

Mixed-Precision Training and Compilation for RRAM-based Computing-in-Memory Accelerators

Rebecca Pelke^{1*}, Joel Klein^{1*},

José Cubero-Cascante*, Nils Bosbach*, Jan Moritz Joseph^{*†}, Rainer Leupers*

*RWTH Aachen University, Germany

[†]RooflineAI GmbH, Germany

*{pelke, klein, cubero, bosbach, joseph, leupers}@ice.rwth-aachen.de [†]joseph@roofline.ai

Abstract—Computing-in-Memory (CIM) accelerators are a promising solution for accelerating Machine Learning (ML) workloads, as they perform Matrix-Vector Multiplications (MVMs) on crossbar arrays directly in memory. Although the bit widths of the crossbar inputs and cells are very limited, most CIM compilers do not support quantization below 8 bit. As a result, a single MVM requires many compute cycles, and weights cannot be efficiently stored in a single crossbar cell.

To address this problem, we propose a mixed-precision training and compilation framework for CIM architectures. The biggest challenge is the massive search space, that makes it difficult to find good quantization parameters. This is why we introduce a reinforcement learning-based strategy to find suitable quantization configurations that balance latency and accuracy. In the best case, our approach achieves up to a $2.48\times$ speedup over existing state-of-the-art solutions, with an accuracy loss of only 0.086% .

Index Terms—RRAM, CIM, Compiler, ML, MPQ

I. INTRODUCTION

To perform Machine Learning (ML) workloads efficiently, new hardware architectures are required. Promising candidates are Computing-in-Memory (CIM) accelerators, which execute Matrix-Vector Multiplications (MVMs) directly in memory. This addresses the von Neumann bottleneck by reducing data movements between memory and processing units. [1]

Resistive Random Access Memory (RRAM) is a well-known technology for CIM crossbars because it offers high device density, low power usage, fast switching speed, and compatibility with standard CMOS processes [2], [3].

A limitation of RRAM crossbars is the restricted bit precision of both, the input activations and the crossbar cells. The cell resolution is limited by device variability, nonlinearity, and drift [4], [5], [6]. This makes it difficult to reliably distinguish between many resistance states. Typical bit widths for RRAM cells are between 1 bit and 4 bits [7], [8], [9], [10]. Input resolution is limited by the area, power consumption, and speed of the Digital-to-Analog Converter (DAC) used to convert the digital inputs to analog voltages [11], [12]. Higher-resolution DACs demand higher-resolution Analog-to-Digital Converters (ADCs), which increases power consumption significantly [8]. As a result, most crossbars use 1 bit DACs to reduce circuit complexity and improve robustness and efficiency [13], [14].

Existing compilers for CIM only support fixed bit width quantization with a minimum of 8 bit for activations and



Fig. 1: Overview of the proposed framework. The main contributions are highlighted in red.

weights [15], [16], [17], [18], [19], [20]. To map such 8 bit models onto low-resolution crossbars, two techniques are used: *weight bit slicing* and *input bit slicing* [21]. Weight bit slicing splits high-bit weights across multiple RRAM cells. Input bit slicing splits high-bit inputs across multiple computing cycles. This increases the inference latency drastically [21].

Mixed-precision Quantization-Aware Training (QAT) with low bit widths is a hardware-friendly alternative to fixed-precision quantization [22]. It combines QAT [23], which takes simulation effects into account during training, and Mixed-Precision Quantization (MPQ), where different layers use different bit widths for weights and inputs. To make use of this quantization scheme on CIM targets, we present our framework shown in Figure 1. It has two main contributions:

1. A MPQ-aware compiler that compiles mixed-precision models from the QAT framework Brevitas [24] for CIM architectures. The supported architectures are described in Sections II and III-A. It can handle crossbars of any size and applies CIM-specific optimizations. The compiler is described in Section IV-B. Results show speedups ranging from $2.20\times$ to $2.48\times$ over 8 bit quantization compilers for ResNet-18, ViT-B/32, and VGG-16 on the ImageNet dataset.

2. A reinforcement learning-guided MPQ optimizer for CIM targets. The exponential search space for MPQ makes manual tuning impractical. Existing automated approaches [25], [26] target only conventional hardware, while open-source frameworks for CIM are missing. We close this gap with CIM-aware Automated Quantization (CIM-AQ), our CIM-aware reinforcement learning-based optimizer. Therefore, we extend the Hardware-aware Automated Quantization (HAQ) framework [25] to enable mixed-precision optimization for CIM. By integrating Brevitas [24] as a new QAT backend, CIM-AQ does not only support a broader range of Neural Networks (NNs), including transformers, but is also approximately 8% faster. CIM-AQ can be found on GitHub².

¹Both authors contributed equally to this work.

²GitHub link: <https://github.com/jmkle/cim-aq>

TABLE I: Comparison with existing CIM compilers.

Compiler	Cell resolution	Crossbar size	Data types	MPQ support
TC-CIM [15]	4 bit	256 × 256	8 bit	✗
TDO-CIM [16]	4 bit	256 × 256	8 bit	✗
OCC [17]	4 bit	variable	8 bit	✗
Jin et al. [18]	2 bit	variable	16 bit	✗
CINM [19]	4 bit	variable	8 bit	✗
CIM-MLC [20]	any	variable	8 bit	✗
Ours	any	variable	1 bit-8 bit	✓

II. RELATED WORK

Many CIM compilers have been proposed in the past [15], [16], [17], [18], [19], [20], [27], [28]. They differ in the targeted CIM architecture, crossbar design, compiler framework, implemented optimizations, and overall flexibility. Table I shows compilers that assume a CIM architecture similar to ours. Figure 2 illustrates this architecture, consisting of a host CPU and a single memory-mapped CIM accelerator. Besides the crossbar, the CIM core also includes registers and control logic, which are omitted here for clarity.

