COMPAQ

Performance Brief

May 2000
12KT-0500A-WWEN

Prepared by Business Critical Servers

Compaq Computer Corporation

Contents

AlphaServer GS Family Benchmark Performance

Abstract: This paper presents benchmark information for the *AlphaServer* GS family. An overview and description of each benchmark is provided, along with the measured performance on that benchmark.

Notice

AlphaServer GS Family Benchmark Performance
Performance Brief prepared by NA Enterprise Computing Group

First Edition (May 2000)
Document Number 12KT-0500A-WWEN

Benchmarks and Performance

Computer hardware performance is measured in terms of the maximum rate the system can achieve in executing instructions. The most common measures have traditionally been Millions of Instructions Per Second (MIPS) and Millions of FLoating point Operations Per Second (MFLOPS).

Hardware specifications are of limited value, as they only address maximum theoretical performance and do not measure realistic system performance. Other factors in the system design, such as memory bandwidth and latency and I/O performance, often limit a system to a small portion of its theoretical performance.

Additionally, hardware performance measurements and system architectures are not standardized, making it difficult to directly compare vendor provided specifications.

Because of this, there has been a strong migration toward measuring actual performance with *benchmarks.*

Benchmarks

Another approach to performance measurement is to develop a program and to measure the performance of the system running this program. This has the advantage of measuring actual performance that can be achieved on the system. If the program is designed to be portable between systems it can also allow direct comparisons between systems. For this reason, organizations, such as SPEC (Standard Performance Evaluation Corporation) and BAPCo (Business Applications Performance Corp.), and academic/research institutions, such as CERN (European Particle Physics Laboratory), have developed and made tools available to provide cross-platform tests that can help users compare the performance of different platforms. We will refer to these tests as industry-standard benchmarks. Other tests, based on specific applications and test sets, we will refer to as “application-specific” benchmarks. These application-specific tests can be either widely-accepted and virtual industry standards, such as Linpack, or simply a comparison of vendor or OEM supplied test files with a given application. We will briefly discuss the industry-standard benchmarks first, and then cover some of the leading application benchmark tools used.

An objection sometimes raised about industry standard benchmarks is that they do not measure total system performance, but focus on the performance of a single subsystem, such as CPU, floating point calculations or memory. This objection is correct but misleading. While most industry standard benchmarks *are* primarily intended to measure subsystem performance, they can be effectively used *in conjunction* with other measurements to determine overall system performance. Few dispute that industry standard benchmarks do an excellent job at what they are intended – measuring the performance that a system (both hardware and software) can actually achieve.

Several benchmarks exist for measuring CPU performance, including SPECint, SPECfp, and Linpack. Other standard benchmarks are used to measure other elements of performance, such as McCalpin Streams for memory bandwidth.

Application Benchmarks

Application benchmarks can be at the same time useful, misleading, and a difficult way to measure performance. Since computers are used to run applications, the only *real* metric for performance is application performance.

Application benchmarks are misleading because they are only valid for that specific application, and sometimes for a specific data set or test file. Even directly competitive applications that do essentially the same things may be dramatically different. Because of this, performance optimization for each application is different. Good performance on one does not necessarily indicate good performance on the other. Likewise, poor performance on one doesn't necessarily indicate poor performance on the other.

Also, different uses (as represented here by data sets or trail files) exercise different parts of the application and may use totally different features and functions.

Application benchmarking requires much work. There are very few broad based, comprehensive benchmarks that span a range of systems and allow easy comparisons. A notable example is TPC-C. This benchmark is run across a wide range of systems and allows easy comparison.

Benchmark Biases

All benchmarks are *biased*. Understanding this fact is critical to effective use of benchmark data. Biased doesn't mean misleading, it simply means that you need to understand *what* the benchmark is measuring, what systems are being compared, *how* they are being measured, and how the *results* are being used. The bias may be subtle or overt, the benchmark may be well designed or poorly designed, the characteristics being measured may be crucial or irrelevant, and the testing methodology may be valid or flawed. Education on the details of the benchmark is the only solution to guide one through the landmines of potential bias.

Good benchmarks are difficult to design. A good benchmark is one that provides a true indicator of the performance for which the system and the application were designed. Developing a benchmark that provides a true indicator of actual performance is: broad based, portable across different hardware and operating systems, easily run, easily reported, and easily interpreted. This is not a simple task.

An additional challenge arises when a benchmark becomes popular, and vendors begin *optimizing for the benchmark* – changing their system to provide higher performance on the benchmark while not improving application performance. This occurs with all popular benchmarks; reviewing the history of TPC-C reveals the efforts that the Transaction Processing Performance Council has gone through to ensure that hardware and software vendors do not implement optimizations specifically for the benchmark.

To summarize, benchmarks are a tool – a tool that can be used or misused. Well-designed benchmarks can provide valuable insights into performance. Poorly designed benchmarks may be highly inaccurate and misleading. And no single figure can capture all the information that is needed for a well-chosen system selection.

The following pages provide more information on some of the most popular benchmarks used today. Recent performance figures for the *Compaq AlphaServer* GS320 and competitive systems help put the benchmarks in perspective.

SPECint and SPECfp

Benchmark

SPEC CPU benchmark suite, with results for SPECint95 and SPECfp95.

Source

SPEC, the Standard Performance Evaluation Corporation, is a non-profit corporation formed to "establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers" (quoted from SPEC's bylaws). The founders of this organization believe that the user community will benefit greatly from an objective series of applications-oriented tests, which can serve as common reference points and be considered during the evaluation process. While no one benchmark can fully characterize overall system performance, the results of a variety of realistic benchmarks can give valuable insight into expected real performance.

SPEC basically performs two functions.

- SPEC develops suites of benchmarks intended to measure computer performance. These suites are packaged with source code and tools and are extensively tested for portability before release. They are available to the public for a fee covering development and administrative costs. By license agreement, SPEC members and customers agree to run and report results as specified in each benchmark suite's documentation.
- SPEC publishes news and benchmark results in *The SPEC Newsletter* and *The GPC Quarterly*. Both are available electronically through http://www.spec.org/ (and available in paper as a quarterly publication if necessary). This provides a centralized source of information for SPEC benchmark results. Both SPEC members and non-SPEC members may publish in the SPEC Newsletter, though there is a fee for non-members. (Note that results may be published elsewhere as long as the format specified in the SPEC Run Rules and Reporting Rules is followed.)

Description

In August 1995, SPEC introduced the CPU95 benchmarks as a replacement for the older CPU92 benchmarks. These benchmarks measure the performance of CPU, memory system, and compiler code generation. They normally use UNIX as the portability vehicle, but they have been ported to other operating systems as well. The percentage of time spent in operating system and I/O functions is generally negligible.

