

# Power Supply Optimization in Sub-130 nm Leakage Dominant Technologies

Man L Mui

Dept. of Electrical and Computer Engineering  
University of Illinois  
Urbana, IL 61801

Kaustav Banerjee

Dept. of Electrical and Computer Engineering  
University of California  
Santa Barbara, CA 93106

Amit Mehrotra

Dept. of Electrical and Computer Engineering  
University of Illinois  
Urbana, IL 61801

## Abstract

In this paper we present a methodology for systematically optimizing the power supply voltage for maximizing the performance of VLSI circuits in technologies where leakage power is not an insignificant fraction of the total power dissipation. For this purpose, we develop simplified empirical equations which describe the transistor behaviour 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  $V_{DD}$  value which minimized delay per unit length in global interconnects and therefore maximizes performance. We show that for 90 nm and 65 nm technologies where leakage power represents a significant fraction of the total power dissipation, optimum  $V_{DD}$  is lower than the ITRS specified supply voltage. This is due to the fact that reducing  $V_{DD}$  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  $V_{DD}$ .

## 1 Introduction

As the channel lengths of MOS devices scale below 180 nm, leakage current becomes non-negligible 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 micron 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 5 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 of 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  $V_{DD}$ . In Section 7 we show that reducing the supply voltage slightly results in an improvement in performance for 90 nm and 65 nm nanometer technology nodes.

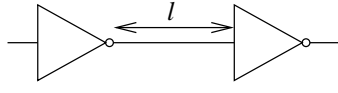
In this work, we develop a methodology to estimate the optimal supply voltage which maximizes circuit performance. For this purpose we first develop simplified empirical models for device equivalent resistance, parasitic capacitance and output capacitance as a function of temperature and  $V_{DD}$  which results in a model for circuit performance as a function of  $V_{DD}$  and temperature (Section 3). 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 4). 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  $V_{DD}$ . By varying  $V_{DD}$  we find the optimal supply voltage which maximizes performance. We consider two typical cases in global interconnect optimization (I) when the buffer insertion can be optimized for the target  $V_{DD}$  and temperature and (II) when buffering scheme is fixed and is designed to be optimal at nominal supply voltage and at temperature of 105°C. We show that the optimal supply voltage which reduces power dissipation is smaller than the nominal  $V_{DD}$  for 90 nm and 65 nm technology nodes.

## 2 Previous Works

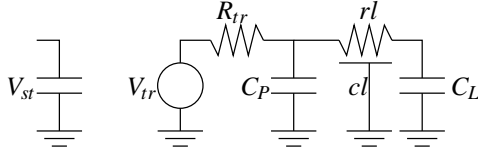
Several techniques have been proposed for reducing the off-state current [5, 6]. These include reducing power supply [7, 8], using non-minimum channel length transistors [9], using stacked transistors [10, 11] and reverse body bias [12]. A comprehensive analysis of the effectiveness of these techniques 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 [13]. 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.

## 3 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



(a) Schematic representation.



(b) Equivalent RC circuit.

Figure 1: Interconnect of length  $l$  between two identical inverters.

Figure 1. Assume that for a minimum sized repeater, the input capacitance is  $c_0$ , the output parasitic capacitance is  $c_p$  and output resistance is  $r_s$ . Therefore for a repeater of size  $s$ , the total output resistance  $R_{tr} = \frac{r_s}{s}$ , the total output parasitic capacitance  $C_P = c_p s$  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 [14]

$$\tau = r_s(c_0 + c_p) + \frac{r_s}{s}cl + r_lsc_0 + \frac{1}{2}rcl^2 \quad (1)$$

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

Now consider a long interconnect of a given length  $L$  which is uniformly buffered with inter-buffer interconnect length  $l$ . Therefore the total number of segments is  $\frac{L}{l}$ . The total delay through that line is given by

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

where  $\frac{\tau}{l}$  is the delay per unit length which is given by

$$\frac{\tau}{l} = \frac{1}{l}r_s(c_0 + c_p) + \frac{r_s}{s}c + rsc_0 + \frac{1}{2}rcl$$

