2007－HPC－111 (47)
2007／8／3

Performance Analysis on Optically-Enabled SMP Servers

Atsuya Okazaki and Yasunao Katayama
IBM Research, Tokyo Research Laboratory
{a2ya, yasunaok}@jp.ibm.com

This paper analyzes the performance and utilization of the cache, memory, and their interconnects in optically-enabled SMP servers, and discusses cases where the optical interconnect is applied to either the cache coherent SMP link or the memory link. The analysis are conducted with a cycle-accurate full system simulation tool running HPC benchmarks, taking care of the extra latency due to the time-of-flight (e.g., 1 meter – 5 ns). The simulation results show that the performance with the optical memory link is better than that with the optical SMP link as the number of nodes increases under realistic configurations. By observing the inside of the processors and the SMP server systems, we identified that the topological differences cause a performance crossover.

1. Introduction

While commercial processor designs are shifting toward homogeneous/heterogeneous multi-core processors to sustain the rate of computational performance improvement predicted by Moore's Law. For a given power budget, the capacity and bandwidth of the memory system per chip are limited due to the electrical wiring and packaging constraints. As a result the memory capacity and bandwidth per core could decrease as the number of cores increases[1]. The situation is most pressing in SMP servers where a large number of cores need to be packaged densely. The memory capacity and bandwidth are also required to accommodate multiple machines and application images in memory. In conventional ccNUMA designs, memory modules must be placed near processor chips to provide enough bandwidth between memory and the processors. In addition, since modern processor chips need to have memory links and SMP cache coherent links as well as standard and proprietary IO links, many lines for signals from and to the processors, and complex carrier and PCB layouts. Since these wiring constrains are dominating the area around processor chips, no room is available to increase the aggregate memory bandwidth. In addition, often memory capacity is also limited by the space. Simply speaking, there are not enough area to accommodate necessary memory modules. Although there are large capacity memory chips using expensive technologies, using these memory modules would result in very expensive server systems.

Optical interconnect technologies are currently widely utilized for high-speed I/O networks such as InfiniBand and Fibre Channel, mostly for rack-to-rack interconnections. Where the electronic links have limitations for bandwidth and wire length, optical interconnects are able to connect between widely separated components without losing much bandwidth. It is natural to believe that this advantage can solve the issue of degraded memory bandwidth and capacity per CPU core in the multi-core era[2),3)]. When optical links are utilized for SMP links, the distances between processor chips can be increased. When optical links are utilized in memory links, the distances between processor chips and memory modules can be increased. Extra latency is required to access memory outside of the CPU core. It will require careful study to assess how these two options effect server performance.

In this paper, we analyze the performance and utilization of the cache, memory, and their interconnects in optically-enabled SMP servers, and discuss cases where the optical interconnect is applied to either the cache coherent SMP link or the memory link. They are evaluated with a cycle-accurate full system simulation tool running HPC benchmarks, taking care of the extra latency due to the transmission time of the optical fiber (e.g., 1 meter = 5 ns). The simulation results show that the performance benefit of the optical memory link exceeds that of the optical SMP link as the number of nodes increases. By observing the inside of the processors and the SMP server systems, we determined that the topological differences cause this performance crossover The present study shows an interesting opportunity where the optical interconnect technology can improve the performance of future SMP servers.

2. Optical Interconnect Opportunities

As is mentioned, optical interconnect technology has essential advantages for longer distances and

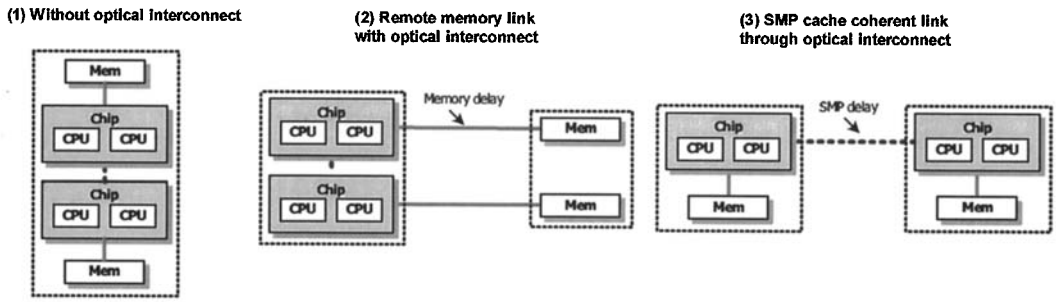

Fig. 1 SMP server architecture options with and without optical interconnects. Optical interconnect technology leads to flexible CPU and memory resource arrangements, but latency overhead needs to be assessed carefully.

