Particle Physics Data Grid Collaboratory Pilot

Quarterly Status Report of the Steering Committee, July - September 2001

31 Oct. 2001

Contents

1. Project Overview 2
   1.1. Project Management and Organization 2
   1.2. Project Web Pages 2
   1.3. End to End Applications and Deployed Systems 2
   1.4. Common Services Development and Integration 2
   1.5. Interactions with other Projects and Activities 3
   1.6. Current Issues and Concerns 3
2. Project Activities 4
   2.1. GDMP (CMS-DataGrid-Globus) 4
   2.2. D0 Job Management (D0-Condor) 4
   2.3. CMS-MOP (CMS-Condor) 5
   2.4. STAR-DDM (STAR-LBNL) 5
   2.5. JLAB-Replication (JLAB-SRB) 5
   2.6. ATLAS distributed data manager, MAGDA (ATLAS-Globus) 6
   2.7. BaBar Database Replication (BaBar-SRB) 7
3. Single Collaborator Efforts and End to End Applications 7
   3.1. ATLAS 7
   3.2. BaBar 8
   3.3. CMS 9
   3.4. D0 10
   3.5. Jlab 11
   3.6. ANL – Globus 11
   3.7. NERSC – SDM 11
   3.8. SDSC – SRB 12
4. Appendix 15
   4.1. SuperComputing 2001 demonstrations related to PPDG 15
   4.2. Draft INTERGRID charge 15

1. Project Overview

Each experiment now has at least one Project Activity in collaboration with a Computer Science group. Reports from each are included in section 2. Some further understanding of the scope and relation of the Project Efforts was achieved. The cross-project work on Certificate Authority and Policy issues got underway. A PPDG Collaboration meeting was held, hosted by the University of Wisconsin (http://www.ppdg.net/mtgs/14aug01-wisc/agenda.htm), and was attended by about 45 people. A collaboration with GriPhyN on related demonstrations for CMS simulation production for SC2001 was started. There are a number of other demonstrations at SC2001 related to PPDG and these are listed in the appendix.

A decision was taken on the project logo – as shown at the top of this report.

The PPDG project plan covering first year goals (roughly FY2002) is finalized and posted at http://www.ppdg.net/docs/PPDGYear1.pdf.

1.1. Project Management and Organization

Project effort reports and briefings have been enhanced and there has been a significant increase in their use for this quarterly report. This input will be of value when we review actual scope and deliverables at the end of the year. The surveys and documents requested by the SciDAC management were delivered (http://www.ppdg.net/docs/SciDAC/) and the Project Management Plan was sent to the DOE project sponsors. Members of PPDG management attended the High Energy and Nuclear Physics Intergrid Coordination Board meeting in Rome on October 7th. The current draft of the groups charge is attached as an appendix to this report. Three members of PPDG - Peter Couvares, Doug Olson, Ruth Pordes – were appointed to the Joint Technical Board by the Coordination Board.

1.2. Project Web Pages

The Project Web pages were enhanced and information added. We are adding information to the documentation web pages to include PPDG related papers and presentations, white papers and reports. More than 90% of the collaboration have submitted their update briefings http://www.ppdg.net/pipermail/effort-reports/2001/ and team effort reports for this quarterly report. This is a significant increase over the last call for submissions.

1.3. End to End Applications and Deployed Systems

All experiments took steps towards one of the main PPDG goals of deployed end to end applications using a DataGrid. About four of the PPDG funded FTEs worked on the data replication infrastructure and deployment – high throughout, robust and transparent. The D0 SAM system is the most advanced in terms of an integrated data replication and distributed system, the other five experiments made progress towards this goal. Discussion started on differences in the infrastructures being developed and deployed. This will continue at a full day focused meeting planned for January at TJNAF. Atlas and CMS have distributed simulation production systems in test at several sites. BaBar continues to develop its architecture of different strategies for intra-site and inter-site replication.

1.4. Common Services Development and Integration

1.4.1. Data Replication and Catalog Services

Reusable data replication and catalog services focus on the Globus and SRB toolkits, together with the GDMP file and object database replication layered above the Globus replica catalog and GridFTP transport service. GridFTP was released in the Globus 2 Alpha releases, GDMP V2.0 Alpha was released integrated with the Alpha release of the Globus replica catalog.

CMS has deployed GDMP/Globus across sites in Europe and the US. ATLAS is moving to integrating these services into its existing prototypes. STAR will deploy a site-to-site data replication using GridFTP

and then evaluate GDMP. Jefferson Lab and SRB are looking to develop a common web services interface to replication services, to be implemented atop SRB and JLab's JASMine. BaBar is working to integrate the SRB data replication services into their experiment system. BaBar has published an application layer file transfer tool (bbcp) which incorporates novel features for high end-to-end throughput (ie between disk blocks) and robustness.

1.4.2. Job Scheduling and Management

Job Scheduling and management using the existing versions of Condor-G and Condor are in test by CMS, ATLAS and D0-SAM. Ideas of how some of the Condor facilities might be made available at the Condor-G layer of the architecture have been fed back to the Condor group. The DAGMAN job dependency service is starting to be used in some of the PPDG applications in collaboration with the GriPhyN project's Virtual Data Toolkit V1.0.

1.4.3. Storage Resource Management

STAR is the first site to use the new DRM implementation from the LBL SRB group. This is providing good feedback to the common services development team. The LBL and JLAB groups are collaborating with FNAL and European Data Grid WP5 to arrive at a common HRM interface across the projects.

1.4.4. Monitoring and Status Reporting

Techniques and services for network monitoring are improving through work at SLAC in collaboration with the PingER project and the BNL ATLAS group. Requirements for monitoring and a catalog of existing monitoring services is being undertaken as a joint project with GriPhyN in a group chaired from the Globus and ATLAS teams.

1.5. Interactions with other Projects and Activities

One of the key interactions with other projects is the GDMP project activity, described below, which is supported by EU DataGrid and CMS as well as PPDG.

