Profusion*

An 8-way Symmetric Multiprocessing Chipset

July 1999
Intel® Corporation

Table of Contents

Figures:

Introduction

The following paper details the architectural features of the Profusion* chipset, a highly integrated 8-way symmetric multiprocessing (SMP) chipset designed for Intel enterprise-class microprocessors such as the Intel® Pentium® III Xeon™ processor family. This paper discusses the buffered non-blocking bus switch, the cache coherency filters, the I/O bridge, RASUM features, and architectural competitive issues.

The proliferation of and reliance on the availability of computer resources is an undisputed fact in business. Businesses are highly motivated to embrace the Internet way of life. New Internet-based business ventures are forming, and old businesses are adopting the new paradigm of e-commerce, Web-based advertising, online trading and services, and so on. The computer industry is expected to facilitate the movement onto the Internet by providing more robust, cost-effective, and viable computer solutions. The computer industry faces a more urgent challenge to make available more powerful, lower cost, open-architecture, versatile and reliable computing platforms. The industry also needs to promote an open architecture and make use of commercially available operating systems and application software.

The Intel® Profusion chipset-based SMP architecture enables computer manufacturers to respond to business' complex, mission-critical applications that require the highest levels of computing and processing power. Such platforms make use of the Intel Pentium III Xeon microprocessor, which is available in high volumes in a single server platform.

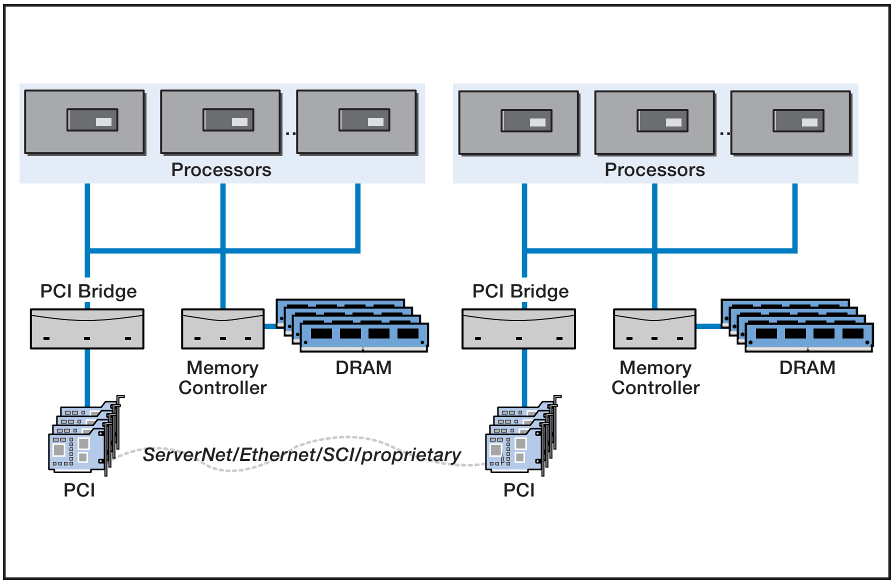
The Profusion chipset addresses the fundamental requirements of enterprise computing by focusing on raw data-crunching performance and high-throughput I/O performance. It also includes reliability, availability, serviceability, usability, and manageability (RASUM) features to lower the total cost of ownership (TCO).

Profusion chipset-based SMP server platforms promote and enable the industry with Standard High-Volume (SHV) servers that incorporate open-architecture hardware and software. This satisfies the needs of the largest corporations to run workloads once only possible using proprietary RISC/UNIX-based systems.

1.0 Shared Memory Multiprocessing

Multiprocessing (MP) can be confined into a single server platform, a node, or a combination of nodes working concurrently. An SMP server allows multiple processors in a single server to work concurrently on multiple tasks . In a clustered architecture, the SMP platforms become the building blocks of the system. In this architecture, multiple SMP platforms are interconnected, usually in a proprietary fashion, to form a larger system. Figure 1 shows a typical SMP-based cluster architecture.

Figure 1. Clustered Architecture



The textbook definition and industry implementation of multiprocessing (MP) servers comes in a wide variety of architectures and sizes, each optimized for a particular application. The most common MP architectures, however, are:

- NUMA Architecture
- Bus-based SMP Architecture
  - Single-bus SMP
  - Multi-bus SMP
- Switch-based SMP Architecture

The Profusion chipset architecture is a hybrid of the multi-bus and switch-based SMP architecture. It is tuned for eight processors and includes a novel approach to providing efficient cache coherency. The Profusion architecture allows up to eight processors to operate at full system bus speed with a cache-coherent, dedicated data path for I/O. Multiprocessing systems based on the Profusion chipset scale cost-effectively, and deliver a sound solution that effectively competes against proprietary RISC-based systems.

2.1 NUMA Architecture for Very Large UNIX Systems

NUMA (Non-Uniform Memory Access), is one type of a cluster architecture. The advantage of NUMA is its ability to use inexpensive standard high-volume (SHV) server building blocks to create a larger system. The architecture offers an ideal implementation for fail-over. However, without aggressive, low-latency, high-bandwidth interconnects or large-node caches, you must restructure the operating system and application software to achieve optimal performance. Compared to SMP, NUMA becomes cost effective only with large numbers of processors.

The figure below shows a typical NUMA architecture connected by the Scalable Coherent Interface (SCI).