TC-CIM [15] uses Tensor Comprehensions [29] and Loop Tactics [30] to detect and offload suitable tensor operations like MVMs and General Matrix Multiplications (GeMMs) to a CIM accelerator. TDO-CIM [16] builds on TC-CIM by detecting patterns at the LLVM-IR level for broader language support and uses Polly [31] and Loop Tactics to offload individual layers to a CIM accelerator. OCC [17] uses MLIR [32] to offload GeMMs and convolutions to a CIM accelerator. They improve endurance through reduced writes and better weight reuse. Jin et al. [18] developed a general-purpose LLVM-based compiler that identifies MVMs, GeMMs, and logic operations, with a runtime application that decides between CPU and CIM execution. CINM [19] offers an end-to-end compiler for heterogeneous CIM systems, using hierarchical MLIR abstractions for progressive lowering and optimizations. CIM-MLC [20] proposes a multi-level compilation stack with progressive lowering and scheduling optimizations tailored to different CIM architecture levels.

As summarized in Table I, current compilers are restricted to a fixed bit width of 8 bit or 16 bit. Since most crossbars have only 2 bit to 4 bit resolution and 1 bit DACs, the overhead of each MVM becomes large because multiple cycles per MVM and many cells per weight are required. As a result, latency increases and efficiency drops. To address these issues, our compiler uses the best bit width for each layer. This reduces cycles and write operations and thereby improves latency.

III. BACKGROUND

This section provides background on RRAM-based CIM architectures, QAT, and the used HAQ framework.

A. Analog MVMs on RRAM Crossbars

RRAM crossbars are used to perform MVMs directly in memory. Figure 2 illustrates the basic architecture of the CIM target used in this work. The topology of the $M \times N$ 1 Transistor 1 Resistor (1T1R) crossbar has a high resilience against wire parasitics [33], [34]. Each column consists of $2N$

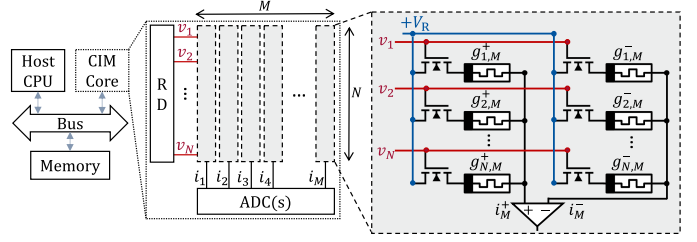


Fig. 2: The CIM architecture used in this work.

cells. Each pair of cells represents a single weight; $g_{j,M}^+$ is the positive and $g_{j,M}^-$ the negative part of the weight. This is called *differential mapping* [21], [35]. The dot product result of a crossbar column i_k , with $k \in [1, M]$, can be written as:

$$i_k = i_k^+ - i_k^- = \sum_{j=1}^N v_j \cdot (g_{j,k}^+ - g_{j,k}^-) \quad (1)$$

The input v_j is used to activate or deactivate row j . The input is binary and requires one cycle per bit. Each pair of cells can store a multi-bit weight. Weight bit slicing is used if the weight has more bits than the cell can store.

B. Quantization-Aware Training

Quantization in ML usually means mapping floating point ranges to integer values, e.g., fp32 ranges to int8 values. In *range-based linear quantization*, this mapping is described by a *scaling factor* s and an offset z called *zero-point*.

In our setup, weights use *symmetric signed quantization*, and activations use symmetric signed or unsigned quantization. In symmetric quantization, the zero-point is set to zero. The symmetric quantization of a floating point value x_f to an integer $x_q \in [-(2^{B-1} - 1), 2^{B-1} - 1]$ is defined as:

$$x_q = \left\lfloor x_f \cdot \frac{2^{B-1} - 1}{\max(|x_f|)} \right\rfloor \quad (2)$$

For small bit widths B , the quantization error can be significant. To improve the accuracy, QAT [23] is commonly used. An important concept in QAT is *fake quantization*. Fake quantization applies rounding and clipping to weights and activations in the forward pass but lets gradients pass through unchanged, so the NN learns the quantization effects.

A well-known framework for QAT is Brevitas [24]. Brevitas is based on PyTorch and also supports MPQ.

C. Reinforcement Learning

Reinforcement learning [36] is a concept where an *agent* learns by interacting with an *environment* over time through *actions* and *rewards*. Figure 3 illustrates the agent-environment interaction. At each time step t , the agent chooses an action a_t according to a *policy* π , with $a_t = \pi(s_t)$ for deterministic policies [37] and $\pi(a_t|s_t)$ for stochastic policies [38]. The environment responds with a new state s_{t+1} and a reward r_t .

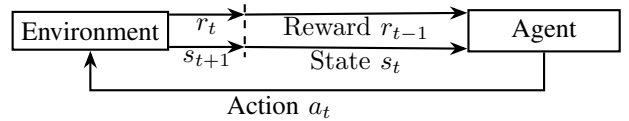


Fig. 3: Agent-environment interaction.

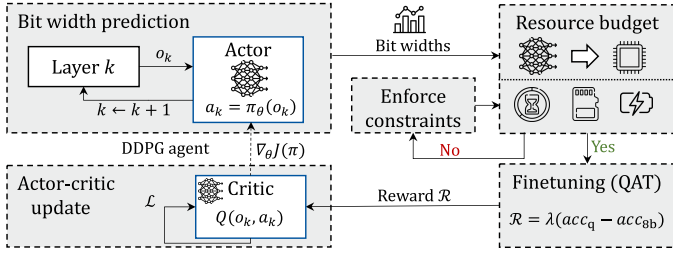


Fig. 4: Learning flow in the HAQ framework.

