Optimizing CMOS technology for maximum performance

D. J. Frank
W. Haensch
G. Shahidi
O. H. Dokumaci

Since power dissipation is becoming a dominant limitation on the continued improvement of CMOS technology, technologists must understand the best way to design transistors in the presence of power constraints. The primary objective is to obtain as much performance as possible for a fixed amount of power, and it is chip performance, not device performance, that matters. In order to investigate this regime, we have captured in simplified models the basic elements for determining chip performance, including intrinsic transistor characteristics, circuit delay, tolerance issues, basic microprocessor composition, and power dissipation and heat removal considerations. These models have been assembled in a processor-level technology-optimization program to study the characteristics of optimal technology across many generations of CMOS. The results that are presented elucidate the limits of future CMOS technology improvements, the optimal energy consumption conditions, and the relative benefits of various proposed technology enhancements, including high-k gate insulators, metal gates, high-mobility semiconductors, improved heat removal, and the use of multiple layers of circuitry.

Introduction

For the past several decades, the semiconductor industry has relied on a progression of smaller, denser, faster, cheaper MOSFETs to provide increasingly better products for digital electronics. This process of shrinking CMOS transistors in order to attain these improvements is known as scaling, and its progress is often characterized by measuring the device speed. As CMOS scaling continues, however, it is increasingly important to analyze potential technology design points for their impact on overall chip performance, and not just for their impact on device speed, because chip-level power constraints as well as device and process variability can seriously diminish the value of device innovations. The high cost of developing new technology options also makes it vital to gain an early understanding of their potential benefit to the final products, so that developments with little benefit can be avoided. This paper describes a high-level technology optimization tool, and the results of using it to perform chip-level analyses of potential technology options for the 45-nm and 32-nm generations. These options include enhanced mobility, high-permittivity gate-insulating materials ("high-*k*"), metal gate workfunctions, and thermal solutions.

Most prior work in this area has focused on system-level power and performance estimation for extrapolating the behavior of future technologies, using estimated critical paths and fairly detailed system descriptions to determine clock frequency, often with much attention paid to the nature and use of the wiring hierarchy. Early examples of prior work are described in [1–3]. Second-generation modeling systems have included GENESYS [4, 5], RIPE [6], BACPAC [7], and, more recently, GTX [8]. Although some of these models are quite detailed at the system level, and optimize various aspects of the wiring and its usage, they have not generally sought to optimize the device technology, preferring to treat information on device technology as a user input. J. D.

©**Copyright** 2006 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/06/$5.00 © 2006 IBM

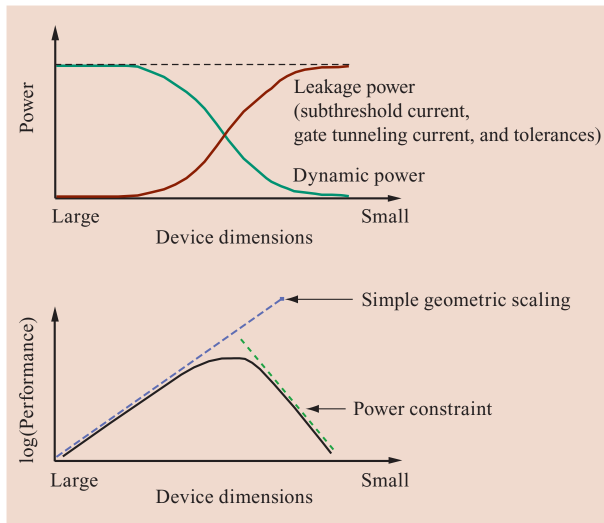

Figure 1

Schematic illustration of the existence of an optimal device miniaturization with maximal processing performance.

Meindl et al. have performed a system-level analysis of the limits to scaling for both devices and wires, looking at limits that are caused by a wide range of physical effects [9]. Threshold- and supply-voltage optimization has frequently been studied as a means to reduce power dissipation (see for example [10, 11]), while other work has focused on minimizing power for a fixed performance, subject to various constraints [5, 12, 13]. Recently it has been argued [14, 15] that power should be considered as the primary constraint that determines how far technology can scale, and that the different power requirements associated with different applications result in different limits to scaling. These prior studies formed the basis for an interdisciplinary CMOS design space study [16], in which many aspects of CMOS design were combined into a single model for optimizing CMOS device and wiring technology, with an emphasis on new device scaling aspects.

The work described in this paper builds on the previous design space studies by adding an improved, calibrated device model, accounting for on-chip tolerance issues, accurately capturing the area and power allocations of real processor chips, and implementing detailed temperature dependences and heat-sink models. This tool is believed to provide the best available analysis of the relative utility of proposed technology options because it attempts to "self-consistently" optimize the devices themselves.

The next section describes the overall optimization approach, followed by an explanation of some of the details of the models. Results of the optimizations are discussed.

Optimization methodology

If Moore's law [17] could provide a path for unbounded progress, as some projections have implied, our goal of seeking an optimum CMOS technology would not be meaningful. The reality, however, is that CMOS technology is bounded. The existence of an optimum technology is illustrated schematically in **Figure 1**. When device dimensions are large and threshold voltages are high (at the left side of the curves), dissipation caused by leakage currents can be low. The shrinking of dimensions reduces capacitance and enables increasing performance at fixed power. However, device scaling eventually leads to increasing leakage current, due to quantum-mechanical tunneling and subthreshold current. When the total power is constrained, this leakage dissipation ultimately dominates the power consumption, leaving very little power left over for active circuit switching, which leads to a loss of overall performance beyond some point in the scaling process, even though the devices may be getting faster (the right side of the curves). The height and position of the maximum performance-versus-scaling curve depends on the power constraint and other system conditions, but a maximum does exist. This performance-versus-scaling situation applies to cases in which the power constraint is associated with a roughly fixed degree of architectural complexity. This argument for the existence of an optimum should generally apply to all computational electronics, because it depends only on some very broad features of device physics: 1) electrostatics imposes geometric constraints on the relative device dimensions; 2) thermodynamics imposes constraints on voltage reduction; 3) quantum-mechanical tunneling effects inevitably cause exponentially increasing leakage currents when dimensions are sufficiently reduced; and 4) various practical considerations limit power dissipation.

An optimization tool has been developed to determine the technology parameters that lead to this optimum performance for CMOS technology. The program involves a collection of models that span the material, device, circuit, and system levels, some aspects of which are described in more detail below [9, 16]. The overall structure of the optimization tool is shown schematically in **Figure 2**. The goal is to find the values of the device technology parameters that will result in the greatest possible processor performance for a given power level. We have chosen to measure performance in terms of a total logic net transition rate, LTR. This is the total number of state changes per second for all of the logic nets in the processor core(s) combined. We have deliberately chosen this metric, rather than a critical path-

delay metric, because it allows a substantial degree of independence from the architectural details, making our results more generally useful for changing architectures. This metric relies on the expectation that the rate at which useful instructions can be executed by the processor will monotonically increase with LTR. An alternate optimization metric also was considered in [16], based on total computation received per dollar spent (on both chip and energy) over the expected life of the chip. Because it was shown in [16] that both optimization approaches give similar results, we have focused on maximizing LTR at fixed power in this paper.

As in [16], to reduce complexity and narrow our focus, only the logic devices in the processor core are actually optimized; it is assumed that the power and speed of the clock and latch circuits, registers, memories, and I/O can ultimately all be optimized with essentially the same power/performance result as the logic part of the processor. To achieve accurate results at the chip level, we can use actual chip data to set the processor core complexity, and use the fraction of the chip area and power that is used for logic. The optimization controls the actual size of the chip, and hence the power density, by adjusting the device and wire sizes.

