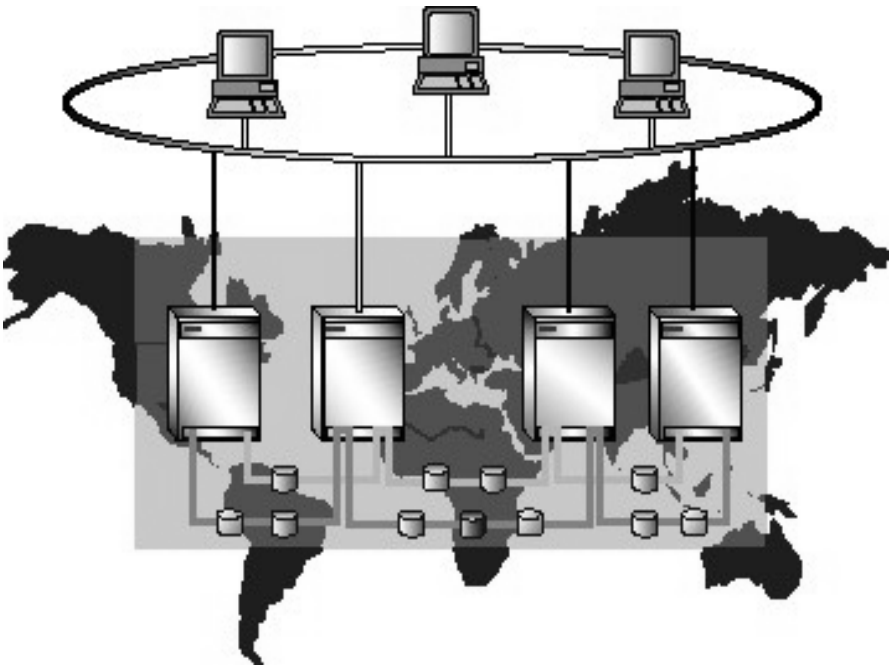
COMPAQ

TruCluster Software

Highly Available and Scalable Solutions on Tru64 UNIX AlphaServer Systems



EC-W9871-43

January 1999

Visit our Web site at

http://www.compaq.com/tru64unix/

Compaq Computer Corporation believes the information in this publication is accurate as of its publication date. Such information is subject to change without notice. Compaq is not responsible for any inadvertent errors.

Compaq conducts its business in a manner that conserves the environment and protects the safety and health of its employees, customers, and the community.

Part Number: EC-Z8404-51

The Challenge

The world of commerce relies on highly available information systems to run and manage core business operations such as data warehousing, online transaction processing, and decision support system solutions. Likewise, the worlds of science and engineering also rely on high-performance computing to provide solutions and solve problems. No matter how fast or available today's computers are, tomorrow's applications will invariably demand more.

Mainframes, supercomputers, and fault-tolerant systems have historically provided the highest levels of service. Unfortunately, their specialized construction and proprietary components contribute to long design cycles, high costs, and a fall off the performance curve—attributes unsuited for today's client/server world.

Today's challenge involves not just providing high availability and optimal performance, but doing so flexibly and inexpensively. The Compaq TruCluster Software products minimize downtime. This document explains how TruCluster Software products address a broad range of system availability and performance requirements.

The Solution: Clusters

A *cluster* is a closely linked group of computers that provides fast, uninterrupted computing service. More technical definitions exist, but they say essentially the same thing: close cooperation among systems can both improve performance and minimize downtime.

Clustering computer systems together is not in itself a remarkable feat. *How* clustering achieves these goals *is* remarkable. The speed and reliability of traditional monolithic systems are based on intensive engineering and research. Many system parts are meticulously designed from scratch. Unfortunately, close tolerances and maximal optimizations also entail long development cycles and high costs. This has led to the search for a better solution.

Clustering results from a fundamental rethinking of the best way to deliver a highly available and scalable system, with no single point of failure. Focusing on the desired results, rather than a specific implementation, allowed designers to innovate. They realized that individual systems and their components do not have to match the characteristics of the mainframe, supercomputers, or fault-tolerant systems—as long as a group of systems can cooperate to achieve similar results.

Instead of custom components, clusters combine the best off-the-shelf, standardized components into cooperative groups. For example, if one cluster member fails, another steps in to assume its workload. If one member cannot accomplish a task quickly, other members help with the workload. Working together, groups of off-the-shelf components approach or surpass the capabilities of mainframes, supercomputers, and fault-tolerant systems. Moreover, they do so at much lower costs.

A Clustering Example

Imagine that several hundred users are accessing your Web site to access flight reservation information. The cluster environment supporting this Web site contains several multiprocessor systems. If a system fails, the server will start up with information still available to users. Any service interruption is brief—a few seconds to a few minutes while the cluster coordination software restarts the application on another cluster member.[1] After the underlying problem is corrected, users are transparently shifted back to their original system.

