RC25145 (W1104-059) April 15, 2011
Computer Science

IBM Research Report

IBM RS/6000 SP

José E. Moreira
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

IBM RS/6000 SP

José E. Moreira
Research Staff Member
jmoreira@us.ibm.com

IBM T.J. Watson Research Center
1101 Kitchawan Rd., Yorktown Heights, NY 10598, USA

SYNONYMS

IBM SP, IBM SP1, IBM SP2, IBM SP3

DEFINITION

The IBM RS/6000 SP is a distributed memory message passing parallel system based on the IBM POWER processors. Several generations of the system were developed by IBM and thousands of systems were delivered to customers. The pinnacle of the IBM RS/6000 SP was the ASCI White machine at Lawrence Livermore National Laboratory, which held the number 1 spot in the TOP500 list from November 2000 to November 2001.

DISCUSSION

Introduction

The IBM RS/6000 SP (SP for short) is a general purpose parallel system. It was one of the first parallel systems designed to address both technical computing applications (the usual domain of parallel supercomputers) and commercial applications (e.g., database servers, transaction processing, multimedia servers). The SP is a distributed memory, message passing parallel system. It consists of a set of nodes, each running its own operating system image, interconnected by a high-speed network. In the TOP500 classification, it falls into the cluster class of machines.

IBM delivered several generations of SP machines, all based on IBM POWER processors. The initial machines were simply called the IBM SP (or SP1) and were based on the original POWER processors. Later, IBM delivered the SP2 machines based on POWER2 and then finally a generation based on POWER3 processors (which was unofficially called the SP3 by some). IBM continued to deliver parallel systems based on later generations of POWER processors (POWER4 and beyond), but those were no longer considered IBM RS/6000 SP systems. The November 2010 TOP500 list shows 16 POWER Systems 575 machines (one based on POWER5 processors and the rest based on POWER6 processors), which can be considered direct follow-on to the RS/6000 SP.

Notable IBM RS/6000 SP systems include the Argonne National Laboratory SP1 (installed in 1993) [5], the Cornell Theory Center IBM SP2, the Lawrence Livermore ASCI Blue Pacific and the Lawrence Livermore ASCI White. This last system consisted of 512 nodes with 16 POWER3 processors each and held the number 1 spot in the TOP500 list from November 2000 to November 2001.

The RS/6000 SP was designed to serve a broad range of applications, from both the technical and commercial computing domains. The designers of the system followed a set of principles [1] that can be summarized as follows: maximize the use of off-the-shelf hardware and software components while developing some special purpose components that maximize the value of the system. As a result, the RS/6000 SP utilizes the same processors, operating systems, and compilers as the contemporary IBM workstations. It also utilizes a special purpose high-speed interconnect switch, a parallel operating environment and message passing libraries, and a parallel programming environment, including a High Performance Fortran (HPF) compiler.

To further enable the system for commercial applications, IBM and other vendors developed parallel versions of important commercial middleware such as DB2 and CICS/6000. With these parallel middleware, customers were able to quickly port applications from the more conventional, single system image commercial servers to the IBM SP.

Hardware architecture

The IBM RS/6000 SP consists of a cluster of nodes interconnected by a switch (Figure 1). The nodes (Figure 2) are independent computers based on hardware (processors, memory, disks, I/O adapters) developed for IBM workstations and servers. Each node has its own private memory and runs its own image of the AIX operating system.

