Signal Processing Methods to Enhance the Energy Efficiency of In-Memory Computing Architectures

Charbel Sakr, *Member, IEEE*, and Naresh R. Shanbhag, *Fellow, IEEE*

***Abstract*—This paper presents signal processing methods to enhance the energy vs. accuracy trade-off of in-memory computing (IMC) architectures. First, an optimal clipping criterion (OCC) for signal quantization is proposed in order to minimize the precision of column analog-to-digital converters (ADCs) at iso-accuracy. For a Gaussian distributed signal, the OCC is shown to reduce the column ADC precision requirements by 3 bits at a signal-to-quantization noise ratio (SQNR) of 22.5 dB over the commonly used full range (FR) quantizer. Next, the input-sliced weight-parallel (ISWP) IMC architecture is presented as a generalization of the popular bit-serial bit-parallel (BSBP) architecture. Quantization noise analysis of the ISWP indicates that its accuracy is comparable to BSBP while providing an order-of-magnitude reduction in energy consumption due to fewer array invocations and smaller ADC precision. Combining OCC and ISWP noise analysis, we map popular DNNs such as VGG-9 (CIFAR-10), ResNet-18 (CIFAR-10), and AlexNet (ImageNet) on a OCC-enabled ISWP architecture and show a reduction in energy consumption by an order-of-magnitude at iso-accuracy over the BSBP architecture that employs FR-based ADCs.**

***Index Terms*—Optimal clipping, quantization, bit slicing, in-memory computing.**

I. Introduction

DEEP neural networks (DNNs) are among most powerful predictive models in many applications such as image [1], [2], speech [3], [4], and language [5], [6] processing. However, their high computational complexity hinders their deployment onto resource-limited devices [7]–[9]. Accentuating the difficulty of DNN deployment is the use of classical von Neumann compute architectures which suffer from the *memory wall* problem whereby the energy and latency costs are dominated by memory access [10].

The in-memory computing (IMC) architecture [11]–[13] strives to eliminate the separation between storage and compute. It does so by realizing functional operations such as dot-products within the bitcell array (BCA) during memory reads. In the process, the energy-delay product (EDP) of inference tasks can be reduced by up to two orders-of-magnitude compared to an equivalent von Neumann architecture [14]. Since IMCs address the memory wall problem, it is particularly attractive for memory-centric workloads such as machine learning algorithms. In recent years, a large number of IMC implementations of DNNs have been proposed [14]–[26].

However, in spite of these advances, the computational precision of IMCs is limited. This is because: 1) IMC computations have been restricted to simple binary operations [27]–[30] in order to adhere to the binary storage formats in memory; 2) mapping of high-dimensional dot-products onto IMCs is often limited by analog noise sources, which are not yet fully understood or characterized [31], [32]; and 3) the dense BCA layout imposes strict area constraints on the column analog-to-digital converters (ADCs) and hence the realizable ADC precision [14]. Today, the IMC precision is limited by the achievable precision of the column ADCs and methods to increase IMC precision remain elusive. Even if ADC precision were to be increased somehow, the impact on the system level energy and latency would be severe [32]. Unfortunately, meeting application-level accuracy requirements with such precision constraints on ADCs is challenging.

Efforts to address some of the above mentioned limitations have relied on ad-hoc trial-and-error methods. The lack of an analytical framework to guide the design of IMCs has led to designs that are overly conservative and therefore sub-optimal in terms of efficiency. For example, the use of the bit-growth criterion (BGC) to set ADC precision [34] avoids loss in fidelity of bitline computations in the BCA but results in much higher precision than necessary. Some IMC designs employ fewer ADC bits than suggested by the BGC and justify it via extensive simulations to ensure that the DNN accuracy is preserved. However, such methods do not provide any guarantees.

Our work addresses the above mentioned precision limits of IMCs by employing quantization noise analysis commonly employed in the design of digital signal processing systems [35]. Specifically, we make the following contributions:

- We propose the Optimal Clipping Criterion (OCC) to minimize the column ADC precision requirements. The signal-to-quantization ratio (SQNR) of OCC is shown to be within 0.8 dB of the well-known optimal Lloyd-Max (LM) quantizer [36]. OCC improves the SQNR by 14 dB over the commonly employed full range (FR) quantizer, which translates to a 3-bit reduction in the ADC precision.

Manuscript received April 27, 2021; revised August 31, 2021 and October 12, 2021; accepted November 15, 2021. Date of publication November 24, 2021; date of current version December 17, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tokunbo Ogunfunmi. This work was supported in part by the Semiconductor Research Corporation and the Defense Advanced Research Projects Agency (DARPA) via the Center for Brain-Inspired Computing and DARPA's Foundations of Novel Compute program. *(Corresponding author: Charbel Sakr.)*

Charbel Sakr was with the University of Illinois at Urbana-Champaign, Champaign, IL 61801 USA. He is now with NVIDIA Research, NVIDIA, Santa Clara, CA 95051 USA. (e-mail: csakr@nvidia.com).

Naresh R. Shanbhag is with the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, Champaign, IL 61801 USA (e-mail: shanbhag@illinois.edu).

Digital Object Identifier 10.1109/TSP.2021.3130488

- We study the quantization noise in a input-serial weight-parallel (ISWP) IMC which generalizes the popular bit-serial weight-parallel IMC of [21]. We show that, using bit slicing techniques, significant energy savings can be achieved with minimal loss in accuracy. Specifically, we prove that an multi-bit IMC dot-product can be computed within a single memory access while suffering no more than 2 dB SQNR drop.
- We apply our analysis on OCC and ISWP to DNN implementation using IMC. We consider mapping of the VGG-9, ResNet-18 and AlexNet networks and contrast our method to common practices in IMC. We show that ADC precision can be lowered by 2-to-3 bits and energy consumption can be reduced by an order of magnitude while maintaining accuracy.

This paper is organized as follows: The problem setup with the corresponding IMC model and architecture is introduced in Section II-A. The OCC method for minimizing column ADC precision is presented in Section III. An analysis of the ISWP architecture is described in Section IV. Numerical results for DNN mapping onto IMC are presented in Section V. Finally, Section VI summarizes and concludes this paper.

II. Problem Setup

Consider an N-dimensional dot-product $y = \mathbf{w}^{\mathsf{T}}\mathbf{x}$ of real valued (signed) weight and (unsigned) input vectors of precision B_W and B_X bits, respectively, given by:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix}; \quad w_i = w_m \left(-w_{i,0} + \sum_{b=1}^{B_W-1} w_{i,b} 2^{-b} \right) \tag{1}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}; \quad x_i = x_m \sum_{b=0}^{B_X-1} x_{i,b} 2^{-b-1}, \tag{2}$$

where $w_{i,b} \in \{0,1\}$ and $x_{i,b} \in \{0,1\}$ are the b^{th} bits of $w_i \in [-w_m, w_m]$ and $x_i \in [0, x_m]$, respectively. The choice of unsigned inputs is to account for the use of activations (e.g., ReLU) in DNNs.

A. The Input-Serial Weight-Parallel (ISWP) IMC

We consider the input-serial weight-parallel (ISWP) architecture (see Fig. 1) [31] which generalizes the architecture [21] by allowing for multi-bit inputs per read cycle.

