Compaq Computer Corporation

Positioning the *AlphaServer* GS Series for High Performance Technical Computing

Contents

The new AlphaServer GS Series systems are at the heart of the next generation of computing. With them, Compaq provides a significant business advantage to customers across a range of environments including e-commerce and the Internet, telecommunications, high-performance technical computing, and business intelligence/enterprise applications.

Abstract: The new Compaq *AlphaServer* GS Series with *Tru64* UNIX or *OpenVMS* software delivers highly scalable application performance along with the availability, capacity for growth, manageability and security required by high-performance computing customers. Based on a modular design, the new GS systems can be scaled from four to 32 processors. Their advanced Non-Uniform Memory Architecture (NUMA) supports symmetric multiprocessing with cache coherency across all processors in the system, using a high-performance, non-blocking switch. Applications can also use message passing protocols such as MPI or PVM instead of SMP parallelism to improve performance scaling. With future releases of the *Tru64* UNIX operating system and Compaq compilers, users will be able to optimize SMP performance using data and thread placement features to take advantage of locality of reference in applications. Very Large Memory (VLM) configurations—up to 256 gigabytes—enable technical users to run the largest and most demanding jobs at high speed with no data partitioning.

New reliability and availability features make these systems virtually unstoppable, supporting on-line repair and maintenance by reducing and minimizing scheduled and unscheduled downtime. The systems offer unsurpassed operational flexibility and hence lower total cost of ownership. Complete “lights out” support means you never have to be with the system in person to manage it.

Take advantage of the new AlphaServer GS systems when:

- Reliability, availability, and manageability are paramount
- SMP parallelism is required
- VLM configurations (more than 32 GB in global memory) are required

Applications center on database management and multi-job throughput

Technical customers looking for scalable performance over a large number of CPUs should also evaluate Compaq’s HPC160 and 320 systems and AlphaServer SC supercomputers. These systems offer the same or larger processor counts, but use clustering technology to support message passing between nodes. On a per processor basis, they deliver message-passing application performance equivalent to the GS Series at a substantially lower cost, while the GS Series offers cache coherency across the entire system and enhanced reliability, availability, and manageability feature.

New GS Series Product Overview

Product Description

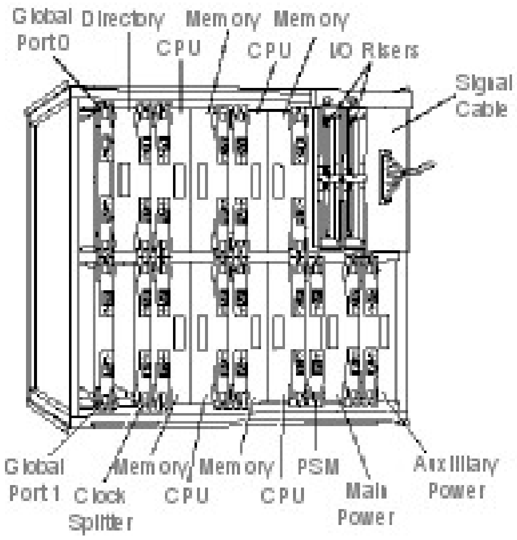
The Compaq *AlphaServer* GS80 is a single-rack system that supports up to eight Alpha EV67 processors, 64GB of memory and 16 PCI buses with 56 PCI slots and more than 13GB/s of aggregate internal bandwidth. The GS80 supports up to two system partitions and will match or exceed the performance of today's *AlphaServer* GS60E and many GS140 systems. The *AlphaServer* GS80 represents a natural growth path with minimal incremental cost for ES40 customers who require more capacity and scalability. The GS80 provides binary compatibility for applications running on the same versions of all operating systems and without any changes to the networking, storage or operational environment. The *AlphaServer* GS80 offers significantly better performance and value than other vendors' eight and 16-way servers such as the HP N-Series or the Sun Ultra Enterprise 3500 or 4500 systems. With reconfiguration, a single-CPU *AlphaServer* GS80 system can grow into a fully configured 32-way *AlphaServer* GS320 system.

The *AlphaServer* GS160 is a two-cabinet system supporting up to 16 CPUs, 128GB of memory and 32 PCI buses with 112 PCI slots, yielding more than 26GB/s of aggregate internal bandwidth. Available first, the GS160 supports up to four system partitions and has demonstrated two to four times the performance of typical *AlphaServer* GS140 systems , for which it represents a replacement and growth path. The GS160 will deliver higher levels of application performance at less cost to customers than 16-, 24- and 32-way systems from Sun (e.g., the UE6500) or HP (e.g., the HPV2600). With minimal operational disruption, the GS160 can grow into a fully loaded 32-way *AlphaServer* GS320 system.

