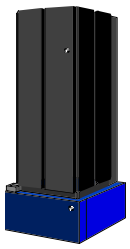


SP POWER3 SMP Node
System Architecture

February 1999

Version 1

G. Mojtabaeezamani (Kazem)

RS/6000
SP Node Hardware Development
Poughkeepsie, NY

Table of contents

Preface

This white paper describes RS/6000® SP™ POWER3 Symmetric Multiprocessor (SMP) Node System Architecture and its major components: System Memory, I/O subsystem, PCI bridge chips, Service Processor, System Firmware, and Packaging.

Acknowledgments

The author wishes to thank Frank May, Bill Mihaltse, Mark Atkins, and Russell Bistline for their contributions to this paper.

Introduction

The SP POWER3 SMP Node is the first scalable processor node which utilizes the POWER3 (630FP) 64-bit microprocessor. The floating-point performance of the POWER3 microprocessor makes this node an excellent platform for compute-intensive analysis applications. The POWER3 microprocessor offers technical leadership for floating-point applications by integrating 2 floating- point, 3 fixed-point, and 2 load/store units in a single 64-bit PowerPC implementation. Since the node conforms to the RS/6000 Platform Architecture, compatibility is maintained for existing device drivers, other subsystems, and applications. The POWER3 SMP Node supports IBM's AIX® operating system, beginning with version 4.3.2.

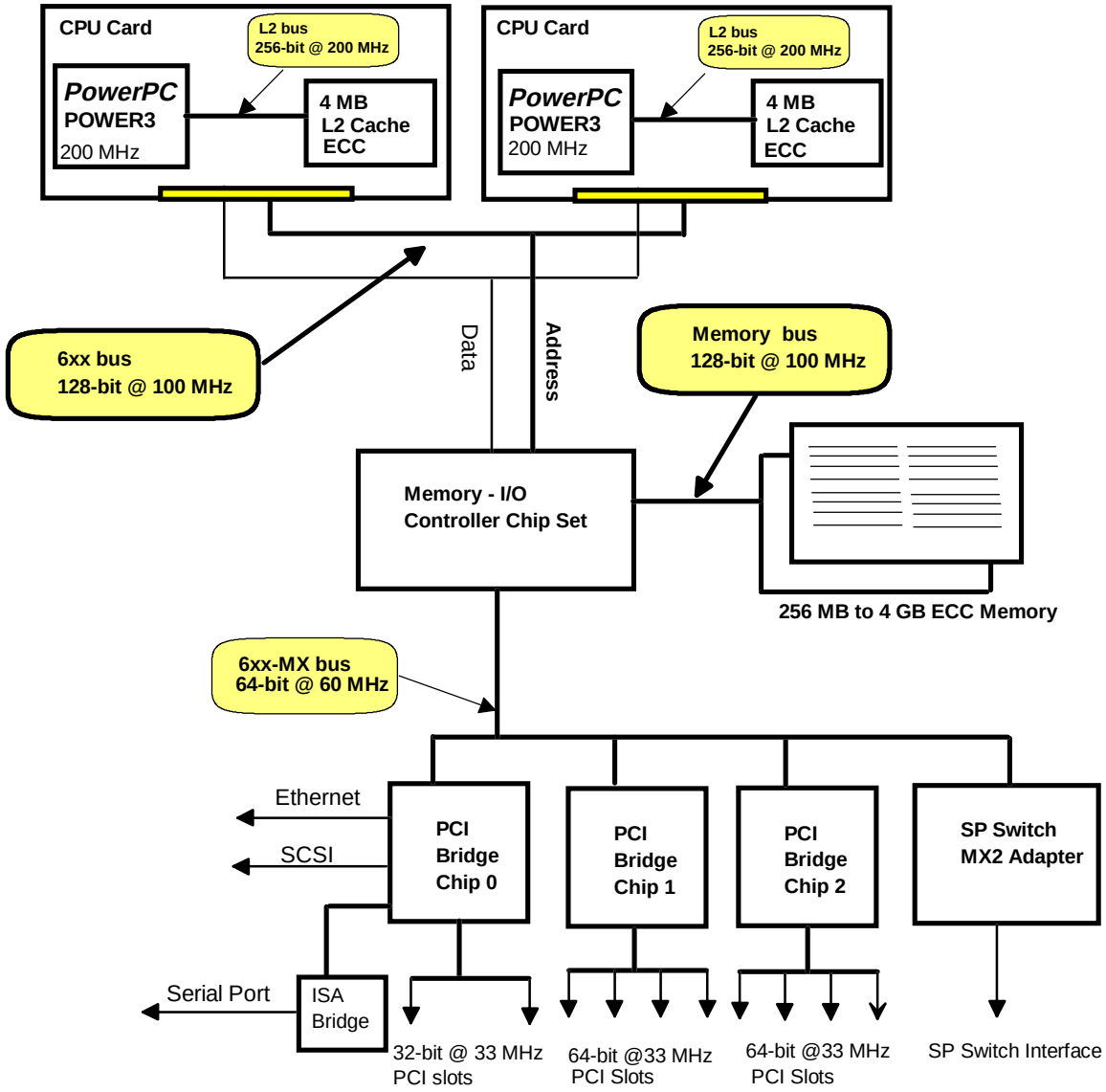
The POWER3 SMP Node is available in two packages: thin and wide. The thin node can accommodate up to two processor cards, two memory cards, and two PCI adapters. It also supports two Ultra SCSI hard files. The node is fully compliant with Revision 2.1 of the Peripheral Component Interconnect (PCI) specifications and implements three PCI buses. The first PCI bus, found in both thin and wide nodes, supports two 32-bit PCI slots running at 33 MHz. Integrated Ultra2 SCSI, Ethernet®, and ISA bridge are also supported by the first PCI bridge chip. The other two PCI buses are located in the expansion I/O unit of the wide node. Each of these buses support four 64-bit PCI slots running at 33 MHz.

