UNIVERSITY OF CALIFORNIA, SAN DIEGO

Optimization of Energy and Throughput for Pipelined VLSI Interconnect

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Electrical Engineering (Electronic Circuits and Systems)

by

Kevin Clark Hamilton

Committee in charge:

Professor Chung-Kuan Cheng, Chair
Professor James Buckwalter
Professor Deli Wang

2010

The Thesis of Kevin Clark Hamilton is approved, and is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Chung-Kuan (C.K.) Cheng, for his support as the chair of my committee. His patience, guidance, and diligence have proved to be invaluable.

I would also like to acknowledge Yulei Zhang, without whom my research would have no doubt taken ten times as long. His support helped me in an immeasurable way.

ABSTRACT OF THE THESIS

Optimization of Energy and Throughput for Pipelined VLSI Interconnect

by

Kevin Clark Hamilton

Master of Science in Electrical Engineering (Electronic Circuits and Systems)

University of California, San Diego, 2010

Professor Chung-Kuan Cheng, Chair

As technology scales, signals may reach proportionally less and less chip area within a single clock cycle, resulting in multi-cycle paths. One solution is to pipeline such signals, being mindful of pipeline throughput. However, pipeline structures can consume substantial energy. The problem is finding the optimal tradeoff between energy and throughput in determining pipeline architecture. We derive a set of pipeline performance metrics, discover that the optimal energy/throughput tradeoff is determined by the pipeline depth, and obtain that depth as technology scales.

1. Introduction

Since Gordon Moore drew a curve on a sheet of graph paper over 40 years ago [1], engineers have worked feverishly to achieve his prophecy. Today, as ultra-deep-submicron semiconductor process technology marches yet deeper beyond the sub-wavelength barrier, and engineers delve yet further beyond the threshold of a billion discrete devices on a single die, increasingly great challenges to continued advancement form on the horizon. One of those challenges relates to global interconnect.

CMOS process scaling has enabled ever-greater integration density and higher clock frequencies. Simultaneously, the complexity of system-on-chip (SOC) designs has enjoyed a corresponding increase. Yet while process scaling has reduced transistor delays, RC wire delays have not seen a corresponding improvement because, unlike transistors, the miniaturization of wires does not improve their performance. The composite result is that as process technology scales, global signals have a shorter time in which to traverse relatively larger digital systems, using relatively slower wires. Still, in a synchronous design, each signal must arrive at its destination within the clock period. While inserted repeaters in the form of inverters or buffers have traditionally been used to meet timing constraints in long wired paths, the improvement in delay offered by this approach is finite because each repeater necessarily adds some delay. Consequently, modern designs have grown to the point where a global signal may require multiple clock cycles to propagate the length of the path.

The data throughput, or bandwidth, of a non-pipelined global signal wire with a multi-cycle path is reduced compared to a single-cycle path because the data must

occupy the multi-cycle path for multiple clock cycles. To improve the bandwidth of multi-cycle global wires, some have proposed wave pipelining of long global signal wires. Others have proposed the more traditional approach of synchronous pipelining using repeaters, such as buffers or inverters, interleaved with sequential elements, such as latches or flip-flops. As it has the most practical application, the traditional synchronous pipelining approach using flip-flops is the focus of our work here.

While the insertion of flip-flops in a multi-cycle path improves bandwidth and may improve operating frequency, it comes at a cost. Flip-flops require energy to operate. Interconnect power already comprises the majority of dynamic power consumption in some modern microprocessors [6]. Power consumption is emerging as the greatest problem in chip design, with some recent high-power processor designs already having become power-limited [14], [15]. Further, such pipelines can be extensive. For example, some predicted that at the 35nm node, the latency to propagate a signal across a chip would be as much as 32 clock cycles, a number that will only increase with further process scaling [7]. The energy required to pipeline such multi-cycle signals is significant and thus cannot be ignored. A practical, efficient pipelined interconnect architecture should attempt to achieve the highest throughput for the least amount of energy.

2. Previous Work

Researchers have studied pipelined interconnect extensively. A great deal of recent research on pipelined interconnect focuses on minimization of pipeline latency. Cocchini proposed a methodology for optimal repeater insertion in pipelined interconnects [9], [10]. The methodology inserts flip-flops and repeaters to meet one of two performance goal options. The first option is to minimize the overall interconnect latency. The second option is to meet a target latency specification. With the second option, Cocchini uses a secondary cost function to minimize allocated routing resource utilization.

Xu and Choudhury developed a set of models to study transparent latch-based interconnect pipelining [11]. Their study explored the implications of pipelined global interconnect at the architecture level, on CAD tool development, and at the circuit level. Ultimately, their algorithms yield a latch-based pipeline featuring the minimum number of sequential elements. Seth, *et al*. also advocate a latch-based approach to pipelining, exploiting retiming principles to meet delay constraints [13]. Like Cocchini, their work emphasizes latency minimization.

Other recent research considers tradeoffs between energy consumption and throughput in pipeline design. Deodhar and Davis contributed a technique reducing power dissipation through supply voltage scaling, increasing the number of repeaters within pipelined stages to offset increased delays due to reduced supply voltage [4], [5]. They recognized that inserting flip-flops to pipeline long global signals with multi-cycle paths could improve communication throughput on such paths. However, their

approach sought to improve the throughput/energy tradeoff by manipulating only the energy side of the ratio, without degrading latency, and did not explore ways to enhance throughput.

