

# Particle Physics Data Grid:

## Proposal to DOE HENP for FY 2000 Funding

**Principal Investigator** Harvey B. Newman, California Institute of Technology  
 (University Collaborators): 1200 East California Blvd., Pasadena, CA 91125  
 (626)-395-6656; Fax (626)-795-3951;  
 Email: newman@hep.caltech.edu

**Principal Investigator** Richard P. Mount, Stanford Linear Accelerator Center  
 (DoE Laboratory Mail Stop 97, P.O. Box 4349, Stanford CA 94309  
 Collaborators) (650)-926-2467; Fax (650)-926-3329;  
 Email: richard.mount@stanford.edu

**Collaborators:**

California Institute of Technology	<b>Harvey B. Newman</b> , Julian J. Bunn, Koen Holtman, Asad Samar, Takako Hickey, Iosif Legrand, Vladimir Litvin, Philippe Galvez, James C.T. Pool, Roy Williams
Argonne National Laboratory	<b>Ian Foster</b> , Steven Tuecke <b>Lawrence Price</b> , David Malon, Ed May
Lawrence Berkeley National Laboratory	<b>Stewart C. Loken</b> , Ian Hinchcliffe, Doug Olson, Alexandre Vaniachine <b>Arie Shoshani</b> , Andreas Mueller, Alex Sim, John Wu
Brookhaven National Laboratory	<b>Bruce Gibbard</b> , Richard Baker, Torre Wenaus
Fermi National Laboratory	<b>Victoria White</b> , Philip Demar, Donald Petravick <b>Matthias Kasemann</b> , Ruth Pordes, James Amundson, Rich Wellner, Igor Terekhov, Shahzad Muzaffar
University of Florida	<b>Paul Avery</b>
San Diego Supercomputer Center	<b>Margaret Simmons</b> , Reagan Moore
Stanford Linear Accelerator Center	<b>Richard P. Mount</b> , Les Cottrell, Andrew Hanushevsky, David Millsom, Davide Salomoni
Thomas Jefferson National Accelerator Facility	<b>Chip Watson</b> , Ian Bird, Jie Chen
University of Wisconsin	<b>Miron Livny</b> , Peter Couvares, Tefvik Kosar

# Contents

Introduction.....	3
HENP's Need for a Data-Intensive Grid .....	3
Project Goals .....	4
Progress: August 1999 to April 2000.....	5
FY 2000 Workplan.....	6
Deploying PPDG Services .....	6
Tests and Demonstrations.....	8
Development of Architecture and Tools .....	11
Longer-Term Vision and Relationship to GriPhyN.....	11
Proposed Tasks per Site: FY 2000 .....	14
Appendix A .....	15
PPDG Status Report, April 2000.....	15

# **Particle Physics Data Grid: Proposal to DOE HENP for FY 2000 Funding**

April 21, 2000

## **Introduction**

The Particle Physics Data Grid (PPDG) proposes to address the long-term data-management needs of high-energy and nuclear physics using a pragmatic rapid deployment of advanced services as the driver for longer-term research and development.

The PPDG project has been underway since August 1999, funded by the DOE Next Generation Internet program. PPDG was planned as a three-year program. The project is a collaborative effort between physicists and computer scientists at six DOE HENP laboratories (Argonne, Berkeley, Brookhaven, Fermilab, Jefferson and SLAC) and Caltech, SDSC and the University of Wisconsin. From its inception, the PPDG has focused on making rapid use of middleware already developed by the collaborators.

This FY00 proposal seeks to build upon the progress of the first year of the PPDG work by extending the development of architecture and tools, by further deployment of PPDG services within the community, and via the implementation of key demonstrations. Funding in FY01 and beyond will also be sought to continue the vital contribution that PPDG can make to HENP exploitation of grid technology.

## **HENP's Need for a Data-Intensive Grid**

High Energy and Nuclear Physics experiments have traditionally pioneered in the need for innovative technology to permit effective analysis of very large quantities of data by geographically dispersed researchers who must nevertheless work closely together. The current (BaBar, Run II, RHIC, JLAB) and future (LHC) large-scale experiments are continuing this trend and indeed have major needs for new capabilities which must be met to permit the full physics potential of these projects to be realized. Demanding data analysis requirements coupled to emerging computational and networking capabilities are giving rise today to a new class of advanced network-based applications. These applications require a coordinated use of distributed computers, high-speed networks, storage resources and sophisticated middleware collectively referred to as a "grid". Much as operating systems have provided a unifying environment for processes on single computers, the "grid" architecture provides a unified approach to applications running on a network of widely separated computing systems. Recognizing the importance of such applications, various federal agencies have started projects to create a grid infrastructure, including the National Science Foundation's Partnerships in Advanced Computational Infrastructure, NASA's Information Power Grid, and the DOE Science Grid Testbed.

The concept of a grid has been embraced by many of the HENP experiments in the United States and in Europe. Such a grid has many features that make it especially suitable for the HENP community with its large, international collaborations and a large volume of data that must be managed, analyzed and displayed.

The Particle Physics Data Grid will eventually permit transparent access to data and facilities across the entire HENP community. Experiment-wide catalogs and transparent caching will ensure that all data requests from widely distributed sites are satisfied in the minimum time. New datasets created as part of analysis can be entered into an experiment-wide catalog so that others may use the same data without re-processing. Analysis and Monte Carlo jobs that require more processing power than is available locally will be sent to any compatible facilities available on the grid.

The ATLAS and CMS experiments at the CERN LHC program will rely on computing resources that are widely distributed around the world among several hundred collaborators. A substantial fraction of the total computing for ATLAS and CMS will be done in this network distributed environment. The diversity of the administrative and technical resources is very large, thus a unified architecture and implementation seen by the ATLAS and CMS computing applications is required. ATLAS and CMS believe a grid to be the best approach. The ATLAS and CMS software architecture is committed to a design that will exploit the grid environment.

