Power-optimal Repeater Insertion Considering Vdd and Vth as Design Freedoms [*]

Yu Ching Chang, King Ho Tam and Lei He
University of California,
Los Angeles, CA 90095, USA
{ychangu, ktam, lhe}@ee.ucla.edu

ABSTRACT

This work first presents an analytical repeater insertion method which optimizes power under delay constraint for a single net. This method finds the optimal repeater insertion lengths, repeater sizes, and V_{dd} and V_{th} levels for a net with a delay target, and it reduces more than 50% power over a previous work which does not consider V_{dd} and V_{th} optimization. This work further presents the power saving when multiple V_{dd} and V_{th} levels are used in repeater insertion at the full-chip level. Compared to the case with single V_{dd} and V_{th} suggested by ITRS, optimized dual V_{dd} and dual V_{th} reduce overall global interconnect power by 47%, 28% and 13% for 130nm, 90nm and 65nm technology nodes, respectively, but extra V_{dd} or V_{th} levels only give marginal improvement. We also show that an optimized single V_{th} reduce interconnect power almost as effective as dual-V_{th} does, in contrast to the need of dual V_{th} for logic circuits.

Categories and Subject Descriptors: B.7.2[Hardware]: Integrated circuits – Design aids

General Terms: Performance, Design

Keywords: Low power, buffer insertion

1. INTRODUCTION

Repeater insertion causes increasingly severe problem of power consumption due to the ever increasing number of repeaters [1]. Traditional approach of repeater insertion optimizes the interconnect in terms of delay, but several works in the literature [2, 3, 4] have made use of the extra tolerable delay (i.e., slack) in nets for significant saving in interconnect power. [2, 3] provide analytical methods to compute unit length power optimal repeater insertion solutions. [4] defines a new figure of merit which allows trade-off between power and delay using repeater insertion legnths, repeater sizes and wire widths as design knobs. None of the above work considers supply voltage V_{dd} and threshold voltage V_{th} as design freedoms. [5] performs dual V_{dd} and dual V_{th} assignments on logic circuits to reduce power consumption, and shows that 20% of power can be saved by going from single V_{th} to dual V_{th} under the dual V_{dd} power supply.

This paper studies the opportunity of power saving by computing power optimal repeater sizes, repeater insertion lengths, and V_{dd} and V_{th} levels for both individual nets and full chips. This paper is organized as follows. Section 2 discusses the delay and the power models. Section 3.1 presents single-net power optimization with V_{dd} and V_{th} tuning. Section 4 studies the full chip power optimization using multiple V_{dd} and V_{th}. We conclude this paper in Section 5.

2. PRELIMINARIES

This section discusses the delay and power models used in this paper. Both models are based on those in [2], which assume fixed V_{dd} and V_{th}. We extend the models to reflect the effects of V_{dd} and V_{th} scaling.

2.1 Delay Model

Consider an interconnect of unit length resistance r, unit length capacitance c, and total length L. Suppose the interconnect is divided into L/l segments and identical repeaters of unit driving resistance r_s, unit input capacitance c_o, unit output capacitance c_p and size s are inserted at the beginning of every segment. The delay of a segment consisting of a repeater driving an interconnect segment of length l terminated with a repeater of the same size is given by

$$\tau = r_s(c_o + c_p) + \frac{r_s}{s}cl + rlsc_o + \frac{1}{2}rcl^2 \qquad (1)$$

and the unit length delay is

$$\frac{\tau}{l} = \frac{1}{l}r_s(c_o + c_p) + \frac{r_s}{s}c + rsc_o + \frac{1}{2}rcl \qquad (2)$$

The total delay of the entire interconnect is $\frac{\tau}{l}L$, assuming continous numbers of buffers and segments. The driving resistance of the repeater depends on the operating V_{dd} and V_{th} levels and is approximated in [3] by

$$r_s = K_1\frac{V_{dd}}{I_{dsat}} \qquad (3)$$

where K_1 is a fitting parameter and I_{dsat} is the saturated drain current of a minimum-sized NMOS or PMOS transistor with both V_{gs} and V_{ds} equal to V_{dd}. According to the alpha-power law model [6], I_{dsat} is modeled as

$$\begin{aligned} I_{dsat} &= K_2(V_{gs} - V_{th})^{\alpha} \\ &= K_2(V_{dd} - V_{th})^{\alpha} \qquad (4) \end{aligned}$$

[*] This paper is partially supported by NSF CAREER award CCR-0306682/0401682, SRC grant 1100, a UC MICRO grant sponsored by Fujitsu Laboratories of America, Intel and Mindspeed, and a Faculty Partner Award by IBM. Address comments to lhe@ee.ucla.edu.

