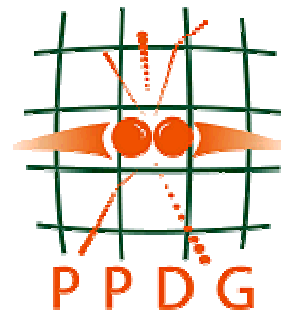


**Particle Physics Data Grid:
From Fabric to Physics**

**Quarterly Status Report of the
Steering Committee,**

July - September 2005

7 November 2005



1 Project Overview	1	6.1.4 D0	10
1.1 Documents and Presentations	1	6.1.5 STAR	11
2 The Common Project	4	6.1.6 ALICE	15
3 CS-11 Data Analysis Working Group	5	6.2 Facilities	15
4 Open Science Grid	5	6.2.1 Jlab	15
5 Collaborations	6	6.2.2 BNL RCF/ACF	15
5.1 EGEE and WLCG	6	6.2.3 FNAL	15
5.1.1 Joint OSG and EGEE Operations Workshop	6	6.2.4 NERSC/PDSF	16
5.1.2 EGEE gLITE	6	6.2.5 SLAC	16
5.2 DOEGrids PKI	7	6.2.6 Collaboration with IEPM, Network Performance Monitoring	16
5.3 TeraGrid	7	6.3 Computer Science & Middleware	18
6 Single Team Reports	7	6.3.1 Condor	18
6.1 Experiments	7	6.3.2 Globus – ANL	18
6.1.1 ATLAS	7	6.3.3 SRM – LBNL	18
6.1.2 BaBar	7	6.3.4 Caltech	20
6.1.3 CMS	7	6.3.5 SRB	20

1 Project Overview

1.1 Documents and Presentations

OSG:

OSG-doc-#	Title	Author	Last Updated
302-v4	Open Science Grid: Beyond the Honeymoon	Dane Skow	25 Oct 2005
193-v1	Edge Services	Frank Wuerthwein	11 Oct 2005
182-v1	Role Based VO Authorization Services	Ian Fisk	11 Oct 2005
205-v1	Accounting System Presentation - Requirements	Sudhir Borra et.	11 Oct 2005

	and Design	<i>al.</i>	
191-v2	Operations Interfaces and Interactions	Rob Quick	11 Oct 2005
190-v1	Storage, Networks and Data Management	Don Petravick	11 Oct 2005
189-v1	Monitoring and Information Services	Mark Green	11 Oct 2005
188-v1	OSG Pre-proposal solicitation	-	11 Oct 2005
187-v2	Rethinking Cybersecurity for Distributed Science	Deb Argawal	11 Oct 2005
204-v1	TAGPMA	Darcey Quesnel	11 Oct 2005
185-v1	Security Update	Bob Cowles	11 Oct 2005
202-v1	Discussion of Community Support Activity	Burt Holzman <i>et. al.</i>	11 Oct 2005
195-v2	Discussion on Partnership of TACC and OSG	Albert Lazzarini <i>et. al.</i>	11 Oct 2005
261-v1	Grid Development and Deployment of LIGO Scientific Collaboration Data Analysis Software	Duncan Brown	11 Oct 2005
256-v1	Virtual Data Toolkit -- Questions Answered	Alain Roy	11 Oct 2005
255-v2	LCG-OSG Interoperability	Shaowen Wang <i>et. al.</i>	11 Oct 2005
254-v1	Ultranet	Don Petravick	11 Oct 2005
260-v1	Sloan Digital Sky Survey - Quasar Spectra	Neha Sharma	11 Oct 2005
253-v2	Enabling LIGO Applications on Scientific Grids	Junwei Cao	11 Oct 2005
257-v1	Grid Application Toolkit	Archit Kulshrestha	11 Oct 2005
252-v1	Installation of the Compute Element software	Rob Quick	11 Oct 2005
251-v1	A Look into Grid Operations Service Infrastructure	Leigh Grundhoefer	11 Oct 2005
259-v2	On demand Monte Carlo Production Service for US CMS	Dave Evans	11 Oct 2005
250-v2	CMS Software Installation on OSG	Bockjoo Kim	11 Oct 2005
248-v1	GUMS and Role Based Authorization at BNL	Gabriele Carcassi	11 Oct 2005
258-v2	Running Virtual Data Workflows on the Open Science Grid	Douglas Scheftner	11 Oct 2005
249-v1	Visual Composition Environment (ViCE) for Grid Workflows	Dave Konerding	11 Oct 2005
181-v1	Interoperability	Laurence Field	11 Oct 2005
200-v2	OSG Interoperability Activity	Fred Luehring	11 Oct 2005
241-v1	Grid Education and Communication	Soma Mukherjee	11 Oct 2005
179-v1	Organizing the Open Science Grid Consortium	Rob Gardner	11 Oct 2005
178-v2	Operating the Open Science Grid	Doug Olson	11 Oct 2005
177-v1	Who Uses the Open Science Grid	Ruth Pordes	11 Oct 2005
173-v1	What is the Open Science Grid	Ian Foster	11 Oct 2005
21-v1	Integration and Certification Activity for the OSG	Deployment	11 Oct 2005
175-v0	Welcome	Vice Chancellor Ourmazed	11 Oct 2005
300-v1	OSG Function Set 0.4 Definition	Dane Skow	07 Oct 2005
299-v1	Grid Storage Management Working Group	Arie Shoshani <i>et. al.</i>	07 Oct 2005
298-v1	MyProxy Integration with PubCookie	Marty Humphrey	05 Oct 2005
296-v1	GridShib: Campus/Grid RBAC Integration	Von Welch	05 Oct 2005
297-v1	Use of Kerberos-Issued Certificates at Fermilab	Dane Skow	05 Oct 2005
295-v1	Purdue Campus Grid	Sebastien	04 Oct 2005

		Goasquen	
288-v1	Multi-Site VOs and Multi-VO Sites in OSG	Abhishek Rana <i>et. al.</i>	03 Oct 2005
294-v1	Grid Computing Applications at the University of Iowa	Shaowen Wang	03 Oct 2005
293-v2	DOSAR	Joel Snow	03 Oct 2005
291-v2	Perspective on Campus Usage and Needs: TTU Experience with Grid Computing	Alan Sill	03 Oct 2005
292-v1	SAMGrid as a Stakeholder of FermiGrid	Valeria Bartsch	03 Oct 2005
289-v1	Grid Laboratory of Wisconsin (GLOW)	Sridhara Dasu	03 Oct 2005
287-v1	Open Science Grid Progress and Status	Doug Olson	28 Sep 2005
285-v1	Edge Services Framework - ESF	Abhishek Rana <i>et. al.</i>	27 Sep 2005
286-v1	EGEE OSG Operations Workshop	Ruth Pordes	27 Sep 2005
282-v1	Towards Storage On-Demand on Petabyte Grids	Abhishek Rana <i>et. al.</i>	13 Sep 2005
283-v1	OSG Storage Day - Agenda	Ruth Pordes	13 Sep 2005
281-v1	OSG-TG Interoperability	Shaowen Wang	08 Sep 2005
280-v1	Opening the Open Science Grid	Dane Skow	06 Sep 2005
174-v1	gPLAZMA: grid-aware PLuggable AuthoriZation MAnagement	Abhishek Rana <i>et. al.</i>	02 Sep 2005
275-v1	Summary and wrapup	Mark Green <i>et. al.</i>	31 Aug 2005
273-v1	Functional goals for OSG Release 0.4	Razvan Popescu	31 Aug 2005
267-v1	Accounting Activity	Philippe Canal <i>et. al.</i>	31 Aug 2005
270-v1	Monitoring and Information Services Interoperability	Shaowen Wang	31 Aug 2005
265-v1	OSG and TeraGrid Interoperability	Shaowen Wang	31 Aug 2005
266-v1	OSG and LCG Interoperability	Oliver Keeble	31 Aug 2005
272-v0	Discussion of homework	Dane Skow	29 Aug 2005
279-v0	Monitoring and Information Services Interoperability	-	29 Aug 2005
278-v0	Clarens Discovery Service	-	29 Aug 2005
277-v0	Monitoring and Information Services Core Infrastructure Activity	-	29 Aug 2005
276-v0	Accounting and Auditing Activity	-	29 Aug 2005
274-v0	Logistics and timetable for OSG Release 0.4	Razvan Popescu	28 Aug 2005
271-v0	Discussion of VO, Operations, Resource Manager Requirements for Monitoring and Information Services	-	28 Aug 2005
269-v0	Clarens Discovery Service	Michael Thomas	28 Aug 2005

