IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 14 June 2024; revised 17 September 2024 and 16 October 2024; accepted 5 November 2024.
Date of publication 11 November 2024; date of current version 6 December 2024.

Digital Object Identifier 10.1109/JXCDC.2024.3496118

System–Technology Co-Optimization for Dense Edge Architectures Using 3-D Integration and Nonvolatile Memory

LEANDRO M. GIACOMINI ROCHA[1] (Member, IEEE), MOHAMED NAEIM[1,2,3], GUILHERME PAIM[4,5] (Member, IEEE), MORITZ BRUNION[1], PRIYA VENUGOPAL[1], DRAGOMIR MILOJEVIC[2], JAMES MYERS[6] (Member, IEEE), MUSTAFA BADAROGLU[7], MARIAN VERHELST[5] (Fellow, IEEE), JULIEN RYCKAERT[1], and DWAIPAYAN BISWAS[1] (Member, IEEE)

[1]imec, 3000 Leuven, Belgium
[2]Bio, Electro, and Mechanical Systems Department (BEAMS), Université Libre de Bruxelles, 1050 Brussels, Belgium
[3]Cadence Design Systems, San Jose, CA 95134 USA
[4]INESC-ID, 1000-029 Lisbon, Portugal
[5]KU Leuven, 3000 Leuven, Belgium
[6]imec, CB2 0AH Cambridge, U.K.
[7]Qualcomm, San Diego, CA 92121 USA

CORRESPONDING AUTHOR: L. M. GIACOMINI ROCHA (leandro.m.giacominirocha@imec.be)

This article has supplementary downloadable material available at https://doi.org/10.1109/JXCDC.2024.3496118, provided by the authors.

ABSTRACT High-performance edge artificial intelligence (Edge-AI) inference applications aim for high energy efficiency, memory density, and small form factor, requiring a design-space exploration across the whole stack—workloads, architecture, mapping, and co-optimization with emerging technology. In this article, we present a system–technology co-optimization (STCO) framework that interfaces with workload-driven system scaling challenges and physical design-enabled technology offerings. The framework is built on three engines that provide the physical design characterization, dataflow mapping optimizer, and system efficiency predictor. The framework builds on a systolic array accelerator to provide the design–technology characterization points using advanced imec A10 nanosheet CMOS node along with emerging, high-density voltage-gated spin–orbit torque (VGSOT) magnetic memories (MRAM), combined with memory-on-logic fine-pitch 3-D wafer-to-wafer hybrid bonding. We observe that the 3-D system integration of static random-access memory (SRAM)-based design leads to 9% power savings with 53% footprint reduction at iso-frequency with respect to 2-D implementation for the same memory capacity. Three-dimensional nonvolatile memory (NVM)-VGSOT allows 4× memory capacity increase with 30% footprint reduction at iso-power compared with 2-D SRAM 1×. Our exploration with two diverse workloads—image resolution enhancement (FSRCNN) and eye tracking (EDSNet)—shows that more resources allow better workload mapping possibilities, which are able to compensate peak system energy efficiency degradation on high memory capacity cases. We show that a 25% peak efficiency reduction on a 32× memory capacity can lead to a 7.4× faster execution with 5.7× higher effective TOPS/W than the 1× memory capacity case on the same technology.

INDEX TERMS 3-D partitioning, edge artificial intelligence (Edge-AI), nonvolatile memory (NVM), system–technology co-optimization (STCO), systolic array, voltage-gated spin–orbit torque (VGSOT), W2W HB.

I. INTRODUCTION

DENSER CMOS technologies have been a catalyst for continuous miniaturization, driving large volumes at low cost and enabling unprecedented system-level innovations. Technology scaling across logic, memory, and 3-D enables higher compute, memory, and bandwidth density and is a key factor toward scalable system design. In the more than Moore era, system–technology co-optimization (STCO) is a promising paradigm for leveraging the synergy between emerging technology and application-driven architectures to achieve higher efficiency and performance at cost parity.

Hardware performance improvements have at large relied on technology scaling and the introduction of scaling boosters following a design–technology co-optimization (DTCO) approach, which is challenged by next-generation artificial intelligence (AI) applications and domain-specific architectures, necessitating a cross-stack codesign. As CMOS scaling stagnates and systems are constantly challenged by memory and power density/thermal bottlenecks, specialism in the form of domain specific architecture and technology (DSAT) tied to an application is taking prominence. Given the multifaceted demands of AI-driven system design, encompassing diverse workloads and architectures, STCO appears as the viable solution based upon cross-stack codesign with technology innovations (Fig. 1).

Workload-driven design-space exploration (DSE) frameworks, especially for Edge-AI inference architectures, such as ZigZag [1], TimeLoop [2], and Maestro [3], provide codesign opportunities to optimize the memory hierarchy and compute blocks of deep neural network (DNN) accelerators considering its large spatiotemporal mapping. However, these frameworks are largely agnostic of emerging technologies, inherently leading to suboptimal design choices.

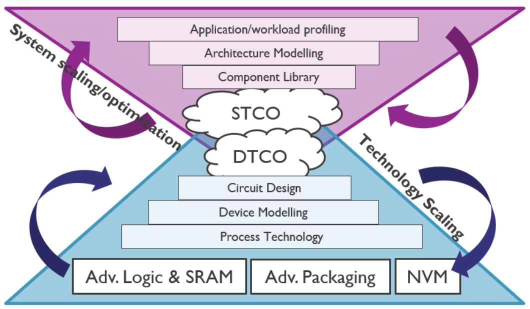

FIGURE 1. STCO: bridging the gap between application-driven systems and technology.

In this article, we present a comprehensive STCO framework for system efficiency insights on high-performance Edge-AI inference architectures driven by emerging, technology-dependent physical design parameters. Our proposed framework enables fast architectural exploration for different workloads, sweeping both compute and memory capacities. In particular, the contributions of this article can be summarized as follows.

1) An STCO framework for system efficiency predictor, considering Edge-AI inference architecture template and mapping of relevant AI workloads, representative of extended reality (XR) applications.
2) Advanced technology annotation of the framework with design evaluation on system scalability is as follows: 1) logic—imec A10 gate-all-around (GAA) nanosheet process design kit (PDK); 2) heterogeneous memory—A10 nanosheet static random-access memory (SRAM) and A10 voltage-gated spin–orbit torque (VGSOT) magnetic memories (MRAM)-based nonvolatile memory (NVM) macros; and 3) advanced 3-D integration using concurrent 3-D IC flow and wafer-to-wafer (W2W) hybrid bonding (HB) at 1.12-μm pitch.

Finally, we demonstrate that within the complex landscape of Edge-AI systems, using advanced technologies does not inherently ensure superior performance. Instead, it requires a comprehensive codesign across the stack (i.e., workload, mapping, memory hierarchy, and technology). Our proposed STCO framework helps in early decision-making driving technology development toward optimal system design.

