Annual Report for Period: 09/2006 - 08/2007
Submitted on: 07/31/2007
Principal Investigator: Livny, Miron .
Award ID: 0621704
Organization: U of Wisconsin Madison
Title:
Sustaining and Extending the Open Science Grid: Science Innovation on a PetaScale Nationwide Facility

Project Participants

Senior Personnel

Name: Livny, Miron
Worked for more than 160 Hours: Yes
Contribution to Project:
and OSG Facility Coordinator

Name: Avery, Paul
Worked for more than 160 Hours: Yes
Contribution to Project:
and OSG Resources Co-Manager

Name: Foster, Ian
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Blackburn, Kent
Worked for more than 160 Hours: Yes
Contribution to Project:
and OSG Resources Co-Manager

Name: Pordes, Ruth
Worked for more than 160 Hours: Yes
Contribution to Project:
and OSG Executive Director

Name: Blatecky, Alan
Worked for more than 160 Hours: Yes
Contribution to Project:
also Engagment Coordinator

Name: Shank, Jim
Worked for more than 160 Hours: No
Contribution to Project:

Name: Cowles, Robert
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Deelman, Ewa
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Dodd, Jeremy
Worked for more than 160 Hours: No
Contribution to Project:

Name: Gardner, Rob
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Integration and Site Coordinator

Name: Gibbons, Lawrence
Worked for more than 160 Hours: No
Contribution to Project:

Name: Gordon, Howard
Worked for more than 160 Hours: No
Contribution to Project:

Name: Wang, Shaowen
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Troubleshooting Coordinator

Name: Wenaus, Torre
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Extensions and Applications Co-Coordinator

Name: Kramer, Bill
Worked for more than 160 Hours: Yes
Contribution to Project:
Chair of the OSG Consortium Council

Name: MacCaulay, Scott
Worked for more than 160 Hours: No
Contribution to Project:

Name: McGee, John
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Engagement Coordinator

Name: Olson, Doug
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Petravick, Don
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Security Officer

Name: Quick, Rob
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Wuerthwein, Frank
Worked for more than 160 Hours: Yes
Contribution to Project:

OSG Extensions and Applications Co-Coordinator

Name: Roy, Alain
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Software Coordinator

Name: Wilde, Mike
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Education Coordinator

Post-doc

Graduate Student

Undergraduate Student

Technician, Programmer

Other Participant

Name: Burne, Cristy
Worked for more than 160 Hours: Yes
Contribution to Project:
Editor of the International Science Grid This Week Newsletter

Name: Green, Chris
Worked for more than 160 Hours: Yes
Contribution to Project:
.

Name: Heavey, Anne
Worked for more than 160 Hours: Yes
Contribution to Project:
OSG Communicator.

Name: Sehgal, Chander
Worked for more than 160 Hours: Yes
Contribution to Project:
.

Name: Rana, Abhishek
Worked for more than 160 Hours: Yes
Contribution to Project:
.

Name: Tannenbaum, Todd
Worked for more than 160 Hours: No
Contribution to Project:

Name: Altunay, Mine
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Bacon, Charles
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Bejan, Alina
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Brieger, Leesa
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Cartwright, Tim
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Chiu, Barnett
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Clifford, Ben
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Britta, Britta
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: DeSmet, Alan
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Frey, Jamie
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Hesselroth, Ted
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Hover,, John
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Huang, Charles
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Kronenfeld, Scot

Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Martin, Terrence
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Oleynik, Gene
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Packard, Jay
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Padmanabhan, Anand
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Pavlo, Andy
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Porter, Jeff
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Rosheck, John
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Rynge, Mats
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Sharma,, Neha
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Sharp, Greg
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Silvers, Tim
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Sim,, Alex
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Teckenbrock, Marcia
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Thapa,, Suchandra
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Vahi, Karan
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Youssef, Saul
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Weigand, John
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Brunelle, John
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Zalokar, Michael
Worked for more than 160 Hours: Yes
Contribution to Project:

Research Experience for Undergraduates

Organizational Partners

Boston University

Brookhaven National Laboratory

California Institute of Technology

Cornell University

Fermi National Accelerator Laboratory

Columbia University

Indiana University

Information Sciences Institute

Lawrence Berkeley National Laboratory

Purdue University

University of North Carolina

Stanford Linear Accelerator Center

University of California San Diego

University of Chicago

University of Florida

University of Illinois at Urbana-Champaign

University of Wisconsin-Madison

University of Wisconsin-Milwaukee

Other Collaborators or Contacts

The OSG relies on external project collaborations to develop the software to be included in the OSG Virtual Data Toolkit and deployed on the shared common distributed infrastructure.

The external projects OSG has worked with over the past year are:
- Community Driven Improvement of Globus Software (CDIGS),
- SciDAC-2 Center for Enabling Distributed Petascale Science (CEDPS),
- the Condor Project,
- the dCache Collaboration,
- the Data Intensive Science University Network (DISUN),
- the Energy Sciences Network (ESnet),
- Internet2,
- LIGO Physics at the Information Frontier,
- Fermilab Gratia Accounting,
- the Storage Resource Management collaboration at LBNL,
- the U.S. Large Hadron Collider experiment Software and Computing Projects.

Activities and Findings

Research and Education Activities:

In the first part of 2007 the Open Science Grid (OSG) consortium established and operated the first generation of a shared national cyberinfrastructure bringing together advanced grid technologies to provide end-to-end distributed capabilities to a broad range of compute intensive application. The 0.6 OSG software stack released early this year offers the community additional capabilities integrated into the Virtual Data Toolkit (VDT). The 0.6 versions offers significant improvements in ease of deployment and maintenance of the software stack as well as advanced accounting and monitoring capabilities.

The OSG Year 1 Project plan provided a set of deliverables and milestones to meet the needs of the stakeholder science communities in high

energy and nuclear physics, gravitational wave physics and astrophysics, as well as engagement with other science communities through the engagement program.

Findings: (See PDF version submitted by PI at the end of the report)

Many challenges in the deployment, evolution and operation of a shared distributed infrastructure at a national scale were exposed and addressed during the first year of the OSG project. Also, missing capabilities of the existing software tools were identified and possible solutions were studied. As new procedures are defined and new organizational structured developed, they are documented and shared with other national and international cyber-infrastructure projects that are driven by the evolving needs of open science. The detailed security plan for the OSG released in early 2007 is a good example of such a document.

Findings enabled by the Distributed Infrastructure:

Science Deliverables Physical Sciences:

LIGO: LIGO has met the two deliverables to date to test the operation of their Inspiral Analysis application using the Pegasus Workflow technologies across multiple sites on the OSG: to run continuously for one week at on the UCSD cluster using 25 or more batch slots; to be able to run continuously for one week using multiple OSG sites with a sustained load of 100 or more batch slots in mid February, roughly four months ahead of schedule and with an average load of nearly twice the required level. Seven OSG sites were typically involved in reaching this throughput.

The LHC: OSG has delivered the necessary throughput and reliability to meet the needs of the US ATLAS and CMS experiments for their data challenges and simulation production in late 2006, early 2007. Interoperability with and contributions to the WorldWide LHC Computing Grid have been a success. Data distribution at rates of more than 1.2 GBytes/sec from the CERN 'Tier-0' to the Brookhaven and Fermilab Tier-1s and fourteen University Tier-2s was achieved.

