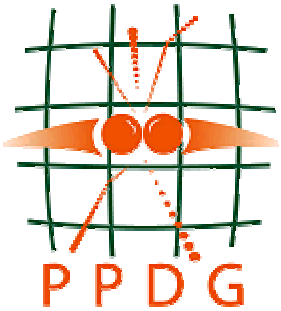
Particle Physics Data Grid Collaboratory Pilot

Quarterly Status Report of the Steering Committee, October - December 2002

31 Jan. 2003



1 Project Overview 2

1.1 Highlights 2

1.2 Project Management and Organization 2

1.3 Trouble Shooting and Diagnosis in Grid Environments 2

1.4 Plans for the next Quarter 4

2 Common Service Areas 4

2.1 CS-1, CS-2 Job Description Languages, Management and Scheduling 4

2.1.1 Collaboration with EDG WP1 4

2.1.2 SAM Job and Information Management (JIM) 4

2.2 CS-3 Information Services 5

2.2.1 Monitoring and MDS work 5

2.2.2 Collaboration with IEPM, Network Performance Monitoring 6

2.3 CS-4 Storage Management 8

2.3.1 LBNL-SRM Development 8

2.3.2 JLab-SRM 9

2.4 CS-5 Reliable File Transfer 10

2.4.1 Globus RFT 10

2.5 CS-6 Robust Replication 10

2.5.1 BaBar Database Replication (BaBar-SRB) 10

2.5.2 Globus ISI, RLS work 10

2.6 CS-7 Documentation 11

2.7 CS-8 Evaluations and Research 12

2.8 CS-9 Security, Authentication, Authorization, Accounting 12

2.8.1 Certificate/Registration Authority 12

2.8.2 Site-AAA 12

2.8.3 Globus Site-AAA work 12

2.8.4 Globus CAS 13

2.8.5 Planned development work 13

2.9 CS-10 Experiment Grids and Applications 13

2.9.1 ATLAS 13

2.9.2 BaBar 14

2.9.3 CMS 15

2.9.3 DZero 16

2.9.4 JLab experiments, and QCD 16

2.9.5 STAR 17

2.10 CS-11 Grid Interface with Interactive Analysis Tools 18

2.10.1 Clarens – Distrbuted Analysis 18

2.10.2 ATLAS DIAL 19

2.11 CS-12 Catalogs and Databases 20

2.11.1 STAR file metadata catalog 20

2.11.2 SDSC – SRB 20

3 Single Collaborator Reports 21

3.1 ANL – Globus 21

3.1.1 Coordination and Support 21

3.1.2 Globus Toolkit updates and bug fixes 21

3.1.3 Globus Toolkit 3.0 21

3.2 SDSC – SRB 21

4 Appendix 22

4.1 List of participants 22

4.2 Meetings 24

1 Project Overview

1.1 Highlights

The project teams continued work on their applications and deployments. There was significant focus on production work for several experiments, and for working demonstrators for SuperComputing 2002 (http://www.ppdg.net/docs/presentations/sc2002-trillium-demos.pdf). This included participation in the WorldGrid interoperability demonstration with iVDGL, GriPhyN and the European DataTag and DataGrid projects.

A successful workshop on Analysis Tools was held at Caltech in association with the combined PPDG, GriPhyN, and iVDGL (Trillium) collaboration meeting. It is planned to hold another workshop for more detailed discussions and interface design near the CHEP conference at the end of March, see Section 2.10.

PPDG organised a workshop to discuss and review Troubleshooting and Diagnosis in Grid Environments. This workshop was attended by about 35 application and computer scientists from a range of Grid projects and application domains. A draft report is under preparation (see documentation below, Section 1.3).

The Site-AA project completed its scheduled work and reported to the PPDG Steering Committee and the funding MICS sponsor, Mary Anne Scott, at the collaboration meeting in December. A summary of recommendations is shown in Section 2.8.2.

A Storage Resource Management (SRM) workshop took place at CERN in early December; to further define the functionality and standardize the interface of SRM – a Grid middleware component. This successful meeting coordinated efforts from PPDG participants (LBNL, Fermilab, JLab) and the European Data Grid (WP2 and WP5). As a result a new functional design document and SRM interface document will be written, which is intended as the next standard interface for SRMs implemented by the different groups. See Section 2.3.1.

1.2 Project Management and Organization

The Executive Team continued with regular phone meetings. Face to face steering meetings were held at SC2002 in Baltimore and at Caltech during the collaboration meeting in December. The Executive Team and Richard Mount (PI) met with representatives of the DOE Nuclear Physics office to discuss the Phenix experiment's interest in collaborating with PPDG, and possible increase in communication between PPDG and that office of the DOE.

1.3 Trouble Shooting and Diagnosis in Grid Environments

Planning for this workshop culminated in a day of presentations and discussion at the Westin Hotel, O'Hare. The scope of the workshop was to review and discuss current needs, practice and work in the area of Error Handling, Diagnosis, Troubleshooting and Problem Diagnosis on production Computational Grids. This workshop was sponsored by the DOE MICS office as part of the SciDAC program, and the NSF. It was a goal of the workshop to prepare a report on what has been learnt, with particular attention to the needs of the Trillium projects. The report is in draft form and includes the following table of contents:

1.1 Introduction and Overview

1.2 Requirements for Grid Monitoring and Troubleshooting

1.3 Deployment, Operation and Troubleshooting

1.4 Propagation, Logging, Interpretation and Response

1.5 System Instrumentation, Probing, Performance Problem Solving

1.6 Necessary Areas of Research and Development

We identified several areas where further research and development is needed:

1) Instrumentation and Analysis Infrastructure with Local through Global Scope

Research and development of tools that support the generation, collection, archiving, analysis and presentation of behavioral information about all Grid layers – application, middleware and fabric, such as:

1. Flexible, extensible and easy to manage instrumentation, collection and storage of behavioral information throughout all software layers and hardware components.
2. Fine and course grained control of the volume, nature and lifetime of the information, at all stages of its lifecycle – generation, collection, storage, presentation and analysis. The needs to be easy and ubiquitous mechanisms to match the characteristics and configuration of the information and functionality to the goals of the monitoring goals. It will be only too easy to generate and record vast quantities of information for which there are insufficient resources - programmatic and human – to synthesize and interpret.
3. Means to ensure that the information is uniquely defined, machine decipherable and understandable. Examples would include a hierarchy of uniquely specific job identifiers throughout the distributed system; universal timestamps etc.
4. Infrastructure for the storage of, management, containment, storage (temporary, durable and long term), mining and analysis of the suite of distributed, structured, heterogeneous log files and/or databases.
5. Curation of long term monitoring and information repositories for long-term trend analysis, prediction and archeology.

2) Consistent and Interpretable Universal and Component Fault and Error Systems

1. Universal naming, interfaces, principals and protocols for error and fault information and codes. This should include standards efforts through such organizations as the Global Grid Forum and W3C.
2. Definition and design of error and fault reporting and handling, synthesis and response mechanisms through all layers of the middleware and application infrastructure. Provision for automated rule, algorithm and human based response and reasoning mechanisms at the component and the system level.
3. Standards for and development of a Grid event service, including the delivery, the interface semantics, and the underlying transport. This should specifically include attention to the expectations of reliability, completeness and fault tolerance of the fault handling and management systems; including allowance for inconsistent state reporting, system guarantees, potential incompleteness of the information, latencies in transport and provision for archeology of the historical record.
4. Definition of aggregation and filtering semantics for reduction and synthesis of information to eliminate redundancy, provide for summary, interpretation and reasoning.

3) Operations and Response Support Technologies and Organizations

Research is needed into the required operational support and response infrastructures – both holistic and technical to:

1. Understand the sociological and organizational constraints and realities of operating integrated and transparent computational infrastructures through national and international locations.
2. Analyse the legal, procedural, and political ramifications of a global cyber infrastructure.
3. Negotiate, define and standardize on policies and procedures for operation, support, and incident response at the geo-political scale.

4. Provide pervasive collaborative environments for distributed independent investigatory teams faced with the full range of hard intermittent, global and/or localized faults and performance anomalies. Often a full complement of experts from every layer of software and hardware is required to work together to identify and solve peripherally coupled but ultimately catastrophic sequences from cause to effect.
5. Develop strategies - both human and automated - to respond to, mitigate and repair fault and failure conditions.
6. Apply automated reasoning and response technologies to grid operations, support and response requirements.

4) Modeling of the System Performance and Response Characteristics

1. Develop models of the distributed system and through the injection of faults and anomalies explore the system and component behavior and performance profiles
2. Through these models develop metrics for the performance and response behaviors.
3. Develop automated comparison and validation tools between the system models and operating infrastructures.

1.4 Plans for the next Quarter

During the next quarter we will be preparing for the SciDAC PI meeting in March and the PPDG review in April. We plan to hold a series of phone meetings to survey the status of the project and assess the needs for the next years work.