This article is further structured as follows. Section II highlights the motivation for STCO, focusing on XR relevant applications and architecture mapping. Section III introduces our proposed STCO framework, based on systolic array architecture, workload mapping, 3-D functional partitioning, and implementation. The results are discussed in Section IV, providing insights into 2-D versus 3-D architecture/design tradeoffs and power, performance, and area (PPA) metrics. The conclusions are drawn in Section V.

II. MOTIVATION FOR STCO

As technology scaling stagnates, improving system performance faces several bottlenecks—memory, power, and bandwidth walls. STCO is intended to mitigate them co-optimizing on two axes: technology and system design. The impact of technology scaling is assessed at the block level through DTCO methodology, synergizing design and process technology to enhance performance, power efficiency, and cost-effectiveness [4]. On the other hand, current system design research is primarily based on new architectures for a target application and design constraints, largely abstracting the impact of underlying technology [5].

A. EMERGING TECHNOLOGIES

Logic DTCO innovations ranging from materials, processes, device architectures, and power delivery mechanisms help maintain area scaling for the next generation of CMOS nodes—progression of FinFET devices to other GAA devices, such as nanosheet and CFET [6], [7] (see Fig. 2).

The progress in 3-D integration in large systems offers potential improvements in PPA compared with traditional 2-D implementations. Technologies, such as microbumps (μBumps) and W2W-HB, have enabled high-density 3-D interconnects, achieving the densities of 10 K and 1 M interconnect/mm^2, respectively [8]. Functional partitioning of the design and stacking the repartitioned dies vertically lead to footprint and total signal wirelength reduction, impacting power and timing. The technology and physical design implications of multidies partitioning, in face-to-face (F2F), face-to-back (F2B), or a combination of both, have been further elaborated in [9] and [10].

As SRAM scaling stagnates, beyond FinFET, MRAM-based NVMs have emerged as a possible candidature for on-chip usage with footprint and leakage power advantages [11]. Spin-transfer torque (STT) and spin–orbit torque (SOT) show significantly smaller bitcell area and standby energy reduction and, however, incur a tradeoff with inherent higher access delay, requiring the system-level architecture modifications to incorporate such characteristics. STT MRAM has been noted for its density advantage over SOT memories, yet it exhibits slower operational speeds. Conversely, SOT memories offer rapid switching times, but their design necessitates two transistors per bit stored in the magnetic tunnel junction (MTJ), leading to a larger area requirement. The emerging VGSOT technology [12], however, shows promise in achieving fast switching at lower currents compared with SOT and uniquely allows for the sharing of a write transistor across multiple MTJ pillars, which can lead to configurations, such as five transistors per 4-bit MTJ, enhancing area efficiency. This leads to shorter interconnect paths, with potential benefits for larger array configurations over SRAM.

There have been several frameworks reporting on optimizing across different logic nodes with design, workload, thermal, and turnaround time (TAT) considerations, but have primarily focused at a block level, excluding system–architecture-level considerations [13], [14].

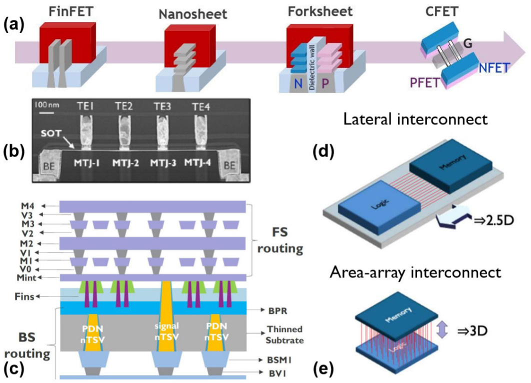

FIGURE 2. Technology scaling boosters. (a) Device scaling from FinFET to CFET. (b) MRAM-based memory cells for denser memories. (c) Wafer backside functionality with buried power tail (BPR). (d) Chiplet approach with interposers. (e) Die-on-die 3-D integration.

B. EDGE-ARCHITECTURE DESIGN CONSIDERATIONS

Edge-AI accelerator design imposes strict boundary conditions with respect to desired performance, power, form factor, and cost tradeoffs [15]. System-wide optimization, as shown in Fig. 1, requires investigation into various aspects.

1) *Application:* AI vision workloads involving GEMM/GEMV kernels (e.g., CNNs).
2) *Architecture:* Integrating compute core and memory hierarchy along with other components.
3) *Mapping (Spatial and Temporal):* Architecture-dependent data movement and utilization.

XR-driven applications drive high-performance computing on edge architectures and have funneled a plethora of neural networks (NNs) for different functionalities, such as vision processing, body movement tracking, depth estimation, and eye tracking among others [16], [17].

Such DNN workloads heavily rely on matrix multiplications, which can be accelerated through array-based accelerators. This capability has been demonstrated through several AI-focused accelerators, many of which are inspired by the Google TPU [18], employing systolic array (SA) with weight-stationary dataflow and dedicated SRAM buffers. Subsequent works, such as Eyeriss [19], proposed a row-stationary, highly energy-efficient design. Also, Meta introduced an augmented reality (AR)/virtual reality (VR) accelerator [20] with an array of 16 × 32 processing elements (PEs) in an output-stationary dataflow, along with 3-D stacked on-chip SRAM buffers for activations and weights.

Due to the data-intensive nature of DNN workloads, multiple DSE frameworks [1], [2], [3] have been proposed to jointly optimize architecture and data movement, taking into consideration different data reuse schemes. Furthermore, optimized mapping schemes for concurrent execution of multiple DNNs on SAs or on segmented versions to optimize for performance and energy have been reported in [21]. Fundamentally, these frameworks rely on dataflow optimization for different workloads to minimize either the energy or latency for a given execution, considering the available compute and memory budget on a given technology node.

Distinction From SoTA: Table 1 summarizes related work and presents a characteristic comparison with our proposed STCO-driven framework. A DTCO study on GPU was performed in [22], whereas [23] focuses on custom accelerators without considering AI workloads. UTOPIA [13] presents a technology–design–system co-optimization approach for mobile SoCs. A workload-aware architectural exploration is presented in [1] without technology implications. Maestro [3] takes a similar approach, incorporating component-level characterization in 28-nm library. The work presented in [24] performs a system-level analysis on systolic array accelerator using BEOL compatible RRAM and CNFETs in 130-nm technology library. However, to the best of authors' knowledge, none of these reported works propose a comprehensive system efficiency predictor that considers heterogeneous memory technologies and advanced 3-D bonding with partitioning schemes along with workload–architecture design-space exploration for AI accelerators.

TABLE 1. Summary of the related work.

