

# DOE Science Grid

## Summary of Progress, Feb., 2002 to Feb. 2003

<b>William Johnston, PI</b>	Lawrence Berkeley National Laboratory	<a href="mailto:WEJohnston@lbl.gov">WEJohnston@lbl.gov</a>
<b>Ray Bair, co-PI</b>	Pacific Northwest National Laboratory	<a href="mailto:RayBair@pnl.gov">RayBair@pnl.gov</a>
<b>Ian Foster, co-PI</b>	Argonne National Laboratory	<a href="mailto:foster@mcs.anl.gov">foster@mcs.anl.gov</a>
<b>Al Geist, co-PI</b>	Oak Ridge National Laboratory	<a href="mailto:gst@ornl.gov">gst@ornl.gov</a>
<b>William Kramer, co-PI</b>	LBL / NERSC	<a href="mailto:WTKramer@lbl.gov">WTKramer@lbl.gov</a>

### Science Grid Engineering Working Group, official site representatives:

**Keith Jackson**, Chair, Lawrence Berkeley National Laboratory

**Tony Genovese**, ESnet

**Mike Helm**, ESnet

**Von Welch**, Argonne National Laboratory

**Steve Chan**, NERSC

**Kasidit Chanchio**, Oak Ridge National Laboratory

**Scott Studham**, Pacific Northwest National Laboratory

### Contents

Introduction .....	2
WBS-b.1 Grid Information Services .....	4
WBS-b.2 Certification Authority (LBL and ESnet) .....	4
WBS-b.3 Deploy Globus and Incorporate Computing Resources .....	5
WBS-b.4 Grid Tertiary Storage .....	6
WBS-b.5 Security Infrastructure (LBL, NERSC, and ANL) .....	7
WBS-b.6 Auditing and Fault Monitoring (ORNL) .....	9
WBS-b.6.1 Job Monitoring Techniques (ORNL) .....	10
WBS-b.6.2 Resource Utilization and Auditing Techniques (ORNL) .....	10
WBS-b.6.6 Fault Notification and Response (ORNL) .....	11
WBS-b.7 User Services (PNNL) .....	11
WBS-b.8 System Services / Science Grid Issues .....	11
b.8.1 Grid Deloplymnet Working Group .....	11
b.8.2 pyGlobus and NetSaint (LBL and NERSC) .....	11
b.8.3 Resource Allocations .....	11
WBS-b.9 Applications .....	12
WBS-c Integration of R&D from other projects .....	14
Longer Term .....	15
Demonstrations at SC2002 .....	15
Collaborations / Liasons .....	16
National Fusion Collaboratory (NFC) .....	16
Particle Physics Data Grid (PPDG) .....	17
Publications .....	17

## Introduction

---

DOE Office of Science laboratories operate a wide range of unique resources, from light sources to supercomputers and petabyte storage systems, that are intended to be used by a large distributed user community. The laboratories' geographically distributed staff are frequently faced with scientific and engineering problems of great complexity, the solution of which requires the creation and effective operation of large multidisciplinary teams. The problems to be addressed are large-scale and highly challenging, often greatly exceeding the limits of traditional computing and information systems approaches.

Recognizing these challenges, DOE's SciDAC program calls for the creation of "a *Collaboratory Software Environment* to enable geographically separated scientists to effectively work together as a team and to facilitate remote access to both facilities and data."

A Collaboratory Software Environment must, by definition, span multiple institutions and link numerous, geographically distributed resources of different types. These characteristics introduce two unique challenges that have, until recently, made the development of collaboratory applications very difficult. First, the diverse resources involved typically feature vendor and site-specific software and management mechanisms; writing code that can deal with the variety of different mechanisms that arise in even a small collaboratory can become prohibitively difficult. Second, the sheer scale of the resources involved, and the fact that they span multiple administrative domains, make such fundamental issues as resource discovery, authentication, and fault detection very difficult—often beyond the abilities of a typical collaboratory application developer.

A new class of distributed infrastructure called Grids<sup>a</sup> address these two challenges directly by (a) defining and deploying standard protocols and services that provide a uniform look and feel for a wide variety of computing and data resources, and (b) providing global services that provide essential resource discovery, authentication, and fault detection capabilities. Together, these capabilities reduce the barriers to the large-scale coordinated sharing and use of distributed resources that are at the heart of collaboratory applications. Grid capabilities are thus fundamental to the efficient construction, management, and use of widely distributed application systems; human collaboration and remote access to, and operation of, scientific and engineering instrumentation systems, and the management and securing of computing and data infrastructure. For these reasons, emerging Grid technologies have been adopted by major collaboratory and supercomputer center access projects worldwide, such as the NSF National Technology Grid [3], NASA's Information Power Grid ([www.ipg.nasa.gov](http://www.ipg.nasa.gov)), the European Union's Data Grid (<http://eu-datagrid.web.cern.ch/eu-datagrid/>), the NSF's Network for Earthquake Engineering Simulation Grid ([www.neesgrid.org](http://www.neesgrid.org)), and the DOE ASCI DISCOM project (<http://www.cs.sandia.gov/discom/>). The Global Grid Forum ([www.gridforum.org](http://www.gridforum.org)) provides an international coordinating and standards definition body, analogous to the Internet's IETF.

