TECHNOLOGY BRIEF

May 1997

Compaq Computer Corporation

ECG Technology Communications

CONTENTS

Performance of Pentium Pro and Pentium II Processor/Cache Combinations

Many factors determine the ultimate performance of servers and workstations. Three of these factors are processor speed, cache size, and cache speed. Understanding the interactions of these factors in system performance is vital in making informed purchasing decisions. This paper outlines the role each of these factors plays in system performance and shows benchmark results for several application environments.

EXECUTIVE SUMMARY

For years, processor speed has been used as an indicator of system performance – the faster the processor, the faster the system. In some cases, this was an accurate assessment, and in other cases it was not. However, system hardware and software have advanced to the point where this relationship often does not hold true. Cache subsystem performance can have a dramatic effect on system performance in a specific application environment. This is especially true in multiprocessor systems.

Ordinarily, memory-intensive applications such as on-line transaction processing (OLTP) will have the best performance from processor/cache combinations with larger cache sizes. Systems with multiple processors also show the greatest performance boost from larger cache sizes. The performance of CPU-intensive applications and applications with small data sets will usually depend more on the processor speed.

The cache architecture of the Intel Pentium II processor differs from the architecture of the Intel Pentium Pro. Although the core processor speed of the Pentium II is higher, the Pentium II cache subsystem may not perform as well as a Pentium Pro cache of the same size in some situations. In addition, the Pentium II processor will not cache system memory in excess of 512 MB, and will not support more than two processors in an SMP system. Pentium II processors are well suited for CPU-intensive applications, systems with single or dual processor system configurations, and systems with no more than 512 MB of system memory. The Pentium Pro is the preferred processor for systems with system memory greater than 512 MB, for systems requiring more than two processors, and for systems used in memory-intensive applications.

Please direct comments regarding this communication to the ECG Technology Communications Group at this Internet address: TechCom@compaq.com

COMPAQ

NOTICE

The information in this publication is subject to change without notice and is provided "AS IS" WITHOUT WARRANTY OF ANY KIND. THE ENTIRE RISK ARISING OUT OF THE USE OF THIS INFORMATION REMAINS WITH RECIPIENT. IN NO EVENT SHALL COMPAQ BE LIABLE FOR ANY DIRECT, CONSEQUENTIAL, INCIDENTAL, SPECIAL, PUNITIVE OR OTHER DAMAGES WHATSOEVER (INCLUDING WITHOUT LIMITATION, DAMAGES FOR LOSS OF BUSINESS PROFITS, BUSINESS INTERRUPTION OR LOSS OF BUSINESS INFORMATION), EVEN IF COMPAQ HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

The limited warranties for Compaq products are exclusively set forth in the documentation accompanying such products. Nothing herein should be construed as constituting a further or additional warranty.

This publication does not constitute an endorsement of the product or products that were tested. The configuration or configurations tested or described may or may not be the only available solution. This test is not a determination of product quality or correctness, nor does it ensure compliance with any federal, state or local requirements.

Performance of Pentium Pro and Pentium II Processor/Cache Combinations

First Edition (May 1997)
Document Number 436A/0597

INTRODUCTION

Processor speed, quoted in megahertz (MHz), is only one factor that determines system performance. Other factors such as memory speed, cache size, I/O bandwidth, internal CPU architecture, and the application environment also play roles. The hardware configuration and application environment will dictate which of these factors have the greatest impact on performance.

This white paper will describe the current Pentium Pro and Pentium II cache architectures and present a model of the relative effects of different speed and cache size options. The impact of these factors will be further explored by examining several system benchmark examples.

CACHE ARCHITECTURE

The Function of Cache Memory

Two basic components of performance are memory access time and instruction execution time. Simply put, memory access time is a measure of the time the processor must wait for data or instructions. One way to improve the memory access time is to use cache memory to store the most recently referenced memory locations. The access time for cache memory is usually faster than system memory, and the cache typically resides physically closer to the processor core. Increasing the cache size can improve the average memory access time by allowing more referenced addresses to be stored in the faster cache memory.