	Summary Results								
Related Work	A	B	C	D	E	F	H	I	J
22 [22]	✓	–	–	–	–	–	–	–	–
23 [23]	✓	✓	–	–	–	–	–	–	–
UTOPIA [13]	✓	✓	–	–	–	–	–	–	–
ZigZag [1]	–	–	–	–	–	✓	✓	✓	✓
Timeloop [2]	✓*	–	–	–	–	✓	✓	✓	✓
MAESTRO [3]	✓*	–	–	–	–	✓	✓	✓	✓
[24]	✓	–	✓	✓	–	✓	✓	✓	–
Our Proposal	✓	✓	✓	✓	✓	✓	✓	✓	✓

Technology: (A) Technology-aware, (B) Advanced Technologies, (C) 3D Integration, (D) Non-Volatile Memories, (E) VGSOT Memory. (*) Partial support

System: (F) Application-aware and AI Workload-aware, (H) Analysis support for any convolutional workload, (I) AI Accelerator Architecture and Memory Hierarchy, (J) Spatial and Temporal Mapping with Scheduling Optimization

III. A SYSTEM–TECHNOLOGY CO-OPTIMIZATION FRAMEWORK

Our proposed STCO framework enables holistic co-optimization, as illustrated in Fig. 3, built upon ZigZag DSE tool [1] and calibrated with PPA metrics extracted from physical design experiments on a template architecture. This leads to a technology-annotated DSE for system efficiency predictor connecting different design knobs. The framework is divided into three engines as follows: 1) the place and route (PnR)-based design–technology characterization generating PPA metrics, feeding to the other two engines; 2) DSE mapping; and 3) system-level efficiency predictor. We bridge the physical characterization with the high-level DSE optimization through a demonstrative accelerator architecture using both a register transfer level (RTL) description and performance model.

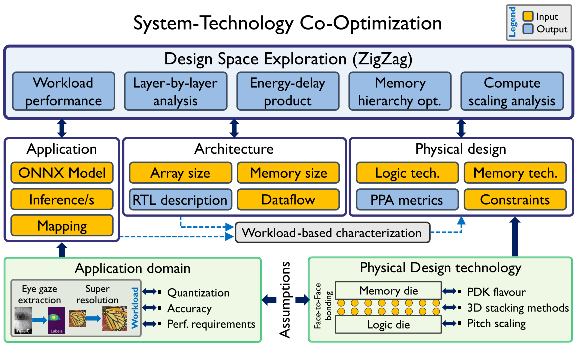

FIGURE 3. **Proposed STCO framework for design-space exploration of edge architectures leveraging *dense* emerging technologies across abstraction levels of the computing stack.**

A. ACCELERATOR TEMPLATE ARCHITECTURE

We propose DenseXR, a single-core configurable squared SA-based accelerator, as illustrated in Fig. 4(a). As the main application focus is on Edge-AI workloads, the array is composed of 32 × 32 PEs with one multiply-accumulate (MAC) operation per PE. We adopted a squared SA inspired on commercial products, such as [18] and [20]. Although nonsquared SAs could be optimized for certain workloads, the unmatched bandwidth requirements would increase the complexity of both system design and dataflow mapping. This choice relies on the versatility of SA-based accelerators on executing general matrix–matrix multiplication (GEMM) operations, the backbone of NN workloads.

It uses 8-bit signed integer (INT8) arithmetic for both activations and weights, with accumulation in 32-bit signed integer (INT32) format. Each PE element utilizes a double-buffered weight-stationary dataflow, allowing data prefetching. To handle the 32-bit partial sums, a 16-kB block of double-buffered 32-bit accumulators is connected to the systolic array [the *Accum* block in Fig. 4(a)]. This 16-kB memory is divided into 32 accumulator units, each equipped with 2 memory banks, with 128 words of 32 bits.

Motivated by Eyeriss [19], the accelerator has a two-level memory hierarchy to store the on-chip data: a shared multiport L2-like scratchpad memory configured with multiple banks, and two L1-like scratchpad memories to store the activations and weights. These memories can be configured to use either SRAM or NVM technologies (discussed in Section III-B). All memories are configured for maximum bandwidth ($32 \times 8b = 256b$) per channel, connected through a crossbar to route the data transfers.

Given that memories can have different latencies as they might be slower than the logic, each memory block can be configured to operate based on time multiplexing using a round-robin scheduling. Each memory bank is divided into smaller banks with pipelined control signals, ensuring that the maximum throughput will be available after the initial ramp-up phase. With proper scheduling and a reasonable multiplexing factor, the initial latency will be negligible on the overall workload execution.

B. DESIGN–TECHNOLOGY CHARACTERIZATION

We perform a block-level characterization to obtain the PPA metrics in the associated technology, running PnR experiments using Cadence tools. This helps to extract block-level energy metrics to annotate the DSE tool.

For 3-D circuit integration, the more mature memory-on-logic (MoL) flow, having memory macros on top and remaining logic at the bottom die, is considered. For our memory-dominated template architecture, the bottom die contains a mix of logic and memory macros, as its size matched the top die for W2W integration considerations. The top die contains most of the L2 sub-banks, while the bottom die contains the remaining L2 banks, all L1 banks, and the compute logic. We have used a delay-annotated post-signoff netlist to simulate the design. The simulation generated a stimuli file to back-annotate the implementation database, allowing the power estimation based on the execution of a real workload. This approach allows a precise characterization of the energy cost per data movement and computation, streamlining the integration with the system exploration. The PnR flows are described in Section IV of the Supplementary Material.

C. DESIGN-SPACE EXPLORATION TOOL AND SYSTEM PERFORMANCE PREDICTION ENGINE

We employ ZigZag [1] as the core DSE mapping engine due to its configurability and extensibility. It takes as input the target workloads specified in Open Neural Network Exchange (ONNX) format and an architecture template describing the memory subsystem and compute capabilities, using such information to find the optimal spatial and temporal operation mapping. The DSE engine supports energy-, latency-, or energy-delay product (EDP)-driven optimizations that rely on the design–technology characterization engine.