---

<sup>a</sup> E.g., see "The Computing and Data Grid Approach: Infrastructure for Distributed Science Applications," William E. Johnston, Lawrence Berkeley National Laboratory and NASA Ames Research Center. Computing and Informatics, To appear. <http://www.itg.lbl.gov/~johnston/Grids/homepage.html#CI2002>

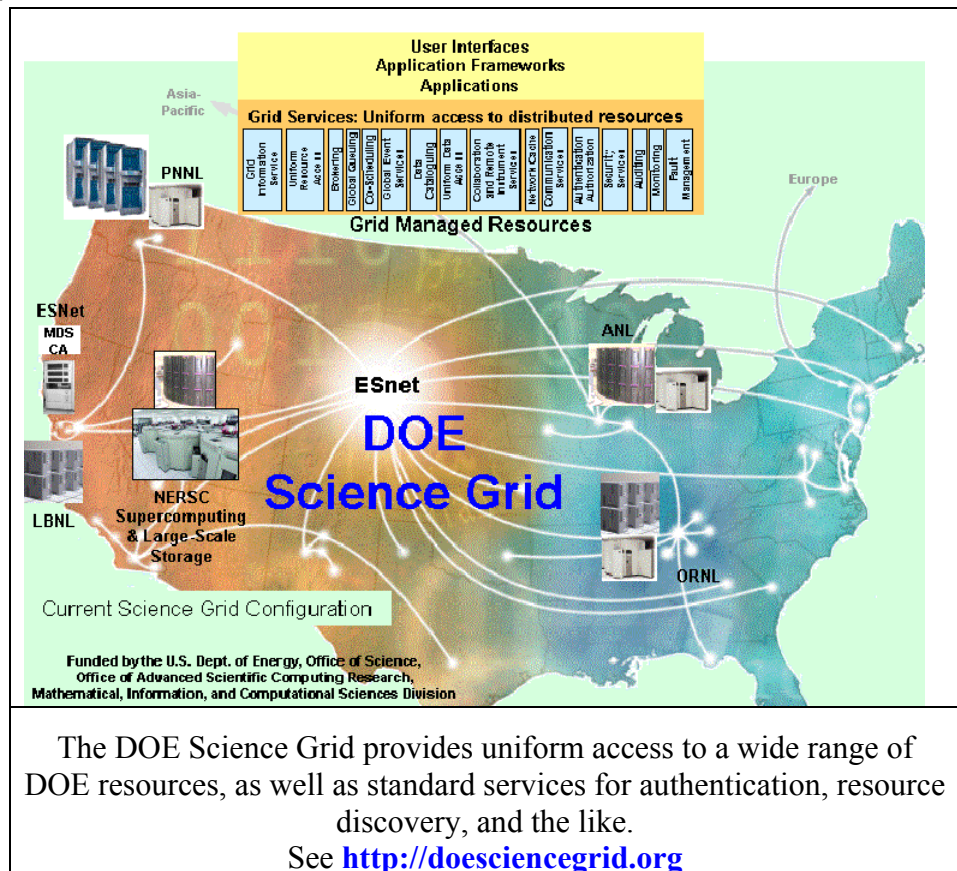
The creation of a DOE-wide distributed computing infrastructure, or *DOE Science Grid* (doesciencegrid.org) will provide this technology to DOE Office of Science programs. The DOE Science Grid that we describe here is designed to reduce or eliminate barriers to the coordinated use of DOE resources, regardless of the physical location of those resources and the users that access them. This infrastructure, as we explain below, will provide a range of new Grid Services addressing issues of resource discovery, security, resource management, instrumentation, and the like. These services will in turn be used to create a range of innovative Grid tools targeting specific application classes. Our goal in creating this infrastructure and tools is to enable innovative approaches to scientific computing based on such concepts as secure remote access to online facilities, distance collaboration, shared petabyte datasets, and large-scale distributed computation; the eventual outcome should be revolutionary changes in a wide range of scientific disciplines across DOE.

The purpose of this proposal, and the DOE Science Grid, is to make Grid technology much more widely appealing and available in DOE.

This project was begun in July 2001, with a kickoff project meeting at the location of the Global Grid Forum in Vienna, VA. At that time we discussed the steps for the first phase of our project, building a multi-site computational and data Grid among the participating institutions, LBNL

(lead), ANL, ORNL, NERSC, and PNNL. We also set up the DOE Science Grid (“DOESG”) Engineering Working Group, with systems staff from each of the institutions. They meet weekly by phone and interact frequently via e-mail, and are key to the success of the project. Funding was received in late August.

Progress is summarized below. (The WBS labels refer to the work plan Gantt chart which may be found at <http://doescienceGrid.org/management/WorkPlan/SciGridWorkplan.2.11.2002.pdf> )



## **WBS-b.1 Grid Information Services (All)**

---

Grid Information Services (GIS) is up and running at each Lab and we have federated these into a larger GIS that allows searches across the combined DOESG resources. We have also constructed several prototype Virtual Organization (VO) based GISs.

The original work plan called for a root GIS operated by ESnet, however current experience is leading to questions about the importance of such a root, and ESnet has not been able to obtain funding for their participation in this. So, at the moment the root GIS is on hold.

