A New Technique for Jointly Optimizing Gate Sizing and Supply Voltage in Ultra-Low Energy Circuits

Scott Hanson Dennis Sylvester David Blaauw
University of Michigan
{hansons,dmcs,blaauw}@umich.edu

ABSTRACT

Mobile applications with battery lifetimes on the order of thousands of days have placed stringent energy requirements on circuits. In this paper, we propose a new energy optimization technique for ultra-low energy circuits operating in the subthreshold regime. Our technique uses simultaneous gate sizing and supply voltage scaling to reduce energy. We demonstrate the effectiveness of our technique on benchmark circuits and offer insight on the roles of the timing distribution and wire capacitance in determining the achievable energy reductions.

Categories and Subject Descriptors: B.6.3 [**Design Aids**]: Optimization

General Terms:Algorithms, Performance, Design

Keywords: Subthreshold circuits, gate sizing, voltage scaling

1. INTRODUCTION

The rise of mobile computing has moved energy and power optimization to the forefront of the semiconductor industry. For a growing class of applications, energy minimization is the overriding priority. The ZigBee Alliance, for example, has specified a low power wireless standard for applications ranging from medical sensing to home security to home environment controllers [1]. While the performance demands of these applications are low, the target battery life is on the order of hundreds or thousands of days. It is therefore very important to investigate techniques that minimize energy, potentially at the expense of performance. In this paper, we describe a new technique that uses simultaneous voltage scaling and gate sizing to achieve optimal energy. We emphasize that the focus of this paper is energy minimization rather than power minimization since energy is a more relevant metric when battery life is the primary concern.

Aggressive voltage scaling has emerged during the past few years as an extremely effective solution to the power and energy minimization problems. Dramatic energy reductions are possible, particularly when voltage is allowed to scale into the subthreshold ($V_{dd} < V_{th}$) regime. Much of the previous work focused on the functional limits of voltage scaling and techniques for extending those limits [2][3][4]. However, recent work has shown that the operating voltage that results in minimum energy consumption is well above the minimum functional voltage [5][6]. As voltage is scaled to subthreshold voltages, leakage energy increases significantly (due to a rapid increase in circuit delay) and places a limit on the energy efficiency of further voltage scaling. Design techniques for maximizing energy efficiency in the subthreshold regime remain largely unexplored and deserve attention.

In the superthreshold regime ($V_{dd} > V_{th}$), it is widely known that minimum energy operation is obtained by setting gates to their minimum sizes, thereby reducing the dynamic energy as much as possible. However, in this paper we find that this is typically not true in the subthreshold regime. We show that increasing the sizes of certain gates in a circuit will reduce the overall leakage energy of the circuit. As a result, the minimum energy operating voltage can be reduced, thereby improving overall energy efficiency. We therefore propose a new optimization technique that alternately sizes gates and scales voltage to achieve minimum energy operation. It is important to note that we are not strictly decreasing leakage. We use leakage reductions to enable supply voltage reductions. This new sizing technique is unique in two respects: it reduces total energy by *increasing* gate sizes and it simultaneously sizes gates and scales supply voltage. Our sizing tool reduces energy by up to 15% in benchmark circuits compared to the case when only voltage scaling is used.

The remainder of this paper is organized as follows. In Section 2, we describe the key implications of low voltage design. We then describe our energy optimal gate sizing/voltage scaling tool in Section 3 and present a detailed analysis of the performance of the tool on a set of benchmark circuits in Section 4. Finally, we summarize the key conclusions of this paper in Section 5.

2. OPTIMIZATION OPPORTUNITIES

We begin our exploration of low voltage optimization opportunities by considering a chain of 50 inverters in a 130nm technology with an activity factor (α) of 0.2. Complex circuits behave similarly to the simple inverter chain, so our discussion is relevant for circuits of varying complexity. Figure 1 shows the energy consumed by the inverter chain per cycle as a function of V_{dd}. As predicted in [5][6], energy reaches a minimum at a voltage (called V_{min}) due to the rise in leakage energy but continues to function below 100mV. The rise in leakage energy is a result of the rapid increase in delay when V_{dd} drops below V_{th}. For this circuit, leakage accounts for 33% of the total energy at V_{min} and offers a unique optimization opportunity.

