RS/6000 SP

SP Switch2 Technology and Architecture

March 2001

March 2001

Contents

Figures

Tables

Preface

This document describes the IBM® RS/6000® SP™ Switch2 subsystem. Information in this document does not apply to any other SP switch design.

Questions or comments

Please send questions or comments to Frank May at fhmay@us.ibm.com

Chapter 1. Introduction

The SP Switch2 is the next step in the evolution of the SP™ interconnection fabric and offers significant improvements in bandwidth, latency, and RAS (Reliability, Availability, and Serviceability) over the SP Switch. The bandwidth gains result from design improvements that allow the switch adapter to be located directly on the node's system bus. RAS improvements are discussed throughout this document and detailed in "Appendix B. SP Switch2 Reliability, Availability, and Serviceability (RAS) enhancements" on page 37. The new SP Switch2 Adapter has several other design improvements including an on-board Rambus™ memory component and a hardware driven, datagram segmentation and reassembly engine (see Figure 1 on page 2).

Note: Information regarding measurement of performance improvements will be available in the SP Switch2 update of the SP Switch Performance paper. This paper (and others) are available from www.ibm.com/eserver/pseries. From that website select:

- Library (tab on left)
- White Papers and Technical Reports

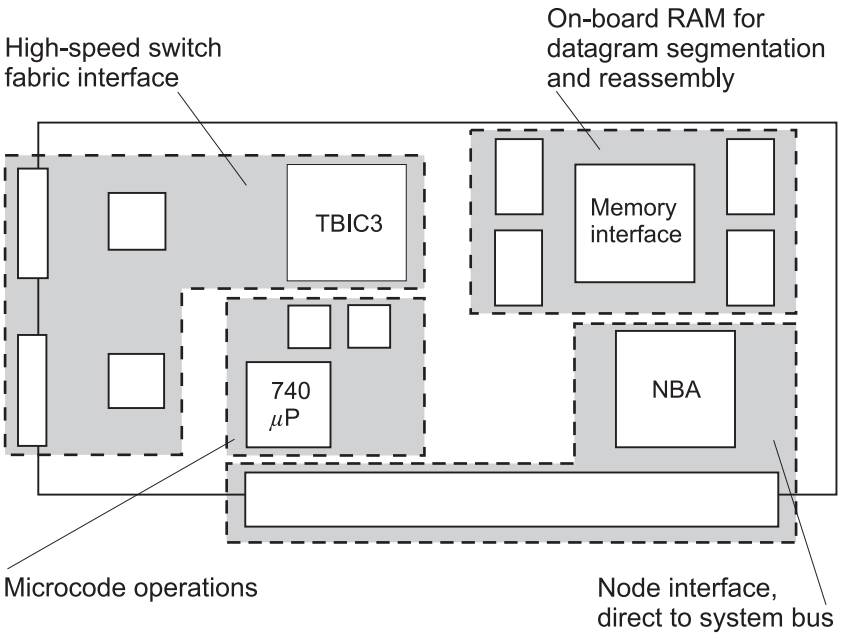

Figure 1. SP Switch2 Adapter functional layout

Because the SP Switch2 design improvements are evolutionary steps on the SP and High Performance switches, the SP Switch2 is fully compatible with applications written for the older switches. The bandwidth improvements of the SP Switch2 are designed to accompany the performance enhancements of the POWER3 SMP High Nodes and 375 MHz POWER3 SMP High Nodes. Because of that, the SP Switch2 is only supported on those nodes.

Design objectives

During the development of the SP Switch2 subsystem, the primary design objectives were to significantly increase communication bandwidth and reduce message latency. Like prior switches, the SP Switch2 software support continues to emphasize usage of MPI (Message Passing Interface), LAPI (Low level Application Interface), and IP (Internet Protocol) but the new switch subsystem enhances message handling so that the switch can keep pace with the increased performance of the 222 MHz POWER3 SMP High Node and the 375 MHz POWER3 SMP High Nodes. To meet these objectives, the design team focused on four primary areas:

- Faster interfaces, data paths, and microprocessors
- Reducing microcode workload using hardware assisted datagram segmentation and reassembly
- Developing an adapter with multiple data paths allowing concurrent microprocessor bus activity and data movement

Faster interfaces, data paths, and microprocessors

In a fully configured POWER3 SMP High Node SP system, SP Switch2 components can offer performance increases (relative to the SP Switch) as shown in Figure 2 on page 3.

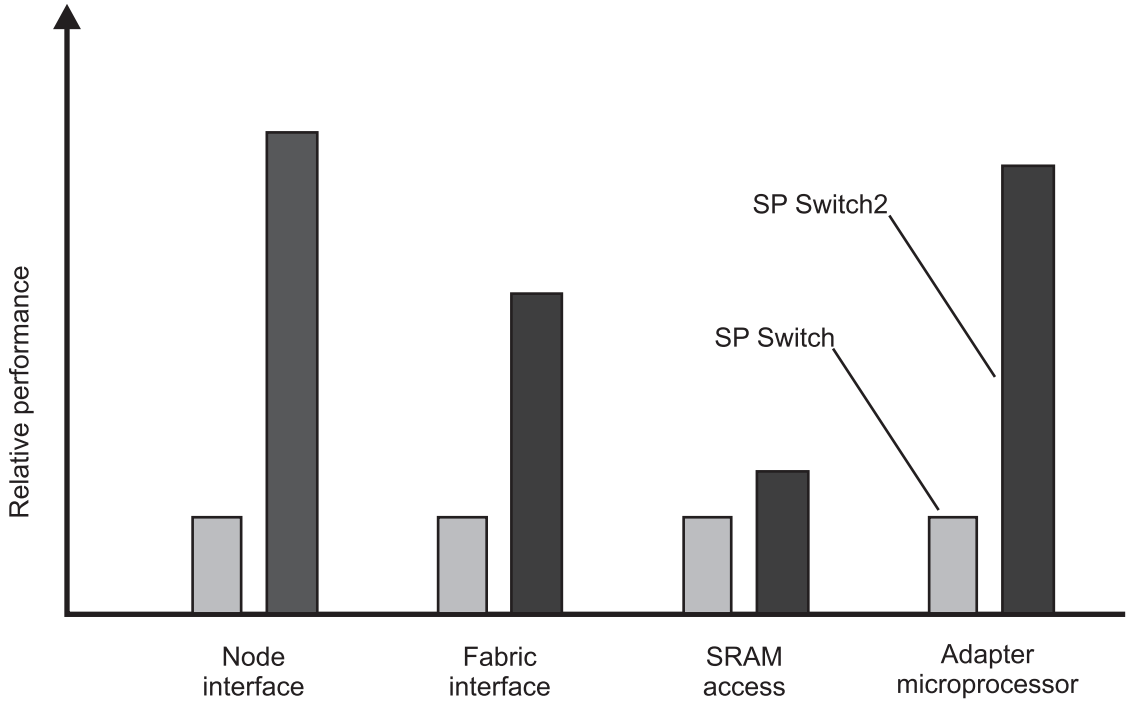

Figure 2. Comparison of SP Switch2 to SP Switch performance

Although the performance gains in logic, memory, and microprocessor technology are very good, additional improvements were needed to provide the required bandwidth. These additional improvements required new data transfer techniques for:

- Incoming message reassembly
- Reducing microcode workload
- Allowing concurrent microprocessor operations

Incoming message reassembly

An SP Switch2 system significantly increases the communication bandwidth over an SP Switch system. Some of this improvement came about by adding a new memory component to the adapter. The new memory component includes 16 MB of Rambus™ RDRAM memory to enable hardware driven message reassembly.

Reducing microcode workload

The SP Switch2 Adapter microcode is C Code running on the adapter's 740 PowerPC microprocessor. The adapter's microcode handshakes with the node's software and enables the exchange of messages between the node and the switch.

The new SP Switch2 Adapter improves both the communications bandwidth at the node interface (6XX bus) and the microprocessor performance by about 5X. In order to avoid performance reductions caused by the 6XX bus working faster than the microprocessor, the design team created new hardware that can take over some of the work formerly done by the microcode. The primary method of reducing microcode workload is through a process called the datagram segmentation and reassembly function.

Segmentation is the process of constructing 1K byte transmission packets from 4K byte software units. Reassembly is the process of converting multiple 1K byte

transmission packets into larger software units processed by microcode. Reassembling the transmission packets back into larger software units results in a significant reduction in the number of times the microcode is invoked. For a typical 64K byte IP datagram, the microcode only has to do processing for the first and last 1K byte switch packet.

Concurrent microprocessor operations

To improve data flow through the SP Switch2 Adapter, the design includes separate paths for microprocessor bus operations and data transfers. Instead of having a single 400 MB/s inter-chip bus used in the older adapters, the SP Switch2 Adapter implements five separate buses using one 576 MB/s bi-directional (simultaneous transfers in both directions) bus and four 1000 MB/s uni-directional buses. Each bus operates independently which enables concurrent data transfers. On the old adapters, the microprocessor cannot use the bus until all data is transferred. With the new adapter, some tasks can be off-loaded from the node to simultaneously process:

- One DMA data transfer to main storage
- One DMA data transfer from main storage
- Two inbound packets from the fabric
- Two outbound packets to the fabric
- Four internal DMA data transfers
- One 740 load instruction
- N 630 load instructions (dependent on pipeline length)
- N 740 store instructions (dependent on pipeline length)
- N 630 store instructions (dependent on pipeline length)

Hardware overview

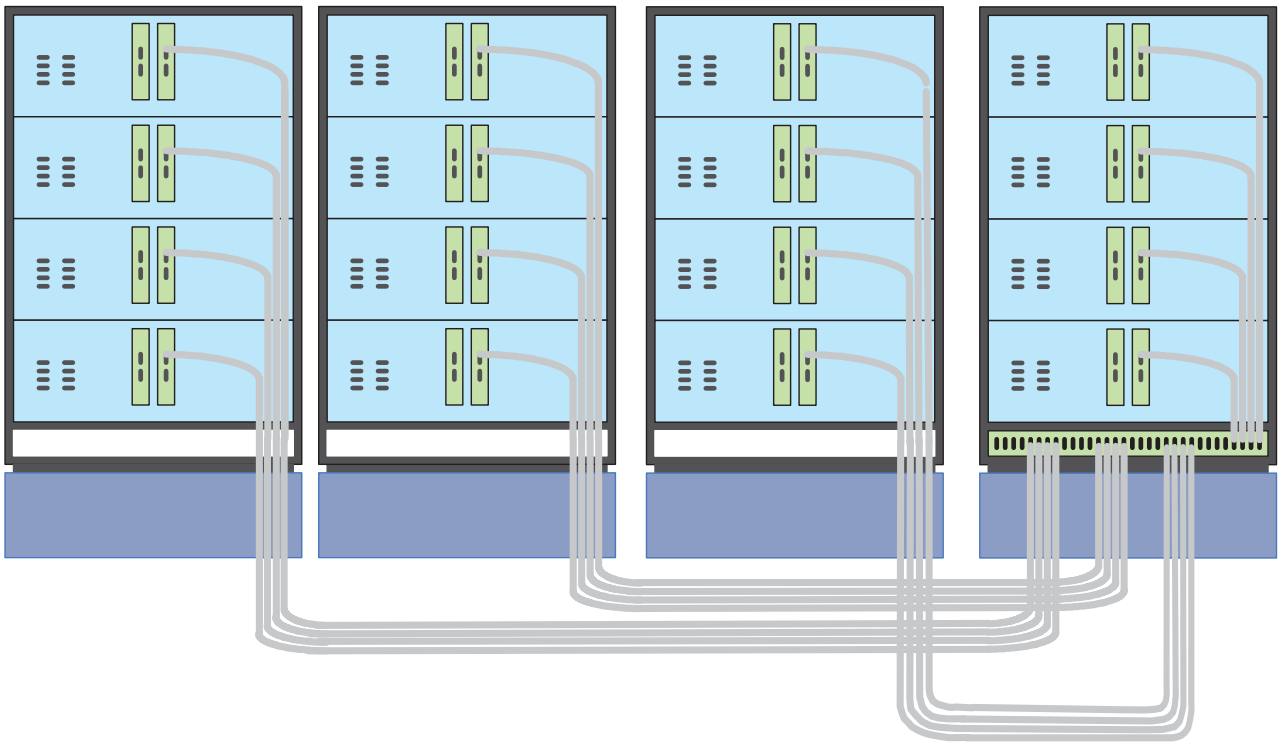

Figure 3. Typical SP Switch2 installation with one switch equipped frame and three non-switched expansion frames

