A 40nm Critical Path Monitor for the Detection of Setup and Hold Time Violations

Hernán Aparicio Pablo Ituero

***Abstract*—In the current context of strict low-power requirements, complex dynamic frequency and voltage scale systems try to constantly push the operating conditions of electronic chips to the lower bound that fulfills the performance requirements. Also, at test time of a synchronous electronic system, any occurrence of timing violations, especially hold time violations, must be identified, located and corrected. Critical path monitors serve these two purposes, they measure the delays where transients are produced in relation to the clock signal for the critical paths of the system. This work introduces a critical path monitor architecture that yields two configurable digital outputs: one for setup time violations, and another for hold time violations. The monitor directly senses the critical path, without the need to introduce synthesized replicas. The architecture has been validated in a 40nm commercial technology, it takes an area of 4028 μm^2 and it is very robust against PVT variations.**

I. Introduction

Extreme miniaturization of electronic circuits has brought uncountable benefits, but it has also exacerbated the uncertainties caused by variability. Fluctuations in the fabrication process introduce random changes in the physical structures that conform transistors and interconnections which affect their electrical characteristics. Degradation of devices caused by usage and the accumulation of radiation doses also increases with the shrinking of technology.

The compaction of the fabrication process has allowed the inclusion of billions of transistors on a single chip. This elevated level of logic density has carried along amplified power densities, and therefore dangerous temperature increments. As a result, energy consumption is one the key parameters that is kept under control during the electronic design process. Power supply levels —on which power consumption depends quadratically— are pushed to the minimum value that warranties safe operation. Actually, current Dynamic Voltage and Frequency Scaling (DVFS) techniques are constantly adapting the power supply and working frequency trying to achieve the minimum energy that safely fulfills the application demands.

Most current digital architectures are based upon the synchronous design methodology. In this methodology the system is divided into combinational circuits, that realize the logical and arithmetic operations and memory elements, namely latches or flip-flops, that store the results of the combinational elements after an edge in the systems clock signal. The longest combinational path in a system is called the critical path (CP) and it is the one that determines the clocking frequency. In order to correctly store the data, the memory elements require a stable incoming signal a certain time before the edge of the clock —setup time— and after the edge —hold time. Guaranteeing this stability has become more and more complicated in the current context of intensified variability. The delay of the combinational circuits is dependent on process, supply voltage and temperature (PVT) variations, it also fluctuates with aging and accumulative radiation effects. DVFS control systems need to be aware of all these delay uncertainties to find the optimal working point.

Hold time violations —caused by too fast combinational circuits— cannot be corrected once the circuit has been fabricated as they are independent on the frequency of the system. For this reason, computer-aided design tools have been designed to assure that these violations do not occur. However, current process fluctuations require extensive prototyping test to detect any of these problems.

In this scenario, monitoring the delay of combinational circuits, especially those that fix the clocking frequency, appeared as a natural and necessary solution. This type of monitor is called critical path monitor (CPM) and is employed at run-time by dynamically adaptable systems, such as DVFS systems, and at test-time by test engineers to detect setup and hold time violations. In the architectures of this type of monitors there have been two tendencies: those that measure the delay directly from the CPs and those that replicate the CP and measure the delay from the replicas. The former approach runs the risk of further increasing the delay of the CP with the inclusion of new logic, thus degrading the performance of the whole system. The latter can incur in accuracy errors as the replicated circuits are not subject to the same variability sources as the real ones.

This work introduces a new CPM architecture that takes measurements from the actual CP and detects signal changes both before the clock edge (setup time violations) and after the clock edge (hold time violations) providing a configurable digital thermometer code. The contributions of the work are the following:

- Introduction of a novel CPM architecture based on a time-to-digital converter centered around the rising edge of the clock signal that is specially efficient in terms of area overhead.

- The CPM employs the real CP, however the delay impact is very reduced, as it just increases the fanout of the last CP stage by an inverter.
- The CPM provides information about both setup time and hold time violations, so that it can be employed at both run- and test-time to detect and identify hazardous timing errors.
- The CPM can be configured and adapted during run-time to the need of the DVFS system.