ISLPED'05, August 8–10, 2005, San Diego, California, USA
Copyright 2005 ACM 1-59593-137-6/05/0008 ...$5.00.

where K_2 is a device parameter and α is about 1.25 for recent technology generations. By plugging Equation (4) into Equation (3), we obtain r_s as a function of V_{dd} and V_{th}, which is given by

$$r_s = K_3 \frac{V_{dd}}{(V_{dd} - V_{th})^{\alpha}} \tag{5}$$

where $K_3 = K_1/K_2$. For a given V_{dd} and V_{th}, we obtain the optimal unit length delay by setting

$$l_{opt} = \sqrt{\frac{2r_s(c_o + c_p)}{rc}} \qquad s_{opt} = \sqrt{\frac{r_s c}{r c_o}} \tag{6}$$

and the optimum unit length delay is given by

$$(\frac{\tau}{l})_{opt} = 2\sqrt{r_s c_o r c}\left(1 + \sqrt{\frac{1}{2}\left(1 + \frac{c_p}{c_o}\right)}\right) \tag{7}$$

Suppose we are given a target delay per length, which is expressed as f% more than $(\frac{\tau}{l})_{opt}$, we can find a family of solutions $\{V_{dd},\ V_{th},\ l,\ s\}$ that satisfy the target delay. In the solution set, there exists a solution that achieves the minimum power. The methodology of finding such solution is presented in Section 3.1.

2.2 Power Model

For an interconnect of length L, the total power dissipated by the repeaters is $\frac{P_{tot}}{l}L$. The power consumption of a repeater comprises three parts: dynamic, leakage, and short circuit. We use the same formulae to compute power as in [2] except that V_{dd} and V_{th} are treated as variables in the expressions. The power models are summarized below.

Dynamic power is dissipated when repeaters charge and discharge their loading capacitances. It is given by

$$P_{switching} = a(s(c_o + c_p) + lc)V_{dd}^2 f_{clk}$$

where a is the switching activity of a repeater, which is assumed to be 0.15, and f_{clk} is the clock frequency.

We consider only the subthreshold leakage as in [2]. The subthreshold leakage current of a minimum-sized NMOS transistor is given by

$$I_{off} = I_{off}^{ref} \cdot 10^{\frac{(V_{th}^{ref} - V_{th})}{S_w}}$$

where I_{off}^{ref} and V_{th}^{ref} are the reference subthreshold leakage current and threshold voltage respectively for a particular technology node, and S_w is the subthreshold swing, which we assume 100mV/decade at the temperature 100^oC. The equation assumes that the transistor is at OFF state when $V_{gs} = 0$ and $V_{ds} = V_{dd}$.

The average leakage power of a repeater is

$$\begin{aligned} P_{leakage} &= V_{dd} I_{leakage} \\ &= \frac{1}{2} V_{dd}(I_{off}^n W_{min}^n + I_{off}^p W_{min}^p)s \end{aligned}$$

where I_{off}^n and I_{off}^p are the subthreshold leakage current for NMOS and PMOS transistors respectively, and W_{min}^n and W_{min}^p are the widths of the NMOS and PMOS transistors in a minimum-sized inverter.

The short circuit power dissipation depends on the transition time at the input and the output of an inverter. Assuming symmetric high-to-low and low-to-high transitions at the input and the output of the repeater, the short circuit power is given by

$$P_{short-circuit} = a t_r V_{dd} W_{min}^n s I_{short-circuit} f_{clk}$$

where a is the same switching factor as in the dynamic power expression, $I_{short-circuit}$ is approximately 65 $\mu A/\mu m$ and $t_r = \tau log_e 3$.

The power per length is therefore given by the sum of all $P_{dynamic}$, $P_{leakage}$ and $P_{short-circuit}$, i.e.,

$$\frac{P_{tot}}{l} = k_1 V_{dd}^2(\frac{s}{l}(c_p + c_o) + c) + k_2 V_{dd}\frac{s}{l} + k_3 V_{dd} s \frac{\tau}{l} \tag{8}$$

where

$$\begin{aligned} k_1 &= a f_{clk} \\ k_2 &= \frac{1}{2}(I_{off}^n W_{min}^n + I_{off}^p W_{min}^p) \\ k_3 &= a W_{min}^n f_{clk} log_e 3 \\ S' &= \frac{S_w}{log_e 10} \end{aligned}$$

We specify the target delay by using $(\frac{\tau}{l})_{opt}(1+f)$, as explained in Section 2.1. By setting the net delay $\tau = (1 + f)(\frac{\tau}{l})_{opt} l$, we can simplify expression (8) by replacing $k_3\frac{\tau}{l}$ with $k_3' = k_3(1+f)(\frac{\tau}{l})_{opt}$.

3. SINGLE NET POWER OPTIMIZATION

3.1 Analytical Solution

Based on the delay and power models discussed previously, we express the problem formulation as

$$\begin{aligned} &min && \left(\frac{P_{tot}}{l}\right)(V_{dd}, V_{th}, l, s) \\ &subject\ to && \left(\frac{\tau}{l}\right)(V_{dd}, V_{th}, l, s) = (1+f)(\frac{\tau}{l})_{opt} \end{aligned} \tag{9}$$

For given V_{dd}, V_{th} and a delay target, the optimal l and s that give the minimum $\frac{P_{tot}}{l}$ can be obtained by solving the following set of nonlinear equations in [2], i.e.,

$$\begin{aligned} \frac{\partial \frac{P_{tot}(l,s)}{l}}{\partial s} &= 0 \\ \left(\frac{\tau}{l}\right)(l,s) - (1+f)(\frac{\tau}{l})_{opt} &= 0 \end{aligned} \tag{10}$$

The insertion length l is a function of the repeater size s under the equality delay constraint in Equation (10). In this problem, both the objective function and the constraint are posynomial functions which are known to be convex under variable transformation. Therefore, there exists a unique minima for such optimization problem, which can be found in polynomial time [7].

