Energy Optimization of Pipelined Digital Systems Using Circuit Sizing and Supply Scaling

Hoang Q. Dao, *Member, IEEE*, Bart R. Zeydel, *Student Member, IEEE*, and Vojin G. Oklobdzija, *Fellow, IEEE*

***Abstract*—We present a systematic method for minimizing the energy of pipelined digital systems, through joint optimization of each pipeline stage and the system. A pipeline stage with a constant load can either be optimized for delay at a given input size, minimized for energy at a fixed delay, or have delay traded off for energy at a fixed input size. The results of these optimizations are combined to yield the design region for energy and delay. At the system level with a fixed throughput constraint, the sensitivities to input size and output load of all pipeline stages form the optimal energy criteria that provide a systematic method to minimize the total system energy. This method is applied to a media datapath, where we show up to 37% energy saving for a fixed performance. The minimal energy–delay curve of the system obtained through application of this method demonstrates similar characteristics as that of a single pipeline stage. With voltage scaling, the optimal solution displays a strong dependency between delay, energy, and supply voltage. The proper tradeoff between these entities makes a fundamental impact on efficient digital design.**

***Index Terms*—Circuit sizing, digital system, energy–delay characteristics, optimal criteria, optimization methodology, pipelined stage, supply voltage effects.**

I. Introduction

TECHNOLOGY scaling, aided by innovative circuit techniques, has produced dramatic improvements in circuit performance [1], [2]. However, with device sizes nearing the physical limits, undesirable effects such as saturated carrier velocity, leakage current, and gate current are starting to hamper performance improvement, requiring more energy to overcome the performance penalty. In addition, power increases at a rate similar to performance improvement, causing power consumption to become a design bottleneck. While circuit techniques may help reduce the energy wasted, power efficiency is fundamentally dependent on circuit size optimization and supply voltage selection.

Several approaches on energy optimization both at pipeline stage and system levels have been proposed. At the circuit level, they focus on obtaining the most energy-efficient design for individual circuit blocks [3]–[7]. However, as different circuit blocks often have conflicting constraints, their individual optimization does not guarantee minimal energy of the entire system.

At the system level, designers must decide how to configure different circuit blocks to deliver the desired performance and achieve minimal energy [8]. Constraints imposed on these two levels are often interdependent and at times conflicting. Zyuban and Strenski [9], [10] developed systematic criteria for the verification of design optimality providing insight into the optimization process. Unfortunately, this methodology has limitations in its applicability, making it difficult to use. It also has some unattainable assumptions in the derivation of optimal criteria that do not lead to an analytical formulation once corrected.

This paper introduces a systematic method that yields the globally minimal energy solution for circuit size optimization. The solution is achieved by exploring different optimization methods for a single pipeline stage. In addition, it formulates the dependency of the circuit energy to its input size (i.e., total size of transistors attached to the input of the circuit) and output load and applies that relationship to the optimization of pipelined systems. The method also provides insight into how supply voltage scaling affects the optimal conditions for which a circuit should be designed.

This paper is organized as follows. Section II provides a detailed overview and examines limitations of prior work that address energy optimization at circuit and system levels. Section III explains the energy and delay optimization of a single pipeline stage for a given load. Section IV examines the relationship of energy to input size and output load for a fixed delay target. Section V presents the optimization method for energy minimization of pipelined systems. Section VI applies the optimization method to a media datapath. Section VII expands the optimization to include supply voltage scaling, explaining the optimality of supply voltage for minimal energy and optimal performance. Section VIII concludes the paper.

II. Prior Work

Digital circuit optimization can be divided into two main levels: pipeline stage and system. At a pipeline stage, the objective is to minimize energy for the performance target under input size and output load constraints.

One approach is transistor-level optimization, such as timed logic synthesizer (TILOS) [3]. The method is simple, using linear delay modeling and direct estimation of area (or energy) based on transistor widths. The main limitation is its long runtime due to the independent adjustment of a large number of transistors, especially on complex designs. More recent methods [4]–[6] proposed optimization at the logic stage level in order to reduce the optimization complexity and allow for a quick calculation of the solution. The energy is minimized by proper distribution of delay to logic stages. The optimization

Manuscript received December 10, 2004; revised July 22, 2005. This work was supported in part by the Semiconductor Research Corporation (SRC) under Research Grant 931.001 and by California MICRO.

The authors are with the Advanced Computer Systems Engineering Laboratory (ACSEL), Electrical and Computer Engineering Department, University of California, Davis, CA 95616 USA (e-mail: hqdao@acsel-lab.com; zeydel@acsel-lab.com; vojin@acsel-lab.com).

Digital Object Identifier 10.1109/TVLSI.2005.863760

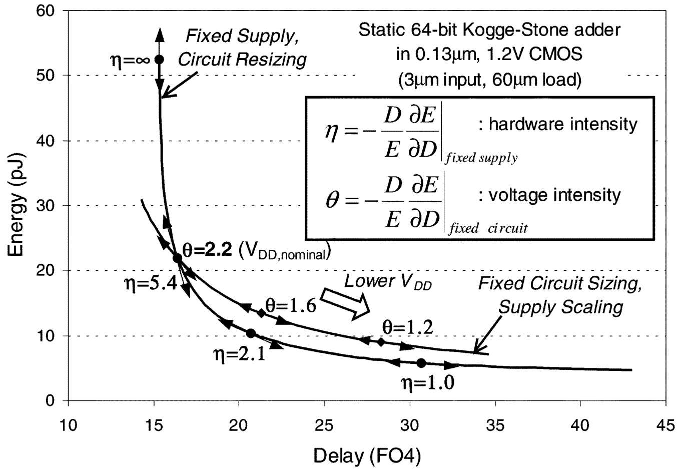


Fig. 1. Hardware intensity and voltage intensity.

could also be extended to include the effects of threshold and supply voltages [7]. However, the locally minimized solutions at individual stages do not necessarily guarantee an optimal solution for the entire system.

Optimization has also been applied at the system level. The unequal monotonic effects of supply voltage scaling on power and performance were studied by Chandrakasan *et al.* [8]. They exploited it to reduce the system energy by using a lower supply voltage while maintaining the same throughput. The study analyzed two architectural approaches: parallelism and pipelining.

Parallelism involves the multiplexing of duplicated circuits at certain blocks. The operating frequency of these blocks can be scaled down by the number of duplications for the same throughput. The frequency reduction allows for lower supply to be used to reduce power. The main disadvantage of parallelism is its high overhead of longer routing and the cost of multiplexers. The second architectural approach is pipelining. Under ideal pipelining conditions, the same throughput can be achieved with reduced length of the critical paths. Therefore, smaller circuits or lower supply can be used. Deep pipelining faces the limitations of increased overhead for extra clock-storage elements that need to be introduced in the system. In addition, the ideal throughput of both architectural approaches may not be achieved due to data dependencies. Thus, the operating frequency must be set higher in order to deliver the desired throughput.

Recently, Zyuban and Strenski [9], [10] proposed a high-level approach to optimize different circuit structures. They introduced the concepts of hardware intensity and voltage intensity that express the effect of sizing and supply scaling, respectively, on the energy–delay relationship. The definition and graphical presentation of these terms are shown in Fig. 1. Note that hardware intensity η defines the energy–delay sensitivity of the circuit due to circuit sizing alone at fixed supply voltage. Its value corresponds to the exponent of the traditional $E \cdot D^{\eta}$ product for the "optimal" design. On the other hand, voltage intensity, θ, refers to the energy–delay sensitivity of the circuit caused by supply scaling on a fixed circuit sizing. Both of these terms can be obtained for a fixed input size and output load of the circuit.

Using these terms, Zyuban and Strenski derived the general criteria for the optimal solution of a pipeline stage and a pipelined system (Appendixes A and B). These criteria were powerful in circuit optimization, as they appeared to reflect the intuitive energy–delay relationship between logic blocks as well as pipeline stages.

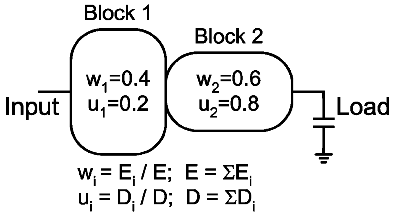


Fig. 2. Sample block sizing for a circuit stage.

The optimal criteria given by Zyuban and Strenski have two primary limitations: their hard-to-use coarse-tuning approach and the restricted assumption of energy and delay dependency among circuit blocks. First, the optimal criteria are difficult to apply and their application is mainly suited for the verification of design optimality. Given a design solution, this criteria can be used to determine if the design is optimal. If the design is not optimal, the criteria may suggest modifications to energy, delay, hardware intensity, or supply voltage.

For example, consider the pipeline stage consisting of two circuit blocks in Fig. 2. The energy and delay percentages of ith block relative to the entire pipeline stage are represented by w_i and u_i, respectively. Zyuban and Strenski have shown that the optimal solution must satisfy

$$\frac{w_1}{u_1}\eta_1 = \frac{w_2}{u_2}\eta_2 \Longleftrightarrow 8\eta_1 = 3\eta_2.$$

If the above condition cannot be met (such as $\eta_1 = \eta_2$ for the actual sizing of the above blocks), the analysis suggests reducing w_1/u_1 and increasing w_2/u_2. However, it is not immediately clear how to change each block. One can fix delay and change energy, fix energy and change delay, or change both. Furthermore, it begs the question of how much the delay or energy of each block should be changed so that the optimal solution is reached.

