

On the accuracy of analog neural network inference accelerators

T. Patrick Xiao[†], Ben Feinberg[†], Christopher H. Bennett[†], Venkatraman Prabhakar*, Prashant Saxena*, Vineet Agrawal*, Sapan Agarwal[†], Matthew J. Marinella^{†,‡}

[†]Sandia National Laboratories, *Infineon Technologies, [‡]Arizona State University
[†]{txiao, bfeinbe, cbennet, sagarwa}@sandia.gov
*{Venkatraman.Prabhakar, Prashant.Saxena, Vineet.Agrawal}@infineon.com
[‡]m@asu.edu

ABSTRACT

Specialized accelerators have recently garnered attention as a method to reduce the power consumption of neural network inference. A promising category of accelerators utilizes non-volatile memory arrays to both store weights and perform *in situ* analog computation inside the array. While prior work has explored the design space of analog accelerators to optimize performance and energy efficiency, there is seldom a rigorous evaluation of the accuracy of these accelerators. This work shows how architectural design decisions, particularly in mapping neural network parameters to analog memory cells, influence inference accuracy. When evaluated using ResNet50 on ImageNet, the resilience of the system to analog non-idealities—cell programming errors, analog-to-digital converter resolution, and array parasitic resistances—all improve when analog quantities in the hardware are made *proportional* to the weights in the network. Moreover, contrary to the assumptions of prior work, nearly equivalent resilience to cell imprecision can be achieved by fully storing weights as analog quantities, rather than spreading weight bits across multiple devices, often referred to as bit slicing. By exploiting proportionality, analog system designers have the freedom to match the precision of the hardware to the needs of the algorithm, rather than attempting to guarantee the same level of precision in the intermediate results as an equivalent digital accelerator. This ultimately results in an analog accelerator that is more accurate, more robust to analog errors, and more energy-efficient.

1. INTRODUCTION

Deep neural networks (DNNs) have grown rapidly in importance in the past decade, enabling image recognition, natural language processing, predictive analytics, and many other tasks to be performed with high accuracy and generalizability [40]. As the size and complexity of DNNs have grown to tackle more challenging problems, so has the demand for increasingly powerful and energy-efficient processors. Hardware that is optimized for DNN processing, which is dominated by matrix operations [15], has been a major enabler of machine learning innovation. But new, more efficient hardware approaches are needed to keep pace with the rapid developments in artificial intelligence and its growing computational needs [64].

Accelerators based on *in situ* computing—utilizing memory for both storage and computation—have attracted sig-

nificant attention as a possible path to order-of-magnitude improvements in energy efficiency [9, 11, 56]. These systems harness the analog properties of non-volatile memory arrays to perform many concurrent multiply-and-accumulate (MAC) operations, enabling the computation of a matrix-vector multiplication (MVM) in a single step.

While analog processing offers intrinsic efficiency benefits, it has historically struggled with accuracy. Unlike digital systems, the solution quality in analog systems is directly degraded by noise, process variations, and various parasitic effects. To provide precision on par with digital systems, many prior analog inference accelerators adopt a hybrid approach known as *bit slicing*, where weight values are spread bitwise across multiple memory devices, and the analog intermediate results are digitized and aggregated [9, 11, 56]. This technique allows weights to be represented more precisely even with low-precision memory devices, but at a higher energy cost than a purely analog approach. Recent work has optimized the performance and energy of bit-sliced accelerators [6, 14, 42, 49], but rarely evaluates the effect of system-level design decisions on inference accuracy.

This work studies how architecture affects accuracy in analog inference accelerators. We use a detailed accuracy model for *in situ* MVMs that includes the effect of various analog errors at the resolution of individual MACs, such as memory cell process variations and array parasitic resistances. The model allows an architectural design space exploration that uses the error sensitivity of end-to-end inference accuracy as the primary figure-of-merit. To provide a sensitivity analysis that can be applied to realistic applications, accuracy is evaluated on ImageNet classification with the ResNet50 neural network from the MLPerf Inference benchmark. This model is also used to benchmark digital systems [51].

Though the accuracy of analog accelerators has been studied [65], the analysis in this work provides a more comprehensive view of how accuracy fits into analog architecture design. This work demonstrates that bit slicing offers a smaller benefit than often assumed and typically does not justify its energy cost; moreover, contrary to the assumptions of prior work, bit slicing cannot be used as a mitigation for highly error-prone analog devices. Just as important, when signed arithmetic is handled in analog, it is possible to obtain a linear or *proportional mapping* between the numerical values in the algorithm and the physical quantities that represent them in the analog hardware. This proportionality is the key to enable

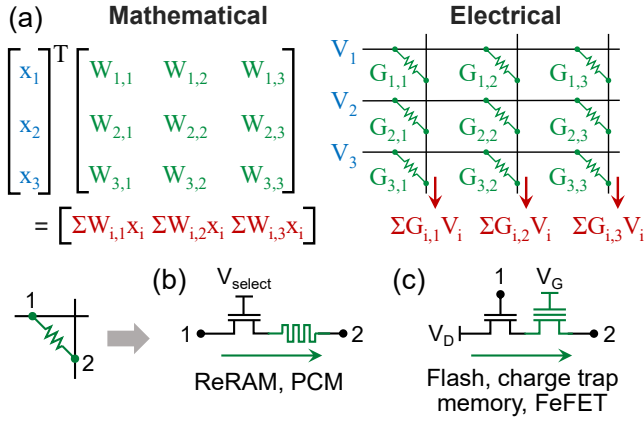


Figure 1: (a) Execution of an MVM $\vec{y} = \mathbf{W}\vec{x}$ within a memory array. (b)-(c) Two implementations of a memory cell that performs analog multiplication.

high inference accuracy and greater robustness to analog errors. Following the end-to-end principle of Saltzer *et al.* [53], this robustness allows hardware requirements (such as the array size and the analog-to-digital converter resolution) to be relaxed while still ensuring high end-to-end accuracy of an application neural network. For state-of-the-art inference applications, proportionality produces significant improvements simultaneously in accuracy, error tolerance, energy efficiency, and area.

This paper is organized as follows. Section 2 introduces analog inference accelerators, and surveys the architectural design space established by prior work in the field. Section 3 lays out what we view to be key principles for designing analog systems that can achieve high accuracy and resilience to errors. These conclusions are supported by the results in the remainder of the paper, which are based on the methodology in Section 4. Sections 5, 6, 7 and 8 discuss how device- and architecture-level design decisions influence an accelerator’s sensitivity to memory cell programming errors, ADC quantization errors, and parasitic resistance errors. Section 9 presents a case study of these principles using an exemplar analog core based on characterized charge trap memory arrays, and evaluate its accuracy, energy efficiency, and area. Section 10 concludes the paper.

2. THE DESIGN SPACE OF ANALOG INFERENCE ARCHITECTURES

Analog accelerators perform matrix computations within the same memory arrays where the neural network weights are stored. In contrast to digital architectures that spend significant energy to read operands from memory, *in situ* computation eliminates the need to move weight data between processing elements. Within an array, individual analog MACs can also be conducted at a lower energy, higher density, and greater parallelism than digital MACs [45]. Due to these potential advantages, *in situ* MVM has attracted significant research attention for neural network inference [6, 11, 14, 42, 49, 56], as well as other applications [9, 18, 58].

Fig. 1(a) shows a conceptual example of an *in situ* MVM array that computes $\vec{y} = \mathbf{W}\vec{x}$. The memory cell conductances

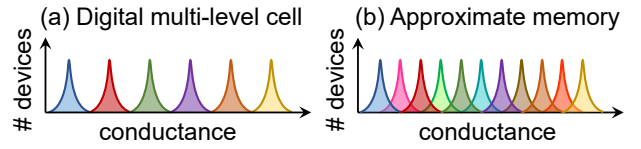


Figure 2: Conductance program distribution of a memory cell when used as two different types of memory.

\mathbf{G} are set proportional to the values of \mathbf{W} , and the rows are driven by input voltages \vec{V} that are proportional to \vec{x} . Each cell’s current is an analog product of its conductance G_{ij} and the applied voltage V_i . Kirchoff’s law then accumulates these products on the bit line (column) current I_j to form the dot product. The analog dot products are subsequently quantized using an ADC.

In situ MVM has been demonstrated using a wide variety of memory cell technologies [55, 61, 68]. Fig. 1(b) shows a 1T1R (1 transistor, 1 resistor) cell, which performs multiplication using Ohm’s law across a two-terminal programmable resistor, such as a resistive random access memory (ReRAM) or phase change memory (PCM) device. During an MVM, the transistor is transparent. Fig. 1(c) shows an alternative cell design, more typically used with transistor-based memories such as flash memory [5, 20, 24], where a select transistor uses the input to gate the flow of current through the memory element (green).

The conceptual example in Fig. 1 elides a number of practical implementation details. Prior work has proposed multiple approaches for data representation (\mathbf{W} and \vec{x}) that differ from the mapping in Fig. 1. Table 1 summarizes the design choices made by several recently proposed *in situ* MVM accelerators, which are explained below. A recent review of analog inference accelerators can be found in Xiao *et al.* [62].

2.1 Weight Bit Slicing vs Unsliced Weights

To represent matrices with more bits than can be reliably programmed in a device, many systems use *bit slicing* [9]. In bit slicing, the bit representation of each matrix element is divided into multiple slices, and the results of bit sliced MVMs are combined via shift-and-add (S&A) reduction [9, 21, 56]. Equation 1 shows how a matrix of 6-bit integers can be divided into two slices of three bits each.

$$\begin{bmatrix} 12 & 58 \\ 29 & 50 \end{bmatrix} = 2^3 \begin{bmatrix} 1 & 7 \\ 3 & 6 \end{bmatrix} + 2^0 \begin{bmatrix} 4 & 2 \\ 5 & 2 \end{bmatrix} \quad (1)$$

Bit slicing admits the use of high-precision weights with more possible values than the number of programmable levels in a memory device. In particular, it allows the use of inherently binary memories such as SRAM that cannot otherwise implement multi-bit weights [17]. Many accelerators use bit slicing as a way to tolerate analog memory cells with arbitrarily low precision, but this assumption has not been thoroughly evaluated on the basis of end-to-end inference accuracy and not just the weight precision.

To avoid the energy and area overheads of reading, digitizing and aggregating multiple bit-sliced arrays, the magnitude of a weight can also be fully encoded in one device [30, 34]. Unsliced weights ostensibly require very precise devices; however, for inference it can be sufficient to use analog mem-

Table 1: Comparison of data representation in selected prior work on analog *in situ* inference accelerators

Accelerator	Bit slicing	Full precision	Negative weights	# rows used per MVM	ADC bits B_{ADC}	DAC bits
Genov <i>et al.</i> [21]	Yes, 1b/cell	No	One’s comp	128	6	1
Memristive Boltzmann Machine [9]	Yes, 1b/cell	Yes	Two’s comp	32	5	1
ISAAC [56], Newton [49]	Yes, 2b/cell	Yes	Offset	128	8	1
PUMA [6]	Yes, 2b/cell	Yes	Offset	128	8	1
PRIME [11]	Yes, 4b/cell	No	Differential	256	6	3
Dot-Product Engine [30]	No	No	Offset	128	4	4
Sparse ReRAM Engine [66]	Yes, 2b/cell	Yes	Offset	16	6	1
CASCADE [14]	Yes, 1b/cell	Yes	–	64	10	1
TIMELY [42]	Yes, 4b/cell	No	–	256	8	8*
FORMS [69]	Yes, 2b/cell	Yes	Retrain [†]	8	4	1
Marinella <i>et al.</i> [45]	No	No	Differential	1024	8	1
Joshi <i>et al.</i> [34]	No	No	Differential	512	8	8
Yao <i>et al.</i> [67] (experimental)	No	No	Differential	128	8	1
Guo <i>et al.</i> [24] (experimental)	No	No	Differential	784	analog	1

[†] The network is trained so that all weights on a column have the same sign. * Inputs are encoded in the temporal duration of a pulse.

ory cells not as multi-bit digital memories as in Fig. 2(a), but as *approximate memories* shown in Fig. 2(b). The conductance state of an analog cell has a nonzero width due to process variations and noise. When used as a multi-bit digital memory, digital levels are mapped to states that have nearly zero overlap to enable statistically reliable readout of a single cell. When used as approximate memory, many more digital levels are mapped to the same conductance range by allowing states to overlap. The number of bits that can be mapped to a cell is ultimately limited by the resolution of the write circuitry and by the intrinsic physical resolution of the cell.

2.2 Input Bit Slicing

Bit slicing can also be applied to multi-bit inputs; each analog MVM processes one slice of \vec{x} and the full input is processed in multiple cycles. One-bit input slices are commonly used to avoid the high overhead of a multi-bit digital-to-analog conversion (DAC) per input on each MVM. A binary input voltage further allows the use of memory devices with a nonlinear I - V curve, since ideally only two points along this curve are sampled [61]. The cell configuration in Fig. 1(c) also relies on one-bit input slices since the select gate functions as a binary switch.

Partial MVM results from multiple input bit slices can be aggregated digitally using a S&A reduction network, similar to weight bit slices. The total number of ADC quantizations required per full MVM (all weight and input bits) is the product of the number of weight and input slices. Alternatively, S&A accumulation of sequentially applied input bits can be conducted by switched-capacitor circuits prior to the ADC, so that only one quantization is needed for all input bit slices [8, 14, 22]. Though feasible for 8-bit inputs, this technique cannot be scaled to arbitrarily many input bits due to the thermal noise floor on the analog signal.

