Enhancing Power, Performance, and Energy Efficiency in Chip Multiprocessors Exploiting Inverse Thermal Dependence

Katayoun Neshatpour, Wayne Burleson, *Fellow, IEEE*, Amin Khajeh, *Senior Member, IEEE*, and Houman Homayoun

***Abstract*— Technology scaling of complementary metal–oxide–semiconductor has resulted in new thermal behavior where increase in operating temperature results in reduced circuit propagation delay. This paper exploits this inverse thermal dependence (ITD) for power, performance, and temperature optimization in single-core and multicore processor architectures for various thermally hot and cold applications. Since ITD increases the maximum achievable operating frequency of a processor at high temperatures, it is used to reduce the execution time of applications. Dynamic thermal management (DTM) techniques, such as activity migration (AM), dynamic voltage frequency scaling (DVFS), and throttling, are modified to leverage ITD to either enhance the performance or the energy efficiency. While recent work observed the ITD effect for 45- and 32-nm technologies, in this paper, we explore future technologies through predictive SPICE models for 20-, 14-, 10-, and 7-nm technologies. The results show that the ITD-aware techniques reduce the execution time, energy-delay product (EDP) and energy-delay-square product by up to 28%, 33%, and 48%, respectively. Moreover, the ITD-aware DVFS yields the lowest execution time while resulting in the most uniform thermal profile and thus enhanced reliability. Overall, the ITD-aware techniques reduce the execution time and EDP, both in combination with DTM techniques or stand-alone, especially at lower than nominal operating voltages.**

***Index Terms*— Dynamic thermal management (DTM), dynamic voltage frequency scaling (DVFS), inverse thermal dependence (ITD), submicrometer complementary metal–oxide–semiconductor (CMOS) technologies.**

I. Introduction

REDUCING the feature sizes has been an ongoing trend in the processor design technology. Supply and threshold voltage scaling is done to control the increase in the power density. It should be noted that various design concerns do not allow the supply and threshold voltages to be scaled at the same rate, as the processor dimensions are scaling. Thus, with technology scaling, the power density is still increased, which results in an increase in the temperature. Such an increase in the power density is more significant for smaller feature sizes as shown in [1]. Huang *et al.* [2] show that with a scaling feature size of 0.755×, the power density is increased by 1.096× at a 45-nm technology node, while the same feature scaling of 0.755× at a 14-nm technology node increases the power density by 1.175×.

Thus, while the dimensions are scaling along with the supply and the threshold voltages, the power density of the processor still increases dramatically, which causes overheating and reliability issues, calling for new solutions to be developed for a low-power and low-temperature design.

As the integrated circuits enter into the deep-submicrometer complementary metal–oxide–semiconductor (CMOS) technologies, their thermal behavior changes significantly. Traditionally, as CMOS temperature increases, a propagation delay of the circuit also increases due to the lower mobility, which slows down the circuit. In the submicrometer technologies however, at low-operating voltage, an opposite behavior is observed—the circuit propagation delay reduces. This phenomenon is called the inverse temperature dependence (ITD) [3].

ITD introduces many new challenges as well as opportunities in the design and optimization of the circuit. Design corners will have to be reinvestigated, and power reduction and thermal management techniques will have to be adapted accordingly to achieve the lowest execution time without introducing any failures. With ITD, at high temperatures, the circuit will be able to benefit from the frequency headroom offered due to the reduction in the propagation delay. For applications in which the performance is heavily impacted by the operating frequency (frequency-sensitive applications), this headroom can be exploited to enhance the performance.

However, in order to observe the ITD and its benefits on frequency scaling, the circuit will have to work at low voltages (typically lower than the nominal), which will in turn decrease the operating frequency of the circuit. On the other hand, operating the circuit at the nominal voltage where no ITD is observed results in higher energy consumption and increased temperature. Therefore, the ITD phenomenon offers several new power and performance tradeoffs at various design corners.

This paper investigates ITD in the future CMOS technologies to understand whether further scaling will increase the significance of the ITD. In this paper, several design optimal points are explored for the 7-, 10-, 14-, and 20-nm

Manuscript received June 5, 2017; revised September 27, 2017; accepted November 12, 2017. Date of publication January 17, 2018; date of current version March 20, 2018. This paper was presented in part at GLSVLSI 2015. *(Corresponding author: Katayoun Neshatpour.)*

K. Neshatpour and H. Homayoun are with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030 USA (e-mail: kneshatp@gmu.edu).

W. Burleson is with the Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA 01003 USA.

A. Khajeh is with Broadcom Limited, San Jose, CA 95131 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2017.2780800

predictive technology models (PTMs) [4]. At the circuit level, we use fully synthesizable LEON3 core to understand the impact of ITD on processor maximum achievable operating frequency.

For circuit simulations, the complete critical paths were explored rather than just a few basic cells (NAND, XOR, and so on) or an inverter-based ring oscillator. The results of circuit characterizations are used at the architectural level to simulate the impact of ITD on dynamic thermal management (DTM) techniques to explore opportunities for improvements.

For architecture simulation, both single-core and multicore architectures are studied. The simulation results of 12 different applications from Media and Spec benchmarks are studied. The studied applications are divided into two different groups, namely, thermally hot and cold applications. Various techniques for thermal management, performance improvement, and energy-efficiency enhancement are proposed for each group of applications accordingly.

The contributions made by this paper are summarized as follows:

1) characterizing propagation delay by performing accurate SPICE-level analysis on LEON3 processor at six voltage levels and various temperature ranges for four PTM technologies to understand ITD trends with technology scaling;
2) carrying out accurate simulations to suggest architecture-level solutions to leverage the ITD effect to enhance the performance and energy efficiency of various applications;
3) proposing stand-alone ITD-aware dynamic frequency scaling (DFS) to enhance the execution time of applications;
4) studying how the DTM techniques change in presence of ITD for the future submicrometer technologies;
5) combining ITD-aware techniques with various DTM techniques to not only enhance performance but also enhance the thermal profile and energy efficiency;
6) analyzing how various operating conditions, including the voltage, frequency, and technology node, affect the ITD and how the DTM techniques have to be modified to leverage the ITD at various operating points.

Based on the simulation results, we make the following observations.

1) The ITD-aware techniques enhance performance in terms of instruction per second (IPS) by up to 28%.
2) The ITD-aware techniques reduce the energy-delay product (EDP) and energy-delay-square product (ED2P) by up to 33% and 48%, respectively.
3) The ITD-aware dynamic voltage and frequency scaling (DVFS) yields the lowest execution time while producing the most uniform thermal profile resulting in enhanced reliability. We show that overall in submicrometer technologies, leveraging ITD speeds up the processor.
4) ITD is more significant at lower operating voltages, and thus, operating the core at lower than nominal voltages and utilizing an ITD-aware DTM technique yields the lowest EDP results.

This paper is organized as follows. Section II introduces the related work. Section III provides the background on ITD. Section IV shows how the propagation delay changes with respect to temperature and operating voltage for various technology nodes. In Section VI, a number of SPEC and media applications are characterized and categorized into cold and hot applications. In Section VII, we introduce ITD-aware thermal management techniques. Section VIII introduces the simulation framework. In Section V, we introduce our ITD-aware algorithm for single-core and multicore applications. Section IX presents the results and evaluates them. Section XI presents our conclusion remarks.

II. Related Work

Prior works have focused on the implication of the ITD on circuit design and how ITD changes the device behavior and design corners. Reference [5] proposes a test structure with a built-in polyresistor-based heater to characterize ITD in digital circuits. Reference [6] carries out a study of the ITD phenomenon at low-voltage supplies. They discuss ITD's consequences for the static timing analysis and propose a proper handling of ITD in an industrial sign-off tool. In [7], possible temperature instabilities in the low-voltage regime are described by using circuit simulation environments incorporating temperature change in time. The experiments are carried out using metal–oxide–semiconductor field-effect transistor and 32-bit adder circuit in a quarter micrometer CMOS technology with a low threshold voltage of 0.25 V. Verma and Mishra [8] show that with increased temperature, propagation delay of LECTOR-based CMOS NAND gate and conventional NAND gate decreases approximately linearly.

Besides studies on the ITD phenomenon, various works have proposed various approaches to leverage the ITD effect at high temperatures to improve energy and delay of the circuits. In [3], the effect of ITD is investigated for the behavior of the clock tree mapped on an industrial 65-nm CMOS technology, and ITD is shown to occur at low-operating voltages. In [9], a dual-v_t circuit is proposed for a commercial low-power 65-nm CMOS technology under ITD to ensure the temperature-insensitive behavior of the circuit. Latif *et al.* [10] use the available power headroom at low temperatures to increase the voltage when the temperature is below a certain level to enhance the performance. While the focus of the latest work is on the effect of ITD in current technology nodes (see [3], [9], [10]), it is important to explore ITD impact on the future nodes as well. This paper investigates ITD in the future CMOS technologies to understand whether further scaling will increase the significance of ITD.

On DTM, processor thermal characteristics at the architectural level have been studied extensively in recent years [11]. Several techniques have been proposed to reduce chip temperature in single-core [11]–[13] and multicore architectures [14]–[20]. These techniques either migrate the processor activity [21] or adapt processor resources to reduce temperature [11]. In the latter, the utilization across processor units is balanced to control the power density. DVFS has also shown to be effective in balancing the temperature of processor [14], [22].

These techniques have been widely studied at the architectural and operating system levels without considering the ITD phenomenon. This paper shows how ITD can be exploited to enhance the benefits of DTM techniques for power, performance, and energy-efficiency improvement.

III. Inverse Thermal Dependence

In order to understand the effect of temperature on the behavior of CMOS transistors, a proper model should be utilized. Numerous models have been used to present the behavior of the CMOS transistors. The square law model in (1) estimates the drain current of CMOS transistor operating in the saturation mode [23]

$$I_D = \mu \times C_{\text{ox}} \frac{W}{L} (V_{\text{gs}} - V_t)^2 \tag{1}$$

where I_D is the drain current, μ is the mobility, C_{ox} is the oxide capacitance, W and L are the channel width and length, V_{gs} is the gate–source voltage, and V_t is the threshold voltage. Mobility and threshold voltage are two parameters that are dependent on the temperature; however, they work on opposite directions. While an increase in the temperature reduces the mobility, and thus slows down the device, it reduces the threshold voltage, which enhances the device speed according to (1).

