Sun Enterprise™ 10000 System Overview Manual

THE NETWORK IS THE COMPUTER™

Sun Microsystems Federal, Inc.
A Sun Microsystems, Inc. Business
901 San Antonio Road
Palo Alto, CA 94303 USA
415 960-1300 fax 415 969-9131

Part No.: 805-0310-13
Revision A, September 1999

Please
Recycle

Sun Enterprise 10000 SSP Attributions:

This software is copyrighted by the Regents of the University of California, Sun Microsystems, Inc., and other parties. The following terms apply to all files associated with the software unless explicitly disclaimed in individual files.

The authors hereby grant permission to use, copy, modify, distribute, and license this software and its documentation for any purpose, provided that existing copyright notices are retained in all copies and that this notice is included verbatim in any distributions. No written agreement, license, or royalty fee is required for any of the authorized uses. Modifications to this software may be copyrighted by their authors and need not follow the licensing terms described here, provided that the new terms are clearly indicated on the first page of each file where they apply.

IN NO EVENT SHALL THE AUTHORS OR DISTRIBUTORS BE LIABLE TO ANY PARTY FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OF THIS SOFTWARE, ITS DOCUMENTATION, OR ANY DERIVATIVES THEREOF, EVEN IF THE AUTHORS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

THE AUTHORS AND DISTRIBUTORS SPECIFICALLY DISCLAIM ANY WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT. THIS SOFTWARE IS PROVIDED ON AN “AS IS” BASIS, AND THE AUTHORS AND DISTRIBUTORS HAVE NO OBLIGATION TO PROVIDE MAINTENANCE, SUPPORT, UPDATES, ENHANCEMENTS, OR MODIFICATIONS.

RESTRICTED RIGHTS: Use, duplication or disclosure by the government is subject to the restrictions as set forth in subparagraph (c) (1) (ii) of the Rights in Technical Data and Computer Software Clause as DFARS 252.227-7013 and FAR 52.227-19.

This is scotty, a simple tcl interpreter with some special commands to get information about TCP/IP networks. Copyright (c) 1993, 1994, 1995, J. Schoenwaelder, TU Braunschweig, Germany, Institute for Operating Systems and Computer Networks. Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that this copyright notice appears in all copies. The University of Braunschweig makes no representations about the suitability of this software for any purpose. It is provided as is” without express or implied warranty.

Please
Recycle

Please
Recycle

Adobe PostScript

Contents

Figures

Tables

Preface

The *Ultra Enterprise 10000 System Overview Manual* provides service personnel with detailed descriptions of the Ultra™ Enterprise™ 10000 system architecture. It is an aid for understanding how the system works and should be read prior to performing service-related procedures on a customer system.

UNIX Commands

This document may not include specific software commands or procedures. Instead, it may name software tasks and refer you to operating system documentation or the handbook that was shipped with your new hardware.

The type of information that you might need to use references for includes:

- Shutting down the system
- Booting the system
- Configuring devices
- Other basic software procedures

See one or more of the following:

- *Solaris 2.x Handbook for SMCC Peripherals* contains Solaris™ 2.x software commands.
- On-line AnswerBook™ for the complete set of documentation supporting the Solaris 2.x software environment.
- Other software documentation that you received with your system.

Typographic Conventions

The following table describes the typographic changes used in this book.

Typeface or Symbol	Meaning	Example
`AaBbCc123`	The names of commands, files, and directories; on-screen computer output	Edit your `.login` file. Use `ls -a` to list all files. `machine_name% You have mail.`
`AaBbCc123`	What you type, contrasted with on-screen computer output	`machine_name% su` `Password:`
AaBbCc123	Command-line placeholder: replace with a real name or value	To delete a file, type `rm` *filename.*
AaBbCc123	Book titles, new words or terms, or words to be emphasized	Read Chapter 6 in the *User's Guide.* These are called *class* options. You *must* be root to do this.

Shell Prompts

The following table shows the default system prompt and superuser prompt for the C shell, Bourne shell, and Korn shell.

Shell	Prompt
C shell	`machine_name%`
C shell superuser	`machine_name#`
Bourne shell and Korn shell	`$`
Bourne shell and Korn shell superuser	#

Related Documents

The following documents contain topics that relate to the information in *Ultra Enterprise 10000 System Overview Manual*.

Application	Title	Part Number
Service	*Sun™ Enterprise™ 10000 System Read Me First*	805-2913
Service	*Sun™ Enterprise™ 10000 System Unpacking Guide*	805-2915
Service	*Sun™ Enterprise™ 10000 System Site Planning Guide*	805-2914
Service	*Sun™ Enterprise™ 10000 Hardware Installation and De-Installation Guide*	805-4651
Service	*Sun™ Enterprise™ 10000 System Service Guide*	805-2917
Service	*Sun™ Enterprise™ 10000 System Service Reference I*	805-3622
Service	*Sun™ Enterprise™ 10000 System Service Reference II*	805-3623

Ordering Sun Documents

SunDocs[SM] is a distribution program for Sun Microsystems technical documentation. Easy, convenient ordering and quick delivery is available from SunExpress™. You can find a full listing of available documentation on the World Wide Web: `http://www.sun.com/sunexpress/`

Country	Telephone	Fax
Belgium	02-720-09-09	02-725-88-50
Canada	800-873-7869	800-944-0661
France	0800-90-61-57	0800-90-61-58
Germany	01-30-81-61-91	01-30-81-61-92
Holland	06-022-34-45	06-022-34-46
Japan	0120-33-9096	0120-33-9097
Luxembourg	32-2-720-09-09	32-2-725-88-50
Sweden	020-79-57-26	020-79-57-27
Switzerland	0800-55-19-26	0800-55-19-27
United Kingdom	0800-89-88-88	0800-89-88-87
United States	1800-873-7869	1800-944-0661

Sun Welcomes Your Comments

Please use the *Reader Comment Card* that accompanies this document. We are interested in improving our documentation and welcome your comments and suggestions.

If a card is not available, you can email or fax your comments to us. Please include the part number of your document in the subject line of your email or fax message.

- Email: `smcc-docs@sun.com`
- Fax: SMCC Document Feedback
  1-415-786-6443

CHAPTER 1

Host Overview

The Ultra Enterprise 10000 system is a SPARC™/Solaris™ (UNIX®-System V Release 4) scalable symmetrical multiprocessing (SMP) computer system. It is an ideal general purpose application and data server for host-based or client/server applications like online transaction processing (OLTP), decision support systems (DSS), data warehousing, communications services, or multimedia services.

The Enterprise 10000 system provides the following capabilities:

- Solaris 2.5.1 and 2.6 compatible.
- Internal frequency of 83.3 MHz with processors running at a clock frequency of 250 or 336 MHz.
- Gigaplane™-XB interconnect: a new generation of interconnect ASICs significantly reduces the cost of the system interconnect architecture.
- Gigaplane-XB interconnect bandwidth: up to 10.5 Gbytes/sec.
- Fast processing: up to 20 GFLOPS/sec.
- More reliability, availability, and serviceability (RAS) than other similarly architected systems.
- Error-correction interconnect: data and address buses are protected by a combination of error correcting codes and parity.
- I/O flexibility: up to 32 independent SBuses and 64 SBus slots or up to 32 independent 33/66 -MHz PCI busses with 32 slots. SBus and PCI can be mixed within a system.
- High I/O bandwidth: up to 3.2-Gbytes/sec aggregate SBus bandwidth. The Enterprise 10000 system's individual SBuses do 64-bit transfers, yielding a sustained data rate of 100 MBytes/sec per SBus.
- Up to 32 I/O slots. PCI adapters are available with 32 or 64 bit data paths and run at a clock frequency or 33 MHz or 66 MHz
- No single points of hardware failure: no single component prevents a properly configured Enterprise 10000 system from automatically reconfiguring itself to resume execution after a failure. This is achieved through a combination of redundancy and alternate pathing architecture. Once exception to this is the control board which requires manual intervention, following a failure, to switch to the alternate control board.

- System domains: groups of system boards can be arranged in multiprocessor domains that can run independent copies of Solaris concurrently. Each domain is completely isolated from hardware or software errors that can occur in another domain.
- Dynamic reconfiguration: enables the system administrator to add, remove, or replace system components online without disturbing production usage.
- Hot swapping: power supplies, fans, and most board-level system components can be exchanged while “hot;” that is, while the system is online.
- Scalable configurations: the Enterprise 10000 system can add memory and I/O slots without displacing processors.
- Service/maintenance process flexibility: the System Service Processor (SSP) connects to the Enterprise 10000 system via conventional Ethernet, permitting system administration from a remote location. The SSP reports status information using the simple network management protocol (SNMP) protocol.

To better acquaint you with the Enterprise 10000 system, this document is presented in the following order:

- The Enterprise 10000 host is a brief overview of how the system is packaged.
- The Enterprise 10000 components including the centerplane interconnect, system board, control board, 48-volt power subsystem, 19-inch AC sequencer subsystem, and cooling subsystem.
- The system service processor (SSP), its role, and how it connects to the system.
- The domain feature, its properties, and how domains are managed.
- The RAS feature and its components.

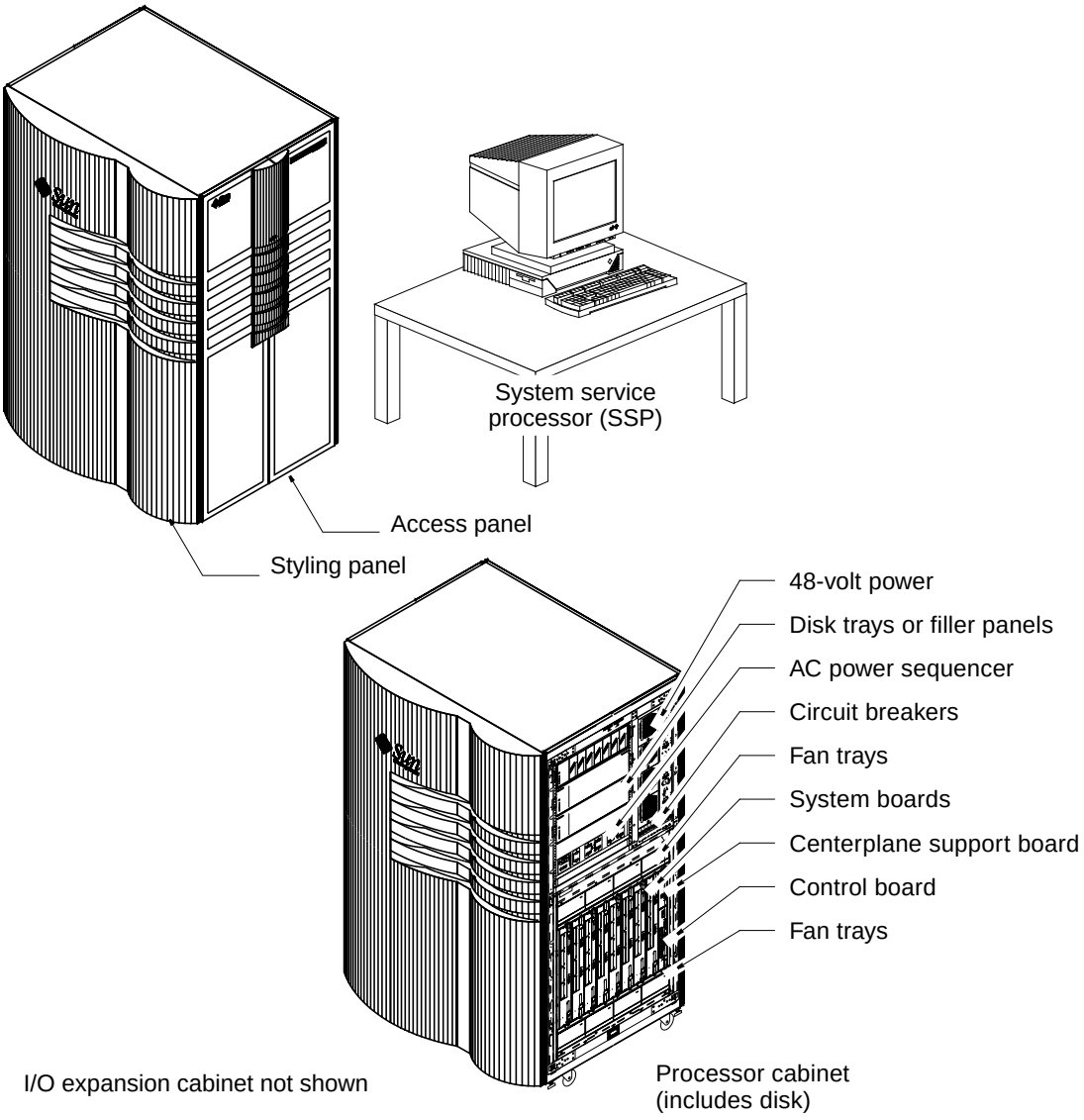

FIGURE 1-1 Enterprise 10000 Cabinet

The Enterprise 10000 system is comprised of a processor cabinet, optional I/O cabinets, and an SSP (FIGURE 1-1). The processor cabinet contains system boards, centerplane interconnect, control boards, 48-volt power subsystem, 19-inch rackmount AC sequencer subsystem, and cooling subsystem. The system boards

house the processors, I/O interface modules, and system memory. Additionally, an area is reserved in the processor cabinet for peripheral trays. The I/O cabinet is used for peripheral expansion and contains peripheral trays and one or more 19-inch rackmount AC sequencer subsystems. The SSP controls and monitors the Enterprise 10000 host hardware and software.

1.1 System Modularity

The Enterprise 10000 system is modular both in functionality and appearance (FIGURE 1-2). The functionality is modular in that system boards, I/O controllers, memory, and processors have no predefined position within the processor cabinet. That is, a system board can be located anywhere within the 16 slots allocated for system boards; and the I/O modules, memory boards, and processor modules can be distributed among the system boards to achieve the customer goals of either high availability, maximum bandwidth, or peak performance. The appearance is modular in that the processor cabinet has no fixed placement within a multicabinet configuration. Also, boards, peripherals, and subsystems are installed from either the front or the back, and all cables can egress from either of these two sides. The front of the system is denoted by the higher-numbered system boards.

To create an attractive, easy-to-service system, access doors for the front and back were designed to be light, easy to open, and easy to remove. A modular styling panel was added to a nonaccess side to provide a distinctive look for the Enterprise 10000 system. This styling panel is not removed for normal servicing.

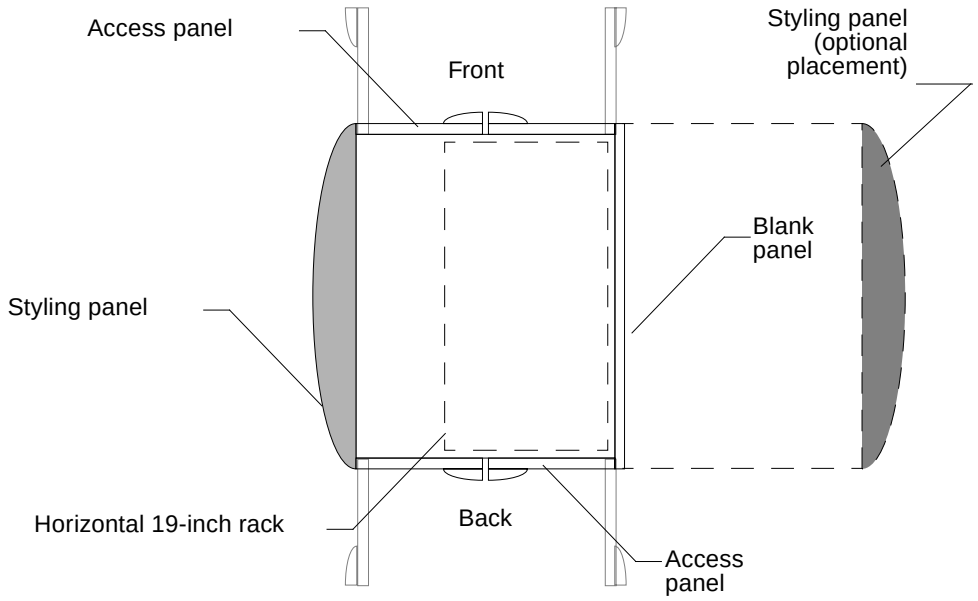

FIGURE 1-2 Styling and Access Panels (Top View)

1.2 Enterprise 10000 Component List

The Enterprise 10000 host is comprised of system boards, a centerplane, centerplane support boards, control boards, peripherals, and power and cooling subsystems (TABLE 1-1).

TABLE 1-1 Internal Component Functions

Component	Function	Quantity per processor cabinet
Centerplane	Contains address and data interconnect to all system boards.	1 (2 logical halves)
Centerplane support board	Provides the centerplane JTAG, clock, and control functions.	Up to 2
System board	Contains processors, memory, I/O subsystem, SBus or PCI boards, and power converters.	Up to 16
Processor modules	Mezzanine boards that contain the UltraSPARC™ processor and support chips.	Up to 64
Memory	Removable DIMMs.	Up to 16 Gbytes
I/O	Removable SBus or PCI boards.	Up to 64
Control board	Controls the system JTAG, clock, fan, power, serial interface, and Ethernet interface functions.	Up to 2
48-volt power system		
AC input module	Receives 220 VAC, monitors it, and passes it to the power supplies.	3 or 4
48-volt power supply	Converts AC power to 48-volt DC.	5 or 8
Circuit breaker panel	Interrupts power to various components within the system.	1
19-inch rackmount universal AC power sequencer	Receives 220 VAC, monitors it, and passes it to the peripherals. This unit can be operated in either master or slave mode enabling the on/off function to be controlled by another AC sequencer.	1 or more
Peripheral power supply	Converts AC power to DC.	Located in peripheral trays
Remote power control module	Connects the remote control line between two control boards and passes it to a master AC sequencer.	1
Fan centerplane	Provides power to the pluggable fan trays.	2
Fan trays	Each fan try contains two fans for system cooling.	10 or 16

1.3 Centerplane

The Enterprise 10000 system boards communicate via the centerplane board. Unlike previous systems using a passive bus architecture, the Enterprise 10000 centerplane contains the Gigaplane-XB interconnect, which is comprised of active address and data routers capable of sustaining a high system board-to-system board address and data bandwidth. This high performance is realized by using address and data routers comprised of 34 ASIC chips that are mounted on the centerplane.

The Gigaplane-XB interconnect has four 48-bit address buses and two 72-bit data arbiters. Each address bus is comprised of four 36-bit wide (plus ECC and parity) ASICs and one address arbiter. Each of the two 72-bit portions of the data router is comprised of six 12-bit data router ASICS and one data arbiter ASIC. The two 72-bit data routers operate in a lock-step mode to produce a single 144-bit data bus. The address bus and data routers are partitioned into two electrically isolated, logically dependent sets consisting of two address buses and one 72-bit data router each. Each set has independent clocks, JTAG interface, and power, such that a failure of a component in one set will not affect the other. In the event that a centerplane component fails, the system can continue to operate in the degraded mode after being reconfigured using JTAG. Degraded modes use one, two, or three address buses and/or one 72-bit data router. System software must be quiescent during a configuration change. See FIGURE 1-3 for a diagram of the centerplane. See FIGURE 1-6 and FIGURE 1-4 for a block diagram of the data address and address arbitration on the centerplane.

1.3.1 Centerplane ASIC Chips

Two types of ASIC chips are used on the centerplane, XMUX and XARB. Mode select inputs to the chips enable them to operate in different applications. The chips and their applications are:

- 16 XMUX chips are used for the four global address routers (GAMUX).
- 4 XARB chips are used to control the global address arbiter (GAARB).
- 12 XMUX chips are used for the global data routers (GDMUX).
- 2 XARB chips are used to control the global data arbiter (GDARB).

The centerplane board-to-board communication is a point-to-point interconnect with ASIC chips used for signal routing. Addresses are routed between system boards using four sets of global address multiplexer chips (GAMUX) controlled by global address arbiter chips (GAARB). Global data is routed using a 16 x 16 nonblocking crossbar controlled by global data arbiter chips (GDARB).

1.3.2 Centerplane Configurability

The centerplane is designed such that a single component failure will not cause a system failure. This is accomplished by partitioning the centerplane into two independent sets of components that operate together unless there is a failure. Should a failure occur, the following degraded modes of operation can be used:

- The system will operate with one, two, or three address buses. Performance degradation when operating with less than four address buses will be application dependent.
- The system can operate with one 72-bit data router. Note that the data router bandwidth is two times the available address bus bandwidth in a fully operational system. Therefore, with only one 72-bit data router the system is balanced for address and data bandwidth.
- The system will operate with one or two address buses and one 72-bit data router with a half centerplane failure.
- The system board slots can be logically isolated and effectively removed from the system.
- Any system board can be hot-swapped from the system as long as the processors and contents of memory have has migrated to other boards and the I/O has been switched to another path. The active clock board cannot be hot swapped without a system boot after configuring the clock source to the alternate clock board.
- The hardware implementation in the centerplane enables the system to be partitioned into multiple domains, which can vary in size, from a single domain with 16 system boards to 16 domains consisting of one system board each.