Reducing leakage results in two benefits, both illustrated in the inset of Figure 1. On one hand, leakage at V_{dd}=266mV is reduced. More importantly, V_{min} is reduced, enabling further energy savings. It is evident that dynamic energy and leakage energy must be optimized simultaneously. Optimization of only dynamic energy (via reduction of V_{dd}) yields a circuit that is very

ISLPED'06, October 4–6, 2006, Tegernsee, Germany.

sensitive to leakage. Addressing leakage is consequently a high priority for energy optimality at low V_{dd}.

The energy consumed by a circuit per clock cycle may be represented as the sum of dynamic and leakage energies as shown in Equation 1. C_s is the switched capacitance, I_{leak} is the leakage current for the circuit, and T_{CLK} is the clock period.

$$E = E_{DYN} + E_{LEAK} = \frac{1}{2} \cdot C_s \cdot V_{dd}^2 \cdot \alpha + I_{leak} \cdot V_{dd} \cdot T_{CLK} \quad \text{(Eq. 1)}$$

By reducing T_{CLK} through gate sizing, the designer can decrease the amount of time that a circuit leaks per instruction. As long as the energy overhead of the sizing technique is low and a relatively small fraction of the total paths are critical, a reduction in T_{CLK} can yield significant leakage energy reductions. This leakage reduction, in turn, will drive further reduction in V_{min}. The notion of increasing gate sizes to reduce energy consumption is counterintuitive. Low power designers have typically chosen minimum-sized gates in order to reduce the dynamic energy consumed by a circuit. The authors of [4] point out that gate sizes may be increased to achieve energy reduction if there are few critical paths but suggest that this is a special case. As we discuss further in Section 4.1, we find that a skewed, unbalanced timing distribution with relatively few critical paths is common for a minimum sized design.

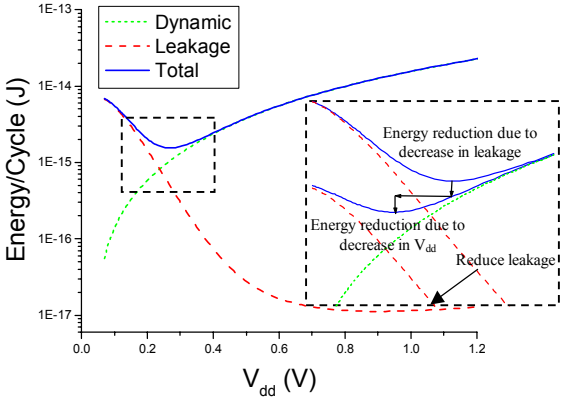

Figure 1. Energy in an inverter chain (n=50, α=0.2) is minimized at V_{dd}=266mV due to the rise in E_{leak}

3. A LOW VOLTAGE SIZING TOOL

The simultaneous scaling of V_{dd} and sizing of gates is a difficult problem since the two degrees of freedom (gate sizes and V_{dd}) are interdependent. A change in the sizing of the circuit causes the energy optimal supply voltage (V_{min}) to change. Conversely, changing V_{dd} alters the timing and leakage characteristics of a circuit and affects the energy optimal sizing point. The problem may be formulated as a multi-dimensional constrained optimization as shown in Figure 2(a), where W is the set of all gate sizes and W_i is the size of gate i. $W_{L,i}$ and $W_{U,i}$ are the lower and upper bounds for the size of gate i. Within those bounds, gate sizes may only assume discrete values set by the standard cell library. We do not initially place a constraint on total circuit area. We will investigate the implications of this decision in Section 4.3. V_L and V_U are the lower and upper bounds on V_{dd}.

A general nonlinear optimizer [7] could be used to solve the constrained optimization problem in Figure 2(a). However, such general optimization methods often incur high runtimes, making optimization of large circuits impractical. In this paper, we re-formulate the problem as two simpler sub-optimization steps that are performed iteratively as shown in Figure 2(b). The two sub-optimizations are: (1) supply voltage optimization, and (2) sizing optimization. Both sub-optmization problems are well known and a number of different methods are available for solving them efficiently. For the supply voltage optimization, we use a binary search and for the sizing optimization we use a simple sensitivity based method similar to the approach in [8].