Because memory actually occupies the majority of the chip in modern processors, it may seem unreasonable not to include it in the optimizations. However, we have not done so because of the previously mentioned observation that different applications must be separately optimized [14]. Memory has very different requirements than logic, which lead to optima that are quite different from those for logic. The best system performance can certainly be obtained by creating a technology that offers different, separately optimized devices (and voltages) for memory and for logic. If economic considerations force memory and logic to use the same devices, optimization across both sets of requirements might very well find different results than those reported here, which are for logic by itself.

The basic optimization methodology starts with definitions for power and delay as functions of the underlying technology parameters. In an inner programming loop, one degree of freedom (usually the supply voltage, V_{DD}) is used to satisfy the power constraint, and then in the outer loop, the remaining variables are optimized to find the maximum possible performance.

The total power (P_{TOT}) calculation includes dynamic switching power (P_{DYN}), power due to subthreshold leakage current (P_{subVT}), power due to gate oxide tunneling current (P_{ox}), and power due to drain-to-body tunneling current (P_{B2B}), as defined in the following equations:

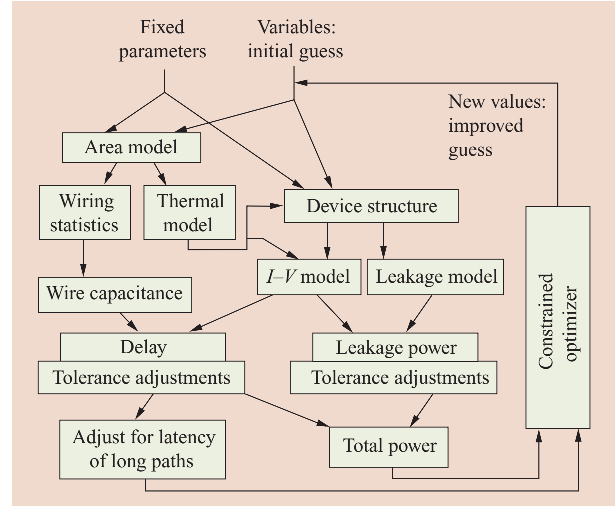

Figure 2

Schematic structure of optimization tool.

$$P_{TOT} = (P_{DYN} + P_{subVT} + P_{OX} + P_{B2B})_{logic} + (P_{DYN} + P_{subVT} + P_{OX} + P_{B2B})_{repeaters}, \quad (1)$$

$$P_{DYN} = \alpha_S N_{CKT} \frac{1}{2} \langle C \rangle (V_H - V_L) V_{DD} \frac{1}{\beta_\tau \ell_D \tau}, \quad (2)$$

$$P_{subVT} = 1.7 \beta_{Ioff} \cdot N_{CKT} \cdot V_{DD} \cdot W \cdot J_{off}(V_T, V_{DD}, t_{ox}, \eta, L_{CH}), \quad (3)$$

$$P_{ox} = N_{CKT} \cdot FI \cdot V_{DD} \cdot L_G \cdot W \cdot J_{ox}(V_T, V_{DD}, t_{ox}, \eta), \quad (4)$$

$$P_{B2B} = N_{CKT} \cdot FI \cdot V_{DD} \cdot \frac{L_G}{2} \cdot W \cdot J_{B2B}(F_{Max}, V_{DD}), \quad (5)$$

where V_{DD}, V_H, V_L, and V_T are the supply voltage, high-logic-level, low-logic-level, and threshold voltage, respectively, and τ is the average switching delay of a loaded logic stage (a NAND gate, with average fan-in, FI, usually set to 2 and average fan-out of 1.65). N_{CKT} is the number of logic gates, $\langle C \rangle$ is the average total load capacitance, α_S is the switching activity factor, ℓ_D is the logic depth, J_{off} is off-current density (at V_L) of a typical logic FET (see next section), J_{ox} is the oxide tunneling current density (at V_H) [15], and J_{B2B} is the band-to-band tunneling current density from drain to body (at V_H, and using junction area $\frac{1}{2} L_G W$) [15, 18]. t_{ox} is the oxide thickness, η is the subthreshold ideality, W is the average FET width, L_G and L_{CH} are the gate length and channel length, F_{Max} is the peak field in the body–drain junction, which depends on the doping and the voltage, and β_τ and

$\beta_{I\text{off}}$ are tolerance-related correction factors. The power contributions are separately computed for logic and for the buffers that are placed in long wires (i.e., repeaters).

The performance metric, LTR, is computed from the delay as

$$LTR = \alpha_S N_{CKT} \frac{1}{\beta_\tau \ell_D \tau} \frac{1}{CPI_{\text{eff}}}, \quad (6)$$

where CPI_{eff} is a correction factor to take into account the impact of long wires and their repeaters, described in a later section. The loaded logic delay computation proceeds in steps as in [5]. First, the basic device delay is computed using a modified CV/I form that has been shown to accurately take into account output conductance effects [19]:

$$\tau_1 = \frac{V_{DD}(C_{\text{parasitic}} + C_{\text{wire}} + C_{\text{gateload}})}{2I_{\text{eff}}}, \quad (7)$$

where $I_{\text{eff}} = \frac{1}{2}[I_{DS}(V_{DD}, \frac{1}{2}V_{DD}) + I_{DS}(\frac{1}{2}V_{DD}, V_{DD})]$, the Cs are average capacitances, as indicated by their subscripts, and $I_{DS}(V_{DS}, V_{GS})$ is drain current as a function of drain and gate voltages, which is described in the next section. Next, the wire RC and time-of-flight delays are computed, and combined using an empirical formula [3, 5]:

$$\tau_2 = R_{\text{wire}}\left(\frac{1}{2}C_{\text{wire}} + C_{\text{gateload}}\right), \quad (8)$$

$$\tau_3 = L_{\text{wire}}/(c/2), \quad (9)$$

$$\tau_4 = (\tau_2^{4/3} + \tau_3^{4/3})^{3/4}, \quad (10)$$

where R_{wire} is the temperature-dependent wire resistance, L_{wire} is the average wire length, and c is the speed of light. Finally, these delays are combined and divided by a rise-time correction factor due to Sakurai and Newton [19]:

$$\tau = \frac{\tau_1 + \tau_4}{0.5 + (1 - V_T/V_{DD})/(1+\alpha)}, \quad (11)$$

where α is the power-law exponent described in the next section.

Model details

Device current–voltage model, calibration, and technology generations

The structural portion of the device model assumes bulk or partially depleted silicon-on-insulator (PD-SOI) FETs, and uses the effective doping, N_{eff}, in conjunction with a first-order analytic 2D Poisson solution to determine V_T, η, and DIBL (drain-induced barrier lowering). This model yields continuous, physically realistic device characteristics for all gate lengths, from punch-through to long-channel. Because the Poisson solution depends on L_{CH}, gate insulator thickness and material, and body doping, we have fully captured the underlying technology dependencies in a very general way. The current–voltage model is a generalization of Sakurai's alpha power-law model [19], in which we use the Fermi–Dirac function $F_{\alpha-1}$ to achieve a smooth transition between an α power law above V_T and an appropriate subthreshold exponential tail so that the same model can be used for both ON and OFF currents. The intrinsic saturation current is given by

$$I_D(V_{GS}) = \frac{W\varepsilon_I}{t_{ox}^{\text{eff}}} \frac{\eta kT}{e} \left(\frac{\eta kT/e}{FI \cdot E_C L_{CH}}\right)^\beta \mu_0 \left(\frac{\mu_x \mu(E_\perp, T)}{\mu_0}\right)^s \times E_C F_{\alpha-1}\left(\frac{V_{GS} - V_T}{\eta kT/e}\right), \quad (12)$$

where V_{GS} is the gate-to-source voltage, ε_I is the gate insulator permittivity, and t_{ox}^{eff} is the effective thickness of the gate insulator, including quantum effects and poly-Si depletion. e is the electronic charge, E_C is the characteristic field in the velocity–field relationship, μ_0 is a calibration parameter with units of mobility, β and s are exponents fitted to available data, $\mu(E_\perp, T)$ is the universal mobility curve [20] as a function of the effective perpendicular field and the temperature, and μ_x is a mobility enhancement factor used to account for technologies that improve mobility. V_{DS} dependence is accommodated by means of a DIBL adjustment to V_T. Source and drain resistance, R_{CS}, is included by using a numerical iteration to self-consistently adjust V_{GS} and V_{DS} to account for the extrinsic voltage drops. **Table 1** gives the values used for the constant-valued parameters.

