

Design in the Power-Limited Scaling Regime

Borivoje Nikolić, *Senior Member, IEEE*

(Invited Paper)

Abstract—Technology scaling has entered a new era, where chip performance is constrained by power dissipation. Power limits vary with the application domain; however, they dictate the choices of technology and architecture and necessitate implementation techniques that tradeoff performance for power savings. This paper examines technology options in the power-limited-scaling regime and reviews sensitivity-based analysis that can be used for the optimal selection of optimal architectures and circuit implementations to achieve the best performance under power constraints. These tradeoffs are examined in the context of power minimization at the technology, circuit, logic, and architecture levels, both at the design and run times.

Index Terms—CMOS, performance, power, technology scaling.

I. INTRODUCTION

TECHNOLOGY scaling reduces the minimum physical dimensions of transistors by a factor of $S = 0.7$ in each generation, and interconnect scaling follows a similar trend. Accordingly, the area needed to implement digital logic functions and memory has been reducing roughly by half with the introduction of each new technology node. In addition, scaled devices have simultaneously been increasing switching speed and lowering switching energy. The ideal scaling scenario proposed by Dennard *et al.* [1] requires that all voltages scale by the same factor of 0.7 in order to maintain constant fields. Consequently, switching energy per transistor has been scaling down by a factor of S^3 , resulting in constant power for a chip with the same area. A limitation of this scaling regime, as pointed out in the original paper, is that kT/q does not scale, resulting in nonscaling of device subthreshold characteristics. Ideal scaling does not account for gate tunneling currents, which are significant with very thin gate oxides.

Practical scaling has not always followed this ideal principle. Supply voltages were maintained at high levels of 12 V, and for an extended period of time at 5 V, to maintain compatibility of chip-to-chip interfaces. Supply-voltage scaling started at approximately the 0.5- μm technology node and, until very recently, has roughly followed the scaling of linear dimensions.

Manuscript received September 20, 2007. This work was supported in part by The National Science Foundation Infrastructure Grant 0403427, by the Center for Circuit & System Solutions (C2S2) Focus Center, by a Semiconductor Research Corporation program, and by an NSF CAREER Grant ECCS 0238572. The review of this paper was arranged by Editor T.-J. K. Liu.

The author is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720-1770 USA (e-mail: bora@eecs.berkeley.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TED.2007.911350

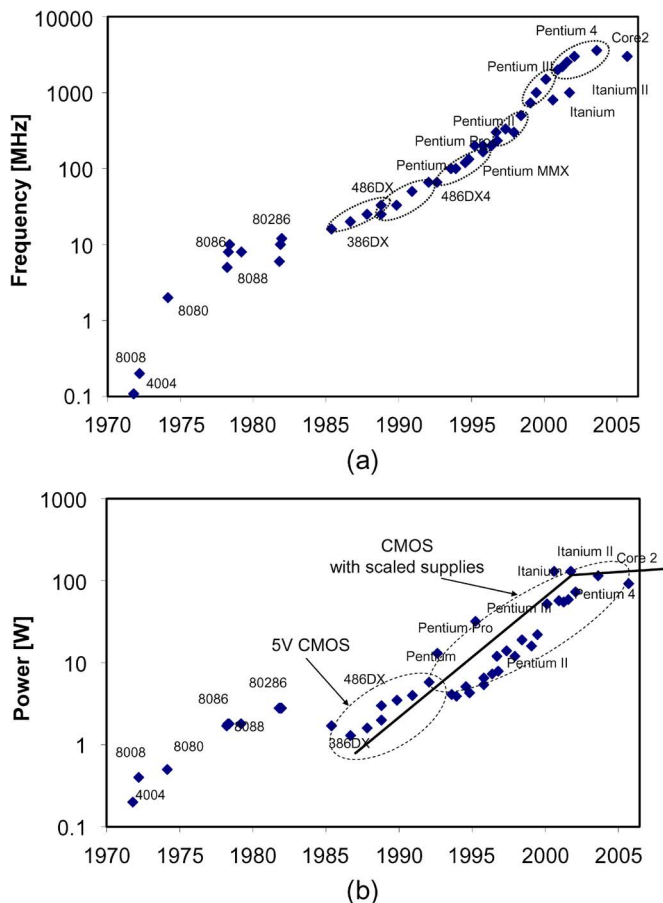


Fig. 1. (a) Frequency and (b) power trends in Intel's microprocessors.

However, designers and manufacturers have often used somewhat higher supply voltages above the ideal values of $V_{DD} = \text{feature size} \times 10 \text{ V}/\mu\text{m}$ to boost the performance within reliability constraints. In addition, chip dimensions have traditionally been increasing rather than staying constant. Specifically in microprocessor design, architectural changes have resulted in increased operating frequencies, beyond the gains achievable by technology scaling alone [2], [3]. Fig. 1(a), as an example, shows the frequency trends in Intel's lead microprocessors over time. All these factors have resulted in a rapid increase in power dissipation. Fig. 1(b) shows the increase in power dissipation in Intel's processors, and the data from other manufacturers follow a similar trend. Over the last 10 years, the power dissipation in the lead microprocessors has increased by a factor of 2.5 per generation, saturating at about 100-W levels. In high-performance (HP) applications, power dissipation is limited

by the practicality and the cost of cooling; in the case of microprocessors with forced-air cooling systems, this limit is in the 100–150-W range. Chips for portable applications often do not allow the use of fans and are limited to about 2 W of power with plastic packaging. Mobile applications are often limited by the battery life, which dictates constraints on both active and leakage powers during the standby and sleep modes. As a result, most of the designs today and all of the designs in the future are power-limited.

These trends in technology scaling have made power dissipation a primary design constraint for both HP and mobile applications. In contrast to the past, fitting within the power budget today is as important for the designers as is achieving maximum performance; instead of targeting the absolute maximum performance, the designers need to maximize the performance for the given power budget. There are many degrees of freedom for trading off performance and power in the design: They can be traded off at the technology selection stage, in circuit and logic design, and at the architecture optimization stage. Many of the decisions in system design are dependent on each other and can involve optimization of both discrete and continuous variables.

Device and Circuit Models: The energy and the delay of a logic gate are functions of its size, supply voltage, and transistor threshold voltage. It is important to model the current in the saturation and subthreshold regions to evaluate the performance of scaled circuits. NMOS current in the saturation region can be accurately modeled as a function of the gate–source and drain–source voltages [4], [5]

$$I_{\text{DSat}} = \frac{W}{L} \frac{\mu_{\text{eff}} C_{\text{ox}} E_C L}{2} \frac{(V_{\text{GS}} - V_{\text{Th}})^2}{(V_{\text{GS}} - V_{\text{Th}}) + E_C L} (1 + \lambda V_{\text{DS}}) \quad (1)$$

where μ_{eff} is the electron/hole mobility, C_{ox} is the gate capacitance per unit area, W and L are the width and the length of the channel, respectively, V_{Th} is the threshold voltage, E_C is a fit for the saturation field, and the parameter λ can be fitted to include the effect of drain-induced barrier lowering (DIBL), as well as of channel-length modulation.

Power dissipation in digital circuits, in general, has four components: active, leakage, short circuit, and biasing

$$P = P_{\text{sw}} + P_{\text{leak}} + P_{\text{sc}} + P_{\text{bias}}. \quad (2)$$

Power dissipation is dominated by the switching and leakage components. The short-current power is proportional to the switching power for well-designed circuits [6], and for the purpose of scaling, analysis can be treated together. Biasing currents often exist for generating reference voltages and currents in memory and I/O circuits, and some less-often-used logic families. Switching power is equal to the rate of energy exchange and is therefore a product of switching energy and the frequency of operation

$$P_{\text{sw}} = E_{\text{sw}} f = \alpha C V_{\text{DD}}^2 f \quad (3)$$

where E_{sw} is the switching energy, C is the total capacitance under consideration (can be a gate, logic path, or an entire chip), α is the switching activity, and f is the operating frequency.

The subthreshold current is a function of gate-to-source and drain-to-source voltages

$$I_{\text{subth}} = \mu_{\text{eff}} C_{\text{ox}} \frac{W}{L} V_t^2 \times \exp\left(\frac{V_{\text{GS}} - V_{\text{Th}}}{n V_t}\right) \left(1 - \exp\left(-\frac{V_{\text{DS}}}{V_t}\right)\right) \quad (4)$$

where $V_t = kT/q$ is the thermal voltage, and n is the subthreshold parameter [5]. The subthreshold slope S_{subth} is often defined as $S_{\text{subth}} = \ln 10 n V_t$.