Switch overview

The SP Switch2 has:

- 32 ports
  - 16 node-to-switch ports for interposer connections to nodes within the switch equipped frame or to nodes in non-switched expansion frames
  - 16 ports for switch-to-switch connections
- N+1 redundancy on:
  - hot-plug* power supplies
  - hot-plug* cooling fans

  (Refer to Figure 4 on page 6)
- Hot-plug* switch supervisor card
- Hot-plug* switch interposers

  (Refer to Figure 5 on page 6)

Notes:

1. Any unused switch ports must have a blank interposer card installed to prevent contamination of the connector and ensure proper cooling air flow.
2. New features are represented by *

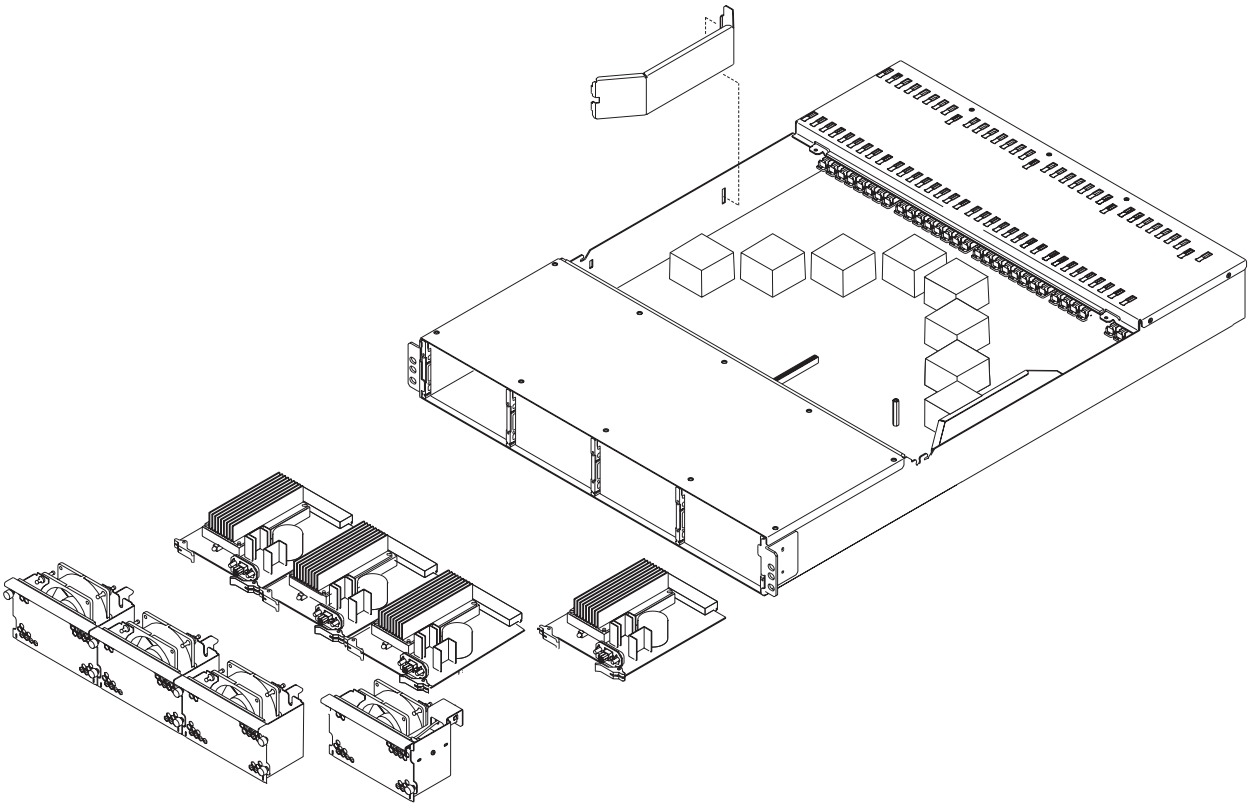

Figure 4. SP Switch2 fan assemblies and power supplies

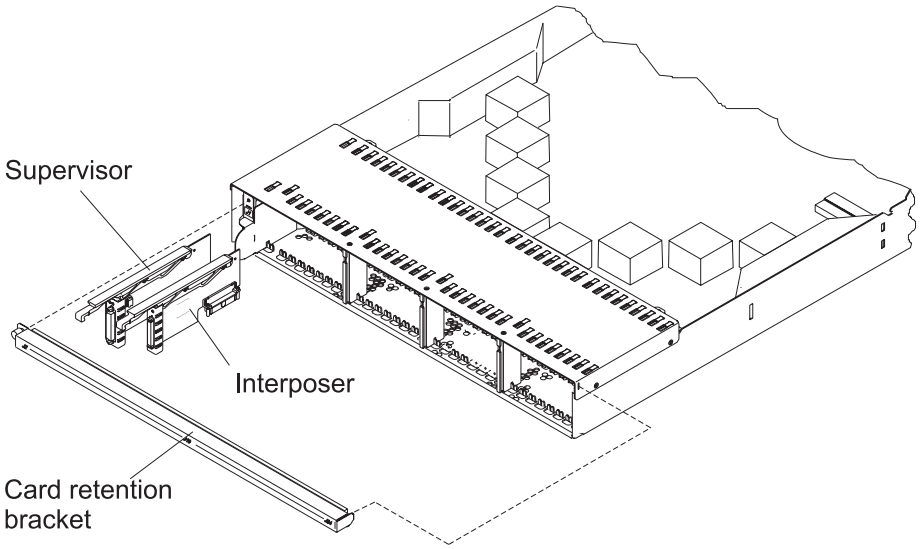

Figure 5. SP Switch2 supervisor and interposer cards

Switch board overview

The SP Switch2 utilizes a 16x16x2 design with a SP Switch form factor. Also, like the previous SP Switch, the new switch uses eight, 8x8 crossbar switch chips to route data through the system.

The eight switch chips on the SP Switch2 switch board are connected to thirty-two interposer chips that make the interface to all external components. Each interposer chip is one full-duplex port (simultaneous input and output) of the 16x16x2 switch and provides a hardware bandwidth of 500 MB/s in each direction. This yields an aggregate hardware bandwidth of 1000 MB/s per interposer. All eight of the 8x8 switch chips use ports 0 through 3 to communicate with the interposer chips, and ports 4 through 7 to communicate with the other switch chips on that board (refer to Figure 6 and Figure 8 on page 12).

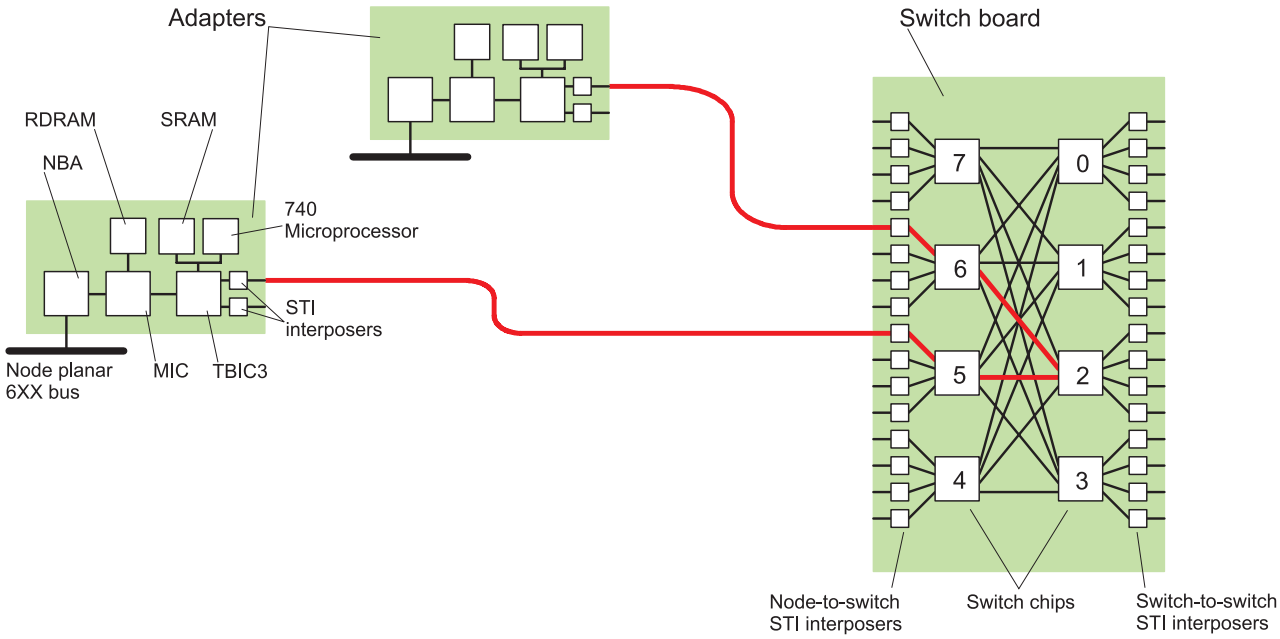

Figure 6. Switch board showing switch chip and interposer connections

Switch fabric improvements

The SP Switch2 subsystem offers the following improvements to system performance compared to the SP Switch:

- Asynchronous clocking rather than previous synchronous clocking (independent clocking for each board)
- Time of Day (TOD) propagation with cable delay compensation (hardware synchronization)
- 4 byte flit (flow control unit) rather than previous 2 byte
- 4 byte External Data Controller (EDC) rather than previous 2 byte
- 4 byte Cyclic Redundancy Check (CRC) rather than previous 2 byte
- 254 tokens rather than previous 61
- Larger Central Queue
- Enhanced software driven installation and configuration features that:
  - Automatically build the node portion of the switch topology file
  - Quickly and easily call out mis-wired switch to switch connections
- New Service Commands

Adapter overview

Like the SP Switch Adapter, the SP Switch2 Adapter is controlled by node software and on-card microprocessor microcode.

When you place the SP Switch2 Adapter into a POWER3 SMP High Node, you will see that the node has two connectors for the SP Switch2 Adapter. Both connectors are functional. As delivered from the factory, all adapters are installed in a specific slot and should not be moved. Once an adapter is placed in the node and PSSP is initialized, the adapter is recognized and assigned a switch node number.

