STIFF: Thermally Safe Temperature Effect Inversion Aware FinFET based Multi-core

Shounak Chakraborty
Norwegian University of Science and Technology, Trondheim, Norway
shounak.chakraborty@ntnu.no

Vassos Soteriou
Cyprus University of Technology
Limassol, Cyprus
vsoteriou@gmail.com

Magnus Själander
Norwegian University of Science and Technology, Trondheim, Norway
magnus.sjalander@ntnu.no

ABSTRACT

FinFET, a non-planar device, has become the prevalent choice for chip-multiprocessor (CMP) designs due to its lower leakage and improved scalability as compared to planar CMOS devices. FinFETs are fundamentally different from conventional CMOS circuits in terms of circuit-delay vs. temperature, i.e., circuit-delay decreases in FinFET at higher temperature even in the super threshold supply-voltage regime. Such characteristic of FinFET is known as temperature effect inversion (TEI). But, a drastic increase in channel temperature may lead to an increase in leakage consumption and may accelerate the circuit aging process due to the self-heating effect (SHE). This paper introduces *STIFF*, which balances the upsides of TEI against the potential hazardous SHE in a FinFET based CMP. Basically, *STIFF* exploits online performance statistics to determine the thermal intensity of cores and local caches, and scales the supply-voltage prudentially to maintain a stable core-frequency and local-cache performance on-the-fly by exploiting TEI, while reducing the SHE. Our simulation results show that, *STIFF* is able to maintain a stable frequency of 3.7GHz of the cores with a small standard deviation of 0.23, while maintaining a safe temperature during execution, and it outperforms a state-of-the-art DVFS technique for the FinFET based cores. *STIFF* also maintains a stable access time at the local L1 caches, while ensuring thermal safety by introducing a cache access cognizant scaling of the supply voltage of the individual L1 cache-banks without any noticeable performance-loss.

CCS CONCEPTS

• **Computer systems organization → Multicore architectures**; *System on a chip*; **Parallel architectures**; • **Hardware → Temperature simulation and estimation**.

KEYWORDS

FinFET, Thermal Management, TEI, CMP, SHE

ACM Reference Format:
Shounak Chakraborty, Vassos Soteriou, and Magnus Själander. 2022. STIFF: Thermally Safe Temperature Effect Inversion Aware FinFET based Multi-core . In *19th ACM International Conference on Computing Frontiers (CF'22), May 17–19, 2022, Torino, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3528416.3530223

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
CF'22, May 17–19, 2022, Torino, Italy
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9338-6/22/05…$15.00
https://doi.org/10.1145/3528416.3530223

1 INTRODUCTION

FinFET devices have become the prevalent technology to alleviate the short channel effects in sub-20nm CMOS devices [1, 4]. The confined 3D-geometry of FinFET makes them special in terms of their power and performance characteristics [20, 36, 39, 41]. Unlike conventional planar MOSFETs, FinFET experiences a reduction in gate delay when temperature is increased, which is known as temperature effect inversion (TEI) [9, 27, 28]. Unfortunately, higher temperature may lead to an increase in leakage consumption, and can also accelerate the circuit aging process, caused by the self-heating effect (SHE). Actually, heating effects in FinFET are even more serious as compared to planar MOSFETs because of their reduced thermal conductivity of the interlayer dielectric material and buried oxide [22]. The presence of thermal insulators in a FinFET device restricts the heat dissipation, which further complicates thermal management [3]. This poor heat transfer between FinFET channel and chip ambience is even give rise to errors in temperature sensing mechanisms [31].

Prior conventional on-chip thermal management focused on reducing the temperature in MOSFET based chip multiprocessors (CMPs) [11, 26, 42], which is in stark contrast to the FinFET based chips, where higher temperature potentially leads to performance benefits. Recently, researchers explored TEI in FinFET, that significantly operates faster in higher temperature even at the super-threshold voltage region [8–10, 21, 24, 27, 28]. These previous works mostly focused on developing circuit level or device level techniques to exploit TEI of the FinFET [24, 28]. However, the impacts of TEI on the performance of multi-cores were first evaluated by Cai and Marculescu [9]. But, enjoying TEI benefits can be catastrophic for the FinFETs, as higher temperature can permanently damage the device due to the SHE [41]. Recent research attempts tried to reduce SHEs at the circuit level to ensure thermal safety in FinFET based CMPs, as FinFET can potentially generate hotspots (experiencing a temperature of more than 80 °C) due to enormous increase in channel temperature along with poor heat transfer to the ambient [1, 3, 22, 31]. Most of these techniques considered circuit/device level characteristics, either to exploit TEI or to reduce SHE, without accounting the impacts of running applications. As applications' runtime behaviors play the most pivotal role in on-chip power consumption, hence temperature, considering application runtime characteristics at the (micro)architectural level is crucial in balancing TEI and SHE in FinFET based CMPs.

In this work, we investigate the duality of increased on-chip temperature in FinFET based chip multiprocessors and present *STIFF*, an architectural level thermal manager that balances the benefits of TEI and the problems of SHE, by analyzing application

runtime characteristics. Recent research on thermal sensors for FinFET based CMPs, analyzed the inaccuracies in temperature sensing, which might lead either to circuit failure in the case of underestimation or to performance degradation in the case of overestimation of the thermal status [15, 31, 32]. Concomitant to the fact of issues in heat-sensing of the FinFET devices, in addition with TEI and SHE, *STIFF* realizes the individual core's as well as L1-cache's thermal status through performance monitoring counters (PMCs). *STIFF estimates the temperature based on PMC values and prudentially scales the supply voltages of the cores and L1 cache banks during execution to maintain thermal safety at the respective components, while exploiting TEI to guarantee a stable performance.*

In summary, *STIFF*:

- as a first study, comprehensively analyzes the benchmark applications to select the relevant PMCs, that can assist in determining the temperature of the FinFET based cores as well as local caches (Sec. 2);
- establishes a relationship between *(a)* register access count and core temperature, and *(b)* L1 access count and temperature of the individual L1 cache banks (Sec. 2);
- determines thermal status of both cores and L1 cache banks through exploitation of the PMCs, and prudentially applies dynamic voltage scaling (DVS) at the cores and the L1-banks, while considering TEI, so that both performance and thermal safety can be maintained (Sec. 3).

Our simulation based analysis (Sec. 5), consisting of a 16-tiled CMP with each tile consisting of an out-of-order (OoO) core with 64KB 4-way set-associative local data and instruction caches, shows that *STIFF* is able to maintain a stable core frequency of 3.7GHz with a small standard deviation of 0.23, and a stable access time is maintained at the L1 caches, while maintaining thermal safety. The proposed DVS mechanism at the cores also surpasses a prior dynamic voltage and frequency scaling (DVFS) based technique [35]. Our benchmark based analytical study also shows that the combination of core- and cache-based techniques offers an overall average energy delay product (EDP) gain of 31% over the baseline.

