This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

Impact of NCFET Technology on Eliminating the Cooling Cost and Boosting the Efficiency of Google TPU

Sami Salamin, *Student, IEEE*, Georgios Zervakis, Florian Klemme, Hammam Kattan, Yogesh Chauhan, *Fellow, IEEE*, Jörg Henkel, *Fellow, IEEE*, and Hussam Amrouch, *Member, IEEE*

Abstract—Recent breakthroughs in Neural Networks (NNs) led to significant accuracy improvements of several machine learning applications such as image classification and voice recognition. However, this accuracy improvement comes at the cost of an immense increase in computation demands. NNs became one of the most common and computationally intensive workloads in today's datacenters. To address these computational demands, Google announced in 2016 the Tensor Processing Unit (TPU), an advanced custom ASIC accelerator for NN inference. Two new TPU versions (v2 and v3) followed in 2017 and 2018 that support also training. Google TPUv3 packs an immense processing power ($90\mathrm{TFLOPS}$ per chip) in a tiny and condensed area, leading to very high on-chip power densities and thus excessive temperature. In this work, superlattice thermoelectric cooling, which is one of the emerging on-chip cooling, is considered as an advanced cooling example for Google TPU and we investigate the impact of Negative Capacitance FET (NCFET), which is one of the recent emerging technologies, on the cooling and efficiency of TPU. Through full-chip design, of the computational core of the TPU, based on $14\mathrm{nm}$ Intel FinFET technology and multiphysics temperature simulations, we demonstrate that NCFET can significantly minimize the required cooling-cost. More than 4000 NCFET configurations are evaluated in order to traverse the entire design space defined by the thickness of the ferroelectric layer of NCFET, the operating voltage, cooling, and the operating frequency, in addition to all possible FinFET's configurations. Moreover, our experimental evaluation shows that by eliminating the cooling cost, NCFET delivers 2.8x higher efficiency compared to the conventional FinFET baseline.

Index Terms—Negative Capacitance Transistor (NCFET), Thermoelectric Cooling, Google TPU, Multiphysics, Neural Processing Unit

1 Introduction

Due to the ever-increasing need to accelerate the training and inference of advanced Neural Networks (NNs), ASIC hardware accelerators have become an integral part of modern computing systems. Google announced in 2016, how their customized hardware called Tensor Processing Unit (TPU) significantly improves the inference throughput of a wide range of NNs compared to traditional approaches typically done in GPUs and/or CPUs. In 2017 and 2018, Google announced the second and third generation of TPU that leverage the bfloat16 number format to support also training. The essential part of TPUv3 is four gigantic systolic multiply-accumulate arrays (namely MXU) that substantially speed up the NN training and inference due to the massive number of multiply-accumulations performed in parallel. Each MXU comprises 16K multiply-accumulate units (128×128 MAC units) and is able to deliver a raw throughput of $22.5\mathrm{TFLOPS}$, which results in a very large amount of power to be consumed in a relatively small silicon footprint. Such high performance in a compact area leads to immense power density and temperature and thus, a vast cooling cost. Compared to TPUv2, in order to maximize the throughput, in TPUv3, Google doubled i) the number of MXUs per TPU chip, ii) the number of TPU's per rack and iii) the number of racks per TPU pod, exacerbating the cooling problem.

The power density (PD) of each MXU can exceed $232\,\mathrm{W/cm^2}$ at the maximum frequency of $787\mathrm{MHz}$. To put this number into perspective, we present in Figure 1 the power density (i.e., heat flux) of the MXU, PULPino [1], and OpenPiton [2]. PULPino is an open-source single-core 32-bit RISC-V architecture with 4-stage pipeline designed mainly for low power applications [1], while OpenPiton is an open-source multi-threaded manycore 64-bit SPARC v9 architecture with 6-stage pipeline designed as general-purpose processor [2]. The PD values in Figure 1 are obtained after full-chip design of the examined architectures at the Intel $14\mathrm{nm}$ FinFET technology following the same methodology as in Section 3.2, and considering operation at their maximum clock frequency. MXU delivers a vast computational power in a very condensed area. As a result, as shown in Figure 1, compared to the OpenPiton and PULPino processors the power density of MXU is 1.6x and 4.0x higher, respectively. Hence, localized hot-spots can rapidly emerge during operation due to excessive on-chip power densities. To avoid unsustainable on-chip tempera-

- *Sami Salamin, Georgios Zervakis, and Jörg Henkel are with Chair for Embedded System (CES), Karlsruhe Institute of Technology (KIT), Karlsruhe 76131, Germany. Hammam Kattan was with CES, KIT from 2018-2020. E-mail: hammam.kattan@hotmail.com, {sami.salamin, georgios.zervakis, henkel}@kit.edu. Salamin and Zervakis have equal contributions.*
- *Yogesh Chauhan is with Electrical Engineer Department, Indian Institute of Technology Kanpur (IITK), Kanpur, India. Email: chauhan@iitk.ac.in.*
- *Florian Klemme and Hussam Amrouch are with the Chair for Semiconductor Test and Reliability (STAR) in the Computer Science, Electrical Engineering Faculty at the University of Stuttgart. E-mail: {klemme, amrouch}@iti.uni-stuttgart.de. The work of F. Klemme and H. Amrouch were done in part at KIT.*

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

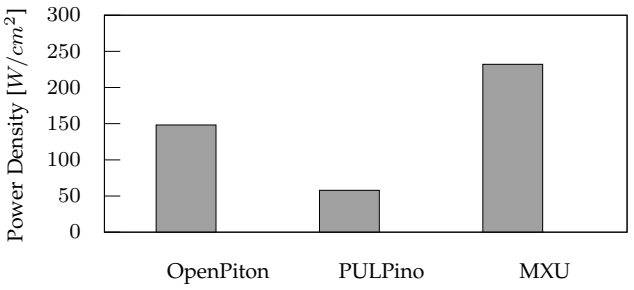

Fig. 1: Power density of the MXU the core of TPUv3, PULPino [1], and OpenPiton [2]. The power density of MXU is **4.0**x higher than the low-power processor PULPino and 1.6x higher than the general-purpose processor OpenPiton. All architectures are designed at 14nm with 0.8V operating voltage. PULPino operates at 763MHz, OpenPiton at 1050MHz, and MXU at 787MHz.

tures, advanced cooling solutions are required and/or the operating frequency of the MXU needs to be scaled down (leading to considerable throughput loss).

The inability of conventional air cooling to dissipate generated heat under the excessive power densities beyond $150\,\mathrm{W/cm^2}$ [3], forced Google to re-engineer the TPU pods and to switch from conventional air cooling in TPUv2 to liquid cooling in TPUv3 [4]. This was obviously so important, that Google used liquid cooling for the first time, due to the excessive on-chip power densities of TPUv3. As the Alphabet CEO announced in his keynote speech: "These chips are so powerful, that for the first time we have had to introduce liquid cooling in our data centers" [5].

The liquid cooling comes with many liquid channels embedded on the surface of the chip with a high thermal conductivity that link on-chip hot-spots to conduct the heat away from the chip [6]. Although the liquid cooling is proven to improve the cooling efficiency compared to conventional air cooling, it has a considerable size, induces significant infrastructure cost, and consumes very high power [7] [8]. The liquid cooling system consumes more than 18% of the overall power in typical datacenters, and 7% of the overall power for air movement for cooling purpose as well, in addition to billions of gallons of water [4], [9].

Recently, many advanced emerging cooling technologies have been proposed to improve cooling efficiency and reduce cooling cost. One of the most promising on-chip cooling technique is Superlattice Thermoelectric cooling (TEC).

Superlattice Thermoelectric cooling (TEC): On-chip cooling in which an ultra-thin layer of a superlattice thermoelectric device is integrated within the chip's packaging itself is one of the very promising advanced techniques for future on-chip cooling. It offers a strong, yet efficient heat dissipation capability to suppress localized on-chip hot-spots. Mature prototypes [10], [11] demonstrated amazing effectiveness in which heat fluxes up to $1300\,\mathrm{W/cm^2}$ can be managed. This brings superlattice thermoelectric devices at the forefront of cooling solutions to be employed in the TPU in order to remove the thermal bottleneck that the MXU creates.

TEC exhibits several advantages in addition to its high cooling efficiency compared to the conventional approaches (e.g., air- or liquid- cooling). Briefly, TEC: 1) features no moving parts such as fans or pumps and hence, no mechanical vibration, no sound, or even mechanical wear out, 2) has low/no maintenance cost, 3) is easy to control as it is current-controlled, and 4) is also flexible in shape and can be tailored for very small areas (less space) [12] [13]. All the aforementioned advantages and features make TEC a very promising option for datacenters and therefore we consider it in our evaluation as the cooling solution for TPU.

The advanced emerging cooling technology allows modern chips to have better cooling efficiency. However, emerging technology nodes can also help to improve the cooling efficiency of the chip.

Negative Capacitance Transistor as an Emerging Cooling Solution: Negative Capacitance Field-Effect Transistor (NCFET) is one of the leading technologies that might replace the existing CMOS technology in the near future. Compared to other emerging technologies, NCFET is fully compatible with the existing CMOS fabrication process as was successfully demonstrated by GlobalFoundries when they implemented the first NCFET-based circuit using their mature commercial 14nm technology node [14]. NCFET replaces the conventional high-κ material within the transistor's gate stack with a thin ferroelectric (FE) material layer. The intrinsic polarization inside the FE layer amplifies the vertical electric field in the transistor and thus it leads to higher switching speed without the need to increase the applied voltage. In practice, the FE layer manifests itself as a negative capacitance and hence an internal voltage amplification, instead of an internal voltage drop as is the case in all MOSFET types, is created. This, in turn, enables the sub-threshold swing of the transistor to go beyond its fundamental limit of $60\mathrm{mV/decade}$ [15]. As a result, NCFET transistor can achieve the same ON current as in the counterpart (conventional) MOSFET transistor but at a lower voltage. In other words, circuits implemented by the NCFET technology can be operated at the same clock frequency but at a much lower operating voltage. Hence, significant power savings (due to reductions in static and dynamic power) are accomplished without any reduction in the clock frequency, i.e., no trade-offs. Therefore, implementing the TPU using NCFET technology mitigates the excessive on-chip power densities that do exist at the MXU and thus elevated on-chip hot-spots can be removed. Importantly, NCFET does not come with an area overhead as the transistor size (length and width) remain exactly the same. This is because the increase in the thickness does not contribute to the area footprint that each FinFET occupies and only the fin height increases [16].