Leakage power is the product of the supply voltage and leakage current I_{leak}

$$P_{\text{leak}} = V_{\text{DD}} I_{\text{leak}} \quad (5)$$

where I_{leak} corresponds to the subthreshold current (I_{subth}), with $V_{\text{GS}} = 0$

$$I_{\text{ds,leak}} = \mu_{\text{eff}} C_{\text{ox}} \frac{W}{L} (n-1) V_t^2 e^{-V_{\text{Th}}/V_t}. \quad (6)$$

In present technologies that use SiO_2 as the gate dielectric (90, 65, and some of 45 nm), there is a significant contribution of the gate leakage current, which has been increasing with thinning of the gate oxides. This trend will be temporarily discontinued with the introduction of high- k dielectrics in the 45-nm node [7], [8]. As a result, in the remainder of this paper, less attention will be paid in treating gate leakage currents.

Constant-Field Scaling: In the conventional scaling model, the device dimensions (W , L , and t_{ox}) scale by the factor of S . The doping concentration increases by the factor $1/S$ in order to scale down the junction depths, and voltages V_{DD} and V_{Th} scale down linearly with S .

The constant-field-scaling regime keeps the chip active power density constant by scaling the active power per device with a factor of S^2 . This model results in constant power dissipation for a fixed area in a scaled technology if the leakage power is negligible. In the early period of technology scaling, the leakage component of the overall power dissipation has been, indeed, small; however, scaling of the threshold voltage has been increasing the leakage exponentially.

Under the constant-field scaling by a factor of S , the leakage power for the chip with a constant area can be shown using (6) to scale with a factor of $(1/S^2) 10^{(V_{\text{Th}}(1-S)/S_{\text{subth}})}$. The relative increase in leakage current is dependent on the actual value of the threshold voltage. Threshold reduction by a factor of $S = 0.7$ increases the chip leakage power density by several orders of magnitude, with high values of threshold voltages (> 0.5 V). However, this traditionally did not affect the overall power consumption as the subthreshold leakage was a very small component of the total power, even smaller than reverse bias junction leakage currents. A continued exponential increase in leakage currents has brought them to a level where they significantly contribute to the overall power budget. In sub-100-nm processes, this increase in leakage is less than an order of magnitude with each technology generation (about 2–3 \times) since $S V_{\text{Th}} < S_{\text{subth}}$.

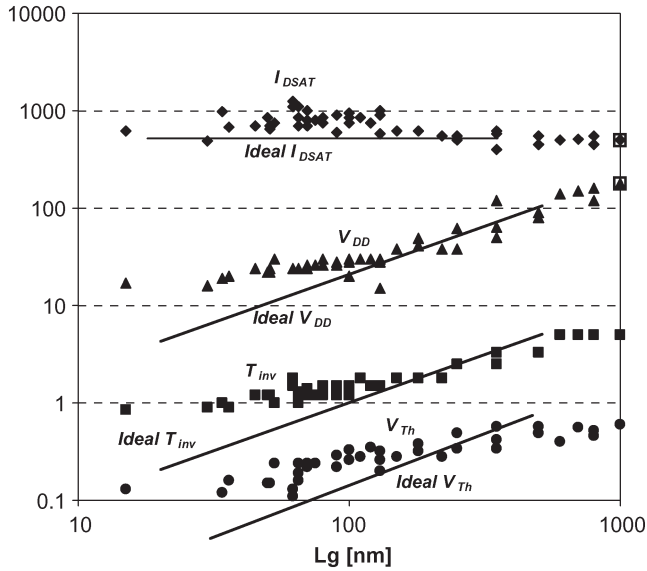


Fig. 2. Historic trends of scaling the saturation current, oxide thickness, supply voltage, and threshold voltage.

II. POWER-LIMITED SCALING

The fundamental limit to constant-field-scaling regime is related to the nonscaling of the subthreshold slope and increase of gate leakage as most of the other limiting factors are under designers' control (voltage, frequency, die size, and architecture) [1], [5].

Reducing the supply voltage significantly reduces the switching power, but lowers the device switching speeds because of lower saturation currents. It is necessary to scale the threshold voltages according to the constant-field model to maintain the performance. Threshold voltage reduction results in an exponential increase in transistor drain leakage currents, which represent a significant portion of the overall power budget today. With scaling of both the supply and threshold voltages, a minimum power is achieved when a balance is struck between the active and leakage power components. This optimum is at the point where leakage contributes to about 30%–40% of the total power during active operation of the circuit [9]–[11]. Many HP designs have reached this point around 130- or 90-nm technology nodes. As a result, continued scaling in the 90-, 65-, and 45-nm nodes and beyond departs from the constant-field model and enters the power-limited-scaling regime. The continued scaling of technology outlined by ITRS still introduces new devices with lower thresholds [12]. The power-limited-scaling regime is characterized by the use of multiple devices in the design optimized for different performance/power targets, together with slowed-down supply- and threshold-voltage scaling, and dramatic changes in chip architectures.

Recent Scaling Trends: Although the technology scaling from the 0.5- μm down to the 0.13- μm technology has involved the reduction in both device dimensions and voltages, it has not been closely following the ideal constant-field-scaling rules. Practical scaling data are plotted against the ideal requirements in Fig. 2. Both the supply voltage and the transistor thresholds have been scaling with feature sizes, but have been generally

falling behind the ideal values. Particularly, threshold scaling has recently been further slowed down, resulting in reduced V_{DD}/V_{Th} ratios. On the other hand, shortened channel lengths and mobility enhancement techniques, such as the use of strained silicon, have increased transistor saturation currents. Similarly, the gate capacitances have been decreasing, instead of staying constant. Both of these trends have been contributing to improvements in transistor switching speeds, despite the slowdown in threshold scaling.

Present technology scaling is characterized by the availability of multiple devices, as outlined by the ITRS. Fig. 3(a) and (b) shows the trends in ON-currents I_{on} , drain-to-source leakage currents I_{off} , gate tunneling currents I_g , oxide thicknesses t_{ox} , supply voltage V_{DD} , and corresponding fan-out-of-four inverter delays FO4 for some of the process options for one foundry. In current technologies, generally, a choice of one of two oxide thicknesses is available for chip core implementation; the thinner oxide is used for HP applications, and the thicker oxide is used in the applications that require lower leakages. This second option is often denoted as “low power” (LP), which may include both the low operating and low standby powers [12]. Within each of the two process options, two, out of two or three, offered threshold voltages are available for implementation. The gate oxide thickness for the high-speed process option (Fig. 2) generally follows the historic trend of scaling by about 20%–25% per technology generation, down to the 45-nm node. Oxide thickness scaling has been slowed down with increased tunneling currents; further scaling of the effective oxide thickness will require advances in the high- k dielectrics. Scaling of the thicker core oxide follows the same trend, lagging by about 1.5 technology generations. In the HP process options, threshold voltages continue to scale, resulting in a continued OFF-current increase in lead, standard-threshold devices by a factor of 2–2.5 per technology generation. In contrast, in the LP process option, threshold voltages are held approximately constant to maintain the battery life requirements in mobile devices. The 65- and 45-nm processes offer a wide variety of devices whose ON-currents with nominal supplies can vary by a factor of four, OFF-currents can vary by three orders of magnitude, and FO4 delays could vary by a factor of three. This variety of process options opens up a possibility for power-performance optimization at circuit and architecture levels using a number of different design variables. By simply mapping a design into a different technology option, large tradeoffs in power performance are possible. For example, up to $3\times$ in delay can be traded off for three orders of magnitude of leakage savings.

Power-Performance Optimization: Performance of an integrated system can be evaluated at different levels of design abstraction. At the system level, benchmark scores for typical application scenarios are often used as a measure of performance, while at the circuit level, the delay is usually the only performance measure. In general, most integrated circuit designs are constrained either by the throughput or the latency, or both, where the throughput is defined as the amount of data processed in a given time, and the latency is the time (often expressed in the number of clock cycles) needed for a certain data input to pass from the input to the output of a system.