ZigZag is originally technology-agnostic, so it reports execution cycles and energy per access metrics, which might overlook the design implementation challenges. Hence, we customized the DSE tool to consider technology characterizations for compute logic or memory levels into its mapping optimization engine and extend it further to do the following: 1) estimate the system performance and efficiency

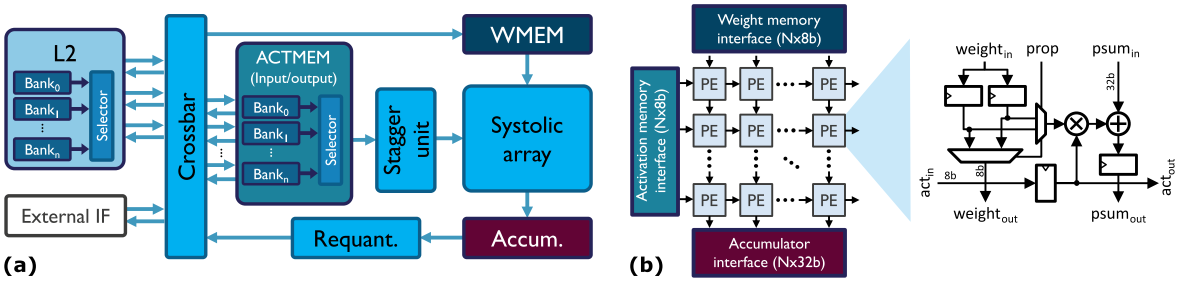

FIGURE 4. DenseXR systolic array. (a) System overview with PE array with high-bandwidth, multilevel memory hierarchy. (b) Array organization with a double-buffered weight-stationary PE architecture.

considering the system frequency and 2) provide a statistical prediction for energy efficiency-related metrics evaluation considering different architectural design points. This approach enables technology-driven, fast architectural explorations for early system efficiency-related design sweeps.

For its first task, the prediction engine assumes the lowest operating frequency point obtained from the design–technology characterization as the baseline iso-frequency performance point. If the selected memory–compute configuration has an annotation point, the engine computes the design peak TOPS/W and estimates the effective TOPS/W, considering the effects of nonideal operation scheduling. However, if the memory–compute configuration is not present, the engine uses a regression-based analysis to estimate the characterization points at iso-frequency, scaling the power and performance according to the compute array size and total memory capacity. The prediction engine is conceived in a modular fashion to support different estimation methods. Further details can be found in Section I of the Supplementary Material.

The framework evaluation comes from the combination of these three engines, which provide optimization knobs at the application, architecture, and technology parameters. Among all the possible optimization knobs, we selected a subset of them as listed in the following sections.

1) APPLICATION KNOB CONSIDERATIONS

We selected FSRCNN and EDSNet workloads, both based on convolution neural network (CNN), to demonstrate the framework capabilities, having vastly different memory and compute requirements, and their functionality is generic enough to be integrated into XR systems.

The FSRCNN workload [25] is an image enhancer that upscales low-resolution images to high-resolution ones with higher fidelity than traditional algorithms. For this experiment, we considered three input resolutions (256 × 256, 512 × 512, and 1024 × 1024) with a 2× upscaling factor. We also considered the EDSNet workload [26], a neural network tailored for eye-tracking purposes. The network is based on top of MobileNet-V2 with an input resolution of 320 × 320 pixels. More details on the inference compute and memory requirements for each workload are described in Section II of the Supplementary Material.

Selected application knob values are as follows.

1) *FSRCNN:* Super-resolution with various resolutions.
2) *EDSNet:* Eye segmentation with single resolution.

2) ARCHITECTURE KNOB CONSIDERATIONS

Workloads have distinct compute/memory resource demands, and the optimal design point requires careful architectural exploration. We consider different compute and memory capacities by changing the on-chip memory and SA sizes based on the underlying template architecture. For the remaining of this article based on physical design explorations, i.e., PnR runs, will be named *physical*. All the PPA numbers obtained from the prediction engine will be marked as *prediction*.

Selected architecture knob values are as follows.

1) *Memory:* Sweep total on-chip capacity from 1× to 4× (physical), extending to 32× (prediction).
2) *Compute:* Use a 32 × 32 array size (physical), considering 16 × 16 and 64 × 64 array sizes (prediction).

3) TECHNOLOGY KNOB CONSIDERATIONS

Although increasing on-chip memory capacity usually leads to better system efficiency due to the reduction of off-chip memory access, the additional on-chip resources can significantly degrade the chip PPA due to the increase in footprint and wirelength, also at an increased die cost. We mitigate this impact by leveraging two technology approaches as follows: 1) moving into the third dimension and 2) integrating emerging memory technology on the same design.

We selected the MoL 3-D partitioning approach assuming advanced W2W HB technology in an F2F configuration, as shown in Fig. 5. This configuration allows for tight 3-D interconnect pitch, at 1.12 μm, as a multiple of the last metal layer pitch of the considered technology (14 × 0.08 μm pitch) that guarantees low RC insertion [27].

Complementary to this approach, we evaluated the partial replacement of SRAM-based memory macros by VGSOT-based macros with the same capacity to benchmark the potential benefits and drawbacks of NVMs on edge devices. All memory macros have an identical capacity of 32 kB, and they are stitched together to create higher capacity memory subsystems. We considered a four-pillar bitcell, as proposed in [11], combined with the memory pipelining scheme described in the previous section to cope with

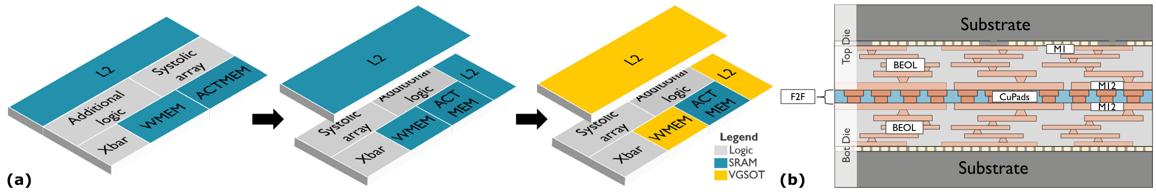

FIGURE 5. Design partitioning experiments. (a) Illustrative 2-D–3-D floor planning with replacement of SRAM macros by NVM. (b) Cross section illustration of F2F 3-D die partitioning, where the top die is composed only of memory macros.

TABLE 2. Overview of PnR experiments.

Block	Memory capacity	Memory technology	Partitioning	3D assumptions
ACTMEM	1× (512kB)	SRAM	2D & 3D MoL	W2W HB 1.12μm pitch
WMEM	1× (256kB) to 4× (1MB)	SRAM & VGSOT		
L2	1× (1.5MB) to 4× (6MB)			

the 4× access delay increase at the macro level when compared with an SRAM implementation. More details on the macro characteristics can be found in Section V of the Supplementary Material.

Selected technology knob values are as follows.

1) *Partitioning Scheme:* Conventional 2-D (single die) and 3-D MoL with two dies and 1.12-μm pitch.
2) *Memory Technology:* SRAM and VGSOT macros with the same macro capacity.

IV. RESULTS AND DISCUSSION

As the main purpose of our framework is to bring together the technology and system exploration, we adopt a two-step evaluation methodology, starting with PnR experiments followed by DSE analysis. For the first part, we characterize the architecture components of our demonstrative systolic array-based system using an advanced nanosheet CMOS PDK. Once this characterization is finished, we annotate the PPA back into the DSE tool, which enables the performance assessment of different workloads as well as the performance prediction for different architecture parameters.

