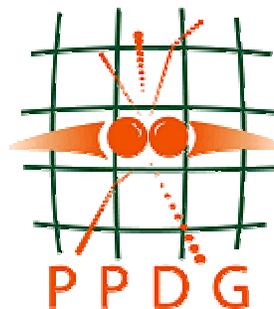


# Particle Physics Data Grid Collaboratory Pilot

## Quarterly Status Report of the Steering Committee, July – Sept 2003

30 Oct. 2003



1 Project Overview .....	2	4.3.2 Networks.....	10
1.1 Progress on the Year 3 Plan.....	2	4.3.3 Grid2003 .....	11
2 Focus Areas and Common Projects.....	3	4.4 D0 .....	11
2.1 Data Management.....	3	4.5 JLab .....	12
2.2 Job Management.....	4	4.6 STAR.....	13
2.3 Production Grids.....	4	4.6.1 Monitoring.....	13
2.4 Data Analysis Working Group .....	5	4.6.2 Hardware & Infrastructure.....	13
2.4.1 JAS-Tech-X SBIR project .....	5	4.6.3 Job submission & scheduling .....	13
2.4.2 Caltech Grid Analysis Environment: ..	5	4.6.4 Catalog and Data Management.....	14
2.4.3 ATLAS DIAL.....	5	4.6.5 Network performance .....	14
2.5 Monitoring.....	5	4.6.6 Registration.....	14
2.6 AAA .....	6	4.6.7 SDM ISIC work and STAR.....	15
3 Collaborations .....	6	4.6.8 Joint JLab/STAR project .....	15
3.1 DOE Science Grid .....	6	4.6.9 Joint STAR/PHENIX activities .....	15
3.2 Trillium and Grid2003.....	7	4.6.10 Other .....	16
3.3 Global Grid Forum and PNPA Research Group.....	7	4.7 Condor .....	16
3.4 Joint Technical Board (JTB) and LHC Computing Grid.....	7	4.8 Globus.....	16
4 Single Team Reports .....	8	4.8.1 Globus Toolkit 2.x updates and bug fixes .....	16
4.1 ATLAS .....	8	4.8.2 Globus Toolkit 3.0.....	16
4.1.1 RLS.....	8	4.8.3 GridFTP .....	17
4.1.2 Grid2003.....	8	4.8.4 Monitoring and MDS work.....	17
4.1.3 Magda.....	8	4.8.5 CAS .....	17
4.1.4 Papers .....	9	4.8.6 Grid Architecture.....	18
4.1.5 DIAL .....	9	4.8.7 Training, Presentations and Papers .	18
4.2 BaBar.....	9	4.9 SRM.....	18
4.3 CMS .....	10	4.10 SRB.....	19
4.3.1 Production.....	10	5 Additional Collaborators .....	20

5.1 IEPM, Network Performance Monitoring .....	20	6.1 List of participants .....	22
5.2 Globus ISI.....	20	6.2 Appendix 2: Additional Information for SRB: .....	24
5.3 ALICE .....	21		
5.4 CDF .....	21		
5.5 PHENIX .....	21		
6 Appendix .....	22		

## 1 Project Overview

The PPDG teams continued with the development, deployment and production activities as planned. The hiatus of the summer months resulted in few weekly phone meetings.

The planned series of meetings on each focus area were initiated.

The joint job management project between STAR and JLAB started with the definition of a User –Job Description Language. Feedback from ATLAS helped to shape the first draft. STAR and PHENIX have also discussed and are currently trying to address the job monitoring aspect of this project and the available existing tools.

STAR has adopted the VDT and will represent PPDG Applications on the VDT software working group, especially during the transition period from GT2 to GT3. D0 made progress towards migrating to using the VDT.

The PPDG coordinators, as well as US ATLAS, US CMS and Condor team members, gave a major focus to the Joint Trillium, US ATLAS, US CMS project Grid2003 (<http://www.ivdgl.org/grid2003>) to deploy a multi-organization functional grid demonstrator. This project is strengthening the ties to iVDGL being a joint project between PPDG and GriPhyN.

The PPDG project activities web page has been revised to reflect the modified projects structure <http://www.ppdg.net/pa/ppdg-pa/project-yr3.htm>

### 1.1 Progress on the Year 3 Plan

Milestones identified in the Year 3 plan for this reporting period are shown with their status and comments:

Date	Team(s)	Task Name	Section + Milestone Number	Date Finished
8/03	Exec	Start monthly job management coordination meetings.	2.7.1	8/03
8/03	ATLAS & CMS	Grid2003 grid with four sites.	2.8.1	8/03
8/03 (to 10/03)	JLAB	SRM Improvements	4.4.1	1)
8/18/03	ATLAS & CMS	Grid3 Integration Week.	2.8.2	8/22/03
9/03	Exec	Restart monthly monitoring coordination meetings.	2.7.2	3).
9/03	Exec	Start monthly data management coordination meetings.	2.7.3	9/03

Date	Team(s)	Task Name	Section + Milestone Number	Date Finished
9/03 (to 10/03)	BaBar	Phased integration with the existing JImport tools replacing the pieces that perform crude SRB-like functions by SRB. With simulated p2p	4.2.2	2)
10/03	D0	1) Finish installation of JIM at initial sites; 2) Begin load testing; 3) Exercise Monte Carlo operation; 4) Resolve issues with user input and output sandbox management; 5) Evaluate VO management options; 6) improved reliability, 7) work with Condor team to transition to VDT releases.	4.5.1	10/1/03

- [1] from Chip: dependent on SRM v2.1 Please redefine to) 10/03 and we'll see if we can make it . I expect the next spec version soon, and we would only need (I believe) to tweak our implementation to declare success.
- [2] from Adil: ..milestone has slipped by 1.5 months due to my not taking into account ccin2p3 holiday schedules We have tested all the pieces, we just need to run them together. I'm not sure how prod it will be by Oct (maybe flaky. now using root files to has some changes to schema (simpler than objy)
- [3] from Ruth: Focus currently on Grid2003 and MonaLisa deployment and integration. Will hold meeting post-sc2003

## 2 Focus Areas and Common Projects

### 2.1 Data Management

A joint activity for a **Replica Registration Service** was initiated, with co-leadership of Don Petravick, Arie Shoshani, and Chip Watson. A meeting was organized by LBNL (September 10, 2003) on the role of the Replica Registration Service (RRS) in the middleware infrastructure. The meeting was also attended by Anne Chevernak representing Globus replica management and RLS, and Jean Philippe Baud from the LHC Computing Grid Project. There was consensus that the RRS service is necessary as an independent service that can be used to standardize replica registration to various replica and file catalogs. A report about this meeting will be written during the next quarter. In the meantime some prototyping work is underway.

**GridFTP rewrite:** The original plan called for a bare bones server in August, and then a subsequent redesign in order to hit required deadlines for external user communities. These requirements were relaxed and the decision was made to take the more efficient route and do the design up front with a single implementation. This did, however, delay the initial release dates. There will be an alpha release of the server with GT3.2 alpha. It will not be feature complete (see features in the Globus report below), but it will have basic functionality in place with the exception of striping support.

The **SRB** team collaborated with the BaBar experiment under PPDG funding, the UK data grid, the NIH Biomedical Informatics Research Network, the NSF National Partnership for Advanced Computational Infrastructure, the National Archives and Records Administration, and the California Digital Library to develop requirements for peer-to-peer federation of data grids. Each of these projects provided funding support for the development efforts listed in the SRB team report. Peer-to-peer federation between data grids is intended to support:

- Collection control by the local data grid while sharing data with another data grid
- Performance improvement by keeping data grid metadata within a local catalog, while supporting global discovery
- Disaster recovery through replication of administrative metadata onto an independent data grid

**Replica Location Service:** RLS was released in the Globus Toolkit GT3.0 release in late June 2003. This

release has made the RLS part of an official Globus toolkit release for the first time. During this quarter, additional testing and some minor bug fixes have been incorporated into the toolkit. Also, some additional functionality was added related to bulk operations. In addition, a new version of the Replica Location Service that supports a hierarchical Replica Location Index was developed this quarter. It is currently undergoing testing and will be included in the Globus Toolkit 3.2 release scheduled for the next quarter.

A series of bi-weekly teleconferences on this data management area were started following the Replica Registration Service meeting in September. A paper (PPDG-33<sup>1</sup>) looking at the replica management tools and practices by the PPDG experiments has been prepared and is currently being reviewed within this data management group.

## 2.2 Job Management.

The Condor team improved the robustness of Condor-G and grid systems, additionally developing and a Condor-G extension designed to improve scalability of Globus Gatekeeper nodes by reducing the use of the globus-jobmanager and to support the gridshell which will improve reporting from jobs and robustness of jobs running on remote Globus resources.

STAR and JLAB are embarking on a joint project to define and implement a user Job Description Language and service based job submission portal.

The JIM team at Fermilab has collaborated closely with the Condor team to add Resource Selection Service (A.K.A Match Making Service – MMS) for general grid usage, plugins or external logic in the MSS, and a three tier job handling architecture.

## 2.3 Production Grids

**STAR, JLAB and BaBar** continued their production high throughput data file transfer production activities throughout this period, servicing users remote from the main data generation and storage facilities. The robustness and efficiency of the transfers were improved in response to network and other operating conditions.

The **D0-CDF SAMGrid** uses the SAM data handling system and has Jobs and Information Management (JIM) to enable the complete Grid functionality. A recent addition is a Resource Selection Service (A.K.A Match Making Service – MMS), as well as a Site configuration management framework, fabric sandbox management, and other important features. The JIM team has concentrated on making the JIM components work at several sites, and are working toward a portable D0 Monte Carlo application to use at Grid sites. The sites where JIM has been installed include the University of Wisconsin, Madison, the IN2P3 Computing Center in Lyon France, GridKa center in Karlsruhe Germany, and U. Texas Arlington. In the UK, the GridPP effort has provided considerable effort to install JIM at Imperial College, RAL, and Manchester, and it is being used for re-reconstructing data which was first reconstructed at Fermilab.