The other primary limitation of the optimal criteria is that their simple forms were derived assuming changes in a particular circuit block did not affect the energy and delay of neighboring ones. While this assumption can be justified in coarse tuning of circuits, it is generally not true for a pipeline stage. One example is a path of inverters where, according to the logical effort method [11], the delay of a gate depends on both input size and output load. A change in the size of any inverter in the path will affect the delay and energy of that inverter and the one driving it. In general, energy and delay dependencies exist among adjacent circuit blocks at their boundaries. Thus, they must be included in the derivation of the optimal solution.

For a pipelined system, Zyuban and Strenski made a similar assumption to enable the coarse tuning of pipeline stages, which is also not true in general. It will be shown in Section III-A that for a given delay with fixed input size and fixed output load, the minimal energy of a pipeline stage is a known value. To cause any change in energy while maintaining the same delay requires a change in either the input size or output load of the pipeline stage that affects the energy and delay of the neighboring stages that connect to it. Therefore, energy and delay dependencies exist between adjacent stages and must be added to the derivation.

The optimal criteria proposed by Zyuban and Strenski for pipeline stages and pipelined systems should be modified to account for these above dependencies. However, due to the nonanalytical form of these dependencies, their inclusion does not lead to an analytical solution.

III. Optimization of a Single Pipeline Stage

When designing a pipeline stage, the objective is either minimal delay [4], [5] or the least possible energy for a fixed delay [6], [7]. The result of either objective is critically dependent on the output load and the input size of the pipeline stage. It is obvious that, given an optimal design for a fixed input size, increasing the output load will either increase the delay or require larger energy to maintain the same delay. However, it is less clear what happens if the output load is fixed and the input size is varied. The effect of changing input size for a fixed load is analyzed next, based on the energy and delay behavior of circuits. Energy and delay quantities are computed using linear models for the energy and delay of gates.

The delay of a gate is approximately linear to its fan-out and is modeled accordingly in [12] and [13]. However, the logical effort model is more widely used due to its relatively technology-independent form [11]. This model can be analytically derived and its accuracy confirmed by simulation. The general form for logical effort delay model is as follows:

$$d = (f + p) \cdot \tau = (g \cdot h + p) \cdot \tau = \left(g \cdot \frac{C_{\text{out}}}{C_{\text{in}}} + p \right) \cdot \tau.$$

The delay d of the gate consists of three components: stage effort f, parasitic delay p, and technology-dependent unit delay τ. The stage effort accounts for the effect of the output load on delay while the parasitic delay represents the effects of the parasitic capacitance inside the gate. In addition, the stage effort is extended as the product of logical effort g (the relative driving capability of the gate), and the electrical effort h (the loading gain normalized to input capacitance $C_{\text{out}}/C_{\text{in}}$). The terms g, p, and τ are constant and can be obtained from simulation.

Similarly, the energy consumed in a gate can be linearly modeled according to its input size and output load as

$$E = (E_g \cdot C_{\text{out}} + E_p \cdot C_{\text{in}}) \cdot \alpha_s + T \cdot P_l \cdot W_{\text{in}} \cdot \alpha_l.$$

The terms α_s and α_l are constant, representing the switching activity and leakage factor of the active and leakage energy, respectively. The term C_{out} refers to the output load, and C_{in} and W_{in} denote the referenced input capacitance and width of the gate. The output load includes capacitances from loading gates and their interconnecting wires to the output node. Within a circuit block, the gate-to-gate wire can be treated as a capacitance because its resistance is mostly negligible. For long wiring, the wire resistance causes a quadratic impact on delay, which is typically minimized using inverter insertion [14]. In our analysis, we assume that such long wires have already gone through this process.

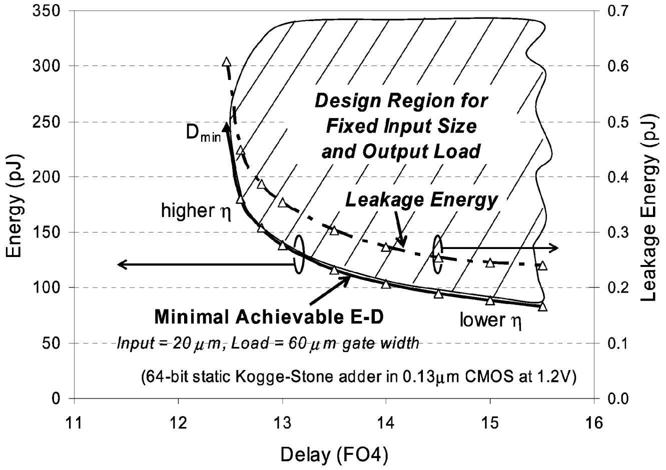


Fig. 3. Minimal achievable energy–delay using circuit sizing for a fixed input size and output load at 1.2-V supply.

The constant terms E_g and E_p represent energy coefficients for the loading capacitance and the internal parasitic capacitance, respectively [4]–[6]. The term $T \cdot P_l$ corresponds to the energy due to leakage current over the clock period, T. E_g and E_p are proportional to the square of supply voltage, while $P_l (= I_0 \cdot e^{-(V_{\text{th}} - \gamma V_{dd})/V_0} \cdot V_{dd})$ is proportional to its exponential. In the past, the leakage energy term $T \cdot P_l$ was negligible. However, in state-of-the-art technologies, all three energy terms are becoming comparable [1]. Nonetheless, for a given operating condition (i.e., fixed supply voltage and temperature), P_l is constant. Therefore, the leakage energy is linearly proportional to gate size and clock cycle. It will be shown later that the total leakage energy is primarily affected by circuit size variation where circuit optimization is desired. In addition, the effect of clock cycle becomes dominant only at very low performance where circuit size variation is insignificant and optimization is generally not needed.

For our analysis, the constant parameters for the energy and delay models were extracted from HSPICE simulation [15] in a 0.13-μm 1.2-V CMOS technology. Using these models, the energy–delay characteristics of three design scenarios for a pipeline stage are analyzed. For our case study, the pipeline stage is a 64-b static Kogge–Stone adder [16] with a 60 μm gate load at its output. The gate-to-gate wire capacitance is included and computed assuming a 4-μm bit pitch.

A. Energy Optimization for Fixed Input Size and Fixed Output Load

Energy optimization for a fixed input size and output load constraint is the most common design scenario for a pipeline stage. Given a fixed input size and fixed output load, the objective is to design the circuit for minimal energy. The design region for a fixed input, fixed output 64-b static Kogge–Stone adder using circuit sizing is shown in Fig. 3. Possible energy–delay points are shown in the area surrounded by the closed dotted curve. The points lying on the lower boundary of this space are most energy efficient for the given input and output constraints and represent the energy–delay curve of interest. Points on this curve can be determined by sizing the circuit for minimal energy under the given input size and output

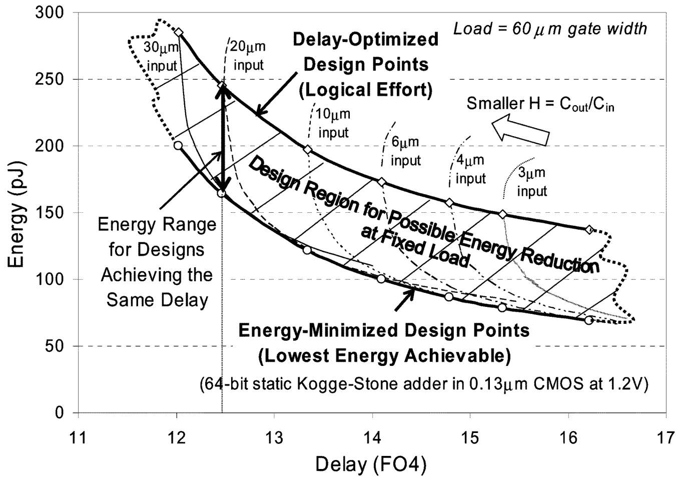


Fig. 4. Design region for possible energy reduction using circuit sizing for a fixed output load at 1.2-V supply.

load constraint for the desired delay target [7]. This curve is often used for energy–delay tradeoff, where a design point is selected based on its cost of energy (ΔE) for a given change in delay (ΔD).

Fig. 3 also shows the leakage energy corresponding to the minimal achievable energy–delay curve. The leakage curve is primarily affected by the large circuit size variation with respect to delay change. The increased leakage associated with a longer clock cycle is substantially less than the leakage reduction obtained from smaller transistor sizes. Therefore, leakage energy behaves as similarly as the active energy. Even when leakage energy becomes comparable to the active energy in future technologies or due to low switching activity of circuits, the characteristics of the minimal achievable energy–delay curve will remain unchanged and no algorithmic change for the optimization is needed.

B. *Delay Optimization for Fixed Output Load*

The minimal energy–delay curve for a fixed input size and fixed output load has an upper bound defined by the minimal delay point $D_{\min}$. It represents the smallest delay that the circuit can possibly achieve for a fixed input size and fixed output load.

