Construction of the MCIL Cyberinfrastructure Lab

Michael Rokitka
Department of Computer Science and Engineering
201 Bell Hall
University at Buffalo, The State University of New York
Buffalo, NY 14260-2000
mrokitka@cse.buffalo.edu
April 18, 2008

Our lab consists of the following hardware:

- Intel Xeon Cluster
  - Head Node: Dell PE1950, with two Dual Core Xeon Processors 5148LV, each with 4MB Cache, 2.33GHz 1333MHz FSB, 16 GB Memory; 146 GB Disk
  - Worker Nodes: Eight Dell PE1950s, each with two Quad Core Xeon E5430 Processors, 2x6MB Cache 2.66GHz, 1333MHz FSB; 160 GB Disk
- AMD Cluster
  - Head Node: Dell PE1950, with two Dual Core Xeon Processors 5148LV, each with 4MB Cache, 2.33GHz 1333MHz FSB, 16 GB Memory; 146 GB Disk
  - Worker Nodes: Eight Dell PowerEdge SC1435s, each with two Dual Core Opteron Processors, 2x1MB Cache, 1.8GHz 1Ghz HyperTransport; 160GB Disk
- Virtual Machine
  - Head Node: Dell PE1950, with two Dual Core Xeon Processors 5148LV, each with 4MB Cache, 2.33GHz 1333MHz FSB, 16 GB Memory; 146 GB Disk
  - VM Nodes: Two PowerEdge R900s, each with 4 quad core X7350 Xeon, 2.93GHz, 8M Cache, 64GB Memory, and 2x300GB 15K SATA Drives
- Storage
  - Three storage systems, each consisting of a Dell NX 1950 with Quad Core Xeon E5430 Processor2x6MB Cache, 2.66GHz and a Dell PowerVault MD3000 external RAID array with 11 TB of Disk. One node contains an expansion using a Dell PowerVault MD1000 (approx. 4TB).
- Networking
  - The clusters are interconnected with both GigE (Dell PowerConnect 6248, 48 GbE PortManaged Switch, 2xDell PowerConnect 3424 24 Port FE with 2 GbE Copper Ports and 2 GbE Fiber SFP Ports) and Infiniband (Dell 24-Port Internally Managed 9024 DDR InfiniBand Edge Switch) switches and cards.
- Condor Flock: 35 PCs
  - 10 Lenovo 3000 J200 Type 9690, Celeron 420, 1GB, 80GB Disk
  - 15 Lenovo 3000 J200 Type 9690, Celeron 420, 2GB, 80GB Disk
  - 10 Lenovo ThinkCenter A61e Type 6417, AMD S LE 1150, 1GB, 80GB Disk
- KVM Switch
- Printer
  - Networked hp LaserJet 4250DTN duplex printer
- Student Workstations
  - 5 Dell workstations

The intent of this lab is to provide an environment for experimentation and learning to advance innovation in cyberinfrastructure. With physical access to the hardware students can simulate

interruptions of service in order to evaluate and test how various middleware reacts under these conditions. Based upon this information, students will be able to devise new strategies for designing and implementing middleware solutions which can detect and recovery appropriately from service interruptions.

In the following sections the steps needed to setup and configure our lab will be described. This information will prove invaluable to future students and others constructing such a system for the first time.

Cluster installation begins with the initial hardware setup and configuration. Racks to house the cluster must be assembled and all units which will reside within the rack must be railed and inserted into the rack. Any additional hardware which needs to be added in to the head node(s) or compute nodes should be done at this point. Additional GigE cards or Infiniband cards fall into this category. The KVM switch should be placed in a central location so that it is accessible by the hardware needing user interaction.

Setup of physical networking devices should be performed next. GigE cabling between cluster nodes and switches should be performed. All cables should be clearly labeled and tied together to prevent entanglement and future confusion. The cluster head node should be connected to the external network so that it accessible via the outside. Compute nodes should be connected to the internal network only. Access to compute nodes should be limited to direct physical access within the lab and remote access via SSH from the head node. Managed switches may need to be configured to operate correctly within the environment, as they typically are setup to block the DHCP broadcasts that the head node uses to establish communication and setup of compute nodes. This may involve direct serial connection to the switch and manipulation via VT100 emulation software. For details on switch configuration refer to the manufacturer's manual [1]. Infiniband cabling should also be performed at this point. All compute nodes should be connected via an Infiniband Fibre Channel switch [2] to ensure a fast interconnect for inter-worker communications. At this point you can setup all the power cabling and prepare to startup the system.