In this work, we investigate, for the first time, the new trade-offs that Negative Capacitance Transistor (NCFET) brings for the future of Tensor Processing Units (TPUs) in the scope of the role that advanced on-chip cooling plays. To achieve this, we perform an extensive design space evaluation that is built upon our most-accurate modeling of NCFET, temperature, and cooling cost using industry-strength tools and multiphysics simulations. In our analysis, we evaluate more than 4000 NCFET and 500 baseline configurations defined by the input current for TEC, voltage value, clock frequency, and size of the ferroelectric layer (only for NCFET).

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

Our contributions in this work are summarized as follows:
(1) Extensive design-space exploration demonstrating the role that NCFET technology may play in reducing or even completely eliminating cooling costs for future TPUs.
(2) To achieve that, 14nm NC-FinFET cell libraries are created after careful calibrations with Intel measurements (concerning the underlying FinFET device) and real experimentally-measured S shaped polarization-electric field curve (concerning the ferroelectric layer).
(3) We present a comprehensive evaluation of the efficiency of Google TPUv3 (i.e., MXU) comparing NC-FinFET with its counterpart Intel 14nm FinFET technology. Superlattice Thermoelectric (TEC) is considered in this work as an example of an advanced cooling solution due to its promises in effective heat dissipation.
(4) We show that by eliminating the cooling cost, NCFET boosts the efficiency[1] of MXU by 2.8x compared to the conventional FinFET.

2 Related Works

The related work can be clustered in two groups: prior works on advanced cooling and prior works on NCFET.

Advanced cooling: Increasing processing power has also increased the amount of energy required for high-end chips. Due to the inability of conventional cooling to dissipate the increasingly on-chip generated temperature, many emerging technologies have been proposed to improve the on-chip cooling. In [7], an implementation of liquid cooling in datacenter has been presented where a cooling energy reduction by over 90% can be achieved. Moreover, the authors also presented a new two phases liquid chip design where the liquid can pass into the chip for better cooling. In [8], authors present different liquid cooling models for thermal management of three-dimensional integrated circuits where many models can efficiently improve the on-chip cooling. In [10], authors show how thermoelectric modules can enhance heat dissipation and reduce the temperatures of electronic chips. They show different types of superlattice to manage excessive power density effectively. In [11], authors present an integration of thermoelectric coolers designed from nanostructured thin-film superlattices into state-of-the-art electronic chip packages, showing how thermoelectric devices can enable energy-efficient solutions. However, all the previously mentioned studies are proposed to improve the cooling capability of the already existing technology and in many cases extra cost and complexity are added [8]. Recently, in [17], authors present a hybrid thermal management technique for neural processing units (NPU) which efficiently employs TEC, precision and frequency scaling to trade-off temperature, power, throughput, and inference accuracy. However, the authors in [17] target only NN inference accelerators. The latter feature signifficantly lower power density compared to the ones we examine in this work that can be used for both trainning and inference.

NCFET: NCFET is an emerging technology that draws high attention from researchers and chip manufactures due to its ability to go beyond the sub-threshold swing limit and its fabrication compatibility with existing technology.

1. Efficiency is defined as $frequency \times \frac{Throughput}{Total\ Power}$

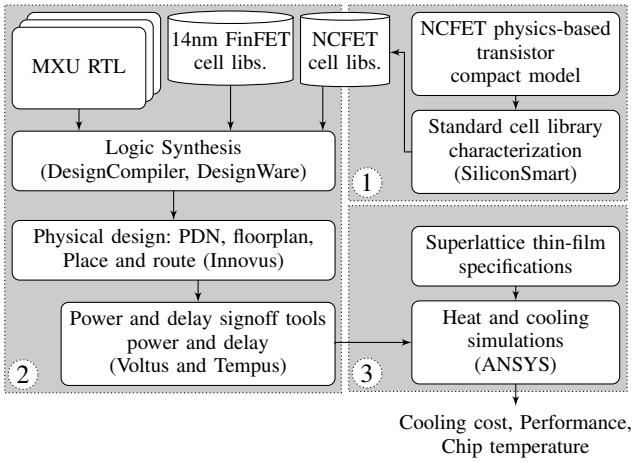

Fig. 2: Our employed holistic methodology to enable accurate modeling of NCFET and temperature and thus, obtain precise characterization of the impact of NCFET on the required cooling cost. MXU modeling (2) is a cross-layer implementation that links the physics- and circuit-level aspects that are involved in the design of an NCFET-based chip. The latter depends on NCFET modeling (1). Multiphysics simulations (3) are performed to obtain the chip temperature.

In [14], GlobalFoundries has announced the first-ever fabricated NCFET-based chip using 14nm FinFET technology. Recently, few works have been presented that explored NCFET chip design and optimization. [18], [19] presented a comparison between conventional FinFET and NCFET processors under different configurations (i.e., ferroelectric thicknesses). Both studies showed how NCFET impacts the performance and power of a processor's circuity. In [20], a dynamic voltage scaling (DVS) technique has been proposed to optimize the power consumption of NCFET many-core systems. In [21], an energy management technique has been presented aiming for minimizing the total energy by selecting the optimal combination of frequency and voltage. In [22], the authors apply approximate computing to mitigate the increased dynamic power of NCFET due to the increase in gate's capacitance and enable operation at high voltage values, further boosting, thus, the performance. Out of these works, only [21] tried to evaluate the impact of NCFET on the temperature and cooling cost. However, [18] uses an abstract model that was calibrated mainly for conventional technology based on total power consumption with conventional cooling. Importantly, the excessive on-chip temperature aggravates many reliability issues [23].

In our work, we leverage the high-accuracy of multi-physics simulation using the finite element method to precisely model the temperature and cooling in NCFET besides the state-of-the-art emerging on-chip cooling technology.

3 MXU, NCFET, and Cooling Modeling

Figure 2 demonstrates an overview of our employed workflow to evaluate the impact of NCFET on the cooling of the TPU. Our workflow is built upon the industry-strength tools to provide the most accurate modeling of NCFET, MXU frequency, and temperature/cooling cost. In this work, we

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

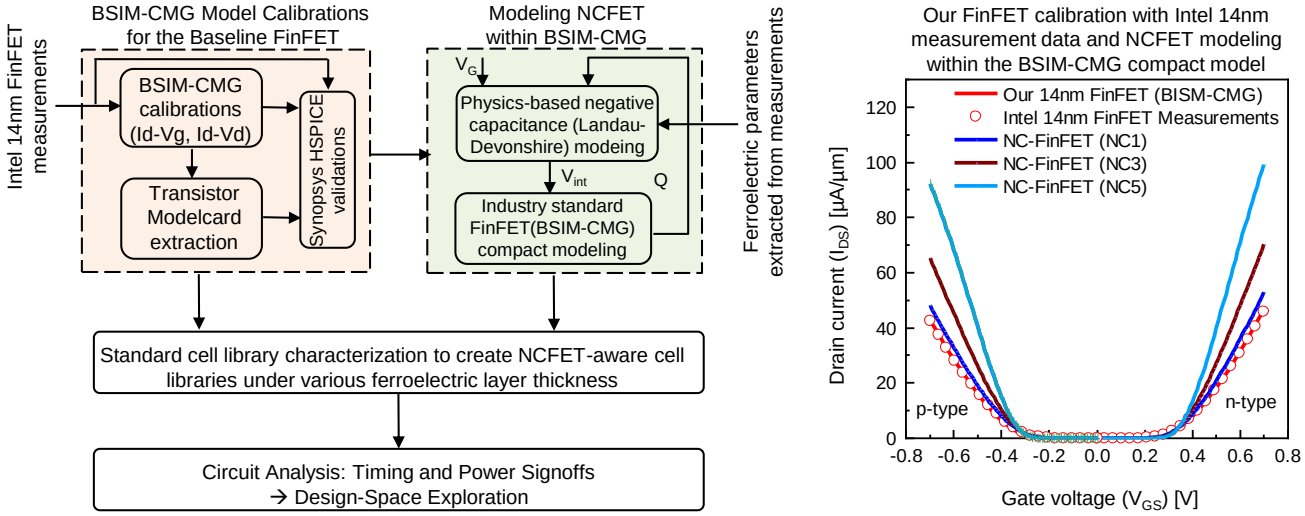

Fig. 3: The right side shows an overview of the NC-FinFET modeling from the device level to the circuit level. First the industry-compact model of FinFET technology (BSIM-CMG) [24] is calibrated to reproduce 14nm measurements from Intel [25]. Then, the ferroelectric physics is modeled using the phenomenological Landau-Devonshire (L-D) theory to accurately capture the effect of negative capacitance. The required parameters of ferroelectric modeling were extracted by fitting the Landau model to a real experimentally measured S shaped polarization-electric field curve [26]. Afterward, the BSIM-CMG along with the ferroelectric modeling are employed to perform SPICE simulations to create NCFET-aware standard cell libraries that are alter employed within commercial timing and power signoff tool flows to accurately analyze the impact of NCFET technology on the power, performance, energy, etc. and perform the required design-space exploration for the Google Tensor Processing Units (TPUs). The right side shows the validation of our used baseline FinFET n-type and p-type transistors against Intel 14nm FinFET measurements, showing an excellent matching (see the red curve and red symbols). The introduction of a ferroelectric layer inside the gate stack boosts the transistor current due to the internal voltage amplification. The obtained gain is proportional to the thickness of the ferroelectric layer. At a thickness of 5nm ($NC5$), the increase in ON current is around 2x, leading to a considerable increase in the transistor switching speed. Results of $NC1$, $NC3$, and $NC5$ are obtained from simulations using BSIM-CMG after integrating physics-based NC modeling.

rely on Intel 14nm FinFET [25] as our baseline technology node and we perform the full chip design (i.e., from RTL to GDSII level) of the MXU. Hence, we can perform an accurate power and frequency evaluation based on state-of-the-art technology. Post-layout gate-level simulations, using input traces obtained from the inference phase of the ResNet [27] network, are performed to obtain realistic switching activity values for the MXU. This enables us to perform accurate power analyses (that consider also the MXU utilization) and thus to obtain accurate power density evaluation. In the following, we present a detailed description of our NCFET, MXU, and Cooling models.

