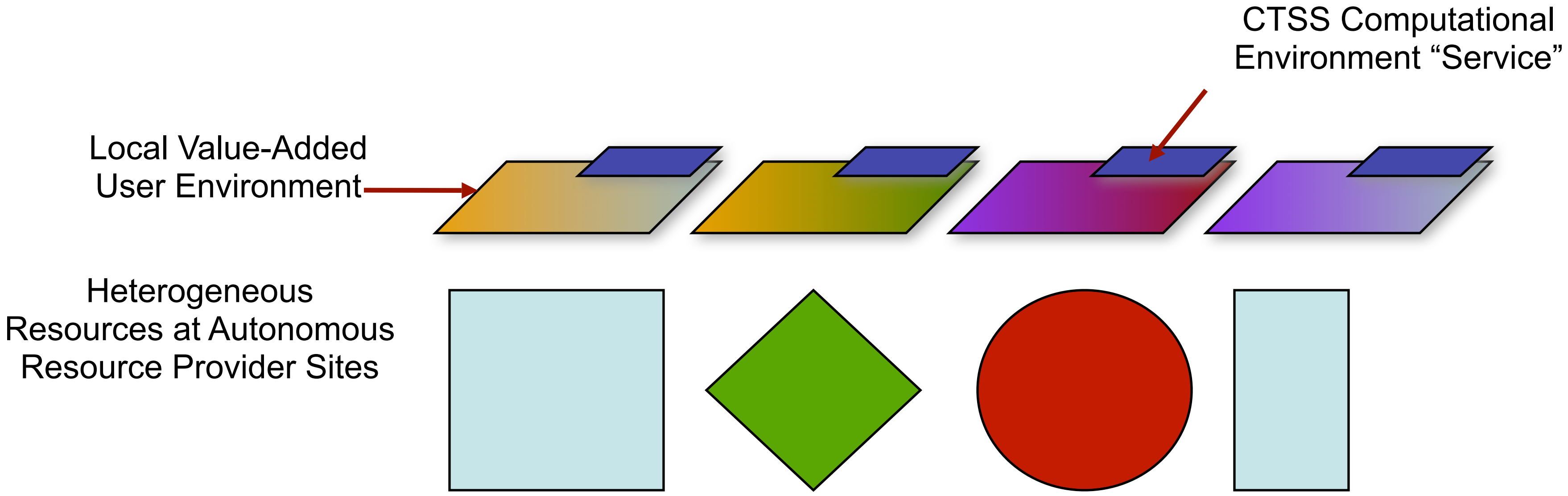
Local Value-Added User Environment

Heterogeneous Resources at Autonomous Resource Provider Sites

- **A single point of contact** for user assistance.
- **A common allocation and accounting infrastructure** that includes a currency usable on all systems, while preserving the need to provide specific machine access to users with specific needs.
- **A common access service and environment** on all platforms, allowing users to readily move from machine to machine - to “roam” - as needed. *Learn Once; Run Anywhere*.
- **Services to assist users in harnessing the right TeraGrid platforms for each part of their work**, ranging from tightly-coupled applications (MPICH-G2) to workflow and parameter sweep (Condor, MyCluster), file staging (GridFTP/RFT) and remote file I/O (GPFS), supported by common authentication (GSI), and Web services via GT4.

▱ CTSS v1 (30+ pkgs)

Coordinated Facilities

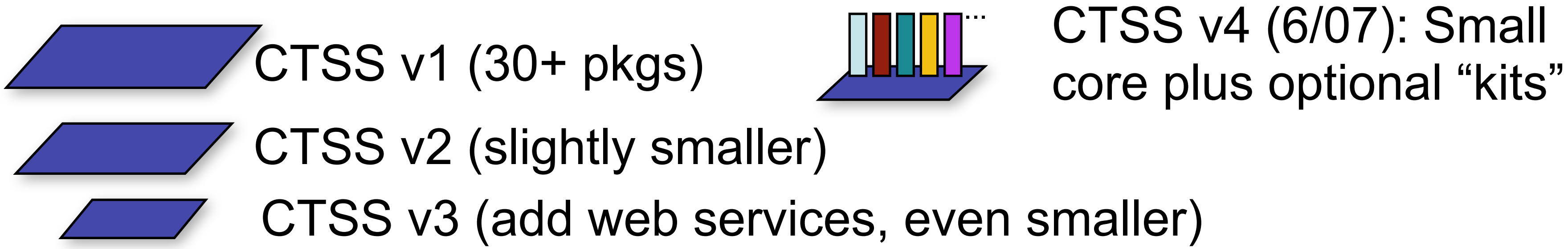


CTSS Computational Environment “Service”

