# Supply and Power Optimization in Leakage-Dominant Technologies

Man Lung Mui, Kaustav Banerjee, *Senior Member, IEEE*, and Amit Mehrotra, *Member, IEEE*

***Abstract*—In this paper, we present a methodology for systematically optimizing the power-supply voltage for either maximizing the performance of very large scale integration (VLSI) circuits or minimizing the power dissipation in technologies where leakage power is not an insignificant fraction of the total power dissipation. For this purpose, we develop simplified empirical equations that describe the transistor behavior as a function of power supply and temperature. We use these models to calculate the full-chip power dissipation as a function of power supply and temperature. We then solve the power and chip thermal equations simultaneously to calculate the chip temperature and power dissipation at a given power supply. By varying the power-supply voltage, we determine the optimum $V_{\rm DD}$ value that minimized delay per unit length in global interconnects and therefore maximizes performance. Using the same framework, by again varying the supply we find the optimum $V_{\rm DD}$ that minimized the total power dissipation while maintaining a given delay per unit length. We show that for 90- and 65-nm technologies, where leakage power represents a significant fraction of the total power dissipation, optimum $V_{\rm DD}$ for maximum performance is lower than the International Technology Roadmap for Semiconductors (ITRS) specified supply voltage. This is due to the fact that reducing $V_{\rm DD}$ results in a large reduction in total power dissipation, and therefore the chip temperature, which improves performance. This improvement in performance is greater than the performance penalty incurred due to reduction in $V_{\rm DD}$. We also show that as the required delay per unit length is increased, total chip power consumption is reduced significantly if the power supply is also reduced as compared to the case when power supply is fixed at the nominal value. This change becomes larger with technology scaling due to the fact that leakage power, which is a very strong function of chip temperature, becomes a larger fraction of the full-chip power dissipation.**

***Index Terms*—Chip temperature, leakage currents, nanometer technologies, power dissipation.**

## I. Introduction

As the channel lengths of metal–oxide–semiconductor (MOS) devices scale below 180 nm, leakage current becomes nonnegligible and off-state current and power dissipation have become important. With technology scaling, the supply voltage needs to be scaled in order to maintain reliable operation of the transistors. This forces the threshold voltage of the transistors to be scaled in order to maintain performance. Off-state leakage current increases exponentially as the threshold voltage is scaled. It has been projected that the transistor off-state current per micrometer of transistor width increases by $\sim 5\times$ per generation [1]. As a result, in the current technology generation, leakage power has become a significant fraction of the total power dissipation and this fraction is projected to increase with technology scaling [2].

Increasing power dissipation increases the cost of the package and may cause reliability concerns and even failures of the chip. In a leakage-dominant technology, power dissipation is extremely critical. For a given package, die temperature is linearly proportional to the total power dissipation. However, leakage current, and therefore leakage power, increases exponentially with temperature. As shown in Section V, if the thermal conductance of the package is not large enough, for a leakage-dominant technology, the exponential dependence of leakage power on temperature will cause thermal runaway where the die temperature increases unbounded and the chip fails. Even if thermal runaway does not occur, the operating temperature of the chip may be larger than the designed value, which will either increase the package cost or degrade the performance as well as the reliability of the chip. Therefore, in leakage-dominant technologies, it is essential to control the leakage power and the temperature of the die.

One viable method for optimizing the performance or minimizing total power dissipation for a given delay of very large scale integration (VLSI) circuits in leakage-dominant technologies is to vary the power supply. Reduction in power supply degrades performance but also results in a quadratic reduction in switching power [3] and an exponential reduction in leakage current, and therefore leakage power, due to reduction in drain-induced barrier lowering (DIBL) [4]. Furthermore, for a given package, reducing power dissipation results in reduction of die temperature, which further reduces the leakage current exponentially [4]. The resulting reduction in temperature will improve the performance and can compensate for the performance degradation due to lowering of $V_{\rm DD}$. Furthermore, for a given interconnect delay, reduction in $V_{\rm DD}$ allows the buffer size to be reduced and interbuffer interconnect length to be increased, which reduces the total repeater power dissipation. Reduction in $V_{\rm DD}$ also results in switching and leakage power of the arithmetic and logic blocks in the circuit and therefore, the overall power dissipation is reduced. In Section VII, we show that reducing the supply voltage slightly results in an improvement in performance for the 90- and 65-nm technology nodes.

In this work, we develop a methodology to estimate the optimal supply voltage that maximizes circuit performance.

Manuscript received August 2, 2004; revised March 21, 2005. This paper was recommended by Associate Editor T. Chen.

M. L. Mui is with the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA (e-mail: manmui@uiuc.edu).

K. Banerjee is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: kaustav@ece.ucsb.edu).

A. Mehrotra is with Berkeley Design Automation, Santa Clara, CA 95054 USA (e-mail: amit.mehrotra@berkeley-da.com).

Digital Object Identifier 10.1109/TCAD.2005.852039

This methodology is unique since it takes into account the change in the die temperature as the total power dissipation of the chip varies. This framework is also used for calculating the minimum supply voltage for a given circuit performance that minimizes the total power dissipation. For this purpose, we first develop simplified empirical models for device equivalent resistance, parasitic capacitance, and output capacitance as a function of temperature and $V_{\rm DD}$, which results in a model for circuit performance as a function of $V_{\rm DD}$ and temperature (Section III). We use the temperature dependence of the leakage current and threshold voltage to derive the temperature dependence of total power dissipation as a function of temperature (Section IV). By solving the power dissipation equation and the package thermal equation, we find the die temperature, power dissipation, and delay per unit length for a given $V_{\rm DD}$. By varying $V_{\rm DD}$, we find the optimal supply voltage that maximizes performance. Furthermore, by minimizing the expression for power dissipation subject to the package thermal equation, we find the die temperature, optimum buffer size and interbuffer interconnect length, and total power dissipation that results in a specified delay per unit length. We consider two typical cases in global-interconnect optimization: 1) when the buffer insertion can be optimized for the target $V_{\rm DD}$ and temperature; and 2) when buffering scheme is fixed and is designed to be optimal at nominal supply voltage and at a temperature of 105 °C. We show that the optimal supply voltage that reduces power dissipation is smaller than the nominal $V_{\rm DD}$ for the 90- and 65-nm technology nodes.

## II. Previous Work

Several techniques have been proposed for reducing the off-state current [5], [6] and optimizing repeaters for reduced delay and power [7], [8]. These include reducing power supply [9], [10], using nonminimum channel-length transistors [11], using stacked transistors [12], [13], and reverse body bias [14]. A comprehensive analysis of the effectiveness of these techniques was presented in [1], but the authors did not take into account the change in temperature due to reduction in power dissipation and therefore the improvement in performance. They concluded that increasing the effective channel length and stacking transistors is the most effective method for reducing leakage power. However, these power minimization techniques did not consider the temperature effect, which is going to be crucial for nanometer-scale technologies where subthreshold leakage can be significant. It has been recently shown that strong electrothermal couplings between supply voltage, frequency, power dissipation, and junction temperature exist in leakage-dominant nanometer-scale technologies, mainly due to the exponential dependence of subthreshold leakage current on temperature, which can significantly impact various power-performance–reliability-cooling cost-optimization schemes [15]. A systematic power-optimal repeater-insertion methodology was proposed in [2] where, for a given delay penalty, optimum repeater size and interbuffer interconnect lengths were calculated for various International Technology Roadmap for Semiconductors (ITRS) technology nodes, which minimized the total interconnect power. A few follow-up works focused on simultaneous repeater area and power optimization [16], buffer insertion and sizing, and wire sizing [17]. However, all these approaches assume that the chip is operating at 100 °C and nominal $V_{\rm DD}$ and do not comprehend the reduction in power dissipation, and therefore temperature (which further reduces leakage power), in a self-consistent manner. In this work we, consider the reduction in temperature due to reduction in power dissipation and, hence, the subsequent improvement in performance due to reduction in power-supply voltage, and unlike [1], we show that for leakage-dominant technologies, reducing power-supply voltage to some extent improves the performance.

