A 300–MHz 64–b Quad-Issue CMOS RISC Microprocessor

Bradley J. Benschneider, Andrew J. Black, *Member, IEEE*, William J. Bowhill, *Member, IEEE*, Sharon M. Britton, Daniel E. Dever, Dale R. Donchin, *Member, IEEE*, Robert J. Dupcak, Richard M. Fromm, Mary K. Gowan, Paul E. Gronowski, Michael Kantrowitz, *Member, IEEE*, Marc E. Lamere, Shekhar Mehta, Jeanne E. Meyer, Robert O. Mueller, Andy Olesin, Ronald P. Preston, *Member, IEEE*, Donald A. Priore, Sribalan Santhanam, Michael J. Smith, and Gilbert M. Wolrich

Abstract—**This 300 MHz quad-issue custom VLSI implementation of the Alpha architecture delivers 1200 MIPS (peak), 600 MFLOPS (peak), 341 SPECint92, and 512 SPECfp92. The 16.5 mm × 18.1 mm die contains 9.3 M transistors and dissipates 50 W at 300 MHz. It is fabricated in a 3.3 V, four-layer metal, 0.5 μm, CMOS process. The upper metal layers (metal-3 and metal-4) are primarily used for power, ground, and clock distribution. The chip supports 3.3 V/5.0 V interfaces and is packaged in a 499-pin ceramic IPGA. It contains an 8-kbyte instruction cache; an 8-kbyte, dual-ported, data cache; and a 96-kbyte, unified, second-level, 3-way set associative, fully pipelined, write-back cache. This paper describes the circuit and implementation techniques that were used to attain the 300 MHz operating frequency.**

I. Introduction

A SECOND-GENERATION Alpha RISC microprocessor has been designed that operates at an internal clock frequency of 300 MHz. The 16.5 mm × 18.1 mm die contains 9.3 million transistors and delivers a peak performance of 1.2 billion instructions per second (BIPS) and 600 million floating point operations per second (MFLOPS). This chip has attained measured performance of 341 SPECint92 and 512 SPECfp92. The chip is implemented in a 3.3 V, 4-layer metal, 0.5 μm, CMOS process and is housed in a 499-pin interstitial pin grid array (IPGA) package. Power dissipation is 50 W from a 3.3 V supply at 300 MHz. Fig. 1 shows a photomicrograph of the chip with an overlay showing all major sections.

The high performance of this second-generation implementation results from many factors, including:

- 0.5 μm CMOS process technology;
- 300 MHz internal clock frequency;
- grid based power and clock distribution;
- fast and versatile latching scheme;
- innovative circuit techniques;
- advanced design and verification tools.

In addition, several architectural improvements over the first Alpha implementation [1] are included in this design. The key architectural performance features are four-way superscalar instruction issue; a high-throughput, nonblocking memory subsystem with low latency primary caches; a large second-level on-chip write-back cache; and reduced operational latencies in all of the functional units.

Manuscript received May 4, 1995; revised August 24, 1995.
The authors are with Digital Semiconductor, Hudson, MA 01749 USA.
IEEE Log Number 9415232.

II. Architecture

As shown in Fig. 2, the chip is functionally partitioned into the following major sections: the instruction unit (I-Box), the integer execution unit (E-Box), the floating point unit (F-Box), the memory management unit (M-Box), and the cache control and bus interface unit (C-Box). The chip features two levels of on-chip cache. The first level consists of an 8-kbyte instruction cache (I-Cache) and an 8-kbyte data cache (D-Cache). The second level is a 96-kbyte unified instruction and data cache.

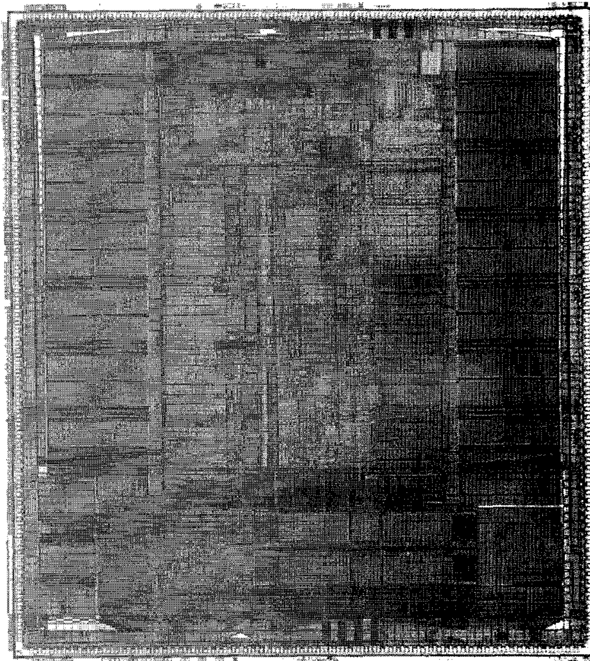
The I-Box contains the 8-kbyte, direct-mapped I-Cache, an instruction prefetcher and associated refill buffer, branch prediction logic, and a 48-entry, fully associative instruction translation buffer. The I-Box can issue up to two integer and two floating point instructions per cycle. Instructions are issued in-order but may complete out-of-order.

The E-Box contains two execution pipelines and a register file for integer operands. Both E-Box pipelines execute load, arithmetic, and logical instructions. In addition, one of the pipelines executes shift and store instructions, while the other pipeline completes jumps and branches. Multiply instructions are executed in a separate unit attached to one of the pipelines. Both pipelines implement full register bypassing, allowing the results from all function units to be available for immediate use. All integer instructions except multiply complete in one cycle.

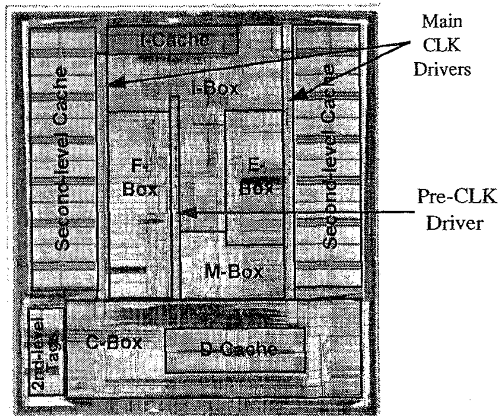
The F-Box contains a register file for floating point operands and two execution pipelines. One pipeline executes multiply instructions while the other executes all remaining instructions. Divide instructions are executed in a separate unit attached to one of the pipelines. All floating point instructions except divide execute in four cycles, a two-cycle reduction from the previous implementation.

The M-Box contains the 8-kbyte, direct-mapped D-Cache, a fully-associative, 64-entry, data translation buffer (DTB), a miss address file for queuing and merging misses from the first-level caches, and a write buffer. The M-Box processes load, store, and memory barrier instructions.

0018–9200/95$04.00 © 1995 IEEE



(a)



(b)

Fig. 1. Photomicrograph with overlay.