When V_{dd} and V_{th} are treated as variables, it is not obvious if the problem is still convex. To visualize this, we can find the power-optimal solution for every point on the V_{dd}-V_{th} space using Equation (10), which solves for power-optimal repeater insertion under fixed V_{dd} and V_{th}. Figure 1 shows the resulting iso-power plot under a delay target of $(1+5\%)(\frac{\tau}{l})_{opt}$. Each contour line represents the continuous combinations of V_{dd} and V_{th} that achieve the same value of $\frac{P_{tot}}{l}$. The optimal value, which is a single point degenerated from a contour, is marked as $(V_{dd}^{opt},\ V_{th}^{opt})$ in Figure 1. This plot shows that there exists a single optimum in the possible range of V_{dd} and V_{th}, which hints that the problem of

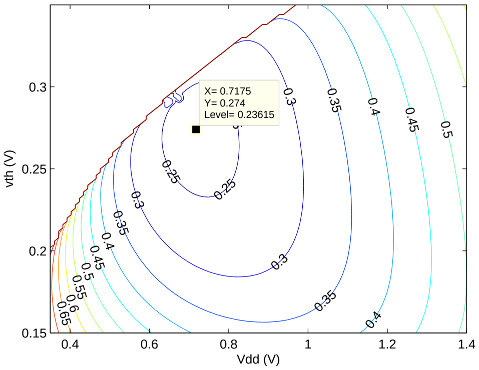

Figure 1: Contour plot of unit length power.

power minimization through V_{dd} and V_{th} optimization can be solved analytically. Our future research will attempt to prove that this problem possesses a unique optimum analytically.

Based on the observation that an optimal point exists, we develop an analytical method to solve this problem. Following the equality delay constraint, one of the variable must be a function of the other three variables. In our derivation, V_{th} is chosen to be the dependent variable, because it is the only variable that can be easily expressed in the closed-form of the other three variables. From Equation (5), V_{th} can be expressed in terms of V_{dd} and r_s as

$$V_{th} = V_{dd} - \left(\frac{K_3 V_{dd}}{r_s}\right)^{\frac{1}{\alpha}}$$

By re-arranging Equation (2), r_s can be expressed as a function of l and s:

$$r_s = \frac{(1+f)(\frac{\tau}{l})_{opt} - rsc_o - \frac{1}{2}rcl}{\frac{c_o+c_p}{l} + \frac{c}{s}}$$

Therefore, when deriving the gradients of the objective function, V_{th} is treated as a function of V_{dd}, l and s. The following equations set the gradients of the objective function with respect to V_{dd}, s and l to zero.

$$\begin{aligned}\frac{\partial \frac{P_{tot}}{l}}{\partial V_{dd}} &= 2k_1 V_{dd}\left(\frac{s}{l}(c_o+c_p)+c\right) + k_2 e^{\frac{-V_{th}(V_{dd},l,s)}{S'_w}}\frac{s}{l} \\ &- \frac{1}{S'}\frac{\partial V_{th}}{\partial V_{dd}} k_2 V_{dd} e^{\frac{-V_{th}(V_{dd},l,s)}{S'_w}}\frac{s}{l} + k'_3 s = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial \frac{P_{tot}}{l}}{\partial s} &= k_1 V_{dd}^2 \frac{c_o+c_p}{l} + k_2 V_{dd} e^{\frac{-V_{th}(V_{dd},l,s)}{S'_w}}\frac{1}{l} \\ &- \frac{1}{S'}\frac{\partial V_{th}}{\partial s} k_2 V_{dd} e^{\frac{-V_{th}(V_{dd},l,s)}{S'_w}}\frac{s}{l} + k'_3 V_{dd} = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial \frac{P_{tot}}{l}}{\partial l} &= -k_1 V_{dd}^2 (c_o+c_p)\frac{s}{l^2} - k_2 V_{dd} e^{\frac{-V_{th}(V_{dd},l,s)}{S'_w}}\frac{s}{l^2} \\ &- \frac{1}{S'}\frac{\partial V_{th}}{\partial l} k_2 V_{dd} e^{\frac{-V_{th}(V_{dd},l,s)}{S'_w}}\frac{s}{l} = 0 \qquad (11)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial V_{th}}{\partial V_{dd}} &= 1 - \frac{1}{\alpha}(\frac{K_3}{r_s})^{\frac{1}{\alpha}} V_{dd}^{\frac{1}{\alpha}-1} \\ \frac{\partial V_{th}}{\partial s} &= \frac{1}{\alpha}(K_3 V_{dd})^{\frac{1}{\alpha}} r_s^{-\frac{1}{\alpha}-1}\frac{\partial r_s}{\partial s} \\ \frac{\partial V_{th}}{\partial l} &= \frac{1}{\alpha}(K_3 V_{dd})^{\frac{1}{\alpha}} r_s^{-\frac{1}{\alpha}-1}\frac{\partial r_s}{\partial l}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial r_s}{\partial s} &= \left(\frac{c_o+c_p}{l} + \frac{c}{s}\right)^{-2}\left\{\begin{array}{l}\frac{c}{s^2}\left((1+f)(\frac{\tau}{l})_{opt} - \frac{1}{2}rcl\right) \\ -\frac{rcc_o}{s} - rc\left(\frac{c_o+c_p}{l} + \frac{c}{s}\right)\end{array}\right\} \\ \frac{\partial r_s}{\partial l} &= -\frac{1}{2}rc\left(\frac{c_o+c_p}{l} + \frac{c}{s}\right)^{-1} \\ &+ \left\{\frac{c_o+c_p}{l^2}\left((1+f)(\frac{\tau}{l})_{opt} - rsc_o - \frac{1}{2}rcl\right)\right\} \\ &\cdot\left(\frac{c_o+c_p}{l} + \frac{c}{s}\right)^{-2}\end{aligned}$$