SP POWER3 SMP Node System Architecture

SP POWER3 SMP Node system design is based on the IBM PowerPC Architecture and the RS/6000 Platform Architecture. The node is designed as a bus-based symmetrical multiprocessor (SMP) system, using a 64-bit address and a 128-bit data system bus running at a 2:1 processor clock ratio. Attached to the system bus (6xx bus) are from 1 to 2 PowerPC 630 microprocessors, and a two chip memory-I/O controller.

The Memory-I/O controller is a general purpose chip set which controls memory and I/O for systems such as the POWER3 SMP Node which implement the PowerPC MP System Bus (6xx bus). This chip set consists of two semi-custom CMOS chips, one for address and control, and one for data flow. The memory-I/O controller chip set includes an independent, separately-clocked "mezzanine" bus (6xx-MX bus) to which three PCI bridge chips and the SP Switch MX2 Adapter are attached. The POWER3 SMP system architecture partitions all the system logic into the high speed processor-memory portion and to the lower speed I/O portion. This design methodology removes electrical loading from the wide, high-speed processor-memory bus (6xx bus) allowing this bus to run much faster. The wide, high-speed 6xx bus reduces memory and intervention latency while the separate I/O bridge bus supports memory coherent I/O bridges on a narrower, more cost-effective bus.

Figure 1. POWER3 SMP System Architectecture block diagram



POWER3 Microprocessor

The POWER3 design contains a superscalar core which is comprised of eight execution units, supported by a high bandwidth memory interface capable of performing four floating-point operations per clock cycle. The POWER3 design allows concurrent operation of fixed-point, load/store, branch, and floating-point instructions. There is a 32 KB instruction and 64 KB data level 1 cache integrated within a single chip in .25 um CMOS technology. Both instruction and data caches are parity protected. The level 2 cache controller is integrated into the POWER3 microprocessor with the data arrays and directory being implemented with external SRAM modules. The POWER3 microprocessor has a dedicated external interface (separate from 6xx bus interface) for the level 2 cache accesses. Access to the 6xx bus and the level 2 cache can occur simultaneously. The level 2 cache is a unified cache (i.e. it holds both instruction and data.), and is configured for direct mapped configuration. The external interface to the 4 MB of level 2 cache has 256-bit width and operates at 200 MHz. This interface is ECC protected. The POWER3 microprocessor is designed to provide high performance floating-point computation. There are two floating-point execution units, each supporting 3-cycle latency, 1-cycle throughput, and double/single precision Mutiply-Add execution rate. Hence, the POWER3 microprocessor is capable of executing four floating-point operations per clock cycle which results in a peak throughput of 800 MFLOPS.

