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 [Hardware]: 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 $Et^n$-optimal system with components optimized for $Et^m$ 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 goodness of one property (delay) versus the badness 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-
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 $E(x) > 0$, and one for delay $T(x) > 0$, we are interested in studying them on the domain $D$ that has the property that if $v, v + dv \in D$, we have $E(v + dv) - E(v) < 0$. In other words, we are interested in the domain where evaluating $E$ and $T$ for a point $v + dv$ different than $v$ results in increasing $E$ while decreasing $T$ or vice-versa. Specifically, we are not interested in domains where $E(v + dv) - E(v) = 0$; 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

$$\Theta_n(v) : D \rightarrow R_+, \Theta_n(v) = E(v)T(v)^n, \quad n \geq 0.$$
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

$$E_n \approx (1 + n)E_0$$

and the delay is

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

where $E_0$ is the total switched wire capacitance of the circuit and $t_\infty$ 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 $\tau$ 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 $Et^n$.

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

$$E(t) = \frac{E_0t}{t - t_\infty}.$$  

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

\[ t(E) = \frac{t_m E}{E - E_0} \]  

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

\[ \dot{P} = E_0 \frac{E}{t_{\infty}}. \]

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 \( E^0 \) through transistor sizing is

\[ \dot{P} = E_0 \frac{E}{t_{\infty}} = n \dot{P}, \]  

This relationship shows that the power consumption increases linearly with the optimization index \( n \). In particular, the power consumption of a circuit optimized for \( E \) is half that of the same circuit optimized for \( E^0 \). 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 \( E^0 \) metric in Section 2. For this reason, we would like to know if new functions are fundamentally different from our previous \( E^0 \) metric? More precisely, if a system were to be optimized using one of these functions or the \( E^0 \) 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 \( E_0 = \min \dot{E} \) given \( t_0 \)” is equivalent to “find the values of \( E \) and \( t \) that minimize \( E^{n_0} \) for \( n_0 = -\frac{1}{\dot{E}_0} \frac{dE}{dt}(t_0) \)” 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 \( E^{n_0} \) for \( n_0 = -\frac{1}{\dot{E}_0} \frac{dE}{dt}(t_0) \)” 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 \( E_0 = \min \dot{E} \) given \( t_0 \)”. Minimizing \( E^0 \) 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 \( E^{n_0} \) for \( n_0 = -\frac{1}{\dot{E}_0} \frac{dE}{dt}(t_0) \). With the help of the minimum-energy function, we can write \( E^0 \) as a single-variable function in \( t \). This function is minimized—given that the minimum-energy function is antimonic—where

\[ \frac{d(E^0)}{dt} = 0 \]
\[ \Rightarrow \frac{dE}{dt} t_0^n + nE_0' t_0^{n-1} = 0 \]
\[ \Rightarrow \frac{t_0}{E_0} \frac{dE}{dt} + n = 0, \]

but for now

\[ n = n_0 = -\frac{t_0}{E_0} \frac{dE}{dt}(t_0) \]
\[ \Rightarrow \frac{t_0}{E_0} \frac{dE}{dt}(t_0) = \frac{t_0}{E_0} \frac{dE}{dt}(t_0). \]

Thus, we found \( E_0 \) and \( t_0 \) that optimize \( E^{n_0} \) 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 \( \Rightarrow t_0 = t \Rightarrow E_0 = E(t_0) = E(t_0) = E_0 \) as well. Thus, when optimizing \( E^{n_0} \) 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 \( E^{n_0} \) 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 \( E^{n_0} \). In particular, the metric \( E^{n_0} \) 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 \( E^{n_0} \) optimization [3] [4] [5] to other types of energy-delay optimizations—optimizations where either the target energy or the delay target 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 \( E^{n_0} \) 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 $E^n$ 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 $t = \max_{1 \leq i \leq m}(t_i)$, 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

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

and

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

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 $E^n$ 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 $E^n$ 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 $E_i(t) = E_{0i}t/(t-t_{oci})$ where $E_{0i}$ is the total switched wire capacitance of subsystem $S_i$ and $t_{oci}$ is the lower bound on the achievable delay of subsystem $S_i$.

Let’s 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

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

Then, we can find

$$n = -\left(\frac{\sum_{i=1}^m \frac{dE_i(t)}{dt}}{\sum_{i=1}^m E_i(t)}\right)\frac{t}{\sum_{i=1}^m E_i(t)} = -\sum_{i=1}^m \frac{E_{0i}t}{(t-t_{oci})^2}.$$  \hspace{1em} (7)

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

$$E^n = \left(\sum_{i=1}^m \frac{E_0i}{t-t_{oci}}\right)^n$$

doing this single-variable function of $t$. It follows that, $\min E^n \Rightarrow \frac{dE^n}{dt} = 0 \Rightarrow t$ 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 $E^n$ optimality. We can obtain this by computing $n_i$ directly using Equation 7.

