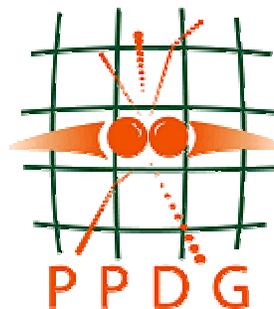


Particle Physics Data Grid Collaboratory Pilot

Quarterly Status Report of the Steering Committee, January - March 2003

25 Apr. 2003



1 Project Overview	2	2.2.1 Information Providers and Glue Schema	10
1.1 SC2002 Collaboration with IEPM-BW Project.....	3	2.2.2 IEPM-BW collaboration.....	11
1.2 Information from Questionnaires to Experiments.....	3	2.2.3 Mona Lisa.....	13
1.2.1 PPDG identifiable contributions to your Experiment goals to date.....	3	2.3 CS-4 Storage Management	13
1.2.2 Where should PPDG concentrate its efforts during the next year and a half?	4	2.3.1 LBNL-SRM Development.....	13
1.2.3 Have the computer scientists better understood your computing requirements?..	5	2.3.2 JLab-SRM.....	14
1.2.4 Issues with the support of the (PPDG) technologies you are deploying?	5	2.4 CS-5 Reliable File Transfer	16
1.3 Questionnaires on Common Service Areas	6	2.5 CS-6 Robust Replication	16
1.3.1 Is Progress as planned in the PPDG proposal?	7	2.5.1 BaBar Database Replication (BaBar- SRB)	16
1.3.2 What work is needed for the next year of PPDG?.....	8	2.5.2 Globus ISI, RLS work	16
1.3.3 At the end of PPDG do you think the technology will be mature and robust enough?	9	2.6 CS-7 Documentation	17
1.3.4 Issues with the support of the technologies being deployed?.....	9	2.7 CS-9 Security, Authentication, Authorization, Accounting	18
2 Common Service Areas	9	2.7.1 Certificate/Registration Authority ..	18
2.1 CS-1, CS-2 Job Description Languages, Management and Scheduling.....	9	2.7.2 Site-AAA.....	18
2.1.1 Job Description Languages.....	9	2.7.3 US CMS, US ATLAS, INFN, iVDGL Joint VO Management Project.....	19
2.1.2 Collaboration with EDG WP1	9	2.8 CS-10 Experiment Grids and Applications	19
2.1.3 STAR Job Scheduling	10	2.8.1 ATLAS	19
2.1.4 SAM Job and Information Management (JIM)	10	2.8.2 BaBar.....	21
2.2 CS-3 Information Services	10	2.8.3 CMS.....	21
		2.8.4 D0	21
		2.8.5 JLab experiments, and QCD.....	22
		2.8.6 STAR.....	22
		2.9 CS-11 Grid Interface with Interactive Analysis Tools	23

2.9.1 CMS Clarens web service layer client and server developments	24	3.1.5 Grid Architecture	26
2.9.2 ATLAS DIAL	25	3.1.6 CS-11 Interactive Data Analysis Tools	26
2.10 CS-12 Catalogs and Databases	25	3.2 Condor Project	27
2.10.1 STAR MySQL database	25	3.3 SDSC – SRB	28
3 Single Collaborator Reports	25	4 Appendix	29
3.1 ANL – Globus	25	4.1 List of participants	29
3.1.1 Coordination and Support	26	4.2 Meetings	31
3.1.2 Training, Presentations and Papers	26		
3.1.3 Globus Toolkit 2.x updates and bug fixes	26		
3.1.4 Globus Toolkit 3.0	26		

1 Project Overview

The project teams in PPDG continued with the technical work planned. The use of Grid technologies in the experiments continued to increase. Members of the PPDG steering committee continued their participation in the High Energy Physics Intergrid and Joint Technical Boards, and the Global Grid Forum in February/March. We are active in the joint projects for Middleware Testing and Glue Schema. PPDG was well represented at the SciDAC PI meeting in March, which meeting included many useful presentations and discussions. As a collaborator PPDG presented at the iVDGL External Advisory committee review, and members of PPDG were involved in the GriPhyN review.

The report from the December Troubleshooting workshop was completed as PPDG-26 and initial preparation for some pilot projects started.

In preparation for the PPDG DOE review in April a set of questions were prepared to explore the benefits and issues of the collaboration as seen by the Experiments and Computer Science Groups. The answers to many of these questionnaires were collected under PPDG-30. We summarize some of the key issues later in this quarterly report.

There were many PPDG projects and participants represented at Globus World in January (<http://www.ppdg.net/docs/news/ppdg-at-globusworld-jan03.htm>) and at the Computing in High Energy Physics conference in March (http://www.ppdg.net/docs/chep03_talks.html) (see lists under http://www.ppdg.net/docs/presentations_list.htm).

Much hardening of middleware continued also with the European DataGrid, DataTAG and Large Hadron Collider Grid Projects as these projects prepared to layer their upper level services on the Virtual Data Toolkit. This work benefits all experiments on PPDG as well as the other application domains using the core Globus, Condor and SRM software and protocols. Members of PPDG were active in the preparation of a new round of proposals for experiment end-to-end distributed analysis and underlying core middleware services from the NSF in the US and for the EGEE proposal in Europe.

PPDG has joined the LHC Computing Grid Project Grid Applications Group (GAG) where there have been relevant and worthwhile discussions to date on the SRM specification and use, and the CS-11 Analysis Tools area.

Information from the individual quarterly effort reports is being very useful in understanding the directions of the project. The reports for Jan-Mar 2003 are posted at <http://www.ppdg.net/pipermail/effort-reports/2003/>

1.1 SC2002 Collaboration with IEPM-BW Project

The breaking of the Internet2 Land speed record has resulted in much publicity, being reported by BBC, CNN, the Times of India, on Tech-TV and ABC Radio among others, and the press release has been made in English, Spanish, Portuguese, French and Dutch. A web site (see <http://www-iepm.slac.stanford.edu/lsr2/>) has been created to provide current information to interested people and reporters. Considerable time was absorbed talking to reporters, making presentations (see for example <http://www.slac.stanford.edu/grp/scs/net/talk/ricoh-lsr.html> and <http://www.slac.stanford.edu/grp/scs/net/talk/chep03-hiperf.html>) and writing publicity reports.

1.2 Information from Questionnaires to Experiments

Below we summarize some of the answers to the questionnaires that were obtained in fairly informal meetings of the Executive Team with each Experiment Team. There was an attempt to include the experiments software and computing management as well as the local PPDG Teams. The questionnaires and answers are available at <http://www.ppdg.net/mtgs/review-apr03/>

1.2.1 PPDG identifiable contributions to your Experiment goals to date.

ATLAS: Magda: is the ATLAS-standard file cataloging and replication tool of ATLAS, used in the ATLAS data challenges in the US, at CERN and internationally. Future VO toolkits: Developed and deployed for US ATLAS production. Interoperability: The PPDG interworking effort in support of SC2002 made quite an identifiable contribution in supporting joint US-Euro grid efforts for ATLAS. Distributed Data analysis: This has gotten on the “radar screen” of both international ATLAS and the LCG.

BaBar: no direct impact to physics analysis yet. PPDG is influencing grid development significantly, job submit, results retrieve, site-AAA proposals necessary for secure access.

CMS MOP – small glue piece. Put grids into the mainstream of CMS. Requirements were written by Koen Holtman which brought the requirements for Grid to the experiment. Demonstration of MOP at SC2001 showed that the grid tools could actually be used. Real collaborative project between the Experiment and CS teams to fill in for people who left. Brought technology to the experiment; brought real application to the attention of the Computer Science developments. Middleware support crucial in the transition from Globus 1.4 to 2.0. facilitated the construction of the US CMS test grid itself. End to end goals were constructive here. Now US CMS testbed brings out problems in the middleware that are not seen elsewhere, MOP and end-to-end focus led to testbed useful not only for the experiment but for the CS developers to learn about real deployment issues. **GDMP** – capabilities did not meet the requirements in the long term, but demonstrated the utility of the grid components. **VO management** – driven by PPDG project use of CAs and Caltech PPDG development of initial infrastructure for VO management. **VDT** – middleware packaging and support now adopted by the EU grid projects for the LHC. UCSD PPDG contribution to the testing. **Clarens/Caigee** – demonstrations of data transfers and analysis on the grid.

D0: PPDG manpower has been used to provide additional functionality to our existing data handling and monitoring systems. Enabling the use of GridFTP within our system has reduced our reliance on bbftp, which we think will prove beneficial in the future. We have also employed PPDG resources to provide a SAM/ROOT interface for our analysis framework. Other work which the CS teams are contributing to other parts of PPDG, like testing of condor and GRAM, will soon directly feed into what we are doing at D0.

JLAB Experiments: Provision of srm tools has been significant help. We have only begun, and have just a single user authenticated at the moment. In the immediate future, improvement of the srm tools would be the most useful development for us. To process monte-carlo for 30M events would take 3 months or 1 year because of other priorities. Able to process at FSU in under a week through replicating it. Joint analysis with RPI. Opened up facility to other users at RPI, JLAB, CMU by giving accounts. Now want to get a common account policy – globus run type tool. Grid is the last thing on the radar and PPDG has enabled this to be jump started.

STAR: the HRM file transfer has provided a production scale data transfer and replication framework which has allowed us to better balance the production load between our two main facilities, namely, RCF/BNL and PDSF/NERSC. In addition, the STAR computing model is currently evolving toward a distributed approach, on a local and global scale. Using solutions presented from within the PPDG collaboration allows us for a smooth transition toward a fully Grid-aware environment without major disruption of our user's habits and analysis methods. We would also like to mention, as an additional positive side effect of working in a collaboration such as PPDG, that the Grid ideals has a profound sociological impact on all experiments and the way we communicate: it brings the experiments together, allow us to discuss and debate our common problems by providing a neutral framework allowing such discussions to take place. In particular, we have successfully organized at BNL a weekly meeting discussing of emerging technologies and common solutions where the four RHIC experiments and Atlas representatives present their success stories and developments.

1.2.2 Where should PPDG concentrate its efforts during the next year and a half?

ATLAS: 1) Continuing and expanded support for the use of RLS and Chimera by both US and Euro grids supporting ATLAS work. 2) The development of a MAGDA-RLS interface. 3) Grid performance, operations, and stability should be addressed. 4) Additional development on grid interfaces with interactive analysis tools; including new or enhanced grid-portals and the ATLAS DIAL effort. 5) Support for the introduction of the new Globus Toolkit 3 (OGSA).

BaBar: interoperability at different sites, BaBar + LHC, necessary for standards, happy to see VDT in LCG-1 interoperability means different versions of s/ware & different s/ware code stability & quality – needs to work in production need some stability & not too fast cycle time on new releases & whole new installations. concern on support time for GT2, dropped end of year needs collaborative arrangement with developers and user groups need the discussion with Globus & Condor teams.

CMS: Continue work on production quality for actively used end-to-end multi-site experiment grid applications, hardening of the middleware and extending of the common toolset further up the application infrastructure in the experiment. There is a proposal that Caigee/Clarens be deployed for a DC04 for the experiment. This would become an end-to-end application PPDG deliverable and a plan and milestones for early deliverables, support and deployment on the CMS testbeds, using short to medium term middleware – aligned with the PPDG mandate. This is aligned with the PPDG changing emphasis from production to analysis for the last year of the project. Would like to see more cross-fertilization and common projects between ATLAS and CMS.

