

# Transistor Sizing of Energy-Delay-Efficient Circuits

Paul I. Péntzes, Mika Nyström, Alain J. Martin

*Computer Science Department*

*California Institute of Technology*

*Pasadena, CA 91125, U.S.A.*

{penzes,mika,alain}@async.caltech.edu

## Abstract

*This paper studies the problem of transistor sizing of CMOS circuits optimized for energy-delay efficiency, i.e., for optimal  $Et^n$  where  $E$  is the energy consumption and  $t$  is the delay of the circuit, while  $n$  is a fixed positive optimization index that reflects the chosen trade-off between energy and delay.*

*We propose a set of analytical formulas that closely approximate the optimal transistor sizes. We then study an efficient iteration procedure that can further improve the original analytical solution. Based on these results, we introduce a novel transistor sizing algorithm for energy-delay efficiency.*

## 1. Introduction

The rapidly increasing complexity of VLSI systems has made it necessary to pay ever more attention to design issues that affect energy consumption. One of the original motivations for CMOS technology was its low energy consumption, and today, there are still no alternatives that approach it in energy efficiency. Nevertheless, energy consumption is more and more often the factor that limits the performance of contemporary CMOS systems.

In order to compare designs that run at different speeds and consume different amounts of energy, we have to combine the energy,  $E$ , and the delay,  $t$ , into a single metric  $\delta$ . The authors have previously proposed  $\delta = Et^2$  as an energy-delay-efficiency metric for VLSI computation [1, 2, 17]. The main reason for choosing this metric over others is that  $\delta$  is to first order constant when we vary the supply voltage of a CMOS system: the delay falls roughly linearly with supply voltage, and the energy consumption increases roughly quadratically; therefore,  $Et^2$  stays roughly constant. Hence, the  $\delta$  metric allows the designer to factor “runtime” voltage scaling out of consideration. The authors have argued that, owing to its voltage independence, the  $\delta$  metric is superior to other efficiency metrics found in the literature, such as  $E$  or  $Et$  [3].

In practice, we can achieve any desired target speed or target energy consumption by adjusting the supply voltage. If we desire to change to a particular delay target  $t$ , we adjust the voltage to meet it, and a circuit optimized

for  $\delta$  would have the best  $E$  for that  $t$ . Likewise, we may choose an energy target  $E$  and get the best  $t$  instead.

The  $Et^2$  metric is a special case of a wider class of energy metrics, which includes  $E$  and  $Et$ , among others. The authors have shown that a metric of the more general form  $Et^n$  for  $n \geq 0$  characterizes any feasible trade-off, not only the trade-off through voltage scaling, between the energy and the delay of a computation [4]. For example, any problem of minimizing the energy of a circuit for a given target delay can be restated as minimizing  $Et^n$  for a certain  $n$ . We call  $n$  the *energy-delay efficiency index*.

In this paper, we study the problem of transistor sizing for energy-delay efficient circuits. Given a transistor netlist where each transistor  $i$  has width  $w_i$  and length  $l_i$ , transistor sizing finds the values of  $w_i$  and  $l_i$  that optimize the target function—in our case  $Et^n$ . While it is true that most layout systems demand that transistor sizes be quantized to some grid, we ignore this constraint.

Also, we can remove the  $l_i$ s from consideration since there is usually no reason to set the lengths of transistors in a digital circuit to anything other than the minimum allowed by the fabrication technology: increasing the length increases both the resistance and the capacitance and hence worsens both the energy and delay.

The sized transistors of a circuit are connected to each other through wires. The capacitance of these wires leads to additional energy and delay. (We ignore wire resistance in this paper.) For delay-only optimization, which can be phrased as the minimization of the metric  $Et^n$  for very large  $n$ , the wire capacitance can be overcome by increasing transistor sizes where appropriate. Conversely, for energy-only optimization, when  $n = 0$ , the transistor widths can be chosen to be minimum size, independently of the wire capacitance. In contrast to these special cases, for  $n$  small but nonzero, wire capacitance cannot be ignored or overcome in a straightforward way, and the optimal transistor sizes depend strongly on this capacitance.

In this paper, we propose an analytic formula for transistor sizing. If the approximate solution is acceptable for the given application, the formula can be used as is (no numerical optimization is then needed); however, if more accuracy is required, the formula can be used to provide a good starting point for numerical optimization. Later in the paper, we propose an efficient iteration procedure that

# Report Documentation Page

Form Approved  
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE <b>2006</b>		2. REPORT TYPE		3. DATES COVERED <b>00-00-2006 to 00-00-2006</b>	
4. TITLE AND SUBTITLE <b>Transistor Sizing of Energy-Delay-Efficient Circuits</b>				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>Defense Advanced Research Projects Agency, 3701 North Fairfax Drive, Arlington, VA, 22203-1714</b>				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT <b>see report</b>					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES <b>7</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

can further improve the accuracy of the original analytical solution. Based on these results, we introduce a novel transistor-sizing algorithm for energy-delay efficient circuits.

The proofs of properties and theorems have been omitted owing to space limitations. They can be found in the first author's Ph.D. dissertation [16].