In this example, the cluster solution helps in several ways. Long before a problem occurs, the clustering capability enables incremental growth—the combination of several modest-cost servers added over time handles a larger, increasing load while

---

[1] TruCluster Available Server software can fail over in as little as 15 seconds. Application-specific recovery times vary from a few seconds to a few minutes.

protecting your investment. If a component fails, the TruCluster environment isolates the fault, and then quickly recovers from the failure conditions. Although an individual system still fails,[2] its impact is minimized.

Tru64 Clusters

Clustering is not just an intriguing idea; it is a reality for Compaq's customers. Since pioneering the concept in the early 1980s, Compaq's AlphaServer customers have installed over 95,000 clusters, containing over 500,000 systems. Our customers' long-term acceptance proves the effectiveness of a well-implemented clustering technology.

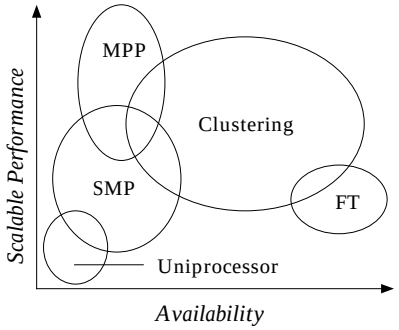
Independent market analysts, such as the Gartner Group and Illuminata, use the capabilities of OpenVMS Clusters as the benchmark for the industry. That leadership continues with the Tru64 UNIX, formerly DIGITAL UNIX, TruCluster products.

In October 1993, the strategy to bring its clustering expertise to Tru64 UNIX systems was announced. Less than a year later, the TruCluster Available Server solution was delivered. The evolution continued in early 1996 with the introduction of low-cost memory channel interconnect cards for the PCI bus and a new high-end product, TruCluster Production Server Software.

Putting Clusters in Perspective

Evaluating systems technologies can be difficult. Symmetrical multiprocessing (SMP), fault tolerance (FT), massively parallel processing (MPP), and clustering all compete for market position based on their respective capabilities.

While each technical approach has a role, SMP and clustering stand out as mature, broadly effective technologies. Working together, as they often do, clustering and SMP address virtually the entire range of customer requirements.



Clusters Deliver Availability

Availability is the proportion of time that a system can be used for productive work. Typically expressed as a percentage, 100% is the best possible rating.

A typical standalone system can achieve about 99% availability. This sounds impressive, until you realize that the missing 1% represents about 90 hours—over three and a half days—of downtime per year. Some of this is "planned" downtime (time the system must be down for maintenance to prevent the more serious "unplanned" downtime), but 99% availability suffices only for forgiving organizations and noncritical applications. Systems upon which a business depends must do much better.

---

[2] Avoiding such failures is also possible with fault-tolerant systems, but at a much higher cost. Also, fault-tolerant systems do little to alleviate planned shutdowns.

At the other end of the spectrum, critical applications such as emergency call centers, telecommunications hubs, air traffic control centers, financial services, and medical equipment must be running 24 hours a day, every day of the year (7 x 24 x 365). *Any* amount of downtime may risk lives, money, and reputations. These situations have been the focus of "fault-tolerant" or "continuous processing" systems that use extensive redundancy and specialized construction in a heroic attempt to prevent service interruptions. Fault tolerant systems can achieve 99.999% availability or better—that is, about five minutes downtime in an average year.

Although everyone wants to eliminate downtime completely, few applications can justify the expense of fault-tolerant systems. The specialized construction and extensive redundancy of fault-tolerant systems makes their cost several times that of conventional systems. More importantly, after a system has been made 99.95% or 99.99% available, all of the likely failures will be software or environmental failures, not hardware breakdowns. Spending money to make the hardware even more reliable is not very cost-effective. Organizations should consider this option *only* in the most availability-sensitive situations.[3]

The solutions of greatest interest to most users are those that experience between five minutes and three days of downtime per year. When attentively managed with a supportive set of systems management tools,[4] conventional standalone systems can achieve between 99.5% and 99.8% availability—or 18 to 44 hours of downtime per year. If this downtime is based on an average cost of $1,400 a minute, then 43 hours of downtime could cost a company $3.6 million per year.[5] Another study estimates that downtime can cost US businesses $4 billion a year in losses, with the average failure costing $140,000 in the retail industry and $450,000 in the securities industry[6].

To go beyond this level of reliability into high availability or fault resilience requires clustering. TruCluster environments can eliminate all but a few minutes of downtime per year. What downtime it cannot eliminate, it mitigates. Unplanned downtime is converted from a serious problem into slightly reduced performance or, at worst, a brief service delay.

%	*Max Downtime*	*Availability*
99	3.5 days/year	Conventional
99.9	8.5 hours/year	High Availability
99.99	1 hour/year	Fault Resilient
99.999	5 minutes/year	Fault Tolerant

Highly available environments suit customers who can tolerate a brief delay while service is being restored. For example, an air-traffic control system requires fault-tolerance, but a reservations system based on highly available or fault-resilient clustering is more than adequate to keep agents selling tickets to satisfied customers.