Halo doping is a fabrication process in which body dopants are implanted at angles from both the source and drain side of a FET. This is very useful because it causes the effective doping, N_{eff}, in the channel of a MOSFET to increase when the gate length becomes shorter, which tends to compensate for the electrostatic short-channel effects. This is accounted for in our model through the fitting function

$$N_{\text{eff}} = N_D \frac{(L_{CH}/x_e)^{n_1}}{1 + (L_{CH}/x_e)^{n_2}}, \quad (13)$$

where N_D sets the doping magnitude, n_1 and n_2 are fitting exponents, and the parameter x_e sets the channel length scale over which N_{eff} varies. x_e should be related to the characteristic length scale of the halo-doping profile. This is one of the parameters that must improve from generation to generation in order for scaling to proceed. The source/drain doping usually causes overlap between the gate and the source and drain, so the channel length is offset from the gate length ($L_{CH} = L_G - x_{\text{ovlp}}$) by an overlap distance, x_{ovlp}, that must also decrease as technology improves.

Table 1 Values for various constant model parameters.

Description	*Symbol*	*Value*
Activity factor over logic depth	α_S/ℓ_D	0.012
I–V curve power law	α	1.462
I–V formula gate-length exponent	β	0.405
I–V formula mobility exponent	s	0.430
Mobility calibration parameter	μ_0	132.3 cm^2/V-s
Critical field	E_C	2.5×10^4 V/cm
Halo exponent 1	n_1	−0.574
Halo exponent 2	n_2	2.18
Maximum logic depth	ℓ_{Dmx}	10
Number of logic stages in typical instruction	n_{LI}	60
Latency penalty weighting factor	γ	0.1

The model is calibrated to 2D drift/diffusion FIELDAY (FInite ELement Device Analysis) [21] simulations at several technology nodes, as shown in **Figure 3**. The correlation between full 2D device simulations and our simple compact model is excellent, considering that only 14 fitting parameters are used to match this entire set of data, which includes three different gate lengths at both the 90-nm and 45-nm technology nodes. The values of most of these parameters are included in Tables 1 and 2.

On the basis of fits to FIELDAY simulations and ITRS roadmap considerations, the set of adjustable parameters for the FET model were chosen for each technology node, as shown in **Table 2**.[1] These are the parameters that are fixed for each node, and are thought of as the best that technology will be capable of at that node. The gate length, oxide thickness, and voltages are *not* fixed by node, but rather are determined by optimization.

Tolerance modeling

The following within-chip tolerances are included in the analysis: discrete dopant V_T variation, random gate-length variation due to line-edge roughness (LER), across-chip gate-length variation (ACLV), V_{DD} variations, and signal coupling noise. The model estimates the impact of these variations on the average subthreshold leakage current and on the worst-case delay.

[1]The International Technology Roadmap for Semiconductors (ITRS) is an assessment of semiconductor technology requirements and is a cooperative effort of manufacturers, suppliers, government organizations, and universities.

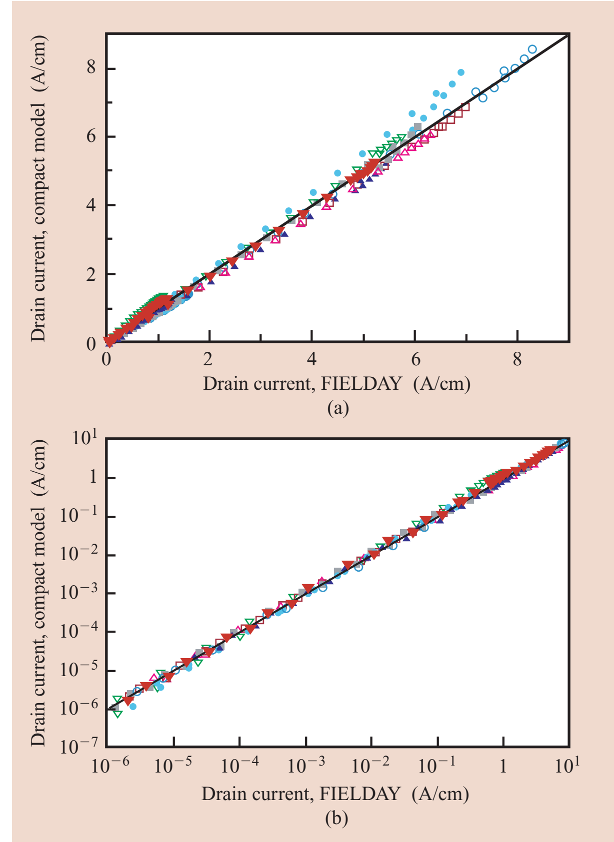

Figure 3

(a) Linear and (b) log plots of FET drain current, showing the correlation between our compact model and the FIELDAY simulations for *I–V* curves for six different FETs. Each point is a different bias condition, and the six different colored symbols correspond to the six different FETs.

It also checks that 6σ V_T and noise shifts do not cause an individual NAND gate to fail, although this is not usually a problem.

The impact on subthreshold leakage current is estimated by observing that the doping variations, gate-length variations, and noise combine to create an approximately Gaussian distribution, ρ, of equivalent threshold voltage with sigma, σ_{VTeff}. When this Gaussian distribution is integrated against the exponential off-current dependence, it yields an average shift [14]:

$$I_{ave} = \int_{-\infty}^{\infty} \rho(V_T) I_{off}(V_T) dV_T = I_{off,nom} \exp\left[+ \frac{\sigma^2_{VTeff}}{2(\eta k T)^2} \right]. \quad (14)$$

Table 2 Fixed technology parameters that vary by node.

Description	*Symbol*	*Technology node* (nm)				
		130	*90*	*65*	*45*	*32*
Wire 1/2 pitch (nm)		175	120	90	70	50
Gate overlap (nm)	x_{ovlp}	19	8.74	3	−1.03	−1.5
Halo scale length (nm)	x_{e}	61.5	53.4	46.3	40.2	34.9
Contact resistance (Ω-cm)	R_{CS}	0.0129	0.0136	0.0144	0.0152	0.0161
LER sigma L_{G} @$W = 1$ μm (nm)		0.28	0.28	0.28	0.28	0.28
ACLV (nm)		3.9	2.7	1.8	1.3	1
Mobility enhancement factor	μ_x	1	1.4	1.7	2	2
Gate depletion (nm)		0.4	0.4	0.3	0.3	0.01
Wiring permittivity	k_{wiring}	3.9	3.5	3.2	2.8	2.5
Permittivity (gate insulator)	ε_{I}	4.7 (oxynitride)	4.7 (oxynitride)	4.7 (oxynitride)	4.7 (oxynitride)	20 (HfO_2)

Thus, the background leakage current increases by the factor

$$\beta_{I\text{off}} = \exp\left[+\frac{\sigma^2_{V\text{Teff}}}{2(\eta kT)^2}\right], \qquad (15)$$

and so does the static power dissipation. If $\sigma_{V\text{Teff}}$ exceeds kT, this factor can become quite large.

The impact of random variations on worst-case delay is estimated on the basis of the following analysis. Each path i in the set of all paths has a distribution $\rho_i(t)$ of worst-case delays, where "worst case" means the longest possible delay that can occur due to any conceivable instruction sequence (i.e., due to worst-case signal coupling and supply noise). The distribution is over random intrinsic device variations (e.g., variations in V_{T}). The distribution should also include across-chip variation. Across-chip variations probably have correlations, but to a first approximation we may treat them as independently random and consider them as part of the intrinsic variations.

