GRID3: AN APPLICATION GRID LABORATORY FOR SCIENCE

R. Gardner*, University of Chicago, Chicago, IL 60637, USA
L. Grundhoefer, Indiana University, Bloomington, IN 47405, USA
J. Rodriguez, University of Florida, Gainesville, FL 32611,USA

On behalf of the Grid3 Collaboration

Abstract

We describe experiences from nine months of continuous operations of the Grid3 Laboratory, deployed as a result of the Grid2003 Project. The project was organized as an application-driven grid platform that would provide the production-scale services needed by the physics experiments of the Large Hadron Collider (LHC) at CERN (ATLAS and CMS), the Sloan Digital Sky Survey project, the gravitational wave search experiment LIGO, the BTeV experiment at Fermilab, applications in molecular structure analysis and genome analysis, and computer science research projects in such areas as job and data scheduling. The deployed infrastructure has been operating since November 2003 with 30 sites, a peak of over 3000 processors, work loads from 10 different applications exceeding 1300 simultaneous jobs, and data transfers among sites of greater than 2 TB/day. We describe the basic components of the infrastructure and results from its use in LHC data challenges.

INTRODUCTION

Grid3 is a persistent, shared, multi-virtual organization (VO) [1], multi-application grid laboratory capable of providing production level services for large-scale computation- and data-intensive science applications. The project (Grid2003 [2]) was organized by representatives of the U.S. “Trillium” projects (the GriPhyN virtual data research project [3], Particle Physics Data Grid, PPDG [4], International Virtual Data Grid Laboratory, iVDGL [5]) and the U.S. ATLAS and U.S. CMS Software and Computing Projects of the Large Hadron Collider (LHC) [6] program at CERN. The goal of the project was to build an application grid laboratory that would provide:

- a platform for experimental computer science research by GriPhyN and other grid researchers collaborating on iVDGL;
- the infrastructure and services needed to demonstrate LHC production and analysis applications running at scale in a common grid environment;
- the ability to support multiple application groups, including the Sloan Digital Sky Survey (SDSS) [7] and the Laser Interferometer Gravitational Wave Observatory (LIGO) [8, 9], core participants in GriPhyN and iVDGL.

*rwg@uchicago.edu

In the rest of this paper, we present the overarching project requirements, related work, applications, grid design, and results.

PROJECT CONSIDERATIONS

Important considerations were to develop a simple architecture that could link many sites, provide software that could be easily installed, and develop an operations center to be used as a focal point for information gathering and dissemination for all aspects of the project. We refine the overall project goals further as follows.

Architecture: We needed a simple grid architecture that would link execution and storage sites and provide services for monitoring, information publication, and discovery.

Software: We opted for a middleware installation based on the Virtual Data Toolkit (VDT) [10], which provides services from the Globus Toolkit [11], Condor [12], GriPhyN, and PPDG, as well as components from other providers such as the European Data Grid Project (EDG) [13]. VDT allows grid facility administrators to configure their sites easily with simple and well-defined interfaces to existing facility configurations, information service providers, and storage elements. Additional services such as Replica Location Service (RLS) [14], Storage Resource Manager (SRM) [15], and dCache [16], could be provided by individual VOs if desired.

Policy management: Experiment groups should be able to run their applications effectively on non-dedicated resources, including resources not controlled by their VO and/or shared with local users. Automated application installation and publication is important so as to impose minimum requirements on grid facility managers.

Federation: Grid3 is one of several large-scale grids in the U.S., Europe, and Asia, and many of the applications targeted by Grid3 are also designed to run on other grids. Thus, efforts were made to ensure consistency with and “federate” with other grid projects where possible, in particular the LHC Computing Grid Project (LCG) [17].

WORLD OF GRIDS

On this last point, it is worth comparing our approach with the many successful grid projects worldwide that encompass a variety of architectures, deployment approaches, and targeted application domains. For example, European efforts include the aforementioned LCG, the European Data Grid (EDG) and its follow-on

(EGEE) [18], and DataTAG [19], which focused on transatlantic grid testbeds and high performance networks. NorduGrid [20] links computational centers in Scandinavia to deliver production services for high-energy physics applications. Early experiences such as the NSF MetaCenter [21], I-WAY [22], and GUSTO [23], and a number of U.S. grids link modest numbers of high-end systems: e.g., NASA's Information Power Grid [24], the NSF PACI grids [25] and TeraGrid [26].

