Compaq AlphaServer SC ANNOUNCEMENT

Photograph of LLNL installation

Compaq Tera Cluster
9-29-99

AlphaServer SC new door design

Table of Contents

Executive Overview

Compaq Computer Corporation's new series of supercomputers, the Compaq *AlphaServer* SC Series, is built from commodity off-the-shelf (COTS) components connected together with an ultra-low-latency, high-bandwidth, interconnect from our partner, Quadrics Supercomputer World Ltd. (QSW). The *AlphaServer* SC system is targeted at the high end of the high-performance computing market and will have performance in the multi-TeraFLOP range.

With the first release, customers will be able to connect up to 128 Compaq *AlphaServer* symmetric multiprocessor nodes to form a single system. The *AlphaServer* SC system is designed and engineered to meet the needs of customers who have the most demanding technical computing requirements. The system will satisfy both the capacity and the capability requirements of the high performance supercomputing market.

The Compaq *AlphaServer* SC system uses Compaq's *Tru64™ UNIX™* operating system and additional Compaq *AlphaServer* SC system software components such as the Compaq cluster file system and the Resource Management System from QSW. In addition, the *AlphaServer* SC development software includes Compaq's well-known high-performance compilers and tools.

Product Availability

The Compaq *AlphaServer* SC series will be announced at Supercomputing '99 in Portland Oregon, on November 16, 1999. *AlphaServer* SC systems, with Compaq *AlphaServer* ES40s as nodes, are orderable now. The first American system, a 128-node machine, was shipped in September 1999 to Lawrence Livermore National Labs (LLNL). The first European machine was shipped to the civilian department of the French Atomic Energy Commission (CEA-Civil). Other machines are currently being built and are scheduled for shipment in Q4 CY'99. Volume manufacturing of *AlphaServer* SC V1.0 systems will commence in Q1 CY'00.

Future *AlphaServer* SC systems will add support for Compaq AlphaServer GS series nodes, node counts greater than 128, faster PCI technology, multi-rail SC Interconnects, and check-point restart. Future systems will also include optimizations to the MPI and Shmem libraries.

Compaq AlphaServer SC Series

Product Overview

The *Compaq AlphaServer* SC family of products is derived from Compaq's involvement in the U.S. Department of Energy Accelerated Strategic Computing Initiative (ASCI) PathForward program. The PathForward program was created with the primary goal of develop interconnect technologies which will accelerate the development of balanced 30- to 100-TFlop systems by 2004 using large numbers of Commodity, Off The Shelf SMP servers. The *AlphaServer* SC program has developed the PathForward technology further, from an experimental demonstrator into a system product. This includes the provision of increased levels of usability, manageability, and reliability.

The *AlphaServer* SC series is Compaq's solution for TeraFLOPS computing. The components of AlphaServer SC systems are:

- Current and future generations of *AlphaServer* systems incorporating multiple generations of the *Alpha™* microprocessor
- A scalable, high-speed interconnect derived from the ASCI/PathForward program in partnership with QSW
- Systems integration software that provides Single System Management (SSM) capability, plus development tools and parallel job management
- Flexible software configuration of the *AlphaServer* SC system, so that it can be used in multiple ways:
    - As a scalable, massively parallel processor (MPP) for single distributed applications that currently run on existing MPP machines like the SGI Cray T3D/E or IBM SP-2

- As a high-capacity throughput machine for large numbers of independent sequential, shared-memory parallel, or message-passing jobs
- As a software development platform for parallel applications using shared- or distributed-memory programming paradigms with OpenMP directives, multi-threading, MPI, or Shmem operations

The *AlphaServer* SC system can be operated concurrently in all three ways. *AlphaServer* SC systems are distinguished from traditional MPP systems by having powerful shared-memory multiprocessors in the place of single-CPU processing elements. This allows the customer to solve problems on the *AlphaServer* SC system that could not be performed on MPP machines. The architecture of an *AlphaServer* SC system is illustrated in Figure 1.