Although the CPU95 benchmarks are sold in one package, they are internally composed of two collections:

- CINT95, integer programs, representing the CPU-intensive part of system or commercial application programs
- CFP95, floating-point programs, representing the CPU-intensive part of numeric-scientific application programs

The actual benchmark runs can be performed two ways: single stream and multi-stream. The single stream test is designed for single-CPU systems. The multi-stream benchmarks, called the *rate benchmarks*, are designed for multi-processor systems.

The rate benchmarks run multiple copies of the core benchmark suite, and are used to measure the performance of large systems running a mix of jobs. Thus, the rate benchmarks are the best indicator of the performance potential of large systems.

Since the core SPEC benchmarks are designed to measure CPU performance, they stress the CPU, cache, and (to a lesser degree) the memory subsystem. They do not measure factors such as multi-processor performance, inter-processor communication or system level memory bandwidth. The surprising result can be that small systems can deliver better performance on the benchmarks than large systems.

The rate benchmarks, on the other hand, apply a load to the *entire system.* These benchmarks measure large system characteristics, fully stress large numbers of processors, and stress total system level memory performance. The rate benchmarks are valid for single CPU systems, and are the benchmark to use for comparing single CPU and multi-CPU systems.

Results are reported for CINT95 and for CFP95 on individual report pages, and results can be reported for either one or for both.

Integer benchmarks: CINT95

This suite contains eight benchmarks performing integer computations, all of them written in C. The individual programs are:

Number and name	Application
099.go	Artificial intelligence; plays the game of "Go"
124.m88ksim	Moto 88K chip simulator; runs test program
126.gcc	New version of GCC; builds SPARC code
129.compress	Compresses and decompresses file in memory
130.li	LISP interpreter
132.ijpeg	Graphic compression and decompression
134.perl	Manipulates strings (anagrams) and prime numbers in Perl
147.vortex	A database program

Floating-point benchmarks: CFP95

This suite contains 10 benchmarks performing floating-point computations. All of them are written in Fortran77. The individual programs are:

Number and name	Application
101.tomcatv	A mesh-generation program
102.swim	Shallow water model with 513 x 513 grid
103.su2cor	Quantum physics; Monte Carlo simulation

104.hydro2d	Astrophysics; Hydrodynamical Navier Stokes equations
107.mgrid	Multi-grid solver in 3D potential field
110.applu	Parabolic/elliptic partial differential equations
125.turb3d	Simulates isotropic, homogeneous turbulence in a cube
141.apsi	Solves problems regarding temperature, wind, velocity and distribution of pollutants
145.fpppp	Quantum chemistry
146.wave5	Plasma physics; electromagnetic particle simulation

What is Measured

Being compute-intensive benchmarks, these benchmarks emphasize the performance of the computer's processor, memory architecture, and compiler. It is important to remember the contribution of the latter two components; performance is more than just the processor.

Units

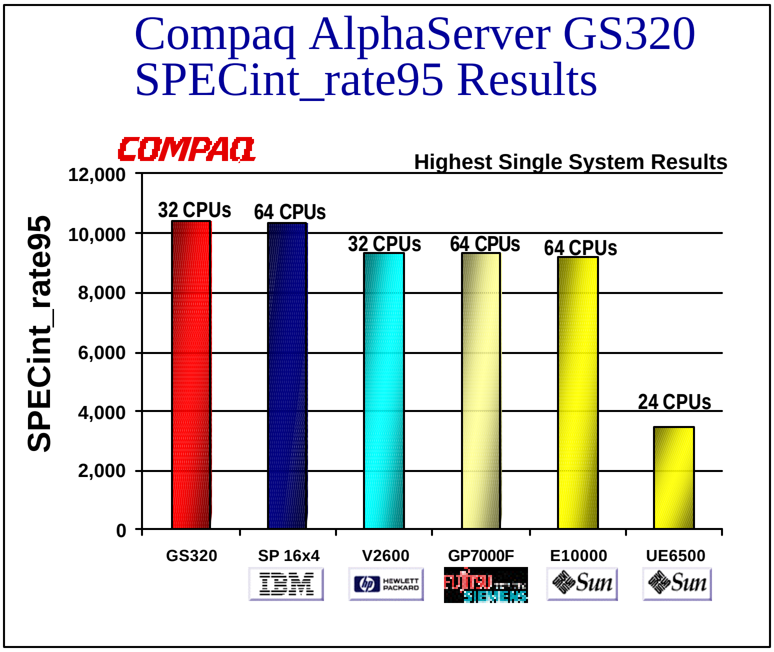
Larger is better. The results ("SPEC Ratio" for each individual benchmark) are expressed as the ratio of the wall clock time to execute one single copy of the benchmark, compared to a fixed "SPEC reference time." For the CPU95 benchmarks, a Sun SPARCstation 10/40 was chosen as the reference machine. A geometric mean of all benchmarks is used, which means that each benchmark contributes equally to the final result.

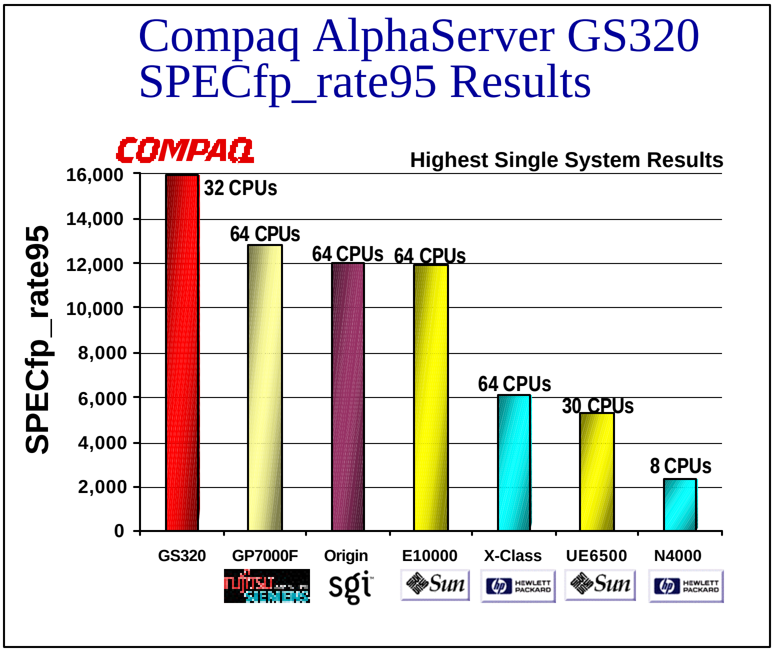
Additional Information on Benchmark

More information is available from http://www.specbench.org.