Grid3 is similar to these projects but differs in several respects. First, it is organized as a consortium among participating stakeholder grid and application software and computing organizations. This structure allows several project objectives to be met simultaneously, and a large scale production environment achieved with the aggregate of resources from the participating groups, while maintaining a development environment for computer science research. Second, the approach taken for construction was aimed to minimize site-specific requirements (e.g., for installation and configuration) while stressing site and VO autonomy. Finally, Grid3 chose a minimum of common services for workload and data management; the VOs provided these components as part of existing (client-side) job schedulers and production frameworks.

GRID DESIGN

With these considerations we adopted a simple two-tier approach, in which each resource (compute, storage, application, site, user) was logically associated with a VO. At each site, a core set of grid middleware services with VO-specific configuration and additions were installed, with registration to a VO-level set of services such as index servers and grid certificate databases. Where appropriate, VO-level services were combined into top-layer services at the iVDGL Grid Operations Center (iGOC), which provided monitoring applications, display clients, and verification tasks and an aggregate view of the collective Grid3 resource and performance. Six VOs (U.S. ATLAS, U.S. CMS, SDSS, LIGO, BTeV, iVDGL) were configured. Appropriate policies were implemented at each local batch scheduler (OpenPBS, Condor, and LSF) and Unix group accounts were established at each site for each VO.

Site Installation Procedures

Procedures for installation, configuration, post-installation testing, and certification of the basic middleware services were devised and documented. The Pacman [27] packaging and configuration tool was used extensively to facilitate the process. A Pacman package encoded the basic VDT-based Grid3 installation, which included:

- The Globus Toolkit's Grid security infrastructure (GSI), GRAM, and GridFTP services;
- Information service based on MDS, with registration scripts to VO-specific information index servers and VO-specific information providers;
- Cluster monitoring services based on Ganglia [28], with provisions for hierarchical grid views; and
- Server and client software for the MonALISA [29] agent-based monitoring framework.

Conventions were documented to provide grid facility administrators and operators with uniform instructions with the goal of obtaining a consistent Grid3 environment over the heterogeneous sites. In particular, information providers were developed for site configuration parameters such as application installation areas, temporary working directories, storage element locations, and VDT software installation locations. Only a few extensions to the GLUE [30] MDS schema were required.

Monitoring and Information Services

The software installed on Grid3 sites included components necessary to monitor the overall behavior and performance of the grid and its applications. Several packages sensed monitoring data and made it available to a distributed framework of services and client tools. The set of information providers deployed was determined by identifying and prioritizing desirable grid-level (such as overall resource availability and consumption) and VO-level (e.g., aggregate CPU usage) performance indicators. Other requirements derived from auditing, scheduling and debugging considerations.

The framework was built by integrating existing monitoring software tools into a simple architecture. Figure 1 shows the components of the framework. Producers provide monitored information, consumers use this information, and intermediaries have both roles, sometimes providing aggregation or filtering functions.

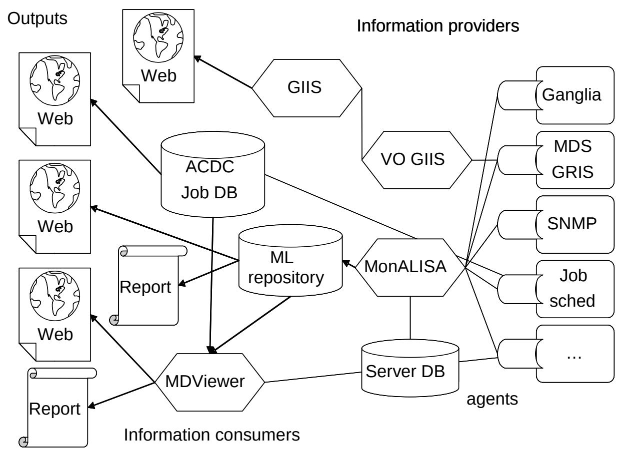

Figure 1: Grid3 monitoring architecture showing information providers and consumers, and the data flows between them.

