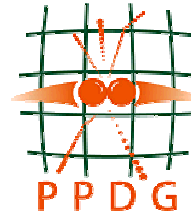


Particle Physics Data Grid: From Fabric to Physics

A Selection of Accomplishments

22 November 2004



1 Introduction.....	1	8 Middleware extended and hardened by PPDG benefits other sciences.....	5
2 ATLAS Data Challenge 2: Grid-based production and analysis.....	1	9 Physics results from the STAR experiment at RHIC benefit from production Grid data services.....	5
3 BaBar DataGrid using Storage Resource Broker.....	2	10 Storage Resource Management (SRM) Standard widely used & adopted.....	6
4 USCMS Data Challenge 04 is Grid Based.....	2	11 References.....	7
5 CDF Distributed Analysis.....	3		
6 DØ Remote Reprocessing.....	3		
7 Grid3.....	4		

1 Introduction

This is a compilation of accomplishments in PPDG, most of which are documented in short individual writeups posted on the News page of the PPDG web site.

2 ATLAS Data Challenge 2: Grid-based production and analysis

The ATLAS experiment is being prepared to take data at the Large Hadron Collider (LHC) at CERN, near Geneva, Switzerland beginning in 2007. A central goal of the LHC program is the discovery and characterization of the Higgs boson, an as yet unseen elementary particle predicted to exist and necessary to the theory describing why particles have mass. The ATLAS collaboration consists of 2000+ scientists around the world with hundreds of scientists participating in the U.S. (USATLAS).

The ATLAS Data Challenge DC2 is a major milestone in the preparation of the ATLAS world wide software environment that includes global distribution of data and access to computing resources in a distributed system. The goals of

DC2 include the operation of the full software environment for ATLAS with multiple, interoperating grids. The results from DCS will provide input for the ATLAS computing model as well as providing input to the overall resource estimates of needs.

The scale of DC2 is based on the simulation of 30 million events, subsequent reconstruction and access by users. This represents an approximate scale of 20% of the size of the data set at the start of data taking in 2007. A significant aspect of DC2 is the merging of events. In a typical event from ATLAS at full beam intensity, there may be as many as forty individual proton-proton collisions occurring. The simulation and reconstruction of these “mixed” events is a major challenge for the computing environment.

The persistent shared grid infrastructure in the U.S., Grid3, is providing about 30% of the global resources used for DC2. The grid-enabling of the ATLAS applications and deployment of the Grid3 environment is a result of USATLAS participation in the Trillium consortium, a cooperative effort of the NSF funded iVDGL and GriPhyN projects and the DOE funded Particle Physics Data Grid project.

3 BaBar DataGrid using Storage Resource Broker

The BaBar experiment at the Stanford Linear Accelerator Center studies electron-positron collisions at asymmetric energies as a way to investigate properties of interactions that produce B mesons. Measurements of the characteristics of how a fundamental symmetry (CP) is violated in these interactions may ultimately lead to an understanding of why the observed universe is dominated by matter over antimatter.

Through participation in the Particle Physics Data Grid, BaBar has been using the Storage Resource Broker (SRB V3) in production for bulk data distribution from SLAC since late last year. Since that time more than 160,000 files corresponding to approximately 60TB of experimental data have been stored in the SRB and distributed to France (CCIN2P3). There have been over 1.5million SRB operations (queries, registration of new files, etc) performed on the metadata catalog (MCAT) during the data taking year by the small number of users (typically 5-6) taking part in the distribution of bulk data. The distributed data has been used extensively for data analysis at CCIN2P3 since May.

SRB V3 has the capability to couple separate metadata catalogs together in a federation (called zoned MCAT) and this is being evaluated and tested for usage in BaBar. It is planned to combine the SRB installations at SLAC and CCIN2P3 in this zoned MCAT mode for the coming run.

This use of SRB is providing a significant improvement to the reliability and level of effort required to distribute BaBar data from SLAC to other centers for data analysis by the scientists.

