Deep In-memory Architectures for Machine Learning - Accuracy vs. Efficiency Trade-offs

Mingu Kang, *Member, IEEE*, Yongjune Kim, *Member, IEEE*
Ameya D. Patil, *Student Member, IEEE*, and Naresh R. Shanbhag, *Fellow, IEEE*

***Abstract*—In-memory architectures, in particular, the *deep in-memory architecture* (DIMA) has emerged as an attractive alternative to the traditional von Neumann (digital) architecture for realizing energy and latency-efficient machine learning systems in silicon. Multiple DIMA integrated circuit (IC) prototypes have demonstrated energy-delay product (EDP) gains of up to $100\times$ over a digital architecture. These EDP gains were achieved *minimal* or sometimes *no loss* in decision-making accuracy which is surprising given its intrinsic analog mixed-signal nature. This paper establishes models and methods to understand the fundamental energy-delay and accuracy trade-offs underlying DIMA by: 1) presenting silicon-validated energy, delay, and accuracy models; and 2) employing these to quantify DIMA's decision-level accuracy and to identify the most effective design parameters to maximize its EDP gains at a given level of accuracy. For example, it is shown that: 1) DIMA has the potential to realize between $21\times$-to-$1365\times$ gains; 2) its energy-per-decision is approximately $10\times$ lower at the same decision-making accuracy under most conditions; 3) its accuracy can always be improved by increasing the input vector dimension and/or by increasing the bitline swing; and 4) unlike the digital architecture, there are quantifiable conditions under which DIMA's accuracy is fundamentally limited due to noise.**

***Index Terms*—In-memory computing, analog processing, machine learning, processor, accelerator.**

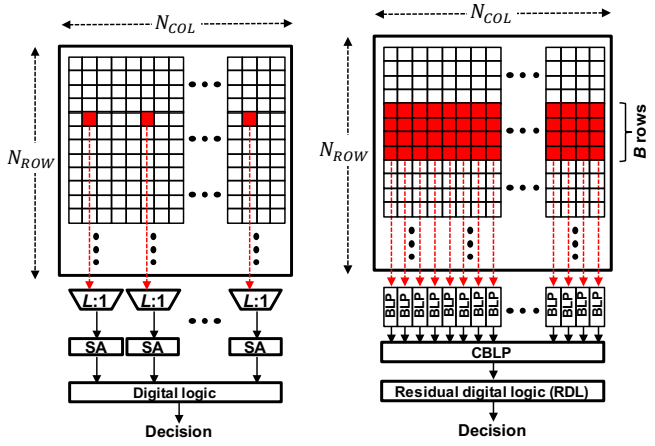
I. Introduction

Emerging applications such as internet of things (IoT), health care, autonomous driving, and sensor-rich platforms demand local decision making capability using machine learning (ML) algorithms. These applications require real-time processing in limited form factor and stringent energy constraints to be performed on autonomous battery-powered platforms. The energy-delay product (EDP) of ML hardware is limited by the energy-delay cost of memory accesses in the von Neumann (digital) architecture, which is the mainstream architecture of choice. This architecture suffers from the well-known *memory wall* problem whereby the energy cost of a single memory access is between 2-to-3 orders-of-magnitude greater than a multiply-accumulate operation in modern process technologies [1]. Though longstanding, the memory wall problem gets severely aggravated for ML workloads since large data volumes need to be processed per inference. The dominant role of memory accesses shows up clearly when realizing systems that incorporate large data models such as deep neural networks (DNNs) and convolutional neural networks (CNNs), e.g., 144 million parameters in popular VGGNet-19 [2].

Nevertheless, several digital architectures [3]–[6] have been proposed to reduce the EDP of inference tasks. These architectures employ specialized data-flow to maximize data reuse and reduce the number of off-chip data accesses, use 16-b fixed-point arithmetic precision, use pruned networks, and focus primarily on the inference (forward) path. Low-power circuit and architectural techniques such as dynamic voltage-frequency scaling [5], RAZOR [6], and power gating [7] have been also employed to achieve energy savings in the digital domain. Static random access memories (SRAMs) customized for ML algorithms [8]–[11] have been proposed which aim to reduce the data access costs in digital architectures. These tend to exploit the intrinsic robustness of ML algorithms to computational errors to reduce the EDP of memory accesses, e.g., protecting MSBs selectively. Though these techniques help alleviate the cost of memory accesses, they do not address its root cause which is the separation between computation and memory in the von Neumann architecture.

The *deep in-memory architecture* (DIMA) (see Fig. 1) first proposed in [12]–[14] strives to eliminate the separation between computation and memory by transforming the conventional memory read process into one in which functions of stored data are fetched. This is achieved by embedding computations into and in the periphery of the memory core of a standard 6T SRAM, i.e., the bitcell array (BCA) *without altering the structure of the bitcell or the BCA* in order to preserve the storage density. DIMA accesses multiple rows of a standard 6T SRAM BCA per precharge via pulse-width or amplitude modulation (PWM or PAM) of wordline (WL) access pulses to generate bitline (BL) voltage discharge $\Delta V_{\rm BL}$ that is now a function of multiple bits in a column rather than just a single one as in conventional digital architectures. The BL voltage discharge $\Delta V_{\rm BL}$ is processed further via analog mixed-signal computations in the periphery of the BCA making DIMA inherently analog in nature. By avoiding explicit data fetching and by computing in analog with low voltage swing per bit, DIMA is able to reduce both the energy and latency costs of both memory accesses and computations.

However, DIMA's analog computations need to be dimensionally matched, i.e., row and column pitch-matched, to the BCA thereby severely limiting the component (transistor and capacitor) sizes. This limitation combined with the presence of process, voltage, and temperature (PVT) variations reduces the signal-to-noise ratio (SNR) of the analog computations raising questions about DIMA's ability to generate accurate inferences. Remarkably, in spite of its inherent low-SNR nature, multiple CMOS integrated circuit (IC) prototypes [15]–[18] have demonstrated more than 2-orders-of-magnitude gains in the EDP of inference over a fixed-function digital architecture with *no/minimal loss in inference accuracy*.



(a) (b)

Fig. 1. The read operation in: (a) conventional system with a $L:1$ column mux and sense amplifiers (SAs), and (b) DIMA [12], [14], [16] with a functional read (FR) with bitline processors (BLPs), and a cross BLP (CBLP). The bitcells marked in red are accessed simultaneously per precharge cycle.

Therefore, question remains - *how robust is DIMA to the typical sources of circuit non-idealities in the low-SNR regime?* and *what is the fundamental EDP vs. accuracy trade-off in DIMA?* Answering these questions is critical if one wants to maximize the achievable EDP gains by operating at the lowest possible SNR. Indeed, a number of in-memory [15], [19], [20] and near-memory architectures [21], [22] strive to avoid this SNR loss via a combination of: 1) implementing severely quantized networks, e.g., BinaryNets [23] and XNOR-NET [24], 2) altering the bitcell architecture [19], [20], [25] and/or partitioning the BCA e.g., *Conv-RAM* [20], 3) using full-swing wordline and bitline voltages [19], all of which reduce either the storage density, or increase the energy consumption.

This paper: 1) derives *silicon-validated* energy, delay and accuracy models to quantify DIMA's low-SNR analog behavior, 2) employs these models to predict the experimentally achieved accuracy of two basic inference algorithms: the support vector machine (SVM) and template matching (TM), and obtain the limits of energy efficiency given accuracy constraints, and 3) provides design guidelines to maximize DIMA's EDP gains over a digital architecture.

The rest of the paper is organized as follows: Section II contrasts DIMA with a digital architecture in terms of its function, energy, and throughput. Section III quantifies the low-SNR attribute of DIMA from first principles to obtain models for accuracy of DIMA's analog computations (component accuracy models), and correlates it to silicon measurements reported in [16]. In Section IV, DIMA's silicon-validated component accuracy models from Section III are employed to predict the experimentally achieved accuracy of inference for SVM and TM algorithms as reported in silicon [16]. In Section V, the energy and delay models from Section II and the component accuracy models from Section III are employed to identify the limits of energy savings feasible from DIMA as well as specific design guidelines to approach those limits.

II. The Deep In-memory Architecture (DIMA): Energy and Delay Benefits

This section first describes DIMA's functionality and then compares its energy and delay gains over a digital architecture using first order models of energy consumption and delay.

A. The DIMA Processing Chain

The DIMA (Fig. 1(b)) employs standard SRAM BCA, read and write circuitry (omitted for simplicity) to preserve conventional read and write functionalities. In addition, DIMA processes both memory access and computation via embedded analog processors. For this mode, the B bits of the scalar weight W are pre-stored in a column-major format vs. row-major used in the conventional read mode. Any overhead associated with writing W in a column-major format is amortized over many inference computations since the weights are computed off-line and written once into the DIMA's BCA.

DIMA based on a $N_{\rm ROW} \times N_{\rm COL}$ BCA has the following features:

- ***Multi-row functional read (FR)***: fetches a word W by reading a function of B rows per BL precharge (read cycle) from each column (Fig. 1(b)). Thus, $N_{\rm COL}$ words are read simultaneously per read cycle.
- ***BL processing (BLP)***: computes the scalar distance (SD) between W and X per column with an analog BLP block operating with low voltage swing based on charge-transfer mechanism [16]. The $N_{\rm COL}$ BLP blocks operate in parallel in a single-instruction multiple data (SIMD) manner.
- ***Cross BL processing (CBLP)***: aggregates the $N_{\rm COL}$ analog SD results from BLP blocks by charge-sharing to obtain the vector distance (VD).
- ***Analog to digital converter (ADC)*** *and* ***residual digital logic (RDL)***: process a thresholding/decision function $f(\cdot)$ and other miscellaneous functions.

The DIMA is well-matched to the data-flow intrinsic to commonly encountered ML algorithms. It is most efficient for inference (forward) path computations. The area overhead of BLP is around 19% of the core area [16] while its energy overhead is 26% (5%) for SVM (TM) for $N_{\rm row} = 512$ and BL swing per bit = 0.14 V. Thus, the BCA dominates the overall read energy and area. The next subsection provides a justification for the energy-delay benefits of DIMA.