Some monitoring components are located on Grid3 sites, some in central servers, and some are the clients of the users accessing the information. An aggregated data summary is available centrally, while more detailed data and streams of updates are available from the sites. The main components of the monitoring framework are: The Globus Toolkit's *Monitoring and Discovery Service* (MDS) [31] is used to maintain site configuration and monitoring information; *Ganglia* is used to collect cluster monitoring information such as CPU and network load and memory and disk usage; *MonALISA* {, 2004 #115} which provides access to monitoring data provided by a variety of information providers, including agents which monitored the GRAM logfiles, job queues, and Ganglia metrics; the *ACDC Job Monitor* collects information from local job managers using a typical pull-based model; the *Site Status Catalog* {, 2004 #117} periodically tests all sites and stores some critical information centrally; and the *Metrics Data Viewer* (MDViewer) [34] allows for the analysis and display of collected metrics information.

Virtual Organization Management

To simplify user access to Grid3 resources and reduce the burden on grid facility administrators, we deployed EDG's Virtual Organization Management System (VOMS) [35]. We also used group accounts at sites, with a naming convention for each VO. We generated the local grid-map files that map user identities presented in X509 certificates to local accounts by calling an EDG script to contact each VO's VOMS server.

Support and Operations

The deployment and operation of the Grid3 environment required a number of centralized support activities. The iGOC hosted centralized services, including the Pacman cache, the top-level MDS index server, the Site Status Catalog, the MonALISA central repositories, and web services for Ganglia. A simple trouble ticket system was used intermittently during the project. An acceptable use policy modeled after that used by the LCG was adopted.

APPLICATIONS AND PERFORMANCE

We describe specific application milestones, in particular the LHC data challenges.

ATLAS Data Challenge

The data challenges for both ATLAS and CMS have been designed to prepare the experiments for global data production and analysis for the start of data taking at the Large Hadron Collider at CERN in 2007. ATLAS DC2 was launched in June 2004, and consisted of three phases: Phase I, large scale production of physics datasets (order 10M events) with a chain of production steps including Pythia event generation, GEANT4-based simulation, digitization and pileup. The produced datasets were stored on Tier1 centers (Brookhaven lab for Grid3) from where they were to be streamed back to the Tier0 at CERN. Phase II focused on Tier0 functionality testing with the aim of reconstructing the events at 1/10 of the full scale operation, with datasets streamed down to the Tier1 centers. Phase III, to be exercised until December 2004, focused on distributed analysis of the events reconstructed and access to event and non-event data from anywhere in the world both in organized and chaotic ways. During Phase I, almost 45 million CPU days were used to execute over 100K jobs, that produced nearly 8 million events and about 30 TB of data. The three Grids produced roughly equal amounts of work. The ATLAS production system was designed to separate the abstract ATLAS job definitions and validation steps (using the Windmill supervisor and Don Quixiote data management interface) from the specifics of the particular Grid which were provided by three "executors".

The Grid3 executor system, Capone, communicates with the supervisor and handles all the interactions with Grid3 resources and services. Job requests from the supervisor are taken by Capone that interfaces to a number of middleware services such as the Chimera and Pegasus virtual data tools [36-38] of the the Virtual Data Toolkit (VDT).

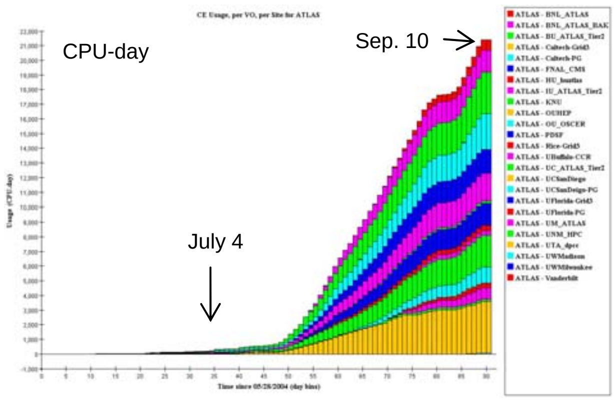

Figure 2: Integrated CPU usage (CPU-day) during the ramp up phase of ATLAS DC2 on Grid3. The plot shows the usage distribution by compute site, approximately 22K CPU-days in aggregate by September.

