RS/6000 SP

SP Switch

and

SP Switch2 Performance

June 2001

Version 5

Table of Contents

Preface

This document presents measurements of SP Switch2 and SP Switch latency and bandwidth between applications running in various AIX® cluster nodes, including the RS/6000® SP™. The measurements discussed in this paper measure inter-node communication performance.

The resulting communication performance is an element of higher-level application performance, measured by benchmarks such as SPECrate, Linpack HPC, and TPC.

Comments and suggestions can be sent to Barbara Butler on the Internet at: butlerbm@us.ibm.com.

Acknowledgments

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

Introduction

The communication performance seen by applications, running in AIX cluster nodes and communicating through the SP Switch and SP Switch2, is comprised of a number of elements. The SP Switch and SP Switch2 provide the base communications performance capability. The performance characteristics of the SP Switch and SP Switch2 adapters in the node, and the time the node takes to process the communication protocol stack, determine how much of the SP Switch and SP Switch2 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 performance capabilities for both the SP Switch and SP Switch2 switch fabrics and their adapters. We also present the current results of application performance measurements. The software path length in processing communication protocol stacks is beyond the scope of this document and is not discussed.

In this document, *switch* refers to any switch type and *adapter* refers to any adapter type, unless a specific switch or adapter is named.

Switch Performance

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 previous generation SP Switch. The bandwidth gains result from design improvements that include locating the switch adapter directly on the 375 MHz POWER3 SMP High Node system bus. These gains enable switch communication performance to keep pace with the increased performance of the high node.

Because the SP Switch2 design improvements are evolutionary steps based on the SP Switch and High Performance switches, the SP Switch2 is fully compatible with applications written for the older switches. As with these prior switches, the SP Switch2 supports the MPI (Message Passing Interface), LAPI (Low level Application Interface), and IP (Internet Protocol).

The raw peak performance of the SP Switch and new SP Switch2 are given in Table 1.

Number of cluster nodes	Switch Type	Latency (µsec)	Bandwidth (MB/sec)	
			Uni-directional	Bi-directional
Up to 16	SP Switch2	1.0	500	1000
17 to 80		1.5		
81 to 512		2.5		
Up to 16	SP Switch	1.3	150	300
17 to 80		1.9		
81 to 512		3.3		

Table 1: Switch peak performance

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

Adapter Performance

The raw peak performance of the SP Switch and SP Switch2 adapters are given in Table 2.

SP Switch Adapter	Bandwidth (MB/sec)	
	Uni-directional	Bi-directional
SP Switch2	500	1000
SPSMX2	150	300
SPSAA	150	150

Table 2: Switch adapter peak performance

Where:

- SP Switch2 adapter
  - The SP Switch2 adapter (SP Switch2, SP feature code #4025) is available on the SP POWER3 SMP High Nodes.
- SP Switch adapters
  - The SP Switch MX2 adapter (SPSMX2, SP feature code #4023) is available on the SP POWER3 SMP Wide/Thin Nodes.
  - The SP System Attachment Adapter (SPSAA, server feature code #8396) is available to attach various servers to the SP Switch. The adapter attaches to a PCI slot on the server and, at the other end, to an SP Switch cable. Servers which can be attached include the @server pSeries 680 and pSeries 660 Models 6H0 and 6H1, and the RS/6000 Enterprise Servers S80, M80, and H80.

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

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

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, when using MPI, a megabyte is defined as 10**6 bytes. When using IP, a megabyte is defined as 2**20 bytes. Historically, these have been the definitions used when measuring these two protocols.

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 in Tables 3, 4, and 5.

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 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 5 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 node were made using the latest release of 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.
- Latencies and bandwidths are measured on two nodes connected to the same switch chip.
- On the 375 MHz POWER3 high node, MPI tasks were bound to CPUs in order to reduce performance fluctuations

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 and 4 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.

The SP Switch2 adapter has the lowest latency of the adapters.

Node Type	Switch Adapter	Latency (µsec)	Bandwidth (MB/sec)	
			Uni-directional	Bi-directional
375 MHz POWER3 High Node	SP Switch2	17.9	350	350
375 MHz POWER3 Thin/Wide Node	SPSMX2	19.7	140	192
p680	SPSAA	36.6	71	87

Table 3: MPI user space performance with a single MPI tasks per node connected to the same switch chip.

The 375 MHz POWER3 SMP High Node with the SP Switch2 adapter has lower latency and higher bandwidth than all other node and adapter combinations. The SP Switch2 adapter connects directly to the node 6xx I/O bus, which results in superior bandwidth compared to earlier adapter-bus (PCI and MX bus) connections.

The results above were obtained using the non-threaded MPI library for the p680 nodes, and used the threaded MPI library with MP_SINGLE_THREAD=yes for all POWER3 results.

Measurements for the SPSAA are included for completeness, even though applications that use attached servers will generally use IP rather than MPI communication. (These 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.

