

Particle Physics Data Grid Quarterly Status Report

July – September 2000

Contents

Introduction.....	2
Summary of activities related to "High performance site-to-site file replication service"	2
Summary of activities related to "Multi-site cached file access service"	3
Summary of activities related to "Development of Grid architecture and APIs"	3
PPDG Quarterly Status Report for Argonne National Laboratory, Mathematics and Computer Science	
Division, Distributed Systems Laboratory	5
Description of activities related to "High performance site-to-site file replication service"	5
Description of activities related to "Multi-site cached file access service"	5
Description of activities related to "Development of Grid architecture and APIs"	5
PPDG Quarterly Status Report for Argonne National Laboratory, Physics Department.....	6
Description of activities related to "High performance site-to-site file replication service"	6
Description of activities related to "Development of Grid architecture and APIs"	6
Description of other activities	6
PPDG Quarterly Status Report for Brookhaven National Laboratory,	7
Description of activities related to "Development of Grid architecture and APIs"	7
PPDG – Quarterly Status Report from the Caltech Group	8
September 2000.....	8
Introduction	8
High Throughput Data Transfer.....	8
Globus Security and Information Infrastructure	9
Tier2 Center Design	9
Distributed Computing and Data Management.....	10
Data Structures and Replication.....	10
CMS Production.....	10
Distributed Task Scheduling.....	11
Distributed System Simulations.....	12
PPDG Quarterly Status Report for Fermilab	13
Description of activities related to "High performance site-to-site file replication service"	13
Description of activities related to "Multi-site cached file access service"	13
Description of activities related to "Development of Grid architecture and APIs"	13
Description of other activities	14
PPDG Quarterly Status Report for Jefferson Lab	15
Description of activities related to "High performance site-to-site file replication service"	15
Description of activities related to "Development of Grid architecture and APIs"	15
PPDG Quarterly Status Report for LBNL, Scientific Data Management.....	16
Description of activities related to "High performance site-to-site file replication service"	16
Description of activities related to "Multi-site cached file access service"	16
Description of activities related to "Development of Grid architecture and APIs"	16
Description of other activities	17
PPDG Quarterly Status Report for LBNL physics	18
Description of activities related to "High performance site-to-site file replication service"	18
Description of activities related to "PPDG Project Coordination"	18
PPDG Quarterly Status Report for SDSC	19
Description of activities related to "High performance site-to-site file replication service and multi-site cached file access".....	19
Description of activities related to "Development of Grid architecture and APIs"	19
Description of other activities	19
PPDG Quarterly Status Report for SLAC	20
Description of activities related to "High performance site-to-site file replication service"	20
Speed Testbed.....	20

BaBar Intercontinental Testbed	20
Description of other activities	21
PPDG Quarterly Status Report for Wisconsin	22
Description of activities related to "Multi-site cached file access service"	22

Introduction

This report describes and summarizes the activities and progress of the Particle Physics Data Grid project for the period July 1 to September 30, 2000. Immediately following the introduction is a summary of activities across all participating sites. More detailed descriptions of these activities are available in the appendices of reports from individual sites (in alphabetical order) regarding activities at those sites.

Summary of activities related to "High performance site-to-site file replication service"

The Globus team has defined a set of replica catalog APIs (based on the LDAP protocol) and replica management APIs. Over the past 3 months the replica catalog APIs have been implemented as both C and Java libraries. These libraries are currently available in alpha release form and are being used in an evaluated in several PPDG prototypes including the CERN Grid Data Management Pilot (GDMP). We have developed a plan for a detailed evaluation of these APIs and libraries in the context of the Babar experiment.

The ANL group has continued testing of the beta version of the Globus LDAP based file replication catalog. ANL is also working with BNL on Objectivity DB replication API and implementation. Additionally, there are discussions with Indiana and Fermilab on "GridCopy" an extension of "FocusCopy" project. This entails adopting some of the grid file replication tools to bulk data transfer work-flow procedures.

Caltech and SLAC are working on a high-speed (dual OC12) point to point connection via NTON with a goal of demonstrating 100 MB/sec effective data transfer. A throughput of >600 Mbps has been measured. There are numerous lessons being learned and this work is ongoing. Once the network throughput has been optimized there will be a demonstration of sending Objectivity/DB database files across this link.

Fermilab has further developed and deployed scripts for replication and cataloging of Focus data files from 8mm tape at Fermilab to Indiana HPSS system. We have developed a set of easy to modify and use scripts that replicates data files between two sites, catalogs the transfers and collects and displays statistics about the copies.

Jefferson Lab has begun to test the use of gsiftp and other tools for transferring experimental data from Jefferson Lab to one or two remote universities. These initial steps have served to understand the use of GSI certificates, and to explore the transfer rates possible to collaborating sites.

LBNL nuclear physics (A. Vaniachine) has prepared a description of the file replication needs and plans for the STAR experiment with regard to data transfer between BNL/RCF and LBNL/NERSC/PDSF (URL here). This was presented at the PPDG collaboration meeting at ANL on July 13,14, 2000. Some additional planning and measurements of network characteristics were carried out during August and September.

SDSC has released a new version of SRB (1.1.7) for PPDG use including new features jointly developed with LBNL to support file staging and status requests. SRB servers are currently running at LBNL and Wisconsin. A capability of registering a replicated file is included in latest release of SRB and enables transferring files by some other means and using SRB as a replica catalog.

From SLAC, throughout the period July – September, BaBar has been transferring about 300 Gbytes per week between SLAC and IN2P3 Lyon. This work is almost totally supported by US and French funding for the BaBar program, but close ties with PPDG promote a valuable exchange of tools and experience. Experience has been gained with enhanced ftp supporting large windows and multiple IP streams (bbftp written by Gilles Farache of CCIN2P3, Lyon) as well as many 'second order' issues of reliability, error handling, resource management that dominate the work of the BaBar scientist who manage the SLAC-Lyon transfers. After discussions with the Globus team in August, trial deployments are planned of new Globus components supporting a replica catalog and streamed ftp.