The CPM has been implemented and validated in a 1.1 V 40 nm commercial CMOS process. The system is characterized by the following features:

- The correct functioning has been characterized for a temperature range between $-40\,^{\circ}\mathrm{C}$ and $125\,^{\circ}\mathrm{C}$.
- The digital output is expressed in terms of buffer delays, independent of PVT variations.
- Maximum layout area: 4028 $\mu\mathrm{m}^2$.

The rest of the paper is organized as follows. Section II describes previous works presented in the literature. In section III we show the architecture, the design concept and the benefits of the CPM proposed. Section IV we present the characterization results of the CPM and its behavior against process, power supply and temperature variations, along with the comparison with previous works. Finally, section V draws some concluding remarks.

II. Previous works

There are two different kinds of approaches to deal with the critical paths in the scientific literature. The most common is the use of a synthesized CP to measure the delay present in the actual CP. The other approach found in the literature employs the real CP to measure the delay.

Both methods have advantages and disadvantages. The use of critical paths replicas incurs in area overhead and just obtains an approximation of the real delay. Besides, the aging processes are highly dependent on the workload and temperature. On the other hand, to perform a measure on the real CP, it is necessary to introduce new logic in the CP, thus it is potentially translated into a global frequency reduction.

A. Synthesized Critical Paths

The design of synthesized CPs avoids to add any extra logic in the real critical path which results in better performance of the systems. Better performance does not means better security in time deadlines. The use of CP replicas emulate the operation of the real CP that is highly dependent on workload and environment conditions. These conditions vary according to the location of the critical path inside the integrated circuit (IC). As known, the voltage supply and temperature are not constant over the IC and each CP presents different workloads which results in different aging rates. Therefore it is very difficult to emulate all these conditions in the synthesized critical paths.

Apart from the difficulty to emulate the real operating conditions of the actual critical path, current electronic systems have multiple CPs and each new technology introduces more uncertainties variations with high spatial correlation. This implies the need for new replicas relatively close to each CP which is translated into an increase in area and power consumption.

CPMs based on synthesized critical path focus their design effort on fine-tuning the replicas so that they accurately reproduce the real behavior of the actual CP. This strategy is convenient and adequate if performance is the principal concern and no logic can be added at the real CP.

As an example, in a previous CPM work [1], the synthesized CP was connected in parallel with the real CP (16 X 16 multiplier) and the phase error between the two paths indicates the delay. A critical path emulator (CPE) was used by a DVS system to adjust the supply voltage according requirements. At different supply voltages, the phase error produces different delays and the system frequency can be adjusted to reduce time margin requirements to enhance energy efficiency.

In a different strategy [2], five parallel paths through nand4, nor3, adder, wire and transmission gates were synthesized. These paths were selected because they have different delay times relative to process variation, temperature and voltage and were used to emulate the behavior of the CPs under PVT variations and aging effects. A time-to-digital converter measures the delay after an edge is launched into the synthesized paths. This method allows to consider local variations that cause delays in the critical path and can identify the slowest path among the five parallel paths.

In other work [3], an algorithm for the design of accurate critical path replicas (CPRs) was introduced. Multiple parallel CPRs are used in a CPM to improve the energy-delay product (EDP) in DVFS systems. The system samples the delay of the CPRs at every clock cycle, if the delay is faster than the clock cycle, the system increases the clock frequency. On the other, hand if the delay is slower than the clock cycle, the frequency is decreased by the system.

Another proposal [4] presents a statistical critical path monitor that senses local path timing variation. The CPM has 15 blocks to measure statistics for different type of path. Each path is formed by 255 replicas of the path under test. The aim of this sensor is to provide information about local variability of digital paths. It can be used to adjust the supply voltage or frequency margin to a particular local variability.

B. Real Critical Paths

The approach which uses the real critical path to measure the critical path delays has the disadvantage to introduce extra logic in the CP reducing the system performance. The most popular approach [5] is implemented using doubled sequential logic, so the single Flip-Flop (FF) at the end of the CP is substituted by a doubled FF. One FF is controlled by the clock frequency system and the other by a delayed clock frequency to create a time window detection. If the doubled FF exhibits the same output, no delay violation is present in the CP and if the FF stores different values a delay error is detected. A XOR gate is used to compare the register values present in the two FF and flag a error signal if a delay is detected.