**CMS** moved into an Era of "Production Grid" from "Integration Grid Testbed". Using all Tier-1/Tier-2 Production Resources. Near end of May/Start of July newer version of DPE (distributed Processing Environment) for CMS grid computing was released, based on VDT 1.1.8. This version was a major enhancement over the previous version 1.0, which was released in late February 2003. Beside an easy to install pacmanized installation, this version included several new features. Enhanced versions of MCRunjob and MOP. Production operation moved into hands of "Operators" instead of "experts". CMS is running production grid almost smoothly now. The DPE deployment testing and operations are smoother than past.

Support for the **ATLAS** Data Challenge Production on the U.S. Grid Testbed. Large part of Simulation and pileup have been done at BNL Computing resource. [Performance, Operations, and Stability] Regression tests, capable of execution on any grid node, were developed and used to validate various common grid services at the node. This set of tests were used initially to debug emerging Grid3 sites and have now been turned over to personnel at the iGOC at Indiana University. A display of how these scripts

---

<sup>1</sup> <http://www.ppdg.net/docs/Papers/PPDG-33.pdf>

have been used to monitor the grid-availability status of a site can be seen at [http://www.ivdgl.org/grid2003/catalog/index.php?site\\_name=grid3](http://www.ivdgl.org/grid2003/catalog/index.php?site_name=grid3)

## 2.4 Data Analysis Working Group

The Data Analysis working group continued to meet biweekly and make good progress on understanding and defining Datasets. The working group reviewed the LHC Computing Grid project HEPCAL-II and HEPCAL-Prime documents, and increased the communication between the projects. <http://cern.ch/fca/HEPCAL-II.doc> and <http://cern.ch/fca/HEPCAL-prime.doc>

The working group will try to finish defining dataset catalog interface and lay out time scale, maybe six months, for defining other interfaces from the CS11 APIs diagram. Members of the GriPhyN Virtual Data system team are regular attendees at these working group meetings.

### 2.4.1 JAS-Tech-X SBIR project

The collaboration of the JAS team at SLAC and the Tech-X phase 2 SBIR project is ramping up. The group has been interviewing for a new hire to be located at SLAC and working closely with the JAS team. In a recent mini-workshop at SLAC work concentrated on the Dataset Catalog Service interface as well as some discussion of the Replica Location Service in GT3.

### 2.4.2 Caltech Grid Analysis Environment:

One of the major outcomes from the GAE workshop at Caltech in June 2003 was agreement on the architectural components for a Grid Analysis Environment. The Caltech team has been documenting and refining the GAE architecture document accordingly. A key component to our CAIGEE architecture, which is a concretized version of the generic GAE architecture, is the use of the Clarens portal for GAE client authentication and Grid resource access. Clarens is now in an advanced stage of development, and server instances are widely deployed across the worldwide HEP community. In this period a Java version of Clarens was created from scratch, in collaboration with NUST (Pakistan). This opens the door to the ease of creating non-Unix based Clarens servers in the future. Another development was a Clarens interface to the POOL metadata catalogue, an effort also made in collaboration with NUST. Incidentally, the Clarens server was also confirmed to interoperate correctly with VOMS-generated certificates, and the distribution and installation procedures for Clarens were completely revised and updated.

Another aspect of the GAE is the user's working environment, which will from time to time include a range of devices from handheld computers through elaborate desktop systems to large clusters at the Regional Centers. The Caltech team has been studying the use of multi-screen desktop displays and servers for use by individual or small groups of physicists engaged in Grid enabled analysis tasks. A prototype system has been deployed at Caltech, and also demonstrated at Telecom World 2003 in Geneva. In collaboration with NUST, the Caltech team has led the development of the handheld version of Java Analysis Studio and WIRED (from SLAC), both into which Grid authentication was included. These PDA-based tools, while still rudimentary, show the promise of portable computing for the LHC physicist. In a related development, the COJAC 3D geometry viewer was updated to use the CMS DDD DTD and XML files.

### 2.4.3 ATLAS DIAL

DIAL advanced significantly during the quarter. Two releases (0.40 and 0.50) were made. These include locally distributed processing (i.e. distributed within a site or farm). The processing may be distributed with fork, LSF, loran or Condor. Datasets may be constructed from logical files cataloged on local disk, NFS, AFS or magda. A line interface was added so that DIAL may be run either standalone or in the ROOT environment. Wensheng Deng contributed heavily to the magda, LSF and Condor pieces.

## 2.5 Monitoring

The MonALISA system, developed by US CMS core application software team, is now included in the VDT (Virtual Data Toolkit) and is being deployed as one of the main monitoring components of Grid2003.

It has been interfaced with Ganglia, PBS and LSF, so allowing jobs in those systems to be fully monitored. A new WSDL/SOAP access method has been provided that allows extraction of all monitored data by suitable clients (browsers etc.). Currently, MonALISA is installed at around 25 HEP sites – including for STAR and the IEPM-BM projects, and throughout the VRVS reflector system.

For STAR the MonaLisa service agents were deployed and tested initially on two nodes at the RCF (*stargrid02.rcf.bnl.gov* and *salonica.itd.bnl.gov*). The first step was to use the Ganglia module to interface the MonaLisa service to gmond to monitor BNL Linux Clusters. As a side note, we found the lookup service to be strikingly stable and with help from the developers, we set up a monitoring STAR group and a [Web repository](#). Appropriate firewall conduits to the hosts running the MonaLisa service have been opened and a LSF monitoring module (based on the PBS module) was developed to collect statistics from the batch system. Further communication with Iosif Legrand attempted to converge (information exchange) as a similar exercise was done in the MonaLisa core developer team. We tried to resolve problems related to firewall issues (this did lead to a new release of the tool where the port range for the service is configurable). We would like to acknowledge and thank the MonaLisa core developers for a good communication and timely support and delivery of solutions for the problems we helped identify.

The Globus team worked closely with the Grid2003 monitoring group to add in the deployment of a Grid monitoring infrastructure for Grid2003. This has involved identifying additional information providers and the proper usage of the GLUE schema within Grid2003. Additional Grid2003 support has been offered in setting up GIS servers. The MDS2 scalability analysis work continued, concentrating on adding netlogger calls to the MDS infrastructure to determine bottlenecks. This is a continuation of the work presented in June at HPDC.

## 2.6 AAA

ANL's contribution to the PPDG Site-AAA effort of an **Authorization callout** for the Globus Toolkit is now being used at FNAL to enable the integration of their implementation of Site Authorization System (SAZ) with their Globus deployment. Other sites, such as NERSC, are also examining this tool to integrate local authorization systems (an AIX-specific account locking feature) into their Globus deployments. Work in the GGF OGSA-Authz working group is well underway to standardize an equivalent callout for use in OGSA and GT3. This standardization effort was greatly helped by the experience of this Site-AAA work.

**BNL GUMS:** The particular problem to address is the need for strong pre-registration of users. The GUMS server is being integrated with the VOMS server. The work is fit in the large collaboration's needs for user registration. All of these developments will be compatible with whatever VO management tools are adopted for LCG. The more information about GUMS can be obtained from Web site: <http://www.atlasgrid.bnl.gov/testbed/gums>.

SLAC, FNAL and BNL now have people from their respective user services organizations participating as “agents” of the PPDG Registration Authority. This helps with the timeliness of processing certificate requests as well as providing experience that will be used to help fully integrate grid identity issues with other user registration procedures. Both JLab and FNAL are planning to set up their own Registration Authorities in conjunction with the DOEGrids PKI.

Participation in the Grid2003 project with iVDGL (see below) has helped sites identify issues around mapping grid identity (from X509 certificates and proxys) to local accounts. Exactly how some of the issues will be resolved, particularly as they relate to differing requirements between the HEP & NP labs and other multi-purpose DOE labs, is still to be determined.

## 3 Collaborations

### 3.1 DOE Science Grid

The DOE Science Grid is collaborating on the Grid2003 project in several areas: in the deployment of a Netlogger application demonstrator, pyGlobus as the grid-enabling interface of LIGO applications, the development and practice of security policies and procedures, and the DOEGrids PKI infrastructure. The

discussions at NERSC and ANL about how existing policies and practices relate to grid users have been very educational.

The user policies of the HEP and NP labs are somewhat different than the other DOE labs. The initial Grid2003 policy of collecting only name, institution and email address for users is not sufficient to satisfy access policies at NERSC and ANL, which also require nationality information about users. Since the type and scale of actual resource access being provided at NERSC and ANL are very similar to that at FNAL and BNL it may be that policies can be adjusted to take into account grid access as being different than complete login access to resources. It is clear that more discussion is required on these issues and Grid2003 is providing an effective motivation to move these discussions forward.

### 3.2 Trillium and Grid2003

The Grid2003 project is moving towards the planned demonstrations at SC2003. <http://www.ivdgl.org/grid2003/>. Two face to face working meetings were held and about 20 people are working on this project for some fraction of their time. There are currently 25 sites and over 100 cpus available on the Grid. Basic operability tests are run every 1/2 hour. There are currently 9 Application demonstrators planned <http://griddev.uchicago.edu/download/grid3/doc.pkg/WIP/Grid3-apps.htm>. The team is working closely together and building an organization for the demonstrators as needed. Further information about the project is given throughout this report.

### 3.3 Global Grid Forum and PNPA Research Group

The Particle and Nuclear Physics Research Group was accepted by the GGF Steering Committee. We are now moving ahead with planning for a workshop in conjunction with GGF10 in Frankfurt. We hope to work with the area directors to involve members of other GGF working groups in this meeting.