higher bandwidth compared to electrical interconnections. These advantages allow server designers to place more components from dense areas to distant sites that have more room without bandwidth degradations. If the free space around processor chips is increased by using optical interconnections, such areas can be used for additional components to improve the performance (e.g., additional memory modules, or additional SMP processor chips). However, at this time, it would not be possible to replace all of the chip-to-chip links in current server systems with optical interconnections due to the limitations of current optical packaging and the high costs. Therefore, in practice, two options are being considered for the first phase of adding optical interconnect technology into server subsystems, memory links and cache coherent SMP links. When using optical interconnects as memory links, it is possible to separate a processor chip from its memory modules. With SMP links, it is possible to separate the processor chips. Both options seem to have natural advantages and disadvantages. This paper quantitatively compares these two options in various SMP server configurations through detailed simulation.

3. Simulation Environment

3.1 Simulation Technique

The originality of our performance simulation methodology is that we add extra latencies to the SMP links and the memory links in the simulation models of the SMP servers to analyze the performances affected by the latencies and the configuration differences (See "Memory delay" and "SMP delay" in Figure 1).

Table 1 summarizes the expected impacts on the cache and memory access times from the CPU cores. When the memory links are long ("Memory delay" in Figure 1), the communication between processors and memory slowed down. Therefore, local and remote memory accesses require long latencies since they pass through the long memory links. When the SMP links are long, the communication between processors is delayed ("SMP delay" in Figure 1). Therefore, remote cache and remote-memory accesses require long latencies since they pass through the long SMP links.

Table 1 The delay effects.

Delayed link	Fast latency	Slow latency
Memory	Local cache, remote memory	Local memory, remote memory
SMP	Local cache, local memory	Remote cache, remote memory

3.2 Simulation Tools

We utilized a full system simulator[4] in cycle-accurate mode to evaluate various SMP server configurations where extra delays are added to SMP links and memory links. The processor model is based on the IBM Power Architecture[8),9)] and each core has private L1 and shared L2 caches. We executed a Linux SMP kernel on the simulator, and utilized LAM/MPI[6),7)] as an MPI environment for a widely used high-performance computing benchmark, the HPC Challenge[5)], where the MPI communication was configured as using shared memory.

3.3 Simulation Models

Figure 2 shows the four SMP configurations we used in the simulations. The first one is a single-node configuration where the node consists of a single chip with two CPU cores. The cores share an L3 cache and a memory controller connected to the external memory. The second one is a two-node configuration where each node consists of a chip with dual cores. The two chips are connected to each other via a cache-coherent SMP link. The third configuration is a four-node configuration. The four

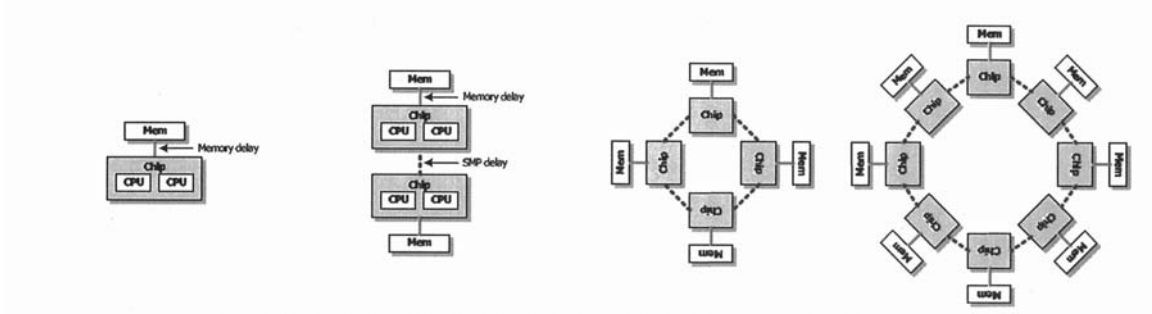

Fig. 2 SMP server models: single, two, four, and eight processor chip systems. Each chip has two CPU cores.

