High-Performance Microprocessor Design

Paul E. Gronowski, *Member, IEEE*, William J. Bowhill, *Member, IEEE*, Ronald P. Preston, *Member, IEEE*, Michael K. Gowan, and Randy L. Allmon, *Associate Member, IEEE*

***Abstract*— Three generations of Alpha microprocessors have been designed using a proven custom design methodology. The performance of these microprocessors was optimized by focusing on high-frequency design. The Alpha instruction set architecture facilitates high clock speed, and the chip organization for each generation was carefully chosen to meet critical paths. Digital has developed six generations of CMOS technology optimized for high-frequency design. Complex circuit styles were used extensively to meet aggressive cycle time goals. CAD tools were developed internally to support these designs. This paper discusses some of the technologies that have enabled Alpha microprocessors to achieve high performance.**

***Index Terms*— Alpha, CMOS digital integrated circuits, computer architecture, flip-flops, integrated circuit design, logic design, microprocessors.**

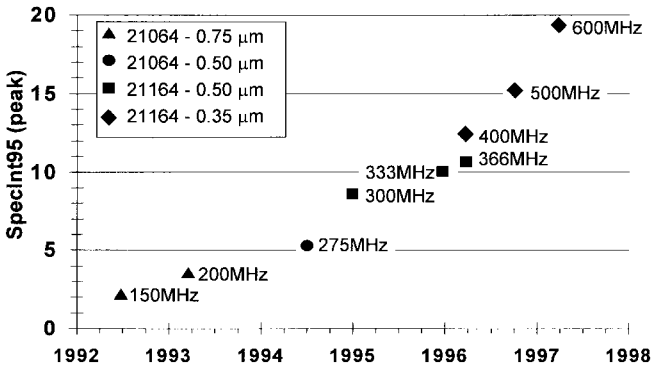

Fig. 1. Alpha performance versus time.

I. Introduction

DIGITAL introduced the Alpha 21064 in 1992, the highest performance microprocessor in the industry at that time [1]. Digital has delivered three generations of high-performance Alpha microprocessors through process advancements, architectural improvements, and aggressive circuit design techniques. Fig. 1 shows the integer performance of the 21064 and 21164 as a function of time. Over the last five years, the clock frequency of the Alpha microprocessor has increased from 150 to 600 MHz. The 21264, the third-generation Alpha, has been designed to operate at 600 MHz with improved performance over the 21164 [2]. Table I contains the key features for each of these microprocessors.

TABLE I
Microprocessor Features

	21064	21164	21264
Transistor Count (million)	1.68	9.3	15.2
Die Size (mm^2)	16.8x13.9	18.1x16.5	16.7x18.8
Process Technology	0.75 µm	0.50 µm	0.35 µm
Power Supply (Volts)	3.3	3.3	2.2
Power Dissipation (Watts)	30	50	72
Target Design Frequency (MHz)	200	300	600
Typical Gate Delays/Cycle	16	14	12
On-chip Cache	8-KB L1-I 8-KB L1-D	8-KB L1-I 8-KB L1-D 96-KB L2	64-KB L1-I 64-KB L1-D
Instruction Issue/Cycle	2	4	6
Execution Flow	in-order	in-order	out-of-order

The 21064 (Fig. 2) was the first implementation of the Alpha architecture. It was designed to operate at 200 MHz in a 0.75-μm n-well CMOS process, allowing for roughly 16 gate delays per cycle including latching. Power dissipation is 30 W from a 3.3-V power supply at 200 MHz. The die measures 2.3 cm^2, and contains 1.68 million transistors, half of which are dedicated to noncache logic.

The second-generation Alpha microprocessor, the 21164 (Fig. 3), is fabricated in a 0.5-μm n-well CMOS process [3]. It was designed to operate at 300 MHz using a 3.3-V supply, and it dissipates 50 W. The number of gate delays per cycle was reduced from 16 to 14 on this design to provide an additional 10% reduction in cycle time beyond process scaling. The die is roughly 3.0 cm^2 and contains 9.3 million total transistors. The noncache transistor count is tripled from the previous generation design to 2.5 million. Although originally designed to operate at 300 MHz, migration of this design to a 0.35-μm process has increased the operating frequency to 600 MHz.

The 21264 (Fig. 4) is the third-generation Alpha microprocessor. It is designed in a 0.35-μm n-well CMOS process, and is targeted to operate at 600 MHz. The number of gate delays per cycle has been further reduced to 12, again providing an additional 10% reduction in cycle time relative to the previous design. A nominal supply voltage of 2.2 V is used to limit power dissipation to an estimated 72 W, but the design and process can operate reliably up to 2.5 V. The die is 3.1 cm^2, and contains 15.2 million transistors. The noncache transistor count is more than double that of the 21164.

To achieve high performance without impacting time-to-market, a careful balance among microarchitectural features, process complexity, and circuit design style was required on each of these microprocessors. The use of high-performance circuit design techniques required the development of many custom CAD tools, and added to the complexity of the circuit verification task.

Manuscript received September 1997; revised December 17, 1997.

The authors are with Digital Semiconductor, Digital Equipment Corporation, Hudson, MA 01749 USA.

Publisher Item Identifier S 0018-9200(98)02228-8.

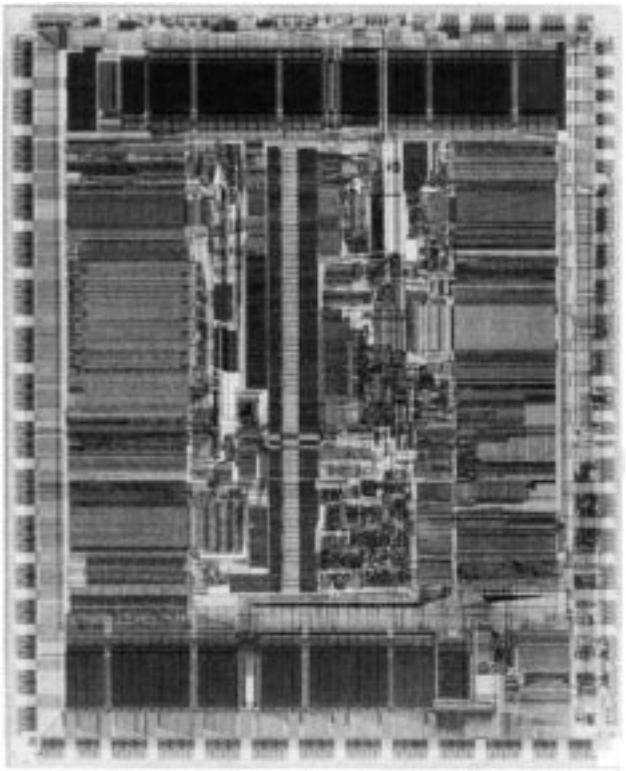

Fig. 2. 21064 die photo.