The ISWP architecture stores $\mathbf{w}$ in the columns of the BCA where the bits of w_i are arrayed across B_W columns in the i^{th} row. When computing a dot-product, ISWP serializes the B_X-bit input vector $\mathbf{x}$ into $N_S = \lceil \frac{B_X}{B_S} \rceil$ *input slices* of B_S bits each, where the i^{th} element x_i of $\mathbf{x}$ is given by

$$x_i = x_m \sum_{s=0}^{N_S-1} x_{i,s}^{(B_S)} 2^{-sB_s} \tag{3}$$

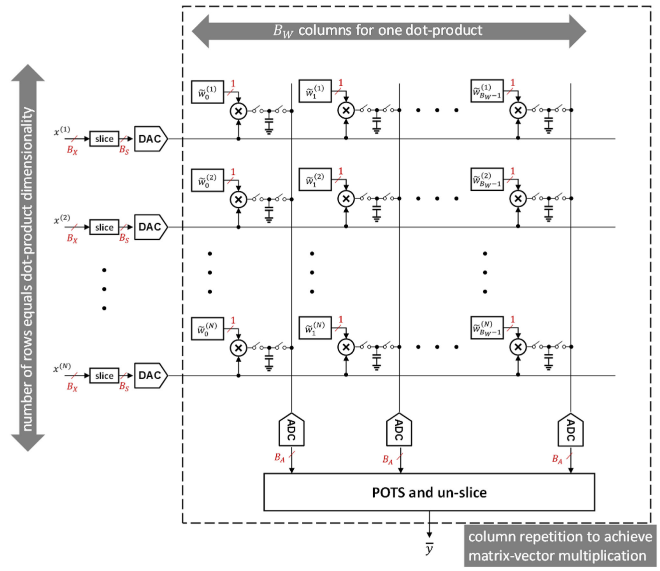


Fig. 1. The input-serial weight-parallel (ISWP) architecture.

with $x_{i,s}^{(B_S)}$ being the s^{th} slice as shown below:

$$x_{i,s}^{(B_S)} = \sum_{b=0}^{B_S-1} x_{i,sB_S+b}^{(1)} 2^{-b-1}. \tag{4}$$

For example, using a bit vector representation, if $x_i = [x_{i,0}, x_{i,1}, x_{i,2}, x_{i,3}]$ is a 4-bit scalar then we can split it into two bit slices $x_{i,0}^{(2)} = [x_{i,0}, x_{i,1}]$ and $x_{i,1}^{(2)} = [x_{i,2}, x_{i,3}]$ with $B_S = 2$ and $N_S = 2$.

Processing the inputs one slice per read cycle, the multi-bit dot-product $y = \mathbf{w}^T\mathbf{x}$ is realized using the following powers-of-two summing (POTS):

$$y = x_m w_m \sum_{s=0}^{N_S-1} \left(-y_{s,0} + \sum_{b=1}^{B_W-1} y_{s,b} 2^{-b} \right) 2^{-sB_S}, \tag{5}$$

where the bitline (BL) dot-product $y_{s,b}$ is computed as:

$$y_{s,b} = \sum_{i=1}^{N} w_{i,b}^{(1)} x_{i,s}^{(B_S)}. \tag{6}$$

on the b^{th} BL.

Thus, the ISWP architecture computes an N-dimensional dot-product between a B_S-bit input and a binary weight and is a generalization of the bit-serial bit-parallel (BSBP) architecture in [21] which computes a fully binarized N-dimensional dot-product, i.e., $B_S = 1$.

B. Quantization and Analog Noise Effects

The BL dot-product $y_{s,b}$ in (6) is computed in the analog domain. Due to noise, the observed BL dot-product $\overline{y}_{s,b}$ is given by:

$$\overline{y}_{s,b} = y_{s,b} + q_{A_{s,b}} + \eta_{a_{s,b}}, \tag{7}$$

where $q_{A_{s,b}}$ and $\eta_{a_{s,b}}$ are the column ADC quantization noise and the analog noise on the b^{th} BL, respectively. An expression

TABLE I
VALUES OF ANALOG NOISE PARAMETERS IN A 65 NM PROCESS

Parameter	Value	Parameter	Value
ρ_1	6.40×10^{-18} F	ρ_3	6.01×10^{-33} F^2
ρ_2	4.14×10^{-21} F	V_{DD}	1 V

for the variance of $q_{A_{s,b}}$ will be presented in Section III since it depends on the quantization method employed in the ADC.

The analog noise term $\eta_{a_{s,b}}$ includes effects from capacitor mismatch, thermal effects, and charge injection. These are fundamental noise sources residing in the core of the ISWP architecture and are hard to overcome via circuit design methods due to the tight area constraints. The variance of the analog noise term $\eta_{a_{s,b}}$ is given by [31]:

$$\sigma^2_{\eta_{a_{s,b}}} = N\left(\frac{\mathbb{E}\left[\left(w^{(1)}x^{(B_S)}\right)^2\right]\rho_1}{(1-2^{-B_S})^2C_o} + \frac{\rho_2}{C_o} + \frac{\rho_3}{C_o^2}\right), \quad (8)$$

where $w^{(1)}$ and $x^{(B_S)}$ are the unindexed weight bits and input slices, respectively, C_o is the nominal extrinsic bitcell (BC) capacitance, and ρ_1, ρ_2, and ρ_3 are technology and layout dependent parameters. For a 65 nm process [31], the values of these parameters are listed in Table I. The capacitance C_o is an extrinsic metal-on-metal (MOM) capacitance that is not a part of a standard SRAM bitcell [21]. This capacitance allows for summing across bitcells within a column in a highly linear fashion. Being an extrinsic capacitance, its value can be assigned independent of the SRAM bitcell. As seen in (8), the noise variance decreases when C_o increases. However, increasing C_o causes higher CV^2 energy consumption and also reduces storage density. Thus, like all IMCs [12], the ISWP architecture exhibits a fundamental trade-off between its energy efficiency and computational accuracy.

The impact of the noise sources in (7) on the accuracy of the dot-product in (5) will be derived in Section IV.

C. Energy Consumption

An IMC's energy efficiency is quantified by the energy per 1-bit multiply-accumulate (MAC) operation E_{OP} [32]:

$$E_{\text{OP}} = \frac{N_S}{B_X}\left(E_{\text{BC}} + \frac{E_{\text{ADC}}}{N}\right), \quad (9)$$

where E_{BC} is the energy consumed by the 1-bit MAC within the bitcell given by

$$E_{\text{BC}} = \mathbb{E}\left[x^{(B_S)}\right]C_oV_{DD}^2 \quad (10)$$

where $\mathbb{E}[x^{(B_S)}]$ is the mean value of an input slice, equal to 0.5, nominally, and V_{DD} is the supply voltage. Equation (9) indicates that the IMC's energy efficiency improves when B_S increases since N_S, number of array accesses, reduces.

For a B_A-bit ADC, E_{ADC} is given by [37], [38]:

$$E_{\text{ADC}} = k_1\left(B_A + \log_2\left(\frac{y_m}{Y_{\text{ADC}}}\right)\right) + k_2\left(\frac{y_m}{Y_{\text{ADC}}}\right)^2 4^{B_A}, \quad (11)$$

where $k_1 = 10^{-13}\,J$ and $k_2 = 10^{-18}\,J$ are fitting parameters, y_m is the maximum value of the BL dot-product $y_{s,b}$, and Y_{ADC} is the ADC input range. These energy models are based on real-world data obtained by curve fitting to silicon measurements of over 700 silicon ADC designs spanning the years 1997-2021 and across various technology nodes from 0.5 um to 16 nm [32], [37], [38]. Equation (11) also indicates that the ADC energy quadruples per bit of increase in its precision B_A, emphasizing the need for minimizing it without impacting accuracy. The next section presents a method to realize this objective.

III. THE OPTIMAL CLIPPING CRITERION

Quantization of a signal $x \in [x_{\min}, x_{\max}]$ to B bits is the process of mapping its value to one of 2^B pre-defined levels $\{r_i\}_{i=1}^{2^B}$. The quantized signal is obtained as:

$$x_q = \arg\min_{\{r_i\}_{i=1}^{2^B}} |x - r_i| \quad (12)$$

and the quantization noise is defined as:

$$q_x = x - x_q \quad (13)$$

The quantization levels r_i are chosen to minimize a fidelity metric such as the mean-squared error (MSE) defined as:

$$J = \mathbb{E}\left[(x - x_q)^2\right] = \sigma^2_{q_x}. \quad (14)$$

For mathematical tractability, we assume q_x is a zero-mean random variable independent of x.

Given a signal distribution $f_x()$, the classical Lloyd-Max (LM) algorithm [36] determines a set of quantization levels $\{r_i\}_{i=1}^{2^B}$ minimizing the quantizer's MSE in (14). Such a quantizer is referred to as the LM quantizer.

Alternatively, it is common to use a full range (FR) uniform quantizer which assigns the quantization levels: $r_i = x_{\min} + (i-1)\Delta$, for $i = 1, \ldots, 2^B$, where $\Delta = (x_{\max} - x_{\min})2^{-B}$ is the quantization step size. The quantization noise q_x as a uniformly distributed random variable [39], [40], i.e., $q_x \sim U[-\frac{\Delta}{2}, \frac{\Delta}{2}]$, and hence it is easy to show that $\sigma^2_{q_x} = \frac{\Delta^2}{12}$.