Local Value-Added User Environment

Heterogeneous Resources at Autonomous Resource Provider Sites

- **A single point of contact** for user assistance.
- **A common allocation and accounting infrastructure** that includes a currency usable on all systems, while preserving the need to provide specific machine access to users with specific needs.
- **A common access service and environment** on all platforms, allowing users to readily move from machine to machine - to “roam” - as needed. *Learn Once; Run Anywhere*.
- **Services to assist users in harnessing the right TeraGrid platforms for each part of their work**, ranging from tightly-coupled applications (MPICH-G2) to workflow and parameter sweep (Condor, MyCluster), file staging (GridFTP/RFT) and remote file I/O (GPFS), supported by common authentication (GSI), and Web services via GT4.



CTSS v1 (30+ pkgs)

CTSS v2 (slightly smaller)

CTSS v3 (add web services, even smaller)

CTSS v4 (6/07): Small core plus optional “kits”

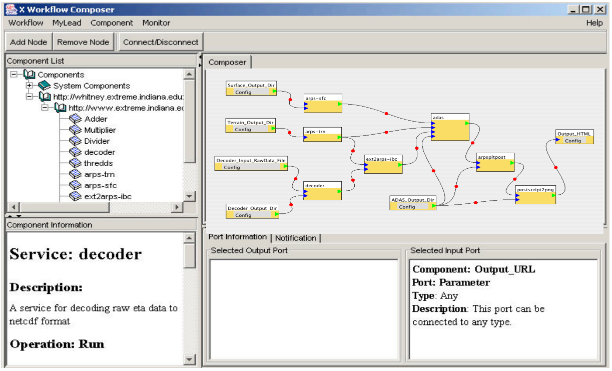
CTSSv4 Core Integration Capability Kit

- The only mandatory CTSS kit
  - Provides the capabilities that are absolutely necessary for a resource to meet the most basic integrative requirements of the TeraGrid.
- *Significantly* smaller than the set of “required” CTSSv3 components.
  - **Security** – Identity, Authentication, Authorization, Auditing
  - **Information** – Capability and Service Registry, System & Service Description, Usage Monitoring & Profiling
  - **Verification & Validation** – System Status and Testing
  - **Software Deployment** – Deployment Tools, Build & Test Capability
- The other CTSS 4 kits will be deployed on the resources where they are appropriate. Some will be widespread, others more specialized.
  - Initial optional kits include:
    - Remote Login
    - Remote Compute
    - Data Movement
    - Data Management
    - Science Workflow Support
    - Parallel Application Support
    - Application Runtime & Development Suite

Source: Lee Liming, JP Navarro, ANL

A service-oriented approach enables entire communities to share software and infrastructure, creating a facility that enables users to innovate w.r.t. grid capabilities and that can be used to educate and grow the scientific workforce.

Science Gateways: Service-Oriented Approach



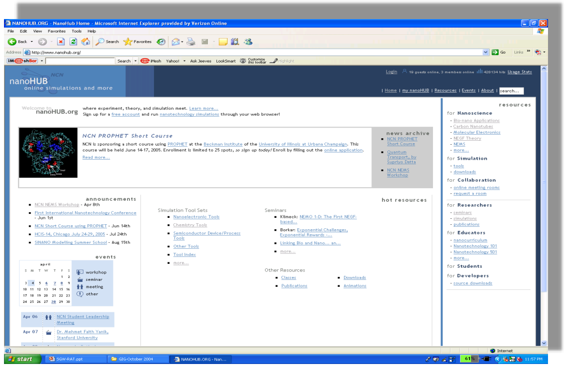
Service: decoder

Description:

A service for decoding raw eta data to netcdf format

Operation: Run

Component: Output_URL
Port: Parameter
Type: Any
Description: This port can be connected to any type.