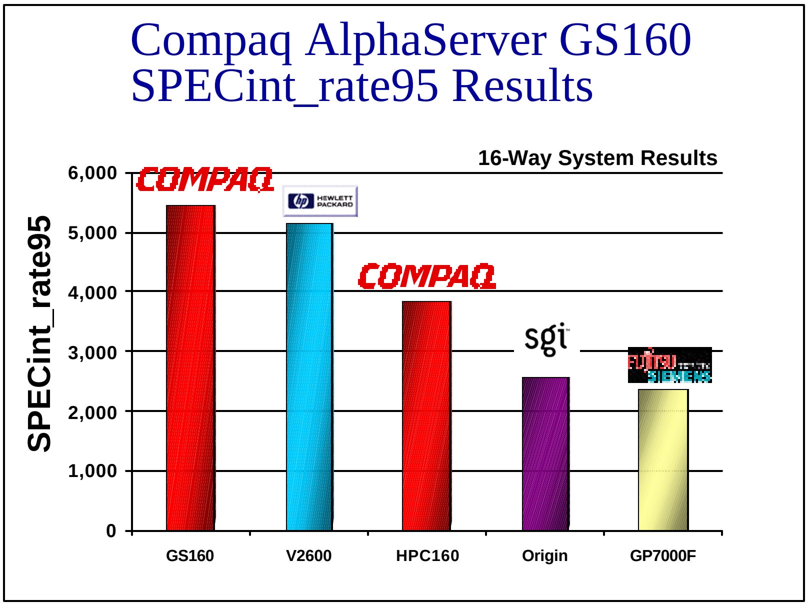
Compaq Results:

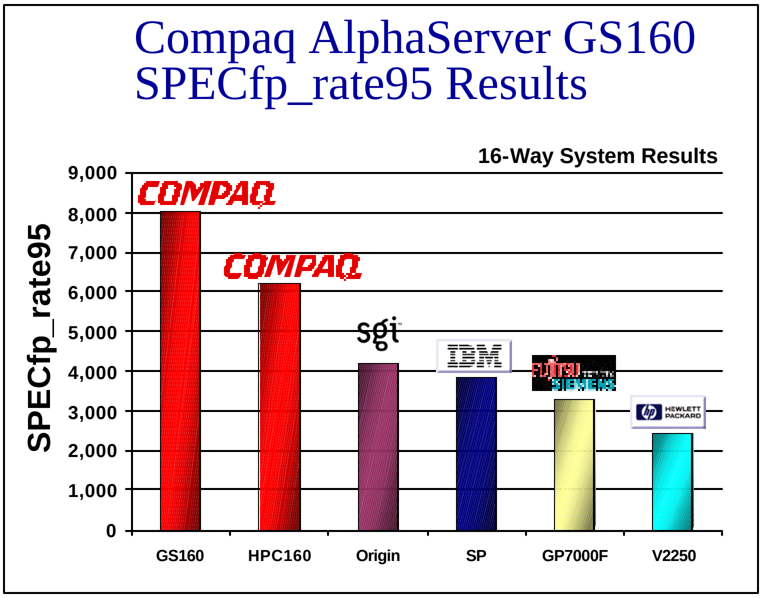
AlphaServer GS320 32 Processor Results:





AlphaServer GS160 16 Processor Results:





Sources:

Industry standard benchmark, available from the Standard Performance Evaluation Corporation (SPEC). More information available at http://www.specbench.org. *AlphaServer* GS family performance, as measured by Compaq and reported to SPEC in accordance with their standard submission process, have not yet been published by SPEC and must therefore be denoted as estimates. Competitive system performance numbers obtained from the SPEC reports contained on their Web site.

Notes

These are industry standard processor benchmarks that are run across a wide range of systems. This test clearly demonstrates the impressive integer instruction and outstanding floating point performance of the Alpha 21264 processor and the added benefit of the high-bandwidth memory and high processor count of the *AlphaServer* GS family.

SPECcpu2000

Benchmark

SPEC CPU benchmark suite, with results for SPECint2000, SPECfp2000, SPECfp_rate2000 and SPECint_rate2000.

Source

SPEC, the Standard Performance Evaluation Corporation, is a non-profit corporation formed to "establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers" (quoted from SPEC's bylaws). The founders of this organization believe that the user community will benefit greatly from an objective series of applications-oriented tests, which can serve as common reference points and be considered during the evaluation process. While no one benchmark can fully characterize overall system performance, the results of a variety of realistic benchmarks can give valuable insight into expected real performance.

SPEC basically performs two functions.

- SPEC develops suites of benchmarks intended to measure computer performance. These suites are packaged with source code and tools and are extensively tested for portability before release. They are available to the public for a fee covering development and administrative costs. By license agreement, SPEC members and customers agree to run and report results as specified in each benchmark suite's documentation.
- SPEC publishes news and benchmark results in *The SPEC Newsletter* and *The GPC Quarterly*. Both are available electronically through http://www.spec.org/ (and available in paper as a quarterly publication if necessary). This provides a centralized source of information for SPEC benchmark results. Both SPEC members and non-SPEC members may publish in the SPEC Newsletter, though there is a fee for non-members. (Note that results may be published elsewhere as long as the format specified in the SPEC Run Rules and Reporting Rules is followed.)

Description

Although the CPU95 benchmarks are sold in one package, they are internally composed of two collections:

- CINT2000, integer programs, representing the CPU-intensive part of system or commercial application programs
- CFP2000, floating-point programs, representing the CPU-intensive part of numeric-scientific application programs

The actual benchmark runs can be performed two ways: single stream and multi-stream. The single stream test is designed for single-CPU systems. The multi-stream benchmarks, called the *rate benchmarks*, are designed for multi-processor systems.

The rate benchmarks run multiple copies of the core benchmark suite, and are used to measure the performance of large systems running a mix of jobs. Thus, the rate benchmarks are the best indicator of the performance potential of large systems.

Since the core SPEC benchmarks are designed to measure CPU performance, they stress the CPU, cache, and (to a lesser degree) the memory subsystem. They do not measure factors such as multi-processor performance, inter-processor communication or system level memory bandwidth. The surprising result can be that small systems can deliver better performance on the benchmarks than large systems.

The rate benchmarks, on the other hand, apply a load to the *entire system.* These benchmarks measure large system characteristics, fully stress large numbers of processors, and stress total system level memory performance. The rate benchmarks are valid for single CPU systems, and are the benchmark to use for comparing single CPU and multi-CPU systems.