D0: For D0, we would very much like to see the projects underway for the Fermi SAM-Grid system be completed and made operational. We are relying on this system to empower the experiment to access the needed resources within the collaboration, and beyond, to meet our computing requirements and fulfill our physics mission. We plan to work with the Condor team, PPDG, and D0 management to establish a succinct work plan which will enable us to do this. With PPDG resources we have made progress toward additional tools needed for job description and job processing orchestration and we would like to see more progress in these areas. We have been building a computing infrastructure that includes several Regional Analysis Centers (RAC) and we are counting on the software provided through this project to make it all work.

JLAB Experiments: I'd first like to see the srm tools improved. At the moment it's still laborious to transfer a lot of files. There's no wildcard facility or ability to transfer a whole directory. These would be useful additions. The ability to copy files based on queries to a run database would be useful. A GUI / web tool to move files. Need to develop a replica catalog. Lack of interoperability between the existing infrastructure and between grid tools from many people. Would propose to develop an interface and then develop their own tools. VO organization support is an issue. Plan written but not yet started to implemented.

STAR: We believe and would like to emphasize that the integration of Database solutions should be a focus of PPDG in the coming year. This effort in grid framework is essential due to the emergence of replica catalog solutions as well as registration and monitoring information using Databases as back-end. The secure propagation of this information on the Grid cannot be left un-attended. We also believe the

PPDG collaboration should focus its effort on monitoring either from an internal effort or through a wider collaborative effort with other Grid collaboration. Finally, our job scheduling program has allowed us to make a contribution to the JDL effort and as we are now into a production level use of our tools, we foresee a greater participation in this activity, share our experience and finalize to/or expand our current knowledge toward a well defined requirement for a JDL (U-JDL, database interaction, etc ...) if not a final proposal

1.2.3 Have the computer scientists better understood your computing requirements?

ATLAS: The fact that international ATLAS decided to run grid production with many/most US tools is a good indication that we're getting a buy-in. I think people are realizing that this is a benefit in the long run, but in the short run, it means a deviation from the "nominal" model of HEP computing. This is an evolutionary process in my opinion. I think that the community is increasingly realizing the benefits of grid technology and deployments - particularly those who are the targets (remote users). The two CS contacts working closely with us are Miron Livny and Jenny Schopf. Collaborations with CS are happening at many levels. the grid-interopability and RLS/Chimera work would not have been successful without the direct involvement of both US and Euro computer science personnel.

BaBar: has an MOU between BaBar and Karlsruhe, which needs grid job submission. Have good interactions with SRB & Globus people.

CMS: Yes. See discussion of Mop above.

D0: We believe the collaboration with the computer scientists in the PPDG initiative has been very fruitful and has enabled physicists to present their needs, in concrete terms, to the IT professionals. We hope this has facilitated a better understanding of the potential for grid technologies. We also feel that the close contact with the computer science groups helps to resolve problems with their supplied software more quickly.

JLAB: Physicist not closely connected to PPDG to date: My general impression is that those developing the grid are (understandably) more excited at developing new protocols etc. than providing user friendly tools to use them. In fact, even system managers like myself, working in different fields of science, have very little idea what the grid might do for them and what they might demand of it. Application developers (applied computer scientists) and domain scientists understand the needs more than the computer scientists. There is some frustration with the CS people interested in the theory of grid computing, rather than the practical application.

STAR: 1) Computer Scientists as seen from a PDDG collaborator stand point: In many cases, we have bumped into a Grid far to ideal vision where it was considered that a global solution would be provided and therefore integrated in current software leaving no room for temporary solutions and plug-in for existing framework. An illustrative example of this is the need by all experiment to have one working File and Replica Catalog solution or the re-use of the existing one. Specifically, the HRM development has been made with the thought of a later universal replica catalog solution and therefore, do not have an easy way to interface the replica management software with any existing Catalog solution. This, in practice, means that more add-hoc solutions and wrapper need to be developed by our own team, diverting important resources to cope for discrepancies in needs. In general, we would require and wish for more flexible approaches and plug-ins in all provided tools/software which would make solutions immediately usable without further time investment on our part.

2) Computer Scientists as seen from an internal to STAR users and programmer perspective.: The potential for grid technology is not quite understood by physicists yet. It is not un-usual for our scientist to perceive the Grid activities as part of a futuristic vision of a far far away world. Consequently, progresses are slow and users need to be lead through small steps into the global effort. This is inherent to running experiments where the stability of the Physics output has priority over development but may not be as relevant to experiments under development phase. The comment however is pertinent to the above first remark: flexibility must be provided to allow running experiments to safely deploy Grid software.

1.2.4 Issues with the support of the (PPDG) technologies you are deploying?

ATLAS: Magda: Slow emergence of production-grade middleware for replica management. If it is adopted as the international standard, then there is a question about the long term support. The demarcation

of who is responsible for what development and support can often be blurred by the overlapping interests of the Trillium collaboration. Long term support and enhancement of robustness of tools that will be adopted internationally. Making sure that efforts are in fact adopted internationally, or can be merged into common tools to avoid duplication of effort.

BaBar: Concern with interoperability and split between US and EDG developments need interoperability between versions or can not possibly make large production grid need support for GT2
SRB team has been responsive & good feedback, s/ware is more mature & fewer problems
if things work no one cares what the code is but when it does not work hooks for troubleshooting are important. We need good connection with developers (50-50 w/ EDG), some fixes have been accepted back into code base & some have not. How to converge on solution with EDG development, LCG development deployment on several sites, VDT in future important to have these meetings & keep improved good communication.

CMS: CMS has proposed in several meetings that PPDG embrace the VDT as the packaging and dissemination mechanism for the core middleware. This provides a focus for discussion of long term technologies and potential funding opportunities. See beneficial influence of PPDG in the computer science projects and related projects such as SRM. Question is whether PPDG is the optimal way of providing development and deployment support for the experiment: No this is not as effective as direct funding and project authority over the full set of funds. How efficient has the PPDG mechanism been. This has been true for the Fermilab part of the project. Should seriously consider the scope, deliverables and goals of any follow on proposal. The goal of engagement of grid middleware groups and the experiments has been a success. The cross experiment fertilization has been lacking and could perhaps be a new focus of the PPDG. PPDG could improve its performance in this area.

D0: Some of the JIM team's schedule for deliverables to D0 have slipped during the last several months. This is partly due to our own underestimates, and partly due to those of the Condor team. We are working to better understand the issues involved and resolve them. We also have some concerns with the level of support provided by the Globus team for GT2 as they move ahead to the new GT3. Although our plans are to transition our projects to Web services, there may be some intermediate steps requiring support of GT2 software.

JLAB Experiments: I'm concerned about the vagueness and abstraction of it all. To computer scientists the grid may be a real, tangible technology. But, to the rest of us it's a rather ethereal entity into which vast sums of research (and commercial) funding are being poured without much to show in the way of *deliverables*. I don't want to be too cynical here - the GRID does seem exciting - albeit in a misty, far off way, and I'm delighted that we've got a certificate which works and got some client software running (with good support and help from the JLAB computer center). As a Linux user / administrator I'm used being able to solve problems by a quick search of the net. The GRID seems to be tightly in the grip of computer scientists and I'm slightly concerned that, for the user, this will result in a loss of flexibility, and ability to customize software and diagnose problems. **JLAB: Pilot deployment.** Support and management of the systems is the largest part of the work. Are concerned with the surrounding support issues and ways of minimizing the impact of the support, management and employment. This is of course the least glamorous, but there must be a way to automate this and reduce the effort load.

STAR: We do and especially about stability in general. We have a total of 1.5 FTE working on Grid technology and doing repetitive testing and stability studies is not practical for these 1.5 FTE. Especially, we foresee a major re-test if not redesign of our approach while we will undergo integration of new versions of middleware (GT3, ...). A suggestion to resolve this problem would be for PDDG to increase or be more pro-active and involved in Globus or middleware development. We would then be in a better position for supporting and developing a long term vision.

1.3 Questionnaires on Common Service Areas

The meetings to discuss the questionnaires on common service areas vary greatly in completeness and quality of the text and review. The CS-4, Storage Resource Management, is by far the most comprehensive and complete - with thanks to Arie Shoshani.

1.3.1 Is Progress as planned in the PPDG proposal?

CS-3: Information and Monitoring Services: The following technologies have been developed, extended and used in PPDG applications: Globus MDS V2 (including interfacing to Ganglia using the GLUE schema), MonaLisa (including integration with MDS and Hawkeye), Condor Hawkeye, IEPM_BW and PingER network monitoring work, site specific fabric management systems (BNL, FNAL(NGOP), etc), D0 SAM mining of monitoring information in ORACLE and text log files. In addition, the GLUE Schema group, a collaboration with Trillium, and the European grid projects, has continued to develop common schema definitions for compute, storage and networks. The pg-monitoring group led by Jennifer Schopf and Brian Tierney identified a preliminary list of requirements for Grid-level scheduling. In addition, a mailing list was started (ppdg-mug for monitoring users group) to aid in the discussion of the initial interests of the application groups which have been focused on fabric monitoring rather than grid monitoring. Work on setting up an MDS2 GIIS for the experiments has been done in coordination with iVDGL and WorldGrid.

CS-4: Storage Management: There are four institutions in PPDG involved in the development and use of SRMs on top of existing storage systems: Fermilab, TJNAF, LBNL, and SDSC (listed alphabetically). It is the goal of CS-4 to have all these storage system accessed through a uniform Grid middleware interface. The key to achieving this goal was to agree on the functionality and interfaces of SRMs, and to have each institution develop their own SRM interface to their specific system. Fortunately, the SRM functionality and specification was already under development by the SRM middleware project at LBNL in collaboration with Fermilab. Building on this initial work, LBNL, Fermilab, and JTNAF staff have developed a specification and an interface that they agreed to follow, called SRM v1.0. The functionality was described in a Joint Specification Design Document and the interface was specified as a web service using WSDL. Furthermore, this functionality was coordinated with two European Data Grid (EDG) working packages: the Data Management Package (WP2), and the Mass Storage System Package (WP5) whose members have developed the Castor Mass Storage System. The benefits of this approach were already demonstrated at SC 2002 by having files accessed using the same interface from multiple diverse Mass Storage Systems (MSS): HPSS at LBNL, Enstore at Fermilab, and Jasmine at TJNAF. Furthermore, since Castor at CERN also implemented a prototype SRM, file transfers between Castor and Enstore have been demonstrated. Furthermore, the LBNL staff that developed the SRM system to HPSS (called HRM) have used two such HRMs at BNL and LBNL (NERSC) to achieve production level robust file replication. This work, performed in collaboration with the PPDG-STAR experiment staff, has been used for over 9 months in a routine way to replicate robustly 100-1000s files per day. Robustness is achieved by monitoring the staging and archiving process by HPSS and recovering from transient failures, as well as monitoring and recovery from network failures. Efficiency of transfer is achieved by using the Globus GridFTP. At SDSC, there is a mature powerful storage brokering system, called SRB (for Storage Resource Broker). Although SRB can be treated as another storage system that can be accessed through an SRM interface, it has other powerful features as being able to access over the WAN various types of storage systems, as well as a flexible metadata catalog. For these reasons, SRM was adapted for deployment with the CMS experiment. The SDSC team developed an Data Movement benchmark between CERN and Fermilab, and integrated the Clarens system (developed at Caltech) on top of SRB to take advantage of its metadata catalog. The SRM design supports the concept of a "site SRM", where the site SRM manages multiple storage system. The SRB fits into this view of SRMs, and consequently there are now plans to use LBNL's SRM as a Grid-enabled gateway to SRB.

