Stack Sizing for Optimal Current Drivability in Subthreshold Circuits

John Keane, Hanyong Eom, Tae-Hyoung Kim,
Sachin Sapatnekar, and Chris Kim

***Abstract*—Subthreshold circuit designs have been demonstrated to be a successful alternative when ultra-low power consumption is paramount. However, the characteristics of MOS transistors in the subthreshold region are significantly different from those in strong-inversion. This presents new challenges in design optimization—particularly in complex gates with stacks of transistors. In this paper, we present a framework for choosing the optimal transistor stack sizing factors in terms of current drivability for subthreshold designs. We derive a closed-form solution for the correct sizing of transistors in a stack, both in relation to other transistors in the stack, and to a single device with equivalent current drivability. Simulation results show that our framework provides a performance benefit ranging up to more than 10% in certain critical paths.**

***Index Terms*—Subthreshold logic, logical effort, ultra low power design**

I. Introduction

Due to the robust nature of static CMOS logic, circuits in this technology family can operate with supply voltages below the transistor threshold voltage (V_{th}), while consuming orders of magnitude less power than in the normal strong-inversion region. The operating frequency of subthreshold logic is much lower than that of regular strong-inversion circuits ($V_{dd} > V_{th}$) due to the small transistor current, which consists entirely of leakage current. The low operating frequency and low supply voltage combine to reduce both dynamic and leakage power, leading to the significant power savings seen in subthreshold designs.

Subthreshold logic holds promise for the growing number of applications in which minimal power consumption is the primary design constraint. Such circuits have received much attention in recent research, and a number of successful designs have been demonstrated. A multiplexer-based SRAM was proposed for subthreshold operation by the authors of [1]. They also introduced new tiny-XOR circuits and demonstrated their performance in a Fast Fourier Transform processor running at a supply voltage of 180mV. The authors of [2] presented a new high-density SRAM system operating down to 200mV at ISSCC 2007. In [3], Kim et al. built an ultra low power adaptive filter for hearing aid applications using subthreshold logic. Subthreshold-friendly logic styles and massively parallel DSP architectures were used in that work to achieve low voltage operation

The characteristics of MOS transistors in the subthreshold region are significantly different from those in the strong-inversion region. The saturation current, which was a near-linear function of the gate and threshold voltages in the strong-inversion region, becomes an exponential function of those values in the subthreshold regime [4]. In this work, we show that the sizing methods used to obtain maximum performance must be reformulated for use in subthreshold designs due to these different characteristics. In particular, we present a framework for choosing the optimal transistor stack sizing factors in terms of current drivability for subthreshold circuits. A closed-form solution for the optimal sizing of stacked transistors is derived and shown to match simulation results. Our theoretical sizing values closely match those found in simulations with Predictive Technology Model (PTM) [5,6] devices ranging from 130nm technology down to the 45nm node. This sizing method is shown to provide a clear benefit in logic paths containing a large number of stacks where the nodal capacitance is not dominated by the increased device sizes used in our method.

Maunuscript received October 22, 2006; revised March 16, 2007.

The authors are with the Department of Electrical and Computer Engineering at the University of Minnesota, Minneapolis, MN 55455 (email: {jkeane, eomxx001, thkim, sachin, chriskim}@ece.umn.edu)

Digital Object Identifier …….

II. Optimal 2-Stack Sizing

A. Optimal Ratio between 2 Stacked Devices

The first step we take in developing the subthreshold stack sizing framework is finding the optimal width ratio between transistors in a stack for maximum drive current. Here we will present a closed-form expression for the relative sizing of two transistors in a stack, showing that it is beneficial to size up the transistor nearest to the supply rail (V_{dd} for PMOS, ground for NMOS). The starting point is the following pair of current equations for upper and lower transistors as situated in an NMOS stack (so the lower device is connected to ground), excluding the common factors that will cancel out when they are equated:

$$I_U = W_U e^{\frac{(V_{dd}-V_X)-(V_{t0}+\gamma V_X+\lambda_d(V_{dd}-V_X))}{mV_T}}\left(1-e^{\frac{-(V_{dd}-V_X)}{V_T}}\right) \quad (1)$$

$$\approx W_U e^{\frac{(V_{dd}-V_X)-(V_{t0}+\gamma V_X+\lambda_d(V_{dd}-V_X))}{mV_T}}$$

$$I_L = W_L e^{\frac{V_{dd}-(V_{t0}+\lambda_d V_X)}{mV_T}}\left(1-e^{\frac{-V_X}{V_T}}\right) \quad (2)$$

