Architectural Considerations for Energy Efficiency

Hoang Q. Dao, Bart R. Zeydel, Vojin G. Oklobdžija
Advanced Computer Systems Engineering Laboratory
Department of Electrical and Computer Engineering
University of California, Davis, CA 95616
(hqdao,zeydel,vojin)@acsel-lab.com

Abstract

The formal analysis of parallelism and pipelining is performed on an 8-bit Add-Compare-Select element of a Viterbi decoder. The results are quantified through a study of the delay and energy behaviors of gates and complex circuits due to supply scaling and circuit optimization on a modified test setup accounting for routing cost. The energy-throughput relationships of both pipelining and parallelism are characterized in connection to their corresponding depth and degree, showing clear advantages of pipelining over parallelism.

1. Introduction

Technology scaling has greatly improved circuit performance. However, it comes at a cost of increasing power density which has become the bottleneck of high-performance designs [1]. Thus, further performance improvement must be achieved by other means, the most important of which is through architectural modifications [2]. However, detailed study was limited to degree-2 parallelism and depth-2 pipelining, while results for higher-degree architectures were very roughly estimated from the base circuit and the assumption of ideal throughput improvement. Zyuban et al. performed more complex analysis for pipelined microprocessors, which involved pipeline depth variation of a functional unit [3]. However, the system under study was too complex to provide an understanding of throughput improvement due to architectural approach alone. This paper revisits and extends the analysis in [2] to higher degrees using circuit optimization with a relevant setup and quantifies the relationship between throughput improvement and energy consumption.

The paper is organized as follows. Section 2 summarizes the delay and energy model of static gates used in our analysis and optimization. Section 3 formularizes the effects of supply scaling on delay and energy in single gates and validates its estimation on more complex circuits. Section 4 examines two architectural approaches for energy-delay efficiency. Section 5 analyzes the energy and delay results for these approaches applied to a Viterbi decoder add-compare-select (ACS) circuit. Section 6 concludes the paper.

2. Gate Delay and Energy Model

For analysis and optimization, models for delay and energy of static gates are needed. The delay is approximated using an RC-network model and is normalized to the per-fanout delay τ of an inverter [4]. The resulting normalized delay of gates is then nearly technology-independent and depends only on the gate size and output load. The modeling allows for quick sizing of circuits [4]. The general form of gate delay is expressed in Eq. 1.

$$d = \tau\left(g\frac{W_{out}}{W_{in}} + p\right) \qquad \text{Eq. 1}$$

where g and p are constant terms characterizing the gate and $\{W_{out},W_{in}\}$ represents its output load and input size in terms of transistor width respectively.

The energy of a gate is primarily consumed in charging its parasitic and loading capacitance. The leakage energy is negligible due to the dominance of active energy over leakage in the available 0.13µm, 1.2V CMOS technology and the high switching activity of the analyzed circuits. These energy elements can be modeled linearly to the gate size and the output load [5] as follow.

$$E = E_g W_{out} + E_p W_{in} + E_l W_{in} \qquad \text{Eq. 2}$$

where E_g, E_p and E_l are constant parameters with the first two terms associated with active energy and the last term leakage energy.

3. Effects of Supply Scaling on Energy and Delay

It is well known that supply scaling allows for large energy reduction at a cost in performance. The trend is quantified from simulation data for the typical process of the 1.2V, 0.13µm CMOS technology at room temperature.

