Networking Considerations for Clustered Servers

Reliable, low-cost switches can enhance the network performance of Intel® processor–based Dell™ PowerEdge™ servers in high-availability failover cluster and high-performance computing (HPC) cluster environments. This article provides guidelines that help IT managers choose the most effective network switches for their cluster interconnects.

By Extreme Networks

Given tight IT budgets, data center managers around the world are rapidly implementing low-cost, high-performance clusters of standard computing and networking components to process compute-intensive tasks. By deploying clusters of relatively inexpensive Intel® processor–based Dell™ PowerEdge™ servers to execute mathematically complex models and simulations, enterprises can eliminate expensive symmetric multiprocessing (SMP) systems in all but the most complex applications.

A cluster can be defined loosely as any two or more servers working together in a coordinated fashion to deliver a common service or set of services. Although specific cluster computing and software elements may vary, the network is one critical component all clusters share. This article presents networking considerations that pertain to high-availability failover clusters and scalable high-performance computing (HPC) clusters.

Scalable clusters are further categorized to address loosely coupled HPC applications, which use distributed memory, and tightly coupled HPC applications, which use shared memory.

High-availability clusters: Providing failover protection

Considering the demands of Internet-based applications, IT organizations must keep the enterprise infrastructure constantly working at a consistent, responsive level of performance, especially during reasonable peaks in user load. High-availability clustering was once the exclusive domain of mainframe and minicomputer systems. Now high-availability clusters are an option for most mission-critical business applications running in Microsoft® Windows® and Linux® operating system environments—including e-mail and messaging, databases, customer relationship management (CRM), enterprise resource planning (ERP), and shared file systems.

High-availability failover clusters are designed to improve system reliability and manageability. IT organizations deploy failover clusters to keep end-user applications operational during both planned system events (such as backups, maintenance, and peak user loads) and unplanned events (such as hardware failures and network outages). As part of the fault-tolerant design, each critical system component sends a short message, or heartbeat, over the network at regular intervals to indicate that it is still online.

Should a computing or networking component in the cluster fail to send a heartbeat within the predetermined time, the operating system will direct other servers in the cluster to perform tests that evaluate the health of the system in question. If a critical component has failed,
cluster members will initiate emergency actions to assume the load and resources of the failed node. System administrators typically define the specific policy to be implemented if emergency failover action becomes necessary in the operating system software that manages the clustered applications.

Administrators can base failover policies on one of two system readiness levels:

- **Hot/Warm**: All nodes in the cluster are either in active service or in standby mode. Nodes in standby mode have access to all resources necessary to assume the load of a failed system with no perceptible downtime in end-user applications.
- **Hot/Cold**: Backup systems require some startup time before assuming the load of the failed system. The hot/cold failover policy potentially can lead to an interruption in end-user service.

High-availability failover clusters are generally deployed in server pairs (see Figure 1). For many applications, system administrators can establish straightforward networking requirements that leverage standard TCP/IP transport protocols, basic network interface cards (NICs), and enterprise-grade network switching equipment. Usually administrators connect each server to two different Ethernet switches. This dual-home configuration helps ensure that the failure of one Ethernet switch or Ethernet link will not cause server failover. The dual-home configuration also helps ensure that the failure of one server will not eliminate failover protection for the high-availability configuration. However, should a switch, link, or server fail, only one Ethernet link to each server would be active at a time.

In addition, multiple Ethernet links can be connected between each Ethernet switch and each server. Using the link aggregation technique, administrators can configure separate physical data channels into a logical channel that performs as one higher bandwidth link. Link aggregation enables administrators to provide another level of failure protection by balancing the load to and from each server across multiple Ethernet connections. In this scenario, the server will retain a network connection even if one link fails, without requiring administrators to reconfigure the network. Link aggregation also increases the total throughput of the active server.

**Loosely coupled HPC clusters: Performing massively parallel computing tasks**

IT administrators can configure stand-alone servers, desktop PCs, and rack-mounted blade server nodes in scalable, loosely coupled HPC clusters to process common, compute-intensive applications. Loosely coupled applications use distributed memory and require minimal communication among cluster nodes; generally, synchronization among nodes is not performance-critical. Any large computing task that can be divided into clean, self-contained segments—that is, a massively parallel computing task—works well in a loosely coupled HPC computing cluster (see Figure 2). Examples of loosely coupled applications running in Windows- and Linux-based environments include chip design simulation, financial modeling, scientific research, and computer graphics.

Whenever possible, developers design applications to take advantage of loosely coupled processing. For example, each frame of an animated movie can be represented by a 3-D model that includes colors, textures, and lighting descriptions. A master node server, which manages communication and job scheduling in the HPC cluster, distributes each frame to an individual compute node for processing. When the processing is completed (or in 3-D terms,
“rendered”), the compute node returns the completed frame to the master node and awaits further instructions. Because massively parallel computing tasks can be executed across individual compute nodes with little or no interprocess communication, the number of cluster members that can help process distributed tasks is virtually unlimited. As a result, HPC clusters rank among the most powerful computers in the world, offering combined processing power measured in hundreds of gigaflops.\(^1\) For examples of such clusters, visit the TOP500 list of the world’s most powerful computer systems (http://www.top500.org).