With the addition of a third CPU cabinet, the *AlphaServer* GS system becomes the *AlphaServer* GS320, capable of supporting up to 32 Alpha EV67 CPUs, 256GB of memory, 64 PCI buses with 224 PCI slots offering over 51GB/s of aggregate bandwidth, and up to eight system partitions. The Compaq *AlphaServer* GS320 represents a high-end replacement and growth path for the *AlphaServer* GS140 systemand will deliver higher levels of application performance than the largest systems from HP (V2500), IBM (S80) or Sun (E10000).

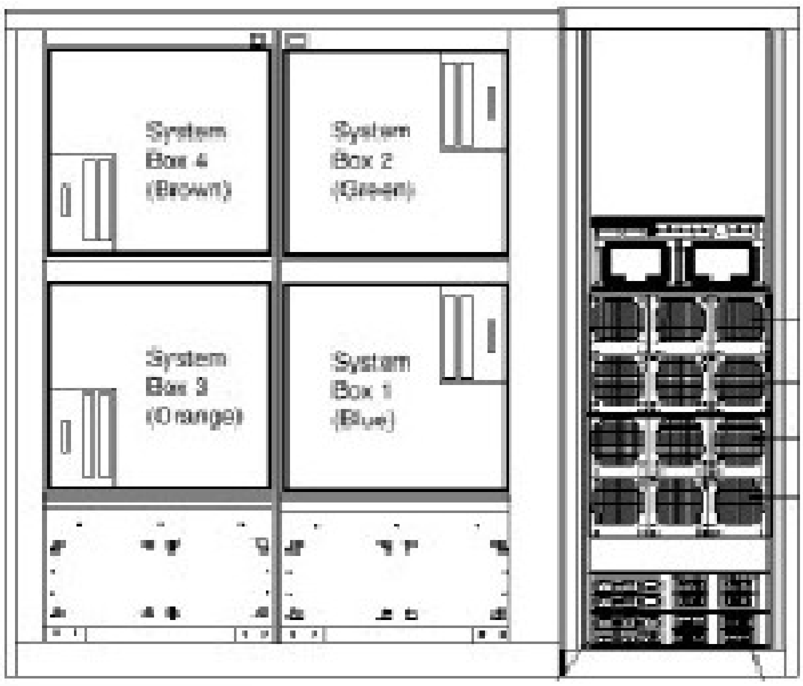
All *AlphaServer* GS systems will support future generations of Alpha microprocessors, including the forthcoming EV7 and EV8 microprocessors. Future Alpha microprocessors will not only dramatically increase application performance, but GS Series systems will also support 48 or more Alpha processors in the same package. This will lead to a 20-fold system performance increase over the next few years and ensure that customers will be able to satisfy their computing needs for even the most "explosive growth" business scenarios.

Product Technologies



The new *AlphaServer* GS Series is based on a fundamentally modular design. The basic unit is a “quad building block” (QBB) and contains up to four Alpha EV67 CPUs, 32GB of memory on 4 carriers (32-way interleaved for high performance) and 28 PCI options on eight PCI buses. CPUs, memory and I/O are connected by a high-performance, low-latency, non-blocking switch with 1.6GB/s I/O bandwidth and 6.4GB/s total CPU-to-memory bandwidth. The switch takes advantage of the EV67 processor’s parallelism , especially its ability to manage a large number of simultaneous outstanding memory requests. There are four QBBs each in front and in back of the *AlphaServer* GS320. Inside the system cabinet, two QBBs are packaged with a “System Box” for connection to power.

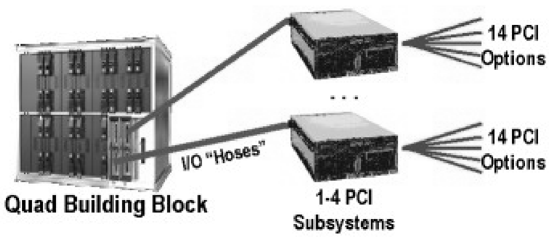
QBBs are aggregated together to build larger systems. Connecting two



QBBs back to back, for example, creates the eight-processor *AlphaServer* GS80. The 16-processor *AlphaServer* GS160 and the 32-processor *AlphaServer* GS320 systems are created by interconnecting four or eight QBBs through a second high-performance, lowlatency, non-blocking Global Switch. In these larger systems, Compaq also offers Capacity on Demand, which means that these resources are acquired with the system and turned on when needed. Customers have instantaneous access to the additional capacitywithout interrupting system operation.

