Global and local properties of asynchronous circuits optimized for energy efficiency

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

Abstract

In this paper we explore global and local properties of asynchronous circuits sized for the energy efficiency metric $E \times t^2$. We develop a theory that enables an abstract view on transistor sizing. These results allow us to accurately estimate circuit performance and compare circuit design choices at logic gate level without going through the costly sizing process. We estimate that the improvement in energy efficiency due to sizing is $2\times$ to $3.5\times$ when compared to a design optimized for speed.

We study sequential composition of circuits and show that for a circuit optimized for $E \times t^n$, the relationship between energy consumption and computation delay of the components is independent of n. When applied to optimizations for $E \times t^2$ via voltage scaling, this relationship implies that the powers of the components must be equalized.

1 Introduction

The continuing decrease in feature size and the corresponding increase in chip density and operating frequency have made energy consumption a major concern in VLSI design. As a consequence, energy efficiency is becoming an important consideration in IC design.

In [1] it was shown that the right measure of energy efficiency of a computation is $E \times t^2$, where E is the energy consumed by the computation and t is the delay (cycle time or latency) of the computation. There are several levels at which a VLSI system can be optimized for $E \times t^2$: architecture, circuit or physical implementation. For the main part of this paper, we focus on the circuit level and study the impact of transistor sizing on a system optimized for $E \times t^2$. We develop a theory that allows an abstract view on transistor sizing. First, the optimum N to P transistor ratio of a logic gate is derived. Then, the relationship of energy to delay is found. We prove that the energy trade-off between a *slow* and a *fast* system computing in parallel is such that only part of the energy saved by slowing down the *fast* system is spent on speeding up the *slow* one. We bound the optimal delay of a circuit between its scaled ($\frac{n+1}{n}\times$) slowest delay and fastest delay. We further prove that the delay of a system optimized for $E \times t^n$ is close to the scaled ($\frac{n+1}{n}\times$) smallest delay of the component that consumes the most energy and that the overall energy consumption is close to the scaled ($(n+1)\times$) energy consumption of this component. Later in this paper, we consider sequential composition of circuits optimized for $E \times t^2$ and infer a general relationship between the energies and delays of the component circuits. When applied to voltage scaling this relationship shows that circuits composed sequentially should be designed so as to equalize their power usage. When applied to transistor sizing the same relationship shows that circuits composed sequentially should be designed so as to make their power usage proportional to the square-root of their *asymptotic power* (to be defined later). While our results are established for QDI asynchronous circuits, they can be applied to synchronous circuits as well.

The paper is organized as follows. Section 2 describes the $E \times t^n$ metric. Sections 3 and 4 present local, and global properties of circuits optimally sized for $E \times t^n$. Section 5 considers the sequential composition of circuits and applies some of the theory developed in section 4. Section 6 elaborates on the practical benefits of energy efficient sizing. Finally, section 7 sums up the main results. Many of the proofs have been omitted due to space limitations. They can be found in [4].

2 Preliminaries

We are looking for an optimization metric that combines energy E and delay t in a way that is independent of voltage. With such a metric δ at hand, if we desire a particular delay target t, we adjust the voltage to meet it, and a circuit optimized for δ would have the best E for that t. Likewise, we may choose an energy target E and get a good t instead. For CMOS, $\delta = E \times t^2$ is the best such metric [1]. Basically, in first approximation $E = CV^2$ and $t = \frac{k}{V}$; thus, Et^2 is roughly constant over a range of voltages. For the purpose of this work, we will generalize the optimization metric $E \times t^2$ to $E \times t^n$, where $n \in N$; n is called the optimization index. This will allow us to compare circuits optimized for an entire range of metrics. For $n = 0$ the optimization metric

is energy only, for $n = 1$ the optimization metric is the energy-delay product, for $n = 2$ the optimization metric is our $E \times t^2$, while for $n \to +\infty$ the optimization metric is speed only. In this paper, optimizing $E \times t^n$ is used as a synonym for minimizing $E \times t^n$.

We explore global and local properties of asynchronous circuits sized for $E \times t^n$, while abstracting away transistor sizing itself. These properties allow us to accurately estimate circuit performance and compare circuit design choices at gate network level without going through the costly sizing process.

Once a circuit has been designed down to a gate network, transistor sizes that ensure correct functionality and performance have to be chosen. The achievable improvement due to sizing is limited ($2\times$ to $3.5\times$ for our types of circuits); however, an improper choice of transistor sizes could significantly affect the efficiency of the design.