At high-operating voltages, the threshold voltage change has little effect on $(V_{\text{gs}} - V_t)$, and thus, the mobility determines the effect of the temperature on the device speed. However, at low-operating voltages close to the threshold voltage, the difference between $(V_{\text{gs}}$ and $V_t)$ is small. Thus, small variations in V_t result in a significant change in $(V_{\text{gs}} - V_t)$, i.e., $(\delta(V_{\text{gs}} - V_t)/(V_{\text{gs}} - V_t))$ is high. The variations in $(V_{\text{gs}} - V_t)$, in turn, translates to variations in the current and the propagation delay. Thus, at low-operating voltages, the influence of variations in V_t on the propagation delay is more significant than the variations in the mobility. In other words, $(\delta I_D/\delta V_t)$ is higher than $(\delta I_D/\delta \mu)$. As a result, the thermal behavior of the transistor is mostly determined by variations in V_t [24], [25].

As the technology scales down, the operating voltage approaches the threshold voltage. Thus, when working at slightly lower than the nominal voltage, increasing the temperature reduces V_t and allows the device to switch faster at higher temperature. This phenomenon is called the ITD.

The ITD effect has been observed for the existing technology nodes for specific cells in [3], [5], and [6]. For instance, [5] shows that for the 130-nm technology, the maximum frequency of an inverter chain oscillator increases at elevated temperatures for lower operating voltages (i.e., 0.55 V). In [6], the delay of a NAND cell at a 90-nm technology is shown to increase at elevated temperatures for the operating voltages of lower than 1.12 V.

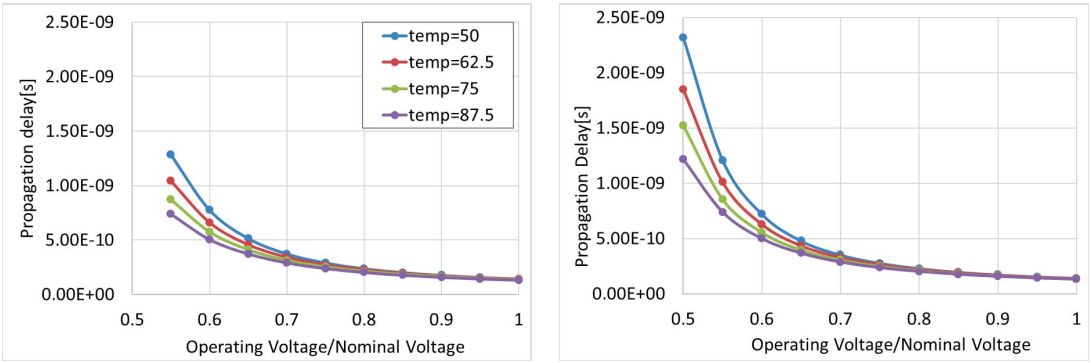
In Section IV, we show that, the propagation delay of the critical paths in a LEON3 processor is actually reduced at high temperature using accurate SPICE models. Moreover, the decrease in the propagation delay is higher at low-operating voltages and smaller technologies. Several works have focused on the implication of the ITD on circuit design and how ITD affects the design corners. Traditionally, in terms of voltage and temperature, the slow corner of a design is verified at low voltage and high temperature, and the fast corner is verified at high voltage and high temperature [6]. The inverse thermal behavior of submicrometer CMOS technologies changes the design corners. In all of these studies, various circuit-level techniques have been investigated for the existing technologies (e.g., 65 nm) in [3] and [9] to prove the existence of ITD or leverage the frequency headroom to enhance the performance [9], [10]. In [1], in this paper, ITD was explored for PTM technologies to not only predict the ITD behavior for emerging technologies but also to explore system-level techniques to dynamically change the frequency of the applications leveraging ITD. In this paper, ITD is explored for various emerging technologies at deep submicrometer and at different temperature levels. Such an analysis allows a high-level exploitation of ITD. An analysis of various applications with a different thermal behavior allows a more profound understanding of the efficiency of ITD-aware solutions for performance, power, and energy-efficiency improvements.

While we discuss and study the ITD phenomenon for CMOS technologies, the same trend has been observed for fin field-effect transistors (FinFETs). According to [26] and [27], fabricated FinFETs operating at subthreshold voltages reportedly have increased current as die temperature rises. Huang *et al.* [28] show that this effect is due the narrowing of the bandgap and changes in carrier mobility, induced by tensile stress effect of the insulator in the FinFET structure (the tensile stress from the insulator layer to the fin body affects the device characteristics more significantly as technology scales down). Since the methods suggested in this paper benefit from increased speed of the device at elevated temperatures, the suggested ITD-aware techniques are applicable to devices operating with the FinFET technology.

IV. Analysis of ITD at the Circuit Level for CMOS Technologies

The latest work has explored the effect of ITD in current technology nodes [3], [9], [10]; however, it is also important to explore ITD impact in the future CMOS technology nodes. This paper first investigates ITD in the future CMOS technologies to understand whether further scaling increases the significance of ITD. We use PTM [4], which is an evolution of Berkeley PTM [29] for our simulations. PTM provides accurate, customizable, and predictive models for future transistor and interconnect technologies. Reference [30] evaluates the PTM models for various studies. For instance, in [31]–[34], the error in prediction of I_D when the transistor is ON is 1%, 1%, 3%, and 4%, respectively.

We synthesize LEON3, a fully synthesizable VHDL model for open SPARC V8 32-bit processor architecture [35] using Synopsys Design Compiler and extract multiple critical paths for further analysis. Synopsis design compiler provides a detailed gate-by-gate report for entire paths in the design. The paths have various delays. We select ten paths with maximum delays and use the PTM transistor models to reconstruct the SPICE models for each gate in the paths. Subsequently, we observe how various temperatures affect the path delay for



(a) (b)

Fig. 1. Propagation delay of a critical path in LEON3 for (a) 7 and (b) 14 nm.

each cell using SPICE simulations. We calculate the modifications to the path delays by combining all the gate delays in the path. To provide accurate analysis of ITD characteristics in a processor design, in these simulations, the complete critical paths were explored rather than just a few basic standard cells, such as NAND and XOR, or an inverter-based ring oscillator.

Additionally, since gate-level netlist from Design Compiler does not include *RC* delay, we simulated dominated paths using π-Models for Metal4–Metal7 layers (that are usually used for signal routing) for 20 nm. For our simulations, we assumed different wire lengths (1, 50, and 100 μm). Moreover, we selected appropriate drivers for each case that resulted in maximum acceptable slew rate as physical design tools do. Our simulation results show that the worst case delay degradation occurs for the lower metal layer (Metal4) and longer wire length. Assuming that 20% of our postlayout top critical path will be *RC* delay and 80% cell delay, the difference between prelayout and postlayout speedup for 25 °C–125 °C will be 7% lower in 20 nm at nominal voltage. In other words, postlayout speedup as a result of increase in temperature is lower than that of prelayout. This is mostly due to the fact that resistance (R) increases with temperature, which in turn increases the *RC* delay at higher temperatures. Since we do not have access to a full library of cells at 14, 10, and 7 nm, for the remainder of this paper, we use the prelayout results.

The focus of this paper is to show the trend of temperature effect as we scale down the technology to build a model to use at the system and the architecture levels. Our proposed methodology is not limited to the particular studied LEON3 design, and is applicable to arbitrary designs, since we carry out an accurate measurement of temperature delay at various process–voltage–temperature. More specifically, we run HSPICE simulation on a set of critical paths extracted from various operating points as an indication of the longest path delays for the LEON3 processor at all operating conditions. Since various factors, including the process variation, could change the critical paths at different design corners, by studying not only a single, but many critical paths in the design, we take the effect of process variation into account. It should be noted that, several factors, such as device aging, wear-out, and process with variation and thermal imbalance, change the critical paths of a design. Therefore, it is important to first identify the impact of these factors on the critical path delays and then analyze the influence of ITD.

We studied 7-, 10-, 14-, and 20-nm technologies, for which the nominal voltages are 0.7, 0.75, 0.8, and 0.9, respectively.

Fig. 1 shows the average path delay for various critical paths in LEON3, for the 7- and 14-nm PTM technologies. Fig. 1(b) shows that for the 14-nm technology, with an operating voltage that is 60% of nominal voltage, the critical path delay at 87.5 °C is 33% lower than the critical path delay at 50 °C. For the 7-nm technology, with the same operating voltage of 60% of the nominal voltage, the critical path delay at 87.5 °C is 38% lower than the critical path delay at 50 °C. This indicates that the impact of ITD on critical path delay increases further as the technology scales down. Thus, for these technologies, there will be a headroom at high temperatures for increasing the frequency as long as overheating of the device does not introduce reliability issues. Moreover, for all the studied technologies, the reduction in the critical path due to the elevated temperature is more significant at low-operating voltages.

The SPICE simulation was carried out for multiple critical paths at various design points at a fine-grained level. Fig. 2 shows the average critical path delays of LEON3 at 7- and 20-nm technologies as a function of the temperature and operating voltage. Fig. 2 shows that the 7-nm technology is faster than the 20-nm technology, which is to be expected. Fig. 2 provides a visualization of the ITD phenomenon. Fig. 2 shows that the variations of the critical path delay with respect to the temperature are insignificant at voltages close to nominal voltages, and are positively correlated with the temperature. However, at low-operating voltages close to the threshold voltage, the correlation gradually changes to negative (i.e., ITD phenomenon), and the changes in the critical path delay are more significant at low voltages. Moreover, Fig. 2 also shows that the sensitivity of the critical path delay to the temperature is more significant for 7 nm in comparison with the 20-nm technology.