Note that optimizing the delay of the interconnect of a fixed length is equivalent to optimizing  $\frac{\tau}{l}$ . This delay per unit length is optimal when [14]

$$l_{opt} = \sqrt{\frac{2r_s(c_0 + c_p)}{rc}} \quad s_{opt} = \sqrt{\frac{r_sc}{rc_0}} \quad (2)$$

and is given by

$$\left(\frac{\tau}{l}\right)_{opt} = 2\sqrt{r_sc_0rc} \left(1 + \sqrt{\frac{1}{2} \left(1 + \frac{c_p}{c_0}\right)}\right) \quad (3)$$

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

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

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

Repeater parameters at various temperatures and supply voltages were extracted using SPICE simulations similar to [15]. A five stage ring oscillator with a given length of global interconnect of width  $W_{min}$  (see Table 2

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  $c_p$  were calculated from these values of  $s_{opt}$ ,  $l_{opt}$  and  $\left(\frac{\tau}{l}\right)_{opt}$  for a given supply voltage and temperature. Figure 2 plots  $r_s$ ,  $c_0$  and  $c_p$  as the power supply is varied  $\pm 20\%$  from the nominal value and the temperature is varied from  $25^\circ \text{C}$  to  $125^\circ \text{C}$ . Note that, as expected, the dependence of  $c_0$  on  $V_{DD}$  and temperature is very weak. Using curve fitting, we generate the expressions of  $r_s$ ,  $c_0$  and  $c_p$  in terms of supply voltage and temperature.

## 4 Power Model

In this work we will consider two cases (I) global interconnects are optimally buffered for the targeted power supply and temperature and (II) global interconnects are optimally buffered for operation at the nominal power supply and temperature. For scenario (I), changing the temperature and supply voltage will change  $s_{opt}$  and  $l_{opt}$  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

$$P_{total} = P_{logic} + P_{repeater}$$

where  $P_{repeater}$  denotes the total power dissipated in the buffers and global interconnects driven by these buffers and  $P_{logic}$  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 [8],

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

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 is varied.

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

$$P_{switching} = \alpha(s(c_p + c_0) + lc)V_{DD}^2 f_{clk}$$

where  $V_{DD}$  is the power supply voltage,  $f_{clk}$  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 [8]. For optimally sized and placed buffers,  $C_L$  is given in [2],

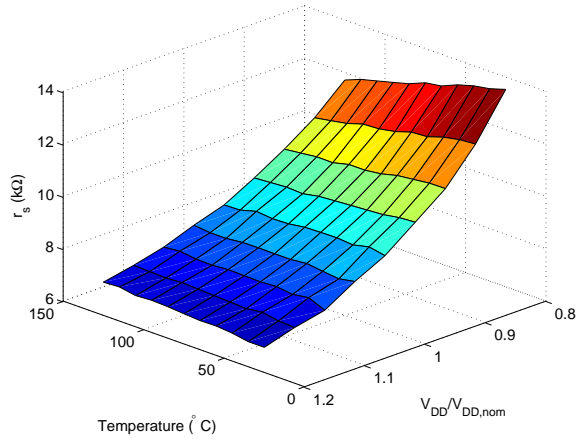
$$C_L = s_{opt}(c_0 + c_p) + cl_{opt}$$

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

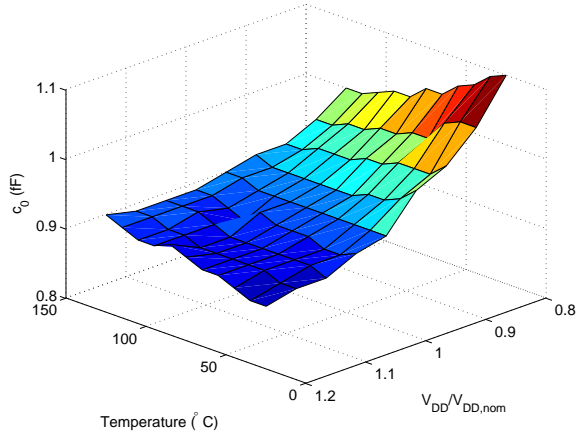
For the logic blocks, we assume that the load capacitance does not vary with temperature and  $V_{DD}$ . It is a valid assumption since the fan-outs of gates of the functional blocks are usually greater than one in general. The loading capacitance, therefore, is dominated by gate capacitance, which has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

The clock frequency  $f_{clk}$  is inversely proportional to the delay of critical path of the circuit. It has been shown in [16, 17] that the performance is dominated by global interconnects. Therefore,  $f_{clk}$  can be assumed to be inversely proportional to  $\left(\frac{\tau}{l}\right)_{opt}$ , which in turn is a function of supply voltage and temperature.

