Journal of Physics: Conference Series **331** (2011) 062009 doi:10.1088/1742-6596/331/6/062009

Integration of cloud, grid and local cluster resources with DIRAC

Tom Fifield[2], Ana Carmona[1], Adrián Casajús[1], Ricardo Graciani[1] and Martin Sevior[2]

[1]The University of Melbourne, Australia, [2] Institut de Ciències del Cosmos (ICC), Universitat de Barcelona, Spain

E-mail: fifieldt@unimelb.edu.au, ana@ecm.ub.es, adria@ecm.ub.es, graciani@ecm.ub.es, martines@unimelb.edu.au

Abstract.

Grid computing was developed to provide users with uniform access to large-scale distributed resources. This has worked well, however there are significant resources available to the scientific community that do not follow this paradigm - those on cloud infrastructure providers, HPC supercomputers or local clusters. DIRAC (**D**istributed **I**nfrastructure with **R**emote **A**gent **C**ontrol) was originally designed to support direct submission to the Local Resource Management Systems (LRMS) of such clusters for LHCb, matured to support grid workflows and has recently been updated to support Amazon's Elastic Compute Cloud.

This raises a number of new possibilities - by opening avenues to new resources, virtual organisations can change their resources with usage patterns and use these dedicated facilities for a given time.

For example, user communities such as High Energy Physics experiments, have computing tasks with a wide variety of requirements in terms of CPU, data access or memory consumption, and their usage profile is never constant throughout the year. Having the possibility to transparently absorb peaks on the demand for these kinds of tasks using Cloud resources could allow a reduction in the overall cost of the system.

This paper investigates interoperability by following a recent large-scale production excercise utilising resources from these three different paradigms, during the 2010 Belle Monte Carlo run. Through this, it discusses the challenges and opportunities of such a model.

1. Introduction

DIRAC[1, 2, 3] had supported job submission to LRMS from its outset, and has been efficiently utilising grid resources for many years now. The recent rise of cloud technology was not initially scoped in the design of DIRAC, however the modular nature of the framework facilitated its integration. This work is described in section 2.

With this completed, several rounds of Monte Carlo production were carried out, using the Belle experiment[4] as a test-case. Belle has a large peak in computational requirements (approximately double its normal) for around three months of the year when Monte Carlo is produced[5]. Belle was purchasing physical servers to accomodate this, which could go unused for the remainder of the time. Cloud computing allows the rental of computing power even on an hourly basis without the expensive capital investment, which opens the possibility of extending the owned infrastructure. Further details are found in 3

Using the three paradigms, instead of purely the grid, Belle was able to have short-term access to resources additional to what it owned. This integration of heterogeneous pieces into a single system implies interoperability of the different pieces, the significance of which is detailed in Section 4.

2. Development

The late resource-payload binding ('pilot') model[6] employed by DIRAC on the grid has been very successful[7]. Pilot jobs deploy and execute a JobAgent to obtain real jobs from a central task queue, once executing on a worker node. This moves the matching of resources to jobs to the worker nodes themselves, greatly reducing the load on central services and allowing quick reaction to changes in the underlying resources pools. There are several auxillary benefits to this approach, including allowing the assignment of priority to user jobs on a virtual-organisation wide basis, verification of environment suitability and confirmation of resource availability.

This model was easily extended to work on the cloud. Rather than using grid mechanisms to get Pilots onto the Worker Node, cloud APIs are used to instantiate Virtual Worker Nodes. The cloud Worker Nodes have the Job Agent pre-installed and are ready to match and execute payloads. Pilots are replaced by instantiating Virtual Worker Nodes but the modular resource discovery, the central control of the scheduling and the late matchmaking paradigm remain. See Figure 1

Three new components were created to facilitate the use of cloud:

- VirtualMachine Scheduler: Monitors DIRAC TaskQueues and requests new VM from resource provider as appropriate
- VirtualMachine Monitor: On-VM module that reports activity and halts VM if no longer needed
- VirtualMachine Manager: Collects information about requested, running and halted VMs, and provides usage monitoring with functionality accessible as an extension to the DIRAC web interface[8]

Cloud computing resources added to DIRAC are automatically integrated with other resources, for instance Grids or local clusters. DIRAC ensures that there is good interoperability among them and provides many additional features to manage, monitor and account all activities and resource usage out of the box.

