HERMES-Core-A 1.59-TOPS/mm^{2}PCM on 14-nm CMOS In-Memory Compute Core Using 300-ps/LSB Linearized CCO-Based ADCs

Khaddam-Aljameh, Riduan; Stanisavljevic, Milos; Fornt Mas, Jordi; Karunaratne, Geethan; Brandli, Matthias; Liu, Feng; Singh, Abhairaj; Muller, Silvia M.; Egger, Urs; More Authors

DOI
10.1109/JSSC.2022.3140414

Publication date
2022

Document Version
Final published version

Published in
IEEE Journal of Solid-State Circuits

HERMES-Core—A 1.59-TOPS/mm^2 PCM on 14-nm CMOS In-Memory Compute Core Using 300-ps/LSB Linearized CCO-Based ADCs

Riduan Khaddam-Aljameh, *Graduate Student Member, IEEE*, Milos Stanisavljevic, *Member, IEEE*, Jordi Fornt Mas, Geethan Karunaratne, *Graduate Student Member, IEEE*, Matthias Brändli, Feng Liu, Abhairaj Singh, *Graduate Student Member, IEEE*, Silvia M. Müller, *Senior Member, IEEE*, Urs Egger, Anastasios Petropoulos, *Graduate Student Member, IEEE*, Theodore Antonakopoulos, *Senior Member, IEEE*, Kevin Brew, Samuel Choi, Injo Ok, Fee Li Lie, *Member, IEEE*, Nicole Saulnier, Victor Chan, Ishtiaq Ahsan, Vijay Narayanan, *Senior Member, IEEE*, S. R. Nandakumar, Manuel Le Gallo, *Member, IEEE*, Pier Andrea Francese, *Senior Member, IEEE*, Abu Sebastian, *Senior Member, IEEE*, and Evangelos Eleftheriou, *Life Fellow, IEEE*

***Abstract*— We present a 256 × 256 in-memory compute (IMC) core designed and fabricated in 14-nm CMOS technology with backend-integrated multi-level phase change memory (PCM). It comprises 256 linearized current-controlled oscillator (CCO)-based A/D converters (ADCs) at a compact 4-μm pitch and a local digital processing unit (LDPU) performing affine scaling and ReLU operations. A frequency-linearization technique for CCO is introduced, which increases the maximum CCO frequency beyond 3 GHz, while ensuring accurate on-chip matrix–vector multiplications (MVMs). Moreover, the design and functionality of the digital ADC calibration procedure is described in detail and the MVM accuracy is quantified. Finally, the measured classification accuracies of deep learning (DL) inference applications on the MNIST and CIFAR-10 datasets, when two IMC cores are employed, are presented. For a performance density of 1.59 TOPS/mm^2, a measured energy efficiency of 10.5 TOPS/W, at a main clock frequency of 1 GHz, is achieved.**

***Index Terms*— Analog computing, deep learning, in-memory computing, phase-change memory.**

Manuscript received June 29, 2021; revised November 12, 2021; accepted December 23, 2021. Date of publication January 28, 2022; date of current version March 28, 2022. This article was approved by Guest Editor Borivoje Nikolić. This work was supported by the IBM Research AI Hardware Center. The work of Riduan Khaddam-Aljameh, Geethan Karunaratne, and Abu Sebastian was supported by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Program under Grant 682675. *(Corresponding authors: Riduan Khaddam-Aljameh; Abu Sebastian.)*

Riduan Khaddam-Aljameh, Milos Stanisavljevic, and Evangelos Eleftheriou were with IBM Research Europe, 8803 Rüschlikon, Switzerland. They are now with Axelera AI, 8038 Zürich, Switzerland (e-mail: riduank@student.ethz.ch).

Jordi Fornt Mas was with IBM Research Europe, 8803 Rüschlikon, Switzerland. He is now with the Barcelona Supercomputing Center, 08034 Barcelona, Spain.

Geethan Karunaratne, Matthias Brändli, Urs Egger, S. R. Nandakumar, Manuel Le Gallo, Pier Andrea Francese, and Abu Sebastian are with IBM Research Europe, 8803 Rüschlikon, Switzerland (e-mail: ase@zurich.ibm.com).

Feng Liu, Kevin Brew, Samuel Choi, Injo Ok, Fee Li Lie, Nicole Saulnier, Victor Chan, and Ishtiaq Ahsan are with IBM Research, Albany, NY 12203 USA.

Abhairaj Singh was with IBM Research Europe, 8803 Rüschlikon, Switzerland. He is now with the Department of Computer Engineering, Delft University of Technology, 2628 CD Delft, The Netherlands.

Silvia M. Müller is with IBM Systems and Technology, 71034 Boeblingen, Germany.

Anastasios Petropoulos and Theodore Antonakopoulos are with the Department of Electrical and Computers Engineering, University of Patras, 26504 Rio-Patras, Greece.

Vijay Narayanan is with IBM Research, Yorktown, NY 10598 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2022.3140414.

Digital Object Identifier 10.1109/JSSC.2022.3140414

I. Introduction

IN-MEMORY computing (IMC) is an emerging non-von Neumann paradigm where computation is performed in the memory array itself [1], [2]. Basically, the conventional memory systems are endowed with in-place computing capabilities, thus eliminating the back and forth shuttling of data between the memory and processing units, which costs energy and latency. Hence, data-centric applications that predominantly use a small set of computational operations can benefit greatly from the novel IMC paradigm. A prominent example is AI applications, and, in particular, deep-neural network inference, where matrix–vector multiplications (MVMs) dominate the workload [3].

In order to accelerate the execution of MVM operations using IMC, the memory system must be repurposed into a single instruction multiple data (SIMD) array of processing elements [4], where the input vector data is broadcasted across the matrix rows and the various partial products are summed up along a column. Standard CMOS logic-based solutions, using multi-bit multipliers and adders, can qualify for IMC operations, but only at an area and latency penalty [5], [6]. Instead, analog processing is used due to its scalability and its ability

to encode multi-bit data in a single physical quantity, such as time [7], [8], electrical current [9]–[12], charge [13]–[16], or voltage [17], [18].

The choice of the underlying memory technology for IMC ranges from conventional memory types, such as SRAM, DRAM, and Flash, to emerging memristive devices, such as metal-oxide-based resistive random access memory (ReRAM) and chalcogenide-based phase change memory (PCM). Compared to ReRAM, PCM device physics is much better understood [19]. It is expected that the volumetric switching in PCM may lead to substantially better array-level variability. PCM has also been commercialized as storage-class memory [20] and embedded memory [21]. Both PCM and ReRAM permit reliable long-term storage of multi-bit quantities in the conductance value of a single device. Subsequently, by employing Ohm's and Kirchhoff's laws, it is possible to perform MVM operations at $\mathcal{O}(1)$ time complexity. This could lead to high throughput and highly energy-efficient systems [22], with the only disadvantage being the time- and energy-consuming programming procedures corresponding to these memristive devices.

Although experimental results on ReRAM-based IMC systems have already been demonstrated [23]–[25], complete IMC systems based on PCM crossbar arrays had been lacking till recently [26], [27]. Prior to this, most of the demonstrations have been based on either simulation studies based on the measured characteristics of individual devices or on experiments based on PCM memory chips that were re-purposed for IMC operations [28]–[30].

One of the main challenges faced during the realization of an IMC system is that the peripheral circuits, especially data converters that interface the crossbar array with the digital world, carry the largest energy overhead and could even dominate the associated latency and area footprint. In addition, voltage-based A/D converters (ADCs) are mostly used [31] that require a voltage to current conversion, usually employing a large capacitor for integration [23], [32]. This has thus far hampered the realization of large fully-parallel on-chip MVM operations at true $\mathcal{O}(1)$ complexity. Besides MVM, neural network applications require a range of other mathematical operations to implement activation functions or to aggregate the results from layers split across crossbar arrays.

In this article, we present a more detailed description of the PCM-based HERMES core which was presented at the VLSI symposium [26]. A schematic overview of HERMES core is presented in Fig. 1. It comprises compact, low-latency, and energy-efficient current-controlled oscillator (CCO)-based ADCs, digital readout blocks, and a local digital processing unit (LDPU) performing affine scaling and ReLU operations.

The remainder of the article is organized as follows. In Section II, we present a new unit-cell design and a crossbar array of such unit cells that supports the storage of signed weights, parallel programming at $\mathcal{O}(N)$ complexity, and the execution of signed MVM in one step. Section III discusses the proposed CCO-based ADC design that employs a linearization technique for the output frequency and that supports built-in shift-and-add operations. Furthermore, the LDPU architecture and implementation is presented. In Section IV, the performance of the analog IMC MVM operation is analyzed for different input modulation schemes. Section V describes the inference applications on the MNIST and CIFAR-10 datasets, hardware-aware training and the experimental results thereof. It also presents a comparison of the proposed PCM-based core with other state-of-the-art IMC designs. Finally, Section VI concludes the article.

