Cluster Systems - An Open-Access Design Solution

K. Ginty. J. Tindle. S.J. Tindle

Department of Computing, Engineering and Technology, Faculty of Applied Sciences,
University of Sunderland. St. Peter's Campus, Sunderland, SR6 0DD, U.K.
(e-mail: john.tindle: kevin.ginty@sunderland.ac.uk)

Abstract: An open-access cluster computing facility, believed to be the first of its type, has been established at the University of Sunderland. This paper describes its computing and networking architecture in terms of hardware and Operating Systems platforms, taking into account the need for low energy consumption and low heat dissipation. The system described operates in an open environment, without the need for air conditioning. A dual boot configuration allows performance comparisons to be made between different Operating System platforms on identical hardware. As a result of its general purpose design, the cluster is able to support a wide range of applications across a number of disciplines, most notably engineering and media.

Keywords: Dual boot cluster computing, low noise and power, networking, grid, Windows Server Compute Cluster Edition.

1. INTRODUCTION

The University of Sunderland recently installed a new large cluster computer. This paper describes the underlying philosophy of the system design. Motivation for the development of this facility centres on efforts to advance earlier work completed by the Network Research Group. This involved the creation of network planning tools which employed numerical optimisation methods based upon evolutionary computing techniques, particularly the genetic algorithm (GA) and particle swarm optimisation (PSO), [Tindle et al. (2003), Tindle et al. (2005), Turner et al. (2006)].

Results from this work indicated that as the complexity of the problems under consideration increased (a problem with more than one thousand variables), the time required to obtain a good solution became unacceptable when using a single processor workstation. The sponsor of the work required a near real time solution in less than five minutes. To allow more complex problems to be addressed in an acceptable time frame, it was apparent that it would be necessary to use a parallel or cluster computer. However, in order maximise its usage, from the outset it was determined that the cluster should be a general purpose machine and not designed for a specific area of application [Foster et al. (2001), Foster et al. (2002)].

2. BACKGROUND

A number of factors were considered important during the creation of the system specification:

- The system should be designed to process real world problems from domains such as engineering and media.
- The system should be capable of running in an open access area without the need for special cooling or air-conditioning.
- The system should be built from standard low cost computing components based upon X86 technology.
- The cluster network should be flexible and programmable using industry standard components based upon Gigabit Ethernet.
- Commodity off-the-shelf components should be used to ensure that the system is inexpensive to maintain and can be managed using mostly existing in-house skills.
- The system should be sufficiently quiet so that soundproofing is not required.
- The system should be designed to minimise the consumption of electrical energy.
- The cluster should employ a dual boot system capable of booting the Compute Nodes into a Windows or Linux Operating System in any required proportion, for example, 60% Windows and 40% Linux. This enhances the versatility of the system and facilitates research into computer networking.

A long detailed specification (a fifty page document) for the DCET Cluster Computer was created by the authors. After the tendering process, Dell proved to be the only approved vendor willing to supply computing and network equipment to satisfy the unique design requirements of the specification. A total of seven potential vendors were initially given the opportunity of tendering for the project.

To produce the system design the authors and engineers from Dell UK and Cisco Systems (San Jose) all collaborated to produce the final system design. To the knowledge of the

authors, the resulting computer cluster is the first system of its type in the world to be delivered and commissioned.

The Dell Configuration Calculator was used during the design process. This tool allows the designer to determine the power consumption and acoustic noise generated for different system configurations, for the whole cluster, in a single model. The calculator is unique; no other approved vendor offers a similar design tool.

The electrical system load of the DCET cluster is 21kW when running at full load. An advantage of this cluster is that the electrical load varies in proportion to the CPU computational load.

Close liaison between the Department of Computing, Engineering and Technology (DCET) and the University Estates Department was required to determine a suitable location for the cluster. The overriding consideration was a new policy from the Estates Department not to install new large computer systems in air-conditioned machine rooms because of wasted heat and high running costs. The David Goldman Informatics Centre, at the University of Sunderland, is a relatively new open-plan building. At its core are the Computing Terraces which are housed in a large open-access atrium approximately 150 meters by 60 meters and four storeys high. This area proved to be the only available space large enough to accommodate the cluster. A small section of the Computing Terraces (8 metres by 8 metres) has been partitioned off to provide secure, lockable accommodation for the cluster (Fig. 1).

Fig. 1. DCET Cluster Computer

To ensure that the heat generated by the cluster on the Computing Terraces did not significantly alter the air flow in the building, a finite element analysis was carried out by external consultants. As anticipated, their report indicated that the computer controlled ducted air flow system would still operate normally when the cluster was installed. The heat energy given off by the cluster is used to provide background heating and is not vented to the external atmosphere.