Summary of activities related to "Multi-site cached file access service"

The Caltech group has been actively involved in many Grid-related activities in Europe. Real use-cases of the HEP community at CERN were evaluated and incorporated in the functionality of Work Package 2 of the EU DataGrid. This research will appear in our paper "Data Management in an International DataGrid". AS has also been actively participating in the CMS-related Grid activities. A project called "Grid Data Management Pilot" (GDMP) is supposed to fulfill CMS's urgent needs of a DataGrid infrastructure and at the same time act as a pilot for the longer term EU DataGrid project. The first version of this software has been released (Version 1.1) and its design and architecture will be presented in the coming ACAT 2000 workshop at FermiLab. Koen Holtman and Asad Samar have had frequent contacts with the Globus team, to support the creation of the requirements for Globus DataGrid components like the Globus Replica Catalog and Globus Replica Manager.

Caltech and UCSD are preparing a plan for Tier2 prototypes and Tier1 interaction, which will involve the purchase and installation of hardware and software. ORCA database file replication between CERN, FNAL and the prototype Tier2 servers at Caltech and UCSD will be one of the first tasks.

CMS is undertaking a large Monte Carlo simulation and reconstruction production run in Fall 2000, with of order 2 to 4 million events planned to be generated, simulated, reconstructed and then analyzed by several different physics groups. The processing of each event involves several stages, each to be performed at different locations, primarily Caltech, Wisconsin, FNAL and CERN. The processed events will be accessed and analyzed by physicists in those and several other locations. This task will be supported by the Globus-based ORCA file replication services being developed by researchers at Caltech and CERN described above, and in collaboration with the European commission DataGrid project.

The D0 SAM system now supports global disk cache management. Automatic caching of data files to be read, and automatic routine of data files to be written, is done based on a central meta-data catalog and distributed resource management services. We are working with our CMS colleagues at CERN, Caltech and the University of Wisconsin on planning and implementation of multi site file caching for the CMS simulation data – initially for the fall production run and for test activities towards the end of this year. The LBNL Scientific Data Management group continued to develop and use the test environment, where an application client at the University of Wisconsin first contacts the Query Interpreter at LBNL to get the list of file that qualify for its logical query, issues file requests to its local SRB client which then contacts the SRB server at LBNL, which in turn requests the HRM (the component that manages file staging from HPSS) to move a file to a staging disk. When this is done, the SRB is notified and it then moves the file in the most efficient way possible to the disk in Wisconsin. This is a major achievement as it proved that a Grid architecture that relies on SRMs is a powerful way to manage Grid storage allocation and coordination.

Wisconsin continued to develop a testbed that will enable the collaboration to perform end-to-end tests of the different file access services provided by members of the PPDG. The testbed is interfaced to our local production environment and allows us to combine data movement and processing.

Summary of activities related to "Development of Grid architecture and API's"

The Globus project continues to help define the architecture, protocols, and APIs that enable grid computing. Members of the Globus project have defined extensions to the FTP protocol, which expand FTP's capabilities to support the needs of grid computing. These extensions include partial file transfer, parallel file transfer, control and data channel security, auto buffer size negotiation, and user definable fault recovery behavior. APIs have been defined which implement these protocols. Of particular interest to PPDG are the Grid Security Infrastructure, Grid Information Service, GridFTP, and replica management work. Several discussions on the subject of replica management were held with key members of the major HEP projects, including ATLAS, BABAR, and CMS and were critical in helping refine these requirements.

The results of these meetings are two documents:

“A Replica Management Service for High-Performance Data Grids”

“Access Control in a Replica Management Service”

The APIs described in these documents are currently under development based on the replica catalog work already completed.

The ANL physics group is actively involved with the ATLAS grid activities including definition of an initial testbed as well as understanding the relation between ATLAS experiment software and grid software.

BNL designed and implemented an Objectivity database class that provides an object oriented API for on-demand database replication. This API is a wrapper around Globus file replication services with added functionality to handle Objectivity-specific database management.

The Caltech group is actively involved with the EU DataGrid project, working with the Data Management work-package team (WP2) of the EU DataGrid project, in the initial design and requirement specification phase.

Fermilab and Wisconsin are collaborating on a test bed application integrating existing components of a grid architecture. This first application will handle reading of data files from the SRB or the SAM systems. This application is being used to test and extend the PPDG HRM API. It will support determining one or more locations of a requested file and determination of whether it should be transferred from SRB or SAM. As part of an activity in exploring the design of grid API's, Jlab as written an object oriented API for a replica catalog, patterned after the globus, and to a lesser extent the SRB, non-OO APIs. In particular, a Java binding was described which is based upon the use of Interfaces (abstract classes) and Factory methods, following the pattern used by Java's database package (JDBC) and a number of other Java packages defined by Sun. This strategy is aimed at cleanly separating API definition from implementation, allowing for wrapping pre-existing grid codes, and migrating to new implementations as they appear.

The Scientific Data Management group at LBNL has been working closely with the SDSC people who developed the Storage Resource Broker (SRB). SRB provides a Grid security and storage infrastructure.

The collaboration's goal is to develop a way for SRB to use the HPSS Resource Manager (HRM) developed at LBNL. The API to the HRM was developed, as well as the software to have SRB communicate with the HRM. Over the last several months, we have been working closely with people at Fermi lab to refine the HRM IDLs so that Fermi can use that as an interface to their SAM storage management system. This resulted in several improvements to the interface definition

LBNL is involved in the development of "Storage Resource Managers" for data grid applications. Our architecture design and implementation is based on our experience with the HENP GC project (where the STACS system was developed). A key concept of adapting this architecture to a distributed grid is the use of Storage Resource Managers (SRMs). Each SRM is associated with a storage resource, such as HPSS, DPSS, or a shared disk cache. The reason that these SRM components are valuable is that the Grid can use these SRMs to request storage reservations, to stage files from tape to a staging disk, and to queue storage transfer requests.