3.1 Gate Level

Figure 1a presents the delay characteristics of gates over supply scaling. It shows that the relative delay characteristics of gates do not vary significantly over a wide range of supply voltage. This implies that supply scaling has little effect on the relative delays among different paths in a circuit. However, supply scaling does affect the reference delay τ. In addition, the short-channel effect (represented by the α-factor [6]) is

reduced at lower supply. The delay can therefore be expressed as:

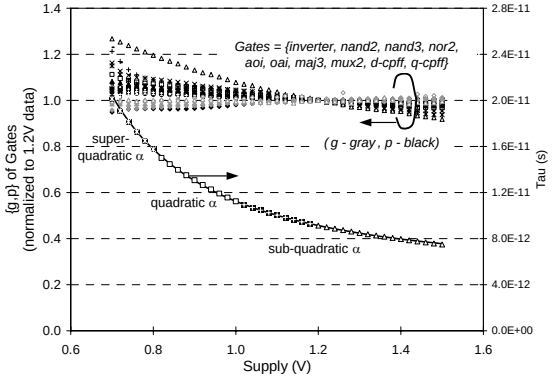
$$d = d_{nom}\frac{\tau(V)}{\tau(V_{nom})} \qquad \text{Eq. 3}$$

where *nom* denotes the results at 1.2V supply.

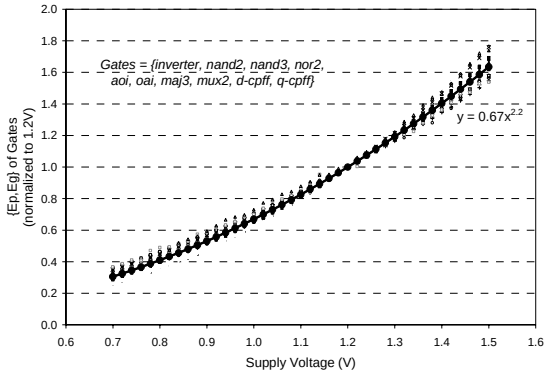
On the other hand, Figure 1b shows that gate energy has a super-quadratic dependency on supply voltage (instead of the traditional quadratic form, CV^2). This behavior is due to input-to-output coupling and higher growth of short-circuit power than the square of voltage. The energy of a gate for a given supply voltage can be estimated using:

$$E = E_{nom}\left(\frac{V}{V_{nom}}\right)^{2.2} \qquad \text{Eq. 4}$$

Thus, when supply voltage is reduced from the nominal 1.2V to 0.7V, gate delay is approximately degraded by half while energy is reduced by more than 65%. Thus, lowering supply can be used to trade delay for energy.



(a) Delay parameters normalized to 1.2V data



(b) Energy parameters

Figure 1. Effects of supply scaling on gates.

3.2 Complex Circuits

The energy delay relationship versus supply voltage in Section 3.1 is also observed in more complex circuits. Figure 2 presents simulation results for the energy and delay of 64-bit adders and 8x8-bit multipliers, compared with the average energy and delay of gates (obtained from Fig. 1) due to supply scaling. The complex circuits match well with the energy-delay relationship observed for individual gates. Thus, more energy-delay efficiency is generally achieved at lower supply voltage. In addition, the simulation results show that leakage energy is three orders of magnitude less than the active energy. This justifies that circuit energy is primarily consumed in the switching of gate and output capacitances for the given technology. In more advanced technologies, leakage energy is becoming comparable to active energy and must be accounted for. However, as Figure 2 suggests, the relative ratio between leakage and active energies is roughly the same over a wide range of supply voltage. Thus, leakage energy can be estimated from active energy and included in energy analysis.

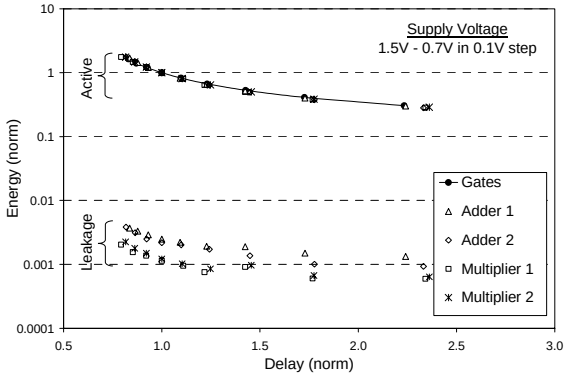


Figure 2. Effect of supply scaling on complex circuits.