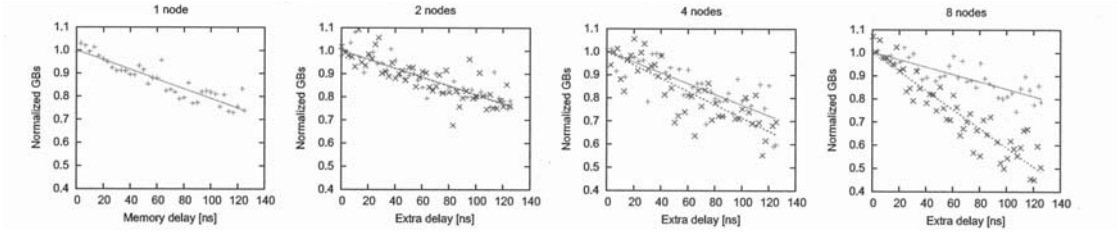

Fig. 3 Performance versus extra delay for PTRANS on single, two, four, and eight node(s). '+' (plus) and 'x' (cross) denote performance with memory delay and SMP delay, respectively. Solid and dashed lines denote least square fitting of the performance with memory delay and SMP delay, respectively.

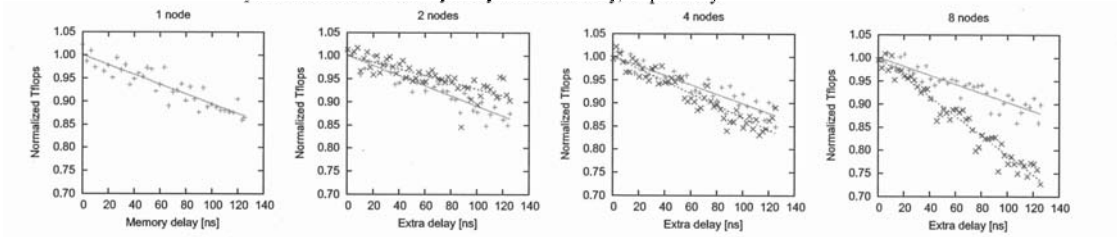

Fig. 4 Performance versus extra delay for HPL on single, two, four, and eight node(s). The notation in these graphs is the same as in Figure. 3.

nodes are connected in a ring topology. The fourth configuration is a eight-node configuration. The eight nodes are connected in a ring topology in the same way as the four node configuration. For each configuration, we added extra delays: *memory delay* between the processor chips and memory and *SMP delay* between processor chips. This enables us to measure the performance impacts due to separating a processor and memory or due to separating the processor nodes. In the single-node configuration, only *memory delay* is added.

4. Results and Observation

4.1 Performance Analysis

Figures 3 and 4 show the simulation results on PTRANS and HPL in the HPC Challenge benchmark for each configuration, one, two, four and eight nodes, respectively. While HPL has temporal locality in its memory access pattern, PTRANS does not have such temporal locality. The single-node data, where only *memory delay* is added, shows that the extra latency overhead due to time-of-flight and other factor does not significantly degrade the overall performance for both benchmarks. Considering that it takes about 5 ns to go through a typical 1-m fiber, the performance is degraded by only a few percent, even when the distance between the memory and processor is 10 m (i.e., 50 ns).

As the number of nodes increases, we can compare how *memory delay* and *SMP delay* affect the performance. The PTRANS benchmark shows that *memory delay* and *SMP delay* effects are comparable for two nodes. With four nodes, the *SMP delay* affects the performance more than the *memory delay*, so, the data captures a clear trend as the number of nodes is increased. The data for eight nodes clearly confirms this trend. The multiprocessor systems utilizing the optical memory bus outperform

those utilizing the optical SMP link. When the SMP system consists of a larger number of nodes, the performance degradation due to multiple hops in the cache communication links becomes significant. In contrast, the performance impact of *memory delay* is less even if the number of SMP nodes is increased.

For the HPL benchmark, the performance impacts of the extra delays are smaller than for the PTRANS benchmark, since the HPL benchmark can effectively utilize the cache memory due to the temporal locality of the memory accesses[5]. In the two-node case, *SMP delay* affects the performance less than *memory delay*. However, in the four-node case, *SMP delay* starts to have larger effects than *memory delay*. For the eight-node case, the trend is even more clear.