The goal in deterministic reinforcement learning is to find a policy π^* that maximizes the expected cumulative reward:

$$\pi^*(s) = \arg \max_{\pi} \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s \right], \quad (3)$$

where $\gamma \in [0, 1)$ is a discount factor that prioritizes rewards.

D. Hardware-aware Automated Quantization (HAQ)

The HAQ framework uses a Deep Deterministic Policy Gradient (DDPG) [39] agent. DDPG is an *off-policy* algorithm, which means that the agent can learn from data generated by a different policy than the one it is currently optimizing. It is based on the actor-critic architecture [40], where the *actor* learns the policy π and the *critic* learns the action-value function $Q(s, a)$. In HAQ, a state is denoted as o_k and contains the layer information of layer k . The policy and the action-value function are learned by NNs. Figure 4 breaks down the main steps of the learning flow of the HAQ framework:

- 1) In each *episode*, the actor determines the quantization parameters $a_k = \pi_{\theta}(o_k)$ for each layer separately.
- 2) It is checked if the quantized NN fits into the resource budget for latency, memory, and power consumption.
- 3) If the resource budget is exceeded, the bit widths are decreased sequentially to enforce the constraints.
- 4) The NN is trained for only one epoch. The reward \mathcal{R} is calculated based on the top-1 accuracy of this epoch.
- 5) All tuples $T_k = (o_k, a_k, \mathcal{R}, o_{k+1})$ are stored in the *replay buffer*, which collects previous experiences.
- 6) The critic network is trained by minimizing a loss based on a variant of the Bellman equation [41]. Therefore, a batch of random samples from the input buffer is used. The actor is updated using gradients from the critic.

After training, a full QAT must be performed using the learned parameters. We will use Brevitas for this step, as its trained NN can be directly imported into the compiler.

IV. OUR APPROACH

Section IV-A describes the CIM-AQ framework, which adapts the HAQ framework to CIM targets and accelerates it. Section IV-B introduces the MPQ-capable compiler.

A. CIM-aware Automated Quantization (CIM-AQ)

In HAQ, the agent searches for the MPQ policy that maximizes accuracy while staying within hardware's resource budgets (memory footprint, latency, power consumption).

TABLE II: Hardware parameters of the CIM target.

Parameter	Description
$M \times N$	Crossbar dimension, as shown in Figure 2
r_{cell}	Precision of a single RRAM cell (in bit)
r_{DAC}	Precision of the DAC (in bit)
t_{write}	Programming time of the crossbar (in μs)
t_{mvm}	Execution time of an MVM (in μs)

In CIM, however, we need to solve the following problem: Find MPQ parameters that **minimize latency** and **preserve high accuracy** for a given CIM target. This leads to the following adjustments of the HAQ framework:

Reward function: The target accuracy acc_t that should at least be achieved by the MPQ is defined as

$$acc_t = acc_{8b} - acc_{\text{loss}} \quad (4)$$

while acc_{8b} is the accuracy of the reference 8 bit NN and acc_{loss} is the maximum tolerated accuracy loss. Unlike the fixed hardware-budget constraints in HAQ, our target accuracy acc_t cannot be enforced as a strict constraint, since the post-quantization accuracy, acc_q , of the MPQ-quantized NN is unknown at the time of constraint evaluation (see Figure 4). Instead, we incorporate acc_t into the reward function to penalize any accuracy loss below the target, thereby enabling a joint optimization of latency and accuracy.

Our reward function \mathcal{R} is defined as:

$$\mathcal{R} = \begin{cases} -\alpha(acc_t - acc_q), & acc_q < acc_t \\ \beta\left(\frac{T_{8b}}{T_q} - 1\right) + \gamma(acc_q - acc_t), & \text{otherwise} \end{cases} \quad (5)$$

The inference latencies of the 8 bit and MPQ quantized NNs are denoted as T_{8b} and T_q , respectively. The hyperparameter α scales the penalty when accuracy falls below the target, whereas β and γ scale the trade-off between speedup and accuracy. They are set to $\alpha = 10$, $\beta = 100$, and $\gamma = 0.1$.

Latency cost modeling: The inference latency T of the NN is determined based on the hardware parameters listed in Table II. These parameters belong to the architecture discussed in Section III-A with a single CIM core. The total latency T can be calculated as follows:

$$T = \sum_{\text{layer } l} r_{\text{repeat},l} (N_{\text{write},l} \cdot t_{\text{write}} + N_{\text{mvm},l} \cdot t_{\text{mvm}}) \quad (6)$$

The repeat factor $r_{\text{repeat},l}$ denotes the repetition factor for operations executed multiple times, such as attention head computations in Multi-Head Attention (MHA) layers.

