Energy Optimization in NCFET-based Processors

Sami Salamin*, Martin Rapp*, Hussam Amrouch*, Andreas Gerstlauer‡, Jörg Henkel*
*Chair of Embedded Systems (CES), Karlsruhe Institute of Technology, Karlsruhe, Germany
‡Department of Electrical and Computer Engineering, University of Texas, Austin, USA
{sami.salamin, martin.rapp, amrouch, henkel}@kit.edu, gerstl@ece.utexas.edu

***Abstract*—Energy consumption is a key optimization goal for all modern processors. Negative Capacitance Field-Effect Transistors (NCFETs) are a leading emerging technology that promises outstanding performance in addition to better energy efficiency. Thickness of the additional ferroelectric layer, frequency, and voltage are the key parameters in NCFET technology that impact the power and frequency of processors. However, their joint impact on energy optimization has not been investigated yet.**

In this work, we are the first to demonstrate that conventional (i.e., NCFET-unaware) dynamic voltage/frequency scaling (DVFS) techniques to minimize energy are sub-optimal when applied to NCFET-based processors. We further demonstrate that state-of-the-art NCFET-aware voltage scaling for power minimization is also sub-optimal when it comes to energy. This work provides the first NCFET-aware DVFS technique that optimizes the processor's energy through optimal runtime frequency/voltage selection. In NCFETs, energy-optimal frequency and voltage are dependent on the workload and technology parameters. Our NCFET-aware DVFS technique considers these effects to perform optimal voltage/frequency selection at runtime depending on workload characteristics. Results show up to 90 % energy savings compared to conventional DVFS techniques. Compared to state-of-the-art NCFET-aware power management, our technique provides up to 72 % energy savings along with 3.7x higher performance.

I. Introduction

Minimizing the energy consumption of a processor is the primary concern in many applications [1]. The energy consumption of any processor depends on its operating frequency (F) and operating voltage (V) as well as on the total execution time of the running workload. Energy consumption for executing a given workload is minimized by carefully selecting V/F pairs to exploit these dependencies. Because these dependencies vary among different technologies, energy optimization techniques should be aware of new technology.

Negative Capacitance Field-Effect Transistors (NCFETs) are a promising emerging technology that provides a considerable improvement in a circuit's performance over conventional FinFETs. This is because NCFETs employ a ferroelectric layer (FL) within the gate stack of the transistor, which manifests itself as a Negative Capacitance (NC). The latter results in a voltage amplification at the internal gate of the transistor, which boosts the electric field. This, in turn, has two key implications [2]: (1) NCFET-based circuits can operate at a higher frequency at the same operating voltage (V), (2) NCFET-based circuits can operate at the same frequency but at lower operating voltage leading to considerable power savings.

Power and performance of NCFET-based processors: The energy consumption of a processor is the integral of the power consumption over the total execution time. Prior work has shown that NCFET-based processors exhibit an observable performance enhancement compared to FinFETs due to voltage amplification. Fig. 1(a) shows how the maximum frequency of a processor at given V increases when a thicker FL is employed. FL thickness is referred to as TFEx, where x is the layer thickness in nanometer. TFE0 refers to conventional FinFET technology with out FL.

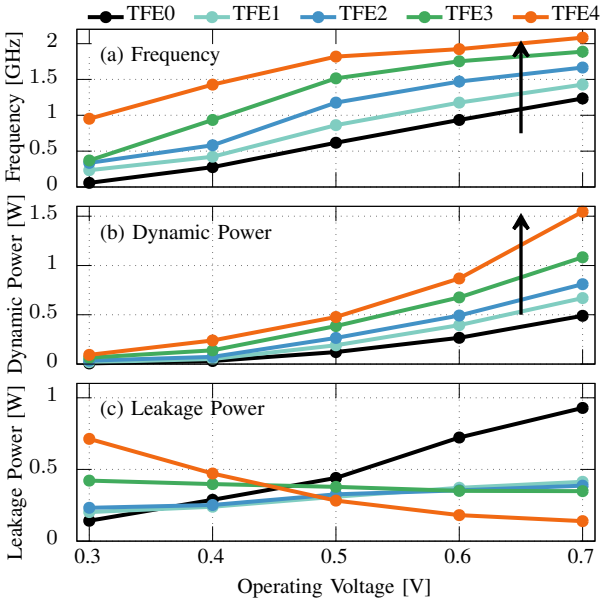

Fig. 1: (a) NCFET boosts the maximum frequency of the processor at given voltage. Gains increase with a thicker ferroelectric layer (FL). (b) NCFET increases the dynamic power due to the increase in the frequency and gate capacitance of the transistor. (c) NCFET with a thin FL weakens the dependency of leakage on voltage. At higher thicknesses, the leakage dependency is reversed [2].

