IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 26, 2021, accepted May 23, 2021, date of publication May 28, 2021, date of current version June 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3084749

Longevity of Commodity DRAMs in Harsh Environments Through Thermoelectric Cooling

DEEPAK M. MATHEW[1], (Member, IEEE), HAMMAM KATTAN[2], CHRISTIAN WEIS[1], (Member, IEEE), JÖRG HENKEL[2], NORBERT WEHN[1], (Senior Member, IEEE), AND HUSSAM AMROUCH[3], (Member, IEEE)

[1]Microelectronic Systems Design Research Group, TU Kaiserslautern, 67663 Kaiserslautern, Germany
[2]Chair for Embedded Systems, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
[3]Chair of Semiconductor Test and Reliability, University of Stuttgart, 70569 Stuttgart, Germany

Corresponding author: Hussam Amrouch (amrouch@iti.uni-stuttgart.de)

ABSTRACT Today, more and more commodity hardware devices are used in safety-critical applications, such as advanced driver assistance systems in automotive. These applications demand very high reliability of electronic components even in adverse environmental conditions, such as high temperatures. Ensuring the reliability of microelectronic components is a major challenge at these high temperatures. The computing systems of these applications rely on DRAMs as working memory, which are built upon bit cells that store charges in capacitors. These commodity DRAMs are optimized for cost per bit and not for high reliability. Thus, very high temperatures impose an enormous challenge for commodity DRAMs as the data retention time and reliability decrease largely, affecting the data correctness. Data correctness can be ensured up to certain temperatures by increasing the refresh rate to counterbalance the retention time reduction. However, this severely degrades the access latencies and the usable DRAM bandwidth. To overcome these limitations, we present for the first time a Thermoelectric Cooling (TEC) solution for commodity DRAMs in harsh-environments, such as automotive. Our TEC solution enables the use of commodity off-the-shelf DRAMs in safety-critical applications by reducing the temperature conditions to a range where they can operate reliably. This TEC solution is applied a posteriori to the DRAM chips without using high-cost package solutions. Thus, it maintains the low-cost targets of such devices, improves the reliability, and at the same time, counterbalances the adverse effects of increasing the refresh rate. To quantitatively evaluate the benefits of TEC on commodity DRAMs in harsh-environments, we performed system-level evaluations with several applications backed up by the measured data on commodity DRAMs. Our experimental results, using accurate multi-physics simulations that employ finite element method, demonstrate that the TEC-based cooling ensures that the maximum temperature of all DRAM chips is always below 85°C despite that the original on-chip temperature (i.e., in the absence of our TEC based cooling) goes beyond 120°C.

INDEX TERMS DRAM, commodity hardware, thermoelectric, harsh-environment, automotive, temperature, bandwidth, reliability, error-rate.

I. INTRODUCTION

Breakthroughs in artificial intelligence using Deep Neural Networks (DNNs) have led recently to make autonomous driving at the forefront of goals of automotive industry [1]. However, when it comes to safety-critical applications like in autonomous driving, the advanced driver assistance systems inevitably demand Dynamic Random Access Memories (DRAMs) to exhibit very low latency, high bandwidth [2], and extreme-low bit error rate [3] to fulfil tight reliability and performance constraints [4], [5]. The key challenge is that the harsh environments in which the operating temperatures go beyond 100°C make electrics in vehicles extremely unreliable [6]. The maximum temperature depends on the mounting location of electronics. For instance, electronic control units (ECUs) in cars are often placed in the

The associate editor coordinating the review of this manuscript and approving it for publication was Yu Wang.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

engine compartment or locations where the temperatures may rise up to 125°C depending on the atmospheric temperature conditions (e.g., tropical summer). Therefore, automotive safety standards, such as ISO 16750-4 [7] and AEC-Q100 [8], demand ensuring reliable operation of electronics up to this temperature.

For DRAMs, where data is stored as charge in capacitors, this affects the data retention time, reliability of internal circuitries, and eventually the data correctness. The key figures of merit of any DRAM chip, such as error rate, latency, refresh rate, leakage currents, and bandwidth considerably degrade when the temperature increases. Major DRAM vendors offer specialized chips where the maximum allowed temperature may reach up to 125°C [9], [10], but at a considerable increase in cost (around 60%)[1] compared to other commercial off-the-shelf DRAM chips that typically operate below 85°C (e.g., in desktop processors). Moreover, for temperatures above 85°C, the refresh rate has to be doubled for every 10°C rise in temperature [4], [9]. This severely degrades the available DRAM bandwidth and considerably increase the average access latencies. Hence, satisfying the strict requirements of autonomous driving with respect to memory becomes profoundly challenging, especially with the ever-increasing demand for higher bandwidths and lower access latencies to ensure execution of real-time tasks under stringent timing constraints.

In order to suppress the deleterious effects of elevated temperatures on DRAMs, we propose to use a thermoelectric device to effectively cool down DRAM DIMMs using the Peltier effect. Our experimental results employing accurate multi-physics simulations after modeling the entire DRAM system demonstrate that when a thermoelectric device is used to dissipate the heat of DRAM DIMMs, the maximum temperature of all DRAM chips can be reduced from 125°C down to 85°C. For various applications, we demonstrate how thermoelectric cooling can ensure that the DRAM on-chip temperature to be always below 85°C which allows the use of commercial off-the-shelf DRAM chips in harsh environments. This has a considerable impact on reducing the cost of automotive electronics as well as ensuring real-time requirements in reliability and bandwidth.

Our key contributions within this paper are as follows:

(1) For harsh environments, where the DRAM chip operate at an ambient temperature of 120°C like in automotive, we propose to employ Thermoelectric-based Cooling (TEC) to effectively manage the temperature of DRAMs. For accurate analysis, we have used commercial multi-physics simulations that employ finite element method.

(2) We evaluate the impact of temperature on figures of merit of DRAM chips in harsh environments like in automotive, and demonstrate how sustaining the required extreme-low error rate of $< 1E{-}9$ in DRAMs results in a significant loss in the DRAM bandwidth due to the exponential increase in the refresh rate. In addition, the rise in refresh rate results in a considerable increase in the average response latency of DRAMs ($\sim$ 50%), which is not tolerated in the majority of algorithms of advanced driver assistance systems.

(3) Using a state-of-the-art DRAM simulation framework after augmenting it with our DRAM measurements and with our multi-physics heat simulations of DRAMs, we study a wide range of applications, including neural network inference, demonstrating the efficacy of our proposed cooling solution in reducing the maximum temperature of all DRAM chips in the DIMM and improving the performance degradation caused by higher refresh rate.

The rest of the paper is organized as follows. In Section II, we concisely summarize our main focus in this work, and in Section III, we provide a brief description of DRAM architecture, refresh operation and its impacts, as well as the DRAM challenges in the automotive industry. Section IV discusses the prior research on TEC as well as DRAM retention errors. Section V demonstrates how the increase in temperature influences the data retention time and retention errors as well as operating currents of DRAMs, based on real measurements. The impacts of temperature-induced retention time reduction on applications' performance and energy are evaluated in Section VI using system-level simulations. Section VII presents our novel TEC technique and shows its benefits on various applications. Finally, Section VIII concludes this paper.

II. THE PRIMARY FOCUS OF THIS WORK

Our cooling solution aims at managing the temperature of DRAMs in electronic systems that operate in harsh environments in which the ambient temperature is excessive (e.g., 120°C) like in automotive. The key goal of our work is to enable, for the first time, the use of commodity DRAMs, which cannot reliably operate beyond 85°C, in such harsh environments. Employing off-the-shelf commodity DRAMs in harsh environments inevitably necessitates reducing the DRAM chip temperature to below 85°C (i.e., the maximum allowed temperature in commodity DRAMs). Otherwise, DRAM's reliability (i.e., the prerequisite probability of error) cannot be ensured anymore.

Existing traditional cooling like forced-convection air or liquid-based cooling will never be able to reduce the temperature of the DRAM chips to below the ambient despite how strong the cooling capability and how significant the provided thermal convection. Therefore, none of the existing cooling solutions in the market can enable the use of off-the-shelf commodity DRAMs in harsh environments.

In such environments, the only possible solution currently is to not use commodity DRAMs and instead using specialized DRAM chips that are fabricated and tested to operate under excessive temperatures (e.g., up to 125°C). However, such specialized (i.e, automotive-graded) DRAM chips are much more expensive (around 60% as mentioned earlier) than off-the-shelf commodity DRAM chips. It is noteworthy that

[1] Based on the data available from https://www.avnet.com and https://www.digikey.de, accessed on Feb. 2021.

