Reduced-voltage power/performance optimization of the 3.6-volt PowerPC 601 Microprocessor

by K. Bernstein
J. E. Bertsch
L. G. Heller
E. J. Nowak
F. R. White

An experimental 2.0-volt low-power PowerPC 601™ Microprocessor built in a modified 3.6-volt, 0.6-μm IBM CMOS technology is described. By using unmodified masks from the 3.6-volt design, a 3× power savings was realized while maintaining nearly the original performance. The use of selective scaling provides high performance at reduced power supply voltage. This technique, applicable to selected existing product designs, may allow early entry into the low-power market while minimizing new process development expense. The technique proposes hyperscaled reductions in specific electrical and physical parameters, while keeping horizontal layout rules unchanged. Static chip designs, which comprise the majority of 601 circuitry, respond well to the alterations. In addition, potential reliability detractors are reduced or eliminated. Challenges to this technique include I/O interfacing and minimizing leakages associated with low device thresholds. The 601 design and its base technology are described, along with the experimental changes. The paper reviews the motivation behind low-power microprocessor development, alternative power-saving techniques being practiced, and opportunities for continued power reduction.

Introduction

The use of fully capable microprocessors in portable consumer electronics represents one of the fastest-growing segments of the electronics market. These applications include computing (tablet, laptop, and notebook computers), entertainment (gameboys, virtual reality toys), and communications (cellular phones, wireless modems). By one estimate [1], the subnotebook computer market alone will grow at a 91% compound annual growth rate through 1997, easily exceeding growth rates of the workstation and deskside/desktop segments.

0018-8646/95/$3.00 © 1995 IBM

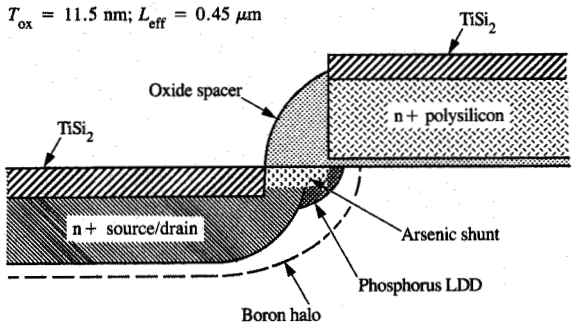

Figure 1

Schematic cross section of the n-type MOSFET source/drain structure used in this experiment. The arsenic shunt provides a low-resistance path from the deep junction to the gate edge, while a phosphorus LDD and a boron halo are used to moderate hot-electron effects and short-channel effects, respectively.

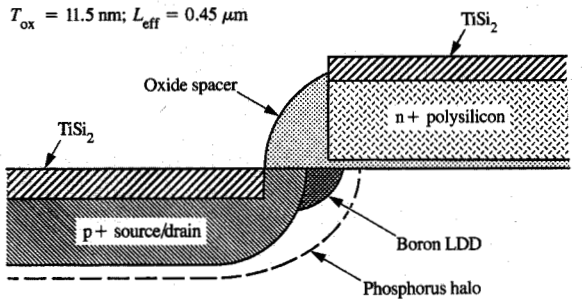

Figure 2

Schematic cross section of the p-type MOSFET source/drain structure used. As in the n-MOSFET, an LDD and a halo are used to control hot electrons and short-channel effects, respectively.

Price sensitivity in many of these consumer products is profound. A difference of just a few dollars in many cases defines the success of the final product. The "chip wars" spawned by this growth have put real limits on the chip's cost, its size, and its complexity, as well as its development expense. Against this backdrop, the portable-application microprocessor developer must provide more performance and function with an order of magnitude less power; battery technology has not been able to keep up with power consumption. Consider notebook computer evolution: A best-of-breed A4-type notebook strives for four times the performance, ten times the storage, half the weight, twice the power consumption, and twice the battery life of its predecessor [2]. The present microprocessor power ceiling for portable applications, generally considered to be less than 3 W, will have to be less than 1 W in the next generation. Clearly, a major shift in microprocessor design is needed to meet these requirements, but it must be done inexpensively and quickly, given the fierce market environment facing the chip producer.

To meet the market need, a CMOS-technology scaling technique has been demonstrated at the IBM Microelectronics chip fabrication facility in Essex Junction, Vermont [3]. This approach may allow microprocessor developers to radically reduce the power consumption of a CPU without sacrificing performance, reliability, or yield, or incurring excessive development expense. The simplicity of the technique will assist in moving certain existing designs into the market with little to no design modification.