This architecture allows to reduce the supply voltage or increase the system frequency keeping the time deadlines. To do that, each time a measurement is taken, the system pipeline must be stalled, thus reducing the performance. This method can not predict or anticipate time failures. It only can detect data errors and correct them before writing to memory.

III. Proposed CPM

In this section we describe the design concept, the benefits and the architecture of the proposed CPM. Its architecture has two main blocks: the delay clock buffer and the time-to-digital converter. The designed CPM realizes the measurements employing the real CP of the system. The monitor receives the clock signal, CLK, and uses it to detect changes before the clock edge; it also produces a delayed clock signal, CLK_d, which is used to detect changes after the clock edge. The monitor yields two digital outputs employing one-hot encoding: out_s, relative to signal changes before the clock edge or setup time violations; and out_h, relative to signal changes after the clock edge or hold time violations. The monitor receives a 2-bit control signal S1,S0 that determine the extension of time before and after the clock edge to be monitored, A schematic of the proposed monitor is shown in Figure 1.

The architecture of the proposed CPM is formed by two main blocks as can be seen in Figure 1. The configurable delay block (CDB) is formed by four buffer chains and a multiplexer and produces the delayed version of CLK, . Each chain is composed by 6 buffers and the multiplexer controls the number of activated buffers chains. This configuration permits the variability of the setup and hold time detection window through the delay change in the CLK delayed signal.

The time-to-digital conversion block showed in Figure 2 is formed by the same number of buffers that compose the CDB. All buffers in the proposed CPM have the same size to expose the data from the critical path output to the same delay suffered by the CLK signal. The buffers delay the output signal of the critical path to allow it to be sampled and to identify potential hazardous delays in the CP. The CPM resolution is based on the delay caused by the buffers, therefore, this delay must be constant to obtain linear and reliable measurements. In order to have equal delays between the buffers, the rise and fall time of the buffers need to be equal or very close. The registers are connected to the output of the buffers. Its function is to store the buffers values at the end of the measurement. At the same time, the registers outputs are the inputs of the XOR gates which are used to convert the output into an edge position.

This configuration makes possible the monitoring of the setup time by the latches controlled by the clock signal and the monitoring of the hold time by the latches controlled by the delayed clock signal. The minimum detection time that this architecture can perceive is the delay of a single buffer. The setup and hold time detection window depends on how many buffers chains are activated. Varying S1 and S0, we can select between 5 slots of detection (one buffer chain activated), 10 slots, 15 slots and 20 slots (four buffer chain activated). If one buffer chain is activated, only the first five XORs outputs are considered in the analysis of the CP delays. In case of four buffer chains activated, all XORs outputs will give information about the CP delay. The CPM can be redesigned to achieve less slots of detections reducing area and decreasing the performance impact in the CP.

Due to the fact that we use a single buffer as a unit of delay to measure the delays in the critical path, there is no need of calibration because PVT variations will affect the combinational logic used in the CP and the CPM in the same way. The slot of detection from the CPM is based on buffers delays. So, to adapt this monitor to any circuit it is only necessary to resize the buffers from the CPM to adapt the buffer delay with the setup and hold times specifications or the safety margin time of the CPs.

Figure 3 shows how the CPM behaves in the presence of four different logic changes inside the setup and hold times considering one buffer chain activated. The slots of detection can be seen in Figure 3 and the XORs outputs indicate the slot where the logic change was detected. The monitor provides two one-hot encoding signals that show where the signal edge was produced in terms of buffer delays. This quantified information can be employed by DVFS systems or by test engineers not only to determine when an error has occurred, but also to have run-time information about how close the delay is to the unsafe zone.

Concerning the increase of delay caused in the CP, as seen in figure 2, the CPM loads the final stage of the CP with the input capacitance of an inverter. Even though this is the smallest capacitance increase that can appear in a digital circuit, it could suppose an unacceptable performance penalty in certain high-performance architectures. For an architecture implementing a DFVS system, this delay degradation is the minimum achievable if the measurement needs to be taken in the real CP.