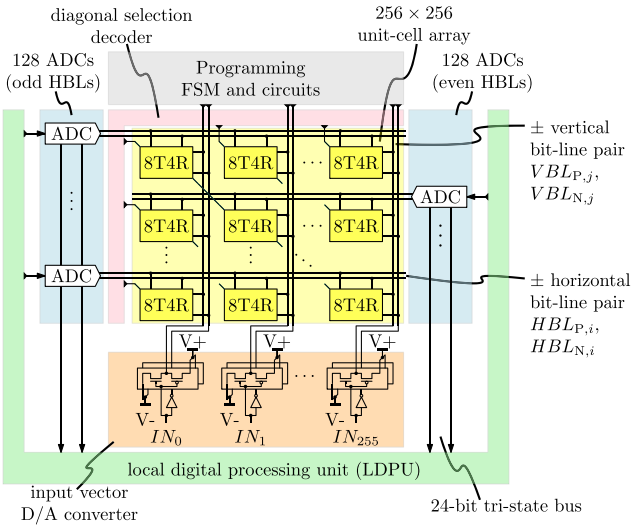

Fig. 1. System overview of the HERMES core. The memory element array (yellow) in the center consists of 256 × 256 8T4R PCM unit cells. The cells are initialized via the programming circuitry (gray), located on top. During an MVM, the input vector is applied as a read voltage pulse via the input modulator (orange) on the vertical bit-lines (VBLs) of the crossbar, while the 256 ADC (blue) digitize the flowing current. The generated ADC outputs arrive via two 24-bit tri-state buses at the local digital processing unit (LDPU), where they are post-processed.

II. Unit Cell and Array Design

A single PCM core is capable of performing a fully parallel analog MVM with 256 8-bit digital inputs at $\mathcal{O}(1)$-complexity. Central to its architecture is an array of 256 × 256 8T4R unit cells. As shown in Fig. 2(a), positive weights are represented by the combined conductance value of two mushroom-type PCM devices G_1^+ and G_2^+, whereas negative weights are represented by the other two PCM devices G_1^- and G_2^-. To support the high voltages beyond 2.5 V that are required for PCM programming, a pair of stacked nMOS devices is used as a selector. The selection signals SEL_1 and SEL_2 are routed diagonally across the array, to ensure uniform current distributions across the horizontal (HBL) and vertical (VBL) bit-lines during parallel write operations [33]. The read procedure for MVMs that involves the parallel read of all the unit cells is executed as shown in Fig. 2(b). Initially, all selection signals are enabled, since the full matrix needs to be active. Also, the HBLs are pulled to the common-mode voltage V_{cm}. Then, based on the input vector signs, the different VBLs are connected to either

$$V_- = V_{\text{cm}} - V_{\text{read}} \tag{1}$$

when current from the respective columns is to be added, or

$$V_+ = V_{\text{cm}} + V_{\text{read}} \tag{2}$$

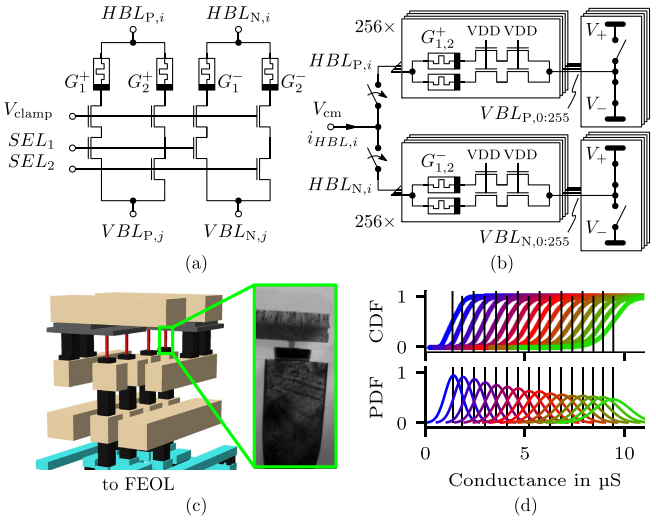

Fig. 2. (a) 8T4R unit-cell schematic. Devices G_1^+ and G_2^+ encode positive weight values and G_1^- and G_2^- negative values. (b) Full array read procedure during MVM for one full row. (c) PCM device insertion point in the upper part of the back-end-of-the-line (BEOL) stack. (d) Cells' storage characteristics obtained from all 256×256×4 PCM devices when encoding 16 representative levels. Note that the number of programmable levels is not limited but instead, a level-dependent standard-deviation of the programming error is encountered.

to subtract current, by applying a negative ΔV across the PCM device. This allows all four combinations of ±inputs and ±weights to be taken into account in a single MVM-step without the need for negative voltages. Fig. 2(c) illustrates the insertion point of the PCM device in the upper part of the back-end-of-the-line (BEOL) metal stack. Therefore, the obtained cell footprint does not yet demonstrate the full potential of the PCM on 14-nm CMOS technology, which will materialize when the insertion point is placed in immediate proximity of the transistors, as done in past technologies [34]. Representative storage characteristics of the employed mushroom PCM device are shown in Fig. 2(d) for 16 distinct levels. The analog nature of the device, however, allows the encoding of more levels, the only limit being ADC precision and allowable programming time [35], [36].

III. Linearized CCO-Based ADC

Unlike the more commonly used voltage ADCs for IMC, time-based current ADCs eliminate the need for additional conversion cycles and are amenable to trading off precision with latency. Furthermore, since large current integration capacitors are avoided and mostly digital circuits are used, this approach facilitates having one converter per column of the crossbar, thus minimizing the overall latency as no resource sharing will be required.

A. CCO-Based ADC Structure

Fig. 3 shows the ADC structure which has been implemented in the HERMES core. While the input D/A converter (DAC) is applying the input data to the VBLs, the generated crossbar currents arrive via the HBL wires first to Class-AB read voltage regulators, as depicted in Fig. 3(a). This type of regulator [37] can keep the HBL potentials on common-mode V_{cm} irrespective of the current polarity. By connecting to a Schmitt trigger, this polarity information is captured in the signal D.

The mirrored HBL current is then fed into the second part of the ADC, which is the CCO [see Fig. 3(b)], to generate a proportional time-encoded signal. Fig. 3(d) illustrates the waveforms of the various signals within the CCO for different amplitudes of i_{HBL} as well as different polarities. The oscillation is controlled by two small capacitors C_1 and C_2 that are alternately charged or discharged until either of their voltages V_{C1} or V_{C2} reaches the threshold voltage v_{th} of the connected cross-coupled inverter pair. Its flipping also toggles the latch state signal A and thus ultimately digitizes the flow of a fixed amount of charge Q_{unit} into the circuit. Based on the signal A, either the first or the second of the two symmetric slices in the oscillator operates, which ensures that always the correct side of the latch toggles. Furthermore, based on the polarity of the incoming HBL current, the signal D controls whether the integration capacitors C_1 and C_2 are discharged until the threshold of pMOSs P_1 or P_2 is reached, or are charged until the nMOSs N_1 or N_2 become active.

Finally, the oscillating signal A is forwarded to the last stage of the ADC, which is the dual 12-bit ripple-counter [see Fig. 3(c)], serving as an integrator for the time-encoded current information. Thus, the frequency f_{CCO} of signal A is captured by the positive and negative counter outputs ADC_P and ADC_N, which are incremented at a rate proportional to the current i_{HBL}.

B. Linearization Technique

A known issue in this CCO circuit is the limited linearity of the output frequency f_{CCO} at high input currents [38]–[40]. This is due to the constant gate delay, t_{delay}, between the time instance when the latch toggles up to the time instance when the current integration proceeds on a second capacitor. t_{delay} is added to the inverse of the oscillator output frequency, which is the time period T_{CCO}, thus interfering with the linear relationship between frequency f_{CCO} and current i_{HBL}

$$T_{CCO} = 1/f_{CCO} = \frac{C_1 \cdot v_{th}}{\alpha \cdot i_{HBL}} + t_{delay}. \tag{3}$$

The solutions proposed in literature range from restriction to low-frequency operation, where the delay is not dominant [40], to extensive digital post-processing using look-up tables [41], digital filters, or other feedback structures [39]. In an IMC system, these solutions would not be ideal, as they incur either significant area or latency penalties. Moreover, the post-processing operations aimed at compensating the nonlinearity only work for dc current measurements and would fail when working with time-varying currents that are integrated over a period of time. In this article, feed-forward compensation is proposed as a solution for the delay-induced nonlinearity issue in CCOs. The underlying idea consists of adapting the threshold voltage to compensate for the delay. This approach is different from the solution presented in [38], which modifies the reference voltage of an attached comparator circuit. Here, the trip-point $v_{th,c}$ of the cross-coupled latch is reduced based