Traditionally, this point was found by sizing all gates along the critical paths of a pipeline stage with a fixed fan-out delay. This view has been reinforced by the logical effort method [11], which states that the delay of a simple chain of gates is optimal when the stage effort (or fan-out delay) of each gate is equal. The same is true for multipath circuits when the off-path gates are linearly proportional in size to the corresponding on-path gates. Nonlinear factors, such as wire effect, minimal gate sizes, unequal numbers of gates along different paths, and parasitic delay difference of gates, will affect the optimality of the result. Nevertheless, only one minimal delay point exists for each input size and output load of the pipeline stage.

The energy–delay curve for these delay-optimized points sets the upper energy limit for the design region [5]. The upper solid gray curve in Fig. 4 shows these points obtained for various input sizes with a fixed output load. The larger the input size is, the smaller delay can be achieved, but at the cost of more energy. All design points above this curve are very inefficient because they use more energy for the same delay and require a larger input size.

C. *Energy Minimization for Fixed Output Load*

As energy consumption becomes more critical, circuit designers are forced to find the globally minimal energy design point for the required delay target. The solution requires the optimization of the pipeline stage for minimal energy while the delay is fixed.

The method proposed in [6] achieves minimal energy by redistributing delay in logic stages and varying input size, such that the changes of energy with respect to delay in all stages are equal. This is achieved at the cost of the increased input size (i.e., reduction of the overall circuit gain, $H = C_{\text{Load}}/C_{\text{Input}}$). The points obtained from energy minimization are shown by the lower solid black curve in Fig. 4. By definition, all other possible energy–delay points of the design must be above this curve.

It is important to observe from this graph that the minimal energy for an arbitrary delay target corresponds to a specific input size of the pipeline stage, which is larger than the delay-optimized one. At this input size, the energy–delay sensitivities among logic stages are balanced. Therefore, increasing the input size beyond this optimal value will result in more energy consumption. This characteristic of the design, with respect to energy, is distinctive compared to its delay characteristic where the delay is continuously improved by increasing input size.

In addition, the bounded area represents the design region for possible energy reduction at a fixed output load. The choice of design point is set by the delay target and the input size condition.

D. *Methodology for Pipeline Stage Optimizations*

In general, the solution to each of the above optimizations is a convex function of all gate sizes. This function can be solved for minimal energy or minimal delay under a set of constraints for input size and output load. This problem is well studied with known polynomial algorithms presented in [3] and [17]. In addition, instead of being optimized individually, the gates can be grouped into logic stages to significantly reduce complexity and improve rate of convergence [6].

IV. Energy Sensitivity to Input Size and Output Load at Fixed Delay

Pipeline stage optimization requires an understanding of the energy sensitivity of a pipeline stage to the variation of its input size and output load for a fixed delay. From the results of delay optimization and energy minimization for a fixed load (previous section), the upper and lower bound of the input size for the design can be obtained. The input size for the design region of a 64-b Kogge–Stone adder with 60-μm gate load in a 0.13-μm, 1.2-V CMOS technology is shown in Fig. 5. The lower bound for input size is found using delay optimization and the upper bound is found using energy minimization. All other energy points for the targeted delay have an input size within this range.

In the case of a pipelined system, the delay is fixed and is determined by the processor clock cycle requirement. Therefore,

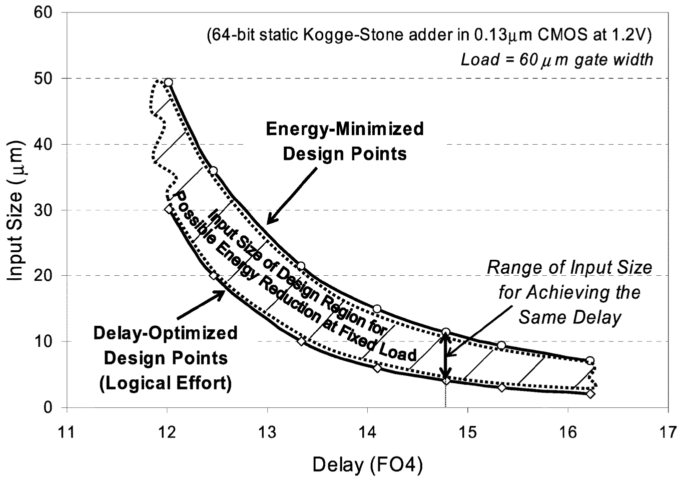


Fig. 5. Input range of design region for possible energy reduction for a fixed output load at 1.2-V supply.

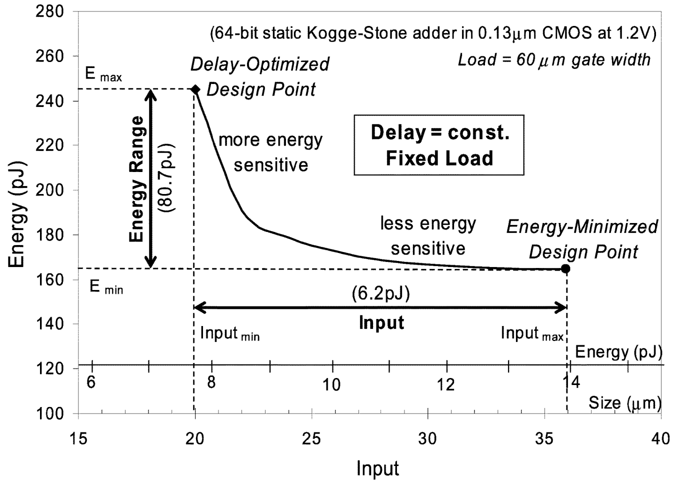


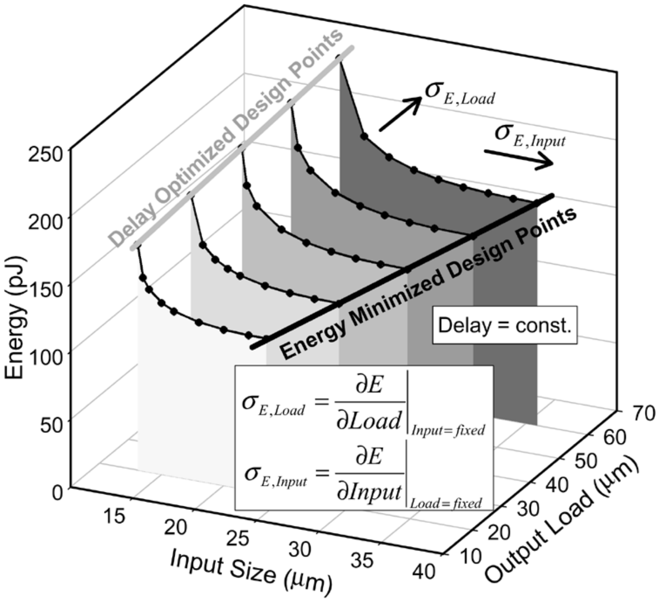
Fig. 6. Energy-input relationship for a fixed output load at a constant delay and 1.2-V supply.

the characteristics of the design region are reduced to the relationship of energy to input size and output load.

A. Relationship of Energy to Input Size at Fixed Delay

A typical energy-input relationship of a pipeline stage at a fixed delay is shown in Fig. 6. The maximal energy point corresponds to the smallest input size and is determined using the delay optimization as presented in Section III-B. The minimal energy point corresponds to the largest input size that is determined using the energy minimization shown in Section III-C. All points in between are computed by minimizing energy for fixed delay, input size, and output load, as discussed in Section III-A.

In general, this form of energy-input relationship is universal to any pipeline stage. It represents the minimal energy design points for the corresponding input range. The actual ranges of energy and input size vary for different pipeline stages and depend on circuit topology and design constraints. The largest energy sensitivity occurs at the smallest input size. This sensitivity demonstrates the high energy cost paid by delay-based optimization techniques (such as logical effort) to achieve the best delay for a given input size. On the other hand, at the maximal input size the energy is fairly insensitive to the input size. The side effect of this insensitivity is that the energy remains relatively flat over the upper half of the input range. This feature can be exploited to reduce the load that a pipeline stage imposes on the preceding stages that drive it.



(64-bit Kogge-Stone adder in 0.13μm CMOS at 1.2V supply)

Fig. 7. Energy relationship to input size and output load at a constant delay and 1.2-V supply.

B. Effect of Output Load on Energy

The energy of a pipeline stage is also affected by output load. The effect of the output load on the energy-input curve is shown in Fig. 7. For a larger load, the curve is shifted toward higher energy and larger input sizes. In addition, the energy dependence on the output load for the delay-optimized and energy-minimized points remains relatively linear. For the delay-optimized case, this linear behavior occurs because the input size and energy are linearly proportional to gate sizes, which are traceable to the output load. For the energy-optimized case, it is unclear why the energy-input relationship can still be linear despite the possible change in delay distribution. It is assumed that such change is a weak function of the output load, so the energy behavior remains linear. The range of input sizes is reduced at smaller load due to the nonlinear effects of wire capacitance, parasitic delays of gates, and secondary effects at minimal transistor sizes.

C. Energy Sensitivity Factors

As observed in Fig. 7, at a given performance the circuit energy is sensitive to both input size and output load. These sensitivities characterize the energy gradient of a pipeline stage and are necessary factors for system optimization.