Clusters Deliver Scalable Performance

In addition to high availability, clusters help achieve high performance. The term *scalability* (an abbreviation for *performance scalability*) is often used to stress the goal of high overall performance. It also hints at the incremental performance growth that clusters provide.

---

[3] This was illustrated by the mid-1994 failure of the NASDAQ stock trading system. Though run on fault-tolerant hardware, the system was brought down by a failure of application software.

[4] This figure is quoted from the 2/4/97 issue of The Financial Times via First! by Individual, Inc.

[6] *Ibid*.

Configuring multiple processors to work in parallel is one of the most direct paths to higher performance. Uniprocessors, multiprocessors, clusters, parallel processors, and distributed computing all use parallelism at a variety of levels. They are not all the same.

Parallelism

Each flavor of parallelism has advantages and limitations. The Compaq goal is to obtain the maximum value from each technique.

The real issue is how well application programs can use the parallelism each approach offers. You can partition some programs into smaller pieces and run each on a separate processor. Such *multithreaded* programs can achieve significant performance gains. The more pieces there are, the better use of parallel components and the bigger the performance gains. The perfect case is *linear scalability:* for N processors, the application runs N times faster than on one processor.

True linearity is rare, particularly as the number of processors grows. Many important programs cannot be extensively multithreaded. They may be partitioned, but only into a limited number of pieces, and decomposition requires significant effort, possibly even rearchitecting the programs.

This inherent difficulty is complicated by the need of the pieces to coordinate amongst themselves. Even if a program can be decomposed, the communications overhead can become a limiting factor, quickly overcoming the system's interprocessor communications facility.

Symmetric Multiprocessing

Symmetric multiprocessing (SMP) is popular because it provides a particularly effective way to scale performance. SMP uses a few parallel processors, connected by a high-speed system bus that are coordinated by an operating system. When well implemented, as it is in Tru64 UNIX, this modest level of parallelism can be handled fairly easily by applications. Developers often do not bother to parallelize individual applications. Users just run many jobs on an SMP system, and the operating system distributes the total workload, one job to a processor. Developers explicitly parallelize those programs with a high payoff from optimization, such as database managers.

SMP is effective with a few processors, but as more processors are added the demand for the shared resources grows. How fast the demand grows, and the number of CPUs that can be used effectively, depends on the level of interthread communications required by the application workload. Most workloads benefit from between two and eight processors.

Because system overhead increases as processors are added, application performance does not improve proportionally with the addition of more processors to an SMP design. At some point, the law of diminishing returns becomes a factor. Contention for shared resources causes a bottleneck. As more processors are added, total performance rises only slightly (if at all) or it may fall. Extending the capabilities of the system bus is not an answer because it is not cost-effective.

Another limitation is that while multiprocessors can be quite reliable, they are not highly available. If a processor fails, the system must be rebooted. If some other component fails (for example, a SCSI disk controller or network adapter), the system cannot be rebooted to solve the problem. Multiprocessors must also be taken off line for maintenance and upgrades.

Massive Parallel Processing

Massive parallel processing (MPP), in contrast to SMP, is not widely used. MPP uses large numbers of processors linked by proprietary interconnects, which are explicitly coordinated by application structuring. This architecture suits problems that

can be decomposed into hundreds or thousands of pieces. Some important tasks qualify, for example, weather forecasting and large-scale text retrieval, but these are not the daily tasks of most organizations. Large-scale parallelism is too hard to implement and the performance gains too haphazard for general use. MPP also does little to ensure system availability. Therefore, for a wide range of applications, scaling via MPP has limited value.

Cluster Coupling

In terms of performance, clusters share characteristics with both SMP and MPP. Depending on the number of processors, the interprocessor interconnect, and the software used to coordinate operation, a cluster can appear to mimic MMP or SMP behavior. *Loosely coupled* clusters provide a large number of processors, potentially hundreds, linked by networking technology. *Tightly coupled* clusters use a smaller number of processors linked by networks, storage channels, or specialized cluster interconnects. (Tight coupling is often used to describe shared-memory SMP systems.)

Because clusters have looser processor-to-processor communications than SMP systems, care must be taken to structure their workloads for scalable performance. However, the logical distance between processors has some advantages. Because systems are isolated from one another, there is less contention for shared resources. Also, one node can fail in a cluster and another node can continue its operations. Any TruCluster environment provides substantially greater availability than either SMP or MPP technologies.

Note that SMP and clusters are not mutually exclusive. TruCluster solutions support the entire AlphaServer product line. In this way, clusters and SMP work together.

Interconnects

Coordinating and sharing data are fundamental to cluster operation. Choosing the right cluster interconnect is crucial to maximizing cluster performance and usefulness.

Tru64 UNIX (formerly DIGITAL UNIX) supports a full range of standard networking media: Ethernet, Token Ring, FDDI, Fast Ethernet, ISDN, ATM, and CATV. TruCluster software, however, uses a higher speed, low latency interconnect.

