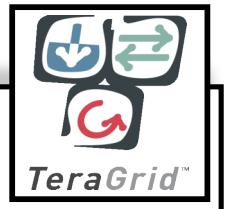
The National Grid Cyberinfrastructure
Open Science Grid and TeraGrid

Introduction

- What we've already learned
  - What are grids, why we want them and who is using them: Intro
  - Grid Authentication and Authorization
  - Harnessing CPU cycles with condor
  - Data Management and the Grid
- In this lecture
  - Fabric level infrastructure: Grid building blocks
  - National Grid efforts in the US
    - TeraGrid
    - The Open Science Grid

Grid Resources in the US

The TeraGrid



Origins:

- National Super Computing Centers, funded by the National Science Foundation

Current Compute Resources:

- 9 TeraGrid sites
- Connected via dedicated multi-Gbps links
- Mix of Architectures
  - ia64, ia32: LINUX
  - Cray XT3
  - Alpha: True 64
  - SGI SMPs
- Resources are dedicated but
  - Grid users share with local and grid users
  - 1000s of CPUs, > 40 TeraFlops
- 100s of TeraBytes

The OSG

Origins:

- National Grid (iVDGL, GriPhyN, PPDG) and LHC Software & Computing Projects

Current Compute Resources:

- 61 Open Science Grid sites
- Connected via Inet2, NLR.... from 10 Gbps – 622 Mbps
- Compute & Storage Elemets
- All are Linux clusters
- Most are shared
  - Campus grids
  - Local non-grid users
- More than 10,000 CPUs
  - A lot of opportunistic usage
  - Total computing capacity difficult to estimate
  - Same with Storage

Grid Building Blocks

- Computational Clusters
- Storage Devices
- Networks
- Grid Resources and Layout:
  - User Interfaces
  - Computing Elements
  - Storage Elements
  - Monitoring Infrastructure…

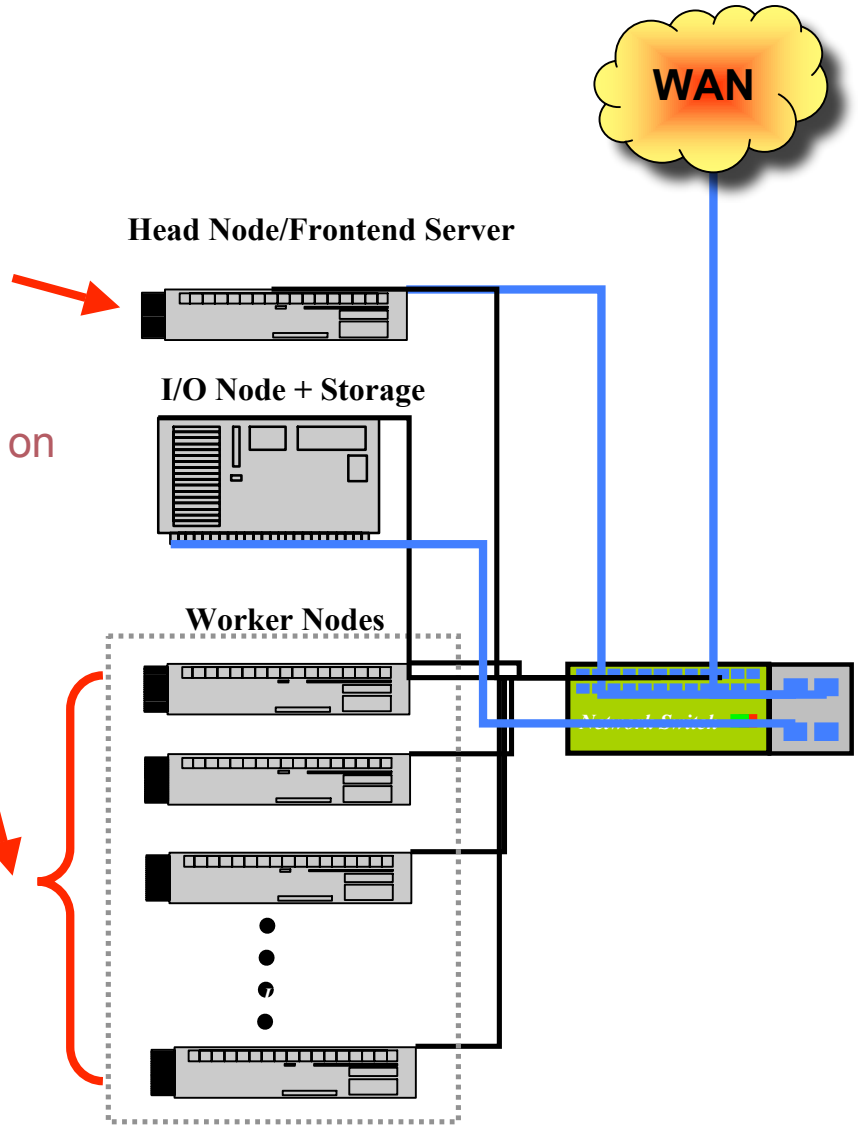
Computation on a Clusters

- Batch scheduling systems
  - Submit many jobs through a head node

```
#!/bin/sh
for each i in $list_o_jobscripts
do
  /usr/local/bin/condor_submit $i
done
```

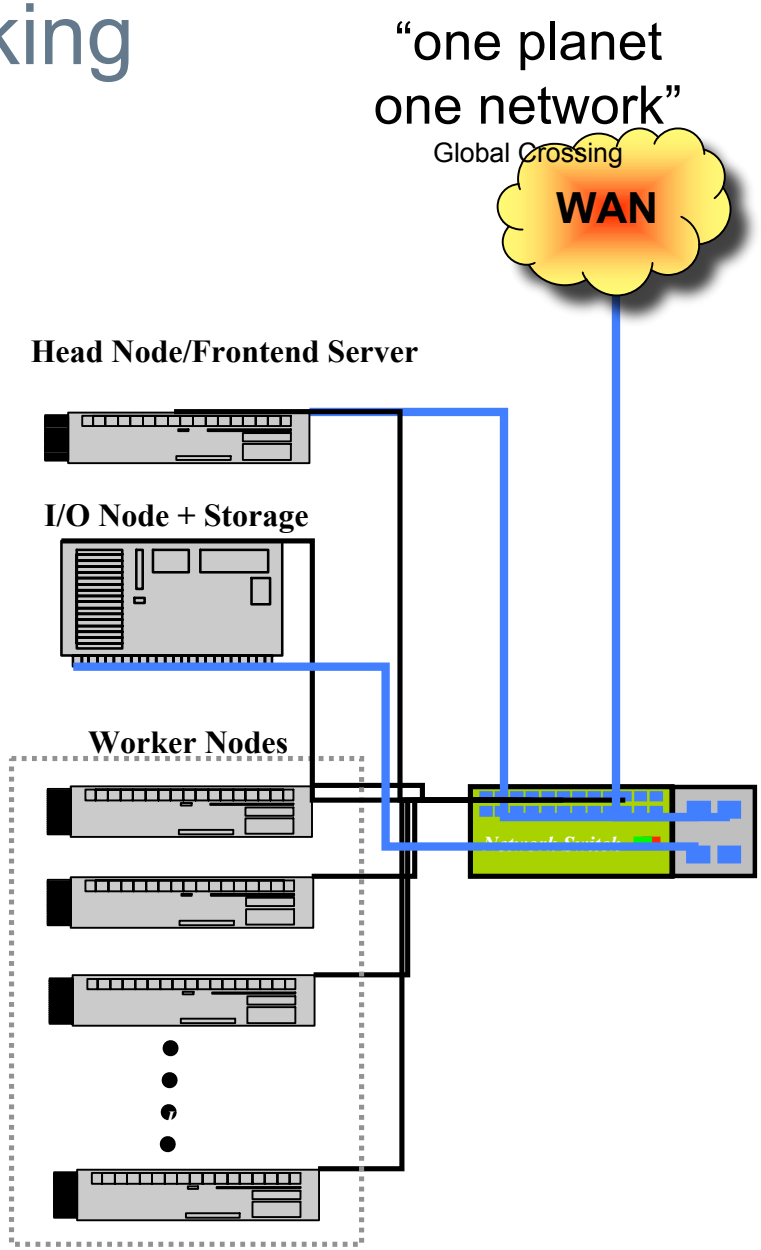
  - Execution done on worker nodes
- Many different batch systems are deployed on the grid
  - condor (highlighted in lecture 5)
  - pbs, lsf, sge…

Primary means of controlling CPU usage, enforcing allocation policies and scheduling of jobs on the local computing infrastructure

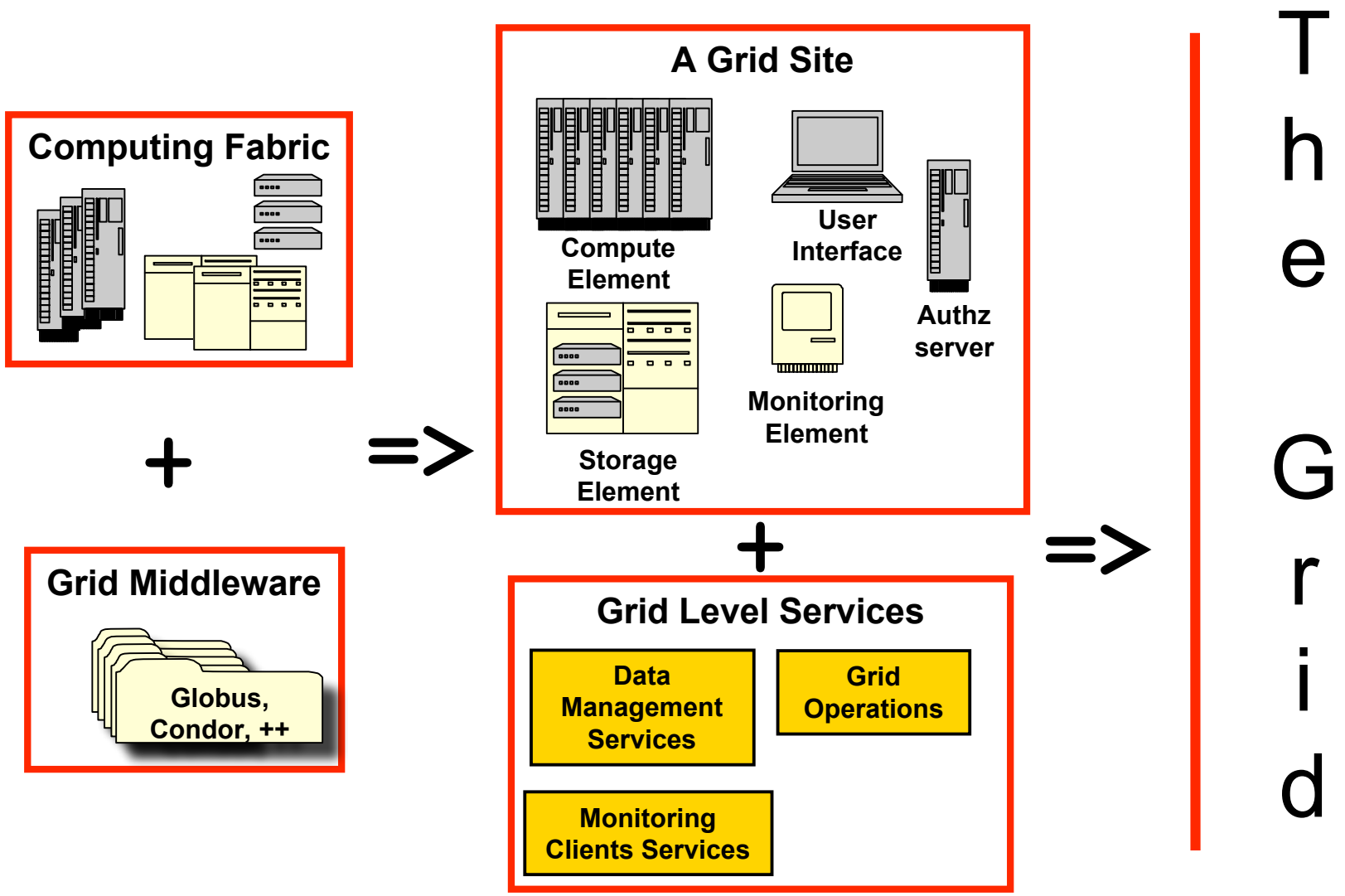


Networking

- Internal Networks (LAN)
  - Private, accessible only to servers inside a facility
  - Some sites allow outbound connectivity via **N**etwork **A**ddress **T**ranslation
  - Typical technologies used
    - Ethernet (0.1, 1 & 10 Gbps)
    - HP, Low Latency interconnects
      - Myrinet: 2, 10 Gbps
      - Infiniband: max at 120Gbps
- External connectivity
  - Connection to Wide Area Network
  - Typically achieved via same switching fabric as internal interconnects