## **WBS-b.2 Certification Authority (LBNL and ESnet)**

---

The use of Public Key Infrastructure (“PKI”), including X.509 identity certificates and the Certification Authorities (“CA”) that issue them, is central to our approach for providing a workable authentication infrastructure that promotes large-scale and widely distributed scientific collaborations.

The policies and practices the DOESG CA are specifically designed to reflect the current human practice in the scientific community for authentication and authorization: People are trusted because they are known, because they have credibility in the science community of a collaboration, and because they belong to an established collaboration, etc.

PKI allows the scientific community to extend the reach of its collaborations by using PKI identity certificates and Grid authentication protocols to manage trust in a widely distributed cyber-environment by providing for remote strong authentication of users.

People are authorized to use resources because they have an agreed upon reason to participate in a collaboration. The agreement about who gets to participate in a collaboration is typically made by the lead scientists of the collaboration.

An important goal for CAs supporting science collaboration is to reflect current the trust practice in the scientific community, and not to go beyond the level of formalism needed to implement this practice in ways that are onerous to that community. Onerous and difficult CA policies would cause the science community to seek infrastructure that does not use the Grid security approach, and is less secure than the Grid infrastructure.

We have designed a scalable approach to providing and deploying a production CA at ESnet to support the scientific community. This CA is issuing certificates for DOESG users and other DOE science related users.

The basic approach to scalability is to push trust decisions out to the people who are directly involved in the Virtual Organization / collaboration. A Registration Authority (RA) at each site/VO has the responsibility of verifying the identity and VO membership of the applicants and then authorizing the CA to issue certificates. Each RA is an individual who also becomes part of the Policy Management Authority (PMA), which is the CA’s policy oversight body. The PMA meets periodically, and has approved the current Certificate Policy and Certificate Practices Statement (CP/CPS).

This policy work was vital in establishing a trust relationship between the DOESG CA and the European Data Grid (EDG) CAs. As a result of this we were able to demonstrate the first

interoperable Grids between the major US and European High Energy Physics collaborations. This work has been vital in enabling closer collaboration and resource sharing between US and European High Energy Physics communities.

### ***Inter-Lab X.509 Certificate Policy***

During February the PMA met and approved the version 2.3 CP/CPS. The CA customer base has expanded considerably in the past year to include a number of DOE science projects, and DOESG sites. Current information about the status of the certificate policy work can be found at the DOESG/ESnet CA website (<http://www.doeGrids.org/>).

### ***ESnet Science Grid CA***

A full production version of the CA has been rolled out. This includes a root CA and the signing CA, and their associated infrastructure, e.g., secure racks, locked room, PIX firewalled subnet, production support, etc.

### ***Registration Authorities Established for Each Lab***

The process of validating that users who apply for certificates conform to the Virtual Organization Certificate Policy is handled by Registration Agents within the VO/site. This means that each site/VO must designate an RA in accordance with the ESnet CA policy. (See ESnet's SciDAC PKI & Directory project Web page: [WWW.DOEGrids.org](http://WWW.DOEGrids.org)). Establishing these RAs typically requires approval of the Lab CIO or network security management. We have established production RAs at each of the DOESG sites, and also for a number of other VOs, including the National Fusion Collaboratory, Particle Physics Data Grid, and International Virtual Data Grid Laboratory (iVDGL).

## **WBS-b.3 Deploy Globus and Incorporate Computing Resources (All)**

Remote job submission, data transfer, and information services are deployed and working at all sites. Interoperability between the sites has been established. We are currently finishing upgrading all of our sites to Globus 2.2.x. In the process we are preparing documentation and guidelines for future upgrades.

### ***ANL***

ANL has made a 20-node Linux cluster available to DOESG users, and is working on bringing a large cluster into the DOESG.

### ***LBNL***

LBNL is currently running Globus 2.2 on four multiprocessor Sun, Solaris machines. Globus is rebuilt from the Globus source tree nightly and tested on a test system.

### ***PNNL***

PNNL has installed Globus 2.2 on Phase 1 of a 1,900 processor, 11.4 teraflops Hewlett-Packard Linux cluster with a high-end Quadrics switch. Phase 1, configured at 1 teraflops of Itanium-2

processors, is providing an HP test platform for Globus and Grid applications. Full Globus applications are being run in trial mode both to and from PNNL. Unrestricted access awaits resolution and implementation of site firewall policy issues. (See firewall issues, below.)

### ***ORNL***

ORNL has installed Globus 2.0 toolkit on a back up node of the 64 nodes XTORC Linux cluster at ORNL. To support HRM<sup>a</sup> installation at ORNL, we have applied latest patches of GridFTP on a Solaris machine, which has a connection to ORNL production HPSS.

### ***NERSC***

Globus 2.0 is in production on PDSF<sup>b</sup>, and is being tested on the Seaborg development IBM/AIX system. (Seaborg is the main production supercomputer for NERSC.) Testbed systems running MyProxy, MDS and service monitoring packages have been installed and tested for suitability to run in production. Pre-production testing of Globus 2.2 has begun on the Seaborg development system, as well as on PDSF and individual Grid testbed machines, with a planned production release of Globus 2.2 across NERSC systems in March 2003.

On PDSF, Globus 2.0 has been used successfully as part of an Atlas testbed for job submission and file moving. Currently PDSF uses GSI-ftp for bulk data transfers from BNL.

