Revisiting Dynamic Thermal Management Exploiting Inverse Thermal Dependence

Katayoun Neshatpour
George Mason University
kneshatp@gmu.edu

Amin Khajeh
Broadcom Corporation
amink@broadcom.com

Wayne Burleson
AMD
University of Massachusetts, Amherst
burleson@ecs.umass.edu

Houman Homayoun
George Mason University
hhomayou@gmu.edu

ABSTRACT

As CMOS technology scales down towards nanometer regime and the supply voltage approaches the threshold voltage, increase in operating temperature results in increased circuit current, which in turn reduces circuit propagation delay. This paper exploits this new phenomenon, known as inverse thermal dependence (ITD) for power, performance, and temperature optimization in processor architecture. ITD changes the maximum achievable operating frequency of the processor at high temperatures. Dynamic thermal management techniques such as activity migration, dynamic voltage frequency scaling, and throttling are revisited in this paper, with a focus on the effect of ITD. Results are obtained using the predictive technology models of 7nm, 10nm 14nm and 20nm technology nodes and with extensive architectural and circuit simulations. The results show that based on the design goals, various design corners should be re-investigated for power, performance and energy-efficiency optimization. Architectural simulations for a multi-core processor and across standard benchmarks show that utilizing ITD-aware schemes for thermal management improves the performance of the processor in terms of speed and energy-delay-product by 8.55% and 4.4%, respectively.

Keywords

ITD, memory-intensive, compute-intensive, IPS, energy-delay-product

1. INTRODUCTION

Reducing the feature sizes has been an ongoing trend in the processor design technology. If the dimensions are scaled without scaling of the supply and the threshold voltage, the power density of the processor will increase dramatically, which will cause overheating and reliability issues. Thus, while heat sinks are used to extract the heat out of the processor, 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, finding techniques for low-power and low-temperature designs has been a hot topic in the processor design.

As the integrated circuits enter into the deep-sub-micron CMOS technologies, their thermal behavior changes significantly. Traditionally, as temperature increases, propagation delay of the circuit also increases due to the lower mobility, which slows down the circuit. In the sub-micron 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).

ITD introduces many new aspects in the design and optimization of the circuit. Design corners will have to be re-investigated and power reduction and thermal management techniques will have to be adapted, accordingly. With ITD, at high temperatures the circuit will be able to benefit from the frequency headroom offered due to reduction in 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.

This paper investigates ITD in future CMOS technologies to understand whether further scaling will increase the significance of the ITD. In this work, several design optimal points are explored for the 7nm, 10nm, 14nm and 20nm predictive technology models (PTM). 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, etc.) or an invert-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, a multi-core architecture is studied. Different temperature-aware techniques including throttling, activity migration (AM), and dynamic voltage frequency scaling (DVFS) are explored to find the best solution in terms of performance, power, thermal behavior, and energy-delay product. To the best of our knowledge, this is the first paper that analyzes the impact of ITD on a fully synthesizable processor core at the circuit level and in deep nanometer technologies, and exploits this new

GLSVLSI'15, May 20–22, 2015, Pittsburgh, PA, USA.

http://dx.doi.org/10.1145/2742060.2742086.

phenomenon to manage power, performance and temperature at the architectural level.

This paper in brief makes the following contributions: (1) ITD characterization by performing spice-level analysis on LEON3 processor at various voltage and temperature corners for PTM technologies to understand ITD trends with technology scaling. (2) Proposing ITD-aware dynamic thermal management including throttling, AM and DVFS. (3) Analyzing simulation results for the comparison of different thermal management techniques in the presence of ITD at various operating points.

The paper is organized as follows: Section 2 provides background on ITD and reports its impact on maximum achievable operating clock frequency in LEON3 processor for four predictive models. In section 3, a number of SPEC and media applications are characterized in terms of frequency-sensitivity and thermal behavior. In section 4, we introduce ITD-aware thermal management techniques. Section 5 introduces the simulation framework. Section 6 presents the results of the simulations and evaluates them. Section 8 presents our conclusion remarks.

2. INVERSE THERMAL DEPENDENCE

2.1 Background

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 model the behavior of the CMOS transistors. The square law model in (1) estimates the drain current of CMOS transistor operating in the saturation mode [1].