Project work will continue as planned across all the experiments and many of the common service areas.

2 Common Service Areas

2.1 CS-1, CS-2 Job Description Languages, Management and Scheduling

2.1.1 Collaboration with EDG WP1

Collaboration continued with EDG WP1 in the completion of the GLUE Schema – common Globus MDS schema for resource discovery and information (in collaboration with iVDGL and DataTAG).

2.1.2 SAM Job and Information Management (JIM)

The prototype version of the JIM (jobs and information management) software was released in early October, and demonstrated to the DZero experiment on Oct 10, 2002. A refined and improved version was shown at SC2002 in Baltimore in November. Features include:

- Remote submission of via the SAM-Grid, Condor-G, Globus layered system.
- Grid job brokering based on the amount of data cached at the participating sites
- Web-based monitoring of the grid system and of grid jobs. The monitoring is viewable at: http://samadams.fnal.gov:8080/prototype/

At each submission site, a user interface is provided which accepts jobs written in a simple Job Description Language (JDL). A parser translates the description into a Condor ClassAd that is delivered to the Condor queuing server (Schedd). A Condor Collector gathers information from each execution site's Grid sensor, or advertising entity, also through ClassAd's. The information for submitted jobs, and available site resources are matched in the Condor Negotiator to rank jobs for submission. The Condor team provided the ability to use an external module for the matching criteria, and this provides the ability to write algorithms using known resource parameters to control the job distribution for the system. The Condor negotiator

sends the jobs through the Condor Grid Manager to the GRAM server (gatekeeper) on the gateway at the appropriate processing resource. Standard GSI mechanisms are used to provide authentication for each user with grid certificates. A kerberos to x509 translator was used to also enable the use of existing Fermilab Kerberos principals.

The jobs that can be submitted include both "vanilla" and "SAM-enabled". Vanilla refers to those jobs which do not take advantage of any SAM provided data management services. SAM-enabled jobs include dataset descriptions and the data handling facilities of the existing SAM infrastructure provide files at the selected processing site for the job to consume. The decision of where to send a processing job is currently based on number of needed files for the project already cached at each execution site. This algorithm will be extended and mature with experience, but its flexibility is a major feature of the system. A "sandboxing" mechanism was built that packages up a complex user job at the submission site, and sends it with the job to the execution site.

For the SC2002 presentation both DZero and CDF submission, execution, and monitoring sites were deployed. This was not really a special demonstration that was prepared, but just a view of the system as it existed at the time. Existing DZero SAM sites were selected and upgraded with JIM software. The SAM-Grid job management was integrated with the CDF Cluster Analysis Facility (CAF) software to provide a working analysis system for CDF. CDF and DZero analysis and test jobs were submitted and resulting histograms were collected from the grid to a web-accessible area for display. For SC2002, there were twelve sites enabled altogether. The DZero sites included 1) Fermilab, 2) the University Texas, Arlington, TX, 3) Michigan State University, East Lansing, MI, 4) The University of Michigan, Ann Arbor, MI, 5) Imperial College, London, UK, and 6) GridKa in Karlsruhe Germany. The CDF sites included 1) Fermilab, 2) Texas Tech University, Lubbock, TX, 3) University of Toronto, Toronto, Canada, 4) Rutgers State University, NJ, 5) Rutherford Appelton Laboratory, Oxfordshire, UK, 6) Kyungpook National University, Daegu, Republic of Korea.

A lot has been learned about the installation and operation of the current software in the last quarter of 2002, but there remains extensive work to be done before the system can be used for production in the spring of 2003. The installation procedures are being improved and automated so the system can be deployed to all of the existing sites by personnel at each location, and not require extensive help form the Fermilab team. Many problems related to various firewall configurations and requirements have been confronted during the testbed deployment and they are resolved and briefly documented. SAM has adapted well to operating on various cluster configurations, including local disk on each node, shared NFS disk, private and public networks. Additional details about the system and installation can be found at http://www-d0.fnal.gov/computing/grid/SAMGridManual.htm. After SC2002, the main concentration has been on the evolution of the software version 1, and this resulted in design discussions on the manipulation of the site xml-based configuration for most of the execution and monitoring side server activities.

The majority of the work for the SAM-Grid effort has been done by Igor Terekhov, Gabriele Garzoglio, and Andrew Baranovski. Major contributions were also made by three students employed for the project over the period. Two Masters students, Siddharth Patil and Abhishek Rana, from the University of Texas, Arlington, and a Coop student, Hannu Kouteniemi, from Espoo-Vantaa Institute of Technology, Finland, have made significant contributions to the project. The students worked on grid job client, grid resource advertisement and web monitoring aspects of the project. The collaboration with the Condor team, especially Todd Tannenbaum and Alain Roy, has been very productive. Tom Rockwell from MSU played a significant role in making the SC2002 presentation work, and Rod Walker at Imperial College London continues to play a major role in the JIM project. Significant work has been done by the CDF team to bring the system up at their sites, the individuals involved in this effort include Frank Wuerthwein and Stefan Stonjek, as well as many others at CDF.

2.2 CS-3 Information Services

2.2.1 Monitoring and MDS work

Globus members have taken a leadership role in the GLUE-schema work that is defining a joint-schema with DataTAG and EDG for interoperability. Information providers using the compute element (CE)

schema were developed and used as part of the WorldGrid SC demos. Work for the storage element (SE) schemas was completed as well. In January it is planned to have the group evaluate the schemas based on their use this past quarter and to update the CE schemas, as well as begin work toward the Network Element schemas.

2.2.2 Collaboration with IEPM, Network Performance Monitoring[1]

2.2.2.1 Web Services

To make PingER results available for Grid applications and to learn about web services, we worked with the University College London to provide node details via web services. Currently the main way Globus applications publish and subscribe to computer information is via MDS. To extend this to access network information we installed MDS/Globus, evaluated the schema used by the EDG for reporting network measurements and are working with collaborators to determine the useful metrics and what else may be required.

To enable adding GridFTP to the IEPM-BW monitoring we worked with Globus and the DOE Science Grid to get certificates. For long-term use we are utilizing the DOE Science Grid certificate. We made regular measurements with GridFTP to compare with bbcp and bbftp. Early results from this were presented at the Edinburgh GGF in
http://www.slac.stanford.edu/grp/scs/net/talk/ggf5_jul2002/NMWG_GGF5.pdf

2.2.2.2 PingER

Modifications to the code received from the European DataGrid (EDG) and NASA have been incorporated into the standard release. The modifications to connectivity.pl and mon-lib.pl provide for an extra color (blue) and a new application type (vnd-ms-excel).

The PingER project has been successfully used to provide information for ESnet, the International Committee for Future Accelerators (ICFA) Standing Committee on Inter-regional Connectivity (SCIC) and to the Abdus Salam International Center for Theoretical Physics (ICTP) eJournals project. The latter two projects are identifying and trying to come up with alternatives to bridge the Digital Divide, an important activity these days, especially for potential Grid collaborators in the Former Soviet Union, India, Pakistan, Latin America, the Caucasus, and S.E. Europe. The collaboration with ICTP is particularly fruitful and has already yielded hosts to monitor in 16 developing countries including Bangladesh, Brazil, China, Columbia, Ghana, Guatemala, India (Kerala, Hyderabad), Indonesia, Iran, Jordon, Korea, Mexico, Moldova, Nigeria, Slovakia, and the Ukraine.

2.2.2.3 IEPM-BW throughput Measurements

To make the measurement code more portable, flexible and powerful, we embarked on two major re-writes to port from Unix, make it easier to add new tests, to improve the reporting (in particular in the areas of predictions and correlation statistics), improve diagnostics, and parameterize the configurations needed for other sites. In addition we developed tools to automate the porting of the monitoring code to new sites, to clean up hung processes and recognize and report pathologies such as measurements failing, running out of disk space or process slots etc. If there are no snags we can now port the toolkit to a new monitoring host in about 30 minutes. As a result of this there are now 10 sites running the code, plus it has been run at iGrid2002, SC2002 and on the Caltech/SLAC/CERN testbed (see below for more on these). Of the ten sites, APAN (Japan), FNAL (Chicago), Georgia Tech, NIKHEF (Amsterdam), INFN (Milan) and SLAC are now running in production (i.e. they have chosen their own remote sites to monitor and maintain their own configurations). U Michigan, Internet2, U Manchester and UCL (London) are evaluating. SLAC is now regularly monitoring over 40 remote sites with high speed connections in 9 countries. The throughputs vary from a few tens of Mbits/s to several hundred Mbits/s providing a valuable testbed for measurement tools (such as packet dispersion techniques and iperf), applications (such as bbcp, bbftp, gridFTP) and new TCP stacks (such as Net100, TCP FAST and HS TCP).

