Chapter 8
Impact of Negative Capacitance Field-Effect Transistor (NCFET) on Many-Core Systems

Hussam Amrouch, Martin Rapp, Sami Salamin, and Jörg Henkel

8.1 Introduction

More than a decade ago, the semiconductor technology had entered the so-called nano-CMOS era, in which the transistor's feature sizes became below 90 nm. Since then, the prior trend of voltage scaling came to an end leading to the discontinuation of Dennard's scaling [7]. In Dennard's scaling, both the dimensions of transistor and the operating voltage are typically scaled by the same factor in order to ensure a constant electric field. Due to the non-scalable voltage, ever-increasing power densities in chips became a substantial obstacle for technology scaling due to the limited ability of existing cooling solutions to dissipate the generated heat [8]. To overcome this fundamental problem, the maximum frequency of processors had stopped increasing with every new generation in order to keep the on-chip power densities under acceptable levels and since 2005 the era of many-core processors had started.

To understand the inability of technology to scale voltage, we need to understand what determines the speed of a processor. As a matter of fact, the drive current (ON current) of a transistor dictates its switching speed and hence it ultimately determines the maximum delay of logic paths that form the processor's netlist. The ON current of a transistor is proportional to $(V_{DD} - V_T)$, where V_T denotes the threshold voltage of transistor and V_{DD} denotes the operating voltage. In order to maintain the same level of current, while V_{DD} is scaled down, V_T must also be reduced by almost the same amount. However, reducing V_T comes with an exponential increase in the leakage current (OFF current) of transistor. This is primarily because that the sub-threshold swing of transistor is fundamentally

H. Amrouch · M. Rapp · S. Salamin · J. Henkel (✉)
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: amrouch@kit.edu; henkel@kit.edu

J.-J. Chen (ed.), *A Journey of Embedded and Cyber-Physical Systems*,
https://doi.org/10.1007/978-3-030-47487-4_8

limited to 60 mV/decade at room temperature akin to “Boltzmann tyranny” [21]. Such a fundamental limit inevitably restricts the minimum possible V_T to be at least 300 mV. To ensure a reliable operation, different kinds of safety margins need to be added on top of the minimum voltage, which enforces the operating voltage to remain almost the same with every new technology generation. As abovementioned, the inability to scale voltage has led to the discontinuation of Dennard’s scaling, which, in turn, had led to preventing the frequency of processors from increasing.

In summary, the fundamental limit of sub-threshold swing of transistor is the primary reason behind not scaling voltage and it is the origin of on-chip power density problems that processor’s designers are facing since more than a decade ago.

8.1.1 Negative Capacitance Field-Effect Transistor (NCFET)

NCFET integrates a ferroelectric layer inside the gate stack of a transistor, which acts as a negative capacitance. Such a layer provides an amplification of the vertical electric field that the transistor perceives. This, in turn, allows the transistor to overcome the fundamental limitation of sub-threshold swing of 60 mV at room temperature. The principle of NCFET was first proposed in 2008 by S. Salahuddin and S. Datta [16]. After which, it very rapidly gained a large popularity due to the remarkable steep switching and high ON current of transistors [1]. Many experiments have consistently proved NCFET [10]. A breakthrough has recently occurred when GlobalFoundries demonstrated NCFET-based circuits using their state-of-the-art industrial 14 nm FinFET technology [9]. This showed, for the first time, that NCFET technology has become compatible with the existing CMOS fabrication process. In fact, such a compatibility is essential for any emerging technology to be adopted by semiconductor companies. Otherwise, massive production will never be possible.

In practice, NCFET technology enables the transistor to reach the same ON current, without increasing the OFF current, but at a much lower voltage [2]. This is only possible due to steeper sub-threshold swing. Therefore, in an NCFET technology, the processor can still meet the same performance (as in the conventional FET) but at a lower operating voltage leading to a significant power saving. Beside the *low-power* usage scenario of NCFET, *high-performance* usage scenario does also exist. NCFET enables the processor to be clocked at a higher frequency (compared to the conventional FET), while it still be operated at the same voltage due to the increase in the ON current. NCFET technology comes with an important side effect in which it increases the total capacitance of transistor. Such an increase can lead to reliability problems caused by IR-drop and voltage fluctuation during circuit’s operation [2, 18]. At the same time, because NCFET technology enables circuits to operate at lower voltages, it is expected that other reliability problems, related to lifetime, to become much less because all the underlying aging mechanisms, such as negative bias temperature instability (BTI) and hot-carrier injection (HCI), strongly depend on the operating voltage [20].

In the following sections, we explain how modeling the NCFET effects from physics all the way to the system level can be done. Then, we explore how a many-core system can profit from the NCFET technology. Finally, we explore the impact that NCFET has on power management schemes and how existing assumptions w.r.t voltage-leakage dependency become not valid anymore when it comes to NCFET, which creates the necessity to develop novel power management techniques.

