

***The IBM @server pSeries 680
Technology and Architecture***

Technical White Paper

1. Abstract

This document describes the system architecture of the IBM @server pSeries 680. The major components; processors, memory, I/O subsystem, and the interconnecting buses are all described. The key aspect of this design is architected balanced performance using the newly developed copper and silicon-on-insulator (SOI) processor technology.

2. Introduction

The pSeries 680 represents IBM's symmetric multiprocessing (SMP) server for high-end commercial performance. It is ideally poised for the future because it supports running concurrently both 32-bit and 64-bit applications. It uses the new SOI technology enhanced RS64 IV processors and is offered in 6-, 12-, 18-, and 24-way configurations. The p680 also includes an enhanced L2 16MB cache and an increased memory size (96GB) for large database transactions. **(See architecture diagram on page 6)**

3. Rack Packaging

The p680 consists of two physical enclosures. The first is the central electronics complex (CEC) that contains processor cards, the memory controller complex backplane, memory cards, power supplies, and cooling units. The CEC cards are packaged in "books" for added reliability. Book packaging is described as cards that are sandwiched between two sheets of metal for protection, proper insertion and retention, and optimum air flow.

The second unit is a standard black 19-inch rack (36U high). It contains the primary I/O drawer and room for two secondary I/O drawers. Up to four I/O drawers are supported in a system. An optional 42U rack is also available. Unused space in the rack can be used for storage and I/O devices.

4. The RS64 IV Processor and Card

The RS64 IV superscalar processors are well optimized for commercial workloads. Target environments are characterized by heavy demands on system memory, both in the form of very large working sets and latency-sensitive serial dependencies. As a result, the design of the RS64 IV processors, which run at 600 MHz, is focused on large cache sizes and data paths having high bandwidth and low latency.

The RS64 IV processor has separate internal 128KB L1 caches for instructions and data. It contains an L2 cache controller and a dedicated 32-byte interface to a private 4-way set associative 16MB L2. The L2 interface runs at half processor speed, but transfers data twice per cycle to provide 19.2 GB/s of bandwidth. RS64 IV internal data paths are 32 bytes wide. The RS64 IV processor also has a separate 16-byte system bus interface.

The RS64 IV has a total of five pipeline execution units. There is a Branch Unit, a Load/Store Unit, a Fixed-point Unit, a Complex Fixed-point Unit and a Floating-point Unit. The processor has a 32-byte interface to the dispatch logic. There is a current

instruction stream dispatch buffer that is 16 instructions deep. The RS64 IV has an eight instruction deep branch buffer. The RS64 IV can sustain a decode and execution rate of up to four instructions per cycle.

All of the arrays in the RS64 IV have either redundancy and ECC or parity and retry to support the requirements for reliability, availability, and integrity. This enables full fault detection and correction coverage.

The RS64 IV p680 processor card has six processors with associated L2 cache contained on each card. There are 16MB of L2 per processor. Each processor card has the six processors on a set of two SMP system buses and that dual bus interface is presented to the p680 backplane. The systems processor cards all need to use the same type and speed of processor.

5. System Bus

The system bus design is optimized for high commercial performance in a multiprocessing environment. The system uses 128-byte coherency blocks. The system bus uses a 128-bit data path and a separate 64-bit address path. Address, data, and control are parity checked. Transfer sequences are validity checked. Each bus request is range checked and positively acknowledged for error detection. The system bus operates in a true split transaction mode and is aggressively pipelined. New requests can be issued before previous requests are snooped or completed. The address path includes status and coherency response signals for returning flow control, error reporting, and coherency information for each request. The coherency protocol used is enhanced MESI (Modified, Exclusive, Shared, and Invalid).

6. Backplane

The p680 has 10 system data buses connected to the memory controller complex. The buses are configured to run at a speed designed to optimize the processor used. The RS64 IV processor has system buses that run at 150 MHz. These busses are used in sets of two. Two of these high-performance buses are used to connect to each of the processor cards and the I/O interface. Four dual bus sets connect to the processor cards and one dual bus set connects to the I/O hub card.