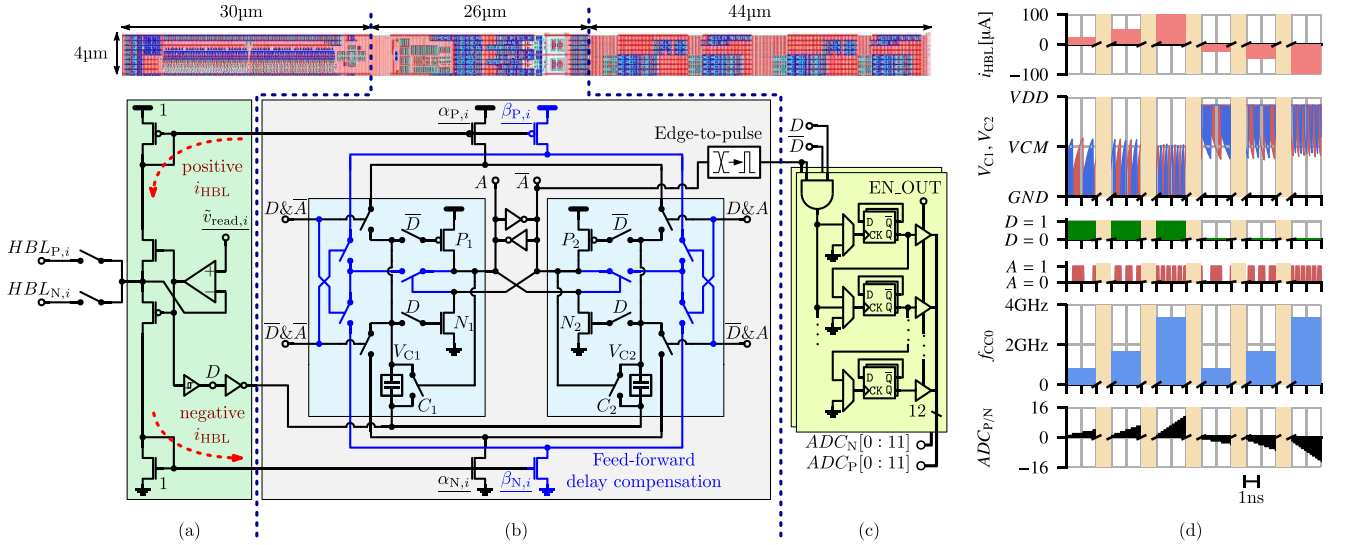

Fig. 3. Circuit diagram and layout of the CCO-based ADC that is used in the HERMES core. (a) Class-AB regulation stage that keeps the HBL on V_{cm} independent of the current polarity. In (b), CCO is shown. It consists of a cross-coupled latch, two pMOS and nMOS current mirrors, and two capacitive integrator stages. Control inputs per ADC for the main pMOS/nMOS current mirror gains ($\alpha_{P/N,i}$), the feed-forward compensation current mirror gains ($\beta_{P/N,i}$), and the trimmed local read voltage ($\tilde{v}_{read,i}$) are underlined. The output signal $\bar{A}$ is fed into an edge-to-pulse converter and forwarded to the last stage, that is, (c) dual 12-bit ripple counter. (d) Waveforms during the operation of the CCO are shown for different HBL current amplitudes ($i_{HBL} = i_{HBL,P} + i_{HBL,N}$).

on the instantaneous current amplitude by using a second current mirror (see Fig. 3(b) in blue)

$$v_{th,c}(i_{HBL}) \approx v_{th}^{*} - t_{delay} \cdot \frac{\alpha \cdot i_{HBL}}{C_1}. \tag{4}$$

As a result, the latch flips by a predefined amount of time earlier, which can be adjusted to compensate for the delay. In this case, the current to frequency relation remains linear even for high current amplitudes

$$T_{CCO} = 1/f_{CCO} = \left(v_{th} = v_{th,c}\right) \approx \frac{C_1 \cdot v_{th}^{*}}{\alpha \cdot i_{HBL}}. \tag{5}$$

Furthermore, this feed-forward compensation technique allows counteracting other saturation effects that appear at the peak oscillator frequency by tweaking the transfer curve through active overcompensation. Such effects include, for example, the read voltage drop due to limited interconnection resistance and regulator output impedance. The drawbacks of this approach include an increased energy consumption and an initial calibration overhead, as the correct compensation gain needs to be set. Moreover, there is an increased area penalty, albeit moderate.

Each ADC measures the outputs of a multi-input–single-output (MISO) system, that is the result of a dot-product between a weight vector and an input vector. Therefore, the adjusted metrics of weight- and input-integral nonlinearity (INL)/differential nonlinearity (DNL) are adopted as proposed in [16]. The measured transfer curves that are shown in Fig. 4 are obtained by fixing one quantity (weight or input) while sweeping the other. Both curves remain bounded within ±1 LSB, thus indicating the absence of significant deterministic errors. Moreover, these INL/DNL plots illustrate the efficacy of the compensation technique. The linear region of the transfer curve in Fig. 4(a) is increased, and the maximum oscillation frequency can exceed 3 GHz.

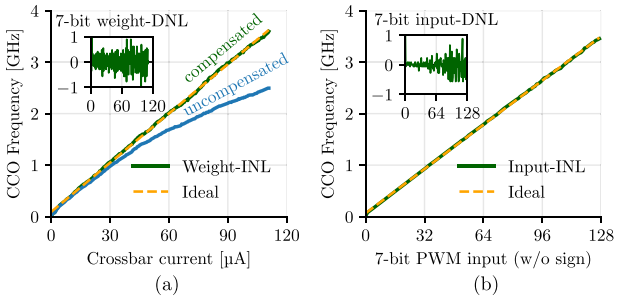

Fig. 4. Measured CCO frequency sweeps. (a) Weight-integral nonlinearity (INL) that is obtained by fixing the input value and varying the number of activated unit cells from 0 to a maximum of 256, such that full designated current range is covered. (b) Contains the input-INL measured for a sweep of the input value while all weights are active. The insets in both graphs contain the differential nonlinearity (DNL).

C. Counter With Variable Increment Size

In the last stage of the CCO-based ADC, the oscillating signal A is integrated in the digital domain using a ripple counter. By selecting the appropriate D flip-flop that receives A, the increment size of the counter can be made variable. This allows the execution of shift-and-add operations within the ADC at a minimal overhead, avoiding dedicated multi-bit adders [42]–[44]. Hence, we enabled bit-serial input modulation in addition to the conventional multi-bit pulsewidth modulation (PWM). The negative ADC output ADC_N that is read from the counter after one integration period T_{int} can be formulated as

$$ADC_N = \left\lfloor \frac{1}{Q_{unit}} \int_0^{T_{int}} f(i_{HBL}(t))dt \right\rfloor \tag{6}$$

$$f(i_{HBL}(t)) = \begin{cases} |i_{HBL}(t)|, & \text{if } i_{HBL}(t) \le 0 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

For the positive output ADC_P, it is vice versa.

D. ADC Calibration Procedure

In order to ensure the best accuracy possible for applications using IMC cores, the errors introduced by gain and mismatch variations are to be kept as small as possible. Given the large number of ADCs that are involved in a large-scale IMC system, an accurate and robust calibration procedure is of high importance.

Besides inaccurate MVM results, the absence of mismatch calibration can also lead to an increased power consumption, due to currents flowing in between ADCs. This happens when the various Class-AB read-voltage regulators settle on different V_{read} values. If the HBLs are on different potentials, then given the crossbar topology, current will flow between them, which manifests itself in an ADC offset and in a decreased energy efficiency. If this is added to the original CCO transfer function (3), a simplified equation for the current-to-frequency relation can be formulated that models all relevant effects

$$f_{\text{CCO}}(i_{\text{HBL}}) = \frac{A_{\text{dc}} \times i_{\text{HBL}}}{1 + B_{\text{NL}} \times i_{\text{HBL}}} + C_{\text{offset}}. \tag{8}$$

Therein, the three variables A_{dc}, B_{NL}, and C_{offset} characterize static gain, nonlinearity, and offset for each ADC. In the employed design, they can be adjusted through three separate types of control inputs [see Fig. 3(a) and (b)]. Main current-mirror gains $\alpha_{P,i}$, $\alpha_{N,i}$ set the static gain A_{dc}. The nonlinearity-related term B_{NL} is reduced by increasing the feed-forward gains $\beta_{P,i}$, $\beta_{N,i}$, and, finally, any read-voltage variation-related offset is compensated using $\tilde{v}_{\text{read},i}$, which is varied by selecting a different tap from a resistor ladder.

Moreover, there is some interrelation between the different parameters. For example, the read-voltage setting $\tilde{v}_{\text{read},i}$ impacts the static gain and the two current mirrors for static gain and feed-forward compensation also exhibit some correlation. The calibration algorithm therefore commences with fixing the offset, then proceeds to setting the static gain, and finally enables and adjusts the feed-forward compensation.

Equation (8) contains three unknown variables that change depending on the calibration settings. They are calculated by generating i_{HBL} currents of three different amplitudes and storing of the measured counter values. To accurately capture the complete transfer curve characteristics, the three current values are chosen so that they include one low and one medium current point as well as one measurement at the largest relevant current amplitude that can be realistically expected in the crossbar. In this case, a value of $I_{\text{HBL,max}} = 100\ \mu\text{A}$ was chosen. The obtained equation system is then solved to obtain the three parameters A_{dc}, B_{NL}, and C_{offset}. By averaging over several measurements, non-systematic disturbances like thermal noise can be filtered out. Note that this process of calculating the three parameters needs to be repeated following each control input change in order to closely monitor the calibration progress. Moreover, positive and negative current mirrors must be calibrated separately.

