Energy-Delay Efficiency of VLSI Computations

Paul I. Pénzes, Alain J. Martin
Computer Science Department
California Institute of Technology
Pasadena, CA 91125, U.S.A.
penzes@async.caltech.edu, alain@async.caltech.edu

ABSTRACT

In this paper we introduce an energy-delay efficiency metric that captures *any* trade-off between the energy and the delay of the computation.

We apply this new concept to the parallel and sequential composition of circuits in general and in particular to circuits optimized through transistor sizing. We bound the delay and energy of the optimized circuit and we give necessary and sufficient conditions under which these bounds are reached. We also give necessary and sufficient conditions under which subcomponents of a design can be optimized independently so as to yield global optimum when recomposed.

We demonstrate the utility of a minimum-energy function to capture high level compositional properties of circuits. The use of this minimum-energy function yields practical insight into ways of improving the overall energy-delay efficiency of circuits.

Categories and Subject Descriptors

B.6 [**Harware**]: Logic Design; B.6.3 [**Logic Design**]: Design Aids—*optimization*

General Terms

Theory

Keywords

Energy-delay optimization, transistor sizing

1. INTRODUCTION

The metric Et^2, where E is the energy and t is the delay of the computation, has been proposed as an efficiency metric for VLSI computation [1]. It has been argued that, due to its voltage independence, the Et^2 metric is superior to other efficiency metrics such as E or Et [2]. In this paper we show that the Et^n metric for the energy-delay efficiency index $n \geq 0$ characterizes *any* feasible trade-off, not only the trade-off through voltage scaling, between the energy and the delay of a computation. For example, any problem of minimizing the energy of a system for a given target delay can be restated as minimizing Et^n for a certain n.

There are many reasons we wish to study this more general metric over Et^2, despite the voltage independence of Et^2 over a wide range. First, we know that the applicability of Et^2 is unfortunately not perfect; it is sometimes better to use Et^n with $n \neq 2$ as the metric when the design performance target would make the Et^2-optimal circuit operate outside the practical range of supply voltages. Second, it is feasible to have a globally Et^2-optimal system with components optimized for Et^n with $n \neq 2$, as suggested by Theorems 2 and 8.

In general, for a VLSI computation implementing a *given* algorithm, the faster the computation the more energy it consumes. This observation points to the existence of a trade-off between the *good*ness of one property (delay) versus the *bad*ness of the other (consumed energy). The goal of our efficiency metric is to quantify such a trade-off. Certainly, there are computations that are both slower (*bad*) and consume more energy (*bad*) than some base case; however, those computations are not interesting implementations and will not be considered. Moreover, there are computations that are both faster (*good*) and more energy efficient (*good*) than some base case (for example the implementation of a transistor network in a newer technology). Again, these cases are of no interest to us since one will always prefer the *good-good* case and there is no trade-off. Later in this paper we formalize the notion of trade-off and the design parameters it applies to.

A VLSI computation can be made slower or faster in several ways: at high-level by using a different architecture, and at low-level by choosing a different supply voltage or different device parameters of the transistor network. In general, any of these choices amounts to trading delay for energy and vice-versa. For example, by operating a circuit at a higher supply voltage its delay decreases, while its energy consumption increases. Conversely, by operating the same circuit at a lower supply voltage its delay increases, while its energy consumption decreases. Thus, voltage scaling is one way to trade delay for energy and vice-versa.

With an efficiency metric at hand, one can define the corresponding optimization problem as of finding a set of parameters in the available parameter space that optimizes the given metric. The parameter space is defined by the freedoms available to the designer. For example, if the op-

GLSVLSI'02, April 18-19, 2002, New York, New York, USA.

erating voltage of the design can be varied, then voltage is part of the parameter space. On the other hand, if operating voltage is fixed, then it is not part of the parameter space. Through this paper, when we are referring to the efficiency metric, we are implicitly referring to the corresponding optimization problem as well.

In Section 2 and Section 3, we define the efficiency metric in two seemingly different ways. In Section 4, we show how these definitions indeed lead to the same trade-off between energy and delay. In Section 5, we quantify the *good*ness of parallel and sequential VLSI computations using the efficiency metric. As an example, we look at the energy-delay efficiency of circuits optimized through transistor sizing. We bound the energy and delay of the optimized circuits and we give necessary and sufficient conditions under which these bounds are reached. We also give necessary and sufficient conditions under which subcomponents of a design can be optimized independently so as to yield global optimum when recomposed. In Section 6, we sum up our results.

Most of the proofs have been omitted due to space limitations and due to the desire to emphasize the practical—as opposed to the mathematical—side of the results. The omitted proofs can be found in [5]. Furthermore, we have omitted the application of the concepts of minimum-energy function (to be defined later) and energy-delay efficiency index to other types of optimizations, besides transistor sizing, such as voltage scaling, branch prediction and performance optimization through buffer insertion. Some of those results can be found in [3] [4] [5].