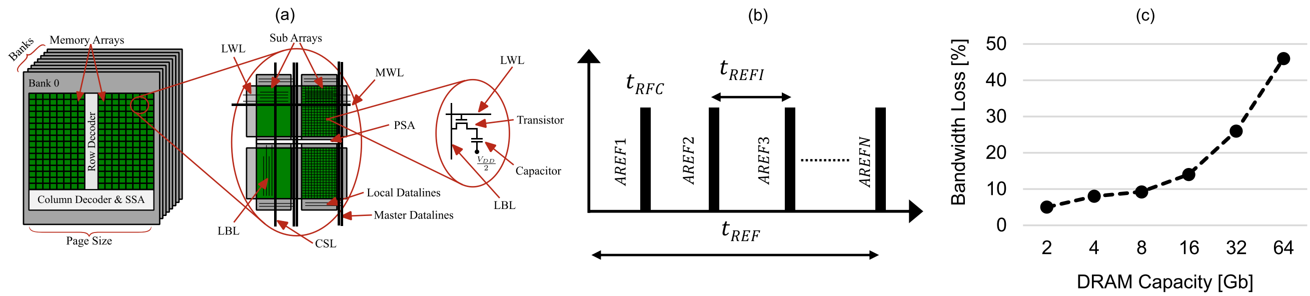

FIGURE 1. (a) An overview of the general architecture of a DRAM chip and how it is hierarchically composed. (b) Refreshing a DRAM device is periodically done at every T_{REFI} interval to ensure data retention (i.e., $P_{fail} < 1E-9$). (c) The periodic refresh of DRAM reduces the available bandwidth. Increasing the DRAM capacity results in large losses in the bandwidth.

in many industries, such as automotive, optimizing the cost of electronic components is a key when it comes to mass production. Hence, enabling the use of commodity DRAMs, while still sustaining the same prerequisite level of reliability, can provide a considerable cost saving.

More importantly, while it is true that specialized DRAM chips that are designed to operate in harsh environments ensure reliability (i.e., keep the probability of error under the prerequisite level), they do suffer from a significant loss in the DRAM bandwidth. This is because sustaining reliability is achieved through increasing the DRAM refresh rate (e.g., from 64 ms (< 85°C) to 8ms at 120°C [9]), which largely reduces the available DRAM bandwidth. Note that losses in DRAM bandwidth are often not tolerated, especially when it comes to applications that demand frequent DRAM access such as in the vast majority of deep learning algorithms in advanced driver assistance systems.

In short, employing commodity DRAMs in harsh environments demands ensuring that the DRAM's temperature does not exceed the specification (i.e., maximum of 85°C). On the one hand, none of the existing traditional cooling solutions can reduce the temperature to beyond the ambient temperature. On the other hand, using specialized DRAMs that can operate in harsh environments results in not only an increase in the cost but also a considerable loss in the DRAM bandwidth.

III. BACKGROUND INFORMATION

First, we explain the basic DRAM architecture and functionality to demonstrate the impact of temperature on DRAM figures of merit. As depicted in Fig. 1(a), a DRAM device is organized as a set of memory banks (e.g., eight) that includes memory arrays (e.g., two). Each memory array has row and column decoders, master wordline drivers and secondary sense amplifiers. Buses, buffers, control signals, voltage regulators, charge pumps, and other peripherals are shared between the different banks. The memory arrays are formed in a hierarchical structure out of sub-arrays for efficient wiring, increased speed and reduced power consumption. Each sub-array is equipped with primary sense amplifiers. A typical memory sub-array consists of e.g., 512 cells × 512 cells = 256 Kb and a 64 Mb DRAM bank is often formed out of two memory arrays, where each memory array consists of 8 × 16 = 128 sub-arrays.

Each single DRAM memory cell is built as a transistor-capacitor pair where the data is stored in the capacitor as electric charges. A DRAM cell has various leakage paths [11], which degrade the stored charge over time. Therefore, DRAM cells inevitably need to be *periodically* refreshed to retain the data stored within the cell capacitors. Otherwise, errors will occur during reading operations and failure probability starts to rapidly increase. The retention time of a DRAM cell is defined as the amount of time that the cell can safely retain its data without being refreshed (i.e., under such a refresh rate the failure probability is guaranteed to be smaller than a certain value). At normal temperature conditions (< 85°C), a typical DRAM cell exhibits a retention time of 64 ms. Hence, it must be refreshed every 64 ms, referred as refresh time (t_{REF}) to retain the stored data. This guarantees that failure probability (P_{fail}) to be smaller than $1E-9$, i.e., < 1 bit errors over the whole data stored in a memory chip with a capacity of e.g., 4 Gbits — as none of the existing automotive standards permit any bit errors from DRAM.

Modern DRAMs are equipped with an *Auto-Refresh* (`AREF`) command to perform this operation. A single `AREF` command does not refresh the entire DRAM at once, but instead it refreshes only a certain number of rows in all banks, depending on the capacity of the DRAM device. Therefore, the memory controller needs to issue `AREF` commands at regular intervals, called *Refresh Interval* (t_{REFI}), to completely refresh the entire DRAM within 64 ms (t_{REF}), as shown in Fig. 1(b). Each refresh command/operation blocks the DRAM during the *Refresh Cycle Time* (t_{RFC}) from performing any read/write operations. As a result, the intermittent refresh operations reduce the *available DRAM bandwidth*[2] (by a factor of t_{RFC}/t_{REFI}), and increase the average memory access latency (typically ranges from 15 ns to 50 ns) because DRAM accesses that arrive during refresh must wait for t_{RFC}. For Double Data Rate 4 (DDR4) DRAMs, this refresh interval is always fixed by JEDEC standard to 7.8 μs [12] for all DRAM densities at normal operating temperatures (< 85°C).

[2]The maximum bandwidth of commodity DIMM-based DDR DRAM solutions is defined by the product of the number of data I/Os (e.g. 64) and the achievable per I/O data rate (e.g. 2.4 Gbit/s).

In total, $N = 8192$ AREF commands must be issued in every t_{REF} window, in order to ensure refreshing the complete DRAM DIMM[3] (a 2 GB DDR4 SO-DIMM composed of *four* 4 Gb DRAM devices) within 64ms. This reduces the available bandwidth from the theoretical maximum limit of 153.6 Gbps down to 140 Gbps (i.e., 8.8% bandwidth loss).

Importantly, with every new generation, DRAM vendors continuously aim at enlarging the total storage capacity on a memory device. The t_{RFC} increases with DRAM capacity due to the increase in the total number of rows and more number of rows have to be refreshed in every AREF command. Hence, the available bandwidth of DRAMs shrinks when the DRAM capacity increases. As shown in Fig. 1(c) [13], [14], the available bandwidth exponentially decreases with the increase in the DRAM capacity. For 64 Gb[4] DRAM the loss in the bandwidth reaches more than 45% compared to merely 8.8% at 4 Gb.

In summary, with the continuous increase in demand for larger DRAM capacity, the available bandwidth shrinks. Hence, tolerating any further reductions in the DRAM bandwidth induced by elevated temperatures (> 85°C), as will be explored later in Section V, becomes profoundly challenging.

How Automotive Industry Challenges DRAMs: Emerging AI applications in advanced driver assistance systems particularly impose fundamental obstacles for the DRAM industry because they indispensably request 1) extreme low failure probability ($P_{fail} < 1E{-}9$) to guarantee safety. 2) very small latency to guarantee meeting hard deadlines, and 3) very large bandwidth to guarantee that the massive number of activation and weight values that DNNs dictates can be loaded in time to the neural processing unit and fulfilling the hard real-time deadlines.

An additional serious challenge that automotive particularly imposes to electronics in general and DRAM specifically is the harsh environment in which the operating temperature is above 100°C due to the very confined structure where electronics in vehicles must be packed together as well as the excessive heat generated from motor and/or radiated from the sun vehicle bodywork. When the temperature goes beyond the nominal operating condition of 85°C, the DRAM cells rapidly become very unreliable because of large increase in various leakage currents within the DRAM cell. As a result, the DRAM capacitor, where the information itself is stored as charges, will discharge rapidly as the temperature increases leading to much higher probability of failures.

To overcome that and to maintain the probability of failure at acceptable levels, the refresh time of DRAM must become smaller ($t_{REF} < 64$ ms), which inevitably necessitates increasing the DRAM refresh rate. This reduces further the available DRAM bandwidth on top of the already large loss incurred due to the increase in the DRAM capacity (e.g., 26% loss for 32 Gb DRAM as shown in Fig. 1(c)).

[3] Dual Inline Memory Module (DIMM) is a PCB module that consists of multiple DRAM devices/chips

[4] At present, maximum available capacity of 32Gb (per chip-package) for LPDDR4 DRAMs in automotive applications.

In addition to the increased cell leakage, various leakage currents in DRAM periphery circuitries also rise with temperature. This increases the operational currents and leads to a higher power consumption.

IV. RELATED WORK

A. THERMOELECTRIC COOLING (TEC)

Due to its strong capability in dissipating heat, employing TEC in managing the temperature of processor chips has attracted a large attention in the last few years for both high-performance as well as mobile devices. [16] proposed to mount a TEC between the two covers of a smartphone and location of hot-spot was determined in a prior using an infrared camera. [16] also proposed to use the same TEC device to harvest energy from the wasted heat when its ideal (i.e., when the temperature of chip is not critical and there is no need to turn the cooling on). Because most of the modern mobile devices have a thickness less than 10 *mm* [17], using a conventional "bulk" TEC device that has a thickness of few millimeters becomes infeasible due to space limitation. Therefore, ultra-thin film TEC has been proposed in which a TEC layer of around 50μm is integrated between the chip's die and the above packaging [18], [45] to manage the temperature localized hot-spots. [19] demonstrated experimentally the advantage of using a TEC in reducing performance losses caused by the thermal throttling in processors when the temperature at run-time goes beyond the critical temperature. The used TEC in [19] was a TEC that has a surface area of 15×15 mm^2 and a thickness of 3.22 *mm* covering the entire processor chip. In [20], the effectiveness of employing ultra-thin film TEC in managing the temperature of Neural Processing Units (NPUs) was recently presented.