A. Clipped Quantization

Recently, we have shown that a uniform quantizer's accuracy can be improved by allowing for signal clipping [41]. Specifically, all quantization levels are placed in a narrow interval $[x_L, x_R]$, with $x_L > x_{\min}$ and $x_R < x_{\max}$. The resulting quantizer has an MSE consisting of quantization and clipping noise terms [41]:

$$J = \frac{\Delta^2}{12} + \sigma_c^2, \quad (15)$$

where, by virtue of the reduced quantization range, the step size is given by $\Delta = (x_R - x_L)2^{-B}$ and the clipping noise variance equals:

$$\sigma_c^2 = \mathbb{E}\left[(x - x_L)^2|\mathcal{A}_L\right]P(\mathcal{A}_L) + \mathbb{E}\left[(x - x_R)^2|\mathcal{A}_R\right]P(\mathcal{A}_R) \quad (16)$$

where $\mathcal{A}_L \triangleq \{x < x_L\}$ and $\mathcal{A}_R \triangleq \{x > x_R\}$ are clipping events. Thus, a clipped uniform quantizer exhibits a fundamental trade-off between its quantization and clipping noise. Hereafter, we demonstrate how to optimally clip a signal.

B. Optimally Clipped Quantization

We present the optimal clipping criterion (OCC) for signals with a Gaussian distribution. Such signals are very prominent in machine learning systems, particularly in high-dimensional dot-product outputs by virtue of the Central Limit Theorem [42]. The following theorem provides a method to compute the optimal clipping levels for a Gaussian signal:

Theorem 1: Given a Gaussian signal $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and a B-bit uniform quantizer, the optimal quantization range is $[\mu_x - \zeta^{(\mathrm{OCC})}\sigma_x, \mu_x + \zeta^{(\mathrm{OCC})}\sigma_x]$ where the optimal clipping level $\zeta^{(\mathrm{OCC})}$ is the converging point of the following recursive expression:

$$\zeta_{n+1} = \frac{\sqrt{\frac{2}{\pi}}e^{-\frac{\zeta_n^2}{2}}}{\frac{4^{-B}}{3} + 2Q(\zeta_n)}, \tag{17}$$

where $Q()$ represents the complementary CDF of a standard Gaussian $\mathcal{N}(0,1)$.

Proof: See Appendix. ■

An important consequence of Theorem 1 is that $\zeta^{(\mathrm{OCC})}$ depends on the number of bits B. Second, (17) does not explicitly compute $\zeta^{(\mathrm{OCC})}$ and requires an initial guess ζ_0. We found that no more than 10 iterations are needed when $\zeta_0 = 4$, i.e., the process is computationally simple.

The OCC quantizer is compared with the LM and FR quantizers in Fig. 2 where a standard Gaussian signal confined to the interval $[-6, 6]$, is quantized to 6 bits. The quantization range $[-6, 6]$ ensures that $> 99.99\%$ of the probability mass of the standard Gaussian signal is included for the purposes of studying quantization effects arising from three methods: (a) uniform, (b) Lloyd-Max, and (c) OCC.

Note, the LM quantizer (Fig. 2(a)) places most of its quantization levels r_i near the mean. Intuitively, most of the representation is allocated to high-density regions which minimizes the MSE. Unfortunately, the non-uniformity of the quantization levels makes it difficult to design efficient arithmetic units to further process the quantizer output [43], [44].

In contrast, the FR quantizer has a large MSE. Indeed, many of its quantization levels are placed on the tails of the distribution which is data deficient as shown in Fig. 2(b). However, FR is popular because its uniformly spaced quantization levels makes it easy to design efficient arithmetic units to process its output.

Fig. 2(c) shows that OCC pin-points the region of high signal probability density and quantizes it uniformly. As a result, the OCC quantizer's accuracy is close to that of the LM quantizer and, similar to the FR quantizer, it has uniformly spaced quantization levels. In this way, OCC preserves the desirable properties of both.

Fig. 3 plots the MSE of an quantized standard Gaussian as a function of the clipping level ζ for different values of B. It illustrates two observations regarding OCC: 1) as suggested by Theorem 1, the optimal clipping level $\zeta^{(\mathrm{OCC})}$ depends on (increases with) the quantizer's precision B; and 2) there is an

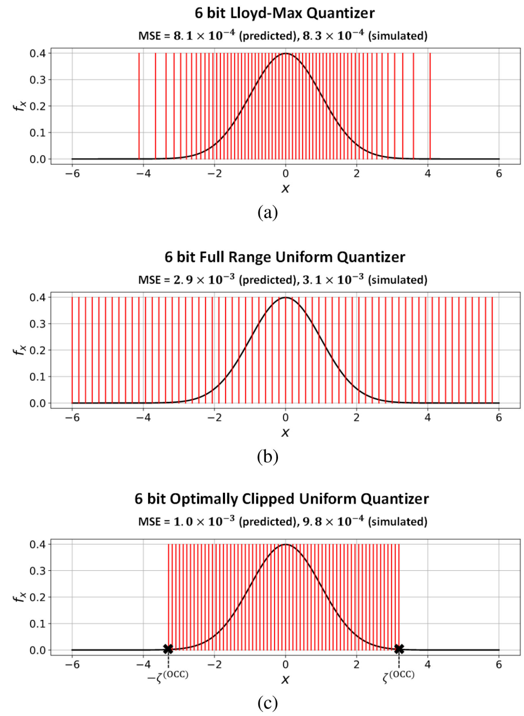


Fig. 2. Illustration of various quantization strategies for a standard Gaussian signal: (a) Lloyd-Max, (b) full range (FR) uniform, and (c) uniform quantizer using the proposed optimally clipped criterion (OCC). The predicted and simulated MSEs are obtained via evaluation of (14) using numerical integration and Monte Carlo simulations, respectively.

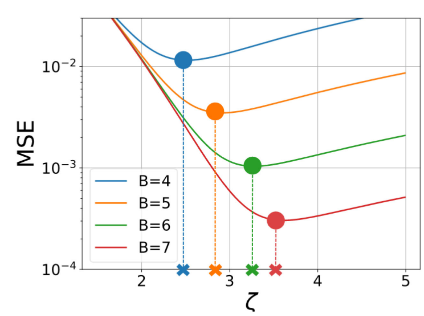
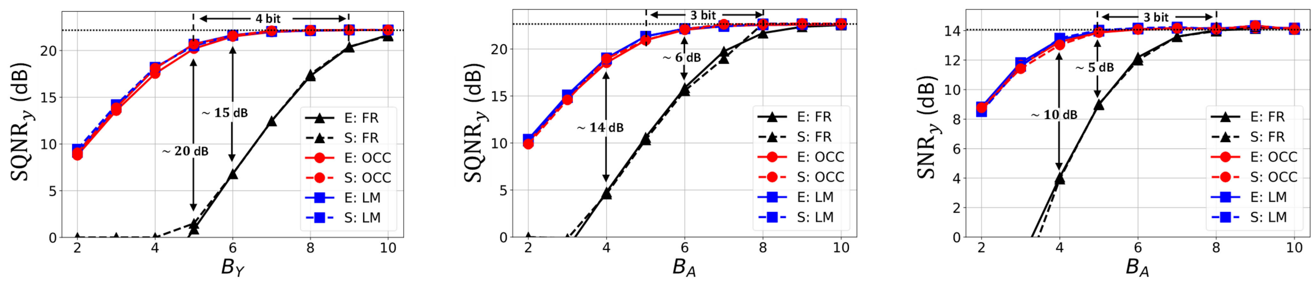


Fig. 3. Trade-off between quantization and clipping noise with OCC and dependence of $\zeta^{(\mathrm{OCC})}$ (marked as crosses) on precision B for a standard Gaussian signal.

intrinsic trade-off between the clipping noise and quantization noise alluded in (15), e.g., when $\zeta > \zeta^{(\mathrm{OCC})}$ clipping noise is reduced at the expense of the quantization noise due to the use of a large step-size Δ and vice versa. Thus, the optimal clipping level $\zeta^{(\mathrm{OCC})}$ is one that balances clipping and quantization noise.

Table II lists $\zeta^{(\mathrm{OCC})}$ for varying values of B and compares $\sigma^2_{(\mathrm{OCC})}$ and $\sigma^2_{(\mathrm{LM})}$, the quantization noise variances for the OCC and LM quantizers, respectively. We find that $\sigma^2_{(\mathrm{OCC})}$ is usually about $\sim 20\%$ higher than $\sigma^2_{(\mathrm{LM})}$ and at worst 56% when $B = 5$. Equivalently, the OCC has an SQNR within 0.8 dB of LM. Thus, the OCC, being a uniform quantizer, is a practical alternative to LM.