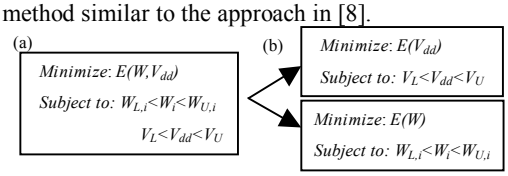

Figure 2. (a)The simultaneous optimization of gate size and supply voltage may be formulated as a two dimensional constrained optimization. (b) We simplify the problem by breaking it into two simpler optimization steps.

Figure 3 shows pseudo-code describing the operation of the sizing algorithm. The top-level optimization iterates between the two sub-optimizations to converge on the optimal solution. The algorithm begins with a circuit composed entirely of minimum-sized gates operating at a sufficiently high voltage (i.e. well above V_{min}). Using this initial approximate solution, the first sub-optimization is solved to find the energy optimal supply voltage (V_{min}) for the unsized circuit using SPICE-generated characterization data. At this voltage, leakage represents a significant portion of total energy, and the sensitivity of leakage to gate sizing is very high. A sizing optimization is performed at this V_{dd}=V_{min}. Each gate in the circuit is evaluated using a sensitivity metric to determine the change in energy that would result from a unit increase in gate size. The gate with the highest sensitivity is then sized up. Sizing continues at the same voltage until gate sizing no longer results in an energy improvement. At this point, a new supply voltage optimization is performed and V_{min} is again determined for the circuit. Iteration between voltage optimization and sizing optimization continues until convergence. We now show that this iterative formulation will converge to an optimal solution.

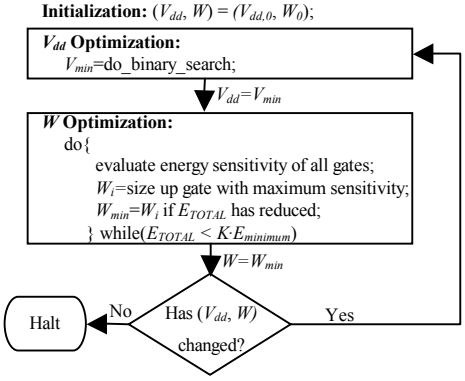

Figure 3. Pseudo-code for a low voltage sizing tool.

To guarantee convergence, we assume that the two sub-optimizations are ideal. In other words, we require: 1) that the V_{dd} optimizer returns the energy optimal supply voltage, V_{min}, for a fixed set of gate sizes, W and 2) that the gate size optimizer returns the energy optimal set of gate sizes, W_{min}, for a fixed supply voltage, V_{dd}. It is also assumed that any change in V_{dd} or

W suggested by a sub-optimizer reduces the total energy. Furthermore the outputs of the V_{dd} optimizer and gate size optimizer may be described by functions $f(W)$ and $g(V_{dd})$, respectively, where $f(W)$ gives V_{min} for a particular set of gate widths and $g(V_{dd})$ gives the set of W_{min} for a particular V_{dd}.

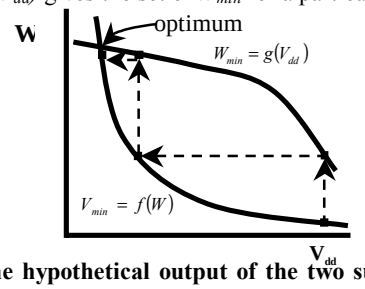

Figure 4. The hypothetical output of the two sub-optimizers is plotted (solid lines). Dashed lines show the path traversed by the top level optimization as the optimum is approached.