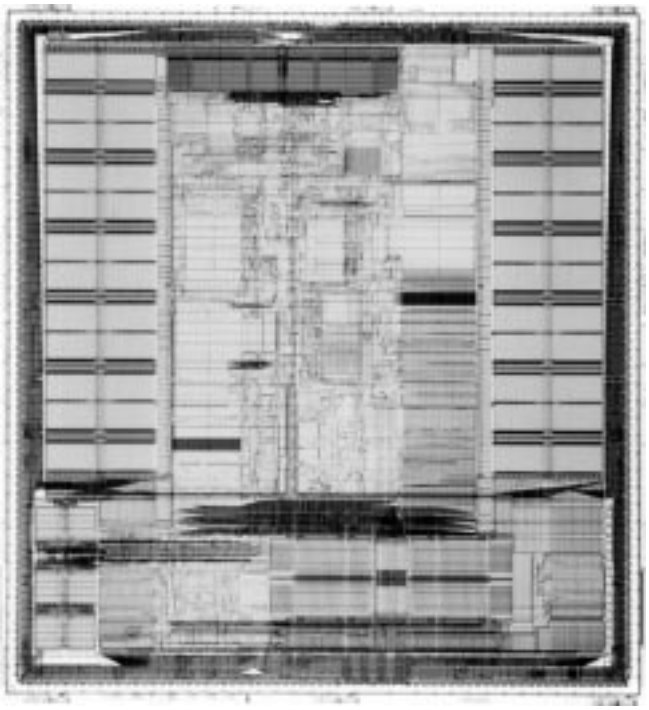

Fig. 3. 21164 die photo.

II. Architecture

The Alpha instruction set architecture is a true 64-bit load/store RISC architecture designed with emphasis on high clock speed and multiple instruction issue [4]. Fixed-length instructions, minimal instruction ordering constraints, and 64-bit data manipulation allow for straightforward instruction decode

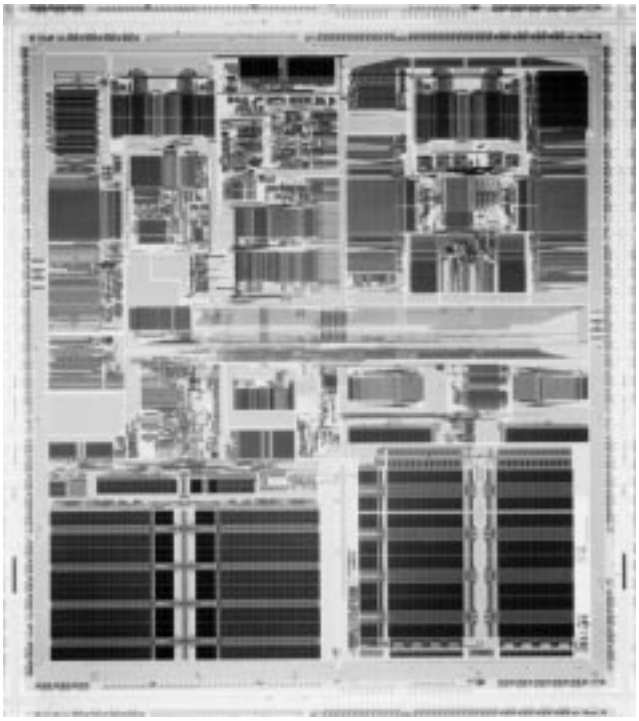

Fig. 4. 21264 die photo.

and a clean microarchitectural design. The architecture does not contain condition codes, branch delay slots, adaptations from existing 32-bit architectures, and other bits of architectural history that can add complexity. The chip organization for each generation was carefully chosen to gain the most advantage from microarchitectural features while maintaining the ability to meet critical circuit paths.

The 21064 is a fully pipelined in-order execution machine capable of issuing two instructions per clock cycle. It contains one pipelined integer execution unit and one pipelined floating-point execution unit. Integer instruction latency is one or two cycles, except for multiplies which are not pipelined. Floating-point instruction latency is six cycles for all instructions except for divides. The chip includes an 8-kB instruction (I) cache and an 8-kB data (D) cache. The emphasis of this design was to gain performance through clock rate while keeping the architecture relatively simple. Subsequent designs rely more heavily on aggressive architectural enhancements to further increase performance.

The quad-issue, in order execution implementation of the 21164 was more complex than the 21064, but simpler than an out-of-order execution implementation [5]. It contains two pipelined integer execution units and two pipelined floating-point execution units. The first-level cache was changed to nonblocking. A second-level 96-kB unified I and D cache was added on-chip to improve memory latency without adding excessive complexity. Integer latency was reduced to one cycle for all instructions, and was roughly halved for all MUL instructions. The floating-point unit contains separate add and multiply pipelines, each with a four-cycle latency [6]. Floating-point divide latency is reduced by 50%.

The trend of increased architectural complexity continues with Digital's latest Alpha microprocessor. The 21264 gains

TABLE II
TECHNOLOGY FEATURES

Technology	CMOS4	CMOS5	CMOS6
Flagship CPU	21064	21164	21264
Feature Size (μm)	0.75	0.50	0.35
Channel L_{eff} (μm)	0.50	0.35	0.25
Gate Oxide (nm)	10.5	9.0	6.0
V_{tn}/V_{tp}	0.5/-0.5	0.5/-0.5	0.35/-0.35
Power Supply (V)	3.3	3.3	2.0-2.5
M1 thick/pitch (μm)	.75/2.25	.85/1.5	.62/1.225
M2 thick/pitch (μm)	.75/2.675	.85/1.75	.62/1.225
Reference Plane 1	n/a	n/a	Vss
M3 thick/pitch (μm)	2.0/7.5	1.53/5.0	1.53/2.8
M4 thick/pitch (μm)	n/a	1.53/6.0	1.53/2.8
Reference Plane 2	n/a	n/a	Vdd

significant performance from six-way-issue and out-of-order execution. It contains four integer execution units and two floating-point execution units. The size of the $L1$ instruction and data caches was increased from 8 to 64 kB, eliminating the need for an on-chip $L2$ cache. Integer multiply latency was reduced and full pipelining improved throughout. The floating-point latency remained at four cycles, but the divide latency was reduced by another 50%. In addition, the ISA was extended to include square root and to support multimedia instructions.

Despite the added architectural complexity, clock frequencies have continued to improve due to circuit design enhancements and advances in process technology.

III. TECHNOLOGY

Digital Semiconductor has developed six generations of CMOS process technology, with a new technology for each major microprocessor design. The microprocessor design occurs in parallel with the development of the manufacturing process. Therefore, close cooperation is required between the process development and microprocessor design teams to perform this concurrent design and ensure optimum chip performance. Table II highlights the key features of the three process technologies used to produce these three microprocessors. The processes were optimized for high-frequency microprocessor design. In particular, emphasis is placed on low V_t's and very short L_{eff}'s which increase drive current at the cost of higher leakage.

Close interaction between the circuit design team and the process development team also results in the following benefits.

1) Early process information and timely updates of technology parameters are provided to the design teams, allowing circuit design to start before the process is fully defined.
2) Early design work provides valuable feedback to the process team to ensure that target process performance is met.
3) Major process features such as number of interconnect layers, interconnect pitch, and device characteristics are managed in the context of the overall chip design.
4) The design of critical structures such as RAM arrays and data paths can be optimized through process and circuit design.