Thus, the probability of path i failing (by taking too long because its delay is longer than the clock time) is $q_i = \int_{t_{\text{CK}}}^{\infty} \rho_i(t)dt$, where t_{CK} is the desired clock delay. Then, the yield for the whole chip is $Y = \Pi\,(1 - q_i)$, where the product is over all independent paths. Each path is considered as a set of stages, so that $\tau_i = \sum_{j=1}^{n_i} \tau_{ij}$, summing over the n_i stages of the ith path. Next, approximate $\langle\tau_{ij}\rangle$ as τ_0, the nominal value of the delay, and assume that the distribution of τ_{ij} is Gaussian and can be characterized by a σ_τ that can be estimated numerically by using a worst-case vector of variations away from the nominal case. Then, set $\tau_i = n_i\tau_0$ and $\sigma_{\tau_i} = \sqrt{n_i}\sigma_\tau$. Now,

$$q_i = \int_{t_{\text{CK}}}^{\infty} \frac{1}{\sqrt{2\pi n_i}\sigma_\tau} \exp\left[-\frac{(t - n_i\tau_0)^2}{2n_i\sigma^2_\tau}\right] dt$$

$$\approx \frac{1}{\sqrt{2\pi}}\left(\frac{1}{u_i} - \frac{1}{u_i^3}\right) e^{-u_i^2/2}, \qquad (16)$$

where

$$u_i = \frac{t_{\text{CK}} - n_i\tau_0}{\sqrt{n_i}\sigma_\tau}\ .$$

Next, treat n_i as a continuous variable, and let $P(n)$ be the density of effectively independent paths. Then,

$$\ln Y = \sum_i \ln(1 - q_i) \approx -\sum_i q_i$$

$$\approx -\int_1^{\ell_{\text{Dmx}}} \frac{1}{\sqrt{2\pi}}\left[\frac{1}{u(n)} - \frac{1}{u^3(n)}\right] e^{-u^2(n)/2} P(n)dn, \qquad (17)$$

where ℓ_{Dmx} is the maximum logic depth. If the $P(n)$ density function can be treated as trapezoidal, then to first order only the value at ℓ_{Dmx} matters, and Equation (16) may be approximately evaluated as

$$Y \approx \exp\left\{-\frac{1}{\sqrt{2\pi}}\left[\frac{1}{u^2(\ell_{\text{Dmx}})} - \frac{3}{u^4(\ell_{\text{Dmx}})}\right]\right.$$

$$\left.\times \frac{e^{-u^2(\ell_{\text{Dmx}})/2} P(\ell_{\text{Dmx}})}{|du/dn|_{\ell_{\text{Dmx}}}}\right\}. \qquad (18)$$

This equation gives yield Y as a function of t_{CK}, which is implicitly present in u.

Since we are given Y and want to find t_{CK}, we can numerically reverse this equation and solve for t_{CK} by iteration. Normalizing by $\ell_{Dmx}\tau_0$, which is the nominal value, gives

$$\beta_\tau = t_{CK}/\ell_{Dmx}\tau_0 \,. \tag{19}$$

Figure 4 shows contours of constant β_τ for a range of σ_τ/τ_0 and yield. The value of $P(\ell_{Dmx})$ is not well known, but fortunately β_τ has only a weak logarithmic dependence on this parameter. Our calculations use $P(\ell_{Dmx}) = 0.125N_{CKT}$.

System composition

As noted before, to reduce complexity, only the logic devices and repeaters in the processor core are actually optimized; we assume that the power and speed of the clock and latch circuits, registers, memories, and I/O will be separately optimized and that the power/performance result of doing so will be essentially the same as for the logic part of the processor (although the optimal devices will be different). The assumed packing density of logic transistors has been adjusted to reflect actual design practices, as have the power allocations. On the basis of analyses of 90-nm- and 65-nm-generation IBM processor chips, we have assumed that 33% of core power and 15% of core area is associated with logic. For this purpose, "logic" excludes latches, clock buffers, registers, and all other forms of RAM. Depending on the processor chip configurations being simulated, we have assumed that 50–75% of the chip power is dissipated in the cores and that 50–75% of the chip area is devoted to cache.

Rent's rule [22] is used to determine average wire length for shorter wires. Repeaters are placed in long wires, with the average repeater width and separation being optimized as part of the overall optimization. In addition, we assume that wires with repeaters are on higher levels of the wiring hierarchy and are two times the size of the regular wires, thus lowering their resistance. This is a very simplified approximation to the detailed optimization of the wiring hierarchy that has been pursued by some [13], but we believe that it is sufficient for addressing the underlying device technology optimization in which we are interested. Furthermore, rather than independently optimizing the repeaters on individual wires [5], which raises questions of which sort of optimization to perform, we have chosen to merge the repeater optimization into that of the whole chip. The impact of the long wires on overall performance is captured in a latency-oriented model described in [16]. The model is based on the observation that long wire delay does not directly influence cycle time when designing a new processor, because it can be absorbed in increased pipelining, but the latency of long wires does contribute to the inefficiency of the processor by increasing the effective CPI (cycles per instruction) due to "instruction misses" (times when the processor must wait for a previous instruction to finish before launching a new instruction because the processor needs the previous result). To account for this latency, an application-dependent latency penalty factor γ is introduced, and an effective CPI (associated with latency issues only) is computed as a weighted average between the case in which instructions can be launched immediately ($CPI = 1$) and the case in which the previous instructions must finish first:

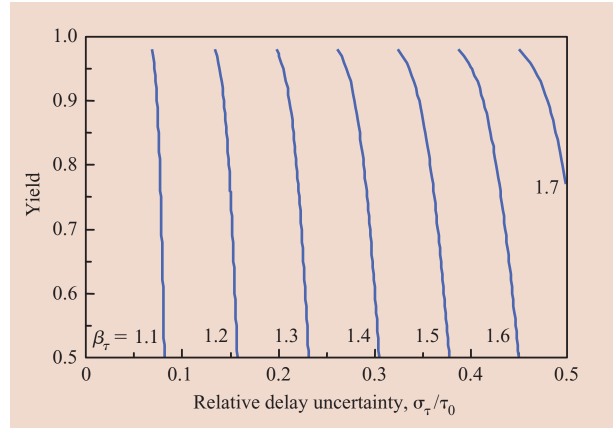

Figure 4

Contour plot of β_τ as a function of σ_τ/τ_0 and yield.

$$CPI_{eff} = (1-\gamma) + \gamma \frac{\tau_{instr}}{\tau_{cycle}} \,, \tag{20}$$

where $\tau_{cycle} = \ell_{Dmx}\tau$ is the cycle time, and $\tau_{instr} = n_{LI}\tau + n_{RI}\tau_R$ is the total time required to complete a typical instruction, from beginning to end, including all the stages of logic (n_{LI}) and the transmission time for long wires, which depends on the repeater stage delay, τ_R. This penalty factor enables the repeater characteristics to be included in the optimizations. The number of repeaters in a typical instruction is taken to be $n_{RI} = 2\sqrt{A_{core}}/S_R$, where S_R is the average spacing between repeaters, and $2\sqrt{A_{core}}$ means that the total wire length requiring repeaters is twice the edge of the processor core. We usually set $\gamma = 0.1$.

Thermal modeling

A thermal model has been implemented that allows the junction temperature to be self-consistently determined from the power dissipation. Temperature dependence is included in the subthreshold leakage current, the mobility model, and the wire resistance model. The heat-sink model is illustrated in **Figure 5**, and can accommodate hot spots, 2D and 3D thermal spreading resistance, and a wide range of materials.