In a precursor to this study, Zhang, *et al*. extensively analyzed the electrical properties of repeated RC wires under various design goals [12]. Important for this study is the contribution of a set of metrics for delay and energy measurement for repeated interconnect. Among the metrics that will prove useful for our study are expressions for delay and energy for repeated RC wire normalized to wire length.

3. Problem Statement

Advances in CMOS process technology have led to increases in the operating frequency, integration density and overall complexity of integrated circuits. With larger dies and faster clocks, the proportion of die area reachable by a global signal within a single clock period is rapidly diminishing. Further advances mean such signals will require increasingly more clock cycles to traverse the modern digital system.

One practical solution for the resulting sequential synchronization issue is to pipeline such multi-cycle global interconnect paths by interleaving routed wire segments with repeaters and flip-flops. As we will see, pipeline insertion on multi-cycle paths also improves data throughput on that path. However, with many large-scale, high performance devices becoming power limited [14], [15], a practical pipeline solution must also consider energy consumption. The goal of this study is to optimize pipelined interconnect simultaneously for energy and throughput to ensure a workable solution in an environment where routing resources and available power are scarce.

Most prior work in this area focuses on latency minimization of pipelined interconnect. For instance, [9]-[11] and [13] offer algorithms that yield pipelines that minimize the latency of pipelined interconnect at the most latent receiver. These approaches guarantee that pipelined data is received as quickly as possible. However, these approaches fail to exploit a key property of a pipeline that enables a dramatic increase in signal throughput without substantially degrading latency.

To understand this property, consider a pipelined signal with minimum latency. Here, each register-to-register path in the pipeline necessarily has zero timing slack.

The signal throughput of the pipelined interconnect is the pipeline clock frequency because a bit of data exits the pipeline each clock cycle. An increase in pipeline clock frequency would yield a like increase in throughput, but it is not possible to increase pipeline clock frequency here because each pipeline stage is a critical timing path. Now, what the previous work did not consider is that the individual pipeline stages can be made faster by making them shorter. Since the distance between the beginning and end of the pipeline is fixed, pipeline wire segments can be made shorter by inserting additional flip-flops in the pipeline. Therefore, by adjusting the pipeline depth to add flip-flops to the pipeline, individual pipeline segments can be made shorter and thus faster, allowing an increase in clock frequency and a corresponding increase in throughput. Latency will increase, but this may be inconsequential.

The latency minimization focus of the prior work may be misdirected. With repeated RC wire, the throughput of a signal is the inverse of its latency, and a reduction in critical path latency enhances throughput because the system clock can be run faster. This explains the emphasis on latency reduction in the study of RC wires. However, with a pipelined signal, an increase in pipeline latency alone has no effect on signal throughput. The reason is that the pipeline removes the long wire from the critical path, allowing the system clock to run independent of the multi-cycle delay. Pipeline insertion causes the pipelined data to be time shifted. If the system can be modified to be made functionally tolerant of the time shift, then pipelining is a practical solution to the multi-cycle delay issue. If the system can not be made to function properly with the time shift, then pipelining is not a feasible solution to the issue. Therefore, pipelining is

only applicable to systems where the added pipeline latency is irrelevant to the overall function of the system. We only consider systems in which the addition of a pipeline in a signal does not destroy the system's functionality. Latency minimization in such circuits has no practical significance.

Instead, the figure of merit worthy of our research is pipeline *throughput*. Throughput is worthy of study because improvement of throughput allows the system to process data faster. Additionally, with global routing resources in scarce supply, maximizing the throughput of global signals could prove highly beneficial in reducing global routing congestion. For example, consider the effect of modifying a pipeline by inserting an additional flip-flop at the midpoint of each repeated wire segment. The delay of each resulting segment would be halved, allowing the pipeline clock rate to rise, roughly doubling the pipeline throughput. By multiplexing the signal so that two signals travel on the same physical path, an entire global signal wire could be made redundant, preserving a valuable resource.

Perhaps counter-intuitively, none of the latency-optimization strategies discussed above addresses pipelined signal throughput. Moreover, the studies ignore energy consumption. For pipelining to be a practical solution to the multi-cycle interconnect issue, we must look beyond latency minimization and optimize the pipeline architecture for throughput with the simultaneous objective of minimizing energy consumption.

Deodhar and Davis recognized the potential to increase data throughput of a pipelined global signal without degrading latency [4], [5]. Their approach maximized

throughput per bit energy of a pipelined signal for a given pipeline architecture. However, they saved power by lowering the supply voltage, while recovering the resulting delay penalties, and sharpening degraded transition times, by increasing the number of repeaters per unit length of interconnect. Using their approach, net energy decreased because, as energy scales with the square of the voltage, energy savings resulting from reducing the source voltage outpaced the added power consumed by the additional repeaters. However, they failed to consider the effect that simply manipulating pipeline depth and clock frequency has on throughput and energy.

This study seeks to offer practical solutions for pipeline architectures that ensure minimum energy per unit of throughput. Improved throughput in global signal channels that traverse long distances can be used to alleviate congestion in global routing layers. The contribution of an energy efficient method to achieve such improved throughput is a worthwhile endeavor.