Tevatron Run II: CDF and D0 have successfully used OSG for Monte-Carlo (both), event production processing (D0) and analysis (CDF). A major success of OSG's Year 1 program was the response to an unanticipated request for significant resources by D0 to do a full re-processing of a multi-terabyte data set (500 million events). . Over 50% of the events were processed using opportunistically available resources on the OSG (see Figure 1). This was an important demonstration of the ability of the OSG Consortium stakeholders to contribute resources to the common infrastructure while still maintaining control for their own use. DZero was able to use more than twelve sites, sustained execution of over 1000 simultaneous jobs, used more than 2 million CPU wallclock hours and moved over 70 Terabytes of data. 'This is the first major production of real high energy physics data (as opposed to simulations) ever run on OSG resources,' said Brad Abbott of the University of Oklahoma, head of the DZero Computing group. Reprocessing was completed in June. Towards the end of the production run the throughput on OSG were more than 5 million events per day?, two to three times more than originally planned.

CDF has moved all its simulation activities to the grid.' Our physicists are submitting jobs to the grid sites on a daily basis without thinking about what resources are on the back end,' says Duke University's Ashutosh Kotwal. ?? Today CDF is actively using five grid sites, has five more in the works, and plans to expand gradually from there.

Nuclear physics: The STAR experiment has continued to use the OSG data movement capabilities between LBNL, BNL and new sites on the OSG at Wayne State, Sao Paolo in Brazil, and in tests at the University of Illinois at Chicago. STAR is converting its simulation production to run on OSG, has worked with the Troubleshooting team to solve problems in efficiency, and through solving problems of robustness and configuration with the Fermilab FermiGrid site has moved its simulation into production on the OSG infrastructure.

Astrophysics: The Sloan Digital Sky survey has continued to use the OSG infrastructure for analysis as needed. The Dark Energy Survey simulation activities are ramping up. Both experiments have been able to achieve their deliverables this part year û issues of availability and robustness are being addressed.

Multi-Disciplinary Sciences: The Rosetta protein-modeling group from the Kuhlman Laboratory, North Carolina now uses the OSG for production simulations (see below). The following non-physics VOs are registered with OSG and are at various stages of contributing to and making use of the distributed facility: CompBioGrid: Virtual Cell biology group from the University of Connecticut; GADU genome analysis database update from Argonne National Laboratory; GLOW multi-disciplinary science from the University of Wisconsin; GRASE: multi-disciplinary science from the University of Buffalo; GROW geographical information systems from University of Iowa; Nanohub nanotechnology (see Engagement report below), SBGrid, structural biology at Harvard University;

Computer Science Research: A collaboration between OSG extensions program, the Condor project, US ATLAS and US CMS is using the

OSG to test new workload and job management scenarios which provide 'just-in-time' scheduling across the OSG sites using 'glide-in' methods to schedule a pilot job locally at a site which then requests user jobs for execution as and when resources are available. This includes use of the 'GLexec' component, which the pilot jobs use to provide the site with the identity of the end user of a scheduled executable.

Findings of the Distributed Infrastructure:

The OSG Facility: The OSG Facility has provided operational, security, troubleshooting, software, integration, and engagement capabilities and support. The usage of the facility varies depending on the needs of the stakeholders and during stable normal operations is providing over 200,000 CPU Wallclock hours a day, of which 30-40% is opportunistic resource sharing. The success rate is difficult to measure and depends as much on the user end-to-end success and failure measurements as on those at the level of the infrastructure. Currently the infrastructure reports success rates at about 85%.

Operations: Operations was able to ensure the continued operation of the OSG facility, provide robust monitoring, registration and documentation services to the community, and improve the services for publishing information to the user organizations and sites. The new 'VORS' service was put into production and upgrades developed in response to the communities' requests. The GOC continues to publish monthly statistics on the number of tickets and their resolution rate.

Security: There were 9 security alerts, which resulted in 3 software patches integrated and deployed on the operating facility. There were no identified security incidents. Several alerts were deemed to not need any software or process modifications. A first active audit of OSG services was performed. The OSG Security Plan was accepted by the Consortium. Working with the Joint Security Policy Group a new policy for Grid Sites defining the expectations and responsibilities was developed.

Troubleshooting: The troubleshooting team has been effective in providing attention to end-to-end problems encountered by the users. They work with many different teams to solve these problems, which can require changes to site configurations and software, software patches and/or development, and also changes to the users application codes and services.

Software: During this past year we have released eleven new versions of the VDTùincluding a production release, which will be used by the next release of the EGEE as well as OSG softwareùand three minor releases of VDT and the OSG software cache to fix security issues in the software. These new releases offered features to enable easier updating of the software, the new OSG accounting service and site 'proxy caching' for the LHC experiments, upgraded versions of >10 middleware packages, added new supported platforms, and fixed numerous bugs. DCache was also integrated into the VDT. The improvement in update and security releases was demonstrated by the time to release of the September update being nearly four weeks, and that in February being three days.

Integration: The integration activity built a Validation testbed of a few sites which provide early testing of new releases of the software stack. The Integration grid provides an at scale system for readying and testing new services and software releases. Integration has helped provision the OSG 0.6.0 software release, as well as provides extensive testing for WS-GRAM, which have resulted in the Globus team developing performance improvements for the OSG, needs.

Engagement: One of our goals in the Open Science Grid (OSG) is to help new user communities benefit from the infrastructure we are putting in place by work closely with these communities over periods of several months to help them derive benefit. During the past year these activities have enabled the Rosetta user community at the Kuhlman Laboratory in North Carolina to smoothly run production opportunistically across more than ten OSG sites. To meet these goals engagement helps in: providing an understanding of how to use the distributed infrastructure; adapting applications to run effectively on OSG sites; engaging the deployment of community owned distributed infrastructures; working with the OSG Facility to ensure the needs of the new community are met; providing common tools and services in support of the engagement communities; and working directly with and in support of the new end users with the goal to have them transition to be full contributing members of the OSG.

Interoperation: OSG interoperates with the EGEE in support of the LHC and other physics VOs. This is now working well based on the correct configuration of the information service. OSG sites must also report the results of site functional tests to the WLCG in support of the LHC infrastructure. WLCG and OSG have worked together on common definitions for the output of such tests, and these are being promulgated to the wider community. The foundation of federation is translation of published information to a format that other grids can use. OSG contributors continue to participate in this activity as part of the Open Grid Forum GLUE work, OSG site functional tests and other information activities.

Training and Development:

Training and outreach to campus organizations, together with targeted engagement activities, are bringing additional users and communities to

use the emerging cyberinfrastructure.

CIDays: OSG collaborates with Educause, Internet2 and TeraGrid to sponsor daylong workshops local to university Campuses? CyberInfrastructure (CI) Days. These workshops bring expertise to the Campus to foster research and teaching faculty, IT facility, and CIO discussions .The first such workshop, held at the University of California Davis in March, had an extremely positive response. At least four more CI days will happen in the fall at the University of New Mexico, Elizabeth College, the University of Arkansas and in collaboration with Clemson University.

