Timing Correction and Optimization with Adaptive Delay Sequential Elements

Abstract

This paper introduces the Adaptive Delay Sequential Element (ADSE). ADSE are simple registers that use nonvolatile, floating-gate transistors to tune their delay characteristics. We propose ADSE for correcting timing violations, as well as performance optimization. This paper presents ADSE system architectures, tuning methodologies, and shows data from fabricated ADSE circuits. It discusses the die-area impact of using ADSE, and presents experimental results that demonstrate the correct operation of simple ADSE circuits. It shows how to use ADSE to correct circuit timing violations, as well as optimizing circuit speed by correcting clock skew.

1. Introduction

Most modern digital circuits rely on synchronous design techniques using sequential elements to synchronize logic operations. The longest delay path between two sequentially adjacent elements limits the maximum operating frequency of the circuit. This delay includes the time to perform the logic operation, the setup and hold time of the flip-flops, clock skew and jitter, and interconnect delays between gates. These delays further depend on circuit layout, process parameters, circuit operating conditions, and even details of the input waveforms. Even when the die layout is complete, circuit parameters such as capacitance, resistance, and inductance as well as transistor behavior can only be modeled with limited accuracy; these characteristics also vary with voltage and temperature. Furthermore, process parameters used in estimating circuit parameters vary statistically in each manufacturing run. The complexity and interactions among these factors make exact timing design of digital circuits an increasingly difficult task, and even with proper design, some percentage of manufactured circuits will fail to meet timing specification or to function at all. We propose ADSE as a method for fine-tuning circuit timing after fabrication. ADSE allows correcting post-fabrication timing errors. ADSE does not correct timing variations due to voltage and temperature changes or noise. However, ADSE does allow post-fabrication adjustment of design margins to accommodate these variations.

ADSE are also powerful optimization instruments and allow efficient implementation of optimal clock skews. The use of intentional clock skew to optimize circuit performance is well known in the literature and was first formulated by Fishburn [1]. Similar ideas have been discussed as cycle stealing [2] and their equivalence to retiming has been illustrated [3]. The solutions to these optimization problems comprise a set of clock skew values for all flip-flops in the circuit. These skews are implemented by adding delay cells to the clock distribution network and/or manipulating the detail routing of the clock signal during the design phase [4]. Unfortunately, adding delay cells increases circuit area and power consumption, and disturbs overall cell placement. Because optimization algorithms require an initial placement for a meaningful estimation of combinatorial delays, disturbing this placement by adding additional delay cells changes the optimization assumptions, and can lead to sub-optimal or even non-functional solutions. Furthermore, precisely manipulating clock routing is often difficult in automated routing tools and, when possible, is wasteful of routing resources. ADSE electrically tunes internal clock delays and obviate the need for using delay cells or clock routing manipulation.

ADSE adjust their timing delays by embedding floating-gate synapse transistors [5] within the sequential cell structure, and using analog values stored on these floating-gate structures to tune the

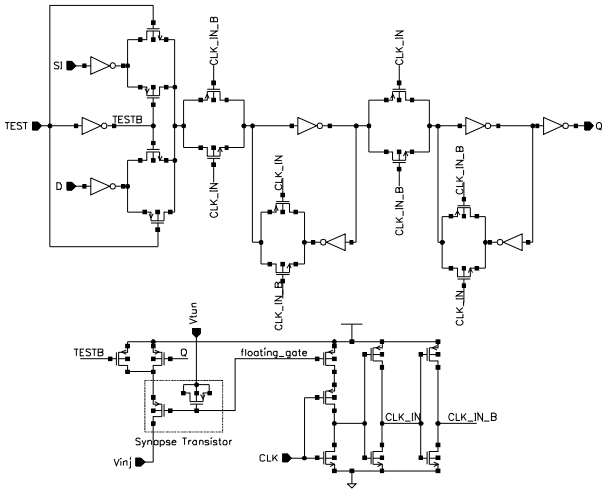

Figure 1. Adaptive delay implicit-pulsed master-slave static flip-flop. The master-slave structure occupies the upper half and the adaptive clock generator the lower half.

circuit delays. The tuning can be performed using minimal external control, without dedicated access to each floating gate.