B. Energy and Delay Models

This section compares the energy and delay models for the digital architecture and DIMA where we assume that the BCA size $N_{\rm ROW} \times N_{\rm COL}$, the BL precharge voltage $V_{\rm PRE}$, and the maximum BL swing $\Delta V_{\rm BL,max}$ are identical in both architectures.

The SRAM in the digital architecture typically includes an $L:1$ column multiplexer (typically $L = 4, ..., 16$) as shown in Fig. 1(a), to accommodate the large footprint of sense amplifiers (SAs). One can employ per-column SAs with narrow footprint for the SRAM in the conventional system by allowing more BL swing ($\Delta V_{\rm BL,max}$) to compensate the

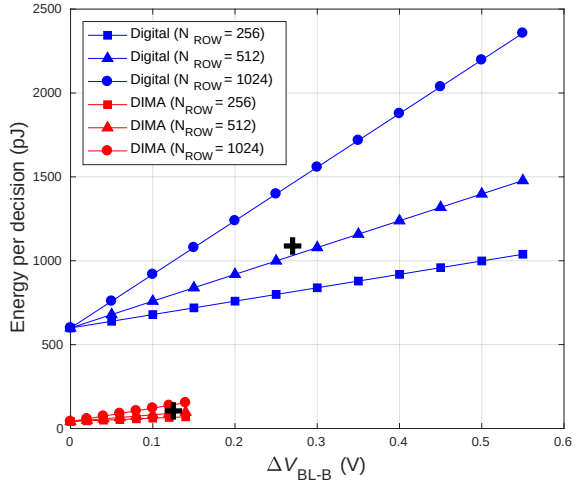


Fig. 2. Comparing the energy consumption of DIMA and the digital architecture obtained from (2) and (3) with $N_{\text{ROW}} = 256, 512$, and 1024, $L = 4$, $B = 4$, $\beta = 1$ in realizing a $N = 128$-dimensional SVM. The plus markers (+) indicate silicon measured results from [16] with $N_{\text{ROW}} = 512$.

increased offset at the cost of energy and latency or by employing advanced SA schemes at the cost of energy and area. Even in such cases, it can be shown that the EDP gain (from (5)) of up to $12\times$ is feasible with $B = 6$, which DIMA can support [16].

The number of bits per read cycle is limited to N_{COL}/L in the conventional system compared to $N_{\text{COL}}B$ in DIMA's FR. Therefore, DIMA needs LB times fewer read cycles to read the same number of bits. However, the read cycle time for DIMA is larger than that of the digital architecture since DIMA's read cycle includes both data read and compute functions via the FR, BLP and CBLP stages. Ignoring the compute delay of the digital architecture, we find that the delay reduction factor in reading a fixed number of bits is given by:

$$\rho_{\text{d}} = LB \cdot \frac{T_{\text{digital}}}{T_{\text{DIMA}}} = \frac{LB}{\gamma} \tag{1}$$

where the T_{digital} and T_{DIMA} are the read cycle times of the digital architecture and DIMA, respectively, and $\gamma = T_{\text{DIMA}}/T_{\text{digital}} \approx 3$ or 6 (in 65 nm CMOS) depending on the BLP function being computed. Previous work [16] has shown that DIMA can realize $B \leq 6$ with $B = 4$ being comfortably realized. Hence, $\rho_d = 5\times$-to-$21\times$ is easily achievable with typical values of $B = 4$, $L = 4, 8, 16$ and $\gamma = 3$ with $\rho_d = 5.3$ demonstrated in silicon [16].

The dominant sources of energy consumption in a SRAM array are the dynamic energy consumed to precharge large BL capacitances C_{BL} every read cycle and due to leakage (e.g., precharge energy takes 74% and 95% of total energy for SVM and TM, respectively [16]). The energy consumed in reading B bits in the digital architecture and DIMA can be expressed as:

$$E_{\text{digital}} = LBC_{\text{BL}}\Delta V_{\text{BL,max}}V_{\text{PRE}} + E_{\text{lk-digital}} \tag{2}$$

$$E_{\text{DIMA}} = \beta C_{\text{BL}}\Delta V_{\text{BL,max}}V_{\text{PRE}} + \frac{E_{\text{lk-digital}}}{\rho_d} \tag{3}$$

where $E_{\text{lk-digital}}$ is the leakage energy of the digital architecture, and the scaling factor L in the first term of (2) accounts for BL discharge of the $L-1$ unselected columns as well since the WL is physically shared across all the bitcells in a row [26]. The coefficient β depends on whether DIMA's FR is a pure B-bit read ($\beta = 1$) or incorporates a $2B$-bit computation in the FR stage ($\beta = 2$) [16]. The leakage energy of DIMA is reduced by a factor of ρ_d since the array can be placed into a standby mode after T_{DIMA} duration.

Since the first term in both (2) and (3) is the dominant component of the energy consumption during active mode (C_{BL} is the largest node capacitance in either architecture), their ratio provides the energy reduction factor ρ_e as follows:

$$\rho_e = \frac{E_{\text{DIMA}}}{E_{\text{digital}}} = \frac{LB}{\beta} \tag{4}$$

with values for $\rho_e = 8\times$-to-$32\times$ is easily achievable for typical values of $B = 4$, $L = 4$, 8, 16 and $\beta = 2$.

Hence, from (1) and (4), the EDP reduction over a digital architecture enabled by DIMA is given by:

$$\rho_{edp} = \rho_e\rho_d = \frac{(LB)^2}{\beta\gamma} \tag{5}$$

which ranges from $21\times$ ($L = 4$, $B = 4$, $\beta = 2$, and $\gamma = 6$)-to-$1365\times$ ($L = 16$, $B = 4$, $\beta = 1$, and $\gamma = 3$) of which the prototype IC in [18] has achieved $100\times$ in the laboratory. This clearly indicates that there is significant room to improve upon DIMA's EDP gains achieved thus far. In this paper, we provide the design guidelines to maximize these EDP gains.

It is also possible to show that when comparing the energy cost of computation only, DIMA's low-swing analog computation is approximately $10\times$ lower than that of the digital architecture.

Though the energy models (2)–(3) are simple, Fig. 2 shows that these correlate well with measured values from silicon [16] for $N_{\text{ROW}} = 512$ and $\Delta V_{\text{BL,max}} = 500\,\text{mV}$ with a modeling error of 11%. Note: the BL capacitance $C_{\text{BL}} = C_{\text{BLC}}N_{\text{ROW}}$ where C_{BLC} is the BL capacitance per cell, i.e., C_{BL} is proportional to the number of rows N_{ROW}. Hence, Fig. 2 also shows that the energy consumption increases linearly with N_{ROW} and BL swing per bit ($= \Delta V_{\text{BL-B}} = \frac{\Delta V_{\text{BL,max}}}{B}$ for DIMA and $\Delta V_{\text{BL,max}}$ for conventional system) for both architectures due to the increased precharge energy. However, DIMA achieves enormous EDP gains by amortizing this precharge energy over the access and processing of $B \times N_{\text{COL}}$ bits compared to N_{COL}/L bits in the digital architecture. In doing so, DIMA sacrifices the SNR of its computations as discussed next.

III. Modeling the SNR of DIMA Computations

DIMA provides significant EDP gains, but these gains are obtained at the expense of the SNR of its analog computations (*compute SNR*). This section presents noise and distortion models of DIMA computations in order to relate its energy and delay to its compute SNR, and hence to the accuracy of ML algorithms realized on DIMA.

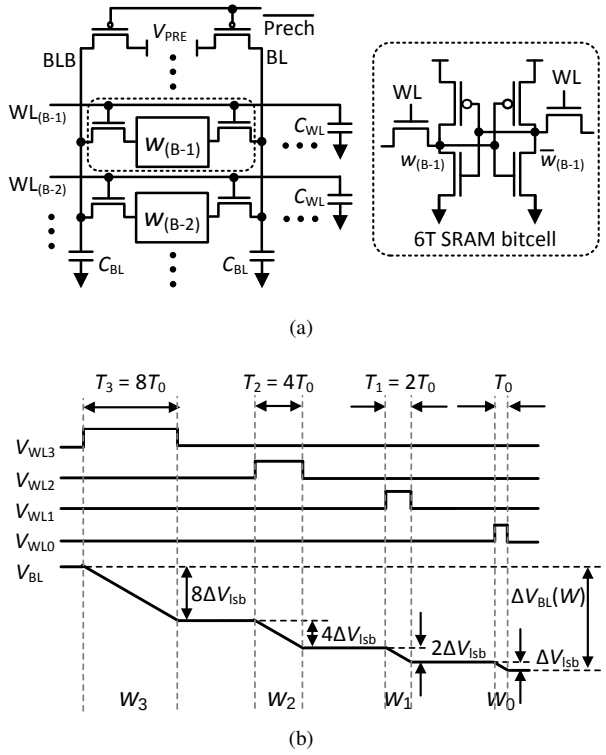


Fig. 3. The functional read (FR) operation: (a) column architecture and bitcell, and (b) idealized FR using pulse-width modulated (PWM) WL access pulses during a $B = 4$-bit word $W = 0000b'$ read-out. In practice, the WL pulses overlap in time but are shown non-overlapped to enhance clarity.

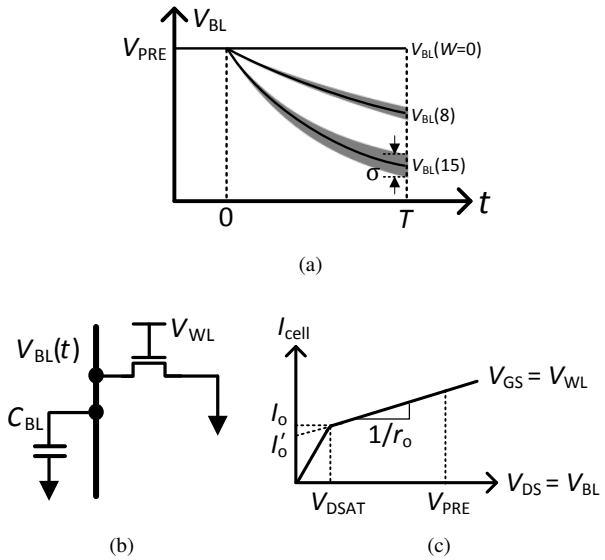


Fig. 4. Modeling BL discharge in the FR stage: (a) BL voltage $V_{\rm BL}(t)$, (b) a simplified RC model, and (c) cell discharge current $I_{\rm cell}$.