GGF is very active in defining the OGSA based grid services which will be implemented over the next months and years. The effects of the transition on PPDG experiments from Globus Toolkit version 2 (GT2) to GT3 are very dependent upon the current and ongoing activities discussed at GGF. In particular, Grid Data Services has recently been identified as a necessary core service in OGSA and work is going on in the OGSA-DAIS working group to develop compliant services. Following these developments and communicating with the particle and nuclear physics experiments is one example of the role of the PNPA research group.

### 3.4 Joint Technical Board (JTB) and LHC Computing Grid

Visits from INFN DataTAG members working on Glue Schema and VOMS resulted in an agreement to the development of a new the **Glue Schema** to include the Grid3 schema and the LCG extensions. Grid3 has deployed the **VOMS** service into production. The LHC Computing Grid projects also plans to deploy VOMS. A request has been made to include VOMS in VDT. A face to face meeting of the Joint Technical Board at GGF resulted in an agreement on a project to work towards merging of the Globus-EDG and EDG-specific RLS versions.

**RLS/RLI:** (From talk by Ian Bird, LCG Grid Deployment Manager:) Two mutually incompatible versions of RLS exist – one being used in US, EDG version is part of LCG-1; This poses a problem for the experiments. operating between different grid infrastructure. The LCG POOL software currently only works with EDG version

RLS/RLI proposal: The aim is to have a common solution as soon as possible but this will not be ready until May/June 2004.

**RLS common solution:** US Condor Project will assist in port of POOL to use both RLS implementations. Globus RLS is being ported to use Oracle (done). A roadmap for convergence has been agreed by both CERN and Globus groups to include agreement on the APIs for RLS and RLI. This discussion should include agreement on the syntax of filenames in the catalog; Implementing the Globus RLI in the EDG RLS, making the EDG LRC talk the “Bob” protocol; Implementing the client APIs; defining and implementing the proxy replica manager service; and updating POOL and other replica manager clients

(e.g. EDG RB). The timescale for this is May 2004; and is joint effort between US and CERN .overseen by LCG and Trillium/Grid3 via the JTB

## 4 Single Team Reports

### 4.1 ATLAS

U.S. ATLAS is testing the computing model to be followed by the LHC experiments: raw data are reconstructed at the CERN Tier 0 site, and transferred to Regional Centers (Tier 1's) for analysis by multiple users. We employ the Grid2003 software suite, along with ATLAS applications, to generate and simulate raw data in the U.S. testbed. Data are cached at BNL and transferred to CERN. At CERN, using LCG-1, the data are reconstructed, cataloged, staged and transferred back to BNL and other sites in the U.S. to test distributed analysis in a grid environment. This exercise tests not only the Grid2003 suite of tools, but tests issues associated with interoperability with CERN. Close discussions have been made between CERN Tier 0 managers and US ATLAS principals on how to configure this exercise. This exercise has been endorsed by International ATLAS (Gilbert Poulard), and makes use of Pacman, Chimera, RLS and VDT to support the grid efforts.

#### 4.1.1 RLS

Installation of Globus RLS Servers at both BNL and the University of Chicago. These installations provide a tiered RLI/LRC framework to support Atlas production efforts. Constructed, tested, and validated a set of Pacman software distribution packages based on Chimera+RLS+VDT to support ATLAS persistent grid challenges. Tested and validated the virtual data services provided by Chimera in support of Atlas simulation and reconstruction efforts.

#### 4.1.2 Grid2003

Participated in the design and development of the Grid3/2003 project. Grid2003 is a coordinated project between iVDGL, GriPhyN, PPDG, and the physics experiments, principally being led by USCMS and USATLAS. The goal of the Grid2003 project is to develop, integrate, deploy and apply a functional grid across (at least) the LHC institutions, extending to non-LHC institutions and to international sites, working closely together with the existing efforts. It is expected that knowledge gained in this effort can be used to address LCG-1 interoperability issues that may arise in its initial deployment.

Deployed Grid2003 software at Brookhaven National Lab, and tested the cross-VO job submission, such as running CMS jobs at the shared ATLAS computing facility, and running ATLAS jobs at Fermi lab. Participation in the VDT testing group to test the new VDT releases. Globus Toolkit 3.\*. is installed on a BNL grid test node.

#### 4.1.3 Magda

File catalog and data management (Magda): Developed new user interface for one-step data replication with gridFTP: A user goes to <http://www.atlasgrid.bnl.gov/magda/dyAddTasks.pl>, specify his `_source_` and `_destination_` locations, and give his replication task a `_name_`. Then the user can go to the command line, runs `dyPrepare.pl`, and `dyReplicate.pl` to start transferring files. If for some reason the transfer crashes, Magda remembers where it crashes. The transfer can be restarted from there. Made a new command `magda_replicate` upon user's request. The command syntax is `magda_replicate <filename> <site:location> [path]` Basically a user provides a logical filename, Magda will find an instance on the grid, and replicate it to the user specified location. This command can be run from any machine since it uses the feature of the gridFTP third party transfer. Setting up a Magda server at `ific.uv.es`. Integration of Magda into Ganga.

Made a pacman package for `edg-gridftp-client` package. This `edg` package provides the functionality of checking and manipulating a file or directory on a remote gridFTP server, so that we do not have to depend on gatekeepers for data management. A new parameter (`no-data-channel-authentication`) needs to be added to the `edg-gridftp-*` commands, in order to make them work to the NERSC HPSS gridFTP server `garchive.nersc.gov`. Supported testbed production and data replication. Tested and testing the NERSC

HPSS gridFTP server. Helped a Atlas user to replicate data files from CERN castor to ifae HPSS with the Magda tool. Did the initial installation of Ganga at BNL.

#### 4.1.4 Papers

Collaboration with a professor from Stony Brook University on Grid scheduling research. Got two paper published at a peer-reviewed conference.

YU, D., AND ROBERTAZZI, T. "Divisible Load Scheduling for Grid Computing". In IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS 2003) (Marina del Rey, CA, Nov. 2003).

WONG, H., YU, D., VEERAVALLI, B., AND ROBERTAZZI, T. "Data Intensive Grid Scheduling: Multiple Sources with Capacity Constraints". In IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS 2003) (Marina del Rey, CA, Nov. 2003).

Work on proposal, "QoS Based Divisible Load Scheduling for Grid Computing": a research program is proposed into distributed scheduling in grid systems with quality of service constraints. A novel feature of this research program is to use analysis techniques based on divisible (i.e. partitionable) computation and communication loads to create and evaluate new QoS scheduling strategies. However aspects of continuing research on indivisible loads will also be incorporated into the proposed research program as described in this proposal.

#### 4.1.5 DIAL

Development of the dataset model for distributed data management and the design and implementation of DIAL for distributed analysis. The web sites for these projects may be found at <http://www.usatlas.bnl.gov/~dladams/dataset> and <http://www.usatlas.bnl.gov/~dladams/dial>

Considerable progress in understanding was made with the dataset model and this understanding is summarized in the note "Datasets for the Grid" which is available at the dataset web site. The properties of datasets have been identified and datasets are categorized by the extend of their location information. The dataset code was updated to reflect some but not all of this new information.

DIAL advanced significantly during the quarter. Two releases (0.40 and 0.50 were made). These include locally distributed processing (i.e. distributed within a site or farm). The processing may be distributed with fork, LSF, loran or Condor. Datasets may be constructed from logical files cataloged on local disk, NFS, AFS or magda. A line interface was added so that DIAL may be run either standalone or in the ROOT environment. Wensheng Deng contributed heavily to the magda, LSF and Condor pieces.

Begun discussions with Gabriele Carcassi of STAR on the definition of a high-level job description language based on some concepts introduced in DIAL. In September, worked out an integration strategy that will enable GANGA users to access the upcoming dataset catalogs and the DIAL schedulers. Participated in many dataset discussions in the CS-11 meetings. provided feedback to the LHC HEPCAL and ARDA groups. Made the DIAL MagdaFileCatalog class which is inherited from the FileCatalog interface. Using MagdaFileCatalog, DIAL users can query the Magda database from a ROOT session, and get a physical file for analysis.

## 4.2 BaBar

Set up the SRB to distribute the BaBar condition snapshots to IN2P3 and other BaBar Tier-A sites. Have been working on some ideas to lock or hide the snapshot files for users while these files are updated. Have been testing the bulk loading and bulk registration of files. Started to install the latest SRB version, 3.0, that allows to federate MCATs.

Further explored the issue of Virtual Smart Cards and the possibility of adding an issuing feature to the card (as opposed to simple signing). I researched the interactions that cooperating VSC's would need in order to make VSC issued certificates portable across the centers compromising a VO. Also research was done into the practicality of allowing jobs to retrieve required data after they start running to increase fault

tolerance when data prestaging does not, for one reason or another, provide all of the data for a job to complete.

Documented the GSI SRB installation procedure for BaBar people. This has been used to install test servers at SLAC, ccin2p3 and GridKA. Have managed to replicate ROOT files to ccin2p3 with SRB (not using Sreplicate, but Sget and then SmodD). Have been working on an set of scripts to update the SRB MCAT with new user DN strings from the BaBar VO in Manchester. The script should be easy to extend to loading DN strings into another system. Have been carrying out scaling tests (loading the MCAT and copying files) to understand what type of server machine we need for production.

## 4.3 CMS

### 4.3.1 Production

At Fermilab, PPDG effort continues to be focused on CMS production developments and integration. The PPDG person at Fermilab working with CMS is Anzar Afaq.

Anzar's work was instrumental to move US CMS production from the "Integration Grid" testbed to the CMS "Production Grid", making full use of the US CMS Tier-1 and Tier-2 facility resources. This required upgrades to the CMS production environment software, DPE, which was moved to VDT 1.1.8, a major enhancement over the previous version. The new DPE release is using Pacman, which gives important improvements for installing it at each site. Anzar was responsible for this new release of the DPE and for integrating and commissioning it in the CMS production environment.