4.2 Cache and Memory Access Patterns

Figure 5 plots "Hit × Cycle" for each memory access operation in the PTRANS benchmark on the two-node and four-node configurations. Hit denotes the number of accesses from the CPU core to the cache or memory. Cycle denotes the average access time of cache and memory. That is, Hit × Cycle means the total cost in time for the CPU core to fetch data from the appropriate storage. Note that the figure does not include some operations where the values of Hit × Cycle are negligible. The figure shows three configuration options on the horizontal axis. The first is that no delay is added anywhere ('0' on the horizontal axis in the figure). Second is *memory delay* ("Mem"), where the *memory delay* is set as 125 ns. Third is *SMP delay* ("SMP"), where the *SMP delay* is also set as 125 ns. L21, L25, and L35 denote the other L2 cache on the same processor chip, the L2 cache on the other chips, and the other L3 caches on the other chips, respectively. LMEM and RMEM also denote the local memory connected to the same processor chip and the remote memory connected to the other chips.

In the two-node case, *memory delay* affects Hit × Cycles more than *SMP delay* for major operations (i.e., User:Data-LOAD, User:Data-STORE, User:L1-Prefetch-LOAD, and User:L2-Prefetch-LOAD). One of the major components affecting this trend is the local memory access, where the value of *memory delay* is much larger than the value of *SMP delay*. However, in the four-node case, the impact of the *SMP delay* exceeds that of the *memory delay* for the major operations. Two major factors are causing this crossover. First is the local memory access which is about half of the two-node case. Second is the remote cache access in *SMP delay*, which is becoming much longer than in the two-node case. Although Figure 5 does not include the eight-node case due to space limitation, its access pattern confirms the above trend. The *SMP delay* affects Hit × Cycle more than the *memory delay* as the number of SMP nodes is increasing. The difference between the values of *SMP delay* and *memory delay* is becoming more clear.

In the HPL benchmark, the same trends as the PTRANS benchmark can be observed. Figure 6 shows Hit × Cycle in the HPL benchmark on the two-node and four-node configurations. While the values of the *memory delay* is larger than that of the *SMP delay* in the two-node case, that of the *SMP delay* exceeds that of the *memory delay* in the four-node case for most operations except for User:Data-LOAD. The value of the *SMP delay* in all of the operations including User:Data-LOAD exceeds that of the *memory delay* in the eight-node case, which graph is not shown here due to the space limitations. In addition, in the eight-node case, the *SMP delay* affected the remote cache access much more than in the four-node case.

4.3 Trends of the Access Patterns to Extra Latency

Figure 7 and 8 focus on the operation User:L2-Prefetch-LOAD in the HPL benchmark on the four-node configuration, and show the difference between the *memory delay* and *SMP delay*.

Figure 7 shows the cycle time and Hit × Cycle versus *memory delay*. The left graph in the figure confirms that the *memory delay* affects only the cycle time of the local and remote memory accesses. The right graph shows that the ratio of the remote memory and local memory accesses is increasing with the *memory delay*. For comparison, Figure 8 shows the cycle time and Hit × Cycle versus *SMP delay*. The left graph in the figure confirms that the *SMP delay* affects only the cycle times of the remote cache and remote memory accesses. The right graph shows that the ratio of the remote cache and remote memory accesses is increasing with the *SMP delay*.

Figure 9 shows cycle and Hit × Cycle versus *SMP delay* for Kernel:Data-LOAD in the PTRANS benchmark on the four-node configuration. The left graph confirms that *SMP delay* affects the cycle times of the accesses utilizing the delayed SMP link, which are not only the remote cache and remote memory accesses but also the accesses requiring to maintain the cache coherency. Therefore, the ratio of the L2 on the other processor chips (i.e., "L25" in the figure) is getting larger with the

