Energy-Delay Tradeoffs in Combinational Logic using Gate Sizing and Supply Voltage Optimization

Vladimir Stojanovic[1], Dejan Markovic[2], Borivoje Nikolic[2],
Mark A. Horowitz[1], and Robert W. Brodersen[2]
[1]Stanford University; [2]University of California, Berkeley

Abstract

This paper relates the potential energy savings to the energy profile of a circuit. These savings are obtained by using gate sizing and supply voltage optimization to minimize energy consumption subject to a delay constraint. The sensitivity of energy to delay is derived from a linear delay model extended to multiple supplies. The optimizations are applied to a range of examples that span typical circuit topologies including inverter chains, SRAM decoders and adders. At a delay of 20% larger than the minimum, energy savings of 40% to 70% are possible, indicating that achieving peak performance is expensive in terms of energy.

1. Introduction

Energy efficient digital systems are typically designed either to minimize the energy consumption subject to a throughput constraint, or to maximize the amount of computation for a given amount of energy. Both these design optimizations can be made if the tradeoffs between energy and delay are known since it is then possible to determine the lowest energy for a given level of performance. System level modifications can then be made to choose the optimal architecture, by choosing the appropriate level of parallelism to achieve the required level of throughput at the lowest energy [1]. In addition, a variety of circuits may exist that can be used to implement a sub-function in the system. Critical system-level decisions rely on delay and energy estimates of the resulting implementations. This paper therefore focuses on the problem of minimizing the energy subject to a delay constraint for a given circuit topology, by utilizing gate sizing and supply voltage optimization.

Gate sizing and supply voltage can be used in various ways to trade off energy and delay. The challenge is to find the most efficient arrangement for a given energy-delay space. The purpose of this paper is to provide methodology for trading off energy and delay, increasing designer's ability to understand available optimization techniques and quantify their effectiveness. To accomplish this task a relationship between the energy profile of a circuit topology (the energy dissipated in each stage of logic) and the potential energy savings has been determined. This analysis reveals the topological properties that have the largest impact on the efficiency of each optimization, providing bounds on energy reduction using sizing and supply optimizations on any logic path.

2. Delay and energy models

The simple model of the drain current, [2], is used as a baseline for the derivation of the gate delay formula:

$$t_d = \frac{K_d W_{in} V_{dd}}{\left(V_{dd} - V_{on}\right)^{\alpha_d}} \left(\frac{W_{out}}{W_{in}} + \frac{W_{par}}{W_{in}} \right) = \tau_{nom} g \left(h + \frac{p}{g} \right). \quad (1)$$

Parameters K_d, V_{on} and α_d are estimated using a least-squares fitting of a FO4 inverter delay across a range of supply voltages. Gate and parasitic capacitance are both linearized and the effects of transistor voltage non-linearities are lumped into the fitting parameters. The gate delay is calculated as $t_d = \tau_{nom} d$, where τ_{nom} is a process-dependent constant, and d is a unitless delay of the gate, [3]. The unitless delay is in turn determined as $d = h_{eff} + p$. There, the effective fanout h_{eff} is the product of logical effort and electrical fanout. The logical effort g describes the relative ability of a gate topology to deliver current. For an inverter, we set $g = 1$. Electrical fanout h is the ratio of the total output to input capacitance, equivalent to the ratio of the width of the transistors loading the gate to the width of the transistors connected to the input, W_{out}/W_{in}. Finally, self-loading delay p is a product of the logical effort and the ratio of the parasitic self-loading to the input gate width, $p = gW_{par}/W_{in}$. This model implicitly includes the impact of the signal slope at the output of the final stage on the delay of the subsequent loading gates.

When there are multiple supply voltages, the perception of a "process constant" to which delays are normalized changes. A gate that operates at a reduced supply voltage has a smaller device current for the same input capacitance. Thus, if V_{dd} is scaled down, the logical effort and the parasitic delay increase, as modelled by the voltage-dependent factor k_v in (2-3) and demonstrated in Fig. 1. When a regular gate is placed at the interface between the high and low supply domains, the pull-up

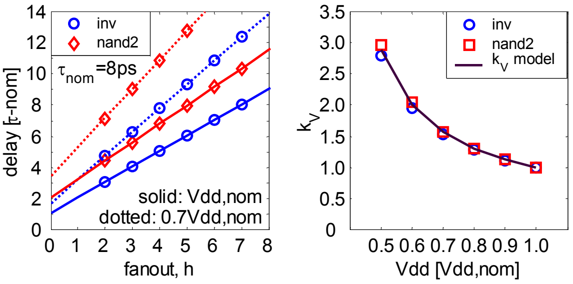

Fig. 1. Delay vs. fanout and supply.