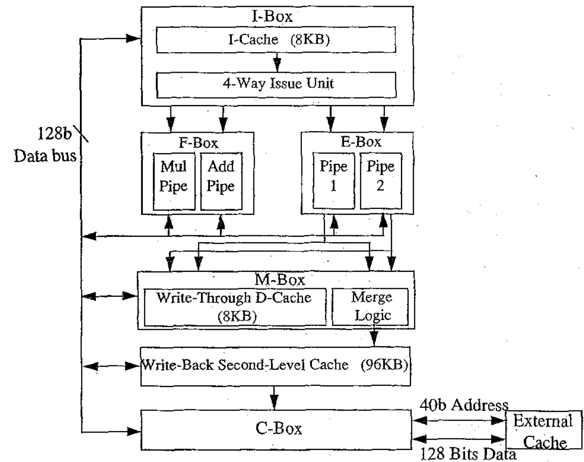


Fig. 2. Block diagram.

The C-Box controls the on-chip, second-level cache and an optional, off-chip, third-level cache and implements a flexible, user-configurable interface to the system. The second-level cache is a fully pipelined, 96-kbyte, three-way set associative, write-back cache. It reads or writes 16 bytes per cycle, providing a peak bandwidth of 4.8 Gb/s at 300 MHz. A 128-b data bus is shared between the optional, off-chip, third-level backup cache and the memory system.

This microprocessor also contains a number of on-chip testability features, including built-in self-test and self-repair of the I-Cache, linear feedback shift registers placed throughout the chip to improve fault coverage, a parallel debug port for monitoring internal chip nodes in real time during chip and system debug, and an IEEE 1149.1 test access port.

TABLE I
TECHNOLOGY CHARACTERISTICS.

Feature size	0.5 μm
Channel length (L_{eff})	0.365 μm
Gate oxide	9.0 nanometers (nm)
Vtn/Vtp	0.5/-0.5 V
Power supply	3.3 V (supports 3.3/5.0 V interface)
Substrate	P-epi with n-well
Salicide	Cobalt di-silicide in diffusions and poly
RAM cell	6T, TiN buried contact
Metal-1	0.84 μm thick AlCu, 1.50 μm pitch (contacted)
Metal-2	0.84 μm thick AlCu, 1.75 μm pitch (contacted)
Metal-3	1.53 μm thick AlCu, 5.00 μm pitch (contacted)
Metal-4	1.53 μm thick AlCu, 6.00 μm pitch (contacted)

III. Global Implementation

The high internal clock frequency, power requirements, and transistor count of the chip required substantial global planning prior to circuit implementation. Technology development, floorplanning, power and clock distribution, and latching strategies were areas of particular importance.

A. Process Technology

The chip is fabricated in a 0.5 μm, 3.3 V, n-well, CMOS process. The major process characteristics are shown in Table I. Technology development occurred in parallel with chip development. Close cooperation between the two development teams allowed both the technology and the chip to be optimized for performance. Metal-4 was added to the technology for power and clock distribution.

B. Floorplanning

The process of floorplanning was initiated during the microarchitectural definition of the chip and was key to achieving the performance goals. Several critical speed paths were iden-

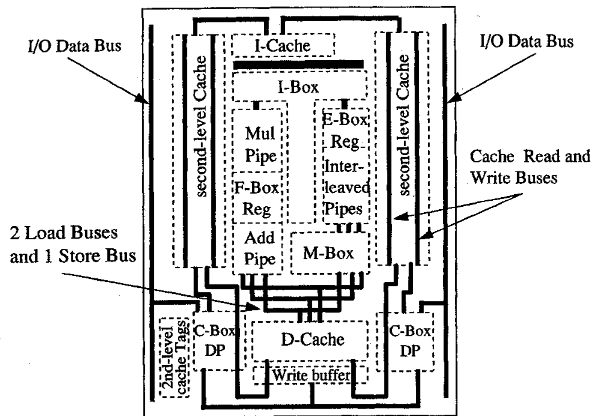


Fig. 3. Floorplan.

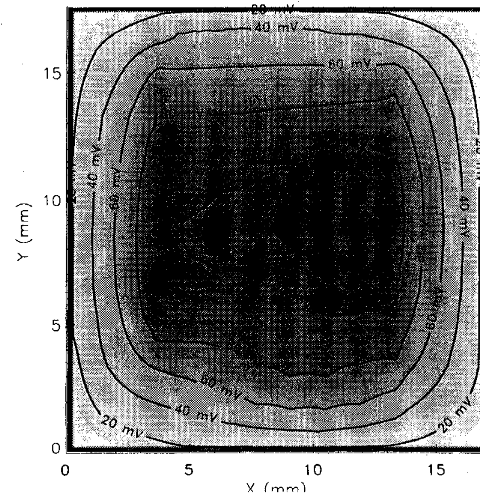


Fig. 4. VDD voltage drop contour plot.

tified and the final floorplan was formulated to optimize these paths. Global interconnect channels were defined early in the project to give priority routing to critical signals. With reference to Fig. 1, metal-4 and metal-2 are routed vertically and metal-3 and metal-1 are routed horizontally. The final floorplan of the chip is shown in Fig. 3.

The second-level cache, the C-Box datapath, and the I/O pins were split into two halves along the left and right edges of the chip. This was done to optimize the routing of data into and out of the chip. The I-Cache was positioned at the top to feed the I-Box decoders and to receive fill data from the pins on the same buses as the second-level cache, minimizing both fill latency and interconnect area. The D-Cache was placed at the bottom of the chip, beneath the E-Box, to minimize the interconnect length of data signals to the E-Box register file. The adder used to calculate the address displacement in the E-Box, and the DTB in the M-Box were placed next to each other. This design allowed the result of the address addition to flow directly into the DTB with little interconnect delay, facilitating a two-cycle D-Cache hit latency.

Different performance requirements led to different floorplan solutions in the F-Box and E-Box. The critical nature of the E-Box data bypass loop required the two integer pipes to be interleaved. This optimized the bypass bus at the expense of widening the pitch of the datapath. The F-Box, having fewer bypass paths, was organized with the register file in the center of the two pipes, effectively sharing vertical metal slots and hence minimizing the area of the datapath.

C. Power Distribution

A high quality power distribution network was essential to meet the power dissipation and performance goals of this chip. Metal-3 and metal-4 are used extensively to distribute power and ground across the chip. These metal layers are approximately twice as thick as lower level metals, thereby offering substantially lower resistance, and are used to form dense grids. Alternating power and ground lines comprise the majority of the grids with a single clock line interspersed every few pairs. The typical drawn line width for V_{DD} and V_{SS} is 12 μm. Power and ground lines in local cells are connected to metal-2 lines that are long enough to span two or more of the metal-3 grid lines. This allowed for a simple and automated procedure to connect local logic to the grid. In addition, 160 nF of on-chip distributed decoupling capacitance was added to minimize switching noise. The clock drivers are closely surrounded by 35 nF of the decoupling capacitance.

The power grid underwent significant electrical verification to ensure its compliance with strict long-term reliability requirements and to guarantee minimal voltage drops in the grid. The analysis was based on full chip capacitance and resistance nodal extracts and used nodal switching data from logic simulations to estimate how much current was conducted through the supply networks. This analysis showed a maximum voltage drop of only 110 mV per rail, demonstrating that the metal-3 and metal-4 grid could effectively distribute power to all devices. Fig. 4 is a contour plot that shows the simulated average voltage drop in the V_{DD} network across the chip. The V_{SS} network exhibited similar characteristics.