6xx Bus

The 6xx bus or System Bus as shown in Figure 1, connects up to two processor cards to the memory-I/O controller chip set. This bus is optimized for high performance and multiprocessing applications. It provides 40 bits of real address and a separate 128-bit data bus. The address, data and tag buses are fully parity checked and each memory or cache request is range checked and positively acknowledged for error detection. Any error will cause a machine check condition and is logged in AIX error logs. The 6xx bus runs at a 100 MHz clock rate and peak data throughput is 1.6 GB/second. Data and address buses operate independently in true split transaction mode and are pipelined, so that new requests may be issued before previous requests are snooped or completed.

System Memory

The SP POWER3 SMP system supports 256MB to 4GB of 10ns SDRAM. System memory is controlled by the memory-I/O chip set via the memory bus. The memory bus consists of a 128-bit data bus and operates at 100 MHz clock cycle. As shown in Figure 1, this bus is separated from the System Bus (6xx bus) which allows for concurrent operations on these two buses. For example, cache to cache transfers can occur while a DMA operation is in progress to an I/O device. There are two memory cards slots in the system. Each memory card contains 16 DIMM slots. Only 128 MB memory DIMMs are supported for GA. Memory DIMMs must be plugged in pairs and at least one memory card with minimum of 256 MB of memory must be plugged in for system to be operational. System memory is protected by Single Error Correction, Double Error Detection ECC code.

I/O Subsystem

The Memory-I/O controller chip set implements a 64-bit plus parity, multiplexed address and data bus (6xx-MX bus) for attaching three PCI bridge chips and the SP Switch MX2 Adapter. The 6xx-MX bus runs at 60 MHz concurrently and independently from the 6xx and memory buses. At 60 MHz clock cycle, the peak bandwidth of the 6xx-MX bus is 480 MB/sec. The three PCI bridge chips attached to 6xx-MX bus provides the interface for 10 PCI slots. (2) 32-bit PCI slots are in thin node and (8) additional 64-bit PCI slots are in wide node.

One of the PCI bridge chips (Bridge Chip0) provides support for integrated Ultra2 SCSI and 10Base2, 100BaseT Ethernet functions. The Ultra2 SCSI interface supports up to 4 internal disks. An ISA bridge chip is also attached to PCI Bridge Chip0 for supporting two serial ports and other internally used functions in the POWER3 SMP Node.

Service Processor

The service processor function is integrated on the I/O planar board in the POWER3 SMP Node. Service processor function is for initialization, system error recovery, and diagnostics. The service processor supports system diagnostics by saving the state of the system in a 128 KB nonvolatile memory (NVRAM). The service processor code is stored in a 512 KB of flash memory and uses 512KB of SRAM to execute. The service processor has access to latches and registers of the POWER3 microprocessors and memory-I/O controller chip set using the serial scan method (JTAG).