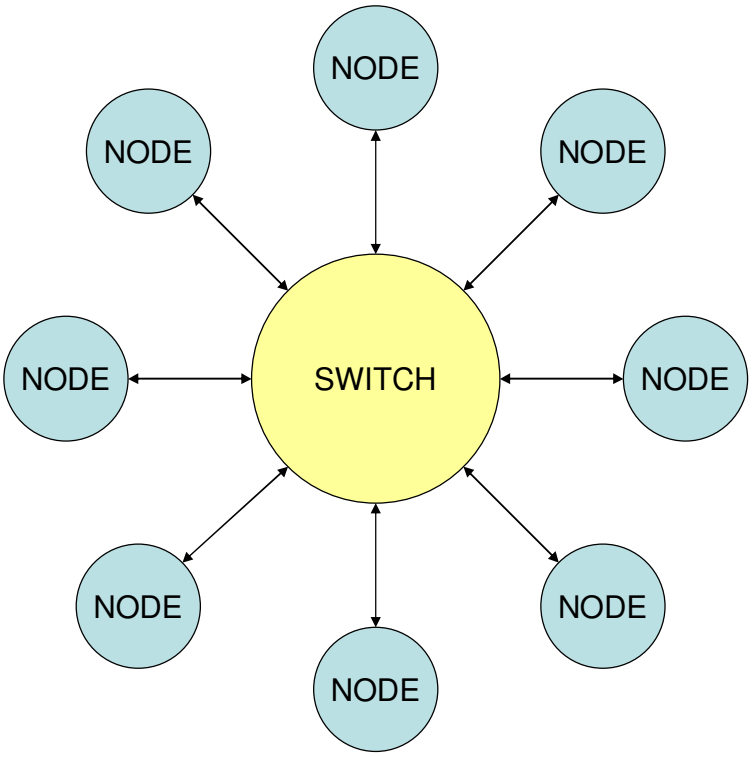

Figure 1: High-level hardware architecture of an IBM RS/6000 SP.

Through the evolution of the RS/6000 SP, different nodes were used. The initial nodes for SP1 and SP2 models where single-processor. Later, with the introduction of PowerPC and POWER3 processors, symmetric multiprocessing (SMP) nodes became available. Nodes could be configured to better serve specific purposes. For example, compute nodes could be configured with more processors and memory, whereas I/O nodes could be configured with more I/O adapters. The RS/6000 SP architecture supports different kinds of nodes in the same system, and it was usual to have both compute-optimized and I/O-optimized nodes in the same system.

The node architecture (illustrated in Figure 2) is essentially the same as contemporary standalone workstations and servers based on POWER processors. Processor and memory modules are interconnected by a system bus that supports memory coherence within the node. An I/O bus also hangs off this system bus. This I/O bus supports off-the-shelf adapters found on standalone machines, such as Ethernet and Fibre Channel. It also supports the switch adapters that connect the node to the network.

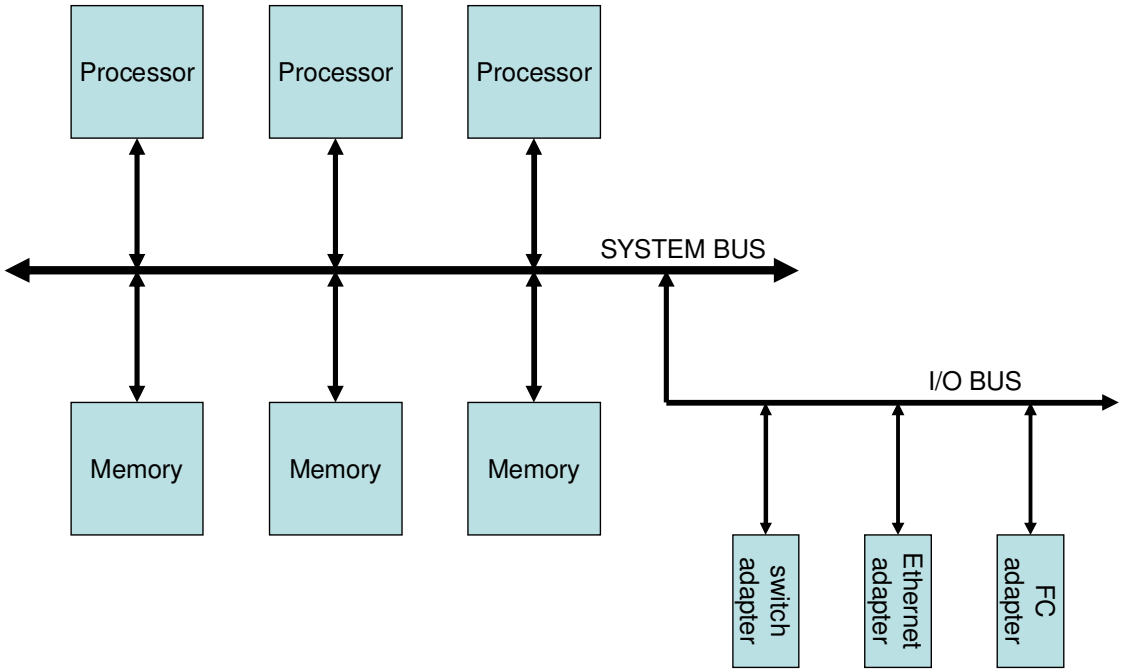

Figure 2: Hardware architecture of an IBM RS/6000 SP node. Different kinds of nodes can and have been used in SP systems.

