NCFET-Aware Voltage Scaling

Sami Salamin*, Martin Rapp*, Hussam Amrouch*, Girish Pahwa‡, Yogesh Chauhan‡, Jörg Henkel*
*Department of Computer Science, Karlsruhe Institute of Technology, Karlsruhe, Germany
‡Electrical Engineering Department, Indian Institute of Technology Kanpur, Kanpur, India
{sami.salamin, martin.rapp, amrouch, henkel}@kit.edu, {girish, chauhan}@iitk.ac.in

***Abstract*—Negative Capacitance Field-Effect Transistor (NCFET) has recently attracted significant attention. In the NCFET technology with a thick ferroelectric layer, voltage reduction increases the leakage power, rather than decreases, due to the negative Drain-Induced Barrier Lowering (DIBL) effect. This work is the first to demonstrate the far-reaching consequences of such an *inverse dependency* w.r.t. the existing power management techniques. Moreover, this work is the first to demonstrate that state-of-the-art Dynamic Voltage Scaling (DVS) techniques are *sub-optimal* for NCFET. Our investigation revealed that the optimal voltage at which the total power is minimized is not necessarily at the point of the minimum voltage required to fulfill the performance constraint (as in traditional DVS). Hence, an NCFET-aware DVS is key for high energy efficiency. In this work, we therefore propose the first NCFET-aware DVS technique that selects the optimal voltage to minimize the power following the dynamics of workloads. Our experimental results of a multi-core system demonstrate that NCFET-aware DVS results in 20% on average, and up to 27% energy saving while still fulfilling the same performance constraint (i.e., no trade-offs) compared to traditional NCFET-unaware DVS techniques.**

I. Introduction

NCFET is an emerging technology that has great potential to replace the CMOS technology in the near future, since it exhibits a considerable improvement in the circuit's performance by overcoming the fundamental limit of sub-threshold swing. The sub-threshold swing of a transistor determines the minimum increase in voltage required to raise the transistor's "on" current by one order of magnitude. Hence, the sub-threshold swing determines how fast the transistor can switch from the "off" to the "on" state. In conventional FET technology, the sub-threshold swing is limited to 60mV/decade at room temperature due to the so-called "Boltzmann tyranny" (i.e., the Boltzmann distribution of charge carriers at the source of the transistor) [1]. To overcome this limitation, NCFET incorporates a ferroelectric layer within the gate stack of the transistor, that behaves as a Negative Capacitance (NC) resulting in a voltage amplification. This allows the sub-threshold swing of the transistor to fall below the 60mV/decade [1], [2]. This, in turn, has two key implications when it comes to high-performance and low-power applications: (1) compared to conventional FET, NCFET-based circuits can be clocked at higher frequencies without the need to increase the voltage. (2) compared to conventional FET, NCFET-based circuits can be clocked at the same frequency but at a much lower voltage [3]. Like any other emerging technology, the compatibility with the existing standard CMOS fabrication process is an indispensable requirement to become a reality. When it comes to NCFET, a breakthrough has been recently achieved when Globalfoundries fabricated NCFET-based circuits using their commercial 14nm FinFET technology [4]. *Therefore, it is now the right time to investigate the implications that NCFET technology has on circuit and system levels.*

978-1-7281-2954-9/19/$31.00

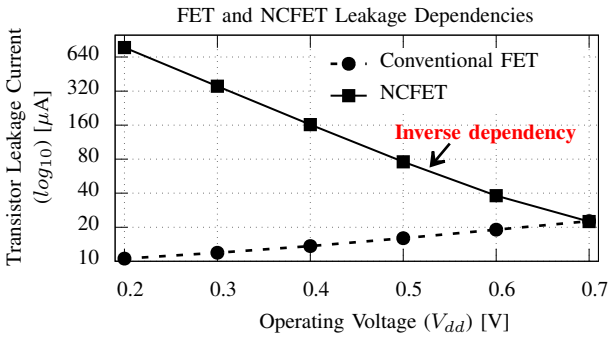

Fig. 1: Leakage current (I_{off}) of a single NCFET transistor in comparison with conventional FET transistor, for 7nm FinFET technology, over a wide range of voltages. NCFET exhibits an *inverse dependency* of leakage current with V_{dd}, unlike conventional FET. This leads to a novel trade-off between leakage and dynamic power as Fig. 2 demonstrates.

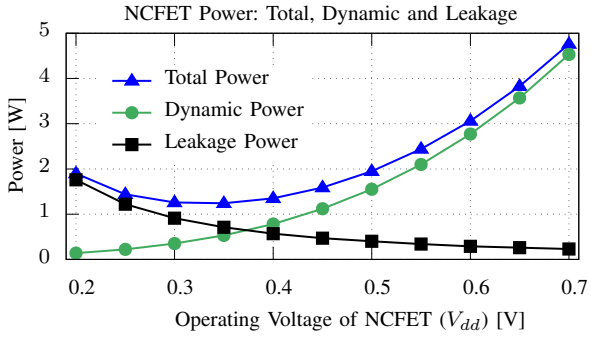

Fig. 2: Total power and its components (i.e., leakage and dynamic) of *canneal* master thread starting from the minimum possible voltage under a performance constraint. In all V_{dd}, the same frequency of 1 GHz is applied. The total power decreases when V_{dd} increases until it reaches an inflection point of *minimum power*. Then it starts to increase dominated by dynamic power. Note that the ability of NCFET to operate at such low V_{dd} (0.2V) is due to the inherent voltage amplification provided by the integrated negative capacitance.