8.2 Modeling NCFET at the System Level

In the following we provide an overview of how NCFET is modeled at the system level, i.e., for the purpose of simulating many-core processors. Fundamentally, the properties of the ferroelectric layer are modeled at the physics level [12]. Figure 8.1 presents our methodology in which we traverse all layers from physics, through device, gate, and processor level, to model NCFET at the system level. The behavior of transistors with varying thickness of the ferroelectric layer is modeled following the industrial-standard compact model (BSIM-CMG) [5, 14]. Based on this model, we created NCFET-aware cell libraries supporting four different thicknesses of the ferroelectric layer under a wide range of the operating voltage [1]. The thickness ranges from 1 nm (called TFE1) up to 4 nm (TFE4). We then implemented a single many-core tile to the GDSII level and performed timing and power signoffs. The results are explained in detail in the next section. Signoff tools allow to compare power and performance of a processor implemented in different NCFET configurations and are used to extract frequency-dependent scaling factors for dynamic and leakage power. These factors serve as an abstraction at the system level and allow to estimate the power of an NCFET-based processor if the power of

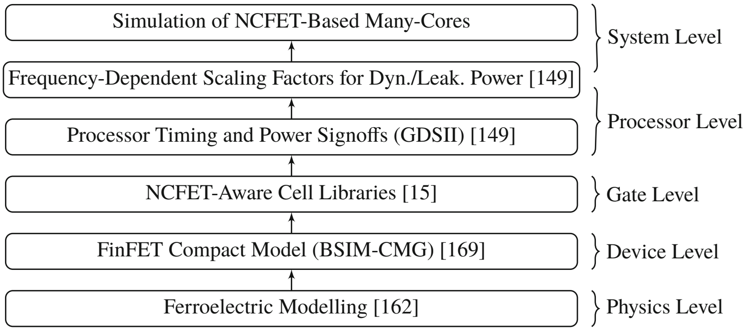
Simulation of NCFET-Based Many-Cores
Frequency-Dependent Scaling Factors for Dyn./Leak. Power [149]
Processor Timing and Power Signoffs (GDSII) [149]
NCFET-Aware Cell Libraries [15]
FinFET Compact Model (BSIM-CMG) [169]
Ferroelectric Modelling [162]
System Level
Processor Level
Gate Level
Device Level
Physics Level

Fig. 8.1 Modeling NCFET at the system level (many-core processors) requires to traverse the whole stack from the physics level, where the effects of the ferroelectric layer are modeled, to the system level, where performance and power of many-core processors are affected

a baseline implementation (conventional FinFET) is known. Finally, these factors are used to simulate a many-core processor (further details in Sect. 8.2.2).

8.2.1 Processor-Level Investigation

This section shows how NCFET affects the performance and power of a single processor. The insights gained from this evaluation are important to build system-level NCFET models and explain observations from system-level simulations. We implemented the layout (GDSII level) of a single tile of the *OpenPiton* many-core [3], which contains a CPU, caches, and a NoC router. Power and timing signoffs are performed for different NCFET configurations (TFE1 to TFE4) and different operating voltages. Further details of the experimental setup can be found in [15].

Figure 8.2a shows how NCFET increases the performance of a processor. It allows to clock a processor at a higher frequency at the same operating voltage or allows to reduce the voltage while still maintaining the same performance (frequency). This is due to the inherent voltage amplification provided by the additional

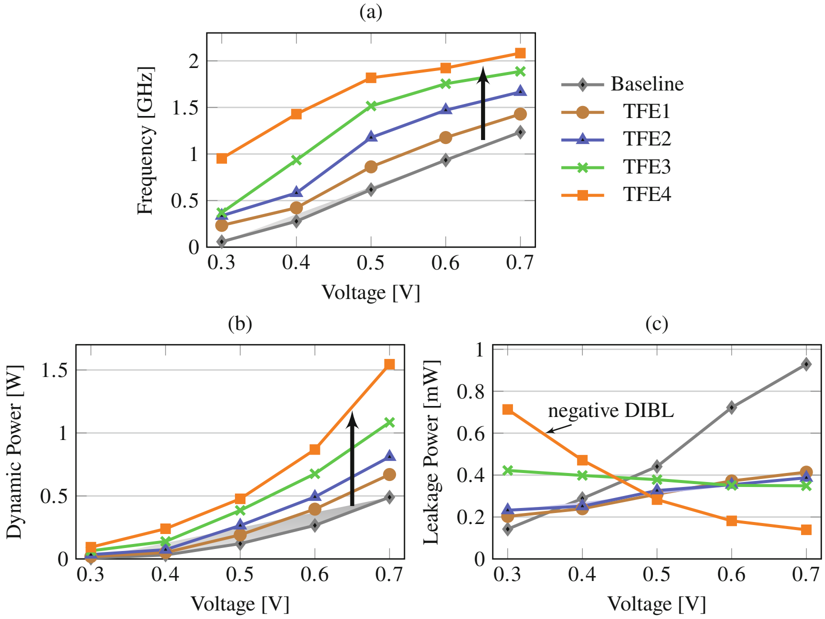

Fig. 8.2 (**a**) NCFET increases the frequency of a processor at a certain operating voltage, but (**b**) also increases the dynamic power consumption due to the increase in the transistor gate capacitance and frequency. (**c**) While leakage increases almost linearly with the operating voltage with conventional FinFET (baseline), this dependency gets weaker with a thin ferroelectric layer and even reverses with TFE4 due to a negative DIBL effect

ferroelectric layer. Like explained earlier, the ferroelectric layer increases the total gate capacity. Together with increased frequency, this increases the dynamic power consumption (Fig. 8.2b). The thicker the ferroelectric layer gets, the higher get the gains in the frequency, but also the higher gets the dynamic power. Figure 8.2c shows that leakage power is affected more severely. NCFET fundamentally changes the trend. With conventional FinFET (baseline), leakage power increases strongly with increasing voltage. When a thin ferroelectric layer is added (TFE1 and TFE2), this dependency becomes weaker, until at TFE3, leakage is almost independent of the voltage. With a thick ferroelectric layer (TFE4), an effect called negative drain-induced barrier lowering (negative DIBL) reverses the leakage dependency on the voltage [13]. Here, leakage increases at lower voltages. We will explain later (Sect. 8.3.3) how this necessitates developing novel power management techniques.