While the problem of transistor sizing for speed only ($n \to \infty$) is relatively well understood, sizing for the more general $E \times t^n$ metric is not. The main difficulty is that gate delays are not independent of wire parasitics and the nice abstraction that the size of a gate is the geometrical mean of its neighbors does not apply.

Interconnects add extra costs and constraints to the optimization problem and they are difficult to accurately predict before layout. For speed-only optimization ($n \to +\infty$), the wire capacitance could in theory be overcome by increasing transistor sizes where appropriate. As we will show later, the parasitic wire capacitance plays a major role in case of optimizing for our target function and cannot, in general, be overcome in a straightforward way.

We model a transistor as a perfect switch in series with a linear resistor. The gate, source, and drain capacitances are proportional to the transistor width w_i and the transistor resistance is inversely proportional to w_i. Thus, a transistor network is modeled by an equivalent RC network. Wire resistance is not taken into account. Gate delays are modeled by Elmore delay (*tau model*) while energy is considered proportional to the sum of the gate and wire capacitance switched during computation (the energy due to leakage and short-circuit currents is ignored).

Within this model, our target function $E \times t^n$ could be written as a function of transistor sizes w_i. This type of function belongs to a special class known as *posynomials*. A *posynomial problem* is the minimization of one posynomial while simultaneously satisfying a collection of upper bound constraints on other posynomials. With the substitution $w_i = e^{x_i}$, a posynomial can be transformed into a *convex function*, and thus a posynomial program is a special case of a *convex program*. A convex program has the special property that a local minimum must necessarily be a global minimum [5]. This observation is exploited by tools that attempt to solve the optimization problem numerically [6].

3 Local properties

In this section we present two important local properties of circuits sized for $E \times t^n$. It is interesting to note that these properties are independent of the optimization index n. This suggests that the $E \times t^n$ optimization has only a global impact on circuits, while the topology of individual logic gates is not affected. As a result, the same logic gate library could be used both for high-speed ($n \to +\infty$) circuits and energy-efficient ($n = 2$) circuits, provided that the library has enough drive range to accommodate the global circuit requirements.

3.1 Synchronization points

A logic gate network can be represented as a directed graph in which each logic gate has a corresponding vertex and each literal a corresponding edge. In such a graph, a *path* corresponds to the sequence of fired (switched) logic gates in a given execution. We call *private* section of a path a maximum length sub-path that is not part of any other path. Similarly, a *public* sections of a path is a sub-path shared among other paths. A *synchronization point* in a logic gate network is any pull-up or pull-down of a gate that synchronizes (effectively waits for) two or more inputs in a given execution. For example, both the pull-up and the pull-down of a Muller C-element constitute a synchronization point, while the pull-ups or the pull-down of a NOR gate do not constitute synchronization points. Finally, a *non-data-dependent* logic gate network is either data-less (control only) or it has the property that each data rail (within a channel) has equal probability of firing in a typical execution. A *non-data-dependent* logic gate network or a non-data-dependent execution of a *data-dependent* logic gate network (most common case) is often a good approximation of the general behavior of the circuit. In this context we can state the following:

Theorem 1 *In a non-data-dependent logic gate network optimally sized for $E \times t^n$, each synchronization point enabled to fire non-vacuously has its input signals arriving simultaneously or, for any early*

transition, each path that contains that early transition has all its private section transistors of minimum size.

If we define a *cycle* in a logic gate network as a closed path and a *normalized cycle* as the ratio between the length of the cycle and the amount of activity on it (number of *tokens*) then we can state the following:

Corollary All *normalized cycles* in a non-data-dependent logic gate network optimally sized for $E \times t^n$ are equal, unless the private sections of the shorter cycles are minimum size.

Theorem 1 suggests a practical way to achieve an $E \times t^n$ optimum: identify all synchronization points with unequal arrival times and slow down the fast path by shrinking the corresponding transistors. If a fast path cannot be further slowed down, all transistors on the private section of that path will be minimum size and the optimization on that path is complete.

3.2 Size of P-transistors relative to N-transistors

The following result shows that there exists a general relationship, when optimizing for $E \times t^n$, between the width of the N-transistors and the width of the P-transistors of the same logic gate. This result depends on the relative mobility μ - the ratio of hole mobility over electron mobility.