4. Presentation of Research

Our objective requires that we construct mathematical models of our pipeline elements (wire, repeaters, and flip-flops - see Figure 1) and analyze their behavior. We will analyze our pipeline's behavior across several process technology nodes.

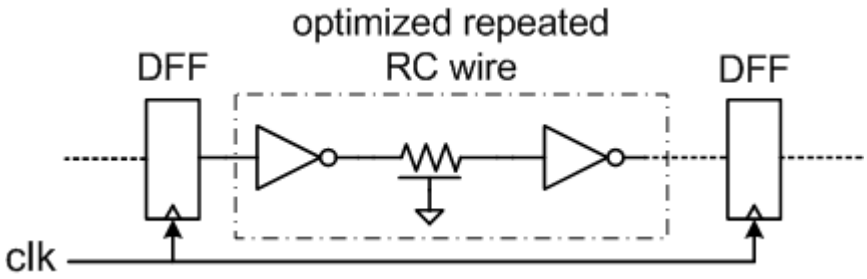

Figure 1. Pipeline Elements

Variables, parameters, and functions used in this study are defined below to clarify further expression.

4.1 Definitions and Technology Data

- N is the pipeline depth, or alternately, the number of flip-flops in the pipeline.
- L is the total wire length of a multi-cycle signal.
- T_{FF} is the total latency of single flip-flop (sum of T_{setup} and $T_{clk \rightarrow q}$).
- C_{eff} is the effective capacitance (sum of the gate capacitances) of a flip-flop.
- t_{RC} is the length-normalized delay for repeated RC wire.
- e_{RC} is the length-normalized energy for repeated RC wire.

To determine T_{FF}, a basic D flip-flop is modeled and HSPICE simulations are run to find T_{setup} and $T_{clk \rightarrow q}$. Figure 2 shows a schematic of the DFF simulated. The DFF incorporates 8X minimum sized inverters and 4X transmission gates. The Q

output is loaded with a 50X inverter. Setup and clock-to-q times are obtained by first finding the clock-to-q delay for the ideal case, where the D input arrival time is large, and then reducing the D input arrival time with respect to the clock such that the clock-

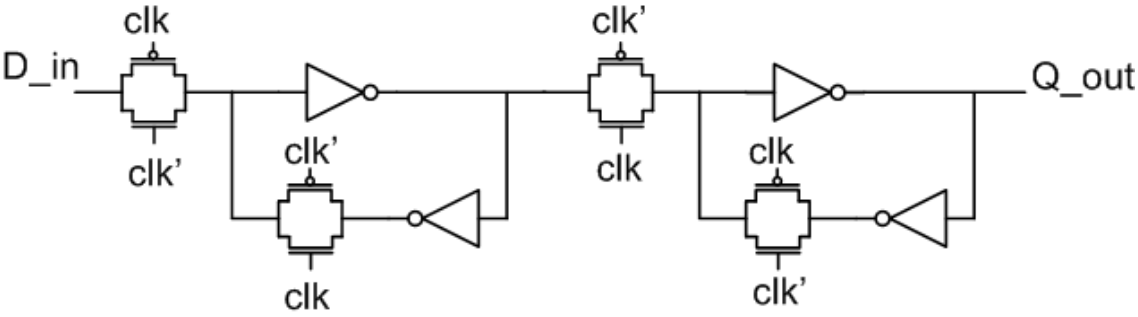

Figure 2. D Flip-Flop

to-q time is degraded by ten percent. $T_{clk \rightarrow q}$ is then the ten percent degraded clock-to-q value, and T_{setup} is the associated D input arrival time. C_{eff} is calculated by simulating the DFF with a clock frequency of 2.1GHz, a D input toggle rate of 100%, and using the formula $P = fCV_{DD}^2$. Simulations were run for 90nm, 65nm, 45nm, 32nm, and 22nm devices, using Predictive Technology Models [2]. Simulation results can be found in Table 1.

Table 1. D Flip-Flop Simulation Results

	90nm	65nm	45nm	32nm	22nm
T_{setup} (ps)	34	24	13	8	4
$T_{clk \rightarrow q}$ (ps)	22	15	10	6	3
C_{eff} (fF)	34.3	23.8	17.0	9.0	4.3

In [12], we derived expressions for delay (t_{RC}) and energy and (e_{RC}) of a minimum-delay-optimized repeated RC wire segment, normalized to wire length.

Using these expressions and ITRS predictions [3], values for length-normalized delay and energy of a repeated RC wire segment were calculated for each technology node in our study. These values can be found in Table 2. Table 3 shows V_{DD} for each process node [3].

Table 2. Wire Length-Normalized Delay and Energy of Repeated RC Wire

	90nm	65nm	45nm	32nm	22nm
t_{RC} (ps/mm)	33.9	42.9	45.9	54.6	54
e_{RC} (fJ/mm)	154.4	135.2	120.6	97.3	97.5

Table 3. V_{DD} by Process Technology

	90nm	65nm	45nm	32nm	22nm
V_{DD}	1.2	1.1	1.0	0.9	0.8

4.2 Pipeline Metrics

Several fundamental performance metrics are essential for our analysis. Those are the pipeline's latency, power, throughput, and energy. Expressions for each are given below.