These equations can be solved numerically using an iterative numerical solver. The optimal solution from the analytical method is verified by exhaustive search and they match each other closely.

3.2 Experimental Results

Equation (11) is used to optimize unit length power for a single net. The parameters for the power and delay models across various technology nodes are taken from [1]. Table 1 compares the results with and without V_{dd} and V_{th} tuning across different technology for target delay $\tau = (1+f)(\frac{\tau}{l})_{opt}$ where f is between 5% and 100%. The results from optimization under fixed V_{dd} and V_{th} are called the reference values in this paper. The reference supply voltage V_{dd}^{ref} used for each technology are obtained from [1] and V_{th}^{ref} values are assumed to be 25% of their respective V_{dd}^{ref} as in [2].

As shown in Table 1, the amount of power saving that can be achieved from V_{dd} and V_{th} optimization depends on the target delay. When $f = 20\%$, the power saving is up to 28% across all technology nodes. When $f = 100\%$, the power saving is more than 50% for all generations. The power saving is mainly achieved by lowering the supply voltage. As we can see, the optimal V_{dd} levels are generally lower than the reference values. When f increases, V_{dd} decreases significantly, showing that V_{dd} provides good trade-off for power by utilizing f. The optimal V_{th} values slowly decreases with increasing f to compensate for the loss of performance from V_{dd} reduction. The reduction in V_{th} causes a moderate increase in leakage power, but is rewarded by a large decrease in the dynamic power from lowering V_{dd}. The performance loss due to V_{dd} reduction is compensated by the increase of repeater size s and the slight decrease of insertion length l when compared to the reference values.

4. FULL-CHIP INTERCONNECT POWER

4.1 Power Calculation

In this section, we propose a methodology to evaluate full-chip interconnect power. In [8], a closed-form analytical expression of the wire-length distribution for on-chip random logic networks based on Rent's rule is developed. We estimate the full-chip power by integrating the unit length power over the wire-length distribution from the smallest

node	f	V_{dd}	$\frac{V_{dd}}{V_{dd}^{ref}}$	V_{th}	$\frac{V_{th}}{V_{th}^{ref}}$	s	$\frac{s}{s_{ref}}$	l	$\frac{l}{l_{ref}}$	$(\frac{P}{l})_{opt}$	$(\frac{P}{l})_{opt}$
	(nm)	(V)		(V)		(× min)		(mm)		(W/m)	saving
130	5%	1.06	0.92	0.27	0.95	59.5	1.12	1.65	0.97	0.16	3 %
	10%	0.97	0.82	0.27	0.95	59.7	1.31	1.74	0.93	0.13	10 %
	20%	0.84	0.70	0.26	0.95	59.1	1.61	1.92	0.92	0.10	25 %
	100%	0.51	0.41	0.24	0.85	42.1	2.60	3.13	1.01	0.04	62 %
90	5%	0.93	0.87	0.23	0.88	57.5	1.12	1.34	0.97	0.25	6 %
	10%	0.85	0.78	0.23	0.88	57.6	1.31	1.41	0.94	0.21	14 %
	20%	0.73	0.66	0.22	0.88	57.0	1.60	1.56	0.92	0.16	28 %
	100%	0.43	0.38	0.19	0.75	40.2	2.54	2.59	1.06	0.06	65 %
65	5%	0.75	1.02	0.20	1.12	39.4	1.08	0.87	0.96	0.23	2 %
	10%	0.69	0.92	0.20	1.11	39.4	1.25	0.92	0.92	0.20	7 %
	20%	0.60	0.79	0.20	1.10	39.0	1.51	1.03	0.89	0.16	18 %
	100%	0.36	0.45	0.16	0.91	27.9	2.35	1.77	0.99	0.07	54 %

