AERO: Design Space Exploration Framework for Resource-Constrained CNN Mapping on Tile-Based Accelerators

Simei Yang, Debjyoti Bhattacharjee, Vinay B. Y. Kumar, Saikat Chatterjee, Sayandip De, Peter Debacker, Diederik Verkest, Arindam Mallik, *Member, IEEE*, and Francky Catthoor, *Fellow, IEEE*

***Abstract*— Analog In-Memory Compute (AIMC) arrays can store weights and perform matrix-vector multiplication operations for Deep Convolutional Neural Networks (CNNs). A number of recent efforts have integrated AIMC arrays into hybrid digital-analog accelerators in a multi-layer parallel manner to achieve energy efficiency and high throughput. Multi-layer parallelism on large-scale tile-based architectures need efficient mapping support at the processing element (PE)-level (*e.g.*, digital or analog processing elements) and tile-level. To find the most efficient architectures, fast and accurate design space exploration (DSE) support is required. In this paper, a novel DSE framework, AERO, is presented to characterize a CNN inference workload executing on hybrid tile-based architectures that supports multi-layer parallelism. Three characteristics can be seen in our DSE framework: (1) It presents a hierarchical Tile/PE-level mapping exploration strategy including inter-layer interaction, and allowing layer fusion/splitting configurations for PE-level mapping optimization. (2) It unlocks different Performance, Power and Area (PPA) exploration points under both sufficient and limited resource constraints, while limited resource case is not considered in prior works of multi-layer parallel architectures. The impact of weight loading and weight stationary mapping are analyzed for better insights into hybrid tile-based architectures. (3) It incorporates a detailed PPA model that supports a broad range of hybrid digital and analog units in a tile. Experimental case-studies are performed for realistic and relevant benchmarks such as MLP, CNNs (Lenet-5, Resnet-18,-34,-50 and -101).**

***Index Terms*— Design space exploration, hybrid digital-analog accelerator, multi-layer parallelism, inference, AIMC, resource constrained mapping, deep learning.**

I. Introduction

DEEP learning applications, such as deep convolutional neural networks (CNNs) have achieved remarkable breakthroughs in various application domains (*e.g.*, image classification and speech recognition). These workloads involve compute intensive matrix-vector multiplication (MVM) operations and huge data communication (*e.g.*, activation and weights of CNNs). As CNNs grow deeper, the ever-increasing computing and communication requirements have led to the need for custom accelerator designs. Dedicated accelerators [1], [2] have been designed to speed up computation using a high degree of parallelism, and optimzed data storage and data movement, in order to improve energy efficiency.

In recent years, analog in memory compute (AIMC) arrays have attracted much attention in the accelerator design [3]. The AIMC array stores the CNN weights in SRAM (or NVM-resistive) array to achieve high storage density. In addition, an AIMC array can effectively execute multiple MVM operations simultaneously, thereby achieving low energy consumption and high throughput [4]. Prior works [5]–[7] integrate AIMC with digital components to create a hybrid digital-analog accelerator to support the execution of full CNN workloads. Digital components are introduced to deal with some non-MVM operations (*e.g.*, Normalization, Pooling, etc.) that are not easily adopted in AIMC arrays.

The state-of-the-art accelerator architectures (can be digital or hybrid digital-analog) can be broadly classified into two categories— *single-layer parallel* and *multi-layer parallel*. Single-layer parallel architectures (*e.g.*, digital [1], [2]) execute a single CNN layer on the architecture at any given point of time, parallelising and speeding up execution of operations within the layer. Execution of a layer is not started until the whole output of the preceding layer(s) has been computed. In such architectures, input and output activations are typically stored in the same memory hierarchies (*e.g.*, DRAM, buffer in Eyeriss [2]). Many DSE frameworks (*e.g.*, Timeloop [8], Interstellar [9], Zigzag [10]) explore loop transformations at memory hierarchies for high data reuse to optimize energy efficiency.

Multi-layer parallel architectures (*e.g.*, ISAAC [5] and PUMA [7]) are capable of processing multiple CNN layers simultaneously, allowing multi-layer pipelines to maximize the throughput of a full CNN workload. Such an architecture is often organized into multiple tiles, where each tile can be composed of multiple processing elements (PE) containing AIMC arrays and digital components. Typically, all the tile resources are partitioned across different CNN layers. A given CNN layer (with MVM operations) is processed by some AIMC arrays, while the outputs of the layer are fetched by other AIMC arrays to process the next CNN layer and so on. This helps in reduction of intermediate result storage

Manuscript received December 14, 2021; revised February 25, 2022 and April 14, 2022; accepted April 18, 2022. Date of publication May 2, 2022; date of current version June 13, 2022. This work was supported by the Flemish Government (AI Research Program). This article was recommended by Guest Editor J.-Y. Kim. *(Corresponding author: Simei Yang.)*

The authors are with IMEC, 3001 Leuven, Belgium (e-mail: simei.yang@imec.be).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JETCAS.2022.3171826.

Digital Object Identifier 10.1109/JETCAS.2022.3171826

TABLE I
COMPARISON OF RELATED WORK ON CNN ACCELERATORS

Reference	**Architecture & Workloads**	**Mapping approaches**			**④ Modeling supports**
		① At PE/AIMC-level	**②At Tile/PE-level**	**③At memory hierarchies**	
Timeloop [8] + Accelergy [12]	-Single-layer parallel architecture - Conv and FC	- Exploration: spatial mapping in AIMC/PE array - No tile-level mapping		- Exploration: loop transformation in L1, L2, DRAM	- PPA: high-level, considering MAC computation and memory access overhead
Interstellar [9]					
Zigzag [4], [10]					
ISAAC [5]	-Multi-layer parallel architecture - Full neural network	- Conventional mapping in AIMC arrays - Exploration: No	- *Exploration:* No (manually) - NF: No	- Analyze buffer requirement between two layers	- PPA: detailed, excluding Instr Mem & WL overheads - RC: sufficient
PUMA [7], [13]			- *Exploration:* HPA (bottom-up)	- Consider register allocation at compiler level	- Instr: custom ISA - PPA: detailed, excluding WL overhead - RC: sufficient
[6]		- Improve AIMC data reuse w.r.t. conventional approach - Exploration: No	- *Exploration:* No - NS: Yes - NF: No	- Discuss buffer requirement between two layers	- PPA: detailed, exlucing Instr Mem & WL overheads - RC: sufficient
[11]		- Improve AIMC utilization w.r.t. conventional approach - *Exploration:* No	- *Exploration:* No - NS: No - NF: No	- No exploration - Consider DP of activations in input Mem	- PPA: high-level, excluding WL overhead - RC: sufficient
This work		- Conventional mapping in AIMC array, aligned to [4] - *Exploration:* No	- *Exploration:* HPA (top-down) - NS: Yes - NF: Yes	- Assume small buffer between two layers as in [5] - Assume ideal DP of activations in input Mem	- Instr: custom ISA - PPA: detailed, considering Instr Mem & WL overheads - RC: sufficient & limited

Note: NS: node splitting, **NF**: node fusion, **HPA**: Hierarchical partitioning algorithm, **Instr**: instructions, **Mem**: memory, **WL**: weight loading, **RC**: resource constraint, **DP**: data placement. * The gray cells highlight the main difference between the closest related work (PUMA) and our work.

space as well as speeding up execution [6]. Unlike DSE frameworks for single-layer parallel architectures (*i.e.*, considering loop mapping at memory hierarchies), the existing DSE frameworks/methodologies for multi-layer parallel architectures (*e.g.*, in [5]–[7], [11]) and they focus more on mapping at AIMC-level and Tile/PE-level.

In this paper, a systematic design space exploration (DSE) framework, AERO, is presented to allow early and fast evaluations of a CNN inference workload executing on a hybrid multi-layer parallel tile-based architecture that integrates multiple AIMC arrays as well as digital components in each tile. The AERO framework presents a complete mapping flow, including *virtual mapping* and *physical mapping* (hierarchical mapping exploration) at tile-level, PE-level and AIMC-level. Two key novelties can be seen in our DSE framework.

- The proposed framework supports hierarchical mapping in multi-layer parallel architecture under both sufficient and limited resource constraints (available tiles/PEs). Existing works consider sufficient tiles/PEs to execute all layers of a full CNN simultaneously, assuming weights (of all layers) are pre-loaded into sufficient AIMCs. AERO unlocks different PPA exploration points under limited resources, which is not possible in prior efforts. Particularly, the impact of weight loading and weight stationary are assessed to gain a better insight into hybrid tile-based architectures.
- The proposed framework supports a broad range of hybrid digital and analog units within a tile, as well as instruction memory and weight loading overhead. Most existing DSE frameworks do not take into account instruction memory, weight loading overhead that have a big overhead or they consider quite high-level estimations (*e.g.*, for MAC computation and memory access overhead as in Timeloop [8], Interstellar [9], ZigZag [10]).

Moreover, experimental case-studies for MLP, (deep) CNNs (Lenet-5, Resnet-18,-34,-50 and -101) offer insights into the impact of design parameters, *e.g.*, (1) weight loading vs weight stationary, (2) external memory access bandwidth of weights, (3) resource constraints, (4) AIMC dimensions and (5) AIMC memory cell technology choices, on hybrid tile-based architectures.

The rest of the paper is organized as follows. Section II compares the existing DSE frameworks for accelerators and highlights the novelties of our work. Section III introduces a hybrid digital-analog tile-based accelerator template as an experimental benchmark in our DSE framework. Section IV describes the mapping methodology and PPA estimation approach of the proposed DSE framework (first contribution). Section V discusses the insights derived from the experimental case-studies (second contribution). Section VI discusses the advantages and limitations of the current AERO framework. Section VII concludes our work.

II. RELATED WORK

Table I compares the state-of-the-art DSE frameworks (or methodologies) related to *single-layer parallel* and *multi-layer parallel* CNN architectures, in terms of *mapping approaches* (*i.e.*, at AIMC level, Tile/PE-level, at memory hierarchies) and *model support* (*e.g.*, on PPA models, resource constraints).

On one hand, multiple systematic DSE frameworks, such as Timeloop [8]+Accelergy [12], Interstellar [9], Zigzag [10], focus on convolution (Conv) or Fully-connected (FC) layers executing on *single-layer parallel* architectures.

These frameworks are primarily designed towards exploration of digital architectural templates (*e.g.*, Eyeriss [2]) and corresponding mapping opportunities. The digital architectural template consists of a digital processing element (PE) array, along with the required memory hierarchies. A variant of Zigzag [4] considers an AIMC array alongside the required memory hierarchies. These DSE frameworks adopt loop transformation approaches into the mapping flow to represent and analyze the design space of Conv/FC layers mapped at PE/AIMC arrays (see ① column in Table I) and memory hierarchies (see ③ column).