4 USCMS Data Challenge 04 is Grid Based

The CMS experiment is being prepared to take data at the Large Hadron Collider (LHC) at CERN, near Geneva, Switzerland beginning in 2007. A central goal of the LHC program is the discovery and characterization of the Higgs boson, an as yet unseen elementary particle predicted to exist and necessary to the theory

describing why particles have mass. The CMS collaboration consists of 2000+ scientists around the world with hundreds of scientists participating in the U.S. (USCMS).

The CMS Data Challenge DC04 is one of the milestones of the experiment scoped to ensure the experiment is ready with its global data distribution and analysis systems for the start of data taking. The performance metrics for DC04 were to provide a baseline to give the experiment input to the Physics and Computing Technical Design Reports in the next two years. These design reports will form the baseline to which the production data processing and analysis systems will be built and must perform. It is essential that these global data distribution and analysis systems work effectively in order to enable scientists and their students to be able to participate fully at their home institutions in addition to having available sufficient computing resources to accomplish the scientific mission of the experiment.

The goal of DC04 was to perform at 25% of the throughput needed at the start of data taking in 2007 (5% of the full LHC rate) including distribution of data from CERN to the primary regional centers in each country. Additional goals were to provide the software infrastructure to manage and distribute the data, and to provide an end to end demonstration of event reconstruction and analysis to show the state of readiness of the experiment's software infrastructure. This was to demonstrate the full chain of data processing and analysis to prove the experiment's software system. These goals were substantially met. About ¼ of the globally produced 70M events were generated on the shared Grid3 environment in the U.S. (25 sites, 3000 CPU's) and achieved 50% greater throughput due to opportunistic use of resources than was possible with only those resources dedicated to CMS. These data were sent over the network to CERN for storage and processing. A typical data sample copied back over the network from CERN to Fermilab in Illinois is 5,200 gigabytes in 440,000 files. These processed data are now being analyzed by scientists in the U.S.

USCMS is participating in the Particle Physics Data Grid (PPDG) project, which together with

the NSF funded GriPhyN and iVDGL projects, form the Trillium consortium. Trillium is a collaboration of several computer science and physics experiment groups that are enabling physicists, computer scientists and computational professionals to integrate, harden, deploy and run national and international scale data intensive distributed computing applications and systems. This work enhances and enables both the physical science and computer science goals of the participants, as well as a broader range of scientists benefiting from improved software and the emerging shared persistent grid infrastructure.

5 CDF Distributed Analysis

The CDF experiment at Fermilab is undergoing an aggressive Trigger and DAQ upgrade which will increase the maximum event rate out of the detector from 80Hz to 360Hz by 2006. The rationale for this increase is to be able to maximize physics output at any luminosity by dynamically filling the bandwidth with low threshold triggers.

The most important physics measurement that benefits from the dynamically filled bandwidth is the measurement of Bs mixing. However, a broad range of other physics topics will also benefit.

The increased data volume translates into a doubling of computing needs which had previously not been budgeted for. To satisfy those needs, CDF has embarked on an ambitious global computing program with the goal of having 50% of all of CDF's computing outside of FNAL by Summer 2005.

The CDF situation is somewhat unique among currently running experiments in particle physics in that its computing need is clearly dominated by user analysis computing. Since January 2004 we have put into production 9 computing centers across the world, all of which are accessible by all of the close to 800 CDF physicists registered at the central system at FNAL. No user logins are required. Instead users interact with the distributed system from their desktop as if their jobs were running locally. Authentication, authorization, and accounting are all based on user credentials, and no user accounts need to be maintained at any of the sites.

For the 2004 summer conference season resources outside of FNAL accounted for roughly 25% of CDF's global resource consumption, and six of the nine centers are located outside the US.

The software infrastructure in CDF today is based on the CDF Analysis Farm (CAF) and SAM projects. CAF provides job management, user interfaces, and monitoring & accounting with SAM providing data management. In addition, there are efforts (FNAL & JHU) underway to deploy a squid based distributed database caching system for calibration metadata. The CAF effort is a UCSD/INFN collaboration while SAM in CDF receives effort from FNAL, the UK, Germany, Rutgers, and INFN. Operations support is co-ordinated by TTU.