2.2 Bus-based Multiprocessing Architecture

There are several approaches to a bus-based SMP solution. Single-bus SMPs have traditionally been the most cost-effective way to build small shared-memory multiprocessing systems. For a long time, modern microprocessors have been designed to enable small-scale (four or fewer processors) bus-based SMPs. For a small number of processors, the bus-based SMP platform offers good latency and bandwidth. However, with the advent of higher speed and more powerful processors, the bus bandwidth and scalability become a bottleneck.

Multi-bus SMP systems are one way to address the performance and the scalability requirements of a multiprocessing platform. The hierarchy of buses and multi-ported memory architecture are also commonly used MP architectures.

Figure 2. NUMA Architecture for Very Large UNIX Systems

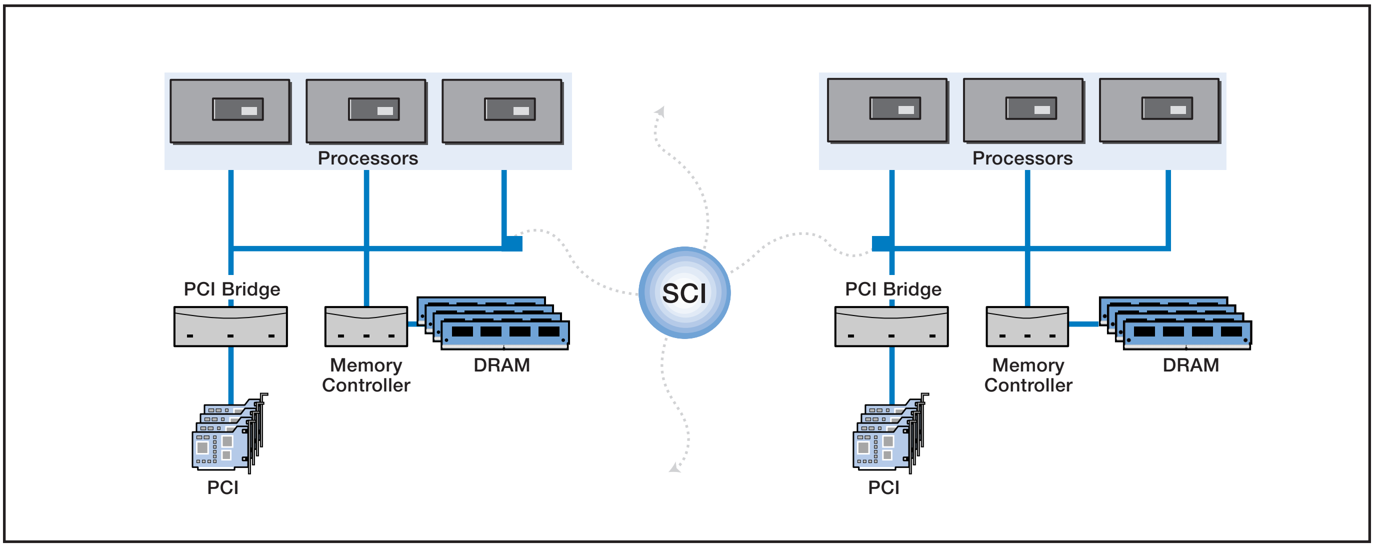
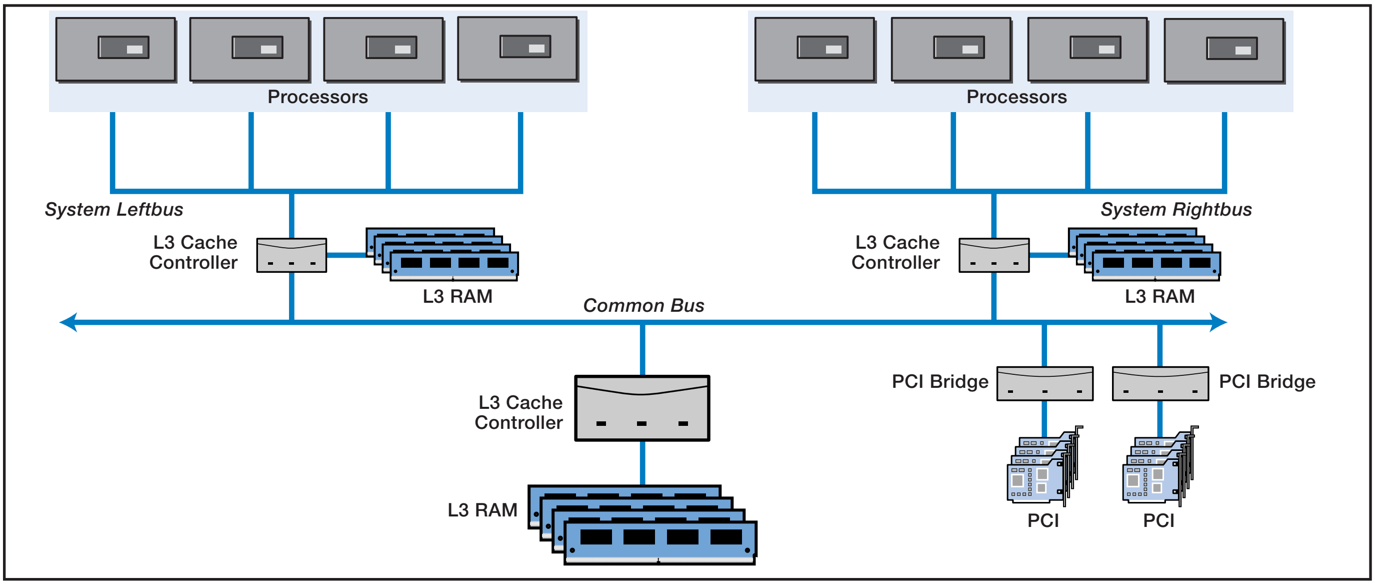