2.3 Handling Negative Numbers

A variety of techniques have been developed to handle signed arithmetic with multi-bit weights and inputs. This work evaluates the two most common implementations of

negative weights: *offset subtraction* and *differential cells*. These schemes will be described in more detail in Section 4.1. Offset subtraction implements signed weights by including an offset in the conductance used to represent a zero. This then allows negative weights to be converted to positive conductances. This offset then needs to be subtracted from the dot product, either digitally or in the analog domain [56].

In the differential cells scheme, a signed weight is represented using the difference in conductance of two memory cells. The specific implementation of the subtraction varies across designs, and can be performed in the analog domain or after digitization. Analog current subtraction can be executed using opposite-polarity voltage inputs and Kirchoff’s law [45], or within the bit line peripheral circuitry [24, 34, 67]. Though this scheme uses two cells per weight (or per weight slice) rather than one, it possesses some advantages over offset subtraction, as summarized in Section 3.

Negative inputs can be handled by applying negative voltages to a resistive array [45], or by using two differential pairs (four cells) per weight [7, 54]. Notably, negative inputs are uncommon beyond the first layer of convolutional neural networks (CNNs) based on rectified linear (ReLU) activations. If both weights and inputs use one-bit slices, it is possible to use a two’s complement representation for both [9].

2.4 The Full Precision Guarantee

The conversion of an analog dot product to a digital result can incur a loss of precision. To provide theoretically digital accuracy from an analog MVM, prior work proposed the *full precision guarantee* (FPG) [9, 56]. The FPG posits that if the ADC has a unique level for every possible output of the MVM operation, then there is no loss of information from digitization. The information content of the analog signal, equal to the number of bits needed to specify all possible dot product values, is a function of the operand widths and the number of summed terms [56]:

$$B_{out} = \begin{cases} B_W + B_{in} + \log_2 N & \text{if } B_W > 1, B_{in} > 1 \\ B_W + B_{in} + \log_2 N - 1 & \text{otherwise} \end{cases} \quad (2)$$

where B_W is the number of weight bits per cell, B_{in} is the number of input bits per ADC operation, and N is the number of rows activated in an MVM. Notably, if input bit slices are accumulated digitally, B_{in} is the number of input bits per slice; if they are accumulated by analog circuitry, B_{in} is the smaller of the full input resolution or the circuit’s resolution. A non-integer value of B_{out} simply means that the number of possible MVM outputs is not a power of two.

The FPG can be stated as:

$$B_{ADC} = \lceil B_{out} \rceil \quad (3)$$

where B_{ADC} is the ADC effective number of bits. Typically, B_{in} , B_W and N are chosen such that $B_{out} \approx 8$ bits. Since the ADC cost rapidly becomes prohibitive with resolution [48], prior work using the FPG has been limited to smaller arrays and/or fewer bits per weight. Shafiee *et al.* [56] proposed a ‘flipped’ encoding of weight values to reduce the required ADC resolution to $\lceil B_{out} \rceil - 1$ bits.

In many systems, the final result after aggregating all slices is truncated before being passed to the next layer. This means that not all B_{out} bits from every slice are useful. Prior work has proposed avoiding this wasted computation by dynamically tuning the ADC resolution on a slice-wise basis [11, 49].

As shown in Table 1, not all systems adopt the FPG. This design choice will be evaluated in Sections 3 and 6.

2.5 Direct Weight Transfer vs Retraining

To improve inference accuracy in the presence of analog errors, many methods have been proposed to integrate these errors into the training process. A common approach is to add noise to weights and activations during forward propagation [28, 32, 34, 36, 44]. However, these techniques incur additional training overhead and are potentially difficult to co-integrate with state-of-the-art training workflows. Therefore, this paper focuses on accuracy with direct weight transfer: weights are mapped as-is to the memory cells, without any retraining or compensation for errors post-training [30, 32]. This work performs some pre-processing to calibrate the ADC limits, which is similar to the process used to quantize neural networks for digital inference [31]. This matches the standard by which the accuracy of digital accelerators is evaluated [51].

3. DESIGN PRINCIPLES FOR AN ERROR-RESILIENT ANALOG ACCELERATOR

This section summarizes the key design principles that enable an error-resilient analog inference accelerator. The remaining sections provide the modeling methodology and the results that support these general conclusions.

3.1 Precise Weight Representation \neq Precise Dot Product Computation

Bit slicing allows weights to be represented with arbitrarily high precision using memory cells that have only a few reliably distinguishable conductance states due to programming errors. However, the most important consequence of cell errors on accuracy is not their effect on the fidelity of individual weights, but rather the effect of summed cell errors on the fidelity of dot products, since these are the quantities that propagate from layer to layer during inference. This is

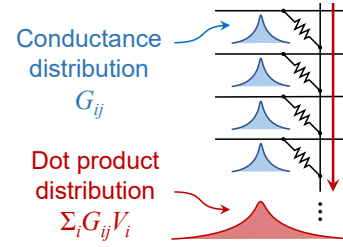


Figure 3: Accumulation of cell errors along a bit line.

illustrated in Fig. 3. Each cell has a random deviation ΔG_{ij} from its target conductance, and these errors are added when currents from multiple cells are summed on a bit line. If $\langle \Delta I_{ij} \rangle$ is the expected error in the product $G_{ij}V_i$ that results from this conductance error, then the expected error in the dot product I_j accumulated on the bit line is:

$$\langle \Delta I_j \rangle^2 = N \langle \Delta I_{ij} \rangle^2 \quad (4)$$

Some prior work uses the dot product error $\langle \Delta I_j \rangle$ as a starting point in an accuracy analysis [14, 52, 65], but this obscures the design choices that affect the size of the accumulated error.

Importantly, the above equation holds whether or not the distributions of conductance states within a cell overlap. To minimize the error in the dot product, the absolute width of the distribution ΔG is more important, and this quantity is not improved by ensuring that the utilized states are well separated. Therefore, *bit slicing does not provide a fundamental advantage to accuracy compared to approximate memories*, and conversely, it cannot be relied upon to save the accuracy when memory cells with inherently large errors are used. Many of the works listed in Table 1 choose to use bit slicing without evaluating the accuracy with unsliced weights, with the implicit assumption that the accuracy would fall significantly without bit slicing. In Section 5, we show that bit slicing does not in fact provide a large advantage to accuracy for the same device conductance precision.

Bit slicing can nonetheless provide a *small* improvement to accuracy, as will be explained in Sections 5.2 and 5.3. The origin of this benefit is subtle and does not stem from having well-separated memory states. Since the benefit tends to be small, it must be considered carefully against the large energy and area overheads of bit slicing, as shown in Section 9.

If the accumulated error ΔI_j can be reduced below the least significant bit (LSB) of the ADC, its propagation to the next layer can be suppressed. This can be achieved by using smaller arrays [43, 66], but this is inefficient as it amortizes the ADC energy cost over fewer MACs. Error correcting codes can correct a fraction of the dot product errors [19], but the simplest and least costly method of reducing these errors is to proportionally map weights to conductances.

3.2 Proportional Mapping Reduces Errors

3.2.1 Weight Proportionality

A very common property of neural networks is the abundance of low-valued or zero-valued weights. This is illustrated in the weight value distributions shown in Fig. 4 of four popular ImageNet neural networks. In digital inference accelerators, this property can be exploited to greatly compress the

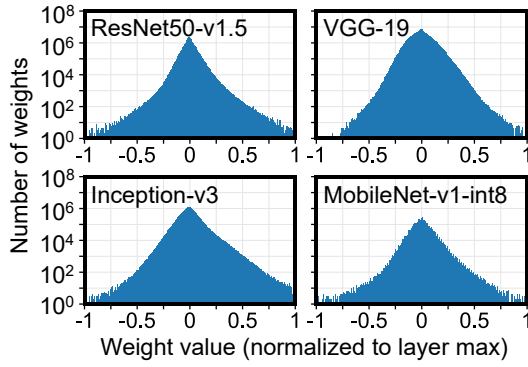


Figure 4: Weight value distributions of several popular ImageNet neural networks.

network size (via pruning) and the resultant sparsity can be used to save computation [25, 26]. Pruning is more difficult to exploit in analog accelerators, due to the rigid structure of a memory crossbar [62]. Nonetheless, it is possible to exploit zero and small-valued weights in analog accelerators by using **proportional mapping**: a linear relationship between numerical values in the algorithm and the physical quantities in the analog hardware.

With proportional mapping, weight values are mapped to conductances in proportion to their magnitude. This is implemented by using differential cells to encode negative weights in the manner described in Section 2.3, and by using cells with high On/Off ratio (G_{\max}/G_{\min}). Together with the strongly zero-peaked weight distributions in neural networks, proportional mapping can reduce the average cell conductance by orders of magnitude, as shown in Section 4.3.

Reduction of the average conductance is important because two types of analog errors tend to increase proportionally with conductance or current. First, the cell programming error ΔG typically increases with the programmed conductance G , as will be described in Section 5. For some technologies, like flash memory, this is a fundamental property of the device. Another source of error that increases with cell conductance is parasitic voltage drops across the columns and/or rows of the array, which nonuniformly distort the elements of a weight matrix as described in Section 8. Proportional mapping mitigates both of these errors, by matching the least-error conductance states to the most-used weight values.

3.2.2 Dot Product Proportionality

Proportional mapping is also important between dot products and analog outputs. Neural networks natively possess some tolerance to low-resolution activations during inference. Activations in ImageNet neural networks, for example, can typically be quantized to 8 bits after training without losing significant inference accuracy [31]. Can analog systems exploit this to perform accurate ImageNet inference with no more than 8 bits of ADC resolution? The answer is yes, and the key enabler is *dot product proportionality*. While an activation may tolerate quantization to 8 bits, this property might be lost if the same information is encoded in a quantity that is not proportional to the original activation. Ensuring proportionality between analog outputs and dot products connects the ADC resolution requirement to the algorithm’s intrinsic

precision requirements. Dot-product proportionality largely follows from weight proportionality, with the requirement that the current subtraction in differential cells be conducted in analog. This will be explored in Section 6.

3.3 The Full Precision Fallacy

The FPG requires the ADC to have a unique level for every possible output of an analog MVM, and thus match the precision of a digital processor. To be compatible with practical ADC resolutions (~ 8 bits), the FPG bounds the amount of computation that can be executed in the analog domain before digitization. This is expressed by Equation (2). There are two fundamental problems with the FPG.

First, the FPG is only meaningful if the accumulated cell error on all bit lines is below the LSB of the ADC. When cell errors are present, the analog input to the ADC does not necessarily have B_{out} bits of precision as given by Equation (2). Thus, in practice, satisfying the FPG requires not only the correct ADC resolution but also sufficiently accurate memory cells to ensure that the ADC resolution is fully utilized. Some early work on *in situ* MVM explicitly set the ADC resolution to match the expected level of accumulated cell error, using fewer bits than required by the FPG [21, 52].

Second, the FPG is imposed at the level of the analog MVM kernel, typically without full consideration of its utility for the accuracy of neural network inference. By focusing on the precision of an individual kernel rather than end-to-end system requirements, the FPG creates inefficiencies, as predicted by the end-to-end argument of Saltzer *et al.* [53]. Specifically, in Section 6, we show that in systems with dot-product proportionality, the FPG is too conservative. In these systems, *the ADC resolution requirement can be decoupled from the hardware configuration and dictated instead by the end-to-end accuracy of the neural network application*. Fortunately, the resolution requirement of ImageNet neural networks is also ~ 8 bits [31]. Removing the constraints of the FPG enables much more analog computation to be done for the same ADC resolution, improving energy efficiency.

4. ACCURACY EVALUATION METHOD

This section describes the methodology for inference accuracy evaluation for the results presented in the remaining sections. Unless otherwise stated, neural networks are quantized to 8-bit precision, a common use case for inference [31, 35].

4.1 Mapping Weights to Conductances

We assume a simple, parameter-less procedure to map a layer’s weights onto device conductances. First, the floating-point weight matrix \mathbf{W}_{FP} is scaled by the maximum absolute value $\max|\mathbf{W}_{\text{FP}}|$ into the range $[-1, +1]$. To quantize the weights to 8 bits, these weights are further scaled to the range $[-127, +127]$, then rounded to integers in this range. The same process is used for the weights of all layers.

The quantized weight matrix \mathbf{W} , with integer values in the range $[-127, +127]$, is then decomposed into one or more non-negative integer-valued matrices that can be mapped to conductances. The specific decomposition depends on the method used to handle negative weights (offset subtraction vs. differential cells) and encode weight precision (with or without bit slicing). Several examples are shown in Fig. 5 for an

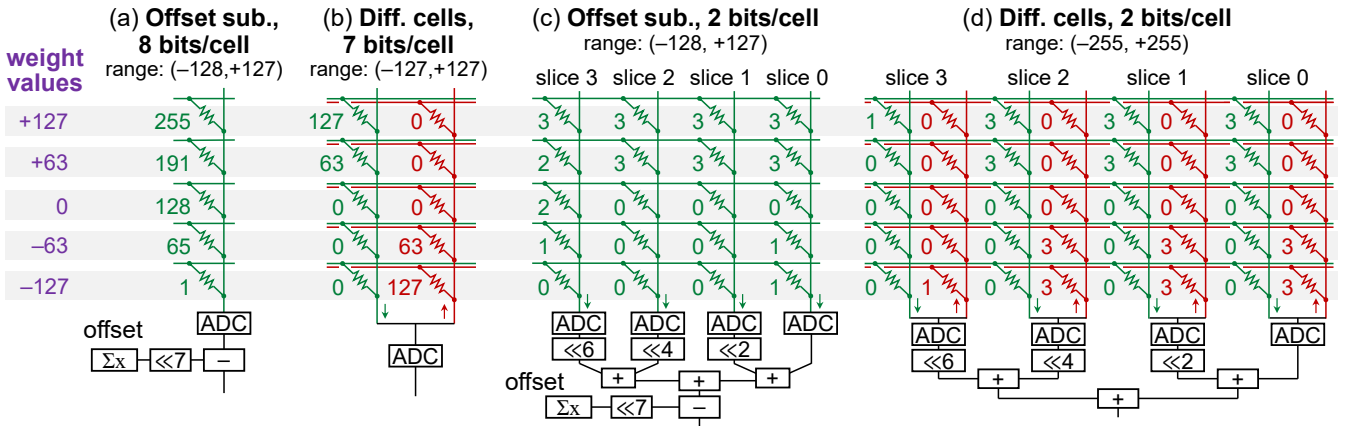


Figure 5: Four schemes for mapping weight values to conductance values in a memory array.