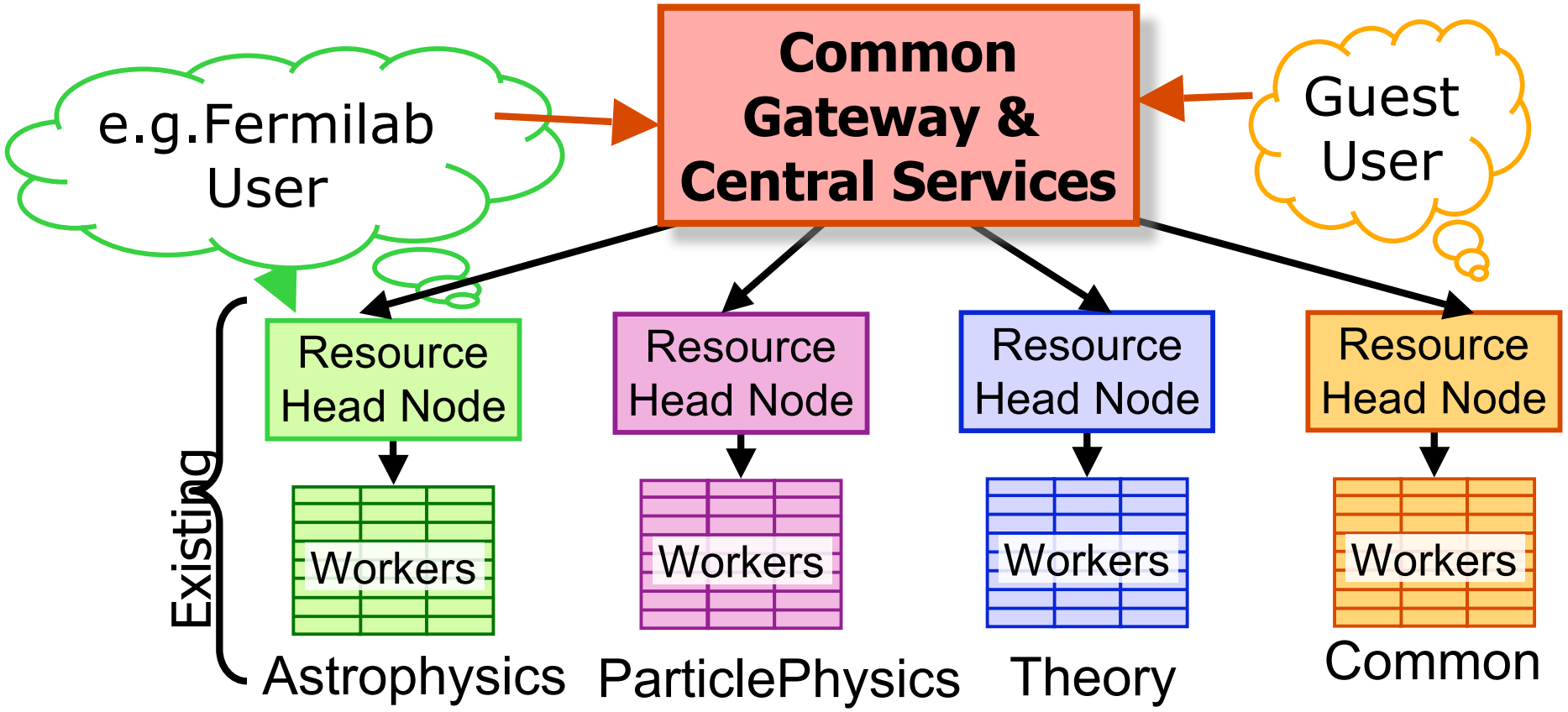


Layout of Typical Grid Site



Local Grid with adaptor to national grid

- Central Campus wide Grid Services
- Enable efficiencies and sharing across internal farms and storage
- Maintain autonomy of individual resources



Next Step: Campus Infrastructure Days - new activity OSG, Internet2 and TeraGrid

Grid Monitoring & Information Services

To efficiently use a Grid, you must locate and monitor its resources.

- Check the availability of different grid sites
- Discover different grid services
- Check the status of “jobs”
- Make better scheduling decisions with information maintained on the “health” of sites

Monitoring provides information for several purposes

- Operation of Grid
  - Monitoring and testing Grid
- Deployment of applications
  - What resources are available to me? (Resource discovery)
  - What is the state of the grid? (Resource selection)
  - How to optimize resource use? (Application configuration and adaptation)
- Information for other Grid Services to use

Monitoring information is either static or dynamic, broadly.

- Static information about a site:
  - Number of worker nodes, processors
  - Storage capacities
  - Architecture and Operating systems
- Dynamic information about a site
  - Number of jobs currently running
  - CPU utilization of each worker node
  - Overall site “availability”
- Time-varying information is critical for scheduling of grid jobs
- More accurate info costs more: it’s a tradeoff.

Open Science Grid Overview

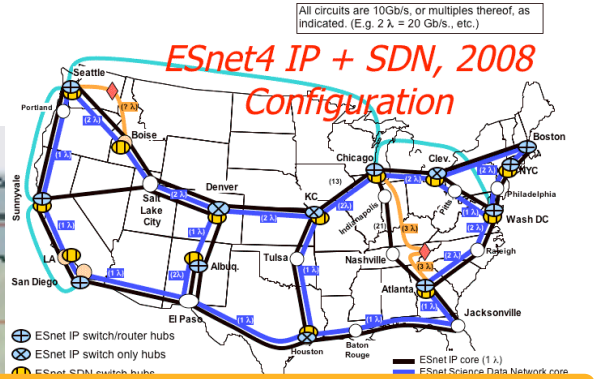
The OSG is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

The Open Science Grid Consortium brings:

- **the grid service providers** - middleware developers, cluster, network and storage administrators, local-grid communities
- **the grid consumers** - from global collaborations to the single researcher, through campus communities to under-served science domains
- into a **cooperative to share and sustain** a common heterogeneous **distributed facility** in the US and beyond.
- Grid providers serve multiple communities, Grid consumers use multiple grids.

OSG Snapshot

96 Resources across
production & integration infrastructures



Using production & research networks

Sustaining through OSG submissions:
3,000-4,000 simultaneous jobs .
~10K jobs/day
~50K CPUhours/day.
Peak test jobs of 15K a day.

~20,000 CPUs (from 30 to 4000)
~6 PB Tapes
~4 PB Shared Disk

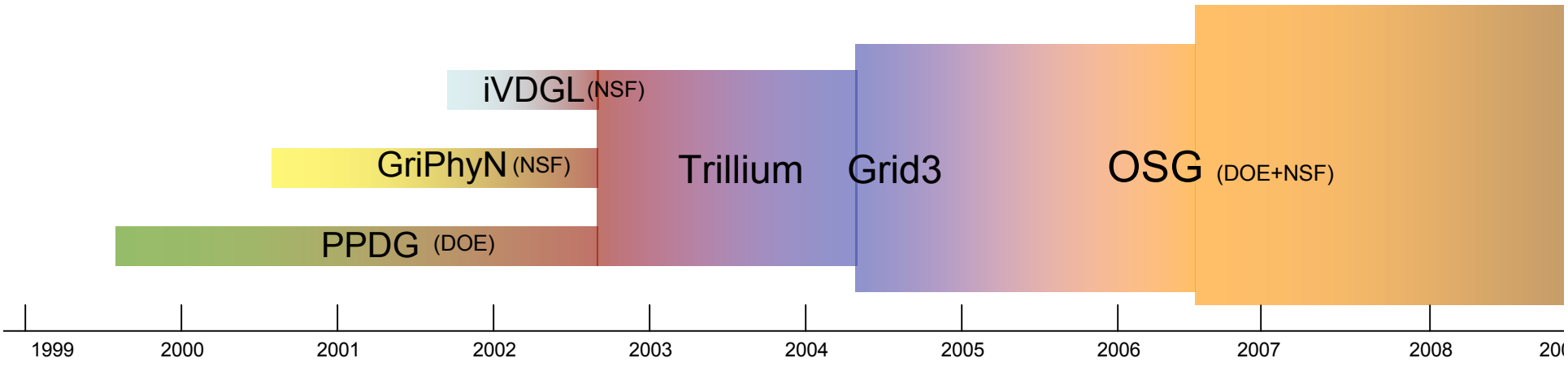
Snapshot of Jobs on OSGs

Farm	Last value	Min	Avg	Max
ATLAS	259	0	338.8	1516
CDF	1278	0	336.6	2086
CMS	579	0	439	3733
DES	39	0	0.385	40
DOSAR	25	0	9.93	192
FERMILAB	4	0	21.38	192
GADU	0	0	23.84	730
GLOW	0	0	35.44	541
GRIDEX	29	0	20.36	268
GROW	41	0	1.434	111
IVDGL	0	0	0.852	73
KTEV	35	0	15.52	260
LIGO	13	0	2.539	88
MINIBOONE	1053	0	128.7	1254
MIPP	2	0	15.32	206
MIS	0	0	0.269	20
NANOHUB	99	0	26.83	187
OPS	2	0	0.017	3
OSG	0	0	0.226	11
SDSS	33	0	3.941	199
STAR	38	0	12.77	150
Total	**3529**		**1434**	