CS-6 Globus Reliable Replica Location Service: There has been extensive development effort by the Globus and DataGrid teams in the implementation, deployment and testing of the Replica Location Service. However, to date, the RLS has not yet been incorporated into the PPDG testbed. Andy Hanushevsky's group at SLAC did participate extensively in the RLS testbed that was constructed for the SC2002 conference. This testbed included over 30 servers on three continents. Whether the BaBar experiment will switch to RLS is still under discussion.

CS-7 Documentation: There is a comprehensive web site for PPDG which is used as a resource by the management teams. The commitment to maintenance and keeping it up to date has been met due to the work of Doug and Ruth. The web site was successfully redesigned by the LBNL web services group. People like the new site layout.

CS-9 SiteAA: not part of original 2001 proposal. why not? because we focused on HENP experiment specific features for grid. did not realize that security & user/vo management technology were so immature. proposal in mar 2002 covered site-aaa requirements, issues definition & prototype implementations – achieved proposed results (<http://www.ppdg.net/mtgs/19dec02-siteaa/>, <http://www.ppdg.net/pa/ppdg-pa/siteaa/SiteAAA-Requirements.rtf>, http://www.ppdg.net/pa/ppdg-pa/siteaa/Site_Issues.pdf)

CS-11 Analysis Tools: Progress amounts to active discussion of domain breakdown and interface definitions. Multiple developers have engaged in the discussion. Refinement following CS-2 (job scheduling) , CS-10 (production grids)

1.3.2 What work is needed for the next year of PPDG?

CS-3: Information and Monitoring Services: For the next year, the primary work should be focused on the use of common tools, common schema and naming conventions. Continuing identification of requirements and the involvement of application groups to work together on common protocols and schemas to allow interoperability of the various tools is needed. In addition, careful thought should be made about the move to OGSA-compliant services, including MDS3, should occur in order to take the application requirements into account. We recommend investigation into possible extensions in order to better integrate monitoring services with decision making services/systems and data flow managers for the applications. In additions, extensions to the alarm, fault reporting and fault tolerance capabilities integrated with the monitoring systems are needed.

CS-4: Storage Management: A new version of an SRM specification was developed over the last year, called SRM v2.0. This version was designed to support dynamic space negotiation and reservations, as well as dynamic creation and support for user directory functionality. In addition, the concepts of “pining” and “lifetime” of files will support three types of files to achieve flexibility of files being in temporary storage on their way to an archive. These features are necessary for request planning and execution, as well as monitoring information about Grid storage usage. The main activity planned for the participating institutions is to develop this new more powerful SRM version. Another important aspect of future work in to have a combined service for SRM functions with registration to Replica Catalogs, such as Globus’ RLS. The need to register to a catalog has also to be expanded for registration into specific experiments file catalogs. A very difficult task ahead is access control to files residing on the Grid storage resources. We foresee the need to collaborate with the Community Authorization Service (CAS), where “virtual organizations” will manage storage quotas for its users, keep track of actual usage, and provide users with certificates for their storage requirements. This capability is needed for request planning and applies to all types of resources, including compute, storage, and network. We expect to develop future versions of SRMs that can accept capabilities, award and monitor space usage, and report to the “virtual organization” on space usage by each user.

CS-6 Globus Reliable Replica Location Service: ISI received special funding last year to work on the RLS and so that PPDG sites would deploy and experiment with RLS servers. To date, very little testing and deployment has actually occurred within PPDG. Extension of this funding would allow ISI to continue to support deployment of RLS servers within PPDG. Further tests have started in the ATLAS-iVDGL environment.

CS-7 Documentation: We need to keep encouraging buy in to the need to publicize PPDG documents and talks on the PPDG web site. This is an area of lack that we have tried to address through extra administrative help at LBNL and discussion with the Steering Committee. Neither of these has really been successful.

CS-9 SiteAA: incident handling (+ accreditation, service level agreements, authentication/authorization for long running jobs. federating identity? is it needed?, who holds it? debugging – robustness. migration to OGSA.

CS-11 Analysis Tools: Continuation of discussion & work. Widening of the community involved: Major players (ROOT, Condor) have been opposed to the underlying idea. Resolution/coordination of work with LCG/ITR/EGEE/GGF work. Nailing down interfaces & how they fit w/ use cases. Trying some real grid services interfaces. Need to identify components than interfaces

1.3.3 At the end of PPDG do you think the technology will be mature and robust enough?

CS-3: Information and Monitoring Services: The experiment requirements are evolving and it is difficult to identify the complete set of needs and technologies needed for experiment wide production grids.

CS-6 Globus Reliable Replica Location Service: In the 18-month time-frame, we expect the RLS to be sufficiently mature to support production applications, assuming that funding continues that allows ISI to work with physicists to deploy the service.

CS-9 SiteAA: no, authorization will still be immature, depends on acceptable risk.

CS-11 Analysis Tools: No. This is more of an on-going effort (ie. a work model) rather than one which results at some point in a finished product ready for production. nHope for good starting points but not done. There are different technologies, some mature & some not

1.3.4 Issues with the support of the technologies being deployed?

CS-3: Information and Monitoring Services: .PPDG needs to review the a) intent and ability of the application groups to contribute to the requirements, design, evaluation and adoption of this technology is still up in the air. b) Effort in the project expected to contribute to this common effort c) Clarify and define the areas of work.

CS-4: Storage Management: The SRM concepts were designed for all aspects of the data production and analysis cycle, including data generation, data replication, data analysis, request planning, execution, monitoring and recovery. We expect this approach to continue to payoff as the Grid matures.

CS-6 Globus Reliable Replica Location Service: The current level of funding is adequate for the support that we are currently required to give to PPDG scientists. However, we would expect that as the use of the RLS increases within PPDG, our requirements for supporting PPDG scientists will increase.

CS-11 Analysis Tools: Concerned in the past with the Grid emphasis (solely in some cases) on file-based data management. This is especially pertinent for streamed (ie non-file) results, which then are completely outside the scope of the Grid's vision of data handling. This isn't a fatal gap, but is something left entirely to the user (at this point). Another concern is with the security effect on interactivity. The presence of firewalls at some Grid sites makes use of arbitrary communication channels difficult to impossible. The "Grid as distributed batch-queue" model, allows very circumscribed crossing of firewall boundaries (although this is still a **current** problem with the Grid). Distributed monitoring, streaming of results, etc are far more difficult to restrict without impacting the work. Understanding how to deal with local administrators' **real** concerns around security needs to be a higher priority for the Grid. Architecture & outside influences are important "support" issues for interactive analysis. PPDG is supporting bringing together these disparate efforts

2 Common Service Areas

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

2.1.1 Job Description Languages

A low level activity across experiments continued in the hope that eventually some common language for describing jobs and workflow management may be forthcoming at a higher level in the application infrastructure layer than ClassAds.

2.1.2 Collaboration with EDG WP1

Collaboration with EDG WP1 work continued. Extensions and hardening work concentrated on Globus and Condor in support of the V2 release and the release of VDT V1.1.7 and 1.1.8 for the European Physics Grid community.

2.1.3 STAR Job Scheduling

The STAR job scheduler project has been a prototype for the submission through Condor-g . Beta test made in evidence some site specific assumption we have now removed. The current class design is being reshaped to allow later integration of monitoring services such as Ganglia/MDS. To fully benefit from this prototype, we plan to focus on our FileCatalog deployment and converge on the monitoring. In principle, the components are ready for a larger test scale.

2.1.4 SAM Job and Information Management (JIM)

We have continued work on D0 job schedule and management. The core of our collaborative project with the Condor team is the match-making on the grid, which we introduced into the system last year. This quarter, we have made major progress receiving and integrating the enhancements that their team implemented, most notably, the rematch, the 3-tier architecture (with spooling etc) and other solidifications for Condor on the WAN, such as usage of TCP. The rematch feature was requested by our team from Condor so that we can correct poor decisions (that are bound to happen especially during the system design evolution) and re-consider the job that could not be executed.

Currently, we're in the process of preparations for the JIM v1 release. We have changed significantly the resource description that is advertised. Things are working properly and we plan to have a production release some time in April. We expect to set up 2-3 analysis and 2-3 monte-carlo sites, possibly mixed. Both analysis and MC sites will benefit from the decision-making that we've been developing.

In the last quarter the work of the SAM-Grid team focused mainly on the release of the JIM software suite V1 (production quality version). The team has addressed various items including the following:

Testing and adapting new Condor-G to JIM : JIM uses a branch of Condor-G v6.5, which has a new architecture based on an n-tier submission scheme plus a few other new features. The JIM team worked with the Condor team to determine the requirements and set priorities for the modifications/addition to old Condor-G. The other features of new Condor-G developed within the collaboration include the ability of re-matching a grid job that failed during submission and management of the security of the condor daemon via GSI.

New JIM advertisement framework: JIM_advertise publishes resource characteristics/policies in the form of classads. The classads are derived locally from a structured description of the resources, created at the time of installation by the administrator. The site description is written in XML and his repository is a local native XML database. XQuery is used internally to express the set of rule that flatten the tree structure of the site description into a set of unstructured classad.

Packaging of JIM V1: The work focuses on the organization of the software in packages and the creation of an installation manager, called samgrid, that drives the administrator through the process of installing and configuring the SAM-Grid software suite. This work is done in collaboration with Tom Rockwell from MSU. A draft of the organization of the packages can be found at http://www-d0.fnal.gov/computing/grid/JIM_V1_Pack_and_Installation_strategy.pdf [.doc]

The JIM V1 security model: since the Condor-G daemons now allow for GSI authentication, the security model for JIM V1 has changed. Gabriele is working with the team on a draft to describe such model.

2.2 CS-3 Information Services

Continued extensions to, interfacing and integration of the various information and monitoring services continued with MDS, Ganglia, MonaLisa, Hawkeye, interfaces being used in various applications.

2.2.1 Information Providers and Glue Schema

Work continued on the GLUE schema. The Compute Element (CE) schema is currently being used in the EDG as part of the MDS deployment and is also available from the Globus MDS website. As soon as licensing issues are resolved with INFN, information providers that use these schemas will be included in

the standard MDS distribution as well. In February several additions were made, and a new version of the UML for the schema will be posted shortly.

The Storage Element (SE) schema has been discussed in detail, and a very good meeting was held at CHEP 2003 in March to finalize the effort. It will be used in EDG as part of the R-GMA deployment in testbed 2 in April.

The Network Element (NE) schema has begun to be addressed, and work by the UK EDG group and the Italian DataTag group is currently being combined. This work is also overlapping with GGF NM-WG effort.

CMS and ATLAS are using MDS 2.2 on their production testbeds. Additional work on information providers for other data than the standard core set is continuing. In addition, a scalability analysis of MDS 2.1 and MDS 2.2 was completed, and successfully accepted for publication in HPDC'03. This work is available at <http://people.cs.uchicago.edu/~hai/hpdcv25.doc>.

Plans for MDS 3, an OGS-compliant set of services, are continuing as well.

In addition, work continued with the ATLAS group and WorldGrid on the virtual-organization-cognizant Ganglia monitoring mechanism, which moved from a prototype to a deployable component.

The Glue Schema are being integrated and used in the EDG V2.0 release and new Information Providers and changes to the schema are required. <http://www.cnaf.infn.it/~sergio/datatag/glue/> An interoperability test between DataTAG and the handsome cluster at Fermilab was demonstrated at the review for the DataTAG project...”real job submissions to machines in US and to LCG-0, using the GLUE schema, authorization via VOMS, and a customized version of the Resource Broker GLUE-aware with special JDL options for the VOMS”

Under STAR direction, worked has been done on the Ganglia Information Provider (IP). This work is made possible by Efstratios Efstathiadis from the BNL IT department. We have mentioned our new collaborative members and are very pleased of their efforts and progresses. We have opened a new section of our web page to report and present this work. It is accessible via our Monitoring (<http://www.star.bnl.gov/STAR/comp/Grid/Monitoring/>) section of our Computing/Grid pages.