Outreach Activities:

Outreach in the US: Workshops at the High Performance and Distributed Computing (HPDC 2007): Joint EGEE and OSG Workshop on Data Handling in Production Grids, and a Grid Monitoring Workshop Presentations to and discussions with NYSGrid, Great Plains Network,, RENCI, and other emerging campus organizations; Participation in the Grid Cookbook project with SURA;

International Outreach: South America: Support for 3 OSG sites in Brazil. Contributions to grid training schools in Brazil, Argentina and Columbia; South Africa: Working with University of Witwatersrand to establish an OSG site as part of the WLCG.

Journal Publications

R. Pordes et al., "The Open Science Grid", To be published in the Proceedings from the 2007 DOE SciDAC PI meeting, Boston 2007., p. , vol. , (2007). Accepted,

A. Aguilar-Arevalo, A. Bazarko, S. Brice, B. Brown, L. Bugel, J. Cao, et al.,, "A Search for Electron Neutrino Appearance at the m2 ~ 1 eV2 Scale.,", Physical Review Letters, p. , vol. , (2007). Submitted,

A. Ramakrishnan, G. Singh, H. Zhao, E. Deelman, R. Sakellariou, K. Vahi, et al.,, "Scheduling Data-Intensive Workflows onto Storage-Constrained Distributed Resources", Seventh IEEE International Symposium on Cluster Computing and the Grid, CCGrid 2007, p. , vol. , (2006). Submitted,

Paul Avery, "Open Science Grid: Building and Sustaining General Cyberinfrastructure Using a Collaborative Approach", First Monday, p. , vol. , (2007). Published, June

Books or Other One-time Publications

Web/Internet Site

URL(s):

www.opensciencegrid.org
www.isgtw.org

Description:

OSG web site for general, publication, and technical information about the project.

OSG sponsored weekly eNewsletter

Other Specific Products

Product Type:

Teaching aids

Product Description:

OSG has developed web based training materials for Grid Schools. These have been reused by other organizations doing training in the field: in particular schools in South America in Columbia, Argentina and Brazil.

Sharing Information:
Shared through being available through the web and personal contacts of Grid school educators and students.

Product Type:
Technical Know-How
Product Description:
OSG is developing an experienced and expert workforce in the operational, management and technical aspects of high throughput production quality distributed infrastructures. This experience includes the use, diagnosis, security and support of distributed computing technologies including Condor, Globus, X509 based security infrastructure, data movement and storage, and other technologies included in the Virtual Data Toolkit.
Sharing Information:
Through engagement and outreach with new site and users of the OSG.

Contributions

Contributions within Discipline:
The OSG has delivered to the science of the physics collaborations who are the major stakeholders and helped to refine and advance the capabilities of distributed computing technologies:

Contributions to Other Disciplines:
In the nine months since the start of the project OSG has provided benefit to:

Protein Modeling and Design: Adaptation and production running opportunistically using more than a hundred thousand CPUhours of the Rosetta application from the Kuhlman Laboratory in North Carolina across more than thirteen OSG sites.

Weather Modeling: Production runs of the Weather Research and Forecast (WRF) application using more than one hundred and fifty thousand CPUhours on the NERSC OSG site at Lawrence Berkeley National Laboratory.

Nanotechnology Simulations: Improvement of the performance of the nanoWire application from the nanoHub project on sites on the OSG and TeraGrid, such that stable running of batches of five hundred jobs across more than five sites is routine.

Molecular Dynamics Simulations: Production running using more than twenty thousand CPU hours of the CHARMM molecular dynamic simulation to the problem of water penetration in staphylococcal nuclease using opportunistically available resources across more than ten OSG sites.

Progress in Campus Grids at Wisconsin and Fermilab, and engagement with the University of California at Davis researchers and campus IT support staff towards a university-based cyberinfrastructure. Triggered exploration of new methodologies for identifying, assessing and resolving security vulnerabilities in multi layered, multi vendor distributed software stacks.

Contributions to Human Resource Development:
The OSG Training program has run 2 grid schools in the US and contributed to the International Summer School for Grid Computing. The schools were held respectively at: the University of Illinois, Chicago, with a cachement area from Alabama to Iowa, and attended by ~10 minority students and faculty; the University of Texas Brownsville, a Minority Service Institution (MSI), and near Stockholm in Sweden. The total number of students (and faculty) trained by these schools this year is around 100. A range of disciplines was represented: Computer Science, Geophysics, Biology, Physics (include high energy physics, IceCube, LIGO), and Space physics.

OSG trains new staff in distributed computing technologies and infrastructure. We have hired and trained more than 7 such staff in the past year.

We are developing web based training material and will be continuing to collaborate with the IceAge organization that runs the International School towards online based training.

Contributions to Resources for Research and Education:
The OSG infrastructure currently provides access to the following resources. It must be remembered that OSG does not own any resources. They are all contributed by the members of the OSG Consortium, and are used both locally and by the owning Virtual Organization. A percentage, that varies between 20 and 40% are in general available for opportunistic use by OSG Communities who are not the direct 'owners'

of the resources:

processing resources on production infrastructure: 60
Grid interfaced data storage resources on prod. infrastructure: 15
Campus Infrastructures interfaced to the OSG: 4
National Grids interoperating with the OSG : 2
processing resources on the Integration infrastructure: 13
Grid interfaced storage resources on integration infrastructure: 2
Cores accessible to the OSG infrastructure: ~30,000
Tape storage accessible to the OSG infrastructure: ~10 Petabytes
Disk storage accessible to the OSG infrastructure: 2 Petabytes

The OSG Virtual Data Toolkit

The OSG Virtual Data Toolkit (VDT) provides the underlying packaging and distribution of the OSG software stack. VDT continues to be the packaging and distribution vehicle for Condor, Globus, myproxy, and common components of the OSG and EGEE software. VDT packaged components are also used by EGEE, the LIGO Data Grid, the Australian Partnership for Advanced Computing and the UK national grid, and the underlying middleware versions and testing infrastructure are shared between OSG and TeraGrid. In the first nine months of the OSG project the VDT has been further extended to include: OSG accounting probes, collectors and a central repository for accounting information, contributed by Fermilab; The EGEE CEMON information manager which converts information from the MDS2 LDIF format to Condor ClassADS; Virtual Organization (VO) management registration software developed for the World Wide LHC Computing Grid (WLCG) and used by most physics collaborations; An additional implementation of storage software, interfaced through the Storage Resource Management (SRM) interface. The dCache software, provided by a collaboration between the DESY laboratory in Hamburg, and Fermilab, is also in use by the WLCG and High Energy Physics experiments in the US.

VDT releases are tested on the OSG Integration Grid before being put in production. VDT is an effective vehicle for the rapid managed dissemination of security patches to the component middleware. Patches and updates are provided to the installation administrators for security and bug fixes.

Contributions Beyond Science and Engineering:

Special Requirements

Special reporting requirements:
OSG has put in place activities that meet the terms of the Cooperative Agreement and Management Plan:

The Joint Oversight Team met to hear about the progress on OSG on February 20th in Washington DC. . OSG is following up on the feedback.