Results are reported for on individual report pages, and results can be reported for either one or for both.

What is Measured

Being compute-intensive benchmarks, these benchmarks emphasize the performance of the computer's processor, memory architecture, and compiler. It is important to remember the contribution of the latter two components; performance is more than just the processor.

Units

Larger is better.

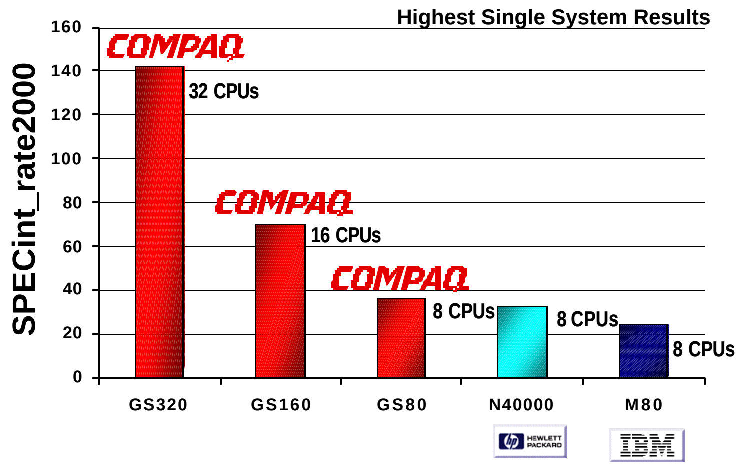
Additional Information on Benchmark

More information is available from http://www.specbench.org.

Compaq Results:

AlphaServer GS320 32 Processor Results:

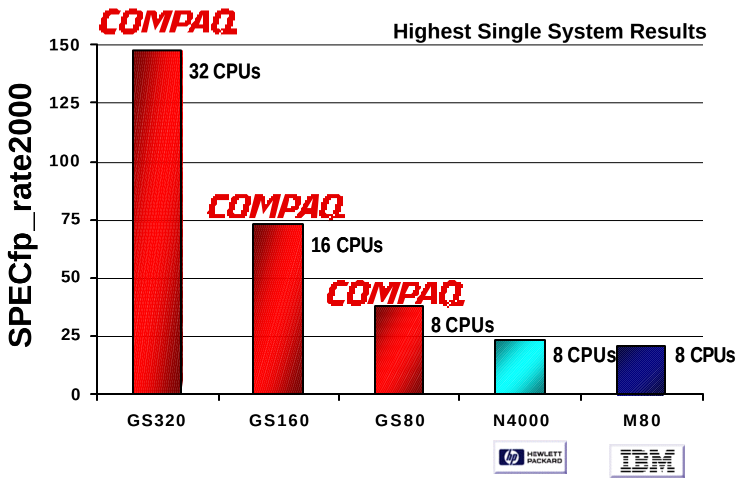
Compaq *AlphaServer* GS320 SPECint_rate2000 Results



Note: Sun results not available

COMPAQ 38 Updated 5/10/00
www.compaq.com
Compaq Confidential: Not for External Use or Presentation

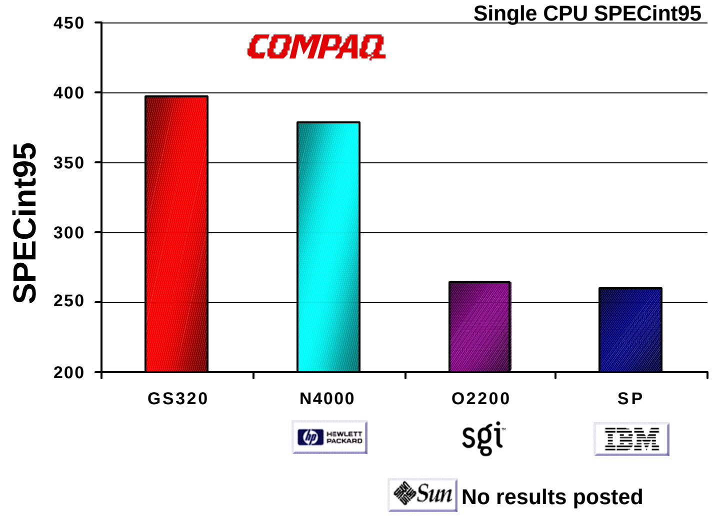
Compaq *AlphaServer* GS320 SPECfp_rate2000 Results



Note: Sun results not available

COMPAQ 36 Updated 5/10/00
www.compaq.com
Compaq Confidential: Not for External Use or Presentation

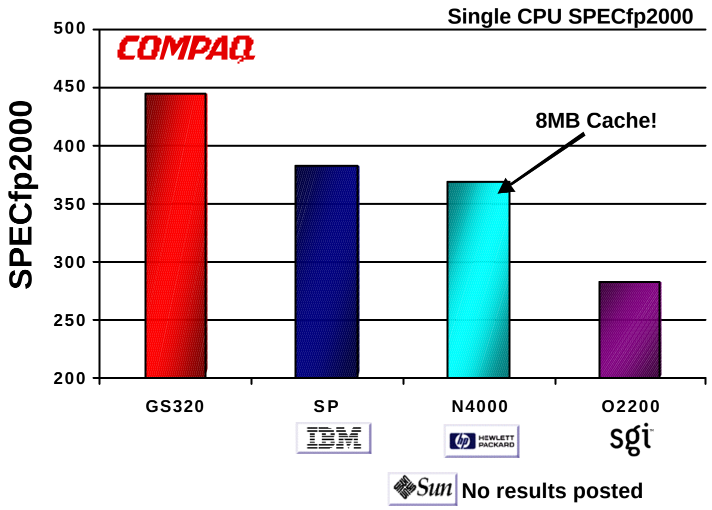
Compaq *AlphaServer* GS320 SPECint2000 Results



Updated 5/10/00

Compaq Confidential: Not for External Use or Presentation

Compaq *AlphaServer* GS320 SPECfp2000 Results



Updated 5/10/00

Compaq Confidential: Not for External Use or Presentation

Sources:

Industry standard benchmark, available from the Standard Performance Evaluation Corporation (SPEC). More information available at http://www.specbench.org. *AlphaServer* GS family performance, as measured by Compaq and reported to SPEC in accordance with their standard submission process, have not yet been published by SPEC and must therefore be denoted as estimates. Competitive system performance numbers obtained from the SPEC reports contained on their Web site.