NC increases the total gate capacitance of FinFETs, together with increased frequency, results in a higher dynamic power at the same operating voltage (Fig. 1(b)). Importantly, increasing the thickness of the FL inverses the dependency of leakage power on V due to the negative drain-induced barrier lowering effect (DIBL) [3], as shown in Fig. 1(c). Therefore, reducing V *increases* the leakage power, instead of decreases as in conventional FinFET. This has a far-reaching impact when it comes to any DVFS-based energy optimization scheme.

Workload dependency: Total power consumption is the sum of dynamic and leakage power. Fig. 1 demonstrates that dynamic and leakage power are differently affected by changes in the voltage and FL thickness. Different workloads have different runtime activities and hence different dynamic to leakage power ratios. *Therefore, the characteristics of the running workloads need to be considered when selecting the FL thickness, voltage and frequency in order to optimize the processor's energy.*

Energy minimization with NCFET: Fig. 2 shows the power consumption of the slave thread of the *PARSEC dedup* benchmark [4] for different frequencies and FL thickness. The

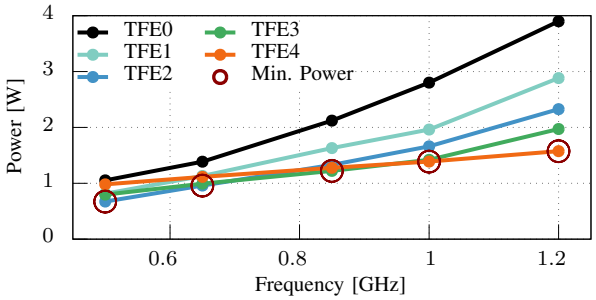

Fig. 2: Total power consumption of the *PARSEC dedup* benchmark depends on the frequency and thickness of the FL. V is selected differently for every combination of thickness TFEx and frequency to sustain the required frequency. Different thicknesses are optimal (minimum power) at different frequencies, showing the importance of selecting the optimal thickness. NCFET weakens the increase of power with frequency, necessitating to revisit the frequency selection in order to minimize energy consumption.

operating voltage at every pair of frequency and FL thickness is selected according to Fig. 1(a) to the minimum voltage that sustains the given frequency.

Fig. 2 gives several key insights into energy minimization in NCFETs. Firstly, it shows the importance of selecting the optimal thickness of the FL. At high frequencies, TFE4 results in the lowest power consumption. The reason is that dynamic power is high but TFE4 suppresses it the most. By contrast, at low frequencies, dynamic power decreases rapidly and therefore leakage becomes more dominant. This is the reason why TFE2 is optimal in this example. *The selection of the FL thickness at design-time strongly affects energy consumption.*

Secondly, the results also confirm the well-known fact that the power consumption in conventional FinFETs (i.e., TFE0) increases stronger than linearly with frequency. Therefore, despite the decrease in runtime, increasing the frequency increases the energy for executing a fixed workload. This leads to a well-known trade-off between energy and performance, where the lowest voltage and frequency levels minimize the total energy of a conventional FinFET processor. However, the trends are different in NCFETs. A thicker FL weakens the power increase with increased frequency. This, in turn, weakens the energy-performance trade-off, such that higher frequencies can potentially lead to lower energy due to their shorter execution time and hence leakage duration (where leakage power itself is potentially reduced at higher voltages). *While a processor's energy is always minimized at the lowest voltage/frequency in conventional FinFET, this does not hold anymore in NCFET. Hence, developing new NCFET-aware energy optimization techniques is indispensable.*

In this work, we present the first energy optimization technique for NCFET-based processors. Our approach models the impact of frequency, voltage, workload characteristics and FL thickness on NCFET energy. Using these models, we present an optimization technique for DVFS operating points in NCFET processors.

Our novel contributions within this paper are as follows:
(1) We present an analytical energy model of NCFET-based processors. The model allows designers to explore the joint effects of voltage, frequency, workload characteristics and ferroelectric layer thickness on NCFET energy.
(2) We present an NCFET-aware DVFS technique for energy optimization that selects the optimal frequency/voltage pair at runtime considering the characteristics of the workloads.
(3) We explore the dependency of DVFS operating points and optimal energy on workloads and technology parameters.