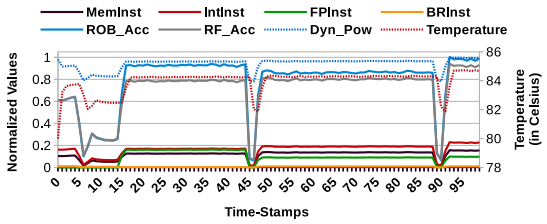
2 BACKGROUND AND PRELIMINARY ANALYSIS

By illustrating our simulation based primary analyses, in this section, we will construct the relevant background for *STIFF*. The relevant empirical studies for background analysis have been performed by simulating a 16 OoO core based homogeneous tiled CMP equipped with 64KB 4-way set-associative local data and instruction caches (at 14nm technology nodes). The CMP configuration and our simulation infrastructure are detailed in Sec. 4.

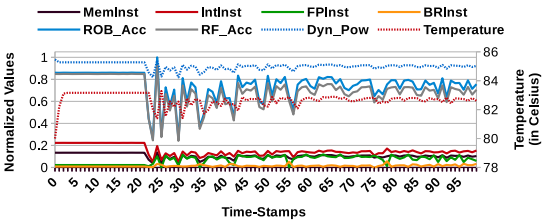
2.1 Analyzing PMCs at the cores

A set of considerable research attempts were taken over the decades that modelled core power by considering PMCs to design several power and thermal management algorithms [13, 18, 19, 40]. In fact, modern processors are equipped with PMCs to count micro-architectural events (e.g., different instruction counts, register file accesses, ROB accesses, etc.), so that processor utilization along with power consumption can be traced online [19]. Here, we first analyze the temporal changes of the following six PMC values: memory instruction count (*MemInst*), integer instruction count (*IntInst*), floating point instruction count (*FPInst*), branch instruction count (*BRInst*), ROB access count (*ROB_Acc*), and register access count (*RF_Acc*). The temporal changes have been traced to a particular core by executing two PARSEC benchmark applications (Bodytrack (*Body*) and Streamcluster (*Stream*)) [5] in our simulation setup (see Sec. 4) for 100M clock cycles[1], and the results are shown in Figure 1. *Dyn_Pow* and *Temperature* in Figure 1 depict how changes in dynamic power consumption and temperature of the core are related to the PMCs.

Out of all these six PMCs, *RF_Acc* and *ROB_Acc* show the most similar behaviors with the temperature changes. In fact, as per prior research attempts [14, 26, 35] register files (RFs) are susceptible to generating on-chip hotspots as they are accessed for every instruction during execution, which makes a register-access-counter useful in determining temperature. Our simulation results not only strengthen the claim of the previous studies, moreover it shows that *ROB_Acc* can also be a useful PMC to determine the core-temperature in case of OoO cores. As RFs and ROBs are accessed during execution of each instruction, either of these two can be used to determine core temperature. In *STIFF*, we use *RF_Acc* values for individual cores to determine the temperature.



(a) *Body*



(b) *Stream*

Figure 1: Temporal changes in PMCs, dynamic power and temperature of core.

Determining temperature by employing PMCs can be a viable option in the case of FinFET based CMPs as the encapsulation of the FinFET channel within thermal insulation might affect the accuracy of thermal sensor [15, 31, 32]. In a recent study, the difficulties related to thermal sensors have been illustrated in the case of FinFET based CMPs [31]. The change in temperature is slower than the change in power consumption caused by dynamic activities, which, especially in the case of FinFET, can negatively impact the accuracy of thermal sensors. Most commercial sensors exploit BJT based sensing mechanism, where temperature induced current or voltage magnitude is sensed to determine the temperature. As FinFET has poor channel to ambient thermal conductivity, such sensors might

[1]Continuous cycles within the region of interest during execution of the applications.

cause inaccurate temperature values, which can differ with up to 5.6 °C lower than the actual temperature for sub-16nm technology nodes [33]. However, placing sensors at the interconnects of the FinFET based circuitry can be a potential solution to improve sensing accuracy [31], but it might incorporate several implementation issues, discussion of which is out of scope of this paper. In *STIFF*, we focus on how temperature can be estimated by employing PMCs in case of FinFET based CMPs.

2.2 Register Accesses vs. Temperature

To establish the relation between register accesses, total power consumption, and core-temperature, we analyzed how changes in *RF_Acc* give rise to changes in the power consumption of the core. Next, we analyze the impact of power consumption on the temperature, as shown in Figure 2. By employing McPAT-monolithic [20], we derived the power consumption, where leakage power has been computed by assuming a fixed temperature of $350K$. We adopt thermal models of FinFET [41] in Hotspot 6.0 [43] to derive temperature for a specific amount of change in power. Thus, we plot the changes in temperature (*Actual_Delta_Temp*) with respect to changes in power consumption. On the other hand, we theoretically analyzed the FinFET power and temperature models [20, 35, 41] to estimate the changes in temperature (*Est_Delta_Temp*) and plot the observation in Figure 2. This temperature estimation is further approximated and the difference between *Actual_Delta_Temp* and *Est_Delta_Temp* are shown in this figure. The graph shows, for most of the cases, that the temperature is overestimated in case of *Est_Delta_Temp* and slightly deviates from the *Actual_Delta_Temp*. Our root-mean-square percentage error (RMSPE) for this estimation is around 3.0%, which is remarkably low. However, by applying linear regression, we derive the equation that shows how the change in register accesses (ΔRF_Acc) will change the power consumption (Δp)-

$$\Delta p = \Delta RF\_Acc \times b_0 + b_1 \tag{1}$$

and, this Δp will change temperature (Δt) as follows:

$$\Delta t = \Delta p \times c_0 + c_1 \tag{2}$$

Both Equation 1 and 2 are used to estimate power and core-temperature for a sufficiently small time-span, which is our sampling interval (1ms). The derived values for the constants, b_0, b_1, c_0 and c_1 are 0.002209, 1.00001437, 0.138906, and $3.1576E-05$, respectively. Note that, the values of these constants will vary with the process node, which we intend to consider in our future work.

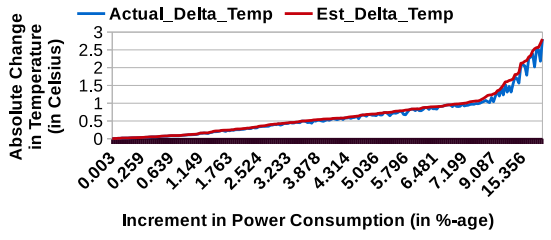


Figure 2: Power consumption of (RFs) vs. temperature

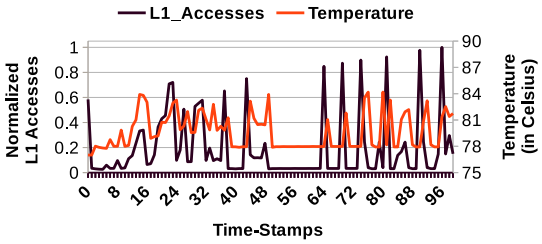


Figure 3: L1 Accesses vs. Temperature.

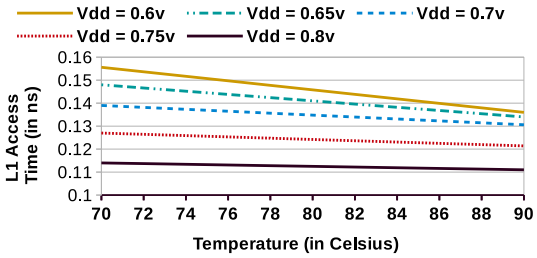


Figure 4: L1-Cache: TEI and Access Time.