Notes

These are industry standard processor benchmarks that are run across a wide range of systems. This test clearly demonstrates the impressive integer instruction and outstanding floating point performance of the Alpha 21264 processor and the added benefit of the high-bandwidth memory and high processor count of the *AlphaServer* GS family.

McCalpin Streams

Benchmark

STREAM

Source

The STREAM benchmark is available at ftp://ftp.cs.virginia.edu/pub/stream/.

Description

The STREAM benchmark is a simple synthetic benchmark program that measures sustainable memory bandwidth (in MB/s) and the corresponding computation rate for simple vector kernels.

The test consists of multiple repetitions of four kernels, described in the table below. The best results of (typically) 10 trials are chosen.

name	kernel	bytes/iter	FLOPS/iter
COPY:	a (i) = b (i)	16	0
SCALE:	a (i) = q*b (i)	16	1
SUM:	a (i) = b (i) + c (i)	24	1
TRIAD:	a (i) = b (i) + q*c (i)	24	2

What is Measured

The STREAM benchmark measures the impact of memory performance on floating point performance. The STREAM benchmark is designed to force problems out of cache and into main memory in the way that large technical computing applications do.

While the results are ostensibly reported in MFLOPS, the performance is more influenced by memory performance than by the floating point computational performance of the CPU.

Multiple threads of the benchmark are run on multiprocessor systems, allowing multiple CPUs and memory controllers to improve performance. The benchmark relies on the compiler to parallelize the benchmark. Thus, performance on multiprocessor systems depends on both hardware and compiler.

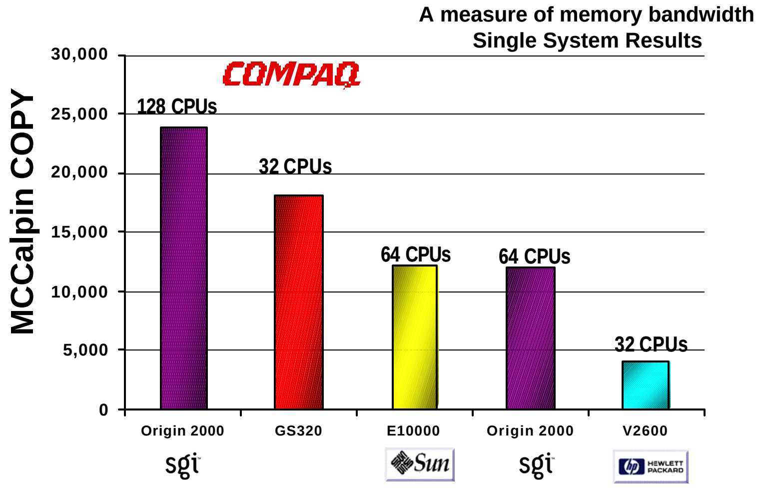
Units

MFLOPS (Millions of Floating point Operations per Second). Larger is better.

Results

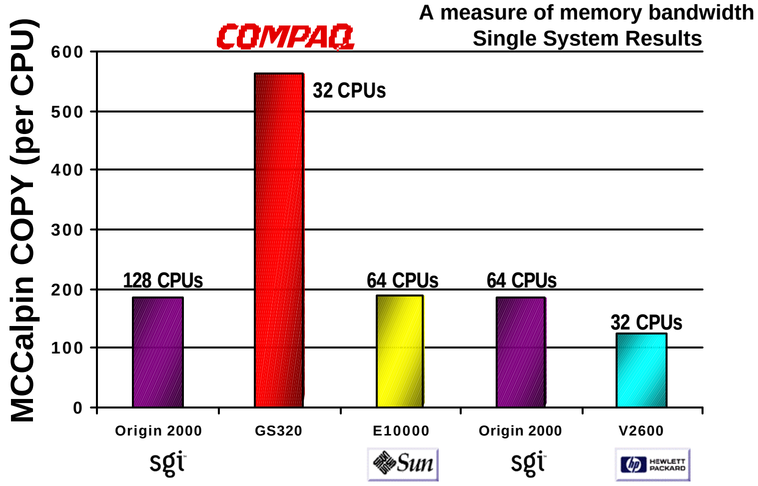
COPY: AlphaServer GS320 absolute performance

Compaq *AlphaServer* GS320
McCalpin “Copy” Results



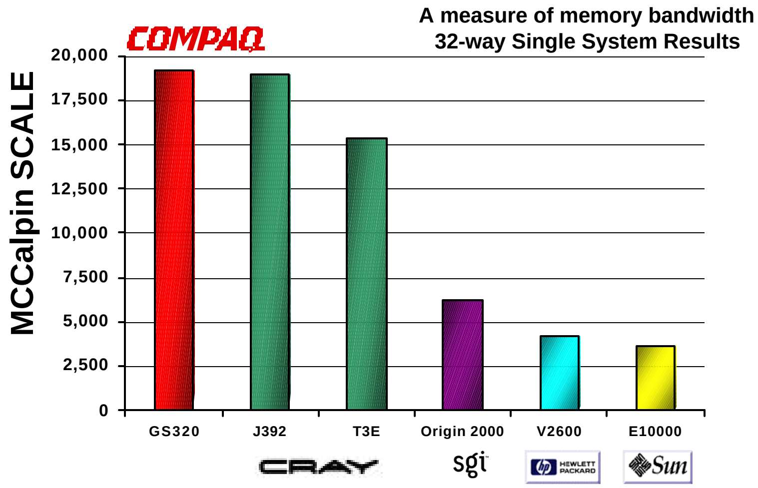
AlphaServer GS320 performance per CPU

Compaq *AlphaServer* GS320
The “King” of Bandwidth!



SCALE: AlphaServer GS320 absolute performance

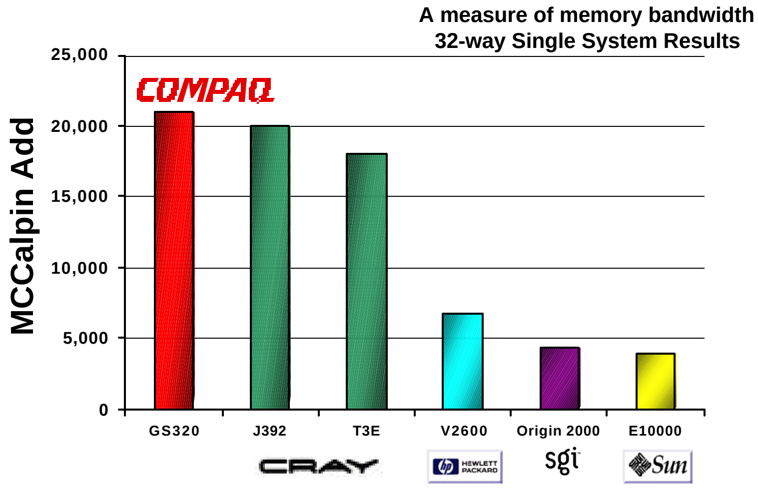
Compaq *AlphaServer* GS320
McCalpin "Scale" Results