2. THE ET^N EFFICIENCY METRIC

As mentioned in the introduction, we are interested in defining an efficiency metric over a set of design parameters, parameters that create a trade-off between energy and delay. More precisely, if we define two functions: one for energy $\mathcal{E}(\star) > 0$, and one for delay $\mathcal{T}(\star) > 0$, we are interested in studying them on the domain D that has the property that if $v, v + dv \in D$, $dv \neq 0$ then

$$\Big(\mathcal{E}(v+dv) - \mathcal{E}(v)\Big)\Big(\mathcal{T}(v+dv) - \mathcal{T}(v)\Big) < 0.$$

In other words, we are interested in the domain where evaluating $\mathcal{E}$ and $\mathcal{T}$ for a point $v + dv$ different than v results in increasing $\mathcal{E}$ while decreasing $\mathcal{T}$ or vice-versa. Specifically, we are not interested in domains where

$$\Big(\mathcal{E}(v+dv) - \mathcal{E}(v)\Big)\Big(\mathcal{T}(v+dv) - \mathcal{T}(v)\Big) \geq 0;$$

since then there is no trade-off and the optimization becomes trivial.

We do not require D to be continuous. It is important that our functions are general enough to be definable on noncontinuous domains. This allows us to use them to reason about noncontinuous parameter spaces like different architectural implementations of a given algorithm or different decompositions of a high level circuit specification. For example, if we want to evaluate the architectural trade-off between adders, the union of each different adder architecture (ripple-carry, carry-lookahead, carry-save, etc) can form the domain D.

With the previous clarifications about D in mind, the first form we propose for an efficiency metric combines the energy consumed by the computation, and the delay (cycle time or latency) of the computation, in the form

$$\Theta_n(v) : D \to R_+, \Theta_n(v) = \mathcal{E}(v)\mathcal{T}(v)^n, \;\; n \geq 0.$$

When the domain D of variable v is clear or irrelevant, we will omit explicitly using v in Θ_n, $\mathcal{E}$ and $\mathcal{T}$. Furthermore, when we will use the value of $\mathcal{E}$ or $\mathcal{T}$ evaluated in a specific point v_0 that follows from the context, we will use E and t as a shorthand for $\mathcal{E}(v_0)$ and $\mathcal{T}(v_0)$, respectively.

Intuitively, the metric $\Theta_n = \mathcal{E}\mathcal{T}^n$ implies that a 1% improvement in speed is worth roughly an n% increase in energy consumption. If one values the computation delay without regard to the consumed energy, the energy-delay efficiency index is $n = \infty$. Conversely, if one values the energy consumed by the computation without regard to the computation delay, the energy-delay efficiency index is $n = 0$. All in all, the metric $\Theta_n = \mathcal{E}\mathcal{T}^n$ quantifies—through the single parameter n—the entire range of feasible preferences in the trade-off between energy and delay.

It has been argued in [1] that Et^2—or with our notation Θ_2 ($n = 2$)—is independent, in first approximation, of the supply voltage. In other words, for $v = (..., V, ...) \in D$ (v includes the supply voltage V), $\Theta_2(..., V_1, ...) = \Theta_2(..., V_2, ...)$ $\forall\ V_1, V_2$. Practically this means that away from velocity saturation and threshold voltages, energy and delay can be freely exchanged through supply-voltage adjustment (within the feasible voltage range) while Θ_2 remains constant. We shall point out that if Θ_n is constant under the variation of a certain design parameter, E and t cannot be determined uniquely (as is the case with voltage scaling).

3. A MINIMUM-ENERGY FUNCTION

We can further refine the energy function $\mathcal{E}(\star)$ and the delay function $\mathcal{T}(\star)$ by defining two implicit functions: energy function of delay and delay function of energy. More precisely, we introduce a single-variable antimonotonic function by defining a *minimum-energy function* $E(t) : R_+ \to R_+$ that describes the minimum energy required for a system to run at a given t. Similarly, we introduce a single-variable antimonotonic function by defining a *minimum-delay function* $t(E) : R_+ \to R_+$ that describes the minimum delay of a system that consumes energy E. Through these two functions, we have abstracted away the original domain D of $\mathcal{E}(\star)$ and $\mathcal{T}(\star)$; however, it should be noted that the choice of D has an impact of the expressions of $E(t)$ and $t(E)$. Furthermore, these two functions depend at high level on the particular computation being implemented and at low level on the circuits and device parameters used.

We shall point out that both of these functions are well defined (in the mathematical sense). In particular, for the minimum-energy function even though there could be several ways to achieve a delay t, yielding several—possibly different—energy values E, by us picking the smallest of them we force this relation to take a unique value for each input t, and thus become a well defined function. A similar argument applies to the minimum-delay function.

The related optimization problem consists of finding these functions over parts or the entire domain of definition.

It can be shown that the minimum-energy function and minimum-delay function represent the same implicit relation between E and t. More precisely, the minimum-energy function and the minimum-delay function are the inverse of each other, i.e., $E \circ t = t \circ E = I$, where I is the identity function. It turns out that the minimum-energy function lends itself better to mathematical manipulation, given the fact that many compositional properties result in relations in terms of the delay t. For this reason, we will use only the minimum-energy function in our reasoning, but one should remember that the same argument can be stated in terms of the minimum-delay function.