The course of a measured calibration procedure is shown in Fig. 5. Initially, the individual read voltage values $\tilde{v}_{\text{read},i}$ are continuously adjusted [see Fig. 5(a)] until the relative offset errors of all ADCs are reduced to almost 0% [see Fig. 5(b)].

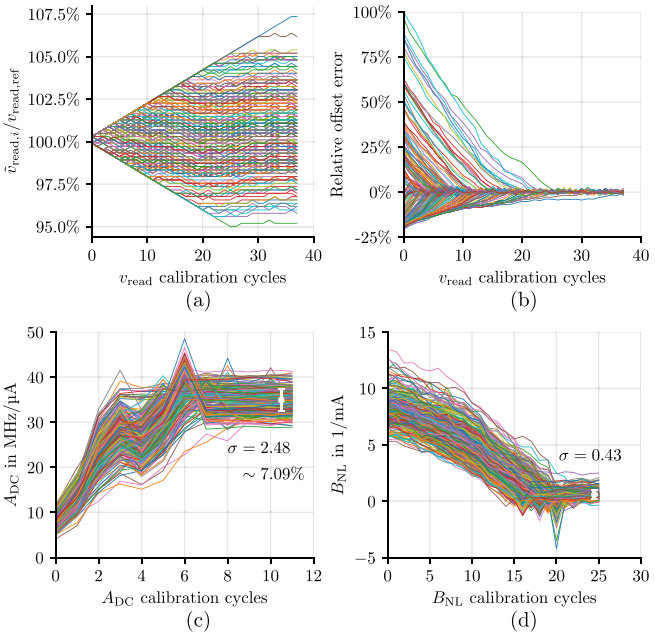

Fig. 5. Calibration procedure for the 256 CCO-based ADCs per HERMES core. (a) Evolution of the per ADC adjustable read-voltage $\tilde{v}_{\text{read},i}$ during calibration until the relative offset errors between the ADCs are eliminated, as shown in (b). In (c), adjustment of the static ADC gain is shown and in (d), nonlinearity-related coefficient B_{NL} is minimized.

Afterward, the static gain A_{dc} is set to ca. 35 (MHz/μA) [see Fig. 5(c)]. Despite the 4-bit gain adjustment circuitry, there is a residual spread of $\sigma = 2.48$ (MHz/μA), corresponding to 7.09% of the static gain reference value. Finally, the nonlinearity-related term B_{NL} that originates from switching delays and voltage drop is reduced to the lowest value possible [see Fig. 5(d)].

E. Local Digital Processing Unit (LDPU)

At the end of the ADC calibration process, the offset differences of the 256 ADCs are effectively eliminated, while the static gain variations are still distributed within $\pm 21\%$ of the reference value. Equalization of the remaining difference, which in the analog domain would require significant silicon area, is performed digitally in the LDPU, which is efficiently combined with any deep learning (DL) inference-required affine scaling. This is done after the left and right blocks of 128 ADCs pass their raw 12-bit positive (ADC_P) and 12-bit negative (ADC_N) output data via two separate 24-bit tri-state buses to the LDPU [see Fig. 6(a)].

Therein, the 2×12 bit data pass through the convert-and-scale blocks [see Fig. 6(b)] that each contains two FP16 multiply–add units. These units apply the scaling and offset factors and also subtract positive and negative ADC values. Moreover, the LDPU supports the combination of results from layers split across multiple crossbars and can also perform residual input additions that are needed for executing ResNets. The INT8 outputs of the LDPU can be transferred to other cores.

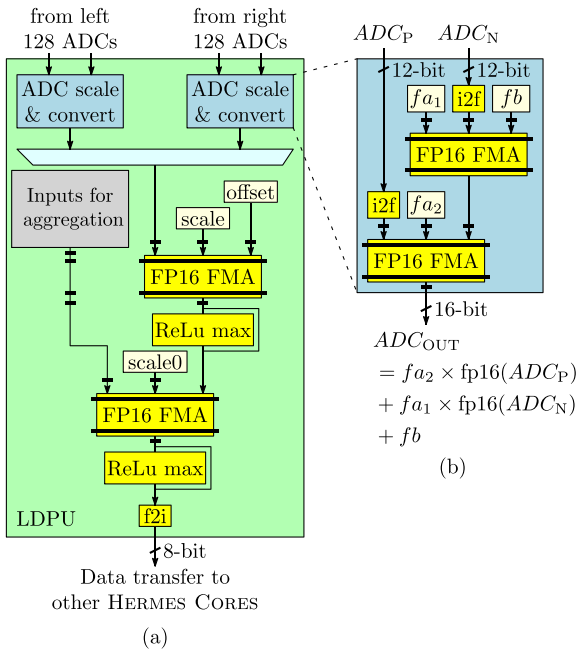

Fig. 6. (a) Block diagram of the LDPU, which receives the output data from the 256 ADCs and post-processes them in the dedicated ADC scale&convert blocks, that are shown in (b). There, remaining gain and offset variations between the different ADCs are equalized. The LDPU can furthermore apply the ReLu activation function and combine results from different HERMES cores.

IV. MVM Operation

We propose a hardware-centric approach to characterize the quality of the analog MVM operation. Using the crossbar's duality of being an MVM engine as well as a cascaded D/A and A/D system, a statistical accuracy metric can be created. First, a set of ideal fixed-point MAC results $\vec{y}$ is defined that spans the full output range. In a second step, random FP32 weight ($\vec{w}$) and INT8 input vectors ($\vec{x}$) are created, as indicated in Fig. 7(a), using a uniform multinomial distribution, so that their dot-products $\vec{w} \times \vec{x}$ yield $\vec{y}$. The weights $\vec{w}$ are then programmed into the crossbar by means of iterative programming [35]. Next, the inputs $\vec{x}$ can be applied as read-voltage pulses using either multi-bit or bit-serial modulation with shift-and-add. Thus, the analog MVM is performed and the resulting HBL currents are digitized in the CCO and finally post-processed in the LDPU, yielding the hardware-obtained MVM-results $\vec{y}_{HW}$. These INT8 results are then compared against the FP32 reference values, as shown in Fig. 7(b) and (c). Finally, the computational accuracy can be characterized by calculating the standard deviation of the error [see Fig. 7(d) and (e)].

A. Multi-Bit and Bit-Serial Input Modulation

Using the HERMES core, the two supported modulation schemes are examined by comparing their respective INT8 LDPU outputs against the ideal FP32 results. The MVM results shown in Fig. 7(b) are obtained by using conventional 8-bit PWM. Following a brief VBL pre-charge procedure, the CCO-based ADCs are activated and continuously integrate the time-varying current while the modulator is active.

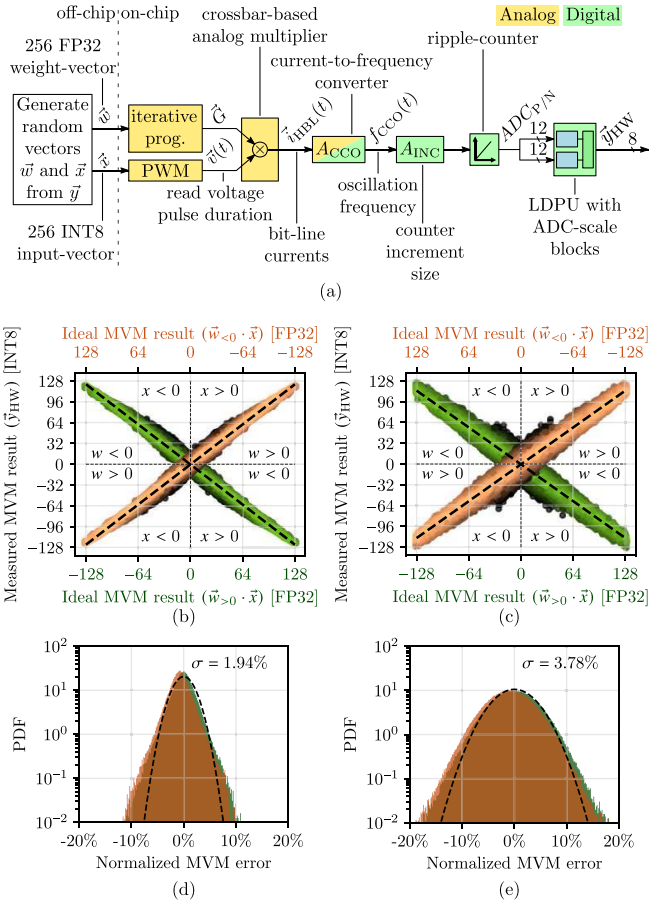

Fig. 7. (a) Experimental flow diagram that is used to determine the MVM accuracy. The measured results are shown in (b) for conventional multi-bit PWM and in (c) for bit-serial input modulation with shift-and-add. Both ±weights and inputs are used to cover all four quadrants. The probability density function (PDF) of the normalized MVM error for both modulation schemes is plotted in (d) and (e).