Particularly for the mapping of data at different levels of memory hierarchies (see ③ column), *single-layer parallel* architectures typically consider input activations coming from DRAM to buffers (*e.g.*, L1 and/or L2 buffer) and PE/AIMC array, while the output activations are stored in the same memory hierarchies. The mapping flow is targeted to find energy-efficient solutions that increase the data reuse within the PE/AIMC array and also at different levels of the memory hierarchy. It is worth mentioning that these DSE frameworks consider high-level PPA models (see ④ column), focusing on MAC computation and memory access overhead.

On the other hand, there has been some works [5]–[7] [11], [13] consider *multi-layer parallel* architecture to support the execution of a full CNN. Typically, the *multi-layer* parallel architectures are organized in a tiled manner – one or more compute arrays (either digital or analog) are grouped together in the form of a tile and tiles communicate using network-on-chip (NoC). These architectures take considerably larger area compared to the *single-layer* parallel architectures and are suitable for high performance scenarios.

For AIMC-level mapping (*i.e.*, mapping weights onto AIMC arrays, see ① column), no exploration is performed in the works of [5]–[7], [11], [13]. These existing works are based on the conventional weight mapping method, aiming to further increase the utilization and data reuse of AIMC arrays to improve computational parallelism and energy efficiency. In the conventional approach, a 3D Conv kernel (*e.g.*, kx, ky, cin as in Fig. 3) is transformed across AIMC rows, and multiple 3D kernels (*e.g.*, cout as in Fig. 3) can be mapped across AIMC columns. Our work also adopts the conventional weight mapping in AIMCs for high utilization. More details about the mapping strategy can be seen in Section IV-A.

For tile/PE-level mapping (see ② column), the objective is to map *nodes* (*e.g.*, layers or operations) to tiles and PEs. The works of [5], [6], [11] perform the tile-level mapping manually without exploration. PUMA [7] explores the optimized tile/PE-level by a hierarchical partitioning strategy in *bottom-up* manner (*i.e.*, from PEs to tiles to accelerator). In contrast, a *top-down* hierarchical partitioning strategy (*i.e.*,from accelerator to clusters, to tiles/PEs) in used in our work, to reduce inter-tile communication traffic within reasonable exploration time. Moreover, our proposed DSE framework supports *node fusion* and *splitting* configurations (at PE-level level) as in TVM compiler[1] [15]. Similar to TVM, splitting factors are configured manually in the current AERO framework. However, fusion and splitting configurations are not supported in PUMA [7], though *splitting* is discussed in [5] and [6].

For mapping at memory hierarchies (*i.e.*, loop transformation, see ③ column), the existing works of [5]–[7], [11] do not perform exploration as in Timeloop [8] and Zigzag frameworks [10] (for *single-layer parallel* architectures). As previously discussed, *single-layer parallel* architectures store input and output activations of a layer in the same memory hierarchies. Differently, *multi-layer parallel* architectures typically store the output activations of one layer in the buffer of a neighboring PE as input activations for the next layer. The next layer can start computation without having to obtain all the outputs of previous layer.[2] This can be further understood as implicit assumptions in existing works ([5]–[7], [11]), that is, all loops related to output activations are ideally mapped to the buffer of a neighboring PE and the next layer can consume data before the buffer is full. This assumption still holds in our work. Besides, the AERO framework currently assume ideal data placement in input activation buffer (*e.g.*, SRAM), where the required data for generating an output is available as needed. More investigations on mapping *loops* at memory hierarchies and orchestrating data in input activation buffer (such as in [11]), will be considered in future work.

Our work is further compared against existing works in two more aspects (see ④ column). (1) While PUMA is a compiler which generates custom ISA (instruction set), our AERO framework can be regarded as a preliminary step before a real compiler. AERO uses custom ISA to generate *approximate* instructions, assuming ideal data placement as previously discussed. The approximate instructions are further used to derive information (*e.g.*, action counts of different hardware components, ISA execution cycles) for detailed PPA (*e.g.*, latency, energy) models. Most of the existing works exclude instruction memory and weight loading overheads (*e.g.*, [5], [6]), or they consider quite high-level estimations (*e.g.*, [8]–[11]). (2) The works of [5]–[7], [11] consider that the number of tiles is always sufficient to execute all neural network layers simultaneously, and weights of all layers are pre-loaded into sufficient AIMCs. This explains why weight loading overhead is excluded in their PPA models. In contrast, AERO considers the cases of *sufficient* and *limited* number of available tiles/PEs, thereby unlocking different PPA exploration points that were previously not possible. AERO also assesses the impact of weight loading to achieve better insight into hybrid tile-based architectures.

III. Architecture Template

This section introduces a multi-layer Tiled Analog In-Memory Accelerator (TANIA) architecture template, which is used for design space exploration. For reference, Table II lists the parameters in the architecture template and other variables used in rest of this paper. Fig. 1(a) illustrates the top level TANIA architecture template. It comprises of multiple clusters,

[1]TVM compiler does not directly/currently support *multi-layer parallel* architectures in the published version [14].

[2]The buffer between two layers can be in small size, which is about $inx \times ky \times cin$, where inx is the number of columns in input feature, ky is the number of rows in kernel, cin is the number of channels in input feature.

TABLE II
NOTATIONS USED IN THIS PAPER

Architecture template	
$N_{cluster}$	Number of clusters in the TANIA template
N_{tile}	Number of tiles in a cluster
$J = N_{cluster} \times N_{tile}$	Total number of tiles in the TANIA
N_a	Number of analog Aicores in a tile
N_d	Number of digital VFUs in a tile
$N_a + N_d$	Number of activation buffers in a tile
Neural network inference workload	
$Conv$	Convolution
$Gemm$	General Matrix Multiply
$Psum$	Partital Sum
$Bias$	Bias add
Add	Residual add
BN	Batch Normalization
NL	Non-Linearity
Virtual mapping	
$node_a$	A virtual node targeting an analog Core
$node_d$	A virtual node targeting a digital VFU
$node_r$	A virtual node targeting data reshape
I	Total number of virtual nodes
K	Total number of virtual communication edges
Physical mapping	
SR	Sufficient resource constraint
LR	Limited resource constraint
WL	Weight loading mode
WS	Weight stationary mode

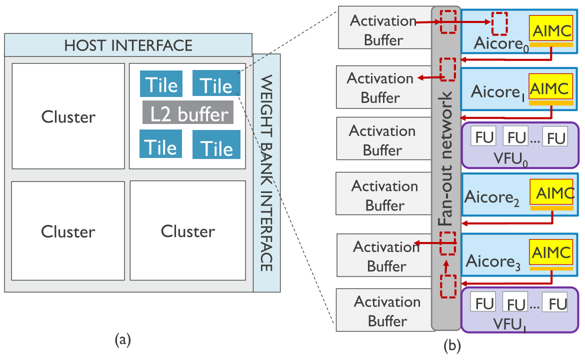


Fig. 1. TANIA architecture template: (a) description of the accelerator-level with multiple clusters; (b) description of the tile-level.

with two external interfaces – host interface and a weight bank interface. Each cluster consists of multiple tiles, which share a local (*e.g.*, L2) buffer for storing intermediate results if needed. The tiles communicate over NoC for transferring activations.

Each tile consists of multiple PEs, consisting of N_a analog Aicores [16] and N_d digital Vector Functional Units (VFUs), as shown in Fig. 1(b). As introduced in [16], each Aicore has an AIMC array to performs MVM operations and it has additional digital circuits for batch-normalization (BN) and non-linearity (NL) operations. This offers the possibility to immediately process the MVM results through BN or NL operations to reduce the amount of intermediate data. For simplicity, the number of BN and NL is considered to be equal to number of AIMC columns. Note that Aicore and AIMC are not the focus of this paper. More details about them can be found in [16] and [17]. Similar to [7], each VFU is a SIMD unit, which concurrently performs digital operation on multi-channel (*e.g.*, 64) activations for high throughput. Analog Aicores and digital VFUs support different operations, which are summarized in Table III. Each PE has a local activation buffer. A low fan-out network is used for communication between the PEs, which offers a much lower energy cost for

TABLE III
SUPPORTED OPERATIONS FOR AICORE AND VFU

Operations	Supported PE types
Conv, Gemm	Aicore
Pooling, Residual Add, Partial Sum (Psum)	VFU
Bias, Batch Normalization (BN), Non Linear (NL)	Aicore or VFU

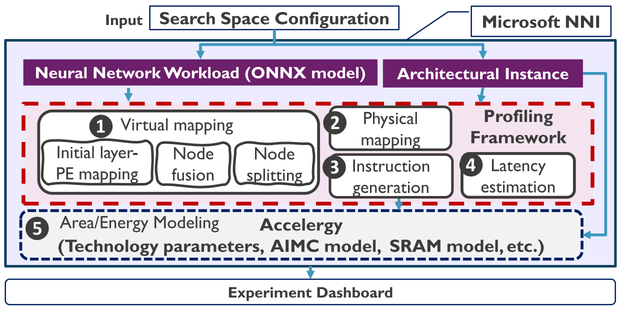


Fig. 2. Overview of AERO DSE framework.

data transfer within the tile, compared to inter-tile data transfer (via NoC). The output of a PE is written to the activation buffer of the destination PE, or to L2 buffer within a cluster.

Each PE has a local instruction memory and program counter. This allow PEs to operate independently, with the flexibility of mapping layers to arbitrary PEs. Since the cost of inter-tile communication is higher than intra-tile communication, a mapping strategy is required to map layers in a way that *reduces inter-tile communication*. The mapping framework will be presented in the following section.

IV. AERO DESIGN-SPACE-EXPLORATION FRAMEWORK

The overview of the proposed DSE framework, AERO, is shown in Fig. 2. The framework takes a hardware architecture template and a neural network (*e.g.*, CNN) inference workload as input. TANIA architecture template (See Section III) is used as the base hardware architecture in the framework. The inference workload is imported in Open Neural Network Exchange (ONNX) format [18]. ONNX format is used to enable interoperabililty among a variety of frameworks [18]. Microsoft NNI [19] is leveraged for performing multiple DSE experiments in parallel locally or on a cluster.