Cache size can be an especially critical factor in the performance of symmetric multiprocessing (SMP) systems. In current SMP servers and workstations each processor has its own cache memory, but all processors share system memory and the bus connections to system memory. As more processors are added, the processors will compete for access to the system memory bus. If two, or more, processors try to access system memory simultaneously, one must wait for the other, increasing its memory latency. At some point, the connections to system memory can become overloaded, creating a bottleneck. Increasing the size of the cache for the individual processors can reduce the number of accesses to system memory and thereby reduce traffic on the memory bus. For many SMP applications, enlarging the cache results in much better system performance than increasing the core processor speed. However, cache memory is typically much more expensive, byte for byte, than system memory and for most applications there is a point of diminishing returns for increases in the cache size. A cost-effective solution must weigh the additional cost for larger cache memory against any performance gain.

Both the Pentium Pro and Pentium II processors provide integrated cache memory as part of the processor module, but as the next two sections explain, the Pentium Pro and Pentium II caches differ in size, speed, and performance.

The Pentium Pro Cache Architecture

The Intel Pentium Pro architecture has two levels of cache memory, the primary or level 1 (L1) cache and the secondary or level 2 (L2) cache. The L1 cache is divided into an 8 KB instruction cache and an 8 KB dual-ported data cache. The L2 cache is connected to the processor through a 64-bit dedicated transaction-oriented bus that supports up to four concurrent cache accesses. Both caches and the processor core operate at the same speed. Increasing the processor speed, for example changing from a 166 MHz Pentium Pro to a 200 MHz Pentium Pro, will speed up both instruction execution time and cache memory access time (Figure 1).

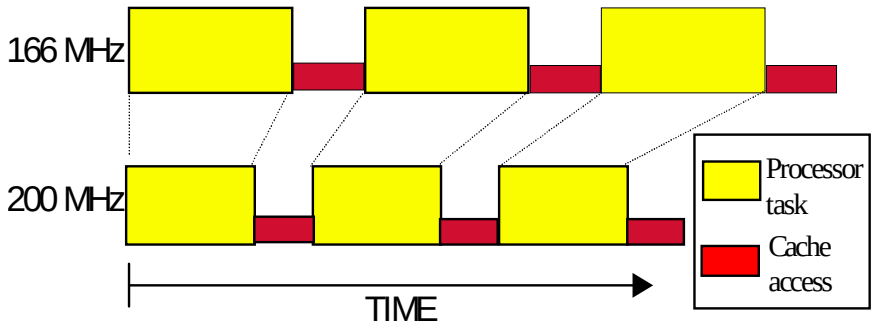

Figure 1. A comparison of instruction and cache timing for 166 MHz and 200 MHz Pentium Pro processors. As the speed of the Pentium Pro processor increases, the time needed to perform processor tasks and to access cache memory is reduced.

The L1 cache is integrated on the processor die, and the size is limited by the amount of space available on the die. The L2 is located on a different die within the Pentium Pro package, which allows more space for memory circuits. Since the Pentium Pro caches operate at the same speed as the processors, the static RAM (SRAM) of the cache must be fast enough to keep up with requests from the processor, and small enough for the cache lookup to be completed within a minimum number of clock cycles.

Pentium II Cache Architecture

The Pentium II employs a different cache architecture from the Pentium Pro. Compared with the Pentium Pro the Pentium II has, among other features:

- Faster CPU core speed
- Larger L1 cache
- Slower L2 cache bus
- 512 MB limit on cacheability

A new manufacturing process being used for the Pentium II die enables the L1 cache size to increase from 16 KB to 32 KB. A larger L1 cache allows more data to be captured in the L1 cache and improves the L1 cache miss rate. The L1 cache continues to run at the same frequency as the processor. However, the Pentium II L2 cache is now comprised of standard commodity SRAM components located outside the processor die on the Pentium II module card (Figure 2).

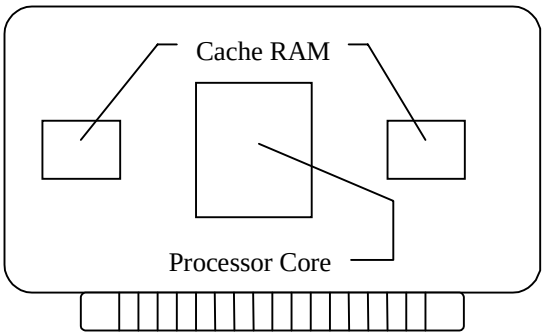

Figure 2. Representation of Pentium II module. The cache RAM components reside separately from the processor core on the Pentium II module card.