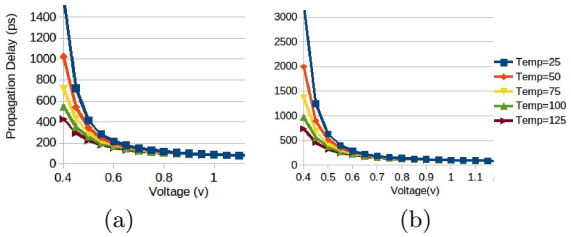
$$I_D = \mu \times C_{ox} \frac{W}{L} (V_{gs} - V_t)^2. \quad (1)$$

In (1), 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 the 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 will enhance the device speed according to (1). At high operating voltages, the threshold voltage change will have little effect on $(V_{gs} - V_t)$ and thus, the mobility will determine the effect of the temperature on the device speed. However, at low operating voltages, the threshold voltage will determine the effect of temperature. As the technology scales down, the operating voltage approaches the threshold voltage. Thus, when working at slightly lower than the nominal voltage, the device is switching faster at higher temperature. This phenomenon is called the inverse thermal dependence.

2.2 ITD in predictive technologies

Latest work have explored the effect of ITD in current technology nodes [2], [3] and [4]; however, it is important to explore ITD impact in the future nodes, as well. This paper first investigates ITD in future CMOS technologies to understand whether further scaling will increase the significance of the ITD. We use PTM for this purpose [5]. PTM provides accurate, customizable, and predictive models for future transistor and interconnect technologies.

We synthesize LEON3, a fully synthesizable VHDL model for open SPARC V8 32-bit architecture [6] using Design Compiler and extract the critical path for further explorations. Based on the PTM transistor models, we run Spice simulations at different temperatures with variable supply voltages to find the effect of temperature on the the critical path delays. To provide more accurate analysis of ITD characteristics in processor design, in these simulations the complete critical paths were explored rather than just a few basic cells. (NAND, XOR, etc.) or an invert-based ring oscillator.



(a) (b)

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

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 20nm. For our simulations, we assumed different wire lengths (1μm, 50μm and 100μm). 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 20% of our post-layout top critical path will be RC-delay and 80% cell-delay, the difference between pre-layout and post-layout speedup for 25°C to 125°C, will be 7% lower in 20nm at nominal voltage. In other words, post-layout speedup as a result of increase in temperature is lower than that of pre-layout. 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 don't have access to a full library of cells at 14nm, 10nm, and 7nm, for the remainder of the paper we will use the pre-layout results.

It should be noted that, several factors such as device aging may change the critical paths of a design. The intent here, is to show the trend of temperature effect as we scale down the technology and build a model for selected case-study in this paper using PTM device models. To apply the proposed technique to other designs, an accurate temperature-delay at different process-voltage-temperature (PVT) should be devised and utilized.

Fig. 1 shows LEON3 critical path delays, for the 7nm and 14nm PTM technologies. The nominal voltages are 0.7v and 0.8v for 7nm and 14nm technologies, respectively.

Fig. 1(b) shows that for 14nm technology, with an operating voltage that is 85% of the nominal voltage, the critical path delay at 75°C, 100°C and 125 °C is 7%, 14% and 17% lower than the critical path delay at 50°C, respectively. For 7nm technology, with the same operating voltage of 85% of the nominal voltage, the critical path delay at 75°C, 100°C and 125 °C is 8%, 18% and 28% 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.

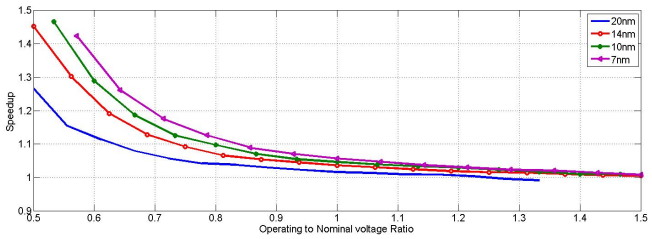


Figure 2: The reduction in propagation delay (speedup) when the temperature rises from 50°C to 75°C.

A comparison of Fig. 1(a) and 1(b) shows that, as the technology scales down, the ITD effect is observed more significantly. Moreover, for each technology the reduction in the critical path due to the elevated temperature is more significant at low operating voltages.

Fig. 2 shows the ratio of the critical path delay at 75°C to the delay at 50°C for the 7nm, 10nm, 14nm and 20nm PTM technologies at different operating voltages, which defines the speedup of the circuit when the temperature rises from 50°C to 75°C. The X-axis shows the ratio of the operating voltage to the nominal voltage. According to the figure, as we move away from the nominal voltage towards threshold voltage, the effect of temperature on the critical path delay increases. Moreover, Fig. 2 shows that the speedup at elevated temperatures is less significant for larger technologies.