Well-engineered Ethernet switches and NICs can handle most communication tasks in loosely coupled clusters. For large data sets such as those produced by 3-D computer graphics processing, administrators can deploy Gigabit Ethernet\(^2\) and 10 Gigabit Ethernet switches. A 10 Gigabit Ethernet switch is usually configured with the master node, where aggregate responses from all the compute nodes could create a system bottleneck. To ease the load on servers during the transfer of large data sets, administrators should consider using Jumbo Ethernet frames. This option improves throughput and network efficiency by increasing the maximum transmission unit (MTU) of the switch to 9018 bytes per packet, compared to 1518 bytes in 10/100/1000 Mbps Ethernet frames.

As administrators scale out the cluster configuration, additional servers increase the demand on Ethernet switch ports. Network switches must offer wire-speed connectivity without blocking traffic at higher aggregate throughputs. Also, clusters require high-speed trunk connections between switches to ensure that performance does not degrade if traffic passes through multiple switches. To handle the aggregate loads, high-speed trunks require higher capacity connections than the server connections. For example, if servers are connected to Fast Ethernet, the trunks should be Gigabit Ethernet; if servers are connected to Gigabit Ethernet, the trunks should be 10 Gigabit Ethernet. Because this arrangement can be expensive, IT organizations should identify network switches that provide the best possible price/performance.

Web content delivery is particularly well suited to loosely coupled HPC clusters. For Web servers, the equivalent of the master node is a server load-balancing (SLB) network switch.\(^3\)

Loosely coupled high-performance computing (HPC) clusters use server load-balancing (SLB) network switches to distribute load among multiple servers. In a loosely coupled architecture, all servers operate independently of one another, even when serving the same content. For Web servers, the equivalent of the master node is a server load-balancing (SLB) network switch, which distributes load among multiple servers. This approach ensures that no single server承担s too much load, improving overall performance.

Figure 3: Loosely coupled SLB cluster architecture

---

\(^1\) One gigaflop equals 1 billion floating-point operations per second.

\(^2\) 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.

---

Web content delivery is particularly well suited to loosely coupled HPC clusters. For Web servers, the equivalent of the master node is a server load-balancing (SLB) network switch.
Assess server health and load by monitoring critical parameters such as response time, CPU utilization, and available memory.

Tightly coupled HPC clusters: Processing complex mathematical models

Some applications—including advanced weather simulations, econometric modeling, and other types of highly complex mathematical programs—are difficult or impossible to decompose, and therefore must remain tightly coupled, requiring shared memory. The overall performance of such applications is highly sensitive to the latency and throughput of communications among the compute nodes. In the past, such applications were deployed either on expensive, proprietary SMP platforms or on lower-cost Intel processor–based compute nodes interconnected using expensive, proprietary networking equipment. Now, by using 10 Gbps Ethernet switches and network switch–enabled Quality of Service (QoS) techniques, administrators can eliminate proprietary network interconnects.

Key QoS provisions for tightly coupled HPC networks perform the following functions:

- Discriminate between communications that have critical latency requirements and less time-critical traffic
- Implement policies to control the granularity of network latency and throughput for critical traffic

For instance, the “sideways” traffic among CPUs in a tightly coupled system, often represented by remote procedure calls (RPCs), can be prioritized over management traffic, external requests, and bulk transfers of administrative reporting data. Emerging network technology and standards will help enterprises to build tightly coupled HPC environments as economically as they can build loosely coupled HPC clusters today. Soon, standards such as Remote Direct Memory Access (RDMA), combined with Gigabit Ethernet and 10 Gigabit Ethernet switches and sophisticated QoS features, will allow administrators to configure inexpensive compute nodes for most tightly coupled HPC cluster applications, thereby avoiding the high cost of proprietary networking gear.

Improved price/performance: Lowering the cost of enterprise computing

Intel processor–based cluster components can dramatically reduce the costs of deploying both mission-critical business applications and compute-intensive scientific models and simulations. Network switches, essential for high-availability failover clusters and HPC clusters, can help administrators achieve faster throughput, greater reliability, and better scalability.

In concert with Dell PowerEdge servers, Extreme Networks’ scalable Gigabit Ethernet switches can help provide the reliability, features, and price/performance that enterprise environments require—whether deployed in a cluster of high-performance computing servers, load-balancing Web servers, or high-availability applications servers.  

Extreme Networks (http://www.extremenetworks.com) delivers effective applications and services infrastructures by creating networks that are faster, simpler, and more cost-effective. Headquartered in Santa Clara, California, Extreme Networks offers its network switching equipment in more than 50 countries.

For More Information

Dell PowerEdge servers: http://www.dell.com/poweredge
Extreme Networks: http://www.extremenetworks.com