path operates at reduced supply, thus with a higher logical effort. The size of the pull-up must increase in order to equalize the logical effort of each path. This is modelled through voltage and gate topology dependent scaling factors k_o and k_{op}.

$$g = g_{nom} \cdot k_v \cdot k_o\,,\ \ p = p_{nom} \cdot k_v \cdot k_{op} \tag{2}$$

$$k_v = \frac{V_{dd,low}}{V_{dd,nom}} \left(\frac{V_{dd,nom} - V_{on}}{V_{dd,low} - V_{on}} \right)^{\alpha_d} \tag{3}$$

The energy of a gate is modelled by its switching component defined as follows:

$$E = K_e \left(W_{out} + W_{par} \right) V_{dd}^2\,, \tag{4}$$

where $K_e W_{out}$ is the load capacitance, and $K_e W_{par}$ is the self-loading of the gate, with the switching activity lumped into the parameter K_e.

3. Energy-delay optimization

System parameters that have the largest sensitivity of energy to delay provide the biggest energy savings for a given delay penalty. By analyzing these sensitivities, the efficiency of sizing and supply optimizations can be estimated from the energy profile of the logic block. The key point here is that the optimization has a tendency to equalize all of the sensitivities on the path towards the solution point. The progress along each dimension (variable) is dictated by the sensitivity of energy to the delay due to the change in that variable. The impact of sizing and supply on the energy sensitivity to delay is formulated by:

$$\frac{dE}{dD}\left(W_i, V_{dd} \right) = \frac{\partial E}{\partial W_i} \bigg/ \frac{\partial D}{\partial W_i} + \frac{\partial E}{\partial V_{dd}} \bigg/ \frac{\partial D}{\partial V_{dd}}\,, \tag{5}$$

where W_i represents the size of the gate at stage i, and V_{dd} is the global supply. E and D represent the total energy and the total delay, respectively.

The sensitivity of energy to delay due to the sizing of stage i is given by (6). There, ec_i represents the energy contribution from stage i, and $h_{eff,i}$ is the effective fanout of stage i. Equation (6a) shows that the largest potential for energy savings occurs at the point where the design is sized for minimum delay, with equal effective fanouts. This extends the variable taper result for an inverter chain, [4], to more complex logic gates and topologies.

$$\frac{\partial E}{\partial W_i} \bigg/ \frac{\partial D}{\partial W_i} = -\frac{K_e}{K_d} \frac{ec_i}{h_{eff,i} - h_{eff,i-1}} \tag{6a}$$

$$ec_i = W_i \left(V_{dd,i-1}^2 + \frac{p_i}{g_i} V_{dd,i}^2 \right) \tag{6b}$$

The sensitivity of energy to delay due to global supply reduction is given by (7). Again, the design sized for the minimum delay at a nominal supply offers the greatest potential for energy reduction. This potential diminishes with the reduction in supply voltage. As the energy decreases, the delay increases and the additional factor in (7) also decreases. The same formula can be applied to dual supply voltage optimization. In that case, E and D would represent the total energy and delay of stages under low supply voltage.

$$\frac{\partial E}{\partial V_{dd}} \bigg/ \frac{\partial D}{\partial V_{dd}} = -\frac{E}{D} \frac{2\left(1 - V_{on}/V_{dd} \right)}{\alpha_d - 1 + V_{on}/V_{dd}} \tag{7}$$

4. Optimization examples

The reference for all the comparisons is a circuit that operates at a nominal supply voltage set by technology, sized for minimal delay d_{min} at a given output load WL. We refer to it as the nominal circuit. Starting with d_{min}, we define a delay constraint $d_{con} = (1+d_{inc}/100)d_{min}$, where d_{inc} is the percentage delay increment. The minimum energy is then found using the supply voltages, gate sizing and the additional buffer insertion as optimization variables. The delay constrained energy minimization represents a geometric program, which can be formulated in a convex form, [5]. Effectiveness of the optimization using different sets of system variables is evaluated on designs over a range of circuit topologies including those with and without off-path loads and reconvergence. Supply optimizations investigated include global supply reduction, multiple discrete supplies, and per-stage supplies that decrease from the input to the output. Gate sizing is investigated both individually and in combination with the supply reduction.