II. Related Work

DVFS is used in almost all modern processors to minimize energy while meeting performance requirements. Conventional DVFS selects the minimum frequency and voltage required under *fixed* performance constraints. When it comes to the optimal energy point, many studies showed that operating processors at a near-threshold voltage achieves such a goal [5]. However, it leads to performance degradation.

Recently, few works explored NCFET processor design and optimization. [2], [6] presented a comparison between conventional FinFET and NCFET processors under different configurations (i.e., FL thicknesses). The study in [2] showed how NCFETs impact the performance, power and temperature of a processor. In [7], a dynamic voltage scaling (DVS) technique has been proposed to optimize the power consumption of NCFET many-core systems under *fixed performance constraint*. The work assumes a constant frequency and hence it only scales the voltage standalone. Furthermore, the work focused solely on power (not energy) minimization and it studied only single FL thickness.

III. NCFET-Aware Energy Models

We first present the application, power and frequency models that are used in this work. Later, we then present our NCFET-aware energy optimization technique.

1) Application Model: The optimal frequency (f_{opt}) is the frequency at which the processor's energy is minimized. $V_{min}(f_{opt})$ is the minimum voltage required to sustain f_{opt}. Note that the minimum energy could be achieved at a higher voltage than V_{min} which is required to sustain f_{opt}. Therefore, $V_{opt}(f_{opt})$ is the optimal voltage for operating at f_{opt} [7].

To simplify the application model, we assume that the performance is linearly affected by frequency. We use the *ratio of dynamic to total power* that a workload exhibits at the highest thickness at the common highest frequency ($\hat{f}$) among all thicknesses (i.e., TFE4 at 1.2GHz) in order to represent a workload. By sweeping this ratio, we explore a large variety of workload domains from memory-bound to compute-bound applications. We assume a single thread is being executed on a single core under a fixed amount of work (W).

2) Power and Frequency Models: To characterize the power and frequency models we follow the same methodology as in [2]. A full SoC [8] is designed entirely from RTL to layout using our NCFET cell libraries [9]. We then use commercial signoff tools to analyze the power and frequency of the full SoC. Finally, and similar to [7], we fitted the results into mathematical equations to use them in our models.

The minimum voltage $V_{min}^{(x)}(f_{min}^{(x)})$ at thickness x required to sustain $f_{min}^{(x)}$ is:

$$V_{min}^{(x)}(f_{min}^{(x)}) = \left(\frac{\frac{1}{f_{min}^{(x)}} - c_{freq}^{(x)}}{a_{freq}^{(x)}} \right)^{\frac{1}{b_{freq}^{(x)}}} \quad (1)$$

$$f_{min}^{(x)}(V_{min}^{(x)}) = \frac{1}{a_{freq}^{(x)} \cdot V_{min}^{b_{freq}^{(x)}} + c_{freq}^{(x)}}, \quad (2)$$

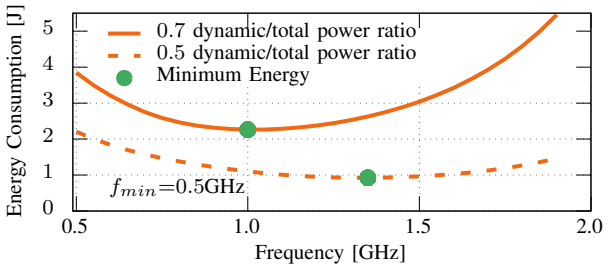

Fig. 3: Energy consumption over frequency of two workloads running on a processor designed in TFE4 and operated at $V_{opt}(f)$. The minimum energy does not appear at f_{min}, but instead at a higher frequency f_{opt}. As the two workloads have different dynamic/total power ratios, the minimum energy appears at different f_{opt}.

where $a^{(x)}_{freq}$, $b^{(x)}_{freq}$, $c^{(x)}_{freq}$ are constant fitting parameters. Minimum leakage and minimum dynamic power when operating at $f^{(x)}_{min}/V^{(x)}_{min}$ are:

$$P^{(x)}_{leak}(V^{(x)}_{min}) = a^{(x)}_{leak} \cdot V^{b^{(x)}_{leak}}_{min} \quad (3)$$

$$P^{(x)}_{dyn,min}(V^{(x)}_{min}) = a^{(x)}_{dyn} \cdot V^{b^{(x)}_{dyn}}_{min} + c^{(x)}_{dyn}. \quad (4)$$

