INVITED PAPER

Practical Strategies for Power-Efficient Computing Technologies

An eightfold improvement in power efficiency can be achieved without loss of performance for modestly parallelizable CMOS-based computer systems.

By Leland Chang, David J. Frank, *Fellow IEEE*, Robert K. Montoye, *Senior Member IEEE*, Steven J. Koester, *Senior Member IEEE*, Brian L. Ji, Paul W. Coteus, *Senior Member IEEE*, Robert H. Dennard, *Fellow IEEE*, and Wilfried Haensch, *Senior Member IEEE*

ABSTRACT | After decades of continuous scaling, further advancement of silicon microelectronics across the entire spectrum of computing applications is today limited by power dissipation. While the trade-off between power and performance is well-recognized, most recent studies focus on the extreme ends of this balance. By concentrating instead on an intermediate range, an $\sim 8\times$ improvement in power efficiency can be attained without system performance loss in parallelizable applications—those in which such efficiency is most critical. It is argued that power-efficient hardware is fundamentally limited by voltage scaling, which can be achieved only by blurring the boundaries between devices, circuits, and systems and cannot be realized by addressing any one area alone. By simultaneously considering all three perspectives, the major issues involved in improving power efficiency in light of performance and area constraints are identified. Solutions for the critical elements of a practical computing system are discussed, including the underlying logic device, associated cache memory, off-chip interconnect, and power delivery system. The IBM Blue Gene system is then presented as a case study to exemplify several proposed directions. Going forward, further power reduction may demand radical changes in device technologies and computer architecture; hence, a few such promising methods are briefly considered.

KEYWORDS | Circuit optimization; CMOS digital integrated circuits; CMOSFETs; integrated circuit design; integrated circuit interconnections; parallel machines; power distribution

I. INTRODUCTION

For several decades, semiconductor technology scaling has enabled manufacturers to produce integrated circuits with ever-increasing levels of performance and functionality—yielding a sustained exponential improvement in cost-per-function and a growing ubiquity of microelectronics in our daily lives. In accordance with trends predicted by Moore [1] and scaling rules set forth by Dennard and coworkers [2], the scaling of silicon complementary metal-oxide-semiconductor (CMOS) technology dimensions has led to simultaneous improvements in performance, density, and power dissipation for digital computing applications. In recent years, however, fundamental physical limitations have caused CMOS scaling to deviate from this path, and, in the interest of maintaining speed and density improvements, power dissipation has become a growing concern. Today, power is already a constraint across all applications [3]—from handheld consumer electronics to high-end servers; projecting scaling trends forward, the issue only becomes more severe. While general strategies to reduce power have been studied for some time [4]–[6], CMOS technologies in the 45 nm node and beyond exacerbate the challenge of power efficiency and require that novel solutions be developed. In addition, widespread acceptance of parallelism in today's computing architectures

Manuscript received August 27, 2009; accepted October 8, 2009. Current version published January 20, 2010. This work was partially supported by DARPA funding under the STEEP program (AFRL contract FA8650-08-C-7806).

L. Chang, **D. J. Frank**, **R. K. Montoye**, **S. J. Koester**, **P. W. Coteus**, **R. H. Dennard**, and **W. Haensch** are with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: lelandc@us.ibm.com; djf@us.ibm.com; montoye@us.ibm.com; skoester@us.ibm.com; coteus@us.ibm.com; dennard@us.ibm.com; whaensch@us.ibm.com).

B. L. Ji was with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA. He is now at 205 Windsor Rd, Fishkill, NY 12524 USA (e-mail: ji2020@gmail.com).

Digital Object Identifier: 10.1109/JPROC.2009.2035451

[7] creates new opportunities for power/performance optimization. In this new landscape, the benefits and trade-offs of potential techniques must be holistically assessed from a perspective that combines technology, circuits, and systems.

Because computing applications span a wide range of power and performance targets as well as activity factors, the term "low power" can suggest many different meanings. This paper will concentrate on low power as it pertains to the active mode of operation, which is important in mainstream computing. This shifts the focus away from low-activity-factor applications dominated by standby power, which range from sensor networks to other portable applications that require only a minimum amount of compute capacity. Many viable techniques to mitigate standby power are known and can effectively achieve low power in such applications, including power gating [8]—reducing the applied voltage across predetermined circuit blocks when idle—and simple adjustment of transistor threshold voltages and gate dielectric thicknesses. Instead, the aim of this paper is to address the more fundamental issue of reducing dissipation in the active mode, which is particularly relevant to applications with high performance requirements and a high activity factor—ranging from compute-intensive wireless consumer electronics to wired high-end server mainframes. While these end products may still present a wide range of performance and power targets, the basic issues that must be solved are shared.

In the analysis to be presented, performance, power, and area are considered to be system-level metrics. The system, as shown in Fig. 1, is assumed to consist of core logic and its associated cache memory, off-chip main memory, and a power delivery system. While such elements conceptualize the primary components of a high-end server, most other applications can be thought to be similarly organized. It is important to note that the total power in such systems includes significant contributions from many sources—not simply the processor itself [9]. Fig. 2 shows that while the specific breakdown varies between different end-user applications, each component must be addressed in order to achieve truly low-power computing.

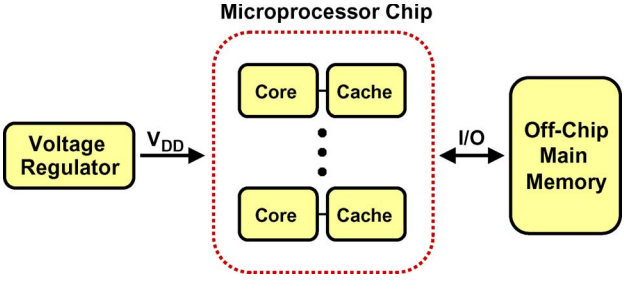


Fig. 1. *Simplified depiction of the exemplary system considered here. This basic system organization is representative of a broad range of end-user applications—from wireless consumer electronics to wired high-end servers.*

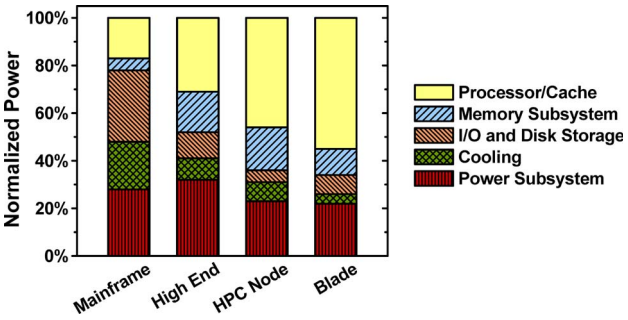


Fig. 2. *Power breakdown for several exemplary server systems. Many different components each contribute a significant amount to total power and thus must all be addressed to achieve a low-power computing technology. Adapted from Rajamani et al. [9].*

An important assumption in this paper is that the majority of applications for which power efficiency is critical can be effectively parallelized to increase system-level performance in the range of interest. The low-power techniques described herein demand an increase in parallelism to compensate for a reduction in operating frequency; however, only a modest amount of additional parallelism is necessary, which minimizes the impact of area and implementation overhead. While it is recognized that systems generally also require single-thread performance for certain applications, it is assumed that a single high-performance core can either be added in parallel to create a heterogeneous system or dynamically created by increasing the voltage of a low-power core. As such, power efficiency considerations for single-thread applications are not focused upon here.

This paper presents the current best understanding of the most effective methods by which power-efficient technologies can be improved under practical performance targets. In considering hardware applications that are ultimately limited by active power dissipation, the base postulate is that power efficiency is fundamentally rooted in voltage scaling [4]. While new algorithms and architectures will no doubt also play a strong role in future systems, the discussion here focuses on the efficiency of the underlying hardware technology as organized today. Voltage scaling in and of itself is not the goal; instead, it is asserted that the ability to reduce the operating voltage throughout the system is the foundation on which a power-efficient technology is based. This paper concentrates on techniques to facilitate and enable low-voltage operation, but other closely related or entirely revolutionary methods of improving power efficiency are also considered. In all proposed solutions, the achievement of optimal performance at optimal power efficiency requires simultaneously solving issues from the technology, circuits, and systems perspectives.

Section II of this paper analyzes in detail the motivation for voltage scaling by performing a comprehensive technology optimization for power efficiency. Starting with current CMOS technologies, a power reduction of $\sim 8\times$ can likely still be achieved by moderate voltage scaling into the 0.5 V range if performance is held constant with parallelism. This voltage range provides a significant improvement over the $\sim$1 V technologies utilized today while avoiding the considerable challenges that arise when the voltage is further reduced into the subthreshold regime. Section III then methodically proposes solutions to the challenges faced in voltage scaling into the 0.5 V range, including those associated with device scaling, low-voltage caches, on-chip digital noise, power delivery, and off-chip connections. In particular, alternate strategies for transistor optimization as well as technology and circuit techniques to enable cache functionality and robust power delivery are presented. In accordance with voltage scaling, these approaches enhance power efficiency in the overall system. Section IV presents a practical case study of the IBM Blue Gene system and its future directions, which is a focal point for the concepts presented in this paper. The prospects for more drastic power reduction using novel technologies and architectures are then described in Section V. Finally, the paper closes with a discussion of caveats in Section VI and a conclusion in Section VII.

II. THE CASE FOR VOLTAGE SCALING

A. MOSFET Scaling Theory

In the early days of CMOS, voltage reduction occurred in conjunction with technology scaling. During this time, the scaling of MOS field effect transistors (MOSFETs) largely followed the theory outlined in Table 1, which was originally proposed in [2]. By applying a suitable scale factor to each technology parameter, constant electric fields can be maintained throughout the device as it shrinks in size. Such a strategy preserves robustness to short-channel effects and device reliability, but more importantly results in improvements in circuit delay without increasing power density. While many significant advances in transistor technology have been made through the years, the basic structure has not changed significantly and these scaling guidelines, first proposed over 35 years ago, are still relevant today.

Table 1 Scaling Theory to Maintain Constant Electric Fields in a MOSFET Device. κ is a Dimensionless Scale Factor. From Dennard *et al.* [2]

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\varepsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

κ is the dimensional scale factor

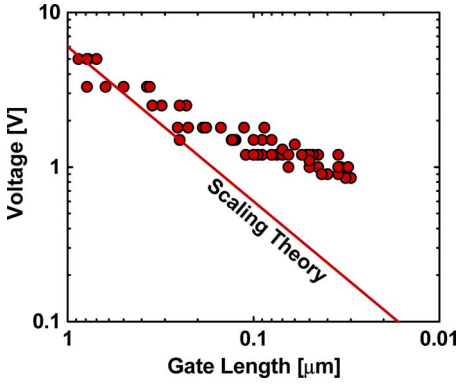


Fig. 3. *Scaling trend for power supply voltages in modern CMOS technologies. Due to leakage and variability constraints, voltage levels have deviated significantly from constant field scaling theory [2]. Adapted from Nowak [11].*

As CMOS technologies entered the submicrometer regime, several fundamental forces led to the modification of these scaling rules [10]. In particular, due to nonscalability of the threshold voltage and underlying limits on the subthreshold slope, supply voltage scaling slowed and in recent years has essentially come to a halt to control leakage power while maintaining device performance. Difficulties in scaling the gate dielectric thickness have also contributed to this trend. In addition, as manufacturing variability has a mounting influence on device characteristics, it has been prudent in some cases to raise voltages as a precaution to preserve operating margins. As a consequence, as shown in Fig. 3, the supply voltage in

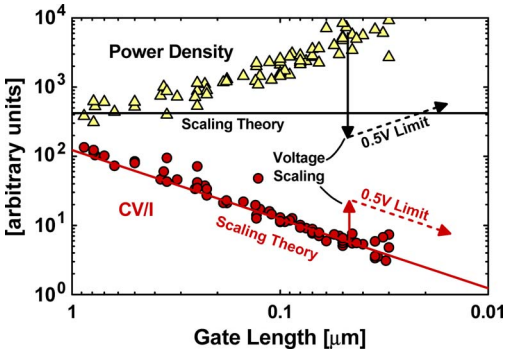


Fig. 4. *To maintain performance (CV/I) trends, voltage scaling has slowed, which results in dramatic increases in power density. A one-time voltage reduction can improve power efficiency for parallelizable applications without system performance degradation, but power densities may increase in future technologies due to limitations in further voltage scaling. Adapted from Nowak [11].*

modern technologies is significantly higher than originally suggested by scaling theory [11]. Fig. 4 emphasizes that performance has been of paramount importance—that the preferred scaling strategy has been to match these targets, which has directly led to dramatic increases in power density. Today, power places severe constraints on technology scaling, and dissipation levels are often raised to the brink of application-dependent cooling limits [3]. As illustrated in Fig. 4, an improvement in power efficiency can be attained via voltage scaling to ~0.5 V at a moderate cost in operating frequency, which can be compensated for by parallelism. While a gradual lowering of operating voltages may also be an acceptable course, this paper hypothesizes an aggressive one-time reduction, which reveals the full potential of voltage scaling. It is believed that ~0.5 V is a practical point of best reward for traditional CMOS technologies, beyond which conventional scaling limits may persist to limit further voltage scaling, inevitably reinstating power concerns in the future.