(a) (b) (c)

Fig. 4. Comparison of FR, OCC, and LM for output and ADC quantization: (a) SQNR$_y$ vs. B_Y in digital dot-products, (b) SQNR$_y$ vs. B_A in IMC dot-products, and (c) SNR$_y$ vs. B_A in IMC dot-products. The dot-product dimension is $N = 256$ and input/weight precisions are set as $B_X = B_W = 4$. The maximum achievable SQNR in (a) and (b) is 22.5 dB (horizontal dotted line). The bitcell capacitance used in (c) is $C_o = 1\,fF$ and the maximum achievable SNR is 14 dB (horizontal dotted line). Solid lines 'E' are obtained via evaluation of (18), (22), (21), (23), and (25); dashed lines 'S' are obtained using Monte Carlo simulations.

TABLE II
COMPARISON OF MSE BETWEEN OCC AND LM FOR A STANDARD GAUSSIAN SIGNAL

B	$\zeta^{(\text{OCC})}$	$\sigma^2_{(\text{OCC})}$	$\sigma^2_{(\text{LM})}$	$\frac{\sigma^2_{(\text{OCC})}}{\sigma^2_{(\text{LM})}}$
2	1.71	1.26×10^{-1}	1.17×10^{-1}	1.077
3	2.15	3.79×10^{-2}	3.45×10^{-2}	1.099
4	2.55	1.16×10^{-2}	9.50×10^{-3}	1.221
5	2.94	3.50×10^{-3}	2.50×10^{-3}	1.560
6	3.29	1.04×10^{-3}	8.14×10^{-4}	1.278
7	3.61	3.04×10^{-4}	2.13×10^{-4}	1.427
8	3.92	8.77×10^{-5}	7.15×10^{-5}	1.227
9	4.21	2.49×10^{-5}	2.02×10^{-5}	1.232
10	4.49	6.99×10^{-6}	5.11×10^{-6}	1.368

C. Application to Dot-Product Computation

In this section, we apply OCC to quantize the output of the dot-product described in Section II-A and compare it with LM and FR quantizers. Since these are general methods for quantization, we consider both digital and IMC computation of dot-products.

Specifically, we consider a $N = 256$-dimensional dot-product where inputs and weights are uniformly distributed in the intervals [0,1] and $[-1, 1]$, respectively. Input and weight precisions are chosen as $B_X = B_W = 4$.

1) Applying OCC to Digital Dot-Products: A digital realization of the dot-product $y = \mathbf{w}^\mathsf{T}\mathbf{x}$ exhibits three sources of noise at its output y: output-referred input quantization noise $q_{x\to y}$, output-referred weight quantization noise $q_{w\to y}$, and output quantization noise q_y. The resulting SQNR is given by:

$$\text{SQNR}_y = \frac{\sigma_y^2}{\sigma^2_{q_{x\to y}} + \sigma^2_{q_{w\to y}} + \sigma^2_{q_y}} \tag{18}$$

where

$$\sigma^2_{q_{x\to y}} = \frac{N x_m^2 \sigma_w^2 4^{-B_X}}{12}; \quad \sigma^2_{q_{w\to y}} = \frac{N w_m^2 \mathbb{E}[x^2] 4^{-B_W}}{3}, \tag{19}$$

and $\sigma^2_{q_y}$ depends on the quantization strategy, i.e., LM, FR or OCC.

An upper bound on the SQNR $\leq 22.5\,dB$ is obtained by setting $\sigma^2_{q_y} = 0$ in (18). This upper bound is achieved when employing the bit-growth criterion (BGC) [45] below:

$$B_Y = B_X + B_W + \log_2(N). \tag{20}$$

This criterion is known to be overly conservative. Instead, we consider three output quantization strategies: (1) FR employing the range $[-N, N]$, (2) OCC, and (3) LM. For each method, the output precision B_Y is swept and the SQNR is evaluated both analytically using (18), (14), and (15)) and empirically using Monte Carlo simulations.

The results in Fig. 4 indicate: (1) OCC's accuracy matches that of LM's. An asymptotic SQNR of $\sim 22.5\,dB$ is attained when $B_Y \geq 6$ for both quantizers. In contrast, the commonly employed FR has a much smaller SQNR. Its gap with respect to LM and OCC can be as high as 20 dB for $B_Y = 5$. In addition, it requires $B_Y \geq 10$ to reach the SQNR asymptote of $\sim 22.5\,dB$. Thus, OCC achieves a 4-bit reduction in output precision over FR, which is substantial.

2) OCC in IMC Dot-Products: An IMC realization of the dot-product $y = \mathbf{w}^\mathsf{T}\mathbf{x}$ exhibits four sources of noise at its output y: (1) output-referred input quantization noise $q_{x\to y}$; (2) output-referred weight quantization noise $q_{w\to y}$, (3) total ADC quantization noise $q_{A\to y}$; and (4) noise due to analog circuit non-idealities $\eta_{a\to y}$. Due to the mixture of quantization and circuit noise sources, we employ the term signal-to-noise ratio (SNR) to quantify the dot-product accuracy, where:

$$\text{SNR}_y = \frac{\sigma_y^2}{\sigma^2_{q_{x\to y}} + \sigma^2_{q_{w\to y}} + \sigma^2_{q_{A\to y}} + \sigma^2_{\eta_{a\to y}}}, \tag{21}$$

where $\sigma^2_{q_{x\to y}}$ and $\sigma^2_{q_{w\to y}}$ are given by (19), while $\sigma^2_{q_{A\to y}}$ depends on the quantization strategy employed by the column ADCs and any subsequent processing such as POTS, and $\sigma^2_{\eta_{a\to y}}$ depends upon the specific circuit style employed in the IMC.

We further define the SQNR of an IMC as an upper bound on the SNR by setting $\eta_{a\to y} = 0$, i.e., zero analog noise in (21):

$$\text{SQNR}_y = \frac{\sigma_y^2}{\sigma^2_{q_{x\to y}} + \sigma^2_{q_{w\to y}} + \sigma^2_{q_{A\to y}}}. \tag{22}$$

We consider the bit-serial bit-parallel (BSBP) architecture, which can be obtained from the ISWP architecture in Section II-A by setting $B_S = 1$ so that $N_S = B_X$. The BSBP architecture is a popular architecture today [21] because of its scalability, i.e., its accuracy is the highest when computing dot-products with large dimensions. For the BSBP architecture, it can be shown that:

$$\sigma^2_{q_{A\to y}} = \frac{4}{9}x_m^2 w_m^2 \left(1-4^{-B_X}\right)\left(1-4^{-B_W}\right)\sigma^2_{q_{A_{s,b}}}, \quad (23)$$

where $\sigma^2_{q_{A_{s,b}}} = \sigma^2_{(\mathcal{B})}\mathrm{Var}(y_{s,b})$ ($\mathcal{B}\in\{\text{OCC,LM,FR}\}$) is the column ADC quantization noise variance (see (7)) which depends on the ADC precision B_A, N and the quantization strategy employed, i.e., LM, FR or OCC.

A special case of (23) is when the input and weight bits are i.i.d. and $Be(0.5)$, i.e., Bernoulli RVs with parameter $p=0.5$. In that case, $\sigma^2_{q_{A_{s,b}}} = \frac{3}{16}\frac{N}{}\sigma^2_{(Q)}$, and therefore:

$$\sigma^2_{q_{A\to y}} = \frac{N}{12}\sigma^2_{(Q)} x_m^2 w_m^2 \left(1-4^{-B_X}\right)\left(1-4^{-B_W}\right) \quad (24)$$

We first study the SQNR in (22). The asymptote of 22.5 dB is identical to the digital dot-product case and can be obtained by setting $\sigma^2_{q_{A\to y}}=0$ in (22). Fig. 4(b) illustrates that as B_A is increased, the SQNR achieved by LM and OCC are nearly identical. Compared to FR, OCC yields a 14 dB improvement in SQNR when $B_A=4$. Furthermore, OCC reaches the SQNR asymptote for $B_A\geq 5$ as compared to FR which requires $B_A\geq 8$. Hence, OCC reduces the column ADC precision of IMCs requirements by 3 bits over the commonly employed FR method.