The number of write operations $N_{\text{write},l}$ is defined as:

$$N_{\text{write},l} = \left\lceil \frac{2 \cdot M_l}{M} \cdot \left\lceil \frac{w_{\text{bit},l}}{r_{\text{cell}}} \right\rceil \right\rceil \cdot \left\lceil \frac{N_l}{N} \right\rceil \quad (7)$$

Here, $w_{\text{bit},l}$ denotes the weight bit-width in layer l . M_l and N_l are the row and column dimensions of the layer's weight matrix. Layers without a 2D weight matrix, e.g., Conv2D, must be transformed into a GeMM operation first. This is usually done with the *im2col* transformation [42], [43], [44]. For instance, a Conv2D layer mapping $(C_{\text{in}}, H_{\text{in}}, W_{\text{in}}) \rightarrow (C_{\text{out}}, H_{\text{out}}, W_{\text{out}})$ with kernel size (K_H, K_W) can be expressed as a (M_l, V_l, N_l) -GeMM operation with dimensions:

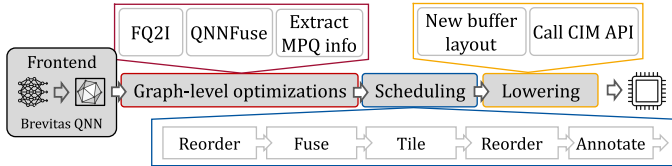


Fig. 5: Overview of the compilation pipeline.

$$M_l = C_{out}, \quad N_l = C_{in} K_H K_W, \quad V_l = H_{out} W_{out} \quad (8)$$

The number of MVM operations $N_{mvm,l}$ is defined as:

$$N_{mvm,l} = V_l \cdot N_{write,l} \cdot \left\lceil \frac{a_{bit,l}}{r_{DAC}} \right\rceil \quad (9)$$

The activation bit width is $a_{bit,l}$. Equation (8) specifies V_l for Conv2D layers. For normal Dense layers, we can simply set $V_l = 1$. For Dense and GeMM layers inside MHA blocks, each token needs to be processed independently, so we set V_l to the sequence length N . We also generate a lookup table for the cost model. It stores the latency for each layer and every combination of weight and activation bit widths in a range from b_{min} to b_{max} , the minimum and maximum bit widths.

HAQ acceleration: In the HAQ framework, one epoch of QAT is performed during every finetuning phase (see Figure 4). One problem is that HAQ relies on a custom quantization implementation, which only supports very few layers. For example, typical layers of transformer architectures are not supported. Another problem is that the time of the whole HAQ framework is dominated by the QAT epoch. To overcome these limitations, we integrate Brevitas as a backend for QAT in the HAQ framework. Brevitas not only supports a wide range of layers, but also speeds up the whole reinforcement learning process.

B. CIM Compiler

Our compiler maps MPQ models trained with Brevitas to the CIM targets described in Section III-A. CIM-specific Application Programming Interface (API) calls are inserted into the ML model. A runtime environment implements those calls to offload MVM executions to the CIM accelerator. The compiler is built on the Tensor Virtual Machine (TVM) [45] framework to reuse existing infrastructure. Figure 5 shows an overview of our compilation pipeline. In the following, we will explain the main steps of the compilation pipeline.

Frontend: After QAT with Brevitas, the MPQ model is converted to the Open Neural Network Exchange (ONNX) format. TVM offers a built-in ONNX frontend. The ONNX graph is then converted into an internal representation called *Relay*. Quantization is encoded in the Quantize and DeQuantize (QDQ) style: `QuantizeLinear` followed by `DeQuantizeLinear` nodes are inserted between the original operators. These nodes allow switching between integer and floating-point domains. At this point, operations like Conv2D are still in the floating-point format.

To enable MPQ in our compiler, we added a `config_update` attribute to Relay operators, and implemented a Relay transformation pass that detects

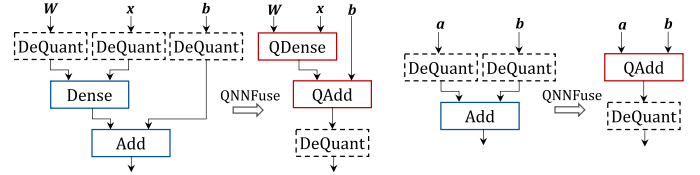


Fig. 6: Optimizations of the QNNFuse pass.

quantized layers, infers activation and weight bit-widths, and determines the bit-splitting scheme for the crossbar. The resulting configuration is stored as JSON in the operator attributes and later used during lowering to generate accelerator function calls.

Graph-level optimizations: We apply a series of high-level transformations to the Relay graph. For execution on the CIM target, all QDQ patterns must be converted to integer-only computations. TVM’s *FQ2I* pass replaces each QDQ pair with a quantized operator and inserts `ReQuantize` nodes to align scales and zero points. However, this pass can leave the final layer or the bias addition partly in floating-point format. To ensure that the last block is also quantized, we implement a custom *QNNFuse* pass. It merges the leaf operations to corresponding quantized operations. Figure 6 illustrates two examples of graph transformations performed by this pass.

Finally, we partition the optimized graph, assigning each operation either to the CPU or to the CIM accelerator. CPU-only parts are lowered through the standard TVM pipeline, whereas CIM operations follow the custom scheduling and lowering flow shown in Figure 5.

Scheduling: Scheduling in TVM controls how an operation’s loop nests are transformed to match a given hardware target. Primitives such as `split`, `reorder`, `tile`, and `fuse` (see Figure 5) help to modify loop nests to improve parallelism and data locality. Since write operations to the crossbar are more costly than MVMs, we adopt a weight-stationary dataflow that keeps weights in the crossbar as long as possible. To target CIM, loops are *annotated* with a predefined set of axis labels (e.g., “outer MVM row”). The idea is to only label the relevant loops in the schedule, so our subsequent lowering passes pick up those labels to insert the proper CIM API calls fully automatically. With this approach, supporting new layers only needs 5-10 lines of code.

Lowering: We provide two custom lowering passes. First, we insert staging buffers of size $M \times N$, $1 \times N$, and $1 \times M$ to hold the weights, inputs, and outputs, respectively. This aligns the memory layout with the required format of the CIM API. Then, we inject API calls that replace the isolated computation in the inner loop nest. Pointers to the buffers are passed as arguments.