Updated 5/13/00
www.compaq.com
49
Compaq Confidential: Not for External Use or Presentation

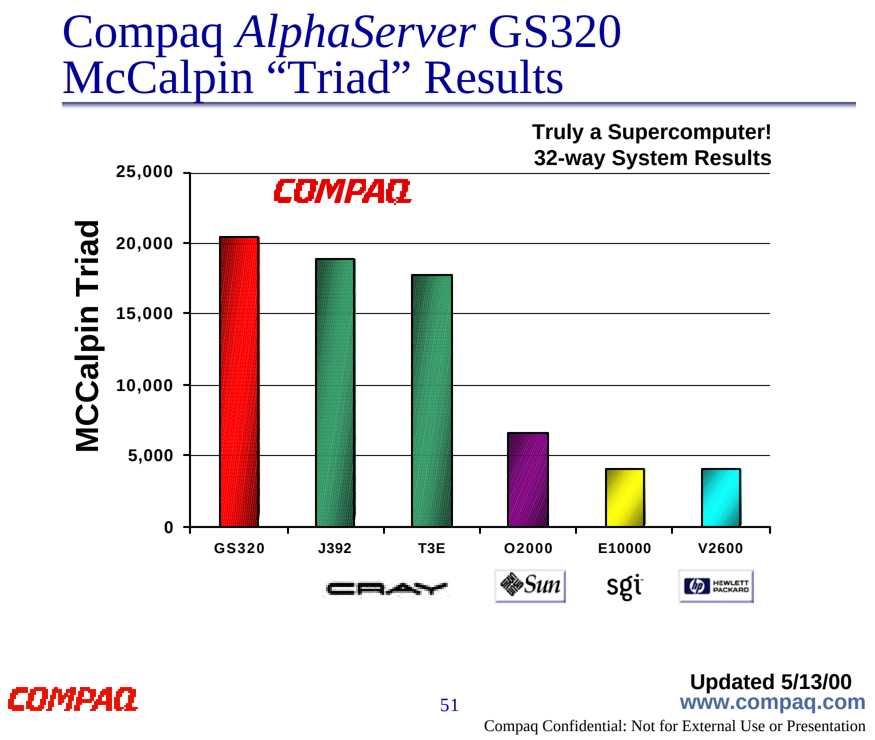
SUM: AlphaServer GS320 absolute performance

Compaq *AlphaServer* GS320
McCalpin "Add" Results



Updated 5/13/00
www.compaq.com
50
Compaq Confidential: Not for External Use or Presentation

TRIAD: AlphaServer GS320 absolute performance



Sources

Industry standard benchmark. STREAM was created and is maintained by John McCalpin mccalpin@cs.virginia.edu. Information is available at http://www.cs.virginia.edu/stream. Competitive information obtained from the web site. *AlphaServer* GS320 data based on testing by Compaq.

Notes

The results of Stream testing show the high system bandwidth of the *AlphaServer* GS320. The GS320 32 processor system delivers over 80% of the bandwidth of a 128 processor SGI Origin 2000, resulting in over three times the bandwidth per processor of the Origin 2000. The 32 processor *AlphaServer* GS320 has greater absolute bandwidth than a 64 processor Sun UE 10000, and much better bandwidth than a 32 processor HP V2600. This bandwidth advantage shows up in application performance.

SAP SD

Benchmark

SAP SD (Sales and Distribution) benchmark.

Source

SAP AG. and the SAP Benchmark Council. Information available at http://www.sap.com.

Description

The SAP R/3 benchmarks are designed to simulate real-life users. The SAP R/3 Standard Application Benchmarks consist of a number of script files that simulate the most typical transactions and workflow of an R/3 user. It uses a predefined SAP client database that contains sample company data that the benchmark is run against.

The SD (Sales and Distribution) benchmark involves creating an order with 5 line items, creating a delivery for this order, displaying the customer order, changing the delivery, post goods issued, list 40 orders for one sold-to party, and creating an invoice.

The SAP benchmarks are designed with a three-tier structure, consisting of a database server, application servers and a presentation server. The *AlphaServer* GS 320 was run as a single system with database and applications layers running on the single *AlphaServer* GS 320. The presentation server was running on a separate machine. This configuration was chosen to demonstrate the ability to integrate the entire server processing onto a single machine while showing the power and capabilities of the *AlphaServer* GS 320.

What is Measured

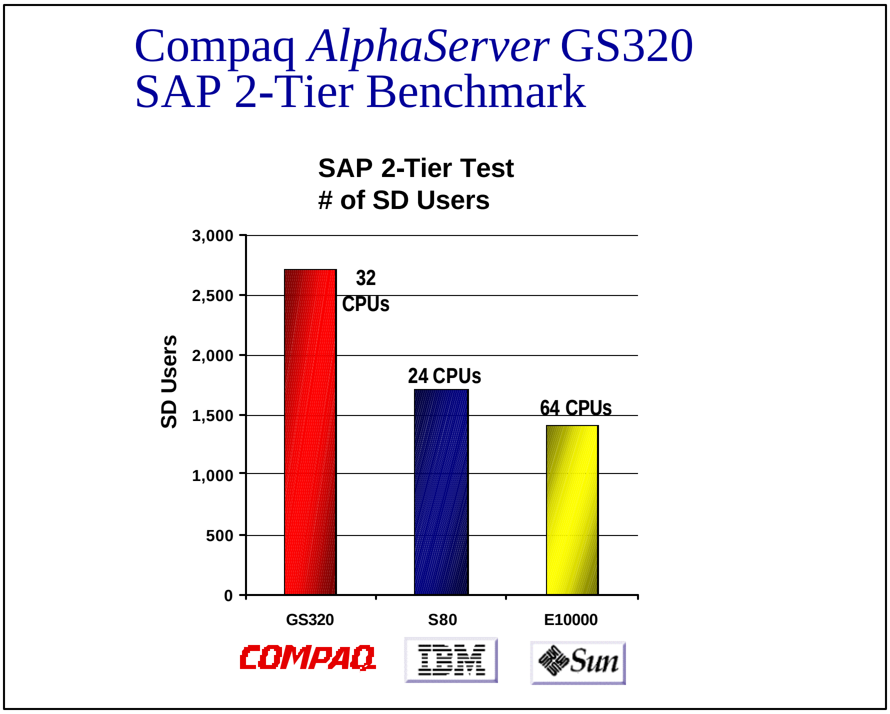
The SD benchmark measures the performance of a complete system running a large commercial workload. The workload includes database processing and applications logic.