Cluster software install begins with the install of the ROCKS [3] cluster distribution software on the head node. You should obtain rocks and burn it onto DVD (or CDs) for installation on the head node. The jumbo roll DVD is preferred since it is the most comprehensive. Follow the instructions in the User Guide [4] for installing the head node. Before beginning the install you should collect all relevant networking information for the head node including: internal IP address (10.X.X.X), external IP address (128.205.X.X), net masks, gateway, DNS server, and NTP server. Please note, ethernet interface 0 MUST be the one connected to the internal network and interface 1 MUST be the one connected to the external network. Once ROCKS has been installed on the head node you can continue to install on the compute nodes. The insert-ethers command should be issued on the head node to signal that it is ready to detect and setup the compute nodes. While insert-ethers is running on the head node, you should boot the compute nodes in PXE boot mode to perform a network install or insert the ROCKS DVD/CDs to begin the install. Note that you should begin with the bottom compute node in the rack and work your way up one at a time with the compute node installs. See the man page for insert-ethers or docs online

for details on configuration settings. You may want to use the --rack and --baseip arguments to avoid compute node naming collisions and issues with managed switches. Please refer to the ROCKS documentation for the ROCKS command list. The CLI commands provided by ROCKS allows you to manage your cluster configuration via the head node without needing to manually interact with individual compute nodes.

Once ROCKS has been successfully installed on the cluster, the drivers and software suite for Infiniband should be installed. This will provide access to a speedy interconnect between the compute nodes, as well as management and visualization software for working with the Infiniband network.

Once the cluster itself has been installed and configured the storage arrays should be setup. This process begins with the cabling of controller heads and the modular disk subsystems. Controller heads for production systems should be connected to the external network via GigE. The controller heads for the internal (unstable) systems should be connected to the internal GigE network only. This will allow the cluster to communicate with the storage systems. Connections between the controller head and modular disk subsystem should be made using the Serial Attached SCSI cables. Depending on the topology you want you can set up the system with multiple controller heads per disk subsystem to provide clustering for failover. Demos of the setup process are available on Dell's website [5]. Note that you may need to make changes to the BIOS and/or SAS configuration utility upon bootup to ensure that the Serial Attached SCSI interfaces are enabled. The steps that had to be followed for our machines were to hit Ctrl+C on bootup to enter the SAS configuration utility to enable the SAS controller. After this was completed, BIOS settings had to be changed via F2 keypress on bootup to auto-detect the SAS devices. Once hardware setup is completed, OS and system configuration must be performed. The Dell PowerVault NX1950 integrated solution ships with Windows Storage Server 2003. Upon system boot you will need to enter networking information by setting static IP addresses, DNS servers, etc. in the TCP/IP configuration of the network interfaces used. Drivers for the MD3000 and the storage system software suite should be installed next for managing and interacting with the storage system.

After all basic hardware and initial software has been setup and configured properly, focus should be placed on setting up and configuring middleware software for use in the distributed environment. The majority of this will be installed on the cluster with ROCKS (depending on which rolls you select), but will still need configuration and setup. These systems include: Globus, Ganglia, PVFS2 (Parallel Virtual File System), Sun Grid Engine, etc. It is recommended that users refer to the roll users guides for ROCKS for information on initial setup and configuration of these systems.

*A UPS is highly recommended for power redundancy and failover for the system.

References

1. Dell PowerConnect 5324 switch documentation
   http://docs.us.dell.com/support/edocs/network/pc5324/en/index.htm
2. Infiniband Trade Association
   http://www.infinibandta.org/
3. ROCKS Cluster Distribution
   http://www.rocksclusters.org/
4. ROCKS User Guide (v4.3)
   http://www.rocksclusters.org/rocks-documentation/4.3/
5. Dell PowerVault NX1950
   http://www.dell.com/content/topics/topic.aspx/global/products/pvaul/topics/en/pv_nx1950_landing?c=us&l=en&s=gen&dgc=IR&cid=14054&lid=400426