TECHNOLOGY BRIEF

November 1997

Compaq Computer
Corporation

CONTENTS

Compaq 8-Way Multiprocessing Architecture

The next logical step-function in the performance curve for high-end standards-based servers is an 8-way architecture with the next-generation Intel 32-bit (IA-32) processor, code-named "Deschutes." Compaq carefully evaluated the options for an 8-way symmetric multiprocessing (SMP) server and has joined with Corollary, Inc., to develop an optimum solution for the next-generation standards-based server. The combination of Compaq's I/O design with Corollary's Profusion 8-way crossbar switch technology and the Deschutes/Slot-2 processor by Intel will yield a balanced, high-performance system architecture for the future. Intel's recent announcement of the acquisition of Corollary further ensures that the architecture will become the industry standard for 8-way servers. This brief explains Compaq's architecture for an 8-way multiprocessing server and compares this architecture to existing 8-way SMP architectures.

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.

Compaq and ProLiant are registered with the United States Patent and Trademark Office.

Pentium is a registered trademark of Intel Corporation.

Profusion is a trademark of Corollary, Inc.

I_2O and I_2O SIG are registered trademarks of the I_2O Special Interest Group.

Other product names mentioned herein may be trademarks and/or registered trademarks of their respective companies.

Portions of this document are copyright of Corollary, Inc.

©1997 Compaq Computer Corporation. All rights reserved. Printed in the U.S.A.

Compaq 8-Way Multiprocessing Architecture
Second Edition (November 1997)
Document Number ECG051.1197

INTRODUCTION

Not long ago, 4-way symmetric multiprocessing (SMP) servers emerged as the leading edge in standards-based servers. Today, the highest end of the technology curve is shifting to a higher number of processors. Why go beyond four processors? First, Compaq has done the engineering work for 4-way servers, and it is now "standard." Second, performance demands of certain applications and environments go beyond what four processors can supply. Third, emerging clustering technologies are demanding greater performance from and more processors for each node. SMP architectures are complimentary to clustering, since a cluster of machines will only be as good as its component nodes. Compaq's plan is to have both.

For SMP servers, the next logical step-function in the performance curve will be 8-way servers using the next-generation Intel 32-bit (IA-32) processor, code-named "Deschutes." The Deschutes processor is the right choice for 8-way implementation for two reasons. First, because of its improved architecture and higher clock speeds, it will greatly outperform the Intel Pentium Pro processor in 8-way implementations. Second, Pentium Pro is nearing the final version of its generation. On the other hand, Deschutes is the first member of the next generation in 32-bit microprocessors from Intel. Compaq will not ask its customers to invest in technology that is short-lived. As an example of Compaq's commitment to investment protection, the groundwork has already been laid for transition to an 8-way server. The Compaq ProLiant 7000 Server, announced in August 1997, is designed to be upgradable to eight processors. For more information, see *8-Way Technology and the Compaq ProLiant 7000*, document number ECG050.1097.

This brief explains Compaq's architecture for an 8-way SMP server and compares this architecture to existing 8-way SMP architectures. This brief assumes the reader understands current SMP designs. For additional information on Compaq technology and architectures, see the list of related documents in the appendix.

8-WAY ARCHITECTURE OVERVIEW

In the history of standards-based SMP machines, two companies have demonstrated a consistent vision and commitment to advancing SMP technology: Compaq and Corollary. Compaq and Corollary recently announced[1]a technology exchange agreement to develop a standards-based 8-way multiprocessing architecture. The architecture is implemented as a chipset that Compaq will use in future system designs. The jointly developed chipset architecture will also be made available to other system providers.

Compaq and Corollary have been working together on the Profusion 8-way chipset technology for more than a year. Corollary has designed the Profusion 8-way crossbar switch technology for the architecture. Compaq has applied its experience with industry-leading I/O subsystems and PCI Hot Plug technology to complete the overall design for a balanced 8-way architecture.