V. Leveraging ITD to Enhance Performance

Based on the simulation results, the frequency headroom available at elevated temperatures due to the ITD effect can

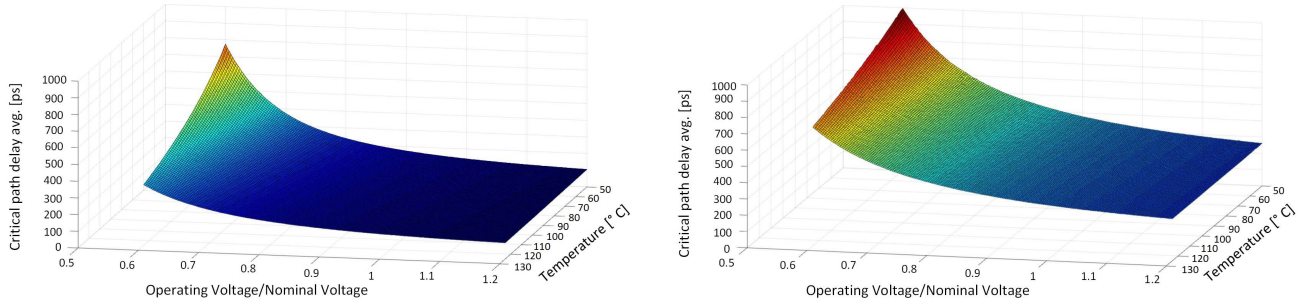


Fig. 2. Average of critical path delay of LEON3 as a function of the temperature and operating voltage for (a) 7 and (b) 20 nm.

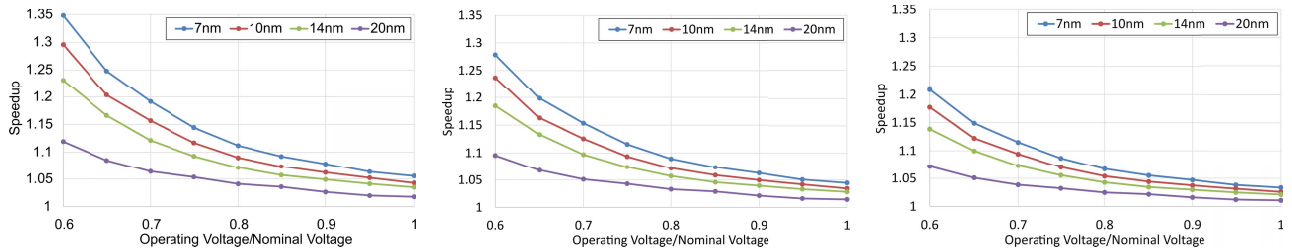


Fig. 3. Reduction in propagation delay (speedup) when the temperature rises from 50 °C to 75 °C. (a) Prelayout. (b) Postlayout assuming 80%–20% cell-RC delay. (c) Postlayout assuming 50%–50% cell-RC delay.

TABLE I
SPEEDUP OF THE CIRCUIT AT ELEVATED TEMPERATURES

Tech.	$\frac{Op.\ Voltage}{Nom.\ Voltage}$	Speedup $50°C \rightarrow 62.5°C$	Speedup $50°C \rightarrow 75°C$	Speedup $50°C \rightarrow 100°C$	Speedup Nom. Voltage
7nm	0.65	1.117	1.261	1.542	3.680
	0.75	1.058	1.15	1.181	2.101
	0.85	1.037	1.089	1.297	1.426
10nm	0.65	1.086	1.186	1.377	3.636
	0.75	1.054	1.118	1.236	1.990
	0.85	1.033	1.070	1.137	1.405
14nm	0.65	1.079	1.166	1.334	3.263
	0.75	1.044	1.091	1.19	1.903
	0.86	1.025	1.052	1.121	1.384
20nm	0.65	1.043	1.089	1.178	2.406
	0.75	1.020	1.043	1.07	1.653
	0.85	1.015	1.035	1.069	1.292

be used to speed up the processor. However, it should be noted that the circuit needs to operate at low voltages (lower than the nominal) in order to observe the ITD effect. Table I shows the range of speedup at elevated temperatures for various operating points for 7-, 10-, 14-, and 20-nm technologies. Arguably, when the circuit is operating at voltages lower than the nominal voltage, the maximum frequency is typically lower. Thus, in order to make a better comparison with the speed of the circuit working at the nominal voltage, Table I also shows the speedup of the circuit, if we run it at the nominal voltage. On the other hand, running the circuit at the nominal voltage results in higher power dissipation and increased temperature. Thus, thermal management techniques have to be used, which in turn slows down the circuit. An overview of Table I shows that, the operating condition that yields the lowest execution time cannot be easily determined, as it heavily depends on power and thermal profile of the processor, which is decided at run time, and is application dependent and requires architecture-level control.

Fig. 3 shows the speedup of the critical path when the temperature rises from 50 °C to 75 °C, assuming prelayout simulation results in Fig. 3(a), as well as postlayout simulation with two assumptions for wire delay [80%–20% cell-RC delay in Fig. 3(b) and 50%–50% cell-RC delay in Fig. 3(c)] in the critical path. The results show improved critical path delays, even with pessimistic wire-delay assumptions. Regardless, the methodology proposed for the analysis presented in this paper remains the same, and as long as increasing the temperature reduces the critical path delay, the proposed ITD-aware techniques enhance the performance. Moreover, Fig. 3 shows that as we move away from the nominal voltage toward threshold voltage, the effect of temperature on the critical path delay increases.

VI. APPLICATION CHARACTERIZATION UNDER ITD

A. Performance Analysis Under ITD

Traditionally, with large feature sizes, the increase in the temperature results in an increase in the propagation delay in all high-temperature paths in the processor. Thus, the operating frequency needs to be reduced to ensure that there is no failing path in the processor. Therefore, processor throughput is affected by its operating temperature. In order to take frequency into consideration for performance estimation, instructions committed per second (IPS) is calculated as well as instructions per cycle (IPCs) (IPS is the frequency × IPC).

With ITD, the increase in the temperature reduces the delay and allows a headroom for increasing the frequency. As long as the critical paths, which generally conclude the operating frequency, lie in the core, an increase in the temperature allows running the processor at higher frequencies. However, it should

TABLE II
ARCHITECTURAL SPECIFICATION

	Specification
Core	8
Issue, Commit width	4
INT instruction queue	24 entries
FP instruction queue	24 entries
Reorder Buffer entries	48 entries
INT registers	128
FP registers	128
L1 cache	64KB, 8-way, 1ns
L2 cache	512KB, 8-way, 7.5ns

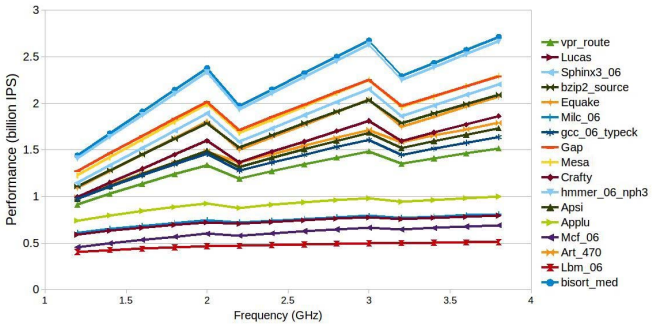


Fig. 4. Performance of Spec benchmarks as a function of frequency.

be noted that colder regions, including the L2 and last level cache, which are normally operating at low temperature [11], still maintain their path delays. For example, for a cache delay of 100 ns, increasing the frequency from 1 to 2 GHz increases the access time from 100 cycles to 200 cycles, which negatively affects IPC. Thus, the IPS does not increase linearly with the frequency, since the IPC might be dropping at higher frequencies.

In order to evaluate the effect of IPC drop on the performance for various applications, we use the SMTSIM simulator [37] with a subset of Spec2000, Spec2006 [38], and Media benchmarks. The applications are carefully selected to represent a wide range of various thermal, performance, and frequency sensitivity behavior. Table II shows the microarchitecture parameters of our studied multicore architecture. The access time for individual SRAM units was captured using McPAT [39].

Based on the operating frequency, the number of cycles to access the memory and cache subsystem is varied, and the new IPC at each frequency level is calculated accordingly. The performance results (i.e., IPS) for a selected group of Spec2000, Spec2006, and Media benchmarks are shown in Fig. 4. The results show that for memory-intensive applications, e.g., lbm_06, the high number of accesses to the cache subsystem reduces the IPC at high frequencies (achieved at high core temperature) and, therefore, limits the potential performance gain of elevated frequency. Thus, for these applications, the increased frequency at elevated core temperatures is not useful, as it only increases the power and temperature, and negatively affects the reliability without enhancing the performance. For compute-intensive applications however, due to the lower number of accesses to the memory, the higher cache access cycles at higher frequencies can be tolerated as it does not affect the IPC significantly. Therefore, for this group

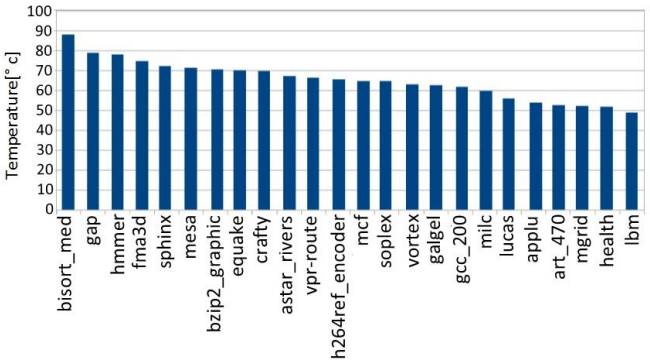


Fig. 5. Steady-state temperature of Spec2000, Spec2006, and Media benchmarks at 2.5-GHz core frequency for the 20-nm CMOS technology with a nominal operating voltage (0.9 V).

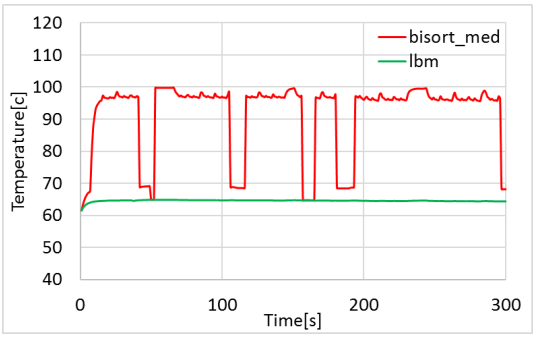


Fig. 6. Transient temperature at 2.5-GHz core frequency for 20-nm CMOS technology with a nominal operating voltage (0.9 V).

of applications, the performance increases with the increased frequencies.