A. Basic Functional Read (FR)

The FR stage generates a BL voltage drop $\Delta V_{\rm BL}(W)$ proportional to *data value* $W = \Sigma_{i=0}^{B-1} 2^i w_i$ of a column-major stored *data bit-vector* $\mathbf{w} = [w_0, ..., w_{B-1}]$ with $w_i \in \{0, 1\}$ (see Fig. 3(a)). This is done via *simultaneous* application of B access pulses with pulse widths T_i $(i \in \{0, \ldots, B-1\})$ one each on the B WLs per precharge cycle. The resulting BL voltage drop $\Delta V_{\rm BL}(W)$ is given by:

$$\Delta V_{\rm BL}(W) = \frac{\Delta Q_{\rm BL}}{C_{\rm BL}} = \frac{I_{\rm cell}}{C_{\rm BL}} \sum_{i=0}^{B-1} T_i w_i \tag{6}$$

where $I_{\rm cell} = V_{\rm PRE}/R_{\rm BL}$ is the average bitcell discharge current, $\Delta Q_{\rm BL} = \sum_{i=0}^{B-1} \Delta Q_i$ is the total charge drawn from $C_{\rm BL}$ by the B bitcells within the total discharge time $T = \max_i \{T_i\}$, and w_i is the ith bit of W. In this way, the FR stage generates a dot-product of a time-coded input vector $\mathbf{t} = (T_0, \ldots, T_{B-1})$ and the data bit-vector $\mathbf{w}$.

If the WL access pulse widths T_i are constrained to be binary-weighted (see Fig. 3(b)) as in [16]–[18], i.e., $T_i = 2^i T_0$, then (6) transforms to

$$\Delta V_{\rm BL}(W) = \frac{I_{\rm cell} T_0}{C_{\rm BL}} \sum_{i=0}^{B-1} 2^i w_i \propto W \tag{7}$$

with the total discharge $T = (2^B - 1)T_0$ thereby accomplishing a crude B-bit D-to-A conversion of the data bit-vector $\mathbf{w}$. The maximum voltage discharge on the BL is denoted as $\Delta V_{\rm BL,max} = (2^B - 1)\Delta V_{\rm BL}(1)$.

The expression for the FR stage output $\Delta V_{\rm BL}(W)$ in (6) assumes a cell discharge current $I_{\rm cell}$ that is both spatially (across bitcells in a column) and temporally invariant. In practice, the access transistor threshold voltage $V_{\rm t}$ varies across bitcells and channel length modulation (CLM) results in spatio-temporal variations in $I_{\rm cell}$. These effects on $\Delta V_{\rm BL}(W)$ are quantified in the next two subsections.

B. Temporal Variations (Distortion)

The discharge current $I_{\rm cell}$ varies over time due to CLM thereby leading to deterministic but non-linear mapping from W to $\Delta V_{\rm BL}(W)$ at the functional read output. This deterministic non-linearity is referred to as distortion and is fixed for a specific die. Therefore, such errors can be overcome by tuning the model parameters (W) before storing them in memory, e.g., via the off-chip [27], [28] or on-chip [18] training of the inference model.

The impact of temporal variations on $\Delta V_{\rm BL}(W)$ (Fig. 4(a)) can be evaluated by assuming a simplified circuit model of the a single bitcell discharging the BL as shown in Fig. 4(b).

Assuming that the source node of the access transistor is close to 0 V during the discharge process, the cell discharge current can be approximated as (see Fig. 4(c))

$$I_{\rm cell}(t) = I_{\rm o} + \frac{V_{\rm BL}(t) - V_{\rm DSAT}}{r_{\rm o}} = I'_{\rm o} + \frac{V_{\rm BL}(t)}{r_{\rm o}} \tag{8}$$

where $I_{\rm o} = k'_{\rm n}(V_{\rm WL} - V_{\rm t})^{\alpha}$ is the current when the access transistor is at the edge-of-saturation, $1 \leq \alpha \leq 2$, and $V_{\rm DSAT}$ is the saturation drain-to-source voltage, $r_{\rm o}$ is the output resistance of the access transistor, and $I'_{\rm o} = I_{\rm o} - V_{\rm DSAT}/r_{\rm o}$ as shown in Fig. 4(c).

Employing (8) and solving the capacitor current-voltage differential equation below:

$$C_{\rm BL} \frac{dV_{\rm BL}(t)}{dt} + I'_{\rm o} + \frac{V_{\rm BL}(t)}{r_{\rm o}} = 0 \tag{9}$$

with initial conditions $V_{\rm BL}(0) = V_{\rm PRE}$, we first obtain an expression for $V_{\rm BL}(t)$, and then substituting it in the relationship $\Delta V_{\rm BL}(t) = V_{\rm PRE} - V_{\rm BL}(t)$, we obtain:

$$\Delta V_{\rm BL}(t) = (V_{\rm PRE} + I'_{\rm o} r_{\rm o})(1 - e^{-\frac{t}{\tau}}) \tag{10}$$

where $\tau = r_{\rm o} C_{\rm BL}$ is the RC time-constant of the discharge path. A large time-constant τ is desirable in order to reduce the impact of temporal distortion. Increasing τ by tuning $r_{\rm o}$, e.g., by adjusting its dimensions, is limited by the tight constraints on transistor sizing in a bitcell. Though $r_{\rm o}$ can be tuned by adjusting this gate bias $V_{\rm WL}$, the gate-bias tunability is constrained by other considerations such as controlling the $\Delta V_{\rm BL,max}$. On the other hand, temporal linearity improves with an increase in the number of rows $N_{\rm ROW}$ of the BCA because the BL capacitance $C_{\rm BL} = C_{\rm BLC} N_{\rm ROW}$ ($C_{\rm BLC}$ is BL capacitance per cell) increases with the number of rows in the BCA. Thus, the discharge curves in Fig. 4(a) shift up as $N_{\rm ROW}$ increases.

Assuming the discharge time $t \ll \tau$ and $r_{\rm o}$ is a constant, we approximate (10) via Taylor series expansion as:

$$\Delta V_{\rm BL}(t) = (V_{\rm PRE} + I'_{\rm o} r_{\rm o})(t/\tau) \tag{11}$$

Substituting $t = T_0 \sum_i 2^i w_i$ in (11) results in:

$$\Delta V_{\rm BL}(W) = \frac{(I_{\rm o} + I_{\rm CLM}) T_0}{C_{\rm BL}} \sum_{i=0}^{B-1} 2^i w_i \tag{12}$$

where $I_{\rm CLM} = (V_{\rm PRE} - V_{\rm DSAT})/r_{\rm o}$. Note that (12) has a form similar to (6). Therefore, we write (12) as:

$$\Delta V_{\rm BL}(W) = \frac{I_{\rm cell} T_0}{C_{\rm BL}} \sum_{i=0}^{B-1} 2^i w_i + g(W,T) \tag{13}$$

where $g(W,T)$ represents the distortion depending on the data W and total discharge time T. In practice, the distortion $g(W,T)$ is measured by calculating the percentage difference between an ideal straight line transfer function of the block and the mean of the measured values over a large number of cells.

Extracting the values for the parameters in (8)–(11) for a 65 nm CMOS process, we find that $k'_n = 220\,\mu$A/V^2, $r_{\rm o} = 74\,$kΩ, $I_{\rm o} = 18.9\,\mu$A, $V_t = 0.4\,$V, $C_{\rm BL} = 270\,$fF, $V_{\rm DSAT} = 0.2\,$V, $\alpha = 1.8$, and $T_0 = 300\,$ps. Therefore, if $V_{\rm BL}(t)$ is allowed to range from 0.5 V to 1 V, then it is straightforward to show that $I_{\rm cell} = I_{\rm o} + I_{\rm CLM}$ in (6) varies up to 26 % over the total discharge time of 4.5 ns. Furthermore, approximating (10) from (11) via Taylor series results in $g(W,T)$ being at most 12 % for these values.

C. Spatial Variations (Noise)

In contrast, spatial variations manifest themselves as spatially distributed noise, i.e., $\Delta V_{\rm BL}(W)$ in (6) can be written as:

$$\Delta V_{\rm BL}(W) = \frac{I_{\rm cell}}{C_{\rm BL}} \sum_{i=0}^{B-1} T_i w_i + g(W,T) + \eta_{\rm BL} \tag{14}$$

where $g(W,T)$ is the distortion as in (13) and $\eta_{\rm BL}$ is the spatial noise contribution. The statistics of this noise can be obtained by substituting $I_{\rm o} = k'_n (V_{\rm WL} - V_t)^\alpha$ into (12) (ignoring $I_{\rm CLM}$ for simplicity) as follows:

$$\Delta V_{\rm BL}(W) \propto \frac{k'_n (V_{\rm WL} - V_t)^\alpha}{C_{\rm BLC} N_{\rm ROW}} \sum_{i=0}^{B-1} T_i w_i \tag{15}$$

where $V_{\rm WL}$, the WL access voltage, is a critical variable in designing DIMA since it controls the discharge path resistance $r_{\rm o}$ and the current. In fact, the impact of V_t variations on the discharge current $I_{\rm cell}$ and hence $\Delta V_{\rm BL}(W)$ increases as $V_{\rm WL}$ approaches V_t. For example, the normalized standard deviation $\left(\frac{\sigma}{\mu}\right)_{\Delta V_{\rm BL,max}}$ of $\Delta V_{\rm BL,max}$ increases from 6 % to 29 % by reducing $V_{\rm WL}$ as shown in Table I. On the other hand, increasing $V_{\rm WL}$ can lead to read upsets (destructive read) in cells storing a '1' if $\Delta V_{\rm BL,max}$ becomes excessively large, e.g., $\Delta V_{\rm BL,max} > 0.7 V_{\rm PRE}$.