Again, we do not require the domain of the minimum-energy function to be continuous. However, when we relate this function to the previously defined Et^n metric, we need—as it will be shown later—to be able to compute dE/dt. If the domain is continuous and the minimum-energy function is differentiable in any point of the domain dE/dt is well defined. However, if the domain is noncontinuous, we would still like to use the concept of dE/dt, even though $E(t)$ is not differentiable in the vicinity of t. To overcome this problem, we define dE/dt on a noncontinuous domain as the derivative of another differentiable function that interpolates E in the vicinity of t.

In the next subsection, we give an example of a minimum-energy function and of a minimum-delay function for a particular type of optimization.

3.1 A Minimum-Energy Function for Transistor Sizing

Transistor sizing is the optimization of a circuit that corresponds to choosing a set of transistor sizes that optimize a given metric. It has been shown in [3] [4] [5] that for optimal transistor sizing for Et^n, the consumed energy is

$$E_n \approx (1+n)E_0 \tag{1}$$

and the delay is

$$t_n \approx \left(1+\frac{1}{n}\right)t_\infty \tag{2}$$

where E_0 is the total switched wire capacitance of the circuit and t_∞ is the lower bound on the achievable delay of the circuit.

Even though Equations 1 and 2 transform to equality for only a very restricted class of circuits, they are in fact good approximations for a much wider class. We have checked the equations against the minimal Et^n obtained by applying an optimization algorithm (gradient descent) to two classes of circuits. In the first class, each circuit consisted of a ring of operators that were chosen at random with a uniform-squared distribution of parasitic capacitances; the number of transistors in series was also chosen according to such a distribution. We used real numbers for both parameters; we optimized the expression for Et^n where E was considered proportional to the total amount of gate and wire capacitance switched during computation and t was expressed using the τ model (Elmore delay). The range of parasitics was [1,100] in normalized units; the range of transistors in series was [1,6].

The results of the simulations for circuits consisting of a ring of 100 operators are summarized in Figure 1. (Simulations for rings of 10 and 1000 operators show similar results.) The figure shows the mean and standard deviation of the error in the estimates of Equations 1 and 2 for a range of different optimization indices ($n \in 1..10$ in Et^n). The estimates get more dependable for larger circuits, where the random variation in operators tends to average out over the cycle. Overall, the estimates are usually good to within five percent of the energy and within two percent of the delay values for the actual optimum Et^n.

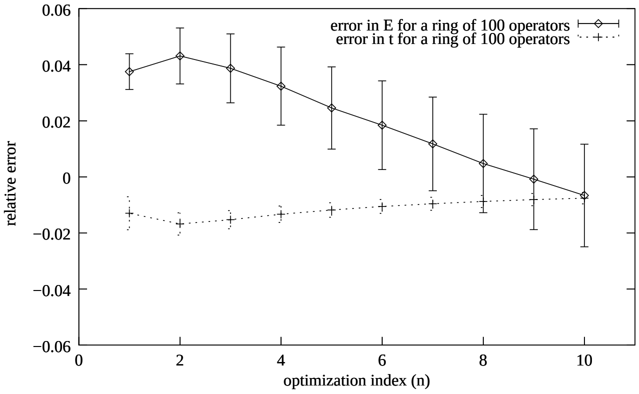

Figure 1: Results of simulating a ring of random gates and parasitics.

The second class of circuits consisted of a closed chain of connected rings of operators with parasitic capacitances, number of transistors in series and number of operators chosen the same way as in the previous experiment. Again, we find the estimates good to within eight percent of the energy and within five percent of the delay values for the actual optimum Et^n. All together, these results show that Equations 1 and 2 hold, with very good accuracy, over a wide range of parasitics, logic-gate types, circuit sizes, and circuit topologies.

Note that Equations 1 and 2 hold not only for a ring, but also for a “chain” of operators, as long as the parameters for the input of the chain are equal to the parameters for the output of the chain (since in this case the equations of 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 to circuit delays both in terms of latency and cycle time. Whenever we will use latency as the measure of delay, we make the assumption that the scrutinized component has its input “drive” equal to its output “drive” (i.e., no amplification). This is a reasonable assumption since most logic-gate chains are part of closed ring topologies.

Equations 1 and 2 establish two important properties of systems optimized for Et^n. First, the consumed energy E_n is independent, in first approximation, of the types of gates (NAND, NOR, etc) used by the circuit and is solely dependent on the optimization index n and the amount of wiring capacitance switched during computation. Second, the circuit speed t_n is independent of the parasitics and depends only on the optimization index n and the types of gates used. Furthermore, Equations 1 and 2 provide a good estimate of both energy and delay of an energy-delay efficient system. They allow an abstract view on transistor sizing and shift the design emphasis to the logical level of circuits.

If we rewrite Equations 1 and 2 with E a function of t—by eliminating n—we get the following function

$$E(t) = \frac{E_0 t}{t - t_\infty} \; . \tag{3}$$

It is easy to prove that Equation 3 satisfies the above definition of the minimum-energy function.