[1] Compaq/Corollary joint press release, *"Corollary and Compaq Announce Technology Exchange Agreement for Breakthrough Eight-Processor Design,"* Oct. 9, 1997

Figure 1 shows a block diagram of the Profusion architecture. The essential features of the architecture are:

- Dual 100-MHz processor buses
- Dedicated 100-MHz I/O bus
- 8–way multiprocessing with Deschutes processors
- Multi-ported system architecture (crossbar switch)
- Dual-ported, interleaved memory
- Uniform memory access for all eight processors
- Dual cache coherency filters
- Up to four Compaq-designed host-to-PCI bridges

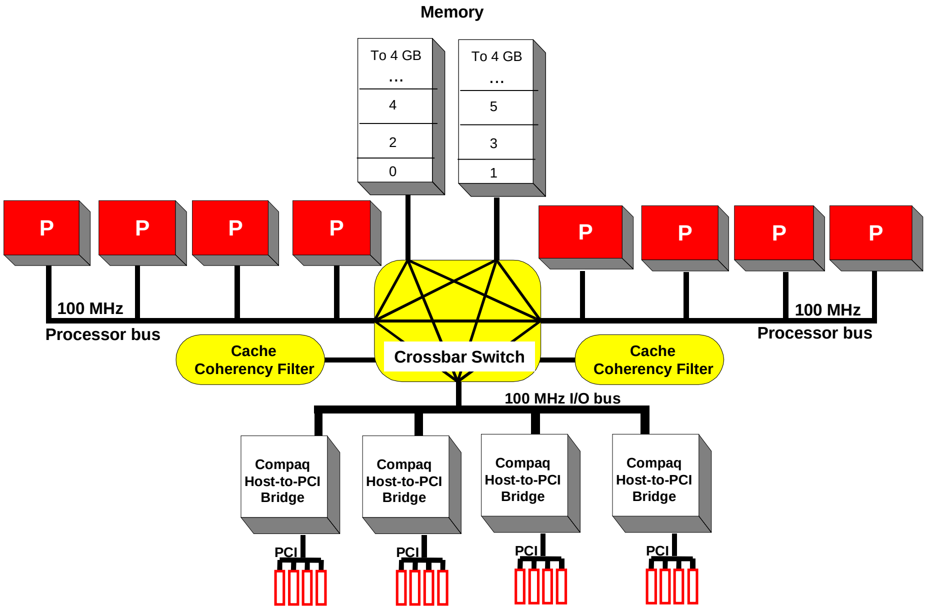

Figure 1: Profusion system architecture

Processor and I/O Bus Design

The two processor buses and the dedicated I/O bus are GTL+, 64-bit buses. GTL+ is the Gunning Transceiver Logic bus used by the Pentium II and Pentium Pro processors. The use of GTL+ allows the buses to operate at higher clock speeds without severely reducing the bus length or number of electrical loads. All three buses run at 100 MHz, whereas the existing Pentium Pro bus runs at only 66 MHz.

The Profusion architecture joins the two processor buses and two memory ports together through a crossbar switch. The otherwise independent processor buses are joined by a logical connection that is made only when required to transfer data. The two independent buses with two independent memory ports allow better scaling than a single host bus with eight processors attached, regardless of the number of memory ports. Using two buses rather than a single common bus reduces the traffic load on each bus, resulting in reduced processor wait states.

In 4-way SMP servers today, the PCI buses are directly attached to the processor bus via a host-to-PCI bridge. However, with the Profusion 8-way architecture, the host-to-PCI bridges are attached to a dedicated high-speed I/O bus. Dedicating a separate bus for I/O reduces the traffic on the processor buses, removing another potential limit to performance.

Because each of the three buses has multiple devices attached with cache data, coherency traffic among the buses could limit the performance and scaling. The Profusion architecture uses coherency filters to eliminate unnecessary traffic on the buses. These filters are discussed in more detail in the "Coherency Filters" section, page 8.

Processor Technology

The Deschutes processor, scheduled for release by Intel in 1998, is based on the Pentium II processor core and cartridge form factor. The version of Deschutes that will be optimized for high-end servers is known by its form factor designation, Slot-2. With an 8-way architecture, there are three compelling reasons to use the Deschutes/Slot-2 processor rather than the Pentium Pro or the Pentium II:

- Faster 100-MHz bus support
- Higher internal (core) frequency
- Larger Level-2 cache size on a full-speed cache bus

100 MHz bus

The Profusion architecture takes full advantage of the 100-MHz capabilities of both processor buses. Because of the increased frequency of the bus, the number of electrical loads must be limited so signals do not degrade and cause a drop in bus performance. The Profusion architecture limits the number of electrical loads on each bus to five. Therefore, it is an optimum solution for the Deschutes processor that will operate at frequencies up to 100 MHz.

Core Frequency

The Pentium Pro processor operates at a maximum core frequency of 200 MHz on a 66-MHz bus. The Deschutes processor, however, will operate at a core frequency above 333 MHz on a 100-MHz bus.