We observed a failure rate of approximately 30%, where failures are defined as jobs experiencing errors in any processing step that prevented perfect completion (pre-stage, job execution producing the output files, post-stage to the final storage element at BNL, and registration to RLS). Approximately 90% of failures were due to site problems: disk filling errors, gatekeeper overloading, or network interruptions. Operational difficulties included the usual problems associated with large, integrated systems. Symptoms in one component often led to discovery of failures within another. Better tools for diagnosing end-to-end grid applications are needed.

Over a time period of roughly two months, more than 60 CPU years were consumed producing DC2 events. This far exceeded the number of resources available solely from CPUs dedicated to ATLAS users, with less than half of the production performed using dedicated resources.

CMS Data Challenge

The CMS Data Challenge DC04 had the full attention of the software and computing program from November 2003 through its completion on May 1 2004. 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.

The CMS Collaboration was able to use Grid3 resources when they came online in October/November 2003 to produce events for their 2004 data challenge. Fifty million events with minimum bias pile-up at a beam luminosity of $2x10^{33}$ were needed in the final sample. CMS detector simulation consists of 3 steps: (1) event generation with Pythia, (2) event simulation with a GEANT-based simulation application, and finally (3) reconstruction and digitization with the additional pile-up events. The sample of simulated events was accumulated at CERN for primary reconstruction, and distributed in real time to Tier1 and Tier2 centers (some being Grid3 sites) for calibration and toy analysis. The software suite includes MCRunJob [39], a CMS tool for workflow configuration, and MOP [40], a CMS DAG writer, which were first grid-enabled during a previous "big n-tuple" production during the fall of 2002. CMS Production jobs are specified by reading input parameters from a control database and converting them to DAGs suitable for submission to Condor-G/DAGMan. All datasets produced were archived through a Storage Element at the Tier1 facility at Fermi National Accelerator Laboratory (Fermilab).

The production was achieved using compute resources on 11 Grid3 sites to simulate more than 14 million GEANT4 full detector simulation events. Figure 3 shows usage over a four month period beginning November 2003, and Figure 4 show the events delivered per day by the system. Efficiency on Grid3 resources is roughly as high as on the original U.S. CMS production grid, once sites are fully validated. The official OSCAR production jobs are long (some more than 30 hours) and not all sites have been able to accommodate running them. The effort required to run the application has been about 2 FTEs, split between the application administrator and site operations support.

Approximately 70% of CMSIM and OSCAR jobs completed successfully, which is consistent with US-ATLAS performance. Jobs often failed due to site configuration problems, or in groups from site service failures. We saw few random job losses: more frequently a disk would fill up or a service would fail and all jobs submitted to a site would die. Service level monitoring needs to be improved and some services probably need to be replaced. For example, storage reservation (e.g., as provided by SRM) would have prevented various storage-related service failures.

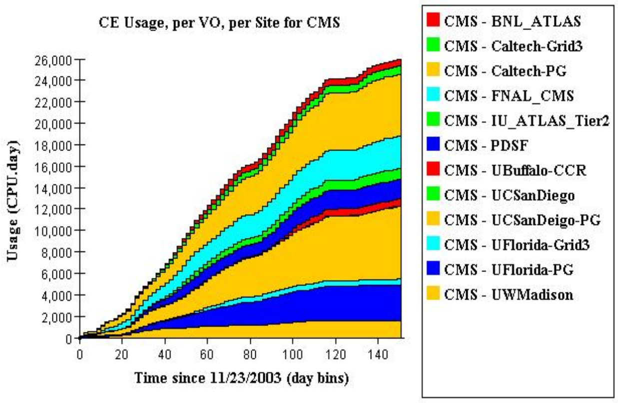

Figure 3: US CMS cumulative use of Grid2003. The chart plots the distribution of usage (in CPU-days) by site in Grid2003 over a 150 day period beginning in November 2003.

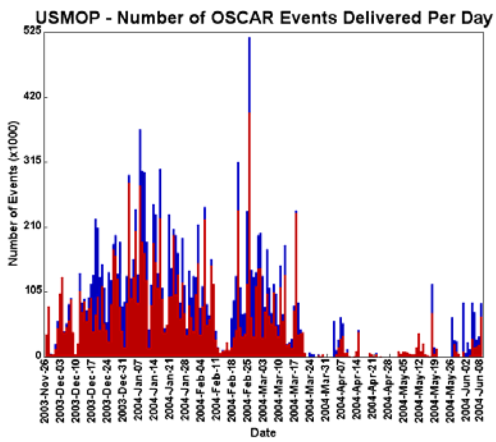

Figure 4: DC04 events delivered per day by the USCMS MOP system executing on Grid3 from November 2003 to June 2004.