19 guests online, 3 members online 428134 hits Usage Stats

Welcome to nanoHUB.org where experiment, theory, and simulation meet. Learn more...
Sign up for a free account and run nanotechnology simulations through your web browser!

NCN PROPHET Short Course

NCN is sponsoring a short course using PROPHET at the Beckman Institute of the University of Illinois at Urbana Champaign. This course will be held June 14-17, 2005. Enrollment is limited to 25 spots, *so sign up today!* Enroll by filling out the online application.
Read more...

news archive
- NCN PROPHET Short Course
- Quantum Transport, by Supriyo Datta
- NCN NEMS Workshop

announcements
- NCN NEMS Workshop - Apr 8th
- First International Nanotechnology Conference - Jun 1st
- NCN Short Course using PROPHET - Jun 14th
- HCIS-14, Chicago July 24-29, 2005 - Jul 24th
- SiNANO Modelling Summer School - Aug 15th

events

april

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

workshop
seminar
meeting
other

Apr 06 NCN Student Leadership Meeting
Apr 07 Dr. Mehmet Fatih Yanik, Stanford University

hot resources

Simulation Tool Sets
- Nanoelectronic Tools
- Chemistry Tools
- Semiconductor Device/Process Tools
- Other Tools
- Tool Index
- more...

Seminars
- Klimeck: NEMO 1-D: The First NEGF-based...
- Borkar: Exponential Challenges, Exponential Rewards -...
- Linking Bio and Nano... an...
- more...

Other Resources
- Classes
- Downloads
- Publications
- Animations

resources

for **Nanoscience**
- Bio-nano Applications
- Carbon Nanotubes
- Molecular Electronics
- NEGF Theory
- NEMS
- more...

for **Simulation**
- tools
- downloads

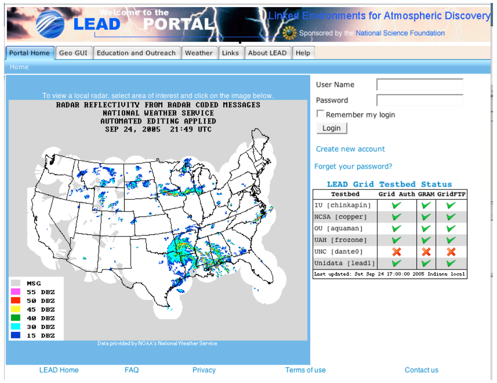
for **Collaboration**
- online meeting rooms
- request a room

for **Researchers**
- seminars
- simulations
- publications

for **Educators**
- nanocurriculum
- Nanotechnology 101
- Nanotechnology 501
- more...

for **Students**

for **Developers**
- source downloads



To view a local radar, select area of interest and click on the image below.