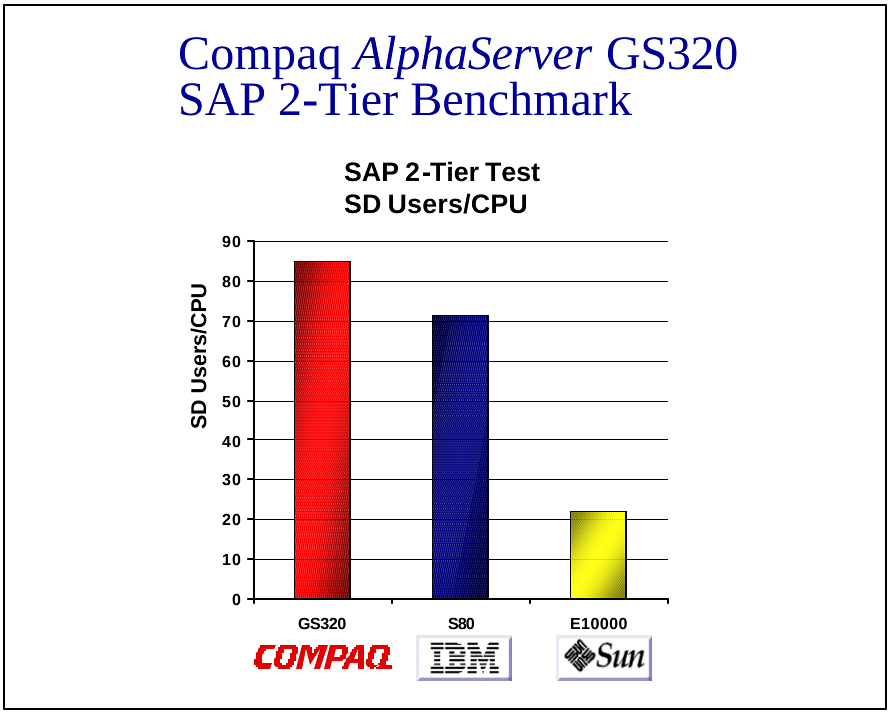
Units

The benchmark is reported in number of users supported, where larger is better. There are a series of conditions specified in the Benchmark Rules. These include a user “thinktime” of 10 seconds and the requirement of an average response time for dialog tasks of less than two seconds.

The throughput is measured by the unit SAPS (SAP Application Benchmark Performance Standard). 100 SAPS are defined as 2,000 fully business processed order line items per hour in the SD benchmark. This throughput is achieved by processing 6,000 dialog steps (screen changes) and 2,000 posting per hour in the SD benchmark.

Results


Compaq *AlphaServer* GS320
SAP 2-Tier Benchmark
SAP 2-Tier Test
# of SD Users
SD Users
3,000
2,500
2,000
1,500
1,000
500
0
32 CPUs
24 CPUs
64 CPUs
GS320
S80
E10000
COMPAQ
IBM
Sun


Compaq *AlphaServer* GS320
SAP 2-Tier Benchmark
SAP 2-Tier Test
SD Users/CPU
SD Users/CPU
90
80
70
60
50
40
30
20
10
0
GS320
S80
E10000
COMPAQ
IBM
Sun

Sources

Application benchmark. Information for this section obtained from the SAP web site at http://www.sap.com. Benchmark descriptions were obtained from Standard Application Benchmarks Descriptions, located on the web site. Competitive performance data is from the SAP web site. *AlphaServer* GS320 performance data provided by Compaq.

Notes

Two-tier and three-tier results cannot be directly compared. Three-tier testing allows clustering dozens or even hundreds of database servers and application servers. Total system size and system count must be considered when examining three-tier results.

Oracle Applications

Benchmark

Oracle Applications Standard Benchmark

Source

Oracle. Web site http://www.oracle.com.

Description

The Oracle Applications Standard Benchmark is focused on ERP applications, from the 11i e-business suite, simulating realistic customer scenarios using a selection of the most commonly used Oracle Applications modules. Definitions of transactions that compose the benchmark load were obtained through collaboration with implementation consultants and are representative of typical customer workloads, with both OLTP and batch components.

The database used in the benchmark has been designed to represent amounts of information that are typical of mid-market type businesses, whose annual revenues range from $100M to $500M. This database is provided with the benchmark kit and is common to all platforms on which the benchmark is available.

What is Measured

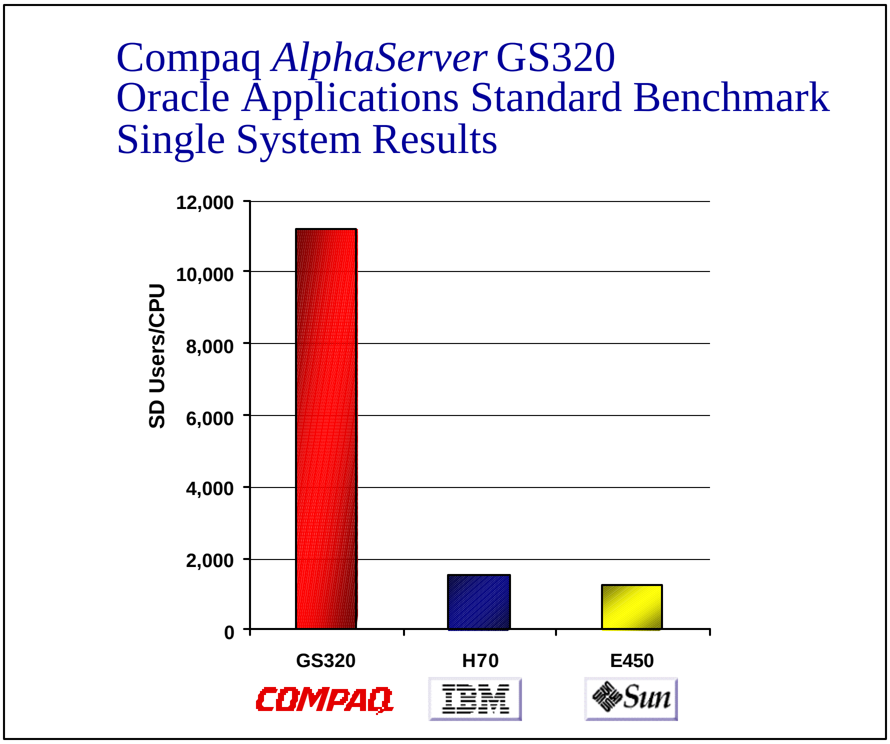
Total system performance running a large commercial workload.