A. TECHNOLOGY CHARACTERIZATION

For the framework calibration, we adopted imec A10 nanosheet CMOS PDK for all PnR experiments. This PDK has 13 back-end-of-line (BEOL) metal layers, and it also offers power delivery through BPR. We combined the architecture and technology knobs defined in Section III, leading to the experiments listed in Table 2. In all explorations, we fixed the SA array size into 32 × 32 and kept the activation memory capacity constant.

We chose to sweep the L2 and WMEM memories for two reasons as follows: 1) they are more tolerant to higher memory latency given the weight-stationary dataflow of the systolic array, with potential for data prefetching and 2) the ACTMEM memory contents could be partially offloaded to the L2. We use NVM macros on these memories for benchmarking purposes, although this is not a limiting factor, since the framework can be extended to other dataflows with the same characterization data points.

The results presented in Fig. 6 include maximum performance and iso-frequency targets. The first target aims to evaluate the effect of technology knobs, while the iso-frequency design configurations facilitate a fairer system energy efficiency comparison (TOPS/W). All comparisons are normalized with respect to the baseline configuration with 1× memory capacity with SRAM macros in a single die.

Employing 3-D partitioning improves the maximum achieved frequency, particularly with SRAM-based implementations. For instance, at 1× memory capacity, the SRAM 3-D configuration increases the maximum achieved frequency by 32%, with diminishing gains to 10% and 2% at 2× and 4× memory capacities, respectively. This is attributed to the design's expansion and extra logic cells accompanying L2 and WMEM macros, requiring longer wires. This is a hint that more than two dies may be considered as an alternative to further improve scaling. In the case of NVM, the 3-D configurations show an approximate 1% increase in frequency compared with 2-D across all memory capacities. This near-constant performance is attributed to the multicycle path constraints (three clock cycles) applied during the implementation of all NVM configurations due to the slower VGSOT macro access time (near 3 ns).

While 3-D NVM does not significantly increase the achieved frequency, it has a significant reduction in total power with approximately 13% power savings across all capacities, compared with 2-D NVM. These memories show a near-linear reduction in power for both 2-D and 3-D configurations, as the memory capacity increases, attributable to the lower achieved frequencies at such capacities. This trend is not present in SRAM-based 3-D implementations, as larger memory capacities (2× and 4×) have higher achieved frequency (10% and 2%, respectively) when compared with their 2-D counterparts while also offering power savings of 10% and 14%, respectively.

For a broader understanding of how the technology choices affect the design peak energy efficiency (TOPS/W), we evaluate the target design at iso-frequency (625 MHz), as shown in Fig. 6. At iso-frequency, 3-D partitioning provides significant switching power reduction due to the signal wirelength reduction. For instance, comparing the SRAM 2× 2-D and 3-D implementations, the 3-D configuration achieves an 18%

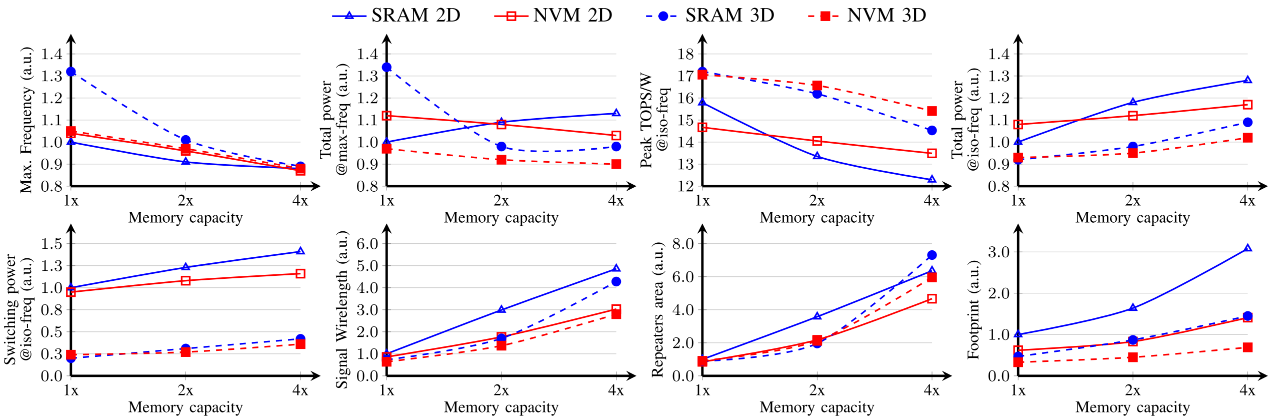

FIGURE 6. **PPA characterization for different PnR configurations. Normalized to the baseline configuration adopting a 2-D partitioning scheme.**

reduction in total power due to a 76% decrease in switching power. This is explained by the 44% reduction in signal wirelength and 36% reduction in the repeaters count.

NVM-based implementations provide better scaling opportunities in terms of energy efficiency and footprint, especially at high on-chip memory capacities. For edge systems where the footprint is also a crucial factor, we observe that the 2-D NVM implementation rivals the 3-D-based SRAM implementation, which can potentially represent cost savings due to the less complex integration scheme. As the memory capacity grows from 1× to 4×, the silicon area gap to the SRAM 2-D implementation grows from 38% to 54%, respectively.

B. SYSTEM-LEVEL ANALYSIS

As each workload has different memory and compute requirements, the effective mapping into the hardware will be greatly affected by the available resources. From the previous section, the more the on-chip resources are available, the higher the degradation on the design PPA and, consequently, the energy efficiency. Yet, looking only at design PPA may hide system-level improvements that workload-dependent. Hence, we assess the system performance considering multiple workloads with different compute/memory characteristics. This evaluation is done for the calibrated designs of Section III-B, and we predict the potential performance for other array sizes and memory capacity, as described in Section III-C.2.

Fig. 7 shows how the workload impacts the effective system energy efficiency at iso-frequency when we take into account the dataflow mapping improvements as a consequence of architectural factors, such as on-chip memory capacity and available compute capability. The overall trend is that more resources can lead to better system utilization even though the system peak performance is lower, as the mapping can better utilize the available memory and compute. As the resources are better utilized, the effective system energy efficiency comes closer to the peak performance estimated for that given architecture.

The experiments in Fig. 7 take the technology-aware framework and estimate the efficiency when more on-chip memory is available. Assuming the downward trend on the peak energy efficiency shown in the previous section, we observe that these mapping gains can offset the worse design PPA. On the EDSNet workload, for instance, the effective TOPS/W decreases until 4× capacity, as the mapping gains are negligible with respect to the PPA degradation, although this trend is reversed at 8× onward, as the mapping gains are much higher than the PPA degradation. Such nonlinear behavior can be explained by the heuristics of the optimizer. However, other workloads have different optimal points, as the mapping optimization is not enough to compensate for the lower peak efficiency.

