Methodology for Energy-Efficient Digital Circuit Sizing: Important Issues and Design Limitations

Bart R. Zeydel and Vojin G. Oklobdzija

Advanced Computer Systems Engineering Laboratory
University of California, Davis,
One Shields Ave, Davis, CA 95616, USA,
{zeydel, vojin}@acsel-lab.com
http://www.acsel-lab.com

Abstract. This paper analyzes the issues that face digital circuit design methodologies and tools which address energy-efficient digital circuit sizing. The best known techniques for resolving these issues are presented, along with the sources of error. The analysis demonstrates that input slope independent models for energy and delay and stage based optimization are effective for analyzing and optimizing energy-efficient digital circuits when applied correctly.

1 Introduction

For several years, the semiconductor industry has been facing a power crisis. Much of the work addressing this issue has focused on system level approaches to reducing power [4][5] or on optimization tools [7]. While these advances have helped to reduce energy in digital circuits, they do not offer insight into how the savings is achieved. Additionally it is not known whether a different realization of the same function would yield better results or how that modification should be made. The opportunity for energy reduction is still significant in digital circuits where even in adders an energy savings of up to 50% without delay degradation has recently been shown [5].

In this work we examine the issues facing digital circuit design methodologies for the analysis and design of energy-efficient circuits. As the intent of a design methodology is to enable short design times, we impose the constraint that the delay and energy of each logic gate must be calculated solely from its output load and size. This inherently limits the types of energy and delay models to those that do not account for input slope variation.

The analysis is organized as follows: Section 2 presents the energy and delay models used for circuit sizing analysis; Section 3 presents the logic families which can be efficiently optimized using these models; Section 4 examines approaches to digital circuit sizing and identifies issues and limitations; Section 5 presents comparison results of different approaches to circuit sizing optimization; Section 6 presents a summary of the analysis.

The work was supported in part by California MICRO.

2 Logic Gate Delay and Energy Models

2.1 Delay and Energy Models

In this section the input slope independent models for delay and energy are presented. To simplify discussion, Logical Effort (LE) terms will be used [1][2]. The delay of a logic gate can be expressed using an RC-model [3]. One such model is the normalized LE form:

$$Td = (gh + p)\cdot \tau$$

The limitation of this model is that it does not account for the dependence of delay on input slope.

Energy can be computed directly from the gate size and output load. The model presented in [11] performs a best fit for energy due to changes in gate size and output load, where E_p and E_g are the fitting terms and $E_{leakage}$ is a function of cycle time:

$$E = E_p C_{in} + E_g C_{out} + E_{leakage}$$

The limitation of this type of energy model is that the change in short-circuit current due to input slope variation is not accounted for. Each of the parameters for the delay and energy models can be obtained from either hand estimation or simulation.

2.2 Effect of Relative Input Arrival Time on Model Parameters

The impact of relative input arrival time on the delay of a logic gate is widely known, however it is ignored in digital circuit sizing optimization and static timing analysis tools [10]. A comparison of LE parameters obtained from hand estimates with those obtained from HSPICE simulation for single input switching and worse case input switching is shown in Table 1.

Table 1: LE parameters obtained using different approaches in a 130nm CMOS technology.

Circuit	LE Hand Estimates (μ_n=2.5μ_p)		Single Input Switching HSPICE		Worst Case Input Switching HSPICE	
	g_{avg}	p_{avg}	g_{avg}	p_{avg}	g_{avg}	p_{avg}
inv	1	1	1.0	1.04	1.0	1.04
nand2	1.29	2	1.13	1.72	1.28	1.95
nand3	1.57	3	1.29	2.68	1.57	3.21
nor2	1.71	2	1.69	2.46	1.87	2.67
nor3	2.43	3	2.36	4.65	2.74	5.06

The values of *g* and *p* are 10-20% larger for worst-case input switching compared to the values obtained from single input switching analysis. Interestingly, the hand estimates for *g* fall in between the single input switching and worst case input switching values, making the values obtained from hand analysis reasonable for use in design methodologies for circuit sizing and analysis. The appropriate characterization environment depends on the goal of the designer. If static timing analysis is acceptable, single input switching can be used. However if the desire is to bound the worst

possible delay, worst case input switching should be used. The impact of input arrival time on circuit sizing will be demonstrated in Section 5.