4.2.1 Latency

Pipeline latency is the total time required for a bit of data to traverse the pipeline. While optimization of pipeline latency is not our ultimate goal, it remains an important metric in our analysis. Pipeline latency can be expressed as the sum of the pipeline stage delays multiplied by the total number of pipeline stages. Alternately,

latency can be written using the length-normalized repeated RC wire delay, as shown in Equation (1).

$$Latency = N(T_{clk \to q} + T_{setup} + T_{RC}^{seg}) = NT_{FF} + t_{RC}L \quad (1)$$

Observing Equation (1), the relationship between latency and pipeline depth becomes apparent. The expression shows that modifying the pipeline depth adjusts latency by the aggregate delay of the flip-flops added or subtracted from the pipeline. However, in a system where pipeline latency only shifts the system output in time, latency has no effect on system throughput, and latency is moot.

4.2.2 Power

We define pipeline power as the total power required to shift a bit of data through the pipeline. Pipeline power can be expressed as is the sum of the repeated RC wire segment power and flip-flop power for a pipeline stage multiplied by the number of such pipeline stages. As shown in Equation (2), total pipeline power can also be written in terms of the energy consumed in the flip-flops and repeated RC wire segments.

$$Power = N(P_{FF}^{seg} + P_{RC}^{seg}) = f_{BW}(NC_{eff}V_{DD}^2 + e_{RC}L) \quad (2)$$

Equation (2) confirms the intuitive conception that increasing pipeline depth increases pipeline power.

4.2.3 Throughput

The signal throughput of a repeated RC wire is the inverse of its latency. However, since a bit of data exits a pipeline each clock cycle, the throughput of a pipelined repeated RC wire is the pipeline clock rate. Throughput can also be written in

terms of the length-normalized repeated RC wire delay, as shown in Equation (3). Note that pipeline throughput can also be thought of as the pipelined signal's bandwidth.

$$Throughput = \frac{1}{T_{cycle}} = \frac{1}{T_{FF} + t_{RC}(L/N)} \tag{3}$$

Table 2 shows that as technology scales, wire-length normalized delay of repeated RC wire actually grows. Additionally, transistor devices are getting faster at a rate greater than that at which wires are becoming shorter [16]. The resulting increase in the dominance of wire delay in Equation (3) suggests that the technique of increasing pipeline depth to improve throughput will become increasingly effective as technology scales.

4.2.4 Energy

Pipeline energy can be expressed as the sum of the energy required to charge the gate capacitances of each pipeline flip-flop's transistors and the total energy dissipated in the repeated RC wire segments for one bit of data, as shown in Equation (4).

$$Energy = \frac{Power}{Throughput} = NC_{eff}V_{DD}^2 + e_{RC}L \tag{4}$$

We know that both power and throughput increase with pipeline depth N. Equation (4) tells us that energy, as well, increases with pipeline depth N, leading to the conclusion that power grows with pipeline depth at a greater rate than does throughput.

A comparison of Equations (3) and (4) reveals a tradeoff between throughput and energy with varying pipeline depth because throughput improves with longer pipeline depth at the expense of energy consumption. Our goal of optimizing pipelined

interconnect simultaneously for energy and throughput can thus be realized by determining the optimal pipeline depth to balance energy and throughput.

4.3 Optimal Pipeline Depth

To balance energy and throughput, we take the ratio of the metrics. The energy/throughput ratio becomes our objective function for optimization. To find the optimal pipeline depth N given our objective function, we take the partial derivative of our objective function with respect to *N*, set the result to zero, and solve for *N*, as shown in Equation (5).

$$\begin{aligned} \frac{\partial}{\partial N}\left[\frac{Energy(N)}{Throughput(N)}\right] &= \frac{\partial}{\partial N}(NC_{eff}V_{DD}^2 + e_{RC}L)(T_{FF} + t_{RC}\frac{L}{N}) \\ &= C_{eff}V_{DD}^2T_{FF} - \frac{1}{N^2}e_{RC}t_{RC}L^2 = 0 \\ N_{opt} &= \sqrt{\frac{e_{RC}t_{RC}}{C_{eff}V_{DD}^2T_{FF}}}L \end{aligned} \quad (5)$$

Equation (5) shows us that optimal pipeline depth for our objective function is directly proportional to wire length. Due to the relatively rapid reduction in transistor delay compared to wire delay as technology scales, (5) also suggests that optimal pipeline depth increases as technology scales.

Figure 3 shows a plot of the ratio of energy and throughput as a function of pipeline depth for the process technology nodes under study. The figure shows the