The major hardware components of the SP Switch2 Adapter are illustrated in Figure 7 on page 9. The primary components of the SP Switch2 Adapter include:

NBA Node Bus Adapter, provides the interface to the node through a 16 byte wide 125 MHz 6XX bus giving the adapter a memory bus type interface. This gives the node the ability to issue load and store instructions to the adapter and the adapter the ability to directly access main storage.

MIC Memory Interface Chip, responsible for either passing data between the NBA and TBIC3 chips or for giving these chips access to the adapter's Rambus data memory component with data transfer rates of up to 2.4 GB/s into Rambus RDRAM.

Rambus™ RDRAM
Provides 16 MB RDRAM memory for packet reassembly using licensed Rambus technology with 11-bit wide data channels passing one byte of data every 1.66 nanoseconds.

TBIC3 Switch fabric interface controller, provides:

- Time of day (TOD) clocking
- Hardware for packet reassembly and segmentation
- Separate busses for concurrent data transfers and microprocessor bus operations

STI interposer chips
Self Timed Interface chip used as switch fabric drivers (on both the adapter and the switch board) providing 1 GB/s full duplex data transfers managed by a pair of macros; a small macro that drives signals a few inches, and an STI macro capable of driving long cables.

740 PowerPC microprocessor
Operates at 468 MHz internal and 72 MHz external to provide microcode control for adapter components, format and decode packet headers, build packet routing information, handle error conditions, and pass control information back to the node bus.

SRAM Data buffer operating at 72 MHz for the 740 PowerPC microprocessor.

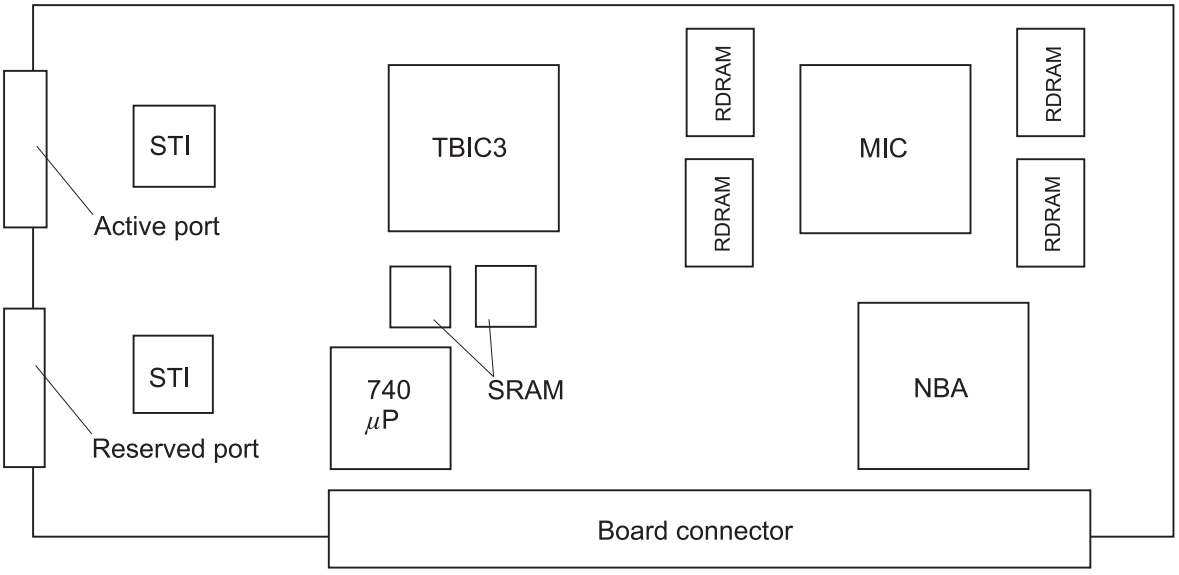

Figure 7. High level diagram of the SP Switch2 Adapter

Software overview

Like the SP Switch, the SP Switch2 also uses service packets between the switch chips and the nodes for switch management. To use the SP Switch2, the system must be operating with PSSP 3.2 at the recommended PTF level and AIX® 4.3.3 or later.

Adapter software overview

This section describes how the adapter microcode controls the various switch adapter components during data transfers.

The SP Switch2 Adapter microcode has been enhanced to support up to 16 user-space process connections (windows).

One of the major tasks performed by the microcode on the SP Switch2 Adapter is to examine incoming packets and then determine:

- When to put the packet in Rambus RDRAM
- Where the packet will eventually end up in main storage
- How to give hardware the right sequence of instructions at the right time to do the appropriate transfers

Hardware functions now replace some of the work that was previously done by microcode. These functions perform a task called packet reassembly. Packet reassembly functions collect several incoming packets and present them as a larger single block to the microcode. Because an adapter could be receiving packets streams from several different sources and the packets from individual streams can be received in any order, packet reassembly can be complicated. To simplify the process, each packet is encoded with several markers identifying each packet as part of a larger stream of packets.

Using these markers, the microcode either:

- Assigns a packet to Rambus memory for reassembly

  –OR–

- Assigns the packet to main storage if it is a small message

If the 1K byte packet is assigned to Rambus, the microcode assigns a location to the first packet and the rest are saved at simple address offsets. With hardware assistance, the microcode is only invoked twice for every reassembly. Depending on the message type (MPI or IP), this means that the microcode is invoked twice for each:

- 4K bytes
- 8K bytes
- 64K bytes

transmitted rather than once per packet. The microcode is initially called when the first packet arrives and then called again when the last packet in the message is reassembled.

SP Switch2 limitations

The SP Switch2 is only supported on (222 MHz) POWER3 SMP High Nodes and 375 MHz POWER3 SMP High Nodes. The SP Switch2 currently cannot be configured into systems using:

- Other node types
- SP Switches or High Performance Switches
- SP-attached servers
- SP Switch Routers
- System partitioning

System partitioning update:
PSSP now supports different levels of AIX™ and PSSP software in a non-partitioned SP system. Because of that, the need for system partitioning is largely if not altogether eliminated. Accordingly, partitioning support is being discontinued for major new SP environments, including the new SP Switch2.

Chapter 2. Hardware

SP Switch2 subsystem

The SP Switch2 subsystem has three main components:

- The SP Switch2 assembly
- Switch interposers
- SP Switch2 Adapters

Note: Switch component feature codes are listed in “Appendix C. Component feature codes” on page 41.

SP Switch2 assembly

The SP Switch2 assembly has five main components:

- The switch planar with switch chips
- Power supplies, four hot-plug units for N+1 redundancy
- Cooling fans, four hot-plug units for N+1 redundancy
- Switch supervisor card
- Circuit breaker assembly

Note: In keeping with the purpose of this document (describing SP Switch2 communication functions, not power and mechanical functions), only the switch planar will be discussed in this section. Additional details are also available in RS/6000 SP: Planning Volume 1, Hardware and Physical Environment (GA22-7280).

Switch planar

The switch planar holds eight, 8 x 8 crossbar switch chips. Four of these switch chips are connected to the node-to-switch ports and four are connected to the switch-to-switch ports. On all switch chips, ports 0 through 3 are connected to the interposer chips on the interposer cards (node or switch) and ports 4 through 7 are connected to the other switch chips on the planar. The chip-to-chip and chip-to-port connections are illustrated in Figure 8. Figure 9 on page 13 shows the layout of the switch chips and the interposer slots on the SP Switch2 planar.

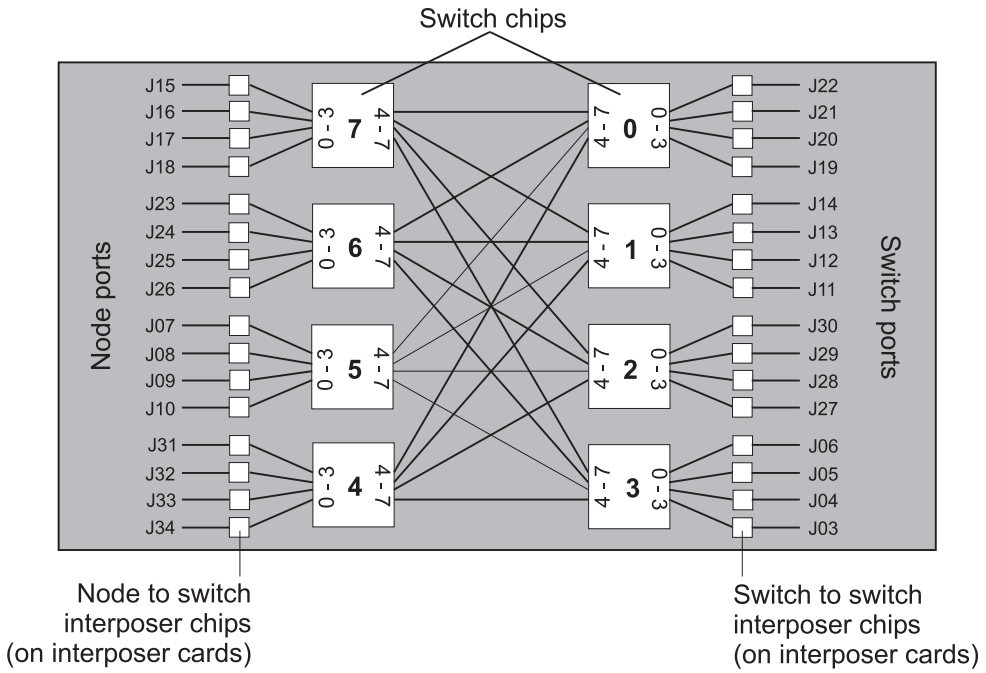

Figure 8. SP Switch2 planar showing chip numbers, chip connections, and port connections

