Kureshi, Ibad, Holmes, Violeta and Liang, Shuo

Hybrid HPC – Establishing a Bi-Stable Dual Boot Cluster for Linux with OSCAR middleware and Windows HPC 2008 R2

Original Citation

Kureshi, Ibad, Holmes, Violeta and Liang, Shuo (2010) Hybrid HPC – Establishing a Bi-Stable Dual Boot Cluster for Linux with OSCAR middleware and Windows HPC 2008 R2. In: UK eScience All-Hands Meeting, 13-16 September 2010, Cardiff, Wales. (Unpublished)

Hybrid HPC – Establishing a Bi-Stable Dual Boot Cluster for Linux with OSCAR middleware and Windows HPC 2008 R2

Ibad Kureshi, Dr Violeta Holmes and Shou Liang
School of Computing and Engineering, University of Huddersfield
Abstract

Abstract

The advent of open source software leading to Beowulf clusters has enabled small to medium sized Higher and Further education institutions to remove the "computational power" factor from research ventures. In an effort to catch up with leading Universities in the realm of research, many Universities are investing in small departmental HPC clusters to help with simulations, renders and calculations. These small HE/FE institutions have in the past benefited from cheaper software and operating system licenses. This raises the question as to which platform Linux of Windows should be implemented on the cluster. As the smaller/medium Universities move into research, many Linux based applications and code better suit their research needs, but the teaching base still keeps the department tied to Windows based applications. In such institutions, where it is usually recycled machines that are linked to form the clusters, it is not often feasible to setup more than one cluster.

This paper will propose a method to implement a Linux-Windows Hybrid HPC Cluster that seamlessly and automatically accepts and schedules jobs in both domains. Using Linux CentOS 5.4 with OSCAR 5.2 beta 2 middleware with Windows Server 2008 and Windows HPC 2008 R2 (beta) a bi-stable hybrid system has been deployed at the University of Huddersfield. This hybrid cluster is known as the Queensgate Cluster. We will also examine innovative solutions and practices that are currently being followed in the academic world as well as those that have been recommended by Microsoft® Corp.

Introduction

As a medium sized HE/FE institution the University of Huddersfield is still finding its place within the research world. A recent research project was undertaken to establish HPC Computing Resources at the Queensgate Huddersfield Campus of the University. Taking the Computing & Engineering Department and the Applied Sciences Department and analysing their software pool it became clear that a cluster based on Unix systems would not adequately cover the universities requirements. Several pieces of software only ran on the Windows platform (e.g. 3d Studio Max for image rendering, Dynamic Studio v3.0 for particle velocimetry), while others had licenses which were platform locked to Windows platform (e.g. Mental Ray). As this project aimed to establish the need of HPC at the University using existing or open source hardware and software, it was not feasible or possible to make 2 clusters of a decent size which would run the two platforms.

To ignore the requirements of Windows users was deemed to be detrimental for the project as a large number of researchers would be left out. Since ANSYS CFD systems and Ferrari have adopted the Windows HPC 2008 R2 Platform for their future applications[1] it was agreed that it would be beneficial to our Mechanical Engineers if we keep our development road map loosely tied to their applications.

There have been several solutions which allow for imaging of desktop and server computers to contain the Linux compute node files and Windows work stations executables. Implemented at Blackburn College Carlinville IL, US, this method allows the machines to be deployed in laboratories as normal user workstation and using scheduling software and CPU idle detection scripts can be used to reboot machines to join a cluster for computation purposes [2]. Other methods involve imaging both the Windows Compute Cluster Pack (CCP)/High Performance Computing (HPC) server on all machines along with the Linux compute node software. These systems are either manually rebooted on a time sharing basis [3] or are automatically rebooted to windows for the job to execute [4]. These nodes immediately reboot back into Linux and wait for further instructions.