B. IMPACT OF TEMPERATURE ON DRAM RETENTION AND INDUCED ERRORS

There are many prior studies [13], [21]–[27] characterizing the retention errors in commercial off-the-shelf DRAMs at different temperatures and refresh times. However all of these studies are limited to either standard operating temperatures below 85°C or short refresh times (< 200 ms). In our previous work [28], [29] we presented a platform to measure the retention errors and operational currents of DDR3 and DDR4 DRAMs for temperatures up to 95°C and refresh times up to 1000 s. This enables us to cover extreme operating conditions, as discussed in this paper.

C. DISTINGUISH FROM EXISTING STATE OF THE ART

In this work, we are the first to propose the use of TEC to manage the temperature of DRAM DIMMs. We demonstrate how the integration of a single TEC device between the DRAM DIMM and above mounted heat-sink can ensure that the maximum temperature of all four DRAM chips (located within the DIMM) is always below 85°C despite the operation in a harsh environment where the ambient temperature is 120°C. This has a far-reaching impact on the automotive

industry because it enables off-the-shelf *commodity DRAMs* to be used in vehicles while tight reliability constraints are still met. This, in turn, removes the large cost overheads stemming from the inevitable need to purchase specialized expensive DRAMs that support the operation in harsh environments. Additionally, our TEC cooling approach eliminates the performance degradation (bandwidth and average access latencies) caused by the increased refresh rates at high temperatures.

V. TEMPERATURE EFFECTS IN DRAMs

To quantitatively analyze the effect of temperature on the retention time and operational currents of commodity DDR4 DRAMs, specifically at temperatures beyond 85°C, we used our measurement platform presented in [29]. These measurements are essential to obtain the required refresh interval and DRAM operational currents at high temperatures, which are used for system-level simulations and accurate multi-physics simulations in Section VI and Section VII, respectively.

A. MEASUREMENT PLATFORM

Fig. 2 shows our measurement platform for performing retention measurements and the operational currents measurement under various temperatures. It is designed to measure retention errors, power consumption, and to heat up the DRAM devices of DDR4 SO-DIMM modules. The heating section consists of a mechanical setup, which is placed on the surface of the DRAM devices to heat them up within a range of 25°C to 95°C. The accuracy of the temperature control was determined to ±2°C using thermal simulations. To analyze the key current sources in DDR4 DRAM (details in Table 1) of V_{DD} and V_{PP} voltage domains, we designed a JEDEC-conform adapter board for DDR4 SO-DIMMs, which is shown in Figure 2. The power lines V_{DD} and V_{PP} are routed across 4 mΩ shunt resistors, whereas the data, address, and control lines are passed through. Due to precise impedance-matched layout design, the adapter board works with DRAM clock frequencies greater than 1 GHz. For our experiments, all currents were measured at 1.2 GHz (DDR4-2400). The voltages across the shunt resistors are amplified with current-sense amplifiers. High-precision 24-bit Analog to Digital Converters are synchronously sampling and converting these voltages into digital values. The measurement platform has a current measurement accuracy of ±0.5 mA. For our tests we used a Xilinx Virtex Ultrascale FPGA-based evaluation platform. The standard *Memory Interface Generator* (MIG) memory controller from Xilinx was customized to generate the required command and data sequences for the current measurements, and to pause the refreshes for retention error measurements. A *Virtual Input/Output* (VIO) core connected to the custom MIG enables the real-time monitoring and control of the internal FPGA signals.

Although we tried to increase the maximum temperature range of the platform to temperatures higher than 95°C (in order to mimic the operation in a harsh environment, as the main focus of this work), we were not able to stabilize the measurement setup at these elevated temperatures because of the deleterious impact of elevated temperature on periphery circuits required for measurement and control. Therefore, we extrapolated the measurement results, as described in Section V-B.

TABLE 1. Key current parameters in commercial DRAMs [15]. The corresponding measurements are in Fig. 3(a).

Name	Explanation
I_{DD0}	*One Bank Active Precharge Current*: Measured across ACT and PRE commands to one bank (other banks remain precharged).
I_{DD2N}	*Precharge Standby Current*: Measured when all banks are precharged (PRE).
I_{DD3N}	*Active Standby Current*: Measured when all banks are active (ACT).
I_{DD4R}	*Burst Read Current*: Measured during read operation, assuming seamless write data burst with all data bits toggling between bursts and all banks open, with the RD commands cycling through all the banks.
I_{DD4W}	*Burst Write Current*: Similar to I_{DD4R}, but with WR commands.
I_{DD2P}	*Precharge Power Down Current*: Measured during precharge power down mode.
I_{DD3P}	*Active Power Down Current*: Measured during active power down mode.
I_{DD5}/I_{DD5B}	*Refresh Current*: Measured during refresh operation, with REF commands issued every t_{RFC}.
I_{DD6N}	*Self Refresh Current*: Measured when the DRAM is in Self Refresh mode.

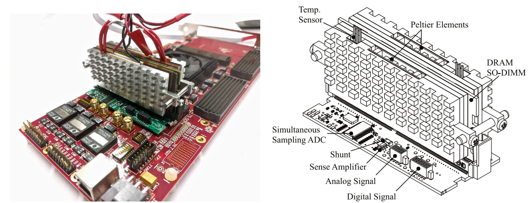

FIGURE 2. Our built measurement platform and adapter board, which enables us to accurately characterize the impact of temperature on failure probability of DRAM cells, different DRAM currents and the retention time [29].

B. MEASUREMENTS OF TEMPERATURE EFFECTS IN DRAMs

Fig 3(a) summarized our measurements of the key sources of current in DRAMs at various temperatures. Due to the aforementioned limitation of the measurement platform, the current values (IDD) at 120°C were determined by extrapolating the currents from 90°C using an exponential function. As can be noticed, temperature rise increases the leakage currents, which, in addition to increasing the total power and energy of the system, decreases the retention time as shown in Fig 3(b). In the latter the constraint of failure probability of $< 1E{-}9$ is considered. However, if the refresh time remains the same, the failure probability will increase as

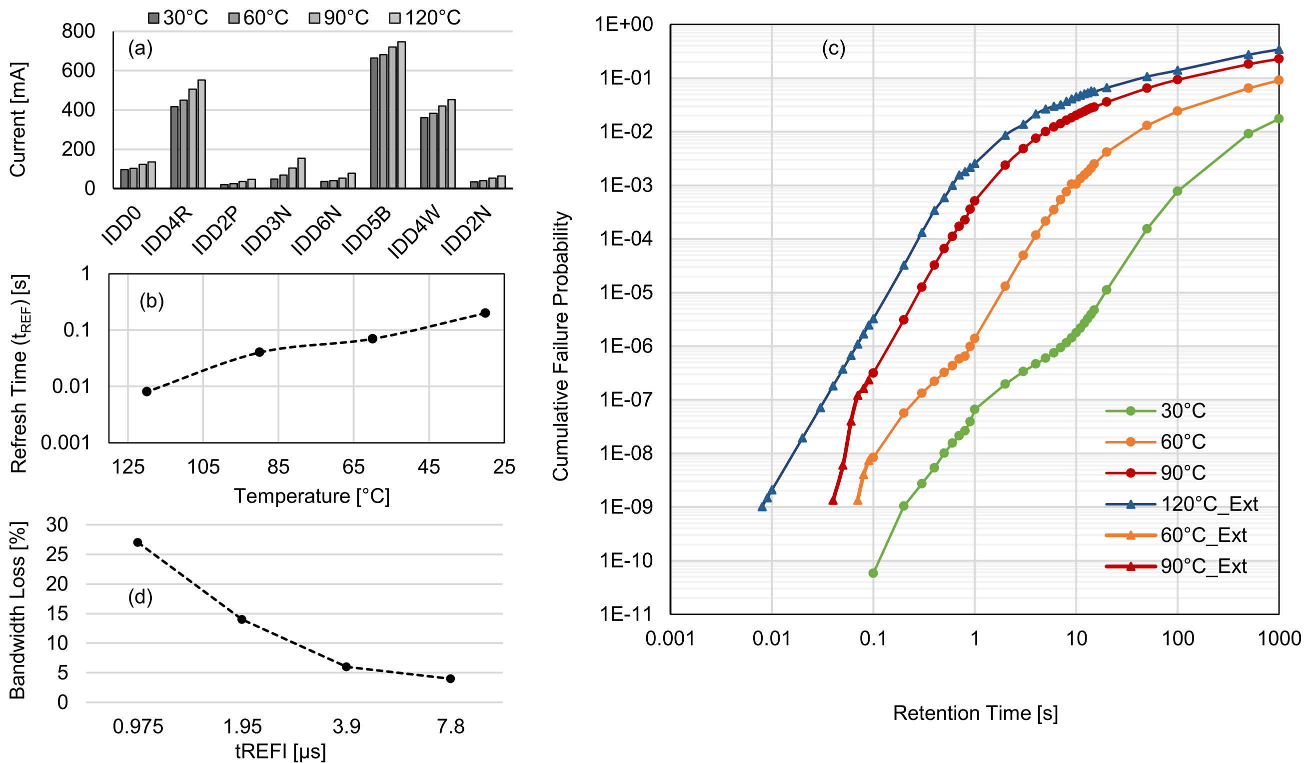

FIGURE 3. **(a) Impact of temperature increase on the various key sources of current (details on Table 1. (b) Impact of temperature rise on considerably decreasing the retention time of DRAM cells. (c) The relation between retention time and cumulative failure probability demonstrating the impact of temperature increase on failure probability in DRAM cells. (d) The relation between the refresh interval (t_{REFI}) and bandwidth loss.**

shown in Fig. 3(c) because the DRAM cells will not be able to retain their stored data until the end of the refresh time interval.