2.3 TEI and Core-Frequency

Due to TEI of FinFET at higher temperatures (t), a core can execute instructions with a frequency (f) that can be represented as [9]:

$$f = d_0.V_{dd}^2 + d_1.V_{dd}.t + d_2.t + d_3.V_{dd} + d_4 \tag{3}$$

where, d_0 to d_4 are the coefficients that model the linear part from V_{dd} and t, and the values for these coefficients are obtained from a prior work [9]. From this equation, we can now derive, for a Δt change in temperature (considering a constant V_{dd}) the frequency change Δf will be:

$$\Delta f = (d_1.V_{dd} + d_2).\Delta t \tag{4}$$

Now, by employing Equation 1, 2, and 4, we estimate the temperature and frequency at some certain time-stamp. This estimation can be employed to judiciously scale supply voltage at the cores to ensure thermal safety, while guaranteeing to maintain a stable as well as high frequency due to the presence of TEI. The reduction in V_{dd} curtails core-temperature due to a significant drop in dynamic power, which has a quadratic dependency on V_{dd} and linear dependency on f (as $Dyn\_Pow \propto V_{dd}^2.f$). Moreover, reduced temperature and lowered V_{dd} reduce the static power, so also the overall power consumption, and hence, the temperature will be reduced [9]. However, by evaluating PMCs, *STIFF* realizes the core-temperature and tactfully scales core-V_{dd}, so that TEI is exploited to maintain a stable frequency while ensuring thermal safety. From above discussion it can be concluded that, thermal management at the FinFET based cores can be implemented by controlling V_{dd} alone through DVS, which will maintain the core-temperature and will ultimately govern the core-frequency, unlike DVFS mechanisms used in conventional MOSFET-based cores. In *STIFF*, we therefore use DVS to govern the core-frequency while maintaining a safe temperature.

2.4 L1 Cache: Accesses, Temperature and TEI

Existing diversity in L1 cache accesses generates diverse power-usage profiles, and so also the temperature, across the execution phases of the applications. Figure 3 shows how access-count during

execution changes temperature of an individual L1-Data cache, for a PARSEC application (*Body*), analyzed with our simulation setup (see Sec. 4). The traces have been periodically collected (at the interval of each 1M clock cycles) for a duration of 100M clock cycles. We have normalized the cache accesses and the respective changes in temperature are plotted in Figure 3. Higher L1 access-count implies an increased dynamic power, which is the principal component of the total power in FinFET based caches [20]. Such increased dynamic power leads to higher temperature of the individual L1 banks. Hence, access-count can be a viable option to determine temperature of the L1-bank during execution.

Figure 3 further shows the existence of hotspots at the L1 caches, where the temperature can be as high as 85 °C, which can be controlled either by employing power gating or by applying DVS at an apt granularity of the cache. As the L1-cache is performance critical, power gating might incur significant performance degradation. Hence, we have decided to apply DVS at the individual L1 banks, and to scale the supply voltage of the banks on-demand to ensure thermal safety. Reducing the supply voltage can cause performance degradation by increasing the access time, however, prudentially applying DVS by considering TEI of the FinFET can reduce such performance drops.

Before devising our algorithm for TEI induced performance aware thermally safe L1-caches, we studied the effects on access-time due to TEI at different viable supply voltage (V_{dd}) regions, and report our observations in Figure 4. For a 64KB 4-way L1 cache built in 14nm FinFET, the range of viable V_{dd} is $0.51 - 1.5$v [38]. A drastic reduction in V_{dd} assists in reducing temperature by reducing dynamic power, but might not increase the access time significantly, thanks to TEI. TEI effects are prominent in case of lower (especially at the sub-threshold region) V_{dd} (< 0.8v). Hence, reducing voltage at the higher operational temperature will have less impact on performance. However, V_{dd} has to be scaled up once the temperature is lower than a preset threshold to maintain performance. In this paper, we maintain the range of V_{dd} within $0.6 - 0.8$v.

3 *STIFF*: PROPOSED MECHANISMS

By applying PMC-induced DVS both at the cores and L1 cache banks, *STIFF* reduces SHEs, while maximizing TEI induced benefits to maintain a stable clock frequency at the cores and to maintain a stable access time at the L1-cache.

Algorithm 1: DVS based Frequency-Governor

```
   Input: PERIOD, Initial V_dd, t, t_Hi, t_Lo, V_Hi, V_Lo
1  while System is running do
2      if PERIOD is completed then
3          Track the RF_Acc at the core and get ΔRF_Acc;
4          Δp = ΔRF_Acc × b_0 + b_1;
5          Δt = Δp × c_0 + c_1;
6          if (t + Δt) ≥ t_Hi then
7              Set V_dd at V_Lo;
8          if (t + Δt) ≤ t_Lo & (V_dd < V_Hi) then
9              Set V_dd at V_Hi;
10         Set t at current temperature (collected from thermal sensors);
```

3.1 Governing core-frequency through DVS

The thermal management of *STIFF* is built on the DVS that determines the thermal intensity by looking at the register usages of individual cores over a stipulated period. Changes in register access count (ΔRF_Acc) help in determining the power usage as well as the current thermal intensity of the core. The determined thermal status is used to maintain a safe core temperature that reduces the SHE, while maintaining performance by not allowing the temperature to drop below a certain value to be benefited by the TEI. *STIFF* periodically checks the register accesses of the individual cores, and determines the changes in power and temperature. Once this determined temperature is more than a certain threshold, *STIFF* reduces the supply voltage to safeguard the underlying circuitry. On the other hand, once the determined temperature is lower than a preset value, the voltage is scaled up further to maintain a higher temperature to be benefited by the TEI effect. Note that, anticipating temperature by employing our analytical model (in Sec. 2) can overestimate the temperature, which might slightly reduce TEI benefits, but ensures thermal safety.

We present the whole process of applying DVS at the individual cores to maintain a stable frequency on-the-fly in Algorithm 1. *STIFF* divides the execution span evenly into multiple segments, each of which is called as a *PERIOD* in this paper. *PERIOD* is taken as input to our frequency-governor algorithm, that is the time-span, at the end of which the algorithm will trace the periodic change in register accesses (ΔRF_Acc) (line 3) at the individual cores. Once ΔRF_Acc is traced, the changes in both power usage (Δp) and temperature (Δt) (line 4 and 5) are determined. To determine the thermal intensity of the respective cores, our frequency-governor algorithm employs Equation 1 and 2. The determined change in temperature (at line 5) is added with the temperature of the core at the beginning of the period to predict the temperature at the end of the period. If the calculated temperature is higher than a preset threshold, t_{Hi}, then the frequency is maintained through TEI at this higher temperature in spite of scaling down the voltage, V_{Lo}. This lower supply voltage reduces the core temperature (line 6 to 7) without impacting the performance. On the other hand, if the temperature is below the lower threshold (t_{Lo}) and the supply voltage is less than its higher threshold (V_{Hi}), then there is still room for the TEI gains. Hence, the algorithm sets the supply voltage at V_{Hi} (line 8 and 9) such that the operating core-frequency can be maintained at the lower temperature (see Equation 3). To limit the deviation in temperature estimation, the current temperature from thermal sensor(s) are tracked [31] and used at the end of the next *PERIOD* (line 10). The values of the thresholds (t_{Hi}, t_{Lo}, V_{Hi}, and V_{Lo}) depend on system parameters and the average expected workload of the system (see Sec. 4).