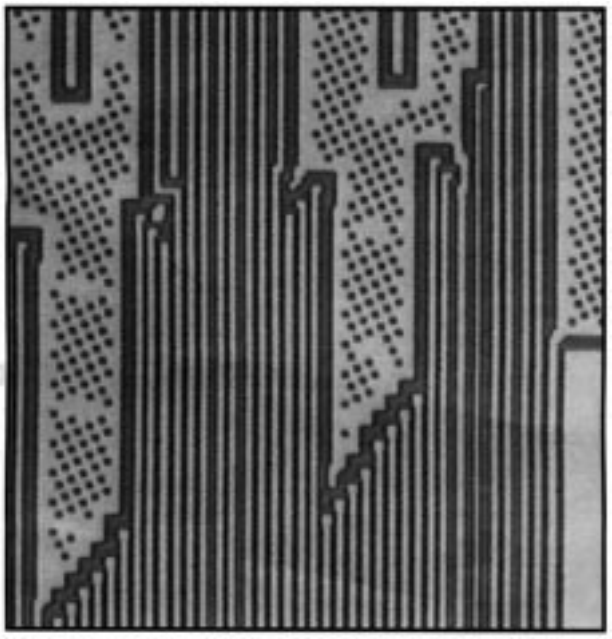

Fig. 5. 21264 metal fill.

5) Scaling issues for future process shrinks may be uncovered.

A. Definition of Design Rules

One of the key areas where close collaboration is required between design and process development teams is the definition of layout design rules. Aggressive design rules can result in increased circuit density, and can potentially improve overall chip performance. However, design rules that are too aggressive will complicate manufacturing, and may impact yield. On the other hand, slack design rules may result in increased die size, resulting in increased distances between critical structures. This increased distance results in higher capacitance, larger RC routing delays, and lower chip performance.

Often, the process team can be more aggressive if limits are placed not only on the minimum widths and spaces of structures, but also on the maximum widths and spaces. The 21264 implements metal fillers to limit the maximum spacing between adjacent lines. The fill metal is automatically placed in the design and tied to V_{DD} or V_{SS}. For large areas of fill metal, stress relief holes are automatically placed in the fill pattern. Metal fill may increase the capacitance of nearby signal lines, but they also result in improved interlayer dielectric uniformity. The improved uniformity allows the process to be targeted more aggressively. Fig. 5 shows filler polygons inserted in the gaps between widely spaced lines.

IV. CIRCUIT DESIGN

Advanced process technologies have allowed the Alpha microprocessor designers to increase performance on each new generation chip through two means, better electrical properties and higher densities. First, scaling the physical dimensions of the transistors and interconnect has reduced the nodal

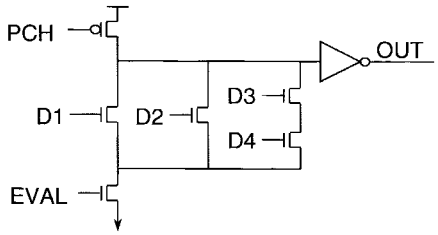

Fig. 6. Dynamic logic example.

capacitance, and has allowed the circuit speed to increase proportionally with the inverse of the technology physical scale factor. Second, the reduced size of the devices and interconnect have enabled designs to double the amount of circuitry that can be implemented in the same die size. The increase in circuit area is inversely proportional to the square of the technology scaling.

A primary objective of each new microprocessor development has been to further increase the performance of each generation of microprocessor by increasing the clock frequency by an amount greater than the above-mentioned process scaling factor. This has been achieved through the use of sophisticated circuit design techniques.

Full-custom circuit design methodologies have been used universally by the microprocessor design teams. All circuits are designed at the transistor level, and are uniquely sized to meet speed and area goals. Custom layout design techniques are used to optimize parasitic capacitance for all circuits. Automatic synthesis approaches for logic and circuit design have been used for fewer than 10% of the circuits.

The use of a full-custom design methodology gives the designer flexibility. The Alpha microprocessors have been implemented with a wide range of circuit styles including conventional complementary CMOS logic, single- and dual-rail dynamic logic, cascode logic pass transistor logic, and ratioed static logic [7].

Dynamic circuits are one of the most commonly used circuit styles, and are present in both data path and random control structures. Dynamic logic has many advantages, but it requires careful analysis to ensure functionality. Fig. 6 shows a simple dynamic domino gate. Dynamic gates allow wide OR structures to be implemented in a single gate which otherwise would require many levels of complementary logic. Dynamic gates are also faster than their complementary gate equivalents for several reasons. First, eliminating the PMOS transistor network reduces both the gate fan-in and fan-out capacitance. Second, the switching point of the dynamic gate is set by the NMOS device threshold voltage. If the timing of the inputs is such that they are not asserted until after the precharge clock has been deasserted, there in no crossover current during the output transition. Finally, removing the PMOS transistor network reduces the layout area, which results in lower interconnect capacitance, further increasing the speed of the circuit. However, dynamic circuits are very sensitive to noise, and require very careful design and extensive verification to ensure functionality. Much of this verification has been automated, and will be discussed in the CAD tool section of this paper.

Another circuit style that is widely used in the microprocessors is dual-rail cascode logic [8]. A simple gate is shown in Fig. 7. Like dynamic logic, cascode logic has been used in both data path and random control logic areas. It has many of the same advantages that dynamic logic possesses over complementary CMOS. The fan-in and fan-out capacitance are both lower, thus reducing delay. Large complex functions such as multiplexers and XOR gates can be easily implemented in a single cascode gate with both true and complement outputs. Finally, a latch function can easily be constructed with the addition of one pair of transistors (see Fig. 7).

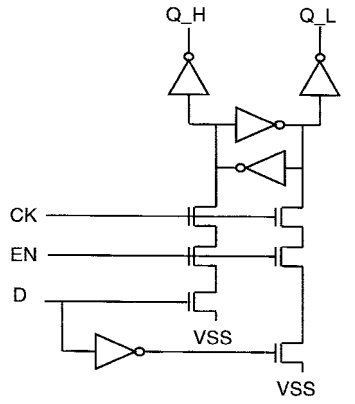

Fig. 7. Cascode logic example.

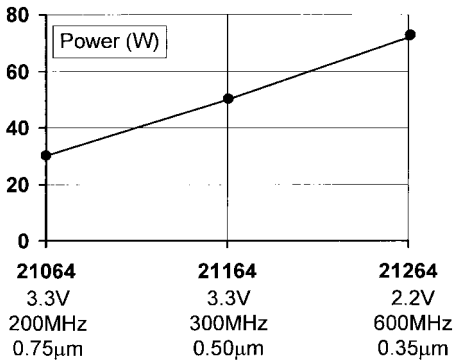

Fig. 8. Power dissipation.

A. Power Dissipation and Supply Distribution

Power consumption has been an important design constraint for the Alpha microprocessor designers. Fig. 8 shows the average power dissipation, process technology, and nominal supply voltage for the first three Alpha microprocessor designs. The graph clearly illustrates a steady increase in power despite the use of advanced technologies and scaled power supply voltages. This increase in power has resulted from a number of factors. First, more complex architectural features have been included into the design of each generation microprocessor. Second, the use of sophisticated circuit techniques has allowed clock frequencies to increase faster than pure process scaling would have provided. Third, aggressive transistor design (high $Id_{\rm sat}$) has increased the magnitude of subthreshold leakage significantly. Finally, improvements in compiler and software technologies have increased the switching activity within the microprocessor.