In Fig. 4, small IPS drops are observed for a number of applications at specific frequency points. For the frequency-sensitive applications, a small increase in the frequency may cause the cache access latency to increase by one cycle, which reduces the IPC consequently. This reduction in the IPC can nullify the effect of frequency increase on the IPS and even can result in a slight reduction in the IPS.

B. Thermal Behavior Analysis

For thermal characterization of studied applications, we used HotSpot [11]. Fig. 5 shows the steady-state temperature of the hottest unit in processor derived from Hotspot for selected applications. As shown, the steady-state temperature of hottest unit (register file in most applications) varies from 48 °C in lbm to 88 °C for bisort_medical in the studied applications, which shows a 30 °C variation.

By comparing the performance and temperature results in Fig. 5, we observe that the applications with higher IPC, which are mostly the compute-sensitive applications, are generally the ones that have higher temperature.

It is also important to gather information about the transient temperatures, as some applications reach critical temperatures during their execution time. Fig. 6 shows the temperature values for the lbm_06 and the bisort_med application, which are the examples of cold and hot applications, respectively. Fig. 6 also shows that there is significant temperature variation during the execution of the hot benchmark.

If the compute-intensive application (hot application) is running on a single core without any feedback from the transient temperature, the core reaches critical temperatures, which damages the device and causes reliability issues. Thus, for such hot applications, a DTM technique needs to be deployed.

Figs. 5 and 6 show that while the steady-state temperature for bisort_medical never reaches 90 °C, the transient temperature exceeds 90 °C frequently. In addition, for smaller CMOS technologies, the power density is higher resulting in higher temperatures. In fact, HotSpot simulations show that the temperature reaches elevated temperatures even when the operating voltage is decreased to 75% of the nominal voltage for smaller technology nodes.

Based on the results in Figs. 4 and 6, the applications are divided into two groups: the memory-intensive applications and the memory-nonintensive applications. For the memory-nonintensive applications, which could also be referred to as compute-intensive applications, using higher frequencies is more desirable, as the increased frequency contributes significantly to the performance. Thus, these applications are categorized as frequency sensitive. For the memory-intensive application, higher frequencies are not desirable as they increase the power consumption with little contribution to the performance.

VII. Dynamic Thermal Management

DVFS [40], AM [21], and throttling [11], [14] are among the most well-studied techniques to manage temperature as well as performance and power in multicore architectures. These techniques have been developed with the assumption that the device is slower at elevated temperatures. In this section, we study how these techniques are effective in managing power, temperature, and performance in the presence of ITD where the behavior of the circuit is changing at various thermal points.

For this study, single-core and multicore architectures are studied. On the single-core architecture, the DTM techniques are carried out on the studied applications. On the multicore architecture, based on the nature of applications, frequency-sensitive applications and frequency-insensitive applications are selected to run simultaneously on a multicore systems. The different thermal behavior of these types of applications is then explored to enhance the thermal, power, and performance of the collective. The integrated platform of the SMTsim and McPAT was used to generate power results for various frequencies for each application. The power results were then fed to the Hotspot at different time intervals to explore various thermal management techniques.

Our simulations show that the ITD phenomenon is observed even at temperatures as low as 62.5 °C (see Table I). Thus, the extent to which ITD is leveraged to enhance the performance depends on two factors: 1) the amount of time the temperature is higher than the threshold values during the execution of the application; and 2) the rate at which ITD allows us to increase the core frequency. Thus, while operating the circuit at lower temperatures results in lower temperatures, and thus reduction in factor 1, for lower voltages as shown in Table I, the speedup rate of the delay paths and hence the increase in the frequency is higher. Thus, selecting the optimal voltage to be able to benefit from the ITD is an important challenge, necessitating the experiments to simulate various operating voltages.

A. Throttling

The simplest way to prevent the processor from reaching critical temperatures is to apply pipeline throttling. With throttling, when an application reaches a threshold temperature, the pipeline is halted until the core cools down [11]. During the throttling time, the core remains idle. This results in an underutilization and, therefore, loss of performance.

B. Activity Migration

AM is another effective technique to reduce temperature. In AM, the application that reaches critical temperature faster is migrated to the colder core in order to avoid high temperatures.

C. Dynamic Voltage and Frequency Scaling

Another effective technique for managing temperature is DVFS, in which various voltage and frequency pairs are utilized in the processor. The processor supply voltage and frequency are adapted dynamically to respond to thermal emergencies [40]. The number of voltage and frequency pairs in DFVS varies from two in Intel SpeedStep technology to 40 or even more in Intel XScale [43].

D. ITD-Aware Throttling, AM, and DVFS

After running a frequency-sensitive application on a single core, the temperature increases. With ITD, this elevated temperature increases the maximum allowable operating frequency on the core. Thus, DFS allows the application to run at higher frequencies when it reaches higher temperature.

Before reaching the critical temperature, the core reaches midpoint thresholds. In the ITD-aware techniques, when the temperature of the core is higher than these midpoint thresholds, we increase the frequency of the core to benefit from the frequency headroom offered by the ITD. Afterward, we keep monitoring the temperature until it reaches the critical temperature at which point, throttling, AM, or DVFS, is done to reduce the temperature. If the temperature drops below the midpoint temperatures at any time, the operating frequency is set back to its lower value. This technique allows us to run the core at higher frequency for the time interval when the temperature is between the midpoint thresholds and the critical value to increase the overall IPS.

Fig. 7(a) and (b) shows the ITD-aware DVFS and throttling algorithms for an n-core system with m levels of temperature-frequency pairs. Array T contains core temperatures, and t_h is a hash table with temperature as its key and task numbers as its value. FL and TL arrays contain frequency and temperature levels, with the first term being the starting/lowest frequency and temperature point. Among the studied DTM techniques, throttling and DVFS could be carried on both

```
Inputs: CoreTemp: T[1:n]              CoreFreq: F[1:n]
        FreqLevel: FL[1:m]   TempLevel: TL[1:m]
Core Hash: t_h[Temp]→{1, 2, …, n}

Begin
Every time interval:
    For i=1 to n/2:
        if T[i]>Critical_temp:
            DTM[t_h(T[i])]→Core[t_h(T[n-i])]  --<perform DVFS/throttle>
        Else:
            For j=1:m:
                If T[i]>TL[j]:
                    F[i]=FL[j]  --<ITD-aware frequency changing>
        Run(benchmark)
        T←new Temperature
end
```

(a)

```
Inputs: CoreTemp: T[1:n]    CoreFreq: F[1:n]
        FreqLevel: FL[1:m]   TempLevel: TL[1:m]
Core Hash: t_h[Temp]→{1, 2, …, n}

Begin
Every time interval:
    Sort_descending(T)
    For i=1 to n/2:
        if T[i]>Critical_temp:
            Task[t_h(T[i])]→Core[t_h(T[n-i])]  --<perform task migration>
        Else:
            For j=1:m:
                If T[i]>TL[j]:
                    F[i]=FL[j]          --<ITD-aware frequency changing>
        Run(benchmark)
        T←new Temperature
end
```

(b)

Fig. 7. ITD-aware algorithm for (a) throttle and DVFS and (b) AM.

single-core and multicore architectures, while AM is specific to multicore architectures. In the algorithm described in Fig. 7, we monitor the temperature of each core and increase its frequency at elevated temperatures to leverage the frequency headroom provided by the ITD at specific time intervals; however, if the core reaches critical temperatures, we activate a DTM techniques, i.e., AM, DVFS, or throttling.

Assuming 50 °C as the minimum and 100 °C as the maximum temperature based on our simulations for the studied operating conditions, the midpoint thresholds are selected based on the number of frequency levels explored for ITD-aware frequency scaling. For instance with a two-level frequency scheme, we use a midpoint temperature of 75 °C. Thus, the frequency of the chip is kept constant for a temperature range of 25 °C. Alternatively, we study four frequency levels, by setting the midpoint thresholds to 62.5 °C, 75 °C, and 87.5 °C, and changing the frequency in temperature steps of 12.5 °C. Increasing the number of midpoint thresholds allows higher exploitation of the ITD; however, since the frequency scaling introduces energy and performance penalties, the large number of midpoint thresholds increases the frequency-switching penalty and nullifies the performance gain of the ITD-aware scheme.

It should be noted that in our simulations, we explore multiple levels of temperature-frequency pairs for multiple midpoint temperatures, which increases the speedup of the ITD-aware DTM techniques. For the DVFS technique, we study various operating voltage–frequency–temperature tuples. However, for AM and throttling, the voltage is fixed and we only change the frequency. For optimization purposes, we study AM and throttling at various operating voltages to find the optimal operating voltage.

VIII. Simulation Platform

In order to study the thermal characteristic of various applications, we study the behavior of SPEC2000, SPEC2006, and Media applications. We use an integrated version of SMTSIM, McPAT [39], and HotSpot-5.02 [11] for performance, power, and temperature simulations. Each benchmark was simulated for 1 billion instructions after fast forwarding for 1 billion instructions.

The DTM techniques introduced in Section VII, including throttling, AM, and DVFS, are implemented in the integrated simulation framework. The ITD-aware modified versions of these techniques are studied to evaluate the performance enhancement given the frequency headroom offered at elevated temperatures in the presence of ITD.

TABLE III
Voltage Frequency Pairs

	Temperature[°C]			
	50	62.5	75	87.5
Voltage[$\frac{V_{op}}{V_{nom}}$]	Frequency[GHz]			
0.7	2150	2308	2462	2611
0.75	2764	2924	3075	3227
0.8	3385	3528	3684	3828
0.85	3978	4127	4265	4418

For this purpose, we study various processor operating voltage points for the 7-nm PTM technology. In order to make a fair comparison, the power and frequencies for different operating voltages are derived based on the values calculated by SPICE simulations from Table III. In Table III, the reported voltages are ($V_{\text{op}}/V_{\text{nom}}$), where V_{op} is the operating voltage and V_{nom} is the nominal voltage.