1.1. Synapse Transistors

A synapse transistor is a four-terminal MOS device comprising two *p*FETs with a common floating gate; the floating-gate charge represents a nonvolatile analog weight. Synapse transistors have three unique attributes: (1) Nonvolatile analog weight storage, (2) locally computed bidirectional weight updates, and (3) simultaneous memory reading and writing. One *p*FET adds electrons to the floating gate using impact-ionized hot-electron injection (IHEI) [6], and allows reading the stored value by measuring the FET's channel current. The second *p*FET has shorted drain and source, and removes electrons from the floating gate by means of electron tunneling [7]. Researchers have used synapse transistors in applications such as weight storage in a learning array [8], trimming a digital-to-analog converter [9], nulling input offsets in a capacitive-feedback operational amplifier [10], and vector quantization [11]. ADSE use synapse transistors to perform timing adjustments in mainstream digital designs. Synapse transistors require a ~70A oxide thickness for reliable charge storage, however ADSE can be fabricated in smaller geometries, where dual gate oxides required for IO pads, make synapse transistors practical.

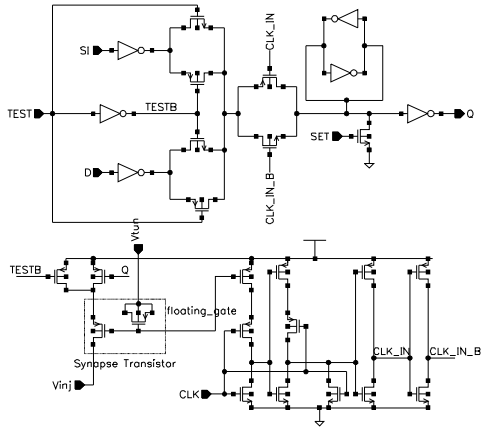

Figure 2. Adaptive delay explicit-pulsed static flip-flop. The adaptive pulse generator on the lower half generates pulses that sample the input. The flip-flop state is preserved by positive feedback.

2. Adaptive Delay Sequential Elements

There exist many different static and dynamic sequential digital circuit elements [12] and adaptive versions of many or perhaps all of them are possible. Here we present only two representative examples of ADSE.

2.1. Implicit-pulsed structures

Figure 1 shows an Adaptive-delay Implicit-Pulsed Master-slave Static (AIPMS) flip-flop. AIPMS is a scan testable, negative edge-triggered flip-flop with an adaptive clock generator. Implicit-pulsed static flip-flops utilize master-slave structures to sample their input on a clock edge. As a result, they have a positive setup time, zero or negative hold time, and a relatively large CLK-Q delay. Consequently, circuits containing master-slave structures are relatively immune to hold violations. We use a synapse transistor in the adaptive clock generator to control flip-flop setup time.

We adjust the setup time by changing the delay between the input clock and the flip-flop's internal clocks. The charge on the floating gate controls the current sourced by the first driver stage and hence the delay to the internal clock edge, delaying he negative clock edge.

2.2. Explicit-pulsed structures

Figure 2 shows an Adaptive-delay Explicit-pulsed Static (AEPS) flip-flop. Explicit-pulsed

static flip-flops are among the fastest and most energy efficient sequential elements. Recently, a dual edge-triggered version of this structure has been proposed for high performance microprocessors running at 3GHz [13]. Whereas implicit-pulsed structures create a transparency window by means of transistor stacks or transmission gates, explicit-pulsed flip-flops generate an explicit clock pulse to sample the data. All timing constraints of explicit-pulsed clocks are directly or indirectly related to the pulse width of the internal clock. We use synapse transistors to adjust this pulse width. The floating-gate charge determines the delay between the input clock and its delayed version, and, hence, the width of the internal clock pulse.

2.3. Tuning Operations

The timing characteristics of ADSE are related to the synapse transistor's floating-gate voltage. We cannot measure the floating-gate voltage directly, but, when the injector *p*FET is saturated, the floating-gate voltage can be inferred from the *p*FET's drain current. We will refer to this drain current as the tuning current. Tuning currents are only active during tuning operations and are read and controlled by accurate off-chip instruments. During normal operation no tuning current flows through injector *p*FET. All ADSE use similar methods to tune their floating-gate voltages. We will explain the tuning operations of AIPMS presented in Figure 1.