20 Virtual Organizations +6 operations
Includes 25% non-physics.

OSG - a Community Consortium

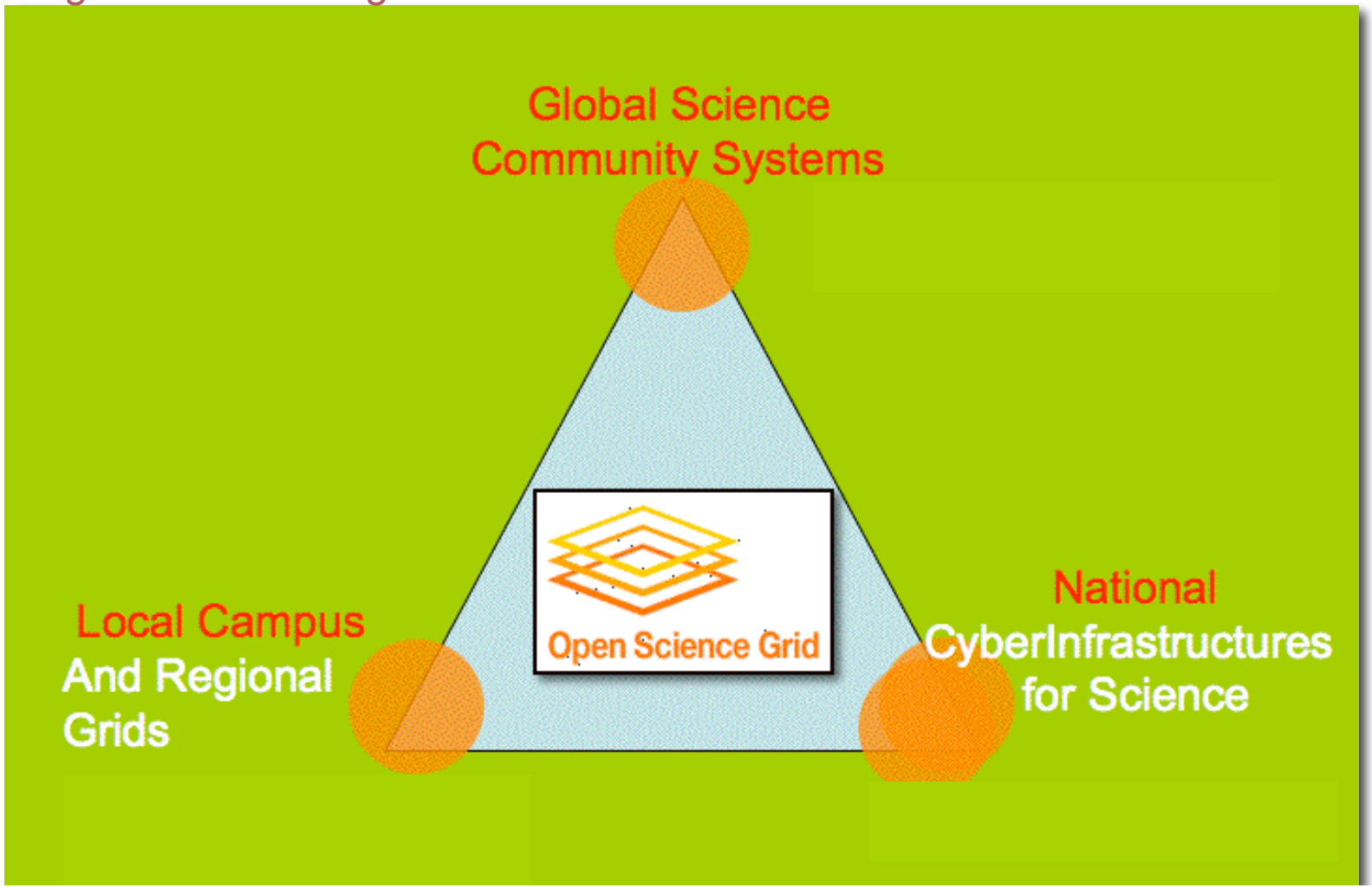
- **DOE Laboratories and DOE, NSF, other, University Facilities** contributing computing farms and storage resources, infrastructure and user services, user and research communities.
- **Grid technology groups:** Condor, Globus, Storage Resource Management, NSF Middleware Initiative.
- **Global research collaborations:** High Energy Physics - including Large Hadron Collider, Gravitational Wave Physics - LIGO, Nuclear and Astro Physics, Bioinformatices, Nanotechnology, CS research….
- **Partnerships:** with peers, development and research groups Enabling Grids for EScience (EGEE),TeraGrid, Regional & Campus Grids (NYSGrid, NWICG, TIGRE, GLOW..)
- **Education:** I2U2/Quarknet sharing cosmic ray data, Grid schools…



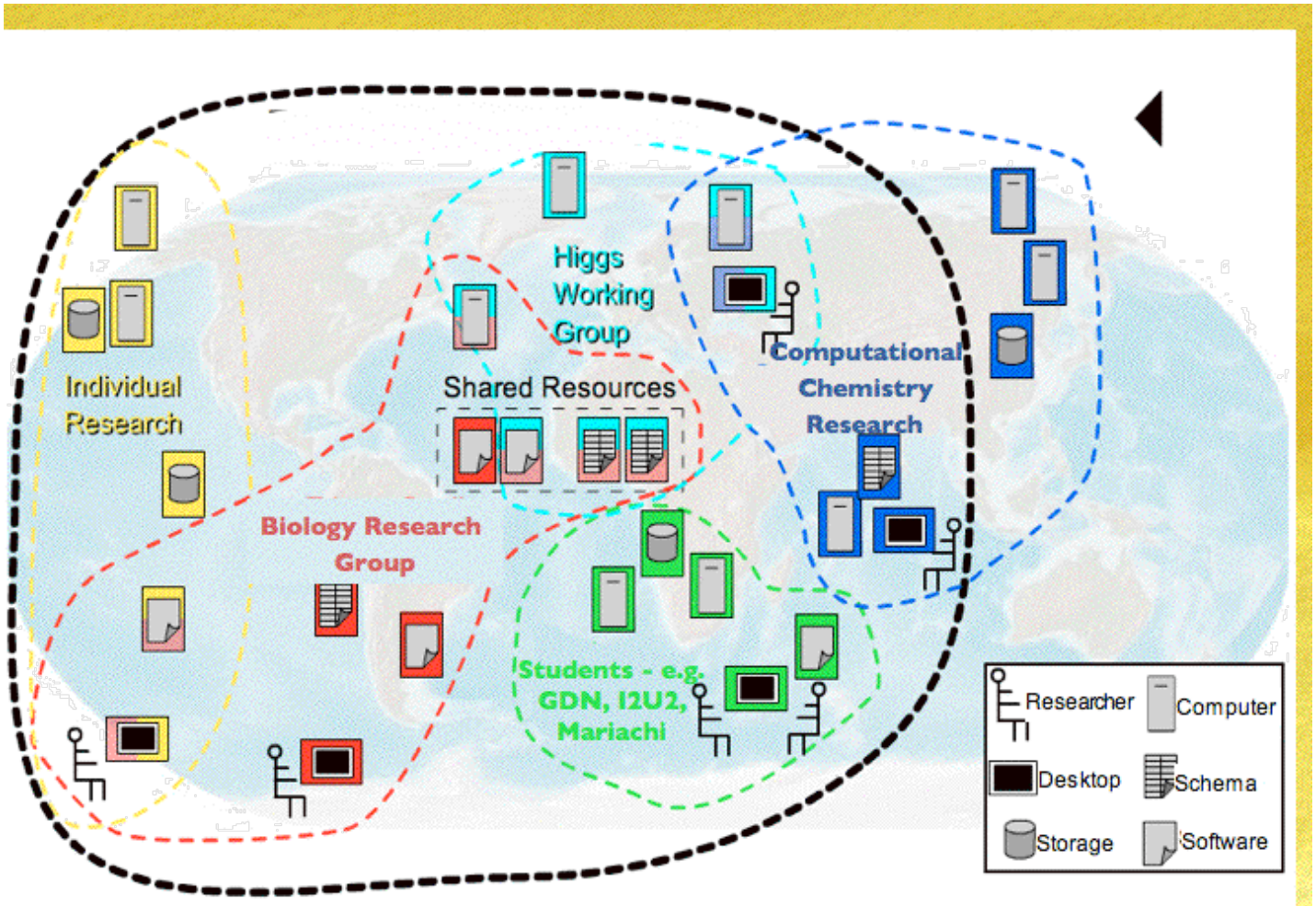
OSG sits in the middle of an environment of a Grid-of-Grids from Local to Global Infrastructures

Inter-Operating and Co-Operating Grids: Campus, Regional, Community, National, International.

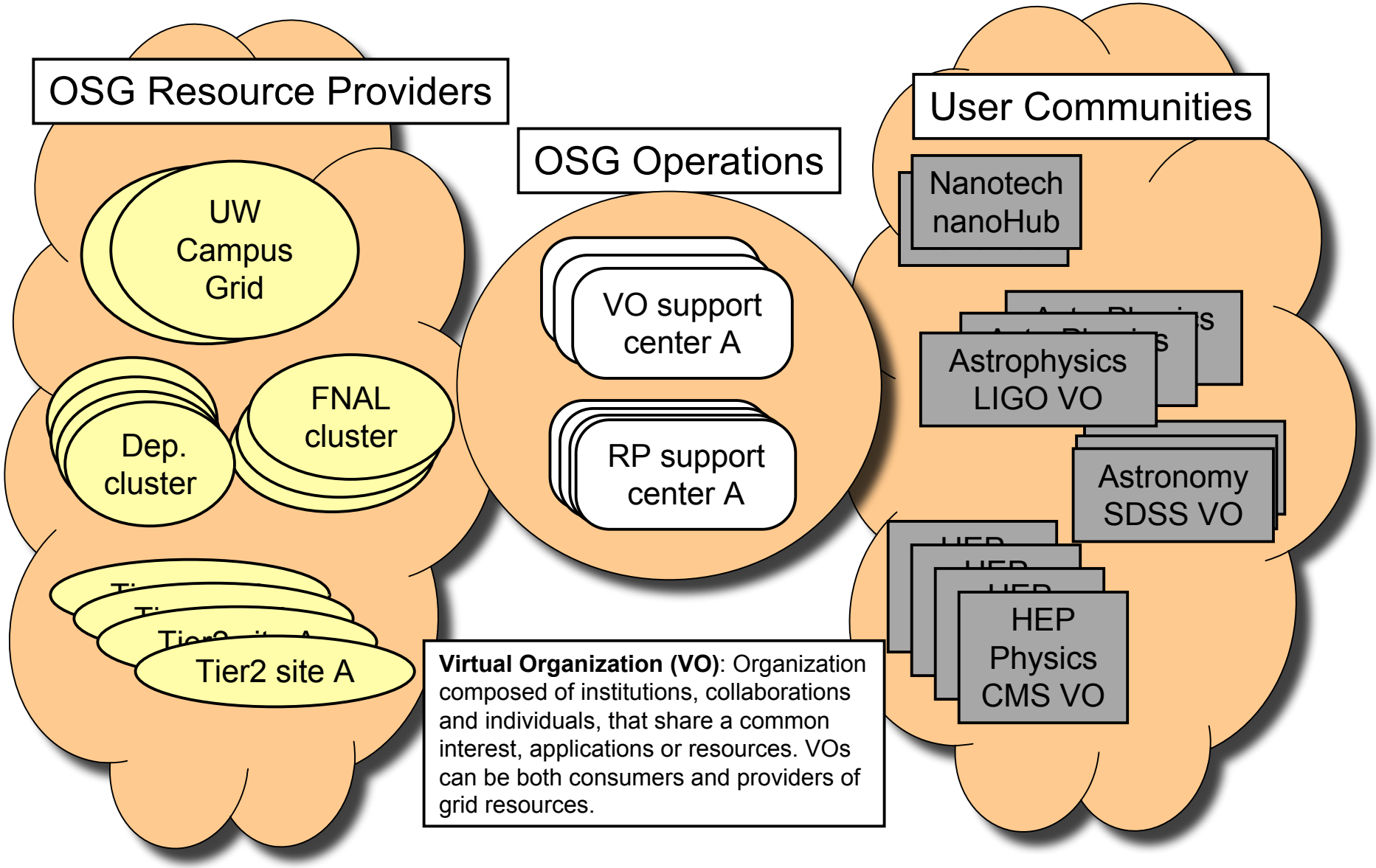
Virtual Organizations doing Research & Education.



