

Enabling Petascale Science with a Nationally Distributed High-Throughput Computing Fabric of Services – The Open Science Grid

Miron Livny¹, Michael Ernst², Dan Fraser³, Ruth Pordes⁴, Chander Sehgal⁴, Frank Würthwein⁵

¹University of Wisconsin-Madison, ²Brookhaven National Laboratory,
³Argonne National Laboratory, ⁴Fermilab, ⁵University of California San Diego
Email: miron@cs.wisc.edu

Abstract: We give a summary of the current state of the Open Science Grid, the principles of distributed high-throughput computing that guide its activities, and an overview of directions for the near future.

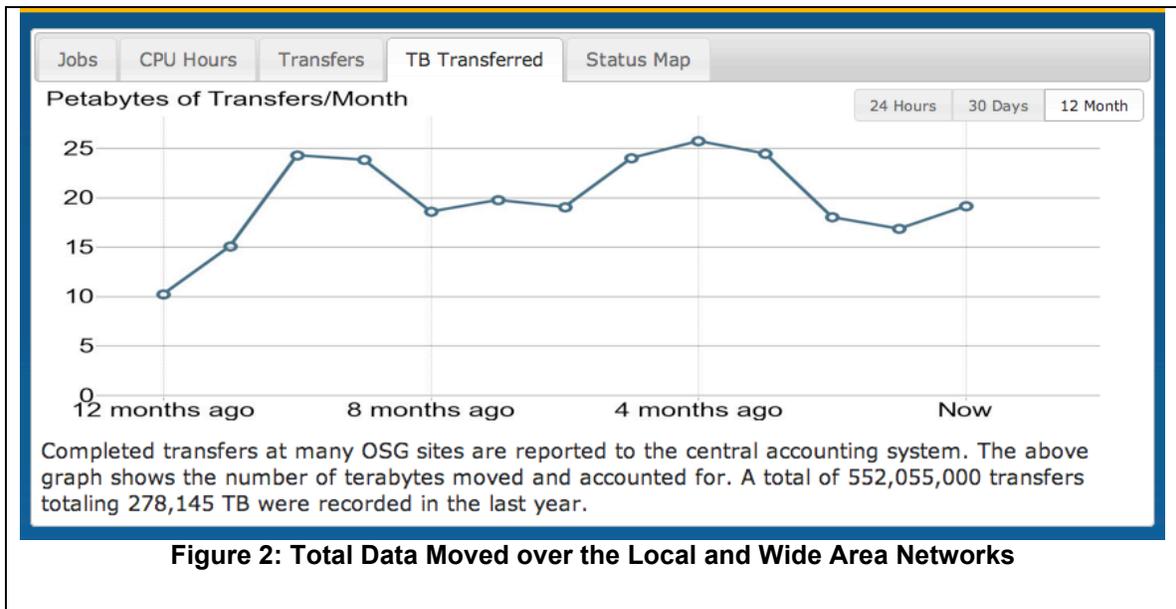
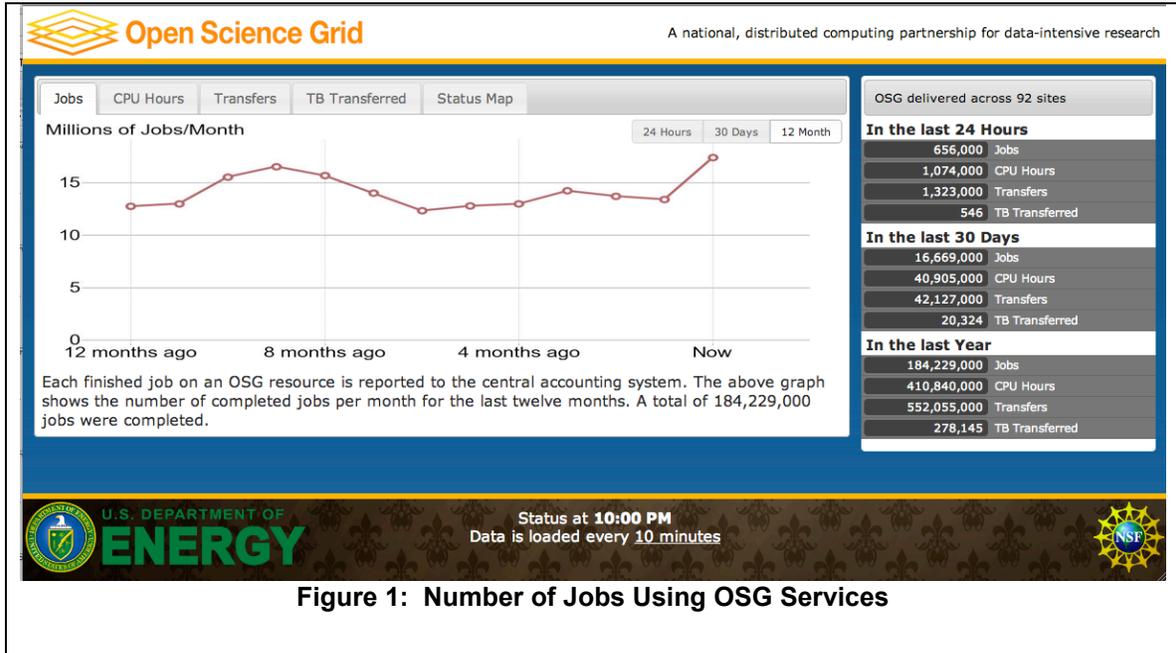
1 Overview