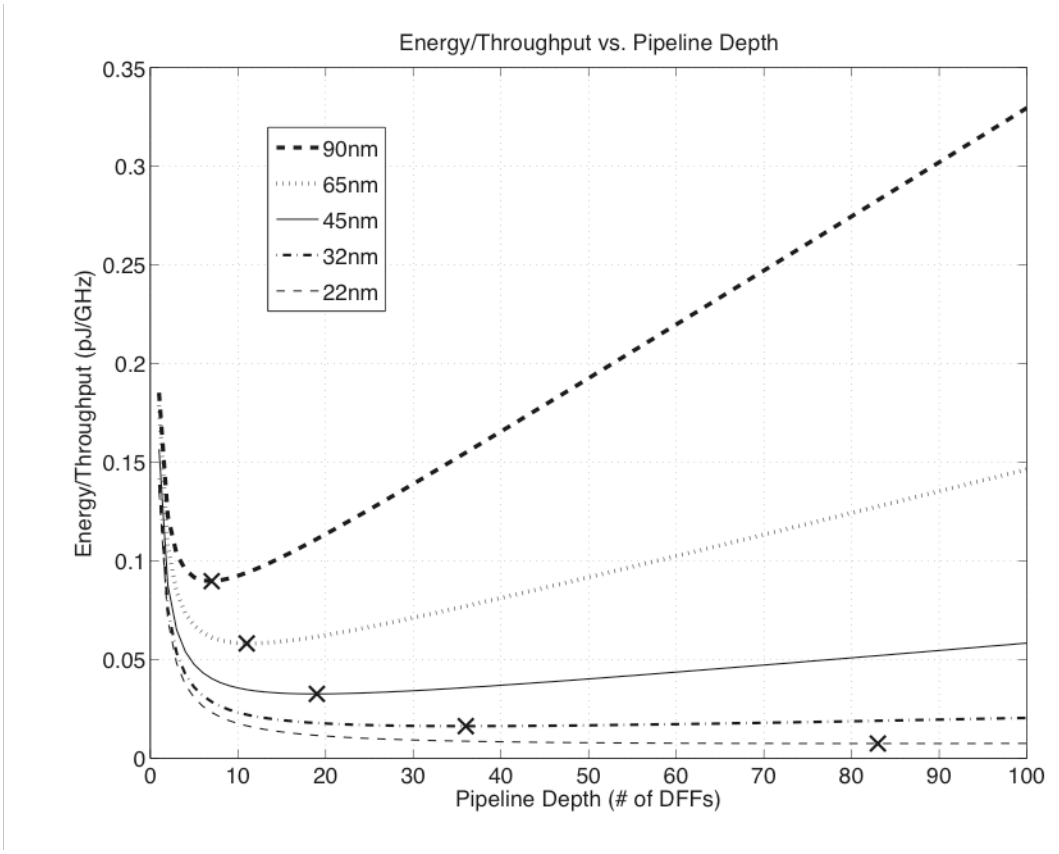

Figure 3. Optimal Pipeline Depth

optimal pipeline depths, occurring at the function minima, for our objective function as technology scales. The figure confirms our prediction that the technique of increasing pipeline depth to improve throughput becomes increasingly effective as technology scales.

Rearrangement of N_{opt}, shown in Equation (6), reveals the physical implications of the optimization. (6) shows that at the optimal pipeline depth, the energy-delay product of the repeated RC wire segments is balanced with that of the flip-flops.

$$e_{RC}\left(\frac{L}{N_{opt}}\right)t_{RC}\left(\frac{L}{N_{opt}}\right) = C_{eff}V_{DD}^2 T_{FF} \tag{6}$$

The ratio of energy to throughput of a repeated RC wire can also be expressed as *delay*^{2}×*power*. The equivalence holds true for repeated wire because throughput is the inverse of latency. In a pipelined repeated RC wire, where throughput is not the inverse of latency, the equivalence does not hold. Nonetheless, an analogy can be drawn between our optimization of pipeline energy-throughput ratio to optimization of the *delay*^{2}×*power* objective function.

4.4 Evaluation of Latency at Optimal Pipeline Depth

Plugging N_{opt} back into our expression for latency yields Equation (7).

$$\begin{aligned} Latency\big|_{PRC} &= N_{opt}T_{FF} + t_{RC}L \\ &= \left(\sqrt{\frac{e_{RC}t_{RC}T_{FF}}{C_{eff}V_{DD}^2}} + t_{RC}\right)L = t_{PRC}L \end{aligned} \tag{7}$$

We know that as technology scales, delay is increasingly dominated by RC wire delay over device delay [16]. We also know from [12] that the length-normalized delay of RC wire increases as technology scales. With this knowledge, Equation (7) suggests that for a given wire length, latency of pipelined RC wires increases as technology scales. Figure 4 confirms this hypothesis.

Note that (7) also reveals the length-normalized latency for pipelined repeated RC wire optimized to our min-energy/throughput objective function. This is a natural extension of our work in [12], where we found length-normalized delay of repeated RC wire optimized to the analogous *delay*^{2}×*power* objective function.