AERO framework encompasses a generic mapping methodology, which consists of virtual mapping (❶ in Fig.2) and physical mapping (❷). This generic mapping strategy can be applied to inference workloads executed on different accelerator templates (*e.g.*, ISAAC [5], PUMA [7]). Based on the mapping solution, preliminary instructions (assuming ideal data placement as discussed in Section II) are generated for each PE (❸). Additionally, a latency characterization approach (❹) is presented to model the start time and finish time of each pixel computation, considering inter-layer data dependencies. To estimate the area of an architecture template, and the energy cost for a given workload mapped to it, the Accelergy [12] framework is integrated into the AERO DSE flow (❺). A number of component-level estimation plugins specific to the architecture (e.g., for AIMC array, NoC, and

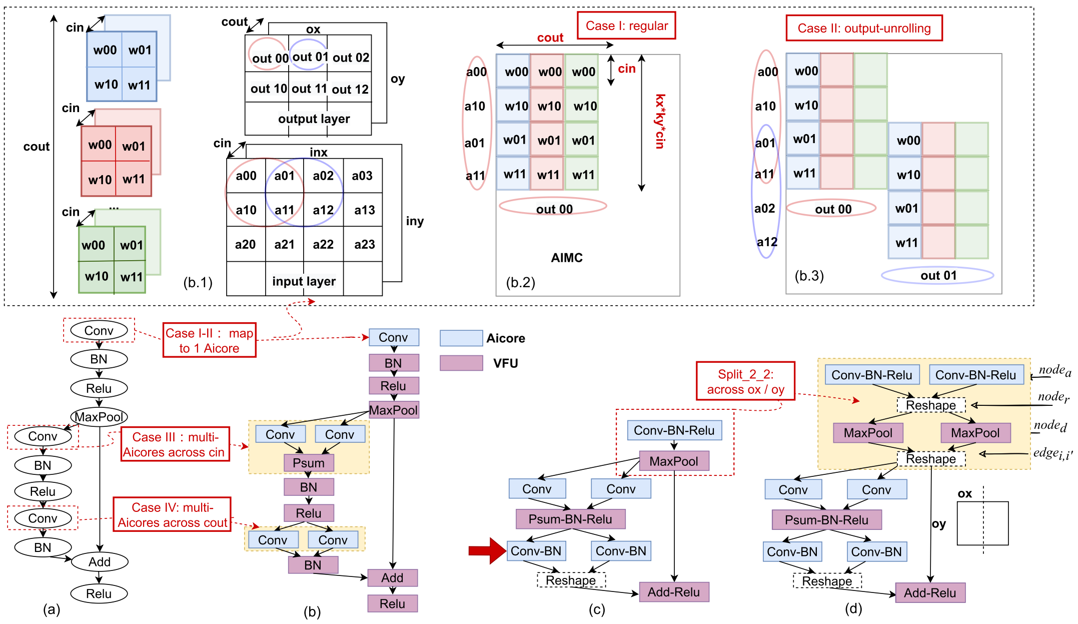


Fig. 3. Overview of virtual mapping. (a) Part of application graph of Resnet-18. (b) Virtual initial layer-PE mapping: Case I-II maps a MVM layer to one AIMC array (*i.e.*, one Aicore); Case III-IV map a MVM layer to multiple AIMC arrays by diving a Conv kernel across *cin* and *cout* dimension, respectively. (b.1) Example of a Conv layer with kernel, input and output features. (b.2) In Case I, AIMC mapping without output unrolling. (b.3) In Case II, AIMC mapping with output unrolling. (c) Virtual node fusion. (d) Virtual node splitting.

other building blocks) are added for energy computation. The generated results are displayed in a dashboard.

The following sections will present the details of mapping and PPA modeling. The custom ISA for instruction generation (❸) is out of the scope of this paper.

A. Virtual Mapping

Virtual mapping (❶ in Fig. 2) aims to map each neural network layer to Aicores or VFUs, depending on the types of layer, and considering AIMC dimension as constraints. Virtual mapping implies that the resource availability (*i.e.*, Aicores and VFUs) of the physical platform is not considered in this stage. Virtual mapping is fulfilled in three steps: *initial virtual mapping*, *virtual-node fusion* and *virtual-node splitting*, as illustrated by the example in Fig. 3.

1) Initial Layer-PE Virtual Mapping: For a neural network graph, initially, each MVM layer (*e.g.*, *Conv* and *Gemm*) is mapped to Aicores, and each digital operation layer is mapped to a VFU (see Fig. 3(a-b)). For Aicores, weights should be appropriately mapped to AIMC arrays (*i.e.*, at AIMC-level). The AIMC-level mapping can be summarized into four cases, depending on the kernel dimension and AIMC dimension [16]. Case I-II map a MVM layer to one AIMC (*i.e.*, in one Aicore). Case II maps multiple copies of weights to a (large) AIMC array, allowing computing multiple outputs simultaneously (*i.e.*, output unrolling in Fig. 3(b.3)). Case III-IV map a MVM layer to multiple AIMC arrays by dividing a Conv kernel across input (cin) and/or output channels ($cout$). More details about AIMC mapping can be found in [16].

2) Virtual-Node Fusion: The mapping of a layer to a Aicore/VFU is denoted (Fig. 3(b)) as a virtual node. The node fusion step fuses the adjacent operations of the same pixel into one PE (see Fig. 3(c)), aiming to reduce intermediate data transfer at memory hierarchies [20]. Based on the hardware supports shown in Table III, AERO performs node fusion according to the following three operation orders.

- Conv-Bias-BN-NL (fuse into Aicore)
- Psum-Bias-BN-NL (fuse into VFU)
- Add (residual)-Relu (fuse into VFU)

These orders are usually observed in most of the existing CNN workloads (*e.g.*, MLP, Resnet-18). Fusion can be performed even when some operations are skipped in each order. For instance, AERO fuses *Conv-BN* operations for Resnet-18 (see the arrow in Fig. 3(c)), while *Bias* and *NL* operations are skipped (in the first order). AERO currently incorporates a first fusion heuristic which will be explored more, as part of future work.

3) Virtual-Node Splitting: AERO performs node splitting to allow different PEs to generate different parts of the output activations (*e.g.*, ox and oy of a Conv layer). This can speed up the the computation of shallow layers (*e.g.*, first couple of layers due to the large output dimensions [6]) and consequently reduce the inference latency. In the example of Fig 3(c), *Split_2_2* refers to using two Aicores and two VFUs for first two layers respectively. The current AERO framework allows setting splitting factor for every layer manually. A good set of splitting factors can be obtained by adding an explicit search space exploration method, which can be elaborated in future.

4) Virtual Mapping Result: A virtual mapping can be characterized by a set of I virtual nodes ($node_i$), and a set of K communication edges ($edge_{i,i'}$) representing data dependencies between two nodes ($node_i$ and $node'_i$). The virtual nodes targeting an analog Aicore, a digital VFU and a *Reshape* operation (*e.g.*, for data organization in activation buffer), as shown in Fig. 3(b-d), are denoted by $node_a$, $node_d$ and $node_r$, respectively.

B. Physical Mapping

Physical mapping (❷ in Fig. 2) aims to map virtual nodes to the physical architecture entities: clusters, tiles, Aicores and VFUs. The objective is to optimize inter-tile communication over the NoCs, subject to resource constraints.

1) Problem Formulation of Physical Mapping: **Given:** a virtual mapping result, with I heterogeneous virtual nodes and K communication edges (see Section IV-A.4), and an architecture template (*e.g.*, TANIA in see Section III).

Find: An optimized physical mapping, which maps virtual nodes to clusters, tiles, PEs (*i.e.*, Aicores and VFUs). Let the matrix of *binary variables* $[p_{i,j}]_{I\times J}$ represent a physical mapping on tiles (including clusters), where $p_{i,j}$ is defined as:

$$p_{i,j} = \begin{cases} 1, & \text{if } node_i \text{ is assigned to } tile_j. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where J refers to the number of all tiles in the architecture ($J = N_{cluster} \times N_{tile}$). Here, the cluster index is hidden into the tile index ($tile_j$) for clarity of the formula.

Objective: To optimize NoC communication traffic, the inter-tile (including inter-cluster) communication is formulated as in Eq.(2), where $edge_{i,i'}$ refers to the edge between nodes $node_i$ and $node_i'$. Then the physical mapping objective can be formulated into Eq.(3).

$$Comm_{noc}(p_{i,j}, p_{i',j'}) = \begin{cases} 0, & \text{if } j = j'. \\ edge_{i,i'}, & \text{otherwise.} \end{cases} \tag{2}$$

$$\min \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} Comm_{noc}(p_{i,j}, p_{i',j'}) \tag{3}$$

Subject to: Resource constraints can be understood by: (1) Each virtual node (except $node_r$) is mapped into one tile (Eq.(4)); (2) The number of $node_a$ and $node_d$ in each tile should respect the tile-level resource availability of Aicores (Eq.(5)) and VFUs (Eq.(6)), respectively.

$$\sum_{j=0}^{J-1} p_{i,j} = 1, \quad \forall i, p_{i,j} \in \{0,1\}, node_i \notin node_r \tag{4}$$

$$\sum_{i=0}^{I-1} p_{i,j} \leq N_a, \quad \forall j, node_i \in node_a \tag{5}$$

$$\sum_{i=0}^{I-1} p_{i,j} \leq N_d, \quad \forall j, node_i \in node_d \tag{6}$$

Here, the physical mapping at PE-level is not shown for sake of simplicity. It is considered that the nodes in each tile can be allocated to any available Aicore or VFU. This is based on the assumption that intra-tile communication (*e.g.*, using fan-out networks in TANIA template) is more energy efficient compared to inter-tile communication via NoC.

2) Hierarchical Physical Mapping Strategy: A hierarchical strategy is proposed to achieve optimized physical mapping solution. Mapping on multi-cores systems has been proved as a NP-hard problem [21]. The large number of CNN layers (*i.e.*, virtual mapping nodes) and the large scale of hardware resources (*i.e.*, Aicores and VFUs in TANIA) lead to a huge design space. Hierarchical mapping approaches [22], [23] are usually used to resolve the scalability issue. The overview of our hierarchical physical mapping strategy is shown in Fig 4(a), which consists of three levels.

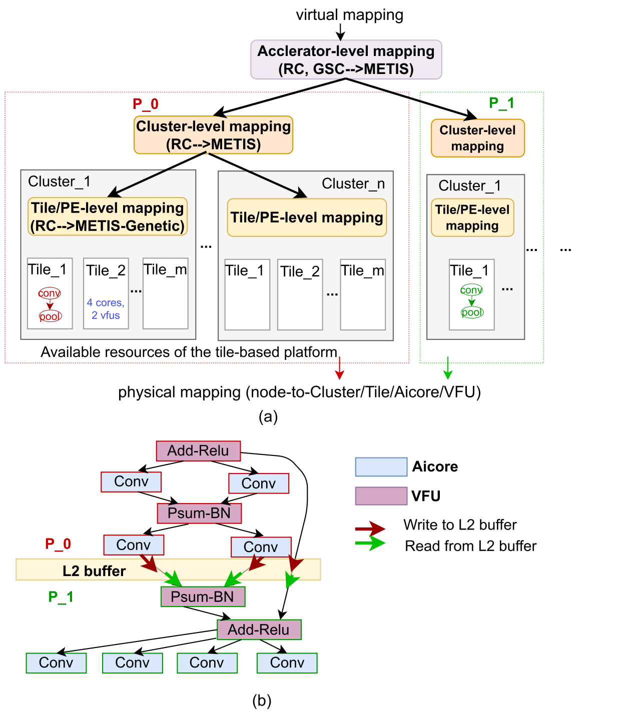


Fig. 4. (a) Overview of hierarchical optimization of physical mapping, under resource constraints (RC) and graph sequence constraints (GSC); (b) An example of accelerator-level physical mapping.