Here, W_U and W_L denote the upper and lower transistor widths, respectively, and V_X denotes the voltage at the node between those devices. The Drain-Induced Barrier Lowering (DIBL) coefficient (a negative number) is represented by λ_d, and γ is the body effect coefficient. The thermal voltage is represented by V_T, while V_{t0} stands for the nominal threshold voltage. According to simulation results, $V_X \approx 10\%$ of V_{dd}. Each V_x term multiplied by the small DIBL coefficient (ranging from roughly -0.01 to -0.2 in current bulk technologies) can then be approximated as ~0. Moreover, note that $e^{-(V_{dd}-V_X)/V_T} \approx 0$. We use the symbol

$$\alpha = e^{\frac{-\lambda_d V_{dd}}{mV_T}}, \quad (3)$$

as well as the fact that $m = 1+\gamma$, to further simplify calculations. Rewriting the two current equations and equating them yields the following relationship:

$$\alpha W_U e^{\frac{-V_X}{V_T}} = W_L\left(1-e^{\frac{-V_X}{V_T}}\right) \quad (4)$$

Solving for V_X and using the definition $V_T = kT/q$ gives us

$$V_X = \frac{kT}{q}\ln\left(1+\frac{\alpha W_U}{W_L}\right) \quad (5)$$

We then define $W_T = W_U + W_L$ to eliminate W_L, which results in the following current equation:

$$I_U = I_L = \frac{\alpha W_U (W_T - W_U)}{\alpha W_U + W_T - W_U} e^{\frac{V_{dd}-V_{t0}}{mV_T}} \quad (6)$$

We find the optimal size for W_U by setting $(\partial I_U / \partial W_U)$ equal to zero. Again using our definition of W_T, we then find the optimal size for W_L. This derivation results in the following equations:

$$W_U = \frac{W_T}{1+\sqrt{\alpha}} \quad (7)$$

$$W_L = \frac{W_T}{1+\sqrt{\alpha}}\sqrt{\alpha} \quad (8)$$

According to these results, we expect to drive a higher current through the two-transistor stack when the lower device is larger than the upper transistor by a factor of $\sqrt{\alpha}$. For example, with an NMOS stack in 90nm PTM technology, when using a W_U of 1µm, the optimal W_L would be 1.23µm at V_{dd} = 0.2V, and 1.30µm at V_{dd} = 0.3V. As shown in equation (3), α is a function of V_{dd}, resulting in the different optimal width ratios for different V_{dd} values.

HSPICE simulations using 45nm through 130nm PTM technology files closely match the results of our derivation, and verify that the benefit of using the $\sqrt{\alpha}$ sizing ratio is more pronounced for larger α values (i.e., when the supply voltage is larger). PMOS transistor stacks exhibited the same sizing trends—optimal sizing requires the upper transistor (adjacent to the power supply) to be sized up by a factor of $\sim\sqrt{\alpha}$. Results for 90nm technology are displayed in Fig. 1, and indicate optimal ratios that are roughly 4% to 6.5% smaller than the theoretical $\sqrt{\alpha}$ factors stated earlier. Due to the small difference in current with the skewed sizing (~0.5% to 1.5% improvement), we will use a 1:1 width ratio in stacks. This reduces the design complexity for a negligibly small performance penalty.

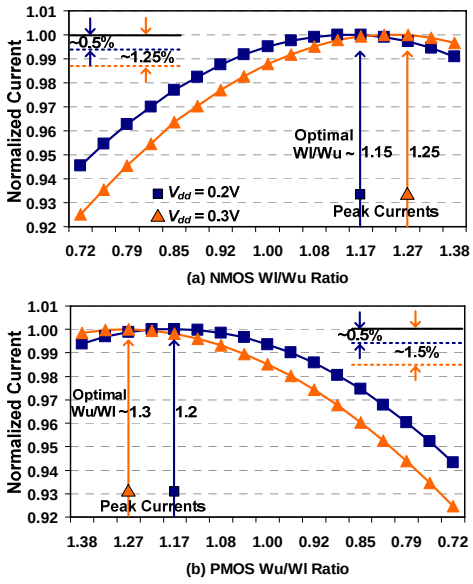


Fig. 1. DC current in stacks of two devices for a range of W_U:W_L sizing ratios. The total width of the stacked devices is held constant at 1um. The small benefits derived by using skewed stack sizing are indicated in the upper corners of the plots.