SDSC has developed and maintains a GSI enabled FTP interface to the HPSS archival storage system. GSI-FTP is used by the Globus environment to access storage systems. In parallel with this, SDSC is examining the interface needed to support Globus access to SRB managed data collections.

PPDG Quarterly Status Report for Argonne National Laboratory, Mathematics and Computer Science Division, Distributed Systems Laboratory

Date: Sept. 15, 2000

Participants: [Ian Foster](#), Steve Tuecke, Bill Allcock, Darcy Quesnel, Joe Bester, John Bresnehan, Sam Meder, Chi Chen (student)

Description of activities related to "High performance site-to-site file replication service"

Replication of data sets is necessary for performance reasons when the data, compute resources, and researchers are distributed over a wide geographic area. The Globus team has defined a set of replica catalog APIs (based on the LDAP protocol) and replica management APIs. Over the past 3 months the replica catalog APIs have been implemented as both C and Java libraries. These libraries are currently available in alpha release form and are being used and evaluated in several PPDG prototypes including the CERN Grid Data Management Pilot (GDMP). We have developed a plan for a detailed evaluation of these APIs and libraries in the context of the Babar experiment.

For further information see: [Globus Replica Management API](#)

Description of activities related to "Multi-site cached file access service"

Once the data has been replicated, there must be fast, reliable methods for accessing the data from application programs. Members of the Globus project have defined extensions to the FTP protocol, which expand FTP's capabilities to support the needs of grid computing. These extensions include partial file transfer, parallel file transfer, control and data channel security, auto buffer size negotiation, and user definable fault recovery behavior. APIs have been defined which implement these protocols. Over the past three months, these APIs have been implemented in the GridFTP client and control libraries. These libraries are available in alpha release form and are being used and evaluated in several PPDG prototypes including the CERN Grid Data Management Pilot (GDMP).

For further information see: [GridFTP Control Lib Doc](#) and [Grid FTP Client Lib Doc](#)

Description of activities related to "Development of Grid architecture and APIs"

The Globus project continues to help define the architecture, protocols, and APIs that enable grid computing. Of particular interest to PPDG are the Grid Security Infrastructure, Grid Information Service, GridFTP, and replica management work. Several discussions on the subject of replica management were held with key members of the major HEP projects, including ATLAS, BABAR, and CMS and were critical in helping refine these requirements. The results of these meetings are two documents:

A Replica Management Service for High-Performance Data Grids
Access Control in a Replica Management Service

The APIs described in these documents are currently under development based on the replica catalog work already completed.

**PPDG Quarterly Status Report for Argonne National
Laboratory,
Physics Department,**

Date: Sept. 15, 2000

Participants: Dave Malon, [Ed May](#)

***Description of activities related to "High performance site-to-site
file replication service"***

Continued testing of the beta version of LDAP based file replication catalog beta software under development by Globus group at ANL-MCS. Work done with Chi Chen a ANL-MCS summer student.

Working with BNL on Objectivity DB replication API and implementation. Discussions with Indiana and Fermilab on "GridCopy" an extension of "FocusCopy" project.

Averaged effort during the interval at ANL-HEP 20% of 1 FTE from people:
Ed May and David Malon.

***Description of activities related to "Development of Grid
architecture and API's"***

Presentations and discussions with and at several US-ATLAS and ATLAS meetings and working groups on the PPDG project and tools. Evaluation and design of an initial US-ATLAS test-bed. Selection and evaluation of initial ATLAS software to be "gridified" and run on a test-bed. Atlas-US Grid Workshop at Indiana U. ATLAS GRID Workshop at CERN July 17,18, 2000.

Description of other activities

Hosted the PPDG collaboration July 13,14 2000 meeting at ANL which included a desktop conference LAN and video conferencing for remote participants.

PPDG Quarterly Status Report for Brookhaven National Laboratory,

Date: Sept. 15, 2000

Participants: Rich Baker, Tom Robertazzi, John Leita, Rich Ibbotson, Ognian Novakov

Description of activities related to "Development of Grid architecture and API's"

We designed and implemented an Objectivity database class that provides an object oriented API for on-demand database replication. This API is a wrapper around Globus file replication services with added functionality to handle Objectivity-specific database management. During July-September, 2000, this activity consumed approximately 3 man-months of effort. No equipment was purchased.

PPDG – Quarterly Status Report from the Caltech Group

September 2000

Julian Bunn, Mehnaz Hafeez, Takako Hickey, Koen Holtman, Iosif Legrand, Vladimir Litvin, Harvey Newman, Asad Samar,

Introduction

The Caltech group is active in several of the PPDG project areas:

- Data Grid Development
 - High throughput data transfer (JB, HN, AS)
 - Globus Security and Information Infrastructure (AS, MH, KH)
 - Tier2 Center Design (HN, JB)
 - Distributed Data Management (JB, HN, AS, MH, KH)
 - Distributed Computing and Task Scheduling (KH, TH, VL, AS, MH)
 - Data Structures and Replication+ (KH, JB, HN)
- Distributed System Simulations (IL; HN, KH)
 - Site and Network Configuration and Throughput Studies
 - Production System and ODBMS Studies and Optimization
 - Tape System requirements and Usage-Modes Study
 - Data Access and Analysis Strategy Studies

These topics are covered in detail in the following sections.

High Throughput Data Transfer

There are ongoing tests of high performance networking between the various sites (Caltech, FNAL, CERN, and several other sites active in CMS software and computing), most notably between SLAC and Caltech in relation to the 100 MB/sec PPDG milestone. It is hoped to use ORCA (Object Reconstruction for CMS Analysis) Objectivity database files (as well as BaBar files) once the network is tuned. Coordination between networking experts at all the collaborating institutes is important to ensure that TCP/IP capabilities match, and routes etc. are set correctly. We have been using the NTON Caltech-SLAC link on which we have a dual OC12 connection between a machine in CACR and two machines at SLAC.