268-v0	Monitoring and Information Services Core Infrastructure Activity	Mark Green	28 Aug 2005
264-v0	Discussion of Scope and Definitions of Monitoring and Information Services	Dane Skow	28 Aug 2005
263-v0	Welcome	Mark Green	28 Aug 2005
262-v1	Requirements for Grid Service Interoperability for the Open Science Grid to Access the TeraGrid	Stuart Martin	26 Aug 2005
28-v6	Open Science Grid Deployment Plan - Spring 2005	Deployment	16 Aug 2005
247-v1	OSG Discussions with NSF and DOE	-	10 Aug 2005
130-v2	Accounting System Requirements Version 1.0	Matteo Melani	03 Aug 2005
212-v1	Ribbon Cutting	Frank Wuerthwein	24 Jul 2005
197-v1	Metrics and Goals	Miron Livny	24 Jul 2005
198-v5	Technology Roadmap	Dane Skow	21 Jul 2005
222-v1	Powerpoint Template for OSG	Jorge Rodriguez	20 Jul 2005
201-v0	Discussion on Partnership of SURA and OSG	Paul Avery et. al.	19 Jul 2005
206-v0	Interoperability between the OSG and LCG, Technical Discussions	Fred Luehring	19 Jul 2005
199-v0	Summary and Wrap Up	Scott Koranda	15 Jul 2005
196-v0	Discussion between LONI and OSG	Dick Greenwood et. al.	15 Jul 2005
194-v0	Discussion between TeraGrid and OSG	Charlie Catlett et. al.	15 Jul 2005
172-v0	Introduction	Scott Koranda	12 Jul 2005

PPDG specific reports:

Reports, Documents and Papers		Date/Version
PPDG-49	SRM Accomplishments in 2005	doc (1 Oct. 2005)

2 The Common Project

The common project effort completed its activities on the Provisioning of the initial Release of Open Science Grid with the “Grand Opening” on July 20, 2005. After a brief respite, the activities of the group turned toward defining and producing our deliverables for Release 0.4 targeted for December 2005. This consist of several updates to previous PPDG Common Project contributions and a few new services.

The Privilege Project products, PRIMA and GUMS, have undergone significant documentation improvement cycles and are now running on many OSG production facilities. Effort continues to be expended with the VDT team and INFN VOMS providers to improve the reliability and usability of this software stack. Migration work to develop a PRIMA module for GT4 WS-GRAM authorization framework has begun. This work has been brought “in house” to the Privilege Project as it was determined that the Globus v2 work would not converge on a production ready component in time for Release 0.4.

The Web Service Discovery Service for OSG, based on Clarens, is deployed in OSG Production and is now being used as the main catalog registry system for VOMS servers. Scripts have been developed for registration of Storage Elements (SRM interfaces). Work is beginning on supporting WS-GRAM and other Web Services generally.

Common Project effort has a leadership role in the Accounting Activity and a workplan toward deployment of a prototype service on the timescale of December 2005 has been developed. FNAL and CMS are contributing large fractions of the effort devoted to this activity. An agreement has been made with all of the OSG Partner grids to standardize on the GGF Usage Record format for accounting records. This should facilitate interoperability and possible reuse of components between the various efforts.

The gPLAZMA Authorization framework for storage (a parallel to the PRIMA framework for GRAM gatekeepers) continues in testing for the dCache deployment. A decision on whether and how to integrate this into production dCache support is expected next quarter. Common Projects activity in this work is largely completed and now responding to requests.

The Grid Exerciser (GridEx) has been instantiated on the OSG Integration Grid and an Activity started to address issues raised in Grid3 and with initial tests. Options for inclusion of a variable payload for GridEx (eg. validation jobs rather than null jobs) are being investigated and this may be one method of addressing an OSG-wide opportunistic scheduler for OSG Release 0.4. A decision on whether to include this in the upcoming Release is expected as part of the Deployment Activity in the next Quarter.

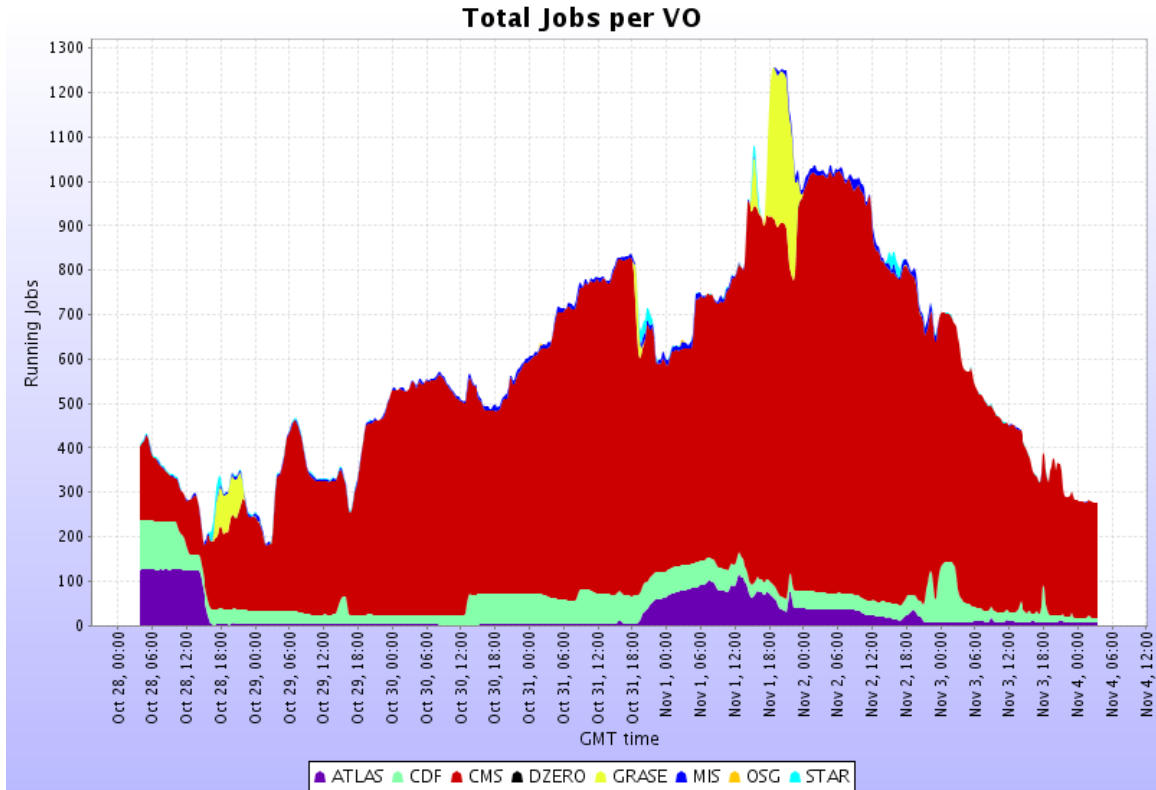
A new OSG Activity, Resource Selection Service, was begun this quarter to communicate work being done by the PPDG Common Project, and solicit other interested parties, on a necessary service to integrate SAMGrid into OSG. This service will initially be supported for and by the DZero participants, but is being developed to be a generalizable service consistent with the gLite and OSG architectures. After the initial evaluation on the ITB and deployment as a DZero VO service (anticipated for Q4 of 2005), stakeholders are expected to evaluate options for its use as a more general component in Spring 2006.