As discussed earlier, the frequency headroom available at elevated temperatures due to the ITD effect can be used to speed up the processor. However, it should be noted that the circuit will have to operate at low voltages (lower than the nominal) in order to observe the ITD effect. Table 1 shows the range of speedup at elevated temperatures for various operating points for the 7nm, 10nm, 14nm and 20nm 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 1 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 will result in higher power dissipation and increased temperature. Thus, thermal management techniques will have to be used, which will in turn slow down the circuit. An overview of Table 1 shows that, the operating condition that yields the best performance cannot be easily determined, as it heavily depends on power and thermal profile of the processor, which is decided at run-time, is application dependent and requires architecture-level control.

Table 1: Speedup of the circuit at elevated temperatures

Tech.	$\frac{Op.\ Voltage}{Nom.\ Voltage}$	Speedup $50°C \rightarrow 75°C$	Speedup $50°C \rightarrow 100°C$	Speedup Nom. Voltage
7nm	0.65	1.261	1.542	3.680
	0.75	1.15	1.297	2.101
	0.85	1.089	1.81	1.426
10nm	0.65	1.186	1.377	3.636
	0.75	1.118	1.236	1.990
	0.85	1.07	1.137	1.405
14nm	0.65	1.166	1.334	3.263
	0.75	1.091	1.19	1.903
	0.86	1.052	1.121	1.384
20nm	0.65	1.089	1.178	2.406
	0.75	1.043	1.07	1.653
	0.85	1.035	1.069	1.292

3. CHARACTERIZATION OF APPLICATIONS UNDER ITD

3.1 Performance analysis under ITD

Traditionally, with large feature sizes (45nm and above), 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. (IPS = 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, will allow running the processor at higher frequencies. However, it should be noted, that colder regions including the L2 and last level cache (LLC), which are normally operating at low temperature [7], will still maintain their path delays. For example, for a cache delay of 100ns, increasing the frequency from 1GHz to 2GHz, will increase the access time from 100 cycles to 200 cycles, which will negatively affect the instruction per cycle (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 SMTSIM simulator [8] with Spec2000, Spec2006 [9], and Media benchmarks. Table 2 shows the micro-architecture parameters of our studied multicore architecture. The access time for individual SRAM units was captured using McPAT [10].

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. 3. 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 affect 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 of applications the performance increases with the increased frequencies.

In Fig. 3, the IPS drops 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 could even result in a reduction in the IPS.

3.2 Thermal behavior analysis

For thermal characterization of studied applications we used HotSpot [7]. The numbers next to each application

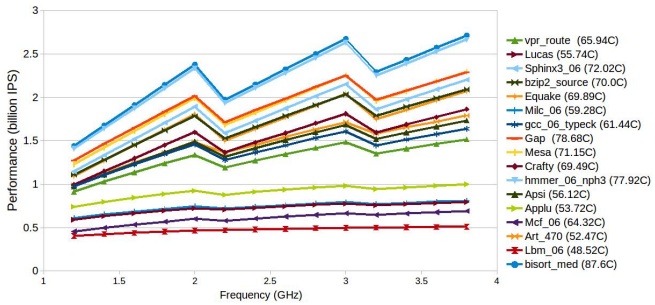


Figure 3: performance of Spec benchmarks as a function of frequency

Table 2: 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

name in Fig. 3 show 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 48.52°C to 78.6°C for these applications, which is about 30°C variation.

By comparing the performance and temperature results in Fig. 3, 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. In such applications, dynamic thermal control techniques are of high importance. If the compute-intensive application (hot application) is running on a single core without any feedback from the transient temperature, the core might reach critical temperatures which may damage the device or cause reliability issues.

Based on the results in Fig. 3, the applications are divided into two groups, the memory-intensive applications and the memory-non-intensive applications. For the memory-non-intensive applications, which could also be referred to as compute-intensive applications, using higher frequencies is more desirable, as the increased frequency will contribute significantly to the performance. Thus, these applications are categorized as frequency-sensitive. For the memory-intensive application, higher frequencies are not desirable as they will increase the power consumption with little contribution to the performance.

4. DYNAMIC THERMAL MANAGEMENT

DVFS [11], AM [12] and throttling [7, 13] are among the most well studied techniques to manage performance, power, and temperature in multi-core architectures. These techniques have been developed with the assumption that the device is slower at elevated temperatures. In this section, we will study how these techniques are effective in managing power, temperature and performance in the presence of ITD.

