Towards Belle Monte Carlo Production on the APAC National Grid Infrastructure

Marco La Rosa, Glenn Moloney and Lyle Winton

School of Physics, The University of Melbourne, VIC 3010 Australia
mlarosa@physics.unimelb.edu.au
glenn@physics.unimelb.edu.au
lylejw@unimelb.edu.au

In 2004 the Belle Experimental Collaboration reached a critical stage in their computing requirements. Due to an increased rate of data collection, an extremely large amount of simulated data was required to correctly analyse and understand the experimental data. In order to meet requirements, the simulated data production was distributed to remote institutions, including those associated with the Australian Belle collaborators. Resources at the Australian Partnership for Advanced Computing (APAC) and partner facilities were leveraged in this effort. As part of this effort, the group began investigating the use of Grid technologies for the Belle collaboration. This paper will detail the effort expended towards deploying the Belle Monte Carlo production on the APAC National Grid.

ACM Classification: J.2 (Physical Sciences and Engineering)

1. INTRODUCTION

Many international scientific collaborations, laboratories, and facilities are investigating ways to make scientific data and results more accessible to widely distributed communities. Australian researchers are leading the investigation of e-Research technologies within the Belle collaboration at KEK, and are contributors to the LHC Computing Grid (LCG) through membership in the ATLAS collaboration (LCG, 2005).

"The Grid" has been described as "*an infrastructure that enables the integrated, collaborative use of high-end computers*" and computing resources.

In practice, rapid development and change have made stable, scalable, Grid solutions difficult to find. We would then suggest that "Grid" is the effort to provide such an infrastructure. Similarly, we suggest that "Data Grid" is an effort to help share, manage, and process large amounts of distributed data within this infrastructure. In international collaborative research, such as Belle, access to shared data is very important so we are naturally interested in contributing to the Data Grid effort.

1.1 The Belle Experiment

The Belle experiment is situated at the Japanese High Energy Accelerator Research Organization (KEK) on the KEKB accelerator (Abe, 2002). The accelerator produces a large rate of B mesons within the Belle detector which are used for the study of a fundamental violation of symmetry in

Manuscript received: 13 December 2006
Communicating Editor: Professor Bill Appelbe

nature, CP violation (Charge-Parity). This study is a part of the ongoing investigation into the outstandingly successful standard model which provides an elegant description of our world at microscopic scales. The most interesting question is whether the standard model can offer a complete description of CP violation or if new physics is needed to explain the phenomenon.

The Belle experimental detector weighs over 1400 tons, and consists of over 100,000 individual detector elements, some measuring subatomic particle tracks down to accuracies of 100 microns. The experiment and collaboration have generated more than 100TB of data. The Belle collaboration consists of 400 people from over 50 institutions around the world, all of which require access to portions of this data. A large fraction of the Belle data set is simulated accelerator collisions (events) using Monte Carlo techniques. These events are essential for prediction, understanding analysis techniques, determining acceptances, and efficiencies. Simulation is computationally intensive, involving: accelerator beam collisions; particle interactions and decays; accurate modelling of the detector and all components; tracking and interaction of particles through all detector materials; all electronics effects such as signal shapes, thresholds, noise, and cross-talk; and the data acquisition system and electronics. To reduce statistical fluctuations in the simulation three times as much simulated data is generated than experimental data.

Due to increases in the efficiency of the KEKB accelerator, the Belle collaboration is collecting data at an increased rate. While this will enable us to probe even further into the physics of CP violation the collaboration is now faced with increased computing needs. In order to facilitate analysis the Belle collaboration required the simulation of 4 x10^9 events during 2004. The level of computing power required for this effort exceeded the computing available to the Belle collaboration at the KEK facility, so the simulated data production was distributed to remote institutions. As the accelerator continues to increase in efficiency the need arises for more institutions to contribute. Our feeling is that, in this situation, the Belle collaboration could benefit from existing Grid technologies. Additionally, the deployment of an existing application of the scale of Belle could help in the development and testing of Grid middleware and infrastructure.