The backplane uses a switch based memory controller complex. The complex contains 10 chips and additional high-function address and data buffers. There are actually two sets of switch chips. Each independent set of chips work on odd or even cache line accesses.

In each set of chips, one chip acts as the data flow control chip for four data switch chips. There is an additional system bus arbiter chip. Crossport traffic is queued at the switch if needed. Each of the processor card dual system buses is directly connected to a port on each of the sets of modules in the switch-based memory controller complex. Each data path is 128 bits wide. Addressing is via a separate 64-bit data path.

There are additional high-function address and data buffer chips that break these logical buses into smaller physical buses. This allows the frequency of the buses and data rates to be increased to the 150 MHz. level. These high-function buffers allow each of the system buses to support up to 2.4 GB/s of throughput.

The memory controller complex is mounted on the two-sided active backplane. The processors and memory are inserted as cards. The I/O subsystem is connected to the complex via the system bus to the Remote I/O (RIO) interface I/O hub chip-II. Two I/O hub chips are mounted on a replaceable card.

The memory controller complex has 14 unique data ports. Each of the four processor cards has its set of dual ports. One dual port set is connected exclusively to the I/O hub card. Four ports connect to the memory. The ports run independently of one another, which allows for another dimension in concurrent operation. Each of the four memory ports are 64 bytes wide. This width gives a very significant advantage when dealing with a large data transfers.

The 10 processor and I/O ports run at 150 MHz, for a total of 24GB/s of available bandwidth. Additionally, the memory ports operate at 75 MHz frequency and the four memory ports together have an aggregate bandwidth of 19.2GB/s. As a whole the total memory controller complex switch bandwidth is an impressive 43.2GB/s. Transfer buffers are available in the switch to queue traffic if the needed connections are not immediately available.

7. Memory

The base configuration includes 4GB of SDRAM based memory for the p680. The maximum configuration is 96GB. The p680 can accommodate up to 16 memory cards. Memory cards are used in sets of four, called quads. Each port connects to a different memory quad. The memory subsystem provides ECC for single-bit correction and double-bit detection.

The p680 family SDRAM memory cards offer redundant modules that will support up to one in 14 memory modules not working. The memory is scrubbed by the controller, a feature designed to eliminate soft error problems. Memory cards are available in 1GB, 2GB, 4GB, and 8GB sizes and utilize 64MB, 128MB, and 256MB technology. SDRAM modules are directly and permanently attached to the memory cards, a feature that minimizes failures and faults caused by connectors or sockets.

8. Remote I/O (RIO) Connections

The entire I/O subsystem is connected via a set of system buses to the RIO hub interface chip-II, which is mounted on a replaceable card. Four RIO connections are supported with a single hub chip. RIO connections are scalable, high-speed, point-to-point interfaces designed for low latency, high bandwidth connections between two boards or boxes. Each RIO bus supports up to 500MB/s total or 250MB/s in each direction concurrently.

The CEC enclosure contains no I/O. RIO cables connect the CEC to the I/O located in the I/O drawers. The RIO connections are set up as loops. The I/O hub chip directs the traffic around the loop in an optimal way for performance, and will redirect traffic if there are link errors.

The RIO hub interface chip-II offers improved buffering to enhance the effectiveness of the I/O interface.

These RIO connections are the key to allowing an expandable number of I/O drawers that are physically separated from the CEC. This feature in turn also enables the high number of PCI buses and slots.

9. I/O Drawer

The p680 I/O drawer offers the advantage of fully redundant power and fans that can be serviced without taking the system down. The drawer also uses robust fans which are especially useful if one of the fans should fail. In addition to the hot-plug fans and power supplies, this drawer offers Ultra SCSI adapters and backplanes that are separately cabled to the two DASD six packs. The drawer has a local display panel and reports more information for status monitoring. For the p680, several PCI adapters are supported in higher capacity configurations as well.