4. Architectural Approaches

Architectural approaches can be used to exploit the improved energy-delay efficiency at lower supply voltage [2]. The impact of pipelining and parallelism on the energy-delay efficiency of a system is analyzed using an 8-bit Add-Compare-Select (ACS) unit. The ACS is a core element of the Viterbi decoder and its design affects the efficiency of the Viterbi decoder by a large margin [7][8]. Figure 3a shows the general block diagram.

4.1 Parallelism

One architectural approach to improving energy-delay efficiency is parallelism, where the hardware is replicated N times and the output is multiplexed as shown in Figure 3b. Unlike [2], a feedback bus is added to reflect the typical setting for real applications. The advantage of parallelism is that the throughput can be improved up to N times. The disadvantages are the addition of an N:1 multiplexer at the output and an N-time increase in circuit area and output load.

4.2 Pipelining

Another approach to improving energy efficiency is to pipeline the circuit (Figure 3c). Similar to parallelism, the throughput is improved by approximately the number of pipeline stages. No multiplexer is needed, nor does output load vary

significantly. The overhead is the addition of extra Clock Storage Elements (CSE) between the pipeline stages, which degrades performance and increases energy consumption.

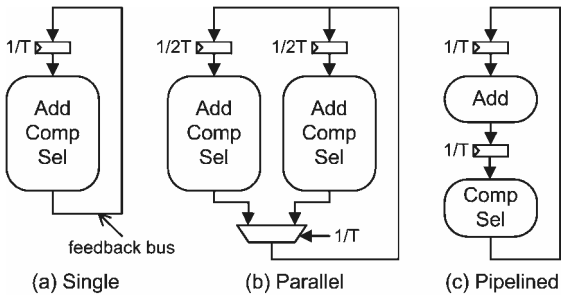


Figure 3. Architectural approaches for an ACS.

5. Results

The ACS unit is implemented with static CMOS gates using the Kogge-Stone scheme for both the adders and comparator [9] under the constraint of a 2μm input size and an equivalent 25μm-gate load. For simplicity, the ACS is considered as an isolated digital system. To achieve the best energy-delay efficiency, the whole system is optimized such that the hardware intensity η (or circuit sizing sensitivity) matches the voltage intensity θ (or voltage scaling sensitivity) as explained in [10][11]. For the 1.2V, 0.13μm CMOS technology used, $\theta \approx 2.1$ at the nominal supply voltage.

To obtain data compatible to the reference ACS unit, the parallel and pipeline designs use the same input size but their output load is adjusted accordingly. In addition, each is optimized for the same hardware intensity η of 2.1 at the nominal supply before scaling the supply voltage.

5.1 Degree/Depth of 2

Due to the limited range of delay variation achieved using supply scaling and the estimated amount of throughput improvement through architectural modifications, degree-2 parallelism or depth-2 pipelining may be sufficient for energy reduction while maintaining the same throughput.

Table 1 presents the energy comparison of the degree-2 parallelism and the depth-2 pipelining at the reference throughput. Voltage scaling results are obtained by first optimizing the designs at the nominal 1.2V supply and then applying voltage scaling formulas to estimate the supply voltage for achieving the reference throughput. Resizing results are found by minimizing energy for the reference throughput at the supply voltage estimated from the above voltage scaling results. There is only a small difference between voltage scaling and resizing data, which justifies the use of the voltage scaling estimation.

Both parallelism and pipelining allow for significant energy reduction, 56% and 44% respectively, for the same throughput as the reference and at much lower supply voltage. Parallelism approach is less efficient than pipelining because the addition of the output MUX and, more importantly, the significant increase of output load in parallelism outweighs the insertion of CSEs in pipelining.