V. RESULTS

In this section, we evaluate the performance of our framework. In summary, the workflow consists of three steps: First, reinforcement learning is employed to determine suitable quantization parameters. Second, our compiler maps the optimized models to the CIM target introduced in Section III-A. Finally, the compiled model is executed on an open-source

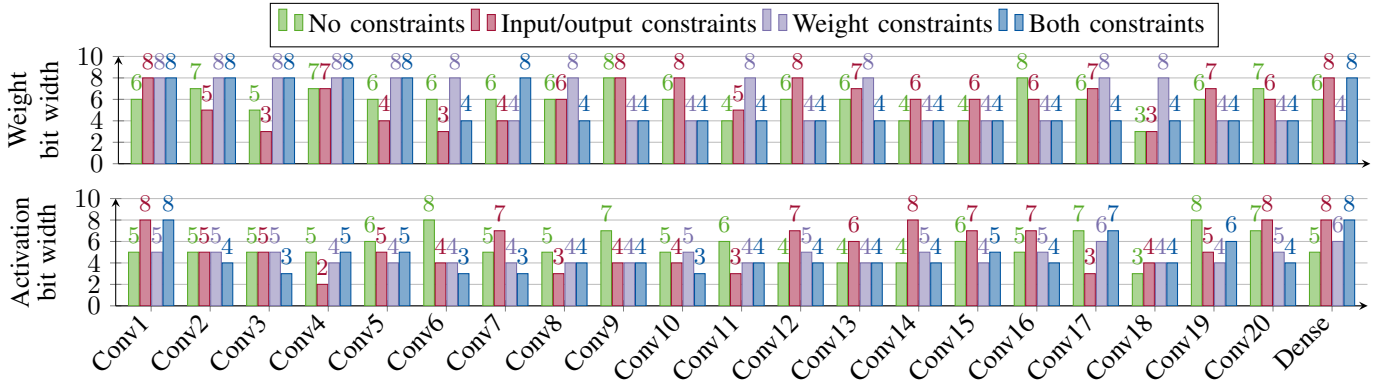


Fig. 7: Final activation and weight bit widths of ResNet-18 layers under different constraints for $r_{\text{cell}} = 4$ bit.

crossbar simulator [35]. The inference latency is determined by using crossbar parameters from [8]: The crossbar size is 256×256 with differential cell mapping and $r_{\text{DAC}} = 1$. The latencies are $t_{\text{write}} = 56 \mu\text{s}$ and $t_{\text{mvm}} = 1.4 \mu\text{s}$ (see Table II). These latencies also include the DAC and ADC conversions. We optimize three representative NNs trained on ImageNet [46]: ResNet-18 [47], VGG-16 [48], and the vision transformer ViT-B/32 [49]. Table III lists the number of layers that are mapped to the crossbar for each benchmark.

A. CIM-AQ - Performance Improvements

To improve the HAQ framework, we replaced the original quantization backend with Brevitas. As mentioned in Section IV, this not only enables the use of a wider range of NN architectures like transformers, but also provides modest runtime improvements. For example, on our NVIDIA L40S GPU (CUDA 12.8, 48 GB memory), we observed a speedup of 8.4% when running reinforcement learning for VGG-16 compared to the original HAQ implementation.

B. CIM-AQ - Evaluation of Constraints

Finding optimal MPQ parameters using CIM-AQ remains time-consuming. This is why we first explore how different constraint strategies affect the quality of the MPQ parameters. By restricting the search space, constraints can help achieve better results. We evaluate the following constraints:

- **Input/output** constraint: The activations and weights in the first and last layer remain at 8 bit, since these layers are typically more sensitive to quantization noise [50].
- **Weight** constraint: The bit widths of the weights are restricted to multiples of the cell resolution r_{cell} .
- **Both** constraints: Both previous constraints.

In all subsequent experiments, CIM-AQ is run for 600 episodes, with 3 QAT epochs per episode (see Section III-D), on a reduced ImageNet100 dataset with 20,000 training and 10,000 validation images. The accuracy threshold acc_{loss} is set to 5% (see Section IV). Once the optimal MPQ parameters

TABLE III: An overview of the benchmarks' layers.

Benchmark	Conv2D	Dense	MatMul
ResNet-18	20	1	-
VGG-16	13	3	-
ViT-B/32	1	49	24

are determined, the NN is further trained for 30 epochs on the full ImageNet dataset. All reported accuracies correspond to the validation accuracy of this final training step.

Figure 7 shows the final MPQ bit widths of the ResNet-18 layers for weights and activations. The cell resolution is set to $r_{\text{cell}} = 4$ bit, meaning that weights exceeding 4 bit are distributed across multiple cells. Although activations are generally considered more sensitive to quantization when using conventional hardware [51], the results indicate that, for CIM architectures, activations are in most cases quantized more aggressively than weights. The reward function (see Equation (5)) explains this behavior because it explicitly rewards low latency. The bit width of the activations has a direct impact on latency: each saved activation bit reduces the MVM latency by one cycle. In contrast, weights affect latency more indirectly: larger bit widths increase the number of weight bit slices, which may no longer fit on the crossbar and can lead to additional crossbar writes and compute cycles.

Another observation is that the activations of the first half of the Conv2D layers are quantized more aggressively than those of the second half. These layers use fewer weights and thus require fewer crossbar write operations. At the same time, they perform more MVMs than later Conv2D layers. Reducing activation bit width therefore speeds up the front layers more.