NCFET and Voltage Dependency: In conventional FET, the leakage current (I_{off}) decreases when the voltage V_{dd} decreases. Hence, DVS techniques always aim at operating the processor at the minimum V_{dd} to minimize power. *However, such a well-known voltage dependency becomes inverse with respect to leakage power in NCFET due to the negative DIBL effect*, which is a typical characteristic of short-channel NCFET [5], [6]. Negative DIBL reduces the threshold voltage

(V_{th}) of the transistor and thus increases I_{off} when the voltage decreases. In practice, when V_{dd} is increased in the "off" state, the gate charge reduces due to the electric field from drain to ferroelectric [7]. As the ferroelectric layer in NCFET is biased in a negative capacitance state, a decrease in charge leads an increase in the voltage drop across the ferroelectric layer. Consequently, the voltage reaching the internal transistor gate decreases, resulting in a rise in the energy barrier to the electrons coming from the source. Thus, I_{off} reduces with V_{dd} instead of increasing (as in conventional FET).

To demonstrate that, we performed simulations for the 7nm FinFET technology node for both baseline (i.e., conventional FET without a ferroelectric layer) and NCFET in which a ferroelectric layer with a 4nm thickness is in use. Results are extracted using the BSIM-CMG, which is the industrial standard compact model for FinFET technologies [8]. We have modified the BSIM-CMG to incorporate the state-of-the-art physics-based model of negative capacitance [9]. As demonstrated in Fig. 1, unlike conventional FET, where reducing V_{dd} results in lower I_{off} and thus lower leakage power, reducing V_{dd} in NCFET results in a noticeable increase in I_{off}. Note that this transistor-level analysis is solely used here to demonstrate the leakage dependency. However, power and timing analysis in the rest of this paper are accurately obtained using signoff tools for a full SoC (see Section IV).

To further demonstrate the consequences of such an inverse dependency at the system level, we show in Fig. 2 the total power and its components (leakage and dynamic power) of the core running the master thread in a multi-core (2×2) system when a benchmark *canneal* from the PARSEC benchmark suite [10] is being executed. Detailed explanation of the employed NCFET modeling and our experimental setup will be presented in Section IV. As shown, scaling down the voltage reduces the dynamic power but also increases the leakage power. As a result, the total power consumption reduces until an inflection point after which it starts to increase again. Therefore, an optimal voltage point exists not at minimum possible operating voltage (around 0.35V for this particular scenario). We will demonstrate in Section II that such a novel trade-off in NCFET breaks the concept of determining Voltage-frequency (V-f) pairs at design time that is typically aimed in conventional FET-based processors [11]. In this work, we demonstrate that NCFET requires voltage scaling that is antagonistic to conventional FET voltage scaling.

Hence, DVS techniques for NCFET must be aware of this property, which offers a novel trade-off. We will demonstrate the scenarios, where an NCFET-aware DVS is required to minimize the power. These are when a CPU is not operated at the peak frequency, or when applications with low CPU utilization are executed, such as memory-bound applications or applications with high synchronization overhead between threads. While there is a small amount of work studying the impact of NCFET, w.r.t. performance and power at circuit, single processor and many-core processor [3], [6], *there is no work on NCFET-aware power management i.e., work that takes the bespoken inverse leakage dependency into consideration. Therefore, we present in this paper the first DVS technique of this kind.*

Our novel contributions within this paper are as follows:
(1) We are the first to demonstrate that voltage scaling in NCFET-based processors leads to a novel runtime trade-off between leakage and dynamic power resulting in an optimal point at which total power is minimized.
(2) The inverse voltage dependency in NCFET leads to different optimal voltages for different workloads based on the share of leakage power from total power.
(3) The aforementioned two points emerge the necessity of developing an NCFET-aware DVS for runtime power optimization. In this work, we present the first technique that dynamically selects the optimal voltage in this new scenario.

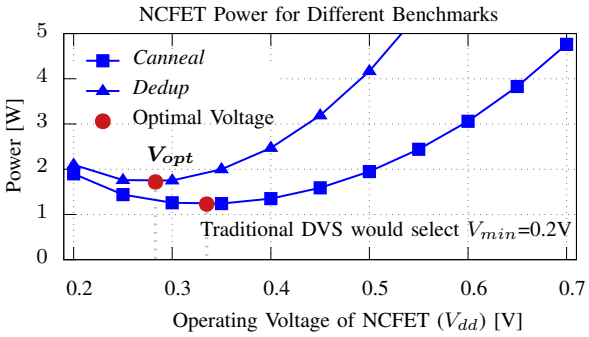

Fig. 3: Total power consumption of different workloads running at the same frequency (1 GHz) over voltage, starting from V_{min} selected by the traditional DVS to achieve the required frequency. Different workloads exhibit different V_{opt} at which the total power is minimized. Such V_{opt} does not exist at V_{min}, which would be the decision of traditional DVS.