During normal operation, we set V_{inj} and V_{tun} to V_{dd}, and TEST to ground, thereby holding the floating-gate charge constant. We increase the delay by applying a large positive voltage (~10V) to the V_{tun} terminal and extracting electrons from the floating gate. We decrease the delay in two distinct modes, GLOBAL and SELETIVE. We use GLOBAL mode to initialize all flip-flops to their minimum delays. We use SELECTIVE mode to correct timing violations and optimize circuit timing. In the GLOBAL mode we assert the TEST signal and set V_{inj} to roughly –1V, causing injection in all floating gates. In the SELECTIVE mode, we set the TEST signal to ground and connect V_{inj} to roughly –1V, causing injection in flip-flops that hold a 0 state. When a cell is selected by holding a 0 state, we can measure its tuning current by applying 1V at V_{inj} and reading the resulting tuning current without causing electron injection. The ability to read tuning currents between injection steps allows for closed-loop stepping of tuning currents and delays. Accurate control of delay is therefore possible by choosing a small enough step size in spite of non-linear dependence of delay on floating-gate voltage.

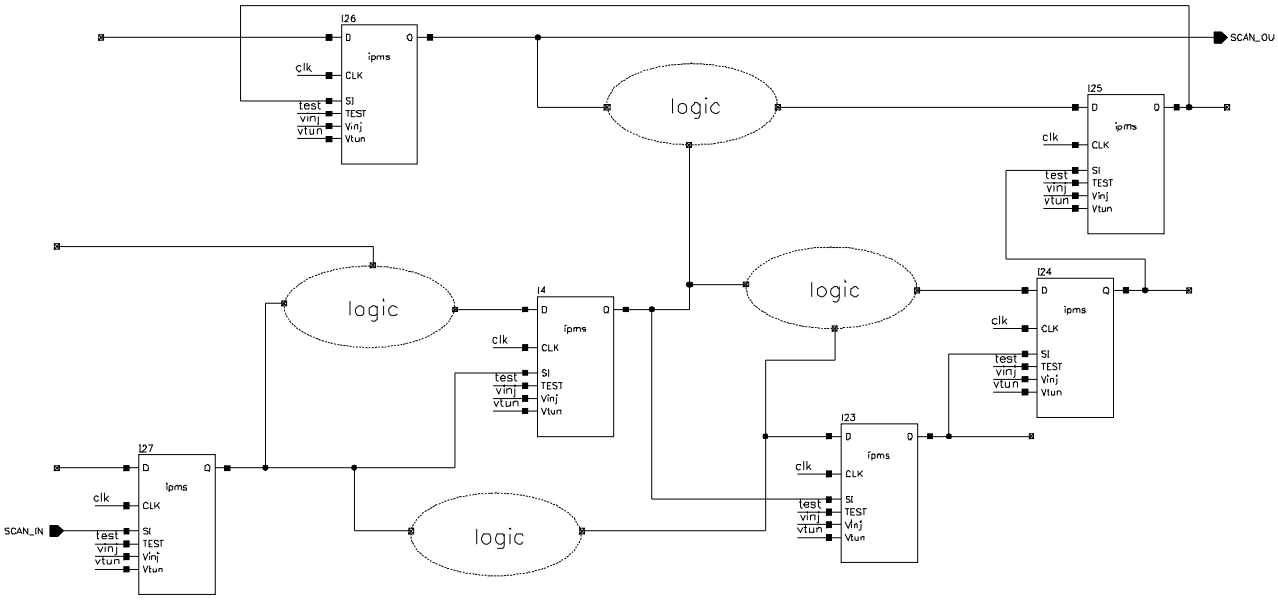

Figure 3. System architecture for embedding ADSE. Flip-flop outputs and scan input are connected together to form tuning chains. When circuit is in scan mode tuning vectors can be shifted in to select floating gates of individual ADSE for injection. Injection and tunneling voltages (vinj and vtun) are global signals.

3. System Architecture and Impact

A major issue in using individually tunable elements in large-scale circuits is the question of addressing, i.e. how to select an element to tune when it is embedded in a large circuit. Previous efforts in on-chip timing adjustment are restricted to groups of flip-flops or regions of the chip since they require significant amounts of additional circuitry [14], [15], [16].

As we have seen, the value stored in each ADSE can control its own internal charge injection. Consequently, we can use tuning chains that combine the use of ADSE and scan Design-For-Test (DFT) technique [17]. To decrement individual cell delays we use the tuning chains to set the value stored in each flip-flop by shifting in tuning vectors. A tuning vector V_i comprises a sequence of bit vectors that hold a "0" for each selected flip-flop and a "1" for unselected flip-flops. V_i also specifies an injection voltage V_{inj} and injection pulse width N_i:

$$V_i = (b_{i0}, b_{i1}, \cdots, b_{im}, V_{inj}, N_i) \qquad b_{ij} \in \{0,1\}, \quad V_{inj}, N_i \in \Re$$

To increment flip-flop delays we use a global tunneling signal that connects the tunneling terminals of all ADSE to a tunneling pad. The delay increment is global and introduces a positive skew on all flip-flops. The combination of global delay increment and selective delay decrement enables us to set individual flip-flop delays without requiring additional addressing circuitry. The system architecture is shown in Figure 3.