3 CS-11 Data Analysis Working Group

There was some progress in the two areas of activity related to interactive data analysis. The Clarens Discovery Service has been integrated into VDT and is being used for publishing software installations and a job monitoring framework used by CMS. And the ATLAS ADA/DIAL software is being integrated with the new ATLAS distributed production system called PANDA.

4 Open Science Grid

PPDG contributes to make significant contributions to the Open Science Grid in Operations, Support Center organization, Deployment, documentation, and delivering new components through the Common Project. While the use of the OSG was lighter than for Grid3 over the summer months, during the last couple of weeks CMS has made good use of the available CPUs for simulation production and analysis.



Work is ramping up towards provisioning of the next OSG Functional Release set 0.4.0 which will include additional storage management, for storage local to a Farm (Compute Element), job monitoring and site information publishing services, increased use of the Discovery Service, a first deployment of the new Accounting service, as well as next steps towards using the Web Service based GT4 GRAM and MDS4.

5 Collaborations

5.1 EGEE and WLCG

5.1.1 Joint OSG and EGEE Operations Workshop

Several people from OSG (& PPDG) participated in the Joint OSG and EGEE Operations Workshop at the end of September¹. This covered several areas of joint work and mutual interest related to grid & inter-grid operations, inter-grid user support, VO-specific edge services at sites, accounting, maintenance of application environment, and fabric management.

5.1.2 EGEE gLITE

The Resource Selection activity is using the CEMON component from the gLITE middleware. CEMON collects information from the Glue Schema in LDIF format and transforms it to ClassADs to be used in matchmaking. CEMON is a new component in VDT in collaboration with the EGEE gLITE middleware group.

¹ <http://agenda.cern.ch/fullAgenda.php?ida=a054670>

5.2 DOEGrids PKI

PPDG at LBNL has started a short term project with DOEGrids on some usability enhancements to the CA service (user interface scripts & a myproxy service) as well as a pilot project testing token-based one-time-password authentication in the OSG environment. This effort, and it's progress, is described at <http://osg.umd.edu/twiki/bin/view/Security/PpdgScriptsMyproxy>.

5.3 TeraGrid

Discussions began this quarter on developing a more detailed understanding of the partnership between Open Science Grid and TeraGrid. Now that the TeraGrid GIG is officially begun, and the OSG has begun formal operations, the context for these discussions is clearer and we are able to discuss more concrete actions and plans. The immediate activities are focussed on establishing proof of principle interoperability with a few test users while trying to identify most promising science needs which cross the grid boundaries as well as areas of possible synergy for operating and developing the necessary grid infrastructure. There are now biweekly management meetings between OSG and TeraGrid, in addition to the collaborative working activities, carrying these efforts forward.

6 Single Team Reports

6.1 Experiments

6.1.1 ATLAS

Effort report from David Adams of ATLAS on Analysis

I continued development of DIAL. Release 1.30 is expected in early October. Important new features added during the last quarter include:

1. Add persistency to analysis services--jobs are no longer lost when a service is restarted.
2. Log files (actually the full run directory) are now tarred up and stored as logical files and can easily be retrieved for examination by the client.
3. The building of tasks is now carried out like any other job.
4. A web based monitoring system makes it easy for users to examine the status, provenance and output of jobs.

Work was begun on the integration with ATLAS DDM and PANDA.

6.1.2 BaBar

6.1.3 CMS

Effort report from Michael Thomas of CMS on Analysis

Summary

This quarter saw a number of improvements to the JClarens Discovery Service in addition to attendance at a number of conferences.

JClarens

Little development occurred on the jclarens core framework this quarter, which I interpret as a sign that it is maturing nicely. Most work was done on the add-on services, such as the oft-mentioned Discovery Service. This month saw a number of improvements to the Discovery Service that will make it more scalable, less memory intensive, and provide a richer API to make it more useful to the OSG community. Based on the meetings in April with Iosif, I finally got an implementation of the Discovery Service to use a

new ClarensRegistry module in MonALISA. This new architecture will make better use of the MonALISA system and greatly reduce the load on the Jini network. At the same time, the web service API for the Discovery Service was extended to include some additional fields that were mentioned in one or more meetings.

Once the API was extended, the OSG VOMS publication scripts were also modified to make use of these new fields. Both the VOMS publication scripts and the Discovery Service improvements were passed on to the VDT team for inclusion in the next VDT release, 1.3.7.

Tier2 Management

The Caltech Tier2 system administrator took a 4 week vacation in August, leaving me in charge of the 3 Caltech Tier2/OSG clusters. During this time a number of issues came up, mostly related to the GUMS configuration and some system logs filling up the disks. This turned out to be a very useful exercise in debugging and maintaining an OSG installation.

Conferences

ICWS in Orlando

I presented a paper on JClarens at ICWS 2005 in Orlando. A more complete summary of this conference can be found at http://ultralight.caltech.edu/gaewiki/ICWS_2005_in_Orlando

MIS Blueprint meeting

I attended the OSG MIS Blueprint meeting in Buffalo in late August. At the meeting there was a good discussion on the requirements for a "yellow pages"-like registry for the OSG. Not coincidentally, the requirements are almost all fulfilled by the Discovery Service. I also had a discussion with Laura Perlman from ISI about possible integration projects between the Discovery Service and Globus MDS. There are a couple of possibilities for integration, both of which are rather straightforward.

DOSAR workshop

I was invited to the DOSAR workshop in Sao Paulo, Brazil in September. At the workshop I presented a live demonstration of some of the GAE components, including BOSS job submission and MonALISA job monitoring. While I was there I also assisted the grid team at USP with preparations for participating in iGrid later in the month.

iGrid

I participated in a GAE demonstration at the iGrid 2005 conference at UCSD in late September. During the demonstration a physics analysis job was run on a remote site. The analysis was performed using the Monte Carlo Production Service (MCPS) installed in Clarens servers running on the remote sites. From the showfloor, we used a web browser to securely access the MCPS interface at USP. A remote dataset was selected and transferred to the Clarens server as the first stage of the job. The second stage involved running a simple data analysis filter on the transferred dataset. Once the analysis was complete we used the ROOT data visualization tool to view a histogram from the results of the data analysis.

Simulated background data transfer jobs were also run on the various networks link using iperf and bbcp. This background traffic showed how the data analysis can be performed on shared high speed network links.

Education and Outreach

I assisted the team at UERJ in Rio de Janeiro with their migration from the OSG ITB grid to the OSG production grid.

Two new students from Pakistan arrived in July: Ahsan Ikram and Waqas-ur-Rehman. During the 6 month stay Ahsan will be working on a Java-based GUI interface for Clarens while Waqas will be working on The Shell, Execution, and Job Monitoring services for Clarens.

Effort report from Suresh Singh of CMS on Production

For the last couple of months efforts were focused on for the necessary preparation of CMS Service Challenge 3 (SC3) exercises. SC3 which was started in beginning of September runs through December of this year. This has covered various aspects of data transfer and data management. Following tasks have been accomplished to enable Caltech Tier2 site to participate in service challenge.