IV. Charcterization Results and Comparison with Previous Works

This section presents the post layout simulation results of the proposed critical path monitor considering parasitics extraction. We have designed the CPM in a 1.1 V commercial 40nm CMOS process. The simulations and layouts have been carried out in the Cadence^{TM} environment. All the results come from transient simulations that cover temperatures from $-40\,^{\circ}$C to $125\,^{\circ}$C, all process corners and mismatch. The layout of the CPM can be seen in Figure 4. The structure in the lower part corresponds to the buffers and the multiplexer that constitute the CDB. The rest of the layout is formed by four equal main blocks. Each block is formed by six buffers, twelve latches and five XORs. The area of the complete sensor is 4028 μm^2. In case of use only one buffer chain (five slots of detection), the area can be reduced to 1225 μm^2 which means an area reduction about 70%.

The buffers used in this design present a delay of 170 ps in nominal conditions (typical corner, nominal power supply and temperature) and considering the layout parasitics. This establishes an upper bound for the error of the estimated value of the delay under these conditions. Parametric and Monte Carlo simulations in layout level were performed varying the temperature between $-40\,^{\circ}$C and $125\,^{\circ}$C and supply voltage

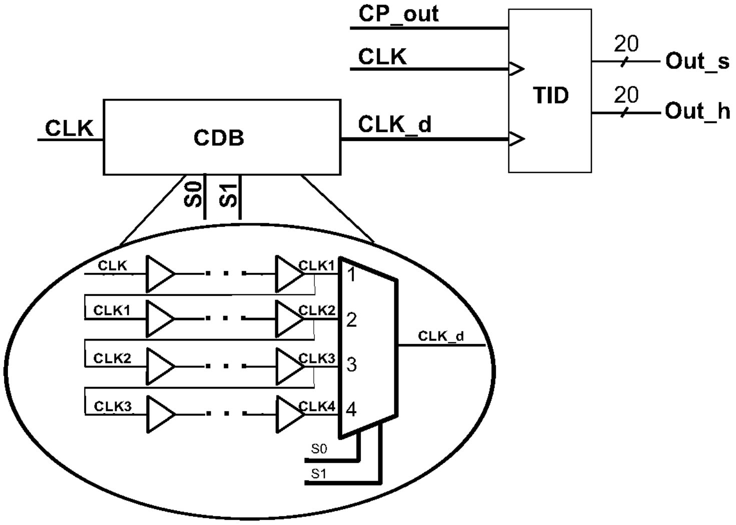

Fig. 1. Architecture of the proposed CPM.

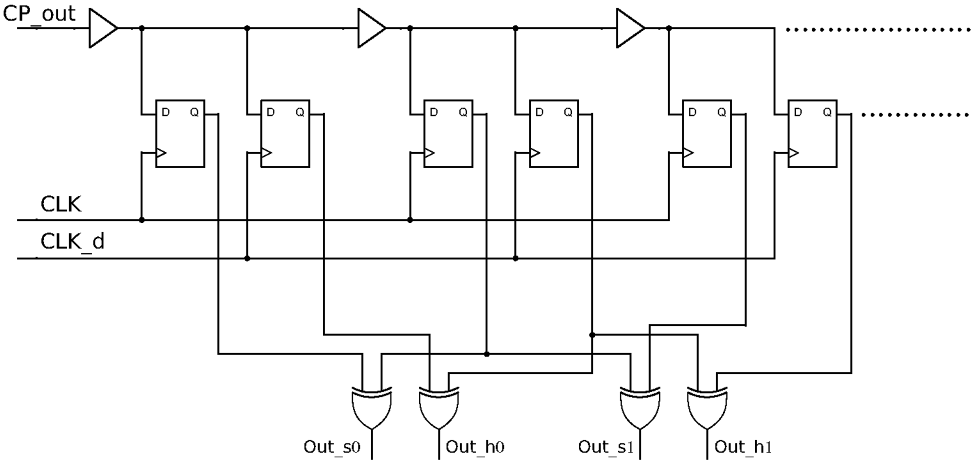

Fig. 2. Time to Digital Conversion Block.

between 0.99 V and 1.21 V. Figure 5 presents the minimum delay detected by the CPM under different process corners, temperature and power supply. As shown, the minimum delay that the CPM can detect is 280 ps in the worst case considering all process corners,temperature and power supply variations.