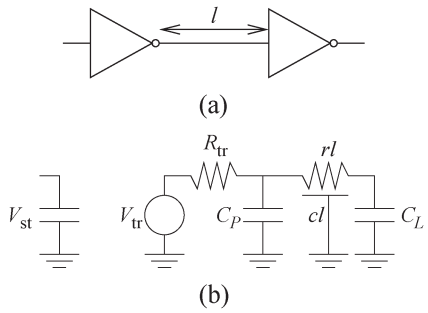



Fig. 1. Interconnect of length $l$ between two identical inverters. (a) Schematic representation. (b) Equivalent resistance–capacitance ($RC$) circuit.

## III. Interconnect Delay Model

Consider a uniform interconnect of resistance $r$ per unit length and capacitance $c$ per unit length buffered by identical repeaters as shown in Fig. 1. Assume that for a minimum-sized repeater, the input capacitance is $c_0$, the output parasitic capacitance is $c_{\rm p}$ and output resistance is $r_s$. Therefore, for a repeater of size $s$, the total output resistance $R_{\rm tr} = (r_s/s)$, the total output parasitic capacitance $C_{\rm p} = c_{\rm p}s$, and the total input capacitance is $C_L = c_0 s$. If the line segment is of length $l$ and the repeater size is $s$, then the time constant of that segment is [18]

$$\tau = r_s(c_0 + c_{\rm p}) + \frac{r_s}{s}cl + rlsc_0 + \frac{1}{2}rcl^2 \tag{1}$$

and the latency or the delay of that section is $\tau \log 2$.

Now, consider a long interconnect of a given length $L$ that is uniformly buffered with interbuffer interconnect length $l$. Therefore, the total number of segments is $L/l$. The total delay through that line is given by

$$\text{delay} = \frac{L}{l} \times \tau \log 2 \propto \frac{\tau}{l}$$

where $\tau/l$ is the delay per unit length, which is given by

$$\frac{\tau}{l} = \frac{1}{l}r_s(c_0 + c_{\rm p}) + \frac{r_s}{s}c + rsc_0 + \frac{1}{2}rcl. \tag{2}$$

TABLE I
INTERCONNECT PARAMETERS AND NOMINAL SUPPLY VOLTAGE FOR DIFFERENT TECHNOLOGY NODES BASED ON ITRS

| Technology Node (nanometer) | 130 | 90 | 65 |
|---|---|---|---|
| $W$ (nanometer) | 335 | 230 | 145 |
| $T$ (nanometer) | 670 | 483 | 319 |
| $\epsilon_{\text{ins}}$ | 3.1 | 2.8 | 2.5 |
| $V_{\text{DD}}$ (volt) | 1.1 | 1 | 0.65 |
| $f_{\text{clk}_{\text{nom}}}$ (gigahertz) | 1.68 | 3.99 | 6.74 |
| $I_{\text{off}_{n_{\text{nom}}}}$ (ampere per meter) | 0.42 | 2.68 | 17.39 |
| $I_{\text{off}_{p_{\text{nom}}}}$ (ampere per meter) | 0.21 | 2.20 | 8.55 |
| $r_{s_{\text{nom}}}$ (kiloohm) | 8.8 | 6.3 | 20.1 |
| $c_{0_{\text{nom}}}$ (femtofarad) | 0.94 | 0.59 | 0.60 |
| $c_{\text{p}_{\text{nom}}}$ (femtofarad) | 2.29 | 1.75 | 0.48 |
| $P_{\text{total}}$ (watt) | 61 | 85 | 104 |

Note that optimizing the delay of the interconnect of a fixed length is equivalent to optimizing $\tau/l$. This delay per unit length is optimal when [18]

$$l_{\text{opt}} = \sqrt{\frac{2r_s(c_0 + c_{\text{p}})}{rc}} \quad s_{\text{opt}} = \sqrt{\frac{r_s c}{r c_0}} \tag{3}$$

and is given by

$$\left(\frac{\tau}{l}\right)_{\text{opt}} = 2\sqrt{r_s c_0 r c}\left(1 + \sqrt{\frac{1}{2}\left(1 + \frac{c_{\text{p}}}{c_0}\right)}\right). \tag{4}$$

Note that the optimal size of repeater $s_{\text{opt}}$, optimal interrepeater length $l_{\text{opt}}$, and optimal delay per unit length $(\tau/l)_{\text{opt}}$ are functions of repeater parameters $r_s$, $c_0$, and $c_{\text{p}}$, and interconnect parameters $r$ and $c$, which, in turn, depend on supply voltage and temperature. Therefore $s_{\text{opt}}$, $l_{\text{opt}}$, and $(\tau/l)_{\text{opt}}$ are functions of supply voltage and temperature. The interconnect resistance per unit length is given by

$$r = r_0\left(1 + \kappa(T - T_{\text{nom}})\right)$$

where $r_0$ is the resistance per unit length at nominal temperature $T_{\text{nom}}$, $\kappa$ is the temperature coefficient with unit of ohms/kelvin, and $T$ is the operating temperature. Interconnect capacitance $c$ is assumed to be independent of $V_{\text{DD}}$ and temperature.

Repeater parameters at various temperatures and supply voltages were extracted using simulation program with integrated circuits emphasis (SPICE) simulations similar to [19]. A five-stage ring oscillator with a given length of global interconnect of width $W_{\min}$ (see Table I for values of $W_{\min}$ for various technology nodes) in between each stage was simulated. The interconnect length $l$ and inverter size $s$ were varied to obtain the minimum stage delay per unit length. $r_s$, $c_0$, and $c_{\text{p}}$ were calculated from these values of $s_{\text{opt}}$, $l_{\text{opt}}$, and $(\tau/l)_{\text{opt}}$ for a given supply voltage and temperature. Fig. 2 plots $r_s$, $c_0$, and $c_{\text{p}}$ as the power supply is varied $\pm$ 20% from the nominal value and the temperature is varied from 25 °C to 125 °C. Note that,