Here, $a^{(x)}_{dyn}$, $b^{(x)}_{dyn}$, $c^{(x)}_{dyn}$, $a^{(x)}_{leak}$, $b^{(x)}_{leak}$ are constant fitting parameters. By operating at a frequency higher than $f^{(x)}_{min}$, dynamic power is scaled linearly.

$$P^{(x)}_{dyn}(V,f) = \frac{f}{f^{(x)}_{min}} \cdot P^{(x)}_{dyn,min}(V^{(x)}_{min}) \quad (5)$$

3) Workload-Dependence and Energy Modeling: Dynamic power consumption $P^{(x)}_{dyn}(V,f)$ is affected by the *running workload*, which is scaled by a factor $r_{dyn}\geq 0$ from the minimum dynamic power $P^{(x)}_{dyn,min}(V^{(x)}_{min})$:

$$P^{(x)}_{dyn}(V,f) = r_{dyn} \cdot P^{(x)}_{dyn,min}(V^{(x)}_{min}) \quad (6)$$

$$P^{(x)}_{total}(V,f) = P^{(x)}_{dyn}(V,f) + P^{(x)}_{leak}(V) \quad (7)$$

r_{dyn} is not constant since it represents the current workload activity that depends on the dynamic/total power ratio as a variable. We define the dynamic/total power ratio as the r_{dyn} observed at $P^{(4)}_{dyn,min}$, which is the peak dynamic power at TFE4 and $\hat{f}$ as shown in Eq. (8):

$$dyn/tot = \frac{r_{dyn} \cdot P^{(4)}_{dyn,min}(V^{(4)}_{min}(\hat{f})}{P^{(4)}_{dyn,min}(V^{(4)}_{min}(\hat{f})) + P^{(4)}_{leak}(V^{(4)}_{min})} \quad (8)$$

$$r_{dyn} = \frac{dyn/tot \cdot P^{(4)}_{leak}(V^{(4)}_{min}(\hat{f}))}{(1-dyn/tot) \cdot P^{(4)}_{dyn,peak}(V^{(4)}_{min}(\hat{f}))}. \quad (9)$$

Therefore, the total energy is:

$$E^{(x)}_{total}(V,f) = (P^{(x)}_{dyn}(V,f) + P^{(x)}_{leak}(V)) \cdot \frac{W}{f} \quad (10)$$

4) Optimal Frequency/Voltage Selection: V_{opt} and f_{opt} that minimize total energy can be obtained from the energy model in the form of a minimization problem:

$$V_{opt}(f, r_{dyn}) = \underset{V_{min}(f)\leq V\leq V_{max}}{\arg\min} E^{(x)}_{total}(V,f) \quad (11)$$

$$f_{opt}(r_{dyn}) = \underset{f_{min}\leq f\leq f_{max}}{\arg\min} E^{(x)}_{total}(V_{opt}(f, r_{dyn}), f) \quad (12)$$

DVFS selection is, therefore, an optimization problem that can be solved by exploring the design space of $E^{(x)}_{total}(V,f)$.

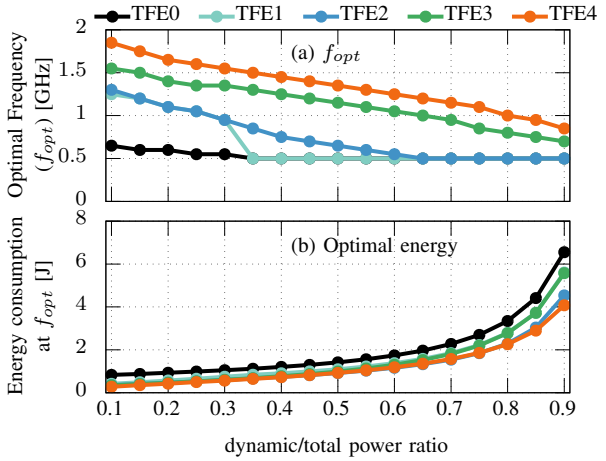

Fig. 4: (a) Optimal frequency selected by our technique over dynamic/total ratio that minimizes energy for thicknesses TFEx using W=10^6. (b) Optimal energy over dynamic/total power ratios for different TFEx operating at optimal frequency f_{opt} from (a).