The buffers of the monitor are expected to suffer the same PVT variations as the CP, therefore the digital word corresponds to a quantized word, being the quantization step the buffer delay (a varying time unit). This means that the CPM informs of how close to the edge in terms of buffer delays a transient at the end of the CP has occurred.

For CP delays that are closer to the clock edge than a buffer delay, the CPM is not able to detect the transient with the flip-flops that employ the system clock. However, the flip-flops fed by the delayed clock signal detect the transient of these signals and they are flagged as hold time violations, activating the last bit of out_h.

Concerning the operating margins of the design, the maximum operating frequency depends on the percentage of the clock period that is required to be monitored and the number of quantization levels needed. For example, a system that monitors the second half of the period for setup time violations and the first half of the period for hold time violations where just five quantization levels are required. Under these conditions, just one buffer chain of the CDB is activated and for the nominal case half of the clock period equals 850 ps, thus the maximum frequency is in the order of 600 MHz.

Table I summarizes the main characteristics of the proposed CPM and compares it with previous works. As shown, most previous works are based on synthesized CP. The special characteristic of this approach is that the measurement is centered around the clock edge and that it provides quantized

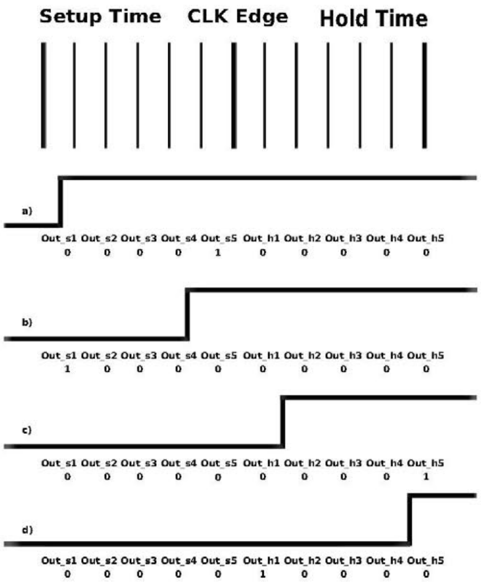
Setup Time
CLK Edge
Hold Time
a)
Out_s1 Out_s2 Out_s3 Out_s4 Out_s5 Out_h1 Out_h2 Out_h3 Out_h4 Out_h5
0 0 0 0 1 0 0 0 0 0
b)
Out_s1 Out_s2 Out_s3 Out_s4 Out_s5 Out_h1 Out_h2 Out_h3 Out_h4 Out_h5
1 0 0 0 0 0 0 0 0 0
c)
Out_s1 Out_s2 Out_s3 Out_s4 Out_s5 Out_h1 Out_h2 Out_h3 Out_h4 Out_h5
0 0 0 0 0 0 0 0 0 1
d)
Out_s1 Out_s2 Out_s3 Out_s4 Out_s5 Out_h1 Out_h2 Out_h3 Out_h4 Out_h5
0 0 0 0 0 1 0 0 0 0

Fig. 3. CPM Behaviour.

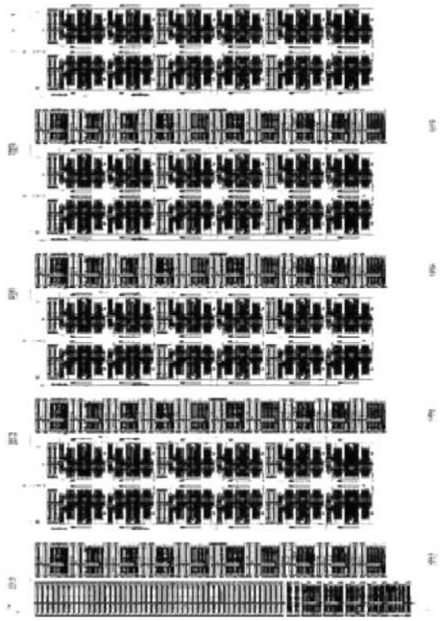

Fig. 4. CPM Layout

information about setup time violations but also about hold time violations. Although the commercial version of a fabricated chip should be free of hold time violations, detecting this hazards during test time is an arduous task that this CPM helps alleviate.