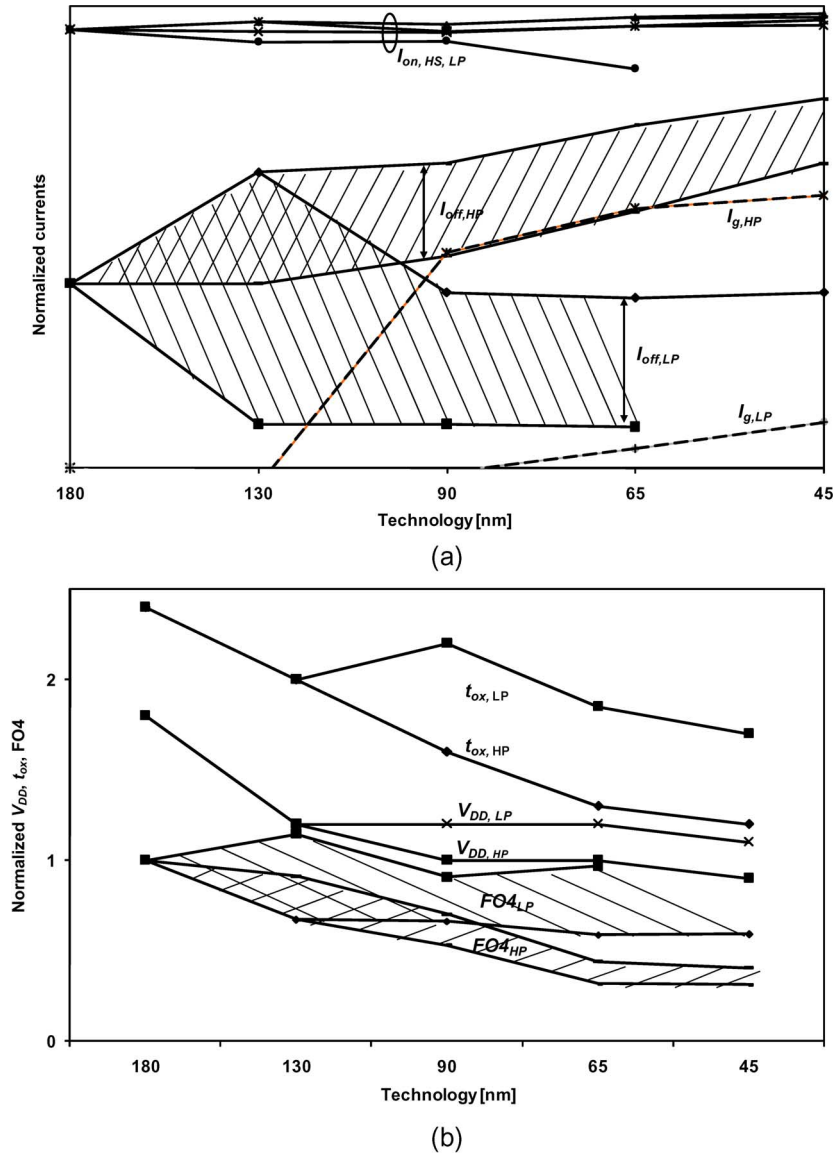


Fig. 3. (a) Trends in ON-currents I_{on} , drain-to-source leakage currents I_{off} , and gate tunneling currents I_g for foundry deep submicrometer processes. (b) Trends in oxide thicknesses t_{ox} , supply voltage V_{DD} , and fan-out-of-four inverter delays FO4. HP represents an HP process option, and LP represents an LP process option.

Similarly, different systems have different priorities for the active and leakage energy consumptions.

Methods for achieving maximum performance have been well explored at all levels of design abstraction. Striking a balance between the energy and performance of a design has been a recent research topic as well. Initially, an optimal system that balances both the energy and performance has been searched for through minimization of objective functions that combine energy and performance. For example, minimizing the energy-delay product at the circuit level [9], [10] results in a particular design point in the energy-delay space, where 1% of energy can be traded off for 1% of delay. Although this composite metric can be used for evaluating performance of different implementations of one function, it may not correspond to an optimum design under desired operating conditions. The $E \cdot D^2$ metric [13] puts more weight on the delay than the energy, and since it is V_{DD} invariant, it presents a good optimization

target for systems that operate with varying supply voltages. It is possible to generalize metrics to the form of $E^m \cdot D^n$; however, designing a system for any particular m and n has limited applicability since it gives only one (E, D) pair in the energy-delay space at which the delay D is minimized for a fixed energy E . Maximizing the performance under energy constraints can be formulated as a constrained optimization problem and has been studied more recently [14], [15]. The system can be optimized to maximize the performance under energy constraints or to minimize the energy under performance constraints. In our recent work [15], sensitivities have been used to formalize the tradeoff between energy and performance. Sensitivity has been defined as the absolute gradient of energy to delay with respect to a change in a particular design variable.

There are several design variables that can be tuned to tradeoff energy for performance at various levels of design

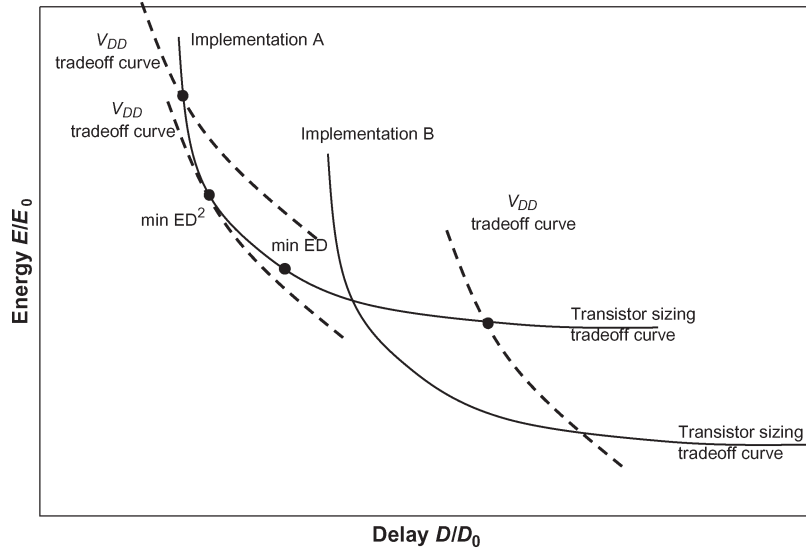


Fig. 4. Illustration of the energy-delay tradeoff curves.

hierarchy. The tradeoff achieved by adjusting a design variable x is given by the energy/delay sensitivity to variable x

$$S_x(X) = \frac{\partial E / \partial x}{\partial D / \partial x} \Big|_{x=X}. \quad (7)$$

This quantity represents the amount of energy that can be traded for delay by tuning variable x around the design point X . The energy-efficient design is achieved when the relative sensitivities to all the tuning variables are balanced [14], [15].

Fig. 4 shows some of the tradeoffs that can be made at the circuit level. In general, the energy-delay tradeoff curve obtained by continuously changing one design variable is convex. An example of such a curve is shown in Fig. 4 for a complex combinatorial function (such as 64-bit addition). A continuous curve can be achieved by optimally changing transistor widths to maximize the performance under varying energy constraints. A different implementation of such a function (e.g., using a different logic design for an adder or a different circuit style) would result in different curves (implementations A and B in Fig. 4). Each point on these curves corresponds to a different design, with different transistor sizes. The slope of the curves changes in each point, and the curves may or may not intersect. The curve that is closer to the coordinate origin is more energy efficient. Dashed lines in the figure illustrate the energy-delay tradeoffs with changing V_{DD} for a fixed sizing. The optimal design is achieved when sensitivity to sizing is equal to the sensitivity to V_{DD} for a particular energy or delay constraint, as is the case for the design that corresponds to ED^2 minimum.

Tradeoff variables differ in each abstraction level. Design variables accessible to the circuit designer include transistor sizing and choice of supply and threshold voltages. Logic designers and architects can alter the logic design, logic depth, pipeline depth, parallelism, etc., to achieve these tradeoffs. The variables can be continuous, such as supply voltage, or discrete, such as the threshold selection or architecture choice.

Optimal Supplies and Thresholds: By relying on the fact that the optimal E_{sw}/E_{Lk} is around two, optimal V_{DD} and V_{Th} can

be determined for a function block [15]. This is a result that fundamentally affects the scaling in the power-limited regime. Supply and threshold voltages are not independent variables as they were in the constant-field regime, and their values are set by the power optimality of the implemented function.

Variability: Increased variability in process technology greatly affects the power-performance optimality of a design. The effective feature sizes, film thicknesses, and doping concentrations vary from their nominal values, affecting the transistor performance and interconnect parameters. The use of subwavelength lithography, coupled with quantum effects in devices, has resulted in the increased process variability in scaled technologies, where feature tolerances do not track scaling of the median values. The transistor performance, for example, is affected by the variations in gate lengths, oxide thicknesses, and random dopant fluctuations that, in turn, reflects in varying chip performance and power [16]. If the power-performance optimality of a design is achieved with the switching/leakage energy ratio of two for typical process parameters, variations can make the actual chip suboptimal since the leakage current is strongly dependent on the gate lengths and threshold voltages. Variations can be classified as within-die or die-to-die (which include wafer-to-wafer and lot-to-lot variations) [17]. Die-to-die variations are often treated as median shifts in parameter values between the dies, where all the transistors within the die vary by the same average amount. Within-die variations can have different radii of spatial correlation. In the two extreme cases, there could be no spatial correlation between the devices (e.g., due to random dopant fluctuations), or all the devices within the die would be correlated (e.g., due to a change in gate length or film thickness). These two scenarios differently affect the chip performance: Uncorrelated variations in the logic path delays result in the reduction of the relative path delay variation through averaging, where using longer paths can be beneficial; however, with correlated variations ratio of the standard deviation over the average delay does not change [18]. The ratio between the switching and leakage energies varies greatly with the process.