1.2 Background on Grid

In addition to the Belle experiment we are participating in the ATLAS experiment situated on the Large Hadron Collider (LHC) in CERN, Switzerland (Amstrong, 1994). The ATLAS experiment will probe the matter and forces making up the universe at higher energies and smaller scales than have ever been observed. Studies at the ATLAS experiment will aid us in understanding the origins of mass and the environment at the early universe beyond the reach of astronomy.

The computing infrastructure to support the five experiments on the LHC will be the Worldwide LHC Computing Grid (WLCG) (LCG, 2005). As of November 2006, the WLCG consists of 205 sites in 47 countries offering over 37,000 CPUs and 13PB of storage – making it the largest international scientific Grid. The ATLAS experiment is under construction and will be operational in early 2007. Access to data from ATLAS will be crucial, especially if new physics is quickly uncovered in this frontier region. As Grid technologies will be used to distribute and analyse this data, using existing Grid middleware for the Belle experiment will provide valuable experience to Australian researchers. Additionally, the size of the Belle collaboration, the increasing need for computing power, and real data that can be used for research, make the Belle experiment an ideal test case for Grid deployment.

Our initial investigations began in 2002 with Version 1 of the Globus Toolkit (Globus, 2005) and the meta-scheduler Nimrod/G (Abramson, 2002). As part of these investigations we were able to demonstrate the use of the Belle Analysis Software Framework (BASF) on Grid enabled resources located at the School of Physics, University of Melbourne, and at the Department of Physics, University of Sydney. BASF was also modified to process data streamed over Globus protocols, such as GASS and GSIFTP. Streamed GSIFTP file processing proved to be as efficient as NFS file processing over the same network.

1.3 High Energy Physics Testbed

To further investigate the existing Grid and e-Science software and infrastructure, in early 2003 an Australian testbed was constructed. Its initial construction occurred over a period of nine days. The testbed consists mainly of hardware donated by IBM Australia in collaboration with IBM Asia-Pacific, with Version 2 of the Globus Toolkit as standard middleware (Globus, 2005). There are five nodes making up the testbed and these are all used for processing and data storage. The nodes are located at the University of Melbourne, the APAC National Facility at the Australian National University (ANU) in Canberra, the South Australian Partnership for Advanced Computing (SAPAC) at the University of Adelaide, and at the University of Sydney. The APAC National Facility mass storage system was also used for storage of data. There are also central services which help to maintain and utilise the testbed, such as a Certificate Authority (CA), a Grid Index Information Service (GIIS), and a Globus Replica Catalog.

Belle simulation and analysis using the testbed was successfully demonstrated at the 4th PRAGMA workshop and ICCS2003 conferences. Further demonstration followed at the international Super Computing conference (SC2003) and associated Grid 2003. The testbed was also utilised in the high performance computing challenge at SC2003. More recently our experience of the Grid through this testbed has been presented to the ATLAS collaboration in CERN and the Global Grid Forum held in Berlin in 2004.

This “High Energy Physics Testbed” continues to be maintained and used for developing and testing new technologies. Access to such a distributed testbed has been essential for this work and has allowed existing production resources to be unaffected and used in parallel.

2. INFRASTRUCTURE

Through the APAC Grid Program we are working with the APAC partners to provide Data Grid infrastructure for high energy physics applications. The Belle simulation production is earmarked to be the first application to be deployed on the APAC National Grid infrastructure.

2.1 Data Management

Our initial investigation into data catalogues and data management involved using the Globus Replica Catalog (RC). The command line interface to the replica catalogue was cumbersome and slow to use, so much so that we developed a simplified interface to help perform common tasks. The rising popularity of Storage Resource Broker (SRB) and installation at the APAC National Facility for use on the MACHO project (Macho, 2005), lead to the investigation of this data catalogue as an alternative. The ease of data management and access, simplicity of user commands, and tighter coupling between logical and physical files has made SRB an attractive solution.

As a result of our SRB investigations we were able to demonstrate international transfer of data using SRB at the “Workshop on High Energy Physics Data Grid” in KEK, Japan late 2004 (KEK, 2004). With the support of Stephen McMahon from the APAC National Facility, we demonstrated the federation of SRB enabled storage facilities within Asia, Poland, and Australia. This federation was used to transfer data for the Belle simulation production throughout Australian facilities. As a result of this meeting the Belle collaboration agreed to further the use of SRB for the global transport of data. This will eventually provide easier access to Belle experimental data for researchers within Australia and throughout the world.