To first order, power dissipation in the active mode can be expressed as

$$P_{\text{active}} = C_{\text{eff}} V^2 f + I_{\text{leak}} V \tag{1}$$

where C_{eff} is the total effective load capacitance of a chip, V is the operating voltage, f is the operating frequency, and I_{leak} is the total aggregate leakage current of active devices when not being switched. The first term is the dynamic power dissipation due to switching, while the second term is the power consumed by leakage. Since C_{eff} is weakly dependent on voltage, the combined effective voltage dependence of $C_{\text{eff}}V^2$ has an exponent closer to 2.5 [12]. Empirically, it has been observed that the maximum operating frequency for a wide variety of circuits is a linear function of voltage in the regime of interest. An expression for frequency can thus be written as

$$f = \alpha(V - V_0) \tag{2}$$

where V_0 is the voltage at which frequency approaches zero (~0.25 V for modern technologies) and α is a constant that depends on the circuit. This same relation also applies to circuits that are optimized at each voltage, but with somewhat higher V_0 (~0.3–0.4 V) since low-voltage technologies generally optimize to higher threshold voltages. Putting these equations together yields

$$P_{\text{active}} = \alpha C_{\text{eff}} V^2 (V - V_0) + I_{\text{leak}} V. \tag{3}$$

While I_{leak} has a strong dependence on voltage, design optimization tends to maintain a consistent ratio between switching and leakage dissipation such that the overall voltage dependence of P_{active} is roughly cubic. Operating voltage is thus clearly the most effective parameter through which power dissipation can be improved. A reduction in voltage, however, limits operating frequencies and inevitably degrades the performance of a given circuit. In accordance with current trends [7], system-level performance can be regained by adding more parallel circuit blocks, which linearly adds to power dissipation. Since the super-linear improvements in power due to voltage scaling outweigh the linear increase in power due to parallelism, the end system can see substantial gains in power efficiency.

B. Full Technology Optimization

Power and performance trade-offs can be more accurately assessed using a power-constrained technology optimization program based on [13], which is conceptually depicted in Fig. 5. This program employs a large number of simple models spanning the device, circuit, and chip levels to estimate the performance of a multiprocessor chip based on underlying technology parameters such as gate length and gate dielectric thickness. Using detailed device models calibrated to 2-D TCAD simulations, these parameters are optimized to obtain the maximum chip performance subject to various power constraints. This program focuses on the active logic circuits of a processor chip, whereas the silicon area and power associated with memory, clock, and I/O portions of the chip are estimated by simple scaling of the logic circuit results as referenced to existing chip

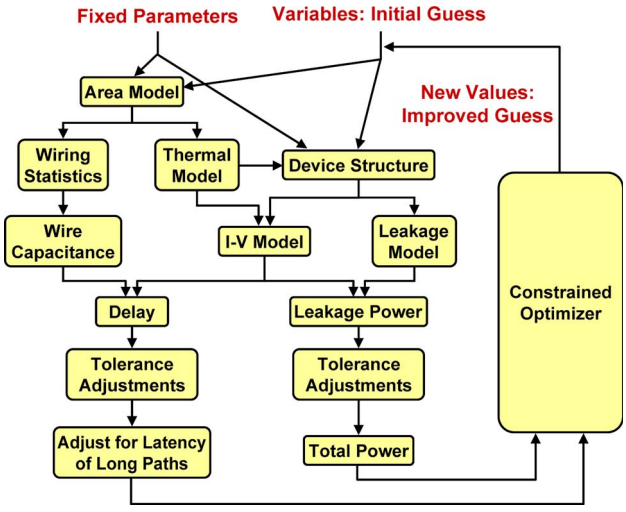


Fig. 5. *Flow diagram depicting operation of an optimization program that can be used to assess technology options using system-level metrics and constraints. By combining a large number of simple models spanning the device, circuit, and chip levels, the performance of a multiprocessor chip can be estimated based on underlying technology parameters.*

designs. Both global and local process variations are accounted for using a most probable worst case vector analysis [14], which is performed separately for delay and power so that worst case power can be constrained at worst case delay conditions. Table 2 summarizes the primary assumptions in device parameters and process tolerances used in this work. The quantitative analysis presented here focuses on 22 nm bulk MOSFET technologies, but similar optimization results are obtained for both future nodes and alternate device structures.

Since this program captures the effects of dynamic switching energy, leakage currents, velocity saturation, chip area, wire length and resistance, parasitic capacitance, variability, and digital noise, it can properly assess the complex trade-offs involved in choosing technology parameters. In considering system-level performance [characterized by millions of instructions per second (MIPS)], power dissipation, and area for various applications, the key trade-off at play in the pursuit of low-power computing is the balance between power efficiency and area efficiency. Power efficiency can be characterized by power/MIPS, which is a metric related to energy per operation. Area efficiency can be quantified as area/MIPS, which essentially describes performance in a parallel system since increased area due to parallelism is used to compensate for a reduction in single-processor performance. Fig. 6 shows the results from the optimizer for these two measures for a representative high-performance processor utilizing 22 nm and 11 nm node bulk MOSFET technologies. Each of the points on this plot represents an optimized chip/technology design in which the performance has been maximized subject to a different total chip power constraint. The variables relative to which performance has been maximized are: gate length, gate dielectric thickness, n- and p-FET body doping (to set the threshold voltages), supply voltage, average device width, and repeater width and spacing. As can be seen, power efficiency improves (power/MIPS decreases) as the voltage is lowered, but since frequency is reduced, area efficiency is degraded (area/MIPS increases). This penalty, however, does not increase dramatically until the voltage scales to below 0.5 V, which can be thought of as an optimum point for power efficiency without significant penalty in area efficiency. As compared with conventional ~1 V technologies, Fig. 6 shows that 0.5 V operation can improve power efficiency by ~8× at a performance penalty of ~4×—values that are somewhat larger than suggested by (2) and (3) due to complete optimization of all technology parameters. The proper optimum between these two metrics depends on the relative importance of power and area constraints, which implies that the ultimate limits in power efficiency will vary for different applications and no doubt be influenced by concerns such as cost and form factor.

As shown in Fig. 7, it can be useful to plot the two metrics in Fig. 6 as a single parametric curve with V_{DD} as an implicit parameter to clearly emphasize the trade-off between area and power efficiency. The point at which $V_{DD} = 0.5$ V lies approximately at the knee of the curve, which, depending on application specifications, is consistent with the optimal balance between power and performance suggested in [15]. Comparing the two curves in Fig. 7, it can be seen that migration from the 22 nm technology node to the 11 nm technology node brings a ~2× improvement while maintaining a consistent trade-off between area and power efficiency at $V_{DD} \sim 0.5$ V.

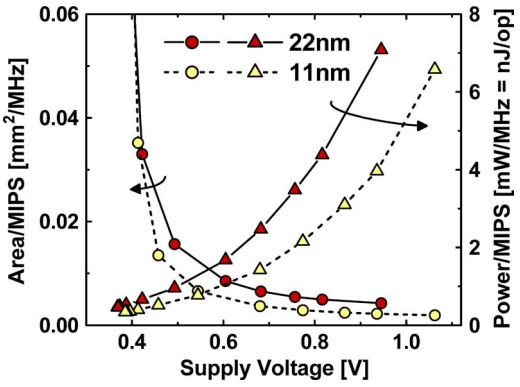


Fig. 6. *Dependence of area/MIPS and power/MIPS on supply voltage based on a full optimization of all technology parameters for bulk MOSFETs in the 22 nm and 11 nm nodes. Area/MIPS is an inverse performance metric in the presence of parallelism, while power/MIPS reflects energy per operation. Voltage scaling from ~1 V (representative of a high-performance 100 W processor chip) to 0.5 V allows considerable improvements (~8×) in power efficiency with only a moderate (~4×) trade-off area efficiency. Below 0.5 V, the performance trade-off is likely too severe for most practical applications. For simplicity, MIPS is calculated as the clock frequency divided by the effective cycles per instruction (~1.6) due to the latency penalty factor [13], so this is an underestimate of the true MIPS, since modern processor cores can usually execute more than one instruction per clock cycle.*

Table 2 Variability and Tolerance Assumptions for the Technology Optimization Program

Parameter	Global Variation	Local Variation
Gate length	σ = 2% of WHP	LER + 2% of WHP
Oxide thickness	σ = 2%	atomistic estimate
Doping	σ = 2%	atomistic estimate
Digital signal noise	N/A	σ = 4% of V_{DD}
Supply noise	15% guardband	σ = 3% of V_{DD}

The technology is assumed to be bulk MOSFETs with high-κ gate dielectrics, band-edge metal gates and raised source/drain contact regions with very shallow implants. Junction temperature is taken as 90°C, and the chips are assumed to be highly sorted for a parametric yield of 50%. WHP is the wiring half-pitch, which is 40 nm for the 22 nm technology node. LER is a width-dependent estimate of line edge roughness.

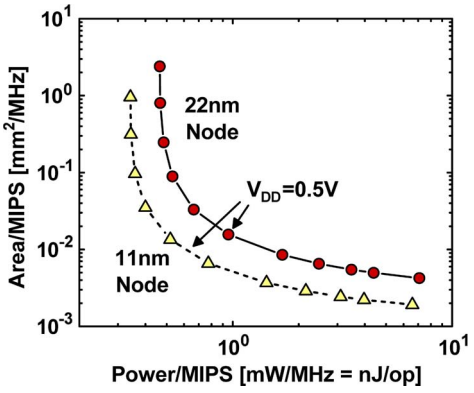


Fig. 7. *Trade-off between area and power efficiency for 22 nm and 11 nm bulk CMOS technologies. Supply voltages can be deduced by referring to Fig. 6.*

Arguments often cited against supply voltage reduction include subthreshold leakage, which may arise due to threshold voltage scaling, and susceptibility to variability and digital noise, which may lead to insufficient operating margins. These effects, however, are accounted for in the optimization program. Fig. 8 shows that the calculated loss of performance due to variability and tolerances increases as the supply voltage is reduced. This penalty, however, is within reason as long as voltages are maintained above ~0.5 V.

Figs. 6 and 7 implicitly assume a linear relation between parallelism and performance. This can be a reasonable approximation if the nonparallelizable portion of a fixed task is relatively small [16] or if the problem size can be grown for a fixed amount of execution time [17]. While there are inevitably overheads associated with increased parallelism due to communications and task partitioning, they are unlikely to be large since the proposed operating point near 0.5 V requires only a modest amount of increased parallelism. Nevertheless, depending on the architecture of future systems, bandwidth and cache capacity limitations may eventually place a limit on the efficacy of parallelization due to I/O congestion and finite chip size.

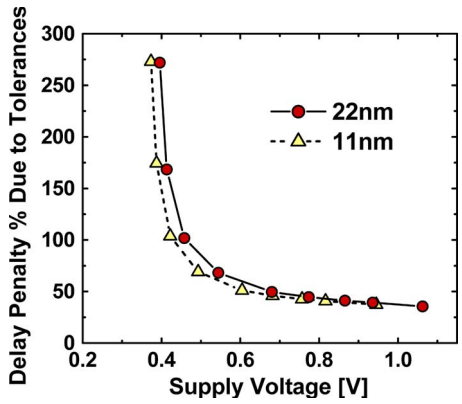


Fig. 8. *Calculated delay penalty due to variations and tolerances as a result of full optimization of technology parameters for bulk MOSFETs in the 22 nm and 11 nm technology nodes. As can be expected, the impact of tolerances grows as the supply voltage is reduced, but severe delay degradation is not observed until much below 0.5 V.*

It can thus be generally asserted that a moderate reduction in operating voltage—from the ~1 V supplies widely used today to ~0.5 V—can provide improvements in power efficiency (~8×) with frequency loss that can realistically be compensated for by parallelism (~4×) while staying in a voltage regime in which technology, circuits, and systems issues are still within control. While more aggressive voltage scaling into the subthreshold regime can enable optimal energy efficiency [18], [19], the associated performance degradation may be unacceptable for the vast majority of applications and many of the issues to be described later in this paper become tremendous challenges. On the other hand, the ~1 V supplies in use today are geared towards maximizing single-thread performance, which may not be appropriate for power efficiency in a parallel system. Thus, this 0.5 V supply voltage regime can be thought of as a practical compromise between power efficiency, performance, and circuit functionality—an operating point that is consistent across technology generations and device structure options. The next section identifies and assesses the issues that must be addressed to enable such advancement. Although the severity of each concern will surely vary across different end-user applications, the challenges and solutions are likely to be common.

III. VOLTAGE SCALING IN CMOS TECHNOLOGIES

While voltage scaling into the 0.5 V range can be straightforwardly motivated by the pursuit of power efficiency, achievement of such a goal requires simultaneous innovation and concession from many aspects of system design. The next several sections directly address each of the individual concerns that must be overcome to enable voltage scaling in 45 nm CMOS technologies and beyond.

A. New Device Scaling Paradigm

In scaling the voltage for a CMOS technology, the first issue to consider is the transistor structure itself. A thorough analysis must consider device-level trade-offs (e.g., scaling, performance, and parasitics) as well as the impact of these technology parameters on circuit density and, ultimately, the power-limited performance of a parallel system. In traditional ~1 V technologies, the gate length is scaled aggressively to fit within the targeted gate pitch, which is based on simple scaling to achieve a 2× density improvement every generation. The resulting MOSFET structure targets maximum device performance with only moderate short-channel effect control, where

DIBL (drain-induced barrier lowering as characterized by the difference between linear and saturated threshold voltages) can be more than 150 mV at a supply voltage of 1 V.