8-bit matrix. Although these representations are functionally equivalent in the absence of analog errors, they differ greatly in their sensitivity to these errors. These methods represent the majority of proposed analog accelerators.

4.1.1 Mapping without Bit Slicing

Fig. 5(a) shows the 8-bit matrix \mathbf{W} is mapped using offset subtraction without bit slicing, following the equation:

$$\mathbf{W}\vec{x} = \mathbf{W}_{\text{prog}}\vec{x} - 2^7 \mathbf{I}\vec{x} \quad (5)$$

where \mathbf{I} is the identity matrix and \mathbf{W}_{prog} is a strictly non-negative 8-bit matrix in the range [1, 255] that can be mapped directly onto conductances. This matrix has an offset such that a zero weight in \mathbf{W} is mapped to a value of 2^7 in \mathbf{W}_{prog} . This offset is subtracted from the MVM result to represent negative weights. Computing the offset term requires summing the elements of \vec{x} , which can be done digitally. Shafiee *et al.* [56] also proposed an analog computation of this offset with a “unit column”, which will be evaluated in Section 5.2. We note that a value of -128 in \mathbf{W} can be mapped by a value of 0 in \mathbf{W}_{prog} , but this state is left unused.

Fig. 5(b) shows how the same matrix is mapped using differential cells, following the equation:

$$\mathbf{W}\vec{x} = \mathbf{W}^+\vec{x} - \mathbf{W}^-\vec{x} \quad (6)$$

where the strictly positive weight matrices \mathbf{W}^+ and \mathbf{W}^- have 7-bit values in the range [0, 127], and are programmed onto the conductances of two sets of memory cells. This definition leaves some ambiguity about how two conductances are decided from a single weight value. This paper evaluates the convention where one cell in the pair encodes the magnitude of positive weights, while the other encodes the magnitude of negative weights. This means that at least one cell in every pair is left in the lowest conductance state. Note that the weight magnitudes in this scheme are directly mapped to the conductances of 7-bit cells, ensuring a proportional mapping.

The integer values in the non-negative matrices on the right sides of Equations (5) and (6) are mapped linearly to conductances. A value of 0 is mapped to the minimum conductance state G_{min} , while the maximum value in the range is mapped to G_{max} . Intermediate integers are linearly mapped to intermediate conductances.

4.1.2 Mapping with Bit Slicing

Fig. 5(c) shows an example of offset subtraction with bit slicing, which implements the following mapping:

$$\mathbf{W}\vec{x} = 2^6 \mathbf{W}_3\vec{x} + 2^4 \mathbf{W}_2\vec{x} + 2^2 \mathbf{W}_1\vec{x} + \mathbf{W}_0\vec{x} - 2^7 \mathbf{I}\vec{x} \quad (7)$$

where \mathbf{W}_i are the 2-bit slices of \mathbf{W} from lowest to highest significance. Each element of \mathbf{W}_i is integer-valued in the range [0, 3] and mapped to the conductance of a single cell. The offset is subtracted after the results of the slices are aggregated. As in the non-bitsliced case, this is not a proportional mapping, since a zero weight is mapped to an intermediate conductance in the top slice.

Fig. 5(d) shows an example of differential cells with bit slicing, which implements the following mapping:

$$\mathbf{W}\vec{x} = 2^6 (\mathbf{W}_3^+\vec{x} - \mathbf{W}_3^-\vec{x}) + 2^4 (\mathbf{W}_2^+\vec{x} - \mathbf{W}_2^-\vec{x}) + 2^2 (\mathbf{W}_1^+\vec{x} - \mathbf{W}_1^-\vec{x}) + (\mathbf{W}_0^+\vec{x} - \mathbf{W}_0^-\vec{x}) \quad (8)$$

where each 2-bit matrix \mathbf{W}_i^\pm is integer-valued in the range [0, 3]. This method uses a sign-magnitude representation and slices the magnitude bits across multiple cells. Within a slice, the magnitudes of positive weights are mapped to \mathbf{W}_i^+ and the magnitudes of negative weights are mapped to \mathbf{W}_i^- , and the resulting bit line currents are subtracted. The most significant slice is proportional to the weight value, and a zero weight is mapped entirely onto the ‘0’ state in all slices, as shown in Fig. 5(d). Because the four slices together represent 8 magnitude bits, the hardware in Fig. 5(d) can map a 9-bit signed weight in the range $[-255, +255]$.

4.1.3 Input Bit Accumulation

For all of the schemes above, one-bit input slices are assumed to simplify the input DAC and device requirements. For differential cells, results from different input bits are sequentially accumulated using analog circuitry as described in Section 2.2, such that $B_{\text{in}} = 8$ bits. For offset subtraction, analog accumulation requires summing all of the 8-bit elements of the input vector \vec{x} to compute the offset, which is more complex than summing the elements of \vec{x} one bit at a time. To avoid this overhead and to provide a baseline that is similar to prior work [6, 49, 56], offset subtraction is evaluated with digital S&A accumulation of input bits ($B_{\text{in}} = 1$ bit).

4.2 Accuracy Simulation of Analog MVMs

For a realistic accuracy simulation of an analog inference accelerator, we extend CrossSim [4] with a highly parameterizable model for an analog MVM array. CrossSim imports a Keras neural network model [13] and maps the weight matrix of each convolution and fully-connected layer to one or more memory arrays, representing different bit slices and matrix partitions, according to a chosen mapping scheme. Every analog MAC is simulated during inference. Digital operations such as the S&A aggregation of weight slices, ReLU activation, and inter-layer communication are assumed to be error-free. Convolutions are unrolled into a sequence of sliding window MVMs, executed on arrays of size $K_x K_y N_{ic} \times N_{oc}$ as described by Shafiee *et al.* [56] ($K_x \times K_y$ is the 2D filter size, N_{ic} and N_{oc} are input and output channel dimensions).

The modeling of random cell programming errors, ADC quantization, and parasitic voltage drops are described in Sections 5, 6, and 8, respectively, where the accuracy impact of each non-ideality is analyzed separately. This work does not study the effect of cycle-to-cycle read noise, which is similar to that of programming errors. Read noise has a weaker effect than programming errors when input bit slicing is used, as explained in Section 5.2. This work also does not study conductance drift over time, which is less generalizable across technologies.

The non-idealities mentioned above grow in severity with the number of cell currents summed on the same bit line, and hence limit the number of rows in the array. The maximum array size is treated as a parameter, and matrices that require more rows are partitioned evenly across equally sized arrays. The results from each array are separately digitized (and possibly clipped) before they are added.

To reduce digital processing overheads, batch normalization parameters are folded into the weight matrix of a convolution for all the evaluated networks [31]. Since the bias weights can lie in a different range from the other weights, representing them together in the same array can cause a loss of precision [31]. Therefore, for all layers the bias weights are stored separately from the array and added digitally to the MVM results.

4.3 ImageNet Neural Network Benchmark

Fig. 6 highlights the importance of using a realistic dataset for accuracy evaluations. The sensitivity to cell errors (described more fully in Section 5) differs dramatically for networks trained on three datasets—ImageNet [16], CIFAR-10 [38], and MNIST [39]—with ImageNet being by far the most sensitive. The validity of any study on accuracy in analog accelerators is therefore bounded by the complexity of the inference task.

To emulate a realistic machine learning application, most of the accuracy evaluation in this paper is based on the ResNet50-v1.5 network for ImageNet, using the reference implementation from the MLPerf Inference Benchmark v0.5 [3, 51]. To compare the error sensitivity of different neural networks (Section 5.4), we include three other popular ImageNet models: VGG-19 [57], Inception-v3 [60], and MobileNet-v1 [29]. For VGG-19 and Inception-v3, we use the reference implementations included in Keras Applications [13]. For MobileNet-v1, we evaluate the quantized implementa-

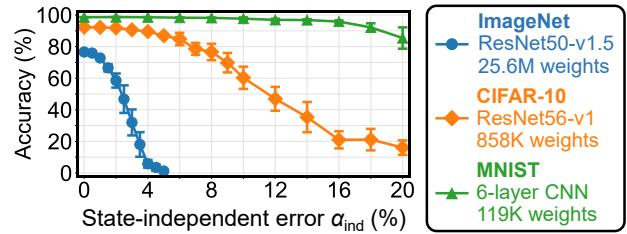


Figure 6: Accuracy sensitivity to state-independent cell errors for networks trained on three datasets. All cases assume unsliced weights with differential 7-bit cells.

Table 2: Evaluated neural networks

Neural network	# weights	ImageNet top-1 accuracy	
		50,000 images	1000 images
ResNet50-v1.5	25.6M	76.466%	77.5%
Inception-v3	23.9M	77.876%	77.8%
VGG-19	143.7M	71.256%	70.2%
MobileNet-v1 (int8)	4.3M	70.614%	71.8%
ResNet50-v1.5 (int4)	25.6M	76.154%	76.6%

tion with 8-bit integer weights that is provided as part of the MLPerf Inference Benchmark v0.5 [3]. The MobileNet model is quantized to 8 bits during training, since the same model trained without quantization loses significant accuracy when quantized after training [51]. Finally, we include a version of ResNet50-v1.5 that was trained by Nvidia at 4-bit precision, which will be described in Section 7. Table 2 shows the accuracy of the evaluated networks on the ImageNet validation set, before applying any errors. The accuracy is also shown on a fixed subset of 1000 images, which is used for the sensitivity analyses in this paper for computational tractability.

Weights are quantized to 8 bits before being mapped to hardware, and activations are quantized to 8 bits during inference. Except in the case of MobileNet, deployment at 8-bit precision does not need retraining, but the numerical range of the activations must be optimized to reduce quantization and clipping errors [31]. This is done by first running the model at floating-point precision and saving the activation values for all layers, using the MLPerf Inference calibration subset of 500 images [2]. For each layer’s collected activations \vec{x} , the range (x_{\min}, x_{\max}) is found that minimizes the L1-norm error $\epsilon = \|\vec{x} - \vec{x}_Q\|_1$, where \vec{x}_Q is obtained by clipping and quantizing \vec{x} to M bits in this range. A value of $M = 12$ was found to yield an optimal inference accuracy for 8-bit activations. The value of M does not correspond to any physical quantity in the system, and differs from the activation resolution because ϵ is not a true proxy for inference accuracy. The same activation ranges are used for all hardware implementations of a neural network. The resolution and range of the activations do not directly correspond to those for the ADCs, since multiple digitizations, a bias, and an activation function may be needed to produce one output activation. The ADC ranges are discussed in Section 6.

Fig. 7 shows the average conductance in each bit slice for the data mapping schemes in Fig. 5 when implementing ResNet50-v1.5. Here, an infinite On/Off ratio ($G_{\min} = 0$) is

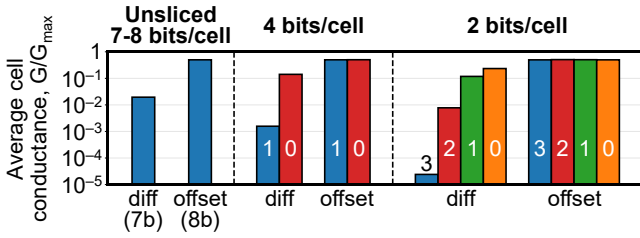


Figure 7: Average cell conductances for several different schemes to map the 8-bit weights in ResNet50-v1.5. The bars are labeled by bit slice (0 is lowest).

assumed. Fig. 7 shows that *using differential cells reduces the average cell conductance by multiple orders of magnitude* in the case of unsliced weights and in the higher slices with bit slicing. In the lower bit slices or when using offset subtraction, the average cell conductance is close to 50% of G_{\max} . As described in Section 3, this conductance reduction is a consequence of proportional mapping and the abundance of low-valued weights in the neural network. The following sections will explore the implications of the conductance distribution on inference accuracy.

5. ROBUSTNESS TO CELL ERRORS

Due to process variations and device and circuit limitations, there is always some uncertainty in the conductance of a programmed cell. This section considers the effect of cell conductance errors on end-to-end inference accuracy. Except in Section 5.4, all results are based on ResNet50-v1.5.

5.1 Error Properties of Memory Devices

Fig. 8 depicts two simple models of conductance error in memory devices. In the state-independent error model, the expected error ΔG does not depend on the conductance G and can be expressed as a fixed fraction of G_{\max} . In the state-proportional error model, ΔG is proportional to the conductance; a smaller conductance has a smaller error. The parameters α_{ind} and α_{prop} are defined such that when the two are equal, the corresponding errors ΔG are the same at the midpoint conductance: $G = 0.5G_{\max}$.