[1] http://www-iepm.slac.stanford.edu/

Following discussions with the INFN (Italian Academic and Research Network for HENP) people at Padova and Trieste, they have approved a project to build and deploy 14 embedded Linux probes to make active high performance throughput measurements at about 12 to 16 INFN sites. They will be using they IEPM-BW toolkit for this deployment.

As proof of the extensibility of the toolkit we added UDPMon (from Manchester) and GridFTP to the suite of test tools.

We studied the effects of the file and disk subsystems on throughput achievable by applications such as bbcp and bbftp. This has been written up in Disk Throughputs (see http://www-iepm.slac.stanford.edu/bw/disk_res.html).

Analyzing the data from IEPM-BW we were able to identify simple ways to evaluate the quantitative extent of the diurnal changes in the performance between many sites. This is documented in http://www.slac.stanford.edu/comp/net/pattern/diurnal.html

As one moves to higher speed links the time spent in TCP slow start increases so one has to make iperf TCP measurements for longer (e.g. for a 1Gbits/s link with an RTT 0f 200msec the time in slow start is about 7-8 seconds) to ensure that most of the measurement is for the stable (AIMD) throughput mode. This can become very intrusive on the network, so we embarked on creating a new version of iperf that uses Web100 to evaluate when it is out of slow-start and make measurements for 1 second after that. This decreases the traffic and time to make the measurement by over 90%. See http://www-iepm.slac.stanford.edu/bw/iperf_res.html for more details on this so-called iperf Quick mode.

We also incorporated Web100 into IEPM-BW and validated the results against the related active measurements, and against Netflow measurements (see http://www.slac.stanford.edu/comp/net/bandwidth-tests/web100/).

2.2.2.4 iGrid2002

We successfully submitted a proposal entitled Bandwidth from the Lowlands (see http://www-iepm.slac.stanford.edu/monitoring/bulk/igrid2002/) together with NIKHEF to simulate an HEP accelerator site replicating data to multiple tier 1 sites. We demonstrated this together with various IEPM measurement tools at iGrid2002 achieving over 2Gbits/s to 32 hosts in 9 countries in N. America, Europe and Japan. We submitted and had accepted a paper “iGrid2002 Demonstration: Bandwidth from the Low Lands” that will be published by Elsevier.

2.2.2.5 SC2002

For SC2002, we proposed and had accepted a demonstration entitled Bandwidth to the World (see http://www-iepm.slac.stanford.edu/monitoring/bulk/igrid2002/). Briefly the demonstration emulated an HENP tier-0 accelerator site distributing large volumes of data to about 40 collaborator sites. This utilized the IEPM-BW infrastructure we have set up (accounts, ssh keys, contacts, configurations etc.) plus about 8 hosts at SC2002.

In addition we demonstrated various IEPM developed network measurement tools including a new real-time Java Applet to display ping RTTs to regions of the world (see http://www-iepm.slac.stanford.edu/tools/pingworld/), PingER animated replays of RTT, loss and derived throughputs (see http://www-iepm.slac.stanford.edu/pinger/perfmap/fromslac/anim/perf_world_anim.gif), IEPM-BW and a new Available Bandwidth Estimation (ABWE, see http://www-iepm.slac.stanford.edu/monitoring/bulk/sc2002/rhmon.jpg) tool based on packet dispersion.

We also collaborated with Caltech to make an SC2002 bandwidth challenge. We achieved over 12Gbits/s with standard 1500 Byte MTUs and 16 cpus, which was the second largest throughput achieved at SC2002. It utilized the new FAST TCP stack (see http://www-iepm.slac.stanford.edu/monitoring/bulk/sc2002/hiperf.htm). We also achieved over 900Mbits/s iperf TCP throughput from a single host at SC2002 (Baltimore) to Sunnyvale using a single stream.

Finally we worked with NIKHEF and Caltech to send over 900Mbits/s from a single host at Amsterdam to a single host at Sunnyvale with a single stream and jumbo frames (9Kbytes). This was documented and submitted to Internet 2 for the Land Speed Record.

2.2.2.6 Testbed

Following discussions at iGrid2002 with Level(3) personnel, Olivier Martin of CERN and Linda Winkler of StarLight, followed by more discussions at CERN with Harvey Newman of Caltech, we successfully worked with Level(3) to get the loan of colocation space at their Sunnyvale colocation gateway plus an OC192 (10Gbits/s) circuit from Sunnyvale to StarLight (Chicago). In addition we successfully requested a loan of a Cisco GSR router plus interfaces (valued at about $1M) that were placed at Sunnyvale. In addition Caltech placed 16 high performance disk and CPU servers at Sunnyvale. This has become part of a wide area high performance testbed including sites at CERN, StarLight and Sunnyvale. We utilized this testbed extensively at SC2002 for the bandwidth challenge, a successful Internet 2 Land Speed Record (LSR) attempt, and to demonstrate various high performance measurement. More details can be found at http://www-iepm.slac.stanford.edu/monitoring/bulk/sc2002/hiperf.htm.

Following SC2002, the loan of the router, circuit and collocation space was extended and we will make further tests of measurement tools as well as testing high speed disk-to-disk measurements, 10GE NIC tests (in collaboration with Wu–chun Feng of LANL), and possibly IPv6 and QoS tests.

2.3 CS-4 Storage Management

2.3.1 LBNL-SRM Development

People involved: Junmin Gu, Alex Sim, Alex Romosan, Arie Shoshani

The following tasks were accomplished:

1) Routine use of HRMs for file replication between BNL and NERSC.

During this quarter, we have continued with routine intensified use of HRMs for file replication between BNL and LBNL as part of the STAR experiment. The system has proven to be extremely robust and worthy of routine use. It handles almost daily transfers of 100s and even more than 1000 files per request that lasts many hours. The transfers can be monitored dynamically over the web with a File Monitoring Tool (FMT). One problem discovered with the FMT was that over time it uses too much memory to the point that it can interfere with the efficiency of file transfers. We have changed the design of FMT so that it keeps minimal information in memory. This work is scheduled for the next quarter.

2) Development of analysis program for the file replication process.

We have developed a program to analyze the log output from HRM runs, and prepare then for plots. The plots can be used to identify bottlenecks in the system. Specifically, the plots show the history of each file staging at the source (from BNL-HPSS to BNL-HRM's disk), transfer using GridFTP (from BNL-HRM's disk to LBNL-HRM's disk), and archiving of the file (from LBNL-HRM's disk to LBNL-HPSS). Many file transfer requests were analyzed, and most have shown that the bottleneck is in the network. The plots also show recovery from transient failures in these stages of transfers by the HRMs. This important recovery feature in HRM is one of the reasons for the robustness of these massive file transfers.

3) Setting up WSDL-based HRM Version 1.1 compatible with Fermi and JLab

The purpose of this work was to develop a version of the HRM interface that is compatible with developments at Fermi and Jlab. We have developed a gateway to our Corba-based interface that accepts SRM v1.1 WSDL requests and translate them to equivalent Corba calls. In trying to achieve this interoperation we have used gSOAP to provide C++ interface based on the WSDL definition, while Fermi and JLab used GLUE, a product that provides a Java interface based on the WSDL definition. We have discovered that the versions generated are not compatible, and that they fail when we use array structures. This led us to a conclusion that we need to find a single product that supports both Java and C++.

4) DRM-NeST integration

After several tries, we were able to install NeST on a Linux machine. This required an upgrade of the Linux system to support a quota capability. In trying to use NeST, we have discovered several bugs that were reported to the NeST team and corrected. We interfaced NeST with DRM, and learned how to use this capability in our test environment. We also developed a design document that explains the

requirements from the NeST system in order to support DRM's reservation of three types of spaces: permanent, durable, and volatile. The main requirement is to be able to assign a space to a client dynamically and establish a directory in that space. The developer of NeST, John Bent, stated that such a capability can be added, and will be done in the next version of NeST.

5) Meeting at CERN to coordinate SRM version 2.0

This was an important activity that involved people from PPDG as well as the European Data Grid (EDG). This 2-day meeting's purpose was to further define the functionality and standardize the interface of Storage Resource Managers (SRMs). The meeting took place at CERN on December 4-5, 2002. In this meeting we focused on additional functionality, especially in the area of dynamic storage space reservation and directory functionality in client-acquired storage spaces. The results of the meeting were described at the PPDG meeting in Caltech in mid December. The LBNL team agreed to put together the documents of the new SRM spec, Version 2.1 during the next quarter. A new web site (http://sdm.lbl.gov/srm-wg) was established to coordinate the activity of the people involved with the SRM design and development. The participants we people from EDG-WP2, EDG-WP5, JLAB, FermiLab, and LBNL. The meeting coordinator was Arie Shoshani.