To compare the constraint strategies, we analyze their effect on the speedup and accuracy loss trade-off. Both values are given relative to the 8 bit baseline used in other CIM compilers. Figure 8 shows the results for ResNet-18. The top row uses

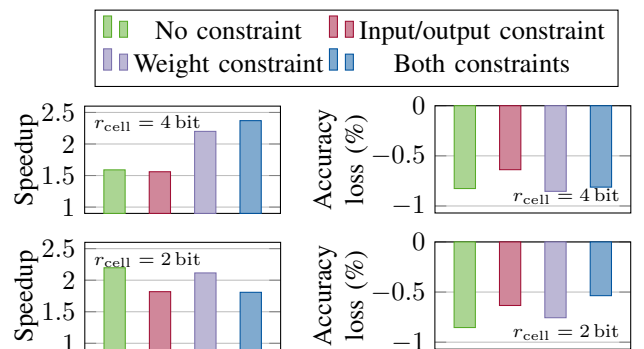


Fig. 8: Speedup and top-1 accuracy loss for ResNet-18 in comparison to the 8 bit baseline for different cell resolutions.

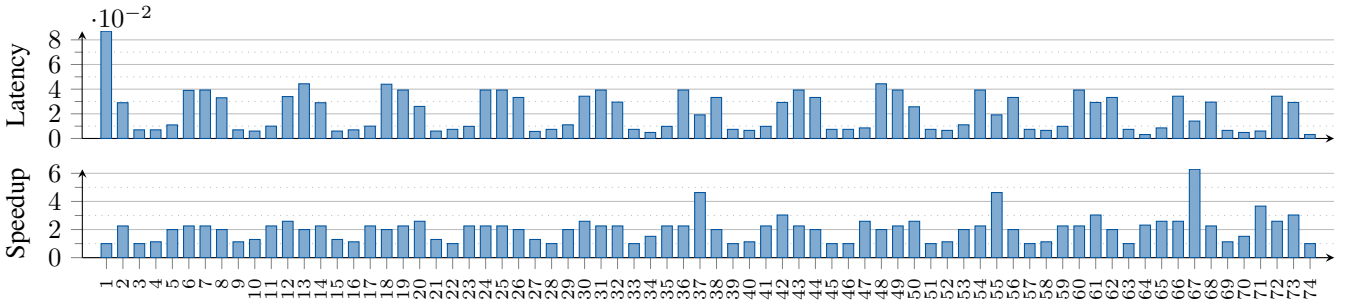


Fig. 9: Latency (in s) for MPQ model and speedup against the 8 bit baseline for each layer of ViT-B/32 with $r_{\text{cell}} = 4$ bit.

4 bit cells, the bottom row 2 bit cells. Accuracy loss is similar for both resolutions, but speedup is slightly higher for 2 bit cells in all cases except when both constraints are applied. This is plausible because the latency also depends on the bit width of the weights: for $r_{\text{cell}} = 4$ bit, only a weight bit reduction from 8 bit to 4 bit improves the latency, whereas for $r_{\text{cell}} = 2$ bit, each reduction of 2 bit increases the speedup.

For 4 bit cells, the highest speedups occur with weight constraints or both constraints, reaching over $2.20\times$. For 2 bit cells, the best speedups are achieved without constraints or with weight constraints. Accuracy loss, however, is lower with input/output or both constraints, in the best case only -0.536% . To capture this trade-off, we define the S/AL score as the ratio of speedup to absolute accuracy loss:

$$S/AL := \frac{\text{Speedup compared to 8 bit NN}}{|\text{Accuracy loss compared to 8 bit NN}|} \quad (10)$$

Table IV lists the S/AL scores for ResNet-18 with 2 bit and 4 bit cells. Without constraints, the score is lowest. Applying both constraints yields the highest score. This confirms that constraints help to reduce the search space and lead to finding better quantization parameters.

TABLE IV: ResNet18: S/AL score in $\frac{1}{\%}$ (higher is better).

	No constraint	Input/Output	Weight	Both
2 bit	2.573 (worst)	2.867	2.798	3.373 (best)
4 bit	1.920 (worst)	2.884	2.575	2.924 (best)

C. Comparison Against Related Work

Finally, we evaluate our approach against state-of-the-art 8 bit solutions discussed in Section II. Since most of the 8 bit compilers operate at a 4 bit cell resolution, our comparison focuses on this configuration. Moreover, we use both constraints in the following because this strategy delivered the best results in the previous experiments. Direct, one-to-one comparisons are not possible because many compilers are either closed source or incapable of executing complete NNs. To address this, we use 8 bit workloads on our own compiler as a proxy reference while adopting the key compiler optimizations described in [17]. Specifically, we assume maximal weight reuse, which is considered the most effective strategy for minimizing costly crossbar rewrites (see Section IV-B).

First, we analyze the results presented in Figure 9, which illustrate the execution latency and speedup against the 8 bit

baseline for each individual layer of the ViT-B/32 model. The speedup shows that the latency is consistently reduced across most layers. Due to the applied constraints, the first layer is not quantized beyond 8 bit. Here, Dense layers have higher latency than MatMul layers, so their weights are reduced to 4 bit, while most MatMul layers remain at 8 bit. Dense layers performing dimensionality reduction in the Multi-Layer Perceptron (MLP) block dominate latency due to their large number of weights. In the second half of the model, their inputs can be reduced to 3 bit, explaining the high speedups in layers 37, 55, and 67.

For all benchmarks, our MPQ models deliver substantial speedups over the 8 bit baseline. The results are summarized in Table V. For VGG16, our approach yields the highest speedup of $2.48\times$, with an accuracy loss of only 0.086% . Across all NNs, we consistently achieve speedups of at least $2.20\times$, while the maximum accuracy degradation does not exceed 2.140% .

TABLE V: Results: Comparison against 8 bit execution.