Units

User Count, where larger is better. The user count measures the number of concurrent Oracle Applications users that the system can sustain while response times are kept under a pre-defined maximum value. User processes are defined by the type of transactions they execute and each user maintains a minimum transaction-per-hour rate.

Results

The *AlphaServer* GS320 delivered performance of 11,200 users. The Oracle Applications Standard Benchmark uses a robust transaction set, aggressive transaction rates and reasonable batch activity. Therefore, we expect that the GS320 benchmark results will translate well into real customer experiences.



Although these results are significantly higher than the majority of Oracle's installed base today, the benefits of Global Business Consolodation with Oracle Applications will quickly drive the requirement for the level of high-end computing provided by the new *AlphaServer* GS Series.

Sources

Application benchmark. Benchmark information obtained from the Oracle web site at http://www.oracle.com/apps_benchmark. Competitive performance results from http://www.oracle.com/apps_benchmark/html/results.html. *AlphaServer* GS320 performance data based on testing by Compaq.

Notes

Benchmark run on *AlphaServer* GS320 with 64 GB of memory. 20 *AlphaServer* ES40 systems were used as application servers, and an additional 20 Intel based systems running Loadrunner used to simulate users.

The GS320 32CPU system result is over 8x the scalability of the UE450 4CPU system result and over 7x the performance of the H70. This demonstrates the near linear scalability of the NUMA software and hardware technology used in the AlphServer GS series.

Fluent

Benchmark

Fluent FL5L1: Transonic Flow around a fighter aircraft.

Source

Fluent, Inc.

Description

Fluent is a leading Computational Fluid Dynamics (CFD) application, used to analyze the flow of gases and liquids. It is used in the automotive and aerospace industries, and in a wide range of industrial engineering applications.

In the FL5L1 benchmark, flow around the AGARD M-151 combat aircraft research model is computed. The simulation geometry contains canards and forward swept wings, but no tail. The conditions modeled were Mach number 0.9 and 10.46 degrees angle of attack.

Number of cells	847,764
Cell type	hexahedral
Models	RNG k-epsilon turbulence
Solver	coupled explicit

The FL5L1 benchmark is classified as a “large” benchmark by Fluent.

What is Measured

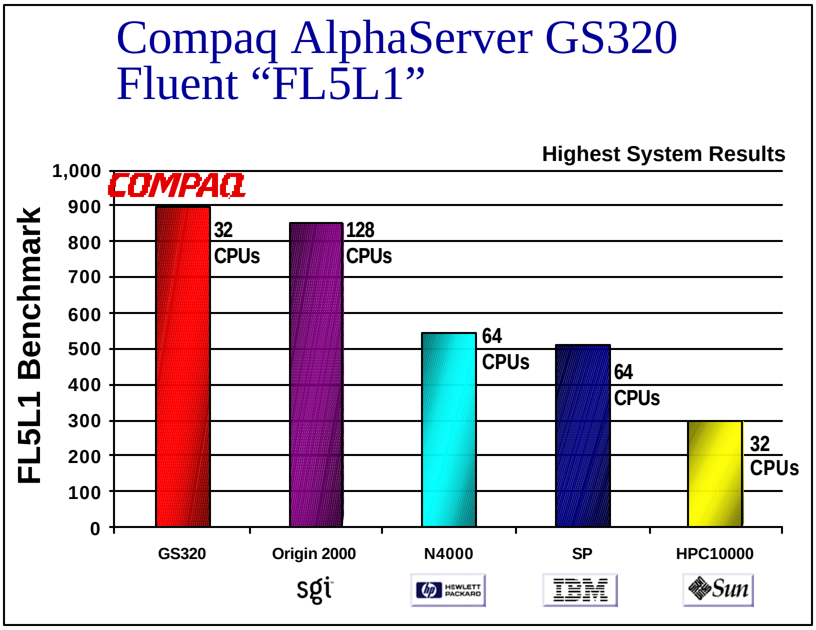
Fluent exercises the entire system, with particular emphasis on CPU, floating point, and memory performance.

Units

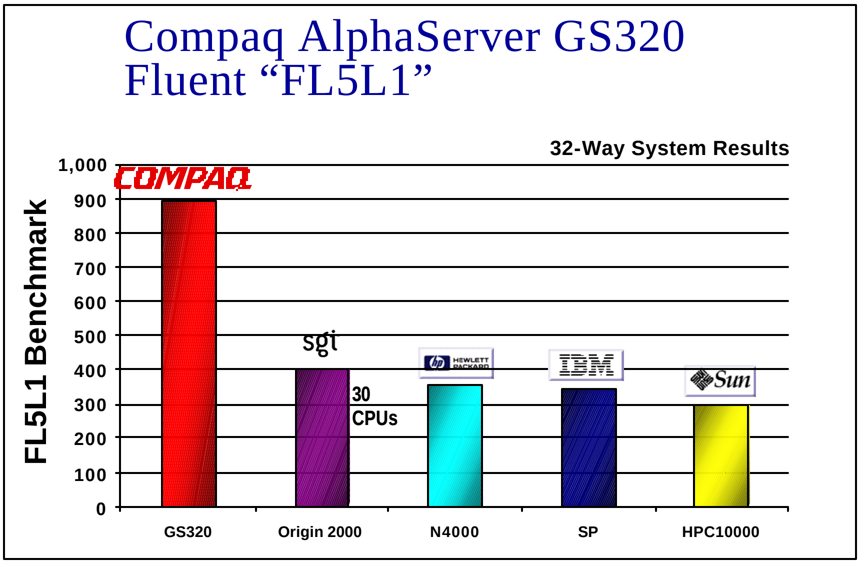
“Rating” is the primary metric used to report performance results of the FLUENT 5 Benchmarks. It is defined as the number of benchmarks that can be run on a given machine (in sequence) in a 24 hour period. It is computed by dividing the number of seconds in a day (86400 seconds) by the number of seconds required to run the benchmark. A higher rating means faster performance.

Results

AlphaServer GS320 absolute performance



AlphaServer GS320 32 CPU comparisons



Sources

Application benchmark. Descriptions and competitive results obtained from the Fluent, Inc. Web site at http://www.fluent.com. Benchmark information available at http://www.fluent.com/software/fl5bench/index.htm.

Notes

While Fluent supports systems with high processor counts, the results clearly show that individual CPU performance, memory bandwidth, and balanced system design yield the best results.