Estimating the leakage energy based on the process corners only would result in wide upper and lower bounds, separated by more than an order of magnitude. To correctly center the design, it is necessary to include the estimation of distributions of both die-to-die and within-die parameter variations [19].

Relative contributions of variability components and their distributions and correlations are generally unknown in advance. In a 90-nm process, lithography-induced channel-length variations dominate other sources; this, in turn, affects logic delays and dramatically changes transistor leakage currents [20]. Reduction in device dimensions, along with the use of restricted design rules to control lithography-induced variations, will increase the relative contribution of random dopant fluctuations because of smaller total dopant counts and can likely become the dominant variability component in small devices [21]. While these variations will be averaged out in long logic paths, they may jeopardize the future of six-transistor SRAM.

To achieve a high product yield, designers add design margins to both power and performance to accommodate the maximum process spread. These margins account for the worst-case process and do not distinguish between the die-to-die and within-die contributions. As a result, the design is often pessimistic, with a large power penalty being paid to meet the performance goals under the worst-case conditions.

In addition to process-induced variability, chip's operation conditions and environment may vary. Changing input activity changes the numbers of inactive logic gates, thus changing the ratio of switching to leakage energy. Temperature during operation changes dramatically in some application areas, slowing down circuit performance and increasing leakage. Supply voltage noise may change with the environment as well, affecting the margins in the design.

III. EXAMPLES OF ENERGY-PERFORMANCE TRADEOFFS AT CIRCUIT AND ARCHITECTURE LEVELS

Techniques for trading off energy and performance can be broadly classified based on the time of their enable time and the targeted energy components [22].

- 1) Enable time: some of the energy-performance tradeoffs can be implemented (or enabled) only at the design time, such as transistor sizing or the logic depth. On the other hand, supply and threshold voltages can be either fixed during the design phase or varied during run time. In the sleep mode, the performance does not matter (energy-delay sensitivity is zero), and the design variables are adjusted to minimize the energy.
- 2) Energy components: design techniques can primarily address either dynamic (switching) or static (leakage) component of energy dissipation and trade them off for performance. Lowering the supply voltage reduces both switching and leakage energies, while the adjustments in the threshold voltage primarily address the leakage energy.

Numerous energy-saving techniques have been proposed in the past. Usually, they are initially presented with the sole

purpose of energy reduction, without much discussion on the impact on performance. For the purpose of this analysis, they are classified in the following three categories.

- 1) Win-win techniques. The best techniques simultaneously increase the performance and lower energy. These techniques are generally based on a major architectural or algorithmic transformation and are often specific to a particular application.
- 2) Zero or near-zero performance penalty. There are many techniques that can lower energy with no impact on performance. These techniques eliminate excess energy in the system, generally consumed in noncritical or redundant operations. Examples include the energy-delay tradeoffs in noncritical paths and clock gating techniques. Some of the techniques that reduce energy in variable-throughput applications are implemented with a very small overall performance cost.
- 3) True energy-performance tradeoffs. Majority of well-known LP techniques reduce the energy dissipation at the expense of a lower system performance. Examples of this include simply slowing down the circuits through downsizing, reducing the supply, or increasing the threshold. Most techniques require an implementation overhead that has to be accounted for in the performance and energy budget.

Several known energy-saving techniques are evaluated in this section, according to the classification in Table I, based on the enable time and performance impact, by roughly evaluating their energy-performance sensitivities.

A. Reducing Switching Activity at the Design Time

There is a large body of work on reducing switching activity or total switched capacitance of circuits. Most notably, techniques include avoidance of unnecessary switching in time-multiplexed resources, exploiting correlation in signal processing data, algorithmic transformation, number representations, and reduction of bus activities [23]–[25]. Most of these techniques fall into categories 1) and 2). A particular example applies in reducing the switching activity of clocks through clock gating, which is a technique that deterministically turns off the clock to units or blocks that are not being active for multiple clock cycles. These techniques generally fall in the second category as their application does not incur any performance penalty, while the range of actual energy savings varies with the application. One exception to this rule is the use of fine-granularity clock gating, where the clock is automatically gated by the data for every flip-flop (or a small group of flip-flops). This is implemented by comparing the new input datum to a flip-flop to its output; if they are the same, the clock edge is being absorbed [26]. This technique, because of its fine granularity, has both the energy and the performance overhead, but can yield overall energy savings when selectively applied [27]. Clock gating and many of other switching activity reduction techniques are well supported by the CAD tools.

TABLE I
SUMMARY OF ENERGY-REDUCTION TECHNIQUES, CLASSIFIED ACCORDING TO THEIR ENABLE TIME AND IMPACT ON PERFORMANCE

Enable time/ Performance impact	Design time	Run time
Near zero performance penalty	Clock gating Architectural switching reduction Multi- V_{Th}	Dynamic V_{DD} scaling Dynamic V_{Th} scaling
True energy- performance tradeoffs	Fine-granularity clock gating V_{DD} , V_{Th} adjustments Multi- V_{DD} Logic styles Sizing (L , W) Stack forcing	Sleep transistors

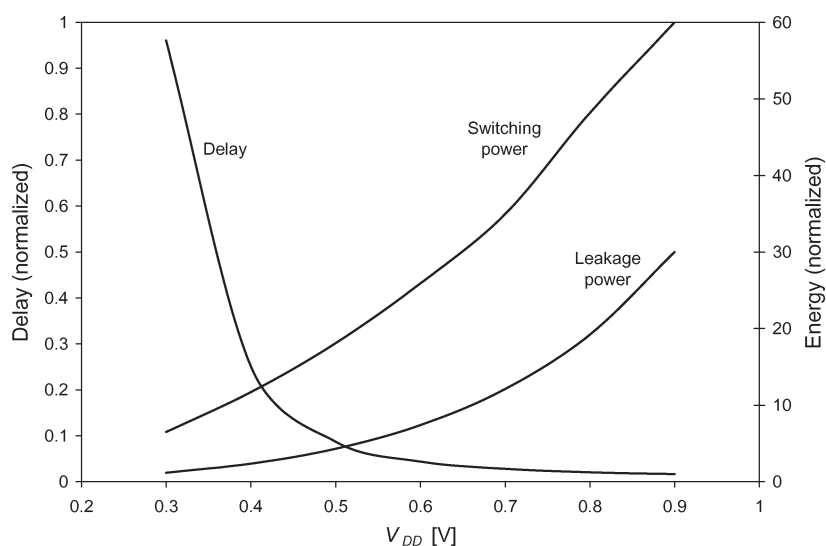


Fig. 5. Switching energy, leakage energy, and propagation delay dependence on the supply voltage, normalized to the values at nominal supply. Leakage power is normalized to be one half of the switching power at the nominal supply.

B. Supply Voltage Selection at the Design Time

Lowering the Supply Voltage: Switching energy is a quadratic function of the supply voltage, and the transistor leakage exponentially depends on it because of the DIBL. Overall, when V_{DD} is close to nominal, from (1) and (3), 1% of supply voltage reduction saves about 2% of energy while increasing the delay by 1%. With significantly lower supplies than typical, the relative delay penalty increases, and the delay variability due to process-induced threshold variations becomes large. As shown in Fig. 5, on a typical 45-nm process, the leakage energy savings are larger than the active energy savings because of the exponential dependence of the leakage current on the supply voltage.

A small adjustment of the core supply voltages has traditionally been used by microprocessor manufacturers during speed binning. With the maximum power being limited by the cooling, the supply voltage is the simplest knob to use to achieve the maximum clock frequency within the power limit for the bin. Chandrakasan *et al.* [25] outlined dramatic energy savings achievable in throughput-constrained applications

through supply voltage reductions, where the throughput was maintained by using hardware concurrence.

Subthreshold Design: Extreme scaling of the supply voltage below the transistor thresholds drives devices into the subthreshold region, resulting in a very low energy of operation [28]. This results in several orders of magnitude of increase in delay compared to operation at the nominal supplies. The supply voltage that is below the transistor threshold is an optimum for certain applications with low throughputs per block.