Interconnect	*MB/s*
FDDI	10
Fast Ethernet	10
ATM	15
SCSI/Ultra SCSI	40
SCI	80+
Memory Channel	100+
Fibre Channel	100-200

Though basic to client/server components, networks are insufficient for all but small, loosely coupled clusters. Even FDDI or ATM, today's fastest common network components, provide only 100-155 MB/s.[5] This is too little bandwidth for the inter-node traffic that clusters can generate. Equally important, the extensive layering of standard network protocols entails high CPU overhead and long message delivery latencies.[6] Networks are often hard pressed to meet the growing needs of client/server communication, much less to handle the increased bandwidth requirements of cluster traffic.

One approach to the lack of available bandwidth is to move cluster communications from networks onto storage channels, which can be faster than networks. Unfortunately, there is not much bandwidth to spare on storage channels, which are already heavily used for data transfer.

Given the inherent bandwidth limitations associated with traditional networks and storage channels, many clusters use high-speed cluster interconnects for intracluster communication. The Tru64 UNIX TruCluster software provides two solutions.

GIGAswitch/FDDI

The first solution, GIGAswitch/FDDI, supercharges the network with switching technology. The TruCluster Available Server product, as an example, uses a dual monitoring system to detect failover by responding to operating system crashes, failed Ethernet controllers, and disconnected network cables. This does not reduce the fundamental latency or CPU overhead issues of networking, but it provides bandwidth to spare in a manner that is transparent to the software. The GIGAswitch/FDDI suits large, loosely coupled clusters.

MEMORY CHANNEL INTERCONNECT

The second solution is to use memory channel, an interconnect that is fundamentally designed for cluster requirements. Using the memory channel interconnect, you can configure your clusters to provide the computing power found in highly specialized MPP systems. These clusters also provide the necessary level of redundancy, while also providing room for growth.

GIGAswitch/FDDI—Turbocharging the Network

GIGAswitch/FDDI is an intelligent, high-speed link between multiple FDDI networks. Using a crossbar switching technology, it can connect up to 34 FDDI ports at full speed.[7] Though each individual communication channel is limited to FDDI speeds, as a whole, GIGAswitch/ FDDI delivers up to 3.6 GB/s of bandwidth. At 360 times the bandwidth of common Ethernet, GIGAswitch/FDDI provides more than enough bandwidth for large-scale data sharing in loosely coupled clusters. Because it strikes the optimal balance of performance and economy, GIGAswitch/FDDI is particularly effective in conjunction with the Compaq TruCluster Production Server product.

GIGAswitch/FDDI is not, however, specific to the TruCluster environment. It is a general networking product for use with standard FDDI equipment managed by any standards-based network manager. Recognizing this innovation, *R&D Magazine*[8] gave the R&D 100 Award to GIGAswitch/FDDI.

---

[5] MB/s means millions of bits per second, a measure of data transfer throughput. Links are measured in (rough) multiples of millions (MB/s) or billions (GB/s).

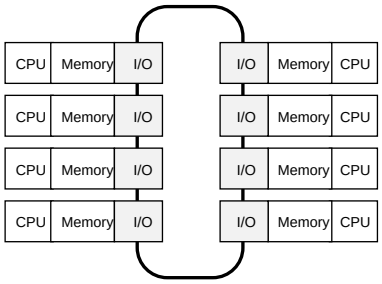
[6] *long* is relative to processor speed. For an Alpha 21164 CPU executing over 1 billion instructions per second, a delay of several hundred microseconds means a lot of wasted CPU cycles.

[7] Actually, being full-duplex, it can transfer 100 MB/s each way.

[8] This puts GIGAswitch/FDDI in select company. Past winners of the award include antilock brakes, automated teller machines, halogen lamps, and fax machines.

The MEMORY CHANNEL Interconnect Breakthrough

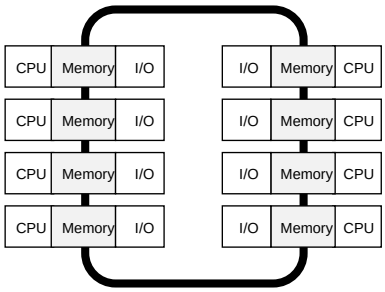
Networks are designed as I/O devices for relatively slow physical links with high error rates, long transmission distances, and an expectation of poor security. Accommodating these conditions requires complex, multilayer protocol suites such as TCP/IP. This layering imposes both delays and overhead for every network packet sent or received.



Cluster interconnects are designed for the opposite environments: fast physical links with low error rates, limited geographic dispersion, and tighter physical security.

Instead of forcing performance-limited networks to carry the additional burdens of cluster traffic, a better approach is to use an interconnect designed from scratch for cluster operations. Such a link will have high bandwidth, low latency, and low overhead. This describes the memory channel interconnect.