While real devices cannot be perfectly described by these models, many memory devices have the property that ΔG increases with G . As will be explained in Section 9, flash memory has approximately state-proportional error properties when operated in the subthreshold regime; this is due to the exponential dependence of source-drain conductance on the amount of stored charge. The property that the error ΔG increases with G has also been seen in PCM [34] and some ReRAM devices [27]. In these cases, ΔG is not strictly proportional to G , so the behavior is a mixture of the two error models analyzed here. Some other ReRAM devices have properties that are closer to state-independent error [46, 67]. Section 9 will evaluate the accuracy of a real memory device with a more complex state-dependent error characteristic.

To model cell errors, the conductance G of every cell in the network is perturbed with an error that is sampled from a normal distribution. The distribution has zero mean and a standard deviation ΔG , based on the equations in Fig. 8. These perturbed conductances are then used to simulate in-

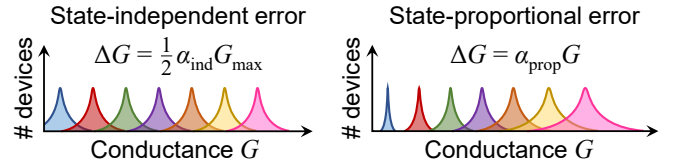


Figure 8: Two models for cell conductance error.

ference on 1000 images. This process is repeated ten times, with re-sampled cell errors, to obtain the variance in accuracy over these images.

To study the effect of cell errors alone, this section does not include ADC quantization. Without ADCs, the accuracy is independent of array size as cell errors are allowed to accumulate over the full size of the weight matrix (up to 4608 inputs in ResNet50-v1.5). Thus, the results here represent the worst-case effect of cell errors. The sensitivity to cell errors in the presence of an ADC will be shown in Sections 7 and 9.

Unless otherwise noted, the following evaluation assumes that $G_{\min} = 0$. The conductance On/Off ratio needed to approximate this idealization is found in Section 5.3. For differential cells, the examples with bit slicing use 9-bit weights to fully utilize the representational range of the hardware, while the unsliced case uses 8-bit weights. The difference in accuracy between 8-bit and 9-bit weights is 0.6% on this subset without cell errors.

5.2 Sensitivity to State-Independent Errors

Fig. 9(a) shows the accuracy sensitivity of offset-subtraction systems to state-independent errors, shown for different slice widths. In all cases, the accuracy is highly sensitive to error, falling nearly to zero at $\alpha_{\text{ind}} = 2.5\%$. The offset term to be subtracted is computed digitally, except in one case where a unit column is used.

The unit column is an additional column in the array whose conductances are all mapped to the center of the weight range [56]. The analog sum in this column is subtracted from all other sums. Fig. 9(a) shows that this method incurs a large accuracy loss. The unit column accumulates error just as the other columns do, and this adds to the error in all other dot products when the offset is subtracted. By correlating the errors in these dot products, the unit column also increases the variance in the accuracy.

Fig. 9(a) also shows that bit slicing can slightly improve accuracy. This is because the random programming errors in different bit slices can cancel. The amount of cancellation is limited, however, because the bit slices are not weighted equally when they are aggregated. The benefit of bit slicing can be analyzed in terms of the signal-to-noise ratio (SNR) of the dot product, following Genov and Cauwenberghs [21].

As an example, consider offset subtraction with two bits per cell. The dot products in different slices add as in Equation (7), while the errors in different slices add in quadrature. As a shorthand, let $D_i = \mathbf{W}_i \cdot \vec{x}$ denote the slice-wise dot products and σ_i the errors in D_i due to accumulated cell errors. The SNR prior to the digital offset subtraction is:

$$\text{SNR} = \frac{2^6 D_3 + 2^4 D_2 + 2^2 D_1 + D_0}{\sqrt{(2^6 \sigma_3)^2 + (2^4 \sigma_2)^2 + (2^2 \sigma_1)^2 + \sigma_0^2}} \quad (9)$$

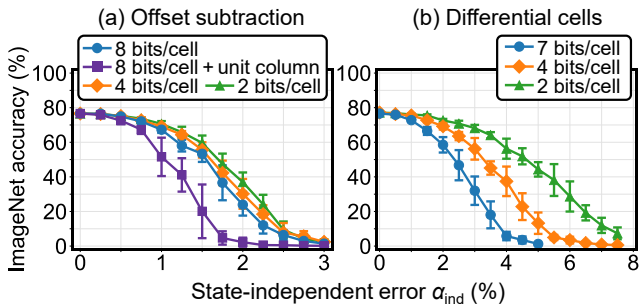


Figure 9: Sensitivity of ResNet50-v1.5 accuracy to state-independent errors using (a) arrays with offset subtraction and (b) arrays with differential cells. Error bars span two standard deviations over ten trials.

This expression can be simplified. With offset subtraction, the expected values of D_i are similar since every slice has close to the same average conductance (see Fig. 7). With state-independent errors, the expected errors σ_i in each slice must also be the same. Therefore:

$$\frac{\text{SNR}}{\text{SNR}_0} \approx \frac{2^6 + 2^4 + 2^2 + 1}{\sqrt{2^{12} + 2^8 + 2^4 + 1}} = 1.286 \quad (10)$$

where $\text{SNR}_0 = D_0/\sigma_0$ is the SNR with unsliced weights. Using four slices slightly increases the dot product SNR. One can show that there is a theoretical maximum SNR increase of $\sqrt{3}$ relative to unsliced weights, obtained in the limit of infinitely many 1-bit slices. Bit slicing thus provides a small accuracy benefit, consistent with Fig. 9(a).

Genov and Cauwenberghs [21] derived a further SNR benefit of up to $\sqrt{3}$ from *input* bit slicing, assuming that dot product errors for the different input bits are also independent. In general, this is not true of programming errors, which are static between analog MVM operations. However, this is true of conductance errors caused by cycle-to-cycle read noise. In systems that use input bit slicing, read noise has a weaker effect than programming errors because noise in different input bits within the same weight slice can cancel.

Fig. 9(b) shows that differential cells are more tolerant to state-independent errors. This is because by using two cells per slice, the dot product SNR improves: the signal range doubles, but the dot product error increases only by $\sqrt{2}$. Also, the improvement in accuracy with the number of bit slices is larger than for the offset case. Recall from Fig. 7 that unlike offset subtraction, the average conductance falls in value from the lowest to highest slice when using differential cells. This trend counteracts the exponential weighting of bit slices in Equation (9) so that the dot products in different slices are more equal in value. This enables greater cancellation of the dot product errors to occur. Because the assumptions that lead to Equation (10) do not fully hold, the SNR improvement from bit slicing can exceed $\sqrt{3}$.

5.3 Sensitivity to State-Proportional Errors

Fig. 10(a) shows the sensitivity of offset subtraction systems to state-proportional cell errors. These systems lack proportionality between weights and conductances, and thus do not substantially discriminate between state-independent and state-proportional errors. A comparison of Fig. 9(a) and

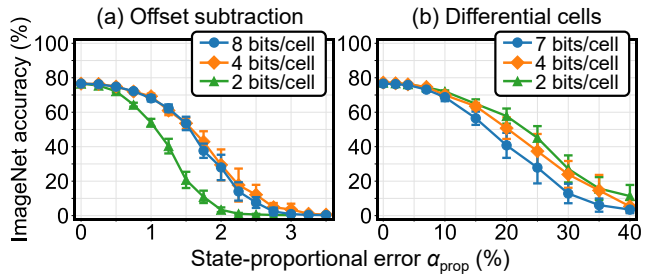


Figure 10: Sensitivity of ResNet50-v1.5 accuracy to state-proportional errors using (a) arrays with offset subtraction and (b) arrays with differential cells.

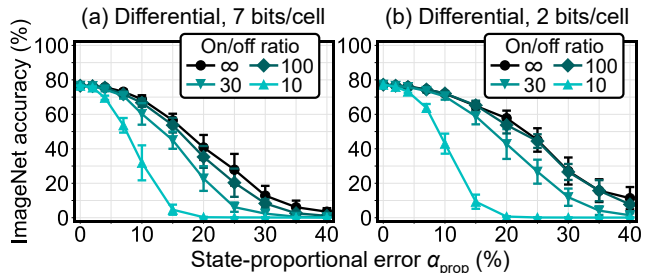


Figure 11: Sensitivity of ResNet50-v1.5 accuracy to state-proportional errors and cell On/Off ratio, using differential cells (a) without bit slicing and (b) with bit slicing.

Fig. 10(a) reveals that the sensitivity is similar between the two error types, but in Fig. 10(a), better accuracy is obtained with fewer slices. This is a subtle effect that stems from the bit representation used by the offset scheme in Fig. 5(c). Specifically, small negative weights use a conductance of G_{\max} in the second most significant slice, so that dot products that use more negative weights will have high state-proportional errors. This slice is weighted more heavily in systems with more slices, leading to a lower accuracy.

Fig. 10(b) shows that *a system with differential cells is very tolerant to state-proportional errors, with $>10\times$ the resilience of offset-subtraction systems*. This large difference results from proportionality; as shown in Fig. 4, most of the weights in ResNet50-v1.5 are close to zero. Consequently, Fig. 7 shows that for both unsliced weights and the top slice in bit-sliced systems, the average cell conductance is a small fraction of G_{\max} and thus has a small error. The lower slices have larger errors, but these are suppressed by the S&A operation. As arrays with fewer bits per cell are used, the top slice becomes more zero-dominated, reducing dot product errors and enabling higher accuracy. Importantly, for bit-sliced systems, a large fraction of the improvement from Fig. 9(b) to Fig. 10(b) can be attributed to the error reduction specifically in the minimum conductance state G_{\min} .

When the cell has a finite On/Off ratio ($G_{\min} > 0$), there is only partial proportionality between the weight magnitudes and cell conductances, and current can flow through cells that encode zero-valued weights. Fig. 11 shows that a low On/Off ratio increases the sensitivity to state-proportional errors, but an On/Off ratio of 100 has nearly the same resilience as an infinite On/Off ratio. The effect is similar in differential cells with and without bit slicing.

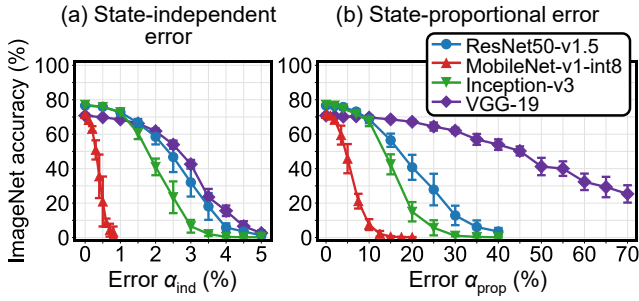


Figure 12: Sensitivity of four ImageNet neural networks to (a) state-independent errors and (b) state-proportional errors. The results assume differential cells without bit slicing and infinite On/Off ratio.

With an On/Off ratio of 100 or more, systems with differential cells see nearly zero accuracy loss for state-proportional errors below $\alpha_{\text{prop}} = 5\%$, even if these errors are allowed to accumulate in an array as large as the weight matrix (up to 4608 rows). The high-accuracy regions of the sensitivity curves in Fig. 9 to 11 correspond to the regime where direct weight transfer can be used with negligible accuracy penalty. Retraining is expected to be useful when the cell error falls in the intermediate-accuracy regions of these curves.

5.4 Error Sensitivity vs. Neural Network

Fig. 12 generalizes the conclusions from the previous sections to three other ImageNet neural networks, whose weights are all quantized to 8 bits. All of the evaluated networks have a much weaker sensitivity to state-proportional errors than state-independent errors. This results from the proportionality of both the cell conductance and the conductance error to the weight value, combined with the fact that all of the networks have a strongly zero-peaked weight value distribution as shown in Fig. 4. To obtain both types of proportionality, analog accelerators should use cell technologies with both *state-proportional errors and high On/Off ratio*. Field-effect transistor memories such as flash can fulfill both requirements, as we show in Section 9.

The differences in sensitivity across the four networks can be explained to first order based on the number of parameters, listed in Table 2. MobileNet has by far the fewest weights and thus the least amount of redundancy in its information content; therefore, its accuracy is more sensitive to errors in these weights. The opposite is true for VGG-19, which has the most weights and is thus the most error-tolerant [70]. The strong sensitivity of VGG-19 to state-independent errors, relative to its larger model size, is due to a very large fully-connected layer (25088 matrix rows) with a large amount of error accumulation. With state-proportional errors, the large matrix size is less consequential, since most of the elements have small or zero values.

Fig. 12 demonstrates the intuitive result that error tolerance can be achieved at the algorithm level by using a larger network with more redundant parameters, such as VGG-19. However, a larger network requires more energy and area to deploy. Reducing the size of the state-proportional cell error α_{prop} allows similar (or superior) accuracy to be achieved using a network with a smaller footprint.

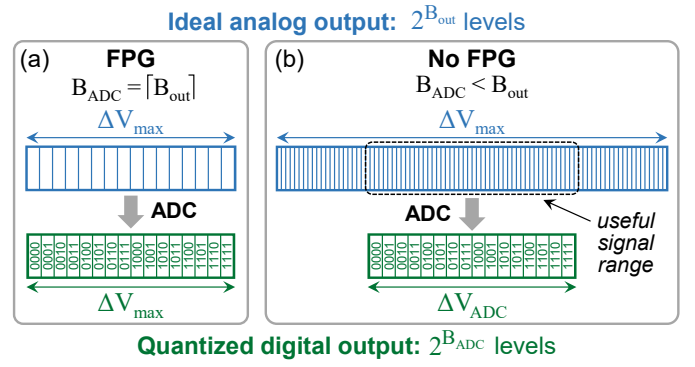


Figure 13: (a) The FPG provides a one-to-one mapping between possible analog outputs and ADC levels. (b) When not using the FPG, the range and resolution of the ADC can be chosen independently from those of the analog signal to minimize quantization and clipping errors.