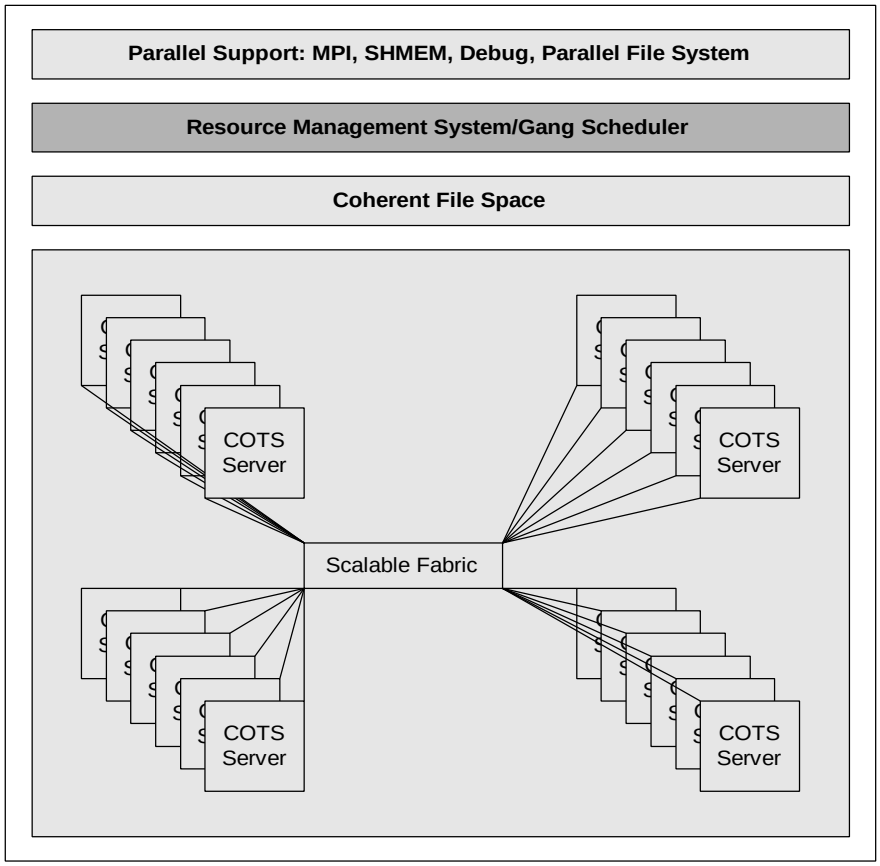
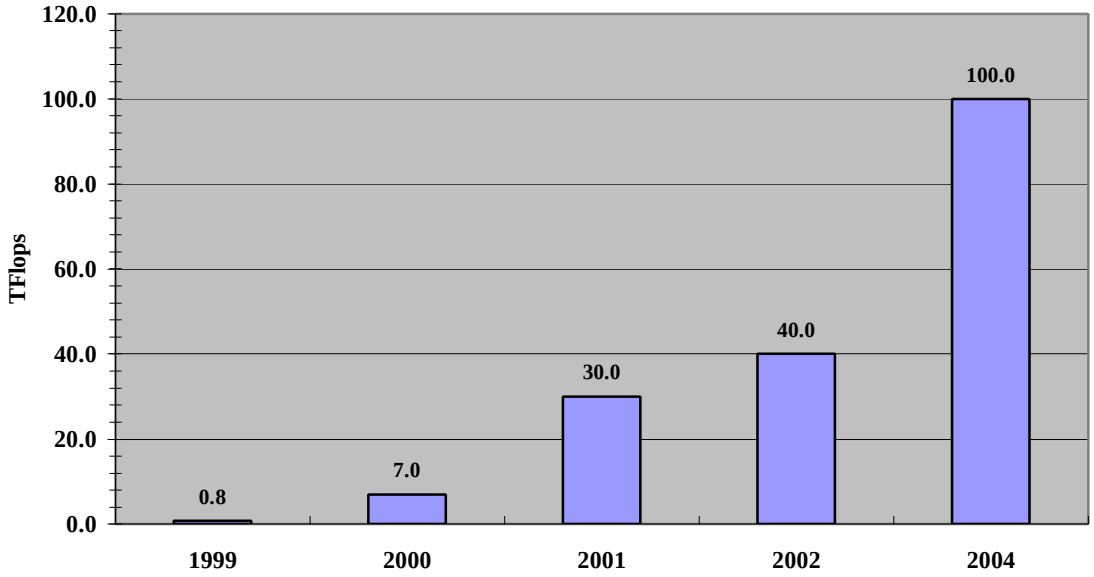


Figure 1: *AlphaServer* SC System Architecture

The Compaq *AlphaServer* SC V1.0 systems are based on *AlphaServer* ES40 nodes, each of which has a peak capability of 5.3 GFlops. Up to 128 nodes can be configured into the *AlphaServer* SC V1.0 system, for a total performance of close to 800 GigaFLOPS. The nodes are connected by *AlphaServer* SC Interconnect , with an internode MPI bandwidth of over 200 MB/s and industry leading MPI latency of under 5.5µ seconds.

The charts below show the increase in power that is derived from the ability to aggregate very large numbers of CPUs into a single manageable and usable system.



Year	Alpha Frequency	Number of SMP CPUs	Number of Nodes	TFlops
1999	667	4	128	0.8
2000	>700	32	128	~7
2001	>1000	64	256	~30
2002	>1200	64	256	~40
2004	~1500	64	256	~100

Figure 2: AlphaServer SC Roadmap

AlphaServer SC systems deliver their enormous computing power through the aggregation of large numbers of SMP nodes. To enable users to harness this power effectively and efficiently, it is necessary to hide elements of the multi-system aggregation from the user. This is achieved by the Single System Management (SSM) capability, provided as part of SC System Software. SC is managed as a single system to users, programmers, and administrators. This capability differentiates SC systems from experimental tera-scale systems. SC systems achieves SSM capability through:

- Leverage of Compaq's enterprise cluster technology

  AlphaServer SC systems incorporate key technologies, such as the Cluster File System (*CFS),* from Compaq's enterprise cluster systems. These systems have evolved from the pioneering VAX Cluster systems through to the current Alpha and *Tru64 UNIX*-based *TruCluster* ™ systems. CFS and its operating system support, provides much of the basis for the SC integrated system management software. By globally sharing the file system directories in a cluster, file system domain, the nodes share system and common configuration files. Management tasks, such as the installation of compilers, only need to be performed once.

- Deployment of Resource Management System

  The SC job management facility, or Resource Management System (RMS) allows the administrator to treat the total set of CPUs as a single large pool, which can then later be subdivided for use by different classes of job. When a user executes a job, RMS is responsible for selecting the set of nodes that that best fits the job's requirements and for creating the requisite processes on each node.

  AlphaServer SC System Software provides optimized implementations of MPI and Shmem. These program protocols are able to directly take advantage of the SC Interconnect's highly efficient communication between processes on different SMP nodes. Direct memory transfers with minimal data movement are used for communication within a node. MPI programs can

be debugged on *SC* systems using the TotalView™ debugger from Etnus Inc. For MPI optimization, message tracing and analysis can be especially revealing. These capabilities can be obtained on the SC systems with Vampir™ and VampirTrace from Pallas GmbH.

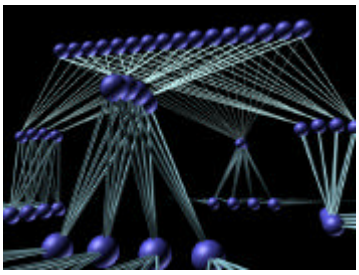
Product Description

AlphaServer SC Interconnect

The *AlphaServer* SC Interconnect (SC Interconnect) is the collection of hardware components from QSW that enables the tightly coupled integration of the SMP nodes into a single system. The SC Interconnect consists of a high-bandwidth crossbar switch and a PCI adapter for each node.