Using a technology known variously as distributed shared memory or reflective memory, memory channel delivers a bandwidth ten times that of FDDI. Its latency and overhead are on the order of 100 times lower. Memory channel operates at near system bus speeds, providing a much larger opportunity for tightly coupled clusters to approach SMP and shared memory performance. The memory channel interconnect represents a fundamental breakthrough in cluster interconnects.

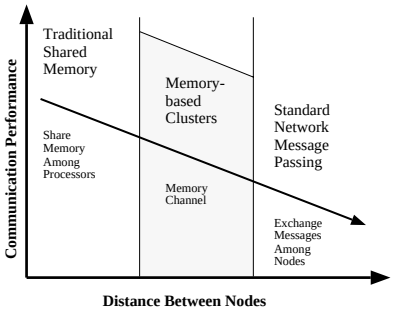


Memory channel is a memory-to-memory connection between multiple systems. Like the shared memory of a multiprocessor, a store instruction executed on one node nearly instantaneously affects the data seen by other nodes. This allows database managers and other cluster-aware software to share information at speeds, latencies, and overheads closely approaching those of memory-to-memory copies across a system bus.[9] This is shown in the following performance figures:

[9] Software that is not cluster-aware can still use the memory channel interconnect with standard TCP/IP protocols. Unfortunately, the overhead inherent in the TCP/IP design will extract a price.

Measure	*Network*	*MC*	*SMP*
Bandwidth (MB/sec)	<10	100	>500
Latency (μsec)	>150	<2.5	<0.5

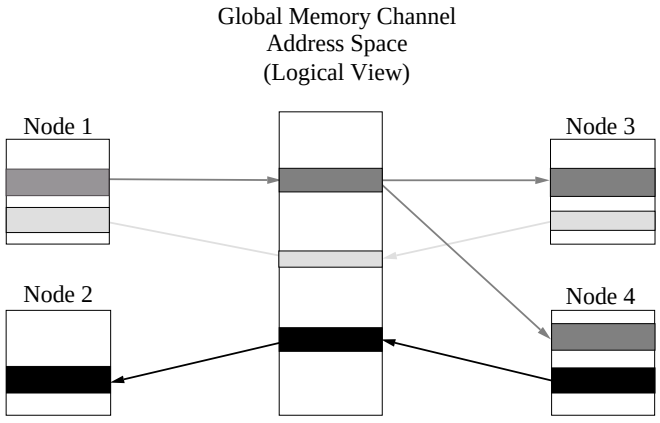
Distributed shared memory is like a system bus extended across multiple systems. Simple in concept, making this idea work requires clearing important engineering hurdles. For example, a simple implementation of a shared bus linking fewer than 100 processors will see a high level of contention and poor performance. The failure of one processor could bring down the entire cluster. Of course, neither situation is acceptable.



(Note: Both axes are log-scaled)

Making memory channel work required a significant engineering effort. At a basic level, this meant handling protocol processing in the hardware of the memory channel connection card, integrating memory channel with system bus technology, and designing hubs to connect more than two systems.

At a higher level, it meant designing a balanced combination of attributes of *intimacy* (like shared memory) and *isolation* (like a network). For example, though data sharing is intimate, the only memory operations transmitted across the memory channel interconnect are those with a cross-cluster impact. For reliability, if one node fails, it is automatically shut off from the remainder of the cluster in order to isolate the failure conditions.



Standard Off-the-Shelf Components

Compaq is committed to using off-the-shelf components and standard interfaces whenever and wherever possible. This is a win-win approach. It improves the time-to-market window and reduces development costs. For customers, it increases flexibility, promotes competition, and reduces long-term costs.

The TruCluster environment is no exception. Unlike many competitive clustering products, there are no unique system configurations, specialized operating system variants, or proprietary storage components. TruCluster products use the same AlphaServers, Tru64 UNIX operating environment, SCSI disks, fibre channel, disk controllers, network adapters, and applications software as our other systems.

Where Compaq has enhanced performance with Compaq designs—for example, with fast Alpha processors or the GIGAswitch/FDDI interconnect—using standard interfaces allows customers to mix and match Compaq and non-Compaq components. The TruCluster Available Server, for example, supports HP, IBM, SGI, and Sun nodes as well as Compaq systems as clients. GIGAswitch/FDDI connects to any vendor’s FDDI network. Even the memory channel interconnect uses a standard PCI bus connector.

In short, Compaq has used and will continue to use standards and commodity components in all facets of the Tru64 UNIX TruCluster products.

Tru64 UNIX TruCluster Products

Business and other enterprises are continuing to move from mainframes and other proprietary systems to UNIX based open systems. UNIX has evolved over the past decade into a solid base for customer's enterprise operations. While UNIX based systems provide a solid foundation, meeting the availability and scalability requirements of mission-critical applications demands a comprehensive, dependable clustering solution. Tru64 UNIX TruCluster products provide this solution.