BaBar, CDF, D0 and the RHIC experiments also recognize the potentially revolutionary benefits of data grid services and are committed to working with the PPDG to perform early trials of new services.

Collaboration tools can be fully integrated into the analysis framework so that groups working on common problems can meet together in a virtual environment and share results. The tools include conferencing, shared visualization and analysis control, and shared notebooks and documents.

## Project Goals

Within the broader vision of grid-enabled data management and access for HENP the specific goals of the Particle Physics Data Grid (PPDG) project are to:

- u Design, develop, and deploy a network and middleware infrastructure capable of supporting data analysis and data flow patterns common to the many physics experiments represented by the participants
- u Adapt experiment-specific software to operate in this wide-area environment and to exploit this infrastructure

To accomplish these goals, the PPDG will deploy two critical services:

- u High-Speed Site-to-Site File Replication Service
- u Multi-Site Cached File Access Service  
(based on deployment of file replica cataloging, transparent cache management, and data movement middleware)

The PPDG collaboration intends to maintain its pragmatic approach centered on moving rapidly from ideas to services. The PPDG will work closely with the BaBar, RHIC and Fermilab Run II experiments to validate ideas by bringing real benefit to physics analysis. Successful deployment and enhancement of services will lean heavily on an ability to monitor network and system activities associated with data transport. The PPDG will develop new network monitoring capabilities that become identified as critical for its success. Collaboration with European sites is integral to the US HENP program and to the PPDG. Several projects are underway or proposed for grid development in Europe. PPDG will provide the main opportunity for U.S. cooperation with

the European projects on testbeds, especially where there are both U.S. and European collaborators on many of our participating experiments.

The long-term goals of the PPDG are to work towards the vision of automated management of data and tasks that has been given the name "Virtual Data". This long-term vision is also that of the GriPhyN<sup>1</sup> Grid Physics Network NSF ITR<sup>2</sup> proposal in which many PPDG collaborators are involved. The goal of virtual data is ambitious and will only be fully achieved through the joint efforts of PPDG and GriPhyN. The pragmatic approach of PPDG is perfectly complementary to the computer-science focus of the GriPhyN ITR proposal. PPDG will work directly with collaborators from ATLAS and CMS to implement, utilize and maintain test-bed grids.

## Progress: August 1999 to April 2000

As detailed in the April 2000 PPDG Status Report (Appendix A), progress towards a data-intensive grid has been made towards each of the above goals.

The PPDG has had access to major computer centers at many of the collaborating sites and to middleware already developed by its collaborators. This has allowed partial but functional implementations of the PPDG architecture to be tested as the architecture was being developed. The HENP Grand Challenge<sup>3</sup> project provided much of the starting point for the PPDG architecture. The developers of the Globus<sup>4</sup> grid toolkit and the Storage Request Broker (SRB)<sup>5</sup> have both collaborated and constructively competed to provide some of the first PPDG services.

Out of this experience and set of components has come an architecture for distributed data handling -- a conceptual view of the various components needed within the data grid. As part of this architecture the collaboration has defined a set of application programming interfaces (APIs) to modularize the system and permit implementation of the architecture in the limited time available for the project. The collaboration has also developed a coordinated work plan under which computer science collaborators integrate grid and data tools into the desired architecture, and HENP collaborators begin to use the data grid system to test and demonstrate data handling capabilities that will be needed in an operational system.

ATLAS and CMS test-beam and simulation data now are in use for trials of multi-site cached file access; BaBar and D0 data are in use for the first point-to-point data file replication services. The Netlogger<sup>6</sup> tool is being exploited to provide the network monitoring essential to achieving and maintaining high performance data flows. The PingER<sup>7</sup> and Surveyor<sup>8</sup> tools are being deployed at all PPDG sites to assist in setting expectations for network performance and to provide information for planning and troubleshooting.

---

<sup>1</sup> GriPhyN: [http://www.phys.ufl.edu/~avery/mre/proposal\\_final.doc](http://www.phys.ufl.edu/~avery/mre/proposal_final.doc)

<sup>2</sup> ITR: Information Technology Research Program of the NSF Directorate for Computer and Information Sciences and Engineering.

<sup>3</sup> HENP Grand Challenge project on data management: <http://www-mc.lbl.gov/GC/>

<sup>4</sup> Globus Project: <http://www.globus.org/>

<sup>5</sup> Storage Request Broker: <http://www.npaci.edu/DICE/SRB/index.html>

<sup>6</sup> Netlogger: <http://www-didc.lbl.gov/NetLogger/>

<sup>7</sup> PingER: Internet response and packet loss monitoring and reporting tools, <http://www-iepm.slac.stanford.edu/pinger>

<sup>8</sup> Surveyor: Program for the deployment of dedicated internet monitoring hardware, <http://www.advanced.org/surveyor/>

# FY 2000 Workplan

The FY2000 proposal has three principal components:

1. Deploying PPDG Services  
In its first nine months the PPDG has been able to demonstrate multi-site cached file access and will soon have demonstrated high-speed point-to-point file transfer. In keeping with the PPDG's close ties to HENP experiments, these demonstrations will be turned into services to current experiments and to LHC test-beam and simulation activities. The real world of HENP experiments, always very different from that of a demonstration, will yield vital direction-setting insight.
2. Tests and Demonstrations  
Continued tests of middleware aiming at enhanced capabilities and performance will remain a major activity. Most tests will be able to exploit the networks, storage and computers already available at the participating sites. Instrumentation and monitoring tools are vital in both the testing and deployment phases.
3. Development of architecture and tools  
The PPDG collaboration includes computer scientists who have developed the tools that will be tested and deployed. Experience from tests and deployment will be used to evolve the architectural design for PPDG software components and to stimulate focused work on new or improved middleware products and monitoring tools for which there is a demonstrated need.

As it moves into a phase of providing real services to experiments, the effectiveness of PPDG will be assured by appointing a full-time project coordinator reporting to the existing PPDG scientific leadership.