3.1 FinFET Calibration and NCFET Modeling

In this work, the underlying FinFET device is first calibrated with the Intel 14nm FinFET measurement data. Then, physics-based modeling for the Negative Capacitance (NC) effect is integrated within the industry-standard compact model of FinFET technology (BSIM-CMG). Validation of the used NC modeling is done using TCAD (Technology CAD) simulations. Next, we explain step by step how NC modeling is done along with several device-level analysis to demonstrate how the NC effect boosts the performance of nFinFET and pFinFET transistors. Further details on our NCFT modeling and how it is employed for NCFET-aware cell library characterization are available in [28].

As shown in Figure 3 (left), the NC-FinFET modeling, in practice, consists of two entities; the baseline constituent FinFET and the added ferroelectric layer inside the transistor gate stack. (1) The underlying FinFET transistor is modeled, using the industry-standard compact model for FinFET technologies (BSIM-CMG) [24], with experimentally-calibrated modelcard parameters of Intel 14nm FinFET technology. As shown in Figure 3 (right), our calibrated FinFET transistors have an excellent matching with Intel measurement data. (2) The ferroelectric physics is then modeled and realized using the phenomenological Landau-Devonshire (L-D) theory [29] in which the voltage across the ferroelectric (V_{fe}) is expressed as a function of its gate terminal charge (Q) as shown in (1). We have considered MFMIS structure for NCFET which is composed of FinFET with a ferroelectric layer in the gate stack. The parameters in the L-D model have been also calibrated with measured P-E loop. It is noteworthy that the NCFET compact model has been originally presented and validated in detail in our previous work in [30], [31].

$$V_{fe} = t_{fe}(2\alpha Q + 4\beta Q^3) \tag{1}$$

where α and β are the ferroelectric material ($Hf_{0.5}Zr_{0.5}O_2$) dependent parameters with α = -1.1×10^9 m/F and β = 2.5×10^{10} $m^5/(C^2F)$ [26]. t_{fe} refers to the thickness of ferroelectric (FE) layer and in this work we explore from 1nm to 5nm. Above which, a hysteresis-free operation which is a prerequisite for logic gates cannot anymore be ensured. Importantly, the ferroelectric parameters were extracted,

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

for the first time ever, by fitting the Landau model to a real experimentally measured S shaped polarization-electric field curve [32]. Afterwards, the L-D equation is solved *self-consistently* inside the BSIM-CMG model, constrained by the equality of terminal charges of the ferroelectric and the baseline gate. In our model, we have assumed a metal-ferroelectric-metal-insulator-semiconductor (MFMIS) structure for NC-FinFET in which a metal layer under the FE layer also exists. Note that our used HfO_2 material for FE layer is compatible with the existing CMOS fabrication process [33]. Various ferroelectric thickness (t_{fe}) are studied to enable the required design-space exploration presented in our evaluation (details in Section 4). Device-level analysis of NC-FinFET is later presented in Figure 5.

After we integrated all the model parameters, the L-D equation is then solved *self-consistently* within the industry-standard BSIM-CMG model, constrained by the equality of terminal charges of the ferroelectric and the baseline gate. It is noteworthy that in the context of integrating the physics-based model of NC effect within the BSIM-CMG, self-consistent manner means that the final obtained solution always satisfies both sets of equations; i.e., the NC equations as well as the equations of BSIM-CMG for the underlying FinFET. In this case, we ensure that all the existing interdependencies are always considered during SPICE simulations. In practice, the integration of NC model within the industry-standard compact model enabled us to efficiently perform standard cell library characterization, while we are still employing standard commercial tool flows. Note that only for simulation purpose, we assume a metal layer between the ferroelectric layer and the underlying baseline gate. The metal-ferroelectric-metal-insulator-semiconductor (MFMIS) for NC modeling in the absence of gate leakage and domain formation exhibits trends that are very similar to the metal-ferroelectric-insulator-semiconductor (MFIS) configuration, as demonstrated in [34], [35]. Despite using the MFMIS over the MFIS may change the absolute values obtained from simulations, the obtained trends remain the same for both (MFMIS and MFIS) configurations. Importantly, the MFMIS, compared with MFIS, enables utilization of already existing and computationally-efficient industry-standard compact models such as BSIM-CMG [24]. In fact, this is inevitable for standard cell library characterization to create NCFET-aware libraries. Note that without such libraries, the impact of NC effects on the performance and power of large circuits such as processors and the examined MAC array (i.e., the focus of this work) cannot be performed. Finally, the developed compact model, described by the Verilog-A language, is employed within the commercial Synopsys SiliconSmart [36] to characterize the NCFET-aware cell libraries (see Section 3.2). Details on our NCFT-aware cell library characterization are available in [28].

NCFET Modeling Validation: Using the commercial TCAD tool flows from Synopsys (Sentaurus TCAD), we first calibrated the baseline FinFET device with the same 14nm measurement data from Intel [25]. Hence, the baseline FinFET device in both TCAD and BSIM-CMG model are calibrated to reproduce Intel 14nm measurement data. Then, we replace in TCAD the high-κ layer with a ferroelectric layer to realize the NC effect. As mentioned earlier, $Hf_{0.5}Zr_{0.5}O_2$ is used as the ferroelectric material with α = -1.1×10^9 m/F and β = 2.5×10^{10} $m^5/(C^2F)$ [26]. These parameters were extracted in [26] by fitting the Landau model to a real experimentally measured S shaped polarization-electric field curve [32].

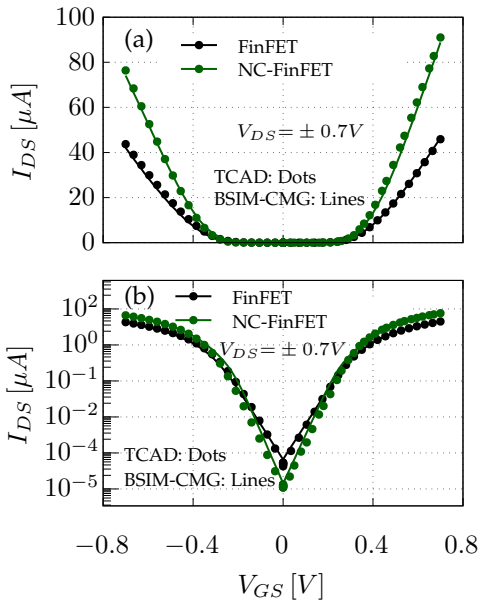

Fig. 4: Comparisons between results obtained from SPICE simulations using our calibrated BSIM-CMG compact model augmented with a physics-based ferroelectric model and results obtained from TCAD simulations. As shown, NC-FinFET results using our calibrated compact model come with an excellent agreement with TCAD results for both linear (a) and logarithmic (b) scales. Note that the baseline FinFET is calibrated with 14nm measurement data from Intel as presented in Figure 3.

In Figure 4(a and b), we present comparisons between results obtained from TCAD simulations and results obtained from SPICE simulations employing our calibrated BSIM-CMG compact modeling including NC modeling. As it can be observed, our calibrated BSIM-CMG model reproduces well the results of TCAD simulations for both baseline nFinFET and pFinFET devices as well as NC-nFinFET and NC-pFinFET devices. As shown, results of SPICE simulations of NC-FinFET come with a very good agreement with results from TCAD in both linear and logarithmic scales, as shown in Figure 4(a) and Figure 4(b), respectively. Note that the same ferroelectric parameters have been used in both TCAD simulations and SPICE simulations.

Switching Speed of NCFET Devices: In our NC modeling, we have assumed the polarization damping constant to be zero within the L-D equation for the doped HfO_2 FE material, which we consider in our work. Importantly, recent theoretical and experimental works suggested that the impact of finite switching time (often called polarization damping) in HfO_2 based FE materials is not critical for NCFET operation [37], [38]. This is because the NCFET digital switching from OFF (ON) to ON (OFF) state does not actually require a complete polarization switching from -P_r (+ P_r) to +P_r (-P_r) such as required in hysteresis loop measurements of FE capacitors or Ferroelectric FETs (FeFETs), which are used for building Non-Volatile Memories

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

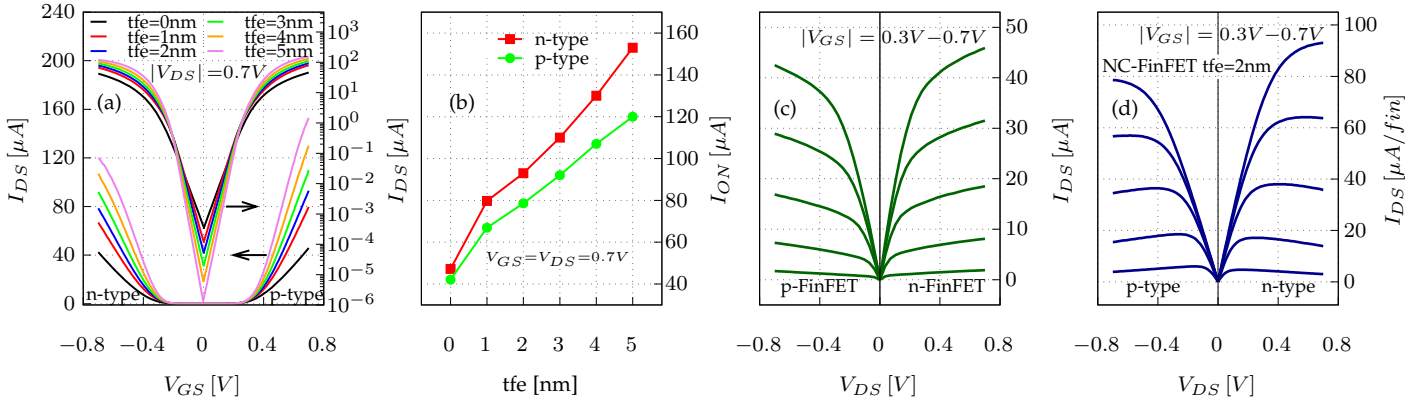

Fig. 5: (a) I_D-V_G comparisons between the baseline FinFET (t_{fe} = 0nm) and NC-FinFET for different ferroelectric thicknesses (t_{fe}). As can be noticed, thicker t_{fe} results in higher ON currents due to the larger intrinsic voltage amplification provided by negative capacitance effect. (b) provides a summary of the improvements in the ON current in NC-FinFET compared to the baseline FinFET. I_D-V_D at various V_{GS} biases for both baseline pFinFET and nFinFET devices (c) and NC-pFinFET and NC-nFinFET devices (d).