Similarly, one can express t as function of E and get the minimum-delay function for optimal transistor sizing

$$t(E) = \frac{t_\infty E}{E - E_0} \tag{4}$$

In the context of transistor sizing, we define the *asymptotic power* as

$$\hat{P} = \frac{E_0}{t_\infty}.$$

The energy and delay used in defining the asymptotic power are of course not simultaneously attainable; yet the asymptotic power is related to the actual circuit power. More precisely, the power consumption of a circuit optimized for Et^n through transistor sizing is

$$P = \frac{E}{t} = n\frac{E_0}{t_\infty} = n\hat{P}. \tag{5}$$

This relationship shows that the power consumption increases linearly with the optimization index n. In particular, the power consumption of a circuit optimized for Et is half that of the same circuit optimized for Et^2. Equation 5 also relates the optimization index n to the ratio between the actual power consumption and the asymptotic power of the circuit.

It should be noted that through the single parameter t—using the minimum-energy function—one can quantify the entire range of feasible preferences in the trade-off between energy and delay. The same holds for the single parameter E using the minimum-delay function. But this outcome was already achieved by the Et^n metric in Section 2. For this reason, we would like to know if these new functions are fundamentally different than our previous Et^n metric? More precisely, if a system were to be optimized using one of these functions or the Et^n metric, would that result in different values of the optimal E and t?

4. METRIC EQUIVALENCE

The answer to the previous question is given by the following

THEOREM 1. *Given an energy-delay optimization of a computation, the problem specified as "find $E_0 = \min E$ given t_0" is equivalent to "find the values of E and t that minimize Et^{n_0} for $n_0 = -\frac{t_0}{E_0}\frac{dE}{dt}(t_0)$"—when such a solution is unique. Similarly, the problem specified as "find $t_0 = \min t$ given E_0" is equivalent to "find the values of E and t that minimize Et^{n_0} for $n_0 = -\frac{t_0}{E_0}\frac{1}{\frac{dt}{dE}(E_0)}$"—when such a solution is unique.*

PROOF. We prove the equivalence of the two statements by showing that one implies the other and vice-versa. First, assume that we are solving "find $E_0 = \min E$ given t_0". Minimizing Et^n for the given t_0 implies—for any n—finding the minimum E given t_0—which in this case is E_0. Second, assume we are solving "find the values of E and t that minimize Et^{n_0} for $n_0 = -\frac{t_0}{E_0}\frac{dE}{dt}(t_0)$". With the help of the minimum-energy function, we can write Et^n as a single-variable function in t. This function is minimized—given that the minimum-energy function is antimonotonic—where

$$\begin{aligned} \frac{d(Et^n)}{dt} &= 0 \\ \Rightarrow \quad \frac{dE}{dt}{t'_0}^n + nE'_0{t'_0}^{n-1} &= 0 \\ \Rightarrow \quad \frac{t'_0}{E'_0}\frac{dE}{dt} + n &= 0, \end{aligned}$$

but for now

$$\begin{aligned} n = n_0 &= -\frac{t_0}{E_0}\frac{dE}{dt}(t_0) \\ \Rightarrow \quad \frac{t'_0}{E'_0}\frac{dE}{dt}(t'_0) &= \frac{t_0}{E_0}\frac{dE}{dt}(t_0). \end{aligned}$$

Thus, we found E'_0 and t'_0 that optimize Et^{n_0} such that

$$\frac{t'_0}{E'_0}\frac{dE}{dt}(t'_0) = \frac{t_0}{E_0}\frac{dE}{dt}(t_0).$$

Clearly, E_0 and t_0 are solutions of this equality. However, by hypothesis the solution to the minimization problem is unique $\Rightarrow t'_0 = t_0 \Rightarrow E'_0 = E(t'_0) = E(t_0) = E_0$ as well. Thus, when optimizing Et^n with $n = n_0$, if a unique solution exists, we find it to be the required E_0 and t_0. □

The uniqueness of the solution minimizing Et^{n_0} is important for non-ambiguously determining E_0 and t_0. It could be the case that there are several (E,t) pairs—including (E_0,t_0)—that minimize Et^{n_0}. In particular, the metric Et^{n_0} accepts infinitely many (E,t) pairs as solution if $E(t) = ct^{-k}$, $c > 0$, $k > 0$. If more than one solution exists, finding the solution pair (E_0,t_0) reduces to choosing from the set of solution pairs (E,t) the one that has $t = t_0$.

Theorem 1 tells us that, for a given system to be optimized in terms of both E and t, one can pose the optimization problem either in terms of an energy-delay efficiency index n, or a desired delay target t and obtain as result the same optimal values of E and t. This seemingly harmless result has the great benefit of allowing the application of the results developed for Et^n optimization [3] [4] [5] to other types of energy-delay optimizations—optimizations where either the target energy or the target delay are fixed. As a concrete example consider finding the optimal transistor sizes of a circuit so as to achieve delay t_0 for minimal energy. Given t_0, one can find the corresponding energy-delay optimization index n_0. With n_0 at hand—using the methodology developed in [5] for optimal Et^n transistor sizing—one can generate directly the transistor sizes that achieve delay t_0 for minimal energy consumption.

In the next section we apply the concept of metric equivalence to the parallel and sequential composition of circuits.

5. COMPOSITION

It is often the case, in practice, that one wishes to decompose the design of a complex system into a set of relatively independent subsystems, which then can be independently designed and implemented. If the optimization problem is defined globally using any of the parameters n, t or E, it is not immediately clear how subsystems of the original design should be optimized in terms of n, t or E, so as to achieve global minimum when the subsystems are recomposed.