Due to the physical separation of the processor and cache, the high processor core frequency, the choice of packaging technology, and the commodity SRAM, the L2 cache bus can no longer run at the processor core frequency. The Pentium II L2 cache runs at 50% of the processor frequency. For example, a 266 MHz Pentium II has an L2 frequency of 133.3 MHz. By comparison, the 166 MHz Pentium Pro has an L2 frequency of 166 MHz; that is 25% faster than the 266 MHz Pentium II L2 cache. The increased size of the Pentium II L1 cache and the faster core frequencies will compensate somewhat for the slower secondary cache. Figure 3 shows a timing comparison of the Pentium Pro and Pentium II. In this example, a slower Pentium Pro executes the instructions and cache access in approximately the same time as the Pentium II, although the Pentium II has a higher core frequency.

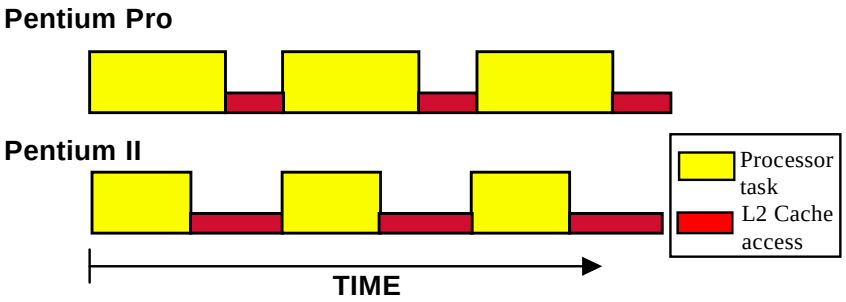

Figure 3. Instruction and cache timing comparison of Pentium Pro and Pentium II processors. The Pentium II processor can execute instructions faster than Pentium Pro processor, but the Pentium II processor's L2 cache access time is longer.

Unlike the Pentium Pro, Pentium II has a cacheability limit of 512 MB of data – meaning memory addresses greater than 512 MB can not be cached in either the L1 or the L2 cache. Because of Pentium II's slower cache speed and the cacheability limit, 266 MHz Pentium II processor-based servers may not perform as well as 200 MHz Pentium Pro-based servers in some memory-intensive or multiprocessing applications.

The Pentium II does have higher core frequencies than the Pentium Pro and a larger L1 cache, which may in some cases more than make up for the slower L2 cache. Many CPU-intensive applications will see better performance from the Pentium II processor family than from the Pentium Pro.

Table 1 shows the different processor/cache combinations: both currently available and future unannounced processors.

Table 1 Current and Future Pentium Pro and Pentium II processor/cache combinations *(Italics indicate unannounced products)*

Processor	Core/L1 Freq (MHz)	L1 Cache Size (KB)	L2 Cache Freq (MHz)	L2 Cache Size (KB)
Pentium Pro	166	16	166	512
	200	16	200	256
	200	16	200	512
	200	*16*	*200*	*1024*
Pentium II	233	32	116.5	512
	266	32	133.3	512

PERFORMANCE MODEL

Model overview

This section describes a model used to predict average memory access time for different processor/cache combinations. While average memory access time is not a direct measure of overall system performance, it is a key component of final system performance and shows the interactions between cache sizes, cache speeds, and processor speeds. The model calculates average memory access time based on the formula:

Average memory access time =

average L1 access time * probability the data is located in L1

+

average L2 access time * probability the data is located in L2

+

average system memory access time * probability the data is located in system memory

The probabilities are expressed in terms of cache miss rates. Cache miss rates are dependent on the operating system, the application software, and the specific data set. The model assumes the data is always located in either the caches or system memory and that all memory is cacheable. It does not include disk access times. More details of the model are given in Appendix A.

Results

Figure 4 shows the average memory access time for two different cases. A lower access time equates to faster (better) memory subsystem performance. In the first case, the miss rates were gathered from a trace of a dual processor Pentium Pro server and a simulation of a Pentium II system both running a widely-used OLTP benchmark under Microsoft SQL Server 6.5 and Microsoft Windows NT 4.0 operating system. The second case was obtained by doubling the miss rates from the OLTP example. A higher miss rate simulates an application environment that places a comparatively higher demand on the memory and cache subsystem. The miss rates for both cases are shown in Appendix A.