Using Multiple Supplies: Using multiple supply voltages for different functional units on a chip at the design time is a design technique to lower the power, with minimum impact on overall performance. Performance-critical blocks (or robustness-critical blocks, such as SRAM) are supplied from the higher supply voltage (V_{DDH}), and the less-critical blocks are supplied from a lower supply voltage (V_{DDL}). Less-critical blocks are often designed not to (or to minimally) affect the system benchmarks by maintaining either the required system throughput or latency. To support the use of multiple voltages, a notion of “voltage domains” (or “voltage islands”) is being used, where each domain needs to be on a separate power grid

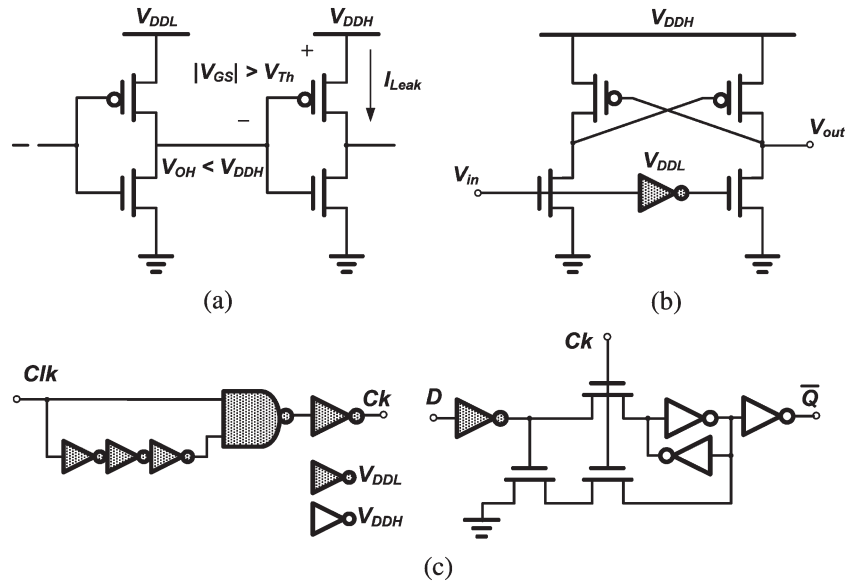


Fig. 6. Level shifting from V_{DDL} to V_{DDH} . (a) Inverter-based. (b) Cross-coupled level converter. (c) Level-converting flip-flop.

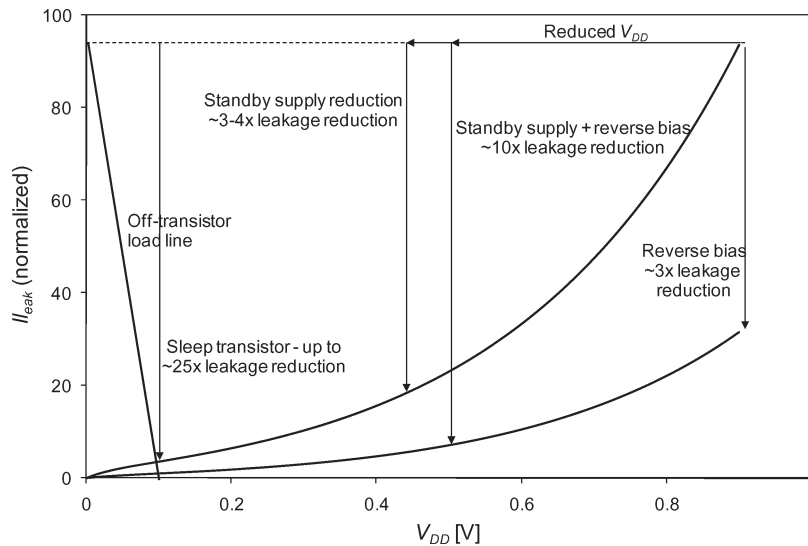


Fig. 7. Normalized leakage current reduction in a general-purpose 45-nm technology. Sleep transistors reduce leakage by 25 \times , reverse bias reduces leakage by 3 \times , lower standby supply reduces leakage by 3–4 \times , and combined lower supply and reverse bias reduces leakage by 10 \times .

[29]. In addition, voltage levels of signals that are crossing the domain boundaries have to be translated. A signal crossing over from a high-voltage domain into a low-voltage island does not require level conversion. However, when a signal from the low-voltage domain needs to drive a gate placed in the high-voltage domain, the pull-down networks in the CMOS gates have a lower overdrive, while the PMOS transistors are not completely turned off, causing increased leakage in that gate [Fig. 6(a)]. The signal level needs to be converted to the high-voltage level, which can be done using regenerative gates, such as a cross-coupled buffer, shown in Fig. 6(b); their added delay is the fundamental performance penalty for using the second supply. Because of the nature of the level converter design, it manifests higher sensitivity of the delay to process and voltage variations than the V_{DDL} gates. Since the level conversion can be conveniently performed both logically and physically at the input of the V_{DDL} block, it can be associated

with the timing boundary and incorporated into a flip-flop, and an example of such a design is shown in Fig. 6(c) [30]. The use of voltage domains is well supported in the contemporary design flows. Practical limitations in generating and distributing supply voltages often limit the number of voltages to two.

By taking the idea of using multiple supplies further, it can be envisioned to use them inside a combinational logic block [31]. To implement this idea, level conversion is necessary for every crossing of a signal from V_{DDL} to V_{DDH} . To simplify the implementation and to minimize its delay sensitivity to disturbances, the idea of clustered voltage scaling can be used, where only one transition from V_{DDH} to V_{DDL} is allowed for each logic path, and the level conversion, if necessary, is performed in the flip-flops. This requires V_{DDH} gates to be clustered in the beginning of each path and simplifies the supply assignment task for each gate. However, the use of two supply voltages, whether implemented in custom or standard-cell

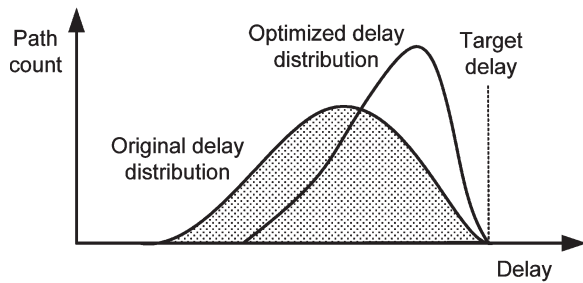


Fig. 8. Impact of a slow down of noncritical paths on the delay distribution of a logic block.

design methodology, requires changes in placement of gates. Using two power rails within the cell is one way of distributing two supplies. Although this has been demonstrated in a custom datapath [32], it is impractical for standard cell designs because of the need for well spacing, increased cell height, and blockages introduced by using low metal layers (M2) for power routing. A practical solution for distribution of power supplies in dual supply blocks is by alternating V_{DDH} and V_{DDL} rows [31]. Overall, the dual supply design saves 15%–35% of energy, excluding the cost of supply generation with minimal delay penalty [30]. Although attractive, these savings are not significant enough to promote the commercial use of this technique because of added design complexity, lack of CAD support, and decreased robustness.

C. Dynamic Energy Management

Dynamic Voltage Scaling (DVS): Most real-time systems are designed to meet maximum throughput requirements in the worst case, but their actual computing requirements greatly vary in time. A general-purpose processor to be used in portable applications, such as notebook computers, electronic organizers, and mobile phones, executes computational functions that fall into three major categories: compute-intensive tasks, low-throughput/high-latency functions, and idle-mode operation. Compute-intensive and short-latency tasks need the full computational throughput of the processor to achieve real-time constraints. MPEG video compression and decompression, or interactive gaming, are examples of such. Low-throughput and long-latency tasks, such as text processing or data entry, operate under far more relaxed completion deadlines and require only a fraction of the maximum processor's throughput. There is no reward for finishing the computation early, and if a task is completed early, it can be considered as a waste of energy. Finally, portable processors spend a large fraction of their time in the idle mode, waiting for a user action or an external event.

Even compute-intensive operations, such as MPEG decoding, show variable computational requirements while processing a typical stream of data. For example, the number of times an MPEG decoder computes an inverse discrete cosine transform per video frame varies widely depending upon the amount of motion in the video scenes [34].

