DOSAR: A DISTRIBUTED ORGANIZATION FOR SCIENTIFIC AND ACADEMIC RESEARCH

J. Snow, Langston U., Langston OK 73050, USA
Z. Greenwood, C. Leangsuksun, J. Steele, Louisiana Tech U., Ruston, LA 71272, USA
E. Gregores, P. Mercadante, S. Novaes, S. Lietti, SPRACE, Sao Paulo, BRAZIL
N. Mondal, N. Panyam, TIFR, Mumbai, INDIA
M. Joy, B. Quinn, U. of Mississippi, Oxford, MS 38677, USA
B. Abbott, K. Arunachalam, P. Gutierrez, H. Severini, P. Skubic, M. Strauss, U. of Oklahoma, Norman, OK 73019, USA
D. Jenkins, H. Kim, P. McGuigan, D. Meyer, J. Smith, M. Sosebee, J. Yu, U. of Texas at Arlington, Arlington, TX 76019, USA

Abstract

Hadron Collider experiments in progress at Fermilab's Tevatron and under construction at the Large Hadron Collider (LHC) at CERN will record many petabytes of data in pursuing the goals of understanding nature and searching for the origin of mass. Computing resources required to analyze these data far exceed the capabilities of any one institution. The computing grid has long been recognized as a solution to this and other problems. The success of the grid solution will crucially depend on having high-speed network connections, the ability to use general-purpose computer facilities, and the existence of robust software tools. A consortium of universities in the US, Brazil, Mexico and India are developing a fully realized grid that will test this technology. These institutions are members of the DØ experiment at the Tevatron and the ATLAS or CMS experiments at the LHC, and form the Distributed Organization for Scientific and Academic Research (DOSAR). DOSAR is a federated grid organization encompassing numerous institutional grids. While founded for HEP research DOSAR forms the nucleus of grid infrastructure organization on the constituent campuses. DOSAR's strategy is to promote multi-disciplinary use of grids on campus and among the institutions involved in the consortium. DOSAR enables researchers and educators at the federated institutions to access grid resources outside the HEP context and is a catalyst in establishing state-wide grid structures. DOSAR is an operational grid which is a Virtual Organization (VO) in the Open Science Grid (OSG). In this talk, we will describe the architecture of the DOSAR VO, the use and functionality of the grid, and the experience of operating the grid for simulation, reprocessing and analysis of data from the DØ experiment. A software system for large-scale grid processing will be described.

INTRODUCTION

Particle physicists employ high energy particle accelerators and complex detectors to probe ever smaller distance scales in an attempt to understand the nature and origin of the universe. The Tevatron proton-anti proton collider located in the Department of Energy's Fermi National Accelerator Laboratory [1], Batavia, Illinois, currently operates at the "energy frontier" and has an opportunity of providing new discoveries through the DØ [2] and CDF [3] high energy physics (HEP) experiments. Future experiments, such as ATLAS [4] at the Large Hadron Collider (LHC) [5] in the European Organization for Nuclear Research (CERN) [6] and a proposed electron-positron linear collider [7] have the prospects of either building on discoveries made at Fermilab, or making the discoveries themselves if nature has placed these new processes beyond the energy reach of the Tevatron. The TeV energies accessed by these experiments coincide with the predicted scale of *electroweak symmetry breaking* (EWSB) [8], lending weight to the belief that new breakthroughs are just around the corner. This EWSB process, although yet to be fully understood, is needed to explain the observation that weak vector bosons are very heavy while photons are massless. Elucidating the mechanism for EWSB stands as one of the top priorities in elementary particle physics.

Current theoretical models of EWSB, including the Higgs mechanism, various super-symmetry schemes, and the presence of extra spatial dimensions share the common feature of predicting the existence of massive new particles [8]. These objects typically decay into complex finals states and are very rarely produced. Selecting these few complicated interesting events from the enormous volume of data requires a detailed understanding of the detector and physics that can only be quantified by the development of sophisticated analysis algorithms and generation of large sets of detailed Monte Carlo (MC) simulations. Even more exciting would be unexpected discoveries, perhaps connected to problems in cosmology such as dark matter and dark energy [9]. Given the small production probabilities of these interesting events, expediting searches for such processes and other critical analyses will require fast, efficient, and transparent delivery of large data sets, together with an

efficient resource management and software distribution system.