The proposed hierarchical physical mapping strategy is in *top-down* manner. The accelerator-level mapping divides virtual nodes into multiple partitions. It is particularly proposed for the case where the number of virtual nodes is larger than the number of physical resources (*i.e.*, Aicores and VFUs). The accelerator-level mapping can be further understood as *partitioning* a neural network to meet resource constraint and graph sequence constraint. It is considered that the nodes of different partitions execute sequentially at available resources, and the execution of each input node should not be later than its output node (*i.e.*, graph sequence constraint). One example is shown in Fig 4 (b), where virtual nodes are divided into two partitions (*i.e.*, P_0 and P_1). The nodes of P_0 execute on available PEs first (*e.g.*, 4 Aicores and 2 VFUs). After P_0 finishes the execution, the intermediate data between the two partitions are stored into L2 buffer. Then, nodes of P_1 load the intermediate data from L2 buffer to corresponding PEs and start their execution. Afterwards, according to resource constraints, the cluster-level mapping divides each set of partitioned nodes to available clusters, while the tile/PE-level mapping divides each set of clustered nodes to available tiles, as well as PEs.

3) Hierarchical Physical Mapping Strategy Implementation: *METIS* [24] graph partitioning program is used at the three hierarchical levels to support efficient explorations. *METIS* is claimed to be one to two orders of magnitude faster than other widely used partitioning algorithms. To further improve

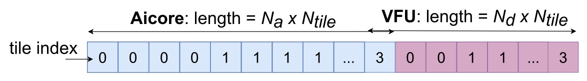

Aicore: length = N_a x N_{tile}
VFU: length = N_d x N_{tile}
tile index
0 0 0 0 1 1 1 1 ... 3 0 0 1 1 ... 3

Fig. 5. The chromosome with two segments for Aicores and VFUs in each cluster. In this example, a cluster has 4 tiles (with tile index from 0 to 3). Each tile has 4 Aicores (4 same tile indexes) and 2 VFUs (2 same tile indexes).

solution quality (*i.e.*, reducing inter-tile communication via NoCs, and exploiting the intra-tile communication), on top of *METIS*, a custom genetic algorithm is implemented at tile/PE-level to have local optimization. The combined strategy is denoted by *METIS-Genetic*.

The developed genetic algorithm is based a standard microbial genetic algorithm (MGA) [25]. MGA takes cluster-level physical mapping (*i.e.*, a set of virtual nodes mapped to a cluster) as input to explore tile/PE-level mapping optimization. Firstly, MGA initializes the population, which consist of a set of random chromosomes. The METIS solution (at tile/PE-level) is inserted as a basic chromosome. The chromosome of a tile/PE-level mapping solution is encoded as in Fig. 5. The chromosome is a vector of PE's length in a cluster (*i.e.*, $N_a \times N_{tile} + N_d \times N_{tile}$ in TANIA), consisting of two segments for Aicores and VFUs respectively. The chromosome is a permutation of tile index of each PE. In each target cluster, the virtual nodes for Aicores (or VFUs) are sequentially allocated to a tile (with a tile index) according to the Aicore (or VFU) chromosomes. Secondly, the fitness function of a chromosome is defined by $\exp(1/(\sum Comm_{noc} + 1))$,[3] where $\sum Comm_{noc}$ refers to the inter-tile NoC communication within a cluster (based on Eq.(3)). The smaller the NoC communication, the better the chromosome's fitness is.

The proposed hierarchical physical mapping strategy implemented by *METIS-Genetic* is able to exploit up to 82.01% communication data (Bytes) within tiles (*i.e.*, intra-tile), which enables more energy-efficient data transfer than inter-tile data transfer via NoCs (recall Section III). For Resnets (-34,-50, and -101), the METIS algorithm takes $0.01 \sim 0.06$ seconds and the genetic algorithm takes $3 \sim 10$ seconds for the exploration per cluster. This indicates that the proposed *METIS-Genetic* is suitable for the architecture template in terms of exploration efficiency and exploration time. The relevant experimental details are not shown due to the limited space in the paper.

C. Latency Characterization Approach

The proposed latency characterization approach (❹ in Fig. 2) describes start time and finish time of pixel computation, on-chip communication (*i.e.*, intra-tile and inter-cluster/tile communication) and off-chip data transfer behaviors, while taking into account inter-layer data dependencies.

Fig. 6 illustrates the latency characterization for two example neural network layers, Layer0 (*e.g.*, Conv layer) and Layer1 (*e.g.*, Pooling layer). At the beginning of system execution, weight loading and off-chip activation transfers are performed in parallel through separate NoCs (see A in Fig. 6). Similar to [5], data transfer via NoCs is assumed to be statically routed without any conflicts. In AERO framework, once

[3]1 is added to avoid the case where divisor is 0.

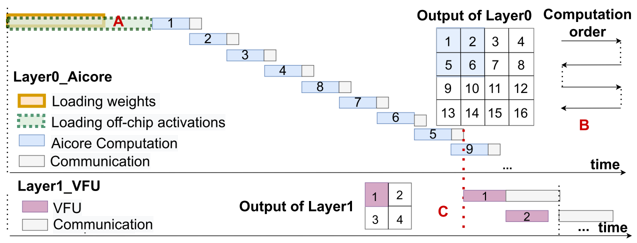


Fig. 6. Latency estimation for Layer0 and Layer1 examples. * For simplicity, the figure characterizes the weight loading (WL) latency in terms of data communication. WL latency caused by AIMC write speed (related to AIMC technology) is discussed in Section V-G.

all off-chip activation (of the first image frame) are loaded from off-chip memory, Layer0 starts to execute according to computation order sequentially (see B in Fig. 6). The pixel computation time is characterized by the average cycles of executing the instructions for one pixel generation, ignoring corner cases (*e.g.*, padding). Computation and inter-tile NoC communication can overlap. NoC communication latency is determined by data size and NoC bandwidth, *i.e.*,

$$\text{NoC communication latency} = \text{data size}/\text{NoC bandwidth}$$

The generated output pixels are written into the activation buffer (of Aicore/VFU) of the next layer. From the next layer (*e.g.*, Layer1), the computation for an output pixel depends on the maximum latency of the required input activations (*e.g.*, 1, 2, 5, 6 of Layer0, see C in Fig. 6).

Fig. 7 characterizes latency for all layers of a full CNN for sufficient and limited resource constraints. *Multi-layer parallel execution on hybrid tile-based architectures under limited resource constraints is the key novelty in this paper.*

1) Sufficient Resources (SR): In this case, the number of physical PEs (Aicores and VFUs) is always sufficient for a full neural network (with all virtual nodes). There is no need to use accelerator-level physical mapping strategy (in Section IV-B.2) to deal with accelerator resource constraints, but using cluster-level and tile/PE-level strategies. For a set of virtual nodes in Fig. 7 (a), part (b.1) illustrates the data transfer behaviors under SR constraints, while part (b.2) characterizes the corresponding latency behaviors. Initially, the off-chip data transfer, including *weight loading* (WL) for all $node_a$ and *image loading* (IL) for the input node (see IL_0 for $node_a^0$), are performed in parallel. When WL is finished for the first Aicore, it continues for the next Aicore (see T_A, $T_{A'}$ in Fig. 7 (b.2)), and so on. Then, all Aicores and VFUs can start execution with on-chip communication (intra-tile and inter-tile/cluster) depending on data availability. The finish time of the last layer refers to network inference latency (see T_B). From the second image, the off-chip IL latency (see IL_1 and IL_2) can be hidden in our assumption, and computation and communication can be performed without WL. This mode is typically called *weight stationary* (WS, see T_C and $T_{C'}$). As a consequence, latency (see $T_{B'}$) can be reduced compared to that of the first image.

2) Limited Resources (LR): In this case, the number of virtual nodes is larger than the number of physical PEs. It requires accelerator-level physical mapping strategy

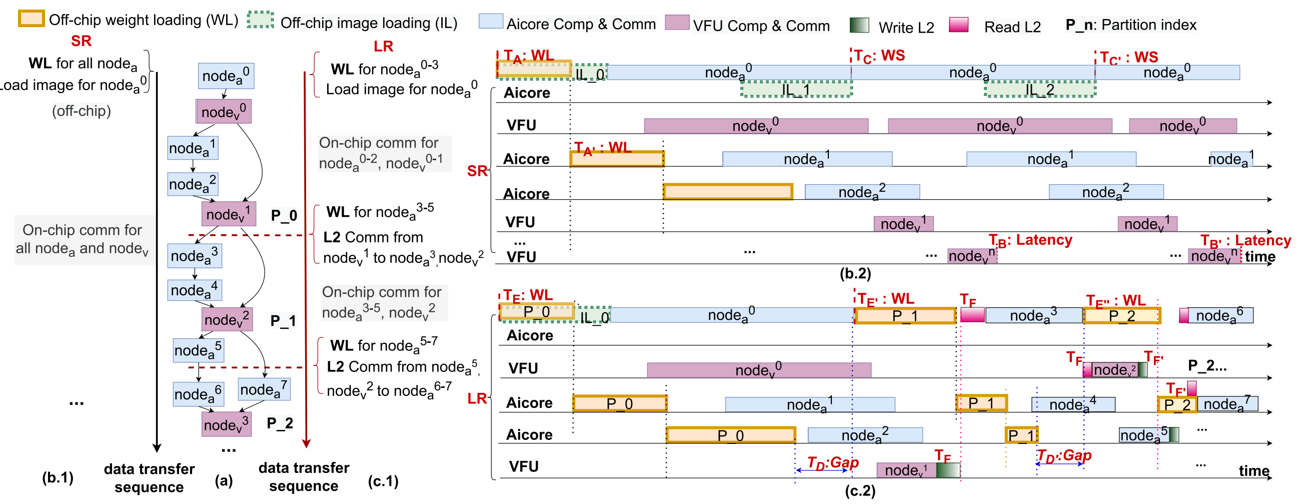


Fig. 7. (a) Example of virtual mapping nodes for a full network. (b.1) Data transfer behaviors of off-chip (WL and IL) and on-chip (intra-tile (via fan-out network), inter-tile/cluster (via NoC), inter-partition (via L2 buffer)) and (b.2) latency characterization, under SR constraint. (c.1) Data transfer behaviors of off-chip and on-chip and (c.2) latency characterization, under LR constraint. **Note:** The *Aicore (or VFU) comm & comm* characterizes the start time of the first pixel and the finish time of the last pixel for an image frame. *Particularly for part (c.2), different nodes mapped to the same Aicore/VFU are displayed in the same row.