By simultaneously lowering the operating frequency and the supply voltage while executing low-throughput/high-latency load, switching energy is quadratically lowered. In order to maintain the required throughput for high workloads and min-

imize energy for low workloads, both supply and frequency must be dynamically varied according to the requirements application that is currently being executed. This technique is called the DVS and is implemented such that the function is operated at the lowest supply voltage that meets the timing constraints [33]–[35]. For successful implementation of this technique, it is necessary to monitor the variation of the critical path delay with supply voltage. Various CMOS circuit delays generally track each other with scaling of the supply voltage [33]. In the simplest implementation, the delay is monitored through a simple ring oscillator that presents a replica of the critical path, with a sufficient margin [33]. Circuits exhibit relative delay variations depending if they are gate-capacitance dominated, diffusion dominated, and wire dominated. To lower the design margin in monitoring, it is necessary to observe replicas of different critical paths, which might end up reordered in criticality with supply and process variations [37].

A practical DVS system, shown in Fig. 9, consists of the following components:

- 1) a circuit block that can operate under a wide variety of supply voltages;
- 2) a supply-regulation loop that sets the minimum voltage necessary for operation at a desired frequency;
- 3) an operating system and scheduler that calculates the required frequencies to meet requested throughputs and task completion deadlines.

A monitored delay is included in the power-supply control loop to provide the translation between the supply voltage and the clock frequency. The operating system digitally sets the desired operating cycle time, and the current value of the delay monitor is compared against it. The difference is used as a feedback error. By adjusting the supply voltage, the supply voltage loop changes the delay to set this error value to zero, incurring no performance penalty. A small performance hit may exist during the transition from one mode of operation to another as it may take several clock cycles to ramp up the supply voltage.

Standby Power Management: Many systems spend most of the time in the standby mode, waiting for an external or internal event to respond to. In order to save energy, it is desirable to operate at the minimum possible supply voltage in the standby. Turning off the supply is not always permissible in the system. The state of the block is erased when the supply is OFF, and restarting the system may take a long time. Furthermore, it is difficult to implement a perfect electronic in-line CMOS switch, which has zero ON-resistance and OFF-current. In standard technologies, this switch is implemented as a thick-oxide longer-channel-length NMOS or PMOS transistor, with associated design tradeoffs. Using higher thresholds and longer channel lengths suppresses the leakage currents through the switch [36]. To further suppress leakage currents, when a negative supply is available, it can be applied to the gate of the switch. The switch needs to have minimal series resistance; this can be accomplished by increasing its size, lowering the threshold, or increasing the gate voltage if the higher voltage is available. The switch resistance is usually targeted to minimize the voltage noise on

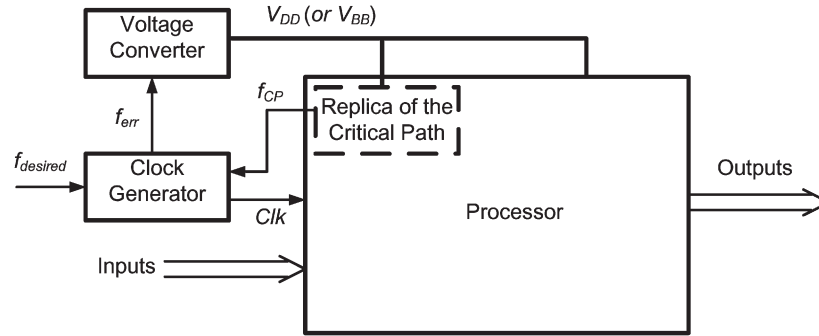


Fig. 9. Variable throughput system. A varying desired operating frequency dictates the required supply voltage or body bias by observing the critical path.

the virtual supply rail. A common design point is to size the switch such that it results in less than 5% of a worst-case voltage spike at the power rails [38]. This, in turn, increases the circuit delay by less than 5%. There is an option of using a decoupling capacitance on the virtual rail, which can absorb some switching noise; the tradeoff is that the activation/deactivation time of the block increases to accommodate charging/discharging this capacitance [39]. The use of a power switch presents a power-performance tradeoff: Performance is degraded by up to 5%, while the leakage power is reduced by an order of magnitude. Entering and leaving the standby mode comes with an energy penalty as well, associated with charging the rail, well, decoupling, and switch capacitances. If the system spends less than about 100 cycles in the sleep mode, entering and leaving the sleep mode yields the overall energy loss [39].

In contrast, when opting not to use power switches, it is necessary to determine the minimum supply for which the state is retained. Among the state elements, flip-flops retain the data generally at lower supply voltages than the SRAM. Conventional six-transistor SRAM exhibits wide variations in the minimum retention voltage, making it difficult to monitor its data retention voltage on the fly [40]. As a result, minimum SRAM voltage needs to be preset during the design, and a sufficient margin needs to be added [37]. The leakage power saving (6–8 \times), is smaller than when using a power switch with an added cost of voltage regulator, but with virtually no performance penalty during the active mode. Fig. 7 illustrates leakage current reduction in a 45 nm process when using sleep transistor, reduced supply voltage and reverse bias.

D. Threshold Voltage Selection and Adjustment

Scaled process technologies offer a thicker core gate oxide, in addition to thick I/O devices, which reduce the gate tunneling currents by two to three orders of magnitude, at the expense of reduced ON-current by about 20%, as shown in Fig. 3(a). For each core oxide, there are often two threshold voltages introduced by ion implantation. The two core thresholds are targeted to be spaced by about 60–100 mV apart, offering about 5–20 \times reduction in drain leakage currents, with about 15%–25% penalty in ON-currents, with corresponding delay increase [Fig. 3(a)]. Migrating an HP leaky design (where leakage is 30% of the total power) from a low-threshold to a high-threshold HP process would reduce the energy consumption by 30%, with 20% delay penalty, offering 1.5% for 1% energy-

delay tradeoff. Although this mapping is certainly a possibility, more commonly, both thresholds are used simultaneously [41]. Low-threshold gates are assigned to critical paths, and high-threshold gates are assigned to less-critical paths to reduce or eliminate the timing slack. Unlike in the case of the dual supply design, any gate can be assigned either of the thresholds. If the timing slack for a path is larger than 20%, all gates can have a high threshold; if it is less than 20%, the assignment process becomes a discrete optimization, with a goal to achieve maximum leakage savings. After the assignment of dual thresholds, in a general case of a typical timing slack distribution, more than 90% of the gates will have a high threshold (Fig. 8).

The threshold voltage can be adjusted through back biasing in bulk CMOS technologies, both statically and dynamically [42]. Static back bias can be used in compensating the process-induced variations, while the dynamic threshold control can be used for energy savings in the variable throughput applications, using a similar concept as shown in Fig. 9. Scaling of the technology has been reducing the body effect, thus limiting the range of threshold adjustments. Furthermore, gate-induced drain leakage limits the range of reverse bias voltage for which the leakage is being reduced [43]. Reverse biasing also increases the device variability, and forward body bias improves it [44]. The range of forward bias is limited by the conduction of drain/source-body p-n junctions to about 0.6 V.

E. Transistor Sizing

At the circuit level, transistor sizing is an effective tool for trading off circuit delay and power. Delay is minimized by using transistors with minimum channel lengths and appropriate widths [22]. Sizing for the minimum delay results in high energy consumption. By departing from the optimum sizing for delay, both switching and leakage energies can be reduced. Examples include adjusting the transistor widths (gate sizes), transistor lengths, and transistor width ratios. Downsizing gates that are off the critical path does not impact the block timing and linearly reduces both the switching and leakage energies. With tightening of the power constraints, automated synthesis tools are being more diligent about reducing the off-critical-path gate sizes. Downsizing the gates that are on the critical path results in a real tradeoff between the energy and the performance. Near the minimum achievable delay point, a large amount of energy can be saved with small impact on performance. In common circuit blocks, the range of performance achievable through

sizing is about $2\times$, where the fastest circuit has about a half of the delay of a minimum-sized one, while the range of energy could vary widely. The sensitivity varies from a very large value down to zero [15].

Increasing the transistor channel lengths reduces the leakage, but increases the gate capacitance, which, in turn, increases switching power and impacts the delay. In a general-purpose 45-nm process, with standard device thresholds, increasing the gate length from the minimum by 5 nm (which is a common manufacturing grid in sub-100-nm technologies) reduces the leakage current by $4\text{--}5\times$ across process corners. If the W/L ratio is preserved, the drive current is reduced by 3.6%, and the input capacitance is increased quadratically (26%), thus slowing down the preceding gate. When placed in a critical path and normalized, the transistor length increase of about 10% decreases the leakage by 4000%, increases the delay by about 30%, and increases the active power by about 25%. The overall effectiveness of this technique largely depends on the fraction of the leakage in the original design. For designs where the leakage contributes more than 25% of the overall power budget, there is an overall win in power savings. For example, if a design dissipates 30% of leakage power, the net balance is that 1% of delay penalty saves about 0.7% of the energy. Assuming that the gate is off the critical path and that can be slowed down just by increasing the transistor length or decreasing the gate width, adjusting gate lengths becomes more effective when leakage contributes more than 25% of the total power.