B. Optimal 2-Stack Sizing Factor

After deciding to use a 1:1 ratio for the two devices in a stack, we must find the amount by which they should be sized up to drive the same current as a single transistor. Defining $W = W_U = W_L$ as the size of each transistor in the stack, we can modify equation (6) as follows:

$$I_U = I_L = \frac{\alpha W^2}{\alpha W + W} e^{\frac{V_{dd}-V_{t0}}{mV_T}} = \frac{\alpha}{1+\alpha} W e^{\frac{V_{dd}-V_{t0}}{mV_T}} \quad (9)$$

For a single transistor, the current equation is:

$$I = W_{eff} e^{\frac{V_{dd}-(V_{t0}+\lambda_d V_{dd})}{mV_T}} = \alpha W_{eff} e^{\frac{V_{dd}-V_{t0}}{mV_T}}, \quad (10)$$

where W_{eff} stands for the effective width of this device. From equations (9) and (10), we have the following relationship:

$$\alpha W_{eff} = \frac{\alpha}{1+\alpha} W \rightarrow W_{eff} = \frac{1}{1+\alpha} W \quad (11)$$

According to this equation, two stacked transistors should be sized up by a factor of (1+α) in relation to a single device for the same current drivability. Tables I and II display (1+α) stack sizing values from this theory and from simulation results, demonstrating the validity of equation (11). DC simulations were performed to find the correct sizing for transistors in a stack which is capable of conducting the same amount of current as a single unit-sized device. Sizing factors found in simulations were slightly smaller than those predicted by the theory derived above due to effects not captured by current equation (1), but the trend with technology scaling is nearly identical in both cases.

Results indicate that stacks need to be sized up by a larger amount in the subthreshold region compared to the strong-inversion region. Also note that NMOS stack sizing factors are significantly smaller in strong inversion due to velocity saturation.

TABLE I
NMOS Stack Sizing Factors

Vdd	Sizing Method	130nm	90nm	65nm	45nm
0.2V	simulation	2.19	2.30	2.42	2.66
	theory	2.39	2.52	2.67	3.04
0.3V	simulation	2.27	2.44	2.64	3.11
	theory	2.50	2.70	2.93	3.57
1.2V	simulation	1.58	1.60	1.63	1.69

TABLE II
PMOS Stack Sizing Factors

Vdd	Sizing Method	130nm	90nm	65nm	45nm
0.2V	simulation	2.33	2.48	2.68	3.00
	theory	2.45	2.66	2.90	3.34
0.3V	simulation	2.60	2.85	3.20	3.95
	theory	2.57	2.88	3.28	4.13
1.2V	simulation	1.98	2.08	2.05	2.15

III. Arbitrary Stack Sizes

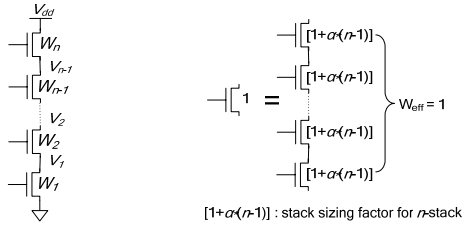
A. Proof of the Symmetry of the Lowest n-1 Device Widths in an n-Stack

Building an extensive cell library based on this stack sizing framework requires an extension of our work to stacks of three or more devices. The derivation for the current equation of a three-stack, which follows a similar method as the derivation in section II.A gives us the following result:

$$I = \alpha\left[\frac{(W_T - W_1 - W_2)W_1W_2}{\alpha(W_T - W_1 - W_2)(W_2 + W_1) + W_1W_2}\right] e^{\frac{V_{dd}-V_{t0}}{mV_T}} \quad (12)$$

W_1 and W_2 stand for the widths of the two lower transistors in the stack of NMOS devices (see notation in Fig. 2). W_T is defined as $W_T = W_1+W_2+W_3$, and is used to eliminate W_3, the width of the upper device. This equation is symmetric with respect to the widths of the W_1 and W_2 transistors, indicating that the optimal sizes for the lower two devices in the stack are equal. We now

extend this finding through a straight-forward direct proof, which confirms the symmetry of the lower *n*-1 transistor widths in a general *n*-stack achieving maximum drive current.