Selective scaling

Conventional CMOS concepts of scaling vertical and horizontal device dimensions and the power supply voltage by a common factor are well known [4]. With the exception of power supply voltage (V_{DD}) and threshold voltage (V_t), the principles of MOS scaling have basically been practiced, within the industry, through several technology generations. Decreases in V_t with technology scaling, however, have been limited. Subthreshold (leakage) current in MOSFETs is due to weak inversion carriers, whose population density is proportional to the Boltzmann factor, $e^{-\phi_s/kT}$, where k, T and ϕ_s are Boltzmann's constant, the temperature (absolute), and the silicon surface potential, respectively. Since ϕ_s is proportional to $(V_G - V_t)$, decreasing V_t leads to exponentially increasing leakage current, thereby limiting the amount of V_t reduction possible. Long-lasting power supply voltage standards, such as the 5-V standard, have historically discouraged the scaling of system-level power supplies. Additionally, high-performance circuit design requirements have limited the allowable reduction in device drive, $V_{DD} - V_t$, and will continue to do so [5]. Active power dissipation has not been a first priority, particularly in the high-performance system/processor design arena, and the leverage of reducing active power quadratically, i.e., in proportion to V_{DD}^2, has not received wide attention. However, in the last few years, chip power has become a more important issue, perhaps first in SRAM, then DRAM, and logic and processors. More recently, proposals for low-power design have included a rapid reduction in V_{DD}, with reduced V_t, scaled technology dimensions, and possible parallel architectures to regain lost circuit performance [6, 7]. Circuit techniques to reduce sub-V_t leakage currents have also been discussed [8–10]. The old notion that low power equals low performance will soon not necessarily be true.

The objective in this work is to quickly demonstrate a 3.5× power reduction in a high-performance, 0.5-μm CMOS product technology, with minor process changes and no mask change. As a demonstration vehicle, we chose the 3.6-V PowerPC 601™ Microprocessor, an existing, well-characterized product. Our approach is to dramatically reduce T_{ox} (gate oxide), V_{DD}, and V_t, i.e., a selective scaling of device parameters chosen to achieve power/performance without mask change and with only minor process changes. Reliability exposures are minimized at the reduced power supply voltage; cf. [11]. Elimination of the need for horizontal scaling allows relatively quick and inexpensive implementation of the technique, compared to the retooling typically required by a full technology scaling.

Base process description and alterations

The technology used to produce the PowerPC 601 chip is a CMOS process with conventional LOCOS isolation. A single n+-doped polysilicon layer acts as the gate electrode for both device types. To provide low sheet resistances, a silicide is formed on diffusions and on the polysilicon simultaneously. A retrograde n-well in 2.5-μm epitaxially grown silicon provides superb latch-up immunity. The n-FET design is an extension of the doubly implanted lightly doped drain (DILDD). As shown in **Figures 1** and **2**, a triple-implant LDD (TILDD) process is used which consists of phosphorus and arsenic implants to serve as the LDD, plus a boron halo implant for short-channel V_t control. Arsenic is employed as an alternate means of solving the gate-edge high-field problem and provides, in effect, a fully overlapped LDD. Hot-carrier and sustaining voltage control are provided by the LDD such that operation at potentials as high as 4.0 volts is feasible.

A low-resistance tungsten local interconnect [12] performs two functions. It is part of the contact path to M1 and also allows interconnection of polysilicon, n+ diffusions, and p+ diffusions without occupying first-level-of-metal wiring channels. This interconnect-and-contact stud requires no wide border at diffusion and polysilicon. In addition, it provides a planar topography for subsequent metal levels. After passivation of the local interconnect, another damascene tungsten stud is used to provide contact between the local interconnect and the first level of metal. All metal levels are aluminum/copper-based. The interlevel dielectric passivation layers are fully planarized between metal levels, allowing aggressive metal pitches of 2.0 μm while maintaining low defect levels. While up to five metal levels can be used by the designer in this technology, the PowerPC 601 chip uses only four in addition to the local interconnect [13].

In order to maintain functionality of the PowerPC 601 at reduced V_{DD}, it was necessary to reduce both the V_t and T_{ox} at least as fast as V_{DD} for the n-FET. The p-FET

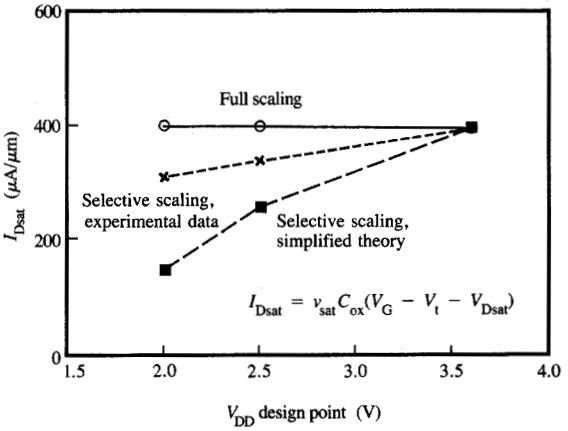

Figure 3

Drain current in an n-type MOSFET (with $V_{GS} = V_{DS} = V_{DD}$) vs. V_{DD} design point for full scaling is compared to the selective scaling case used here. The diamonds represent the simplified theory in which it is assumed that V_{Dsat} is not changed by T_{ox} and V_t. Experimental data (×) show that such an approximation is much too pessimistic.