	Speedup	Accuracy loss (%)	S/AL score ($\frac{1}{\%}$)
ResNet-18	2.37	-0.812	2.924
VGG16	2.48	-0.086	28.862
ViT-B/32	2.20	-2.140	1.028

VI. CONCLUSION

In this work, we introduced a MPQ training and compilation framework for RRAM-based CIM accelerators. We first extended and improved the HAQ framework to create CIM-AQ, a reinforcement learning-based optimizer that explores the large MPQ search space for CIM architectures. CIM-AQ balances latency and accuracy. We also introduced constraint strategies to further improve search efficiency. Our experiments with ResNet-18, VGG-16, and ViT-B/32 on ImageNet demonstrate that, in the best case, the proposed approach achieves up to a $2.48\times$ speedup over existing 8 bit compilers, with an accuracy loss of only 0.086% . These results confirm that our MPQ framework can significantly reduce inference latency on CIM while preserving high accuracy.

At this stage, the observed accuracy losses stem solely from quantization effects. Future work will extend the framework to also include crossbar non-idealities in the training process.

ACKNOWLEDGMENTS

This work was funded by the Federal Ministry of Research, Technology and Space in the projects NeuroSys II (03ZU2106CA), NEUROTEC-II (16ME0399), and EXIST (03EFWNW338).