(see Section IV-B.2) to divide the virtual mapping nodes into multiple partitions, followed by cluster-level and tile/PE-level strategies. For the virtual noes in Fig. 7 (a) under LR constraints, part (c.1) describes the data transfer behavior and (c.2) shows the corresponding latency characterization. Due to LR constraints, the virtual nodes are divided into 3 partitions. The off-chip WL is firstly performed for $node_a^{0-3}$ in P_0 (partition index), and IL is performed for $node_a^0$. Then all $node_a$ and $node_v$ in P_0 start computation and on-chip communication (intra-tile, inter-tile/cluster) considering data availability. When any physical Aicores are released by $node_a$ in P_0, WL can be performed for $node_a^{3-5}$ in P_1. This can lead to a time gap (see T_D) between the WLs for different partitions. Additionally, the AERO framework stores the on-chip intermediate data between partitions into L2 buffer. The intermediate data (from $node_v^1$, see T_F) is then used for the computation (of $node_a^3$ and $node_v^2$) in P_1. Similar off-chip WL and on-chip intermediate data (via L2 buffer) should be performed for different partitions until the end of the full network. As a result, the inference latency can be increased to get area benefits. The L2 intermediate data transfer depends on the applied accelerator-level physical mapping strategy.

The AERO framework considers both WL-mode and WS-mode for SR constraints. However, for LR constraints, the framework focuses on WL-mode (as previously discussed) and currently does not support WS-mode. WS-mode is possible for a set of partitioned nodes in multiple image frames, but a large L2 buffer space is required to store the intermediate data of the multiple frames. This is out of the context of this paper.

D. Energy and Area Estimation Approach

For area estimation of a given architectural instance and energy estimation corresponding to a specified workload, AERO integrates Accelergy [26], a component-level area and energy estimation methodology (❺ in Fig. 2). In this approach, a system architecture is hierarchically described in terms of components of various classes. For each component class, there needs to be a component-level estimator plugin (CEP) that can be queried—given the specified attributes on a component instance—for area, and energy corresponding to all actions defined on that component class. A CEP may support as fine a level of granularity for the feasible actions on a component (e.g,. read action on SRAM with the same address, or compute action on the AIMC array with specified array utilization). Accelergy also supports constructing *compound components* from the primitive ones, the actions on which are defined in terms of actions on the constituent primitive components.

Accordingly, the TANIA architecture template (Fig. 1) is described as a set of yaml files, in terms of various components such as: the external storage for weights-banks and host-side, input-activations, inter-tile NoCs (mesh), the analog AIMC arrays of various compute-cell types, a compound component containing the SIMD lanes in a VFU (composed in turn from primitive components like integer adders, multiplers, dividers, comparators, barrel-shifter, etc.,), the intra-tile fan-out network, and the on-chip activation (L1) and L2 level SRAM buffers. While Accelergy does include a few CEPs covering primitive components (e.g., arithmetic/combinational operations, registers, etc.,) with analytical models. Favoring more specificity, the estimation methods of some components where updated as necessary and new estimation plugins (CEPs) were added, for the components specific to Fig. 1. In particular, (highlighted in *blue text* in Fig. 8): an in-house model for SRAM buffers (corresponding to 22nm) is used; a model for AIMC array (modeling this as a *compound component* as in [27] was an option, but modeling it as a primitive with a dedicated CEP is a better choice in this case); the NoC is also modeled as a primitive component, and DSENT [28] was introduced as a new plugin; for the arithmetic and other digital components, the existing table-based plugin is leveraged

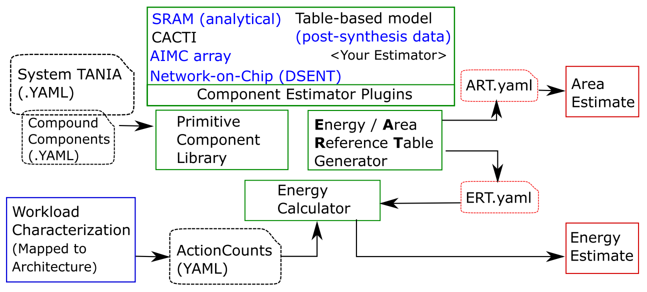


Fig. 8. Area and Energy estimation in AERO through Accelergy [26].

TABLE IV
CONSIDERED WORKLOADS

Networks	#FC	#Conv	#Pooling	NL types	Weights (MB)
MLP-3	3	0	0	Relu	0.159
Lenet-5	2	3	2	Tahn	0.015
Resnet-18	1	20	2	Relu	2.784
Resnet-34	1	36	2	Relu	5.193
Resnet-50	1	53	2	Relu	6.080
Resnet-101	1	104	2	Relu	10.596

* Ternary weights are considered for all experiments.

together with post-synthesis data (targeting the 22nm process and with Cadence Genus) based on RTL descriptions of respective components.

V. CASE-STUDIES

In this section, the AERO framework uses TANIA architecture template to study the impact of: (1) WL and WS, (2) external memory access bandwidth of weights, (3) SR and LR constraints, (4) AIMC dimensions and (5) AIMC memory cell technology choices. Notice that the experimental section does not include estimating the impact of host DRAM bandwidth and NoC bandwidth,[4] as they have less impact on inference latency than weight DRAM bandwidth.

A. Experimental Setup

The AERO framework was developed in Python and Microsoft NNI [19] was used for exploration of the design space. The experiments were executed on a machine running *Red Hat Enterprise Linux* with Intel CPU E5-2690 at 2.60GHz configuration. ONNX format is used [18] for input workloads. Table IV lists the workloads that are considered in experiments, including a MLP (denoted by MLP-3 with 3 Fully-connected layers), Lenet-5 (a small CNN) and multiple deep Resnets (-18,-34,-50 and -101). The MLP-3 [29] and Lenet-5 [30] workloads are exported to ONNX format based on their open-source code, while the Resnets are obtained from ONNX Zoo [31]. All the operations including digital operations, skip connection are modelled in the current framework. The table summarizes the number of different layers and the total weight size of each workload, considering ternary weights. In the experimental case-studies, 2 PEs are allocated to each of the first two layers (Split_2_2, recall Fig. 3 (d)) of the neural networks by default to optimize the inference latency. Split_2_2 can reduce inference latency by up to 70% with less than 13.5% increase in energy consumption (for Lenet-5 and Resnets). The relevant experimental details are not shown due to the limited space in the paper. Splitting is not performed for MLP-3, since the output dimension (*ox, oy*) of FC layer is 1. Note that the hardware technology nodes used for energy and area estimation haven been described in Section IV-D.

[4]The multi-layer parallel architecture offers the possibility to hide the latency of loading input images from host DRAM and intermediate data (*e.g.*, input, output and accumulation) via NoC.

TABLE V
BASELINE ARCHITECTURE PARAMETERS

Abstraction	Hardware Resources			
Accelerator-level	4 Clusters	Host DRAM (LPDDR4, 10 Bytes/cycle)	★Weight DRAM (HBM2, 40 Bytes/cycle)	Inter-cluster & Inter-tile NoC (60 Bytes/cycle)
Cluster-level	4 Tiles	L2 buffer (512KB)		
Tile-level	4 Aicores	2 VFUs	Tile Shuffler	
PE-level	Activation Buffer	Instruction Memory	AIMC array for Aicore (1152 rows × 512 cols)	

*Digital frequency=1GHz, Analog AIMC frequency=0.1GHz
*Activation Buffer: 1536 KB; Instruction Memory: 128 KB.
*★ is set for demonstration purpose (see Section V-D). This bandwidth is sufficient for CNNs (Lenet-5, Resnet-18), but leads to weight loading stalls for deep CNNs (Resnet-34,-50,-101).

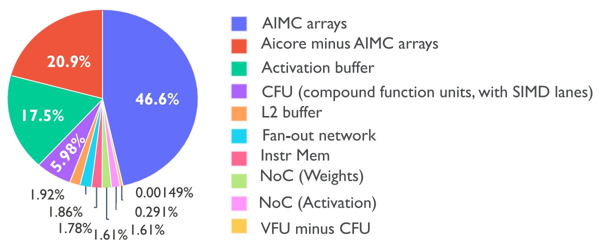


Fig. 9. Area breakdown of TANIA baseline architecture (4-cluster, 103.60 mm^2). Note: **CFU**: Compound function units (SIMD lanes).

B. Baseline Architecture and Area Analysis

Table V presents the baseline architecture used for the experiments. Given an area budget of 200 mm^2, 4 clusters can fit when using different dimensions (Row ∈ {512, 576, 1024, 1152} and Col ∈ {256, 512, 1024}) in each AIMC array. 1152 rows × 512 columns dimension is selected for the baseline architecture (as explained in Section V-F).

The area breakdown of the baseline architecture is presented in Fig. 9. The AIMC arrays contribute significantly (46.6%). to the overall area. The digital blocks inside the Aicores (pie in red) account for 20.9%. Activation buffers are another major contributor to area (17.5%). In modern day CNNs, the number of MAC operations dominate the total percentage of computations. For example, Resnet-18 has 3628.15 *million* analog MAC operations and 9.65 *million* digital operations. The area is dominated by AIMC arrays, which are responsible for performing the MAC operations that constitute the bulk of operations in the CNNs.

C. Impact of Weight Loading and Weight Stationary

This case-study assesses the impact of WL and WS under SR constraints. As previously discussed in Section IV-C.1, sufficient clusters allow loading weights of all layers at once.

TABLE VI
COMPARISON OF WL AND WS WITH SUFFICIENT CLUSTERS

Networks	#Clusters (SR)	Area (mm^2)	Energy efficiency ($TOPs/W$)			Pipeline throughput[5] (#inference / s)		
			WL[6]	WS[7]	Ratio=WS/WL	WL	WS	Ratio=WS/WL
MLP-3	1	25.90	0.442	262.267	593.364	23618	2×10^7	846.811
Lenet-5	1	25.90	1.087	1.707	1.570	239291	283849	1.186
Resnet-18	4	103.60	29.723	80.313	2.702	10836	17173	1.585
Resnet-34	7	181.31	28.739	92.165	3.207	6415	17174	2.677
Resnet-50	8	207.21	21.077	50.562	2.399	5426	11104	2.046
Resnet-101	15	388.52	17.048	48.789	2.745	3303	11105	3.362

* Pipeline throughput is calculated at $1GHz$ based on latency cycles.