In the area where the cluster system is situated ten workstations have been connected to the cluster networks. Researchers are able to control the cluster from any one of the workstations to run experiments. To view the output generated by the cluster, four high definition LCD display screens have been mounted high up on a wall. In many cases this output is in the form of large numerical data files. To visualise this in a concise format, colour graphics are often deployed to display the data as bar graphs and icons.

As the cluster significantly reduces energy loss caused by heat dissipation, the design has attracted considerable attention. In some articles in the press, it has been given the nickname the ‘green grid’. A Case Study about the design of the cluster has been carried out by Dell UK (2008).

Since it was designed as a general purpose system, the cluster may be used by researchers working in various disciplines, for example: (i) media researchers may use the system for 3D computer graphics rendering; (ii) engineering researchers for problem solving in fields such as structural mechanics and computational fluid dynamics; (iii) software engineers may use the cluster to research web based systems and Internet search engines; (iv) network researchers may use the system to investigate multi channel standard and high definition video streaming. Potentially, the cluster may be deployed in many more areas beyond those listed. While it will be employed primarily for research work, the possibility for use in commercial activities also exists.

The DCET has close working links with Cisco Systems and operates a Cisco Network Academy. Consequently, a flexible network architecture was designed to facilitate network experimentation. The central switch may be programmed to provide multiple VLANs (virtual local area networks) and link rate limiting.

Furthermore, the cluster also provides VPN (virtual private network) support that allows users to gain secure remote access to the system to start and stop the processes and schedule tasks. It is therefore possible for collaborating researchers to gain access to this powerful computing resource from any location in the world.

DCET has run a thriving Masters course in Network Systems for several years, the aim of which is to provide students with as much hands-on practical experience of networking as possible. As a consequence, many Masters as well as PhD students have been given access to the cluster computer facility.

In addition, DCET is a Microsoft Training Academy. For the purposes of this paper, therefore, the focus will be primarily upon Microsoft Windows 2003 High Performance Computing (HPC).

3. THE NETWORK

Windows Server 2003 Compute Cluster Edition supports five different network topologies with one, two or three network interface cards (NICs) in each Compute Node. The cluster is based upon this three network interface model, Microsoft (2006).

The cluster is primarily a Linux based system. A user is able to control processes and operations on the cluster via a Scali (Scalable Linux) management system.

The performance of a cluster computer is directly influenced by the architecture and bandwidth of the associated network. The cluster employs three main Class C networks, namely the Data, IPC and IPMI networks:

Data: The data bus manages the traffic generated by user applications. Effectively, it is an Intranet that is not connected to the main Campus Network. A gateway and firewall provide access to the Campus Network.

IPC: The IPC (Interprocessor Communication) bus provides a communications link between the Operating Systems running on the cluster Compute Nodes. In addition, it may be used during system upgrades to load Operating System images on to Compute Nodes.

IPMI: The IPMI (Intelligent Platform Management Interface) bus is used to control and monitor the operation of the cluster. It is used to monitor speed, CPU temperatures and the start up and shutdown of the Compute Nodes.

The physical design of the cluster incorporated a reduction in the packing density of Compute Nodes by a factor of two, where every other rack slot remains empty. The purpose of this strategy is to reduce the power density, CPU operating temperatures and consequently the cooling fan speeds. Acoustic noise, primarily caused by the cooling fans, meets the standards set by the Health and Safety requirements for the university.

The Compute Nodes are rack based units of 2U height. All Compute Nodes have six low noise, large diameter motorised fans. The low power density ensures that the fans run at a low speed and do not generate too much noise.

The Baseboard Management Controller (BMC) is a specialised microcontroller embedded on the motherboard of Compute Nodes. In the cluster, the BMC provides the intelligence in the Intelligent Platform Management Interface (IPMI) architecture. The BMC manages the interface between the cluster system management software and platform hardware.

A number of sensors built into the Computer Nodes monitor parameters such as temperature, cooling fan speeds, power mode and Operating System status. The BMC monitors the sensors and can send alerts to the cluster administrator via the network if any measured parameters do not remain within preset limits, indicating a potential failure of the system. The administrator can also communicate with the BMC using a remote desktop protocol (RDP) to take corrective action, for example, to reset or power cycle a Compute Node that has crashed. The BMC normally communicates with a BMC management utility (BMU) on a remote client using IPMI protocols.

3.1 Head Nodes

In the cluster there are two Head Nodes. The Windows Head Node runs Win2003 Server CCE. A program named the Grid Administrator is used to manage the Windows Compute Nodes. The Grid Administrator can send control commands to selected groups of Windows Compute Nodes.