The two major composition techniques used in VLSI design are parallel composition and sequential composition.

In the following, we show how the energy-delay efficiency metrics have to be applied to subcomponents so as to yield global minimum when recomposed in parallel or serially.

We will assume that each subsystem S_i has its own optimization index n_i (to be determined), and its own minimum-energy function $E_i(t)$.

5.1 Parallel Composition

Let us consider the parallel composition of m subsystems S_i. Let us assume a computation that runs in parallel all S_i's to completion before starting a new computation. We want to know at what t_i to run S_i or which n_i to optimize S_i for, so as to obtain the best E for a given t or to minimize Et^n for a given n, respectively.

Let us consider the first case, i.e. when we would like to find the minimal E for a given t. Knowing that S_i will complete after delay $t = \max_{1 \le i \le m}(t_i)$, there is no reason to run any of the subsystems faster than t, in other words $t_i = t$, $\forall i \in 1..m$. Under these circumstances, the energy consumption of S_i is $E_i(t)$ and the total energy consumption is $E = \sum_{i=1}^{m} E_i(t)$. Using Theorem 1 we can determine

$$n = -\Big(\sum_{i=1}^{m} \frac{dE_i(t)}{dt}\Big) \frac{t}{\sum_{i=1}^{m} E_i(t)}$$

and

$$n_i = -\frac{dE_i(t)}{dt} \frac{t}{E_i(t)}.$$

On the other hand, if n is given—noting again that $t_i = t$, $\forall i \in 1..m$—we can use the minimum-energy functions of subsystems S_i to write Et^n as a single-variable expression in t. If this single-variable function is continuous and differentiable, we find its minimum using the methods of mathematical analysis. If Et^n is not continuous—because the underlying domain is not continuous—one can still find the minimum by enumeration. Once the point of minimum is known, all other unknowns can be determined the same way as in the previous case.

In the following, we consider a relevant example of energy-delay optimization in the context of parallel composition.

5.1.1 Parallel Composition and Transistor Sizing

Consider a system consisting of the parallel composition of m subsystems S_i optimized through transistor sizing for energy-delay efficiency. Given the nature of the optimization parameter (transistor sizing), the minimum-energy function of S_i is given by Equation 3 as $E_i(t) = E_{0i}t/(t - t_{\infty i})$ where E_{0i} is the total switched wire capacitance of subsystem S_i and $t_{\infty i}$ is the lower bound on the achievable delay of subsystem S_i.

Lets consider the first instance of the optimization problem, namely when t is given and we want to find the minimum E that achieves this t. Based on the previous discussion on parallel composition, using the minimum-energy functions we can compute E directly as

$$E(t) = \sum_{i=1}^{m} E_i(t) = \sum_{i=1}^{m} \frac{E_{0i}t}{t - t_{\infty i}} \ . \tag{6}$$

Then, we can find

$$n = -t\frac{\sum_{i=1}^{m} \frac{dE_i(t)}{dt}}{\sum_{i=1}^{m} E_i(t)} = \frac{\sum_{i=1}^{m} \frac{E_{0i}t_{\infty i}}{(t-t_{\infty i})^2}}{\sum_{i=1}^{m} \frac{E_{0i}}{t-t_{\infty i}}}$$

and

$$n_i = -t\frac{\frac{dE_i(t)}{dt}}{E_i(t)} = \frac{t_{\infty i}}{t - t_{\infty i}} \ . \tag{7}$$

On the other hand, if we are given n and asked to find E and t that optimize Et^n, we use Equation 6 to write

$$Et^n = \Big(\sum_{i=1}^{m} \frac{E_{0i}t}{t - t_{\infty i}}\Big)t^n$$

and then minimize this single-variable function of t. It follows that, $\min Et^n \Rightarrow \frac{d(Et^n)}{dt} = 0 \Rightarrow t$ as the solution to a $2m-1$ order polynomial equation. With the computed t, E and E_i follow. Lastly, we would like to find what n_i to optimize S_i for, so as to yield global Et^n optimality. We can obtain this by computing n_i directly using Equation 7.

With the help of the minimum-energy function and the energy-delay efficiency index, we can infer several properties of parallel composition optimized through transistor sizing without the need to solve a $2m-1$ order polynomial equation to compute t. These properties are presented next.

THEOREM 2. *For the parallel composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$n = n_i \ \forall i \in 1..m \iff t_{\infty i} = t_{\infty j} \ \forall i, j \in 1..m.$$

Theorem 2 tells us that the parallel components of a system can be optimized independently for Et^n, yielding global optimum when recomposed, if and only if all $t_{\infty i}$'s are equal. Otherwise, even if one is globally optimizing for n, locally one needs to be able to optimize for $n_i \neq n$.

Consider, as an example, two subsystems S_1 and S_2 that have $t_{\infty 1} = 1$, $E_{01} = 2$, $t_{\infty 2} = 3$, and $E_{02} = 1$. If the parallel system composed of subsystems S_1 and S_2 is globally optimized for Et^2 then S_1 is locally optimized for Et while S_2 is locally optimized for Et^3.