Level-2 Cache

Deschutes will have a Level-2 (L2) cache size greater than 512 kilobytes. As with Pentium Pro, the cache will operate on a full speed bus. While the Pentium II processor has a faster core frequency than Pentium Pro, it only has a half-speed cache bus. Deschutes' large cache size combined with the full speed bus will allow more of the recently requested data to be retrieved more quickly. The processors will use the bus more efficiently, will have fewer wait states, and will greatly enhance the performance of an 8-way SMP server.

Table 1 shows a complete comparison of Deschutes to the Pentium II and Pentium Pro processors.

Table 1: Comparison of Pentium Pro, Pentium II, and Deschutes processors

Feature	Pentium Pro	Pentium II	Deschutes
Core Speed	166, 180, 200 MHz	233, 266, 300 MHz	>333 MHz
System Bus Speed	66 MHz	66 MHz	100 MHz
L2 Cache Bus Speed	Full Speed	Half Speed	Full Speed
L2 Cache Size	256K, 512K, 1M	512K	>512K
Cacheable Address Space	4 GB	512 MB	>512 MB
MMX™ Technology	No	Yes	Yes

Multi-Port Crossbar Switch

An important element of the Profusion architecture chipset is Corollary's system controller design, which is based on a crossbar switch. The five port switch is schematically represented in Figure 2.

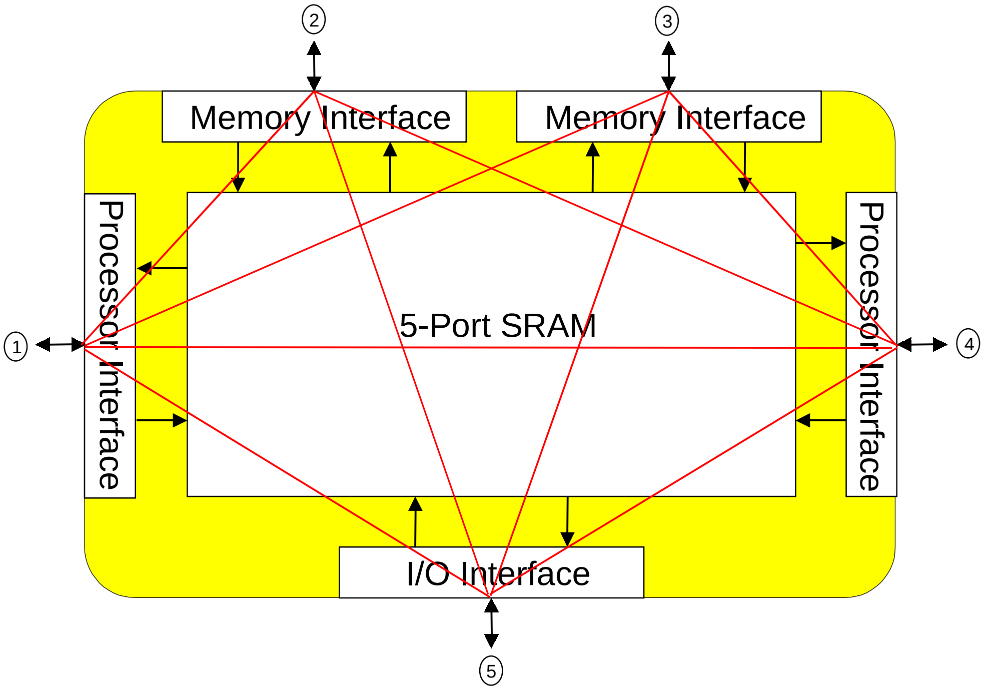

Figure 2: Multi-port crossbar switch

The switch is non-blocking and multi-ported. It allows simultaneous read and write paths so that each bus can perform reads while writes complete on another bus. There are nine possible information paths between the processors, memory arrays, and I/O. These direct paths reduce the amount of time it takes to perform a read, since the information has to travel through only one device. This is a significant advantage over other architectures that use multiple chips (for example, two memory controllers on two different buses) to transfer information between processor buses or between memory and I/O.

Figure 3 shows the Application Specific Integrated Circuit (ASIC) partitioning in the Profusion crossbar switch. After a cache miss, a data request goes through the crossbar switch to main memory. While the crossbar switch consists of two physical chips, the functions are partitioned so that for any single command in that data request, information must traverse only one of the chips: either the memory address controller (MAC) or the data interface buffer (DIB).

