PA-RISC Symmetric Multiprocessing in Midrange Servers

By making a series of simplifying assumptions and concentrating on basic functionality, the performance advantages of PA-RISC symmetric multiprocessing using the HP PA 7100 processor chip were made available to the midrange HP 9000 and HP 3000 multiuser system customers.

by Kirk M. Bresniker

The HP 9000 G-, H-, and I-class and HP 3000 Series 98x servers were first introduced in the last quarter of 1990. Over the lifetime of these systems almost continual advances in performance were offered through increases in cache sizes and processor speed. However, because of design constraints present in these low-cost systems, the limits of uniprocessor performance were being reached.

At the same time, the HP PA 7100 processor chip was being developed. Its more advanced pipeline and superscalar features promised higher uniprocessor performance. Advances in process technology and physical design also promised higher processor frequencies.

Part of the definition of the PA 7100 is a functional block that allows two PA 7100 processors to share a memory and I/O infrastructure originally designed for a single processor. This functional block provides all the necessary circuitry for coherent processor communication. No other system hardware resources are necessary. This feature of the PA 7100 processor made it technically feasible to create a very low-cost two-way symmetric multiprocessing processor board for the HP 9000 and HP 3000 midrange servers. However, significant design trade-offs had to be made to create a product in the time frame necessary.

This article describes the design of this new processor board, which is used in the HP 9000 Models G70, H70, and I70 servers. The HP 3000 Series 987/200 business computer is based on the same processor board.

Design Goals

The design goal of the system was to provide the advantages of symmetric multiprocessing in the midrange servers both to new customers in the form of a fully integrated server and to existing customers in the form of a processor board upgrade. The only constraint was that existing memory, I/O cards, and sheet metal had to be used. Everything else was open to possible change. However, a strong restoring force was provided by the need to minimize time to market and the very real staffing constraints. There simply weren't time or resources to enable us to provide all the features associated with symmetric multiprocessing. The decision was made to make the performance advantages of symmetric multiprocessing the primary design goal for the midrange servers.

Development History

The I-class server was chosen as the initial development platform for the PA 7100 processor. An I-class processor board was developed that accepts a PA 7100 module consisting of the processor package and high-speed static RAMs. In addition, an extender board was developed that allows two PA 7100 modules to be connected to the I-class processor board. This four-board assembly, which was the first prototype of the eventual design, booted and was fully functional within five months of the initial PA 7100 uniprocessor turn-on. This short time period allowed all the basic operating system changes and performance measurements to be made at the same time as the uniprocessor work was being done, by the same design team, with only a small incremental effort.

At this point, the efforts of the design team were centered on introducing the PA 7100 uniprocessor servers. However, since the initial performance measurements of the symmetric multiprocessing prototype were so encouraging, the team continued to refine and develop the initial prototype into a manufacturable product.

The first decision of the design team was to implement the design using 1M-byte instruction and data caches, a fourfold increase over the initial PA 7100 designs. This decision was driven by the initial performance measurements made on prototypes, which showed that the larger caches optimized the utilization of the shared processor memory bus. The same measurements also showed that the most desirable performance levels would require the design to match the previous processor frequency of 96 MHz. This would be the first of the large-cache, high-speed designs for the PA 7100 processor, and would therefore carry considerable design risk.

The next decision was to implement the design not with modules, but as a single board. This was done to lower the cost and technology risk of the design. The shared processor memory bus would be twice as long as in previous designs, but it would not have to bear the additional signal integrity burden of two module connector loads. This was the first of the simplifying assumptions, but it led to several key others.

A great deal of the complexity in symmetric multiprocessing systems arises not just from the problems of maintaining the

