

Designing High-Performance Computing Clusters

High-performance computing (HPC) clusters have evolved from experimental computing architecture to mainstream supercomputing architecture. Today, HPC cluster architecture integrates multiple state-of-the-art, industry-standard components to provide aggregated, cost-effective supercomputing power. This article examines the influence of component selection on cluster performance and manageability.

BY YUNG-CHIN FANG; SAEED IQBAL, PH.D.; AND AMINA SAIFY

Related Categories:

Beowulf clusters

Dell PowerEdge servers

High performance

High-performance computing (HPC)

Industry standards

Intel EM64T processors

Intel IA-32 processors

Memory

Parallel systems

Visit www.dell.com/powersolutions for the complete category index to all articles published in this issue.

The TOP500 supercomputer sites list¹ is a snapshot of worldwide supercomputer implementations. The June 2000 list included 11 high-performance computing (HPC) clusters—which translated to 2 percent of the installations on the list. However, by December 2004, the HPC cluster tally had reached 294 of the 500 installations (see Figure 1), indicating that 58.8 percent of the systems used cluster architectures. This gain shows that cluster computing architecture has rapidly advanced from the status of experimental technology to mainstream supercomputing architecture.²

Each data center has its own specific technical requirements, job distribution patterns, and management schemes. In addition, organizations must consider budget constraints when designing HPC clusters. Application behavior also plays a significant role in HPC cluster design choices. In particular, the main components in an HPC cluster design are the processor architecture, platform architecture, memory subsystem, communication subsystem, storage subsystem, and management framework.

Selecting an appropriate processor architecture

Processor selection is an important performance factor for compute-intensive applications. Selecting a suitable architecture for the applications task can help save procurement costs and enhance computing performance. The three most popular Intel® processor architectures are Intel Architecture 32-bit (IA-32), Intel Extended Memory 64 Technology (EM64T), and Intel Architecture 64-bit (IA-64).

Intel 32-bit processor architecture

The current IA-32 architecture includes Intel Pentium® 4 and Intel Xeon™ processors. IA-32 processors each have different frontside bus (FSB) speeds and cache sizes. Usually, a high FSB clock rate or a large cache can enhance memory performance. For integer-intensive applications, a processor with a fast clock rate typically provides high performance; for floating-point-intensive applications, the Streaming SIMD (single instruction, multiple data) Extensions 2 (SSE2) instruction set and SIMD registers can provide high performance if the code is compiled to utilize these features. Intel Hyper-Threading Technology

¹ The TOP500 supercomputer sites list is available at www.top500.org.

² For more information about the history of HPC clusters, see www.beowulf.org/overview/history.html.

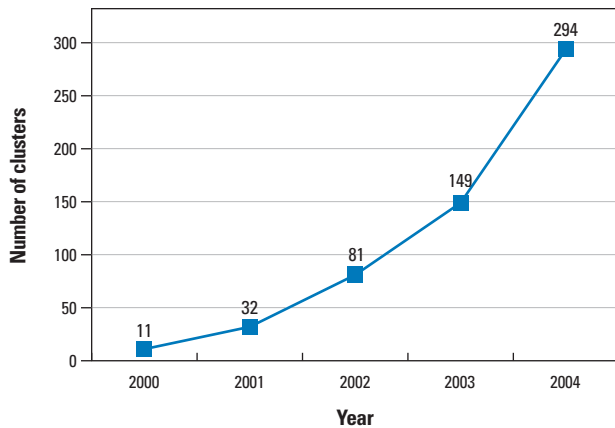


Figure 1. From 2000 to 2004, the proportion of TOP500 supercomputer sites that are HPC cluster-based has grown dramatically

can provide high performance for applications that can schedule threads to execute simultaneously on the logical processors in a physical processor, and it is designed to enable on-chip execution resources such as integer units and floating-point units to be utilized at a high level.