1. dCache/SRM:

A dCache storage system with SRM interface has been deployed at Caltech production cluster using latest dCache code from dcache.org. A dedicated 3.2 GHz Xeon system has been used for dCache admin as well as door (GridFTP) node. This node has a gigabit connection to Campus backbone network. All the 30 compute nodes which are equipped with second 320 GB hard drives have been use as dCache pool node. This has provided us approximately 9 TB for storage. A replica management system has been implemented on it for the proper balance of files spreading on several pool nodes. The dCache URL is given by <http://pnfs.cacr.caltech.edu:2288>.

2. Phedex:

A latest CMS data transfer/management software (Phedex version 2.2) has been installed on a separate dedicated system. This software is composed of various transfer agents written in Perl, does many functions of CMS data transfer and management. So far 2.2 TB of datasets have been pull from FNAL cache to Caltech dCache system.

3. Adding more Storage:

As per SC3 requirement, a 4.5TB of storage have been added to production cluster using three Dell PowerEdge 1800 servers.

4. PubDB:

In order to publish available datasets, PubDB for SC3 instance has been installed in same server as Phedex. Necessary additional softwares like CMSGLIDE and SHAKAR are also installed on it. Efforts are underway to publish datasets into PubDB.

5: Site Policy:

Caltech Tier2 site policy has been drafted and is given by URL:
<http://www.cacr.caltech.edu/projects/tier2-support/SitePolicy.htm>.

Effort report from Iosif Legrand of CMS on Instrumentation & Monitoring

Based on the MonALISA framework we developed a first prototype for the CMS dashboard (<http://monalisa.cacr.caltech.edu:9090>) that collects information about all the CMS jobs. The monitoring information from the CMS jobs is collected based on the ORCA instrumentation with the ApMon package. The CMS repository is analyzing all the monitoring information from the jobs and creates different aggregate views : total number of jobs running at each site, total number of different type of jobs or the total number of events processed per site sites

Together with the ARDA team at CERN, we are working to enhance the information we collect for each individual job. CRAB, the CMS job submission tool was instrument with ApMon to provide complete information for each job. In this way it is now possible to trace each individual jobs, from the time it was submitted, to know the data sets it is using and to monitor in real time its execution on a computing element and its exit status. The system allows to determine correlation for the jobs that were not successfully finished, and to identify the data sets or the computing nodes were the probability to have such problems is higher.

In collaboration with the ARDA team, we also developed monitoring modules to collect information from XROOTD servers and clients. This is very useful to analyze the IO performed by analysis jobs, to understand the data distribution between different xrootd servers. The Monalisa framework provides the functionality to aggregate this information and to present in real time global views for the analysis activities.

We developed monitoring modules to collect and analyze NetFlow information from routers and to generate aggregate traffic information among different sites. The module is currently used to analyze the traffic between CERN and all the GRID sites in US. and we will use it soon to present the same information for the sites in Europe and Asia

Effort report from Conrad Steenberg of CMS on Analysis

Overview

During the reporting period a major milestone was reached in the development of the Clarens server, namely it's inclusion into the Virtual Data Toolkit (VDT) that is distributed to numerous US Grid sites.

Technical Development

During this quarter a Clarens server and web interface release 0.7.1 was made.

The 0.7.1 release contained updates to the web interface to dramatically improve loading time, better expiry and cleanup of session database caches. The server will also better detect and log corrupt databases, allowing for timely error recovery. new functionality was added to the shell and proxy services to allow the convenient use of proxy certificates in server-side sandboxes by shell commands that need access to proxies. A full changelog is available on the Clarens web page at http://clarens.sourceforge.net/index.php?intro+release_notes/server_0_7_1

Numerous small changes were made to prepare a release that is compatible with the relocatable, user-installable nature of the VDT. These changes are available in both the VDT and the rpm-based releases numbered 0.7.2. Other changes include updates to the web interface to allow parsing of XML return values from web services, improved Safari and Internet Explorer browser compatibility, and the use of a new Discovery Service to publish Clarens service instances. The new Discovery Service allows for much more flexible publication of service parameters, including arbitrary key value pairs.

A successful effort was started to integrate the Clarens server and all it's dependencies in to the Virtual Data Toolkit. This work is included in the VDT 1.3.7 release as a Clarens 0.7.2 package.

The VDT Clarens package serves as the basis for deploying the JobMon service, which was also packaged by me, and included in VDT 1.3.7. Some time was also spent on moving towards a unified service container framework in collaboration with colleagues from the LHCb experiment. No releases of this work have been made.

Collaboration

I attended the OSG Monitoring and Information Systems Technical group in Buffalo, NY from Aug 29 - 30, and subsequently took over chairmanship of the weekly MIS meetings. Although only tangentially related to monitoring, the Clarens framework was proposed as part of a planned system to securely distribute accounting information from OSG sites to requesting clients.

A remote analysis demonstration was given at the iGrid 2005 conference held in San Diego from September 26 - 29. This demonstration was a collaboration between Caltech, UFL, and Fermilab to show analysis of production data from the CMS experiment stored at various locations around the country, all connected via high-bandwidth links.

During the demonstration, data was transferred to the show floor for local analysis, but was also analyzed remotely using the Clarens Monte Carlo Processing Service (MCPS) service and browser interface. The results of these analyses were subsequently visualized using ROOT.

6.1.4 D0

Effort report from Gabriele Garzoglio of D0 on Workload Mgmt & Scheduling

Working on the production of reprocessed data for DZero using the SAM-Grid. The project consist in reprocessing 250 TB of data using 1600 CPU years of computing cycles. As of today, more than 90% of the data has been reprocessed.

Started working on the Resource Selection Service Project (ReSS) to allow the use of OSG resources from the SAM-GRid for DZero. VO with similar requirements on job management will be able to use the service as a general OSG service.

6.1.5 STAR

6.1.5.1 General

During this quarter, we have mainly concentrated on consolidating our support for remote sites which have come into play as part of global Grid deployment. The continuous presence of our Instituto de Fisica da Universidade de São Paulo institution is surely a reassuring sign of the definite interest. Several new institutions have since come forward in this quarter, perhaps raising questions as per the level of support needed for infrastructure in comparison to our own funding the available manpower, i.e. 1 FTE, has almost entirely shifted to remote support:

- São Paulo is looking toward expanding their local setup with an additional 50 CPU and 50 TB in order to support local analysis and development. It appears that the setup will also benefit from enhancements to the cyber-infrastructure and dedicated fibers may be allocated upon resource expansion. All resources are planned to be accessible via an OSG gatekeeper.
- Wayne State University's interest and presence was achievable due to the hard effort of Elizabeth Atems providing now constant local support. Following a similar deployment path than São Paulo, SGE and basic OSG + SGE add-ons were deployed. Additionally, ML monitoring services and local Ganglia information is up and running. Liz did not however complete the registration of the new site in GridCat (will be completed shortly) but all is set to have the new site gatekeeper (`rhic.physics.wayne.edu`) ready to be appearing as the STAR site on the map.
- The University of Birmingham is looking into a local deployment as well but more (for now) in line with the model demonstrated by our previous success from Frankfurt: the initial activity is to achieve a transparent job submission to the offsite resources while bringing the results to local cluster and storage, a currently modest size cluster of 50 CPUs. Prospects are also along the possible use of the recently created regional eScience cluster. Only modest achievements were made on this not directly managed cluster to date (basic grid functionalities were tested at least and were at appropriate quality of service). Perhaps a pairing with the GridPP collaboration (also using SRM tools for example) could make a strong case to get attention and resources needed for a valuable local contribution.

The next quarter will most likely see our Grid then enhanced by two more site.

6.1.5.2 Infrastructure

User Support

Continuous upgrade and maintenance of our grid infrastructure has been provided by Jason Smith (RCF) and by Wayne Betts (STAR). We note that the current arrangement seem to be a better match of our need, the overall stability being provided by multiple gatekeepers and at least the two administrator scheme. We would like to acknowledge the help from several members of the Condor team therein mentioned and extend our thanks for their attention.