Whereas the SP nodes are built primarily out of off-the-shelf hardware (except for the switch adapter), the SP switch is a special purpose design. At the time the SP was conceived, standard interconnection networks (Ethernet, FDDI, ATM) delivered neither the bandwidth nor the latency that a large scale general purpose parallel system like the SP required. Therefore, the designers decided that a special purpose interconnect was necessary [1,9].

The IBM RS/6000 SP switch [9] is an any-to-any packet-switched multistage network. The bisection bandwidth of the switch scales linearly with the size (number of nodes) of the system. The available bandwidth between any pair of communicating nodes remains constant irrespective of where in the topology the two nodes lie. These features supported both system scalability and ease of use. The system could be viewed as a flat collection of nodes, which could be freely selected for parallel jobs irrespective to their location. Selection could focus on other features, such as processor speed and memory size. As a consequence, the IBM RS/6000 SP did not suffer from the fragmentation problem that was observed in other parallel systems of the time [4].

The SP switch is built from basic 4×4 bi-directional crossbar switching elements, which are grouped eight to a board to form a 16×16 switch board, as shown in Figure 3. A switch board connects to nodes on one side and to other switch boards on the other side. Systems with up to 80 nodes can be assembled with just one layer of switch boards, whereas larger systems require additional layers.

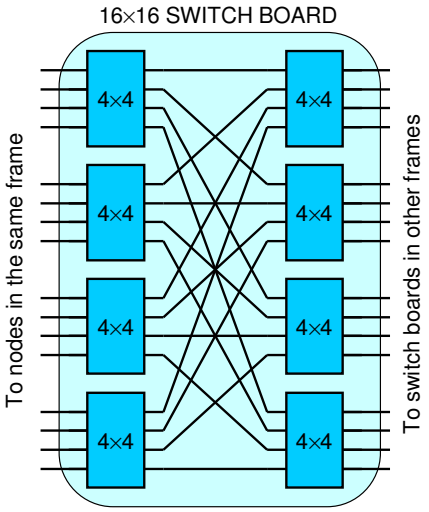

Figure 3: A 16×16 switch board is built by interconnecting eight 4×4 bi-directional crossbar switching elements. The switch board connects to nodes on one side and other boards on the other side.

Software architecture

Figure 4 illustrates the software stack of the RS/6000 SP. That software stack is built upon off-the-shelf UNIX components (in green) and specialized services for parallel processing (in red). Each node runs a full AIX operating system instance. That operating system is complemented at the bottom layer of the software stack by high-performance services that provide connectivity to the SP switch.