Theorem 2 *Consider any cycle of a logic gate network implementing a QDI circuit. Assume that each logic gate i on this cycle has $k_{ni} \in N^*$ N-transistors of width w_{ni} in series, and $k_{pi} \in N^*$ P-transistors of width w_{pi} in series. Under these circumstances, when optimized for $E \times t^n$:*

$$w_{pi} = w_{ni}\sqrt{\mu \frac{k_{pi}}{k_{ni}}}$$

It is important to note that in a QDI asynchronous circuit, in general both the rising transition and the falling transition of a logic gate are on the same - possibly critical cycle. Theorem 2 is a consequence of this property. Theorem 2 is a local relationship: the N-device to P-device ratio only depends on the topology of the gate (through k_{ni} and k_{pi}) and the relative mobility μ. It does not depend on the output fanout or output wire parasitic, neither on the global E or global t, and is also independent of the optimization index n.

One consequence of Theorem 2 is that the number of free variables in the search space for optimum sizing could be reduced roughly by half, by eliminating the free variables corresponding to either the N-transistors or the P-transistors. Theorem 2 also eases the way to any transistor sizing abstraction.

4 Global properties

4.1 Global properties of E and t under $E \times t^n$ sizing

Consider a circuit optimized for $E \times t^n$ by transistor sizing. We make two main claims in this context. First, the consumed energy is independent, in first approximation, of the types (NAND, NOR, C-element, etc) of gates 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 is independent of the parasitics and depends only on the optimization index and the types of gates used. These results allow an abstract view on transistor sizing and shift the design emphasis to the logical level of circuits.

Theorem 3 *For a circuit composed of a ring of inverters with equal output wire parasitics p, the optimization for $E \times t^n$ yields a total gate capacitance of $w = w_{ni} + w_{pi} = np$ per logic gate.*

Proof. If we write E and t as functions of w_{ni}, w_{pi} and p then $\min(E \times t^n)$ implies $\frac{\partial(Et^n)}{\partial w_{ni}} = \frac{\partial(Et^n)}{\partial w_{pi}} = 0$ $\Rightarrow$ $w_{ni} = \frac{np}{1+\sqrt{\mu}}$ and $w_{pi} = w_{ni}\sqrt{\mu} = \frac{np\sqrt{\mu}}{1+\sqrt{\mu}}$ $\Rightarrow$ $w = w_{ni} + w_{pi} = np$.□

Theorem 3 shows that the total gate capacitance of an operator is equal to the output wire capacitance times the optimization index. Thus, a circuit optimized for $E \times t^2$ will have, on average, transistors twice as big as the same circuit optimized for $E \times t$. Theorem 3 also suggests a strong dependence of transistor sizes on wire capacitance, a dependency that is in general ignored when sizing for speed only. For $E \times t^2$ optimization, wire capacitance plays a major role and needs to be dealt with explicitly.

One can notice that Theorem 3 holds not only for a ring topology, but also for a chain topology, given that the input drive of the chain is equal to the output drive of the chain (since in this case the E and t equations for a chain have the same form as the ones for a ring). This is an important observation, it makes our results for transistor sizing applicable

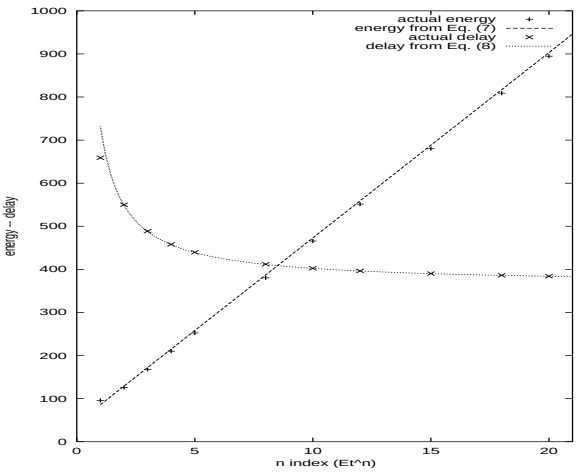

Figure 1: Actual and estimated energy E and delay t for optimal $E \times t^n$ as a function of n

to circuit delays both in terms of latency and cycle time. Whenever we will use latency as the measure of delay, we make the salient 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.

Under the conditions of Theorem 3 we can show the following:

Theorem 4 *The energy E and delay t of a circuit optimized for $E \times t^n$ are given by:*

$$E = (n+1)E_0 \tag{1}$$

$$t = \frac{n+1}{n}\tau_\infty \tag{2}$$

where E_0 is the total switched wire parasitic and τ_∞ is the smallest achievable delay.