One way to address this issue is to set T_0 to the smallest value such that the rise and fall times of the WL pulses are a small fraction of T_0. We choose the rise and fall times to be $< 0.2 T_0$ to ensure that the sum of the rise and fall times is less than half LSB, thereby setting $T_0 \approx 200\,$ps to 300 ps. Once T_0 and B (number of active rows) are fixed, then $\Delta V_{\rm BL,max}$ can be controlled by tuning $V_{\rm WL}$ within the upper limit set by destructive read considerations. Note that this upper limit on $V_{\rm WL}$ can be increased by increasing the number of rows $N_{\rm ROW}$ and hence the BL capacitance (see Fig. 5). This is because the BLs discharge slower due to their higher BL capacitance thereby allowing one to increase $V_{\rm WL}$ which leads to reduced impact of process variations at a given $\Delta V_{\rm BL}(W)$.

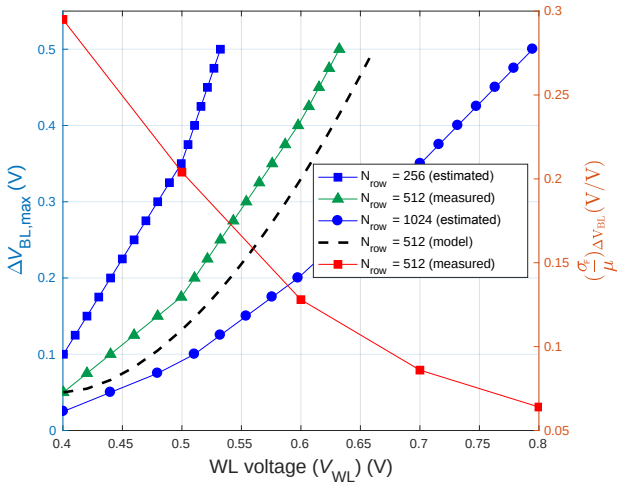


Fig. 5. BL swing vs. BL noise ($\sigma_{\rm F}/\mu$) with respect to WL voltage $V_{\rm WL}$. The noise ($\sigma_{\rm F}/\mu$) and BL swing with $N_{\rm ROW} = 512$ is measured from silicon prototype [16]. BL swings with $N_{\rm ROW} = 256$ and 1024 are estimated from (15) and the measured BL swing with $N_{\rm ROW} = 512$.

D. Noise and Distortion Models

This section presents two compute-intensive distance metrics employed pervasively in ML algorithms - the dot-product and the sum-of-absolute difference (SAD) - in the presence of DIMA's non-ideal behavior described in Sections III-B (distortion) and III-C (noise).

An N-dimensional dot-product computation is given by:

$$y = \mathbf{w}^{\mathsf{T}}\mathbf{x} = \sum_{i=1}^{N} W_i X_i \tag{16}$$

where $\mathbf{w} = (W_1, \ldots, W_N)$ and $\mathbf{x} = (X_1, \ldots, X_N)$ are two N-dimensional real-valued vectors of precision B_W and B_X, respectively. A digital architecture will realize (16) via N $B_X \times B_W$-b multiply-accumulate (MAC) operations with quantization noise as the primary source of non-ideal behavior.

In DIMA, due to its mixed-signal attribute, the dot product (16) is computed as (assuming a fixed total discharge time T):

$$\widehat{y} = \frac{1}{N}\sum_{i=1}^{N}(W_i + g(W_i, T) + \eta_{\text{wF},i})X_i + \eta_{\text{yB}} + \eta_{yC} \tag{17}$$

$$= y + \eta_{\text{y}} \tag{18}$$

where $\eta_{\text{wF},i}$ is the spatial noise (variance σ^2_{wF}) at the ith column due to FR, η_{yB} (variance σ^2_{yB}) and η_{yC} (variance σ^2_{yC}) are the noise contributions from the BLP and CBLP, respectively, as seen at the output y, and η_{y} is the composite of noise and distortion on y at the CBLP output, respectively (see Fig. 1(b)).

We map the FR output in (18) into the voltage domain in order to relate its algebraic computation to the circuit realization in DIMA, as follows:

$$\widehat{s} = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta V_{\text{BL}}(W_i) + \eta_{\text{F},i}\right)X_i + \eta_{\text{B}} + \eta_{\text{C}} \tag{19}$$

where $\Delta V_{\text{BL}}(W_i) \in \{0, \delta, \ldots, \Delta V_{\text{BL,max}}\}$ denotes the voltage swing corresponding to the ith weight W_i with $\Delta V_{\text{BL,max}} = (2^B - 1)\delta$. The η_{F}, η_{B}, and η_{C} are the non-idealities in the voltage domain including both distortion and noise contributions from FR, BLP, and CBLP stages, respectively. $\widehat{s}$ denotes DIMA's output in the presence of the non-idealities.

Table I quantifies both distortion and noise contributions, which indicates that the noise variances $\sigma^2_{\text{F}} \gg \sigma^2_{\text{B}} \gg \sigma^2_{\text{C}}$, i.e., FR noise dominates. This is to be expected since FR processing involves discharge of a BL via minimum-sized transistors in the bitcell and operates close to near-threshold voltage regime with low V_{WL} thereby incurring large spatial mismatch as described in Section III-C. Therefore, in the rest of this paper, we will focus on the noise contributions from the FR stage.

Similarly, the ideal SAD computation between $\mathbf{w}$ and $\mathbf{x}$, and its noisy (DIMA) version is given by:

$$y = \frac{1}{N}\sum_{i=1}^{N}|W_i - X_i| \tag{20}$$

$$\widehat{y} \approx \frac{1}{N}\sum_{i=1}^{N}|W_i - X_i| + 2\eta_{\text{yF}} = y + 2\eta_{\text{yF}} \tag{21}$$

where the factor of 2 in (21) appears since both W_i and X_i are read using FR [16]. Mapping (21) to the voltage domain gives:

$$\widehat{s} \approx \frac{1}{N}\sum_{i=1}^{N}(|\Delta V_{\text{BL}}(W_i) - \Delta V_{\text{BL}}(X_i)| + 2\eta_{\text{F},i}) \tag{22}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(s_i + 2\eta_{\text{F},i}) = s + \eta \tag{23}$$

where $|\Delta V_{\text{BL}}(W_i) - \Delta V_{\text{BL}}(X_i)| \in \{0, \delta, \ldots, \Delta V_{\text{BL,max}}\}$ denotes the voltage swing corresponding to the ith absolute difference $|W_i - X_i|$ with $\Delta V_{\text{BL,max}} = (2^B - 1)\delta$, $\widehat{s}$ denotes DIMA's output in the presence of equivalent FR noise η with variance $2\sigma^2_{\text{F}}/N$ at the CBLP output, and s is the voltage-domain version of the ideal SAD output y in (20).

TABLE I
NOISE AND DISTORTION IN DIMA STAGES [16].

Error type	FR (η_{F})	BLP (η_{B})		CBLP (η_{C})
		DP	SAD	
% distortion (μ)	$2.6^{(1)}$	$2.1^{(1)}$	$2.5^{(1)}$	$0.8^{(3)}$
% noise (σ/μ)	(6 -to- 29)$^{(2)}$	$2.8^{(2)}$	$3.2^{(2)}$	$0.2^{(3)}$

Row 1: obtained as an average over all 16 4-b data values.
Row 2: obtained for maximum discharge $\Delta V_{\text{BL,max}}$.
(1) silicon measured; (2) Monte Carlo simulations with $0.4\,\text{V} \leq V_{\text{WL}} \leq 0.8\,\text{V}$; (3) estimated from the capacitor sizes in [16].

IV. Prediction of Inference Accuracy

In this section, DIMA's noise models from Section III are employed to predict its system-level classification accuracy. DIMA's accuracy is compared with that of the digital architecture operated at the same voltage swing per bit $\Delta V_{\text{BL-B}} = \frac{\Delta V_{\text{BL,max}}}{B}$. In this way, the BL discharge energy per bit is made identical for both architectures. Two different tasks are considered: 1) template matching (TM) using the SAD kernel and, 2) SVM using the dot product kernel.

A. Template Matching

The TM algorithm computes the SADs between a query input $\mathbf{x}$ and a set of M candidate images $\left\{\mathbf{w}^{(0)}, \ldots, \mathbf{w}^{(M-1)}\right\}$ and outputs the index corresponding to the one with minimum SAD, as shown below:

$$j^* = \arg\min_j y^{(j)} \tag{24}$$

where $y^{(j)}$ represents the SAD between $\mathbf{x}$ and the jth candidate image $\mathbf{w}^{(j)}$ as in (20).

1) Digital Architecture: In the digital architecture, spatial mismatch and the low $\Delta V_{\text{BL-B}}$ can result in bit flipping errors caused by insufficient input swing to the sense amplifiers. Hence, $y^{(j)}$ is computed as

$$\widehat{y}^{(j)} = \frac{1}{N}\sum_{i=1}^{N}(y_i^{(j)} + e_i) = y^{(j)} + \overline{e} \tag{25}$$

where $e_i \in \{-2^B + 1, \ldots, 2^B - 1\}$ denotes the numerical error in ith element and $\overline{e}$ denotes the sample mean of e_i:

$$\overline{e} = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{26}$$

where $\mathbb{E}(e_i) = 0$ and the variance $\text{Var}(e_i)$ is given by:

$$\text{Var}(e_i) = \sigma_e^2 = \left(\frac{4^B - 1}{3}\right) p \tag{27}$$

where p denotes the bit error probability which depends on $\Delta V_{\text{BL-B}}$ as follows [10]:

$$p = Q\left(\frac{\Delta V_{\text{BL-B}}}{\sigma_{\text{read}}}\right) \tag{28}$$

where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$, and σ_{read} is the standard deviation of a composite noise source that includes spatial variations across bit-cells during the BL discharge and the SAs input offset [29]. By the Central Limit Theorem, $\overline{e}$ can be modeled as a Gaussian for large N, i.e.,

$$\overline{e} \sim \mathcal{N}\left(0, \left(\frac{4^B - 1}{3}\right) \cdot \frac{p}{N}\right) \tag{29}$$

where $\text{Var}(\overline{e}) = \frac{\sigma_e^2}{N}$.