THEOREM 3. *For the parallel composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$E = \frac{1}{n}\sum_{i=1}^{m} n_i E_i,$$

or equivalently

$$\sum_{i=1}^{m} (n_i - n)(1 + n_i)E_{0i} = 0$$

or equivalently

$$P = \frac{1}{n}\sum_{i=1}^{m} n_i^2 \hat{P}_i.$$

Theorem 3—in its first form—relates the total consumed energy, as defined by Equation 6, to the optimization indexes of the components and their respective energies, or—using the second form—it relates the optimization indexes to the minimal energies E_{0i} of the components. The last form of Theorem 3 relates the total power of the system to the optimization indexes and asymptotic powers of its components.

THEOREM 4. *For the parallel composition of m systems $S_i(E_{0i},\ t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$E \le (n+1)\sum_{i=1}^{m} E_{0i}$$

with equality if and only if all $t_{\infty i}$'s are equal.

PROOF. The optimal Et^n of this composed system is reached for E and t that satisfy

$$\frac{d(E(t)t^n)}{dt} = 0 \ ,$$

which is achieved when

$$(n+1)\sum_{i=1}^{m} \frac{E_{0i}}{t - t_{\infty i}} = t \sum_{i=1}^{m} \frac{E_{0i}}{(t - t_{\infty i})^2} \ . \qquad (8)$$

We may now invoke the Cauchy-Schwarz inequality

$$\left(\sum_{i=1}^{m} l_i r_i\right)^2 \le \left(\sum_{i=1}^{m} l_i^2\right)\left(\sum_{i=1}^{m} r_i^2\right),$$

where equality holds if and only if l_i/r_i has the same value for all i. If we substitute $l_i \leftarrow \frac{\sqrt{E_{0i}}}{t-t_{\infty i}}$ and $r_i \leftarrow \sqrt{E_{0i}}$, we get that

$$\left(\sum_{i=1}^{m} \frac{E_{0i}}{t - t_{\infty i}}\right)^2 \le \sum_{i=1}^{m} \frac{E_{0i}}{(t - t_{\infty i})^2} \sum_{i=1}^{m} E_{0i} \qquad (9)$$

with equality if and only if all $t_{\infty i}$'s are equal. Using Equation 8, we replace $\sum \frac{E_{0i}}{(t-t_{\infty i})^2}$ with $\frac{(n+1)}{t} \sum \frac{E_{0i}}{t-t_{\infty i}}$ in Equation 9, and we get the following result:

$$\left(\sum_{i=1}^{m} \frac{E_{0i}}{t - t_{\infty i}}\right)^2 \le \frac{(n+1)}{t} \sum_{i=1}^{m} \frac{E_{0i}}{t - t_{\infty i}} \sum_{i=1}^{m} E_{0i} \ .$$

By Equation 6, then,

$$E(t) = t\sum_{i=1}^{m} \frac{E_{0i}}{t - t_{\infty i}} \le (n+1)\sum_{i=1}^{m} E_{0i} \ .$$

And therefore

$$E \le (n+1)\sum_{i=1}^{m} E_{0i} \ .$$

□

In Theorem 4, equality holds if and only if all $t_{\infty i}$'s are equal; in this situation, we also have that $E_i = (n+1)E_{0i}$. In practice, generally all bits within a datapath pipeline are identical and different datapath pipelines have similar structure, thus it could be assumed that—for most well designed circuits—the cycles formed by these bits have very similar (or identical) t_∞'s. So, we should expect that usually $E \approx (n+1)\sum E_{0i}$. The existence of some potentially faster cycles (due possibly to buffers or fast control) will not have a significant impact on the global speed and energy of the system.

Let us consider a numerical example to illustrate Theorem 4. If $n = 2$, $m = 2$, $t_{\infty 1} = 1$, $t_{\infty 2} = 1.2$ and $E_{01} = E_{02} = 10$ then $t = 1.70$ and $E = 58.37$ ($E = E_1 + E_2 = 24.31 + 34.06$). Notice that $(1+\frac{1}{n})t_{\infty 1} = 1.5$, $(1+\frac{1}{n})t_{\infty 2} = 1.8$, $(n+1)E_{01} = 30$ and $(n+1)E_{02} = 30$. Thus, the optimal running speed of the system is between $(1+\frac{1}{n})t_{\infty 1}$ and $(1+\frac{1}{n})t_{\infty 2}$ (as claimed by the next theorem). The way t is reached is by running the faster system S_1 slower than its own speed target $(1+\frac{1}{n})t_{\infty 1}$ — thus saving energy (from $(n+1)E_{01} = 30$ to $E_1 = 24.31$), and running the slower system S_2 faster than its own speed target $(1+\frac{1}{n})t_{\infty 2}$ — thus spending more energy (from $(n+1)E_{02} = 30$ to $E_2 = 34.06$). What Theorem 4 is saying is that the energy trade-off between the slow and the fast systems is done such that only part of the energy saved by slowing down S_1 is spent on speeding up S_2; i.e. $(n+1)E_{01} + (n+1)E_{02} = 60$ is always greater that $E = 58.37$.

