Redbooks Paper

Ruzhu Chen
Mike Ebbers

PMaC Benchmarking on Three POWER4 Platforms

The performance tests of three POWER4™ platforms, IBM @server™ pSeries™ 690, 655, and 650 were evaluated using the PMaC Benchmark Suite. The PMaC benchmark tests can be used to meaningfully compare and predict machine performance on parallel applications. This IBM® Redpaper summarizes the results on memory bandwidths and latency for load and store, I/O performance, communication bandwidths, and mega floating-point operations per second (MFLOPS) for loop applications.

Please visit the PMaC Web site at:

http://www.sdsc.edu/PMaC/

Introduction

The performance evaluation of parallel applications on a High Performance Computing (HPC) machine is complicated. However, two main factors, single processor performance and the communication performance among processors, are usually sufficient for evaluating the performance of an application.[1] To model the single processor performance of an application, the bandwidth and latency of data storing and loading into the memory hierarchy (main memory, caches), as well as the performance of total floating points and memory operations, are measured. The communication performance is evaluated by measuring communication bandwidth and latency on and off nodes, and the percentage of communication required to distribute and gather data in a parallel application.

The Performance Modeling and Characterization, PMaC, HPC Benchmark Suite[2] is a set of orthogonal benchmarks that enable performance evaluation of essential HPC hardware and system features, such as I/O, CPU, network, and memory.[3] Because total floating points and memory operations of an application are the same, the comparison of the performance of different system can be decided mainly by machine profiles, such as the load and store bandwidths of single processor and communication bandwidths between processors on or off nodes. This paper describes the performance measurements of our symmetric multiprocessor (SMP) POWER4 machines pSeries 690, 655, and 650 using the PMaC Benchmark Suite.

The PMaC Benchmark Suite is derived from the synthetic benchmarks of the Department of Defense High Performance Computer Modernization Office (DoD HPCMO) and Pallas Message Passing Interface (MPI) Benchmarks (that is, PMB benchmarks).[4] The suite consists of six benchmarks: MAPS (memory load and store bandwidth), EFF-BW (effective bandwidth), MAPS-CG, MAPS-PING (MPI communications), I/O-BENCH (I/O), and PEAK (floating point and memory operations). The purpose of the PMaC benchmark testing is to meaningfully compare and predict machine performance on certain applications. The following describes the benchmarks:[2]

- **MAPS**: MAPS measures the bandwidths of a specific machine. The memory measurements of MAPS are designed to measure the bandwidth for loads and stores from all levels of memory. These measurements include stride one and random access.
- **MAPS-CG**: MAPS-CG is an extension of the MAPS benchmark for measuring bandwidth. MAPS-CG measures the rate at which a system does an MPI all-to-all using blocking sends and non-blocking receives. The benchmark is modeled after the communication pattern that occurs in the conjugate gradient (CG) kernel of the NAS Parallel Benchmarks.
- **MAPS-PING**: MAPS-PING is a network benchmark, an extension of the PMaC MAPS bandwidth benchmark. MAPS-PING measures the startup and throughput of a simple message sent between two processes. This ping-pong test is slightly altered from the original PMB ping-pong kernel and calculates latency and bandwidth. It focuses on a single message transfer between two processes.
- **EFF-BW**: The effective bandwidth for a given machine and processor is calculated by sending messages through MPI_Send, which are then received through MPI_Recv by its partner process, and then sent back to the original process. EFF-BW reports a cumulative result for all participating processors, reporting the performance for small and large messages.