(NVMs) [37]. In practice, the inversion charge to be switched in state-of-the-art FETs (around $1\mu C/cm^2$) is generally much smaller than P_r value of practical ferroelectric materials. Taking this fact into account, [37] reported that the intrinsic delay of Z_r doped HfO_2 FE material to be merely 270 fs. This is, in fact, considerably smaller than the clock delay of modern GHz frequency processors. In addition, [38] showed that the measured delay of fabricated 14nm NC-FinFET based circuits (ring oscillators) can be directly modeled from the DC modeling of NC-FinFET using the static L-D theory, without any need to invoke transient effects. [38] reported very good matching between the model and the delay measurements up to very high frequencies. Therefore, the static L-D theory used in our work is a very reasonable approach to analyze NCFET based digital circuits (i.e., MAC array circuits as our focus). It is noteworthy that the underlying Landau-Khalatnikov (LK) modeling to capture NC effects accounts for large-signal as well. Furthermore, the intrinsic capacitance of the FE layer (background permittivity) is considered in our NC model. As we have used BSIM-CMG model, the capacitances of baseline FinFET are already captured by the BSIM-CMG model. Note that capacitance behavior in the NCFET changes due to gain from ferroelectric layer and this is already captured by the presented self-consistent simulation scheme (further details are available in [39]).

Device-Level Analysis: In our work, we explore the impact of NC-FinFET on boosting the performance of TPU while reducing the required cooling cost for given temperature constraints. In our analysis we consider various ferroelectric thicknesses (t_{fe}), i.e., 1nm up to 5nm. In Figure 5(a), we present the I_D-V_G characteristics of NC-nFinFET and NC-pFinFET devices for t_{fe} = 1, 2, 3, 4, 5nm in comparison to the baseline line nFinFET and pFinFET in which no ferroelectric layer is in use. In Figure 5, we refer to the baseline as t_{fe} = 0nm. The presented results in Figure 5(a) include both linear and logarithmic scales. As can be noticed, increasing the thickness of the employed ferroelectric layer increases the effect of negative capacitance, which leads to higher ON current due to the larger intrinsic voltage amplification. In Figure 5(b), we summarize the ON current as a function of ferroelectric thickness, for both p-type and n-type devices. As can be noticed, the ON current increases as the thickness of ferroelectric increases. For instance, at t_{fe} = 5nm, the ON current increases by around 3x compared to the baseline case (i.e., t_{fe} = 0nm).

In addition to the transfer characteristics presented in Figure 5(a), we present the output characteristics (I_D-V_D) in Figure 5(c and d) for both baseline FinFET and NC-FinFET, respectively. As an example to demonstrate the impact of negative capacitance, we show here the analysis for t_{fe} = 2nm. In Figure 5(c and d), various V_{GS} biases from 0.3V to 0.7V are used. As can be noticed, NC-FinFET provides a higher current compared to the baseline FinFET due to the effects of negative capacitance.

In Figure 3 and Figure 5, we report how NC boosts the transistor ON current which in turn leads to a higher switching speed. As can be noticed, the larger the FE layer thickness the higher the gain due to the higher internal voltage amplification.

In Figure 6, as a first approach to evaluate the impact of NCFET on Google TPU, we use the bfloat16 MAC unit (i.e., the basic building block of the MXU) as our driving circuit and we examine the frequency-power trade-off delivered by NCFET. Three thicknesses of FE layer are considered in Figure 6, i.e, 1nm, 3nm, and 5nm ($NC1$, $NC3$, and $NC5$, respectively) beside the conventional FinFET as baseline. As shown in Figure 6, at each voltage, as the thickness increases, the NCFET-based MAC achieves higher frequency, but also it consumes more power. As the voltage decreases, the frequency difference (between $NC1$, $NC3$, and $NC5$) becomes higher, while the power difference becomes less significant. Compared to the baseline, at each voltage value, NCFET achieves significantly higher frequency at the cost of increased power consumption. However, as shown in Figure 6, leveraging this frequency gain, we can decrease the operating voltage and obtain considerable power savings for similar performance with the baseline. For example, compared to the baseline at 0.8V, $NC1$, $NC3$, and $NC5$ at 0.6V achieve 4%, 8%, and 11% higher frequency, respectively, while delivering a power reduction of 31%, 21%, and 4%. Finally, as illustrated in Figure 6, compared to the

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

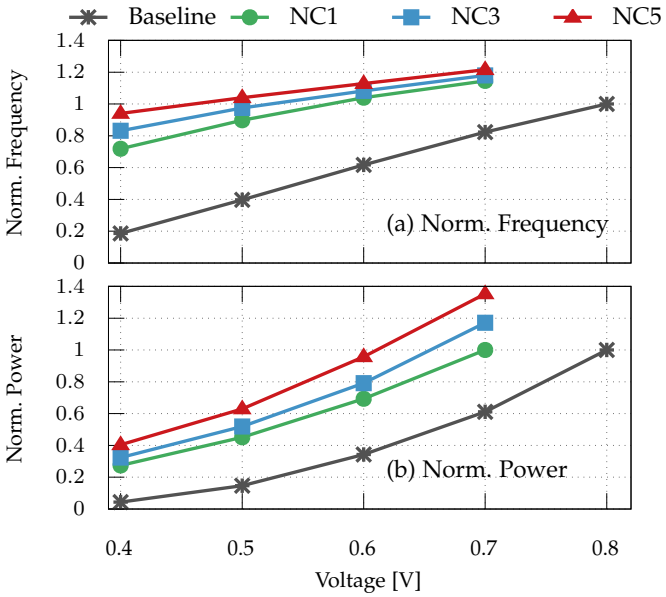

Fig. 6: The (a) maximum operating frequency and (b) total power consumption of the bfloat16 MAC circuit w.r.t. the voltage value for FinFET (baseline) and NCFET ($NC1$, $NC3$, and $NC5$). The frequency and power consumption are reported as normalized values over the respective ones of the baseline at 0.8V. NCFET boosts the maximum frequency of the MAC at any given voltage but increases the total power due to the increase in the frequency and total gate capacitance of the transistor.

baseline at 0.8V, as the voltage value decreases, the power overhead of NCFET decreases much faster than the delivered frequency gain. As a result, by lowering the supply voltage, NCFET is able to deliver better frequency-power trade-offs. These promising results motivate us to further evaluate the impact of NCFET on TPU and to this end, we perform a full-chip design of the MXU.

3.2 MXU Modeling

In this work, we implement the computational core of Google TPUv3, i.e., the MXU [40], [41]. MXU is a systolic 128×128 MAC array. The MXU RTL is developed in Verilog based on the optimized arithmetic components within the industrial Synopsys DesignWare library. Each MAC unit consists of a bfloat16 multiplier followed by a float32 accumulator [40], [41].

To model our designed chip for both FinFET and NCFET and enable fair comparisons when evaluating the power and frequency, we first calibrate the industry-standard compact model of FinFET technology (BSIM-CMG) [24] to have the same characteristics of Intel 14nm FinFET [25]. The calibrations are done for both nmos and pmos types of FinFET transistors to obtain model-cards. Model-cards are essential to perform an accurate SPICE simulations and then to characterize our standard cell-libraries. We then use a setup similar to the one used in [42]. We characterize our FinFET standard cells for a wide range of voltages between 0.4-0.8V by performing SPICE-accurate simulations. Similarly, we repeat the same flow to characterize NCFET-aware cell libraries. We integrate the physics-based NCFET model (see Section 3.1) within the industry-standard compact model of FinFET technology (BSIM-CMG). We then characterize our NCFET-aware cell-libraries for a wide range of voltages between 0.4-0.7V. Three thicknesses of the employed ferroelectric layer have been selected 1nm, 3nm, and 5nm ($NC1$, $NC3$, and $NC5$). Our libraries are fully compatible with the existing EDA tool flows (e.g., Synopsys and Cadence). Therefore, we can directly deploy them to perform a full-chip design all the way from logic synthesis to the GDSII level.

For chip design and implementation, we follow the standard EDA tools flow. We start with logic synthesis using Synopsys Design Compiler to synthesize the MXU under tight constraints to maximize the performance (e.g., compile_ultra, zero slack, etc). Then, the physical implementation of the full chip design (i.e., GDSII level) is performed using Cadence tools. Using Cadence Innovus, we designed the chip floorplan and the power delivery network. Then, the place and route including clock tree synthesis are done targeting the maximum performance. For accurate power and timing analysis, we use timing and power signoff tools to report the delay and power of the designed chip. For this purpose, we employ Cadence Tempus Timing Signoff tool for delay analysis and Voltus IC Power Integrity signoff tool for power analysis. Power and delay analyses are performed for FinFET for the whole voltage range with 100mV step and similarly for NCFET covering also all thickness values ($NC1$, $NC3$, and $NC5$). Moreover, we enabled the on-chip signal and power integrity to consider the impact of the RC-parasitic of the entire chip on delay and power. Finally, for dynamic power analysis, we employ QuestaSim to perform timing circuit simulations using realistic input datasets and extract the switching activity of the final designed chip which is used as input for the power signoff tool for accurate power estimation. Note that, all designed chips, FinFET and NCFET based ones, are fully identical and only the employed transistors differ (i.e., the same netlist, layout, packaging, etc).