## 2. Previous Work

Classical numerical methods, such as the conjugate gradient descent method, have been applied to the transistor-sizing problem: there exist several transistor sizing programs that minimize power consumption while maintaining performance specifications [5, 6, 7]. More recently, several specialized numerical techniques have been proposed [8, 9, 10]. On the analytical side, Cong and Koh have studied the related problem of simultaneous gate and wire optimization for optimal delay and power [13]. Cong and Koh's solution space and optimization metric are different from what we shall see in the present paper. A different analytic approach to the transistor sizing problem, for the performance metric  $Et$ , is given by Hu [11] and another by Horowitz, Indermaur, and Gonzalez [12]. Both Hu and Horowitz *et al.* present qualitative results; they only analyze basic inverter gates. To the best of the authors' knowledge, the present paper is the first one that goes beyond such a qualitative approach, both in terms of the generality of the optimization metric and in terms of the generality of the considered circuits.

## 3. $Et^n$ -optimal circuits

Let  $t$  be the cycle time of the critical cycle of the circuit whose transistor sizes we wish to optimize. We assume that the circuit is designed so that all cycles are critical; this is true in many well designed circuits, and it is true for any optimally sized circuit in the absence of additional constraints on transistor sizes (such as minimum-size constraints or slew-rate constraints). Let  $E$  be the energy consumption of the critical cycle. Let us further assume that  $E$  is a constant proportion of the total energy consumption; in this case, optimizing the energy  $E$  of the critical cycle optimizes the total energy of the circuit, and vice-versa.

Using the  $\tau$ -model [14, 15], we can write the energy as

$$E = \sum_{i=0}^{m-1} (w_{ni} + w_{pi} + p_i), \quad (1)$$

and the delay as

$$t = \sum_{i=0}^{m-1} \frac{k_{ni} f_{i+1} (w_{n(i+1)} + w_{p(i+1)} + p_{i+1})}{w_{ni}} + \sum_{i=0}^{m-1} \frac{\mu k_{pi} f_{i+1} (w_{n(i+1)} + w_{p(i+1)} + p_{i+1})}{w_{pi}}, \quad (2)$$

where  $w_{ni}$  and  $w_{pi}$  are the nFET and pFET (nMOS and pMOS transistor) widths of logic gate  $i$ ;  $k_{ni}, k_{pi} > 0$  are the numbers of nFETs and pFETs in series in logic gate  $i$ ;  $p_{i+1} > 0$  represents the wire parasitics at the output of logic gate  $i$ ;  $f_{i+1} > 0$  is the fanout of logic gate  $i$ ;  $\mu$  is the ratio of electron mobility to hole mobility;  $m$  is the length of the cycle, and  $i \in 0..m-1$  with all indices modulo  $m$ . In writing Equations 1 and 2 we have made several simplifying assumptions. We ignored the energy consumption due to short-circuit and leakage currents. Furthermore, we have constrained all devices in a series transistor network to have the same width. Finally, we have ignored the wire RC and time-of-flight delays.

## 4. Properties of transistor sizes in $Et^n$ -optimal circuits

**Property 1** *If  $w_i$  are the values that minimize  $Et^n$  for a given set of wire parasitics  $p_i$  and gate topologies  $k_i$ , then  $\alpha w_i$ ,  $\alpha > 0$  are the values that minimize  $Et^n$  for the set of wire parasitics  $\alpha p_i$  and gate topologies  $k_i$ .*

**Property 2** *If  $w_i$  are the values that minimize  $Et^n$  for a given set of wire parasitics  $p_i$  and gate topologies  $k_i$ , then  $w_i$  also minimize  $Et^n$  for the set of wire parasitics  $p_i$  and gate topologies  $\alpha k_i$ .*

If we ignore special constraints on transistor sizes, such as minimum-size and minimum-slew-rate constraints, and if we further assume that every transition on every circuit node matters to the circuit's overall speed (this last assumption is especially relevant in asynchronous circuits), then we can show that, when a system is optimized for  $Et^n$ , the widths of the nFETs and pFETs of each gate  $i$  are related as follows [16]:

$$w_{pi} = w_{ni} \sqrt{\mu \frac{k_{pi}}{k_{ni}}}. \quad (3)$$

Equation 3 is a local relationship; it does not depend on either  $E$ ,  $t$  or  $n$ . Equation 3 allows us to eliminate either the nFETs or the pFETs from the transistor-sizing problem. In particular, with the notation

$$w_i = w_{ni} + w_{pi} = w_{ni} \left( 1 + \sqrt{\frac{\mu k_{pi}}{k_{ni}}} \right) \quad (4)$$

and

$$k_i = f_{i+1} k_{ni} \left( 1 + \sqrt{\frac{\mu k_{pi}}{k_{ni}}} \right)^2, \quad (5)$$

by eliminating the pFET sizes from Equations 1 and 2, we get

$$E = \sum_{i=0}^{m-1} (w_i + p_i) \quad (6)$$

and

$$t = \sum_{i=0}^{m-1} k_i \frac{w_{i+1} + p_{i+1}}{w_i}. \quad (7)$$

We shall use these simpler formulas in the expressions for  $Et^n$ .

## 5. Preliminaries for $Et^n$ -optimal transistor sizing

We formalize the sizing problem of a transistor netlist for minimal  $Et^n$  as the minimization, over the  $w_i$ s, of  $Et^n$  where  $E$  and  $t$  are given by Equations 6 and 7.

$$Et^n = \left( \sum_{i=0}^{m-1} (w_i + p_i) \right) \left( \sum_{i=0}^{m-1} k_{i-1} \frac{w_i + p_i}{w_{i-1}} \right)^n \quad (8)$$

Note that Equation 8 holds not only for a ring, but also for a chain of gates, as long as the widths and parasitics for the input of the chain are equal to the widths and parasitics for the output of the chain (since in this case the  $E$  and  $t$  for a chain have the same form as the ones for a ring). This is an important observation, as it makes our results for transistor sizing applicable both to latency and cycle-time minimization.