For temperature study, we simulate 400 s of transient thermal behavior. For every time interval (for this study, 1 and 0.1 ms), the transient temperature is checked, and if it reaches to 90 °C, a DTM mechanism is activated. This includes throttling, AM, or DVFS. To reduce the throttling or migration cost, we assume the switching process to happen at operating system time-slice intervals in power-gated cores as suggested in [44].

Thus, when a throttle is decided, a user state is saved to memory and a cache flush is triggered. The core is powered down. Thus, the core suffers no static leakage or dynamic switching power. However, after the throttling interval (during which power is set to zero and no instruction is executed), bringing the core back up introduces a latency, which includes power-up time, software overhead, and cache cold start effects. In [44], the performance penalty is set to 1k clock cycles. Other works (see [45], [46]) use various approaches to reduce the switching costs (1–3 μs and 32 cycles, respectively). However, in our simulations, we use the pessimistic assumption for performance penalty of each throttle (1k clock cycles per throttle).

In case of migration in the multicore architecture, the state of both cores is saved to the memory, and the user state saved by the cores is loaded from memory and switched for the two cores. Thus, the power traces used for the simulation of the rest of the applications are extracted from switched applications until the next migration. An overhead penalty of 1000 cycles is assumed as overhead for each migration, which again is a pessimistic assumption considering the software overhead and cache cold start effects.

For the DVFS, five levels of voltage-frequency pairs are used for each operating point (i.e., nominal—80%, 85%, 90%, and 95% of the nominal voltage.) We pessimistically assume 20-μs performance and 20-μJ energy penalty for a frequency and voltage transition based on [47]. In all the reported results, for performance, power, and energy, this overheard has been taken into account for the whole system.

It should be noted, that since we are experimenting with deep submicrometer technologies, the increase in power density increases the core temperature and necessitates the need for a DTM technique to control the temperature. Thus, the overhead penalties due to various DTM techniques are unavoidable. However, the overhead penalties increase in our proposed ITD-aware scheme due to the increase in the number of switchings. However, the results show that savings in performance, power, and energy efficiency are still considerable taking into account the increase in the overhead penalties.

The simulation results are illustrated for workloads combining hot and cold applications. It is important to note that pairing hot–hot applications and cold–cold applications yields no performance gain, as there is no noticeable thermal difference to exploit ITD. Therefore, in this paper, we only focus on the pairing of hot–cold applications, where there are opportunities to leverage ITD phenomena for performance and energy-efficiency improvements.

IX. Simulation Results

As mentioned earlier, the DTM techniques are mostly utilized when the applications reach a high critical temperature. However, not all applications always reach an extreme temperature. In general, the temperature of an application varies during its execution, allowing it to benefit from the frequency headroom allowed by the ITD by performing DFS. Thus, we divide the studied application into two groups, namely the cold and hot applications. The temperature of cold applications never reaches critical values, and thus, no DTM technique is performed on them. On the other hand, the temperature of the hot applications reaches critical values, necessitating a DTM technique to be applied.

For the studied applications, operating voltages as low as 80% of the nominal voltage yield the best results. However, in our experiments with lower voltages (i.e., the operating voltages of 75% and 70% of the nominal voltage), the temperature of most of the applications never reaches values higher that the selected thresholds and no improvement was observed.

A. Cold Applications

1) ITD-Aware Dynamic Frequency Scaling: In the DFS technique, we increase the frequency of the cold applications when the temperature rises above certain threshold values shown in Table III. In Fig. 8, the improvement in the performance of ITD-aware DFS over ITD-agnostic DFS scheme is depicted. The reported voltages are the ratio of the operating voltage to the nominal. The performances are compared based on the IPS values. As mentioned, we study various operating voltages. Fig. 8 shows that the performance enhancement gained through the ITD-aware DFS is improved more significantly at lower voltages, which was expected. As reported in Table III, the ITD is more significant at lower operating voltages close to the threshold voltage. At the nominal voltage, the gain of ITD-aware DFS is significantly lower. Overall, the performance of the studied applications has improved by 10.3%, 4.65%, 3.54, 2.07%, and 1.38% at 0.8, 0.85, 0.9, 0.95, and 1 of the nominal voltage, respectively, on average. It should be noted that at low voltages, regardless of temperature, the operating frequency is lower. On the other hand, the frequency headroom at elevated temperatures due to ITD is more significant. Moreover, at low-operating voltages, the energy consumption of the core is lower. Thus, for selecting the optimal operating point, not only should we take into account the energy-delay tradeoff, but also the benefits we get from ITD-aware dynamic thermal methods.

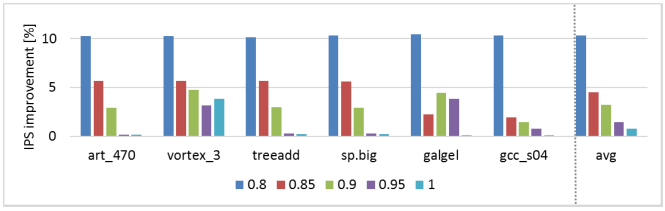


Fig. 8. IPS improvement using ITD-aware DFS for cold applications.

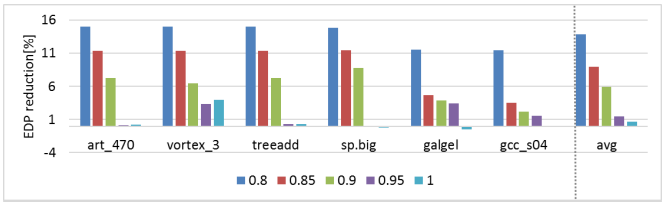


Fig. 9. EDP improvement using ITD-aware DFS for cold applications.

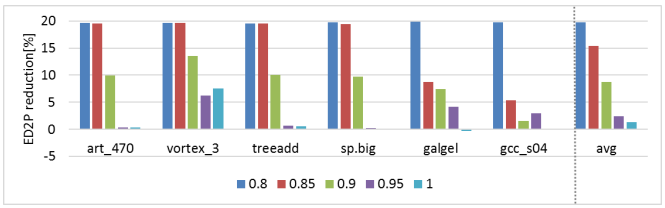


Fig. 10. ED2P improvement using ITD-aware DFS for cold applications.

Figs. 9 and 10 show the relative EDP and ED2P reductions, respectively, after applying ITD-aware DFS technique. Fig. 9 shows an average reduction of 15%, 11.4%, 7.2%, 0.16%, and 0.19% and 1.38% at 0.8, 0.85, 0.9, 0.95, and 1 of the nominal voltage, respectively. It should be noted that increasing the frequency at elevated temperatures increases the power consumption of the applications. However, since the performance is enhanced, the application finishes faster, resulting in little or no increase in the energy and, therefore,

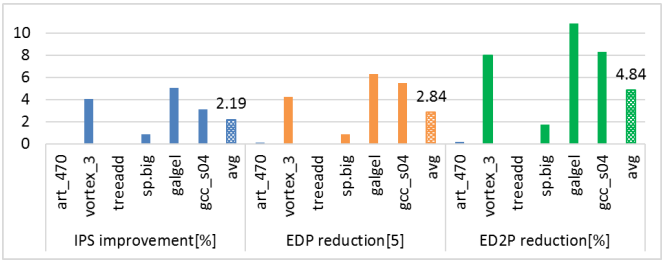


Fig. 11. IPS, EDP, and ED2P improvement using ITD-aware DVFS for cold applications.

reduction in the EDP. The ED2P results show even more reductions. Based on Fig. 10, the ED2P is reduced by 19.66%, 19.55%, 9.94%, 0.32%, and 0.33% at 0.8, 0.85, 0.9, 0.95, and 1 of the nominal voltage, respectively. At the nominal voltage, the ITD-aware DFS shows negligible gain in the energy efficiency. However, at the operating voltage of 0.8 of the nominal voltage, the energy efficiency is enhanced by up to 19%.

2) ITD-Aware Dynamic Voltage Frequency Scaling: DVFS has been traditionally used not only to prevent the processor from reaching critical voltages but also to enhance the energy efficiency of the processor. In an ITD-aware DVFS technique, for each voltage-level multiple frequencies, options are available, which are selected based on the temperature. More specifically, at each time interval, first we select the voltage level. Subsequently, taking into account the frequency headroom at the operating temperature, we select the frequency. Fig. 11 shows the results for the studied applications. It should be noted that for the cold applications in which temperature never reaches critical values, after the start of the application, the configuration will be set to the highest voltage to achieve the highest performance, and thus, the improvement observed from the ITD-aware DVFS over the ITD-agnostic scheme is not significant.

B. Hot Applications

In this section, we explore the applications that reach extreme temperatures at some point during their execution. For this class of applications, a DTM technique is required.

1) Throttling: Throttling is one of the DTM techniques utilized to control the temperature of the processor. When the core reaches critical temperatures, the application execution is halted to cool down the processor. After the temperature drops, the application is resumed. In our simulations, we halt the application whenever its temperature reaches 90 °C, and resume the execution as soon as the temperature is back to lower temperatures (a trigger temperature). We experiment with multiple trigger temperatures (65 °C, 70 °C, 75 °C, 80 °C, and 85 °C) in our simulations. To get the best performance for both AM and ITD-aware AM, we set the trigger temperature to 80 °C.

The ITD-aware algorithm for throttling is depicted in Fig. 7. In this algorithm, we increase the frequency of the hot applications when the temperature rises above the values shown in Table III. We analyze various operating points. Fig. 12 shows that the performance gain using ITD-aware throttling

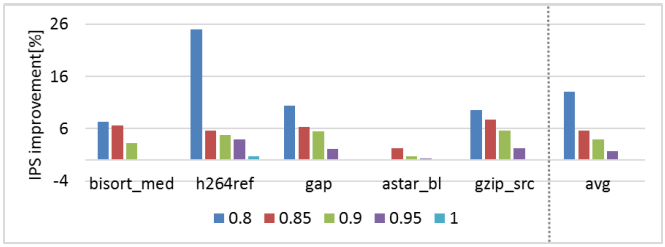


Fig. 12. IPS improvement using ITD-aware throttling for hot applications.