Given a fixed load, the energy–input sensitivity will increase as the input decreases. This sensitivity is infinite at the delay-

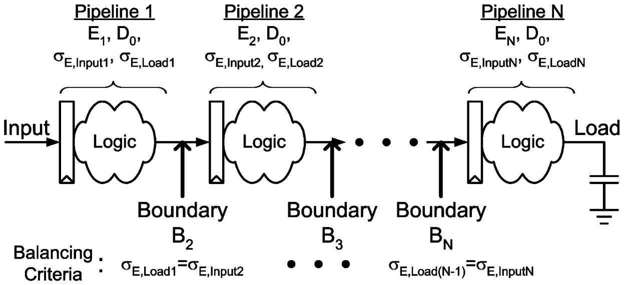

Pipeline 1
E_1, D_0,
$\sigma_{E,\mathrm{Input1}}$, $\sigma_{E,\mathrm{Load1}}$
Pipeline 2
E_2, D_0,
$\sigma_{E,\mathrm{Input2}}$, $\sigma_{E,\mathrm{Load2}}$
Pipeline N
E_N, D_0,
$\sigma_{E,\mathrm{InputN}}$, $\sigma_{E,\mathrm{LoadN}}$
Input
Logic
Logic
Logic
Load
Boundary B_2
Boundary B_3
Boundary B_N
Balancing Criteria :
$\sigma_{E,\mathrm{Load1}}=\sigma_{E,\mathrm{Input2}}$
$\sigma_{E,\mathrm{Load(N-1)}}=\sigma_{E,\mathrm{InputN}}$

Fig. 8. Block diagram of a simple pipelined digital system.

optimized point and zero at the energy-minimized point. We define the energy–input sensitivity as

$$\sigma_{E,\mathrm{Input}} = -\left.\frac{\partial E}{\partial \mathrm{Input}}\right|_{\mathrm{Load=fixed}}.$$

It represents the rate of energy change with respect to input variation at the current input size for a fixed output load. Note that the minus sign indicates the opposite behavior of energy to input size.

The output load also affects the energy of the pipeline stage. We define the energy-to-output sensitivity on the output load as

$$\sigma_{E,\mathrm{Load}} = \left.\frac{\partial E}{\partial \mathrm{Load}}\right|_{\mathrm{Input=fixed}}.$$

It refers to the rate of energy change with respect to load variation at the given output load for a fixed input size. For a chosen input size, the energy–load sensitivity will increase as the load increases.

These two energy sensitivities play important roles to the optimization of pipelined systems.

V. Pipelined Digital System Optimization

Pipelined systems are the most common digital systems. The diagram of a simple pipelined system is shown in Fig. 8. Each pipeline stage i consists of the logic circuits for the stage and the clock storage elements driving them. The boundary B_i between the $(i-1)$th and ith pipeline stages is defined at their interface. In other words, the boundary represents the end of $(i-1)$th stage and the beginning of ith stage. The load at the boundary B_i consists of the ith-stage input size and the interconnect wire between the two stages.

The constraints of the pipelined system are the system I/O, throughput, and energy. System designers will need to translate these requirements into the estimated input size and output load for each individual pipeline stage which are then given as implementation requirements to circuit designers. By trading the energy sensitivities to the input size and output load among pipeline stages, the minimal energy of the system is achieved.

A. Problem Definition

The energy minimization of a pipelined system typically begins with an initial implementation for a delay target D_0 (set by the system cycle time). The energy of the entire system needs to be minimized under this delay constraint.

The initial implementation is obtained by applying the following steps:

1) estimate an acceptable input size for each pipeline stage and compute its corresponding output load (i.e., the summation of output wire capacitance and the input sizes of next pipeline stages);
2) minimize energy of each stage for the above input size and output load, using the optimization method described in Section III-A.

Note that the delay-optimization and energy-minimization methods (Section III-B and III-C) should be used to verify if the initial choices for the input sizes are acceptable in each particular stage. For example, the delay-optimized method can determine if the input size is too small to achieve the delay target for the given load. If this is the case, either the initial input value must be increased or the output load must be reduced by lowering the input sizes of next stages. On the other hand, the energy-minimized method can determine if the initial size is too large. If so, its value can then be reduced to that provided by the energy-minimized method in order to avoid unnecessary loading to the preceding stage.

It is expected that the energy of the system resulting from the initial choice of input sizes and output loads will not be optimal. That is, energy sensitivities among different pipeline stages at some pipeline boundaries are not equal. The next step is to find the criteria that yield the minimal energy for the system.

B. Optimization Criteria for Simple Pipelined Systems

Energy can be improved if energy sensitivities to input size and output load are not balanced at an arbitrary pipeline boundary B_i. In general, the energy of a system is minimal if and only if the energy sensitivities are equal at each pipeline boundary. That is

$$\sigma_{E,\mathrm{Load\ P_{i-1}}} = \sigma_{E,\mathrm{Input\ P_i}}$$

for all boundaries B_i.

This fact can be proven via the two cases where energy sensitivities are not balanced.

Case 1: $(i-1)^{th}$ pipeline shows more energy sensitive to its output load than ith pipeline to its input size. That is, $\sigma_{E,\mathrm{Load\ P(i-1)}} > \sigma_{E,\mathrm{Input\ P(i)}}$.
Reducing the input size of the ith stage ($\Delta\mathrm{Input_i} < 0$) will result in less load to the $(i-1)^{th}$ stage and allow for total energy reduction. The mathematical proof is shown below.

$$\begin{aligned}\Delta E =& E_{\mathrm{new}} - E_{\mathrm{init}}\\ =&\Delta E_{\mathrm{Pipeline(i-1)}} + \Delta E_{\mathrm{Pipeline(i)}}\\ \approx&(\sigma_{E,\mathrm{Load\ P(i-1)}} - \sigma_{E,\mathrm{Input\ P(i)}})\Delta\mathrm{Input_i}\\ <&0.\end{aligned}$$

As the input size is reduced, $\sigma_{E,\mathrm{Load\ P(i-1)}}$ will decrease while $\sigma_{E,\mathrm{Input\ P(i)}}$ will increase. Their values will begin to approach each other. When they are equal $(\sigma_{E,\mathrm{Load\ P(i-1)}} = \sigma_{E,\mathrm{Input\ P(i)}})$, further reduction in energy is not possible, as is illustrated in the next case.

Case 2: $(i-1)$th pipeline is less energy sensitive to its output load than ith pipeline is to its input size. That is $\sigma_{E,\mathrm{Load\ P(i-1)}} < \sigma_{E,\mathrm{Input\ P(i)}}$.

Similarly and in the opposite manner, increasing the input size of the ith stage $(\Delta\mathrm{Input_i} > 0)$ will yield less energy as shown mathematically.

$$\begin{aligned}\Delta E =& E_{\mathrm{new}} - E_{\mathrm{init}}\\ =&\Delta E_{\mathrm{Pipeline(i-1)}} + \Delta E_{\mathrm{Pipeline(i)}}\\ \approx&(\sigma_{E,\mathrm{Load\ P(i-1)}} - \sigma_{E,\mathrm{Input\ P(i)}})\Delta\mathrm{Input_i}\\ <&0.\end{aligned}$$

In addition, $\sigma_{E,\mathrm{Load\ P(i-1)}}$ will approach $\sigma_{E,\mathrm{Input\ P(i)}}$ with the increasing input size. The energy will not reduce any more when these sensitivity terms are equal.

Combining the results of the above two cases, energy is minimal when the energy sensitivities are equal at the given boundary. These conditions can be extended to the general system, where feedback from multiple pipeline stages may occur.

C. Optimal Criteria for General Pipelined Systems

In a general pipelined system, a pipeline stage can be driven by several other pipeline stages via its different inputs. Since the delay from each input to the output must be the same, any change at one input will affect all the other inputs of the stage. Consequently, the energy of the stages driving these inputs will also change and a single energy sensitivity to input can be used.

On the other hand, a pipeline stage may drive several other stages that have independent input sizes. Therefore, the energy sensitivities caused by these loading stages can be accounted for separately.

The conditions for optimal energy are extended to include the above general pipeline stages. Using similar reasoning as in Section V-B, the optimal criteria for minimal energy must satisfy

$$\sum_k \sigma_{E,\mathrm{Load\ A_k}} = \sigma_{E,\mathrm{Input\ B}}$$

at the boundary of any arbitrary stage B where stage A_k drives its k^{th} input.

D. Optimization Algorithm

The optimization of the pipelined system is a recursive process where optimization switches between individual pipeline stages and the pipelined system. The outlined algorithm in Fig. 9 represents the process flow of the optimization.

The load of each pipeline is computed from the initial (arbitrarily) chosen input size of pipeline stages. Each pipeline stage is then minimized for energy for the chosen input size, output load, and delay target. With respect to input size and output load, its energy sensitivities can be computed. Next, the energy sensitivity criteria are applied to each pipeline boundary to verify the energy balance between pipeline stages. Should there be an energy imbalance at a boundary, the input sizes of pipeline stages attached to the boundary are adjusted toward improving the energy balance. The entire process is repeated until the energy is balanced at all pipeline boundaries.