The Open Science Grid (OSG) has evolved into an internationally recognized key element of the US national cyber infrastructure, enabling scientific discovery across a broad range of disciplines. This has been accomplished by a unique partnership that cuts across science disciplines, technical expertise, and institutions. Building on novel software and shared hardware capabilities, the OSG has been expanding the reach of high-throughput computing (HTC) to a growing number of communities. The OSG is open in its investments in new communities, new resources, and new services that all form part of the growing computational society.

Through a comprehensive, dependable, and cost-effective suite of distributed high-throughput (job and data) computing services, the OSG underpins the US contribution to the World Wide Large Hadron Collider (LHC)¹ Computing Grid (WLCG).² This global shared computing infrastructure of unprecedented scale and throughput has facilitated a transformation in the delivery of results from the most advanced experimental facility in the world—enabling the public presentation of results days to weeks after the data is acquired rather than the months to years it took in the recent past. Today our stakeholders include the LHC collaborations, the Laser Interferometer Gravitational Wave Observatory (LIGO),³ more than eight other physics experiments, and groups across more than five other science domains. The members of the OSG continue to be united by a commitment to promote the adoption and to advance the state of the art of *distributed* high-throughput computing (DHTC)—shared utilization of autonomous resources where all the elements are optimized for maximizing computational throughput.

On a typical day, our fabric of DHTC services supports the launch of more than half a million jobs (see Figure 1), and the transfer of more than a quarter of a petabyte of data, across more than 60⁴ US universities and Department of Energy (DOE) laboratories (see Figure 2). In 2010, 249 scientific papers were published that depended on direct use of OSG services and software,⁵ many of which are early LHC results, and 10% of which are nonphysics. The number of users of the OSG has risen substantially over the five years, with more than 2,000 end-users accessing the OSG computing resources (see Figure 3). More than 160 students and 80 system administrators have attended technical training

and education, and the number of university resources accessible through the OSG has risen from 40 to over 90.