The primary I/O drawer for p680 systems contains the I/O planar and the PCI service processor/native I/O card. In addition, the primary I/O drawer contains 12 hot-plug DASD bays, one available media bay, one floppy disk drive, one CD-ROM, 14 PCI slots, one keyboard port, one mouse port, two serial ports and one parallel port. The I/O subsystem is expandable by attaching up to four I/O drawers to a single CEC.

The 14 PCI I/O slots consist of five 64-bit and nine 32-bit PCI slots. Three of the 14 slots in the first I/O drawer are used for the service processor, storage, and media support. This provides 11 additional slots in support of graphics, communications, and storage in the initial I/O drawer configuration.

On the advanced I/O drawers, the majority of the function is implemented on the I/O planar. A single RIO to I/O bridge bus chip converts the RIO bus to the local mezzanine bus or the I/O bridge bus. One RIO to I/O bridge bus chip drives four PCI bridge chips through the local I/O bridge bus. Each bus works independently of the others. Each RIO to I/O bridge chip has an "IN" and "OUT" RIO port to support a redundant loop connection of RIO devices that can be chained together. The I/O bridge bus runs at 66 MHz and has a 528MB/s bandwidth.

10. I/O Bridge Bus

The local I/O bridge bus is a reduced signal version of the system bus that has been optimized for I/O. The I/O bridge bus uses a multiplexed 64-bit address and data path. The I/O bridge bus is parity checked for address, data, and control errors. Each bus

request is range checked and positively acknowledged for error detection. The I/O bridge bus operates in a pipeline mode. New requests can be issued before previous requests are snooped or completed. The bridges and other chips in the I/O path provide significant queuing.

11. PCI Busses

The PCI bridge chips convert the I/O bridge bus to PCI. There are 14 PCI compliant slots running at 33 MHz per I/O drawer. PCI 2.1 cards are supported. Four PCI bridges are used per I/O planar. One of the PCI bridge chips drive two 64-bit PCI slots. The other three PCI bridges each drive one 64-bit PCI slot and three 32-bit PCI slots. This configuration performance balances the load. Both 5V and 3.3V are available at the slots. Five volt PCI signaling conventions are used.

The 64-bit/8-byte PCI slots have a maximum throughput of 266MB/sec. The 32-bit/4-byte PCI slots have a maximum throughput of 133MB/sec. It is important to note that no PCI to PCI bridges are used in this performance optimized design. PCI to PCI bridges significantly limit the useful bandwidth of the related PCI slots.