Table 1: Comparision of unit length power with and without Vdd and Vth tuning

wire length with non-negligible power to the longest global interconnect assumed by the wire-length distribution model. We use the delay optimal segment length l_{opt} given by Equation (6) to define the shortest interconnect which requires at least one repeater to be inserted. Nets shorter than l_{opt} are not considered as they do not need repeaters. The delay of each net is bounded by 90% of the clock period T_{clk} as in [9]. For an interconnect of length L operating at V_{dd} and V_{th}, the optimal delay is

$$D_{opt} = \left(\frac{\tau}{l}\right)_{opt} (V_{dd}, V_{th}) L$$

where $(\frac{\tau}{l})_{opt}(V_{dd}, V_{th})$ is given by Equations (5) and (7). The difference between D_{opt} and $0.9 \cdot T_{clk}$ is the slack that we can use to optimize its power. We define L_{max} to be the longest interconnect length which satisfies the target delay with delay optimal repeater insertion, i.e.,

$$L_{max} = \frac{0.9 \cdot T_{clk}}{\left(\frac{\tau}{l}\right)_{opt}}$$

We pipeline the interconnects of lengths larger than L_{max} so that the length of each segment is smaller than L_{max}. We assume that the delay overhead of pipelining flip-flops is amortized in $0.1 \cdot T_{clk}$. Therefore, the power for the full-chip is given by

$$P = \int_{\nu_{opt}}^{2\sqrt{N}} \mathbf{R}(\nu) \left(\frac{P}{l}\right)_{opt} (f) l_\beta \beta \, d\nu \qquad (12)$$

where

ν	wire length in terms of gate pitches;
ν_{opt}	l_{opt} in terms of gate pitches;
N	number of logic gates;
β	number of pipelining stages;
l_β	wire length per stage;
$\mathbf{R}(\nu)$	wirelength distribution function;
$\left(\frac{P}{l}\right)_{opt}(f)$	power per length function defined in the Problem Formulation (9);
f	slack in terms of multiple of $\left(\frac{\tau}{l}\right)$;

The length in terms of gate pitches is obtained by

$$\nu = \frac{l}{\sqrt{AF}\mathbf{T}} \qquad (13)$$

where AF is the gate area factor, which is 320 across all technology nodes [1] and $\mathbf{T}$ is the technology node in terms of minimum local metal's half-pitch dimension. The number of pipelining stages β and the wire length per stage l_β are given by

$$\beta = \lceil \frac{\nu\sqrt{AF}\mathbf{T}}{L_{max}} \rceil,$$
$$l_\beta = \frac{\nu\sqrt{AF}\mathbf{T}}{\beta}$$

The optimal power per length $(\frac{P}{l})_{opt}$ is a function of the target delay, and is obtained using Equation (10) discussed in when V_{dd} and V_{th} are fixed and Equation (11) when V_{dd} and V_{th} are design variables, both discussed in Section 3.1. Target delay of an interconnect of length l_β is again specified by $\tau = (1+f)\left(\frac{\tau}{l}\right)_{opt}(V_{dd}, V_{th}) l_\beta$. Therefore f can be computed from l_β by

$$f = \frac{0.9 \cdot T_{clk}}{(\frac{\tau}{l})_{opt} \cdot l_\beta} - 1$$

Technology Node (nm)	130	90	65	45
# transistors (M)	97	193	276	1546
T_{clk} (ps)	594	251	148	86.9
V_{dd} (V)	1.1	1	0.7	0.6
V_{th} (V)	0.28	0.25	0.17	0.15
L_{max} (mm)	6.94	2.30	1.06	0.513
l_{opt} (mm)	1.32	1.06	0.67	0.540

Table 2: List of parameters based on 2001 ITRS.
Note: The number of gates N is assumed to be # transistors/4

4.2 Vdd and Vth Optimization

To optimize the full-chip interconnect power, we consider various cases of V_{dd} and V_{th} assignment for nets. Practical assignment has limited number of V_{dd} and V_{th} levels throughout the chip. Multiple V_{dd} levels are provided either by having multiple power distribution networks or by inserting pass transistors to create lower V_{dd} supplies than the system V_{dd}. Multiple V_{th} can be achieved either through selective transistor doping or through substrate biasing. The V_{dd} and V_{th} pair for a net can be formed from any one of the available V_{dd} and V_{th} levels. Therefore, increasing V_{dd} and V_{th} levels improves the power saving it can achieve due to more fine-grained control to V_{dd} and V_{th} for each net. We are interested in maximizing the power saving that can be achieved by the minimum number of V_{dd} and V_{th} levels available at the full-chip level, since extra V_{dd} and V_{th} levels increase area and manufacturing costs. We compare the optimal full-chip global interconnect power of each combination (N_{dd}, N_{th}), where N_{dd} is the number of V_{dd} levels and

N_{th} is the number of V_{th} levels. The theoretical optimum power occurs at $N_{dd} \rightarrow \infty$ and $N_{th} \rightarrow \infty$, i.e., the V_{dd} and V_{th} of each net can be taylored. Such comparison provides us with an idea of the potential power saving by increasing N_{dd} and N_{th}.