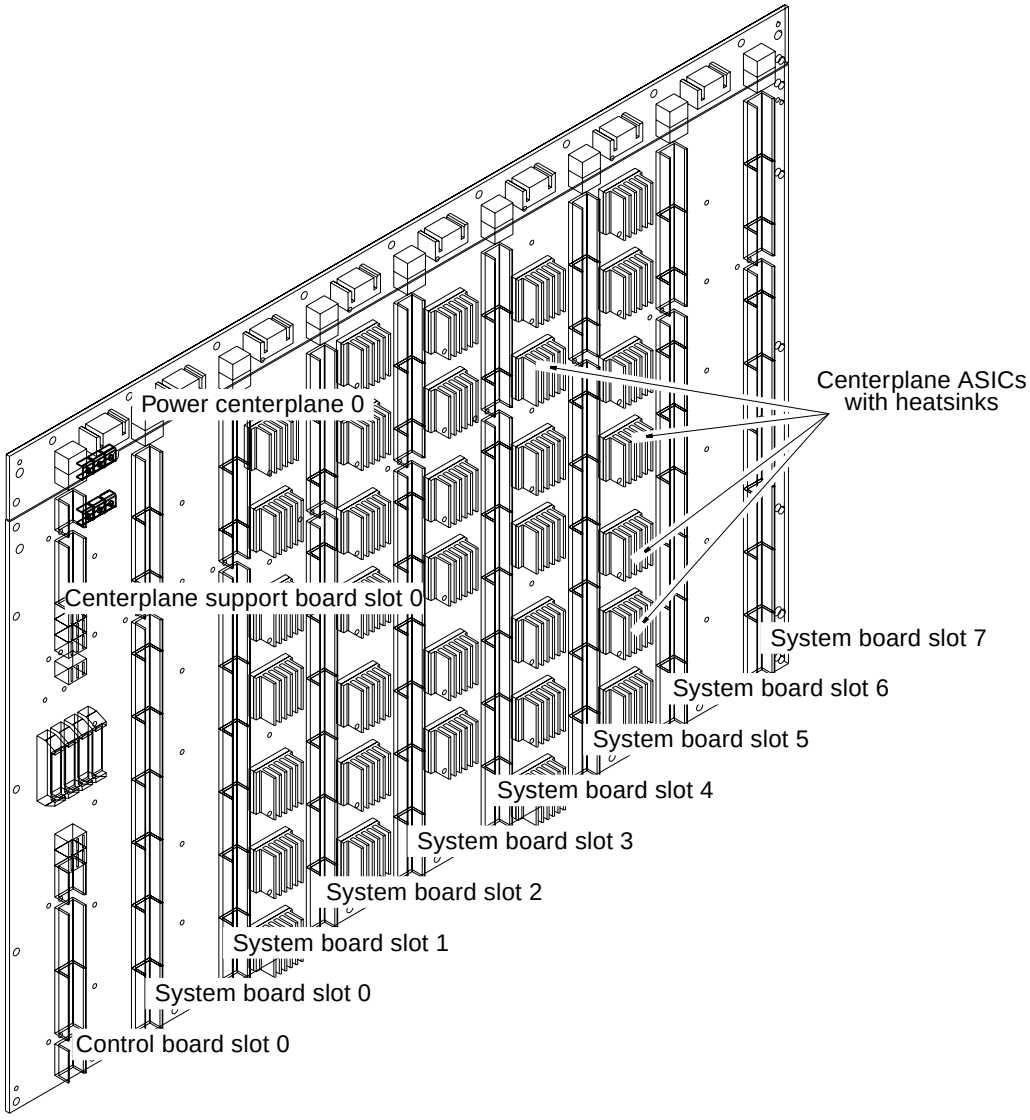

FIGURE 1-3 Centerplane (Back View)

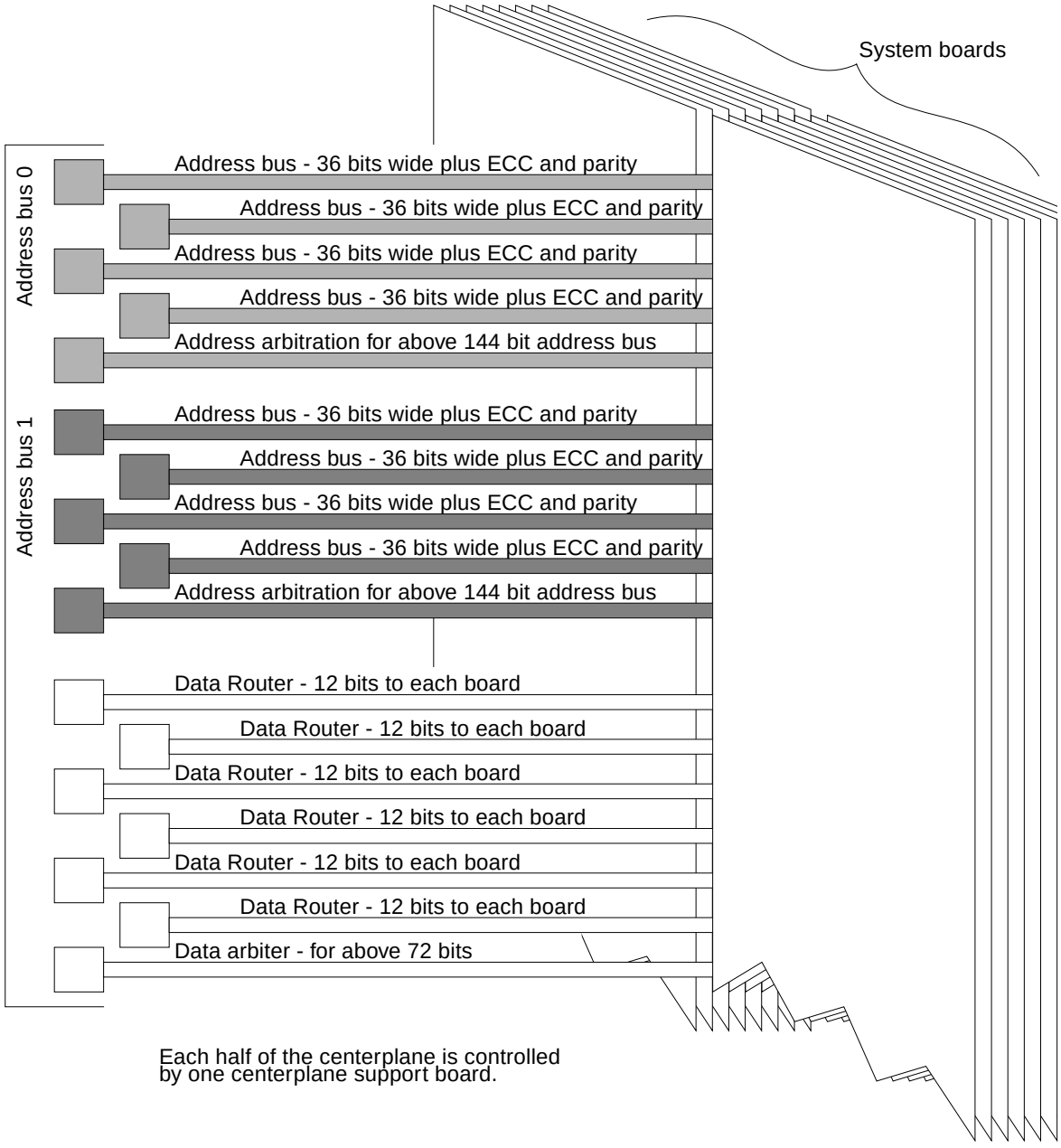

FIGURE 1-4 Block Diagram of One-half of the Centerplane

1.3.3 Ultra Port Architecture

The Ultra Port Architecture (UPA) defines a separate address and data interconnect. Usually on a bus-based system only about 70% of the wire bandwidth is available for data, with the rest being used for address and control. Separating the functions lets both addresses and data each have 100% of the wire bandwidths on their separate paths, and lets the wire topology of each function be optimized differently. Snoop addresses need to be broadcast simultaneously to all the boards, while data packets can be sent point-to point.

The UPA defines the processor and dynamic memory access interface to shared memory through a cache coherent interconnect. A UPA port has four interface properties. A UPA port:

- Can optionally be a master interface.
- Must have a slave interface.
- Can optionally be an interrupter.
- Can be an interrupt handler.

FIGURE 1-5 shows the UPA interconnection model. The model comprises four components:

- UPA ports
- Address interconnect and coherency control
- Data interconnect
- Memory

A UPA module logically plugs into a UPA port. The UPA module can contain a processor, an I/O controller with interfaces to I/O buses, and so forth. A UPA port has separate packet-switched address and data buses, and the address and data paths operate independently.

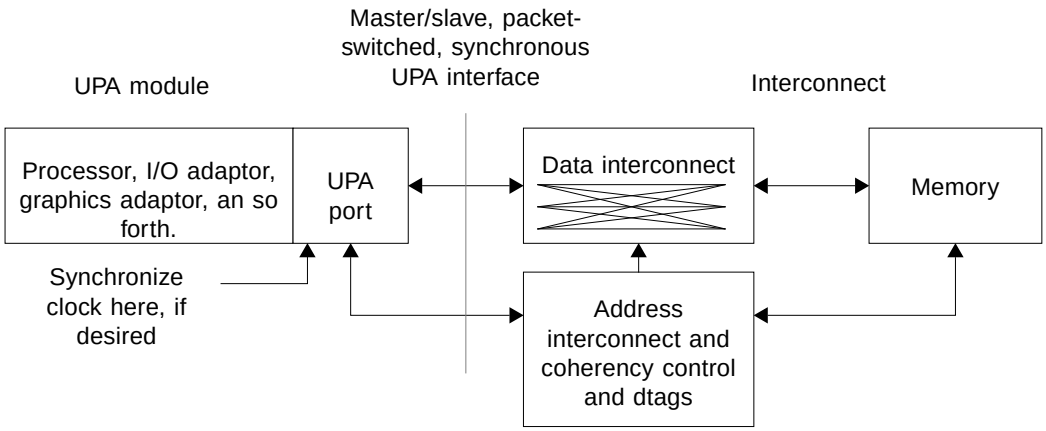

FIGURE 1-5 UPA Model

1.3.4 Enterprise 10000 System Interconnect

A combination of improvements have been implemented to increase interconnect bandwidth by tenfold over high-end bus-based systems. This amount of bandwidth is enough to keep memory latency nearly constant for data-intensive processing on full 64-processor configurations—with some headroom leftover for faster processors in the future.

1.3.4.1 Separate Address and Data Lines

The Gigaplane-XB interconnect is designed using the UPA, which defines the processor and DMA interface to shared memory through a cache coherent interconnect for a family of multiprocessor and uniprocessor systems designed around the UltraSPARC V.9 processor. The Enterprise 10000 system is the largest implementation in the Ultra Port Architecture family.

1.3.4.2 Sixteen-Byte-Wide Data Path

Relative to prior XDBus systems, doubling the data path width halved the number of cycles needed to transmit a 64-byte data packet from eight to four.

1.3.4.3 Sixteen Data Paths

The Enterprise 10000 system has 16 data paths, which provide a separate data path connection to each board.

1.3.4.4 Four Snoop Paths

The Enterprise 10000 system has four separate address snoop paths that provide enough address bandwidth to match the data bandwidth. One quarter of the memory space is snooped on each of the four address buses. A snooping chip (the coherency interface controller) keeps a set of duplicate cache tags for each of the processors on its board. It snoops its address bus looking for required cache reads or invalidates.

1.3.4.5 Point-to-Point Connections

In a multidrop bus, all the processors, I/O devices, and memory modules attach to a single set of wires. As the number of connections rises, the clock rate must be lowered to maintain reliability in the face of increasing electrical load. A failure of any component on the bus can bring down the entire bus, not just the connections to the failing component.

Changing to point-to-point connections reduces the electrical load to a minimum, enabling a faster clock rate. Point-to-point connections enable partitioning the machine into independent multiboard domains.

We call the address paths “buses” because they act like a familiar snoop-type bus for coherency transactions. Even so, the addresses are actually implemented using point-to-point switching ASICs.

1.3.4.6 Multistage Address and Data Routers

Connecting all 64 processors, 32 I/O ports, and 16 memory modules together in a giant crossbar would have been unmanageable. Instead, the Enterprise 10000 system has a two-stage routing topology based upon the physical board partitioning. Local many-to-one routers gather on-board requests, and connect them to one off-board port. A global data interconnect connects one port from each board together. Four point-to-point address buses broadcast addresses to all the boards (FIGURE 1-6).

1.3.4.7 100-MHz System Clock

The combination of faster logic gates, and point-to-point instead of bussed signals, enabled a design for a 100-MHz system clock. The UltraSPARC-I processor enables the two clocks to be either a 2:1, 3:1, or 4:1 ratio. The initial Enterprise 10000 system uses a system clock of 83.3 MHz and 250 MHz processors. However, the design is capable of using a 100-MHz system clock, which will be implemented pending release of the future, faster processors.

1.3.4.8 Pipelined UPA Implementation

Typically, when the processor does a ReadToShare UPA transaction after a load miss, the UPA protocol optimizes for the possibility that the processor will later want to modify the data, and gives the requester exclusive access if no other processor already has the data cached. This optimization enables the processor to write the data without further interconnect transactions.

On the Enterprise 10000 system, this process was modified slightly. Since the Enterprise 10000 system is a distributed, pipelined system, it cannot pipeline a conditional reply from the system of either a ReadBlockShared or a ReadBlockUnshared fast enough for our hardware implementation of the UPA. Therefore, the Enterprise 10000 system always replies with a ReadBlockShared. Thus, the processor will never enter the exclusive state, but instead uses the shared clean state. If a processor decides to store data, it must do a ReadToOwn coherency transaction to get the cache line into the exclusively modified state, and to invalidate any other caches copies.

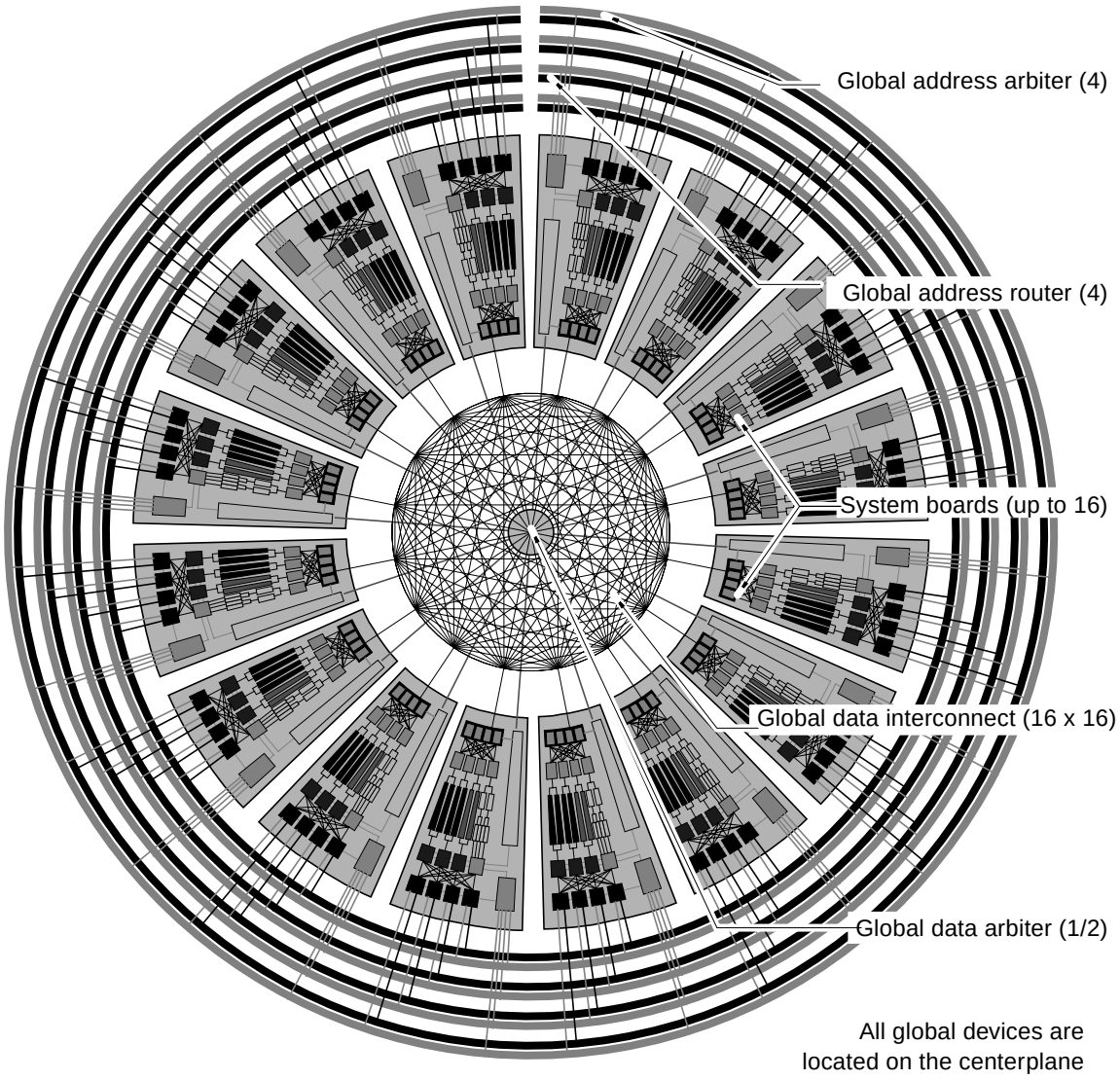

FIGURE 1-6 Enterprise 10000 System Interconnect Diagram

1.3.5 Centerplane Support Board

The centerplane support board provides the centerplane with the JTAG, clock, and control functions. There are two centerplane support boards in the Enterprise 10000 system; should one fail, the centerplane runs in a degraded mode with two of its four global address buses functioning (FIGURE 1-7).

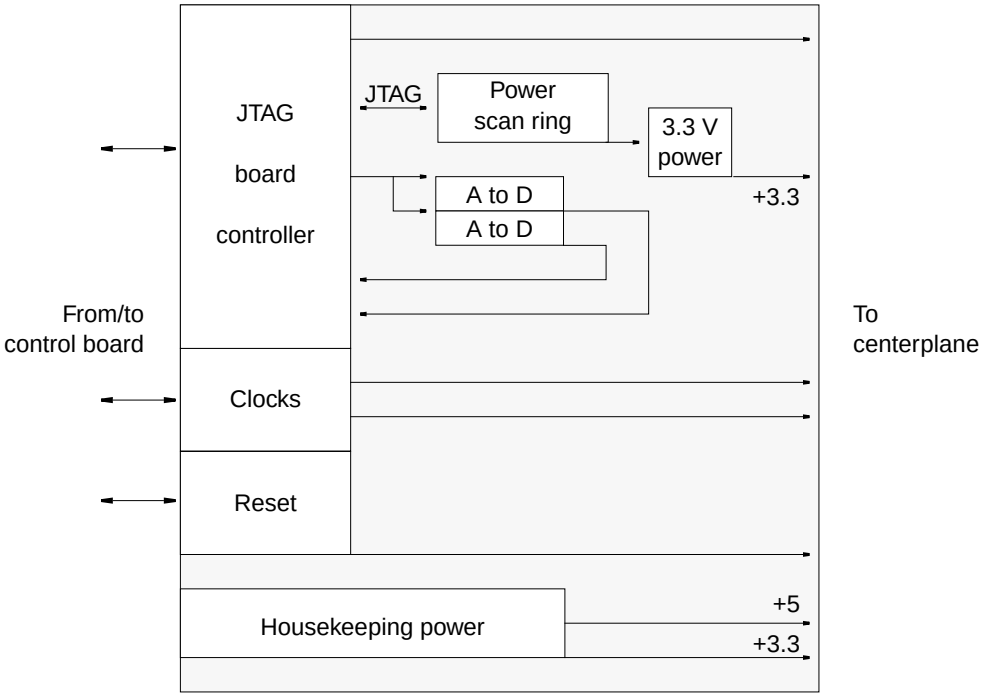

FIGURE 1-7 Centerplane Support Board Block Diagram

1.4 System Board

The Enterprise 10000 system is partitioned into system boards interconnected by a centerplane. A single system cabinet holds up to 16 of these system boards, each of which can be independently configured with processors, memory, and I/O channels, as follows:

- Four 336-MHz UltraSPARC microprocessor modules with supporting second level 4-Mbyte cache per module. Modules with faster processors and larger caches will be available as a future extension.
- Four memory banks with a capacity of up to 4 Gbyte per system board (64 Gbytes per Enterprise 10000 system).
- Two SBuses per system board; each with slots for up to two controllers for networking and I/O (32 SBuses or 64 slots per system).
- A PCI module can be used in place of the standard SBus module. The PCI module has two 66 MHz buses and each can accommodate one PCI adapter or up to 32 adapters per system. (These adapters are the 6.5-in. cards, not the 11-in. cards).

Note – The Enterprise 10000 system may have a mix of SBus and PCI.

The system board is a multilayer printed circuit board that connects the processors, main memory, and I/O subsystems to the centerplane (FIGURE 1-8). A total of 18 application-specific integrated circuits (ASICs), composed of six distinct (ASIC) types, reside on the system board. ASIC breakdown according to type is listed in TABLE 1-2. Mode select control bits in the XARB and XMUX enable them to operate in different applications. The two XARBs, when installed on the system board, operate in local address arbitration mode and local data arbitration mode. All four XMUX instances operate in the local data multiplexor (router) mode.

TABLE 1-2 Quantities of ASICs per System Board

ASIC Name	Quantity per system board
Coherency interface controller (CIC)	4
Memory controller (MC)	1
Port controller (PC)	3
Arbiter (XARB)	2
Data buffer (XDB)	4
Multiplexer (XMUX)	4