6. ROBUSTNESS TO QUANTIZATION ERRORS

This section shows that by digitizing analog outputs at a precision that matches the network’s inherent precision requirements, the ADC requirements can be relaxed to the minimum feasible resolution that still yields high accuracy. The analog output is assumed here to be a voltage V .

6.1 ADC Errors

The bit resolution B_{out} given by Equation (2) is the resolution contained in the analog output of the array if the analog computation were free of errors. Equivalently, B_{out} measures the amount of computation that is done by the array before leaving the analog domain. Under the FPG, the ADC resolution is set equal to B_{out} , so that there is a one-to-one mapping from the possible analog outputs to the ADC’s digital levels, as shown in Fig. 13(a). Ideally, this guarantees no loss of information upon digitization. However, because the ADC resolution must be kept moderately low (typically 8 bits) due to energy considerations, the FPG limits the amount of analog processing and its associated energy benefits.

Alternatively, the ADC resolution B_{ADC} can be kept well below B_{out} . In this case, the ADC *compresses* a higher-resolution analog output into a lower-resolution digital output, as shown in Fig. 13(b). This compression induces two potential kinds of error: (1) quantization error, due to the potentially larger separation between ADC levels compared to the minimum separation between analog output levels, and (2) clipping error, if the signal range spanned by the ADC is smaller than the possible range of the signal. Any signal lying outside the ADC range is assumed to clip to the highest or lowest ADC level.

6.2 Calibrating the ADC Range

A way to eliminate clipping errors altogether is to make the ADC range equal to the maximum possible range of output voltages: $\Delta V_{\text{ADC}} = \Delta V_{\text{max}}$. In a practical inference application, however, the analog outputs may be much smaller on average than ΔV_{max} , and this design choice would leave most of the ADC levels heavily underutilized, increasing quantization errors. To better utilize the ADC levels, the range

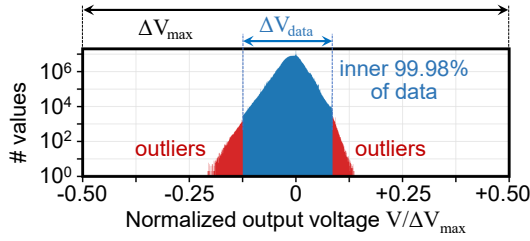


Figure 14: Distribution of output voltages for the sixth convolution layer of ResNet50-v1.5 (res2b_branch2b), using differential cells with unsliced weights, collected using the MLPerf calibration set.

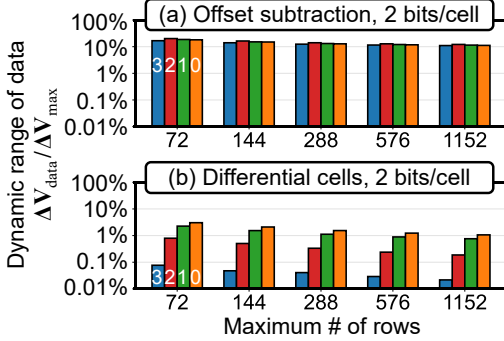


Figure 15: Output voltage statistics for ResNet50-v1.5 with (a) offset subtraction and (b) differential cells, using 2 bits/cell, with different constraints on array size. Ranges are averaged over all layers. Bars are colored by bit slice index (0 = lowest).

of values quantized by the ADC should be calibrated to the *useful* range of the analog signal, as illustrated in Fig. 13(b).

Our process for determining the optimal ADC quantization ranges is qualitatively similar to that used by Jacob *et al.* for low-precision digital inference [31], where the optimal activation quantization ranges are found from the activation statistics seen during training. Similar ideas have also been proposed for analog systems [23, 34]. We collect the statistics on every array’s output voltages (ADC inputs) by simulating inference on the MLPerf calibration subset of 500 ImageNet images [2]. The ADC limits of each layer are separately calibrated, as are the ADC limits of different bit slices within a layer, whose outputs can differ greatly in range.

Fig. 14 shows an example distribution of normalized output voltages, for a layer in ResNet50-v1.5. ADC range calibration relies on a single statistical property of these distributions: the range ΔV_{data} that contains the inner $P = 99.98\%$ of all collected values of V . This was empirically determined to be the useful signal range for ResNet50-v1.5, as clipping the remaining 0.02% of outlier values had a negligible effect on accuracy. The ADC limits are chosen to be just large enough to contain the useful signal range (i.e. $\Delta V_{\text{ADC}} \geq \Delta V_{\text{data}}$). For bit-sliced systems, the ADC limits of different slices are constrained to differ only by a power of two; this ensures that their results can still be aggregated via S&A operations without any complex scaling steps. With unsliced weights, there is no such constraint on the ADC limits. We note that in general, the definition of the useful signal range (set by

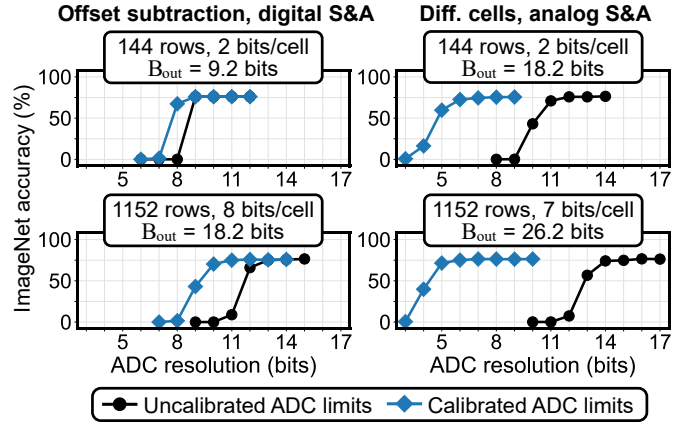


Figure 16: ImageNet accuracy using ResNet50-v1.5 vs ADC resolution for different weight mapping schemes, without calibrated ranges ($\Delta V_{\text{ADC}} = \Delta V_{\text{max}}$) and with calibrated ranges ($\Delta V_{\text{ADC}} \approx \Delta V_{\text{data}}$).

the single parameter P) may need to be tuned to optimize the accuracy for a given neural network, dataset, mapping scheme, and ADC resolution.

Comparing the useful signal range ΔV_{data} (normalized to ΔV_{max}) of different mapping schemes and bit slices reveals important insights about their output voltage distributions: this is shown in Fig. 15. When using offset subtraction, as in Fig. 15(a), the useful signal occupies 10-20% of ΔV_{max} ; the remainder is used only by outlier values. This somewhat low percentage results from the fact that input activations tend to concentrate near zero, especially in ReLU networks [47]. Thus, the ADC levels can be safely re-allocated to cover only this smaller range to offer better signal resolution.

Fig. 15(b) shows that the useful signal range is orders of magnitude smaller for differential cells: less than 0.1% of ΔV_{max} for the top slice. This is principally a result of proportional mapping: since differential cells use much lower conductances as shown in Fig. 7, the output voltages are reduced correspondingly. There is also a significant signal reduction from the analog cancellation of positive and negative bit line currents. The smaller signal range enables a much more aggressive reduction of the ADC range.

6.3 Matching the ADC to the Algorithm’s Precision

Fig. 16 shows the ADC resolution sensitivity of ImageNet accuracy for different mapping schemes. ADC quantization is assumed to be deterministic, and cell errors are not included in order to isolate the ADC’s effect. As described in Section 4.1.3, input bits are aggregated with digital circuitry for offset subtraction ($B_{\text{in}} = 1$ bit) and with analog circuitry for differential cells ($B_{\text{in}} = 8$ bits). This leads to a much higher analog resolution B_{out} for differential cells.

In all cases, calibration of the ADC range allows high accuracy to be obtained at a reduced resolution. By confining the ADC range to the useful signal range, quantization errors are reduced for the same number of levels. The benefit of calibration is greater for differential cells, which have a much smaller useful signal range as explained in Section 6.2.

Fig. 16 further shows that after range calibration is ap-

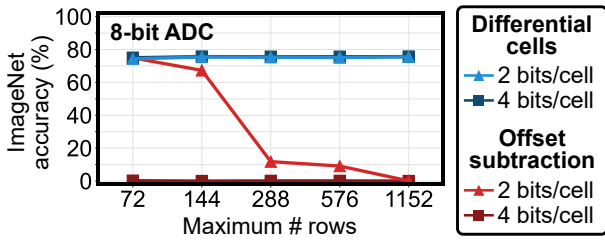


Figure 17: ImageNet accuracy using ResNet50-v1.5 with a calibrated 8-bit ADC for different mapping schemes, cell precision, and array size. The ADC range was separately calibrated for each point.

plied, differential cells can tolerate an ADC with several fewer bits of resolution than offset subtraction systems: evidently, differential cells are more resilient to quantization errors. This resilience can be understood by again considering the proportional weight mapping property of differential cells. Combined with the analog subtraction of currents, this results in *dot product proportionality*: the voltages at the ADC input are proportional to the numerical values of the dot products (or the slice-wise dot products). Offset subtraction systems lack this critical proportionality, since an offset must be subtracted after the ADC to obtain the true dot products.

Dot product proportionality implies that the data compression function of the ADC is effectively applied to the numerical dot products. Therefore, the required ADC resolution is directly connected to *the neural network’s inherent sensitivity to data precision*, which is hardware-independent and is fully decoupled from B_{out} . The effect is most striking with unsliced weights, where $B_{\text{out}} = 26.2$ bits but high accuracy is maintained down to $B_{\text{ADC}} = 7$ bits. This is close to the inherent precision sensitivity of ImageNet neural networks, which is typically about 8 bits [31].

Fig. 17 compares systems with offset subtraction (without dot product proportionality) and differential cells (with dot product proportionality) at a fixed ADC resolution of 8 bits. Offset-subtraction systems can only tolerate an 8-bit ADC when the array is small (≤ 144 rows) and the weights are finely sliced (≤ 2 bits/cell), which together bring B_{out} close to 8 bits. Differential cells suffer almost no accuracy loss with an 8-bit ADC regardless of the bits per cell and array size; this again illustrates that the accuracy is decoupled from B_{out} . Dot product proportionality makes a practical ADC resolution of 8 bits compatible with a much larger analog resolution B_{out} . Equivalently, much more computation can be done in the analog domain before the signal is ever converted to digital. This has significant consequences for energy efficiency, discussed in Section 9.

7. SUPPRESSING ERROR PROPAGATION

As shown in Section 5, the accuracy loss due to accumulated cell errors can be minimized by using sufficiently precise memory cells and exploiting state-proportional errors. With less precise cells, some prior work has relied on ADC quantization to cut off the propagation of cell errors from layer to layer in a DNN [30, 66, 71]. Yang *et al.* [66] activated only a few rows per MVM, such that on average, the accumulated errors on a bit line fall below the separation of levels in an

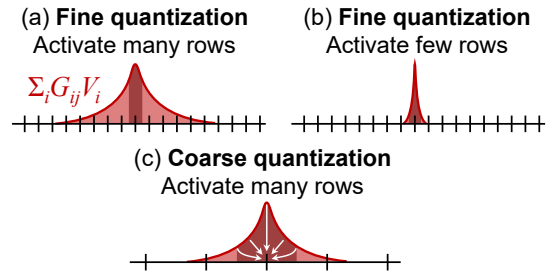


Figure 18: (a) The distribution of accumulated cell errors can span multiple digital levels. Propagation of these errors can be suppressed if they fall within a single level. For the same cell error, this can be achieved by (b) activating fewer rows per MVM or by (c) using coarse ADC or activation quantization.

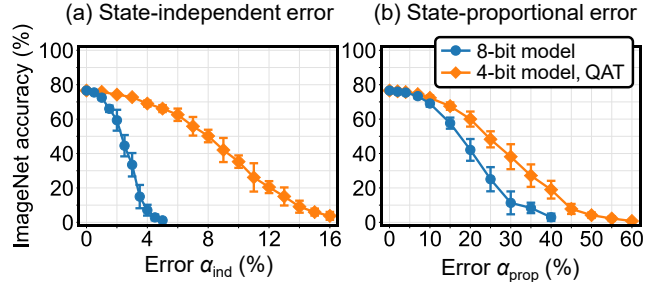


Figure 19: Sensitivity of two ResNet50-v1.5 models at 8-bit and 4-bit precision to (a) state-independent errors and (b) state-proportional errors. Simulations include calibrated 8-bit ADCs and assume differential cells with unsliced weights and 1152 maximum rows.

ADC, as shown in Fig. 18(b). While this approach succeeds in suppressing error propagation, it reduces energy efficiency since many more analog MVMs (and ADC operations) are needed to process each layer. A coarse ADC can provide the same benefit without reducing the number of rows, as shown in Fig. 18(c), but the accuracy would suffer due to quantization errors, as discussed in Section 6. A purely hardware solution cannot solve this problem, but it is possible to eliminate the quantization errors by training a DNN to tolerate *low-precision activations* during inference. This would combine the benefits of low quantization errors, greater resilience to cell errors, and high energy efficiency.