Table 3 shows our searching algorithm for the power optimal V_{dd} and V_{th} levels at the full-chip level. Given N_{dd} and N_{th}, the algorithm first generates all possible combinations of V_{dd} and V_{th} for the full-chip at line 3. For a particular N_{dd} levels of V_{dd} and N_{th} levels of V_{th}, any combination of (V_{dd}, V_{th}) that has lower delay per length than the reference combination $(V_{dd}^{ref}, V_{th}^{ref})$, which provides the best delay performance, is discarded. Combinations which cannot even achieve the delay bound at the shortest wire length $l_{opt}(V_{dd}^{ref}, V_{th}^{ref})$ in our defined global interconnect are also discarded. These are implemented in line 5. The algorithm then evaluates $L_{max}(V_{dd}, V_{th})$, which is the maximum wire length that satisfies the $0.9 \cdot T_{clk}$ delay bound, for every (V_{dd}, V_{th}) combination. The combinations are then sorted as in line 6, after which nets of different lengths are assigned with V_{dd} and V_{th} according to the sorted order, as illustrated in Figure 2. Finally, the power of each of these regions with different (V_{dd}, V_{th}) assignments are computed in lines 9–14. Note that wires of length larger than $L_{max}(V_{dd}^{ref}, V_{th}^{ref})$ have to be broken down into segments by means of pipelining as discussed, which is implemented by looping on the number of pipeline stages at line 10 and by folding the integration bounds in lines 11–12. ν is simply the length in terms of gate pitches, and the conversion between ν and length in absolute dimensions are done using Equation (13). Also note that the optimal power per length function $\left(\frac{P}{l}\right)_{opt}(f, V_{dd}, V_{th})$ in line 13 refers to the power optimal repeater insertion with fixed V_{dd} and V_{th} using Equation (10).

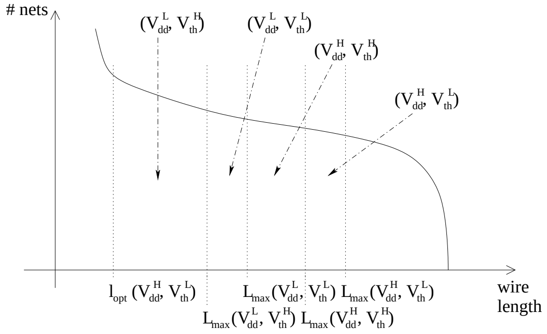
# nets
(V_{dd}^{L}, V_{th}^{H})
(V_{dd}^{L}, V_{th}^{L})
(V_{dd}^{H}, V_{th}^{H})
(V_{dd}^{H}, V_{th}^{L})
$l_{opt}(V_{dd}^{H}, V_{th}^{L})$
$L_{max}(V_{dd}^{L}, V_{th}^{L})$
$L_{max}(V_{dd}^{H}, V_{th}^{L})$
wire length
$L_{max}(V_{dd}^{L}, V_{th}^{H})$
$L_{max}(V_{dd}^{H}, V_{th}^{H})$

Figure 2: (V_{dd}, V_{th}) assignment in a net distribution

The ideal case in which $N_{dd} \rightarrow \infty$ and $N_{th} \rightarrow \infty$ can be computed by the same algorithm with some modification. Even though some smart pruning has been done to the search space as shown in Table 3, the algorithm fundamentally performs exhaustive search, in which the number of combinations for (V_{dd}, V_{th}) grows exponentially as N_{dd} and N_{th} increase. We have found that N_{dd} and N_{th} beyond 3 is impractical from the runtime perspective. Therefore, instead of using large N_{dd} and N_{th}, the power per length function is changed to our analytical repeater insertion solution considering both V_{dd} and V_{th} optimization in Equation (11), and set $N_{dd} = N_{th} = 1$. This is equivalent to finding the optimum repeater insertion with numerically computed optimum V_{dd} and V_{th} for each net.

```
Algorithm:  $ComputeOptPower(N_{dd}, N_{th})$
1.  $S(V_{dd})$ = the set of $V_{dd}$ levels to search
2.  $S(V_{th})$ = the set of $V_{th}$ levels to search
3.  $S(\{V_{dd}\}, \{V_{th}\})$ = ${}_{|S(V_{dd})|}C_{N_{dd}} \times {}_{|S(V_{th})|}C_{N_{th}}$
4.  for each $\{V_{dd}\}, \{V_{th}\} \in S(\{V_{dd}\}, \{V_{th}\})$
5.    remove combinations $(V_{dd}, V_{th}) \in \{V_{dd}\} \times \{V_{th}\}$
        s.t.  $L_{max}(V_{dd}, V_{th}) < l_{opt}(V_{dd}^{ref}, V_{th}^{ref})$ or
              $\left(\frac{\tau}{l}\right)_{opt}(V_{dd}, V_{th}) > \left(\frac{\tau}{l}\right)_{opt}(V_{dd}^{ref}, V_{th}^{ref})$
6.    $\mathcal{S}$ = sorted $(V_{dd}, V_{th})$ combinations in the
            ascending order of $L_{max}(V_{dd}, V_{th})$
7.    P = 0
8.    LB = $\nu_d^{opt}$
9.    for each $\{V_{dd}, V_{th}\} \in \mathcal{S}$
10.     for p = 0 to $\beta - 1$
11.       $\top = \min(2\sqrt{N}, (p+1)\nu_{max}(V_{dd}, V_{th}))$
12.       $\perp = \max((p+1)\text{LB}, (p+1)\nu_{max}(V_{dd}, V_{th}))$
13.       P += $\int_{\perp}^{\top} \mathbf{R}(\nu) \left(\frac{P}{l}\right)_{opt}(f, V_{dd}, V_{th}) l_{\beta} \beta \, d\nu$
14.     LB = $\nu_{max}(V_{dd}, V_{th})$
15.  mark the set $\{V_{dd}\}, \{V_{th}\}$ as optimal
      if P is the minimum power found
```