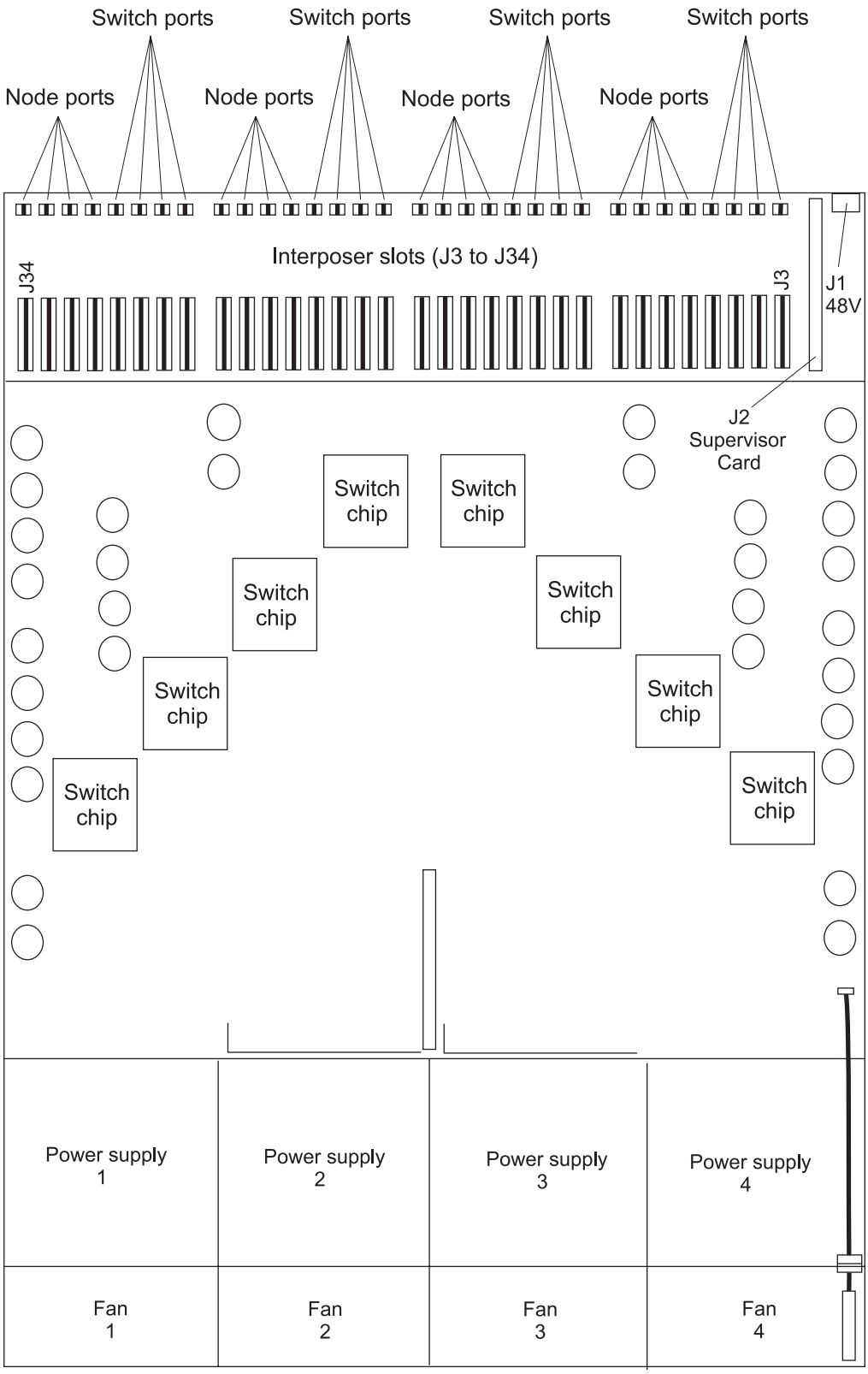

Figure 9. High level diagram of the SP Switch2 planar

Switch chip description:

Functional areas: The switch chips in the SP Switch2 have four functional areas:

Receiver logic
Handles link protocols, performs EDC checking, tracks outstanding tokens, controls receive buffer, and routes data to send ports or central queue

Sender logic
Receives data from the central queue or bypass paths, generates EDCs, tracks available tokens, and controls the send buffer

Central queue logic
Arbitrates between receivers that want to write into the central queue and senders that want to read from the central queue, serializes and deserializes data from flits into chunks

Service logic
Receives error information from other logic areas then generates error and status packets

In addition to the logic areas, each chip also maintains Time of Day (TOD) control logic and performance monitoring logic. Neither of these functions existed on previous switches.

TOD clocking: Unlike previous switch designs, the SP Switch2 subsystem does not have a master system clock. Instead, each switch planar has two redundant, internal clocks. Each switch chip also has its own internal clock operating at a unique frequency. The switch chips use a Phase Lock Loop (PLL) to synchronize the planar clocks to the chip's internal clock. To synchronize all switch chips, the system software selects one switch chip to act as a primary TOD chip which then propagates a TOD signal to all switch boards and adapters.

Each switch chip handles differences in phase and frequency of the transmitted TOD signal by using its receiver logic function. Part of the receiver logic (called the ping state) sends a ping character and counts the cycles until the return character comes back. The hardware calculates the delay, determines the cable lengths, and adjusts the TOD for delays due to cable length and intermediary switches.

Every 872 microseconds, the primary TOD chip sends out an additional TOD signal to maintain synchronization. Using the primary TOD signal and the calculated delays, synchronization across the system is maintained to about 1 microsecond.

During data transfers, the switch chip transmits the send clock to the interposer (along with the send data) and obtains the receive clock off the cable (with the receive data).

SP Switch2 interposer

The SP Switch2 has thirty-two interposer card slots and requires one interposer card for each switch connection. Sixteen of the card slots are used for node-to-switch connections and sixteen are for switch-to-switch connections. If a switch port is not in use, an interposer is not required for that port however, a blank interposer is required to maintain cooling air flow. Wrap plugs are only required during some diagnostic operations. If an interposer card should fail, they are hot-pluggable meaning they can be inserted or removed while the system is operating.

The primary component of the SP Switch2 interposer card is the STI interposer chip. Figure 10 illustrates an interposer card and shows the relative position of the interposer chip.

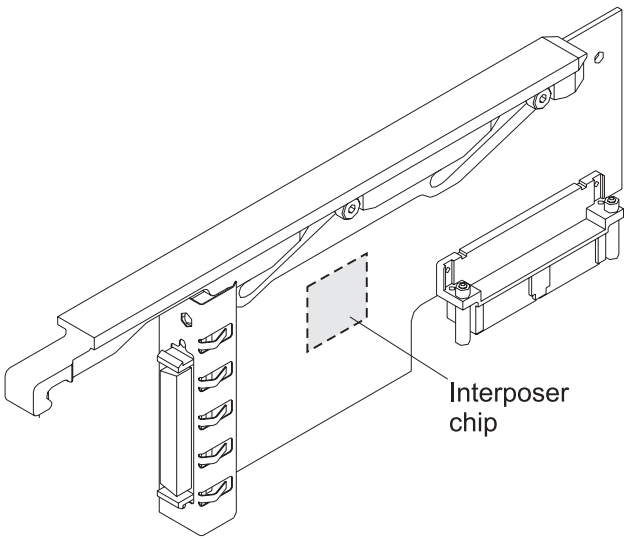

Figure 10. Interposer card showing position of interposer chip

Interposer chip

The SP Switch2 STI interposer is a high speed, byte wide (9 bits + clock) transceiver chip. The chip performs simple buffering operations and is capable of transmitting and receiving data at a full duplex rate of 1000 MB/s (500 MB/s/direction) over local and cable interfaces. The chip design has a 2 nanosecond data period using proprietary STI link protocol and hardware for unidirectional, differential STI I/O.

STI Interposer chip has two interfaces, one to communicate to a local chip with STI interfaces (the local interface) and a second to communicate over long distances using copper cables to another interposer chip (the cable interface). The optional cable lengths for the cable interface are:

- 2.6 m and 10 m for switch to node connections
- 5 m, 10 m, 15 m, and 20 m for switch to switch connections

Interposer chip logic: The STI link protocols and hardware macros are:

- Cable receiver macro
- Local receiver macro
- Sender macro

Cable receiver macro: The cable receiver macro de-skews the clock and data signal off of the cable interface using two operating states:

Timing mode
: Adjusts the phase of incoming data signals to compensate for system delays. In this mode, the chip can de-skew up to three bit times between any two bits. Once the receiver is timed, the macro enables the operating mode.

Operating mode
Adjusts fine delay elements to de-skew data bits and the clock without requiring a re-time. During operating mode, latency through the receive path logic is held to about eight clock cycles.

Local receiver macro: The local receiver logic aligns the clock edge with the center of the data. This latches the correct data and then deserializes the data back into a four byte flit. Data skew between bits is less than 500ρs.

Sender macro: The sender macro acts as a serializer. This macro receives four bytes of data every 8 nanoseconds and transmits one byte of data every 2 nanoseconds. Total latency (logic plus driver) is three clock cycles.

SP Switch2 Adapter

The primary components of the SP Switch2 Adapter are:

- Node Bus Adapter (NBA)
- Memory Interface Chip (MIC)
  - Rambus RDRAM
- TBIC3 switch fabric controller
- 740 PowerPC microprocessor
  - SRAM
- Self Timed Interface (STI) interposer chips

“Adapter overview” on page 8 contains background information you may need when reading this section.

Chip and driver information

Table 1. SP Switch2 Adapter chip information

Chip	Type	Core Voltage
NCD	CMOS 6S	2.5 V
NBA	CMOS 6S	2.5 V
PGA	NA	3.3 V
MIC	CMOS 5S	3.3 V
RDRAM	NA	3.3 V
TBIC3	CMOS 5S	3.3 V
740 microprocessor	CMOS 6S	2.5 V
SRAM	NA	3.3 V
Note: The NCD chip is part of the node controller and resides on the node not the adapter.		

Table 2. SP Switch2 Adapter driver information

Boundary	Driver Type	Driver Voltage
NCD to NBA	CMOS 6S	1.8 V or 2.5 V
NBA to MIC	Special CMOS 5S compatible CMOS 6S drivers	3.3 V
NBA to PGA	NA	3.3 V
MIC to RDRAM	NA	1.0 V
MIC to TBIC3	CMOS 5S	3.3 V
TBIC3 to SRAM	NA	2.5 V
TBIC3 to 740 microprocessor	Special CMOS 5S compatible CMOS 6S drivers	3.3 V
TBIC3 to STI	CMOS 5S	1.0 V

Node Bus Adapter

The SP Switch2 Adapter differs from all previous switch adapters in that the new adapter plugs directly into the 6XX bus on POWER3 SMP High Nodes. The interface for this connection is through the Node Bus Adapter (NBA) chip. The adapter interfaces with the 6XX bus through the NBA chip buffers. The NBA slave buffer receives decoded bus operations and returns read data. The NBA master buffer passes bus transfer requests and receives data from fetch requests.

The NBA chip is also segmented into two partitions, the bus interface logic (Left Hand Side or LHS) and the adapter interface logic (Right Hand Side or RHS). LHS and RHS operate synchronously to the clock with which they interface and asynchronously to each other. The asynchronous interface between LHS and RHS is performed through arrays for both master and slave operations. NBA uses full array addressing for 6XX master in order to accommodate out of order events on the 6XX bus.

Memory Interface Chip and RDRAM

Figure 11 on page 18 illustrates the flow of data through the Memory Interface Chip.

The Memory Interface Chip (MIC) passes data between the NBA and the TBIC3. The MIC also allows other adapter components to access the Rambus RDRAM memory using protocols developed by the Rambus Corporation. The adapter uses the 16 MB RDRAM memory to buffer large blocks of data between the switch fabric and main storage.

There are four paths in and out of RDRAM memory:

- NBA request
- NBA response
- TBIC3 request
- TBIC3 response

All four of these paths are buffered into the Memory Buffer Unit (MBU), pass through the Memory Protocol Unit (MPU), and then enter the Rambus ASIC Cells (RACs). The RACs synchronize and create high speed signaling for data transfers in and out of the Rambus RDRAM.