2.2 Grid Middleware

Traditional high performance computing almost guaranteed an application developer a homogeneous computing environment. That is, the developer could design their application for a targeted computing environment where the architecture was the same across the system. Increasingly however, computing is facilitating collaboration and data sharing in such a way that these assumptions no longer hold true. Indeed the interconnection of heterogenous resources across enterprise boundaries is allowing the development of a new computing paradigm – namely, Grid computing (Foster, 2002).

To this end, grid middleware allows applications to access heterogenous resources in a uniform and consistent manner. The Globus toolkit may be considered a first attempt at connecting disparate, heterogenous resources in a defined and consistent manner. Sometimes termed 2nd generation middleware, application developers can use Globus services and programming interfaces to submit jobs to available resources, manage data and retrieve output in a defined way which is independent of the architecture of the underlying resource.

This new computing paradigm is defined by the sharing and coordination of diverse resources amongst dynamic, distributed "virtual organisations" (VO) of users. Perhaps more importantly, this model facilitates the development of technological solutions outside of the domain of traditional high performance computing. For example, a distributed Data Grid.

Experiences with Globus and the High Energy Physics testbed have proven an invaluable introduction to Data Grid techniques. A large number of problems were encountered and overcome, each helping to better understand an ideal Data Grid model for our requirements. Some of the problems included: the middleware was difficult and time consuming to install; the middleware required the loosening of network security; validation of middleware installations was difficult; configuration and maintenance became time consuming on growing Grids; existing data catalogues were difficult to maintain; job dispatch tools were rudimentary and not user friendly; intermittent failure of middleware due to network problems; lack of a virtual organisation information system; and lack of job status and error reporting.

Indeed as time has passed and the model has matured, it has been shown that from a users point of view, the system must provide simpler, more well defined interfaces. 3rd generation middleware (LCG, 2005; Grid3, 2005) has attempted to address the limitations observed in the 2nd generation toolkits.

The LCG middleware provides the end user with a number of important services including Virtual Organisation (VO) membership, resource brokering, meta-scheduling and information services (VOMS, 2005). Membership of a virtual organisation allows the user to identify herself to the system as being authorised to use that VO's allotted resources. User queries to the information system (as a member of a specific VO) will return available resources which could potentially satisfy the users requirements. Resource brokering actually compares the users requirements with the available resources and prepares the users' request for submission to specific resources. Meta-schedulers extend resource brokering services by serving a specific VO. When a user requests resources, the meta-scheduler matches that request with available resources in a manner that is consistent with, and operates within the guidelines agreed upon by the user's VO and the resource provider. As a 3rd generation platform, the LCG grid middleware provides users with a higher level view of available resources.

As an aside, some technologies which have been developed outside the domain of grid computing integrate well with the fundamental ideas of 3rd generation middleware. For example, consider wiki software for collaboration, access-grid for meetings and conferences, PACMAN for code management and software deployment, meta-data Catalogue Service (MCS) for the description of file and data.

2.3 APAC National Grid

The Australian Partnership for Advanced Computing (APAC) has commissioned the development of a National Grid infrastructure to support Australian eResearch (Francis, 2005). As a mechanism for federating national computational resources, data repositories and expertise, the National Grid project will allow Australian research communities to effectively participate in the global arena.

Indeed this infrastructure is particularly important for the Australian High Energy Physics (HEP) community. As members of both the Belle experiment at KEK in Japan and the forthcoming ATLAS experiment at CERN in Switzerland, the Australian HEP community is taking a lead role in the deployment and use of grid technology for analysing and processing large amounts of data.

As a first step towards developing expertise in Grid middleware technology, the HEP group is working to deploy the Belle Monte Carlo production on the national infrastructure. This exercise, which has traditionally been achieved via successive login and submission to various disparate resources, has been the first application to be deployed on the national Grid. Deployment of the Belle production on the Grid will result in real time and effort savings arising from the single sign-on nature of the infrastructure and the ease of job submission to various resources.