Table 3: Optimal V_{dd} and V_{th} levels search

4.3 Experimental Results

The methodology discussed above is used to optimize the full-chip power of chip sizes reported in [1] for various technology generations. N_{dd} and N_{th} are enumerated only up to three for the sake of runtime. V_{dd} and V_{th} search range are minimized without compromising the power optimality. Figure 3 shows the full-chip power of various V_{dd} and V_{th} configurations, where each pair on the x-axis is (N_{dd}, N_{th}). The highest performance (the most power consuming) combination $(V_{dd}^{ref}, V_{th}^{ref})$ is always retained in all configurations by default, therefore the configuration (1, 1) refers to the optimal full-chip power with fixed reference V_{dd} and V_{th} for all nets. The “ideal” combination refers to the continuous V_{dd} and V_{th} assignment, i.e., $N_{dd}, N_{th} \rightarrow \infty$. Power reduces by 47%, 28% and 13% for 130nm, 90nm and 65nm technology nodes respectively by going from the single V_{dd}, single V_{th} configuration to the dual V_{dd}, dual V_{th} configuration. Using dual V_{th} instead of single V_{th} under dual V_{dd} only gives ~3% power reduction, as opposed to the 20% plus reduction reported for logic circuits in [5]. This suggests that optimizing the single reference V_{th} may just perform as well as the dual V_{th} configuration in terms of interconnect power consumption. The dual V_{dd} and dual V_{th} configuration has the total power just 17%, 12% and 5% from the theoretical power optimum configuration which allows infinite V_{dd} and V_{th} levels. Moreover, we observe no significant power reduction by moving to combinations with more V_{dd} and V_{th} levels in all technology generations.

The power breakdown of the optimized full-chip interconnect for each (N_{dd}, N_{th}) configuration is shown in each bar in Figure 3. Multiple V_{dd} configurations (i.e., $N_{dd} > 1$) in 130nm and 90nm technology nodes achieve significant dynamic power saving by aggressively reducing the second V_{dd} level, as shown in Table 4. The threshold voltage of the second V_{th} level slightly decreases to compensate for the loss of performance due to V_{dd} reduction, at the expense of slight increase in the leakage power. On the other hand, the leakage power in 65nm technology node is comparatively a lot larger in the $(1, 1)$ configuration. From Table 4, the second $V_{th} = 0.2V$ leaps above the reference level of $0.175V$ to limit the growth of leakage power. This can be seen in Figure 3, where the block of leakage for the 65nm bars slightly reduces from the single V_{dd}, single V_{th} combination to the

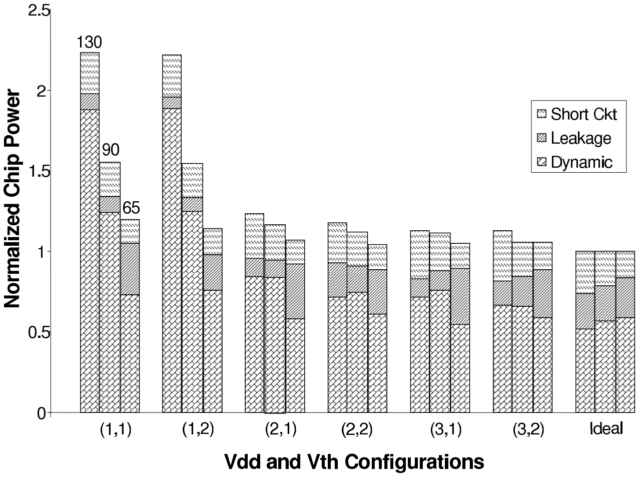

Figure 3: Power of optimized nets under different N_{dd} and N_{th}. Each group of bars contain results for 130nm, 90nm and 65nm technology nodes.

other multi-V_{dd}/V_{th} configurations. From this, we see that in order to get the right balance between dynamic power and leakage power for total power reduction in interconnect, we must consider both V_{dd} and V_{th} optimization.