3 Logic Families Included in Analysis

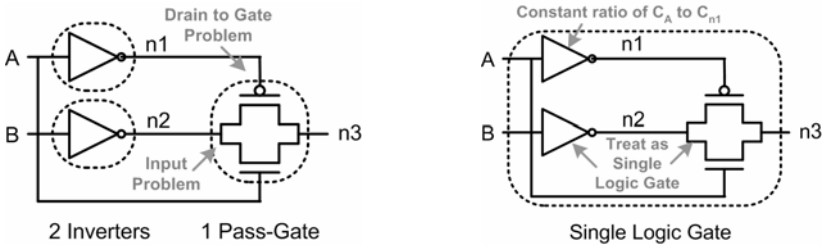
To be compatible with the models in Section 2, digital circuit sizing will only be allowed on an entire logic gate. That is, if the size of a logic gate changes by a factor α, the size of each transistor in the logic gate will also change by the same factor α. To achieve the best possible circuit sizing using these models, a logic gate will be defined as a group of transistors with the following characteristics: each input is attached to the gate of a transistor (i.e. no input can be attached to the source/drain of a transistor), for each valid combination of inputs there exists a path to V_{dd} or V_{ss}, and no source/drain can be connected to the gate of a transistor within the logic gate. The following sub-sections describe the issues (if any) that commonly used logic families have with meeting these criteria, along with techniques for resolving these issues.

3.1 Static and Domino CMOS Logic Families

Static CMOS logic gates require no modification to meet the logic gate definition. Domino gates are similar to static CMOS, except they have a keeper which violates the source/drain connection to the gate of a transistor within the logic gate. To handle the keeper, a separate inverting logic gate which does not drive the output is used for feedback to the keeper. The size of the inverting gate and the keeper scale directly with α, resulting in extra parasitic delay for the gate, *Note: The dynamic gate is treated as an individual gate separate from the static gate that follows it.*

3.2 Static CMOS with Pass-Gates

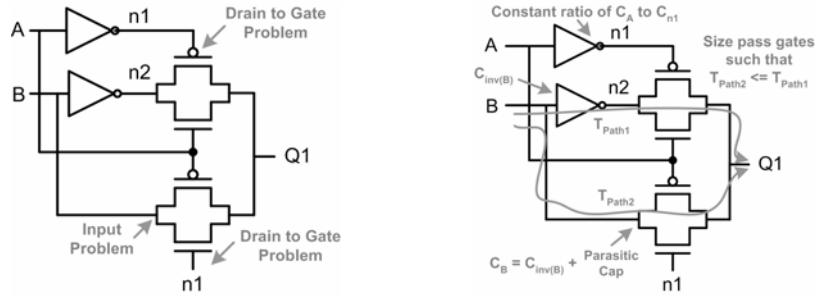
Since pass-gates are commonly used in digital circuits for implementing MUX's and XOR's it is essential that they can be analyzed and sized using the models of section 2. One of the problems with pass-gates arises from the fact that they are often treated as a logically separate entity from a static CMOS logic gate (Fig. 1a). This issue can be resolved by treating the pass gate as part of the logic gate that is driving it as is the case with the path through input B in Fig. 1b. Input A is more difficult to handle. Since the pass-gate begins to turn on when input A arrives, it is impossible to treat input A and n1 as independent inputs to the logic gate. As a result, a fixed ratio of C_{n1} to C_A must be chosen to create a roughly constant relative arrival time of signal A and n1. This exception to the logic gate definition constrains the circuit sizing optimization by fixing the relative arrival of A to n1, regardless of delay target. To achieve the best results, this relative arrival time should be selected based on the expected performance of the circuit (larger for low-speed and smaller for high-performance).



(a) (b)

Fig. 1. (a) Logic Gate definition issues when separating pass-gates from static CMOS gates (b) Modifications required to handle as a single Logic Gate.