In the bit-serial case, however, only static currents are measured. Therein, the modulator applies the seven magnitude bits of each entry in the read-voltage vector $\vec{x}$ from LSB to MSB to the crossbar, so that dc currents develop. After a defined settling time, these currents are measured by enabling the ADCs for a constant time period and integration of the CCO output in the attached counter. The significance of the applied input-bit is taken into account by selecting which of the first seven counter bits to increment. Thus, a shift and add operation is realized in the counter. The obtained MVM results are shown in Fig. 7(c).

In the employed conventional multi-bit modulation scheme, an integration time of 128 ns is used, given by the 1-GHz operation frequency of the modulator and the 7-bit input magnitude. Due to the peak CCO frequency of ca. 3.3 GHz, more than 420 different charge levels can be quantized and, thus, a resulting resolution of more than 8 bit is ensured. The measured MVM results in Fig. 7(b) and (d) do not show any significant systematic errors and the error distribution, if modeled as a random normal distributed Gaussian, yields a $\sigma = 1.94\%$.

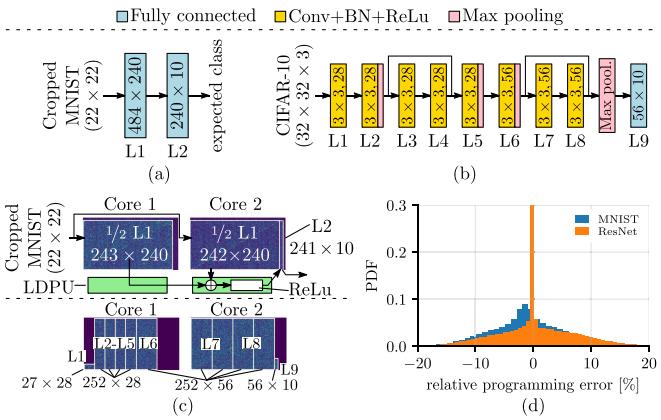

Fig. 8. Implemented network applications. (a) Two-layer MLP that is used for MNIST image classification and (b) ResNet-9 network used for CIFAR-10 image classification. (c) Layer map of MNIST and ResNet-9 networks within two HERMES cores each. (d) PDF of the relative weight programming error for the MNIST and ResNet-9 networks.

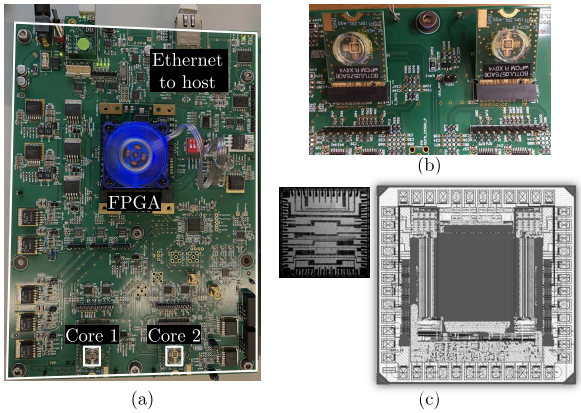

Fig. 9. (a) FPGA-based test board with two PCB-packaged HERMES cores that was used to run the neural network applications. (b) Die holder sockets and a bonded HERMES cores on holder boards with a plastic lid. (c) Chip micrograph and snapshot from the EDA tool.

Regarding the bit-serial modulation scheme, the maximum intermediate result per input bit must be limited to 4 bit to avoid saturation or overflow in the 12-bit counters. This is achieved by limiting the integration time per input-magnitude bit to ca. 5 ns. As a consequence, only the upper 4 bit of the final result remains usable, while the remaining bits mostly consist of quantization noise. This is also reflected in the results shown in Fig. 7(c) and (e), which show a broader error histogram with twice the standard deviation than for the multi-bit case.

Therefore, in the remainder of this article, the conventional multi-bit modulation scheme will be used for the application studies. However, note that the encountered precision limit of bit-serial modulation is specific to the presented design and could be easily circumvented through adjustments to the counter.

V. Applications and Results

For experimental validation of the inference performance, a two-layer MLP [see Fig. 8(a)] and a ResNet-9 network [see Fig. 8(b)] were trained to perform MNIST or CIFAR-10 image classification, respectively. In both cases, the networks were designed so that the weights can be mapped onto the two HERMES cores, which constitute the employed demo system, as is shown in Fig. 9.

A. Hardware-Aware Training

Prior to mapping the networks onto analog IMC cores, it is essential to perform a hardware-aware custom training in software as described in [30]. Due to device variability and noise, the networks need to be trained in a specific way so that transferring the digitally trained weights to the analog PCM devices will not result in significant loss of accuracy.

1) Two-Layer MLP: For the two-layer MLP case, the MNIST input images were cropped to 22 × 22 and scaled between 0 and 1. The hidden layer of the MLP network comprises 240 neurons and employs the ReLU activation function, while the output layer comprises ten neurons and a softmax-operation step. Initially, the network was trained in floating-point arithmetic to minimize cross entropy loss with a batch size of 4 and initial learning rate of 0.0005 for 50 epochs. The weights and biases were clipped at 3.5 times the standard deviation σ_W of the weight distributions. The learning rate was reduced by half after every 20 epochs. Furthermore, the network was retrained while adding 15% noise to the weights for 35 epochs and clipping of the weights and biases at $3 \times \sigma_W$ after every training batch. Following this training approach, the network achieved a 98.6% accuracy on the 10 000 cropped MNIST image test set.

2) ResNet-9: In the second application example, a ResNet-9 with 1 00 726 trainable parameters is used for CIFAR-10 image classification, as shown in Fig. 8(b). The training images were normalized to have zero mean and unit standard deviation. In addition, the images are augmented by random cropping to 32 × 32 squares after padding by four pixels on each boarder and random horizontal flipping. The network was first trained in floating-point arithmetic with a batch size of 200 and initial learning rate of 0.5 for 300 epochs of 50 000 training images. The learning rate was reduced to one-tenth after every 30 epochs. As for the previous application, the network was further retrained by adding 6.67% noise to the weights for 270 epochs and clipping of the weights and biases at $2 \times \sigma_W$ after every training batch. Eventually, the network achieved an 88.4% accuracy on the 10 000 test images of CIFAR-10.

B. Hardware Experiment

The trained weights of all layers were then mapped on two HERMES cores and iteratively programmed on the PCM unit-cell arrays. Fig. 8(c) shows the layer mapping of both networks on the two cores and Fig. 8(d) shows the resulting weight programming error. The iterative programming convergence rate was 100% for the MLP and 99.9% for ResNet-9. The relative weight programming error standard deviations of 4.8% (MLP) and 5.3% (ResNet-9) are related to the difficulty in reading accurately the individual conductance values with the on-chip ADC when iteratively programming them. This could be mitigated by increasing the ADC resolution, using

TABLE I
COMPARISON TABLE OF RECENT ANALOG IMC-BASED MVM/MULTIPLY-ACCUMULATE (MAC)-OPERATION ACCELERATORS

Metric	This work	ISSCC'21 [45]	ISSCC'21 [44]	ISSCC'20 [46]
CMOS technology	14 nm	22 nm	16 nm	7 nm
Memory technology	PCM	ReRAM	SRAM	SRAM
Non-volatile	Yes	Yes	No	No
Operating Voltage in V	0.8	0.8	0.8	0.8
Operation Frequency	1 GHz	-	200 MHz	-
ADC design	CCO-based ADC	Sense amplifier	8bit SAR ADC	4bit Flash-ADC
Memory size	65.5 K	4 M	4.5 MB	4 KB
Unit-cell	8T4R	1T1R	10T1C	8T
Number of input/weight/output-bits	8b/Analog/8b	8b/8b/14b	4b/4b/8b	4b/4b/4b
Peak Throughput (TOPS)	1.008	0.035	11.8 | 5.90[1]	0.372 | 0.186[1]
Energy Efficiency (TOPS/W)	10.5	11.91	121 | 60.5[1]	351 | 175.5[1]
Area Efficiency (TOPS/mm^2)	1.59	0.013	2.67 | 1.34[1]	116.3 | 58.13[1]

[1] normalized to 8b input

drift-mitigation schemes, or increasing of the maximum number of iterations during programming.

After the weights were programmed, the inference experiment was conducted by providing as input the test images to the two-core platform. For the MNIST MLP experiment, the test images were flattened and split over cores 1 and 2 to process the first layer. The partial MVM output from core 1 was sent to the LDPU of core 2, added to the output of core 2 via the LDPU circuitry, after which the ReLU was performed on the summed output. The resulting LDPU output from the first layer was then input to core 2 for processing the second layer and to obtain the final classification result. Hence, all the inference operations for processing the MNIST test images were performed on-chip. The FPGA was used solely for control and data propagation between the cores. The resulting on-chip inference accuracy obtained on MNIST was 98.3%, which is only 0.3% lower than the software accuracy obtained after training.