## Deploying PPDG Services

The primary goal of this effort is to deploy the PPDG Software within a number of the major HENP experiments. These efforts will involve multiple laboratory and university sites and will focus on the analysis of testbeam or simulated data. The result will be experience and feedback on the use of grid technology in real experimental analysis situations and the early use of grid tools across HENP that will grow into a fully coordinated grid implementation across all of HENP.

### **ATLAS**

The ATLAS groups will focus on distributed analysis of TileCal testbeam data using the new PPDG tools. In addition, the groups will use the tools to analyze GEANT Monte Carlo to test detector reconstruction code. They will develop plans for use of the ATLAS Monte Carlo with the ATLAS control-framework currently under development across the PPDG testbeds. As part of this effort the ANL group will establish a disk cache with a GSIFTP<sup>9</sup> interface at ANL.

The ATLAS core database project, under non-PPDG auspices, is developing software for database replication and distribution, and for appending data to remote databases. This software will be deployed for distribution of this year's testbeam data to remote sites. ATLAS PPDG participants plan to integrate this software with PPDG middleware, so that standard ATLAS software will both take advantage of PPDG's "High-Speed Site-to-Site File Replication Service" between grid-enabled sites, and make use of PPDG's file replica cataloging and management. Subsequent development will exploit PPDG's "Multi-Site Cached File Access Service" to provide data access for physicists at sites remote from the testbeam data caches. This effort will build

---

<sup>9</sup> GSIFTP: Globus Security Infrastructure File Transfer Protocol

upon last year's PPDG work on ATLAS tile calorimeter testbeam data distribution, providing integration with production ATLAS data distribution services, and extension to support a much broader data analysis community.

### **CMS**

CMS will incorporate the exploitation of PPDG tools into its data management plans for this year. The plans include the production of the order of 10 Terabytes of fully simulated and reconstructed data, with several Terabytes of data in the Objectivity ODBMS<sup>10</sup>. In the process of generating this data, including the handling of overlapping background events characteristic of the high luminosity environment of the LHC, several hundred Terabytes must be handled. Data is being generated at several sites in the US, including PPDG participants at Caltech, Fermilab, and the University of Wisconsin (and possibly extended to UC Davis), as well as at CERN. The US centers will make available more than 10 Terabytes of mass storage and a few Terabytes of disk cache to support a distributed base of user analysis programs. Scripts are available to provide scheduled synchronization of local data bases and caches with the main CERN repository.

### **CDF**

The CDF experiment will utilize some of the PPDG tools to permit scientists at Berkeley to access Run II data stored at Fermilab. This will require the implementation of Storage Resource Managers (SRM) and integration of these SRMs with the CDF Data Handling System (DHS). Because the CDF DHS is already designed as a file-delivery system, this integration should not require significant redesign. The Berkeley and Fermilab groups will also install NetLogger at both sites to measure network bandwidth achieved in end-to-end data analysis tasks.

In addition, the Berkeley group is now one of the major producers of the CDF simulation data. At present, these data are moved to storage at Fermilab. In the future such simulated data may be moved to Fermilab or they may be stored at Berkeley and entered into a global catalog that is managed using PPDG tools.

### **D0**

The D0 experiment uses the SAM<sup>11</sup> system for data access and resource management. SAM has already been used as a component of PPDG and work is underway to further integrate SAM with additional storage management systems, with Condor<sup>12</sup> and Condor resource management, and with the Globus toolkit. Multi-site caching of data and further development of the resource management component of PPDG will be carried out with D0 Monte Carlo data.

### **STAR**

The STAR experiment has its major computational sites at the RHIC Computing Facility (RCF) at Brookhaven Lab. and at the NERSC facility at Berkeley Lab. At present there are significant volumes of data being copied between these facilities. Over the past 2 years this has consisted of about 5 Terabytes of simulated data and detector test data. Over the next two years this is expected to increase to about 20 Terabytes of data per year replicated between these centers. The services of the PPDG are very well suited to this task. The participants in PPDG from BNL and LBNL will establish the LBNL-BNL link in the data grid and work with STAR to establish this data replication service. In the future (FY01 and beyond) this service will be extended to include other RHIC experiments as well as collaborators at university sites.

### **JLAB**

The participants at Jefferson Lab will work to deploy the PPDG tools within the Experimental Nuclear Physics community. Two university user sites will take part in testing of GSIFTP and other grid tools with a disk cache at Jefferson Lab of data from the CLAS detector. In addition, file

---

<sup>10</sup> Object Database Management System

<sup>11</sup> SAM: Sequential Access via Metadata, <http://d0ora2.fnal.gov/sam/>

<sup>12</sup> Condor Project, High Throughput Computing, <http://www.cs.wisc.edu/condor/>

replication and caching services will be deployed in support of the Lattice Hadron Physics Initiative, a multi-site lattice QCD computing effort with shared data production and storage at Jefferson Lab and MIT.

### ***BaBar***

BaBar is already using *ad hoc* versions of the PPDG site-to-site file replication service. Production service will be deployed initially between SLAC and CCIN2P3 Lyon at the relatively low speed (around 4 Megabytes/s) possible over the current US-CERN-CCIN2P3 link.

### ***General***

The PPDG collaboration will work to develop GSIFTP into a data transfer utility that supports truly excellent performance in a variety of settings. One approach is to provide GSIFTP with the ability to select an optimal window size automatically, based on predicted network performance.

## **Tests and Demonstrations**

### ***High-speed site-to-site file replication***

Caltech and SLAC are focusing on meeting or exceeding the 100 Megabytes/sec PPDG milestone using NTON<sup>13</sup>. They are each setting up servers equipped with fiber channel RAID arrays, two OC-12<sup>14</sup> Fore Systems interfaces, and multiple 64 bit 66 MHz PCI buses, and (for the near future) Gigabit Ethernet connections to meet this goal. In the near term, the Caltech Exemplar (~3000 SpecInt95) will be used to combine data transport and some processing during the milestone tests. Various data transfer tools including the BBFTP, GSIFTP and the SLAC-developed sfc<sup>15</sup> will be tried out. As soon as possible, these tests will be expanded to include ESnet sites (LBNL, ANL and Fermilab).

