RS/6000 SP

SP Switch Performance

August 1999

Version 3

Table of Contents

Preface

This document presents measurements of SP Switch latency and bandwidth between applications running in various SP node types. The measurements are discussed, as are the node processor and memory subsystem characteristics which affect inter-node communication performance. The resulting communication performance is, in turn, an element of higher-level application performance, measured by benchmarks such as SPECrate, Linpack HPC, and TPC.

This version of the document has added information about the
- POWER3 SMP Wide and Thin Nodes and
- SP-attached servers

and deleted information about the 160 MHz Thin Node which has been withdrawn.

Comments and suggestions can be sent via
- Lotus Notes to: Frank May/Somers/IBM@IBMUS
- Internet to: fhmay@us.ibm.com.

Acknowledgments

The information presented in this document was provided by Frank Johnston and Bernard King-Smith, of the SP Parallel Systems Performance group, Poughkeepsie, New York.

Introduction

The communication performance seen by applications, running in SP processor nodes and communicating through the SP Switch, is comprised of a number of elements. The SP Switch provides the base communications performance capability. The performance characteristics of the SP Switch adapter in the processor node, and the time the node takes to process the communication protocol stack , determine how much of the SP Switch performance capability can be sustained during application-to-application communication. The time for processing the communication protocol stack is, in turn, determined by the software path length and the performance characteristics of the processor.

This document presents the SP Switch and SP Switch adapter performance capabilities, discusses processor performance characteristics, presents the results of application performance measurements, and discusses these measurements. The software path length in processing communication protocol stacks is beyond the scope of this document and is not discussed.

SP Switch Performance

The raw peak performance of the SP Switch is given in Table 1.

Number of SP Nodes	Latency (µsec)	Bandwidth (MB/sec)	
		Uni-directional	Bi-directional
up to 80	1.2	150	300
81 to 512	2.0		

Table 1: SP Switch peak performance

This peak (i.e., not-to-exceed) performance cannot be achieved by an application.

SP Switch Adapter Performance

The raw peak performance of the SP Switch adapters is given in Table 2.

SP Switch Adapter	Bandwidth (MB/sec)	
	Uni-directional	Bi-directional
SPSMX2	150	300
SPSMX	150	300
SPSSA	150	150

Table 2: SP Switch adapter peak performance

- The SP Switch MX2 adapter (SPSMX2, feature code #4023) is available on the 200 POWER3 MHz SMP nodes. This adapter has an onboard 125 MHz 603e processor.
- The SP Switch MX adapter (SPSMX, feature code #4022) is available on the 332 MHz SMP nodes. This adapter has an onboard 100 MHz 603e processor.
- The SP System Attachment Adapter (SPSSA, feature code #8396), available on the Enterprise Server Models S70 and S70 Advanced, attaches on one end to a PCI slot in the server I/O drawer and on the other to the SP Switch cable.

These peak performance numbers represent the maximum rate at which data can be given to or taken from the SP Switch by the node and, effectively, become the base communications performance capability of the SP Switch subsystem. Similar to the SP Switch peak performance, the SP Switch adapter peak (i.e., not-to-exceed) performance cannot be achieved by an application.

When communication is between nodes with unlike adapters, the effective peak performance of the SP Switch subsystem is that of the slower adapter.

Processor Performance

The key processor characteristics affecting application communication performance are clock frequency and memory bandwidth. They are discussed for the top-of-the-line SMP and P2SC processors.

200 MHz POWER3 Processor

The IBM 200 MHz POWER3 processor operates at a clock frequency of 200 MHz. And delivers the highest sustained floating-point performance of any SP processor.

The POWER3 SMP processor chip has a 128-bit wide memory bus. This 16-byte bus width and a bus frequency of 100 MHz (1:2 compared to the clock frequency) delivers a peak memory bandwidth of 1.6 GB/sec. In this document, for bandwidth, gigabyte is defined as 10**9 (vs. 2**30) bytes, except where noted otherwise.

The POWER3 is capable of executing up to 2 FMA instructions per cycle (i.e. four floating-point operations per cycle) for a total peak floating-point performance of 800 MFLOPS (Mega FLoating-point Operations Per Second). In addition to better peak floating-point performance, the 200 MHz POWER3 can sustain higher memory bandwidth than the 160 MHz P2SC processor.