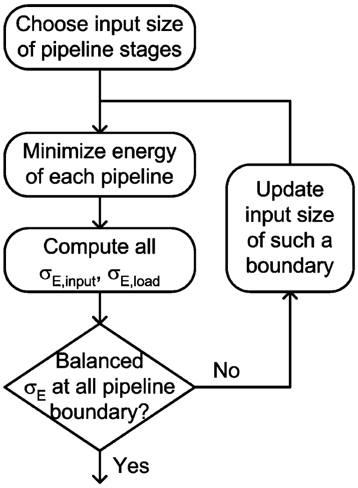

Choose input size of pipeline stages
Minimize energy of each pipeline
Compute all $\sigma_{E,\mathrm{input}}$, $\sigma_{E,\mathrm{load}}$
Balanced σ_E at all pipeline boundary?
No
Yes
Update input size of such a boundary

Fig. 9. Algorithm for energy optimization of pipelined digital systems.

Note that the total energy of the system is continuously improved after each iteration of the algorithm until it reaches the optimal solution. The effectiveness of the algorithm depends on two selections: the unbalanced pipelined boundary to be updated and the amount of input size variation at that boundary. The boundary selection will depend on the amount of energy sensitivity difference at this boundary and the energy weight of its pipelines. A brute-force approach using the most energy sensitivity difference can be used. More complex algorithms may be found in [18]. The input size update at the boundary can be estimated from the difference of energy sensitivities with respect to input size and output load at the boundary. The amount of input size variation at this boundary may not need to be exact when many boundaries have unbalanced sensitivities because adjustments in the latter can alter its balanced energy sensitivities. Notice that inexact input adjustment will not affect the energy improvement of the system. It only affects the converging time. One the other hand, input size adjustment should be gradually tightened when the energy sensitivities at boundaries reduce.

VI. Optimization of a Media Datapath

An example of a simplified datapath used in a media processor [19] was chosen to demonstrate the validity of the optimization criteria. The datapath allows operations on the operands of 8-, 16-, 32-, and 64-bit sizes, as required for processing different media data types. Fig. 10 shows the main building blocks of the selected datapath. The main components include two 16×16-b multipliers and one 64-b partitioned adder, separated by register pairs (X, Y), (U, V), (C, S) and output register Z. The multipliers assume signed operands and employ Radix-4 Booth encoding [20]. The adder is implemented using a configurable parallel-prefix structure for the carries and 4-bit conditional summation [21]. The registers are implemented with the conditional-precharged flip-flop [22].

For energy optimization, the pipeline stages of the datapath are defined as shown in Fig. 10. Note that some of the registers $(X, Y, U,$ and $V)$ are split into different pipeline stages so that each pipeline can be optimized separately. In addition, due

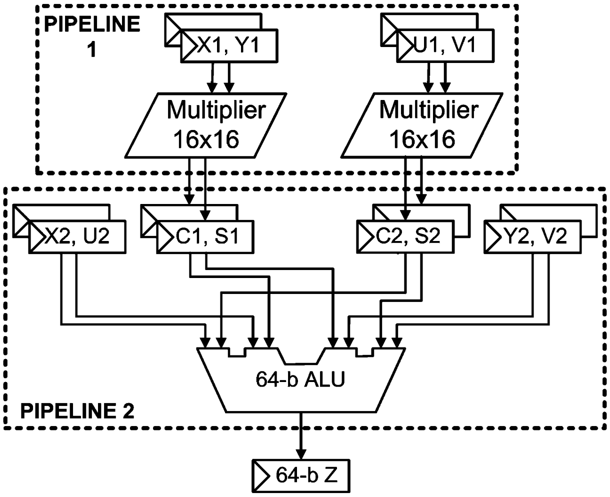


Fig. 10. Block diagram of the media processor datapath.

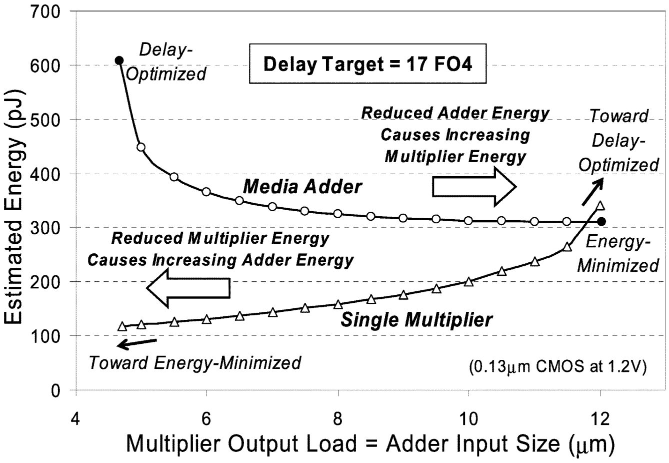


Fig. 11. Minimal energy versus boundary conditions for the media adder and the multiplier at 17FO4 delay and 1.2-V supply.

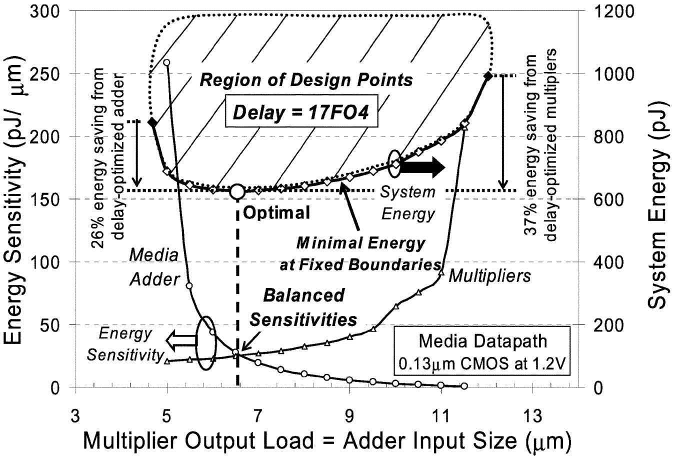


Fig. 12. Minimal energy solution for the media datapath at 17FO4 delay and 1.2-V supply.

to the symmetry of the system, the multipliers in pipeline stage 1 are identical. The maximal input size of the system is determined by the largest of the register pairs, $\{X1, X2\}$, $\{Y1, Y2\}$, $\{U1, U2\}$, and $\{V1, V2\}$. The system load is set by register Z.

The system is optimized for the following constraints in a 1.2-V 0.13-μm CMOS technology. The performance target is set at 17 FO4 delay. The system load (applied to the media adder) is fixed and equivalent to 60-μm gate width. The system input size set by registers X, Y, U, and V is no more than 30-μm gate width.

At the optimal solution, the following boundary equations must be met:

- $\mathrm{MAX}\{\mathrm{Input}\ X,\ \mathrm{Input}\ Y,\ \mathrm{Input}\ U,\ \mathrm{Input}\ V\} \leq 30\ \mu\mathrm{m}$;
- $\sigma_{E,\mathrm{Load\ (Pipeline\ 1)}} = \sigma_{E,\mathrm{Input\ (Pipeline\ 2)}}$.

The adder has a fixed load of 60-μm gate width (or system load). Its energy versus input size can be obtained using the optimization methods in Section III. The results are shown in Fig. 11. There exists a two-times difference between the maximal and minimal energies for the media adder, with a corresponding range of input size from 4.7 to 12 μm.

The multipliers are directly loaded with the input size of the adder. The maximal input size condition (summation of adder input size and multiplier input size) allows for the computation of the multiplier input size. The energy for each multiplier has a three-times range versus its output load (or the media adder input size) under the given delay constraint using the optimization method in Section III-A (Fig. 11). In addition, the exponential increase of energy at higher load indicates that the input size of the multiplier is pushed closer to its delay-optimized value.

The energy characteristics of the adder and the multipliers show opposite behavior. The energy sensitivity of the adder is higher at a small input size and lower at a larger one. Conversely, the energy sensitivity of the multipliers is lower at small output load (or small input size of the adder) and higher at large output load (or large input size of the adder). Therefore, a minimal energy solution exists within the input range of the adder.

The optimal input size of the adder may be found using the binary search algorithm over the input range of the adder. First, the midpoint of the input range is selected to compute the energy sensitivities of the multipliers and the media adder. Then, the search range is reduced to the half range where the energy sensitivity behavior at the midpoint is opposite to that of the corresponding end. The process is repeated on the new search range until the solution is reached (i.e., where the energy sensitivities match). This search reflects the gradually tightening approach of input adjustment toward the correct value, as suggested in Section V-D.

Fig. 12 shows the energy sensitivities of the multipliers and adder and the total energy of the system. The optimal energy occurs when the input sizes of the adder and the multipliers are set to 6.5 and 23.5 $\mu\mathrm{m}$, respectively. The minimal energy solution corresponds to the optimal criteria where the output energy sensitivity of the multipliers is equal to the input energy sensitivity of the adder.

Significant energy is saved compared to those at the edges, 26% and 37% when delay optimization is applied to the media adder and the multipliers respectively. It indicates that optimization focusing only on individual pipeline stages could actually lead to very high energy consumption of the whole system. In addition, more energy can be saved compared with other possible design points when input sizes are not correctly chosen as shown in the shaded region.