128-way SC Interconnect

The *AlphaServer* SC 16-Port or 128-Port Switch connects the nodes together using a fat-tree routing network. In principal, fully connected crossbars provide the most scalable bisection bandwidth; however, their scalability is effectively limited by their N^2 growth in component count. The SC interconnect technology avoids this limitation by using a fat-tree topology. The fat-tree has the same theoretical bisection bandwidth as a crossbar, but only an N*Log(N) growth in component count and cost. With random traffic patterns, the fat-tree will experience performance that is somewhat less than the theoretical peak. However, the SC Interconnect hardware employs routing techniques to improve throughput for random patterns. Inter-node bandwidths have been measured at rates of over 200 MB/s/rail per node-pair. In future releases, multiple rails of the interconnect are planned on SMP nodes with higher CPU counts. This will allow the ratio of FLOPS to interconnect bandwidth to be maintained for bigger systems.



Artist's rendition of SC Interconnect

The *AlphaServer* SC Elan Adapter is a PCI card installed in each of the host nodes. The adapter employs a virtual DMA capability that allows short or long transfers to be carried out directly by the adapter without host CPU or operating system intervention. It is possible to source or sink data anywhere within the virtual address space of a process. There is no requirement to lock down pages or copy data to intermediate buffers. The adapter supports get and put operations, enabling data retrieval from a remote peer without intervention of the remote system. User-level latencies for remote operations have been measured at less than 3µ seconds.

AlphaServer Nodes

Commodity *AlphaServer* machines running the *Tru64 UNIX* operating system are used as the building blocks for *AlphaServer* SC systems. Access to successive generations of Alpha processors and their associated servers provide increasing microprocessor speed, performance, memory bandwidth and node CPU count. The system nodes run the *Tru64 UNIX* operating system as well as the *Compaq AlphaServer* SC System Software, which provides additional features, required for the management and use of the SC.

Compaq AlphaServer SC System Software Version 1.0 – Summary

Compaq *AlphaServer* SC System Software Version 1.0 provides the utilities and libraries for the operation and management of the SC systems. This software includes capabilities, from QSW, which supply the infrastructure for extreme network performance of parallel applications. In addition, the software provides users and system administrators with highly available single system management (SSM) of the nodes of the SC system, and thus minimizes the effort and complexity of both the administration and the usage of the system.

AlphaServer SC System Software organizes a SC system into a small number of cluster file system domains of up to 32 nodes. Each with a single namespace for files and directories, including a single root file system that is shared by all members of the domain. A Parallel File System can be layered on each cluster file system domain. The SC System Software also includes a cluster alias feature for the Internet protocol suite (TCP/IP) so that a domain appears as a single system to its network clients. Because, under SC System Software, the nodes in a domain share system and common system configuration files, management tasks need to be performed only once within the domain, rather than repeatedly for each individual node. For example, you install Compaq FORTRAN just once per SC domain rather than once on each node. In addition, most network applications need to be configured only once for the domain.

SC System Software arranges a SC system into parallel execution partitions. The partitions are managed and scheduled by a job management facility, RMS, which implements gang scheduling of parallel jobs, partition management, user quotas and job accounting. These functions may be further enhanced through the integration of other resource and batch-management products.

The system software, including the SC System Software, the *Tru64 UNIX* Operating System, and the Advanced File System Utilities, resides primarily on storage systems served by one of two designated system nodes of each of the domains of the *SC* system. The other node is capable of serving the storage system to provide load balancing and in case of failover. System utilities are provided for easy replication of system files between domains, thereby further simplifying overall management of the entire SC system.

Single system management, performance visualization, and hardware diagnostics are each done from centralized user interfaces, which extend the system management utilities of *Tru64 UNIX*. These management tools include, in addition to the *Tru64 UNIX* functions, a network-diagnostics utility.

SC System Software includes the drivers and low-level libraries for the *AlphaServer* SC Elan cards. The low-level Shmem library is provided for direct access to the memory of remote processes. In addition, a highly optimized MPI message-passing library is included. MPI jobs are deployed by the *prun* execution command.

File Systems

Through the system software, *AlphaServer* SC systems provide access to three primary file systems:

- Cluster File System (CFS)
- Parallel File System (PFS)
- Externally Served Network File System (NFS)

These file systems are illustrated in Figure 3.

