High Performance PA-RISC Snakes Motherboard I/O

Andy DeBaets, Wayne Ashby, Sharon Ebner, Maria Lines, Danny Lu, Monish Shah, Rob Snyder, Paul Zimmer

Hewlett-Packard Co.
Entry Systems Lab/User Interface Hardware Lab/ System Interface Lab
19447 Pruneridge Avenue, MS 47L9
Cupertino, Ca 95014

Abstract: *A new low-cost high-performance motherboard I/O design has been developed. The design consists of 93K gates spread across three ASIC chips and a number of PC boards. Key features of the desktop and deskside design include 20 MByte/sec Fast-Wide SCSI-2, link rate FDDI, and CD quality audio.*

1. Introduction

Customer response to the first "Snakes" products, introduced March 1991, was very positive. Design teams began work on upgrades to these products (720/730/750) prior to the March launch. Data from customers indicated an ever increasing need for more MIPs, MFLOPS, vectors/sec, as well as greater disk and network I/O throughput. Market requirements also determined a need to include CD-Audio functionality.

The need for more MIPs, MFLOPS, and greater graphics performance was addressed by HP's PA7100 CPU chip running at 99 MHz. This CPU design is described in a companion paper [1]. The graphics performance scales up with the increased CPU performance [2]. This paper describes the design of a high performance I/O subsystem to keep pace with the dramatic increases in CPU performance. This design provides up to 14 MB/sec of disk throughput and up to 98 million bits per second of network throughput.

The overall design was guided by the following principles: optimized motherboard architecture via ASICs, highest possible performance, lowest possible cost, direct connection to fast system buses, overall minimization of complexity, direct and total control of hardware by the software driver, I/O performance that scales with the CPU, and use of off-the-shelf I/O controllers.

The functionality of the design was implemented in a number of PC board assemblies and three ASICs. These will be described in turn.

2. Board Design

The I/O board contains and controls the I/O interface functions of the computer. The complete assembly includes a network interface module board, a bulkhead with all the mounted I/O connectors, and an extension connector board. The extension board allows a second row of connectors to be mounted to the bulkhead.

The second generation "Snakes" I/O includes all the functionality of the first generation (SCSI, 802.3 LAN, two serial ports, bi-directional parallel port, HP-HIL input device port, EEPROM, real-time clock, and EPROM), plus 16-bit differential SCSI (also known as "fast-wide" SCSI), CD quality audio, and optional FDDI. This functionality is organized as shown in the block diagram (Figure 1). A control chip is the central controller for the I/O peripheral chips, and provides other resources such as arbitration, DMA and interrupt registers. There is one local shared address bus and two 32-bit data buses. FDDI or 802.3 (or both in the 755) sit on the 32-bit bus dedicated to networking; everything else attaches to the second 32-bit bus. A data path chip buffers data between SGC and either of the two 32-bit buses. The

1063-6390/93 $3.00 © 1993 IEEE

disk and network interfaces are separated on these two data buses, allowing maximum system throughput on NFS, and other I/O intensive applications.

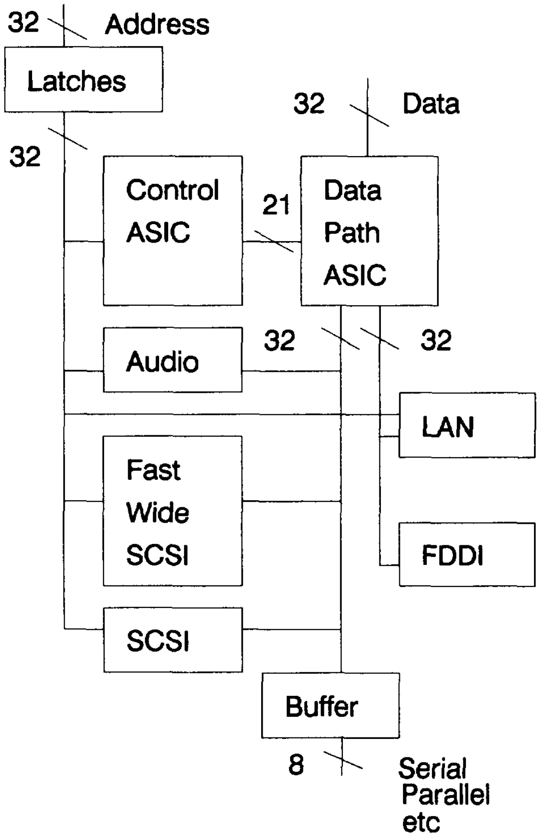


Figure 1

The audio functionality provided includes mono speaker output, stereo headphones output, stereo line input, stereo line output, and mono microphone input. In addition, there is an internal mono speaker mounted to the board and a piezo-electric "beeper".