User Requirements

The most appropriate cluster configuration depends on your specific situation and needs. Most users, however, see the following themes, with availability and scalable performance woven throughout their requirements:

- Availability
- Performance
- Affordability
- Manageability

The TruCluster Solution

The TruCluster product suite meets business and enterprise requirements with the following products:

- TruCluster Available Server provides high availability
- TruCluster Production Server combines high availability and high performance

TruCluster Available Server

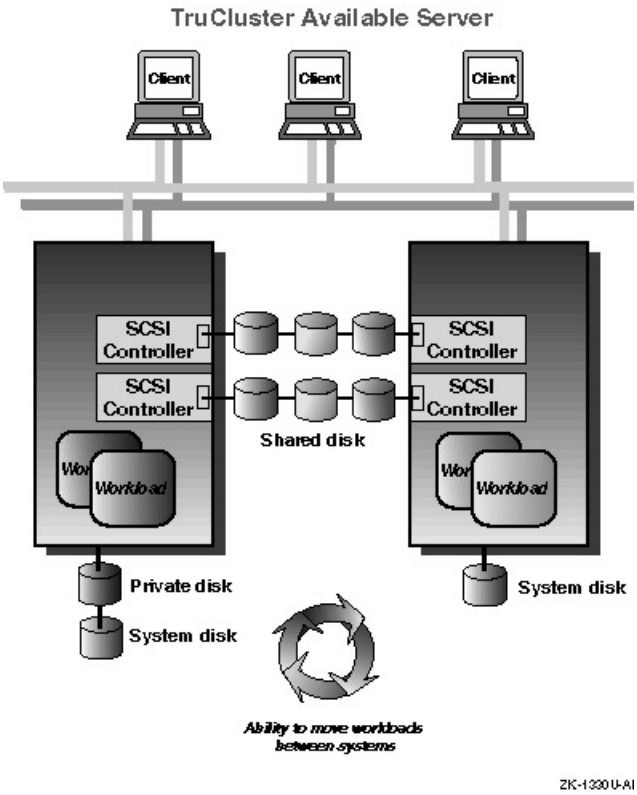
The TruCluster Available Server makes mission-critical computing a reality by keeping applications and network services running even when systems or components fail. TruCluster Available Server is designed for computing environments that can tolerate a short disruption, but need critical applications automatically restarted. By the time a system manager can be paged to report a problem, TruCluster Available Server has failed over to another server and users are back at work.

TruCluster Available Server suits customers who can occasionally tolerate a few seconds to a minute of downtime. The advantage of tolerating occasional downtime is economy: TruCluster Available Server adds only modest cost, as compared to the high costs for truly fault-tolerant systems.

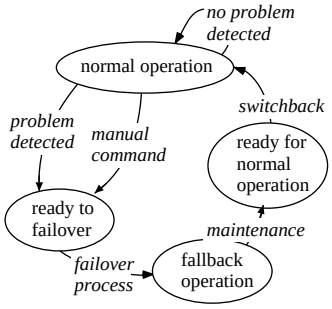
TruCluster Available Server combines the advantages of symmetric multiprocessing and fault resilience, as well as providing multihost access to shared disks and a generic failover mechanism, making applications and data highly available.

An Available Server cluster can support up to four nodes in a cluster configuration using any of the broad range of AlphaServer systems.

The following figure shows a typical high-availability configuration. Here, two servers each run a set of applications or network services. Each system runs an independent workload.



Each system monitors the health of the other by watching for “heartbeat” signals sent over both network and SCSI channels. This dual monitoring system ensures reliable failure detection, while differentiating among network, I/O, and host failures.



If one of the systems stops signaling, TruCluster Available Server recognizes the problem and initiates a failover of applications to the remaining system(s). A failover can also be initiated by limited failures such as a bad network card or software problem, or even on a system manager's command for planned shutdowns or online load balancing. It is not necessary to fail over entire systems. Individual services can be failed over as needed for maintenance—or just to balance the load evenly. A flexible failover capability is crucial to smooth and effective cluster operation.

After a failover is initiated, a recovering system takes over the failed unit's storage devices and network identity. At a high level, this involves either the Advanced File System (AdvFS) or a database management system (DBMS) from companies such as Oracle, Informix, Sybase, Computer Associates, or SAP AG. This recovery takes just a few seconds for AdvFS. DBMS recovery time will vary, depending on the DBMS and the size of the database involved. TruCluster Available Server configurations accomplish failover within 15 to 30 seconds, which is the fastest failover rate in the industry.

TruCluster Available Server configurations can run in two modes: Concurrently Active, where each node in a cluster can run a service, and Master Standby, where one system in the cluster runs all services and services are failed over to a designated Standby node in case of failure. In terms of fast file recovery, there is storage management software functionality built inside the Tru64 UNIX operating system that enables fast file recovery and increased data integrity. AdvFS is a journaled, local file system that provides higher availability, greater flexibility, and recovery than traditional UNIX file systems. The Logical Storage Manager (LSM) is an integrated, host-based solution to data storage management.