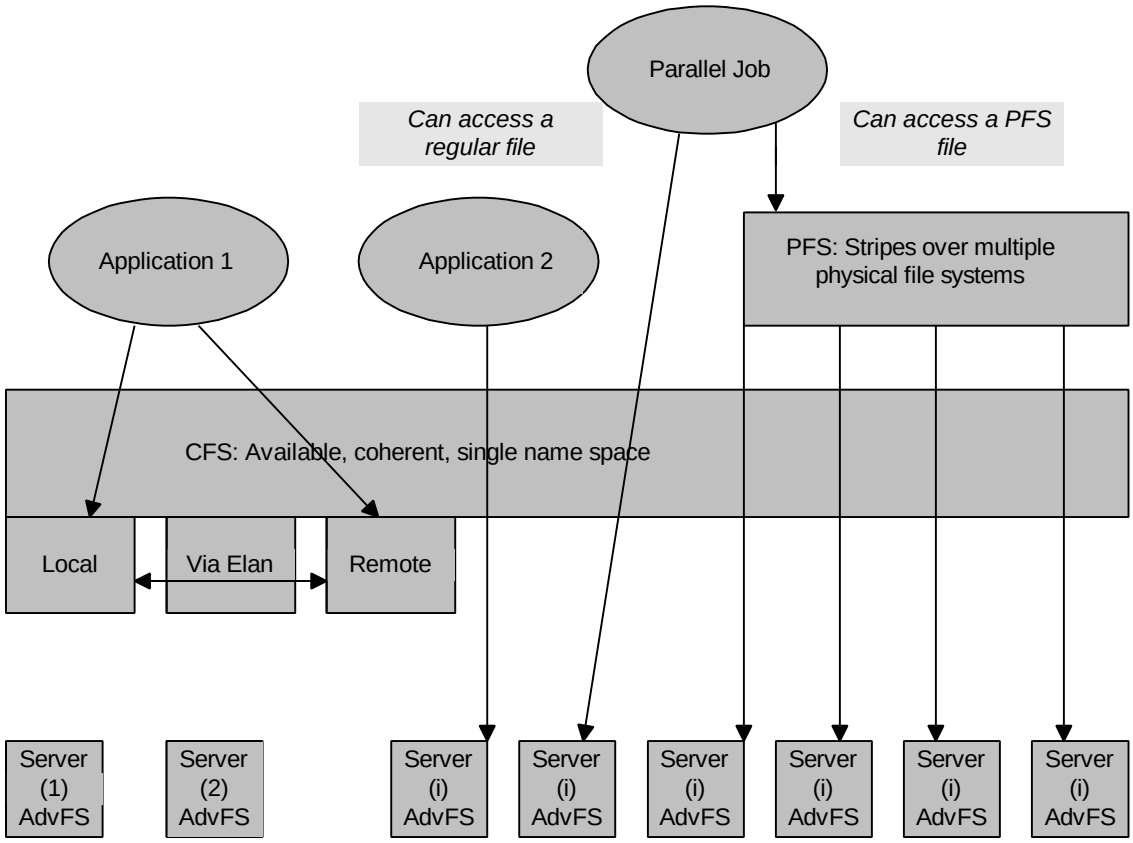


Figure 3: File Systems

Cluster File System (CFS)

CFS is a file system service that integrates all of the underlying file systems within a CFS domain. CFS does not provide disk-structure management; it uses the capabilities of the serving file system for this. The underlying serving file system used is the standard Compaq AdvFS product, with no changes to on-disk structures.

CFS is a POSIX and X/Open compliant file system. CFS provides the following capabilities:

- **A single coherent name space**: The same pathname refers to the same file on all nodes. A file system mount on any node is a global operation and results in the file system being mounted at the same point on all nodes.
- **Global root**: The point of name space coherency is at the root of the file system and not at a subordinate point; therefore, all files are global and common. This enables all nodes to share the same files, for example, system binaries, and global configuration and administration files.

- **Failover**: Because the file system capability is global, CFS will detect the loss of a service node. CFS will automatically move a file service from a failed node to another node that has a path to the same storage. In-flight file system operations are maintained.
- **Coherent access**: Multiple accesses of the same file will give coherent results. (Though this mode of access is less common with high-performance applications, and incurs a performance penalty, it is essential for enterprise applications.)
- **Client/Server file system architecture:** Each node's local file system acts a server to other nodes. Each node is also a client of other nodes.
- **Support for node-specific files with the same pathname on each node**: This is implemented through a Context Dependent Symbolic Link (CDSL) – a file link with a node identifier in the link name. CDSL is a feature of CFS. The node identifier is evaluated at runtime and can be resolved to a node-specific file. This can be used to provide, for example, a node-specific */tmp* directory. This feature support is used to provide node-unique files and to optimize for local performance.

The diagram in Figure 4 illustrates a sample CFS file hierarchy. The member boot partitions are mounted on node local storage. The system components of the root partiton will be mounted on external storage on one of the system file servers. Other nodes serving external storage can mount their file systems at other points within the hierarchy; this is illustrated by */a* and */b* in the diagram.

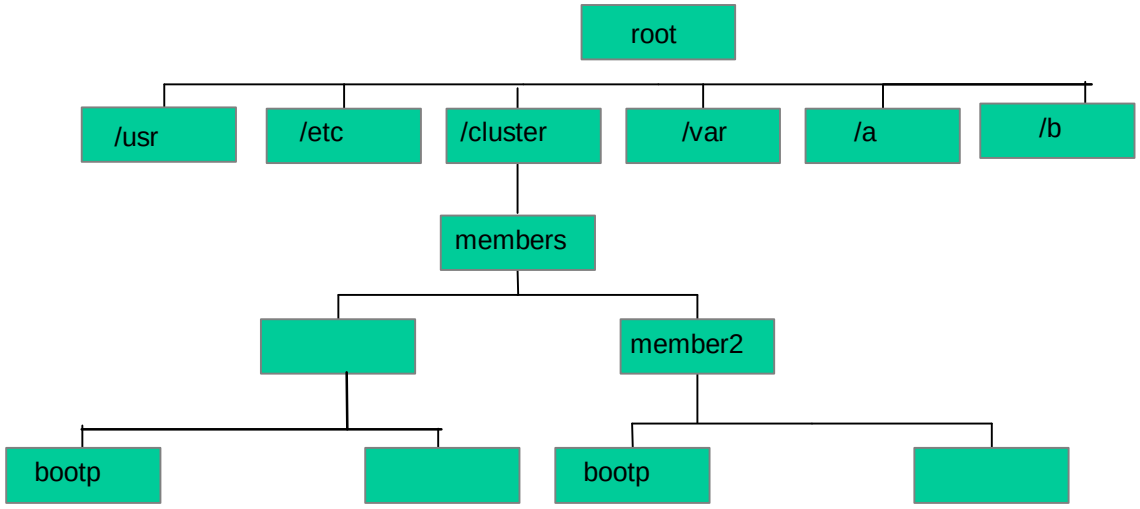


Figure 4: Sample *AlphaServer* SC CFS File Hierarchy

The operation of CFS is described in detail in the Compaq *AlphaServer* SC System Administration Guide

Parallel File System (PFS)

CFS provides the ability to scale the total file system bandwidth for accesses to files on different servers. PFS delivers scalable bandwidth capability to a single job within a CFS domain by striping the data of a single parallel file system over multiple underlying file systems. This allows accesses from different processes within a single parallel job to proceed in parallel and to be served by concurrent file servers. Contrast this with multiple processes accessing the same file served by a single file server.

When a parallel file system is created by an administrator, the constituent file systems and the default stripe size are specified.

For the programmer, PFS uses POSIX-compliant syntax and semantics. It supports additional file system commands (implemented as *ioctl*s) that allow a job to interrogate and modify attributes of the file system. For example, determining which file system components are local to a process, modifying the stripe size of a file, and so on. This enables a parallel job to optimize its I/O with respect to the underlying PFS structure. The operation of PFS is described in detail in the Compaq *AlphaServer* SC System Administration Guide.