Power is supplied to the package through 205 of the 499 pins. The connection between the die and the package is achieved using 282 bond wires. The package outer lead bond pads are arranged in two tiers around the die cavity. The die bond pads are evenly spaced around the chip using an alternating pattern of power and signal (i.e., V_{SS} signal V_{DD} signal V_{SS} etc. ...). This configuration allowed all supply pads to be bonded to the lower tier of landing pads in the package with short, low inductance bond wires. To help dissipate the heat generated on chip, a custom ceramic IPGA package was used. It incorporates an intrusive copper tungsten slug that provides a low thermal impedance contact between the die and a removable heat sink. The slug results in a θ_{JC} for the package of only 0.45°C/W. Although the chip dissipates 50 W, it is effectively cooled by conventional thermal management techniques.

D. Clock Distribution

Another critical aspect of the global chip implementation was the buffering and distribution of the two-phase, single wire

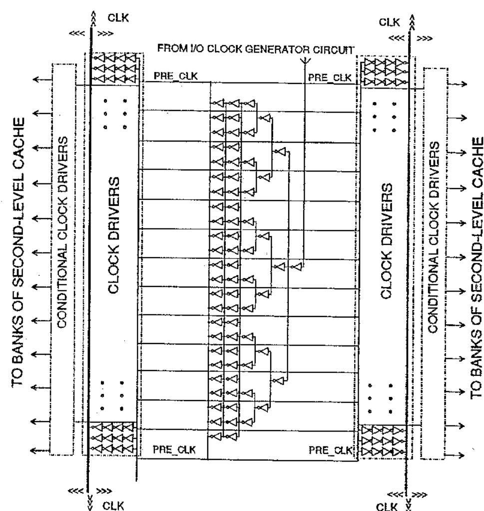


Fig. 5. Clock generation network.

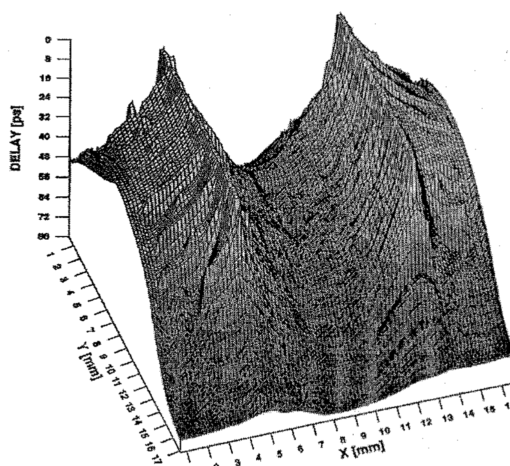


Fig. 6. Clock *RC* contour plot.

clock. The chip receives a 600 MHz differential ECL oscillator signal that is level shifted and divided by two to produce a 300 MHz, 50% duty cycle clock. As shown in Fig. 5, this clock is routed to the center of the chip where it is buffered up and fanned out through a balanced tree of inverters to generate the PRE_CLK signal that drives the main clock drivers. The main clock drivers provide four additional levels of buffering to generate the single wire clock (CLK) that is distributed over the entire chip using a grid of metal-3 and metal-4. The final CLK driver inverter has a transistor width of 58 cm and drives a load of 3.75 nF.

In addition to the single wire clock, two sets of conditional clocks are generated for the second-level cache. These clocks conserve power by conditioning the bank enable signals with preliminary address decode information when accessing the cache. This is described in more detail in Section IV-D.

A low-skew clock signal was critical to meeting the cycle time goal of the chip. The CLK network was designed to minimize the following three components of clock skew: driver and RC delay variations in the PRE_CLK driver network; variations in the transistor characteristics of the main CLK drivers; and RC delay through the global metal grid. Techniques used to reduce the individual clock skew components included: connecting the common nodes of the PRE_CLK and CLK drivers with low resistance upper level metal; tuning the PRE_CLK driver by adjusting the widths of metal wires to balance the RC delays; working with the process development group to design a modular clock driver block to minimize polysilicon processing variations; and specifying design rules that limited the RC delay of the global metal grid.

The CLK driver network was subjected to extensive simulation to evaluate the impact of RC delay on the chip's various circuits. The simulation used layout extracted resistance and capacitance data and contained one million resistors and two million capacitors. Fig. 6 is a three-dimensional inverted contour plot that shows the results of the RC delay simulations. The two peaks seen on the contour plot represent the locations of the main clock drivers. As shown on the plot, the maximum RC delay is 80 ps. The PRE_CLK network and the distributed CLK drivers were also analyzed with SPICE to assess the impact of the power supply noise and cross-chip device characteristic variations on CLK skew.

E. Latch Design

This chip primarily uses dynamic, level-sensitive, pass-transistor latches. This type of latch was chosen to minimize the propagation delay of data through the latch. Since the chip uses a single wire, two-phase clocking scheme, two types of latches are utilized in the chip: an A-phase latch, which is open when CLK is a logic-1; and a B-phase latch, which is open when CLK is a logic-0. Fig. 7(a) and (b) shows examples of the standard A-phase and B-phase latches. These latches were carefully designed and laid out to minimize capacitive coupling effects onto the dynamic node (ZZ_XD in the figures). Several basic latches were designed prior to implementation and were maintained in a library for use by all designers to minimize the number of latch variants.

The overhead associated with latching data in critical paths was further reduced by building simple logic functions into the input and output gates of the latch. An example of a standard A-phase latch with a NAND function at the front end is shown in Fig. 7(c). Fig. 7(d) shows a sample configuration where logic has been built into both stages of the latch, effectively reducing the latching delay to that of a single pass transistor. When this configuration was used, steps were taken to prevent the output node from coupling back onto the ZZ_XD nodes (through the output transistors) while the ZZ_XD nodes are in the dynamic state. The inputs to the final logic gate are required to come from the same latch type (A-phase or B-phase), and the output nodes of the pass transistors must only drive the final 2-input logic gate using minimal routing. These requirements, together with the latch setup time, ensure that the output gate will switch while the ZZ_XD node is statically driven.