(a) *n*-stack notation (b) *n*-stack sizing for equivalent width

Fig. 2. NMOS *n*-stack

The following equations hold for the drive-current through the transistors in an *n*-stack:

$$I_n = \alpha W_n \beta e^{-V_{n-1}/V_T} = \alpha W_n \beta \nu_{n-1} \quad (13)$$

$$I_{n-1} = W_{n-1}\beta(e^{-V_{n-2}/V_T} - \nu_{n-1}) = W_{n-1}\beta(\nu_{n-2} - \nu_{n-1}) \quad (14)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$I_3 = W_3\beta(e^{-V_2/V_T} - \nu_3) = W_3\beta(\nu_2 - \nu_3) \quad (15)$$

$$I_2 = W_2\beta(e^{-V_1/V_T} - \nu_2) = W_2\beta(\nu_1 - \nu_2) \quad (16)$$

$$I_1 = W_1\beta(1 - \nu_1) \quad (17)$$

The ν_i variables are shorthand for e^{-Vi/V_T}, and β stands for $e^{(V_{dd}-V_{t0})/V_T}$.

Step 1: By setting equation (16) equal to equation (17), we can show that

$$\nu_1 = \frac{W_1 + W_2\nu_2}{W_1 + W_2} \quad (18)$$

Step 2: Next, by setting (15) equal to (16), and solving it for ν_2, we have

$$\nu_2 = \frac{W_3\nu_3 + W_{2\|1}}{W_3 + W_{2\|1}}, \quad (19)$$

where $W_{2\|1} = \frac{W_2 W_1}{W_2 + W_1}$ is called the parallel combination of W_1 and W_2. *Step 2* is now repeated to move up through the stack until we reach the equation $W_{n-1}(\nu_{n-2} - \nu_{n-1}) = W_{n-2}(\nu_{n-3} - \nu_{n-2})$. From this we find

$$\nu_{n-2} = \frac{\nu_{n-1}W_{n-1} + W_{\{n-2\|1\}}}{W_{n-1} + W_{\{n-2\|1\}}} \quad (20)$$

where $W_{\{n-2\|1\}}$ is the parallel combination of transistors 1 through *n-2*.

Step 3: Finally, setting equation (13) equal to (14), we can solve for ν_{n-1}

$$\nu_{n-1} = \frac{W_{\{n-1\|1\}}}{\alpha W_n + W_{\{n-1\|1\}}} \quad (21)$$

We now have the following current equation:

$$I_n = \alpha W_n \beta \nu_{n-1} = \beta\left[\frac{(\alpha W_n)W_{\{n-1\|1\}}}{\alpha W_n + W_{\{n-1\|1\}}}\right] \quad (22)$$

Defining $W_T = \sum_{i=1}^{n} W_i$ and substituting for W_n in equation (22), we get:

$$I_n = \beta\left[\frac{\alpha\left\{W_T - \sum_{i=1}^{n-1} W_i\right\}W_{\{n-1\|1\}}}{\alpha\left\{W_T - \sum_{i=1}^{n-1} W_i\right\} + W_{\{n-1\|1\}}}\right] \quad (23)$$

An examination of equation (23) shows that the variables W_1 through W_{n-1} appear symmetrically in the expression. Therefore, when I_n is optimized, W_1 through W_{n-1} must have identical values, since setting the partial derivative of I_n with respect to each W_i, for i = 1 to *n*-1, will result in a symmetric set of *n*-1 equations.