8.2.2 Simulation of NCFET-Based Many-Core

We use the *Sniper* many-core simulator [6] to simulate many-core processors. *McPAT* [11] is used to periodically estimate the power consumption of each core. Since *McPAT* does not support NCFET, it is used to estimate the power with conventional FinFET instead. We develop frequency-dependent scaling factors for dynamic and leakage power based on the processor-level investigation explained earlier.

Figure 8.3 shows the dynamic and leakage power of the single processor studied in the previous section depending on the *frequency*, as opposed to voltage like in Fig. 8.2. Two effects play a role for the dynamic power: NCFET technologies increase the dynamic power at a certain operating voltage (Fig. 8.2b), but also

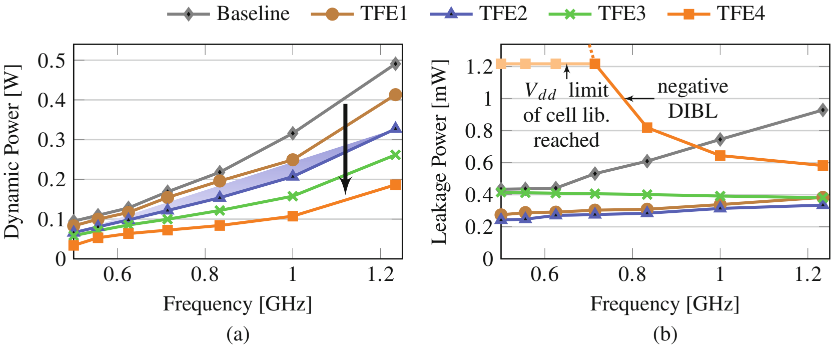

Fig. 8.3 (**a**) While NCFET technologies increase the dynamic power at iso-voltage, they also lower the required operating voltage at iso-frequency, which in total decreases the dynamic power at the same frequency. (**b**) NCFET technologies with a thin ferroelectric layer lower the leakage power, whereas leakage increases with a thick layer (TFE4). Most importantly, the negative DIBL effect reverses the leakage dependency, where lowering the V/f-levels increases leakage

allow to go to a lower operating voltage (Fig. 8.2a) while still maintaining the same frequency. Lowering the operating voltage has a stronger effect on the dynamic power. Consequently, NCFET technologies lower the dynamic power when operating at the same frequency (Fig. 8.3a). Figure 8.3b shows how leakage power depends on the V/f-level. The reverse leakage dependency with TFE4 strongly increases the leakage power. Below 700 MHz, TFE4 would allow to reduce the voltage below 0.2 V, which is the lower limit of the cell library.

Figure 8.3a,b allows to estimate the dynamic and leakage power consumption of a processor that is implemented in NCFET, if the power consumption in the baseline (conventional FinFET) is known. We extract frequency-dependent scaling factors for both dynamic and leakage power. These factors serve as an abstraction that allows simulation of complex benchmark applications, like *PARSEC* [4], on many-core processors with dozens of cores. We thereby scale the leakage and dynamic power that is estimated by *McPAT* to estimate the power consumption of NCFET-based many-cores. For brevity, details on this approach are omitted here and can be found in [15].

8.3 Performance, Power, and Cooling Trade-Offs with NCFET-based Many-Cores

NCFET fundamentally changes the characteristics of transistors and therefore also changes the performance and power of circuits [19], single-core processors [1], and many-core processors [15]. This section demonstrates the impact of the thickness of the ferroelectric layer on the power and performance of a many-core processor. We show that the optimal thickness depends on many factors, such as the application characteristics and the cooling scenario. This section evaluates performance, power, and cooling of a 25-core many-core operating under a thermal constraint of 80°C. We study *PARSEC* [4] tasks with up to eight slave threads. Their characteristics range from highly memory-bound (e.g., *canneal*) to highly compute-bound (e.g., *swaptions*).

8.3.1 Impact of NCFET on Performance

Due to high power densities (failure of Dennard's scaling) and limited cooling capabilities, it is not always possible in modern technology nodes to simultaneously operate all cores at the peak V/f-levels without violating the thermal constraint. This study investigates the use-case in which cores with an active thread are operated at the peak V/f-levels and cores without a thread mapped to it are power-gated. In this use-case, four factors affect the thermally sustainable utilization (i.e., the number of cores that can be turned on): the application characteristics (power consumption), the mapping of threads to cores, the cooling system, and the transistor technology.

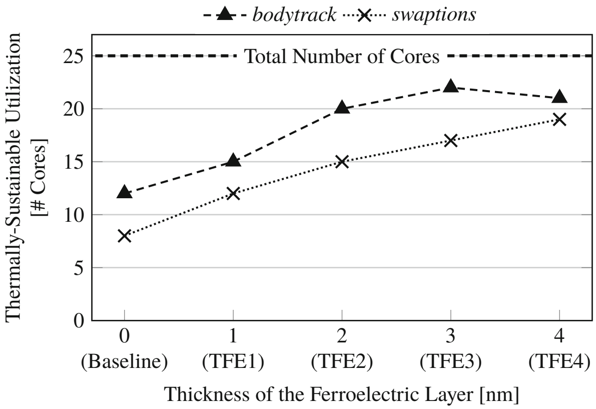

Fig. 8.4 NCFET technologies increase the thermally sustainable utilization of a 25-core many-core, i.e., the number of usable cores without violating the thermal constraint, compared to the baseline (conventional FinFET). The optimal thickness of the ferroelectric layer depends on the application characteristics

We use an Integer Linear Program to obtain the thermally-optimal mapping of threads to cores, which minimizes the formation of hotspots and, thereby, maximizes the thermally sustainable utilization. We study the use-case of a passive cooling, i.e., there is no fan on top of the heat sink.