Tech Node (nm)	(N_{dd}, N_{th})	V_{dd}s (V)	V_{th}s (V)
130	(2, 1)	1.1, 0.572	0.275
	(2, 2)	1.1, 0.506	0.226, 0.275
90	(2, 1)	1, 0.64	0.25
	(2, 2)	1, 0.64	0.2, 0.25
65	(2, 1)	0.7, 0.532	0.175
	(2, 2)	0.7, 0.532	0.175, 0.2

Table 4: V_{dd} and V_{th} levels for each (N_{dd}, N_{th})

Figure 4 shows the breakdown of total wire length being assigned to (V_{dd}, V_{th}) marked on each region of the figure for the dual V_{dd}, dual V_{th} case. The regions are ordered in the increasing power (the decreasing delay) (V_{dd}, V_{th}) combinations from the bottom to the top. A large portion of the net is assigned to the combination which has V_{th}/V_{dd} ratio way above the default 0.25, particularly for 65 nm technology. This implies that the V_{th}/V_{dd} ratio has to be increased in order to attain power optimality. This is in line with the conclusion made by other works in the literature [10], which suggests that the V_{th}/V_{dd} ratio shall be made larger than that current designs use for power efficiency.

5. CONCLUSIONS

This paper studies the opportunity of power saving by computing power optimal repeater sizes, repeater insertion lengths, and for the first time V_{dd} and V_{th} levels for both single nets and a full chip. We have derived a set of analytical formulae which finds the optimal interconnect power given the amount of the timing slack on a single net. Compared to [2] which does not consider V_{dd} and V_{th} as design variables, our method that customizes V_{dd} and V_{th} for each net can reduce power by more than 50% for both single nets and at the chip level. We have also studied the power saving of using multiple V_{dd} and V_{th} levels for buffering interconnects. Power reduces by 47%, 28% and 13% for 130nm, 90nm and 65nm technology nodes respectively by going from the single V_{dd}, single V_{th} configuration to the dual V_{dd}, dual V_{th} configuration. The fact that majority of the nets favors a V_{dd} to V_{th} ratio of more than 0.35 across all generations suggests that the ratio of 0.25 as suggested by other works in the literature is too low for power optimality. We show that the dual V_{dd} and dual V_{th} configuration is within 17%, 12% and 5% of the theoretical optimal power computed from our analytical method for 130nm, 90nm and 65nm technology node; and that extra V_{dd} or V_{th} level beyond dual V_{dd} and dual V_{th} only gives marginal improvement. Our experiment also shows that multiple V_{th} does not improve power of interconnect as much as that of logic circuits.

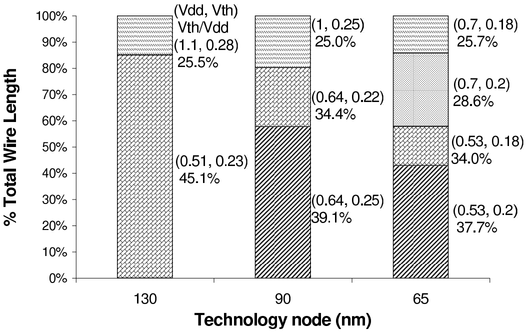

Figure 4: Net length distribution for dual Vdd, dual Vth configuration

6. REFERENCES

[1] Semiconductor Industry Association, *http://public.itrs.net*, 2001.

[2] K. Banerjee and A. Mehrotra, "A power-optimal repeater unsertion methodology for global interconnects in nanometer designs," *IEEE Trans. on Electron Devices*, vol. 49, pp. 2001–2007, November 2002.

[3] G. Chen and E. Friedman, "Low power repeaters driving RC interconnects with delay and bandwidth constraints," in *IEEE International ASIC/SOC Conference*, pp. 335–339, August 2004.

[4] M. Mui, K. Banerjee, and A. Mehrotra, "A global interconnect optimization scheme for nanometer scale vlsi with implications for latency, bandwidth, and power dissipation," *IEEE Trans. on Electron Devices*, vol. 51, pp. 195–203, February 2004.

[5] A. Srivastava and D. Sylvester, "Minimizing total power by simultaneous vdd/vth assignment," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 23, pp. 665–677, May 2004.

[6] T. Sakurai and A. Netwon, "Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas," *IEEE Trans. on Electron Devices*, vol. 25, no. 2, pp. 584–594, 1990.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[8] J. Davis, V. De, and J. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)–part I: Derivation and validation," *IEEE Trans. on Electron Devices*, vol. 45, pp. 580–589, Mar. 1998.

[9] J. Davis, V. De, and J. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)–part II: Application to clock frequency, power dissipation, and chip size estimation," *IEEE Trans. on Electron Devices*, vol. 45, pp. 590–597, Mar. 1998.

[10] M. Hamada and et al, "A top-down low power design technique using clustered voltage scaling with variable supply-voltage scheme," in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 495–498, 1998.