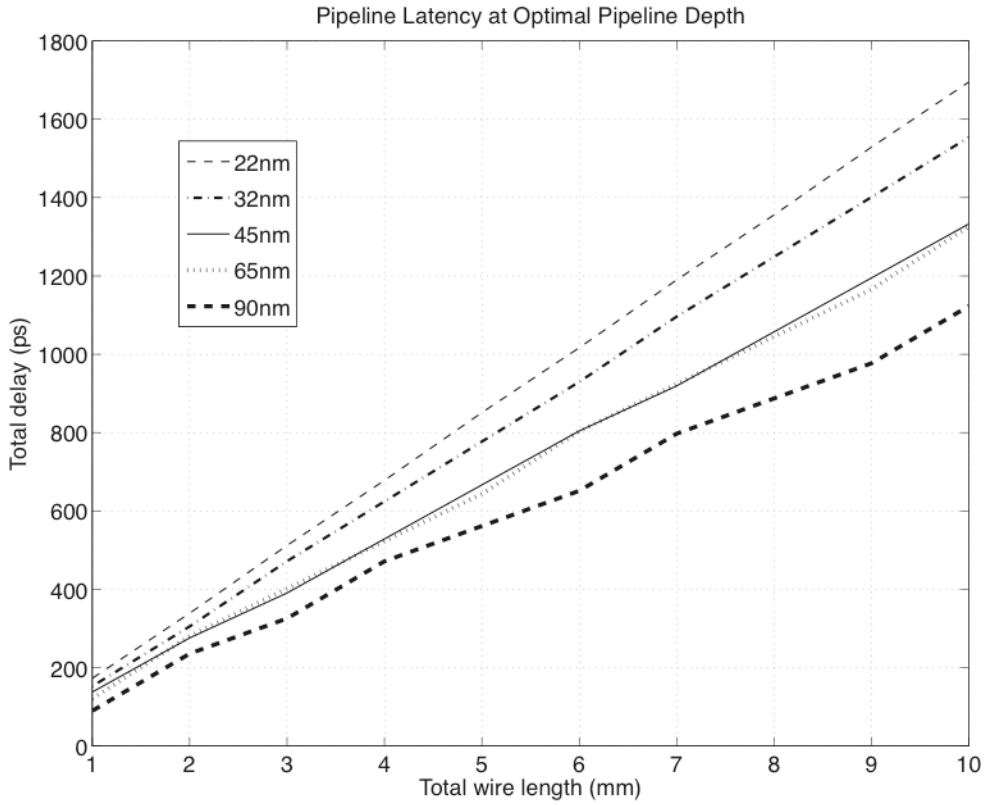

Figure 2. Latency at Min-Energy/Throughput Objective

4.5 Evaluation of Power at Optimal Pipeline Depth

Plugging N_{opt} into our power expression yields Equation (8). A comparison of Table 1 to Table 2 reveals that flip-flop delay decreases with technology scaling quicker than does the length-normalized repeated RC wire energy. Thus, for a given wire length, power tends to increase as technology scales. Figure 5 gives us confirmation.

$$Power\big|_{PRC} = f_{BW}(N_{opt}C_{eff}V_{dd}^2 + e_{RC}L) = \frac{e_{RC}L}{T_{FF}}p_{PRC}L \qquad (8)$$

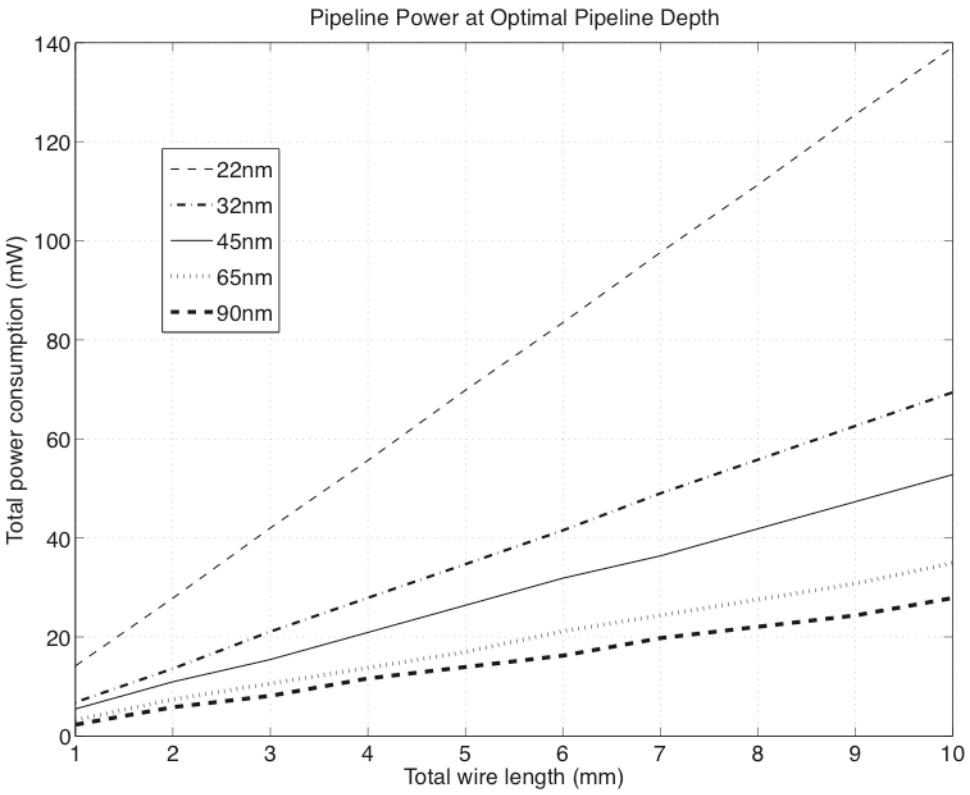

Figure 5. Power at Min-Energy/Throughput Objective

4.6 Evaluation of Throughput at Optimal Pipeline Depth

Substituting N_{opt} into our expression for throughput yields Equation (9). We learn from (9) that throughput is independent of wire length. Throughput is instead technology dependent. Equation (9) suggests an increase in throughput as technology scales. A plot of throughput versus wire length for different technologies can be found in Figure 6.