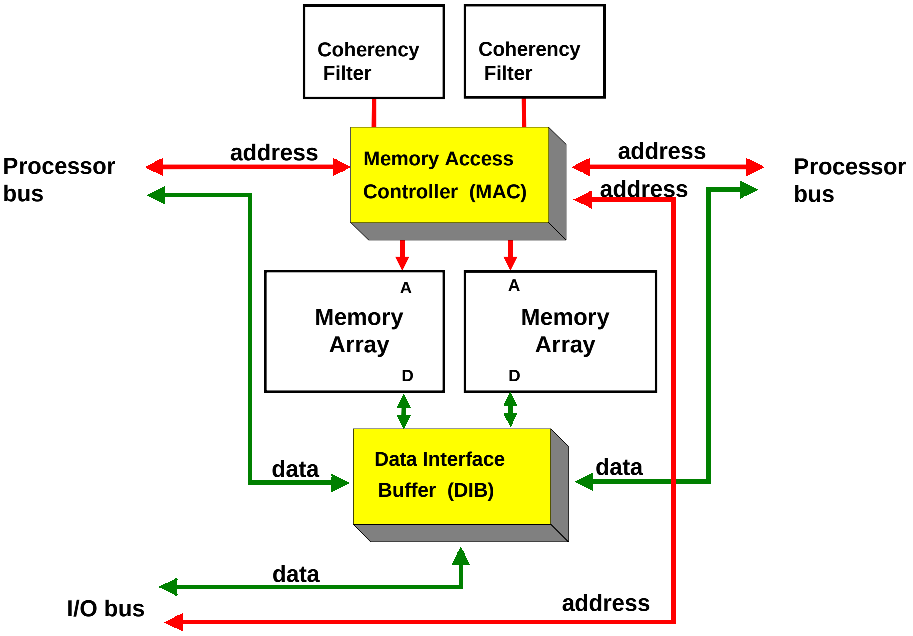

Figure 3: ASIC partitioning in Profusion architecture

If an address or control signal travels through the crossbar switch, it goes through the MAC. The MAC manages the external coherency filters and tracks the information stored in the data interface buffer. The MAC supports up to 8 Gigabytes (GB) of main memory.

If data travels through the crossbar switch, it goes through the DIB. The DIB allows simultaneous data transfer on all five ports, has 64 cache line buffers, and uses error correcting code to ensure that data is correct. The cache buffers can be used for any function. The efficiency of the buffers is high because there are no dedicated queues between buses.

Finally, with the 100-MHz bus, there is a peak throughput on each port of 800 Megabytes per second (MB/s). This is a 33 percent improvement over the 533 MB/s throughput on a 66-MHz bus. With all five ports, this gives a total peak throughput of 4.0 Gigabytes/s.

Dual Interleaved Memory and Uniform Memory Access

As noted in Figures 1, 2, and 3, the crossbar switch has two ports to main memory. Having two memory ports increases memory bandwidth, reduces access conflicts, and increases the maximum memory supported. The two memory arrays are cache line interleaved; they share a common address range. One memory port responds to even-numbered cache lines, and the other port responds to odd-numbered cache lines. This configuration has the highest performance because it allows the two memory arrays to be used in a balanced fashion. It is especially advantageous for common applications that access memory in a random manner. In random accesses, roughly half the requests at any one time are even-numbered lines, while the other half are odd-numbered lines.

Because the memory is dual-ported, processors from either bus have equal access to memory. This uniform memory access reduces latency. In non-uniform memory access (NUMA) architectures, a processor has quick access to one memory array, but a lag time, or latency, to a second memory controller on a separate processor bus. This type of architecture is described in more detail in the NCR architecture section, page 11.

Coherency Filters

One of the design challenges of SMP systems is to maintain a consistent view of memory by all the processors and the I/O subsystems. This is typically referred to as maintaining cache coherency. Because data is shared between multiple caches on the processor buses and main memory, there is the possibility that copies of the data may be inconsistent. Intel architecture processors support a snooping protocol to solve the cache-coherency problem. In a typical two-bus design, a memory transaction on one processor bus would have to "snoop" the other bus to make sure that the most recent data is used. Every snoop cycle adds traffic to the bus and potentially limits the performance of the system.