Figure 8.4 shows the thermally sustainable utilization of two benchmarks *bodytrack* and *swaptions* during the parallel section of the benchmarks (Region of Interest) for different NCFET technologies. Other benchmarks are available in [15]. *Swaptions* is a highly compute-intensive task, which results in high power consumption and therefore, the thermally sustainable utilization in the baseline is low (only 8 out of 25 cores). Dynamic power forms the major part of the total power consumption and therefore, thicker ferroelectric layers increase the thermally sustainable utilization because dynamic power is reduced (compare Fig. 8.3a). Consequently, the highest performance is observed with the thickest ferroelectric layer (TFE4). *Bodytrack* is less compute-intense and has lower dynamic power consumption and consequently lower total power. This results in a higher thermally sustainable utilization compared to *swaptions*. However, due to lower dynamic power, leakage power accounts for a larger fraction of the total power. As demonstrated in Fig. 8.3b, TFE4 increases the leakage significantly over TFE3. Consequently, TFE4 results in a lower thermally sustainable utilization than TFE3 for *bodytrack* and the highest performance is observed with TFE3.

These investigations show that *the optimal thickness of the ferroelectric layer depends on the application characteristics*. Further investigations on how NCFET affects the performance in the case that cores are not operated at the peak V/f-levels can be read in [15]. These investigations additionally study forced-convection cooling (a heat sink with a fan) and reveal that *the optimal thickness of the ferroelectric layer also depends on the cooling scenario*.

8.3.2 Impact of NCFET on Cooling Requirements

This section studies how NCFET reshapes the existing trade-off between cooling costs and achievable performance, where higher performance comes at the cost of higher power dissipation and therefore higher cooling costs. We study the use-case in which the many-core is operated at its peak performance, i.e., all cores are active at the peak V/f-levels and determine the cooling capabilities that are required to make this use-case thermally safe. The cooling capabilities are measured by the inverse of thermal resistance of the heat sink $1/R_{th}$. Varying this value corresponds to changing the air convection.

Figure 8.5 shows the required cooling capabilities for the three *PARSEC* benchmarks *swaptions*, *bodytrack*, and *canneal* during the parallel section of the benchmarks (Region of Interest). NCFET technologies allow to reduce the cooling capabilities over the baseline (conventional FinFET). Most importantly, the required cooling capabilities are minimized at different thicknesses of the ferroelectric layer depending on the application. *Swaptions* is highly compute-intensive and consequently, dynamic power accounts for the majority of the total power. Increasing the thickness of the ferroelectric layer reduces the dynamic power (see Fig. 8.3) and therefore reduces the required cooling. *Canneal* on the other side is highly memory-bound and therefore, the power consumption is dominated by leakage. Leakage is minimized at TFE2, which consequently minimizes the cooling requirements. *Bodytrack* shows intermediate values for the dynamic power and therefore, TFE3 is optimal. *This investigation shows again that the optimal thickness of the ferroelectric layer depends on the application characteristics and ranges from 2 nm to 4 nm.*

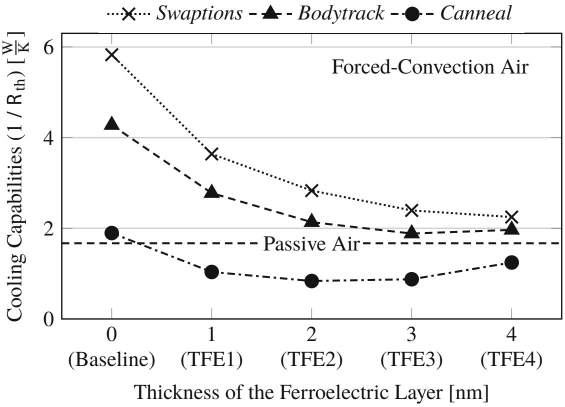

Fig. 8.5 NCFET technologies decrease the required cooling capabilities while maintaining the same maximum temperature of 80°C under full system utilization (all cores active at peak V/f-levels). The thickness of the ferroelectric layer that results in the lowest cooling costs depends on the application characteristics and ranges from 2 nm (with *canneal*) to 4 nm (with *swaptions*)

8.3.3 Impact of NCFET on Power Management Techniques

The above investigations use the well-established concept of V/f-pairs that are determined at design time by selecting the operating voltage for a given frequency as the lowest voltage that makes operating at this frequency reliable. This is a reasonable approach with conventional transistor technologies, because using a higher voltage would unnecessarily increase both dynamic and leakage power. However, this is no longer true with NCFET with a thick ferroelectric layer (TFE4). Here, increasing the voltage decreases the leakage power. This leads to new optimization potential by selecting the operating voltage for a given frequency, which is demonstrated in the next section.

8.4 NCFET-Aware Voltage Scaling

Dynamic voltage scaling (DVS) technique for processor power management is considered to be one of the most effective ways to reduce the energy consumption of an application. DVS technique typically selects the minimum operating voltage $V_{\min}$ that sustains the operating frequency of the processor at runtime based on the frequency demands of the application being executed. Reducing the operating voltage, in conventional FET, results in reducing the total power consumption, which implicitly reduces both dynamic and leakage power. However, such a well-known voltage dependency becomes inverse with respect to leakage power in NCFET due to the negative DIBL effect (see Sect. 8.2.1). With such opposed dependencies (dynamic and leakage) to the operating voltage, total power follows the dominant component when voltage changed which leads to a novel trade-off. Consequently, power is not necessarily minimized at the minimum voltage $V_{\min}$, which traditional DVS selects, but at another voltage V_{opt}. Unawareness of NCFET and its trade-off could lead to not minimize the total power consumption. Therefore, in this section, a novel NCFET-aware voltage scaling technique is presented [17] to overcome the shortness that traditional DVS has in NCFET-based processors.