$Et^n$  is a *posynomial* function of the transistor widths. A posynomial in variables  $w_i$  is a function of the form  $\sum_{0 < i \leq q} \alpha_i w_0^{\beta_0} w_1^{\beta_1} \dots w_{m-1}^{\beta_{m-1}}$  where  $\alpha_i \geq 0$ . A *posynomial problem* is the minimization of one posynomial while simultaneously satisfying a set of upper-bound constraints on other posynomials. With the substitution  $w_i = e^{x_i}$ , any posynomial can be transformed into a convex function; therefore the unique optimum of  $Et^n$  is achieved when  $\forall i : \frac{\partial Et^n}{\partial w_i} = 0$ .

This implies that the optimum is achieved when

$$\begin{aligned} \forall i : \quad & \frac{k_{i-1}}{w_{i-1}} - \frac{k_i(w_{i+1} + p_{i+1})}{w_i^2} \\ & = -\frac{1}{n} \frac{\sum_{i=0}^{m-1} k_{i-1} \frac{w_i + p_i}{w_{i-1}}}{\sum_{i=0}^{m-1} (w_i + p_i)} = -\frac{1}{n} \frac{1}{P}, \end{aligned} \quad (9)$$

where  $P = E/t$  is the power consumption of the chosen cycle. If  $\forall i : p_i = 0$  (no wire parasitics) and  $n$  is very large (delay-only optimization), Equation 9 reduces to

$$k_i \frac{w_{i+1}}{w_i} = k_{i-1} \frac{w_i}{w_{i-1}},$$

which is the known condition of equal stage delays for delay-only transistor sizing [14]. If we were able to solve Equation 9 analytically for any  $p_i$ s and  $k_i$ s, we could compute the optimal  $w_i$ s directly and our transistor-sizing problem would be solved. Unfortunately, this is not the case. We can compute an exact analytical solution of Equation 9 only for a restricted class of  $p_i$ s and  $k_i$ s [16]. In particular, we can show that if  $\forall i : k_i = k$ , i.e., the case of homogeneous circuits, and  $\forall i : p_i = p$  then

$$\forall i : w_i = np \quad \forall p, k > 0. \quad (10)$$

Equation 10 states that the transistor widths  $w_i$  of a homogeneous circuit with equal wire parasitics  $p$ , optimized for  $Et^n$ , are all equal to  $np$ , independently of  $k$  [17, 18].

## 6. $Et^n$ -optimal transistor sizes

So far we have explored some general properties of transistor sizes for circuits optimized for  $Et^n$ . Based on these properties, we now develop a simple analytical formula that approximates the transistor sizes of an  $Et^n$ -optimal circuit.

We start by finding an approximate formula for the transistor sizes  $w_i$  that optimize  $Et^n$ , in Equation 8, when  $\forall i : k_i = k$ . We then extend this formula to the case when the  $k_i$ s are no longer equal to each other.

### 6.1. Homogeneous Circuits

For the case when  $\forall i : k_i = k$ , we propose an approximate solution of the  $w_i$ s, of the following form:

$$w_i = \alpha_1 p_{i+1} + \alpha_2 p_{Avg} \quad (11)$$

where

$$p_{Avg} = \frac{1}{m} \sum p_i \quad (12)$$

and  $\alpha_1$  and  $\alpha_2$  are constants to be determined later. First, let us motivate Equation 11. Based on Property 2, we know that finding the  $w_i$ s when  $\forall i : k_i = k$  is equivalent to finding the  $w_i$ s when  $\forall i : k_i = 1$ . In other words, the value of the  $w_i$ s is independent of the  $k_i$ s, when all  $k_i$ s are equal. Conversely, based on Property 1, we know that the  $w_i$ s scale linearly with the  $p_i$ s. This suggests that the  $w_i$ s should not have terms that are independent of the  $p_i$ s. Based on our experience of sizing, we know that—while the transistor sizes of gate  $i$  depend mostly on  $p_{i+1}$ —the effect of a particular  $p_i$  gets distributed to some degree to all other gates. As a consequence, we would like Equation 11 to depend linearly on both  $p_{i+1}$  and some average of all other  $p_i$ s and one such choice is  $\alpha_1 p_{i+1} + \alpha_2 p_{Avg}$ . We use the arithmetic mean for  $p_{Avg}$  since the  $p_i$ s correspond physically to wire capacitances that are manipulated additively both in terms of delay and energy. With these clarifications in mind, we state the following:

**Theorem 1** *For a neighborhood  $\mathcal{V}_p = [p - \eta, p + \eta]$  of  $p > 0$ ,  $\eta > 0$ , the values of  $\alpha_1$  and  $\alpha_2$  that minimize  $Et^n$  given the  $w_i$ s of the form defined by Equation 11, where  $\forall i : p_i \in \mathcal{V}_p$ ,  $k_i = k > 0$  and  $\eta \rightarrow 0$ , are*

$$\alpha_1 = \frac{\frac{1}{2}}{\frac{1}{n} + \frac{m}{m-1}} \text{ and } \alpha_2 = n - \frac{\frac{1}{2}}{\frac{1}{n} + \frac{m}{m-1}}.$$

If the problem is large, i.e.,  $m \rightarrow \infty$ ,  $\frac{m}{m-1} \approx 1 \Rightarrow \alpha_1 = \frac{n}{2(1+n)}$  and  $\alpha_2 = \frac{n(1+2n)}{2(1+n)}$ , thus  $\frac{\alpha_2}{\alpha_1} = 1 + 2n$ . What is particularly surprising about Equation 11 is that the strength of a given gate depends far more strongly ( $5 \times$  for  $Et^2$  optimization) on the *average* parasitic load ( $\alpha_2 = 5/3$ ) than it does on the load on that *particular* gate ( $\alpha_1 = 1/3$ ). Furthermore,  $\lim_{n \rightarrow 0} \alpha_1 = \lim_{n \rightarrow 0} \alpha_2 = 0 \Rightarrow \forall i : w_i = 0$  for  $n = 0$  regardless of the  $p_i$ s. In other words, for energy-only optimization, Equation 11 yields minimum-size transistors, as one might expect.