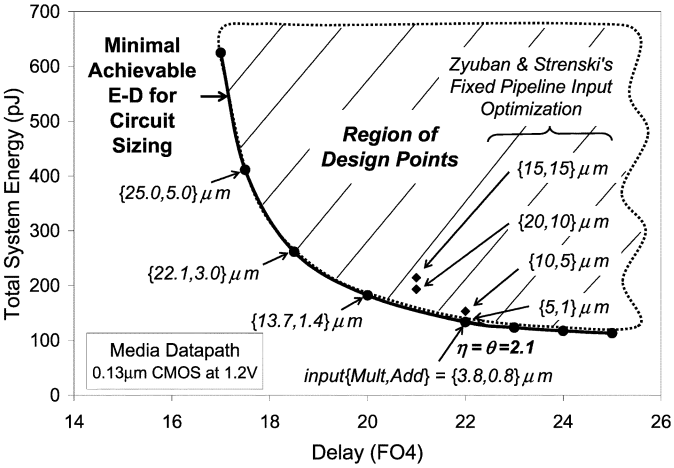


Fig. 13. Minimal achievable E-D for the media datapath using circuit sizing at 1.2-V supply.

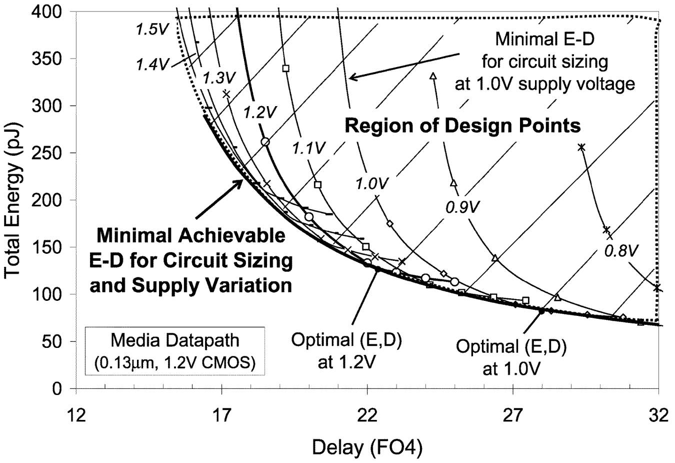


Fig. 14. Minimal achievable energy–delay for the media datapath using circuit sizing and supply variation.

Furthermore, due to the flat energy characteristics near the minimal energy solution, coarse tuning of the system energy is possible over a wide input range of the adder (5.5, 8.5 μm), where energy is within 5% of the optimum. Therefore, a less strict algorithm can be used to speed up the convergence process. Nonetheless, to guarantee the closeness of the result, our proposed optimal criteria must have the primary role during the optimization.

VII. Delay and Supply Scaling of a System

An important aspect of energy efficiency is the energy behavior of a system as its delay target changes. This behavior can be influenced by circuit sizing and supply scaling.

A. System Energy–Delay Characteristic Due to Circuit Sizing

Intuitively, improved performance of a system comes at the cost of more energy. This can be explained using the energy–delay characteristics of individual pipelined stages from Section III and proof of contradiction. The derivation is omitted for its triviality.

The minimal energy delay curve for the media datapath at 1.2-V supply voltage is shown as a solid line in Fig. 13. As the system delay increases, its energy is reduced in a similar manner as for a single pipeline stage. In addition, the maximum input size of the system is also smaller. This is possible because the delay increase can be traded for less energy by not only circuit resizing but also output load reduction by reducing the input size of loading pipeline stages.

The "optimal" solutions using the system criteria in [9] and [10] are also shown and represented by filled diamonds. They are obtained for the shown input sizes of the multipliers and adder, and correspond to the energy–delay points where the system criteria in [9] and [10] are satisfied. The system results show that the energy–delay curve obtained with our method represents the minimal achievable energy–delay points for the system. In addition, the results obtained from the criteria in [9] and [10] are not guaranteed to be optimal.

The above deduction assumes no significant change in system energy behavior. That is, less energy is achieved by increasing the delay. In reality, this energy–delay behavior can be altered in two ways. One way is the change of switching activities, which can occur when gate resizing causes different arrival times of input signals and may, therefore, cause multiple switching. Another influence to energy behavior is leakage power. At a larger delay, the circuit size variation is smaller but the leakage time is longer. The energy savings due to the decreased circuit size may eventually be offset by the increasing leakage time. These two effects are important only in the low-power region of design [23]. They determine the delay at which the absolute minimal energy occurs for the given supply voltage. This delay also sets the lower limit for low-power performance.

While the effect of extra circuit switching at low power does not change with technology scaling, the effect of leakage does [1]. Technology trends show an increasing percentage of leakage over the total energy due to the reduction of threshold voltage. Because of increased leakage, the absolute minimal energy is pushed toward high performance design, essentially reducing the useful range of system performance.

B. Effects of Supply Scaling

Supply voltage can also provide an important source of energy efficiency. Zyuban and Strenski [9], [10] generalized the effects of supply voltage on energy and performance of different circuits and represented them by voltage sensitivity terms, $E_V = (V/E) \cdot (\partial E/\partial V)$, $D_V = -(V/D) \cdot (\partial D/\partial V)$, and $\theta = E_V/D_V$, where the last was called voltage intensity. Assuming the terms E_V and D_V were equal for all gates and circuit stages, they derived the minimal-energy criteria for different types of pipeline structures that related the voltage scaling and circuit sizing (i.e., voltage intensity and hardware intensity, respectively).

The implication of their results is profound at the system level, which reveals the dependencies between supply voltage, optimal circuit sizing, and optimal performance, as demonstrated in Fig. 14. The minimal energy–delay curve for the media datapath is shown at different supply voltages. For a given supply voltage, the circuits can be designed for minimal energy over a range of delay targets, using optimization methods discussed previously. At different supply voltages,

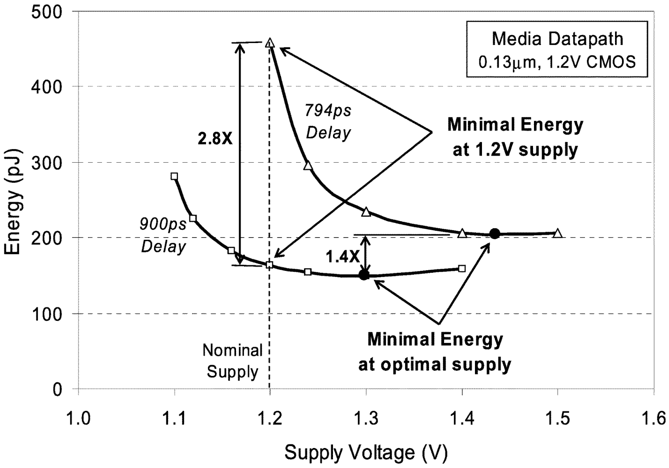


Fig. 15. Minimal energy at optimal supply voltage for the media datapath.

different optimal energy–delay curves can be obtained. At higher supply, better performance is achieved while more energy is consumed. The result is the overlap of the optimal energy–delay curves. Consequently, there exists a minimal achievable energy–delay curve over all possible supply values and circuit sizing.

There is only one single point on the optimal energy–delay curve for each supply where the energy and delay are globally optimal. This point occurs when the hardware intensity of the system matches the voltage intensity. That is

$$\eta_{\text{system}} = \theta.$$

This implies that minimal energy, optimal delay, and supply voltage are interdependent. Therefore, once the supply voltage is chosen for a system, its optimal performance and energy consumption have been determined. Likewise, once the delay target is fixed (set by the system cycle), there is an optimal supply voltage that yields the minimal energy design. Should the performance (such as for a desired throughput) be incorrectly assigned for the nominal supply, a design of smaller energy can be found with a different power supply voltage.

In practice, there exists no method to directly determine the optimal delay for a supply voltage. Another iterative step needs to be added to the optimization algorithm (Section V-D) in order to adjust the delay toward the optimal value. The hardware intensity of the system can be estimated using (7) in Appendix B and is accurate only for an infinitesimal change in delay, where input size and output load of pipeline stages are assumed constant. The result cannot be used to determine the exact delay change. Instead, it can only help to decide if the system delay should be increased or decreased when estimated η_{system} is larger or smaller than θ, respectively. Based on typical energy–delay characteristics of systems, experienced system designers may make a good educated guess to set the delay closer to the optimal value.

The potential energy saving of designs optimized for different supply voltages is observed in Fig. 15, which shows the estimated energies for the media datapath at 794 and 900 ps over a range of supply voltages. The global minimal energy design point for each delay occurs at a specific supply voltage. The energy increases significantly at lower supply voltage than the optimal value because the design is then pushed closer toward the delay-optimized design point where the energy sensitivity increases exponentially. On the other hand, more energy is consumed at higher supply voltages because the quadratic energy increase due to voltage change slightly exceeds the energy savings due to sizing reduction. In addition, Fig. 15 shows that supply-scaling optimization enables significant energy reduction between delay points compared to those at a fixed supply voltage. Furthermore, it also allows for larger delay range of the system.

VIII. Conclusion

In this paper, we have presented a systematic optimization process for pipelined digital systems. For each pipeline stage, different optimizations can be performed depending on the design objective. The analysis of these optimizations reveals the design region for the energy and delay of a pipeline stage where design choice should be made. For a pipelined system, the design choice depends on the sensitivity of each pipeline stage to its input size and output load. These sensitivity factors allow for energy tradeoffs amongst stages to minimize the total system energy. At the minimal energy design point, the derived optimal criteria show that the energy sensitivities at each pipeline boundary must be balanced. An example of this optimization process applied to a media datapath leads to the minimal energy consumption of the design under the given performance and I/O constraints. Energy savings up to 37% are obtained through the correct optimization of pipeline stages and their boundaries.