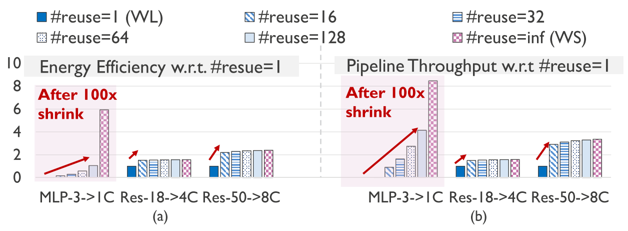


Fig. 10. (a) Normalized energy efficiency (TOPs/W) and (b) Normalized pipeline throughput (#inference/s) at different #reuse, under SR constraints. Note Res-18→4C (Resnet-18 mapped to 4 cluster). Note: since the values of MLP-3 are large, the plotted values are 100× smaller than the original values.

Table VI shows the resource requirements, indicating that as the application layer gets deeper (*e.g.*, from Resnet-18 to Resnet-101), the number of required clusters and corresponding area increase.

To reduce energy/latency overheads due to WL, weights can be loaded once for the first input (*e.g.*, image) and reused for the next multiple inputs (*i.e.*, denoted by *weight reuse*). When *#reuse>1*, multiple input images can be processed simultaneously and the pipeline throughput is dependent on the critical path (*i.e.*, layer with the longest computation time). When *#reuse=infinite*, the mapping can be approximated as WS.

Fig. 10 compares the energy efficiency (TOPs/W) and pipeline throughput (#inference/s) with increasing number of weight reuses. It can be observed that WL mode (*#reuse=1*) has the lowest energy efficiency. MLP-3 workload (with 3 FC layers) is sensitive to *#reuse* in terms of energy efficiency and performance, as FC layers are typically dominated by WL. Compared to WL mode (*#reuse=1*), WS mode (*#reuse=inf*) is ∼100× better in energy efficiency and performance.

On the other hand, for Resnet-18 and Resnet-50 (in Fig. 10), their energy efficiency and pipeline throughput approaches up to 90% of WS mode (*#reuse=inf*) when #reuse=16. This indicates that *weight reuse can enable energy efficiency and performance for CNNs close to WS mapping, even for low values.* Hence, the theoretical optimum can be approached with our architecture template and our mapping strategy. A broader comparison for all the considered workloads can be seen in Table VI. The table shows that WS mode has 1.570× ∼ 3.207× better energy efficiency and 1.186× ∼ 3.362× better performance than WL mode (for CNNs, excluding MLP-3).

D. Impact of External Memory Access of Weights

This case-study assesses the impact of external memory access bandwidth (BW) on latency in WL mode. An intuition overview can be seen Fig. 11(a), which describes the latency behavior of Res-34→7C when BW=40 Bytes/cycle. Indexes A and B refer to the behaviors of initial layers, where WL latency is significantly smaller than computation latency. Nevertheless, possible WL stalls (index C in Fig. 11(a)) can defer the execution of later layers. For better performance, multiple HBMs could be interfaced to achieve higher bandwidth, similar to TPUs [32].

Fig. 12 shows the impact of increasing weight DRAM bandwidth on latency. It can be observed that the latency of MLP-3 is sensitive to the increase in bandwidth due to its weight-rich FC layers. Besides, the latency of Lenet-5→1C and Res-18→4C do not change with increasing bandwidth, which implies that small CNNs are dominated by computation activities (w.r.t WL). Lastly, for deep neural networks with a huge amount of weights (*e.g.*, Resnet-34,-50,-101), an increase in weight DRAM bandwidth can reduce latency. Thus, the tiled architecture can take benefits of higher available weight bandwidth to reduce the inference latency of deep CNNs. But in the end, latency saturates to a critical path, which happens around bandwidth of 60 Bytes/cycle for Resnets.

E. Impact of Resource Constraint

This case-study assesses the impact of SR and LR constraints. In the baseline architecture with 4 clusters, the deep Resnets (Resnet-34,-50,-101) are resource constrained. For an intuitive overview, Fig. 11 compares the latency behavior of Res-34→7C and Res-34→4C (*i.e.*, SR and LR). Under SR constraints, WL is performed for all Aicores (mapped MVM layers) in layer depth order (see index A). Under LR constraints, the virtual nodes of Resnet-34 are divided into 2 partitions (see P_0 and P_1 in part (b-c)). WL is first performed for nodes in P_0 (see index D). Once Aicore resources are released by one or more node(s) in P_0, WL can be started for P_1 (see index E, with a gap between D and E). As previously illustrated in Section IV-B, the intermediate data between two partitions are transferred via the shared L2 buffer (see F and G, corresponding to *Add-Relu_3*, *Conv_35_0* and *Conv_35_1* in part (c)). Nodes in different partitions execute sequentially until the end of the last layer.

Fig. 13 shows the comparison of three cases: *LR-WL* (*i.e.*, limited resources in weight loading mode), *SR-WL* and *SR-WS* for the deep Resnets (-34,-50,-101). Part (a) indicates that *LR-WL* has the highest latency (about 18% higher than SR-WL, 35% ∼ 47% higher than SR-WS), with area saving

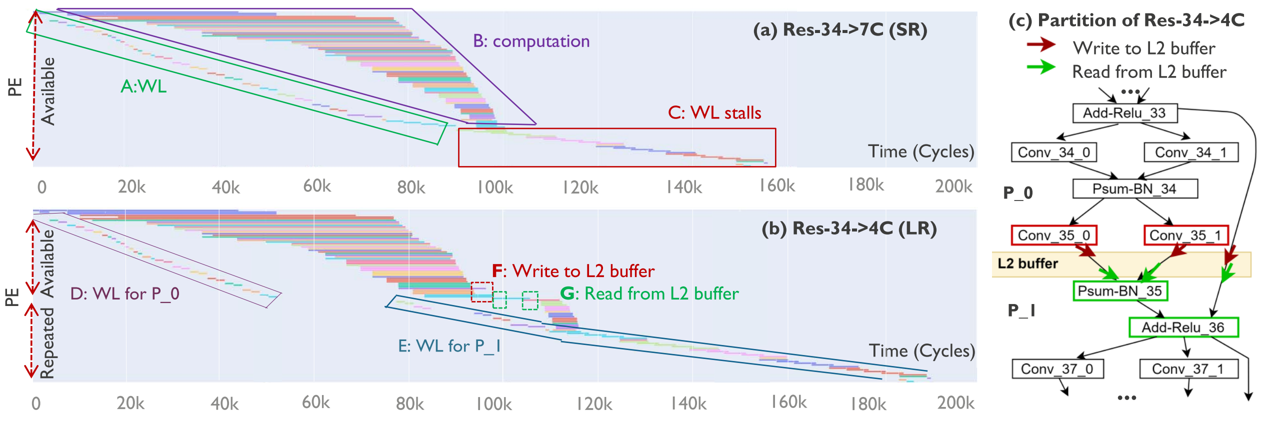


Fig. 11. Latency behavior Resnet-34 with of weight access bandwidth of 40 Bytes/cycle: (a) Res-34→7C (SR) and (b) Res-34→4C (LR). (c) Partition of Res-34→4C (*e.g.*, for *Conv_x_y*, *x* is the layer index and *y* is the mapping index on multiple Aicores (recall Section IV-A)). **Note**: **A**: weight loading (WL). **B**: computation. **C**: WL stalls. **E**: write to L2 buffer. **F**: read from L2 buffer. **G**: WL for P_0 (*e.g.*, partition index). **H**: WL and WL stalls for P_1. *Particularly for part (b), different nodes mapped to the same Aicore/VFU are displayed in different rows.

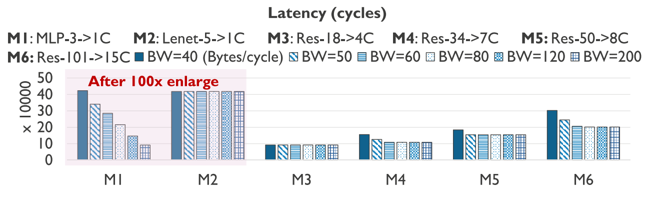


Fig. 12. Impact of external memory access of weights at increasing bandwidth (BW, Bytes/cycle) in WL mode, under SR constraints. Note: since the values of M1 and M2 are small, the plotted values are 100× the original values.

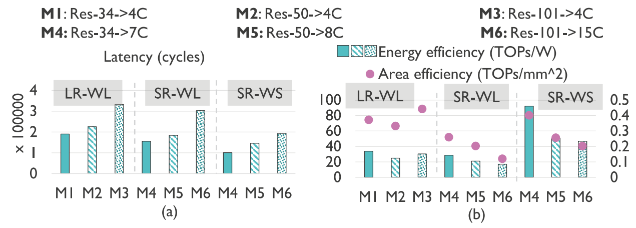


Fig. 13. (a) Latency (cycles) and (b) energy efficiency (TOPs/W) and area efficiency (TOPs/mm^2) in three cases: LR-WL (*i.e.*, limited resources in weight loading mode), SR-WL and SR-WS.

43% ∼ 73% (4 clusters w.r.t. 7, 8 and 15 clusters). This is because in *LR* case, a full network is divided into multiple partitions. WL and intermediate data transfer (via the shared L2) are performed for each set of partitioned nodes, which leads to extract overheads (*e.g.*, latency, energy). Fig. 13 (b) shows that *SR-WS* has the highest energy efficiency, followed by *LR-WL* (15% higher than *SR-WL*). Additionally, *LR-WL* has the highest area efficiency for Resnet-50,-101 (up to 54% higher than *SR-WS*). Since Resnet-34 is less resource constrained in 4 clusters than Resnet-50,-101, the area efficiency of Resnet-34 in *LR-WL* is close to (*i.e.*, 7% lower than) *SR-WS*. This proves that the AERO framework unlocks different PPA exploration points under LR constraints.

F. Impact of AIMC Dimensions

This case-study assesses the impact of AIMC dimensions. Fig.14 (a-b) shows the latency evolution of Res-18→4C (SR) and Res-50→4C (LR) across different AIMC dimensions (represented by Rows×Columns), considering WL overheads. Larger AIMC dimensions typically have lower latency, since higher output-unrolling factors (*i.e.*, more copies of weights in AIMC, recall Fig. 3 (b.3)) allow computing multiple output activations simultaneously. However, higher output-unrolling factors in a AIMC can take more time for WL. This explains why the latency of 1152 × 1024 is higher than 1152 × 512.