It turns out to be difficult to even get iperf or ttcp rates up to the necessary > 800 Mbits/sec that would encourage one to attempt a real file transfer. We currently see an aggregate of a little more than 600 Mbits/sec on NTON, with some mysterious packet retransmits that we are investigating.

CACR has recently ordered new equipment, including a Juniper M160 Router. This will allow us to move from ATM to Gbit Ethernet, and to route traffic between Caltech, JPL, SLAC and other sites over NTON at OC-48 (2.5 Gbps), with the ability to upgrade our setup at CACR to OC-192 in the future. We will then use dual Gbit Ethernet connections across NTON. At some point we will get > 800 Mbits/sec and then the challenge is to:

- a) choose our Objectivity database files,
- b) decide which direction they need to go in,

- c) put them on a disk array that can sustain > 100 MBytes/sec read
- d) pump them across NTON to a disk array that can sustain > 100 MBytes/sec write

We were initially enthusiastic to use all eight of the OC3 adapters on the Caltech X class Exemplar, which were multiplexed into the two OC12 lines to SLAC. This turned out to be too slow. Only the OC3 adapter on the primary Exemplar node could achieve a substantial fraction of the nominal 155 Mbit/sec, and the maximum rate achievable from the secondary nodes (where data was routed across the Exemplar's inter-node buses) was unacceptably low. After initial tests by Davide Salomoni and discussion with Alex Szalay (JHU) on the I/O capabilities of various servers, we purchased a Dell PowerEdge 4400 with dual 860MHz CPUs, 1GB of RAM, twin OC12 ATM cards, and very fast disk arrays (capable of at least 150 MBytes/sec), running Windows 2000 Server.

We chose to run Windows 2000 initially because its TCP/IP stack is highly efficient (as our tests with Gbit ethernet on a LAN confirm) as it takes maximum advantage of all the hardware capabilities of latest generation network cards (we have had excellent results with those from SysKconnect).

Globus Security and Information Infrastructure

The Caltech group has been actively involved in many Grid-related activities in Europe. We have been working with the Data Management work-package [1] team (WP2) of the EU DataGrid project [2], in the initial design and requirement specification phase. We evaluated real use-cases of the HEP community at CERN and incorporated these in the functionality that this work package offers. This research will appear in our paper "Data Management in an International DataGrid" [3]. AS has also been actively participating in the CMS-related Grid activities. We carried out a project called "Grid Data Management Pilot" (GDMP) [4] which is supposed to fulfill CMS's urgent needs of a DataGrid infrastructure and at the same time act as a pilot for the longer term EU DataGrid project. The first version of this software has been released (Version 1.1) and its design and architecture will be presented in the coming ACAT 2000 workshop at FermiLab. Koen Holtman and Asad Samar have had frequent contacts with the Globus team [5], to support the creation of the requirements for Globus DataGrid components [6] like the Globus Replica Catalog and Globus Replica Manager [7].

[1] <http://cern.ch/grid-data-management>

[2] <http://grid.web.cern.ch/grid/>

[3] IEEE, ACM International Workshop on Grid Computing [Grid'2000], 17-20 Dec. 2000, Bangalore, India

[4] <http://cmsdoc.cern.ch/cms/grid>

[5] www.globus.org

[6] <http://www.globus.org/datagrid/>

[7] <http://www.globus.org/datagrid/deliverables/default.asp>

Tier2 Center Design

Caltech and UCSD are preparing a plan for Tier2 prototypes and Tier1 interaction, which will involve the purchase and installation of hardware and software. ORCA database file replication between CERN, FNAL and the prototype Tier2 servers at Caltech and UCSD will be one of the first tasks. The database files are each typically several hundred MBytes in size. The Tier2 prototypes will probably offer ~2 TByte of online disk storage. It is hard at this stage to estimate accurately the WAN traffic from CERN or FNAL to the Tier2 servers. However, we can postulate a half-fill of the available capacity at each site over a couple of days at the start of an analysis or re-reconstruction task, i.e. an average of ~50 Mbits/sec to both sites, followed by replication between the two sites to fill the remainder of available capacity. This second phase will soak up available bandwidth on the SDSC-Caltech link.

Use of the Tier2 for CMS simulated event production, distribution and analysis will involve groups at UC Davis, UC Riverside and UCLA, as well as Caltech and UCSD. The "California Tier2" concept of a distributed center linked over CALREN2 and NTON will be further developed during 2001, based on fund sharing at some of the university sites mentioned above.

Distributed Computing and Data Management

Data Structures and Replication

Objectivity object level and user-collection replication R&D at Caltech is focussing on CMS (and SDSS astrophysics) data. Initial designs for object level replication tools were presented in February 2000 at the CHEP 2000 conference [1]. A design for a first prototype was completed in July 2000 [2] [3]. Coding began in August 2000. The initial prototype will replicate CMS ORCA physics objects, will use GLOBUS middleware [4] for security and fast data transport, and Objectivity/DB [5] as the underlying storage layer. First results will be presented at ACAT'2000 [6] (at Fermilab) in October, and the prototype will be demonstrated at Supercomputing 2000 [7] (Dallas) in November. The demonstration will involve transparent data replication for access and processing with improved throughput between the SC2000 conference site, Caltech, CERN and potentially other sites. The Caltech - Dallas path will be instrumented to support data transfers in the Gbps range (up to OC-48). This R&D is progressing in close collaboration with the Globus team of Ian Foster et al., and Johns Hopkins University team of Alex Szalay et al.

A visit has been made to the SDSS Science Archive team at JHU [8] to exchange knowledge and experience in implementing large science archives using Objectivity/DB. Development copies of the SX query tool and server [9] have been successfully installed at Caltech in July 2000. This is a first step towards installing a complete replica of the production SDSS SX data at Caltech.