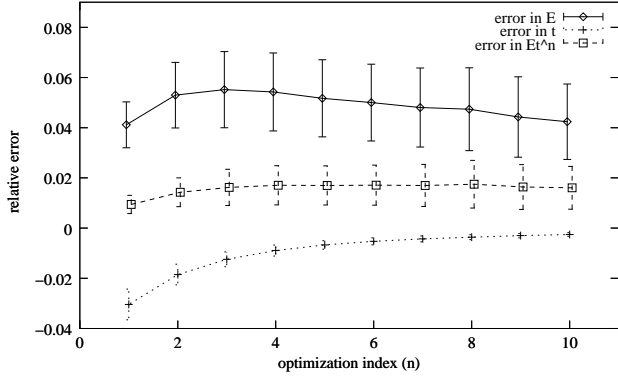


Figure 1. Accuracy in  $E$ ,  $t$  and  $Et^n$  of Equation 11 with  $\alpha_1$  and  $\alpha_2$  given by Theorem 1.

Theorem 1 yields the optimal values of  $\alpha_1$  and  $\alpha_2$  in a close neighborhood of  $p$ , or equivalently when the  $p_i$ s are close to each other. We want to check now if the form of the  $w_i$ s given by Equation 11 and Theorem 1 yields a practical approximation of the  $w_i$ s when  $\forall i : k_i = k$  but the  $p_i$ s are no longer close to each other. We use a numerical optimizer to compute the error between the optimal and the predicted  $Et^n$  for a given  $n$ ,  $m$  and a set of  $p_i$ s. We varied  $m \in [2, 1000]$ ,  $n \in [1, 10]$  and used three different distributions (uniform, uniform-squared, and uniform-cubed) for  $p_i \in [1, 100]$ . The observed errors are practically independent of the problem size  $m$  and the distribution chosen for the  $p_i$ s; the errors only depend on  $n$ . Figure 1 shows the relative error in  $E$ ,  $t$  and  $Et^n$  for  $m = 31$ ,  $n \in [1, 10]$ ,  $k_i = 1$  and  $p_i \in [1, 100]$  chosen randomly through a uniform-squared distribution. The average error in  $E$  is between 4.1% and 5.5%, the average error in  $t$  is between -3.0% and -0.3%, and the average error in  $Et^n$  is between 1.0% and 1.7%.

## 6.2. Non-homogeneous Circuits (first form)

The formula resulting from Theorem 1 yields excellent results when all  $k_i$ s are equal. We would like to extend it to incorporate the case when the  $k_i$ s are no longer all equal. To do this, we assume that the cumulative effect of the  $p_i$ s and the  $k_i$ s on the  $w_i$ s can be viewed as the product between the individual effect of the  $p_i$ s (wire capacitances) on the  $w_i$ s and the individual effect of the  $k_i$ s (gate topologies) on the  $w_i$ s. Hence, we propose an approximate solution of the  $w_i$ s of the following form:

$$w_i = (\alpha_1 p_{i+1} + \alpha_2 p_{Avg}) r_i(k_0, k_1, \dots, k_{m-1}) \quad (13)$$

where  $\alpha_1$  and  $\alpha_2$  are given by Theorem 1, while functions  $r_i$  will be determined later. When all gates are identical, i.e.,  $\forall i : k_i = k$ , we know from Equation 10 that the  $w_i$ s are independent of the  $k_i$ s. For this reason, we choose  $r_i(k_0, k_1, \dots, k_{m-1})$  such that  $\forall k : r_i(k, k, \dots, k) = 1$ .

Based on our experience on delay-only transistor sizing, we know that—while the transistor sizes of gate  $i$  depend strongly on  $k_i$ —the effect of a particular  $k_i$  gets

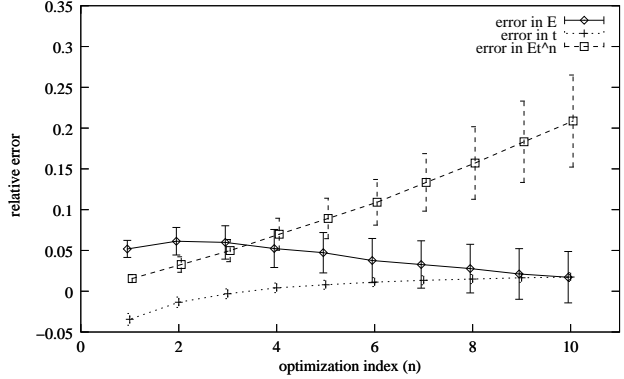


Figure 2. Accuracy in  $E$ ,  $t$  and  $Et^n$  of Equation 13 with  $\alpha_1$ ,  $\alpha_2$ , and  $r_i$ s given by Theorem 2.