3.2 L1-Cache: TEI-aware Thermal Management

Our previous discussion in Sec. 2.4 shows the salient presence of hotspots at the FinFET based L1 caches, and cache temperature directly depends upon the accesses and on the supply voltage. At the end of each *PERIOD*, *STIFF* collects the access counts for the last *PERIOD* and compares them against the previously collected counts. The access count for an L1 bank at the end of *PERIOD* is denoted as $Acc(PERIOD)$, and the change in access count (D) is calculated as $D = \frac{Acc(PERIOD)}{Acc(PERIOD-1)}$. To control the voltage, *STIFF* uses D to assist in switching the voltage based upon two thresholds: x and y ($x > y$). Our DVS policy makes L1 caches to be operated in three different voltage regions: V_H, V_M and V_L. Initially, the cache

will start operating at V_H, and based upon the changes in D, the voltage for an L1 cache bank is changed.

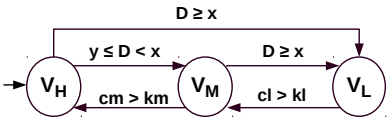


Figure 5: Thermal Management at L1-Cache through DVS.

The events that trigger the voltage to switch are illustrated in Figure 5. If the voltage is set to V_H and $D \geq x$, then the voltage will be scaled down to V_L. However, a comparatively smaller increment in the access-count (i.e., $y \leq D < x$) will scale down the voltage from V_H to V_M. At V_M, if $D \geq x$, then the voltage is scaled down to V_L. The increased access-count implies a temperature increment during the last *PERIOD*, hence, by operating the L1 bank at a lower voltage will reduce the power consumption, and similarly the temperature. If the operating temperature is high at the end of a certain *PERIOD*, reduction in voltage for the immediate next *PERIOD* does not increase the access time, due to the TEI property of the FinFET. While operating at the lower voltages (V_M and V_L), *STIFF* observes a stipulated number of *PERIOD*s (km at V_M and kl at V_L) before scaling up the voltage to the next higher levels. *STIFF* tracks such *PERIOD*-counts by implementing a few number of counters (cm and cl) for two different voltage levels. This provides the L1 bank with a sufficient time for temperature reduction and helps in reducing the SHE.

3.3 Efficient Voltage Switching

To enable the per core DVS in a faster manner, we employ on-chip voltage regulators (VRs), which have significantly smaller timing overheads than their off-chip counterparts [2, 7, 16, 25, 29]. Details about the voltage switching speed, and the type of VR adopted are discussed in Sec. 4. We also analyze the power consumption of the on-chip VRs, as they might pose their own challenges regarding power consumption and hotspots, which can be addressed by techniques like ThermoGater [23]. To implement Algorithm 1, we employ a monitor for scaling the supply voltage at the cores. This monitor triggers DVS once the calculated temperature ($t + \Delta t$) crosses either of the thresholds and there exists room for regulating the supply voltage. The hardware cost for implementing such a monitor is very limited and is also trivial to implement [19, 37]. Additionally, for implementing DVS at the L1 banks, the hardware techniques proposed by Flautner et al. can be adopted [17].

4 METHODOLOGY

We simulate a homogeneous tiled CMP having 16 Alpha 21364 OoO cores (shown in Figure 6), in gem5 [6] full system simulator. Along with a core, each tile contains a data and an instruction private L1 cache and an L2 cache bank. The L2 cache is logically shared, yet physically distributed into 16 banks of the same size. A 2D-mesh-NoC connects the tiles, hence, each tile is equipped with a router. For complete performance-power-thermal analysis, the periodic performance traces of the multithreaded PARSEC benchmarks (16 threads with large input set) [5] (collected from gem5) are fed to McPAT-monolithic [20] for simulating power. The power traces are further sent to HotSpot 6.0 [43] for generating thermal traces. For improved accuracy in generating thermal traces, we adopt thermal properties of FinFET [9] in HotSpot 6.0. Note that, with an interval of 1ms the periodic *performance traces* are collected from gem5. Although TEI effect in FinFET makes frequency no longer fixed at different temperature but for our simulation, we assume a fixed temperature during the whole span of a *PERIOD*, i.e., 1ms [9]. The default parameters used in our simulations are listed in Table 1.

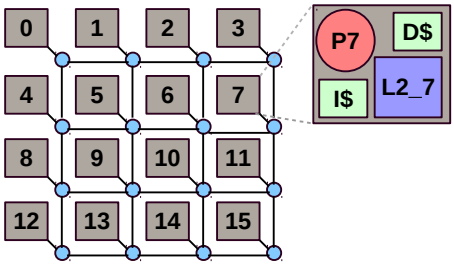


Figure 6: Tiled CMP Architecture.

Table 1: System Parameters

Parameters	Values
Number of Cores	16
Core Model	Alpha 21364
Nominal Frequency	3.5GHz
V_{Hi}, V_{Lo} (at cores)	0.8v, 0.65v
L1 D/I Cache	Private 64KB, 4W SA, LRU
V_H, V_M, V_L (at L1-caches)	0.8v, 0.7v, 0.6v
Shared L2 Cache bank (16 banks)	512KB, 16W SA, LRU
DRAM	8GB
Ambient Temperature	47 °C
Technology Node	14nm (FinFET)

By considering a prior TEI induced frequency model [9], we set the threshold values used in Algorithm 1 as follows: $t_{Hi} = 80\,^{\circ}C$, $t_{Lo} = 77\,^{\circ}C$, $V_{Hi} = 0.8v$ and $V_{Lo} = 0.65v$. Note that, we have assumed a maximum safe temperature of 82 °C, hence, we set $t_{Hi} = 80\,^{\circ}C$ to restrict the thermal overshoot beyond 82 °C. Khan et al. have shown that the operating temperature of FinFET can reach as high as 80–85 °C, which can be considered as a hotspot [22][2]. To maintain an average frequency of 3.7GHz during execution, we set these values, so that the lowest and the highest frequencies can be maintained at 3.0GHz (for t_{Lo}, V_{Lo}) and 3.9GHz (for t_{Hi}, V_{Hi}), respectively. Note that, all of these threshold values are tunable and can be set by considering technical details of the underlying circuitry. However, our employed on-chip VR assumed to be installed at the cores as well as at the L1-caches has a switching speed of 20 mV/ns [16], and the respective area and power overheads are based on a prior regulator-power model [25]. In our evaluation, we use three operational voltage levels for the L1 caches, as mentioned in Table 1 [38].

We have evaluated the following core-based techniques:

- **Baseline** – the default model with system parameters according to Table 1;
- **STIFF** – implementation of the techniques as described in Sec. 3.1 (Algorithm 1);

[2]By considering technology node and process variation, both temperature values for determining hotspots and threshold temperatures can be adjusted, which we intend to analyze in our future work.