Equations (1) and (2) can be generalized (in first approximation) to rings with arbitrary logic gates and arbitrary parasitics [4]. This is illustrated in Figure 1 for a generic circuit composed of different gate types with unequal output parasitics. Based on (1) the consumed energy is independent of the types of gates used by the circuit and is solely dependent on the optimization index and the amount of wiring capacitance switched during computation. On the other hand, based on (2) the circuit speed is independent of the parasitics and depends only on the optimization index and the types of gates used.

The generalization of equations (1) and (2) is based on the mathematical elaboration of the following observation. Consider a complex circuit with given wire parasitics p_i and gate sizes w_i that optimize $E \times t^n$. Assume that this system consumes energy E and operates with delay t. One can notice that if all parasitics are scaled with $\alpha \in R_+$ then αw_i are the gate sizes that optimize $E \times t^n$ of the new circuit. This is because E and t depend on w_i and p_i in such a way that α drops out of the partial derivative equations. It follows that the energy of the new circuit is $E' = \alpha E$ and the new delay is unchanged $t' = t$. Similarly, if all gate drives are scaled by $\beta \in R_+$ the actual gate sizes w_i that optimize $E \times t^n$ stay unchanged. It follows that the energy of the new circuit is unchanged $E' = E$ and the new delay is $t' = \beta t$. These observations show that the consumed energy is directly proportional to the switched parasitics through the scaling factor α, while the operation delay is independent of the wire parasitics. On the other hand, the operation delay is directly proportional to the type of gates through the scaling factor β while the consumed energy is independent of the type of gates.

If $n \to +\infty$ in (2) then $t = \tau_\infty$, an expected result in case of speed-only optimization. On the other hand, if $n = 0$ in (1) then $E = E_0$, i.e. transistors should be sized as small as possible for minimum energy consumption - another expected result. For energy efficiency ($n = 2$) the cycle time of the circuit should be chosen $\frac{3}{2}\tau_\infty$; while the minimum energy to achieve this delay will be $3E_0$ - the loss in speed is more than compensated by the energy savings.

Equations (1) and (2) provide an elegant way to analyze E and t independently at circuit level, while optimizing $E \times t^2$ (or in general $E \times t^n$). In particular, the speed of an $E \times t^n$ optimal system can be directly derived from the absolute speed τ_∞ of the same system - i.e. a well studied problem. Similarly, the energy consumption of the system can be directly derived from the total switched wire capacitance, wire capacitance that could be estimated for example using Rent's Rule.

4.2 Energy vs delay under optimal $E \times t^n$ sizing

In this section, we characterize the function $E(t)$, i.e. the *minimum* energy consumption of a circuit given the delay of operation t (at constant voltage). There is an absolute lower bound on the delay (cycle time or latency) at which a given circuit can operate, no matter how big its transistors are sized. We called this limit τ_∞ in (2) ($t \to \tau_\infty \Rightarrow w_i \to +\infty$). As a result, $E(t)$ has a vertical asymptote to $+\infty$ at

$t = \tau_\infty$ ($lim_{t \to \tau_\infty} E(t) = +\infty$). If the parasitics p_i of the circuit are considered fixed (which is the case in a wire-limited design), and there is no upper bound on the delay ($t \to +\infty \Rightarrow w_i \to 0$), the minimum energy with which the circuit can operate is $E_0 = \sum p_i$ - the total switched wire capacitance (assuming that the minimum width of transistors is 0). E_0 is the same term as in (1). As a result, $E(t)$ has a horizontal asymptote E_0 at $t = +\infty$ ($lim_{t \to +\infty} E(t) = E_0$). The simplest function that fulfills these two requirements is

$$E(t) : (\tau_\infty, +\infty) \to R_+, E(t) = \frac{E_0 t}{t - \tau_\infty} \qquad (3)$$

Interestingly, (3) can be verified using (1) and (2).

Consider a system composed of m subsystems S_i (E_i, τ_i) executing in parallel such that each subsystem has its *minimum* energy function of the form $\frac{E_i t}{t - \tau_i}$ (in particular, these subsystems can be chains or rings of arbitrary logic gates). If the subsystems are synchronized, then all delays affected by the synchronization stabilize to the same delay t (cycle time or latency depending of what is synchronized). As a consequence, the total energy function is:

$$E(t) : (\max_{i \in 1..m} \tau_i, +\infty) \to R_+, E(t) = t \sum_{i=1}^{m} \frac{E_i}{t - \tau_i} \qquad (4)$$

Note that (4) has the same asymptotic behavior as (3) and also that it is closed under addition, a property that (3) lacks.

Theorem 5 *For a system composed of m subsystems S_i (E_i, τ_i) as specified above, if the system is optimally sized for $E \times t^n$ then*

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

with equality iff all τ_i's are equal.