While these level-sensitive latches are fast, they are susceptible to data race-through. This problem was managed by a

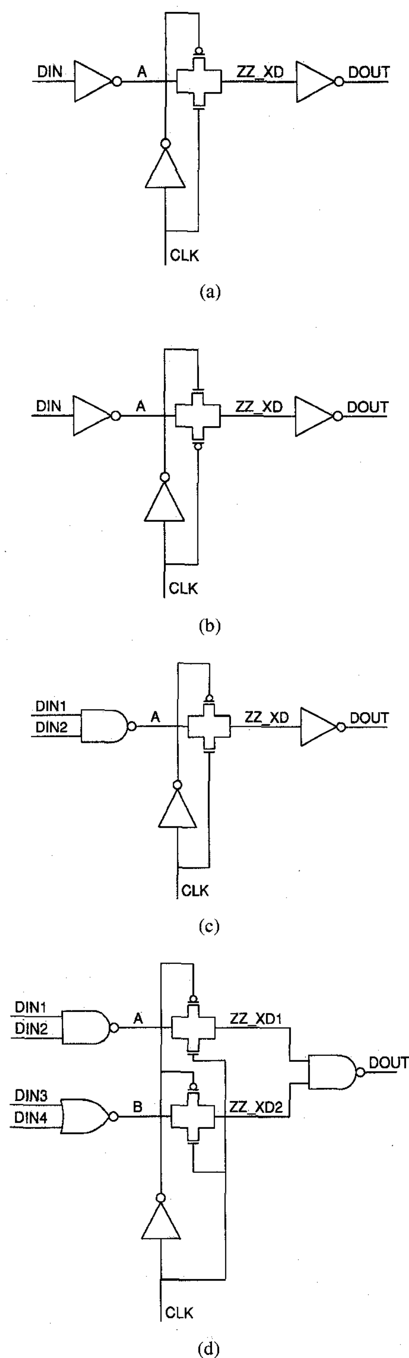


Fig. 7. (a) A-phase latch. (b) B-phase latch. (c) A-Phase latch with NAND function. (d) A-Phase latch with embedded logic.

combination of techniques, including controlling the skew on CLK, precisely sizing the local clock buffer inside the latch, placing strict rules on the use of the locally buffered clock, and requiring a minimum number of gate delays between latches. A custom verification tool was developed to check that the latching rules were followed throughout the design. This CAD tool performed a static analysis that identified places where buffered clocks were used outside of the standard latch structure and where there was an insufficient number of delays between latches. A suite of postlayout checks were performed to ensure that the internal dynamic latch nodes were not adversely affected by noise sources.

IV. Implementation Examples

The operating frequency of this chip lead to some very challenging implementation problems that were encountered during the design of the chip. This section of the paper describes four example circuits.

A. I-Box Issue Stage Domino Logic Circuit

The issue stage of the I-Box coordinates the release of instructions into the E-Box, F-Box, and M-Box pipelines. Issuing four instructions per cycle in a machine with deep pipelines and a complex memory system presented several design challenges. The four result and eight source registers of the issuing instructions must be compared against the 37 possible outstanding instructions (seven integer instructions, nine floating point instructions, and 21 load instructions that missed) within the machine. Concurrent with the register checks, 44 possible data bypass calculations must also be performed to ensure the most up-to-date data is forwarded to the issuing instruction.

Domino logic was used to implement the register scoreboard and bypass structures in order to meet the chip performance and area requirements. During instruction issue, each source and result register address is decoded and loaded into a 31-b wide pipeline that mimics the execution pipeline. Checks are performed for stalls and bypasses by selecting the appropriate bits from each stage of the pipeline and comparing them to the decoded register addresses of the new instructions. Integer and floating point instructions are handled in separate pipelines.

Decoding the register addresses into a 31-b wide vector, prior to entering the pipeline, allows the result register addresses to be logically ORed together to create a dirty register vector. This allows the stall calculations to be performed using only 38 comparators and the bypass calculations using only 44 comparators.

The comparators were implemented in three dynamic domino stages, as shown in Fig. 8. The first stage is a two-input multiplexer that selects the source/result decode field for the new instruction or the source/result decode address of the previous cycle based on whether or not a stall had been detected. The dirty bit vector is created in a similar logical OR structure. The second stage detects if there is a register conflict. The register conflict wire is discharged when there is a matching source/result decode and dirty vector bit. The transmission gate qualifies the detected register conflict with an instruction valid signal. The third stage is used to further qualify the detected conflict with instruction type decode information and to start combining the 38 conflict outputs down to a single stall wire. In the case of bypasses, the third stage is used to ensure that only the most up-to-date data is bypassed by priority-encoding the individual bypass indications.

The dynamic circuit implementation of the issue logic required careful analysis of several key circuit issues, such

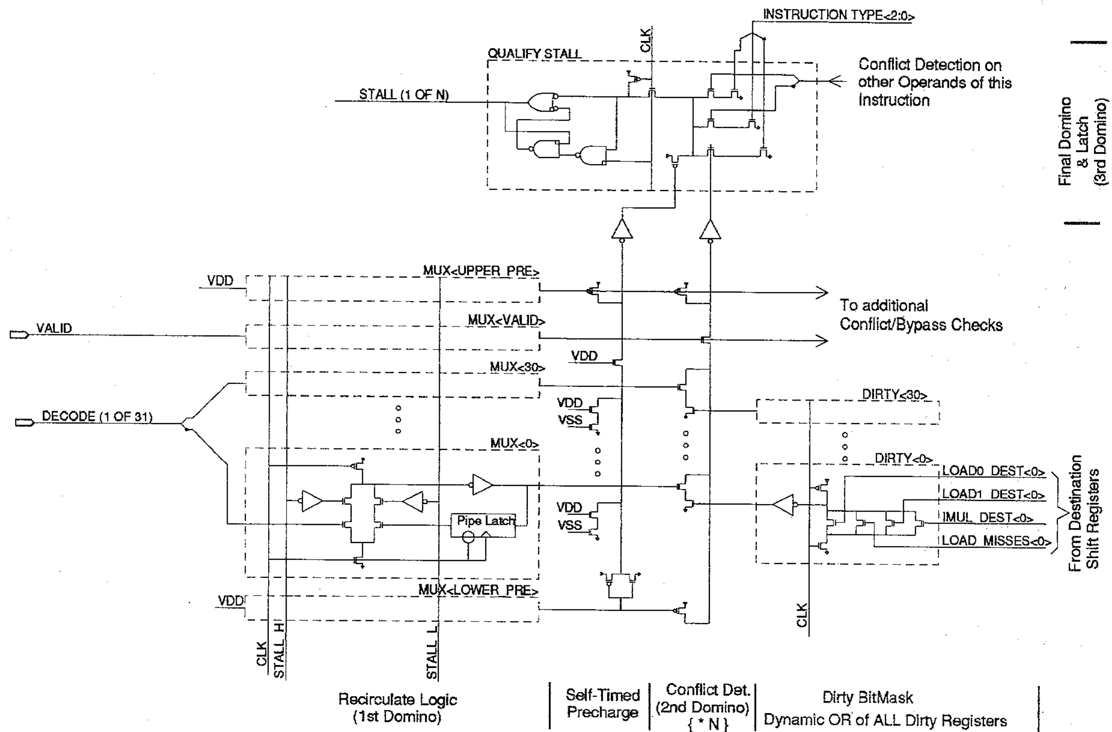


Fig. 8. I-Box domino logic circuit diagram.

as power dissipation, noise margins, and coupling. Power dissipation resulting from overlap between the precharge and the evaluate phases of the domino stages was minimized by self-timing the precharge enable signals for the second and third domino stages. Noise margins were maintained by locally buffering the input signals of dynamic gates and ensuring a common ground. To reduce the lateral capacitive coupling between the many dynamic signals, the space between the signals was maximized and wires were arranged to take advantage of signals with mutually exclusive switching characteristics.