332 MHz SMP Processor

The IBM 332 MHz 604e SMP processor operates at a clock frequency of 332 MHz and delivers the highest integer performance of any SP processor.

The SP 604e processors are part of the RS/6000 PowerPC family of processors. The 332 MHz 604e SMP processor chip has a 128-bit wide memory bus. This memory bus width and a bus frequency of 83 MHz (1:4 compared to the clock frequency) delivers a peak memory bandwidth of 1.33 GB/sec.

The 332 MHz 604e SMP processor is capable of executing one floating-point multiply-add (FMA) instruction per cycle (i.e., two floating-point operations per cycle), for a total peak floating-point performance of 664 MFLOPS.

Application Performance

High-performance inter-node communication is a key component of the overall performance of many user applications. The most basic measurements to characterize the performance of the communication subsystem are *latency* and *bandwidth*. Latency is the overhead associated with sending data between two processors, and is usually quantified in microseconds (μsec). Bandwidth is the rate at which data can be transmitted between two processors, and is typically measured in megabytes per second (MB/s). In this document, for bandwidth, megabyte is defined as 10**6 (vs. 2**20) bytes, except where noted otherwise.

In this section, we will characterize inter-node communication performance over the SP Switch for the two SP Switch communication protocols: the so-called *user space* protocol, used to support the industry standard Message Passing Interface (MPI), and the industry standard *IP (Internet Protocol) family* of communication protocols (which include TCP/IP and UDP/IP). The user space protocol is sometimes referred to as a lightweight protocol because it requires fewer processor cycles to transmit a given amount of data compared to heavier protocols like TCP/IP.

User space is most commonly used for scientific and technical computing applications via a message-passing interface. MPI was used to measure user space performance. TCP/IP is utilized in socket communication for many commercial applications, and is the basis for popular network protocols such as Network File System (NFS) and File Transfer Protocol (FTP). Performance measurements for these two protocols are presented for the SMP nodes in Tables 3, 4, and 6, and for the P2SC nodes in Tables 5 and 7.

Several factors contribute to the communication performance that is obtained by a user application. Inter-node communication performance depends on the processor, the memory subsystem, the switch adapter, and the SP Switch fabric. Therefore, when considering communication performance measurements, it is extremely important to understand the exact configuration of the system to which the data applies. In addition to hardware considerations, the system software contributes to the overhead involved in sending data between processors.

Please note the following:

- The latency and bandwidth measurements Tables 3 through 7 represent performance as seen by an application.
- Measured bandwidth increases asymptotically as message size grows very large. Each bandwidth measurement presented in these tables represents the asymptotic values for very large messages.
- The measurements for a given SP node were made using the latest release of SP software generally available at the time of the announcement of that node.
- The latencies and bandwidths for older nodes are included in these tables for reference.

MPI/user space

On a distributed-memory system like the SP, parallel applications perform inter-processor communication via some form of message passing. IBM fully supports MPI as an industry standard. This standardized interface for message-passing greatly improves the portability of parallel application codes among different parallel systems. We will discuss the performance of the SP Switch for parallel applications in terms of what can be measured using MPI.

Tables 3 through 5 show inter-processor communication performance measurements from a FORTRAN program with MPI calls, using the user space protocol. Latency is a measure of the time of sending a zero byte message between two processors using `mpi_send` and `mpi_recv` from the MPI library. It is calculated as half the time for a round trip between the processors for that zero byte message. Latency represents the time taken to set up a single message for transfer at the level of an application, and may be seen as the initialization overhead for transferring information between applications.

These tables also contain data for bandwidth measurements using MPI over the user space interface. The uni-directional bandwidth, sometimes called *point-to-point* bandwidth, was measured for messages of several megabytes in size. The bi-directional bandwidth, sometimes called *exchange bandwidth*, implies simultaneous sending and receiving of messages between processors, thereby achieving a slightly higher data rate. The bi-directional data rate is the sum of the simultaneous data rates in both directions.

SP Nodes and Attached Servers

The SPSMX2 adapter latency is the lowest achieved on the SP.