REFERENCES

- [1] H. Amrouch, N. Du, A. Gebregiorgis, S. Hamdioui, and I. Polian, "Towards Reliable In-Memory Computing: From Emerging Devices to Post-von-Neumann Architectures," in *Proc. 29th IFIP/IEEE Int. Conf. VLSI-SoC*, 2021, pp. 1–6.
- [2] F. Zahoor, T. Z. Azni Zulkifli, and F. A. Khanday, "Resistive Random Access Memory (RRAM): an Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications," *Nanoscale Res. Lett.*, vol. 15, no. 1, p. 90, 2020.
- [3] P.-C. e. a. Wu, "A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices," in *Proc. IEEE Int. Conf. ISSCC*, vol. 65, 2022, pp. 1–3.
- [4] A. Grossi *et al.*, "Impact of Intercell and Intracell Variability on Forming and Switching Parameters in RRAM Arrays," *IEEE Trans. Electron Devices*, vol. 62, no. 8, pp. 2502–2509, 2015.
- [5] A. Ding, Y. Qiao, and N. Bagherzadeh, "BNN An Ideal Architecture for Acceleration With Resistive In Memory Computation," *IEEE Trans. Emerg. Topics Comput.*, vol. 11, no. 2, pp. 281–291, 2023.
- [6] R. Pelke *et al.*, "The show must go on: a reliability assessment platform for resistive random access memory crossbars," *Philos. Trans. A*, vol. 383, no. 2288, p. 20230387, 2025.
- [7] C.-X. Xue *et al.*, "24.1 A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," in *Proc. IEEE Int. Conf. ISSCC*. IEEE, 2019, pp. 388–390.
- [8] W. Wan *et al.*, "A Compute-in-Memory Chip Based on Resistive Random-Access Memory," *Nature*, vol. 608, no. 7923, 2022.
- [9] J. Read *et al.*, "Neurosim v1. 5: Improved software backbone for benchmarking compute-in-memory accelerators with device and circuit-level non-idealities," *arXiv preprint arXiv:2505.02314*, 2025.
- [10] R. Pelke *et al.*, "A Fully Automated Platform for Evaluating ReRAM Crossbars," in *Proc. 25th IEEE LATS*, 2024, pp. 1–6.
- [11] S.-T. Wei *et al.*, "Trends and challenges in the circuit and macro of RRAM-based computing-in-memory systems," *Chip*, vol. 1, no. 1, p. 100004, 2022.
- [12] L. Ni *et al.*, "An energy-efficient matrix multiplication accelerator by distributed in-memory computing on binary RRAM crossbar," in *21st ASP-DAC*, 2016, pp. 280–285.
- [13] T. Chou, W. Tang, J. Botimer, and Z. Zhang, "CASCADE: Connecting RRAMs to Extend Analog Dataflow In An End-To-End In-Memory Processing Paradigm," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. MICRO*, 2019, pp. 114–125.
- [14] C. Bengel, L. Dixius, R. Waser, D. J. Wouters, and S. Menzel, "Bit slicing approaches for variability aware ReRAM CIM macros," *it-Information Technology*, vol. 65, no. 1-2, pp. 3–12, 2023.
- [15] A. Drebes *et al.*, "TC-CIM: Empowering Tensor Comprehensions for Computing-In-Memory," in *Proc. 10th IMPACT*, 2020.
- [16] K. Vadivel *et al.*, "TDO-CIM: Transparent Detection and Offloading for Computation In-memory," in *Proc. DATE*, 2020, pp. 1602–1605.
- [17] A. Siemieniuk *et al.*, "OCC: An Automated End-to-End Machine Learning Optimizing Compiler for Computing-In-Memory," *IEEE TCAD*, vol. 41, no. 6, pp. 1674–1686, 2022.
- [18] H. Jin *et al.*, "A Compilation Tool for Computation Offloading in ReRAM-based CIM Architectures," *ACM Transactions on Architecture and Code Optimization*, vol. 20, no. 4, pp. 1–25, 2023.
- [19] A. A. Khan *et al.*, "CINM (Cinnamon): A Compilation Infrastructure for Heterogeneous Compute In-Memory and Compute Near-Memory Paradigms," in *Proc. 29th ACM ASPLOS*, 2024, pp. 31–46.
- [20] S. Qu *et al.*, "CIM-MLC: A Multi-Level Compilation Stack for Computing-In-Memory Accelerators," in *Proc. 29th ACM ASPLOS*, 2024, pp. 185–200.
- [21] T. P. Xiao *et al.*, "On the Accuracy of Analog Neural Network Inference Accelerators," *IEEE Circuits Syst. Mag.*, vol. 22, no. 4, pp. 26–48, 2022.
- [22] S. Huang *et al.*, "Mixed precision quantization for ReRAM-based DNN inference accelerators," in *26th ASP-DAC*, 2021, p. 372–377.
- [23] B. Jacob *et al.*, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," in *IEEE CVPR*, 2018, pp. 2704–2713.
- [24] A. Pappalardo *et al.*, "Xilinx/brevitas: Release v0.12.0," May 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.15375017>
- [25] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "HAQ: Hardware-Aware Automated Quantization With Mixed Precision," in *Proc. CVPR*, 2019, pp. 8612–8620.
- [26] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "HAWQ: Hessian AWare Quantization of Neural Networks With Mixed-Precision," in *Proc. IEEE/CVF ICCV*, 2019, pp. 293–302.
- [27] J. Ambrosi *et al.*, "Hardware-Software Co-Design for an Analog-Digital Accelerator for Machine Learning," in *Proc. IEEE ICRC*, 2018, pp. 1–13.
- [28] J. Han, X. Fei, Z. Li, and Y. Zhang, "Polyhedral-Based Compilation Framework for In-Memory Neural Network Accelerators," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 18, no. 1, pp. 1–23, 2022.
- [29] N. e. a. Vasilache, "The Next 700 Accelerated Layers: From Mathematical Expressions of Network Computation Graphs to Accelerated GPU Kernels, Automatically," *ACM Trans. Archit. Code Optim.*, vol. 16, no. 4, pp. 1–26, 2019.
- [30] L. Chelini, O. Zinenko, T. Grosser, and H. Corporaal, "Declarative loop tactics for domain-specific optimization," *ACM Trans. Archit. Code Optim.*, vol. 16, no. 4, pp. 1–25, 2019.
- [31] T. Grosser, A. Groesslinger, and C. Lengauer, "Polly—Performing Polyhedral Optimizations on a Low-Level Intermediate Representation," *Parallel Process. Lett.*, vol. 22, no. 04, p. 1250010, 2012.
- [32] C. Lattner *et al.*, "MLIR: Scaling Compiler Infrastructure for Domain Specific Computation," in *Proc. 19th IEEE/ACM CGO*. IEEE, 2021, pp. 2–14.
- [33] T. P. Xiao *et al.*, "Analysis and mitigation of parasitic resistance effects for analog in-memory neural network acceleration," *Semicond. Sci. Technol.*, vol. 36, no. 11, p. 114004, 2021.
- [34] J. Cubero-Cascante *et al.*, "Evaluating the Scalability of Binary and Ternary CNN Workloads on RRAM-based Compute-in-Memory Accelerators," *arXiv preprint arXiv:2505.07490*, 2025.
- [35] R. Pelke *et al.*, "Optimizing Binary and Ternary Neural Network Inference on RRAM Crossbars using CIM-Explorer," 2025. [Online]. Available: <https://arxiv.org/abs/2505.14303>
- [36] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.
- [37] D. Silver *et al.*, "Deterministic Policy Gradient Algorithms," in *Proc. ICML*, 2014, pp. 387–395.
- [38] P.-W. Chou, D. Maturana, and S. Scherer, "Improving Stochastic Policy Gradients in Continuous Control with Deep Reinforcement Learning using the Beta Distribution," in *Proc. ICML*, 2017, pp. 834–843.
- [39] H. Tan, "Reinforcement Learning with Deep Deterministic Policy Gradient," in *Proc. CAIBDA*. IEEE, 2021, pp. 82–85.
- [40] I. Grondman, L. Busoni, G. A. Lopes, and R. Babuska, "A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients," *IEEE Trans. Syst., Man, Cybern., Pt. C (Appl. Rev.)*, vol. 42, no. 6, pp. 1291–1307, 2012.
- [41] R. Bellman, "On the Theory of Dynamic Programming," *Proc. Natl. Acad. Sci. USA*, vol. 38, no. 8, pp. 716–719, 1952.
- [42] Y. Zhang, G. He, G. Wang, and Y. Li, "Efficient and Robust RRAM-Based Convolutional Weight Mapping With Shifted and Duplicated Kernel," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 40, no. 2, pp. 287–300, 2020.
- [43] R. Pelke *et al.*, "Mapping of CNNs on multi-core RRAM-based CIM architectures," in *Proc. 31th IFIP/IEEE Int. Conf. VLSI-SoC*, 2023, pp. 1–6.
- [44] R. Pelke *et al.*, "CLSA-CIM: A Cross-Layer Scheduling Approach for Computing-in-Memory Architectures," in *Proc. DATE*, 2024, pp. 1–6.
- [45] T. Chen *et al.*, "TVM: An Automated End-to-End Optimizing Compiler for Deep Learning," in *Proc. 13th OSDI*, 2018, pp. 578–594.
- [46] J. Deng *et al.*, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE CVPR*, 2009, pp. 248–255.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.
- [48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.
- [49] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
- [50] J. Lee, M. Yu, Y. Kwon, and T. Kim, "Quantune: Post-training quantization of convolutional neural networks using extreme gradient boosting for fast deployment," *Future Gener. Comput. Syst.*, vol. 132, pp. 124–135, 2022.
- [51] S. Zhou *et al.*, "DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients," *arXiv preprint arXiv:1606.06160*, 2016.