The Output of the Ganglia IP was modified to match the CE-GLUE Schema. Now all the output from the Ganglia IP gets into MDS properly : the initial try showed a problem with the provider. A bug for this work was submitted to Globus bugzilla as bug# 745). After testing the initial provider, a new Ganglia IP was developed. Written in Perl, it has the many advantages over the original Ganglia IP as *it has the option to connect to the Ganglia Meta Daemon or the Ganglia Monitoring daemon*, its output is compatible with the Glue schema and finally, we believe a perl to be simpler and more flexible.

We also worked on testing the local MDS installation and trying to understand the authorization issues to LDAP. What we needed is to accept authorized access only, turning off anonymous binding and setting up appropriate ACLs (MDS anonymous access is ON by default). This work was in collaboration with IEPM, Network Performance Monitoring¹.

Last, we developed the reversed engineering script. From the MDS information, we are now able, using Java and COG to produce a ldif output of the information. Getting the information is done with authorization (not anonymous). We are planning to further prospect the reverse engineering approach as in direct connection to our program of having a monitoring information for the job scheduling and resource brokering work.

2.2.2 IEPM-BW collaboration

2.2.2.1 Web Services

We worked with Internet2 folks to put a web service front-end on their one-way measurement (OWAMP) data. The WSDL is available from

¹ <http://www-iepm.slac.stanford.edu/>

http://thunderbird.internet2.edu/~cottrell/wsd/get_owamp.wsd. A next step is to work with Jim Ferguson of NLANR/DAST to modify their Advisor to access the data. We are also trying to build a similar front end for the NASA iperf data. We met with the MonaLisa developer at CHEP and are looking to integrate our measurements with their toolkit. We met with Eric Boyd and Russ Hobby of the Internet2 PiPES project at SLAC to discuss coordination and collaboration. We prepared a presentation on our work with the Internet2 PIPES group on making data available and trouble detection (see <http://www.slac.stanford.edu/grp/scs/net/talk/chep03-pipes.html>). In order to proceed with publishing network monitoring data in a standard fashion, we are part of a GGF working group (NMTF) working on developing a standard naming hierarchy for the measurements. This was presented to GGF7 in Tokyo.

2.2.2.2 PingER

PingER has been turned into Perl module format, to enable much easier installation of the monitoring code at remote sites.. It will be released on the web in April 2003. The number of sites being monitored as part of the Digital Divide study has been increased to 27 located in 20 countries. We met with Enrique Canessa of the Abdus Salam International Center for Theoretical Physics to go over future plans for extending PingER. We made both a vocal and a poster presentation at CHEP04 on "Monitoring the Digital Divide" based on the PingER measurements. In order to further spread the publicity on the Digital Divide we arranged to make a presentation at the "Hard to Network" BOFF at Internet2 meeting Washington in April. We produced a formal report for the ICFA/SCIC on Network Monitoring with an emphasis on the Digital Divide, see <http://www.slac.stanford.edu/xorg/icfa/icfa-net-paper-dec02/>

2.2.2.3 IEPM-BW throughput Measurements

IEPM-BW has been re-written to improve robustness, and simplify adding new sensors. The FNAL improved IEPM-BW time series plots have been incorporated into the production release. FNAL has become much more active in IEPM-BW and is now energetically deploying it to enable monitoring of the major FNAL projects (D0, CDF and CMS). With LBNL, ICIR and PSC we put together a proposal called MAGGIE to build and provide initial deployment of a new measurement infrastructure for both repetitive and on-demand measurements. We made a presentation on MAGGIE (see http://www.slac.stanford.edu/grp/scs/net/talk/i2_feb03/e2e_Miami.pdf) at the Internet2 Joint Tech Meeting. As part of the pre-work we brought the NIMI measurement host at SLAC back to life. We made presentations on IEPM-BW at the Protocols For Long Distance (PFDL) Networks meeting in Geneva (see for example <http://www.slac.stanford.edu/grp/scs/net/talk/pfdl-feb03.ppt>), and the SCAMPI meeting in Amsterdam. We also had papers on this accepted at PFDL and PAM03. We made a presentation on the IEPM project at the ESnet Site Coordinators Committee meeting (see http://www.slac.stanford.edu/grp/scs/net/talk/i2_feb03/escc_Miami.pdf)

2.2.2.4 Bandwidth Estimation

The ABwE lightweight bandwidth estimation toolkit has been carefully evaluated and now provides good bandwidth estimates in over 80% of the cases. It can make an estimate in real time (< 1 second) with minimal impact (40kbits). It is getting to the point where we now use it to indicate when we need to re-evaluate the parameters (windows & streams) used for our heavier weight iperf estimator. With its real-time capability and low impact it is very suitable for providing real time feedback of anomalous changes in bandwidth performance. We had a paper on ABwE accepted at the PAM03 conference.

2.2.2.5 Traffic Characterization

Following a request from DoE headquarters, we worked with ORNL and NERSC to make public our work on NetFlow passive monitoring of the traffic flows at the SLAC border. Much of the work went into ensuring the reports did not expose private (to SLAC) information, and making the data understandable at a higher level.

2.2.2.6 Testbed

The loan of the Sunnyvale to Chicago testbed and link was kindly extended by Cisco and Level(3) until early March. We were also successful in obtaining the loan of several Intel 10GE NICs. These were successfully used to break the Internet Speed Record once again, achieving 2.36Gbits/s over a 2.5Gbits/s bottleneck. We installed 3 new TCP stacks on the Sunnyvale testbed hosts, and compared their

performance with the standard Linux TCP stack, as well as with the use of multiple parallel streams and also with jumbo frames.

2.2.2.7 IPv6

With the renewed academic interest in IPv6, we have are reviving our IPv6 testbed so we can make tests and understand the ramifications etc. AS part of this we are working on procuring a second monitoring host from the AMP project, for IPv6 work.

2.2.3 Mona Lisa

A dynamic pseudo-client framework for the MonaLisa service was developed. This allows the creation of a dedicated WEB Repository containing information on global services, by using selected information from groups of MonaLisa services. The technology used is a set of JINI Lookup Discovery Services that discover all active MonaLisa Services from a specified set of groups, and then subscribes to these services using a set of predicates and filters. These predicates or filters depend on the information the pseudo client wishes to be collected. All the values received from the running services are stored into a local database. The database technology is MySQL, coupled with Java thread procedures that compress old data. These procedures calculate mean and minimum/maximum values in the data, which then replace the stored data. This allows any large fluctuations to be observed over any required history period. The servlet engine of the Web repository uses the database to generate a set of customizable charts showing current and historical data. The engine is also capable of generating Wireless Access Protocol (WAP) pages showing the latest data. These can be accessed from Mobile Phones. Using this infrastructure, multiple Web repositories could be created that describe the dynamic services running in a globally distributed environment.

Currently, the US-CMS repository contains aggregate data history for all US-CMS sites, as well as the traffic across the major networks we are monitoring. Compute farm usage is also stored and being made available from WAP-capable mobile devices.

We have deployed MonaLisa at an ATLAS center in Taiwan, where we monitor a farm of approximately 80 nodes, and a newly established WAN connection to Chicago running at 622Mbits/sec. In addition, an interface to the LSF batch system at CERN has been developed and is being used to collect data from the LXBATCH cluster (~600 nodes) as well as information about the cluster jobs (e.g. running, pending, executed, exited with errors).

We have also deployed MonaLisa on a set of VRVS reflectors. We have built specialized modules for collecting information related to the VRVS system. These modules allow us to dynamically discover the topology of connections between VRVS reflectors. They also collect useful information such as lost packets, numbers of clients and active Virtual Rooms. More general modules that collect CPU load data and I/O traffic rates are also used. For the VRVS system a dedicated global Jini Client was developed that presents the VRVS specific information.

The work on MonaLisa will continue its goal of providing flexible user interfaces to monitoring and control of such complex distributed systems with real-time constraints. <http://monalisa.cern.ch/MONALISA/>
<http://monalisa.cern.ch:8080/CMS/>.

2.3 CS-4 Storage Management

Progress was made on the definition and implementation of SRM V2.0 by various sites. The EDG WP5 and Castor project made plans for implementation of an SRM V2.0 interface. A joint paper between US CMS, SRM and the LHC Computing Grid project on Grid File I/O protocols – which is related to some of the PPDG work - was circulated for comment and discussion. http://cern.ch/pkunszt/GridFAP_gag_PK.doc

2.3.1 LBNL-SRM Development

People involved: Junmin Gu, Alex Sim, Vijaya Natarajan, Arie Shoshani.

2.3.1.1 Enhancements to HRM to support directory replication

Currently, Hierarchical Resource Managers (HRMs) are used to support file replication between HPSS systems at BNL and LBNL for the STAR project. The multi-file replication is invoked through a Command-Line Interface, called HRM-CLI. The HRM-CLI is designed to support multi-file requests, as well as a directory structure. However, to move an entire directory, the HRM-CLI needs to get the names of the files from the source HPSS. This is not possible if the HPSS is behind a firewall as is the case in BNL. To circumvent this difficulty, we have added a function call to HRM to list the files in an HPSS directory. Thus, the HRM-CLI calls the source HRM at BNL, which in turns calls the HSI-interface to HPSS to get the names of file in the directory. During this quarter, this capability was added to the HRM. Also, at the target HRM, we added the capability of creating a directory (mkdir). These new features will permit a flat directory to be replicated in a single call. Testing will start during the next quarter.

2.3.1.2 Development of an enhanced File Monitoring Tool

The Storage Resource Management (SRM) File Monitoring Tool (FMT) is a tool that permits a web-based tracking of progress of a large multi-file replication request. The FMT has two components: the FMT-server that is a daemon co-located where the target HRM is, and the FMT-Client that is invoked by the user as a web-service. The FMT tools worked very well, but since it was designed to be only memory-based, it gradually took over more and more memory. Also, the frequency of updating was interfering with the operation of HRM making it less efficient. In the new design, we minimized the amount of information kept in memory, and saved the rest to a file on disk. We also synchronize the FMT client and server infrequently (every 20-30 sec), but display the results by using interpolation more often (as low as a 1 sec refresh). This new version of FMT is much less intrusive, and therefore we expect to have it in routine used in the STAR project to track file replication. We also added summary features, where the user can find a summary status either during transfer or after the completion of file transfers. The graphical user interface was also enhanced, so that the display represents files in the order of there transfer initiation. This helps tracking visually the transfer of files regardless of their size.

2.3.1.3 Revising versions of SRM 2.0 interface spec document

SRM v2.0 is a new version that is designed to support space reservation and directory structures on demand. These are essential features for the next version of SRMs that can support request planning and execution. A time consuming activity during this quarter was the refining of the SRM v2.0 specification. We received a large number of comments that required rethinking of the specification. Three new interim versions were developed.

2.3.1.4 Various tasks related to SRM and PPDG

- a) WSDL prototype using gSoap with GSI plugin. Our plan is to develop an SRM wrapper based on SRM version 1.0 WSDL. To this end we are developing a prototype using gSoap with GSI plugin. We also tested this version with Fermi's Axis implementation as well as testing with GT3's Axis implementation. Our plan is to develop a Java based wrapper that will also use Axis during the next quarter.
- b) Continued coordination with NeST. At our request, a new feature was added to NeST, where a file can be associated with a single lot. When using a DRM on top of NeST, we can take advantage of this feature, by having NeST check that files written into a lot do not exceed the allocated space. During this quarter, we continued to coordinate with the NeST developers and tested the newest release of NeST and reported bugs and desired features.
- c) We also performed debugging for some HPSS-pftp problem on PDSF for that Eric Hjort has encountered.
- d) Preparing a 2-pager at the request of SciDAC office.
- e) Preparing an SRM poster for the SciDAC PI meeting.