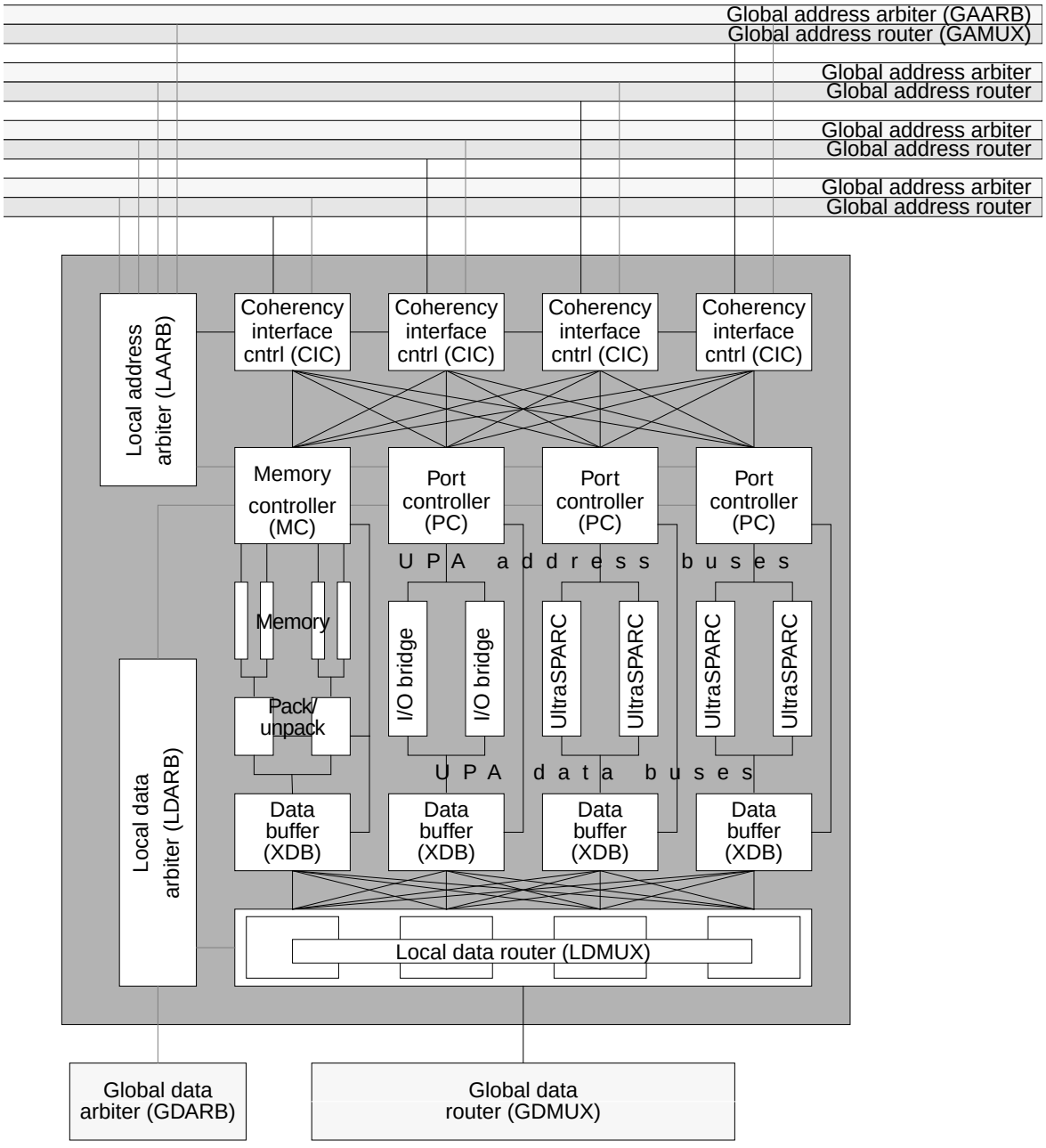

FIGURE 1-8 System Board Block Diagram

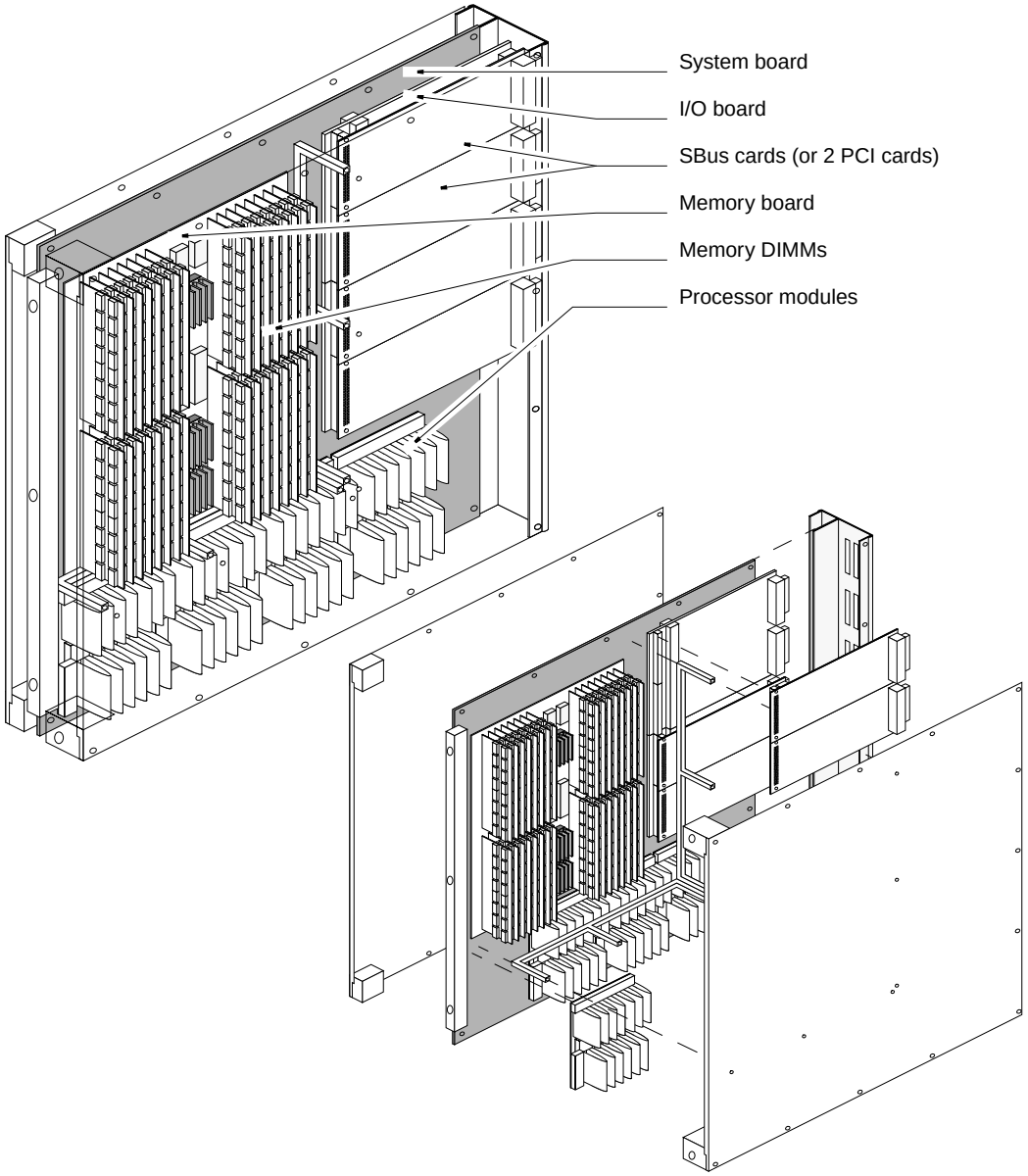

FIGURE 1-9 System Board Subassembly

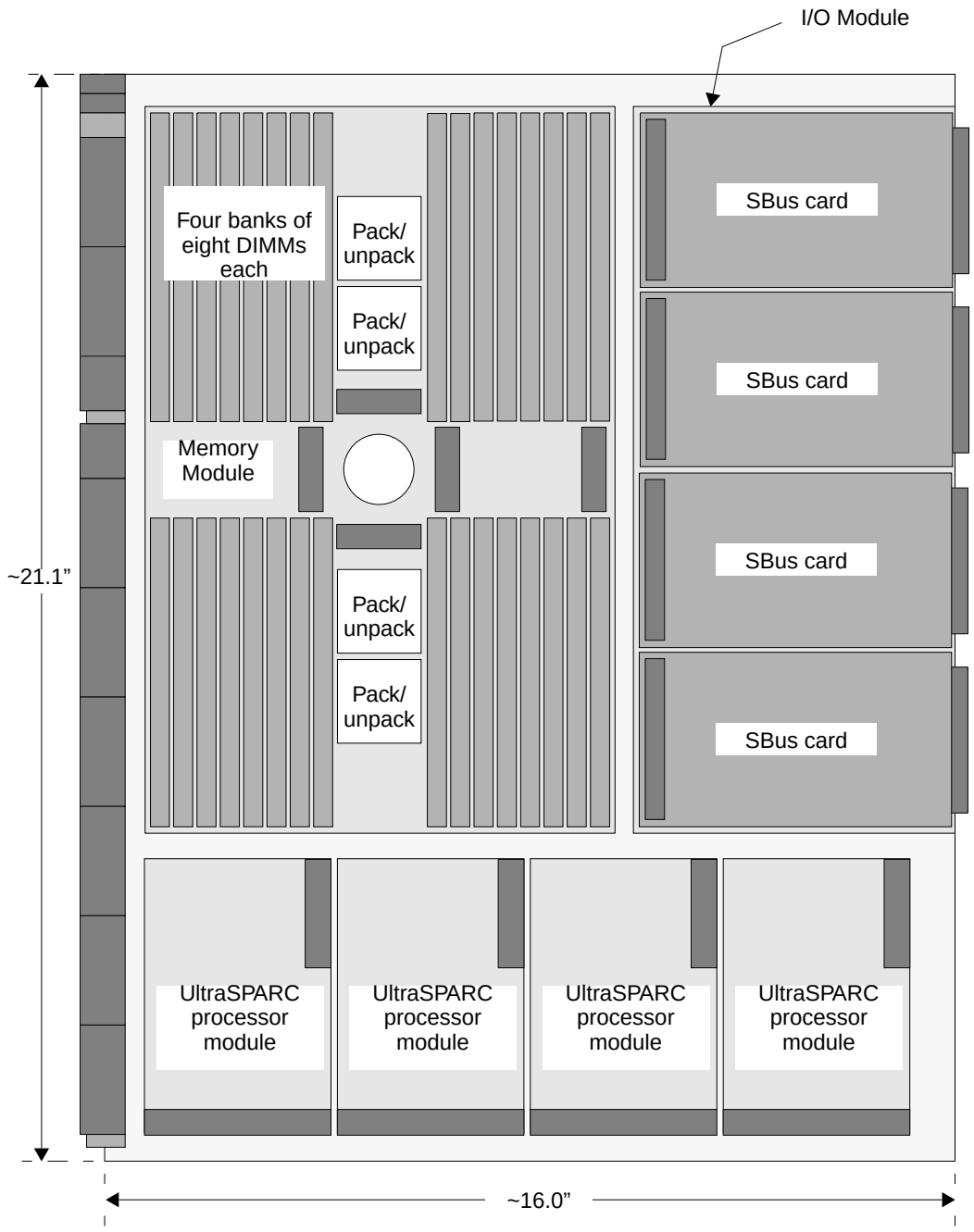

FIGURE 1-10 System Board Layout (Primary Side)

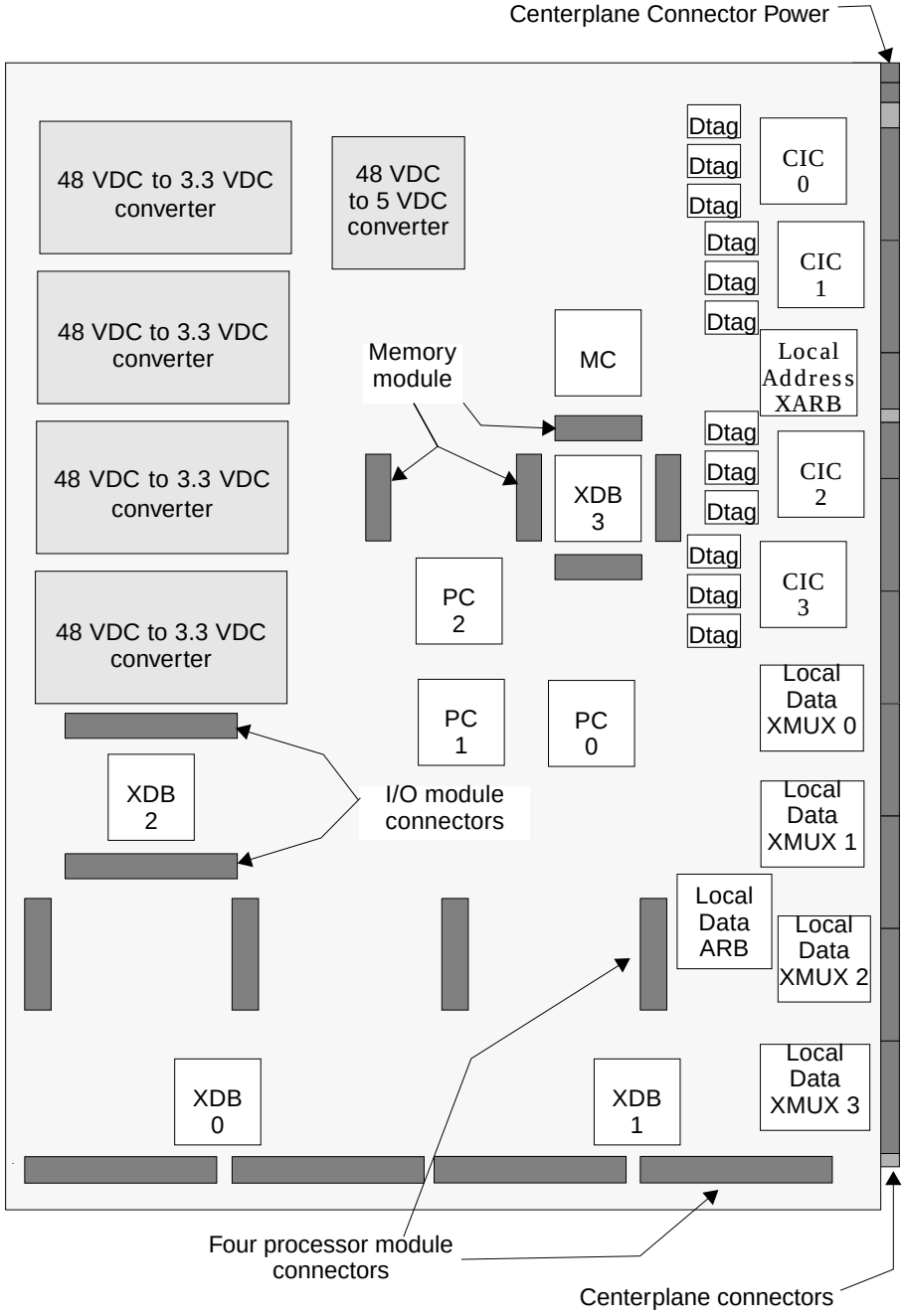

FIGURE 1-11 System Board Layout (Secondary Side)

1.4.1 Control and Arbitration

The port controller (PC) serves as the onboard UPA controller (see Section 1.3.4, “Enterprise 10000 System Interconnect), which is responsible for handling all UPA address and data requests to the system. The memory controller (MC) handles all accesses to main memory.

All data and address arbitration on the system board is handled by an arbiter while in local data arbitration (LDARB) and local address arbitration (LAARB) mode, respectively.

1.4.2 Data Interconnect

The global data router is a 144-bit-wide, 16 x 16, crossbar that steers data packets between the 16 system boards and their ecaches located on the processor modules (FIGURE 1-12). From a hardware point of view, the global data router consists of 12 multiplexer ASICs situated on the centerplane board. As a whole, the data is organized in 64-byte packets. The system requires four clock cycles to transmit each packet.

Data routing is carried out with a two-stage topology based on the Enterprise 10000 system physical board partitioning. Local “many-to-one” routers on the system boards gather on-board requests and direct them to one port (per board). The global data crossbar connects 16 system board ports together. With the 16 x 16 crossbar, any port can be connected to any other throughout the centerplane.

The local data router (LDMUX) is used to build a system board data path of 144 bits. The LDMUX is organized as a five port by 36-bit multiplexor with each port comprised of one 36-bit input and one 36-bit output. One port connects to the centerplane global data multiplexor (GDMUX) and the other four connect to data buffers (XDBs). Data transfers between the XDBs and the LDMUX are controlled by the LDARB and either the PC (for processors) or the MC (for memory).

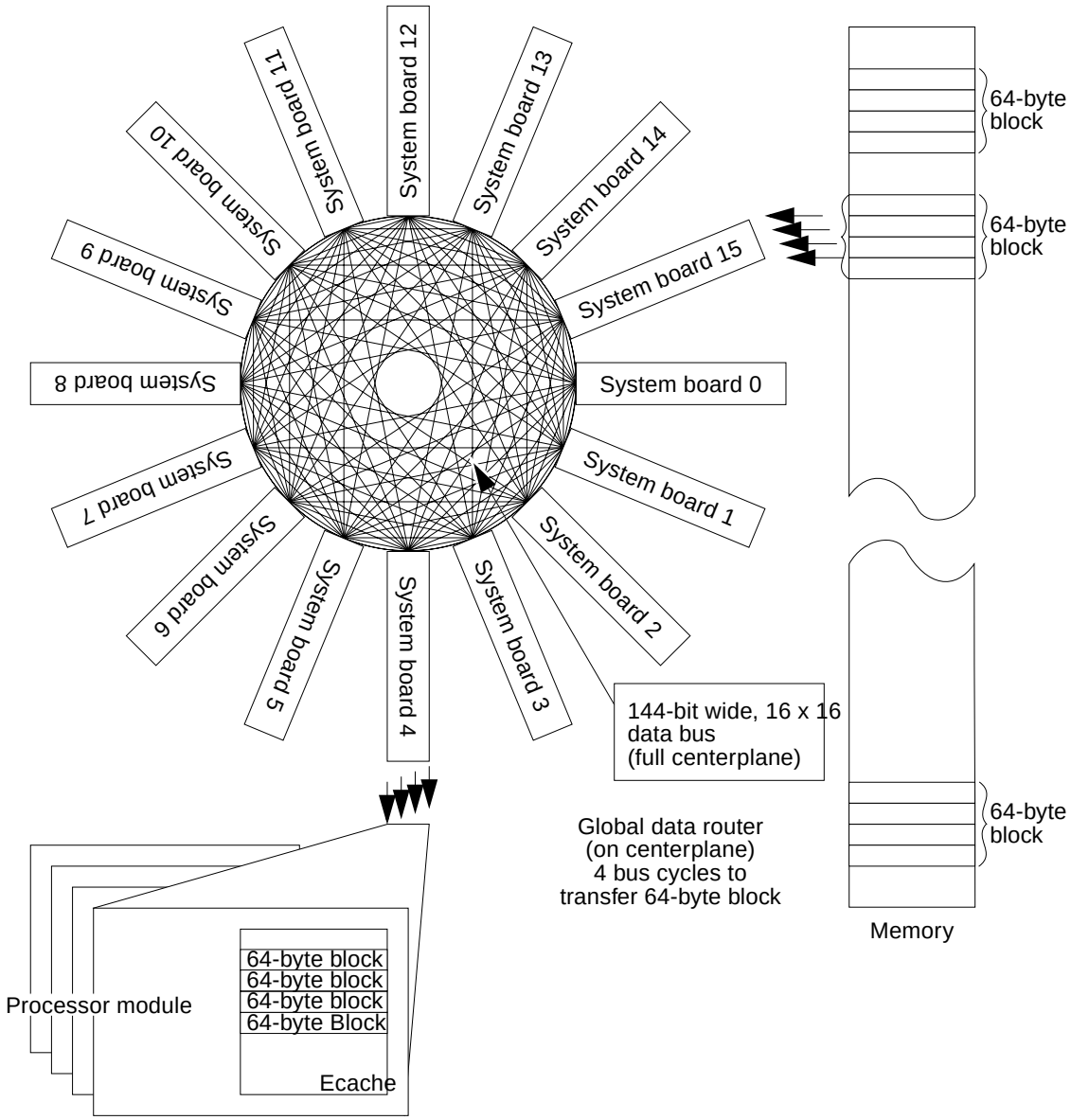

FIGURE 1-12 Data Routing

1.4.3 Address Interconnect

The Enterprise 10000 system address routing is implemented over a separate set of four global address buses (FIGURE 1-13). Although called *address buses* to convey that addresses are broadcast, the implementation is as a point-to-point router. The significance of this is that routers have more inherent reliability than does a bus, as there is only one load per driver. The buses are 48 bits wide including error correcting code bits. Each bus is independent, meaning that there can be four distinct address transfers simultaneously. An address transfer takes two clock cycles, equivalent to a snoop rate of 167 million snoops per second. Should an uncorrectable failure occur on an address bus, degraded operation is possible using the remaining buses.

The four coherency interface controllers (CICs) serve as the interface between the local address buses and the global address buses. Each CIC connects to a separate global address router through one of the four global address buses.

The CICs maintain cache coherency. Each CIC maintains an external set of duplicate tags (Dtags) in SRAM chips for the ecache of each of the four processors on its system board. There are three SRAM chips per CIC. Each CIC also maintains a one line, internal Dtag entry for each of the two I/O boards.

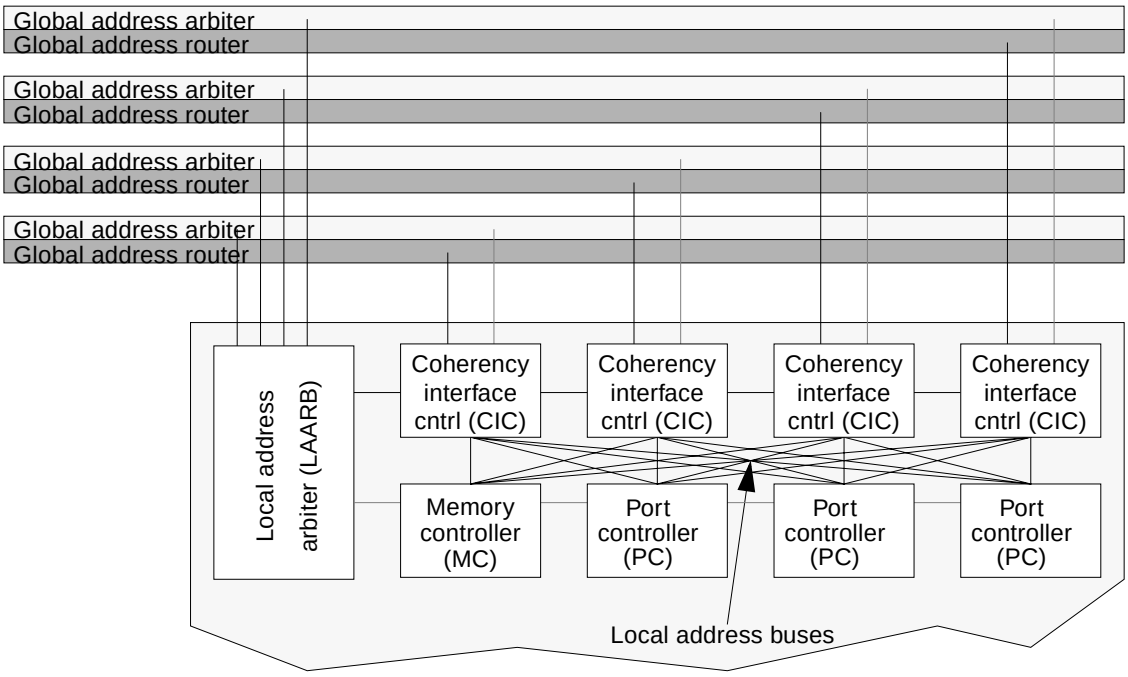

FIGURE 1-13 Address Interconnect

1.4.4 UltraSPARC-I Processor Module

The main components on the UltraSPARC-I processor module are an UltraSPARC processor, two data path chips, one 32K x 36 tag SRAM, eight 64K x 36 (for a 1-Mbyte external cache) data SRAMs, and an MC10ELV111 clock driver.

The module runs synchronously with the system interface at a 4:1 frequency ratio (336 MHz). The UPA interface to the module is through a 110-pin dual high-speed edge connector.

1.4.5 Memory Subsystem

The major components of the memory subsystem include one memory controller (MC), one Enterprise 10000 data buffer (XDB), thirty-two 32-Mbyte or 128-Mbyte dual in-line memory modules (DIMMs), and four multiplexers in "pack/unpack" mode (PUMUX). The DIMMs and PUMUX are located on a daughter card, which connects to the system board via Augat compression connectors.

Using 16-Mbit DRAM chips, a fully configured system offers 16 Gbytes of system memory or 64-Gbyte of system memory using 64-Mbit DRAM chips. The Solaris operating system has been enhanced to provide scalability consistent with this memory capacity. High memory performance is ensured by offering extensive interleaving. The entire memory data path is protected by error correcting code (ECC) mechanisms, and DIMM organization is specifically designed such that each DRAM chip contributes only one bit to a 72-bit word of data. Thus the failure of a DRAM chip causes only correctable single-bit errors.

Main memory is configured in multiple logical units. All units—memory, processors, and I/O buses—are equidistant in the architecture and all memory modules comprise a single global shared memory. Each memory unit consists of four banks of physical memory connected to the memory data buffer and the local data router to the crossbar switch (the global data router). Memory addresses are assigned through software at boot time.

A memory module can do a cache-line read or a write every four system clocks (48 ns), a bandwidth of 1300 MBytes/sec. This matches both the Global Data Router port bandwidth, and that of the processor UPA port.

1.4.5.1 Physical Overview

A system board can utilize zero, two, or four banks of memory. Two sizes of JEDEC standard 72-pin 8-byte DIMMs are supported: 32 Mbytes and 128 Mbytes. This flexibility enables the amount of memory on a given system board to vary from zero to 4 Gbytes. The connector used for the DIMMs is a JEDEC compliant through-hole component.

1.4.5.2 Functional Overview

The MC controls one to four banks of DIMMs, which serve as the main memory for the Enterprise 10000 system. The MC accepts memory address packets from the four CIC ASICs and data packets from the global data bus (via the XDB ASIC) and performs the reading and writing of 64-byte blocks of main memory.

There are two independent 576-bit data paths between the DIMM banks and the PUMUX ASICs. One path connects banks 0 and 2 to PUMUX 0 and PUMUX 2 while, the other connects banks 1 and 3 to PUMUX 1 and PUMUX 3. These are three-load bidirectional nets used to both store and load data.

1.4.6 I/O Subsystem

The I/O subsystem is a mezzanine module that resides on the system board. It provides a connection between I/O controllers and the rest of the Enterprise 10000 system (FIGURE 1-14).

1.4.6.1 Physical Overview

The major components of the Enterprise 10000 I/O interface subsystem include one port controller (PC), one Data Buffer (XDB), two I/O ASICs (SYSIO) and various standard I/O interface hardware. A block diagram of the I/O interface subsystem is shown in FIGURE 1-15. The interface between the PC and XDB (which reside on the system board) and the remaining I/O interface subsystem components (which reside on the I/O Module) logically follows the Ultra Port Architecture (UPA).

A system board can have from zero to four I/O ports. All components that make up the I/O Subsystem, except the PC and XDB, reside on the I/O Module. The I/O Module attaches to the primary side of the system board with two 288-pin Augat compression connectors.