I/O Subsystem

I/O cables run from the QBBs to PCI subsystems.
These I/O hose cables can run as long as 10 meters to ease the cable congestion that often accompanies the connectivity requirements of enterprise-class systems. . I/O racks can even be positioned across a walkway or in another room.
Each of the PCI card cages supports four PCI buses and 14 slots. For very large I/O subsystems or for higher I/O bandwidth, customers can add an expansion cabinet to house the additional PCI card cages. Each PCI expansion box is connected to a QBB by two cables.



Availability and Manageability

The design center for high performance technical computing systems is, of course, scalable performance, and Compaq's new GS Series provides unprecedented capability and capacity in running SMP workloads. However, technical customers running production workloads require levels of availability and reliability found in the most demanding commercial systems as well. The GS Series excels in these qualities as well.

The detailed availability and manageability features of the new GS Series are documented in detail elsewhere (reference). Here are the highlights:

- On-line maintenance and upgrades
- Internal network of independent service processors
- Resilient power
- External subsystems (storage, networks, system consoles) with no single points of failure and resilience in the server to their failure
- Complete hardware and environmental reporting
- Physical system design to facilitate serviceability and simplicity

Non-Uniform Memory Architecture

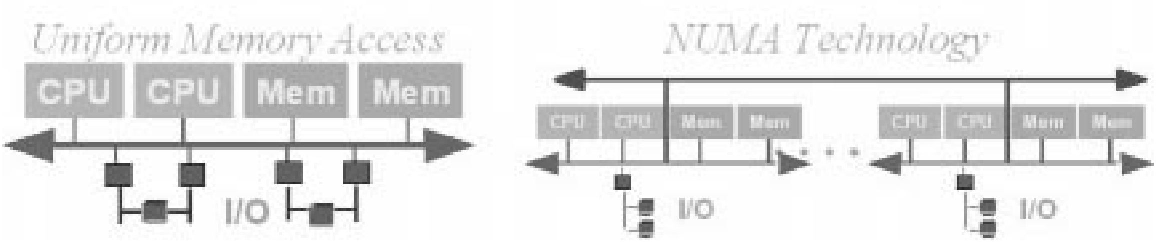
Processors in the GS Series can access both memory within the same QBB ("local memory") and memory in other QBBs ("remote memory"). Local access is through the QBB's on-board switch , whereas remote access involves transactions across both local and global switches.

That characteristic classifies the system as a Non-Uniform Memory Architecture (NUMA) system. NUMA is the critical architectural element for growing the system beyond simple bus and/or switch models of the past.

A primary determinant of computer system performance is the memory latency perceived by processors. Bus-based UMA systems feature a single memory latency value, but tend to exhibit latency growth of three to four times their nominal latency under load. The AlphaServer GS Series latency profile shows remote reference latencies only a factor of three times local reference latencies, and the latency profile exhibits very little growth under system load. The AlphaServer GS Series is therefore able to continue delivering high sustained performance to applications when under heavy load.

Systems that use older architectures also face physical limits on scalability. As processors are added, the interconnect becomes overloaded. Once saturation occurs, adding more processors will actually decrease overall system performance by increasing interprocessor communication overhead.

In the *AlphaServer* GS Series, the switch, memory, and processors within a QBB are balanced. Adding QBBs also adds interconnect and memory bandwidth, so the balance can be maintained. The Global Switch maintains cache coherency and provides the bandwidth required for remote memory transactions across a fully loaded GS320 system. The result is scalable performance from four to 32 CPUs.



Memory transactions that traverse the Global Switch have, of course, a much longer path than local transactions that traverse only one QBB switch., Therefore, remote memory transactions have greater latencies than local memory transactions.. In earlier (non-Compaq) NUMA implementations, this remote memory latency could increase five to 10 times because of several switch levels and inefficient routing protocols. The GS Series' global switch, however, combines a single switch level with highly efficient routing algorithms and sufficient bandwidth to keep the remote memory latency penalty relatively low and nearly constant under heavy load. The result is a breakthrough—a NUMA system with the performance characteristics and software features of a well-balanced SMP system.

In a NUMA system, by definition, there is some performance penalty when processes access non-local memory (that's what "non-uniform" means). Intelligent placement of data and processes in the same QBB, if possible, by the operating system or the application programmer will minimize such penalties. Compaq is developing enhancements to its application development environment that allow high performance technical computing customers to take full advantage of this new architecture. These features include:

- Additional system functions available from *Tru64* UNIX
- Enhancements to the Compaq C, C++, and Fortran programming languages
- Improvements in debugging on NUMA machines
- New features for profiling of parallel applications

Techniques for developing parallel applications

Developers of HPTC codes employ a number of techniques to implement parallel applications. For distributed memory systems, message passing, implemented with message passing libraries such as MPI or PVM, is used to share data between processes.. Using Compaq MPI and Compaq PVM, this message passing approach can also be used with Compaq's shared memory systems and will continue to work well with the new Alpha GS series.