To study the impact of various quantization noise strategies on the SNR, the total analog noise variance for the BSBP architecture is given by:

$$\sigma^2_{\eta_{a\to y}} = \frac{4}{9}x_m^2 w_m^2 \left(1-4^{-B_X}\right)\left(1-4^{-B_W}\right)\sigma^2_{\eta_{a_{s,b}}} \quad (25)$$

with $\sigma^2_{\eta_{a_{s,b}}}$ given by (8). Employing a practical value of $C_o = 1\,fF$ results in an asymptotic SNR of 14 dB obtained by setting $\sigma^2_{q_{A\to y}}=0$ in (21). Fig. 4(c) shows that the SNR achieved with OCC is close to that of LM for all values of B_A. Furthermore, OCC improves upon FR by up to 10 dB when $B_A=4$. The asymptote of $\sim 14\,dB$ is attained when $B_A\geq 5$, implying a 3-bit reduction compared to FR which requires $B_A\geq 8$ to reach the SNR asymptote.

These results indicate that OCC is a practical alternative to LM and results in a non-trivial reduction in the column ADC precision in IMCs.

IV. Accuracy Analysis of the ISWP Architecture

The analysis in Section III focused on the SQNR of individual dot-products using OCC, LM and FR to quantize the column ADC inputs. However, the ISWP architecture computes multi-bit dot-products by slicing the inputs, computing multiple lower-precision dot-products and then combining their outputs via POTS (see Fig. 1). In this section, we investigate how the choice of input slice precision B_S and the use of OCC affects the total ADC quantization noise $q_{A\to y}$ at the output of a multi-bit dot-product computed by the ISWP architecture.

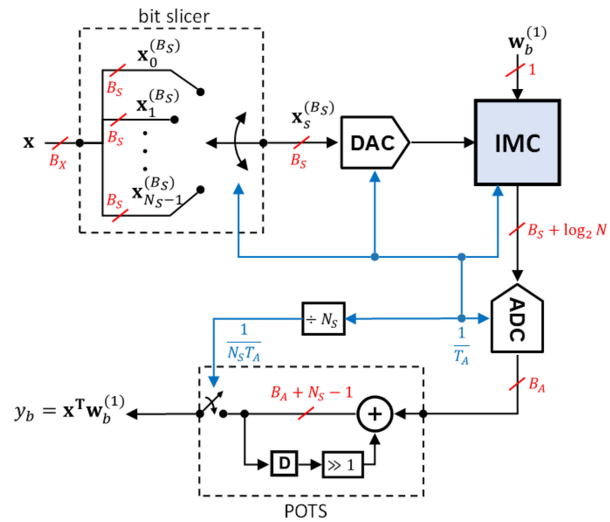


Fig. 5. Multi-bit dot-product computation in an ISWP architecture. The latency per array invocation is denoted as T_A.

A. Noise Analysis for Bit-Sliced Computation

Fig. 5 shows that an ISWP architecture computes a $B_X\times B_W$-bit dot-product by slicing the input into N_S slices of B_S bits each, using a DAC to convert each slice sequentially into the analog domain, computing a $B_S\times B_W$-bit N-dimensional BL dot-product with a maximum (BGC) precision of $B_S+\log_2(N)$ bits, using a B_A-bit column ADC to quantize the analog dot-product, and finally accumulating the digitized BL dot-products over N_S array invocations. Thus, N_S intermediate ADC quantizations occur and its impact on the final output $q_{A\to y}$ needs to be analyzed.

We prove in the Appendix that, when bits of the input $x^{(1)}_{i,b}$ and the weights $w^{(1)}_{i,b}$ are i.i.d and $Be(0.5)$ distributed, the total ADC quantization noise when employing OCC is given by:

$$\sigma^2_{q_{A\to y}} = \frac{N}{36}\times\sigma^2_{(\text{OCC})}x_m^2 w_m^2\left(1-4^{-B_X}\right)\left(1-4^{-B_W}\right)\times\beta, \quad (26)$$

where β is the *bit slicing gain* and is given by:

$$\beta = \frac{5-2^{-B_S}}{1+2^{-B_S}}. \quad (27)$$

Comparing (24) with (26)–(27) shows that $\beta/3$ is the factor by which the total ADC quantization noise $q_{A\to y}$ is amplified over the case when $B_S=1$. Furthermore, (27) indicates that β approaches a value of 5 as $B_S\to\infty$. This implies that bit slicing causes at most a $1.6\times$ increase in total ADC quantization noise variance corresponding to a 2 dB SQNR worst-case drop, equivalent to a third of an least-significant bit (LSB) [46].

Contrast this with the popular choice of $B_S=1$ used to obtain the BSBP architecture. This choice is motivated in part to minimize the impact of ADC quantization noise, also indicated by (26)–(27). In doing so, however, the BSBP architecture requires $N_S=B_X$ array invocations vs. $N_S=\lceil B_X/B_S\rceil$ invocations required by the ISWP architecture. Since (9) shows that the energy efficiency is proportional to N_S, the BSBP architecture incurs a significant energy and latency penalty for limited gains in accuracy. This conclusion runs counter to the prevalent practice and rationale for using BSBP. Our analysis indicates that there is a better option: the ISWP architecture with $B_S = B_X$.

We also analyze the impact of bit slicing on analog noise. In the Appendix, we prove that:

$$\sigma^2_{\eta_{a\to y}} = \frac{4x_m^2 w_m^2 \left(1-4^{-B_X}\right)\left(1-4^{-B_W}\right)}{3\left(1-4^{-B_S}\right)} \sigma^2_{\eta_{a_{s,b}}} \tag{28}$$

with:

$$\sigma^2_{\eta_{a_{s,b}}} = N\left(\frac{\rho_1\left(2-2^{-B_S}\right)}{12\left(1-2^{-B_S}\right)C_o} + \frac{\rho_2}{C_o} + \frac{\rho_3}{C_o^2}\right). \tag{29}$$

Thus, when B_S increases, $\sigma^2_{\eta_{a\to y}}$ decreases, though not drastically. This is not surprising since higher B_S leads to fewer array invocations and hence less accumulation of analog circuit non-idealities.

B. Impact on Multi-Bit Dot-Product Accuracy

We use the same setup as in Section III, but consider higher input precision B_X to increase the choices for B_S. Specifically, we keep $B_W = 4$, $N = 256$ but use $B_X = 8$ and $B_X = 10$. For each case, we sweep the value of $B_S = 1, \ldots, B_X$. The column ADC precision B_A is also fixed to 3, 4, or 5 bits and the quantization method is OCC.

Fig. 6(a) shows that the choice of B_S has a minor impact on the SQNR, e.g., when $B_A = 3$, the SQNR lies between $\sim 14\,dB$ for $B_S = 1$ and $\sim 12\,dB$ for $B_S \to B_X$. This validates our contention that single bit slicing offers no more than a $2\,dB$ SQNR boost. In general, when the ADC precision B_A increases, SQNR is insensitive to B_S.

Fig. 6(b) shows that the SNR in (21) with $C_o = 1\,fF$, is minimally affected by the choice of B_S, e.g., when $B_A = 3$, the SNR lies between $\sim 11.5\,dB$ for $B_S = 1$ and $\sim 10.5\,dB$ for $B_S \to B_X$. As expected the SNR is lower than the corresponding SQNR due to the presence of analog noise. Thus, in the case of the SNR too, the loss in accuracy due to multi-bit slicing is just $1\,dB$. In fact, when $B_A = 5$, the SNR$\sim 14\,dB$ more or less independent of B_S.

To summarize, the analysis in this section recommends choosing $B_S = B_X$ whenever possible. Such fully sliced (FS) IMC designs significantly improve energy efficiency with negligible loss in accuracy. Since, it is accepted that deep nets can be implemented with activations being quantized to $B_X \sim 4-6$ bits, the resulting savings in energy and latency will be significant.

V. Realizing DNN on the ISWP Architecture

We illustrate the application of our analyses in Sections III and IV to characterize the accuracy and energy efficiency of mapping various DNNs on the ISWP architecture.