Standard policies for host based packet filters have been specified and custom Bro firewall extensions have been implemented for operating GSI-ftp in production, and to allow Globus services to operate in firewalled environments.

A production GSIftp gateway to NERSC HPSS is in the final testing stages, and is available for use by “early adopters” and NERSC staff. In addition, basic Globus authorization control is being completed with the internal NIM account management system, allowing users to specify their certificate information using the standard NERSC accounting framework. In the next stage of NIM integration, an automated process for certificate generation and signing will be implemented.

The Alvarez cluster is being readied for Globus testing with PNNL, and the open access visualization server, Escher, is available for Grid applications.

## **WBS-b.4 Grid Tertiary Storage (NERSC and PNNL)**

---

### ***NERSC***

Storage resources are being tested with GridFTP and/or GSIFTP at NERSC. LBNL, in collaboration with the Mass Storage Group at NERSC, has contributed to the design,

---

<sup>a</sup> See the Storage Resource Management (SRM) Middleware Project at <http://sdm.lbl.gov/indexproj.php?ProjectID=SRM>

<sup>b</sup> The PDSF is a networked distributed computing environment used to meet the detector simulation and data analysis requirements of large scale High Energy Physics (HEP) and Nuclear Science (NS) investigations. <http://pdsf.nersc.gov/>

prototyping, and testing of a version of an HPSS tertiary storage system, parallel ftp (pftp) server that supports GSI authentication, and tcp buffer tuning. This prototype is being used by the Earth Systems Grid to move data between the ORNL and NERSC.

NERSC staff are working on a native HPSS, GridHPSS, but based on very preliminary estimates, we do not expect to see “released” GridHPSS until 3Q CY2004. Development versions should be available before then. Overall the HPSS support for Grid software has been slow due to funding limits within the HPSS collaboration.

### ***PNNL***

PNNL is deploying it’s new data archive (13TB Disk based system) Globus has been included as an authentication method.

## **WBS-b.5 Security Infrastructure (LBNL, NERSC, and ANL)**

---

### ***Firewall Issues***

The communication ports used by Globus are subject to local firewall policies, so the DOESG project has been examining the ports used, and how they are used, and has developed a best practices document for firewall administrators. This work has been done in close collaboration with the production security administrators at the DOESG sites to ensure the relevance of this document. The document is available at: <http://www.globus.org/Security/v2.0/firewalls.html>.

As a case study of the issues and approach, the National Fusion Collaboratory ran into immediate problems when they attempted to demonstrate their Grid enabled remote job submission software at the national fusion physics conferences: TTF (U.S. Transport Task Force) Apr 2002, and Sherwood Fusion Theory Meeting, Apr 2002. The client side was behind NAT firewalls at the conference hotels and the server side was behind firewalls at General Atomics and the Princeton Plasma Physics Lab. The DOESG project helped to document in detail exactly which ports were used, how to limit the number of ports and in some cases how to work around the port restrictions. In the course of this a test firewall environment was set up at LBNL to test the assumptions.

The firewall document was presented by DOE SG and NFC members at the ESnet ESCC meeting Jul. 31 - Aug. 1, 2002 where it was part of a discussion between the users of Grid software and the network administrators of the large DOE labs. Hopefully this dialogue has raised the awareness on both sides of the need to be able to securely run Grid applications across firewalls.

To support the open usage of Grid middleware PNNL is actively evaluating dividing the laboratory into two or more enclaves, each with network security policies attuned to the nature of the work and sensitivity of the information.

### ***Authorization***

DOE SG efforts supported the development of Globus job submission software used by the National Fusion Collaboratory in its SC002 demonstration. This demo featured the integration of two existing fusion applications distributed between an exhibit floor machines in the LBNL

and Argonne booths and server machines at the Princeton Plasma Physics Lab and General Atomics, including the use of a new authorization callout in GRAM to allow fine-grained authorization decisions to be done by an Akenti authorization server. A design team consisting of DOE SG and NFC researchers met at SC2002 to design the next iteration of the GRAM common authorization callout. The goal of this new interface is to allow either an authorization service such as Akenti or module that verifies user provided authorization assertions from a CAS or VOMS<sup>a</sup> server to be called to make fine-grained authorization decisions.

***Risk Identification and Mitigation***

Risk identification and mitigation is a critical issue for Supercomputer centers, which are especially concerned about root compromise because of the potentially very long downtime (weeks) needed to rebuild compromised a system. Because the Grid exposes Supercomputers to the Internet in many more ways than they are currently exposed, it is very important to monitor that exposure to make sure that it does not present an unacceptable vulnerability.

NetSaint is a python based system for monitoring remote systems. This monitoring is based on identifying the systems of interest and then periodically probing those systems. The probes are pluggable python modules, and this allows us to build modules that probe all of the Globus service related ports to see which services are available on which systems. A prototype of this is complete and illustrated in the figure for some of the Science Grid systems. Future directions for this will include obtaining the list of resources that need to be monitored by querying the Grid Information Service (MDS), and semantic monitoring. By “semantic monitoring” we mean that not only will the service port be probed, but the monitor will connect up to the point where the service identity certificate can be retrieved and validated at a trusted CA. This could be followed by validating the service protocol if the service authorization is late enough in the connection interaction to allow identifiable protocol exchange. An intermediate step is to have a monitoring admin that has limited authorization on all machines of interest so that the connection can be completed and checked for the proper protocol.

