Guidelines for
Building a High-Performance Computing Cluster

This article provides guidelines for planning an efficient, cost-effective HPC cluster based on Dell™ PowerEdge™ servers, including recommendations for management and Message Passing Interface (MPI) software.

BY HÅKON BUGGE

High-performance computing (HPC) clusters have become increasingly popular in a wide range of industries. IT organizations considering implementing a Dell™ server–based HPC cluster can benefit from understanding the steps involved in this task. This article describes these steps and explains how to determine the appropriate cluster configuration and management software to help maximize performance and lower total cost of ownership (TCO).

Considering physical requirements for building an HPC cluster

The first step in building an HPC cluster—before choosing the hardware and software—is planning the physical environment. An HPC environment must support the cluster, while leaving room to scale out as necessary. The space required varies with the number of processors. For example, rack-mounted systems can house a 128-CPU cluster in approximately the same space as that required for a large UNIX®-based server that typically accommodates 32 processors. Weight is also important; for example, a 500-node cluster requires a reinforced floor. Other considerations include cooling requirements, power sources, and backup.

Choosing the hardware architecture

The appropriate processor architecture for a cluster depends on performance requirements and budget. Traditional clustering software runs only on a homogeneous architecture. However, some communication and management middleware—such as Scali Manage™ and Scali MPI Connect™ software—can accommodate hardware heterogeneity. Using more flexible software allows administrators to design clusters with a mix of node architectures that best meet requirements.

Heterogeneity permits greater design flexibility and scalability. IT organizations can also realize cost savings from the ability to use legacy technology. A heterogeneous cluster can be designed to handle various tasks with optimized performance, such as pre- and post-processing. Additionally, administrators can replace failed nodes with the latest technology, yielding more power at or below the original cost. The ability to have heterogeneous architectures in a single cluster also allows joining two separate clusters into a single large unit for increased processing power.

To select an appropriate node architecture, consider the number of units required and how fast a single node should be able to execute its processing tasks. Choosing 64-bit rather than 32-bit architecture provides more
processing power. This decision should be made considering the applications that will run on the cluster and the cluster’s performance requirements and scale-out characteristics. For example, as administrators scale out an application, adding 32-bit servers may provide increasingly less incremental processing, so the more powerful 64-bit architecture could be a better choice.

Dell offers several core 32- and 64-bit HPC platforms. For more details and server configuration information, visit http://www.dell.com/hpcc.

Factory configuration streamlines cluster installation
Configuration is another hardware-related consideration. For cluster deployment, the boot order often must be changed to instruct the node to boot from the network rather than a CD or disk drive. Administrators can establish the BIOS setting in firmware or they can order a custom configuration from Dell to save setup time.

During system configuration, Dell also can set the BIOS option for enabling or disabling Intel Hyper-Threading Technology in the Dell PowerEdge™ 1750 and PowerEdge 650 to better meet application requirements. In Hyper-Threading, one physical CPU appears as two logical processors. Because this arrangement might require application licenses for twice the number of CPUs, and because Hyper-Threading seldom delivers a performance boost of more than about 30 percent, disabling Hyper-Threading can be more cost-effective if the application cannot take advantage of it.

Ultimately, hardware TCO will largely be determined by non-technical issues such as service and support, so consider a vendor that provides a single point of contact for support and has proven competency in the HPC arena.

Determining which interconnect to use
Nodes use interconnects to function as a cluster. Available products range from legacy interconnects such as Gigabit Ethernet¹ to the more advanced interconnects such as Myricom® Myrinet®. Gigabit Ethernet is less expensive, but Myrinet offers higher performance, lower latency, and better scalability. Organizations often run multiple applications on a single cluster, making the choice of interconnects a complex one without a single right answer.

For example, a single-processor PowerEdge 650 is suitable for Gigabit Ethernet. A dual-processor PowerEdge 1750 offers more computational power, suggesting the choice of a faster interconnect. A cluster of PowerEdge 3250 servers may work best with a fast, low-latency interconnect to leverage the computing power of the Intel® Itanium® 2 processor and prevent the communication channel from becoming a bottleneck.

Another factor influencing interconnects is the Message Passing Interface (MPI) software (see “Resolving cluster communication software issues” later in this article). Scali MPI Connect can simplify cluster interconnect decisions because it handles both legacy and advanced protocols and standards: TCP/IP, Gigabit Ethernet with Direct Ethernet Transport (DET), Myrinet, InfiniBand™, and Scalable Coherent Interface (SCI).

Selecting the operating system
Open source software is increasingly available with professional support and maintenance. Therefore, more companies are using commercial versions of the Linux® operating system, such as Red Hat® Linux, which ships with Dell HPC hardware. Choosing a

SOFTWARE SIMPLIFIES MANAGING AND RUNNING DELL SERVER–BASED CLUSTERS

Dell provides turnkey HPC cluster configurations that incorporate powerful cluster management and interconnect software—such as products from Scali, a provider of high-performance clustering software.