---

[1] Snavely, et al., *Modeling Application Performance by Convolving Machine Signatures with Application Profiles*, San Diego Supercomputing Center, 2001, available at: http://www.sdsc.edu/pmac/Papers/micro34.pdf

[2] See: http://www.sdsc.edu/PMaC

[3] Carrington, et al., *A Framework For Application Performance Prediction to Enable Scalability Understanding*, available at: http://www.sdsc.edu/pmac/Papers/PSC_Scaling/Scaling_paper.pdf

[4] See: http://www.pallas.com

- **PEAK**: PEAK measures the MFLOPS of a given machine by timing loop applications including divide, DAXPY, DOT product, and a fifth degree polynomial over a series of loop sizes.
- **IO-BENCH**: I/O-BENCH measures the rate a machine performs reads and writes to an arbitrary state of disk. These tests are designed to “mimic the I/O requirements for a given shared resource center site's I/O workloads.” The benchmark runs a series of sequential, backward, and random read and write tests from a single interactive script.

System configurations

Three of the IBM POWER4 (and plus) platforms (p690, p655, and p650) were tested for performance using the PMaC Benchmark Suite. The p690 system is introduced as a high-end server for technical computing and commercial server workloads. The p690 system uses four 8-way multi-chip modules (MCM), which have a full complement of L2 and L3 caches, and is optimized for scientific and technical computing.[5] Both the p655 and p650 systems are introduced as mid-range servers, where p655 is an ultra-dense packaged server specifically designed for building high-performance clusters.[6,7] The p655 system is a one-MCM machine. The p650 machine is the first pSeries using POWER4+™ chip technology that is packaged on a single chip module (SCM) as part of the processor “book” that incorporates a L3 cache and eight memory DMM slots.[7] The detailed configurations of these three platforms used for this benchmarking are listed in Table 1.

All the tests were performed in a local file system without modification of benchmark codes. Each benchmark was repeated three times.

Table 1 System and hardware configurations

Configurations		**p690**	**p650**	**p655-1.3**
Processor		1.3 GHz POWER4	1.45 GHz POWER4+	1.3 GHz POWER4
Processors/node		32	8	4
Memory/node		160 GB	32 GB	32 GB
Mem (GB)/processor		5	4	8
Caches	L1	64/32 KB	64/32 KB	64/32 KB
	L2	1.44 MB/card[a]	1.5 MB/card	1.44 MB/card
	L3	16 MB/card	32 MB/card	128 MB
OS		AIX® 5.1.0.0	AIX 5.1.0.0	AIX 5.1.0.0
AIX kernel		32-bit	32-bit	32-bit
File system		JFS	JFS	JFS
Fortran compiler		xlf 7.1	xlf 7.1	xlf 7.1
C/C++ compiler		vac 6.0	vac 6.0	vac 6.0

a. Two processors per card.

[5] Mathis, et al., *IBM e-Server pSeries 690 Configuring for Performance*, available at: http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.pdf
[6] See: http://www-1.ibm.com/eserver/pseries/hardware/midrange/p655_desc.html
[7] Olszewski, *IBM eServer pSeries 650 Performance and Tuning*, available at: http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p650_perf.pdf

Measurement and results

In this section, we discuss the various benchmarks that were run and the results of each one.

MAPS benchmark to test the memory performance of a single processor

To obtain the memory performance of a single processor, the memory access pattern signature, MAPS, benchmark was executed to collect memory access information (such as bandwidth) for all levels of memory, using various size working sets and memory access patterns (stride one and random access). Thus, the MAPS benchmark results give sustainable rates of memory load or store, depending on the access pattern and the sizes of the working problem.

Figure 1, Figure 2, Figure 3, and Figure 4 show the memory bandwidths versus the problem sizes for load and store, displaying a three-stair pattern on all three systems. When the size of a problem was smaller than the L1 and L2 cache sizes, the bandwidths of load and store were closely related to the processor speed. The bandwidths of the 1.3 GHz p690 and p655 were ~1.2 (1.3/1.1) higher than that of the 1.1 GHz p655, while the bandwidth of the 1.45 GHz p650 was about 1.1 times that of the 1.3 GHz system. When the size of a problem surpassed the cache size, the bandwidths were dramatically decreased. The higher bandwidths were observed on p690 with a size more than 32 K.