Fig. 3(c) demonstrates the plots the Cumulative Failure Probability (CFP)/bit-error rate of a DRAM cell at different temperatures and refresh times, based on the measured data of a commodity DDR4 DRAM using the measurement platform described in the previous section. Note that failure probabilities for 120°C were extrapolated based on the measured data for 30, 60, and 90°C. Different regions of the 90°C curve (1) from 0.1s to 1s (linear behaviour), (2) from 1s to 50s (exponential curve), and (3) from 50s to 1000s (linear behavior) were analysed. Then the additional increase of the CFP due to the larger temperature at 120°C was extrapolated based on the increase in CFP from 30°C to 60°C and from 60°C to 90°C. A failure probability of $< 1E{-}9$ corresponds to a $t_{REF} = 64$ ms at temperatures below 85°C. At higher temperatures, the retention time is reduced. Thus, t_{REF} must be accordingly reduced to sustain the same bit-error rate. As can be noticed from Fig. 3(b and c), to sustain the same bit-error rate of below $1E{-}9$, the refresh time must be decreased from 64 ms to 32 ms when the temperature is increased from 60°C to 90°C. When the temperature increases further to 120°C, the refresh time must be decreased to below 10 ms. In order to compensate for this refresh time reduction, the refresh rate has to be quadrupled and t_{REFI} is decreased from 7.8 μs to .97 μs).

This, in turn, leads to a severe collapse of the available bandwidth (see Fig 3(d)), and a non-tolerable increase in the response latency as the DRAM device will be for the vast majority of the execution time unavailable because it will be busy in keep refreshing the stored data. The problem becomes even worse when the DRAM capacity is larger because the available bandwidth become much smaller at the first place as explained earlier in Section I (see Fig. 1(c)). It is noteworthy that decreasing the refresh time, additionally, considerably increases the total power and energy that the DRAM device consumes due to the massive increase in the refresh power and due to the temperature-dependent rise in leakage power in the periphery circuitries of DRAM.

VI. IMPACT OF ELEVATED TEMPERATURES ON APPLICATIONS

A. SIMULATION SETUP

We employ the DRAM simulation framework DRAMSys [30], [31] and the power modeling tool DRAMPower [32], [33] to evaluate the impact of refresh on the performance and energy of DRAMs while running realistic applications. We employ trace-based simulation approach for the majority of experiments due to the fast simulation time, but also validate our results against full-system simulations for some applications. For trace-based simulations, traces were generated from the CHStone [34] and the Mediabench [35] benchmarks through the SimpleScalar

simulator (ARM instruction set) [36] with 16 KB L1 D-cache, 16 KB L1 I-cache, 128 KB shared L2 cache and 32-byte cache line configuration. We filter out the L2 cache misses for instructions and data, and obtained a trace of the DRAM transactions. The generated traces were then executed on the DRAMSys framework.

Three different applications are selected from the CHStone and the Mediabench benchmark suites. Those are *h263encode*, *chstone-aes* and *chstone-sha*. The applications chstone-aes and chstone-sha have dense memory access pattern while the h263encode application exhibits a sparse memory access pattern. To simulate realistic scenarios in multicore CPUs, we also simulated a mixture of these applications by running them parallel on different CPU cores. Those are *mixed1* (chstone-aes + chstone-sha) and *mixed2* (chstone-aes + chstone-sha + h263encode).

For the full-system simulations, the applications were executed on the Gem5 simulator [37] that is linked with the DRAMSys framework. In these experiments, we use the STREAM benchmark (*stream*) [38] and a Neural Network (NN) inference application (*nn_inference*) [39]. The neural network application performs an inference task on the MNIST data set for recognizing handwritten digits using a fully connected neural network (single-precision).[5] Gem5 was configured with an ARM High-Performance In-order (HPI) CPU model (v8-A architecture) with two CPU cores, and each clocked at 4.0 GHz. All CPU cores have their own private 64 KB L1 D-cache and I-cache. The L2-cache (1 MB) with a 64 -Byte cache line was shared between the cores. Multicore scenarios were simulated by running the same application simultaneously on two different cores (i.e., *nn_inference_2cores*, *stream_2cores*). In addition, we have executed the STREAM benchmark on a single core as well (i.e., *stream_1core*).

The simulated memory system consists of a 2 GB DDR4 SO-DIMM composed of *four* 4 Gb DRAM devices, which is the same DRAM configuration adopted in our measurement platform (referring to Section V-A) for consistency. Then, the timing specifications of DRAM devices are taken from the datasheet [40]. Importantly, the various DRAM operational currents and the required refresh rates to sustain a Bit-Error Rate (BER) of $< 1E{-}9$ under different temperatures were obtained from our measurements as presented in Fig 3(a and b), respectively. The t_{REF} must be reduced for temperatures $>$ 85°C to sustain the required BER. For temperatures within the normal operating condition (i.e., $<$ 85°C), the t_{REF} is kept 64 ms as commercially done

B. EXPERIMENTAL RESULTS

Fig. 5(a) shows the average bandwidth of the simulated applications at different temperatures. Since the *h263encode* application is less memory intensive, it exhibits relatively lower average bandwidth compared to the *mixed_trace*. Note that for temperatures up to 85°C, the bandwidth remains constant because the t_{REF} is fixed within the normal operating condition as explained earlier. However, the bandwidth starts to drastically reduce from 90°C. The large reduction in bandwidth from 90°C to 120°C is due to the 4× increase in the refresh rate in order to compensate the retention time reduction (see Fig. 3(b and c)). In practice, refresh operations block the DRAM from performing reads and writes until the refresh cycle is completed. Hence, increasing the refresh rate reduces the available bandwidth. This reduction in bandwidth is prominent for the mixed trace due to its dense memory access pattern.

In addition to the bandwidth reduction, it also increases the average response latency of the DRAM, which seriously degrade the performance of applications and increases the likelihood of missing hard deadline especially when it comes to safety-critical application such as in automotive. As can be noticed in Fig. 5(b), the average response latency does not change for temperatures below 85°C as the t_{REFI} is constant. However, it rapidly grows from 90°C due to the increased refresh rate (4×). For instance, the average response latency increases by 37% and 31% for the *h263encode* and the *mixed_trace* applications respectively. When it comes to the neural network inference application (*nn_inference_2cores*), the average response latency increases by 82% when the temperature rises from 90°C to 120°C. Note that such a significant increase in the response latency largely degrade the inference performance which cannot be tolerated in real-time image processing that are very frequently executed in the advanced driver assistance systems. Fig. 5(c) presents the average power consumption at different temperatures for every application. Although the power increases gradually with temperature, it rises rapidly after 90°C. The gradual increase in the power from 30°C to 90°C is due to the increase in DRAM operational currents as shown in Fig. 3(a). The more abrupt change in the average power from 90°C to 120°C is attributed to the increased refresh power while performing refresh operations at a much higher rate to compensate retention time reductions induced by temperature. The power consumption of *h263encode* is relatively lower than the other two applications due to its sparse memory access pattern and long execution time. Fig. 5(d) plots the total energy for every application. The increase in the energy is due to the longer execution time caused by the larger response latency and the higher power consumption. As can be noticed, temperature increase considerably increases the total energy and for the case of *stream* application it is larger than the other applications due to its very long execution time.

VII. SUPPRESSING TEMPERATURE EFFECTS IN DRAMs USING THERMOELECTRIC COOLING

A. OUR PROPOSED TEC-BASED COOLING

A Thermoelectric Cooler (TEC) is a solid-state device that consists of several pairs of N-type and P-type semiconductor

[5] We selected MNIST for simplicity. The real-time images from automotive cameras have much higher resolution, and therefore require significantly higher memory bandwidth. The performance impact of refresh will be more severe in those scenarios compared to our sample application.