In the absence of bit flips, given two candidate images $\mathbf{w}^{(k)}$ and $\mathbf{w}^{(l)}$ and their respective SAD outputs $y^{(k)}$ and $y^{(l)}$, if $y^{(k)} < y^{(l)}$, the correct decision is $j^* = k$ since $\mathbf{w}^{(k)}$ is closer to query image $\mathbf{x}$ than $\mathbf{w}^{(l)}$. However, in the presence of bit flips, it may turn out that $\widehat{y}^{(k)} > \widehat{y}^{(l)}$ in which case the incorrect decision $j^* = l$ will be output by the digital architecture. The mismatch probability that $\mathbf{w}^{(l)}$ is incorrectly chosen is given by (see Appendix A):

$$p_{\text{m-digital}}^{(k\to l)} = Q\left(\alpha_{\text{TM,digital}}^{(k,l)} \sqrt{\frac{3(2^B - 1)}{2(2^B + 1)} \cdot \frac{N}{p}}\right) \tag{30}$$

where $0 \le \alpha_{\text{TM, digital}}^{(k,l)} \le 1$ denotes the normalized decision margin given by:

$$\alpha_{\text{TM,digital}}^{(k,l)} = \frac{\left|y^{(k)} - y^{(l)}\right|}{2^B - 1} \tag{31}$$

where $2^B - 1$ represents the maximum of the difference $|y^{(k)} - y^{(l)}|$.

Without loss of generality, suppose $j^* = 0$, i.e., $y^{(0)}$ has the minimum SAD. Then, the detection (accuracy) probability of the digital architecture is given by:

$$P_{\text{det-digital}} = \prod_{m=1}^{M-1} (1 - p_{\text{m-digital}}^{(0\to m)}) \tag{32}$$

where $p_{\text{m-digital}}^{(0\to m)}$ can be obtained from (30). We observe that $P_{\text{det-digital}}$ improves (increases) with the decision margin $\alpha_{\text{TM,digital}}^{(i,j)}$ and the vector dimension N, but worsens (reduces) if bit error probability p increases ($\Delta V_{\text{BL-B}}$ reduces).

2) DIMA: Consider two images $\mathbf{w}^{(k)}$ and $\mathbf{w}^{(l)}$ and their ideal voltage domain SAD outputs being $s^{(k)}$ and $s^{(l)}$, respectively. If $s^{(k)} < s^{(l)}$ then the correct decision is $j^* = k$. However, in the presence of noisy DIMA computations described by (23), it is possible that $\widehat{s}^{(k)} > \widehat{s}^{(l)}$, in which case an incorrect decision $j^* = l$ will be made. The mismatch probability that DIMA incorrectly chooses $\mathbf{w}^{(l)}$ (see Appendix A), is given by:

$$p_{\text{m-DIMA}}^{(k\to l)} = Q\left(\alpha_{\text{TM,DIMA}}^{(k,l)} \sqrt{\frac{N \cdot \text{SNR}_{\text{DIMA}}}{2}}\right) \tag{33}$$

where the decision margin $0 \le \alpha_{\text{TM,DIMA}}^{(k,l)} \le 1$ in the voltage domain is given by:

$$\alpha_{\text{TM,DIMA}}^{(k,l)} = \frac{\left|s^{(k)} - s^{(l)}\right|}{\Delta V_{\text{BL,max}}} \tag{34}$$

which is equivalent to (31), and SNR_{DIMA} is defined as:

$$\text{SNR}_{\text{DIMA}} = \frac{\Delta V_{\text{BL,max}}^2}{\sigma_{\text{F}}^2}. \tag{35}$$

As in the derivation of (32), the DIMA's detection probability of DIMA is given by:

$$P_{\text{det-DIMA}} = \prod_{m=1}^{M-1} (1 - p_{\text{m-DIMA}}^{(0\to m)}) \tag{36}$$

where $p_{\text{m-DIMA}}^{(0\to m)}$ can be obtained from (33). Note that $P_{\text{det-DIMA}}$ improves with the decision margin $\alpha_{\text{TM}}^{(i,j)}$ and the dimension N, and with a sufficiently large N, accurate decisions can be made even in a low-SNR regime.

B. Support Vector Machine

A binary SVM computes the sign of the dot product as follows:

$$\text{sign}(s) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x}) \tag{37}$$

where the weight vector $\mathbf{w}$ is chosen to maximize the margin between the decision hyperplane and the input vectors in the training set. For ease of analysis, we assume that $\mathbf{w} = (W_1, \ldots, W_N)$ in (37) denotes the normalized weight vector, i.e., $0 \le |W_i| \le 1\ \forall i$, and we omit the bias term.

1) Digital Architecture: Bit flips in a digital architecture results in the dot product computation of (37) being transformed to:

$$\widehat{y} = \sum_{i=1}^{N} (W_i + e_i) X_i = y + \widetilde{e} \tag{38}$$

where W_i is distorted to $W_i + e_i$ due to bit flips, e_i denotes the numerical error as in (25), and $\widetilde{e}$ denotes the weighted sum of e_i, i.e.,

$$\widetilde{e} = \sum_{i=1}^{N} X_i e_i. \tag{39}$$

The mismatch probability of the digital architecture is given by:

$$\begin{aligned} p_{\text{m-digital}} &= \Pr(\text{sign}(y) \ne \text{sign}(\widehat{y})) \\ &= Q\left(\frac{N\alpha_{\text{SVM}}}{\sigma_e}\right) \end{aligned} \tag{40}$$

where $\sigma_e^2 = \left(\frac{4^B - 1}{3}\right) p$ as in (27), and the normalized decision margin α_{SVM} is given by:

$$\alpha_{\text{SVM}} = \frac{1}{N} \cdot \left| \sum_{i=1}^{N} W_i \cdot \frac{X_i}{\|\mathbf{x}\|} \right|. \tag{41}$$

Note that α_{SVM} depends on the only trained weights and the input data vector (see Appendix B for details).

TABLE II
DESIGN AND MODEL PARAMETERS.

Parameter	Values	Parameter	Values
V_{DD}	1 V	V_{WL}	0.4 − 0.9 V
V_{PRE}	1 V	L	4
N_{ROW}	256 - 1024	N_{COL}	256
T_0	300 ps	N	128 - 1024
B	8	M	64

2) DIMA: DIMA's SVM computation is well-modeled by (19) with signal s and noise η terms given by:

$$s = \frac{1}{N}\sum_{i=1}^{N} \Delta V_{\text{BL}}(W_i) X_i \tag{42}$$

$$\eta = \frac{1}{N}\sum_{i=1}^{N} X_i \eta_{\text{F},i} \tag{43}$$

where $\eta_{\text{F},i} \sim \mathcal{N}(0, \sigma_{\text{F}}^2)$. The mismatch probability of DIMA can be shown (see Appendix B) to be:

$$p_{\text{m-DIMA}} = Q\left(N\alpha_{\text{SVM}}\sqrt{\text{SNR}_{\text{DIMA}}}\right) \tag{44}$$

where $\text{SNR}_{\text{DIMA}} = \frac{\Delta V_{\text{BL,max}}^2}{\sigma_{\text{F}}^2}$.

C. Experimental Model Validation

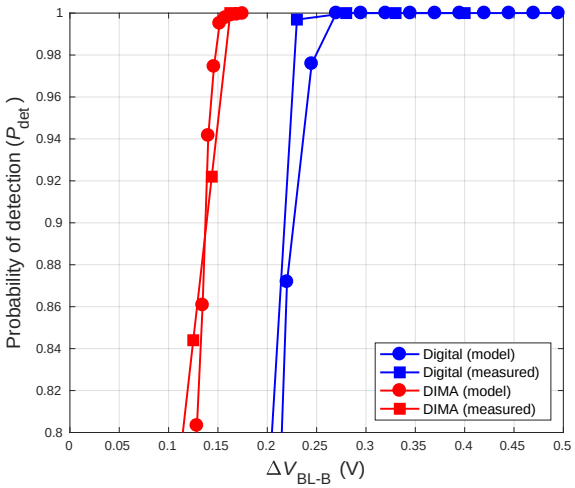
In this section, we validate the system-level accuracy prediction models (32), (36), (40), and (44) by comparing their predictions of the detection probability P_{det} with silicon measured results in [16]. We consider the MIT CBCL dataset with the design parameters listed in Table II, and σ/μ with respect to BL swing per bit in Fig. 5 to evaluate the accuracy prediction models.

For the TM algorithm, one of the 64 candidate images was chosen as the template. The accuracy P_{det} is calculated by averaging the 64 detection probabilities $P_{\text{det-t}}$ (one per template), where $P_{\text{det-t}}$ is obtained by counting the the number of correct detections in multiple (>1000) trials. For SVM, 800 query images (400 faces and 400 non-faces) are tested for face detection task and the overall P_{det} is obtained by averaging the 800 query-specific P_{det} values. Various values of the decision margin $\alpha_{\text{TM or SVM},k\to l}$ are tried to better evaluate modeling accuracy.

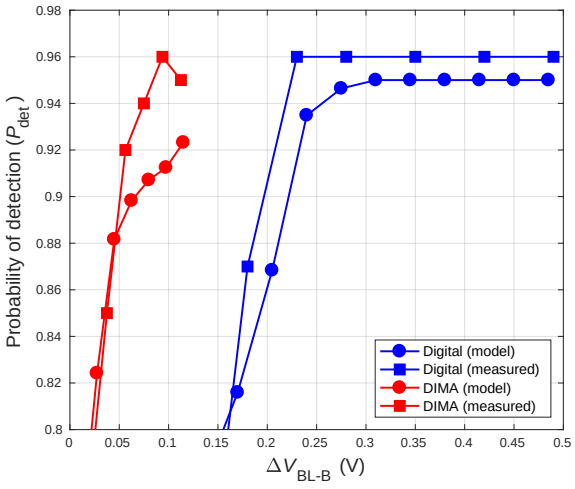
Figure 6 shows that the predictions of system-level accuracy from ((32) and (36)) (TM) and ((40) and (44)) (SVM) match very well with values obtained from silicon [16] with a modeling error of $< 10.5\%$. In general, the model estimates of accuracy are worse than the measured accuracy because these models consider the worst case scenario.