The Science Advisory Group met on June 12th. The presentations are posted as OSG Event-68. The OSG Executive Board is addressing feedback from the advisory panel.

Two intermediate progress reports were submitted to the DOE and the NSF:

OSG Progress Report - February 2007 OSG Document 570
http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=540

Report on OSG Engagement Activity OSG Document 632.
http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=632

Change in Objectives or Scope: None
Unobligated funds: $ 0.00

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Any Book

Contributions: To Any Beyond Science and Engineering

OSG - 625
June 12th 2007
Version 2

Document Name	OSG Engagement Status
Authors	OSG Executive Board

Summary

One of our goals in the Open Science Grid (OSG) is to help new user communities benefit from the infrastructure we are putting in place. The Engagement activity's mission, coordinated by John McGee the Engagement Coordinator, is to work closely with these communities over periods of several months to help them derive this benefit. These activities include: providing an understanding of how to use the distributed infrastructure; adapting applications to run effectively on OSG sites; engaging in the deployment of community owned distributed infrastructures; working with the OSG Facility to ensure the needs of the new community are met; providing common tools and services in support of the engagement communities; and working directly with and in support of the new end users with the goal to have them transition to be full contributing members of the OSG.

In the nine months since the start of the OSG project engagement activities have succeeded in the[1]:

1) Establishment of a supported community infrastructure under which Engagement communities can use the OSG (Mats Rynge, Chris Green).

2) Adaptation and production running opportunistically using more than a hundred thousand CPUhours of the Rosetta application from the Kuhlman Laboratory in North Carolina across more than thirteen OSG sites (Mats Rynge, John McGee).

3) Production runs of the Weather Research and Forecast (WRF) application using more than one hundred and fifty thousand CPUhours on the NERSC OSG site at Lawrence Berkeley National Laboratory (LBNL); (Leesa Brieger, John McGee).

[1] Names in () are those OSG Staff contributing directly to each activity.

4) Improvement of the performance of the nanoWire application from the nanoHub project on sites on the OSG and TeraGrid, such that stable running of batches of five hundred jobs across more than five sites is routine (Jamie Frey, Miron Livny).

5) Production running using more than twenty thousand CPU hours of the CHARMM molecular dynamic simulation to the problem of water penetration in staphylococcal nuclease using opportunistically available resources across more than ten OSG sites (Torre Wenaus).

6) Developing working relationships with additional research groups to fill the pipeline of potential new users. These upcoming production runs include WRF modeling from UC Davis, and two senior researchers in molecular and biochemistry (John McGee).

7) Progress in Campus Grids at Wisconsin and Fermilab, and engagement with the University of California at Davis researchers and campus IT support staff towards a university-based cyberinfrastructure. (many participants).

Preparation and Infrastructure

Initially we spent significant time reaching out to new communities to understand their applications and the benefits that could be gained by access to OSG resources. Some of these have not yet led to a fruitful engagement: As an example, approaches to researchers at the National Renewable Energy Center were productive but it was clear that significant effort is needed to adapt their Windows based applications to run effectively on the OSG Linux based sites. We plan to revisit this in the next year with Clemson and Purdue Universities providing a Windows based resource to OSG.

The OSG engagement and user support teams have established a Virtual Organization Management Service (VOMS) for the Engage VO. We monitor the OSG sites and aim to increase the amount of opportunistically available resources. We use the common Grid Exerciser and other customizable VO tests to probe the applicability and availability of sites to running engagement applications. The OSG operations and troubleshooting activities work with the sites to fix any problems found. We also use the common resource selection and matchmaking services on OSG to improve the success rate in scheduling jobs to run on particular sites, and Engagement activities have led to enhanced requirements and functionality of this system.

Initial documentation for new Engagement users has been written on the OSG@Work Twiki site: https://twiki.grid.iu.edu/twiki/bin/view/Engagement/WebHome. This documentation is now used during the discussions with potential new users, and is being extended based on their feedback.

Rosetta at the Kulhman Laboratory

The Rosetta application was successfully brought to production on OSG in April 2007 and the users are submitting jobs as needed for their research. All cycles on OSG used by Rosetta are opportunistic and to date fifteen sites have contributed to their production runs.

The Rosetta team members working in the Kuhlman Lab (http://www.unc.edu/kuhlmanpg/) are using OSG to help design proteins that adopt specific three-dimensional structures and more ambitiously bind and regulate target proteins important in cell biology and pathogenesis. Rosetta is the molecular modeling program that was originally created in David Baker's laboratory at the University of Washington. Rosetta is

now maintained and developed by over eight groups at universities across the world (http://www.rosettacommons.org/). Designing novel protein structures is computationally demanding because it requires searching both amino acid sequence space and conformational space for low energy sequence-structure pairs. The strategy used by Rosetta to search these multi-dimensional spaces is to perform thousands of independent simulations that sample different regions of structure space. These simulations can be farmed out to large numbers of independent processors, and the OSG is ideal for such needs.

Rosetta has now used more than one hundred thousand CPUHours on opportunistically available OSG resources. The Rosetta researchers themselves have responsive results from the simultaneous submission of a large number of jobs. The engagement team monitors the latency of the jobs and the success rate to ensure that the response remains adequate.

Use of the OSG during April and May 2007 resulted in structure predictions for ten proteins. The table shows the sum of the wallclock hours used.

Site	CPUHours 4/1-5/19
Texas Tech U	22,410
Caltech CMS T2	14,848
USCMS-Fermilab	11,632
Fermilab General Purpose Farm	8,312
Nebraska	7,093
GLOW (Wisconsin)	4,692
U Wisconsin Milwaukee	4,282
Academia Sinica Taiwan	3,914
Vanderbilt University	3,885
FermiGrid	3,660
Fermilab-CDF	1,933
UCSD	1,717
UCalifornia Riverside	1,311
Purdue-RCAC	413
CPUHoursTotal	**90,102**

Table 1: Use of the OSG by the Rosetta Application, April-May 2007

The figure shows a typical use across OSG sites over the final week in this period. The peak on May 17th is due to the scheduling of jobs submitted in the previous 2 days directly to the Milwaukee site.