With the help of the minimum-energy function and the energy-deay 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 $S_i(E_{0i}, t_{oci})$, if the composed system is optimized for $E^n$ through transistor sizing, then

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

Theorem 2 tells us that the parallel components of a system can be optimized independently for $E^n$, yielding global optimum when recomposed, if and only if all $t_{oci}$’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 $t_{oci1} = 1$, $E_{01} = 2$, $t_{oci2} = 3$, and $E_{02} = 1$. If the parallel system composed of subsystems $S_1$ and $S_2$ is globally optimized for $E^2$ then $S_1$ is locally optimized for $E$ while $S_2$ is locally optimized for $E^3$.

**Theorem 3.** For the parallel composition of $m$ systems $S_i(E_{0i}, t_{oci})$, if the composed system is optimized for $E^n$ 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

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

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 \( S_i(E_{o_i}, t_{o_i}) \), if the composed system is optimized for \( E^o \) through transistor sizing, then

\[
E \leq (n + 1) \sum_{i=1}^{m} E_{o_i}
\]

with equality if and only if all \( t_{o_i} \)'s are equal.

Proof. The optimal \( E^o \) of this composed system is reached for \( E \) and \( t \) that satisfy

\[
\frac{d(E(t)E^o)}{dt} = 0,
\]

which is achieved when

\[
(n + 1) \sum_{i=1}^{m} \frac{E_{o_i}}{t - t_{o_i}} = t \sum_{i=1}^{m} \frac{E_{o_i}}{(t - t_{o_i})^2}.
\]

We may now invoke the Cauchy-Schwarz inequality

\[
\left( \sum_{i=1}^{m} l_i r_i \right)^2 \leq \left( \sum_{i=1}^{m} l_i^2 \right) \left( \sum_{i=1}^{m} r_i^2 \right),
\]

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

\[
\left( \sum_{i=1}^{m} E_{o_i} \frac{1}{t - t_{o_i}} \right)^2 \leq \sum_{i=1}^{m} \frac{E_{o_i}}{t - t_{o_i}} \sum_{i=1}^{m} E_{o_i},
\]

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

\[
\left( \sum_{i=1}^{m} E_{o_i} \frac{1}{t - t_{o_i}} \right)^2 \leq (n + 1) \sum_{i=1}^{m} E_{o_i} \sum_{i=1}^{m} E_{o_i},
\]

By Equation 6, then,

\[
E(t) = t \sum_{i=1}^{m} \frac{E_{o_i}}{t - t_{o_i}} \leq (n + 1) \sum_{i=1}^{m} E_{o_i}.
\]

And therefore

\[
E \leq (n + 1) \sum_{i=1}^{m} E_{o_i}.
\]

\( \Box \)

In Theorem 4, equality holds if and only if all \( t_{o_i} \)'s are equal; in this situation, we also have that \( E_i = (n + 1)E_{o_i} \). 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) \( t_{o_i} \)'s. So, we should expect that usually \( E \approx (n + 1) \sum_{i=1}^{m} E_{o_i} \). 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 \( n = 2, m = 2, t_{o_1} = 1, t_{o_2} = 1.2 \) and \( E_{o_1} = E_{o_2} = 10 \) then \( t = 1.70 \) and \( E = 58.37 \) (\( E = E_1 + E_2 = 24.31 + 34.06 \)). Notice that \( (1 + \frac{1}{n}) t_{o_1} = 1.5, (1 + \frac{1}{m}) t_{o_2} = 1.8 \). \( (n + 1)E_{o_1} = 30 \) and \( (n + 1)E_{o_2} = 30 \). Thus, the optimal running speed of the system is between \( (1 + \frac{1}{n}) t_{o_1} \) and \( (1 + \frac{1}{m}) t_{o_2} \) (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 \( (1 + \frac{1}{n}) t_{o_1} \)—thus saving energy (from \( (n + 1)E_{o_1} = 30 \) to \( E_1 = 24.31 \)), and running the slower system \( S_2 \) faster than its own speed target \( (1 + \frac{1}{m}) t_{o_2} \)—thus spending more energy (from \( (n + 1)E_{o_2} = 30 \) 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. \( (n + 1)E_{o_1} + (n + 1)E_{o_2} = 60 \) is always greater that \( E = 58.37 \).