A pass-gate implementation of an XOR logic gate creates further complications as seen in Fig. 2. The issues associated with input A can be handled as in Fig. 1. Accurate handling of input B requires that the input of a logic gate be allowed to connect to the source/drain of a transistor. The drawback of this approach is that the output load of a gate is potentially a mix of gate capacitance and a pass-gate connection, making modeling complex. Instead of increasing the complexity, an approximation can be made. The worst case delay for input B is through the inverter and upper pass-gate (which can be handled in the same fashion as in Fig. 1). However, the load that is presented to input B depends on the state of input A. Assuming only a single input switches at a time, it is seen that only one pass-gate can be "on" at a time. Therefore the load presented to input B for the worst case delay of the gate is the sum of the input capacitance of the inverter and capacitance of the lower pass-gate. To allow this approximation, the upper and lower pass-gates must be sized to guarantee that the path from B through the lower pass-gate will be faster than the path from B through the upper pass-gate regardless of the logic gate driving input B or the load at the output.



(a) (b)

Fig. 2. (a) Pass-Gate XOR logic gate definition issues. (b) Modifications and approximations required to treat as a single Logic Gate.

4 Optimization of Digital Circuit Sizing

In this section, the input slope independent energy and delay models are used to examine the optimization of digital circuit sizing.

4.1 Optimization of a Chain of Logic Gates

Sutherland and Sproull developed Logical Effort based on the characteristic that the optimal sizing for a chain of gates occurs when each gate has the same stage effort [1][2]. They demonstrated this using a RC-delay model which does not account for delay variation due to changes in input slope. From this the optimal stage effort, f_{opt}, could be found by taking the N^{th} root of the product of stage efforts (where N is the number of stages), which by definition yields equal stage effort. This solution allows for input slope independent energy and delay models to be very accurate for a chain of identical gates.

In simulation it is observed that the delay of a logic gate is approximately linear with respect to h when the ratio of input slope to output slope is constant [2][8]. As a result, the delay and energy models are accurate for each gate in the chain because equal f corresponds to equal h in a chain of identical gates. If the types of gates in a chain differ, the parasitic capacitance will differ resulting in different slopes which introduce error into the delay estimate.

The delay optimal sizing is not always the minimum energy solution for a circuit. Depending on the system that a circuit is used in, it is possible that the optimal solution might occur at a relaxed delay target (while still maintaining the same input size and output load) [6]. The energy optimization of a chain of digital circuits for a fixed delay, input size, and output load can be performed by redistributing the stage efforts from the delay optimal solution [8][6]. The error introduced by this optimization to the input slope independent energy and delay models is best observed by comparing the energy minimized sizing to the delay optimal sizing of a circuit with a smaller input size. In order for the energy minimized case to have the same delay as the delay optimized sizing the summation of the stage efforts must be the same, resulting in some stage efforts that are larger and some that are smaller than the delay optimized stage efforts. From this it can be seen that the error introduced is not additive.

In Fig. 3, a comparison of estimated minimum energy points for different delay targets versus HSPICE simulation is shown for two different chains of gates. Each circuit has the same output load and input size (with $C_{out} = 32C_{in}$). The energy error is small, demonstrating that the change in short circuit energy due to changes in input slope does not constitute a large portion of the total energy. The delay error on the other hand grows as the performance target is relaxed from delay optimal.

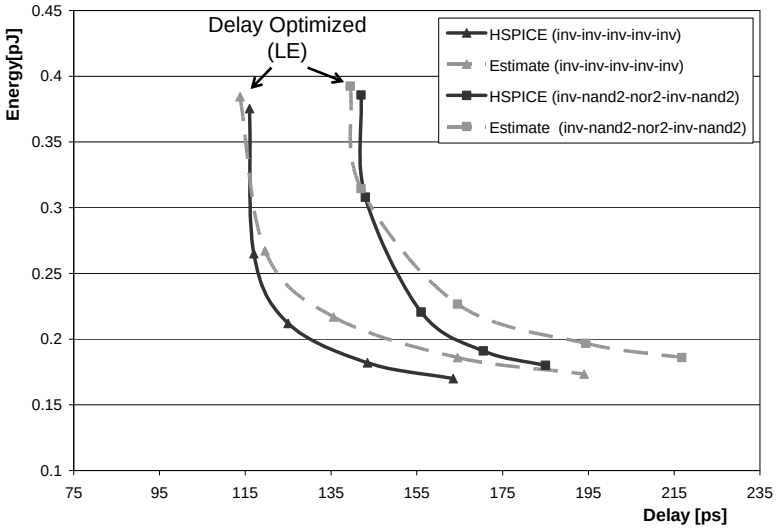


Fig. 3. Estimated Energy and Delay vs. HSPICE for two chains of logic gates optimized for different performance targets with a fixed output load and input size..