3.3 Cooling Modeling

Advanced on-chip cooling, using superlattice (ultra-thin film) thermoelectric, is on top of the promising solutions for effective heat dissipation [10], [11]. It offers on-demand cooling for localized hot-spots that might form a thermal bottleneck for the entire chip. In principle, when an electrical current flows through the superlattice thin-film Thermoelectric cooler (TEC), a large thermal gradient between the upper and lower surfaces is rapidly formed due to the Peltier effect. This, in turn, enables effective and fast heat-pumping with a capability to cool heat fluxes up to $1300\,\mathrm{W/cm^2}$ as recent prototypes and measurements demonstrated [10], [11].

Importantly, when it comes to the microarchitecture of Google TPUs, the MXU has by far the largest on-chip power density due to the significant amount of power that is consumed in a confined area compared to other existing microarchitectural components such as the on-chip SRAM memory array and/or the peripheral circuits (e.g., PCI interface, control units, I/O, etc.). For instance, at the Intel $14\,\mathrm{nm}$ technology node which we consider in our work, the

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

power consumption of the MXU reaches ~64W (at 787MHz) within a very small area of ~0.275cm^2. Hence, this leads to an immense power density of $232\,\text{W}/\text{cm}^2$. Therefore, using superlattice TEC to perform on-chip localized cooling for the MXU provides an efficient solution to suppress such a thermal bottleneck that can deteriorate the performance and reliability of the entire chip.

In this work, we employ state-of-the-art superlattice TEC as described in [11], which consists of 3×3 thermoelectric couples where each couple has n-type and p-type leg. The surface area of each couple is $250\,\mu m \times 500\,\mu m$ and the thickness is $8\,\mu\text{m}$. The total area of the superlattice TEC itself is $1.75 \times 1.75\,\text{mm} = 0.03\,\text{cm}^2$. However, superlattice TEC consumes power for cooling and its power is proportional to the cooling capability, the higher the power (i.e., more current pass through) the higher the cooling capability. This additional power consumed to cool the chip, represents the *cooling cost* of TEC. The input current for superlattice TEC ranges from 0A-7A. Zero current is the baseline temperature when no active cooling (i.e., electric current) is applied. Note that in this case the effects of passive cooling associated with the superlattice TEC take place due to e.g., the lesser thickness of the thermal interface material. Hence, TEC does not result in a higher temperature compared to the baseline chip in which no TEC is implemented. This has been demonstrated from multiphysics simulations [43] as well as from experimental measurements [11]. Note that the 7A is TEC saturation as no further heat dissipation can be achieved even with feeding more current due to compensations caused by Joule heat effects. Given the area of MXU (0.275cm^2 as earlier mentioned) and the area of a single TEC device, nine TEC devices are used to cover the MAC array and provide the required MXU cooling.

For a more realistic heat/cooling simulations (i.e., thermal and thermal-electric simulations), we model a full system stack that consists of a silicon chip, superlattice thin-film Thermoelectric (TE), heat spreader and heat sink. All the material-dependent parameters of TEC and its dimensions were obtained from [11], [43]. The heat spreader and heat sink were made from aluminum and the base area of each is $0.16\,\text{cm}^2$ and $4\,\text{cm}^2$, respectively. The heat sink design was carefully 3-D modeled to fully replicate a commercial heat sink including all fins. Despite that the heat sink has a base surface area of $4\,\text{cm}^2$, the total surface area reaches $3112\,\text{mm}^2$ due to the existing fins. This allows having a proper heat dissipation. To simulate the maximum capability that conventional cooling using forced-convection of air provides, we consider a Heat Transfer Coefficient (HTC) of $100\,\text{W}/\text{m}^2\text{K}$, representing the maximum capability. All the heat and cooling simulations (with and without the cooling effect of superlattice TEC) including the above-described system was completely built in ANSYS®, a commercial multiphysics simulation tool flow using finite element methods, which can very accurately model the complex interactions between various parts including their different thermal properties (i.e., the silicon die, Peltier effect, Joule heat due to the flowing electrical current, heat flux, heat spreader/sink, and air convection on top of the heat sink etc.). The accuracy of the performed multiphysics simulations depends on many aspects such as the mesh sizing, which must have sufficient intrinsic temporal and spatial resolutions for the studied problem of interest, the boundary conditions, and material properties. In our work, we have calibrated the required meshing, after several iterations to maximize the resulted accuracy of the simulations. Furthermore, our analysis considers and simulates the entire system (i.e., silicon die, superlattice thermoelectric TE device including all intrinsic parts of TE, heat spreader, thermal material interface, full heat sink including several fins, etc.). This ensures a very high accuracy in our heat simulations while considering the various interactions between the different materials.

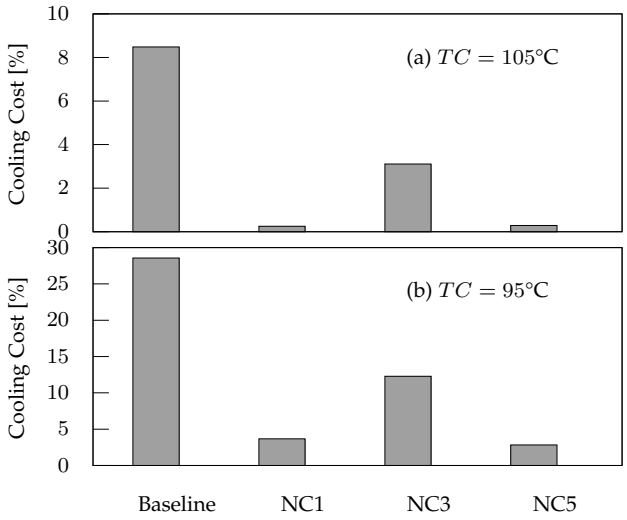

Fig. 7: The cooling costs required to sustain the maximum operating frequency 787MHz which the baseline MXU can achieve (at 0.8V) while satisfying given thermal thresholds. The temperature constraints are: (a)TC=105°C (b) TC=95°C.

4 Evaluation and Analysis

In this section we experimentally evaluate how employing NCFET can mitigate and even eliminate the mandatory cooling cost that is required to enable the operation of the MXU under sustainable temperature thresholds. More than 4000 NCFET different configurations are generated and evaluated in order to explore the entire design space defined by the NCFET thickness size, the operating voltage, the input current to TEC, and the operating frequency.

The maximum operating frequency that the baseline MXU can achieve is 787MHz at $0.8V$. However, in this case, the MXU power density reaches $232\,\text{W}/\text{cm}^2$, resulting in an unsustainable on-chip temperature of 138°C (considering conventional air-cooling) that goes beyond the critical temperature of 105°C, as typically defined by Intel [44]. As a result, in order to decrease the temperature to acceptable values we must either i) decrease the operating frequency that results in a significant throughput loss, or ii) apply emerging cooling (TEC) and pay the respective cooling cost, or iii) apply a combination of emerging cooling and frequency decrease to trade-off between cooling cost and performance.

Targeting the maximum performance, we examine in Figure 7 the required cooling cost in order to enable the operation at the maximum frequency (787MHz) while satisfying the temperature constraints of 105°C (Figure 7a) and

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

95°C (Figure 7b). Note that even when considering the maximum cooling cost (i.e., 7A input current for TEC) the baseline cannot achieve a temperature below 93°C at 787MHz. In Figure 7, the cooling cost is reported as a percentage value with respect to the power consumption of the baseline MXU at 787MHz. Hence, as shown in Figure 7, in order to satisfy the temperature constraints of 105°C and 95°C, the baseline requires a cooling cost equivalent to 8.5% and 28.6%, respectively, of the MXU's power consumption. On the other hand, as shown in Figure 7, for all the examined thicknesses, NCFET requires a significantly lower cooling cost to enable operation at 787MHz. When considering the 105°C constraint, the cooling cost of the NCFET with thickness 1nm, 3nm, and 5nm ($NC1$, $NC3$, and $NC5$) is 0.25%, 3.11%, and 0.29%, respectively. Similarly, for 95°C, the respective values are 3.67%, 12.29%, and 2.82%. Both $NC1$ and $NC3$ attain 787MHz at 0.7V and thus, $NC3$ requires higher cooling cost than $NC1$ due to its higher power consumption that originates from the increase in gate capacitance with higher thickness. This is verified by the fact that at 787MHz and 0.7V the PD of $NC1$ and $NC3$ is $183\,\mathrm{W/cm^2}$ and $211\,\mathrm{W/cm^2}$, respectively. However, $NC5$ requires less cooling cost than both $NC1$ and $NC3$. $NC5$ achieves 787MHz at a lower voltage, namely at 0.6V and hence, at lower PD, i.e., $180\,\mathrm{W/cm^2}$. As a result, when targeting operation at the maximum frequency of the baseline, NCFET with thickness 5nm ($NC5$) is highly preferred due to the minimal cooling cost, *10x smaller compared to the baseline*, which is required to satisfy the temperature constraint of 95°C as shown in Figure 7b.