Memory Filters

The Profusion architecture uses two coherency filters to minimize the number of snoop cycles that occur between the processor buses. Each coherency filter holds the addresses of data stored in the four L2 processor caches on each bus. Each filter also holds information about the state of the data, for example, whether the data is shared between two caches and whether the data is current. By holding the state of the data in addition to the address, the coherency filter can direct snoop traffic to the other bus only when it is required to maintain coherency. This filtering results in overall lower bus utilization on the other bus.

I/O Filter

The Profusion architecture also has a coherency filter for the I/O bus. This coherency filter is specifically designed to work with up to four Compaq host-to-PCI bridges. When a PCI bus master requests data, the I/O coherency filter checks the buffers in the host-to-PCI bridge before going to main memory. This reduces snoop traffic on the I/O bus whenever a PCI bus master requests data. The cooperative design of the memory and I/O subsystems is a clear example of the performance advantages to be gained through the technology exchange between Corollary and Compaq.

I/O Technology

Compaq has designed a host-to-PCI bridge (host bridge) to address the needs of next-generation servers. It is specifically engineered to enhance the Profusion performance.

Compaq's host bridge includes the following important features:

- Delayed transactions support as defined in the PCI 2.1 specification[2]
- Compatibility with 64-bit/66-MHz PCI bus
- Asynchronous design to accommodate multiple bus frequencies
- Multiple prefetch buffers
- PCI Hot Plug controller integrated into the bridge
- Peer-to-peer operations supported on a single PCI bus segment and across the I/O bus to other PCI segments

[2] *PCI Local Bus Specification,* Revision 2.1, June 1, 1995, PCI Special Interest Group

Delayed Transactions

One of the most important features of Compaq's host bridge is its support for PCI delayed transactions. A delayed transaction is a modified version of a split transaction. Splitting one transaction into two separate transactions frees the bus and the processor for other activities while waiting on data requests to be completed.

Compaq's design performs two essential tasks. First, it supports delayed transactions, as described in the PCI 2.1 specification, which improves bus performance. However, in a PCI delayed transaction, the processor must poll the host bridge to determine when its data is there, rather than waiting for an interrupt, as with a split transaction. Therefore, Compaq incorporated additional features to reduce the amount of processor polling required by the PCI delayed transaction. Reducing the processor polling further maximizes bus efficiency.

In contrast, other host-to-PCI bridge devices do not support delayed transactions at all and can often slow transactions when both a read and a write request occur simultaneously.

A note on PCI 2.1 compliance and delayed transaction support: the PCI 2.1 specification includes delayed transaction support. The specification was also written to be backward compatible with the PCI 2.0 specification. Therefore, the user must be cautious when the term "PCI 2.1 compliant" is used. The term may be used to denote support for delayed transactions. On the other hand, since 2.1 is backward compatible with 2.0, the term may mean nothing more than the device is PCI 2.0 compliant. The Compaq host bridge truly supports PCI 2.1, including delayed transaction support.

PCI Bus

The Compaq host bridge is fully compatible with the 64-bit/66-MHz PCI bus and therefore provides four times the maximum bandwidth of today's standard 32-bit/33-MHz PCI bus.

Asynchronous Design

The host bridge is split into two sections: upstream (host I/O side) and downstream (PCI side). All the functions on the upstream side are in the host processor clock domain at 100 MHz. All functions on the downstream side are in the PCI clock domain at 66 MHz. This asynchronous design allows the bridge to accommodate any speed required on any of the buses. This will provide maximum I/O performance as clock rates increase with new processor releases.

Multiple Prefetch Buffers

The Compaq host bridge is designed with multiple prefetch buffers to ensure optimum I/O to processor performance, and each can hold multiple cache lines. These buffers have been sized to provide optimal performance at a reasonable and cost-effective die size. Because of the delayed transaction support, multiple PCI devices can request data and the bridge can get the data concurrently. This is unlike other host-to-PCI bridges that can only hold a single cache line and can only handle a single request at a time.

PCI Hot Plug

The host bridge supports PCI Hot Plug technology, pioneered by Compaq. PCI Hot Plug technology allows a PCI adapter to be removed and replaced while the server is up and running. Slots are powered down individually, impacting only the desired slot. Other adapters continue functioning while service is occurring on that slot. The host bridge includes the electronics to control the PCI bus during these functions.

Peer-to-Peer Transactions

The host bridge also supports PCI peer-to-peer transactions as defined in the $I_2O^{®}$ specification. The host bridge allows communications between two devices on the same PCI bus segment. It also allows peer-to-peer communication across the I/O bus to PCI devices on other PCI bus segments (Figure 4).