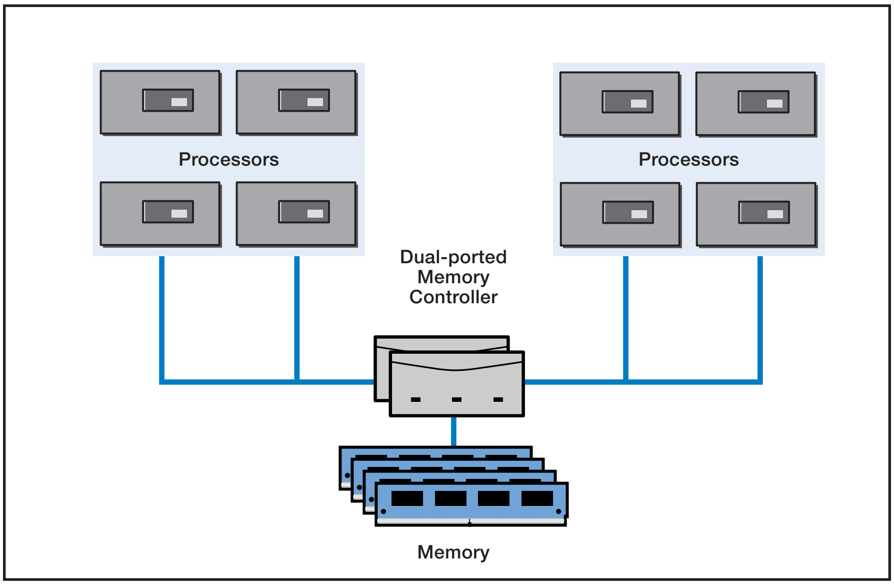
Figure 3. Hierarchy of Bus Architecture



2.2.1 Hierarchy of Buses

Perhaps the most obvious multi-bus architecture is the hierarchy of buses. In this scheme, the processors are organized into single bus nodes that are connected through hierarchical buses to form a single system. This method uses a system bus as a common bus, then attaches four-way subsystems to the bus with a third-level cache to filter memory references before reaching this common bus (see Figure 3). While this architecture supports eight processors (and potentially more), it requires a high-cost cache and greater memory latency due to its multi-level architecture. In a four-CPU configuration, it could cost more and deliver lower performance than a standard 4-way platform.

Figure 4. Dual-ported Memory Architecture



2.2.2 Dual-ported Memory Architecture

The dual-ported memory system architecture delivers significantly higher performance than the hierarchy of buses architecture and eliminates the need for a third-level cache. The performance is achieved at a lower cost premium over a clustered 4-way system.

A dual-ported memory design (see Figure 4) overcomes the problems inherent in a hierarchy of buses design. The architecture bypasses the need for expensive level-3 cache-impeding performance. The straightforward implementation offers only two system buses, which overcomes the worst-case transaction of traversing three buses in a hierarchy of buses. In a dual-ported memory design, all CPUs connect to a common memory, but each system bus segment operates independently for maximum performance. The implementation issues to be solved in any such design include:

- Addressing the total system memory bandwidth
- Designing the topology of the I/O bridge connectivity
- Providing for cache coherency

2.3 Switch-based SMP Architecture

The switch-based architecture departs from the bus-based approach by implementing a massive central switch that connects processors and memory. While switch-based SMPs enable the construction of large shared-memory system, the average latency, and the cost of this solution, increases dramatically with the number of processors.

Today, switch-based SMPs represents a point solution to the scalability requirements of a specific market segment. A chipset designed with a particular number of processors in mind is not necessarily appropriate for a different number of processors. As the number of processors increases, the bandwidth through the switch must increase accordingly. The challenge of switch-based SMP solutions remains providing a balanced platform that is not over-designed and will scale. Figure 5 illustrates a switch-based SMP diagram.

2.4 Profusion Chipset System Architecture

The Profusion chipset architecture is a hybrid between an enhanced version of the dual-ported memory and a switch-based SMP, optimized for eight processor systems. It creates a “fusion” of three Pentium III Xeon processor buses and two main memory subsystems (see Figure 6). This design allows a number of CPUs and I/O bridges to access the synchronous, shared, interleaved SDRAM memory, through a high-speed channel.

Two of the system buses are dedicated to CPU attachment, while the third system bus is utilized exclusively for I/O traffic. Further improving system performance, a novel cache coherency mechanism reduces conflicts generated by accesses from one system bus to another. Although there are three system buses, no transaction ever traverses more than two buses, providing low latency and enhancing performance. The Profusion architecture is specifically designed to overcome the dual-ported memory system architecture bottlenecks by providing two independent 64-bit data paths to the interleaved memory. It also provides a high-throughput I/O channel to avoid I/O bottlenecks.

2.5 Profusion Chipset and Cluster Architecture

SMP and clustering have different strengths and can be combined to achieve both high performance and high availability. Profusion chipset-based SMP platforms can be used as the building blocks of a clustered system architecture when an application requires more than eight processors. Profusion chipset-based platform clustering is possible but not validated by Intel.