PPDG continues to spend time and resources on inter-working with peer projects in the field. Many papers and talks related to PPDG work (http://www.ppdg.net/docs/chep01_abstracts.htm) were given at the Computing in High Energy Physics (CHEP01) conference in Beijing (http://www.ihep.ac.cn/~chep01/).

There have been discussions with DOE Science Grid and ESnet personnel on the topic of PKI and possibility of a certificate authority that could issue certificates for PPDG participants. A recent decision by DOE and ESnet that ESnet can operate such a certificate authority should be of great benefit to PPDG and the experiments and groups participating in PPDG. Details of how this works out and becomes implemented will be available in the next quarterly report.

1.6. Current Issues and Concerns

While overall the project developments are going well, practical means of meeting our longer term goals of having common and reusable software are still not clearly planned. The lack of resources for cross project activities and funding is seen as potentially reducing our ability to persuade in this direction, and is the subject of Steering Committee discussion.

There is a current, hopefully very short term issue, to encourage the collaboration to avoid having completely out of date and divergent PPDG web pages.

There has always been overlap in deliverables and effort between GriPhyN, iVDGL and PPDG. Coordination amongst the projects is going well, and there is progress towards ensuring that there will not be duplication of effort. As a result, however, identifying which component deliverables are the result of PPDG funded efforts per se is becoming a complex matrix.

Some sites are still ramping up on their project funded effort. It is expected that hiring can be completed and the project will be operating at full strength by the end of 2001.

2. Project Activities

The major part of the PPDG work is organized as Project Activities (http://www.ppdg.net/pa/ppdg-pa/projects.htm). These are formed as working teams of Experiments and Computer Science Groups – with an identified Project Leader and a Liaison/Coordinator from one experiment and one Computer Science Group (where the project leader may or may not be the same person as one of these). The goal is that each project defines deliverables, milestones and necessary effort to allow them to work semi-independently towards their goals. PAs report technical progress and details at the regular bi-weekly phone conferences.

2.1. GDMP (CMS-DataGrid-Globus)

The main effort of this quarter was to produce a new release of GDMP for production as well as for use in the European DataGrid test-bed in autumn 2001 and the MOP demo at SuperComputing 2001. Major effort has been done to finalise the release for GDMP 2.0alpha for 30 September 2001. The software has been tested extensively and is now ready to use in a production environment. The GDMP web page at http://cmsdoc.cern.ch/cms/grid has been redesigned and all the code as well as detailed documentation for GDMP 2.0alpha are available on that web site. Having finalised the release on time means that GDMP is a major part of the European DataGrid software as well as for production use in physics experiments within PPDG (in particular CMS). Interactions with the DataGrid Integration team (WP6, DataGrid) make sure that the software is well integrated into the entire project.

Starting with the new release, the software package has now a slightly new name: instead of Grid Data Management Pilot ("pilot" was okay for the beginning of the project phase) we renamed it to Grid Data Mirroring Package (GDMP).

Based on several fruitful discussions with colleagues in PPDG, GDMP is extended by a notification system that notifies local and remote sites about successful file transfers. This notification system can further be extended to a fully automatic replication system where replication can be triggered by a production site: as soon as files are published, a remote site (a consumer) gets notified and thus can immediately trigger file transfers. In addition, the notification system also allows for some additional disk space management: since a site is notified when files are transferred from a remote site, local files can be deleted when they have arrived safely at a remote site. For details refer to the GDMP User Guide version 2.0alpha on the GDMP web page.

In addition to the notification system, file transfer states have been introduced that allow to monitor and check several sites in the file replication process. These transfer states include successful file transfer, registration, notification, insertion into file and replica catalogues etc. For details refer to the GDMP User Guide.

An early version of GDMP 2.0alpha has been used in the PPDG-CMS MOP project.

The GDMP team prepared an informational document about the status and features of GDMP to be presented at the Global Grid Forum (GGF3) in Frascati, Italy, October 2001. (see http://cmsdoc.cern.ch/cms/grid and go to "Documents").

Asad Samar has left High Energy Physics and cannot participate actively to GDMP anymore. We all want to thank him for his great effort!

2.2. D0 Job Management (D0-Condor)

D0 has continued to extend, deploy and support SAM for data taking for the D0 community and to work with the remote sites on integration of SAM with their local mass storage and fabric systems. Until additional manpower is hired the amount of effort available for PPDG funded work has been minimal. As part of the D0 /Condor Job Management project Imperial College have interfaced SAM to Condor and are running D0 Monte Carlo jobs on a local Linux cluster, with data sent back through SAM to the Fermilab mass storage system.
Modifications have been made to the D0 monte carlo framework to store meta data at intermediate stages in the total job pipeline and prepare the framework for parallel job processing and checkpointing using

Condor. DAGMAN was installed on a D0 development machine at Fermilab as well on collaborators sites at NIKHEF and ICL

2.3. CMS-MOP (CMS-Condor)

The past quarter of the MOP project has been devoted primarily to testing, integration and deployment. Some parts of the CMS production chain have been run through MOP at Fermilab, the University of Wisconsin, the University of California, San Diego and Caltech. The entire chain has been run at Fermilab. We are preparing a demonstration of the MOP system at Supercomputing 2001.

The integration of MOP with the CMS production system, recently officially named IMPALA, was implemented by James Amundson. The IMPALA code is available from the Fermilab Computing Division CVS repository (FNAL-CDCVS), :pserver:anonymous@cdcvs.fnal.gov:/cvs/cd_read_only, in the module cms_prod_util. It can also be viewed on the web at <http://cdcvs0.fnal.gov/cgi-bin/public-cvs/cvsweb-public.cgi/>. The MOP modifications to the IMPALA code currently live in the branch mop-branch. We plan to merge the branch back into the main trunk when the testing phase is completed.