V. FUNDAMENTAL TRADE-OFFS AND LIMITS

In this section, we study the fundamental trade-offs between energy-efficiency and accuracy as a function of input vector dimension (N), the decision margins (α_{TM} and α_{SVM}), the array size (N_{ROW}) for both the digital architecture and DIMA.



(a)



(b)

Fig. 6. Probability of detection (P_{det}) vs. measured results from [16] with $N_{\text{ROW}} = 512$ for: (a) TM, and (b) SVM. The voltage discharge $\Delta V_{\text{BL-B}}$ is swept by tuning V_{WL}, choosing σ_{F} from Fig. 5 and σ_{read} from measurements. Note that $\Delta V_{\text{BL,max}}$ is maintained to be less than $0.7V_{\text{PRE}}$ to avoid destructive read, e.g., $\Delta V_{\text{BL-B}}$ <0.18 V for DIMA and <0.7 V for the digital architecture.

Furthermore, we identify conditions under which DIMA provides significant EDP gains over a digital architecture, and those that limit its accuracy.

The conventional digital architecture comprises an SRAM of the same size as the one in the DIMA prototype [16] with 4:1 column muxing ($L = 4$), and a digital block synthesized separately for realizing an SVM and a TM. Then, the energy and delay of the digital block were estimated from post-layout simulations. The energy and delay of the SRAM in the conventional system were measured from the DIMA prototype [16] operating in the conventional read mode.

The analysis in this section employs typical parameters values listed in Table II. The value of $T_0 = 300\,\text{ps}$ is chosen to prevent destructive read while ensuring that WL access pulses with the rise and fall times of $< 0.2T_0$ can be easily realized. The value of $\Delta V_{\text{BL,max}}$ is set by tuning V_{WL} to be less

than $0.7V_{\text{PRE}}$ to avoid destructive read, e.g., $\Delta V_{\text{BL-B}}$ <0.18 V for DIMA and <0.7 V for the digital architecture. Numerical values for each term in (2) and (3) are obtained from circuit simulations in a 65 nm CMOS process technology. The total energy estimates from (2) and (3) are validated by comparing against those measured from the IC prototype in [16].

A. Energy Efficiency vs. Accuracy Trade-offs

Figure 7 shows that DIMA achieves the same accuracy as the digital architecture but an energy-per-decision cost that is lower by approximately $10\times$ for most operating conditions. Coupled with a latency reduction of $5\times$-to-$20\times$ (see Section II-B), DIMA can achieves a decision-making EDP reduction of $50\times$-to-$200\times$ over a digital architecture, of which close to $100\times$ EDP reduction has already been achieved [18].

Figure 7 also shows that the accuracy P_{det} improves with decision margin $\alpha_{\text{TM or SVM}}$ and input vector dimension N for the same BL swing $\Delta V_{\text{BL-B}}$. However, if the decision energy cost is kept fixed, then accuracy in fact reduces when the input dimension N is increased.

Finally, unlike the digital architecture, DIMA's accuracy is seen to be limited when $\alpha_{\text{TM/SVM}}$ and N are small, e.g., $\alpha_{\text{TM}} = 0.05$, $\alpha_{\text{SVM}} = 0.2$ and $N = 128$. This is because DIMA's analog computations lead to various non-idealities introduced in Section III which have a greater impact on the accuracy of inference when $\alpha_{\text{TM/SVM}}$ and N are small.

B. Impact of Array Size

The number of columns N_{COL} in the bitcell array is limited only by constraints on the rise and fall times of the WL access pulses, and by the available area. However, the number of rows N_{ROW} directly impacts the system-level accuracy and energy-efficiency. This is because the BL capacitances increase in proportion to N_{ROW} requiring a higher value of the WL access pulse voltage V_{WL} to obtain the same $\Delta V_{\text{BL-B}}$. A higher value of V_{WL} implies that the impact of transistor threshold voltage variations on the discharge current is reduced leading to improved accuracy (see Fig. 8(a)) but at the cost of higher energy consumption (see Fig. 8(b)) for both DIMA and the digital architecture. Conversely, this also implies that N_{ROW} needs to be sufficiently large for DIMA to achieve an accuracy comparable to that of a digital architecture, e.g., Fig. 8 also shows that DIMA is unable to achieve maximum accuracy for $N_{\text{ROW}} = 256$. Though Fig. 8 shows these trends for SVM, similar trends were observed for TM as well.

C. DIMA Design Space

Based on the results presented in the previous sections, the following conclusions can be drawn regarding the conditions under which DIMA will perform favorably over a digital architecture:

- for a specific array size, DIMA has a minimum BL swing $\Delta V_{\text{BL-B}}$ to achieve maximum accuracy, e.g., Fig. 6 shows this to be $\approx 100\,\text{mV}$ for DIMA and $\approx 250\,\text{mV}$ for the digital architecture for $N_{\text{ROW}} = 512$.
- DIMA's decision accuracy improves with a higher value of decision margin $\alpha_{\text{TM/SVM}}$ at the same decision energy due to the intrinsic robustness of the ML algorithm, and for all values of the decision margin, DIMA consumes approximately $10\times$ less energy than digital for the same decision accuracy (see Fig. 7).
- DIMA's decision accuracy can be improved by increasing N at the same $\Delta V_{\text{BL-B}}$ and decision margin, but at a higher energy cost, e.g., approximately $4\times$ more energy when increasing N from 128 to 512, which is still much less than that of the digital architecture at the same accuracy.
- DIMA is unable to achieve an accuracy comparable to a digital architecture when $\alpha_{\text{TM or SVM}}$, N and/or N_{ROW} are small, e.g., α_{TM} = 0.05 and $N = 128$ or when both N_{ROW} and N are small, e.g., $N_{\text{ROW}} = 256$ and $N = 128$.

Therefore, we can conclude that DIMA is favorable when the vector length N and the number of rows N_{ROW} are large enough for a classification task with moderate difficulty, i.e., moderate values of $\alpha_{\text{TM or SVM}}$. Since ML algorithms tend to have high inherent error immunity (large $\alpha_{\text{TM or SVM}}$) and require number of model parameters to be stored, one can expect that DIMA will continue to exhibit high decision-level EDP gains over the digital architecture in most scenarios.

D. Impact of Technology Scaling

In advanced CMOS process nodes, we expect to see improved energy and delay due to reduced capacitance and increased I_{cell}. However, advanced nodes also exhibit increased process variations which will be reflected in a higher σ/μ of $\Delta V_{\text{BL,max}}$ leading to a loss in accuracy. From (35) and (44), we find that this loss in accuracy can be recovered by increasing either N and/or $\Delta V_{\text{BL,max}}$ but at the expense of increased energy costs. Hence, it will be interesting to study how the decision-level EDP gains due to technology scaling are offset by the mechanisms to compensate for the corresponding loss in accuracy.

E. Extension for Other In-memory Architectures

Though this paper focuses on the DIMA as implemented in [16], [18], the models in this work can be re-purposed to analyze other in-memory architectures. For example, the architectures in [7], [15], [30] suffer from similar noise sources as this work, e.g., process variations during the BL discharge. Thus, the accuracy models provided in this paper are applicable. In addition, the SNR_{DIMA} in (44) can be modified to consider the dominant noise sources in the specific architecture, e.g., temporal noise in switched capacitors [31], ADC noise [32], comparator [33], and others, to cover those in-memory architectures.

VI. Conclusions

This paper analyzed an SRAM-based DIMA in terms of its decision energy, delay, and accuracy via theoretical modeling and analysis, validation with measured results of our earlier published silicon IC prototypes [16], [18], and identified