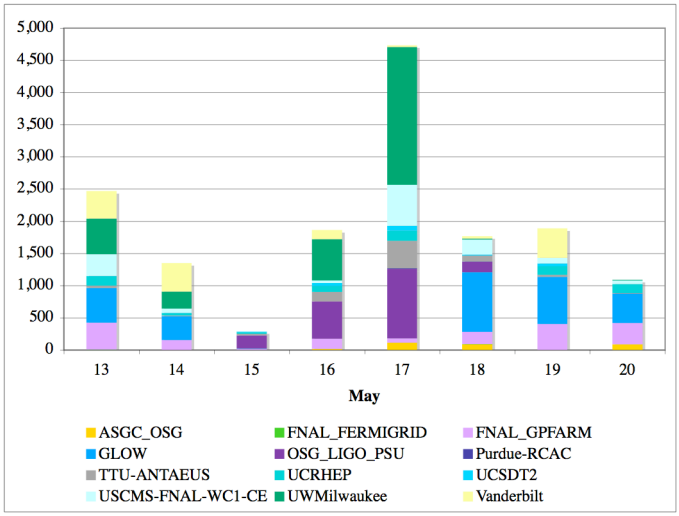


Figure 1: CPUHours/Day for a week of Rosetta Production

The usage pattern shows that once jobs are submitted they run quickly across the grid. We have so far tested the robustness of the system to the submission of up to about 3,000 jobs simultaneously. Experience shows that once a site has been "commissioned" it is fairly stable against errors unless, and until, a scheduled maintenance occurs. A typical profile showing jobs submitted and waiting in the queue (idle) and then running in a few hours, is shown below

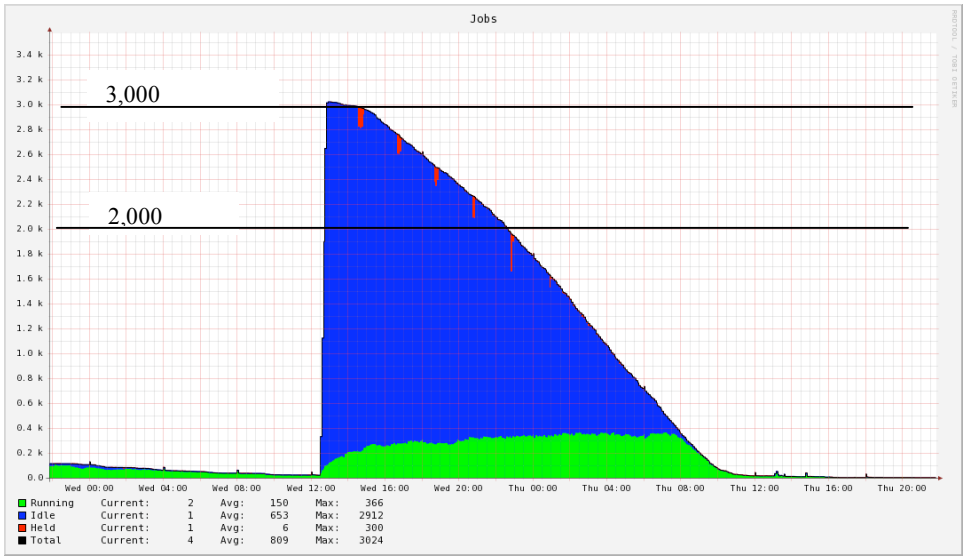


Figure 2: Profile of Rosetta Jobs Submitted and Run

This new community, which continues to run steadily across OSG, is providing our first engagement deliverable of a new science community relying on the infrastructure for production.

Weather Research and Forecast Model

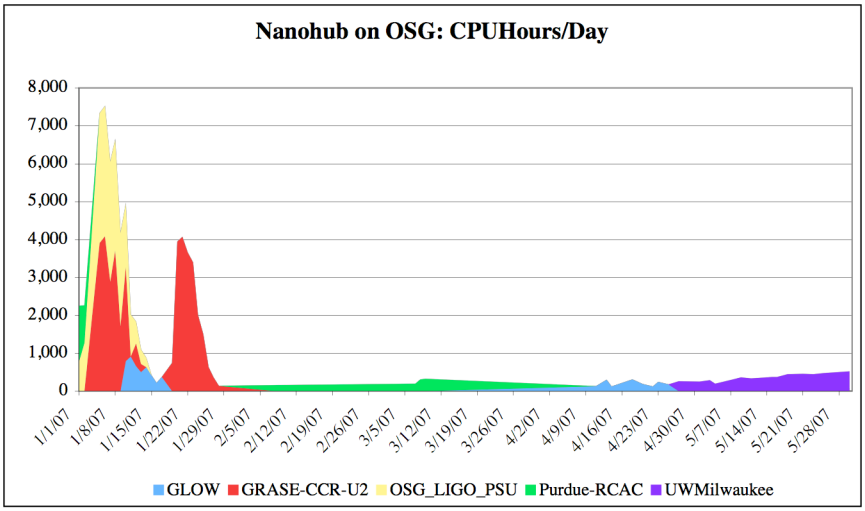
Researchers at RENCI are now running the Weather Research and Forecast (WRF) (http://www.wrf-model.org/index.php) model on a remote OSG site. WRF is a next-generation mesocale numerical weather prediction system designed to serve both operational forecasting and atmospheric

research needs. OSG as an organization has received an overall allocation on the NERSC site at LBNL, which is targeted for MPI applications. WRF is the first MPI application that OSG is supporting and together with NERSC forms an ideal first test case. The initial allocation of 25,000 CPUHours has been completed, and an increase of ~150,000 hours has been approved and consumed in support of a researcher producing results for publication at a conference in July 2007. With this initial success we are now looking to other OSG sites that support MPI applications, e.g. Purdue University, the University of New Mexico, and GLOW, on which we can run WRF jobs and begin testing scheduling and resource selection for such applications.

nanoHUB: BioMoca Application and nanoWire Testing

The nanoHUB wants to use grid resources to give the users of the "hub" access to more computing power than its own cluster of machines at Purdue can offer. The OSG has been a major source of this extra computing time. Work was done in support of nanoHUB scientists to use a OSG resources to run BioMoca simulation jobs last year and the first couple months of this year.

In the past few months, nanoHUB has used OSG resources to run daily tests of their grid software using an application (nanoWire) that simulates the electronic properties of silicon nanowire transistors. They have used OSG machines at UW-Madison, UW-Milwaukee, Fermilab, Vanderbilt and Purdue, and plan to use additional sites in the future. These tests have allowed them to find expose limitations in the grid "backend" of nanoHUB and to make it more resilient to both internal and external failures. They have now transitioned to production mode, running users' jobs on OSG sites. The major user of nanoWire is currently from the University of Cincinnati (http://www.nanohub.org/resources/1307/)



CHARMM

The ATLAS workload management system, PANDA, is being adapted to manage data and jobs for general applications. The first such application is for production runs of the Chemistry at Harvard Molecular Mechanics (CHARMM) program (http://www.charmm.org/) for macromolecular simulations, in this case for the study of water penetration in staphylococcal nuclease (http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=15971206&dopt=Abstract). Two thirds of the production needed by Dr A. Damjanovic has been done by her local staff at Johns Hopkins

University. They have opportunistically used over thirty thousand CPUHours on twelve OSG resources over the last two months.