II. Workload Dependency

In this section, we show that with NCFET voltage selection should be chosen based on the running workloads. General purpose processors run a variety of applications, which typically exhibit different characteristics at runtime resulting in different runtime power consumption. In the context of this work, the optimal voltage (V_{opt}) is defined as the operating voltage at which the processor's total power is minimized. Total power in this work is the peak total power consumption when a core is active. In traditional DVS, which was designed for conventional FET, the minimum voltage (V_{min}) under a performance constraint always equals V_{opt}. Importantly, the selected V_{min} does not depend on the running workload and hence it can be obtained at design time. Consequently, a set of V-f pairs are typically determined at design time and then be later employed by the DVS technique at runtime. However, in the context of NCFET, V_{opt} does not necessarily occur at V_{min}. In an NCFET-based processor, V_{opt} depends on the share of leakage power from the total power and hence V_{opt} will vary at runtime based on workloads. *Therefore, in NCFET technology, different workloads will exhibit different V_{opt}. Hence, selecting V_{opt} becomes a runtime decision.*

To demonstrate how different workloads may have different V_{opt}, we present in Fig. 3 the total power across a wide range of voltages for two different workloads (obtained from the PARSEC benchmark suite [10]). As can be observed, different workloads exhibit different V_{opt} at which the total power is minimized and such V_{opt} is not equal to the V_{min}.

All in all, NCFET-aware DVS is necessary due to the change in the behavior of total power consumption over voltage scaling which stems from the inverse dependency

w.r.t. leakage power. The new behavior results in a novel trade-off between leakage and dynamic power. Based on the leakage share (which is workload dependent) V_{opt} differs from V_{min}. However, our work exploits the trade-off between the dynamic and leakage power at *runtime*. Therefore, it is fundamentally different from any existing *design-time* trade-offs, e.g., changing the threshold voltage of transistor V_{th} [12].

III. Our NCFET-Aware DVS Technique

In the following, we first present the developed power and performance models for selecting V_{opt} at runtime. Then, we present our novel NCFET-aware DVS technique. Power and performance models are derived based on [6].

A. Design Time: Power and Performance Modeling

The maximum, yet sustainable operating frequency $f_{max}(V)$ depends on the voltage V through the minimum delay $d_{min}(V)$:

$$d_{min}(V) = a_{del} \cdot V^{b_{del}} + c_{del}; \quad f_{max}(V) = \frac{1}{d_{min}(V)} \tag{1}$$

$a_{del}>0$, $b_{del}<0$, $c_{del}\geq 0$ are constants fitting parameters obtained at design time. Operating at the peak frequency with maximum CPU activity results in the following leakage and peak dynamic power consumption:

$$P_{leak}(V) = a_{leak} \cdot V^{b_{leak}} \tag{2}$$

$$P_{dyn}^{peak}(V, d_{min}(V)) = a_{dyn} \cdot V^{b_{dyn}} + c_{dyn} \tag{3}$$

$a_{dyn}>0$, $b_{dyn}>1$, $c_{dyn}\geq 0$, $a_{leak}>0$, $b_{leak}<0$ are constant fitting parameters. Both $P_{dyn}^{peak}(V, d_{min}(V))$ and $P_{leak}(V)$ are *convex* in V. It is possible to operate the CPU at lower frequency (higher delay) than the maximum sustainable frequency (minimum sustainable delay). Since leakage power is independent from CPU activity, it is not affected. Dynamic power decreases proportionally with the increase in the delay.

$$\begin{aligned} P_{dyn}^{peak}(V, d) &= \frac{d_{min}(V)}{d} \cdot P_{dyn}^{peak}(V, d_{min}(V)) \\ &= a_{del}a_{dyn}V^{b_{del}+b_{dyn}} + a_{del}c_{dyn}V^{b_{del}} \\ &\quad + a_{dyn}c_{del}V^{b_{dyn}} + c_{del}c_{dyn} \end{aligned} \tag{4}$$

$P_{dyn}^{peak}(V, d)$ is convex in V (for constant d) if $b_{dyn}+b_{del}>1$.