CMS Production

CMS is undertaking a large Monte Carlo simulation and reconstruction production run in Fall 2000, with of order 2 to 4 million events planned to be generated, simulated, reconstructed and then analyzed by several different physics groups. The processing of each event involves several stages, each to be performed at different locations, primarily Caltech, Wisconsin, FNAL and CERN. The processed events will be accessed and analyzed by physicists in those and several other locations. This task will be supported by the Globus-based ORCA file replication services being developed by researchers at Caltech and CERN described above, and in collaboration with the European commission DataGrid project. These services will be implemented as a first prototype in time for the fall 2000 production. The prototype will allow replication of the data and meta data in streaming or on-demand modes. Once replicas of the produced events have been made, additional processing steps will be executed at the primary sites, followed by further replication of the new results. At that stage, results can be analyzed by the distributed groups of CMS physicists.

So far, the main focus has been on ORCA4 and CMSIM installation on the CALTECH and Wisconsin facilities:

- *jasper.cacr.caltech.edu*: A SUN Ultra-250 dual CPU machine at CACR, running the full CMS ORCA and User Analysis Environment. It has been used as a test-bed for the first versions of the automatic file transfer system, which will be used in the Fall 2000 September CMS production. The GLOBUS toolkit has been installed and tested. This will be used together with the Grid-enabled GDMP application as part of a next-generation system for CMS data production and distribution.

- *X-Class Exemplar* at Caltech/CACR - Full version of the automatic file transfer system has been installed together with an adapted CMSIM 116 (CMS simulation program) version. After the CMSIM 120 release (which includes a complete representation of the latest all-Silicon tracker), we will adapt the program for massive simulation production runs (using 240CPUs) in support of the Higher Level Trigger studies which are a major focus of CMS work this year and 2001. 250,000 minimum bias events using the CMSIM 116 version have been generated and stored on the CALTECH HPSS system by means of this system. An automatic event-transferring system (which is to be replaced by the application based on GDMP (see above)) was extensively tested during this production run.
- *Linux part of Condor cluster* at Wisconsin: A full version of the automatic file transfer system has been installed together with recent CMSIM versions. The ORCA 4.2.0 release was successfully used to build and populate Objectivity/DB federations of simulated hits and digits (ooHits & ooDigis); both in a standalone application and in jobs running on the Condor flock (without checkpointing). Approximately 0.5M QCD background events were produced during these runs. We are planning to use the Linux part of the Condor flock at Wisconsin for the Fall CMS production. We also have plans to utilize the Solaris part of Condor and make changes inside ORCA 4.2.0 (and future releases) to enable it to run smoothly in the Condor flock. Additional disk space has been installed on Condor for this purpose.
- *naegling.cacr.caltech.edu* - CMSIM/ORCA 4.2.0 has been installed on this Beowulf-class cluster. We tested managed queuing systems, which were developed in the GIOD[10] project framework. That was successfully tested together with CMSIM 118, and very soon we will make tests with ORCA 4.2.0. 70GB additional disk space has been installed for that purpose. Work has begun on using naegling as a test-bed for our future Linux PC Farm. The intent is to verify all three major components together - ORCA itself, the automatic file transfer system and the managed queuing system from the GIOD project.
- *future PC Farm*. We have plans to create a test-bed for distributed computations between Caltech-SDSC after purchase of the Tier2 PC Farm in October 2000 (see Section on Tier 2 Centers). A 155Mbps network connection will be leased.

Transparent migration of data in and out of the Caltech HPSS system will be added as part of the production procedures in the future, based on work to be done in collaboration with EU DataGrid WP5 (on Mass Storage System integration) starting in November 2000.

- [1] K. Holtman, H. Stockinger. Building a Large Location Table to Find Replicas of Physics Objects. Proceedings of CHEP 2000, Padova, Italy. http://kholtman.home.cern.ch/kholtman/olt_long.ps
- [2] http://www.cacr.caltech.edu/ppdg/meetings/ppdg_collab/holtman/objrepl.pdf
- [3] http://kholtman.home.cern.ch/kholtman/globusretreat_objrepl.ppt
- [4] www.globus.org
- [5] www.objy.com
- [6] <http://conferences.fnal.gov/acat2000/>
- [7] <http://sc2000.org/>
- [8] <http://www.sdss.jhu.edu/>
- [9] <http://www.sdss.jhu.edu/ScienceArchive/doc.html> Resource/Job Management Services
- [10] <http://pcbunn.cacr.caltech.edu/>

Distributed Task Scheduling

In the past quarter a job management service has been developed and prototyped on the naegling Linux cluster. The service allows clients to submit, monitor, and terminate jobs as a set. It has a scheduling mechanism (to be refined) that allows selection of processors based on processor type, load, available data sets, etc. The service maintains replicated states, so that computations will not be lost if a server fails. The state tolerates network partition failures, so clients will not lose long running jobs when a partition heals. The original implementation of the service [4] used the group communication toolkit Ensemble [2,3]. As

the service software was ported to the 65 processor naegling cluster at CACR [1] some scalability problems were encountered. The system was thus redesigned to use group communication only for a small set of servers. To preserve a consistent membership and replication state an additional mechanism was employed, termed reliable RPC. The new design runs well over 65 servers. The earlier version that run up to 32 server was tested with ORCA production software. A paper describing the new design is under way [5].

- [1] Intel PentiumPro Beowulf Cluster (naegling), <http://www.cacr.caltech.edu/resources/naegling>
- [2] Kenneth P. Birman, "Building Secure and Reliable Network Applications", Manning Publishing Company and Prentice Hall, Jan 1997.
- [3] Mark Hayden, "The Ensemble System", Ph.D. thesis, Cornell University, Jan 1998.
- [4] Takako M. Hickey and Robbert van Renesse, "An Execution Service for a Partitionable Low Bandwidth Network", In Proceedings of the Twenty-Ninth International Symposium on Fault-Tolerant Computing, Madison, Wisconsin, USA, June 1999. (Also available as http://www.hep.caltech.edu/~takako/pubs/pex_ftcs.ps)
- [5] Takako M. Hickey, "Augmenting Group Communication to Handle Membership of Larger Groups", in preparation.