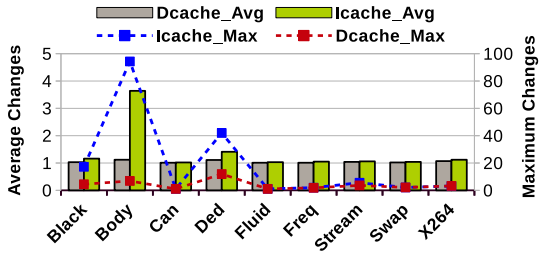


Figure 7: Average and maximum changes in L1 accesses between two consecutive *PERIOD*s. Values for *Max* and *Avg* should be measured with the scales *Maximum Changes* and *Average Changes*, respectively.

- **ENPASS** – a state-of-the-art DVFS technique for the FinFET based cores proposed by Neshatpour et al. [35].

Before applying DVS at the individual L1 cache banks, we first determine the following parameters: x, y, km, and kl, mentioned in Figure 5. Figure 7 reports the maximum and average changes (in terms of times(×)) in L1 accesses between two consecutive *PERIOD*s having a length of 1M cycles. For both data and instruction caches, this average change has a value of less than 2, i.e., the access-count doubles, and has a maximum value of 94.31 (for ICache (*Body*)). In most of the cases, the maximum value is much higher than the average, which indicates that the majority of the times the changes take place in proximity to the average. For a wider design-space exploration, we perform a sensitivity analysis of our cache based mechanism (Sec. 3.2) by considering the following values for x and y: **A**(1.0, 2), **B**(1.25, 4), **C**(1.5, 6), **D**(1.75, 8), and **E**(2, 10). Hypothetically, smaller values/ranges indicate a more frequent change in voltage, while larger ones will incur comparatively fewer occurrences of voltage switching. For all of our simulations, we set $km = 2$, and $kl = 1$, i.e., V_M and V_L can be the viable voltage levels for 2 consecutive *PERIOD*s and 1 *PERIOD*, respectively.

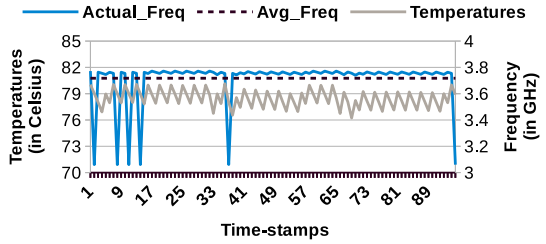
5 RESULTS & ANALYSIS

Now, we will empirically validate the efficacy of our *frequency governor algorithm* at the cores and *TEI-aware thermal management* at the L1-caches.

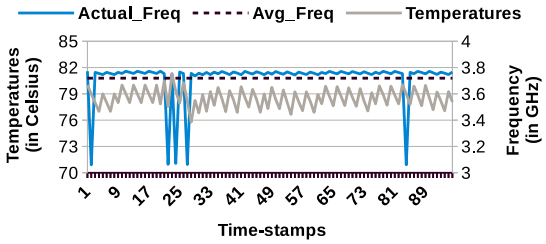
5.1 Effects on Core Frequency & Temperature

Algorithm 1 attempts to exploit TEI to maintain a stable core-frequency by prudentially applying a PMC aware DVS while maintaining a safe core temperature. By tracking the register-access-counter during execution at the individual cores, power and temperature are surmised, and DVS is accordingly applied at the cores. For nine PARSEC applications, *STIFF* is able to maintain a core temperature approximately between our stipulated thermal limits (t_{Lo} and t_{Hi}). The overshooting of core temperature is tackled by reducing the voltage, whereas performance is further maintained by scaling up the voltage in addition with TEI, once the core temperature reaches below t_{Lo}.

We report the temporal changes in core temperature and frequency (for *Body* and *Stream*) in Figure 8. The results are shown for 100 time-stamps during execution, where sampling interval is 1ms (for which, we assumed an unchanged thermal status [9]). For both



(a) *Body*



(b) *Stream*

Figure 8: Temporal change in frequency and temperature.

Body and *Stream*, *STIFF* is able to maintain core temperature between 77 °C and 80 °C (see Figure 8a and 8b). The PMC driven DVS incurs more changes in voltage supply at the initial phases in case of *Body*, whereas *Stream* has more stable register accesses over time resulting into lesser changes in core frequency (see Figure 8). Both *Body* and *Stream* also experience undershoot in frequency at some places, however, for both applications *STIFF* is able to maintain a stable frequency. We further summarize the average core frequency and peak temperature for a particular core for all nine PARSEC applications in Table 2. For all of these nine applications, *STIFF* maintains an average frequency of 3.71GHz, whereas the range for the same is in 3.65 – 3.74GHz. To notice the deviation in frequencies at the individual cores, we further derive the standard deviation for each of the applications. The average standard deviation is 0.23 with a range of 0.16 – 0.33. Higher standard deviation values (e.g. *Fluid*, *Stream*, etc.) are basically caused by more diverse register access pattern of the application on-the-fly whereas lower values (e.g. *Can*, *Swap*, etc.) indicate comparatively better stability in register accesses. We also plot overall reduction in peak temperature of the CMP for all nine applications in Figure 9, which shows *STIFF* reduces peak temperature by 2.3–3.8 °C over the baseline.

Table 2: Average frequency and peak temperature of a core.

Applications	Black	Body	Can	Ded	Fluid	Freq	Stream	Swap	X264	Avg.
Freq. (GHz)	3.73	3.71	3.72	3.74	3.65	3.68	3.72	3.73	3.72	3.71
Std. Dev.	0.21	0.24	0.16	0.19	0.32	0.29	0.33	0.18	0.22	0.23
Temp. (°C)	79.23	78.59	78.78	79.27	78.98	78.51	78.52	79.94	79.82	79.07

5.1.1 ***Comparison with the State-of-the-art.*** We further compared core-based technique of *STIFF* with *ENPASS* [35], that enhanced energy efficiency in a FinFET based CMP with an objective to maintain thermal safety while leveraging through TEI. *ENPASS* integrates pipeline throttling, DVFS and activity migration together. The activity migration can only take place if spare core(s) are available, contrary to *STIFF*, where we assume all cores are always active. Additionally, pipeline throttling leads to significant aggravation in

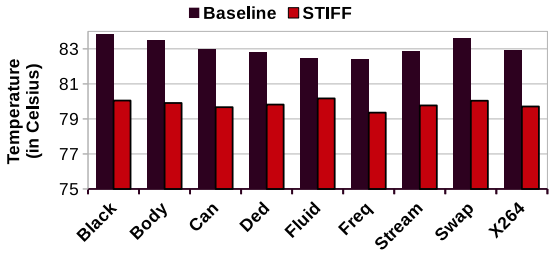


Figure 9: Reduction in peak temperature of the CMP.

performance. Hence, we restrict ourselves to compare *STIFF* only with the DVFS technique of the *ENPASS*. For *ENPASS*, we considered V/F pairs at 4 different threshold temperature values listed in Table 3, where the temperature of the cores can be in the range of 65–80 °C with a couple of midpoint values. Each time the core temperature reaches a midpoint, the V/F is set accordingly so that thermal safety can be maintained, while exploiting TEI to enhance performance. *ENPASS* reduces V/F to the lowest possible level (0.65v, 2.94GHz) as soon as the temperature reaches the highest allowable value (80 °C in this case), and increases the V/F to the next possible level once it reaches a temperature level below 80 °C at the scale of 5. *ENPASS* does not switch V/F in between two levels of temperature (mentioned in Table 3).