B. Runtime: Workload-Dependent Power Modeling

The workload (application) executed at runtime affects the dynamic power consumption $P_{dyn}(V, d)$, which is reduced by a factor $0\leq r_{dyn}\leq 1$ from the peak dynamic power $P_{dyn}^{peak}(V, d)$:

$$P_{dyn}(V, d) = r_{dyn} \cdot P_{dyn}^{peak}(V, d) \tag{5}$$

$$P_{total}(V, d) = P_{dyn}(V, d) + P_{leak}(V) \tag{6}$$

r_{dyn} represents the current workload activity and therefore is not constant. When measuring the total power consumption $P_{total}(V_c, d)$ at the current voltage V_c, r_{dyn} can be calculated:

$$\begin{aligned} r_{dyn} &= \frac{P_{dyn}(V_c, d)}{P_{dyn}^{peak}(V_c, d)} \\ &= \frac{P_{total}(V_c, d) - P_{leak}(V_c)}{P_{dyn}^{peak}(V_c, d)} \end{aligned} \tag{7}$$

Algorithm 1 Our NCFET-aware voltage scaling algorithm to select the optimal voltage (V_{opt}) at runtime

Input: Power and performance models: $P_{dyn}^{peak}(c, d)$ and $P_{leak}(V)$, current supply voltage V_c and delay d, current power consumption P_{curr}, min. voltage resolution ϵ
Output: Optimal supply voltage V_{opt}

```
r_dyn ← (P_curr − P_leak(V_c)) / P_dyn^peak(V_c, d)        ▷ Eq. (7)
V_opt ← V_min(d)                                           ▷ Eq. (8)
repeat
    ΔV_opt ← −P_total(V_opt, d)′ / P_total(V_opt, d)″
    V_opt ← V_opt + ΔV_opt                                 ▷ iterative update
    if V_opt < V_min(d) then return V_min(d)               ▷ out of bounds
    if V_opt > V_max then return V_max                     ▷ out of bounds
until ΔV_opt < ε                                           ▷ Termination criteria
return V_opt
```

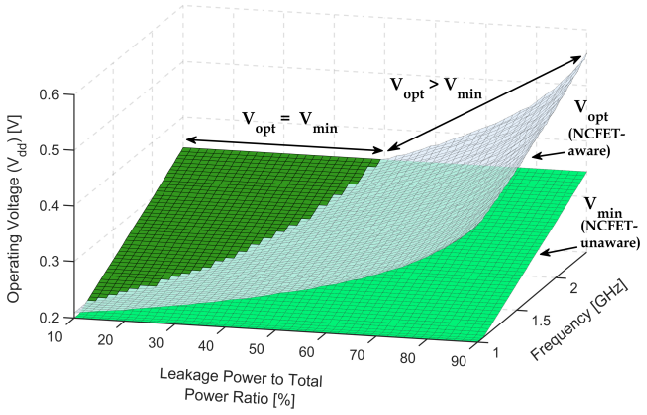

Fig. 4: Comparison of V_{dd} selected by NCFET-unaware (V_{min}) and NCFET-aware (V_{opt}) DVS based on frequency and leakage to total power ratio. NCFET-unaware DVS always selects V_{min} that sustains the required frequency. NCFET-aware DVS selects higher voltages ($V_{opt}>V_{min}$) for low frequencies or high leakage to total power ratio to minimize total power.

C. Runtime: Optimal Voltage Computing

V_{opt} that minimizes the total power, can be then obtained from the power and performance models:

$$V_{min}(d) = \left(\frac{d - c_{del}}{a_{del}}\right)^{\frac{1}{b_{del}}} \tag{8}$$

$$V_{opt}(d, r_{dyn}) = \underset{V_{min}(d)\leq V\leq V_{max}}{\arg\min} P_{total}(V, d) \tag{9}$$

In our implemented algorithm, we exploit that $P_{total}(V, d)$ is convex in V, because it is composed of convex functions. Convexity guarantees that $P_{total}(V, d)$ has exactly one minimum w.r.t. V within the range $[V_{min}(d), V_{max}]$. Therefore, we can use any convex optimization algorithm to efficiently obtain V_{opt}. We choose Newton's method to achieve fast convergence. Algorithm 1 summarizes our implemented DVS technique.

D. NCFET-Aware and NCFET-Unaware DVS

Fig. 4 shows the design space with NCFET-aware (V_{opt}) and NCFET-unaware DVS (V_{min}). NCFET-unaware DVS sets the minimum voltage that is needed to sustain the required frequency and therefore it does not consider workload behavior. Contrarily, NCFET-aware DVS does consider the workload as

it depends on the ratio of leakage to total power measured at V_{min}. The explored design space reveals the following:
(a) Two trends can be observed: (1) the higher the frequency increases, the higher V_{opt} (i.e., the selected voltage) is. This is completely consistent with NCFET-unaware DVS. (2) the higher the leakage to total power ratio increases, the higher V_{opt} is. *This is because leakage power gets prominent and therefore, it should be reduced by selecting a higher V_{dd}.*
(b) Two distinct regions exist: (1) For low leakage to total power ratio and for high frequencies, both techniques (NCFET-aware and traditional NCFET-unaware) select the same voltage (i.e., V_{opt}=V_{min}). (2) For high ratios of leakage to total power or low frequencies, NCFET-aware DVS selects higher voltages than the minimum voltage to minimize the total power (V_{opt}>V_{min}).

As shown in Fig. 4, the shape of the operating voltage for both techniques differ and they have different trends (i.e., not just shifted or scaled). For NCFET-aware, V_{opt} almost always differs than V_{min}. Hence, it is indispensable to develop an NCFET-aware DVS to optimize the total power.