Cluster finding in SDSS

SDSS has executed several challenge problems on the Grid3 platform over the past year in which have been used to advance development of their distributed cluster finding tools and computations. A search for galaxy clusters in SDSS data resulted in workflows with several thousand processing steps organized by Chimera virtual data tools. A second application involved a pixel-level analysis of astronomical data, such as analysis of cutouts of images about galaxies with the aim of adding more information to existing catalogs. Other applications included a search for near earth asteroids, which calls for examining complete SDSS images in search of highly elongated objects.

Blind Gravitational Wave Searches

The LIGO experiment has used Grid3 to perform extensive, all-sky, blind searches for continuous wave (pulsar) signals in the LIGO "S2" data set. Each search required that a conventional binary short Fourier transform data file be accessible containing the frequency band that the target signal spans during the observation time. Additional data files containing the ephemeris data for the year are staged from LIGO facilities to Grid3 sites using GridFTP. The location of the staged data (on average 4 GB per job) is published in RLS so that its location is available to the job. The last job in the workflow stages the output results back to the LIGO facility and updates database entries. Each workflow instance runs for several hours on an average processor. The GriPhyN-LIGO working group developed the necessary infrastructure using Chimera and Pegasus to generate and execute the workflows.

Computational Chemistry and Biology

SnB [41, 42], a computer program based on the *Shake-and-Bake* method for chemical structure analysis, is used by laboratories worldwide in local cluster settings. The "*SnB*" program uses a dual-space direct-methods procedure for determining crystal structures from X-ray diffraction data. This program has been used in a routine fashion to solve difficult atomic resolution structures, containing as many as 1000 unique non-Hydrogen atoms, which could not be solved by traditional reciprocal-space routines. It has been successfully ported to Grid3 and executed on a large number of sites, providing en successful runs at much larger scales than previously attempted.

GADU [43] is a Genome Analysis and Databases Update Tool developed by the Mathematics and Computer Science division at Argonne National Laboratory, used to perform a variety of analyses of genome data. The program is designed to address the first and most crucial step in genome analysis which is the assignment of function to genes. The efficiency and accuracy of such predictions is achieved by the use of a variety of bioinformatics tools and approaches (e.g. analysis of global similarities, domain and motif analysis, analysis of the relevant structural and functional information). This process can be extremely tedious, time-consuming, and prone to human error if it were to be done by manually scheduled computations. GADU is an automated, high-performance, scalable computational pipeline for the data acquisition and analysis of sequenced genomes that allows efficient automation of the major steps of genome analysis: data acquisition and analysis by variety of tools and algorithms, as well as result storage and annotation. The group successfully integrated the GADU analysis modules to a grid backend using Chimera virtual data tools, which provided access to the resources of Grid3.

Computer Science Challenge Problems

Computer science groups worked with experiment developers to provide the application middleware (e.g., Chimera and Pegasus, Globus client libraries, Condor-G, RLS) required by grid-based application frameworks. Various computer science groups also used Grid3 as a vehicle for research studies. In addition, the following three demonstrators were provided.

A data transfer study was performed to evaluate whether we could perform large-scale reliable data transfers between Grid3 sites. A Java-based plug-in environment (Entrada) was used to generate simulated traffic between a matrix of sites in a periodic fashion [44].

NetLogger-instrumented GridFTP was used to monitor the Globus Toolkit GridFTP server and [45] URL copy program. NetLogger events were generated at program start, end, and on errors (the default) and for all significant I/O requests (by request).

An exerciser backfill application provided by the Condor group tested the status of the batch systems and operation characteristics of each Grid3 site. This application ran repeatedly with a low priority at 15 minute intervals.

RESULTS

An important strategic goal for Grid3 was to provide the infrastructure and services needed to demonstrate LHC production and analysis capabilities at scale in a common, shared grid environment. A full summary of lessons learned and metrics analysis is available in Ref [2]. As it turned out, US CMS learned many lessons during the first six months of Grid3 which carried over for the ATLAS DC2 production during the latter months. Figure 5 shows the CPU usage during the transitional period.