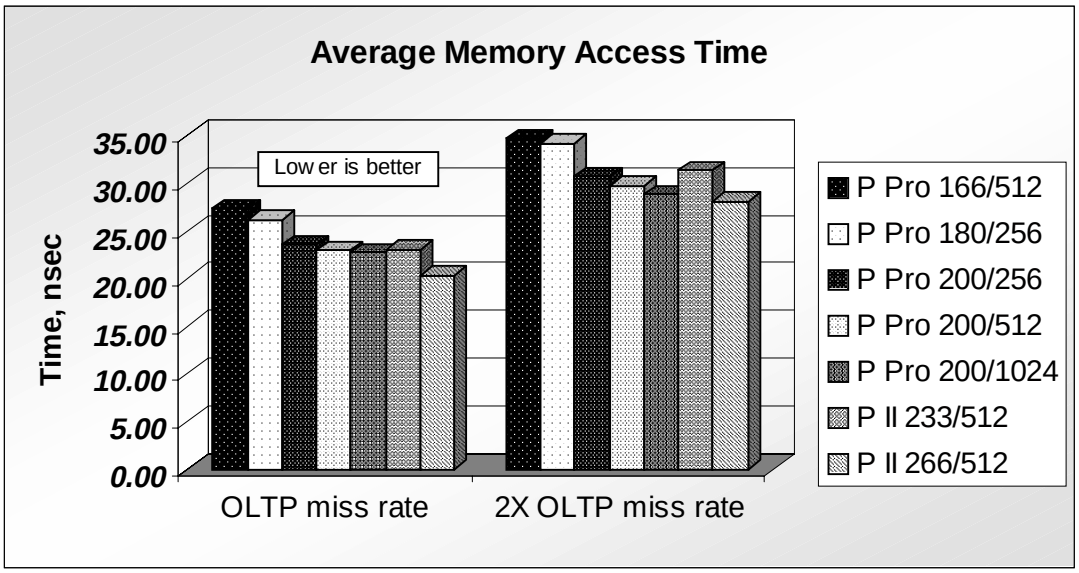

Figure 4. Model results showing average memory access times for various Pentium Pro and Pentium II processor cache combinations.

As shown in Figure 4, in the OLTP example, changing from a 166 MHz/512 KB cache processor to a 180 MHz/256 KB process resulted in a 4.3% performance increase. However, when the miss rate was doubled, the performance difference between the combinations narrowed to 1.6%. Keeping the cache size constant at 512 KB, a 200 MHz Pentium Pro was 18.7% faster than a 166 MHz Pentium Pro for the OLTP case.

The performance differences between the Pentium II and the Pentium Pro changed dramatically when the miss rates were increased. In the OLTP case, the 200 MHz/512 KB Pentium Pro and the 233 MHz/512 KB Pentium II performed virtually the same. However, when the miss rate doubled, the Pentium Pro had a 5.4% better access time than the Pentium II 233 MHz/512 KB did. Although the clock speed of the 266 MHz Pentium II is 33% faster than a 200 MHz Pentium Pro, it's memory access time was only 13% faster for the OLTP case, and only 5.8% for the higher miss rate case.

The performance relationship between the Pentium Pro and Pentium II is further illustrated in Figure 5. If the miss rate is lowered to 0 simulating a very CPU-intensive environment, the 233 MHz Pentium II has a 16.6% faster memory access time than a 200 MHz/512 KB Pentium Pro. As the miss rate increases, the memory performance of the two processors converge, and then switch. For this model, the crossover point occurred at the same miss rate as the OLTP example – a Pentium II L2 miss rate of 10.8%.

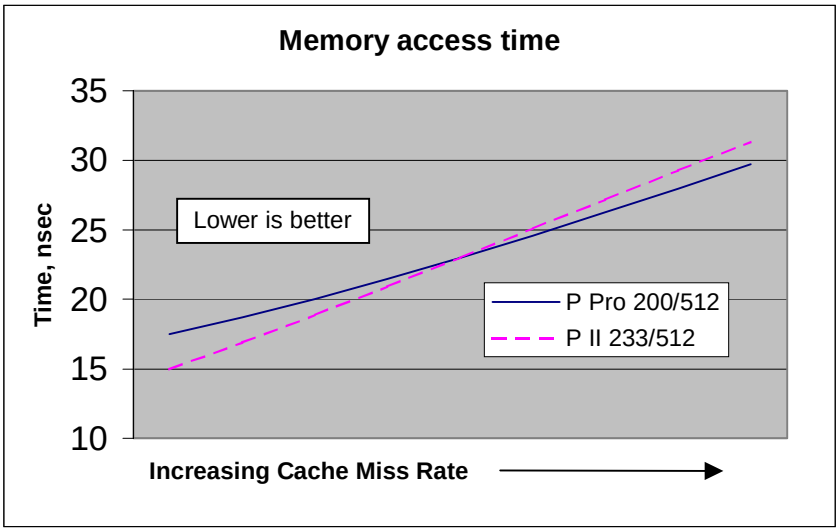

Figure 5. The access time of Pentium Pro and Pentium II processors varies as a function of cache miss rate. Cache miss rate varies by application and data set.