Fig. 14 (c-d) shows energy efficiency ($TOPs/W$) and area efficiency ($TOPs/mm^2$) for the two considered networks. Good energy efficiency can be achieved in large array dimension (*i.e.*, 1152 × 256, 1152 × 512, 1152 × 1024), out of which, 1152 × 512 AIMC dimension has the highest energy efficiency. However, the AIMC array dimensions with largest number of rows (1024) have low area efficiency, due to severe under-utilization of array. High area efficiency is achieved in the AIMC arrays with small number of rows (*e.g.*, with 256 rows). Similar trade-offs between latency, energy efficiency and area efficiency can be seen in other Resnets (*i.e.*, kernel dimension, $kx, ky = 3$). On the other hand, the small networks, MLP-3 and Lenet-5, have small variations in latency, energy efficiency, and area efficiency for different AIMC dimensions. For MLP-3→1C, 1024 × 512 has the best PPA, while 1152 × 512 is the second best.[5] For Lenet-5→1C (kernel dimension, $kx, ky = 5$), 1024 × 1024 has the lowest latency, while 1152 × 512 has higher energy and area efficiency.[6]

Overall, 1152 × 512 is selected in the baseline architecture, due to its good trade-off in latency, energy efficiency and area efficiency for the considered networks in this paper. To the best of our knowledge, AERO is the only framework that explicitly reports such trade-offs, under both SR and LR constraints, on hybrid tile-based architectures.

[5] 1152 × 512: 4234 cycles, 0.443 $TOPs/W$ and 0.013 $TOPs/mm^2$; 1152 × 512: 4234 cycles, 0.442 $TOPs/W$ and 0.012 $TOPs/mm^2$

[6] 1024 × 1024: 3451 cycles, 0.878 $TOPs/W$ and 0.002 $TOPs/mm^2$; 1152 × 512: 4179 cycles, 1.004 $TOPs/W$ and 0.005 $TOPs/mm^2$

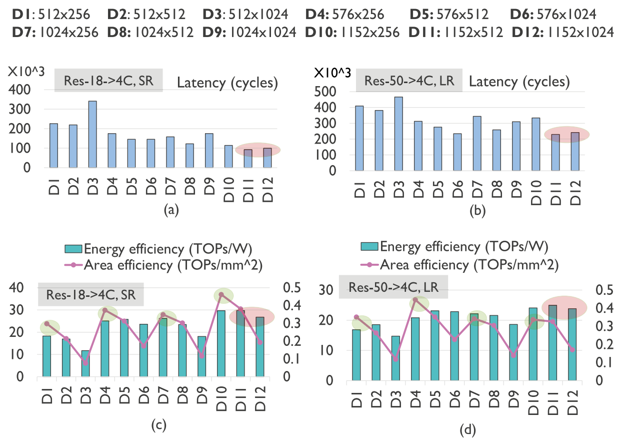


Fig. 14. (a-b) Latency evolution, (c-d) energy efficiency and area efficiency of Resnet-18 and Resnet-50 mapping in 4 clusters, considering different AIMC dimensions (Rows × Columns) and WL overheads.

TABLE VII
AIMC TECHNOLOGY CHOICES

Parameter	SRAM	IGZO	SOT-MRAM
Area	0.79 um^2	0.26 um^2	10.37 um^2
Write Energy	71.12 nJ	686.91 nJ	1191.50 nJ
Compute Energy	1.06 nJ	0.693 nJ	1.66 nJ
Write Latency	1 ns	10 ns	3 ns [33]
Compute Latency	2.3 ns	0.5 ns	2 ns [34]

*Write energy is for writing a row of cells. Compute energy is for MVM operations in one AIMC cycle (10 ns, 0.1GHz in the baseline architecture).
*Compute latency of SOT-MRAM is in 45nm node [34], while other parameters are in 22nm node (in-house data). The largest parameter in each row is highlighted.

G. Impact of AIMC Memory Cell Technology Choices

This case-study evaluates the impact of AIMC compute cell technology choices, *i.e.*, SRAM, IGZO and SOT-MRAM [17]. For the three technology choices, Table VII compares their area, energy and latency parameters.

In the baseline architecture, the digital frequency (f_D) and analog AIMC frequency (f_A) are set to 1 GHz and 0.1 GHz, respectively (see Table V). In our assumptions, when the AIMC write latency (L_W, affecting the speed of WL) is the same as the digital cycle (1 ns), there is no WL stall due to AIMC write (*e.g.*, for SRAM). The higher L_W of IGZO and SOT-MRAM results in additional WL stalls at f_D, and their WL speeds are 10× and 3× slower than SRAM, respectively. On the other hand, AIMC compute latency (L_C) limits the maximum analog frequency ($max(f_A) = \frac{f_D}{L_C}$) and has a large impact on the overall latency for a given inference workload. In WS mode, ideally, the inference latency of the three technologies is approximately proportional to their L_C. Considering DAC/ADC requirement between analog and digital components, this case-study fixes $f_A = 0.1GHz$ (with 10 ns analog cycle). Since the L_C of the three technologies are within the analog cycles, the compute latency of SOT-MRAM (45nm node [34]) caused by different technologies (w.r.t. 22nm node, see Table VII) does not affect the AIMC computation nor the experimental results.

Fig. 15 compares normalized energy and latency, energy efficiency and area efficiency (w.r.t. SRAM) for different AIMC technologies. Two observations can be made.

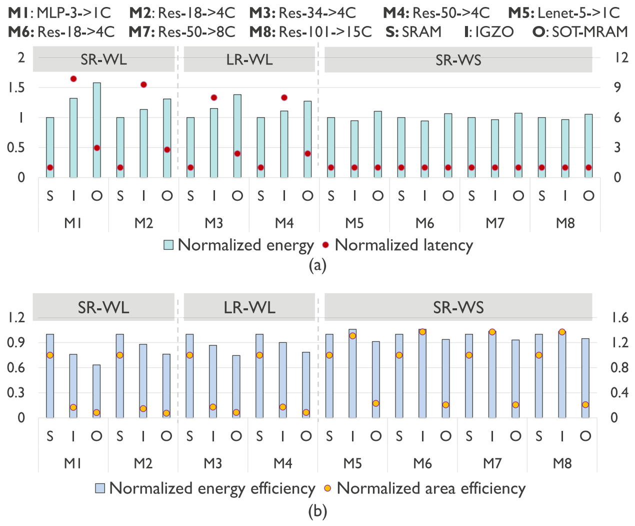


Fig. 15. (a) Normalized energy and latency, along with (b) normalized energy efficiency (TOPs/W) and area efficiency (TOPs/mm^2) for SRAM (S), IGZO (I) and SOT-MRAM (O) AIMC technology choices.

- In WL mode (SR-WL and LR-WL, under SR and LR constraints), SRAM leads to the lowest energy and lowest latency. This is because SRAM has the minimum AIMC write energy and write latency. Consequentially, SRAM has the highest energy efficiency and area efficiency.
- In WS mode, IGZO results in the lowest energy, due to its lowest compute energy. The latency is comparable across the three AIMC technologies. Besides, IGZO has the highest energy efficiency and area efficiency.

VI. DISCUSSION

This section discusses the advantages and limitations of the proposed AERO DSE framework. As presented in the previous sections, AERO is able to assess the impact of several design parameters and derive insights into tile-based hybrid accelerators, such as:

- Low values of weight reuse (*e.g.*, 16) can still enable energy efficiency and performance for CNNs close to weight stationary mapping.
- Higher weight access bandwidth can reduce the inference latency of deep CNNs. But in the end, latency saturates to a critical path, which happens around 60 Bytes/cycle.
- *LR-WL* is a PPA trade-off solution. It has 18% higher latency than *SR-WL*, with 15% higher energy efficiency and 73% lower area. Compared to *SR-WS*, *LR-WL* has close or even 54% higher area efficiency.
- The 1152 col×512 row AIMC has a good trade-off in inference latency, energy efficiency and area efficiency for Resnets as well as Lenet-5 and MLP-3.
- For AIMC cell technologies (SRAM, IGZO and SRAM), SRAM has lowest energy/latency in WL mode, while IGZO has the best energy/area efficiency in WS mode.

Further, AERO supports a broad range of DSE options to facilitate co-exploration of mapping, architecture and technology decisions. Table VIII summarizes the currently supported DSE options in AERO.

The current version of AERO has two main limitations. First, latency estimation at the NoC level does not model

TABLE VIII
SUPPORTED DSE OPTIONS IN AERO

Abstraction	DSE options
Accelerator-level	#Clusters; NoC design / technology parameters; Instr Mem Dim (**E, A**); Bandwidths of host & weight bank & NoC (**L**)
Cluster-level	#Tiles in a cluster; L2 buffer dimensions
Tile-level	Aicores-VFU organization in a tile; Actbuf Dim
PE-level	#Rows & #Colums of AIMC; AIMC cell technology; ADC & DAC precision; AIMC & VFU compute speed
Mapping	Fusion operation order; Splitting factor configuration

* **Instr Mem**: Instruction memory, **Actbuf**: Activation buffer, **Dim**: dimension, **L**: Latency-only, **E**: Energy-only, **A**: Area-only

communication contention. Second, loop transformation exploration over the memory hierarchy is not supported. It must be noted, however, that existing loop transformation DSE tools (*e.g.*, Timeloop [8], Interstellar [9], Zigzag [10] for single-layer parallel architecture) are too pessimistic for multi-layer parallel architectures as they do not take into account the dynamic data-flow between two neighboring layers.

VII. CONCLUSION

This paper presents a DSE framework, AERO, for a hybrid digital-analog CNN accelerator architecture that supports multi-layer parallel execution. The AERO framework presents a general mapping flow, including virtual mapping (allowing node fusion, node splitting configurations) and physical mapping (a hierarchical strategy) to fulfil the mapping at accelerator-level, cluster-level, tile/PE-level and AIMC-level. Besides, AERO incorporates detailed PPA models to characterize the latency, energy and area of a full neural network executing under sufficient and limited resource constraints. The experimental case-studies on MLP, Lenet-5 and Resnets (-18, -34, -50, and -101) derive insights of design parameters, such as weight loading scenarios, and resource constraints, on the tile-based hybrid architectures.

REFERENCES

[1] Y. Chen *et al.*, "DaDianNao: A machine-learning supercomputer," in *Proc. 47th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2014, pp. 609–622.

[2] Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 367–379, 2016.

[3] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-memory chips for deep learning: Recent trends and prospects," *IEEE Circuits Syst. Mag.*, vol. 21, no. 3, pp. 31–56, Aug. 2021.

[4] P. Houshmand *et al.*, "Opportunities and limitations of emerging analog in-memory compute DNN architectures," in *IEDM Tech. Dig.*, Dec. 2020, pp. 1–29.

[5] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with *in-situ* analog arithmetic in crossbars," *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 14–26, 2016.

[6] X. Peng, R. Liu, and S. Yu, "Optimizing weight mapping and data flow for convolutional neural networks on RRAM based processing-in-memory architecture," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.