SUMMARY AND OUTLOOK

We have discussed the deployment and use of a persistent, shared, multi-virtual organization, multi-application grid. The infrastructure remains in place and is currently evolving within the OSG Consortium framework with the goal of providing fully functional, production quality grid that supports Peta-scale operations.

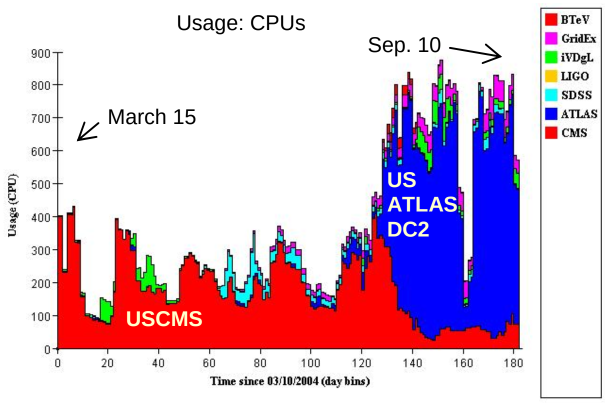

Figure 5: Daily CPUs used, organized by VO, during the six month period ending in September 2004.

REFERENCES

[1] I. Foster, Kesselman, C. and Tuecke, S., "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *International Journal of Supercomputer Applications*, vol. 15 (3), pp. 200-222, 2001.

[2] *The Grid2003 Project*, http://www.ivdgl.org/grid2003/.

[3] P. Avery, I. Foster, *Towards Petascale Virtual Data Grids (GriPhyN Project)*, http://www.griphyn.org/.

[4] *Particle Physics Data Grid*, http://www.ppdg.org/.

[5] P. Avery , I. Foster, R. Gardner, H. Newman, A. Szalay, "An International Virtual-Data Grid Laboratory for Data Intensive Science," *Technical Report: GriPhyN-2001-2*, 2001, http://www.griphyn.org/.

[6] *The Large Hadron Collider Project at CERN*, http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/.

[7] *The Sloan Digital Sky Survey Project (SDSS)*, http://www.sdss.org/sdss.html.

[8] *Laser Interferometer Gravitational Wave Observatory*, http://www.ligo.caltech.edu/.

[9] B.C. Barish and R. Weiss, "LIGO and the Detection of Gravitational Waves," *Physics Today*, vol. 52, pp. 44, 1999.

[10] *The Virtual Data Toolkit (VDT)*, http://www.lsc-group.phys.uwm.edu/vdt/.

[11] I. Foster, Kesselman, C., "Globus: A Metacomputing Infrastructure Toolkit," *International Journal of Supercomputer Applications*, vol. 11(2), pp. 115-129, 1998.

[12] M.J. Litzkow, Livny, M. and Mutka, M.W., "Condor - A Hunter of Idle Workstations," *8th International Conference on Distributed Computing Systems*, pp. 104-111, 1988.

[13] *The European Data Grid Project (EDG)*, http://eu-datagrid.web.cern.ch/eu-datagrid/.

[14] A. Chervenak, E. Deelman, et al., "Giggle: A Framework for Constructing Scalable Replica Location Services," presented at SC'02: High Performance Networking and Computing., 2002.

[15] A. Shoshani, Sim, A. and Gu, J., "Storage Resource Managers: Essential Components for the Grid," in *Resource Management for Grid Computing*, J. Nabrzyski, Schopf, J. and Weglarz, J., Ed., 2003.

[16] *The dCache Project*, http://www.dcache.org/.

[17] *The LHC Computing Grid Project (LCG)*, http://lcg.web.cern.ch/LCG/.

[18] *EGEE: Enabling Grids for E-Science in Europe*, http://public.eu-egee.org/.

[19] *Research and Technological Development for a Data TransAtlantic Grid*, http://datatag.web.cern.ch/datatag/project.html.

[20] *NorduGrid: Nordic Testbed for Wide Area Computing and Data Handling*, http://www.nordugrid.org/.

[21] L. Smarr C. Catlett, "Metacomputing," *Communications of the ACM*, vol. 35, pp. 44-52, 1992.

[22] T. DeFanti, Foster, I., Papka, M., Stevens, R. and Kuhfuss, T., "Overview of the I-WAY: Wide Area Visual Supercomputing," *International Journal of Supercomputer Applications*, vol. 10 (2), pp. 123-130, 1996.