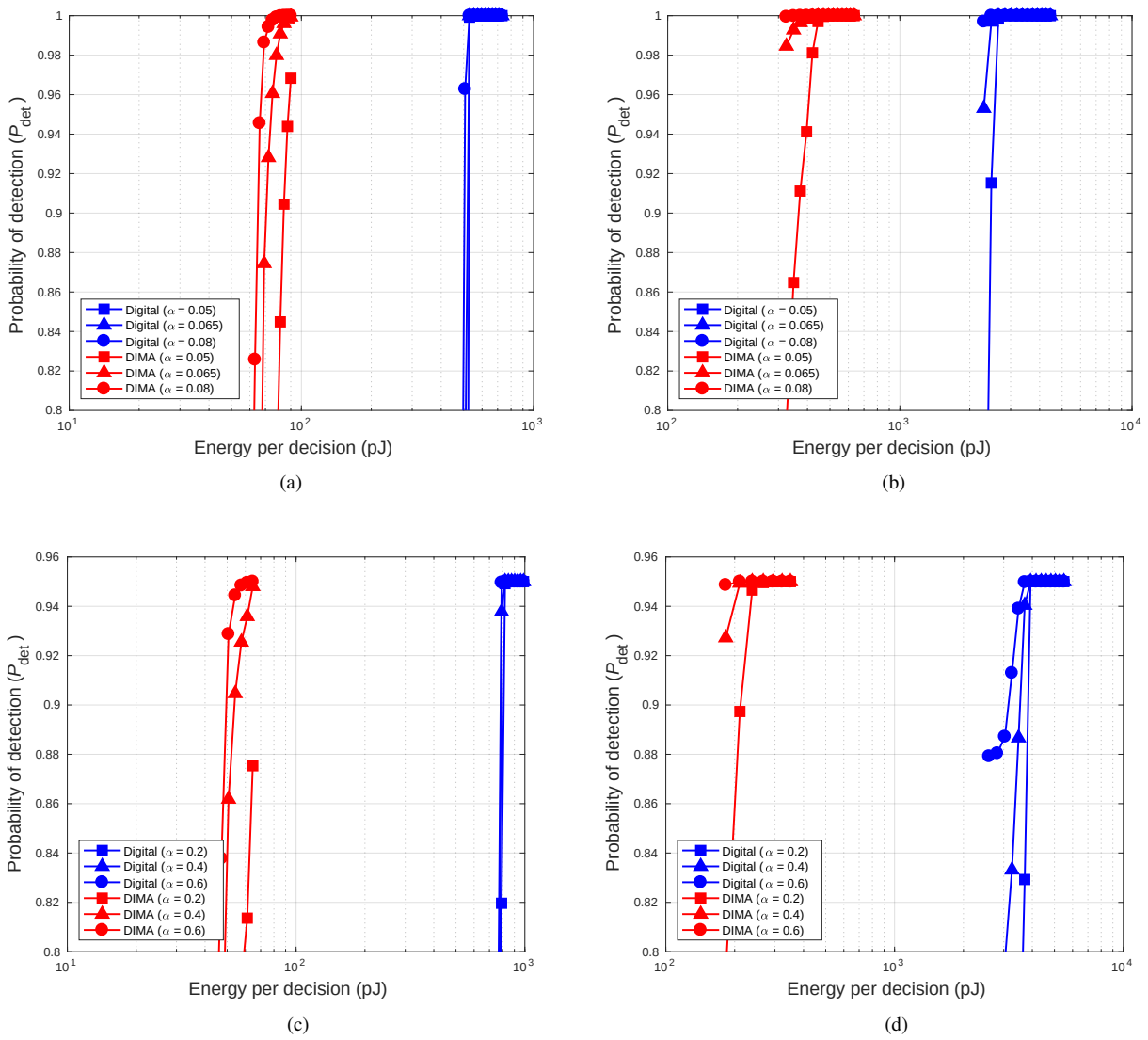


Fig. 7. System-level accuracy vs. energy-efficiency trade-offs obtained via accuracy prediction models ((32) and (36)) by sweeping $\Delta V_{\text{BL-B}}$ for TM with $\alpha_{\text{TM}} \in [0.05, 0.08]$ and vector dimensions (a) $N = 128$, (b) $N = 512$, and via ((40) and (44)) for SVM with $\alpha_{\text{SVM}} \in [0.2, 0.6]$ and vector dimensions (c) $N = 128$, (d) $N = 512$.

conditions under which DIMA will be superior to a digital architecture.

The DIMA energy, delay and accuracy models presented in this paper, can be employed to study the benefits of a DIMA-based system for other bitcell architectures (8T or 10T), in the context of new ML algorithms and applications, e.g., natural language processing using long-short term memory (LSTM) and/or process technologies (CMOS at smaller nodes, resistive RAM and magnetic RAM). Furthermore, given DIMA's regular structure, these models along with design principles can be encapsulated into platform design tools such as a DIMA memory compiler to automatically synthesize DIMA macros. The accuracy model indicates that larger vector length and decision margin allow analog operations in DIMA to be more accurate and energy-efficient. In addition, large number of rows in bitcell array leads to better application-level accuracy at the cost of degraded energy efficiency.

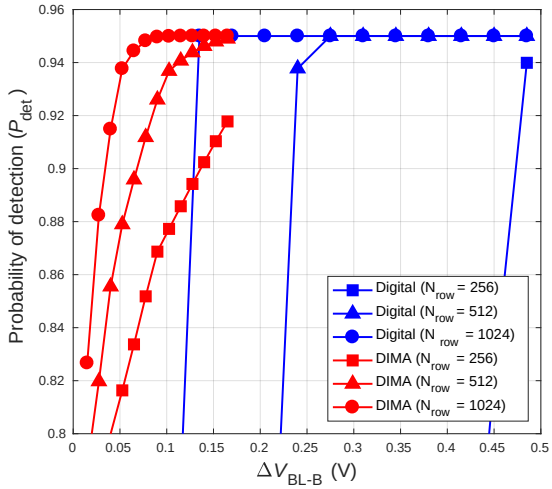
APPENDIX A
TEMPLATE MATCHING

A. Conventional Digital Architecture

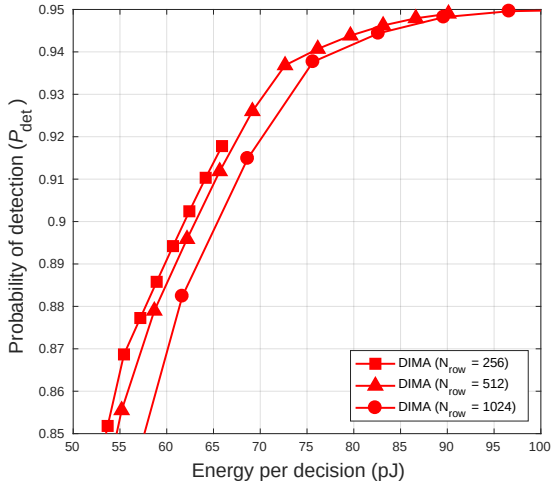
Suppose that there are two images $\mathbf{w}^{(k)}$ and $\mathbf{w}^{(l)}$ such that $y^{(k)} < y^{(l)}$. The mismatch probability that $\mathbf{w}^{(l)}$ is incorrectly chosen instead of $\mathbf{w}^{(k)}$ is given by

$$\begin{aligned} p_{\text{m-digital}}^{(k\to l)} &= P(\widehat{y}^{(l)} > \widehat{y}^{(k)}) \\ &= P(y^{(l)} - y^{(k)} < \overline{e}^{(k)} - \overline{e}^{(l)}) \end{aligned} \tag{45}$$

where $y^{(l)} - y^{(k)}$ denotes the decision margin and $\overline{e}^{(k)} - \overline{e}^{(l)}$ denotes the effective error in TM's digital computation. If the effective error is greater than the decision margin, then the mismatch occurs.



(a)



(b)

Fig. 8. Impact of number of rows in the bitcell array N_{ROW} on the system-level accuracy as a function of: (a) BL swing, and (b) energy per decision for SVM with $\alpha_{\text{SVM}} = 0.4$ and $N = 128$.

If $\overline{e}^{(k)}$ and $\overline{e}^{(l)}$ are independent, then $\left(\overline{e}^{(k)} - \overline{e}^{(l)}\right)$ can be modeled by Gaussian distribution as follows:

$$\left(\overline{e}^{(k)} - \overline{e}^{(l)}\right) \sim \mathcal{N}\left(0, 2 \cdot \left(\frac{4^B - 1}{3}\right) \cdot \frac{p}{N}\right). \quad (46)$$

and the mismatch probability is given by

$$p_{\text{m-digital}}^{(k\to l)} = Q\left(\frac{\left|y^{(k)} - y^{(l)}\right|}{\sqrt{\text{Var}(\overline{e}^{(k)} - \overline{e}^{(l)})}}\right) = Q\left(\sqrt{\frac{N}{p}\left(\frac{3}{2(4^B-1)}\right)} \cdot \left|y^{(k)} - y^{(l)}\right|\right). \quad (47)$$

Substituting (31) into (47) results in the final expression (30) for the mismatch probability:

$$p_{\text{m-digital}}^{(k\to l)} = Q\left(\alpha_{\text{TM,digital}}^{(k,l)} \sqrt{\frac{3(2^B-1)}{2(2^B+1)} \cdot \frac{N}{p}}\right). \quad (48)$$

B. DIMA

Suppose that there are two images $\mathbf{w}^{(k)}$ and $\mathbf{w}^{(l)}$ such that $s^{(k)} < s^{(l)}$. Similar to (47), we can derive the mismatch probability that $\mathbf{w}^{(l)}$ is incorrectly chosen instead of $\mathbf{w}^{(k)}$ by DIMA's TM computation as follows:

$$p_{\text{m-DIMA}}^{(k\to l)} = P(\widehat{s}^{(k)} > \widehat{s}^{(l)}) = P(s^{(l)} - s^{(k)} < \overline{n}^{(k)} - \overline{n}^{(l)}) = Q\left(\sqrt{\frac{N}{2\sigma_{\text{F}}^2}} \cdot \left|s^{(k)} - s^{(l)}\right|\right) \quad (49)$$

which comes from $\left(\overline{n}^{(k)} - \overline{n}^{(l)}\right) \sim \mathcal{N}\left(0, \frac{2\sigma_{\text{F}}^2}{N}\right)$. Substituting (34) into (49) results in final expression (33) for the mismatch probability shown below:

$$p_{\text{DIMA}}^{(k\to l)} = Q\left(\alpha_{\text{TM,DIMA}}^{(k,l)} \sqrt{\frac{N\Delta V_{\text{BL,max}}^2}{2\sigma_{\text{F}}^2}}\right) = Q\left(\alpha_{\text{TM,DIMA}}^{(k,l)} \sqrt{\frac{N \cdot \text{SNR}_{\text{DIMA}}}{2}}\right) \quad (50)$$

where $\text{SNR}_{\text{DIMA}} = \frac{\Delta V_{\text{BL,max}}^2}{\sigma_{\text{F}}^2}$.

Appendix B
Support Vector Machine

A. Conventional Digital Architecture

The mismatch probability for an SVM is given by

$$p_{\text{m-digital}} = \Pr(\text{sign}(y) \neq \text{sign}(\widehat{y})) = \Pr(|y| < \widetilde{e}). \quad (51)$$

Note that $\mathbb{E}(\widetilde{e}) = 0$ because of $\mathbb{E}(e_i) = 0$. The variance of $\widetilde{e}$ is given by

$$\text{Var}(\widetilde{e}) = \sum_{i=1}^{N} X_i^2 \cdot \sigma_e^2 = \|\mathbf{x}\|_2^2 \cdot \sigma_e^2 \quad (52)$$

where $\sigma_e^2 = \left(\frac{4^B-1}{3}\right) p$ as in (27).

By the Central Limit Theorem, we claim that

$$\widetilde{e} \sim \mathcal{N}\left(0, \|\mathbf{x}\|_2^2 \cdot \sigma_e^2\right). \quad (53)$$

Then, the mismatch probability is given by

$$p_{\text{m-digital}} = Q\left(\frac{|y|}{\sqrt{\text{Var}(\widetilde{e})}}\right) = Q\left(\frac{|y|}{\|\mathbf{x}\|_2 \sigma_e}\right) = Q\left(\frac{\left|\sum_{i=1}^{N} W_i \cdot \frac{X_i}{\|\mathbf{x}\|_2}\right|}{\sigma_e}\right) = Q\left(\frac{N \cdot \alpha_{\text{SVM}}}{\sigma_e}\right) \quad (54)$$

where the normalized decision margin is given by $\alpha_{\text{SVM}} = \frac{1}{N} \cdot \left|\sum_{i=1}^{N} W_i \cdot \frac{X_i}{\|\mathbf{x}\|_2}\right|$, which is the same as (40)–(41).