[7] A. Ankit *et al.*, "PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Proc. 24th Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, Apr. 2019, pp. 715–731.

[8] 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.

[9] X. Yang *et al.*, "Interstellar: Using Halide's scheduling language to analyze DNN accelerators," in *Proc. 25th Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, Mar. 2020, pp. 369–383.

[10] 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.

[11] M. Dazzi, A. Sebastian, L. Benini, and E. Eleftheriou, "Accelerating inference of convolutional neural networks using in-memory computing," *Frontiers Comput. Neurosci.*, vol. 15. p. 63, Aug. 2021.

[12] Y. N. Wu, J. S. Emer, and V. Sze, "Accelergy: An architecture-level energy estimation methodology for accelerator designs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2019, pp. 1–8.

[13] J. Ambrosi *et al.*, "Hardware-software co-design for an analog-digital accelerator for machine learning," in *Proc. IEEE Int. Conf. Rebooting Comput. (ICRC)*, Nov. 2018, pp. 1–13.

[14] *TVM: Open Deep Learning Compiler Stack*. Accessed: Dec. 1, 2021. [Online]. Available: https://github.com/apache/tvm

[15] T. Chen *et al.*, "TVM: An automated end-to-end optimizing compiler for deep learning," in *Proc. 13th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2018, pp. 578–594.

[16] D. Bhattacharjee *et al.*, "Design-technology space exploration for energy efficient AiMC-based inference acceleration," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.

[17] S. Cosemans *et al.*, "Towards 10000TOPS/W DNN inference with analog in-memory computing—A circuit blueprint, device options and requirements," in *IEDM Tech. Dig.*, Dec. 2019, pp. 2–22.

[18] *Open Neural Network Exchange*. Accessed: May 3, 2022. [Online]. Available: https://onnx.ai/

[19] *Neural Network Intelligence*. Accessed: Nov. 9, 2021. [Online]. Available: https://www.microsoft.com/en-us/research/project/neural-network-intelligence/

[20] M. Alwani, H. Chen, M. Ferdman, and P. Milder, "Fused-layer CNN accelerators," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2016, pp. 1–12.

[21] H. Aydin and Q. Yang, "Energy-aware partitioning for multiprocessor real-time systems," in *Proc. Int. Parallel Distrib. Process. Symp.*, 2003, p. 9.

[22] W. Quan and A. D. Pimentel, "A hierarchical run-time adaptive resource allocation framework for large-scale MPSoC systems," *Design Autom. Embedded Syst.*, vol. 20, no. 4, pp. 311–339, Dec. 2016.

[23] M. Götzinger, A. M. Rahmani, M. Pongratz, P. Liljeberg, A. Jantsch, and H. Tenhunen, "The role of self-awareness and hierarchical agents in resource management for many-core systems," in *Proc. IEEE 10th Int. Symp. Embedded Multicore/Many-Core Syst.-Chip (MCSOC)*, Sep. 2016, pp. 53–60.

[24] *METIS—Serial Graph Partitioning and Fill-Reducing Matrix Ordering*. Accessed: May 23, 2022. [Online]. Available: http://glaros.dtc.umn.edu/gkhome/metis/metis/overview

[25] I. Harvey, "The microbial genetic algorithm," in *Proc. Eur. Conf. Artif. Life*. Berlin, Germany: Springer, 2009, pp. 126–133.

[26] Y. N. Wu, J. S. Emer, and V. Sze, "Accelergy: An architecture-level energy estimation methodology for accelerator designs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2019, pp. 1–8.

[27] Y. N. Wu, V. Sze, and J. S. Emer, "An architecture-level energy and area estimator for processing-in-memory accelerator designs," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Aug. 2020, pp. 116–118.

[28] C. Sun *et al.*, "DSENT—A tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *Proc. IEEE/ACM 6th Int. Symp. Netw.-Chip*, May 2012, pp. 201–210.

[29] MNIST. *Multi-Layer Perceptron*. Accessed: May 23, 2022. [Online]. Available: https://github.com/iam-mhaseeb/Multi-Layer-Perceptron-MNIST-with-PyTorch/blob/master/mnist_mlp_exercise.ipynb

[30] *Lenet-5*. Accessed: May 23, 2022. [Online]. Available: https://github.com/activatedgeek/LeNet-5/blob/master/lenet.py

[31] *ONNX Model Zoo*. Accessed: May 23, 2022. [Online]. Available: https://github.com/onnx/models

[32] *High-Bandwidth Memory (HMB)*. Accessed: Dec. 1, 2021. [Online]. Available: https://www.amd.com/system/files/documents/high-bandwidth-memory-hbm.pdf
[33] M. R. Sarkar, M. M. A. Bappy, M. M. Azmir, D. M. Rashid, and S. I. Hasan, "VG-SOT MRAM design and performance analysis," in *Proc. IEEE 12th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2021, pp. 715–719.
[34] Z. He, Y. Zhang, S. Angizi, B. Gong, and D. Fan, "Exploring a SOT-MRAM based in-memory computing for data processing," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 4, no. 4, pp. 676–685, Oct. 2018.

Simei Yang received the B.Sc. and M.Sc. degrees in microelectronics (integrated circuit engineering) from Sun Yat-sen University, Guangzhou, China, in 2014 and 2016, respectively, and the Ph.D. degree in the embedded system with IETR UMR CNRS 6164, Polytech Nantes, Université de Nantes, France, in 2020. She is currently holds a post-doctoral position at IMEC and KU Leuven. Her research interests include DNN evaluation framework on compute-in-memory based accelerators, run-time management of energy efficiency on multi/many-core systems, and system-level modelling and simulation.

Debjyoti Bhattacharjee received the B.Tech. degree in computer science and engineering from the West Bengal University of Technology (WBUT), Kolkata, West Bengal, India, in 2013, the M.Tech. degree in computer science from the Indian Statistical Institute, Kolkata, India, in 2015, and the Ph.D. degree in computer science and engineering from Nanyang Technological University, Singapore, in 2019. He worked as a Research Fellow with Nanyang Technological University for a year. During his doctoral studies, he worked on design of architectures using emerging technologies for in-memory computing. He is currently a Research and Development Team Leader at IMEC, Leuven. He developed novel technology mapping algorithms, technology-aware synthesis techniques, and proposed novel methods for multi-valued logic realization. His current research interests include machine learning accelerator using analog hardware, hardware design automation tools, and application-specific accelerator design, with emphasis on emerging technologies.

Vinay B. Y. Kumar received the B.Tech. and M.Tech. degrees microelectronics and the Ph.D. degree in EE from the Indian Institute of Technology Bombay in 2008 and 2019, respectively. Prior to Ph.D., he was at the Davinci Innovation Center, ASUS, Taipei, where he worked on applied speech processing. He is currently a Research and Development Team Leader at IMEC, Leuven, with a focus on uncore system modeling, to enable pathfinding future compute systems. Prior to this, he was a Post-Doctoral Research Fellow with the School of CSE, Nanyang Technological University, Singapore, where he contributed to the SOCure project, a programme toward secure SoC design.

Saikat Chatterjee received the B.E. degree in instrumentation and electronics engineering from Jadavpur University (JU), India, the M.S. degree in instrumentation from the Indian Institute of Science (IISc), Bengaluru, India, and the Ph.D. degree in intelligent systems from Bielefeld University, Germany. He is currently a Research Engineer at IMEC, Leuven, Belgium. His research focuses on computer architecture, ASIC design, block, and chip level verification.

Sayandip De received the B.Tech. degree in electronics and communication engineering from the Kalyani Government Engineering College (KGEC), India, and the M.Tech. degree in VLSI design from the Indian Institute of Engineering Science and Technology (IIEST) Shibpur, India. He is currently a Researcher at IMEC, Leuven, Belgium, and a Guest Research Candidate with the Electronic Systems Group, Eindhoven University of Technology (TU/e). His research focuses on computer architecture, approximate computing, design automation for low power circuits and systems, and optimizing machine learning inference.

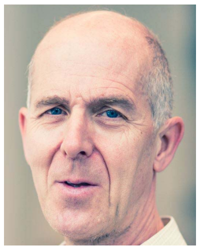
Peter Debacker received the M.Sc. (Hons.) degree in electrical engineering from the Katholieke Universiteit Leuven, Leuven, Belgium, in 2004.

Before joining IMEC, Leuven, in 2011, he worked at Philips as a System Engineer and at Essensium as a System Architect. At IMEC, he is currently a Program Member and leads a team that researches architecture, algorithms, and circuits and devices to create efficient AI hardware ranging from large and tiny DNN accelerators to in-memory compute and neuromorphic hardware. Before that, he has worked on IMEC's low-power digital chip and processor architectures and implementation in advanced technology nodes to optimize power-performance-area (PPA) optimization of scaled CMOS technologies (for 3 nm and beyond). His current research interests include AI, machine learning and neuromorphic computing, processor and computer architectures, design methodologies, design-technology co-optimization, and digital chip design.

Dr. Debacker was the Co-Chair of the tinyML EMEA Technical Forum 2021 and will be the Chair of the 2022 tinyML EMEA Forum.

Diederik Verkest received the Ph.D. degree in applied sciences from the University of Leuven, Belgium. He started working with the VLSI Design Methodology Group, IMEC, Leuven, Belgium, on hardware/software co-design, re-configurable systems, and multi-processor system-on-chips in the domain of wireless and multimedia. From 2010 to 2020, he was responsible for IMEC's Logic Institute Research Program in which the leading design and process companies jointly work on co-optimization of CMOS design and process technology for N+2 nodes. In recent years, he focused on design and process technology optimization for ML accelerators. He published and presented over 150 papers in international journals and at international conferences.

Arindam Mallik (Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering and computer science from Northwestern University, USA, in 2004 and 2008, respectively. He leads the Future System Exploration (FuSE) Group in the Compute System Architecture (CSA) Research and Development Unit. He is a technologist with 20 years of experience in semiconductor research. He has authored or coauthored more than 100 papers in international journals, conference proceedings, and holds number of international patents. His research interests include novel computing systems, design-technology co-optimization, and economics of semiconductor scaling.



Francky Catthoor (Fellow, IEEE) received the Ph.D. degree in EE from the Katholieke Universiteit Leuven, Belgium, in 1987. From 1987 to 2000, he has headed several research domains in the area of synthesis techniques and architectural methodologies. Since 2000, he has strongly involved in other activities at IMEC, including co-exploration of application, computer architecture and deep submicron technology aspects, biomedical systems and IoT sensor nodes, and photo-voltaic modules combined with renewable energy systems, all at IMEC, Leuven, Belgium, where he is currently a Senior Fellow. He is also a part-time Full Professor with the EE Department, KU Leuven. He has been an associate editor for several IEEE and ACM journals and was elected an IEEE fellow in 2005.