For the ResNet-9 inference experiment, all the MVMs required for performing the convolutions on the CIFAR-10 images were performed on-chip. The FPGA was used to send data from one layer to the next after each layer was processed by the HERMES cores. The pooling operations, not supported by the LDPU, were performed in software. The experimentally obtained accuracy was 85.6%, which is less than 3% lower than the software accuracy. Although this accuracy difference is higher than that observed for the MNIST dataset, it is expected that the ResNet-9 on CIFAR-10, with its inherent low number of parameters, will be more sensitive to weight programming errors and also additional errors introduced by the ADC conversion. Reducing the weight programming errors and the ADC nonlinearity would be needed to bridge the accuracy gap between experiment and software. Using a larger network in addition would improve the overall robustness to these nonidealities.

C. System-Level Performance

At an operation frequency of 1 GHz and a supply voltage of 0.8 V, the HERMES core shows a peak throughput of 1.008 TOPS at an efficiency of 10.5 TOPS/W, when running the MNIST-based experiment as described above. Compared to the state-of-the-art (shown in Table I), the measured throughput density of 1.59 TOPS/mm^2 is significantly higher than recent non-volatile ReRAM-based designs [24], [45] and also slightly higher than recent SRAM + capacitor-based designs [44], when 8-bit input quantization is used. Only the SRAM-based design in [46] shows a better throughput density, given by its compact 8T SRAM unit-cell design employing push-rules and the advanced manufacturing node. However, it uses a lower precision, 4 bit-based computation mode and offers no persistence throughout power cycling due to the volatile SRAM cells.

VI. CONCLUSION

In this article, an IMC-based MVM accelerator is presented that uses a novel PCM on 14-nm CMOS process. The compact CCO-based ADCs allow the MVM operation to be executed at $\mathcal{O}(1)$-complexity, since the pitches of ADC and unit-cell match. Through linearization of the ADC's current-to-frequency transfer curve, a resolution of 300 ps per LSB is demonstrated. Thus, a system performance of 1.008 TOPS at an energy efficiency of 10.5 TOPS/W is achieved.

Furthermore, the usage of a time-based ADC architecture allowed efficient implementation of shift-and-add operations within the ADC, thus obviating the need for dedicated digital adders. In addition, this enables bit-serial input modulation to be implemented natively. Both supported input modulation schemes, conventional multi-bit PWM and bit-serial modulation with shift-and-add, are compared with respect to the achievable MVM operation precision. Finally, the suitability for DNN applications is demonstrated through successful on-chip implementation of a two-layer MLP and a ResNet-9 for MNIST and CIFAR-10 image classification, respectively.

The presented system demonstrates the feasibility of a high-throughput IMC MVM accelerator system using non-volatile memristive devices. Despite the integration of the PCM element at a relatively high position in the metal stack, a throughput density of 1.59 TOPS/mm^2 is achieved,

comparable to existing SRAM-based accelerators. It is conceivable that by integrating the PCM devices closer to the transistor-level at a denser pitch, substantially higher throughput density can be achieved.

ACKNOWLEDGMENT

The authors would like to thank G. Burr, M. Dazzi, J. Demarest, Y. Kohda, P. Adusumilli, C. Goldberg, L. Clevenger, W. Haensch, R. Haas, K. Hosokawa, A. Curioni, J. Burns, R. Divakaruni, L. Benini, and M. Khare for technical and management support.

REFERENCES

[1] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 529–544, Mar. 2020.

[2] B. Crafton, S. Spetalnick, Y. Fang, and A. Raychowdhury, "Merged logic and memory fabrics for accelerating machine learning workloads," *IEEE Des. Test. Comput.*, vol. 38, no. 1, pp. 39–68, Feb. 2021.

[3] M. Kang, S. K. Gonugondla, and N. R. Shanbhag, "Deep in-memory architectures in SRAM: An analog approach to approximate computing," *Proc. IEEE*, vol. 108, no. 12, pp. 2251–2275, Dec. 2020.

[4] H. T. Kung, "Why systolic architectures?" *Computer*, vol. 15, no. 1, pp. 37–46, Jan. 1982.

[5] H. Kim, T. Yoo, T. T.-H. Kim, and B. Kim, "Colonnade: A reconfigurable SRAM-based digital bit-serial compute-in-memory macro for processing neural networks," *IEEE J. Solid-State Circuits*, vol. 56, no. 7, pp. 2221–2233, Jul. 2021.

[6] Y.-D. Chih *et al.*, "16.4 An 89TOPS/W and 16.3TOPS/mm^2 all-digital SRAM-based full-precision compute-in memory macro in 22 nm for machine-learning edge applications," in *IEEE ISSCC Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 252–254.

[7] Y. Toyama, K. Yoshioka, K. Ban, S. Maya, A. Sai, and K. Onizuka, "An 8 bit 12.4 TOPS/W phase-domain MAC circuit for energy-constrained deep learning accelerators," *IEEE J. Solid-State Circuits*, vol. 54, no. 10, pp. 2730–2742, Oct. 2019.

[8] M. Yamaguchi, G. Iwamoto, Y. Nishimura, H. Tamukoh, and T. Morie, "An energy-efficient time-domain analog CMOS binaryconnect neural network processor based on a pulse-width modulation approach," *IEEE Access*, vol. 9, pp. 2644–2654, 2021.

[9] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[10] A. Jaiswal, I. Chakraborty, A. Agrawal, and K. Roy, "8T SRAM cell as a multibit dot-product engine for beyond von Neumann computing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 11, pp. 2556–2567, Nov. 2019.

[11] S. K. Gonugondla, A. D. Patil, and N. R. Shanbhag, "SWIPE: Enhancing robustness of ReRAM crossbars for in-memory computing," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design*, Nov. 2020, pp. 1–9.

[12] G. W. Burr *et al.*, "Neuromorphic computing using non-volatile memory," *Adv. Phys. X*, vol. 2, no. 1, pp. 89–124, 2017.

[13] M. Kang, S. K. Gonugondla, A. Patil, and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 642–655, Feb. 2018.

[14] S. K. Gonugondla, M. Kang, and N. R. Shanbhag, "A variation-tolerant in-memory machine learning classifier via on-chip training," *IEEE J. Solid-State Circuits*, vol. 53, no. 11, pp. 3163–3173, Nov. 2018.

[15] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.

[16] R. Khaddam-Aljameh, P.-A. Francese, L. Benini, and E. Eleftheriou, "An SRAM-based multibit in-memory matrix-vector multiplier with a precision that scales linearly in area, time, and power," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 2, pp. 372–385, Feb. 2021.

[17] M. F. Ali, A. Jaiswal, and K. Roy, "In-memory low-cost bit-serial addition using commodity DRAM technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 1, pp. 155–165, Jan. 2020.

[18] V. Seshadri *et al.*, "Ambit: In-memory accelerator for bulk bitwise operations using commodity DRAM technology," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2017, pp. 273–287.

[19] M. Le Gallo and A. Sebastian, "An overview of phase-change memory device physics," *J. Phys. D, Appl. Phys.*, vol. 53, no. 21, May 2020, Art. no. 213002.

[20] P. Fantini, "Phase change memory applications: The history, the present and the future," *J. Phys. D, Appl. Phys.*, vol. 53, no. 28, May 2020, Art. no. 283002.

[21] F. Arnaud *et al.*, "High density embedded PCM cell in 28 nm FDSOI technology for automotive micro-controller applications," in *IEDM Tech. Dig.*, Dec. 2020, pp 24.2.1–24.2.4.

[22] P. Houshmand *et al.*, "Opportunities and limitations of emerging analog in-memory compute DNN architectures," in *IEDM Tech. Dig.*, Dec. 2020, pp. 29.1.1–29.1.4.

[23] Q. Liu *et al.*, "33.2 A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2020, pp. 500–502.

[24] C.-X. Xue *et al.*, "15.4 A 22 nm 2 Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit MAC computing for tiny AI edge devices," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2020, pp. 244–246, iSSN: 2376-8606.

[25] J. Yue *et al.*, "14.3 A 65 nm computing-in-memory-based CNN processor with 2.9-to-35.8TOPS/W system energy efficiency using dynamic-sparsity performance-scaling architecture and energy-efficient inter/intra-macro data reuse," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2020, pp. 234–236.

[26] R. Khaddam-Aljameh *et al.*, "HERMES core—A 14 nm CMOS and PCM-based in-memory compute core using an array of 300ps/LSB linearized CCO-based ADCs and local digital processing," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2021, pp. 1–2.

[27] P. Narayanan *et al.*, "Fully on-chip MAC at 14 nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format," *IEEE Trans. Electron Devices*, vol. 68, no. 12, pp. 6629–6636, Dec. 2021.

[28] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Feb. 2015.

[29] A. Sebastian *et al.*, "Computational memory-based inference and training of deep neural networks," in *Proc. Symp. VLSI Technol.*, Jun. 2019, pp. T168–T169.

[30] V. Joshi *et al.*, "Accurate deep neural network inference using computational phase-change memory," *Nature Commun.*, vol. 11, no. 1, p. 2473, May 2020.