Figure 5. Switch-based SMP Architecture

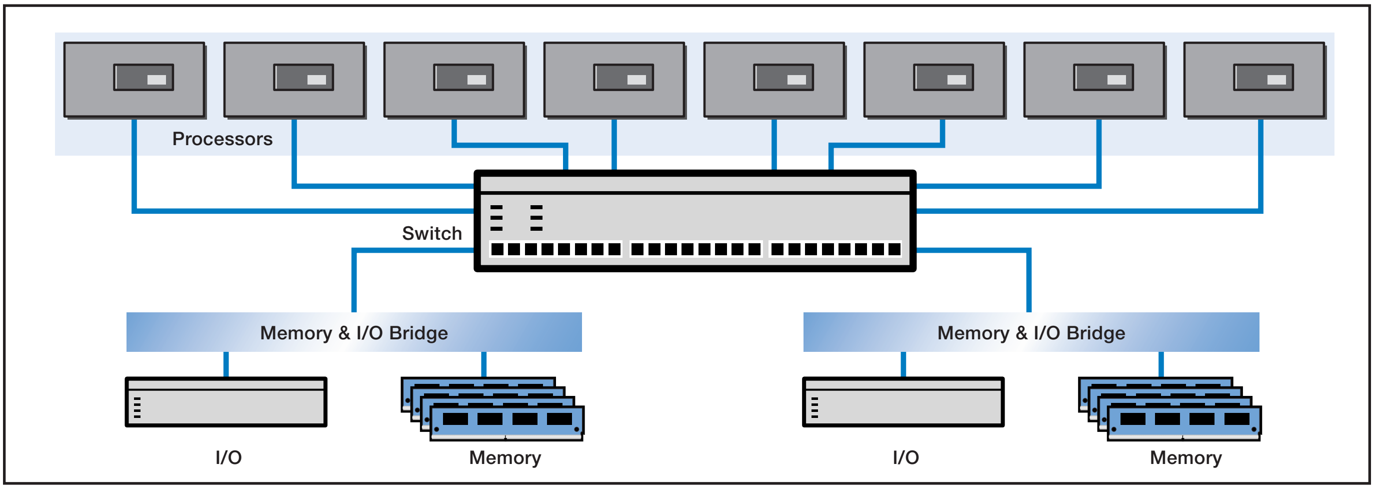
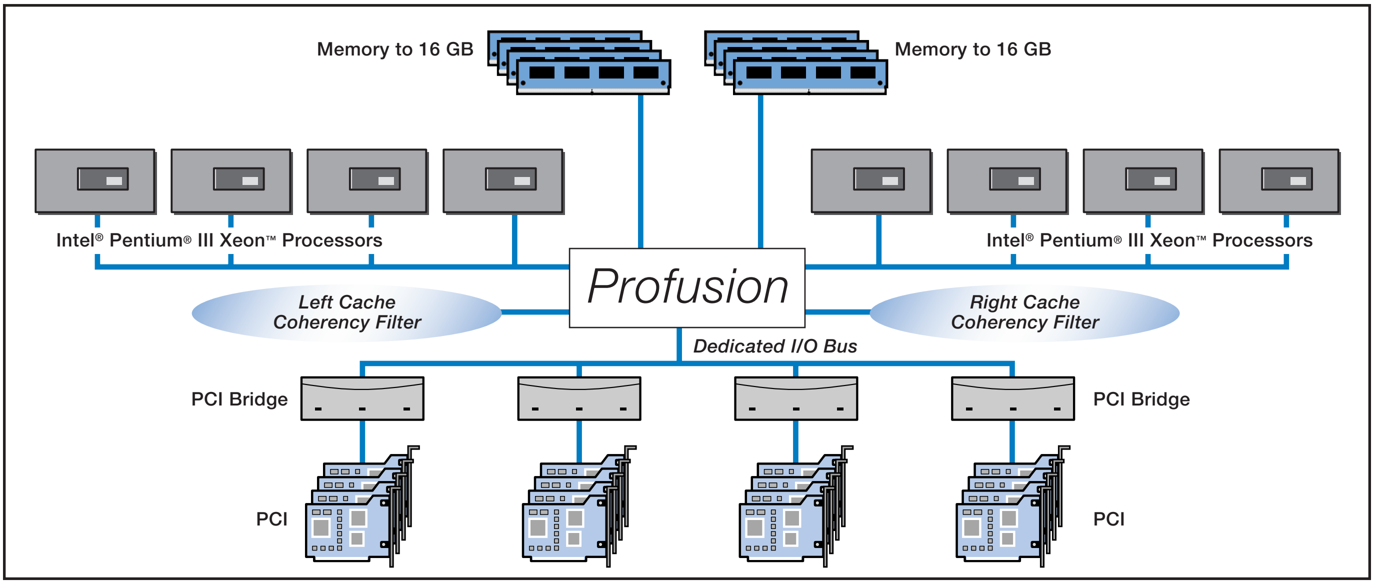


Figure 6. Profusion* Chipset Multi-ported System Architecture


Memory to 16 GB
Memory to 16 GB
Intel® Pentium® III Xeon™ Processors
Intel® Pentium® III Xeon™ Processors
Profusion
Left Cache Coherency Filter
Right Cache Coherency Filter
Dedicated I/O Bus
PCI Bridge
PCI Bridge
PCI
PCI

3.0 The Profusion Chipset Architecture

The Profusion chipset is a high-performance, tightly coupled, highly integrated 8-way SMP chipset that supports Intel Pentium III Xeon processors. The chipset features:

- Dual 100 MHz processor buses
- Dedicated 100 MHz I/O bus
- Non-blocking, multi-ported buffered switch
- Dual interleaved SDRAM memory controllers
- Up to 32 GB of SDRAM
- Dual Cache-coherency Filters
- Proven SMP software model

The Profusion chipset architecture is a bus-based solution that enables high-end server platforms to be developed around the Intel's enterprise class Pentium III Xeon processors. The Profusion architecture multiprocessing technology employs a unique multi-ported system architecture that scales and extends the system processors from one to eight processors. The Profusion chipset-based system architecture meets Intel MPS 1.4 specification, taking advantage of major operating systems, such as Microsoft Windows NT* Server and SCO UnixWare.