Intel 64-bit extension architecture

Intel Corporation introduced the Intel EM64T architecture in 2004. This 64-bit extension technology is an enhancement to Intel IA-32 architecture. An IA-32 processor equipped with 64-bit extension technology is designed to be compatible with existing IA-32 software; enable future 64-bit extended software to access a larger physical memory space than 32-bit software; and allow coexistence of software written for a 32-bit linear address space with software written for a 64-bit linear address space. In addition, 64-bit extension technology includes the IA-32 extension (IA-32e) operating mode. A processor with 64-bit extension technology is designed to run in either Legacy IA-32 mode or IA-32e mode. The IA-32e mode includes two sub-modes: Compatibility mode, which enables a 64-bit OS to run most existing legacy 32-bit software unmodified; and 64-bit mode, which enables a 64-bit OS to run applications written specifically to access a 64-bit address space.³ Recent Intel Xeon processors feature EM64T, Hyper-Threading Technology, and the Streaming SIMD Extensions 3 (SSE3) instruction set. For applications that require large addressing space, double-precision floating-point calculation, or threading capability, EM64T processors can be suitable candidates.

Intel 64-bit processor architecture

IA-64 is based on the Very Long Instruction Word (VLIW) architecture. It implements the Explicitly Parallel Instruction Computing (EPIC)

instruction set and can issue multiple EPIC instructions in the same cycle to achieve instruction-level parallelism (ILP). Intel Itanium® processors are based on IA-64. Because of architectural and instruction-set differences, IA-32 applications must be recompiled to run on IA-64 architectures. EM64T applications (which are based on IA-32 instructions) are incompatible with IA-64 architectures. However, the Itanium 2 processor includes an IA-32 execution layer, which can support IA-32 applications. For applications that require large memory addressing space, large cache size, intensive double-precision floating-point calculations, and ILP, the Itanium processor can be a suitable choice.

Choosing an effective cluster platform

Cluster architects can use one-, two-, or four-processor servers as building blocks. The main concern for building-block selection is the price/performance trade-off. For example, given comparable processors, a single-processor server is generally less expensive than a two-processor server. If a single-processor server is used for applications that require a high-performance interconnect subsystem such as Myrinet or InfiniBand, the overall interconnect cost can be close to or even exceed the overall server cost. Usually, a two-processor server is well suited for price/performance balance in an HPC cluster configuration. However, for extremely memory-intensive applications, a single-processor server can be a viable choice. For transaction-intensive but not communication-intensive applications, four-processor servers are often suitable.

As an alternative to monolithic rack or tower servers, blade servers can be used for certain applications. A blade is a simplified, single-board server. Usually, multiple blades share a blade server chassis to help reduce cost. The chassis is equipped with a shared power supply; a keyboard, video, mouse (KVM) switch; a network switch; and a pass-through switch. When comparing equivalent processing power in monolithic server versus blade server architectures, a blade server architecture can be designed to require less cooling and can provide higher rack density, a smaller footprint, and enhanced add-on chassis management features. Monolithic servers usually provide more room for expansion slots than single-board server blades, but the computing power for a two-processor server blade is otherwise the same as that of a two-processor monolithic server.

Determining memory subsystem needs

Key factors affecting memory subsystem performance are CPU data bus bandwidth, chip set, memory architecture supported by the chip set, types of memory, and clock rate. In relative terms, a faster FSB can usually provide better bandwidth. Certain chip sets can support one or more types of interleaved memory subsystem; server designers decide which to implement. When FSB bandwidth

³For more information, see *Intel Extended Memory 64 Technology Software Developer's Guide Volume 1 of 2* at www.intel.com/technology/64bitextensions/30083402.pdf.

is matched, two-processor interleaved memory subsystems can be designed to use less-expansive dual in-line memory modules (DIMMs) and provide potentially better bandwidth than one-processor memory subsystems. Cluster architects should select the DIMM type that matches the server specification to help provide good price/performance.