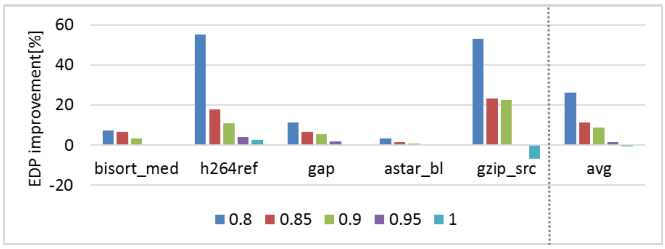


Fig. 13. EDP improvement using ITD-aware throttling for hot applications.

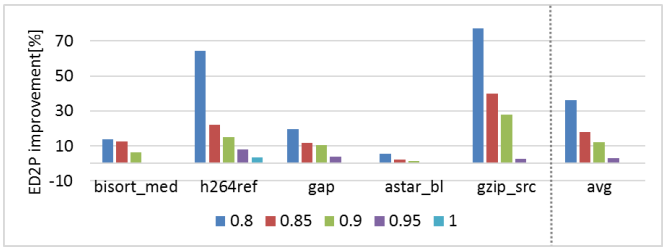


Fig. 14. ED2P improvement using ITD-aware throttling for hot applications.

is more significant at lower voltages, which was expected. At the nominal voltage, the gain of ITD-aware DFS is much lower. On average, the performance of the studied applications improves 13.1%, 5.7%, 4%, 1.76%, and 0.21% at 0.8, 0.85, 0.9, 0.95, and 1 of the nominal voltage, respectively.

A comparison of IPS improvement leveraging ITD between the hot and cold applications shows that while hot applications generally work at elevated temperatures, and therefore allows us to operate the processor at higher frequencies in the ITD-aware techniques, the temperature of the processor reaches critical points more frequently for these applications, which results in IPS reduction due to more frequent throttling. Moreover, while cold applications mostly achieve similar improvements in terms of IPS, the IPS gain of the ITD-aware technique for the hot applications varies significantly across different applications, and this is due to the more uniform thermal behavior of cold applications.

Figs. 13 and 14 show the relative EDP and ED2P reductions, respectively when applying ITD-aware throttling. Fig. 13 shows an average reduction of 26.1%, 11.19%, 8.7%, 1.39%, and −0.79% at 0.8, 0.85, 0.9, 0.95, and 1 of the nominal voltage, respectively. Again, increasing the frequency at elevated temperatures increases the power consumption of the applications. Given the enhancement in the performance, the energy and/or the energy efficiency is also improved.

An important observation from Fig. 13 is that for some applications (i.e., gzip_source and gap), the EDP is actually increased (negative reduction in the EDP shown in Fig. 13). This means that for these particular applications, the increase

in the power consumption at higher temperatures enabled by the ITD-aware frequency increase cancels out the performance enhancement in the EDP. EDP is a function of both the delay and power of the design. The ITD-aware AM technique allows the frequency to increase by a small amount at elevated temperatures, reducing the delay and, thus, the EDP. On the other hand, the increase in the frequency increases the power and, thus, the EDP. Whether the overall EDP is reduced or increased is based on the interplay effect of the power and delay of the ITD-aware technique. For this application, the increase in the power is dominant, and thus, no reduction in the EDP is observed. Moreover, the energy overhead of the ITD-aware technique at high temperature is more significant, as the elevated temperature induces more frequency switching. In fact, for this application, EDP has increased by 4%. However, for the same application, ED2P remains the same.

Based on Fig. 14, on average, the ED2P is reduced by 36.11%, 17.8%, 12.27%, 3.08%, and 0.65% at 0.8, 0.85, 0.9, 0.95, and 1 of the nominal voltage, respectively. At the nominal voltage, the ITD-aware throttling shows negligible gain in the energy efficiency. However, at the operating voltage of 0.8 of the nominal voltage, the energy efficiency is enhanced by up to 64%. Moreover, the results show that while the IPS improvement values are comparable with the ones for the cold applications, the EDP and ED2P values are generally more significant for the hot applications. This is due to the fact that for hot applications, the increase in the power consumption when applying ITD-aware technique is less significant compared with cold applications.

2) DVFS: For the hot applications, DVFS is used to control the temperature of the processor and enhance the energy efficiency. Whenever the processor reaches high temperatures, the voltage and frequency are scaled down by changing processor configuration to a lower voltage-frequency pair. At low temperatures, the voltage and frequency are scaled up by changing processor configuration to a higher voltage-frequency pair. In the ITD-aware technique, the voltage–frequency pairs are determined by the temperature too. Thus, always the highest frequency allowed by the temperature is used. The reduction in the voltage manages the power consumption and acts as the dynamic technique to reduce the temperature. For the hot applications, the voltage changes dynamically throughout the execution of the application.

In Fig. 15, the improvement in the IPS and energy efficiency by using ITD is reported. Five voltage levels are explored for the studied DVFS. Based on Fig. 15, the gain of the ITD-aware technique is different across various applications. It ranges between 2% and 28% for gap and bisort_med. It should be noted that based on Fig. 15, bisort-medical application maintains a high temperature during its execution. Thus, for this application, the ITD-aware technique is more beneficial. The EDP and ED2P results ranges between 5.4% and 33% and 7.9% and 48%, respectively for the same two applications, i.e., gap and bisort_med.

The results reported in this section are for single-core architecture. In a multicore architecture that executes the same application on one of the cores, the same results are expected.

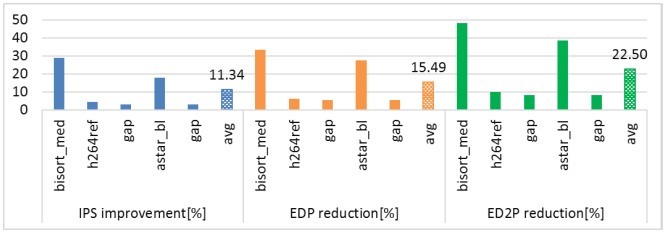


Fig. 15. IPS, EDP, and ED2P improvement using ITD-aware DVFS for hot applications.

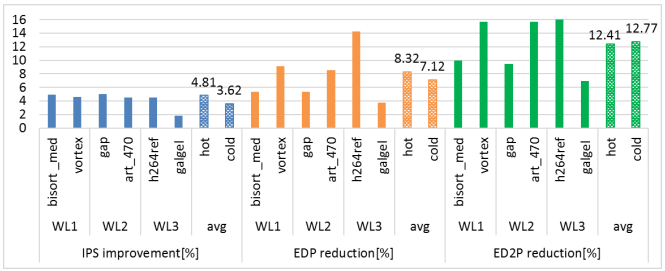


Fig. 16. IPS, EDP, and ED2P improvement using ITD-aware AM in a multicore architecture.

This is due to the fact that, we use the temperature of each individual core to make decision about the operating frequency. However, the AM can be deployed in only a multicore architecture, where the applications are migrated to colder cores when reaching critical temperatures.

3) Activity Migration: To study the ITD-aware techniques for multicore architectures, we simulate a dual-core processor simultaneously running workloads combining hot and cold applications. With an operating voltage of 85% of v_{nom}, the frequency is increased from 3.90 to 4.4 GHz at elevated temperatures. The studied work loads include WL1: [bisort_med, vortex_03], WL2: [gap, art_470], and WL3: [*h*264ref, galgel]. Among the studied workloads, [bisort_med, vortex_03] and [h264ref, galgel] have the highest and lowest contrast in their maximum temperature, respectively.

Fig. 16 shows the results of improvement in the IPS and reduction in the EDP and EDP2 after applying an ITD-aware policy to the AM. The results show that both hot and cold applications benefit from the ITD-aware technique. We studied ITD-aware techniques for the single-core and multicore architecture. It should be noted that the basic idea of proposed ITD-aware scheme is to leverage the reduction in the path delays at elevated temperatures due to the characteristics of submicrometer CMOS technologies. Since the core temperature depends on the level of activities on a core, whether the activities and the corresponding temperature come from one thread or multiple, multithreaded processors can benefit from the proposed ITD-aware techniques; however, the migration overheard of moving all the running applications is higher.

C. Thermal Behavior

To observe how the system evolves along time with the proposed thermal management policies, we have depicted the thermal behavior for 35 ms in a dual-core system for WL1 using AM. Fig. 17(a) shows the ping-pong thermal

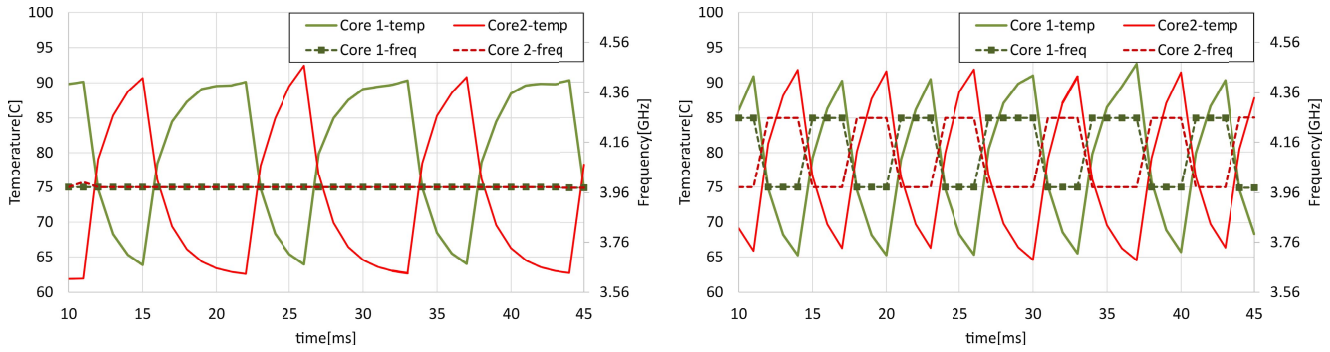


Fig. 17. Thermal behavior for dual-core system pairing lbm and bisort-medical using (a) AM and (b) using ITD-aware AM.