B. E-Box Bypass Circuit

The E-Box contains two 64-b integer pipelines. Each pipeline produces a result from an operation requiring two input operands. A total of four distinct input operands must be provided to the two execution pipelines. The E-Box performance is significantly increased by implementing a bypass scheme that allows the result from any function unit in either pipeline to be bypassed to each of these four input operands. Through this scheme, every result is immediately available to the next instruction in each pipeline.

Fig. 9 shows the block diagram for one of the four E-Box operand bypasses. The operand bus can be sourced from the register file or from one of seven other bypasses. A bypass select signal from the I-Box controls that bypass input to use and a B-phase latch stores the chosen data. The function unit then performs its operation, latches the result with an A-phase latch, and drives it onto the result bus completing the loop.

The circuit implementation of one of the E-Box operand bypass buses is shown in Fig. 10. The A-phase latch is

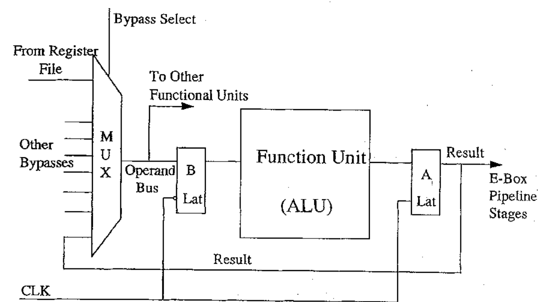


Fig. 9. E-box bypass block diagram.

immediately obvious but no explicit B-phase latch is used. The operand bus is implemented as a dynamic differential bus that spans the height of the E-Box datapath. It is driven by one of eight clocked operand bus drivers forming a distributed dynamic multiplexer. The bus is precharged in the A-phase and evaluated in the B-phase when one of the bypass enable signals is asserted.

The function units use a static receiver followed by a dynamic gate to receive the operand data. The dynamic gate evaluates in the A-phase and precharges during the B-phase. A race exists between the precharge of the operand bus and the evaluation of the operand bus in the function unit, since they both occur in the same phase. The race could have been avoided by placing a B-phase latch between the bypass bus and function unit receivers. However, the latch was eliminated to save time in the bypass critical speed path. The race was managed by including a bus precharge delay that provides additional hold time for the data so that all receivers have

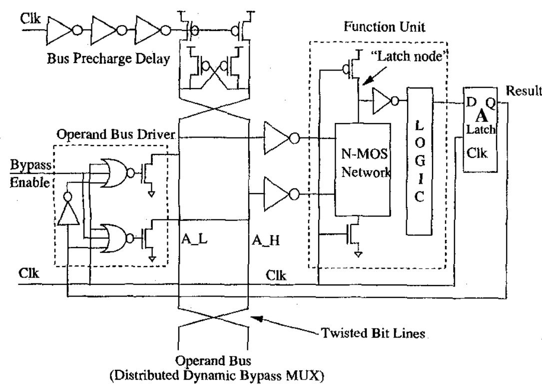


Fig. 10. E-box bypass circuit diagram.

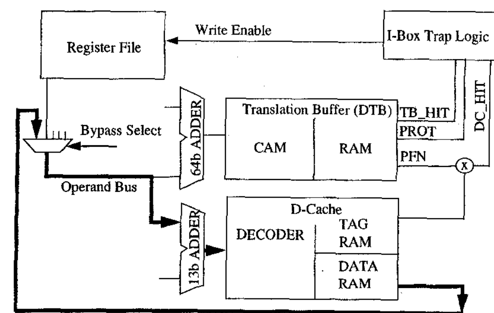


Fig. 11. D-cache access critical path.

enough time to capture the operand data before the bus is precharged.

A further concern of the bypass scheme is the possibility that crosstalk could corrupt data on a bus due to interconnect coupling capacitance. This is a consequence of the E-Box datapath being densely populated with eight dynamic differential buses per bit. Three design techniques were used to manage this problem. First, cross-coupled P-devices were connected across each differential bus pair to retain the precharged "1" value of the dynamic bus during the evaluate phase. Second, the beta-ratios of the bypass receiver gates were skewed to maximize the noise margins. Lastly, the operand buses were arranged in a manner similar to memory arrays with twisted bit lines, thus reducing worst-case capacitive crosstalk by approximately one half.

C. D-Cache Access Critical Paths

This microprocessor has a two-cycle latency for load instructions that hit in the D-Cache. A dependent instruction can issue two cycles after the load instruction has issued. This is a one cycle improvement over the previous implementation. This gives rise to two major critical speed paths. Fig. 11 shows a logical representation of these two paths. The D-Cache is accessed with a 13-b index and returns the data and its associated address tag. Simultaneously, the DTB translates the full address, which is then compared with the tag. Signals that indicate whether the load hit in the D-Cache are sent to the I-Box, which drives write enable signals to the E-Box register file. Data returned to the E-Box may be speculatively used by a consuming instruction, but it is only written into the register file if the load hits in the D-Cache.

One critical path delivers the lower 13 b of the address that is used as the index to the D-Cache. The index is calculated by a dedicated 13-b adder that drives the large capacitive load of the D-Cache decoders. The 13-b adder is placed outside of the E-Box datapath, between the register file and the D-Cache. The index output from the adder is routed to the D-Cache in metal-3 and metal-4.

The other critical path determines whether a reference hits in the D-Cache. In the first of the two cycles needed for this determination, a 64-b adder in the E-Box calculates the full address that is then compared against all entries in the CAM array of the DTB. This array is carefully positioned below the E-Box adders to avoid unnecessary RC delays. The adder, which is also used for basic integer arithmetic instructions, normally drives its output to the E-Box's result bus. However, in this case, the output is driven directly from the adder into the CAM array with a large dynamic driver. To save power, this dynamic driver is conditioned to evaluate only during address calculations.

In the first phase of the next cycle, an address tag is read from the D-Cache and delivered to a comparator prior to the arrival of the translated address. This address is read from the DTB and driven directly into the comparator by the sense-amps at the output of the DTB. The remainder of the cycle is used to perform the address comparison, to check the success of the address translation, and to return data back to the E-Box.

D. Cache Design

Special design considerations were given to the three on-chip caches. Redundancy was included in all the arrays to improve yield and the caches were designed to minimize power consumption. In addition, the general cache design was shared among the three on-chip caches to reduce the amount of design and verification effort.

The I-Cache and D-Cache both include two sets of fuse-programmable redundant rows to improve yield. These fuses are constructed of a titanium nitride layer that can be electrically blown using a laser. The I-Cache features built-in self-test logic to identify bad rows and built-in self-repair logic to automatically map the redundant rows over failing rows during wafer probe [2], [3]. This approach allows a more extensive test of the chip at wafer probe. After wafer probe the redundant rows are permanently mapped over the failing rows by blowing the fuses with the laser.