A. Setup

We consider the following networks and datasets: VGG-9 [47] and ResNet-18 [2] deployed on CIFAR-10 [48], and AlexNet [1]

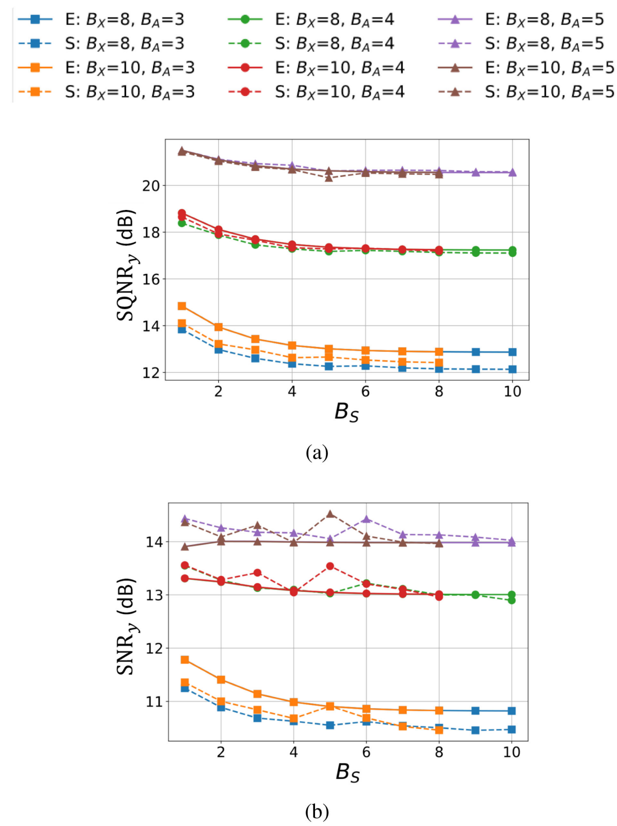


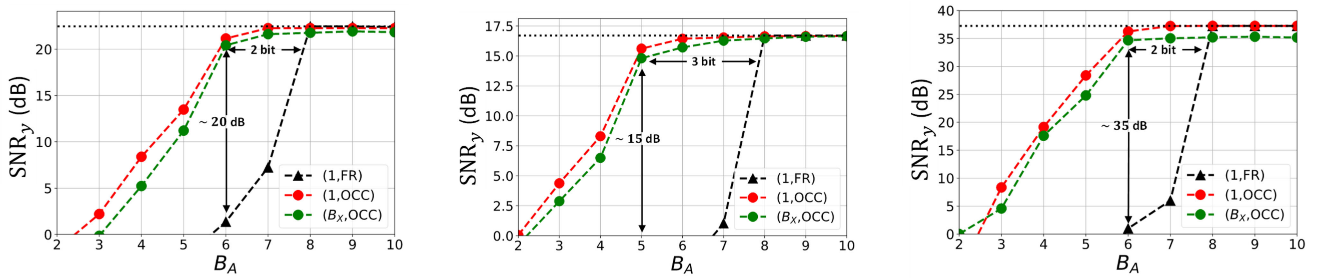
Fig. 6. Impact of bit slicing on the accuracy of IMC dot-products: (a) SQNR$_y$ vs. B_S and (b) SNR$_y$ vs. B_S. The legend is included at the top of the figure and lists various values of B_X and B_A used. The dot-product dimension is $N = 256$ and weight precision is set as $B_W = 4$. The bitcell capacitance used in (c) is $C_o = 1\,fF$. Solid lines 'E' are obtained via evaluation of (18), (22), (21), (26), (28), and (29); dashed lines 'S' are obtained using Monte Carlo simulations.

TABLE III
Accuracy, Precision, and the Dot-Product SQNR

Network (Dataset)	Accuracy FL (%)	Accuracy FX (%)	$B_X = B_W$ (bits)	SQNR$_y$ (dB)
VGG-9 (CIFAR-10)	87.71	87.47	6	22
ResNet-18 (CIFAR-10)	94.53	93.74	6	17
AlexNet (ImageNet)	56.55	55.60	10	37

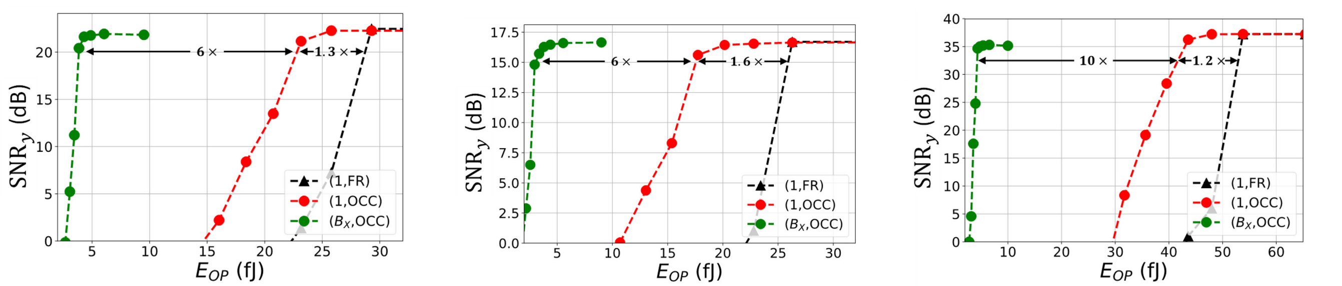
deployed on ImageNet [49], and employ the following methodology:

1) For each pre-trained floating-point (FL) network, we employ the methodology in [50] to obtain the smallest activation precision (B_X) and weight precision (B_W) such that $B_X = B_W$ and the fixed-point (FX) network accuracy remains within 1% of that of the FL baseline (see Table III).
2) For each network, we randomly select 4000 dot-products from all layers to mapped on an ISWP architecture with an array size of $N_{\text{row}} = 256$ rows. Since, DNN dot-products have very high dimensions, i.e., $N > 1000$ is not uncommon, we partition the dot-product computations across multiple banks as required.



(a) (b) (c)

Fig. 7. The trade-off between SNR and ADC precision in DNNs: (a) VGG-9 on CIFAR-10, (b) ResNet-18 on CIFAR-10, and (c) AlexNet on ImageNet. The dotted black line corresponds to the output-referred input SQNR in each case and sets an upper bound on the achievable SNR.



(a) (b) (c)

Fig. 8. The trade-off between SNR and energy consumption (measured using E_{OP} in (9)) in DNNs: (a) VGG-9 on CIFAR-10, (b) ResNet-18 on CIFAR-10, and (c) AlexNet on ImageNet.

3) We estimate the SNR via ensemble averaging over dot-products within each network where this averaging is performed both spatially, i.e., across dot-products sampled from the network, and temporally, i.e., over randomized network inputs. In this way, the Monte Carlo simulations are emulating the ISWP architecture to compute dot-products from which the SNR is estimated numerically.
4) We estimate the energy per operation E_{OP} from (9) and (11).

We consider three implementation methods: (1) (1,FR) which is the conventional BSBP architecture using $B_S = 1$ and FR quantization in the ADC; (2) (1,OCC) which employs $B_S = 1$ and OCC quantization in the ADC; and (3) (B_X,OCC) which is the FS (fully-sliced) architecture with $B_S = B_X$ and OCC quantization in the ADC.

B. SNR vs. ADC Precision

We set bitcell capacitance to be sufficiently high so that the analog noise term $\sigma_{\eta_{a\to y}}$ in (21) is negligible. Fig. 7 shows that (1,OCC) requires an ADC precision B_A that is between 2-3 bits lower than the conventional approach of using an FR quantizer (1,FR) across the three networks. Alternatively, for the same ADC precision, (1,OCC) achieves between 15 dB to 35 dB higher SNR$_y$ than (1,FR). As mentioned earlier, the energy consumption of of ADCs is exponential in B_A (see (11)) since these operate in the noise-limited regime. Hence, the aforementioned reduction in ADC precision results in a significant savings in energy consumption of the ADCs.

Fig. 7 also indicates that (B_X,OCC) exhibits minimal loss in SNR$_y$ as compared to (1,OCC) implying that the use of $B_S = B_X$, i.e., processing a B_X-bit input in one array invocation ($N_S = 1$) is feasible. Doing to leads to an additional reduction in array energy consumption by a factor of B_X.

Thus, the use of (B_X,OCC) reduces both E_{ADC} and N_S in (9) leading to significant overall energy savings. Next, we quantify these energy savings.