The memory bandwidth is highly influenced by the memory access pattern. Problems with random access patterns ran much slower than those that accessed memory sequentially (stride one). Smaller problem sizes yielded significant cache reuse. Of the four machines being tested, the 1.45 GHz p650 had better memory performance for small size problems, and the p690 had the best performance for large size problems.

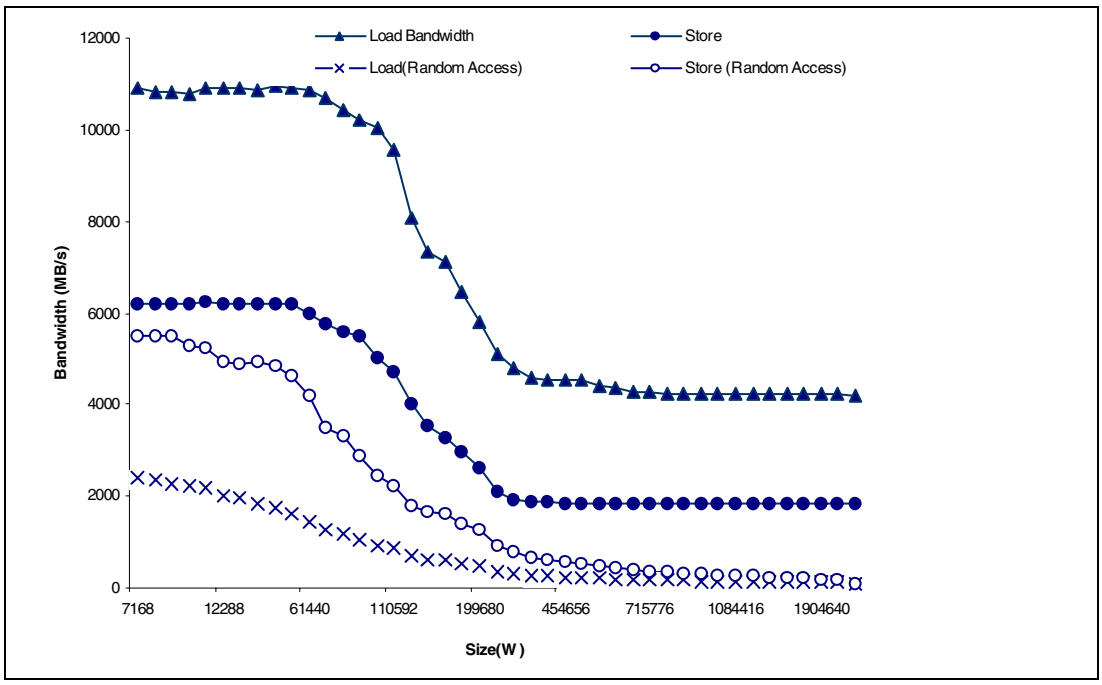


Figure 1 Memory bandwidth versus size for load and store on p690 (two streams)