BENCHMARK PERFORMANCE EXAMPLES

This section contains specific examples illustrating the relative performance impact of processor speed and cache size.

OLTP Example

In general, the overall performance of database applications is limited by the CPU and memory performance. Performance data shows that these applications are typically more sensitive to cache size than processor speed. This is especially true for multiprocessor systems.

Figure 6 shows the relative performance of a Compaq ProLiant 2500 two processor (2P) system running SQL Server 6.5 (Server Pack 2.0) and a widely-used OLTP performance benchmark. The chart shows relative performance for configurations with dual 180 MHz Pentium Pros, dual 200 MHz/256 KB Pentium Pros, dual 200 MHz/512 KB Pentium Pros, and a mixed configuration of one 200 MHz/256 KB and one 200 MHz/512 KB Pentium Pro. In this example, the 180 MHz/256 KB processors ran the benchmark 11% slower than the 200 MHz/256 KB processors. Upgrading the L2 size from 256 KB to 512 KB for a 200 MHz Pentium Pro increased the benchmark performance by 14%.

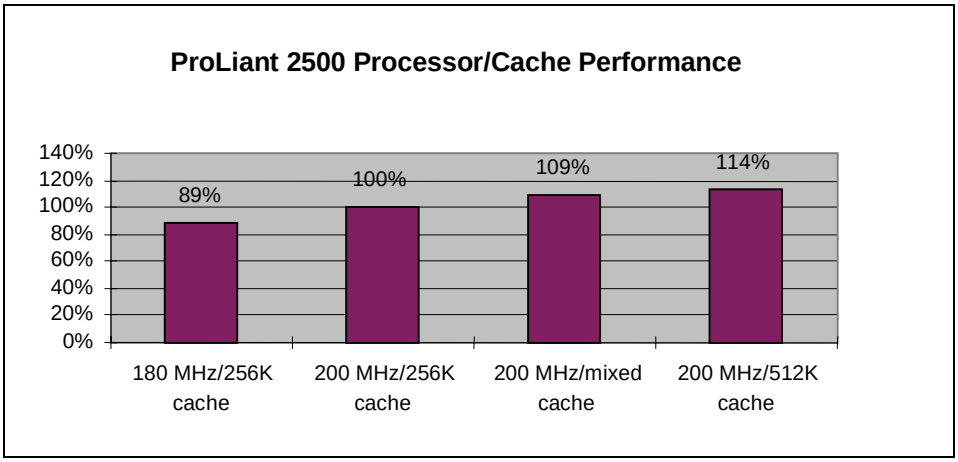

Figure 6. Compaq ProLiant 2500 Performance

Scalability - An Application Server Example

Like database servers, application servers tend to be far more sensitive to cache size than to other factors. Figure 7 shows the relative performance of single processor and four processor ProLiant 5000 configurations running the ServerBench benchmark under Microsoft Windows NT 3.51. The systems were configured with 256 MB of RAM, a SMART-2/P with 6x 2 GB wide drives, and a 1 GB narrow boot drive. Configuration details are shown in Appendix B. Figure 7 illustrates the impact of cache performance on multiprocessor systems. Notice that in the single processor case, the 200 MHz Pentium Pro processor with a 256K cache outperformed the 166 MHz Pentium Pro with a 512K cache. In the 4P configuration, however, the 166 MHz/512 KB processors outperformed the 200 MHz/256 KB model by an average of 16%.