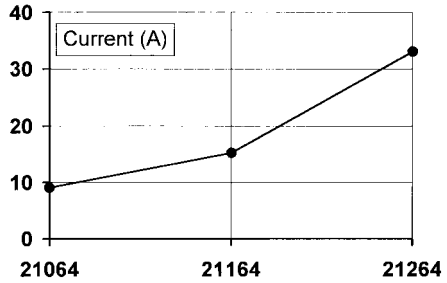

Fig. 9. Power supply current.

Power supply scaling has been an important lever to slowing the rise in power consumption. However, if current rather than power is considered, a much more disturbing trend is seen (see Fig. 9). Supply current is roughly doubling with each new generation. The problem of on-chip power distribution is increasing with each generation microprocessor. A design goal has been to limit the combination of dc IR drops and inductive ringing on the chip to 10% of the total supply voltage. As a side effect of scaling V_{dd} to achieve lower power levels, the amount of acceptable power supply noise is also reduced. Therefore, each generation of microprocessor has required additional process options to be added to each technology generation to lower the power supply impedance.

In CMOS logic, circuit speed is directly related to supply voltage. Therefore, in order to achieve high clock frequencies, the power supply networks must be designed to supply the required current with minimal IR drops. The 21064 consumed 30 W, and is fabricated in a 0.75-μm process. Distributing the necessary supply current across the 2.3-cm^2 die, with an acceptable IR drop, would not have been possible using the existing two-level metal process. Therefore a third, low-resistance aluminum interconnect layer $(M3)$ was added to the process. Adding layers to a fabrication process increases the cost of the die both through extra masking steps and reduction in yield. Therefore, relatively large minimum geometries were selected for the third metal layer to minimize yield impact. This new layer was used mainly for power, ground, and clock distribution.

V_{dd} and V_{ss} $M3$ lines were alternated to form two interleaved combs as shown in Fig. 10. To further reduce the effective resistance of the power grid, the $M3$ lines were routed with a pitch of 15 μm or about twice the process minimum. The second metal layer $(M2)$ was used to strap together the $M3$ lines, forming a grid for power, ground, and clock. One disadvantage of this power-routing scheme is that the V_{dd} and V_{ss} bond pads on the left and right sides of the die are connected to the grid using the higher resistance $M2$.

The 21164 consumed nearly twice the power of the 21064. The $M2/M3$ grid structure used on the 21064 was not adequate to meet the 21164 power requirements. Additionally, reliability concerns in connecting V_{dd} and V_{ss} to $M3$ would have required dedicating all of the bond pads on two sides of the chip to power and ground. The floor plan constraints of this bond pad allocation would have significantly complicated chip layout. Therefore, a fourth layer $(M4)$ of aluminum interconnect was added to the new 0.5-μm process. $M4$ was

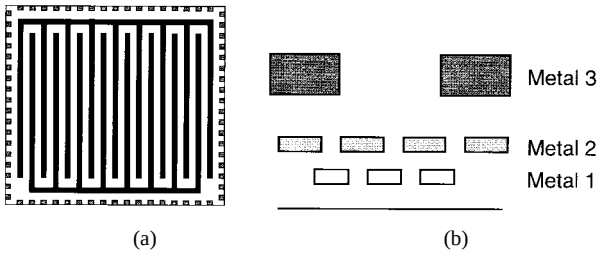

Fig. 10. (a) 21064 metal layers and (b) power distribution.

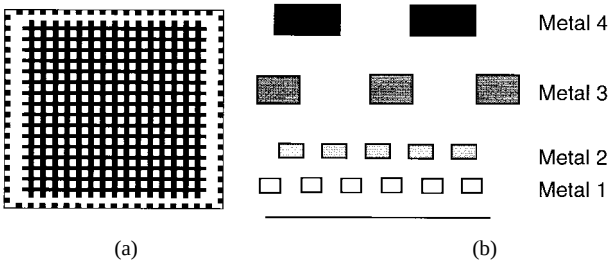

Fig. 11. (a) 21164 metal layers and (b) power distribution.

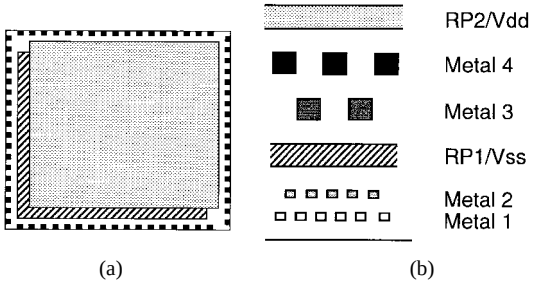

Fig. 12. (a) 21264 metal layers and (b) power distribution.

routed perpendicular to the $M3$ to form a two-dimensional (2-D) grid of V_{dd} and V_{ss} as shown in Fig. 11. A CAD tool was used to automatically contact the intersections between $M3$ and $M4$ with a maximum number of contacts. The 2-D grid allowed for bond pads on all four sides of the die to contribute to supplying the current demand of the 21164. The $M3$ and $M4$ routing layers were also used to rout a limited number of global signals and for clock distribution.

The power dissipation on the 21264 increased to 72 W despite a V_{dd} reduction from 3.3 to 2.2 V, increasing the supply current to 33 A. In addition, conditional clocking exaggerates the cycle-to-cycle current variation causing a maximum delta-I_{dd} of 25 A between adjacent cycles. As a result, the two-dimensional grid used on the 21164 was no longer sufficient.

In order to meet the very large cycle-to-cycle current variations, two thick low-resistance aluminum reference planes were added to the process. Reference plane 1, tied to V_{ss}, was added between $M2$ and $M3$. Reference plane 2 was added above $M4$, and is connected to V_{dd}. Fig. 12 provides a cross section of the metallization and power routing of the 21264. Contacting the reference planes to adjacent layers was automated.

The use of planes has several beneficial effects. Nearly the entire die area is available for power distribution. Second, the

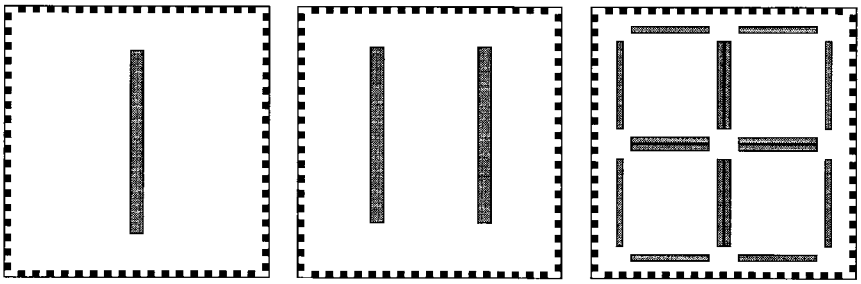

Fig. 13. Final clock driver location: (a) 21064, (b) 21164, and (c) 21264.

lower reference plane inductively and capacitively decouples $M2$ from $M3$ signal lines. This reduces on-chip crosstalk and simplifies CAD tool parasitic extraction. Finally, solid planes provide excellent current return paths which minimize inductive noise caused by signal-switching events [9].

The additional metal layers significantly improve the power distribution on the chips, and help to reduce the voltage loss in the center of the die due to dc IR drop. However, the high clock frequencies of these microprocessors result in large fluctuations in I_{dd} current. This current must be supplied through the package lead and bond wire inductance, and which results in power supply noise on chip. On-chip decoupling capacitance is implemented to help reduce this noise.