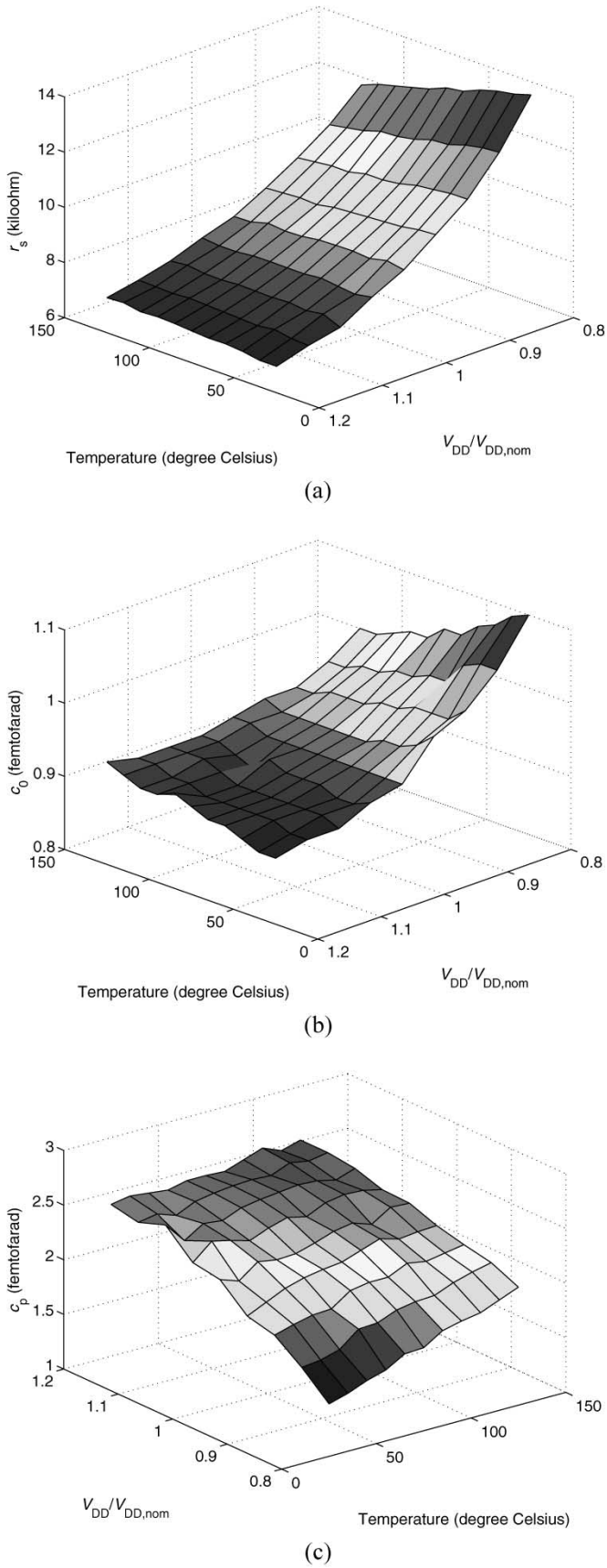



Fig. 2. Temperature and supply-voltage dependence of buffer parameters for the 130-nm technology. (a) $r_s$ as a function of temperature and $V_{\text{DD}}$. (b) $c_0$ as a function of temperature and $V_{\text{DD}}$. (c) $c_{\text{p}}$ as a function of temperature and $V_{\text{DD}}$.

as expected, the dependence of $c_0$ on $V_{\rm DD}$ and temperature is very weak. Using curve fitting, we generate the expressions of $r_s$, $c_0$, and $c_{\rm p}$ in terms of supply voltage and temperature.

## IV. Power Model

In this work, we will consider two cases: 1) global interconnects are optimally buffered for the targeted power supply and temperature; and 2) global interconnects are optimally buffered for operation at the nominal power supply and temperature. For scenario 1), changing the temperature and supply voltage will change $s_{\rm opt}$ and $l_{\rm opt}$, which, not only changes the power dissipation of each repeater, but also changes the number of repeaters. We therefore separate the full-chip power consumption into two parts

$$P_{\rm total} = P_{\rm logic} + P_{\rm repeater} \tag{5}$$

where $P_{\rm repeater}$ denotes the total power dissipated in the buffers and global interconnects driven by these buffers and $P_{\rm logic}$ is the remaining power. For this work, we assume that for each technology node, 30% of total power dissipation is repeater power.

The power consumption of both logic circuits and repeaters can be expressed as the following [10]:

$$P = P_{\rm switching} + P_{\rm short\ circuit} + P_{\rm leakage}.$$

We need to determine the switching, short-circuit, and leakage power for logic circuits and repeaters. We assume that for logic blocks, the load capacitance is dominated by input capacitance of logic gates whereas the load capacitance of repeaters will have both interconnect capacitance and input capacitance of other repeaters. Therefore, the percentage of switching, short-circuit, and leakage power will be different for logic gates and repeaters. We also need to determine how each of the above three components of power change as temperature and supply voltage are varied.

The switching power of a repeater in Fig. 1(a) is given by [3]

$$P_{\rm switching} = \alpha\left(s(c_{\rm p}+c_0)+lc\right)V_{\rm DD}^2\, f_{\rm clk}$$

where $V_{\rm DD}$ is the power-supply voltage, $f_{\rm clk}$ is the clock frequency, and $\alpha$ is the switching factor (or activity factor), which is the fraction of repeaters on a chip that are switched during an average clock cycle. $\alpha$ can be taken as 0.15 [10]. For optimally sized and placed buffers, $C_L$ is given in [2]

$$C_L = s_{\rm opt}(c_0 + c_{\rm p}) + cl_{\rm opt}$$

which is a function of supply voltage and temperature since $s_{\rm opt}$ and $l_{\rm opt}$ are functions of supply voltage and temperature.

For the logic blocks, we assume that the load capacitance does not vary with temperature and $V_{\rm DD}$. It is a valid assumption since the fan-outs of gates of the functional blocks are usually greater than 1 in general. The loading capacitance, therefore, is dominated by gate capacitance, which has a very weak dependence on temperature and $V_{\rm DD}$.

The clock frequency $f_{\rm clk}$ is inversely proportional to the delay of the critical path of the circuit. It has been shown in [20] and [21] that the performance is dominated by global interconnects. Therefore, $f_{\rm clk}$ can be assumed to be inversely proportional to $(\tau/l)_{\rm opt}$, which in turn is a function of supply voltage and temperature.

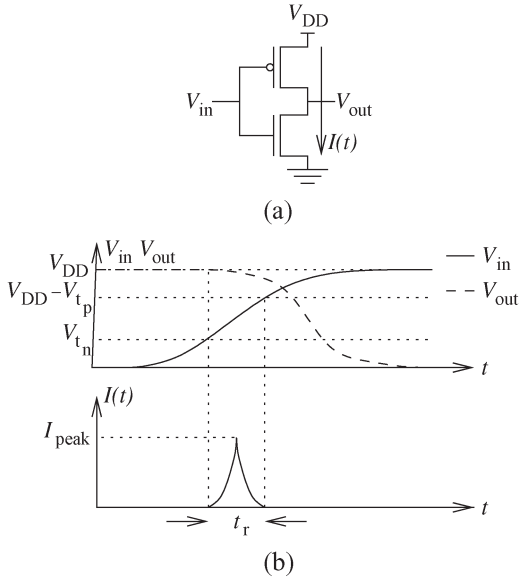



Fig. 3. Voltage and current waveforms of a CMOS inverter. (a) CMOS inverter. (b) Voltage and current waveforms.

The second component is the short-circuit power. This power consumption is incurred when both pull-up and pull-down networks are simultaneously on. Consider the simplest static complementary MOS (CMOS) logic circuit, an inverter, which is shown in Fig. 3(a). When the N-channel MOS (NMOS) transistor turns on due to a rising waveform at the input and the P-channel MOS (PMOS) transistor continues to conduct current until the input voltage becomes greater than $V_{\rm DD} - |V_{\rm t_p}|$, both transistors are on simultaneously. Hence, there is a direct current (dc) flowing from supply to ground, and is called short-circuit current. Note that the current not only depends on the input voltage, but also depends on the output voltage. The input- and output-voltage waveforms and the current waveform are shown in Fig. 1(b). The short-circuit current waveform can be approximated as a triangular wave [22]. The total charge that flows in this period can be found by calculating the area of this triangle. Let $t_{\rm r}$ denote the time for the input voltage to rise from $V_{\rm t_n}$ to $V_{\rm DD} - |V_{\rm t_p}|$. Assuming symmetric high-to-low and low-to-high transitions for both input and output of the logic gate, the total short-circuit power for a single logic gate is given by

$$\begin{aligned} P_{\rm short\ circuit} &= \alpha t_{\rm r} V_{\rm DD} I_{\rm peak} f_{\rm clk} \\ &= \alpha t_{\rm r} V_{\rm DD} W_{n_{\rm min}} s I_{\rm short\ circuit} f_{\rm clk} \end{aligned}$$

where $\alpha$ is the same switching factor as in the switching-power expression. $I_{\rm short\ circuit}$ is the peak current per transistor width. Assuming that the output waveform is a single time-constant exponential, $t_{\rm r}$ is given by [2]

$$t_{\rm r} = \tau \log_{\rm e}\left(\frac{V_{\rm DD} - |V_{\rm t_p}|}{V_{\rm t_n}}\right)$$

where $\tau$ is the time constant for the output node, which is defined in Section III. For repeaters, $\tau$ is given by (1). For logic blocks, since the interconnect delay is very small, $\tau$ for these circuits can be expressed as

$$\tau_{\text{logic}} \approx r_s(c_0 + c_{\text{p}}).$$

Note that $I_{\text{short circuit}}$ for both logic circuits and buffers is the same and is temperature dependent since the mobility and threshold vary with temperature.