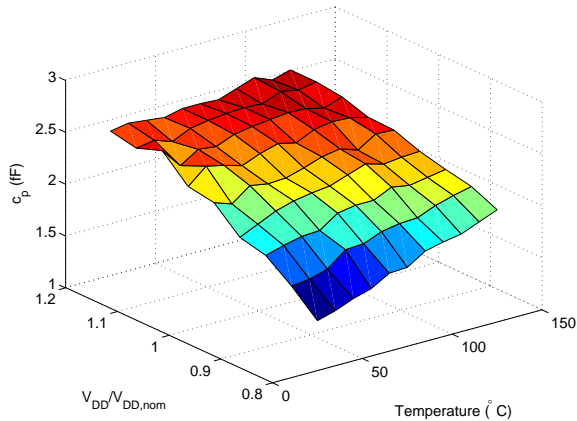
The second component is the short circuit power. This power consumption is incurred when both pull up network and pull down network are simultaneously on. Consider the simplest static CMOS logic circuit, an inverter, which is shown in Figure 3(a). When the NMOS transistor turns on due to a rising waveform at the input and the PMOS transistor continues to conduct current until the input voltage becomes greater than  $V_{DD} - |V_{tp}|$ ,



(a)  $r_s$  as a function of temperature and  $V_{DD}$

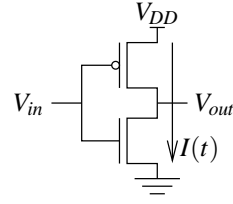


(b)  $c_0$  as a function of temperature and  $V_{DD}$

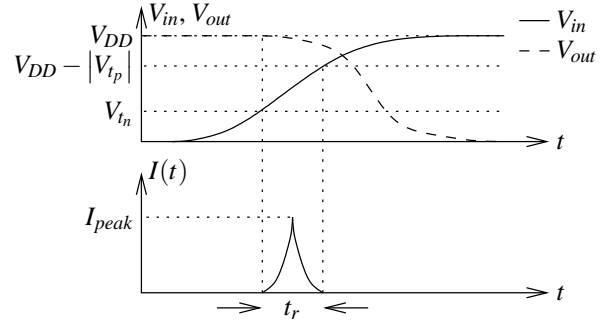


(c)  $c_p$  as a function of temperature and  $V_{DD}$

Figure 2: Temperature and supply voltage dependence of buffer parameters for 130 nm technology.



(a) CMOS inverter.



(b) Voltage and current waveforms.

Figure 3: Voltage and current waveforms of a CMOS inverter.

both transistors are on simultaneously. Hence, there is a DC current 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 waveform, and the current waveform are shown in Figure 1(b). The short circuit current waveform can be approximated as triangular wave [18]. The total charge that flows in this period can be found by calculating the area of this triangle. Let  $t_r$  denote the time for the input voltage to rise from  $V_{in}$  to  $V_{DD} - |V_{tp}|$ . 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

$$P_{short\ circuit} = \alpha t_r V_{DD} I_{peak} f_{clk} = \alpha t_r V_{DD} W_{nmin} s I_{short\ circuit} f_{clk}$$

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

$$t_r = \tau \log_e \left( \frac{V_{DD} - |V_{tp}|}{V_{in}} \right)$$

where  $\tau$  is the time constant for the output node, which is defined in Section 3. 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

$$\tau_{logic} \approx r_s (c_0 + c_p)$$

Note that  $I_{short\ circuit}$  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],

$$V_t = -\frac{E_g}{2q} + \phi_B + \frac{\sqrt{4\epsilon_{Si}qN\phi_B}}{C_{ox}} \quad (4)$$

where  $\epsilon_{Si}$  is the permittivity of silicon,  $N$  is the doping concentration,  $q$  the single electron charge,  $C_{ox}$  is gate-oxide capacitance,  $E_g$  is band-gap energy, which has the following temperature dependence [19]

$$E_g = 1.166 - \frac{4.73 \times 10^{-4} T^2}{T + 636}$$

$E_g$  in the above expression is in the units of eV.  $\phi_B$  is defined as

$$\phi_B = \frac{kT}{q} \log_e \left( \frac{N}{n_i} \right) = \frac{kT}{q} \log_e \left( \frac{N}{4.66 \times 10^{15} T^{1.5} \exp\left(-\frac{E_g}{2kT}\right)} \right) \quad (5)$$

where  $k$  is the Boltzmann constant and  $N$  is the doping concentration in  $\text{cm}^{-3}$ .