As Grids increase exposure of supercomputers to Internet threats, NERSC and the Science Grid have undertaken a risk and compromise mitigation study to see what can be done in the Grid environment to help counteract the increased exposure.

The following is a brief summary from that study (“Compromise Mitigation for Grid Applications”), which may be found at <http://doescienceGrid.org/Grid/papers> .

***Threat Analysis and Mitigation***

Threat	Countermeasures
Alternate services running on ports assigned to Grid services - i.e., the ports opened for Grid applications, may be used by other applications,	<ul style="list-style-type: none"> <li>• No clear text logins</li> <li>• No anonymous logins</li> <li>• Logging and audit trail for all transactions</li> </ul>

---

<sup>a</sup> “Virtual Organization Membership Service Provides information on the user's relationship with her Virtual Organization: her groups, roles and capabilities.” See <http://hep-project-Grid-scg.web.cern.ch/hep-project-Grid-scg/voms.html>

benign or otherwise	
Vulnerable user code <ul style="list-style-type: none"> <li>Authenticated user privilege escalation</li> <li>Unauthenticated users getting access via jobs or code</li> </ul>	<ul style="list-style-type: none"> <li>Non-executable stack and heap in OS</li> <li>Wrapped libraries and/or sandboxing</li> <li>Build hardened binaries w/Stackguard</li> <li>No path to shell/exec</li> <li>Audit trail</li> <li>No anonymous logins</li> <li>Check for anomalous behavior</li> <li>Educate users about “best practices”</li> </ul>
Vulnerable commodity code (services, libraries) <ul style="list-style-type: none"> <li>httpd, ftpd, SOAP, SSHD, LDAP, SSL, etc...</li> </ul>	<ul style="list-style-type: none"> <li>Non-executable stack and heap in OS</li> <li>Protocol identification and analysis</li> <li>Check for anomalous behavior</li> <li>Controls on staged code origin</li> <li>Build hardened binaries</li> </ul>
Unauthorized / unauthenticated use of Grid accounts	<ul style="list-style-type: none"> <li>No clear text logins</li> <li>No anonymous logins</li> <li>Logging and audit trail for all transactions</li> </ul>

We now have to determine which of these can be addressed in the near-term, which require development, and which are inherent risks.

### **WBS-b.6 Auditing and Fault Monitoring (ORNL)**

In the past year, ORNL’s major accomplishment has been the design and development of prototype software that provides a fault-monitoring system for a computer center or a virtual organization in DOE Science Grid. The framework has demonstrated reliability, efficient fault detection capability, and interoperability with other software in a distributed environment.

There are two research contributions in the current accomplishment. First, new group membership protocol with a practical specification has been developed. The protocol is for an asynchronous distributed system and its specification is primarily focused on preventing a single point of failure and providing a consistent management over the distributed system. The protocol follows a basic concept of leader-follower relationships. The leader is entitled to make decisions for the group. In case the leader fails, the followers will elect a new leader to resume the responsibility. Significant efforts have been made to make sure that the protocol performs correctly in an asynchronous distributed environment.

Second, prototype software has been developed for a fault-monitoring system that consists of three components: the distributed system kernel, a monitoring data table, and a monitoring manager process. The *distributed system kernel (dkernel)* is a core runtime component of the fault monitoring system. The group membership protocol has been incorporated into its. The dkernel consists of daemon processes running on every computer in its system. These daemons collaborate and provide group membership services. To add a computer into the dkernel, the Globus GRAM protocol is used to spawn a daemon process on the new computer.

A *monitoring data (MD) table* has been implemented to store information about jobs, services, or software components to be monitored by the fault monitoring system. A *monitoring manager*

process maintains correct information in the MD table, performs monitoring operations, and interoperates with other software to collect useful data for the fault monitoring system.

At runtime, the dkernel creates an MD table and starts the monitoring manager at a leader computer of the fault-monitoring system. If the leader computer fails, the dkernel will elect a new leader. Then, it will reconstruct a new MD table and restart a monitoring manager automatically on the new machine.

### ***Deployment on Science Grid Resources (ORNL)***

A fault-monitoring system has been developed and deployed on DOE Science Grid testbed at ORNL. The prototype software runs on five heterogeneous computers under Linux, Solaris, and AIX platforms. These computers are also under different system administrations: three belong to CSM, and two belong to CCS division at ORNL. All machines are behind the ORNL firewall.

At SC 2002, the fault-monitoring system was configured to interoperate with tomcat web server and report resource availability information as well as job status to the web server in html format.

#### **WBS-b.6.1 Job Monitoring Techniques (ORNL)**

There are three accomplishments in the area of job monitoring techniques. First, a set of C API's and command line tools have been developed to register jobs into the fault monitoring system and record their information into the MD table. Users can apply these API's to register their jobs, services, or any process to the fault monitoring system. Second, two mechanisms to monitor jobs have been implemented. If a job is running on a computer within the fault-monitoring system, a local dkernel daemon will monitor the job and report its status to the MD table. On the other hand, if users run their jobs on computers outside the fault monitoring system, they need to register the jobs to the monitoring manager, which will monitor the jobs by periodically probing their status and report results to the MD table. Finally, a number of API's and command line tools have been provided to query and manipulate information in the MD table.