The full technical details of the modifications to DIRAC to enable work on the cloud can be found in [9].

3. Results

The Belle Monte Carlo simulation task is divided into individual sub-tasks closely following the data taking of the experiment. Each individual sub-task corresponds to a detector run[1] and thus there is a huge dispersion in the cpu time and output data requirements, depending on the duration and conditions of the corresponding run. The input for the simulation is taken from the official sets of scripts and input data files provided by the computing group of the collaboration. And, at the end of the execution, the simulated data and log files must be transferred to Belle grid Storage Elements (SEs) and File Catalog (FC). The exercise was divided into three parts.

3.1. Cloud-only

The aim of this part was to verify the robustness of the newly developed cloud components, described in section 2. 160 cores were run for two weeks, with a twenty-four hour ramp up using

[1] A run is the collision data registered by a detector in an uninterrupted manner with almost identical conditions, its duration and the accumulated data may vary by several orders of magnitude.

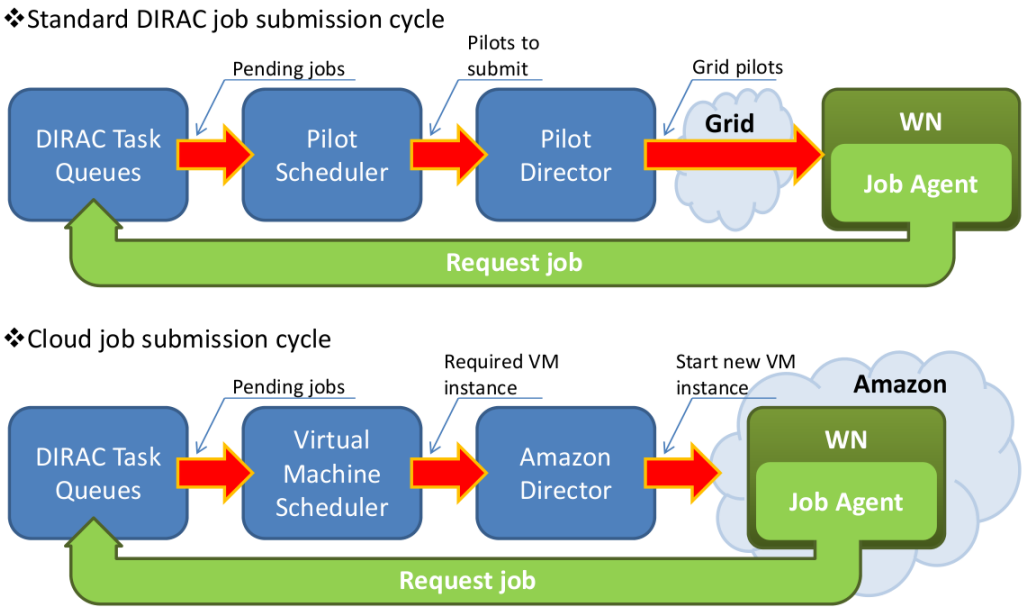

Figure 1. Applying the pilot model(top) to cloud-based Worker Nodes(bottom)

250 8-core virtual machines on Amazon EC2. Output was put to grid storage elements - at KEK in Japan, GridKa in Germany and SiGNET in Slovenia. Though initially Grika was used, two hours before the peak of the run, this storage stopped responding. In a good demonstration of the DIRAC Configuration Service[10][11], one variable was changed to send data instead to KEK. Highlights:

- Peaked at 2000 cores from Amazon EC2
- Data outbound to grid at 50MB/s
- Ran input data from cloud-based DIRAC SE

3.2. Cloud and HPC

This second section was used to investigate a new EC2 product - spot instance pricing, and integrate non-grid enabled clusters. Where previously, we pre-staged input data to a cloud storage element, in this instance we copied it directly from grid storage elements with no significant performance impact. Highlights:

- 8.5 CPU years over 5 days
- 2 HPC centres in Barcelona (40%), Amazon EC2 (60%)
- Input data from the grid

3.3. Grid, Cloud and HPC

In our final run, we placed some emphasis in brokering jobs to different resources based on their characteristics. The cloud and clusters were ideal for running long (>24 hours) jobs, and on the cloud we were able to take advantage of the multiple core support in our software - something that is near-impossible on the current grid infrastructure as detailed in user feedback on EGI infrastructure and the WLCG[18][19]. During this, DIRAC was running on a pair of low-specification (1 core, 2GB RAM) virtual machines in Barcelona and scaled without issue.