In order to pursue full utilization of grid concepts, we are establishing an operational regional grid called *DOSAR-Grid* using all available resources, including personal desktop computers and large dedicated computer clusters, in six DOSAR universities. We will initially construct and operate DOSAR-Grid utilizing a framework called *SAM-Grid* [10] being developed at Fermilab. DOSAR-Grid will subsequently be made interoperable under other grid frameworks such as LCG [11], Teragrid [12], Grid3 [13] and Open Science Grid (OSG) [14]. Wherever possible, we will exploit existing software and technological advances made by these other grid projects. We plan to develop and implement tools to support easy and efficient user access to the grid and to ensure its robustness. Tools to transfer binary executables and libraries efficiently across the grid for environment-independent execution will be developed. DOSAR will implement the Grid for critical physics data analyses, while at the same time subjecting grid computing concepts to a stress test to its true conceptual level, down to personal desktop computers. Many of the proponents of this project are members of the LHC ATLAS [4] experiment and a future experiment under study by the American Linear Collider Physics Group [15], and will apply the experience gained from DOSAR to these projects.

CURRENT ENVIRONMENT

We outline here the current environment that this project will be building upon. We discuss various DØ software packages, the DØ Remote Analysis Model, the virtual organization, and current simulation efforts.

SAM. The data access system for DØ offline analyses is managed by a database and cataloging system called Sequential Data Access via Metadata, or *SAM* [16]. SAM is used in conjunction with the combined software and hardware systems, known as *Enstore* [17, 18], that provide actual access to the data stored in the tape robot. It is designed to accommodate a distributed computing environment, and the combination of Enstore and SAM is well suited to the grid. SAM provides high-level collective services for reliable data storage and replication through efficient cataloging, recording, and retrieval of data. In service for the past four years, SAM's strengths and weaknesses have been thoroughly assessed; opportunities to improve upon its features and stability through the actual use of the system in data analyses have been identified. The system also allows for offsite remote access to the data, a critical functionality for a successful computing grid. Localized offsite data access features also have been in use through the past three years, primarily in MC production.

DOSAR has been very effective in rapidly mobilizing significant additional computing resources into the production of MC events and data reprocessing for DØ, helping to meet a critical need of the experiment. The DOSAR institutions have been operating MC production facilities ("farms"). These farms are localized computing clusters that share a common storage and user space with external Internet connections. Some of the farms are comprised of dedicated rack systems while others use linked personal desktop computers [19]. These farms operate as independent entities that communicate to the central site (Fermilab) via SAM to transfer MC requests and output data.

MC Production and Management. The *MC_Runjob* package [20] provides a low-level MC manager that coordinates data files used by and produced with the executables of the MC chain, via scripts produced for the job. This package is also brought to the execution site from SAM storage by SAM-Grid at the time the job is run. This grid based model results highly efficient MC production. SAM-Grid provides high-level services such as bookkeeping, job parallelization, data retrieval and storage at the Fermilab central facility, interfacing to the specific cluster batch system, and farm performance and request status monitoring information. The system is capable of obtaining an official request from the central site, breaking the request into small jobs to better utilize computing resources, communicating with SAM to acquire any necessary input files from Fermilab, submitting the jobs to the local batch system, collecting and storing the output data and metadata at Fermilab via SAM, and monitoring job progress and farm status. The high degree of automation allows efficient management and oversight of an operating farm.

MC Production Monitoring. DOSAR utilizes web accessible applications to monitor the farm clusters in a graphical manner. The Ganglia Cluster Toolkit [21] is used to provide resource monitoring of all the DOSAR clusters. *MonALISA*, developed at Caltech, provides detailed information about the DOSAR Virutal Organization (VO) that is registered with OSG. DOSAR activity may be monitored through the OSG Repository site. Since its inception, DOSAR farms have produced and stored millions events of Monte Carlo and reprocessed data, a significant contribution to the DØ total production.

PanDA. DOSAR institutions have contributed to the development of the ATLAS Production and Distributed Analysis (PanDA) software system. PanDA is expected to be extensively used for the next ATLAS data challenge that will simulate production rates after data taking begins. PanDA utilizes the full capabilities of the grid computing environment. For example a supervisor running on a machine at OU accepted jobs from CERN and distributed them to various grid executors such as the one for OSG. PanDA is described in further detail in other talks at this conference.