Solving Eq. (12) using two different workloads on TFE4 processor results in curves shown in Fig. 3. Following a conventional technique, the processor would run at f_{min}/V_{opt} to minimize energy. However, increasing the frequency further increases the operating voltage. This will increase the dynamic energy, but stronger decreases the leakage energy and hence the total energy decreases. This will continue until an inflection point appears where the dynamic energy becomes prominent and therefore increasing the frequency further increases the total energy. At this point, f_{opt} is observed. Importantly, it shows how two applications have different optimal frequencies.

IV. Exploration and Optimization

In the following, we present our NCFET-aware DVFS technique for energy optimization. We then perform a design space exploration to determine the impact of FL thickness on optimal energy as a function of workload parameters.

1) Frequency and Voltage Selection: f_{opt}/V_{opt} selection following Eq. (12) is an optimization problem that can be solved using a search algorithm by sweeping across all possible frequency and voltage steps to minimize energy. We then examine how the optimal frequency that minimizes energy using our technique depends on possible workload characteristics. To cover a wide range of workloads, we examine dynamic/total power ratios in the range of 0.1-0.9 for W=10^6 cycles. The optimal frequencies are shown in Fig. 4(a). Results show that TFE4 exhibits the best performance (i.e., highest frequency) over all thicknesses.

2) Thickness Exploration: Using the optimal frequencies from Fig. 4(a), we can now examine the dependency of FL thickness on the minimum energies. Minimum energy results for different thicknesses and application characteristics are shown in Fig. 4(b). The energy of TFE4 is always the minimum among all thicknesses. However, the preference is for TFE4 as it shows the best performance (see Fig. 4(a)) in addition to the minimum energy. As a result, *TFE4* shows the optimal energy and best performance (i.e., higher f_{opt}) among all thicknesses.

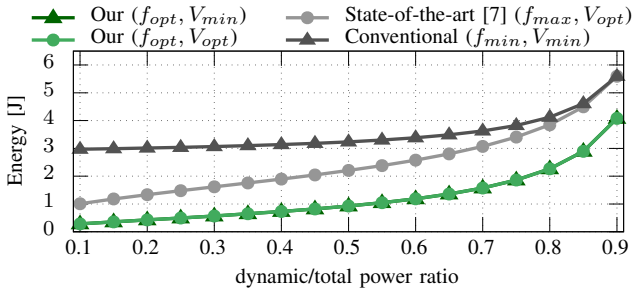

Fig. 5: Optimal energy of TFE4 over dynamic/total power ratio of the four used scenarios. Our scenarios, operating at f_{opt} regardless of voltage, show the minimal energy among all cases. The conventional technique using f_{min} is the worst scenario. A state-of-the-art [7] approach selecting V_{opt} to achieve a trade-off between leakage and dynamic power when operating at f_{min} is sub-optimal.

V. Evaluation and Comparisons

In the following, we examine the achievable energy savings using our NCFET-aware frequency and voltage selection in comparison with conventional DVFS and state-of-the-art. As shown previously, TFE4 shows the minimum energy over all thickness at f_{opt}. TFE4 also shows the highest frequency over all thicknesses (i.e., best performance). Therefore, we will only show the energy savings for TFE4.

We examine the energy of TFE4 for different scenarios: (1) *NCFET-aware voltage and frequency selection (our)*: the processor operates at f_{opt} with $V_{opt}(f_{opt})$ selected using the technique published in [7]. (2) *NCFET-aware frequency selection (our)*: the processor operates at f_{opt} using the $V_{min}(f_{opt})$ required to sustain that frequency. (3) NCFET-aware voltage selection (state of the art) [7]: the processor operates at f_{min}, which is required to meet performance goal, and $V_{opt}(f_{min})$ that minimizes the power consumption at f_{min}. (4) Conventional DVFS technique where the processor operates at f_{min} required to meet a performance goal and V_{min} required to sustain that frequency.