8.4.1 Importance of NCFET-Aware DVS

With traditional DVS, a set of voltage-frequency pairs are typically selected at design time and later are employed by the DVS technique at runtime to optimize the power. In this case, the lower the selected voltage is, the lower the total power is. Due to the new inverse dependency in leakage power that NCFET exhibits, this is not always valid with respect to NCFET. To demonstrate the consequence of such an inverse dependency at the system level, we plot the total power consumption and its components of the master thread of PARSEC *canneal* benchmark in Fig. 8.6.

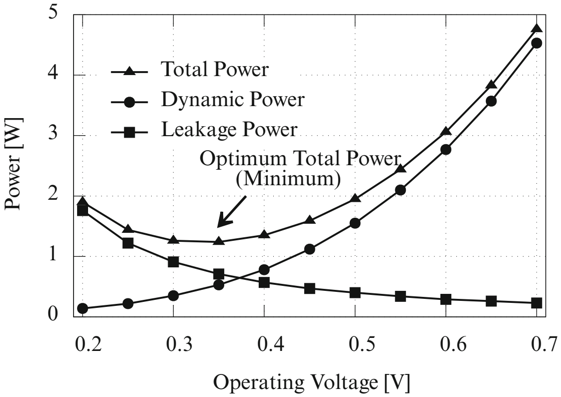

Fig. 8.6 Total power and its components (i.e., leakage and dynamic) of *canneal* master thread starting from the minimum voltage $V_{\min}$ required to sustain 1.0 GHz frequency. The total power is not minimized at $V_{\min}$. The operating voltage required to minimize total power V_{opt} appears at a higher voltage than $V_{\min}$ due to leakage increases in NCFET

The power examined starting from $V_{\min}$, that traditional DVS selects to sustain the required frequency, and then to overscale the operating voltage. The result shows that the power is not minimized at $V_{\min}$ (i.e., $V_{opt} \neq V_{\min}$).

Different workloads exhibit different characteristics and hence different total power. Therefore, the contribution of power components differs. Traditional DVS neglects this difference as both contributions (leakage and dynamic) are affected in the same manner with voltage (both are reduced). With NCFET, the contribution of leakage to the total power cannot be neglected because it affects the operating voltage selection when DVS tries to minimize total power. Hence, based on the leakage share, V_{opt} could differ from $V_{\min}$.

For the aforementioned reasons, NCFET-aware DVS is crucial due to the change in the behavior of total power consumption over voltage scaling which emerges from the inverse dependency with respect to leakage power in NCFET.

8.4.2 NCFET-Aware DVS Technique

To enable runtime voltage selection, DVS first needs to determine workload characteristics and then V_{opt} can be correctly selected. Therefore, determining V/f-pairs at runtime, like in traditional DVS techniques, is not possible here. Instead, the results from Sect. 8.2.1 have been used to build the power and performance analytical models at design time. Then, these models can be integrated with our new NCFET-aware DVS technique for runtime voltage selection.

8.4.2.1 Design-Time Models

Power and Performance Modeling The maximum operating frequency $f_{\max}(V)$ depends on the voltage V over the minimum delay $d_{\min}(V)$:

$$d_{\min}(V) = a_{del} \cdot V^{b_{del}} + c_{del}; \quad f_{\max}(V) = \frac{1}{d_{\min}(V)} \tag{8.1}$$

$a_{del}>0$, $b_{del}<0$, $c_{del}\geq 0$ are constants fitting parameters obtained at design time. Peak leakage and peak dynamic power consumption results by operating at maximum frequency are

$$P_{leak}(V) = a_{leak} \cdot V^{b_{leak}} \tag{8.2}$$

$$P_{dyn}^{peak}(V, d_{\min}(V)) = a_{dyn} \cdot V^{b_{dyn}} + c_{dyn} \tag{8.3}$$

$a_{dyn}>0$, $b_{dyn}>1$, $c_{dyn}\geq 0$, $a_{leak}>0$, $b_{leak}<0$ are constant fitting parameters obtained at design time. Both $P_{dyn}^{peak}(V, d_{\min}(V))$ and $P_{leak}(V)$ are *convex* in V. By lowering the operating frequency of the CPU (higher delay), dynamic power decreases. However, since leakage power is independent from CPU activity, it is not affected.

$$P_{dyn}^{peak}(V, d) = \frac{d_{\min}(V)}{d} \cdot P_{dyn}^{peak}(V, d_{\min}(V)) \tag{8.4}$$

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

8.4.2.2 Runtime Models

Workload-Dependent Power Modeling Dynamic power consumption $P_{dyn}(V, d)$ is affected by the running workload, which is reduced by a factor $0\leq r_{dyn}\leq 1$ from the peak dynamic power $P_{dyn}^{peak}(V, d)$:

$$P_{dyn}(V, d) = r_{dyn} \cdot P_{dyn}^{peak}(V, d) \tag{8.5}$$

$$P_{total}(V, d) = P_{dyn}(V, d) + P_{leak}(V) \tag{8.6}$$

r_{dyn} is not constant since it represents the current workload activity. Therefore, total power consumption $P_{total}(V_c, d)$ at the current voltage V_c, r_{dyn} is

$$r_{dyn} = \frac{P_{dyn}(V_c, d)}{P_{dyn}^{peak}(V_c, d)} = \frac{P_{total}(V_c, d) - P_{leak}(V_c)}{P_{dyn}^{peak}(V_c, d)} \tag{8.7}$$