distributed to some degree to all other gates. As a consequence, we would like  $r_i$  to depend on both  $k_i$  and some average of all other  $k_i$ s. We use the geometric mean  $k_{Avg} = \sqrt[n]{\prod k_i}$  as an average of the  $k_i$ s, since it has physical meaning—it is proportional to the theoretical minimal delay of the cycle. In this context, we introduce the following

**Theorem 2** For a neighborhood  $\mathcal{V}_p = [p - \eta, p + \eta]$  of  $p > 0$ ,  $\eta > 0$ , and a neighborhood  $\mathcal{V}_k = [k - \eta, k + \eta]$  of  $k > 0$ , the values of  $\alpha_1$ ,  $\alpha_2$ ,  $\beta_1$  and  $\beta_2$  that minimize  $Et^n$  given the  $w_i$ s of the form defined by Equation 13 with  $r_i(k_0, k_1, \dots, k_{m-1}) = \beta_1 \frac{k_i}{k_{Avg}} + \beta_2$ , where  $\forall i : p_i \in \mathcal{V}_p$ ,  $k_i \in \mathcal{V}_k$ , and  $\eta \rightarrow 0$ , are

$$\alpha_1 = \frac{\frac{1}{2}}{\frac{1}{n} + \frac{m}{m-1}}, \alpha_2 = n - \frac{\frac{1}{2}}{\frac{1}{n} + \frac{m}{m-1}}, \text{ and } \beta_1 = \beta_2 = \frac{1}{2}.$$

Theorem 2 yields the optimal values of  $\alpha_1$ ,  $\alpha_2$ ,  $\beta_1$  and  $\beta_2$  when the  $p_i$ s are in a close neighborhood of  $p$ , and the  $k_i$ s are in a close neighborhood of  $k$ . We would like to verify now how good these values are in minimizing  $Et^n$  when the  $p_i$ s and the  $k_i$ s are no longer close to each other. We use again a numerical optimizer to compute the error between the optimal and the estimated  $Et^n$  for a given  $n$ ,  $m$  and a set of  $p_i$ s and  $k_i$ s. We vary  $m \in [2, 1000]$ ,  $n \in [1, 10]$  and use three different distributions (uniform, uniform-squared, and uniform-cubed) for  $p_i \in [1, 100]$  and  $k_i \in [1, 3.3]$  (if we assume  $k_{ni} \in [1, 6]$  and  $k_{pi} \in [1, 2]$ , then with  $\mu = 2.5$  we get  $k_i \in [6.66, 21.95]$  or equivalently, using Property 2,  $k_i \in [1, 3.3]$ ). As for Equation 11, the observed errors are practically independent of the problem size  $m$  and the distribution chosen for the  $p_i$ s and the  $k_i$ s; the errors only depend on  $n$ . Figure 2 shows the relative error in  $E$ ,  $t$  and  $Et^n$  for  $m = 31$ ,  $n \in [1, 10]$  and  $p_i \in [1, 100]$ ,  $k_i \in [1, 3.3]$  chosen randomly through a uniform-squared distribution. The average error in  $E$  is between 1.7% and 6.1%, the average error in  $t$  is between -3.4% and 1.7%, while the average error in  $Et^n$  is about 3.3% for  $n = 2$ , but increasing about linearly with  $n$ , ow-

ing to the error amplifying artifact of  $Et^n$  (if  $t = t_0(1+\Delta) \Rightarrow t^n \approx t_0^n(1+n\Delta)$  for small  $\Delta$ ).

### 6.3. Non-homogeneous Circuits (second form)

The main intended use of Equation 13 in energy-delay efficient design is to find approximate transistor sizes when  $n \approx 2$ , i.e., when voltage scaling is a design parameter. As Figure 2 shows, the equation stated by Theorem 2, i.e., a particular case of Equation 13, does this reasonably well—i.e., within a few percent of the optimum. On the other hand, one might want to use such a sizing formula for large  $n$  as well—i.e., predominantly delay-only optimization. Getting a close approximation of  $Et^n$  when  $n$  is large requires a very good delay estimate, since even a small error  $\Delta$  in  $t$  gets linearly amplified to  $n\Delta$  in  $Et^n$ . For this reason, we study the behavior of Equation 13 and the delay estimate resulting from it, when  $n \rightarrow \infty$ .

For now, consider a simpler problem, namely finding the transistor widths  $w_{\infty i}$  that minimize  $t$  given by Equation 7. This is a special case of the  $Et^n$  optimization problem for  $n \rightarrow \infty$ . In [16] we have shown that the optimal delay  $t_{\infty} = mk_{Avg}$  is reached for transistor widths that have the property

$$\forall i : \frac{w_{\infty(i+1)}}{w_{\infty i}} = \frac{k_{Avg}}{k_i}. \quad (14)$$

We would like the  $w_i$ s given by Equation 13 to have property (14) for large  $n$ . More precisely,

$$\lim_{n \rightarrow \infty} \frac{w_{i+1}}{w_i} = \frac{w_{\infty(i+1)}}{w_{\infty i}} \quad (15)$$

or equivalently, using  $\alpha_1$  and  $\alpha_2$  given by Theorem 2,

$$\lim_{n \rightarrow \infty} \frac{w_{i+1}}{w_i} = \lim_{n \rightarrow \infty} \frac{r_{i+1}(k_0, k_1, \dots, k_{m-1})}{r_i(k_0, k_1, \dots, k_{m-1})} = \frac{w_{\infty(i+1)}}{w_{\infty i}}. \quad (16)$$