The threshold voltage is given by [4]

$$V_{\text{t}} = -\frac{E_{\text{g}}}{2q} + \phi_B + \frac{\sqrt{4\epsilon_{\text{Si}} qN\phi_B}}{C_{\text{ox}}} \tag{6}$$

where $\epsilon_{\text{Si}}$ is the permittivity of silicon, $N$ is the doping concentration, is $q$ the single electron charge, $C_{\text{ox}}$ is gate-oxide capacitance, $E_{\text{g}}$ is bandgap energy, which has the following temperature dependence [23]:

$$E_{\text{g}} = 1.166 - \frac{4.73 \times 10^{-4} T^2}{T + 636}.$$

$E_{\text{g}}$ in the above expression is in electronvolts. $\phi_B$ is defined as

$$\begin{aligned}\phi_B &= \frac{kT}{q} \log_{\text{e}} \left(\frac{N}{n_i}\right) \\ &= \frac{kT}{q} \log_{\text{e}} \left(\frac{N}{4.66 \times 10^{15} T^{1.5} \exp\left(-\frac{E_{\text{g}}}{2kT}\right)}\right)\end{aligned} \tag{7}$$

where $k$ is the Boltzmann constant and $N$ is the doping concentration per cubic centimeter.

The last component is leakage power. In our model, we are only concerned with the subthreshold leakage power, which is given by [2]

$$P_{\text{leakage}} = V_{\text{DD}} I_{\text{leakage}} = V_{\text{DD}} \frac{1}{2} \left(I_{\text{off}_{\text{n}}} W_{\text{n}} + I_{\text{off}_{\text{p}}} W_{\text{p}}\right)$$

where $I_{\text{off}_{\text{n}}}(I_{\text{off}_{\text{p}}})$ is the leakage current of NMOS (PMOS) transistor per transistor width, which is given by [1]

$$I_{\text{off}} = \mu_{\text{eff}} C_{\text{ox}} \frac{W}{L_{\text{eff}}} \left(\frac{kT}{q}\right)^2 \exp(1.8) \exp\left(\frac{-V_{\text{t}} + \eta V_{\text{DD}}}{n\frac{kT}{q}}\right) \tag{8}$$

where $\eta$ is the DIBL coefficient and $n$ is the transistor subthreshold swing coefficient. The temperature dependence of mobility is given by [24]

$$\begin{aligned}\mu_{\text{n,eff}} &= 88 T_n^{-0.57} + \frac{1250 T_n^{-2.33}}{1 + \frac{N_a}{1.26\times 10^{17} T_n^{2.4}} \times 0.88 T_n^{-0.146}} \\ \mu_{\text{p,eff}} &= 54.3 T_n^{-0.57} + \frac{407 T_n^{-2.33}}{1 + \frac{N_d}{2.35\times 10^{17} T_{\text{n}}^{2.4}} \times 0.88 T_n^{-0.146}}\end{aligned} \tag{9}$$

where $N_a$ and $N_d$ are bulk doping concentrations and $T_n = T/300$ where $T$ is the temperature in Kelvin. $\eta$ is assumed to be independent of temperature and $V_{\text{DD}}$ and is taken to be 50 mV/V for all technologies. $n$ can be related to temperature as follows:

$$n = 1 + \frac{\sqrt{\frac{\epsilon_{\text{Si}} qN}{4\phi_B}}}{C_{\text{ox}}}$$

where $\phi_B$ is a function of temperature [see (7)].

Note that the leakage current per unit transistor width is the same for both logic circuits and buffers. In addition, $I_{\text{off}}$ is a strong function of temperature. Therefore, temperature reduction can result in large savings in leakage power.

To summarize, for each technology node.

1) Assuming $V_{\text{t}_{\text{nom}}} = (1/4) V_{\text{DD}_{\text{nom}}}$, $N_a$ and $N_d$ are calculated using (6) and (7).
2) $\mu$ and $I_{\text{off}}$ are calculated at nominal temperature and $V_{\text{DD}}$ using (8) and (9).
3) $f_{\text{clk}}$ is assumed to be inversely proportional to $(\tau/l)_{\text{opt}}$. At nominal $V_{\text{DD}}$ and $T$, $f_{\text{clk}}$ is assumed to be the ITRS specified clock speed. This value of $f_{\text{clk}}$ and $(\tau/l)_{\text{opt}}$ are used to determine the proportionality constant.
4) Switching, leakage and short-circuit power are calculated using the above assumptions for logic circuits for a minimum-sized inverter driving a fan-out of four identical minimum-sized inverters at nominal $V_{\text{DD}}$ and temperature. This determines the fraction of switching, leakage and short-circuit power for the logic blocks at nominal $V_{\text{DD}}$ and temperature (see Table II).
5) Assuming that 30% power is consumed by the repeaters at each technology node, the above ratio is used to calculate the total switching, leakage, and short-circuit power for logic blocks. This is used to back calculate $C_{L_{\text{logic}}}$, $W_{\text{n}}$, and $W_{\text{p}}$ for each technology node.
6) Total repeater power and the power dissipation of a single repeater is used to estimate the number of repeaters $(M_{\text{repeater}})$. This is used to determine the fraction $p$ of global lines that are optimally buffered at nominal $V_{\text{DD}}$ and temperature as follows:

$$M_{\text{repeater}} = p \frac{L}{W_{\text{int}} + S_{\text{int}}} \times \frac{L}{l} \times G$$

where $L$ is the chip edge, $S_{\text{int}}$ is the global interconnect spacing, and $G$ the total number of global interconnect levels.

TABLE II
RELATIVE CONTRIBUTION OF THE THREE COMPONENTS OF OVERALL POWER DISSIPATION FOR LOGIC BLOCKS AND REPEATERS AT NOMINAL $V_{\text{DD}}$ AND TEMPERATURE

| | Logic Blocks | | | Repeaters | | |
|---|---|---|---|---|---|---|
| Technology Node (nanometer) | 130 | 90 | 65 | 130 | 90 | 65 |
| Switching | 0.874 | 0.791 | 0.445 | 0.811 | 0.763 | 0.551 |
| Short-circuit | 0.092 | 0.087 | 0.062 | 0.170 | 0.167 | 0.152 |
| Leakage | 0.035 | 0.123 | 0.493 | 0.018 | 0.069 | 0.297 |

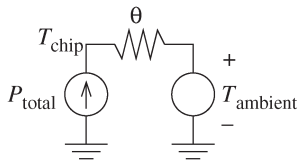



Fig. 4. Package thermal model.

## V. Chip Thermal Model

We saw in the previous section that power dissipation is a strong function of temperature. The chip temperature, however, is linearly dependent of the total power dissipation of the chip. The thermal equivalent circuit of the chip and the package is shown in Fig. 4, where $T_{\rm chip}$ is the chip temperature, $T_{\rm ambient}$ is the ambient temperature, $\vartheta$ is the package thermal coefficient, and $P_{\rm total}$ is the total chip power consumption. In this model, the total power consumption of a chip corresponds to the value of the current source, the temperature corresponds to the node-voltage value, and the package thermal coefficient corresponds to the resistor value. Therefore, for a given package

$$T_{\rm chip} = T_{\rm ambient} + \vartheta P_{\rm total}. \tag{10}$$

This model assumes that the whole chip is at a uniform temperature. In the above model, $\vartheta$ is computed using the die temperature and power dissipation in present-generation VLSI listed in the ITRS roadmap.