Distributed System Simulations

The Caltech group continued the development of the MONARC [1][2] simulation toolset and its validation by simulating the CMS - High Level Trigger Farm (Spring 2000) setup. This first attempt to simulate a large production farm based on Objectivity to store data provided encouraging results in understanding and optimizing large scale distributed systems [3]. Dedicated modules were developed for the simulation framework to allow the study of cost effective solutions in using tapes for different access patterns [4]. Efficient job scheduling policies in very large distributed systems, which evolve dynamically, as the off-line data processing for LHC experiments, is a challenging task. Currently, we are evaluating a possible approach for such a scheduling middle-layer system as a self organizing Neural Network structure which is based on competitive learning from past experience, and which evolves dynamically while trying to optimize the resource utilization and the efficiency for those jobs of high priority.

- [1] http://www.cern.ch/MONARC/sim_tool/
- [2] http://www.cern.ch/clegrand/MONARC/WSC/monarc_wsc2000.pdf
http://www.cern.ch/clegrand/MONARC/CHEP2k/sim_chep.pdf
- [3] http://www.cern.ch/MONARC/sim_tool/Publish/CMS/publish/
http://www.cern.ch/MONARC/sim_tool/Publish/CMS/publish/sim_cms_hlt.pdf
- [4] http://www.cern.ch/MONARC/sim_tool/Publish/TAPE/publish/

PPDG Quarterly Status Report for Fermilab

Date: Sept. 15, 2000

Participants: Jim Amundson, Phil Demar, Don Petravick, [Ruth Pordes](#), Igor Terekhov Rich Wellner, Vicky White

Description of activities related to "High performance site-to-site file replication service"

Fermilab has further developed and deployed scripts for replication and cataloging of Focus data files from 8mm tape at Fermilab to Indiana HPSS system <http://grid.fnal.gov/ThreadCopy/>. We have developed a set of easy to modify and use scripts that replicates data files between two sites, catalogs the transfers and collects and displays statistics about the copies. These scripts can be left to run for several days without attendance. They have been used to copy 2Terabytes of specified Focus data files. Atlas has shown interest in these scripts for extension for their use. The work for this was done by a summer student for three man months.

Description of activities related to "Multi-site cached file access service"

The D0 SAM system now supports global disk cache management. Automatic caching of data files to be read, and automatic routine of data files to be written, is done based on a central meta-data catalog and distributed resource management services. Tests are underway in support of the D0 Monte Carlo Challenges between the SAM installation at IN2P3 and the D0 Central Analysis Systems at Fermilab. Simulated events will be sent from Fermilab to IN2P3 where the full monte carlo is performed – with the resulting output being transferred back to Fermilab for archiving.

We are working with our CMS colleagures at Cern, Caltech and the University of Wisconsin on planning and implementation of multi site file caching for the CMS simulation data – initially for the fall production run and for test activities towards the end of this year. We have deployed Globus and done throughput tests between Fermilab and Cern http://home.fnal.gov/~muzaffar/data_transfer/

Description of activities related to "Development of Grid architecture and API's"

Fermilab and Wisconsin are collaborating on a test bed application integrating existing components of a grid architecture. This first application will handle reading of data files from the SRB or the SAM systems. This application is being used to test and extend the PPDG HRM API whose architecture is defined at <http://gizmo.lbl.gov/ppdg/> (the current IDL is at http://cdcvs.fnal.gov/cgi-bin/public-cvs/cvsweb-public.cgi/ppdg_idl/) It will support determining one or more locations of a requested file and determination of whether it should be transferred from SRB or SAM. Five man months has been spent on this application.

Fermilab will participate in the definition of the PPDG File Replication Service and provide applications to prove and test the proposed interfaces and functionality.

Description of other activities

A CDF front end VME crate and readout system has been set up in the Feynman Computer Center as a test bed for the CDF INFN group for testing of remote monitoring of online data using QOS.

<http://www.cnaf.infn.it/~ferrari/quadis/> Tests will start using this system over the next few months.

Fermilab hosts one of the CVS repositories accessed by collaborators on the PPDG project.

<http://grid.fnal.gov/ppdg/ppdg-cvs/> Access to this repository is readily available.

PPDG Quarterly Status Report for Jefferson Lab

Date: Sept. 15, 2000

Participants: Chip Watson, Jie Chen

Description of activities related to "High performance site-to-site file replication service"

Jefferson Lab has begun to test the use of gsiftp and other tools for transferring experimental data from Jefferson Lab to one or two remote universities. These initial steps have served to understand the use of GSI certificates, and to explore the transfer rates possible to collaborating sites. We are in the process of defining two testbeds: (1) experimental data transfer from Jefferson Lab to a university site involved in analysis of CLAS data; (2) transfer of lattice QCD simulation data between Jefferson Lab and MIT. This second testbed will eventually evolve into a testbed for a multi-site cached file access service. This activity has consumed approximately 2 man-months in this quarter.

Description of activities related to "Development of Grid architecture and API's"

As part of an activity in exploring the design of grid API's, we have written an object oriented API for a replica catalog, patterned after the globus, and to a lesser extent the SRB, non-OO APIs. In particular, a Java binding was described which is based upon the use of Interfaces (abstract classes) and Factory methods, following the pattern used by Java's database package (JDBC) and a number of other Java packages defined by Sun. This strategy is aimed at cleanly separating API definition from implementation, allowing for wrapping pre-existing grid codes, and migrating to new implementations as they appear. A presentation of this API can be found at <http://www.jlab.org/~watson/OOReplicaCatalog.ppt>. This activity has consumed approximately 0.5 man-months.

PPDG Quarterly Status Report for LBNL, Scientific Data Management

Date: Sept. 15, 2000

Participants: Arie Shoshani, Alex Sim, Andreas Mueller, Ekow Otoo

Description of activities related to "High performance site-to-site file replication service"

The Scientific Data Management group at LBNL has been working closely with the SDSC people who developed the Storage Resource Broker (SRB). SRB provides a Grid security and storage infrastructure. The collaboration's goal is to develop a way for SRB to use the HPSS Resource Manager (HRM) developed at LBNL. The API to the HRM was developed, as well as the software to have SRB communicate with the HRM.