There are three possible network interface modules: 802.3 LAN BNC, 802.3 LAN AUI, and FDDI. The network interface module boards are 2.3" by 5.5". The network module also has its own mini-bulkhead, which is mounted flush to the main bulkhead by two screws.

The main board utilizes a double sided surface mount assembly process with topside through-hole components and 20 mil lead pitch surface mount devices on the bottom side. This board is .062 inches thick and composed of 8 layers; 3 power planes and 5 signal layers. The dimensions are 8.25" by 11.2". The design rules support 6 mil line widths and 6 mil spaces. The solder mask is dry film, and the board material is standard FR-4 solder mask on bare copper with hot air leveling.

The final PCB is a small 1.3 x 8.24 inch extension board which contains the connectors for the parallel port, two RS232 ports, HP-HIL, five audio jacks and microphone analog buffer circuitry. Its connection to the main board is accomplished through a "riser" connector.

3. Control ASIC

The motherboard I/O connects into the system via a high-speed Standard Graphics Connection (SGC) interface. The SGC is a pipelined protocol that connects the processor and memory with built-in I/O, graphics, and the EISA interface. Like the first generation "Snakes" machines, the 735 and 755 have a central I/O subsystem controller which manages the interfaces to SGC and all the peripheral controllers, and also provides local arbitration, interrupt registers, and a DMA/FIFO engine for the parallel port.

The AMD FDDI chipset lacks a complete DMA engine for accessing system memory. This function is provided and optimized inside of the control chip. This control ASIC strobes data between the FDDI chipset FIFO and the data path ASIC's FIFO. The NCR53C720 fast-wide SCSI controller has a built-in FIFO and a complete DMA state machine. Neither the FDDI nor fast-wide SCSI DMA engines can transfer directly to SGC. A common state machine in the control chip acts on their behalf to generate addresses during SGC burst transfers. Local address bandwidth requirements can be met with only one address bus.

3.1 Local Arbitration

To achieve high performance, multiple data buses and simultaneous activity of those buses is supported. This required increasing the complexity of the local arbitration function within the control chip. The arbitration block becomes active under the following circumstances: LAN, FDDI, fast-wide SCSI, SCSI, audio, or the parallel port attempt DMA, a programmed I/O transaction takes place, fast-wide SCSI needs to do a local data transfer, or FDDI needs to transfer data locally. The local devices request ownership to the control chip, and the control chip requests bus ownership globally while prioritizing between local requestors.

Priorities are set in the following order: CPU programmed I/O transaction, audio, LAN, fast-wide SCSI data transfer, FDDI inbound, FDDI outbound, SCSI, and parallel port. Priorities are based upon the maximum latency that a device can tolerate and the severity of system impact if the latency requirement is not met. Other factors included bus ownership time, and frequency of bus requests. The system bus is released after each local bus ownership to minimize the latency seen in other parts of the system. The characteristics of the devices in the system, combined with the prioritization scheme, ensure that starvation never occurs.

	Latency (usec)	bus hold (usec)	transfer rate MB/sec	% bus used
Audio	83	1.3	12	2
LAN	52	5.1	12	10
fwSCSI	large	.5	64	31
FDDI	23,000	.9	79	16
SCSI	large	.6	27	19
parallel	large	2.9	6	7

This standard cell control chip is a 32,146 gate chip in a 240 PQFP extra fine pitch package. The die is 10.3 mm by 11.2 mm, simultaneously area and pad limited. This chip is fabricated in HP's high volume 1.0 micron process (CMOS-34). 33 Mhz operation is supported and power dissipation is less than 1 watt.

4. Data Path ASIC

The introduction of new functions dictated the incorporation of several novel design features. High data rates for fast-wide SCSI and FDDI DMA (20 MB/s and 12 MB/s) require the most efficient use of available SGC bandwidth by using burst mode DMA. This led to the design of a semi-custom data path chip. The overall design achieves a transfer rate of 1 word per cycle while incurring only a 2 cycle penalty per arbitration period.