PPDG has contributed to CDF via Condor, Globus, and SAM, and efforts are under way to understand both technical as well as support issues with regard to adding storage elements with SRM interfaces at each of the sites. At present, the main FNAL as well as the UCSD site are based on Condor, and the remaining sites are expected to transition to Condor within the next 6 months. Condor effort has been provided in three areas so far: scalability & hardening, computing on demand, and security & encryption. In addition, work on policy implementation and "pull model" for job management is ongoing. CDF benefits from Globus in the sense that all data movement between sites is based on gridFTP. CDF benefits from SAM in that all data access outside of FNAL is based on SAM.

Future directions are focused in the US on SAMGrid within the context of the Open Science Grid. However, as 50% of future CDF hardware funding presently appears to be located outside the US, interoperability between and federation of grids will be a primary concern for CDF in the future.

6 DØ Remote Reprocessing

The Fermilab RunII program is the world-leading elementary particle physics program colliding protons and anti-protons, until the LHC program takes over at the end of this decade. The physics program for RunII is testing the limits of the "Standard Model" of particle physics (known to

be incomplete), investigating the nature of the electroweak interaction, and CP symmetry violation. The RunII experiments (DØ and CDF) each have about 700 physicists participating from around the world. One of the significant challenges of the RunII program compared to the previous run is a greatly increased luminosity and data rate from the experiments, leading to a much greater need for computing resources to process and analyze the data.

The strategy adopted by DØ to meet this challenge is to integrate advances in grid computing technology with a distributed computing model so that data analysis and simulations activities can be shared at many computing facilities around the world. A major milestone achieved, in January 2004, on the path to this goal is the successful reprocessing of over 500 million events using a distributed system with integrated grid technology, using six computing facilities in six countries, with 20% of the resources coming from outside of Fermilab. This repeated reconstruction of all its recorded data created a homogeneous dataset for physics analysis based on an up to date understanding of the detector and enabled analyses to show improved results at conferences in March 2004.

As participants in the Particle Physics Data Grid and in collaboration with the Condor team at the University of Wisconsin, matchmaking for grid scheduling was enhanced and integrated with Condor-G as part of the SAMGrid distributed computing system along with additional existing grid technology (DAGMan, Globus Gatekeeper, GridFTP, MDS). This was tested and hardened into a real production system that could be deployed, maintained and monitored. Additional deployments are continuing beyond the six sites in January 2004. The failure rate of the grid job planning and management infrastructure of <1% is very good and the overall end-to-end application job failure rate is tolerable compared to that achieved in a non-grid local cluster environment. The result being that DØ is well on it's way to the expanded and distributed computing resources that enable to most effective scientific results for the RunII program.

7 Grid3

The Particle Physics Data Grid (PPDG) project, in cooperation with the NSF funded iVDGL and GriPhyN projects, and the U.S. CMS and U.S. ATLAS Software and Computing projects have deployed a first shared Grid environment, Grid3, across U.S. Laboratories and Universities to run applications from the participating experiments and computer science groups. The Condor and Globus computer science teams, central participants in the grid projects, provide the core middleware for this grid environment that is packaged and distributed as the VDT.

Grid3 is used by ATLAS and CMS for their "data challenges," in preparation for the experiments' start of data taking in 2007, as well as numerous other applications including: LIGO gravity wave search, SDSS galaxy cluster detection, BTeV proton-antiproton simulation, SnB biomolecular analysis, GADU/Gnare genome analysis, and Computer Science experiments testing resource planning and data movement.

By late 2003, Grid3 consisted of 26 sites providing over 2500 processors. A project team of more than 30 people part time constructed and deployed Grid3 over a span of about 4 months and operations and usage has continued as a production environment for the science teams involved. The iVDGL Grid Operations Center at Indiana University provides a focal point for day-to-day operations support as well as running some central servers for information and monitoring services and virtual organization management.