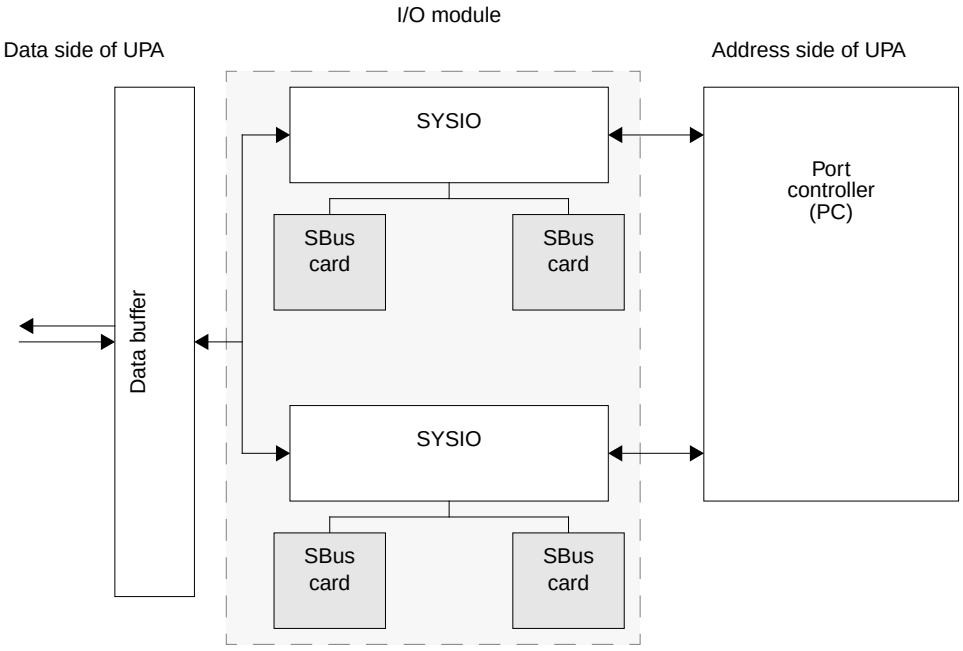

FIGURE 1-14 SBus I/O Interface Subsystem

1.4.6.2 Functional Overview

The SBus I/O port interconnection is controlled by an SMCC developed ASIC called the SYSIO. The SYSIO has a UPA interface that provides an address and data interface to the system. The address is interfaced through the PC (FIGURE 1-14), which accepts request packets from the UPA address port and controls all the data transfers to and from the UPA data port. FIGURE 1-15 provides a block diagram of the I/O module. The Reset Interrupt Scan Controller (RISC) provides handshaking between the SBus or PCI card and the SYSIO chips. The shift register and EEPROM provide board type, serial number, and revision via JTAG. The UPA address and control connector as well as the UPA data connector physically reside on the system board.

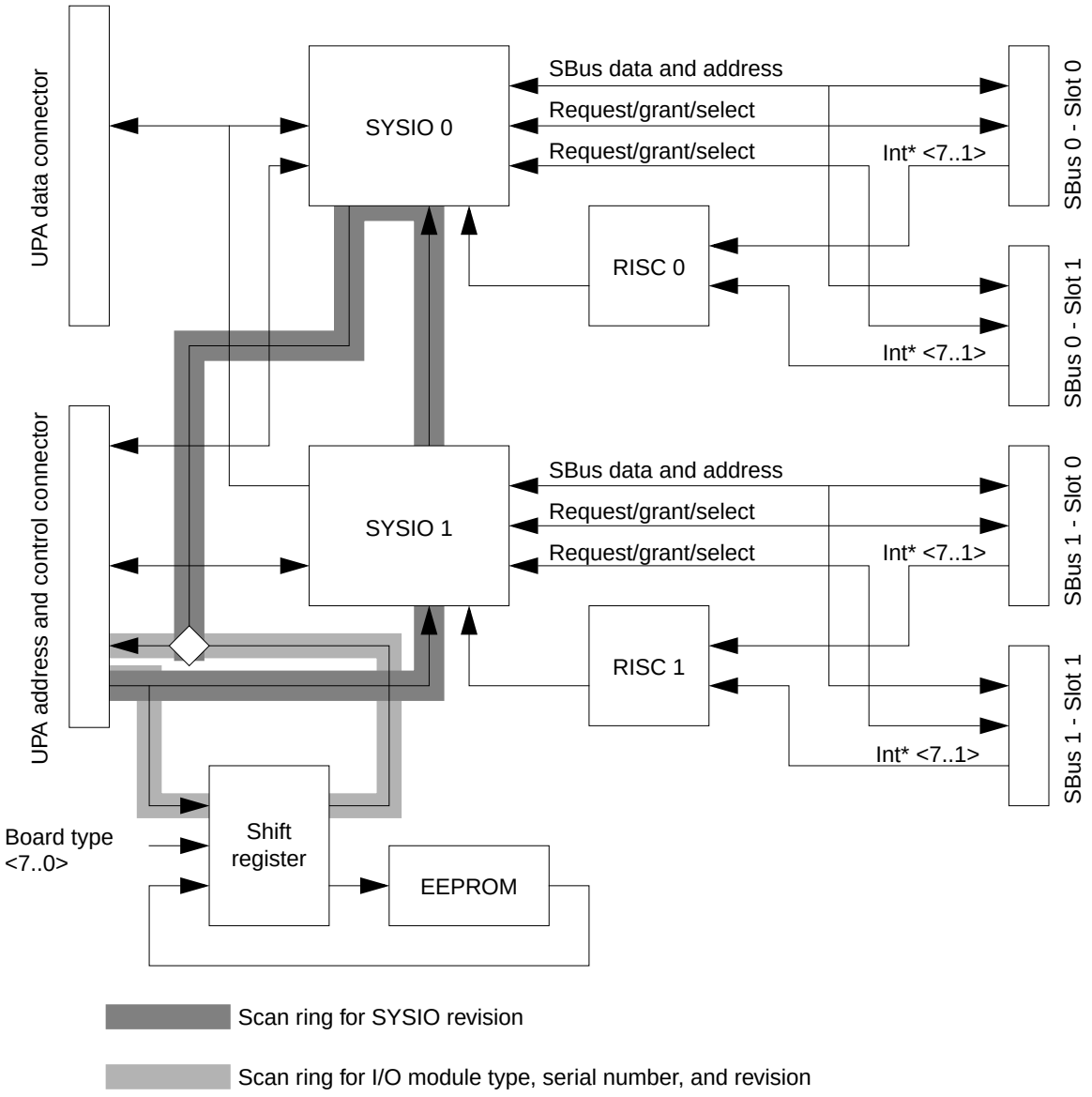

FIGURE 1-15 I/O Module

Although both logically follow the Ultra port architecture (UPA), the Enterprise 10000 system I/O interface subsystem port differs from the Enterprise 10000 processor subsystem port. For example, in the case of the I/O, support for Dtags is not relevant. Additionally, for expandability, two modes are supported in the I/O instance: 72-bit and 144-bit. Unlike a processor port, there is one physical connection for the UPA, but it supports two logical connections. The most overwhelming difference between the two implementations is the intended use for the UPA port. In this application, the UPA serves to provide a common adaptor to various industry standard I/O interfaces.

1.4.6.3 UPA to I/O Interface (SYSIO)

The SYSIO ASIC is the primary connection between the 64-bit-wide UPA system bus and the SBus I/O subsystem.

The SYSIO is designed to optimize data transfers between the UPA and the external I/O. There are dedicated paths for PIO and DVMA requests and replies. This reduces the amount of unnecessary arbitration for internal paths to some degree. The performance target is to support over 100-MBytes/sec for DVMA reads and writes.

All of the functional parts of the SYSIO can be put into one of the following four categories:

- UPA
- External I/O
- Mondo interrupt
- Internal control/miscellaneous

The UPA interface block provides address and data control and generates and checks ECC on the 64-bit data path.

The I/O interface block performs all of the standard protocol functions along with arbitration, flow control, and error handling. Additionally, the I/O interface block maps I/O virtual addresses to the UPA physical address and contains a streaming cache that accelerates I/O DVMA activity.

The Mondo interrupt block accepts interrupt requests from the I/O and SYSIO sources and dispatches interrupt packets to the UPA.

The internal control block contains a merge buffer, PIO control, BUS control, DMA control, timer counter, and JTAG control.

The graphics capability in the Ultra Enterprise 6000 system, known as a Fast Frame Buffer (FFB) is not available for the Enterprise 10000 system.

1.4.6.4 SBus Interface Hardware

The SBus is a chip-level interconnect between components for use in integrated computer systems. The SBus is designed to span only a small physical distance; it is not designed for use as a general-purpose backplane interconnect. The SBus offers the following features:

- An IEEE standard (number 1496-1993).
- A 32-bit hardware bus with 32 bidirectional data lines, parity and 28 address lines. 64-bit data transfers are possible by timesharing the address lines (and some control lines) for the upper 32 data bits.
- A three-stage information transfer cycle:
  - Arbitration to gain control of the SBus.
  - Translation of virtual addresses to physical before sending the addresses out on the bus.
  - Data transfer; this phase is extended for 64-bit operation.
- The clock rate is 25 MHz which enables a sustained data rate of 50 Mbytes/sec for 32-bit data transfers and 80 Mbytes/sec for 64-bit data transfers.

Like most computer buses, the SBus can be used in a variety of configurations. With the exception of the small amount of logic used for the SBus-controller-to-host interface, the SBus functions are independent of the host computer. This enables many types of SBus-based devices to be used on a variety of computers.

The system board is designed with two SBus controllers (SYSIO) that connect to SBus slots. These slots are populated with SBus boards that contain either SBus masters or SBus slaves. The SBus controller arbitrates over transfer requests, translates virtual to physical addresses and generally oversees the entire SBus operation. The SBus master initiates SBus transfers, supports slaves of varying widths, and provides error recovery and reporting functions. The SBus slaves respond to data transfer requests by the SBus masters. The Enterprise 10000 system supports SBus controllers for attachment of peripherals and for networking.

1.4.6.5 PCI Interface Hardware

PCI is an open I/O standard to which Sun products are moving. The main advantage of using PCI, in a server application, is PCI's higher speed compared to SBus. PCI adapters are available with 32 or 64 bit data paths and run at a clock frequency or 33 MHz or 66 MHz.

System Boards for the Enterprise 10000 system are available with a PCI module in place of the standard SBus module. This PCI module has two 66 MHz buses and each can accommodate one PCI adapter. (These adapters are the 6.5-in. cards, not the 11-in. cards).

The following figure shows how the PCI module is mounted to a system board:

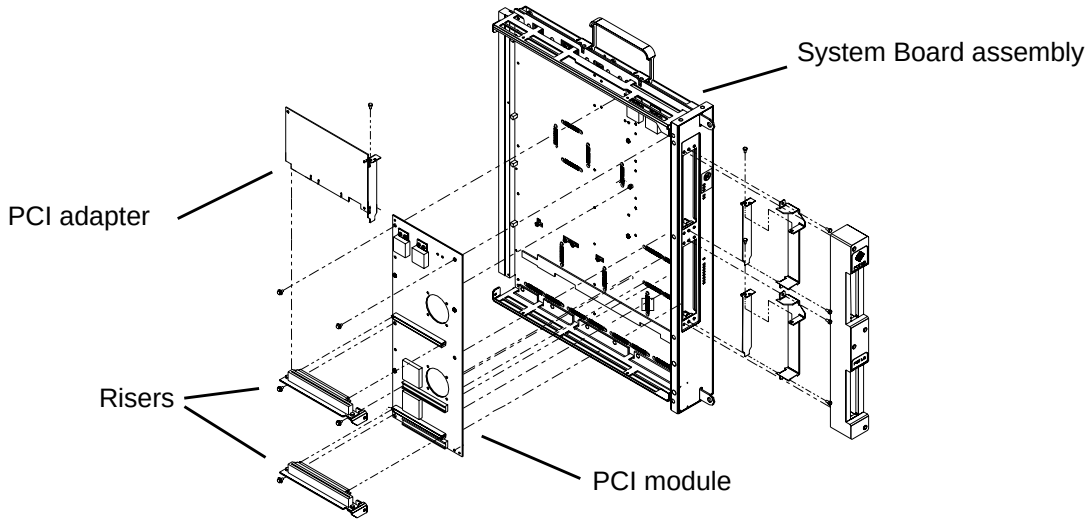

Figure 1. System board with PCI module

The “risers” allow the PCI adapters to be mounted in the same plane as the PCI module.

Because it is only possible to package two PCI adapters per system board (in contrast to four SBus adapters), PCI is not cost effective where there is not a performance requirement. Therefore the Enterprise 10000 system will remain basically as an SBus-based system with PCI available for selected uses. For instance, customers will prefer to use SBus for the interfaces detailed in the previous section (e.g. SCSI, Ethernet, fiber channel, FDDI, ATM, etc). PCI will be used for fast data transfer situations such as HIPPI and Sun Cluster interconnect. (Both of these are future capability). PCI will also be used for customer-supplied adapters.

1.4.7 Boot Bus

The Enterprise 10000 system boot bus is a byte-wide synchronous bus running at half the Enterprise 10000 system clock. Its primary purpose is to provide access to the SRAM that is local to the boot processor. Because the port controller is not directly on the Enterprise 10000 system data interconnect path, processor access to the boot bus data is through the data buffer. Because the port controller must provide JTAG access to this SRAM, however, the port controller is also on the boot

bus data bus. This interconnection is used for the secondary function of the boot bus, which is to provide processor access to control and status registers in the port controller, for which there is no normal data path.

The secondary use of the boot bus for port controller command and status register access requires that the boot bus, without the SRAM, be implemented on the port controllers/data buffers servicing the I/O ports, and an abbreviated version is implemented in the memory controller to enable access to its command and status registers through its associated data buffer.

There are four separate boot buses on each Enterprise 10000 system board; one for each PC, and one for the MC.

1.5 Control Board

The control board facilitates communication between the SSP and Enterprise 10000 processors, while controlling support subsystems such as power and cooling. Additionally, it monitors the health of the system using JTAG hardware located on each system board. The control board is a self-contained computer consisting of a SPARClite processor, serial interface, Ethernet controller, JTAG controller, 10BASE-T interface, and reset and control logic. It controls the system JTAG, clock, fan, system and I/O power, and Ethernet interface functions.

The control board contains a custom serial bus to control the fans and bulk power supplies. This same bus is used to remotely power on and off the peripheral cabinets via the five DB9 connectors, RPC0-RPC4, located on the front panel. The Ethernet controller provides the link between the SSP and the control board. The JTAG controller scans and controls the power to all of the Enterprise 10000 components. Reset and control logic performs various functions such as monitoring the inlet airstream of the ambient air, updating the heartbeat logic in a timely function to prevent the heartbeat logic from resetting the control processor, and writing to the power control rings on the system board and centerplane support board to select the active clock source.

The Enterprise 10000 system contains either one or two control boards. In the event of a control board failure, the other control board can take over full management of the system following a reboot of the system. See FIGURE 1-16 and FIGURE 1-17 for a control board illustration and block diagram.

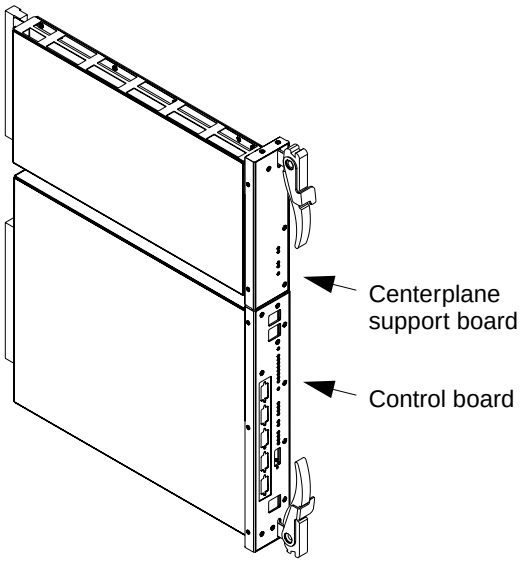

FIGURE 1-16 Centerplane Support Board and Control Board

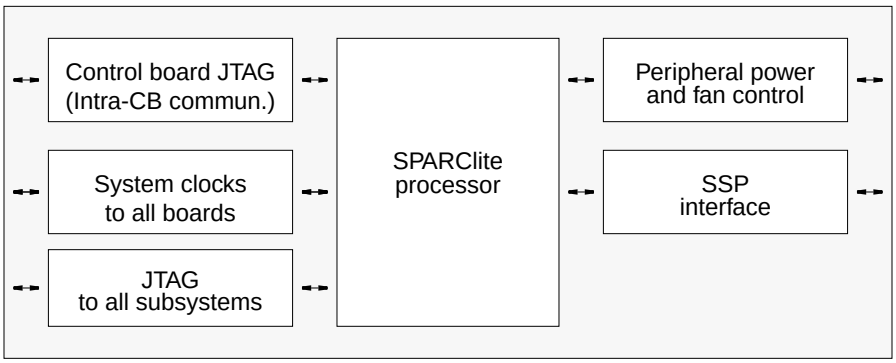

FIGURE 1-17 Control Board Block Diagram

FIGURE 1-18 and FIGURE 1-19 show the power, cooling, and control board functions fully integrated with each other. In particular, the major points to note are:

- The Enterprise 10000 system has fault-tolerant power and cooling.
- Power to all peripheral cabinets is individually controlled.
- All housekeeping functions are managed by the control board. For redundancy, a second control board can be configured.

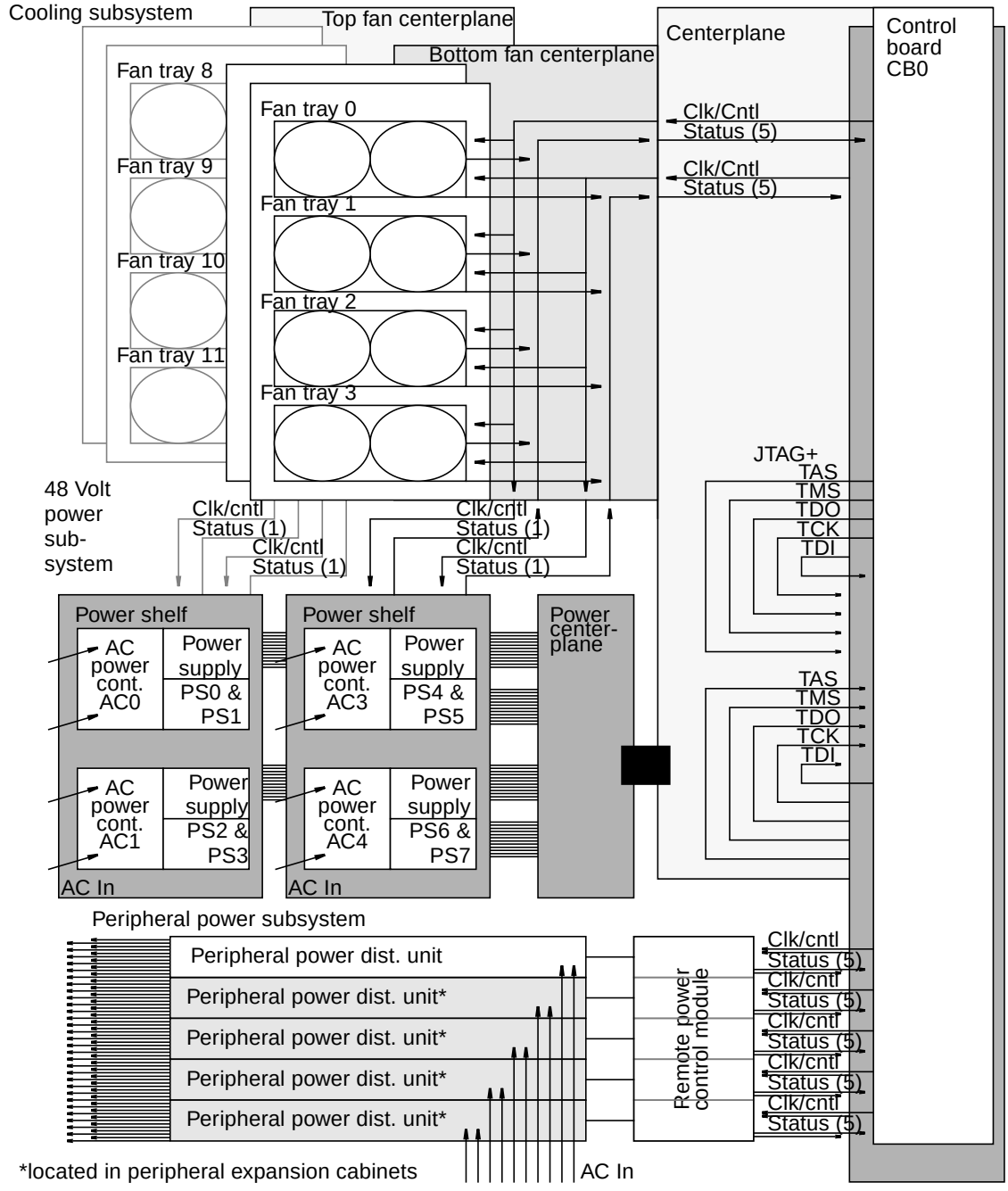

FIGURE 1-18 Control Block Diagram I

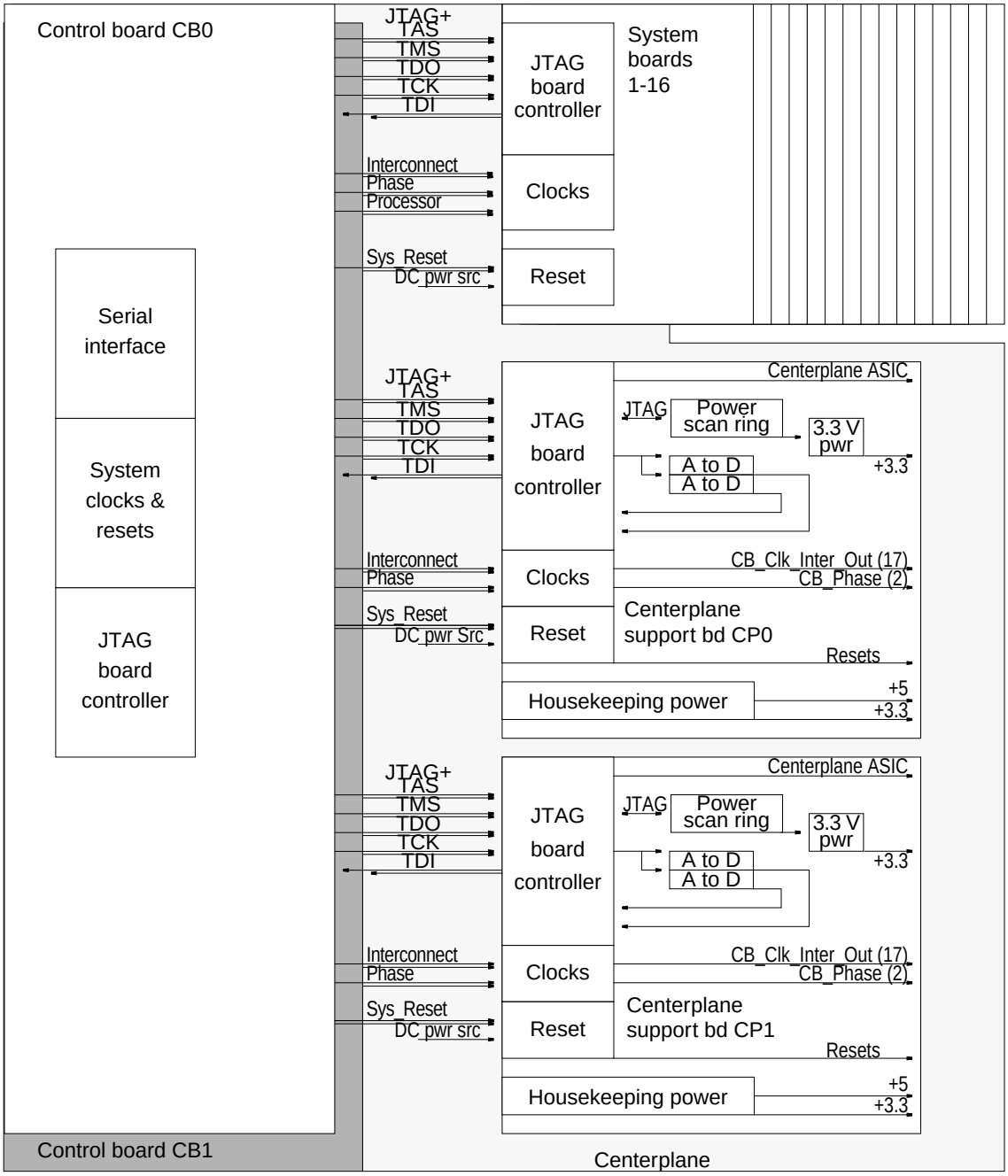

FIGURE 1-19 Control Block Diagram II

1.6 48-Volt Power Subsystem

The Enterprise 10000 power system is distributed and consists of three functional pieces as shown in FIGURE 1-20.