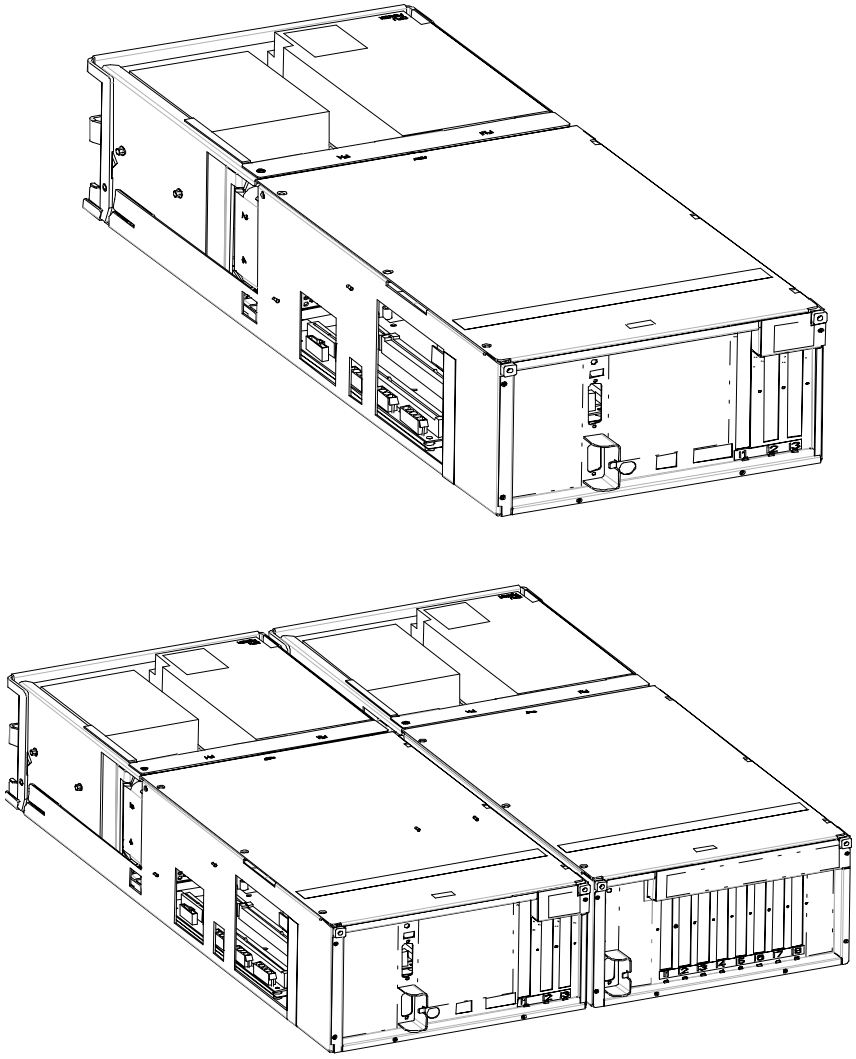
System Firmware and RTAS

The POWER3 SMP Node system firmware flash memory is located on the I/O planar. System firmware contains code which is executed by the POWER3 microprocessor during the initial program load (IPL) phase of the system boot. It also supports various interactions between the AIX operating system and hardware. The extent and method of interaction is defined in the RS/6000 Platform Architecture (RPA). The Run Time Abstraction Software (RTAS) defined by RPA, provides support for AIX and hardware for specific functions such as initialization, power management, time of day, I/O configuration, and capture and display of hardware indicators. RTAS and system IPL code are contained on 1 MB of flash memory.

System Packaging

The POWER3 SMP Node system packaging is somewhat different from that of the earlier 332 MHz SMP Node. However, the external dimensions and many internal mechanical packaging features have remained the same. One noticeable difference is the absence of flex cables in the POWER3 SMP Node. Flex cables were used in 332 MHz nodes to connect the thin node drawer to the expansion I/O drawer. In the POWER3 SMP nodes, connection between the main CPU drawer and expansion I/O drawer is made directly by mating the connectors on I/O planars located in each drawer.

Figure 2. POWER3 SMP Thin and Wide Nodes mechanical packaging.



Special Notices

This publication was produced in the United States. 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. Consult your local IBM business contact for information on the products, programs, services and features available in your area. Any reference to an IBM product, program, service or feature is not intended to state or imply that only IBM's product, program, service or feature may be used. Any functionally equivalent product, program, service or feature that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program, service or feature.

Information in this presentation concerning non-IBM products was obtained from the suppliers of these products, published announcement material or other publicly available sources. Sources for non-IBM list prices and performance numbers are taken from publicly available information including D.H. Brown, vendor announcements, vendor WWW Home Pages, SPEC Home Page, GPC (Graphics Processing Council) Home Page and TPC (Transaction Processing Performance Council) Home Page. IBM has not tested these products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this presentation. The furnishing of this presentation does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, NY 10594 USA.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your IBM local Branch Office or IBM Authorized Reseller for the full text of a specific Statement of General Direction.

The information contained in this presentation has not been submitted to any formal IBM test and is distributed AS IS. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. The use of this information or the implementation of any techniques described herein is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. Customers attempting to adapt these techniques to their own environments do so at their own risk.

The information contained in this document represents the current views of IBM on the issues discussed as of the date of publication. IBM cannot guarantee the accuracy of any information presented after the date of publication.

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

Any performance data contained in this document was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements quoted in this presentation may have been made on development-level systems. There is no guarantee that these measurements will be the same on generally-available systems. Some measurements quoted in this presentation may have been estimated through extrapolation. Actual results may vary. Users of this presentation should verify the applicable data for their specific environment.

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

Biographies

G. Mojtabaeezamani (Kazem) is the SP Node Hardware System Design Engineer. He is a member of the IBM Server Group, Poughkeepsie, New York.