The ESnet PKI service provides digital identity certificates for the people and the resources participating in Grid3. Virtual Organization Management services are provided using the EDG/DataTAG Virtual Organization Management Software (VOMS) for central registration of user information. Job Submission and Management services are provided through the use of Condor-G using the GRAM protocol. Condor-G manages the submission of the jobs, the Globus Gatekeeper submits the jobs to the local Compute Element batch queue, the Condor GridMonitor manages the load on the local head node, and all components cooperate to manage

retries, report errors and monitor the job submissions. Data Movement services are provided through the use of GridFTP. The MonALISA system was used as the core monitoring system. It was extended for Grid3 to collect and present information based on each VO. This allowed accounting for usage of resources by each VO on each Grid2003 site.

Many participants in Grid3 are founding members of the Open Science Grid Consortium and are planning an evolution of the Grid3 environment (especially the addition of the SRM interface to storage services) to become the first version of Open Science Grid in the spring of 2005.

8 Middleware extended and hardened by PPDG benefits other sciences

A focus of the Particle Physics Data Grid (PPDG) project is to integrate, extend and harden middleware for distributed computing with the end to end applications of several experiments in high energy and nuclear physics. PPDG, together with the NSF funded GriPhyN and iVDGL projects (the Trillium consortium), has adopted the use of the VDT packaging of grid middleware and the enhancements, debugging and bug fixing resulting from this broad deployment across many particle, nuclear and astrophysics experiments is benefiting a much wider science community.

Extensions and robustness improvements in the Condor DAGMAN software, developed through the support of PPDG experiments, is now benefiting the biology community at the University of Wisconsin in their execution of BLAST. They have been able to increase the number of comparisons per run from the millions to over 4 billion. For the CNS/Cyana group successful computational runs have increased from several thousand to over 25000 CPU hours.

Accomplishments for the GADU, the Genome Analysis and Database Update system, benefited directly from several key deliverables of PPDG contributing scaling enhancements, reliability improvements, and feature development to the Grid services on which the GADU system relies:

GRAM, Condor-G, DAGMan, GridFTP, and the Replica Location Service (RLS). From Aug 2003 through Mar 2004 more than 7.5M genome sequences were processed by GADU on Grid2003 resources at a throughput more than 5 times faster than the pre-Grid capabilities of this tool.

Storage Resource Management (SRM) middleware software was initially tested, deployed and improved by PPDG projects, and is now also included in the VDT. This technology was applied to the Earth Science Grid (EDG), and an SRM was adapted to work with a legacy mass storage system at NCAR. It is now deployed in several institutions for use by ESG projects, including ORNL, NERSC, NCAR, and LBNL. It is also installed and being use by a climate scientist at the University of Colorado, who is an ESG collaborator.

9 Physics results from the STAR experiment at RHIC benefit from production Grid data services

The STAR experiment at Brookhaven National Laboratory, Long Island, New York, is one of four experiments at the Relativistic Heavy Ion Collider (RHIC) accelerator whose primary motivation is the observation and characterization of the Quark Gluon Plasma (QGP). The QGP is a form of matter consisting of a hot and dense soup of quarks and gluons thought to have existed briefly when the universe was about one microsecond old, just before a phase transition formed the protons and neutrons that make up atomic nuclei in the universe today. Data taking started in 2000 and by March 2004 a wealth of scientific results have come out including 26 publications in refereed journals.

STAR utilizes primarily two computing facilities, one at Brookhaven Lab, the RHIC Computing Facility (RCF), and another at Berkeley Lab, the National Energy Research Scientific Computing Facility (NERSC). As part of the Particle Physics Data Grid project, STAR has been developing its datagrid connection between these two facilities since 2001 so that it is now routine to have automated transfers of 1,000's of gigabytes of data (10,000's of files), and it's associated

metadata (description of the data) over a day's time. This allows "next day" access to fresh data for analysis and physicists using the facilities at Berkley and Brookhaven are able to collaborate more effectively on the analyses that have led to the recent physics results.

This datagrid connection is implemented using Storage Resource Manager and Globus Toolkit software, developed with support from other DOE middleware projects, as well as the open source database software, MySQL. The data transfer procedures used before the current grid-based implementation were tedious, manpower intensive and error prone. It would take 10 days to transfer 1,000 gigabytes of data, not because of insufficient network bandwidth, but due to a lack of automation and fault tolerance. After transferring a dataset it was common to have an inconsistency in the files of 1% between the two sites, compromising the integrity of the data. Today this level of inconsistency is less than 0.02% (50 times improvement).