Unlike training techniques that are specialized for analog systems (see Section 2.5), quantization-aware training (QAT) benefits digital accelerators by reducing the computational load at inference time. Therefore, there has been much recent work on 4-bit or lower resolution networks with nearly no accuracy loss relative to floating-point networks [12, 59, 70]. Importantly, the broad applicability of low-precision networks increases the likelihood that QAT methods can be integrated into state-of-the-art training workflows.

This section evaluates a 4-bit QAT network with the ResNet50-v1.5 topology, submitted by Nvidia to the MLPerf Inference Benchmark [1]. The network uses 4-bit weights and activations in all layers except the first and last, which use 8-bit weights. Each ReLU output is multiplied by 16-bit scaling

factors before quantizing to 4 bits. The digital software accuracy of this network is shown in Table 2. When simulating the analog accuracy of this network, these scaling steps are processed digitally between *in situ* MVMs.

Fig. 19 compares the error sensitivity of the 4-bit QAT model with the floating-point ResNet50-v1.5 model, whose weights are quantized to 8 bits after training. For a fair comparison, an 8-bit ADC is included for both cases; in the 4-bit model, this higher-resolution ADC helps minimize errors prior to the 4-bit quantization step, which is performed digitally. Fig. 19(a) shows that the 4-bit model is substantially more resilient to state-independent errors than the 8-bit network. This results entirely from activation quantization, and not weight quantization. The large separation between the 16 activation levels effectively cuts off the propagation of accumulated cell errors from one layer to the next. The same cell error results in a smaller dot product error on average.

Fig. 19(b) shows that the 4-bit network is also more resilient to state-proportional error, but here the benefit is smaller. This can be explained by the different weight value distributions of the two networks. With only 16 levels, the distribution of the 4-bit weights cannot peak as sharply at zero as the 8-bit weights, which have 256 levels. As a result, the memory cells in the 4-bit network have a significantly higher average conductance (7.52% of G_{\max}) than the cells in the 8-bit network (1.95% of G_{\max}). Thus, deeply quantized *weights* can actually be harmful for analog systems, as it leads to a higher state-proportional error per cell. This is ultimately outweighed by the benefit of quantized activations, so there is a net improvement in error sensitivity.

Notably, for the 8-bit network, the sensitivity to both types of error with an 8-bit ADC remains largely unchanged from the case with no ADC quantization, shown in Fig. 9(b) and Fig. 10(b). As shown in Section 6, a calibrated 8-bit ADC on its own induces very little accuracy loss for this network. When the ADC levels are relatively finely spaced, as depicted in Fig. 18(a), deterministic quantization neither suppresses nor compounds the effect of accumulated cell errors.

8. MITIGATING PARASITIC RESISTANCE

Errors in the current conducted by a cell can arise not only from conductance errors, but also from voltage errors. A major source of voltage errors is the parasitic metal resistance, which induces voltage drops along the array’s rows and columns. The resulting errors in the cell currents are spatially non-uniform and input-dependent. The effect grows super-linearly with array size, as each new row contributes both a line resistance and a source of current. Together with accumulated cell errors, this effect limits the size of an *in situ* MVM. It is well known that parasitic resistance degrades MVM accuracy and some compensation methods have been proposed [30,32,33,71]. Here, we evaluate the end-to-end accuracy impact of parasitic voltage drops and their dependence on architecture-level design choices.

Since parasitic voltage drops are proportional to bit line currents, they can be compared to state-proportional cell errors, described in Section 5.3. Like state-proportional cell errors, parasitic resistance errors can be reduced by using a proportional mapping. Proportionality exploits the zero-peaked distribution of the weights to reduce the average cell

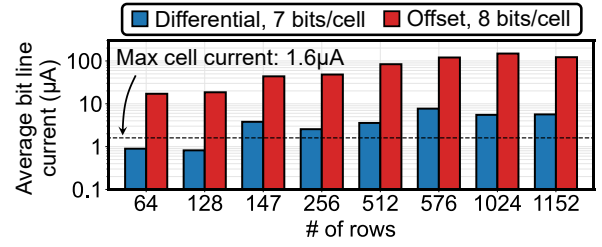


Figure 20: Bit line currents vs array size and mapping for ResNet50-v1.5. Layers that use the same # rows are averaged together. Inputs are applied bit-wise based on the cell in Fig. 1(c), and results are computed for the input LSB, which activates the most rows. The cell current varies from 0 to $I_{\max} = 1.6 \mu\text{A}$ (infinite On/Off ratio).

conductance, and hence the accumulated bit line currents. Fig. 20 shows that for ResNet50-v1.5, differential cells reduce the average bit line current by more than an order of magnitude. Even in arrays with as many as 1152 rows, the average bit line current (before analog subtraction) is only a few times the maximum current of a single cell. The lower bit line currents directly lead to smaller parasitic voltage drops.

To a greater degree than cell errors or ADC errors, errors induced by parasitic resistance depend on the specific array topology. The following analysis assumes the memory cell in Fig. 21(a), which is the same as that in Fig. 1(c). Input bits are applied to the gates of select transistors that draw nearly zero current [5, 9, 10, 20]. All cells source current from a low-resistance power distribution network (V_D). Thus, only the parasitic resistance of the bit line is considered, whose value between two adjacent cells is denoted R_p . The bottom of the bit line is held at virtual ground by the peripheral circuitry, such as a current integrator [45] or transimpedance amplifier [41]. For computational tractability, neural network simulations use the approximate circuit in Fig. 21(b). Select transistors are modeled as ideal switches, and a small-signal approximation is made to model the memory devices as linear resistors. To generalize to any memory cell or metal interconnect technology, the sensitivity analysis uses the normalized parasitic resistance \hat{R}_p , defined as the ratio of R_p to the minimum cell resistance $1/G_{\max}$.

When applying inputs one bit at a time, the effect of parasitic resistance varies considerably with bit position. This is because activations are typically skewed heavily toward low values, which makes the higher bits more sparse [47]. The lower bits have less sparsity, activate more rows, and have the largest parasitics-induced errors. However, these errors are suppressed to some degree by the input S&A operation.

Fig. 21(c) shows the sensitivity of ImageNet accuracy to parasitic resistance for three different weight mapping schemes. The array is limited to at most 1152 rows. The offset subtraction case is more than two orders of magnitude more sensitive to parasitic resistance than differential cells. This large difference can be attributed to three causes. First, due to the lack of proportional mapping, the offset case has a much higher bit line currents (see Fig. 20) and thus larger parasitic voltage drops. Second, for the same voltage drops, cells with high conductance contribute a larger error current to the bit line than cells with low conductance. Third, in the

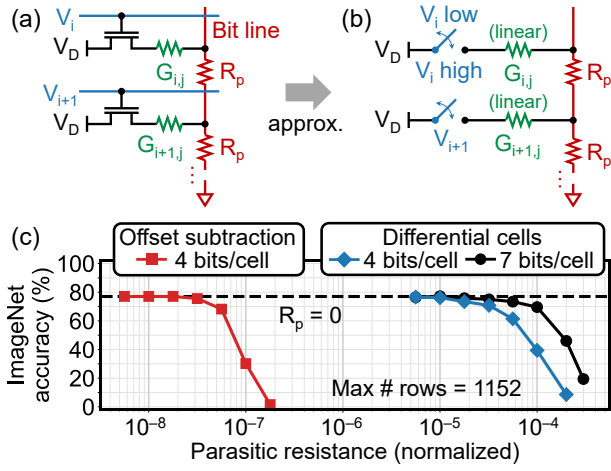


Figure 21: (a) Modeled unit cell and (b) its approximation. (c) ImageNet accuracy using ResNet50-v1.5 (500 images) versus normalized parasitic resistance.

differential case, parasitic resistance perturbs the current on both the positive and negative bit lines in the same direction: downward. When these currents are subtracted, a significant portion of the error induced by the parasitic voltage drops cancels. In the offset case, this cancellation does not occur since the subtracted offset is computed digitally. For the systems that use differential cells, the case with 4 bits/cell is slightly more sensitive to parasitic resistance than unsliced weights. This is because the lower bit slice has higher conductances due to its lack of proportionality (see Fig. 7), making it more sensitive to parasitic resistance.

Using differential cells, the accuracy loss is negligible for $\hat{R}_p \leq 10^{-5}$. This ratio is realistically achieved using cell resistances above ~ 100 k Ω and metal interconnects used in scaled process nodes, which can have a resistance of ~ 1 Ω per cell in a memory array [50]. Analog MVMs can thus be scaled to large arrays (~ 1000 rows) without being limited by parasitic resistance.

9. CASE STUDY: SONOS MVM CORE

This section demonstrates the design principles outlined in the previous sections using a real memory technology. The case study is a SONOS (silicon-oxide-nitride-oxide-silicon) charge trap memory that has been fabricated in an embedded 40nm process, and for which arrays have been electrically characterized to obtain the cell error properties. This section also examines the effects of the previously described design principles on energy efficiency and area, which can be generalized to other technologies.

9.1 SONOS Approximate Memory Device

The two-transistor SONOS flash memory cell has the configuration in Fig. 1(c). The SONOS device is programmed by adding or removing charge from the nitride storage layer, which shifts the threshold voltage V_T of the transistor channel. The threshold voltage in turn modulates the cell's drain current I_D . The SONOS gate stack and write process were optimized for operation as approximate memory [5, 63]. Histograms of the cell drain current, measured at $V_G = 0V$ and

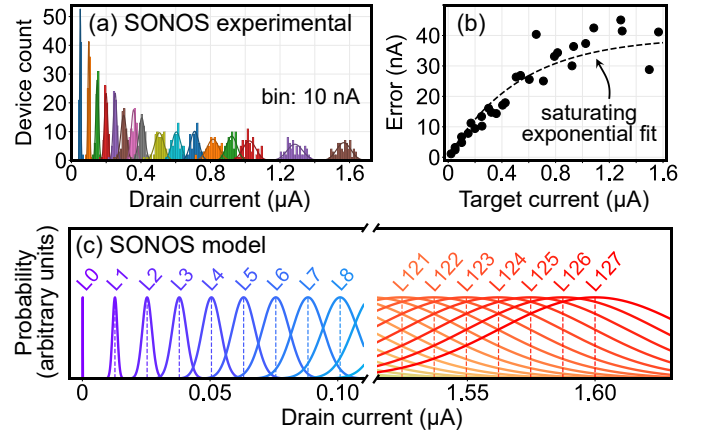


Figure 22: (a) Measured current distributions of SONOS memory cells. Each color is a group of 64 cells programmed to the same target current. (b) Standard deviation of the distribution for various target currents, including histograms not shown in (a). (c) Modeled program error distribution of a 7-bit SONOS cell, based on fit to measurements.

$V_D = 0.1V$, are shown in Fig. 22(a) for various target currents with $I_{max} = 1.6$ μA . The same biases are used during an MVM. Each histogram is fit to a normal distribution whose width is the expected programming error in a cell.

Fig. 22(b) shows the error as a function of current. The cell error is approximately state-proportional below 0.5 μA with $\alpha_{prop} \approx 6\%$, and saturates at high conductance. This property comes from the fact that the SONOS transistor is designed to operate in the subthreshold regime at a fixed bias of $V_G = 0V$:

$$I_D = I_0 \exp\left(-\eta \frac{qV_T}{kT}\right) \quad (11)$$

where I_0 is a constant, η is the gate efficiency, q is the electron charge, k is the Boltzmann constant, and T is the temperature. Differentiating the above with respect to V_T gives:

$$dI_D = -\frac{q\eta}{kT} dV_T \times I_D \quad (12)$$

Since the amount of stored charge is related linearly to V_T , the error in the charge injection or removal process is proportional to the error dV_T . Equation (12) shows that for the same error in the write process, the error in the cell current dI_D is proportional to the cell current I_D . This is consistent with the data in Fig. 22(b). At currents above around 0.8 μA , corresponding to lower V_T , the device leaves the subthreshold regime and the error consequently increases sublinearly with the current. As discussed in Section 5, state-proportional error is highly advantageous for neural network inference, as it matches the most frequently used weight values to devices with the least error.

The state-dependent error of the SONOS device is modeled within CrossSim using a saturating exponential fit to the data, shown in Fig. 22(b). The SONOS cell is used as an approximate 7-bit memory; the modeled program error distributions of the 128 target current levels are shown in Fig. 22(c). The device is programmed into deep subthreshold (highest possible V_T) for the lowest state to realize an On/Off

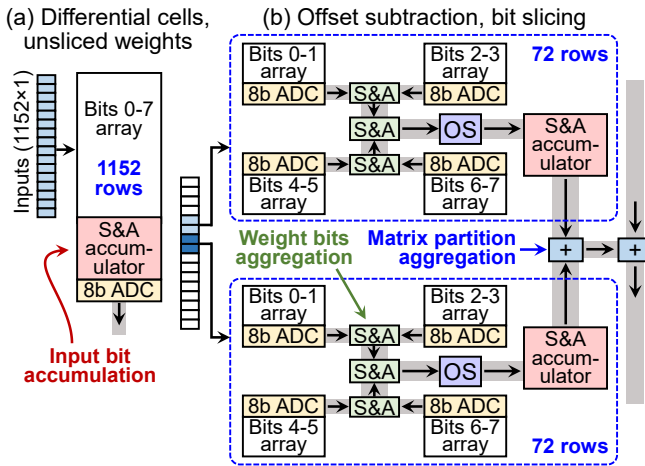


Figure 23: Two core configurations used in the energy and area evaluation. Based on results in Section 6, a matrix with 1152 rows must be partitioned across multiple arrays to maintain accuracy with offset subtraction (OS).

current ratio of 10^7 . The ImageNet accuracy with this device will be evaluated in Section 9.4.