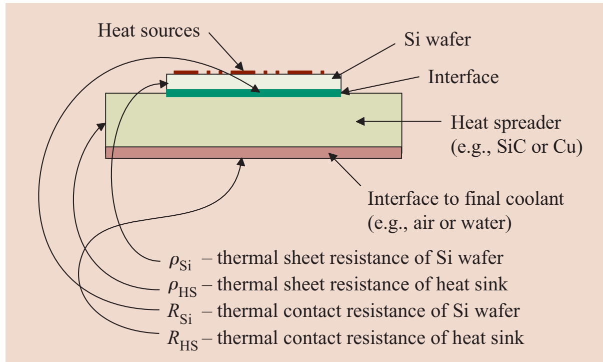

Figure 5

Geometric and interface aspects of the heat-sink model.

Junction temperature constraints can also be imposed on the optimizations, to reflect realistic reliability concerns. At low power levels, when the junction temperature does not reach the constraint value, there is no effect, but at high power levels, when the temperature would exceed the constraint value, the chip area is increased by the addition of non-dissipating, unused areas. This increase in area reduces the power density just enough to keep the temperature at the constraint value. Such design points are undesirable, but represent the best that could be done if one insisted on dissipating excessive power.

Optimization results

Figure 6 shows the detailed results of optimizations for the 90-nm to 32-nm-technology nodes, using the node characteristics shown in Table 2. These optimizations are performed for a dual-core processor chip with aggressive air cooling. Seven variables have been optimized: gate length, oxide thickness, halo doping, mean width, mean repeater spacing, mean repeater width, and V_{DD}. The peak in performance versus power seen in Figure 6(a) occurs because the heat-sink technology is fixed and the temperature rise is constrained. The peak corresponds to the power at which the maximum temperature is first reached, as can be seen from the constant temperature contours in the figure. In this case, the maximum temperature rise is 60°C. The only way to increase power further, without increasing temperature, is to spread the chip out, as described in the previous section. This lengthens wires and slows down the chip even though the power level is higher. Design points at power levels beyond the peak are undesirable and should be avoided in practice. Low-power designs require larger, less scaled devices [Figures 6(b), 6(c)] in order to reduce leakage currents, indicating that only the highest-power applications can utilize extremely scaled devices. **Figure 7**

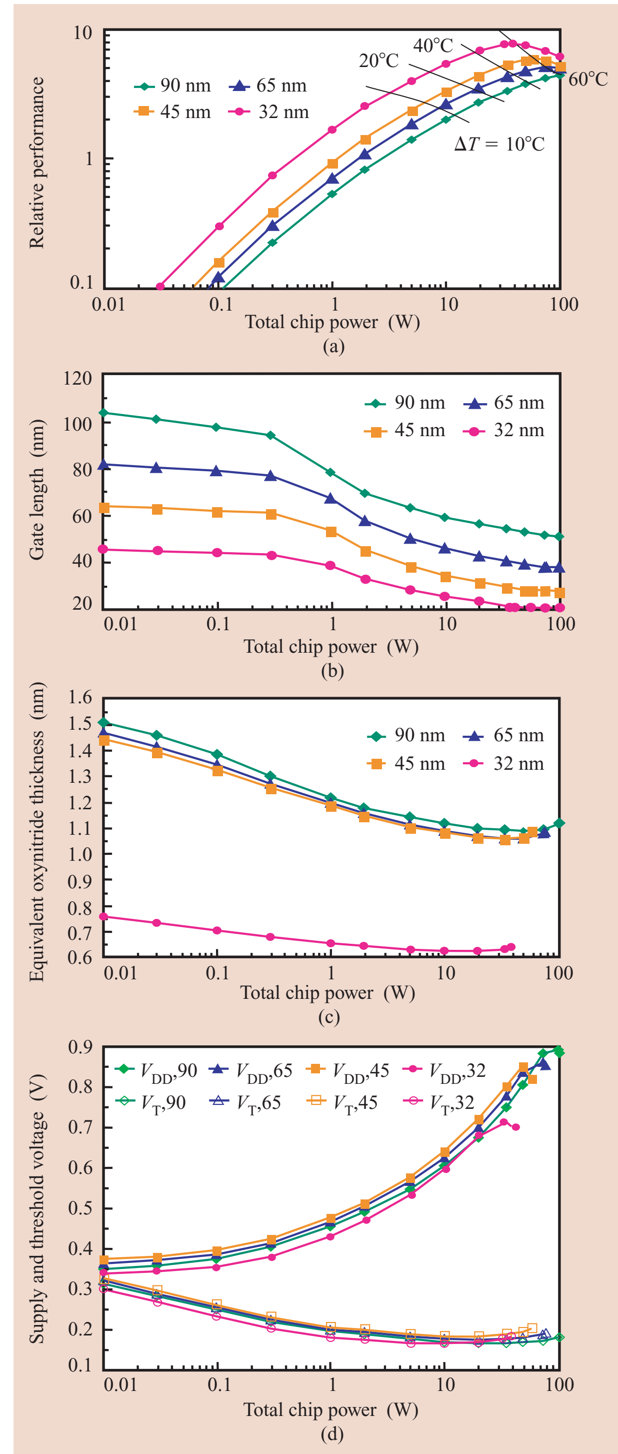

Figure 6

(a) Optimized performance versus power for 90-nm to 32-nm nodes. Junction temperature rise is indicated by the added contours. (b) Optimum gate lengths corresponding to part (a). (c) Optimum equivalent oxynitride (gate insulator) thickness corresponding to part (a). The 32-nm-node case uses a high-*k* material. (d) Optimum supply and threshold voltage (V_{TSAT}) corresponding to part (a).

shows the optimal allocation of power dissipation among the various mechanisms for a processor using water cooling (which allows higher power dissipation), from which it can be seen that gate leakage dissipation should not exceed a few percent, but optimal subthreshold leakage can exceed 50% for very high-power designs.

In an effort to understand the accuracy of our calculations, we have checked to see how our model predictions for past technology generations compare with what was actually built. We have found that the gate lengths [e.g., Figure 6(b)] agree reasonably well at the power levels for which technology generations have been targeted, but our supply and threshold voltages tend to be lower than what was used in practice, rising only slowly as the lithographic dimension increases. Two main reasons exist for this voltage discrepancy: 1) We have not yet included process variations in our analysis, which would undoubtedly slightly increase our optimized voltages; and 2) voltages used in past designs were probably not optimal. Ten to fifteen years ago, supply voltages were determined by external considerations, such as the "industry standard" five-volt power supply, and there was much resistance to the idea of lowering voltages, even when it became clear that reliability concerns demanded a change [23]. Furthermore, technologists tended to think that standby power should be quite low (unlike the optimized results in Figure 7), which required higher V_T, and commensurately higher V_{DD}. Consequently, we believe that past technology generations had non-optimally high voltages, making our comparison partially unsuccessful. On the basis of comparisons with more recent technologies, we expect our modeling to accurately predict trends, but exact optimum values for a specific scenario may be less accurate than the trend predictions because of the many simplifying assumptions.

Next we consider future technology options. Metal gates are simulated by removing the poly-Si depletion effects and adjusting the workfunction. High-*k* gate stacks are simulated using a double-layer bandgap-dependent tunneling model. As can be seen in **Figure 8**, high-*k* combined with metal-gate can potentially yield excellent chip-level performance enhancement, as is also seen in the 32-nm node in Figure 6(a), but metal gates by themselves do not offer much benefit over poly-Si, even for workfunctions that are equivalent to poly-Si. As the workfunction shifts from band edge toward midgap, a significant loss of performance occurs for both metal-gate and high-*k* combined with metal-gate. This loss occurs because the optimizations compensate for midgap workfunctions by lower doping (which increases depletion depth) and by raising the supply voltage (which necessitates thicker oxide). According to these optimizations, the benefits of the use of high-*k* are lost for PD-SOI by the time the workfunction reaches quarter-gap.

Many future technology options involve increasing mobility, such as the use of strain, hybrid-orientation substrates, and SiGe layers. **Figure 9** shows that mobility increases can indeed increase chip performance, though with diminishing returns for large increases in mobility.