- Universal 220 VAC to 48 VDC front end power supplies
- 48-volt filter, distribution and protection box
- Centerplane and system logic boards.

A 48-volt distribution system was selected because of its safety. Distribution voltages of less than 60 volts are defined by the safety agencies as extra-low voltages (ELVs) because they pose no shock hazard.

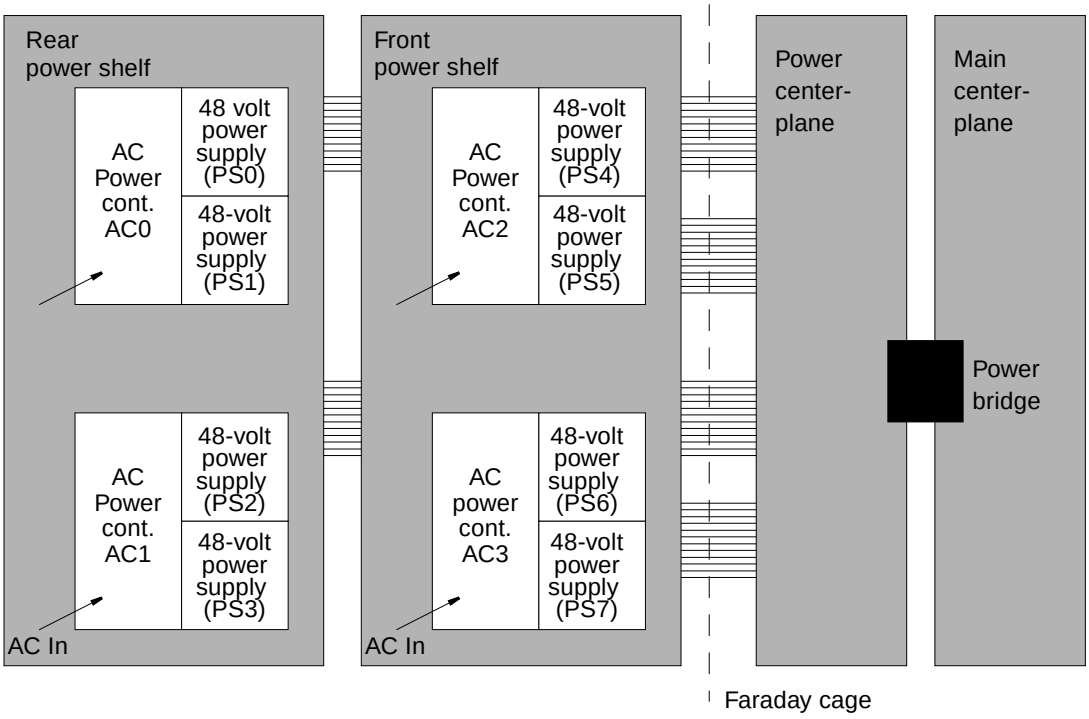

FIGURE 1-20 Enterprise 10000 Power System Block Diagram

The 48-volt distribution system supports up to twenty 750 watt loads, which includes 16 system boards, two control boards/centerplane support boards combinations and two groups of fan subsystems. It is designed so that no individual load is capable of keeping the system from operation.

To facilitate EMI filtering within the Enterprise 10000 Faraday cage while using leveraged off-the-shelf power supplies, the 48-volt power supplies are isolated from the system noise through an EMI filter. This EMI filter is placed on each load and is located at the point of 48-volt power distribution for the entire system.

The 48-volt power system converts the AC input power to DC voltages, which are distributed to the power centerplane. Its modular design makes troubleshooting easier while providing redundant power in the event of a failure.

The Enterprise 10000 power system is implemented as a distributed 48-volt front-end power system with DC/DC converters on each logic board to provide local DC voltages as necessary.

The 48-volt power supplies are rackmounted into the power shelf, which contains a line cord rated at 200 VAC at 24A. Each line cord provides power to two supplies, and up to four power shelves can be configured to provide redundant AC connections to the cabinet. By configuring a system with a redundant AC connection and two redundant power supplies, one AC connection can fail and be concurrently serviced during normal system operation.

Power is distributed from the 48-volt power supplies through two separate buses, with each bus supplying power to one side of the cabinet. These buses are tied together. Power is distributed individually to each circuit board through a DC circuit breaker and cable harness, and power supplies on each circuit board are also individually fused. This enables a 48-volt power failure on one circuit board to be isolated from the main system and serviced at a later time.

1.6.1 48-Volt Power Shelves

The 48-volt power supply system consists of up to two 48-volt power shelves. These drawers are self-contained, rackmounted, hot-pluggable units that are powered from a universal single-phase 220-volt power source (180-264 VAC @ 47-63 Hz). The power shelves are comprised of up to two AC input modules and up to four 48-volt power supplies. The power shelves supply 48 VDC power to the 48-volt power distribution system.

Each of the power shelves is capable of having multiple supply modules, which can be used in parallel to provide N+1 redundancy across the drawer and across multiple drawers when multiple drawers are used. When used in an N+1 redundant manner, each of the individual supplies is capable of being serviced without interrupting the 48-volt output of the N+1 configuration (TABLE 1-3).

TABLE 1-3 Configuring the Power Supplies

	Number of System Boards	Max System Power	Required Power Supplies for Redundancy	Required 200V, 30 A single phase circuits
Standard configuration	1	2,070	3	2
	2	2,815	3	2
	3	3,560	3	2
	4	4,305	4	2
	5	5,050	4	2
	6	5,795	5	3
	7	6,540	5	3
	8	7,285	5	3
Standard configuration with option power expansion	9	8,030	6	3
	10	8,775	6	3
	11	9,520	7	4
	12	10,265	7	4
	13	11,010	7	4
	14	11,755	8	4
	15	12,500	8	4
	16	13,245	8	4

1.6.2 AC Input Module

The AC input module receives the AC input from the outside and passes it to the power supplies. Additionally, the AC input module monitors and senses any disturbances and reports these to the control board.

On each AC input module, there are two yellow LEDs (for redundancy), one green LED, and one circuit breaker for each of two associated power supplies. The green LED indicates 48-volt DC output is all right. The yellow LEDs indicate that power is applied to the unit, and it is NOT safe to service. When the yellow LEDs are off, the corresponding supply is powered off and it is safe to service concurrently with normal system operation.

1.6.3 Power Supply

The 48-volt power is developed using multiple 1728W AC/DC modular power supplies. The outputs of these supplies are connected in parallel to a common 48-volt power bus, which distributes power to the rest of the system. Each supply current shares with the others, and redundant supplies can be configured to provide concurrent service capability for the 48-volt power subsystem.

When a system is configured with N+1 redundant power supplies, a supply can be serviced concurrently with normal system operation by powering it off (if it has not already been powered off by the SSP as a result of the failure), removing it, and replacing it with a new supply. The new supply can then be powered on without disturbing the running system.

1.6.4 Circuit Breaker Panel

The distribution system uses individual circuit breakers to protect system components from potential faults associated with the insertion of new and potentially shorted loads. Separate circuit breakers protect each of the twenty loads so that no individual load is capable of keeping the system from operation.

1.7 19-Inch Rackmount Power Distribution Subsystem

The rackmount power distribution subsystem provides power to all of the I/O peripherals either in the processor cabinet or in an I/O expansion cabinet. It is comprised of AC sequencers and remote power control modules.

1.7.1 AC Sequencer and Remote Power Control Module

AC power is received from the outside by the AC sequencer. The AC sequencer then provides AC power to the peripheral trays (disk or tapes) via twelve 10A accessory outlets. Various remote configurations are supported to assure maximum redundancy (FIGURE 1-21). The AC sequencers can be controlled remotely by the SSP by connecting the control board to a remote power control module, which in turn is connected to the AC sequencer.

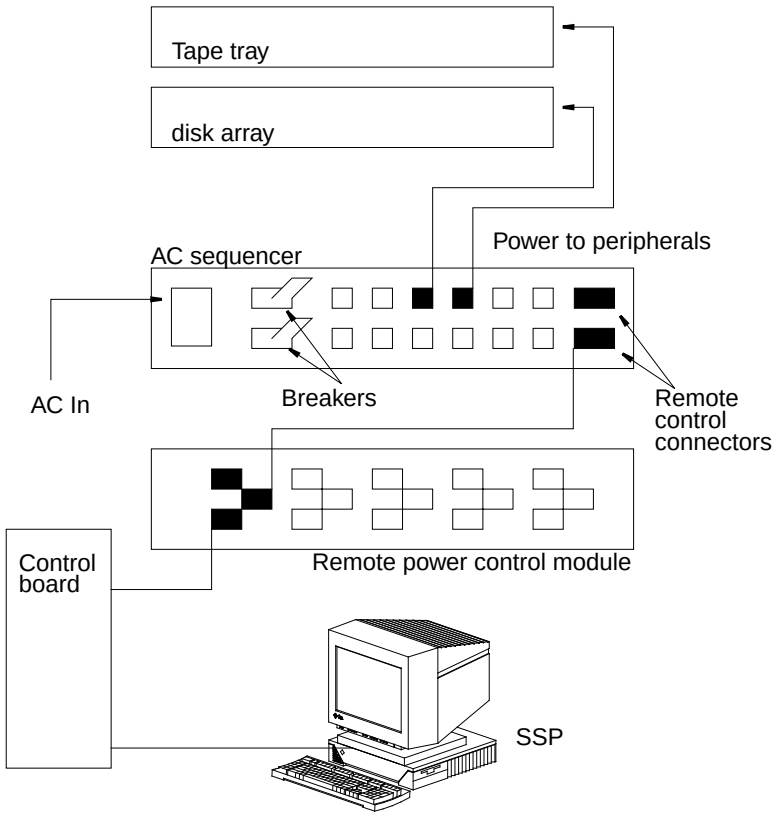

FIGURE 1-21 Power Distribution and Remote Power Control Configuration

1.7.2 Power Configurations

FIGURE 1-22 illustrates the use of a redundant control board. In this configuration, either control board can control the power to the trays. Additionally, multiple AC sequencers can be daisy chained together enabling the SSP to control up to five AC sequencers.

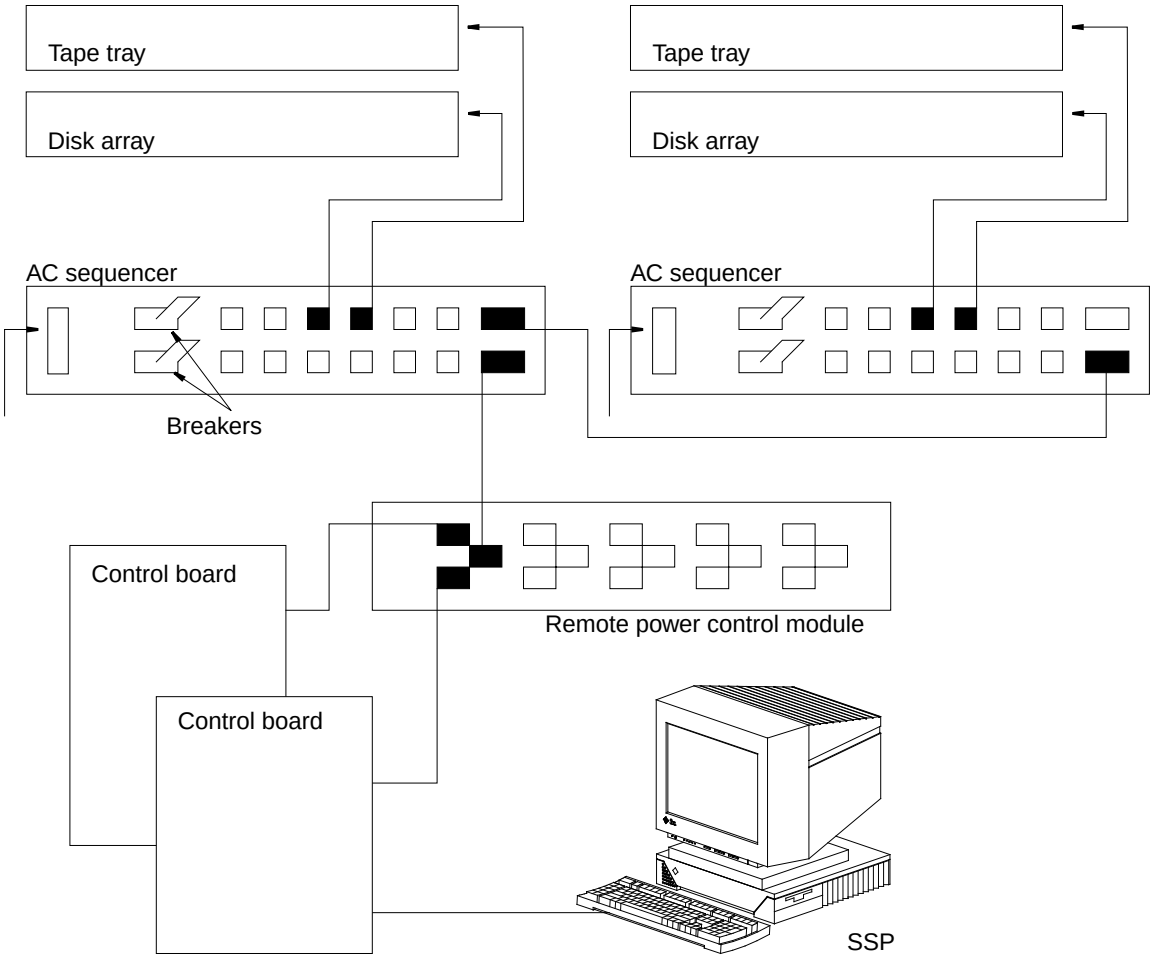

FIGURE 1-22 Remote Power Control Configuration

1.8 Cooling Subsystem

The processor cabinet is cooled by the flow of pressurized air drawn into the bottom of the cabinet from the ambient environment, and exhausted out the top to the outside environment. The flow of cooled air is produced by a redundant system consisting of 16 fan trays operating through four circuit breakers to alleviate a source of single point failure (SPF). Each fan tray contains two fans and, in the event of a failure, can be removed while the system is operating (FIGURE 1-23).

For cooling system monitoring purposes, two types of sensors are mounted onto the bottom of each system board; a temperature sensor and an airflow sensor. The air flow sensor detects that the blowers are functioning (regardless of temperature) and the temperature sensor verifies that the ambient air is within the specified range. The monitoring system is such that a failure of either mode will notify the user and enable orderly resolution or shut down prior to loss of data.

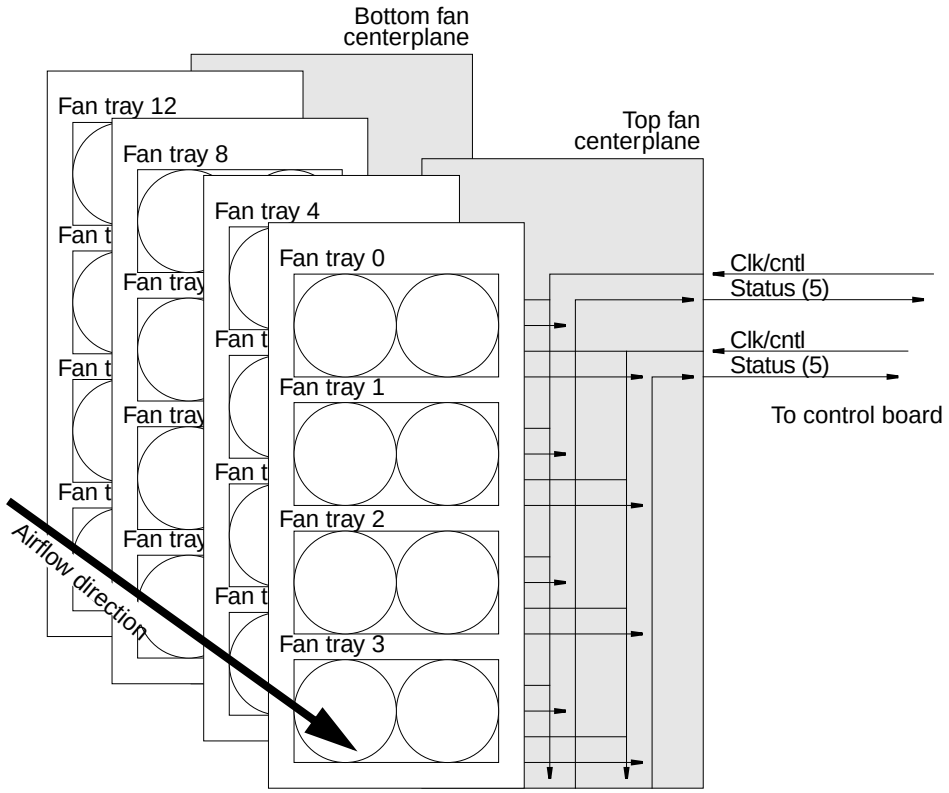

FIGURE 1-23 Cooling Subsystem

CHAPTER 2

System Service Processor

The system service processor (SSP) is the access port to the Enterprise 10000 system for system administration and service providers. A key point is that it provides a level of intelligence physically separate from the Enterprise 10000 system. The SSP sequences the system boot process, assists in configuration decisions, monitors environmental conditions, assists in the auto-reboot process following a system interruption, and sends information to the service provider about any abnormal conditions. The SSP also is the controlling mechanism for the Enterprise 10000 domains.

The SSP is built from a SPARCstation™5 with 64 Mbytes of RAM and a 2.1-Gbyte disk. A CD-ROM is included for loading software onto the Enterprise 10000 system. The SSP runs standard Solaris environment plus the following software:

- Hostview

This is a graphical user interface (GUI) that assists the system administrator with management of the Enterprise 10000 system hardware.

- Power-on Self-test (POST)

This software performs diagnosis on the hardware and configuration management based on the diagnosis.

- OpenBoot™ PROM (OBP)

OBP reads a structure created by POST and constructs a device tree for use by the operating system.

- Network console (netcon)

The netcon software enables the system administrator to access the Enterprise 10000 system console remotely, from anywhere where there is a network connection, via an X-Window interface.

2.1 Connection to the Enterprise 10000 System

The SSP is connected via Ethernet to the Enterprise 10000 system control board. The control board has an embedded control processor that interprets the TCP/IP Ethernet traffic and converts it to JTAG control information. (JTAG stands for Joint Test Action Group, the group that developed the IEEE 1149.1 serial scan interface standard).

JTAG is a serial-scan protocol that is supported by many ASIC suppliers. It enables a wealth of opportunity with concurrent diagnostics and configuration control. It also enables interconnect tests which can replace a bed-of-nails tester in manufacturing.

You can interact with the Enterprise 10000 system by using the Hostview GUI that runs on the SSP.

2.2 Network Connections

The Enterprise 10000 system requires 10BASE-T or 100BASE-T Ethernet connections on the customer network for an SSP and each host domain.

Additional Ethernet connections are required if any of the following Enterprise 10000 options are ordered.

- Optional redundant control board
- Optional redundant SSP
- Additional domains are used
- Alternate pathing (AP) of Enterprise 10000 network connections will be configured

To prevent general purpose Ethernet traffic from negatively affecting the SSP to Enterprise 10000 host communication, the following configuration rules should be observed:

- Connect the SSP and control boards via a private 10BASE-T network (separate subnets). This will connect the one (or two) SSPs with the one (or two) control boards.
- Connect the SSP and each of the host domains via a second network. To facilitate net booting a domain from the SSP, the network between the domain and the SSP must be either 10BASE-T or 100BASE-T Ethernet.

FIGURE 2-1, FIGURE 2-2, and FIGURE 2-3 illustrate three possible network configurations. These configurations use the hubs that are packaged in manufacturing with the control boards and an SSP that is configured in manufacturing with an additional quad Ethernet card. Customer networks and their hubs, however, are site dependent and, therefore, not supplied by Sun Microsystems.