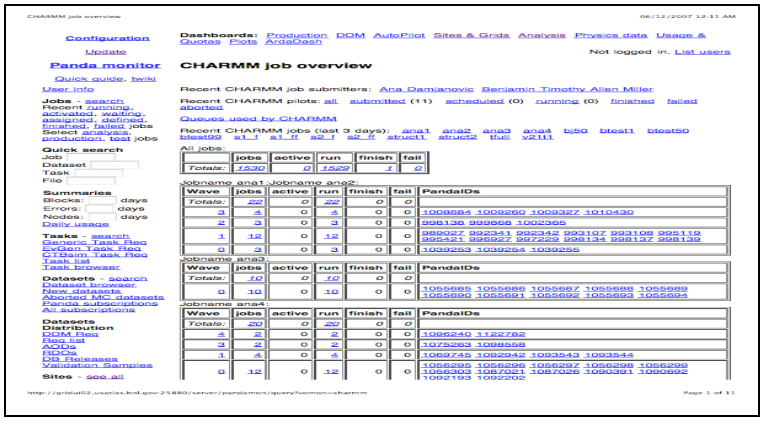


Figure 3: PANDA CHARMM monitoring

Campus Grids

GLOW

The Condor group is working as members of the OSG and the Grid Laboratory at Wisconsin (GLOW) to enable applications to be automatically "elevated" to OSG if no suitable resources are available locally. This is challenging because the security infrastructures are different on the two facilities – GLOW users do not have X509 identity certificates – they use their local kerberos identities. These are mapped by the infrastructure to GLOW group certificates at the boundary between GLOW and OSG. A second challenge is the handling of data and access to storage on the OSG. The goal is for the infrastructure to transparently handle the transport and access to data on OSG from the local data pools on GLOW. Better automated handling of data storage reservation, access and clean-up is a major goal for the OSG Year 2 program of work.

The "football problem" from LeHigh University, was the first such application to be thus "elevated". Since the initial runs in 2006 LeHigh University has become an OSG site and is running the application both locally and remotely. A second activity which is under development is a biology application with the group of Bret Payseur http://www.cs.wisc.edu/condor/PCW2007/presentations/payseur_CondorPayseur.ppt. We are also working with a Ph.D student, Frank DiMaio, and his advisor on restructuring a bio application to run on GLOW and OSG. It involves restructuring the application to be self contained as far as input data goes and is part of the work to understand how best to address the data issues between Campus Grids and OSG.

FermiGrid

The Fermilab Campus Grid has integrated six clusters belonging to different communities at Fermilab. The Campus infrastructure provides a uniform interface to OSG and dispatches jobs to available resources on site. A shared data area allows sharing of data across the local sites.

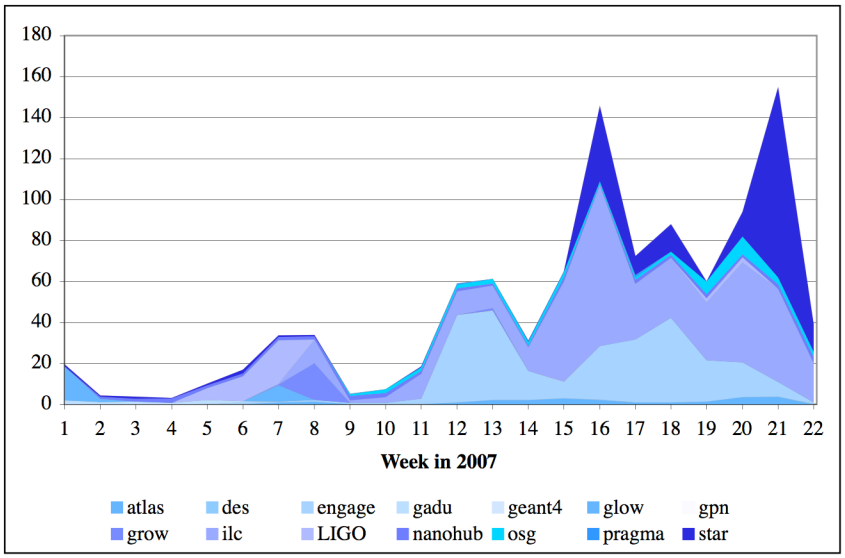


Figure 4: CPUWeeks Usage of FermiGrid by non-Fermilab VOs during 2007

Campus Infrastructure (CI) Days

The OSG Engagement activity is collaborating with Educause, Internet2 and TeraGrid to sponsor day-long workshops local to university Campus' (http://teragrid.org/er/cidays/index.php/Main_Page). The goal of these projects is to bring their expertise to the Campus to foster research and teaching faculty, IT facility, and CIO support and discussions for local campus-wide cyberinfrastructure.

The first such workshop was held at UC Davis in March http://vpiet.ucdavis.edu/cyberinfrastructure.cfm and the response was extremely positive. We are currently following up with two individual researchers whose applications may be ready to take advantage of available distributed resources in the next few months. We have also begun discussions with Rich Wolski at UC Santa Barbara, who is incubating a campus CI by building on top of the grid appliance developed at UCLA. By integrating our activities and sharing technologies with these three campuses, we hope to catalyze a well-integrated OSG enabled regional CI.

The organization of at least four more CI days is going forward for the fall at the University of New Mexico, at Elizabeth College, the University of Arkansas and in collaboration with Clemson University.

Appendix: Engagement Year 1 Program of Work

The text below is from the published Program of Work for Year 1 of OSG. It shows our original plans, which of course were, and continue to be modified through experience and feedback. We use the weekly OSG Executive Team meetings as the forum for discussing what has happened and revisiting plans:

"The program of work for Engagement is laid out in the Year 1 program work thus:
The engagement activity provides focused effort to support additional science disciplines in the use of the OSG and for regional and campus infrastructures to federate with the OSG Facility. The engagement activities include recruitment of users and collaborators, in addition to technical work to help these communities and infrastructures to adapt and interface to the OSG, and provide requirements to the rest of the OSG activities to support their use.

We understand that there are significant sociological and organizational changes that are needed to bring new communities to fully rely on and trust a shared, common distributed infrastructure such as the OSG. To meet these needs, the Engagement activities include meetings, workshops and one on one collaborative efforts working with users and research groups.
The Year 1 activities will include

- Adaptation of some of the components of the RENCI Bioportal workflow to enable access through this interface to OSG resources. This depends on external development of the Bioportal to include security and authorization.
- Collaboration with the National Renewable Energy Laboratory. Three separate applications are being evaluated for integration:

  An "Energy+" application coded in Fortran-90.

  The offshore wind project CCON that has 50k 10 minute simulations for each of three turbine configurations.

  A Center for Transportation Technologies and Systems vehicle simulation. This involves a detailed model, which depends on Matlab and requires solving licensing issues for this application to run on remote OSG sites.
- Researching recent NSF grants that included a non-trivial ($50k to $1M) computational hardware award. This set of investigators is particularly interesting from a perspective of federating these NSF awarded systems with OSG.
- Adapting applications and infrastructure on the GLOW Campus Grid to bring chemistry and nanotechnology simulations to run over OSG.
- Work with the Crimson Grid project to enable the resources to be used by OSG application. Collaboration is progressing more slowly than hopes due to the sociological challenges mentioned above.
- Collaboration with TeraGrid and I2 on a Campus Grid initiative as it evolves.
- Other engagement activities as they are available e.g. with the SNS or GridChem projects.

Deliverables and Milestones

Our goal is to demonstrate use of and dependence on the OSG by one new discipline after each six months of the project:

1.1.6.2.4 Production use of OSG by one additional science community.
This milestone will be met by providing access to the OSG infrastructure and resources to one or more Bioscience user communities.

1.1.6.2.5 Production use of OSG by a 2nd additional science community.
This milestone will be met by the activities with the Renewable Energy Laboratory and the GLOW campus communities.