2.3.2 JLab-SRM

Jefferson Lab has continued to develop and deploy data grid web services, with a particular focus on Storage Resource Management (SRM) software (server and client). JLab is working with Arie Shoshani

and the SRM group at LBNL to develop an interoperable specification for this particular web service, with an expectation that additional interoperable specifications will allow the specification of a complete web services data grid.

At the present time, Jefferson Lab has developed, as prototyping activities, two separate data grids: one using the lattice QCD project as the customer, and the other using experimental physics as the customer. The lattice QCD grid work has been a key prototyping platform within the broader SRM international activity because of its early incorporation of capabilities beyond those specified in the SRM version 1 document. The experimental physics SRM product was intentionally geared towards SRM version 1 interoperability (demonstrated last quarter in collaboration with FNAL and LBNL).

During this quarter, the effort was focused upon furthering the design of SRM version 2 to support all of the features actively used by Jefferson Lab (particularly directory related functions). The lab has actively contributed API design ideas, and provided feedback on multiple versions of this document.

In anticipation of a more complete API, work has also begun on merging the two prototypes into a single product, so that in version 2 of the Jefferson Lab SRM (JSRM) the following design features will be present:

- single SRM supporting multiple disk management policies:
 - strict (files owned by daemon, as implemented in Jasmine)
 - auto-delete / auto-migrate (cache daemon), but files (and space) owned by users, modifiable via NFS operations without SRM interaction
 - user managed: conventional NFS user file system, no daemon, but files visible to SRM clients
 additional management policies can be implemented by inheriting from a base class
- support for the Jasmine silo & disk management backend
- abstract interface for the connection between disk management and tertiary storage, so that a different silo system could be plugged in below the SRM services
- reliable file transfer services, including protocol negotiation (currently a separate service within the J-SRM prototype; this implementation will be folded into JSRM as per the draft SRM v2.x specification for “copy” operations)
- interoperability for GSI based web service invocations (J-SRM previously used an incompatible method for certificate delegation)

Most of the design work is now finished, and parts of the SRM version 2 specification have already been implemented, providing additional feedback to finish that design.

2.3.2.1 JLab-QCD

Jefferson Lab is serving as the liaison between PPDG and the Lattice QCD SciDAC project, and also with the International Lattice Data Grid, a collaboration between US Lattice researchers and their counterparts in the UK, Japan, Germany, Italy and eventually other sites.

These different collaborations have agreed to base their data grids (at least for standard, collaborative activities) on web services for data grids, with the SRM as the first component to be deployed. Interoperability, not shared software, is the expected path for this work. As an example, the UKQCD data grid is currently implemented as multi-site, replicated, disk resident data files with custom (UK specific) interfaces. To this software they will add an SRM version 2 conforming “face”, allowing interoperability within the International Lattice Data Grid.

The pace of activity for this international work is expected to grow in the coming quarter as the version 2 specification is finished, with a virtual workshop planned for May, and additional discussions taking place in the following quarter at the annual Lattice 2003 conference in Japan in July.

Within the US, interoperability is also the driving force. FNAL will also be a major LQCD site, and the lattice work there will likely leverage the Enstore system. As already demonstrated for SRM v1, FNAL will have an SRM v2 interface, which can be used for the FNAL lattice grid node.

2.3.2.2 JLab-Experiments

During the past quarter, additional Jefferson Lab user sites have been added to support data grid testing, now including one site in Europe. Even the version 1 prototype, which contains only basic file copy capability, is already proving to be a useful tool. As the 2 existing JLab prototypes are merged, the experimentalists will gain additional features not present in the version 1 design deployed for the CLAS collaboration.

2.4 CS-5 Reliable File Transfer

Work on reliable file transfer middleware applications has been lowered in priority within PPDG in order to concentrate on robustness and hardening of file and data transfer technologies. Work continued in improving the reliability and multi-stream efficiency of GridFTP. Work over the last quarter focused primarily on completing the design and coding of XIO. XIO will amongst other things, resolve the hanging issues with globus-url-copy (bug 256) that are currently being worked around. The design is complete, the framework, TCP, and file drivers are done, and a GSI driver is in progress. This is part of a larger effort to completely reimplement the Globus GridFTP server, to improve its stability, maintainability and to ease addition of features. It also resolves licensing issues discovered with the wuftp code.

2.5 CS-6 Robust Replication

2.5.1 BaBar Database Replication (BaBar-SRB)

The focus has been to set up a system to transfer data from SLAC to IN2P3 (Lyon, France) using the SRB. Together with Liliana Martin from the Univ. Paris, we are working on scripts that assist in transferring many files from SLAC to IN2P3. These scripts allow to select files by attribute, retrieve the files from HPSS at SLAC and transfer them to IN2P3. Liliana was visiting SLAC in January for a week and we worked together on developing and testing the scripts. A new version of SRB (SRB2.0) was released in February. We installed this version at SLAC and IN2P3. For the SLAC SRB a new MCAT was created in order to test the old and new SRB versions independently. The new SRB version implements a new parallel stream file transfer protocol that increases the data transfer rate significantly compare to the SRB versions prior to 2.0. High transfer rates is a critical requirement for us. First test were performed which show that the new SRB version improves the data transfer, a rigorous evaluation is underway. We encountered fire wall problems with the new SRB version, which requires us to reconfigure the setup at SLAC. During CHEP2003, we set up a demo that transferred data from SLAC to SDSC. Transfer rates up to 18 MByte/s were achieved. Test were done to understand the performance, but we still have to compare it to other tools (bbcp, bbft, gridftp) in order to evaluate the performance.

2.5.2 Globus ISI, RLS work

2.5.2.1 Continued development, packaging, testing and deployment of the Replica Location Service

This quarter saw intensive continued testing of the Replica Location Service, including debugging and extensive functionality and performance testing. The RLS was packaged for release in Globus Toolkit 2.4 and in GT3 Alpha. The RLS was used by an increasing number of groups, including the LIGO physics application, the Earth Systems Grid and the Chimera system.

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

Some progress was made on creating OGSA services for the Replica Location Service. This was delayed somewhat by the lack of a C hosting environment in OGSA and by discussions within the Global Grid Forum about what the correct interfaces should be for the RLS services.

2.5.2.3 Plans for next quarter

1. Resolving licensing issue for RLS
A licensing problem was found with the MySQL database currently used as the back end for RLS services. Unless this problem is resolved, it will be necessary for us to switch in the coming quarter to the postgresql database backend. This will require additional debugging and testing.
2. Release of RLS in GTR
While the issue of the backend database is being resolved, the RLS will be released in the Grid Technology Repository (GTR).
3. Release of RLS in GT3.2 and possibly GT2.6
Once the database issues are resolved, we plan to release the RLS in stable Globus releases. This includes GT3.2 and also GT2.6, if there is such a release.
4. OGSA RLS Services
In this quarter, we will provide OGSA wrapper services around the existing RLS Local Replica Catalog and Replica Location Index services. These wrapper services will be deployed in the current OGSA Java hosting environment.
5. OGSA Copy and Registration Service
This quarter, we will also create an OGSA service that performs copy operations using the Reliable File Transfer Service and registers new replicas in the RLS. These operations will be reliable, with rollback of unsuccessful or partial operations.

2.5.2.4 Presentations Given

January 17, 2003: Replica Location Service presentation at GlobusWorld.

February 5, 2003: Replica Location Service presentation at NASA Information Power Grid Meeting.

March 5, 2003: Birds of a Feather Session on OGSA Data Replication Services at Global Grid Forum in Tokyo, Japan.

2.6 CS-7 Documentation

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

Reports, Documents and Papers		Date/Version
PPDG-30	Questionnaires to and Answers from Experiment and Common Service areas	html
PPDG-29	A globally-distributed grid monitoring system to facilitate HPC at D0/SAM-Grid MS Thesis, The University of Texas, Arlington; Abhishek S. Rana	pdf
PPDG-28	Site-AAA: Requirements List	3/03 (rtf)
PPDG-27	Site-AAA: Issues List	1/25 (pdf)
PPDG-26	Report from the TroubleShooting Workshop	V1.0 pdf

		doc
PPDG-25	Site-AAA: Recommendation for Future Activities	1/03 (doc , pdf)

Talks and presentations: (http://www.ppdg.net/docs/presentations_list.htm)

Presentations & Publications

March 2003	Talk at HICB meeting , Doug Olson; DOESG Review , Doug Olson; PPDG-related presentations at CHEP2003 ; Slides for SciDAC PI meeting: 4page, 1-page . CS-11 Workshop
January 2003	Talks at GlobusWorld ; talk at the iVDGL/GriPhyN EAC review

2.7 CS-9 Security, Authentication, Authorization, Accounting

2.7.1 Certificate/Registration Authority

In addition to the routine operation of issuing and renewing certificates the PPDG RA was involved in the PMA discussions about the new CA being set up for the doegrids.org domain. This will involve some changes to the namespace to be documented in an update to the CPS. There are also modifications to the user interface for certificate requests.

2.7.2 Site-AAA

2.7.2.1 GGF7 site-aaa working group meeting.

The accounting section of the Site-AAA requirements document was reviewed and discussed at the Site-AAA working group meeting at GGF7 in Tokyo. After some discussion it became apparent that accounting in the security area is better called auditing in order to distinguish from the accounting done for resource utilization reports.

2.7.2.2 Globus Site-AAA work

While the Site-AAA WG has concluded regular meetings, work continues on implementing the authorization callout interface designed in conjunction with this group. The goal of this work is to allow easy addition of policy evaluation modules for site-specific (e.g. FNAL's centralized certificate policy validator) or third-party (e.g. VOMS, CAS) authorization systems authorization systems to the Globus Toolkit at deployment time. Coding has been initiated and we estimate it should be available to PPDG sites in the next quarter.

2.7.2.3 Globus CAS

Working with Doug Olson and Craig Tull, we successfully experimented with using CAS as a VO group server. These results were presented at CHEP03: <http://chep03.ucsd.edu/files/441.ppt>

A general update on the CAS architecture was also presented at CHEP03: <http://chep03.ucsd.edu/files/518.ppt>

Work has started on integrating CAS as a production service along with a CAS-enabled GridFTP server in the main Globus Toolkit release. It is unclear at this time if this work will appear in GT3.0 or the subsequent release.

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

2.7.2.4 BNL ATLAS VO Management

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

We developed the Grid User Management System (GUMS). The particular problem that we intend to address is the need for strong pre-registration of users, and we believe that this builds nicely on the existing grid VO management software. Our focus has been on developing a user account management system and tools to allow sites to keep track of users. We intend all of these developments to be compatible with whatever VO management tools are adopted for LCG. GUMS downloads certificates from VO server, then stores the certificates in the local MYSQL database. The local account manager maps the newly discovered Distinguish Name to a grid account and groups based on local policy. The mapping information will be stored in MYSQL database also. A simple Python script scans the MYSQL database and creates the Gridmap file. The software package can be obtained from Web site:

<http://www.atlasgrid.bnl.gov/testbed/gums>. The presentation is available at

<http://chep03.ucsd.edu/files/363.ppt>. The paper is in the writing processing. It will be published in CHEP03.