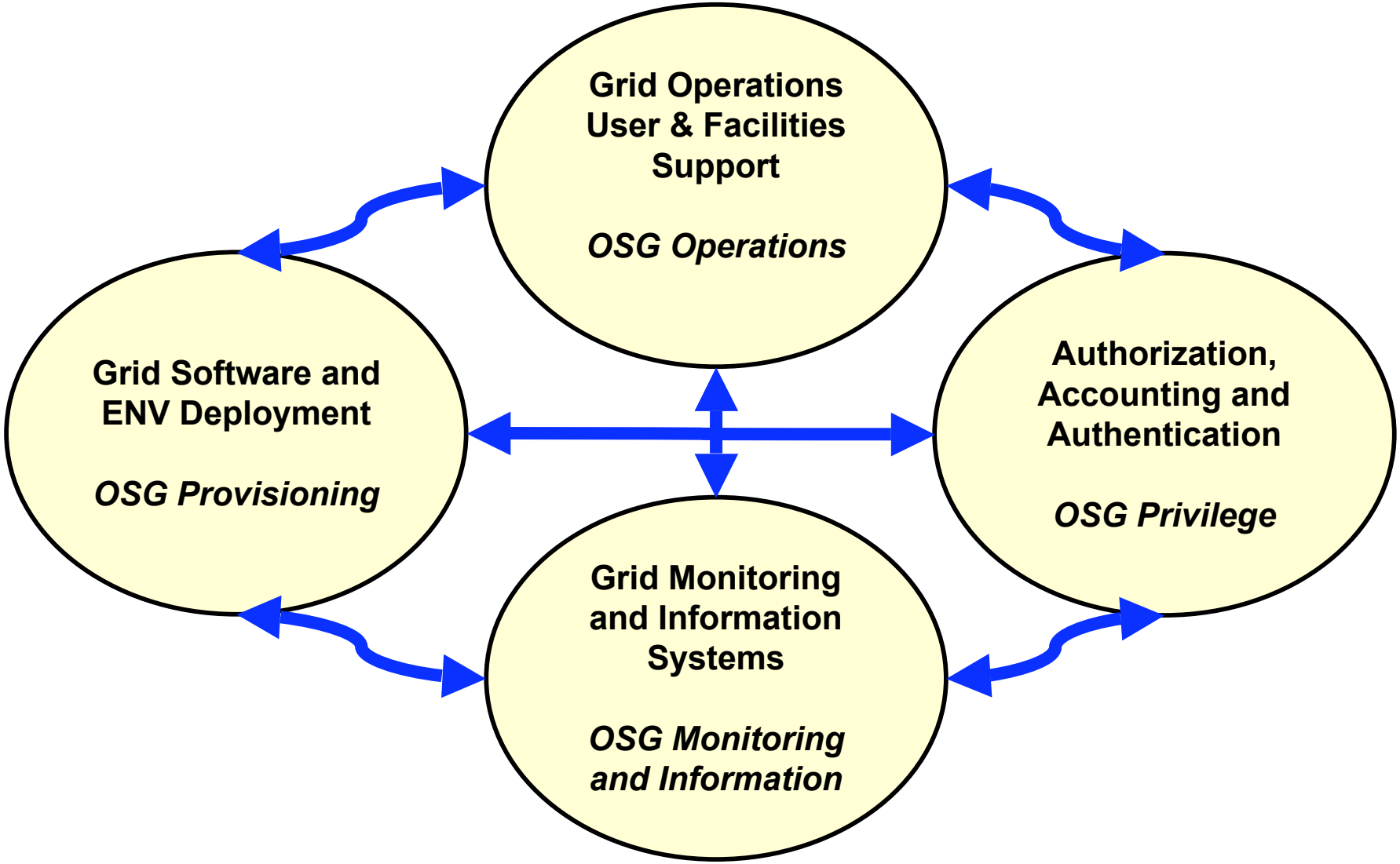
Overlaid by virtual computational environments of single to large groups of researchers local to worldwide

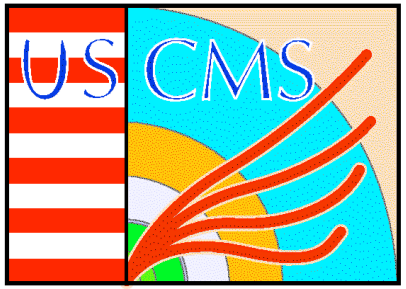


The Open Science Grid



The OSG: A High Level View



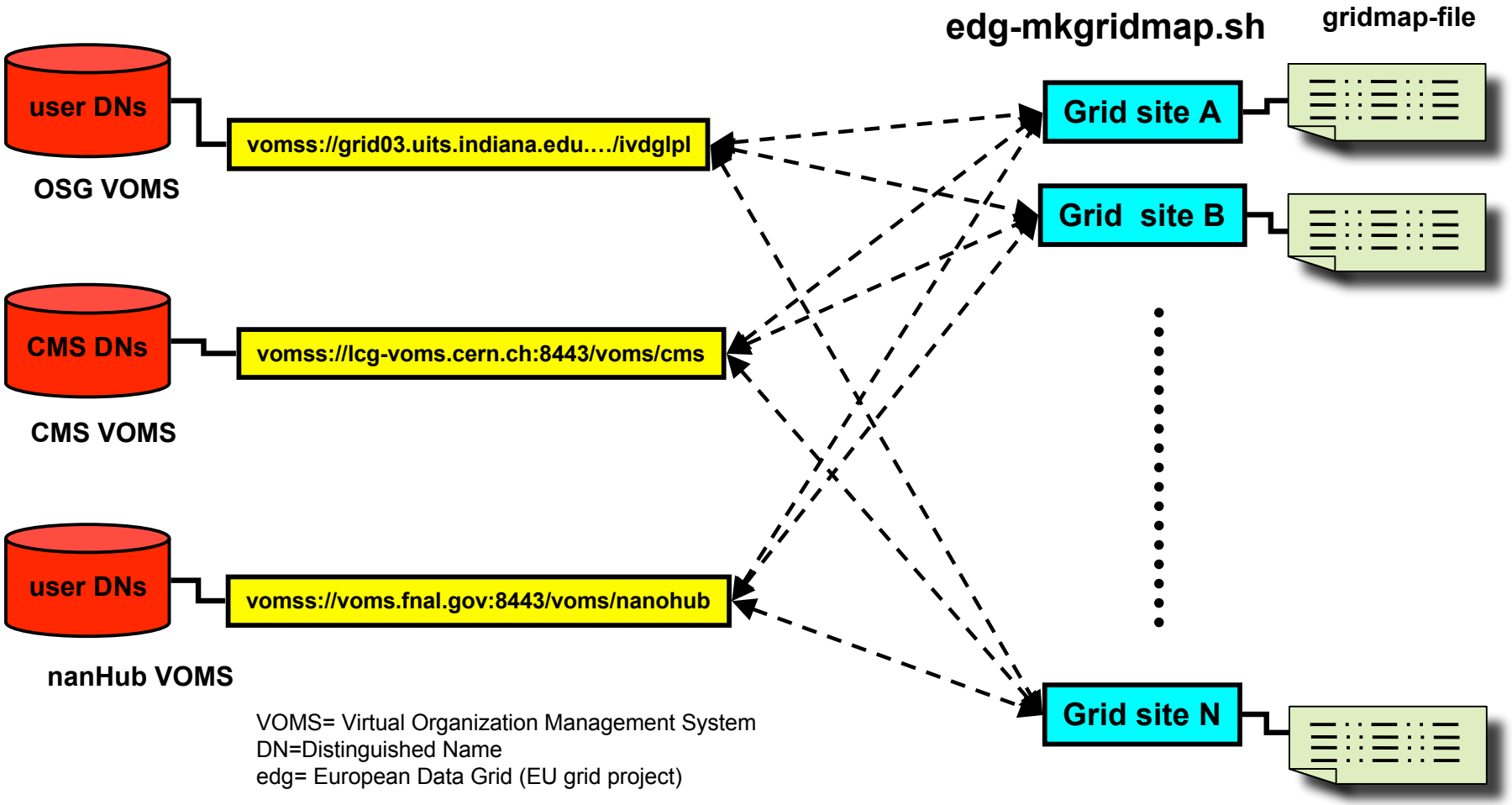


The Privilege Project

Application of a
Role **B**ased **A**ccess **C**ontrol model for OSG

An advanced authorization mechanism

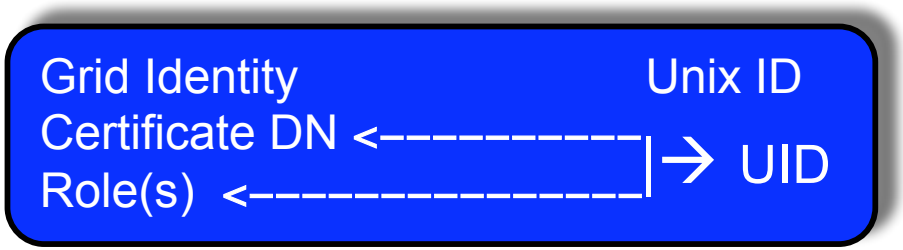
OSG Authentication (2)



VOMS= Virtual Organization Management System
DN=Distinguished Name
edg= European Data Grid (EU grid project)