4.1. Inverter chain: no off-path load or reconvergence

Figure 2a illustrates the use of gate sizing to minimize the energy of a fixed length inverter chain. Initially, all stages have the same delay. Due to the geometric progression in size, most of the energy is dissipated in the last few stages, with the largest energy stored in the final load. Starting from the minimal delay point where all of the sensitivities are infinite, we change the gate sizes along the chain so that all of the sensitivities decrease uniformly. This leads to increase in effective fanout toward the output where most of the energy is consumed, as shown in (6). The biggest energy savings, for a fixed delay increase, are achieved by downsizing the largest gates in the chain first.

If the number of stages can be varied, the delay constraint may be met with a smaller number of stages leading to a larger energy reduction. Intuitively, as the final stage is downsized to gain the biggest energy savings for a given delay increase, the size and the number of the remaining stages adjust to meet the delay constraint, Fig. 2b. Since the number of stages must be an integer, as the fanout of the final stages grows, the fanout of the earlier stages must decrease, sometimes

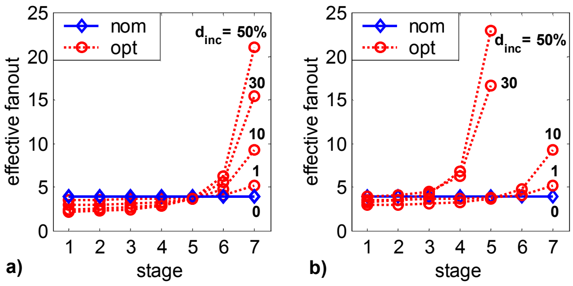

Fig. 2. Sizing: a) fixed, b) variable number of stages.

pushing them below the optimal fanout per stage. This effect is clear in Fig. 2a, where the number of stages is fixed. In Fig. 2b, with a variable number of stages, the effect is smaller but still exists. Since most of the energy is consumed in driving the fixed final load, the maximum energy saving from sizing is limited to about 30%.

Scaling the supply, on the other hand, directly affects the energy needed to charge the final load capacitance, and therefore can have a larger effect on the total energy. In the nominal case, where the delay of each stage is equal, the supply sensitivity of each stage only depends on the energy of that stage, as indicated in (7). Like the sizing, supply voltage optimization therefore adds incremental delay first to the stages with the highest energy consumption (stages towards the end of the chain) and increases the effective fanout of these stages by lowering their supply voltage. Figure 3 shows the optimized per-stage supply and the resulting effective fanout.

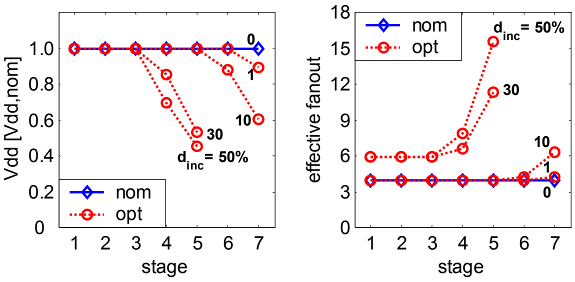

Fig. 3. Per-stage supply, variable number of stages.

The worst-case effective fanout increase occurs when sizing optimization is utilized to minimize the energy of an inverter chain. The supply optimization requires less change in the effective fanout for the same energy reduction. In practical designs, the effective fanout of the gate is bounded by the signal slope constraints to around 10 to 15.

4.2. Memory decoder: off-path load without reconvergence

A buffer chain has a particularly simple energy distribution, one which grows geometrically to the final stage. This type of profile drives the optmization over sizing and V_{dd} to focus on the final stages first. Most practical circuits have a more complex energy profile, for example an SRAM decoder. The decoder shares some characteristics with a simple inverter chain – the total capacitance at each stage grows geometrically – but the number of active paths decreases geometrically as well. As a result, the peak of the energy distribution is often in the middle of the structure. In particular, this decoder has the energy peak at the output of the predecoder.

Figure 4 shows the critical path of a 256 wordline SRAM decoder. The multiplication factor *m* denotes the number of active gates at each stage. Branching occurs at the input of each NAND gate and the number of active gates per stage decreases in a geometric fashion and selects only one wordline at the output. Sizing optimization effectively reduces the internal energy peaks through direct gate sizing or increase in nominal number of stages, as shown in Fig. 5. The initial sizing for minimum delay does not require an extra buffer at the output of the decoder.