6) Developing 3 demos for SC 2002

We put together 3 demos at SC 2002 that are relevant to PPDG:

a) Robust File Replication of Massive Datasets on the Grid. This demo consisted of showing the file replication between BNL and LBNL, described above using the web-based FMT tool.

b) Access of GridFTP-to-HPSS through HRM. This was a demo of a capability developed in the previous quarter of using gridFTP to access files from HPSS. This was achieved by modifying the gridFTP daemon to access HRM. HRM staged the file into its disk, and returned to the gridFTP daemon to complete the transfer from that disk location. The reverse capability of writing into HPSS with gridFTP was also demonstrated.

c) Uniform Grid Access to Different Mass Storage Systems. This demo consisted of accessing files using the newly developed WSDL interface described above.

2.3.2 JLab-SRM

Jefferson Lab has continued to develop and deploy web services based Storage Resource Management software (server and client) on two fronts, one using the lattice QCD project as the customer, and the other using experimental physics as the customer. The major activity this quarter has been collaborating in the specification of the next version of the SRM specification, particularly so that it includes the directory and file operations already in use in the J-SRM implementation and GFM (grid file manager) client.

2.3.2.1 JLab-QCD

For the Lattice Data Grid, development of J-SRM continued, particularly the interface to Jasmine for silo put/get, and the automatic migration of files to Jasmine as the managed disk cache filled. Also, work on the recursive grid tree copy/merge was completed and preliminary testing done.

Input to the specification of the next SRM version was provided, based upon experience gained in developing and using the Grid File Manager (GFM) with J-SRM. Arie Shoshani incorporated many of these suggestions into the draft document, and additional discussions were carried out by email. The intention is to merge the work on J-SRM and the Jasmine-only SRM (which implement different disk management policies) to a single product (to support multiple policies). The end product would possibly be a proper superset of the next SRM specified API (to be determined in the first quarter of 2004, depending upon whether all desired features make it into the next SRM specification).

2.3.2.2 JLAB-Experiments

For the experimental physics program, this quarter the two primary efforts were achieving interoperability with collaborating PPDG sites to support the SC2002 demo, and the SRM joint design meeting. Bryan Hess and Andy Kowalski participated in the second SRM joint design meeting held in December at CERN. This meeting, organized by Arie Shoshani, included both PPDG and EDG collaborators. There was agreement to

produce independent implementations of SRMs that could continue to securely interoperate at the protocol level. Additional JLab details are in section 2.9.4.

2.4 CS-5 Reliable File Transfer

2.4.1 Globus RFT

The Reliable File Transfer Service (RFT) will be Globus's first OGSA service, completely up to the current specification. It currently is built on top of the GSI security, and allows single-file transfer only (URL to URL), although there are plans to develop a separate queue service.

An alpha-release, and solicitation of alpha testers, is planned for Fall. The schedule of these releases, which is also the schedule for the Globus Toolkit 3.0 technology releases, is:

- GT3 Alpha: GlobusWORLD, Jan 13, 2003
- GT3 Beta: End of April, 2003
- GT3.0: End of June, 2003

Additional information on RFT is available at http://www.mcs.anl.gov/~madduri/ogsa_docs/reliable_transfer.html.

2.5 CS-6 Robust Replication

2.5.1 BaBar Database Replication (BaBar-SRB)

BaBar demonstrated data replication from SLAC to IN2P3 at the SC2002 conference, copying ~400GB of Objectivity event data using the standard SRB replication tools running with multiple streams (by means of multiple replicate commands spawned by a Python script). The replication ran reliably for the duration of SC2002.

BaBar have been scanning the newly processed data, using an automated procedure to register files in SRB. The tools and procedures for production transfer from SLAC to IN2P3 of the next bulk processing (12-series data) are now being finalised.

An expression of interest in setting up SRB tests at RAL has been received from the e-Science group there.

2.5.2 Globus ISI, RLS work

This quarter saw intensive continued development on the Replica Location Service, including debugging for greater robustness, more expensive testing and performance evaluation, and packaging of the RLS with the Globus Toolkit version 2.2.3. The RLS was deployed widely, including an RLS testbed of over 30 machines on three continents that was used for demonstrations during the SC2002 Conference in Baltimore, MD in November, 2002. The RLS was also used in GriPhyN and Earth Systems Grid demonstrations at SC2002. The RLS saw greater interest and testing from other groups, including PPDG, the European DataGrid project, IBM and others.

2.5.2.1 Turning the Replica Location Service into an Open Grid Services Architecture Service

This quarter, we developed a first draft of a specification document for an Open Grid Services Architecture Replica Location Service. This document was presented at the Replication Research Group of the Global Grid Forum in October, 2002. We proposed an OGSA Data Replication working group of the Global Grid Forum through which we intend to standardize interfaces to replication grid services.

2.5.2.2 Plans for next quarter

Release of RLS as part of the GT3 Alpha release, January 2003

RLS was packaged for release in the GT3 Alpha release, which occurred on January 13, 2003.

Widespread deployment of Replica Location Services

An increasing number of projects (LIGO, ESG, EDG, etc.) are deploying and using the RLS. With the GT3 alpha release of the RLS code, we expect to see the deployment of RLS increase dramatically in the coming quarter. With wider use, we expect to spend extensive time in the coming quarter doing additional debugging, documentation and testing of the software as well as adding some new features.

Development of Replica Location Grid Service Specification and GGF Working Group

We are awaiting final approval on the existence of the Replication Services Working Group within GGF. We will present a new version of the Replica Location Services Specification at the March meeting in Tokyo, Japan.

2.5.2.3 Papers Published or in Progress

Giggle: A Framework for Constructing Scalable Replica Location Services. Ann Chervenak, Ewa Deelman, Ian Foster, Leanne Guy, Wolfgang Hoschek, Adriana Iamnitchi, Carl Kesselman, Peter Kunszt, Matei Ripeanu, Bob Schwartzkopf, Heinz Stockinger, Kurt Stockinger, Brian Tierney. Published in Proceedings of the SC2002 Conference in Baltimore in November, 2002.

2.5.2.4 Presentations Given

October 2, 2002: Presented RLS design as part of PPDG teleconference on reliable replication.

October 15-17, 2002: Presented update on RLS to Global Grid Forum Replication Research Group. Presented first draft of grid service interface for replica location service.

November 21, 2002: Presented paper on Replica Location Service to SC2002 Conference in Baltimore, MD.

November 18-21, 2002: Gave demonstrations of Replica Location Service testbed at SC2002 conference in Baltimore, MD.

2.6 CS-7 Documentation

Document below are posted at http://www.ppdg.net/docs/documents_and_information.htm.

Reports, Documents and Papers		Date/Version
PPDG-27	Site-AAA: Issues list	1/25 (pdf)
PPDG-26	Report from the TroubleShooting Workshop	Draft/1/29/03
PPDG-25	Site-AAA: Recommendation for Future Activities	1/03 (doc, pdf)
PPDG-24	US HEPCAL (Use Case Requirements) Response and Joint (US, EDG) HEPCAL response to the LCG PEB and SC2	US V3 Joint V8 11/02

Talks and presentations:

Presentations & Publications

December 2002	Joint Trillium Collaboration meeting: Doug's presentation on PPDG collaboration with iVDGL, Interoperability ,Site AA report,Analysis Tools, Newsletter
November 2002	GLUE at the MAGIC meeting; PPDG at SC2002 SciDAC booth; PPDG discussion with DOE/NP.
October 2002	Large Site AA Research Group at GGF6; Talk to the LHC Computing Grid Project Grid Deployment Board and to the HICB; Report at D0 Collaboration meeting, Lee Lueking. SiteAA ESSC

2.7 CS-8 Evaluations and Research

2.8 CS-9 Security, Authentication, Authorization, Accounting

2.8.1 Certificate/Registration Authority

The current doesciencegrid.org certificate authority has been in operation for one year, and the certificates, issued with a one year lifetime, are starting to be renewed. This has brought out some issues that the ESnet staff are working on. User certificates are rather straightforward to renew but host certificates, while possible to renew, do not have a simple procedure. This has also brought forward a discussion of whether or not the key-pair for a renewed certificate should be changed or re-used.

2.8.2 Site-AAA

The Site-AA project presented reports of its work at the December collaboration meeting. They are collating a list of issues and next steps, a summary of which is included below:

0) Expand the detailed discussion between site security infrastructure and Grid middleware security developers to include European counterparts.

1) Implement common authorization callout in Globus gatekeeper and gridFTPd.

2) Virtual organization as registrars for multiple sites.

3) Long Running Jobs. - Solutions for this will necessarily involve services acting on behalf of users. The definition of where reauthentication and reauthorization can/must be performed in a Grid and what communications are required is needed.

4) Proxy Generation Services - This is basically a different kind of user proxy generation (ie. authentication method) where the user doesn't maintain the private key (or at least that copy of it in the MyProxy case).