#### **WBS-b.6.2 Resource Utilization and Auditing Techniques (ORNL)**

In collecting resource utilization information, one first has to monitor the availability of the resources. If a computer is in the fault-monitoring system, the availability information is captured automatically by the membership protocol. On the other hand, if users want to monitor a computer outside the fault-monitoring system, then the monitor manager can be configured to probe its availability directly and report results to the MD table.

Information about jobs and resources may be retrieved from the MD table and used for auditing or analyzing resource utilization in the DOE Science Grid environment.

The monitoring manager also saves old job execution and resource availability information into a log file, which can be retrieved for future uses.

The fault monitoring system can also collect information about resource and jobs in the Grid environment by retrieving the information from GIS, filtering it, and storing the results in the

MD table. Currently the system works with MDS 2.1, however, the design does not depend on a specific GIS implementation and it can be reconfigured to work with other protocols as well.

The Reliable Grid Monitoring Services (RGMS) has also been implemented. The RGMS is an example application of the group membership services. In it, the MDS 2.1 index daemon associated with the leader computer. When the leader fails, the dkernel is configured to automatically start a new MDS index on a new leader computer. Then, dkernel will reconfigure the rest of the MDS daemon to register with the new index.

#### **WBS-b.6.6 Fault Notification and Response (ORNL)**

Based on the fault-monitoring system, the fault notification system will be developed as a new software component.

#### **WBS-b.7 User Services (PNNL)**

---

A production version of PNNL's online trouble ticket system, ESHQ, has been installed and is in testing on <http://doesg.emsl.pnl.gov>, and a web page has been created to guide users to helpdesk support. ESHQ is a Java application, enabling support staff to query and respond to support queue items from any computer. ESHQ has a graphical user interface and supports a plethora of techniques for database searching. This not only helps the support staff, it also allows users to search for solutions for similar problems.

#### **WBS-b.8 System Services / Science Grid Issues**

---

##### **b.8.1 Grid Deployment Working Group (All)**

The Science Grid Engineering Working Group, chaired by Keith Jackson (LBNL) meets weekly. This group coordinates all community Science Grid activities, resolves cross-site Grid problems, etc.

##### **b.8.2 pyGlobus and NetSaint (LBNL and NERSC)**

The LBNL python wrapped Globus services toolkit ([www-itg.lbl.gov/Grid/projects/pyGlobus/](http://www-itg.lbl.gov/Grid/projects/pyGlobus/)) was originally developed in the DOESG and is being used to build some experimental Grid system administration tools to help support the production usage of Grid resources across the DOESG sites. These include graphical tools for checking the status of Grid resources, managing Globus configuration, adding new Grid users, etc.

A monitoring infrastructure based on the NetSaint framework ([www.netsaint.org](http://www.netsaint.org)) has been deployed. LBNL developed a series of plug-ins built using pyGlobus to support the monitoring of Grid Services, including GRAM, GridFTP, GIS, etc.

##### **b.8.3 Resource Allocations (All)**

The DOESG has been allocated 200,000 MPP hours on the NERSC-3 production system through the regular NERSC resource allocation process – 100,000 hours are being used for a large simulation using the Cactus code to study a high resolution simulation of colliding black holes and the resulting gravitational waves, and 100,000 hours remain to be allocated. The DOE

Science Grid also has an allocation of 2,000 Storage Resource Units on the NERSC production HPSS system, as well as use of the PROBE/HPSS environment.

DOE Science Grid investigators will be able to use the LBNL Alvarez Linux cluster, however the level has yet to be determined. Alvarez is a cluster of two-processor IA-32 nodes running at 866 MHz connected via a Myrinet 2000 switch.

ORNL has reached an initial agreement with the Computer Science and Mathematics Division, which will allow Science Grid users to access (1) a Linux cluster of 64 Pentium 4, 2 GHz computers with Gigabit Ethernet network, currently being tested; (2) a Pentium 4, 1700 MHz computer running Linux 7.1; and (3) a cluster consisting of 4 Pentium II dual processor machines. In an agreement with the ORNL Center for Computational Sciences, users are allowed to access a Sun E250 running Solaris 8 and HPSS on their PROBE testbed. As mentioned earlier, this resource has already been deployed in the ESG project.

Through the user proposal system for the Environmental Molecular Sciences Laboratory, resources may be requested on production and development systems in the Molecular Science Computing Facility. Users can request allocations at <http://mscf.emsl.pnl.gov>. At this time, pilot projects have been set up for initial Grid applications. As the large systems are released for production Grid use, any authorized EMSL user will be able to run Grid applications.

## **WBS-b.9 Applications (All)**

---

The Extensible Computational Chemistry Environment (Ecce) is an interactive problem-solving environment for an extensive set of scalable computational chemistry capabilities in the NWChem and Gaussian code suites. In the first quarter of FY2002, PNNL's Ecce project team (supported by EMSL operations) completed the Linux port of Ecce, which is expected to greatly expand its user community. In collaboration with the DOE Science Grid, the capabilities of the upcoming April release have been upgraded to use version 2.2 of Globus to launch computations and retrieve results. The Ecce team has identified further opportunities to integrate Grid capabilities as Grid services evolve (see: K. Schuchardt, B. Didier, and G. Black, "Ecce—a problem-solving environment's evolution toward Grid services and a Web architecture," *Concurrency and Computation: Pract. Exper*, 14:1221-1239 (2002) – available at <http://aspen.ucs.indiana.edu/gce/>).