2 Distributed High-Throughput Computing

We define DHTC to be the shared utilization of autonomous resources toward a common goal, where all the elements are optimized for maximizing computational throughput. Sharing of such resources requires a framework of mutual trust and maximizing throughput requires dependable access to as much processing and storage capacity as possible. The inherent stress between the requirements for both trust and broad collaboration underpins the challenges that the DHTC community faces in developing frameworks and tools that translate the potential of large-scale distributed computing into high throughput capabilities accessible by a diverse group of users ranging from international collaborations to single-PI research teams. The OSG addresses these challenges by following a framework that is based on four underlying principles:

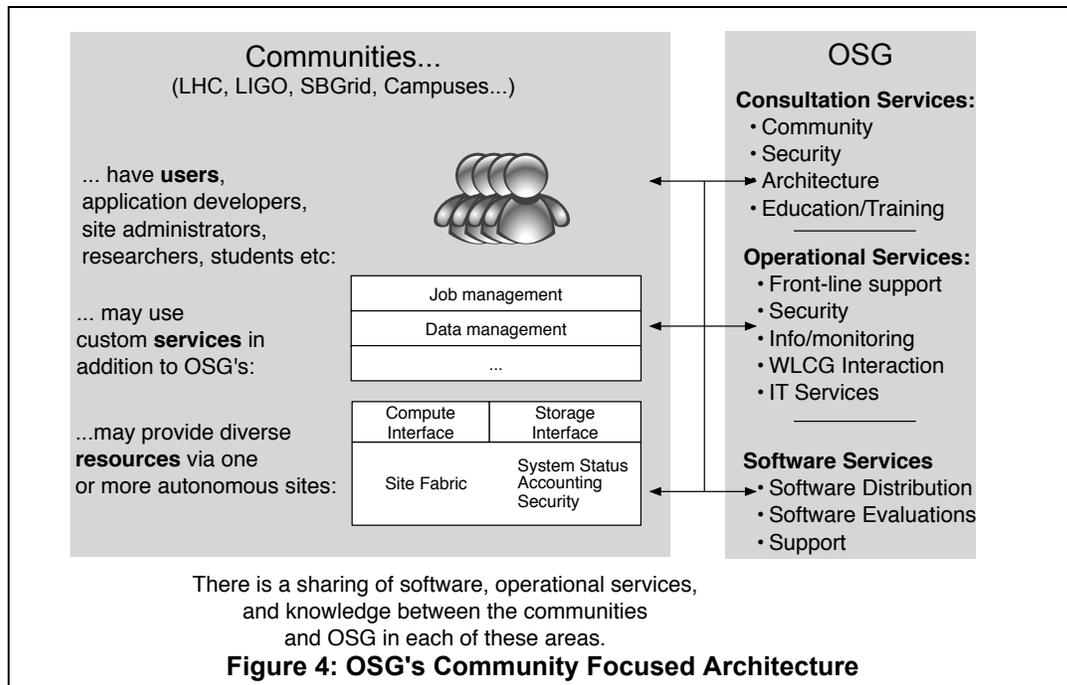
- **Resource Diversity:** Maximizing throughput needs the flexibility to make use of many types of resources and the integration of multiple layers of software and services.
- **Dependability:** Throughput must be tolerant to faults since the scale and distributed nature of high-throughput environments means there will always be unavailable services and malfunctioning resources.
- **Autonomy:** Users and resource providers from different domains and organizations pool and share resources while preserving their local autonomy to set policies and select technologies.
- **Mutual Trust:** The formulation and delivery of a common goal through sharing require a web of trust relationships that crosses the boundaries of organizations and between software services.

Guided by these principles, the OSG maintains a blueprint of the definitions and architecture describing the technologies and methods we deploy.

3 The OSG Architecture and Fabric of Services

Guided by the DHTC principles, the OSG provides a highly available fabric of services that enables sites to accept jobs from many different communities and enables science

communities to harness the power of ten, twenty, some as high as forty sites simultaneously. This fabric of services is composed of three groups—software services (the Virtual Data Toolkit - VDT),⁶ intellectual support services (e.g., education, training, consulting in the best practices of DHTC), and an infrastructure of DHTC services (production) for those who would like to join the OSG DHTC environment (Figure 4). Services in the first two groups serve the broader community that builds and operates their own DHTC environments (e.g., LIGO), as well as supporting the DHTC environment of the OSG.



The OSG both sustains the services we provide (which involves enhancements along the way!) and extends new and innovative services. Doing so requires following careful change control protocols to maintain the dependability and throughput of the OSG DHTC environment, as well as a balance of effort between providing dependable production services and extending them.