The job submission portion of MOP, mop_submitter, now uses Condor DAGMan for robust and restartable job staging, running and publishing. DAGMan is also used to track publishing dependencies between multiple Objectivity jobs. mop_submitter continues to use Condor G for remote job submission and control. This work was done by James Amundson with help from Peter Couvares and the rest of the Condor team. The mop_submitter code is not CMS specific and could potentially be used outside of MOP. The mop_submitter is also available from FNAL-CDCVS. The repository also contains a module named mop_master, which contains the very small amount of glue needed to set up and install IMPALA and mop_submitter together in addition to some installation documentation.

Substantial integration work has been done at all of the MOP sites. Shahzad Muzaffar has provided GDMP installation and integration support for all sites. At Fermilab, James Amundson and Greg Graham have worked on local issues with the assistance of the rest of the CMS department. The CMS software is being packaged for distribution by Natalia Ratnikova. The University of Wisconsin installation was handled by Rajesh Rajamani and Peter Couvares, along with the rest of the Condor team. Installation work at UCSD was done by Ian Fisk. Caltech installation was handled by Suresh Singh, Koen Holtman and Takako Hickey. Some of the MOP installation can be easily accomplished remotely; these installation steps were done by James Amundson.

2.4. STAR-DDM (STAR-LBNL)

In this quarter HRM version 3.0 was released. It includes an enhanced Tape Resource Manager (TRM) component that can perform both "reads" and "writes" from/into HPSS systems. The TRM can queue requests if HPSS is not accepting PFTP requests, and when writing files TRM generates a second call-back to notify the client. DRM version 1.1 was also released and is integrated into HRM 3.0. This is a completely new DRM component that gives HRM the same behavior as DRM along with the capability of reading and writing to HPSS with TRM. API's for HRM were also released both as CORBA IDLs and as C++ API's.

Utilization of HRM 3.0 requires installation of the latest Globus software in both sites, the Orbacus CORBA ORB, and the correct version of compilers. These upgrades to our present systems are in progress, and will be followed by installation of the HRMs at BNL and at LBNL along with a "Simulated Command-line Client" that is available for testing the system. In the meantime Globus 1.1.3 has been installed at both sites and STAR-DDM members have learned to use it and have transferred data between the sites using gsincftp.

2.5. JLAB-Replication (JLAB-SRB)

During the 3 month period July-Sept 2001, a modest amount of progress was made on the PPDG/Jlab-Replication task, much of it as independent efforts at Jefferson Lab and SDSC/SRB to move towards an XML based interface to data management systems. The two sites have also exchanged some information

on the meta data being used in their respective systems so that some analysis work can be done to arrive at a common schema.

Jefferson Lab is currently using java servlets to implement web services carried over raw XML messages (this will be converted to SOAP during the next reporting period). This interface has been refined and improved to support an advanced remote data management client, implemented as a Java application and deployed (with automatic updates) using Java Web Start[1]. SDSC is doing their prototyping work using web forms to gather information from the user, with one cgi script converting that to an XML query, and another cgi script processing that query and returning an XML structured result. Currently the two sites do not have a common XML structure, but these prototyping exercises will provide the necessary experience to enable a good common design to emerge.

Both sites have agreed to participate in a testbed in which a common web services interface to two disparate system (SRB and JASMine) can be demonstrated. This activity will also involve the Global Grid Forum Data Working Group, and most likely a soon-to-be-formed Web Services working group.

2.6. ATLAS distributed data manager, MAGDA (ATLAS-Globus)

Development of the Magda (formerly DBYA) distributed data management system continued. Magda is being developed to fulfill the principal ATLAS PPDG deliverable for year 1, a production distributed data system deployed to users.

Several enhancements were made to the file and replica cataloging in Magda. A Globus replica catalog loader was developed to migrate the Magda replica catalog content to Globus and evaluate, but it remains to be tested. Scalability tests of Magda cataloging were done; catalog size was increased from the current stable count of 160k up to ~400k and then up to 1.5M cataloged files. After minor bugs were fixed the system performed well at 1.5M files with a lookup performed on the entire catalog taking ~30sec. Input was given to the replication requirements document based on Magda and earlier experience. Support for several types of file collections was added.

Support for file replication between distributed sites was added to Magda. The Globus gsiftp tool is used for replication among US ATLAS grid testbed sites, while scp is used at the moment between CERN and BNL. A multi-stage automated process moves a file collection (in the most complex case) out of a source-side mass store into a cache, over the network into a destination cache, and into a destination mass store. The system has so far been used to replicate ~100GB of ATLAS simulation data between CERN and BNL, and small volumes have been replicated to other sites. Cataloging and replication were extended to support the Castor mass storage system at CERN.

An 'SQL accelerator' was developed and integrated into Magda to expedite processing of MySQL commmands from remote client sites. SQL commands are accumulated on the client side and dispatched in bulk to the database as an SQL text blob, which is processed on the server side by a script triggered (via HTTP) by the client. This eliminates per-command network latencies and speeds up bulk catalog operations over WANs by orders of magnitude. With the accelerator, cataloging 1.5M files over a WAN was shown to be practical.

Deployment of Magda was extended beyond BNL and CERN to ANL and LBNL, and partially to Boston University.

Development plans for Magda were coordinated with PPDG, GriPhyN and the CS projects at GriPhyN and PPDG collaboration meetings in August. Jennifer Schopf now acts as liaison with the CS projects. The description and documentation of the system was improved. Further information (the documentation page) and a talk is available at http://atlassw1.phy.bnl.gov/magda/info The system itself is at http://atlassw1.phy.bnl.gov/magda/dyShowMain.pl

Near term plans include completion of command-line tools providing a file access interface to production jobs; tools to monitor throughput and gather statistics in a production environment; ATLAS framework (Athena) integration; further integration of Globus tools (remote command execution, replica catalog);

[1] http://java.sun.com/products/javawebstart/

exploration of other data movers (GDMP, bbcp); and application and testing in ATLAS Data Challenges commencing in December. Discussions on the application of Magda within the ATLAS Data Challenges began during the period. Development of a DC production scenario for simulation data using Magda also began during the period.