Profusion chipset-based platforms create a tightly coupled fusion of three system buses and two main memory subsystems, which provides the processors and I/O bridges with independent, high-speed access to shared, interleaved SDRAM memory.

Two of the system buses are utilized for processor attachment, and the third bus is dedicated for the I/O subsystem traffic. Further improving system performance, each bus is equipped with a Cache Coherency Filter to reduce cache coherency cycles between buses.

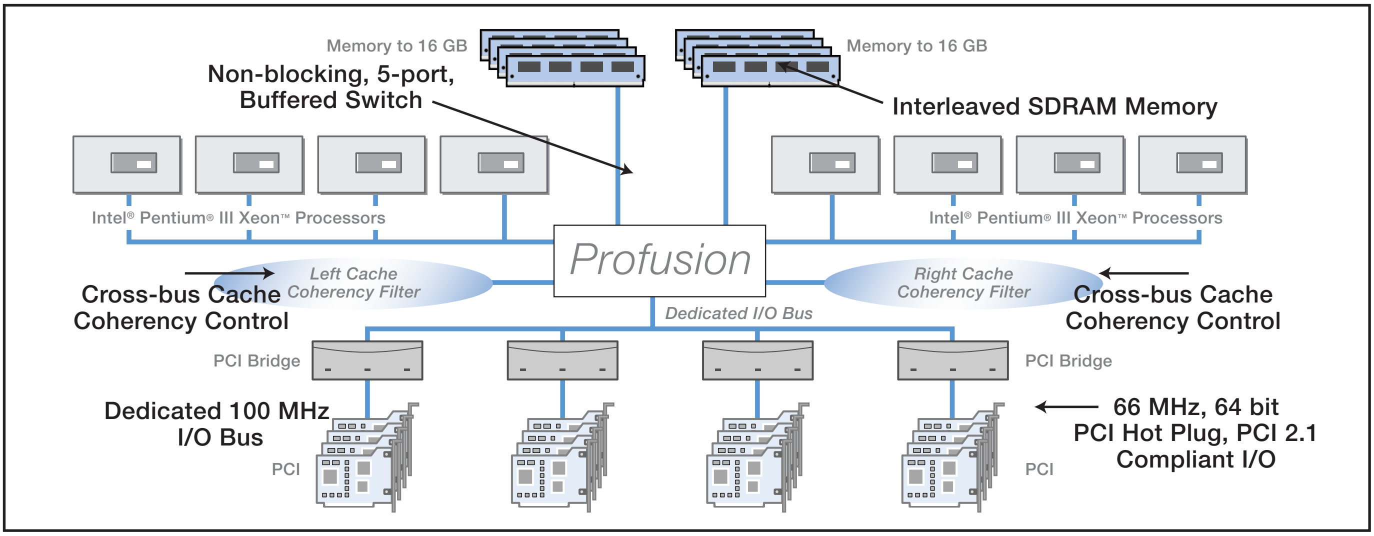
3.1 Profusion Chipset-based System Architecture

Figure 8 illustrates the Profusion chipset-based system architecture. The architecture's balanced topology allows all resources to evenly share the system bus' aggregated bandwidth. The architecture is suited for scalability, reliability, and performance.

The architecture offers balanced system performance. The burst and aggregate bandwidth are as follows:

Resource Type	Burst Bandwidth	Aggregate Bandwidth
System Bus (Right & Left)	800MB/Sec	640 MB/Sec
Total Memory Interface	1.6 GB/Sec	1.28 GB/Sec
Dedicated I/O Bus, Read	800 MB/Sec	631 MB/Sec
Dedicated I/O Bus, Write	800 MB/Sec	434 MB/Sec

Figure 7. Profusion Chipset-based System Architecture


Memory to 16 GB
Non-blocking, 5-port, Buffered Switch
Memory to 16 GB
Interleaved SDRAM Memory
Intel® Pentium® III Xeon™ Processors
Intel® Pentium® III Xeon™ Processors
Left Cache Coherency Filter
Cross-bus Cache Coherency Control
Profusion
Right Cache Coherency Filter
Cross-bus Cache Coherency Control
Dedicated I/O Bus
PCI Bridge
PCI Bridge
Dedicated 100 MHz I/O Bus
PCI
PCI
66 MHz, 64 bit PCI Hot Plug, PCI 2.1 Compliant I/O

The memory aggregated bandwidth is for an estimated mix of reads, write and refresh operations. The I/O bus write aggregate bandwidth is impacted by strong write ordering.