3.1. Correcting timing violations

In most circuits only small subset of all timing paths are near critical. Modeling approximations or manufacturing uncertainties can cause delay variations in these near-critical paths that lead to timing violations that render the entire circuit non-functional. We can correct these timing violations in circuits using AIPMS. We first set all ADSE delays to their minimum value using the GLOBAL injection mode. This happens automatically since injection is a self-limiting process. We then detect timing violations on near-critical paths by applying timing test vectors. Assume a failure on flip-flop F is detected. We build a tuning vector V_i that has "0"s corresponding to all predecessors of F and "1"s in all other positions. We tunnel all cells for a predefined period to introduce a positive skew. We apply V_i which causes charge injection and delay decrement in all predecessors of F. The flip-flops that are not selected by V_i hold a "1" that blocks their charge injection. After applying V_i, we use the same timing test vector to examine the violation again. We repeat this process until the timing violation is corrected. If the number of timing violations is large or many correlated paths exits, it is necessary to solve the optimal skew problem as discussed in section 5.

Table I. Area comparison of adaptive and non-adaptive flip-flops on ISCAS89 benchmark circuits.

Circuit	#FF	IPMS	AIPMS	%diff	EPS	AEPS	%diff
s400	21	2149	2511	16.8	2149	2471	15
s526	21	2555	2948	15.4	2596	2858	10.1
s444	21	2186	2596	18.7	2228	2511	12.7
s382	21	2072	2471	19.2	2108	2428	15.1
s526n	21	2555	2948	15.4	2596	2858	10.1
s953	29	3923	4465	13.8	3978	4354	9.4
s838	32	4465	5043	12.9	4465	4980	11.5
s1423	74	8064	9389	16.4	8135	9227	13.4
s5378	179	24140	27219	12.8	24401	26943	10.4
s9234	211	41367	45201	9.3	41708	44659	7.1
s15850	534	79009	88186	11.6	79481	87193	9.7
s13207	638	73029	84506	15.7	73697	83051	12.7
s38584	1426	194149	219610	13.1	195629	216479	10.7
s38417	1636	196371	225097	14.6	197859	221555	12
s35932	1728	182553	213371	16.9	183988	209156	13.7

3.2. Area impact

ADSE can simply replace regular scan flip-flops in a design without requiring additional addressing and support circuitry. The tunneling and injection signals require only two additional global signals and IO pads. The tunneling signal requires no special routing provision, since in normal operation it is tied to Vdd and during tuning when tunneling voltages of ~10V are applied the magnitude of tunneling current is small (~nA). The injection signal routing has to accommodate the total tuning current in global injection mode. This is on the order of 1μA times the number of flip-flops in the circuit. ADSE layouts are larger than their non-tunable counterparts since they contain more transistors and tunneling devices require wider spacing. However, chip area is usually constrained by routing feasibility rather than transistor area. Furthermore, even though ADSE are larger than their non-adaptive counterparts, their impact on the overall chip area is much smaller since they constitute only a fraction of the total area. To show the impact of using ADSE on overall block-area we use the ISCAS89 benchmark circuits [18] and design-rule-correct, fabrication-tested layout areas from a typical 0.35-micron cell library. We

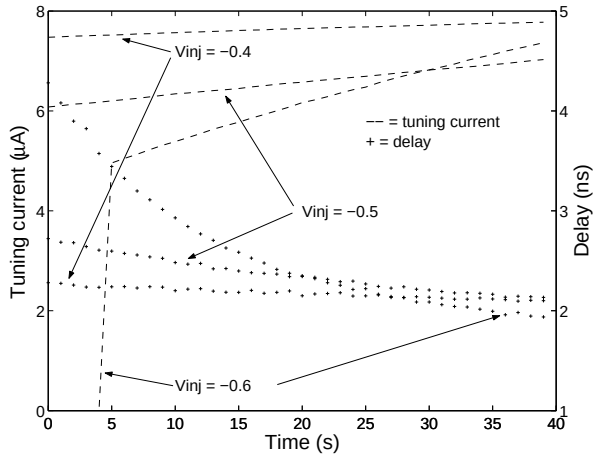

Figure 5 Delay and tuning current variation with injection pulse width for the AIPMS flip-flop in a 0.35μ process.