The data bus for the network function (FDDI and/or 802.3) is separated from the rest of the subsystem. This resulted in the reduction of data bandwidth contention and enabled simultaneous disk and network transfers. Putting FDDI and 802.3 on the same data bus also facilitated implementation of the modular network interface by minimizing the number of signals on the connector. The datapath chip facilitates this structure by integrating muxes for the two local data buses with high-speed FIFOs for FDDI and fast-wide SCSI burst DMA. The FIFOs allow the design to match the SGC burst pipeline with the FDDI and fast-wide SCSI controller protocols. The data path ASIC also has as a simple buffer path around the FIFOs for directed I/O and non-burst DMA.

A block diagram for the data path ASIC for the I/O interface is shown in Figure 2. It contains a 64 byte FIFO for fast-wide SCSI and two 88 byte FIFOs for FDDI, one for each direction. It latches data for slave transactions and single-word DMAs with registers between the local data buses and the SGC backplane in both directions.

The 64 byte fast-wide SCSI FIFO was designed to be as flexible and fast as possible. It can be used in either direction and must be emptied or reset before switching directions. It can be filled and emptied at the same time, to allow data to be written through. It can run at 133 MBytes/sec and directly monitors the bus to determine when to clock in data. Any number of pieces of data can be written in or read out in one operation.

Two FIFOs are provided to perform simultaneous FDDI inbound and outbound DMA. These FIFOs provide the data paths for the full-duplex DMA controller. The FIFOs cannot be written through (written and read at the same time), but both can be written into at the same time. Electrical noise is minimized by preventing the simultaneous read of both FDDI FIFOs.

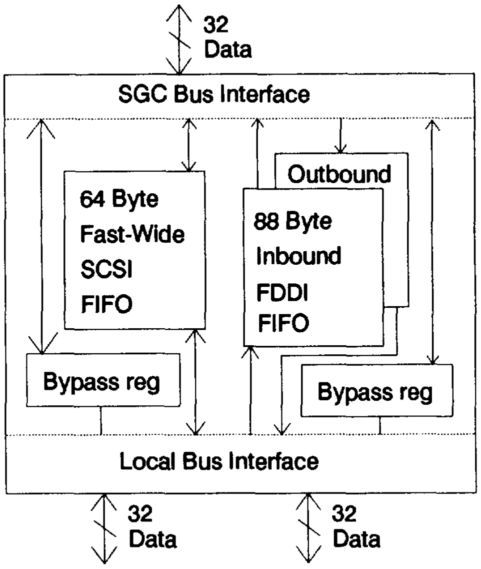


Figure 2

The FDDI FIFOs are asynchronous, with the I/O side running at 25 MHz and the backplane side running at the SGC speed (typically 33 MHz). The data entering the FDDI FIFO from the network side is also accumulated as a checksum inside the data path chip. These 88 byte FIFOs provide efficient transfers to and from the system as well as providing clock domain synchronization.

For programmed I/O transfers, as well as single-word DMA transfers to FWSCSI, data is latched into a register. There are four 32-bit registers latching data between the two local buses and SGC. All data going to or from the I/O board must go through the data path chip, either through the latches or through the FIFOs.

This standard cell data path chip consists of 56,174 gates on a 8.3 mm by 8.4 mm die packaged in a 208 pin PQFP extra-fine pitch package. This chip is designed in HP's high-performance .8 micron, 3-level metal CMOS-26B process and dissipates just under 1.5 watts.

5. Disk Interfaces

The disk interfaces on the Outfield board consist of a 16 bit differential SCSI port and a separate 8 bit single-ended SCSI port. The former connects high speed disk drives/arrays through a 68 pin high density "P" connector (per the SCSI-3 specification) and the latter connects slower disk drives or peripherals through a 50 pin high density A connector.

The 16 bit differential-ended SCSI port is also known as the fast-wide SCSI interface. The wide data path allows a maximum data rate of 20 MByte/sec on the SCSI bus and 15 SCSI target devices. The differential feature provides better signal noise margin, greater reliability, and longer cable lengths (up to a maximum length of 25 meters).

Main components of the fast-wide SCSI subsystem are the NCR53C720, the control ASIC, and the data path ASIC. The NCR53C720 is an intelligent SCSI controller. It has a 32 bit DMA interface on the host bus side which transfers data at a maximum rate of 97 MByte/sec. On the SCSI bus side, it has a 16 bit differential data bus with maximum data rate of 20 Mbyte/sec. To maximize transfer rate, a burst mode of operation of the NCR53C720 is used.