Proof. The optimal $E \times t^n$ of this composed system is reached for E and t that satisfy:

$$\frac{d(Et^n)}{dt} = 0 \;\Rightarrow\; (n+1) \sum_{i=1}^{m} \frac{E_i}{t - \tau_i} = t \sum_{i=1}^{m} \frac{E_i}{(t - \tau_i)^2} \qquad (5)$$

For the next step in the proof we use the Cauchy-Schwarz inequality:

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

which transforms to equality iff all $\frac{l_i}{r_i}$ terms are equal. With the substitutions $l_i = \frac{\sqrt{E_i}}{t - \tau_i}$ and $r_i = \sqrt{E_i}$ the previous inequality becomes:

$$\sum_{i=1}^{m} \frac{E_i}{(t - \tau_i)^2} \sum_{i=1}^{m} E_i \geq \left(\sum_{i=1}^{m} \frac{E_i}{t - \tau_i} \right)^2 \qquad (6)$$

with equality iff all τ_i's are equal. Using (6) to bound the right-hand-side of (5) we get:

$$(n+1) \sum_{i=1}^{m} \frac{E_i}{t - \tau_i} \geq \frac{t}{\sum_{i=1}^{m} E_i} \left(\sum_{i=1}^{m} \frac{E_i}{t - \tau_i} \right)^2$$

$$\Rightarrow\; (n+1) \sum_{i=1}^{m} E_i \geq t \sum_{i=1}^{m} \frac{E_i}{t - \tau_i} \overset{(4)}{=} E(t)$$

$$\Rightarrow\; (n+1) \sum_{i=1}^{m} E_i \geq E(t) \qquad (7)$$

with equality iff all τ_i's are equal. □

If for all i, $\tau_i = \tau_\infty$ we get $E(t) = (n+1) \sum_{i=1}^{m} E_i$ and $t = \frac{n+1}{n} \tau_\infty$ (a generalization of (1) and (2) to systems composed in parallel).

Let us consider a numerical example to illustrate (7). If $n = 2$, $m = 2$, $\tau_1 = 1$, $\tau_2 = 1.2$ and $E_1 = E_2 = 10$ then $t = 1.70$ and $E = 58.37$ ($E = E_{S_1} + E_{S_2} = 24.31 + 34.06$). Notice that $\frac{n+1}{n} \tau_1 = 1.5$, $\frac{n+1}{n} \tau_2 = 1.8$, $(n+1)E_1 = 30$ and $(n+1)E_2 = 30$. Thus, the optimal delay of the system is between $\frac{n+1}{n} \tau_1$ and $\frac{n+1}{n} \tau_2$ (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 ($\frac{n+1}{n} \tau_1$) - thus saving energy (from $(n+1)E_1 = 30$ to $E_{S_1} = 24.31$), and running the slower system S_2 faster than its own speed target ($\frac{n+1}{n} \tau_2$) - thus spending more energy (from $(n+1)E_2 = 30$ to $E_{S_2} = 34.06$). What (7) 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. $(n+1)E_1 + (n+1)E_2 = 60$ is always greater that $E = 58.37$.

Theorem 6 *For a system composed of m subsystems S_i (E_i, τ_i) as specified above, if there exists $j \in 1..m$ such that $|\frac{E_i}{E_j}| < \varepsilon$, $\forall\ \varepsilon > 0$ and $\exists\ \delta > 0$ such that $|\frac{n+1}{n} \tau_j - \tau_i| > \delta$, $\forall\ i \in 1..m, i \neq j$ then, for optimal $E \times t^n$: $t = \frac{n+1}{n} \tau_j$ and $E = (n+1)E_j$.*

The technicality of ε and δ in Theorem 6 is needed to avoid a division by zero for a case with no practical importance. More importantly, Theorem 6 tells us that the composed system runs close to the target delay $\frac{n+1}{n} \tau_j$ of the component (S_j) that consumes

the most energy E_j and that the overall energy consumption is close to $(n+1)E_j$. In practice, generally all bits within a datapath pipeline are identical and different datapath pipelines have similar structure, thus it could be assumed that the cycles formed by these bits have very similar (or identical) τ_∞'s. These τ_∞ cycles will generate a dominant term in the energy expression (since most of the energy is consumed in the datapath) and will bound the optimal cycle time of the system to $\frac{n+1}{n}\tau_\infty$ and its energy consumption to $(n+1)E_0$. The existence of some potentially faster cycles (due possibly to slack matching buffers or fast control) will not have a significant impact on the global speed and energy of the system. Theorem 6 allows us to use, under certain circumstances, the simpler formula (3) in our global performance analysis.