The urgent need for automated bi-directional replication between SLAC and CCIN2P3 Lyon will be exploited to demonstrate early versions of automated devices. Automated replication of databases also will be implemented between CERN, Caltech and Fermilab in support of CMS data challenges, and in a CMS interactive analysis prototype accessing data seamlessly from an object database<sup>16</sup>.

### ***Multi-site cached file access***

Largely independent of PPDG and other grid development efforts, ATLAS will soon (once again) begin generating massive amounts of simulation data at widely distributed sites. A goal of the ATLAS database effort is to provide an environment in which all of this data is available from anywhere, without the necessity of moving all data to a central store. ATLAS PPDG collaborators plan to use PPDG middleware to build a wide-area storage grid, in which generated data are centrally registered (possibly at more than one registry), but are not required to be moved to a single site, though replication will be possible. PPDG and related grid middleware would identify the appropriate data source(s) for a physics client, and would handle the relevant site authorization, data access, and data delivery.

---

<sup>13</sup> NTON: National Transparent Optical Network, a fiber-optic network in the Western US dedicated to network research, <http://www.ntonc.org/>

<sup>14</sup> OC-12: Optical Carrier link at 622 Mbits/s

<sup>15</sup> sfc: Secure Fast CoPy, <http://www.slac.stanford.edu/~abh/sfc/>

<sup>16</sup> To be demonstrated between Caltech, CERN and Yokohama at INET2000 in July 2000

Caltech, Fermilab and the University of Wisconsin will integrate the CMS framework, ORCA<sup>17</sup>, with the PPDG middleware to provide a demonstration test bed to support user program access to the distributed data storage system and incorporate the operation of more intelligent data movement and caching techniques. The LBNL HRM<sup>18</sup> API will be modified and extended to interface to the Fermilab mass storage system, Enstore<sup>19</sup>, and thus allow support for a common user access to this store of simulation data. PPDG middleware data caching policies and resource management, including for example the Condor<sup>20</sup> teams request planning and execution, the Enstore caching, and SAM resource optimization, will be integrated into the data access and retrieval layers. A vertically integrated interactive analysis prototype based on CMS' IGUANA<sup>21</sup> system, accessing an ensemble of networked databases over transoceanic links, is in preparation and is planned to be demonstrated, embedded in a high quality remote collaborative environment, at the Inet2000<sup>22</sup> conference in July 2000.

This work augments the development efforts in CMS and CMS GriPhyN by continuing to provide operational test beds to understand, measure and extend the efficiency gains from incorporating such techniques into the analysis environment.

Figure 1 indicates the close relationship between PPDG and the ATLAS/CMS component of GriPhyN. This sites named in bold are PPDG members. All sites apart from JLAB host members of the GriPhyN collaboration.

---

<sup>17</sup> ORCA, CMS OO reconstruction, <http://cmsdoc.cern.ch/orca/>

<sup>18</sup> HPSS Resource Manager, an interface to HPSS developed at LBNL

<sup>19</sup> Enstore: Fermilab-developed mass storage system, <http://www-hppc.fnal.gov/enstore/>

<sup>20</sup> Condor Project, High Throughput Computing, <http://www.cs.wisc.edu/condor/>

