

Title:	A Multi-Functional In-Memory Inference Processor Using a Standard 6T SRAM Array
Archived version	Accepted manuscript: the content is identical to the published paper, but without the final typesetting by the publisher
Published version DOI :	10.1109/JSSC.2017.2782087
Journal homepage	http://sscs.ieee.org/publications/ieee-journal-of-solid-state-circuits-jssc
Authors (contact)	Mingu Kang (mkang17@illinois.edu) Sujan K. Gonugondla (gonugon2@illinois.edu) Ameya D. Patil (adpatil2@illinois.edu) Naresh R. Shanbhag (shanbhag@illinois.edu)
Affiliation	University of Illinois at Urbana Champaign

Article begins on next page

A Multi-functional In-memory Inference Processor using a Standard 6T SRAM Array

Mingu Kang, *Member, IEEE*, Sujan Gonugondla, *Student Member, IEEE*,
Ameya Patil, *Student Member, IEEE*, and Naresh R. Shanbhag, *Fellow, IEEE*

Abstract—A multi-functional in-memory inference processor integrated circuit (IC) in a 65 nm CMOS process is presented. The prototype employs a deep in-memory architecture (DIMA) which enhances both energy efficiency and throughput over conventional digital architectures via simultaneous access of multiple rows of a standard 6T bitcell array (BCA) per precharge, and embedding column pitch-matched low-swing analog processing at the BCA periphery. In doing so, DIMA exploits the synergy between the data-flow of machine learning (ML) algorithms and the SRAM architecture to reduce the dominant energy cost due to data movement. The prototype IC incorporates a 16 KB SRAM array and supports four commonly used ML algorithms - the support vector machine, template matching, k -nearest neighbor, and the matched filter. Silicon measured results demonstrate simultaneous gains (dot-product mode) in energy efficiency of $10\times$ and in throughput of $5.3\times$ leading to a $53\times$ reduction in the energy-delay product (EDP) with negligible ($\leq 1\%$) degradation in the decision-making accuracy, as compared to conventional 8-b fixed-point single-function digital implementations.

Index Terms—in-memory processing, machine learning, accelerator, inference, analog processing, associative memory.

I. INTRODUCTION

There is much interest in embedding data analytics into sensor-rich platforms such as wearables, biomedical devices, autonomous vehicles, robots, and Internet-of-Things, to provide these with decision-making capabilities. Such platforms need to implement machine learning (ML) algorithms under severe resource-constraints. Energy efficiency is critical for embedded battery-powered and autonomous platforms. As a result, a number of integrated circuit (IC) implementations of ML kernels and algorithms have appeared recently [1]–[10] to address the problems of designing energy-efficient ML systems in silicon.

ML algorithms require processing of large data volumes. Hence, the energy efficiency of ML hardware is limited by the energy cost of memory accesses [11] especially for large systems such as deep neural networks (DNNs) [1], [12]. Reusing data read from external memory (data reuse) [12], [13] is highly effective in saving energy as shown by DianNao [12] and Eyeriss [1], [14], but results in on-chip memory access still accounting for 35% to 45% of total energy, and does not

address the energy and delay of SRAM reads. Low-power circuit and architectural techniques such as dynamic voltage-accuracy-frequency scaling [9], RAZOR [8], [15], and power gating [7] focus on energy savings in the digital domain. These miss out on the opportunities afforded by analog domain processing. Low-voltage (few hundred mVs) SRAM techniques [16], [17] reduce the memory read energy but at a significant loss in throughput and a catastrophic loss in inference accuracy due to possible most significant bit (MSB) errors by reduced noise margin [18], [19]. Hence, SRAMs customized for ML algorithms [18], [19] were proposed to protect MSBs selectively. Modifying the bitcell array (BCA) to enable efficient data fetching via 7T SRAM [10] or embed computations [20] was proposed, but this leads to reduced storage density.

This paper employs an alternative referred to as the *deep in-memory architecture* (DIMA) [21], [22]. DIMA accesses multiple rows of a standard 6T SRAM BCA per precharge via pulse width modulated (PWM) wordline (WL) enabling signals, and processes the resulting BL voltage drops ΔV_{BL} via column pitch-matched low-swing analog circuits in the periphery of the BCA. Thus, DIMA reduces the energy cost due to data access without reducing the storage density and is able to simultaneously enhance energy efficiency and throughput over conventional architectures. Previously, the DIMA’s versatility was demonstrated by mapping ML algorithms such as a template matching [21], CNN [23], and sparse distributed memory [24], [25]. However, these works have relied on *simulations* using idealized system models of various stages, and do not address the numerous challenging design issues and trade-offs that arise in a practical IC realization due to DIMA’s intrinsic mixed-signal nature and stringent BCA pitch-matching requirements. Recently, DIMA ICs have been reported [26], [27] where some of these challenges have been alleviated by focusing on a single/fixed-function or limited precision. In [27], the AdaBoost algorithm was implemented using 1-b weight and 5-b input data in a 130 nm CMOS by employing pulse amplitude modulated (PAM) WL enabling signals to demonstrate a $113\times$ improved energy efficiency. Similarly, the Random Forest algorithm was implemented in a 65 nm CMOS [26] demonstrating a $6.8\times$ EDP reduction. Extension to multi-function and higher (8-b) precision weight and data realization of DIMA ICs is made non-trivial due to the need to address a host of design

The authors are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: mkang17@illinois.edu; gonugon2@illinois.edu; adpatil2@illinois.edu; shanbhag@illinois.edu).