A complementary look at this analysis is shown in Fig. 8, where the technology assumptions are fixed to a 3-D-based implementation with NVM memories while varying the available compute and memory resources at iso-frequency. The trends per workload are quite diverse, as the EDSNet implementation sometimes is compute-bound, and sometimes, it is memory-bound. Focusing on the 32× memory capacity for the 64 × 64 array, it shows a 25% peak system efficiency degradation with respect to the 1× configuration, yet the workloads can execute faster (2.6× on EDSNet and 7.4× on FSRCNN) with better effective system energy efficiency, with improvements ranging from 2× (EDSNet) to 5.7× (FSRCNN).

C. OFF-CHIP MEMORY ANALYSIS

The major premise around increasing on-chip memory is to reduce the DRAM access, since it often has a lower bandwidth and a very high energy/bit access cost. Hence, we provide some early insights on how the relative off-chip access can be minimized by increasing the on-chip memory and the potential impact on the system energy efficiency. We assume a relative approach to measure the additional accesses needed to handle the memory spillover, i.e., the temporary storage of partial activations in the main memory and the fetching of the same input/weight data from off-chip.

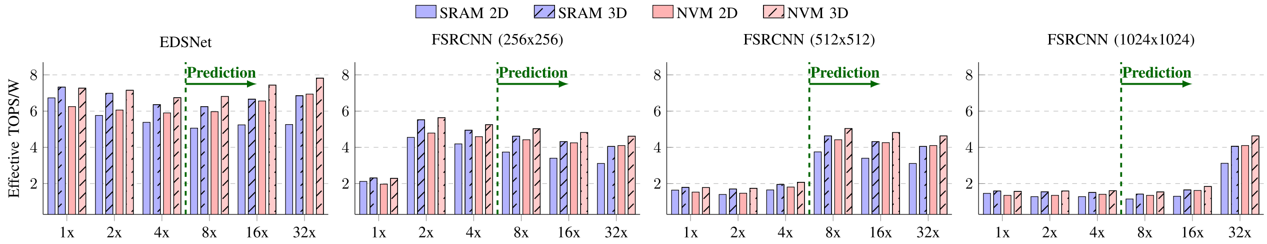

FIGURE 7. **Iso-compute analysis: system efficiency evaluation for multiple workloads at different memory and technology selections, estimating the system energy efficiency up to 32× memory capacity.**

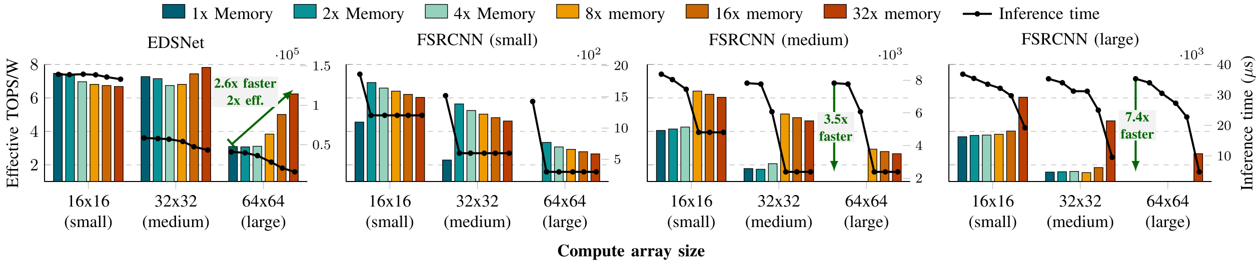

FIGURE 8. **Iso-technology analysis: system efficiency evaluation for compute and memory scale-up considering NVM memories and 3-D integration.**

Assuming a dataflow optimizer configured to minimize the compute latency at any cost and a similar DRAM memory bandwidth as on-chip memory bandwidth, we can estimate the extra off-chip memory accesses needed. Focusing on the FSRCNN (large) model for the 64 × 64 compute engine, the 32× memory capacity is the only configuration able to hold the entire model on-chip. Considering the 1× memory capacity, it would need 18× more off-chip accesses, significantly offsetting the overall system energy efficiency.

V. CONCLUSION

This work is a first of its kind to provide a framework for workload- and technology-aware design-space exploration, considering a template AI accelerator, 3-D exploration at the fine-grained pitch with NVM, and advanced nanosheet A10 CMOS technology. It shows the feasibility of STCO as a methodology for evaluating system implications of future technology decisions and vice versa. Other works [28], [29] highlighted the potential benefits of adopting such advanced 3-D integration over diverse architectures. Results show that increasing the memory capacity leads to worse peak energy efficiency due to the increase of on-chip resources. However, the workloads and mapping engine have a crucial role in the effective system efficiency, and they can compensate for the adverse effects on PPA. For instance, we demonstrated that for a 64 × 64 array, increasing the memory from 1× to 32× has a 25% degradation on peak efficiency, yet the effective system energy efficiency increases 2× with 2.6× lower execution time on the EDSNet workload with similar findings to other cases. Future work will aim to improve the framework, incorporating detailed cost analysis for off-chip main memory access, adoption of emerging technology elements (advanced logic, MRAM, 3-D bonding, and other DTCO design knobs), and thermal assessment into the system efficiency analysis.

ACKNOWLEDGMENT

(Leandro M. Giacomini Rocha and Mohamed Naeim contributed equally to this work.)

REFERENCES

[1] L. Mei, P. Houshmand, V. Jain, S. Giraldo, and M. Verhelst, "ZigZag: Enlarging joint architecture-mapping design space exploration for DNN accelerators," *IEEE Trans. Comput.*, vol. 70, no. 8, pp. 1160–1174, Aug. 2021.

[2] A. Parashar et al., "Timeloop: A systematic approach to DNN accelerator evaluation," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Mar. 2019, pp. 304–315.

[3] H. Kwon, P. Chatarasi, M. Pellauer, A. Parashar, V. Sarkar, and T. Krishna, "Understanding reuse, performance, and hardware cost of DNN dataflow: A data-centric approach," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchit.*, Oct. 2019, pp. 754–768.

[4] L. W. Liebmann and R. O. Topaloglu, "Design and technology co-optimization near single-digit nodes," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2014, pp. 582–585.

[5] A. B. Kelleher, "Celebrating 75 years of the transistor a look at the evolution of Moore's Law innovation," in *IEDM Tech. Dig.*, Dec. 2022, pp. 1.1.1–1.1.5.