375 MHz POWER3 SMP High Node

- Two 16-processor 375 MHz POWER3 High Nodes
- One SP Switch2 adapter per node
- SP Switch2

375 MHz POWER3 Thin/Wide Node

- Two 2-processor 375 MHz POWER3 Thin/Wide Nodes
- One SPSMX2 adapter per node
- SP Switch

p680 node

- One 24-processor, 600 MHz p680 server
- One SPSAA adapter per p680
- One 200 MHz POWER3 Thin/Wide Node
- One SPSMX2 adapter per POWER3 node
- SP Switch

Starting with Release 3.1 of the Parallel Systems Support Programs (PSSP) we allow multiple user space processes per adapter (MUSPPA). Table 4 shows unidirectional and exchange bandwidth for multiple MPI tasks per node. As the number of tasks per node increases, the aggregate memory to memory copy rate increases. The bandwidth through the SP Switch2 adapter increases up to 4 MPI tasks per node, where the throughput limits of the SP Switch2 adapter are reached.

On the 375 MHz POWER3 Thin/Wide Nodes and the p680 nodes, there is no increase beyond a single task due to the bandwidth limits of the SPSMX2 and SPSAA adapters.

Node Type	Switch Adapter	Number of MPI tasks per node	Bandwidth (MB/sec)	
			Uni-directional	Bi-directional
375 MHz POWER3 High Node	SP Switch2	1	350	350
		2	445	680
		4	445	720
		5 or more	No increase	No increase
375 MHz POWER3 Thin/Wide	SPSMX2	1	140	192
		2 or more	No increase	No increase
p680	SPSAA	1	71	87
		2 or more	No increase	No increase

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

TCP/IP

TCP/IP is a more common industry standard communication protocol used to transfer information between two systems running the IP family of protocols. 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. This increases the CPU overhead for the same bandwidth when compared to MPI.

The performance of the TCP/IP socket protocol on various nodes was measured using a version of the Netperf public-domain benchmark, and the results are listed in Table 5. All Netperf measurements were memory-to-memory to eliminate slower devices, such as disks, from impacting the performance. As with the user space measurements, TCP/IP bandwidths are largely determined by the speed of the processor memory copy rate.

The TCP/IP bandwidths on the 375 MHz POWER3 SMP High Node are the highest TCP/IP data rates achievable on any SP node.

Number of processors[1]	Bandwidth (MB/sec), uni-/bi- directional					
	375 MHz POWER3 High Node		375 MHz POWER3 Thin/Wide Node		p680 Node	
	Uni-	Bi-	Uni-	Bi-	Uni-	Bi-
1	132	253	135.4	174.6	74.2	90.7
2	252	472	134.8	175.8	74.3	90.7
4	440	650	135.4	174.0	74.4	90.7
5 or more	No increase		N/A	N/A	No increase	
Switch adapter	SP Switch2		SPSMX2		SPSAA	

Table 5: TCP/IP performance

The 375 MHz POWER3 SMP High Node delivers the highest bandwidth of all node and adapter combinations for the switches. This performance can primarily be attributed to the difference in memory bandwidth and processor speed of the node. The p680 and 375 MHz POWER3 SMP Wide/Thin Nodes deliver excellent single process throughput, however, multiple processor performance is limited by the throughput of the bus to which the adapter is connected.

The single-processor performance on the SP Switch and SP Switch2 adapters on POWER3 nodes is identical. This is due to the single-CPU memory-to-memory copy rate. When using the SP Switch2 adapter and switch, you get higher bandwidth when using more than one CPU.

[1] see description of configuration differences, below

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

The nodes can take advantage of multiple processors if there are multiple IP connections running at the same time. 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.

375 MHz POWER3 SMP High Node

- Two 16-processor 375 MHz POWER3 High Nodes
- One SP Switch2 adapter per node
- SP Switch2

375 MHz POWER3 Wide/Thin Node

- Two 2-processor 375 POWER3 Wide/Thin Nodes
- One SPSMX2 adapter per node
- SP Switch

p680 node

- One 24-processor, 600 MHz p680 server
- One SPSAA adapter per p680 node
- One 375 MHz POWER3 SMP Wide/Thin Node
- One SPSMX2 adapter per POWER3 node
- SP Switch

SP Switch Router Node

The SP can send SP Switch traffic to external networks using the SP Switch Router (machine type 9077), through the SP Switch Router Adapter. In this discussion, the SP Switch Router Adapter is considered to be an SP node. Only IP communication is supported and the SP Switch2 is not supported. Performance was measured using the Netperf benchmark described earlier, and Table 6 shows the peak aggregate throughput .

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

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

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.

SP Switch Router

- 1 SP Switch Router
- 2 SP Switch Router Adapters
- 10 120 MHz Thin Nodes
- 8 135 MHz Wide Nodes
- One SP Switch adapter per each of the 18 SP nodes
- SP Switch

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

Produced in the United States of America
04-01

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

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.