For shared memory systems, parallel applications are developed using multiple threads that share data in memory ,frequently implemented by using a threads library, such as PTHREADS, or through compiler directives such as OpenMP. However, in NUMA systems, all memory doesn't behave the same, even though it can be addressed as a single address space, as in a conventional SMP system. In particular, memory latency for access to memory near a CPU is shorter (better) than the latency to access memory that is far from the CPU . Optimal performance is obtained when computations occur on CPUs that have fast access to data . Thus, to optimize performance on NUMA architectures, programmers need features called "data placement," the ability to control the placement of data into a set of memories, and "thread placement," the ability to assign a thread to a CPU for which the required data is in local memory.

Tru64 UNIX is providing additional services for the GS series that will enable both data placement and thread placement. For threaded applications, these features allow programmers to assign memory pages and threads closely, thereby minimizing use of the global switch.In addition, *Tru64* UNIX provides memory and thread allocation techniques that aid in allocating threads and data for existing applications.

Programming Language Extensions

Compaq is adding a set of directives to Compaq Fortran for *Tru64* UNIX that support data placement and thread placement in OpenMP applications for the GS series. They extend the OpenMP Fortran API to allow the programmer to place data near OpenMP threads and, if desired, to control the placement of both data and OpenMP threads onto memories in the machine.

Compaq C and C++ will provide features that permit programmers to place data and threads explicitly. Features will also be provided that permit application developers to control the migration of pages of data to the threads that use them, thereby moving the data close to the threads.

Tool Support

To facilitate code development, capabilities have been added to the Compaq software development tools. For debugging, both Ladebug and TotalView will be enhanced to support the data layout extensions.

Key debugging capabilities will support Fortran OpenMP extensions and represent different data layouts. TotalView, a product from Etnus, Inc., already has extensive support for debugging OpenMP codes.

Additional profiling capabilities are provided via DCPI. The Compaq (formerly Digital) Continuous Profiling Infrastructure provides continuous, low-overhead profiling for *AlphaServer* systems. Still other features statistically estimate timings for non-local memory referencesto help developers identify bottlenecks and tune applications for optimal performance.

Availability of Optimized Parallel Solutions on GS Series

Optimized parallel capabilities will be available in twostages. *Tru64* UNIX support will be included with the 5.1 release, currently scheduled for June 2000. The compilers and tools will undergo extensive beta testing during the second half of 2000 and will be released in January 2001. Fully optimized end user applications from ISVs and public domain developers should be available during the first half of 2001.

HPC160 and 320 Product Overview

Compaq's family of clustered servers, the HPC160 and the HPC320, are 16- and 32-CPU systems configured and tuned for high-performance technical computing applications. Based on Compaq's industry-leading *AlphaServer* ES40 servers that link four of the latest 64-bit Alpha 21264 processors via a high-bandwidth crossbar switch, the HPC160 and HPC320 are configured to deliver computational power for HPTC applications that use message passing protocols for parallelism.

These parallel servers harness the SMP power of four or eight ES40 servers in an integrated system using Compaq's Memory Channel II system area network. In this computational environment, large, distributed-memory parallel applications can run on as many as 16 or 32 processors for maximum application performance, or several smaller, single-processor or shared-memory parallel applications can run within the quad-processor building blocks for maximum application throughput.

By using Memory Channel II and Compaq’s *TruCluster* environment, the parallel servers are able to take advantage of Compaq’s extensive experience with system area networks. Compaq’s parallel software development environment includes MPI and PVM libraries optimized for Memory Channel and, combined with workload management software from ISVs such as Platform Computing (LSF) and Gridware (Codine, GRD), it has enabled us to deliver large-scale systems for many years. The HPC series packages a System Area Network as a single system, configured and tuned for HPTC applications, and provides simplified system management through the *TruCluster* Version 5 cluster file system and associated single system image management tools.

Many HPTC applications support either multi-job, shared-memory parallelism on the building blocks or single-job, distributed-memory parallelism across the HPC160/320 systems in areas such as:

- Government, academic and military R&D
- Engineering simulation
- Biological and chemical sciences
- Weather forecasting
- Digital content creation
- Satellite image processing

The HPC160 and HPC320 provide a lower-cost alternative to the GS160 and 320 for applications that run well in a distributed-memory environment, or for SMP applications that run most effectively in a quad-processor SMP system but need more CPUs for throughput.