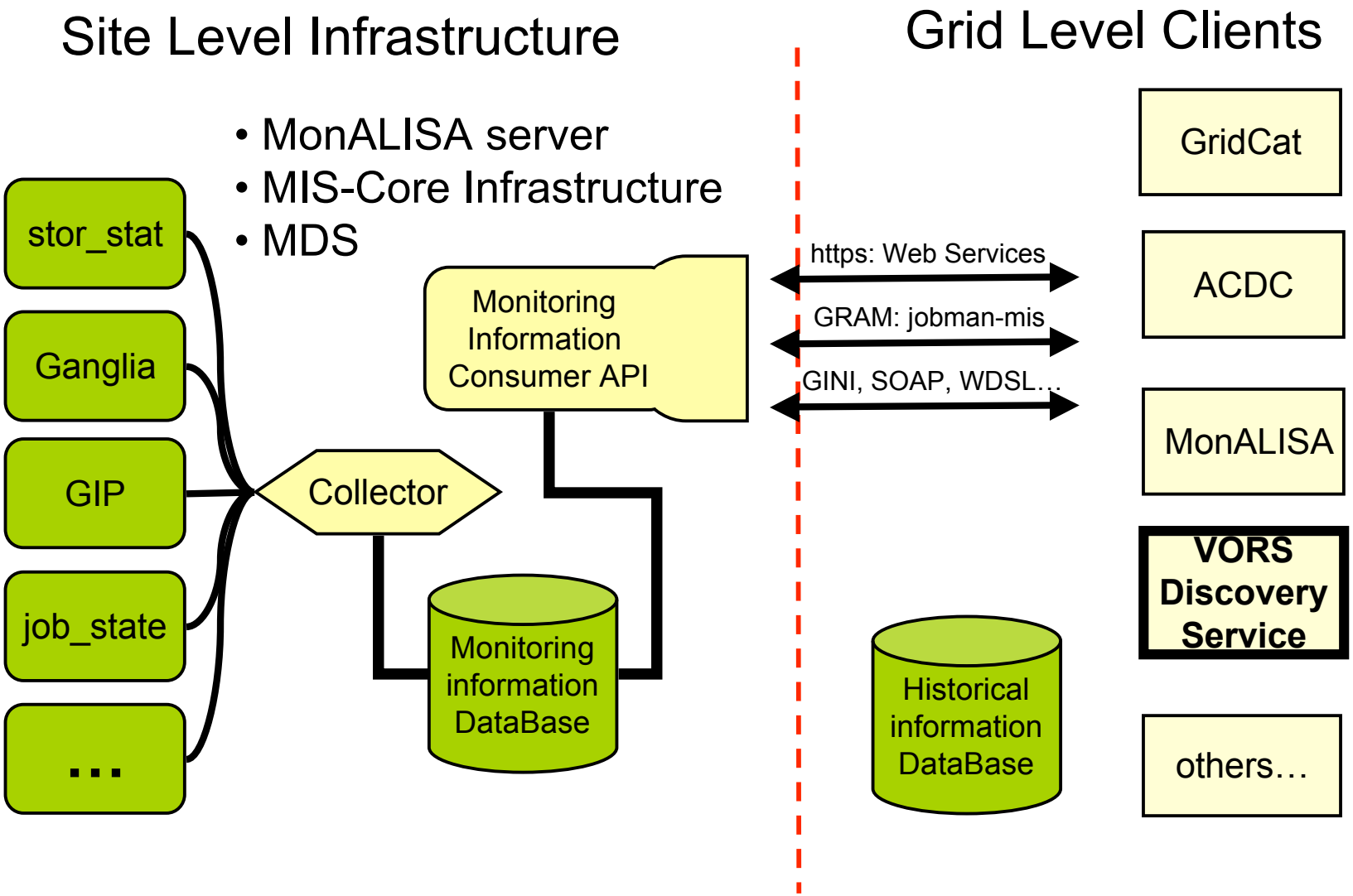
The Privilege Project Provides

- A more flexible way to assign DNs to local UNIX qualifiers, (uid, gid…)
  - VOMSes are still used to store grid identities
  - But gone are the static gridmap-files
  - voms-proxy-init replaces grid-proxy-init
    - Allows a user to specify a role along with unique ID
  - Access rights granted based on user's
    - VO membership
    - User selected role(s)



OSG Grid Monitoring

OSG Grid Monitoring



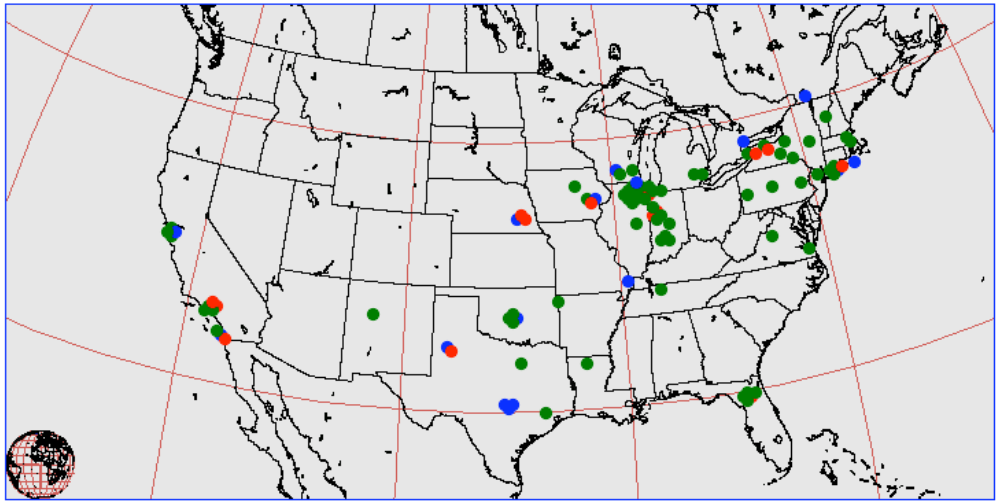
http://vors.grid.iu.edu/cgi-bin/index.cgi?grid=0&VO=0&res=0&region=1

Virtual Organization Resource Selector

Open Science Grid

All	OSG	TeraGrid	EGEE	OSG-ITB

Open Science Grid



Virtual Organization Selection

All	CDF	CMS	CompBioGrid	DES	DOSAR	DZero	Engage	Fermilab	fMRI	GADU
	geant4	GLOW	GPN	GRASE	GridChem	GridEx	GROW	i2u2	iVDGL	LIGO
mariachi	MIS	nanoHUB	NWICG	Ops	OSG	OSGEDU	SDSS	STAR	USATLAS	

Resources

Name	Gatekeeper	Type	Grid	Status	Last Test Date
BNL_ATLAS_1	gridgk01.racf.bnl.gov:2119	compute	OSG	PASS	2006-12-08 14:57:13
BNL_ATLAS_2	gridgk02.racf.bnl.gov:2119	compute	OSG	PASS	2006-12-08 14:58:43
BU_ATLAS_Tier2	atlas.bu.edu:2119	compute	OSG	PASS	2006-12-08 15:00:44

Virtual Organization Resource Selector - VORS

- Custom web interface to a grid scanner that checks services and resources on:
  - Each Compute Element
  - Each Storage Element
- Very handy for checking:
  - Paths of installed tools on Worker Nodes.
  - Location & amount of disk space for planning a workflow.
  - Troubleshooting when an error occurs.

VORS entry for OSG_LIGO_PSU

Gatekeeper: grid3.aset.psu.edu

Scheduler Types	jobmanager is of type fork jobmanager-fork is of type fork jobmanager-mis is of type mis jobmanager-pbs is of type pbs
Path to Condor Binaries	
Path to MIS Binaries	/opt/osg-ce-0.4.1/MIS-CI/bin
MDS Port	2135
VDT Version	1.3.10b
VDT Location	/opt/osg-ce-0.4.1
$APP Location	/usr1/grid3/app
$DATA Location	/usr1/grid3/data
$TMP Location	/usr1/grid3/data
$WNTMP Location	/tmp
$OSG_GRID Location	/usr1/grid3/osg-wn-0.4.1
$APP Space Available	179.065 GB
$DATA Space Available	179.065 GB
$TMP Space Available	179.065 GB

VORS is developing a grid-scanner for Storage Elements (coming soon) for OSG 0.6.0

Testing for SE SRM control protocol : YES control type = srm_v1 end point = srm://fndca1.fnal.gov:8443/ full path = /pnfs/fnal.gov/usr/fermigrid/volatile/mis executing srmls -retry_num=0 srm://fndca1.fnal.gov:8443//pnfs/fnal.gov/usr/fermigrid/volatile/mis 2>&1

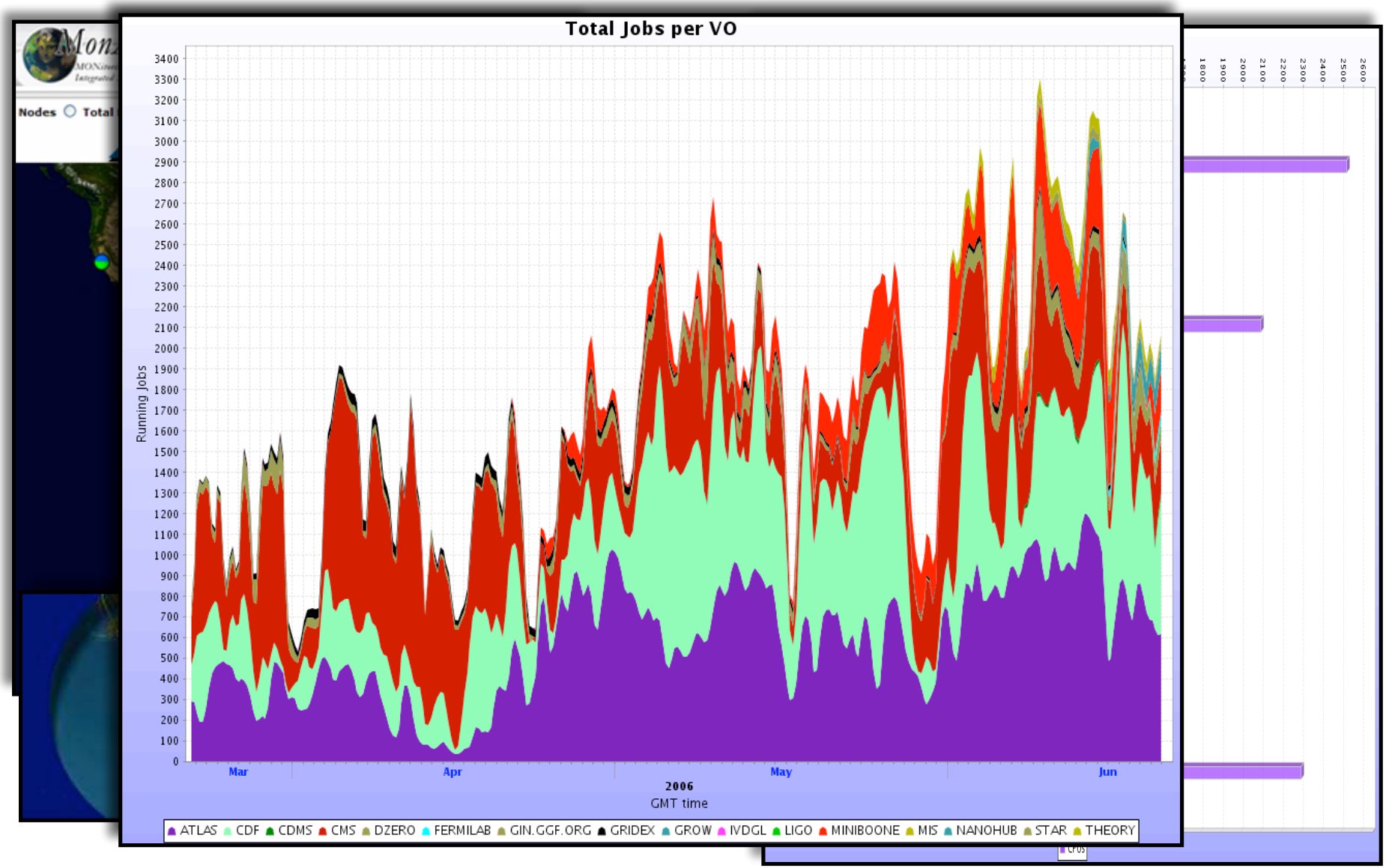
….