We discussed of the difficulties with the local support center scheme (a multi-layer CTS based system requiring diverse tokens and authentication method). It appears we are back to the initial problem: the system is accessible by the Atlas community. This issue being opened since May 23rd 2005, we decided to create a VO specific ticket system which could then forward a copy of the Emails to any RCF later supported systems when it will become ready, available and usable beyond Atlas VO support. We actually see advantages in this approach: a two layer system would also make possible, from a centralized place, to support any of our remote facilities in a convenient manner and we see the requirements for such system to allow for the following:

- each site would be provided with a 'site problem queue'

- each 'site problem queue' requires a site primary contact person whose role is to either determine this problem is not a site specific problem (and therefore redirect to another queue) or direct the ticket to the appropriate site support mailing list
- both site support mailing list and VO service assumes Email support (receive/send) rather than Web-interface (could be either but Email is the minimal requirement) ; since remote ticket system may require diverse authentications, it appears to us that a strong requirement for scalability and viability is for the remote systems to accept Emails not originated by authenticated Emails. Otherwise, each user would have to be granted authentication methods on N systems, N being greater or equal to the number of system (site) in a VO.
- Our VO-based system would provide VO specific support queues such as SRM issues, Scheduler issues. Each of those categories could either have the entire category add an administrator Email (including Condor support, VDT support or similar etc ...) or have the ability to have such allocation be done on a per ticket basis.

This scheme would allow the STAR VO site, local facilities and remote Grid support center and mailing list to interact with our system without the need for individual effort dispersion over all existing lists. This two layer mechanism would also provide independence in terms of readiness and complexity in general of remote site adopted systems.

Condor-G

We noticed regular crashes of our Condor deployment, version 6.6.6 also confirmed in later version (including 6.6.9 and version chosen for the OSG deployment). Wayne Betts, our computer support specialist managing part of Grid infrastructure first noticed this issue, the death of both condor_master and condor_schedd processes at regular timely intervals. We initially monitored the issue and found the crashes were Nessus Scan related. One of the test ("Gain a Shell remotely") issues several buffer overflows crashing the service with no recovery and complete job loss. While a cronjob could restart the service with some penalties, this issue was reported to the Condor team through Email to condor-admin #12749 in the beginning of September. We also found that earlier this year, a similar report from John McCarthy and Jason Smith mentioning the same behavior was filed as VDT support [ticket # 470](#). Alain Roy has provided since a patched version 6.6.11 deployed and under testing. We will report in the next quarter the results and follow-up. We believe this problem being of major importance in a hacker-rich environment we live in.

SGE & Grid issues

We found another problem with SGE in Grid mode: the default for Condor-G for error, input, output I understandably /dev/null. However, SGE jobmanager tries to suppress streams by redirecting to local input/output and error 'files' which are made of the initial file name post-fixed with ".real". Consequently, users or administrator would get lots of confusing messages along the line of 'error: can't open output file "/dev/null.real": Permission denied'. We have assembled a temporary web page with all bundle instructions as per the problems encountered with SGE jobmanager and how to solve them. We will release this page when it will be moved to a permanent Web site but we will provide the information to others upon request. In the meantime, we have been in contact with Tim Cartwright from the Condor project for the best course of action. It was agreed we (STAR) will first attempt to contact the initial (very silent on the issue) developers of the SGE jobmanager and, if in case of no answer, support the patching ourselves. This is a natural move for STAR as our NERSC/PDSF site relies on SGE as its resource management system.

Authentication and Grid job submission

With help from Alain Roy (VDT support), Iwona Sakrejda resolved a long term issue but specific to a few accounts. While globus commands would work from the command line, submission via Condor would fail for "some" users. This was due to a confusion condor has when the ~/.globus/certificates directory exists. This was reported in the VDT support [ticket #782](#).

Other

We would like to mention the continuous help and support provided by Gabriele Carcassi (BNL/RCF) as per our VO support (VOMRS). A recent corruption of records of our repository required live discussion with Tanya Levshin (FNAL). Thanks to both for their continuous support.

MonaLisa server was down a few times at one site or another. This seems to require constant watch (same as GridCat) and a clear reporting mechanism would definitely help our teams to identify the problems as soon as they appear. We tend to notice those problems later than earlier mainly since if jobs are running fine and files are being transferred regularly, we are “happy” (perhaps a side effect of being light weight in terms of component needs). We understand this to be a problem though if/when we open our sites to others and will try to have our basic support center mechanism and ideas in place and consolidate.

We started to reshape our Web pages into a [Drupal](#) based system. Drupal offers the usual Content Management System (CMS) feature, blogs, meeting support and configurable model. We will provide space for our remote sites for site specific information. The evaluation is done pending the setup of a new a more robust Web server to replace our several years main server.

6.1.5.3 SRM and production Grid demonstrator

In the last report, we reported a scheme by which we intend to support analysis scenario in which a user at a given site access Grid resources but need its output to be brought back to the initiator site in an efficient manner but (a) nor necessarily in a non-synchronous manner (no CE/SE coupling and lock-down) and (b) in a way compatible with file registration regardless of the catalog and approach. We described an SRM+RRS scheme well in line with the basic requirements as well as the STAR data transfer model. This work took longer to come to completion due to unexpected authentication problems. Alex Sim and Eric Hjort were the prime investigator of this. At the end, a simplified model where a script handling the copy back (SRMCopy.csh) was put in place and integrated to SUMS. As planned, jobs are submitted from BNL to PDSF and run on worker nodes. After the job's execution is complete and for each file created within that job, the files first moves the file into the local DRM cache using `srms-put` and then initiates a second transfer with `srms-copy` from the DRM at the execution site (PDSF) back to the job submission site (BNL). Testing thus far has shown excellent reliability on a limited scale. The version of DRM (1.2.9) that we need for this use case has been included in VDT 1.3.7. We will have to proceed with a scalability test in the next quarter.

Production data transfer with HRM+RRS during this period totaled 18.5 TB transferred from BNL to NERSC. This corresponds to about 190k files.

6.1.5.4 SUMS

Several reshape of the Star Unified Meta-Scheduler (SUMS) occurred to provide diverse level of support. The modifications were on the improvement, extension and architecture re-structure for better globalization.

First, Pavel Jakl (our summer student from the Czech Republic) completed the first phase of work on the integration of Xrootd with SRM by delivering a fully functional deployment of Xrootd on the RCF STAR analysis farm. This work was nicely and impressively completed, in record time, with code support in SUMS for Xrootd. In Xrootd syntax implemented in SUMS and in conjunction with the changes to Xrootd by Pavel mentioned in the last quarterly, access to local SE is not only much simpler but unified: we can now access either local files and PFN (rootd mode) or files through the redirector using LFN (xrootd mode).

This development also allowed revisiting the rootd approach which was incomplete. Our past implementation strongly bonded SE and CE by submitting the job where the files are locally placed. As a consequence, node containing popular datasets (statically populated) as well as nodes having larger storage would start to see a heavier queue of pending jobs waiting for the computing slot to open to access the files. However, this is unnecessary when (x)rootd is deployed. LSF allows for node prioritization a mechanism by which nodes could be ordered in arbitrary scheduling priorities. The preferred node could be declared as the node which have the most files locally attached, then the second node and so on; our dispatching rule takes into account a weighting proportional to the number of files accessed on of nodes. This implementation is not needed in Xrootd mode since Se and CE are completely decoupled. Levente Hajdu has made this modification.