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

$$P_{\text{leakage}} = V_{DD} I_{\text{leakage}} = V_{DD} \frac{1}{2} (I_{\text{off}_n} W_n + I_{\text{off}_p} W_p)$$

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

$$I_{\text{off}} = \mu_{\text{eff}} C_{\text{ox}} \frac{W}{L_{\text{eff}}} \left( \frac{kT}{q} \right)^2 \exp(1.8) \exp \left( \frac{-V_t + \eta V_{DD}}{n \frac{kT}{q}} \right) \quad (6)$$

where  $\eta$  is the DIBL coefficient and  $n$  is the transistor sub-threshold swing coefficient. The temperature dependence of mobility is given by [20]

$$\begin{aligned} \mu_{n,\text{eff}} &= 88 T_n^{-0.57} + \frac{1250 T_n^{-2.33}}{1 + \frac{N_a}{1.26 \times 10^{17} T_n^{2.4}} \times 0.88 T_n^{-0.146}} \\ \mu_{p,\text{eff}} &= 54.3 T_n^{-0.57} + \frac{407 T_n^{-2.33}}{1 + \frac{N_d}{2.35 \times 10^{17} T_n^{2.4}} \times 0.88 T_n^{-0.146}} \end{aligned} \quad (7)$$

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

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

where  $\phi_B$  is a function of temperature (see (5)).

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,

- Assuming  $V_{t,\text{nom}} = \frac{1}{4} V_{DD,\text{nom}}$ ,  $N_a$  and  $N_d$  are calculated using (4) and (5).
- $\mu$  and  $I_{\text{off}}$  are calculated at nominal temperature and  $V_{DD}$  using (6) and (7).
- $f_{\text{clk}}$  is assumed to be inversely proportional to  $\left(\frac{\tau}{T}\right)_{\text{opt}}$ . At nominal  $V_{DD}$  and  $T$ ,  $f_{\text{clk}}$  is assumed to be the ITRS specified clock speed. This value of  $f_{\text{clk}}$  and  $\left(\frac{\tau}{T}\right)_{\text{opt}}$  are used to determine the proportionality constant.
- 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 4 identical minimum sized inverters at nominal  $V_{DD}$  and temperature. This determines the fraction of switching, leakage and short-circuit power for the logic blocks at nominal  $V_{DD}$  and temperature (see Table 1).
- 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  $C_{L,\text{logic}}$ ,  $W_n$  and  $W_p$  for each technology node.
- 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 which are optimally buffered at nominal  $V_{DD}$  and temperature as follows

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

	logic blocks			repeaters		
Tech. node (nm)	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

Table 1: Relative contribution of the three components of overall power dissipation for logic blocks and repeaters at nominal  $V_{DD}$  and temperature.

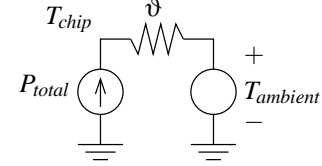


Figure 4: Package thermal model.

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

## 5 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 Figure 4, where  $T_{\text{chip}}$  is the chip temperature,  $T_{\text{ambient}}$  is the ambient temperature,  $\vartheta$  is the package thermal coefficient and  $P_{\text{total}}$  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

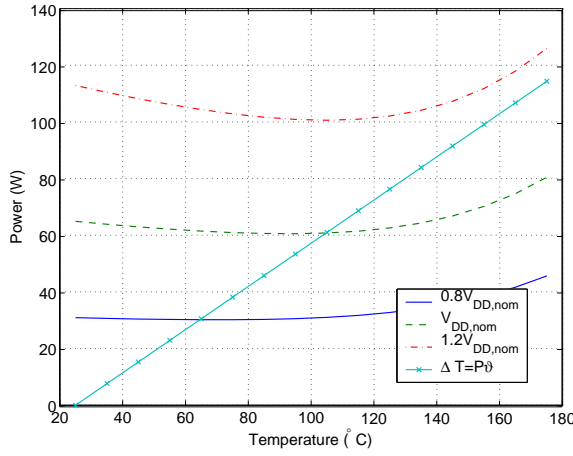
$$T_{\text{chip}} = T_{\text{ambient}} + \vartheta P_{\text{total}} \quad (8)$$

This model assumes that the whole chip is at a uniform temperature.