SMP Processor Type	RS/6000 Model Equivalent	SP Switch Adapter	Latency (µsec)	Bandwidth (MB/sec)	
				Uni-directional	Bi-directional
200 MHz. POWER3 SMP	43P-260	SPSMX2	21.7	139	170
332 MHz SMP	H50	SPSMX	23.5	83	86
262 MHz RS64 II	S7A	SPSSA	37.3	70	87

Table 3: MPI user space performance with SMP nodes with a single MPI tasks per node

The 200 MHz. POWER3 node with the 125 MHz SPSMX2 adapter has a slightly lower latency than a 332 MHz SMP node with the 100 MHz SPSMX adapter. This is due to the increased speed of the processor in the adapter. The SPSMX2 adapter connects directly to the system *Mezzanine* (MX) bus. The 200 MHz. POWER3 has better MPI bandwidth than the 332 MHz. SMP node because the POWER3 CPU can perform memory-to-memory copies faster. The results above were obtained using the non-threaded MPI library.

The 332 MHz SMP node with the SPSMX Adapter has lower latency and higher bandwidth than the SPS Adapter, or the SP-attached server with the SPSSA adapter. The superior memory bandwidth of the 332 MHz SMP node contributes to the increased MPI bandwidth. The SPSMX adapter also connects directly to the system MX bus. The superior design of the SPSMX leads to improved MPI performance relative to SPS.

Measurements for the SPSSA are included for completeness, even though applications that use SP-attached servers will generally use IP rather than MPI. (Attached servers will generally be used for commercial computing applications, while MPI is generally used by scientific and technical computing applications.)

The performance data shown in Table 3 were generated using the following hardware configurations. Because the measurements were memory-to memory, the processor memory and node internal disk storage configurations did not affect the results.

200 MHz. POWER3 node
- Two 2-processor 200 MHz. POWER3 nodes
- One SPSMX2 adapter per node
- SP Switch

332 MHz SMP node
- Two 4-processor 332 MHz SMP nodes
- One SPSMX adapter per node
- SP Switch

262 MHz SP-attached server

- Two 4-processor 262 MHz SP-attached servers
- One SPSSA adapter per server
- SP Switch

Release 3.1 of the Parallel Systems Support Programs (PSSP) allows multiple user space processes per adapter (MUSPPA). Table 4 shows unidirectional and exchange bandwidth for multiple MPI tasks per POWER3 and 332 MHz SMP nodes. As the number of tasks per node increases, the aggregate memory to memory copy rate increases. The bandwidth through the SPSMX/SPSMX2 adapter also increases up to four MPI tasks, where the throughput limits of the adapters are reached.

The SPSMX2 adapter is the limiting factor on POWER3 so the bandwidth with two MPI tasks per node is not much better than with one MPI task per node.

SMP Processor Type	Number of MPI tasks	Bandwidth (MB/sec)	
		Uni-directional	Bi-directional
332 MHz SMP	1	83	86
	2	127	149
	4	128	162
200 MHz POWER3 SMP	1	139	170
	2	140	185

Table 4: MPI user space performance on SMP nodes with multiple MPI tasks per node.

TCP/IP

TCP/IP is a more common industry standard communication protocol used to transfer information between any two systems running IP. It is a robust protocol that supports multiple users and reliable transport of data. However, since it supports networking function not currently used by MPI, such as multiplexing, it requires higher processor overhead compared to the user space protocol using MPI.

The performance of the TCP/IP socket protocol on various nodes was measured using Netperf, a public-domain benchmark, and the results are listed in Tables 6 and 7. All Netperf measurements were memory-to-memory to eliminate slower devices, such as disks, from impacting the performance. Note that in these tables, megabyte is defined to mean 2**20 (vs. 10**6) bytes, due to the way Netperf calculates its results. As with the user space measurements, TCP/IP bandwidths are largely determined by the speed with which the TCP and IP protocol stacks are processed. The processor memory copy rate also affects the maximum throughput rate.

It must be emphasized that the performance of the IP protocols family is a strong and complex function of the characteristics of the network, the processor, the processor's memory bandwidth, as well as a lengthy list of IP stack tuning parameters termed *network options*.

SP Nodes and Attached Servers

The TCP/IP bandwidths on the POWER3 SMP nodes are the highest TCP/IP data rates achievable on any SP node.