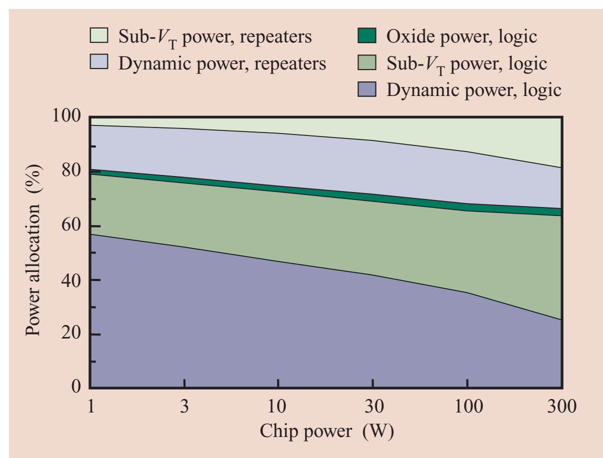

Figure 7

Cumulative power allocation fractions for the logic in processor cores, based on 45-nm-node optimizations. (At less than 0.5%, the graph segment for oxide power in repeaters is too small to be visible.)

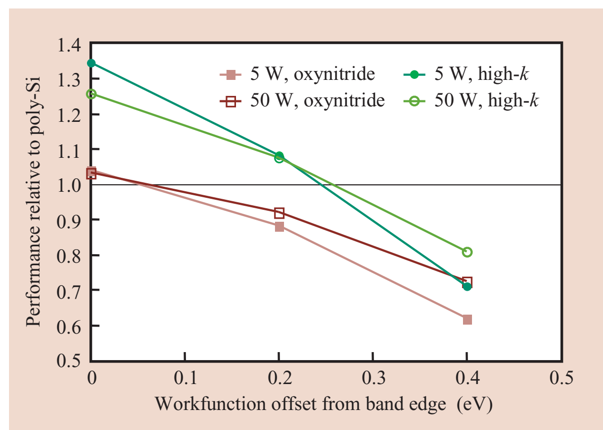

Figure 8

Relative chip performance versus workfunction offset from the band edge, for oxynitride and high-*k* gate dielectrics at the 45-nm node, and for two different chip power levels (5 W and 50 W). Workfunction offset values are measured from the Si band edge toward midgap, so that 0 is equivalent to the poly-Si offset. Performance is relative to a poly-Si gate, oxynitride case.

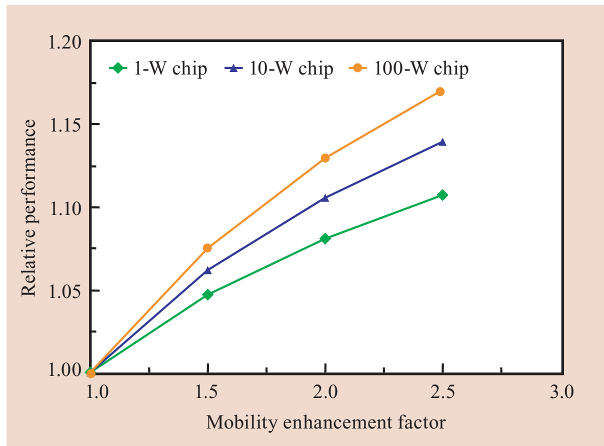

Figure 9

Relative chip performance increase versus mobility enhancement for the 45-nm node, for three different chip power targets. Mobility is normalized to its value in unstrained FETs.

This performance increase is larger for high-power chip designs than for low-power designs. It is not yet clear, however, how much mobility improvement will be possible at 32 nm, because much of the available increase will already have been achieved in previous nodes.

If we pessimistically suppose that some improvements will not be manufacturable, we obtain the results shown in **Figure 10**. Figure 10(a) serves as a baseline, in which improvements are successfully implemented according to Table 2. This is the same data as for Figure 6(a), on a linear scale, and we have added a high-*k* option for the 45-nm node, which clearly illustrates the potential benefits of an optimal high-*k* solution. Figure 10(b) shows the results if we are unable to improve the wiring dielectric constant and it remains fixed at 2.8. This reduces the peak performance of the 32-nm generation to the same value as for the 45-nm node. Finally, in Figure 10(c) we assume that the device technology is also fixed, with $L_G = 36$ nm, $t_{ox} = 1.1$ nm, and the wiring dielectric constant $k_{wiring} = 2.8$. In this case, the generation-to-generation changes involve only the packing density, as driven by the widths and wiring pitch, which are not fixed. A significant peak performance loss occurs in future generations, with very little gain even at low power, making it clear that density improvements alone are insufficient for future technology generations.

In **Figure 11**, the optimizer is used to assess and compare the potential performance gain achievable by a variety of proposed technology options. In this figure, the "base" case is the "baseline" 65-nm-node technology to which the other cases should be compared. Each point plotted is the peak performance that is possible with the given heat sink and the specified temperature rise,

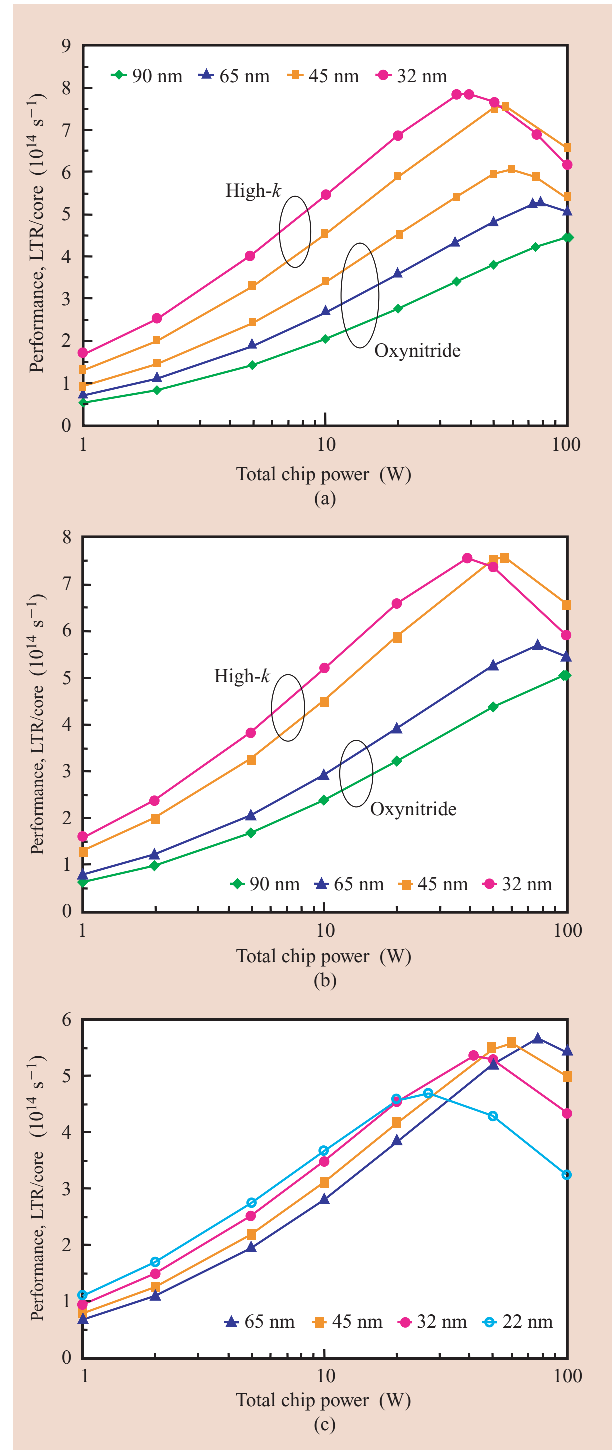

Figure 10

Chip performance versus power across technology generations, using different assumptions: (a) Full technology enhancements at each generation. (b) k_{wiring} fixed at 2.8. (c) L_G, t_{ox}, and k_{wiring} fixed at 45-nm-node values for all future generations; only voltages, widths, and wire sizes vary. Assumes dual-core processor with aggressive air cooling. (LTR: logic net transition rate.)