Highlights:

- CPU efficiency >95%
- Up to 2,400 cores
- 6 grid sites(KEK in Japan, GridKa in Germany, IJS in Slovenia, CYFRONET in Poland, CESNET in Czech Republic and KISTI in South Korea), 3 HPC clusters and Amazon EC2

Figure 2 is perhaps the best depiction of the interoperability work - showing a clear grid baseline (50%), supplemented by cloud(28%) and local clusters(23%). All of the cloud capacity from EC2 in this excercise was derived from a product known as Spot Instances. These resources are provided at a lower rate, but availability is dependant on the price at which they are aquired. The resilience of DIRAC enabled us to use this without issue, as any jobs failing due to loss of a cloud virtual machine were automatically resubmitted.

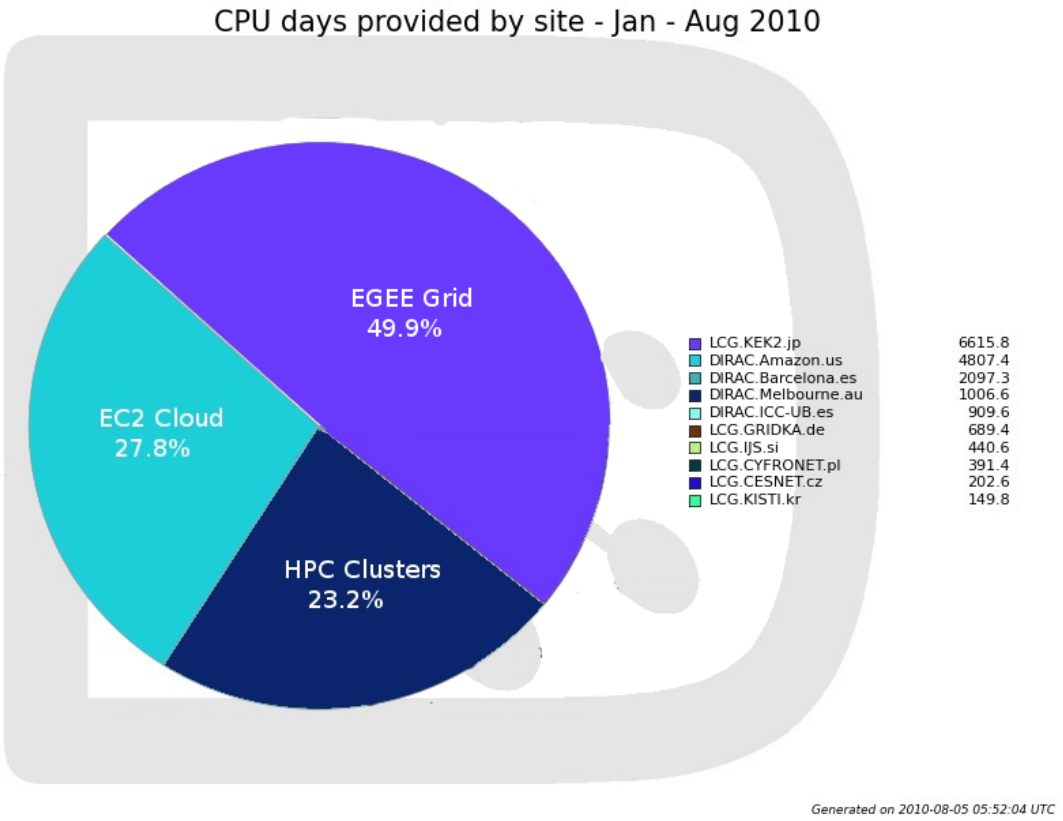

Figure 2. Distribution of CPU days by computing paradigm

Further interpretation of these results, with a detailed TCO analysis will be provided in a future publication.

4. Interoperability

The interoperability between grids and clouds has been under investigation for some time. The beginnings of the research can be seen in the work of adding the ability to make batch systems 'dynamic' - that is, adding and removing (often virtualised) Worker Nodes from a batch system queue based on certain events[12, 13, 14, 15, 16, 17]. This has resulted in such modern projects as INFN's Worker Node on Demand[20], which deploys virtualized resources at the WLCG Tier 1 in Italy. This methodology, aiming to make computing resources 'elastic' matches well with cloud computing. According to a recent StratusLab survey[21], most administrators who

responded were already using cloud or virtualization technologies or planning to deploy them in the next 12 months.