[6] S. K. Moore, "Keeping Moore's Law going is getting complicated," *IEEE Spectr.*, 2023. [Online]. Available: https://spectrum.ieee.org/stco-system-technology-cooptimization

[7] A. Spessot et al., "Device scaling roadmap and its implications for logic and analog platform," in *Proc. IEEE BiCMOS Compound Semicond. Integr. Circuits Technol. Symp. (BCICTS)*, Nov. 2020, pp. 1–8.

[8] S. B. Samavedam et al., "Future logic scaling: Towards atomic channels and deconstructed chips," in *IEDM Tech. Dig.*, Dec. 2020, pp. 1.1.1–1.1.10.

[9] M. Naeim et al., "Design enablement of 3-dies stacked 3D-ICs using fine-pitch hybrid-bonding and TSVs," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, May 2023, pp. 1–4.

[10] G. Sisto et al., "Design enablement of fine pitch face-to-face 3D system integration using die-by-die place & route," in *Proc. Int. 3D Syst. Integr. Conf. (3DIC)*, Oct. 2019, pp. 1–4.

[11] M. Gupta et al., "Ultimate MRAM scaling: Design exploration of high-density, high-performance and energy-efficient VGSOT for last level cache," in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4.

[12] K. Cai et al., "Selective operations of multi-pillar SOT-MRAM for high density and low power embedded memories," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 375–376.

[13] S. C. Song et al., "Unified technology optimization platform using integrated analysis (UTOPIA) for holistic technology, design and system co-optimization at <= 7–nm nodes," in *Proc. IEEE Symp. VLSI Circuits (VLSI-Circuits)*, Jun. 2016, pp. 1–2.

[14] L. Liebmann et al., "DTCO acceleration to fight scaling stagnation," *Proc. SPIE*, vol. 11328, pp. 62–76, Mar. 2020.

[15] O. Ali, M. K. Ishak, M. K. L. Bhatti, I. Khan, and K.-I. Kim, "A comprehensive review of Internet of Things: Technology stack, middlewares, and Fog/Edge computing interface," *Sensors*, vol. 22, no. 3, p. 995, Jan. 2022.

[16] M. Abrash, "Creating the future: Augmented reality, the next human-machine interface," in *IEDM Tech. Dig.*, Dec. 2021, pp. 1.2.1–1.2.11.

[17] H. Kwon et al., "XRBench: An extended reality (XR) machine learning benchmark suite for the metaverse," *Proc. Mach. Learn. Syst.*, vol. 5, pp. 1–20, Mar. 2023.

[18] N. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in *Proc. 44th Annu. Int. Symp. Comput. Architect. (ISCA)*, New York, NY, USA, 2017, pp. 1–12.

[19] Y.-H. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[20] T. F. Wu et al., "11.2 A 3D integrated prototype system-on-chip for Augmented reality applications using face-to-face wafer bonded 7–nm logic at $< 2\mu$m pitch with up to 40% energy reduction at Iso-area footprint," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, vol. 67, Feb. 2024, pp. 210–212.

[21] H. T. Kung, B. McDanel, S. Q. Zhang, X. Dong, and C. C. Chen, "Maestro: A memory-on-logic architecture for coordinated parallel use of many systolic arrays," in *Proc. IEEE Int. Conf. Appl. Specific Syst., Archit. Process. (ASAP)*, 2019, pp. 42–50.

[22] J. R. Hu, J. Chen, B.-k. Liew, Y. Wang, L. Shen, and L. Cong, "Systematic co-optimization from chip design, process technology to systems for GPU AI chip," in *Proc. Int. Symp. VLSI Design, Autom. Test (VLSI-DAT)*, Apr. 2018, pp. 1–2.

[23] C.-K. Cheng, C.-T. Ho, C. Holtz, and B. Lin, "Design and system technology co-optimization sensitivity prediction for VLSI technology development using machine learning," in *Proc. ACM/IEEE Int. Workshop Syst. Level Interconnect Predict. (SLIP)*, Nov. 2021, pp. 8–15.

[24] T. Srimani et al., "Ultra-dense 3D physical design unlocks new architectural design points with large benefits," in *Proc. Design, Autom. Test Eur. Conf. Exhibit. (DATE)*, Apr. 2023, pp. 1–6.

[25] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.

[26] V. Parmar, S. S. Sarwar, Z. Li, H.-H. S. Lee, B. D. Salvo, and M. Suri, "Exploring memory-oriented design optimization of edge AI hardware for extended reality applications," *IEEE Micro*, vol. 43, no. 6, pp. 40–49, Nov. 2023.

[27] N. Pantano, M. Stucchi, G. Van der Plas, and E. Beyne, "Impact of pitch scaling on 3D die-to-die interconnects," in *Proc. IEEE 74th Electron. Compon. Technol. Conf. (ECTC)*, May 2024, pp. 1064–1071.

[28] S. Das et al., "3D partitioning with pipeline optimization for low-latency memory access in many-core SoCs," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2024, pp. 1–5.

[29] L. M. G. Rocha et al., "Multidie 3-D stacking of memory dominated neuromorphic architectures," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 32, no. 11, pp. 2144–2148, Nov. 2024.

LEANDRO M. GIACOMINI ROCHA (Member, IEEE) received the Engineering degree in electronic integrated systems (Hons.) from the Grenoble INP, France, in 2015, the Engineering degree (Hons.) in computer engineering from Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, in 2016, and the Ph.D. degree (cum laude) in microelectronics from the Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, in 2020.

In 2019, he was a Visiting Researcher with Katholieke Universiteit Leuven (KUL) and imec, Leuven, Belgium. He is a Research and Development Engineer at imec, Leuven, Belgium since 2021, focusing on the System-Technology Co-Optimization (STCO) methodology for edge systems. He has authored or co-authored more than 30 journal and conference articles, and served as a technical program committee member to several conferences in the field.

Dr. Rocha received the Best Ph.D. Thesis Award from Brazilian Microelectronics Society (SBMicro) in 2021.

MOHAMED NAEIM received the B.Sc. degree from Zewail University of Science and Technology, Giza, Egypt, in 2019, and the M.Sc. degree in nanotechnology and nanoelectronics from KU Leuven, Leuven, Belgium, and TU Dresden, Dresden, Germany, in 2021. He is currently pursuing the Ph.D. degree in thermal-aware design enablement of sub-2-nm CMOS technology nodes and 3-D integration technology with ULB, Brussels, Belgium.

At TU Dresden, he worked as a Research Assistant with the Institute of Neuromorphic VLSI Systems and the Institute of Adaptive Dynamic Systems. The work is carried on as part of a joint degree program between Cadence Design Systems, San Jose, CA, USA, and Imec, Leuven.

GUILHERME PAIM (Member, IEEE) received the bachelor's degree (Hons.) in electronics engineering from the Universidade Federal de Pelotas (UFPel), Pelotas, Brazil, in 2015, and the Ph.D. degree (summa cum laude) in microelectronics from the Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, in 2021.