5) Incident handling - Methods and responsibilities for identifying, investigating, responding to, and following up on incidents of attack and misuse need to be determined across the interconnected grid.

2.8.3 Globus Site-AAA work

Von Welch continues to actively lead ANL participation in the PPDG Site-AAA group and attended the Site-AAA meeting at CalTech in December. In addition, in collaboration with Markus Lorch, work has begun on implementing the authorization callout interface this group agreed to at the October GGF meeting.

2.8.4 Globus CAS

In collaboration with Globus Project personnel, LBNL and NERSC have begun evaluation of the second release of our Community Authorization Service (CAS) prototype with a presentation of the work proposed for CHEP.

Additional information on CAS can be found at http://www.globus.org/Security/CAS/.

2.8.5 Planned development work

As stated above, for the next quarter our planned development work includes:

- Work towards the authorization interface for CAS that is now being discussed
- Finish evaluation of the GridFTP performance
- Continued RFT development to stay in agreement with the still changing OGSA spec
- Extended support in RFT for queuing service (next 6 month time frame)
- Additional GT3 development work related to experiment needs

2.9 CS-10 Experiment Grids and Applications

2.9.1 ATLAS

2.9.1.1 ATLAS distributed data manager, Magda (ATLAS-Globus)

One important event in this period is the US Atlas grid testbed demo in the SuperComputing 2002 conference. Magda as an indispensable component on the US Atlas testbed, was presented to the visitors of the conference at BNL booth. Three power point files were made and showed in the demo:

1. To show how Magda organizes the grid-based data (hyperlinks embedded),

   http://www.atlasgrid.bnl.gov/magdademo/archit.ppt

2. To show how Magda was used in the US testbed production for the Atlas Data Challenge (DC) 1 phase1,

   http://www.atlasgrid.bnl.gov/magdademo/jobdata.ppt

3. To present an overall picture of Magda, our achievement with Magda and near term plan,

   http://www.atlasgrid.bnl.gov/magdademo/sc2002_poster.ppt

Magda developer actively participated in the Atlas DC, supported its usage and replied users' emails. As before Magda was used to transfer files between BNL HPSS and CERN castor. The good thing is that physicists from the Physics Working Group tried Magda for the first time on the Internet. They ran Magda to move the DC1 data files around the net in their analysis activities and gave very interesting feedbacks.

As last quarter Magda's usage in the US Atlas testbed DC1 production was continuously supported. Magda itself has been improved as more feedbacks from the testbed developers came in.

The command line tool 'magda_putfile' has been further developed during this period.

1. Extended to manage the files distributed on each node of Linux farms. As discussed before, a farm can been seen as a special Magda 'site', and a Magda farm 'location' is the directory path preceded by the hostname.

2. Extended to work with the files on Lyon HPSS for the Atlas experiment, the third mass storage system that Magda manages.

3. Support third-party transfer with the BNL HPSS, both source and destination can be on BNL HPSS. To users, command syntax for interacting with BNL HPSS is exactly the same as that for working with disk storage. The gridFTP server names are completely hidden from users. Now magda_putfile can invoked with three ways:

magda_putfile <site:location/filename> <site:location>

magda_putfile <filename> <site:location>

magda_putfile <gsiftp:/host/path/filename> <site:location>

4. A new parameter called 'rftpdelay' was added. When trying to move the production output to BNL HPSS, if BNL HPSS is down for some reason, a user could run magda_putfile with this parameter switched on, and then the file will be put to the BNL disk cache which is in the front of HPSS. Later on another script 'magda_promote' could be invoked to send the file to the final destination on HPSS. This feature was found to be very usefull and added flexibility to the US Atlas testbed production.

The command line tool 'magda_getfile' has one more parameter called 'usagehour'. When putting a file to a disk cache defined by the environment variable MAGDA_CACHE, user can specify how many hours he expects this file to be on the cache. After that another script 'magda_cleancache' (ran by a cron job) will delete the file.

A magda user guide is in preparation. It is intended to be a complete reference. Please see http://www.atlasgrid.bnl.gov/magdadoc/userguide.htm

A new Magda server was setup for the NCG group of Stony Brook for the Phenix experiment.

A basic authentication mechanism was developed for the users of the web interface. Normally the Magda web pages are for viewing and querying file information. With the web forms, users could obtain an account for using the web interface, login as members to do editing, modify their profiles and logout. More edit functionalities (edit replication taks, location, site and host) will be developed at the next step.

Kaushik De made invaluable suggestions to the Magda development. Jason Smith maintained the Magda database server and web server, and backuped the database. Dantong Yu maintained the gatekeeper and gridFTP server. Alex Undrus made the second copies of the backup of the database. Torre Wenaus and David Adams helped to write this report.

2.9.1.2 Monitoring

This effort is being led by Dantong Yu of Brookhaven National Laboratory.

Initial steps in organizing a monitoring effort were taken during this period. U.S. ATLAS participates in the joint PPDG/GriPhyN effort for Grid monitoring. Use cases and requirements for a cross-experiment testbed were developed and collected. Work now focuses on developing facilities monitors and MDS information providers.

1. Work on PPDG/SC2002 Grid Monitoring Project.

a. Deployed US-ATLAS Map Center for SC2002 ATLAS demo. Grid Map Center has been designed to logically and graphically represent all elements, applications and services running over grids. It polls grid entities and services (GridFtp, MDS, Gatekeeper), check their status and builds aggregated views of difference types of grid entities.

b. Modify US-ATLAS Map Center for the need of SC2002. The Web site can be found at:

http://www.atlasgrid.bnl.gov/mapcenter/

c. Testing GLUE Schema.

2.9.2 BaBar

2.9.2.1 Distributed batch

An application to make a deep copy of Objectivity collections has been run through the distributed batch system, using the EDG software. Improvements and user testing will be made in the very near future.

We have been working with Alasdair Earl (UK GridPP student) to create a simple script to help users install the client-side Globus software, and are working on extensions to allow users to install the client-side EDG software.

2.9.3 CMS

2.9.3.1 CMS Production Accomplishments

USCMSSC delivered the following sets of simulated and reconstructed events, as a contribution to the CMS-wide production effort in support of physics studies on higher level triggers and the preparation of the DAQ TDR.

• 100K jetmet events generated, digitized, analyzed

• Three TB muon federation data mirrored at FNAL (600K events)

• 50K muon data events analysed so far

The following production tools and systems were developed in US CMS to support these efforts:

• McRunjob accepted for Summer minbias production

• Grid-enabled version of Production (MOP) is being used successfully

• Production tools interfaced locally at Fermilab with dCache

2.9.3.2 Deployment of USCMS Grid

As a result of the general acceptance of the concept of LHC computing grid, USCMSSC is moving rapidly to accommodate necessary changes. Several months ago, USCMSSC initiated the concept of grid test bed. During the past quarter, the feasibility of the grid test bed based on Tier 1/Tier 2 centers was proven. The next step is the deployment of a robust production quality grid. In order to achieve this objective, it is necessary to design a seamless migration path beginning with the development grid environment to the final production grid environment. An important step in this procedure is the deployment of an intermediate integration grid environment, the Integration Grid Testbed (IGT). The final production grid environment, USCMS Production Grid (UPG), will be focused on production, without any testing or testbed activities. In this quarter, USCMSSC members worked hard to deploy the Integration Grid Testbed. They also defined plans for the USCMS Production Grid, the final destination in this migration path.

The USCMS grid testbed will operate a portion of its resources in "production mode" so that running jobs can be supported in a 24x7 fashion. The integration grid testbed will serve as the USCMS production grid in the first few months of operation. Components of the distributed production environment will be run and tested on IGT in preparation for rollover to the production grid. At the time of writing of this report, the team successfully ran the following MOP jobs using MOP master at Fermilab: 100 jobs submitted to the master itself, 100 jobs submitted to UFL. These submissions included the following steps: stage-in, run (CMKIN), stage-out and publish.

VO administration tools were deployed. A complete set of VO administration tools GroupMan and mkgridmap are already deployed. The current implementation of the US-CMS User Database is an LDAP server (based at Fermilab) containing the DN's for each testbed user. The European Data Grid "mkgridmap" scripts are used as a client to map the account information to local group accounts. Finally, the "GroupMan" administrative software (developed and supported by the PPDG) provides a Graphical User Interface and is used to manage the central LDAP server at Fermilab. The US-CMS Grid Testbed will maintain a version of the GroupMan package until it is included in (and hence maintained by) the GriPhyN/iVDGL Virtual Data Toolkit. An initial job execution and job manager systems was deployed using Globus GRAM and Condor-G.

The MonaLisa Monitoring Framework was deployed on the testbed and API for MonaLisa to MDS was developed. GridFTP is already deployed with ftsh from Condor. An alpha version of the Chimera Virtual Data System is already deployed. Alpha versions of several software packages were deployed on the testbed including Virtual Data Catalog, Abstract Planner and Concrete Planner.