Flexible Configuration

You can customize the TruCluster Available Server to meet practically every situation. It comes with configuration files that support common services such as the Network File System (NFS) and popular relational databases. You can use just about any program or network service. Configurations are flexible to meet growing needs.

Up to four systems are currently supported, each in either concurrently active or standby mode.[10] In Concurrently Active mode, each system performs useful work. If one system fails, its workload will be assumed by the remaining system(s). The only drawback to this configuration is that once a failover occurs, N systems will be doing the work of N+1 systems, leading to temporarily diminished performance. Diminished performance may be avoided by using standby systems, or by configuring excess capacity to handle multiple workloads.

In Master Standby mode, one host remains on line but idle, ready to immediately pick up the load of another system if a problem occurs. This allows for predictable, steady post-failover performance. The cost is having one generally idle system. You can minimize this cost with three-way "pair and spare" or four-way "trio and spare" configurations. Of course, the

[10] Soon to be extended to eight nodes.

standby need not be completely idle. It can run development work or other lower-priority tasks that can be temporarily pushed aside in the event of a primary system failure.

The Cluster Monitor provides a graphical view of the cluster configuration. This allows you to determine the current state of availability and connectivity within a configuration. A system administrator can invoke management tools, which will allow the configuration to be managed from a single location.

Shared tape, a feature for high availability backup servers, provides support for the NetWorker Server, as well as allowing for the configuration of a single high availability NetWorker Server. Shared tape also allows shared device allocation to be controlled by availability/failover functionality. Failover situations will trigger tape service failover.

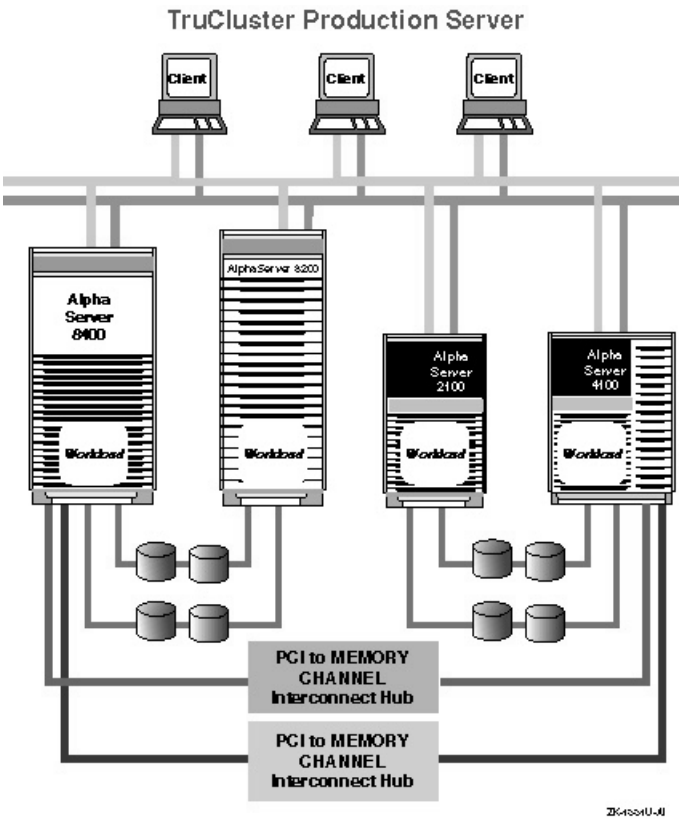
Online service modification makes the task of modifying storage management components much easier, because modifications can be made without taking any service off line. Therefore, the cluster will be kept highly available and other services will continue running. These modifications can also be made without taking the service off line.

Coupled with the reliability features in AlphaServers and Tru64 UNIX, TruCluster Available Server provides such a high level of availability that applications software, rather than the system, becomes the weak link. Thus the purchase of an expensive fault-tolerant system improves overall availability only marginally.

TruCluster Production Server

TruCluster Production Server software consists of the TruCluster Available Server Software and the TruCluster Memory Channel Interconnect Software to provide both high availability for mission-critical applications as well as high-performance database applications.

This product combines the advantages of SMP, distributed computing, and fault resilience to achieve high availability, scalability, centralized cluster management, and high performance.



TruCluster Production Server provides support for eight nodes using any of the broad range of AlphaServer systems.

Database management systems (DBMS) represent the core of many users' client/server applications. Clustering solutions must take special care to accommodate DBMS requirements. Making databases highly available is very useful, but it would be spectacular if database performance were to increase steadily as systems are added to a cluster.

TruCluster Production Server can achieve this for databases that can be partitioned. The database must be able to be explicitly split into several pieces, each of which can run independently of the others on a different system. Transaction management packages such as Tuxedo and ACMSxp can make this partitioning easier.

The Connection Manager defines cluster membership. It forms a cluster, adds nodes, detects node failures, and reconfigures a cluster around membership changes. It also establishes and maintains communication paths between the cluster members.