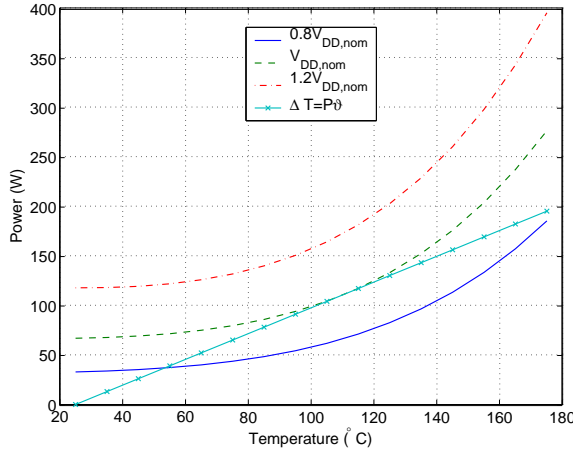
Figure 5 plots the total power dissipation at various supply voltages as a function of temperature and (8) for 130nm and 65nm technology nodes. Note that the total power consumption of 65nm node is a stronger function of temperature than that of 130nm node. This is due to that fact the leakage power is a more significant fraction of total power dissipation for 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 total power dissipation curve with (8). This intersection point can be numerically obtained by simultaneously solving (8) and the power equation using Newton-Raphson's method. When the supply voltage is 20% higher than the nominal  $V_{DD}$ , 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  $1.2V_{DD,\text{nom}}$  and this results in thermal runaway and failure of the chip.

## 6 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 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  $V_{DD}$ , the power dissipation and therefore the chip temperature will increase rapidly which will degrade performance. We therefore want to determine the *optimal* value of  $V_{DD}$  where the performance will be maximum, i.e., the delay per unit length will be minimum.



(a) Power consumption vs temperature for different supply voltages for 130nm technology



(b) Power consumption vs temperature for different supply voltages for 65nm technology

Figure 5: Chip power dissipation and temperature

As pointed out earlier, we consider two cases: (I) chip design is not complete and therefore  $s_{opt}$  and  $l_{opt}$  can be chosen for optimal delay per unit length at the desired  $V_{DD}$  and temperature and (II) the chip has been designed and optimally buffered using  $s_{opt}$  and  $l_{opt}$  calculated for the nominal  $V_{DD}$  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

$$P_{logic} = k_1 \sum C_{logic} + k_2 r_s (c_0 + c_p) \sum W_n + k_3 \left[ I_{off_n} \sum W_n + I_{off_p} \sum W_p \right]$$

$$P_{repeater} = M_{repeater} \left( k_1 (s(c_0 + c_p) + lc) + k_2 s (I_{off_n} W_{n_{min}} + I_{off_p} W_{p_{min}}) \right. \\ \left. + k_3 \left( r_s (c_0 + c_p) + \frac{r_s}{s} cl + rlsc_0 + \frac{1}{2} rcl^2 \right) s W_{n_{min}} \right)$$

where  $k_1 = \alpha V_{DD}^2 f_{clk}$ ,  $k_2 = \frac{3}{2} V_{DD} I_{off_n} W_{n_{min}}$  and  $k_3 = \alpha V_{DD} W_{n_{min}} I_{short\_circuit} f_{clk} \log_e \left( \frac{V_{DD} - |V_{tp}|}{V_{in}} \right)$ . For case II, we assume the buffer scheme is designed to be optimal at nominal supply voltage and at temperature of 105°C and therefore  $M_{repeater}$ ,  $s$  and  $l$  are fixed. For case I, we generate expressions of  $s_{opt}$ ,  $l_{opt}$  and  $\left(\frac{\tau}{l}\right)_{opt}$  in terms of supply voltage and temperature by SPICE simulation. For a given

Tech. node (nm)	130	90	65
W (nm)	335	230	145
T (nm)	670	483	319
$\epsilon_{ins}$	3.1	2.8	2.5
$V_{DD}$ (V)	1.1	1	0.65
$f_{clk_{nom}}$ (GHz)	1.68	3.99	6.74
$I_{off_{nom}}$ (A/m)	0.42	2.68	17.39
$I_{off_{pnom}}$ (A/m)	0.21	2.20	8.55
$r_{s_{nom}}$ (k $\Omega$ )	8.8	6.3	20.1
$c_{0_{nom}}$ (fF)	0.94	0.59	0.60
$c_{p_{nom}}$ (fF)	2.29	1.75	0.48
$P_{total}$ (W)	61	85	104

Table 2: Interconnect parameter and nominal supply voltage for different technology nodes based on ITRS.