From the results of the optimization program described in Section II-B, a new scaling paradigm is observed for low-voltage technologies: optimal low-power devices exhibit significantly more short-channel effect control than is normally targeted for today's technologies. This is illustrated in Fig. 9, which shows that the optimal DIBL for a technology should be lowered in accordance with supply voltage scaling. Much of this reduction in DIBL comes directly from reducing the voltage, but part of it also comes from improving the electrostatic aspect ratio, which is the ratio of the channel length to the characteristic scale length, λ [20]. This quantity, which is essentially a measure of how well short-channel effects are controlled in a given device, can be thought of as an aspect ratio since the scale length is a quantity determined by device parameters perpendicular to the channel length. Increasing the electrostatic aspect ratio decreases the two-dimensionality of the FET, and, as shown in Fig. 9, the optimal ratio increases significantly with a push to lower power levels. The improved ratio is accomplished by utilizing gate lengths somewhat longer than expected, which necessarily increases gate pitch. This adjustment is preferred because reduced short-channel effects enables lower voltage, which quadratically improves power consumption, while increasing gate length only linearly worsens capacitance. The increased ratio also mitigates the impact of variability and leakage, which further helps to enable robust low-voltage operation. The optimizations show that it is worthwhile to use larger gate lengths because doing so trades off only a slight decrease in density for a significant reduction in voltage and power.

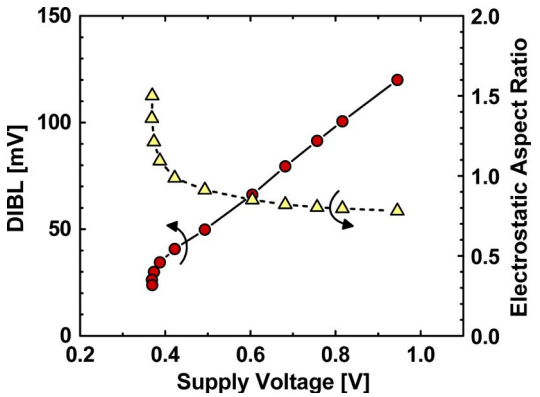


Fig. 9. *In an optimized technology, voltage reduction should be accompanied by significant improvements in the control of short-channel effects as evidenced by DIBL (a measure of short-channel threshold voltage rolloff) and the electrostatic aspect ratio (a measure of how well short-channel effects are controlled in a given device). This example assumes a 22 nm bulk MOSFET technology, but similar trends hold for different technology nodes and device structures.*

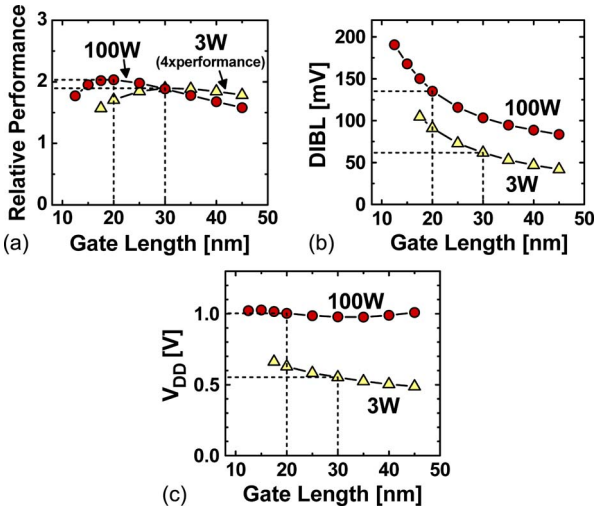


Fig. 10. *Gate length dependence of (a) performance, (b) short-channel effects, and (c) voltage in 22 nm bulk MOSFET technologies optimized for processor chips of two different power levels. For the low-power case, optimal performance occurs at much larger gate lengths, dramatically reduced DIBL, and lower voltages. Similar trends hold for different technology nodes and device structures.*

Fig. 10 compares the results of optimizing all technology parameters other than gate length for bulk MOSFETs at two different chip power levels. The maximum point on the performance curves [Fig. 10(a)] defines the optimal gate length. If the gate length is too short, short-channel effects [DIBL, Fig. 10(b)] degrade rapidly, which forces higher voltage [Fig. 10(c)], wastes energy, and ultimately worsens performance in a power-constrained scenario. Conventional scaling wisdom drives device design to smaller gate lengths and much larger values of DIBL, but the low-power optimization case shows that the highest performance comes from using a longer gate length and a lower DIBL—a target that becomes more extreme as power levels are reduced. Though not presented here, FinFETs and other device types show qualitatively the same sort of behavior. In particular, the improved electrostatic integrity of device structures such as the FinFET is best used to enable better short-channel effect control and lower voltages rather than to scale to the minimum possible gate length.

B. Low-Voltage Caches

In addition to a logic device technology optimized for low-voltage operation, practical computing systems must also have available a complementary low-voltage embedded memory. Static random access memory (SRAM) as depicted in Fig. 11(a) has long been the embedded memory of choice due to its inherent process compatibility and fast access time. However, SRAM cell transistors are especially vulnerable to variability in the manufacturing process due to a combination of aggressively scaled dimensions and sheer numbers. Since basic operation of the memory cell is

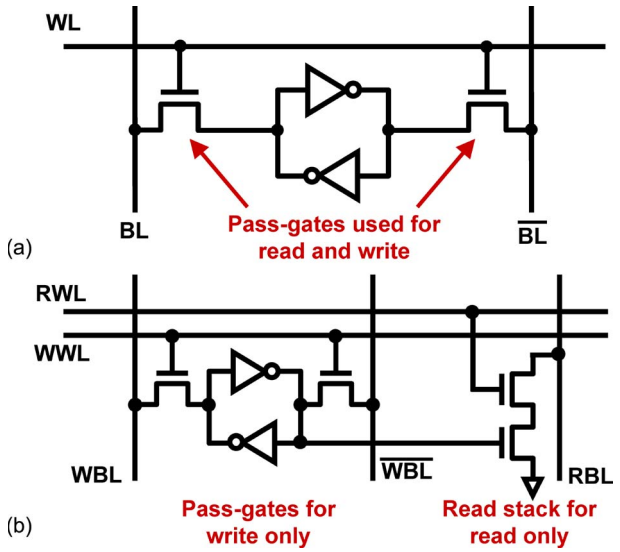


Fig. 11. *SRAM memory cell circuit diagrams for (a) standard 6T-SRAM, and (b) 8T-SRAM. Because the read and write ports are controlled by separate devices, the 8T cell can improve variability-limited yield, thus enabling voltage scaling. From Chang et al. [26].*

dependent upon carefully chosen device strength ratios, variability-limited yield presents the primary restriction in low-voltage SRAM operation. This is in stark contrast to low-voltage logic operation, which is limited primarily by performance degradation. Already, SRAM minimum voltage limits are a significant concern at ~1 V supplies; voltage scaling to 0.5 V will undoubtedly necessitate new, low-voltage SRAM solutions.

In modern CMOS technologies, SRAM device variability is dominated by discrete dopant fluctuation [21] and is most visibly manifested as random distributions in transistor threshold voltage. The standard deviation of these Gaussian distributions can approach $\sigma_{Vt} \sim 50$ mV for advanced technologies, which results in an intrinsic threshold voltage range of several hundred millivolts across a megabit-scale memory array. To ensure sufficient margins for both read and write operations, the minimum operating voltage for an SRAM array must be quite high—a value that will only increase as technology continues to scale and variability continues to intensify. While device optimization can reduce the impact of variability, only limited robustness can be gained, and such improvements generally come at the expense of either performance or leakage power. Consequently, the optimum design varies by application depending on the importance of performance and leakage, thus resulting in a growing divergence between cell design for high-speed first-level (L1) and dense second-level (L2) caches.

A powerful yet brute force solution to SRAM voltage scaling is simply not to scale, but rather to add a dedicated SRAM supply voltage higher than the standard logic supply [22]. The higher voltage as well as the offset between the two supplies can not only help to enable SRAM scaling to future technologies, but also presents a strategy to maintain SRAM functionality when logic voltages scale. In some ways, this can still be compatible with a low-power strategy since in this dual-voltage scheme, bit line voltages, which often dominate active power dissipation, can scale with the logic supply. Such a solution, however, does not scale well into new technologies and architectures as this secondary voltage may further increase in future technology nodes and must also be distributed throughout the chip and the system. For some applications, such as L1 caches, which may be scattered throughout a processor core, it may be difficult or undesirable to add another power grid for a separate supply. In addition, especially in massively parallel systems, the penalties associated with generating, distributing, and regulating another voltage may also be prohibitive. An embedded memory that can instead scale in voltage along with logic and share a common supply is thus the most desirable solution. While many techniques have been proposed to achieve this goal, most add considerable complexity to the peripheral circuits of the memory array and result in penalties in performance, power, and area [23]–[25]. Instead, it may be more effective to fundamentally change the circuits or technologies used to build SRAM in order to enable voltage scaling.

A simple low-voltage SRAM solution is the use of the 8T-SRAM cell depicted in Fig. 11(b) [26]–[28], which adds two transistors to the conventional 6T-SRAM cell. The additional devices form a read access port to the cell that is decoupled from the pass-gate devices, which form a write access port. Because the read and write functions are performed by different transistors in an 8T cell, each can be optimized independently to maximize read and write operating margins. In contrast, in a 6T cell, the same access devices are used for both read and write operations, which leads to a fundamental optimization trade-off between the two. As such, the operating margins for 8T-SRAM can be dramatically improved over that for 6T-SRAM, which enables increased robustness to variability and thus lower voltage operation, as evidenced by the wide operating range in Fig. 12. In [26], an 8T array was demonstrated down to 0.41 V operating voltages in a 65 nm CMOS technology with, as shown in Fig. 13, a linear frequency dependence on voltage. This suggests that the adoption of this cell configuration can allow SRAM to share the same scaled power supply as logic in future low-voltage technologies. Despite an increased transistor count, the cell size for 8T-SRAM is not necessarily larger than that 6T-SRAM. Due to reduced operating margins, a 6T cell must maintain large transistor dimensions to alleviate variability, which greatly increases cell size; in contrast, these dimensions can be continually scaled in an 8T cell. In addition, array efficiency (the fraction of array area dedicated to the memory cells as opposed to peripheral logic) is often improved in an 8T-SRAM array such that the total area of the memory array can be quite dense. As

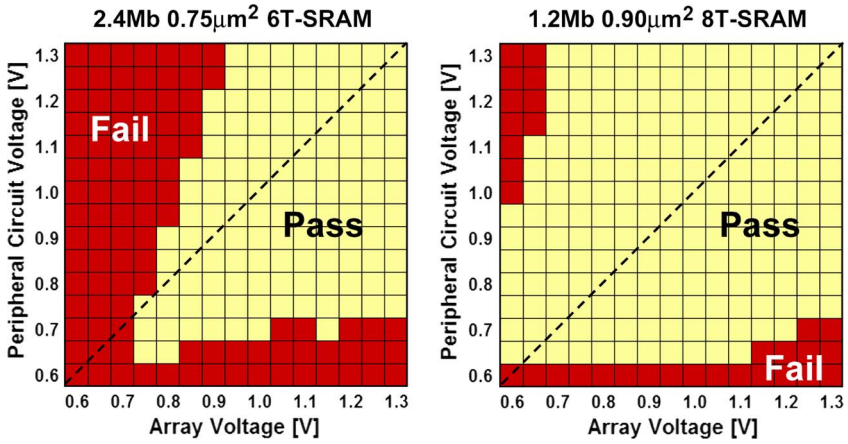


Fig. 12. *65 nm SRAM yield monitors show that 8T-SRAM improves low-voltage functionality. Voltage scaling down to the Vdd limit of the peripheral logic (∼0.6 V) is observed. From Chang et al. [26].*

CMOS technology approaches the 22 nm node, the area requirements for comparable 6T-SRAM and 8T-SRAM arrays begin to converge and may soon cross over such that 8T-SRAM is, in fact, more area-efficient.

Low-voltage SRAM could also be enabled by modifying the underlying transistor structure to mitigate variability. Thin-body MOSFET structures, such as extremely-thin silicon-on-insulator (ET-SOI) and the FinFET as shown in Fig. 14, have been studied for some time as solutions to continue gate length scaling beyond the 22 nm node [29], but significant benefits may also lie in the ability to build practical device structures with no doping in the channel. Since short-channel effects can be well-controlled by the thickness of the silicon channel itself, channel dopant profile engineering, such as commonly achieved by halo implants, is not necessary. This eliminates any variability that would otherwise arise from random dopant fluctuation and thus removes a substantial portion of the σ_{Vt} that is observed in CMOS technologies today. Fig. 15 demonstrates that this can enable significant yield enhancement and voltage reduction for 6T-SRAM; implemented in an 8T-SRAM cell, maximum voltage scalability could be attained. These thin-body device structures could thus be an ideal solution for SRAM scaling [30]. However, since the performance of these device structures often suffers from parasitic resistance and capacitance [31], there may be a divergence between logic performance needs and memory yield requirements such that a hybrid technology—traditional MOSFET structures for logic and thin-body structures for SRAM—may become a desirable option.