As part of our collaboration in the ATLAS experiment, we are working on deploying the LCG middleware at our partner computing sites. Conducting the Belle Monte Carlo production on an LCG infrastructure allows us to develop the expertise required to participate within ATLAS. Given that we are collaborating on an international scale, these skills are essential for us to take full advantage of experimental data as quickly as possible.

2.4 LCG-2 Deployment

The LCG-2 grid middleware is provided as packages for installing and configuring specific grid components. This includes packages for worker nodes (traditionally known as cluster compute nodes), compute elements (traditionally known as server/management nodes), storage elements, workload management system, site information system, monitoring nodes and user interfaces.

Initial deployment involved installing and configuring all of the components on a single host. As a proof-of-concept, this deployment showed that the LCG-2 middleware had reached a critical level of maturity and was easily deployable. Indeed, connecting the host to a local PBS resource allowed us to test the functionality of the middleware and understand its design. From this initial deployment we have found that we are able to construct ‘gateway’ machines consisting of compute and storage element components which can then be interfaced to existing PBS and data store resources.

LCG-2 gateway machines have been deployed at Advanced Research Computing Services (ARC), The University of Melbourne and the Victorian Partnership for Advanced Computing (VPAC). These machines are connected to local PBS clusters and provide an LCG-2 Grid interface to these resources.

In addition to the Grid-gateways, a Workload Management System (WMS) host has also been deployed. As a core component of the LCG-2 middleware, the WMS provides a number of essential services including a network server, a workload manager, resource broker, a job adaptor/controller and a logging and bookeeping service. The WMS host accepts job submissions from user interfaces, queries the information system for available hosts which match the users requirements, prepares the job for submission and then submits the job to the chosen resource.

These deployments have provided a minimal, yet functional LCG-2 grid in Australia. Overall, this trial has been successful. Although difficulties have been encountered. Through the deployment of these nodes it has been seen that the LCG middleware implicitly assumes that grid tools will be available on the compute nodes of the cluster.

When a job is submitted to the Workload Management System (WMS), the job adaptor component creates a shell script which is executed as part of the PBS job script. The purpose of this script is to copy the input sandbox from the WMS to the compute node, and, to copy it back once the job has completed. This is achieved using globus-url-copy and gsiftp. However, the clusters to which these gatekeepers interface do not have Grid tools available on the compute nodes. Furthermore, the compute nodes are on private networks and are not configured for external access. In order to get around this, the job adaptor scripts have been modified to execute the globus-url-copy commands on the LCG-2 gateways via Secure Shell (SSH).

Although a simple solution, it will not scale readily. Consider that the Belle Monte Carlo production typically requires 150MB of input data and may produce up to 1GB of output. Staging of the data onto the gateway prior to or following the calculation for every job submitted to the resource simply will not scale. For the ATLAS analysis where the input and output will typically be an order of magnitude greater, this solution will not suffice at all. Furthermore, for the ATLAS analysis, at any point it may be necessary for the calculation to retrieve more data. Clearly it is not feasible to stage all of the available data onto the gateway prior to job submission. Accordingly, the ATLAS analysis software framework is being designed to operate on streaming data. Thus the requirement of grid tools on the compute nodes.

3. PRODUCTION

Prior to Belle Monte Carlo production in 2004 it was felt there was a need for a more traditional mechanism for job dispatch and management, to be deployed in parallel with the ongoing Grid deployment. In this way Grid deployment problems would not greatly effect urgently needed production.

The production required the use of the Belle Analysis Software Framework (BASF), 2 million lines of code written by the Belle collaboration for simulation and analysis. This framework includes over 200 “pluggable” modules, around 60 of which are used in Monte Carlo simulation. Simulation occurs in two stages. The first stage is “event generation” which generates a small amount of “evtgen” data and takes little CPU. This stage is done centrally at KEK to reduce overlap due to poor random seed choices. The second stage is “simulation” and “reconstruction” which takes “evtgen” data, performs a full simulation of the Belle experiment, and summarises the output in a similar way to experimental data.

A typical production run consisted of around 1700 “evtgen” data files. These files represented four different types of simulation required within Belle. Each “evtgen” file was processed as a single job together with signal background data files called “addbg”. The script to process the job was 10kB, “evtgen” files ranged from 3 to 50MB, and “addbg” files ranged from 20 to 100MB. Jobs produced 50 to 1000MB of simulated data, 4MB histogram files, and 3 to 30MB log files kept for reference. Jobs typically took hours to days. The wide variation in data size and processing time was due to variation in the four different data types.