Another memory subsystem consideration is fault tolerance. Standard error-correcting code (ECC) systems are designed to provide for automatic correction when a single data bit is in error and for guaranteed detection of two data bits in error—known as single error correction/double error detection (SEC/DED). However, SEC/DED cannot detect multi-bit errors (more than 2 bits), and this can lead to data integrity problems and system downtime. Another technology used for fault-tolerant memory is IBM Chipkill, which is designed to provide three implementation methods: each data bit of a memory device is included in a separate *ECC word*; more ECC bits are provided than in SEC/DED so that each ECC word can correct a multi-bit failure; or a combination of these two techniques. The preceding implementation methods enable Chipkill to help prevent costly downtime caused by a memory bit error. Choosing the fault-tolerant memory feature and matching memory with server specifications can further help protect against cluster downtime caused by memory error. Usually, memory bit error also can be observed by using the management software utility provided by the platform vendor.

Selecting an appropriate communication subsystem

Many parallel applications that are run on HPC clusters generate frequent communication traffic over the cluster fabric, and the type of interconnect can affect the overall performance of communication bandwidth and time-sensitive applications. The most common interconnect choices are Fast Ethernet, Gigabit⁴ Ethernet, Myricom Myrinet, and InfiniBand. The cost of high-performance interconnects such as Myrinet and InfiniBand is higher than the cost of commodity interconnects such as Fast Ethernet and Gigabit Ethernet.

HPC cluster application behavior can determine which interconnect is the best fit. For distributed and independent jobs or embarrassingly parallel jobs with minimal data communication, Fast Ethernet can be a cost-effective choice. For communication-intensive, bandwidth/latency-sensitive applications, high-performance interconnects such as Myrinet or InfiniBand are appropriate. Data centers that need to run several kinds of applications may require both high-performance and commodity interconnect fabrics.

Because most Dell™ PowerEdge™ servers are equipped with embedded auto-sensing Gigabit Ethernet network ports, the cost difference between Fast Ethernet and Gigabit Ethernet configurations is primarily the network switch. Usually, nonblocking switch

configurations are more expensive but perform better than network switches designed for over-subscription because of differences in the backplane bandwidth.

Finding a suitable storage subsystem

A cluster environment in which data must be distributed among multiple nodes requires special storage technology considerations. Cluster architects have several choices in storage subsystem architectures.

Direct attached storage (DAS) connects storage directly to the master node, and the Network File System (NFS) can be used to serve data to the compute nodes. This structure is suitable for applications with minimal I/O access. Advantages of using NFS are that it is easy to use and does not require special training. The main disadvantage is that the NFS server can become a bottleneck for I/O-intensive applications and can be a single point of failure.

Another option is to serve data in parallel using a storage area network (SAN). Data can be stored in a central location using Fibre Channel storage, and multiple nodes can access data simultaneously via Fibre Channel switches. This approach is suitable when high throughput is required, but it can be expensive to provide data in parallel because Fibre Channel-based SAN components such as host bus adapters, storage devices, and switches can be expensive.

Storage subsystem performance can be improved by using various RAID levels. Several RAID levels (such as RAID-0, RAID-1, RAID-3, and RAID-5) can be used based on the I/O subsystem requirements. RAID-0 implements a striped disk array. This approach can significantly enhance I/O performance because the load is spread across many channels and drives. However, RAID-0 does not provide fault tolerance; failure of one drive will result in data loss. RAID-1 stripes data across multiple disks with mirroring. It provides good read performance and 100 percent data redundancy, but has the highest disk overhead of all RAID types. RAID-3 stripes the data and writes on data disks. Stripe parity is generated on writes, recorded on the parity disks, and checked on reads. RAID-3 provides good performance as well as fault tolerance. RAID-5 writes each data block on disks. Parity for the blocks on the same disk is generated during the writes and checked during the reads. RAID-5

The main components in an HPC cluster design are the processor architecture, platform architecture, memory subsystem, communication subsystem, storage subsystem, and management framework.