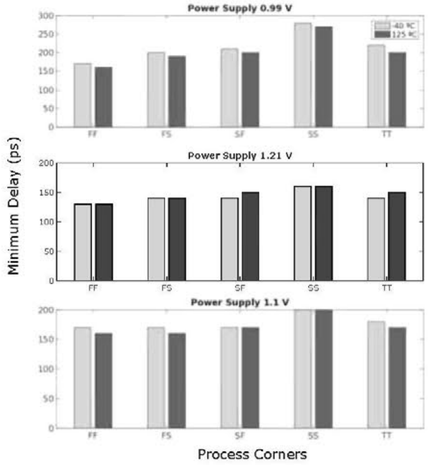

Fig. 5. Minimum delay detected.

V. Conclusions and Future Work

Combinational delays in digital electronic systems are nowadays subjected to several sources of variation. Aging and radiation effects join process, temperature and supply voltage fluctuations in the long list of physical and electrical uncertainties that jeopardizes the safe operation of synchronous systems. During test time designers need to identify critical paths that violate setup time restrictions for flip-flops and, especially, hold time restrictions that cannot be corrected by adjusting the working frequency. Furthermore, the need to find the minimum allowable energy consumption for each computational load make DVFS control systems change the clocking frequency constantly. In this context a CPM provides with feedback information about the combinational delay of a CP that can used both a test and run-time operations.

This work has introduced a new CPM architecture based on a time-to-digital converter working with both the system clock and a delayed version of it. The structure provides a means to obtain a quantized measure of the last transition produced in the CP centered around the clock edge. The monitor provides information for both setup time and hold time violations. The monitor can be set to different configurations depending on the portion of clock period that needs to be sensed, also, the proposed CPM is able to work with any logic family.

The monitor has been validated in a 40nm 1.1V commercial technology realizing post-layout parasitic-extracted simulations, it takes an area of 4028 μm^2. The monitor provides two configurable digital outputs, one for setup time violations and another for hold time violations. The outputs are quantized by the delay of a buffer (170 ps in the nominal case), independent of the operating conditions. Compared to previous works in the literature, the CPM occupies little area, as there is no need to replicate the CP, and the impact in the CP consists just on the addition of the input capacitance of an inverter at the end of the CP. Furthermore, this monitor provides information on hold time violations, which can be used at test time.

As future work, we are working on reducing the quan-

TABLE I

COMPARISON TABLE OF CPM AS DESCRIBED IN LITERATURE.

	This Work	[1]	[2]	[3]	[4]	[5]
Architecture	Not applicable	16x16 multiply	IBM Power6	32-bit MIPS	Not applicable	64-bit Alpha processor
Signal generation	Not applicable	Flip-flop	Flip-flop	Toggle Flip-Flop	VCO and CLK	Not applicable
Critical path synthesis	Embedded in critical path	1 serial	5 parallel	Critical Path Replicas	Critical Path Replicas	Embedded in critical path
Data out	5-20 bits	1 bit	12 bits	3 bits	PWM signal	1 bit
Time-to-Digital Conversion	Thermometer code	Flip-flop	Thermometer code	Thermometer code	Not applicable	Flip-flop
Technology	40 nm	180 nm	65 nm	45 nm	14nm	180 nm
Frequency	1 GHz	90 MHz	4-5 GHz	1.5 GHz	—	200 MHz
Area	53x76 μm^2	350x450 μm^2	90x38 μm^2	—	—	—

tization step down to inverter delays or even fractions of inverter delays, this will certainly increase the complexity of the system, especially the tuning throughout the temperature and process spread, however it will provide the monitor with enhanced characteristics, improving its range of operation.