Energy Savings with NCFET-Aware DVFS: The results of the four scenarios are demonstrated in Fig. 5, showing the energies over dynamic/total power ratios. Results show that our scenarios (1) and (2) (i.e., f_{opt}) result in the minimum energy regardless of voltage. The two scenarios have exactly the same energy as results show that empirically, V_{min}=V_{opt} at f_{opt}. *This shows that frequency selection is more important than voltage selection for minimizing energy in NCFETs.*

Moreover, results compared to scenario (3) [7] highlight the importance of selecting the optimal frequency. Our scenarios are orthogonal to scenario (3) as [7] targets minimum power under fixed performance while we target minimum energy. Crucially, our results show that, depending on the workload, minimal energy is potentially achieved at a higher frequency than any performance constraint would require. In other words, even optimal power management may necessitate more complex frequency optimizations than investigated in [7]. The energy savings using our optimization over state-of-the-art can reach up to 72 %.

Finally, the conventional scenario (4) shows the highest energy consumption among all cases for all dynamic/total power ratios as it is completely NCFET-unaware. This highlights, again, that existing power management techniques cannot be used for NCFET-based processors. Instead, new NCFET-aware technique need to be developed, which we present in this work. The energy gains using our technique compared to a conventional DVFS can reach up to 90 %.

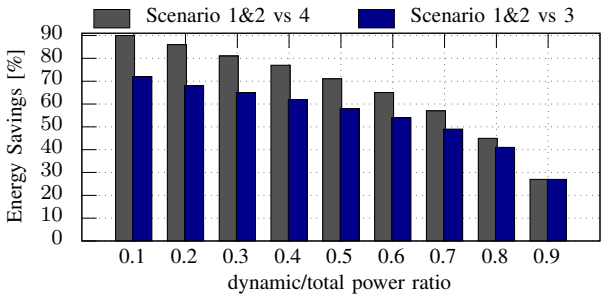

Fig. 6: Energy savings by operating at f_{opt} selected in our scenarios (1&2) in comparison with conventional technique (4) and state-of-the-art (3). Saving is up to 90 % and up to 72 % compared to conventional and state-of-the-art [7] scenarios, respectively.

Energy savings results are summarized in Fig. 6. A state-of-the-art scenario results in higher savings than a conventional DVFS approach, as the state-of-the-art is NCFET-aware albeit for voltage selection only.

VI. Conclusions

NCFETs are a promising emerging technology that provides outstanding performance in addition to better power optimization compared to conventional FinFET technology. As conventional energy minimization techniques are unaware of the inverse dependency that leakage power exhibits in NCFETs, they become sub-optimal. In this work, we presented the first NCFET-aware DVFS technique to optimize the energy of NCFET-based processors. We showed how optimal frequency and voltage can be selected. The optimal frequency to achieve minimal energy is larger than the minimum frequency. The largest FL thickness provides both the best energy and performance. Our analysis further demonstrated a design space for selecting the optimal operating frequency f_{opt} and voltage V_{opt} to minimize energy based on thickness and application characteristics. Compared to conventional DVFS techniques, our approach results in up to 90 % and up to 72 % energy savings compared to conventional and state-of-the-art NCFET-aware voltage scaling, respectively.

References

[1] J. Lee, Y. Zhang *et al.*, "19.2 a 6.4pj/cycle self-tuning cortex-m0 iot processor based on leakage-ratio measurement for energy-optimal operation across wide-range pvt variation," in *ISSCC*, Feb 2019.

[2] M. Rapp, S. Salamin *et al.*, "Performance, Power and Cooling Trade-Offs with NCFET-based Many-Cores," *DAC*, 2019.

[3] G. Pahwa, T. Dutta *et al.*, "Designing energy efficient and hysteresis free negative capacitance FinFET with negative DIBL and 3.5x IONusing compact modeling approach," in *ESSDERC*, Sep. 2016.

[4] C. Bienia, S. Kumar *et al.*, "The PARSEC Benchmark Suite: Characterization and Architectural Implications," in *PACT*, 2008.

[5] S. Salamin, H. Amrouch *et al.*, "Selecting the optimal energy point in near-threshold computing," in *DATE*, March 2019.

[6] S. K. Samal, S. Khandelwal *et al.*, "Full chip power benefits with negative capacitance FETs," in *ISLPED*, July 2017.

[7] S. Salamin, M. Rapp *et al.*, "NCFET-Aware Voltage Scaling," *ISLPED*, pp. 1–6, July 2019.

[8] J. Balkind, M. McKeown *et al.*, "OpenPiton: An Open Source Manycore Research Framework," in *ASPLOS*, 2016.

[9] H. Amrouch, G. Pahwa *et al.*, "Negative Capacitance Transistor to Address the Fundamental Limitations in Technology Scaling: Processor Performance," *IEEE Access*, vol. 6, 2018.