THEOREM 5. *For the parallel composition of m systems $S_i(E_{0i},\ t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$\max\Bigl(\max_{i\in 1..m} t_{\infty i},\ \Bigl(1+\frac{1}{n}\Bigr)\min_{i\in 1..m} t_{\infty i}\Bigr) \le t \le \Bigl(1+\frac{1}{n}\Bigr)\max_{i\in 1..m} t_{\infty i}.$$

with equality if and only if all $t_{\infty i}$'s are equal.

In Theorem 5, equality holds if and only if all $t_{\infty i}$'s are equal; in this situation, we also have that $t = (1+\frac{1}{n})t_\infty$. Theorem 5 bounds the optimal running speed of a circuit between its scaled $(1+\frac{1}{n})\times$ slowest cycle ($\min_{i\in 1..m} t_{\infty i}$) and fastest cycle ($\max_{i\in 1..m} t_{\infty i}$). If those cycles are close to each other—as is the case in a balanced design—both bounds on t are tight. If $n \to \infty$ then $\max_{i\in 1..m} t_{\infty i} \le t \le \max_{i\in 1..m} t_{\infty i}$ $\Rightarrow t = \max_{i\in 1..m} t_{\infty i}$, i.e. the delay of a circuit optimized for speed only is limited by the delay of its critical cycle; an expected result for speed-only optimization.

Based on Theorems 4 and 5, we can find an upper bound on the minimum Et^n, as suggested by the following

THEOREM 6. *For the parallel composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$\min(Et^n) \le (n+1)\Bigl(\sum_{i=1}^{m} E_{0i}\Bigr)\left(\Bigl(1+\frac{1}{n}\Bigr)\max_{i\in 1..m} t_{\infty i}\right)^n$$

with equality if and only if all $t_{\infty i}$'s are equal.

As mentioned earlier, in a well designed system, the $t_{\infty i}$'s are close to each other; thus, the upper bound given by Theorem 6 is tight and can be used as a good approximation of the actual minimal Et^n.

5.2 Sequential composition

Let us now consider the sequential composition of m subsystems S_i. Let us assume a sequential computation that runs S_1 to completion, then S_2 to completion, all the way to the completion of S_m; we assume the delay between the end of S_i and the start of S_{i+1} to be negligible. Again, we want to know at what t_i to run S_i or which n_i to optimize S_i for, so as to obtain the best E for a given t or to minimize Et^n for a given n, respectively.

THEOREM 7. *For the sequential composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for minimum energy E given a delay t or for Et^n, then*

$$\frac{dE_i(t_i)}{dt_i} = \frac{dE_j(t_j)}{dt_j}\ \forall i, j \in 1..m.$$

Theorem 7 is a very general result, it holds for any energy function $E(t)$ (as defined earlier) and any optimization index n. It extends to the more general case of sequential composition where each subsystem S_i is used repetitively with probability p_i.

Using Theorem 7 one can determine t_i, E_i and E. If n is not given, it can be determined from

$$n = -\frac{dE_i(t_i)}{dt_i} \frac{\sum_{i=1}^{m} t_i}{\sum_{i=1}^{m} E_i(t_i)},$$

while

$$n_i = -\frac{dE_i(t_i)}{dt_i} \frac{t_i}{E_i(t_i)}.$$

In the following, we consider a relevant example of energy-delay optimization in the context of sequential composition.

5.2.1 Sequential Composition and Transistor Sizing

Consider a system consisting of the sequential composition of m subsystems S_i optimized through transistor sizing for energy-delay efficiency.

THEOREM 8. *For the sequential composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$n = n_i \ \forall i \in 1..m \iff \hat{P}_i = \hat{P}_j \ \forall i,j \in 1..m$$

Theorem 8 is the equivalent, for sequential composition, of Theorem 2. Theorem 8 tells us that the sequential components of a system can be optimized independently for Et^n, yielding global optimum when recomposed, if and only if all $\hat{P}_i$'s are equal. Otherwise, even if one is globally optimizing for n, locally one needs to be able to optimize for $n_i \neq n$.

THEOREM 9. *For the sequential composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n, then*

$$P = \frac{n_i^2}{n} \hat{P}_i \ \forall i \in 1..m.$$

Theorem 9 is the equivalent, for sequential composition, of the third form of Theorem 3. Theorem 9 relates the total consumed power of the system to the optimization indexes and asymptotic powers of its components.

THEOREM 10. *For the sequential composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$E \leq (n+1) \sum_{i=1}^{m} E_{0i}$$

with equality if and only if all $\hat{P}_i$'s are equal.

Theorem 10 is the equivalent, for sequential composition, of Theorem 4. In Theorem 10, equality holds if and only if all $\hat{P}_i$'s are equal; in this situation, we also have that $E_i = (n+1)E_{0i}$. Given that Theorems 4 and 10 have the same form, it follows that for any parallel-sequential composition of circuits a property of the same form holds.

THEOREM 11. *For the sequential composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$t \leq \left(1 + \frac{1}{n}\right) \sum_{i=1}^{m} t_{\infty i}$$

with equality if and only if all $\hat{P}_i$'s are equal.

Theorem 11 is the equivalent, for sequential composition, of Theorem 5. In Theorem 11, equality holds if and only if all $\hat{P}_i$'s are equal; in this situation, we also have that $t_i = (1 + \frac{1}{n})t_{\infty i}$. The same way as Theorems 4 and 10 give an upper bound on the energy E for a parallel-sequential composition, Theorems 5 and 11 give an upper bound on the delay t of the same composition.