Testing srmls : PASS - read 12 lines 512
srm://fndca1.fnal.gov:8443//pnfs/fnal.gov/usr/fermigrid/volatile/mis 1715
srm://fndca1.fnal.gov:8443//pnfs/fnal.gov/usr/fermigrid/volatile/mis/file1 1715
srm://fndca1.fnal.gov:8443//pnfs/fnal.gov/usr/fermigrid/volatile/mis/file2 41767

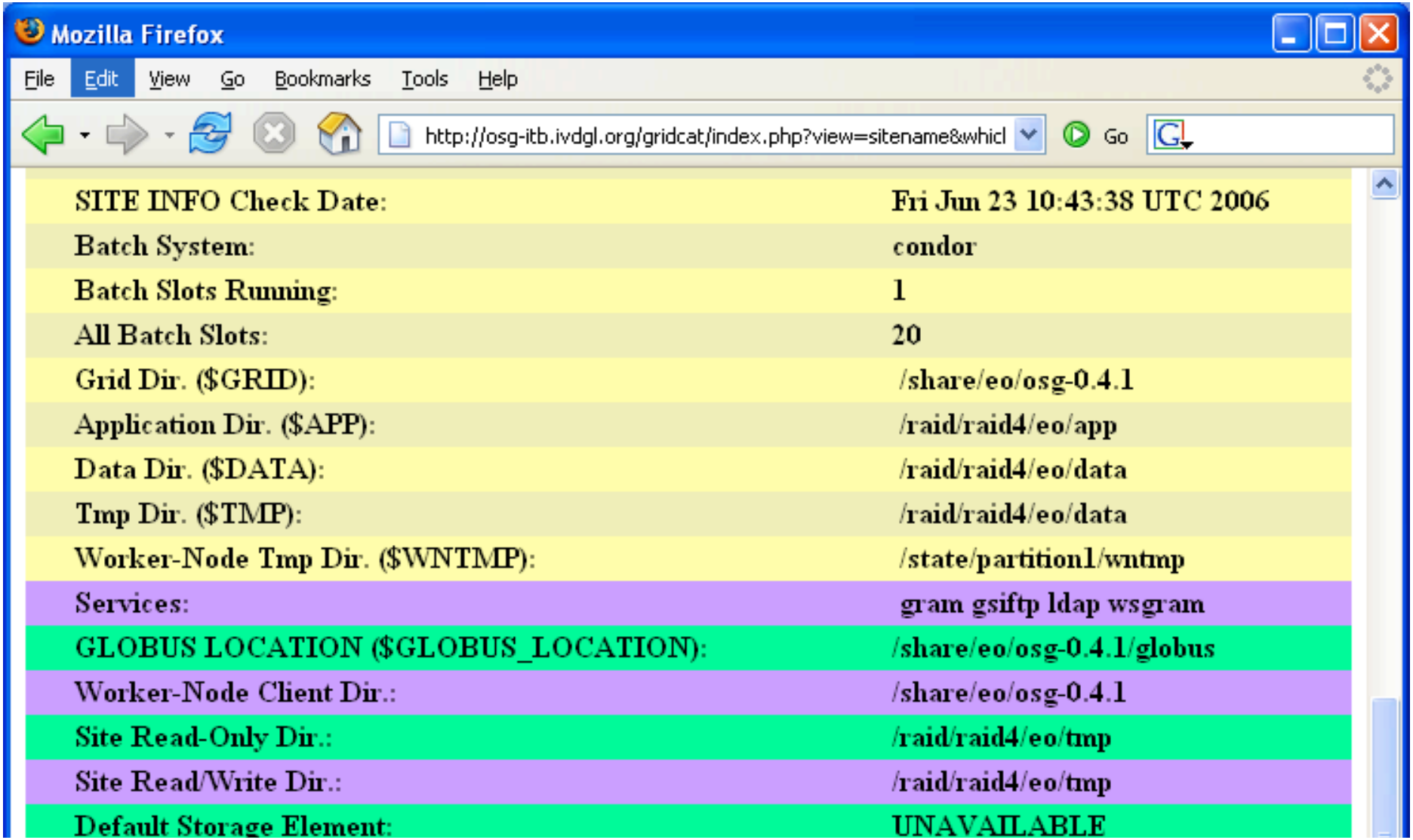
….

Testing srmcp (from SE) : FAIL - returns error code 256

MonALISA


Nodes
Total
Total Jobs per VO
Running Jobs
3400
3300
3200
3100
3000
2900
2800
2700
2600
2500
2400
2300
2200
2100
2000
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Mar
Apr
May
Jun
2006
GMT time
ATLAS CDF CDMS CMS DZERO FERMILAB GIN.GGF.ORG GRIDEX GROW IVDGL LIGO MINIBOONE MIS NANOHUB STAR THEORY
1800
1900
2000
2100
2200
2300
2400
2500
2600

The OSG **ENV**ironment

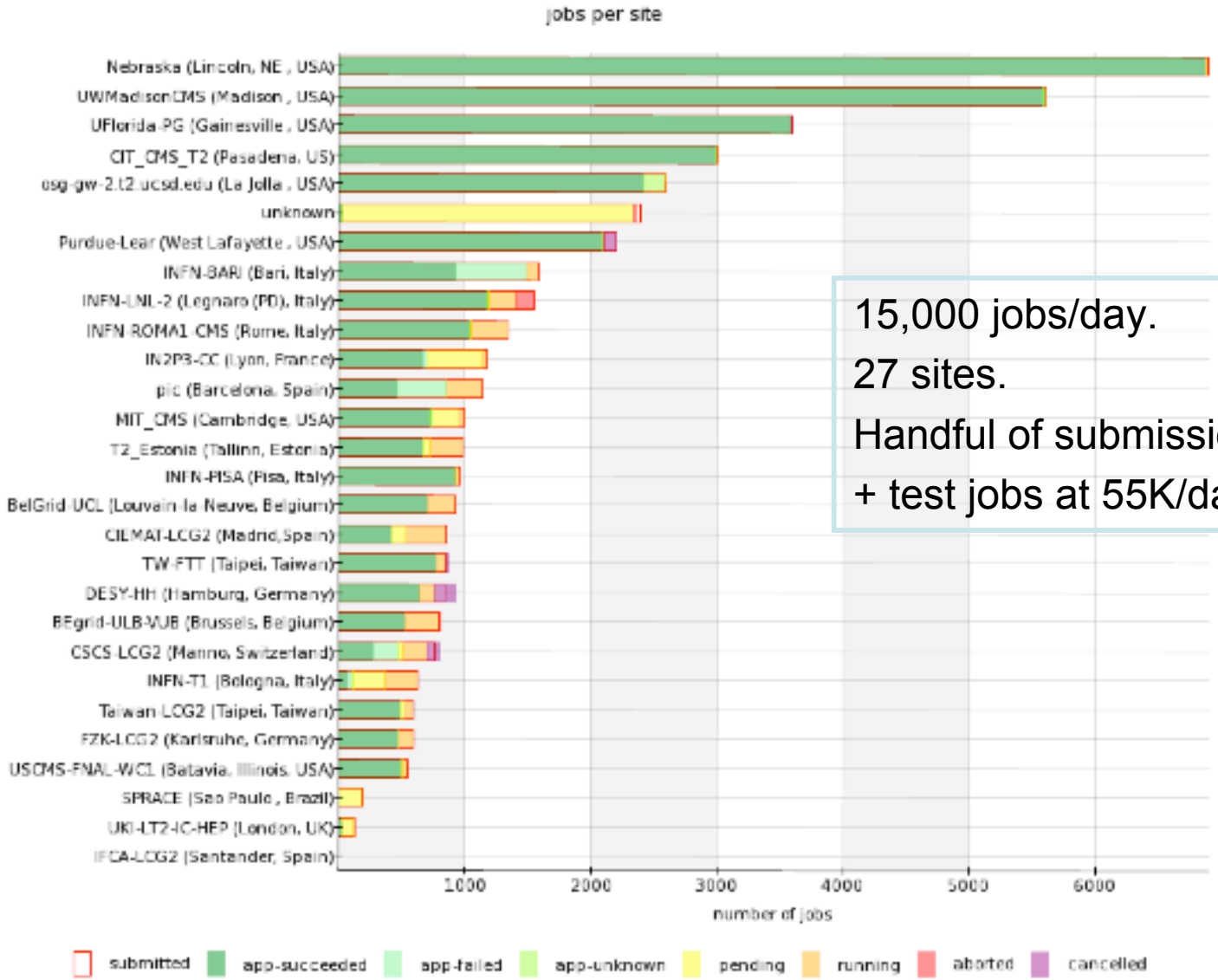


SITE INFO Check Date:	Fri Jun 23 10:43:38 UTC 2006
Batch System:	condor
Batch Slots Running:	1
All Batch Slots:	20
Grid Dir. ($GRID):	/share/eo/osg-0.4.1
Application Dir. ($APP):	/raid/raid4/eo/app
Data Dir. ($DATA):	/raid/raid4/eo/data
Tmp Dir. ($TMP):	/raid/raid4/eo/data
Worker-Node Tmp Dir. ($WNTMP):	/state/partition1/wntmp
Services:	gram gsiftp ldap wsgram
GLOBUS LOCATION ($GLOBUS_LOCATION):	/share/eo/osg-0.4.1/globus
Worker-Node Client Dir.:	/share/eo/osg-0.4.1
Site Read-Only Dir.:	/raid/raid4/eo/tmp
Site Read/Write Dir.:	/raid/raid4/eo/tmp
Default Storage Element:	UNAVAILABLE

OSG Grid Level Clients

- Tools provide basic information about OSG resources
  - Resource catalog: official tally of OSG sites
  - Resource discovery: what services are available, where are they and how do I access it
  - Metrics Information: Usage of resources over time
- Used to assess scheduling priorities
  - Where and when should I send my jobs?
  - Where can I put my output?
- Used to monitor health and status of the Grid

Submitting Locally, Executing Remotely:



15,000 jobs/day.

27 sites.

Handful of submission points.

+ test jobs at 55K/day.

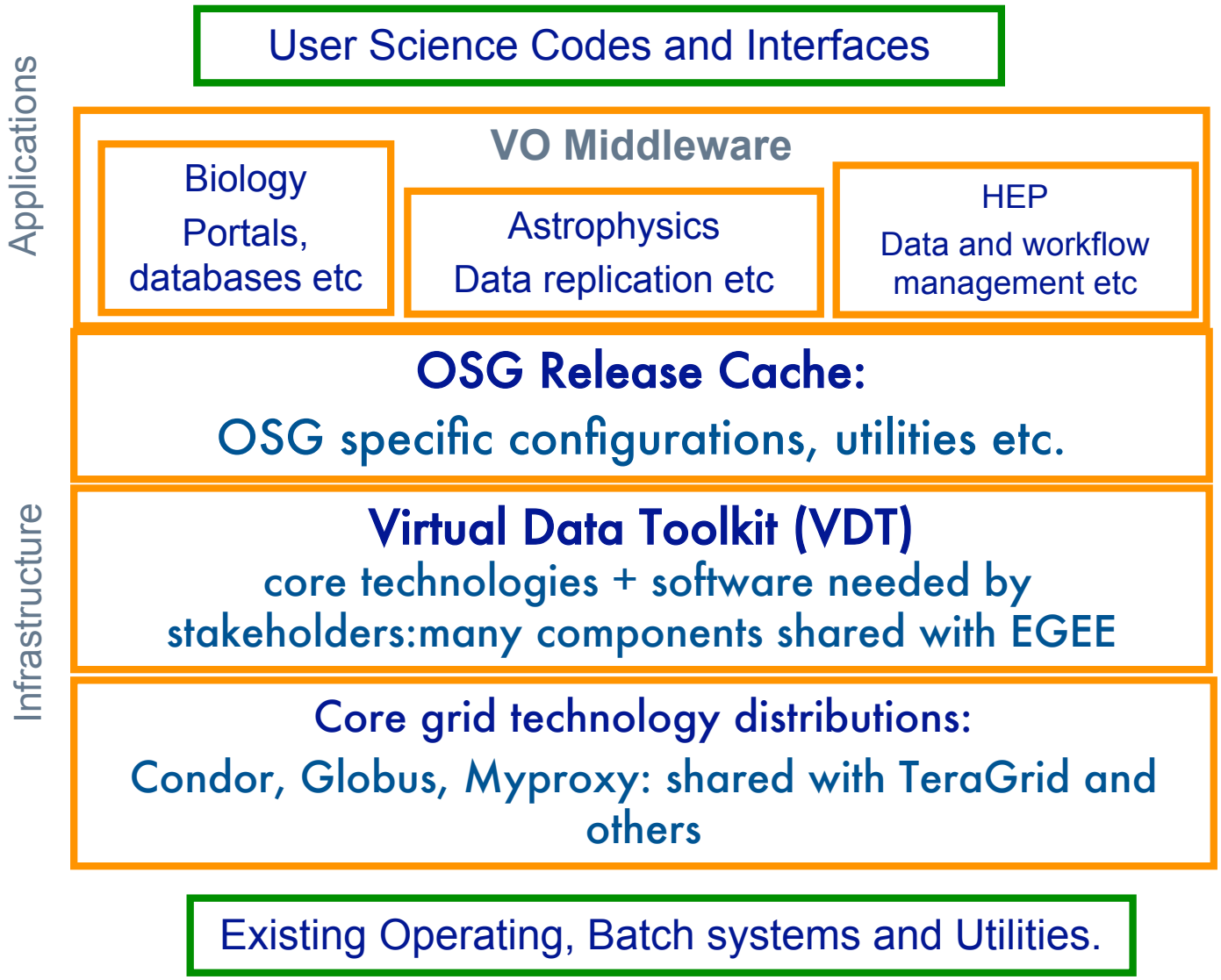
Managing Storage

- A Solution: SRM (Storage Resource Manager)
- Grid enabled interface to put data on a site
  - Provides scheduling of data transfer requests
  - Provides reservation of storage space

```
$> globus-url-copy srm://ufdcache.phys.ufl.edu/cms/foo.rfz \
 gsiftp://cit.caltech.edu/data/bar.rfz
```