[31] P. Yao *et al.*, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, 2020.

[32] M. Ali, A. Jaiswal, S. Kodge, A. Agrawal, I. Chakraborty, and K. Roy, "IMAC: In-memory multi-bit multiplication and accumulation in 6T SRAM array," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 8, pp. 2521–2531, Aug. 2020.

[33] R. Khaddam-Aljameh *et al.*, "A multi-memristive unit-cell array with diagonal interconnects for in-memory computing," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 12, pp. 3522–3526, Dec. 2021.

[34] G. F. Close *et al.*, "A 256-Mcell phase-change memory chip operating at 2+bit/cell," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 6, pp. 1521–1533, Jun. 2013.

[35] N. Papandreou *et al.*, "Programming algorithms for multilevel phase-change memory," in *Proc. ISCAS*, May 2011, pp. 329–332.

[36] S. R. Nandakumar *et al.*, "Precision of synaptic weights programmed in phase-change memory devices for deep learning inference," in *IEDM Tech. Dig.*, Dec. 2020, pp. 29.4.1–29.4.4.

[37] M. Haberler, I. Siegl, C. Steffan, and M. Auer, "A bidirectional current-mirror-based potentiostat using a slice-based class-AB amplifier," *IEEE Solid-State Circuits Lett.*, vol. 3, pp. 298–301, 2020.

[38] K. R. Raghunandan, T. L. Viswanathan, and T. R. Viswanathan, "Linear current-controlled oscillator for analog to digital conversion," in *Proc. Custom Integr. Circuits Conf. (CICC)*, Sep. 2014, pp. 1–4.

[39] P. Prabha *et al.*, "A highly digital VCO-based ADC architecture for current sensing applications," *IEEE J. Solid-State Circuits*, vol. 50, no. 8, pp. 1785–1795, Aug. 2015.

[40] J. Tsai, Y. Chen, and Y. Liao, "A power-efficient bidirectional potentiostat-based readout IC for wide-range electrochemical sensing," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2018, pp. 1–5.

[41] J. Kim, T. K. Jang, Y. G. Yoon, and S. Cho, "Analysis and design of voltage-controlled oscillator based analog-to-digital converter," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 1, pp. 18–30, Jan. 2010.
[42] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *Proc. 43rd Int. Symp. Comput. Archit.*, Jun. 2016, pp. 14–26.
[43] H. Jia, H. Valavi, Y. Tang, J. Zhang, and N. Verma, "A programmable heterogeneous microprocessor based on bit-scalable in-memory computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 9, pp. 2609–2621, Sep. 2020.
[44] H. Jia *et al.*, "A programmable neural-network inference accelerator based on scalable in-memory computing," in *IEEE ISSCC Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 236–238.
[45] C.-X. Xue *et al.*, "16.1 A 22 nm 4 Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7TOPS/W for tiny AI edge devices," in *IEEE ISSCC Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 245–247.
[46] M. E. Sinangil *et al.*, "A 7-nm compute-in-memory SRAM macro supporting multi-bit input, weight and output and achieving 351 TOPS/W and 372.4 GOPS," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 188–198, Jan. 2021.

Riduan Khaddam-Aljameh (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in electrical engineering and information technology from the Swiss Federal Institute of Technology in Zürich (ETH Zürich), Zürich, Switzerland, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the Digital Circuits and Systems Group.

From 2017 to 2021, he has been a Predoctoral Fellow with the In-memory Computing Group, IBM Research in Zürich, Zürich, where he conducted research towards his Ph.D. degree in memory and ADC design for deep neural network accelerators using non-volatile memories and SRAM. In 2021, he joined Axelera AI, Eindhoven, The Netherlands, as a Founding Member working on mixed-signal circuit design.

Milos Stanisavljevic (Member, IEEE) received the M.S. degree in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 2004, and the Ph.D. degree in microelectronics from the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2009.

Before joining Axelera AI, Eindhoven, The Netherlands, as a Founding Member, he worked with IBM Research for ten year as a Research Scientist and Research Staff Member. He has published one book, three book chapters, and numerous papers in prestigious journals and conference proceedings. He holds more than 20 U.S. and European patents in the areas of solid-state memory, AI hardware, and AI applications. His current research interests include system architectures and digital circuit design in the area of AI hardware technology and design and optimization of algorithms for signal processing and machine learning.

Jordi Fornt Mas received the B.S. degree in industrial electronics from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2019, and the M.S. degree in electrical engineering and information technology from the Swiss Federal Institute of Technology in Zürich (ETH Zürich), Zürich, Switzerland, in 2021. He is currently pursuing the Ph.D. degree with the High-Performance Integrated Circuits (HIPICS) Group, UPC.

In 2020, he worked as a Research Intern with the In-memory Computing Group, IBM Research–Zürich, Zürich. In 2021, he joined the Barcelona Supercomputing Center (BSC), Barcelona, where he also conducts research towards his Ph.D. degree in deep learning accelerator design.

Geethan Karunaratne (Graduate Student Member, IEEE) received the B.Sc. degree in electronic and telecommunication engineering from the University of Moratuwa, Moratuwa, Sri Lanka, in 2014, and the M.Sc. degree in information technology and electrical engineering from ETH Zürich, Zürich, in 2018, where he is currently pursuing the Ph.D. degree.

He joined IBM Research–Zürich, Zürich, in 2018, where he is currently a member of the In-Memory Computing Group, Zürich. His research interests include in-memory computing and brain-inspired computing.

Matthias Brändli received the Dipl.Ing. (M.Sc.) degree in electrical engineering from the Swiss Federal Institute of Technology in Zürich (ETH Zürich), Zürich, Switzerland, in 1997.

From 1998 to 2001, he was with the Integrated Systems Laboratory, Swiss Federal Institute of Technology in Zürich (ETH Zürich), working on deep-submicron technology VLSI design challenges, digital video image processing for biomedical applications, and testability of CMOS circuits. In 2001, he joined the Microelectronics Design Center, ETH Zürich, where he was involved in numerous digital and mixed-signal ASIC design projects, worked on EDA design automation, and contributed to teaching. In 2008, he joined the IBM Zürich Research Laboratory, Rüschlikon, Switzerland, where he has been working on multi-gigabit/s, low-power communication circuits in advanced CMOS technologies.

Feng Liu (Frank) was a Senior Circuit Layout Designer with IBM and Aragio Solutions design house. He worked on many cutting technologies functional IP layout design, including 22-, 16-, 14-, 12-, 7-, and 2-nm tech nodes. He is currently the design automation lead with IBM Research, where he developed the IBM testsite macro design automation methodology that helped to increase the design efficiency by more than three times. He is also leading the design team at IBM Albany research center and working on 2-nm nano-sheet and vertical stacked FET technology.

Abhairaj Singh (Graduate Student Member, IEEE) received the master's degree *(cum laude)* in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2019. He is currently pursuing the Ph.D. degree with the Computer Engineering Laboratory, Delft.

Prior to his master's, he worked as a SRAM/ROM/RF Circuit Design Engineer with ARM Technologies for more than four years. He worked on memory-based circuit techniques, verification methodologies, design automation, assist circuits, CMOS sensors, low-power (IoT) techniques, and silicon debug. He holds one patent during his time at ARM. His current research interests include circuit design for hardware AI, computation-in-memory, and emerging technologies.

Silvia M. Müller (Senior Member, IEEE) received the M.Sc. degree in mathematics and the Ph.D. degree in computer science from University of Saarland, Saarbrücken, Germany, in 1989 and 1991, respectively.

She is currently an IBM Distinguished Engineer, is leading the development of competitive arithmetic units for POWER and z Systems. She has been driving a shift in IBM's Systems value towards stack solutions based on hardware/software co-design and co-optimizations, specialized arithmetic accelerator engines, and hardware differentiation. Prior to joining IBM in 1999, she was a Professor of processor arithmetic and processor design with the University of Saarland. She has authored over 200 issued patents, three books, and over 30 articles.

Urs Egger completed a four-year apprenticeship as an Electronics Technician in 1984, a three-year study as an Electronic Hardware Engineer at HTL Luzern, Horw, Switzerland, in 1990, and a three-year study as a Software Engineer at HTL Luzern in 1995. He worked for ten years with ETH Zürich, Zürich, Switzerland, where he developed electronic hardware for astrophysics-related research. Since 2010, he has been with IBM Research–Zürich, Zürich, where he develops the test infrastructure for memory and next generation AI hardware.

Anastasios Petropoulos (Graduate Student Member, IEEE) received the Diploma degree in electrical and computer engineering and the M.Sc. degree in signal processing and communication systems from the University of Patras, Patras, Greece, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree.

His research interests include in-memory and high-performance computing.

Theodore Antonakopoulos (Senior Member, IEEE) joined the Department of Engineering, Hellenic Air Force Academy (HAFA), Acharnes, Greece, in 1979, where he studied for two years. He received the Diploma degree in electrical engineering and the Ph.D. degree from the Department of Electrical Engineering, University of Patras, Patras, Greece, in 1985 and April 1989, respectively.