A suite of testing procedures for the VDT, was developed which are summarized on the web page http://hepweb.ucsd.edu/vdt/testing-1.1.3.html. New versions of the VDT were validated before release to the general community

A Makefile was provided to install the VDT Client, Server and SDK packages on the test systems, which were chosen to provide a range of different Red Hat operating system versions and different system configurations, from Red Hat 6.1, the current CMS standard, up to Red Hat 7.3, a recent version.

Tests were made of the installation of the individual software packages in the VDT as well as their interoperability. Individual tests were made of Globus, GDMP (including pinging and registering files), Condor, and the interoperability of Globus and Condor. All tests with kernels from the CERN standard 2.2.19-6.2.1.1 upward were successful. However, the Globus GRAM bridge to Condor was not setup properly in the installation of the VDT when both the Server and Client packages were installed. A pacman package to reconfigure the Globus GRAM bridge is provided in http://hepweb.ucsd.edu/vdt/vdt_cache/configure-gram-bridge-to-condor.pacman.

DZero

DZero continues to operate the existing SAM infrastructure to provide the needs of the experiment. There are currently four dozen SAM sites and a team of DZero collaborators from around the world monitor the system and respond to user problems. We have tested GridFTP as the transfer protocol for the system, possibly to replace bbFTP, and half a dozen sites have transitioned to using GridFTP for all extra-domain file transfers. See http://d0db.fnal.gov/sam/documents.html#admin for data transfer performance measurements. Gabriele implemented the sam_gsi_config version 1, the product is responsible for keeping the user authorization list at the various grid sites consistent with a central VO (Virtual Organization) repository. Gabriele also continues to support sam_client_api, the product that interfaces root with sam, and the version that can be compiled together with root analysis programs has been tested by the users and is now being used. We are discussing grid authentication schemes to use for our Database and in conjunction with the middle tier sam database server used in our three tier database architecture. Work is ongoing to integrate dCache with SAM for use in DZero data handling. We have a test dCache server in place and plan to transition it to limited production operation soon.

Work is ongoing with the GridKa team to grow the operation of the prototype Regional Analysis Center (RAC) at the GridKa Center in Karlsruhe Germany.. DZero. DZero collaborators Daniel Wicke, Christian Schmitt (Wuppertal), and Christian Zeitnitz (Mainz) have been working with the technical staff at Karlsruhe to integrate the DZero system and keep it operational. The DZero experiment plans to establish between six and ten fully functional RAC's in the coming years to provide additional computing and intellectual resources to the experiment

2.9.4 JLab experiments, and QCD

Jefferson Lab continues to support two test grids, one for the lattice QCD collaboration, and one for experimental physics simulation and analysis.

The lattice grid still has operating nodes only at MIT and Jefferson Lab. Expansion to the University of Maryland was delayed by a fire effecting their computer room, and will resume the first quarter of 2003. To facilitate future deployments, a set of RPM's (Redhat Package Manager) have been developed and tested for many of the components used in deploying a data grid node. This packaging of components will continue in the next quarter.

The experimental physics side, in November, LBL, Fermilab, and JLab demonstrated SRM interoperability at Supercomputing 2002. Using GSI authentication, independent SRM implementations transferred data over the grid, linking diverse mass storage systems such as Jasmine as JLab and Fermi's Enstore.

The University of Glasgow has now successfully transferred experimental data using the Jasmine-based SRM. To setup this interaction, agreement to accept UK eScience certificates was required. Once we worked through the mechanics of this, we successfully moved data from tape, through the SRM, to Glasgow.

Jefferson Lab's work for the Site-AAA sub-project included specification of site requirements for AAA and revised CSPP policies for inclusion of grid users. A summary report was presented on December 19 at the GriPhyN/iVDGL/PPDG Technical Planning Meeting, including JLab's AAA work done to date and plans for continuing efforts. JLab participated in the preparation of the Site-AAA reports on issues and requirements and on recommendations for continuing work. One man-month of effort was devoted for this period.

2.9.5 STAR

2.9.5.1 New members

In this quarter, we are welcoming three new members from the BNL/ITD department to our Grid team effort : Dave Stampf, Richard Casella and Efstratios Efstathiadis. We thank them for allowing us to explain our STAR Grid program directions and interests and for further joining trusting and supporting our efforts by joining the BNL team.

2.9.5.2 STAR DDM Addendum & Infrastructure changes

In the last quarterly report, we reported on network performance and progress report. In section 2.10.6, we reported a 10.6 MB/sec for 14 sessions maximum performance between NERSC/LBNL and BNL while Dantong Yu clearly measured a 30 MB/sec data transfer network performance to Oklahoma. Many thanks to Shane Canon who has investigated this inconsistency, several problems were found on the PDSF side amongst which, transfer congestions on pdsfgrid2, out of order and dropped packets at the NERSC border router. We hope to eliminate those problems but also to be able to upgrade the two side Gatekeepers with the Web100 kernel patch allowing for peers to self-optimize and further expanding our Grid gateways.

Dantong Yu helped us with the upgrade of stargrid01 to Linux 7.2 we can now use in conjunction with LSF 4.2 . Using `globus-job-submit` and the LSF job manager, the NCG at Stony Brook (Phenix Collaboration) successfully submitted jobs on the RCF cluster.

2.9.5.3 Deployment and tests of HRM-based File Replication

Use of HRM and GridFTP for production data movement between RCF/BNL and NERSC/LBNL continues to work well. Some network issues at NERSC were observed and diagnosed that prevent achieving theoretical limits on data transfer but the achieved values are meeting STAR's needs so fixing these has not yet become a high-priority item. Typical usage is to submit requests to move up to 1000 files (100's of GB) and the system runs without interruption for many hours to completion. A demonstration of these data transfers was shown at SuperComputing 2002.

In the coming quarter the STAR file-metadata catalog will be installed at NERSC and the file replication will be integrated with this catalog. It is planned to evaluate the Globus Replica Location Service (RLS) for use as part of this data replication service.

2.9.5.4 Monitoring

The Ganglia monitoring system has been set up at RCF/BNL and NERSC/LBNL (thanks to respectively Jason Smith and Shane Canon) to provide system load and performance information for individual nodes as well as summaries for the entire clusters. It is being evaluated how to utilize this information with the job scheduling system so that choices of location for job execution can be made to better optimize resource utilization. We also hope to finalize the effort of propagating the Ganglia information through MDS, our new team members have shown interest in evaluating this approach as a startup project and report in the next quarter.

2.9.5.5 Job scheduling

The job scheduling system for STAR is being extended to submit jobs via Condor-G. The system was originally developed to allow for scheduling jobs taking into account the location of data so that jobs execute on nodes with the data is resident. This job submission was directly to an LSF batch queue. By extending this system to submit jobs to Condor-G it will be possible to submit jobs that will run at one of two sites, RCF/BNL or NERSC/LBNL.

In November a meeting was held at LBNL with attendance from the STAR BNL and LBNL groups, and the LBNL Scientific Data Management (SDM) group. The purpose was to discuss co-scheduling in a shared-nothing cluster and possible integration efforts between the STAR scheduler and a research project of the SDM group on co-scheduling in a shared-nothing cluster that uses Condor for the scheduling system. It was estimated that some integration work may be feasible in early CY2003.

2.10 CS-11 Grid Interface with Interactive Analysis Tools

The group at SLAC worked towards a Grid-Enabled version of the Distributed JAS analysis system. Demonstrated this system at SC2002 with identical versions of the demonstration run at the SLAC/Fermilab booth and at the Sun booth, staffed by personnel from SLAC and from SLAC's SBIR partner, Tech-X Corporation. Jobs controlled by the JAS analysis front-end using SRB to locate data were distributed to grid nodes at two different locations, one being SLAC, the other being a temporary grid farm set up on a cluster of Linux machines at the Sun booth. The SLAC/Fermilab booth also included a demonstration of PROOF. See http://sc2002.slac.stanford.edu/10.htm.

In phone meetings, CS11 participants talked with David Adams about his excellent paper on Dataset specifications, http://www.usatlas.bnl.gov/~dladams/dataset. This is the best formalization of data sets we've seen thus far and fills a need identified in our paper Grid Service Requirements for Interactive Analysis, http://www.ppdg.net/pa/ppdg-pa/idat/papers/analysis_use-cases-grid-reqs.doc.

Presented a PPDG Analysis Tools day as part of the Trillium Joint Session at Caltech Dec 19th. See http://www.ppdg.net/mtgs/19dec02-cs11/index.html.

First part of the day was a series of prepared talks on analysis topics.

Second half of the day was a discussion on how to move forwards. Participants were encouraged by the manner in which the different tool makers who participate in the AIDA project had managed to find common ground to work together and to make their tools interoperate through a set of shared interfaces.