C. On-Chip Digital Noise

While the solutions discussed in the previous sections can help to maintain logic performance and memory circuit functionality at low supply voltages, it must also be ensured

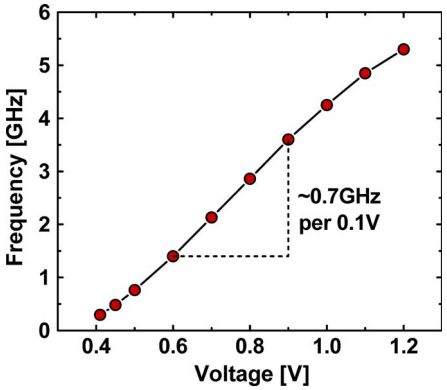


Fig. 13. *A 65 nm 8T-SRAM subarray optimized for low-voltage operation demonstrates linear scaling of frequency down to 0.41 V. Adapted from Chang et al. [26].*

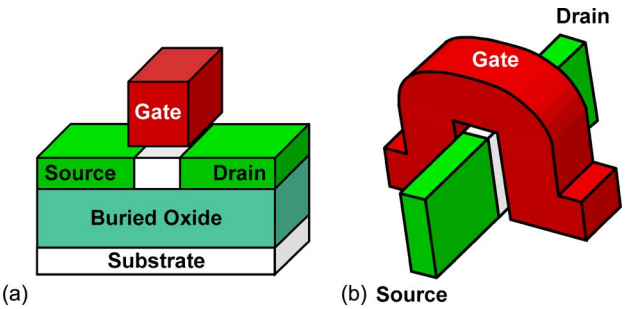


Fig. 14. *Thin-body transistor structures such as (a) single-gate extremely-thin SOI (ET-SOI) or (b) the double-gate FinFET have been proposed for gate length scaling beyond the 22 nm node, but also enable viable devices without channel doping, which eliminates random dopant fluctuation—the primary source of variability in SRAM.*

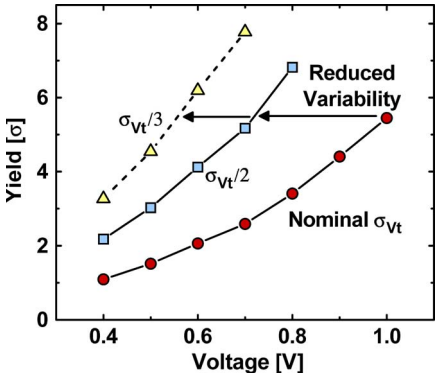


Fig. 15. *Simulated yield projection for a generic 6T-SRAM cell compatible with 22 nm ground rules. Reduced threshold voltage variation (characterized by a Gaussian distribution of length σ_{Vt}) enables voltage scaling since cell operating margins can be maintained at low voltages. This can be achieved in ET-SOI or FinFET device structures that eliminate discrete dopant fluctuation.*

that signals being propagated at these low voltages are resilient to those noise sources relevant to digital circuits. A common initial reaction to voltage scaling is that susceptibility to noise is increased—that at a constant noise level, a reduction in operating voltage could compromise margins. It is important to remember, however, that the noise sources relevant to digital circuit operation will scale with voltage and that the end impact on circuit functionality may, in fact, improve. It will thus be argued in this section that such on-chip digital noise should not be of concern if voltages are scaled to ~0.5 V as proposed.

Generally, digital CMOS circuits are quite tolerant of noise, but operating margins for some circuits can be small. In particular, dynamic logic can suffer from charge leakage problems while latches may see degraded setup and hold times. Sources of such voltage noise can be caused by resistive drops, capacitive charge coupling, and inductive transients. With appropriate consideration of scaling factors, each of these noise sources decreases in importance as voltages are lowered. It should be noted that this discussion neglects mechanisms such as thermal, shot, and 1/f noise. While important for analog circuits, such noise is not generally a concern for digital circuits.

Resistive voltage drops as a fraction of the power supply voltage are related to the current drive I of a given device and the characteristic resistance R of the wiring path in question:

$$\frac{\Delta V_R}{V_{DD}} = \frac{IR}{V_{DD}} \propto \frac{(V_{DD} - V_T)^{1.5}}{V_{DD}}. \tag{4}$$

Since R is not a function of voltage, the expression above depends only on I, which, for the purposes of this discussion, can be expressed as a power law function of the gate over-drive voltage [32], where the exponent is assumed here to be ~1.5 for modern CMOS technologies. Since I is a super-linear function of V_{DD}, it scales faster than V_{DD}, and the expression above can be seen to decrease with voltage scaling for relevant values of V_{DD} and V_T. Thus, resistive voltage drops become less of a concern with voltage scaling.

Voltage noise due to capacitive coupling occurs when an aggressor of parasitic capacitance C_{agg} switches by a potential difference of V_{DD} and acts on a victim load capacitance C_{vic}:

$$\frac{\Delta V_C}{V_{DD}} = \frac{\frac{C_{\text{agg}} V_{DD}}{C_{\text{vic}} + C_{\text{agg}}}}{V_{DD}} = \frac{C_{\text{agg}}}{C_{\text{vic}} + C_{\text{agg}}}. \tag{5}$$

The magnitude of this coupling noise as a fraction of V_{DD} is related to the capacitive divider between the victim and aggressor. Since charge is directly proportional to V_{DD}, this ratio is not a function of voltage. Thus, voltage noise due to capacitive coupling scales with V_{DD} and does not worsen.

Inductive voltage noise arising from current transients can be calculated as

$$\frac{\Delta V_L}{V_{DD}} = \frac{L\frac{\partial I}{\partial t}}{V_{DD}} \propto \frac{LI}{V_{DD}\tau} \propto \frac{(V_{DD} - V_T)^{1.5}(V_{DD} - V_0)}{V_{DD}} \tag{6}$$

where τ is the characteristic time of such current spikes, which is related to the operating frequency of the circuit in question, which, from (2), is linearly dependent upon V_{DD}. This overall expression is a super-linear function of V_{DD}, which means that inductive noise scales faster than V_{DD} and thus only improves with voltage scaling.

D. Power Delivery

Assuming that low-voltage logic and memory solutions are available and digital noise is contained, the next requirement is to ensure that this low-voltage supply is efficiently and accurately delivered to the chip. Without appropriate consideration of the power delivery system, excessive voltage margins may be needed, which can counteract the gains achieved by successful on-chip voltage scaling. Already today, at ~1 V supplies, Fig. 2 shows that the power loss and noise in the path from the external power source to the circuits on a chip can be significant. When voltage is reduced to improve power efficiency at constant performance, total power is lowered, which improves supply efficiency and stability. However, in scenarios that increase parallelism beyond this point to improve system-level performance (such as through future density scaling or in a power-constrained application), these issues become severely degraded and require

advancements in chip packaging or point-of-load power conversion.

In a traditional power delivery system, the supply voltage is normally regulated by a dc–dc converter located off-chip, then delivered to and distributed throughout the chip via a power grid. Nonnegligible power loss occurs in the power delivery network due to Joule heating, which degrades power efficiency. For a system to deliver a power P at a voltage V and total current I through a power delivery line of effective resistance R, the power loss is given by

$$\frac{P_{\text{loss}}}{P} = \frac{I^2R}{P} = \frac{(P/V)^2R}{P} = \frac{PR}{V^2}. \tag{7}$$

While a reduction in voltage could increase power loss, the corresponding drop in power dissipation levels more than compensates since the dependence of power on voltage in (3) is more than cubic. For the optimizations discussed in Section II-B, scaling operating voltages from 1 V to 0.5 V and reoptimizing the device technology yields a power density reduction of $\sim 30\times$. This results in degradation of the operating frequency, which must be offset by introducing an additional $\sim 4\times$ in parallelism. Assuming that parallelism is achieved at the system level and not by growing chip size, it can be assumed that the resistance remains constant. Thus, as shown in Table 3, power delivery efficiency for a fixed design may not degrade, but might in fact be improved with voltage scaling.

Supply variation due to sudden load changes can result in a voltage drop, which can be calculated as

$$\frac{\Delta V_L}{V} = \frac{L\frac{dI}{dt}}{V} \propto \frac{\Omega LP}{V^2} \tag{8}$$

where L is the inductance of the power distribution network and Ω is the characteristic frequency over which current loads change (which might be related to the distribution network rather than the voltage-dependent chip operating frequency). As with power loss, Table 3 shows that supply variation scales well with voltage due primarily to reduced power dissipation levels. The additional dependence on frequency may further suppress supply noise at low voltages—rendering instability a less critical issue than power delivery efficiency. However, as shown in Fig. 8, it should be remembered that circuits operating at low voltage may be more sensitive to supply variations. It should also be noted that this issue can also be improved by the addition of more decoupling capacitance.

Table 3 Impact of a Moderate Voltage Reduction (From 1 V to 0.5 V) of a Fixed Design Based on a Full Technology Optimization and Subsequent Constant Voltage Technology Scaling

Device or circuit parameter	Relative change due to voltage reduction	Scaling Factor
Chip operating voltage	0.5	1
Number of parallel units	4	κ^2
Device density	1	κ^2
Power density	0.03	κ^2
Power line loss	0.12	κ^2/α^2
Power line variation	0.12	κ^2/α^2

κ is the dimensional scale factor. α is the on-chip voltage down-conversion factor.

While voltage scaling appears not to cause problems in the power delivery system due to aggressive reduction of total power dissipation, this situation may not hold with continued technology scaling to future nodes. As might be expected from the trends in Fig. 4, after this one-time drop in supply voltage, significant further voltage scaling is unlikely, which may well lead to increasing power density due simply to lithography and ground rule scaling. Under the assumption that chip sizes will stay constant, future nodes will allow increased parallelism, which translates to increasing chip power levels. At lower voltages, power delivery efficiency and variation become more sensitive in future technology nodes. Thus, assuming realistic scaling factors, Table 3 suggests that the improvements in both power loss and supply noise achieved by initial voltage scaling might be conceded in just a few technology generations.

In the Section II-B optimizations, power efficiency is improved at constant system performance, which, as stated above, only results in power delivery efficiency and supply variation issues when scaled to future technology nodes. However, the savings in power efficiency due to voltage scaling could instead be used to maximize the number of parallel units for a given power budget, which improves system performance at constant total power. Just as with density scaling, the number of parallel units could be increased dramatically—constrained instead by cost and physical chip size limits. In this case, (7) and (8) indicate that the issues of power delivery efficiency and supply variation could worsen significantly with voltage scaling. Thus, whether due to density scaling or the desire to maximize system performance at constant total power, new methods of power delivery may be needed.

Unless chip packaging techniques can be dramatically changed to reduce both resistance and inductance, a new strategy is required to accommodate the delivery of efficient and stable low-voltage power supplies. The most effective solution is to combine moderately high voltage power delivery and on-chip voltage down-conversion in a scheme illustrated in Fig. 16. Since noise and efficiency both depend strongly on voltage, increasing the voltage at which power is distributed can present tremendous benefits. Such a solution, however, requires the development of highly efficient on-chip dc–dc voltage conversion techniques to down-convert this large distribution voltage to the desired operating level. Resistive series regulators

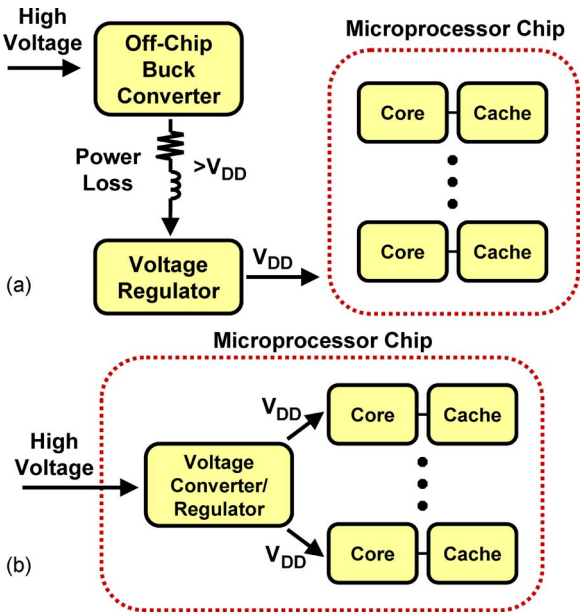


Fig. 16. *(a) Traditional power delivery system, which can experience significant power loss as voltage conversion is performed off-chip. (b) On-chip voltage conversion and regulation allows for moderately high voltage power delivery to the chip, which minimizes power loss and improves supply stability.*

are fundamentally limited to low (~50%) conversion efficiencies due to the inherent resistive divider network and are thus unsuitable for on-chip voltage conversion. Buck converter techniques utilizing on-chip inductors are more efficient, but practical implementations are limited to ~75% due to difficulties in achieving on-chip inductors with high quality factors. Instead, switched-capacitor circuits as generally depicted in Fig. 17 may be an effective solution for on-chip voltage conversion. Such circuits traditionally suffer from limitations in efficiency [33], but

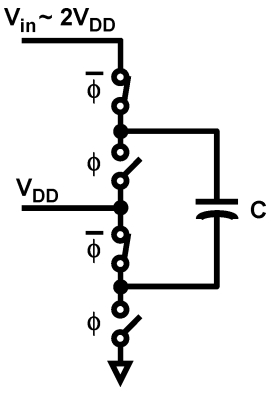


Fig. 17. *Switched-capacitor circuit for on-chip voltage conversion. By utilizing high-performance SOI MOSFET technologies and low-parasitic trench capacitors, high conversion efficiencies can be attained.*

recent advancements in process technology can potentially enable on-chip conversion efficiencies of more than 90%. The improvement is primarily derived from the availability of trench capacitor structures in a high-performance CMOS process. Trenches used for embedded DRAM [34] yield capacitors of very high density with minimal stray parasitics. In addition, with technology scaling, the MOSFETs used as switching devices become quite efficient at the 45 nm node and beyond. Furthermore, the introduction of SOI substrates enables the down-conversion of higher voltages without well isolation leakage concerns. As shown in Table 3, maximizing the voltage of the power delivery system and thus the on-chip down-conversion factor is the most effective means by which to mitigate efficiency and stability concerns.