RADAR REFLECTIVITY FROM RADAR CODED MESSAGES
NATIONAL WEATHER SERVICE
AUTOMATED EDITING APPLIED
SEP 24, 2005 21:49 UTC

Data provided by NOAA's National Weather Service

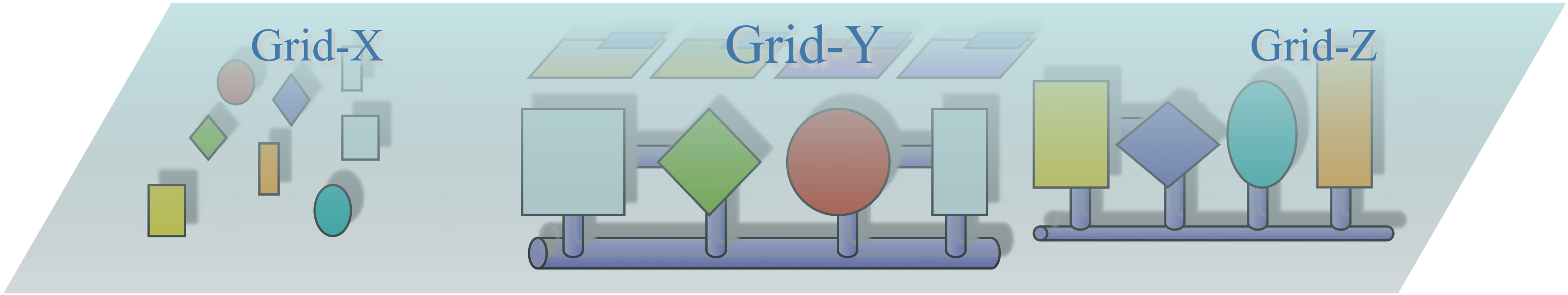
User Name
Password
□ Remember my login

LEAD Grid Testbed Status

Testbed	Grid Auth	GRAM	GridFTP
IU [chinkapin]	✓	✓	✓
NCSA [copper]	✓	✓	✓
OU [aquaman]	✓	✓	✓
UAH [frozone]	✓	✓	✓
UNC [dante0]	✗	✗	✗
Unidata [lead1]	✓	✓	✓

Last updated: Sat Sep 24 17:00:00 2005 Indiana local

Web Services

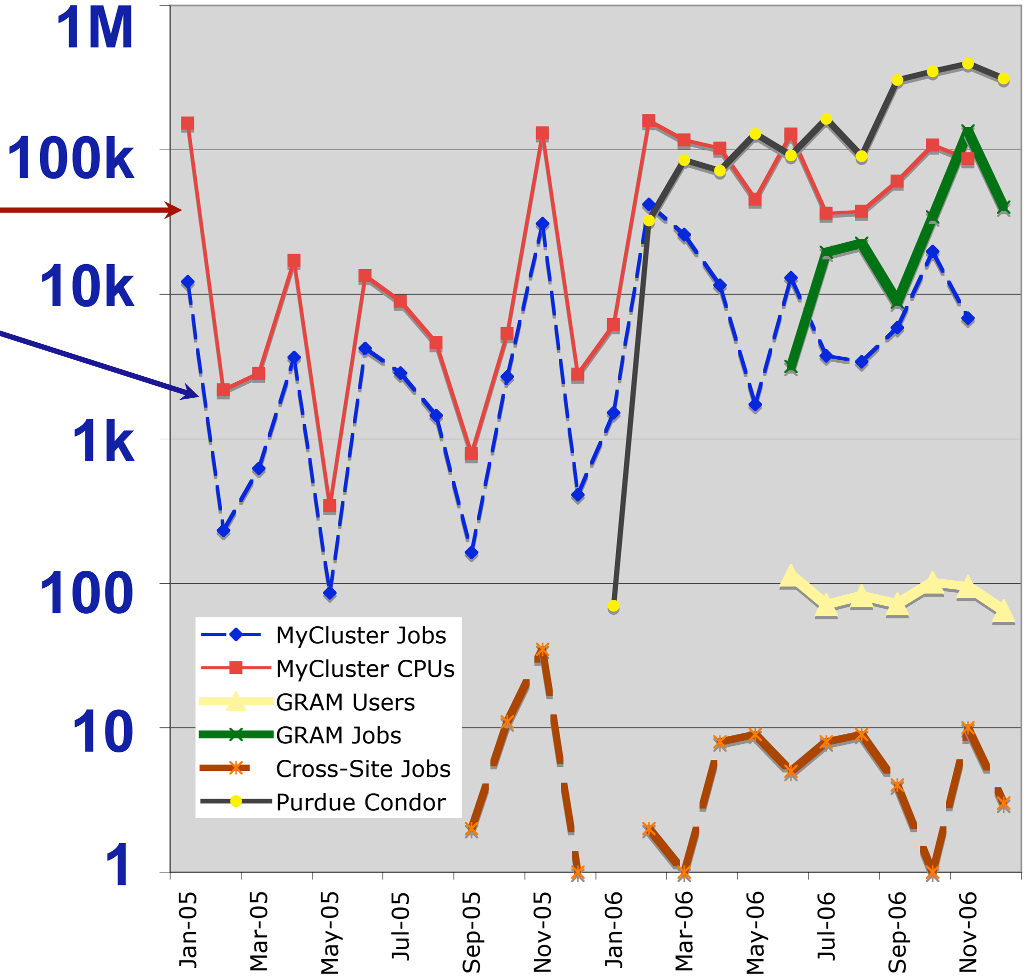


We’ve built a distributed facility with exponential user growth and usage growth.

Are people really using this grid stuff?