REFERENCES

[1] Mohamed Elgebaly and Manoj Sachdev. Variation-aware adaptive voltage scaling system. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 15(5):560–571, 2007.

[2] Alan Drake, Robert Senger, Harmander Deogun, Gary Carpenter, Soraya Ghiasi, Tuyet Nguyen, Norman James, Michael Floyd, and Vikas Pokala. A distributed critical-path timing monitor for a 65nm high-performance microprocessor. In *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pages 398–399. IEEE, 2007.

[3] Junyoung Park and Jacob A Abraham. A fast, accurate and simple critical path monitor for improving energy-delay product in dvs systems. In *Proceedings of the 17th IEEE/ACM international symposium on Low-power electronics and design*, pages 391–396. IEEE Press, 2011.

[4] Bruce Fleischer, Christos Vezyrtzis, Karthik Balakrishnan, and Keith A Jenkins. A statistical critical path monitor in 14nm cmos. In *Computer Design (ICCD), 2016 IEEE 34th International Conference on*, pages 507–511. IEEE, 2016.

[5] Dan Ernst, Nam Sung Kim, Shidhartha Das, Sanjay Pant, Rajeev Rao, Toan Pham, Conrad Ziesler, David Blaauw, Todd Austin, Krisztian Flautner, et al. Razor: A low-power pipeline based on circuit-level timing speculation. In *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, pages 7–18. IEEE, 2003.

[6] A. Drake. *Adaptive Techniques for Dynamic Processor Optimization*, chapter 7: Sensors for Critical Path Monitoring, pages 145–174. Springer, 2008.

[7] Alan J Drake, Robert M Senger, Harmander Singh, Gary D Carpenter, and Norman K James. Dynamic measurement of critical-path timing. In *Integrated Circuit Design and Technology and Tutorial, 2008. ICICDT 2008. IEEE International Conference on*, pages 249–252. IEEE, 2008.

[8] Masakatsu Nakai, Satoshi Akui, Katsunori Seno, Tetsumasa Meguro, Takahiro Seki, Tetsuo Kondo, Akihiko Hashiguchi, Hirokazu Kawahara, Kazuo Kumano, and Masayuki Shimura. Dynamic voltage and frequency management for a low-power embedded microprocessor. *IEEE journal of solid-state Circuits*, 40(1):28–35, 2005.

[9] Todd Austin, David Blaauw, Trevor Mudge, and Krisztián Flautner. Making typical silicon matter with razor. *Computer*, 37(3):57–65, 2004.

[10] Tim Fischer, Jayen Desai, Bruce Doyle, Samuel Naffziger, and Ben Patella. A 90-nm variable frequency clock system for a power-managed itanium architecture processor. *IEEE Journal of Solid-State Circuits*, 41(1):218–228, 2006.

[11] Hendrik F Hamann, Alan Weger, James A Lacey, Zhigang Hu, Pradip Bose, Erwin Cohen, and Jamil Wakil. Hotspot-limited microprocessors: Direct temperature and power distribution measurements. *IEEE Journal of Solid-State Circuits*, 42(1):56–65, 2007.

[12] Victor Avendano, Victor Champac, and Joan Figueras. Signal integrity verification using high speed monitors. In *Proceedings of the European Test Symposium, Ninth IEEE*, pages 114–119. IEEE Computer Society, 2004.

[13] Haihua Su, Frank Liu, Anirudh Devgan, Emrah Acar, and Sani Nassif. Full chip leakage estimation considering power supply and temperature variations. In *Proceedings of the 2003 international symposium on Low power electronics and design*, pages 78–83. ACM, 2003.

[14] Jason Howard, Saurabh Dighe, Yatin Hoskote, Sriram Vangal, David Finan, Gregory Ruhl, David Jenkins, Howard Wilson, Nitin Borkar, Gerhard Schrom, et al. A 48-core ia-32 message-passing processor with dvfs in 45nm cmos. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pages 108–109. IEEE, 2010.

[15] Wonyoung Kim, Meeta Sharma Gupta, Gu-Yeon Wei, and David Brooks. System level analysis of fast, per-core dvfs using on-chip switching regulators. In *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on*, pages 123–134. IEEE, 2008.