Prototype scenarios for grid-enabled data access from Athena, the ATLAS experiment's control framework, were investigated. Two approaches, in particular, were explored, one involving registration of files containing event collections with the Globus replica catalog, the other involving use of GDMP 1.2.2. The latter approach was exercised on EU Data Grid testbed nodes in Geneva and Milan by Silvia Resconi, using the ATLAS fast simulation program Atlfast running under Athena, with the object database product Objectivity/DB as the underlying storage technology. This work was described at the CHEP'01 conference in Beijing.

2.7. BaBar Database Replication (BaBar-SRB)

The BaBar database replication effort has been focused on the re-design of the current BaBar specific data distribution tools. These tools are in urgent need of redesign as the current set are difficult to maintain and are breaking.

The current tools lie on top of a set of ASCII files that contain the meta data information necessary for database distribution within BaBar. The new design has to continue to support this legacy system as well as allow us to seamlessly (if that's ever possible) replace the ASCII files with metadata catalogs, file moving services and whatnot.

We are still at the design stage of the new set of tools, we expect to make progress on this (design and implementation) within the next quarter.

3. Single Collaborator Efforts and End to End Applications

As well as the identified Project Activities, PPDG effort focuses on end-to-end applications and demonstrators. This is in keeping with our mission of a short to medium term focus, with working systems used to input requirements to current and future projects, feedback to ongoing designs and implementations, and a basis from which to discuss future needs and work. In some cases a fraction of the work to be done and reported is through off-project effort. PPDG benefits significantly from this synergistic work and uses it as input to the discussions and decisions for on-project efforts and goals.

3.1. ATLAS

3.1.1. US ATLAS Grid Testbed

GDMP 1.2.2 was installed and tested at the ANL-HEP node. Installation and testing of Globus DataGrid beta tools for gsiftp, data replica catalog, data replica manager also took place at the ANL HEP gatekeeper. MDS 2 was installed at the ANL HEP gatekeeper. Testing of the GRIPE account request management system continued; the system was found to be too immature for public deployment and will be further developed (at Indiana U) in light of feedback. Testing of Objectivity servers was done on the ANL,BU,IU and BNL gateways. The testbed now contains 8 gatekeepers at BNL, Boston U, Indiana U, LBNL, ANL, Oklahoma U, U Mich, UT Arlington. A PHP front end for Tilecal Production and Testbeam SQL databases is in development. These tables store meta-data and replication information for Tilecal.

See http://www.usatlas.bnl.gov/computing/grid/ for more testbed information.

3.1.2. Monitoring

A PPDG working group on instrumentation and monitoring was organized, co-chaired by Dantong Yu (BNL) and Jennifer Schopf (ANL, CS rep for ATLAS).

Initial steps in organizing a monitoring effort were taken during this period. Monitoring tools and instruments which are available or under development were cataloged. Requirements from information

consumers were collected and compared to existing capabilities to identify missing functionality. Prioritization of the essential services and resources to be monitored in the grid infrastructure was done.

3.1.3. Distributed job management

A program of work and schedule was developed for the initiative (joint with GriPhyN) to study and test the capabilities of Condor to manage a hierarchical job management infrastructure incorporating the various tiers of grid sites. Discussions are underway towards possibly making this a PPDG project.

See http://physics.bu.edu/~youssef/atlas/notes/ for more information on the Condor scheme being investigated.

3.1.4. Data signature

An enumeration was done of the information to be contained in a 'data signature' recording the history of a data set in sufficient detail and completeness that it could be reproduced. Design issues (such as a global identifier scheme to identify the history objects making up a data signature) have begun to be addressed. See http://www.usatlas.bnl.gov/~dladams/data_history for details about this work in progress.

3.2. BaBar

3.2.1. Network throughput performance

We have extended measurements of bulk network throughput between SLAC and major BaBar and collaborator sites using iperf, to include PPDG and major HENP sites. We are now monitoring these sites on a regular basis by iperf for 10 seconds each hour. These measurements are reported in http://www.slac.stanford.edu/comp/net/iperftests-html/International-Iperf-Tests.html. As part of understanding how best to make these measurements we have also made spot measurements to understand how to achieve high performance and the impact of measurement duration, large windows and multiple parallel streams (see http://www-iepm.slac.stanford.edu/monitoring/bulk/window-vs-streams.html). From this we can set realistic expectations of what is achievable and how to go about achieving it. Our current work involves making the automated measurements more robust, improving resistance to denial of service attacks, and providing better reports and better ways to view them.

We have extended the measurements to the application level, and are using the SLAC written bbcp file copy program to make measurements of file transfer performance between sites. We have worked with the author of bbcp (Andy Hanushevsky) to extensively test it, identify bugs and define new features. The goals of this are to see what extra constraints are imposed by the application on top of the network layer (e.g. security, disk access etc.), and how close one can get to network (iperf) performance. The early results are available at http://www-iepm.slac.stanford.edu/monitoring/bulk/bbcp.html. We presented details of this work at CHEP01 in Beijing (see http://www.slac.stanford.edu/grp/scs/net/talk/chep01-throughput_files/frame.htm).

We are looking into how to use QBone Scavenger Service (QBSS) to reduce the impact of the high throughput on other users. We reported on this at the Virtual Internet 2 meeting (see http://www.slac.stanford.edu/grp/scs/net/talk/qbss-i2-oct01_files/frame.htm).

As part of this we have set up 2 QBSS testbeds. More details can be found at http://www-iepm.slac.stanford.edu/monitoring/qbss/measure.html where we show how well QBSS manages traffic and how well interactive traffic works in the presence of heavy QBSS marked bulk throughput traffic.

We have also put together a proposal for the SC2001 Bandwidth Challenge. See http://www-iepm.slac.stanford.edu/monitoring/bulk/sc2001/ for more details. This proposal has been accepted. It includes over 20 collaborating sites (including all the PPDG sites) to which we will be sending large amounts of bulk throughput from the SLAC/FNAL booth at SC2001. We also hope to demonstrate the effectiveness of QBSS for very high-speed links (2Gbits/sec). Following SC2001 we expect to use some of the infrastructure (accounts, privacy keys, installed software) put together for this demonstration in order to provide long term monitoring of bulk throughput between PPDG and some other key sites.