4.1 Throttling

The simplest way to prevent an application from reaching critical temperatures is throttling. In throttling when an application reaches a threshold temperature, the pipeline is halted until the core cools down [7]. During the throttling time the core remains idle. This results in an underutilization and therefore loss of performance.

4.2 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.

4.3 Dynamic Voltage and Frequency Scaling

Another effective technique for managing temperature is DVFS, in which the processor supply voltage and frequency are adapted dynamically to respond to thermal emergencies [11]. The number of voltage and frequency pair varies from two in Intels SpeedStep technology to 40 or even more in Intel XScale [14].

4.4 ITD-aware throttling, AM and DVFS

After running a frequency-sensitive application on a single core, the temperature will increase. With ITD, this elevated temperature will increase the maximum allowable frequency on the core. Thus, dynamic frequency changing will allow the application to run at higher frequencies when it reaches higher temperature.

Before reaching the critical temperature, the core will reach a mid-point threshold (In our simulations we set this mid-point temperature to 75°C). In the ITD-aware schemes, when the temperature of the core is higher than this mid-point threshold, we increase the frequency of the core to benefit from the frequency headroom offered by the ITD. Afterwards, we keep monitoring the temperature until it reaches the critical temperature (i.e., 90°C in this paper), at which point, throttling, AM or DVFS is done to reduce the temperature. If the temperature drops below the mid-point temperature 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 mid-point threshold and the critical value, to increase the overall IPS.

Note, that in addition to two different frequencies for 2 temperature thresholds, more level of temperature-frequency pair may be explored. Fig. 4 shows the ITD-aware AM algorithm for an n-core system with m levels of temperature-frequency pairs. Array T contains core temperatures, 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.

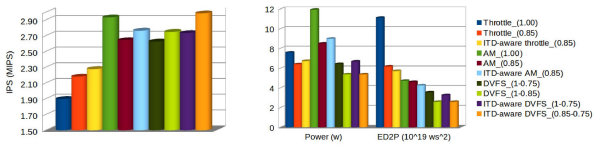
5. SIMULATION PLATFORM

In order to study the thermal characteristic of various applications, we study the behavior of SPEC2000, SPEC2006, and Media applications. Due to space limitation we only report the results for a representative subset of these benchmarks. We use an integrated version of SMTSIM, McPAT [10], and HotSpot-5.02 [7] 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 4 including throttling, AM and DVFS are implemented in the integrated

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

Figure 4: ITD-aware acitivty migration algorithm



(a) (b)

Figure 5: (a) Performance(b) Power and ED2P of thermal control techniques for various operating voltages for the 7nm PTM technology

simulation framework. The ITD-aware modified versions of these techniques are also studied to evaluate the performance enhancement from the frequency headroom offered at elevated temperatures in the presence of the ITD.

For this purpose, we study various processor operating voltage points for the 7nm PTM technology. In order to make a fair comparison, the power and frequencies for different operating voltages are derived based on the values form Table 1. E.g., for the 7nm technology, if the maximum frequency is 3.565 GHz at the nominal voltage at temp=50°C, the maximum frequency at 85% of nominal voltage is 2.5GHz and 2.722GHz at temperatures of 50°C and 75°C, respectively. Moreover, at 75% of the nominal voltage the maximum frequency is 1.696GHz and 1.951GHz at temperatures of 50°C and 75°C, respectively.

For temperature study, we simulate 400ms of transient thermal behavior. Every time-interval, (for this study, 1m and 0.1ms) the transient temperature is checked and if it reaches to 90°C, a DTM mechanism is activated. This include throttling, AM, or DVFS. For the AM, the cost of each migration is set one thousand cycles according to [15], which is the cost to bring up a completely power-gated core. For the DVFS, two levels of voltage frequency pairs are used for each operating point, the higher being the operating point (either nominal or 85% of the nominal voltage) and the lower being 75% of the nominal voltage, which is used when the core reaches critical temperatures.

As described in Section 2.2, in order to see the ITD phenomenon, we have to operate the core at lower than nominal voltages. Thus, for the ITD-aware schemes the operating point of 85% and 75% of the nominal voltage are investigated. For $v = 0.85v_{nom}$, the frequency is increased from 2.50GHz to 2.722GHz and for $v = 0.75v_{nom}$, the frequency is increased from 1.696GHz to 1.951GHz for the frequency-sensitive application when it reaches a temperature of 75°C.