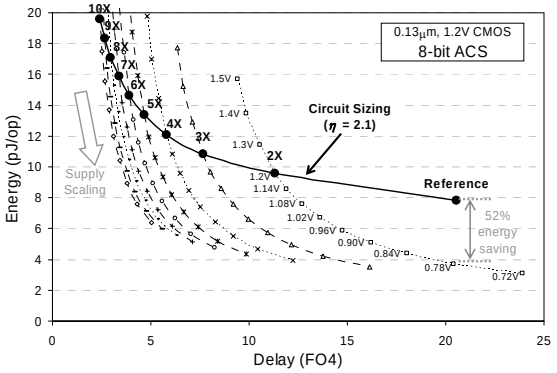
Table 1: Energy comparison versus architectures.

Design		*Supply (V)*	*Energy (pJ/op)*	*Norm*
Reference		1.2	7.83	1.00
Parallel	V scaling	0.78	3.74	0.48
	Resizing		4.36	0.56
Pipeline	V scaling	0.78	3.04	0.39
	Resizing		3.47	0.44

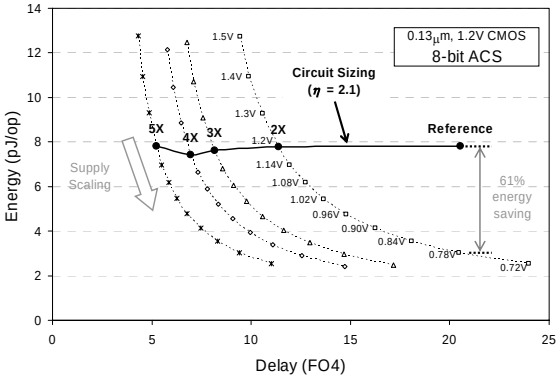
It is observed that the amount of energy reduction is less compared to [2]. This may be attributed to the smaller difference between supply and threshold voltages of the technologies.

5.2 Higher Degree/Depth

Figure 4a shows the results for parallelism of degree N from 2 to 10. The filled symbols represent the optimized solutions of different architectural degrees at nominal supply (where NX corresponds to degree N). At the nominal supply voltage the throughput is improved by 0.87N. Both energy and throughput improvement are degraded at higher degrees of parallelism due to the increasing output load.



(a) Parallelism vs. supply scaling



(b) Pipelining vs. supply scaling

Figure 4. Higher architectural degree/depth.

Figure 4b presents the energy and throughput results for pipelining for pipelining depth of 2 through 5 (where NX represents depth N). It is perhaps a coincidence that the energies do not vary much at the nominal supply voltage. However, that is possible because the CSE overhead is offset by sizing reduction in the ACS logic. The throughput improvement is 0.79N (less than parallelism) at the nominal supply voltage and is also degraded at higher depth due to increased overhead of adding more circuit stages (i.e. CSEs).

In both cases, the throughput improvement can be traded using supply scaling to achieve better energy efficiency. Furthermore, regardless of the extent of parallelism or pipelining, it is always more energy efficient to use lower supply voltage. The combined use of supply scaling and optimal selection of architectural degree allows for more than 52% energy reduction.

In addition, pipelining is more efficient than parallelism in terms of energy per operation for the given setup. This indicates that the cost of adding CSEs in pipelining is consistently less than the cost of the increase in output load due to longer routing and MUX addition when using parallelism.

The costs of parallelism and pipelining can be generalized further, as shown in Figure 5. The plot of the $Energy \cdot Delay^{\eta}$ product normalized to the reference ACS shows an interesting formulation. The exponent η is chosen because all designs are optimized for the same η. Ideally, the results should be inversely proportional to $Degree^{\eta}$. However, due to the extra loading in parallelism or cost of CSEs in pipelining, the exponent term is reduced from 2.1 to approximately 1.75 for pipelining and 1.58 for parallelism. The exponents clearly express the reduced energy efficiency of parallelism in comparison to pipelining for the given circuit setup.