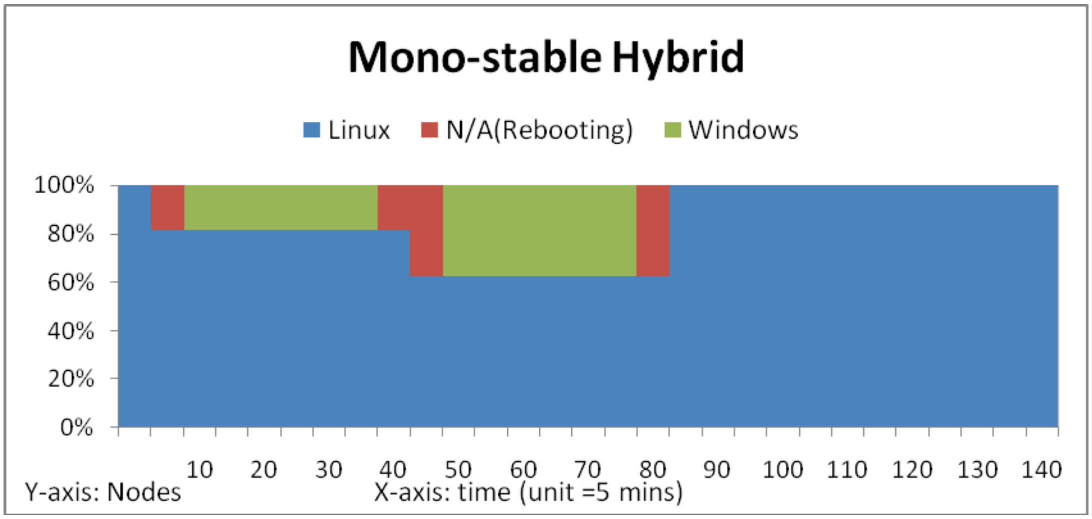

Figure 1: Reboot times in a Mono-stable Hybrid Cluster

As figure 1 illustrates, the nodes spend to much time rebooting. From an all Linux state a Job requiring 20% of the nodes executes forcing the machines to switch to Windows at 5 minutes. Once completed at time 40 these nodes reboot back to Linux where they are instructed to execute another Windows based Job requiring 40% of the nodes. The machines must reboot again. This system also forces users to submit the job in the Linux environment rather than using the Windows based Interface.

The Queensgate Hybrid Cluster

The Beowulf-type Queensgate cluster comprises of two head nodes and 16 compute nodes. Windows Server/ HPC 2008 R2 is deployed on the 'winhead' machine while CentOS/OSCAR is deployed on the 'linhead' machine. Both head nodes must ascertain first whether they have nodes available in their native OS. If not each node must communicate with the other asking it to set a flag in the boot loader and to reboot the machine. Windows provides an API to detect node status while on the Linux side a script has been implemented that parses the text from PBS and then both operating systems use a script that enables TCP based communication. If the Windows server (tx-node) requires 2 nodes it will submit 2 'reboot' jobs into the Linux (rx-node) queue and vice versa. This enables the clusters to keep the first come first served scheduler rule. The *'rx-node'* appends the boot loader and reboots an idle compute node.

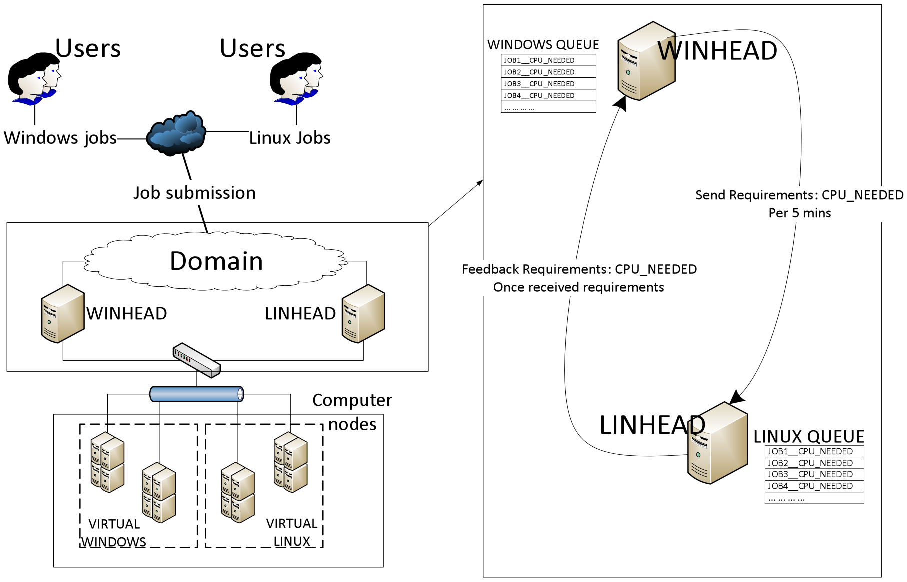

Figure 2:Queensgate Cluster, System and Job Scheduler structure

Each compute node therefore has to be carefully deployed using a prescribed method so adequate boot loaders can be setup to enable the reboot. As the Windows MBR doesn't perform well with Linux, and Windows cannot make changes to GRUB, a tool called GRUB for DOS has to be implemented. A FAT partition is required to hold the Grub and Grub for DOS boot loaders so that both operating systems can access and modify it.