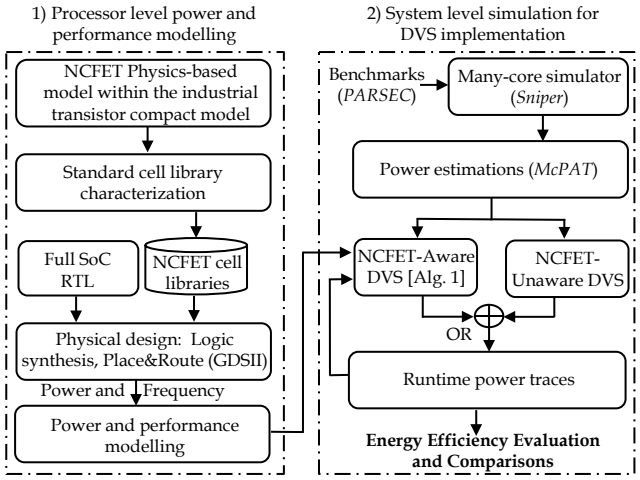

Fig. 5: General overview of our design flow demonstrating the implementation at different abstraction levels to investigate the efficiency of a multi-core system when our NCFET-aware DVS versus traditional (NCFET-unaware) DVS is employed.

IV. EXPERIMENTAL EVALUATION

In the following, we present different evaluation phases to examine the effectiveness of our NCFET-aware DVS. Starting from the experimental setup to show the used tools and models, then we investigate the scenarios when our algorithm is better and lastly, we show the energy gains and energy savings obtained when employing our algorithm.

A. Experimental Setup

As summarized in Fig. 5, our experimental setup consists of two main parts: (1) processor-level power and performance modeling and (2) system-level simulation to evaluate the efficiency of a multi-core system under the effects that traditional (i.e., NCFET-unaware) DVS and our NCFET-aware DVS have.
Processor-Level Power and Performance Modeling: To obtain detailed power and performance modeling of the processor under the effects of NCFET, we implement a full tile of the state-of-the-art *OpenPiton* SoC [13]. *OpenPiton* is an opensource multi-core processor based on the *OpenSPARC* T1 architecture. Using a physics-based NCFET model [9], integrated within the industrial compact model of FinFET (BSIM-CMG) [8], NCFET-aware cell libraries were created [3], [14] based on the open-source 7nm FinFET PDK [15]. Our libraries are fully compatible with existing EDA tool flows (Synopsys and Cadence). This allows us to directly deploy them to perform a full chip design starting from logic synthesis all the way to the GDSII level (i.e., full chip layout design). To accurately estimate how the power and performance of the processor are affected by NCFET for a wide range of voltages, we created our cell libraries for different voltages at the 7nm FinFET technology. Moreover, the room temperature of $25°C$ is assumed in all analysis. Since the optimal voltage V_{opt} depends on the application's characteristics, which might be unknown at prior, we analyze the whole voltage range from 0.2 to 0.7V that is supported by the cell libraries. The thickness of the employed ferroelectric layer is selected to be 4nm because with larger thicknesses, a hysteresis-free operation in NCFET cannot be ensured anymore [3] in the employed 7nm FinFET transistor. To accurately estimate the resulting power and delay, industrial signoff tools (Cadence Voltus [16] and Tempus [17]) are employed. Afterwards and based on the obtained processor analysis (as in [6]), we develop power and performance analytical models as described in Section III. These models are integrated later within a system-level simulator for runtime voltage selection.

System-Level Simulation: We evaluate our proposed NCFET-aware DVS technique on a multi-core (2×2) system. Each of the cores has private L1-I and L1-D caches with 32 KB, each. The per-core private L2 caches have a size of 256 KB, each. The 8 MB L3 cache is shared among all four cores. We use the *HotSniper* tool-chain [18] to simulate our multi-core system. It combines the *Sniper* multi-core simulator [19] with a periodic invocation of *McPAT* [20] for runtime power estimation. We run tasks from the PARSEC benchmark suite [10], which is commonly used to evaluate multi-core system. The benchmarks *facesim* and *raytrace* do not offer a *simsmall* input, the benchmarks *ferret*, *freqmine* and *vips* have unresolvable instrumentation errors. Therefore, we had to skip these five benchmarks in our analysis. Since *McPAT* does not support the NCFET technology (investigated in this work), we estimate power at 45nm using *McPAT* and then scale dynamic and leakage power to 7nm NCFET. We therefore implemented the *OpenPiton* SoC additionally at the 45nm technology node [21]. The frequency-dependent scaling factors are obtained by comparing the dynamic and leakage power consumption of both technology nodes based on our previous work in [6]. The frequencies are set between 1.0 GHz and 2.4 GHz. The maximum frequency limit of 2.4 GHz comes from our scaling-based approach in which we first employ *McPAT* to estimate the power at 45nm. Therefore, we limited our frequency range to up to 2.4 GHz. V_{dd} is set between 0.2 V and 0.7 V. It is important to note, that low V_{dd} (i.e., 0.2V) does not result in sub-threshold computing due to the inherent voltage amplification provided by the negative capacitance.