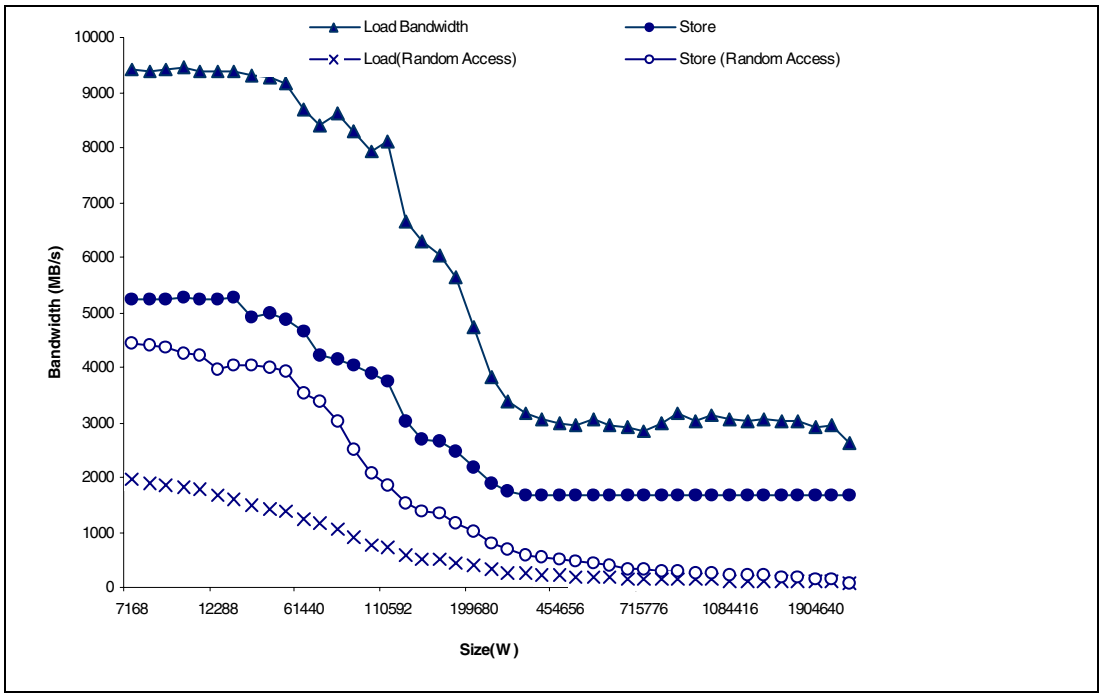


Figure 2 Memory bandwidth versus size for load and store on p655-1-1 GHz (two streams)

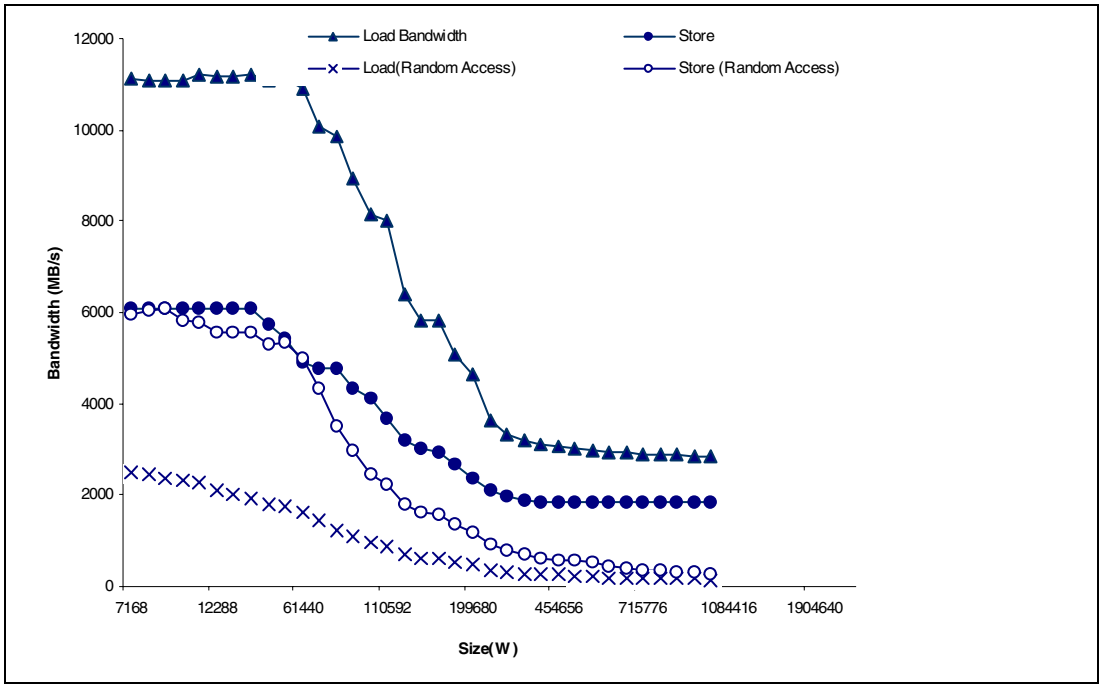


Figure 3 Memory bandwidth versus size for load and store on p655-1.3 GHz (two streams)