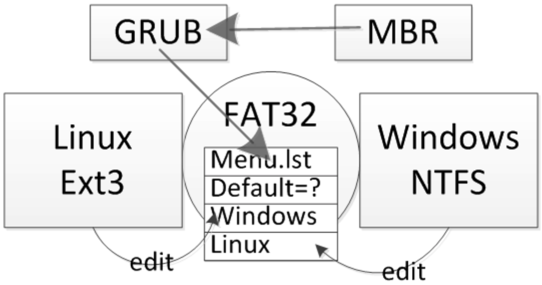

Figure 3: Compute Node Partition Information

It is possible to image the client nodes Windows first or Linux first. In the Windows first method the client image is created and only the first half of the hard drive is partitioned in NTFS format. After deploying the image using PXE boot, the Linux image is created in OSCAR. The partition information is given to the *'systemimager'* package to state that the first half of the hard drive is NTFS. Once the image is created the installer scripts within *'systemimager'* are edited to ensure that the Windows partition is not formatted again. For a Linux first deployment all the partitions can be created as before and the Windows Server DVD installer is added to the image with an 'autoinstall' script. This will image the Node with Linux and Windows with out any user intervention. Once the installation is complete the HPC pack has to be manually configured to connect to the '*winhead*' machine.

Conclusions and Future Work

The Bi-Stable Dual-Boot Linux/Windows system, we have developed and deployed, has allowed the School of Computing and Engineering at the University of Huddersfield to provide a high performance computing resource for both Windows and Linux based applications. As shown in Figure 4 the throughput of the system improves with this bi-stable arrangement.

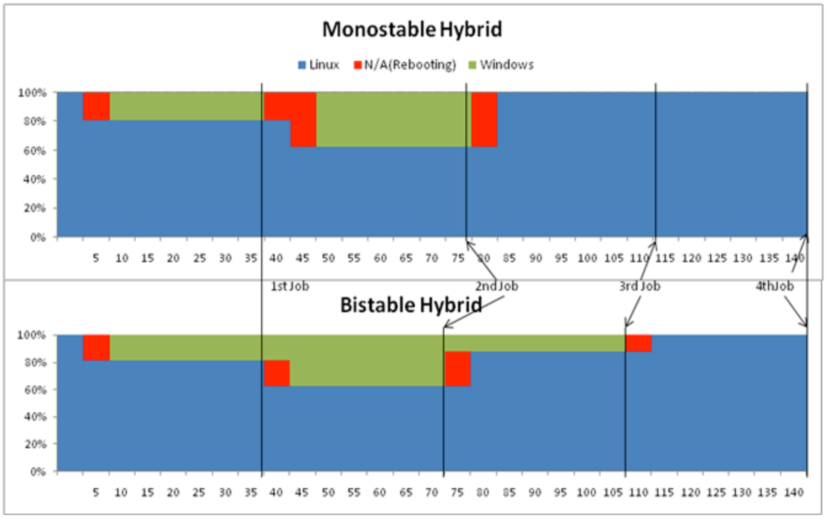

Figure 4: Throughput of Bi-Stable Hybrid Cluster

This system therefore has met its requirements of being economical by using open source software and the software currently available in the department, and has utilised existing hardware to provide the maximum performance in both Linux and Windows based environments.

References

[1] Baker, C. (n.d.). 'Turnkey Dual-Boot Clusters: Clustercorp and Microsoft present Rocks+Hybrid, a simple Windows/Linux Cluster Solution'. [online] Available from: <http://www.microsoft.com/hpc/en/us/community/hpc-forums-blogs.aspx> [Accessed June 1, 2010]

[2] Carrigan, T. (2002). 'Setting up an Oscar cluster where the nodes will alsobe'. *Open Source Cluster Application Resource*. [online] Available from: <http://www.mail-archive.com/oscar-users@lists.sourceforge.net/msg00662.html> [Accessed June 1, 2010]

[3] (2007). 'Dual Boot White Paper'. [online] Available from: <http://www.microsoft.com/downloads/en/results.aspx?freetext=Dual-boot&displaylang=en&stype=s_basic> [Accessed June 1, 2010]

[4] Bucholtz, J. and Zebrowski, M. (2007). 'Dual Boot: WCCS and Rocks Cluster Distribution'. [online] Available from: <http://www.microsoft.com/downloads/details.aspx?displaylang=en&FamilyID=e73a468e-2dbf-4782-8faa-aaa20acb63f8> [Accessed June 1, 2010]

[5] Calegari, P. and Varlet, T. (2008). 'Dual-Boot and Virtualization with Windows HPC Server 2008 and Linux Bull Advanced Server for Xeon'. [online] Available from: <http://www.microsoft.com/downloads/details.aspx?displaylang=en&FamilyID=978fd733-6c92-49f6-95a8-89fba4422e56> [Accessed June 1, 2010]

[6] Gray, P. and Miller, S. (2004). 'In Search of Clusters for High Performance Computing education', in *Proceedings of the 5th International Conference on Linux Clusters: the HPC Revolution*,

[7] Meredith, M. et al. (2003). 'Exploring Beowulf clusters'. *J. Comput. Small Coll.*, Volume 18, Part 4, pp. 268-284. [online] Available from: <http://portal.acm.org/citation.cfm?id=767598.767641> [Accessed May 29, 2010]