The Linux Head Node runs Scientific Linux which is a version of Linux supported by CERN. The Scali management system operates on Scientific Linux. A Scali management program named the Parallel Shell is used to manage the Linux Compute Nodes. The Parallel Shell can send control commands to selected groups of Linux Compute Nodes.

A PXE (Preboot eXecution Environment) boot system is used to start the Operating System installed on selected groups of Compute Nodes. To start up and run the Compute Nodes in Windows, the following sequence is necessary:

- Parallel Shell - select group of Compute Nodes
- Parallel Shell - boot selected Compute Nodes into Scientific Linux
- Parallel Shell - run a script in Scientific Linux and set next boot into Windows
- Parallel Shell - power cycle selected group of Compute Nodes
- Grid administrator - run Windows applications

With the current arrangement, if the Linux Head Node crashes then it is not possible to run and control the Windows system. In the near future the authors intend to devise an improved start up boot sequence to allow the Windows and Linux systems to be controlled separately.

The Windows Head Node also acts as a domain controller for the cluster. In the domain controller DNS (domain name system) forwarding is configured to send requests to the main Campus DNS server. In addition, a DHCP (dynamic host configuration protocol) server is also configured on the domain controller.

A recent design note from Microsoft recommended that the Head Node and domain controller should be installed on separate nodes to reduce system complexity and CPU load. In the future the authors intend to reallocate these two functions to different nodes.

In normal operation the Grid Administrator running in the Head Nodes is used to start and stop worker processes in the Compute Nodes. Large applications that can run in parallel are normally executed on one of the workstations. The main task in the application is broken down into smaller sized jobs and placed into an embedded job scheduler. The scheduler manages the execution of jobs by the worker nodes.

3.2 Central Switch

As described in a typical cluster it is normal practice to break down a job into a number of tasks of equal size and allocate them to the Compute Nodes. However, a common problem is that a communications bottleneck can occur when all of the Compute Nodes are ready to write their data to a shared storage device at the same time. In the DCET cluster, a very high bandwidth programmable switch has been installed to help eliminate this potential bottleneck.

The Cisco Catalyst 6500/Cisco 7600 Series Supervisor Engine 720 integrates a high-performance 720 Gbps crossbar switch fabric with a forwarding engine in a single module, delivering 40 Gbps of switching capacity per slot (4*48-port

Gigabit Ethernet density line cards). With hardware-enabled forwarding for IPv4, the system performance is capable of 400 Mpps for IPv4.

At present the main switch in the cluster provides 192 (4*48) network ports and the system can support a total of 385 ports, with additional line cards. With regard to future expansion of the system the network capacity is not expected to be a limiting factor. In the cluster area there is space to accommodate about two additional cabinets each with ten Compute Nodes. However, there may be insufficient duct space to accommodate the extra Gigabit Ethernet cables. More detailed technical information relating to the cluster components and network is given in Appendix A.

3.3 Direct Attach Storage (DAS)

A DAS system is connected to the Head Nodes to provide a large shared control storage facility for both the Linux and Windows systems. Disk storage is split nearly evenly between both Operating Systems. The bandwidth of the DAS network connection is 12Gbps for each operating system.

3.4 Compute Nodes

In the cluster, there are a total of forty Compute Nodes, each with two dual core CPUs. This gives a total of (40 x 4) 160 CPU cores. When compared to a computer with a similar single processor CPU, the theoretical speedup is of the order 160. To put this into perspective, a task that requires one hour to run on the cluster will require about one week to run on a single processor CPU computer.

3.5 Updates

The management of the cluster has proved to be a very time-consuming task, particularly in relation to updates. It is necessary to update all nodes at regular intervals to ensure that the system is not corrupted by a malicious virus. In the near future the authors aim to install a local update server to manage this process in a more orderly manner.

At certain times it is necessary to install a new Operating System image on all Compute Nodes. An automated process is normally deployed to update the cluster; however, to fully complete this task can take a whole day.

4. VIRTUAL COMPUTING

The authors are currently experimenting with the introduction of virtual computing methods to aid the management of the system. From the outset the cluster nodes were allocated a relatively large amount of memory to facilitate the application of virtual computing methods. All cluster nodes have 8GB of RAM memory installed and four CPU cores. It is anticipated that Windows 2008 HPC will be installed on all nodes in the near future. The introduction of virtual computing methods will simplify the update process and allow users to install Operating Systems preconfigured for their own particular needs.