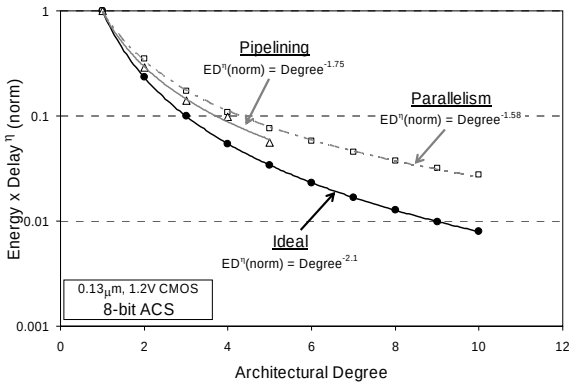


Figure 5. Effects of architectural scaling.

6. Conclusion

Architectural approaches present an important degree of freedom in achieving energy-delay efficiency. Our analysis has quantified the performance and energy improvement that parallelism and pipelining can provide. For the given circuit setup, parallelism enables slightly better throughput improvement while pipelining achieves much better energy reduction. Using the metric set by the optimization criteria, pipelining is clearly more efficient than parallelism. In addition, the energy-throughput relationship of parallelism and pipelining to degree or depth of the architecture is characterized. This knowledge in combination with degradation factors due to instruction and data dependency will allow designers to determine the optimal architectural implementation for their system.

7. Acknowledgements

This work has been supported by SRC Research Grant No. 931.001, California MICRO, Fujitsu Ltd. and Intel Corp.

References

[1] S. Borkar, “Design Challenges of Technology Scaling,” IEEE Micro, Vol. 19, No. 4, pp. 23-29, July-August 1999.

[2] A. Chandrakasan, “Low-Power Digital CMOS Design,” PhD thesis, University of California at Berkeley, UCB/ERL Memorandum No. M94/65, August 1994.

[3] V. Zyuban, D. Brooks, V. Srinivasan, M. Gschwind, P. Bose, P. N. Strenski, P. G. Emma, “Integrated Analysis of Power and Performance for Pipelined Microprocessors,” IEEE Transactions on Computers, Vol. 53, No. 8, August 2004.

[4] D. Harris, R. F. Sproull, and I. E. Sutherland, “Logical Effort: Designing Fast CMOS Circuits,” Morgan Kaufmann Publishers, 1999.

[5] V. G. Oklobdžija, Bart R. Zeydel, H. Q. Dao, S. Mathew, R. Krishnamurthy, “Comparison of High-Performance VLSI Adders in Energy-Delay Space,” IEEE Transaction on VLSI Systems, in press, 2005.

[6] T. Sakurai, A. R. Newton, “Alpha Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and Other Formulas,” IEEE J. Solid-State Circuits, vol. 25, no. 2, pp. 584-594, April 1990.

[7] P. Black, T. Meng, "A 140 MB/s 32-state radix-4 Viterbi decoder", IEEE Journal of Solid-State Circuits, vol. 27, no. 12, pp. 1877-1885, December 1992.

[8] A. Yeung, J. Rabaey, “A 210 MB/s Radix-4 Bit-Level Pipelined Viterbi Decoder,” IEEE International Solid-State Circuits Conference, Digest of Technical Papers, pp. 88-89, 1995.

[9] P.M. Kogge, H.S. Stone, “A Parallel Algorithm for the Efficient Solution of General Class of Recurrence Equations,” IEEE Trans. Computer, vol. C-22, no. 8, pp. 786-793, Aug 1973.

[10] V. Zyuban, P. Strenski, “Unified Methodology for Resolving Power-Performance Tradeoffs at the Microarchitectural and Circuit Levels,” Proc. Int. Symp. on Low Power Electronics and Design, Aug. 2002, pp. 166-17.

[11] H. Q. Dao, B. R. Zeydel, V. G. Oklobdžija, “Energy Optimization of Digital Pipelined Systems Using Circuit Sizing and Supply Scaling,” IEEE Transaction on VLSI Systems, submitted for publication.