For fair comparisons, we configured our simulator for both DVS cases to have: the same frequencies, voltage range, and architecture, in addition to running the same benchmarks. Thus, only voltage selection differs based on DVS decision.

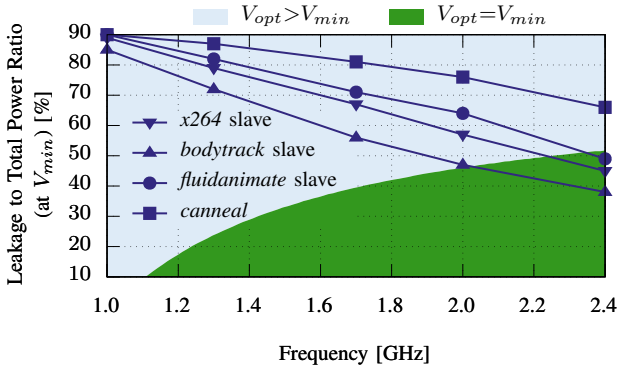

Fig. 6: Traditional DVS only selects the optimal V_{dd} (V_{opt}=V_{min}) for some PARSEC benchmarks when operated at very high frequency. In all other cases total power is minimized at higher operating voltage (V_{opt}>V_{min}). This demonstrates the necessity for NCFET-aware DVS.

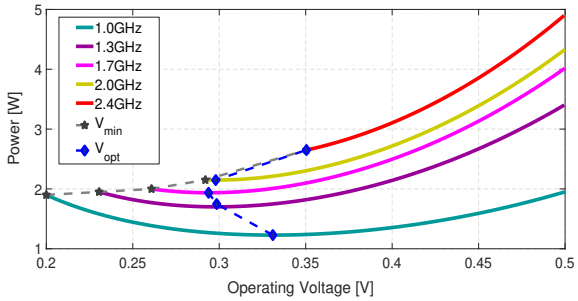

Fig. 7: This figure illustrates the total power consumption over voltage of *fluidanimate* running at different frequencies. This example emphasizes again that power is minimized not at V_{min}, but at a higher voltage.

B. When does NCFET-Aware DVS result in Power Saving?

In the following, we demonstrate the conditions in which our presented NCFET-aware DVS technique results in total power reduction and hence more energy savings compared to the baseline (i.e., traditional DVS). As explained in Section III-D, NCFET-aware DVS selects a higher V_{dd} than traditional DVS for low frequency or high ratio of leakage to total power. This area is highlighted in Fig. 6 (V_{opt}>V_{min}). Fig. 6 also shows the ratio of leakage power to total power for a representative set of PARSEC benchmarks operating at different frequencies. Different workloads exhibit different ratios of leakage to total power. This ratio decreases with increasing frequency because dynamic power consumption increases more strongly than leakage. We do not show all PARSEC benchmarks in this figure to maintain the readability. It can be noticed that for almost all scenarios (i.e., running a workload at a certain frequency), it is required to select a higher operating voltage than V_{min} to minimize the total power. *This experiment demonstrates that NCFET-aware DVS is required not only in some corner-cases, but in almost all execution scenarios of workloads.* In the case where V_{min}=V_{opt}, traditional DVS already selects the optimal operating voltage, which is also selected by our proposed NCFET-aware DVS, i.e., *there are no power losses induces by our proposed NCFET-aware DVS*. Furthermore, Fig. 7 illustrates the impact of frequency on V_{dd} selection and the total power consumption as well for *fluidanimate*. For low frequency V_{opt}>V_{min} and the difference decrease until V_{opt}=V_{min}.

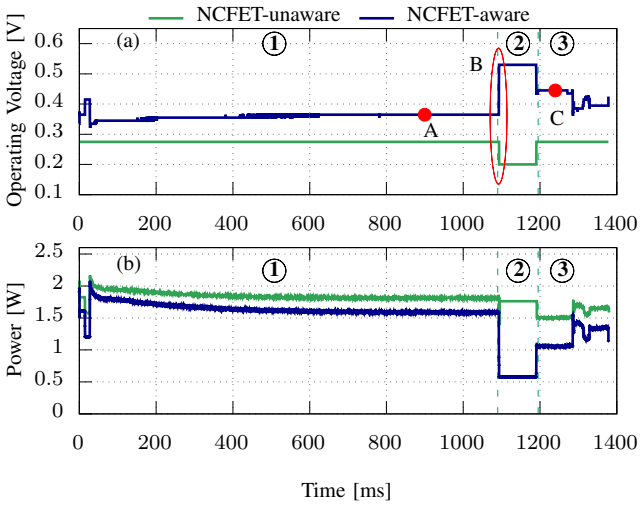

Fig. 8: (a) operating voltage V_{dd} and (b) total power consumption of the *canneal* master thread with our NCFET-aware DVS and NCFET-unaware DVS. NCFET-aware DVS overscales the voltage based on the workload and thereby reduces the total power by up to 67 % in phase-2 at the same CPU frequency and results in total energy savings of 17 %. Traditional DVS (using V-f pairs determined at design time) fails when it comes to NCFET. As shown for point A and C, they have the same frequency but different voltages. At point B, voltage is *antagonistically* selected between the two DVS techniques.