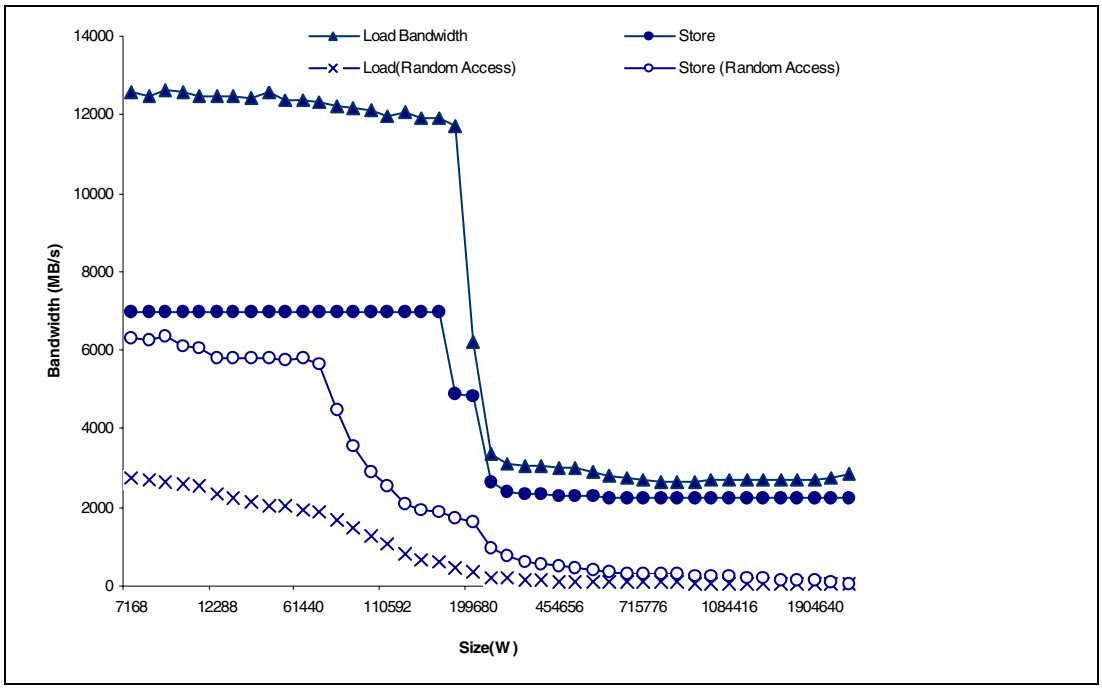


Figure 4 Memory bandwidth versus size for load and store on p650 (two streams)

MAPS-CG to collect communication bandwidth in conjugate gradient (CG) kernel

Using MPI all-to-all blocking sends and receives, the MAPS-CG benchmark tests communication bandwidths and latency among processors. Table 2 shows that the p690 and 1.3 GHz p655 machines have higher bandwidths, but the p655 had lower latency.

Table 2 MAPS-CG benchmark results

Platform	Number of nodes	CPUs/node	Bandwidth (MB/sec)/CPU	Latency (µsec)
p690	1	4	614.28	9.2
	1	8	343.33	9.6
	1	32	211.92	9.8
p650	1	4	557.54	8.5
p655-1.1	1	4	540.79	9.5
	2	8	407.94	10.0
p655-1.3	1	4	606.62	3.8
	4	4	638.23	8.6

MAPS-PING to test communication bandwidth between two processors

The MAPS-PING benchmark was used for a performance evaluation of communication between processors. The test measures the bandwidths and latency between two processors on a node or cross node. The resulting data on two processors, either randomly picked or close together in one chip, was collected (see Table 3). The results show that the performance of two processors in one chip is better where higher bandwidth and lower

latency were observed. Of the four machines being tested, the p650 has the best performance. These results indicate that using logical processor binding to execute parallel applications might result in better performance.