B. Optimal n-Stack Sizing Factor

Given the symmetry of the lower *n*-1 device sizes, i.e., $W_X = W_1 = W_2 = \ldots = W_{n-1}$, we have the following general form for I_n in an *n*-stack:

$$I_n = \beta\left[\frac{\alpha\{W_T - (n-1)W_X\}W_X}{\alpha\{W_T - (n-1)W_X\}(n-1) + W_X}\right] \quad (24)$$

To optimize I_n, we set $\partial I_n/\partial W_x$=0 to obtain

$$W_X = \frac{(\alpha n - \alpha - \sqrt{\alpha})}{(\alpha n^2 - 2\alpha n + \alpha - 1)}W_T \quad (25)$$

Using the definition of W_T, i.e., $W_n = W_T - W_X(n-1)$, we get

$$\frac{W_X}{W_n} = \left[\left(\frac{\alpha n^2 - 2\alpha n + \alpha - 1}{\alpha n - \alpha - \sqrt{\alpha}}\right) - (n-1)\right]^{-1} = \sqrt{\alpha} \quad (26)$$

Thus, we have proven that the $\sqrt{\alpha}$ sizing ratio holds for the general *n*-stack case.

As in the two-transistor stack case, the scaling factor of $\sqrt{\alpha}$ leads to a trivial performance benefit (e.g., a 0.3% increase in current through a PMOS or NMOS stack in 90nm technology with a total stack width of 1um), so sizing all stacked transistors equally is the best choice in terms of overall design complexity. Using equation (24) and following the example of equation (11), we find that each device in an *n*-stack should then be scaled up by a factor of [1+α*(*n*-1)] to set the effective width of the stack equal to that of a single unit transistor (Fig. 2). Note that all work done here again applies to PMOS stacks in a similar manner. The discrepancies between the larger sizing factors predicted by this theory and those found with simulations become slightly more pronounced as the stack size grows. For PMOS three stacks, the difference stays within the ~4-7% range, while for large alpha values, NMOS sizing factors are overestimated by up to ~15% due to second order effects not captured in equations (1) and (2).

IV. Simulation Results

A. Critical Path: A Chain of Stacks

We tested our sizing with 130nm, 90nm, 65nm, and 45nm PTM simulations using simple chains of logic gates that are representative of those that may be found in the critical path(s) of ultra low power circuits. In order to isolate the benefits of using the larger stack sizing in subthreshold operation, a consistent beta ratio (PMOS to NMOS width ratio) of 1.5 was employed across all simulations. This nominal value is close to that used in advanced CMOS processes. Stack sizing factors found with DC simulations as described in section II.B were used. These experimentally determined numbers closely match our theoretical results, as stated earlier.

The logical effort sizing method was used as a straight-forward means of quickly optimizing the delay though a logic path [7]. Logical effort is defined as the ratio of the input capacitance of a gate to that of an inverter driving the same amount of output

current. Fig. 3 displays logical effort values based on our stack sizing parameters, as well as the corresponding parasitic delay values. Parasitic delay represents the delay of a gate driving no load, and is set by the parasitic junction capacitance.

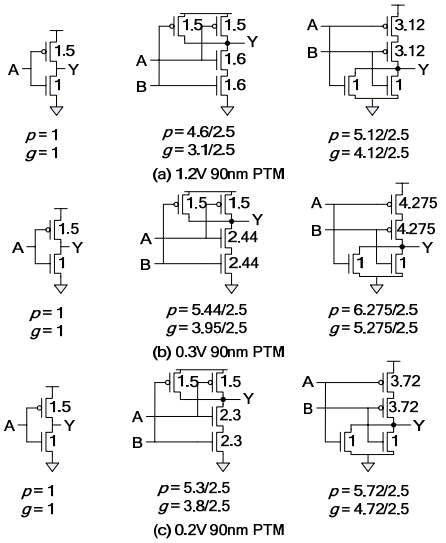


Fig. 3. Parasitic delay (*p*) and logical effort (*g*) values

While the additional loading on previous stages created by the larger stack sizes here can degrade the performance of some logic chains, critical paths driving substantial fanout capacitance, and particularly those containing paths dominated by stacks, do benefit from this sizing. The simple circuit illustrated in Fig. 4 is an example of a critical path whose delay is improved with our stack sizing framework. The fanout inverter widths were kept constant across all experiments, and their loading effect was taken into account through the branching factor [7]. The minimum width (i.e., the NMOS width in the unit-sized inverter) was held at 1um. The gate capacitance of the inverters indicated in Fig. 4 served as the input and output capacitance parameters for the logical effort calculations (Cin and Cout, respectively).