Scali Manage. This software includes comprehensive tools for cluster installation, configuration, management, and monitoring, and supports the leading cluster interconnects and platforms. Scali Manage includes the following features:

- Central system for managing cross-enterprise cluster resources
- Platform-independent cluster management
- Advanced productivity tools for effective maintenance
- Ease and speed of cluster installation, configuration, and expansion
- Ability to manage both small and large systems
- Secure architecture

Scali MPI Connect. This software offers an interconnect-independent, integrated architecture that allows interaction with a single MPI implementation. Third-party applications need to be compiled only once to run on the leading interconnects. Binary programs linked with Scali MPI Connect can run on any of the supported interconnects without recompilation or relinking. Other features include:

- High-bandwidth, low-latency performance
- Advanced application debugging, tuning, and optimization
- Dynamic binding to operator-selected interconnects
- High reliability and system scalability

¹ This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.
professionally developed and maintained operating system rather than freeware can help keep TCO low by sparing system administrators time-consuming manual maintenance and troubleshooting tasks.

**Resolving cluster communication software issues**

The message-passing middleware layer encapsulates the complexity of the underlying communication mechanism and shields the application from different methods of basic communication. Today, MPI has become the de facto standard for message passing. Although MPI is commonly used for parallel applications, developers face a significant challenge: virtually every brand of interconnect requires its own particular implementation of the MPI standard.

Furthermore, most applications are statically linked to the MPI library, which can create the following problems:

- **Conflicts**: If two applications run on the cluster and different versions of MPI link the applications, a conflict might occur.

This inconsistency is solved by having one of the application vendors relink, test, and qualify its application for the other MPI version—a procedure that may be significantly time-consuming.

- **Disconnects**: Evolving demands from applications or errors detected and corrected in the MPI implementation can force one of the applications to use a newer version. In this case, a disconnect between MPI versions and application versions occurs, which again requires an application vendor to relink, test, and qualify its application for the other MPI version.

- **Inflexibility**: Administrators may wish to change MPIs. For example, with a Gigabit Ethernet interconnect, the TCP/IP stack may impose overhead that restricts application scalability. To switch to an MPI that can use leaner, more efficient protocols—such as Remote Direct Memory Access (RDMA)—again requires approaching the application vendors and asking for help with an upgrade or evaluation. This hurdle can deter IT organizations from implementing major improvements afforded by newer, more innovative communications software or interconnect hardware.

Administrators can avoid problems caused by static links to the MPI library by using software that offers dynamic binding between the application and the MPI middleware, and between the MPI middleware and device drivers for various types of interconnects. Offerings such as Scali MPI Connect can help enable administrators to develop the MPI implementation and the application independently of one another, because the application can take advantage of different interconnects or protocols without being changed or relinked.

**Managing systems using the Scali Manage software package**

Management software is the final layer in the cluster stack. IT organizations can benefit from a full-featured management package such as Scali Manage, which includes comprehensive tools for systems installation, configuration, management, and monitoring. In addition, compatibility with leading cluster interconnects and platforms can enable a single management application to support the broad range of systems architectures, nodes, and interconnects that may be dispersed throughout the enterprise. Management software that is designed to be centralized across clusters can help administrators create an integrated computing environment that can reduce costs, increase efficiencies, and improve overall price/performance.

**Scali Manage facilitates Dell HPC cluster administration**

For administrators managing Dell server–based HPC clusters, Scali Manage scales to support both small and large systems. Scali Manage streamlines the installation process with rapid installation of cluster
software, including the operating system and middleware. It offers controlled installation of third-party applications and reinstallation of nodes—for example, when a node has been replaced.

Scali Manage works with the Dell Embedded Remote Access option (ERA/O), which enables remote management of critical servers. With ERA installed on the compute nodes, Scali Manage allows administrators to perform the following functions:

- Power on, power off, or toggle power on single or aggregated compute nodes
- Monitor and connect to the console ports of the nodes individually or collectively
- Replicate commands to several consoles by using a special broadcast window, avoiding the time-consuming and error-prone task of typing the same command several times

Scali Manage is also integrated with the Dell OpenManage™ software suite, which allows administrators to manage, monitor, and control the health of Dell PowerEdge servers from a central or a remote location. Incorporating information from Dell OpenManage into Scali Manage provides both a detailed view of the individual server nodes and a global view of the health and performance of the cluster as a whole.

Operationally, Scali Manage streamlines ongoing activities by providing a single management system for mixed nodes and interconnects; it is a single point of management for one or many clusters. Its easy-to-use, flexible interface enables advanced network and user administration as well as preventive management such as hardware monitoring for early fault detection. Finally, Scali Manage offers out-of-band management and disaster recovery.

Simplifying cluster management and achieving lower TCO
Computing clusters built with products from Dell and Scali have shown real-world success in compute-intensive environments. HPC clusters provide a university consortium with the processing power to analyze large amounts of data cost-effectively, and enable an oil exploration company to achieve an oil strike in half the usual time. For details, see “Exploring the universe with HPC” and “Improving oil exploration using HPC clusters.”

Software components such as message-passing interfaces and management systems are important considerations when building an HPC cluster from the ground up. For today’s IT environments, software like that from Scali simplifies cluster management and operation and helps save time, effort, and money—making administrators more productive and helping to achieve business goals by lowering TCO.

Håkon Bugge (hakon.bugge@scalli.com) is a founder and vice president of product development at Scalli (http://www.scali.com), a provider of management software and MPI middleware for high-performance clustering. Håkon has focused on developing clustering software, interconnect technology, and advanced CPU architectures. He has an M.S. from the University of Oslo, Norway, and has also served as a part-time lecturer in computer architecture at the University Graduate Center (Unik), Kjeller, Norway.