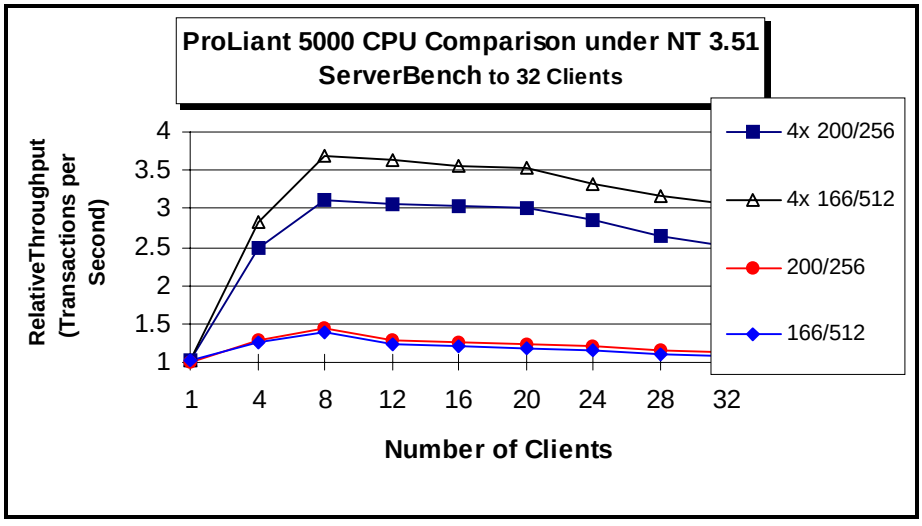

Figure 7. ServerBench Performance of ProLiant 5000 Servers

Scalability - An OLTP Example

Figure 8 compares the scalability of a ProLiant 5000 server with 166 MHz/512k Pentium Pro processors to the scalability of a ProLiant 5000 server with 200 MHz/512k Pentium Pro processors. The ProLiant was running a widely used OLTP benchmark under Microsoft Windows NT 4.0 and SQL Server 6.5 with Service Pack 3. In this example, both the 166 MHz and 200 MHz Pentium Pro systems show excellent scalability. Notice that adding an additional processor boosted the performance much more dramatically than increasing the processor speed.

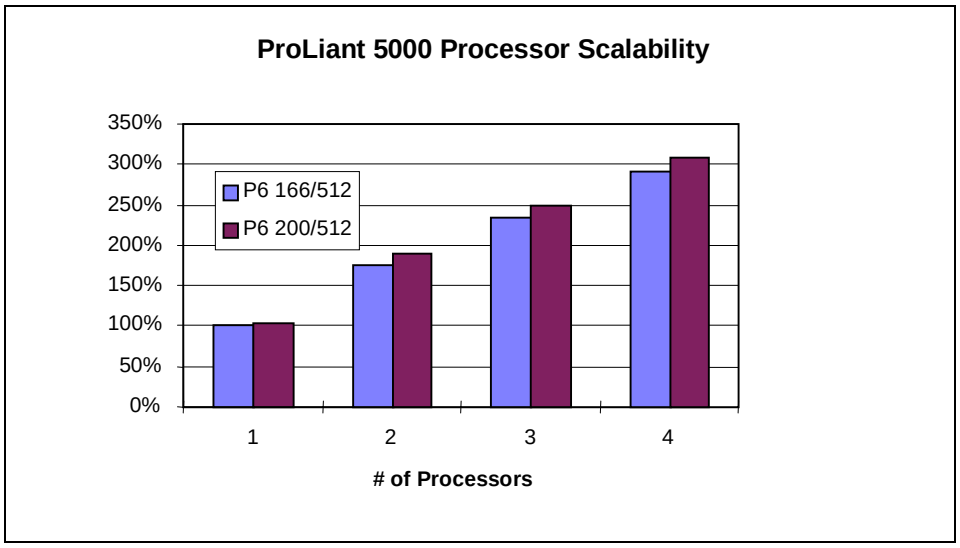

Figure 8. ProLiant 5000 Scalability

Workstation Performance

Unlike large database applications, many workstation applications show a strong performance increase from Pentium Pro to Pentium II processors, as these applications tend to be more CPU-intensive. Figure 9 shows a comparison of SPECint_rate95 benchmark results of a Pentium Pro workstation (Tulane) compared with a Pentium II-based workstation (Poydras). SPECint_rate95 is component-level benchmark that is designed to measure the overall performance of the workstation's processor, memory, internal architecture, and compiler. In the chart, the performance numbers are normalized to a 200 MHz/512 KB Pentium Pro. For the SPECint_rate95 benchmark, the single and dual processor configurations with Pentium II processors outperformed the similar Pentium Pro systems. However, 3P and 4P Pentium Pro processor configurations outperformed all other combinations. As stated previously, Pentium II processors are limited to single and dual configurations due to the processor internal architecture. Note: These results have not been submitted to SPEC. They have not appeared in SPEC literature.