The large 96-kbyte, 3-way set associative, second-level cache is 128 b wide and is partitioned into two 64-b wide data arrays each containing twelve 4-kbyte banks (4 per set). The two arrays are placed on the left and right sides of the chip. Read and write buses pass over the two arrays in metal-4, connecting the data banks together and allowing access to the primary caches at the top and bottom of the chip. Fig. 12 is a block diagram showing the arrangement of the twelve 4-kbyte banks on the right side of the chip. There are also three separate tag arrays (1 per set) that are placed in the bottom left corner of the chip (see Fig. 1). Each bank of the tag and

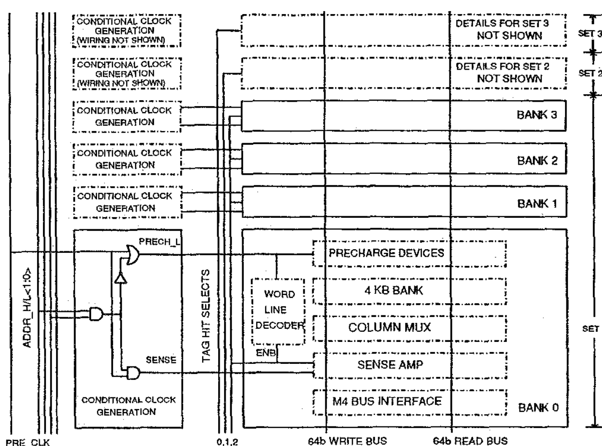


Fig. 12. Block diagram of right-half of the second-level cache.

data arrays implements row redundancy. The data arrays also implement column redundancy.

The second-level cache operates in a three-stage pipeline, two stages for tag lookup and modification, and one for data access. Pipelining the tag access ahead of the data access limits the number of data banks that are concurrently accessed. Partial address decoding during the tag lookup enables two banks per set (six total) by using the conditional clocks discussed earlier. Banks that have not been selected are held in the precharge state. The "hit" signals from the three tag arrays gate the word lines and sense amplifiers of the banks. Therefore, of the six banks enabled, only the two banks for the set that hit are activated and discharged. This results in an estimated power savings of 10 W.

The 8-kbyte D-cache supports two loads per cycle, requiring a dual-read-ported design. The D-cache was implemented as two single-ported caches containing identical data instead of one dual-ported cache. The major consideration that led to this decision was the ability to share the single-ported design with the I-Cache at the cost of a small increase in area. Sharing the design also reduced the overall analysis and verification required. In addition, the aspect ratio of this D-Cache configuration was ideal for the floorplan.

V. CAD Tools and Verification

A. Design Methods and Tools

An extensive suite of proprietary in-house CAD tools contributed significantly to the successful design of this chip. These tools were particularly effective in supporting design entry and rigorously checking the large amount of full-custom circuitry that was employed.

Tools that aided schematic generation included a schematic editor, a logic synthesis tool, and a device sizing tool. Postschematic tools included a latching methodology checker, a circuit verifier that highlighted electrical design methodology violations (e.g. dynamic node noise susceptibility), and a timing verifier that identified and analyzed potential critical paths. The use of the design tools varied across the chip based on the degree of customized logic required. For example, synthesis tools were not heavily used in the F-Box because of the need for optimized circuit structures. However, these tools were used extensively in the C-Box to produce initial schematics, which were then modified by hand as necessary.

Timing analysis was done statically on a per-section basis and the results were maintained in a database that tracked intersection signal timing information. Significant productivity improvement resulted from the timing analyzer's ability to automatically invoke SPICE[1] to verify suspected critical paths. This capability produced an order of magnitude more SPICE simulations than would have been manually possible, resulting in more accurate timing analysis and increased design confidence.

As previously discussed, minimizing clock skew across the chip was critical in meeting the operating frequency target. This could not have been achieved without robust simulation tools for accurately analyzing and reporting the skew during the clock design. Extracted RC time constants were analyzed by an asymptotical waveform evaluation tool. Clock skew results were displayed as a function of X, Y, and time using three-dimensional visualization tools that generated an animated display of the clock as it propagated from the drivers to the receivers across the chip. Fig. 6 shows a representation of the results of this analysis.

Detailed physical verification and reliability checks included analysis of electromigration, hot carriers, latch-up, power supply noise, coupling, and the detection of structures susceptible to charge damage from plasma etch processing during manufacturing. The CAD programs used to automate these tasks were frequently updated to provide additional features and enhanced circuit analysis. It is noteworthy that no electrical sensitivities were found in the chip during prototype debug.

B. Functional Verification

The complexity of the chip necessitated extensive functional verification prior to mask generation. A two-state RTL behavioral model was the primary verification vehicle. This model provided a balance of accurate design representation and simulation speed. The RTL model of the CPU was augmented with an abstract behavioral model of the remaining system components, including an off-chip cache, main memory, and I/O devices. This allowed normal system traffic to be applied to the simulation model. Once the circuit design was complete, a two-state gate-level model was extracted from the transistor netlist. This was used to verify that the schematics matched the RTL model. In addition, a three-state switch-level model was used to check for proper reset initialization.

The simulation models were verified with a variety of stimuli. The most effective techniques used targeted-random exercisers. These exercisers consisted of developing an outline of a test and allowing the specifics to be generated randomly every time the test was run. The general outline could be

[1] SPICE is a general-purpose circuit simulator program developed by Lawrence Nagel and Ellis Cohen of the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley.

executed repeatedly, generating a different test stimuli each time. These outlines could be combined with each other to create complicated random test generators called exercisers. In addition to the random exercisers, hand-crafted focused tests were also created to verify specific areas of the design.

Coverage analysis techniques were used to guide the verification process. This process consisted of analyzing each piece of logic and determining which sequences needed to be stimulated. For instance, in a state machine, it is important to exercise all transitions between states. For sections of the design where specific coverage goals were not achieved, either additional focused tests were created or random exercisers were tuned to fill in the coverage holes.

By the time the design was released to manufacturing, over 14 billion cycles of random stimuli were run on the RTL model, and over 500 million cycles were run on the gate-level models. In addition, over 400 focused tests were developed and executed on the RTL and gate-level models.

Many of the focused tests were used to generate manufacturing test patterns. Fault simulation was performed using these test patterns. The results indicated that tests that do an excellent job covering design faults achieve about 85% coverage of gate-level stuck-at faults. The fault simulation data is being used to direct test enhancements, leading to a steady increase in fault coverage.

VI. Conclusion

Employing a custom circuit design style, coupled with an optimized high-performance 0.5 μm technology, this second-generation, high-end microprocessor was designed to operate at a target clock frequency of 300 MHz. This was accomplished by employing a dense, low resistance power grid, maintaining minimal clock skew, using fast latches, and employing high-speed circuit techniques. The chip encompasses an area of 299 mm^2 and contains 9.3 million transistors, including a 96-kbyte second-level cache.

Within four weeks of silicon fabrication, the OpenVMS[2] operating system was booted, quickly followed by Digital UNIX®[3] and Windows NT[4] operating systems. Currently, all three operating systems are running successfully on a number of different system platforms operating at 300 MHz. This microprocessor has been incorporated into a number of available system products since April 1995.