3.2 The Profusion Chipset

The heart of the Profusion chipset is the two-chip ASIC system controller, which consists of a Memory Access Controller (MAC) and a Data Interface Buffer (DIB). The Profusion chipset system controller is a multi-ported controller that provides interconnection among two Intel Pentium III Xeon processor buses, one dedicated I/O bus, and two shared system memory resources. The following functions are built in the Profusion chipset system controller chipset:

- Three independent Intel Pentium III Xeon processor bus interfaces
- Two independent SDRAM memory controllers
- Five-port data path connects the processor and memory ports
- Large pool (64) of cache line buffers
- Concurrent data transfers on all ports
- Two Cache Coherent Filter interfaces to external SRAM
- An integrated I/O Bus Coherent Filter

Figure 9 shows the block diagram of the 5-port non-blocking crossbar switch. The architecture allows simultaneous read/write access through all paths.

The MAC-DIB chipset interconnect is shown in Figure 10 (see page 14).

Figure 8. Profusion Chipset Non-Blocking 5-port Crossbar Switch

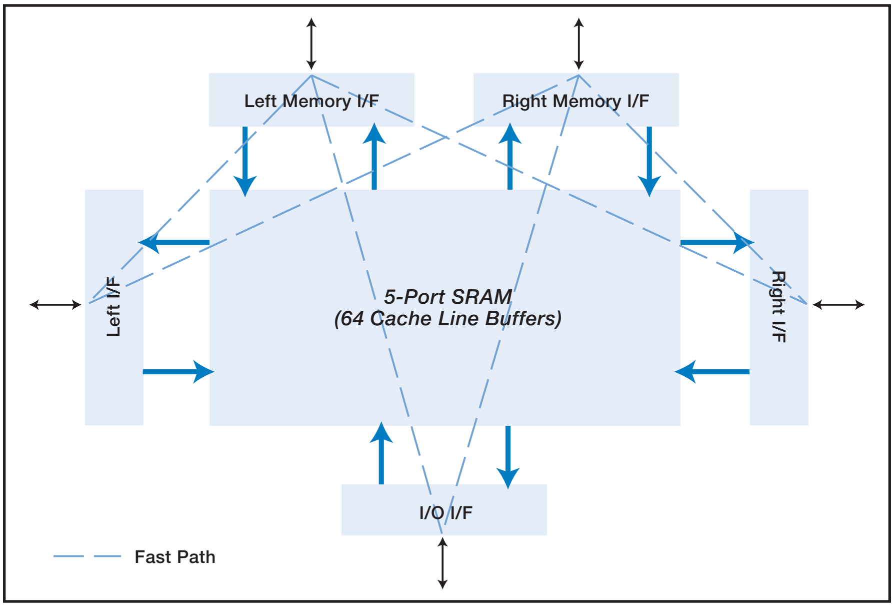
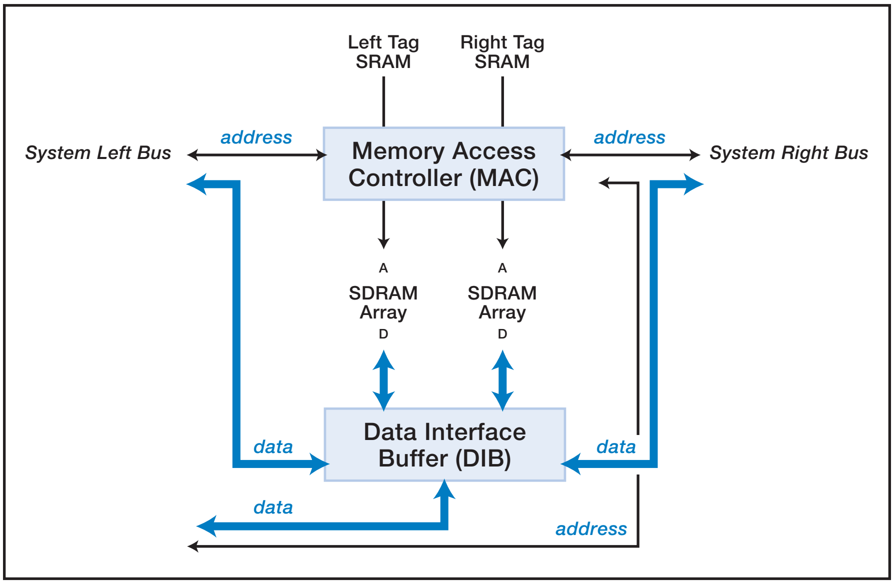


Figure 9. Profusion Chipset Interconnect



3.2.1 Memory Access Controller (MAC)

The MAC provides the control function for the Profusion chipset controller and includes the following features:

- Multi-ported memory controller
- Three-way Pentium III Xeon processor bus bridge
- Support for up to 32 GB SDRAM
- External Coherency Tag SRAM Management

The Central Request List is the major structure in the MAC. It contains the Buffer Manager, which maintains the status of each of the 64-cache line buffers inside the DIB chip. An address associated with each buffer is snooped during each memory operation to ensure coherent data. Additionally, the Central Request List is the controlling structure that forwards transactions as required to maintain coherency. This structure, working with the corresponding Coherency Tag Control Module, prevents coherency traffic from reaching the Pentium III Xeon processor bus unnecessarily.

3.2.2 Data Interface Buffer (DIB)

The DIB provides the data path for the Profusion chipset controller. It is controlled by the MAC and is used to move data between the system memory and each of the Intel Pentium III Xeon processor buses. The DIB performs all data ECC checking and generation. The DIB includes the following features:

- Three 72-bit Intel Pentium III Xeon processor bus data ports (AGTL+)
- Two 72-bit SDRAM data ports (LVTTL)
- Concurrent data transfer on all ports
- ECC generation / single bit correction
- 64 cache line buffers

The basic DIB structure is a multi-ported SRAM used as a memory buffer between the left system bus, right system bus, the dedicated I/O bus, and the two SDRAM data buses. Data can be accessed on each bus independently and concurrently with the other four buses.