OSG Operations services are responsible for a fabric of production and support services. The OSG Operations and Systems Engineering team manages a set of mission-critical discovery and information systems⁷ that underpin the OSG DHTC fabric of services. Even a short interruption in some of these services demonstrably affects the dependability and thus the throughput of the infrastructure. These are highly specialized DHTC services that are operated in failover type environments according to service-level agreements with the stakeholders, as well as subject to change management processes to ensure dependable operation even while services are upgraded.

The DHTC principles—diverse resources, dependability, autonomy and mutual trust—that OSG advances and implements at a national level map well to a campus environment. We are working on technologies that will allow local deployment of high-throughput computing capabilities at the nation's campuses, intra- and intercampus sharing of computing resources and when interfacing to the OSG infrastructure. Our

approach aims to eliminate key barriers to the adoption of HTC technologies by small research groups on our campuses through the following efforts:

- Support for local campus identity management services, removing the need for the researchers to fetch and maintain additional security credentials such as grid certificates.
- An integrated software package that moves beyond current cookbook models that require campus IT teams to download and integrate multiple software components. This package does not require root privileges and thus can be easily installed by a campus researcher.
- Coordinated education, training, and documentation activities and materials that cover the potential best practices and technical details of DHTC technologies.
- A campus job submission point capable of routing jobs to multiple heterogeneous batch scheduling systems (e.g., Condor,⁸ LSF,⁹ PBS¹⁰). Existing campus DHTC models required that all resources be managed by the same scheduler.
- An OSG submission point that can route jobs directly to on-campus resources, hence providing a natural mechanism for existing OSG sites to expand into a campus DHTC infrastructure.

We have begun working with campus communities at Clemson, Nebraska, Notre Dame, Purdue and Wisconsin-Madison on a prototyping effort that includes enabling the formation of local HTC partnerships and dynamic access to shared local (intracampus) and remote (intercampus) resources using campus identities. This is accomplished by locally deploying “pilot factories”¹¹ and by leveraging the “flocking” capabilities of Condor.

The OSG provides three software services: (1) VDT distributions—the end-to-end pipeline that begins with user requirements and results in a deployable software infrastructure, including prioritization, planning, building, testing, configuration, documentation, and packaging to integrate the individual components into a deployable distribution; (2) support—full support for installation and configuration, first-level support for all the software components in the VDT, assistance in determining which component is the problem and provide fixes or workarounds, working with the software provider for second-level support and resolution of bugs; and (3) evaluation—testing of the capabilities of the VDT software components for scalability, reliability, and usability.

4 Science on the OSG

We distinguish three types of beneficiaries of the services and software provided by the OSG: core constituencies, domain science communities, and campus and regional cyber infrastructure. We define as core constituencies those who either are the major resource consumers (LHC experiments) or benefited from “embedded user support” to port their applications (LIGO). Domain science communities range from science collaborations that benefit from the OSG, and also contribute back, to small groups and individuals who use the services and software as they are delivered. A number of scientific collaborations and projects contribute effort in the form of software tools, services, and technical guidance.

In 2010, the LHC experiments produced their first physics publications and discoveries.¹²¹³ Computing has proven to be the enabling technology scientists were hoping for, providing an agile environment for scientific discovery.

LIGO benefits from the OSG software infrastructure to operate the LIGO Data Grid and from OSG services to share LIGO Data Grid computing resources with other communities.¹⁴ LIGO also benefits from opportunistic computing, with the OSG project providing the embedded user support effort to port of LIGO applications to a DHTC environment.

In nuclear physics, we have well-established relationships with the heavy ion physics program at the Relativistic Heavy Ion Collider¹⁵ (RHIC) and the LHC. Apart from these core constituencies, a number of domain science communities from physics, biology, chemistry, mathematics, medicine, computer science, and engineering have benefited from the DHTC services and software provided by the OSG.¹⁶ Among them is the structural biology community (SBGrid)¹⁷ centered at Harvard Medical School, a collaboration of more than 140 x-ray crystallography, NMR, and electron microscopy laboratories, including groups at more than 50 academic institutions in 12 countries. SBGrid is a young community in OSG, consuming a modest 6 million CPU-hours in 2010 across more than 20 sites, leading to their first scientific publication¹⁸ derived solely from opportunistic computing on OSG. Some of the most active communities today preceded the OSG and have adapted technologies from OSG into their computing operations, or even moved their operations onto OSG as their primary HTC environment (CDF, D0¹⁹).