The control chip converts the burst mode protocols of the NCR53C720 to SGC and provides control signals to the data path chip's fast-wide SCSI FIFO. For register accesses and non-burst DMA to the NCR53C720, data goes through the simple buffer path in the data path ASIC. During disk reads, data from the SCSI bus are first stored in the data path FIFOs allowing the NCR53C720 to continue to gather data from the disks without penalties or system arbitration delays. For disk writes, data is prefetched (up to 64 bytes) into the data path chip and then provided to the NCR53C720 at the cache burst rate.

The 8 bit single-ended SCSI design is taken from the 720/730/750 workstations. This SCSI port has a maximum data rate of 5 Mbyte/sec. The heart of this interface is the NCR53C700 SCSI controller. Refer to [3] for more details.

6. FDDI

The FDDI interface is designed to provide link-rate networking performance for workstation clients in a Client/Server environment at minimum cost. Link-rate FDDI is media-limited throughput, 98 million bits per second. The focus of the design is on helping the CPU achieve high-throughput, low-latency networking through appropriate hardware assists, while keeping overall complexity low.

One key to high-performance networking is to avoid "touching" the data. Ideally, inbound data should be split from the header information, with the data arriving page-aligned and fully checksummed in host memory. (A checksum is a value calculated using all transmitted data - it is used to detect data loss). In this way, the data may be passed to the application by simply re-mapping the page into user space. In the HP735/755, moving the data by remapping the virtual pages takes a fifth of the time of moving the data by copying memory to memory.

The FDDI interface assists the CPU in doing header/data split and checksumming on inbound packets. The driver programs the inbound DMA controller to bring up enough of the packet to get the header and some amount of data. A variable byte offset in the AMD 79C830 FDDI controller chip is used to get the data word-aligned in memory. When the DMA completes, the driver parses the header under interrupt, copies over any data to a new, page-aligned buffer, then uses the linked Address/Count pairs in the inbound DMA controller to deposit the remainder of the data in this buffer, with any residual data and the controller's descriptor word going to a third buffer. All data is checksummed as it goes through. When the final DMA is complete, the driver reads the checksum value, corrects it by subtracting header data, and passes it up to the network transport layer without ever touching the data. This results in two interrupts of overhead for large packets, with very small data movement overhead. The size of the initial header/data DMA is chosen based on network workload data to pull in most small packets with only one interrupt. The small packets are then moved directly by the CPU.

Outbound packets are handled directly by the CPU with little hardware assist. Accessing the data is minimized by combining the network protocol copy with a simultaneous checksum calculation. In keeping with the theme of simple, direct control by the CPU, there is no "node processor" in the FDDI interface to run the Station Management (SMT) code. Instead, the FDDI Interface provides a direct path to the controller and physical interface chips, and the SMT protocol is run on the host CPU. SMT protocols handle functions such as adding or removing systems on the FDDI network.

To avoid the complexity associated with sharing a single DMA channel between inbound and outbound data flow, the FDDI Interface provides independent inbound and outbound DMA channels. Any contention between the two channels, such as access to common resources like the system bus or the AMD controller chip, is handled directly in the hardware. The hardware ensures that neither channel can be blocked by the other. In this way, the inbound and outbound portions of the driver can be designed and optimized separately, without the burden of special-case code for sharing the channel.

7. CD Audio

To enable multimedia functionality, all second generation PA-RISC workstations contain a built-in audio sub-system. This enhances the workstation's ability to act as a communications machine. Adding an audio sub-system to a HP-UX (or any Un*x) workstation introduces some unique design challenges. Audio is a continuous-time data type, requiring a predictable response from the system to keep an audio stream going.

HP-UX does not guarantee predictable response times because most tasks do not have any such requirements. Since changing HP-UX was not practical on a tight schedule, the problem needed to be solved in the existing framework. The solution was to come up with a reasonable upper bound on the response time, instead of an absolute upper bound. The system was designed around that upper bound, with plenty of margin. That meant using adequately large buffers for audio data at every point in the system.

The audio sub-system can record and playback audio simultaneously. Recording refers to digitization of analog input, while playback means converting digital audio back to analog form. While record and playback paths are generally separate, a path is provided for feeding recorded data back to the playback path with no delay. This allows the user to monitor what is being recorded. Both mono and stereo modes are supported. The audio data is represented either as 16 bits per sample or as 8 bits per sample. The 16 bit values are simply linear values, while 8 bit values are encoded as either μ-law or A-law (μ-law and A-law are telecommunications standards to encode voice audio in digital form). The audio sub-system supports many different sample rates, each of which has some legacy in either consumer audio or telecommunications. Mechanisms for controlling gain are provided for both inputs and outputs. For inputs, they act as recording level controls. For outputs, they are volume controls. Independent controls are provided for right and left channels, so left/right balance can be controlled. Since the gain controls are implemented digitally, they are under software control.