$$Throughput\big|_{PRC} = \frac{1}{T_{FF} + t_{RC}(L/N_{opt})} = \frac{1}{T_{FF} + t_{RC}\sqrt{C_{eff}V_{DD}^2 T_{FF}/e_{RC}t_{RC}}} \qquad (9)$$

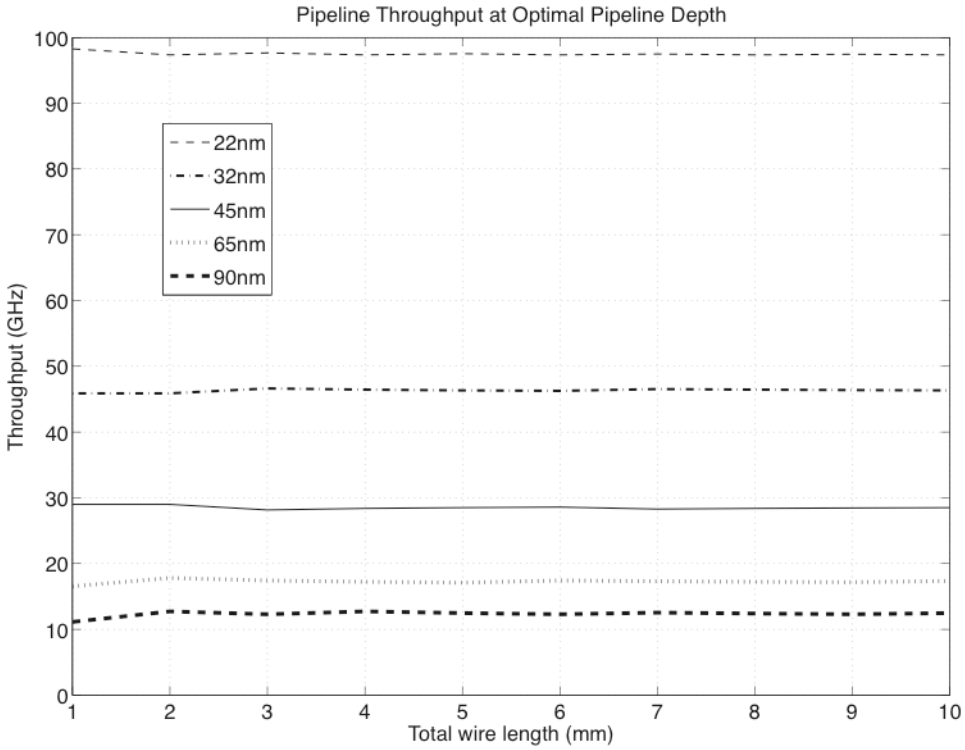

Figure 6. Throughput at Min-Energy/Throughput Objective

4.7 Evaluation of Energy at Optimal Pipeline Depth

Substitution of N_{opt} in our bit energy expression yields Equation (10). Since length-normalized repeated RC wire energy decreases as technology scales, given a wire length, optimized pipeline energy decreases as technology scales. Note that (10) also reveals an expression for the length-normalized energy of pipelined repeated RC wire optimized to our min-energy/throughput objective function. Figure 7 shows pipeline energy per bit versus wire length as technology scales.

$$Energy\big|_{PRC} = N_{opt} C_{eff} V_{DD}^2 + e_{RL} L = \left(\sqrt{\frac{e_{RC} t_{RC} C_{eff} V_{DD}^2}{T_{FF}}} + e_{RC} \right) L = e_{PRC} L \qquad (10)$$