2.7.3 US CMS, US ATLAS, INFN, iVDGL Joint VO Management Project

In response to the immediate needs of the US CMS, iVDGL and US ATLAS grid applications a joint project for Registration, VO management and site authorization callouts is underway with Tanya Levshina as the project leader. Collaboration is ongoing with the Globus CAS and INFN VOMS teams

<http://www.uscms.org/s&c/VO/>.

2.8 CS-10 Experiment Grids and Applications

2.8.1 ATLAS

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

The Magda effort is led by Wensheng Deng.

Magda has been adopted by the International ATLAS Collaboration as the primary metadata and replication service for ATLAS (data challenges and user interface). The system is based on an SQL (MySQL) database at the core of the system. DB interactions occur via perl, C++, java and cgi (perl) scripts. A web interface for browsing files resulting from the current data challenge for the high level trigger has been implemented, along with command line tools (magda_findfile, magda_getfile, magda_putfile, and magda_validate).

For the current high level trigger data challenge for international ATLAS, data are cataloged from the entire US ATLAS Grid Testbed, Alberta, CERN, Lyon, INFN (CNAF, Milan), FZK, IFIC, NorduGrid, and RAL, and made available to the Collaboration. So far, 264k files have been cataloged, representing 65.5 TB of data. The system has now been tested for scalability up to 1.5M files.

Magda has now been in stable operation since May 2001,. It is now being employed by the Phenix experiment at RHIC and is being evaluated by other collaborations.

Current and near term work includes an implementation of Magda as an option for a catalog back-end to the LCG POOL persistency framework. It is being tested by the EDG testbed. Further details can be found in a talk presented by Dr. Deng at the CHEP2003 conference:

http://www.atlasgrid.bnl.gov/magdadoc/magda_chep2003.ppt

2.8.1.2 Monitoring

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

In the last quarter, there was work to integrate Ganglia (large scale fabrication monitoring toolkits) into Globus Monitoring and Discovery Service. Here, we implemented the Ganglia information provider which

publishes the data from the collector daemon of Ganglia (gmetad) into MDS. The existing Ganglia information provider only works with individual daemons and it does not support hierarchical monitoring while ours does. The front-end of our information provider uses Glue-schema (<http://www.cnaf.infn.it/~sergio/datatag/glue/>), and its back-end uses XML.

The work will be published by CHEP03: "GridMonitoring: Integration of Large Scale Facility Monitoring with Meta Data Service in Grid Environment" This abstract and full version of paper describe a Grid Monitoring Architecture that captures and makes available the most important information from a large computing facility. The presentation can be downloaded at: <http://chep03.ucsd.edu/files/321.ppt>.

2.8.1.3 Production support

The PPDG portion of this effort is led by Dantong Yu of Brookhaven National Laboratory.

A major activity during this quarter was the support of high level trigger studies, which required a very data-intensive effort to make "pile-up" events. At high luminosities, as many as 40 minimum bias events can overlap in one beam crossing. It is critical to take this into account in the trigger studies. This effort was done as part of the US ATLAS Grid Testbed.

About 4TB data was successfully produced at BNL site during the production. We also experienced and fixed numerous problems:

- a). The ATLAS LINUX farm alone can not provide enough computing resource for the peak production. We managed to setup a batch queue which harvests the unused computing resource at RHIC sites. The batch queue could provide twice computing power as the existing ATLAS linux farm if no higher priority jobs is running in the nodes of the batch queue. Many of ATLAS production can be re-directed to this job queue. It is proved that the computing resource at different collaborations can be efficiently and effectively shared.
- b). During the production, BNL experienced twice power failure, twice system hung, Network outages, Gatekeeper, out of disk space. We coordinated the site administrator to fix these problems.
- c). The Globus software have cache corruption, version incompatibility, scalability and robustness issues during the production. More stable globus middleware is needed for large scale production. These issues are being communicated back to the globus team. Our understanding is that the scalability issues have been observed by other groups as well, and the Globus team is working on future releases with better scalability.

2.8.1.4 Prototype testing of emergent grid tools and interoperability

This effort is being led by Jerry Gieraltowski of Argonne National Laboratory.

US ATLAS has established a prototype subset of its grid testbed in order to evaluate emerging middleware and coordinate interoperability tests. Work was done to assess the availability of the EDG testbed compute queues by creating a script to submit and monitor simple jobs. The results showed that, on average only 60% of the published resources are really available. Work is being done to evaluate details of RLS on one of the EDG development testbeds.

Chimera is a major Virtual Data service that is being evaluated by US ATLAS. At this point, we are becoming familiar with its capabilities. The interfaces to RLS and ATLAS software have been established. In particular, Chimera has been deployed on a number of US ATLAS grid testbed sites via the standard installation procedures. Chimera and RLS have been used to successfully execute simple event simulation programs in order to understand its behavior. In the next quarter, we intend to test Chimera/RLS with increasing challenges in terms of CPU and data requirements.

2.8.1.5 Distributed Data Management

This effort is being led by David Adams of Brookhaven National Laboratory.

Work was done on providing a collective view of even data and a framework for distributed interactive analysis of such data. The data collections (termed datasets) are described at <http://www.usatlas.bnl.gov/~dladams/dataset>. The overall DIAL project is described at: <http://www.usatlas.bnl.gov/~dladams/dial>.

A substantial effort was spent on the design and development of datasets and DIAL. The code was ported to gcc3.2 and Xerces 2.1. It is expected that ATLAS will move to the newer versions of gcc and Xerces in the second quarter of this year. LCG is already using these versions.

All of the dataset and DIAL classes can now be imported into the ROOT dictionary using ACLiC. This required some effort on the part of ROOT team. The package dial_root builds the associated libraries and provides a ROOT macro to do the import. This is an important in that it allows ROOT to be used as the interactive interface to DIAL.

2.8.2 BaBar

We have managed to successfully launch applications to make a deep copy Collection of events of interest using a BaBar BdbCopyJob application through the grid (from Edinburgh to SLAC), but just using pure GLOBUS commands. So far, we have only tried 2 BdbCopyJobs through the grid and we kept track of the jobs manually. We have not tried the extraction of the deep-copied data yet. We've started to try to stress test the gatekeeper with simple /bin/date commands (currently using the job-fork command) to understand where the bottlenecks could be. Alasdair presented at CHEP a poster on the BdbServer++ plans: <http://www-conf.slac.stanford.edu/chep03/register/report/abstract.asp?aid=456> (the poster should become available soon from the BaBar web pages). and also presented plans at the recent BaBarGrid meeting in Karlsruhe:

<http://www.slac.stanford.edu/BFROOT/www/Computing/Offline/BaBarGrid/meetings/Karlsruhe-April03/agenda.html>

I have spent ~10-20% of my time on re-designing the BaBar specific oracle metadata tables that will be used for BaBar data distribution using the Grid. The code needed a major re-write due to a lot of code duplication making maintenance difficult. This task is now completed, we should be able to populate the new tables over the coming weeks and hook them up to SRB.

2.8.3 CMS

2.8.3.1 CMS Production Accomplishments

I have been involved in development of MOP, improving its functionality. The functionality available through commandline calls only is also available through Python functions. But the most important changes are the MOP abilities to create complex DAGs consisting of several smaller DAGs. This new facility is used to write a MOP-DAG-Creation module for CMS Tool MCRunJob to create comprehensive DAGs for chained production. Micro-DAGs are created for ever-step of chained production, and then tied-up in a MegaDAG. Then another tool built over this helps in submitting these DAGs to any MOP site, and even Multi-Site submission is tested successfully. Multi-Site DAGs allowed few parts of chained production to run on one MOP site and then rest of it completed on other site. This helps accommodate various limitations of the MOP sites, still avoiding under-utilization of resources.

In addition CMS DGT and IGT activities continued, Production of 300K events requested by some FermiLab Physists was achieved over Integration Grid (IGT). Development Grid (DGT) and IGT have abandoned GLOBUS Certificates, using DOESG certs, and also attempts are being made to use KCA Certificates, actual deployment and testing of KCA Certs is underway. New version of VDT and planned to be deployed and tested, while LCG-0 test deployment is underway and I am very much part of all these activities.

2.8.4 D0

The D0 Collaboration has further developed its plans to establish Regional Centers for data production and analysis operations. These centers have significant computing resources for use by institutions within the geographic/political region. They also represent significant resources for the D0 experiment at large. The first such working center is at GridKa, located at Forschungszentrum Karlsruhe (FZK) in Germany, has provided resources used for analysis presented at the winter HEP conferences. We also have centers in

Lyon France (CCIN2P3), the UK, and will have one at U. Texas Arlington in Summer of 2003. Among the essential components of each of these centers are the SAM-Grid (SAM and JIM) servers.

2.8.4.1 D0 JIM Deployment

We will begin deployment of the first phases of the JIM software to D0 in the coming months. Generally, we expect JIM deployment to follow that of SAM, so that the new services are deployed wherever there already is a need and use of data access. Depending on the collaboration needs, however, we may, with the help of the core SAM team, deploy JIM together with SAM. Though theoretically possible, we will not in practice have JIM installations at sites without at least the file storing services from SAM. In what follows, SAMGrid refers to the network of sites having both JIM and SAM.

The initial production JIM deployment is planned for April 2003, with about five participating execution and twenty submission sites (see below), excluding FNAL. We anticipate that SAMGrid will grow to about 50 execution and 200-400 submission sites within a year. A site can join the SAM-Grid installing a combination of the following services:

1. Monitoring site AND/OR
2. Execution site AND/OR
3. Submission site (includes client machine and spooling machine, optionally combined)

Service 1 allows the monitoring of jobs and data handling services (e.g. a SAM Station) via the web; service 2 enables a site to accept and execute user jobs; service 3 enables the users at a site to submit jobs to the Execution sites that are part of the Grid. The standard (typical) site will deploy all three groups of services. We expect the roles of the execution sites to overlap such that about three sites are available for each of the following job types: analysis, reconstruction, Monte-Carlo.

2.8.5 JLab experiments, and QCD

Jefferson Lab continues to support two test grids, one for the Lattice Hadron Physics collaboration (a component of the SciDAC collaboration), and one for experimental physics simulation and analysis.

The lattice grid still has operating nodes only at MIT and Jefferson Lab. Expansion to the University of Maryland (delayed by a fire effecting their computer room), and started again at the end of this reporting period. They are now testing a set of RPM's (Redhat Package Manager) that have been developed at JLab for deploying a compute cluster and a data grid node.

The list of experimental physics sites using the Jasmine SRM v1 prototype includes Florida State University, the University of Glasgow, Old Dominion University, and the University of Regina. Carnegie Mellon University will be added in the coming quarter.

2.8.6 STAR

2.8.6.1 File Replication

HRMs are in constant use to transfer data from the RCF to NERSC, as always. In addition, two new ways for using HRM data transfer were tested. First, we tested and implemented the reverse direction copy, that is, from NERSC to the RCF. This differs from usual RCF to NERSC transfer in the HRMs are not allowed to write into RCF/HPSS. So instead, the destination has been chosen to be a disk and a DRM is used.

The second feature we have tested involves transfers we were accustomed too i.e. from the RCF to NERSC. Instead of getting files out of HPSS, the files are retrieved from an NFS mounted disk. We needed to test this mod of operation as the STAR experiment is undergoing data production of the d+Au collision and we needed to test data transfer of data as they are produced. No HRM or DRM is needed at the RCF in this case. This mode appears to provide a much better transfer rate than using the intermediate step of getting files out of HPSS and it is used for as many files as possible. Only when files are not accessible from a disk, they are fetched from RCF/HPSS and replicated on a centralized storage.