YES

Monthly Use of Selected Grid Capabilities



January 2005 through December 2006

Parameter Search w/ MyCluster:

MyCluster CPUs

MyCluster Jobs

Purdue Condor jobs

Remote Job Submission (e.g. workflow) w/ Globus:

Globus GRAM Jobs

Globus GRAM Users

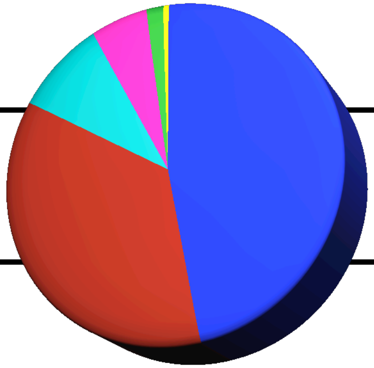
Synchronous multi-site jobs

TeraGrid User Community in 2006

Use Modality	Community Size (est. number of projects)
Batch Computing on Individual Resources	850
Exploratory and Application Porting	650
Workflow, Ensemble, and Parameter Sweep	160
Science Gateway Access	100
Remote Interactive Steering and Visualization	35
Tightly-Coupled Distributed Computation	10

TeraGrid User Community in 2006

Use Modality	Community Size (est. number of projects)
Batch Computing on Individual Resources	850
Exploratory and Application Porting	650
Workflow, Ensemble, and Parameter Sweep	160
Science Gateway Access	100
Remote Interactive Steering and Visualization	35
Tightly-Coupled Distributed Computation	10