Over the last several months, we have been working closely with people at Fermi lab to refine the HRM IDLs so that Fermi can use that as an interface to their SAM storage management system. This resulted in several improvements to the IDL, which are now posted at <http://gizmo.lbl.gov/ppdg/> and at the CVS repository in Fermi at <http://grid.fnal.gov/ppdg/ppdg-cvs/index.html>.

Another accomplishment achieved is the extension the type of file requests that can be sent to the Grid by the client when accessing HRM. In addition to asking for a file to be transferred, the client can ask for files to be pre-staged, for the status of files (how long before they are staged), and the cancellation of a file request. To support this functionality, both the SRB and the HRM components were further developed/modified. The work on SRB was done by SDSC staff, and the work on the HRM by LBNL staff.

Description of activities related to "Multi-site cached file access service"

We continued to develop and use the test environment, where an application client at the University of Wisconsin first contacts the Query Interpreter at LBNL to get the list of file that qualify for its logical query. Then it issues file requests to its local SRB client. The SRB client then contacts the SRB server at LBNL, which in turn requests the HRM (the component that manages file staging from HPSS) to move a file to a staging disk. When this is done, the SRB is notified and it then moves the file in the most efficient way possible to the disk in Wisconsin. This is a major achievement as it proved that a Grid architecture that relies on SRMs is a powerful way to manage Grid storage allocation and coordination. A slide presentation describing this work as well as an introduction to SRMs is posted at: <http://gizmo.lbl.gov/ppdg/>

Description of activities related to "Development of Grid architecture and API's"

LBNL is involved in the development of "Storage Resource Managers" for data grid applications. Our architecture design and implementation is based on our experience with the HENP GC project (where the STACS system was developed). A key concept of adapting this architecture to a distributed grid is the use of Storage Resource Managers (SRMs). Each SRM is associated with a storage resource, such as HPSS, DPSS, or a shared disk cache. The reason that these SRM components are valuable is that the Grid can use these SRMs to request storage reservations, to stage files from tape to a staging disk, and to queue storage transfer requests. This makes it possible to plan and schedule an efficient use of the network as well as take advantage of network bandwidth reservations.

Several meetings took place for the purpose of defining and planning joint work in this project. These include a meeting at Caltech, two at LBNL, and one at Fermi. Summaries of these meetings can be found at <http://gizmo.lbl.gov/ppdg/>, and at <http://www.cacr.caltech.edu/ppdg/>.

Description of other activities

Future plans include:

- 1) The development of the IDL for a generic Disk Resource Manager (DRM). This include the capability to request that DRM will get a file from another DRM (or HRM), that DRM can be asked to pin a file till it is transferred. A document describing the DRM functionality will be distributed for comments.
- 2) We plan to start the development of a DRM. It will be implemented as a CORBA server based on the functionality of the IDL described above.
- 3) We plan to change the HRM-HPSS (an HRM that accesses HPSS) interface we developed previously to accept the new HRM IDL recently defined. In this way, the same HRM interface can b e used to access the HPSS system (currently at LBNL) or the SAM system at Fermi.

PPDG Quarterly Status Report for LBNL physics

Date: Sept. 15, 2000

Participants: Stu Loken, Doug Olson, A. (Sasha) Vaniachine

Description of activities related to "High performance site-to-site file replication service"

A. Vaniachine has prepared a description of the file replication needs and plans for the STAR experiment with regard to data transfer between BNL/RCF and LBNL/NERSC/PDSF (URL here). This was presented at the PPDG collaboration meeting at ANL on July 13,14, 2000. Some additional planning and measurements of network characteristics were carried out during August and September. One particular concept being considered is "store and forward" architecture is beneficial and if so, where the stores should be located.

This effort was about 0.5 person-month during July-Sept. 2000.

Description of activities related to "PPDG Project Coordination"

Doug Olson started as project coordinator in May 2000. Since then he has been running the bi-weekly phone meetings and organized a collaboration meeting hosted at ANL in July. At the July collaboration meeting a decision was taken to implement a file replication service as a first step toward bringing grid services to the experiments. There have been numerous activities following this decision including descriptions of the file replication plans by the experiments, meetings by various working partners to discuss aspects of file replication to help clarify both the functionality of file replication as well as first definitions of programming interfaces for a file replication service. In discussions of file replication, particular between the physicist and computer scientist partners, it became clear that a white paper providing an overall context and vocabulary in which to discuss file replication scenarios will be valuable and this work has started.

There was also a decision to acquire the ppdg.net domain name and use this for hosting the primary web pages (www.ppdg.net), email distribution list and archives for the collaboration. This has been set up and will go "public" in early October following some testing.

This effort was about 1.5 person-months in July-September and \$2K was expended for hardware to host the ppdg.net domain.

PPDG Quarterly Status Report for SDSC

Date: Sept. 15, 2000

Participants: Reagan Moore, Bing Zhu, Arcot Rajasekar

Description of activities related to "High performance site-to-site file replication service and multi-site cached file access"

The Storage Resource Broker (SRB) software that is used by the PPDG group has been upgraded from version 1.1.4 to version 1.1.7. New features include additional S-commands, which were jointly developed by SDSC and LBNL, to support staging and status requests. Version 1.1.7 also provides a table interface for directly accessing object-relational databases, and returning tabular data as formatted records through use of either XML or HTML. A similar interface may be of interest for use with Objectivity.

Performance tests have been conducted on version 1.1.7 of the SRB. Ingestion rates over 10 million data objects per day have been measured on a 4-node E10k server. By scaling the server size, higher data ingestion rates can be achieved.

Currently two SRB 1.1.7 servers are running at LBNL and Wisconsin. The servers are checked on a daily basis, to guarantee that the system is running after machine reboots. In the process, the SRB logs are checked and any bugs are resolved.