In Figure 7, operation at the maximum frequency of the baseline is considered. However, operating at such high frequency stresses the baseline, leading to very high PD and thus, highly increased requirements for cooling cost in order to mitigate the increased temperature. In Figure 8, as a mean to decrease the temperature, we consider both TEC as well as operation at a lower frequency. Figure 8 presents the cooling cost–frequency trade-off for the baseline and NCFET under varying temperature thresholds. For the baseline, in addition to 0.8V we also examine operation at 0.7V. The maximum operating frequency of the baseline at 0.7V is only 641MHz. In Figure 8, we set the minimum presented frequency to 610MHz, i.e., 10% throughput loss compared to the advertised raw throughput of Google TPU[2]. Moreover, we consider a frequency step of 10MHz and we, therefore, examine frequencies that range from 610MHz up to the maximum frequency of each design. For the input current of TEC, a step of 0.25A is considered and the examined values range from 0A to 7A. In addition, three temperature constraints of 100°C, 85°C, and 75°C are evaluated. In order to enable direct comparisons in Figure 8, the cooling cost is reported as an absolute value in Watts. As shown in Figure 8, for each frequency (same vertically), NCFET requires lower cooling cost in order to satisfy the examined temperature constraints. The ability of NCFET to achieve the same frequency with the baseline at lower voltage values, leads to significantly lower PD and thus, significantly lower cooling cost. It is noteworthy that in

2. The MXU of TPUv3 operates at ~680MHz delivering a raw throughput of 22.5TFLOPS [40], [41].

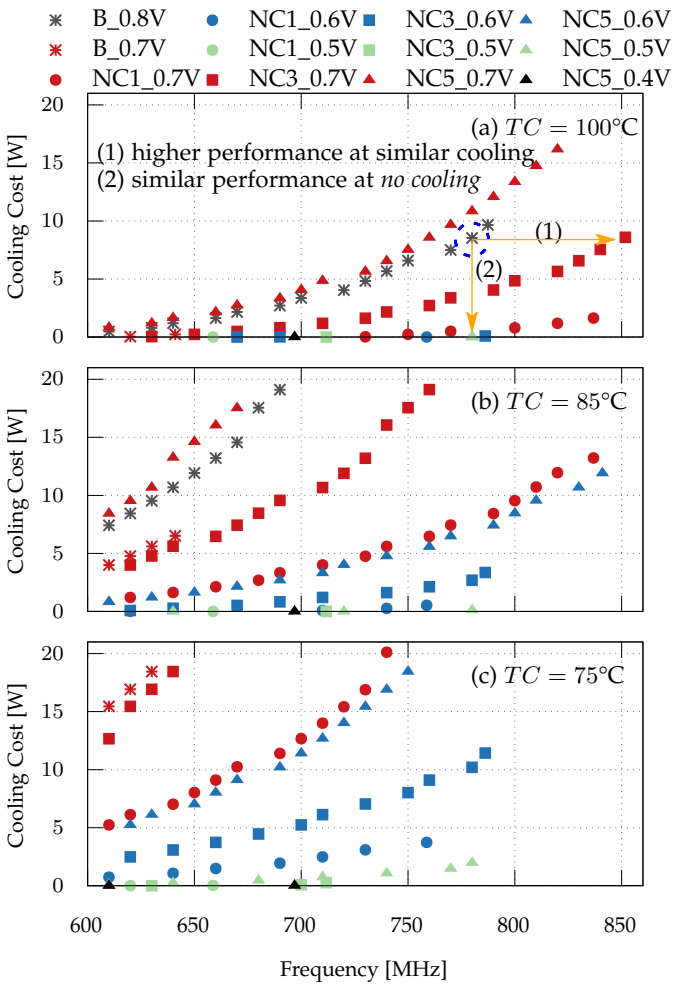

Fig. 8: The cooling cost–frequency trade-off using the emerging cooling (TEC) and frequency scaling to mitigate the increased temperature of the MXU. Three temperature constraints of (a) 100°C, (b) 85°C, and (c) 75°C are evaluated. For the baseline, we examine operation at 0.8V and0.7V while for NCFET 0.4V- 0.7V. A step of 10MHz is used for the frequency scaling and a minimum frequency threshold of 610MHz is considered. Importantly, (a) shows two use cases: (1) NCFET increases the performance at the same cooling cost, (2) NCFET can entirely eliminate the cooling cost at the same frequency with the baseline. A symbol-color coding is used: points with the same color feature the same voltage supply and points with the same symbol feature the same thickness size.

many cases NCFET not only mitigates (i.e., reduces) the cooling cost requirement but also it eliminates it. For 100°C temperature threshold, NCFET achieves 780MHz at zero cooling cost ($NC5$ at 0.5V). Similarly, for 85°C constraint, NCFET enables operation at 711MHz ($NC3$ at 0.5V) with zero cooling cost. For 75°C threshold, NCFET ($NC1$ at 0.5V) delivers 620MHz without any cooling cost requirements. For the respective frequencies, the cooling cost required by the baseline is 8.54W (at 0.8V) for 100°C and 16.92W (at 0.7V) for 75°C. For 85°C constraint, even with maximum cooling, the baseline can achieve up to 690MHz for more than 19W cost. Furthermore, as also shown in Figure 8, compared to the baseline and for the same cooling cost

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

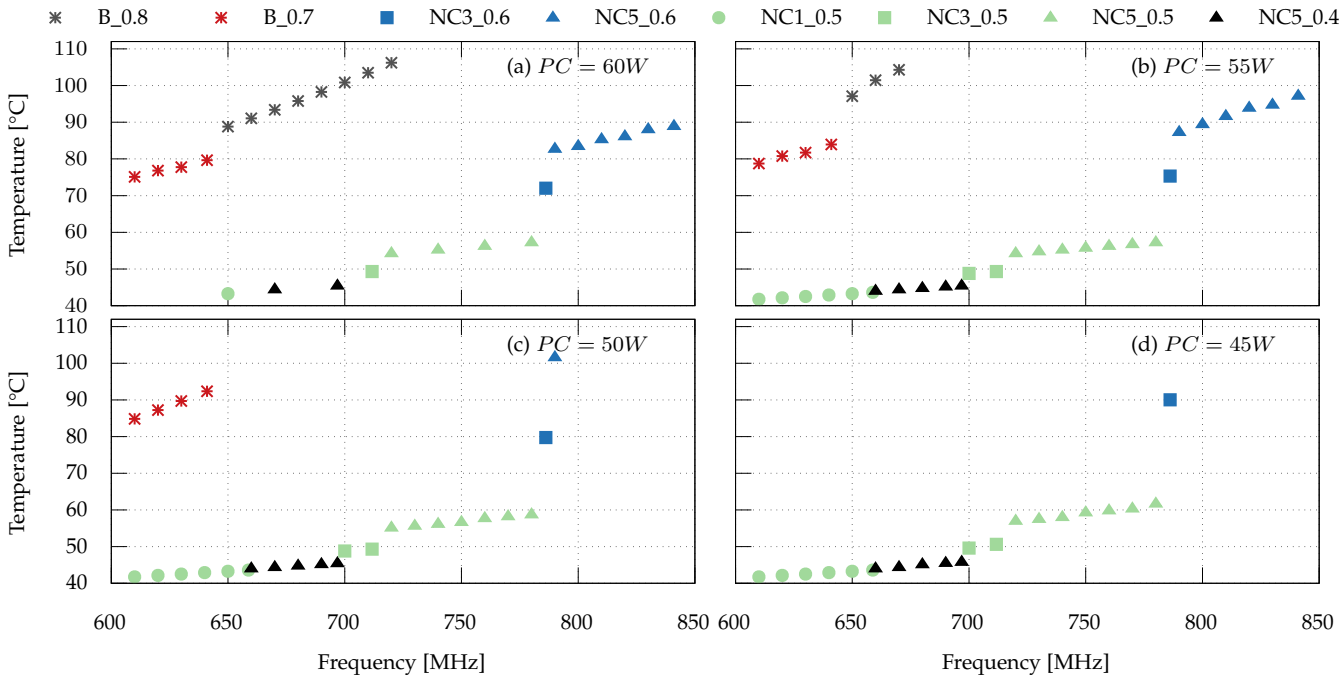

Fig. 9: The temperature–frequency Pareto frontier of the baseline and NCFET for four different power constraints: (a) 60W, (b) 55W, (c) 50W, and (d) 45W. A step of 10MHz is used for the frequency scaling and a minimum frequency threshold of 610MHz is considered. Note that, the power constraint incorporates both the MXU power consumption and the cooling cost. As shown, for all the examined power constraints, NCFET significantly outperforms the baseline, achieving lower temperature at the same frequency and higher frequency at the same temperature. The same symbol-color coding with Figure 8 is used.

(same horizontally), NCFET can achieve significantly higher operating frequency. For example, considering 10W cooling cost, NCFET achieves 852MHz ($NC3$ at 0.7V) for the 100°C constraint. Similarly, for 85°C temperature threshold, NCFET delivers 810MHz ($NC5$ at 0.6V), while for 75°C the respective value is 780MHz ($NC5$ at 0.5V). On the other hand, for up to 10W cooling cost, the baseline achieves 787MHz (0.8V) at 100°C and 641MHz (0.7V) at 85°C. In other words, NCET delivers 8% and 26% higher frequency respectively. For 75°C threshold, the baseline cannot satisfy the 10W cost limit and the cooling cost is more than 15W for 610MHz operating frequency at 0.7V. Finally, the above analysis and Figure 8 highlight the importance of properly selecting, at design time, the NCFET thickness size with respect to the desired operating frequency and target temperature constraint at runtime.

In Figure 8, we demonstrate that emerging cooling (TEC) can efficiently decrease the MXU temperature at the cost of increased power consumption. As shown, NCFET eliminates this cooling cost (e.g., Figure 8a(2)) and/or achieve significantly higher operating frequency for the same cost(e.g., Figure 8a(1)). However, the operation of TPU is not only limited by the temperature value but also by the total power consumption that incorporates cooling cost. In Figure 9, we plot the temperature–frequency Pareto frontier of the baseline and NCFET for four different power constraints: (a) 60W, (b) 55W, (c) 50W, and (d) 45W. Similar to Figure 8, we set a lower frequency threshold of 610MHz, a frequency step of 10MHz, and a TEC current step of 0.25A. All the designs that feature temperature less than 110°C are depicted in Figure 9. Given an acceptable power budget, Figure 9 can be used to select the frequency-optimal design at each temperature threshold. As shown in Figure 9, for all the examined power constraints, NCFET significantly outperforms the baseline, achieving lower temperature at the same frequency and higher frequency at the same temperature. As the power constraint decreases, for the same frequency, the temperature of both the baseline and NCFET increases since less power budget is available to spend for cooling. For example, the minimum temperature of the baseline increases from 75°C (at 0.7V) for 60W power constraint to 85°C (at 0.7V) for 50W available power budget. On the other hand, due to the inherent ability of NCFET to achieve lower PD at the same frequency, compared to the baseline, NCFET is less affected by the power constraint. Consider also that, as shown in Figure 8, NCFET eliminates in many cases the cooling cost, and thus, the available power budget impacts only its maximum operating frequency (since power is proportional to the frequency) and not the cooling efficiency of TEC. Even for 45W power budget, NCFET delivers 780MHz ($NC5$ at 0.5V) while the respective attained temperature is lower than 70°C. In the meantime, for the 45W threshold, the baseline *cannot go below* 110°C and thus it is not depicted in Figure 9d. Similarly, for 60W power constraint (Figure 9a), NCFET achieves 800MHz ($NC5$ at 0.6V) for less than 85°C, while the baseline delivers up to 641MHz for the same power and temperature constraints.