Levente also modified SUMS configuration model to accommodate for multiple-site and the rapid expansion of our infrastructure. Every time we had to add a site, a new block needed to be added, declaring not only the local resource but its relation (path to submission and result retrieval) to the other resources. The configuration growth would then be geometrical (a major flaw in the initial quick design). The approach was then changed allowing for a more linear growth with the major improvement of having SUMS now determining automatically if jobs should be executed locally or via a Grid submission. In other words, within the same configuration file, N+M jobs could be scheduled as N local and M grid jobs on site A while on site B, containing it would be scheduled as M local and N grid jobs. Later implementation will make the algorithm as part of a dynamic resource discovery (sites would add minimal amount of information and SUMS will immediately take advantage of the available sites).

Because there is a non-standard approach across Resource Management Systems (RMS) as per in which directory jobs are started, and since we have many supported RMS (SGE, LSF, Condor, SGE, PBS and their Grid symmetric) a major change occurred: the startup directory for any dispatcher is now defined to be the scratch directory defined by SUMS. While this change presents a major change in user's job approach, it also gets our user closer to a Grid approach since the abstract request now needs to be shaped without assumptions they were making before. But to ensure backward compatibility, a new tag needs to be introduced to hide possible change of directory users maybe tempted to add verbatim.

Since the RHIC/Phenix have renewed their usage of SUMS for Grid scheduling and exercise, we resumed support for their needs. In fact, all is in place since the implementation of the PBS dispatcher. The sole additional change was the renaming of the previous Grid dispatcher to a generic GRSLDispatcher covering for any job manager.

6.1.5.5 Other activities

As mentioned in the previous section, Xrootd was deployed in our framework (Pavel) and beta testers reported equal stability of the Xrootd comparing to rootd. We however seem to have local issues with authentication, possibly related to LDAP. The symptoms seem to be present in both rootd and Xrootd: the authentication module fails from time to time while retries later would succeed. Not grid related, this shows how local scalabilities could be affected by infrastructure choices and consequently its relation to grid stability.

During the meeting organized at LBNL with Arie Shoshani and Andy Hanushevsky, we agreed of several immediate implementation and interface changes such as the LFN to PFN approach of Xrootd (and make it more universal, plu-I oriented and experiment customizable) and the SRM API / Interface. Alex Sim provided a first draft API for Xrootd=SRM in August while discussion were carried on as per the LFN to PFN API requirements for Xrootd (this would allow to move Pavel's support for rootd/Xrootd simultaneous syntax from hack to optional component library).

PACMAN integration with SUMS is not complete: we discussed of implementation of a sandbox description in the U-JDL and use in the RDL context. We hope to have it fully implemented in the next quarter.

While the SRM+RRS scenario was slower coming than expected, Grid users started to implement their own scheme to bring resulting files back to sites. Non-grid approach (with low bandwidth) such as using scp were used with success (not a single loss) with the same merit reported in the past: Grid and secure copy of results back is more stable than running locally on the RCF resources where the disk space is shared and often filling up to maximum occupancy. With the convenience of having the results 'back home', this has been and continues to be a major drive for users to go toward the Grid approach.

A meeting was held with members of the ROOT team at the end of September related to the GridCollector and the Bitmap-Index developed by the SDM center at LBNL and implemented in STAR. Between Rene Brun, Kesheng Wu, Fons Rademakers, Philippe Canal and Jerome Lauret, the integration of the index bitmap and the ROOT framework was discussed in detail. The Alice team also seems to be interested in the GridCollector technology.

6.1.5.6 Perspectives

Eric Hjort participated to TG-Storage and the SRM collaboration meeting at J-Lab.

Participation in the OSG has continued through the council work, meetings, feedback and in particular participation in the management plan for the OSG (Lothar Bauerdick is the EB main contact for this section, Jerome Lauret provides feedback and counsel as per the model and possible improvements). We sadly expect a dramatic downsize of our community participation especially when it comes to gathering information from the diverse corners of the NP field for an attempt to present through written documents or sections a NP vision and perspective. We will however complete our current opened commitments to OSG and maintain our participation on the OSG council as a STAR VO.

6.1.6 ALICE

6.2 Facilities

6.2.1 Jlab

Effort report from Michael Haddox-Schatz of JLAB on Data Management

In the July-September 2005 timeframe I continued to work on the SRM v.3. This included coming up with an approach that could be used for ensuring forward compatability of SRMs. The prototype that I created used this approach to validate that it could work. I presented this approach along with experiences from creating the prototype at the SRM collaboration meeting that was held at JLab in September. I have also participated in collaborative conversations about SRM v.3 functionality.

6.2.2 BNL RCF/ACF

6.2.3 FNAL

The Fermilab Campus Grid FermiGrid has 3 components in operation:

- FermiGrid Common Grid Services: Supporting common Grid services to aid in the development and deployment of Grid computing infrastructure by the supported experiments at FNAL. These are now in operation hosting the following services: A "site wide" globus gateway; VOMS / VOMRS and GUMS services and Myproxy. A FermiGrid user guide has been written. <http://fermigrid.fnal.gov/user-guide.html>. Metrics are being kept for the basic services. e.g. number of GUMS calls since

29 Aug 2005 22:13:33,797 - 4809
 23 Aug 2005 10:13:40,565 - 66200
 18 Aug 2005 02:43:53,642 - 65612
 13 Aug 2005 09:51:47,763 - 72575
 10 Aug 2005 09:28:00,379 - 83087
 08 Aug 2005 13:13:23,484 - 89124
 05 Aug 2005 13:42:33,542 - 84315
 03 Aug 2005 01:41:02,951 - 92603
 30 Jul 2005 16:25:49,375 - 93497
 27 Jul 2005 17:48:46,486 - 93047
 25 Jul 2005 11:12:31,836 - 94290

- FermiGrid Stakeholder Bilateral Interoperability: Facilitating the shared use of central and experiment controlled computing facilities by supported experiments at FNAL: CDF, D0, CMS, GP Farms.
- FermiGrid Development of OSG Interfaces for Fermilab: Enabling the opportunistic use of FNAL computing resources through Open Science Grid (OSG) interfaces. The following interfaces are enabled: FNAL_GPFARM, FNAL_FERMIGRID, SDSS_TAM, USCMS-FNAL-WC1-CE

6.2.4 NERSC/PDSF

NERSC has installed a gatekeeper node for the OSG production grid providing access to the production PDSF cluster via it's SGE batch system. PDSF is being used as a test platform in hardening the LBNL DRM version of SRM in the VDT and supporting the STAR use cases for running on other OSG sites. The issue of how best to interface between the NERSC user account management and GUMS/VOMS is being investigated, at somewhat low priority.

6.2.5 SLAC

SLAC has installed a gatekeeper on the OSG production grid providing access to the BaBar compute resources (4000+ CPUs), and are allowing access to selected CMS users to assist with the CMS production jobs.

SLAC is also involved in the accounting development work described in the Common Project section.

6.2.6 Collaboration with IEPM, Network Performance Monitoring

This section, including reference links is also posted at <http://www-iepm.slac.stanford.edu/about/status/ppdg-2005-09.html>.

6.2.6.1 Bandwidth/Throughput Monitoring

The DataGrid Wide Area Network Monitoring Infrastructure (DWMI) now has IEPM-BW monitoring successfully installed, making measurements, collecting, analyzing and reporting results at: BNL, Caltech, CERN, FNAL, and SLAC.

We are now using the plateau method, of detecting significant, persistent drops (events) in network performance, in production. It is now used to generate email alerts. Typically we are seeing a couple of alerts/week. These are being carefully reviewed and case studies (see Network Problem Case Studies) are being developed. The results are encouraging, next we need to carefully quantify the success of the method in terms of false positives, missed events etc. We are also working on gathering extra relevant information to report in the alerts.

We are studying a new packet train method pathneck that appears to work better at high speeds than packet pair techniques. We are hoping to use it to gather information on path bottlenecks after detecting an event.