Optimal Voltage Computing V_{opt} that minimizes the total power can be obtained from the power and performance models:

$$V_{\min}(d) = \left(\frac{d - c_{del}}{a_{del}}\right)^{\frac{1}{b_{del}}} \tag{8.8}$$

Algorithm 1 NCFET-aware voltage scaling algorithm to select the optimal voltage (V_{opt}) at runtime [17]

```
Require: Power and performance models: $P_{dyn}^{peak}(c, d)$ and $P_{leak}(V)$, current operating voltage
    $V_c$ and delay $d$, current power consumption $P_{curr}$, min. voltage resolution $\epsilon$
Ensure: Optimal operating voltage $V_{opt}$
$r_{dyn} \leftarrow (P_{curr} - P_{leak}(V_c)) / P_{dyn}^{peak}(V_c, d)$                ▷ Eq. (8.7)
$V_{opt} \leftarrow V_{\min}(d)$                                                      ▷ Eq. (8.8)
repeat
    $\Delta V_{opt} \leftarrow -P_{total}(V_{opt}, d)' / P_{total}(V_{opt}, d)''$
    $V_{opt} \leftarrow V_{opt} + \Delta V_{opt}$                                     ▷ 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 $\Delta V_{opt} < \epsilon$                                                     ▷ Termination criteria
return $V_{opt}$
```

$$V_{opt}(d, r_{dyn}) = \underset{V_{\min}(d) \leq V \leq V_{\max}}{\arg\min} \; P_{total}(V, d) \tag{8.9}$$

Since $P_{total}(V, d)$ is composed of convex functions, our implemented algorithm exploits that $P_{total}(V, d)$ is convex in V. This guarantees that $P_{total}(V, d)$ has exactly one minimum w.r.t. V within the range $[V_{\min}(d), V_{\max}]$. Algorithm 1 summarizes our implemented DVS technique and obtaining V_{opt}.

8.4.3 Operating Voltage Selection

Both DVS techniques differ in the way they select the operating voltage. Therefore, to show the different behavior between both techniques in operating voltage selection, the design space of the operating voltage selection with NCFET-aware (V_{opt}) and NCFET-unaware DVS ($V_{\min}$) has been explored in Fig. 8.7. NCFET-unaware DVS sets $V_{\min}$ that is needed to sustain the required frequency and therefore workload characteristic is not considered. Contrarily, NCFET-aware DVS considers the workload characteristic as it depends on the ratio of leakage to total power measured at $V_{\min}$. The explored design space in Fig. 8.7 shows two distinct regions: (1) For low leakage to total power ratio and for high frequencies, the same voltage is selected (similar action) by both techniques (i.e., $V_{opt}{=}V_{\min}$). (2) For high ratios of leakage to total power or low frequencies, NCFET-aware DVS selects a higher voltage ($V_{opt} > V_{\min}$). Moreover, Fig. 8.7 reveals that: the higher the required frequency or the higher the leakage to total power ratio, the higher V_{opt} is.

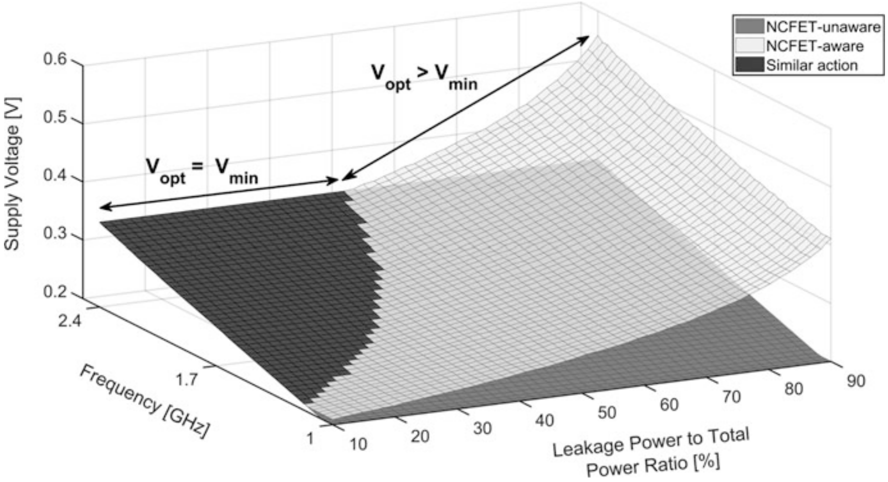

Fig. 8.7 Operating voltage selection using both DVS techniques. Two regions appear: (1) NCFET-aware selection differs from NCFET-unaware ($V_{\min} \neq V_{opt}$). (2) Similar action is done by both DVS as they select the same operating voltage ($V_{\min} = V_{opt}$). NCFET-unaware DVS selects $V_{\min}$ (that sustains the required frequency) and NCFET-aware selects V_{opt} to minimize the power depending on the frequency and the ratio of leakage to total power. NCFET-aware DVS selects higher voltages when leakage power becomes prominent or at lower frequency

8.4.4 Evaluation

8.4.4.1 Experimental Setup

Using the same setup in Sect. 8.2.1, power and delay results were examined using the highest ferroelectric thickness (4 nm). Afterwards, the power and performance analytical models have been developed as described in Sect. 8.4.2.

For system-level simulation, relying on the setup described in Sect. 8.2.2, the NCFET-aware DVS technique (Algorithm 1) has been used to select the operating voltage when a set of tasks were examined from the PARSEC benchmark suite [4]. The frequencies are set between 1.0 GHz and 2.4 GHz. V_{dd} is set between 0.2 V and 0.7 V. The low operating voltages V_{dd} in NCFET are lower than traditional FET due to the inherent voltage amplification in NCFET provided by the negative capacitance. For fair comparisons, simulators for both DVS cases were configured to have: the same frequencies, and architecture, in addition to running the same benchmarks. Hence, only voltage selection differs based on DVS decision.