A new capability for registering a replicated file into a SRB collection was also integrated into SRB release 1.1.7 for the PPDG group. The new capability allows a data object to be moved independently of the SRB data handling system, and then registered as a replica of an existing data object. Since the SRB does not manage the data movement in this case, the "replica" cannot be guaranteed to be an exact copy of the original.

Description of activities related to "Development of Grid architecture and API's"

SDSC has developed and maintains a GSI enabled FTP interface to the HPSS archival storage system. GSI-FTP is used by the Globus environment to access storage systems. In parallel with this, SDSC is examining the interface needed to support Globus access to SRB managed data collections.

Description of other activities

Development activities include integration of code to tune the TCP window size within SRB server and clients. After analysis of the source code within Iperf, a network tuning tool developed by NLANR, we have decided to use their approach. Tests were conducted between Wisconsin and SDSC using Iperf to validate their window size tuning. The results shows that the file transfer rate can be significantly improved with an appropriate window size compared with using the default window size. We have been given permission by NLANR to use their source code in the SRB.

PPDG Quarterly Status Report for SLAC

Date: Sept. 15, 2000

Participants: Bob Cowles, Andy Hanushevsky, Adil Hasan, Steffen Luitz, David Millsom, Richard Mount, Davide Salomoni.

Description of activities related to "High performance site-to-site file replication service"

Speed Testbed

Within the PPDG framework, SLAC and Caltech have done several high-performance file transfer tests using the experimental NTON (National Transparent Optical Network) infrastructure.

The goal was to demonstrate file transfers at speeds of 100MB/s or more. SLAC and Caltech are connected via NTON with a 2xOC-12 (2x622Mbit/s) ATM link. Using ATM has proved to be not very easy.

Before trying to reach the proposed goal, we tried to understand the limitations of the existing gigabit network cards / operating systems installed on the servers we were about to use. This has been a very useful task, which has shown that we need to use applications where one can adapt the TCP window size and the number of parallel streams; it has also shown significant differences in the throughput one can get from different cards/operating systems, with Sun Servers and Solaris delivering noticeably poorer performance than Intel Pentium IIIs and Linux.

On the WAN side, at SLAC we connected the 2 OC-12 links to 2 ATM cards on a Cisco GSR 12012; the Cisco GSR was then connected to SLAC servers using Gigabit Ethernet links. At Caltech, lacking a router, we have started connecting the 2 OC-12 links to an ATM switch, and from there to 8 OC-3 (155Mbit/s) ATM cards installed on an HP Exemplar, connected to an NSTOR FC Array. The performance with this setup was very low, and in practice only one of the 8 OC-3 adapters on the HP Exemplar was fully utilized.

Caltech then connected a Dual-Pentium III 833 MHz running Windows 2000 with 2 OC-12 ATM adapters to NTON in substitution of the HP Exemplar; this has resulted in a great improvement in the achievable transfer rate. The most recent measurements showed that we could get roughly 80% of an OC-12 link (i.e. ~500 Mbit/s); we have not been able to go any higher, and this measurement was only memory-to-memory. The reason why we could not utilize more than a single OC-12 was under investigation when NTON went dead for a very long time (it is still not operational between SLAC and Caltech as of 9/25/00). Given also the current NTON plans to migrate hopefully soon to OC-48 POS, we have decided to wait until the new infrastructure is in place. SLAC already has an OC-48 POS cards for its Cisco GSR router, and Caltech has ordered a Juniper router to support this speed as well. We expect to resume the tests as soon as the new equipment and network connectivity is available.

Presentation at the July PPDG Meeting:

http://www-rnc.lbl.gov/PPDG/mtgs/13jul00-anl/salamoni/High_Perf_PPDG_Jul2000.ppt

BaBar Intercontinental Testbed

Throughout the period July – September, BaBar has been transferring about 300 Gbytes per week between SLAC and IN2P3 Lyon. This work is almost totally supported by US and French funding for the BaBar program, but close ties with PPDG promote a valuable exchange of tools and experience. Experience has been gained with enhanced ftp supporting large windows and multiple IP streams (bbftp written by Gilles Farache of CCIN2P3, Lyon). Experience has also been gained in the many ‘second order’ issues of reliability, error handling, resource management that dominate the work of the BaBar scientist who manage the SLAC-Lyon transfers.

After discussions with the Globus team in August, trial deployments are planned of new Globus components supporting a replica catalog and streamed ftp.

Personpower used at SLAC July – September: 2 person-months.

Equipment bought July – September: Dell servers with gigabit Ethernet interfaces: \$12k

Description of other activities

Organized a PPDG File Replication Meeting at SLAC on September 20, 2000.

<http://www-rnc.lbl.gov/PPDG/mtgs/20sep00-slac/>

PPDG Quarterly Status Report for Wisconsin

Date: Oct. 15, 2000

Participants: Peter Couvares , Tevfik Kosar , Miron Livny

Description of activities related to "Multi-site cached file access service"

We continued our effort to develop a testbed that will enable the collaboration to perform end-to-end tests of the different file access services provided by members of the PPDG. The testbed is interfaced to our local production environment and allows us to combine data movement and processing. As part of this effort we participated in the following activities:

1. Visited the LBNL group to discuss protocol and API issues related to the HRM and DRM components.
2. Had several meetings with the Fermi group to discuss protocol and API issues related to interfacing our local environment with the SAM storage system at Fermi.
3. Designed and implement a testbed that links our local Condor pool with SAM. This links enables to move data managed by SAM to Wisconsin.
4. Integrated the new staging services added to SRM in our environment and started to test them jointly with the LBNL group.
5. Developed scripts for FTP based data transfers between Wisconsin and Caltech to support the CMS production runs on the Wisconsin Condor pool.

Distributed Computing and Data Management

We continued to support CMS production runs from Caltech on our Condor pool. We have been working closely with the Caltech group on developing the infrastructure needed to support these runs. We added support for GSIFTP services to our local software and conduct some intial tests involving the GSIFTP server at Caltech.