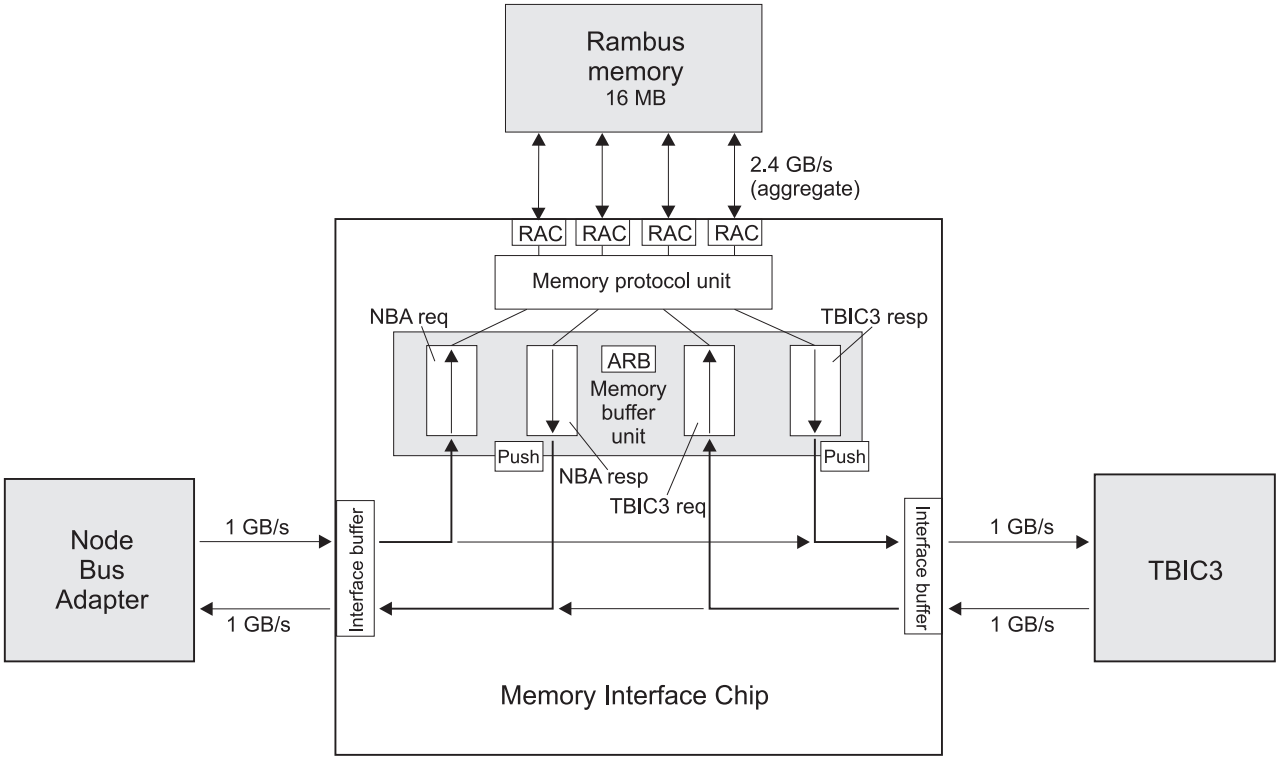

Figure 11. High level view of message passing through the MIC

The MIC also has input and output connections with both the NBA and TBIC3. Each of these four paths is capable of handling up to 1 GB/s of data transfers. When storing to Rambus, hardware flow control is used for transfers to and from Rambus. If the memory is not capable of holding the packets, the write transaction will not proceed. When this happens, the packet is buffered until memory is available.

Rambus reads can be done as either a load or a push DMA (Direct Memory Access). During a load, the data that is read out gets returned to the requester. During a push, the requester pushes the data to a different destination. Pushes are usually used with large data chunks (greater than 256 bytes).

Data flow through the MIC generally accommodates multiple parallel streams that keep load traffic independent from push traffic. Buffers are partitioned to accommodate six full push transactions and only two loads. The intent of this hardware controlled mechanism is to prevent the push stream from blocking the load stream.

Rambus™ RDRAM: The communications buffer memory uses licensed technology from Rambus Inc. that allows high speed operation over a few wires but with high bandwidth. The memory modules connect to the MIC through Rambus channels which transfer a byte of data every 1.66 nanoseconds. The 10-bit wide channels passes control, address, and data using a packet based protocol in order to move data back and forth. Each channel supports bi-directional transfer rates of up to 600 MB/s which provides a peak bandwidth of 2.4 GB/s. Total RDRAM memory capacity is 16 MB.

Rambus memory

LGS 60 nanosecond RDRAM
250/267/300 MHz
32 pin 25 mm SHP package
Total power 9.2 W
3.3 V

TBIC3 chip

The TBIC3 chip is an interface between the two STI interposer chips, a 60X bus (to the 740 microprocessor), and a high speed Interchip Bus (IB). The architecture of the adapter is similar to that of the SP Switch MX and SP Switch MX2 adapters. The primary difference is the addition of a memory control chip (MIC) and local RDRAM memory. Also, the TBIC3 control path for the 60X bus is not shared with the data path, which uses the IB bus. An exception is when control information must be passed between the 6XX memory or Rambus memory and the local 60X memory; those transactions must go over the IB bus.

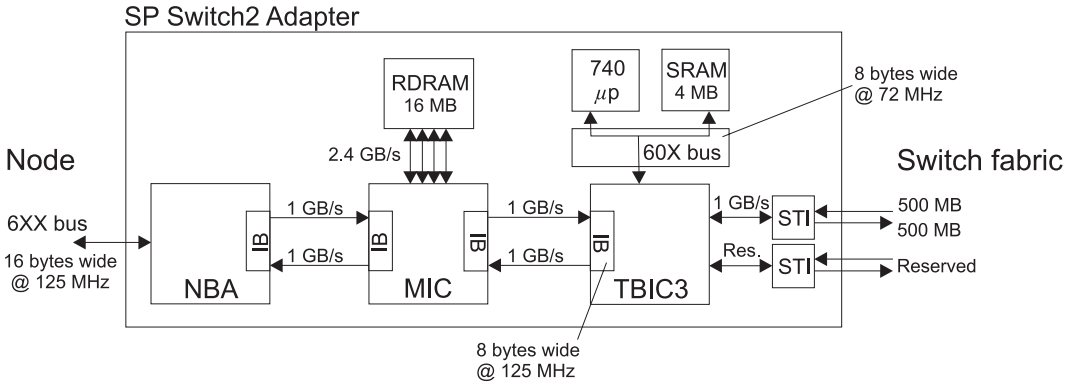

Figure 12. SP Switch2 Adapter showing TBIC3 interfaces

Sending messages: To send a message, the 6XX notifies the 740 by putting a work ticket into SRAM. When ready, the 740 reads the work ticket out from SRAM and sets up a DMA into the NBA chip and a push from the NBA into the TBIC3. When the data arrived completion occurs on the DMA, a header is written into the TBIC3 header buffer to move the data out of the TBIC3 into the switch. When the Send logic sees the header and associated data in the buffer, it pulls them out, generates CRC if necessary, and sends the packets out the STI link.

Receiving messages: When receiving a message, the receive logic checks CRC (if needed) on the incoming packet and separates the header, data, and service information into the correct queues. The reassembly logic examines the header and if:

- The header information has not been assigned a buffer in RDRAM
- The reassembly engine has been disabled
- It is a service packet

then the header is forwarded to a queue which is read by the 740 which then assigns a buffer in RDRAM.

If an RDRAM buffer has been assigned, the reassembly logic automatically sets up and executes the DMA from the TBIC3 into the assigned RDRAM buffer. When the

entire message has been received and transferred into the RDRAM, an entry is placed into a queue which is read by the 740. The 740 then requests a DMA to the node and notifies the 6XX by writing information back to the node memory indicating the message completion.

Additional functions: The TBIC3 also performs the following functions:

- Maintains a 75 MHz time-of-day counter used for application synchronization
- Implements buffer thresholding for both the send and receive buffers.
- Performs address bus snooping

  Snooping reduces the number of memory addresses that need to be read by looking for register bits indicating that address contains data.
- Segmentation and reassembly engines

  These engines break data streams into packets for transmission and reassemble them into memory when packets are received.
- Acts as a 60X bus arbitrator and SRAM controller

Comparing the TBIC3 (SP Switch2) to the TBIC2 (SP Switch):

Table 3. TBIC chip comparison

TBIC3 (SP Switch2)	TBIC2 (SP Switch)
Technology	
• CMOS 5S • 90/125 MHz Internals • 8 byte internal data paths • 12.7 mm chip • 42 mm CCGA • 507 Signal I/O used	• CMOS 5L • 50/75 MHz internals • 8 byte internal data paths • 10.90 mm chip • 32 mm CBGA • 380 Signal I/O used
60X interface	
• 72 MHz bus/468 MHz 740 PowerPC • Bus slave and master • Bus arbitration in TBIC • R/W TBIC registers from 740 • Snoop RAM accesses from node • Slave 32 or 64 bit transfers • External DMA to/from TBIC buffers with extended burst • Monitors 740 for errors • Multiple wait-state operation • R/W to/from 740 RAM from node through TBIC - single beat • Only Control reg hard reset • Error registers reset only by write xor	• 50 MHz bus/125 MHz 603e PowerPC • Bus slave • Bus arbitration in XILINX • R/W TBIC registers from 603 • Snoop RAM accesses from node • 32 or 64 bit transfers • External DMA to/from TBIC buffers with extended burst • Monitors 603 for errors • 0 or 1 wait state operation • Xilinx DMA between node and RAM • Control reg hard and soft resets • Error registers reset by reset signal/bit

Table 3. TBIC chip comparison (continued)

TBIC3 (SP Switch2)	TBIC2 (SP Switch)
Sideband interface	
• 2 x 8 bytes data paths at 125 MHz (send and receive IB Bus) • Access to all TBIC3 addresses • External DMA to send buffer, Internal DMA from receive buffer • DW or W access to registers • Complete bridge function to 60X bus • Hardware assisted message segmentation and reassembly	• 8 byte data path at 50 MHz 64 bit bidi data bus • Access only to buffers • External DMA to/from buffers • DW access on DW boundary • 603 bus access via a mailbox reg • Software message segmentation and reassembly
Fabric interface	
• 1 GB/s STI port driving 20 m • Asynchronous clocking • Adaptive or non-adaptive routing • 4 byte flit • 254 tokens • 4 byte EDC • 4 byte CRC on header and data • Special SP Switch2 packets • TB4 message packet • TB4 acknowledge packet • TBS service packets • Cable connectivity test	• 300 MB/s STI port driving 20 m • Synchronous clocking • Non-adaptive routing • 2 byte flit • 61 tokens • 2 byte EDC • 2 byte CRC • TB2 packet • TB4 message packet • TB4 acknowledge packet • TBS service packet • Special research packets
Internal buffer sizes (depth and width in bits)	
• 128 x 72: Send Header • 256 x 72: Send Data • 256 x 73: Send TBFC (x2 ports) • 256 x 73: Receive TBFC (x2 ports) • 1024 x 72: Receive Data • 128 x 72: Receive Header • 128 x 72: Receive Service • 16 x 73: Receive IB bus • 16 x 73: Master 60X Response • 16 x 72: Master 60X Request • 16 x 73: Slave 60X Request • 64 x 72: Reassembly key • 64 x 72: Reassembly data • 8 x 72: DMA request queue • 8 x 72: Receive look-ahead (x3) • 8 x 72: Segmentation route table	• 64 x 72: Send Header • 512 x 72: Send Data • 128 x 80: Send TBFC (per port) • 128 x 80: Receive TBFC (per port) • 512 x 72: Receive Data • 64 x 72: Receive Header • 64 x 72: Receive Service

Table 3. TBIC chip comparison (continued)

TBIC3 (SP Switch2)	TBIC2 (SP Switch)
DMA operation	
• External DMA to TBIC buffers • Internal DMA engine from TBIC buffers • Hardware flow control with Rambus/BiFIFO – BiFIFO transfers require whole packet – To Rambus as packet is received – From Rambus as there is space	• External DMA To TBIC buffers • External DMA From TBIC buffers • Software flow control with BiFIFO: – Source must contain whole packet – Destination must have space • Complete memory to BiFIFO before transmission to TBIC • Complete fabric to TBIC before transmission to BiFIFO

740 PowerPC microprocessor and SRAM

Microcode on the 740 microprocessor is responsible for:

- Passing control information between the adapter and node software
- Formatting or decoding packet headers
- Building packet routing information

This requires the microcode to handle multiple traffic streams for each open communication domain. To accomplish this, the 740 microcode and node software work together to present an IP, MPI, or LAPI interface to the user and to share the adapter among several independent node processes. Node software is responsible for initializing the adapter and for allowing access to main storage.