8.4.4.2 NCFET-Aware DVS Results and Analysis

To show the effectiveness of the NCFET-aware DVS, we first show how NCFET-aware DVS actually operates to save power and later to report the energy savings

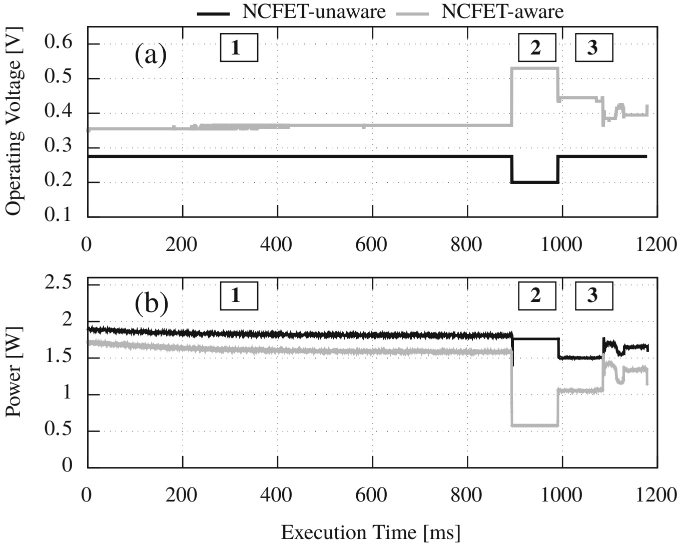

Fig. 8.8 (**a**) Operating voltage and (**b**) total power consumption during an interval of the execution time of the *canneal* master thread with NCFET-unaware and NCFET-aware DVS. NCFET-aware DVS selects higher voltage most of the time (in this particular example) and reduces the power further at the same CPU frequency. Voltage selection is based on workload characteristics

for different benchmarks in comparison with NCFET-unaware DVS. Accordingly, an illustrative example of the master thread of PARSEC *canneal* benchmark was selected. Figure 8.8 shows distinct phases during an interval of the execution time. In phase-1, in Fig. 8.8b, it shows the total power consumption when the frequency is set at 1.7 GHz. Traditional DVS sets V_{dd} to the minimum voltage (0.28 V) which required to sustain this frequency. Thus, dynamic power is minimized but the leakage power is not. NCFET-aware DVS sets V_{dd} to a higher value to guarantee a better trade-off. This will increase the dynamic power but strongly decreases leakage power resulting in a power saving. In phase-2, the master thread is idle and waits for the termination of the slave threads. Therefore, frequency is reduced to the minimum frequency (1.0 GHz). Traditional DVS reduces V_{dd} to 0.2 V due to the low required frequency in which it increases the leakage power. NCFET-aware DVS, instead of reducing V_{dd}, increases the voltage to 0.53 V, which decreases the leakage power. Thereby, the total power consumption in phase-2 is reduced by 67 % compared to the traditional DVS. In phase-3, after the slaves terminated, the master resumes operation and its frequency is boosted again to 1.7 GHz. It is worth to mention that the performance obtained with both DVS techniques is the same. This is because they do not affect the frequency, but only set the V_{dd} under performance constraint.

To reveal the energy savings, different PARSEC benchmarks were examined when active threads are operated at 1.7 GHz and idle cores are suppressed to 1.0 GHz. Figure 8.9 summarizes the energy savings. Energy savings range as shown in Fig. 8.9 from 14 % up to 27 % and in average are up to 20 %.

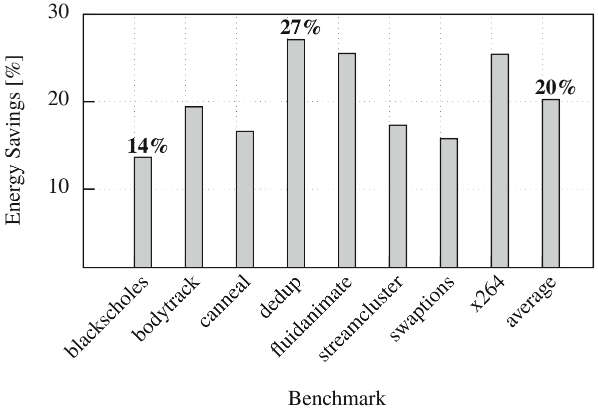

Fig. 8.9 Energy saving results of different benchmarks using the NCFET-aware DVS compared to NCFET-unaware DVS. Energy savings range from 14 % up to 27 % , and in average 20%

8.5 Conclusion

In this chapter, we investigated how NCFET technology impacts the existing trade-offs in processors and how it can reshape the future of many-core systems. Compared to the existing FinFET technology, NCFET technology allows the processor to operate at a much lower voltage while it still meets the same performance. This results in a considerable power saving and as a result the total number of cores, that can be simultaneously turned on at the maximum frequency, increases without violating the predetermined thermal constraints. We also showed how NCFET inverses the leakage-voltage dependency and proposed a new NCFET-aware DVS technique that provides an energy saving of 20% on average compared to conventional DVS techniques, which are unaware of the new leakage-voltage dependency that NCFET brings.

Acknowledgments The authors would like to thank Yogesh S. Chauhan, Girish Pahwa, and Amol Gaidhane from the Indian Institute of Technology Kanpur (IIT Kanpur) for their valuable contribution and support in the NCFET analysis at the device level and compact modeling.