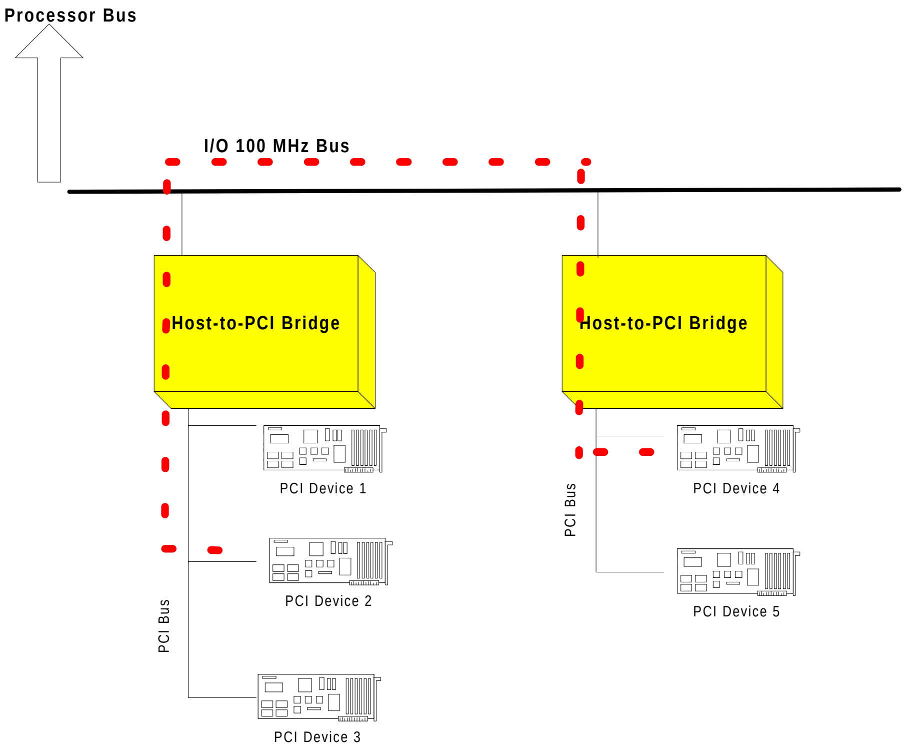

Figure 4: Example of peer-to-peer transaction

Microprocessors embedded in the PCI devices will make the devices “intelligent” and will perform the communication tasks previously done by the main processors. Data requests that would normally go to the processor bus are handled by the coherency filters within the crossbar switch. Offloading work from the main processors is expected to bring significant performance improvements to the system overall.

OTHER 8-WAY ARCHITECTURES

Unlike Compaq, other companies have chosen to implement 8-way technology using the Pentium Pro processor and existing chipsets. Several weaknesses are common to these other architectures:

- **Intel 82450GX chipset**. This chipset is designed for a 66-MHz bus and is not compatible with the future 100-MHz bus.
- **Delayed transactions.** No support for delayed transactions exists with the Intel 82450GX host-to-PCI bridge. While this is acceptable in today’s 4-way SMP servers, it is likely to cause an I/O bottleneck with 8-way servers.

- **Pentium Pro to Deschutes transition.** The transition to a 100-MHz bus from the 66-MHz bus requires significant engineering. In particular, a redesign may include reducing the number of loads on the bus.

The following sections describe the 8-way SMP architectures designed by NCR and Axil. Both use the Pentium Pro processor as the basis for their architectures.

NCR

NCR's 8-way system using the Pentium Pro processor is known as OctaScale. The architecture uses a NUMA (non-uniform memory access) architecture. The design links, or chains, two processor buses together through an interconnect (Figure 5). Thus, it is also known as a chained architecture.