Other communities that have been exploring (with evaluations and/or preproduction or production runs) DHTC technologies for their applications include DES,²⁰ GlueX,²¹ IceCube,²² SCEC,²³ NEES,²⁴ and LSST.²⁵

We have found that bringing HTC capabilities to new communities beyond the above national-scale collaborations is most effective and sustainable via campus and regional affiliation. The original model for campus-based HTC preceding OSG was the Grid Laboratory of Wisconsin,²⁶ followed by FermiGrid and NYSGrid, showing that the shared HTC capabilities that are part of a national cyber infrastructure can be successfully implemented at universities, national laboratories, and even at the state or regional level. An excellent example of a more recent adoption of HTC at the campus level is the Holland Computing Center (HCC)²⁷ providing the research computational resources for the University of Nebraska system. In 2010, HCC began offering HTC capabilities based on the OSG and was able to harness 11 million CPU-hours on non-HCC resources. Six groups from multiple campuses (largest users are mathematics and biochemistry) use the OSG. In 2010 OSG resources contributed to four publications, three of which would have been impossible without OSG, given their tight schedule.

5 Future Areas of Work

We have learned that it is not trivial to transfer capabilities from large international collaborations to a smaller scientific community. The following are key challenges we see today that we aim to address in the future: (1) The heterogeneity of the resource environment of OSG makes it difficult for smaller communities to operate successfully at a significant scale. We made a breakthrough in this area in 2010 by offering a job submission service that creates community specific overlay batch systems across OSG sites, thus providing meta-cheduler functionality across the entire infrastructure. Today, six different communities share a single instance of this service, providing an economy of scale and centralizing the support across these communities.

We will offer this service as a core feature of OSG to all communities in the next few years. (2) The complexity of the grid certificate-based authorization infrastructure presents a non-negligible barrier of entry for smaller communities. During the past year, we saw multiple promising approaches emerging in different contexts: LIGO is pioneering a more integrated approach that includes federated identity management through Shibboleth,²⁸ and the OSG campus infrastructure group has developed a prototype in which local identity management mechanisms are extended to a regional cross-campus infrastructure spanning campuses in six Midwest cities. (3) While the LHC communities are moving petabytes worldwide, smaller communities find it exceedingly difficult to manage terabytes. Over the past five years we have seen this gap in capability grow rather than shrink. The first step in reversing this trend is the “Any Data, Anytime, Anywhere” initiative,²⁹ a collaboration of OSG, WLCG, U.S. ATLAS, and U.S. CMS computing communities. The initiative aims to reduce the problem from one of moving data around within the OSG fabric to one of getting the data onto the OSG fabric in the first place; once data is anywhere on the fabric, it is accessible remotely from everywhere on the fabric.

We will extend the software services in the following two dimensions: (1) sustainable distribution methodology, supporting standard, community-supported native packaging mechanisms (e.g., RedHat³⁰ RPM) as well as cluster management tools such as ROCKS³¹; and (2) community-focused software distributions, where a single VDT distribution cannot meet the different and sometimes conflicting expectations of some science communities. We are also working to extend the VDT to incorporate key technologies requested by the stakeholders including: campus infrastructures, cloud computing, virtualization, and new identity management solutions.