Trading cooling cost for temperature reduction enables the MXU to achieve a significantly higher frequency for given temperature constraints. However, the cooling cost can become considerable for high frequencies and/or low temperature constraints, significantly increasing the total

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

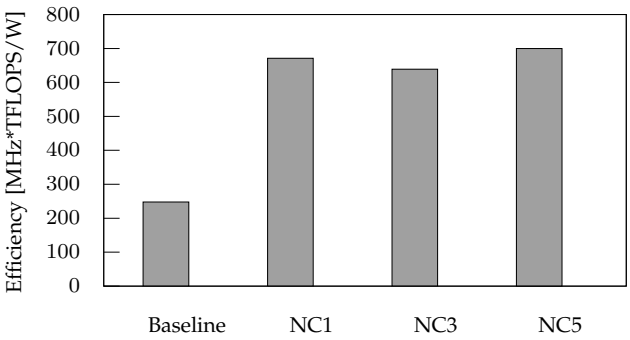

Fig. 10: The maximum efficiency of the baseline and NCFET for 85°C temperature constraint. Compared to the baseline $NC1$, $NC3$, and $NC5$ deliver 2.71x, 2.58x, and 2.82x higher efficiency, respectively. Note that, the respective NCFET designs feature zero cooling cost, boosting the efficiency of the NCFET-based MXUs.

power consumption of the MXU. For this reason, in Figure 10 we examine the efficiency of the MXU. The efficiency is defined as: $frequency \times \frac{Throughput}{Total\ Power}$ (MHz $\times$ TFLOPS/W). In Figure 10, we demonstrate the maximum efficiency of the baseline and NCFET for 85°C temperature constraint and similar results are obtained for the other temperature values. As shown in Figure 10, for all the thickness values, NCFET features significantly higher efficiency than the baseline. Specifically, $NC1$, $NC3$, and $NC5$ achieve 2.71x, 2.58x, and 2.82x higher efficiency, respectively, compared to the baseline. Note that, all the NCFET designs in Figure 10 feature zero cooling cost (see Figure 8b). The ability of the NCFET to eliminate the cooling cost, even at high frequency values and low temperature constraints, boosts the efficiency of NCFET since it can achieve very high performance without the need to pay for additional power for cooling. On the other hand, as shown in Figure 8, given a temperature constraint, the cooling cost of the baseline increases very fast w.r.t. the frequency. As a result, the latter significantly contributes to the total power consumption of the baseline, significantly limiting thus its efficiency.

5 Discussion and Summary

In this work, we investigated, for the first time, the impact of Negative Capacitance (NCFET) technology, which is one of the recent emerging technologies, on the cooling cost and efficiency of future MAC arrays (i.e., as the heart of Google TPU) in combination with superlattice thermoelectric cooling, which is one of the emerging on-chip cooling solutions. When it comes to modeling the impact of NCFET on the systolic MAC array, our work aims at very fine-grained modeling in which we start from transistor modeling to standard cell modeling, to MAC circuit, all the way up to MAC array modeling. It is noteworthy that the systolic MAC array forms the major part and the heart of state of the art NN accelerators such as Google TPU. Other components like on-chip memory were out of the scope of this work due to the lack of NCFET modeling for them.

To achieve our goal and provide robust analysis, we followed an accurate evaluation procedure based on industry strength tool flows and we precisely modeled the core components of our examined setup, i.e., the MXU and the superlattice thermoelectric cooling. However, the granularity of the presented analysis in Section 4 depends also on the granularity of the design space exploration. Overall, four parameters define the granularity of our design space search: frequency step, thickness step, voltage value step, and input current step for TEC. For the frequency value (i.e., our main optimization goal) we used a very fine-grained approach, i.e., only 10MHz. In addition, the input current step for TEC, our second optimization goal (i.e., cooling cost) is also reasonably selected, i.e., 0.25A, in which we observed a marginal gain in cooling when a smaller power/current step is considered. Our analysis demonstrates that NCFET is able to boost (even maximize w.r.t. the temperature constraint) the performance for the same cooling cost as the baseline or to completely eliminate the cooling cost for the same performance. On the other hand, the steps for the thickness (2nm) and the voltage value (100mV) are more coarse-grained. Selecting smaller steps for the voltage and thickness values would deliver a more fine-grained exploration and more design points for the examined system. *However, even with the current setup, we explored 4000 NCFET-based configurations*. Considering the increased time required for full-chip design, post-layout simulation, and/or heat simulations [17], examining a finer granularity for the voltage value and the thickness would explode the design space as well as the exploration time required. Concluding, in our analysis, we used a fine-grained system modeling and we selected a very fine-grain search approach for our major optimization metrics (i.e., frequency and cooling cost) and a more coarse-grain one for the rest parameters that affect the evaluation (i.e., voltage value and thickness). Nevertheless, note that the presented design points are sufficient to demonstrate the high impact of NCFET on the cooling and efficiency of future NN accelerators (e.g., TPUs). For example, for all the examined temperature constraints, NCFET eliminates the cooling cost for the same performance as the baseline.

Finally, to evaluate the impact of NCFET on the cooling and efficiency of future NN accelerators, we used as our use case scenario the matrix multiply unit (MXU) employed in Google TPUv3. Our evaluation framework is not dependent on the evaluated MXU, however, it can be used with any circuit/microarchitecture. The only requirement is to provide the register transfer level (RTL) description. The origin of NCFET is at the transistor level and thus, similar results are expected for any digital circuit [22], [42]. Moreover, note that in our analysis, we first examined the performance and power evaluation of the MAC unit (see Fig. 6). The power gain at the MAC level is translated to power gain at the microarchitecture level, and thus, to lower cooling cost. Nevertheless, evaluating the power density, temperature, and cooling cost at only the MAC unit level (i.e., common part in any Neural Network accelerator), would result in imprecise conclusions. To obtain an accurate evaluation of the impact of NCFET on the cooling cost and efficiency of NN accelerators, we need to evaluate a larger and realistic microarchitecture (e.g., to accurately capture the area, switching power, and power density). We consider Google TPUv3 to be a perfect candidate for our evaluation since it is a state-of-the-art chip and can be used for both NN inference and training. Therefore, although the performed analysis is

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

highly correlated with the bfloat16 MAC array we examined (MXU), it provides valuable and clear insight regarding the impact that NCFET can bring on the temperature and efficiency of future NN accelerators.

6 Conclusion

In this work, we consider the emerging Superlattice Thermoelectric cooling as the cooling solution for Google TPU and we investigate the impact of Negative Capacitance FET (NCFET) on mitigating the increased cooling cost of TPU that originates from its very high power density. To this end, we perform a thorough design space evaluation that is built upon industry-strength tools, full chip design of the MXU based on 14nm Intel FinFET technology and its NCFET counterpart, and multiphysics simulations for accurate modeling of the temperature and cooling cost. Our analysis demonstrates that NCFET is able to mitigate the increased power density of the MXU. Compared to the conventional FinFET baseline, we demonstrate that, for any given temperature constraint, under iso-frequency conditions NCFET minimizes and even eliminates the cooling cost, while under iso-cost conditions NCFET boosts the MXU performance. Finally, our evaluation shows that by eliminating the cooling cost, NCFET delivers 2.8x higher efficiency compared to the conventional FinFET baseline.

Acknowledgment

Authors would like to thank Om Prakash from KIT for his valuable help in TCAD simulations and NCFET analysis.

References

[1] "PULPino Processor Platform," December 2017. [Online]. Available: http://www.pulp-platform.org/

[2] 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)*, 2016, DOI:10.1145/2872362.2872414.

[3] W. Huang, M. Stan, S. Gurumurthi *et al.*, "Interaction of scaling trends in processor architecture and cooling," in *26th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, 03 2010, DOI:10.1109/STHERM.2010.5444290.

[4] L. A. Barroso, U. Holzle, P. Ranganathan *et al.*, *The Datacenter as a Computer: Designing Warehouse-Scale Machines, Third Edition*, 2018.

[5] datacenter knowledge, "Google Brings Liquid Cooling to Data Centers to Cool Latest AI Chips," July 2018. [Online]. Available: https://www.datacenterknowledge.com/google-alphabet/google-brings-liquid-cooling-data-centers-cool-latest-ai-chips

[6] H.-M. Tong, Y.-S. Lai, and C. Wong, *Advanced Flip Chip Packaging*, 08 2013, DOI:10.1007/978-1-4419-5768-9, isbn:978-1-4419-5767-2.

[7] T. J. Chainer, M. D. Schultz, P. R. Parida *et al.*, "Improving data center energy efficiency with advanced thermal management," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 2017.

[8] S. Wang, Y. Yin, C. Hu *et al.*, "3d integrated circuit cooling with microfluidics," in *Micromachines*, 06 2018.

[9] Rosie Frost , "Giant data centres use billions of litres of water to cool down computers," July 2020. [Online]. Available: https://time.com/5814276/google-data-centers-water/

[10] G. Bulman, P. Barletta, J. Lewis *et al.*, "Superlattice-based thin-film thermoelectric modules with high cooling fluxes," *Nature Communications*, vol. 7, 01 2016.

[11] I. Chowdhury, R. Prasher, K. Lofgreen *et al.*, "On-chip cooling by superlattice-based thin-film thermoelectrics," *Nature nanotechnology*, vol. 4, 05 2009.

[12] D. Zhao and G. Tan, "A review of thermoelectric cooling: Materials, modeling and applications," in *Applied Thermal Engineering*, vol. 66, 05 2014, DOI:10.1016/j.applthermaleng.2014.01.074.

[13] T. U. Uttam Shyamalindu Ghoshal, Austin, "Highly reliable thermoelectric coolingapparatus and method," 1999, uS Patent US6266962B1. [Online]. Available: https://patents.google.com/patent/US6266962B1/en

[14] 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.

[15] S. Salahuddin and S. Datta, "Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices," in *Nano Letters*, vol. 8, no. 2, 2008, DOI:10.1021/nl071804g.