EGEE, the peak general purpose grid infrastructure in Europe at the time, itself acknowledged the potential of cloud computing in their 2008 study[22], concluding that "a roadmap should be defined to include cloud technology in current e-Infrastructures ...". As a result, currently there are several[23, 24, 25] EU-funded projects working on various aspects of cloud-grid interoperability, and providing simplified access cloud paradigm to the community.

However, despite this large amount of work on the infrastructure angle, little work has been completed from the Virtual-Organisation side. In addition to our own previous work[26][27][9], notable projects include clobi[28] which provides a cloud backend to Ganga[29] for the ATLAS experiment[30].

The potential reasons for the lack of effort on this front is perhaps derived from the work of Field, Laure and Schulz[31]. The paper details experience in grid deployment with a focus on five interoperability mechanisms: Virtual Organisation Driven, Parallel Deployment, Gateways, Adaptors and Translators and Common Interfaces. Field, Laure and Schulz note that "For Grid Computing to achieve its full potential, different infrastructures must offer interoperable services which a user can access in a seamless way...", which is the overarching aim of our work. However, several downsides are attributed to the VO-driven approach. The authors state that:

- it places significant effort on the VO in their development framework
- effort required also increases with the number of Grid infrastructures
- it often results in a keyhole approach where the minimum common subset of functionality is used

However, if one notes that the DIRAC framework already has a large userbase with developers commited to supporting it - the points about development effort become less relevant. The ideal solution would be for middleware to support common standards, however the design effort involved shows this trailing the user demand for the infrastructure, often by years.

In this work, Belle was able to utilise grid, cluster and cloud resources to contribute to its Monte Carlo production - without any changes to job code. It is worth noting that other virtual organisations with use of similar pilot framworks already supporting multiple grid backends could make similar modifications to gain this functionality.

5. Conclusion

The cloud paradigm has gained enormous traction in the IT world, and DIRAC is ready to provide seamless integration of these resources to its existing grid and cluster users today. High Energy Physics is in the position to take a leading role in the integration, but scale will be a key factor in deciding suitability.

Acknowledgements This work would not have been possible without the economic support of Centro Nacional de Fsica de Partículas, AstroFísica y Nuclear, CPAN (reference CSD2007-00042 from Programa Consolider-Ingenio 2010), Programa Nacional de Física de Partículas, FPA (reference FPA2007-66437-C02-01 from Plan Nacional I+D+i), the Australian Research Council Discovery Project (reference DP0879737), and KEK.

References

[1] A Tsaregorodtsev, M Bargiotti, N Brook, A C Ramo, G Castellani, P Charpentier, C Cioffi, J Closier, R G Diaz, G Kuznetsov, Y Y Li, R Nandakumar, S Paterson, R Santinelli, A C Smith, M S Miguelez, and S G Jimenez. Dirac: a community grid solution. *Journal of Physics: Conference Series*, 119(6):062048, 2008.

[2] A C Smith and A Tsaregorodtsev. Dirac: reliable data management for lhcb. *Journal of Physics: Conference Series*, 119(6):062045, 2008.

[3] A. Tsaregorodtsev et al. DIRAC - Distributed infrastructure with Remote Agent Control. 2003.

[4] A. Abashian et al. The Belle detector. *Nucl. Instrum. Meth.*, A479:117–232, 2002.