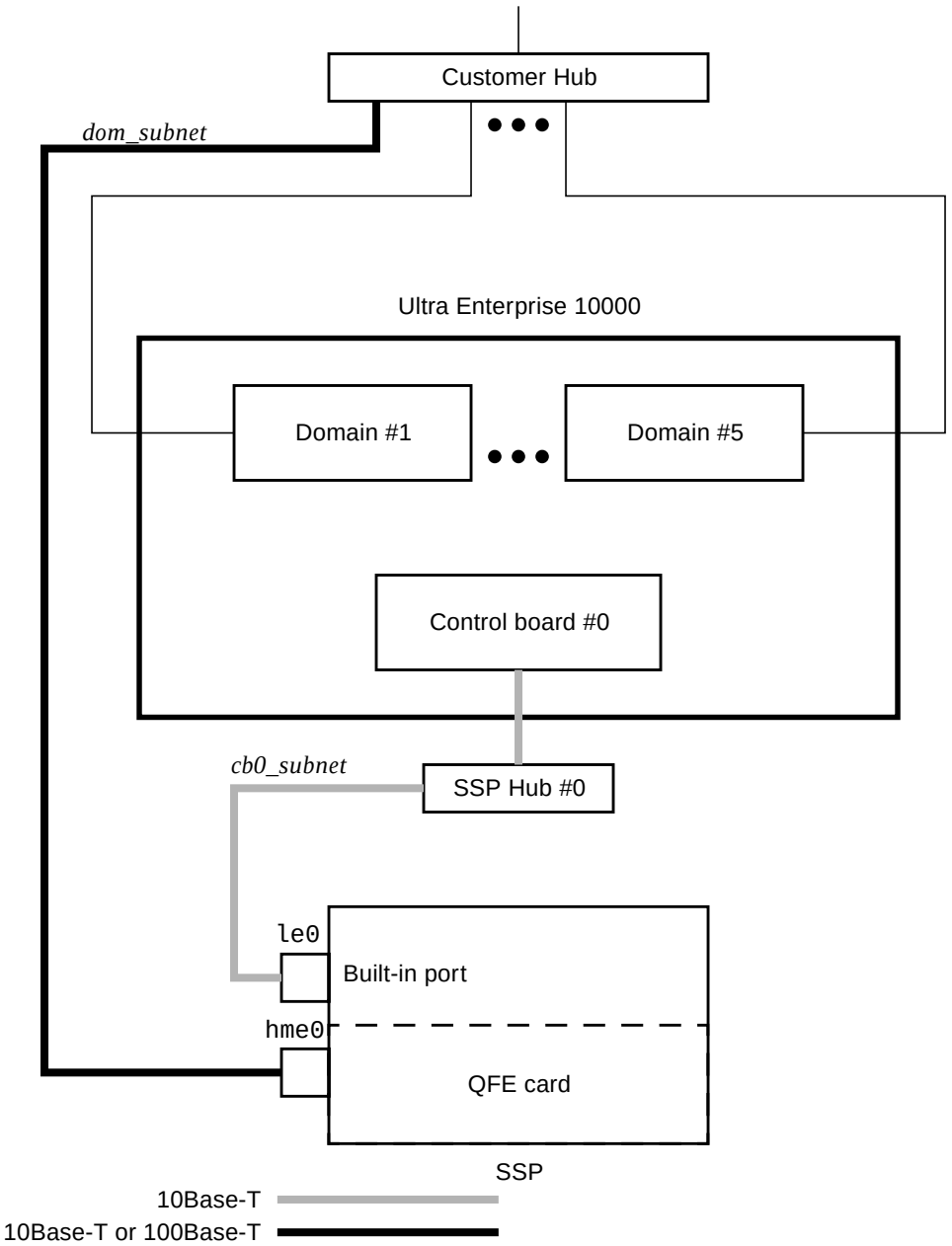

FIGURE 2-1 Network Configuration (Base Case)

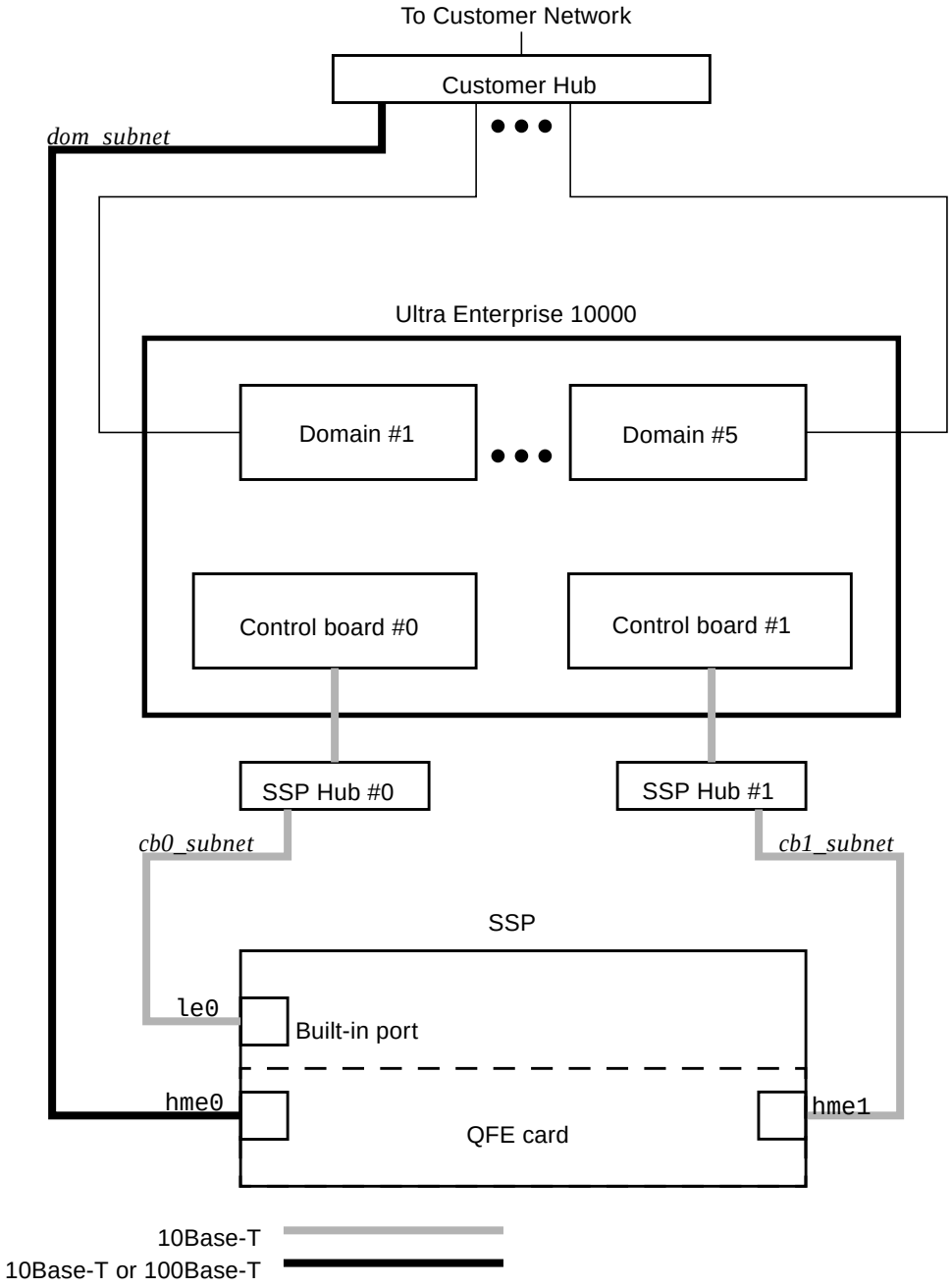

FIGURE 2-2 Network Configuration with Redundant Control Boards

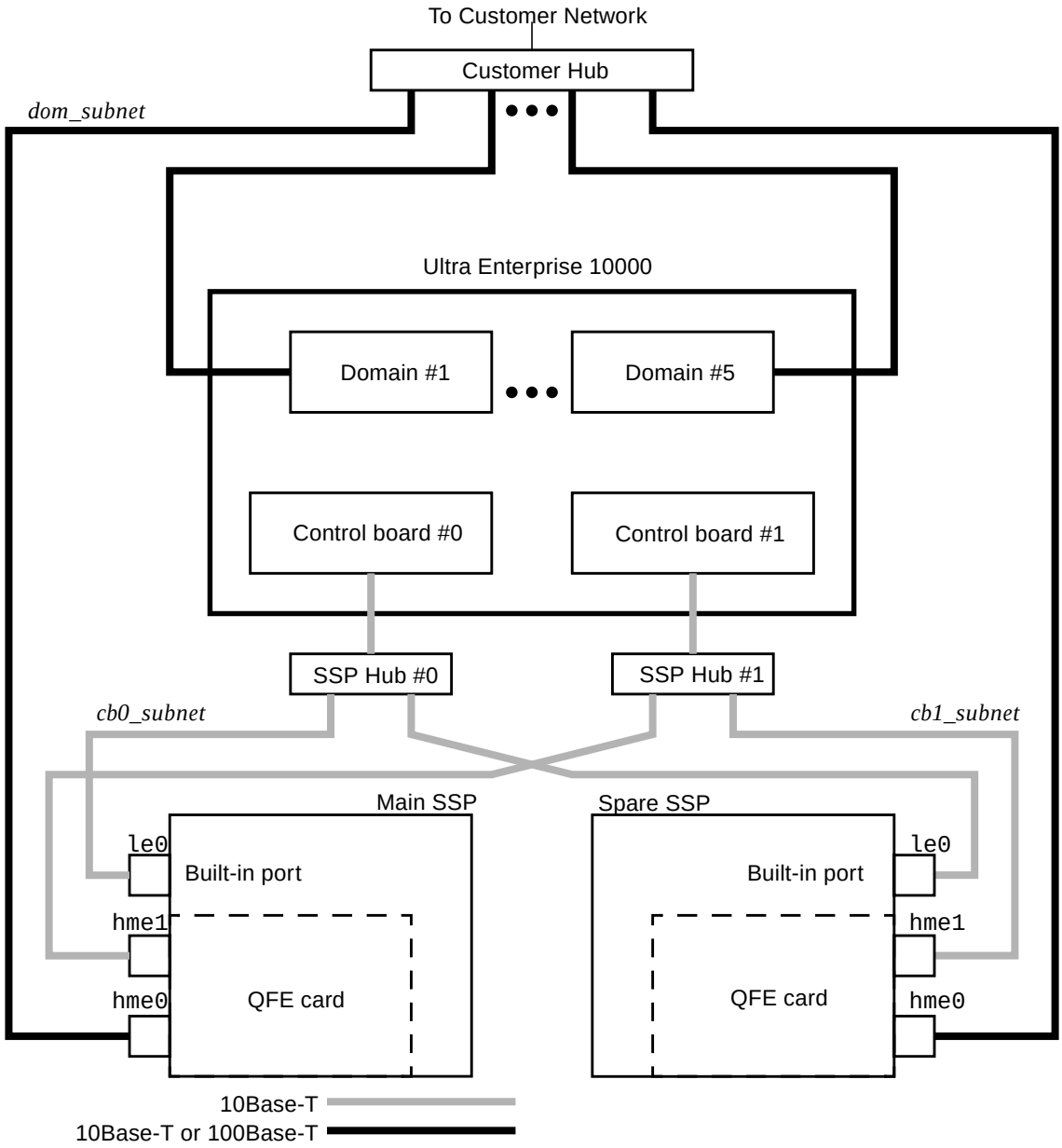

FIGURE 2-3 Network Configuration with Redundant Control Board and a Spare SSP

2.3 Hostview

Hostview enables you to perform the following actions:

- Power the Enterprise 10000 system on and off.
- Power peripherals and I/O expansion cabinets on and off.
- Dynamically group system boards into domains. Each domain can carry its own workload and has its own log messages file.
- Boot the operating system on a domain.
- Dynamically reconfigure the boards within the Enterprise 10000 system, logically attaching or detaching them from the operating system, resetting them, and running diagnostics on them.
- Start a netcon console window for each domain.
- Access the SSP log messages file for each domain.
- Remotely log in to each domain.
- Edit the blacklist file to enable or disable hardware components on each domain.
- Monitor the Enterprise 10000 system status.

2.4 Network Console

Network Console (netcon) provides a *console* for single-user operations without requiring an asynchronous dedicated connection between the Enterprise 10000 control board and the SSP. The service affords sessions, similar to `rlogin` sessions, that can be provided to X-Windows clients on the same network as the SSP. This enables system administrators to access the SSP from any location on the same subnetwork as the SSP.

Netcon end-user sessions link to the SSP, which funnels input to, and output from, OpenBoot PROM (OBP) and Solaris. This duality is provided in a seamless fashion to emulate the behavior of a serial line, where OBP and Solaris share a single physical nonmultiplexed medium. Multiple end-user sessions can be started simultaneously, enabling administrators to observe console access in real-time. This same feature enables service providers to log in and provide service without disruption to the customer console session.

2.5 Configuration Control

The SSP provides two forms of configuration control: user control, which is provided through a *blacklist* concept, and recovery control, which is provided automatically through a *figure of merit* concept but with predefined user control of the decision process through a boot heuristic and a GUI.

2.5.1 Blacklist

The blacklist is a file for user modification. It can be modified graphically by using supplied GUIs, or edited directly with an ASCII file editor. The system does not modify the blacklist file automatically under any circumstances. The blacklist contains the names of parts not to be used in the next configuration established.

2.5.2 Figure of Merit

The SSP determines the best configuration by assigning a figure of merit to each technically feasible configuration, and then choosing the configuration with the highest figure of merit. The figure of merit is a number that can be affected by giving different weighting factors to different architectural features.

CHAPTER 3

Dynamic System Domains

The dynamic system domain feature enables one Enterprise 10000 system to be logically broken down into up to 8 smaller systems. Three examples of the benefits of this are:

1. To simultaneously and safely run a production environment on one part of the system while running one or more development environments within the same system cabinet.
2. To divide the system up so that different functional organizations can have their own smaller Enterprise 10000 system matched to the task to be performed.
3. To dynamically tier data on the same server for use in data warehouse applications.

Each domain has its own instance of Solaris as well as its own hostid, boot disk, network connection, memory, and disk space. The Solaris license issued with the Enterprise 10000 system covers all copies of Solaris running in the different domains. Software errors in a domain are confined to that domain and will not affect the rest of the Enterprise 10000 system. Administration of each domain is done from one SSP that services all the domains.

FIGURE 3-1 shows Domain 1 running an early version of Solaris 2.7. This is a four-processor domain with a boot disk and a network connection. It uses system board 4. Domain 2 is a eight-processor domain using system boards 9 and 10 (system boards do not have to be adjacent to be in the same domain). This domain is equipped with a disk array and an Ethernet connection. It could be running a test version of Oracle under Solaris 2.6. The rest of the Enterprise 10000 system is a nine-system board (36 processors) domain that is used for production work.

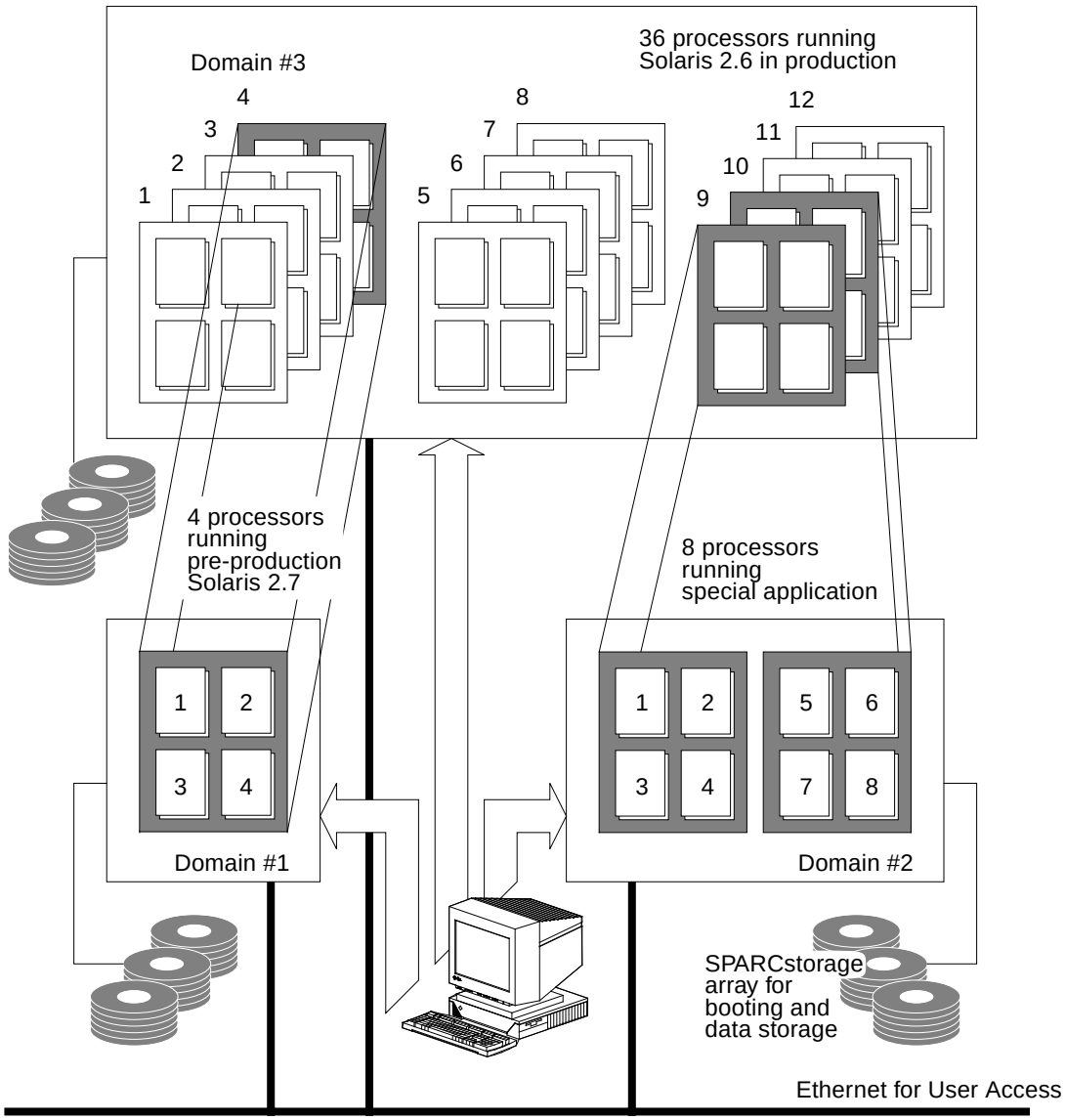

FIGURE 3-1 Example of Enterprise 10000 Domains

3.1 Properties of Domains

Boards in the same system domain:

- Have a common physical address space.
- Can access each other's noncacheable (command and status) address space.
- Execute the same copy of the Solaris operating system.

Domains in the same domain group:

- Can re-use physical addresses except for any interdomain shared-memory ranges.
- See each other's arbstops, holds, interrupt NACKs, and aborts. This means that if one domain in a group has an arbstop, all other domains within that group will also be arbstopped.
- See each other's interrupt messages.
- Can export memory to another domain within the domain group.

Domain groups are composed of one or more system domains. A single system domain is, thus, also a minimal domain group.

3.2 Management of Domains

By issuing dynamic reconfiguration commands at the SSP, the administrator can specify that one or more system boards be assigned to a domain. This takes place without interruptions to the users of the Enterprise 10000 system. Then Solaris can be booted up on the domain, and it is available for immediate use. Each domain has its own hostid, and appears to be a complete system on its own. The SSP remains as the system console for each of the domains.

Each domain must be configured with a disk to boot from, a network connection, and sufficient memory and disk space to accomplish the intended task. For the Enterprise 10000 system, the domain size is limited only by the number of system boards configured.

3.3 Domain Filtering

Global and local filtering (in two stages) is performed via the system board and centerplane to determine if a board is in a domain or domain group.

3.3.1 Global Filtering

The global arbiters do the first-level filtering that identifies all the boards in a domain group. Each global arbiter contains a 16 x 16 bit set of Shared Memory Domain Control Registers as shown in FIGURE 3-2. There is a register for each system board that, when the bits are set to one, establish the set of system boards in the domain group of that particular board.

The global address arbiter sends a signal to all the memory controllers in a domain group, since the memory controllers only pay attention to addresses with this signal asserted. Holds, arbstops, and recordstops are also propagated to all the boards in a domain group. The shared memory domain control registers are readable and writable only via the JTAGs.

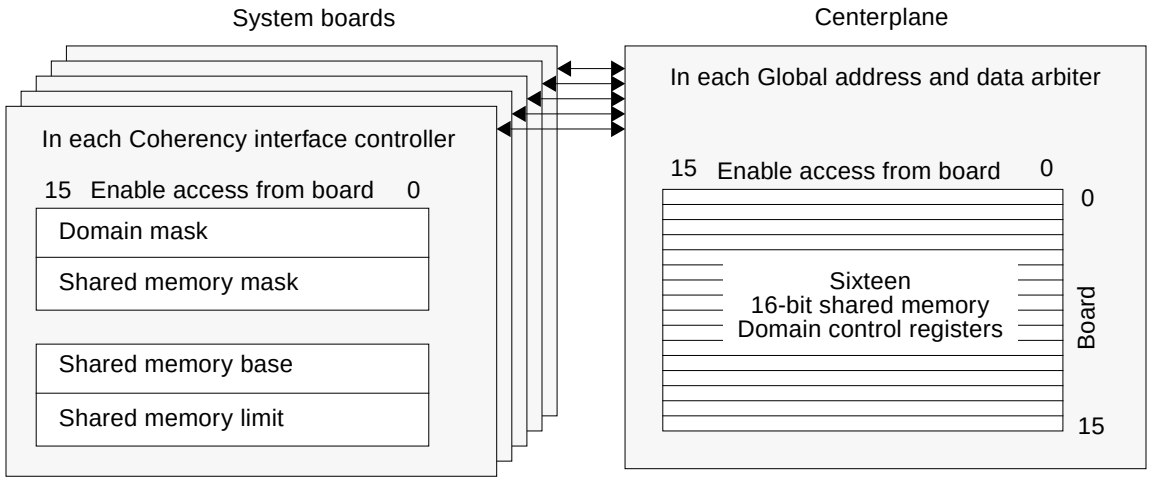

FIGURE 3-2 Address Filtering Hardware for Domains

For example, suppose system boards 2, 4, 8, and 10 are in one domain group, and that system boards 5, 12, 13, and 14 are in a second domain group. The shared memory domain control register settings are shown in FIGURE 3-3. A board must always list itself in its own domain group to obtain centerplane arbitration services. At times it is desirable to remove the listing of a system board from its own register to deconfigure the system board for removal.

After a bus transaction request is granted, the global address arbiter examines the source board domain group register. Only boards within the domain group will receive a valid signal. If a valid signal is received, the board will be able to look at the bus information and further decode it.

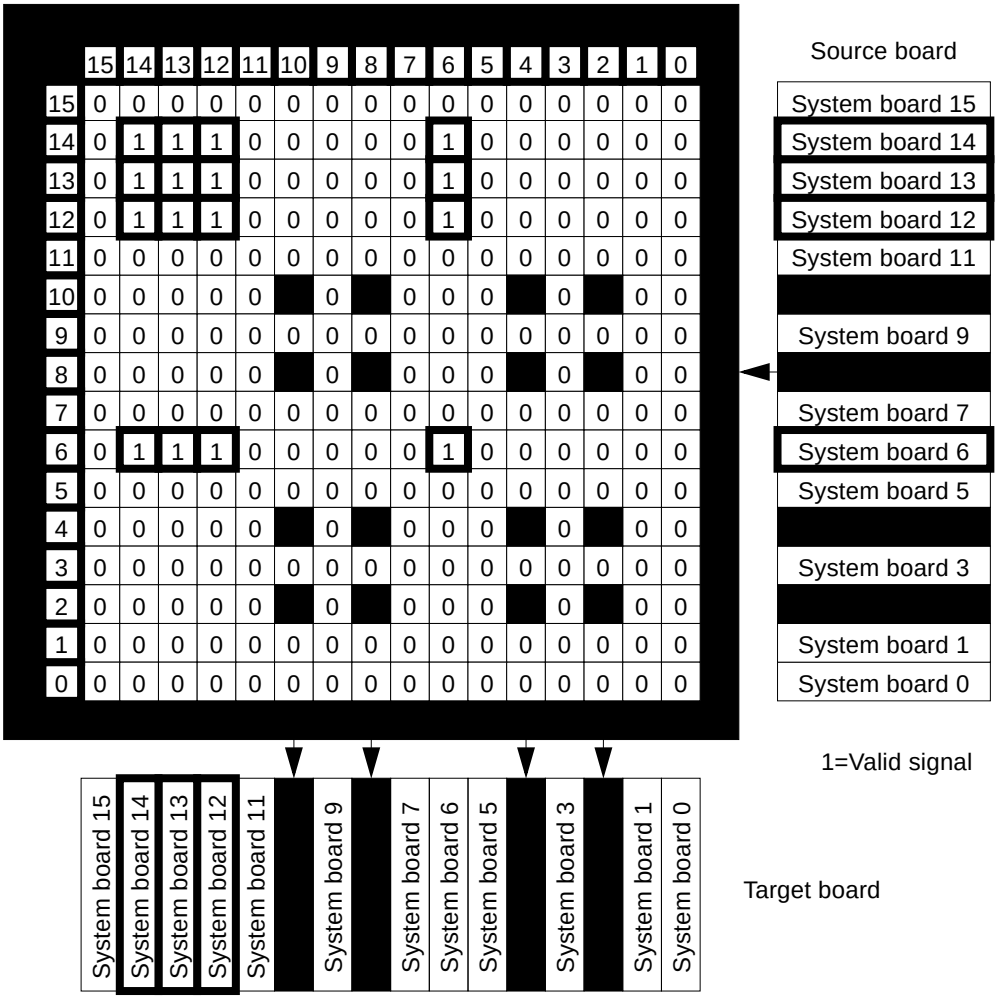

FIGURE 3-3 Shared Memory Domain Control Register

3.3.2 Local Filtering

Local filtering removes interdomain addresses that are outside the shared-memory ranges of a domain group. All four coherency interface controllers on a system board have identical copies of the following registers:

- Shared memory mask. Sixteen bits identify what other system boards are in this board's domain group. Only coherent requests from these system boards, whose addresses are within the shared-memory base and limit, must be processed. The shared memory mask register is readable and writable only via the JTAGs.
- Shared memory base and limit registers. These registers contain the high 25 bits [40:16] of the lower and upper physical address of the memory on this board that are visible to other domains of a domain group. The granularity of these addresses is 64 Kbytes. The shared memory base and limit registers are readable and writable via the control and status address space.
- Domain mask. Sixteen bits identify what other system boards are in this board's system domain. All coherent and noncoherent requests from these boards must be processed. The domain mask register is readable and writable only via the JTAGs.