Concerning new developments, the legacy Impala script generation system was abandoned and a new "ImpalaLite" script generation was introduced. MOP is the CMS software that generates DAGs, that represent the workflow to be executed in the Grid environment. This software was changed to incorporate the new MCRunjob architecture, and to add new features making DAG generation more generalized, site independent and helping to easily maintain and run production.

Using the MCRunjob "Configurator" architecture, several configurators were developed which integrates the meta-data driven MCRunjob script generation with the MOP workflow generation of DAGs. This system provides submission to the Grid. Other new features like Condor\_G based match-making and integration with the dCache data access system were completed. Tests with the SRM storage management interfaces and test with the Configuration Monitoring tasks were completed.

In preparation for DC04, the Caltech group is developing the required MONARC-based simulation codes that can properly estimate the behavior of this large data challenge. One significant addition in the covered period is the new support for DAGs in the simulation.

### 4.3.2 Networks

The Caltech group continues to be at the vanguard of high speed WAN developments. With strong collaboration from other groups (SLAC, DataTAG, Internet2 ...) and vendors (notably Cisco) we have been able to break all existing Internet2 land speed records, and show rapid data transfers across the Atlantic between Switzerland and the USA corresponding to whole DVD movie transfer rates of just a few seconds. Our focus is of course on the scientific use of high speed networks, and not just on making memory-to-memory transfers. Accordingly we have been configuring, building, testing and tweaking a variety of hardware RAID servers based on dual processor Xeon, Itanium and Opteron

CPUs, all equipped with the latest 10Gbit network cards from Intel. We have demonstrated around 5.65 Gbps between a dual 3.6 (or 3.2) GHz Xeon server and an Itanium2 server. It does not matter if the servers are side by side or separated by 10,000 km (LA - Geneve) using 1 TCP stream. This was done including a path that used Abilene between Chicago and LA. Abilene traffic (several hundred Mbps) was not disrupted. We ran a VRVS conference at the same time as one of the tests, such that the Chicago-Indianapolis part of the path was shared. The videoconference continued to work. The above is memory to memory. We will be working on optimizing storage to storage transfers. The issue there is still running out of CPU power. In future 10 GbE network interfaces that offload the CPU are expected to solve this problem. We expect this to change the Computing Model, by making high speed data transfers more routine. Other than the

activities mentioned above, the Caltech group has been (and continues to be) very active in preparations for various demonstrations at events including Telecom World 2003 (Geneva, October), Supercomputing (Phoenix, November) and WSIS (Geneva, December).

### 4.3.3 Grid2003

Caltech is one of the core sites for Grid3 that will participate in forthcoming demonstration in SC2003 in Phoenix, AZ. The main objective of demonstration is to show interoperability between different VOs (experiments). The participating experiments are Atlas, CMS, Ligo, SDSS and BTeV. I have set up a 6 node Grid3 cluster using ROCKS 2.3.2 operating on Red Hat 7.3. The cluster consists of a 3U frontend server built with dual Intel 2.4 GHz CPUs, 2GB RAM and 1 Terabyte of raid storage on Supermicro X5DPE-G2 motherboard. The other 5 nodes consist of dual Intel 2.4 Xeons, 1 GB RAM on Supermicro motherboard. The grid middleware has been installed from Grid3 cache and using DOEGrids host certificate. The monitoring tools installed are MDS, Ganglia and Monalisa. The tests for running jobs from various experiments are going on.

Moving production cluster to Grid3: After the success of running jobs from different Grid3 experiment sites, USCMS has decided to move Production Grid Cluster to Grid3. The idea is to see how seamlessly a production cluster can be integrated into Grid3 environment. Our 33 node production cluster has been configured with Grid3 cache. A light weight USCMS DPE-client will also be installed to take advantage of some features like configuration monitor that will publish site specific resource information to grid users. One prime objective would be to see how dynamically experiment specific application can be installed, run jobs, fetch data product and delete these after the job completion. This cluster is still using DOESCIENCE Grid host certificate, but will also accept jobs submitted by users whose corticated are signed by different CAs.

Tests on DGT: Our Developmental Grid Testbed termed DGT has truly become a test ground for various grid middleware. Any Grid3 and DPE related efforts have been well tested before they are installed in actual Grid3 and production cluster. Recently, this cluster was upgraded with ROCKS 2.3.2. Analysis users are also working on this cluster to test their software on Red Hat 7.3 platform.

## 4.4 D0

D0 has met the following milestone as deifned in the project Year 3 plan:

1. Finished initial JIM deployment at U. Wisconsin Madison, 3 sites in UK including Imperial College, RAL, and Manchester, CCIN2P3 in Lyon, and GridKa in Karlsruhe. Currently the most active site is UWM .
2. load testing is being done at UWM.
3. A portable D0 MC application has been developed and is being run at UWM. This will be improved and operated at additional JIM sites.
4. A lot of progress has been made developing the input and output sandboxing. It is working quite well and many additional features have been provided to get applications to unprepared worker nodes, and sending job outputs back to a secure repository where the users can access it.
5. We looked more at the VOX project being developed for USCMS and SDSS at Fermilab. No decisions made yet.
6. Our goal is still to move to using VDT, but no progress has been made in this area yet.

Recently, the focus has been on achieving true Grid-style Monte Carlo processing at UW Madison with considerable work going into making the D0 MC application portable, and perfecting the input and output sandboxing capabilities. Wisconsin is a non-dedicated cluster where we cannot have preinstalled software on the worker nodes, nor can we run any dedicated processes. The sandboxing mechanisms overcome these obstacles, and using the sandboxing mechanism, we are now able to run sam analysis projects on the grid at Wisconsin. To achieve this, we have modified the mc\_runjob (workflow management tool) local interface to use sam\_batch\_adapter to submit and monitor jobs. This interface has then been integrated

with the job-managers. With the JIM system we also have the ability to retrieve job output and perform job cancellation via the web. The job manager and sandbox frameworks were connected together and it is important to ensure that success of the GRAM job is dependent on the condition of the underlying job output being successfully handled.

Making the D0 Monte Carlo application portable has been a major challenge over the last several weeks. The program comprises a chain of 3 to 5 stages, and each stage requires a binary executable which is 50 to 200 MB of dynamically linked components. There are about 40 packages which are needed, some just for dummy parameter files and many dependencies on libraries such as Root, Motif, X11, ACE are included for various reasons. The D0 experiment has distributed the code in the past with a tarball, which was hand crafted and bug prone for each distribution. When this is unpacked it sometimes requires around 2GB of space. In the D0 code, many advanced features of C++ are used, every libc library call, and even system calls are needed. This makes the code difficult to port, and it was noted that the code behaved differently on two LINUX systems with RedHat 7.2 installed. The total release tree can take up to 20 hours to build, so it cannot be done dynamically at each site. This has been an enormous challenge to make this an automated process and there is ongoing work to achieve this for the MC , as well as other D0 experiment codes.

## 4.5 JLab

During this quarter we participated in a final round of comments and revisions to the SRM v2.1 specification, particularly in the area of recursive directories and large directories. We have done additional internal planning to prepare to migrate to this specification once it is finalized. As the SRM specification does not contain the WSDL for the web services, this will be one remaining task and specification to complete in the coming quarter, concurrent with the migration of the implementation, so that interoperability is again achieved.

Jefferson Lab is now beginning to focus on the next area of a web services infrastructure, which is batch services. As presently envisaged, this will be a layered system, with a higher level component providing a user-level description of a job (domain specific description of work), and a lower level component providing a batch system view of a job (PBS, LSF, Condor, etc.) This work is in the context of a collaboration with STAR, and like the SRM work will support both the experimental physics program and the Lattice QCD program at Jefferson Lab.

For the high level piece, we are collaborating with STAR on a user job description language (uJDL or RDL for Request Description Language), aiming at a solution which could be common across multiple experiments. It is envisioned that both STAR and the JLab Grid Job Scheduler will make use of this common uJDL. We have already begun work on a prototype of a Grid Job Scheduler, and as part of this project, has investigated Globus Toolkit version 3 and Java CoG Kit version 1.1. The project makes use of web services and plans to incorporate the GSI security model using the Java CoG.

Four phases of development are currently planned for the Grid Job Scheduler. Early phases will focus on simple scheduling and deployment of jobs, whereas the final phase will include file lookup, file movement (input and output files), and scheduling based on file locations and the estimated time for execution of jobs at different sites. This later phase will integrate use of the SRM and ReplicaCatalog. After some initial discussion, STAR / BNL is heading up the uJDL definition, and we are providing input and feedback. The current state of the uJDL is that an initial proposal is nearly ready for comments (see STAR section below for additional details).

For the interface between the two levels, we have had some initial discussions on how a scheduling and dispatching architecture that could be composed from simpler components, including the SRM and other data grid web services. As part of this interface definition and decomposition, we've started investigating a specification of the lower level System Job Description Language (sJDL).

In other areas of data grid developments, Jefferson Lab participated in the Replica Registration Service (RRS) meeting at LBNL (described below). Planning for a migration of the JLab Replica Catalog to GSI is also underway.

## 4.6 STAR

Lots of work has been done in STAR since the last quarterly report.

### 4.6.1 Monitoring

This work is supported by Efstratios Efstathiadis from the ITD.

In July, we did continue with more scalability tests on the Ganglia/MDS/GRIIS monitoring approach. The Ganglia information was made available into GIIS at PDSF (thanks to Steve Chan and Shave Canon). GIIS was also setup on stargrid02 at BNL. The first result showed that the recovery of the information was slow and did not fit our needs (although we cannot quantify it at this stage). One possible direction (and part of our year3 plan) of this activity would be to go directly toward a testing with a GT3 context without further GT2 testing.