B. DIMA

Since $n_{\mathrm{F},i} \sim \mathcal{N}(0, \sigma_{\mathrm{F}}^2)$, we obtain $\mathbb{E}(\eta) = 0$ and the variance of η is given by

$$\mathrm{Var}(\eta) = \frac{1}{N^2} \sum_{i=1}^{N} |X_i|^2 \sigma_{\mathrm{F}}^2 = \frac{\|\mathbf{x}\|_2^2 \sigma_{\mathrm{F}}^2}{N^2}. \tag{55}$$

Hence, $\eta \sim \mathcal{N}\left(0, \frac{\|\mathbf{x}\|_2^2 \sigma_{\mathrm{F}}^2}{N^2}\right)$ and the mismatch probability of DIMA in (44) can be derived as

$$\begin{aligned} p_{\text{m-DIMA}} &= Q\left(\frac{|s|}{\sqrt{\mathrm{Var}(\eta)}}\right) \\ &= Q\left(\frac{\left|\sum_{i=1}^{N} \Delta V_{\mathrm{BL}}(W_i) \cdot \frac{X_i}{\|\mathbf{x}\|_2}\right|}{\sigma_{\mathrm{F}}}\right) \\ &= Q\left(\frac{\left|\sum_{i=1}^{N} W_i \cdot \frac{X_i}{\|\mathbf{x}\|_2}\right| \cdot \Delta V_{\mathrm{BL,max}}}{\sigma_{\mathrm{F}}}\right) \\ &= Q\left(N \alpha_{\mathrm{SVM}} \cdot \sqrt{\mathrm{SNR}_{\mathrm{DIMA}}}\right) \end{aligned} \tag{56}$$

where $\Delta V_{\mathrm{BL}}(W_i) = W_i \Delta V_{\mathrm{BL,max}}$ due to the normalized trained weights W_i.

ACKNOWLEDGEMENT

The authors acknowledge funding from Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-18-2-7866 (ERI FRANC program).

REFERENCES

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

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, May 2015.

[3] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," in *Proc. Int. Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, Mar. 2014, pp. 269–284.

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

[5] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "Envision: A 0.26-to-10 TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Pap.*, Feb. 2017, pp. 246–247.

[6] P. N. Whatmough, S. K. Lee, H. Lee, S. Rama, D. Brooks, and G.-Y. Wei, "A 28nm SoC with a 1.2 GHz 568 nJ/prediction sparse deep-neural-network engine with >0.1 timing error rate tolerance for IoT applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Pap.*, Feb. 2017, pp. 242–243.

[7] M. Price, J. Glass, and A. P. Chandrakasan, "A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Pap.*, 2017, pp. 244–245.

[8] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, and M. Alioto, "SRAM for error-tolerant applications with dynamic energy-quality management in 28 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 5, pp. 1310–1323, May 2015.

[9] F. Frustaci, D. Blaauw, D. Sylvester, and M. Alioto, "Approximate SRAMs with dynamic energy-quality management," *IEEE Trans. VLSI Syst.*, vol. 24, no. 6, pp. 2128–2141, Jun. 2016.

[10] Y. Kim, M. Kang, L. R. Varshney, and N. R. Shanbhag, "Generalized water-filling for source-aware energy-efficient SRAMs," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4826–4841, Oct. 2018.

[11] ——, "SRAM Bit-line Swings Optimization using Generalized Water-filling," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1670–1674.

[12] 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. (ICASSP)*, May 2014, pp. 8326–8330.

[13] M. Kang, S. K. Gonugondla, M.-S. Keel, and N. R. Shanbhag, "An energy-efficient memory-based high-throughput VLSI architecture for convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2015, pp. 1037–1041.

[14] N. Shanbhag, M. Kang, and M.-S. Keel, "*Compute memory*," Feb. 5 2015, US Patent App. 14/614,743.

[15] J. Zhang, Z. Wang, and N. 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] 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.

[17] M. Kang, S. K. Gonugondla, S. Lim, and N. R. Shanbhag, "A 19.4-nJ/decision, 364-K decisions/s, in-memory random forest multi-class inference accelerator," *IEEE J. Solid-State Circuits*, vol. 53, no. 7, pp. 2126–2135, Jul. 2018.

[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] Z. Jiang, S. Yin, M. Seok, and J.-s. Seo, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," in *Proc. IEEE Symp. VLSI Tech. (VLSIT)*, Jun. 2018, pp. 173–174.

[20] 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. (ISSCC) Dig. Tech. Pap.*, Feb. 2018, pp. 488–490.

[21] D. Bankman and B. Murmann, "An 8-bit, 16 input, 3.2 pJ/op switched-capacitor dot product circuit in 28-nm FDSOI CMOS," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2016, pp. 21–24.

[22] J. Wang, X. Wang, C. Eckert, A. Subramaniyan, R. Das, D. Blaauw, and D. Sylvester, "A compute SRAM with bit-serial integer/floating-point operations for programmable in-memory vector acceleration," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Pap.*, Feb. 2019, pp. 224–226.

[23] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," *arXiv preprint arXiv:1602.02830*, 2016.

[24] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: Imagenet classification using binary convolutional neural networks," in *Proc. European Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 525–542.

[25] 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 (VLSIC)*, Jun. 2018, pp. 141–142.

[26] N. Verma, "Analysis towards minimization of total SRAM energy over active and idle operating modes," *IEEE Trans. VLSI Syst.*, vol. 19, no. 9, pp. 1695–1703, Aug. 2010.

[27] Z. Wang, K. H. Lee, and N. Verma, "Overcoming computational errors in sensing platforms through embedded machine-learning kernels," *IEEE Trans. VLSI Syst.*, vol. 23, no. 8, pp. 1459–1470, Aug. 2015.

[28] M. Kang, S. Lim, S. Gonugondla, and N. R. Shanbhag, "An in-memory VLSI architecture for convolutional neural networks," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, pp. 494–505, Apr. 2018.

[29] M. H. Abu-Rahma, Y. Chen, W. Sy, W. L. Ong, L. Y. Ting, S. S. Yoon, M. Han, and E. Terzioglu, "Characterization of SRAM sense amplifier input offset for yield prediction in 28nm CMOS," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, Sep. 2011, pp. 1–4.

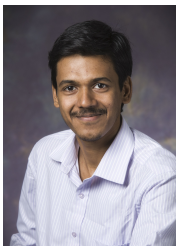
[30] L. Yang, D. Bankman, B. Moons, M. Verhelst, and B. Murmann, "Bit error tolerance of a CIFAR-10 binarized convolutional neural network processor," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.

[31] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8μJ/86% CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, Jan. 2018.

[32] J. Zhang, Z. Wang, and N. Verma, "A matrix-multiplying ADC implementing a machine-learning classifier directly with data conversion," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Pap.*, Feb. 2015, pp. 1–3.
[33] Z. Wang and N. Verma, "A low-energy machine-learning classifier based on clocked comparators for direct inference on analog sensors," *IEEE Trans. Circuits Syst. I*, vol. 64, no. 11, pp. 2954–2965, Jun. 2017.

Mingu Kang (M'13) received the B.S. and M.S. degrees in Electrical and Electronic Engineering from Yonsei University, Seoul, South Korea, in 2007 and 2009, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2017. From 2009 to 2012, he was with the Memory Division, Samsung Electronics, Hwaseong, South Korea, where he was involved in the circuit and architecture design of phase change memory. Since 2017, he has been with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, where he designs machine learning accelerator architecture. His current research interests include low-power integrated circuits, architecture, and system for machine learning, signal processing, and neuromorphic computing.

Yongjune Kim (S'12–M'16) received the B.S. and M.S. degrees in Electrical and Computer Engineering from Seoul National University, Seoul, South Korea, in 2002 and 2004, respectively, and the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2016. From 2016 to 2018, he was a postdoctoral scholar in Coordinated Science Laboratory at University of Illinois at Urbana-Champaign, Urbana, IL, USA. Since 2018, he has been with Western Digital Research, Milpitas, CA, USA. His research interests include machine learning, coding theory, information theory, computing, and storage. He was a recipient of the IEEE Data Storage Best Student Paper Award, the Best Paper Award of the 2016 IEEE International Conference on Communications (ICC), the Best Paper Award (honorable mention) of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS), and the Best Paper Award of the 31st Samsung Semiconductor Technology Symposium.



Ameya D. Patil (S'15) received the B.Tech. degree in 2014 from the department of Electrical Engineering at Indian Institute of Technology (IIT) Hyderabad, India. He received his M.S. degree in 2016 from the department of Electrical and Computer Engineering (ECE) at the University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, USA, where he is currently pursuing his Ph.D. degree. His research interests lie at the intersection of machine learning, circuits, and architecture. He is a recipient of the Joan and Lalit Bahl Fellowship from the ECE department at UIUC in 2015-16 and 2016-17.

Naresh R. Shanbhag (F06) is the Jack Kilby Professor of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He received his Ph.D. degree from the University of Minnesota (1993) in Electrical Engineering. From 1993 to 1995, he worked at AT&T Bell Laboratories at Murray Hill where he led the design of high-speed transceiver chip-sets for very high-speed digital subscriber line (VDSL), before joining the University of Illinois at Urbana- Champaign in August 1995. He has held visiting faculty appointments at the National Taiwan University (Aug.-Dec. 2007) and Stanford University (Aug.-Dec. 2014). His research focuses on exploring fundamental energy-latency-accuracy trade-offs in the design of integrated circuits and systems for communications, signal processing and machine learning. He has more than 200 publications in this area and holds thirteen US patents.

Dr. Shanbhag received the 2018 SIA/SRC University Research Award, became an IEEE Fellow in 2006, received 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, Dr. Shanbhag co-founded and served as 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 2013-17, he was the founding Director of the Systems On Nanoscale Information fabriCs (SONIC) Center, a 5-year multi- university center funded by DARPA and SRC under the STARnet program.