3.2.2. Replica Catalogs in the Globus Framework

A study has been performed of various replica catalog approaches consistent with the Globus framework. For example replica catalog entries that refer to a set of files as opposed to individual files and dynamic protocol definition to increase the flexibility so that new replica catalog schema can be tested and deployed within the framework.

3.3. CMS

Major activities of the CMS group in PPDG included work on the distributed Monte Carlo Production system (MOP, Jim Amundson) and the Grid Data Mirroring Package (GDMP, Shahzad Muzzafar) at Fermilab, and work on the Clarens data server (Conrad Steenberg), on a monitoring system using distributed services (Iosif Legrand), on Robust Execution Services (RES, Takako Hickey), on preparation of the "Bandwidth Challenge" for the Supercomputing 2001 Conference (Koen Holtman and Julian Bunn) and on the documentation of the CMS data model (Koen Holtman) at Caltech.

The Fermilab CMS group, together with the U.S. CMS prototype Tier-2 center at UCSD (Ian Fisk) and Caltech (Suresh Man Singh) and the University of Wisconsin CMS group are working with the Condor team on developing a prototype distributed Monte Carlo production system. The MOP efforts during this quarter are described in some detail in section 2.3.

The Fermilab group is also involved in GDMP. This effort is described above in section 2.1.

At Caltech the work on Robust Execution Service (RES) is progressing. The project goal is to identify the needs and to provide fault-tolerance for grid systems. In contrast to traditional approach to fault-tolerant system, in the current Grid systems fault-tolerance is not part of the design of many components. Instead the Grid is assembled from already existing components, which often are built with different or non-existent fault-tolerance properties.

This work includes studying existing Grid components such as Condor and Globus to analyze the fault-tolerance properties they provide. Ways to extend the fault-tolerance of these components are explored, to provide desired additional fault-tolerance properties in the form of fault-tolerance plug-ins. Rather than trying to replace the entire system with a fault-tolerant equivalent, we are investigating small fault-tolerant components that can be plugged into existing or evolving grid systems. This will allow applications to experiment and investigate fault-tolerant components without committing to them.

The integration of RES with the MOP system was started and a first plug-in is being tried in this context. The plan is to complete the installation of PBS/GRAM at Caltech, then to follow the example to enable GRAM access to RES, thereby enabling the use of RES inside MOP when submitting jobs to the Caltech Tier-2 facility.

The next plug-in that is being designed is a fault-tolerant version of the Condor master worker tool. The Condor M/W tool has a single point of failure: if the master gets partitioned away from most of workers, it will be unable to utilize most of system resources. We are investigating to create a master that consists of multiple peers that can progress despite partition failures. Because the Condor master/worker potentially expands over a large number of sites, we are looking into using unreliable multicast for communication among peers. This introduces an additional challenge: while unreliable multicast increases the scalability compared to reliable group communication, it makes providing fault-tolerance more difficult. Other plug-ins being considered include to replicate DAGMan and replica catalogue services.

A new design of RES was completed which will enable it to run on a larger set of processors, like a Tier-0/1 center. The design is based on a hierarchical set of servers. The implementation of this design is not yet completed. Collaboration with Keith Marzullo's team at UCSD was started on research on fault-tolerance for the Grid in connection with the GriPhyN project.

Clarens is a data server for remote analysis of tag- and histogram data, developed at Caltech. During this quarter the grid-enabled user environment was enhanced with tag selection code, by using query strings from remote clients. New clients were developed for Lizard (or any Python language script), Java Analysis Studio, and C++ (from within user analysis code) and a web browser client. This work is described at http://heppc22.hep.caltech.edu.

Clarens testbed activities include installation and test the data server on the Caltech Tier2 prototype and to setup tests to monitor distributed production at Fermilab, as part of the PPDG SC2001 demo. A live demonstration of the Clarens server was given at the PPDG meeting, showing remote access of histograms and tags stored at the Caltech Tier-2 prototype.

Work is progressing on the Distributed Service System for Grid monitoring at Caltech. During this quarter, the SNMP part was added to the monitoring system. A lightweight SNMP library (from OpenNMS - http://www.opennms.org) is being used. A structure is used to allow dynamically loading of monitoring modules from a distributed file system or http servers, and to deploy them to perform certain monitoring tasks on selected nodes. The system allows to control all these modules from a GUI. The functionality of the GUI was improved.

A real-world test will soon start at CERN. A PC was set up with RedHat 7.2 and J2EE to provide a dedicated Web sever, able to support Web services. Work is in progress on a Web page for the distributed services.

The system is described in detail in a paper submitted to the CHEP2001 conference at:http://clegrand.home.cern.ch/clegrand/CHEP01/chep01_10-010.pdf . The talk at the CHEP conference can be found here: http://clegrand.home.cern.ch/clegrand/CHEP01/chep01_10-010_slides.pdf. The architecture of the monitoring service is described in this paper http://clegrand.home.cern.ch/clegrand/CMS_Monitor/MMonitorTool.doc.

The “Network Bandwidth Challenge” for the Supercomputing 2001 Conference has the motto "Bandwidth Greedy Grid-enabled Object Collection Analysis for Particle Physics." It is planned to demonstrate a client/server application that will allow particle physicists to define, replicate and analyze collections of objects stored in a multi-TeraByte object or relational database. The client application components include a Grid-aware tool for communication with central servers, located across the WAN at Caltech's Center for Advanced Computing Research (CACR), and at the San Diego Supercomputer Centre, and several Java codes that allow the user to select and analyze physics event object of interest. The object collection defined by the user is communicated to and assembled on the CACR server. The object collection is then replicated over high speed WAN links to a caching database on the client on the conference floor. For subsequent selections, only objects missing from the cached collection are replicated from the server. The software embodies previews of several techniques that are being developed to support the analysis of PetaBytes of event data due to be collected at the LHC.