Table 3: V/F pairs for *ENPASS* [35].

Voltage	Temperature (in °C)			
	65	70	75	80
	Frequency (in GHz)			
0.65	2.94	2.98	3.02	3.06
0.7	3.19	3.23	3.27	3.32
0.75	3.43	3.47	3.51	3.55
0.8	3.64	3.68	3.73	3.77

STIFF uses only PMC induced DVS technique that shows a better proactive thermal management, and is useful in maintaining a stable core frequency as well as cache performance. As thermal management in *ENPASS* reacts upon detection of temperature, and both voltage and frequency will be set to the lowest possible value, the fluctuation in frequency is significantly higher than *STIFF*, might accelerate the aging process. Figure 10 shows the maximum and average frequencies maintained by both *STIFF* and *ENPASS*. For all applications, *STIFF* maintains a higher average frequency than *ENPASS*, as cores experience a sudden drop in V/F at the critical temperature, while applying *ENPASS*. In fact, sudden voltage drop and a large range of allowable core temperature lead to a more fluctuation in core-frequency and shows higher standard deviation for *ENPASS* than *STIFF*. On average, the standard deviation in case of *ENPASS* for the nine PARSEC applications is 0.31, which is 0.23 in the case of *STIFF*. This is because *STIFF* regulates only voltage but not the frequency like *ENPASS*, and attempts to maintain a stable safe temperature, which results in a sustainable performance.

5.2 L1-Cache: Impacts on Temperature & Access-Time

We summarize the deviation in access time during execution with and without *STIFF*, where the proposed policy is evaluated for five different pairs of values of x and y (**A** to **E**). Figure 11 and 12

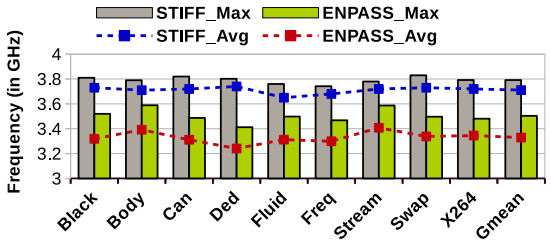


Figure 10: Max. and Avg. frequency: *STIFF* vs. *ENPASS*.

exhibit the average access time and thermal efficiencies of the L1 caches, respectively, gained by *STIFF* for **A** to **E** over the baseline. According to Figure 11, the access time for the L1 cache for **A** to **D** has slightly been increased than baseline, as *STIFF* maintains a lower temperature of the L1 cache (see Figure 12), hence the lower TEI, in addition with lower supply voltage for a number of periods. However, with more stringent values of x and y in **E**, *STIFF* is able to maintain lower temperature for more number of periods than **A** to **D** (average temperature values in Figure 12 indicate the same), resulting in a marginally increased access time due to lower temperature (so the lower TEI), however this leads to better thermal efficiency. Out of all these five settings (i.e. **A** to **E**), ***D** can be considered as a promising choice, that offers a comparatively better thermal efficiency without much degradation in access time.*

5.3 Performance and EDP

Finally, we analyzed the impact of *STIFF* on performance while applying PMC induced DVS at the cores and L1-caches. The core-frequency is maintained by DVS in *STIFF*, whereas no such governor is active in case of baseline. Although we have considered a nominal frequency of 3.5GHz for our baseline system, however, the change in temperature changes the frequency in baseline. The implementation of *STIFF* is although able to maintain thermal safety at the cores, but, at the cost of slight (average) reduction in core frequency. The similar effects are also observed for L1-caches, and for our final comparison, we have considered **D** as the threshold values for x and y. To summarize, we evaluate overall performance loss by *STIFF*, in terms of instructions per second (IPS), that includes both core and L1-cache performance. On an average, *STIFF* experiences a reduction in IPS of 3.2% over the baseline for nine PARSEC applications. Reduction in supply voltages at L1 caches and cores also reduces the overall power consumption, that leads to lower EDP. We report the EDP gains (that include both cores and L1 caches) for all nine applications in Figure 13. On average, *STIFF* shows a noticeable EDP gain of 28 – 35% with an average of 31% over baseline.

6 STATE-OF-THE-ART

The advantage of reduced gate delay, in FinFET devices, is achieved if the circuit is operated at higher temperature, however, thermal safety has to be guaranteed to avoid the device from breaking down. DVFS in combination with dynamically power-gated caches are widely used in thermal management of the high-end computer systems [12, 26, 42]. The existing arts are mostly based on conventional MOSFET designs, where reduced temperature benefits both the performance and the energy efficiency of the system. On the contrary, the performance of the FinFET based designs is benefited from an increased temperature, while energy and reliability get

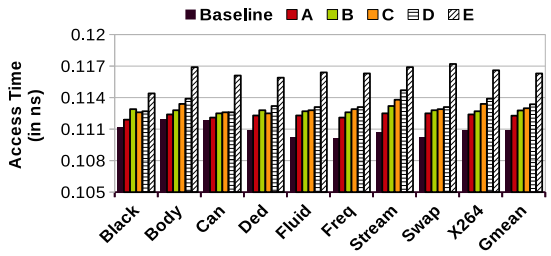


Figure 11: Average access time of L1-cache.

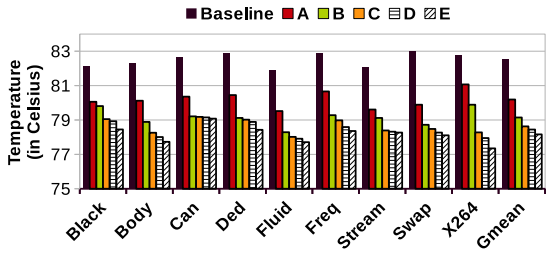


Figure 12: Average temperature of L1-cache.

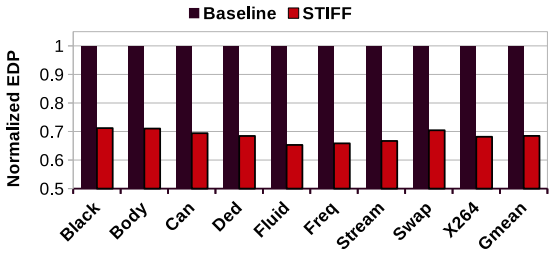


Figure 13: EDP: Baseline vs. *STIFF*.

worse [35]. Over a decade, researchers are investigating TEI in FinFET, that significantly lowers circuit delay in higher temperature even at the super-threshold voltage region [8–10, 21, 24, 27, 28]. Kim et al. have comprehensively analyzed several circuit and device-characteristics for understanding TEI in-depth [24]. Through online scaling of the supply voltage, Lee et al. [28] proposed a thermal management technique for the FinFETs, while exploiting TEI. Lee and Jha have proposed a static power-area-timing model at the micro-architectural level for the FinFET based caches [27]. However, most of these prior techniques focused on the TEI effects, but, its impacts on the performance of multi-cores were first evaluated by Cai and Marculescu [9]. Later, a TEI-aware DVFS was proposed [35], where voltage/frequency (V/F) switching mechanism exploits on-chip thermal sensors, where poor heat transfer at the FinFET devices might lead to inaccurate temperature sensing [31, 32].