Stratos has provided detailed documentation on the [MonaLisa deployment](#). This documentation greatly benefited a rapid deployment of the service at PDSF and Stratos provided support and hand-on experience for the deployment there also used for the Grid2003 demo. We would like to thank Shane Canon and Iwona Sakrejda for their effort on setting this and working with us and the MonaLisa team to find a solution which accommodates both STAR and Grid2003 need: in fact, the information is currently registered in both groups. Also, because of the two site deployment, we are now able to monitor the interconnection between sites as well. Finally, the information service is being deployed at a third site (in Europe).

Finally, work has been done in setting up and testing MonaLisa as a Webservice to allow programmatic access to monitored values (interface with the Scheduler).

In our effort to disseminate information and knowledge, Stratos has presented his work on [Grid Monitoring](#) at the BNL Technology Meeting.

### 4.6.2 Hardware & Infrastructure

Several upgrades and deployment were done in order to consolidate our grid infrastructure and backbone. Two new gatekeepers were deployed (stargrid03 and stargrid03) to support respectively a vanilla VDT 1.1.10 and Globus toolkit 3.0 (the latest has not yet been deployed). While working with Globus 2.2.4, Dantong Yu noticed several problems with the LSF job manager ([Bug #950](#)) which led to a patch release in Globus 2.4.0 ([Bug #931](#)). We greatly benefited from this experience and stargrid01 was upgraded to Globus 2.4 for support of the Grid job submission (see next section for details) and in an attempt to get the benefits from a bug fix.

### 4.6.3 Job submission & scheduling

Since the RCF team has deployed Condor 6.4.7 on all farm nodes, we also modified the STAR Scheduler to take advantage of the existing Condor Pool. Although not apparently a Grid related activity, submission to two independent queuing systems allows us to test the resource brokering implemented from within the Scheduler in our path to better understand submission to multiple independent systems. Over 5k jobs were submitted in stress test mode to a Condor pool, 220 jobs did not make it (completely lost) but it was rapidly discovered that the problem was related to a local Condor wrapper intended to support and separate the four RHIC experiment by resource while having a single Condor Pool.

Gabriele also implemented submission through Condor-G in the STAR Scheduler. This exhibited a few issues right at the beginning: for example, the default LSF job manager does not propagate all the information one may need (such as resource usage declaration or submission to specific nodes). This was circumvented by extending the RSL and the LSF module with new attributes. This was extensively discussed in the [ppdg-condor@ppdg.net](mailto:ppdg-condor@ppdg.net) mailing list. The submission to Condor-G 6.5.3 using Globus 2.5 then resumed and tested on the local cluster for a better understanding and control of the problems and caveats. The next problem was found to be related to submission without the use of the Condor Grid-monitor. This led to an increase load of the gatekeeper as each submitted job would start its own job manager on the gatekeeper (lesson learned, don't even try without the Grid-Monitor). Without it, we observed

a consequent loss of the output from most of the jobs (~ 53 % of the jobs did not produce any output, 3% did not run at all and of these some were not registered or known to Condor-G).

After carefully considering the “[Golden Rules to Happiness when submitting lots of Jobs to Globus](#)”, starting the Grid-monitor as suggested and driven to use VDT 1.1.10 (GT 2.2.4 and Condor-G 6.5.3), tests were resumed and 0.2% of the jobs did not run and only one job was not submitted which sounded like an improvement. However, we still had ~ 50% of the job output were lost.

Considering the scope of this test and the usage profile for job submission in STAR (which is an order of magnitude higher than this test), we consider at this stage that the use of Condor-G is not ready for our prime time users and would need to bring closure to this issue. However, we do not foresee any major show stoppers as it seems to us that the issue is mostly related to the packaging of several versions and releases of components working with one another (GT, Condor, etc ...). We will soon resume testing based on a most recent VDT release and continue the reporting of problems and statistics. It is noteworthy to mention that this testing was supported by the Condor developer team: special thanks to Alan, Alain and Peter.

On another topic, a new version of the scheduler, supporting a flexible queue configuration including Condor and Condor-G support and an arbitrary/flexible organization of queues and pools, was deployed at both PDSF and the RCF. PDSF has also deployed the Web interface to usage statistics.

#### 4.6.4 Catalog and Data Management

The STAR FileCatalog was fully deployed at PDSF and included a support for the distributed disk approach which provides an independent way to test robustness of a distributed data approach. So far, the system has proven itself robust, the commands and functionalities the same (including the interface of the FileCatalog with the Scheduler). A new version of the API was deployed at BNL: the main change is that the API now takes into account the connection to the site of interest via a XML schema design describing the possible connections. Its implementation also supports connecting to the federation of databases (untested for now) which will complete the first wave of deployment and implementation of a truly distributed Catalog. This work is tight to the Replica Registration Service work from the SDM group. As we now have the basic infrastructure for our FileCatalog, the next steps are to test in real life scenario the transfer of data with automatic Catalog registration and the use of the information for job submission via the Scheduler and sites Condor-G.

The new HRM version release was delayed by a few weeks due to some problems found with the interaction of gridftp and HRM. This release is happening as this report is being written so it will be discussed in the next quarterly report.

#### 4.6.5 Network performance

While pursuing our production level data transfer between BNL and NERSC, the network performance was found to be less than desirable. We started an effort to understand and tune the performance to meet basic expectations (). This investigative work started with Doug Olson and Brian Tierney and continued with help by many people from the LBL, RCF and ITD team (Eric Hjort, Alex Sm, Shane Canon, Brian Tierney, Jin Guojun, Brent Draney, Doug Olson, Dantong Yu, Jerome Lauret, John Bigrow, Terry Healy and Scott Bradley). A network sniffer was deployed on pdsfgrid2 and it was discovered that window scaling does not seem to occur between BNL and PDSF as it does to other sites. Some summary results were made by Alex Sim and summarized [here](#). The problem was identified on the PDSF side. Hopefully fully understood, we hope to close this issue in the next quarter and push toward spawning activities on setting a [self configuring network](#).

#### 4.6.6 Registration

Robert Petkus and Richard Casela from ITD have worked on setting up the registration agent for handling the PPDG DOE grid certificate. This activity was helped and guided by Doug Olson.

#### 4.6.7 SDM ISIC work and STAR

We would like to mention that this work is not funded by PPDG but correspond to a cross-project Grid activity we mentioned in our Year3 planning document, section 1 and presented at the last PPDG collaboration meeting in the Catalog and Data sets section ([Event Collection Manager](#)). Named the “*Grid Collector*”, this work is supported by Kensheng (John) Wu from the SDM group and developed within the STAR framework. This work attempts to solve the problem of managing large datasets and avoid the need that other approaches have to go through a data preparation and management phase, by using a bitmap index that can find desired events based on selection criteria. It is intrinsically a set of software that works together to address those issues. It automates the file management tasks and provides “direct” access to a selected sub-set of events for analysis.

To speed up the development of this project, John visited us at BNL and worked closely with the STAR core developer team. Fully integrated with the STAR FileCatalog, the local HRM and DRM and the STAR analysis software, we successfully accomplished several milestones including:

- A real time analysis examples of events selected for an anti-3He and search for strangelets
- Implementation of event selections based on the second pass production files (Micro-DST)

This work period was very fruitful and we hope to be able, in the coming month, to consolidate it including finalizing the micro-DST implementation and stress testing the system.

#### 4.6.8 Joint JLab/STAR project

The scope and goal of this activity is described in section 2.9 of the PPDG work plan.

Work and discussion has begun primarily focused on delivering the first milestone namely, providing a first draft of the U-JDL (user JDL) to the PPDG collaboration for a first wave of comment and feedback gathering. We would like to thank David Adams for his comments and hope for enhanced interest from PPDG participants.

#### 4.6.9 Joint STAR/PHENIX activities

First, the STAR collaboration and team would like to particularly thank Barbara Jacak for making this joint project and activity possible.

The first meeting was held in late September. The goal of the first meeting was to identify commonalities in the respective activities and try to shape a strategy for the future, including basic logistics and scope for the collaborative activities. It was agreed to create a common mailing list to discuss cross project activities (done) and we identify several initial activities as of common interest amongst which:

- STAR Scheduler support for PHENIX needs and concept integration as needed (broadening)
- Job monitoring activity carved as an activity which is being supported by PHENIX (Andrey Shevel)

Subsequently, Gabriele (STAR) and Alex Withers (PHENIX) worked together to make the STAR Scheduler handle PHENIX job and file syntax: in PHENIX, the file syntax requires sequence of logical file names resolved at application level to physical files while in STAR, the scheduler interfaces with the FileCatalog and resolve logical collections into lists of physical files depending on where the job is to be run.

At the second meeting, we discussed the recent developments of the Scheduler, had detailed discussions on our respective approach and convention for passing file information to an experiment specific application and defining datasets. The PHENIX collaboration showed interest in our Monitoring work and strategy. The STAR data management system and our FileCatalog concepts (definition of logical collection, logical name etc ...) also seemed of interest and a topic for a next meeting.

#### 4.6.10 Other

We continued the support of the BNL [Technology Meeting](#). We mentioned earlier a presentation from Efstratios (Stratos) Eftathiadis. Dave Stampf also gave us series of lectures and tutorials on WebServices. We invited Andrey Shevel for a presentation and demonstration of his work on Job Monitoring from within Phenix. We also started to use the AccessGrid system and hope by then to broaden our audience and speakers.

Due to the Grid2003 needs, it was discussed within a the PPDG executive committee and the party involved (Rob Gardner, Alain Roy, ...) that experiment's Milestones related to VDT and GT3 may slip in time by a few weeks. Although not clear at this stage by how many weeks and the STAR Year3 plan being affected by this new time table and imperative, we agreed to leave to Grid2003 a higher priority and will therefore soon revisit our milestones accordingly. We would feel concerned however if few weeks would become few months.