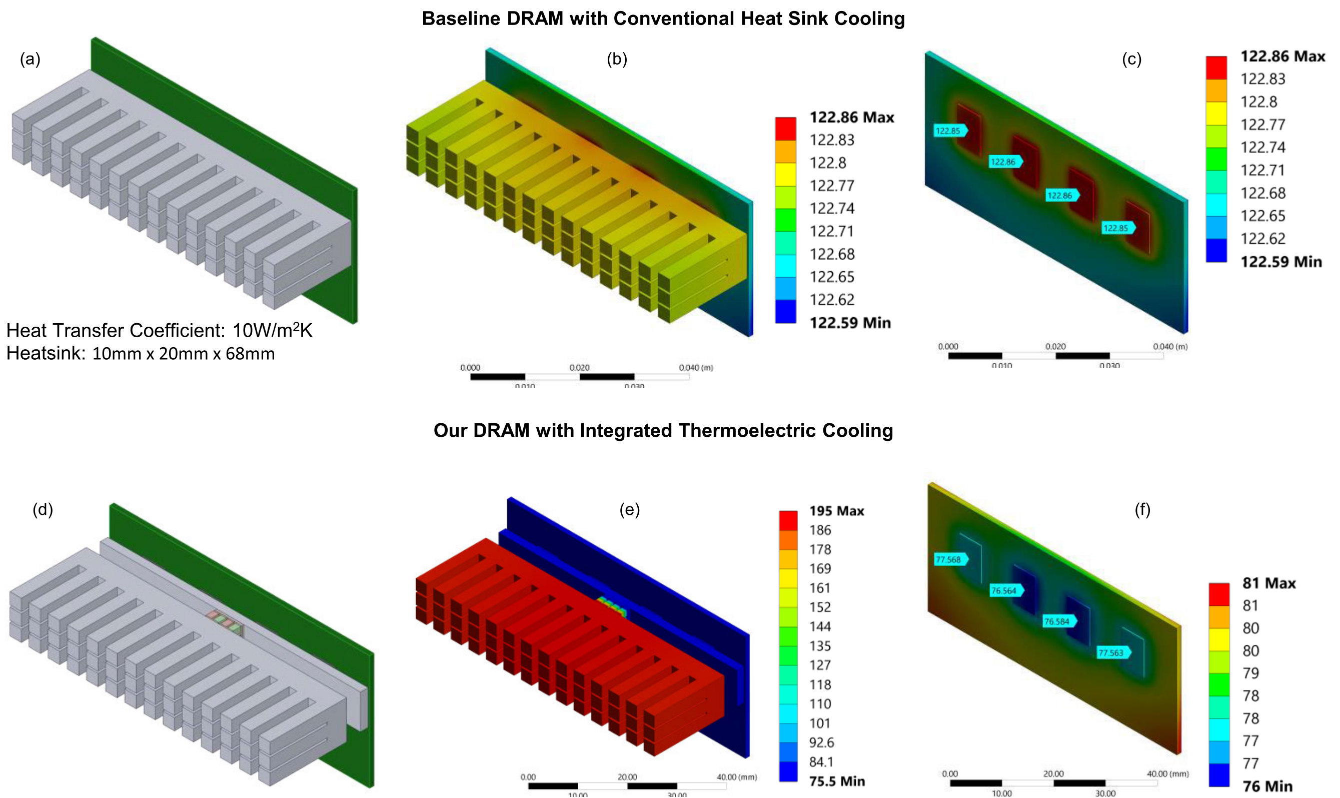

FIGURE 4. Simulated thermal profile using multiphysics simulation tool flows from ANSYS for the entire DRAM DIMM when the original baseline cooling is in use and when our proposed TEC-based cooling is in use. (a, b and c) demonstrate the case when the DRAM DIMM is cooled down using the original cooling that has only a heat sink in which the temperature of the four DRAM dies reach around 122.8°C. (d, e, and f) demonstrate the case of the DRAM DIMM when it is cooled down using a TEC device as we propose showing how the temperature of the four DRAM dies is reduced to just 77.5°C. The heat flux corresponding to the consumed power by the application "neural network inference" is considered in this simulation example. In both cases (baseline cooling and our TEC-based cooling), we consider a Heat Transfer Coefficient (HTC) of merely 10 W/m²K, which represents the convection of *static air* in order to estimate the resulting on-chip temperature under a minimum capability that conventional static air-based cooling provides (i.e., in the absence in any fan that gives air forced-convection).

elements connected electrically in series and thermally in parallel [41]. When electrical current flows through a TEC, Peltier effect takes place in which a thermal gradient across the upper and lower sides of TEC is created. One side of TEC (where heat is absorbed) becomes cold and the other side of TEC (where heat is released) becomes hot. The capability of TEC to dissipate heat depends on several factors, such as the number of N-P couples and the Seebeck coefficient.

In this work, we propose to employ a TEC device to cool down the DRAM chips within one DIMM. We demonstrate how such a cooling solution can effectively keep the maximum temperature of all DRAM chips below 85°C and hence it can be used as countermeasure to the significant reduction in bandwidth and increase in response latency induced by elevated temperatures as explained earlier. This, in turn, open doors to employ commodity off-the-shelf DRAMs in a harsh environment like in automotive electronics in which the operating temperature goes beyond 120°C.

In order to accurately model and investigate the impact of TEC, we employ a commercial multiphysics tool from ANSYS [42] for accurate thermal and thermal-electric analysis, which allows us to simulate complex heat and cooling effects using finite element method. The results of all the thermal simulations done in this work are extracted from steady-state analysis. For a more realistic modeling, we consider a complete system of a full DIMM that consists of four DRAM chips on top of a printed circuit board (PCB).

In the baseline scenario in which a specialized DRAM (i.e, automotive-graded DRAM that may operate up to 125°C) is employed, a heat sink is directly mounted on top of the DRAM DIMM to dissipate the generated heat (see Fig. 4(a)). To illustrate the requirement of a heat-sink in the baseline scenario, we first performed a thermal analysis, comparing the temperature of the DRAM with and without a heat-sink. The analysis showed that the temperature of the DRAM chips increase from 122°C (in the presence of a heat-sink) to above 130°C (in the absence of a heat sink). In this analysis, we have considered a small power consumption in DRAM chips of merely 0.078W. The temperature rise becomes even larger when a higher power consumption (i.e., heat flux) is considered. This demonstrates that even automotive-graded DRAMs will also demand a small heat-sink to operate. Hence, the heat-sink in our cooling solution is not an additional prerequisite.

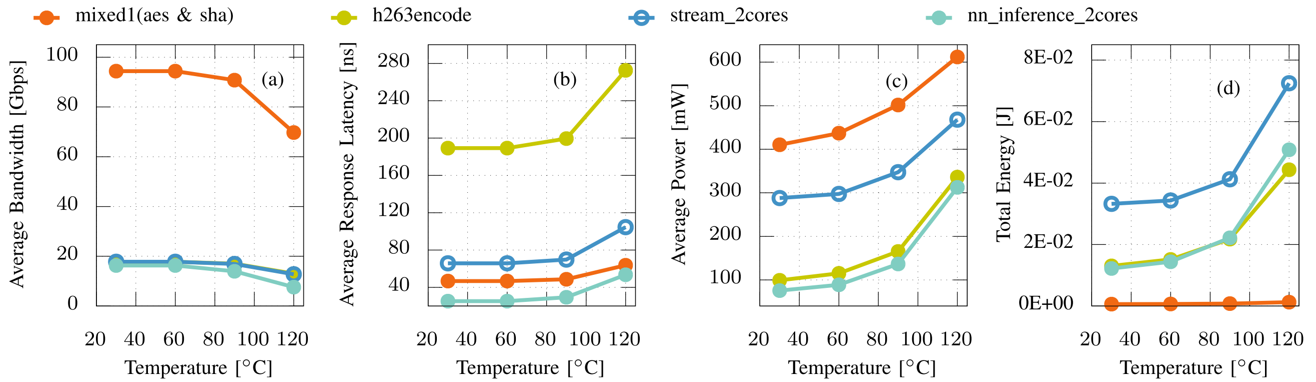

FIGURE 5. (a) Average bandwidth reduction due to temperature increase. (b) Resulting increase in the average response latency due to temperature. (c) Impact of temperature on the average power consumption of DRAM. (d) Impact of temperature rise on increasing the total energy of DRAM akin to the larger response latency and higher average power.

In our proposed TEC-based cooling, a very thin aluminum metal plate is additionally added to cover the DRAM chips and distributes uniformly the cooling effect (generated by the TEC above) among them, as shown in Fig. 4(d). On top of the metal plate, the TEC is attached at the center and the generated heat from the upper side of TEC is dissipated through the heat sink that is attached on top of the TEC. Because our proposed cooling solution still employs the same heat-sink, the only increase in volume that will incur is due to the thermoelectric device itself, which is included between the DRAM and the heat-sink. The thickness of our used TEC device (which is commercially available) is merely 3.2mm. Furthermore, the TEC device can be even inserted inside the base of the heat-sink itself in the case of the incurred increase in the volume (i.e., 3.2mm) could not be tolerated in a very tight area scenario.

To ensure fair comparisons, both baseline cooling and our cooling employ the same heat sink as shown in Fig. 4(a and d) as well as the same Heat Transfer Coefficient (HTC) of merely 10 W/m^{2}K. This low HTC, in fact, represents the convection of *static air* in order to estimate the resulting on-chip temperature in the DRAM DIMM under a minimum capability that conventional static air-based cooling provides, i.e., in the absence in any fan that gives air forced-convection. This necessary to mimic the situation of electronic systems in very tight and confined areas in which available volume is very limited.

The complete simulated system is presented in Fig. 4(a-f). The sizes of all employed parts such as PCB, DRAM chips, etc. have been obtained from existing data sheets of commercial products. Similarly, the TEC device has been carefully modeled in ANSYS replicating a commercial TEC device as provided in the data sheet [43]. In our designed TEC, 20 N- and P-Legs providing 10 TE couples have been considered. In all simulations, the ambient temperature of 120°C is assumed to consider and mimic the DRAM operation in harsh environments like in automotive as the focus on this work.