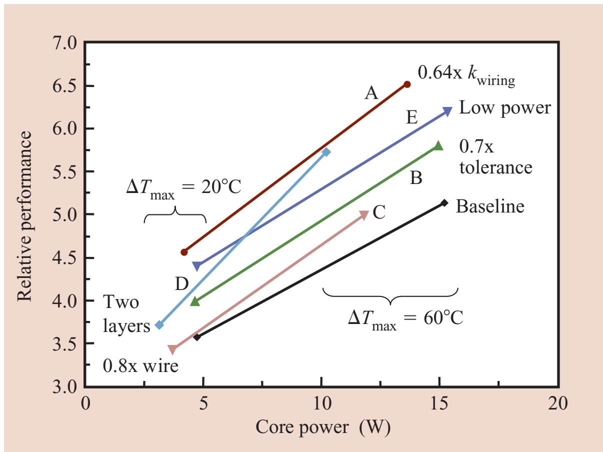

Figure 11

Peak performance versus power for aggressive air cooling and two different maximum allowed temperature rises, comparing various technology options (different lines) with the 65-nm-node technology baseline. Options compared are improvements in k_{wiring} (A), tolerance (B), and wire size (as indicated on plot) (C), 3D integration using two layers of active circuitry (D), and use of low-power circuit techniques to eliminate two thirds of passive power by turning off inactive circuit blocks (E).

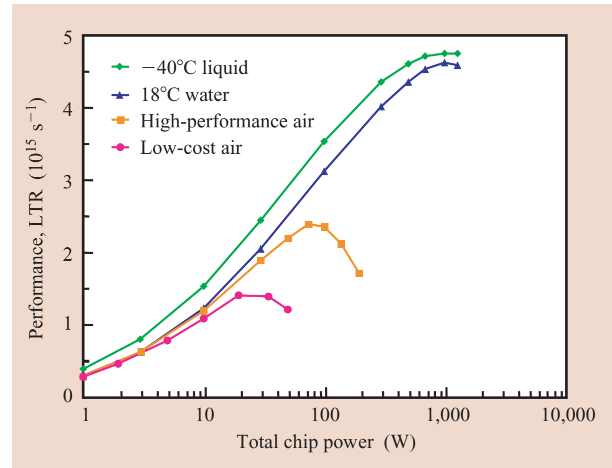

Figure 12

Dependence of total optimized chip performance on cooling technology, for a 45-nm-node four-core processor. The four cooling cases are very low-cost air cooling, high-performance air cooling, chilled water cooling through a microchannel heat sink, and −40°C liquid cooling through a microchannel heat sink. The low-temperature case does not include refrigerator power.

corresponding to the highest points on curves similar to those in Figure 6(a). Because the power level at which the peak occurs varies depending on the technology option, the data points are somewhat scattered along the x-axis. Among the options compared, the following appear to be effective for improving performance: reducing the wiring permittivity by 0.64x, using 3D integration with two layers of active circuitry, and turning off the supply to inactive logic. Reducing variability by 0.7x also helps performance somewhat, while simply making the wiring smaller does not appear very beneficial, as has already been discussed. Overall, technology changes that truly lower the switching energy appear useful, while changes that only make devices faster or denser at the same switching energy are not valuable when the circuits are power-limited. Because power is the controlling factor, improved heat removal is also quite effective. Note that maximum performance should be achieved by implementing all of the favorable changes simultaneously.

As noted above, improved heat-sink technology offers a direct path to larger performance gains than those provided by improved device technology, as shown in **Figure 12**. One may also gain a modest performance increase by decreasing the heat-sink temperature, but this performance gain generally disappears if the refrigerator power must be taken into account. Thermal solutions are difficult because such high-power processors are very inefficient, and performance is only increasing as roughly the log of the power. Note that it may not actually be possible to reliably deliver such high power levels to the chip, but experiments have shown that microchannel liquid cooling can remove the associated heat [24, 25].

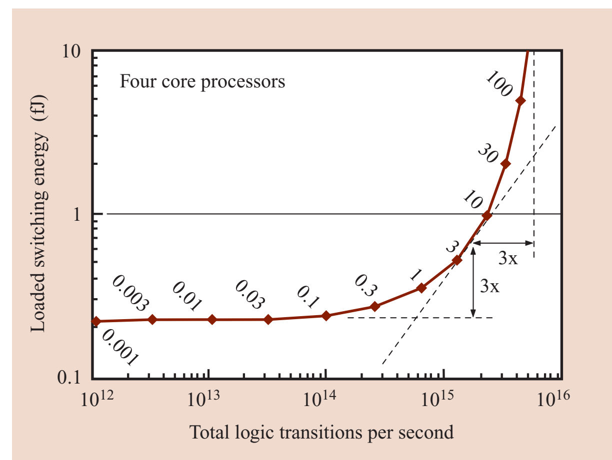

Figure 13

Average loaded switching energy versus performance for cross-generational optimization (involving nine parameters). Number label on each point gives total chip power for that point.

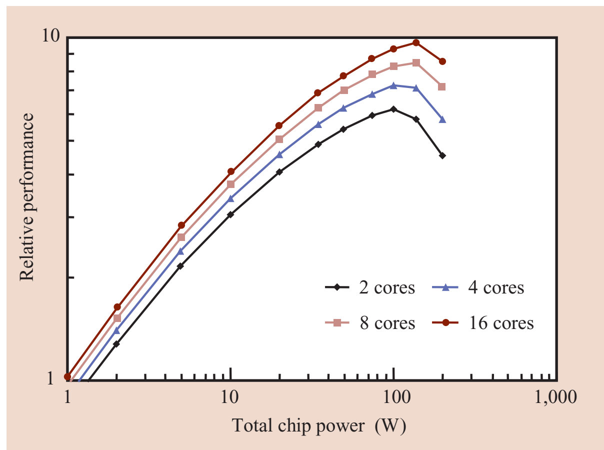

Figure 14

Dependence of performance on power for different numbers of cores for a 32-nm-technology node.

This efficiency challenge is captured in **Figure 13**, which shows average energy dissipated per logic transition (total logic power divided by LTR) versus overall performance, for optimizations that cross technology generations by also including the wire pitch and the halo behavior among the optimized variables, yielding a total of nine variables being optimized. These optimizations are considered for four-core processor chips. Clearly, the very high-power designs are quite energy-inefficient, on a logic transition basis, compared to what is possible at lower power. The knee in this curve is very interesting, because it turns out to be within a factor of ~3 from both the lowest-energy designs and the highest-performance designs. Architectural innovations may allow most applications below and above the knee of this curve to efficiently utilize the device design at the knee, making the knee a very important technology design point.

One way to address the energy inefficiency of the high-power design points involves the use of smaller, lower-power cores in parallel. This is examined in **Figure 14**, in which a fixed number of transistors is divided into different numbers of processor cores. The cases with the higher numbers of cores have higher performance because the smaller cores result in relatively less wiring capacitance due to the shorter wires. This basic performance increase must of course be adjusted for architecture and system effects associated with increased parallelism, but these issues are outside the scope of this work.

Conclusion

Our results show very clearly that power constraints have a great effect on technology scaling. It is no longer possible to scale CMOS technology from one generation to another without taking into account power dissipation. Many of the proposed technology enhancements do show promise, but careful optimizations are necessary at every power level to ensure that the most appropriate technology is being used. The optimizations show that there is still room for significant progress in high-performance CMOS out to at least the 32-nm generation, especially for high-power applications, but low-power applications will require less-scaled devices. In the future, the dominant CMOS market may include technologies such as those with characteristics near the knee of Figure 13; however, smaller markets will undoubtedly continue to exist for both high-performance logic and very low-power technology.

Acknowledgments

The authors wish to thank M. Wisniewski, M. Scheuermann, P. Restle, S. Kosonocky, E. Colgan, and J. Magerlein for useful discussions and information.