The simulations are done for workloads combining hot and cold applications (e.g., bisort_med and lbm_06). Based on the simulation results in Section 3, these two applications showed high contrast in their thermal profile. Moreover, one of them is frequency-sensitive and the other is frequency-insensitive. While other cold-hot applications can be paired with the same simulation framework, pairing hot-hot application and cold-cold applications will yield no performance gain, as there is no thermal difference to exploit ITD.

6. SIMULATION RESULTS

Fig. 5(a) shows the overall performance of the frequency-sensitive application (i.e. bisort_med) for different thermal control techniques and various operating voltages for the 7nm PTM technology. The values in the parenthesis show the ratio of the operating voltage to the nominal voltage.

Table 3: Simulation Results

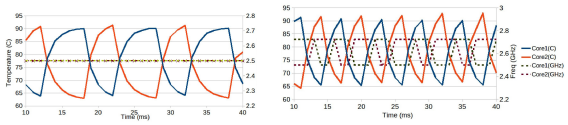
Scheme	$\frac{op.voltage}{nom.voltage}$	IPS (MIPS)	power (w)	EDP ($10^{13}ws^2$)	ED2P ($10^{19}ws^3$)	max. temp. ($°C$)
Throttle	1	1.898	7.59	21.10	11.11	90.6
Throttle	0.85	2.182	6.40	13.5	6.17	91.16
ITD throttle	0.85	2.277	6.73	13.00	5.71	92.32
AM	1	2.934	11.93	13.86	4.72	90.59
AM	0.85	2.644	8.50	12.16	4.60	91.10
ITD AM	0.85	2.765	8.98	11.75	4.25	92.86
DVFS	(1,0.75)	2.627	6.43	9.32	3.55	90.07
DVFS	(0.85,0.75)	2.749	5.37	7.10	2.58	90.13
ITD DVFS	(1,0.75)	2.735	6.68	8.93	3.27	90.07
ITD DVFS	(0.85,0.75)	2.984	6.11	6.86	2.30	90.41

As mentioned earlier, with the operating point of $v = v_{nom}$, the change in the operating frequency at different temperatures is negligible, thus for this operating point, dynamic frequency changing is not useful and only the results for throttling and activity migration are reported. For $v = 0.85v_{nom}$, the results for all techniques in Section 4 are reported. For $v = 0.75v_{noml}$, the cores will never reach critical temperatures, thus no thermal control technique is used. Fig. 5(a) shows that the ITD-aware techniques yield better performance compared to their corresponding DTM technique.

Table 3 shows the summary of the results for various DTM techniques at different operating points. In Table 3, EDP stands for the energy delay product. The cost of migration has been included in the results reported in the table. The number of task migrations during a 400ms simulation in the AM technique was 126 and 550 at $v = 0.85v_{nom}$ and $v = v_{nom}$, respectively, while it was 126 for the ITD-aware technique at $v = 0.85v_{nom}$.

Fig. 5 illustrates the power and EDP comparison between the DTM techniques. According to Table 3 and Fig. 5, using the ITD-aware scheme will enhance the performance of all DTM techniques. For throttling, AM and DVFS, the ITD-aware scheme increases the IPS by 4.39%, 4.59% and 8.55%, respectively at $v = 0.85v_{nom}$. At the operating voltage of $v = v_{nom}$ the ITD-aware scheme is used when DVFS technique in working at a lower voltage level (i.e. $v = 0.75v_{nom}$) and it enhances the performance of DVFS technique by 3.9%. Among the DTM techniques studied in this paper, AM and DVFS show better performance in comparison to the throttling.

While the AM at the nominal voltage yields the best performance, its behavior in other design metrics including power and EDP is not desirable. Moving to a lower operating voltage point would result in better power profile at the cost of decreased performance. The ITD-aware algorithms allow us to compensate for this reduced performance by exploiting the frequency headroom at high temperatures. The DVFS algorithm shows a behavior close to the AM algorithm in terms of IPS when equipped with ITD-aware scheme; however, its lower power and EDP makes it a better candidate for low-power designs.