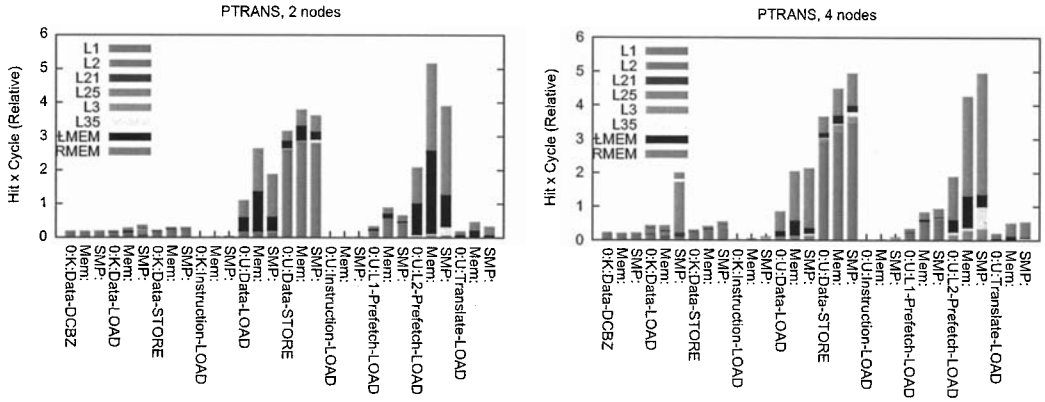

Fig. 5 Cache and Memory Access Patterns in PTRANS on two-node (left) and four-node (right). O = no extra delay, Mem = Memory delay (125 ns), SMP = SMP delay (125 ns): K = Kernel, U = User: *Requester-operation*

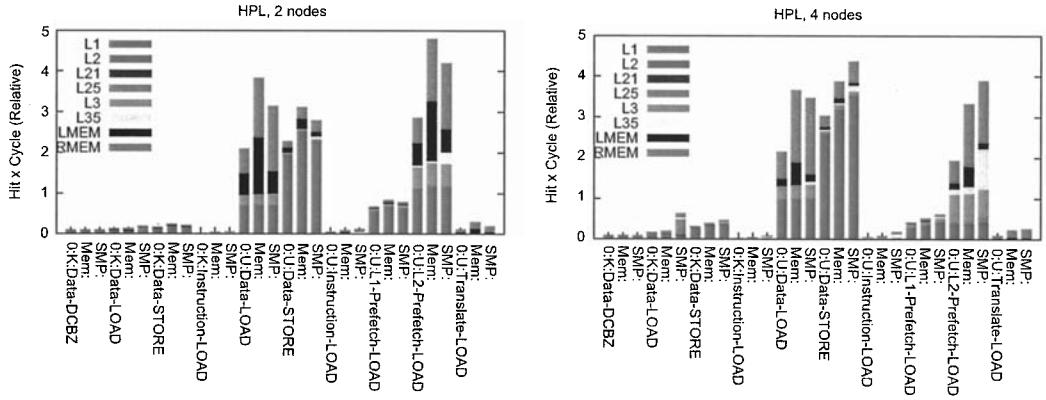

Fig. 6 Cache and Memory Access Patterns in HPL on two-node (left) and four-node (right).

SMP delay compared to the other accesses in the Kernel:Data-LOAD.

Though the data shown is for the four-node system. the two-node and eight-node simulation results confirmed that the trends are similar.

4.4 Observation

The performance with *memory delay* was better than that with *SMP delay* as the number of the SMP nodes increases, as described in Section 4.1. Section 4.2 shows that the same trends exist in the Hit × Cycle metric. As the number of SMP nodes increases, the utilization of the cache is becoming large due to the increasing total size of the available cache memory in the system. The *SMP delay* affects not only the remote cache access but also the local cache access due to the cache coherent protocols. In addition, the *SMP delay* also affects the remote memory accesses that pass through the delayed SMP link. These impacts cause performance degradation, that cannot be ignored in the large SMP server systems. On the other hand, the optical memory link has advantages since it affects only memory access latencies with high-speed cache coherent SMP links. The optical memory link would be a reasonable option for future SMP servers with optical interconnections.

5. Conclusion

We analyzed the performance and utilization of the cache, memory, and their electronic and optical interconnects in SMP servers, and discusses cases where the optical interconnects are applied to either the cache coherent SMP link or the memory link. They were evaluated with a cycle-accurate full system simulation tool using HPC benchmarks, taking care of the extra latency due to the time-of-flight. The simulation shows that the performance of the optical memory link exceeds that of the optical SMP link as the number of the nodes increases. By observing the inside of the processors

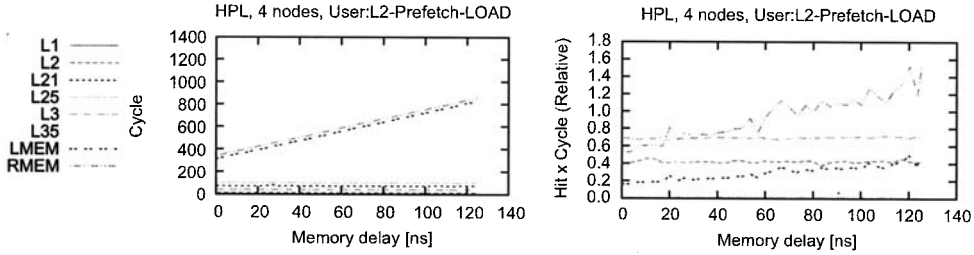

Fig. 7 Cycle and Hit × Cycle versus memory delay for User:L2-Prefetch-LOAD in HPL on four-node.