The power supply is decoupled using the gate oxide of NMOS transistors to form a capacitor. This type of structure was chosen as it provides the highest efficiency in capacitance per unit area without introducing additional process steps. The capacitor design provides a low-impedance path to the V_{ss} terminal, improving the bandwidth of the capacitor enough to efficiently decouple high-frequency noise on the power grids. A decoupling capacitor standard cell was designed and automatically placed in the chip. Whenever possible, the capacitor was placed in vacant areas of the chip where it did not impact die area, e.g., underneath global metal 1 buses. However, in areas close to the clock generators, it was required to dedicate die area to decoupling capacitance to supply the large clock switching currents. In total, 15–20% of the die area is used for decoupling.

The 21264's conditionally clocking scheme and the superscalar architecture of the microprocessor exaggerated the data and program dependencies that result in large variations in supply current. For the 21264, It was not possible to integrate enough decoupling capacitor on chip to manage this noise. Therefore, an additional source of decoupling was added to the chip and package network. A 1-μF 2-cm^2 wirebond attached chip capacitor (or WACC), implemented as a p-type accumulation mode MOS capacitor, was bonded on top of the microprocessor die [3]. This silicon capacitor helped control power supply noise.

B. Clock Distribution

The high frequencies of the Alpha microprocessors have required the generation and distribution of a very high-quality clock signal and the use of fast (low-latency) latches. The primary objective of the clock system is to not limit the performance of the microprocessor. Uncertainties in clock edges resulting from power supply noise, process variation, and interconnect RC delay lower the maximum clock frequency of the microprocessor. In addition, slow clock edges introduce uncertainties in latch timing which further limit performance and can lead to functional failures due to latch race-through.

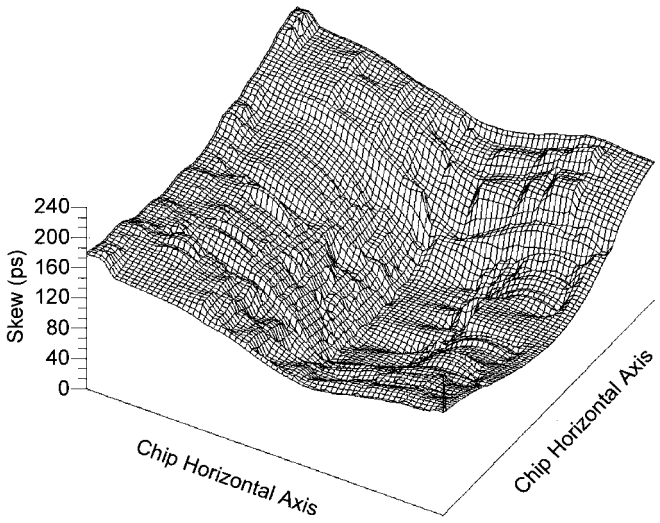

Fig. 14. 21064 clock skew.

The 21064 uses a two-phase single wire clocking scheme. The driver is located in the center of the die as shown in Fig. 13(a). The final clock load was 3.5 nF, and it required a final driver with a gate length of 35 cm. To handle the large di/dt transient currents in the power grid when the clock driver switched, on-chip decoupling structures (NMOS transistor with the gate tied to V_{dd} and the source and drain tied to ground) were placed around the clock driver. Roughly 10% of the chip area was allocated to decoupling capacitance. Fig. 14 shows the results of the 21064 clock skew analysis.

Fig. 15 shows the approximate power breakdown of the first two microprocessors. Since the main clock drivers consumed 40% of the chip power, thermal management was a major concern. Fig. 16 shows that the temperature of the main clock driver is elevated about 30°C relative to the rest of the die. The elevated temperature in the clock driver area reduces the performance of the clock drivers and other local logic, directly impacting performance.

The primary goals of the 21164 clock design were to reduce the clock skew by 30% and to reduce the thermal gradients.

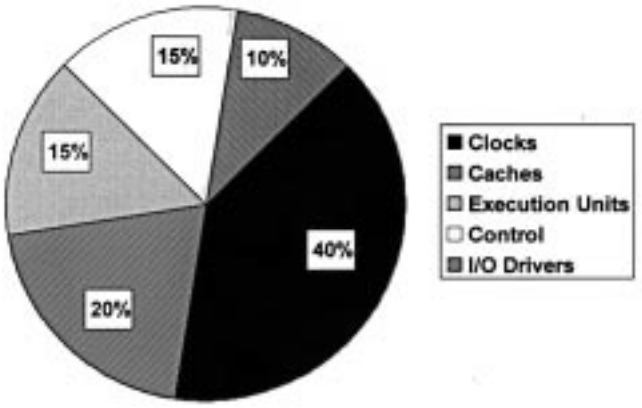

Fig. 15. Approximate power breakdown of the 21064 and the 21164.

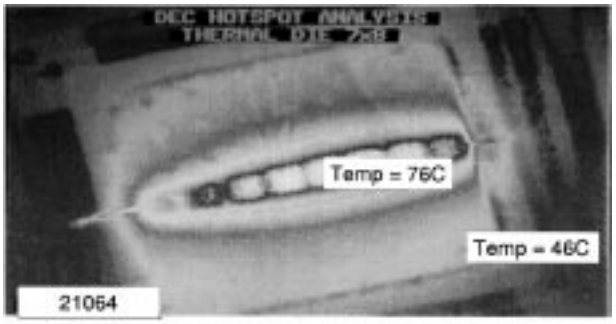

Fig. 16. 21064 thermal image.

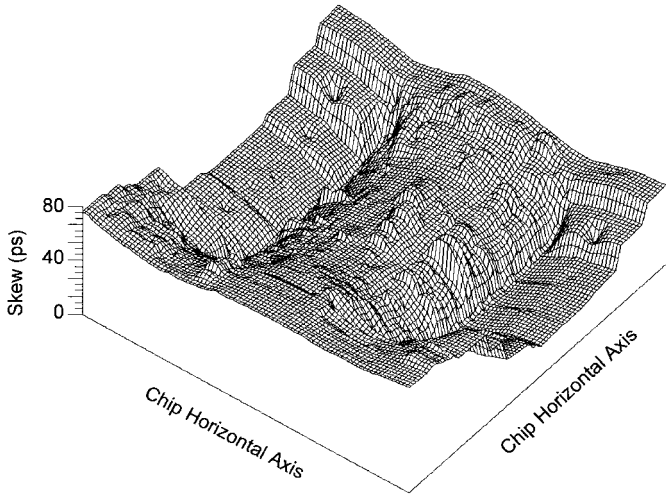

Fig. 17. 21164 clock skew.

Fig. 13(b) illustrates the location of the main clock drivers on the 21164. The main clock driver is split into two banks and is placed midway between the center of the die and the edges. A predriver is located in the center of the die to distribute the clock to the two main drivers. The clock skew was reduced by a factor of 2 using this approach (Fig. 17). In addition, by distributing the main clock driver over a larger area, the localized heating seen on the 21064 was reduced. The thermal image of the 21164 in Fig. 18 shows reduced temperature gradient.