Number of processors[1]	Bandwidth (MB/sec), uni-/bi- directional					
	200 MHz. POWER3 SMP node		332 MHz SMP node		262 MHz SP-attached Server	
	Uni-	Bi-	Uni-	Bi-	Uni-	Bi-
1	114.3	156.2	63.9	101.0	73.5	88.5
2	134.8	174.0	114.5	156.0	73.7	89.9
4	N/A	N/A	128.6	156.5	73.9	89.9

Table 5: TCP/IP performance with SMP nodes

The POWER3 SMP node, compared to the 332 MHz. SMP node, delivers between 12% and 78% better bandwidth. These improvements are primarily attributable to the difference in memory bandwidth of the nodes. The 262 MHz SP-attached server delivers excellent single process throughput. However, it is limited by the throughput of the PCI bus to which the SPSSA adapter is connected. The SPSSA adapter uses a 132 MB/s PCI bus and a single adapter in that bus can only get 90 MB/s under a real application.

Bandwidth is the maximum obtainable both uni-directional and bi-directional over TCP between two identical applications running on two identical SMP nodes. All Netperf measurements were memory-to-memory to eliminate slower devices such as disks from impacting the performance.

The SMP nodes can take advantage of multiple processors if there are multiple IP connections running at the same time. The results in Table 5 show that as you increase the number of processors or TCP streams, the aggregate throughput increases for all but the SP-attached server. If only one TCP/IP socket is used, the maximum throughput will be similar to the single-processor throughput no matter how many processors are configured in the node. A single TCP/IP socket currently cannot take advantage of multiple processors, due to the single-threaded nature of memory-to-memory copies and the TCP/IP stack.

The performance data shown in Table 5 were generated using the following hardware configurations. The processor memory and node internal disk storage configurations did not affect the results.

200 MHz POWER3 node
- Two 2-processor 200 POWER3 nodes
- One SPSMX2 adapter per node
- SP Switch

SP Switch Router

The SP can send switch traffic to outside networks through an SP Switch Router. Sold exclusively by IBM (as machine type 9077), this is a combination of the Lucent Technologies (formerly Ascend Communications) GRF router and the IBM SP Switch Router Adapter which connects to the SP Switch fabric. The SP Switch Router node only supports IP traffic. The performance of the SP Switch Router Adapter was measured using the Netperf benchmark described earlier. Table 6 contains the peak aggregate throughput through the SP Router node .

Adapter type	Bandwidth (MB/sec)	
	Uni-directional	Bi-directional
SP Switch Router Adapter	100	200

Table 6: TCP/IP performance through the SP Router nodes

The performance data shown in Table 6 were generated using the following hardware configuration. The node internal disk storage configuration did not affect the results. Not all nodes in the test configuration were needed to sustain the peak throughput of the SP Router node.

SP Switch Router
- 1 SP Switch Router
- 2 SP Switch Router Adapters
- 10 P2SC 120 MHz nodes
- 8 P2SC 135 MHz nodes
- One SPS adapter per node
- SP Switch

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 paper 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 paper. 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 paper 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 paper 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.

All prices shown are IBM's suggested list prices; dealer prices may vary.

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

Any performance data contained in this paper was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements quoted in this paper 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 of International Business Machines Corporation in the United States and/or other countries: AIX, RS/6000, SP.

Microsoft, Windows, Windows NT and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation. 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, which may be denoted by a double asterisk (**), may be trademarks or service marks of others.

Notes on Benchmarks and Values

The benchmarks and values shown here were derived using particular, well configured, development-level computer systems. Unless otherwise indicated for a system, the values were derived using 32 bit applications and external cache if external cache is supported on the system. All benchmark values are provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration. Buyers should consult other sources of information to evaluate the performance of systems they are considering buying and should consider conducting application oriented testing. For additional information about the benchmarks, values and systems tested, please contact your IBM local Branch Office or IBM Authorized Reseller or access the following on the Web:

- SPEC http://www.specbench.org
- Linpack http://www.netlib.no/netlib/benchmark/performance.ps
- NAS http://science.nas.nasa.gov/Software/NPB/Reports/NAS-96-018.html