2.8.6.2 File Catalog

We have started the deployment of our metadata Catalog at PDSF/NERSC. Although we have not finalized the deployment (currently involving a complete inventory in the Catalog of the files accumulated over the past years), we have initiated discussions with the SRM group and will help defining the requirements for the “pluggin” of Catalogs into the SRM tools. Our goal is to have our files automatically registered as they are transferred from site to site. Implementation pending, we have out the basic foundation for this work we will resume after the PPDG review.

Our FileCatalog work and automatic registration of files at BNL and how it is related to the Scheduler, monitoring and database work was presented at CHEP in the presentation "Using distributed resource in STAR. An overview of our tools and architecture".

2.8.6.3 Job Scheduling

The STAR Job Scheduler was installed at PDSF. This allows users to have the same interface for submitting jobs at both STAR main computing sites. However, at PDSF not all of the underlying infrastructure is in place so the full functionality of the scheduler is not yet implemented. In particular there is not yet a functioning file catalog. Installation is in progress and for the time being, users are provided with premade lists of files for the jobs scheduler (the STAR scheduler can either access files via a query, use a dataset or loop over a pre-declared list of files).

At the RCF, we have developed a set of [Web based monitoring](#) tools to gather feedback on usage, users habits, preferred access method for files (centralized storage or local distributed disk). This prototype will further evolve.

We mentioned in section 2.1 that work and progress was also made in developing a prototype for submitting jobs through Condor-G .

The STAR Scheduler conceptual design was presented at CHEP03 as a Poster "The STAR Scheduler Project. A Job submission tool for distributed resource environment".

2.9 CS-11 Grid Interface with Interactive Analysis Tools

A CS-11 workshop was held at UCSD just before Chep 03 <http://www.ppdg.net/mtgs/20mar03-cs11/> with an achieved goal of expanding the list of identified interfaces and specifying at least two interfaces in some detail. We covered an introduction to OGSA (borrowed slides from GlobusWorld and OGSA-DAI Tutorial). In principle OGSA provides a framework for specifying the interfaces useful for interactive work. At the moment however, it appears that it will help application developers considerably when more tools are available to assist developing OGSA-compliant grid services. We agreed that long term we will move to OGSA, but may have to make other short term choices.

We have reviewed sequence diagrams for various prototype grid-enabled interactive analysis systems

The San Diego meeting made progress in defining the set of interfaces needed by interactive analysis tools on the grid. We extended the list of what APIs are needed, a general description of each API and pseudo-code drafts of at least one or two specific APIs:

- o API for generic access to datasets.
- o API for access to results including partial results and collating of results from multiple servers.

We chose to focus on two specific APIs because this was the smallest number that leaves us something to work on if we get stuck on one of them. But even getting one API written would be sufficient. The point was to come out of the San Diego workshop with something concrete.

We chose these particular two APIs because, based on our previous discussions, these seem to be two areas that have particular relevance to interactive analysis.

Our white board discussions in San Diego specify interfaces in Java Pseudo-Code rather than any specific interface language such as WSDL or CORBA so that we could focus on the intellectual content of the

interfaces rather than their expression in any particular language (though we will, of course, be sensitive to the need for whatever we write to be expressible in specific languages).

Our document, Grid Service Requirements for Interactive Analysis, summarizes use cases as seen by the end user physicist. The next step was to see how those requirements translate into interfaces required between the analysis tool and other grid services. We have a very rough API diagram. This is just presented as a starting point to encourage discussion. A useful step to refine that diagram would be for each of the analysis tools to provide sequence diagrams showing how they use those components/interfaces. Some of these were presented in phone meetings, and a more complete set, including presentations from DIAL, JAS, PROOF and GANGA, were shown at the San Diego meeting.

2.9.1 CMS Clarens web service layer client and server developments

The Clarens security mechanism continued to be a major focus of work done during this period, with many refinements to the virtual organization structure in the server, and certificate verification. Support for self-signed certificates was implemented, where no certificate authority (CA) certificate could be relied upon for verification. Clients with self-signed certificates are authenticated by their presence in the virtual organization, i.e. in the authorization stage.

A Java-based client was developed for use as a stand-alone application or as a browser applet. This turned out to be a major undertaking due to the security restrictions placed on browser applets, namely lack of file system access, the lack of built-in cryptographic services and the inability to install cryptographic providers inside the applet framework. The standard cryptographic providers was also found to be immature and overly complex compared to other implementations, even those implemented in lower level languages like C++. These limitations were worked around by using a third-party cryptographic library from <http://www.bouncycastle.org> that did not require the cumbersome Java security provider infrastructure. The lack of filesystem access meant that users' proxy certificate or certificate/key files could not be loaded by the applet for authentication. Instead an HTML form was constructed that could submit the necessary files to a servlet on the Clarens server that could mediate access to the Clarens proxy escrow service. A second web page can then be used to retrieve the proxy or certificate/key files in such a way that the Clarens client applet could access them for use in authentication with the Clarens server.

Server-side support for connections to the Storage Resource Broker (SRB) of the San Diego Supercomputing Center (SDSC) was added in collaboration with the SRB development team. This allows Clarens clients to access their SRB storage space transparently using the same protocol as with other Clarens services, i.e. eliminating the need to install SRB client-side libraries. On the server side connections to SRB is handled with Clarens as a client to SRB. Network connections with SRB are persistent and stateful as opposed to the stateless connection model used by Clarens. Currently every request made by a Clarens client results in a new connection to SRB being opened, the request being processed, and the connection closed again. This results in subpar performance. It was decided that this could be resolved by having a light-weight Clarens server start up in response to an SRB connection request. This process could hold a persistent connection to SRB and therefore provide improved performance to clients.

Clarens was selected to be part of the CMS first data challenge for remote analysis. An architecture of stateful connections to CMS analysis processes was decided upon, similar to that described above. In this architecture, the main Clarens server would start a CMS analysis job using a cluster scheduler, e.g. Condor or OpenPBS. The analysis process would be scripted using the Python scripting layer, which is part of the CMS software. This scripting layer would in turn start up a small Clarens server available only to the original requester through which the analysis would be controlled and data returned to the client.

A first public release of the Clarens server and client software was made available for download at the main web site at <http://clarens.sourceforge.net>. This site was also updated with a new front page layout and logo. More documentation was also added to the site to aid developers of client and server functionality.. Two other projects now make use of the Clarens infrastructure, namely the SOCATS RDBMS analysis project by Eric Aslakson at Caltech, and a project at the National University of Science and Technology (NUST) in Pakistan to provide access to CMS analysis through a PDA (handheld) interface

2.9.2 ATLAS DIAL

My PPDG-related efforts are focused on providing a collective view of event data and a framework for distributed interactive analysis of such data. The data collections are called datasets and are described at <http://www.usatlas.bnl.gov/~dladams/dataset>. The analysis project is called DIAL and is described at <http://www.usatlas.bnl.gov/~dladams/dial>.

I spent 15% of my time on the design and development of datasets and DIAL. The code was ported to gcc 3.2 and Xerces 2.1. The code can still be built under gcc 2.95 but support for Xerces 1.7 has been dropped. It is expected that ATLAS will move to the newer versions of gcc and Xerces the second quarter of this year. LCG is already using these versions.

All of the dataset and DIAL classes can now be imported into the ROOT dictionary using ACLiC. This required some effort on the part of ROOT team. The package dial_root builds the associated libraries and provides a ROOT macro to do the import. This is an important in that it allows ROOT to be used as the interactive interface to DIAL.

I spent many weeks at CERN this quarter. The LCG applications project is advancing rapidly and I took advantage of time there to meet with members of the SEAL and POOL sub-projects. POOL is developing a hierarchical collection model. This is a promising candidate to provide the connection between the serial event collection stream expected by the ATLAS event-processing framework and the composite structure of datasets used for distributed dataprocessing.

2.10 CS-12 Catalogs and Databases

2.10.1 STAR MySQL database

As the STAR collaboration has started to deploy its metadata catalog, authorization became in our minds an important problem to address. When it comes to database however, not much work exists on this topic and gridification (securing the information transferred in between database servers or sites and the login authentication granting database privileges) is not done to our knowledge. The Meta data catalog implementation itself is a good toy for STAR to test what in principle, we would like to achieve as per the Gridification of MySQL. But most important is our analysis habits themselves: some of our code may need to be granted access at run time to a database and authorization is an issue we will soon be facing.

To this end, Richard Casella from the ITD department, started the evaluation and feasibility of the Grid-enable MySQL. We have documented the steps necessary to setup a MySQL version 4.0.0 or later on our [Grid Enabling MySQL](#) pages, a new opened activity in our group. At the end, our goal would be to be able to authenticate to MySQL without additional credentials over an OpenSSL connection after doing a grid-proxi-init. Currently, our initial tests using X509 certificates created as described in the mysql documentation have been successful both from the command-line and from Perl::DBI over an SSL connection. We have also checked that the data replication in MySQL 4.0 and later uses SSL encryption for data transfer (which is required for many application needing either confidentiality data access restriction). Our next step is to investigate the application of the GSI patch to mysql.

We have also participated to the MySQL conference (April 10-12) and took contact with the developers as changes we would make in the course of this work would need to be re-integrated into the development main stream.

3 Single Collaborator Reports

3.1 ANL – Globus

Parts of the Globus work are described above in sections on Monitoring, Reliable File Transfer and Security.

3.1.1 Coordination and Support

NOTE: RLS work is listed in the ISI-Globus report.

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.

3.1.2 Training, Presentations and Papers

GlobusWorld, a week-long training event and conference, was held in January. It consisted of tutorials for GT2 and the newly released GT3 alpha, as well as three tracks of presentations over 4 days. The fifth day consisted of additional tutorials as well as Workshops on DataGrid technologies, Security and Life Sciences Application work. A full agenda and set of talks can be found at http://www.globusworld.com/globusworld_web/jw2_program_tut.htm.

Other papers published this quarter include:

"Using CAS to Manage Role-Based VO Sub-Groups", Shane Canon, Steve Chan, Doug Olson, Laura Pearlman, Craig Tull, and Von Welch, CHEP 2003. <http://chep03.ucsd.edu/files/441.ppt>

"The Community Authorization Service: Status and Future", Ian Foster, Carl Kesselman, Laura Pearlman, Steven Tuecke, and Von Welch, CHEP 2003. <http://chep03.ucsd.edu/files/518.ppt>

"A performance Study of Monitoring and Information Services for Distributed Systems", Xuehai Zhang, Jeffrey L. Freschl, and Jennifer Schopf, to appear in HPDC '03. <http://people.cs.uchicago.edu/~hai/hpdcv25.doc>

3.1.3 Globus Toolkit 2.x updates and bug fixes

This quarter we released numerous bug fixes, and version 2.2.4 can be downloaded at our website. In addition, a number of advisories are available at <http://www-unix.globus.org/toolkit/2.2/advisories/>. We closed 9 bugs listed in Bugzilla (256, 542, 596, 625, 713, 722, 741, 745, 823) and have only 5 open PPDG-related bugs still open in our system (53, 260, 347, 398, 872). Additional information about Bugzilla bugs can be found at <http://bugzilla.globus.org>.

3.1.4 Globus Toolkit 3.0

The Globus Toolkit 3.0 Alpha release was made available for download, most recently Alpha-3. It contains an open source implementation of OGSi, several OGSi-compliant services corresponding to familiar GT2 services, and the ability to create new OGSi-compliant services. For more information, please visit <http://www.globus.org/ogsa/releases/alpha/>.