Control processor

PowerPC 740
468 MHz internal operations
72 MHz external 60X bus operations
255 pin 21 mm BGA package
2.5 V core and 3.3 V I/O's

SRAM: The control memory is high speed synchronous burst SRAM. This memory holds:

- Control microcode
- Several control and message structures
- A number of interface structures.

Message data is not held in this memory since that is the primary task of the Rambus RDRAM memory.

Control memory

Pipelined Burst SRAM
72 MHz clock
3.3 V core, 2.5 V I/O's
Power consumption about 1 watt
100 pin, TQFP package

STI interposer chips

The SP Switch2 Adapter uses two STI chips that provide the interface to the switch fabric. The Self Timed Interface chips used on the adapter are identical to the STI chip used in the switch interposer card. For more information, refer to “Interposer chip” on page 15.

Note: Only one of the two STI interposer chips on the adapter card is supported. The second chip and the associated switch port are reserved.

Flow control and clocking

Flow control

The SP Switch2 Adapter uses various levels of flow control that either make it easier for the microcode to coordinate operations or relieve the microcode of managing some operations entirely. The first three mechanisms listed are hardware based; microcode has no involvement in their operation. The last four are hardware capabilities that allow the microcode to control the flow of operations relative to higher level operations within the code. The flow control mechanisms are:

Interface handshaking
: Handshaking signals indicate the sender has valid send data and the receiver is ready to accept it or has already accepted it using VALID and READY signals. These type of interfaces are found at the FIFO buffers, the IB macros, and the internal bus switching within the MIC.

Token counting
: Tokens are used to indicate an amount of space at the receiver. These counts are incremented and decremented, usually on a cycle by cycle basis, and are constantly communicated back and forth between the sender and receiver. With this mechanism, the sender maintains count of the number of free slots available in the receive buffer.

Almost full checking
: Almost full indicators are signals produced by the major receive buffers that indicate when a predefined amount of space is left available. These signals are used by the Push engine moving data across IB buses so that the receiving end will have enough room to absorb an entire transfer, before the transfer is started. This prevents bottlenecks at the IBs due to Pushes that cannot progress.

Outstanding Rambus requests
: The MIC has two streams of dataflow. One stream is made up of loads and stores, and the other stream is made up of Pushes out of Rambus. Although the MIC is designed to keep these streams separate, the buffer within the MIC (the MPU) merge both the load and store stream and the Push stream. This means that until the data streams leave the buffer and resume independent paths, there can be interference. Because the MIC uses a sequence number for each request sent to Rambus, the MIC can control the number of outstanding requests at any given time.

Buffer and queue counts
: Buffer and queue counts allow microcode to read the available space left in various queues or buffers. Using these counts, operations don’t overutilize or underutilize critical adapter resources. If these values are not used and overflowing or underflowing occurs, no data will be lost but status bits will indicate system operation is less than optimal.

Push completion reports

The completion report is used to synchronize the end of one operation with the start of another. Completion reports automatically generate an indicator when a Push operation completes. This can be used to manage the number of outstanding Pushes in progress and maintain ordering between sequential Push operations.

Ordering and synchronization

There are a number of ordering and synchronization methods possible with the SP Switch2 Adapter. These functions implement various types of high level microcode flow control.

Adapter clocking

The SP Switch2 Adapter has five internal clock domains (originating on the adapter) and two external domains.

Internal domains

Rambus domain

Based on two 14 MHz on-card crystals which generate two, 300 MHz clocks that are fed separately onto all four Rambus channels. Operations occur at twice the frequency (600 MHz) by using both edges of the base oscillator. The RAC on the MIC takes this oscillator and divides it by 4 to 75 MHz for the internal logic and calls it SYNCLK. The four RAC's develop four individual SYNCLK's which are independent but close in phase and frequency.

Processor domain

Contains the 740 microprocessor, the SRAM and a portion of the TBIC3. This is all based on an on-card 72 MHz oscillator which drives the 740 at 468 MHz and drives the bus, SRAM and TBIC3 logic at 72 MHz. Both parts of this domain are synchronous with each other. The processor domain borders the base domain through the TBIC3's master and slave buffers.

Base domain

Covers the bulk of the adapter logic and controls any portions of the chips that don't fall into the first two domains. This domain is sourced by a single 63.26 MHz on-card oscillator and is brought onto the chips through the PLL's where it is multiplied up to 126 MHz. These clocks can be shut off using an external module pin.

TOD domain

Driven by an on-card 75 MHz oscillator that is fed directly to the TBIC3 for incrementing the on-chip timebase (time-of-day clock).

PGA domain

Handles the JTAG interface and test mode controls is based on a 20 MHz on-card oscillator.

External domains

6XX bus domain

Contains the portion of the NBA chip that communicates across the 6XX bus and is bordered by the bus interface logic (LHS) master and slave arrays. Everything within this domain runs with the same frequency (111 to 125 MHz) and phase. The source of this clock is the 6XX bus.

Switch domain

Contains part of the TBIC3 and part of the STI interposer chip. This domain is driven by a 250 MHz clock that may be slightly higher or lower than the clock on the base domain. Operations within this domain happen either at 500 MHz on the STI wires or at 125 MHz within the STI macros. The boundary for this domain is within the STI macros. The 250 MHz source clock for this comes over the link from the Interposer chip.

Node bus design

SP Switch2 installations are limited to systems using 375 MHz POWER3 SMP High Nodes and POWER3 SMP High Nodes. The I/O planar in these nodes has two slots, J2 and J3 (correspondingly labeled W3 and W1 on the rear of the node), that are dedicated to SP Switch2 Adapters. Although both slots are fully functional, only one adapter may be installed at this time. The J4 slot (correspondingly labeled W2 on the rear of the node) is dedicated to the SP Switch MX2 adapter used with the older SP Switches, however the two switch types cannot be mixed. Refer to Figure 13.

As delivered from the factory, all adapters will be installed in the same node slot. In order to maintain consistency in your system configuration, you should maintain a similar strategy. Once an adapter is placed in the node and PSSP is initialized, the adapter is recognized and assigned a node number. If the adapter is removed from the node and placed back in the wrong slot, you will get system errors.

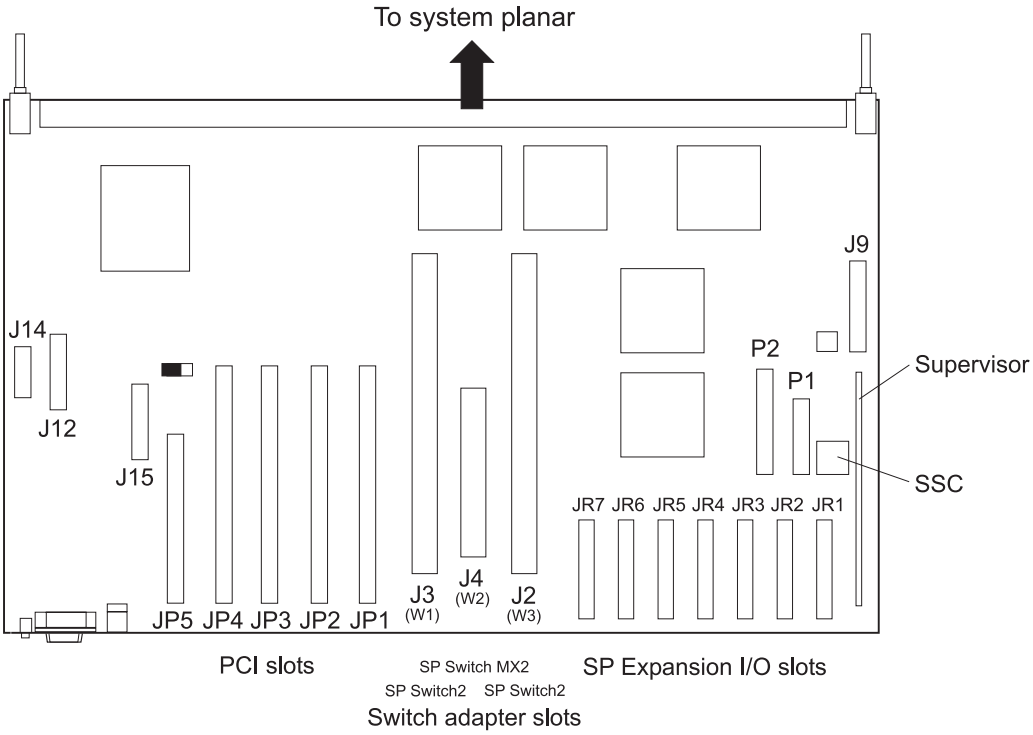

Figure 13. POWER3 SMP High Node I/O planar

The difference between the SP Switch2 slots and the SP Switch MX2 slot lies in the system architecture. The older SP Switch MX2 adapters share a 500 MB/s I/O bus with:

- The internal SCSI bus

- Internal PCI slots
- Ethernet
- Serial ports

The performance of the SP Switch2 is enhanced since its data is transferred directly through the node controllers (and therefore direct to the node's main memory) over dedicated bi-directional (simultaneous transfers in both directions) connections with the switch. These connections offer an aggregate hardware memory bus bandwidth of 2 GB/s to the I/O planar. Refer to Figure 14.