Theorem 5. For the parallel composition of \( m \) systems \( S_i(E_{o_i}, t_{o_i}) \), if the composed system is optimized for \( E^o \) through transistor sizing, then

\[
\max_{i \in 1..m} \left( \frac{1}{n} \right) \min_{i \in 1..m} \left( \frac{1}{n} \right) \leq t \leq \left( \frac{1}{n} \right) \max_{i \in 1..m} t_{o_i},
\]

with equality if and only if all \( t_{o_i} \)'s are equal.

In Theorem 5, equality holds if and only if all \( t_{o_i} \)'s are equal in this situation, we also have that \( t = (1 + \frac{1}{n}) t_{o_i} \). Theorem 5 bounds the optimal running speed of a circuit between its scaled \((1 + \frac{1}{n}) \times \) slowest cycle \( \min_{i \in 1..m} t_{o_i} \) and fastest cycle \( \max_{i \in 1..m} t_{o_i} \). 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 \( \max_{i \in 1..m} t_{o_i} \leq t \leq \min_{i \in 1..m} t_{o_i} \) \( \Rightarrow t = \max_{i \in 1..m} t_{o_i} \), 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 \( E^o \), as suggested by the following

Theorem 6. For the parallel composition of \( m \) systems \( S_i(E_{o_i}, t_{o_i}) \), if the composed system is optimized for \( E^o \) through transistor sizing, then

\[
\min(E^o) \leq (n + 1) \sum_{i=1}^{m} E_{o_i} \left( \left( \frac{1}{n} \right) \max_{i \in 1..m} t_{o_i} \right)^n
\]

with equality if and only if all \( t_{o_i} \)'s are equal.

As mentioned earlier, in a well designed system, the \( t_{o_i} \)'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 \( E^o \).

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 \( t \) to run \( S_i \) or which \( n \) to optimize \( S_i \) for, so as to obtain the best \( E \) for a given \( t \) or to minimize \( E^o \) for a given \( n \), respectively.

Theorem 7. For the sequential composition of \( m \) systems \( S_i(E_{o_i}, t_{o_i}) \), if the composed system is optimized for minimum energy \( E \) given a delay \( t \) or for \( E^o \), then

\[
\frac{dE_i(t_i)}{dt_i} = \frac{dE_j(t_j)}{dt_j} \quad \forall i, j \in 1..m.
\]
5.2.1 Sequential Composition and Transistor Sizing

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

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

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 $n$ subsystems $S_i$ optimized through transistor sizing for energy-delay efficiency.

**Theorem 8.** For the sequential composition of $n$ systems $S_i(E_0, t_{oci})$, if the composed system is optimized for $E_i^n$ through transistor sizing, then

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

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 $E_i^n$, 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 $n$ systems $S_i(E_0, t_{oci})$, if the composed system is optimized for $E_i^n$, then

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

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 $n$ systems $S_i(E_0, t_{oci})$, if the composed system is optimized for $E_i^n$ through transistor sizing, then

$$E \leq (n + 1) \sum_{i=1}^{m} E_0,$$

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 $E_i = (n + 1)E_0$. 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 $S_i(E_0, t_{oci})$, if the composed system is optimized for $E_i^n$ through transistor sizing, then

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

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 $t_i = (1 + \frac{1}{n})t_{oci}$. The same way as Theorems 4 and 10 give an upper bound on the energy $E_i$ 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 $S_i(E_0, t_{oci})$, if the composed system is optimized for $E_i^n$ through transistor sizing, then

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

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 $S_i(E_0, t_{oci})$, if the composed system is optimized for $E_i^n$ through transistor sizing, then

$$\min(E_i^n) \leq (n + 1) \left(\sum_{i=1}^{m} E_0 \left(1 + \frac{1}{n}\right) \sum_{i=1}^{m} t_{oci}\right)^n$$

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 $E_i^n$ metric around the optimum it is a good approximation of the absolute minimum.

While it is rather obvious what it means to have all $t_{oci}$’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, $E_0 = N_1E_0$, $t_{oci} = t_{oci}$ and $E_0 = N_2E_0$, $t_{oci} = t_{oci}$. This assumption is reasonable for pipelines with comparable per-bit-complexity and similar latency—so as to operate in the same clock mode. When we compute the asymptotic powers of $S_1$ and $S_2$ we get that $\hat{P}_i = N_1\hat{P}_i$ and $\hat{P}_i = N_2\hat{P}_i$. 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 $E_0$'s and $t_{ns}$'s. Transistor sizing is not able to change the $t_{ns}$'s or the $E_0$'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 $E_0$ 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, $t_{ns}$'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