3.1.5 Grid Architecture

A meeting was held at Argonne between members of Condor team and GriPhyN and PPDG CS researchers to examine Grid planner architecture issues and identify new features of Condor, Condor-G, and DAGman that can be used to implement a testbed to be used to experiment with new approaches. Problems addressed included job flow control, deferring the binding of a job to an execution site until the job actually needs to be placed at the site, and the various strategies and points in the planning process where data transfer can be requested and performed. Work on a follow-on document to the Data Grid Reference Architecture that focuses on the approaches to and mechanisms for late planning is underway.

3.1.6 CS-11 Interactive Data Analysis Tools

Attended teleconference meetings of this group to gather requirements of interactive tools for OGSA services, to provide information on OGSA to ATLAS, and to coordinate the evolving ideas for dataset representation and processing between PPDG and GriPhyN.

3.2 Condor Project

We extended ClassAds to support plug-ins for providing user-defined functions (via shared libraries, scripts, etc.) This is already being used to extend the classad matchmaking process to incorporate external, dynamic data from outside the classad system. Specifically, the Dzero experiment is using a custom classad function to import information from the SAM system on the current location of cached data sets, to assist in matching jobs with resources (see below).

We added grid matchmaking capabilities to Condor-G. This means users are no longer forced to hard-code a specific target site (i.e. gatekeeper) for each grid job at submit time. Instead, each job can simply specify requirements and preferences which, via ClassAd matchmaking, Condor-G uses to decide which Globus site to submit to. Using the new user-defined ClassAd plug-in feature to import dynamic information from SAM, the Dzero experiment can now submit grid jobs which dynamically prefer the grid site with the most already cached data, reducing startup i/o.

We added a number of custom policy control expressions to Condor-G which allow users to define how and when grid jobs should be retried in the event of failures. This allows Condor-G to automatically and robustly recover from many grid job failures in a batch environment without an interactive user having to manually debug or restart them. For example, these expressions can be used to define whether a failed job should be retried immediately, suspended and retried later, re-matched with a new site and retried, etc. Furthermore, these expressions can take into account which site a job has matched with in the past. This formerly manual process was one of the single most time-consuming elements of last fall's US-CMS MOP-based production runs, and its automation should result in a major improvement in the operational efficiency of large grid runs.

We added multi-tier support to Condor-G, allowing users to submit grid jobs from completely stateless machines which host no persistent services (e.g., a laptop, web page, etc.) to one or more centralized servers on which a reliable queue is maintained, and which manage the computation on the user's behalf. This required us to implement a complete job "sandbox", which Condor-G moves around with the job between tiers, and to provide new flexible policy expressions for garbage collection of impartially-submitted jobs, etc. on the reliable queue.

We spent a considerable amount of time hardening the security of the (now multi-tier) system to ensure that all tool/daemon and daemon/daemon connections can be secure and authenticated with Globus GSI security. We also implemented secure forwarding of the user proxy from the submission point to the reliable queue.

We enhanced Condor-G to run under an unprivileged account but still manage a multi-user queue, using GSI-security mechanisms.

We enhanced the communication between Condor-G and the Condor Central Manager (which is used during matchmaking) to support TCP updates rather than only UDP. This solved a problem the Dzero experiment was experiencing whereby UDP network updates were being lost between Fermilab and a site in Asia.

In addition to these bullet-items, for the last few months a portion of the Condor team has been working on the scalability, reliability, and correctness of Condor-G, Globus, and the Condor-G/Globus interaction. This work, partially funded by the PPDG, will benefit all users of Globus and Condor-G. EDG work is seeing immediate benefits. The CMS testbed work can expect to see improvements in throughput and reductions in administrative effort as a result.

During this work the Condor Project identified and found solutions for a number of scalability and correctness issues. Over 40 such issues were logged in the Globus incident tracking system. The vast majority of these have been resolved by the Globus team. In many cases the change made to the Globus system was directly based on changes suggested by the Condor Project. Problems include a number of minor correctness issues that had serious ramifications in systems like Condor-G that use Globus as middleware. (A full list of the incidents reported to Globus by the Condor project over the last few months can be found here:

<http://bugzilla.globus.org/bugzilla/buglist.cgi?emailreporter1=1&emailtype1=substring&email1=cs.wisc.edu>

[u&bugidtype=include&chfield=%5B%5Bbug+creation%5D&chfieldfrom=2002-12-01&chfieldto=Now&cmdtype=doit](http://www.ppdg.net/bugidtype=include&chfield=%5B%5Bbug+creation%5D&chfieldfrom=2002-12-01&chfieldto=Now&cmdtype=doit))

The Condor Project has also developed tools to improve the use of Globus. For example, in a variety of situations state files created on a Globus Gatekeeper machine can be left behind after the job finishes. It is not feasible for Globus itself to handle all of the situations in which this can occur, as they often require decisions by an administrator. The Condor Project developed a tool to that can be used to automate this administration or to simplify the process during human intervention.

We also found scalability problems when a single Globus gatekeeper was acting as a submission front end to large clusters. Any cluster with more than a hundred nodes will likely have problems, and clusters with more than a thousand nodes will likely require that jobs be carefully divided among multiple Gatekeeper machines by users. We spent considerable effort understanding the extent of the problem and seeking solutions. While we made many minor improvements, serious improvements were not feasible in the short term. To solve this, we added new functionality to Condor-G to help minimize use of the non-scalable parts of Globus. This effort involved significant changes to Condor-G. This work heavily exercised Globus' functionality, in the process we uncovered and fixed many additional correctness issues in Globus. As a result of this work, we have seen a forty-fold improvement in run times. A typical test run of 400 jobs that run concurrently for 15 minutes each now takes 1 hour to run on our test site. During this time the gatekeeper is heavily loaded but usable. Prior to our changes such a run took two days, during which time the gatekeeper machine was frequently so heavily loaded as to be unusable.

We also worked to add support for multiple proxy certificates to Globus and Condor-G. This allows a user which desires to use multiple proxies, perhaps as many as one per job, to do so.

3.3 SDSC – SRB

The activities at the San Diego Supercomputer Center in support of the PPDG have focused on the development of the data grid capabilities required by current high-energy physics experiments. Collaborations are now being conducted with both the BaBar experiment and the CMS experiment on use of the Storage Resource Broker technology. In February, a major release of the SRB was provided (version 2.0) and used to test data grid functionality. The release contains multiple capabilities that have been requested by the High –Energy Physics community, including:

- Support for server-initiated parallel I/O, server-to-server parallel I/O, and client directed parallel I/O. When interacting with archives such as HPSS, the number of I/O streams is determined by the number of servers across which HPSS has striped the data. The client must respond to the streams initiated by the server. The SRB parallel I/O has been optimized for interactions with the HPSS mover protocol. It is worth noting that GridFTP is being promoted as the next generation mover protocol for HPSS. For third-party transfer, the parallel I/O streams are supported directly between servers.
- Support for bulk file registration, bulk data load, and bulk data unload. Tests were conducted at SDSC demonstrating bulk file registration at better than 400 files per second.
- Support for direct tape access, and management of a tape robot.
- Support for the Postgres open source database technology.
- Improved Java Gui for administration, and improved installation procedures.

The new functionality was tested with BaBar and CMS.

- Support for the BaBar experiment. SDSC participated in a demonstration of the transfer of data between Stanford SLAC and UCSD for the Conference on High-Energy Physics, which was held at UCSD. Equipment provided by Sun was used to interface to an existing network connection between Stanford and UCSD. Data was moved using SRB parallel I/O at a rate of 18 Mbytes/sec.

- Support for the CMS experiment. Ian Fisk installed an alpha release of version 2.0 of the SRB, and demonstrated data movement between CERN, Fermi lab, and UCSD at rates from 80% to 90% of the available bandwidth. The transfers were done using 4 parallel I/O streams and a window size of 700 kBytes. Higher rates can be achieved by using more I/O streams, and by going to a 4 MB window size.

SDSC is now working on the next set of functional requirements that have been requested. The major requirement is support for peer-to-peer federation of logical name spaces, or MCAT metadata catalogs. A design is being developed in collaboration with the UK data grid and other projects. The first implementation has demonstrated the ability of the SRB to forward operations to the designated MCAT catalog, irrespective of the origin of the request.

The next major requirement has been improved operation with the Globus GSI and GridFTP environment. An effort is underway to integrate GSI version 2.2 into the SRB authentication environment. SDSC has verified that it is possible to interact with both versions of the GSI environment through the SRB. GridFTP drivers and a GridFTP API are being tested both to use the SRB to access a Globus managed resource, and to provide the ability for Globus to access data within a SRB collection.

The final major requirement is the integration of the new access mechanisms with SRB managed data collections. Discussions on the integration possibilities for SRM have been started with Arie Shoshani. An interesting approach is to use the SRB to mediate interactions with storage repositories within the SRM, making it possible for the SRM to manage caching on distributed caches while accessing data from any of the storage repositories supported by the SRB. A related effort is the integration of the CERN tape management capability with the SRB tape management system. This may make it possible to provide alternate interfaces to the CERN collections.

Development of a WSDL interface is progressing rapidly at SDSC, with services available for the access and manipulation of files. The WSDL interface for manipulating and managing metadata is in progress. Both implementations will be upgraded to track the evolution of the OGSA-DAIS web service interfaces promoted by the Global Grid Forum.

4 Appendix

4.1 List of participants

TEAM	Name	F	Current Role CS	1	2	3	4	5	6	7	8	9	10	11	12
Globus/ANL	Ian Foster	Y	Globus Team Lead, GriPhyN PI, iVDGL, GriPhyN						x	x					
	Mike Wilde	N	GriPhyN coordinator					x					x		
	Jenny Schopf	Y	GriPhyN, iVDGL, Globus team liason, ATLAS-CS liason			x				x	x		x		
	William Alcock	Y							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		

SRB/UCSD	Reagan Moore	Y	SRB Team Lead. GriPhyN collaborator					X		X	x						
	Wayne Schroeder	Y	CS-8: Web Services					x			X						
JLAB	William Watson	Y	JLAB Team Lead				x	x	x		x						
	Sandy Philpott	N	facilities									X	X				
	Andy Kowalski	N					X										
	Bryan Hess	Y	Web Services				x				X						
	Ying Chen	Y	Web Services				X		X		x		X				
	Walt Akers	N	Web Services					x					X				
STAR	Jerome Lauret	N	STAR Team Lead		x						X		x				
	Gabrielle Carcassi	Y			x									x			
	Dave Stampf	N				X								X			
	Richard Casela	N				X								X			
	Efratios Efstathiadis	N				X								X			
	Eric Hjort	Y					x	x						x			
	Doug Olson	N					X	X			X		X				
Condor/U.Wisconsin	Miron Livny	Y	PPDG PI, PPDG Coordinator. GriPhyN collaborator	x	x	X	x		x		x						
	Peter Couvares	Y		X	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	Y								x							

- CS-1 Job Description Languages
- CS-2 Job Management and Scheduling
- CS-3 Information Services
- CS-4 Storage Management
- CS-5 Reliable File Transfer
- CS-6 Robust File Replication
- CS-7 Documentation
- CS-8 Evaluations and Research
- CS-9 Authentication and Authorization
- CS-10 Experiment Grids and Applications
- CS-11 Analysis Tools
- CS-12 Catalogs
- CS-13 Troubleshooting and Error Handling

4.2 Meetings

CS-11 Workshop <http://www.ppdg.net/mtgs/20mar03-cs11/>

US CMS+iVDGL+SiteAA VO Management Project Meetings
<http://www.uscms.org/s&c/VO/meeting/meet.html>

GGF Working Groups <http://www.gridforum.org/Meetings/ggf7/default.htm>

IVDGL Operations Planning Meeting <http://www.ivdgl.org/Planning/2003-04-operations/>