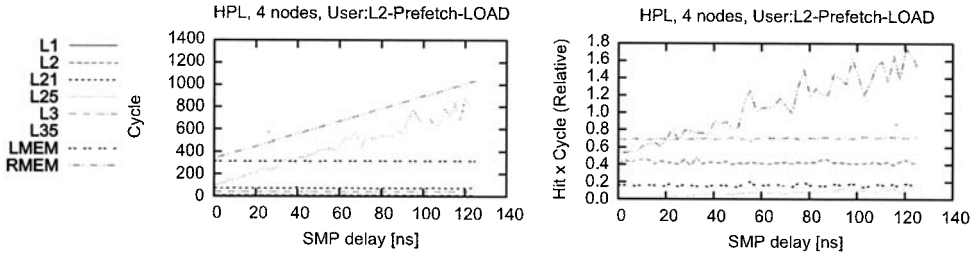

Fig. 8 Cycle and Hit × Cycle versus SMP delay for User:L2-Prefetch-LOAD in HPL on four-node.

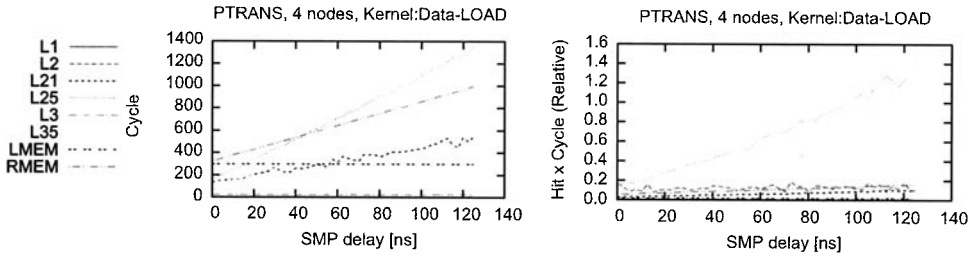

Fig. 9 Cycle and Hit × Cycle versus SMP delay for Kernel:Data-LOAD in PTRANS on four-node.

and the SMP server systems, we determined that the topological differences cause this performance crossover.

We wish to acknowledge helpful advice from Lixin Zhang, James Peterson, and Patrick Bohrer in the IBM Austin Research Laboratory.

References

1) J. Huh, S. Keckler, and D. Burger, "Exploring the Design Space of Future CMPs," *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques (PACT-2001)*, 2001.

2) A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, and M. B. Ritter "Exploitation of optical interconnects in future server architectures." *IBM J. Res. Dev.* 49, 4/5, pp.#755 – 775 (Jul. 2005).

3) Y. Katayama and A. Okazaki, "Optical Interconnect Opportunities for Future Server Memory Systems," *13th International Symposium on High-Performance Computer Architecture (HPCA-13)*, 2007.

4) P. Bohrer, et.al., "Mambo–A Full System Simulator for the PowerPC Architecture," *ACM SIGMETRICS Performance Evaluation Review*, vol.31(4) pp.#8-12, March 2006.

5) Dongarra, J., Luszczek, P. "Introduction to the HPCChallenge Benchmark Suite," *ICL Technical Report, ICL-UT-05-01*, 2005.

6) Greg Burns, Raja Daoud and James Vaigl, "LAM: An Open Cluster Environment for MPI," *Proceedings of Supercomputing Symposium*, pp.#379–386, 1994

7) Jeffrey M. Squyres and Andrew Lumsdaine, "A Component Architecture for LAM/MPI," *Proceedings, 10th European PVM/MPI Users' Group Meeting*, pp.#379–387, 2003

8) J. M. Tendler, J. S. Dodson, J. S. Fields, Jr., H. Le, and B. Sinharoy, "POWER4 System Microarchitecture," *IBM J. Res. & Dev.* 46, No. 1, 5 – 25 (January 2002).

9) Sinharoy, B., Kalla, R. N., Tendler, J. M., Eickemeyer, R. J., and Joyner, J. B, "POWER5 System Microarchitecture," *IBM J. Res. & Dev.* 49, No. 4/5, 505 – 521 (July 2005).