4.2 Optimization of Multiple-Output Digital Circuits

The exact optimization of digital circuit sizing for digital circuits with branching and multiple inputs is a complex problem. While recent advances have been made in computing power and numerical methods, such as the application of geometric programming [7], no commercial tools are available which solve this problem exactly or quickly. However, characteristics of the optimal solution are known. It is possible to simplify the optimization process using the characteristics of the optimal sizing solution. In a circuit optimization with a fixed output load and fixed maximum input size, the delay from input to output of each dependent path in the circuit should be equal [10]. This is true whether the circuit has a single input (Fig 4a) or multiple inputs (Fig. 4b). The condition holds whether the circuit is being optimized for delay or optimized for energy at a fixed delay. *Note: paths that are equalized can not be a subset of another path (i.e. each path must contain a unique logic gate).*

In general, for this condition to be achieved there must be no minimum size gate constraint and the inputs can not be constrained to be the same size. In the case where a minimum sized gate occurs, the gate cannot be optimized further and therefore the path it is on will be faster than the paths without minimum sized gates. Additionally, the critical path will be slower when minimum gate sizes occur as the off-path loading can not be reduced further. The case of having all inputs the same size is similar. By forcing the inputs to be the same size, certain paths are overdriven. As a result, these paths will be faster than the paths that require the largest input (which is used as

the input size for all paths). In practice these two constraints do occur, minimum sizes because of physical limitations of technology and equal input sizes because of a desire to create a regular layout.

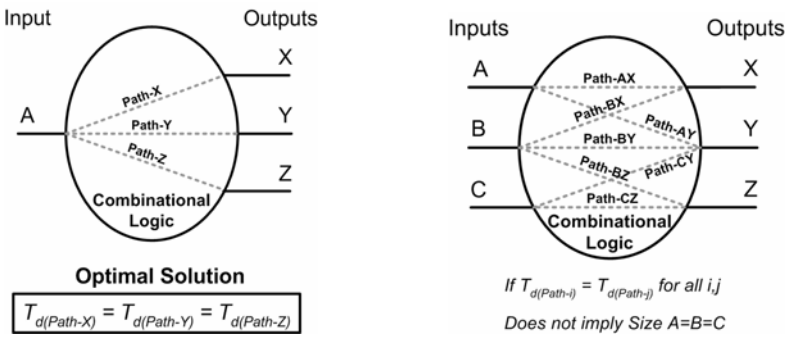


Fig. 4. Unconstrained delay optimal sizing solutions for (a) Single input multiple-output circuit and (b) Multiple input and multiple output circuit.

The optimization of a path that branches to multiple gates is examined on the circuit structure shown in Fig. 5 for a fixed output load, C_{out}, and fixed input size C_1. To allow for application of LE, the parasitic delay difference between paths is ignored (this is why the hand estimate for parasitic delay does not need to be very accurate, as it does not affect circuit sizing when using LE, it only affects the reported delay). The optimality of the solution depends on the parasitic delay difference between the paths. If the delay difference is small, the circuit sizing solution will be close to optimal however as the delay difference grows, the optimality of the solution degrades. It is important to note that the calculation of branching can be prohibitively complex in multi-stage circuits. Often a simplification is made where the branch value is treated as the number of paths attached to a node. The problem with this simplification is that it does not relate to a physical sizing, resulting in inaccurate energy estimates.

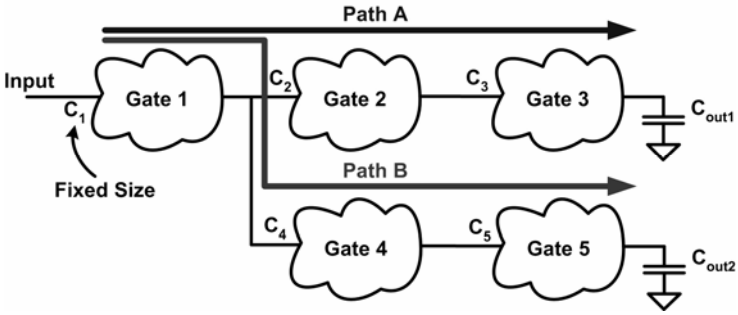


Fig. 5. Multiple output circuit with equal number of stages on each branch.