Leakage reduction is very steep near the minimum channel length, potentially offering an attractive technique to achieve large leakage energy savings with linear switching power increase, by slightly increasing transistor lengths of gates that are off the critical path [45]. This cannot be done in the design phase as the transistor lengths are usually restricted to be placed on a manufacturing grid that is 5 nm or more. However, smaller channel-length biases (less than 10%) can be made directly in the mask layouts, making this approach possible [46].

F. Transistor Stacks

Stacking transistors is another way of trading off active power and delay for leakage power savings [47], besides the use in the sleep mode. For example, a stack of two OFF devices in 45-nm CMOS leaks about $25\times$ less than a single device. Adding the third OFF-transistor to the stack reduces the leakage by another factor of two, resulting in a net leakage that is $50\times$ smaller than the leakage of a single transistor.

Transistor stacks occur naturally in CMOS logic—a two-input NAND gate has a stack of two NMOS transistors. To reduce the leakage in the sleep mode, both transistors in the stack should be OFF; however, bringing the logic block into a state where all the appropriate signals are set for minimum leakage may take several clock cycles.

Stacks can be forced when needed by adding another transistor in the stack [47]. Adding an NMOS transistor to the stack in the inverter reduces the leakage by $25\times$, and without increasing the transistor width, it doubles the input capacitance

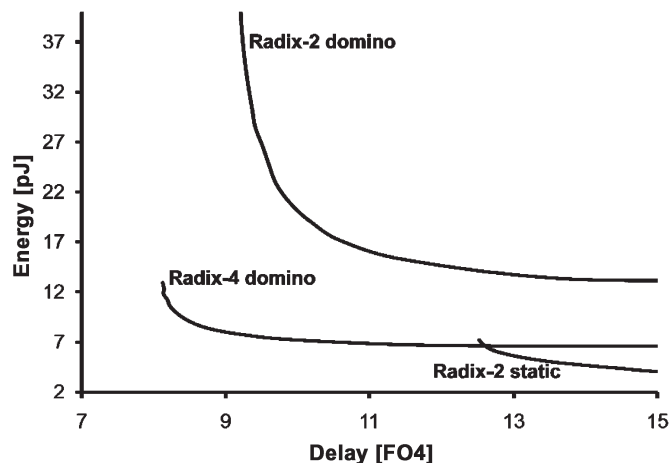


Fig. 10. Energy delay for various implementations of carry-lookahead adders. Included are static and domino implementations, and radix-2 and radix-4 logic designs.

and reduces the drive current by 30%. As a result, this technique trades off the leakage reduction for a 30% increase in switching energy and delay. This tradeoff yields energy savings for low activity gates that are not on the critical path.

G. Logic Styles

Implementing the same logic function using alternate logic styles will result in different performance and power tradeoffs [24], [48]. Power limitations, design complexities, and concerns for robustness in deeply scaled technologies have narrowed down the choices of circuit design styles. High switching factors and high power dissipation have eliminated differential logic families from datapaths. Similarly, high design costs associated with assuring the required robustness of pass transistor logic styles have limited their use. Today's designs are dominated by the standard static CMOS, which is inherently robust and can be synthesized. Only where the extreme performance is needed and when the power budgets allow for it, custom domino logic paths are being used. Each design needs to be optimally mapped into a circuit implementation, which can be done for some classes of logic blocks, for example adders [49]. Fig. 10 shows energy-delay tradeoffs for optimized 64-bit carry-lookahead adders. Radix-2 design is easily mapped into static CMOS and has LP, but limited speed. Mapping the same radix-2 design into domino logic (where the dynamic logic stage is followed by an inverter) reduces the delay with added power penalty. Changing the architecture to a radix-4 or modifying the circuit implementation to compound domino (where the dynamic logic gate is followed by a static gate) can further improve the delay and can lower power.

H. Architecture

Power-limited design requires rethinking of digital architectures. At the architecture level, different metrics are used in evaluating the performance and energy. In microprocessors, performance is evaluated through the number of instructions or operations per second, graphics processors are evaluated,

e.g., by the number of vertices processed per second, and signal processors have their own application-dependent aggregate performance metrics. Finally, the system performance is measured by a common benchmark score.

Energy and performance can be effectively traded off at the system level by increasing the level of concurrency in the execution, which involves tradeoffs in the number of functional units and their organization. Concurrency can also be increased by increasing the depth of pipelining or parallelism within a functional unit. Increasing concurrency often results in increased chip area, thus increasing the system cost. To establish the relationship between the performance and energy, synthetic metrics can be used, such as energy and area efficiencies, defined as the number of operations per unit of energy and the unit of area, respectively.

A classical example of an interaction of a higher level architectural design variable, such as the degree of parallelism or pipelining, with a technology design variable, such as a supply voltage, has been studied in [25]. By using parallelism or deeper pipelining, a datapath can have the same throughput at a lower supply voltage, reducing the active power dissipation. With optimized transistor thresholds and supplies, parallelism and pipelining reduce the total power. It is important to point out that, because of a larger number of inactive circuits in parallel datapaths, optimal thresholds in parallel datapaths are higher than in pipelined. The ultimate benefit of increased concurrency implemented through parallelism or pipelining is limited by practical constraints, like the minimum practical supply voltage or logic depth, and on the range of energy-delay tradeoffs for implementing the desired function.

Architectural changes can have a very large impact on power dissipation of a system, and power-conscious rearchitecting of a system might be able to improve the performance. Microprocessors in the past have been increasing the pipeline depth while reducing the logic depth to about 10–15 to increase the performance through increasing the clock frequency, at a dramatic cost in power. Power-limited designs prefer somewhat deeper logic. In an example study, it has been shown that the logic depth of 22 FO4 inverters is optimal for PowerPC architecture [50]. Practical trends in mainstream microprocessors have followed this trend as well. Intel has backed off from the Pentium 4 architecture and opted for longer logic paths in multicore processors. Modern processors utilize multiple processor cores and larger cache memories, operating at lower supplies.

IV. CONCLUSION

Today, power limitations are as important in the design as is the performance. Design in the power-limited-scaling regime requires continuous changes in the architectures, circuit implementation, and technology choices to maximize the performance under power constraints. Many techniques for lowering power consumption are well known, but their implementation often incurs a performance penalty. An optimum implementation is achieved when the energy/delay sensitivity of the design is equal for all the design and technology variables. Implementation of LP techniques increases the design and

verification complexity and often requires special technology features, which increases the design cost. Ultimately, scaling will end when the increase in the design cost stops being manageable.

ITRS recognizes the emerging device structures as possible candidates in replacing planar CMOS device structures. Double-gate transistors offer better OFF-currents, and if the velocity enhancement techniques are applied, it offers better ON-currents as well. Their application in circuit design favorably shifts the basic energy-performance tradeoff curves (but does not fundamentally alter them), which would provide a one-time performance boost/energy reduction at the point of their introduction.

ACKNOWLEDGMENT

The author would like to thank the students, faculty, and sponsors of the Berkeley Wireless Research Center for their contributions and D. Markovic, V. Stojanovic, and R. Zlatanovici, in collaboration with M. Horowitz and R. Brodersen, for the sensitivity-based optimization.

REFERENCES

- [1] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SSC-9, no. 5, pp. 256–268, Oct. 1974.
- [2] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul./Aug. 1999.
- [3] P. P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunities and the new frontiers," in *Proc. ISSCC Dig. Tech. Papers*, San Francisco, CA, Feb. 5–7, 2001, pp. 22–25.
- [4] K. Toh, P. Ko, and R. Meyer, "An engineering model for short-channel MOS devices," *IEEE J. Solid-State Circuits*, vol. 23, no. 4, pp. 950–958, Aug. 1988.
- [5] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
- [6] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. SSC-19, no. 4, pp. 468–473, Aug. 1984.
- [7] *Intel's Transistor Technology Breakthrough Represents Biggest Change to Computer Chips in 40 Years*, Jan. 2007. Intel Corp., press release. [Online]. Available: <http://www.intel.com/pressroom/archive/releases/20070128comp.htm>
- [8] *IBM Advancement to Spawn New Generation of Chips*, Jan. 2007. IBM Corp., press release. [Online]. Available: <http://www-03.ibm.com/press/us/en/pressrelease/20980.wss>
- [9] J. Burr and A. M. Peterson, "Ultra low power CMOS technology," in *Proc. NASA VLSI Des. Symp.*, Oct. 1991, pp. 4.2.1–4.2.13.
- [10] R. Gonzalez, B. Gordon, and M. A. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, Aug. 1997.
- [11] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for low-power and high-speed applications," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2000, pp. 469–474.
- [12] *International Technology Roadmap for Semiconductors*, 2006. [Online]. Available: <http://public.itrs.net>
- [13] A. J. Martin, "Towards an energy complexity of computation," *Inf. Process. Lett.*, vol. 77, no. 2–4, pp. 181–187, Feb. 2001.
- [14] V. Zyuban *et al.*, "Integrated analysis of power and performance for pipelined microprocessors," *IEEE Trans. Comput.*, vol. 53, no. 8, pp. 1004–1016, Aug. 2004.
- [15] D. Marković, V. Stojanović, B. Nikolić, M. A. Horowitz, and R. W. Brodersen, "Methods for true energy-performance optimization," *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1282–1293, Aug. 2004.
- [16] K. Bernstein *et al.*, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM J. Res. Develop.*, vol. 50, no. 4/5, pp. 433–450, Jul.–Sep. 2006.