The audio sub-system uses DMA to access data. DMA reduced the buffering requirements because hardware response time for mastering the bus is much faster than the CPU's response time to an interrupt. The FIFOs required to address hardware latency were small enough to fit inside a small gate array. Figure 3 shows the block diagram of the audio sub-system.

The DMA hardware is responsible for writing recorded data and reading playback data to/from main memory. For playback, a "current address" register points to the next location from which the data must be read within a 4K byte page. When the pointer advances to the end of the page, the hardware interrupts the CPU for the next address. However, to buffer the CPU interrupt latency, a "next address" register keeps track of the next page address, which was supplied by the CPU in a previous interrupt. In effect, the "next address" is transferred to the "current address", and then the CPU writes a new value into the "next address" register. This allows the CPU to stay sufficiently ahead of the hardware. A similar pair of registers are provided for recording.

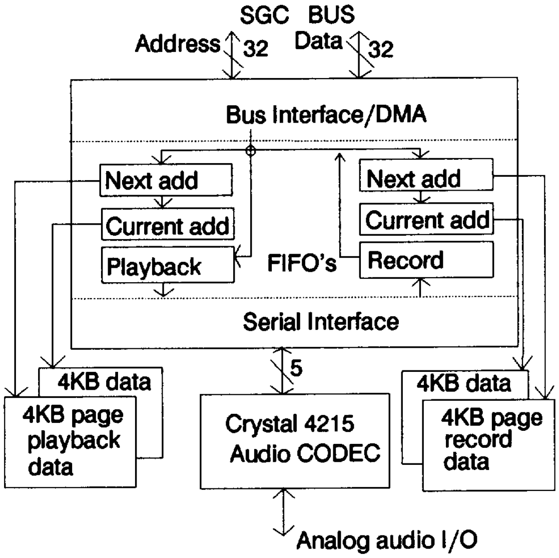


Figure 3

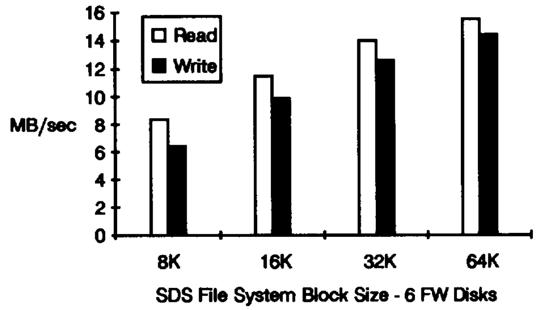
Separate FIFOs are provided for recording and playback. The FIFOs are 32 bytes, while DMA transactions transfer 16 bytes at a time. During playback, a DMA transaction is initiated when the playback FIFO is half empty. This ensures that there is enough room in the FIFO for all 16 bytes fetched by the DMA. Also, as long as the transfer is completed before the remaining 16 bytes in the FIFO are consumed by the Crystal 4215 audio CODEC (coder-decoder), there will be no breaks in the audio. There is a similar protocol for recording.

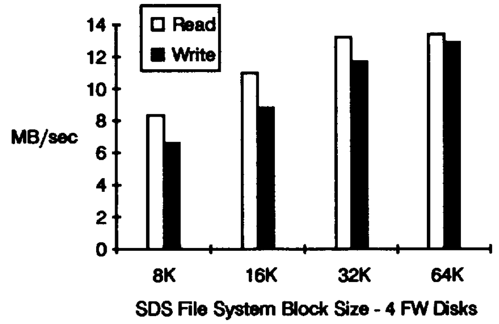
The Crystal 4215 audio CODEC consists of two A/Ds (analog to digital converters) and two D/As (digital to analog converters). These allow simultaneous record and playback in stereo mode. The CODEC also contains the logic required for filtering, oversampling, gain controls, and clock generation. The CODEC has a serial port for communicating the digital audio data. Hence, a serial interface is required to communicate data between the FIFOs and the CODEC. As indicated in the figure, the DMA logic, the FIFOs, and the serial interface are all implemented in a gate array. This gate array consists of 4,952 gates and is packaged in a 120 pin PQFP package.