- Technologies in the OSG pipeline
  - dCache/SRM (disk cache with SRM)
    - Provided by DESY & FNAL
    - SE(s) available to OSG as a service from the USCMS VO
  - DRM (Disk Resource Manager)
    - Provided by LBL
    - Can be added on top of a normal UNIX file system

OSG Middleware



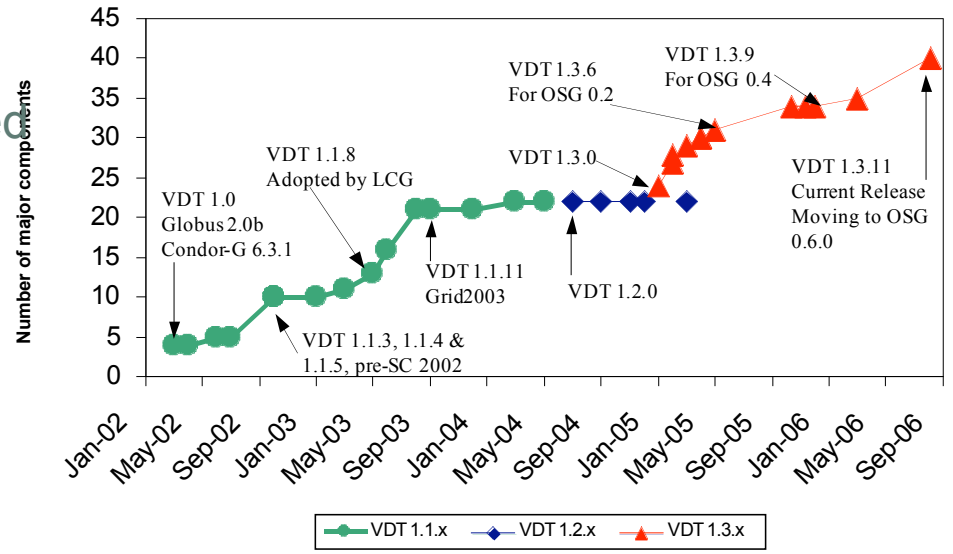
The OSG Software Cache

- Most software comes from the Virtual Data Toolkit (VDT)
- OSG components include
  - VDT configuration scripts
  - Some OSG specific packages too

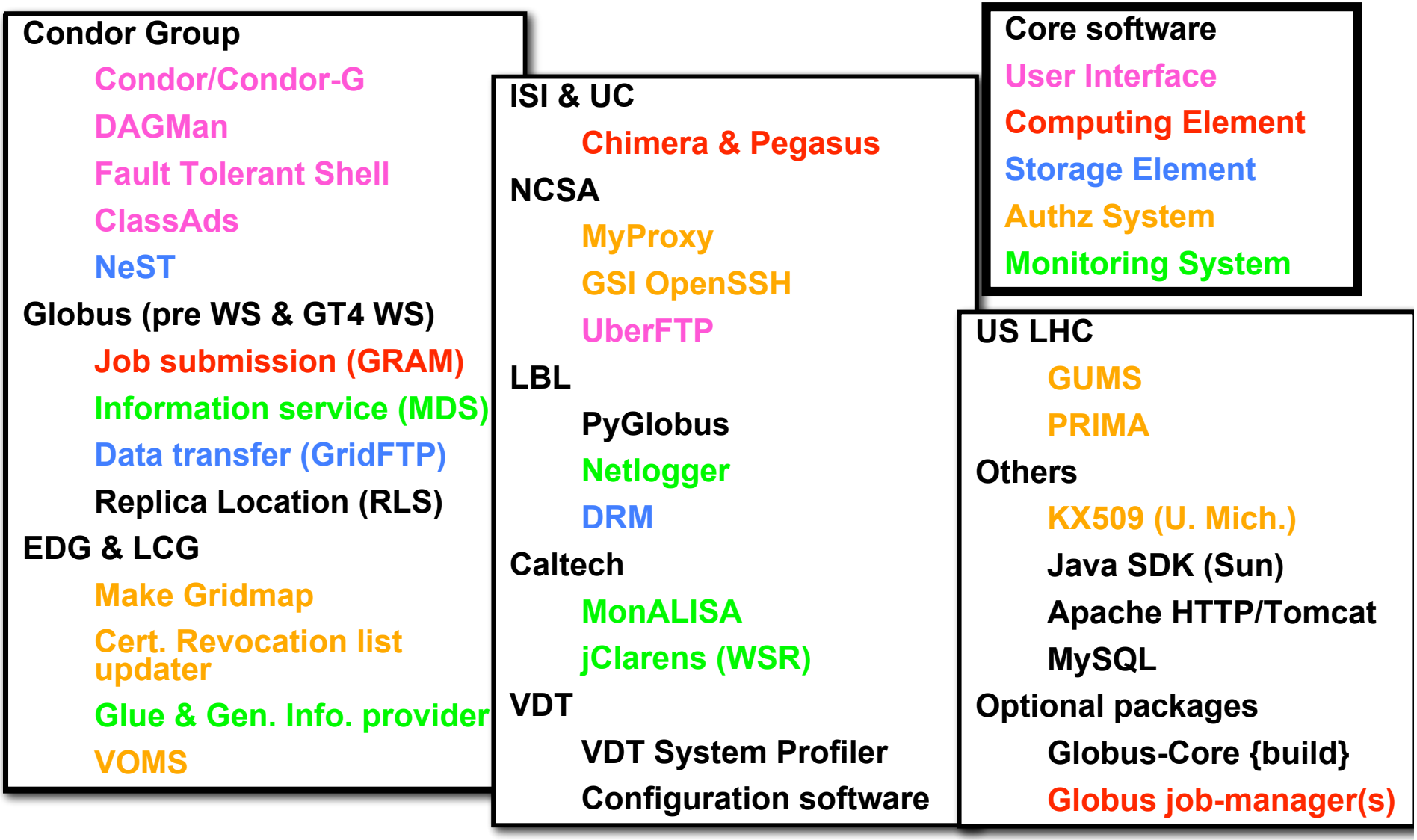
- Pacman is the OSG Meta-packager
  - This is how we deliver the entire cache to Resource Providers

What is the VDT?

- A collection of software
  - Grid software: Condor, Globus and lots more
  - Virtual Data System: Origin of the name “VDT”
  - Utilities: Monitoring, Authorization, Configuration
  - Built for >10 flavors/versions of Linux
- Automated Build and Test: Integration and regression testing.
- An easy installation:
  - Push a button, everything just works.
  - Quick update processes.
- Responsive to user needs:
  - process to add new components based on community needs.
- A support infrastructure:
  - front line software support,
  - triaging between users and software providers for deeper issues.



What is in the VDT? (A lot!)



Pacman

- **Pacman is:**
  - a software environment installer (or Meta-Packager)
  - a language for defining software environments
  - an interpreter that allows creation, installation, configuration, update, verification and repair of installation environments
  - takes care of dependencies
- **Pacman makes installation of all types of software easy**

LCG/Scram Globus/GPT Nordugrid/RPM

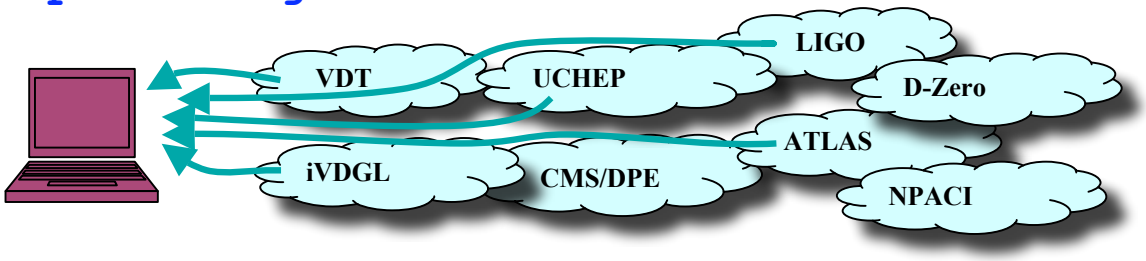
ATLAS/CMT NPACI/TeraGrid/tar/make

LIGO/tar/make D0/UPS-UPD Commercial/tar/make

OpenSource/tar/make CMS DPE/tar/make

Enables us to easily and coherently combine and manage software from arbitrary sources.

```
% pacman -get OSG:CE
```



Enables remote experts to define installation config updating for everyone at once.

Pacman Installation

1. Download Pacman
   - `http://physics.bu.edu/~youssef/pacman/`
2. Install the “package”
   - `cd <install-directory>`
   - `pacman -get OSG:OSG_CE_0.2.1`
   - `ls`

```
condor/   globus/   post-install/   setup.sh
edg/      gpt/      replica/        vdt/
ftsh/     perl/     setup.csh       vdt-install.log
/monalisa ...
```