1.1.6.3.3 – 1.1.6.3.5 Deploy running of GLOW and GROW jobs into production, Enable OSG jobs to run in production on Crimson Grid resources.
These milestones depend significantly on external effort to develop the technologies to enable the smooth flow of jobs and work between the Campus infrastructures and OSG, as well as the availability of active members of the target community to work with us.

Open Science Grid

Open Science Grid

In the first part of 2007 the Open Science Grid (OSG) consortium established and operated the first generation of a shared national cyberinfrastructure bringing together advanced grid technologies to provide end-to-end distributed capabilities to a broad range of compute intensive application. The 0.6.0 OSG software stack released early this year offers the community additional capabilities integrated into the Virtual Data Toolkit. These capabilities were used to deliver opportunistic computing to the DZero Tevatron Experiment at Fermilab, demonstrated their power in the data challenges for the CERN LHC experiments, are being tested at scale by the LIGO gravitational wave and STAR nuclear physics experiments, and provide a basis for the use of OSG by non-physics communities. Training and outreach to campus organizations, together with targetted engagement activities, are bringing additional users and communities to use the emerging cyberinfrastructure.

D0 Reprocessing on Opportunistically Available Resources

During the first half of 2007 DZero reprocessed their complete dataset. Over 50% of the events were processed using opportunistically available resources on the OSG (see Figure 1). This was an important demonstration of the ability of the OSG Consortium stakeholders to contribute resources to the common infrastructure while still maintaining control for their own use. DZero was able to use more than twelve sites on the OSG including the LHC Tier-1s in the US (BNL and Fermilab) and university Tier-2 centers, LIGO and other university sites.

On OSG, DZero sustained execution of over 1000 simultaneous jobs, and overall moved over 70 Terabytes of data. "This is the first major production of real high energy physics data (as opposed to simulations) ever run on OSG resources," said Brad Abbott of the University of Oklahoma, head of the DZero Computing group. Reprocessing was completed in June. Towards the end of the production run the throughput on OSG was more than 5 million events per day—two to three times more than originally planned. In addition to the reprocessing effort, OSG provided 300,000 CPU hours to DZero for one of the most precise measurements to date of the top quark mass, and to achieve this result in time for the spring physics conferences.

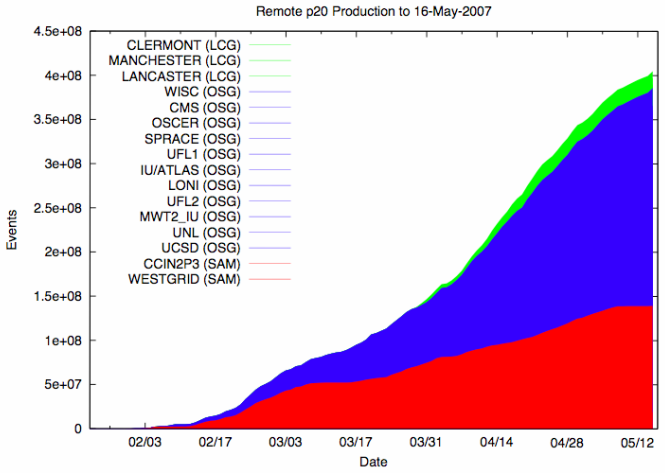


Figure 1: D0 Event Processing on Different Grids

LHC Data Challenges: Simulated Data Distribution and Analysis

As part of the preparations for data acquired from the accelerator at CERN, the ATLAS and CMS experiments organize "data challenges" which test the performance and functionality of their global data distribution and analysis systems. The latest round of activities covered the managed distribution and placement of data around the world, including moving data between storage resources at CERN to more than 10 Tier-1 sites for each experiment worldwide. Movement of data between the Tier-1s and the OSG University Tier-2 sites was also part of these exercises (see Figure 2). The sustained performance is as important, and more difficult to achieve in many cases, as the peak throughput delivered. Each experiment uses the Globus GridFTP protocols to distribute the data, the Enabling Grids for EsciencE (EGEE) gLITE File Transfer Service (FTS) to manage contention for and policies within the network pipes, and an experiment specific data placement service (known as DQ2 for ATLAS and PHEDEX for CMS) to orchestrate the data catalogs, namespaces and management.

End-to-end analysis and production job scheduling and throughput are another important aspect of the exercises, which

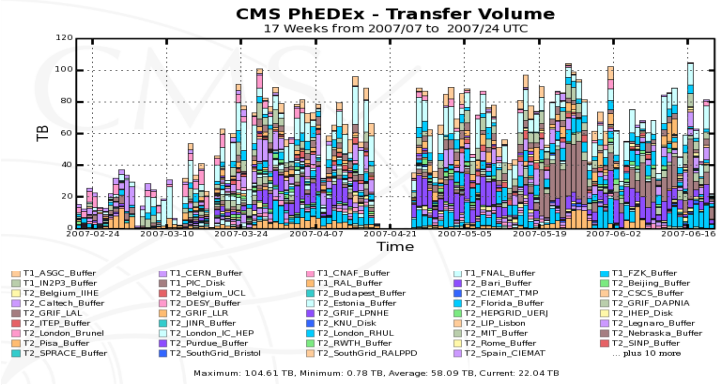


Figure 2: CMS Data Transfer

included support from OSG. Both experiments achieved throughputs of more than 10K jobs a day, with success rates of more than 80%.. For data taking next year all these rates must increase by factors of 2-5. The tests are continuing and the sites and middleware are being scaled to meet the deliverables.

Virtual Data Toolkit Extensions

The OSG Virtual Data Toolkit (VDT) provides the underlying packaging and distribution of the OSG software stack. The distribution was initially built and supported by the Trillium projects – GriPhyN, iVDGL and PPDG. VDT continues to be the packaging and distribution vehicle for Condor, Globus, myproxy, and common components of the OSG and EGEE software. VDT packaged components are also used by EGEE, the LIGO Data Grid, the Australian Partnership for Advanced Computing and the UK national grid, and the underlying middleware versions and testing infrastructure are shared between OSG and TeraGrid. The VDT distribution (See Figure 3) is available as either a set of pacman caches or RPMs, with specific distributions available for making processing farms or storage resources accessible from a Grid infrastructure.

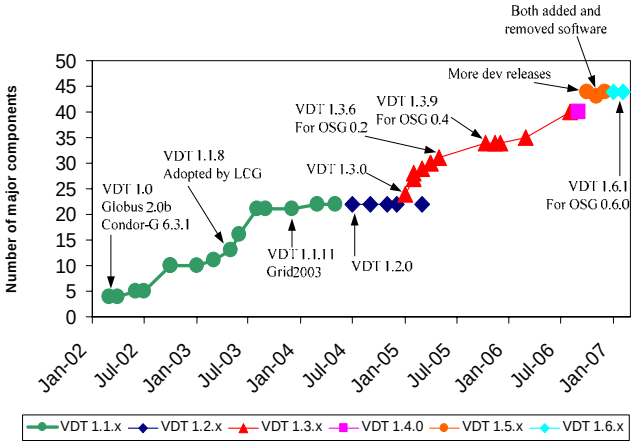


Figure 3: Timeline of VDT Releases

In the first nine months of the OSG project the VDT has been further extended to include: OSG accounting probes, collectors and a central repository for accounting information, contributed by Fermilab; The EGEE CEMON information manager which converts information from the MDS2 LDIF format to Condor ClassADS; Virtual Organization (VO) management registration software developed for the World Wide LHC Computing Grid (WLCG) and used by most physics collaborations; An additional implementation of storage software, interfaced through the Storage Resource Management (SRM) interface. The dCache software, provided by a collaboration between the DESY laboratory in Hamburg, and Fermilab, is also in use by the WLCG and High Energy Physics experiments in the US. VDT releases are tested on the OSG Integration Grid before being put in production. VDT is an effective vehicle for the rapid managed dissemination of security patches to the component middleware. Patches and updates are provided to the installation administrators for security and bug fixes.