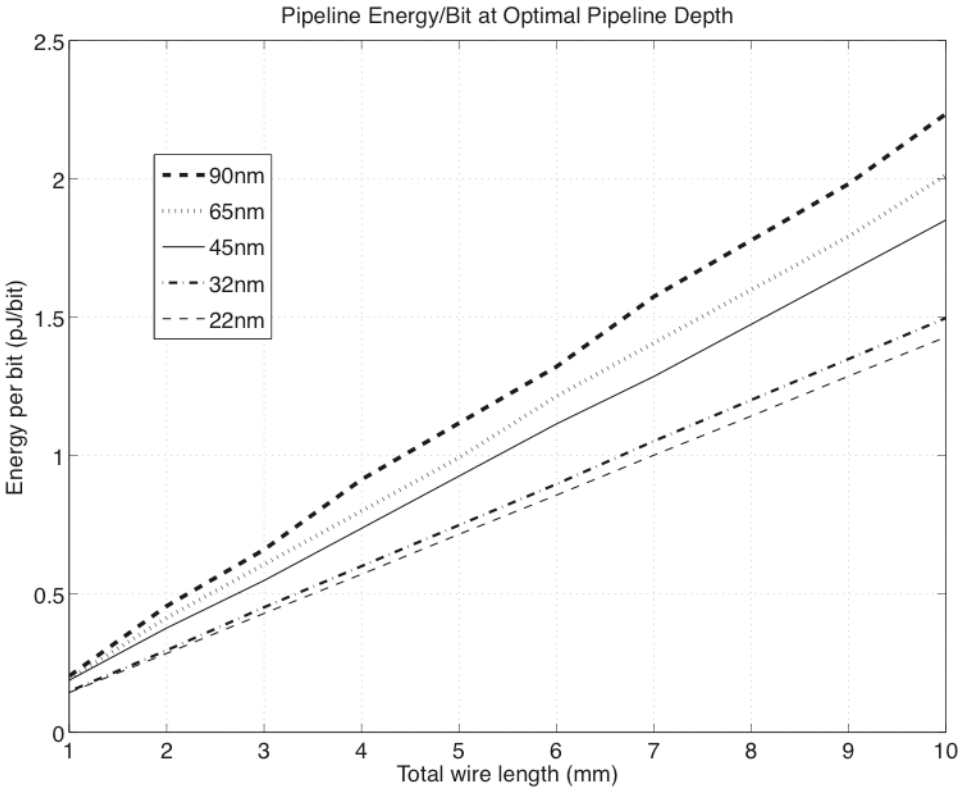

Figure 7. Pipeline Energy/Bit at Min-Energy/Throughput Objective

5. Conclusion

This study presents a new optimization strategy for pipelining long global interconnects comprising repeated RC wire. Pipelined interconnects are designed to remove such interconnect paths from the critical path. Systems tolerant of inserted pipeline latency can benefit from improved throughput and reduced energy. Pipeline depth may be modified to achieve an optimal tradeoff between energy and throughput. Optimal pipeline depth is proportional to wire length, and increases as technology scales. At the point of optimal pipeline depth, the energy-delay product consumed on repeated RC wire segments and the pipeline flip-flops is equal.

This study has practical implications, including the potential to make more energy-efficient use of costly pipeline structures. Additionally, the increased pipeline throughput made possible by flip-flop insertion can be used to reduce global routing congestion. Since optimal pipeline depth grows rapidly as technology scales, this technique will become increasingly practical and important.

References

[1] Intel Corporation (2009, Dec.). Moore's Law: Made Real by Intel Innovation. [Online]. Available: http://www.intel.com/technology/mooreslaw/

[2] Nanoscale Integration and Modeling Group, Arizona State University, Predictive Technology Model. [Online]. Available: http://www.eas.asu.edu/~ptm/

[3] ITRS. International Technology Roadmap for Semiconductors, 2008 edition. [Online]. Available: http://www.itrs.net/Links/2008ITRS/Home2008.htm

[4] Deodhar, V.V.; Davis, J.A., "Voltage scaling and repeater insertion for high-throughput low-power interconnects," *Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium on*, vol. 5, pp. 349-352, May 2003.

[5] Deodhar, V.V.; Davis, J.A., "Optimization of throughput performance for low-power VLSI interconnects," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 13, no .3, pp. 308-318, March 2005.

[6] Magen, N.; Kolodny, A.; Weiser, U.; and Shamir, N., "Interconnect power dissipation in a microprocessor," in *SLIP '04: Proc. of the Int. Workshop on Syst. Level Interconnect Prediction*, pp. 7-13, 2004.

[7] Carloni, L.; Sangiovanni-Vincentelli, A.L., "On-chip communication design: roadblocks and avenues," *Hardware/Software Codesign and System Synthesis, 2003. First IEEE/ACM/IFIP International Conference on*, pp. 75-76, Oct. 2003.

[8] Agarwal. V.; Hrishikeesh, M.S.; Keckler, S.W.; and Burger, D., "Clock rate versus IPC: The end of the road for conventional microarchitectures," in *27th Annual Intl. Symposium on Computer Architecture*, pp. 248-259, June 2000.

[9] Cocchini, P., "Concurrent flip-flop and repeater insertion for high performance integrated circuits," *Computer Aided Design, 2002. ICCAD 2002. IEEE/ACM International Conference on*, pp. 268-273, 10-14 Nov. 2002.

[10] Cocchini, P., "A methodology for optimal repeater insertion in pipelined interconnects," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 22, no. 12, pp. 1613-1624, Dec. 2003.

[11] Jingye Xu; Chowdhury, M.H., "Latch Based Interconnect Pipelining For High Speed Integrated Circuits," *Electro/information Technology, 2006 IEEE International Conference on*, pp. 295-300, May 2006.

[12] Ling Zhang; Hongyu Chen; Bo Yao; Hamilton, K.; and Chung-Kuan Cheng, "Repeated On-Chip Interconnect Analysis and Evaluation of Delay, Power, and

Bandwidth Metrics under Different Design Goals," *Quality Electronic Design, 2007. ISQED '07. 8th International Symposium on*, pp. 251-256, March 2007.

[13] Seth, V.; Min Zhao; and Jiang Hu, "Exploiting Level Sensitive Latches in Wire Pipelining," *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*, pp. 283-290, Nov. 2004.

[14] Horowitz, M.; Alon, E.; Patil, D.; Naffziger, S.; Rajesh Kumar; Bernstein, K., "Scaling, power, and the future of CMOS," *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pp. 7-15, Dec. 2005.

[15] Horowitz, M.; Stark, D.; and Alon, E., "Digital circuit design trends," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 757-761, April 2008.

[16] Yulei Zhang; Xiang Hu; Deutsch, A.; Engin, A.E.; Buckwalter, J.F.; and Cheng-Kuan Chen, "Prediction of High-Performance On-Chip Global Interconnection," in *SLIP '09: Proc. of the 11th Int. Workshop on Syst. Level Interconnect Prediction*, pp. 61-68, 2009.