8. Performance

8.1 Disk throughput

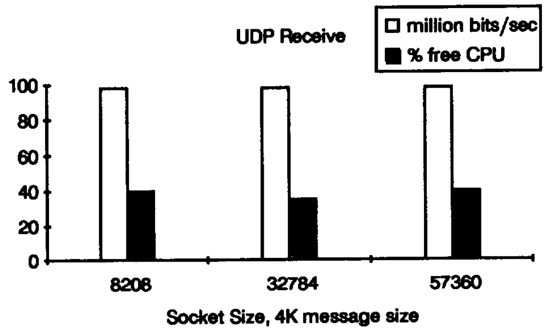
The graphs below show the disk performance achieved through this new design. Sequential disk throughput is measured with HP-UX 9.01, using HP's SDS (Software Disk Striping) technology. The disks used are HP's new Fast-Wide SCSI C2247's. The throughput levels shown reach up to 14 MB/sec, twice the level of the first generation machines while not using any expansion slots. The design improves both components of the price/performance vector: performance is approximately doubled, while price is reduced by the low-cost motherboard implementation and the fact the functionality can be provided in a lower-priced desktop system.

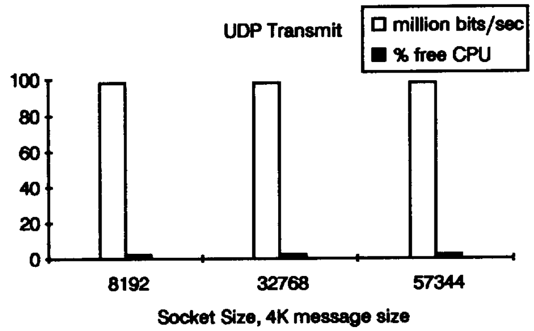


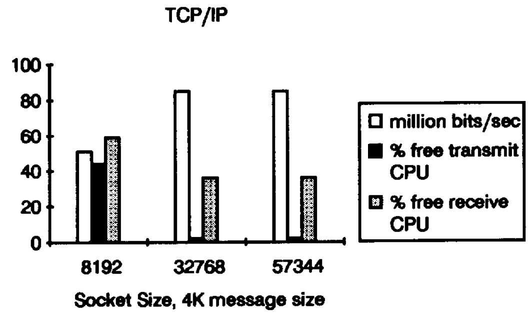


8.2 Network throughput

Shown below are graphs showing UDP and TCP/IP memory-to-memory performance. These measurements were taken between two 735s running "netperf" an HP network performance measurement tool, which has been contributed to the public domain.







The graphs above show link rate performance on UDP (98 million bits per second), and just below link rate (85 million bits per second, limited by the transmitting system) on TCP/IP. For comparison, the motherboard FDDI performance is approximately three times the performance of HP's previous generation FDDI offering, at a lower price.

9. Conclusion

This design provides the high performance I/O needed to keep up with the industry's ever-increasing RISC CPU performance. Overall motherboard disk performance was increased by a factor of four and motherboard network performance by a factor of ten. This was done while adding new CD-Audio features. The overall hardware and software design was also completed on a tight schedule. The first hardware prototype booted the operating system on the very first attempt (no blue wires and no firmware bugs). All of the chips described released to production on the initial mask sets.

Acknowledgments

We would like to acknowledge the following people who directly contributed to this design: Byron Alcorn, Rich Carr, Pia Chamberlain, Carl Haney, Bob Hansen, Mohammad Hatami, Seno Judaprawira, Daniel Li, Victor Martin, Rayka Mohebbi, Tom Parker, Arlen Roesner, Deborah Savage, Ravi Sharma, Joe Steinmetz, and Christie Wilde. Additional thanks go to the verification team of Ali Ahi, Greg Burroughs, Audrey Gore, Steve LaMar, Robert Lin, and Alan Wiemann. Many thanks to our software partners: John Marvin, Rich Testardi, Venkat Rao, Fatima Yu, Alan McGowen, and Annie Wang. Finally, thanks to the management team of Steve Foster, Cliff Loeb, Marlu Allan, and Denny Georg for providing support for this effort.

References

[1] Keane, E., McGuire, P., "HP9000/735 Second Generation PA-RISC Snakes Workstations" Compcon Spring 93: Digest of Technical Papers (Feb 93).

[2] Dowdell, C., Thayer, L., "Scalable Graphics Enhancements for PA-RISC Workstations" Compcon Spring 92: Digest of Technical Papers (Feb 92).

[3] Li, D., Gore, A., "HP9000 Series 700 Input/Output Subsystem" Hewlett Packard Journal, Vol. 43, no. 4, August 1992, pages 26-31.

[4] Lee, R., B, "Precision Architecture", IEEE Computer, Vol. 22, No. 1, Jan 1989, pp. 78-91.