Theorem 7 *For the composed system considered above we have:*

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

Theorem 7 bounds the optimal delay of a circuit between its scaled ($\frac{n+1}{n}\times$) slowest delay ($\min_{i\in 1..m}\tau_i$) and fastest delay ($\max_{i\in 1..m}\tau_i$). If those delays are close to each other - as it is the case in a balanced design, both bounds on t are tight. Based on Theorem 6, any of the bounds for t in Theorem 7 could be reached if the energy consumption of the respective component is dominant. If $n \to +\infty$ then $\max_{i\in 1..m}\tau_i \le t \le \max_{i\in 1..m}\tau_i \Rightarrow t = \max_{i\in 1..m}\tau_i$, i.e. the speed of a circuit optimized for delay only is limited by the speed of its critical path; an expected result for speed-only optimization.

So far, it was considered that all components participate in the computation. In general, some parts of a circuit are only activated under certain conditions; for example, a branch adder will be used only when a branch instruction is being executed. Thus, some paths are not active on every computation cycle. The question is how shall these paths be sized to ensure global $E \times t^n$ optimality.

Our results could be extended to a related, but less general problem. Assume that a given path is activated every 1 out of $f \in N^*$ computation cycles. The corresponding component perceives the system timing t as ft. Thus, its contribution to the total energy is $\frac{E_0 ft}{ft-\tau_\infty} = \frac{E_0 t}{t-\frac{\tau_\infty}{f}}$. If we consider $\tau'_\infty = \frac{\tau_\infty}{f}$, then all properties inferred for paths activated every computation cycle are true also for paths activated only every 1 out of f computation cycle. In particular, (7) achieves equality when $\frac{\tau_\infty}{f}$ is equal to the other τ_is normalized by their usage frequency.

5 Application: sequential composition

This section reveals some remarkable global properties of sequential systems optimized for $E \times t^n$. Consider two programs A and B implemented by the circuits S_A and S_B, respectively. Assume a sequential computation that runs repetitively program A to completion and then program B to completion - the delay between the end of one program and the start of the other is assumed negligible. We would like to know at what t_A, t_B to run circuits S_A, S_B as to optimize the metric $E \times t^n$.

For the next theorem, assume the existence of a general energy function $E(t)$ - *minimum* energy consumed by a system given the system's operation delay t. This is a more general energy function than the one defined in section 4.2, since it is valid not only at circuit level but at any optimization level. Each system has its own $E(t)$, since the energy function will depend at high level on the particular computation being implemented and at low level on the circuits used.

Theorem 8 *For the sequential composition of two systems S_A and S_B, if the composite system is optimized for $E \times t^n$, then:*

$$\frac{dE_A}{dt_A} = \frac{dE_B}{dt_B}$$

independently of n.

Proof. The latency of the composed system is $t = t_A + t_B$, while its energy is $E = E_A(t_A) + E_B(t_A)$; thus, we minimize $f(t_A, t_B) = (E_A(t_A) + E_B(t_B))(t_A + t_B)^n$. f reaches its minimum where $\frac{\partial f}{\partial t_A} = \frac{\partial f}{\partial t_B} = 0 \Rightarrow \frac{dE_A(t_A)}{dt_A}(t_A + t_B) + n(E_A(t_A) + E_B(t_B)) = 0 \wedge \frac{dE_B(t_B)}{dt_B}(t_A + t_B) + n(E_A(t_A) + E_B(t_B)) = 0 \Rightarrow \frac{dE_A}{dt_A} = \frac{dE_B}{dt_B} = -\frac{n(E_A(t_A)+E_B(t_B))}{t_A+t_B}$. □

Theorem 8 is a very general result, it holds for any energy function $E(t)$ (as defined earlier) and any optimization index n. It extends to any number of sequential circuits S_i and to the more general case of sequential composition where each circuit S_i is used repetitively with probability p_i.

5.1 Sequential composition and voltage scaling

Assume the optimization parameter is V (voltage scaling), then $E(t) = \frac{\delta}{t^m}$. In [1] it is shown that

$E \times t^m$ is constant over a wide voltage range for $m = 2$. Let us define P_A and P_B to be the power consumed by component S_A, respectively S_B.