Externally Served Network File System (NFS)

Just as with any SMP server, *AlphaServer* SC systems can mount file systems exported by other file servers. This means that user home directories can be located outside the *AlphaServer* SC system, and data on other servers can be accessed. Multiple external file systems can be mounted. It is possible to mount different external servers on different *AlphaServer* SC nodes.

Nodes that mount external file systems act as clients to those external systems. These nodes also act as servers of the imported file system to other nodes within the *AlphaServer* SC system – that is, external file systems only need to be mounted by a single node to be available to all nodes.

External Networking and I/O

AlphaServer SC systems connect to external LANs using network interfaces on a subset of the nodes. Standard network interfaces are supported, including Fast Ethernet, FDDI, ATM, and HiPPI.

Any subset of nodes can support external interfaces, ranging from one node to all nodes.

AlphaServer SC systems employ a facility known as Cluster Alias (CA) that allows each CFS domain to be represented by a single IP address or hostname. Multiple aliases can be defined for each CFS domain. The administrator can define additional aliases and their constituent nodes. The administrator can also define the alias capabilities of the services that run on the system, as follows:

- A service can be designated to run on a single node only, with transparent failover to another node.
- A service can be defined to run on multiple nodes; the CA facility will load-balance between those nodes running the service.
- A service can be defined as accessible only through an IP alias.
- A service can be defined as accessible only through a real IP address.

Services establishing outward connections will, by default, use the default alias as their source address.

The Cluster Alias capabilities enable the system administrator to configure subsets of nodes to provide specific services in available single server mode or in multiserver load-balanced mode. Clients of these services do not need to be exposed to the multiple node nature of the *AlphaServer* SC system.

Job Management

The RMS job management facility allows the administrator to treat the total set of CPUs on the *AlphaServer* SC system as a single large pool. The administrator can subdivide the pool into smaller partitions for use by different classes of job. When a user executes a job, the job management system is responsible for selecting the set of nodes that best fits the job's needs and for creating the requisite processes on each node. This process is referred to as **gang scheduling**, that is, scheduling the gang, or set, of processes that constitute a job over multiple nodes. The user does not have to be directly aware of the individual nodes on which the job is running.

The total computing space provided by the nodes can be subdivided into non-overlapping partitions. A partition is a collection of SMP nodes; it is not possible to subdivide the resources of a single SMP node between partitions. The system administrator can use partitions to allocate compute resources according to enterprise needs; for example, a system administrator may create partitions for development, production, and interactive work. The system administrator can define multiple configurations, each with a different set of partitions. Different configurations can be applied, for example, for successive shifts. Partition types include "interactive" and "parallel". Login is disabled for nodes defined as being part of parallel partitions. Interactive nodes can be used for login and code development. When a user runs a parallel job, the following can be specified:

- The partition that is to be used for the job.
- The number of processes in the job.
- The number of CPUs per process.
- The allocation of processes to nodes. The user can specify maximum granularity, finer granularity, or "don't care". If no granularity is specified, the gang scheduler will run the job as long as there are sufficient CPUs available.

If sufficient CPU resources are available, the scheduler creates the required processes on the specified nodes. The processes that make up a parallel job are constrained to run on the set of CPUs specified by the gang scheduler. If processes within the job create

child processes, these child processes are subject to the same constraints. When a job is terminated or de-scheduled, all processes are handled in concert, including any child processes created after job startup. This ensures that CPU resources can be managed reliably.

The job management system supports the time-sharing of a partition between different long-running jobs: the system deschedules the set of processes that constitute the job and schedules the set of processes for the job that is to be run. The gang scheduler works by giving the operating system scheduler no work to do; that is, it ensures that there are no other active processes competing for a CPU's cycles. Ensuring that all processes of a parallel job make even progress is critical to job performance.

When a partition is designated as time-shared, the gang scheduler will group jobs of equal priority using an administrator-specified time slice. (The time slice is specified in seconds or minutes.) When jobs have been grouped in this way, the scheduler will schedule a new job only if sufficient memory resources are available to run the new job and the existing jobs. Swapping and paging are not practical because of the large memory sizes and slow page rates involved. The job management system accounts for resource usage on a per-job basis; that is, total resource usage is aggregated. In addition, the job management system can be used for access control. The resources in each partition are managed by a Partition Manager, which mediates user requests, checking access permissions and resource limits before scheduling the user's jobs. RMS accounts for resource usage on a per-job basis. Total resource usage is aggregated. The RMS database is accessible through an SQL interface.

RMS can be integrated with batch systems such as LSF from Platform Computing. This capability allows users to run parallel jobs by submitting them to a batch queue, which dispatches jobs according to the scheduling policies that, are managed by the batch system software. More information about LSF can be found at the URL: http://www.platform.com.

The components of *AlphaServer* SC systems job management are described in the Compaq *AlphaServe*r SC System Administration Guide.

System Administration

Tru64 UNIX SysMan Menu and SysMan Station together provide a single-system interface to management of the *AlphaServer* SC configuration, and may be used to determine the state of availability and connectivity in the system. A choice of graphical, Web-based, or command-line user interfaces makes management and file-manipulation tasks easier for the administrator, flexible for those with large configurations, and streamlined for users.

Most of the system management and configuration files used by the system are global and common. With the use of CFS, operations on these files only need to happen once for the changes to be visible to all nodes within each CFS domain. For example, the password file is global-adding a user to the password file automatically creates an account for the user on all nodes within each CFS domain. Apart from system installation and physical maintenance, the system can be managed remotely. The system can be managed and monitored as a single entity. Monitor display provides the ability to aggregate collections of components into a single entity. Any fault or attention state in a sub-component is reflected in the parent entity. The management user interface allows the system administrator to drill down from the wide system view into nodes and within a node.

System administration utilities can have a global CFS domain or local focus. When the global focus is selected, commands are applied to the entire system. Frequently, this will mean updating a single file in global CFS space; at other times, daemons may need to be restarted on all nodes. When the local focus is selected, the administration command is applied to the specified node. This would apply, for example, to changing configuration settings on a network interface card on a particular node. The system management utilities are *locus aware*, that is, they are aware of whether they apply to the global or local context and will report an error if used incorrectly.