Fig. 5 plots the total power dissipation at various supply voltages as a function of temperature and (10) for the 130- and 65-nm technology nodes. Note that the total power consumption of the 65-nm node is a stronger function of temperature than that of the 130-nm node. This is due to that fact the leakage power is a more significant fraction of total power dissipation for the 65-nm node. These curves predict that reduction in temperature results in significant amount of power savings in future technology. The chip temperature and the actual total power consumption of the chip with a given supply voltage are determined by the intersection of the total power dissipation curve with (10). This intersection point can be numerically obtained by simultaneously solving (10) and the power equation using Newton–Raphson's method. When the supply voltage is 20% higher than the nominal $V_{\rm DD}$, note that those two curves do not intersect for either of the technology nodes. This shows that the package is not adequate to maintain the die temperature and $1.2V_{\rm DD_{nom}}$ and this results in thermal runaway and failure of the chip.

## VI. Optimization Methodology

It was shown in the previous section that each component of power consumption is a function of temperature. Reduction of the supply voltage reduces the chip total power consumption, which reduces the chip temperature. As the chip temperature reduces, the leakage power reduces dramatically. It has been empirically observed from SPICE simulation that the

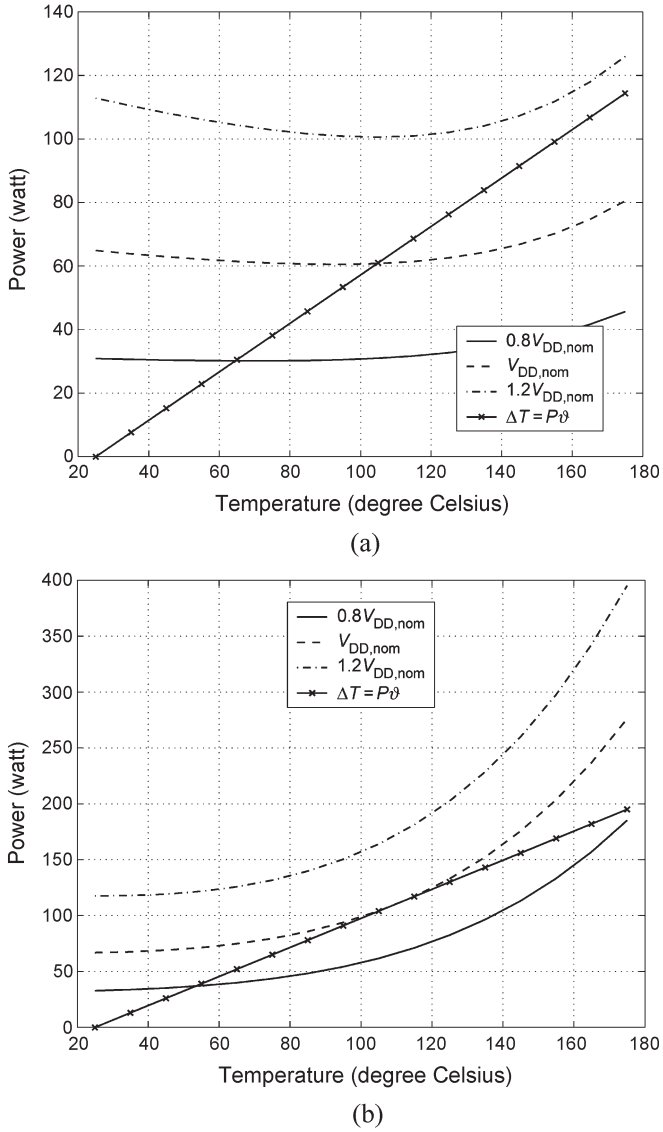



Fig. 5. Chip power dissipation and temperature. (a) Power consumption versus temperature for different supply voltages for the 130-nm technology. (b) Power consumption versus temperature for different supply voltages for the 65-nm technology.

performance improves as the device temperature is reduced. Reduction of supply voltage, however, reduces the on-state current, which degrades the performance. Therefore, as the power supply is increased from a very small value, initially the performance will improve but beyond a certain value of $V_{\rm DD}$, the power dissipation, and therefore the chip temperature, will increase rapidly, which will degrade performance. We therefore want to determine the optimal value of $V_{\rm DD}$ where the performance will be maximum, i.e., the delay per unit length will be minimum.

As pointed out earlier, we consider two cases: 1) chip design is not complete and therefore $s_{\rm opt}$ and $l_{\rm opt}$ can be chosen for optimal delay per unit length at the desired $V_{\rm DD}$ and temperature; and 2) the chip has been designed and optimally buffered using $s_{\rm opt}$ and $l_{\rm opt}$ calculated for the nominal $V_{\rm DD}$ and temperature of 105 °C, but its power supply can be externally varied for optimal performance.

The power consumption of logic blocks and the repeaters are

$$P_{\text{logic}} = k_1 \sum C_{\text{logic}} + k_2 r_s (c_0 + c_{\text{p}}) \sum W_{\text{n}} + k_3 \left[ I_{\text{off}_{\text{n}}} \sum W_{\text{n}} + I_{\text{off}_{\text{p}}} \sum W_{\text{p}} \right]$$

$$P_{\text{repeater}} = M_{\text{repeater}} \Bigg( k_1 \left( s(c_0 + c_{\text{p}}) + lc \right) + k_2 s \left( I_{\text{off}_{\text{n}}} W_{\text{n}_{\min}} + I_{\text{off}_{\text{p}}} W_{\text{p}_{\min}} \right) + k_3 \left( r_s (c_0 + c_{\text{p}}) + \frac{r_s}{s} cl + rlsc_0 + \frac{1}{2} rcl^2 \right) s W_{\text{n}_{\min}} \Bigg)$$

where

$$k_1 = \alpha V_{\text{DD}}^2 f_{\text{clk}}$$

$$k_2 = \frac{3}{2} V_{\text{DD}} I_{\text{off}_{\text{n}}} W_{\text{n}_{\min}}$$

$$k_3 = \alpha V_{\text{DD}} W_{\text{n}_{\min}} I_{\text{short circuit}} f_{\text{clk}} \log_{\text{e}} \left( \frac{V_{\text{DD}} - |V_{\text{t}_{\text{p}}}|}{V_{\text{t}_{\text{n}}}} \right).$$

For case 2), we assume the buffer scheme is designed to be optimal at nominal supply voltage and at a temperature of 105 °C and therefore $M_{\text{repeater}}$, $s$, and $l$ are fixed. For case 1), we generate expressions of $s_{\text{opt}}$, $l_{\text{opt}}$, and $(\tau/l)_{\text{opt}}$ in terms of supply voltage and temperature by SPICE simulation. For a given supply voltage, we can find the chip temperature by solving the following equation

$$T = T_{\text{nom}} + \vartheta \left( P_{\text{logic}}(T) + P_{\text{repeater}}(T) \right) \tag{11}$$

where $P_{\text{logic}}(T)$ and $P_{\text{repeater}}(T)$ are nonlinear functions of $T$. $\vartheta$, the package thermal resistance, is chosen such that the total power dissipation is ITRS predicted power at the nominal supply voltage and 105 °C.

The above methodology can also be modified for other optimization scenarios. For instance, consider minimizing the total power dissipation similar to the one presented in [2], i.e., minimizing total power dissipation for a given delay per unit length, but taking into account the chip temperature change due to the change in total power dissipation. For a given $\tau/l$ and $V_{\text{DD}}$, we find $s$ and $l$ that minimize $P_{\text{total}}$ in (5) subject to (2) and (11). Within this framework, various scenarios are considered:

1) $V_{\text{DD}}$ fixed at $V_{\text{DD}_{\text{nom}}}$;
2) $V_{\text{DD}}$ reduced to minimum supply voltage for which $\tau/l$ can be achieved; and
3) $V_{\text{DD}}$ reduced to minimum supply voltage for which $\tau/l$ can be achieved but $V_{\text{DD}} \geq V_{\text{DD,min}}$.