[16] S. Salamin, M. Rapp, A. Pathania *et al.*, "Power-effecient heterogeneous many-core design with ncfet technology," in *IEEE Transactions on Computers (TC)*, 2020.

[17] H. Amrouch, G. Zervakis, S. Salamin *et al.*, "Npu thermal management," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2020.

[18] M. Rapp, S. Salamin, H. Amrouch *et al.*, "Performance, Power and Cooling Trade-Offs with NCFET-based Many-Cores," *Design Automation Conference (DAC)*, 2019.

[19] S. K. Samal, S. Khandelwal, A. I. Khan *et al.*, "Full chip power benefits with negative capacitance fets," in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, July 2017.

[20] S. Salamin, M. Rapp, H. Amrouch *et al.*, "NCFET-Aware Voltage Scaling," *The International Symposium on Low Power Electronics and Design (ISLPED)*, 2019.

[21] S. Salamin and M. Rapp and H. Amrouch and A. Gerstlauer and J. Henkel, "Energy optimization in ncfet-based processors," in *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2020.

[22] G. Zervakis, H. Amrouch, and J. Henkel, "Design automation of approximate circuits with runtime reconfigurable accuracy," *IEEE Access*, vol. 8, 2020.

[23] J. Henkel and N. Dutt, *Dependable Embedded Systems*, 2021, ISBN 978-3-030-52016-8, DOI: https://doi.org/10.1007/978-3-030-52017-5.

[24] BSIM, "BSIM-CMG Technical Manual," October 2018. [Online]. Available: http://www-device.eecs.berkeley.edu/bsim/?page=BSIMCMG

[25] S. Natarajan, M. Agostinelli, S. Akbar *et al.*, "A 14nm logic technology featuring 2nd-generation finfet, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm2 sram cell size," in *2014 IEEE International Electron Devices Meeting*, 2014.

[26] M. Hoffmann, B. Max, T. Mittmann *et al.*, "Demonstration of high-speed hysteresis-free negative capacitance in ferroelectric hf0.5zr0.5o2," in *2018 IEEE International Electron Devices Meeting (IEDM)*, Dec 2018.

[27] K. He *et al.*, "Deep residual learning for image recognition," in *IEEE Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*, June 2016.

[28] H. Amrouch, G. Pahwa, A. D. Gaidhane *et al.*, "Impact of variability on processor performance in negative capacitance finfet technology," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2020.

[29] Devonshire and A. Frederick, "XCVI. Theory of barium titanate: Part I," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 40, no. 309, pp. 1040–1063, 1949, doi:10.1080/14786444908561372.

[30] 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 I: Model Description," *IEEE Transactions on Electron Devices*, vol. 63, no. 12, pp. 4981–4985, 2016.

[31] 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," *IEEE Transactions on Electron Devices*, vol. 63, no. 12, pp. 4986–4992, Dec. 2016, doi:10.1109/TED.2016.2614436.

[32] M. Hoffmann, F. P. Fengler, M. Herzig *et al.*, "Unveiling the double-well energy landscape in a ferroelectric layer," *Nature*, vol. 565, no. 7740, 2019.

[33] S. Mueller, J. Mueller, A. Singh *et al.*, "Incipient ferroelectricity in Al-doped HfO_2 thin films," *Advanced Functional Materials*, vol. 22, no. 11, 2012.

[34] G. Pahwa, A. Agarwal, and Y. S. Chauhan, "Numerical investigation of short-channel effects in negative capacitance mfis and

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TC.2021.3065454, IEEE Transactions on Computers

mfmis transistors: Subthreshold behavior," *IEEE Transactions on Electron Devices*, vol. 65, no. 11, pp. 5130–5136, 2018.
[35] ——, "Numerical investigation of short-channel effects in negative capacitance mfis and mfmis transistors: Above-threshold behavior," *IEEE Transactions on Electron Devices*, vol. 66, no. 3, pp. 1591–1598, 2019.
[36] Synopsys, "Synopsys EDA Tool Flows," October 2018. [Online]. Available: https://www.synopsys.com/
[37] K. Chatterjee, A. J. Rosner, and S. Salahuddin, "Intrinsic speed limit of negative capacitance transistors," *IEEE Electron Device Letters*, vol. 38, no. 9, pp. 1328–1330, 2017.
[38] D. Kwon, Y. Liao, Y. Lin *et al.*, "Response speed of negative capacitance finfets," in *2018 IEEE Symposium on VLSI Technology*, 2018, pp. 49–50.
[39] G. Pahwa, T. Dutta, A. Agarwal *et al.*, "Designing energy efficient and hysteresis free negative capacitance finfet with negative dibl and 3.5x ion using compact modeling approach," in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*. IEEE, 2016, pp. 49–54.
[40] Google. (2019) Bfloat16: The secret to high performance on cloud tpus. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus
[41] Google. (2019) Hotchips 2019 tutorial. [Online]. Available: https://www.hotchips.org/hc31/HC31_T3_Cloud_TPU_Codesign.pdf
[42] H. Amrouch, G. Pahwa, A. D. Gaidhane *et al.*, "Negative capacitance transistor to address the fundamental limitations in technology scaling: Processor performance," *IEEE Access*, 2018.
[43] S. H. Choday, M. S. Lundstrom, and K. Roy, "Prospects of thin-film thermoelectric devices for hot-spot cooling and on-chip energy harvesting," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 3, no. 12, 2013.
[44] Intel, 2016. [Online]. Available: https://www.intel.com/content/dam/support/us/en/documents/joule-products/intel-joule-thermal-management.pdf

Sami Salamin received his B.Sc. degree in computer systems engineering and M.Sc. degree (first rank) from Palestine Polytechnic University, Hebron, Palestine in 2005 and 2012, respectively. Since 2016, he is pursuing his Ph.D. degree with the Chair of Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. He holds HiPEAC Paper Award. ORCID 0000-0002-1044-7231

Georgios Zervakis is a Research Group Leader at the Chair for Embedded Systems (CES) at the Karlsruhe Institute of Technology (KIT), Germany. He received the Diploma and the Ph.D. degree from the Department of Electrical and Computer Engineering (ECE), National Technical University of Athens (NTUA), Greece, in 2012 and 2018, respectively. Before joining KIT, Georgios worked as a primary researcher in several EU-funded projects as member the Institute of Communication and Computer Systems (ICCS), Athens, Greece. His research interests include approximate computing, low power design, design automation, and integration of hardware acceleration in cloud.

Florian Klemme (M'20) is a Doctoral Researcher at the Chair of Semiconductor Test and Reliability, University of Stuttgart. He received the B.Sc. in System Integration from the University of Applied Sciences Bremerhaven, Germany, in 2014 and the M.Sc. in Computer Science from the Karlsruhe Institute of Technology, Germany, in 2018. He is currently working towards the Ph.D. degree at the Chair of Semiconductor Test and Reliability, University of Stuttgart. His research interests include cell library characterization and machine learning techniques in electronic design automation and computer-aided design. He is a member of the IEEE. ORCID 0000-0002-0148-0523.

Hammam Kattan received his B.S. degree in mechanical engineering from Aleppo University, Aleppo, Syria, in 2011, and his M.S. degree in mechanical engineering from Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2017. He was working as research assistant at Chair for Embedded Systems at Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. His research interests include advanced cooling techniques, thermoelctrics, modeling and simulation using Finite Element Method (FEM).

Yogesh Singh Chauhan (SM'12-F'21) is currently a full Professor with IIT Kanpur, Kanpur, India. He is also the developer of industry standard BSIM-BULK (formerly BSIM6) model for bulk MOSFETs and ASM-HEMT model for GaN HEMTs. He has authored or co-authored over 200 papers in international journals and conferences. His current research interests include characterization, modeling, and simulation of semiconductor devices.

Jörg Henkel (M'95-SM'01-F'15) is the Chair Professor for Embedded Systems at Karlsruhe Institute of Technology. Before that he was a research staff member at NEC Laboratories in Princeton, NJ. He received his diploma and Ph.D. (Summa cum laude) from the Technical University of Braunschweig. His research work is focused on co-design for embedded hardware/software systems with respect to power, thermal and reliability aspects. He has received six best paper awards throughout his career from, among others, ICCAD, ESWeek and DATE. For two consecutive terms he served as the Editor-in-Chief for the ACM Transactions on Embedded Computing Systems. He is currently the Editor-in-Chief of the IEEE Design&Test Magazine and is/has been an Associate Editor for major ACM and IEEE Journals. He has led several conferences as a General Chair incl. ICCAD, ESWeek and serves as a Steering Committee chair/member for leading conferences and journals for embedded and cyber-physical systems. Prof. Henkel coordinates the DFG program SPP 1500 "Dependable Embedded Systems" and is a site coordinator of the DFG TR89 collaborative research center on "Invasive Computing". He is the chairman of the IEEE Computer Society, Germany Chapter, and a Fellow of the IEEE.

Hussam Amrouch (S'11-M'15) is a Junior Professor for the Semiconductor Test and Reliability (STAR) within the Computer Science, Electrical Engineering Faculty at the University of Stuttgart as well as a Research Group Leader at the Karlsruhe Institute of Technology (KIT), Germany. He received his Ph.D. degree with distinction (Summa cum laude) from KIT in 2015. His main research interests are design for reliability and testing from device physics to systems, machine learning, security, approximate computing, and emerging technologies with a special focus on ferroelectric devices. He holds seven HiPEAC Paper Awards and three best paper nominations at top EDA conferences: DAC'16, DAC'17 and DATE'17 for his work on reliability. He currently serves as Associate Editor at Integration, the VLSI Journal. He has served in the technical program committees of many major EDA conferences such as DAC, ASP-DAC, ICCAD, etc. and as a reviewer in many top journals like T-ED, TCAS-I, TVLSI, TCAD, TC, etc. He has 110+ publications in multidisciplinary research areas across the entire computing stack, starting from semiconductor physics to circuit design all the way up to computer-aided design and computer architecture. He is a member of the IEEE. ORCID 0000-0002-5649-3102.

0018-9340 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on April 15,2021 at 14:59:52 UTC from IEEE Xplore. Restrictions apply.