THEOREM 12. *For the sequential composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$\frac{P_i}{\sqrt{\hat{P}_i}} = \frac{P_j}{\sqrt{\hat{P}_j}} \ \forall i,j \in 1..m.$$

Theorem 12 tells us that when optimizing through transistor sizing, circuits composed sequentially should be designed so as to make their power usage proportional to the square-root of their *asymptotic power.*

The achievable upper bound through transistor sizing of a sequential composition follows from Theorem 10 and 11 and is given by the next

THEOREM 13. *For the sequential composition of m systems $S_i(E_{0i}, t_{\infty i})$, if the composed system is optimized for Et^n through transistor sizing, then*

$$\min(Et^n) \leq (n+1)\left(\sum_{i=1}^{m} E_{0i}\right)\left(\left(1 + \frac{1}{n}\right)\left(\sum_{i=1}^{m} t_{\infty i}\right)\right)^n$$

with equality if and only if all $\hat{P}_i$'s are equal.

Theorem 13 is the equivalent, for sequential composition, of Theorem 6. The given upper bound is, in practice, a tight bound and due to the flatness of the Et^n metric around the optimum it is a good approximation of the absolute minimum.

While it is rather obvious what it means to have all $t_{\infty i}$'s equal for parallel composition, it is not immediately clear what all $\hat{P}_i$'s equal imply. Consider two pipeline stages S_1 and S_2 composed sequentially, and assume that S_1 operates on N_1 bits while S_2 operates on N_2 bits. Further assume—for simplicity—that, per bit, the minimal energies and the minimal delays are the same for both pipeline stages, respectively. In other words, $E_{01} = N_1 E_0$, $t_{\infty 1} = t_\infty$ and $E_{02} = N_2 E_0$, $t_{\infty 2} = t_\infty$. This assumption is reasonable for pipelines with comparable per-bit-complexity and similar latency—so as to operate in the same clock domain. When we compute the asymptotic powers of S_1 and S_2 we get that $\hat{P}_1 = N_1 \frac{E_0}{t_\infty}$ and $\hat{P}_2 = N_2 \frac{E_0}{t_\infty}$. For these two values to be equal—as required for equality in Theorems 10, 11 and 13—we need to have $N_1 = N_2$, i.e. the number of bits each pipeline operates on should be the same. This suggests that the bounds are tighter for pipeline chains that average out more evenly the number of bits operated on in each individual stage.

Theorem 6 together with Theorem 13 provide an upper bound to the energy-delay efficiency of any parallel-sequential composition of circuits. Furthermore, they suggest a practical way to improve the energy-efficiency of these circuits by reducing the E_{0i}'s and $t_{\infty i}$'s. Transistor sizing is not able to change the $t_{\infty i}$'s or the E_{0i}'s, since they depend on other variables than transistor sizes—such as circuit micro-architecture, supply voltage, and fabrication technology. Thus, improving these other factors will ultimately impact the efficiency of the final design. In particular, it should be noted that E_{0i} depends on the wiring of system S_i; thus, compact hand layout or good layout tools can make a difference on the energy-efficiency of circuits. Similarly, $t_{\infty i}$'s can be directly improved by a proper choice of transistor netlist topology.

6. CONCLUSIONS

In this paper we introduced an energy-delay efficiency metric that captures *any* trade-off between the energy and the delay of the computation. We have presented two—seemingly different—ways to capture this trade-off and we have shown that these two forms ultimately yield the same circuit solution.

We applied this new concept to the parallel and sequential composition of circuits in general and in particular to circuits optimized through transistor sizing. We gave necessary and sufficient conditions under which subcomponents of a design can be optimized independently so as to yield global optimum when recomposed. We bounded the delay and energy of the optimized circuit and we gave necessary and sufficient conditions under which these bounds are reached. When applied to transistor sizing, we found that circuits composed sequentially should be designed so as to make their power usage proportional to the square-root of their asymptotic power. Many of the results inferred for parallel and sequential composition apply directly to the more general parallel-sequential composition of circuits.

We have demonstrated the utility of the minimum-energy function and its capacity to capture high level compositional properties of circuits. The use of the minimum-energy function gave us practical insight into ways to improve the overall energy-delay efficiency of the studied design.

7. ACKNOWLEDGMENTS

We wish to thank the members of the Asynchronous VLSI Group at Caltech for many stimulating discussions: Mika Nyström, Catherine Wong, and Karl Papadantonakis.

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

8. REFERENCES

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

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

[3] Paul I. Pénzes and Alain 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.

[4] Alain Martin, Mika Nyström, Paul I. Pénzes. *ET^2: A Metric for Time and Energy Efficiency of Computation.* Power-Aware Computing, Kluwer Academic/Plenum Publishers, 2002

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

[6] José A. Tierno. *An Energy-Complexity Model for VLSI Computations*, Ph.D. Thesis, California Institute of Technology, 1995.

[7] Anantha P. Chandrakasan, Robert W. Brodersen *Low Power Digital CMOS Design* Kluwer Academic Publishers, 1995