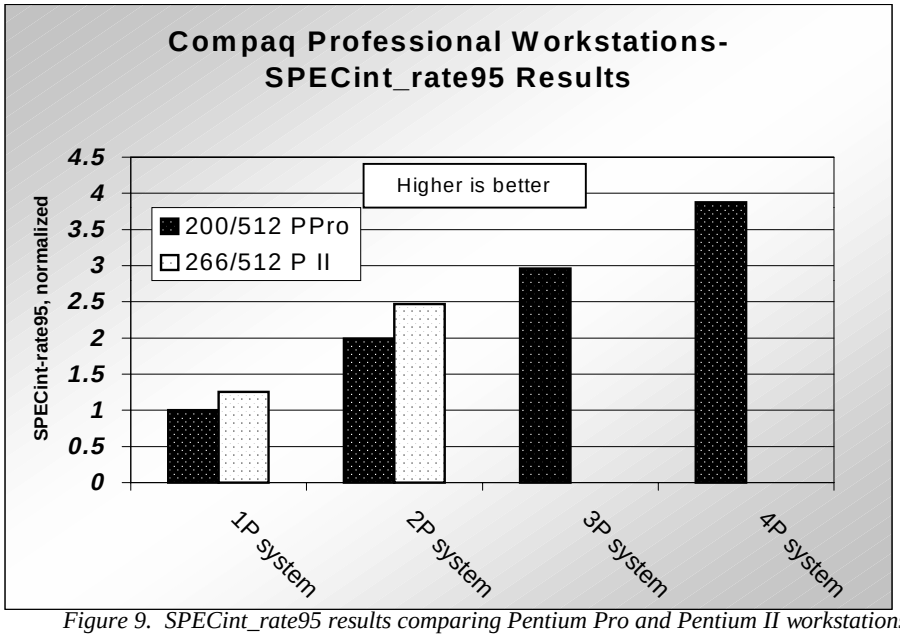

Figure 9. SPECint_rate95 results comparing Pentium Pro and Pentium II workstations.

Other benchmarks also show Pentium II outperforming similar Pentium Pro systems. A single processor 266 MHz Pentium II workstation outperformed a similarly configured 200 Mhz/512 KB Pentium Pro by 19% in the Viewperf CDRS benchmark. Viewperf is a benchmarking application for measuring 3D rendering performance of systems using OpenGL. Similarly, a Pentium II 266 MHz workstation beat out a similarly configured 200 MHz/512 KB Pentium Pro by 29% in the CADalyst '96 benchmark, a measure of AutoCAD performance. Note: These are unreleased performance numbers using pre-release drivers and are subject to change.

FUTURE TRENDS

Pentium Pro 1024 KB L2 Processor

Intel is expected to announce a 200 MHz Pentium Pro processor with a 1024 KB L2 cache in late second quarter. Initial TPC-C, and AIM numbers show significant performance increases compared with the 512 KB L2 cache. Performance numbers for this processor configuration are still changing as the product comes closer to announcement and more studies are performed. The 1024 KB Pentium Pro should be particularly well suited for multiprocessor configurations and for memory-intensive applications such as large data base environments.

Pentium II

The Pentium II processor is a viable processor for many workstation and single and dual processor server environments due to its larger L1 cache and faster core. However, due to its architectural limit to 2 processors and 512 MB cacheability limit, it is not a suitable processor for larger SMP environments.

Conclusion

Processor speed alone does not determine the performance of systems. Cache sizes and cache speeds are also important to the performance of servers and workstations, particularly in systems with more than one processor. The performance relationships between the processor/cache combinations change depending on the specific application used and the hardware configuration.

Memory-intensive applications such as OLTP and application servers will be more sensitive to changes in cache size, while CPU-intensive applications will get a relatively higher performance boost by increasing processor speed. SMP systems will almost always show a performance sensitivity to cache size because of the competition for system memory bus access. Figure 10 summarizes the different application environments most suitable for Pentium Pro and Pentium II processor systems. Customers should weigh all factors in determining which server or workstation best meets their particular requirements.