Tech. node (nm)	Case I			Case II		
	$\left(\frac{\tau}{l}\right)_{opt} / \left(\frac{\tau}{l}\right)_{nom}$	$\frac{P_{opt}}{P_{nom}}$	$\frac{V_{DD_{opt}}}{V_{DD_{nom}}}$	$\left(\frac{\tau}{l}\right)_{opt} / \left(\frac{\tau}{l}\right)_{nom}$	$\frac{P_{opt}}{P_{nom}}$	$\frac{V_{DD_{opt}}}{V_{DD_{nom}}}$
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

Table 3: The ratio of total power consumption and delay per unit length with optimal supply voltage and with nominal supply voltage for various technology nodes.

supply voltage, we can find the chip temperature by solving the following equation

$$T = T_{nom} + \vartheta (P_{logic}(T) + P_{repeater}(T))$$

where  $P_{logic}(T)$  and  $P_{repeater}(T)$  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.

## 7 Results

Figure 6 shows the delay per unit length as a function of power supply voltage for the 130 nm, 90 nm and 65 nm technology nodes. Note that the optimal supply voltage for both Case I and Case II 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 1). 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 I and case II of 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  $V_{DD}$  and temperature, operating it at a lower supply voltage can improve performance. Note that the optimum values of  $\frac{\tau}{l}$  are very similar for case I and case II for every technology node.

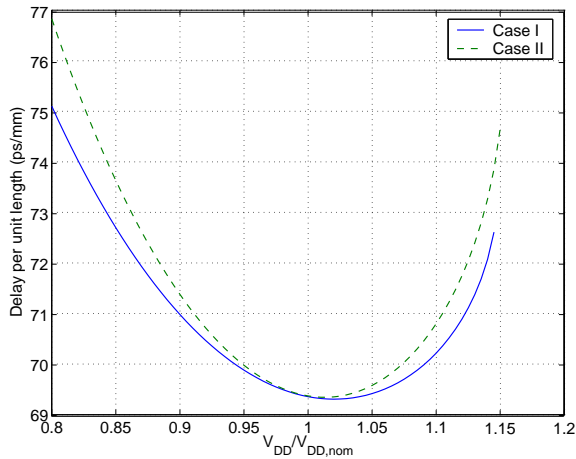
The interconnect parameters and the nominal supply voltage are based on ITRS [16], and are shown in Table 2.  $S_{int}$  is assumed to be equal to the minimum width of the global interconnect. The absolute value of  $V_{in}$  and  $V_{tp}$  are assumed to be the same and are equal to  $\frac{1}{4} V_{DD}$  at nominal supply voltage and temperature of 105°C. Table 3 shows  $\left(\frac{\tau}{l}\right)_{opt} / \left(\frac{\tau}{l}\right)_{nom}$ , i.e., the ratio of delay per unit length at optimum  $V_{DD}$  and the delay per unit length at the nominal  $V_{DD}$  and temperature, and  $\frac{P_{opt}}{P_{nom}}$ , i.e., the ratio of total power consumption at the optimum  $V_{DD}$  and the total power consumption with the nominal  $V_{DD}$  and temperature. Note that both performance and total power consumption improve at the optimal supply voltage for 90nm and 65nm technology.