#### 4.7 Condor

The team worked on improving and extending Condor-G interface to Globus and improved the robustness of Condor-G. Work was done on developed and maintenance of the `grid_monitor`, a Condor-G extension designed to improve scalability of Globus Gatekeeper nodes by reducing the use of the `globus-jobmanager`. Work was done on Condor-G changes to support the gridshell which will improve reporting from jobs and robustness of jobs running on remote Globus resources. The Condor team provided support to various projects using Condor-G, while continuing to integrate, and debug DZero, USCMS, and Grid3 software stacks, particularly with respect to Condor-G/Globus/Batch-System interactions. The team also provided VDT support across ATLAS, CMS and STAR. The Condor team worked with Fermilab collaborators to install DZero SAMGrid development and production sites at UW-Madison.

#### 4.8 Globus

Continuing interactions in terms of coordination and support of the PPDG applications included weekly phone meetings and email lists for Atlas and CMS, following the grid emails lists of D0, and providing support for the Argonne-Chicago ATLAS team in their efforts to perform "data challenge on demand" event generation using VDT, RSL, and Chimera. Support through the discuss lists and Bugzilla are available to all experiments.

##### 4.8.1 Globus Toolkit 2.x updates and bug fixes

In the past quarter, Globus release a new version of the GT2.x code line, namely release 2.4, which is packaged as part of the 3.x release. Several improvements were made after the release, and users can currently download the 2.4.3 release from <http://www.globus.org/gt2.4/download.html>. This stable release features a version of GSI capable of understanding a new proxy format that is used in the Globus Toolkit 3.0 release. Additional advisories and bug fixes can be found at <http://www-unix.globus.org/toolkit/advisories.html?version=2.4>

We closed 9 bugs listed in Bugzilla (53, 347, 398, 542, 713, 872, 951, 1076, 260) and have only 3 open PPDG-related bugs still open in our system (950, 1283, 1284), two added during the quarterly report writing process.

##### 4.8.2 Globus Toolkit 3.0

The Globus Toolkit 3.0 final release was made available for download. This stable release contains an open source implementation of OGSI, several OGSI - compliant services corresponding to familiar GT2 services, and our latest GT 2.4 release. The release is available from <http://www-unix.globus.org/toolkit/download.html> with additional advisories posted at <http://www-unix.globus.org/toolkit/advisories.html>.

### 4.8.3 GridFTP

**GridFTP rewrite:** The original plan called for a bare bones server in August, and then a subsequent redesign in order to hit required deadlines for external user communities. These requirements were relaxed and the decision was made to take the more efficient route and do the design up front with a single implementation. This did, however, delay the initial release dates. There will be an alpha release of the server with GT3.2 alpha. It will not be feature complete (see features in the Globus report below), but it will have basic functionality in place with the exception of striping support.

XIO is complete and will be released with GT3.2. GSI, TCP, file, and UDP drivers are complete. Work on Mode E and GridFTP drivers is under way, but they will not make the official GT3.2 release. They can (and will) be released as update packages when they are ready.

Significant improvements were made to the wuftp based server. Added are structured directory listings (MLST, MLSD), checksum support (CKSM), and a switch to have RFC 1738 URL support (paths are relative to where you log in rather than root relative). File globbing support has been added to globus-url-copy so that you can now specify \*.dat or a directory and globus-url-copy will move the entire directory and will utilize channel caching for efficiency. Support for the chmod command has also been added. These are the final feature enhancements added to the wuftp. Future wuftp efforts will be limited to bug fixes. There is not yet an end-of-life date for the wuftp based server. It should be noted that large file support was not working in versions 2.4.0 thru 2.4.2. It was fixed in 2.4.3.

### 4.8.4 Monitoring and MDS work

The Globus team worked closely with the Grid2003 monitoring group to add in the deployment of a Grid monitoring infrastructure for Grid2003. This has involved identifying additional information providers and the proper usage of the GLUE schema within Grid2003. Additional Grid2003 support has been offered in setting up GIIS servers. The MDS2 scalability analysis work continued, concentrating on adding netlogger calls to the MDS infrastructure to determine bottlenecks. This is a continuation of the work presented in June at HPDC.

While current monitoring plans involve MDS2, it has been clearly identified PPDG will require the capabilities of the OGSi-based MDS3, part of the Globus Toolkit version 3 (GT3) release, for its full deployment. The new version of MDS3 was released as part of the GT3.0 release in June. MDS3 represents a re-architecting of the MDS: information is now represented as XML, much of the functionality is subsumed by the OGSi core framework and is compliant with the GGF OGSi specification, some information sources are merged with domain-specific resource-layer services, and some of MDS is manifested as higher-level services, such as a collective-layer Index Service (comparable to an MDS2 GIIS). While MDS3 does not, as yet, present new radical new functionality to the end user, it lays the foundation for new features, such as registration, which will allow entities to request notification of changes in information instead of having to poll for such changes. We expect notification and archiving, as well as other services, to be a part of the upcoming 3.2 release.

### 4.8.5 CAS

ANL's contribution to the PPDG Site-AAA effort of an **Authorization callout** for the Globus Toolkit is now being used at FNAL to enable the integration of their implementation of Site Authorization System (SAZ) with their Globus deployment. Other sites, such as NERSC, are also examining this tool to integrate local authorization systems (an AIX-specific account locking feature) into their Globus deployments. Work in the GGF OGSA-Authz working group is well underway to standardize an equivalent callout for use in OGSA and GT3. This standardization effort was greatly helped by the experience of this Site-AAA work.

A production release of CAS, implemented as an OGSi service, is due out with GT3.2 Additional information on CAS can be found at <http://www.globus.org/Security/CAS/>

#### 4.8.6 Grid Architecture

Work on grid architecture in this period has focused on the adaptation of the virtual data language (VDL) and its implementation in the Chimera system to some specific requirements of the PPDG community as well as general simplification of the VDL dataset model.

Development has begun on deferred-binding of workflow nodes to sites, to enable more effective resource allocation and to facilitate research on data placement. It is expected that this approach will be tested in a Grid2003 environment on ATLAS production (along with challenge problems from SDSS and from DOE computational biology projects).

Initial design work has been done to fit Chimera into the US-ATLAS production system and then into the production system being managed within international ATLAS from CERN. Work on resource-utilization policy management for Grid2003 has progressed, with the refinement of earlier models and prototypes to monitoring and policy control mechanisms that will be suitable for Grid3 deployment. We plan to test this in the US-ATLAS GCE production management framework.

We are coordinating this work with workflow management progress taking place in GriPhyN, where a model of interacting workflow "refiners" or "editors" is being created which will provide a runtime environment for the execution of workflows produced by VDL and other workflow-generating clients. We are working to use this model to structure both late-binding and distributed planning of production workflows, and to address the needs of managing, tracking, updating, and monitoring long-running dataset production in HEP data challenges.

#### 4.8.7 Training, Presentations and Papers

The following Globus related presentations and tutorials are scheduled for the Supercomputing Conference this year. See the Supercomputing Conference web site <http://www.sc-conference.org/sc2003/> for more details.

1. The Grid: Software Standards for Cyberinfrastructure, November 16, 8:30AM -12:00PM, Carl Kesselman (USC Information Sciences Institute)
2. How to Build a Grid Service Using the Globus Toolkit ® 3 November 16, 8:30AM -5:00PM or November 17, 8:30AM 5:00PM (the latter is already sold out), Lisa C. Childers (Argonne National Laboratory), Charles A. Bacon (Argonne National Laboratory), Ravi K. Madduri (Argonne National Laboratory), Ben Z. Clifford (The Globus Project)
3. A Tutorial Introduction to High Performance Data Transport, November 16, 8:30AM 5:00PM, Bill Allcock (Argonne National Laboratory), Robert Grossman (University of Illinois at Chicago), Steven Wallace (Indiana University)
4. Grid Services for Data Management and Virtual Data, November 16, 1:30PM 5:00PM, Ann Chervenak (University of Southern California), Ewa Deelman (University of Southern California), Mike Wilde (Argonne National Laboratory)

### 4.9 SRM

Development Tasks:

A new capability was developed in order to support recursive directory-to-directory file replication. This includes several tasks. 1) The HRM was enhanced to support recursive 'ls' command. Since HPSS does not support recursive 'ls', the HRM gets the information one level at-a-time, parses the result, and repeats this till it reaches all the leaves of the directory. 2) the HRMs has to be enhanced to support a "mkdir" to the HPSS, and 3) the DataMover was enhanced to create a recursive directory at the target site. Here, again, this is performed by performing a "mkdir" command one level at-a-time until the entire directory is established.

RRS for STAR – the Replica Registration Service (RRS) is a new service developed to support file registration into the STAR File Catalog. This capability is very important to the STAR project and to HENP experiments in general. The RRS can perform registration in different modes, such as a file-at-a-time or in bulk, based on the success of the entire file replication request. At this time, we are only

targeting the STAR File Catalog (based on MySQL). In the future, we plan to target the RLS as well. Currently, the RRS is driven by the target HRM – it is notified each time a file is successfully replicated and archived to HPSS. This capability works with target DRMs as well, as soon as files are transferred to the target disk system. A working version is ready and waiting for STAR to provide a test database.

A new version of FMT was developed, tested and ready to be deployed. This version is more efficient in terms of memory usage and communication with the DRM. This reduces greatly the use of system resource to a negligible level relative to the DRM/HRM resource usage. This used be a barrier which for the use of FMT for large volume data replication.

#### Testing tasks

We scaled the testing of file replication from BNL (stargrid02.rcf.bnl.gov) to PDSF (pdsfgrid2.nersc.gov). Large volume tests involving about 500 files each were run 4 times. Scaling tests revealed some race condition bugs that were fixed.

Retrieving files from BNL-HPSS into disk cache. These tests were performed to test robustness of the HRM running BNL-HPSS.