We worked with the author of the achievable TCP throughput tool thrulay to specify required new features. Google funded development of thrulay over the summer so the enhancements have been added. We now need to evaluate the enhancements.

The integration of IEPM-BW into MonALISA to provide improved navigation and visualization has been completed.

6.2.6.2 Passive Monitoring

Passive monitoring provides data from real user applications making real transfers, file to file, for real users, and to real collaborating sites. It adds no extra traffic to the network, does not require us to make reservations or get accounts/passwords/keys/certificates. We are evaluating its effectiveness for providing estimates of achievable throughput (e.g. for grid middleware) by looking at Netflow records at the SLAC border router for large (>1 MByte) flows from the SLAC border router for the last 9 months. Daily there are about 30K of passive Netflow measurements to about 70 sites. Comparisons with the active measurements (where available) show good agreement and aggregating multiple parallel streams is

relatively simple and accurate. From the active measurements 90% of the paths have negligible seasonal variation so the data can be aggregated over long periods. Over a 9 month period, 40% of throughput distributions of the flows between SLAC and a given site are single mode 30% have two modes and 30% have three or more modes. We are evaluating the causes for the multi-modality, e.g. hosts with different network connections, cpu speed, configurations. We are also looking at what to report in terms of percentiles etc.

6.2.6.3 PingER and Developing Region Monitoring

The focus this quarter is on providing better management tools for PingER so we can more easily ensure the data is of high quality. To check that hosts are where we believe they are we are building a tool to make round trip measurements to selected hosts from landmarks (e.g. PingER monitoring sites) so we can triangulate to determine the real position of the host. To support this we put together a secure ping server to be deployed at PingER monitoring sites.

We put together a case study of the fiber outage to Pakistan June 27th to July 8, 2005.

We added a monitoring site in S. Africa, and monitored sites in four African countries, in Manaus Brazil and Israel. We are working with contacts to get sites in Palestine. We validated the data being measured from S. Africa and configured it to measure to a suitable set of sites.

6.2.6.4 Testbeds

The 10Gbps wide area network testbed at Sunnyvale is still in place with a connection to UltraLight.

With Caltech, Manchester, FNAL, CERN and others, once again we are preparing to participate in the SC2005 (in Seattle) BandWidth Challenge (BWC). We have put together a web site to publicize our efforts. Equipment loans have been secured from Sun, Cisco, Boston Computers, QLogic, Neterion, and Chelsio. We have arranged for seven 10 Gbits/s waves to the SLAC/FNAL booth (2 from SLAC, 4 from FNAL and one from the UK). At SLAC we are installing a xrootd cluster of ten Sun v20z dual 1.8GHz Opterons, plus 4 file servers. At SC2005 we will have eight file servers from Boston Computers, a cluster of ten Sun v20z with dual 2.4GHz Opterons, 40Gbits/ fibre channel connection to 20 TBytes in the StorCloud booth at SC2005. We are hoping to win the BWC for the third year in succession.

We have made contact with Microsoft and are working on an MOU to evaluate a new TCP stack on real networks.

6.2.6.5 Admin, visits, papers, presentations, proposals etc.

Article on PingER published in Science Grid this week.

Submitted proposal to USAID for the SLAC/NIIT collaboration to provide monitoring for PERN/NTC.

Submitted paper on "Anomalous Event Detection" to NOMS 2006.

We made the following presentations:

- Terapaths: Datagrid Wide Area Monitoring Infrastructure (DWMI) presented by Les Cottrell at the DoE Network Research PI meeting BNL, Sept '05.
- Report from ICFA Digital Divide WorkshopCommand: Daegu, Korea, May 23-27 05 presented by Les Cottrell at the Internet2 Fall Members meeting, Philadelphia Sept 21, 2005.
- Network Monitoring for SCIC prepared by Les Cottrell for the ICFA meeting September 2005.
- Network Monitoring Tools for High Performance Networks presented by Les Cottrell at the Internet Fall 2005 Members meeting, Philadelphia, Sep 19, 2005.
- Network Monitoring for ICFA/SCIC presented by Les Cottrell, at ICFA/SCIC meeting 8/24/05
- SLAC Site Report, presented by Les Cottrell for ICFA/SCIC meeting 8/24/05
- Monitoring 10Gbits/s and Beyond presented by Les Cottrell at the LHC tier0, tier1 meeting, CERN July 19 '05.

6.3 Computer Science & Middleware

6.3.1 Condor

The Condor Project produced four Developer's series releases for the 3rd quarter of 2005.

Release 6.7.9 marked the debut of the "Parallel Universe". Condor's Parallel universe is a mechanism to support a wide variety of parallel programming environments, including most implementations of MPI. This universe also supports jobs which need to be co-scheduled, that is, jobs where more than one process must be running at the same time to be correct.

Release 6.7.10 included enhancements for Condor-C (Condor to Condor job submission), and the DAGMan workflow manager.

The Quill job queue database mirror debuted in release 6.7.11. Quill replicates the Condor job queue, job history and pool status in an independent RDBMS. With Quill, the Condor scheduler and Central Manager are insulated from user status queries. Further, a new SQL API is now available for querying job and pool status.

Release 6.7.12 was primarily devoted to bug fixes. In fact, bug fixes are included in all Developer's series releases. Throughout the quarter, all releases have also seen steady enhancements of Condor-C and Globus compatibility. Condor-G now fully supports the Globus 4 Toolkit for Web Services.

Several active efforts have been begun towards enhancing Condor future technology. Work has begun on the next generation SchedD.V7. One of the first capabilities of the SchedD.V7 will be dynamic translation of the job to a remote Grid pool, when no local resources are available. the SchedD.V7 will have more autonomous capabilities for locating suitable job resources, either local or remote.

Another research effort is exploring the ability to deploy a SchedD on the fly. Our implementation of a mobile SchedD has reached a point where we can dynamically install and start a SchedD on an arbitrary machine. This SchedD can later be reliably shutdown and uninstalled automatically. This functionality has been modularized and will be used to provide more dynamic capabilities for current Condor functionality such as condor_glidein.

Research is under way towards adding a dynamic matchmaking capability to the Stork data placement scheduler. This will enable data storage resources to advertise capabilities to Stork, and for Stork to schedule jobs based upon a configurable policy.

July saw the initial operational status of the new Open Science Grid. The Condor VDT group has been instrumental in developing software releases for the Open Science Grid, and contributors to the OSG's initial operational status.

6.3.2 Globus – ANL

6.3.3 SRM – LBNL

Participants: Alex Sim, Junmin Gu, Viji Natarajan, Arie Shoshani

The following activities took place during the last quarter:

OSG activities

- STAR analysis scenario
The STAR analysis scenario consists of the following four steps: 1) files are requested through DRM on the PDSF cluster to get files from BNL's HRM; 2) the files are analyzed on worker nodes at PDSF

through posix I/O; 3) the output files go to the DRM cache; and 4) the output files are pushed to BNL's HPSS through HRM.

Several changes were made to the DRM in order to assign a site URL to files being written through posix I/O. The site URL is required in order to take the next step of copying the files to the target site at BNL. We are collaborating with Eric Hjort on this activity. This scenario was setup, and works correctly with test cases.

- srm-tester 1.0 released to VDT/OSG
A new package was developed and introduced into VDT. The package is an tester program to test any SRMs. It is the only test program now available to test SRMs v1.1. This test program can test the following functions: Get, Put, Copy, Ping, GetRequestStatus, and getProtocols. The srm-tester was used to test the LBNL's DRM and Fermilab's SRM/dCache. Both tested successfully.
- DRM 1.2.9 released to VDT
This new version includes the changes needed to support posix I/O by DRM, necessary for the analysis scenario mentioned above. The requirement is to open, close, and read files that were put into DRM cache. However, immediately after it is written it should have the write permission removed and only read allowed. The DRM was enhanced to include this feature to ensure the integrity of files written into DRMs (as well as HRMs). The new version runs on multiple platforms including: Redhat 7.3, Redhat 8, Redhat 9, Redhat Enterprise-Linux 3, Suse 9, and Debian.