The client device is a standard dual-Pentium III PC running Linux. It is equipped with 512MB memory, Gbit and Fast Ethernet NICs, and several Ultra SCSI3 disks. The software installed includes the C++-based object replication codes, the Java-based object analysis and selection codes, the Objectivity DBMS, and the Globus software. The servers are at LHC/GriPhyN/PPDG "Tier2" prototype compute farms in Caltech and UCSC, each consisting of about 20 dual processor Pentium III slave nodes, two dual processor master nodes, 3 TeraBytes of very fast RAID disk, Gbit and Fast Ethernet NICs, interconnected on an HP ProCurve switch. More details on the U.S. CMS prototype Tier-2 regional center at Caltech and UCSD is available at http://pcbunn.cacr.caltech.edu/Tier2_Overall_JJB.htm.

Further work at Caltech included documenting the CMS data model. A paper entitled “Views of CMS Event Data: Objects, Files, Collections, Virtual Data Products.' (CMS note 2001/047) was written by Koen Holtman (Caltech) and is available here: http://kholtman.home.cern.ch/kholtman/note01_047.pdf . This is a contribution to and provides common (terminological) reference material for the CMS architectural efforts and the Grid projects PPDG, GriPhyN, and the EU DataGrid. The paper outlines the current CMS production and future CMS grid plans taking a data-centric viewpoint. It has an up-to-date picture of where everything fits in the complete vertical chain from the high-level physics view down to bits on hardware devices.

3.4. D0

3.4.1. Integration with GridFTP

We have started on modifications to SAM to allow use of GridFTP. The Fermilab security group are starting to understand the requirements for integration of the CA/PKI authentication with the local

Kerberos authentication. The SAM monitoring system has been extended for a demonstration of replication for a demonstration at SC2001.

3.4.2. Distributed Analysis

A first version of making SAM files available through the ROOT analysis framework has been done and is in use at the Experiment.

3.4.3. Test Harness

The test harness has been extended and is used as a regression test for new SAM releases. This is of particular importance for ensuring backwards compatibility for new versions of a system in use by several hundred users at five different sites.

3.5. Jlab

In addition to the collaborative work with the SRB group (section 2.5), Jefferson Lab has been active in two areas: Disk Cache Work, and an exploration of common HRM functionality needs.

The development of a document describing common HRM operations continues via discussions between with Jlab, LBL, FNAL, and WP5. An early draft was distributed earlier in September. There have been significant revisions since then. A final copy should be distributed in the next quarter with prototype implementations to follow.

Jefferson Lab's disk cache software component, a part of JASMine (see http://cc.jlab.org/scicomp/JASMine) will be a building block of our grid-related services involving file replication. Maintenance changes to this software are under way to make this integration cleaner.

3.6. ANL – Globus

A PPDG working group on instrumentation and monitoring was organized, co-chaired by Jennifer Schopf (ANL, CS rep for ATLAS) and Dantong Yu (BNL). This will be in coordination with the GriPhyN project as well.

Initial steps in organizing a monitoring effort were taken during this period. The cataloging of monitoring tools and instruments which are available or under development was started. Requirements from information consumers are beginning to be collected and compared to existing capabilities to identify missing functionality. Prioritization of the essential services and resources to be monitored in the grid infrastructure is ongoing.

Jennifer also participated in ATLAS planning and development discussions and meetings. Results of this work has been the solidification of a testbed plan for ATLAS, joint with PPDG and GriPhyN, as well as continued integration of the Globus Toolkit components in ATLAS software components.

Globus developer John Bresnahan worked on a POSIX-like open / read / write interface to GridFTP, as well as then layering an RFIO interface on top of that. We made bug fixes, and worked on packaging, in preparation for the upcoming Globus 2.0 beta release.

The fourth Alpha release of Globus 2.0 was released; a Beta release is expected before the second week of November, although the exact date is not yet finalized.

An informal design for a reliable file transfer service to run on top of GridFTP has been created and the implementation of a prototype is in progress. This service will (as a goal) accept and maintain file transfer requests persistently across crashes of all servers and clients involved, and restart requests as necessary to attempt request completion across a broader set of failure modes than GridFTP alone can do.

3.7. NERSC – SDM

People inlvolved: Junmin Gu, Alex Sim, Arie Shoshani

We have completed HRM version 3.0, and DRM version 1.1. Both DRM and HRM use a uniform API according to the design performed in the previous quarter. This work involved the following parts:

1) The development of a completely new DRM component that can perform both "reads" and "writes" of requests, each having multiple files.

2) The development of an enhanced TRM (Tape Resource Manager) component that can perform both "reads" and "writes" from/into the HPSS system.

3) The integration of the DRM and TRM into a single HRM module.

We have written 3 documents describing this work: a) A short 5-page introductory document on SRMs and their functionality; b) A SRM design document that delineates our considerations in the design stage. c) A document containing the APIs expressed both as CORBA IDLs, as well as C++ APIs.

In addition to the above, two additional activities took place:

1) We have started the work of collaborating with the STAR project people at BNL and LBNL on the installation of the software necessary to setup the "distributed file management" (DFM) project. This includes the installation of the latest version of Globus software in both sites, the Orbacus ORB, and the correct version of compilers. This will be followed by the installation of HRMs in both locations, as well as a "Simulated Commands-line Client" (SCC) that will help us test the system.

2) We have been working with the people in Fermi and JLAB on a joint document that describes the reasoning and the design of HRM APIs for PPDG. This work was initiated by the JLAB people, was followed by changes/additions by Fermi, and after a joint conference call, we have modified and added sections of the document.

Both DRM and HRM implementations have the following functionality.

1) They are capable of managing both "read" and "write" requests. They both use a single uniform API.

2) They allocate space subject to a quota per user, based on both "total space", and "total number of file" limits.