E. Off-Chip Connections

Combined, solutions to the aforementioned issues can enable low-voltage operation of a chip to improve power efficiency. However, any chip must always communicate with the rest of the system, which, as shown in Fig. 2, can dissipate significant power. In particular, in many applications, much of the power associated with the memory subsystem can be attributed to such interconnections. Especially for future exascale applications that stress extreme memory bandwidth, it is imperative that along with voltage scaling, solutions need to be found to reduce power in off-chip connections. For lossless, short-range connections, voltage scaling of the signals driving the capacitive load can effectively improve power efficiency. Such a strategy, however, may not be effective for lossy, long-reach connections as the circuits needed to compensate for channel attenuation tend to dominate power, thus driving the need for alternate solutions.

Off-chip connections that are relatively short or otherwise operating in a high-quality channel can be thought of as lossless. Depending on available packaging strategies, such connections can comprise a significant portion of overall I/O power—especially with rising needs in cache bandwidth close to the processor. Without attenuation concerns, the driver and receiver circuits are relatively simple, and the power needed to drive the connection itself can dominate. For these short connections, the active power can be expressed as

$$P_{I/O} = C_{I/O} V^2 f_{\text{eff}} \tag{9}$$

where $C_{I/O}$ is the interconnect capacitance, V is the operating voltage, and f_{eff} is the effective frequency—considering activity factors—at which the connection is operated. Clearly, scaling of the output voltage range in these interconnect driver circuits is an effective method by which power could be reduced. The introduction of a locally generated and regulated low-voltage supply (e.g., as

discussed in Section III-D) can enable a power-efficient, low-voltage driver. On the receiving end, a single-ended sense amplifier, such as enabled by a gated diode device [35], can provide efficient, low-voltage signal recovery. Together, as shown in Fig. 18, these components can minimize power in low-loss connections. As discussed in Section III-C, as long as the voltages of all connections are scaled together, signal crosstalk can be minimized. For short interconnections that follow (9), it may also be possible to reduce interconnect capacitance. In particular, advanced packaging techniques such as three-dimensional integration via wafer bonding [36] or silicon carriers [37] bring chips closer together, which can eliminate transmission line effects, reduce capacitance, and decrease power as compared with traditional I/O pins and board-level wiring. Ultimately, continued density scaling and single-chip integration can shorten many off-chip connections.

For longer reach interconnections that suffer from losses due to high-frequency attenuation, it may be possible to utilize low signal swing to reduce power [38], but ultimately, channel quality limits the practicality of such techniques as transceiver circuits can dominate total power. Recent work on low-power serial links has focused on equalization techniques [39]–[41], which may benefit somewhat from the general CMOS voltage scaling strategies described in this paper; however, voltage scaling in analog circuits may be limited and parallelism is likely not a viable solution in this case. Thus, ultimately, optical interconnect [42] may be needed to achieve significant power reduction in long-range links.

It should be noted that the power associated with off-chip connections can also be dramatically affected by the design and organization of the overall system. For example, since significant energy is consumed in moving data between main memory and the computational engine, the most power-efficient solutions directly attach DRAM chips to the processor chip without intervening address/control or data redrive circuits, hub chips, or other JEDEC [43] standardized devices. In addition, the availability of sufficient I/O pins enables operation of off-chip connections at a modest data rate, which allows for source-terminated interconnects and removes the need for far-end bus termination, thus further reducing power.

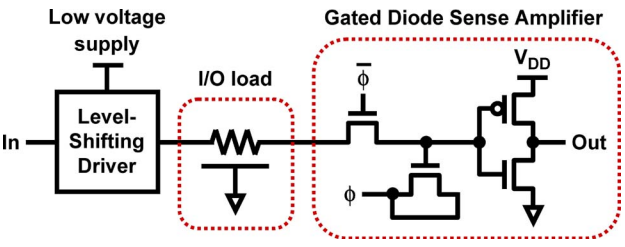


Fig. 18. *Low-voltage signaling circuit concept. Combined with a low-power supply source, a level-shifting driver circuit can efficiently limit the swing of a large capacitive I/O load. A single-ended sense amplifier based on capacitive coupling in a gated diode can be used to recover the signal.*

IV. BLUE GENE CASE STUDY

IBM's series of Blue Gene supercomputers—notably Blue Gene/L [44] and most recently Blue Gene/P [45]—combine many tens of thousands of low-power computing nodes of modest performance to yield massive supercomputers that are not only the lowest power per computation [46], but also the fastest in the world. Blue Gene machines held the top position in the Top500 list [47] from 2004 to 2007; with a planned third platform, it is possible that this architecture will again achieve top marks. Fundamentally based upon massive parallelism, Blue Gene systems provide a practical framework within which to discuss the strategies outlined in this paper. While the Blue Gene/L and Blue Gene/P systems take some important initial strides towards power efficiency, future systems such as Blue Gene/Q and beyond will likely make more widespread use of the concepts discussed in Section III.

A. Voltage Scaling

Blue Gene systems utilize efficient, voltage-scaled processors combined with system-on-a-chip designs that integrate memory controllers, a network router, and an I/O adapter alongside the processor and local cache. The Blue Gene/P quad-core processor chip, fabricated in a 90 nm node process, operates at 850 MHz—a frequency well below that of other processors in similar technologies since performance will be compensated for at the system level by parallelism. Due to low power dissipation levels, 4096 processor cores can be placed in parallel within a single air-cooled cabinet to deliver a peak performance of 13.9 Tflops per rack. On the LINPACK benchmark, the system delivers 82.3% of peak performance at a power of 31.5 kW per rack, which translates to a system-level power efficiency of 364 Mflops/W. While current Blue Gene systems employ low voltages that are within technology specifications, future designs will leverage subnominal supply voltages to attain greater power efficiency—eventually driving the need for low-voltage device and memory techniques such as those outlined in Sections III-A and III-B.

B. Power Delivery

In a massively parallel system, generating, monitoring, and preserving the requisite supply voltages can be challenging—due both to the sheer number of voltages required as well as stringent requirements on supply robustness and redundancy. With increased parallelism and voltage scaling, these issues will become more severe in future generation systems.

In Blue Gene/P, to facilitate efficiency in data communication and cycle reproducibility in massively parallel

applications, a single system clock is distributed to all compute nodes. As a consequence, since each processor chip must run at the same frequency, tailored supply voltages are used to minimize power in the presence of process variation. Chips are binned by performance into three groups—each with a different supply voltage—with the fastest chips running at a lower voltage and the slowest chips running at a higher voltage. Three separate high-current supply voltages must therefore be delivered to the chips in the system.

With a multitude of supplies operating at high current levels and energy densities, redundancy, supply stability, and IR losses must be carefully addressed. To ensure high system reliability, redundant supply voltages are provided by point-of-load converters distributed on a large circuit card. In Blue Gene/P, four power supplies offer ~800 A per 32-processor node card while a fifth redundant supply is available to cover failures. Supply failures create the largest transient response, but near-instantaneous changes of up to 40% in processor power due to synchronous clocks and power-gating techniques also create large voltage transients. The power converter loop response and recovery from a supply failure must be fast enough so that the voltage droop does not fall below the minimum voltage required to run the processor. To ensure this, the nominal voltage is raised high enough to cover the worse case voltage droop and thus adds to the nominal power dissipation. Power planes on the circuit card are designed to minimize these drops and match the voltage delivered to each processor chip. Processor nodes far from the power supply have additional conduction paths in other circuit card layers to reduce resistance while parallel connections may be removed from other nodes; as a result, supply equipotentials are created at each processor chip location. The number of power planes on a circuit card is fixed, however, such that as extra power supply voltages are added (as needed for DRAM and I/O), distribution losses inevitably worsen.

In future systems, as discussed in Section III-D, the final stage of dc–dc voltage conversion may be performed directly on the system planer or ultimately on the processor chip itself. By delivering higher voltages closer to the chip, supply stability and IR losses can be improved, which may be especially important as operating voltages are scaled. In addition, local voltage generation and regulation could significantly simplify the power delivery system by reducing the number of supplies needed.

C. Off-Chip Connections

In Blue Gene/P, each processor chip contains a 32 B interface to directly attached SDRAM-DDR2 memory—a system design that reduces power consumption. Source-terminated I/O cells matched in impedance to the transmission lines between the processor and DRAM eliminate the need for other data line termination. The power dissipated in these lines follows (9), where the capacitance is minimized by placing memory immediately adjacent to the processor chip. By combining variable voltage I/O cells and power supplies, low-voltage memory can be introduced as it becomes available—a trend to be continued in the next generation machine with SDRAM-DDR3.

Going forward, the power dissipated in the connections between the processor and external memory will become an increasingly large part of the total power budget unless the interconnect capacitance can be reduced to compensate for increasing bus frequencies. By utilizing high-I/O-count DRAM stacks based on through-silicon-vias or other high wiring density interconnect media, large amounts of external DRAM can be placed adjacent to the processor with greatly reduced wiring path lengths.

The longest off-chip connections in Blue Gene/P are used to communicate with processors in other racks. These links, which can be up to 8 m in length, are currently managed by electrical cables using differential signaling at 3.2 Gb/s, but may eventually be well-handled by optical techniques. In optical connections, attenuation is substantially less than that in electrical connections, which means that links of up to 100 m may be reached with little or no more power than links of a few meters—a characteristic that may lead to new system design paradigms.

V. FUTURE STRATEGIES FOR POWER EFFICIENCY

Several new ideas to solve or circumvent the issues described in Section III are currently being pursued in the research community and merit discussion. These concepts, which combine innovations in technology, circuits, and systems, may be key enablers for more drastic power reduction than can be achieved by scaling CMOS voltages to the ~0.5 V limit. The back-gated MOSFET is first discussed as a technology that could enable more efficient system-level power delivery. Steep subthreshold slope devices are then described as a new technology to enable further scaling of operating voltages without performance loss. Finally, reversible computing techniques are considered as a fundamental change to circuits and architecture that can approach the ultimate limits of power efficiency.

A. Back-Gated MOSFET for Variability Compensation

As depicted in Fig. 19, the back-gated MOSFET [48] has long been studied as a device that can offer both scaling advantages and threshold voltage modulation. Similar in structure to an ET-SOI device, it incorporates a thin buried oxide to allow a conductive back-gate electrode to modulate the channel current. The device concept is akin to a double-gate MOSFET, but it instead assumes that the back-gate terminal is not involved in switching operation of the device, but only biased to set the threshold voltage of the front channel. The device structure is thus normally optimized with a back-gate oxide considerably thicker than

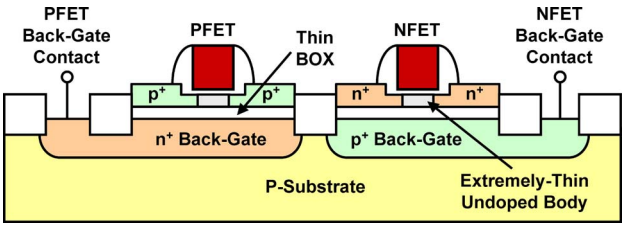


Fig. 19. *The back-gate MOSFET structure leverages a thin buried oxide (BOX) and extremely-thin undoped body to enable threshold voltage modulation by NFET and PFET back-gate electrodes. Such adjustment can compensate for systematic process variation, which facilitates power delivery in massively parallel systems.*

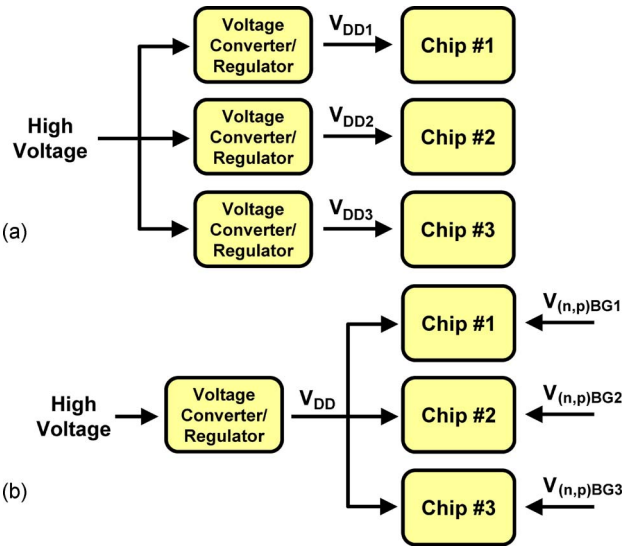


Fig. 20. *Exemplary schemes to compensate for chip-to-chip variability due to manufacturing processes. (a) Standard technique of applying different voltages to different chips. (b) Scheme utilizing the back-gate MOSFET, in which a single voltage converter/regulator can supply the same voltage to all chips. Because they do not draw significant current, individual back-gate voltages can be efficiently generated and regulated.*

the front-gate oxide to minimize parasitic capacitance to the source and drain. The device shares scaling characteristics with ET-SOI, but integrates a back-gate terminal that is essentially an idealized version of a well or substrate bias for a bulk MOSFET. With a thin, fully depleted silicon body, capacitive coupling from the back-gate electrode can be very strong and thus have an appreciable effect on the device threshold voltage. In addition, due to oxide isolation of the back-gate from the channel, there are no limitations on the voltage range of the applied bias; however, due to the practical range of electric fields, threshold voltage tuning is perhaps best kept to a few hundred millivolts.