3.1 Traditional Job Dispatch

During the 2004 production we had access to a number of resource facilities throughout Australia: APAC National Facility in Canberra; AC3 in Sydney; ARC at the University of Melbourne; and VPAC in Melbourne. The traditional method of dispatching jobs to these resources is connection to a user "gateway" machine via *ssh*, transferring required configuration files via *scp*, then submission of the job script to the local batch system. In this case all resources provided PBS (Portable Batch System). The Belle Analysis Software Framework (BASF) was installed in a user home directory on each system prior to production.

A job dispatcher daemon was developed to enable the centralised and coordinated submission of PBS jobs via *ssh*. This included tools for submission, deletion, post-processing, and monitoring of jobs across all resources. The daemon centrally recognised jobs in various stages: Holding, not to be submitted; Pending, awaiting submission; Doing, submitted to remote resource; Post, awaiting post-processing; and Done, post-processing completed. The status of jobs submitted to remote resources was periodically monitored by parsing PBS command output executed via *ssh*. Differences between resources and facilities required the job dispatcher to be aware of limitations and settings for each. The dispatcher ensured the number of jobs submitted to any resource was kept within an optimal or acceptable range. Required environment variables, PBS resource requirements, and PBS attributes were configured for each host separately.

The 2004 Belle Monte Carlo production occurred over several months during which we estimate having utilised around 120 CPUs constantly. We currently have 4TB of data on the APAC National Facility Mass Storage System, and are requesting 10TB for ongoing Belle production. Data on this storage facility is managed and accessed via SRB.

3.2 Grid Job Dispatch

We are developing a simple data resource brokering tool and meta-scheduler to provide intelligent decisions about job location within a Data Grid. It is often advantageous to execute jobs in a location as close as possible to the required data. Using grid resource information services (GRIS, BDII) and data catalogue services (Globus RC, SRB), free computing resources can be allocated and mapped to the most appropriate copy of the data. It is our eventual goal to implement these brokering and scheduling techniques into an existing economic grid scheduling tool, GridBus (Venugopal, 2004). The tool, named "GQSched", was originally designed as a simple prototype Data Grid job scheduler and dispatcher. It was designed with a parameter sweeping functionality similar to that of the Nimrod/G tool, with the additional ability to sweep over Grid files and collections. Grid files can be either physical or logical files stored on remote resources. Physical files are made accessible via Grid protocols (eg. GSIFTP, GASS). Logical files are registered within data catalogues such as Globus RC and SRB. Logical files are more complex than simple parameter variables as each can be stored at a number of physical locations.

Since its inception in mid 2002, GQSched has progressed to a stable and usable tool, enabling quick command-line access to Grid processing and data resources. One advantage of this tool is the use of very low level, common, Grid services. While many tools are moving toward using 3rd generation Grid services, 2nd generation services are still the most common and likely to be available. By using only a minimal subset of low level services we can ensure the most amount of compatibility with existing Grid resources.

The deployment of LCG infrastructure throughout the APAC National Grid aids the Belle production effort in two ways. First, the LCG middleware provides the low level Grid services that we require to access resources. Second, we are investigating the use of higher level LCG services to aid in resource brokering, job dispatch, and job monitoring.

Together with the GridBus project we are investigating how our experiences and development can be contributed back to the LCG project.

We have begun initial testing of Belle Monte Carlo production using the GQSched tool, running in parallel with traditional job dispatch. If successful, we hope to use the APAC National Grid deployment of the Belle production as a model for our international collaborators.

4. CONCLUSIONS

A trial of Belle Monte Carlo production has been successfully deployed on the Australian High Energy Physics testbed. We have shown that data stored in an SRB federation can be transferred from Japan to an Australian data centre, staged to local facilities for processing, and the results transferred back to the collaboration in Japan. Until now the production effort has utilised traditional job submission techniques and SRB for data management. Following the success of this trial we are currently implementing an LCG infrastructure on which to deploy the Belle Monte Carlo production using a meta-scheduler developed by High Energy Physics for Data Grid.