- [17] J. W. Tschanz *et al.*, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.
- [18] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, Feb. 2002.
- [19] S. Narendra, V. De, S. Borkar, D. A. Antoniadis, and A. P. Chandrakasan, "Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18- μm CMOS," *IEEE J. Solid-State Circuits*, vol. 39, no. 3, p. 510, Mar. 2004.
- [20] L. T. Pang and B. Nikolić, "Impact of layout on 90 nm CMOS process parameter fluctuations," in *Proc. Symp. VLSI Circuits, Dig. Tech. Papers*, Honolulu, HI, Jun. 15–17, 2006, pp. 84–85.
- [21] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proc. IEEE*, vol. 89, no. 3, pp. 259–288, Mar. 2001.
- [22] J. M. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital Integrated Circuits: A Design Perspective*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2003.
- [23] A. P. Chandrakasan, M. Potkonjak, R. Mehra, J. Rabaey, and R. W. Brodersen, "Optimizing power using transformations," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 14, no. 1, pp. 12–31, Jan. 1995.
- [24] A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE*, vol. 83, no. 4, pp. 498–523, Apr. 1995.
- [25] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, Apr. 1992.
- [26] M. Hamada *et al.*, "Flip-flop selection technique for power-delay trade-off," in *Proc. IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, San Francisco, CA, Feb. 15–17, 1999, pp. 270–271.
- [27] D. Marković, B. Nikolić, and R. W. Brodersen, "Analysis and design of low-energy flip-flops," in *Proc. ACM/IEEE ISLPED*, Huntington Beach, CA, Aug. 6–7, 2001, pp. 52–55.
- [28] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778–1786, Sep. 2005.
- [29] D. E. Lackey, P. S. Zuchowski, T. R. Bednar, D. W. Stout, S. W. Gould, and J. M. Cohn, "Managing power and performance for system-on-chip designs using voltage islands," in *Proc. IEEE/ACM ICCAD*, San Jose, CA, Nov. 10–14, 2002, pp. 195–202.
- [30] F. Ishihara, F. Sheikh, and B. Nikolić, "Level-conversion for dual-supply systems," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 2, pp. 185–195, Feb. 2004.
- [31] K. Usami and M. Horowitz, "Clustered voltage scaling technique for low-power design," in *Proc. Int. Symp. Low Power Des.*, Dana Point, CA, Apr. 23–26, 1995, pp. 3–8.
- [32] Y. Shimazaki, R. Zlatanovici, and B. Nikolić, "A shared-well dual-supply-voltage 64-bit ALU," *IEEE J. Solid-State Circuits*, vol. 39, no. 3, pp. 494–500, Mar. 2004.
- [33] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Brodersen, "A dynamic voltage scaled microprocessor system," *IEEE J. Solid-State Circuits*, vol. 35, no. 11, pp. 1571–1580, Nov. 2000.
- [34] V. Gutnik and P. Chandrakasan, "Embedded power supply for low-power DSP," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 5, no. 4, pp. 425–435, Dec. 1997.
- [35] T. Kuroda *et al.*, "Variable supply-voltage scheme for low-power high-speed CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 33, no. 3, pp. 454–462, Mar. 1998.
- [36] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS," *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, Aug. 1995.
- [37] H. Mair, "A 65-nm mobile multimedia applications processor with an adaptive power management scheme to compensate for variations," in *Proc. Symp. VLSI Circuits*, Honolulu, HI, Jun. 2007, pp. 224–225.
- [38] J. Kao, A. Chandrakasan, and D. Antoniadis, "Transistor sizing issues and tool for multi-threshold CMOS technology," in *Proc. 34th Des. Autom. Conf.*, Jun. 1997, pp. 409–414.
- [39] J. W. Tschanz, S. G. Narendra, Y. Ye, B. A. Bloechel, S. Borkar, and V. De, "Dynamic sleep transistor and body bias for active leakage power control of microprocessors," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1838–1845, Nov. 2003.
- [40] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Proc. 5th Int. Symp. Quality Electron. Des.*, 2004, pp. 55–60.
- [41] L. Wei, Z. Chen, K. Roy, Y. Ye, and V. De, "Mixed- V_{th} (MVT) CMOS circuit design methodology for low power applications," in *Proc. 36th Des. Autom. Conf.*, Jun. 1999, pp. 430–435.
- [42] T. Kuroda *et al.*, "A 0.9-V, 150-MHz, 10-mW, 4 mm², 2-D discrete cosine transform core processor with variable threshold-voltage (VT) scheme," *IEEE J. Solid-State Circuits*, vol. 31, no. 11, pp. 1770–1779, Nov. 1996.
- [43] A. Keshavarzi, S. Narendra, S. Borkar, C. Hawkins, K. Roy, and V. De, "Technology scaling behavior of optimum reverse body bias for standby leakage power reduction in CMOS IC's," in *Proc. Int. Symp. Low Power Electron. Des.*, Aug. 1999, pp. 252–254.
- [44] J. Tschanz *et al.*, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.
- [45] S. Rusu *et al.*, "A 65-nm dual-core multithreaded Xeon processor with 16-MB L3 cache," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 17–25, Jan. 2007.
- [46] P. Gupta, A. B. Kahng, P. Sharma, and D. Sylvester, "Gate-length biasing for runtime-leakage control," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 8, pp. 1475–1485, Aug. 2006.
- [47] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan, "Scaling of stack effect and its application for leakage reduction," in *Proc. Int. Symp. Low Power Electron. Des.*, Aug. 2001, pp. 195–200.
- [48] S. Kosonocky *et al.*, "Low power circuits and technology for wireless digital systems," *IBM J. Res. Develop.*, vol. 47, no. 2/3, pp. 283–298, Mar.–May 2003.
- [49] S. Kao, R. Zlatanovici, and B. Nikolić, "A 250 ps 64-bit carry-lookahead adder in 90 nm CMOS," in *Proc. IEEE ISSCC, Dig. Tech. Papers*, San Francisco, CA, Feb. 4–8, 2006, pp. 438–439.
- [50] V. Srinivasan *et al.*, "Optimizing pipelines for power and performance," in *Proc. 35th Annu. IEEE/ACM Int. Symp. Microarchitecture, (MICRO)*, Nov. 2002, pp. 333–344.



Borivoje Nikolić (S'93–M'99–SM'06) received the Dipl. Ing. and M.Sc. degrees in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 1992 and 1994, respectively, and the Ph.D. degree from the University of California at Davis, Davis, in 1999.

He was with the faculty of the University of Belgrade from 1992 to 1996. He spent two years with Silicon Systems, Inc., Texas Instruments Storage Products Group, San Jose, CA, working on disk-drive signal processing electronics. In 1999, he joined the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, where he is now an Associate Professor. His research activities include high-speed and low-power digital integrated circuits and VLSI implementation of communications and signal processing algorithms. He is the coauthor of *Digital Integrated Circuits: A Design Perspective*, 2nd ed. (Prentice-Hall, 2003).

Dr. Nikolić received the IBM Faculty Partnership Award in 2005–2007, NSF CAREER award in 2003, College of Engineering Best Doctoral Dissertation Prize, and Anil K. Jain Prize for the Best Doctoral Dissertation in Electrical and Computer Engineering, University of California at Davis, in 1999, as well as the City of Belgrade Award for the Best Diploma Thesis in 1992. For works with his students and colleagues, he received the Best Paper Award at the ACM/IEEE International Symposium of Low-Power Electronics in 2005 and the 2004 Jack Kilby Award for the Outstanding Student Paper at the IEEE International Solid-State Circuits Conference.