While the back-gated MOSFET is often proposed as a power-gating method that dynamically adjusts threshold voltages to place circuit blocks into a standby state, other techniques to achieve such functionality exist and perform adequately [7]. Instead, the back-gate structure may hold much greater benefits in variability compensation for efficient power delivery. As described in Section IV, in a massively parallel system comprised of many separate processor chips, due to manufacturing variability, chips may be binned by performance and different supply voltages may be applied to each such that the performance of each chip in the parallel system is, in the end, constant. This requires efficient regulation of multiple power supplies—the granularity of which is limited by cost considerations. With a back-gated MOSFET technology, the back-gate voltage may be used to tune the threshold voltage of each chip to its desired value, thus enabling every chip in the system to operate at a common supply voltage. Such a scheme, as shown in Fig. 20, can help to reduce cost and complexity in the power subsystem and facilitate supply redundancy. Separate back-gate voltages can be applied to NFETs and PFETs, which enables compensation of systematic variation between the two device types. In this scenario, each chip must have separate back-gate voltages, but since these supplies draw very low current (back-gate electrodes are isolated by oxide), generation and regulation of this voltage level can be far more efficient than that for the multiple power supplies that would otherwise be needed. At these low current levels, even the simplest of on-chip converter circuits could provide suitable efficiency. In the end, while availability of such a device may not fundamentally alter the overall power efficiency of a computing system, it holds the promise of simplifying the power subsystem, which could lead to tangible benefits in efficiency.

B. Sub-60 mV/decade Inverse Subthreshold Slope Tunneling FETs

From the analysis in Section II, voltage scaling is ultimately limited by the need to maintain sufficiently high threshold voltages to control device leakage. Conventional CMOS technologies are subject to a 60 mV/decade minimum constraint on the inverse subthreshold slope at room temperature due to inversion charge generation as determined by the thermal distribution of carriers in the conduction and valence bands. As shown in Fig. 21(a), reduction of the inverse subthreshold slope to below this limit enables faster turn-on of the device with gate voltage, which, at a given leakage specification, allows for voltage scaling to improve power efficiency.

Several classes of devices can potentially offer sub-60 mV/decade inverse subthreshold slopes by altering carrier transport or generation in the device channel. Ferroelectric gate materials [49] have been proposed to leverage an effective negative oxide capacitance to magnify channel surface potential modulation with gate voltage; although tantalizing, such an effect has yet to be confirmed experimentally. Very steep subthreshold characteristics

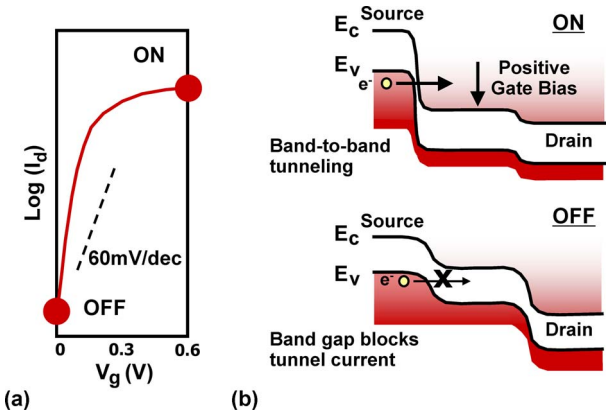


Fig. 21. *(a) Steep subthreshold slope devices enable sharp turn-on characteristics such that at a given leakage current specification, large on-current can be achieved even at low voltages. The most promising structure to achieve such characteristics is the tunneling FET, as depicted by (b) band diagrams in the source-to-drain direction.*

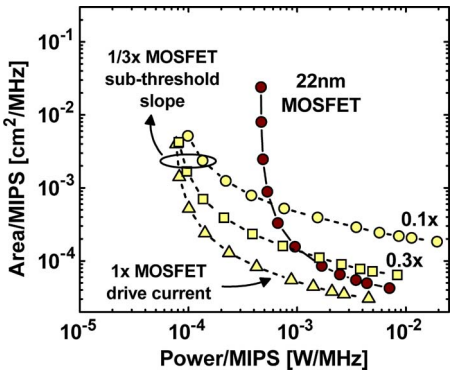


Fig. 22. *Area vs. power trade-off analysis for theoretical devices with a 3× improvement in subthreshold slope over traditional MOSFETs. For most practical applications, it is likely that a corresponding drive within an order of magnitude of current MOSFETs will be required. This analysis represents the range of characteristics that might be achievable with tunneling FETs.*

have been demonstrated [50], [51] using nonthermionic mechanisms based on impact ionization or positive feedback, but these devices require large (> 1 V) source-to-drain biases and suffer from fundamental drawbacks in switching speed and reliability. The most promising steep subthreshold slope devices appear to be tunneling transistors, which operate via the principle of gate-controlled band-to-band tunneling. As shown in Fig. 21(b), a positive gate bias creates tunneling paths between the valance band in the source and the conduction band in the channel, while reduction of the gate bias abruptly shuts off the tunneling mechanism. While experimental results to date suffer from low drive current [52], the use of heterojunction band gap engineering could potentially enhance drive currents to desired levels [53].

The ultimate limits in drive current and subthreshold slope that can be achieved in tunneling transistors are not yet understood, as uncertainties remain in the physics of band-to-band tunneling and nonidealities in the device structure itself. Due to improved subthreshold slope, however, it may not be necessary for tunneling transistors to achieve the same drive current levels as today's MOSFETs. Using the optimization program described in Section II-B, the trade-off between subthreshold slope and drive current is depicted in Fig. 22. To roughly emulate a tunneling transistor, a MOSFET model is modified using simple multipliers for the subthreshold slope and drive current. This does not capture the unique properties of tunneling FETs, but can still provide guidance on power and performance trade-offs for a generic device with improved subthreshold slope. For applications with practical operating frequencies (indicated by low area/MIPS), drive currents for a device with a 3× improvement in subthreshold slope must be within a factor of three of conventional MOSFET levels to yield an overall benefit. However, even with a 10× degradation in drive current, steep subthreshold slope devices can open up a new power/performance space that was previously unachievable. While such devices may only provide incremental gains for high-performance systems, power efficiency improvements of nearly an order of magnitude could be achieved for ultra low-power applications.

C. Reversible Computing

In addition to new device concepts, novel circuit and architecture approaches can also be used to achieve low power. In particular, the reversible computing paradigm, which has been investigated since the early 1990s [54]–[62], can be implemented using conventional CMOS FETs and adiabatic charging techniques to dramatically reduce energy consumption without voltage scaling. Such a radical change in architecture may enable computing technologies to approach the ultimate physical limits of energy consumption. The theoretical basis of reversible computation is that the laws of physics only require energy to be dissipated by computation when information is erased [63]. If computation can be cleverly performed in a logically reversible manner, it can be arranged such that nothing needs to be erased [64], which means that energy dissipation can, in principle, be arbitrarily low. While quite challenging to achieve, such reversible techniques might be the only viable alternative to dramatically improve power efficiency beyond voltage scaling of conventional circuits [65].

A simple example of reversible logic without erasure is shown in Fig. 23, in which adiabatic charging is used with CMOS to implement an XOR logic function. During each cycle of the LC resonator (essentially a clock), the logic signal A $\oplus$ B is created and then removed from the output node capacitor. If A = B = "0" or A = B = "1", the

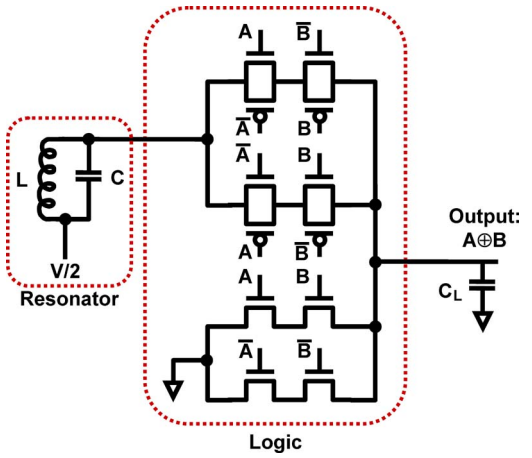


Fig. 23. *XOR logic implementation utilizing adiabatic charging. The LC resonator causes the output to oscillate between A ⊕ B and "0" with relatively little dissipation. Inputs A and B are assumed to be static during this oscillation.*

resonator oscillates, but charge does not pass through the transmission gates and the output is held at zero; in this case, energy dissipation occurs only in the LC resonator. If A = "1" and B = "0", or A = "0" and B = "1", then charge flows smoothly back and forth from the resonator to the output load. Here, the circuit is effectively an LCR resonator and the dissipation is related to the quality factor, Q, by

$$E_{\text{diss}} = \frac{\pi}{8}\frac{C_L V^2}{Q} = \frac{\pi^2}{4} f R C_L^2 V^2 \tag{10}$$

where E_{diss} is the energy dissipated in a half cycle (a single transition), V is the peak-to-peak oscillation voltage, C_L is the load capacitance, and R is the effective resistance of two series transmission gates. For simplicity, the capacitance of the transmission gates and the dependence of R on voltage are neglected. As the equation shows, dissipation decreases as the frequency decreases—asymptotically approaching zero—even if the supply voltage is not reduced.

The preceding example always performs the same function and typifies single stage reversible combinational logic. Multistage reversible combinational logic, as does sequential logic, requires more complex waveforms, which can be created by methods proposed in [57], [58], [60], [62]. In general, these rules must be followed:

1) All logic transitions must be directly driven by a clock waveform passing through FETs instead of rippling through statically powered gates as in conventional logic.
2) The ramp rate of the clock waveforms must be low to save energy.
3) FETs should not be turned on while there is a voltage difference between source and drain since dissipation would otherwise result.

Under rule 3, to avoid dissipation, input signals should only be changed when the resonator is at its low point. This can be achieved if the inputs are driven out of phase with the clock. This relies upon the generation and synchronization of multiple clock phases, which can be efficiently derived from sinusoidal resonators using circuits such as proposed in [60].

Implementation of adiabatic CMOS, however, incurs an initial penalty (extra FETs, the need to uncompute information, the need to create complex control signals to drive every transition, etc.) as compared with conventional CMOS. This is shown in Fig. 24, which qualitatively indicates the performance versus energy trade-offs for conventional and adiabatic logic [60], [61]. High-performance applications (large operations/sec) suffer a substantial energy penalty for adiabatic circuits due to implementation overhead. At lower operations/sec, however, conventional CMOS hits a wall in minimum dissipation because, as can be inferred from Fig. 6, the optimal supply voltage cannot be reduced to below 0.3–0.4 V. On the other hand, adiabatic circuits do not hit this wall, so for applications that can tolerate low speeds, such approaches may be promising. Until solutions are found to dramatically reduce implementation overheads for reversible computing, initial applications may be limited to those at the extreme end of the power spectrum. Fig. 24 also shows a dashed curve for partially adiabatic CMOS. There has been much work in the past 15 years on techniques aimed at an intermediate regime between conventional CMOS and fully reversible circuits, in which some energy is saved by adiabatic switching and energy recovery into resonant supplies and some energy is dissipated conventionally [59], [62]. They are generally irreversible from a logical point of view but have fewer overhead penalties and thus provide

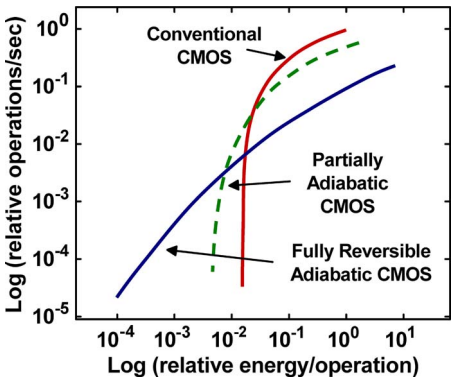


Fig. 24. *Qualitative comparison of conventional CMOS and adiabatic approaches. While likely limited in speed, adiabatic CMOS can hope to achieve dramatic improvements in energy per operation.*

a stepping stone towards fully reversible computing. Complete reversibility and energy recovery is unlikely to be implemented because it does not seem to be practical to entirely avoid data erasure (e.g., in memory), but dramatic improvements in energy dissipation may still be achievable through the judicious use of adiabatic techniques.

While the basic concept of reversible computing has been known for some time, the technique may prove to be more viable today due to recent trends in CMOS technology. With scaling, device and parasitic resistances and capacitances are reduced, which, with proper design, can yield circuits with sufficiently high quality factors to minimize energy loss. As parallelism gains general acceptance, the low operating frequency needed to reduce dissipation in adiabatic circuits can be compensated for by established techniques to achieve system performance targets. In addition, the severe power constraints that limit operating frequencies in conventional circuits (which, as described in Fig. 4, will return even with voltage scaling) are likely to be avoided in reversible circuits, which may reduce the effect of overheads associated with reversible computing.

VI. DISCUSSION

Section III presented the major issues in moderate voltage scaling—from the ~1 V supplies normally used today to ~0.5 V. Undoubtedly, the list of issues addressed is not exhaustive, but it does represent those that are well-understood and believed to be most important. Many related issues that need more investigation remain, including reliability and design tools. In addition, while system power is likely dominated by the digital circuits for parallel computation, further study is needed to identify the best means of integrating other essential functions, including single-thread performance and analog circuits.

With voltage scaling, many reliability mechanisms are mitigated due to a reduction in electric fields and current densities. However, radiation-induced single-event upsets caused by alpha particles and cosmic rays become more important as voltages are reduced since the critical charge (Q_{crit}) needed to upset a memory cell or latch is lowered. In addition, the need for parallel units to achieve desired performance levels could enhance fail rates for the overall system. Since memory cells are well-protected from both single and multibit fails by error correction codes and bit interleaving, latches instead tend to determine overall resiliency to soft errors [66]. The net quantitative impact of voltage scaling on soft error rates, however, requires further investigation in light of potential changes in device technology, since alternate device structures such as those as pictured in Fig. 14 greatly reduce charge collection volume. In addition, latch circuit sensitivities may change, since low-voltage functionality may require new topologies. While any increase in soft error rate is of concern, the magnitude of the increase as a result of voltage scaling may be in a range that can still be effectively controlled by device, circuit, and architecture techniques.