AlphaServer SC System Software includes highly automated system installation capabilities with supporting components, including console software. The capabilities provide an interface to the management functions that build the system software on the disks local to each node. They automate the setting of SRM (System Resources Manager) console parameters on all nodes. They are also used to set up and manage DECserver terminal servers into which the individual node console ports are connected, to manually connect to the node consoles and to log console output.

AlphaServer SC System Software can perform hardware error diagnostics and initiate remedial actions in the case of node failures in the *AlphaServer* SC system. Hardware-failure events are logged and can be monitored with the SysMan utilities. In addition, the events can also be configured to trigger actions such as sending email to the system administrator. When appropriate, events will initiate console functions that cause failed hardware components to automatically fail over to replacement components and then to automatically restart the failed hardware. Network behavior is controlled and monitored by the Switch Manager, which samples the network control interface, checking for network errors and monitoring performance. When network problems are found, events are posted and can be monitored.

Performance Visualizer

Performance Visualizer is a performance and resource-monitoring tool. Developers of parallel applications can use Performance Visualizer to monitor execution of their applications on the nodes of the *AlphaServer* SC system. With this information, they can redesign their applications to spread the load more evenly and reduce overall execution time. System managers can use Performance Visualizer by selecting one or more graphical displays to help them to quickly identify overloaded hosts, underutilized resources, active users, and busy processes.

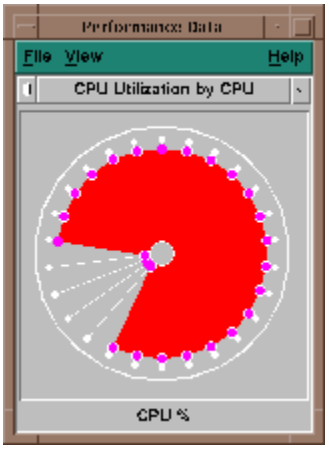


Figure 5: Performance Analyzer Kiviat Chart

System Installation

The system installation process provides an automated way to install a consistent set of operating systems and binaries on each system node. Automated system installation contributes to system usability, scalability and robustness. The system installation process is fully described in the *Compaq AlphaServer* SC Installation Guide.

The goal of the system installation process is to install a complete 128-node system within eight hours. *AlphaServer* SC system installation can begin when all physical infrastructures are in place, connected, and powered. The system installation process happens in three stages; data collection, manual installation of the server nodes, and automated installation of the remaining nodes.

At the data collection stage of system installation, all of the necessary configuration data is gathered through user and hardware interrogations. This central data collection saves time, avoids the tedium associated with repeated requests for the same information, and ensures that the configuration data used at different steps is consistent.

Within each CFS domain, one node is designated as the server node. The server node is installed manually according to the standard server installation processes. Installation is complete once the base operating system, the cluster integration system, and any additional layered products have been installed.

When installation of the server nodes has been completed, the remaining nodes are automatically installed. The installation process verifies that each node's hardware configuration is correct. The installation processes operate in parallel; the goal is to complete the automatic phase of the installation in one half shift.

Program Development and Tools

Programmers can view an *AlphaServer* SC system as a pool of serial processors, a pool of shared memory SMPs, or a distributed memory multi-processor. Each of these views is supported by one or more programming models, compilers, libraries, and tools. Typically, programs will be written using Compaq FORTRAN, Compaq C, or Compaq C++. Compaq FORTRAN and Compaq C both include support for the industry-standard OpenMP directives, which are widely used for, parallel development on shared-memory SMP systems

Most large-scale jobs that use the full capability of the *AlphaServer* SC system will be programmed using a message-passing model (multiple processes and address spaces) *via* the Message Passing Interface (MPI) library. In some cases, higher performance will be achieved through the deployment of a hybrid-programming model, where a set of shared address space threaded processes (written, for example, with OpenMP) pass messages *via* MPI. For the highest possible performance, but at the expense of additional complexity, programmers can use the shared-memory Shmem facility to take the fullest advantage of the *AlphaServer* SC Interconnect's extremely low-latency one-sided communication capability. Yet another alternative for parallel development and execution, the Compaq FORTRAN compiler includes High Performance Fortran (HPF).

The *AlphaServer* SC MPI library is an optimized implementation of the MPI-1 specification and is based on MPICH Version 1.1.1 from Argonne National Laboratory and Mississippi State University. Fortran and C interfaces are provided. The *AlphaServer* SC MPI library is layered on top of tagged message passing routines that are especially designed for the *AlphaServer* SC Elan cards, and that cause highly efficient communication between nodes connected by the *AlphaServer* SC Interconnect. In addition, highly efficient communication within a node occurs through direct memory transfers done with minimal data movement. A number of environment variables can be set to help optimize the performance of MPI programs, and these are described in the software documentation.

Support is included for a large subset of the MPI-2 I/O interface via the ROMIO package from Argonne National Laboratory. Performance of the MPI-2 I/O operations will be related to the attributes of the Parallel File System.

For more information on MPI, you can access the following URL: http://www-unix.mcs.anl.gov/mpi.

The *AlphaServer* SC Shmem library provides direct access, via *put* and *gets* calls, to the memory of remote processes. Fortran and C interfaces are provided. Whereas MPI, for example, requires that the remote process issue a receive to complete the transmission of each message, the Shmem library, by contrast, provides the initiating process with direct access to the target memory. The one-sided communication of Shmem maps well onto the DMA hardware in the Compaq *AlphaServer* SC Elan card. A consequence of this is that Shmem latencies are very low and bandwidths are high. The *AlphaServer* SC Shmem library is highly compatible with the CRAY™ Shmem library. Further information can be found in the software documentation.

Debugging both MPI and threaded (OpenMP and POSIX threads) applications on *AlphaServer* SC systems can effectively be done with the TotalView debugger from Etnus, Inc. For information on TotalView, access the following URL: http://www.etnus.com. MPI message tracing and analysis, primarily for purposes of optimization, can be done with the VampirTrace and Vampir tools from Pallas GmbH. For more information, access the following URL: http://www.pallas.com/pages/vampir.htm.