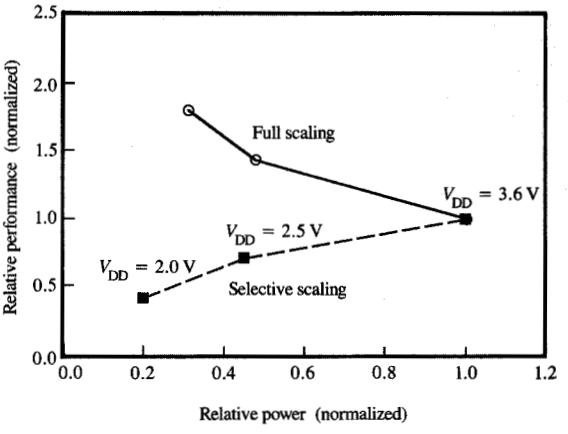

Figure 4

Performance (normalized to the 3.6-V-design case) is plotted vs. active power (also normalized to the 3.6-V-design case) for selective scaling (experimental data) vs. full scaling (theoretical expectation). Significant power reduction is achieved with selective scaling while minimizing performance loss. Full scaling would also deliver performance gains, but at the cost of retooling.

Table 1 Process conditions for standard and selectively scaled products.

V_{DD} (V)	3.6	2.5	2.0
n-FET ion implant (cm^{-2})	2.4×10^{12} BF_2	3.4×10^{12} BF_2	3.4×10^{12} BF_2
p-FET ion implant (cm^{-2})	2.4×10^{12} P^{31}	4.4×10^{12} As^{75}	4.2×10^{11} As^{75}
	2.4×10^{12} BF_2	6.8×10^{12} BF_2	3.4×10^{12} BF_2
T_{ox} (nm)	11.5	7.5	5.5
E_{ox} (MV/cm)	3.1	3.3	3.6

Table 2 IBM PowerPC 601 characteristics.

Die size	10.95 mm × 10.95 mm
Performance	80 MHz; 66 MHz in experiment*
Power consumption	8.5 W; 7.7 W in experiment*
Device count	2.8 million
Signal I/O	184
Power supply	3.6 ± 5% V
Packaging	304-pin C4 ceramic quad flat pack
Temperature range	0 to 100°C

*An earlier vintage of the product was used in the experiments.

Table 3 Composite plot of performance as a function of V_{DD} of 321 modules across all experimental wafer splits. The number of modules passing at the conditions specified by the row and column headings is shown.

V_{DD} (V)	*Cycle time* (ns)						
	12	*15*	*18*	*21*	*24*	*27*	*30*
3.6	0	0	13	5	3	0	0
3.3	0	0	120	89	75	6	0
3.0	0	0	168	140	141	124	78
2.7	0	49	202	123	202	184	155
2.4	0	40	85	43	111	184	198
2.1	0	5	30	7	40	48	50
1.8	0	0	0	0	0	0	3

shares the same gate oxide process with the n-FET. Since the load capacitance would not be scaled in this experiment, we would need to keep the MOSFET drain currents constant with (selective) scaling just to keep constant performance with reduced V_{DD}. A simple analysis gives, for the MOSFET current,

$$I_{Dsat} = W v_{sat} C_{ox} (V_G - V_t - V_{Dsat}),$$

where v_{sat} (8×10^6 cm/s) is the saturation velocity in silicon, W is the MOSFET width, C_{ox} the gate capacitance per unit area, and V_{Dsat} the drain voltage at which the MOSFET enters velocity saturation [14]. In full scaling, V_G, V_t, and V_{Dsat} would scale in proportion to V_{DD}, while C_{ox} would scale inversely. This clearly gives constant current per width. In the selective scaling discussed here, the outcome differs from the conventional scaling case in that V_{Dsat} decreases more slowly than V_{DD}, since L_{eff} is not reduced. In **Figure 3**, full scaling is compared to our selectively scaled approach. For a worst-case theoretical analysis, V_{Dsat} was assumed to remain constant with selective scaling cases. In reality, there is some reduction of V_{Dsat} in the selective scaling case. We expect performance f to behave roughly as

$$f = I_{Dsat}/(C_{load} V_{DD}),$$

while power consumed from switching circuits behaves as

$$P = N_{sw} f C_{load} V_{DD}^2,$$

where N_{sw} is the number of circuits switching with capacitance C_{load}. By using the selective scaling (including V_t) case as approximated above, as well as the traditional scaling case, power versus performance is shown in **Figure 4**. Note that selective scaling improves power consumption while attempting to minimize the loss in performance, while full scaling simultaneously does both. This advantage (of full scaling) is tempered, however, by the investment requirements demanded by scaled lithography.

Process changes were made to target 2.5-V and 2.0-V design points. Since the p-FET has a compensated channel, arsenic was substituted for phosphorus in the channel in an effort to maintain short-channel behavior in the scaled processes. In the 3.6-V process, T_{ox} was reduced from 11.5 nm to 4.9 nm and 7.0 nm for the 2.0-V and 2.5-V design points, respectively. Polysilicon gate-electrode depletion resulted in electrical equivalent gate-oxide thicknesses of 5.5 and 7.5 nm for the 2.0-V and 2.5-V cases, respectively. **Table 1** summarizes the experimental details. The n-FET I_{Dsat} achieved is also included in Figure 3. Note that nearly constant I_{Dsat} was achieved by reducing the equivalent oxide thickness t_{eq} and V_t slightly faster than V_{DD}. The p-FET I_{Dsat} was significantly reduced with selective scaling, however, as discussed below.