Table 3 MAPS-PING benchmark results

Platform	On node		Off node	
	Bandwidth (MB/s) per CPU	**Latency (µsec)**	**Bandwidth (MB/s) per CPU**	**Latency (µsec)**
p690	1513.19[a]	4.42		
	1425.13	5.00		
p650	1858.24[a]	3.94		
	1654.11	4.60		
p655-1.1				
	1464.05	4.96	412.36	19.9
p655-1.3	1633.00[a]	4.18		
	1606.38	4.25	414.17	19.2

a. Between logical CPUs close to each other.

EFF-BW to report effective communication bandwidth and latency

EFF-BW evaluates the communication performance of all the participating processors by cumulating the results for small and large messages. The results in Table 4 show that both p655 machines have the best effective-bandwidth and lowest latency using simple and cross mapping, while the latency of the p690 is nearly double that of the p655s. These results suggest that the p655 might be well designed for clustering.

Table 4 Effective bandwidths of the three platforms

Platform	Number of nodes	PEs/node	EFF-BW (MB/sec)	EFF-BW/pe (MB/sec)	Latency (µsec) (simple mapping)	Latency (µsec) (cross mapping)
p690	1	2	631.88	315.94	4.78	4.78
	1	4	1224.32	306.08	5.01	4.91
	1	8	2306.53	288.31	5.03	5.23
	1	16	3701.90	231.37	4.57	5.19
	1	32	5726.83	178.96	5.00	5.11
p655-1.3	1	2	856.73	428.37	2.26	2.25
	1	4	1651.62	412.91	2.39	2.61
p655-1.1	1	2	720.96	360.48	2.82	2.82
	1	4	1337.71	334.43	3.03	2.82
	1	8	2147.74	268.47	2.80	2.95

Platform	Number of nodes	PEs/node	EFF-BW (MB/sec)	EFF-BW/pe (MB/sec)	Latency (µsec) (simple mapping)	Latency (µsec) (cross mapping)
p650	1	2	692.17	346.09	4.31	4.30
	1	4	1223.28	305.82	4.47	4.35
	1	8	1560.90	195.11	4.46	4.71

PEAK to test single processor performance on loop applications

The PEAK benchmark measures the performance of floating and memory operations on a single processor. Four loops (division, DAXPY, DOT, and fifth degree polynomial) that represent computation-intensive to memory-intensive applications were executed. The results are shown in Table 5. For short loops, the p650 (1.45 GHz) has the best performance, while the p650 and 1.3 GHz p655 and p690 have very close performance for long loops.

Table 5 PEAK performance results on all three platforms

Platform	Loop length	Division (MFLOPS)	DAXPY (MFLOPS)	DOT (MFLOPS)	Fifth degree poly (MFLOPS)
p690	1024	68.04	1462.85	1920.00	2718.58
	4096	68.27	1204.71	1346.63	2720.59
	16384	68.51	1198.83	1363.76	2705.61
	65536	68.27	747.56	1134.28	2553.35
	262144	64.46	298.45	620.87	170389
	524288	62.42	289.66	559.24	1718.98
	1046576	62.38	289.82	559.56	1723.02
	4194304	62.29	288.60	551.08	1715.85
p650	1024	76.99	1638.40	2296.82	3056.72
	4096	77.04	1342.95	1520.17	3046.61
	16384	77.40	1346.63	1532.01	3024.74
	65536	77.40	973.31	1123.47	2877.19
	262144	71.82	389.33	457.01	2016.49
	524288	67.65	333.94	379.00	1728.42
	1046576	67.28	334.01	378.03	1733.59
	4194304	67.65	341.92	387.17	1747.63