Enough tool makers expressed interest in searching for common ground in the CS11 realm that it seemed worthwhile to try to organize a follow-up meeting of interested tool developers to facilitate identification and specification of common interfaces.

The two major CS11 meetings last year (one at LBL, the other at Caltech) accomplished the initial steps of familiarizing ourselves with the existing tools and projects and setting forth a set of requirements.

The plan forward is to convene a meeting of analysis tool makers at which:

a) everyone is assumed to have already read background material on one another's tools

b) everyone brings their own component diagram that works for their tool

c) a full day or two is spent looking for what components we have in common and starting to define some of the interfaces.

It was suggested that this meeting might make sense in San Diego immediately before CHEP at the end of March. This could be followed by a birds of a feather get together at CHEP.

2.10.1 Clarens – Distrbuted Analysis

An analysis display was created using Root to display CMS data produced as part of the Fall 2002 production process on the so-called Integration Grid Testbed (IGT). For the conference an extra step was added to the production process to convert the Ntuple data files to Root files. Files from Caltech, the University of Florida, Fermilab and UC San Diego were remotely analyzed through Clarens servers installed at these sites. The Root package was Clarens-enabled using loadable modules produced as part of the Clarens project. On the show floor a live display of the analysis was shown as new data files were being created at the remote sites. Results of the demonstration will be published in the Computing in High Energy Physics 2003 conference volume as part of the IGT contribution.

An application of distributed analysis was demonstrated for the case where the analysis was completed at remote sites, and the results were aggregated on the show floor. CMS event data was stored at the Starlight

POP in Chicago and at Caltech in relational databases, and data queries were passed to the databases through a Clarens server running at each site. The analysis results were produced in the form of Root files using the Sql2Root package. The files were browsable from the show floor through the Clarens-enabled Root client.

A talk entitled Experience with Clarens SC demo was presented. Briefly the SC2002 demo showed that installation and configuration of the Clarens server needs to be simplified, that the Root client should be protected from corrupt files at the remote sites since these caused crashes. Since the CACR booth was so equipped, an attempt was made on the show floor to port the Clarens-Root modules to the Intel IA-64 (Itanium) platform, which was largely successful, with the exception of the http library (libcurl) which failed in unpredictable ways. Consequently the CACR booth demo used a Pentium III-based laptop exporting an X11 display to the IA-64 machine. Both used the Red Hat Linux OS. Most importantly, the Clarens servers exhibited no crashes or other problems despite being accessed constantly by two client machines over a period of five days.

2.10.1.1 Clarens web service layer client and server developments

A version of the server was developed that uses the well-known SOAP protocol instead of the XML-RPC protocol used so far.

A more complete security implementation including integrated Virtual Organization (VO) management and stricter X-509 certificate chain verification, was developed. This implementation is available as a module replacing the standard Clarens security setup if installed. The module implements as a hierarchical structure of access control lists (ACLs) with separate domains for RPC methods, files, VOs (i.e. user and group administration is controlled by ACLs). This provides a poweful, yet easy to administer alternative to the standard grid-map file concept. ACL verification is done via a so-called ternary-tree structure providing fast look-ups in the code's critical path.

Proxy escrow service implemented, similar to the MyProxy service, see http://www.ncsa.uiuc.edu/Divisions/ACES/MyProxy/ . Briefly, this enables a user to store a proxy certificate and private key in a database on the Clarens server, which can be retrieved from any machine. The proxy credentials are password-protected and enrypted using the RC4 stream cipher.

Job submission service implemented. Jobs can be submitted as users of the host system, and the standard output and standard error messages can be retrieved asynchronously while the job is running. Access to user accounts is controlled by a separate ACL domain, allowing flexible group and user mapping to individual user accounts on the host system. Note that this is meant mainly to submit commands to more complex batch schedulers, not replace these schedulers.

As always, the client and server implementations were further polished and releases was made on the Clarens web pages at http://clarens.sf.net for public download. Documentation for server installation, Python client access, Root client access, as well as server extension module writing was added to the above location.

2.10.2 ATLAS DIAL

Most of the pieces needed for the first implementation of DIAL (Distributed Interactive Analysis of Large datasets) were put in place during the quarter. In addition to all the supporting components, these include a local scheduler which can be used to run jobs on the local machine.

DIAL provides a connection between an analysis framework and a data-processing application. Our first choice for the former is ROOT and we plan to integrate DIAL with ROOT in the next quarter.

The obvious data-processing application in ATLAS is Athena. ATLAS plans to move much of its data into ROOT files and so ROOT itself might be a candidate for this application. Most of the terabytes of DC1 data generated during the quarter are in zebra format so atlsim might also be a candidate. During the next quarter, we plan to choose one of these and integrate it to provide users with the DIAL-based capability to directly analyze ATLAS data from ROOT.

The next step will be to provide a distributed scheduler which makes it possible to carry out an analysis task on multiple nodes at a site. After this, the plan is to GRID-enable DIAL (or more precisely the DIAL scheduler) so that it can locate and process data over the GRID.

2.11 CS-12 Catalogs and Databases

2.11.1 STAR file metadata catalog

The STAR file metadata catalog was modified slightly as part of the integration work with job scheduling. The mechanisms for database replication in MySQL are planned to be used for implementing a distributed version of this catalog. It is planned to install the catalog at NERSC/LBNL in the next quarter and synchronize between BNL and LBNL using the built-in MySQL replication features. The evaluation of the integration of this meta data catalog with the Globus RLS will be later pursued.

2.11.2 SDSC – SRB

Discussions were held with Daresbury Laboratory in England about use of the SRB technology in the UK data grid. They are interested in gaining access to BaBar data. The UK data grid provided a list of requirements requiring new capabilities within the SRB, including:

- Collection ownership changes – (version 2)
- Checksum verification of uploads – (version 2)
- MCAT replication and collection management – (version 2)
    - Access rights replication
    - Replication of data only supported by the receiving MCAT database
    - Use of existing native database functionality
- Encrypted data transfers - (version 2.1)
- Disjoint access permissions between metadata and data - (version 2.1)
- Simplified installation - (version 2.1)
- Federation of MCAT collection catalogs - major development for summer
- Modularization of the SRB
    - Access to Spitfire
    - Giggle replication management
    - Resource description using XML/WSDL
- Support for a “non” permission for groups to restrict access
- Support for access to a CVS repository

Of these requirements, the highest priorities are the release of version 2 of the SRB, and the federation of MCAT collection catalogs. SRB version 2 includes capabilities that are already developed. The creation of version 2 requires the integration of the capabilities into a common release, which is scheduled for February 2003.

The federation of MCAT collection catalogs is needed to improve the ability of the BaBar experiment to manage access to data distributed between Stanford and Lyons. The development of the ability to federate collections is being designed in collaboration with the UK data grid, the NPACI distributed teragrid facility, and the Globus grid team. Two designs are under consideration, creation of mount points to access remote collections, and creation of publication links to point to data that has been linked from a remote collection to coordinate metadata.

3 Single Collaborator Reports

3.1 ANL – Globus

3.1.1 Coordination and Support

Continuing interactions in terms of coordination and support of the PPDG applications included weekly phone meetings with Atlas and CMS, SuperComputing (and it's planning activities), and participation in the trillium meetings in southern California in December. In addition, we contributed to the joint Trillium architecture document for Grid execution planning.

3.1.2 Globus Toolkit updates and bug fixes

This quarter we released numerous bug fixes, and version 2.2.5 can be downloaded at our website . We closed 11 bugs listed in Bugzilla (82, 230, 246, 256, 257, 291, 324, 328, 370, 394, 480) and have only 4 open PPDG-related bugs still open in our system (260, 398, 347, 542). Additional information about Bugzilla bugs can be found at http://bugzilla.globus.org .

3.1.3 Globus Toolkit 3.0

Work continued on the development of the OGSA-based Globus Toolkit 3 (GT3) implementation. Technology preview 4 was released on Oct 31, 2002, followed by Technology Preview 5 on December 14, 2002. Information is available here http://www.globus.org/ogsa/releases/TechPreview/index.html .

The GT3 alpha code will be released as part of GlobusWORLD on January 13, 2002. Further information on GT3 can be found here http://www.globus.org/ogsa/.

3.2 SDSC – SRB

The activities at the San Diego Supercomputer Center in support of the PPDG continue to be focused on collection management, and support for sharing of data. Explicit activities include:

Support for the BaBar experiment. SRB servers are installed at both Stanford and Lyons, France. Details of the requirements and consequent Improvements planned and Implemented In SRB are

- Support for a non-proprietary database. We had many requests from PPDG for the ability to run the SRB metadata catalog on a non-proprietary database. We chose to implement a port to PostgresSql, in collaboration with NASA. The port has been completed and tested for functionality and performance. The performance is slower than Oracle, but the functionality is the same. This version of the MCAT system will be released with version 2.0 of the SRB in February.