As more aggressive circuit techniques and complex microarchitectural features were implemented in the 21264, power

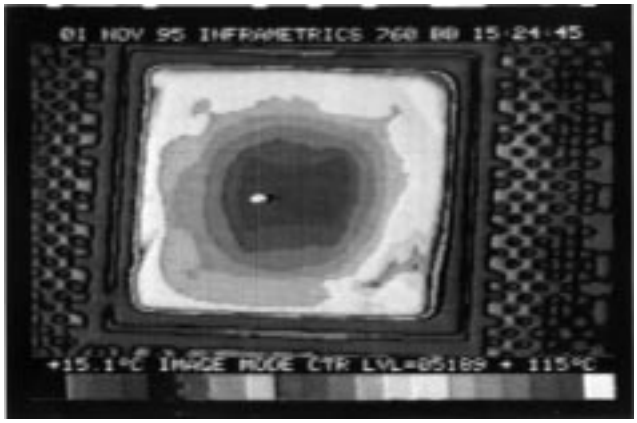

Fig. 18. 21164 thermal image.

consumption became a major concern in designing the clocking system. A single wire global clock (GLK) is routed over the entire chip as a global timing reference. The GCLK drivers are distributed in a window pane pattern as shown in Fig. 13(c) to reduce clock grid RC delay and distribute clock power.

GCLK is the root of a hierarchy of thousands of buffered and conditioned local clocks used across the chip as shown in Fig. 19. There are several advantages to this clocking scheme. First, conditioning the local clocks saves power. Second, circuit designers can take advantage of multiple clocks. For example, a phase path can be extended by initiating it with GCLK and terminating it with a delayed section clock. This approach significantly complicates race and timing verification, which will be discussed later in this paper. Finally, using local buffering significantly lowers the GCLK load, which reduces GCLK skew. Extensive electrical analysis was performed on the clock distribution network. The GCLK skew is less than 75 ps as shown in Fig. 20.

C. Latch Design

Latch design is another important element of the microprocessor circuit design strategy [10]. In order to ensure proper operation across all operating conditions, clock and latch circuits cannot be designed independently. Each generation of the Alpha microprocessor has utilized latches with improved characteristics combined with the improvements to the clock distribution networks previously described.

The high clock frequencies and small number of gate delays available per cycle on Alpha implementations has made low-latency latch design essential. In addition to high speed, other primary goals in latch design are minimal area and clock loading, low power dissipation, and low setup and hold times. The capability of embedding a logic function directly in the latch also helps reduce the number of gate delays per cycle. The 21064 was Digital's first microprocessor to use a two-phase, single wire-clocking scheme. This was a radical departure from the four-phase scheme that allowed for race-free design in previous versions of Digital's microprocessors. This major change in design methodologies has required designers to develop new strategies to manage noise and race-through issues.

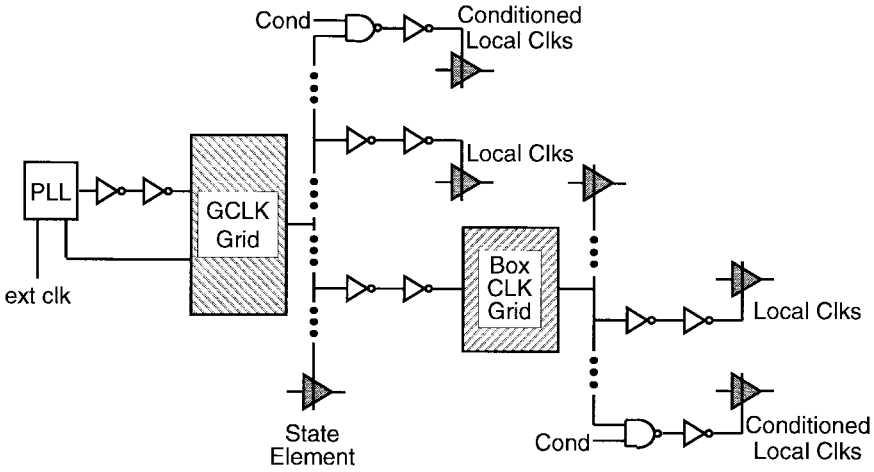

Fig. 19. 21264 clock hierarchy.

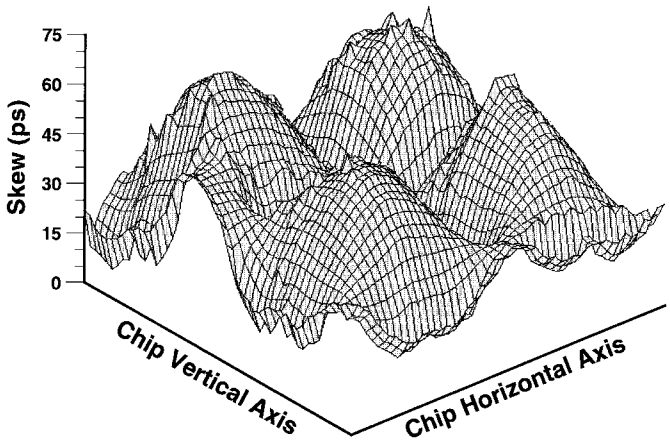

Fig. 20. 21264 GCLK skew.

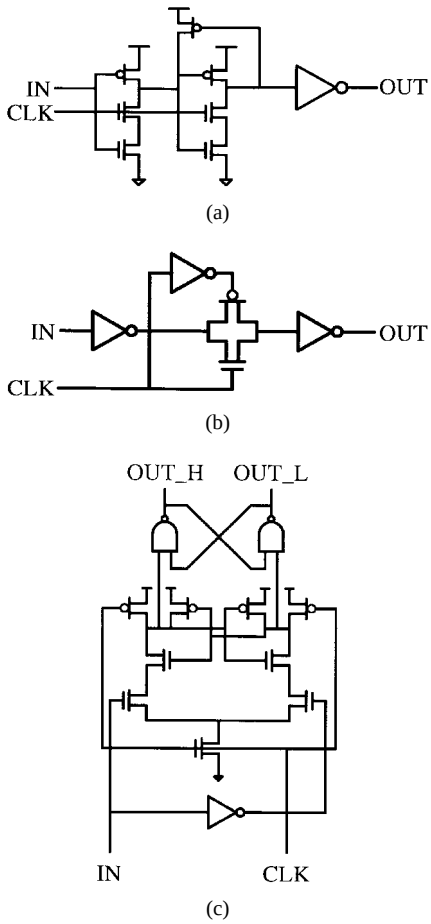

Fig. 21. Latches.

To reduce the chance of data race-through on the 21064, a variation of the true single-phase clocked (TSPC) level-sensitive latch developed by Svensson and Yuan was used [11], [12]. An example of this latch is shown in Fig. 21(a). These latches use the unbuffered main clock directly, significantly increasing race immunity. As long as the main clock edge rate is kept reasonably fast compared to the latch delay, there is little chance for data race-through. An additional benefit of this latch is that the first stage can incorporate a simple logic function. The combined delay of the high- and low-phase latches consumed about 25% of the cycle time.

One goal of the 21164 design was to increase the clock frequency by more than could be provided by process scaling. To accomplish this goal, the number of typical gate delays per cycle was reduced from 16 to 14. To offset the reduction in available gates per cycle, a lower latency latch was employed. Fig. 21(b) shows the basic dynamic CMOS transmission gate latch used on the 21164 [13]. This latch requires true and complementary clock signals, one of which is generated locally. The clock buffer for each latch type was custom designed so that its delay and edge rate characteristics could be tightly controlled. This additional buffer delays the clocking of the latch by one gate delay after the global clock transitions. However, since the preceding latch opens with the global clock transition, the possibility of latch race-through is significantly increased. In order to minimize the possibility of data race-through with the use of these latches, at least one minimum logic delay element was required between all latches. This constraint was easily verified with a simple CAD tool.