3) If the requested file is not in the disk cache, they get the file from the source location. In the case of DRM, it gets files from remote DRMs, HRMs, or directly from disks; in the case of HRM, it gets files from HPSS.

4) They "pin" files as soon as files get to the disk cache.

5) They enforce a time-out per file set by the local administrator.

6) They "unpin" files as soon as they are released by the client.

7) They queue requests if the storage system is busy or the cache is currently fully in use.

8) Client can either be called-back when files are cached or archived, or can find out the status of files by issuing status calls.

9) They permit files to be designated as "permanent" by the SRM administrator.

10) In the case of HRM, it provides 2 kinds of call-backs when a file is written to it. when the file has transferred to cache, and later when the file is migrated to tape.

3.8. SDSC – SRB

The San Diego Supercomputer Center is collaborating on multiple projects that are developing data grid technology. The explicit PPDG activities include support for database replication for the BaBar experiment, and collaboration with JLab on the definition of replication attributes. Associated efforts that are funded through other projects that potentially can impact the PPDG data grid include development of a Grid Portal that is integrated with the Storage Resource Broker / Metadata CATalog collection management software and collaboration with a United Kingdom data grid that will also access BaBar data. Finally, extensions to the SRB have been made to support parallel I/O, support asynchronous bulk metadata load, provide

collection management capabilities through a Web interface, and to implement the soft links required by the BaBar project. Each PPDG project is described in detail in the effort reports on the PPDG web site.

3.8.1. SDSC - JLab Replica management interface

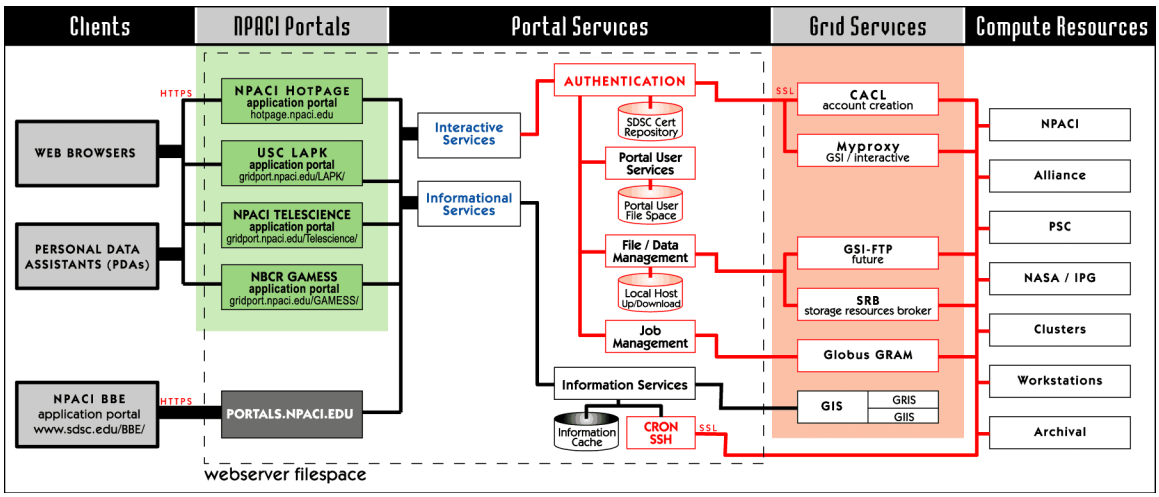
A common interface is being defined for access to Jefferson Laboratory's replica catalog and the San Diego Supercomputer Center Storage Resource Broker, Metadata Catalog replication system. The goal is to develop a set of standard attributes to describe the logical name space in each data grid and the operations that can be done on that name space.

One approach for determining the common set of attributes is to define the capabilities that each system provides, and then to identify the attributes that are used to characterize each capability. As a start towards this effort, a comparison has been done between the SRB and the Globus replica catalog and GridFTP transport system. This will be extended to include the Jefferson Laboratory replica management capabilities.

3.8.2. GridPortal project at SDSC.

The GridPortal team has been successfully using the GSI-enabled SRB to upload and download file(s) into a storage location. The system combines the ability to read data from a user-ID under Globus remote-proxy authentication, import the data into a SRB collection, store the data in a remote storage system through the SRB data handling system, and support replication and discovery of files in the collection.

The GridPortal provides a web interface to both the Globus execution environment and to data stored in the SRB collections. This has served as a demonstration system for proving the feasibility of web-based interfaces to Globus. The data flows that are driven in this environment are demonstrated below.



A second interface, called mySRB, has been created for managing collections. The mySRB interface supports collection creation, attribute definition, attribute creation, data set import, data set replication, browsing, and querying of the resulting collection.

Both Jefferson Laboratory and SDSC have agreed to participate in a testbed in which a common web services interface to the two disparate systems (SRB and JASMine) can be demonstrated. This activity will also involve the Global Grid Forum Data Working Group, and most likely a soon-to-be-formed Web Services working group.

3.8.3. BaBar Support

The status of the Storage Resource Broker Prototype for SLAC Data Replication has been supplied by Adil Hassan of SCS/BaBar.

1) Setup of SRB software (R. Schmitz, Oct-June2001) – done. Included installation of SRB and

Mcat software on solaris7 machine, rudimentary testing of the catalog and SRB software (using srb tools to insert, move, delete entries in the catalog).

2) Setup of test federations containing test data (A. Hasan June 2001) – done. Included setup and load of two federations with data used in analysis, and attaching the collections. An NFS exported file system was used as the second storage system (initial copies used simple unix tools).

3) Development of prototype replication scripts (A.Hasan, R.Schmitz Apr-July 2001) – done. Simple scripts to identify databases for the disk copy based on the results from dumping the catalog have been developed, which hcurrently work off of flat ascii files. The next step is to integrate these scripts to interact with the MCAT. Scripts are needed to write a log of each step of the replication process. Scripts also are needed to query the catalog in a user intuitive way.