As demonstrated by a recent Google ExaCycle initiative³² to offer ten scientists with high-throughput computing (HTC) applications 100 million compute-hours and the deployment of a 10K HTC Condor pool on the Amazon Cloud by Genentech to provide 80K hours of computing time for protein analysis jobs,³³ a growing number of scientific domains measure computing productivity in large units such as millions of simulations completed per month or protein structures modeled per week. These newcomers to the world of high-throughput computing join the physics community in an ever-growing demand for computational capacity and for tools to manage and use it effectively. To meet this demand for cost-effective computing capabilities, we must continue to expand and enhance a research computing infrastructure that can operate effectively 24 hours a day and 365 days a year with minimal manual intervention. Doing so will transition the OSG toward an exascale distributed infrastructure capable of transparently tolerating failures, accepting incremental changes, managing resources, and adapting to changing workloads.

While we already collaborate with scientists at seven of the sixteen DOE national laboratories, we see and seek opportunities to establish partnerships with new communities especially in genomics and structural biology.

¹ <http://public.web.cern.ch/public/en/lhc/lhc-en.html>

² <http://lcg.web.cern.ch/lcg/>

³ <http://www.ligo.caltech.edu/>

⁴ <http://display.grid.iu.edu/>

-
- ⁵ [“OSG Annual Report to DOE”](#), Jan. 13, 2010.
- ⁶ Alain Roy et al., Building and testing a production quality grid software distribution for Open Science Grid, J. Physics: Conf. Ser. 180, SciDAC 2009, 012052, 2009. Edited by Horst Simon.
- ⁷ <http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=1010>
- ⁸ <http://www.cs.wisc.edu/condor/>
- ⁹ <http://www.platform.com/workload-management/high-performance-computing>
- ¹⁰ <http://www.pbsworks.com/>
- ¹¹ I. Sfiligoi et al., glideinWMS—a generic pilot-based workload management system, J. Phys.: Conf. Ser. 119, 062044 doi: 10.1088/1742-6596/119/6/062044, 2008.
- ¹² Observation of Long-Range, Near-Side Angular Correlations in Proton-Proton Collisions at the LHC / CMS Collaboration arXiv:1009.4122, J. High Energy Phys. 09 (2010) 091.
- ¹³ Observation of a Centrality-Dependent Dijet Asymmetry in Lead-Lead Collisions at $\sqrt{s(NN)} = 2.76$ TeV with the ATLAS Detector at the LHC. arXiv:1011.6182.
- ¹⁴ B. Allen, Einstein@Home search for periodic gravitational waves in LIGO S4 data. Physical Review D. LIGO Scientific Collaboration, corresponding author, H. J. Pletsch. "Einstein@Home search for periodic gravitational waves in early S5 LIGO data. Physical Review D. Phys. Rev. D 80, 042003 (2009)
- ¹⁵ <http://www.bnl.gov/rhic/>
- ¹⁶ R Pordes (for the Open Science Grid Executive Board), Analysis of the current use, benefit, and value of the Open Science Grid, J. Phys. Conf. Ser. 219, 062024 doi: 10.1088/1742-6596/219/6/062024, 2010.
- ¹⁷ <http://sbgrid.org/>
- ¹⁸ I. Stokes-Rees, P. Sliz, Protein structure determination by exhaustive search of Protein Data Bank derived databases, Proc. Nat'l Academy of Sciences doi:10.1073/pnas.1012095107, 2010.
- ¹⁹ <http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=1010>
- ²⁰ <http://www.darkenergysurvey.org/>
- ²¹ <http://portal.gluex.org/>
- ²² <http://icecube.wisc.edu/>
- ²³ <http://www.scec.org/>
- ²⁴ <https://twiki.grid.iu.edu/bin/view/Engagement/EngageOpenSeesB>
- ²⁵ [LSST Workflow / Orchestration Middleware Meeting](#), September 21-22, 2010.
- ²⁶ <http://www.cs.wisc.edu/condor/glow/index.html>
- ²⁷ <http://hcc.unl.edu/main/index.php>
- ²⁸ <http://shibboleth.internet2.edu/>
- ²⁹ <http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=1025>
- ³⁰ <http://www.redhat.com/>
- ³¹ <http://www.rocksclusters.org/>
- ³² http://research.google.com/university/exacycle_program.html
- ³³ http://www.theregister.co.uk/2011/04/06/cycle_computing_hpc_cloud/