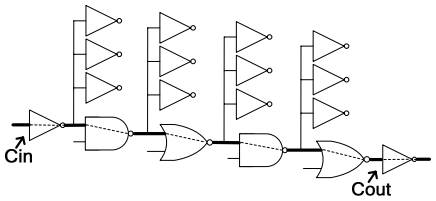


Fig. 4. Representative chain of logic gates with FO4 at each output

Delays were found for both the path through this circuit consisting entirely of stacks (the "Stacks" path), and that containing no stacks (the "Fast" path), using the worst-case input pattern for each. Critical path delay results for V_{dd} = 0.3V and V_{dd} = 0.2V are shown in Tables III and IV, respectively. As indicated here, the critical path shifts from the Stacks path to the Fast path when using the optimized subthreshold sizing, and the critical delay is consistently reduced. Also note that the 1.2V sizing scheme was optimal when operating in strong-inversion, with improvements over subthreshold sizing performance ranging from <1% to 12.3%.

In logic paths where there are not chains of stacks driving each other in sequence, the larger subthreshold stack sizing becomes less beneficial, or even detrimental in terms of performance, due

TABLE III
Critical Path Delay Improvement for V_{dd} = 0.3V

Technology	Conventional 1.2V sizing		Subthreshold 0.3V sizing	
	Delay	Crit. Path	Speedup	Crit. Path
130nm	14.86n	Stacks	7.3%	Fast
90nm	14.10n	Stacks	6.0%	Fast
65nm	16.14n	Stacks	8.1%	Fast
45nm	24.23n	Stacks	4.6%	Fast

TABLE IV
Critical Path Delay Improvement for V_{dd} = 0.2V

Technology	Conventional 1.2V sizing		Subthreshold 0.2V sizing	
	Delay	Crit. Path	Speedup	Crit. Path
130nm	98.12n	Stacks	6.6%	Fast
90nm	96.25n	Stacks	6.2%	Fast
65nm	113.8n	Stacks	8.1%	Fast
45nm	174.6n	Stacks	10.4%	Fast

to its loading effect on the previous stage. For instance, if inverters are inserted between each NAND/NOR pair in the circuit in Fig. 5, improvements in subthreshold with our larger stack sizes are reduced to ~1%. In a chain of just NAND gates, the smaller stack sizes used in superthreshold were generally better choices across all supply levels. In detailed optimization schemes, care must be taken to account for transient effects, including the variance of load capacitances as operating conditions change. DC sizing schemes such as the one presented here provide us with intuition about the devices we are constructing circuits with, and a starting point for thorough optimization procedures.

V. Conclusion

We have presented a new stack sizing framework for circuits operating in the subthreshold region. A closed-form solution for the optimal width ratio between different devices within a stack, as well as the sizing factor for stacked transistors was presented and shown to closely match experimental results. Our optimization scheme resulted in performance gains of up to 10+% in simulations of critical paths where internal node capacitance is not dominated by the increased stack sizing factors.

References

[1] A. Wang, A.P. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology", *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310-319, Jan. 2005.

[2] T. Kim *et al*., "A High-Density Subthreshold SRAM with Data Independent Bitline Leakage and Virtual Ground Replica Scheme", *Int. Solid-State Circuits Conf*., pp. 330-331, Feb. 2007.

[3] C.H. Kim *et al*., "Ultra-low-power DLMS adaptive filter for hearing aid applications", *IEEE Transactions on VLSI Systems*, vol. 11, no. 6, pp. 1058-1067, Dec. 2003.

[4] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations", *IEEE J. Solid-State Circuits*, vol. 12, no. 3, pp. 224-231, June 1977.

[5] Predictive Technology Model, online: http://www.eas.asu.edu/~ptm/.

[6] W. Zhao and Y. Cao, "New generation of Predictive Technology Model for sub-45nm Design Exploration," *Int. Symp. On Quality Electronic Design*, pp. 585-590, 2006.

[7] I. Sutherland et al., *Logical Effort: Designing Fast CMOS Circuits*. San Francisco, CA: Morgan Kaufmann, Jan. 1999.