FUTURE PLANS

We intend to integrate the resources in DOSAR institutions into an operating grid to be utilized in critical data analyses at the DØ and LHC experiments, taking grid computing one step further to the realization of the conceptual grid. We will implement a system that aims at robustness and portability. We will develop and implement monitoring capabilities and diagnostics for this system. We will develop smart binary transport tools to dramatically increase the efficiency of the grid and interfaces to simplify use of the grid. These activities will also provide system services that complement and improve existing functionality. The DOSAR Grid will use all locally available computing resources, including personal desktop computers and large dedicated computer clusters. It will leverage software advancement by other grid projects, and be applied to cutting-edge high energy physics research.

Robustness and Stability

A major initiative DOSAR is to study and implement ways to provide greater stability and robustness in operations using the grid. System failures are unavoidable in the information technology field, and their elimination has been the subject of much work. If not resolved in a timely fashion, failures can often result in unavailability of service, thereby significantly reducing productivity. We intend to address such unavailability issues while maintaining a high-performance computing environment. The ultimate goal to strive for is zero downtime.

A common weakness in many cluster designs is the existence of sources of single-point failures. For example, a single-point failure can occur as a result of the remote site gatekeeper malfunctioning. If this unit is lost, the site is no longer available for job distribution.

The chance of single-point failures can be greatly reduced by including multiple-head nodes in each site. For DØ, this entails augmenting SAM stations with a second node at each site that will transparently assume station functions in case of the loss of the primary node. Similarly the present SAM-Grid design will be extended to include a back-up head node that will automatically assume control in case of loss of the primary head node.

In both examples described above, reliability is improved without changing or interfering with grid resource management. Performance can be sustained in the entire grid by providing each local site with transparently self-healing and redundancy mechanisms, e.g., the implementation of multiple head gatekeepers and other important services. Much research in this area that has been performed on non-grid systems [22] can be applied in a similar fashion at DOSAR. A particularly attractive idea is that of Leangsuksun and his collaborators on *HA-OSCAR* (High Availability-Open Source Cluster Application Resources) [23]. A dual-head architecture has been created that includes reliable communications with a heartbeat mechanism to maintain cluster health. Periodic transmission of heartbeat messages traverse across a dedicated serial link, as well as an IP channel, and monitor the health of the working server. When a failure of the primary server occurs, the system can automatically transfer the control of services to the failover server, allowing services to remain available with minimal interruption.

The HA-OSCAR self-healing architecture operates as follows: Each head node runs monitoring, self-healing, and System-Imager daemons. These daemons work together forming adaptive self-healing mechanisms that detect undesirable events and attempt to shield and recover potential site outages, especially at the head nodes. There are three types of daemons monitoring system health, services, and resources, respectively. The health daemon detects hardware, OS and network outages, and performs recovery with a failover server. Several critical services such as grid gatekeeper, GRIS services, local schedulers such as PBS and MAUI, as well as NFS and DHCP, are monitored by the HA-OSCAR service monitoring daemon. Once a service failure occurs, a corresponding service recovery procedure is triggered, smoothly taking over the operations.

Data Mangangement and Proxy Utilities

The DØ experiment uses very large programs for data processing and analyses; for example, the experiment's software libraries and other executables occupy more than seven gigabytes [24] of storage space. As understanding of the detector is improved and more sophisticated data analysis algorithms are developed, the size of executables will only increase. Large program size requires significant time to transfer updates over networks to remote analysis sites, which must be done frequently in an evolving experiment. This utilizes large amounts of network bandwidth and time. Network bandwidths at many remote analysis centers improve at a much slower pace than the overall network, and also typically suffer from restrictive network policies. These issues pose difficulties in keeping the data and libraries up-to-date at remote sites. Problems will be exacerbated in the ultimate implementation of grid since computing tasks must be performed in a self-contained and self-sufficient manner in foreign environments that do not have prior knowledge of particular software executables. This necessitates the need to reduce and manage the transfer of large repetitive libraries. One of the key requirements in resolving these issues is efficient transport of software and other binary data.

We thus propose to build an effective data management system, which transports only the minimal required software needed to perform a given computing task. We will investigate various existing compression [25] and string-matching techniques to identify the most optimal techniques. We intend to analyze these existing software components and, if necessary, develop new ones to

complement them in order to handle the transport of software executables in the DOSAR Grid environment. We will exploit the RAC at UTA and the DØRAM architecture as a binary management hierarchy where software libraries may be temporarily cached, and where decisions on what version to transfer to which remote sites within the region will be optimized.