Property 1 *For the sequential composition of two systems S_A and S_B, if the composite system is optimized for $E \times t^n$ through voltage scaling, then $P_A = P_B$.*

Proof. Using Theorem 8 with $E_A(t_A) = \frac{\delta_A}{t_A^m}$ and $E_B(t_B) = \frac{\delta_B}{t_B^m}$ we get $\frac{\delta_A}{t_A^{(m+1)}} = \frac{\delta_B}{t_B^{(m+1)}} \Rightarrow \frac{E_A}{t_A} = \frac{E_B}{t_B}$ $\Rightarrow P_A = P_B$.□

For $m = 2$ this correlation was first suggested by Mika Nyström. Property 1 tells us that if the used optimization is voltage scaling, circuits composed sequentially and optimized for $E \times t^n$ should be designed as to equalize their power usage.

Property 2 *For the sequential composition of two systems S_A and S_B, if the composite system is optimized for $E \times t^n$ through voltage scaling, then if $n > m$ $\min Et^n$ is $\left(\sqrt[m+1]{\delta_A} + \sqrt[m+1]{\delta_B}\right)^{(n+1)} \left(\frac{\min t_A}{\sqrt[m+1]{\delta_A}}\right)^{(n-m)}$, if $n < m$ $\min Et^n$ is $\left(\sqrt[m+1]{\delta_A} + \sqrt[m+1]{\delta_B}\right)^{(n+1)} \left(\frac{\max t_A}{\sqrt[m+1]{\delta_A}}\right)^{(n-m)}$, and if $n = m$ then $\min Et^n$ is $\left(\sqrt[n+1]{\delta_A} + \sqrt[n+1]{\delta_B}\right)^{(n+1)}$ or $\sqrt[n+1]{\min(Et^n)} = \sqrt[n+1]{\delta_A} + \sqrt[n+1]{\delta_B}$.*

This property was first proved by Karl Papadantonakis for $n = m = 2$. Property 2 gives a lower bound on the achievable optimum for sequential composition using voltage scaling.

As a side note, for parallel composition of circuits S_A and S_B - as a consequence of Theorem 1 - $t_A = t_B$ for optimal $E \times t^n$. Assuming $E(t) = \frac{\delta}{t^m}$, $\min Et^n = \min(E_A(t_A) + E_B(t_A))t_A^n = (\delta_A + \delta_B)\min t_A^{(n-m)}$. If $n > m$ $\min Et^n$ is reached for $\min t_A$ (highest feasible voltage), while if $m > n$ $\min Et^n$ is reached for $\max t_A$ (lowest feasible voltage). If $n = m$ then $\min Et^n = \delta_A + \delta_B$, i.e. the lower bound on the achievable optimum for parallel composition using voltage scaling is the sum of the bounds of the individual components.

5.2 Sequential composition and transistor sizing

Assume the optimization parameters are transistor sizes, then $E(t) = \frac{E_0 t}{t - t_\infty}$. Let us define the *asymptotic power* of a circuit S (E_0, τ_∞) as $P_f = \frac{E_0}{\tau_\infty}$. If the circuit S is optimized for $E \times t^n$, its power consumption is $P = \frac{E}{t} \overset{(1)(2)}{=} n\frac{E_0}{\tau_\infty} = nP_f$. This relationship shows that the power consumption of a circuit optimized for $E \times t^n$ increases linearly with the optimization index n. In particular, the power consumption of a given circuit optimized for $E \times t$ is half that of the same circuit optimized for $E \times t^2$ (at constant voltage). Using this new definition, we can show the following:

Property 3 *For the sequential composition of two systems S_A and S_B, if the composite system is optimized for $E \times t^n$ through transistor sizing, then $\frac{P_A}{\sqrt{P_{fA}}} = \frac{P_B}{\sqrt{P_{fB}}}$.*

Proof. Using Theorem 8 with $E_A(t_A) = \frac{E_{0A} t_A}{t_A - \tau_A}$ and $E_B(t_B) = \frac{E_{0B} t_B}{t_B - \tau_B}$ we get $\frac{\sqrt{E_{0A}\tau_A}}{t_A - \tau_A} = \frac{\sqrt{E_{0B}\tau_B}}{t_B - \tau_B} \Rightarrow$ $\frac{E_A}{t_A}\frac{\sqrt{E_{0A}\tau_A}}{E_{0A}} = \frac{E_B}{t_B}\frac{\sqrt{E_{0B}\tau_B}}{E_{0B}} \Rightarrow \frac{P_A}{\sqrt{\frac{E_{0A}}{\tau_A}}} = \frac{P_B}{\sqrt{\frac{E_{0B}}{\tau_B}}} \Rightarrow$ $\frac{P_A}{\sqrt{P_{fA}}} = \frac{P_B}{\sqrt{P_{fB}}}$.□