Recent past experience has shown that only supported applications should be installed on the cluster. The installation of applications not on the supported list can cause the system to become unstable. Great care is taken therefore with regard to updates and modifications to the cluster.

5. UNIQUE FEATURES OF DCET CLUSTER

The design of the DCET Cluster Computer is not standard. The special features of the system are discussed below in a concise format.

- The system operates in an open access area without the need for air conditioning or special cooling.
- The network system uses a managed (programmed) switch, a Cisco 6509 switch, providing a very high bandwidth.
- The system has a versatile dual boot capability and Windows or Linux Operating Systems may be started in any proportion.
- The main switch provides support for VLANs, VPNs, QoS control, IPv4 and IPv6 switching in hardware and rate limiting to support experimentation.
- The standard designs proposed by other vendors normally employed a set of daisy chained unmanaged switches to minimise costs. In these inflexible network designs the uplink connecting the daisy chain was identified by the authors to be a potential bottleneck.
- The authors selected a managed switch design to provide a flexible high bandwidth network.
- The managed switch design employed in the DCET cluster is very much more costly to implement than the unmanaged alternative design.
- The system employs three Gigabit Ethernet networks (Data, IPMI and IPC). The allocation of roles to the networks was determined by the authors and engineers employed by Dell and Cisco.
- A subset of six Compute Nodes have five network connections, two Infiniband connections (10 Gbps full duplex) plus the three Ethernet networks described above.

6. CONCLUSIONS

The DCET cluster computer has been operating successfully since May 2007. The dual boot facility has been proven to be reliable and users are now able to select between the Windows (Server 2003 CCE) and the Linux (Scientific Linux) Operating Systems to meet their research needs. Consequently, the initial design goal to create a general purpose cluster computer has been achieved in that the system is now used by staff and researchers with varying research interests, as well as Graduate project students.

The cluster has found application in many different areas, for example; animated computer graphics rendering, web services, Java network programming, client server systems, video streaming, parallel genetic algorithms and evolutionary algorithms based upon message passing Java (MPJ). A companion ICSE 2009 paper entitled “Rendering 3D Computer Graphics on a Parallel Computer” describes the application of the DCET cluster computer for 3D computer

graphics rendering. In the future, it is anticipated that the system will be employed increasingly to solve engineering problems such as automotive design based upon computational fluid dynamics and structural design using finite element analysis.

It is the intention of the authors to carry out experiments using virtual computing methods. These techniques will make it possible to dynamically load a version of any Operating System preconfigured to meet the requirements of a particular researcher. In addition, it will facilitate easier management of the cluster with regard to Operating System version control and system updates.

Recently the concept of cloud computing has created considerable interest. This involves the provision of a dynamically scalable computing service, which may also offer virtual computing facilities, over the Internet. Users themselves do not need to understand how the cloud operates in order to take advantage of the services it provides. One of the objectives for the development of the cluster is to establish it as an online computing resource that is reliable, secure and easy to use. Consequently, the authors are investigating methods based upon web services that will allow secure remote access and control of the cluster. The remote user will upload a project, schedule a task and download the results with the minimum involvement of DCET technical support staff. It is expected, therefore, that this facility will be particularly useful for collaborative research projects.

For interest, a companion ICSE 2009 paper entitled "Rendering 3D Computer Graphics on a Parallel Computer", compares the performance of four rendering applications.

ACKNOWLEDGEMENTS

The Science Research Investment Fund (SRIF) is a joint initiative by the UK Office of Science and Technology (OST) and the Department for Education and Skills (DfES). The purpose of SRIF is to contribute to higher education institutions' (HEIs) long-term sustainable research strategies and address past under-investment in research infrastructure. The University of Sunderland Cluster Computer was purchased with the support of the SRIF III fund.

REFERENCES

Dell UK (2008). Eco Friendly Super Computing – Power efficient "green" super computing solution puts British university at the forefront of research, A Case Study by Dell UK, Dell Corporation Ltd. http://www1.euro.dell.com/content/topics/global.aspx/casestudies/en/emea/uk/fy2008_q4_id798?c=uk&cs=RC1105026&l=en&s=pad. (Last accessed 7 July 2009.)

Foster I., Kesselman C., Tuecke S. (2001). The Anatomy of the Grid: Enabling Scalable Virtual Organizations, *International J. Supercomputer Applications*, 15(3), 2001.

Foster I., Kesselman C., Nick J., Tuecke S. (2002). The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, *Global Grid Forum*. 22 June 2002. The Globus Alliance, http://www.globus.org/alliance/publications/papers.php. (Last accessed 7 July 2009.)