Since 1991, he has been on the faculty of the Electrical and Computers Engineering Department, University of Patras, where he is currently a Professor. He is leading the Communications and Embedded Systems Laboratory focusing his research activities on efficient hardware implementation and rapid prototyping of embedded systems. In 2001 and 2002, he spent his sabbatical year in the IBM Research Laboratory (IBM-ZRL), Rüschlikon, Switzerland, where he was involved in hardware development for probe-based storage devices. Since then, he has long-standing research collaboration with IBM-ZRL. He also participated in various research projects funded by European and international companies and institutes.

Kevin Brew received the Ph.D. degree in chemical engineering from Purdue University, West Lafayette, IN, USA, in 2015, with an Honors B.ChE. with distinction, *(summa cum laude)*, from the University of Delaware, Newark, DE, USA, in 2010.

His more than ten years of experience spans across materials and device research fields, including catalysis, photolithography, photovoltaics, and most recently, memristors. He is currently with the AI Hardware Center, IBM Research, Albany, NY, USA. He has authored or coauthored over 50 patents and publications, and in 2020, was awarded the title of IBM Master Inventor. He continues serving as the unit process lead for phase change materials at the IBM AI Hardware Research Center.

Samuel Choi received the Ph.D. degree in physics from the University of Massachusetts at Lowell, Lowell, MA, USA.

He is currently with IBM Research, Albany, NY, USA. As a Senior Engineer, he is also responsible for integration and delivery of PCM technology on Foundry CMOS logic wafer. His prior responsibilities are in Si-CMOS BEOL/MOL process integration development for IBM server chips expanding eight generations of Copper-nodes. His interests are in exploring and developing new materials integration for leading edge computation. He has 40 patents in the field of CMOS and published numerous articles.

Dr. Choi received three of IBM's Outstanding Technical Achievement awards for 7-nm IBM server processor qualification, Cu/ULK BEOL extendibility at 7-nm node and beyond, and 65-nm node Technology.

Injo Ok received the Ph.D. degree in electrical and computer engineering with The University of Texas at Austin, Austin, TX, USA.

In April 2008, he joined Sematech, Albany, NY, USA, where he worked on fabrication and characterization of III–V and SiGe FinFET transistors for 22-nm node and beyond. He is currently with the IBM Albany Research, Albany. He was a Lead Integrator for POC and RMG for 14-, 10-, 7-nm technology. He currently serves as the Integration Leader of the PCM, IBM AI Center. He has authored or coauthored more than 102 technical papers in international journals and conference proceedings and 102 U.S patents as the author and the coauthor. His research interests include non-volatile memory technologies, and neuromorphic and in-memory computing

Fee Li Lie (Member, IEEE) received the Ph.D. degree in chemical engineering from The University of Arizona, Tucson, AZ, USA.

She is currently with IBM Research, Albany, NY, USA. Her research interests include patterning and integration for semiconductor technology in logic, analog AI, and advanced packaging.

Nicole Saulnier received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA.

She joined IBM upon graduation in 2011. She is currently the Manager of AI materials and process integration with IBM Research, Albany, NY, USA. Since the launch of the AI Hardware Center, she and her team have been developing Analog-AI technologies, including PCM and RRAM.

Victor Chan received the Ph.D. degree in electrical engineering from The Hong Kong University of Science and Technology, Hong Kong.

He is currently with AI Hardware Center, IBM Research, Albany, NY, USA. His research and development interests include emerging devices for analog computing, CMOS device design and yield learning from 90 to sub-10 nm technologies.

Ishtiaq Ahsan received the bachelor's degree from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 1999, and the master's and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, AZ, USA.

He has over 20 years' experience in semiconductor research, development, and manufacturing across seven technology nodes. He currently leads the yield/characterization group in IBM's semiconductor research facility in Albany, NY, USA.

Vijay Narayanan (Senior Member, IEEE) received the B.Tech. degree in metallurgical engineering from the Indian Institute of Technology at Madras, Chennai, India, in 1995, and the M.S. and Ph.D. in materials science and engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1996 and 1999, respectively.

After completing the Post-Doctoral Researcher at Arizona State University, Tempe, AZ, USA, he joined IBM Thomas J. Watson Research Center, Albany, NY, USA, in 2001, where he pioneered high-*k*/metal gate research and development from the early stages of materials discovery to development and implementation in manufacturing. These high-*k*/metal gate materials form the basis of all recent IBM systems processors and of most low-power chips for mobile devices. He is currently an IBM Fellow and a Senior Manager with IBM Research, where he is the strategist for physics of AI and leads a worldwide IBM team developing Analog Accelerators for AI applications within the IBM Research AI Hardware Center. He is the author of over 100 journal articles and conference papers, holds more than 230 U.S. patents, and has edited one book: "*Thin Films on Silicon: Electronic and Photonic Applications*."

Dr. Narayanan was elected a fellow of the American Physical Society in 2011.

S. R. Nandakumar received the M.Tech. degree in microelectronics from the Indian Institute of Technology Bombay (IIT Bombay), Mumbai, India, in 2015, and the Ph.D. degree in electrical engineering from the New Jersey Institute of Technology, Newark, NJ, USA, in 2019.

He is currently a Post-Doctoral Researcher with IBM Research–Zürich, Zürich, Switzerland. His research interests include computational intelligence, neuromorphic engineering, and in-memory computing for non-von Neumann architectures.

Manuel Le Gallo (Member, IEEE) received the dual bachelor's degree from the Ecole Polytechnique de Montréal, Montreal, QC, Canada, and the Ecole Polytechnique (X), Palaiseau, France, and the M.Sc. and Ph.D. degrees from ETH Zürich, Zürich, Switzerland, in 2014 and 2017, respectively.

He joined IBM Research Europe, Rüschlikon, Switzerland, in 2013, where he is currently a Research Staff Member. His current research interests include using phase-change memory devices for non-von Neumann computing.

Pier Andrea Francese (Senior Member, IEEE) received the degree *(cum laude)* in electrical engineering from the Politecnico di Milano, Milan, Italy, and the Ph.D. degree from the Swiss Federal Institute of Technology in Zürich (ETH Zürich), Zürich, Switzerland, in 1993 and 2005, respectively.

He has more than 20 years of experience in the research and the development of integrated circuits in advanced CMOS technologies. He is currently a Principal Research Staff Member with the IBM Zürich Research Laboratory, Rüeschlikon, Switzerland, where he also serves as the Technical Leader of the high-speed interconnect technology group. His research interests include high-speed data converters, analog equalization, and clock-data-recovery circuit techniques.

Abu Sebastian (Senior Member, IEEE) received the B.E. degree (Hons.) in electrical and electronics engineering from BITS Pilani, Pilani, India, in 1998, and the M.S. and Ph.D. degrees in electrical engineering (minor in mathematics) from Iowa State University, Ames, IA, USA, in 1999 and 2004, respectively.

He is currently a Distinguished Research Staff Member with IBM Research–Zürich, Zürich, Switzerland. He was a contributor to several key projects in the space of storage and memory technologies and also manages the research effort on in-memory computing with IBM Research–Zürich. He is the author/coauthor of over 200 publications in peer-reviewed journals/conference proceedings and holds over 70 U.S. patents. He has also the co-edited a book titled "*Memristive Devices for Brain-Inspired Computing*" by Elsevier in 2020.

Dr. Sebastian was awarded the European Research Council (ERC) Consolidator Grant in 2015. In 2020, he was awarded an ERC Proof-of-Concept Grant. He has been an IBM Master Inventor since 2016. In 2019, he received the Ovshinsky Lectureship Award for his contributions to "Phase-change materials for cognitive computing." He has served on the technical program committees of several conferences, including IEDM, AICAS, and E\PCOS. He was selected as the Distinguished Lecturer in the IEEE Council on Electronic Design Automation for 2022–2023.

Evangelos Eleftheriou (Life Fellow, IEEE), received the Ph.D. degree in electrical engineering from Carleton University, Ottawa, ON, Canada.

In 1986, he joined IBM Research–Zürich, Zürich, Switzerland, as a Research Staff Member, and over the years he has held various management positions. Since October 2021, he has been the Chief Technology Officer (CTO) and the Co-Founder of Axelera AI. He has authored or coauthored over 250 publications and holds over 160 patents (granted and pending applications). His research interests include AI and machine learning, including emerging computing paradigms, such as neuromorphic and in-memory computing.

Dr. Eleftheriou was appointed an IBM Fellow in 2005, and was inducted into the IBM Academy of Technology in the same year. In 2018, he was inducted into the U.S. National Academy of Engineering as a Foreign Member. He was a co-recipient of the IEEE ComS Leonard G. Abraham Prize Paper Award in 2003. He was also a co-recipient of the 2005 Technology Award of the Eduard Rhein Foundation. In 2009, he was also a co-recipient of the IEEE Control Systems Technology Award and the IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY Outstanding Paper Award. In 2016, he received a Honoris Causa professorship from the University of Patras, Patras, Greece.