<sup>21</sup> IGUANA: CMS Interactive Graphical User Analysis

<sup>22</sup> Inet2000: <http://www.inet2000.com/public/>

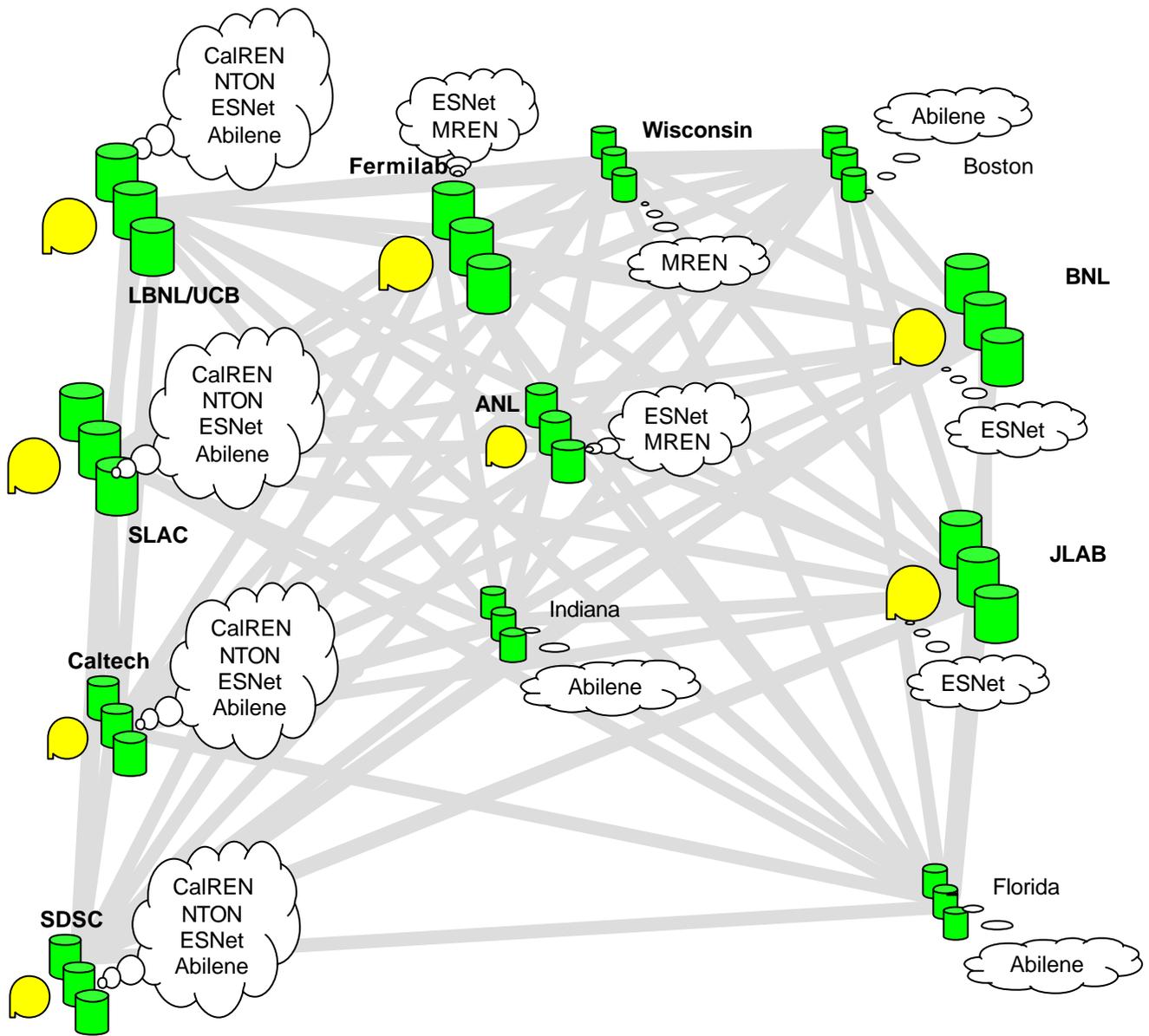


Figure 1: Sites participating in PPDG and GriPhyN/LHC (PPDG sites are in bold). The gray lines indicate the global access to data that is the goal of the PPDG. CERN and CCIN2P3 Lyon are also collaborating with PPDG members

## Development of Architecture and Tools

A major goal of the next six months is to have the client application request files from several locations on the grid, including files stored at LBNL on HPSS<sup>23</sup>, and at Fermilab, ANL, and SDSC on disk caches. This component will be developed by the U. Wisconsin staff in collaboration with ANL, LBNL and the SRB group. The goal is to make decisions on file transfers so as to optimize the network use and the throughput to the clients.

It is also important to extend the type of requests that will be sent to the grid by the client. 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 cancellation of a file request. To support this functionality, both the SRB and the HRM components need to be further developed or modified. The work on SRB will be done by SDSC staff, and the work on the HRM by LBNL staff.

NetLogger and other tools will be used to characterize the achievable transaction and transmission rates for transferring files over the network after they are staged by the HRM. Experience indicates that the OC-12 network connection between LBNL and ANL needs constant monitoring to ensure that it provides high-speed performance. In addition, new program interfaces will be created to obtain additional network performance information from sources such as MDS<sup>24</sup>, NWS<sup>25</sup>, etc.

Other efforts in the project aimed at fast file transfers include the use of DPSS<sup>26</sup> at LBNL as a staging disk for HPSS files. A fast parallel transfer capability (using multiple sockets and large FTP window size) is used to move data from DPSS to the client.

In the future, the collaboration will deliver a robust, high-performance set of replica catalog and data movement functions, and an API for programs to access the replica catalog.

## Longer-Term Vision and Relationship to GriPhyN

Due, in large measure, to the stimulus of the first months of collaborative work on PPDG, both the computer science and the physics communities identified *virtual data* as a vision to steer the development of Data Grids serving the physical sciences. The GriPhyN collaboration, which includes many PPDG members, has submitted an "Information Technology Research" proposal to NSF aimed at funding the computer science research that will be vital to move towards a *Petascale Virtual-Data Grid* on the LHC timescale.

The concept of *virtual data* recognizes that all except irreproducible raw experimental data need 'exist' physically only as the specification for how they may be derived. In high-energy physics today, over 90% of data access is to derived data. Thus the grid may instantiate zero, one, or many copies of derivable data depending on probable demand and the relative costs of computation, storage, and transport. The virtual-data grid will ensure that computation is performed by 'agents' at the optimum locations. Full implementation of the virtual-data grid would maximize the speed and flexibility of physics analysis. A user's view of the virtual-data grid architecture is show in Figure 2.

---

<sup>23</sup> HPSS: High Performance Storage System, <http://www.sdsc.edu/hpss/>

<sup>24</sup> MDS: Metacomputing Directory Service, <http://www-fp.globus.org/mds/>

<sup>25</sup> NWS: Network Weather Service, <http://nws.npaci.edu/NWS/>

<sup>26</sup> DPSS: Distributed Parallel Storage System, <http://www-itq.lbl.gov/DPSS/DPSS.html>

To move towards a realization of the wider virtual-data grid vision it is essential to remain in close contact with the reality of supporting current needs with the best existing tools. This is the key role of PPDG. The PPDG focus has been and will be on gaining insight by providing real grid support for current experiments. Vital to this insight will be the deployment of monitoring and instrumentation so that the inevitable exploration of the failure modes of the PPDG middleware leads to understanding of value to both PPDG and GriPhyN.