References

1. H. Amrouch, G. Pahwa, A.D. Gaidhane, J. Henkel, Y.S. Chauhan, Negative capacitance transistor to address the fundamental limitations in technology scaling: processor performance. IEEE Access **6**, 52754–52765 (2018)
2. H. Amrouch, S. Salamin, G. Pahwa, A.D. Gaidhane, J. Henkel, Y.S. Chauhan, Unveiling the impact of IR-drop on performance gain in NCFET-based processors. IEEE Trans. Electron Devices **66**(7), 3215–3223 (2019). https://doi.org/10.1109/TED.2019.2916494

3. J. Balkind, M. McKeown, Y. Fu, T. Nguyen, Y. Zhou, A. Lavrov, M. Shahrad, A. Fuchs, S. Payne, X. Liang, M. Matl, D. Wentzlaff, OpenPiton: an open source Manycore research framework, in *Architectural Support for Programming Languages and Operating Systems (ASPLOS)* (2016), pp. 217–232. https://doi.org/10.1145/2872362.2872414
4. C. Bienia, S. Kumar, J.P. Singh, K. Li, The PARSEC benchmark suite: characterization and architectural implications, in *Parallel Architectures and Compilation Techniques (PACT)* (2008), pp. 72–81
5. BSIM-CMG Model. http://bsim.berkeley.edu/models/bsimcmg
6. T.E. Carlson, W. Heirman, L. Eeckhout, Sniper: exploring the level of abstraction for scalable and accurate parallel multi-core simulation, in *High Performance Computing, Networking, Storage and Analysis (SC)* (ACM, New York, 2011), p. 52
7. R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, A.R. LeBlanc, Design of ion-implanted MOSFET's with very small physical dimensions. IEEE J. Solid State Circuits **9**(5), 256–268 (1974)
8. L.B. Kish, End of Moore's law: thermal (noise) death of integration in micro and nano electronics. Phys. Lett. A **305**(3), 144–149 (2002)
9. Z. Krivokapic, U. Rana1, R. Galatage, A. Razavieh, A. Aziz, J. Liu, J. Shi, H. Kim, R. Sporer, C. Serrao, A. Busquet, P. Polakowski, J. Müller, W. Kleemeier, A. Jacob1, D. Brown, A. Knorr, R. Carter, S. Banna, 14 nm ferroelectric FinFET technology with steep subthreshold slope for ultra low power applications, in *IEEE International Electron Devices Meeting (IEDM)* (2017), pp. 15.1.1–15.1.4
10. D. Kwon, K. Chatterjee, A.J. Tan, A.K. Yadav, H. Zhou, A.B. Sachid, R.D. Reis, C. Hu, S. Salahuddin, Improved subthreshold swing and short channel effect in FDSOI n-channel negative capacitance field effect transistors. IEEE Electron Device Lett. **39**(2), 300–303 (2018). https://doi.org/10.1109/LED.2017.2787063
11. S. Li, J.H. Ahn, R.D. Strong, J.B. Brockman, D.M.Tullsen, N.P. Jouppi, The McPAT framework for multicore and manycore architectures: simultaneously modeling power, area, and timing. Trans Archit. Code Optim. (TACO) **10**(1), 5 (2013)
12. J. Müller, T.S. Boscke, U. Schröder, S. Mueller, D. Bräuhaus, U. Böttger, L. Frey, T. Mikolajick, Ferroelectricity in simple binary ZrO_2 and HfO_2. Nano Lett. **12**(8), 4318–4323 (2012). https://doi.org/10.1021/nl302049k
13. G. Pahwa, T. Dutta, A. Agarwal, Y.S. Chauhan, 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), pp. 49–54
14. G. Pahwa, T. Dutta, A. Agarwal, S. Khandelwal, S. Salahuddin, C. Hu, Y.S. Chauhan, Analysis and compact modeling of negative capacitance transistor with high ON-current and negative output differential resistance—Part II: model validation. IEEE Trans. Electron Devices **63**(12), 4986–4992 (2016)
15. M. Rapp, S. Salamin, H. Amrouch, G. Pahwa, Y. S. Chauhan, J. Henkel: Performance, power and cooling trade-offs with NCFET-based many-cores, in *Design Automation Conference (DAC)* (2019)
16. S. Salahuddin, S. Datta, Use of negative capacitance to provide voltage amplification for low power nanoscale devices. Nano Lett. **8**(2), 405–410 (2008). https://doi.org/10.1021/nl071804g
17. S. Salamin, M. Rapp, H. Amrouch, G. Pahwa, Y. S. Chauhan, J. Henkel, NCFET-Aware voltage scaling, in *The International Symposium on Low Power Electronics and Design (ISLPED)* (2019)
18. S. Salamin, V.M. van Santen, H. Amrouch, N. Parihar, S. Mahapatra, J. Henkel, Modeling the interdependences between voltage fluctuation and BTI aging. IEEE Trans. Very Large Scale Integr. VLSI Syst. **27**(7), 1652–1665 (2019)
19. S.K. Samal, S. Khandelwal, A.I. Khan, S. Salahuddin, C. Hu, S.K. Lim, Full chip power benefits with negative capacitance FETs, in *International Symposium Low Power Electronics and Design (ISLPED)* (2017)

20. V.M. van Santen, H. Amrouch, J. Henkel, Modeling and mitigating time-dependent variability from the physical level to the circuit level. IEEE Trans. Circuits Syst. I Regul. Pap. **66**(7), 2671–2684 (2019)
21. V.V. Zhirnov, R.K. Cavin, Nanoelectronics: negative capacitance to the rescue? Nat. Nanotechnol. **3**(2), 77–78 (2008)