TABLE IV
IPS RESULTS FOR ALL THE DTM TECHNIQUES

	IPS [MIPS]													
	no DTM (cold apps)/ throttle (hot apps)					ITD-aware DFS (cold apps) / throttle (hot apps)					DVFS	ITD-aware DVFS	AM	ITD-aware AM
$\frac{v_{op}}{v_{nom}}$	0.8	0.85	0.9	0.95	1	0.8	0.85	0.9	0.95	1			0.85	0.85
art_470	1.62	1.68	1.70	1.72	1.79	1.79	1.77	1.75	1.72	1.80	1.69	1.69	1.68	1.75
vortex_3	1.41	1.52	1.54	1.63	1.68	1.55	1.60	1.62	1.68	1.74	1.49	1.55	1.41	1.47
sp.big	1.63	1.68	1.69	1.7	1.71	1.79	1.77	1.74	1.71	1.72	1.72	1.74	-	-
galgel	1.63	1.73	1.94	2.14	2.34	1.80	1.83	2.04	2.27	2.46	2.15	2.26	1.73	1.77
gcc_s04	1.63	1.74	1.82	1.90	1.94	1.80	1.78	1.90	1.97	1.95	1.68	1.73	-	-
bisort_med	1.30	1.43	1.62	1.81	1.90	1.43	1.46	1.64	1.82	1.90	1.65	2.12	1.77	1.86
h264ref_foreman	1.36	1.52	1.69	1.95	2.01	1.36	1.52	1.69	1.95	2.01	1.59	1.66	1.87	1.96
gap	1.250	1.44	1.51	1.63	1.73	1.25	1.44	1.51	1.63	1.73	1.61	1.65	1.78	1.87
astar_biglakes	1.35	1.45	1.58	1.81	1.90	1.35	1.45	1.58	1.81	1.90	1.17	1.38	-	-
gzip_source	1.15	1.22	1.29	1.36	1.46	1.15	1.22	1.29	1.36	1.46	1.35	1.35	-	-

TABLE V
EDP RESULTS FOR ALL THE EDP TECHNIQUES

	EDP [$10^{12} ws^2$]													
	no DTM (cold apps)/ throttle (hot apps)					ITD-aware DTM (cold apps)/ throttle (hot(apps)					DVFS	ITD-aware DVFS	AM	ITD-aware AM
$\frac{v_{op}}{v_{nom}}$	0.8	0.85	0.9	0.95	1	0.8	0.85	0.9	0.95	1			0.85	0.85
art_470	10.7	13.8	14.8	15.8	18.1	9.1	12.2	13.7	15.8	18.1	16.6	16.6	13.8	12.6
vortex_3	9.3	12.4	14.7	17.6	19.9	8.0	11.0	13.7	17.0	19.2	13.1	12.6	12.4	11.3
sp.big	10.7	12.0	13.0	13.6	14.0	9.1	10.7	12.1	13.6	14.0	13.6	13.4	-	-
galgel	8.1	9.4	11.7	14.2	16.4	6.9	8.4	11.2	13.2	15.6	11.8	11.0	9.4	9.1
gcc_s04	10.8	12.2	14.6	17.1	19.2	9.6	11.7	14.0	16.5	19.3	13.7	13.0	-	-
bisort_med	9.0	10.6	14.0	17.1	19.6	8.0	10.2	13.7	16.8	19.6	14.8	9.8	9.5	9.0
h264ref_foreman	8.7	10.8	13.7	17.5	19.9	7.0	9.1	12.4	16.8	19.4	11.7	11.0	9.8	8.3
gap	8.2	10.1	12.3	14.4	16.9	8.2	10.1	12.3	14.4	16.9	12.4	11.8	9.1	8.7
astar_biglakes	9.3	10.7	13.0	16.0	18.1	9.3	10.7	13.0	16.0	18.1	10.7	7.8	-	-
gzip_source	4.3	5.1	6.0	6.9	8.1	4.3	5.1	6.0	6.9	8.1	6.9	6.9	-	-

behavior expected in AM. The cores take turns executing the hot applications, while the other cores cool down during the execution of the cold application. The core frequency is fixed in this ITD-ignorant scheme. Fig. 17(b) shows the temperature and frequency in the ITD-aware scheme. The ITD-aware scheme uses only two frequency levels with the lower frequency for core temperatures under 75 °C and the higher frequency for core temperatures above 75 °C. As expected, the number of migrations increases in this scheme, since the frequency is increased at temperatures higher than 75 °C, necessitating a faster migration for the hot application. In this scheme, the frequency of the core running the hot application is increased at elevated temperatures, which in turn increases the IPS for the hot application.

X. CROSS COMPARISON

In Section IX, we compared the relative speedup, EDP, and ED2P results for various studied techniques. However, in order to find the best thermal management method, we need to compare the absolute values. Tables IV and V show the absolute values for the IPS and EDP, respectively, for both hot and cold applications. The first five columns in each table show the execution time of the application at $v_{\text{op}}/v_{\text{nom}}$ of 0.8, 0.85, 0.9, 0.95, and 1, where v_{op} is the operating voltage and v_{nom} is the nominal voltage. For these results, throttling is used in response to thermal emergency. The second five columns show the absolute values when the ITD-aware algorithm is utilized. Moreover, the last four columns show the results for DVFS, ITD-aware DVFS, AM, and ITD-aware AM.

Table IV shows that the ITD-aware technique at the nominal voltage yields the highest IPS for most applications. This was expected, as operating the processor at the nominal voltage allows operating at higher frequency. However, for some cases (e.g., bisort_medical and spbig), the ITD-aware DVFS yields higher IPS. Table V shows that the lowest EDP values are achieved with the ITD-aware techniques at low-operating voltages (e.g., ($\text{vol}_{\text{op}}/\text{vol}_{\text{nom}} = 0.8$)), due to the low power dissipation. Moreover, the ITD impact on increasing

the frequency is more significant, resulting in lower EDP and better energy efficiency.

XI. Conclusion

In this paper, we performed a thorough analysis of the ITD at both circuit level and application level. We simulated the critical path delays of the LEON3 processor. We showed that the ITD is more significant at lower voltages, allowing a frequency headroom at elevated voltages, which can be exploited to speedup of the execution of applications. We studied several known DTM techniques and proposed an enhanced ITD-aware solution for each of studied DTM techniques to enhance their performance, both in terms of IPS and energy efficiency. The results show that using the proposed ITD-aware techniques, we can enhance the IPS of the DTM techniques by up to 28%. For IPS optimization purposes, operating the processor at the nominal voltage or using DVFS yields the lowest execution time. For energy-efficiency purposes, operating the processor at lower than nominal voltages is more beneficial. Overall, the ITD-aware techniques result in lower execution time and EDP.

References

[1] K. Neshatpour, H. Homayoun, A. Khajeh, and W. Burleson, "Revisiting dynamic thermal management exploiting inverse thermal dependence," in *Proc. 25th Ed. Great Lakes Symp. VLSI (GLSVLSI)*, 2015, pp. 385–390.

[2] W. Huang, K. Rajamani, M. R. Stan, and K. Skadron, "Scaling with design constraints: Predicting the future of big chips," *IEEE Micro*, vol. 31, no. 4, pp. 16–29, Jul. 2011.

[3] A. Sassone *et al.*, "Investigating the effects of inverted temperature dependence (ITD) on clock distribution networks," in *Proc. DATE*, 2012, pp. 165–166.

[4] Y. K. Cao, "Predictive technology models," Nanoscale Integr. Model. (NIMO) Group, Arizona State Univ., Tempe, AZ, USA, Tech. Rep. V2.1, 2012.

[5] M. Cho, M. Khellah, K. Chae, K. Ahmed, J. Tschanz, and S. Mukhopadhyay, "Characterization of inverse temperature dependence in logic circuits," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2012, pp. 1–4.

[6] A. Dasdan and I. Hom, "Handling inverted temperature dependence in static timing analysis," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 11, no. 2, pp. 306–324, Apr. 2006.

[7] K. Kanda, K. Nose, H. Kawaguchi, and T. Sakurai, "Design impact of positive temperature dependence on drain current in sub-1-V CMOS VLSIs," in *Proc. IEEE Custom Integr. Circuits Conf.*, May 1999, pp. 563–566.

[8] P. Verma and R. Mishra, "Temperature dependence of propagation delay characteristic in LECTOR based CMOS circuit," *Int. J. Comput. Appl. Electron., Inf. Commun. Eng.*, pp. 28–30, 2011.

[9] A. Calimera, R. I. Bahar, E. Macii, and M. Poncino, "Temperature-insensitive dual-Vth synthesis for nanometer CMOS technologies under inverse temperature dependence," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 11, pp. 1608–1620, Nov. 2010.

[10] M. A. A. Latif, N. B. Z. Ali, and F. A. Hussin, "Design for cold test elimination—Facing the inverse temperature dependence (ITD) challenge," in *Proc. ISCAS*, May 2012, pp. 3081–3085.

[11] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan, "Temperature-aware microarchitecture," in *Proc. ISCA*, 2003, pp. 2–13.

[12] K. Sankaranarayanan, S. Velusamy, M. Stan, and K. Skadron, "A case for thermal-aware floorplanning at the microarchitectural level," *J. Instruct.-Level Parallelism*, vol. 7, no. 1, pp. 8–16, 2005.

[13] J. Donald and M. Martonosi, "Temperature-aware design issues for SMT and CMP architectures," in *Proc. Workshop Complexity-Effective Design*, 2004, pp. 1–11.

[14] J. Donald and M. Martonosi, "Techniques for multicore thermal management: Classification and new exploration," *Comput. Archit. News*, vol. 34, no. 2, pp. 78–88, 2006.

[15] X. Yin *et al.*, "Efficient implementation of thermal-aware scheduler on a quad-core processor," in *Proc. IEEE 10th Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Nov. 2011, pp. 1076–1082.

[16] T. Ebi, M. A. Al Faruque, and J. Henkel, "TAPE: Thermal-aware agent-based power econom multi/many-core architectures," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2009, pp. 302–309.