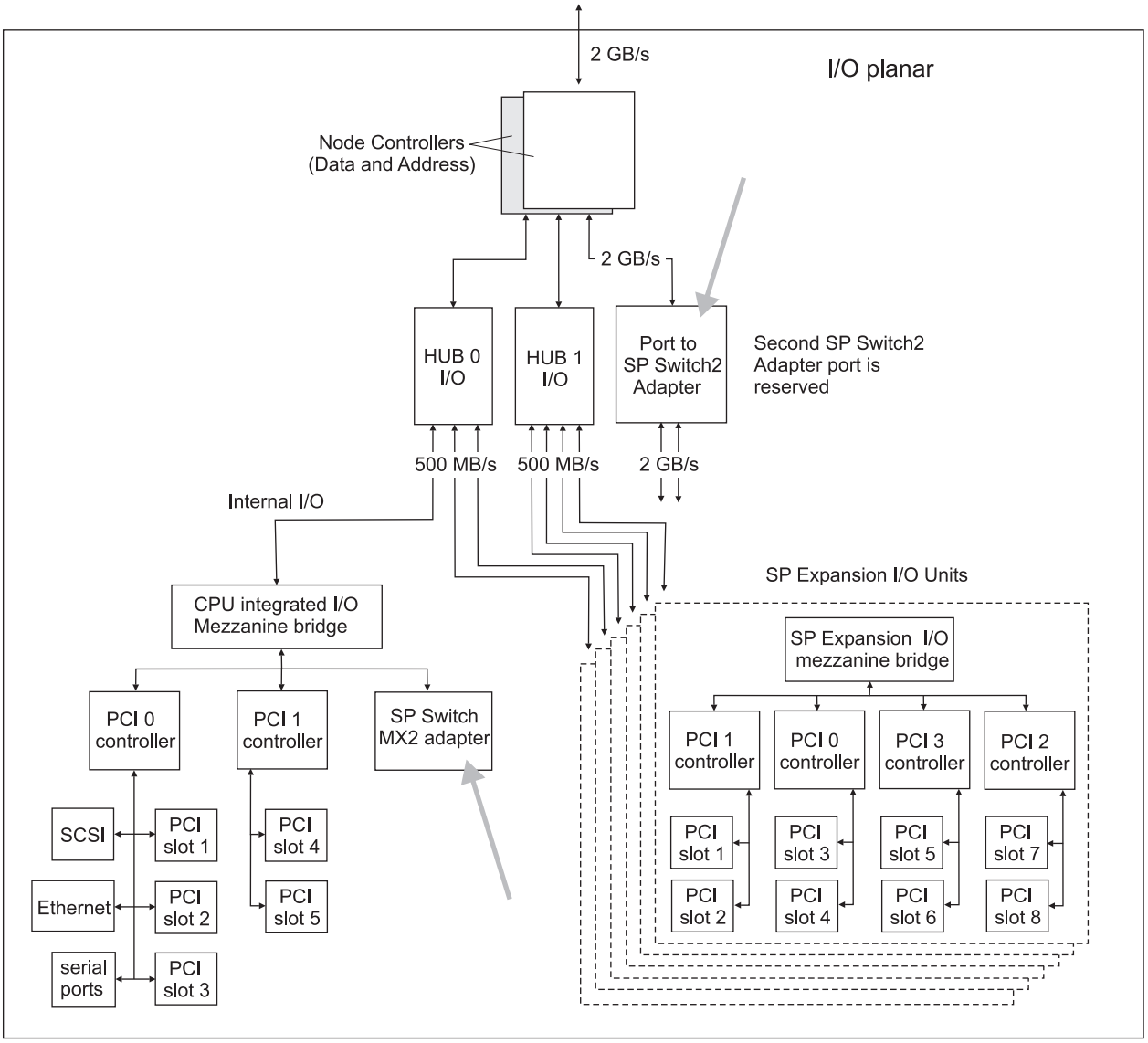

Figure 14. 375 MHz POWER3 SMP High Node and POWER3 SMP High Node block diagram

Network topology

The network topology for the SP Switch2 is essentially the same as the SP Switch. SP Switch2 adapters in either a 375 Mhz POWER3 SMP High Node or the (222 Mhz) POWER3 SMP High Node connect the nodes to the SP Switch2 and switch boards are interconnected like SP Switch boards (see Figure 3 on page 5). Both switch subsystems use the same cables, however the two switch types cannot be configured into the same system.

Switch cabling

"Appendix A. Switch connections" on page 35 contains a quick overview of switch cabling.

For detailed information on configuring switch networks, refer to RS/6000 SP Planning Volume 1, Hardware and Physical Environment (GA22-7280) located at the following URL:
http://www.rs6000.ibm.com/resource/aix_resource/sp_books/planning/index.html

Hardware diagnostics

The SP Switch2 subsystem improves system Reliability, Availability, and Serviceability (RAS) by adding a number of new hardware diagnostic capabilities including:

- Interposer hot-plugging
- An interposer wrap card that temporarily substitutes for a switch interposer
- Nodes can now be plugged into **any** available node port on the switch
- Switches can now be connected through **any** available switch-to-switch port

Note: Additional RAS details are available in "Appendix B. SP Switch2 Reliability, Availability, and Serviceability (RAS) enhancements" on page 37.

Interposer hot-plugging

As a diagnostic aid, interposer hot-plugging allows you to remove an SP Switch2 interposer without powering down or otherwise halting switch operations. By replacing a suspect interposer, you can identify or eliminate the interposer as the problem source.

Hot-plugging SP Switch2 interposer cards also works in conjunction with wrap testing the switch port and switch cables to eliminate those items as a possible problem.

Interposer wrap card

The interposer wrap card is an interposer card without switch ports. It can be hot-plugged into the SP Switch2 to test interposer function and help detect possible planar failures.

Node plugging

With the SP Switch2, nodes can be plugged into **any** available node port on the switch. This allows you to localize subsystem failures or to reroute a node to a working circuit should a failure occur.

Switch plugging

With the SP Switch2, switch-to-switch connections can be made through **any** available switch port on the switch. This allows you to localize subsystem failures or to reroute a node to a working circuit should a failure occur.

Chapter 3. Software

Software enhancements

Several software components have been enhanced to support new functions available with the SP Switch2 and SP Switch2 Adapter including:

- The Perspectives interface
- Switch management
- Software diagnostics
- KLAPI

Perspectives interface

The Perspectives interface includes the following updates for the SP Switch2:

- System Notebook: Identification Page shows the number of switch planes in use
- Switch Notebook:
  - Configuration Page
    - Switch name displays as SP_Switch2
    - Switch supervisor type defined as 132
  - Switch Notebook Status Page
    - Shows the status of the four switch power supplies
    - Shows the status of the four cooling fans
- Node Notebook:
  - Status Page
    - Need to monitor ″switchResponds0″
    - ″switchResponds″ attribute only applies to SP Switch (shows but does not apply to SP Switch2)
  - Configuration Page
    - Switch port number is no longer displayed (moved to SP Adapter page)
    - Switch chip is no longer displayed (moved to SP Adapter page)
    - Switch chip port is no longer displayed (moved to SP Adapter page)

– SP Adapter Page
  - This is a new page listing Ethernet, CSS (switch information), Token Ring, and FDDI information
  - CSS attributes for SP Switch2 include:
    Adapter
    Adapter port
    Switch port number
    Network address
    Netmask
    Subnet
    Switch chip
    Switch chip port
    Switch number
    Configuration status

Note: Even though an SP Switch2 configured system cannot be partitioned, the System Partitioning Aid remains visible on Perspectives. If you attempt to use this function, a message box stating ″Partitioning not supported″ opens, and the System Partitioning Aid window closes.

Switch management

Threads

Node threads

Node threads handle work requests (such as Ecommands) from interface routines and then direct these requests to the appropriate adapter work request queue. The node thread also manages daemon initialization as well as any synchronization or shutdown that may be necessary.

Adapter service threads

Adapter service threads are spawned for each installed adapter. The thread's function are associated with the associated adapter and include:

- Receiving service packets from the adapter via HAL (Hardware Abstraction Layer) and directing them to the correct port management thread
- Recovering from local adapter errors
- Passing work requests from the adapter's queue to the targeted port management thread's work request queue
- Merging the individual route tables of the port management threads and downloading the combined route table to the adapter

Port management threads

Port management threads are spawned as necessary for each port on installed the SP Switch2 Adapters. The port management threads perform the core functions of the SP fault service daemon including:

- Switch fabric initialization
- Switch fault recovery
- Route table generation
- Switch connectivity status update

Switch administration daemon

The switch administration daemon is enhanced in SP Switch2 systems to provide Node and Global recovery. If the switch is started without a switch Primary node, the switch administration daemon will select a new switch Primary (or Backup) and then run Estart.

Switch installation and configuration

Switch Management software provides enhanced installation and configuration features for SP Switch2 systems. These enhancements use a ″Discovery Node″ to run a switch initialization (Estart) checkout that:

- Calls out mis-wired switch to switch connections
- Automatically builds the node portion of the switch topology (out.top) file

When the switch completes the initialization process, the node information is distributed to all nodes attached to the switch plane.

Logging

Event and error logging have been restructured for improved usability. With this new structure, there is now little or no file sharing between threads.

- Node level files (***/var/adm/SPlogs/css*** directory)
  - daemon.log
  - rc.switch.log
  - css.snap.log
  - Ecommands.log
  - logevnt.out
  - daemon.log (contains log of node thread activity)
- Adapter level files (***/var/adm/SPlogs/css0*** directory)
  - adapter.log
  - dtbx.trace
  - dtbx_failed.trace
  - scan_out.log
  - scan_save.log
- Port level files (***/var/adm/SPlogs/css0/p0*** directory)
  - flt
  - fs_daemon_print.file (also contains information that was in worm.trace file)
  - out.top
  - router.log
  - topology.data
  - cable_miswire

Software diagnostics

The SP Switch2 subsystem improves system Reliability, Availability, and Serviceability (RAS) by adding a number of new software diagnostic capabilities including:

- Node location
- Built-in port and cable wrap tests
- Built-in and functional self-tests

Software

- Differential line testing
- Mis-wire detection

Note: Additional RAS details are available in “Appendix B. SP Switch2 Reliability, Availability, and Serviceability (RAS) enhancements” on page 37.

Node location

The switch management software for the SP Switch2 allows greater flexibility with respect to node connections to the switch fabric. With prior switches, nodes were restricted to specific slots. With the SP Switch2, nodes can be placed into any valid slot. During switch initialization, the switch management software detects the node and updates system configuration information.

Built-in wrap tests

There are two built-in wrap tests for the SP Switch2 that run during diagnostics:

Local Wrap Mode
Both data and clocks from the **local** STI Receiver are wrapped back inside the chip to the local STI Sender. Placing the Local Wrap signal active low enables the interposer chip to act as a local wrap plug.

Cable Wrap Mode
Data from the **cable** STI Receiver and **incoming** cable clock are wrapped back inside the chip to the Cable STI Sender port. Placing the Cable Wrap signal active low enables the interposer chip to act as a remote wrap-plug for the chip on the other end of the cable.

Built-in and functional self-tests

The SP Switch2 contains two test capabilities that are controlled completely through the Supervisor interface:

Built-in self-tests
The built-in self-tests (BIST) are initiated automatically during the power on sequence. The self-test scans data into all pattern generator latches, then clocks the data and scans the results into the MISR. When the process completes, the BIST repeats the procedure.

The logic also runs the array self-test (ABIST). The results of the logic self-test (LBIST) and array self-test are combined to create a signature that is returned in all Error and Status packets. These tests do not test the logic at speed, nor do they test any chip I/Os. However, they do test the error checking logic which is not tested by the functional self-test.

Functional self-tests
During diagnostics, the functional self-tests stress the switch chips by exercising the switch board at speed with as much data contention (packet interaction) as possible. This is done by writing packet data into the receivers through the supervisor interface. To run the functional self-tests, the switch ports need to be wrapped. When the run test bits are set, the receivers see waiting packets and transmit the packets. The supervisor watches for errors and monitors the registers to confirm correct operation.

Differential line testing

The SP Switch2 contains logic to test for open circuits and high impedance on the STI network. Differential line testing also tests for weak differential drivers by

pushing the driver to a specific value. The receiver then adds a significant load to the signal. If there is a broken single leg of a differential, the receiver will receive the opposite value.

Issuing the **cable-test** command on the control workstation initiates differential line testing. When this test is running, the supervisors of both switches write to the switch chips with a command to place the appropriate ports into test mode. The interposers activate their test drivers and sample the outputs of the differential receivers. After the link test, the register are checked for errors and the switch chips are returned to normal operation.

Differential line tests will need to be run multiple times to isolate the exact failing component. For example, run with a wrap plug on the switch board to isolate interposer problems and run with a wrap plug at the end of the cable to try to detect cable problems.

Mis-wire detection

Switch initialization (Estart) triggers the mis-wire detection sequence. Mis-wire detection is possible because the SP Switch2 does not use the SDR topology file to label switch chip IDs. Instead, the SP Switch2 system management software initializes the switch chips before initializing the rest of the network. This enhancement allows the switch chip initialization to set the frame and switch location and assign an ID for each chip. Any connectivity information (what devices are connected and where they are connected) is discovered by the switch management software during exploration. If the discovered connection does not match the topology file, the switch management software can accurately point out the endpoints of the mis-wire.

Note: Chip and switch initialization occurs every time the SP Switch2 is powered on.