Several hundreds of file replications were tested over a period of 1-2 weeks

We started to conduct measurements of transfer rates between sites, in order to track the bottleneck in the network. We have observed a slow transfer rate from BNL to PDSF, and testing the network bandwidth among sites would give us some clues where the slow connection is happening. We plan to produce plots that will reveal the source of the bottleneck.

## 4.10 SRB

The BaBar and IN2P3 project teams have been major advocates for the development of the Federated MCAT system and we hope and expect that this will serve them well. The BaBar SLAC team is running a single-MCAT SRB system, called a zone, and the IN2P3 team in France is running another single-MCAT SRB system or SRB zone. SRB version 3 will allow them to integrate these two SRB zones into a single Federated SRB system.

SRB 3.0 implements the ability to send commands between zones, and was released on October 1, 2003. The new version, called the zoneSRB, is described in the Release notes at <http://www.npaci.edu/dice/srb/CurrentSRB/ReleaseNotes3.0.html>, the accompanying guiding document at: <http://www.npaci.edu/dice/srb/FedMcat.html>, and a readme file at <http://www.npaci.edu/dice/srb/README.zones>.

The challenge in the design of the peer-to-peer federation has been the multiple user scenarios that will need support. One can use the Zone SRB in multiple ways. The examples described in Appendix 2 illustrate some of the possibilities, although one can also use the Zone SRB in other creative ways to achieve your goals of collaboration without losing autonomy.

A WSDL interface has been developed for the Storage Resource Broker, including support for both file manipulation and metadata manipulation. The WSDL interface is being upgraded to an OGSA compliant interface, to track the evolution of the OGSA-DAIS web service interfaces promoted by the Global Grid Forum.

In work unfunded by PPDG but related to the collaborating experiments, SDSC (Michael Wan) collaborated on the development of SRB servers for three archives used in the CMS experiment: Castor, Dcache, and the Atlas Data Store. SDSC also provided support to resolve problems seen in the deployment of the SRB for the CMS data challenge. In particular, the very high latencies seen in the CMS data grid caused a race condition during parallel transfer that required a code change.

## 5 Additional Collaborators

### 5.1 IEPM, Network Performance Monitoring

**Web/Grid Services:** To enable easier ways to search through the IEPM-BW data we built several tools to push the data into an Oracle back-end database. We have been working with the UCL developers (Yee-Ting Lee and Paul Meallor) to discuss and understand the OGSi schema. Following this Warren worked on porting the MAGGIE Web service to an OGSi type grid service using the perl module. We co-authored and had accepted a paper from the GGF/NMWG on Enabling Network Measurement Portability through a Hierarchy of Characteristics.

**Data Presentation and Visualization:** We have developed some new web based methods to visualize the traceroutes that we run every 10 minutes to all the remote IEMP-BW hosts. This provides a table of time of day versus remote host showing the route numbers and identifying changes. It also provides the ability to select hosts and times and show graphical topologies of the routes. See for example Traceroute Analysis for 10/07/2003. PingER

There is increasing interest from the HENP and other scientific communities to understand and do something about the Digital Divide, i.e. the difference in Internet performance to developing and developed countries. Since 10-20% of HEP collaborators on the major experiments come from countries in developing nations, this is very important to HENP. In order to provide a more balanced view seen from Europe, following a recommendation from the ICFA/SCIC we added ICTP/eJDS sites to CERN's PingER monitoring.

In preparation for a series of presentations, talks and papers this winter at the WSIS, RSIS, ICTP/eJDS etc. we energetically worked on cleaning up and replacing broken links, adding new countries and more nodes in existing countries. New countries include: Phillippines, Cuba, Tajikistan, Turkmenistan, Khirgizstan, Namibia. Hosts have also been clean out and added in Macedonia, Serbia/Montenegro, Belarus, Turkey, Armenia, Mexico, Azerbaijan, South Africa, Saudi Arabia. We are also working on getting a monitoring host in Iran. The Internet2 Land Speed Record is now in the Guinness Book of records. We also submitted and had published an article on High Speeds are Good for Guinness for the DoE Pulse magazine. We co-authored a paper with LANL and Caltech on Optimizing 10-Gigabit Ethernet in Networks of Workstations, Clusters, and Grids: A Case Study.

For the SC2003 Bandwidth Challenge we submitted a proposal on "Bandwidth Lust": Distributed Particle Physics Analysis using Ultra high speed TCP on the Grid with Caltech. We are now working with Cisco, Level(3), Stanford, QWest and CENIC to get a 10Gbits/s from SLAC to the SC2003 show in Phoenix.

**Advanced TCP Stack Evaluation:** With the emergence of many new advanced TCP (FAST, HS, Scalable, LP, H, Bic, Westwood+ ...) stacks that are trying to provide high-, fair-, responsive-performance on fast long-distance paths without needing to resort to using multiple parallel streams, it is important to evaluate and compare these stacks on high-speed production links to understand their domains of applicability etc. We (Les and Hadrien Bullot a summer intern) have put together tools to enable evaluating and comparing the performance of new TCP stacks. The tools included two new fast GE connected dual 3GHz cpu Dell 2650s, a modification to iperf to provide sinusoidally varying UDP streams, plus automated measurement scheduling scripts. We also worked with the stack developers to install, configure FAST-TCP, HS-TCP, HSTCP-LP, Bic-TCP, H-TCP and Westwood+. With these stacks we could saturate networks, so we worked carefully with administrators at several sites (e.g. CERN, UFL, Manchester, UIUC, Caltech and UMich) to get access to 2 fast hosts (the second to use for cross-traffic generation) at each site and to carefully schedule our tests. To facilitate the installation and testing of HSTCP-LP (a fusion of HS-TCP and TCP-LP) Aleksander Kuzmanovic, a student from Rice University, spent 3 fruitful weeks at SLAC assisting with the installation and providing fixes, enhancements as we tested and found out more about the stack on our fast, production links.

### 5.2 Globus ISI

Also this quarter, we have implemented a client tool that performs copy and registration operations, invoking gridFTP servers to copy data items and registering the resulting copies in the RLS. This tool is

intended to provide the same functionality that was formerly provided by the replica management API in earlier versions of the Globus toolkit.

Finally, we have been developing a simple Grid service wrapper around the existing Replica Location Service implementation. An initial implementation is complete, and we expect to release this grid service through the GTR (Grid Technology Repository) in the first half of October 2003.

An ongoing effort related to the RLS is to work with the European DataGrid project to overcome the problem of two implementations of RLS that are not interoperable. This situation arose last December after the EDG group diverged from our common design and implementation efforts. During the last quarter, we traveled to Italy for a meeting and devised a plan for achieving interoperability. While the general plan was agreed upon during that meeting, we are still waiting for resource commitments from the EDG/LCG projects before progress can be made. Discussions are actively ongoing.

### **5.3 ALICE**

### **5.4 CDF**

### **5.5 PHENIX**

The Phenix experiment has collaborated with STAR on the development of tools and strategies for job scheduling including:

- Adopted STAR scheduler for trial use by PHENIX.
- Collaborated with STAR to specify and implement a split between "master" and "subjobs" applicable for both collaborations. Began joint development with STAR of a generalized file specification syntax and file location query at run-time as well as at submission time.

Other activities included:

- Developed a GUI for PHENIX users to ease and standardize the specification of parameters for simulation jobs. The GUI produces as output JDL files appropriate for the STAR scheduler. We have performed successful limited-scale testing of job submission at RCF using this tool.
- Researched existing options for GRID job monitoring. We have selected and tested aspects of BOSS, and its web interface BODE, as candidates for testing and customization for PHENIX needs. We believe that splitting the job monitoring task from the task of job submission can simplify the development of monitoring components for PHENIX. A "registration" component will be developed to maintain independence of the two components. We have also started exploring commonality with STAR needs to determine whether a jointly useful tool can result from this work.
- Submitted and monitored 3000 test jobs to 3 PHENIX GRID sites - RCF, Stony Brook and the Univ. of New Mexico (UNM) - via a Globus gateway at each site, and using scripts and the BOSS/BODE tools. We use a Ganglia Module to interface to the Globus Resource Information Service and retrieve Stony Brook cluster machine status and load information.
- Researched options for PHENIX data management automation. Are currently studying the feasibility of combining existing single-site solutions with multi-site data transfers using Globus tools.

## 6 Appendix

### 6.1 List of participants

TEAM	Name	F	Current Role CS	Systems and Production Grids	Job Mgmt	Data Mgmt	AAA	Grid Analysis and Catalogs	Other: Web Services, Evaluations Interoperation, etc.
Globus/ANL	Ian Foster	Y	Globus Team Lead, GriPhyN PI, iVDGL, GriPhyN			x			
	Mike Wilde	N	GriPhyN coordinator, ATLAS- CS liaison	x		x			
	Jenny Schopf	Y	GriPhyN, iVDGL, liaison,	x		x			x
	Greg Nawrocki		PPDG Globus liaison	x		x			x
	William Allcock	Y	GridFTP	x		x			x
	Von Welch	Y	CAS				x		
	Stu Martin	Y		x	x				
ATLAS	John Huth	N	ATLAS Team lead, GriPhyN Collaborator	x					
	Torre Wenaus	N	LCG Applications liason		x	x		x	
	L. Price	N	Liaison to HICB, HICB Chair						
	D. Malon	N	Database/POOL Liason					x	
	A. Vaniachine	N						x	
	E. May	N	Testbed applications	x		x			
	Rich Baker	N	Testbed applications, VO tools	x			x		
	Kaushik De	N	Testbed applications	x					
	David Adams	Y	Distributed analysis					x	
	Wensheng Deng	Y	Metadata catalogs			x		x	
	R. Gardner	N	iVDGL coordinator, Atlas Grid Tools		x	x			x
	G. Gieraltowski	Y	Interoperability	x				x	x
	Dantong Yu	Y	Monitoring and VO	x			x		
BaBar	Richard Mount	N	PPDG PI, BaBar Team co- Lead						
	Tim Adye	N	BaBar Team Co-Lead						
	Robert Cowles	N					x		
	Andrew Hanushevsky	Y				x			
	Adil Hassan	Y				x			
	Les Cottrell	N	IEPM Liaison	x					
	Wilko Kroeger	Y				x			
CMS	Lothar Bauerdick	N	CMS Team Lead. GriPhyN collaborator						
	Harvey Newman	N	PPDG PI. GriPhyN collaborator, Co-PI iVDGL						
	Julian Bunn	N	CMS Tier 2 manager, GriPhyN & iVDGL collaborator	x				x	
	Conrad Steenberg	Y	CS-8:Analysis Tools, GriPhyN collaborator					x	x
	Iosif Legrand	N	CS-8:Monitoring Tools						x