The memory controller processes the address stream in parallel with the coherency interface controller (FIGURE 3-4). The coherency interface controller must send an abort to its memory controller for all addresses from boards identified by the shared memory mask that are outside the range of the shared memory base and limit registers.

To separate domains and their operating systems, address information containing a physical address, transaction type, and the source board number is received by the coherency interface controller and filtered locally during a bus transaction. The coherency interface controller compares the information against its internal registers to determine if the information came from a board within its domain. If it did, the transaction will be honored. If not, the transaction will be ignored.

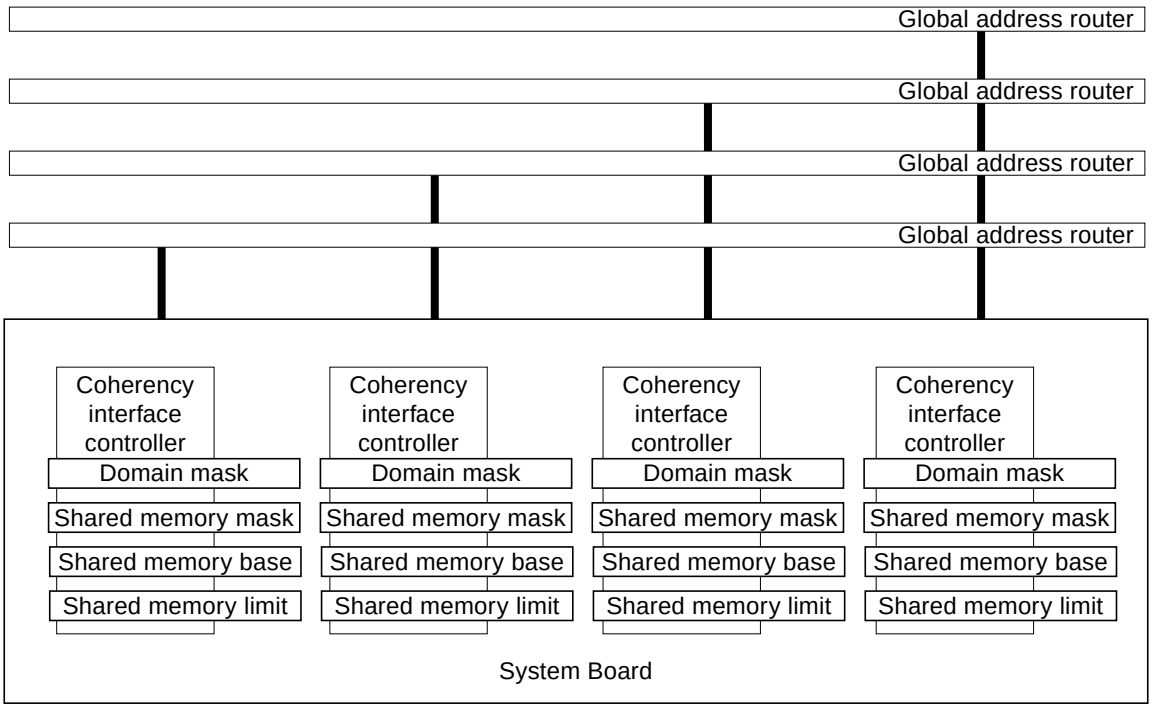

FIGURE 3-4 Registers that Control Domains

CHAPTER 4

Reliability, Availability, and Serviceability (RAS)

On the Enterprise 10000 system, SunTrust represents the RAS features reliability, availability, and serviceability, which are the interrelated components that translate to system up-time. These features are described as:

- *Reliability* is a function of the care with which the hardware and software design was executed, the quality of the components selected, and the quality of the manufacturing process (for example, ESD protection, clean rooms, and so forth).
- *Availability*, is the percentage of time the customer's system is able to do productive work for them.
- *Serviceability* of a system ensures that repair time (down time) is kept at a minimum.

4.1 4.1Reliability

The Enterprise 10000 system reliability features fall into three categories:

1. Features that eliminate or reduce the probability of errors.
2. Features that correct errors on the fly, and thus avoid any availability reduction.
3. Features that detect uncorrected errors so that data integrity is not compromised even if a system failure ultimately requires a Solaris automatic reboot.

4.1.1 Reducing the Probability of Error

All ASICs are designed for worst case temperature, voltage, frequency, and airflow combinations. The high level of logic integration in the ASICs reduces component and interconnect count.

A distributed power system improves power supply performance and reliability.

Extensive self-test upon power-on reboot after a hardware failure screens all of the key logic blocks in the Enterprise 10000 system:

- Built-in self-test logic in all the ASICs applies pseudo-random patterns at system clock rate providing at least 80% single-stuck-at-fault coverage of combinatorial logic.
- The Power-On Self-Test-controlled from the SSP-tests each logic block first in isolation, then with progressively more and more of the system. Failing components are electrically isolated from the centerplane. The result is that the system is booted only with logic blocks that have passed this self-test and that should operate without error.

The memory subsystem continues to operate even if an entire DRAM chip fails. The DRAM chips are four bits wide, with the bits distributed across four different words, each of which is protected by separate error-correcting-code bits. Thus, a complete DRAM failure appears as single bit errors in four separate words, which are all correctable.

All I/O cables have a positive lock mechanism and a strain relief support.

4.1.2 Correcting Errors on the Fly

The Enterprise 10000 system contains a number of subsystems that are capable of recovering from errors without failing. Subsystems that have a large number of connections have greater odds of failure. The Enterprise 10000 subsystems that have the highest probability of errors are protected from transient errors through the use of on-the-fly single-bit error correction using an error-correcting-code.

4.1.2.1 Error-Correcting-Code Protection of the Data Interconnect

The entire data path from the UltraSPARC data buffer, through the Enterprise 10000 data buffers, local data routers, global data router, and the memory subsystem is error-correcting-code protected. Single-bit-data errors detected in these subsystems are corrected by the receiving UltraSPARC module, and the system is notified for logging purposes that an error has occurred.

The error-correcting-code protection is designed such that the failure of an entire DRAM chip will result in correctable, single-bit errors.

The error-correcting-code is generated by the data sourced and checked and, if necessary, corrected by the UPA port receiving the data. The memory subsystem does not check or correct errors, but just provides the extra storage bits. The Enterprise 10000 data buffer chips utilize the error correcting codes to assist in fault isolation.

If a correctable error is detected by the interconnect, the SSP is notified and enough information is saved to isolate the failure to a single net within the interconnect system. The data containing the error is sent through the interconnect unchanged to the destination UPA port where the error is reported to software.

Memory errors are logged by software so that DIMMs can be identified and replaced during scheduled maintenance.

4.1.2.2 Using Error Correcting Code in the Address Interconnect

The address interconnect is error-correcting-code protected from the sending coherency interface controller across a global address router to other coherency interface controllers and memory controllers. These errors are corrected dynamically, and the error information is captured by the coherency interface controller(s). The SSP will subsequently be notified of the error. The local address bus, which connects the UltraSPARC processor to the port controller, is parity protected, as is the path from a port controller to a coherency interface controller. These are short paths on individual boards and parity protection is considered sufficient at this level.

4.1.3 Detecting Uncorrectable Errors

Almost all internal system paths are protected by some form of redundant check mechanism. Transmission of bad data is thus detected, preventing propagation of bad data without notification. All uncorrectable errors will result in an arbstop condition. The recovery requires a Solaris automatic reboot.

4.1.3.1 Multiple-Bit Error-Correcting-Code Data Errors

Multiple-bit error-correcting-code errors are detected by the receiving UPA port, which notifies the operating system, so that depending upon what process is affected, the system as a whole can avoid failure.

Parity errors on external cache reads to the interconnect become multibit error-correcting-code data errors, and so are handled as other multibit errors.

4.1.3.2 Processor External Cache Read Parity Error

An external cache parity error on a processor read will cause a trap in that processor, for potential retry of the operation.

4.1.3.3 Multiple-Bit Error-Correcting-Code Address Errors

Multiple-bit error-correcting-code errors detected in the address interconnect are unrecoverable; therefore, fatal to the operating system, and will cause an arbstop condition.

4.1.3.4 Fatal Parity Errors

Parity errors in the following Enterprise 10000 subsystems are fatal and will cause an arbstop condition:

- All control signals within the Enterprise 10000 system
- The UPA address bus between an UltraSPARC processor and its port controller n The EData bus on the UltraSPARC processor module between the external cache, the UltraSPARC processor, and the UltraSPARC data buffer
- The point-to-point address connections between the port controllers and the coherency interface controllers on a given system board
- The address and data of the duplicate tag SRAM connected to each coherency interface controller
- The secondary cache tag SRAMs for each UltraSPARC processor

4.1.3.5 Paranoid Errors

The Enterprise 10000 ASICs have paranoid logic that checks for anomalous conditions indicating an error has occurred (such as queue overflows, invalid internal states, and missing events) rather than let the error propagate and become corrupted data or access time-outs that can be difficult to correlate with the actual failure

4.1.3.6 System Time-Out Errors

Time-out errors detected by the port controller or memory controller are an indication of lost transactions. Time-outs are, therefore, always unrecoverable and will cause an arbstop condition.

4.1.3.7 Power Corrected Failures

The Enterprise 10000 system uses a highly reliable distributed power system. Each system, control, or centerplane support board within the system has DC to DC converters for that board only, with multiple converters for each voltage. When a DC-to-DC converter fails, the SSP is notified. The system board reporting the failure will then be deconfigured from the system. No guarantee is made regarding continued system operation at the time of the failure.

4.1.4 Environmental Sensing

The system cabinet environment is monitored for key measures of system stability, such as temperature, airflow, and power supply performance. The SSP is constantly monitoring the system environmental sensors in order to have enough advance warning of a potential condition so that the machine can be brought gracefully to a halt-avoiding physical damage to the system and possible corruption of data.

4.1.4.1 Temperature

The internal temperature of the system is monitored at key locations as a fail-safe mechanism. Based on temperature readings, the system can notify the administrator of a potential problem, begin an orderly shutdown, or power the system off immediately.

4.1.4.2 Power Subsystem

Additional sensing is performed by the Enterprise 10000 system to enhance the reliability of the system by enabling constant *health* checks. DC voltages are monitored at key points within the Enterprise 10000 system. DC current from each power supply is monitored and reported to the SSP.

The reset signals in the Enterprise 10000 system are sequenced with the DC power levels to guarantee stability of voltage throughout the cabinet prior to removing reset and enabling normal operation of any of the Enterprise 10000 system logic.

4.2 Availability

For organizations whose goal it is to make information instantly available to users across the enterprise, high levels of availability are essential. This is especially true for a large shared resource system such as the Enterprise 10000 system. In fact, the larger the system and user base, the more important high availability becomes. This is because as clients and users are added, the consequences of down time increase.

4.2.1 Integrity and Availability Goals for the Enterprise 10000 System

The RAS goals for the Enterprise 10000 system were, first, to protect the integrity of the customer's data, and, second, to maximize availability. Engineering efforts focused on three areas:

- Problem detection and isolation-knowing what went wrong and ensuring the problem is not propagated
- Tolerance and recovery-absorbing abnormal system behavior and fixing it, or dynamically circumventing it
- Redundancy-replicating critical links

To ensure data integrity at the hardware level, all data and control buses are protected by parity checks right out to the data on the disks. These checks ensure that errors are contained.

For tolerance to errors, resilience features are designed into the Enterprise 10000 system to ensure the system continues to operate, even in a degraded mode. Being a symmetrical multiprocessing system, the Enterprise 10000 system can function with one or more processors disabled. In recovering from a problem, the system is checked deterministically but quickly (to ensure minimum down time). The system can be configured with redundant hardware to assist in this process. And to minimize the repair time, major components of the Enterprise 10000 system can be exchanged while the system is online and in use.

4.2.2 High Availability Features of the Enterprise 10000 System

This section discusses the Enterprise 10000 system product features that together raise the availability of the Enterprise 10000 system from the normal commercial category to the *high availability* category. These features are best grouped as follows:

- Some specific areas of the Enterprise 10000 system are *fault tolerant*, meaning that any one single point of failure is totally transparent to users. Users see no loss of performance or capability.
- *Resiliency features* are built into the Enterprise 10000 system. These are features that enable processing and access to data to continue in spite of a failure, possibly with reduced resources. Use of these resiliency features usually requires a reboot, and this is counted as repair time in the availability equation.
- *Serviceability features*. Failures do happen but the serviceability features lower or eliminate the repair time.

4.2.2.1 Cooling

The Enterprise 10000 system is air-cooled with fan trays, each of which consists of two fans. Should one fan fail, the remaining fan will automatically increase its speed, thereby enabling the system to continue to operate-even at the maximum specified ambient. Therefore, operation need not be suspended while a fan is being ordered. Also, fan replacement can be done while the system is operating, again assuring no adverse impact on the availability metric. Of course, the Enterprise 10000 system has comprehensive and fail-safe temperature monitoring to ensure there is no over temperature stressing of components in the event of a cooling failure.

4.2.2.2 AC Power Supply

AC power is supplied to the Enterprise 10000 system through up to four independent 30 ampere single-phase plugs. Each AC plug carries power to a pair of 2,000-watt bulk power supplies. Only as many bulk supplies as are needed for the particular system configuration need be configured, plus an optional extra supply for N+1 redundancy. The standard redundant configurations are five power supplies for up to eight system boards and eight power supplies for a maximum configuration.

The AC connections should be controlled by separate customer circuit breakers, and can be on isolated power grids if a high level of availability is required. Optionally, third-party battery backed up power can be used to provide AC power in the event of utility failure.

4.2.2.3 ECC

On the Enterprise 10000 system, data errors are detected, corrected, and/or reported by the data buffer on behalf of its associated processor. Additionally, data errors passing through the interconnection will be detected and cause a record stop condition for those ASICs that can detect and initiate this condition. These history buffers and record stop condition bits can then be read via JTAG and used by offline diagnostics.

Data errors can be generated on an access from either memory, I/O, or from another processor (in a processor-to-processor interrupt). When a data error is detected by the data buffer, if it is correctable, the data buffer will correct the error. The data buffer will then report the data error to the processor, which will receive the address of the data in error in the asynchronous fault address register and the type of fault in the asynchronous fault status register. In addition, there will be an information in the data buffer error register as to whether the error was correctable or uncorrectable and the ECC syndrome that was used.

The processor can be programmed to ignore correctable data errors, in which case the address and status are simply registered in the asynchronous fault address and status registers. In this case the processor is not interrupted. The processor can be programmed to receive a correctable data fault upon detection of a correctable data error by enabling the correctable error enable bit in the E-cache error enable register.

For I/O accesses which read from memory, the SYSIO chip is responsible for detecting both correctable and uncorrectable data errors. There are corresponding interrupt vectors (uncorrectable error interrupt mapping register and correctable error interrupt mapping register) as well as uncorrectable error and correctable error fault address and status registers. When the SYSIO detects a fault, it will interrupt the processor indicated by the corresponding interrupt mapping register which, in turn, maps through the PC Interrupt mapping registers.

4.2.3 Automatic Recovery

The Enterprise 10000 system is a shared memory multiprocessing system. There can be up to 64 processors, and the current maximum memory of 16 Gbytes is split into multiple banks. FIGURE 1-6 shows how the processors and memory are interconnected by the crossbar. A key point is that all processors and all sections of memory are equal and the Enterprise 10000 system can continue processing with less than the total processors or memory configured.

A processor failure can usually be detected by parity checking. As explained above, single bit memory errors are automatically corrected and double bit errors, and worse, are detected. Should any of these errors occur, a diagnostic routine called POST is automatically invoked. POST stands for power-on self-test but does a lot more than that. POST is invoked at system boot time or when a reboot is required

following an error being detected. For each system board in the Enterprise 10000 system, POST can isolate problems to processor, memory, and I/O. If a problem is found, POST reconfigures the Enterprise 10000 system to eliminate use of any faulty items. POST is designed to pick the best system configuration by exploring different legal ways of doing it. A figure of merit (FOM) is calculated for each different way and the greatest FOM value wins. However, the system administrator can set the overall objectives for POST; for instance to keep all system interconnects up and exclude a processor, if necessary. In fact, this is the normal way of organizing POST, as an interconnect failure has a more drastic effect on performance than the loss of a single processor. However, if more appropriate, the global address bus, for instance, can degrade from four to three to two to one. In the same way, the Enterprise 10000 system can withstand loss of several processors or banks of memory and continue to function. There is a reduced level of service to the user population but work can continue.

The reboot time discussed above is, of course, dependent on the amount of hardware configured on the Enterprise 10000 system. For a typical Enterprise 10000 system of 16 processors and 1 Gbyte of memory, the reboot time is 15 minutes. This time includes integrity checking of the customer's data on disk. This checking, and subsequent repair if required, is virtually instantaneous as all previous transactions have been logged by the logging file system. Reboot time is down time for customers and counts against availability.

4.2.4 Resiliency Features

4.2.4.1 DC Power

TheEnterprise 10000 system logic DC power system is modular at the system board level. Bulk 48-VDC is supplied through a circuit protector to each system board. The 48 volts is converted through several small DC-to-DC converters to the specific low voltages needed on the board. Failure of a DC-to-DC converter will affect only that particular system board.

4.2.4.2 Logic Boards

Logic boards can be removed from and inserted into (hotswap) a powered on and operating Enterprise 10000 system for servicing the on-board supplies.

The control board contains the SSP interface as well as the clock source, JTAG master, and emergency shutdown logic. Optionally, two control boards can be configured in the system for redundancy.

A centerplane support board holds the DC-to-DC converters for one side of the centerplane, which power the global address and data router ASICs. Should one centerplane support board be operational, the centerplane will continue to operate in a degraded mode which includes two of the four address buses and a 72-bit wide data bus instead of the full 144 bits (see Section 4.2.4.8, “Data Interconnect Availability”).

4.2.4.3 Processor

In the event of a failure of a UltraSPARC processor, the UltraSPARC data buffer ASIC, the external cache SRAMs, or the associated port controller, the failed processor can be isolated from the remainder of the system by a power-on self-test (POST) configuration step. As long as there is at least one functioning processor available in the configuration, an operational system will result.

If one of the coherency interface controllers in a system board fails, the global address router on which the failing coherency interface controllers sits can be deconfigured. The resulting configuration keeps all the processors available.

4.2.4.4 Memory

There is one memory controller on each system board. If any one fails, it can be isolated from the bus by a POST. A failed memory controller will also make the corresponding memory on the respective system board unavailable.

When POST completes the testing of the memory subsystem, faulty banks of memory will have been located. POST can then configure a maximum memory configuration using only reliable memory banks by taking advantage of the highly configurable nature of the address match logic in the memory controller.

Another resiliency feature connected with the memory uses word mapping. At the lowest level, the memory is composed of multiple data storage chips called DRAMs (dynamic random access memories). The mapping of words to DRAMs is done in a fashion so that if a complete DRAM should fail, four successive memory words will have single bit (correctable) errors. This is clearly superior to a 4-bit failure in one word, which can result in a system crash.

4.2.4.5 I/O Interface Subsystem

If a SYSIO chip fails, an entire SBus is inaccessible (each SBus is capable of supporting up to two SBus cards). This is discovered by power-on self-test and that SBus is isolated from the system. This will result in a loss of connectivity to some I/O interface subsystem resources.

However, the Enterprise 10000 system facilitates maintaining connectivity by enabling alternate SBus controllers to be configured. This enables the connections to networks to have redundant paths. The same can be done with disks as detailed in Section 4.2.5, “I/O Redundancy,” later in this section.

4.2.4.6 Interconnect Availability

Both the address and data global interconnect are implemented from point-to-point routers, instead of the multidrop buses used on the CS6400 system. This means that a chip failure on one of the ports to the interconnect does not render the entire interconnect unusable, only that particular port.

4.2.4.7 Address Interconnect Availability

Addresses are broadcast between boards via four independent, error-correcting-code protected global address buses. Should one or more of these buses (or the attached coherency interface controllers) experience a chip failure, the bad bus or buses can be configured out of the system after a reboot. The system can be configured to operate with four, three, two, or one global address buses, providing 100%, 75%, 50%, or 25% of the normal address bandwidth. The interconnect is normally address-router limited only for configurations of more than 13 or more boards, so it is very large configurations that are degraded by address-bus failures.

4.2.4.8 Data Interconnect Availability

Data is moved between boards via the global data router. It is a 16 x 16 crossbar, with each port consisting of two unidirectional 8-byte-wide error-correcting-code-protected paths. The data router is physically divided into two 32-byte wide halves. Should a failure occur in one half of the data router, the system can be reconfigured after reboot to operate using only the good half of the router width. Data transmissions then take eight cycles of 8-bytes each, instead of the normal four cycles of 16-bytes each-providing half the normal data bandwidth. The interconnect is normally data-router limited for configurations of 12 or less boards, so it is small and medium configurations that are degraded by data-router failures.

4.2.5 I/O Redundancy

There are also resiliency features that protect the I/O (including the networks) of the Enterprise 10000 system from down time. The significant difference between protection of the I/O and that of the system boards discussed above, is that for the

I/O there is no concept of having the machine run in a degraded fashion. The strategy is to switch to alternate hardware instead. FIGURE 4-1 shows the options available.