Vehicle description

The vehicle for the experiment, the IBM PowerPC 601 microprocessor, is the first implementation of a series of reduced-instruction-set computer (RISC) processors. The primary characteristics of the PowerPC 601 are described in **Table 2**.

Physically, the part comprises three distinctive design styles and components built hierarchically. The first is a 32KB combined instruction and data cache which is eight-way set-associative and has an eight-word-wide fetch bus.

The cell comprising the array is a standard six-device type, 65 μm^2 in area. In addition, there are smaller general-purpose register (GPR) arrays on the chip. The 32-bit data flow on the chip is handled in a "bit stack" of custom dataflow macros which processes the 32-bit word in parallel. Control logic steering the stack is implemented in random logic macros (RLMs). Each RLM is composed of books executing basic combinatorial logic. Array bit line redundancy, error correction coding, and word parity are built into the product.

The circuit style on the chip is predominantly static. There is no self-timed, dynamic domino, or DCVS-style circuitry on board. Clock buffers and redrivers on the chip shape input clocks, but do not run autonomously. There are no phase-locked loops on the product. A limited amount of ratioed logic circuitry is used, as discussed below.

The logic function of the part is organized into floating-point, fixed-point, and branch execution unit partitions. Up to three instructions may be dispatched per cycle by the instruction queue via the dispatch unit in order to keep the RISC pipeline full. Preservation of complex timing relationships was a benchmark of the technique's success, as evidenced by functioning modules.

There were no alterations to the mask or the design for the experiment. Wafer-level horizontal dimensions are identical to those of the standard production product.

Experimental results

The product response to the experiment is shown graphically in **Table 3** and **Figures 5** and **6**. Table 3 shows the functionality window for the experiment including both design points. Minimum chip cycle time is plotted against the range of supply voltage. Figures 5 and 6 are histograms of maximum chip operating frequencies across 2.0-V and 2.5-V wafer experimental groups. **Figure 7** shows the voltage/performance relationship for a typical part in the 2.0-V experimental group.

Generally, performance of up to 68.4 MHz was observed at room temperature for V_t as low as 2.1 V. The average performance measured on experimental hardware was 62 MHz. The masks used were from an earlier version of the 601 which achieved 80-MHz performance at room temperature and 66 MHz at 85°C, with a standard process. Standby currents, which on the standard product rarely exceed 100 μA, were observed to be between 25 and 45 mA in the experiment. Active power was found to be 2 W on average while the standard production test patterns were running, as compared to 7.5 W seen on standard production hardware at 80 MHz. This is shown graphically in **Figure 8**. Functional module yield of experimental hardware was equivalent to that found on standard 3.6-V production hardware. Except for the modification of input-drive and output-sense voltage levels, standard production

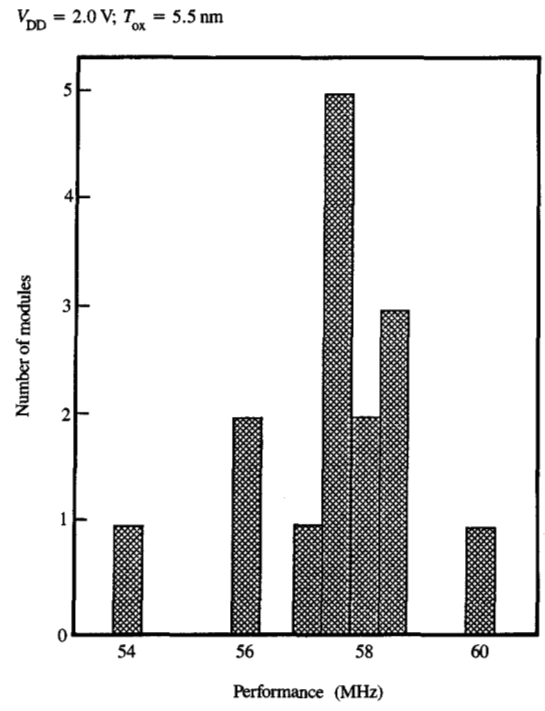

Figure 5

Performance histogram of 2.0-V design.

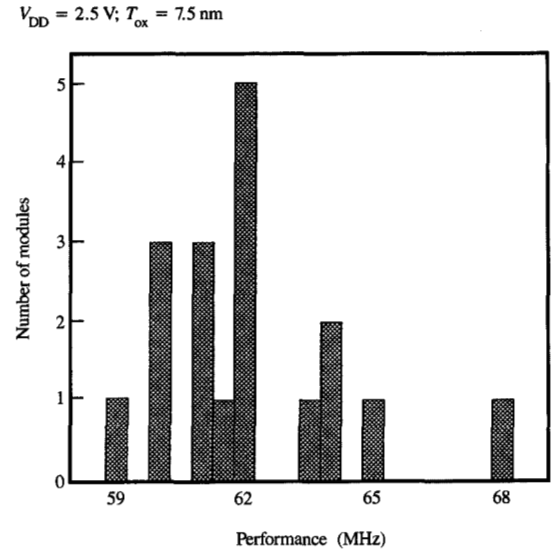

Figure 6

Performance histogram of 2.5-V design.