KLAPI

The KLAPI (Kernel Low level Application Interface) subsystem is available for use by kernel subsystems (currently being used by VSD). KLAPI takes advantage of the SP Switch2 Adapter to provide an efficient and reliable communication protocol. The SP Switch2 Adapter supports DMA (Direct Memory Access) engines that are used by KLAPI to avoid copies by the VSD kernel subsystem from the I/O buffers into the network buffers or vice versa. Using these copy avoidance mechanisms, VSD/KLAPI achieve ″zero copy transport.″

Using zero copy transport to avoid copies enhances overall system performance by:

1. Freeing the CPU from performing copy tasks and allowing the CPU to be available for other applications.
2. Eliminating copy tasks on the node reduces memory bandwidth loading and therefore increasing the memory bandwidth available to other applications.

In order to implement copy avoidance, KLAPI takes advantage of the SP Switch2 Adapter's on-board RDRAM (Rambus) memory. The adapter has facilities for hardware segmentation and reassembly of large datagrams to maximize communication efficiency. KLAPI uses the extra memory available on the adapter (in RDRAM) to stage the incoming data associated with zero-copy messages.

Any coordination required for datagram reassembly is done locally on the receiving node and does not involve extra messages between the sender and receiver. This is done by using the RDRAM to stage zero-copy messages on the receiver until the application can determine the target for the data. By using the RDRAM to stage data, other network traffic is not affected and the SP Switch2 adapter can continue processing incoming and outgoing messages without having to wait for a response from the host. The use of RDRAM helps avoid an extra message while maintaining maximum parallelism in the adapter to achieve full throughput.

The KLAPI subsystem permits flow control on messages sent using KLAPI. Lack of flow control may cause switch congestion which will degrade overall system performance. For example, earlier systems that did not have KLAPI available had instances where high loads (more than the switch bandwidth capability) between VSD servers and GPFS clients were adversely affected by the lack of flow control. The flow control on KLAPI is designed so that applications such as GPFS which use VSD experience fewer switch congestion problems.

Appendix A. Switch connections

Definitions

Switch-equipped frame
Frame containing nodes and one switch.

Non-switched expansion frame
In an SP system with a switch, this is a frame containing nodes (only) that are connected to a switch in a switch-equipped frame.

Dedicated switch frame
Frame containing either four or eight switches (nodes not permitted). All switches in this frame are connected to switches in switch-equipped frames.

The SP Switch2 dedicated switch frame (FC 2032) is the same physical frame as the 1.93 m tall frame. FC 2032 comes with four second stage switches and 132 SP Switch2 interposers. Dedicated switch frames are required for SP systems having from 65 to 128 nodes. Additional second stage switches can be added for use in special-order SP systems having more than 128 nodes.

Note: For SP systems with less than 65 nodes but anticipated growth to over 65 nodes, using a dedicated switch frame can greatly simplify adding switches. Refer to RS/6000 SP: Planning Volume 1, Hardware and Physical Environment (GA22-7280) for details.

Single-stage configuration
All switches are located in switch-equipped frames. Switch connections are a combination of node-to-switch connections and switch-to-switch connections.

Two-stage configuration
Some switches (first stage) are located in switch-equipped frames and other switches (second stage) are located in dedicated switch frames. First stage switch connections are node-to-switch connections. Second stage switch connections are switch-to-switch connections. The two-stage configuration is not required, but is recommended if the system is expected to grow beyond five switches.

Systems with one switch

Systems with one switch illustrate the basic single-stage switch network. All nodes in these systems are connected through the 16 node-to-switch ports. These nodes may be in the switch-equipped frame or in non-switched expansion frames. SP systems with one SP Switch2 do not use the 16 ports reserved for switch-to-switch connections.

Systems with two to five switches

SP systems configured with two to five switches are connected with each other in either a single-stage or two-stage configuration.

Single-stage configuration
In a single-stage configuration, every node has a node-to-switch connection and all switches are directly connected to each other through switch-to-switch connections.

Two-stage configuration

In a two-stage configuration, every node has a node-to-switch connection (first stage connection). Each first stage switch is then connected to switches in a dedicated switch frame (second stage connection). This simplifies system expansion by centralizing switch-to-switch connections at the dedicated switch frame.

In either configuration, some or all of the node-to-switch ports are used. The number of node-to-switch ports used depends on the number of nodes in the SP system and their configuration. All of the switch-to-switch ports are used on each switch; this provides redundant paths between switches. Refer to RS/6000 SP: Planning Volume 1, Hardware and Physical Environment (GA22-7280) for detailed descriptions and illustration of these configurations.

Systems with six or more switches

SP systems using six or more switches require a two-stage configuration. This ensures that the SP design objectives for performance (such as bi-sectional bandwidth) and availability (such as a minimum of four paths between any two nodes) are met. In a two-stage configuration, all nodes are connected to the first stage switches and the first stage switches are connected to each other through the second stage switches.

Appendix B. SP Switch2 Reliability, Availability, and Serviceability (RAS) enhancements

Design improvements in the SP Switch2 have made the new switch extremely reliable and produced an interconnect fabric that will continue to operate with most component failures. In addition to the Reliability, Availability, and Serviceability (RAS) improvements mentioned throughout this paper, the following sections describe additional usability and system level RAS enhancements incorporated into the SP Switch2 design. These enhancements include:

- Switch management simplification
- Switch subsystem fault tolerance
- System diagnostics improvements
- System problem determination

Switch management simplification

The SP Switch2 subsystem has the following characteristics:

- N+1 redundancy on the switch power supply. Should any single module in the power supply fail, the power supply has sufficient capacity to maintain switch operation.
- If a power supply does fail, the replacement can be hot-plugged (installed concurrent with system operation).
- The SP Switch2 supervisor card, all interposer cards, and switch cables are hot-pluggable.
- Intermediate switch boards (for switch-to-switch connections) are configured with redundancy allowing full system operation and concurrent repairs should a switch board fail.
- If a primary switch board (for node-to-switch connections) should fail, only the nodes directly attached to that switch are affected.
- If a switch adapter (mounted in the node) should fail, only the node in which the adapter is mounted will be affected.
- Event Management provides a summary of switch errors that is available as a log file on the Control Workstation.
- Code patches for the SP Switch2 supervisor card and switch adapters can be installed online.
- Each switch chip in the SP Switch2 has its own oscillator. This eliminates the need for external clock sources and the Eclock command. Although external clocking has been eliminated, each oscillator is closely synchronized to all other oscillators (plesiochronous clocking) which prevents clock tree failures.

Switch subsystem fault tolerance

One of the design goals for the SP Switch2 was to set up mechanisms that would detect errors that might cause data corruption or that would stop data flow. To move toward this goal, the following mechanisms were improved for the SP Switch2:

- Error Correction Circuitry (ECC)

  ECC has been incorporated into the design of the Rambus Memory on the SP Switch2 adapter card. ECC uses an encoding scheme and redundant memory bits to correct both hard and soft memory failures. A hard memory failure occurs when a memory bit or group of bits can no longer be written

correctly with data. A soft memory failure occurs when data read from memory is wrong, but the memory location can be rewritten correctly with new data. Using ECC, the SP Switch2 Adapter enables the memory to continue operating in the presence of both single-bit soft and hard memory failures and prevents those failures from causing system outages. In contrast, most computer memory systems implementing ECC detect single and double bit errors but only correct single bit errors.

- Parity checking

  Parity checking determines the number of ″1's″ or ″0's″ in a data stream prior to sending it to the next logic stage. When the data is sent, a ″parity bit″ is also sent which indicates if there is an even or odd number of ″1's″ or ″0's″ in the data stream. When the data reaches the next logic stage, the system regenerates the parity data and checks that parity data against the parity bit sent with the data stream. This check verifies that bits have not changed. Parity checking has also been added on all data paths through the switch chip including: all counters, all linked list pointers in the central queue, and on all arrays.

- Cyclic Redundancy Checks (CRCs)

  Link CRCs are used on the SP Switch2 subsystem to detect link errors. CRCs are more complicated than a simple parity check. CRCs use a hashing code generated on the data and several bits are sent (instead of a single bit parity check) to represent an odd or even count of bits.

- Switch fencing

  Using the SP Switch2 Adapter, nodes attached to the SP Switch2 can be unfenced even during correctable transient port errors.

System diagnostics improvements

A fault service daemon automatically monitors failures on the SP Switch2. In addition, error logging and recovery on the SP Switch2 Adapter provide reliability and availability enhancements over the SP Switch adapter. On the SP Switch, permanent adapter errors required administrative intervention to reset the SP Switch adapter and exit of the fault service daemon. With the SP Switch2 Adapter, this is no longer true. The SP Switch2 Adapter automatically resets during permanent adapter errors, and does not require administrative intervention. Also, a new level of errors that are recoverable without resetting the adapter have been defined. These errors, called critical adapter errors, enable recovery from a large range of errors. The SP Switch2 Adapter also employs bit tunable thresholds on recoverable errors. This allows two things:

- Better recovery for non-critical, transient errors
- Fault isolation for critical but transient adapter errors

The SP Switch2 Adapter has improved definitions for:

- Error classification types
- Error sources
- Bit descriptions for system error logging

In addition, label improvements were added to the following recoverable errors:

- Hardware
- Microcode
- Threshold
- Bad packet

- Transient
- Service queue full

This reduces the time to isolate a failure, and increases adapter availability.

System problem determination

Improvements to the SP Switch2 provide thourough mis-wire detection. The system now calls out mis-wired links and cables, and gives visual aids to the field engineer in solving these problems. Additional features were added to help resolve switch subsystem software problems. These include:

- A supervisor location register is used to alert the software to mis-wires and helps determine mis-wire locations.
- The ″Return on Same Route Option″ provides a response even on a mis-wire.
- The ″Port Received on Register″ function specifies the switch port that received a particular service.
- Sender hang detect logic determines when a sender is unable to make progress and re-times the link.

Error detection and fault isolation have also been enhanced by using the SP Switch2's hot-plug feature to test switch interposers, interposer connectors, and switch cables. This is done using a wrap card and wrap plug included with the switch. The wrap plug can test the cables, and isolate a fault in the cable. If the wrap plug does not identify a fault with the cable, then the wrap card can isolate the fault to either a faulty interposer or the switch planar.

Appendix C. Component feature codes

SP Switch2 components can be ordered using the feature codes listed in Table 4.

Table 4. SP Switch2 component feature codes

SP Switch2 component	Feature code
SP Switch2	4012
SP Switch2 Adapter	4025
SP Switch2 interposer	4032
SP Switch2 blank interposer	9883
SP Switch2 frame	2032

Notices

IBM Corporation
Marketing Communications
Server Group
Route 100
Somers, NY 10589

Produced in the United States of America
03-01 All Rights Reserved

More details on IBM UNIX hardware, software and solutions may be found at ibm.com/servers/unix/

You can find notices, including applicable legal information, trademark attribution, and notes on benchmark and performance at ibm.com/rs6000/hardware/specnote.html

AIX, RS/6000, and SP are registered trademarks or trademarks of the International Business Machines Corporation in the United States and/or other countries.

Other company, product and service names may be trademarks or service marks of others.

IBM may not offer the products, programs, services or features discussed herein in other countries, and the information may be subject to change without notice.

General availability may vary by geography.

IBM hardware products are manufactured from new parts, or new and used parts. Regardless, our warranty terms apply.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Any reliance on these statements is at the relying party's sole risk and will not create any liability or obligation for IBM.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Any performance data contained in this document was determined in a controlled environment. Results obtained in other operating environments may vary significantly.