The components of AlphaServer SC software development are described in the *Compaq AlphaServer* SC User Guide.

Physical Infrastructure

This section describes the physical infrastructure that is used to create an *AlphaServer* SC system, including the internal and external network connections and as support components.

Hardware Components

The following list identifies the main hardware components of an *AlphaServer* SC system:

- System nodes: AlphaServer ES40s-based 21264a (EV67) severs with a minimum of 1 microprocessor and 1 GB of memory to a maximum of 4 microprocessors and 16GB of memory. Memory options should be configured so that full interleaving is enabled for optimal performance

- SW Interconnect adapters, cables and switch
- Internal management network: adapters, cables and hubs
- System Storage: required external RAID storage for system management and availability

- [optional] External support systems: for example, file servers, development systems
- [optional] External network connections: adapters and cables
- [optional] SC Management Station: locates central management functions and database

System Nodes

The system nodes provide the computing capability of the system. The system nodes can be flexibly configured, using the job management system's partition capability. The system can be partitioned such that some subset of nodes can be configured to provide file serving and/or interactive capability.

Networks

Each system node is connected to the following networks:

AlphaServer SC Interconnect: The interconnect provides the high-speed message passing and remote memory access capability for parallel applications. Multiple rails of this network can exist. Multiple rails are used to increase aggregate throughput and to reduce queuing delays on systems with large numbers of CPUs per SMP node.

Internal management network: This network is used to monitor and control all of the major system components (system nodes, data switches, and so on). This traffic is separated from the data network to avoid perturbing parallel application performance. If the SC Management Station node is present, it is connected to the internal management network as well to the external site network. This allows the SC Management Station to act as external portal for control of the system.

Internal console network: The serial port of each system node is connected to a terminal server. The terminal servers are connected to the internal management network. This configuration enables each node's console port to be accessed through IP over the management network. This facility provides management software with access to a node's console subsystem (for boot, power control, configuration probes, firmware upgrade, and so on).

Root Consoles

The system is also equipped with a graphics terminal that acts as the root console (the console on system nodes 0 or 1). This is used during the initial installation process to install and configure the root node. After the installation procedure, console access to all nodes is possible *via* management software. At this stage, the root console can be used to log in to the system.

The core components of the *AlphaServer* SC system are illustrated in Figure 6

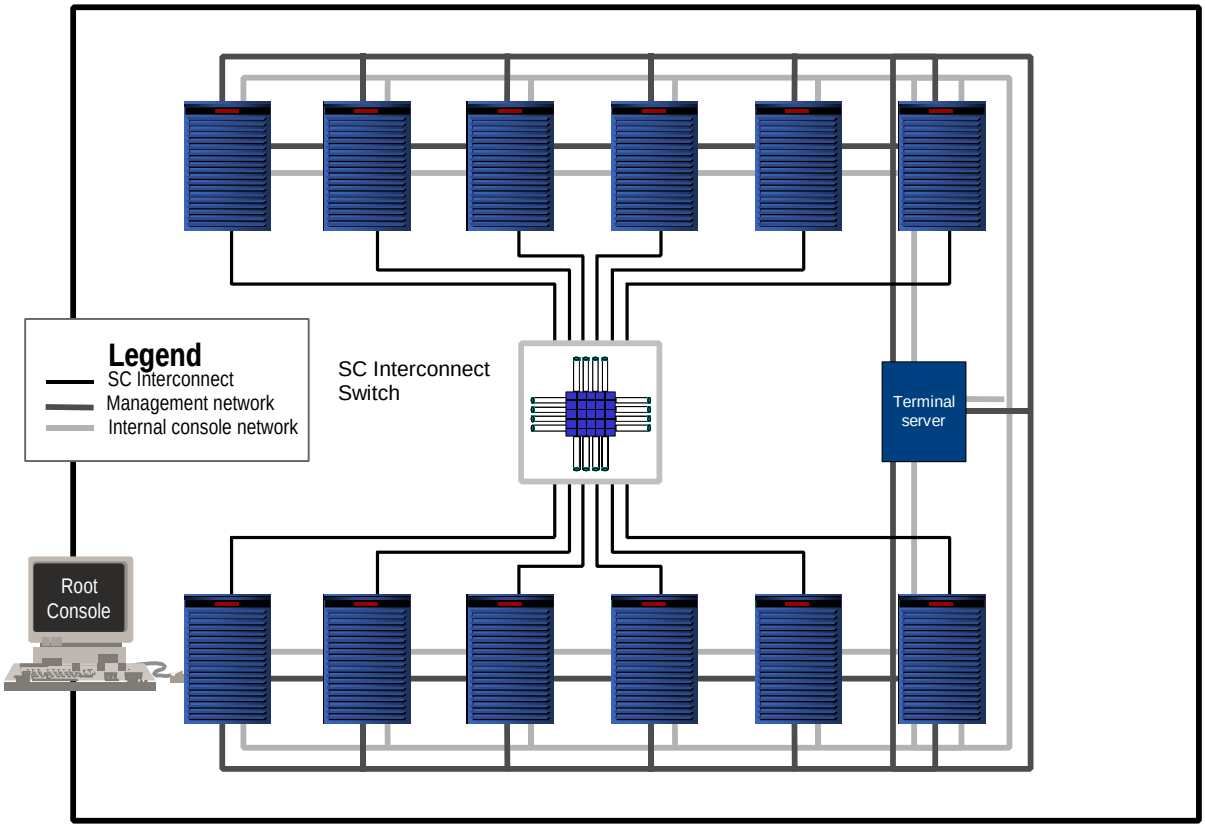


Figure 6: *AlphaServer* SC System Components

Node Local Storage

Node Local storage is internal to a node's cabinet. Typically, it used to store data that is intimately associated with the node itself; this includes the node's bootable operating system, swap space, and node temporary files. This internal storage should be considered volatile; it is not a system goal to make this storage highly available. All of the system data (the bootable operating system) can be automatically regenerated in case of complete hard disk failure. Data stored on local storage is node-specific and the loss of a node should not have an impact on the data requirements of the rest of the system.