The characteristic curves for $f(W)$ and $g(V_{dd})$ are plotted in Figure 4. For clarity of illustration, the total width is shown on the Y-axis, although the figure can be extended to a multi-dimensional space where each individual gate size is an individual gate width. The solution to the top level optimization lies at the point where $f(W)$ and $g(V_{dd})$ intersect. At this point, both gate sizing and supply voltage are optimal. Our optimization algorithm will iterate back and forth between $f(W)$ and $g(V_{dd})$ until convergence as shown with arrows on the diagram. The optimization will converge as long as the following conditions are satisfied:

$$\frac{\partial f(W)}{\partial W} = \frac{\partial V_{min}}{\partial W} \le 0, \quad \frac{\partial g(V_{dd})}{\partial V_{dd}} = \frac{\partial W_{min}}{\partial V_{dd}} \le 0, \qquad \text{(Eq. 2)}$$

$$\frac{\partial^2 f(W)}{\partial W^2} = \frac{\partial^2 V_{min}}{\partial W^2} \ge 0, \quad \frac{\partial^2 g(V_{dd})}{\partial V_{dd}^2} = \frac{\partial^2 W_{min}}{\partial V_{dd}^2} \le 0$$

Note that the first two conditions in Equation 2 tend to push optimization toward larger gate sizes and lower V_{dd}. In other words, a reduction in V_{dd} causes an increase in W, which causes further reduction in V_{dd}. The first condition requires that an increase in W always leads to a reduction in V_{min}. It is clear an increase in W results in larger dynamic energy consumption. However, we know that the total energy consumption is reduced by such a gate size increase since this is the objective of the optimizer. Consequently, an increase in W results in a higher dynamic energy/leakage energy ratio. In [5] and [6], it was shown that a circuit with a larger dynamic/leakage energy ratio tends to have a lower V_{min}. This shows that an increase in transistor width by the sizing optimizer will result in a V_{min} that is lower, satisfying the first condition in Equation 2. The second condition means that a reduction in V_{dd} always leads to an increase in W_{min}. We saw in Section 2 that a reduction in V_{dd} causes leakage energy to increase. To mitigate this leakage, the gate size optimizer will reduce T_{CLK} further by increasing the sizes of the gates along the critical path. Hence, a V_{dd} reduction will lead to a gate size increase, which satisfies the second equation in Equation 2. The third and fourth conditions, which place requirements on the second derivatives of $f(W)$ and $g(V_{dd})$ are satisfied because the energy overheads of gate sizing and V_{dd} scaling force $f(W)$ and $g(V_{dd})$ to saturate near the optimum.

To model delay and energy in our sizing algorithm, we use SPICE characterization data for a 130nm standard cell library at twelve voltage points ranging from 130-350mV. This range of voltages is sufficient to contain all V_{min} values for a variety of benchmarks under widely varying switching activity conditions. We verify our sizing algorithm using a set of benchmarks that includes the ISCAS85 benchmarks as well as two more complex circuits. The standard cell library contains 38 cells including inverters, 2 and 3-input NAND gates, 2 and 3-input NOR gates, and buffers of various sizes. Each benchmark is originally synthesized using only minimum-sized gates. The average switching activity values for each node are extracted from a Verilog simulation assuming an input switching activity of 0.2. In addition, we use a simple wireload model of the form $k(1+0.4(FO-1))$, where k represents the wire capacitance for a gate with one fanout, and FO is the number of fanout gates at the node of interest [9]. For our technology we choose $k=5fF$ which corresponds to a wire length of approximately 20μm.

Table 1. Results of sizing various benchmarks

Benchmark	Number of Gates	ΔE_{TOTAL} (%)	$\Delta E_{DYN} / \Delta E_{TOTAL}$ (%)	ΔArea (%)
c432	161	8.6	85	39.9
c499	544	5.9	99	20.1
c880	366	6.1	71	32.1
c1908	507	5.6	91	26.7
c1355	582	8.4	96	30.4
c2670	860	11.4	80	27.9
c3540	984	7.6	80	15.5
c5315	1668	11.7	84	19.8
c6288	2480	5.7	80	21.8
c7552	2087	15.0	81	39.8
SOVA	17559	14.7	85	47.2
R4	35039	13.9	96	60.9