B. IMAPCT OF OUR PROPOSED TEC-BASED COOLING ON THE FIGURES OF MERIT OF DRAMs

The thermal profiles shown in Fig. 4 demonstrate the impact of our proposed TEC-based cooling on the temperature of DRAM DIMM, while running the neural network inference application. Fig. 4 (a, b and c) show the case when the DRAM DIMM is cooled down using the original cooling that has only a heat sink. Fig. 4 (d, e and f) show the case of the DRAM DIMM when it is cooled down using a TEC device as we propose. In the former case the temperature of DRAM chips lie around 122 °C, and will never be below the ambient temperature (i.e., 120 °C) despite how strong the applied convection is. However, in the latter case when an TEC-based active cooling is applied, the maximum temperature of all four DRAM chips in the DIMM goes below 85°C (see Fig. 4(c), which is much lower than the temperature in the baseline scenario (> 120°C) as shown in Fig. 4(f). It is noteworthy that in both cases (i.e., the baseline cooling and our TEC-based cooling), we consider a Heat Transfer Coefficient (HTC) of merely 10 W/m^{2}K, as mentioned earlier along with the same heat sink. Note that in the case of TEC-based cooling, the heat sink becomes hot with a maximum temperature of around 195°C (Fig. 4(e)) due to the excessive heat generated on the upper side of the TEC device and the very low HTC (i.e., static air). However, although the heat sink becomes very hot, the metal plate attached on top of the four DRAM chips is cold and much lower than the ambient temperature due to the effects of Peltier. This, in turn, effectively reduces the temperature of DRAM chips down to 78°C, as can be seen in Fig. 4(c) and Fig. 4(f). In this example, the TEC device is fed with a 3A current.

Fig. 6(a) highlights the impact of TEC cooling in reducing the temperature of DRAM chips while executing several application benchmarks. The maximum temperature among the four chips has been here reported. As can be noticed, the higher the current fed to the TEC the larger the cooling effect. For all studied benchmarks the maximum temperature has dropped from around to 125°C to below 85°C. In Fig. 7(a, b), we present the impact of TEC cooling on

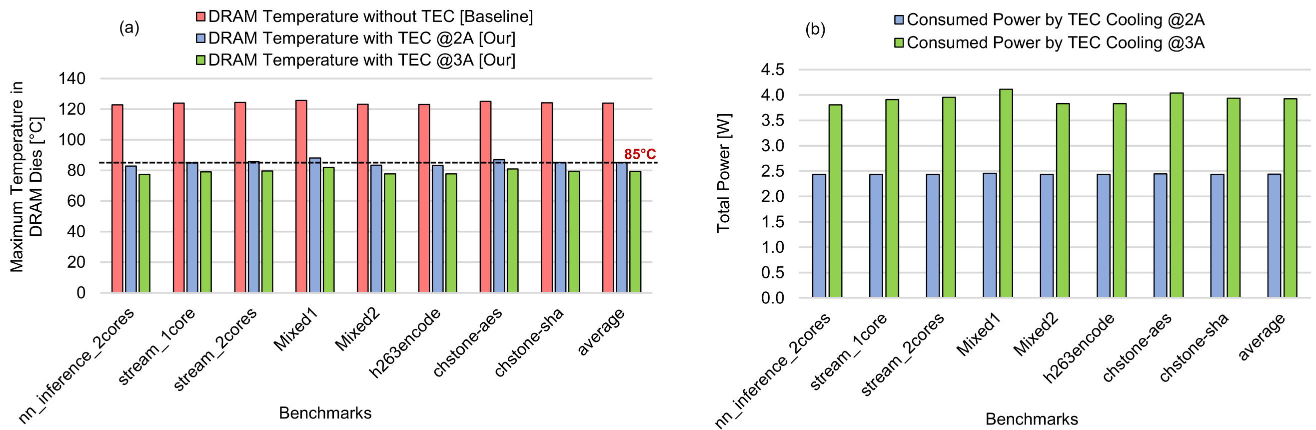

FIGURE 6. (a) Impact of TEC cooling in reducing the temperature of DRAM chips when the TEC device is operated at two different currents. Two different current intensities are examined (2 and 3A). The maximum temperature of DRAM chips is reduced, on average, down to 85°C for the case of 2A and further down to 79°C for the case of 3A. The higher the provided current, the larger the cooling effects and hence the larger the temperature drop. (b) Reports the corresponding power cost of TEC device for each case. The power consumption is estimated using the multiphysics simulations in which the impact of temperature on the internal electrical resistance of TEC is also considered. As expected, the higher the provided current to TEC, the larger the power consumption.

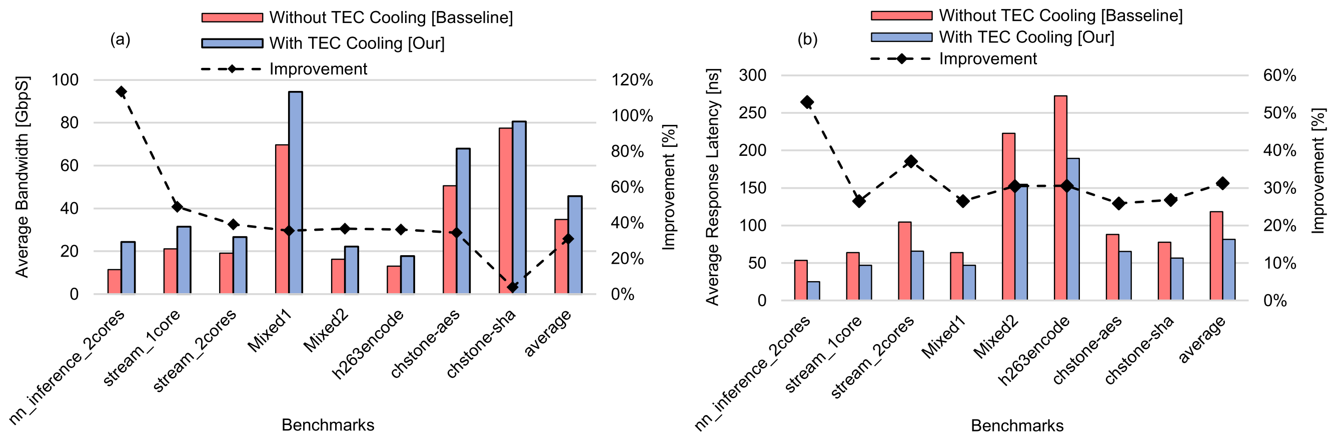

FIGURE 7. (a) Impact of TEC cooling in improving the available bandwidth of DRAM. (b) The impact of TEC cooling on reducing the average response latency of DRAM. As shown, Our TEC-based cooling noticeably increases the average bandwidth as well as reduces the average response time due to the considerable reduction in temperature as presented above in Fig. 6(a).

improving the available bandwidth of DRAM as well as the average response latency for various benchmarks, respectively. As can be noticed in Fig. 7(a), reducing the temperature from above 120°C (as it was the case originally in the baseline DRAM system in the absence of our TEC cooling) to below 85°C (as it is the case in the DRAM system in the presence of our TEC cooling) considerably increases the average bandwidth of all benchmarks. The improvement gain depends on the nature of running benchmark and how much DRAM bandwidth it demands. On average, the bandwidth improvement is about 31% and for some benchmarks such as "neural network inference", the bandwidth improvement is very high reaching 120%. Similarly, reductions in temperature obtained by our TEC cooling also considerably improve the average response latency of DRAM, as demonstrated in Fig. 7(b). The response latency reduction for studied benchmarks is up to 60% and it reaches, on average, 30%.

In Fig. 6(b), we demonstrate the corresponding power cost of TEC device. The power cost has been accurately calculated from the used multiphysics simulations in which the impact of temperature on changing the internal electrical resistance of TEC was also considered. As expected, using a larger current results in a higher power consumption. However, the used current needs to be carefully selected. Increasing the TEC current from 2A to 3A leads to doubling the power cost. However, the corresponding impact and benefit on heat dissipation is not that large. As can be seen in Fig. 6(a), the temperature reduction and benefit from operating the TEC device at 3A instead of 2A is merely around 5°C in which the average temperature of benchmarks drops from 85°C to 79°C. This is because when operating TEC at higher currents, Joule heat effects inside the TEC becomes much stronger and therefore the heat sink attached at the top of the TEC starts to fail in dissipating the generated heat rapidly.

Hence, the overall capability of TEC in cooling the DRAM DIMM becomes much less. As earlier mentioned, we rely in our work on a commercial TEC device, and we have followed the specifications available in the datasheet [43]. Note that the maximum input current of any TEC device is dictated by the generated Joule heat effects inside the TEC device. The latter is subject to 1) the internal resistance of TEC and 2) the available cooling on top of the TEC to dissipate the generated heat induced by Joule heat effects. In our work, we assume a small heat-sink to be attached on top of the TEC device, as shown in Fig. 4(a). This enables the TEC to operate up to 3A input current. Above 3A, the obtained cooling from TEC becomes negligible because Joule heat effects inside TEC become dominant.