The current generation of FinFETs are suffering from electro-thermal issues, known as SHE, which has become a serious design concern for the FinFET based CMPs, especially those built in sub-14nm technology [41]. Recently, researchers tried to reduce SHEs to ensure thermal safety in FinFET based CMPs [1, 3, 22, 30, 34]. Authors in [22] discussed SHE and reliability issues, where FinFET can potentially experience a temperature more than 80 °C due to increase in gate and drain temperature. The underlying causes of such thermal issues have further been studied [34], where authors comprehensively analyzed the dielectric breakdown of the FinFET by considering the physical characteristics. The confined geometry of FinFET devices is one of the prime factors of SHE, hence, it has to be modelled with due consideration to the non-planar geometric shape in addition with the power consumption to slow down the aging process [1, 30]. In another work, a circuit level analytical solution was proposed [41] to improve the tolerance against within chip ambient temperature induced SHE for sub-14nm technology nodes. SHE induced design challenges however lead researchers towards designing novel SRAM cells and processor-cores, so that circuit aging can be prolonged with enhanced performance. Towards that, researchers have recently realized the interaction between SHE, IR-drops and aging for a FinFET based SRAM [3].

Most of these prior arts focused at the circuit/device level, either to exploit TEI benefits or to tackle SHEs, without concentrating on application characteristics, the prime factor that impacts on-chip power-thermal status. In this work, *STIFF*, as a first study, attempts to exploit TEI benefits while tackling SHEs with a PMC based temperature estimation by accounting dynamic behaviors of the applications.

7 CONCLUSIONS AND FUTURE WORK

With the continuation of the scaling of process technology nodes, FinFET, a non-planar device, has become a prevalent choice for CMP designers, due to its lower leakage and reduced delay in higher temperature, even in the super-threshold voltage region, called as TEI [10]. However, a noticeable increase in channel temperature of the FinFET may drastically increase the leakage power consumption, which can escalate the circuit aging process due to SHEs. *STIFF* presents a TEI-conscious performance improvement for FinFET based CMPs, while reducing SHEs by maintaining a safe temperature. *STIFF* is uses register-filer access and L1 access counters to periodically monitor and apply DVS, such that cores and caches are thermally safeguarded. At higher temperature, TEI is exploited, to lower the supply voltage without curtailing the performance noticeably. Our simulation-results exhibit that *STIFF* is able to maintain a safe temperature and an average core frequency of 3.7GHz with a small average standard deviation of 0.23 during execution, and outperforms a prior DVFS technique [35]. For 64KB L1 data and instruction caches, *STIFF* maintains thermal safety without any significant performance loss. To the best of our knowledge, *STIFF is the first technique that employs PMCs to combat SHEs in FinFET based CMPs, while considering TEI to maintain a stable performance at the cores and at the L1-caches.*

In the future, we intend to perform an analytical study by considering process variations of the FinFET and its effects on thermal characteristics and performance of CMPs. Additionally, as thermal conductivity at the channel is a major design concern for the FinFET devices, our future work will also include a detailed comparative study between PMC based and thermal sensor based temperature determination techniques by considering process variation, supply voltage, and temperature based analyses for different technology nodes with various configuration choices.

ACKNOWLEDGMENTS

This work was funded by Marie Curie Individual Fellowship (MSCA-IF), EU (Grant Number 898296).

REFERENCES

[1] W. Ahn, S.H. Shin, C. Jiang, H. Jiang, M.A. Wahab, and M.A. Alam. 2018. Integrated modeling of Self-heating of confined geometry (FinFET, NWFET, and NSHFET) transistors and its implications for the reliability of sub-20nm modern integrated circuits. *Microelectronics Reliability (Elsevier)* (2018).

[2] Mohammad Alian, Ahmed H. M. O. Abulila, Lokesh Jindal, Daehoon Kim, and Nam Sung Kim. 2017. NCAP: Network-Driven, Packet Context-Aware Power Management for Client-Server Architecture. In *HPCA*.

[3] Hussam Amrouch, Victor M. van Santen, Om Prakash, Hammam Kattan, Sami Salamin, Simon Thomann, and Jörg Henkel. 2019. Reliability Challenges with Self-Heating and Aging in FinFET Technology. In *IOLTS*.

[4] A. Bansal, M. Meterelliyoz, S. Singh, Jung Hwan Choi, J. Murthy, and K. Roy. 2006. Compact thermal models for estimation of temperature-dependent power/performance in FinFET technology. In *ASPDAC*.

[5] Christian Bienia, Sanjeev Kumar, Jaswinder Pal Singh, and Kai Li. 2008. The PARSEC Benchmark Suite: Characterization and Architectural Implications. In *PACT*.

[6] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R. Hower, Tushar Krishna, Somayeh Sardashti, Rathijit Sen, Korey Sewell, Muhammad Shoaib, Nilay Vaish, Mark D. Hill, and David A. Wood. 2011. The gem5 Simulator. *ACM SIGARCH CAN* (2011).

[7] Edward A. Burton, Gerhard Schrom, Fabrice Paillet, Jonathan Douglas, William J. Lambert, Kaladhar Radhakrishnan, and Michael J. Hill. 2014. FIVR — Fully integrated voltage regulators on 4th generation Intel® Core™ SoCs. In *APEC*.

[8] E. Cai and D. Marculescu. 2015. TEI-Turbo: temperature effect inversion-aware turbo boost for FinFET-based multi-core systems. In *ICCAD*.

[9] E. Cai and D. Marculescu. 2017. Temperature Effect Inversion-Aware Power-Performance Optimization for FinFET-Based Multicore Systems. *IEEE TCAD* (2017).

[10] Ermao Cai, Dimitrios Stamoulis, and Diana Marculescu. 2016. Exploring aging deceleration in FinFET-based multi-core systems. In *ICCAD*.

[11] S. Chakraborty and H. K Kapoor. 2019. Exploring the role of large centralised caches in thermal efficient chip design. *ACM TODAES* (2019).

[12] Shounak Chakraborty and Magnus Själander. 2021. WaFFLe: Gated Cache-WaYs with Per-Core FIne-Grained DVFS for Reduced On-Chip Temperature and LeAkage Consumption. *ACM Trans. Archit. Code Optim.* 18, 4 (2021).

[13] G. Contreras and M. Martonosi. 2005. Power Prediction for Intel XScale® Processors Using Performance Monitoring Unit Events. In *ISLPED*.

[14] J. Donald and M. Martonosi. 2006. Techniques for Multicore Thermal Management: Classification and New Exploration. In *ISCA*.

[15] M. Eberlein and H. Pretl. 2019. A No-Trim, Scaling-Friendly Thermal Sensor in 16nm FinFET Using Bulk Diodes as Sensing Elements. *IEEE Solid-State Circuits Letters* 2, 9 (2019), 63–66.