- Development of web services. The creation of WSDL and OGSA based access mechanisms to data and collections is of interest to PPDG, GriPhyN, NPACI DTF, NSF SCEC, and NASA IPG projects. Using funding from the SCEC project and the GriPhyN project, the development effort has accelerated. The immediate driver is the development of a WSDL interface for the data access mechanisms required by the NPACI Grid Portal. A project is being initiated to identify the common services between the NPACI Grid Portal, the UK data grid, and the PPDG.

- The testing of the GridFTP driver interface to the SRB is still being done. Version 2.2 of GridFTP has been installed at SDSC, and will be tested in December 2002. The goal is to demonstrate the ability to move 1000 files consecutively without failure. We will then look at performance issues, comparing the access times between going through the SRB data transport and going through the GridFTP data transport. Of particular interest is the comparison of parallel I/O based access to archives and the performance of server-initiated parallel I/O.

4 Appendix

4.1 List of participants

TEAM	Name	F	Current Role CS	1	2	3	4	5	6	7	8	9	10	11	12
Globus/ANL	Ian Foster	Y	Globus Team Lead, GriPhyN PI, iVDGL, GriPhyN						x	x					
	Mike Wilde	N	GriPhyN coordinator					x					x		
	Jenny Schopf	Y	GriPhyN, iVDGL, Globus team liason, ATLAS-CS liason			x				x	x		x		
	Willliam Alcock	Y							x		x		x		
	Von Welch		CAS									x			
	Stu Martin				x								x		
ATLAS	John Huth	N	ATLAS Team lead										x		
	Torre Wenaus	N			x			x							
	L. Price	N	Liaison to HICB, HICB Chair												
	D. Malon	N													
	A. Vaniachine	Y													
	E. May	N						x					x		
	Rich Baker	N													
	Alex Undrus	Y													
	Dave Adams	Y													
	Wengshen Deng								x						
	G. Gieraltowski	Y									x		x	x	
	Dantong Yu	Y	Monitoring			x							x		
BaBar	Richard Mount	N	PPDG PI, BaBar Team co-Lead												
	Tim Adye	N	BaBar Team Co-Lead												
	Robert Cowles	N										X			
	Andrew Hanushevsky	Y						x	X						
	Adil Hassan	Y						x	X						
	Les Cottrell	N	IEPM Liaison			X									
	Wilko Kroeger	Y						x	x						
CMS	Lothar Bauerdick	N	CMS Team Lead. GriPhyN collaborator												
	Harvey Newman	N	PPDG PI. GriPhyN collaborator, Co-PI iVDGL												
	Julian Bunn	N	CMS Tier 2 manager, GriPhyN & iVDGL collaborator										x	x	
	Conrad Steenberg	Y	CS-8:Analysis Tools, GriPhyN collaborator								x			x	
	Iosif Legrand	N	CS-8:Monitoring Tools								x				
	Vladimir Litvin	N	GriPhyN collaborator	x	x										
	James Branson	N	CMS Tier 2 manager										x		
	Ian Fisk	N	CMS Level 2 CAS manager, iVDGL liaison										x		
	James Letts	Y	Working on VDT testing scripts										x		
	Eric Aslakson	Y	job execution, grid monitoring		x	x									
	Edwin Soedarmadji	N	web services prototyping											x	

	Saima Iqbal	N	web technology evaluation											x	
	Suresh Man Singh	N	grid deployment										x		
	Anzar Afaq	Y			x								x		x
	Greg Graham	N			x								x		x
Coordination	Ruth Pordes	Y	PPDG coordinator							x					
	Doug Olson	Y	PPDG coordinator				x			x		x		X	
	Miron Livny	Y	PPDG coordinator	x	x		x	x				x			
D0	Lee Leuking	N	SAM Project co-manager		x	x									
	Igor Terekhov	Y	JIM Team Lead	x	X	x									
	Andrew Baranovski	Y				X									
	Gabriele Garzoglio	Y		X	X	X			X						
	Siddharth Patil	Y	Through contract with UTA CSE Department	X	X	X									
	Abhishek Rana	Y	Through contract with UTA CSE Department	X	X	X									
	Hannu Koutaniemi	N	FNAL Finnish coop Student	X	X	X									
	John Weigand	Y			x										
HRM/LBNL	Arie Shoshani	y	SRM Team Lead. GriPhyN collaborator				X								
	Alex Sim	Y					X								
	JunminGu	Y					X								
	Alex Romosan	Y					x								
SRB/UCSD	Reagan Moore	Y	SRB Team Lead. GriPhyN collaborator					X		X	x				
	Bing Zhu	Y	CS-8: Web Services					x			X				
JLAB	William Watson	Y	JLAB Team Lead				x	x	x		x				
	Sandy Philpott	N	facilities									X	X		
	Andy Kowalski	N					X								
	Bryan Hess	Y	Web Services				x				X				
	Ying Chen	Y	Web Services				X		X		x		X		
	Walt Akers	N	Web Services					x					X		
STAR	Jerome Lauret	N	STAR Team Lead		x						X		x		
	Gabrielle Carcassi	Y			x								x		
	Dave Stampf	N				X							X		
	Richard Casela	N				X							X		
	Eftratios Efstathiadis	N				X							X		
	Eric Hjort	Y					x	x					x		
	Doug Olson	N					X	X			X		X		
Condor/U.Wisconsin	Miron Livny	Y	PPDG PI, PPDG Coordinator. GriPhyN collaborator	x	x	X	x		x		x				
	Paul Barford	Y				X									
	Peter Couvares	Y		X	x							x			
	Rajesh Rajamani	N			x						X				
	Alan DeSmet	Y			x							x			
	Alain Roy	N			x										

	Todd Tannenbaum	N			X										
Globus/ISI	Carl Kesselman	N	Globus/ISI lead												
	Ann Chervenak	Y							x						
SiteAAA (if not listed above)	Dane Skow	Y	FNAL, Co-Lead									x			
	Bob Cowles		SLAC, Co-Lead												
	Booker Bense	Y	SLAC									x			
	Tomasz Wlodek	Y	UT Arlington									x			
	Igor Mandrechenko	Y	FNAL									x			
	Steve Chan,		LBNL									x			
	Steve Lau		LBNL									x			
	Craig Tull		LBNL									x			
	Matt Crawford		FNAL									X			
	David Bianco		JLab									X			
	Hao Wang		U Wisconsin									X			

CS-1 Job Description Languages
CS-2 Job Management and Scheduling
CS-3 Information Services
CS-4 Storage Management
CS-5 Reliable File Transfer
CS-6 Robust File Replication
CS-7 Documentation
CS-8 Evaluations and Research
CS-9 Authentication and Authorization
CS-10 Experiment Grids and Applications
CS-11 Analysis Tools
CS-12 Catalogs

4.2 Meetings

Wednesday, October 2, 2002

PPDG weekly phone meeting

URL: http://www.ppdg.net/mtgs/phone/021002/default.htm

Wednesday, October 9, 2002

PPDG weekly phone meeting

URL: http://www.ppdg.net/mtgs/phone/021009/default.htm

Wednesday, October 23, 2002

PPDG weekly phone meeting

URL: http://www.ppdg.net/mtgs/phone/021023/default.htm

Wednesday, October 30, 2002

PPDG steering committee meeting

Wednesday, November 6, 2002

PPDG weekly phone meeting

URL: http://www.ppdg.net/mtgs/phone/021106/default.htm

Wednesday, November 13, 2002

PPDG weekly phone meeting

URL: http://www.ppdg.net/mtgs/phone/021113/default.htm

Wednesday, November 27, 2002

PPDG weekly phone meeting (cancelled)

Wednesday, December 4, 2002

Storage Resource Manager Workshop, CERN

PPDG weekly phone meeting

URL: http://www.ppdg.net/mtgs/phone/021204/default.htm

Friday, December 6, 2002

PPDG Analysis Tools Interface phone meeting

Wednesday, December 11, 2002

TroubleShooting Workshop

URL: http://www.ppdg.net/mtgs/Troubleshooting/agenda.htm

Friday, December 13, 2002

Phone mtg, Re-scheduled BaBar-grid update

URL: http://www.ppdg.net/mtgs/phone/021213/default.htm

Monday, December 16, 2002 through Wednesday, December 18, 2002

GriPhyN, iVDGL, PPDG at ISI

URL: http://www.ivdgl.org/events/view_agenda.php?id=3

Thursday, December 19, 2002

PPDG meeting at Caltech

URL: http://www.ivdgl.org/events/view_agenda.php?id=3

Grid Interface to User/Analysis Tools session, Caltech

URL: http://www.ppdg.net/mtgs/19dec02-cs11/

PPDG Site-AA Caltech

URL: http://www.ppdg.net/mtgs/19dec02-siteaa/

Friday, December 20, 2002

PPDG steering committee meeting at Caltech