Compared to other existing work in the state of the art, the maximum current could be different. For instance, on the one hand, the work in [19] employs the TEC for the scenario of mobile devices in which the available space and volume is extremely limited. Hence, the generated heat from the TEC is dissipated through merely a thin heat spreader and not through a heatsink as in our work. Consequently, the TEC device in [19] could operate only up to around 1A. On the other hand, the work in [20] employs a superlattice TEC and assumes a much larger heat sink to be attached on top of the chip. Superlattice TEC devices feature a very low internal resistance compared to bulk TEC devices. Hence, compared to our work, the used TEC in [20] has a larger heatsink that prevents Joule heat effects from becoming dominant at relatively low currents as well as it features a smaller internal resistance that leads, in general, to lower Joule heat effects. Consequently, the superlattice TEC in [20] was able to operate up to around 7A.

C. DISCUSSION ON THE POWER OVERHEAD

Our proposed Thermoelectric (TEC)-based cooling solution enables the reduction of temperature to below the ambient through using the Peltier effect-induced active cooling. Our investigation using multi-physics simulations demonstrated that the DRAM's temperature using our cooling (for a wide range of applications) is always below 85°C despite the excessive ambient temperature (120°C). Our TEC-based cooling comes with an additional power as reported in our evaluation. Importantly, this additional power is only required when the temperature exceeds the 85°C, which is the specified temperature above which commodity DRAMs will start to exhibit higher error rates. Hence, the additional power overhead is not always paid and the TEC cooling works on-demand (i.e., the TEC is activated only when it is needed). Therefore, although the additional power of (2.4W – 3.9W on average as shown in Fig. 6(b)) might seem too high, in reality the average cooling power is much lower since the cooling has only to be activated if the ambient temperature is too high (i.e., above 85°C). Furthermore, such additional power could be tolerated in systems like automotive as a tradeoff with 1) cost reduction, 2) bandwidth increase, and 3) reliability guarantee, which are primary optimization goals that cannot be compromised. It is noteworthy that the power consumption of advanced electronic control units (ECUs) like TESLA's FSD compute unit is in the order of $70 - 80$W [44]. Hence, such an additional power (3W - 6W consumed by the TEC device) is affordable.

In the case of a varying ambient temperature, the available temperature sensors will provide the thermal management unit with the current ambient temperature in order to perform on-demand cooling. In practice, our TEC-based cooling needs to be activated only when the temperature exceeds the specification of the commodity DRAM chip (e.g., >85°C). Therefore, under a varying ambient temperature, there is no need to continuously turn on the TEC and instead on-demand cooling can be applied once the ambient temperature exceeds a certain predetermined threshold.

All in all, the TEC-based cooling solution opens doors, for the first time, to employ commodity DRAMs in harsh environments in which neither reliability nor bandwidth is sacrificed.

VIII. CONCLUSION

In this work, we investigated thoroughly the impact of excessive temperatures on DRAMs. Using accurate measurements obtained from commercial DRAM devices, we demonstrated how increasing the temperature beyond the nominal operating range (i.e., >85°C) significantly reduces the available bandwidth as well as increases the average response latency. For many domains, such as safety-critical applications in automotive, such induced degradations cannot be tolerated due to the necessity of meeting hard real-time deadlines.

Using accurate multiphysics simulations, we demonstrated how Thermoelectric-based cooling can ensure that the maximum temperature of all DRAM chips within one DIMM to be below 85°C despite that the original temperature was above 120°C. This opens door for the first time to still use off-the-shelf commercial DRAMs (which were supposed to operate within nominal temperature ranges, i.e., <85°C) in harsh environments (where the operating temperature is above 120°C), while tight reliability and timing constraints are still fulfilled. This, in turn, illuminates for the automotive industry, the large costs associates with the need for using expensive DRAMs that are fabricated and designed specifically to tolerate excessive temperatures while the probability of error is below $1E-9$.

ACKNOWLEDGMENT

Deepak M. Mathew and Hammam Kattan contributed equally to this work. The work of Hammam Kattan was conducted at KIT, from 2018 to 2020. The work of Hussam Amrouch was done in part at KIT during 2020.

REFERENCES

[1] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020, doi: 10.1002/rob.21918.

[2] A. Spessot, R. Ritzenthaler, E. D. Litta, E. Dupuy, B. O'Sullivan, J. Bastos, E. Capogreco, K. Miyaguchi, V. Machkaoutsan, Y. Yoon, P. Fazan, and N. Horiguchi, ''80 nm tall thermally stable cost effective FinFETs for advanced dynamic random access memory periphery devices for artificial intelligence/machine learning and automotive applications,'' *Jpn. J. Appl. Phys.*, vol. 60, no. SB, May 2021, Art. no. SBBB06.

[3] G. Boschi, E. Spano, H. Grigoryan, A. Kumar, and G. Harutyunyan, ''Die-to-die testing and ECC error mitigation in automotive and industrial safety applications,'' in *Proc. IEEE Int. Test Conf. (ITC)*, Nov. 2020, pp. 1–6.

[4] M. Jung, S. A. McKee, C. Sudarshan, C. Dropmann, C. Weis, and N. Wehn, ''Driving into the memory wall: The role of memory for advanced driver assistance systems and autonomous driving,'' in *Proc. Int. Symp. Memory Syst.*, Oct. 2018, pp. 377–386.

[5] S. Woo, ''Autonomous vehicles: Memory requirements & deep neural net limitations,'' Rambus, Sunnyvale, CA, USA, White Paper, 2019. [Online]. Available: https://www.rambus.com/blogs/autonomous-vehicles-memory-requirements-deep-neural-net-limitations/

[6] M. Greenberg, ''Understanding Automotive DDR DRAM,'' Synopsys, Mountain View, CA, USA, White Paper, 2017. [Online]. Available: https://www.synopsys.com/designware-ip/technical-bulletin/automotive-ddr-dram.html

[7] *Road Vehicles—Environmental Conditions and Testing for Electrical and Electronic Equipment—Part 4: Climatic Loads*, 3rd ed, document ISO 16750-4, 2010.

[8] *AEC-Q100—Rev-G, Failure Mechanism Based Stress Test Qualification for Integrated Circuits*, document, Automot. Electron. Council-Compon. Tech. Commitee, New York, NY, USA, 2007.

[9] Micron. (2021). *8Gb Automotive DDR4 SDRAM (MT40A1G8)*. Accessed: Feb. 2021. [Online]. Available: https://www.micron.com/-/media/client/global/documents/products/data-sheet/dram/ddr4/8gb_ddr4_sdram.pdf

[10] Samsung. (2021). *Specifications of 32Gb Automotive LPDDR4 SDRAM (K4FBE3D4HM-GUCL)*. Accessed: Feb. 2021. [Online]. Available: https://www.samsung.com/semiconductor/dram/lpddr4/K4F6E3S4HM-GUCL/

[11] K. Kraft, D. M. Mathew, C. Sudarshan, M. Jung, C. Weis, N. Wehn, and F. Longnos, ''Efficient coding scheme for DDR4 memory subsystems,'' in *Proc. Int. Symp. Memory Syst.*, Oct. 2018, pp. 148–157.

[12] *JEDEC Standard, DDR4 SDRAM*, Standard JESD79-4B, JEDEC, 2017.

[13] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu, ''RAIDR: Retention-aware intelligent DRAM refresh,'' in *Proc. 39th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2012, pp. 1–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=2337159.2337161

[14] I. Bhati, M.-T. Chang, Z. Chishti, S.-L. Lu, and B. Jacob, ''DRAM refresh mechanisms, penalties, and trade-offs,'' *IEEE Trans. Comput.*, vol. 65, no. 1, pp. 108–121, Jan. 2016.

[15] *DDR4 SDRAM (JESD 79-4)*, Jedec Solid State Technology Association, Arlington County, VA, USA, 2012.

[16] Y. Dai, T. Li, B. Liu, M. Song, and H. Chen, ''Exploiting dynamic thermal energy harvesting for reusing in smartphone with mobile applications,'' in *Proc. 23rd Int. Conf. Architectural Support Program. Lang. Operating Syst.*, Mar. 2018, pp. 243–256.

[17] (2019). *Smartphones Online Comparsion*. Accessed: Sep. 15, 2019. [Online]. Available: https://www.gsmarena.com

[18] F. Kaplan, S. Reda, and A. K. Coskun, ''Fast thermal modeling of liquid, thermoelectric, and hybrid cooling,'' in *Proc. 16th IEEE Intersoc. Conf. Thermal Thermomech. Phenomena Electron. Syst. (ITherm)*, May 2017, pp. 726–735.

[19] Y. Lee, E. Kim, and K. G. Shin, ''Efficient thermoelectric cooling for mobile devices,'' in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2017, pp. 1–6.

[20] H. Amrouch, G. Zervakis, S. Salamin, H. Kattan, I. Anagnostopoulos, and J. Henkel, ''NPU thermal management,'' *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 11, pp. 3842–3855, Nov. 2020.

[21] K. Kim and J. Lee, ''A new investigation of data retention time in truly nanoscaled DRAMs,'' *IEEE Electron Device Lett.*, vol. 30, no. 8, pp. 846–848, Aug. 2009.

[22] C.-H. Lin, D.-Y. Shen, Y.-J. Chen, C.-L. Yang, and M. Wang, ''SECRET: Selective error correction for refresh energy reduction in DRAMs,'' in *Proc. IEEE 30th Int. Conf. Comput. Design (ICCD)*, Sep. 2012, pp. 67–74.