Condition 16 guarantees that the delay estimate resulting from Equation 13 is optimal for large  $n$ . An obvious choice of the  $r_i$ s that fulfills (16) is  $\forall i : r_i(k_0, k_1, \dots, k_{m-1}) = \beta w_{\infty i}$ , where  $\beta > 0$  is a constant scaling factor. The role of  $\beta$  is to normalize the  $w_i$ s to the right energy level; its optimal value is stated by the following

**Theorem 3** For a neighborhood  $\mathcal{V}_p = [p - \eta, p + \eta]$  of  $p > 0$ ,  $\eta > 0$ , and a neighborhood  $\mathcal{V}_k = [k - \eta, k + \eta]$  of  $k > 0$ , the values of  $\alpha_1$ ,  $\alpha_2$ ,  $\beta$  that minimize  $Et^n$  given the  $w_i$ s of the form defined by Equation 13 with  $r_i(k_0, k_1, \dots, k_{m-1}) = \beta w_{\infty i}$ , where  $\forall i : p_i \in \mathcal{V}_p$ ,  $k_i \in \mathcal{V}_k$ , and  $\eta \rightarrow 0$ , are

$$\alpha_1 = \frac{\frac{1}{2}}{\frac{1}{n} + \frac{m}{m-1}}, \alpha_2 = n - \frac{\frac{1}{2}}{\frac{1}{n} + \frac{m}{m-1}}, \text{ and}$$

$$\beta = \frac{S_2}{2} \left( \left(1 - \frac{1}{n}\right) + \sqrt{\left(1 - \frac{1}{n}\right)^2 + \frac{4}{nS_1S_2}} \right),$$

where

$$S_1 = \frac{1}{m} \sum_{i=0}^{m-1} w_{\infty i} \text{ and } S_2 = \frac{1}{m} \sum_{i=0}^{m-1} \frac{1}{w_{\infty i}}.$$

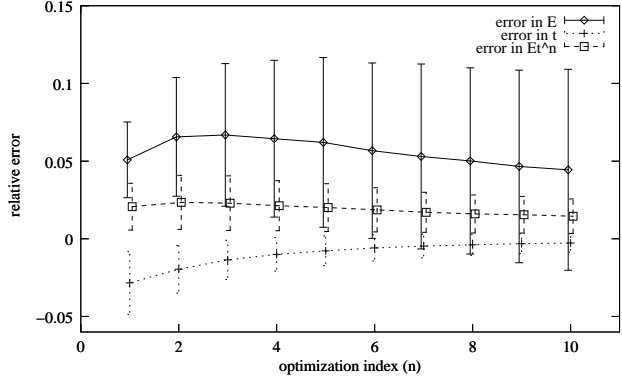


Figure 3. Accuracy in  $E$ ,  $t$  and  $Et^n$  of the approximation given by Theorem 3.

Assuming  $\forall i : p_i = p$ , and the  $w_i$ s given by Theorem 3, Equations 6 and 7 yield

$$E = (1 + nS_1\beta)E_0 \text{ and } t = \left(1 + \frac{S_2}{n\beta}\right)t_{\infty},$$

where  $E_0$  is the theoretical minimal energy (i.e. total switched wire parasitic) and  $t_{\infty}$  is the theoretical minimal delay. In [16, 17] we have shown that for a wide class of circuits

$$E \approx (1 + n)E_0 \text{ and } t \approx \left(1 + \frac{1}{n}\right)t_{\infty}.$$

Given the value of  $\beta$  from Theorem 3, we have that  $\forall n \geq 0 : \frac{1}{S_1} \leq \beta \leq S_2$  with  $\beta = \frac{1}{S_1}$  if  $n \rightarrow 0$  and  $\beta = S_2$  if  $n \rightarrow \infty$ . If we choose  $\beta = \frac{1}{S_1}$ , the error in  $E$  is reduced by bringing  $E$  close to  $(1 + n)E_0$ , while if we choose  $\beta = S_2$ , the error in  $t$  is reduced by bringing  $t$  close to  $(1 + \frac{1}{n})t_{\infty}$ .

The formula resulting from Theorem 3 works extremely well in practice for small  $m$ , i.e., it keeps the error in  $Et^n$  very low for the entire range of  $n$ , including large  $n$ . However, for  $m$  large the accuracy of the formula deteriorates somewhat due to the fact that  $E$  becomes consistently overestimated, while the estimate in  $t$  stays very accurate. This is a consequence of the choice of the  $r_i$ s, where we have intentionally favored the accuracy of the delay estimation. For large  $m$ s, the difference between  $\frac{1}{S_1}$  and  $S_2$  becomes large enough so that the resulting  $\beta$  pulls  $E$  noticeably away from the optimum  $(1 + n)E_0$ .

Figure 3 shows the relative error in  $E$ ,  $t$  and  $Et^n$  for the approximation given by Theorem 3 for  $m = 9$  (an 18 transitions per cycle circuit),  $n \in [1, 10]$  and  $p_i \in [1, 100]$ ,  $k_i \in [1, 3.3]$  chosen randomly through a uniform-squared distribution. The average error in  $E$  is between 4.4% and 6.7%, the average error in  $t$  is between -0.2% and -2.8%, and the average error in  $Et^n$  is between 1.4% and 2.3%. It is interesting to point out that for  $n = 100$ , the average error in  $E$  is about 1.2%, the average error in  $t$  is about -0.003%, and the average error in  $Et^n$  is about 0.5%.