The last scenario disallows the reduction of $V_{\text{DD}}$ beyond a certain limit determined by $IR$ and $L(\text{d}I/\text{d}t)$ drops and noise margins of the logic.

## VII. Results

Fig. 6 shows the delay per unit length as a function of power-supply voltage for the 130-, 90-, and 65-nm technology nodes. Note that the optimal supply voltage for both cases 1) and 2) is slightly higher than the nominal supply voltage for the 130-nm node. This is due to the fact that the leakage power only contributes approximately 3.5% of the total power consumption for this node (Table II). As leakage power becomes a significant portion of the total power consumption, this optimum point shifts to the left. It is found that the optimal supply voltage is only 96% of the nominal supply voltage for case 1) and case 2) of the 65-nm technology node. This implies that as leakage power becomes dominant, decreasing the supply voltage from the nominal value improves performance. This also has the added benefit of decreasing the power dissipation and chip temperature, and therefore improving the reliability of the chip. These results also suggest that even if the chip is optimized for operation at nominal $V_{\text{DD}}$ and temperature, operating it at a lower supply voltage can improve performance. Note that the optimum values of $\tau/l$ are very similar for cases 1) and 2) for every technology node.

The interconnect parameters and the nominal supply voltage are based on ITRS [20], and are shown in Table I. $S_{\text{int}}$ is assumed to be equal to the minimum width of the global interconnect. The absolute value of $V_{\text{t}_{\text{n}}}$ and $V_{\text{t}_{\text{p}}}$ are assumed to be the same and are equal to $(1/4)V_{\text{DD}}$ at nominal supply voltage and temperature of 105 °C. Table III shows $(\tau/l)_{\text{opt}}/(\tau/l)_{\text{nom}}$, i.e., the ratio of delay per unit length at optimum $V_{\text{DD}}$ and the delay per unit length at the nominal $V_{\text{DD}}$ and temperature, and $P_{\text{opt}}/P_{\text{nom}}$, i.e., the ratio of total power consumption at the optimum $V_{\text{DD}}$ and the total power consumption with the nominal $V_{\text{DD}}$ and temperature. Note that both performance and total power consumption improve at the optimal supply voltage for the 90- and 65-nm technologies.

Now consider power optimization for fixed delay per unit length. As indicated earlier, we consider three scenarios:

Case 1) $V_{\text{DD}}$ fixed at $V_{\text{DD}_{\text{nom}}}$;
Case 2) $V_{\text{DD}}$ reduced to minimum supply voltage for which $\tau/l$ can be achieved; and
Case 3) $V_{\text{DD}}$ reduced to minimum supply voltage for which $\tau/l$ can be achieved but $V_{\text{DD}} \geq 0.8 V_{\text{DD}_{\text{nom}}}$.

For comparison purposes only, we also consider the case when both $V_{\text{DD}}$ and $T$ are fixed at their nominal values [case 1)]. Fig. 7 plots the power dissipation as a function of delay per unit length $(\tau/l)$ for the three technology nodes for the four cases. $\tau/l$ is varied from $(\tau/l)_{\text{opt,nom}}$ at nominal $V_{\text{DD}}$ and temperature to $1.1(\tau/l)_{\text{opt,nom}}$. Note that for all the technology nodes, power dissipation reduces as $\tau/l$ increases and the power dissipation is the least in case 4) and the highest in case 1). Even if $V_{\text{DD}}$ is not allowed to decrease below $0.8V_{\text{DD}_{\text{nom}}}$, the power dissipation in cases 3) and 4) are very similar. The power dissipation of case 2) is somewhere in between case 1) and cases 3)–4). Also note that for all three technology nodes, the power dissipation in case 2) is lower than case 1) even at $\tau/l = (\tau/l)_{\text{opt,nom}}$. This is due to that fact that a slight decrease in temperature results in the same $\tau/l$ but lower power dissipation.

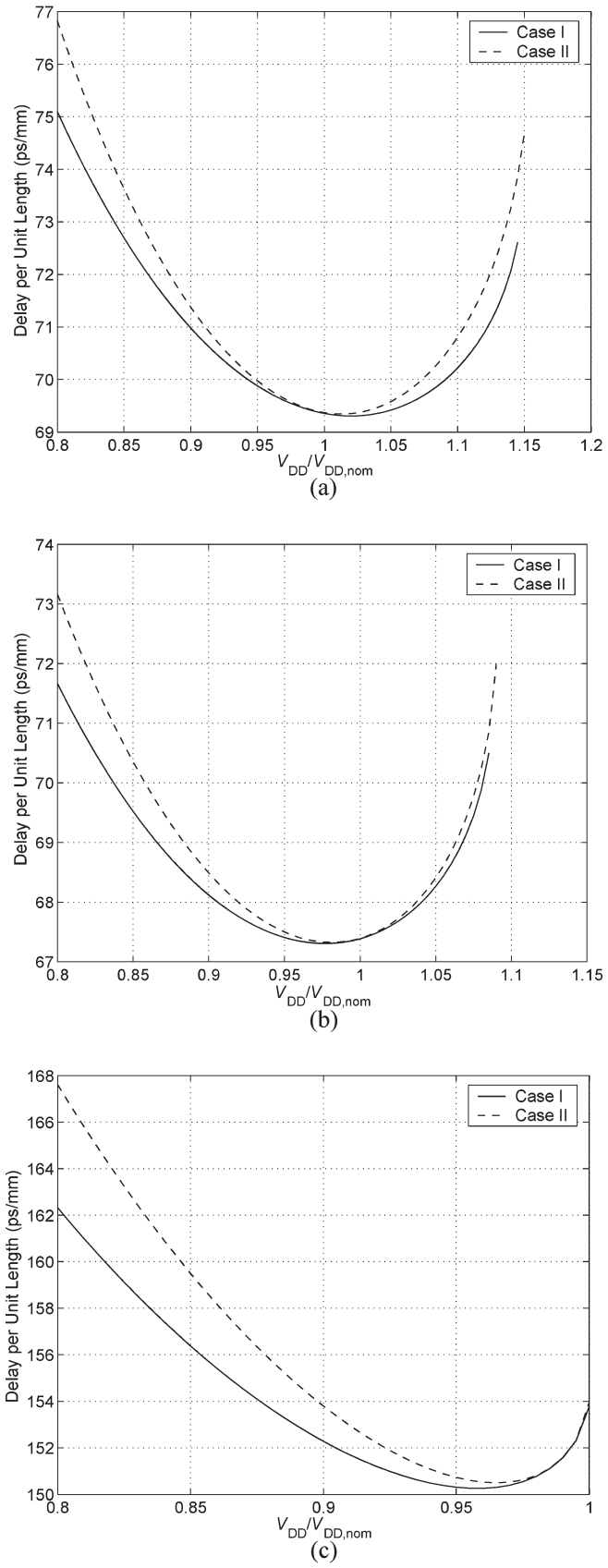



Fig. 6. Performance versus supply voltage for various technologies. (a) Delay per unit length as a function of supply voltage for the 130-nm technology. (b) Delay per unit length as a function of supply voltage for the 90-nm technology. (c) Delay per unit length as a function of supply voltage for the 65-nm technology.