Support of LBNL's SRMs

- Upgrade of HRM on various sites
The HRM was upgraded to include the changes made for posix I/O to support the analysis scenario. The new HRM was installed on BNL and PDSF. In addition, the File Monitoring Tool (FMT) was upgraded to work with the new version. This will be used in the demo at SC2005.
- Helping fix the srmcp client program
This work was performed in collaboration with FNAL to fix srmcp for gsi authentication and file transfer problems. The srmcp is a command-line client program (that allows Get, Put, and Copy) developed at FNAL, and it was tested to work with SRM/dCache. However, it did not work properly with LBNL's SRMs. The authentication problem required adding gsi to functions not tested previously. The transfer problem came about because gsiFTP was not previously used. These problems were fixed and now srmcp interoperates correctly with LBNL's SRMs.

New development

- Berkeley SRM v2.1.1
The development of SRM v2.1.1 continued during this quarter, and now includes HPSS (BTW, it is also used to interface to the MSS system from NCAR). This version is a unified version that can include access to a disk system only or to a disk system that interfaces to an MSS, such as HPSS. It now includes all the functionality in the specification except the authorization functions.
- New srm-client tool
A new SRM command-line client was developed by LBNL for SRM v2.1.1 that also works with SRM v1.1. It can be invoked with a parameter to work with either version. It's primary purpose is to run with SRM v2.1.1 functionality including the new "space reservation" and "directory" functions. It was tested with the current implementation of SRM 2.1.1 from LBNL. It is available on-line for future

implementations of SRM v2.1.1.

- **xrootd integration with DRM**
xrootd is a system developed to support high performance and fault-tolerance of distributed ROOT files. It provides asynchronous parallel requests, stream multiplexing, data pre-fetch, automatic data segmenting, and the framework for a structured peer-to-peer storage model that allows massive server scaling and client recovery from multiple failures. We have discussed joint activity to have xrootd use Disk Resource Managers (DRMs) to manage the space of server nodes, and to use the Hierarchical Resource Managers (HRMs) to access HPSS. Other SRMs could be used as well. We plan to develop a prototype of the combined technologies.

Other activities

- **Grid Collector**
The Grid Collector is supported as part of the SDM ISIC activities, but it is related to PPDG, since it addresses analysis at the event level, rather than a at the file level. This work was performed in collaboration with the STAR team at BNL. We report here the culmination of work that took place over the last two years. The Grid Collector combines an efficient indexing technology, called FastBit, with a Grid file management technology (SRMs) to speed up common analysis jobs on high-energy physics data and to enable some previously impractical analysis jobs. To analyze a set of high-energy collision events, one typically specifies the files containing the events of interest, reads all the events in the files, and filters out unwanted ones. Since most analysis jobs filter out significant number of events, a considerable amount of time is wasted by reading the unwanted events. The Grid Collector removes this inefficiency by allowing users to specify more precisely what events are of interest and to read only the selected events. This speeds up most analysis jobs. In existing analysis frameworks, the responsibility of bringing files from tertiary storage to disk falls on the users. This forces most of analysis jobs to be performed at centralized computer facilities where commonly used files are kept on disks. The Grid Collector automates file management tasks and makes it easy to perform analyses on data files that are not already on disk. This enables some analysis jobs that were previously too time-consuming. The Grid Collector was deployed in the STAR analysis framework. The following publication on the Grid Collector received a best paper award at the International Super Computing (ISC) 2005 conference.

publications

Grid Collector: Facilitating Efficient Selective Access from Data Grids, In Proceedings of International Supercomputer Conference 2005, Heidelberg, Germany, best paper award. K. Wu, J. Gu, J. Lauret, A. Poskanzer, A. Shoshani, A. Sim, W. Zhang.

6.3.4 Caltech

6.3.5 SRB

A standard characterization of grid technology is emerging that uses the concept of virtualization to explain how use of distributed resources is managed. In this characterization:

- **Grids** – provide support for workflow virtualization. This is the ability to manage the execution of processes on remote compute resources, independently of their execution environments.
- **Data Grids** – provide support for data virtualization. This is the ability to manage a shared collection that is distributed across multiple storage systems, independently of the remote administrative domains. All properties of the shared collection are managed by the data grid.

- Semantic Grids – provide support for information virtualization. This is the ability to reason across inferred meanings of attributes in distributed collections. Examples are the ability to do distributed joins across attributes in disjoint collections.

The management of shared collections also requires the ability to support Trust Virtualization. This is the assignment of “ownership” of the shared data to the data grid, which then stores the data under account Ids that correspond to the data grid. This minimizes the administrative support required to install data grids, as only one account needs to be installed at each remote storage system. The data grid then authenticates users independently of the remote storage system, manages access controls independently of the remote storage system, and checks the access controls on each request. The SRB data grid supports access controls on data, metadata, and resources. Hierarchical Trust Virtualization also is needed in SRB data grid federation. In this case, a trust relationship is established between each data grid in the federation. Each user is associated with a “home” data grid. A request by a user for a file in a remote data grid then goes through the following steps:

- User authenticates to the “home” data grid
- User issues a request to the “home” data grid for a file in a remote data grid
- The request is forwarded to the remote data grid which checks the “home” data grid identity
- The remote data grid asks the “home” data grid to authenticate the User’s identity
- The remote data grid then checks its own access controls for permissions
- The data is returned if all checks are satisfied.

This federated environment is being installed on the BaBar data grid. Specific support from SDSC has included fixing bugs in the current version (SRB v 3.3.1), expanding the SRB test suite, extending the set of platforms that support the SRB, and providing public access to the status page of the SRB automatic test system at <http://www.sdsc.edu/srb/tinderbox.html>

The next release of the software will include fixes to over 50 SRB bug reports, and the initial code needed to support queries across SRB data grid federations.

Effort report from Wayne Schroeder of SRB-SDSC on Data Management

Assistance has been provided to the BaBar high energy physics experiment team, as they make use of the SRB software for federation of two independent SRB data grids between Stanford and Lyons, France. This included fixing some bugs, such as SRB Bugzilla item 172 (a Sphymove -P failure). Some of the support provided was done on our srb-chat email list, which is archived at <https://lists.sdsc.edu/pipermail/srb-chat/>.

SRB version 3.4 is being finalized and documented, and will be released after that is done and it successfully completes a series of final tests, probably in the next few weeks. This version has many small, and some large, improvements and bug fixes, which we have been developing over the last few months. Over 50 SRB bugzilla items (bugs and enhancements) have been resolved since our previous 3.3.1 release on April 6 (see <http://srb.npaci.edu/bugzilla>).

Pre-release testing of SRB code (from our CVS repository) has been significantly extended. Adil Hasan, formerly with BaBar and now with the UK E-Science project, developed and contributed a set of SRB test scripts (see the last item on <http://www.sdsc.edu/srb/contributed.html>), written in Python, which we integrated into the SRB automatic testing system (Tinderbox based) as yet another set of tests. This testing has discovered two or three previously unknown bugs and, on occasion, would quickly determine that a change caused a new problem; all of which we have now fixed. We have also extended our automatic testing system to run different tests on the various hosts depending on the time of day. We also recently added the ability to build SRB via gcc on a Solaris host (in addition to via the Solaris compiler) and so were able to get a Solaris host functioning as part of the automatic test system. The status page of our automatic test system is now available for viewing from our home page (www.sdsc.edu/srb) via a link to a description (<http://www.sdsc.edu/srb/tinderbox.html>) which links to the actual status page.