Each node is configured with two internal drives. Under normal operation, the secondary drive holds a copy of the primary drive's boot partition. A node can run with a single drive. During the upgrade process, the secondary drive stores the boot partition for the upgraded operating system. This allows fast reversion to the original operating system version, if necessary.

The internal drives are configured at system installation time, but they can be reconfigured by the administrator.

Applications can use data that is stored on file systems on the internal drives, on the understanding that this data is not highly available.

External Storage

External storage is used to hold file systems served to the rest of the system (and potentially, to other external clients). External storage is also used to hold system files and customer data. Any number of nodes, from two nodes to all nodes, can be configured as file server nodes with external storage.

External storage is highly available through the use of redundant paths, RAID, and the failover capabilities of CFS. External storage and associated file systems will remain available and accessible if any single component fails, because of the following:

- Each storage subsystem is connected to at least two hosts

- Storage is RAID-protected
- Physical disks are connected to dual RAID controllers

System storage is configured as external storage for availability.

External Support Systems

Using the external network, the system can connect to a variety of external support servers, for example, file servers and development systems. Alternatively, it is possible to configure an internal partition (consisting of some subset of the computational nodes) to provide login/compile services and user file directories.

External Network Connections

System nodes can also be connected to external networks. The number of nodes used to provide such connections and the type of network interconnect used can be specified. Standard system interconnects (FDDI, Ethernet, Gigabit Ethernet, ATM and HiPPI) are supported.

IP aliasing is used to present a system with multiple external interconnects as a single network entity.

Systems can also be configured with an optional front end. The *AlphaServer* SC Management Station is a server running the *Tru64 Unix* operating system, but it does not have to be the same server type as the system nodes. If this node is present, it can be configured to do the following:

- Act as a Remote Installation Server (RIS) for the installation process
- Serve user home directories
- Act as a development server
- Host the central management functions and daemons for the system

Key Features

Up to 512 667MHz Alpha 21264a (EV67) Microprocessors at launch

Up to 16,000 planned Alpha 21364 (EV7) Microprocessors within two years

More than 6 GB/sec of bisection bandwidth

Less than 3µ seconds of user-level latency

Optimized MPI and Shmem libraries

Comprehensive parallel development software

Very Large Memory Support

Hot swap and hot add components

Redundant remote and local console subsystem options

High availability and reliability features engineered throughout

The Compaq Advantage

- Proven Technology
- The best Price/Performance in the industry
- High Availability
- Scalability and Growth
- Incredible Flexibility
- Investment Protection
- Full Service Worldwide

Model Specifications

Form Factor	Standard RETMA Racks, H9A15-MD, 230/240V
Processor/ES40	Choice of 1 to 4 667MHz Alpha 21264a processors; each with 64KB I-cache, 64Kb D-cache on chip, and 8MB per processor of L2 cache
Memory/ES40	Choice of 1 – 16 GB, ECC, 4-way interleaved industry-standard DIMM memory.
ES40 Architecture	Advanced dual 256-bit wide memory data paths and crossbar switch technology providing 5.2 GB/sec peak memory bandwidth; dual 64-bit PCI busses providing over 500 MB/sec peak I/O throughput
Performance	For the latest performance numbers visit our Web site at: http://www.compaq.com/hpc
ES40 PCI	10 64-bit PCI slots; 2-4 available with initial configurations
Storage Controllers	Integrated single-channel Ultra2-SCSI, Ultra SCSI RAID, HIPPI, and Fibre Channel
Network Controllers	Includes dual 10/100 Fast Ethernet, and asynch. Communications, optional Gigabit Ethernet and ATM configurations
Drive Bays	Includes 1 internal hot-swap StorageWorks Ultra2 SCSI drive bay in each ES40, with 2 18.2 GB1.6" drives included. Four removable media bays: one 3.5" bay for diskette drive; one 5.25" for CD-ROM; and two open HH 5.25" bays for tape or hard disk drives
Power Supply	Includes three hot-swap 750-watt (N+1) power supplies per ES40.
Cooling	Air, Six hot-swap redundant variable speed fans per ES40
High Availability	Hot-swap redundant power and cooling, auto reboot, thermal management software, remote system management, RAID, hot-swap drives, memory fail over, ECC memory, ECC cache, SMP CPU fail over, error logging, optional Uninterruptible Power Supply (UPS), and UPS Power Management Software
Service and Support	Compaq provides a 3-year on-site, 5-day x 9-hour warranty with next business day response. Optional service options for up to 4-hour same-day response time are available, as well as a complete portfolio of worldwide service offerings to maximize your critical system environment
Operating System	Tru64 UNIX 5.0 with Java
Required Software	*AlphaServer* SC System Software is required with all *AlphaServer* SC Series systems. Also highly recommended is *AlphaServer* SC Development Software, which includes FORTRAN, C++, and Developers Toolkit Extensions for each node. For the Development Software, all nodes must be licensed if ordered.
SC Interconnect	DMA driven; get and put ; Bandwidth 200MB/sec/rail bi-directional; 0.035µsec switch Latency; <3 µsec DMA/Shmem and <5.5µsec MPI Latency
SC Interconnect Switch	8-way x-bar chips; 16 or 128 port packages; up to 20m cables
Single System Management	*AlphaServer* SC System Software supports system wide gang scheduling, resource management and performance monitoring. PFS, and Cluster File System software enable the management of up to 4 –32-node CFS domains.
Value-added Implementation Services (VIS)	*AlphaServer* SC standard package includes Staging and Integration of the *AlphaServer* SC System, storage devices, QSW interconnect switch and other peripheral devices, software load of Tru64 UNIX operating systems and associated layered products. Optional services include custom site planning, custom freight arrangements, VIS engineering team travel and installation of *AlphaServer* SC, and tailored Customer Configuration Documentation
Support and Training	90 days of Consultant level onsite support is included to help address the planning and introduction of the *AlphaServer* SC into your environment. A full range of Training Courses are available at additional charge
Installation	Installation is included with each of the System Building Blocks and is available for quotation on additional and optional equipment