12. Summary

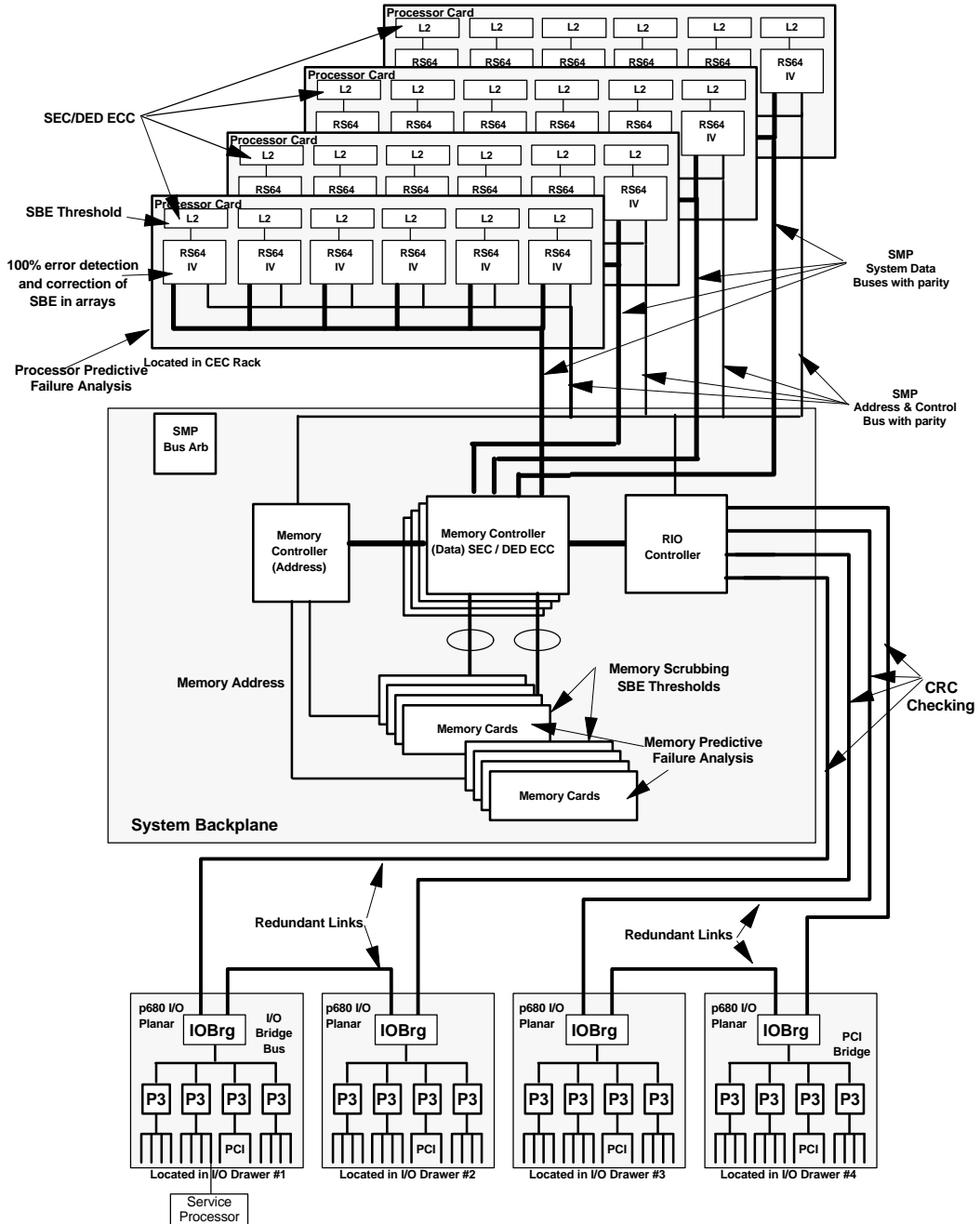
The p680 with the new high-performance copper and SOI technology, RS64 IV processors, offers significant performance improvement over previous designs of commercial SMP servers. The packaging, processors, high-function memory controller, and expandable/ adaptable I/O subsystem all contribute to a well integrated system implementation. Attention to RAS is present in the complete design. The I/O drawer includes redundant and hot-plug power and cooling for the I/O subsystem. The quality of the total integrated design is demonstrated by p680 world-class performance results, proof positive that the p680 design is focused on the proper system performance attributes.

13. For More Information:

The RS64 IV is further discussed in the “7th Generation 64-bit PowerPC Compatible Commercial Processor Design” White Paper.

The p680 also has a rich set of reliability features that are discussed in “The IBM @server p680 RAS White Paper”.

pSeries 680 System Architecture



© International Business Machines Corporation 2000

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

Produced in the United States of America
11-00 All Rights Reserved

More details are available at ibm.com/servers/unix/

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

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

IBM, the IBM logo, the e-business logo, AIX, CICS, DB2, DB2 Universal Database, IMS, MQSeries, pSeries, RS/6000 and WebSphere are registered trademarks or trademarks of the International Business Machines Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries, licensed exclusively through X/Open Company, Limited.

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