In addition to the data paths through the SRAM, dedicated fast data paths bypass the SRAM and provide low latency delivery of data from the memory ports directly to the processor and I/O ports.

The SRAM has five write ports and five read ports. A pair of configuration and status data registers in the DIB facilitate access to the configuration and status registers in the MAC. Additional DIB configuration and status registers hold the ECC syndromes of the first correctable ECC error detected (and corrected) for each of the five data ports.

3.2.3 The Coherency Filters

To keep data coherent on all three buses, memory transactions on the local bus must be snooped on all remote buses, creating unwanted bus traffic that can negatively impact the system throughput. The processor coherency filter-one for each bus-holds a superset of the processor cache tags, reducing the frequency of the dispensable snoop cycles from the remote bus. The coherency filters operate using the inclusion rule, where the coherency tag size encompasses all the processor cache line. The estimated performance gain through the coherency filter can be up to 20 percent, depending on the application.

3.3 SDRAM Memory Interface

The Profusion chipset memory interface is a dual-port synchronous design that uses commercially available PC/100 SDRAM on registered DIMM memory modules. The memory interface allows for up to 32 GB maximum memory in the platform. The memory bandwidth is 800 MB/sec per port, for a total bandwidth of 1.6GB/sec, at an idle access time of 140 nsec.

3.4 The I/O Bus

The Profusion architecture effectively addresses the server platform's need for high-throughput I/O performance and meets the requirement to support a large number of I/O peripherals. The Profusion chipset dedicates the third Intel Pentium III Xeon processor bus for I/O traffic. At 100 MHz and 64-bit, the I/O bus is capable of handling a burst bandwidth of 800 MB/sec.

The Profusion chipset enhances the I/O subsystem performance and implementation by furnishing a high-performance, single-chip 64-bit PCI bus bridge, the PB64. The PB64 I/O bridge has been developed with Compaq Computer Corporation to provide high performance, robust and balanced I/O throughput for enterprise class applications. The PB64 PCI bridge supports:

- PCI 2.1 compliant
- 66.6/33.3 MHz, 64-bit PCI
- PCI burst data throughput of 533 MB/sec
- PCI Hot Plug
- Supports up to eight 66.6 MHz, 64-bit PCI bus master I/O adapters
- Supports up to sixteen 33.3 MHz, 64-bit PCI bus master I/O adapters

The PB64 enhances system performance by implementing outbound Cache-Coherent Buffers that help eliminate memory access to the previously read data, allowing concurrent transfers to occur on all system bus segments. The PB64 can support both standard peer-to-peer transactions when the PCI bus master and target are on the same PCI segment, and traversing peer-to-peer transactions when the PCI bus master and the target are on a separate PCI bus segments. The balanced architecture of the Profusion chipset allows platforms to support large number of I/Os at the highest I/O throughput.

4.0 RASUM

The hardware implementation of system management focuses on providing fault prediction, detection and resilience. The Profusion chipset system management features enable a cost-effective approach to implementing a high RASUM system. These features include:

- Processor scaling from 1 to 8 processors
- Dual processor ports, capable of operation with one port
- System bus ECC generation and correction
- Memory bus ECC generation and correction
- I/O bus ECC generation and correction
- Dual-memory port, capable of operating with one port
- Integrated PCI Hot-Plug support
- System Management Bus (SMB) support
- PCI bus address and data parity generation
- Extensive system bus error logging for system serviceability
- Performance monitoring function

4.1 Reliability

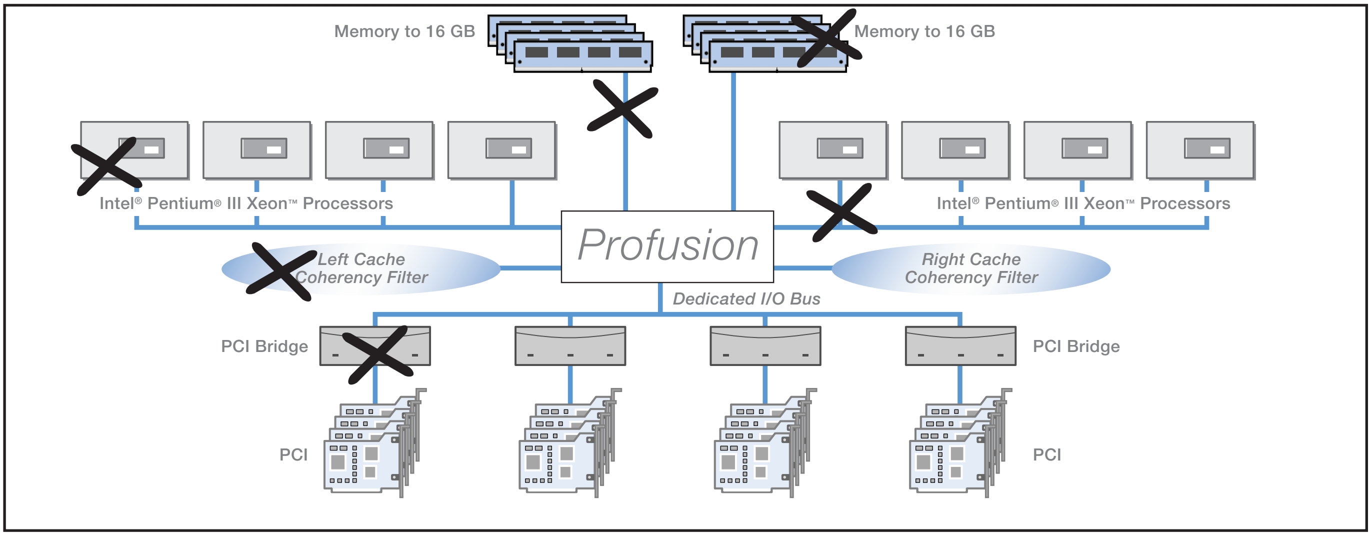
The Profusion chipset ensures system reliability by supporting Error Checking and Correction on all five ports, the two system buses, two memory buses, and the dedicated I/O. The I/O bridge supports address and data parity on the PCI bus and reports all errors on the PCI bus to the system bus, allowing implementation of a graceful error recovery mechanism by the system management software.