In addition, we demonstrate the interdependencies between supply voltage, minimal system energy, and optimal performance and their tradeoffs. Given a desired performance of the system, there is an optimal supply voltage and sizing where globally minimal energy can be achieved. Similar conclusions can be made when the supply voltage is fixed.

This work provides a fundamental improvement to digital pipelined system design, by demonstrating what can and should be done in each pipeline stage and at the system level as well as how to achieve the solution. The presented methods provide designers and tool developers with a systematic approach toward finding the minimal energy solution.

Appendix A
Derivation of Optimal Criteria for a Single Pipeline Stage [9], [10]

The general block diagram of a composite stage (such as individual stages of a pipelined system) is shown in Fig. 16. It consists of N logic stages. Each stage i is represented by energy E_i, delay D_i, supply voltage V, and circuit sizing η_i.

Zyuban and Strenski assumed that $E_i = E_i(V, \eta_i)$ and $D_i = D_i(V, \eta_i)$. That implies energy and delay of individual logic stages are independent of one another. This is only possible by fixing the input size and therefore output load of the stages.

The objective is to minimize the total energy of the whole stage $(E = \Sigma E_i)$ while keeping its total delay unchanged $(D = \Sigma D_i = D_0 = \text{constant})$.

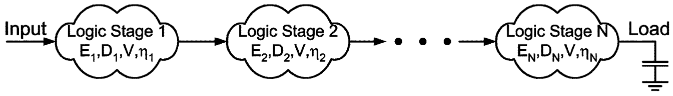


Fig. 16. Block diagram of a pipeline.

That is

- Minimize:

$$E(V,\eta_1,\ldots,\eta_N)=\sum_i E_i(V,\eta_i)$$

- Constraint:

$$D(V,\eta_1,\ldots,\eta_N)=\sum_i D_i(V,\eta_i)=D_0=\text{const.}$$

Solution is found from minimizing LaGrange function

$$L=\sum_i E_i(V,\eta_i)+\lambda\left(\sum_i D_i(V,\eta_i)-D_0\right)$$

(a) Solving the LaGrange function with respect to the sizing of stage i, we will get

$$\frac{\partial L}{\partial \eta_i}=0,\qquad \text{for } 1\to N$$

$$\Longleftrightarrow \lambda=-\frac{\partial E_i(V,\eta_i)}{\partial \eta_i}\frac{\partial \eta_i}{\partial D_i(V,\eta_i)}=\left.\frac{\partial E_i}{\partial D_i}\right|_{\text{fixed V.}}$$

Multiply both sides with D/E

$$\begin{aligned}\frac{D}{E}\lambda&=-\frac{D}{E}\left.\frac{\partial E_i}{\partial D_i}\right|_{\text{fixed V}}\\&=-\frac{\left(\frac{E_i}{E}\right)}{\left(\frac{D_i}{D}\right)}\frac{D_i}{E_i}\left.\frac{\partial E_i}{\partial D_i}\right|_{\text{fixed V}}\\&=\frac{w_i}{u_i}\eta_i.\end{aligned}\tag{1}$$

The left term is the same regardless of value i. Therefore, at optimal energy

$$\frac{w_i}{u_i}\eta_i=\frac{w_j}{u_j}\eta_j\qquad \forall\,(i,j).$$

In addition, note that the total intensity of the stage

$$\eta_{\text{stage}}=-\frac{D}{E}\frac{\partial E}{\partial D}=-\frac{D}{E}\frac{\sum_i \partial E_i}{\sum_i \partial D_i}$$

cannot be related to (1). Therefore, η_{stage} has no connection to the optimal solution.

(b) Solving the LaGrange function with respect to supply V, we will get

$$\frac{\partial L}{\partial V}=\sum_i\frac{\partial E_i(V,\eta_i)}{\partial V}+\lambda\sum_i\frac{\partial D_i(V,\eta_i)}{\partial V}=0$$

$$\Longleftrightarrow \lambda=-\frac{\sum_i \frac{\partial E_i(V,\eta_i)}{\partial V}}{\sum_i \frac{\partial D_i(V,\eta_i)}{\partial V}}=-\frac{\frac{\partial E}{\partial V}}{\frac{\partial D}{\partial V}}=-\left.\frac{\partial E}{\partial D}\right|_{\text{fixed size}}=\frac{E}{D}\theta.\tag{2}$$

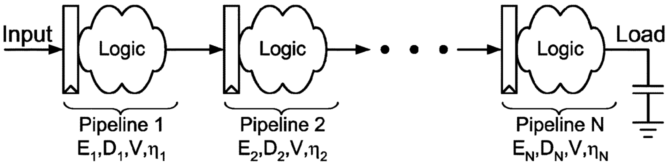


Fig. 17. Block diagram of a simple pipelined system.

Combining (1) and (2), we will get the optimal criterion for minimal energy of the composite stage

$$\boxed{\frac{w_i}{u_i}\eta_i=\theta\qquad \text{for } i=1\to N.}\tag{3}$$

Note that (3) is exact and very simple in form. Unfortunately, the result is undermined by the energy and delay dependencies between adjacent stages via their input sizes and output load (Section III). Accounting for these dependencies leads to no analytical solution.

Appendix B
Derivation of Optimal Criteria for a Pipelined System [9], [10]

Fig. 17 shows the general diagram of a simple pipelined system. It consists of N pipeline stages. Each stage i is represented by energy E_i, delay D_i, supply V, and circuit sizing η_i. Similar to the composite-stage case, energy and delay of individual stages are assumed independent of one another, which is only possible by fixing input size and output load of each pipeline. That is $E_i=E_i(V,\eta_i)$ and $D_i=D_i(V,\eta_i)$.

The goal is to minimize the total energy of the system ($E=\Sigma E_i$) while maintaining the same delay ($D_i=D_0$) in each pipeline stage. That is

- Minimize:

$$E(V,\eta_1,\ldots,\eta_N)=\sum_i E_i(V,\eta_i).$$

- Constraint:

$$D_i(V,\eta_i)=D_0=\text{const}\qquad \text{for } i=1\to N.$$

Solution is found from minimizing the LaGrange function

$$L=\sum_i E_i(V,\eta_i)+\sum_i \lambda_i\left[D_i(V,\eta_i)-D_0\right].$$

(a) Solving the LaGrange function relative to the sizing of stage i, we will get

$$\frac{\partial L}{\partial \eta_i}=\frac{\partial E_i(V,\eta_i)}{\partial \eta_i}+\lambda_i\frac{\partial D_i(V,\eta_i)}{\partial \eta_i}=0\qquad \text{for } i=1\to N$$

$$\Longleftrightarrow \lambda_i=-\frac{\partial E_i(V,\eta_i)}{\partial \eta_i}\frac{\partial \eta_i}{\partial D_i(V,\eta_i)}=\left.\frac{\partial E_i}{\partial D_i}\right|_{\text{fixed } V.}\tag{4}$$

Sum up all (4) and multiply both sides with D/E

$$\begin{aligned}\frac{D}{E}\sum_i\eta_i&=-\frac{D}{E}\sum_i\left.\frac{\partial E_i}{\partial D_i}\right|_{\text{fixed } V}\\&=-\sum_i\left(\frac{E_i}{E}\frac{D}{E_i}\left.\frac{\partial E_i}{\partial D_i}\right|_{\text{fixed } V}\right)\\&=\sum_i w_i\eta_i,\quad \text{with } w_i=\frac{E_i}{E}\qquad \forall i.\end{aligned}\tag{5}$$

In addition, the equal-delay constraint allows that $\partial D_i = \partial D$ $\forall\, i$. Then

$$\begin{aligned}\frac{D}{E}\sum_i \eta_i &= -\frac{D}{E}\sum_i \frac{\partial E_i}{\partial D_i}\bigg|_{\text{fixed } V}\\ &= -\frac{D}{E}\,\frac{\sum_i \partial E_i}{\partial D}\bigg|_{\text{fixed } V}\\ &= -\frac{D}{E}\,\frac{\partial E}{\partial D}\bigg|_{\text{fixed } V}\\ &= \eta_{\text{system}}.\end{aligned} \tag{6}$$

Combining (5) and (6), we will get

$$\eta_{\text{system}} = \sum_i w_i \eta_i. \tag{7}$$

The system hardware intensity is computable from the hardware intensity of individual stages.