Design tools will likely need to be modified and recalibrated to realize the full benefit of voltage scaling. Device models developed for high-voltage operation are often not well-calibrated for lower voltages and may need reevaluation. With only moderate voltage scaling, it is unlikely that fundamental changes to standard cell libraries are needed, but timing tools, which are used to determine margins for variability and tolerances, will need adjustment since these issues change significantly with voltage scaling. As shown in Fig. 8, these concerns could be magnified at low voltage and thus require more accurate assessment to enable proper design optimization. In particular, design automation depends heavily upon these tools to generate efficient and robust circuits without overly conservative margins.

As stated in Section I, the implications of voltage scaling on single-thread performance have not been considered here. Applications that cannot be parallelized will depend strongly on the frequency of a single processor core—for which a high supply may be maintained. It might thus be optimal to build a hybrid system, whether as a heterogeneous parallel system with a single high-power/high-frequency core and many low-power/low-frequency cores or as a dynamically adjustable parallel system in which the voltage of one or more cores can be raised to improve single-thread performance. While the former option enables separate optimization of low-power and high-performance cores, the latter enables more flexible system organization. In either case, as only one of many parallel units, it can be expected that the power dissipated in this high-performance core does not dominate overall system power such that the techniques as outlined in this paper can still effectively improve power efficiency. Due to density scaling and a desire to continue increasing clock frequencies for this core, however, even with optimized total power, at some point, power density in this one core will become a concern as available cooling techniques may be inadequate. This suggests that density and frequency scaling for the high-performance core will eventually be limited.

Finally, while some analog circuits are present in most integrated circuits, power consumption is usually dominated by the digital circuits used for computation. As such, the discussion here has focused on these digital circuits, while it is assumed that analog circuits could, in the worst case, continue to operate at a higher voltage supply. Low-voltage operation carries a different set of constraints for analog circuits due to reduced voltage headroom, which can significantly degrade device transconductance, circuit gain, and signal-to-noise ratio, and limit the effectiveness of commonly used cascode circuit topologies that depend on stacked transistors. For these reasons, it may be difficult to scale analog voltages. Since many applications already

offer a separate analog supply, it could be straightforward to maintain larger voltage levels.

VII. CONCLUSION

Voltage scaling is the key to power efficiency. For parallelizable applications, the voltage in current CMOS technologies can be reduced by a moderate amount to yield a significant improvement in power without loss of system performance. Well-recognized concerns arise when voltages are reduced, but in the range of interest—from ~1 V to ~0.5 V, such issues can be contained without significant trade-offs. The potential improvements in power efficiency achievable through voltage scaling may be an important step towards the enablement of future exascale computing systems.

Such an advance, however, requires simultaneous improvements that span technology, circuits, and systems issues. In a low-voltage technology, the transistor structure itself should be optimized for extreme short-channel effect control to limit the influence of variability and leakage. Embedded memory functionality at low voltages, which can otherwise be limited by random statistical variation, can be achieved with modification of the core cell circuit or the integration of new, less variable device technologies. Issues such as on-chip digital noise and power delivery may scale suitably or, in fact, improve at low voltages. In the worst case, solutions such as high-voltage power delivery with on-chip conversion can mitigate future problems. At the system level, since voltage scaling improves the power efficiency of core computation, the power associated with off-chip connections must correspondingly be improved; here, low-voltage signaling can potentially provide significant gains, but alterations to the fundamental system organization may also be valuable.

The analysis presented here is by no means comprehensive, but is merely a summary of the most important issues based on current understanding. Some concerns, such as the detailed impact and importance of voltage scaling on soft-error rates, warrant further study. Even for the issues that have been discussed, however, there exists a wide range of consequence that depends upon the specific application in question. In general, the intensity of the need for power efficiency will drive expedition of the development of solutions. Trade-offs in area, performance, and ultimately monetary costs will determine the limits for each application.

The challenges associated with CMOS scaling will continue; the ideas proposed here do not forever solve power issues. Where conventional CMOS is limited, the future strategies discussed herein provide hope to continue to keep power dissipation at bay. Each carries its own set of new trade-offs, however, and acceptance may depend strongly on the evolution of end-user applications. Going forward, power efficiency will no doubt remain important. We can only hope to understand the limits and trade-offs associated with current and future technologies. ■

Acknowledgment

The authors would like to thank all of their colleagues at IBM for invaluable discussions, criticism, and guidance. In particular, the authors would like to acknowledge A. Gara for direction on Blue Gene trends, D. Friedman for advice on low-power serial links, D. Heidel for insight into soft error rate scaling issues, R. Nair for input on system-level parallelism effects, and G. Shahidi for management support. The authors would also like to thank E. Nowak and K. Rajamani for kindly providing the original data used in their previously published work.

REFERENCES

[1] G. E. Moore, "Progress in digital integrated electronics," in *IEDM Tech. Dig.*, 1975, pp. 11–13.

[2] R. H. Dennard, F. H. Gaensslen, H. N. Yu, 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. SC-9, pp. 256–268, Oct. 1974.

[3] 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, pp. 259–288, Mar. 2001.

[4] A. P. Chandrasakan, S. Sheng, and R. W. Broderson, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–483, Apr. 1992.

[5] M. Horowitz, T. Indermauer, and R. Gonzalez, "Low-power digital design," in *IEEE Symp. Low Power Electron.*, 1994, pp. 8–11.

[6] J. Meindl, "Low power microelectronics: Retrospect and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.

[7] M. Horowitz and W. Dally, "How scaling will change processor architecture," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2004, pp. 132–133.

[8] J. Tschanz, S. Narendra, Y. Ye, B. Bloechel, S. Borker, and V. De, "Dynamic-sleep transistor and body bias for active leakage power control of microprocessors," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2003, pp. 102–103.

[9] K. Rajamani, C. Lefurgy, S. Ghiasi, J. Rubio, H. Hanson, and T. Keller, "Power management solutions for computer systems and datacenters," in *Tutorial in Proc. Int. Symp. High-Performance Comput. Archit.*, 2008.

[10] G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized scaling theory and its application to a 1/4 micrometer MOSFET design," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 452–462, Apr. 1984.

[11] E. J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM J. Res. Dev.*, vol. 46, pp. 169–180, Mar./May 2002.

[12] R. H. Dennard, J. Cai, and A. Kumar, "A perspective on today's scaling challenges and possible future directions," *Solid-State Electron.*, vol. 51, pp. 518–525, Apr. 2007.

[13] D. J. Frank, W. Haensch, G. Shahidi, and O. Dokumaci, "Optimizing CMOS technology for maximum performance," *IBM J. Res. Dev.*, vol. 50, no. 4/5, pp. 419–431, Jul./Sep. 2006.

[14] C. Wann, R. Wong, D. J. Frank, R. Mann, S.-B. Ko, P. Croce, D. Lea, D. Hoyniak, Y.-M. Lee, J. Toomey, M. Weybright, and J. Sudijono, "SRAM cell design for stability methodology," in *Proc. IEEE VLSI-TSA Int. Symp. VLSI Technol.*, 2005, pp. 21–22.

[15] V. Zyuban and P. Strenski, "Unified method for resolving power-performance tradeoffs at the microarchitectural and circuit levels," in *Proc. Int. Symp. Low Power Electron. and Design*, 2002, pp. 166–171.

[16] G. M. Amdahl, "Validity of the single-processor approach to achieving large-scale computing capabilities," in *Proc. Amer. Federation Inf. Process. Soc. Conf.*, 1967, pp. 483–485, AFIPS Press.

[17] J. L. Gustafson, "Reevaluating Amdahl's law," *Commun. ACM*, vol. 31, pp. 532–533, May 1988.

[18] R. Gonzalez, B. M. Gordon, and M. A. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1210–1216, Aug. 1997.

[19] S. Hanson, B. Zhai, K. Berstein, D. Blaauw, A. Bryant, L. Chang, K. K. Das, W. Haensch, E. J. Nowak, and D. M. Sylvester, "Ultralow-voltage, minimum-energy CMOS," *IBM J. Res. Dev.*, pp. 469–490, Jul./Sep. 2006.

[20] D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFET's," *IEEE Electron Device Lett.*, vol. 19, pp. 385–387, Oct. 1998.

[21] R. W. Keyes, "Effect of randomness in the distribution of impurity ions on FET thresholds in integrated electronics," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 245–247, Aug. 1975.

[22] J. Pille, C. Adams, T. Christensen, S. Cottier, S. Ehrenreich, T. Kono, D. Nelson, O. Takahashi, S. Tokito, O. Torreiter, O. Wagner, and D. Wendel, "Implementation of the CELL broadband engine in a 65 nm SOI technology featuring dual-supply SRAM arrays supporting 6 GHz at 1.3 V," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2007, pp. 322–323.

[23] A. J. Bhavnagarwala, S. V. Kosonocky, S. P. Kowalczyk, R. V. Joshi, Y. H. Chan, U. Srinivasan, and J. K. Wadhwa, "A transregional CMOS SRAM with single, logicV_{DD} and dynamic power rails," in *Proc. Symp. VLSI Circuits Dig.*, 2004, pp. 292–293.

[24] M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De, "Wordline and bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65 nm CMOS designs," in *Proc. Symp. VLSI Circuits Dig.*, 2006, pp. 9–10.

[25] H. Pilo, J. Barwin, G. Braceras, C. Browning, S. Burns, J. Gabric, S. Lamphier, M. Miller, A. Roberts, and F. Towler, "An SRAM design in 65 nm and 45 nm technology nodes featuring read and write-assist circuits to expand operating voltage," in *Symp. VLSI Circuits Dig.*, 2006, pp. 15–16.

[26] L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek, "An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches," *IEEE J. Solid-State Circuits*, vol. 43, pp. 956–963, Apr. 2008.

[27] N. Verma and A. Chandrakasan, "A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy," *IEEE J. Solid-State Circuits*, vol. 43, pp. 141–149, Jan. 2008.

[28] Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto, "An area-conscious low-voltage-oriented 8T-SRAM design under DVS environment," in *Symp. VLSI Circuits Dig.*, 2007, pp. 256–257.

[29] L. Chang, Y.-K. Choi, D. Ha, P. Ranade, S. Xiong, J. Bokor, C. Hu, and T.-J. King, "Extremely scaled silicon nano-CMOS devices," *Proc. IEEE*, vol. 91, pp. 1860–1873, Nov. 2003.

[30] H. Kawasaki, M. Khater, M. Guillorn, N. Fuller, J. Chang, S. Kanakasabapathy, L. Chang, R. Muralidhar, K. Babich, Q. Yang, J. Ott, D. Klaus, E. Kratschmer, E. Sikorski, R. Miller, R. Viswanathan, Y. Zhang, J. Silverman, Q. Ouyang, A. Yagishita, M. Takayanagi, W. Haensch, and K. Ishimaru, "Demonstration of highly scaled FinFET SRAM cells with high-k/metal gate and investigation of characteristic variability for the 32 nm node and beyond," in *IEDM Tech. Dig.*, 2008, pp. 237–240.

[31] M. Guillorn, J. Chang, A. Bryant, N. Fuller, O. Dokumaci, X. Wang, J. Newbury, K. Babich, J. Ott, B. Haran, R. Yu, C. Lavoie, D. Klaus, Y. Zhang, E. Sikorski, W. Graham, B. To, M. Lofaro, J. Tornello, D. Koli, B. Yang, A. Pyzyna, D. Neumeyer, M. Khater, A. Yagishita, H. Kawasaki, and W. Haensch, "FinFET performance advantage at 22 nm: An AC perspective," in *Symp. VLSI Technol. Dig.*, 2008, pp. 12–13.

[32] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584–594, Apr. 1990.

[33] G. Patounakis, Y. W. Li, and K. L. Shepard, "A fully integrated on-chip DC-DC conversion and power management system," *IEEE J. Solid-State Circuits*, vol. 39, pp. 443–451, Mar. 2004.

[34] G. Wang, K. Cheng, H. Ho, J. Faltermeier, W. Kong, H. Kim, J. Cai, C. Tanner, K. McStay, K. Balasubramanyam, C. Pei, L. Ninomiya, X. Li, K. Winstel, D. Dobuzinsky, M. Naeem, R. Zhang, R. Deschner, M. J. Brodsky, S. Allen, J. Yates, Y. Feng, P. Marchetti, C. Noris, D. Casarotto, J. Benedict, A. Kniffin, D. Parise, B. Khan, J. Barth, P. Parries, T. Kirihata, J. Norum, and S. S. Iyer, "A 0.127μm^2 high performance 65 nm SOI-based embedded DRAM for on-processor applications," in *IEDM Tech. Dig.*, 2006.

[35] W. K. Luk and R. H. Dennard, "Gated diode amplifiers," *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 52, no. 5, pp. 266–270, May 2005.

[36] S. J. Koester, A. M. Young, R. R. Yu, S. Purushothaman, K.-N. Chen, D. C. La Tulipe, Jr., N. Rana, L. Shi, M. R. Wordeman, and E. J. Sprogis, "Wafer-level 3D integration technology," *IBM J. Res. Dev.*, vol. 52, pp. 583–597, 2008.