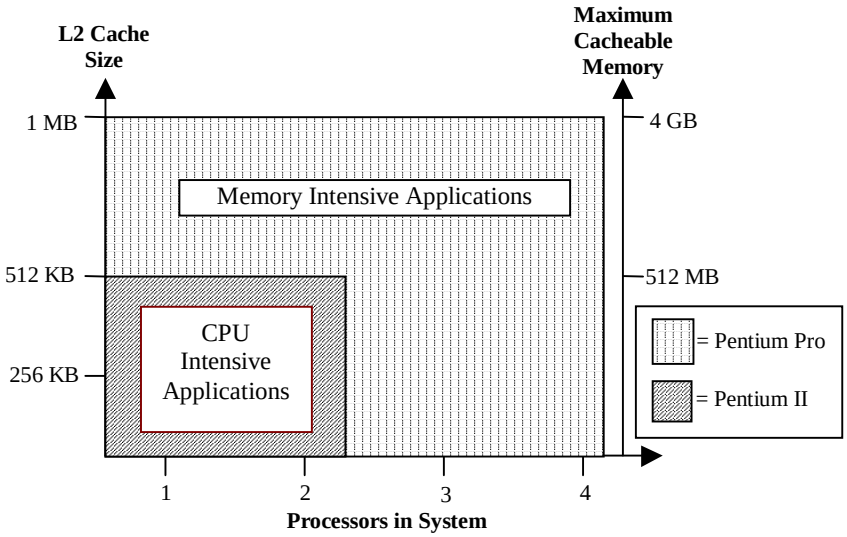

Figure 10. Recommended application environment for Pentium Pro and Pentium II processors

Appendix A

MEMORY ACCESS TIME MODEL

The model predicts the average memory access time based on the following equations:

Average memory access time = (1-L1_miss_ratio)*(3.5*L1_clock_time) + {L1_miss_ratio*{[(1-L2_miss_ratio)*(15*L2_clock_time)]+[L2_miss_ratio*(memory_access_time*memory_clock_time)]}}

Average Execution time = [Average_Instrution_Execution_time + (Average_memory_Access_time*memory_accesses/instruction)]

Memory access time = 15 clock cycles

The following table shows the miss ratios derived from a system simulation of a dual processor server running an OLTP benchmark under Microsoft SQL Server 6.5 and Microsoft Windows NT 4.0.

Table A-1. Data used in Memory Access Model

Processor	L1/Core Freq (MHz)	L2 Cache Size (KB)	L1 Miss Ratio %	L2 Miss Ratio %	Memory access time - OLTP Miss rate (nsec)	Memory access time- 2X OLTP Miss rate (nsec)
Pentium Pro	166	512	8%	8.5%	27.30	34.66
	180	256	8%	15.7%	26.16	34.1
	200	256	8%	15.7%	23.54	30.68
	200	512	8%	8.5%	22.99	29.74
	200	1024	8%	5.3%	22.68	28.84
Pentium II	233	512	6.6%	10.8%	23.01	31.36
	266	512	6.6%	10.8%	20.34	28.09

APPENDIX B: SERVERBENCH 3.0 TEST CONFIGURATION DISCLOSURE

Server Disclosure

Machine name	ProLiant 5000
Size of hardware CPU cache	256 KB / processor or 512 KB /processor
Amount of memory	256 MB
Type of I/O bus	2 PCI / 1 EISA
Number and type of hard disk controllers	1 SMART-2 PCI Array Controller and Compaq Integrated 32-bit SCSI Controller
Number and type of hard disks	6 x 2GB Seagate Wide SCSI-II 1 x 1GB HP Narrow SCSI-II
Disk organization (striped, mirrored, RAID, etc.)	RAID 0
Disk controller driver version	cpqarray.sys 6/27/96
Number and type of network controllers	1 x Netelligent 10/100
Network controller driver version	netflx3.sys 6/27/96
Network operating system name and version	NT 3.51
Any relevant modifications to default network operating system parameters	Drives formatted NTFS

TestBed Disclosure

Network type (10Base T, Token Ring, etc.)	100 Base-TX
Number and type of clients	60 Compaq ProLineas
Number and type of hubs/concentrators (full duplex, switching, etc.)	5 Bay Network 28115 Switched hubs
Number of clients/segment	15
Client CPU type and speed in percentages	90MHz Pentiums - 72%, 100MHz Pentium - 5%, 100MHz i486 - 5%, 50MHz i486 - 3%
Client network controller broken down by percentages	Intel Pro/100 - 100%
Client network software name and version (drivers, protocols, redirector)	Win95, TCP/IP
Size of any client network cache	none
Disk controller software	Win95 driver for IDE drives
Network controller software	Intel Pro/100 Driver

Controller Disclosure

Controller Operating System	Win95
Network type (10Base T, Token Ring, etc.)	100 Base-TX
Disk controller software	Win95 driver for IDE drives
Network controller software	Intel Pro/100 Driver

ServerBench Disclosure

ServerBench version	3.0
Description of the test parameters for each mix in the test suite	SYS_60.TST