While the HPC160 and 320 are targeted at ISV applications requiring up to 32 CPUs, there is a need for much larger systems with hundreds or thousands of CPUs to support traditional supercomputing applications. Compaq’s *AlphaServer* SC systems meet this need.

AlphaServer SC Series Product Overview

The *AlphaServer* SC Series, Compaq's solution for TeraFLOPS computing, is a family of supercomputers offering unique scalability and upgradability. By using a combination of existing processor and system designs, high-performance interconnects and innovative software, Compaq is able to deliver true supercomputer performance at affordable (in supercomputer terms!) price points.

The flexible software configuration of the *AlphaServer* SC system, enables it to be usedin multiple ways:

- A scalable, massively parallel processor (MPP) for single distributed applications that currently run on existing MPP machines like the Cray T3D/E or IBM SP-2
- A high-capacity throughput machine for large numbers of independent sequential, shared-memory parallel, or message-passing jobs
- A software development platform for parallel applications using shared or distributed-memory programming paradigms with OpenMP directives, multithreading, MPI, or Shmem operations.

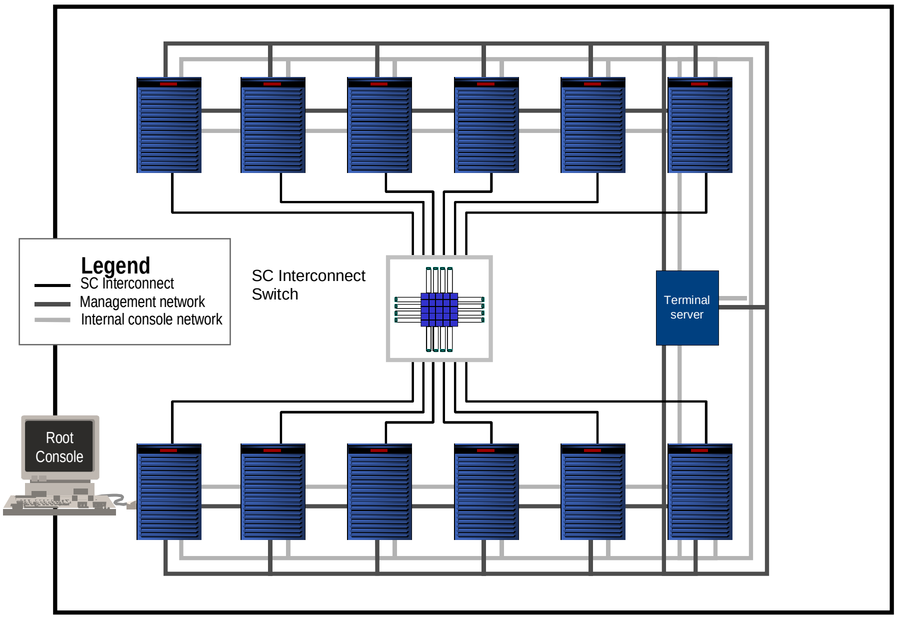


Figure 1: *AlphaServer SC* System Architecture

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 peak performance of close to 800 GFLOPS. The nodes are connected by the *AlphaServer* SC Interconnect, with an inter-node MPI bandwidth of over 200 MB/s and industry-leading MPI latency of under 5.5 microseconds.

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, elements of the multi-system aggregation must be hidden from the user. Single System Management (SSM) capability, provided as part of SC system software manages the *AlphaServer* SC 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:

- 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 VAXcluster systems to current Alpha and *Tru64* UNIX *TruCluster* systems. CFS and its operating system support provide much of the basis for the SC integrated system management software. By globally sharing 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, need be performed only once.

- 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 jobs. When a user executes a job, RMS is responsible for selecting the set of nodes 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 directly take advantage of the AlphaServer 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 SC systems with Vampir and VampirTrace from Pallas GmbH.

Comparing the GS, HPC, and SC Series

Table 1 provides a detailed comparison of the GS, HPC, and SC Series of *AlphaServer* systems. The GS Series ranges from eight to 32 processors, the HPC Series from 16 to 32, and the SC Series from 64 to 512.

All three series use four-way SMP systems as building blocks (ES40s in the HPC and SC Series, QBB subsystems in the GS Series), but only the GS Series provides cache coherency to support the SMP programming model across a full system. The HPC and SC Series deliver SMP programming within four-way nodesand message passing for parallelism between nodes.