References

1. H. B. Bakoglu and J. D. Meindl, "A System-Level Circuit Model for Multi- and Single-Chip CPU's," *Proceedings of the International Solid-State Circuits Conference*, 1987, pp. 308–309.
2. H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*, Addison-Wesley Publishing Co., Reading, MA, 1990.
3. G. A. Sai-Halasz, "Performance Trends in High-End Processors," *Proc. IEEE* **83**, 20–36 (1995).
4. J. C. Eble, V. K. De, D. S. Wills, and J. D. Meindl, "A Generic System Simulator (GENESYS) for ASIC Technology and Architecture Beyond 2001," *Proceedings of the 9th Annual IEEE International ASIC Conference*, Rochester, NY, 1996, pp. 193–196.
5. J. C. Eble, *A Generic System Simulator with Novel On-Chip Cache and Throughput Models for Gigascale Integration*, Ph.D. Thesis, Georgia Institute of Technology, Atlanta, November 1998.
6. B. M. Geuskens and K. Rose, *Modeling Microprocessor Performance*, Kluwer Academic Publishing Co., Boston, MA, 1998.
7. D. Sylvester and K. Keutzer, "System-Level Performance Modeling with BACPAC—Berkeley Advanced Chip Performance Calculator," workshop notes, ACM International Workshop on System-Level Interconnect Prediction, 1999, pp. 109–114.
8. Y. Cao, C. Hu, X. Huang, A. B. Kahng, I. Markov, M. Oliver, D. Stroobandt, and D. Sylvester, "Improved A Priori Interconnect Predictions and Technology Extrapolation in the GTX System," *IEEE Trans. VLSI Syst.* **11**, No. 1, 3–14 (2003).
9. J. D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proc. IEEE* **83**, No. 4, 619–635 (1995).
10. Z. Chen, J. Burr, J. Shott, and J. D. Plummer, "Optimization of Quarter Micron MOSFETs for Low Voltage/Low Power Applications," *IEDM Tech. Digest*, pp. 63–65 (1995).
11. D. J. Frank, P. Solomon, S. Reynolds, and J. Shin, "Supply and Threshold Voltage Optimization for Low Power Design," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, Monterey, CA, August 1997, pp. 317–322.
12. A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A Minimum Total Power Methodology for

Projecting Limits on CMOS GSI," *IEEE Trans. VLSI Syst.* **8**, No. 3, 235–251 (2000).
13. J. D. Meindl, J. A. Davis, P. Zarkesh-Ha, C. S. Patel, K. P. Martin, and P. A. Kohl, "Interconnect Opportunities for Gigascale Integration," *IBM J. Res. & Dev.* **46**, No. 2/3, 245–263 (2002).
14. 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* **89**, No. 3, 259–288 (2001).
15. D. J. Frank, "Power-Constrained CMOS Scaling Limits," *IBM J. Res. & Dev.* **46**, No. 2/3, 235–244 (2002).
16. D. J. Frank, "Power Constrained Device and Technology Design for the End of Scaling," *IEDM Tech. Digest*, pp. 643–646 (2002).
17. P. K. Bondy, "Moore's Law Governs the Silicon Revolution," *Proc. IEEE* **86**, 78–81 (1998).
18. P. M. Solomon, J. Jopling, D. J. Frank, C. D'Emic, O. Dokumaci, P. Ronsheim, and W. E. Haensch, "Universal Tunneling Behavior in Technologically Relevant PN Junction Diodes," *J. Appl. Phys.* **95**, No. 10, 5800–5812 (2004).
19. 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 Circ.* **25**, No. 2, 584–594 (1990).
20. S. Takagi, M. Iwase, and A. Toriumi, "On the Universality of Inversion-Layer Mobility in N and P-Channel MOSFETs," *IEDM Tech. Digest*, pp. 398–401 (1988).
21. E. Buturla, J. Johnson, S. Furkay, and P. Cottrell, "A New 3-D Device Simulation Formulation," *Proceedings of NASCODE VI: Sixth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits*, Boole Press, Dublin, Ireland, 1989, p. 291.
22. J. A. Davis, V. K. De, and J. D. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI)—Part I: Derivation and Validation," *IEEE Trans. Electron Devices* **45**, 580–589 (March 1998).
23. B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS Scaling, the Next Ten Years," *Proc. IEEE* **89**, 595–606 (1995).
24. E. G. Colgan, B. Furman, M. Gaynes, W. Graham, N. LaBianca, J. H. Magerlein, R. J. Polastre, M. B. Rothwell, R. J. Bezama, H. Toy, J. Wakil, J. Zitz, and R. Schmidt, "A Practical Implementation of Silicon Microchannel Coolers for High Power Chips," *Proceedings of the 21st Annual IEEE Semiconductor Thermal Measurement and Management Symposium,* San Jose, CA, March 15–17, 2005, pp. 1–7.
25. C. J. M. Lasance and R. E. Simons, "Advances in High Performance Cooling for Electronics," *Electron. Cooling* **11**, (November 2005). See *http://www.electronics-cooling.com/html/2005_nov_article2.html.*

Received October 3, 2005; accepted for publication December 22, 2005; Internet publication August 6, 2006

David J. Frank *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (djf@us.ibm.com)*. Dr. Frank received a B.S. degree from the California Institute of Technology in 1977 and a Ph.D. degree in physics from Harvard University in 1983. Since graduation, he has worked at the IBM Thomas J. Watson Research Center, where he is a Research Staff Member. His studies have included non-equilibrium 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. Dr. Frank is an IEEE Fellow; he has served as chairman of the Si Nanoelectronics Workshop and is an associate editor of the *IEEE Transactions on Nanotechnology*. He has authored or co-authored more than 90 technical publications and holds nine U.S. patents.

Wilfried Haensch *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (whaensch@us.ibm.com)*. In 1981, Dr. Haensch received his Ph.D. degree from the Technical University of Berlin, Germany, in the field of theoretical solid-state physics. In 1984 he joined Siemens Corporate Research in Munich to investigate high-field transport in MOSFET devices, and in 1988 he joined the DRAM development team at the Siemens Research Laboratory to investigate new cell concepts. In 1990, he joined the DRAM alliance between IBM and Siemens to develop quarter-micron 64M DRAM. In this capacity, Dr. Haensch was involved with device characterization of shallow-trench bounded devices and cell-design concerns. In 1996, he moved to a manufacturing facility to build various generations of DRAM. His primary mission was to transfer technologies from development into manufacturing and to guarantee a successful yield ramp of the product. In 2001, he joined the IBM Thomas J. Watson Research Center to lead a group concerned with novel devices and applications. He is currently responsible for post-45-nm-node device design and its implications for circuit functionality.

Ghavam Shahidi *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (shahidi@us.ibm.com)*. Dr. Shahidi received his B.S., M.S., and Ph.D. degrees, all in electrical engineering, from MIT. In 1989 he joined the IBM Thomas J. Watson Research Center, where he initiated the SOI development program. Over the following years, Dr. Shahidi led the development of SOI CMOS technology. From the mid-1990s, first as manager and later as Director of High-Performance Logic Development in the IBM Microelectronics Division, he led the development of several generations of high-performance CMOS technology in the Advanced Silicon Technology Center in Hopewell Junction, New York, until 2003. He is currently the Director of Silicon Technology in the IBM Research Division and an IBM Fellow.

Omer H. Dokumaci *IBM Systems and Technology Group, 2070 Route 52, Hopewell Junction, New York 12533 (dokumaci@us.ibm.com)*. Dr. Dokumaci received his Ph.D. degree in electrical engineering from the University of Florida, joining the Process and Device Modeling Group at the IBM facility in Hopewell Junction, New York. Dr. Dokumaci's research has concentrated on modeling and simulation of dopant diffusion and activation, and advanced devices such as FinFETs, ultrathin silicon, metal-gate, back-gate, and ground-plane devices.