[37] P. S. Andry, C. K. Tsang, B. C. Webb, E. J. Sprogis, S. L. Wright, B. Dang, and D. G. Manzer, "Fabrication and characterization of robust through-silicon vias for silicon-carrier applications," *IBM J. Res. Dev.*, vol. 52, pp. 571–581, 2008.

[38] R. Palmer, J. Poulton, W. J. Dally, J. Eyles, A. M. Fuller, T. Greer, M. Horowitz, M. Kellam, F. Quan, and F. Zarkeshvari, "A 14 mW 6.25 Gb/s transceiver in 90 nm CMOS for serial chip-to-chip communications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2007, pp. 440–441.

[39] Y. Liu, B. Kim, T. O. Dickson, J. F. Bulzacchelli, and D. J. Friedman, "A 10 Gb/s compact low-power serial I/O with DFE-IIR equalization in 65 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2009, pp. 182–183.

[40] Y. Hidaka, W. Gai, T. Horie, J. H. Jiang, Y. Koyanagi, and H. Osone, "A 4-channel 10.3 Gb/s backplane transceiver macro with 35 dB equalizer and sign-based zero-forcing adaptive control," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2009, pp. 188–189.

[41] J. F. Bulzacchelli, T. O. Dickson, Z. T. Deniz, H. A. Ainspan, B. D. Parker, M. P. Beakes, S. V. Rylov, and D. J. Friedman, "A 78 mW 11.1 Gb/s 5-tap DFE receiver with digitally calibrated current-integrating summers in 65 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2009, pp. 368–369.

[42] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, pp. 728–749, Jun. 2000.

[43] Joint Electron Devices Engineering Council. [Online]. Available: http://www.jedec.org/

[44] "Blue Gene," *IBM J. Res. Dev.*, vol. 49, Mar./May 2005, entire issue.

[45] IBM Blue Gene Team, Overview of the IBM Blue Gene/P project," *IBM J. Res. Dev.*, vol. 52, pp. 199–220, Jan./Mar. 2008.

[46] The Green500 List. [Online]. Available: http://www.green500.org/

[47] TOP500 Supercomputing Sites. [Online]. Available: http://www.top500.org/

[48] I. Yang, C. Vieri, A. Chandrakasan, and D. Antoniadis, "Backgated CMOS on SOIAS for dynamic threshold voltage control," *IEEE Trans. Electron Devices*, vol. 44, pp. 822–831, May 1997.

[49] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nanoletters*, vol. 8, pp. 405–410, Feb. 2008.

[50] K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, "I-MOS: A novel semiconductor device with a subthreshold slope lower than kT/q," in *IEDM Tech. Dig.*, 2002, pp. 289–292.

[51] A. Padilla, C. W. Yeung, C. Shin, C. Hu, and T.-J. K. Liu, "Feedback FET: A novel transistor exhibiting steep switching behavior at low bias voltages," in *IEDM Tech. Dig.*, 2008, pp. 171–174.

[52] W. Y. Choi, B.-G. Park, J. D. Lee, and T.-J. King Liu, "Tunneling field-effect transistors (TFETs) with subthreshold swing (SS) less than 60 mV/dec," *IEEE Electron Device Lett.*, vol. 28, pp. 743–745, Aug. 2007.

[53] O. M. Nayfeh, C. N. Chleirigh, J. Hennessy, L. Gomez, J. L. Hoyt, and D. A. Antoniadis, "Design of tunneling field-effect transistors using strained-silicon/strained-germanium type-II staggered heterojunctions," *IEEE Electron Device Lett.*, vol. 29, pp. 1074–1077, Sep. 2008.

[54] C. L. Seitz, A. H. Frey, S. Mattisson, S. D. Rabin, D. A. Speck, and J. L. A. van de Snepscheut, "Hot-clock nMOS," in *Proc. 1985 Chapel Hill Conf. VLSI*, 1985, pp. 1–17.

[55] J. S. Hall, "An electroid switching model for reversible computer architectures," in *Proc. Workshop Physics Comput.*, 1992, pp. 237–247.

[56] J. G. Koller and W. C. Athas, "Adiabatic switching, low energy computing, and physics of storing and erasing information," in *Proc. Workshop Physics Comput.*, 1992, pp. 267–270.

[57] S. G. Younis and T. F. Knight, Jr., "Practical implementation of charge recovering asymptotically zero power CMOS," in *Proc. 1993 Symp. Integr. Syst.*, 1993, pp. 234–250.

[58] P. Solomon and D. J. Frank, "The case for reversible computation," in *Proc. Int. Workshop Low Power Design*, 1994, pp. 93–98.

[59] J. S. Denker, S. C. Avery, A. G. Dickinson, A. Kramer, and T. R. Wik, "Adiabatic computing with the 2N-2N2D logic family," in *Proc. Int. Workshop Low Power Design*, 1994, pp. 183–187.

[60] D. J. Frank and P. M. Solomon, "Electroid-oriented adiabatic switching circuits," in *Proc. Int. Workshop Low Power Design*, 1995, pp. 197–202.

[61] D. J. Frank, "Comparison of high speed voltage-scaled conventional and adiabatic circuits," in *Proc. Int. Workshop Low Power Electron. Design*, 1996, pp. 377–380.

[62] W. Athas, "Practical considerations of clock-powered logic," in *Proc. Int. Workshop Low Power Electron. Design*, 2000, pp. 173–178.

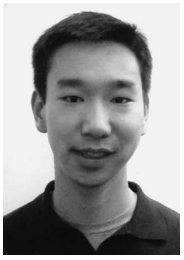
[63] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Dev.*, vol. 5, pp. 183–191, 1961.

[64] C. H. Bennett, "Logical reversibility of computation," *IBM J. Res. Dev.*, vol. 6, pp. 525–532, 1973.

[65] M. P. Frank, "Introduction to reversible computing: Motivation, progress, and challenges," in *Proc. 2nd Conf. Comput. Frontiers*, 2005, pp. 385–390.

[66] D. F. Heidel, K. P. Rodbell, E. H. Cannon, C. Cabral, Jr., M. S. Gordon, P. Oldiges, and H. H. K. Tang, "Alpha-particle-induced upsets in advanced CMOS circuits and technology," *IBM J. Res. Dev.*, vol. 52, pp. 225–232, 2008.

ABOUT THE AUTHORS



Leland Chang received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer sciences in 1999, 2001, and 2003, respectively, from the University of California, Berkeley.

He joined the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, in 2003 and is now manager of Design and Technology Solutions. His research has spanned topics ranging from silicon CMOS technology and circuits to nonvolatile memory and RF MEMS. His early efforts focused on thin-body MOSFETs for CMOS scaling, including early demonstration of the double-gate FinFET structure. More recently, he has pursued scaling issues for embedded memory, including SRAM cell scaling to record sizes and the demonstration of an 8T-SRAM cell for variability tolerance and low-voltage operation. His current work focuses on new technologies for power efficiency in high-performance systems. He has authored or coauthored more than 50 technical articles and holds 10 patents.

David J. Frank (Fellow, IEEE) received the B.S. degree from the California Institute of Technology, Pasadena, in 1977 and the Ph.D. degree in physics from Harvard University, Cambridge, MA, in 1983.

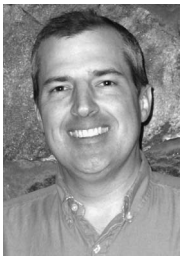
Since graduation he has been employed at the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he is a Research Staff Member. His studies have included nonequilibrium superconductivity, III-V devices, and exploring the limits of scaling of silicon technology. His recent work includes the modeling of innovative Si devices, analysis of CMOS scaling issues such as power consumption, discrete dopant effects, and short-channel effects associated with high-k gate insulators, exploring various nanotechnologies, investigating the usefulness of energy-recovering CMOS logic and reversible computing concepts, and low-power circuit design. He has authored or coauthored over 100 technical publications and holds 11 U.S. patents.

Dr. Frank has served as chairman of the Si Nanoelectronics Workshop and as associate editor of IEEE TRANSACTIONS ON NANOTECHNOLOGY.

Robert K. Montoye (Senior Member, IEEE) received the B.S. degree in physics and the M.S. and Ph.D. degrees in computer science from the University of Illinois, Urbana.

Joining IBM in 1983, he designed and implemented the RS/6000 floating-point unit. After pursuing interests outside IBM from 1990 to 1995, he returned to IBM to focus on finding a lower supply circuit family with state-of-the-art performance and its impact on overall microarchitecture and architecture. He is currently with the IBM T. J. Watson Research Center, Yorktown Heights, NY. He has published 25 technical papers and holds more than 50 patents.

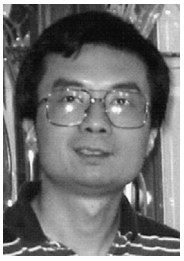
Dr. Montoye is a member of the IBM Academy of Technology.



Steven J. Koester (Senior Member, IEEE) received the B.S.E.E. and M.S.E.E. degrees from the University of Notre Dame, Notre Dame, IN, in 1989 and 1991, respectively, and the Ph.D. degree from the University of California, Santa Barbara, in 1995.

He is a Research Staff Member and Manager of the Exploratory Technology group at the IBM T. J. Watson Research Center, Yorktown Heights, NY. He joined IBM in 1995, and has performed research on a wide variety of semiconductor materials and devices, including SiGe and strained Si FETs, III-V MOSFETs, scaled CMOS integration, group-IV photonic devices, and three-dimensional integration. He has authored or coauthored over 125 technical publications and conference presentations and holds 32 U.S. patents.

Dr. Koester is a member of the AAAS and Tau Beta Pi. He was recently the General Chair for the Device Research Conference in 2009.



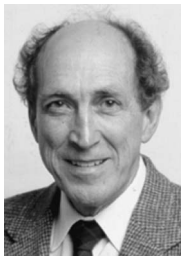
Brian L. Ji received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1984 and the Ph.D. degree in physics from Harvard University, Cambridge, MA, in 1991.

From 1991 to 1994, he was a Research Scientist at SUNY at Stony Brook, where he studied nanofabrication, single electron memory, and superconducting devices. He was a Visiting Scientist in physics at IBM in 1995. From 1996 to 2005, he worked at IBM for the developments of semiconductor memory and high speed serial link products. From 2005 to 2009, he was a Research Staff Member at IBM T. J. Watson Research Center, researching in silicon devices, statistical variation, power conversion, and exploratory circuit designs. He is currently an independent scientist with research interests in statistical dynamics for physics and economic analysis. He holds 21 U.S. patents.

Paul W. Coteus (Senior Member, IEEE) received the Ph.D. degree in physics from Columbia University, New York, in 1981.

He continued at Columbia to design an electron–proton collider, and spent from 1982 to 1988 as Assistant Professor of Physics at the University of Colorado, Boulder, studying neutron production of charmed baryons. In 1988 he joined the IBM T. J. Watson Research Center, Yorktown Heights, NY, as Research Staff Member. He has managed the Systems Packaging Group since 1994, where he directs and designs advanced packaging for high-speed electronics, including I/O circuits, memory system design and standardization of high-speed DRAM, and high-performance system packaging. He currently leads the Blue Gene future systems packaging team. He has authored more than 90 papers in the field of electronic packaging, and holds 79 U.S. patents.

Dr. Coteus is a member of IBM's Academy of Technology, and an IBM Master Inventor. He was Chairman in 2001 and Vice-chairman from 1998–2000 of the highly influential JEDEC Future DRAM Task Group. He led the system design and packaging of the Blue Gene family of Supercomputers, recently honored with the National Medal of Technology and Innovation.



Robert H. Dennard (Fellow, IEEE) was born in Terrell, TX, in 1932. He received the B.S. and M.S. degrees in electrical engineering from Southern Methodist University, Dallas, TX, in 1954 and 1956 respectively, and the Ph.D. Degree from Carnegie Institute of Technology, Pittsburgh, PA, in 1958.

He then joined the IBM Research Division, where his early experience included the study of new digital devices and circuits for logic and memory applications, and the development of advanced data communication techniques. Since 1963, Dr. Dennard has been at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, where he has been involved in microelectronics research and development from the early days onward. His primary work has been in MOS transistors and integrated digital circuits using them. In 1967 he invented the dynamic RAM memory cell used in most all computers today. With coworkers he developed the concept of MOS transistor scaling in 1972, which is often cited as a guiding principle for microelectronics. He has contributed numerous papers on advances in CMOS technology and on prospects and challenges of scaling that technology to very small dimensions.

Dr. Dennard received from the IEEE the Cledo Brunetti Award in 1982, the Edison Medal in 2001, and the Medal of Honor in 2009. He is a member of the National Academy of Engineering and the American Philosophical Society, and he has received many honors, including the National Medal of Technology in 1988 and induction into the National Inventors Hall of Fame in 1997. He was appointed an IBM Fellow in 1979.

Wilfried Haensch (Senior Member, IEEE) received the Ph.D. degree in theoretical solid-state physics from the Technical University of Berlin, Germany, in 1981.

He started his career in Si technology in 1984 at Siemens Corporate Research Munich. There he worked on high field transport in MOSFETs. In 1990 he joined the DRAM alliance between IBM and Siemens to develop quarter micrometer 64 M DRAM. From there he moved in 1996 to Infineon's manufacturing facility in Richmond, VA, to be involved in the production of various generations of DRAM. In 2001 he joined IBM T. J. Watson Research Center, Yorktown Heights, NY, to lead a group for novel devices and applications. He is currently responsible for the exploration of device concepts for 15 nm node and beyond, and new scaling concepts for memory and logic circuits. He is the author of a text book on transport physics and author/coauthor of more than 100 publications.

Dr. Haensch was awarded the Otto Hahn Medal for outstanding research in 1983.