CONCLUSIONS

Historically, experimental particle physics has pushed information technology in directions that have had profound broader impact. The most spectacular example is CERN's development of the World Wide Web, originally intended to facilitate communication among collaborating physicists spread around the globe. DOSAR's goal of efficiently using computing resources scattered across the globe for tasks such as complex data reduction and simulation will produce grid computing solutions with broad applicability. DOSAR initiatives will also likely affect how we disseminate and analyze data from future experiments at facilities such as the LHC, and perhaps fundamentally enhance the collaborative nature of high energy physics and other large international scientific endeavors. DOSAR will help improve the cyber-infrastructure within the member's region; five participating institutions are located in states that are traditionally under-funded for R&D activities.

In addition to technology development, there is also significant development of human resources in the region. Students at the undergraduate, graduate, and postdoctoral levels receive interdisciplinary training in physics and computer science. Faculty are able to pass on the knowledge gained to a wide spectrum of students in their classrooms. The broader scientific and technical community will be reached by student theses, refereed journal articles, and conference talks resulting from DOSAR's work. Information for the broader public will be disseminated on the web.

Finally, DOSAR will realize the conceptual grid to the level of personal desktop computers, demonstrating the performance of a grid at its fullest level. The DOSAR Grid will be exploited in critical physics analyses of real and simulated data from the DØ experiment, and later from LHC and other future experiments. These data will serve to advance our understanding of nature at the smallest distance scales.

REFERENCES

[1] Fermilab National Accelerator Laboratory, http://www.fnal.gov/.

[2] The DØ Experiment, http://www-d0.fnal.gov/.

[3] The CDF Experiment, http://www-cdf.fnal.gov/.

[4] The ATLAS (A Toroidal LHC ApparatuS) Experiment, http://atlasinfo.cern.ch/Atlas/.

[5] The Large Hadron Collider (LHC) project, http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/.

[6] European Organization for Nuclear Research (CERN), http://www.cern.ch/.

[7] The Linear Collider Research and Development Working Group, http://www.hep.uiuc.edu/LCRD/.

[8] For reviews, see The Review of Particle Physics, K. Hagiwara et al., Phys. Rev. D66, 010001 (2002); E.D. Commins and P.H. Buchksbaum, Weak Interactions of Leptons and Quarks, (Cambridge Univ. Press, Cambridge, 1983); J.M. Conrad, M.H. Shaevitz, and T. Bolton, Rev. Mod. Phys 70, 1341 (1998).

[9] V. Trimble, Ann. Rev. Astron. Astrophys. 25, 425 (1987); B. Sadoulet, Rev. Mod. Phys. 71, S197 (1999); M.S. Turner and J.A. Tyson, Rev. Mod. Phys. 71, S197 (1999); N.A. Bahcall et al., Science 284, 1481 (1999).

[10] The DØ Grid Group, http://www-d0.fnal.gov/computing/grid/

[11] The Computing Grid Project, http://lcg.web.cern.ch/LCG/

[12] Wired News, http://www.wired.com/news/technology/0,1282,45977,00.html

[13] The Grid 3 project, http://www.ivdgl.org/grid2003/

[14] Open Science Grid, http://www.opensciencegrid.org/

[15] American Linear Collider Group (ALCPG), http://blueox.uoregon.edu/~lc/alcpg/

[16] The Sequential Data Access via Metadata (SAM) project, http://d0db.fnal.gov/sam/

[17] J. Bakken, E. Berman, C.H. Huang, A. Moibenko, D. Petravick, R. Rochenmacher, K. Ruthmansdorfer, "Enstore Technical Design Document," Fermilab-JP0026 (1999).

[18] The Enstore project, http://www-isd.fnal.gov/enstore/

[19] H. Severini and J. Snow, "Preparation of a Desktop Linux Cluster for DØ Monte Carlo Production", DØ Note #4208, unpublished (2003), http://www-hep.nhn.ou.edu/d0/grid/docs/DesktopMcFarmHowto.pdf

[20] MC_Runjob Home Page, http://www-clued0.fnal.gov/runjob/current/

[21] The Ganglia project, http://ganglia.sourceforge.net/

[22] C. Leangsuksun, et al, "Availability Prediction and Modeling of High Availability OSCAR Cluster", IEEE International Conference on Cluster Computing, December 1-4 2003, Hong Kong.

[23] The DØ Offsite Reprocessing project, http://www-d0.fnal.gov/computing/reprocessing/

[24] Adaptive Huffman Encoding, http://www.nist.gov/dads/HTML/adaptiveHuffman.html

[25] An extremely fast Ziv-Lempel data compression algorithm, Williams, R.N., http://ieeexplore.ieee.org/xpl/abs_free.jsp?arNumber=213344.