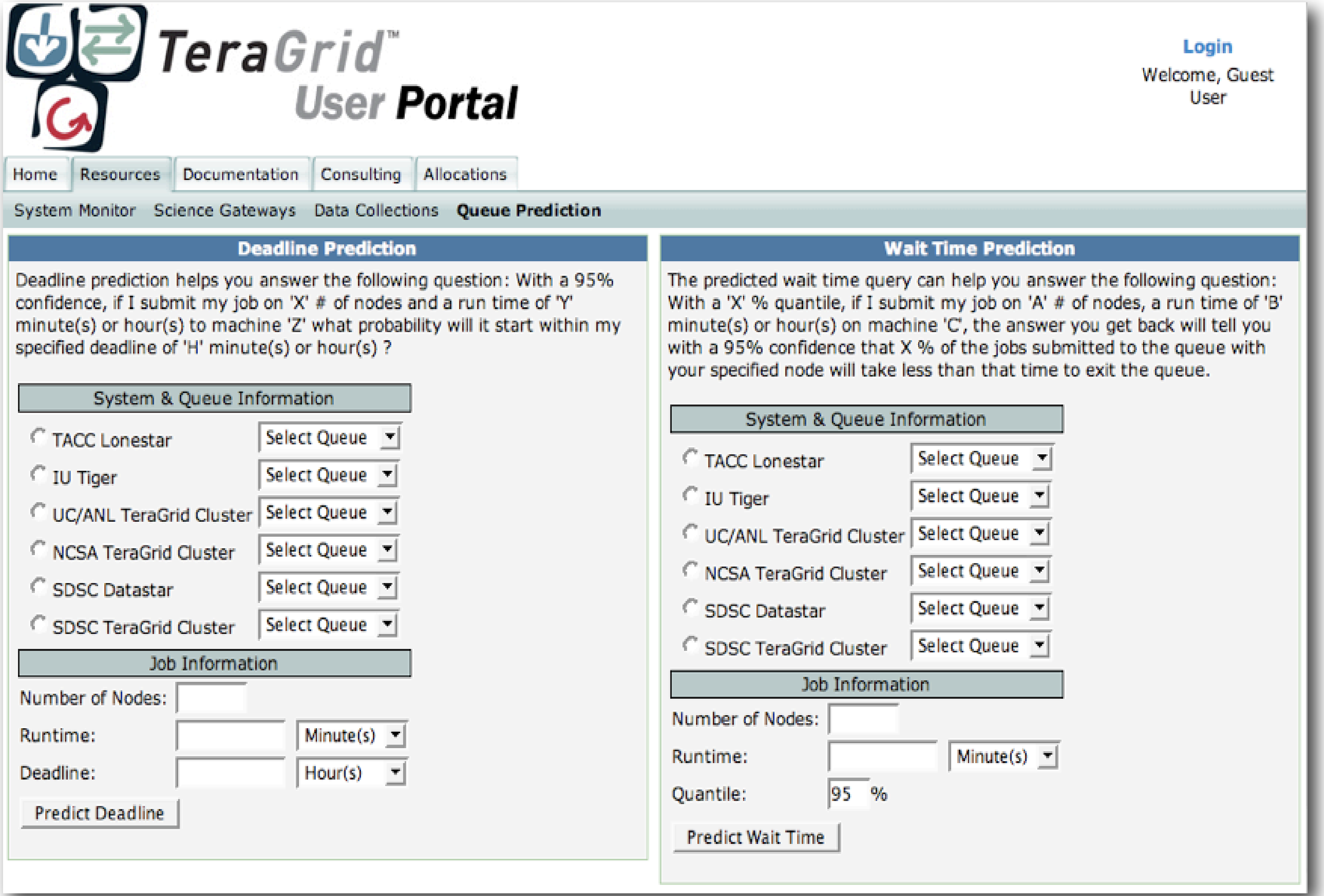
TeraGrid User Community in 2006

Use Modality	Community Size (est. number of projects)
Batch Computing on Individual Resources	850
Exploratory and Application Porting	650
Workflow, Ensemble, and Parameter Sweep	**250**
Science Gateway Access	**500**
Remote Interactive Steering and Visualization	35
Tightly-Coupled Distributed Computation	10

With a service oriented infrastructure, users and community infrastructure providers can begin to build advanced capabilities (rather than waiting for us to do it).

Consider what hooks, knobs, and outlets to provide - let a broader community build the bells and whistles they need.

Co-Scheduling? Advanced Reservation?



TeraGrid™ User Portal

Login

Welcome, Guest User

Home | Resources | Documentation | Consulting | Allocations

System Monitor | Science Gateways | Data Collections | **Queue Prediction**

Deadline Prediction

Deadline prediction helps you answer the following question: With a 95% confidence, if I submit my job on 'X' # of nodes and a run time of 'Y' minute(s) or hour(s) to machine 'Z' what probability will it start within my specified deadline of 'H' minute(s) or hour(s) ?

System & Queue Information

○ TACC Lonestar — Select Queue
○ IU Tiger — Select Queue
○ UC/ANL TeraGrid Cluster — Select Queue
○ NCSA TeraGrid Cluster — Select Queue
○ SDSC Datastar — Select Queue
○ SDSC TeraGrid Cluster — Select Queue

Job Information

Number of Nodes:

Runtime: Minute(s)

Deadline: Hour(s)

Predict Deadline

Wait Time Prediction

The predicted wait time query can help you answer the following question: With a 'X' % quantile, if I submit my job on 'A' # of nodes, a run time of 'B' minute(s) or hour(s) on machine 'C', the answer you get back will tell you with a 95% confidence that X % of the jobs submitted to the queue with your specified node will take less than that time to exit the queue.

System & Queue Information

○ TACC Lonestar — Select Queue
○ IU Tiger — Select Queue
○ UC/ANL TeraGrid Cluster — Select Queue
○ NCSA TeraGrid Cluster — Select Queue
○ SDSC Datastar — Select Queue
○ SDSC TeraGrid Cluster — Select Queue

Job Information

Number of Nodes:

Runtime: Minute(s)

Quantile: 95 %

Predict Wait Time

Network Weather Service “Batch Queue Prediction” (BQP) - **Rich Wolski, UCSB** (rich@cs.ucsb.edu)

Current Campus Partnership Areas

- Integrated Authorization & Authentication
  - Improve CI usability for scientists and engineers on campuses, simultaneously increasing the security of CI
    - S. Goasguen (Clemson), J. Kyriannis (NYU), C. McMahon (LSU)
    - Testbeds at Purdue, University of Chicago using Shibboleth
- Federated HPC and Data Management
  - Develop and deploy frameworks to support access to the increasingly powerful campus and national HPC investments, providing both capability and capacity services, and a storage and data management infrastructure to support open, extensible, evolvable science and engineering data collections
    - J. Boisseau (TACC), V. Agarwala (PSU), S. Corbato (Utah/ Internet2)
    - Partnership with Open Science Grid, University of Wisconsin