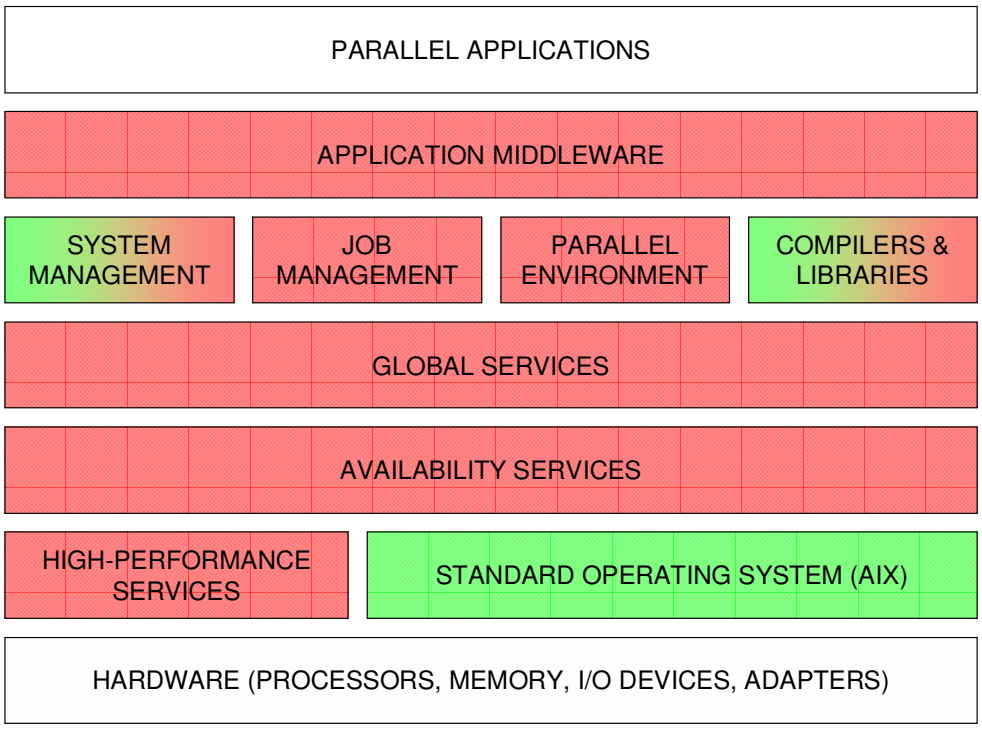

Figure 4: Software stack for RS/6000 SP.

The availability and global services layers implement aspects of a single-system image. They are intended to be the basis for parallel applications and application middleware. The availability services of the IBM SP support heartbeat, membership, notification and recovery coordination. Heartbeat services implement the monitoring of components to detect failures. Membership services allow processors and processes to be identified as

belonging to a group. Notification services allow members of a group to be notified when new members are added or old members are removed from that group. Finally, recovery coordination services provide a mechanism for performing recovery procedures within the group in response to changes in the membership.

The global services of the IBM SP provide global access to resources such as disks, files and networks. Global access to files is provided by networked file solutions, either with a client-server model (e.g., NFS) or with a parallel file system model (e.g., GPFS). With this approach, processes in every node see have access to the same file space. Global network access is implemented through TCP/IP and UDP/IP protocols over the IBM SP switch. Gateway nodes, connected to both the SP switch and an external Ethernet network, allow all nodes to access the Ethernet network. Global access to disks is implemented by virtual shared disk (VSD) functionality. VSD allows a process running on any SP node to access any disk in the system as if it were locally attached to that node. Another global service is the system data repository (SDR). The SDR contains system-wide information about the nodes, switches and jobs currently in the system.

The job management system of the IBM SP supports both interactive and batch jobs. Batch jobs are submitted, scheduled and controlled by LoadLeveler [2]. For interactive jobs, a user can login directly to any node in the SP, since the nodes all run a full version of the AIX operating system.

System management for the IBM SP is built upon components used for management of RS/6000 AIX workstations. It also includes extensions developed specifically for the SP to facilitate performing standard management functions across the many nodes of an SP. The functions supported include system installation, system operation, user management, configuration management, file management, security management, job accounting, problem management, change management, hardware monitoring and control, and print and mail services. The system management functions can be performed via a control workstation that acts as the system console.

The compilers and run-time libraries for the IBM RS/6000 SP are based on the standard software stack for IBM AIX augmented with certain features specific to the SP. Fortran, C and C++ compilers and run-time libraries for POWER-based workstations and servers can be directly used in the SP, since the nodes of the latter are based on hardware developed for the former. The software stack for the SP also includes message-passing libraries that implement both IBM proprietary models, such as MPL, and standard models such as PVM and MPI [5,8]. Also available for the IBM RS/6000 SP is an implementation of the High Performance Fortran (HPF) programming language [7].

In addition to middleware like MPI libraries that cater to scientific applications, the RS/6000 SP software stack also includes middleware targeted at enabling parallel commercial applications. The main example is DB2 Parallel Edition (PE) [2], an implementation of the DB2 relational database product that runs in parallel across the nodes of the SP. DB2 PE is a shared-nothing parallel database system, in which the data is partitioned across the nodes. DB2 PE splits SQL queries into multiple operations that are then shipped to the nodes for execution against their local data. A final stage combines the results from each node into a single result. DB2 PE enables database applications to use the parallelism of the RS/6000 SP without changes to the application itself, since the exploitation of parallelism happens in the database middleware layer.

Example applications

Thousands of IBM RS/6000 SP systems were delivered over the product lifetime, ranging in size from as few as 2 nodes all the way up to 512 nodes. The availability of a full workstation- and server-compatible software stack on the SP nodes allowed it to run off-the-shelf AIX applications with zero porting effort. The availability of standard message passing libraries (such as MPI), High Performance Fortran and parallel commercial such as DB2 PE also meant that existing parallel applications could be moved to the IBM RS/6000 SP with relative ease. Furthermore, several applications were specifically developed or optimized for the SP.