4. RESULTS

In this section, we examine the energy reductions achieved using our new optimization technique. Table 1 summarizes the performance of our sizing tool on a number of benchmarks. The column labeled "ΔE_{TOTAL}" lists the energy reductions achieved using our technique. It compares the energy of the unsized design at $V_{dd}=V_{min,unsized}$ and the energy of the sized-up design at $V_{dd}=V_{min,sized\text{-}up}$. Table 1 also includes a column labeled "$\Delta E_{DYN} / \Delta E_{TOTAL}$." This quantity represents the fraction of energy savings attributed to dynamic energy improvement. Though we initially target leakage reduction, most of the energy benefits are a result in V_{min} reduction. We also list the area penalty (ΔArea), which is the increase in total transistor width as compared to the unsized design. Across all of the benchmarks, we observe that energy improves by 5.6% to 15% with the area penalty ranging from 15.6% to 60.9%. In the remainder of this section we look closely at the key factors affecting the efficiency of our algorithm.

4.1 The Effect of the Timing Distribution

The proposed sizing technique ultimately requires some timing slack to be effective. Timing slack is not easily exploited in a well-balanced circuit with many near-critical paths since many gates must be sized up to achieve small changes in T_{CLK}. As a result, as with traditional power optimization techniques our approach is more effective at reducing energy when a circuit has few critical paths. Figure 5 shows the timing distributions for c499 and c7552 before and after the completion of all gate sizing. The "before" and "after" distributions are measured at different V_{dd} but are normalized to their respective T_{CLK} values to facilitate

comparison. It is obvious from the initial timing distributions that c499 has many more critical timing paths than c7552 since the average relative path delay in the c499 initial distribution is much higher than that of the c7552 distribution. The shapes of the distributions suggest that c499 is well balanced compared to c7552. After gate sizing, the timing distribution of c499 moves slightly, but the shift in the c7552 distribution is far more significant. It is not surprising that the energy reduction achieved in c499 (5.9%) is much lower than in c7552 (15.0%). These observations show that the shape of the timing distribution is a strong indicator of whether or not our technique is effective at reducing energy. We can confirm this observation by finding the correlation between the "criticality" of a timing distribution and the observed energy savings. We quantify the "criticality" of an initial timing distribution using the average path delay and find that energy savings and average path delay are related with a correlation coefficient of 0.62. Only nine benchmarks are included in these calculations because the runtime to generate path distributions is prohibitive for large circuits.

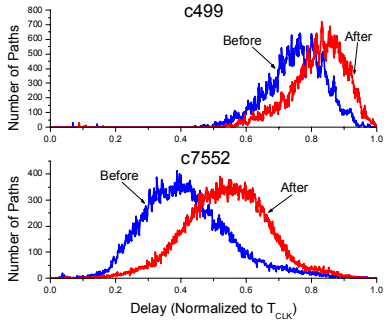

Figure 5. Timing distributions for c499 and c7552 behave differently. Delay is normalized to the clock period (T_{CLK}).

4.2 The Role of Wire Capacitance

The dynamic, leakage and total energy savings achieved for the c7552 benchmark are plotted as a function of wire capacitance in Figure 6(a). The change in transistor width after gate sizing is also plotted. At small wire capacitances, the change in total width is small (<20%). In this region, gate capacitance tends to dominate total load capacitance, minimizing the benefit of gate sizing. As wire capacitance grows, gate sizing becomes beneficial, as evidenced by the increase of both transistor width and energy savings in Figure 6(a). Designs with long routes will therefore benefit from combined V_{dd} scaling and gate sizing. Most of the energy reduction in Figure 6(a) is due to dynamic energy reduction. This is consistent with the results of Table 1, which shows that reductions in dynamic energy are responsible for 71-99% of total energy reductions. Put another way, the primary target of simultaneous V_{dd} scaling and gate sizing is reduced dynamic energy rather than reduced leakage energy.