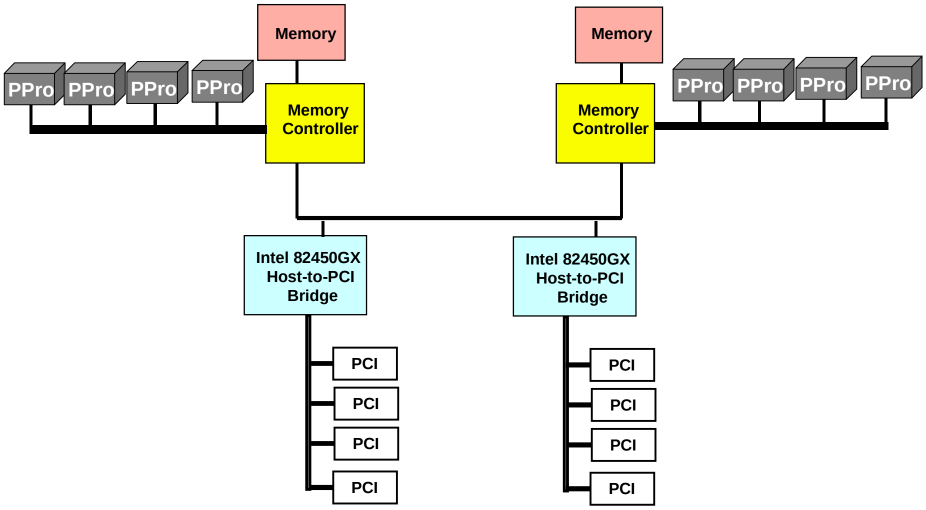

Figure 5: NCR chained architecture

NCR is using its own memory controller design in OctaScale. The controller has reportedly[3] been the cause of shipment delays in the product. Because there are two memory controllers on two different buses, a transaction has to pass through both controllers before transferring data between memory on one bus and I/O on another bus. The Profusion architecture has a distinct advantage over this architecture because the multi-ported crossbar switch allows all transactions to go through a single chip.

The NCR design uses the Intel 82450GX for its I/O controller. It is likely that the I/O controller will cause an I/O bottleneck with 8-way SMP architectures. The Intel 82450GX is also limited for future 8-way applications because it has no delayed transaction support and no support for a 100-MHz bus.

Performance estimates made by Corollary[4] shows that a chained bus architecture such as NCR's is effective in a four processor system. However, when eight processors are used, performance does not improve proportionally (Table 2).

[3] "NCR Rolls Out 8-Way Windows NT Server," John McCright, *PC Week*, Sept 5, 1997.
[4] *Profusion Technical Overview*, Corollary, Inc., Document 57-00167-07, 1996.

Table 2: Comparison of chained architecture and Profusion architecture

Architecture	Case	Memory Accesses with 4 processors	Clock Cycles	Memory accesses with 8 processors	Clock cycles
Chained	Low Memory	100%	10	40%	10
	High Memory	N/A	N/A	40%	25
	Other-Side Intervene	N/A	N/A	20%	50
	Average Access Time		**10**		**24**
Profusion	Memory Reference	100%	10	80%	10
	Other-Side Intervene	N/A	N/A	20%	25
	Average Access Time		**10**		**13**

The data in Table 2 compares a chained architecture such as NCR's to the existing Profusion architecture, modeled with 66 MHz Pentium Pro processors for appropriate comparison. The model only evaluates memory transactions. Since the architecture uses non-uniform memory access, a memory access can take significantly longer when a processor requires data from the other side of the system, as noted by the Other-Side Intervene data. With the inclusion of the Compaq host bridge in the Profusion architecture, the performance difference will be even more significant.

Axil

Axil Computer uses its Adaptive Memory Crossbar technology for its 8-way Pentium Pro system (Figure 6).

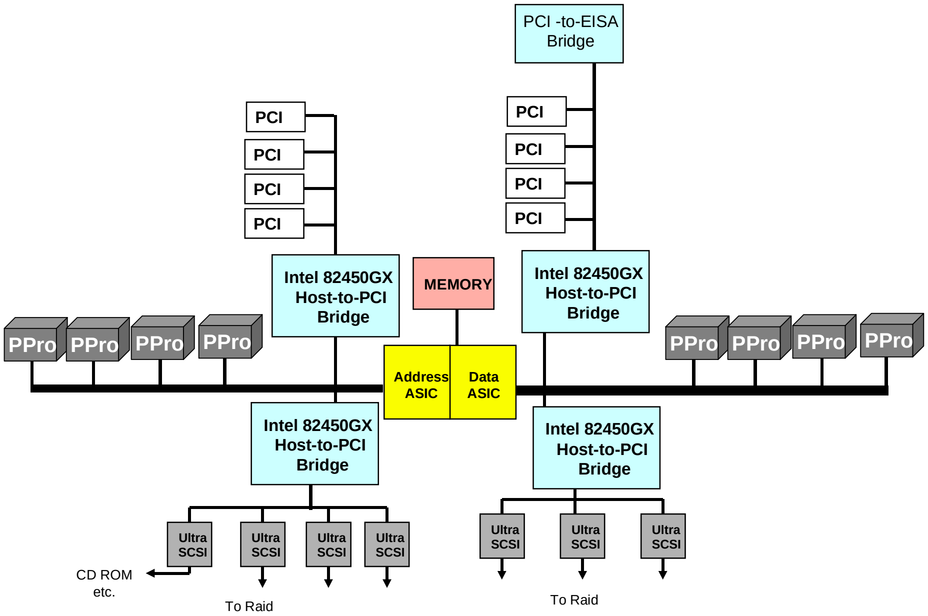

Figure 6: Axil dual-ported architecture

This design is a dual-ported memory design, similar to the Profusion multi-ported architecture. However, there are some important differences.