Property 3 tells us that under transistor sizing optimization, circuits composed sequentially should be designed as to make their power usage proportional to the square-root of their *asymptotic power*. For the relevant case of $P_{fA} = P_{fB}$ we can state the following:

Property 4 *For the sequential composition of two system S_A and S_B, if circuits S_A and S_B have equal asymptotic powers (i.e. $\frac{E_{0A}}{\tau_A} = \frac{E_{0B}}{\tau_B}$) then for optimal $E \times t^n$ under transistor sizing, $t_A = \frac{n+1}{n}\tau_A$, $E_A = (n+1)E_{0A}$ and $t_B = \frac{n+1}{n}\tau_B$, $E_B = (n+1)E_{0B}$.*

Property 4 tells us that under the equal asymptotic power assumption, each circuit of a sequential composition of circuits could be optimized - under transistor sizing, independently and the composition of their local optimum results in a global optimum.

The achievable lower bound through transistor sizing of a sequential composition is given by the next two properties:

Property 5 *For the sequential composition of two systems S_A and S_B, if the composite system is optimized for $E \times t$ through transistor sizing, then $\sqrt{\min(Et)} = \sqrt{E_{0A}\tau_A} + \sqrt{E_{0B}\tau_B} + \sqrt{(E_{0A} + E_{0B})(\tau_A + \tau_B)}$.*

Property 6 *For the sequential composition of two systems S_A and S_B, if the composite system is optimized for $E \times t^n$ through transistor sizing, then $\min(Et^n) \le (n+1)(E_{0A} + E_{0B})\left((1 + \frac{1}{n})(\tau_A + \tau_B)\right)^n$ with equality iff $P_{fA} = P_{fB}$.*

While Property 5 gives an exact minimum for the special case of $n = 1$, Property 6 gives an upper

bound on this minimum. This upper bound is a tight bound and due to the flatness of the $E \times t^n$ metric around the optimum this bound is a good approximation of the absolute minimum.

6 Improvement in $E \times t^2$ metric due to transistor sizing

In theory, sizing for speed only requires $n \to +\infty$. In practice n is chosen big, but finite. This is because wire parasitics are not fully washed away, since that would result in impractically large transistors. While the miniMIPS microprocessor (an asynchronous version of a MIPS R3000) [3] was designed and sized for high speed, we estimate its optimization index n to be between 10 to 20. If the same design had been optimized for $E \times t^2$, the expected energy improvement would have been $\frac{11}{3}$ to $\frac{21}{3}$ while the speed slow-down between $\frac{7}{10}$ to $\frac{11}{15}$ of the original. This would have resulted in an overall $E \times t^2$ improvement of $2\times$ to $3.5\times$. We would expect this improvement everywhere on the chip except for the cache core cells which are sized based on different considerations. Based on the notion of asymptotic power, the power consumption would decrease $5\times$ to $10\times$.

7 Conclusions

In this paper we have explored global and local properties of asynchronous circuits sized for the metric $E \times t^n$. We have developed a theory that allows an abstract view on transistor sizing under this type of optimization. We have shown that the target cycle time of an optimized circuit should be $t = \frac{n+1}{n}\tau_\infty$, where τ_∞ is the smallest achievable cycle time of the component consuming the most energy, while the energy to achieve this delay should be $E = (n+1)E_0$ where E_0 is the total wire parasitic switched during computation. For the miniMIPS microprocessor, we estimated that the improvement in energy efficiency ($E \times t^2$ improvement) due to transistor sizing is $2\times$ to $3.5\times$ when compared to a high-speed design.

We have considered a sequential composition of circuits optimized for $E \times t^n$ at any level of design and inferred a general relationship between the energies and delays of the component circuits. When applied to voltage scaling this relationship shows that circuits composed sequentially should be designed as to equalize their power usage. When applied to transistor sizing the same relationship shows that circuits composed sequentially should be designed as to make their power usage proportional to the square-root of their *asymptotic power*.

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, and José Tierno from IBM, TJ Watson Research Center.

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.

References

[1] Alain J. Martin, "Towards an energy complexity of computation," Information Processing Letters 77, (2001) p181-187

[2] Alain J. Martin, Andrew Lines, Rajit Manohar, Mika Nyström, Paul Pénzes, 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.

[3] José Tierno, "An energy-complexity model for VLSI computations", Ph.D. Thesis, California Institute of Technology, 1995

[4] Paul I. Pénzes - Ph.D. Thesis (in preparation), Caltech

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

[6] 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