Another issues arises when the length of branches differ (i.e. if another gate were added at the end of path A in Fig. 5). In this case, LE's solution of equal effort per stage is not optimal (even when parasitic delay difference of paths is ignored). Clearly if each gate had the same stage effort, the delay associated with stage effort of Path A and B would differ. Additionally, the optimal solution does not occur when the stage effort of Gate 1 is equal to the stage effort of Gate 2 or 4. The exact solution to this case, even when ignoring the parasitic delay difference of paths, must be found using numerical methods. As larger circuits are analyzed, the complexity of this type of optimization becomes prohibitive for use in design methodologies and even CAD tools. To reduce complexity, the analysis and circuit size optimization are often decoupled in modern CAD tools, where the critical path of the circuit is identified using static timing analysis and the sizes of the circuits along that path are optimized. This process is repeated until the result converges.

4.3 Stage-Based Optimization

Instead of optimizing each logic gate individually in a circuit, the sizes can be optimized based on logic stages. The following steps describe the setup.

1) Each gate is assigned to the earliest possible logic stage.
2) Delay is equalized by assigning the same stage effort to each logic gate in the stage.
3) For logic gates that are attached only to an output, yet are not in the last stage of the circuit, the stage efforts of subsequent stage are added to the stage effort of the logic gate in order to equalize the delay of each path.

The stage assignment equalizes the delay associated with stage effort of each path except for in the following cases: an input is attached to a logic gate that is not in the first logic stage; the output of a logic gate attaches to multiple logic stages; minimum gate sizes occur; input sizes are constrained to be equal; and wires. Proper handling of these cases requires additional constraints on the optimization.

Once stage assignment is complete and additional constraints are added, the circuit can be optimized for delay or energy at a fixed delay using numerical optimization of the N variables (where N is the number of stages). The optimality of the result is dependent on the parasitic delay differences between the logic gates along a path and between paths as well as the difference in length of the branches within the circuit. Using this approach, delay and energy estimate can always be obtained which correspond to a physical sizing of the circuit, unlike with LE.

5 Results

A comparison of the optimal delay and total width for several chains of logic gates using the characterization parameters of Table 1 is shown in Table 2. The results demonstrate an increase in optimal delay of 13-18% for worst case input switching

compared to single input switching. The sizing (shown by total width) does not vary significantly for each setup. The sizing does not differ much because $f = gh$. Thus, for a chain of identical gates, regardless of the value of g, the delay optimal sizing of the chain will be the same. Differences in sizing occur because of differences in g per stage (as seen in the third and fourth cases).

Table 2: Comparison of Delay results and sizing of paths sized for delay optimization using the characterization parameters for $\underline{g}$ and p in Table 1 ($C_{out} = 12C_{in}$ for all paths).

Circuit	Hand Estimates ($\mu_n=2.5\mu_p$)		Single Input Switching		Worst Case Input Switching	
	Delay (τ)	Width (C_{in})	Delay (τ)	Width (C_{in})	Delay (τ)	Width (C_{in})
nand3-nand3-nand3	19.78	8.53	16.9	8.53	20.41	8.53
nor3-nor3-nor3	25.69	8.53	30.16	8.53	34	8.53
nand3-nor3-nand3	21.47	8.18	20.84	8.08	24.46	8.11
nand2-nand3-nor2	17.39	7.49	16.14	7.21	18.51	7.22

A comparison of gate based optimization and stage based optimization is shown in Table 3. The comparison is performed on a circuit with an inverter driving two branches (similar to Fig. 5). One branch has 8 inverters, while the other branch ranges from 2 to 8 inverters creating a path length difference of 6 to 0 inverters respectively. The output of each path is loaded the same for all cases, with $C_{out} = 128C_{in}$.