4.3 Mitigating the Area Penalty

Though we assume energy to be the most important metric, we cannot ignore the area of the circuit. An increase in area is accompanied by an increase in cost. For many low energy applications (for example, a widespread sensor network) cost and area are overriding priorities [10]. The area penalties listed in Table 1, ranging from 15.6-60.9%, may result in an intolerable cost for such applications. By adding an area constraint to the minimization problem expressed in Section 3, the cost of sizing can be reduced. Figure 6(b) highlights the effectiveness of our sizing technique when an area constraint is asserted for several benchmarks. In the case of the R4 benchmark, the area penalty can be reduced from 60.9% to 26.4% with only 10% reduction in the energy savings. The assertion of an area constraint is simple and should be considered by designers with area limitations.

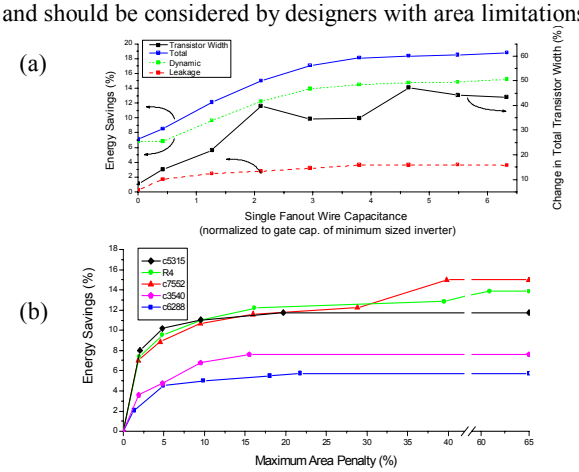

Figure 6. (a) Energy savings (% of total energy) and change in total transistor width as a function of wire capacitance. (b) Energy savings (% of total energy) as a function of the maximum allowable area penalty.

5. CONCLUSIONS

In this paper, we describe a technique that uses a combination of gate sizing and V_{dd} scaling to reduce energy. Our technique uses leakage energy reductions to drive extended voltage scaling and achieves total energy reductions of up to 15% across a set of benchmark circuits. Our results show that different design types respond differently to our technique. We find that designs with a wide timing distribution and few critical paths experience significant energy reductions using our technique. Since V_{dd} reduction ultimately drives energy reduction, we also find that the energy consumption in designs with wire-dominated load capacitances may be improved substantially using our technique. Finally, we show that the area penalty of our technique can be mitigated by applying an area constraint during optimization.

References

[1] ZigBee Alliance. http://www.zigbee.org/.

[2] J. Meindl, J. Davis, "The Fundamental Limit on Binary Switching Energy for Terascale Integration (TSI)," *IEEE Journal of Solid-State Circuits* **35**, No. 10, 1515-1516 (October 2000).

[3] A. Wang, A. Chandrakasan, "A 180mV FFT Processor Using Subthreshold Circuit Techniques," *Int. Solid-State Circuits Conf. (ISSCC)*, pp. 292-529, 2004.

[4] B. Calhoun, A. Wang, A. Chandrakasan, "Device Sizing for Minimum Energy Operation in Subthreshold Circuits," *Custom Integrated Circuits Conf. (CICC)*, pp. 95-98, 2004.

[5] B. Zhai, D. Blaauw, D. Sylvester, K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," *Design Automation Conf. (DAC)*, pp. 868-873, 2004.

[6] B. Calhoun, A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," *Int. Symp. on Low Power Electronics and Design (ISLPED)*, pp. 90-95, 2004.

[7] B.A. Murtagh, M.A. Saunders, "MINOS 5.4 User's Guide, Report SOL 80-20R," *Systems Optimization Lab., Stanford University*, Dec. 1983 (revised Feb. 1995).

[8] J.P. Fishburn, A.E. Dunlop, "TILOS: A Posynomial Programming Approach to Transistor Sizing," *Int. Conf. on Comp. Aided Design (ICCAD)*, pp. 326-328, 1985.

[9] G.A. Sai-Halasz, "Performance Trends in High-End Processors," *Proc. of the IEEE* ***83***, 20-36 (Jan. 1995).

[10] J.M. Kahn, R.H. Katz, and K.S.J. Pister, "Emerging Challenges: Mobile Networking for Smart Dust," *Journal of Comm. And Networks* **2**, No. 3, 188-196 (Sep. 2000).