		Cycle time (ns)																									
		10.00	10.75	11.50	12.25	13.00	13.75	14.50	15.25	16.00	16.75	17.50	18.25	19.00	19.75	20.50	21.25	22.00	22.75	23.50	24.25	25.00
	2.500	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	2.450	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	2.400	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	2.350	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	2.300	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
V	2.250	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
D	2.200	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
D	2.150	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	2.100	-	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*
s	2.050	-	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*
u	2.000	-	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*	*
p	1.950	-	-	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*	*
p	1.900	-	-	-	-	-	-	-	-	-	-	*	*	*	*	*	*	*	*	*	*	*	*
l	1.850	-	-	-	-	-	-	-	-	-	-	*	-	-	-	*	*	*	*	*	*	*	*
y	1.800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*	*	*	*
	1.750	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	1.700	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(V)	1.650	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	1.600	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	1.550	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	1.500	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

*denotes one or more functional modules

Figure 7

Functional envelope for a typical module from the 2.0-V experimental wafer group. Asterisks represent functionality.

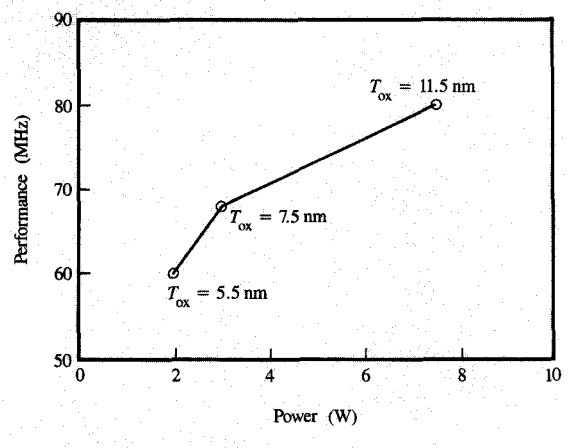

Figure 8

Performance vs. power for standard and experimental modules, as measured at their targeted operating voltages.

criteria were used in testing the modules. Finally, other than the well implanting and gate oxide growth, the parts were fabricated using the standard 601 fabrication process, and received normal handling through the rest of the wafer and module fabrication process.

Design response

The predominant concern in moving an existing design into a new fabrication process is its functionality, and the loss of functionality margin with changes in conditions. To quantify those changes in PowerPC 601, the shift in the unity gain point of a typical static circuit was analytically determined, and is plotted in **Figure 9**. To anticipate changes in noise immunity and in worst tolerable most-positive-down levels (MPDL) and least-positive-up levels (LPUL), the normalized minimum voltage needed for functionality is plotted against the V_{DD} center for each experiment wafer group, and is shown in **Figure 10**. While both measures indicate some loss of window, the remaining margin is considered sufficient to provide good noise immunity. On a reduced-voltage planar, noise generally remains proportional to V_{DD}, so Figures 9 and 10 accurately reflect the cumulative change in use conditions. On planars where reduced-voltage components are added, extra care must be taken to isolate supply and signal noise associated with higher-voltage components on the card.

The composition of the capacitive load driven by each net shifted as a result of the experiment. The average net on the standard PowerPC 601 production part was found to be 76.5% gate load and 24.5% wire load. Of the total chip delay, an average of 17.4% was due to the combined *RC* of input gate and wire. One would anticipate that decreasing gate-oxide thickness without reducing L_{eff} would increase the percentage of total load due to gate inputs.

The standard CMOS circuit composition of the 601 chip adapts itself well to selective scaling. Its static circuitry made acceptable MPDL/LPUL more likely and motivated its selection as the test vehicle. There are no logic paths on the chip which anticipate delay in other paths without proper interlocking. Its performance is determined by the maximum speed at which its components can develop output and accurately capture it in a register after each cycle. That performance, however, depends on recapturing the device saturation current of the base process at reduced V_{DD}.

The biggest threat to 601 functionality in the experiment was in the limited number of ratioed logic circuits used. Beta-ratio-dependent circuitry relies on the n-FET–p-FET device current ratio to develop acceptable MPDL and LPUL levels. Most are used in wide NORs, and use grounded p-FETs, as shown in **Figure 11**. This circuit style may present the one greatest challenge to robust selective scaling. To retain functionality, the MPDL must remain at

or below the value found on production hardware. The p-FET device thresholds were not modified to ensure good down levels on grounded p-FET NOR circuits. Retuning of circuit device ratios would allow p-FET device threshold reduction as well.

The PowerPC 601 chip is rich in NAND structures. Good CMOS design technique exploits logic NANDs by stacking multiple n-FETs in series to ground, rather than NORs which stack multiple p-FETs to V_{DD}. In CMOS technology, n-FETs have more than twice the current of p-FETs. The choice not to modify p-FET thresholds fortunately has only a mild impact on overall performance in this case.

Products using charge-retention-sensitive circuitry were considered poor candidates for selective scaling. On this experiment, subthreshold leakage currents were observed to increase by more than three orders of magnitude. While that change may be tolerable for product running at normal speeds, many contemporary microprocessors are used in a variety of applications where fixed clock speed or presence may not be guaranteed. The presence of high leakage reduces the minimum clock period under which precharge or developed logic levels may discharge below minimum detectable up levels. There are a number of techniques in the literature to help counter the effects of higher subthreshold leakage. Their application to low-power/high-performance products is promising [8–10, 15].