9.2 MVM Core Design

Since this work addresses the design of the analog core, the energy and area results here will be restricted to the core level for generalizability. A core is defined as the collection of processing elements that perform a full-precision MVM; all bit slices, input bits, and matrix partitions. Fig. 23 shows the evaluated core design for two example weight mapping schemes. An 8-bit ADC is used for all of the considered design points.

As shown in Section 6, cores that use differential cells can achieve high accuracy with an 8-bit ADC independent of the array size and bits per cell. Therefore, both parameters can be swept without affecting accuracy. A maximum array size of 1152 rows is assumed to control the parasitic voltage drops, based on the results in Section 8 for ResNet50-v1.5.

Offset-subtraction cores have a more limited design space. Fig. 17 shows that due to 8-bit ADC quantization alone, offset subtraction can reach high accuracy only with a small array (≤ 144 rows) and finer bit slices (≤ 2 bits/cell). Sections 5 and 8 showed that they are also more sensitive to cell errors and parasitic resistance, which might further reduce the array size. Additionally, the absence of proportional mapping requires larger and more power-hungry peripheral circuits that can support larger bit line currents, as shown in Fig. 20. For these reasons, only four design points are evaluated for offset subtraction, both using digital input bit accumulation: 72 and 144 rows with 1 and 2 bits per cell, which are close to the design point in ISAAC [56]. One of these designs is shown in Fig. 23(b) and requires multiple digital steps to aggregate partial results produced by the analog hardware.

All energy and area estimates are based on SONOS arrays and peripheral circuits that are designed and simulated in an embedded 40nm process compatible with SONOS memory [5, 37]. The energy consumption of the array and row drivers is based on the average cell conductances in Fig. 7 and

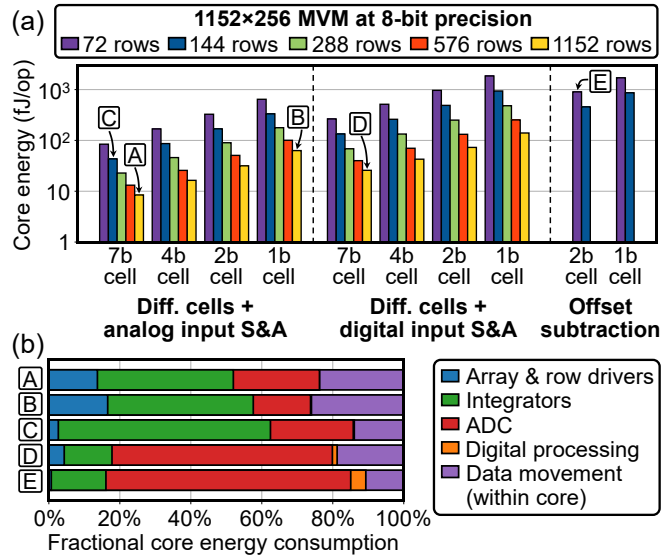


Figure 24: (a) Core energy per operation (1 MAC = 2 operations) for various core configurations applied to an MVM of size 1152×256 . (b) Breakdown of energy use among core components for selected configurations.

the average activity factors for each input bit when running ResNet50-v1.5 on ImageNet. The core uses a current conveyor that integrates each input bit for 10 ns [45], a switched-capacitor circuit for analog input bit accumulation [8], and a power- and area-efficient 8-bit ramp ADC clocked at 1 GHz [45]. ADC range calibration is implemented with a tunable operational-amplifier gain stage after the integrator. Digital component energies are derived from a standard cell library. Since all array outputs are simultaneously available with a ramp ADC, as many S&A units are allocated as needed to process these results in parallel. Area is estimated from the sum of circuit block areas rather than a physical layout.

9.3 Energy and Area Evaluation

Fig. 24(a) shows the energy efficiency of various core configurations. To estimate the peak efficiency, a 1152×256 weight matrix is evaluated that utilizes every cell in each array. Table 3 details the area and energy efficiency of five labeled configurations in Fig. 24(a), whose energy breakdown among core components is shown in Fig. 24(b). These results reveal several trends:

(1) Unsliced weights are more efficient than bit slicing because the bit line peripheral circuit costs increase roughly linearly with the number of slices. The area increases linearly with bit slicing due to having more cells per weight.

(2) Larger arrays are more efficient since the integrator and ADC energies are amortized over more operations, and less computation is done in the less efficient digital domain. Density also improves since these circuits are shared by more matrix elements. While smaller arrays can offer better area utilization when mapping small matrices [49, 65], large arrays are necessary to extract the efficiency benefits of analog processing. Neural networks that more fully utilize large arrays will have superior system-level energy efficiency when deployed in an analog accelerator.

Table 3: Efficiency of selected core configurations

Design	A	B	C	D	E
Negative values	Diff.	Diff.	Diff.	Diff.	Offset
Weight resolution	8	9	8	8	8
Bits / cell	7	1	7	7	2
# rows	1152	1152	144	1152	72
Input bit S&A [†]	A	A	A	D	D
ADC resolution	8	8	8	8	8
Ideal analog resolution B_{out}	26.2	20.2	23.2	18.2	8.2
Core area (mm ²)	0.24	2.02	1.30	0.27	11.14
Core energy (fJ/op)	8.4	63.1	43.3	25.8	902.0

[†]A = analog, D = digital.

Table 4: ResNet50-v1.5 accuracy with SONOS errors using selected core designs (1000 images, 10 runs each)

Design	A	B	C	D	E
Ideal cells	76.3%	75.3%	75.4%	76.3%	74.9%
SONOS	74.0% ±1.0%	75.4% ±0.3%	73.6% ±0.7%	74.1% ±1.0%	50.2% ±5.3%

Table 5: Accuracy of the analog inference accelerator on the full ImageNet test set (50,000 images, 10 runs each)

ResNet50-v1.5	Floating-point	4-bit, QAT
Fully digital	76.466%	76.154%
Design A, ideal cells	76.082%	76.038%
Design A, SONOS	74.296% ±0.348%	75.294% ±0.192%

(3) Analog input bit accumulation yields a 2-4 \times energy improvement. The technique increases integrator energy, but reduces the number of ADC conversions by 8 \times . When each input bit requires a digitization step, the ADC dominates the energy cost, as shown in Fig. 24(b) for designs D and E; this is consistent with prior work [11, 56].

For the reasons summarized in Section 9.2, systems that rely on offset subtraction cannot exploit any of the above techniques to reduce energy. Design A is the most efficient design for differential cells, while Design E is an offset-subtraction design that very nearly satisfies the FPGA (a pre-requisite for high accuracy using offset subtraction, as explained in Section 6.3). Design E has 107 \times higher energy consumption and 46 \times larger area. The higher energy comes from having 4 \times as many bit slices, 8 \times as many ADC conversions per input value, and 16 \times as many arrays to map a large matrix. The actual ratio of energy consumption is smaller than the product of these factors since only part of the total energy scales with these factors.

9.4 Accuracy Evaluation

Table 4 compares the ImageNet accuracy with ResNet50-v1.5 obtained using the same five design points. The simulations include 8-bit weight and activation quantization, 8-bit ADCs calibrated separately for each design, and random SONOS programming errors following the full state dependence in Fig. 22(b), sampled ten times as described in Section

5.1. The small differences in the baseline accuracy using ideal cells result from the varying effectiveness of the 8-bit ADC calibration across designs.

To keep the computations tractable, parasitic resistance was not included. Relative to the SONOS cells, the metal interconnects in the 40nm process have a normalized resistance of $\hat{R}_p \approx 10^{-5}$. Fig. 21 shows that this resistance has negligible effect on the accuracy of Designs A, C, and D, which use differential cells and unsliced weights. For the other designs, the accuracies in Table 4 are best-case estimates; with a realistic parasitic resistance, the accuracy of Design B may be slightly lower, and that of Design E is likely to be much lower.

The designs with differential cells and unsliced weights (A, C, and D) all have similar accuracies, losing roughly 2% on ImageNet by using SONOS cells. Design B, which uses 1-bit slices, is less sensitive to SONOS errors than unsliced weights. This result is consistent with Fig. 10(b), and is due to the fact that finer bit slicing creates greater sparsity in the most significant slice, as explained in Section 5.3. However, this design requires nearly 8 \times larger energy and area than Design A. Whether this small difference in accuracy is worth the considerable overhead is dependent on the end-to-end application requirements.

Design E loses more than 20% in accuracy from SONOS cell errors. This design is the least robust because it uses offset subtraction, which lacks weight proportionality and thus does not exploit the state-proportional error property of the cells. At the average cell current used by this system ($0.5I_{max}$, or 0.8 μ A), the SONOS error properties are intermediate between state-independent error and state-proportional error (with $\alpha_{ind} \approx \alpha_{prop} \approx 4\%$). Notably, however, the accuracy of this design is much higher than that predicted in Fig. 9(a) or Fig. 10(a) for offset subtraction with 2 bits/cell. This is due to the ADCs, which cut off the analog accumulation of cell errors beyond 72 rows: this effect is depicted in Fig. 18(b). As noted earlier, the true accuracy of Design E is likely much lower than listed in Table 4 due to parasitic resistance.

Table 5 shows the accuracy of the most efficient design, Design A, on the full ImageNet test set of 50,000 images. The 2.17% accuracy loss on ResNet50-v1.5 is relatively small for a system that uses direct weight transfer. By comparison, the PCM devices in Joshi *et al.* lose 7.8% ImageNet accuracy using ResNet34 [34]. The main accuracy advantage of the SONOS device over PCM is state-proportional error, as explained in Section 9.1. Table 5 further shows that by using a standard quantization-aware training scheme with 4-bit activations, the propagation of this error can be significantly suppressed as described in Section 7. This reduces the accuracy loss induced by the SONOS device to only 0.86%.

10. CONCLUSIONS

Error resilience can be built into analog accelerators by designing the system to leverage the properties of the application neural network. A proportional mapping of numerical values in the algorithm to physical quantities in the accelerator exploits a feature common to many networks: a weight distribution that is skewed toward low values. This paper showed that a proportional mapping reduces sensitivity to several categories of analog errors. The critical building

blocks of a proportional system are differential cells for mapping signed weights, a memory technology with high On/Off ratio, and programming errors that scale with conductance.

This paper also evaluated the popular design choices made by prior analog accelerators from the perspective of accuracy and robustness to errors. Bit slicing has only a small accuracy benefit, which is unlikely to outweigh the considerable energy and area overhead needed to support it. The full-precision guarantee is also too conservative a choice for neural network inference, and leads to smaller arrays or greater ADC overheads than needed. Proportional systems can perform a much larger share of the computation in analog, and allow the algorithm to dictate the precision with which the analog outputs are digitized.

In analog systems, where algorithmic accuracy depends on device-level effects, hardware design should ultimately be guided by a rigorous evaluation of the *end-to-end* accuracy. While the evaluation of a design choice on the basis of intermediate results (such as MVM-level precision) can yield valuable insights, an end-to-end accuracy evaluation is needed to avoid unnecessary bottlenecks for accuracy and efficiency. An end-to-end design approach results in a harmonization of the hardware and the algorithm that ultimately delivers the order-of-magnitude energy efficiency benefits promised by analog accelerators.