[5] Ichiro Adachi et al Belle computing system. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 534(1-2):53 – 58, 2004. Proceedings of the IXth International Workshop on Advanced Computing and Analysis Techniques in Physics Research.
[6] Adrian Casajus, Ricardo Graciani, Stuart Paterson, Andrei Tsaregorodtsev, and the Lhcb Dirac Team. Dirac pilot framework and the dirac workload management system. *Journal of Physics: Conference Series*, 219(6):062049, 2010.
[7] Stuart Paterson, Joel Closier, and the Lhcb Dirac Team. Performance of combined production and analysis wms in dirac. *Journal of Physics: Conference Series*, 219(7):072015, 2010.
[8] A Casajus Ramo and M Sapunov. Dirac: Secure web user interface. *Journal of Physics: Conference Series*, 219(8):082004, 2010.
[9] R. Graciani et al. Belle-DIRAC Setup for using Amazon Elastic Compute Cloud. *Journal of Grid Computing*, 2010.
[10] A. Casajus and R. Graciani. DIRAC Services and Agents. In *CHEP 2007*.
[11] R. Graciani and A. Casajus. DIRAC Agents and Services. In *CHEP 2007*.
[12] Jeffrey S. Chase, David E. Irwin, Laura E. Grit, Justin D. Moore, and Sara E. Sprenkle. Dynamic virtual clusters in a grid site manager. *High-Performance Distributed Computing, International Symposium on*, 0:90, 2003.
[13] I. Foster, T. Freeman, K. Keahy, D. Scheftner, B. Sotomayer, and X. Zhang. Virtual clusters for grid communities. *Cluster Computing and the Grid, IEEE International Symposium on*, 0:513–520, 2006.
[14] Michael A. Murphy, Brandon Kagey, Michael Fenn, and Sebastien Goasguen. Dynamic provisioning of virtual organization clusters. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, CCGRID '09, pages 364–371, Washington, DC, USA, 2009. IEEE Computer Society.
[15] Hideo Nishimura, Naoya Maruyama, and Satoshi Matsuoka. Virtual clusters on the fly - fast, scalable, and flexible installation. *Cluster Computing and the Grid, IEEE International Symposium on*, 0:549–556, 2007.
[16] Wesley Emeneker, Dave Jackson, Joshua Butikofer, and Dan Stanzione. Dynamic virtual clustering with xen and moab. In *Frontiers of High Performance Computing and Networking ISPA 2006 Workshops*, volume 4331 of *Lecture Notes in Computer Science*, pages 440–451. Springer Berlin / Heidelberg, 2006. 10.1007/11942634_46.
[17] Wesley Emeneker and Dan Stanzione. Dynamic virtual clustering. In *Proceedings of the 2007 IEEE International Conference on Cluster Computing*, CLUSTER '07, pages 84–90, Washington, DC, USA, 2007. IEEE Computer Society.
[18] https://documents.egi.eu/public/RetrieveFile?docid=211 EGI-InSPIRE User Feedback and Recommendations Gergely Sipos, Members of EGI-InSPIRE collaboration
[19] http://www.eu-emi.eu/events-of-2011#WLCG2011 EMI at WLCG 2011 Collaboration Workshop
[20] http://web.infn.it/wnodes/index.php. Worker Nodes on Demand.
[21] http://stratuslab.eu/lib/exe/fetch.php/documents:stratuslab-d2.1-v1.2.pdf. StratusLab Deliverable 2.1.
[22] Marc-Elian Bégin. An EGEE comparative study: Grids and clouds - evolution or revolution. Technical report, CERN - Engeneering and Equipment Data Management Service, June 2008.
[23] StratusLab. https://www.stratuslab.eu.
[24] Venus-C. https://www.venus-c.edu.
[25] http://www.reservoir-fp7.eu. RESEVOIR.
[26] Martin Sevior, Tom Fifield, and Nobuhiko Katayama. Belle Monte-Carlo Production on the Amazon EC2 Cloud. *Journal of Physics: Conference Series*, 219(1):012003, 2010.
[27] Adria Casajus and Ricardo Graciani. DIRAC Community Grid Solution. In Lorena Bello, editor, *IBERGRID, 4th Iberian Grid Infrastructure Conference Proceedings*, pages 153–164, May 2010.
[28] http://code.google.com/p/clobi/. Clobi.
[29] A Maier. Ganga a job management and optimising tool. *Journal of Physics: Conference Series*, 119(7):072021, 2008.
[30] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.
[31] L. Field, E. Laure, M. Schulz Grid Deployment Experiences: Grid Interoperation. *Journal of Grid Computing*, 287-296, 2009.
[32] T. Abe et al. Belle II Technical Design Report. *ArXiv e-prints*, November 2010.
[33] A Casajus, R Graciani, and the Lhcb Dirac Team. Dirac distributed secure framework. *Journal of Physics: Conference Series*, 219(4):042033, 2010.
[34] R. Graciani and A. Casajus. DIRAC Framework for Distributed Computing. In *CHEP 2006*.
[35] H Nakazawa. Grid Efforts in Belle. *Proceedings of CHEP2007*, 2007.