Reliability results

Three predominant mechanisms limit the field reliability of any CMOS product. Random defects not screened out before shipment can emerge once in the field. The hot-electron effect gradually slows product performance by elevating the device threshold. Conductor electromigration gradually increases the resistance of the most heavily used interconnections on chip. All three effects have a supply voltage dependence.

By using a standard analytic technique for modeling random defects, it has been determined that the 2.0-volt 601 product enjoys twice the reliability of the standard 3.6-volt part, when burned in at $1.5 \times V_{DD}$ for five hours. Most defects are associated with the back-end-of-line interconnect process, which was not modified. Reliability calculations reveal that although fewer defects are screened out at burn-in with lower voltage, the lower V_{DD} seen by the product over its lifetime offsets the difference. The experiment aggressively scaled gate oxides to recapture performance. The dielectric strength of the thinner oxide used is equivalent to more conventional thicknesses and is not expected to be a reliability concern.

Electromigration (EM) was a persistent concern on the standard product part. Pulsed-dc-averaged contact currents and wire current density concerns are regularly associated with ratioed logic and supply connections of high-power-book outputs. Peak instantaneous ac currents and joule heating have limits often approached by outputs of high-power-level books. Electromigration severity is proportional to a net's average current:

$$I_{signal} = s_f \, f \, C_{load} \, V_{DD},$$

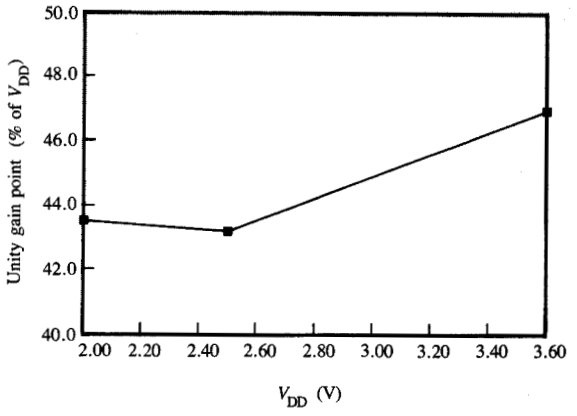

Figure 9

Inverter unity gain point as a function of V_{DD} design point.

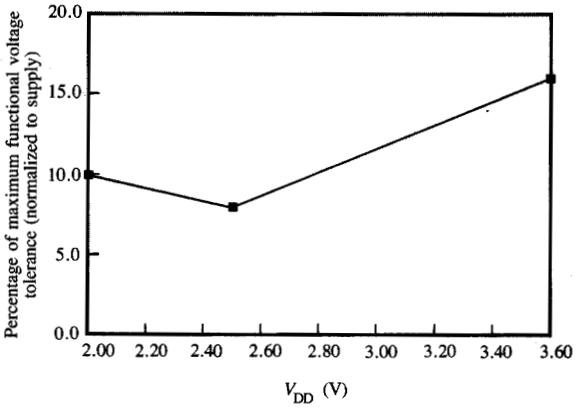

Figure 10

Maximum allowable V_{DD} tolerance vs. design point V_{DD}.

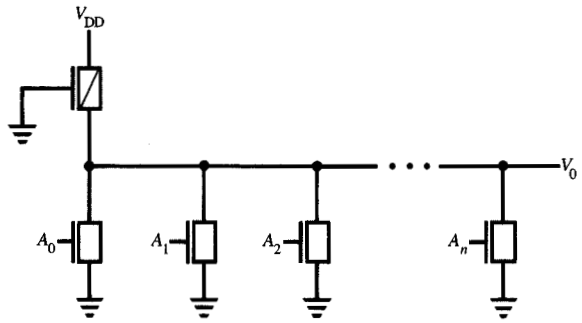

Figure 11

Schematic diagram of grounded p-FET NOR circuit. The maintenance of adequate down levels in ratioed logic is a challenge to selectively scaled designs.

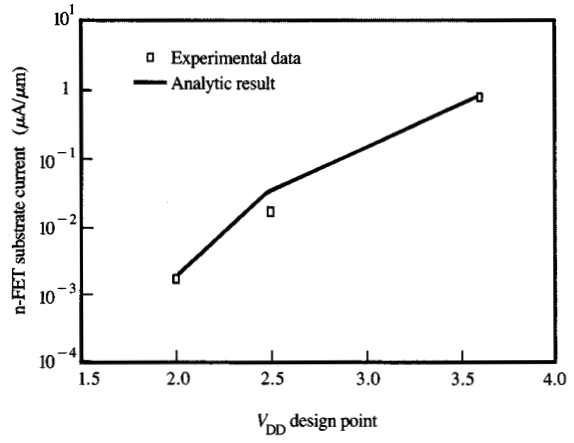

Figure 12

Substrate current in an n-type MOSFET (@ $V_{DS} = V_{DD}$, $V_{GS} = V_{DD}/2$) vs. V_{DD} design point.

where s_f is the switch factor of the net, f is the operating frequency, C_{load} is the total capacitance of the net, and V_{DD} is the supply voltage. In this experiment, the effect of decreasing gate oxide thickness is balanced by the reduced V_{DD}. The switch factor remains constant, and frequency is slightly decreased. The result is fairly EM-neutral. Electromigration varies with temperature, however, as $\exp(1/T_{max})$ and so enjoys the lower heat dissipation associated with lower V_{DD}.