[17] D.-C. Juan and D. Marculescu, "Power-aware performance increase via core/uncore reinforcement control for chip-multiprocessors," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED)*, 2012, pp. 97–102. [Online]. Available: http://doi.acm.org/10.1145/2333660.2333686

[18] S. Herbert and D. Marculescu, "Analysis of dynamic voltage/frequency scaling in chip-multiprocessors," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2007, pp. 38–43. [Online]. Available: http://doi.acm.org/10.1145/1283780.1283790

[19] R. Mukherjee and S. O. Memik, "Physical aware frequency selection for dynamic thermal management in multi-core systems," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2006, pp. 547–552.

[20] E. Kursun and C.-Y. Cher, "Variation-aware thermal characterization and management of multi-core architectures," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, Oct. 2008, pp. 280–285.

[21] V. Hanumaiah, S. Vrudhula, and K. S. Chatha, "Performance optimal online DVFS and task migration techniques for thermally constrained multi-core processors," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 11, pp. 1677–1690, Nov. 2011.

[22] C. Zhu, Z. Gu, L. Shang, R. P. Dick, and R. Joseph, "Three-dimensional chip-multiprocessor run-time thermal management," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 27, no. 8, pp. 1479–1492, Aug. 2008.

[23] B. Razavi, *Design of Analog CMOS Integrated Circuits*, 1st ed. New York, NY, USA: McGraw-Hill, 2001.

[24] J. M. Daga, E. Ottaviano, and D. Auvergne, "Temperature effect on delay for low voltage applications [CMOS ICs]," in *Proc. Design, Autom. Test Eur.*, Feb. 1998, pp. 680–685.

[25] I. M. Filanovsky and A. Allam, "Mutual compensation of mobility and threshold voltage temperature effects with applications in CMOS circuits," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 48, no. 7, pp. 876–884, Jul. 2001.

[26] S.-Y. Kim *et al.*, "Temperature dependence of substrate and drain–currents in bulk FinFETs," *IEEE Trans. Electron Devices*, vol. 54, no. 5, pp. 1259–1264, May 2007.

[27] W. Lee, Y. Wang, T. Cui, S. Nazarian, and M. Pedram, "Dynamic thermal management for FinFET-based circuits exploiting the temperature effect inversion phenomenon," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2014, pp. 105–110. [Online]. Available: http://doi.acm.org/10.1145/2627369.2627608

[28] X. Huang *et al.*, "Sub-50 nm p-channel FinFET," *IEEE Trans. Electron Devices*, vol. 48, no. 5, pp. 880–886, May 2001.

[29] Y. Cao, T. Sato, M. Orshansky, D. Sylvester, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, May 2000, pp. 201–204.

[30] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2816–2823, Nov. 2006.

[31] H. Ohta *et al.*, "High performance 30 nm gate bulk CMOS for 45 nm node with Σ-shaped SiGe-SD," in *IEDM Tech. Dig.*, Dec. 2005, pp. 1–4.

[32] K. Goto *et al.*, "High performance 25 nm gate CMOSFETs for 65 nm node high speed MPUs," in *IEDM Tech. Dig.*, Dec. 2003, pp. 1–27.

[33] Z. Luo *et al.*, "High performance and low power transistors integrated in 65 nm bulk CMOS technology," in *IEDM Tech. Dig.*, Dec. 2004, pp. 661–664.

[34] V. Chan *et al.*, "High speed 45 nm gate length CMOSFETs integrated into a 90 nm bulk technology incorporating strain engineering," in *IEDM Tech. Dig.*, Dec. 2003, pp. 3–8.

[35] A. Gaisler. (2010). *LEON3 Processor*, Nanoscale Integration Modeling (NIMO) Group. [Online]. Available: http://www.gaisler.com/index.php/products/processors/leon3

[36] H. Tajik, H. Homayoun, and N. Dutt, "VAWOM: Temperature and process variation aware wearout management in 3D multicore architecture," in *Proc. ACM DAC*, 2013, Art. no. 178.

[37] D. M. Tullsen, "Simulation and modeling of a simultaneous multithreading processor," in *Proc. CMG Conf.*, 1996, pp. 819–828.

[38] Open Systems Group. (2013). *Standard Performance Evaluation Corporation*. [Online]. Available: http://www.spec.org

[39] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. MICRO*, Dec. 2009, pp. 469–480.
[40] Y. Liu, H. Yang, R. P. Dick, H. Wang, and L. Shang, "Thermal vs energy optimization for DVFS-enabled processors in embedded systems," in *Proc. ISQED*, Mar. 2007, pp. 204–209.
[41] H. Homayoun, M. Makhzan, and A. Veidenbaum, "Multiple sleep mode leakage control for cache peripheral circuits in embedded processors," in *Proc. Int. Conf. Compil., Archit. Synth. Embedded Syst.*, 2008, pp. 197–206.
[42] D. Zhao, H. Homayoun, and A. V. Veidenbaum, "Temperature aware thread migration in 3D architecture with stacked dram," in *Proc. 14th Int. Symp. IEEE Quality Electron. Design (ISQED)*, Mar. 2013, pp. 80–87.
[43] A. K. Coskun, R. Strong, D. M. Tullsen, and T. S. Rosing, "Evaluating the impact of job scheduling and power management on processor lifetime for chip multiprocessors," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 37, no. 1, pp. 169–180, 2009.
[44] R. Kumar, K. Farkas, N. P. Jouppi, P. Ranganathan, and D. M. Tullsen, "Processor power reduction via single-ISA heterogeneous multi-core architectures," *IEEE Comput. Archit. Lett.*, vol. 2, no. 1, p. 2, Jan./Dec. 2003.
[45] R. Strong, J. Mudigonda, J. C. Mogul, N. Binkert, and D. Tullsen, "Fast switching of threads between cores," *ACM SIGOPS Oper. Syst. Rev.*, vol. 43, no. 2, pp. 35–45, 2009.
[46] S. Heo, K. Barr, and K. Asanović, "Reducing power density through activity migration," in *Proc. Int. Symp. Low Power Electron. Design*, 2003, pp. 217–222.
[47] M. Kondo and H. Nakamura, "Dynamic processor throttling for power efficient computations," in *Proc. Int. Workshop Power-Aware Comput. Syst.*, 2004, pp. 120–134.

Katayoun Neshatpour received the M.S. degree from Sharif University of Technology, Tehran, Iran, with a focus on the VLSI implementation of an MIMO detector applied to the long-term evolution. She is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA.

Her current research interests include the hardware acceleration of big data applications, with a focus on the implementation of several machine learning algorithms in Apache Hadoop and efficient implementation of convolutional neural networks.

Ms. Neshatpour was a recipient of the three-year Presidential Fellowship and a one-year supplemental Electronics and Communication Engineering Department Scholarship advised by Dr. H. Homayoun and co-advised by Dr. Sasan.

Wayne Burleson (S'87–M'89–SM'01–F'11) received the B.S. and M.S. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA and the Ph.D. degree from the University of Colorado, Denver, CO, USA.

He has been a Professor of Electrical and Computer Engineering at the University of Massachusetts Amherst, Amherst, MA, USA, since 1990. He was a Custom Chip Designer and a Consultant in the semiconductor industry with VLSI Technology, San Jose, CA, USA; DEC, Maynard, MA, USA; Compaq/HP, Shrewsbury/Hudson, MA, USA; Intel, Shrewsbury/Hudson, MA, USA; Rambus, Sunnyvale, CA, USA; and AMD, Boston, MA, USA. He was a Visiting Professor at École Nationale Supérieure des Télécommunications, Paris, France, from 1996 to 1997; the Laboratory of Informatics, Robotics and Microelectronics of Montpellier, Montpellier, France, in 2003; and the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, from 2010 to 2011. He is currently a Senior Fellow at AMD Research, Boston. He teaches courses in VLSI design, embedded systems, and security engineering. He has authored over 200 refereed publications in these areas. His current research interests include the general area of VLSI, including circuits and CAD for low power, long interconnects, clocking, reliability, thermal effects, process variation, noise mitigation, hardware security, reconfigurable computing, content-adaptive signal processing, radio frequency identification, and multimedia instructional technologies.

Amin Khajeh (S'01–M'11–SM'14) received the B.Sc. degree in electrical engineering and communication from Shiraz University, Shiraz, Iran, in 2002, the M.Sc. degree in electrical engineering from The University of Texas at Arlington, Arlington, TX, USA, in 2005, and the Ph.D. degree in electrical engineering and computer sciences from the University of California at Irvine, Irvine, CA, USA, in 2010.

He was an Intern at the Research and Development Division, Siemens Company, Austin, TX, USA, in 2002, where he was a Research Staff Member from 2002 to 2003. He was with the Qualcomm Low Power DSP Team, Austin, TX, USA, from 2010 to 2012, where he was involved in researching and implementing advance low-power techniques for DSP cores for wireless multimedia applications. He was with Circuit Research Labs, Hillsboro, OR, USA, and Intel Research Labs, Hillsboro, from 2012 to 2014, where he was involved in the low-power system-on-chip (SoC) design. He is currently a Principal Scientist with Broadcom Limited, San Jose, CA, USA, where he is involved in the low-power design and methodology for network switches. He has authored over 30 technical papers and holds five patents on these subjects. His current research interests include the low-power design and methodology for SoCs, design of low-power high-performance circuits, and high-performance high-yield memory design.

Houman Homayoun received the B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2003, the M.S. degree in computer engineering from the University of Victoria, Victoria, BC, Canada, in 2005, and the Ph.D. degree from the Department of Computer Science, University of California at Irvine, Irvine, CA, USA, in 2010.

He was an NSF Computing Innovation Fellow at the University of California at San Diego, San Diego, CA, USA, for two years. He is currently an Assistant Professor at the Electronics and Communication Engineering Department, George Mason University, Fairfax, VA, USA. He also holds a joint appointment with the Computer Science Department. He is currently leading a number of research projects, including the design of heterogeneous architectures for big data and nonvolatile logics to enhance design security, which are funded by the National Science Foundation, General Motors Company, and the Defense Advanced Research Projects Agency.

Dr. Homayoun was a recipient of the NSF Computing Innovation Fellowship from the CRA and CCC.