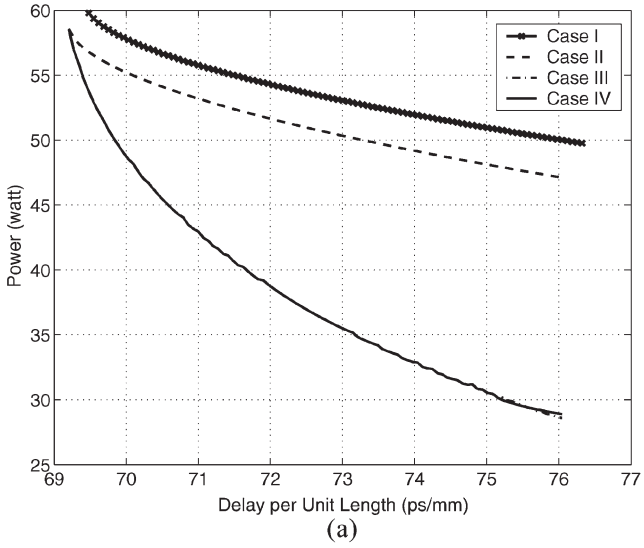
TABLE III
RATIO OF TOTAL POWER CONSUMPTION AND DELAY PER UNIT LENGTH WITH OPTIMAL SUPPLY VOLTAGE AND WITH NOMINAL SUPPLY VOLTAGE FOR VARIOUS TECHNOLOGY NODES

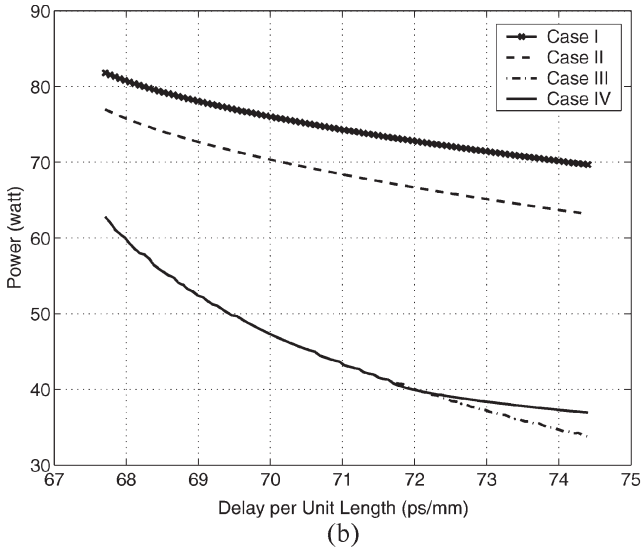
| | Case I | | | Case II | | |
|---|---|---|---|---|---|---|
| Technology Node (nanometer) | $\frac{\left(\frac{\tau}{l}\right)_{\rm opt}}{\left(\frac{\tau}{l}\right)_{\rm nom}}$ | $\frac{P_{\rm opt}}{P_{\rm nom}}$ | $\frac{V_{\rm DD\ opt}}{V_{\rm DD\ nom}}$ | $\frac{\left(\frac{\tau}{l}\right)_{\rm opt}}{\left(\frac{\tau}{l}\right)_{\rm nom}}$ | $\frac{P_{\rm opt}}{P_{\rm nom}}$ | $\frac{V_{\rm DD\ opt}}{V_{\rm DD\ nom}}$ |
| 130 | 0.9996 | 1.0468 | 1.015 | 0.9993 | 1.0619 | 1.02 |
| 90 | 0.999 | 0.9331 | 0.98 | 0.9988 | 0.9189 | 0.975 |
| 65 | 0.9776 | 0.7393 | 0.965 | 0.9772 | 0.7315 | 0.96 |

Note that the difference in power dissipation between cases 1) and 2) increases with technology scaling. Recall that the only difference between these two cases is that for case 2), the temperature is calculated in a self-consistent manner using (5) and (11). The resulting temperature reduction results in a large reduction in power dissipation for the 65-nm technology node where the leakage power contribution is the highest. Furthermore, note that as $\tau/l$ is increased, power dissipation for cases 3) and 4) reduces very rapidly for the 130-nm technology node whereas the reduction in the 90- and 65-nm technology nodes is not that large. To further explain this observation we have plotted $V_{\rm DD}$ versus $\tau/l$ for the three technology nodes as the delay per unit length is varied (Fig. 8). Note that at $\tau/l = (\tau/l)_{\rm opt,nom}$, power-supply voltage for the 130-nm node is almost $V_{\rm DD}$ whereas for the 90- and 65-nm nodes, it is 91% and 88%, respectively. This is due to the fact that for the 90- and 65-nm nodes, due to the dominance of leakage power, a reduction in the supply voltage causes a large reduction in leakage power and die temperature, which causes a net improvement in $(\tau/l)_{\rm opt}$ with the reduction in $V_{\rm DD}$. Therefore, even for $\tau/l = (\tau/l)_{\rm opt,nom}$ (i.e., at nominal $V_{\rm DD}$ and temperature), $V_{\rm DD}$ can be reduced and significant power savings can be obtained for these nodes. However, as $\tau/l$ is increased, further power optimization does not yield as impressive a power gain as for the 130-nm technology node.

## VIII. CONCLUSION

In conclusion, we have developed a methodology for calculating the optimal supply voltage that either minimizes the delay per unit length or minimizes power dissipation for a given delay per unit length while considering the total chip power dissipation and temperature rise in a consistent manner. The performance-optimization methodology is demonstrated for two cases: 1) the design can be optimally buffered for the target $V_{\rm DD}$ and temperature and 2) when the design is buffered for a fixed $V_{\rm DD}$ and temperature. Using this methodology, we have computed the optimal operating voltage for the 130-, 90-, and 65-nm technology nodes for both cases. Furthermore, we have shown that as the technology is scaled beyond the 130-nm technology, the supply voltage at which the performance is optimal is below the nominal supply voltage. This is due to the fact that leakage power is becoming a significant fraction of the total power consumption. As the supply voltage reduces to the optimal point, the chip's temperature is reduced, which



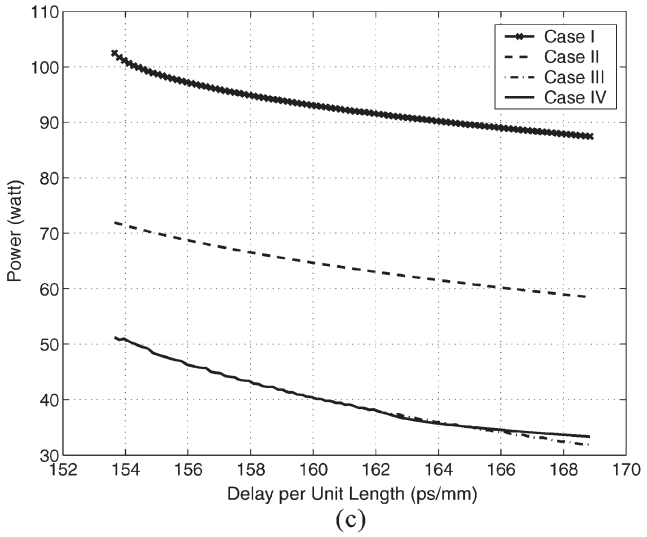




Fig. 7. Total power dissipation as a function of delay per unit length for the 130-, 90-, and 65-nm ITRS technology nodes. (a) Power versus $\tau/l$ for the 130-nm node. (b) Power versus $\tau/l$ for the 90-nm node. (c) Power versus $\tau/l$ for the 65-nm node.

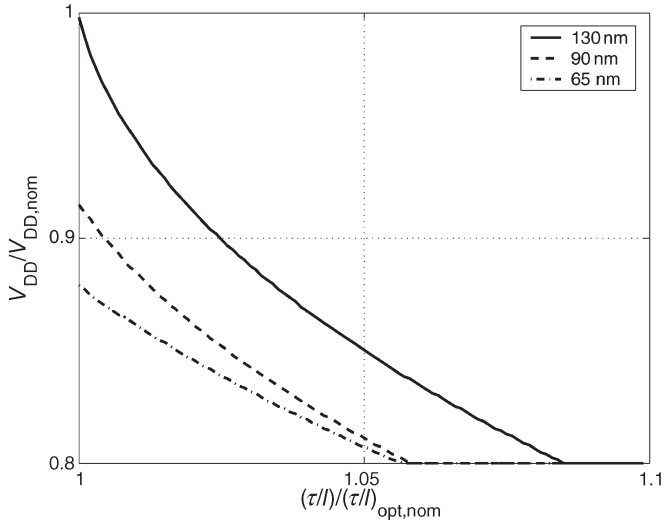



Fig. 8. $V_{\rm DD}$ versus $\tau/l$ for the three technology nodes for case 4).

results in reduction of the leakage power and improvement of performance. It is also shown that increasing the supply voltage beyond a certain threshold for a given package results in thermal runaway and failure of the chip. The power-optimization methodology has been demonstrated for various constraints on the supply voltages. We have shown that lowering the power supply results in a large reduction in the total power dissipation, and therefore chip temperature, while still maintaining performance, i.e., delay per unit length. This reduction increases with technology scaling because leakage power, which increases exponentially with die temperature, becomes a larger fraction of the total power dissipation.