Fig. 13 shows a shmoo plot of operating frequency versus supply voltage for this microprocessor functioning at a case temperature of 85°C. The plot shows the pass/fail boundary for various speeds and voltages for a chip with typical target process parameters. The plot demonstrates that the chip functions at frequencies greater than 300 MHz under normal operating conditions. Additionally, the results demonstrate that full speed operation is attainable even at high temperatures and voltages less than the normal 3.3 V supply.

[2] OpenVMS is a trademark of Digital Equipment Corporation.

[3] UNIX is a registered trademark in the United States and other countries licensed exclusively through the X/Open Company Limited. Digital UNIX for Alpha is a UNIX 93 branded product.

[4] Windows NT is a trademark of Microsoft Corporation.

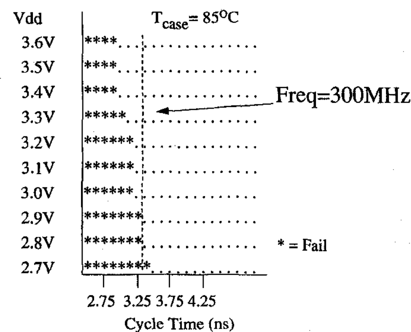


Fig. 13. Shmoo plot.

At 300 MHz, the four-way superscalar microprocessor is capable of achieving a peak execution rate of 1.2 BIPS and 600 MFLOPS. It has also attained measured performance of 341.4 SPECint92 and 512.9 SPECfp92 [4].

Acknowledgment

The authors would like to acknowledge the contributions of many people who helped make this chip possible. These include W. Herrick and P. Rubinfeld for project management support; F. Fox for technical consultation; A. Cave and R. Cvijetic for invaluable CAD assistance; and L. Bair, N. Arora, L. Gruber, and B. Zetterlund for device and technology modeling. Designers include R. Allmon, R. Badeau, P. Bannon, S. Bell, T. Benninghoff, R. Blake-Campos, D. Brasili, K. Broch, T. Broch, R. Castelino, M. Charnoky, E. Cooper, J. Edmondson, H. Fair, T. Fischer, A. Jain, J. Keller, J. Kowaleski, P. Kroesen, T. Mast, S. Mehta, A. Murphy, J. Mylius, T. Pham, V. Rajagopalan, T. Shedd, C. Somanathan, S. Strickland, S. Thierauf, and J. White.

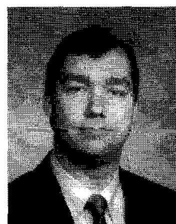
References

[1] D. Dobberpuhl *et al.*, "A 200-MHz 64-b dual-issue CMOS microprocessor," *IEEE J. Solid State Circuits*, vol. 27, pp. 1555–1564, Nov. 1992.

[2] D. K. Bhavsar and J. H. Edmondson, "Testability strategy of the Alpha AXP 21164 microprocessor," in *Int. Test Conf.*, Oct. 1994, pp. 50–59.

[3] D. K. Bhavsar and R. M. Fromm, "Testability features and testability access of the Alpha 21164 microprocessor," in *Custom Integrated Circuits Conf.*, May 1995.

[4] SPEC Newsletter: vol. 7, no. 2, June 1995.



Bradley J. Benschneider received the B.S.E.E degree from the University of Cincinnati, Cincinnati, OH.

Since joining Digital Semiconductor, Hudson, MA, in 1987, he has contributed to several custom chip designs in the VAX 6000 family and the early Alpha implementations. He is a Principal Hardware Engineer in the Semiconductor Engineering Group. He was responsible for designing various sections of the memory management unit on this chip, as well as defining the latching methodology for the chip. He is currently leading the implementation effort of the memory management unit for a next generation Alpha CPU. He has one patent and coauthored four papers.

Andrew J. Black (S'90–M'92) received the B.S.E.E. degree from Pennsylvania State University, University Park and the M.S.E.E. degree from the University of Southern California, Los Angeles.

After working for International Solar Electric Technology, he joined Digital Semiconductor, Hudson, MA, in 1992. He is a Senior Hardware Engineer in Digital's Palo Alto Design Center where he is designing the bus interface unit for a microprocessor chip. During his work on the CPU, he was a member of the design team for the memory management unit and for the chip's clock design.

Mr. Black is a member of Tau Beta Pi and Eta Kappa Nu.

William J. Bowhill (M'93) received the B.Eng. degree in electronic engineering from the University of Liverpool, Liverpool, UK, in 1981.

Before joining Digital Semiconductor, Hudson, MA, in 1985, he worked for Standard Telecommunications Laboratories UK, where he designed VLSI chips for telecommunication applications. He is a Consultant Engineer in Digital Semiconductor's High Performance CPU Group. He led the implementation of the microprocessor that is described in this paper. He was also the design organization's representative for the development of the 0.5 μm CMOS process used to fabricate the chip. His previous responsibilities have included technical contributions to both the VAX 6000 Model 400 and Model 600 chip sets.

Sharon M. Britton received the B.S.E.E. degree from Boston University, Boston, MA and the M.S.E.E. degree from the Massachusetts Institute of Technology, in 1983 and 1990, respectively.

She joined Digital in 1983 to work on the design and development of optical disk drive controllers. Since joining Digital Semiconductor, Hudson, MA, in 1990, as a Principal Hardware Engineer, she has contributed to the design of the floating-point unit on the Alpha 21064 chip and led the implementation of the M-Box load/store unit for this CPU. She is currently a member of the design team working on the instruction issue unit for the next-generation Alpha chip.

Daniel E. Dever received the B.S. degree in electrical engineering from the University of Cincinnati, Cincinnati, OH, in 1988.

Since joining Digital in 1988, he has worked on the design and logic verification of CMOS VAX and Alpha microprocessors. He is currently involved in the design of the memory management unit for the next-generation Alpha microprocessor.

Dale R. Donchin (S'75–M'78) received the B.S.E.E. and M.S.E.E. degrees from Rutgers University College of Engineering, New Brunswick, NJ, in 1976 and 1978, respectively.

He was previously a Development Manager in the RSX Operating System Group. In 1986, he joined Digital Semiconductor, Hudson, MA, where he is an Engineering Manager and Technical Contributor. He designed several circuits related to the clock and cache and contributed to and coordinated CAD tool use for this CPU. He is presently performing these duties for the development of the next-generation Alpha microprocessor.

Mr. Donchin is a member of ACM.

Robert J. Dupcak received the B.S.E.E and the M.Eng. degrees from Cornell University, Ithaca, NY, in 1992 and 1993, respectively.

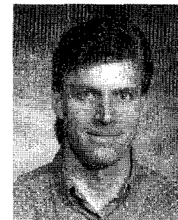
After joining Digital Equipment Corporation in 1993, he worked on the floating point unit of this microprocessor. He is currently involved with the design of the floating point unit of the third generation Alpha CPU.

Richard M. Fromm received the B.S.E.E. degree from Cornell University, Ithaca, NY.

After cooping in the Digital Video Group at Bell Communications Research, he joined Digital Equipment Corporation in 1991. Currently, he is a Senior Hardware Engineer in the High Performance CPU Group of Digital Semiconductor, Hudson, MA. On this microprocessor, he contributed to the design, implementation, and verification of the testability section, and he has assisted in the testing and debugging of chip prototypes. He is currently working on the memory management unit for the next generation Alpha microprocessor.