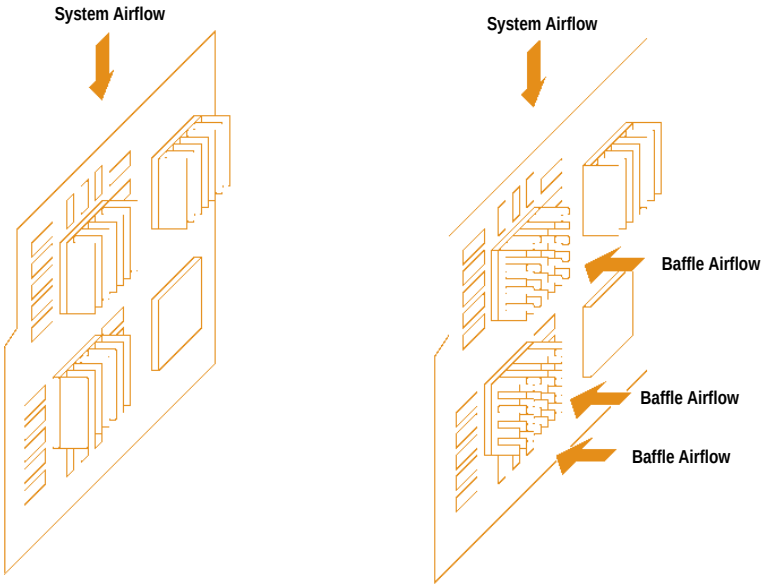

Fig. 1. On the left is the unmodified airflow pattern showing the second processor in the thermal shadow of the first. On the right is the revised airflow pattern showing the impingement cooling provided by the baffle fan.

processors during normal operation, but from handling special operating conditions like failures or booting. Since in this case both processors are always installed, one processor is designated as the "monarch" and is allocated special responsibilities. The second processor is designated as the "serf," and is not allocated any special responsibilities. This obviates the need for a complex method of determining which processor should maintain control during exceptional circumstances. Also, since both processors are on the same board and cannot be replaced independently, it was decided that if one processor should fail, the other would not continue to operate. This removes an entire class of complex interactions that would have had to be discovered, handled, and tested, considerably shortening the firmware development life cycle.

One negative implication of the single-board solution was that one processor was in the direct airflow path of the other (see Fig. 1). This meant that a new solution for cooling had to be devised, but in such a way that the upgrade to the new design would not impact the existing sheet metal. A passive solution of diverting the airflow using air baffles did not prove to be effective enough, so the mechanical design team devised an active solution. A forced-air baffle was devised that is essentially a box occupying the airflow volume next to the processor board. It has three openings centered above the processors and the worst-case cache components. The box is pressurized by a miniature fan. This causes air to impinge directly on the critical components without disturbing the airflow to the rest of the processor board. Since the primary airflow is now normal to the processor board, a new heat sink consisting of a grid of pins was devised to allow the impinging air to cool the processors most efficiently.

One drawback of this active airflow solution is that it relies so heavily on the miniature fan to maintain the processor temperature in a safe range. Of all component classes used in these systems, fans have some of the higher failure rates. Since so much of the air volume next to the processor board is committed to the forced-air baffle, failure of the forced-air baffle fan can cause permanent damage to the processors if not detected in time. In fact, the overheating of the processors was measured to be so rapid in the event of the baffle fan failure that the existing overtemperature protection could not be activated quickly enough. For this reason, the fan is continuously monitored. If the fan stops spinning or rotates slower than a preset limit, the system power supplies are shut down immediately. In addition to providing maximum protection to the processors, this solution also removes the need to burden the software and firmware development with status checking routines.

All of these decisions were made in the background, while the uniprocessor design was being readied for release. In fact, some of the impetus for making the simplifications was the lack of time. However, it was clear that the desire for the system was strong enough for the team to continue. Within one week of the release of the final revision of the uniprocessor system, the initial revision of the multiprocessor processor board was also released. This functional prototype proved to be extremely stable, with no hardware failures reported during the design phase.

Verification

It was at this point that the electrical verification of the design began, and with it the challenging phase of the project as well. The design risks of the large, high-speed caches imagined early on turned out to be all too real. The most problematic aspect of the cache design is that the read access budget for the cache access is one and one half clock cycles (15.6 ns, assuming 96-MHz operation). During that time, the address must be driven to the SRAMs, the SRAMs must access the data, and the data must be driven back to the processor. Current SRAM technology consumes almost 60% of the read budget in internal access time. This budget needs to be maintained over all possible operating conditions, and a single fault can cause either a reload (in the case of instructions) or a system panic and shutdown (in the case of data). The unique problem with this design was that caches this large had never before been run with the PA 7100 processor.

The test methodology used was to run tests tailored to stress the caches while varying the system voltage, temperature, and frequency. Although functional testing at normal conditions had yielded no failures, the initial cache design quickly

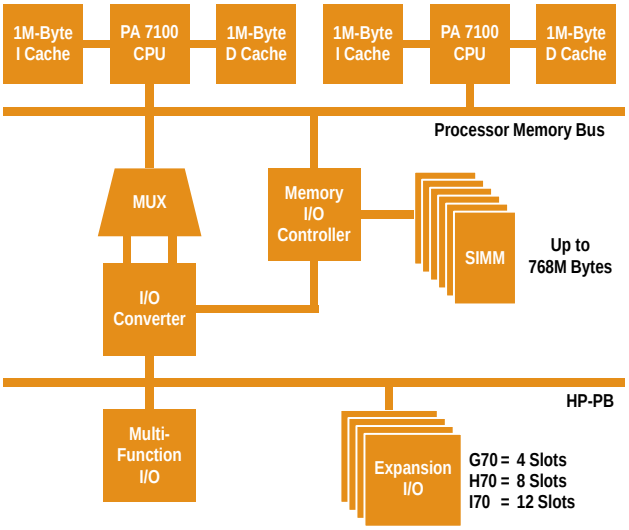

Fig. 2. Block diagram of the HP 9000 Model I70 computer system.