C. Energy Savings with Our NCFET-Aware DVS

In this section, we evaluate the effectiveness of our NCFET-aware DVS. First, we show how NCFET-aware DVS saves power, then we show how total power saving varies at runtime. Lastly, to demonstrate the effectiveness of NCFET-aware DVS technique, we report the energy saving for different benchmarks in comparison with NCFET-unaware DVS.

Fig. 8 presents an illustrative example. The master thread of PARSEC *canneal* shows distinct phases during execution. The total power consumption during phase-1 gradually decreases as shown in Fig. 8b. The frequency is set at 1.7 GHz. Traditional DVS sets V_{dd} to the minimum voltage (0.28 V) required to sustain this frequency. Thereby, dynamic power is minimized, but the leakage power is high. Our NCFET-aware DVS sets V_{dd} to a higher value (up to 0.37 V), which increases the dynamic power but stronger decreases leakage power resulting in a lower total power compared to traditional DVS. As can be noticed, V_{dd} is not constant, but it increases slightly over time. This is because dynamic power decreases and therefore it is more beneficial to decrease the leakage power. It is important to note, that operating voltages in NCFET are lower than traditional FET due to the inherent amplification in NCFET provided by the ferroelectic layer.

During phase-2, the master thread is then idle and awaits termination of the slave threads. The frequency is reduced to the minimum frequency (1.0 GHz). Here, leakage power dominates the total power. Traditional DVS would reduce V_{dd} down to 0.2 V due to the low required frequency. Operating at such a low voltage strongly increases the leakage power in NCFET. Our NCFET-aware DVS, instead of reducing V_{dd}, boosts the voltage to 0.53 V to minimize the leakage power. Thereby, the total power consumption during phase-2 is decreased by 67 % compared to the traditional DVS. Once

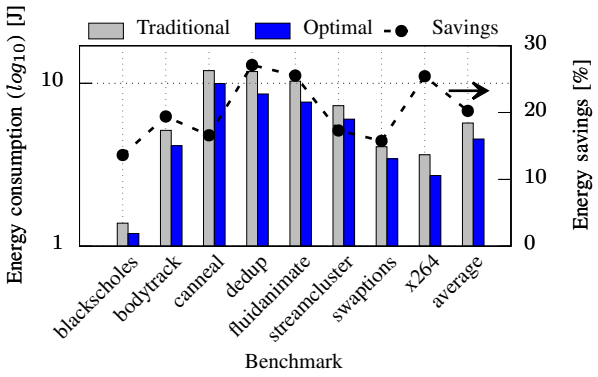

Fig. 9: Energy results and energy savings of different benchmarks running at 1.7GHz using our NCFET-aware DVS compared to NCFET-unaware DVS. Our NCFET-aware DVS technique results in up to 27 % energy savings (20 % on average) while still providing the same performance.

the slaves terminated, the master resumes operation in phase-3 and its frequency is increased again to 1.7 GHz. However, CPU activity here is very low due to the frequent memory accesses. Our NCFET-aware DVS is able to exploit this by using a higher V_{dd} than in phase-1, even though the frequency is the same. The total energy consumption of all phases has been reduced by 17 %. *It is important to notice that our technique does not simply statically increase V_{dd}, but in fact it results in opposite behavior. Traditional DVS decreases V_{dd} in phase-2, whereas V_{dd} needs to be increased to minimize the total power as shown at point B in Fig. 8a where V_{dd} is antagonistically selected. Furthermore, the V-f pairs model, which is used in traditional DVS, does not hold anymore with NCFET. As shown in Fig. 8a for points A and C, the CPU operates at the same frequency but have different selected V_{dd}.*

Fig. 9 summarizes the energy savings for different PARSEC benchmarks with *simsmall* inputs when active threads are operated at 1.7 GHz and idle cores are throttled to 1.0 GHz. The DVS techniques (i.e., NCFET-aware and NCFET-unaware) do not affect performance since the frequency is the same with both techniques. Therefore, the only difference is the selected V_{dd}, which results in a different total power. Energy savings range from 14 % for *blackscholes* up to 27 % for *dedup*.

Two factors affect the observed gains: (1) the CPU utilization which affects the dynamic power consumption of active threads and (2) the idle times of threads which result from synchronization between threads. The higher the total power consumption of active threads, the lower are the possible gains for these threads. This is the reason why e.g., *swaptions* results in low gains. Long idle times of threads result in higher gains since the total power consumption during idle times mainly consist of leakage that is reduced by our NCFET-aware DVS technique. This is a reason why *fluidanimate* has high savings. Overall, the average energy saving is 20 %.