Interoperability and Federation

Campus Grids

OSG includes within its scope the support and gateways between campus and the OSG infrastructures:

- FermiGrid: The Fermilab Campus Grid provides a uniform interface to OSG and dispatches jobs to available resources on site. A shared data area allows sharing of data across the local sites.
- Grid Laboratory Of Wisconsin (GLOW): Work continues to enable applications to be automatically “elevated” to OSG—challenging because the security infrastructures are different on the two facilities, and to allow GLOW users to use their existing local kerberos identities. The "football problem" from LeHigh University, is the first such application to be thus "elevated".

OSG collaborates with Educause, Internet2 and TeraGrid to sponsor day-long workshops local to university Campuses—CyberInfrastructure (CI) Days. These workshops bring expertise to the Campus to foster research and teaching faculty, IT facility, and CIO discussions .The first such workshop, held at the University of California Davis in March, had an extremely positive response. At least four more CI days will happen in the fall at the University of New Mexico, Elizabeth College, the University of Arkansas and in collaboration with Clemson University.

Interoperability

OSG interoperates with the EGEE in support of the LHC and other physics VOs. This is now working well based on the correct configuration of the information service. OSG sites must also report the results of site functional tests to the WLCG in support of the LHC infrastructure. WLCG and OSG have worked together on common definitions for the output of such tests, and these are being promulgated to the wider community. The foundation of federation is translation of published information to a format that other grids can use. OSG contributors continue to participate in this activity as part of the Open Grid Forum GLUE work.

Engagement of Non-Physics Communities

The OSG Engagement activity's mission is to work closely with new communities over periods of several months to help them use the production infrastructure and transition to be full contributing members of the OSG.
In the nine months since the start of the OSG project engagement activities have succeeded in:

- Production running opportunistically using more than a hundred thousand CPU hours of the Rosetta application from the Kuhlman Laboratory in North Carolina across more than thirteen OSG sites. Experience shows that once a site has been "commissioned" it is fairly stable against errors unless, and until, a scheduled maintenance occurs, and once jobs are submitted they run quickly across the grid (See Figure 4).
- Production runs of the Weather Research and Forecast (WRF) application using more than one hundred and fifty thousand CPUhours on the NERSC OSG site at Lawrence Berkeley National Laboratory.
- Improvement of the performance of the nanoWire application from the nanoHub project on sites on the OSG and TeraGrid, such that stable running of batches of five hundred jobs across more than five sites is routine.
- Adaptation of the ATLAS workload management system, PANDA, for the Chemistry at Harvard Molecular Mechanics (CHARMM) program for macromolecular simulations, in this case for the study of water penetration in staphylococcal nuclease by Dr A. Damjanovic at Johns Hopkins University who has used over thirty thousand CPUHours on twelve OSG resources over the last few months.

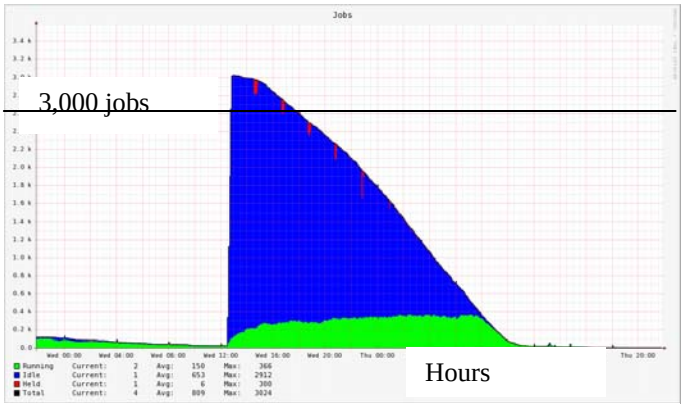


Figure 4:Rosetta Jobs Submitted across the OSG

Grid Schools and the Education Virtual Organization

The heart of the Open Science Grid education training builds on the successful annual grid schools run by the Trillium projects. Each of the OSG grid schools is two to three days of lectures and hands on practicals. Schools have so far been held in at the University of Illinois at Chicago and at the University of Texas Brownsville (UTB, a Minority Serving Institution), with a third school taking place at the beginning of August at the University of Nebraska, Lincoln (an Espcor state). The schools are focused on the graduate level student but at each session several faculty members have signed up with their students. This has resulted in good follow up after the classes as the teachers provide a foundation for continuing to use the material. As well as schools in the US, international organizations—to date in Argentina, Columbia and Brazil— have used the material provided (See Figure 5).

Module #	Time	Lecturer	Topics
			Saturday, March 24
	0800		*Registration and Continental Breakfast*
1	0900	Ruth Pordes, Mike Wilde	Welcome and Introductions
	0915	Mike Wilde	Introduction to Grid Computing (PDF) Network Primer (PDF)
	1000	Dr. Joe Mambretti	Overview of Grid Networks (PDF)
	1030		Intro Lab (1) – 60 min
2	1130	Ben Clifford	Grid Security (PDF) (45 mins)
	1215		*Lunch* – 1 hour
	1315		Security Lab (2) – 75 min
3	1430	Todd Tannebaum	Job Management (PDF)
	1500		Job Management lab (3) – 90 min
4	1630	Bill Allcock	Data management (PDF)
	1700		Data Mgt. Lab (4) – 90 min
	1830	Adjourn	Organize rides to dinner
	1930		Dinner at Maggiano's, Clark and Grand (starts)

Figure 5: Sample of Grid School Curriculum

For the first time OSG has contributed to the annual International Summer School for Grid Computing, which is organized by the National Escience Research Center in Edinburgh. Ten students who have taken the short OSG course were able to spend two weeks immersive hands on training, with close contact with the staff, and in a group of more than 60 students worldwide. The concepts and challenges of distributed computing are taught in tandem with hands on exercises using today's technologies and systems. After attending a school participants can register, with the OSG Virtual Organization and access opportunistically available resources. At the University of Texas Brownsville, for example, students are continuing to work with LIGO on their data analysis.

Acknowledgements

OSG is supported by the Department of Energy Office of Science SciDAC-2 program from the High Energy Physics, Nuclear Physics and Advanced Software and Computing Research programs, and the National Science Foundation Math and Physical Sciences, Office of CyberInfrastructure and Office of International Science and Engineering Directorates.