[16] S. Eyerman and L. Eeckhout. 2011. Fine-Grained DVFS Using on-Chip Regulators. *ACM TACO* (2011).

[17] K. Flautner, Nam Sung Kim, S. Martin, D. Blaauw, and T. Mudge. 2002. Drowsy caches: simple techniques for reducing leakage power. In *ISCA*.

[18] Bhavishya Goel, Sally A. McKee, Roberto Gioiosa, Karan Singh, Major Bhadauria, and Marco Cesati. 2010. Portable, scalable, per-core power estimation for intelligent resource management. In *ICGC*.

[19] Bhavishya Goel, Sally A. McKee, and Magnus Själander. 2012. Chapter two - Techniques to Measure, Model, and Manage Power.

[20] A. Guler and N. K. Jha. 2020. McPAT-Monolithic: An Area/Power/Timing Architecture Modeling Framework for 3-D Hybrid Monolithic Multicore Systems. *IEEE TVLSI* (2020).

[21] Xuejue Huang, Wen-Chin Lee, C. Kuo, D. Hisamoto, Leland Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Yang-Kyu Choi, K. Asano, V. Subramanian, Tsu-Jae King, J. Bokor, and Chenming Hu. 2001. Sub-50 nm P-channel FinFET. *IEEE T-ED* (2001).

[22] Muhammad Imran Khan, Abdul Rehman Buzdar, and Fujiang Lin. 2014. Self-heating and reliability issues in FinFET and 3D ICs. In *ICSICT*.

[23] S. Karen Khatamifard, Longfei Wang, Weize Yu, Selçuk Köse, and Ulya R. Karpuzcu. 2017. ThermoGater: Thermally-aware on-chip voltage regulation. In *ISCA*.

[24] Sang-Yun Kim, Young Min Kim, Kwang-Ho Baek, Byung-Kil Choi, Kyoung-Rok Han, Ki-Heung Park, and Jong-Ho Lee. 2007. Temperature Dependence of Substrate and Drain–Currents in Bulk FinFETs. *IEEE T-ED* (2007).

[25] Wonyoung Kim, Meeta S. Gupta, Gu-Yeon Wei, and David Brooks. 2008. System level analysis of fast, per-core DVFS using on-chip switching regulators. In *HPCA*.

[26] Joonho Kong, Sung Woo Chung, and Kevin Skadron. 2012. Recent Thermal Management Techniques for Microprocessors. *ACM Comput. Surv.* (2012).

[27] C. Lee and N. K. Jha. 2011. CACTI-FinFET: An integrated delay and power modeling framework for FinFET-based caches under process variations. In *DAC*.

[28] Woojoo Lee, Yanzhi Wang, Tiansong Cui, Shahin Nazarian, and Massoud Pedram. 2014. Dynamic thermal management for FinFET-based circuits exploiting the temperature effect inversion phenomenon. In *ISLPED*.

[29] Yunsup Lee, Andrew Waterman, Henry Cook, Brian Zimmer, Ben Keller, Alberto Puggelli, Jaehwa Kwak, Ruzica Jevtic, Stevo Bailey, Milovan Blagojevic, Pi-Feng Chiu, Rimas Avizienis, Brian Richards, Jonathan Bachrach, David Patterson, Elad Alon, Bora Nikolic, and Krste Asanovic. 2016. An Agile Approach to Building RISC-V Microprocessors. *IEEE Micro* (2016).

[30] Y.-H. Lee, J H. Lee, Y.F. Wang, R. Hsieh, Y.S. Tsai, and K. Huang. 2016. Consideration of BTI variability and product level reliability to expedite advanced FinFET process development. In *IEDM*.

[31] Xin Li, Zhi Li, Wei Zhou, and Zhemin Duan. 2020. Accurate On-Chip Temperature Sensing for Multicore Processors Using Embedded Thermal Sensors. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 28, 11 (2020), 2328–2341.

[32] S. L. Liu, J. J. Horng, Amit kundu, B. S. Lien, C. H. Lee, Y. H. Chen, Chester Kuo, Y. C. Peng, and Mark Chen. 2019. Hot-Spot Thermal Sensor Design in FinFETs. In *2019 IEEE International Electron Devices Meeting (IEDM)*. 5.2.1–5.2.4.

[33] K A A Makinwa. 2018. *Temperature Sensor Performance Survey*. http://ei.ewi.tudelft.nl/docs/TSensor_survey.xls

[34] S. Mei, N. Raghavan, M. Bosman, D. Linten, G. Groeseneken, N. Horiguchi, and K. L. Pey. 2016. New understanding of dielectric breakdown in advanced FinFET devices — physical, electrical, statistical and multiphysics study. In *IEDM*.

[35] Katayoun Neshatpour, Wayne Burleson, Amin Khajeh, and Houman Homayoun. 2018. Enhancing Power, Performance, and Energy Efficiency in Chip Multiprocessors Exploiting Inverse Thermal Dependence. *IEEE TVLSI* (2018).

[36] Alireza Shafaei, Yanzhi Wang, Xue Lin, and Massoud Pedram. 2014. FinCACTI: Architectural Analysis and Modeling of Caches with Deeply-Scaled FinFET Devices. In *ISVLSI*.

[37] Magnus Själander, Margaret Martonosi, and Stefanos Kaxiras. 2014. Power-Efficient Computer Architectures: Recent Advances. *Morgan & Claypool* (2014).

[38] Taejoong Song, Woojin Rim, Jonghoon Jung, Giyong Yang, Jaeho Park, Sunghyun Park, Yongho Kim, Kang-Hyun Baek, Sanghoon Baek, Sang-Kyu Oh, Jinsuk Jung, Sungbong Kim, Gyuhong Kim, Jintae Kim, Youngkeun Lee, Sang-Pil Sim, Jong Shik Yoon, Kyu-Myung Choi, Hyosig Won, and Jaehong Park. 2015. A 14nm FinFET 128 Mb SRAM With V_{MIN} Enhancement Techniques for Low-Power Applications. *IEEE Journal of Solid-State Circuits* (2015).

[39] Aoxiang Tang, Yang Yang, Chun-Yi Lee, and Niraj K. Jha. 2015. McPAT-PVT: Delay and Power Modeling Framework for FinFET Processor Architectures Under PVT Variations. *IEEE TVLSI* (2015).

[40] V. Tiwari, S. Malik, A. Wolfe, and M.T.-C. Lee. 1996. Instruction level power analysis and optimization of software. In *VLSID*.

[41] Sankatali Venkateswarlu, Akhil Sudarsanan, Shiv Govind Singh, and Kaushik Nayak. 2018. Ambient Temperature-Induced Device Self-Heating Effects on Multi-Fin Si n-FinFET Performance. *IEEE T-ED* (2018).

[42] W. Zang and A. Gordon-Ross. 2013. A Survey on Cache Tuning from a Power/Energy Perspective. *ACM Comput. Surv.* (2013).

[43] R. Zhang, Mircea R. Stan, and K. Skadron. 2015. HotSpot 6.0: Validation, Acceleration and Extension.. In *University of Virginia, Tech. Report CS-2015-04*.