Hot-electron degradation is greatly reduced by the reduction in electric fields afforded by the lower V_{DD}. The behavior of channel hot-electron degradation is usually well monitored by the substrate current, I_{sx}. It accompanies operation of the n-FET, which is due to avalanche multiplication in a high-field region close to the MOSFET drain. I_{sx} is approximately

$$I_{sx} = A_0 I_{DS} \exp[-\lambda E_{crit}/(V_{DS} - V_{Dsat})]$$

[14]. In **Figure 12** the simplified selectively scaled result above is compared to our experimental data. Indeed, a significant reduction in I_{sx} is achieved with the selective scaling as expected.

Additional technology opportunities

While the thrust of this investigation was to explore the leverage available without significant device or design alterations, additional improvement with only minor changes to both is readily available. A simple enhancement can be achieved in this case by omission of the LDD phosphorus in the n-FET; this is afforded through the electric field reduction due to the reduced V_{DD}. The lighter grading in turn results in significantly improved short-channel effects. Hence, operation of the modified n-FET at L_{eff} as short as 0.25 μm is possible. Similarly, the p-FET can afford a higher phosphorus halo dose to improve its short-channel characteristics, although in this case the minimum L_{eff} achieved is limited to 0.28 μm. These changes would allow a gate-level shrinkage of 0.05 μm (allowing for no change in manufacturing tolerances), which could be accomplished either through a new mask design or through lithographic techniques such as resist trim or etch-back. We expect an increase in performance of about 15% above the selective scaling actually demonstrated, which would reduce the performance loss (in the 2.0-V design point) to only 10% from the base (3.6-V) case. While power increases from the increased operating frequency, this will be countered by the decreased gate capacitance. We thus expect to maintain the 3.5× power reduction.

Future design practices

Earlier generations of microprocessor circuit design often showed little regard for power consumption. Despite more recent advances in power–delay product and in battery energy density, successful future products will have to integrate active power management at every level of design.

Circuit improvements will include schematic improvements such as cascoding which limit subthreshold leakage, and minimized-capacitance circuit implementations.

Architectural power management in future microprocessors will more distinctly identify and respond to the quiescent periods of smaller sections of the design. Chip logic configuration decisions will weigh added performance and function against its cost in power consumption.

Finally, system-level recognition of power concerns will be evident in changes to communication protocols as well as to compiler algorithms.

Summary

A new direction in CMOS technology modification has been explored for reducing power. The technique maintains nearly equivalent performance by selectively scaling specific chip parameters rather than improving performance by full scaling. Motivated by market demand for inexpensive products with lower power–delay product, this development extends the lifetime of existing semiconductor fabricator tooling. Shifts in design focus, along with reconsideration of new technology scaling, will provide the basis for a new generation of lower-power, higher-speed computing.

Acknowledgments

The authors wish to thank D. Cook, B. Corrow, A. Johnson, and S. Lawrence for assistance in the process and device work. Furthermore, the authors acknowledge J. Blatt, D. Flye, M. Hawkes, C. Jallipalli, M. Maurice, and T. Welch for their product and test support, and D. Bouldin for EM advice.