## 8 Conclusions

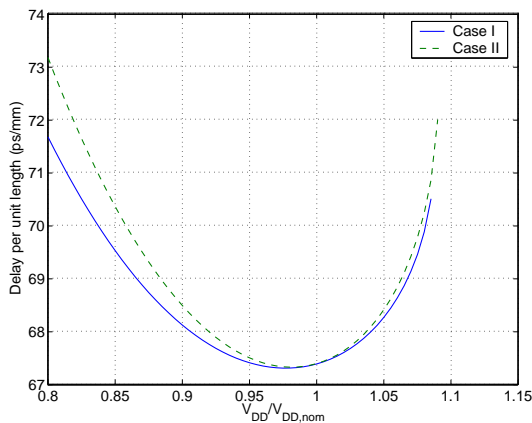
In conclusion, we have developed a methodology for calculating the optimal supply voltage which minimizes the delay per unit length while considering the total chip power dissipation and temperature rise in a consistent manner. The methodology is demonstrated for two cases (I) the design can be optimally buffered for the target  $V_{DD}$  and temperature and (II) when the design is buffered for a fixed  $V_{DD}$  and temperature. Using this methodology, we have computed the optimal operating voltage for 130 nm, 90 nm and 65 nm technology node for both cases. Furthermore, we have shown that as the technology is scaled beyond 130nm 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 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.

## References

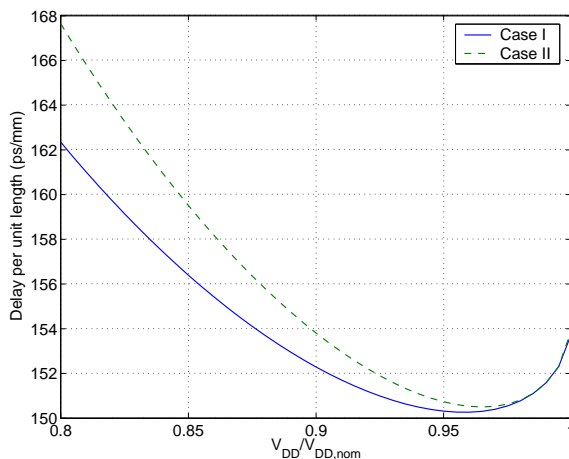
- [1] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, and S. Borkar, "Effectiveness and scaling trends of leakage control techniques for sub-100nm CMOS technologies," in *International Symposium on Low-power Electronics*, 2003.
- [2] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Transactions on Electron Devices*, vol. 49, no. 11, pp. 2001–2007, 2002.
- [3] J. M. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital Integrated Circuits : A Design Perspective*. Prentice Hall, 2003.
- [4] Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge University Press, 1998.
- [5] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.
- [6] R. W. Brodersen, M. A. Horowitz, D. Marković, B. Nikolić, and V. Stojanović, "Methods for true power minimization," in *Digest of Technical Papers, IEEE/ACM International Conference on Computer Aided Design*, pp. 35–42, 2002.
- [7] M. R. Stan, "Optimal voltages and sizing for low power," in *Proceedings. Twelfth International Conference On VLSI Design*, pp. 428–433, 1999.
- [8] A. Chandrakasan and R. W. Brodersen, *Low Power Digital CMOS Design*. Kluwer Academic Publishers, 1995.
- [9] N. Sirisantana, L. Wei, and K. Roy, "High-performance low-power CMOS circuits using multiple channel length and multiple oxide thickness," in *Proceedings. 2000 International Conference on Computer Design*, pp. 227–232, 2000.
- [10] A. P. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*. IEEE Press, 2000.
- [11] M. Johnson, D. Somasekhar, and K. Roy, "Models and algorithms on bounds of leakage in CMOS circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, no. 6, pp. 714–725, 1999.
- [12] 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 *International Symposium on Low Power Electronics and Design*, pp. 207–212, 2001.
- [13] 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 *Proceedings, IEEE International Electron Devices Meeting*, pp. 887–890, 2003.
- [14] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Addison-Wesley, 1990.
- [15] K. Banerjee and A. Mehrotra, "Analysis of on-chip inductance effects for distributed RLC interconnects," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, pp. 904–915, Aug. 2002.
- [16] "International Technology Roadmap for Semiconductors (ITRS)," 1999.
- [17] 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," *Proceedings of the IEEE*, vol. 89, pp. 602–633, May 2001.
- [18] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE Journal of Solid-State Circuits*, vol. 19, pp. 468–473, 1984.
- [19] S. M. Sze, *Physics of semiconductor devices*. Wiley, 1981.
- [20] R. S. Muller and J. I. Kamins, *Device Electronics for Integrated Circuits*. Wiley, 1984.



(a) Delay per unit length as a function of supply voltage for 130nm technology



(b) Delay per unit length as a function of supply voltage for 90nm technology



(c) Delay per unit length as a function of supply voltage for 65nm technology

Figure 6: Performance vs supply voltage for various technologies.