V. SUMMARY AND CONCLUSIONS

NCFET is a promising emerging technology that has recently become compatible with standard CMOS technology. In this work, we presented the first NCFET-aware DVS. The necessity for a novel DVS technique stems from the inverse voltage dependency that leakage power has in NCFET technology caused by the negative DIBL effect. Our NCFET-aware DVS, implemented at the 7nm FinFET, selects the optimal voltage at runtime following the induced dynamics by running workloads. Compared to traditional (NCFET-unaware) DVS, our technique results in up to 27% energy saving because it does consider the novel runtime trade-off between leakage and dynamic power that NCFET brings. *NCFET-aware DVS is key to achieve the highest level of power efficiency in this new emerging technology.*

ACKNOWLEDGMENTS

This work (except the NCFET part) was supported in parts by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 146371743 – TRR 89 “Invasive Computing”.

REFERENCES

[1] S. Salahuddin and S. Datta, “Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices,” *Nano Letters*, vol. 8, no. 2, 2008.
[2] M. Hoffmann, F. P. Fengler, M. Herzig *et al.*, “Unveiling the Double-Well Energy Landscape in a Ferroelectric Layer,” *Nature*, 2019.
[3] H. Amrouch, G. Pahwa, A. D. Gaidhane *et al.*, “Negative Capacitance Transistor to Address the Fundamental Limitations in Technology Scaling: Processor Performance,” *IEEE Access*, vol. 6, 2018.
[4] Z. Krivokapic, U. Rana, R. Galatage *et al.*, “14nm Ferroelectric FinFET Technology with Steep Subthreshold Slope for Ultra Low Power Applications,” in *IEEE Int. Electron Devices Meeting (IEDM)*, Dec 2017.
[5] G. Pahwa, T. Dutta, A. Agarwal *et al.*, “Designing Energy Efficient and Hysteresis Free Negative Capacitance FinFET with Negative DIBL and 3.5 XI ON Using Compact Modeling Approach,” in *European Solid-State Circuits Conference (ESSCIRC)*, 2016.
[6] M. Rapp and S. Salamin and H. Amrouch and G. Pahwa and Y. S. Chauhan and J. Henkel, “Performance, Power and Cooling Trade-Offs with NCFET-based Many-Cores,” *Design Automation Conference (DAC)*, 2019.
[7] G. Pahwa, A. Agarwal, and Y. S. Chauhan, “Numerical Investigation of Short-Channel Effects in Negative Capacitance MFIS and MFMIS Transistors: Subthreshold Behavior,” *IEEE Transactions on Electron Devices (TED)*, vol. 65, no. 11, 2018.
[8] “BSIM-CMG Technical Manual,” October 2019, http://www-device.eecs.berkeley.edu/bsim/?page=BSIMCMG.
[9] G. Pahwa, T. Dutta, A. Agarwal *et al.*, “Analysis and Compact Modeling of Negative Capacitance Transistor with High ON-Current and Negative Output Differential Resistance—Part II: Model Validation,” *Transactions on Electron Devices (TED)*, vol. 63, no. 12, 2016.
[10] C. Bienia, S. Kumar, J. P. Singh *et al.*, “The PARSEC Benchmark Suite: Characterization and Architectural Implications,” in *Parallel Architectures and Compilation Techniques (PACT)*, 2008.
[11] Choi, Jung Hwan and Murthy, Jayathi and Roy, Kaushik, “The Effect of Process Variation on Device Temperature in FinFET Circuits,” in *International Conference on Computer-aided Design (ICCAD)*, 2007.
[12] T. Kuroda, “Optimization and Control of VDD and VTH for Low-power, High-speed CMOS Design,” in *International Conference on Computer-aided Design*, ser. ICCAD. ACM, 2002.
[13] J. Balkind, M. McKeown, Y. Fu *et al.*, “OpenPiton: An Open Source Manycore Research Framework,” in *Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, ser. ASPLOS, 2016.
[14] H. Amrouch and S. Salamin and G. Pahwa and A. Gaidhane and J. Henkel and Y. S. Chauhan, “Unveiling the Impact of IR-drop on Performance Gain in NCFET-based Processors,” *Transactions on Electron Devices (TED)*, 2019.
[15] L. T. Clark, V. Vashishtha, L. Shifren *et al.*, “ASAP7: A 7-nm FinFET predictive process design kit,” *Microelectronics Journal*, vol. 53, 2016.
[16] “Voltus IC Power Integrity Solution,” https://www.cadence.com.
[17] “Tempus Timing Signoff Solution,” https://www.cadence.com.
[18] A. Pathania and J. Henkel, “HotSniper: Sniper-Based Toolchain for Many-Core Thermal Simulations in Open Systems,” *IEEE Embedded Systems Letters (ESL)*, 2018.
[19] T. E. Carlson, W. Heirman, and L. Eeckhout, “Sniper: Exploring the Level of Abstraction for Scalable and Accurate Parallel Multi-Core Simulation,” in *Int. Conf. for High Performance Computing, Networking, Storage and Analysis (SC)*. ACM, 2011.
[20] S. Li, J. H. Ahn, R. D. Strong *et al.*, “The McPAT Framework for Multicore and Manycore Architectures: Simultaneously Modeling Power, Area, and Timing,” *TACO*.
[21] NanGate, “Open Cell Library,” https://www.silvaco.com/.