Switches connecting four-way building blocks vary considerably among the series. The GS Global Switch delivers 12.8 GB/second of system bisection bandwidth versus only 800 MB/second for the MC II interconnect in the HPC Series. The QSW Elan switch in the SC Series, however, delivers scalable bandwidth; each connection or “rail” to a node adds 200 MB/second to the system. A 128-way *AlphaServer* SC with one rail per node delivers roughly 13 GB/second bisection bandwidth, but adding rails can scale this up indefinitely.

The GS Series is managed as a true single system, with one copy of the operating system running across all CPUs (unless, of course, the system is partitioned). The HPC and SC Series are cluster implementations, although *TruCluster* V5 provides tools to manage the clusters as single systems. The SC Series adds resource management features like soft partitioning and gang scheduling that supercomputer users require.

Performance Comparisons

Figures 1-4 provide examples of GS, HPC, and SC Series performance compared to competitive systems. These results are based on initial versions of the GS hardware and software; further optimizations should improve performance.

Figure 1 demonstrates the GS Series leadership performance in multi-stream applications that require high memory bandwidth and no data sharing. GS systems deliver two to three times the performance of HP, Sun, and IBM systems running the same workload. Customers can take advantage of this leadership in workloads that involve running many jobs on different processors concurrently.

Figure 2 shows GS and HPC Series performance running Fluent, Inc.’s large Computational Fluid Dynamics benchmark. GS320 performance ranges from 1.3 times that of IBM’s SP2 Power3 375MHz system to 3.4 times that of the Sun UE10000. Note that the HPC320 500MHz system is slightly lower in performance than the GS 320 but still higher than all competitive systems. Since the 667 MHz ES40 shows better performance from one to four CPUs than the GS Series, a 667 MHz HPC320 is likely to be somewhat higher in performance than the GS320 as well.

Figure 3 shows performance data for LSTC’s LS-Dyna crash test code. The GS Series shows slightly better performance from four to 16 processors than a 667 MHz AlphaServer SC, primarily because this application generates many small messages and is therefore sensitive to interconnect latency. Both Compaq systems are substantially faster than the competitive systems shown.

Table 1: System Comparison

	AlphaServer SC Series	HPC160 / 320 (ES40)	GS160 / 320
Building Block			
Building Block	AlphaServer ES40, four processor switch-based SMP server	AlphaServer ES40, four processor switch-based SMP server	Quad building block (QBB), four processor switch-based SMP subsystem
CPU	EV6 at 500MHz or EV67 at 667MHz	EV6 at 500MHz or EV67 at 667MHz	EV67 at 729MHz
L2 Cache	4MB/CPU (EV6) 8MB/CPU (EV67)	4MB/CPU (EV6) 8MB/CPU (EV67)	4MB/CPU
Memory Capacity	Up to 4GB/CPU now, 8GB/CPU later	Up to 4GB/CPU now, 8GB/CPU later	Up to 8GB/CPU
Aggregate Memory Bandwidth	5.2GB/sec	5.2GB/sec	6.4GB/sec
Memory Interleaving	Up to 4 ways	Up to 4 ways	Up to 32 ways
PCI Busses / Slots	2 PCI busses, up to 10 slots	2 PCI busses, up to 10 slots per ES40	Up to 8 PCI busses per QBB, 224 slots per system
Upgrade path	Board upgradeable from EV6 to EV67, EV68 and EV69 System upgrade to EV7	Board upgradeable from EV6 to EV67, EV68 and EV69 (with 8 MB L2 cache) System upgrade to EV7	Board upgrade from EV67 to EV68 and EV69 Cage upgrade to EV7
System (32 CPUs)			
Architecture	Distributed Memory, SMP nodes	Distributed Memory, SMP nodes	ccNUMA
System Interconnect	QSW Elan 8-way switch	8-way Memory Channel II	Global Switch
Bi-Section Bandwidth	Scalable up by adding rails at 200MB/sec bi-directional per rail	800MB/sec	12.8GB/sec
Programming model	SMP within the building blocks, Message passing (MPI, PVM, etc.) across the building blocks	SMP within the building blocks, Message passing (MPI, PVM, etc.) across the building blocks	SMP with NUMA extensions for data locality, or distributed memory (MPI, etc.)
Software			
Operating system	Tru64 UNIX V5	Tru64 UNIX V4.0f or later (TruCluster V5 requires Tru64 UNIX V5.0a)	Tru64 UNIX V4.0g or later (NUMA extensions require V5.1)
System management	AlphaServer SC software supports system wide gang scheduling, resource management and performance monitoring. PFS and Cluster File System software enable the management of up to four 32-node CFS domains	Tru64 UNIX V4.0f / TruCluster V1.6: Each ES40 building block has its own system disk. Parallel System Management Utility simplifies management of multiple systems Tru64 UNIX V5.0a / TruCluster V5: Via CFS, the cluster has a single system disk (single system image), such that the cluster is effectively managed as one system (requires shared FiberChannel storage)	Managed as a single system, although can be partitioned to run different operating systems or versions of operating systems in each QBB
Workload / Batch Queue Management	Various workload management products including LSF from Platform Computing and Codine/GRD from Gridware (formerly Genias)	Various workload management products including LSF from Platform Computing and Codine/GRD from Gridware (formerly Genias)	Various workload management products including LSF from Platform Computing and Codine/GRD from Gridware (formerly Genias)