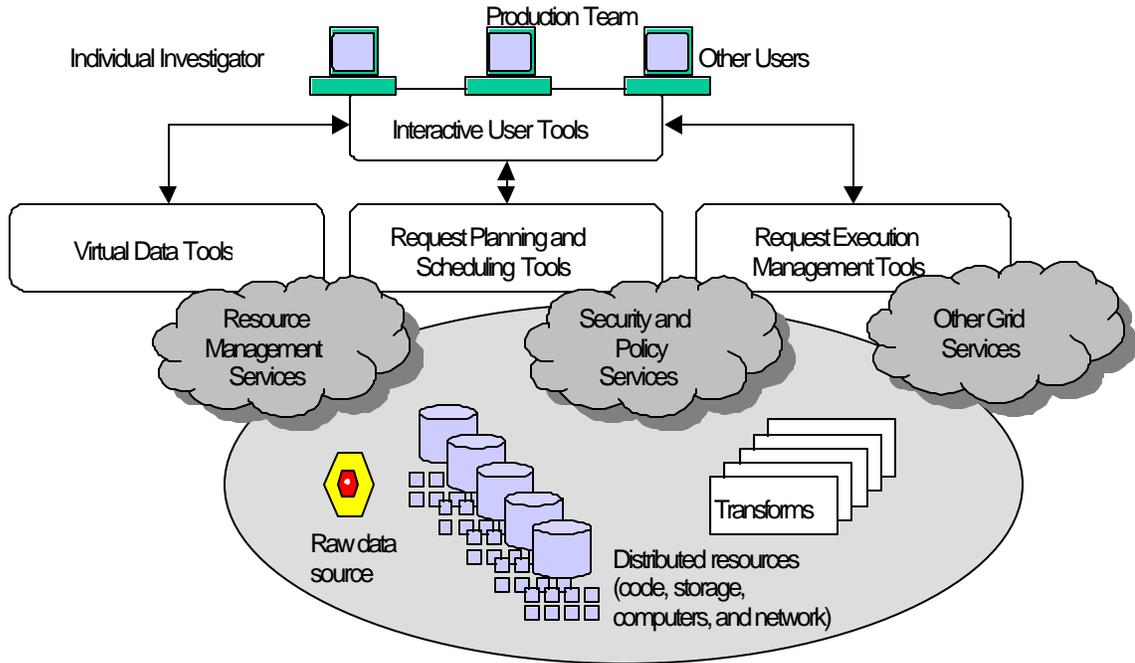


Figure 2: User view of "Petascale Virtual Data Grid" architecture, showing the diverse users that compete for resources, the tools used to formulate and execute requests, the services used to coordinate use of Grid resources, and the resources themselves.

The PIs of PPDG (Mount, Newman) and of GriPhyN (Avery, Foster) will set up a PPDG-GriPhyN Coordination Board to ensure that the programs of work remain as complementary as they are now. The PPDG Project Coordinator and the GriPhyN Project Coordinator will be tasked to ensure that PPDG facilities and the results of PPDG experience are available to GriPhyN. Conversely, as new tools are created within GriPhyN, they will be evaluated by PPDG.

Specific developments currently planned for FY 2001 on are:

- Development of a generalized file-mover framework that is aware of, and can effectively exploit the relevant types of network service including a high-priority, low latency service for control functions and multiple high latency (or scheduled availability) bulk transfer services;

- Implementation/generalization of the cataloging, resource broker, storage resource managers and matchmaking services needed as foundations for both transparent write access and agent technology;
- Implementation of transparent write access for files;
- Implementation of limited support for 'agents': automatic scheduling of data analysis operations to the most appropriate CPUs at Data Grid sites;
- Implementation of distributed resource management for the Data Grid. This will require that network and storage systems at each site be instrumented to support resource discovery;
- Instrumentation of all Data Grid components in support of a systematic approach to measurement of and modeling of Data Grid behavior. Modeling studies are currently expected to be supported by other projects that will be closely coordinated with the work on the Data Grid;
- Major efforts on robustness and rapid problem diagnosis, both at the component level and at the architectural level;

Continuing efforts to support the services, such as transparent support for persistent objects, most appropriate for end users performing data analysis. We expect that the majority of this support will come from other sources such as the particle-physics experiments themselves. However, our focus on ensuring that useful services are really delivered will remain.

## Proposed Tasks per Site: FY 2000

ANL	<p>Establish disk cache with GSIFTP interface</p> <p>Operate PPDG services for ATLAS TileCal testbeam data</p> <p>Continue architecture development</p> <p>Enhance Globus enhancements for PPDG</p>
BNL	<p>Set up Grid node (RCF) and install and operate file-replication software (STAR/BNL with STAR/LBNL)</p> <p>Use PPDG services to grid-enable US ATLAS Tier 1 resources</p>
Caltech	<p>Test PPDG services for simulated CMS data</p> <p>Globus enhancements to PPDG</p> <p>Instrumenting PPDG testbed nodes with SNMP-based network monitoring and host workload tracking</p>
Fermilab	<p>Continue integration of SRB, Condor, SAM and parts of Globus toolkit into PPDG</p> <p>Set up PPDG data distribution for ORCA Production data for HLT milestones</p> <p>Operate PPDG file-replication services for ORCA production of CMS testbeam data</p> <p>Test and support PPDG services for simulated CMS data</p> <p>Install Netlogger</p>
U. Florida	<p>Ensure coordination with GriPhyN</p>
JLAB	<p>Set up PPDG data distribution for CLAS data and for the Lattice Hadron Physics Initiative</p>
LBNL	<p>Set up Grid node (RCF) and install and operate file-replication software (STAR/LBNL with STAR/BNL)</p> <p>Implement PPDG SRMs for CDF Data Handling System</p> <p>Develop Netlogger for PPDG</p> <p>Continue architecture development</p>
SDSC	<p>Enhance SRB for PPDG</p> <p>Participate in SRB deployment</p>
SLAC	<p>Operate site-to-site file replication service (with CCIN2P3 Lyon)</p> <p>Set up 100 Megabytes/s testbed (with Caltech)</p> <p>Develop improved file-replication middleware</p> <p>Install Netlogger</p>
U. Wisconsin	<p>Continue architecture development</p> <p>Test Multimulti-site cached file access tests</p>
Project Coordinator	<p>To be appointed at SLAC or LBNL</p>

# Appendix A

## PPDG Status Report, April 2000

Within the broader vision of grid-enabled data management and access for HENP, the specific goals of the Particle Physics Data Grid (PPDG) are the following:

- u Design, develop, and deploy a network and middleware infrastructure capable of supporting data analysis and data flow patterns common to the many particle physics experiments represented by the participants
- u Adapt experiment-specific software to operate in this wide-area environment and to exploit this infrastructure

Instantiate and deliver an operating infrastructure for distributed data analysis by participating physics experiments

To accomplish these goals, the PPDG will deploy two critical services:

- u High-Speed Site-to-Site File Replication Service
- u Multi-Site Cached File Access Service  
(based on deployment of file replica cataloging, transparent cache management, and data movement middleware)

### Progress in PPDG

The initial focus of PPDG work was the definition of an architecture for distributed data handling. This was accomplished through a series of meetings at a number of PPDG sites. Summaries of these meetings can be found at <http://gizmo.lbl.gov/ppdg>, and at <http://www.cacr.caltech.edu/ppdg/>.