Microsoft Corporation (2006). Getting Started with Compute Cluster 2003, Appendix 2 Network Requirements. http://technet.microsoft.com/en-us/library/cc720109(WS.10).aspx#BKMK_Appendix2. (Last accessed 7 July, 2009.)

Tindle S.J., Fletcher I., Mellis .J, Mortimore D., Tann P., Tindle J. (2003). Automated Planning for Broadband Passive Optical Networks, In Burnham K.J. and Haas O.C.L. (eds.) *Proceedings of 16th International Conference on Systems Engineering*, Vol. 2, pp. 703-707, 9-11September 2003, Coventry University, ISBN: 0905949919.

Tindle S.J., Tindle J., Fletcher I., Turner S., Mortimore D. (2005). Network Planning: Optimisation of Equipment Deployment in Broadband Optical Networks, In Mitchell J.E. and Faulkner D.W. (eds.) *Proceedings of 10th European Conference on Networks and Optical Communications*, 5-7 July 2005, University College London, ISBN: 0953886387.

Turner S., Tindle J., Tindle S.J., Mellis J., Fletcher I., Mortimore D. (2006). Planning of Complex Industrial Systems using a Novel Parallel Genetic Algorithm, In Burnham K.J. and Haas O.C.L. (eds.) *Proceedings of 18th International Conference on Systems Engineering*, pp. 493-498, 5-7 September 2006, Coventry University, ISBN: 1846000130.

APPENDIX A. DCET CLUSTER SPECIFICATION

DCET Cluster Computer Head and Compute Nodes

- Number of Compute Nodes: 40
- Head nodes (Linux/Windows): 2
- Compute Nodes based upon Dell Server type 2950
- Head node (i) boots Scientific Linux
- Head node (ii) boots Win2003 Server HPC version (64bit)
- Central processor unit Xeon 5100 64bit, 2.66GHz, 4Mbyte L2 cache per processor
- Intel Greencreek chipset FSB1333MHz
- Number of CPU processor per node: 2
- Number of CPU cores per node: 4
- Total number of compute cores in the cluster: 160
- RAM per node: 8Gbyte, per CPU core: 2Gbyte
- RAM type FBD DIMM ECC FSB 1333MHz
- Number of network interface cards NICs per node: 3
- Number of network (i) data, (ii) control and (iii) monitoring: 3
- Network bandwidth Gigabit Ethernet 1Gbps using copper cable
- Support for TCP off load engine TOE
- Data storage SATA drive, 250Gbyte per node
- Total distributed storage 40 * 250Gb, 10Tb
- VLAN support for three buses, Data, IPC and IPMI
- NICs are allocated static IP addresses

Main Switch Unit

- Main switch type Cisco 6509
- The Cisco Catalyst 6500/Cisco 7600 Series Supervisor Engine 720 integrates a high-performance 720 Gbps crossbar switch fabric with a forwarding engine in a single module, delivering 40 Gbps of switching capacity per slot (4*48-port Gigabit Ethernet density line cards). With hardware-enabled forwarding for IPv4, the system performance is capable of 400 Mpps for IPv4.
- Cisco 7600 controller: 1
- Cisco 720 switches: 2
- Number of line cards: 4
- Line card type 48 port 10/100/1000Mbps
- Total number of 1Gbps ports: 192
- Expandable up to 285 ports
- Content services, firewall, NAT, intrusion detection, IPSec/VPN, network analysis, and SSL acceleration, MPLS and QoS
- Support for GRE and VLAN trunking

Infiniband Overlay

- Number of nodes: 6
- Number of CPU cores: 24
- Host channel adapters HCA per node: 2
- HCA interface PCI Express
- HCA bandwidth 10Bbps
- Cisco SFS 7000P InfiniBand Server Switch: 1
- Nonblocking cross-sectional bandwidth 10Gbps, port to port latency less than 200 nanoseconds

Central Data Storage

- Central storage 15 * 500Gbyte disk, 8 Linux and 7 Windows
- Total central storage on Dell MD1000: 7.5Tbyte
- Bandwidth 2 * 12Gbps
- SAS harddrives 3.0Gbps at 15000rpm

Operating Systems

- All Compute Nodes support dual boot
- Compute nodes can boot Linux or Win2003 server in any proportion

Control Software

- Scali Manage provides support for managing the cluster, monitoring performance and controlling services
- Linux Parallel Shell
- Windows Grid Administrator

Web Servers

- Number of web servers providing access and support for users: 5
- A total of five web server provide various web services
- Virtual private networks (VPN) for remote access
- Gateways and firewall, FTP server for secure file transfer
- Server type Dell 1950

Peak Performance

- 1.7024 Tflops