For clarity, Theorems 1, 2, and 3 were formulated to refer to the transistor sizing problem of a single-cycle system. However, these theorems can be easily extended to multi-cycle systems. We extend formula 11, and as a consequence Theorem 1, to multi-cycle systems by redefining  $p_{Avg}$  for each gate  $i$  to be the average parasitic of all simple cycles gate  $i$  is part of. Theorem 2 extends to multi-cycle systems by substituting  $mk_{Avg}$  with  $t_\infty$  (the minimum achievable delay of the circuit). Given the definition of  $w_{\infty i}$ s and  $p_{Avg}$ , Theorem 3 generalizes straightforwardly to multi-cycle systems, with the only remark that  $m$ —in the expression of  $S_1$  and  $S_2$ —represents the total number of transistors in the considered circuit, not just the ones on a given cycle.

Remembering the derivation of Section 4, the values of the  $w_i$ s are per gate  $i$ ; but they can be transformed into the effective nFET and pFET sizes directly, using Equations 3, 4 and 5.

## 7. An iterative approach to $Et^n$ -optimal transistor sizing

With the help of Theorems 2 and 3, we can compute approximate transistor sizes of an  $Et^n$ -optimal circuit. As we have seen, the approximate solution yields energy and delay values within a few percent of the optimum. However, if the accuracy of such a solution is not acceptable for the given application, one might wish to employ an iterative procedure to “fine tune” the initial transistor sizes.

Using Equation 9, we can compute  $w_i$ —for a fixed  $i$ —as a function of the other  $w$ s. More precisely, if we call

$$a_2 = \frac{b_1 + nb_0b_2}{(n+1)b_2}, \quad a_1 = \frac{(-n+1)b_3}{(n+1)b_2}, \quad \text{and} \quad a_0 = \frac{-nb_0b_3}{(n+1)b_2},$$

where

$$b_0 = \sum_{i=1, i \neq j}^m w_j + \sum_{i=1}^m p_j,$$

$$b_1 = \sum_{i=1, i \neq j, i \neq j+1}^m k_{j-1} \frac{w_j + p_j}{w_{j-1}} + k_{i-1} \frac{p_i}{w_{i-1}},$$

$$b_2 = \frac{k_{i-1}}{w_{i-1}},$$

and

$$b_3 = k_i(w_{i+1} + p_{i+1});$$

we can compute  $w_i$  as the positive solution to the cubic equation

$$w_i^3 + a_2w_i^2 + a_1w_i + a_0 = 0. \quad (17)$$

(Equation 17 has a single positive root for  $n \geq 1$ ; this can be found using Cardan’s method.)

The iterative procedure starts with an initial solution and then repetitively computes each  $w_i$  as the positive solution of Equation 17 with coefficients computed from the current value of all other  $w$ s. It is easy to see that such a procedure converges to the  $Et^n$ -optimal solution. First, the recomputed value of  $w_i$  yields a better  $Et^n$  than the

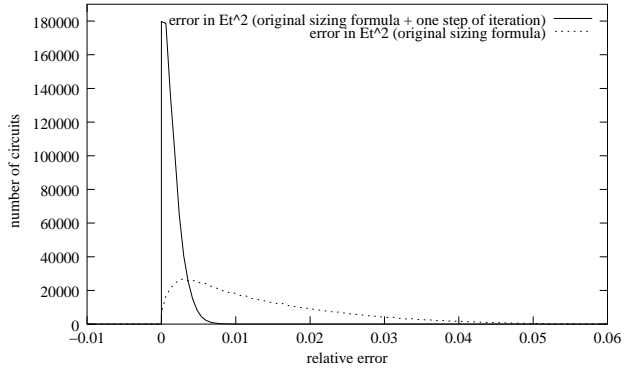


Figure 4. Error in  $Et^2$  when exhaustively simulating an entire class of circuits.

pre-iteration value. This is because  $\frac{\partial Et^n}{\partial w_i} = 0$ , i.e., the new  $w_i$  is  $Et^n$ -optimal when all other  $w$ s are fixed at their current value. Secondly, the  $Et^n$  optimization problem is convex in the  $w$ s, hence a local minimum reached by the iteration procedure is indeed the global minimum.

To fully appreciate the benefit of the proposed iteration procedure when applied to the initial solution given by Theorems 2 or 3, we exhaustively analyze a particular case of Equation 8 with  $n = 2$ ,  $m = 5$ ,  $p_i \in \{1, 2, 3, 4, 5\}$  and  $k_i \in \{1, 2, 3\}$ . Figure 4 shows a histogram of the relative error in  $Et^2$  between the optimal values (computed with an optimization algorithm) and the estimated values based on Theorem 3, and also between the optimal values and the values computed by one step of the iteration procedure starting with the approximate solution given by Theorem 3. One step of iteration assigns one new value to each  $w_i$ . We observe that the already small maximal error of the original sizing formula is reduced about ten-fold by a single step of the iteration procedure. Of course, one can repeat the same procedure and get an even smaller error. However, this second step does not have the same impact on reducing the error as the first step had. Given that the transistor sizes of a real circuit are integer multiples of a technology dependent constant, there is not much point in trying to find the zero-error solution. That solution is unlikely to be implementable in practice, since it will likely have non-integer components.

We have done several experiments in which we tested the dependence of the iteration procedure on the initial starting point. We have found that the applicability of the method strongly depends on the initial solution’s proximity to the optimal solution. Without a good initial solution like the one given by Theorem 2 or 3, the method still converges eventually to the optimum. However, the first step of iteration yields a solution that has an error spread two orders of magnitude greater than the solution resulting from the first step of the iteration executed on the good initial solution.