These meetings formed the basis for a coordinated work plan under which Computer Science collaborators integrated grid and data tools into the desired architecture and HENP collaborators began to use the data grid system to test and demonstrate data handling capabilities that will be needed in an operational system. Biweekly telephone conferences have provided global coordination, while several specialized subgroups have had their own meetings, both in person and via remote technologies.

The collaboration then completed an evaluation of available tools and defined a set of APIs that would permit implementation of the architecture in the limited time available for the project.

Another important component of the recent PPDG effort has been to work with high-energy physics data using the currently available tools. This permits the collaboration to investigate the tools and their applicability to HENP. In addition, it gives the collaboration valuable expertise that will be needed when it comes time to deploy the grid for HENP. We now have in place sets of ATLAS and CMS data with which to test the PPDG Architecture and Tools.

To understand the results of the data movement experiments it is necessary to monitor the state of the underlying network infrastructure. To accomplish this, the collaboration has also deployed tools to monitor the state of the underlying network infrastructure.

This report outlines progress in the areas: realizing the architecture, data movement tests and network measurement. The remaining work is summarized in a separate section.

### Realizing the Architecture:

This is the focus of the Computer Science component of the collaboration at LBNL, SDSC, Wisconsin and Argonne. The Scientific Data Management group at LBNL has addressed the problem of "request management", the software infrastructure to support distributed data intensive applications on a wide-area grid. The emphasis is on the management of moving files efficiently from any storage resource on the grid to the application that needs these files.

The architecture design implementation is based on experience with the HENP Grand Challenge project. 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. These SRM components are valuable because 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.

The Storage Request Broker (SRB) provides a data handling system for collection-based access to remote storage systems. The SRB supports caching of data sets, replication of data sets, and supports interoperability between SRMs. The SRB also uses the standard (ANL/ISI-developed) Grid Security Infrastructure to authenticate users to a collection and then to authenticate collection access to the remote storage infrastructure. The approach 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. A test environment was set up, where an application client at the University of Wisconsin first contacts the Query Interpreter at LBNL to get the list of files 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 proved that a Grid architecture that relies on SRMs is a powerful way to manage Grid storage allocation and coordination.

Another recent accomplishment was the replication of a subset of the files that are on HPSS at LBNL at SDSC and Fermilab and ANL. The replication was managed through simple SRB commands that directed the creation on copies of the data sets at each site. In addition, SRB was installed at ANL, Fermilab, U. Wisconsin, and Caltech to facilitate file sharing. This software will be used for multi-site transfer tests.

SDSC has developed a SRB/HRM interface to support access to the HPSS Resource Manager at LBNL. The interface has been used to investigate interactions between local storage management policies and the global storage requests. In collaboration with LBNL, they have identified a need to request staging of data, and a user interface is being developed. Design issues include specification of the local cache policy for amount of data that can be cached, specification of time estimates for creation of the cached copy, and error returns for handling deletion of data sets from the cache.

SDSC continues development of new versions of the SRB. Version 1.1.7 is being integrated during March 2000, and will include support for authentication through the Grid Security Infrastructure, and support for returning XML-annotated information from the metadata catalog.

The Condor team at the University of Wisconsin has implemented a technique for "request planning and execution", the software structure that runs at the client site and is responsible for feeding the application with data files. They have developed a simple version of a software module that can locate files needed by a high level query and transfer them to a local disk cache at Wisconsin. They are running tests to evaluate the APIs, functionality and robustness of the different software layers. These tests involved the installation of new software on their machines and the distribution of data files among several PPDG sites. They have also installed most of the available network monitoring tools on their machines.

The Distributed Systems Lab at ANL is working with USC/ISI to develop tools for managing data distributed across a wide area network. They have chosen as their initial focus three important problems: security (the Grid Security Infrastructure, referenced above in the context of SRB), standards-based mechanisms for high-speed data movement, and tools for the creation and management of file replicas in distributed environments. These tools are based on the capabilities of Globus. They are working with various groups including the Scientific Data Management Group at LBNL, the Condor group at U. Wisconsin, and the SRB group at SDSC, and with other PPDG collaborators at ANL to understand the requirements.

They have completed the design of a catalog for recording replica information, and are currently developing script-based tools to load and maintain the catalog. (An early prototype was demonstrated at SC'99 in November 1999; this comprised a prototype file transfer API that supports multiple protocols and a GUI function that demonstrates replica management by combining capabilities for browsing a replica catalog with file transfer capabilities.)

Working with collaborators at NCSA, they have developed FTP clients and servers that are enhanced with Grid Security Infrastructure (public key) mechanisms as well as support for large TCP window sizes. This GSIFTP utility has proved to be a valuable Data Grid building block, enabling good (not yet outstanding) performance in wide area settings with standard interfaces. They have supported the installation of GSIFTP at several PPDG sites, as well as the addition of Globus Security Infrastructure (GSI) to the HPSS FTP server.

ANL has also have been developing a reservation architecture called GARA, which can serve as the basis for reservation services for networks and disk space. End-to-end reservation capabilities have been demonstrated between ANL and LBNL.

### Data Movement Tests:

An important part of the PPDG effort is to work with real high energy physics data using the currently available tools. This permits the collaboration to investigate the tools and their applicability to HENP. In addition, it gives the collaboration valuable expertise that will be needed when it comes time to deploy the grid for HENP.

The ANL ATLAS Group has moved 100 Gigabytes of TileCal testbeam raw data from CERN to HPSS at NERSC. It has prepared a relational database to describe raw data files and database file contents and locations (metadata) and has set up client ends of GLOBUS to be able to use GSI based tools, e.g. GSIFTP.