succumbed to the pressures of this type of electrical verification. Analysis of the failures indicated that the read budget was being violated at the combined extremes of low voltage, high temperature, and high frequency. The 1M-byte SRAMs had higher capacitive loads and were physically larger than their lower-density counterparts. This greatly increased the address drive time. The team did not have recourse to faster high-density SRAMs from any vendor, and caches built out of faster lower-density SRAMs would not have provided the symmetrical multiprocessing performance we desired.

What followed was an exhaustive analysis by all three contributors to the design: the PA 7100 design team, the board design team, and the SRAM vendor design teams. Each team worked at pulling fractions of nanoseconds out of the read access. The board design team experimented with termination designs and new layouts to improve address drive time. The PA 7100 team pushed their chip faster to increase the read time budget. They also identified which critical signals had to be faster than all the rest and simulated the board team's changes. The SRAM vendor design teams pushed their processes to achieve faster components. All three teams pushed their designs to the limits, and it took contributions from all three teams to succeed. In the end, it took over six months of constant design refinement and testing to achieve the final result, a design that meets the team's initial electrical verification requirements. This turned out to be the only significant electrical design problem that the processor board team had to solve.

While the board design team worked out the electrical design issues, a separate team was formed to verify the multiprocessing functionality of the PA 7100 processor. This formal verification was the last step in the development cycle for the systems.

System Overview

A block diagram of the Model I70 system appears in Fig. 2. Both PA 7100 CPUs are configured with 1M bytes of instruction cache and 1M bytes of data cache. The processors run at a speed of 96 MHz. The shared processor memory bus is operated at a fixed ratio of 3:2 with respect to the processors, or 64 MHz, and connects the processors to the single memory and I/O controller. The memory and I/O controller interfaces to a maximum of 768M bytes of error corrected memory. The I/O adapter connects a demultiplexed version of the shared processor memory bus to a four-slot (Model G70), eight-slot (Model H70), or twelve-slot (Model I70) HP-PB (Hewlett-Packard Precision Bus) I/O bus.

In addition to the processor board, the base system consists of the HP-PB backplane, a memory extender, a fan baffle, and a multifunction I/O card.

System Specifications

The following specifications are for the 12-slot Model I70 server.

Processors	2 PA 7100 superscalar processors with integrated floating-point unit
Cache	1M-byte instruction cache per processor. 1M-byte data cache per processor
Processor Clock	96 MHz
System Clock	64 MHz
Maximum Memory	768M bytes
I/O Bus	1 12-slot HP-PB
Maximum Integrated Storage	6G bytes
Maximum External Storage	228G bytes
Maximum LANs	7
Maximum Users	3500

Summary

The success of bringing PA-RISC symmetric multiprocessing to the HP 9000 and HP 3000 midrange servers was the result of implementing simplified symmetric multiprocessing functionality. The PA 7100 team integrated all the functionality for two-way symmetric multiprocessing into their design. The system design team followed their lead by creating a system around the two processors that includes only the core hardware and firmware functionality absolutely necessary for operation.

Acknowledgments

The author wishes to acknowledge the members of the hardware and firmware development team under Trish Brown and Ria Lobato: Jim Socha, Janey McGuire, Cindy Vreeland, and Robert Dobbs. Special thanks go to Lin Nease, who implemented the original firmware and operating system changes. The operating system work was done by Anastasia Alexander and Steve Schaniel. Also, thanks go to Jeff Yetter's PA 7100 team, especially Greg Averill, implementor of the processor memory bus sharing functionality, and Tom Asprey, Bill Weiner, and Tony Lelm for their help in the electrical verification of the cache design. Also instrumental in the verification of the cache design were Bill Hudson and the engineers of the Motorola fast SRAM group. The success of the multiprocessor functional verification is the result of the efforts of the multiprocessor verification team of Akshya Prakash. The excellent work of both the electrical and multiprocessor verification teams was evidenced when the Model H70 was chosen as the minicomputer product of the year in the 1993 *VAR Business* reader survey. When asked to explain the nomination, one reader quipped, "It doesn't crash."