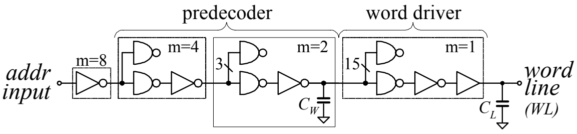

Fig. 4. Critical path, 256 wordline SRAM decoder.

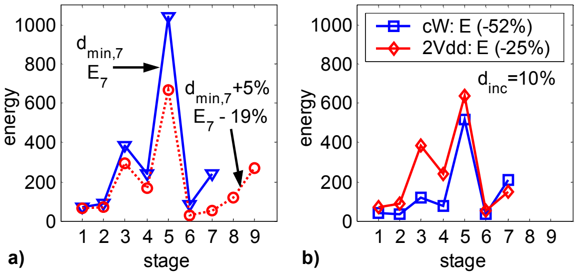

Fig. 5. Energy distribution, SRAM decoder, *WL* = 128.

Inserting a buffer stage at the output reduces the effective load presented by the 256 decoder/word driver cells. This decreases the load on the predecoder gate and effectively reduces the energy consumption at that node, Fig. 5a. Alternatively, optimization by direct gate sizing minimizes the size of the word driver input and achieves the same effect, as shown in Fig. 5b. This relates to heuristic [6] that divides the sizing problem into two sub-problems: a) sizing of predecoder logic to drive the minimum word driver input, and b) sizing of word driver to drive the wordline.

The supply optimization is less effective in designs where the peak of the energy consumption occurs inside the block. In order for the supply to affect the energy peak, the delay of all stages after the peak needs increase, reducing the marginal return, Fig. 5b.

4.3. Adder: off-path load and reconvergence

More complex designs may have reconvergent fanouts and multiple active outputs qualified by paths with various logic depth. As an example, we analyze a 64-bit Kogge-Stone tree adder. There are many paths through an adder, and unlike the decoder, not all of these paths are balanced. This raises two questions: how to size the initial design, and how to display the resulting energy maps. To be fair, the initial sizing makes all paths in the adder equal to the critical path. As a result, further reductions in size would cause the delay of the adder to increase. Since the paths through an adder roughly correspond to different bit slices, we allocate each gate in the adder to a bit slice. This partition works well for tree adders, and Fig. 6 shows the resulting energy map for the minimum delay, as well as the situation when a 10%

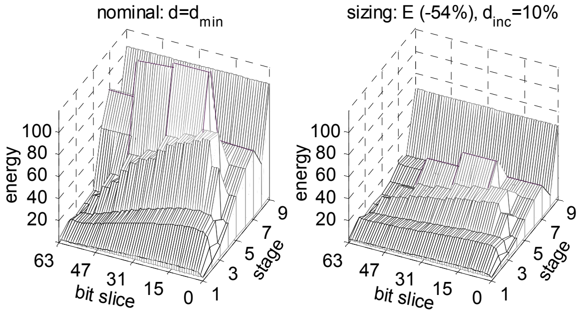

Fig. 6. Energy distribution, adder, *WL* = 32.

delay increase is allowed. Like the decoder, the dominant energy peaks are internal, which makes transistor sizing more effective than V_{dd} scaling. The data indicates that a 54% decrease in energy is possible using transistor sizing, while only 27% is saved using two supplies.

5. Energy reduction bounds

In delay-constrained energy reduction using sizing optimization, an inverter chain represents the worst case for any single logic path initially sized for minimal delay. Compared to any other gate, the inverter has the smallest delay for any given amount of energy when driving a load. Moreover, sizing optimization cannot minimize the energy dissipated in driving the output load, which represents the largest portion of the total energy. In any other design involving gates with higher logical effort, the energy dissipated in driving the final load is a smaller portion of the total energy since the capacitance ramp-up is smaller. Therefore, the inverter chain establishes a lower bound on the effectiveness of delay constrained energy minimization through gate sizing on a single logic path, as shown in Fig. 7a. This result for a single logic path can be used to predict savings in design with multiple paths. As in the adder example, all paths should initially be sized as critical, in which case some of the paths already start away from the minimum delay point, producing smaller marginal returns.