Consistent with the previous latch family, simple logic elements were built into the input stage, thus lowering the overhead associated with the use two of these latches to

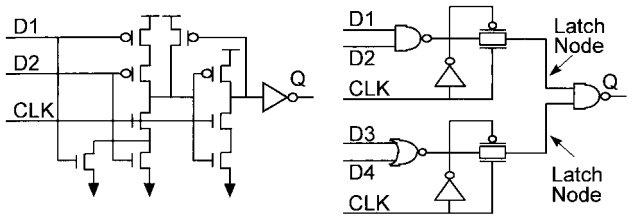

Fig. 22. Embedding logic into latches: (a) 21064 Function Latch: one level of logic; (b) 21164 Function Latch: two levels of logic.

15% of the cycle time. In some instances, both inverters were replaced with logic gates, further reducing the latching overhead. Examples of embedded logic in the two latch families are shown in Fig. 22. It can be seen that the cost of latching in the 21164 could be reduced to a minimum of one pass transistor.

Again, a primary goal of the 21264 design was to further increase the clock frequency by more than the process scale factor. Additionally, the 21264 utilized many conditional clocks to reduce power, thus requiring a static latch design. A family of edge-triggered flip-flops, based on the dynamic flip-flip show in Fig. 21(c), was developed to simplify the timing and race issues that were exacerbated by the addition of conditional clocking. Despite reducing the cycle time, the latching overhead was kept constant by using a flip-flop-based design that required only one latching element per clock cycle. The change from level-sensitive to edge-triggered design techniques and the use of multiple clock buffers introduced a number of new timing issues to the design.

The capability of buffering and conditioning the main clock is possible as long as each circuit satisfies both its critical path and race requirements. The left example in Fig. 23 illustrates the use of two buffered section clocks, ECLK and FCLK, in a one-cycle path. The circuit on the right of the figure combines a buffered local clock and a conditioned local clock to define a one-phase path. For both examples, critical path and race analysis start with the identification of the common clock initiating both the receive and drive paths, denoted R and D, respectively. Every critical path or race is defined by a single common clock and a pair of receive and drive paths. In the examples, GCLK and FCLK are the respective common clocks. Critical path analysis verifies that the difference in delay between the drive path D plus the receiver setup time and the receive path R does not exceed the phase or cycle time of the common clock. For worst case analysis, effects that minimize R and maximize D are considered. The converse is true for races; effects that maximize R and minimize D are considered. When the ratio of delay of the drive path D to the receive path R, including hold time, is equal to 1, the circuit is on the verge of failing. Ratios (denoted by X) exceeding 1 imply a margin that may be used to account for effects not included in the analysis. Given the clocking hierarchy and timing methodology used on the 21264, the ability to control and predict, in relative terms, the minimum and maximum path delays was essential in accurately predicting

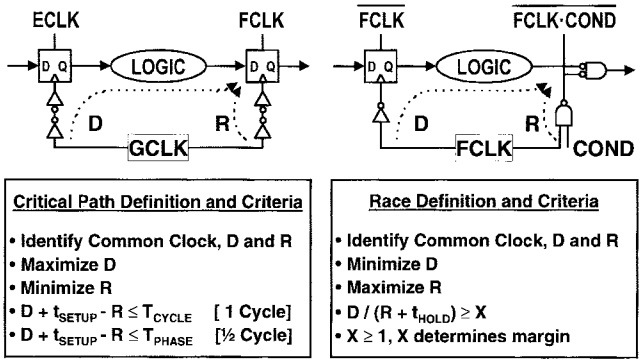

Fig. 23. Critical path and race analysis.

cycle time and ensuring functionality. Factors that could not be managed by the design were accounted for by additional margin.

V. CAD Tools and Verification

Custom circuit techniques allow designers to build very high-speed circuits. However, the use of these custom circuits requires design expertise and detailed postlayout electrical verification. Commercially available EDA design systems and point CAD tools have very limited support for these techniques. Therefore, Digital developed an extensive suite of in-house CAD tools to facilitate the design of custom VLSI microprocessors. Internally developed CAD tools are used extensively in all phases of microprocessor design, from initial performance evaluation, through circuit implementation and final design verification. The internally developed CAD suite includes tools for schematic and layout entry, two-state and three-state logic simulation, RTL versus schematic equivalence checking, static timing analysis and race verification, parameter and netlist extraction, and electrical analysis and verification.

A. Electrical Verification

Electrical verification covers all circuit issues that are not related to logic functionality such as timing behavior, electrical hazards, and reliability. Electrical hazards result from noise sources interfering with the logical functions of the chip, and include charge sharing, interconnect capacitive and inductive coupling, power supply IR noise, and noise-induced minority-carrier charge injection [14]. Reliability checks include metal and via electromigration, transistor hot-carrier damage, ESD, and latch-up failure.

The primary goal of the electrical verification tools is to verify that all circuits conform to the project design methodology. The design methodology defines an acceptable set of circuit styles and sizing rules that, when followed, ensures functionality with minimal analysis. The methodology also forces a consistent design style to be used project-wide, which has the added benefit of simplifying CAD design. However, occasionally, there is a need to design circuits outside the methodology to meet performance or area goals. In these instances, additional manual verification is required to ensure functionality.

Detailed analysis requires complex models with many process and circuit variables. Many checks are complex and hard to define as procedures that can be completely automated. Exact chip-wide analysis is impractical; instead, the tools perform design filtering. The tools filter out all circuits that can easily be validated while identifying the small number of circuits that may have problems and require additional analysis. This approach focuses design attention on potential problem areas, and therefore helps improve overall design efficiency. The CAD tools perform over 100 unique electrical checks. Some of the major areas of focus are circuit topology violations, dynamic node checks, including charge sharing, IR noise, injection, and leaker usage, interconnect coupling, noise margin checks, writeability checks, latch checks, beta ratio checks, gate fan-in and fan-out restrictions, transistor size and stack height limitations, max/min edge rates and delays, and power consumption. Some of the checks are applied to all circuit styles, while other checks are required for specific circuit types. The CAD tools require a large amount of design information to perform these checks, including electrical parameter extraction from layout, device electrical characteristics, relative circuit locations, timing information, and transistor connectivity.

B. Functional and Logical Verification

The functional complexity and time-to-market pressures of microprocessor design necessitated the development of an extensive functional verification strategy covering all phases of the design process from functional definition through manufacturing tests.

During the microarchitectural design phase of chip development, a two-state RTL behavioral model is the primary verification vehicle. This model provides a balance between design detail and simulation speed. The model is combined with abstract behavioral models of the other system components to verify correct operation of the processor in the system environment. Once logic design is complete, a two-state gate-level simulation model is extracted from the circuit schematics. This model is used to ensure that the schematics match the RTL model. Finally, to verify correct initialization of the circuits at power-up a three-state switch-level model is extracted from the circuit schematics.

A wide variety of simulation stimuli is used to verify the design, including hand-coded test patterns and randomly generated test patterns. Coverage analysis guides the verification process. Many of the manufacturing test patterns are derived from the simulation stimuli. Fault simulation is used to direct test enhancement.