Large scale inverse modeling is one of the most computationally demanding analyses performed by PNNL in the characterization of subsurface contaminant behavior. Field scale three-dimensional models of flow and transport at the Hanford site are being calibrated with inverse modeling technology that scales up from coarse-grained parallel processing on clusters to massively parallel processing on DOESG resources. Work in progress includes fine-grained parallelization of the subsurface flow and transport simulator to exploit tightly coupled parallel architectures that provide the required computational efficiency and spatial resolution. Overall the intent is to substantially improve the understanding of subsurface flow and transport at the Hanford Site and predictive uncertainty by enabling a larger range of hydrostratigraphic conceptualizations to be tested.

At LBNL several applications are being prototyped on DOESG compute resources. These include a regional air quality model that has been converted to MPI so as to be able to use the DOESG clusters, and a Supernova Cosmology data analysis application that is using a Grid scheduler.

As a collaboration between Globus project and the Argonne bioinformatics group, the GADU (Genome Analysis and Databases Update tool) has been developed: an automated, high-performance, scalable computational pipeline for data acquisition and analysis of the newly sequenced genomes with DOE Science Grid backend. GADU allows efficient automation of major steps of genome analysis: data acquisition, data analysis by variety of tools and algorithms, as well as data collection, storage and annotation.

During the past decade, the scientific community has witnessed an unprecedented accumulation of gene sequence data and data related to the physiology and biochemistry of organisms. Sequencing of more than 120 genomes has been completed and genomes of 587 organisms are at various levels of completion [<http://wit.integratedgenomics.com/GOLD/>]. However, in order to utilize the enormous scientific value of this data for understanding of biological systems, this information must be integrated, analyzed, graphically displayed and ultimately modeled computationally [Lee Hood 2001]. An emerging systems biology approach requires the development of high-throughput computational environments that integrate (1) large amounts of genomic and experimental data (2) powerful tools and algorithms for knowledge discovery and data mining. However, most of these tools and algorithms are very CPU-intensive and require substantial computational resources that are not always available to the researchers. The large-scale, distributed computational and storage infrastructure of the DOE Science Grid offers an ideal platform for mining such large volumes of biological information.

In collaboration with Globus project an implementation of GADU is being developed that is based on distributed computing technology as a computational backend for genome analysis. This Grid version will use DOE Science Grid resources: PNNL clusters, DataGrid at MCS, ANL, and NERSC.

ANL is working on Grid enabling the Genome Analysis and Databases Update tool (GADU) <http://www-wit.mcs.anl.gov/Alex/GADU/Index.cgi> to use Condor-G and Chimera to submit jobs instead of custom job submission scripts. They are currently using resources at both ANL and PNNL, and hope to BLAST 40 DOE genomes in ~15 hours by the end of February by adding other Science Grid resources.

The goal of the application, in combination with the DOE Science Grid is to address existing bottlenecks

1. Exponential growth of sequence and experimental data
2. Need for the periodic updates and re-analysis of the genomic data -> high CPU, data storage and human time requirements

by:

1. Developing scalable, Grid based system (CPU+storage)
2. Automated (reduces human intervention)
3. Reliable (GADU automated checkers of dataflow + Chimera)

Grid based use of the GADU server for analysis of 40 bacterial genomes by BLAST is projected to take 15 hours of a run time on O(300) processors.

The GADU server will be available for use by the scientific community in May 2003.

## **WBS-c    Integration of R&D From Other Projects**

---

This outline is just intended to indicate the range of things being worked on.

- CoG tools and Web interface
  - o Java CoG based interface to HPSS
  - o Components for domain specific portals.
  - o The use of portals is being discussed at NERSC
- Data replica management
  - o Using SRM with HPSS
  - o ORNL is installing the Replica Location Service (<http://www.globus.org/rls/>)
- STACS tertiary storage management
  - o Get Shreyas to look at this for HPSS Grid interface.
- Workflow management
  - o Fusion collab?
- Grid Monitoring Architecture, Network monitoring
  - o Ganglia (distributed monitoring and execution system <http://ganglia.sourceforge.net>) – GMA integration
- CPU Resource reservation
  - o Not currently used , though there are schedulers on several DOESG systems that can support this
- Certificate based authorization
  - o Akenti and the Fusion Grid
- CAS / Restricted delegation
  - o Von is looking at an application that need this
  - o NERSC is interested in testing this
- Condor-G
  - o Need to install
  - o Add to testing suite.
- GridFTP on MSS
  - o Extensions to GridFTP
  - o HPSS
  - o ADSM
  - o Parallel file system (SAN)
    - PNNL is planning this
  - o GUPFS (NERSC)
    - Steve will check on this