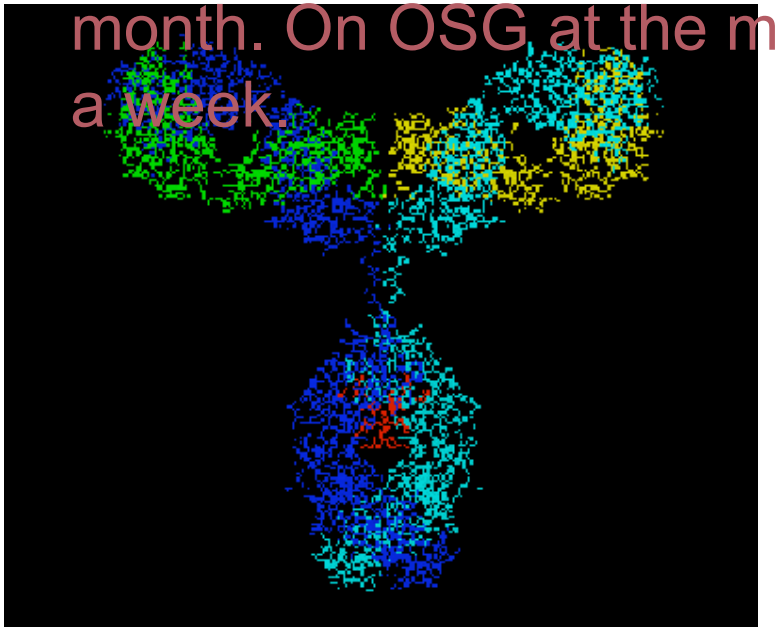
Grid Operations Center

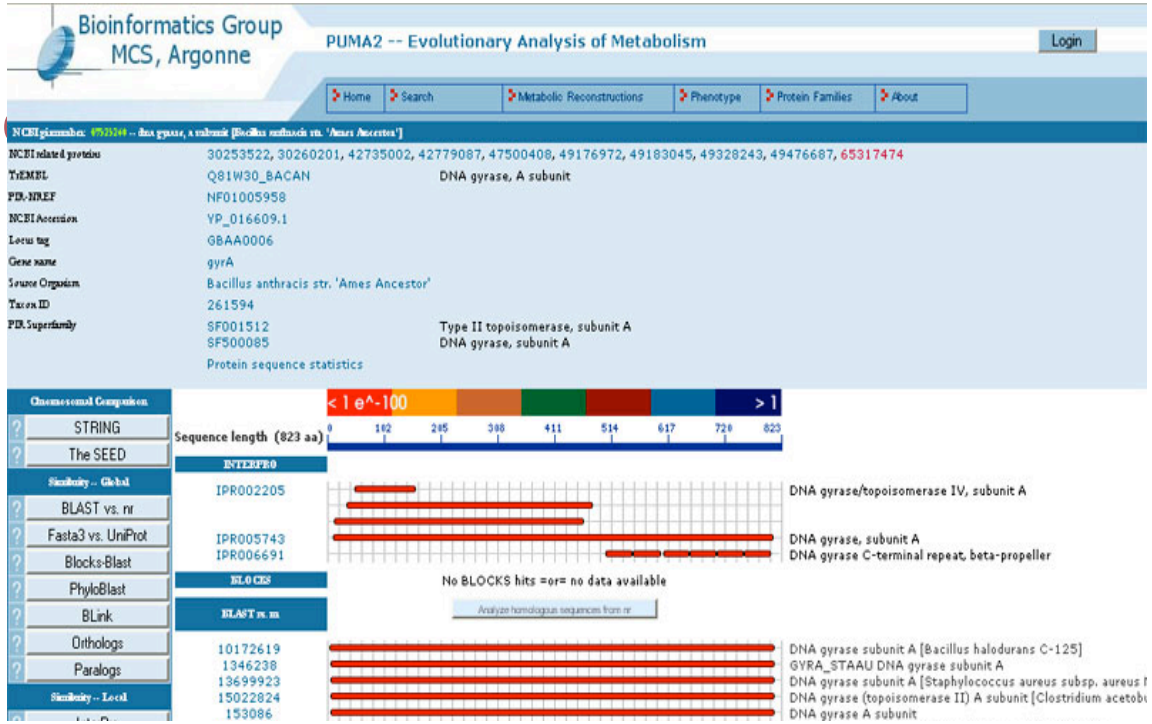
- Based at Indiana University and provides a central repository of staff and monitoring systems for:
  - Real time grid monitoring.
  - Problem tracking via a trouble ticket system.
  - Support for developers and sys admins.
  - Maintains infrastructure – VORS, MonALISA and registration DB.
  - Maintains OSG software repositories.

Applications can cross infrastructures e.g: OSG and TeraGrid

Genome Analysis and Database Update system

- Runs across TeraGrid and OSG. Uses the Virtual Data System (VDS) workflow & provenance.
- Pass through public DNA and protein databases for new and newly updated genomes of different organisms and runs BLAST, Blocks, Chisel. 1200 users of resulting DB.
- Request: 1000 CPUs for month. On OSG at the m a week.





Summary of OSG today

- Providing core services, software and a distributed facility for an increasing set of research communities.
- Helping Virtual Organizations access resources on many different infrastructures.
- Reaching out to others to collaborate and contribute our experience and efforts.

TeraGrid Overview

Open Science Grid

TeraGrid™

What is the TeraGrid?

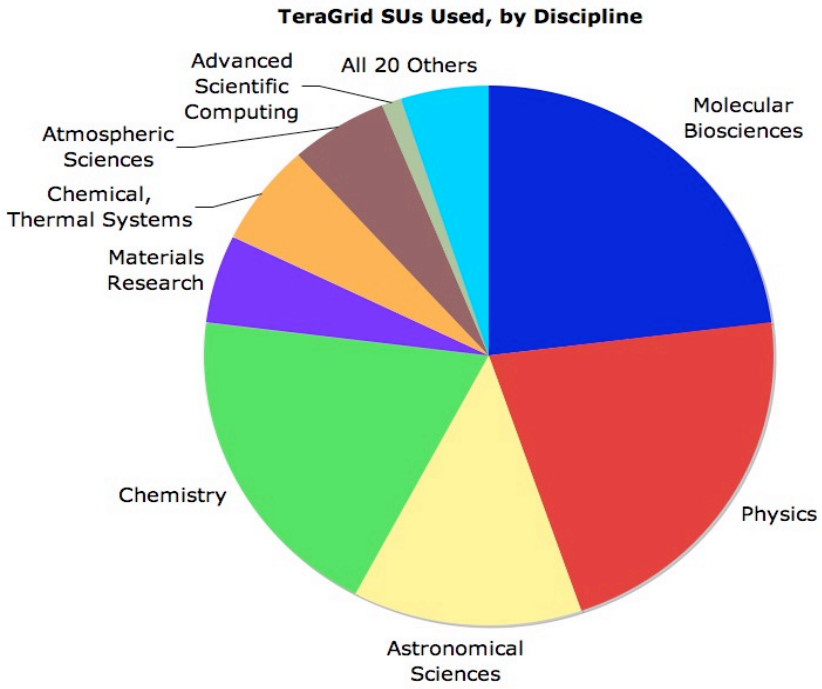
Technology + Support = Science

The TeraGrid Facility

- Grid Infrastructure Group (GIG)
  - University of Chicago
  - TeraGrid integration, planning, management, coordination
  - Organized into areas
    - User Services
    - Operations
    - Gateways
    - Data/Visualization/Scheduling
    - Education Outreach & Training
    - Software Integration
- Resource Providers (RP)
  - Currently NCSA, SDSC, PSC, Indiana, Purdue, ORNL, TACC, UC/ANL
  - Systems (resources, services) support, user support
  - Provide access to resources via policies, software, and mechanisms coordinated by and provided through the GIG.

NSF Funded Research

- NSF-funded program to offer high end compute, data and visualization resources to the nation's academic researchers
- Proposal-based, researchers can use resources at no cost
- Variety of disciplines



TeraGrid Hardware Components

- High-end compute hardware
  - Intel/Linux clusters
  - Alpha SMP clusters
  - IBM POWER3 and POWER4 clusters
  - SGI Altix SMPs
  - SUN visualization systems
  - Cray XT3
  - IBM Blue Gene/L
- Large-scale storage systems
  - hundreds of terabytes for secondary storage
- Visualization hardware
- Very high-speed network backbone (40Gb/s)
  - bandwidth for rich interaction and tight coupling

TeraGrid Resources

	ANL/UC	IU	NCSA	ORNL	PSC	Purdue	SDSC	TACC
Computational Resources	Itanium 2 (0.5 TF) IA-32 (0.5 TF)	Itanium2 (0.2 TF) IA-32 (2.0 TF)	Itanium2 (10.7 TF) SGI SMP (7.0 TF) Dell Xeon (17.2TF) IBM p690 (2TF) Condor Flock (1.1TF)	IA-32 (0.3 TF)	XT3 (10 TF) TCS (6 TF) Marvel SMP (0.3 TF)	Hetero (1.7 TF) IA-32 (11 TF) *Opportunistic*	Itanium2 (4.4 TF) Power4+ (15.6 TF) Blue Gene (5.7 TF)	IA-32 (6.3 TF)
Online Storage	20 TB	32 TB	1140 TB	1 TB	300 TB	26 TB	1400 TB	50 TB
Mass Storage		1.2 PB	5 PB		2.4 PB	1.3 PB	6 PB	2 PB
Net Gb/s, Hub	30 CHI	10 CHI	30 CHI	10 ATL	30 CHI	10 CHI	10 LA	10 CHI
Data Collections # collections Approx total size Access methods		5 Col. >3.7 TB URL/DB/ GridFTP	> 30 Col. URL/SRB/DB/ GridFTP			4 Col. 7 TB SRB/Portal/ OPeNDAP	>70 Col. >1 PB GFS/SRB/ DB/GridFTP	4 Col. 2.35 TB SRB/Web Services/ URL
Instruments		Proteomics X-ray Cryst.		SNS and HFIR Facilities				
Visualization Resources RI: Remote Interact RB: Remote Batch RC: RI/Collab	RI, RC, RB IA-32, 96 GeForce 6600GT		RB SGI Prism, 32 graphics pipes; IA-32		RI, RB IA-32 + Quadro4 980 XGL	RB IA-32, 48 Nodes	RB	RI, RC, RB UltraSPARC IV, 512GB SMP, 16 gfx cards

100+ TF

8 distinct architectures

3 PB Online Disk

>100 data collections

TeraGrid Software Components

- Coordinated TeraGrid Software and Services "CTSS"
  - Grid services
  - Supporting software
- Community Owned Software Areas "CSA"
- Advanced Applications

Coordinated TeraGrid Software & Services 4

- CTSS 4 Core Integration Capability
  - Authorization/Accounting/Security
  - Policy
  - Software deployment
  - Information services
- Remote Compute Capability Kit
- Data Movement and Management Capability Kit
- Remote Login Capability Kit
- Local Parallel Programming Capability Kit
- Grid Parallel Programming Capability Kit
- <more capability kits>

Science Gateways
A new initiative for the TeraGrid

- Increasing investment by communities in their own cyberinfrastructure, but heterogeneous:
  - Resources
  - Users – from expert to K-12
  - Software stacks, policies
- Science Gateways
  - Provide "TeraGrid Inside" capabilities
  - Leverage community investment
- Three common forms:
  - Web-based Portals
  - Application programs running on users' machines but accessing services in TeraGrid
  - Coordinated access points enabling users to move seamlessly between TeraGrid and other grids.

Gateways are growing in numbers

- 10 initial projects as part of TG proposal
- >20 Gateway projects today
- No limit on how many gateways can use TG resources
  - Prepare services and documentation so developers can work independently

- Open Science Grid (OSG)
- Special PRiority and Urgent Computing Environment (SPRUCE)
- National Virtual Observatory (NVO)
- Linked Environments for Atmospheric Discovery (LEAD)
- Computational Chemistry Grid (GridChem)
- Computational Science and Engineering Online (CSE-Online)
- GEON(GEOsciences Network)
- Network for Earthquake Engineering Simulation (NEES)
- SCEC Earthworks Project
- Network for Computational Nanotechnology and nanoHUB
- GIScience Gateway (GISolve)
- Biology and Biomedicine Science Gateway
- Open Life Sciences Gateway
- The Telescience Project
- Grid Analysis Environment (GAE)
- Neutron Science Instrument Gateway
- TeraGrid Visualization Gateway, ANL
- BIRN
- Gridblast Bioinformatics Gateway
- Earth Systems Grid
- Astrophysical Data Repository (Cornell)

- Many others interested
  - SID Grid
  - HASTAC

For More Info

- Open Science Grid
  - http://www.opensciencegrid.org
- TeraGrid
  - http://www.teragrid.org

it’s the people…that make the grid a community!