(b) Solving the LaGrange function relative to the supply

$$\begin{aligned}\frac{\partial L}{\partial V} =& \sum_i \frac{\partial E_i(V,\eta_i)}{\partial V}\bigg|_{\text{fixed size}}\\ &+\sum_i \lambda_i \frac{\partial D_i(V,\eta_i)}{\partial V}\bigg|_{\text{fixed size}} = 0.\end{aligned} \tag{8}$$

By definition

$$\begin{aligned}\frac{\partial E_i(V,\eta_i)}{\partial V}\bigg|_{\text{fixed size}} &= \frac{E_i}{V}\left(\frac{V}{E_i}\frac{\partial E_i(V,\eta_i)}{\partial V}\right)\bigg|_{\text{fixed size}}\\ &= \frac{E_i}{V}E_{V,i}\\ \frac{\partial D_i(V,\eta_i)}{\partial V}\bigg|_{\text{fixed size}} &= \frac{D_i}{V}\left(\frac{V}{D_i}\frac{\partial D_i(V,\eta_i)}{\partial V}\right)\bigg|_{\text{fixed size}}\\ &= -\frac{D_i}{V}D_{V,i}\\ &= -\frac{D}{V}D_{V,i}.\end{aligned} \tag{9}$$

Note that E_V and D_V terms refer to the normalized sensitivity of energy and delay with respect to voltage. The expressions can be seen directly in (9). Substituting these expressions into (8)

$$\frac{\partial L}{\partial V} = \sum_i \frac{E_i}{V}E_{V,i} - \sum_i \eta_i \frac{D}{V}D_{V,i} = 0. \tag{10}$$

Zyuban and Strenski assumed that E_V *and* D_V *for all stages of the pipeline [system] are equal* [9], [10]. Then, (10) can be written as

$$\begin{aligned}E_V\sum_i E_i - DD_V\sum_i \lambda_i = EE_V - DD_V\sum_i \lambda_i = 0\\ \sum_i \lambda_i = \frac{E_V}{D_V}\frac{E}{D} = \theta\frac{E}{D}\\ \theta = \frac{D}{E}\sum_i \lambda_i = \eta_{\text{system}}.\end{aligned} \tag{11}$$

Combining (7) and (11), we will get the optimal criterion for minimal energy of a system

$$\boxed{\textstyle\sum_i w_i\eta_i = \eta_{\text{system}} = \theta.} \tag{12}$$

The above equation appears quite impact and simple. Similar to the composite stage, the result is undermined by the assumption on independency between pipeline stages. Inclusion of the dependencies will not lead to an analytical solution.

ACKNOWLEDGMENT

The authors would like to thank V. Zyuban for discussions and suggestions.

REFERENCES

[1] S. Borkar, “Design challenges of technology scaling,” *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul.–Aug. 1999.

[2] V. G. Oklobdzija, *High-Performance System Design: Circuits and Logic*. New York: IEEE Press, 1999.

[3] J. P. Fishburn and A. E. Dunlop, “TILOS: A polynomial programming approach to transistor sizing,” in *Proc. Int. Conf. Computer-Aided Design*, Nov. 1985, pp. 326–328.

[4] V. G. Oklobdzija, B. Zeydel, H. Dao, S. Mathew, and R. Krishnamurthy, “Energy–delay estimation of high-performance microprocessor VLSI adders,” presented at the 16th Symp. Computer Arithmetic, Santiago de Compostela, Spain, Jun. 2003.

[5] V. G. Oklobdzija, B. R. Zeydel, H. Q. Dao, S. Mathew, and R. Krishnamurthy, “Comparison of high-performance VLSI adders in the energy–delay space,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 6, pp. 754–758, Jun. 2005.

[6] H. Q. Dao, B. R. Zeydel, and V. G. Oklobdzija, “Energy minimization method for optimal energy–delay extraction,” presented at the Eur. Solid-State Circuits Conf., Estoril, Portugal, Sep. 2003.

[7] D. Markovic, V. Stojanovic, B. Nikolic, M. Horowitz, and R. Brodersen, “Methods for true energy-performance optimization,” *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1282–1293, Aug. 2004.

[8] A. Chandrakasan, S. Sheng, and R. Brodersen, “Low-power CMOS digital design,” *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, Apr. 1992.

[9] V. Zyuban and P. Strenski, “Unified methodology for resolving power-performance tradeoffs at the microarchitectural and circuits levels,” in *Proc. Int. Symp. Low Power Electronics and Design*, Aug. 2002, pp. 166–171.

[10] ——, “Balancing hardware intensity in microprocessor pipelines,” *IBM J. Res. Develop.*, vol. 47, no. 5/6, pp. 585–598, Sep./Nov. 2003.

[11] D. Harris, R. F. Sproull, and I. E. Sutherland, *Logical Effort: Designing Fast CMOS Circuits*. San Mateo, CA: Morgan Kaufmann, 1999.

[12] V. G. Oklobdzija and E. R. Barnes, “On implementing addition in VLSI technology,” *J. Parallel Distrib. Comput.*, no. 5, pp. 716–728, 1988.

[13] T. Sakurai and A. R. Newton, “Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas,” *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.

[14] N. Hedenstierna and K. Jeppson, “CMOS circuit speed and buffer optimization,” *IEEE Trans. Comput.-Aided Des.*, vol. CAD-6, no. 2, pp. 270–281, Mar. 1987.

[15] *HSPICE Simulation and Analysis User Guide*, Version W-2005.03, Synopsys, Mountain View, CA, 2005.

[16] P. M. Kogge and H. S. Stone, “A parallel algorithm for the efficient solution of general class of recurrence equations,” *IEEE Trans. Comput.*, vol. C-22, no. 8, pp. 786–793, Aug. 1973.

[17] P. M. Vaidya, “A new algorithm for minimizing convex functions over convex sets,” presented at the 30th Annu. Symp. Foundations of Computer Science, Research Triangle Park, NC, Oct.–Nov. 1989.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer, 1999.

[19] A. A. Farooqui and V. G. Oklobdzija, “General data-path organization of a MAC unit for VLSI implementation of DSP processors,” presented at the IEEE Int. Symp. Circuits and Systems, Monterey, CA, May-Jun. 1998.

[20] F. Chehrazi, V. G. Oklobdzija, and A. A. Farooqui, “High performance universal multiplier,” U.S. Patent 6 353 843, Mar. 5, 2002.

[21] A. A. Farooqui, V. G. Oklobdzija, and F. Chehrazi, “Multiplexer based adder for media signal processing,” in *IEEE Int. Symp. Very Large Scale Integr. (VLSI), Technol., Syst., Applicat.*, Taipei, Taiwan, R.O.C., Jun. 1999, pp. 100–103.

[22] N. Nedovic, M. Aleksic, and V. G. Oklobdzija, "Conditional techniques for low power consumption flip-flops," in *Proc. 8th IEEE Int. Conf. Electronics, Circuits and Systems*, Sep. 2001, pp. 803–806.

[23] T. Gemmeke, M. Gansen, H. J. Stockmanns, and T. G. Noll, "Design optimization for low-power high-performance DSP building blocks," *IEEE J. Solid-State Circuits*, vol. 39, no. 7, pp. 1131–1139, Jul. 2004.

Hoang Q. Dao (S'00–M'05) received the B.S. degree (*summa cum laude*) in electrical engineering and computer engineering from the University of California at Davis in 1997, where he is currently pursuing the Ph.D. degree.

He was an intern with IBM Research Laboratory, Austin, TX, in 2000 and 2001, and with Intel Circuit Research Laboratory, Hillsboro, OR, in 2002. His expertise is digital circuit research with focus on development of energy-efficient arithmetic circuits and design methodology. He has coauthored ten conference papers and two journal papers.

Bart R. Zeydel (S'00) received the B.S. degree in computer engineering from the University of California at Davis in 2001, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

In 2000, he worked at Mentor Graphics on the VRTX real-time operating system. In 2001, he worked at Fujitsu Microelectronics where he designed datapath elements for a VLIW processor and at Telairity Semiconductor, where he developed portable hard-IP datapath blocks. In 2003, he was an intern at Intel Corporation's Circuits Research Laboratories, Hillsboro, OR, where he designed datapath elements for DSPs. His research interests include high-performance and low-power datapath circuits, design methodologies for energy-efficient high-performance and low-power digital circuits, and the development of CAD tools for design in the energy–delay space.

Vojin G. Oklobdzija (M'82–SM'88–F'96) received the Dipl. Ing. degree in electrical engineering from the University of Belgrade, Belgrade, Yugoslavia, in 1971, and the Ph.D. degree from the University of California at Los Angeles in 1982.

From 1982 to 1991, he was at the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he made contributions to the RISC processors and superscalar computer design resulting in several patents, the most notable one on register renaming, which enabled a new generation of computers. From 1988 to 1990, he was an IBM visiting faculty at the University of California at Berkeley. Since 1991, he has been a Professor at the University of California at Davis, and has served as a consultant to many companies, including Sun Microsystems, Bell Laboratories, Hitachi, Fujitsu, SONY, Texas Instruments Incorporated, Intel, Samsung, and Siemens Corporation, where he was a Principal Architect for the Infineon TriCore processor. He holds 14 U.S., 7 international, and 5 other patents pending. He has published more than 150 papers, three books, and many book chapters in the areas of circuits and technology, computer arithmetic and computer architecture. He has given over 150 invited talks and short courses in the U.S., Europe, Latin America, Australia, China, and Japan. He directs the ACSEL Laboratory (http://www.ece.ucdavis.edu/acsel), which is involved in digital circuit's optimization for low-power and ultra low-power, high-performance system design and sensor nodes.

Prof. Oklobdzija is a Distinguished Lecturer of the IEEE Solid-State Circuits Society. He serves as Associate Editor for the IEEE TRANSACTIONS ON COMPUTERS, the IEEE *Micro*, and the *Journal of VLSI Signal Processing*. He served as Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS from 1995 to 2003, the ISSCC digital program committee from 1996 to 2003, first A-SSCC in 2005, and numerous other conference committees. He was a General Chair of the 13th Symposium on Computer Arithmetic and IASTED Conference on Circuits, Signals and Systems.