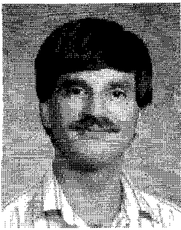
Mary K. Gowan received the B.S. degree in electrical engineering from the University of Illinois at Urbana.

Since joining Digital Semiconductor, Hudson, MA, in 1987, she has been involved in the design of three CPU chips, including the memory management unit on this microprocessor. She is a Senior Hardware Engineer in Digital Equipment Corporation's Semiconductor Engineering Group.



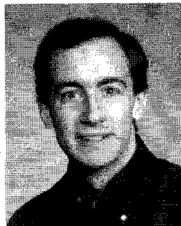
Paul E. Gronowski received the B.S. degree in electrical engineering from the University of Cincinnati, Cincinnati, OH.

Since joining Digital Semiconductor, Hudson, MA, in 1984, he has contributed to the design of several high-performance microprocessors. For this CPU, he was responsible for the integer execution unit and led the physical chip verification effort. He is currently responsible for the technical design and management of the next-generation processor. He is the coauthor of several ISSCC papers and holds one patent.



Michael Kantrowitz (S'80–M'84) received the B.S.E.E. degree from Stevens Institute of Technology, Hoboken, NJ and the M.S.E.E. degree from Worcester Polytechnic Institute, Worcester, MA.

Before joining Digital Semiconductor, Hudson, MA, in 1988, he worked at Raytheon Company. A Consulting Engineer, he is currently leading the verification effort for a new Alpha microprocessor and developing new verification tools and methods. Prior to this project, he was coleader of the verification of this chip, responsible for the instruction fetch and execute units. He has also contributed to the verification of the Mariah, NVAX+, 21064 floating-point unit, and FAVOR vector unit.



Marc E. Lamere (A'92) received the B.S.E.E. degree from Rensselaer Polytechnic Institute, Troy, NY and the M.S.E.E. degree from Northeastern University, Boston, MA, in 1983 and 1988, respectively.

In 1984, he joined Digital Semiconductor, Hudson, MA as an ECL Circuit Designer on the VAX 9000 project and helped design custom and semicustom bipolar chips. A Principal Hardware Engineer in Digital Semiconductor, he is currently a CMOS Circuit Designer for the next-generation Alpha microprocessor. In his work on this CPU, he was responsible for the integer execution unit shifter and other circuit designs as well as the physical and electrical verification of the chip.

Shekhar Mehta received the M.S.E.E. degree from the University of Wisconsin, Madison, in 1988.

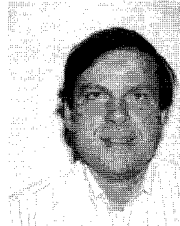
Before joining Digital Semiconductor, Hudson, MA, in 1988, he was an Engineer at Larsen & Toubro, Bombay, India. He is a Senior Hardware Engineer in Digital Semiconductor's High Performance Computing Group. He designed the miss address file on the memory subsystem of this CPU and was responsible for the electromigration checks of the chip. He is currently leading the design of the caches on a future Alpha microprocessor.

Jeanne E. Meyer received the B.S.E.E. degree from the University of Cincinnati, Cincinnati, OH, in 1982.

Since joining Digital Semiconductor, Hudson, MA, in 1989, she has worked on the implementation, behavioral modeling, and logic verification of several microprocessor chips. In her work on this CPU, she was responsible for PALcode verification, maintenance, and support. She also contributed to the microarchitecture definition and behavioral model of the chip's memory management unit. She is currently leading the design of the memory management unit for a new Alpha microprocessor. She holds two patents.

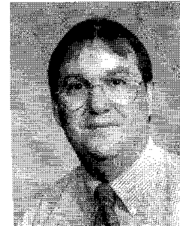
Robert O. Mueller received the B.S. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY.

As a Senior Hardware Engineer at Digital Semiconductor, Hudson, MA, he is currently involved in the design and implementation of the pad ring for a new Alpha microprocessor. In his work on this CPU, he has contributed to the design, implementation, and electrical verification of the pad ring, the cache control, and the bus interface unit.



Andy Olesin received the B.S.E.E. degree from Rutgers University, New Brunswick, NJ in 1970.

Prior to joining Digital Equipment Corporation in 1983, he was a Design Section Manager at Motorola Semiconductor. At Digital, he was the Circuit Design Leader for the company's first CMOS VAX microprocessor. He is currently a Consulting Engineer in the Semiconductor Engineering Group and was responsible for the design of the CPU Clock Logic and Floating Point Multiplier. He holds six patents related to CMOS circuit design.



Ronald P. Preston (S'84–M'88) received the B.S.E.E. and M.S.E.E. degrees from Rensselaer Polytechnic Institute, Troy, NY, in 1984 and 1988, respectively.

Since joining Digital Semiconductor, Hudson, MA, in 1988, where he is a Principal Engineer, he has worked on the design of several microprocessors and was the implementation leader for the instruction unit on this microprocessor. He was also responsible for the architecture and implementation of the issue/bypass/scoreboard logic. He is the coauthor of several articles on hot carrier analysis of CMOS circuits.

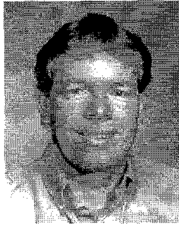
Mr. Preston is a member of Eta Kappa Nu.

Donald A. Priore received the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1984.

He joined Digital Semiconductor, Hudson, MA, in 1984, where he is a Consultant Engineer and has worked on high-performance Alpha microprocessor design since its inception. His interests include on-chip signal and power integrity and embedded memory design.

Sribalan Santhanam received the B.E. degree in electrical engineering from Anna University, Madras, India, and the M.S.E. degree in computer science and engineering from the University of Michigan, Ann Arbor, in 1987 and 1989, respectively.

In 1989, he joined Digital Semiconductor, Hudson, MA, where he worked on the design of the floating-point unit of the 21064 CPU and subsequently on the design of cache control unit of this CPU. He is currently a Senior Hardware Engineer in Digital's Palo Alto Design Center where he is responsible for the cache design for the StrongArm PDA microprocessor chip.



Michael J. Smith received the B.S.E.E. degree from the Rochester Institute of Technology, Rochester, NY.

A Principal Engineer at Digital Semiconductor, Hudson, MA, where he has been since 1986, he was a member of the instruction unit design team for this microprocessor, responsible for floorplanning, logic, and circuit design. Prior to this, he was involved in the design of two memory controller/bus adapter chips for the VAX 4000 Models 300 and 600. Currently, he is a member of the bus interface and instruction unit teams of the next-generation Alpha microprocessor.

Gilbert M. Wolrich received the B.S.E.E degree from Rensselaer Polytechnic Institute, Troy, NY and the M.S.E.E. degree from Northeastern University, Boston, MA.

A Consultant Engineer at Digital Semiconductor, Hudson, MA, he was the leader and architect for the floating-point unit on this chip.