4.2 Availability

Server platforms based on the Profusion chipset can achieve high up-time via its inherently flexible architecture. The system architecture allows platform operation to resume in the event of failure on any of the following subsystems:

- **Any one of the processor system buses:** In the event of processor failure on any of the system buses or system bus failure, the platform can isolate that bus, and operate via the right or left system bus only.
- **Any one of the memory ports:** In the event of soft or hard memory failures, the system can adjust the memory subsystem's configuration setting from dual-port to a single-port memory configuration.
- **Any of the coherency cache filters:** In the event of soft or hard-failure, the cache coherency filters can be disabled to allow proper system operation. The system performance is compromised with system availability.

Figure 10. Profusion Chipset-based Platform High Availability


Memory to 16 GB
Memory to 16 GB
Intel® Pentium® III Xeon™ Processors
Intel® Pentium® III Xeon™ Processors
Profusion
Left Cache Coherency Filter
Right Cache Coherency Filter
Dedicated I/O Bus
PCI Bridge
PCI Bridge
PCI
PCI

- **Any of the I/O bridges:** In the event of I/O failure, each one of the PB64 I/O bridges can be isolated and disabled. The three remaining PB64 systems can continue the I/O functions. The legacy I/O bridge is the exception, since the system BIOS is accessed through this port.
- **Any of the PCI devices:** In the event of PCI I/O card failures, the PCI Hot-Plug feature allows a compliant PCI card to be replaced while the system is up and running. The system uptime is greatly improved without a major impact on system performance.

4.3 Serviceability

Fault detection and rapid repair improves system availability. This is achievable via mechanisms available to the server platform designer in an effort to detect, prevent, pinpoint and repair the ailing components of the platform. These features include:

- **Error Logging:** The Profusion chipset's extensive error-logging feature aids the system management software with the details of the ailing subsystems symptoms, which can result in rapid identification of the problem and its resolution.
- **Hot Plug:** Integration of PCI Hot-Plug into the chipset is a versatile and cost-effective feature in the chipset, allowing platform-level diagnostics without a real impact to the system operation.

4.4 Useability

Flexibility of the Profusion chipset architecture comes from the scalability of the surrounding components. Scalability exists at the processor, memory, and I/O subsystems levels.

- **Processor scalability:** The server platforms built around the Profusion chipset can start as a 4-way SMP and migrate to an 8-way SMP once the system management indicators alert IT of processing bottlenecks in the network.
- **Memory Scalability:** The dual-ported memory subsystem also allows extending the number of the memory cards and memory density.
- **I/O scalability:** The I/O bus expansion is possible at the platform level, allowing addition of clients to a network powered by higher-powered 8-way servers.

4.5 Manageability

Platforms built around the Intel Pentium III Xeon processor and the Profusion chipset support system management by making available detailed information about the platform components and characteristics. This aids IT managers in asset management, inventory tracking, configuration management and performance management. The following manageability features can be found in Profusion chipset-based platforms:

- **Performance:** the performance monitoring capabilities of the Profusion chipset enables IT managers to monitor system usage patterns so they can optimize system responsiveness and network throughput, monitor CPU utilization, capture performance utilization, and plan for capacity.

- **Asset Management:** The error-logging capability enables IT managers to detect, tag and eliminate the ailing components, and quantify spares for effective network asset management.

5.0 Summary

The choice of replacing proprietary systems with open architecture-based SMP 8-way platforms is now available to the industry with the introduction of the Profusion chipset. The Profusion chipset pushes the limits and raises the performance threshold on SHV servers by extending the concept from 4-way to 8-way SMP designs.

Profusion chipset-based platforms redefine the price/performance benchmarks for enterprise-class SHV server platforms and effectively address the complex computing needs of today's Internet-linked businesses. The architecture offers longevity, high performance, compatibility, high RASUM, and support for extended numbers of industry-standard I/O expansion slots such as PCI, and exceptional I/O bandwidth. It offers all this within the framework of commercially available, high-volume, enterprise-class processors from Intel and operating systems from industry leaders.

Information in this document is provided in connection with Intel® products. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document. Except as provided in Intel's Terms and Conditions of Sale for such products, Intel assumes no liability whatsoever, and Intel disclaims any express or implied warranty, relating to sale and/or use of Intel products including liability or warranties relating to fitness for a particular purpose, merchantability, or infringement of any patent, copyright or other intellectual property right. Intel products are not intended for use in medical, life saving, or life sustaining applications.

Intel may make changes to specifications and product descriptions at any time, without notice.

Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them.

The Pentium® III and the Profusion* chipset may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an ordering number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725 or by visiting Intel's web site at http://www.intel.com