He is currently an Assistant Professor with the Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal. He is also an Integrated Researcher with the INESC-ID, Lisbon. He developed part of his thesis at Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in collaboration with the University of Stuttgart, Stuttgart, Germany, from 2019 to 2020. He is also a Post-Doctoral Researcher with MICAS Labs, KU Leuven, Leuven, Belgium, under the supervision of Prof. Verhelst, with a special focus on multicore AI system-on-chips in close collaboration with the system-to-technology co-optimization (STCO) program of the research and development, imec, Leuven. He is also a Technology Transfer Consultant for Microelectronics Projects with Brazilian Ministry of Science, Technology, and Innovation (MCTI), Brasilia, Brazil. He holds about 100 research papers in conferences and journals on Circuits, Architectures and Systems. He has been collaborating and leading work packages on several projects, such as EU H2021 CONVOLVE, ERC BINGO, CAPES/FCT ML-DSIV, and CAPES/PROBRAL ReACT.

Dr. Paim received the Best Ph.D. Thesis Award from Brazilian Microelectronics Society (SBMicro) and the Honor Thesis Award from the CAPES Research and Development Agency in 2022. His website is gppaim.wordpress.com.

MORITZ BRUNION received the bachelor's and master's degrees in electrical and computer engineering from the University of Bremen, Bremen, Germany, in 2019 and 2022, respectively.

He is currently a Researcher with Imec, Leuven, Belgium. His current research focuses on technology-driven interconnect architecture design.

PRIYA VENUGOPAL received the B.Tech. degree in electronics and communication engineering from Anna University, Chennai, India, in 2009, and the Ph.D. degree in electrical engineering from IIT Hyderabad, Hyderabad, India, in 2021.

She worked as a VLSI Design Engineer with Innovation Communication Systems, Hyderabad, India, from 2010 to 2012. She was with Global Foundries as a Principal Engineer, Bengaluru, India, from May 2021 to September 2021. She was a Post-Doctoral Researcher and currently is a Research and Development Engineer with imec, Leuven, Belgium. Her research interests include low-power power management circuits, and integrated circuits and systems in advanced technology nodes.

Dr. Venugopal was a recipient of the KIRAN Women Scientist Fellowship from the Department of Science and Technology, Government of India, from 2017 to 2019.

DRAGOMIR MILOJEVIC received the M.S. and Ph.D. degrees in electrical engineering from the Université libre de Brussels (ULB), Brussels, Belgium, in 1994 and 2004, respectively.

He holds the position of a Professor of Digital Electronics and Digital Systems Design with ULB. In 2004, he joined Imec, Leuven, Belgium, where he first worked on multiprocessor and network-on-chip architectures for low-power multimedia systems. Since 2008, he has been working on design enablement of 3D-IC. Today, part of STCO and 3-D programs with Imec, where he is working on system and design technology co-optimization of advanced technology nodes and technology-aware design of 3-D integrated circuits. He has authored or co-authored more than 100 journal and conference articles, and served as a technical program committee member to several conferences in the field.

JAMES MYERS (Member, IEEE) received the M.Eng. degree in electrical and electronic engineering from Imperial College, London, U.K., in 2004.

He spent 15 years at Arm, leading research from low-power circuits and systems, through printed electronics, to DTCO activities. He joined Imec, Leuven, Belgium, in August 2022 to lead the System Technology Co-Optimization Program, with the aim to build upon established DTCO practices to overcome the numerous scaling challenges foreseen for future systems. He holds 60 U.S. patents, has taped out 20 SoCs, and has presented at ISSCC and VLSI Symposium. He has published in IEDM and Nature.

MUSTAFA BADAROGLU received the master's degree in industrial management and the Ph.D. degree in electrical engineering from KU Leuven, Leuven, Belgium.

He is currently a Staff Program Manager with Qualcomm, Leuven, and a Site Manager with Imec, Leuven, where he is in charge of managing the realization of advanced technologies through leading semiconductor suppliers.

Dr Badaroglu is the Chair of the More Moore Section in IRDS.

MARIAN VERHELST (Fellow, IEEE) received the Ph.D. degree from Katholieke Universiteit Leuven, Leuven, Belgium, in 2008.

From 2008 to 2010, she was a Research Scientist with Intel Labs, Hillsboro, OR, USA. She is currently a Full Professor with MICAS Laboratories, Katholieke Universiteit Leuven, and the Research Director of imec, Leuven. Her research interests include embedded machine learning, hardware accelerators, HW-algorithm codesign, and low-power edge processing.

Dr. Verhelst is a member of the board of directors of tinyML and active in the TPCs of DATE, ISSCC, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, and ESSCIRC; and was the Chair of tinyML2021 and the TPC Co-Chair of AICAS2020. She received the Laureate Prize of the Royal Academy of Belgium in 2016, the 2021 Intel Outstanding Researcher Award, and the André Mischke YAE Prize for Science and Policy in 2021. She is an IEEE SSCS Distinguished Lecturer; a member of the Young Academy of Belgium and the STEM Advisory Committee to the Flemish Government; and an Associate Editor of IEEE Transactions on Very Large Scale Integration (VLSI) Systems, IEEE Transactions on Circuits and Systems—II: Express Bsriefs, and IEEE Journal of Solid-State Circuits.

JULIEN RYCKAERT received the M.Sc. degree in electrical engineering from the University of Brussels (ULB), Brussels, Belgium, in 2000, and the Ph.D. degree from the Vrije Universiteit Brussel (VUB), Brussels, in 2007.

He joined Imec, Leuven, Belgium, as a Mixed-Signal Designer in 2000, specializing in RF transceivers, ultralow power circuit techniques, and analog-to-digital converters. In 2010, he joined the process technology division in charge of design enablement for 3DIC technology. Since 2013, he has been in charge of Imec's design-technology co-optimization (DTCO) platform for advanced CMOS technology nodes. In 2018, he became the Program Director focusing on scaling beyond the 3 nm technology node as well as the 3-D scaling extensions of CMOS. Today, he is a Vice President of Logic in charge of Compute Scaling.

DWAIPAYAN BISWAS (Member, IEEE) received the M.Sc. degree in the system on chip and the Ph.D. degree in electrical engineering from the University of Southampton (UoS), Southampton, U.K., in 2011 and 2015, respectively.

From 2015 to 2016, he was a Post-Doctoral Research Fellow with UoS. In 2016, he joined Imec, Leuven, Belgium, as a Researcher on Digital IC Design for Biomedical Applications. He is currently the System Technology Co-optimization (STCO) Program Manager with Imec. He has authored several peer-reviewed journals, conferences, and edited a book. His current research interests include exploring the interception of Imec's advanced semiconductor technology on future compute system challenges, low-power VLSI design, and advanced technology for system optimization.