VI. Future Challenges

As clock frequencies continue to increase with each new technology, many significant challenges are on the horizon. Power consumption has increased, despite voltage scaling, to the point of being a first-order concern. Designers must identify innovative solutions to lower power without significantly impacting performance. In addition, the distribution of a chipwide timing reference with extremely fast edge rates poses a significant challenge. The size and architectural complexity of the next generation microprocessor will require improved design and verification methodologies. Finally, designer productivity must increase significantly, requiring innovative CAD tools to meet time-to-market constraints.

VII. Conclusion

This paper has reviewed Digital's approach to high-performance microprocessor design. Three generations of Alpha microprocessors have been designed and optimized for performance by focusing on high-speed design using fully custom circuit design techniques, incorporating state-of-the-art architectural features, and utilizing high-performance CMOS technology for fabrication. An extensive suite of in-house CAD tools is used to analyze complex electrical and logical behavior to ensure functionality. All three microprocessors booted multiple operating systems using first-pass silicon, validating the design and verification methodologies.

Acknowledgment

The microprocessors described in this paper have resulted from the work of a tremendous number of people. The authors would like to recognize the technical contributions of the 21064 design team, including D. Dobberpuhl, R. Witek, J. Montanaro, L. Madden, and S. Samudrala; the 21164 design team, including J. Edmondson and P. Rubinfeld; and the 21264 design team, including B. Gieseke, J. Keller, and D. Priore. Finally, they would also like to acknowledge the significant contributions made to these microprocessors by the verification, CAD, process development, and manufacturing teams.

References

[1] D. Dobberpuhl *et al.*, "A 200 MHz 64 b dual-issue CMOS microprocessor," *IEEE J. Solid-State Circuits*, vol. 27, pp. 106–107, Nov. 1992.

[2] W. Bowhill *et al.*, "A 300 MHz 64 b quad-issue CMOS microprocessor," in *ISSCC Dig. Tech. Papers*, Feb. 1995, pp. 182–183.

[3] B. Gieseke *et al.*, "A 600 MHz superscalar RISC microprocessor with out-of-order execution," in *ISSCC Dig. Tech. Papers*, Feb. 1997, pp. 176–177.

[4] R. Sites and R. Witek, *Alpha AXP Architecture Reference Manual*, 2nd ed. Boston, MA: Digital, 1995.

[5] J. Edmondson *et al.*, "Superscalar instruction execution in the 21164 Alpha microprocessor," *IEEE Micro*, vol. 15, pp. 33–43, Apr. 1995.

[6] J. Kowaleski *et al.*, "A dual-execution pipelined floating-point CMOS processor," in *ISSCC Dig. Tech. Papers*, Feb. 1996, pp. 358–359.

[7] B. Benschneider *et al.*, "A 300-MHz 64-b quad-issue CMOS RISC microprocessor," *IEEE J. Solid-State Circuits*, vol. 30, pp. 1203–1214, Nov. 1995.

[8] L. Heller and W. Griffin, "Cascode voltage switch logic: A differential CMOS logic family," in *ISSCC Dig. Tech. Papers*, Feb. 1984, pp. 16–17.

[9] D. Priore, "Inductance on silicon for sub-micron CMOS VLSI," in *1993 Symp. VLSI Circuits, Dig. Tech. Papers*, May 1993, pp. 17–18.

[10] P. Gronowski and W. Bowhill, "Dynamic logic and latches: Practical implementation methods and circuit examples used on the ALPHA 21164," in *1996 Symp. VLSI Circuits—Proc. VLSI Circuits Workshop.*

[11] D. Dobberpuhl *et al.*, "A 200-MHz 64-bit dual-issue CMOS microprocessor," *Digital Tech. J.*, vol. 4, no. 4, pp. 35–50, 1992.

[12] P. Larsson and C. Svensson, "Impact of clock slope on true single phase clocked (TSPC) CMOS circuits," *IEEE J. Solid-State Circuits*, vol. 29, pp. 723–726, June 1994.

[13] W. Bowhill *et al.*, "Circuit implementation of a 300-MHz 64-bit second-generation CMOS alpha CPU," *Digital Tech. J.*, vol. 7, no. 1, pp. 100–118, 1995.

[14] P. Gronowski *et al.*, "A 433 MHz 64 b quad-issue CMOS RISC microprocessor," in *ISSCC Dig. Tech. Papers*, Feb. 1996, pp. 222–223.

Paul E. Gronowski (M'96) received the B.S.E.E. degree from the University of Cincinnati, Cincinnati, OH, in 1984.

He joined Digital Equipment Corporation, Hudson, MA, in 1984, and has worked on both VAX and Alpha microprocessor designs in both NMOS and CMOS technologies. He was responsible for the chip physical verification effort and the integer execution unit on the 300-MHz 0.5-μm CMOS version of the Alpha 21164. He led the implementation of the 0.35-μm CMOS implementation of the Alpha 21164. He is currently a Consulting Engineer in Digital Semiconductor's Advanced Development Group, coleading the design of a future generation Alpha microprocessor. He jointly holds one patent and has coauthored several papers.

William J. Bowhill (M'93) received the B.Eng. (honors) degree in electrical engineering from the University of Liverpool, U.K., in 1981.

He is a Senior Consulting Engineer in Digital Semiconductor's Advanced Development Group, Hudson, MA. He was a lead designer on the VAX 6000/400, VAX 6000/600, and Alpha 21164 CPU's, and is currently working on the advance development of a future generation Alpha microprocessor. Prior to joining Digital, he worked for Standard Telecommunications Laboratories, Harlow, U.K., where he designed VLSI chips for telecommunication applications. He jointly holds three patents and has coauthored 12 technical papers.

Ronald P. Preston (M'88) received the B.S.E.E. and M.Eng. degrees from Rennselaer Polytechnic Institute, Troy, NY.

He is a Consulting Engineer in the Alpha Microprocessor Advanced Development Group at Digital Semiconductor, Hudson, MA. Since joining Digital in 1988, he has been involved in the design of six microprocessors in four different CMOS processes. Most recently, he was the Implementation Leader for the 550-MHz Alpha 21164PC chip. He is currently the coimplementation leader for a future high-performance Alpha microprocessor design. He has coauthored several papers on microprocessor design and reliability verification of high-performance CPU's.

Michael K. Gowan received the B.S.E.E. degree from Purdue University, West Lafayette, IN, and the M.S.E.E. degree from North Carolina State University, Raleigh.

He is a Principal Engineer at Digital Semiconductor, Hudson, MA. He was responsible for the System and B-Cache Controller and Bus Interface Unit designs on the Alpha 21264 microprocessor. He previously led the implementation of the 0.5-μm version of the Alpha 21066 microprocessor. For the past ten years at Digital Equipment Corporation, he has made technical contributions to several microprocessor and network chip designs.

Randy L. Allmon (A'92) received the B.S.E.E. degree from the University of Cincinnati, Cincinnati, OH, in 1981.

He is a Consulting Engineer in Digital Semiconductor's Advanced Development Group, Hudson, MA. He is currently the Implementation colead of the 0.35-μm Alpha 21264 microprocessor. In the past 17 years at Digital, he was a lead designer and implementation leader on several high-performance VAX and Alpha microprocessors. He has one patent and has coauthored eight technical papers.