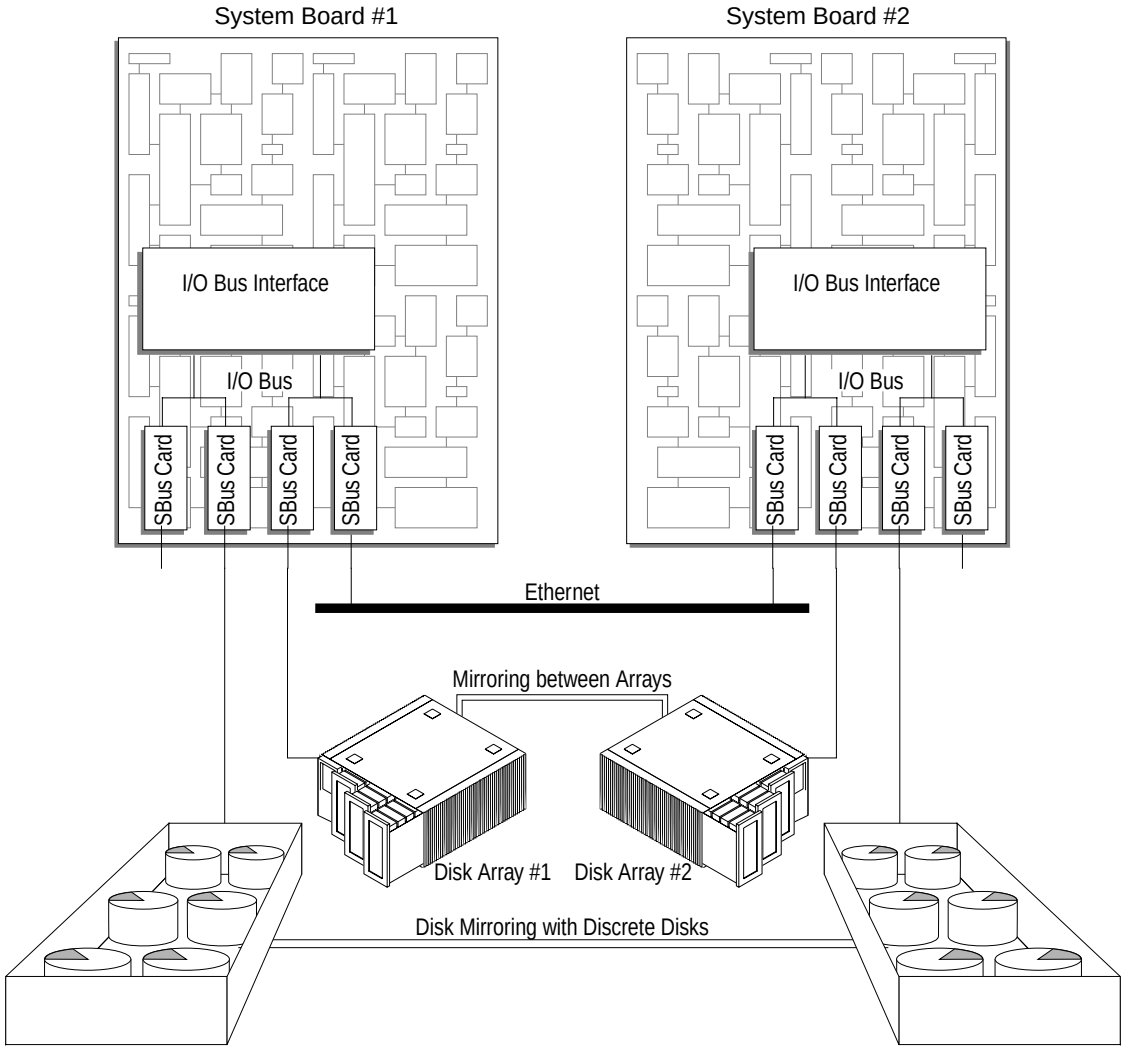

FIGURE 4-1 Alternate Hardware Diagram

Each of the system boards in the Enterprise 10000 system has two SBuses. For each SBus, there can be up to two I/O controllers. Protection against a failure of an I/O controller can be provided by using a duplicate controller, as shown above for Ethernet and disks. The duplicate controller is configured on a different system

board than the primary controller to protect against the system board being taken out of service or failing. The alternate pathing feature will assist the System Administrator to cleanly switchover from one I/O controller to the other.

4.2.6 Disk Redundancy

The Enterprise 10000 system offers two types of disk storage. *Discrete* disks are packaged six to a tray, and there are disk arrays with up to 30 disks arranged in a matrix. Mirroring (often called RAID 1) will protect against the failure of a disk by having the information duplicated on an alternate disk. To protect against the failure of the actual disk tray, the mirrored disks should be in separate trays as shown in FIGURE 4-1. The array is also shown mirrored by a second array. This gives protection against failures of the array logic board, power supply, and fans, as well as the disks themselves.

The disk arrays will also support a RAID 5 arrangement of disks. While mirroring requires there to be twice as many disks to gain security against a failure, RAID 5 requires only 20% more disks. For instance, for every five disks being used to store the customer's data, there is an extra parity disk that enables the data to be reconstructed should one of the data disks fail. (In practice, the data and the parity are distributed amongst all six disks. This achieves better performance than having a dedicated parity disk). If any of the six disks in a group should fail, the missing information is reconstructed from the surviving disks.

4.2.7 Redundant Components

Both the customer mean time to interrupt and the customer availability measures of the system are enhanced by the ability to configure redundant components. There are no components in the system that cannot be configured redundantly if the customer desires. Each system board is capable of independent operation. The Enterprise 10000 system is built with multiple system boards and is inherently capable of operating with a subset of the configured boards functional.

In addition to the basic system boards, redundantly configurable components include:

- Control boards
- Centerplane support boards
- Disk data storage
- Bulk power subsystems
- Bulk power supplies
- Peripheral controllers and channels

- System service processors and interfaces

It is the system administrator's option to configure the data storage capacity in several different modes of redundancy from mirrored disks to RAID 5 to simple nonredundant storage. Replacement of failing disks in redundant systems can be accomplished without interfering with system operation.

Systems can also be configured with multiple connections to the peripheral devices enabling redundant controllers and channels. Software maintains the multiple paths and can switch to an alternate path on the failure of the primary.

The SSP is controlled via the control board. Redundant SSPs and interfaces can be configured if the customer desires.

4.2.8 Enterprise 10000 System Failover

The previous discussions shows how features built into the Enterprise 10000 system enable 99.95% system availability to be achieved. To move up to the next availability level requires duplication of the system itself. FIGURE 4-2 shows a pair of Enterprise 10000 system with failover provisions using Sun's Energizer technology.

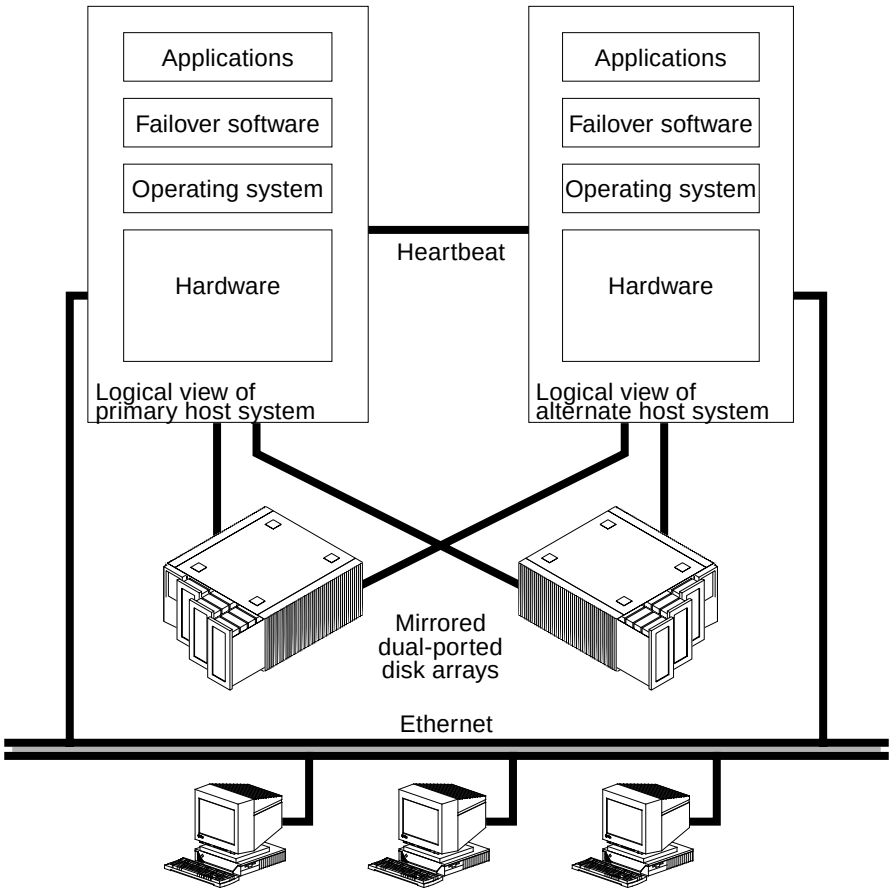

FIGURE 4-2 Failover Configuration

On the left is the primary Enterprise 10000 system. Its data is stored on two dual ported disk arrays, one of which is a mirror of the other. User access is over Ethernet. The alternate Enterprise 10000 system is identically configured from a hardware, operating system, and application software point of view. Between the two is a private link, usually Ethernet, that enables the alternate system to monitor the primary via a *heartbeat monitor.* Failover software resides on both systems. Should a problem develop with the primary Enterprise 10000 system, there is a switchover to the alternate Enterprise 10000 system and users will generally regain service within one minute.

It is expensive to keep an alternate Enterprise 10000 system idle waiting for the primary machine to fail. In practice, the alternate can normally be assigned to perform less critical tasks that can be suspended in the event of a failover. For

instance, sites can use a second system for development and testing purposes. Another arrangement is to accept a lower level of service from the alternate and configure it with less hardware.

There are some side advantages of a failover arrangement. One is that scheduled maintenance periods are no longer relevant. A deliberate switchover to the alternate can be performed to enable, for instance, an operating system upgrade to be made to the primary machine. Another is that the primary and alternate Enterprise 10000 system can be installed up to a mile apart to protect against site disasters (fiber optic connections to the disk arrays enable this).

4.2.9 Domains

A final Enterprise 10000 system feature that contributes to *all* the RAS goals of the Enterprise 10000 system is *domains*-a system within a system. The Enterprise 10000 system can be logically divided into multiple sections. Each domain is comprised of one or more system boards so a domain can be a minimum of one processor or up to 64 processors. Each section runs its own copy of the operating system and has its own peripherals and network connections. Domains are useful for testing new applications or operating system updates on the smaller sections, while production work continues on the remaining (and usually larger) section. There will not be any adverse interaction between any of the sections and customers can gain confidence in the correctness of their applications without disturbing production work. When the testing work is complete, the Enterprise 10000 system can be rejoined logically without a Solaris reboot (There are no physical changes when using domains).

4.3 Serviceability Features

To support the lowest possible mean time to repair, the Enterprise 10000 system has been designed with a number of maintenance features and aids. These are used by the Enterprise 10000 system administrator and by the service provider.

There are several features that enable service to be performed without forcing scheduled down time. Failing components are identified in the failure logs in such a way that the field-replaceable unit is clearly identified. All boards and power supplies in a properly configured system can be removed and replaced during system operation without scheduled down time.

4.3.1 Diagnostics and Monitoring Tools

The Enterprise 10000 system utilizes several type of diagnostics and monitoring tools. Several of them are SSP-based diagnostics, indicating they run out of the SSP to test the Enterprise 10000 host. These types of tests include bring-up, OpenBoot PROM, and power-on self-test (POST). Other tests run directly on the Enterprise 10000 host. These tests include exercisers and `hpost` (host POST).

1. Bring-up

Bring-up diagnostics provide static, repeatable testing that catch most hard errors. The SSP provides site-configurable levels setting the duration to run bring-up diagnostics. If bring-up diagnostics find a failure at a given level, then no event will cause a lower level to be executed until some positive action is taken by a system administrator or customer service stating that it is OK to return to the lower level. Bring-up diagnostics log all failures to the system log file.

2. OpenBoot PROM

The Enterprise 10000 system uses an SSP-resident version of the OpenBoot firmware. The Enterprise 10000 version of OpenBoot resides on the SSP hard disk, not in hardware PROM as in other Sun systems. The primary task of the OpenBoot firmware is to boot the operating system from either a mass storage device or from a network. The firmware also provides extensive features for testing hardware and software interactively.

3. Power-On Self Test

Power-on self-test (POST) is the principal bring-up diagnostic. It runs on the individual UltraSPARC processors, under the direction of the SSP. POST exercises the Enterprise 10000 system logic at a level below that of the field replaceable unit (FRU), and with a high degree of accuracy finding failing components-enabling isolation to the field-replaceable unit.

POST has four objectives:

a. To provide a highly available platform for customer applications, even in the face of hardware failures.

b. To provide the lowest-level start-of-day configuration services, including detailed interaction with specific hardware components. This removes the need for higher levels such as OBP to deal with the hardware at this level.

c. To provide support for manufacturing bring-up of the hardware.

d. To record sufficient information about failed and marginal components so that both field replacement and subsequent factory repair are expedited.

4. Control Board Self-Tests

Upon power-on/reset, the control board executes a series of diagnostic self-tests. As each test executes, the corresponding diagnostic LED is illuminated. As each test successfully completes, the LED for that test is extinguished. In the event of a test failure, the corresponding diagnostic LED will illuminate and then extinguish or flash as the test completes or fails.

5. Status Monitoring and Displays

The SSP provides services related to monitoring and reporting on the status of the machine. All displays minimally only require an X server; in this mode the display provider is run on the SSP with the DISPLAY variable set to select the X server. Optionally, display programs such as netcon and Hostview can actually be run on any X windows platform, obtaining information from the SSP and providing actions via the network.

6. SunVTS™

SunVTS, the online validation test suite, tests and validates hardware functionality by running multiple diagnostic hardware tests on configured controllers and devices. SunVTS is the Solaris 2.5 replacement for what was known as SunDiag™. Refer to the SunVTS AnswerBook™ or sunvts(1M) man page for additional information.

7. Redx

`Redx` is an interactive debugger built by the POST group. Redx shares portions of its code with `hpost`. Online reference documentation for redx is built into the program.

4.3.2 Error Logging

When uncorrectable errors occur, information about the error is saved to help with isolation. The Enterprise 10000 system has extensive error logging capabilities. When any hardware error (excluding memory data) is detected, the error is logged in the detecting component and, if applicable, the history buffers are frozen in the multiplexer, coherency interface controller, port controller, and global data arbiter chips. The SSP can then detect these errors by polling these components. If the error is fatal (for example, control bus parity error), the system is stopped, error log information is collected by the SSP, and the system is automatically rebooted. If the error is correctable (for example, single-bit interconnect error) the error information is logged and system operation continues.

The goal is to be able to isolate to a failing chip or to a single wire between chips. Parity protection does not lend itself intrinsically to single-wire isolation. However by saving all the bits in a parity error, a strategy of repeated error samples and knowledge of the semantics of the datum being communicated can often narrow the fault down considerably beyond the notification of parity error.

4.3.3 Standard JTAG Capabilities

The standard JTAG instruction set is implemented in all of the Enterprise 10000 ASICs. The command and status register chains are ASIC-specific in length and make up the JTAG control/status functions. These are similar to the shadow registers in the CS6400 system. Additionally the command and status register JTAG registers can share bits with the system registers. In this case, several system registers in the part can be combined into one JTAG command and status register.

The history chains are ASIC-specific in length. Multiple history register sets can be combined into one of two chains. Two chains are provided to enable minimal configurability in the length (divide by two).

4.3.4 Mechanical Serviceability

Connectors are keyed so that boards cannot be plugged in upside down. Special tools are *not* required to access the inside of the system. This is because all voltages within the cabinet are considered extra-low voltages (ELVs) as defined by applicable safety agencies. A torque wrench is required for the system board mezzanine modules.

No jumpers are required for configuration of the Enterprise 10000 system. This makes for a much easier installation of new and/or upgraded system components. There are no slot dependencies other than the special slots required for the control and centerplane support boards.

There is a special keying feature that prevents control boards, centerplane support boards, and system boards from being plugged into each other's slots. The only slot difference is the address of the JTAG chain used to access a slot from the SSP. System level addresses (deviceIDs) are assigned by POST and scanned into the global address router client chips prior to boot of the system. There are no slot dependencies for the deviceIDs.

The Enterprise 10000 cooling system design includes features that provide strength in the area of RAS. Standard proven parts and components are used wherever possible. Field replaceable units (FRUs) and subassemblies are designed for quick and easy replacement with minimal use of tools required.

Power on replacement (hot swap) of electronic modules and cooling system components is such that any one component can be removed for an indefinite period of time. This can be done without causing impairment or damage to the system due to overheating caused by system pressure drop due to the removed module, or removal of a portion of the cooling system. Finally, air filters are replaceable while the system is operational.

4.3.5 Concurrent Serviceability

The most significant serviceability feature of the Enterprise 10000 system is to replace system boards online as a concurrent service. Concurrent service is defined as the ability to service various parts of the machine without interfering with a running system. Failing components are identified in the failure logs in such a way that the field replaceable unit (FRU) is clearly identified. With the exception of the centerplane, all boards and power supplies in the system can be removed and replaced during system operation without scheduled down time. Replacing the control board that is currently active, or switching control to the redundant control board, requires a UNIX shutdown and reboot.

The ability to repair these items without incurring any down time is a significant contribution to achieving higher availability. A by-product of this online repairability of the Enterprise 10000 system concerns upgrades to the on-site hardware. Customers may wish to have additional memory or an extra I/O controller. These operations can be accomplished online with users suffering only a brief (and minor) loss of performance while the system board affected is temporarily taken out of service. Concurrent service is a function of the following hardware facilities:

- All centerplane connections are point-to-point making it possible to logically isolate system boards by dynamically reconfiguring the system.
- The Enterprise 10000 system uses a distributed DC power system, that is, each system board has its own power supply. This type of power system enables each system board to be powered on/off individually.
- All ASICs that interface to the centerplane have a loopback mode that enables the system board to be verified before it is dynamically reconfigured into the system.

4.3.5.1 Dynamic Reconfiguration of System Boards

The online removal and replacement of a system board is called dynamic reconfiguration (DR). DR can be used to remove a troubled board from a running system. For example, the board can be configured in the system even though one of its processors failed. In order to replace the module without incurring down time, DR can isolate the board from the system, hot swap it out, and then replace the failing processor module. Therefore, the DR operation has three distinct steps:

- Dynamic detach
- Hot swap
- Dynamic attach

DR enables a board that is not currently being used by the system to provide resources to the system. It can be used in conjunction with hot swap to upgrade a customer system without incurring any down time. It can also be used to replace a defective module that was deconfigured by the system and subsequently hot swapped and repaired or replaced.

Dynamic deconfiguration and reconfiguration is accomplished by the system administrator (or service provider) working through the SSP. The first step is the logical detachment of the system board by the Hostview program. The Solaris operating system's scheduler is informed, for the board in question, that no new processes should start. Meanwhile, any running processes and I/O operations are enabled to complete, and memory contents are rewritten into other Enterprise 10000 memory banks. A switchover to alternate I/O paths then takes place so that when the system board is removed, its I/O controllers will not be missed. The next step in the process is for the system administrator to manually remove the now inert system board from the Enterprise 10000 system-a *hot swap* operation. The removal sequences are controlled by the SSP, and the system administrator follows instructions given by Hostview. The removed system board is then repaired, exchanged, or upgraded, and the second half of hot swap is employed to reinsert it into the Enterprise 10000 system. Finally, the replaced system board is dynamically configured in by the operating system. The I/O can be switched back, the scheduler assigns new processes and the memory starts to fill.

So with a combination of dynamic reconfiguration and hot swap, the Enterprise 10000 system can be repaired (or upgraded) with minimal user inconvenience. There can also be, for a 32 processor Enterprise 10000 system, a loss of 12% of processing power until the detached system board is back in service. Hot swap minimizes this interval to minutes by on-site exchange of system boards.

An interesting additional advantage of dynamic reconfiguration and hot swap is that online system upgrades can be performed. For instance, when a customer purchases an additional system board, it too can be added to the Enterprise 10000 system without disturbing any user activity.

4.3.5.2 Control Board Removal and Replacement

A control board that has been deconfigured from the system (so it is not supplying system clocks) can be removed from the system with no system interruption.

4.3.5.3 Centerplane Support Board Removal and Replacement

A centerplane support board that has been deconfigured from the system (so it is not supplying clocks to the centerplane and the centerplane is running is a degraded mode) can be removed from the system with no system interruption.

4.3.5.4 Bulk Power Supply Removal and Replacement

Bulk 48-volt power supplies can be hot swapped with no interruption to the system. This assumes a standard configuration from the factory, which is configured for power supply redundancy.

4.3.5.5 Fan Tray Removal and Replacement

Following the failure of a fan, the remaining working fans are set to high speed operation by the SSP in order to compensate for the reduced airflow. The system is designed to operate normally under these conditions until such time as the failed fan assembly can be conveniently serviced. The fan trays can be hot swapped with no interruption to the system.

4.3.5.6 Interconnect Domain Isolation

The Enterprise 10000 system has an interconnect domain facility that enables the system boards to be assigned to different software systems. For example, one interconnect domain can be doing production, while a second interconnect domain is experimentally running the next revision of Solaris, and a third interconnect domain is exercising a suspected bad board with production-type work.

4.3.5.7 Disks

The disk arrays available with the Enterprise 10000 system have some interesting serviceability features. The arrays can be configured with up to 30 disks and one (or more) disks can be designated as a hot spare. If the remaining disks in the array are mirrored, and if one should fail, the hot spare is automatically brought into service. Then a rebuild operation takes place to bring the replacement drive up to date using the information from its mirror. The failed drive can be replaced at a later time using the *warm swap* feature of the array. The array has three individual disk trays that hold up to ten disks each. Any tray can be removed or inserted into the array while the power is live, as long as data traffic to the tray is quiesced. So by bringing down one half of the disk mirror, a disk can be replaced while the other half of the mirror continues to provide service.

4.3.6 Nonconcurrent Service

Nonconcurrent service requires the entire system to be powered down. This category is limited to replacement of the centerplane or components on the centerplane.

4.3.7 Remote Service

The optional capability exists for automatic reporting (to customer service headquarter sites) of unplanned reboots and error log information via email.

Every SSP has remote access capability that enables remote login to the SSP. Via this remote connection, all SSP diagnostics are accessible. Diagnostics can be run remotely or locally on deconfigured system boards while the Solaris is running on the other system boards.

4.3.8 Field Replaceable Units

The Enterprise 10000 system has a hierarchy of field-replaceable units (FRUs). While this complicates the field service strategy, it also minimizes cost by replacing the particular failing system component when possible, rather than the entire expensive system board. The Customer Service goal is to ship all of the Enterprise 10000 system FRUs via common carrier. An item-by-item analysis was performed to make certain that this was possible.

Hot swap repairs can be carried out while the system is running, while nonconcurrent service requires the entire machine to be shut down.

Index

Reader Comment Card

Your comments and suggestions are important to us. Please let us know what you think about the *Ultra Enterprise 10000 System Overview Manual*, part number 805-0310-12.

1. Were the procedures well documented? Yes ❑ No ❑

   Please explain: _____

   _____

2. Were the tasks easy to follow? Yes ❑ No ❑

   Please explain: _____

   _____

3. Were the illustrations clear? Yes ❑ No ❑

   Please explain: _____

   _____

4. Was the information complete and easy to find? Yes ❑ No ❑

   Please explain: _____

   _____

5. Do you have additional comments about the Ultra Enterprise 10000 System Overview Manual?
   You can send detailed comments via email to `smcc-docs@sun.com`, or send a fax to *SMCC Doc Feedback* at (415) 786-6443.

   _____

   _____

   _____

   _____

   _____

   _____

   _____

Your Name: _____

Title: _____

Company Name: _____

Address: _____

City: _____ State / Province: _____

Country: _____ Zip / Postal Code: _____

Email Address: _____

Telephone: _____

Thank you.

AIR MAIL
PAR AVION

IBRS/CCRI No. 808

Sun

NE PAS AFFRANCHIR

NO POSTAGE
NECESSARY
IF MAILED
TO THE
UNITED STATES

INTERNATIONAL BUSINESS REPLY MAIL/REPONSE PAYEE

PERMIT NO. 808 MOUNTAIN VIEW CA

POSTAGE WILL BE PAID BY ADDRESSEE

INFORMATION PRODUCTS
M/S MPK14-108
SUN MICROSYSTEMS INC
2550 GARCIA AVE
MOUNTAIN VIEW CA 94043-9551
UNITED STATES OF AMERICA