The first difference is the cache coherency mechanism in the memory subsystem. Axil has a mechanism that selectively turns off the coherency checking and forwards all transactions for a particular processor to the opposite bus. Axil claims a 2 to 5 percent overall performance increase due to this feature.[5] Since even performing a snoop operation causes a bandwidth and latency penalty, how can the performance increase by forwarding all transactions across the bus? The conclusion is that during normal coherency checking, the Axil design significantly slows performance by “stretching” the coherency-checking phase and adding clock cycles. The Profusion architecture, on the other hand, does not stretch the coherency-checking phase, does not cause a backlog in the bus transactions, and does not add unnecessary traffic to the other processor bus.

Second, there are differences in the I/O subsystem. Axil attaches an I/O controller onto each of the two processor buses, while the Profusion architecture has a third bus dedicated to I/O. Axil estimates that I/O could be using as much as 20 percent of the bandwidth on the processor bus.[6] Thus, each processor bus in Axil’s design has to deal with the bandwidth load of four processors, associated snoop traffic, and associated I/O. In contrast, the Profusion design has a dedicated I/O bus with a direct path into memory. The design gives a higher effective I/O bandwidth. As with other SMP systems, Axil uses the Intel 82450GX chipset, with its limitations for aggressive 8-way SMP implementations.

Finally, the architecture differs in the processor bus design. The Axil design has seven loads on each processor bus (four processors, two host-to-PCI bridges, and the memory controller). While seven loads are feasible on the current 66-MHz bus, it will be difficult to upgrade to a full speed processor in the Deschutes timeframe with this many loads.

SUMMARY

Compaq carefully evaluated the options for designing an 8-way SMP server and chose the optimum solution for the next-generation standards-based server. By entering a technology exchange agreement, Compaq and Corollary are ready to take full advantage of the Deschutes/Slot-2 processor in SMP servers. The combination of Compaq’s I/O design with Corollary’s Profusion system controller and the next-generation Deschutes/Slot-2 processor by Intel will yield a balanced, high-performance system architecture for the future.

The resulting architecture will be made available to standards-based system providers. Several major system suppliers have already licensed this architecture from Corollary, including Data General, Hitachi, and Samsung. Intel’s recent announcement to acquire Corollary further enhances the longevity of this architecture. Customers are assured that applications will be optimized for this architecture as the industry standard for IA-32 based servers beyond four processors.

[5] *“Axil’s Adaptive Memory Crossbar” Architecture for 8-Way SMP Pentium Pro Servers Provides Breakthrough Scalability and Price/Performance*, Axil Computer, Inc., Document Number NX102, 1997.
[6] *ibid*

APPENDIX: RELATED DOCUMENTS

These documents are available on the Compaq web site. The URL address for all the documents is http://www.compaq.com/support/techpubs/whitepapers/________.html. To access one of these documents on the web, replace the blank in the URL address with the URL file name.

8-Way Technology and the Compaq ProLiant 7000, document number ECG050.1097, URL ECG0501097

Compaq Highly Parallel System Architecture, document number 597A/0697, URL 597A0697

Deploying PCI Hot Plug on Compaq Servers in a Microsoft Windows NT Environment, document number 064A/0797, URL 064A0797

ECC Memory, document number 100A/0395, URL 100A0395

PCI Hot Plug Technology, document number 398A/1196, URL 398A1196

PCI Hot Plug Technology with Novell Architecture, document number 131A/0397, URL 131A0397

PCI Hot Plug Technology with SCO Software Architecture, document number ECG078.0997, URL ECG0780997

Pentium II Processor Technology, document number ECG046.0897, URL 046_0897

Performance of Pentium Pro and Pentium II Processor/Cache Combinations, document number 436A/0597, URL 436A0597

Positioning Pentium II and Pentium Pro in Server Environments, document number 235A/0797, URL 235A0797