References

1. *Dataquest Worldwide Computer Systems Forecast*, technical report, Dataquest Corporation, San Jose, CA, August 1993.
2. E. Harris, "Technology Directions for Low-Power Systems," presented at the IEEE VLSI Symposium on Low Power Electronics, Phoenix, AZ, August 25, 1993.
3. J. E. Bertsch, K. Bernstein, L. G. Heller, E. J. Nowak, and F. R. White, "Experimental 2.0 V Power/Performance Optimization of a 3.6 V-Design CMOS Microprocessor—PowerPC 601," *Proceedings of the IEEE Symposium on VLSI Technology*, 1994, pp. 83–84.
4. R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of Ion-Implanted MOSFET's with Very Small Dimensions," *IEEE J. Solid-State Circuits* **SC-9**, 256 (1974).
5. E. J. Nowak, "Ultimate CMOS ULSI Performance," *IEDM Tech. Digest*, pp. 115–118 (1993).
6. A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-Power CMOS Digital Design," *IEEE J. Solid-State Circuits* **27**, No. 4, 473–484 (April 1992).
7. D. Liu and C. Svensson, "Trading Speed for Low Power by Choice of Supply and Threshold Voltages," *IEEE J. Solid-State Circuits* **28**, No. 1, 10–17 (January 1993).
8. G. Kitsukawa, M. Horiguchi, Y. Kawajiri, T. Kawahara, T. Akiba, Y. Kawase, T. Tachibana, T. Sakai, M. Aoki, S. Shukuri, K. Sagara, R. Nagai, N. Hasegawa, N. Yokoyama, T. Kisu, H. Yamashita, T. Kure, and T. Nishida, "256 Mb DRAM Technologies for File Applications," *ISSCC Digest of Technical Papers*, pp. 48–49 (February 1993).
9. T. Kawahara, M. Horiguchi, Y. Kawajiri, G. Kitsukawa, T. Kure, and M. Aoki, "Subthreshold Current Reduction for Decoded-Driver by Self-Reverse Biasing," *IEEE J. Solid-State Circuits* **28**, No. 11, 1136–1144 (November 1993).
10. M. Horiguchi, T. Sakata, and K. Itoh, "Switched-Source-Impedance CMOS Circuit for Low Standby Subthreshold Current Giga-Scale LSI's," *IEEE J. Solid-State Circuits* **28**, No. 11, 1131–1135 (November 1993).
11. D. Bouldin, "Reliability Issues in Multilevel Interconnects," *1994 Multilevel Interconnection Seminar Proceedings*, Santa Clara, CA, June 1994, p. 5.
12. F. White, W. Hill, S. Eslinger, E. Payne, W. Cote, B. Chen, and K. Johnson, "Damascene Stud Local Interconnect in CMOS Technology," *IEDM Tech. Digest*, p. 301 (1992).
13. R. Uttecht and R. Geffken, "A Four-Level-Metal Fully Planarized Interconnect Technology for Dense High Performance Logic and SRAM Applications," *Proceedings of the VLSI Multilevel Interconnect Conference*, 1991, p. 20.
14. T. Y. Chan, P. Ko, and C. Hu, "Dependence of Channel Electric Field on Device Scaling," *Electron Device Lett.* **6**, No. 10, 551–553 (October 1985).
15. T. Sakata, M. Horiguchi, and K. Itoh, "Subthreshold-Current Reduction Circuits for Multi-Gigabit DRAM's," *IEEE Symposium on VLSI Circuits Digest of Technical Papers*, pp. 45–46 (May 1993).

Received May 6, 1994; accepted for publication November 4, 1994

Kerry Bernstein *IBM Microelectronics Division, Burlington facility, Essex Junction, Vermont 05452 (KERRY at BTVLABVM, kbernstein@vnet.ibm.com).* Mr. Bernstein is an advisory engineer in the IBM PowerPC Microprocessor Development group, currently responsible for PowerPC technology performance assessment and application. He previously provided technology design support for the IBM RS/6000 Workstation's microprocessor suite. Mr. Bernstein joined IBM in 1978 after receiving the B.S.E.E. degree from Washington University, St. Louis, Missouri, and continued with graduate work at the University of Vermont. He is an inventor of six U.S. patents and has co-authored papers on the design of Multiport Register File Architectures.

John E. Bertsch *IBM Microelectronics Division, Burlington facility, Essex Junction, Vermont 05452 (BERTSCH at BTVLABVM, bertsch@vnet.ibm.com).* Dr. Bertsch received the B.E.E. and M.Eng. degrees from the University of Detroit in 1975 and 1976, respectively, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, in 1983. He joined IBM in 1983 and has worked in the areas of device reliability, technology applications, and process development. At present he works in the area of device engineering, developing advanced CMOS technologies for logic, SRAMs, and microprocessors.

Lawrence G. Heller *IBM Microelectronics Division, Burlington facility, Essex Junction, Vermont 05452 (HELLER at BTVLABVM, heller@btvlabvm.vnet.ibm.com).* Dr. Heller is a senior engineer currently involved in CMOS circuits and technology for high-performance and low-power application, and Intellectual Assets technical support. He received the B.S.E.E. degree from the New Jersey Institute of Technology, Newark, in 1964 and the M.S. and Ph.D. degrees in electrical engineering from Iowa State University, Ames, in 1967 and 1969, respectively. Dr. Heller has published several papers, including work on CCD and Bucket Brigade Device modeling and charge-transfer sense amplifiers for DRAM; he is an author of the original paper on DCVS logic. He is an inventor of 23 issued U.S. patents. Dr. Heller is a Senior Member of the IEEE.

Edward J. Nowak *IBM Microelectronics Division, Burlington facility, Essex Junction, Vermont 05452 (A680005 at BTVLABVM, ejnowak@vnet.ibm.com).* Dr. Nowak received a B.S. degree in physics from M.I.T. in 1973, and M.S. and Ph.D. degrees in physics from the University of Maryland in 1974 and 1979, respectively. Following postdoctoral research at New York University, he joined IBM in 1981, working in MOSFET device design on the first IBM 1Mb DRAM. He currently works in the IBM Microelectronics Division, Essex Junction, Vermont, where he is engaged in CMOS logic process development in support of low-power/high-performance device design.

Francis R. White *IBM Microelectronics Division, Burlington facility, Essex Junction, Vermont 05452 (FWHITE at BTVVMOFS, fwhite@btvvmofs.vnet.ibm.com).* Mr. White received his B.S. and M.S. degrees in electrical engineering from the University of Maine at Orono in 1977 and 1978, respectively. He joined IBM in 1978 and has worked in the areas of dry etching of silicon compounds and process development. He has also held a management assignment and is a member of the American Vacuum Society, where he has been active in local chapter activities. At present he works in the area of process development for advanced CMOS technologies.