Table 3: Comparison of Gate Based Optimization and Stage Based Optimization

Path Length Difference	Gate Based Optimization		Stage Based Optimization		% worse for Stage Based	
	Delay [τ]	Energy [pJ]	Delay [τ]	Energy [pJ]	Delay	Energy
6	27.03	2.57	28.25	2.90	5%	13%
5	25.71	2.72	26.82	3.15	4%	16%
4	25.31	2.88	26.15	3.28	3%	14%
3	25.19	3.01	25.78	3.35	2%	11%
2	25.23	3.16	25.57	3.39	1%	7%
1	25.38	3.34	25.51	3.43	0%	3%
0	25.67	3.52	25.67	3.52	0%	0%

As can be seen in the table, the delay error of stage based optimization compared to gate based optimization is relatively small, regardless of the path length difference. The energy error can be up to 16%. If the path length difference is only 2 gates, the energy result is only 7% worse. Typically in digital circuits the difference in lengths of branches is only large when the off-path circuit is minimum size. In that case the stage based optimization will yield similar energy as gate based optimization. As a result, for many digital circuits, stage based optimization gives comparable results to gate based optimization. The case when stage based optimization suffers is when the resulting sizing gives off-path gate sizing that is not minimum size and the path length difference between the branches is large.

6 Conclusion

It is shown that the error introduced using delay and energy models that do not account for input slope variation is relatively small. The true challenge facing design methodologies is that the accurate application of these models requires the analysis of the entire circuit. Since system optimization requires energy and delay values for each input size and output load of a digital circuit [6], it is essential that the reported energy and delay of a design methodology have a one-to-one correlation with the resulting circuit sizing (meaning that LE branching simplifications can not be used). The analysis also showed that stage based optimization yields comparable results to gate based optimization, while reducing the complexity of the optimization. Stage based optimization also appears promising for future technologies where circuits will have a more regular layout [9] (i.e. each logic gate in a logic stage has the same size). Under this condition, it is possible to use a stage based optimization approach where the size of each logic stage is used as the variable instead of the stage effort.

References

1. R. F. Sproull, and I. E. Sutherland, “Logical Effort: Designing for Speed on the Back of an Envelop,” IEEE Adv. Research in VLSI, C. Sequin (editor), MIT Press, 1991.
2. I. E. Sutherland, R. F. Sproull, and D. Harris, “Logical Effort: Designing Fast CMOS Circuits,” Morgan Kaufmann Publisher, c1999.
3. M. Horowitz, “Timing Models for MOS Circuits,” PhD Thesis, Stanford University, December 1983.
4. V. Zyuban, P. N. Strenski, “Balancing Hardware Intensity in Microprocessor Pipelines”, IBM Journal of Research and Development, Vol. 47, No. 5/6, 2003.
5. B. R. Zeydel, T.T.J.H. Kluter, V. G. Oklobdzija, "Efficient Energy-Delay Mapping of Addition Recurrence Algorithms in CMOS", International Symposium on Computer Arithmetic, ARITH-17, Cape Cod, Massachusetts, USA, June 27-29, 2005.
6. H. Q. Dao, B. R. Zeydel, V. G. Oklobdzija, "Energy Optimization of Pipelined Digital Systems Using Circuit Sizing and Supply Scaling", IEEE Transaction on VLSI Systems, to appear, 2006.
7. S. Boyd, S. -J. Kim, D. Patil and M. Horowitz, “Digital Circuit Sizing via Geometric Programming,” Operations Research, Nov.-Dec. 2005, Vol. 53, Issue 6, pp.899-932.
8. H. Q. Dao, B. R. Zeydel, V. G. Oklobdzija, "Energy Optimization of High-Performance Circuits", Proceedings of the 13th International Workshop on Power And Timing Modeling, Optimization and Simulation, Torino, Italy, September 10-12, 2003.
9. P. Gelsinger, “GigaScale Integration for Teraops Performance -- Challenges, Opportunities, and New Frontiers”, 41st DAC Keynote, June 2004.
10. S. Sapatnekar, “Timing,” Kluwer Academic Publishers, Boston, MA, 2004.
11. V. G. Oklobdzija, B. R. Zeydel, H. Q. Dao, S. Mathew, R. Krishnamurthy, "Comparison of High-Performance VLSI Adders in Energy-Delay Space", IEEE Transaction on VLSI Systems, Volume 13, Issue 6, pp. 754-758, June 2005.