C. SNR vs. Energy-Efficiency Trade-off

Fig. 8 shows that (B_X,OCC) enhances the fundamental energy-efficiency metric E_{OP} in (9) by a factor of 7.8×-to-12× over the conventional (1,FR) or BSBP architecture. The bulk (6×-to-10×) of energy savings arises from (B_X,OCC) invoking the array once ($N_S = 1$) compared to (1,FR). For example, the higher input precision of 10 bits in AlexNet results in a 10× savings in energy consumption. The rest of the energy savings are from the reduction in ADC precision from the use of OCC. These results are observed to be consistent across all three networks.

In summary, we have demonstrated that OCC and bit slicing reduces the energy consumption of the ISWP architecture by an order-of-magnitude compared to the conventional approach at iso-accuracy.

VI. Conclusion

We have presented the optimal clipping criterion (OCC) method for minimizing the ADC precision in IMCs and found that it saves between 2-3 bits in ADC precision. The ISWP architecture, a generalization of the popular BSBP IMC, was proposed to reduce the number of array invocations. Application of OCC to the ISWP architecture is shown to provide about an order-of-magnitude reduction in the energy cost per operation at iso-accuracy. Since IMCs have already shown to be close to two orders-of-magnitude more efficient than digital architectures at iso-accuracy [18], our work extends these gains and empowers IMC designers to push the limits of the energy vs. accuracy trade-off intrinsic to IMCs. Though OCC is particularly useful for IMCs, it is also highly effective when minimizing the output precision of digital filters and dot-products since it provides a theoretically justified alternative to the bit-growth criterion (BGC) commonly employed by digital designers.

Many types of IMC designs are being proposed today. However, not much work is being done in comprehending their energy vs. accuracy trade-off primarily due to the challenging nature of this problem. This paper has formulated a framework based on quantization noise analysis employed in digital signal processing systems in order to analyze the energy vs. accuracy trade-off in IMCs and employed this analysis to motivate the ISWP architecture. We believe the framework in this paper can be repurposed to analyze other IMCs resulting in significantly improved IMCs in the future. Future work can also include further validating the results of the proposed methods on real-life integrated circuit prototypes of IMCs.

Acknowledgment

We thank Sujan Gonugondla for helpful discussions and suggestions.

Appendix

Proof of Theorem 1: Without loss of generality, we consider B-bit uniform quantization of a unit Gaussian signal $x \sim \mathcal{N}(0,1)$ in the range $[x_L, x_R]$. A necessary condition for optimality is $x_R = -x_L = \zeta$ by virtue of the distribution's symmetry. The MSE in (15) can be written as the following function of ζ:

$$f(\zeta) = \frac{\zeta^2 2^{-2B}}{3} + 2\int_{\zeta}^{\infty} \frac{1}{\sqrt{2\pi}}(x-\zeta)^2 e^{-\frac{x^2}{2}}dx, \qquad (30)$$

where we used $\Delta_x = \zeta 2^{-B}$ and $\sigma_c^2 = 2\mathbb{E}[(x-\zeta)^2 \mathbb{1}_{\{x>\zeta\}}]$. Our task is to find $\zeta^{(\mathrm{OCC})}$ minimizing $f(\zeta)$ in (30) which can be written as:

$$f(\zeta) = f_0(\zeta) + \sqrt{\frac{2}{\pi}}\left(f_1(\zeta) + f_2(\zeta) + f_3(\zeta)\right) \qquad (31)$$

with $f_0(\zeta) = \frac{\zeta^2 2^{-2B}}{3}$, $f_1(\zeta) = \int_\zeta^\infty x^2 e^{-\frac{x^2}{2}}dx$, $f_2(\zeta) = -2\zeta\int_\zeta^\infty x e^{-\frac{x^2}{2}}dx$, and $f_3(\zeta) = \zeta^2\int_\zeta^\infty e^{-\frac{x^2}{2}}dx$. It follows that:

$$f_0'(\zeta) = 2\zeta\frac{2^{-2B}}{3} \quad \text{and} \quad f_1'(\zeta) = -\zeta^2 e^{-\frac{\zeta^2}{2}} \qquad (32)$$

$$f_2'(\zeta) = -2\int_\zeta^\infty x e^{-\frac{x^2}{2}}dx + 2\zeta^2 e^{-\frac{\zeta^2}{2}} = 2(\zeta^2-1)e^{-\frac{\zeta^2}{2}} \qquad (33)$$

$$f_3'(\zeta) = 2\zeta\int_\zeta^\infty e^{-\frac{x^2}{2}}dx - \zeta^2 e^{-\frac{\zeta^2}{2}}. \qquad (34)$$

Combining (31), (32), (33), and (34) yields:

$$\begin{aligned} f'(\zeta) &= 2\zeta\frac{2^{-2B}}{3} + \sqrt{\frac{2}{\pi}}\left(2\zeta\int_\zeta^\infty e^{-\frac{x^2}{2}}dx - 2e^{-\frac{\zeta^2}{2}}\right) \\ &= 2\left[g_0(\zeta) + \sqrt{\frac{2}{\pi}}\left(g_1(\zeta) + g_2(\zeta)\right)\right], \end{aligned} \qquad (35)$$

where $g_0(\zeta) = \zeta\frac{2^{-2B}}{3}$, $g_1(\zeta) = \zeta\int_\zeta^\infty e^{-\frac{x^2}{2}}dx$, and $g_2(\zeta) = -e^{-\frac{\zeta^2}{2}}$. It follows that:

$$g_0'(\zeta) = \frac{2^{-2B}}{3} \quad \text{and} \quad g_2'(\zeta) = \zeta e^{-\frac{\zeta^2}{2}} \qquad (36)$$

$$g_1'(\zeta) = \int_\zeta^\infty e^{-\frac{x^2}{2}}dx - \zeta e^{-\frac{\zeta^2}{2}}. \qquad (37)$$

Combining (35), (36), and (37) yields:

$$f''(\zeta) = 2\left(\frac{2^{-2B}}{3} + \sqrt{\frac{2}{\pi}}\int_\zeta^\infty e^{-\frac{x^2}{2}}dx\right), \qquad (38)$$

which is strictly positive for any ζ. Hence, $f(\zeta)$ is convex and can be minimized using Newton's algorithm [51] via the following recursion:

$$\zeta_{n+1} = \zeta_n - \frac{f'(\zeta_n)}{f''(\zeta_n)}. \qquad (39)$$

Replacing (35) and (38) into (39) and substituting $\sqrt{\frac{2}{\pi}}\int_\zeta^\infty e^{-\frac{x^2}{2}}dx = 2Q(\zeta)$ yields (17) in Theorem 1 which concludes our proof.

Derivation of (26): Combining (5) and (7), we have:

$$q_{A\to y} = x_m w_m \sum_{s=0}^{N_S-1}\left(-q_{A_{s,0}} + \sum_{b=1}^{B_W-1} q_{A_{s,b}} 2^{-b}\right)2^{-sB_S}$$

and it follows that:

$$\begin{aligned} \sigma^2_{q_{A\to y}} &= x_m^2 w_m^2 \sum_{s=0}^{N_S-1}\sum_{b=0}^{B_W-1} \sigma^2_{q_{A_{s,b}}} 4^{-b} 4^{-sB_S} \\ &= \frac{4x_m^2 w_m^2}{3}\sigma^2_{q_{A_{s,b}}}\frac{\left(1-4^{-B_W}\right)\left(1-4^{-B_X}\right)}{1-4^{-B_S}}. \end{aligned} \qquad (40)$$

Recall the column ADC uses the OCC so that:

$$\sigma^2_{q_{A_{s,b}}} = \mathrm{Var}(y_{s,b})\sigma^2_{(\mathrm{OCC})} = N\mathrm{Var}(x_s^{(B_S)} w_b^{(1)})\sigma^2_{(\mathrm{OCC})}. \qquad (41)$$

From the equiprobable bitwise representation assumption we have $w_b^{(1)} \sim Be(0.5)$ is a Bernoulli random variable and $x_s^{(B_S)} = \frac{u_s}{2^{B_S}}$ where $u_s \sim U(0, 2^{B_S}-1)$ is a discrete uniform random variable. Hence, it can be shown that:

$$\mathrm{Var}\left(x_s^{(B_S)} w_b^{(1)}\right) = \frac{\left(1-2^{-B_S}\right)\left(5-2^{-B_S}\right)}{48}. \qquad (42)$$

Substituting (42) and (41) into (40) yields (26) which concludes our proof.