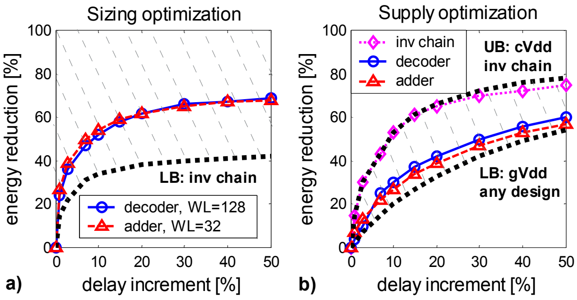

Fig. 7. Energy reduction bounds: a) sizing, b) supply.

A lower bound on energy reduction through supply optimization is given by global supply reduction, Fig. 7b. Delay of a logic gate and a logic path scales by the same factor k_v with global supply resulting in the same energy savings for a given delay increment. An upper bound on energy savings achieved through supply reduction on any single logic path initially sized for minimal delay, is also found in an inverter chain, for the same reason it was a lower bound for sizing. The upper bound is defined by per-stage supply optimization of an inverter chain. Using only two supplies in the case of an inverter chain is almost as effective as the upper bound, Fig. 7b.

The performance of joint to individual sizing and supply optimizations is compared in Fig. 8. The joint optimization has an additional degree of freedom to choose a more efficient method between sizing and supply at each point towards the optimal solution. It is, therefore, always better than any of the individual optimizations, regardless of the circuit topology and energy profiles. As a side effect, if increasing V_{dd} above the nominal value is possible, energy savings without any delay penalty can be achieved. The increased V_{dd} combined with sizing optimization results in lower energy because the marginal returns of V_{dd} are smaller than those of sizing. In adder example, 15% higher V_{dd} with sizing optimization provides 40% energy savings with no delay penalty.

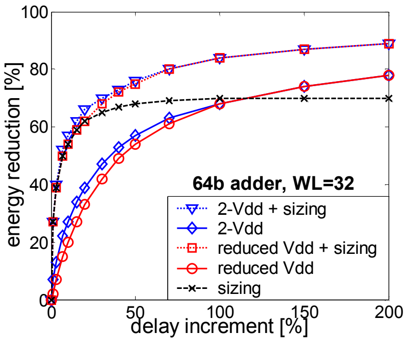

Fig. 8. Comparison of optimization techniques.

6. Conclusions

In topologies with a monotonic increase in energy toward the output, such as an inverter chain, supply reduction achieves the largest energy savings with sizing being much less effective. If however off-path load and reconvergent fanout are present, this raises the internal energy, favoring the sizing optimization.

In a design optimized for speed, the nominal clock cycle should be set around 10% higher than the theoretical minimum due to the large energy benefit offered for a small delay penalty. The returns from sizing quickly fall off, and above 20% the return is very small. In contrast, global supply reduction is the least effective energy reduction technique for small delay increments, but is quite useful when the delay increment is sizeable. The dual-supply design closely tracks the theoretical limit of the supply reduction on a per-stage basis. It is found that for the circuits analyzed, at a delay increment of 20%, at least 30% energy savings can be achieved by sizing and 30%-60% by supply optimization. A combination of sizing and supply can provide 40-70% savings.

7. Acknowledgments

This research is supported in part by MARCO contracts: CMU 2001-CT-888, GSRC 98-DT-660, and Georgia Tech B-12-D00-S5.

8. References

[1] A.P. Chandrakasan and R.W. Brodersen, “Minimizing Power Consumption in Digital CMOS Circuits,” *in Proc. IEEE*, pp. 498-522, Apr. 1995.

[2] T. Sakurai and R. Newton, “Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas,” *IEEE JSSC*, pp. 584-594, Apr. 1990.

[3] I. Sutherland, B. Sproul, and D. Harris, “Logical Effort: Designing Fast CMOS Circuits,” San Francisco, CA: *Morgan Kaufmann*, 1999.

[4] S. Ma and P. Franzon, “Energy Control and Accurate Delay Estimation in the Design of CMOS Buffers,” *IEEE JSSC*, pp. 1150-1153, Sept. 1994.

[5] J-M. Shyu *et al.*, “Optimization-Based Transistor Sizing,” *IEEE JSSC*, pp. 400-409, April 1988.

[6] B.S. Amrutur “Design and Analysis of Fast Low-Power SRAMs,” *Ph. D. dissertation*, Stanford Univ., Aug. 1999.