At a relatively early point in the product lifetime (1995), the SP was already being used in many different areas, including computational chemistry, crash analysis, electronic design analysis, seismic analysis, reservoir modeling, decision support, data analysis, on-line transaction processing, local area network consolidation, and as workgroup servers. In terms of economic sectors, SP systems were being used in manufacturing, distribution, transportation, petroleum, communications, utilities, education, government, finance, insurance, and travel [1].

The IBM RS/6000 SP played an important role in the Accelerated Strategic Computing Initiative by the U.S. Department of Energy. That program was responsible for several of the fastest computers in the world, including two SPs: the ASCI Blue-Pacific and ASCI White machines. ASCI Blue-Pacific was the largest SP in number of nodes. It consisted of 1,464 nodes, each with four PowerPC 604e processors. Each processor had a clock speed of 332 MHz and a peak floating-point performance of 664 Mflops. As indicated in the TOP500 list, the machine had a peak performance (Rpeak) of 3856.5 Gflops and a Linpack performance (Rmax) of 2144 Gflops. It was ranked #2 in the November 1999 and June 2000 lists. ASCI White was the largest SP in number of processors (or cores). It consisted of 512 nodes, each with 16 POWER3 processors. Each processor had a clock speed of 375 MHz and a peak floating-point performance of 1.5 Gflops. As indicated in the TOP500 list, the machine had a peak performance (Rpeak) of 12288 Gflops and a Linpack performance (Rmax) of 7304 Gflops. It was ranked #1 in the November 2000, June 2001 and November 2001 lists.

RELATED ENTRIES

TOP500; LINPACK; POWER processor; MPI

BIBLIOGRAPHIC NOTES

For a thorough discussion of the system architecture of the RS/6000 SP, the reader is referred to [1]. Details of the RS/6000 SP interconnection network can be found in [9]. An overview of the system software for the RS/6000 SP is given in [8] while details for the MPI environment and the job scheduling facilities are described in [5] and [3] respectively. The HPF compiler for the RS/6000 SP is described in [7]. Commercial middleware is covered in [2] and user experience in a scientific computing environment is described in [6]. Finally, additional information on the machine fragmentation problem is available in [4].

BIBLIOGRAPHY

1. T. Agerwala, J.L. Martin, J.H. Mirza, D.C. Sadler, D.M. Dias, M. Snir. SP2 System Architecture. IBM Systems Journal, vol 34, no 2. 1995.

2. C.K Baru, G. Fecteau, A. Goyal, H. Hsiao, A. Jhingran, S. Padmanabhan, G. P. Copeland, W. G. Wilson. DB2 Parallel Edition. IBM Systems Journal, vol 34, no 2. 1995. pp 292-322.

3. S. Dewey, J. Banas. LoadLeveler: A Solution for Job Management in the UNIX Environment. AIXtra, May/June 1994.

4. D. G. Feitelson, M. A. Jette. Improved utilization and responsiveness with gang scheduling. Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP '97). 1997. Springer-Verlag LNCS. pp 238-261.

5. H. Franke, C. E. Wu, M. Riviere, P. Pattnaik, M. Snir. MPI programming Environment for IBM SP1/SP2. Proceedings of the 15th International Conference on Distributed Computing Systems (ICDCS '95). May 30 – June 2, 1995. Vancouver, BC, Canada. pp 127-135.

6. W.D. Gropp, E. Lusk. Experiences with the IBM SP1. IBM Systems Journal, vol 34, no 2. 1995.

7. M. Gupta, S. Midkiff, E. Schonberg, V. Seshadri, D. Shields, K.-Y. Wang, W.-M. Ching, Wai-Mee, T. Ngo. An HPF compiler for the IBM SP2. Proceedings of the 1995 ACM/IEEE conference on Supercomputing. San Diego, CA, United States.

8. M. Snir, P. Hochschild, D. D. Frye, K. J. Gildea. The communication environment and parallel environment of the IBM SP2. IBM Systems Journal, vol 34, no 2. 1995. pp 205-221.

9. C. B. Stunkel, D. G. Shea, B. Abali, M. G. Atkins, C. A. Bender, D. G. Grice, P. Hochschild, D. J. Joseph, B. J. Nathanson, R. A. Swetz, R. F. Stucke, M. Tsao, P. R. Varker. The SP2 High-Performance Switch. IBM Systems Journal, vol 34, no 2. 1995.