Platform	Loop length	Division (MFLOPS)	DAXPY (MFLOPS)	DOT (MFLOPS)	Fifth degree poly (MFLOPS)
p655-1.3	1024	68.27	1517.04	1861.82	2706.61
	4096	68.27	1164.74	1322.48	2720.59
	16384	68.03	1177.29	1325.45	2693.26
	65536	67.80	768.00	907.42	2542.25
	262144	63.94	355.85	450.46	1767.26
	524288	62.05	349.53	420.83	1709.64
	1046576	61.56	360.89	424.93	1755.13
	4194304	61.68	363.67	414.82	1747.63
p655-1.1	1024	57.53	1226.35	1649.40	2301.12
	4096	57.56	1015.54	1134.28	2296.82
	16384	57.66	992.97	1145.30	2295.03
	65536	57.49	610.58	732.70	2137.04
	262144	54.80	291.27	413.01	1547.08
	524288	53.23	277.40	364.72	1423.41
	1046576	53.22	277.03	362.28	1427.15
	4194304	52.87	311.46	375.61	1559.87

IO-BENCH to collect I/O performance information

The IO-BENCH tests a machine's I/O performance through different read/write patterns. The results are summarized in Table 6.

Table 6 I/O benchmark test results on all three platforms

Platform	Test	File size (MB)	Buffer size (KB)	Bandwidth (MB/sec) Max.	Min.	Avg.
p690	Sequential write	10	4	166.67	142.86	154.76
	Sequential read	10	4	1000.00	1000.00	1000.00
	Random read rewrite	10	4	200.00	200.00	200.00
	Random read	10	4	1000.00	1000.00	1000.00
	Write backwards	10	4	333.33	333.33	333.33
	Backward read	10	4	200.00	200.00	200.00

Platform	Test	File size (MB)	Buffer size (KB)	Bandwidth (MB/sec) Max.	Min.	Avg.
p650	Sequential write	10	4	200.00	166.67	181.82
	Sequential read	10	4	41.67	38.46	40.05
	Random read rewrite	10	4	166.67	166.67	166.67
	Random read	10	4	1000.00	500.00	750.00
	Write backwards	10	4	333.33	250.00	292.65
	Backward read	10	4	71.43	71.43	71.43
p655-1.3 GHz	Sequential write	10	4	66.67	66.67	66.67
	Sequential read	10	4	500.00	500.00	500.00
	Random read rewrite	10	4	166.67	142.86	154.77
	Random read	10	4	500.00	500.00	500.00
	Write backwards	10	4	250.00	250.00	250.00
	Backward read	10	4	83.33	76.92	80.13
p655-1.1 GHz	Sequential write	10	4	66.67	66.67	66.67
	Sequential read	10	4	1000.00	500.00	750.00
	Random read rewrite	10	4	142.86	142.86	142.86
	Random read	10	4	500.00	500.00	500.00
	Write backwards	10	4	250.00	250.00	250.00
	Backward read	10	4	76.92	71.43	74.18

Conclusion

The PMaC Benchmark Suite test results indicate that it could be utilized as a model to compare machines and predict performance. Of the four systems on the three platforms being evaluated, the performance on the p690, p650, and 1.3 GHz p655 is close, although memory bandwidths and communication bandwidths are slightly different from each other.

The team that wrote this Redpaper

This Redpaper was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

Ruzhu Chen is an Advisory Engineer/Scientist at the IBM @server pSeries and HPC Benchmark Center. He joined the HPC benchmarking team in 2001 and works on benchmarking and performance analysis of scientific and technical applications for IBM @server pSeries and xSeries® servers. Ruzhu received his Ph.D. in Life Sciences and Master of Computer Science at the University of Oklahoma, and then performed two years postdoctoral research on gene design and cloning before joining IBM.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those 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.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

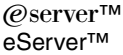
COPYRIGHT LICENSE:
This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Send us your comments in one of the following ways:

- Use the online **Contact us** review redbook form found at:
  ibm.com/redbooks
- Send your comments in an Internet note to:
  redbook@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYJ Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®	IBM®	pSeries™
	ibm.com®	Redbooks™
	POWER4™	Redbooks(logo)  ™
eServer™	POWER4+™	xSeries®

The following terms are trademarks of other companies:

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