Figure 1

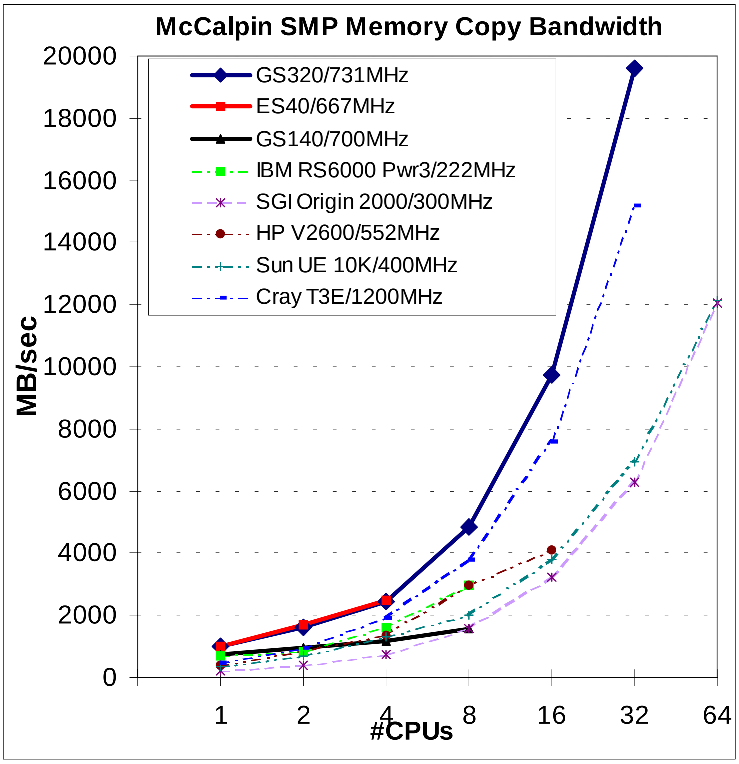


Figure 2

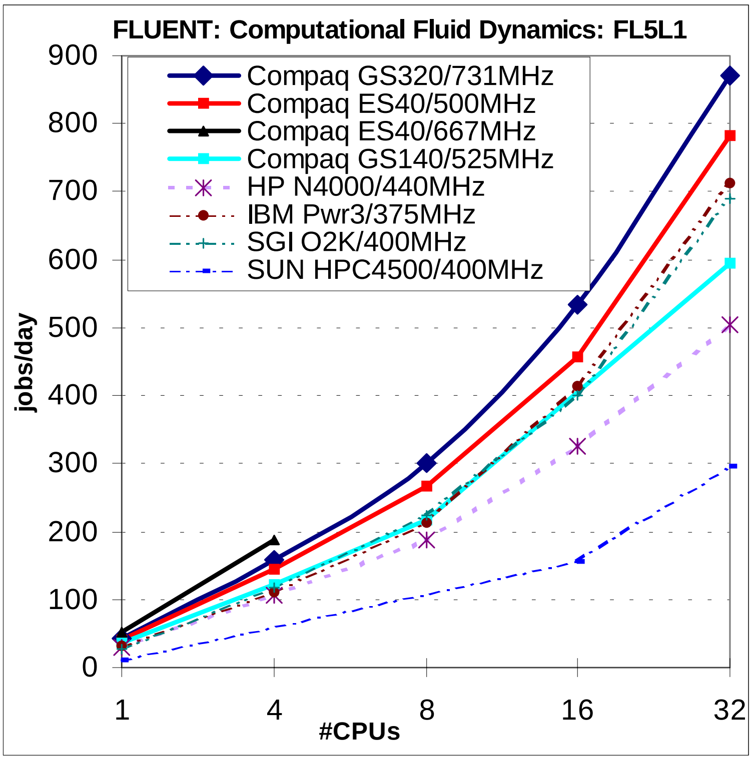


Figure 3

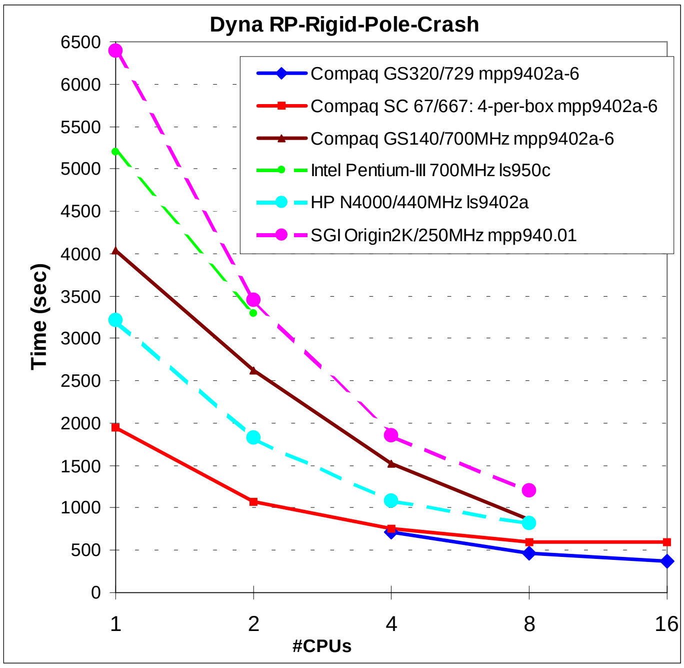
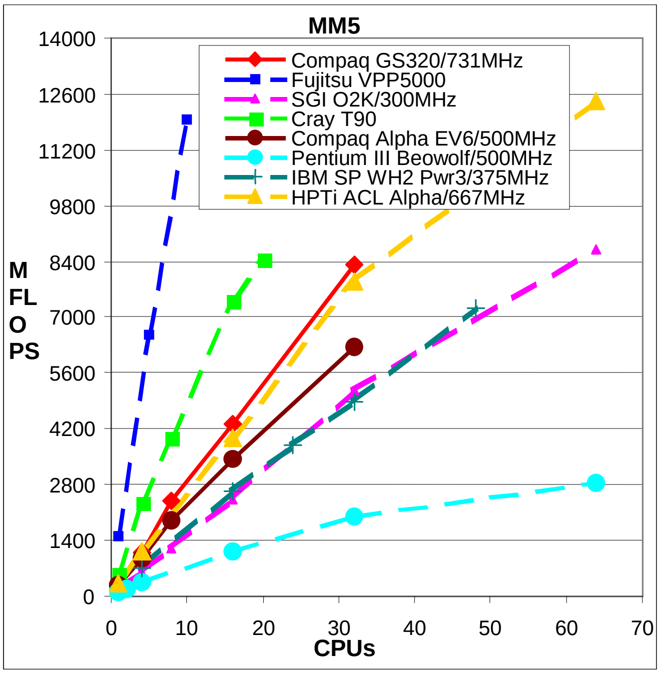


Figure 4 shows system performance running MM5, a weather modeling code developed by NCAR and Penn State. Here, the GS and HPC Series are compared with a combination of vector systems, IBM and SGI servers, and Linux clusters. While the vector systems still provide the highest performance per CPU, the GS320 offers performance equal to a Cray T90. A 667 MHz Alpha Linux cluster from HPTi actually demonstrates the highest absolute performance with 64 CPUs. Here again, the 500 MHz HPC320 is slightly below the GS320 in performance, but well above competitive systems.

Figure 4



Summary

Perhaps the most fundamental benefits of the new GS Series are these: support of the true SMP programming model across a full 32-CPU environment, support for a single cache-coherent shared memory of up to 256GB, and breakthrough system management, serviceability, and availability features. The SMP performance of the GS Series leads the industry today for data management and data throughput applications and will evolve to industry leadership in SMP parallel HPTC applications, as Compaq's software tools for NUMA optimization are released in the second half of 2000. Expect fully optimized SMP parallel ISV applications to be available on the GS Series in the first half of 2001.

The most fundamental benefit of the HPC Series is cost-effective performance scaling of distributed memory applications or workloads from 16 to 32 processors. HPC Series cannot run SMP parallel jobs across nodes and do not provide the large global memory or RAMP features of the GS Series.

The SC Series offers higher performance and dramatically higher scalability than the HPC or GS Series for distributed memory applications or workloads. The SC Series starts where the GS and HPC Series stop — at 64 CPUs — and currently scales to 512 CPUs. SC offers a more robust set of resource management features than either the GS or HPC Series, including a parallel file system and gang scheduling, to support the needs of supercomputing centers.

Compaq *AlphaServer* GS Series Product White Paper: An Introduction to and overview of the new Compaq *AlphaServer* GS Series