(a) (b)

Figure 6: Transient temperature for (a) AM (b) ITD-aware AM at $v = 0.85v_{nom}$

Fig. 6 shows the thermal behavior of the AM and ITD-aware AM techniques for $v = 0.85v_{nom}$. The dashed lines show the changes in the operating frequency. In the conventional AM algorithm in Fig. 6(a), the frequency is kept unchanged; however, in the ITD-aware algorithm, the frequency increases whenever the temperature of the core increases above 75 °C. Moreover, these figures show that increasing the frequency of the core at the mid-point temperature threshold, will result in the core temperature reaching the critical temperature faster. Thus, in case of the ITD-aware scheme, the number and thus the cost of migration increases.

7. RELATED WORK

Prior works have focused on the implication of the ITD on circuit design and how it changes the device behavior and design corners. In [2], the effect of ITD is investigated for the behavior of the clock tree mapped on an industrial 65nm CMOS technology, and ITD is shown to occur at low operating voltages. In [3], a dual-v_t circuit is proposed for a commercial low-power 65 nm CMOS technologies under ITD to ensure the temperature-insensitive behavior of the circuit. In [4], the authors use the available power headroom at low temperatures to increase the voltage when the temperature is bellow a certain level to enhance the performance. Latest work have explored the effect of ITD in current technology nodes [2], [3] and [4]; however, it is important to explore ITD impact in the future nodes, as well. This work investigates ITD in future CMOS technologies to understand whether further scaling will increase the significance of the ITD.

On dynamic thermal management, processor thermal characteristics at the architectural level have been studied extensively in recent years [7]. Several techniques have been proposed to reduce the chip temperature in single core [7] as well as multicore architectures [13]. These techniques either migrate the processor activity [12] or adapt processor resources to reduce temperature [7]. In the latter, the utilization across processor units are balanced to control the power density. DVFS has also shown to be effective in balancing the temperature of processor [13, 16].

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

8. CONCLUSION

This paper exploited ITD for power, performance, and temperature optimization in processor architecture. It investigated ITD in future CMOS technologies showing further scaling would increase the significance of the ITD. In Future CMOS technologies, while working at the nominal voltage yields the best performance, due to high power densities the power profile and thermal behavior of a chip is not desirable. On the other hand, moving to a lower operating voltage results in better power profile at the cost of lowering the performance. The ITD-aware algorithm proposed in this paper, allows us to compensate for this reduced performance by exploiting the frequency headroom at high temperatures, while maintaining a low power profile. Simulation results were provided for a workload consisting of a hot-cold combination of applications for the proposed ITD-aware schemes.

9. REFERENCES

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

[2] A. Sassone et. al, "Investigating the effects of inverted temperature dependence (ITD) on clock distribution networks," *DATE*, 2012.

[3] A. Calimera et. al, "Temperature-insensitive dual- vth synthesis for nanometer cmos technologies under inverse temperature dependence," *TVLSI*, 2010.

[4] M. Latif and et. al., "Design for cold test elimination - facing the inverse temperature dependence (ITD) challenge," *ISCAS*, 2012.

[5] Y. K. Cao, "Predictive techonology models," Nanoscale Integration and Modeling (NIMO) Group, 2012. [Online]. Available: http://www.ptm.asu.edu

[6] A. Gaisler, "Leon3 processor," Nanoscale Integration and Modeling (NIMO) Group, 2010. [Online]. Available: http://www.gaisler.com/index.php/products/processors/leon3

[7] K. Skadron et. al, "Temperature-aware microarchitecture," *ISCA*, 2003.

[8] D. M. Tullsen, "Simulation and modeling of a simultaneous multithreading processor," *In Proc. of CMG Conference*, 1996.

[9] "Standard performance evaluation corporation," Open Systems Group (OSG), 2013. [Online]. Available: http://www.spec.org

[10] L. Sheng et. al, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," *MICRO-42*, 2009.

[11] Liu Yongpan et. al, "Thermal vs energy optimization for DVFS-enabled processors in embedded systems," *ISQED*, 2007.

[12] V. Hanumaiah et. al, "Performance optimal online DVFS and task migration techniques for thermally constrained multi-core processors," *TCAD*, 2011.

[13] J. Donald and M. Martonosi, "Techniques for multicore thermal management: Classification and new exploration," in *Computer Architecture News*, 2006.

[14] Coskun et. al, "Evaluating the impact of job scheduling and power management on processor lifetime for chip multiprocessors," in *ACM SIGMETRICS Performance Evaluation Review*, 2009.

[15] R. Kumar et. al, "Processor power reduction via single-isa heterogeneous multi-core architectures," *Computer Architecture Letters*, 2003.

[16] C. Zhu and et. al., "Three-dimensional chip-multiprocessor run-time thermal management," *TCAD*, 2008.