In addition to bulk data transfer STAR has interfaced its computational workload job submission tool, the STAR Unified Meta Scheduler (SUMS) with the grid and has been using it to run 1000's of simulation jobs at NERSC and automatically transport and archive the output at the RCF, making the results available to the collaboration for data analysis. This utilizes the Condor middleware from the University of Wisconsin as well as the Globus Toolkit.

10 Storage Resource Management (SRM) Standard widely used & adopted

Storage Resource Managers (SRMs) are middleware components whose function is to provide dynamic space allocation and file management on shared storage components on the Grid. They complement Compute Resource Managers in providing storage allocation and dynamic management of files and storage resources for execution of Grid jobs.

The SRM functional specification effort was initiated by the Scientific Data Management Group (SDM) at LBNL and supported by the

DOE SciADC middleware program. Because the SDM group is also participating to deploy SRMs in two of the National Collaboratory projects, Earth System Grid and especially Particle Physics Data Grid, SRM has developed into an internationally coordinated effort between several DoE laboratories including LBNL, Fermilab and TJNAF, as well as European institutions including CERN and RAL in the UK. This coordinated effort has resulted in the adaptation of the standard specification, and the development of multiple SRM middleware components that inter-operate. This approach is particularly essential for providing Grid access to complex Mass Storage Systems (MSSs). To date, there are multiple MSSs that provide SRM Grid interfaces, including HPSS (developed by LBNL), Enstore (developed at Fermilab), JASMine (developed at TJNAF), and Castor (developed at CERN). In addition, there are several SRM implementations to disk systems, including disk systems on Solaris and Linux platforms (developed at LBNL), dCache (developed at Fermilab), and J-SRM (developed at TJNAF). An SRM was even developed for a legacy MSS system at NCAR, which enabled it to be accessed from the Grid.

SRMs have been used in production by several facilities including BNL, NERSC, Fermilab, CERN, and JTNAF. Some are being used for access to files or for storing files in remote systems, and some are used for intensive data movement between storage systems. For example, The BNL to NERSC setup using SRMs that access HPSS, takes advantage of robustness features of the SRMs. They use the Berkeley-SRMs to move about 10,000 files per month (about 1 GB each) in an automated fashion. The main benefits are great reduction in the error rates, and great savings of human effort. This arrangement, called a DataMover, is also being used by the Earth Systems Grid to move robustly a large volume of simulation production data from NERSC to NCAR, as well as ORNL to NCAR.

The dream of having multiple, diverse, storage system interoperate was fulfilled by this effort, thanks in great part to goodwill and commitment of the international participants. The SRM standard is now widely adopted by various

international efforts, including several HENP experiments, including STAR, ATLAS, and CMS, as well as other efforts, such as NorduGrid. SRMs are also being used by other Grid Collaboratory projects, most notably by the Earth Systems Grid for production use. Although the SRM specification has become a de-facto standard, there is an ongoing effort to standardize this functional specification through the Global Grid Forum (GGF).

11 References

BaBar – <http://www.slac.stanford.edu/BFROOT/>

CDF – <http://www-cdf.fnal.gov>

Condor – <http://www.cs.wisc.edu/condor>

DØ – <http://www-d0.fnal.gov>

Globus - <http://www.globus.org>

Grid3 – <http://www.ivdgl.org/grid3>

GriPhyN - <http://www.griphyn.org>

iVDGL – <http://www.ivdgl.org>

PPDG – <http://www.ppdg.net>

SRB – <http://www.sdsc.edu/DICE/SRB>

SRM – <http://sdm.lbl.gov/srm-wg/>

STAR – <http://www.star.bnl.gov>

USATLAS – <http://www.usatlas.bnl.gov>

USCMS – <http://uscms.fnal.gov>

VDT – <http://www.cs.wisc.edu/vdt>