⁴This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

also provides good performance and fault tolerance. If a failure occurs, RAID-5 can recover the data using parity, but performance may degrade during the rebuild process. Even if implementing the same RAID level, embedded controllers and add-on controllers can perform differently. RAID controllers that use Peripheral Component Interconnect Express (PCI Express) buses are usually faster than PCI Extended (PCI-X) bus-based RAID controllers.

Determining a well-suited hardware management framework

For a large-scale deployment, remote hardware manageability is necessary to help minimize system downtime and keep maintenance costs low. This capability falls into two categories: in-band management and out-of-band management. In-band management requires the presence of an OS to execute required management features, while

A cluster environment
in which data must be
distributed among multiple
nodes requires special
storage technology
considerations.

out-of-band management is independent of the OS. Several remote hardware management tools can use both in-band and out-of-band management components. A complete HPC cluster management design should include and integrate both in-band and out-of-band management.

In-band management

A centralized management console application, such as Dell OpenManage™ IT Assistant, can run on a dedicated management system to remotely monitor and manage cluster health. All nodes must have an OS-level agent, such as Dell OpenManage Server Administrator, running on them. This type of management system shares the OS network bandwidth. When a node's OS hangs or network difficulty occurs, the management console will assume the affected node is down. Except for the watchdog mechanism, no in-band management feature can be used to recover a hung node. Out-of-band management tools are designed to overcome such drawbacks.

Out-of-band management

An out-of-band management system consists of embedded hardware and firmware and add-on software. The embedded hardware connects to motherboard management buses to monitor and manage node hardware health. The software component, such as a browser, is used to interact remotely with the hardware component, such as the embedded remote controller. The management system can be either an embedded Intelligent Platform Management Interface (IPMI) implementation or an add-on controller such as a Dell Remote Access Controller (DRAC). An add-on out-of-band management system usually is designed to provide more features—such

as virtual media or digitized-graphics/text console redirection over IP—than an embedded management system. For a large-scale HPC cluster, an out-of-band management system is preferred because it can be used to remotely manage a node without the presence of an OS. For example, an out-of-band management system enables system administrators to remotely power cycle a hung node.

Designing an effective HPC cluster

HPC cluster components include processor architecture, platform architecture, memory subsystem, communication subsystem, storage subsystem, and management framework. Significant effort is required to verify the compatibility and performance permutations of these components—and most organizations do not have access to the latest hardware, firmware, and software. Moreover, many organizations do not have a complete understanding about the serviceability of all these components and lack the personnel to verify their compatibility and performance. Therefore, to efficiently and cost-effectively configure and design an HPC cluster, a cluster architect should begin with a standardized configuration in which all components are supported and verified to help protect against compatibility or performance problems. Then, the architect can add or remove components to tailor the HPC cluster to best fit the organization's needs. ↩

Yung-Chin Fang is a senior consultant in the Scalable Systems Group at Dell. He specializes in cyberinfrastructure management and high-performance computing. He participates in open source groups and standards organizations as a Dell representative and has published dozens of technical, conference, and journal papers.

Saeed Iqbal, Ph.D., is a systems engineer and advisor in the Scalable Systems Group at Dell. His current work involves evaluation of resource managers and job schedulers used for standards-based clusters. He is also involved in performance analysis and system design of clusters. Saeed has a Ph.D. in Computer Engineering from The University of Texas at Austin. He has an M.S. in Computer Engineering and a B.S. in Electrical Engineering from the University of Engineering and Technology in Lahore, Pakistan.

Amina Saifi is a member of the Scalable Systems Group at Dell. Amina has a bachelor's degree in Computer Science from Devi Ahilya University (DAVV) in India, and a master's degree in Computer and Information Science from The Ohio State University.

FOR MORE INFORMATION

Dell HPC clusters:
www.dell.com/hpcc