## REFERENCES

[1] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, and S. Borkar, "Effectiveness and scaling trends of leakage control techniques for sub-100 nm CMOS technologies," in *Proc. Int. Symp. Low-Power Electronics*, Seoul, Korea, 2003, pp. 122–127.

[2] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Trans. Electron Devices*, vol. 49, no. 11, pp. 2001–2007, Nov. 2002.

[3] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*. Upper Saddle River, NJ: Prentice-Hall, 2003.

[4] Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[5] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[6] R. W. Brodersen, M. A. Horowitz, D. Markovic, B. Nikolic, and V. Stojanovic, "Methods for true power minimization," in *Dig. Tech. Papers, IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 2002, pp. 35–42.

[7] V. Adler and E. G. Friedman, "Repeater design to reduce delay and power in resistive interconnect," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 45, no. 5, pp. 607–616, May 1998.

[8] A. Nalamalpu and W. Burleson, "A practical approach to DSM repeater insertion: Satisfying delay constraints while minimizing area and power," in *Proc. 14th Annu. IEEE Int. Application-Specified Integrated Circuit/System-on-a-Chip (ASIC/SOC) Conf.*, Arlington, VA, 2001, pp. 152–156.

[9] M. R. Stan, "Optimal voltages and sizing for low power," in *Proc. 12th Int. Conf. VLSI Design*, Goa, India, 1999, pp. 428–433.

[10] A. Chandrakasan and R. W. Brodersen, *Low Power Digital CMOS Design*. Boston, MA: Kluwer, 1995.

[11] N. Sirisantana, L. Wei, and K. Roy, "High-performance low-power CMOS circuits using multiple channel length and multiple oxide thickness," in *Proc. Int. Conf. Computer Design*, Austin, TX, 2000, pp. 227–232.
[12] A. P. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*. New York: IEEE Press, 2000.
[13] M. Johnson, D. Somasekhar, and K. Roy, "Models and algorithms on bounds of leakage in CMOS circuits," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 18, no. 6, pp. 714–725, Jun. 1999.
[14] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS ICs," in *Proc. Int. Symp. Low Power Electronics and Design*, Huntington Beach, CA, 2001, pp. 207–212.
[15] K. Banerjee, S.-C. Lin, A. Keshavarzi, S. Narendra, and V. De, "A self-consistent junction temperature estimation methodology for nanometer scale ics with implications for performance and thermal management," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, 2003, pp. 887–890.
[16] G. S. Garcea, N. P. van der Meijs, and R. H. J. M. Otten, "Simultaneous analytic area and power optimization for repeater insertion," in *Dig. Tech. Papers, Int. Conf. Computer-Aided Design*, San Jose, CA, 2003, pp. 568–573.
[17] R. Li, D. Zhou, J. Liu, and X. Zheng, "Power-optimal simultaneous buffer insertion/sizing and wire sizing," in *Dig. Tech. Papers, Int. Conf. Computer-Aided Design*, San Jose, CA, 2003, pp. 581–586.
[18] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addision-Wesley, 1990.
[19] K. Banerjee and A. Mehrotra, "Analysis of on-chip inductance effects for distributed RLC interconnects," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 21, no. 8, pp. 904–915, Aug. 2002.
[20] *International Technology Roadmap for Semiconductors (ITRS)*, Semiconductor Industry Assoc., San Jose, CA, 1999.
[21] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and system-on-chip integration," *Proc. IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.
[22] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. 19, no. 4, pp. 468–473, Aug. 1984.
[23] S. M. Sze, *Physics of Semiconductor Devices*. New York: Wiley, 1981.
[24] R. S. Muller and J. I. Kamins, *Device Electronics for Integrated Circuits*. New York: Wiley, 1984.

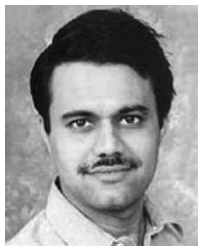
**Man Lung Mui** was born in Hong Kong, SAR. He received the B.S. degree in electrical engineering in 2002 from the University of Illinois, Urbana–Champaign, where he is currently pursuing the M.S. degree in electrical engineering with an emphasis in integrated circuit design.

In 2002, he joined the Illinois Center for the Integrated Micro-Systems Group, Coordinated Science Laboratory, University of Illinois, as a Research Assistant. His research focuses on interconnect performance and modeling for very large scale integration (VLSI) circuit designs.

**Kaustav Banerjee** (S'92–M'99–SM'03) received the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley in 1999.

He was with Stanford University from 1999 to 2002 as a Research Associate at the Center for Integrated Systems. In July 2002, he joined the faculty of the Electrical and Computer Engineering Department at the University of California, Santa Barbara (UCSB), where he is currently an Associate Professor. From February 2002 to August 2002, he was a Visiting Professor at the Circuit Research Labs of Intel in Hillsboro, OR. In the past, he had also held summer/visiting positions at Texas Instruments, Fujitsu Labs, and the Swiss Federal Institute of Technology in Lausanne (EPFL). His research has been chronicled in over 100 journal and refereed international conference papers and a book chapter. He also coedited a book titled *Emerging Nanoelectronics: Life with and after CMOS* (Boston, MA: Kluwer, 2005). His present research interests focus on nanometer-scale issues in high-performance VLSI as well as on circuits and systems issues in emerging nanoelectronics.

Dr. Banerjee serves or has served on the technical program committees of the IEDM, IRPS, EOS/ESD Symposium, and ISPD. He has also served on the organizing committee of the IEEE International Symposium on Quality Electronic Design (ISQED), at various positions including Technical Program Chair (2002) and General Chair (2005). He has been recognized through the ACM SIGDA Outstanding New Faculty Award (2004) as well as a Best Paper Award at the Design Automation Conference (2001). He is listed in *Who's Who in America* and *Who's Who in Science and Engineering*.



**Amit Mehrotra** (S'96–M'99) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1994 and the Masters and Ph.D. degrees from the Department of Electrical Engineering and Computer Science, University of California at Berkeley in 1996 and 1999, respectively.

In August 1999, he joined the University of Illinois at Urbana-Champaign as an Assistant Professor with the Department of Electrical and Computer Engineering and as a Research Assistant Professor with the Illinois Center for Integrated Micro-Systems group at the Coordinated Science Laboratory. In January 2004, he joined Berkeley Design Automation, Santa Clara, CA, a company he cofounded in 2003, as the Chief Technology Officer. His research interests include radio frequency (RF), analog and mixed-signal circuit design for mobile communication systems, simulation techniques for RF and mixed-signal circuits and systems, interconnect performance and modeling issues in VLSI and novel circuits and physical design issues for high-performance VLSI designs, model-order reduction of linear and nonlinear circuits. He has authored and coauthored over 40 technical papers in archival journals and refereed international conference proceedings. He also coauthored a book titled *Noise Analysis of Radio Frequency Circuits* (Boston, MA: Kluwer, 2004).

Dr. Mehrotra has served as the Technical Program Committee Member of the International Symposium on Quality Electronic Design since 2002. He received Best Paper Awards at the 1997 International Conference on Computer Design and 2001 Design Automation Conference.