The Distributed Raw Disk (DRD) allows a raw disk-based, user-level application to run within a cluster, regardless of where in the cluster the physical storage is located. DRD also allows applications, such as distributed database systems, parallel access to storage media from multiple cluster members. DRD can provide all cluster members access to RAID volumes.

The Distributed Lock Manager (DLM) synchronizes access to resources that are shared among cooperating resources throughout the cluster. DLM also provides a software library that applications use to implement a resource sharing policy. DLM provides services that notify a process owning a resource that it is blocking another process requesting the resource. An application can also use DLM routines to coordinate the application's activities efficiently with the state of the cluster.

Parallel Commercial Software

Many key components of the commercial processing infrastructure support parallel operations. These include Oracle Parallel Server (OPS) and Oracle8, and Informix Extended Parallel Server (XPS), which decompose the DBMS task into a series of server processes that can be spread across multiple processors or cluster nodes. Clever decoupling of cache coherency and transaction integrity allows these parallel DBMSes to scale performance in cluster deployment. Because the database designers have parallelized their server code, application performance increases without requiring application recoding.

Commercial applications, such as the SAP R/3 comprehensive business management package, also run in parallel mode with Oracle Parallel Server.

MEMORY CHANNEL INTERCONNECT

The memory channel interconnect is a primary feature of the TruCluster Production Server product. As discussed earlier, the memory channel interconnect is a high-performance, PCI-based interconnect that cluster members use to pass low-overhead messages among themselves. TCP/IP is supported over the memory channel interconnect. A software library provides a special set of application programming interfaces (APIs) for high-performance data delivery over the memory channel interconnect.

Memory channel has an API library that you can use to create a clusterwide address space that is visible to all hosts within a cluster. Processes on the cluster communicate with each other by writing data to this address space, and reading data that other processes write to the address space. To use the memory channel address space, a process maps a region of the address space into its own process virtual address space.

The high bandwidth and low latency of the memory channel interconnect allow it to function just as in-system shared memory. The Oracle OPS and Informix XPS products integrate the memory channel interconnect into normal operations almost as though running on an SMP system. The result is SMP-class scalability that is extended beyond the half-dozen or so processors generally used in SMP database configurations. Unlike most SMP servers, if any TruCluster node fails, the entire cluster keeps running.

Disaster Tolerance

TruCluster technology provides continuous computing, even in the event of disaster. Fire, floods, earthquakes—they can strike without warning and leave you without access to mission-critical applications. With TruCluster technology you will have a predictable recovery strategy. You can locate disaster-tolerant cluster nodes miles apart, linked by high-performance network products. This ensures the availability of applications and data despite the complete loss of systems in your local facility. The Compaq CustomSystems Division will consult with you to discuss a wide range of performance and pricing options to determine your precise requirements, as well as the appropriate levels of availability to meet these requirements.

Tru64 UNIX Storage Management

The exceptional storage management tools in Tru64 UNIX—the Tru64 UNIX Logical Storage Manager (LSM), Advanced File System (AdvFS), and NetWorker—are fully utilized. The Tru64 UNIX storage information is quite extensive. A partial list of capabilities includes extremely flexible storage layout (including mirroring, file-by-file striping, and physical place-

ment), fast system startup (based on journaling), enormous files (up to multiple terabytes[11]), extensive online performance optimizations, defragmentation, disk load balancing, and automated backup and restore capability.[12]

Future Enhancements

Tru64 UNIX TruCluster products solve users' availability and scalability problems today, and they do so in a cost-effective manner. However, Compaq will continue to evolve the TruCluster products, incorporating new features and technologies. Enhancements and new features that are being worked on now include:

- Dynamic load balancing for network services and connections for the TruCluster technology.
- Continue to achieve applications and network services to take advantage of TruCluster Software. However, some software—particularly database managers, middleware servers, and high-end technical applications—benefit from explicit tuning. Over time, more off-the-shelf software will be cluster-tuned.
- TruCluster Software will provide a cluster file system that is optimized for concurrent data access in a cluster environment through load balancing, enhanced disaster tolerance, and cluster aliasing. The TruCluster File System will be similar to OpenVMS Clusters, including root, which makes the cluster easier to administer, allows applications to scale more easily, and makes management of the cluster much easier.
- Integration with NonStop Clusters, including additional interconnect support such as Servernet, enabling TruCluster Software to provide fault tolerant application programming interfaces.

Conclusion

The Compaq TruCluster products provide the highly scalable, highly available software needed in today's client/server environments. Whether you want to accelerate power-hungry applications, run thousands of transactions every minute, build multiterabyte data warehouses, rapidly provide data to thousands of users, ensure maximum uptime, or do all of these simultaneously, Compaq can help with affordable, dependable solutions.

---

[11] Terabytes are thousands of gigabytes.

[12] For more information, see "AdvFS: The Advanced File System for Commercial UNIX" (a Compaq white paper).