[23] P. J. Nair, D.-H. Kim, and M. K. Qureshi, ''ArchShield: Architectural framework for assisting DRAM scaling by tolerating high error rates,'' in *Proc. 40th Annu. Int. Symp. Comput. Archit.*, Jun. 2013, pp. 72–83, doi: 10.1145/2485922.2485929.

[24] J. Liu, B. Jaiyen, Y. Kim, C. Wilkerson, and O. Mutlu, ''An experimental study of data retention behavior in modern DRAM devices: Implications for retention time profiling mechanisms,'' *ACM SIGARCH Comput. Archit. News*, vol. 41, no. 3, pp. 60–71, Jun. 2013, doi: 10.1145/2508148.2485928.

[25] J. Lucas, M. Alvarez-Mesa, M. Andersch, and B. Juurlink, ''Sparkk: Quality-scalable approximate storage in DRAM,'' in *The Memory Forum*. N.A., Jun. 2014. [Online]. Available: http://www.redaktion.tu-berlin.de/fileadmin/fg196/publication/sparkk2014.pdf

[26] C. Weis, M. Jung, P. Ehses, C. Santos, P. Vivet, S. Goossens, M. Koedam, and N. Wehn, ''Retention time measurements and modelling of bit error rates of WIDE I/O DRAM in MPSoCs,'' in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2015, pp. 495–500.

[27] M. Patel, J. S. Kim, and O. Mutlu, ''The reach profiler (REAPER): Enabling the mitigation of DRAM retention failures via profiling at aggressive conditions,'' in *Proc. 44th Annu. Int. Symp. Comput. Archit.*, Jun. 2017, pp. 255–268.

[28] M. Jung, D. M. Mathew, C. C. Rheinlander, C. Weis, and N. Wehn, ''A platform to analyze DDR3 DRAM's power and retention time,'' *IEEE Des. Test. IEEE Des. Test. Comput.*, vol. 34, no. 4, pp. 52–59, Aug. 2017.

[29] D. M. Mathew, M. Schultheis, C. C. Rheinlander, C. Sudarshan, C. Weis, N. Wehn, and M. Jung, ''An analysis on retention error behavior and power consumption of recent DDR4 DRAMs,'' in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 293–296.

[30] M. Jung, C. Weis, N. Wehn, and K. Chandrasekar, ''TLM modelling of 3D stacked wide I/O DRAM subsystems: A virtual platform for memory controller design space exploration,'' in *Proc. Workshop Rapid Simulation Perform. Eval. Methods Tools (RAPIDO)*, 2013, pp. 5:1–5:6, doi: 10.1145/2432516.2432521.

[31] M. Jung, C. Weis, and N. Wehn, ''DRAMSys: A flexible DRAM subsystem design space exploration framework,'' *IPSJ Trans. Syst. LSI Des. Methodol.*, vol. 8, pp. 63–74, Aug. 2015.

[32] K. Chandrasekar, B. Akesson, and K. Goossens, ''Improved power modeling of DDR SDRAMs,'' in *Proc. 14th Euromicro Conf. Digit. Syst. Design*, Aug. 2011, pp. 99–108.

[33] D. M. Mathew, É. F. Zulian, S. Kannoth, M. Jung, C. Weis, and N. Wehn, ''A bank-wise DRAM power model for system simulations,'' in *Proc. 9th Workshop Rapid Simulation Perform. Eval., Methods Tools*, Jan. 2017, pp. 1–7, doi: 10.1145/3023973.3023978.

[34] Y. Hara, H. Tomiyama, S. Honda, and H. Takada, ''Proposal and quantitative analysis of the CHStone benchmark program suite for practical C-based high-level synthesis,'' *J. Inf. Process.*, vol. 17, pp. 242–254, Oct. 2009.

[35] M. Consortium. (2015). *Mediabench*. [Online]. Available: http://euler.slu.edu/fritts/mediabench/

[36] T. Austin, E. Larson, and D. Ernst, ''SimpleScalar: An infrastructure for computer system modeling,'' *Computer*, vol. 35, no. 2, pp. 59–67, 2002, doi: 10.1109/2.982917.

[37] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, ''The gem5 simulator,'' *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, May 2011, doi: 10.1145/2024716.2024718.

[38] J. D. McCalpin, ''Memory bandwidth and machine balance in current high performance computers,'' in *Proc. IEEE TCCA Newslett.*, Dec. 1995, pp. 19–25.

[39] A. Carter. *Mnist Neural Network in C*. Accessed: 2020. [Online]. Available: https://github.com/AndrewCarterUK/mnist-neural-network-plain-c

[40] Samsung. (2016). *Datasheet 4Gb E-die DDR4 SDRAM (K4A4G165WE)*. Rev. 1.4. Accessed: Jun. 2016. [Online]. Available: https://www.samsung.com/semiconductor/dram/ddr4/K4A4G165WE-BCRC/

[41] M. J. Dousti and M. Pedram, ''Platform-dependent, leakage-aware control of the driving current of embedded thermoelectric coolers,'' in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, Sep. 2013, pp. 311–316.

[42] (2019). *Ansys Software Version 19.1*. Accessed: 2020. [Online]. Available: https://www.ansys.com

[43] T. Device. *Txl-127-03l*. Accessed: 2020. [Online]. Available: https://txlgroup.com/wp-content/uploads/2018/06/TXL-127-03L_data_sheet.pdf

[44] Auto Pilot Review. (2019). *Tesla Hardware 3 (Full Self-Driving Computer) Detailed*. [Online]. Available: https://www.autopilotreview.com/tesla-custom-ai-chips-hardware-3/

[45] H. Kattan, S. W. Chung, J. Henkel, and H. Amrouch, ''On-demand mobile CPU cooling with thin-film thermoelectric array,'' *IEEE Micro*, early access, Feb. 23, 2021, doi: 10.1109/MM.2021.3061335.

DEEPAK M. MATHEW (Member, IEEE) is currently pursuing the Ph.D. degree in electrical and computer engineering with the Microelectronic Systems Design Research Group, TU Kaiserslautern, Germany. His research interests include main memory subsystems, heterogeneous memories, memory controller hardware design, and emerging nonvolatile memory technologies.

HAMMAM KATTAN received the B.S. degree in mechanical engineering from Aleppo University, in 2011, and the M.S. degree in mechanical engineering from the Karlsruhe Institute of Technology (KIT), in 2017. From 2018 to 2020, he worked as a Research Assistant with the Chair for Embedded Systems (CES), KIT. His research interests include advanced cooling techniques, thermoelectrics, and modeling and simulation using the finite element method (FEM).

CHRISTIAN WEIS (Member, IEEE) received the Ph.D. degree in electrical engineering from TU Kaiserslautern, Germany, in 2014. From 1998 to 2009, he was with Siemens Semiconductor, Infineon Technologies AG, and Qimonda AG, Munich, Germany, in DRAM design. During this time frame, he was involved in DRAM design for graphics and commodity DRAM products. Since 2009, he has been with the Microelectronic System Design Research Group, TU Kaiserslautern. He holds more than 20 patents related to DRAMs and DRAM design, and published more than 60 articles. His current research interests include DRAM controller design, near- & in-memory processing, 3D-integrated DRAMs, heterogeneous memory architectures, physical design, and MPSoCs.

JÖRG HENKEL received the Diploma and Ph.D. *(summa cum laude)* degree from the Technical University of Braunschweig. He is currently the Chair Professor of embedded systems with the Karlsruhe Institute of Technology. Before that he was a Research Staff Member with NEC Laboratories, Princeton, NJ, USA. His research interest includes co-design for embedded hardware/software systems with respect to power, thermal and reliability aspects. He has led several conferences as a General Chair, including ICCAD and ESWeek, and serves as a steering committee chair/member for leading conferences and journals for embedded and cyber-physical systems. He 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. 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 also the Editor-in-Chief of the *IEEE Design & Test* magazine and is/has been an associate editor for major ACM and IEEE journals.

NORBERT WEHN (Senior Member, IEEE) received the Diploma and Ph.D. degrees from the TU Darmstadt, Germany. He currently holds the Chair for Microelectronic Systems Design, Department of Electrical and Computer Engineering, TU Kaiserslautern, Germany. He has more than 350 publications in various fields of microelectronic system design and holds 20 patents. His research interests include VLSI architecture for mobile communication, forward error correction techniques, low-power techniques, advanced SoC and memory architectures, 3-D integration, reliability issues in SoC, the IoT, and hardware accelerators for big data applications. He is a member of several scientific industrial advisory boards. He is an associate editor of various journals.

HUSSAM AMROUCH (Member, IEEE) received the Ph.D. degree *(summa cum laude)* from the Karlsruhe Institute of Technology (KIT), Germany, in 2015. He is currently a Junior Professor with the Semiconductor Test and Reliability (STAR) Chair, Computer Science, Electrical Engineering Faculty, University of Stuttgart, as well as a Research Group Leader with KIT. He has more than 115 publications (including 45 journals) 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. His main research interests include 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, such as DAC'16, DAC'17, and DATE'17 for his work on reliability. He also serves as an 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, and ICCAD. He served as a reviewer in many top journals, such as IEEE Transactions on Electron Devices, IEEE Transactions on Circuits and Systems—I: Regular Papers, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, and IEEE Transactions on Computers.