REFERENCES

ABE, K. *et al* (2002): The belle collaboration. *Nucl. Inst. Meth. A*, 479:117.

ABRAMSON, D., BUYYA, R. and GIDDY, J. (2002): A computational economy for grid computing and its implementation in the NIMROD-G resource broker. *Future Generation Computer Systems (FGCS) Journa*l, 18:1061–1074.

ARMSTRONG, W.W. *et al* (1994): ATLAS Technical Proposal. Technical Report CERN/LHCC/94-43 LHCC/P2, CERN, http://atlas.web.cern.ch/Atlas/TP/NEW/HTML/tp9new/tp9.html LCG (2005): The LHC Computing Grid project (LCG). http://lcg.web.cern.ch/LCG/. Accessed 14 December 2005.

FOSTER, I. and KESSELMAN, C. (1997): Globus: A metacomputing infrastructure toolkit. *Intl J. Supercomputer Applications*, 11(2):115-128.

FOSTER, I. and KESSELMAN, C. (1999): The Grid: Blueprint for a new computing infrastructure. Morgan Kaufmann.

FOSTER, I. and KESSELMAN, C., TUECKE, S. (2001): The anatomy of the grid: Enabling scalable virtual organizations. *Intl J. Supercomputer Applications*.

FOSTER, I., KESSELMAN, C., NICK, J. and TUECKE, S. (2002): The physiology of the grid: An open Grid services architecture for distributed systems integration. *Global Grid Forum*, June.

FRANCIS, R. (2005): The APAC national grid program plan. http://www.vpac.org/twiki/pub/APACgrid/WebHome /APACGridProgramPlan.pdf. APAC.

GLOBUS (2005): The Globus Toolkit. http://www.globus.org/toolkit/. Accessed 14 December 2005.

GRID3 (2005): The grid3 collaboration. http://www.ivdgl.org/grid3/. Accessed 14 December 2005.

KEK (2004): KEK Computing Research Center. Workshop on High Energy Physics Data Grid. *High Energy Accelerator Research Organisation (KEK)*, December.

MACHO (2005): The MACHO Project. http://wwwmacho.mcmaster.ca/. Accessed 14 December 2005.

LCG-UI (2005): LCG-2 user guide. https://edms.cern.ch/file/454439//LCG-2-UserGuide.html. Accessed 14 December.

VENUGOPAL, R., BUYYA., R. and WINTON, L. (2004): A Grid service broker for scheduling distributed data-oriented applications on Global Grids. *5th International Middleware Conference*, October.

VOMS (2005): Virtual organisation management server (VOMS). http://edg-wp2.web.cern.ch/edg-wp2/security/voms/. Accessed 14 December 2005.

BIOGRAPHICAL NOTES

Marco La Rosa graduated with a PhD in Computational Chemistry from the University of Melbourne in 2003. He used molecular dynamics and Monte Carlo techniques to simulate the behaviour of surfactants at the oil/water interface. Following completion, he decided to pursue his interest in computing and has since built beowulf clusters for chemistry and physics research. He is now focused on administering the first Australian node of the Worldwide LHC Computing Grid.

Marco La Rosa

Glenn Moloney graduated with a PhD in Physics from the University of Melbourne in 1999. He has contributed to several high energy physics experiments, including the Belle experiment in Tsukuba, Japan and the ATLAS experiment at the CERN laboratory in Switzerland. He has been the project leader of the High Energy Physics Application Project of the APAC National Grid Program, and has led planning for Australian participation in the Worldwide LHC Computing Grid (WLCG). Glenn has also led the Australian Silicon Vertex Detector program for the Belle experiment.

Glenn Moloney

Lyle Winton began postgraduate research with the University of Melbourne's Experimental Particle Physics group in 1993. His PhD involved research on an international experiment at the world's largest particle physics laboratory, CERN. Subsequently he worked for five years as an IT professional for several companies in software design and project management. Most recently Lyle worked as a research fellow with the Experimental Particle Physics group at the University of Melbourne. He is now working as a senior research support officer at the University supporting e-Researchers and investigating higher education ICT architecture for teaching and research.

Lyle Winton