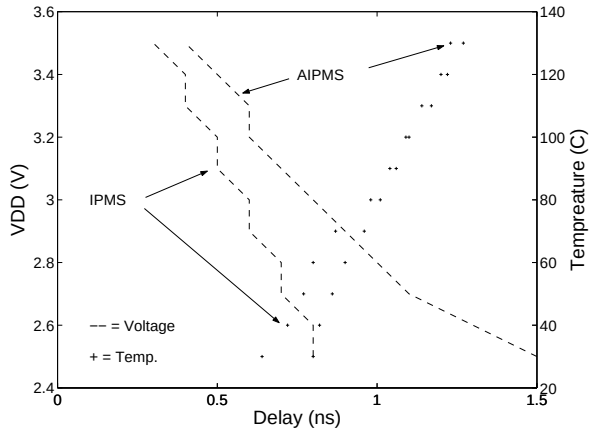

Figure 4 Delay variations with supply voltage and temperature for AIPMS and IPMS flip-flops in a 0.35μ process.

map each benchmark circuit to this library and calculate the total cell area. The flip-flops are first mapped to regular Implicit-Pulsed Master-slave Static (IPMS) flip-flop and then to its adaptive version (AIPMS). We repeat these experiments with Explicit-pulsed Static (EPS) flip-flop and its adaptive counterpart (AEPS). Table I shows the layout area in square microns and the percentage difference. We would like to point out that even though AIPMS area is larger than IPMS by 39%, the overall area increase in benchmark circuits is between 9-17%. The area of AEPS is larger than EPS by 32% while the area increase of the circuits is 7-15%.

Table II. Timing optimization Algorithm

```
Optimize C ()
    Estimate delays and form LP;
    Solve LP to get skew vector S;
    Convert S to I_opt ;
    Initialize all flip-flops by FN-tunneling;
    For each flip-flop Fi in C
        Generate tuning vector V_i = (v_ij) = { 0  i = j
                                              { 1  i ≠ j
        Shift-in V_i;
        While I_i ≠ I_i^opt
            Lower V_inj for δt;
            Measure I_i;
        end;
    end;
```

4. Experiments

We fabricated ADSE test circuits in a 0.35μ process. Figure 5 shows variations in CLK-Q delay and tuning current of AIPMS flip-flop with injection pulse width. Clearly, lower injection voltages correspond to larger adjustment steps for the same pulse width. We would like to note that when the injector transistor is just below threshold and the tuning current is too small to reliably measure, injection can still take place and the transistor eventually reaches threshold. We have also verified the clock pulse width adjustments in AEPS but are unable to present the data due to space limitations.

Also, we studied ADSE sensitivity to supply voltage and temperature variations. Figure 4 shows that supply voltage and temperature sensitivity of ADSE are comparable to their non-adaptive counterparts.

We verified that we can change tuning currents with a resolution of less than 1nA by using short pulses (~ms) to inject small amounts of charge onto the floating gates. Our experimental data show that for channel currents greater than 5μA a 1nA change in tuning current corresponds to an output delay change of 1ps or less. Direct delay measurements at this resolution, however, are not possible with our current experimental setup.

5. Timing Optimization

If many correlated timing violations exist or maximum clock frequency is desired we need to solve an optimal skew problem as follows. Given a circuit C estimate all combinatorial path delays between its sequential elements after placement

and formulate the optimal clock skew problem as a linear program as suggested by Fishburn [1]. This problem can be efficiently solved by methods suggested in [19] to provide a vector of clock skew values, S, for all flip-flops. Translate this vector to equivalent tuning current values, I^{opt}, using a table lookup method. Initialize the ADSE by FN-tunneling and use tuning vectors to set each flip-flops' tuning current as prescribed in I^{opt}. Table I summarizes the steps for timing optimization with ADSE.

6. Conclusions

Adaptive Delay Sequential Elements (ADSE) embed floating-gate synapse transistors within sequential elements. ADSE can be used to solve a range of timing problems from correcting a few timing violations to implementing an optimal clock skew solution without the need for delay cell insertion or manipulating clock routing. We have presented two sample ADSE circuits and the system architecture. We discussed the ADSE tuning operations and demonstrated the impact of using ADSE on block area. We presented results from tuning experiments using our sample circuits fabricated in a 0.35µm process, and compared their sensitivity to voltage and temperature variations to their non-adaptive counterparts. ADSE can be fabricated in smaller geometries where dual gate oxides required for IO pads, make synapse transistors practical.