The ATLAS Group has worked with the Globus Group at ANL to design and test an LDAP<sup>27</sup>-based replication service. They have built and tested a process-level parallel application to look up and fetch raw data files any of the stores at ANL, NERSC or SDSC (in both tape robots and user level disk caches). The application can build, and then re-store Objectivity, database files. Both the applications and control script are now stable and run reliably.

---

<sup>27</sup> LDAP: Lightweight Directory Access Protocol, a set of protocols for accessing information directories.

They have made performance measurements that give typical throughput of 3 MB/sec (ESnet links) and 8 MB/s (test-bed link) between LBNL and ANL for 8-way process parallel database-build jobs. They have also completed some network-link and disk low level performance studies (using lperf<sup>28</sup> and IOzone<sup>29</sup>) to understand how to set the TCP window parameter and why the performance is limited to only 8 MB/sec. [A draft-working document describing the work and result in more detail can be found at <http://www.hep.anl.gov/may/PPDG>] Fermilab has also worked mainly on integration of existing middleware components and use of test beds. They have performed bulk file transfers with capture of the metadata in the SAM catalog, of Fermilab Fixed Target Data (FOCUS) in collaboration with Indiana University, from multiple 8 mm data tapes to the HPSS system at IU. The next steps include use of the HPSS GSI-enabled ftp server, transfer and distributed caching of data at other Focus sites in support of the experiment analyses.

Fermilab is collaborating with the University of Wisconsin on the integration of SAM and Condor. This will provide an interface from Condor to the SAM storage management system, such that application file names will be automatically mapped to a SAM file names, and the SAM catalog and data delivery system used to provide the file to the Condor I/O layer. In addition it will provide an interface from the SAM Optimizer and Project Master, to the Condor matchmaker and scheduling system, to provide for the dispatching and execution of a distributed set of Run II DO physics analysis applications.

Fermilab is also working with CERN and Caltech to provide a replicated subset of the production CMS simulation data for analysis by ORCA 4. In the next six months we will provide local caching of data for CMS data stored locally in Enstore, as well as providing access for ORCA applications to data cached at CERN, Caltech or Fermilab.

The main focus of the Caltech group has been on wide area access to large ODBMS stores of fully simulated CMS events produced by the CMS reconstruction software, ORCA. The current version of ORCA, has been run in production at CERN to create a large Objectivity federated database of several thousand database files containing about 500,000 events. The ORCA software has been installed at Caltech and Fermilab in preparation for an analysis pass over subsets of the event database. The intent is to test WAN access by running analysis tasks at the two sites. These tasks require access to a database located at the other site, thus causing a copy of a portion of the database to have to be made. Work has just begun on arranging large-scale ORCA simulations on the Condor facility at Wisconsin. The resulting database files will be shipped to Caltech and attached to the Caltech ORCA federation. Use of the Condor facility from Caltech will stress test the PPDG system components being developed in Wisconsin.

A major undertaking later this year will be the processing of ~1 million newly simulated events using ORCA. These events will be simulated at Caltech using the latest version of the CMS simulation program CMSIM. ORCA will also be run at Caltech and hopefully on Condor. Concurrently to the ORCA activities, explorations have begun on the use of the Globus GSIFTP tool, which allows large TCP windows to be configured to maximize WAN throughput. Tests from Caltech to Argonne are now complete, as are initial tests with INFN in Italy. In these tests, large ORCA database files have been used. The eventual scheme will be to incorporate the GSIFTP transfer mechanism in a system that requests a copy of a database across the WAN, attaches it to a local federation, and then allows an analysis job to execute on the newly available events. Prototype tools to facilitate this operation have been developed in collaboration with CERN's CMS group.

At SLAC the needs of the BaBar Collaboration are driving intense work on both local and wide-area data transfer. GSIFTP has been installed and is supplemented by SFCP (written at SLAC)

---

<sup>28</sup> lperf: A tool for measuring maximum TCP and UDP throughput, <http://dast.nlanr.net/Features/lperf/>

<sup>29</sup> iozone: A file-system benchmark tool, <http://www.iozone.org/>

for local transfers and by BBFTP<sup>30</sup> (written at CCIN2P3 Lyon) for transfers to Europe. SFCP achieves single file transfers at up to a hardware-limited 45 Megabytes/s using ssh authentication and multithreaded asynchronous disk I/O over multiple stream sockets. BBFTP is also multithreaded and fully exploits the network path to Lyon achieving data transfers at over 1 Mbyte/s averaged over 24 hours.

### Network Measurement:

To understand the results of the data movement experiments, the collaboration is also devoting effort to monitor the state of the underlying network infrastructure.

The collaboration has now installed IEPM<sup>31</sup>/PingER monitoring at 6 of the 9 PPDG sites, i.e. we are collecting data from U Wisconsin, ANL, BNL, LBNL, Fermilab & SLAC. We are in contact with the remaining 3 sites (Caltech, JLab and UCSD/SDSC). There is a web site set up for this activity, see <http://www-iepm.slac.stanford.edu/monitoring/ppdg/>

The next steps will be to extend the monitoring code and the archiving to allow customizing the monitoring for these sites to gather more detailed information (i.e. increase the ping frequency, and enable saving of the extra information). SLAC will investigate extracting new information that may be available from the more detailed measurements (e.g. measurements of jitter), provide better methods to analyze and report on the more detailed information.

SLAC is also working with CERN, IN2P3, Caltech & SDSC to understand how to improve and predict bulk data transfer performance. This includes evaluating the effect of increasing the number of simultaneous streams and increasing the buffer/window sizes, and comparing new and existing tools for fast bulk data transfer. At the same time we are looking for simple ways to predict the bulk data transfer performance from simple measurements available by PingER or other mechanisms.

---

<sup>30</sup> BBFTP: BaBar File Transfer Protocol

<sup>31</sup> IEPM: (Internet End-to-end Performance Monitoring) a SLAC led, DOE/MICS funded field work proposal, <http://www-iepm.slac.stanford.edu/>