Derivation of (28)–(29): First, (28) follows from combining (5) and (7) in a similar fashion as was done to obtain (40). Then, (29) is obtained from (8) by evaluating:

$$\mathbb{E}\left[\left(x_s^{(B_S)} w_b^{(1)}\right)^2\right] = \frac{1}{12}\left(1 - 2^{-B_S}\right)\left(2 - 2^{-B_S}\right).$$

This result itself is a consequence of the equiprobable bitwise representation assumption discussed in the derivation of (26).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[3] H. Sak *et al.*, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 4280–4284.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A. neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 4960–4964.

[5] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu, "Bridging the gap between training and inference for neural machine translation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4334–4343.

[6] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *Proc. 41st Int. ACM SIGIR Conf. Res. & Develop. Inf. Retrieval*, 2018, pp. 1371–1374.

[7] On-Device Intelligence, "Conference workshop," in *Proc. Int. Conf. Mach. Learn.*, 2016.

[8] N. Shanbhag, "Energy-efficient machine learning in silicon: A communications-inspired approach," in *Proc. Int. Conf. Mach. Learn. On-device Intell. Workshop*, 2016.

[9] B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," Google Research Blog, vol. 3, 2017.

[10] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *Proc. IEEE Int. Solid-State Circuits Conf. Digest Tech. Papers*, 2014, pp. 10–14.

[11] M. Kang, M. Keel, N. R. Shanbhag, S. Eilert, and K. Curewitz, "An energy-efficient VLSI architecture for pattern recognition via deep embedding of computation in SRAM," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 8326–8330.

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

[13] M. Kang, S. Gonugondla, and N. R. Shanbhag, *Deep In-Memory Architectures for Machine Learning*. Berlin, Germany: Springer, 2020.

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

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

[16] S. Yin, Z. Jiang, J. Seo, and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," *IEEE J. Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, Jun. 2020.

[17] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2018, pp. 488–490.

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

[19] H. Dbouk, S. K Gonugondla, C. Sakr, and N. R Shanbhag, "KeyRAM: A. 0.34 uJ/decision 18k decisions/s recurrent attention in-memory processor for keyword spotting," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2020, pp. 1–4.

[20] W.-S. Khwa *et al.*, "A 65nm 4Kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3ns and 55.8 TOPS/W fully parallel product-sum operation for binary DNN edge processors," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2018, pp. 496–498.

[21] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement," in *Proc. IEEE Symp. VLSI Circuits*, 2018, pp. 141–142.

[22] J. Kim *et al.*, "Area-efficient and variation-tolerant in-memory BNN computing using 6T SRAM array," in *Proc. IEEE Symp. VLSI Circuits*, 2019, pp. 118–119.

[23] Q. Dong *et al.*, "A 351 TOPS/W and 372.4 GOPS compute-in-memory SRAM macro in 7nm FinFET CMOS for machine learning applications," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2020, pp. 242–243.

[24] J.-W. Su *et al.*, "A 28nm 64Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2020, pp. 240–241.

[25] X. Si *et al.*, "A 28nm 64Kb 6T SRAM computing-in- memory macro with 8b MAC operation for AI edge chips," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2020, pp. 246–247.

[26] C.-X. Xue *et al.*, "A 22nm 2Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit MAC computing for tiny AI edge devices," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2020, pp. 244–245.

[27] M. Courbariaux *et al.*, "BinaryConnect: Training deep neural networks with binary weights during propagations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3123–3131.

[28] I. Hubara, M. Courbariaux, D. Soudry, R. EI-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4107–4115.

[29] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Visi.*, 2016, pp. 525–542.

[30] M. Kim and P. Smaragdis, "Bitwise neural networks," 2016, *arXiv:1601.06071*.

[31] S. K. Gonugondla, C. K. Sakr, H Dbouk, and N. R. Shanbhag, "Fundamental limits on energy-delay-accuracy of in-memory architectures in inference applications," 2020, *arXiv:2012.13645*.

[32] B. Murmann, "Mixed-signal computing for deep neural network inference," *IEEE Trans. Very Large Scale Integration Syst.*, vol. 29, no. 1, pp. 3–13, Jan. 2021.

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

[34] A. Rekhi *et al.*, "Analog/mixed-signal hardware error modeling for deep learning inference," in *Proc. 56th Annu. Des. Automat. Conf.*2019, 2019, pp. 1–6.

[35] K. Parhi, *VLSI Digital Signal Processing Systems: Design and Implementation*. Hoboken, NJ, USA: Wiley, 2007.

[36] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[37] B. Murmann, "ADC performance survey," 1997-2019. [Online]. Available: https://web.stanford.edu/ murmann/adcsurvey.html

[38] B. Murmann, "A/D converter trends: Power dissipation, scaling and digitally assisted architectures," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2008, pp. 105–112.

[39] J. B. Evans, P. Xue, and B. Liu, "Analysis and implementation of variable step size adaptive algorithms," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2517–2535, Aug. 1993.

[40] C. Caraiscos and B. Liu, "A roundoff error analysis of the LMS adaptive algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 1, pp. 34–41, Feb. 1984.

[41] S. K. Gonugondla, C. Sakr, H. Dbouk, and N. R. Shanbhag, "Fundamental limits on the precision of in-memory architectures," in *Proc. 39th Int. Conf. Comput.-Aided Des.*, 2020, pp. 1–9.

[42] O. Johnson, *Information Theory and the Central Limit Theorem*. Singapore: World Scientific, 2004.

[43] R. Narasimha, M. Lu, N. R. Shanbhag, and A. C. Singer, "Ber-optimal analog-to-digital converters for communication links," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3683–3691, Jul. 2012.

[44] J. Mo and R. W. Heath, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498–5512, Oct. 2015.

[45] S. Zhang and N. R. Shanbhag, "Embedded algorithmic noise-tolerance for signal processing and machine learning systems via data path decomposition," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3338–3350, Jul. 2016.

[46] M. Goel and N. Shanbhag, "Finite-precision analysis of the pipelined strength-reduced adaptive filter," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1763–1769, Jun. 1998.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[48] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Tech. Rep., 2009.
[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[50] C. Sakr, Y. Kim, and N. Shanbhag, "Analytical guarantees on numerical precision of deep neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3007–3016.
[51] T. J. Ypma, "Historical development of the Newton-Raphson method," *Soc. Ind. Appl. Math. Rev.*, vol. 37, no. 4, pp. 531–551, 1995.

Charbel Sakr (Member, IEEE) received the B.E. degree (with High Distinction) from the American University of Beirut, in 2015, and the M.S. and Ph.D. degrees from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2017 and 2021, respectively. He is currently a Research Scientist with ASIC & VLSI Research Group, NVIDIA. His research interests include resource-constrained machine learning, with a particular focus on analysis and implementation of reduced precision models and algorithms and their co-design with machine learning accelerator hardware. Dr. Sakr was awarded the Best in Session Award at Techcon 2017 and the Rambus Fellowship from the University of Illinois during 2018–2019 and 2019–2020.

Naresh R. Shanbhag (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1993. He is currently the Jack Kilby Professor of Electrical and Computer Engineering with the University of Illinois at Urbana-Champaign, Champaign, IL, USA. From 1993 to 1995, he was with AT&T Bell Laboratories, Murray Hill, NJ, USA, where he led the design of high-speed transceiver chip-sets for very high-speed digital subscriber line, before joining the University of Illinois at Urbana-Champaign in August 1995. He has held visiting faculty appointments with National Taiwan University, Taipei, Taiwan, during August–December 2007 and Stanford University, Stanford, CA, USA, during August–December 2014. He holds 13 US patents, is the coauthor of two books and multiple book chapters, and more than 200 publications in his research field, which include the design of energy-efficient systems for machine learning, communications, and signal processing, spanning algorithms, architectures, and integrated circuits. Dr. Shanbhag was the recipient of the 2018 SIA/SRC University Researcher Award, the 2010 Richard Newton GSRC Industrial Impact Award, the IEEE Circuits and Systems Society Distinguished Lecturership in 1997, the National Science Foundation CAREER Award in 1996, and multiple best paper awards. In 2000, he co-founded and was the Chief Technology Officer of the Intersymbol Communications, Inc., which introduced mixed-signal ICs for electronic dispersion compensation of OC-192 optical links, and became a part of Finisar Corporation in 2007. From 2013 to 2017, he was the Founding Director of the Systems on Nanoscale Information fabrics Center, a five-year multi university center funded by DARPA and SRC under the STARnet Program.