References

[1] J. P. Fishburn, "Clock skew optimization," IEEE Trans. Computers, vol. 39, no. 7, pp. 945-951, 1990.

[2] E. G. Friedman, "Clock Distribution Networks in VLSI Circuits and Systems," Piscataway, NJ, IEEE Press, 1995.

[3] S. S. Sapatnekar and R. B. Deokar, "Utilizing the retiming skew equivalence in a practical algorithm for retiming large circuits," IEEE Trans. Computer-Aided Design of Circuits and Systems, vol. 15, no. 10, pp. 1237-1248, 1996.

[4] J. L. Neves and E. G. Friedman, "Design methodology for synthesizing clock distribution networks exploiting nonzero localized clock skew," IEEE Trans. VLSI Systems, vol. 4, no. 2, pp. 286-291, 1996.

[5] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "Floating-gate MOS synapse transistors," in T. S. Lande (ed.), Neuromorphic Systems Engineering: Neural Networks in Silicon, Boston, MA: Kluwer Academic Publishers, pp. 315–337, 1998.

[6] P. Hasler, Foundations of Learning in Analog VLSI, Ph.D. thesis, Department of Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 1997.

[7] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO2," J. of Applied Phys., vol. 40, no. 6, pp. 278–283, 1969.

[8] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A floating-gate MOS learning array with locally computed weight updates," IEEE Trans. Electron Devices, vol. 44, no. 12, pp. 2281–2289, 1997.

[9] M. Figueroa, J. Hyde, T. Humes, and C. Diorio, "A floating-gate trimmable high-resolution DAC in standard 0.25µm CMOS," Proc. IEEE Nonvolatile Semiconductor Memory Workshop, Monterey, CA, pp. 46–47, 2001.

[10] P. Hasler, B. A. Minch, and C. Diorio, "An autozeroing floating-gate amplifier," IEEE Trans. Circuits and Systems II, vol.48, no. 1, pp. 74–82, 2001.

[11] D. Hsu, M. Figueroa, and C. Diorio, "A silicon primitive for competitive learning," Advances in Neural Information Processing Systems 13, T. K. Leen, T. Dietterich, and V. Tresp, eds., MIT Press, pp. 713–719, 2001.

[12] V. Stojanovic, V. G. Oklobdzija, "Comparative analysis of master-slave latches and flip-flops for high-performance and low-power systems," IEEE Journal of Solid-State Circuits, vol. 34, no. 4, pp. 536-548, 1999.

[13] J. Tschanz, S. Narendra, Z. Chen, S. Borkar, M. Sachdev, V. De, "Comparative delay and energy of single edge-triggered & dual edge-triggered pulsed flip-flops for high-performance microprocessors," Proc. IEEE International Symposium on Low Power Electronics and Design, pp. 147-152, 2001.

[14] M. Mizuno, M. Yamashina, K. Furuta, H. Igura, H. Abiko, K. Okabe, A. Ono, H. Yamada, "A GHz MOS adaptive pipeline technique using MOS current-mode logic," IEEE Journal of Solid-State Circuits, vol. 31, no. 6, pp. 784-791, 1996.

[15] S. Rusu and S. Tam, "Clock generation and distribution for the first IA-64 microprocessor," Proc. International Solid-State Circuits Conference, pp. 176-177, 2000.

[16] N. A. Kurd, J. S. Barkatullah, R. O. Dizon, T. D. Fletcher, P. D. Madland, "A multigigahertz clocking scheme for the Pentium 4 microprocessor," IEEE Journal of Solid-State Circuits, vol. 36, no. 11, pp. 1647-1653, 2001.

[17] M. Abramovici, M. A. Breuer, and A. D. Friedman, "Digital Systems Testing and Testable Design", New York, NY, Computer Science Press, 1990.

[18] F. Brglez, D. Bryan, K. Kozminski, "Combinational Profiles of Sequential Benchmark Circuits", Proc. IEEE International Symposium on Circuits And Systems, pp. 1929- 1934, 1989.

[19] R. B. Deokar and S. S. Sapatnekar, "A graph-theoretic approach to clock skew optimization," Proc. IEEE International Symposium on Circuits and Systems, pp. 1.407-1.410, 1994.