## Longer Term

- HRM
- Brokering
- Execution environment management
  - o Restricted execution environment
  - o Finding software they want
  - o Application meta-data
- CCA
  - o Wait for OGSi integration
- Instrument integration
  - o Mass Spec at PNNL
- WebDAV
  - o Portal related
  - o Ecce
- Dynamic application / Web service aggregation (e.g. IBM WebSphere)
  - o Depends on OGSi
- Generalized data subscription, publication, selection
  - o Depends on OGSi

## Demonstrations at SC2002

---

DOE SG efforts supported the development of Globus job submission software used by the National Fusion Collaboratory in its SC002 demonstration. This demo featured the integration of two existing fusion applications distributed between an exhibit floor machines in the LBNL and Argonne booths and server machines at the Princeton Plasma Physics Lab and General Atomics. Secure Globus I/O was integrated with two of the fusion community's legacy codes that were restructured so that they communicate directly with each other rather than via asynchronous file transfers. This included using a new authorization callout in GRAM that allows fine-grained authorization decisions to be done by an Akenti authorization server. A design team consisting of DOESG and NFC researchers met at SC2002 to design the next iteration of the GRAM common authorization callout. The goal of this new interface is to allow either an authorization service such as Akenti or module that verifies user provided authorization assertions from a CAS or VOMS server to be called to make fine-grained authorization decisions.

This is an important first step towards real-time analysis of data between tokamak pulses.

ANL also demonstrated the usage of CAS in an Earth Systems Grid demonstration showing controlled access to climate data. In addition, ANL demonstrated a prototype implementation of GT3.0.

In another exhibit, PNNL demonstrated an 80 GigaFlop version of their new HP parallel supercomputer running Globus 2.2 on the DOE Science Grid, including 10 Hewlett-Packard RX2600 nodes each with two 1GHz Itanium 2 processors, 4 GB RAM, Quadrics Elan 3 Interconnect, plus two 3-Terabyte Lustre file system servers. PNNL demonstrated a Grid-enabled version of taskr.pl, which parallelizes Ucode, an inverse-modeling code used for groundwater flow modeling used in subsurface transport and environmental remediation studies.

ORNL demonstrated a fault monitoring system at SC 2002. The system consists of a number of daemon processes working together to monitor DOE Science Grid resources at ORNL. We have shown that the system does not have a single point of failure and can efficiently monitor jobs and resources. We have also shown its interoperability with other software including Globus GRAM, MDS, and tomcat web server. This work emphasizes fault tolerance and job monitoring while the NetSaint approach emphasizes monitoring port usage for security purposes.

## **Collaborations / Liaisons**

---

### **National Fusion Collaboratory (NFC)**

The National Fusion Collaboratory (NFC) was an early collaborator with the DOE Science Grid. NCF was one of the founding members of the DOEGrids PKI that was established and is now managed by the DOE Science Grid to support the X509 certificate needs of Grid infrastructure users in the DOE science community. The continuing leadership and management of this Grid CA has relieved the NCF of the responsibility of having to manage its own CA, thereby allowing more resources to be concentrated on fusion specific services.

The DOE Science Grid was used as an early test bed to adapt fusion software to run on a Grid. This need has lessened as the NFC own machines are running the Globus infrastructure, but the DOE SG nodes are still a useful source of machines running the latest versions of the Grid software.

The DOE SG members were instrumental in documenting the issues of running Globus software across firewalls. The NFC ran into immediate problems when they attempted to demonstrate their Grid enabled remote job submission software at the national fusion physics conferences: TTF (U.S. Transport Task Force) Apr 2002, and Sherwood Fusion Theory Meeting, Apr 2002. The client side was behind NAT firewalls at the conference hotels and the server side was behind firewalls at General Atomics and the Princeton Plasma Physics Lab. Members of the DOE SG helped to document in detail exactly which ports were used, how to limit the number of ports and in some cases how to work around the port restrictions. In the course of this, a test firewall environment was set up at LBNL to test the assumptions. This proved to be a big help for the NFC.

This document was presented by DOE SG and NFC members at the ESnet ESCC meeting Jul. 31 - Aug. 1, 2002 where it was part of a discussion between the users of Grid software and the

network administrators of the large DOE labs. Hopefully this dialogue has raised the awareness on both sides of the need to be able to securely run Grid applications across firewalls.

The two projects continue to learn from each other by sharing experience and expertise.

### **Particle Physics Data Grid (PPDG)**

(Report from the PPDG project.)

The Particle Physics Data Grid (PPDG) has an active collaboration with the DOE Science Grid. PPDG was one of the early drivers and a founding member of the DOEGrids PKI that has been established by the DOE Science Grid to support the X509 certificate lifecycle for users and host computers participating in the developing Grid infrastructure for the DOE science community. The leadership provided by DOE Science Grid was and is essential to establishing this critical infrastructure that is now supporting international data Grids for several (8) of the largest high-energy and nuclear physics experiments funded by DOE. As the data Grids in the PPDG community mature and evolve, the interactions with the DOE Science Grid team are extremely helpful on a variety of topics ranging from security/authorization issues, user support questions, monitoring and troubleshooting issues to recent developments of Open Grid Services Architecture. We look forward to a continuing the beneficial collaboration with DOE Science Grid as our prototype-production Grids transition to the robust, ubiquitous services we are counting on.

### **Publications**

---

Please see <http://doesciencegrid.org/Grid/papers/>