[23] S. Brunett, Czajkowski, K., Fitzgerald, S., Foster, I., Johnson, A., Kesselman, C., Leigh, J. and Tuecke, S., "Application Experiences with the Globus Toolkit," presented at 7th IEEE International Symposium on High Performance Distributed Computing, 1998.

[24] W.E. Johnston, Gannon, D. and Nitzberg, B., "Grids as Production Computing Environments: The Engineering Aspects of NASA's Information Power Grid," presented at In 8th IEEE International Symposium on High Performance Distributed Computing, 1999.

[25] R. Stevens, Woodward, P., DeFanti, T. and Catlett, C., "From the I-WAY to the National Technology Grid," *Communications of the ACM*, vol. 40 (11), pp. 50-61, 1997.

[26] C. Catlett, *The TeraGrid: A Primer*, http://www.teragrid.org/.

[27] S. Youssef, *Pacman, a Package Manger*, http://physics.bu.edu/~youssef/pacman/.

[28] Mason J. Katz Federico D. Sacerdoti, Matthew L. Massie, David E Culler, "Wide Area Cluster Monitoring with Ganglia," presented at IEEE Cluster 2003 Conference, Hong Kong, 2003.

[29] I.C. Legrand H.B. Newman, P.Galvez, R. Voicu, C. Cirstoiu, "MonALISA: A Distributed Monitoring Service Architecture," presented at CHEP 2003, La Jola, California, 2003.

[30] DataTAG and iVDGL Interoperability Working Group, *Grid Laboratory Uniform Environment*, http://grid.infn.it/datatag/wp4/doc/glue-v0.1.2.pdf.

[31] K. Czajkowski, Fitzgerald, S., Foster, I. and Kesselman, C., "Grid Information Services for Distributed Resource Sharing," *10th IEEE International Symposium on High Performance Distributed Computing*, pp. 181-184, 2001.

[32] *Grid3 MonALISA frontend, 2004*, http://gocmon.uits.iupui.edu:8080/index.html.

[33] *Grid2003 site catalog, 2004*, http://www.ivdgl.org/grid2003/catalog.

[34] M. Mambelli and D. Bury, *MDViewer: a Metrics Data Viewer of the Grid*, http://grid.uchicago.edu/metrics/.

[35] EU DataGrid Java Security Working Group, *VOMS Architecture v1.1*, http://grid-auth.infn.it/docs/VOMS-v1_1.pdf.

[36] I. Foster, J. Voeckler, et al., "Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation," presented at 14th Intl. Conf. on Scientific and Statistical Database Management, Edinburgh, Scotland., 2002.

[37] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, S. Patil, M.-H. Su, K. Vahi, and M. Livny, "Pegasus : Mapping Scientific Workflows onto the Grid," presented at 2nd EUROPEAN ACROSS GRIDS CONFERENCE, Nicosia, Cyprus, 2004.

[38] James Blythe Ewa Deelman, Yolanda Gil, Carl Kesselman., "Pegasus: Planning for Execution in Grids," *GriPhyN Technical Reports*, vol. 2002-20, 2002.

[39] Dave Evans Gregory E. Graham, Iain Bertram, "McRunjob: A High Energy Physics Workflow Planner for Grid Production Processing," presented at CHEP 2003, La Jolla, California, 2003.

[40] *The MOP Project*, http://www.uscms.org/s&c/MOP.

[41] *The SnB Program*, http://www.hwi.buffalo.edu/SnB/.

[42] G.T. DeTitta C.M. Weeks, R. Miller, & H.A. Hauptman, "Applications of the minimal principle to peptide structures," *Acta Cryst.*, vol. D49, pp. 179-181, 1993.

[43] D. Sulakhe A. Rodriguez, E. Marland, V. Nefedova, G. X. Yu, and N. Maltsev, "GADU - Genome Analysis and Database Update Pipeline," *Preprint ANL/MCS-P1029-0203*, 2003.

[44] Scott Gose, *Entrada, a Lightweight Application Hosting Environment*, http://www-unix.mcs.anl.gov/~gose/entrada/.

[45] K. Jackson, "pyGlobus: A Python Interface to the Globus Toolkit," *Concurrency and Computation: Practice and Experience*, vol. 14, pp. 1075-1083, 2002.