PPDG Status Report, Jul.– Sep. 2003

	Vladimir Litvin	N	GriPhyN collaborator		x				
	James Branson	N	CMS Tier 2 manager	x					
	Ian Fisk	N	CMS Level 2 CAS manager, iVDGL liaison	x					
	James Letts	Y	Working on VDT testing scripts	x					
	Eric Aslakson	Y	job execution, grid monitoring	x	x				
	Saima Iqbal	N	web technology evaluation					x	
	Suresh Man Singh	N	grid deployment	x					
	Anzar Afaq	Y		x	x			x	
	Greg Graham	N		x	x			x	
Coordination	Ruth Pordes	Y	PPDG coordinator			x			
	Doug Olson	Y	PPDG coordinator			x	x	x	
	Miron Livny	Y	PPDG coordinator		x	x	x	x	
	Joseph Perl	Y	CS-11 co-coordinator					x	
	Craig Tull	Y	CS-6 Robust File Replication Common interface specification			x			
D0	Lee Leuking	N	DO PPDG liason	x	x				
	Igor Terekhov	Y	JIM Team Lead	x	x				
	Andrew Baranovski	Y		x					
	Gabriele Garzoglio	Y		x	x	x			
	Sankalp Jain	Y	Through contract with UTA CSE Department	x	x				
	Aditya Nishandar	Y	Through contract with UTA CSE Department	x	x				
HRM/LBNL	Arie Shoshani	y	SRM Team Lead. GriPhyN collaborator			x			
	Alex Sim	Y				x			
	JunminGu	Y				x			
	Viji Natarajan	Y				x			
SRB/UCSD	Reagan Moore	Y	SRB Team Lead. GriPhyN collaborator			x			x
	Wayne Schroeder	Y	CS-8: Web Services			x			x
JLAB	William Watson	Y	JLAB Team Lead			x			x
	Sandy Philpott	N	facilities	x			x		
	Andy Kowalski	N				x			
	Bryan Hess	Y	Web Services			x			x
	Ying Chen	Y	Web Services	x		x			x
	Walt Akers	N	Web Services	x		x			
STAR	Jerome Lauret	N	STAR Team Lead	x	x				x
	Gabrielle Carcassi	Y		x	x				
	Dave Stampf	N		x					
	Richard Casela	N		x					
	Efratios Efstathiadis	N		x					
	Eric Hjort	Y		x		x			
	Doug Olson	N		x		x			x
Condor/U.Wis consin	Miron Livny	Y	PPDG PI, PPDG Coordinator. GriPhyN collaborator	x	x	x			x

	Peter Couvares	Y			X		x		
	Rajesh Rajamani	N			x				x
	Alan DeSmet	Y			x		x		
	Alain Roy	N			x				
	Todd Tannenbaum	Y			x				
Globus/ISI	Carl Kesselman	N	Globus/ISI lead						
	Ann Chervenak	N				x			
PHENIX	David Morrison								
CDF									
ALICE									

## 6.2 Appendix 2: Additional Information for SRB:

The challenge in the design of the peer-to-peer federation has been the multiple user scenarios that will need support. One can use the Zone SRB in multiple ways. The following examples illustrate some of the possibilities, although one can also use the Zone SRB in other creative ways to achieve your goals of collaboration without losing autonomy.

- **First Model: Occasional Interchange** - This is the simplest model in which two or more zones operate autonomously with very little exchange of data or metadata. The two zones exchange only user-ids for those users who may go across from one zone to another. Most of the users stay in their own zone accessing resources and data that managed by their zone MCAT. Inter-zone users will occasionally cross zones, browsing collections, querying metadata and accessing files that they have permission to read. These users can store data in remote zones if needed but these objects are not accessible to users in their local zone unless they cross into other zones. This model provides the greatest of autonomy and control. The cross-zone user registration is done not for every user from a zone but for selected users only. The local SRB admins control who is given access to their system and can restrict these users from creating files in their resources. (NPACI Zones)
- **Second Model: Replicated Catalog** - In this model, even though there are multiple MCATs operating distinct zones, the overall system behaves as though it is a single zone with replicated MCATs. The MCATs synchronize metadata between them, so that each contains the same information as any of its sister MCATs. Metadata about the tokens being used, users, resources, collections, containers and data objects are all synchronized between all MCATs such that any file or resource is accessible from any Zone as though it is locally available without going across to another zone. An object created in a zone is registered as an object in all other sister zones and any associated metadata is also replicated. Hence, the view from every zone is the same. This model provides a completely replicated system which has a high degree of fault-tolerance for MCAT failures. The user will not miss any access to data even if their local MCAT becomes non-functional. The degree of synchronization though very high in principle, in practice, the MCATs might be out of sync on newly created data and metadata and will be constantly catching up with her sisters. The periodicity of synchronization is decided by the cooperating administrators and can be as long as days if the systems can tolerate them. An important point to note is that because of these delayed synchronizations, one might have occasional clashes. For example, a data object with the same name and in the same collection might be created in two zones almost at the same time. Because of delayed synchronization both will be allowed in their respective Zones. But when the synchronization is attempted, the system will see a clash when registering across zones. The resolution of this has to be done by mutual policies set by the cooperating administrators. In order to avoid such clashes, policies can be instituted with clear lines of partitioning about where one can create a new file in a collection. (NARA)

- **Third Model: Resource Interaction** - In this model resources are shared by more than one zone and hence they can be used for replicating data. This model is useful if the zones are electronically distant, but want to make it easier for users in the sister zone to access data that might be of mutual interest. In this model, a user in a zone creates a data replicated in these multi-zonal resources (either using synchronous replication or asynchronous replication as done in a single zone), then the metadata of these replicated objects get synchronized across the zones. The user list of the zones need not be completely synchronized. (BIRN)
- **Fourth Model: Replicated Data Zones** - In this model two or more zones work independently but maintain the same `data across zones, i.e., they replicate data and related metadata across zones. In this case, the zones are truly autonomous and do not allow users to cross zones. In fact, user lists and resources are not shared across zones. But data stored in one zone is copied into another zone along with related metadata, by a user who has accounts in the sister zones. This method is very useful when two zones are operating at considerable (electronic) distance, but want to share` data across zones. (BaBar Model)
- **Fifth Model: Master-Slave Zones** - This is a variation of the 'Replicated Data Zones' model in which new data is created at a Master site and the slave sites synchronize with the master site. The user list and resource list are distinct across zones. The data created at the master are copied over to the slave zone. The slave zone can create additional derived objects and metadata but this may not be shared back to the Master Zone. (PDB)
- **Sixth Mode: Snow-Flake Zones** - This is a variation of the 'Master-Slave Zones' model, In this case, one can see this as a ripple- model, where a Master Zone creates the data and which is copied to the slave zones, whose data in turn gets copied into other slave zones in the next hierarchy. Each level of the hierarchy can create new derived products of data and metadata and have their own client base and propagate only a subset of their holdings to their slave zones. (CMS)
- **Seventh Model: User and Data Replica Zones** - This is another variation of the 'Replicated Data Zones' where not just the data get replicated but also user lists are exchanged. This model allows user to go across zones and use data when they operate in that zone. This model can be used for wide-area enterprises where users travel across zones and would like to access data from their current locations. (Roving Enterprise User)
- **Eighth Model: Nomadic Zones - SRB in a Box** - In this model, a user might have a small zone on a laptop or other desktop systems that are not always connected to other zones. The user during his times of non-connectedness can create new data and metadata. The user on connecting to the parent Zone, will then synchronize and exchange new data and metadata across the user-zone and the parent zone. This model is useful for users who can have their own zones on laptops but also for zones that are created for ships and nomadic scientists in the field who might go on scientific forays and come back and synchronize with a parent zone. (SIOExplorer)
- **Ninth Model: Free-floating Zones - myZone** - This is a variation of the 'Nomadic Zone' model having multiple stand-alone zones but no parent zone. These zones can be considered peers and possibly having very few users and resources. These zones can be seen as isolated systems running by themselves (like a PC) without any interaction with other zones, but with a slight difference. These zones occasionally "talk" to each other and exchange data and collections. This is similar to what happens when we exchange files using zip drives or CDs or being occasional network neighbors. This system has good level of autonomy and isolation with controlled data sharing. (peer-to-peer, Napster)
- **Tenth Model: Archival Zone, Backup Zone** - In this model, there can be multiple zone with an additional zone called the archive. The main purpose of this is to be an archive of the holdings of the other zones that can designate which collections need to be archived. This provides for having a backup copy for a set of zones which by themselves might be fully running on spinning disks. (backup)

The development was done by Michael Wan (NARA funding), Arcot Rajasekar (NPACI and BIRN funding), and Wayne Schroeder (PPDG and DOE SciDAC funding). Wayne Schroeder developed extensions to the GUI administration tool, extensions to the installation script, developed the Zone synchronization script, created the Zone authority system, coordinated and developed much of the documentation, and did some of the integrated testing.

The features provided in zoneSRB are:

- New metadata for Zones
- New Zone table
- Add zone to user metadata
- Authentication across zones
- Resource and data access across zones