## 8. An algorithm for $Et^n$ -optimal sizing

As we have seen, the transistor sizes  $w_i$  of a system optimized for  $Et^n$  depend strongly on the wire parasitics  $p_i$ . Unfortunately, these parasitics are not known *a priori*, since they are attributes of wires that connect transistors whose dimensions have not yet been found.

A two-phase algorithm solves the problem of the unknown parasitics. In the first phase, given the transistor netlist, each wire is assigned an initial wiring cost. The more is known about the structure of the transistor netlist and about a future floorplan, the more accurate such an assignment will be. Based on these initial wire parasitics, we can then compute an initial estimate for the  $w_i$ s with the formulas established by Theorems 2 and 3.

In the second phase, we wire up the pre-sized transistors and extract the actual wire capacitances from the layout. With these new parasitics, we recompute the transistor widths  $w_i$ . Finally, we may fine-tune the solution by iterating once as described in Section 7.

If the accuracy of the final solution should not be deemed acceptable, we can add a pass through a classical numerical optimizer. Given the proximity of the current solution to the optimum, such an optimization will converge quickly. In this last phase, a more accurate transistor model (e.g., a BSIM model) can be employed, so as to bridge the gap between the simplified transistor model used in this paper and the actual transistor behavior.

## 9. Conclusions

We have proposed a set of analytical formulas that closely approximate the optimal transistor sizes for circuits optimized for  $Et^n$ . We have justified the validity of these formulas both mathematically and experimentally. We have proposed an iterative procedure that can further improve the accuracy of the original analytical solution. Experiments show that, when the procedure is applied on the analytical solution, it converges much more quickly than with an arbitrary starting point. Based on these results, we have introduced a novel transistor sizing algorithm for energy-delay efficiency.

## 10. Acknowledgments

The authors thank Catherine Wong and Karl Papadantonakis for many stimulating discussions.

The research reported in this paper was sponsored by the Defense Advanced Research Projects Agency and monitored by the Air Force under contract F29601-00-K-0184.

[1] Alain J. Martin, Andrew Lines, Rajit Manohar, Mika Nyström, Paul Pézses, Robert Southworth and Uri Cummings. *The Design of an Asynchronous MIPS R3000 Microprocessor*. Proceedings of the 17th Conference on Advanced Research in VLSI, IEEE Computer Society Press, p164-181, 1997.

[2] Alain J. Martin, *Towards an Energy Complexity of Computation*. Information Processing Letters, 77, 2001.

[3] R. Gonzalez, M. Horowitz, *Supply and threshold voltage scaling for low power CMOS*. IEEE Journal of Solid-State Circuits, August 1997.

[4] Paul I. Pézses, Alain J. Martin, *Energy-Delay Efficiency of VLSI Computations*. 12th Great Lakes Symposium on VLSI, New York, USA April 18-19, 2002

[5] P.E.Gill, W.Murray, M.H.Wright, *Practical Optimization*. Academic Press, 1981.

[6] D.P.Marple, *Transistor size optimization in the Tailor layout system*. Design Automation Conference, 1989, pp. 43-48

[7] J.P.Fishburn, A.E.Dunlop, *TILOS: A posynomial approach to transistor sizing*. Proceedings of the 1985 International Conference on Computer-aided Design, Nov. 1985, pp. 326-328

[8] B. Hoppe, G. Neuendorf, D. Schmitt-Landsiedel, W. Specks, *Optimization of high-speed CMOS logic circuits with analytical models for signal delay, chip area, and dynamic power dissipation*. IEEE Transactions on Computer-Aided Design, 9(3):236-247, 1990.

[9] Y. Tamiya, Y. Matsunaga, M. Fujita, *LP based Cell Selection with Constraints on Timing, Area and Power Consumption*. in Proc. ICCAD Conf., Nov. 1994.

[10] G. Chen, H. Onodera, K. Tamaru, *An Iterative Gate Sizing Approach with Accurate Delay Evaluation*. Proc. IEEE Int'l. Conf. on CAD, 1995.

[11] Chenming Hu, *Device and Technology Impact on Low Power Electronics*. Low Power Design Methodologies, Kluwer Academic/Plenum Publishers, 1996

[12] M. Horowitz, T. Indermaur, R. Gonzalez, *Low-power digital design*. Symposium on Low Power Electronics, October 1994, pages 8-11.

[13] J. Cong, C. K. Koh, *Simultaneous Driver and Wire Sizing for Performance and Power Optimization*. IEEE Trans. on VLSI Systems, pp. 408-425, December 1994.

[14] C. Mead, L. Conway, *Introduction to VLSI systems*. Addison Wesley, 1980

[15] J. Rubenstein, P. Penfield, M. A. Horowitz, *Signal delay in RC tree networks*. IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems 2 (1983), no. 3, 202-211.

[16] Paul I. Pézses, *Energy-delay Complexity of Asynchronous Circuits*. Ph.D. Thesis (in preparation), California Institute of Technology, 2002.

[17] Alain J. Martin, Mika Nyström, Paul I. Pézses. *ET<sup>2</sup>: A Metric for Time and Energy Efficiency of Computation*. Power-Aware Computing, Kluwer Academic/Plenum Publishers, 2002

[18] Paul I. Pézses, Alain J. Martin, *Global and Local Properties of Asynchronous Circuits Optimized for Energy Efficiency*. IEEE Workshop on Power Management for Real-time and Embedded Systems, Taipei, Taiwan, May 29th, 2001.