Acknowledgments

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

- [1] "MLPerf Inference Benchmark Suite," <https://github.com/mlcommons/inference>, 2019.
- [2] "MLPerf Inference ImageNet calibration set," https://github.com/mlcommons/inference/blob/master/calibration/ImageNet/cal_image_list_option_1.txt, 2019.
- [3] "MLPerf Inference v0.5 NVIDIA-optimized implementations for Open Division," https://github.com/mlperf/inference_results_v0.5/tree/master/open/NVIDIA, 2019.
- [4] S. Agarwal, S. J. Plimpton, R. K. Schiek, I. Richter, A. H. Hsia, D. R. Hughart, R. B. Jacobs-Gedrim, C. D. James, and M. J. Marinella. (2017) CrossSim. [Online]. Available: <https://cross-sim.sandia.gov/>
- [5] V. Agrawal, V. Prabhakar, K. Ramkumar, L. Hinh, S. Saha, S. Samanta, and R. Kapre, "In-memory computing array using 40nm multibit SONOS achieving 100 TOPS/W energy efficiency for deep neural network edge inference accelerators," in *Intl. Memory Workshop (IMW)*, May 2020.
- [6] A. Ankit, I. E. Hajj, S. R. Chalamalasetti, G. Ndu, M. Foltin, R. S. Williams, P. Faraboschi, W.-m. W. Hwu, J. P. Strachan, K. Roy, and D. S. Milojicic, "PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Intl. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, April 2019, p. 715–731.
- [7] M. Bavandpour, M. R. Mahmoodi, and D. B. Strukov, "Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond," *IEEE Trans. on Circuits and Systems II: Express Briefs (TCAS-II)*, vol. 66, no. 9, pp. 1512–1516, 2019.
- [8] M. Bavandpour, S. Sahay, M. R. Mahmoodi, and D. Strukov, "Efficient mixed-signal neurocomputing via successive integration and rescaling," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 3, pp. 823–827, 2020.
- [9] M. N. Bojnordi and E. Ipek, "Memristive Boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning," in *Intl. Symp. on High Performance Computer Architecture (HPCA)*, March 2016.
- [10] W.-H. Chen, K.-X. Li, W.-Y. Lin, K.-H. Hsu, P.-Y. Li, C.-H. Yang, C.-X. Xue, E.-Y. Yang, Y.-K. Chen, Y.-S. Chang, T.-H. Hsu, Y.-C. King, C.-J. Lin, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors," in *IEEE Intl. Solid-State Circuits Conf. (ISSCC)*, 2018, pp. 494–496.
- [11] P. Chi, S. Li, S. Li, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *Intl. Symp. on Computer Architecture (ISCA)*, June 2016.
- [12] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "PACT: Parameterized clipping activation for quantized neural networks," *arXiv preprint arXiv:1805.06085*, 2018.
- [13] F. Chollet *et al.*, "Keras," <https://keras.io>, 2015.
- [14] 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 *Intl. Symp. on Microarchitecture (MICRO)*, Oct. 2019, p. 114–125.
- [15] A. Coates, B. Huval, T. Wang, D. J. Wu, A. Y. Ng, and B. Catanzaro, "Deep learning with COTS HPC systems," in *Intl. Conf. on Machine Learning (ICML)*, 2013, pp. 1337–1345.
- [16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2009.
- [17] Q. Dong, M. E. Sinangil, B. Erbagci, D. Sun, W. Khwa, H. Liao, Y. Wang, and J. Chang, "A 351TOPS/w and 372.4GOPS compute-in-memory SRAM macro in 7nm FinFET CMOS for machine-learning applications," in *Intl. Solid-State Circuits Conf. (ISSCC)*, Feb. 2020, pp. 242–244.
- [18] B. Feinberg, U. K. R. Vengalam, N. Whitehair, S. Wang, and E. Ipek, "Enabling scientific computing on memristive accelerators," in *Intl. Symp. on Computer Architecture (ISCA)*, 2018, pp. 367–382.
- [19] B. Feinberg, S. Wang, and E. Ipek, "Making memristive neural network accelerators reliable," in *Intl. Symp. on High Performance Computer Architecture (HPCA)*, Feb. 2018, pp. 52–65.
- [20] L. Fick, D. Blaauw, D. Sylvester, S. Skrzyniarz, M. Parikh, and D. Fick, "Analog in-memory subthreshold deep neural network accelerator," in *Custom Integrated Circuits Conf. (CICC)*, May 2017, pp. 1–4.
- [21] R. Genov and G. Cauwenberghs, "Charge-mode parallel architecture for matrix-vector multiplication," in *Midwest Symp. on Circuits and Systems (MWSCAS)*, Aug. 2000.
- [22] S. Ghodrati, H. Sharma, S. Kinzer, A. Yazdanbakhsh, J. Park, N. S. Kim, D. Burger, and H. Esmailzadeh, "Mixed-signal charge-domain acceleration of deep neural networks through interleaved bit-partitioned arithmetic," in *Intl. Conf. on Parallel Architectures and Compilation Techniques (PACT)*, 2020, p. 399–411.
- [23] S. K. Gonugondla, C. Sakr, H. Dbouk, and N. R. Shanbhag, "Fundamental limits on the precision of in-memory architectures," in *Intl. Conf. on Computer Aided Design (ICCAD)*, 2020, pp. 1–9.
- [24] X. Guo, F. M. Bayat, M. Bavandpour, M. Klachko, M. R. Mahmoodi, M. Prezioso, K. K. Likharev, and D. B. Strukov, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *Intl. Electron Devices Meeting (IEDM)*, Dec. 2017, pp. 6.5.1–6.5.4.
- [25] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient inference engine on compressed deep neural network," in *Intl. Symp. on Computer Architecture (ISCA)*, June 2016, p. 243–254.

- [26] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
- [27] Y. Hayakawa, A. Himeno, R. Yasuhara, W. Boullart, E. Vecchio, T. Vandeweyer, T. Witters, D. Crotti, M. Jurczak, S. Fujii, S. Ito, Y. Kawashima, Y. Ikeda, A. Kawahara, K. Kawai, Z. Wei, S. Muraoka, K. Shimakawa, T. Mikawa, and S. Yoneda, "Highly reliable TaOx ReRAM with centralized filament for 28-nm embedded application," in *Symp. on VLSI Technology*, June 2015, pp. T14–T15.
- [28] Z. He, J. Lin, R. Ewetz, J.-S. Yuan, and D. Fan, "Noise injection adaption: End-to-end ReRAM crossbar non-ideal effect adaption for neural network mapping," in *Design Automation Conf. (DAC)*, June 2019, pp. 57:1–57:6.
- [29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [30] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams, "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Design Automation Conf. (DAC)*, June 2016, pp. 1–6.
- [31] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 2704–2713.
- [32] S. Jain and A. Raghunathan, "Cx2DNN: Hardware-software compensation methods for deep neural networks on resistive crossbar systems," *ACM Trans. Embed. Comput. Syst.*, vol. 18, no. 6, Nov. 2019.
- [33] Y. Jeong, M. A. Zidan, and W. D. Lu, "Parasitic effect analysis in memristor-array-based neuromorphic systems," *IEEE Transactions on Nanotechnology (TNANO)*, vol. 17, no. 1, pp. 184–193, Jan 2018.
- [34] V. Joshi, M. L. Gallo, S. Haefeli, I. Boybat, S. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, "Accurate deep neural network inference using computational phase-change memory," *Nature Communications*, vol. 11, 2020.
- [35] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-I. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snellman, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, "In-datacenter performance analysis of a tensor processing unit," in *Intl. Symp. on Computer Architecture (ISCA)*, June 2017, p. 1–12.
- [36] M. Klachko, M. R. Mahmoodi, and D. Strukov, "Improving noise tolerance of mixed-signal neural networks," in *Intl. Joint Conf. on Neural Networks (IJCNN)*, 2019, pp. 1–8.
- [37] I. Kouznetsov, K. Ramkumar, V. Prabhakar, L. Hinh, H. Shih, S. Saha, S. Govindaswamy, M. Amundson, D. Dalton, T. Phan *et al.*, "40 nm ultralow-power charge-trap embedded NVM technology for IoT applications," in *Intl. Memory Workshop (IMW)*, 2018, pp. 1–4.
- [38] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [39] Y. LeCun, "The MNIST database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>.
- [40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [41] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Analogue signal and image processing with large memristor crossbars," *Nature Electronics*, vol. 1, no. 1, pp. 52–59, 2018.
- [42] W. Li, P. Xu, Y. Zhao, H. Li, Y. Xie, and Y. Lin, "TIMELY: Pushing data movements and interfaces in PIM accelerators towards local and in time domain," in *Intl. Symp. on Computer Architecture (ISCA)*, June 2020, p. 832–845.
- [43] M.-Y. Lin, H.-Y. Cheng, W.-T. Lin, T.-H. Yang, I.-C. Tseng, C.-L. Yang, H.-W. Hu, H.-S. Chang, H.-P. Li, and M.-F. Chang, "DL-RSIM: A simulation framework to enable reliable ReRAM-based accelerators for deep learning," in *Intl. Conf. on Computer-Aided Design (ICCAD)*, 2018, pp. 1–8.
- [44] Y. Long, X. She, and S. Mukhopadhyay, "Design of reliable DNN accelerator with un-reliable ReRAM," in *Design, Automation Test in Europe Conf. Exhibition (DATE)*, 2019, pp. 1769–1774.
- [45] M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, and C. D. James, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 8, no. 1, pp. 86–101, 2018.
- [46] V. Milo, F. Anzalone, C. Zambelli, E. Pérez, M. K. Mahadevaiah, S. G. Ossorio, P. Olivo, C. Wenger, and D. Ielmini, "Optimized programming algorithms for multilevel rram in hardware neural networks," in *Intl. Reliability Physics Symp. (IRPS)*, 2021, pp. 1–6.
- [47] D. Miyashita, E. H. Lee, and B. Murmann, "Convolutional neural networks using logarithmic data representation," *arXiv preprint arXiv:1603.01025*, 2016.
- [48] B. Murmann, "ADC performance survey 1997-2020," <http://web.stanford.edu/~murmman/adcsurvey.html>, 2020.
- [49] A. Nag, R. Balasubramonian, V. Srikumar, R. Walker, A. Shafiee, J. P. Strachan, and N. Muralimanohar, "Newton: Gravitating towards the physical limits of crossbar acceleration," *Intl. Symp. on Microarchitecture (MICRO)*, vol. 38, no. 5, pp. 41–49, 2018.
- [50] P. Narayanan, G. W. Burr, R. S. Shenoy, S. Stephens, K. Virwani, A. Padilla, B. N. Kurdi, and K. Gopalakrishnan, "Exploring the design space for crossbar arrays built with mixed-ionic-electronic-conduction (MIEC) access devices," *IEEE Journal of the Electron Devices Society (J-EDS)*, vol. 3, no. 5, pp. 423–434, 2015.
- [51] V. J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, R. Chukka, C. Coleman, S. Davis, P. Deng, G. Damos, J. Duke, D. Fick, J. S. Gardner, I. Hubara, S. Idgunji, T. B. Jablin, J. Jiao, T. S. John, P. Kanwar, D. Lee, J. Liao, A. Lokhmotov, F. Massa, P. Meng, P. Micikevicius, C. Osborne, G. Pekhimenko, A. T. R. Rajan, D. Sequeira, A. Sirasao, F. Sun, H. Tang, M. Thomson, F. Wei, E. Wu, L. Xu, K. Yamada, B. Yu, G. Yuan, A. Zhong, P. Zhang, and Y. Zhou, "MLPerf Inference Benchmark," in *Intl. Symp. on Computer Architecture (ISCA)*, June 2020.
- [52] A. S. Rekhi, B. Zimmer, N. Nedovic, N. Liu, R. Venkatesan, M. Wang, B. Khailany, W. J. Dally, and C. T. Gray, "Analog/mixed-signal hardware error modeling for deep learning inference," in *Design Automation Conf. (DAC)*, 2019.
- [53] J. H. Saltzer, D. P. Reed, and D. D. Clark, "End-to-end arguments in system design," *ACM Transactions on Computer Systems (TOCS)*, vol. 2, no. 4, pp. 277–288, 1984.
- [54] C. R. Schlottmann and P. E. Hasler, "A highly dense, low power, programmable analog vector-matrix multiplier: The FPAA implementation," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 1, no. 3, pp. 403–411, 2011.
- [55] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnology*, vol. 15, no. 7, pp. 529–544, 2020.
- [56] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *Intl. Symp. on Computer Architecture (ISCA)*, June 2016.
- [57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [58] L. Song, Y. Zhuo, X. Qian, H. Li, and Y. Chen, "GraphR: Accelerating graph processing using ReRAM," in *Intl. Symp. on High Performance Computer Architecture (HPCA)*, 2018, pp. 531–543.
- [59] X. Sun, N. Wang, C.-Y. Chen, J. Ni, A. Agrawal, X. Cui, S. Venkataramani, K. El Maghraoui, V. V. Srinivasan, and K. Gopalakrishnan, "Ultra-low precision 4-bit training of deep neural networks," *Advances in Neural Information Processing Systems*

- (*NeurIPS*), vol. 33, 2020.
- [60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.
- [61] H. Tsai, S. Ambrogio, P. Narayanan, R. M. Shelby, and G. W. Burr, "Recent progress in analog memory-based accelerators for deep learning," *Journal of Physics D: Applied Physics*, vol. 51, no. 28, p. 283001, jun 2018.
- [62] T. P. Xiao, C. H. Bennett, B. Feinberg, S. Agarwal, and M. J. Marinella, "Analog architectures for neural network acceleration based on non-volatile memory," *Applied Physics Reviews*, vol. 7, no. 3, p. 031301, 2020.
- [63] T. P. Xiao, B. Feinberg, C. H. Bennett, V. Agrawal, P. Saxena, V. Prabhakar, K. Ramkumar, H. Medu, V. Raghavan, R. Chettuvetty, S. Agarwal, and M. J. Marinella, "An accurate, error-tolerant, and energy-efficient neural network inference engine based on SONOS analog memory," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–14, 2022.
- [64] X. Xu, Y. Ding, S. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi, "Scaling for edge inference of deep neural networks," *Nature Electronics*, vol. 1, pp. 216–222, 2018.
- [65] T. Yang and V. Sze, "Design considerations for efficient deep neural networks on processing-in-memory accelerators," in *Intl. Electron Devices Meeting (IEDM)*, 2019, pp. 22.1.1–22.1.4.
- [66] T.-H. Yang, H.-Y. Cheng, C.-L. Yang, I.-C. Tseng, H.-W. Hu, H.-S. Chang, and H.-P. Li, "Sparse ReRAM engine: Joint exploration of activation and weight sparsity in compressed neural networks," in *Intl. Symp. on Computer Architecture (ISCA)*, 2019, p. 236–249.
- [67] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, 2020.
- [68] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-memory chips for deep learning: Recent trends and prospects," *IEEE Circuits and Systems Magazine*, vol. 21, no. 3, pp. 31–56, 2021.
- [69] G. Yuan, P. Behnam, Z. Li, A. Shafiee, S. Lin, X. Ma, H. Liu, X. Qian, M. N. Bojnordi, Y. Wang, and C. Ding, "FORMS: Fine-grained polarized reram-based in-situ computation for mixed-signal DNN accelerator," in *Intl. Symp. on Computer Architecture (ISCA)*, 2021, pp. 265–278.
- [70] D. Zhang, J. Yang, D. Ye, and G. Hua, "LQ-Nets: Learned quantization for highly accurate and compact deep neural networks," in *European Conf. on Computer Vision (ECCV)*, 2018, pp. 365–382.
- [71] F. Zhang and M. Hu, "Mitigate parasitic resistance in resistive crossbar-based convolutional neural networks," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 16, no. 3, May 2020.