4) Metadata catalogs (A. Hasan, A. Hanuschevsky, R.Schmitz, D. Boutigny May-June 2001) – done. Determined the information to be recorded in the metadata catalog.

5) Test of prototype using SRB (A. Hasan, Aug 2001) in progress. Testing is being done using a source and target federation containing a portion of the analysis federation. Testing will entail replication (for use use cases 1 - 8), testing of robustness of the system, and implementation of error recovery procedures and monitoring tools.

All of the features required to support the BaBar collections have been implemented in the SRB software.

4. Appendix

4.1. SuperComputing 2001 demonstrations related to PPDG

CMS Simulation Production – IMPALA and GDMP	FNAL, Caltech	Demonstration of current CMS simulation production tools and GDMP replication tools
CMS Distributed Simulation Production (MOP)	Caltech, FNAL,Wisconsin, ANL, UCSD	Use of Condor-G/DAGMAN to automatically run CMS simulation production at multiple sites
Bandwidth Greedy Grid-enabled Object Collection Analysis for Particle Physics	Caltech, UCSD	Demonstration of the use of Grid tools and virtual data to support interactive physics analysis.
Reliable Transport	ANL	Extensions to the transport layer of GridFTP to support retry
Proxy Server Demo	ANL, SLAC	Demonstration of replica catalog proxy server
GriPhyN Virtual Data (CMS)	ANL, Florida	Generation of CMS simulation scripts from definition of physics parameters
Globus CAS prototype	ANL	Use of Community Authorization Service in Earth Sciences Grid
"Letting Scientists Concentrate on Science: Providing a Transparent View of Data on the Grid"	LBNL	http://gizmo.lbl.gov/~arie/sc2001.demo/slides/index.htm http://gizmo.lbl.gov/~arie/sc2001.demo/poster.pdf
“Bandwidth to the World”	SLAC/FNAL	The "Bandwidth to the World" project is designed to demonstrate the current data transfer capabilities to about 25 sites with high performance links, worldwide. (http://www-iepm.slac.stanford.edu/monitoring/bulk/sc2001/)
SDSC Grid Portals Architecture	SDSC	The SRB team has been working with Grid Portal Architecture group to use SRB in building Grid Portal services.

4.2. Draft INTERGRID charge

GW/Draft 2OCT01

International HENP Grid Coordination and Joint Development Framework
{PRIVATE }

The HEPN Grid R&D projects (initially DataGrid, GriPhyN, and PPDG, as well as the national European Grid projects in UK, Italy, Netherlands and France) have agreed to coordinate their efforts to design, develop and deploy a consistent standards-based global Grid infrastructure. The guidelines for coordination and joint development by the projects are enunciated below. This collaborative effort can be referred to as INTERGRID.

Preamble

The consortia developing Grid systems for current and next generation high energy and nuclear physics experiments, as well as applications in the earth sciences and biology, have recognized that close collaboration and joint development is necessary in order to meet their mutual scientific and technical goals. A framework of joint technical development and coordinated management is therefore required to ensure that the systems developed will interoperate seamlessly to meet the needs of the experiments, and that no significant divergences preventing this interoperation will arise in their architecture or implementation.

To that effect, their common efforts will be organized in three major areas:

- **An InterGrid Management Board (IGMB) for high level coordination**
- **A Joint Technical Board (JTB)**
- **Common Projects, and Task Forces to address needs in specific technical areas**

A/ IGMB (Intergrid Management Board)

A.1 IGMB Role

- **Information exchange on the status, plans and issues facing national and regional Grid initiatives**
- **Periodic review of key developments and directions in the Grid projects, with particular attention to maintaining convergence and interoperability, including review of the Common Projects**
- **Set up a legal framework for collaboration, covering intellectual property rights and associated issues**
- **Organizes Common Events (Workshops, Seminars, etc.)**
- **Proposes joint submissions of items to external bids, where appropriate**
- **Receives regular reports from the Joint Technical Board**
- **Approves the list of common projects and ad hoc task forces, proposed by the JTB**

A.2 IGMB Composition

The IGMB is presently composed of the combined Management Boards of the DataGrid, PPDG and GriPhyN projects. It will be extended to represent new Grid projects as they come along.

A.3 IGMB Meetings

Three times per year, synchronised as much as possible with Global Grid Forum meetings

A.4 Chairmanship

The IGMB will elect a chairman , who will serve for one year.

{PRIVATE }B. Joint Technical Board

B.1 Role

- Ensure compatibility and interoperability of Grid tools
- Clearly identify API, interfaces
- Launch task forces on specific issues (such as networking, architectural issues, security, ...)
- Reviews the common projects
- Reports to the InterGrid Management Board
- Ensures good contact with the various Grid forums, especially the Global Grid Forum working groups

B.2 Composition

6 members for European GRID projects, 6 for US GRID projects and 2 for Asian Pacific projects

B.3 Chairmanship

One year term

B.4 Meetings

At least 4 times per year, using teleconferencing as needed

Common Projects

Common projects are specific well-focused joint efforts on a small number of key issues, or sets of issues.

C.1 Scope

Common projects will normally take one of two forms:

- Joint development of specific Grid services or components targeted at one or more large HENP experiments involving US and Europe partners
- Realisation of dedicated transatlantic testbeds for software development, network tests, etc.

Testbeds will normally be linked to a well-specified development program with deliverables, and will be targeted at near or medium term goals of the targeted experiment(s)

C.2 Liaison Team

A liaison team will be appointed for each project by the IGMB and the relevant partners.

- Role

The role of the liaison team will be to develop a reasonable work plan with precise milestones and deliverables for each partner (Grid consortium and HENP experiment), and manpower requests from each partner . The work plan will be reviewed by the Joint Technical Board and approved by the IGMB.

- Composition

The liaison team for a specific project will include one member from each Grid consortium involved, and if a HENP experiment is involved, one European member and one US member of this experiment. The liaison team will designate its chair for interaction with the Technical Panel

C.3 Reviews

The project will be reviewed at regular intervals by the Joint Technical Board.