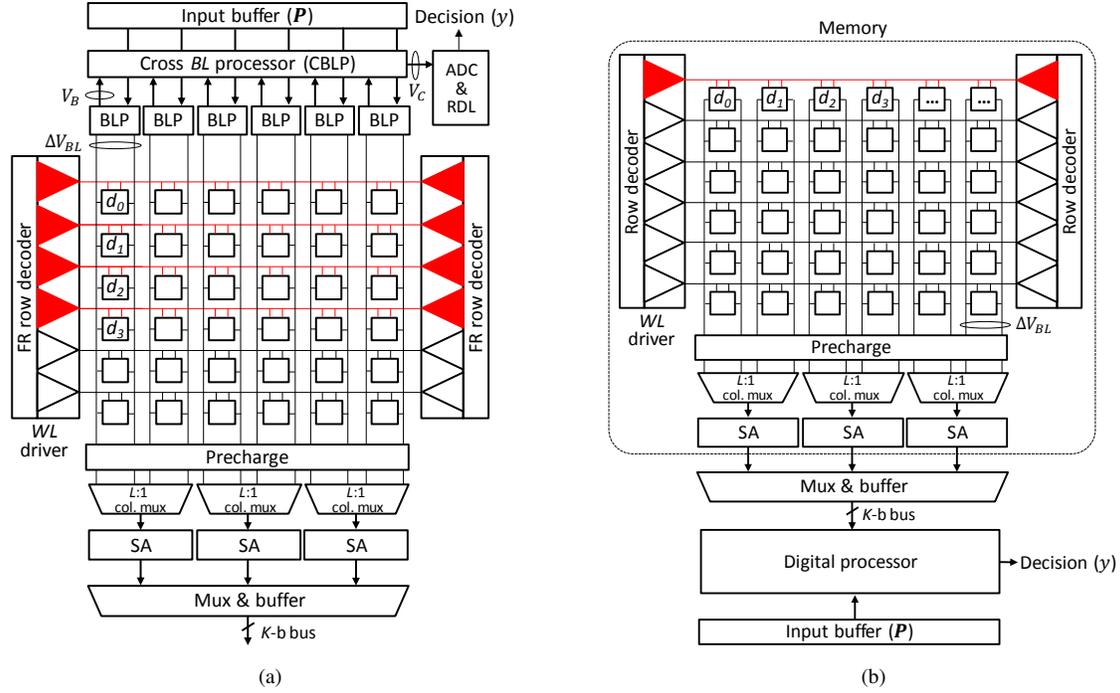


Fig. 1: A comparison of: (a) DIMA showing functional read (FR), bitline processing (BLP), cross BLP (CBLP) and sense amplifier (SA), and (b) the conventional architecture. Wordline (WL) drivers shaded in RED are turned on simultaneously.

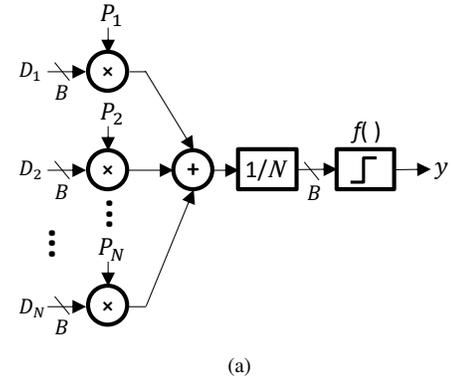
challenges.

This paper presents a multi-functional DIMA prototype IC in a 65 nm CMOS process with multi-bit precision (8-b) for both weights and input data. The multi-functional DIMA supports four different algorithms: support vector machine (SVM), template matching (TM), k -nearest neighbor (k -NN), and matched filter (MF). Measured results of prototype IC demonstrate up to $10\times$ and $5.3\times$ energy and delay reductions leading to $53\times$ energy-delay product (EDP) reduction with negligible ($\leq 1\%$) accuracy degradation as compared to the conventional 8-b fixed-point single-function digital (SRAM+digital MAC) architecture. Preliminary measured results of the multi-functional DIMA IC were reported in [28]. The contributions of this paper are to: (a) describe the critical issues underlying the design of robust DIMA ICs, (b) propose design techniques to overcome those issues, and (c) demonstrate the use of these techniques to design a multi-functional DIMA prototype IC in a 65 nm CMOS.

This paper is organized as follows. Section II presents an overview of DIMA, the design issues, and design guidelines/techniques to address those issues. Section III describes the prototype multi-functional IC architecture. Measured results including accuracy, energy, and delay trade-offs are presented in section IV. Section V concludes the paper.

II. THE DEEP IN-MEMORY ARCHITECTURE (DIMA)

This section describes DIMA and issues related to the design of robust DIMA ICs. Design techniques are also introduced to overcome those issues.



Algorithm	Scalar distance (SD)	Vector distance (VD)	$f()$
SVM	Multiplication	Dot product	sign
TM	Absolute difference	Manhattan distance	min
k -NN	Absolute difference	Manhattan distance	majority vote
MF	Multiplication	Dot product	max

(b)

Fig. 2: Data-flow of typical inference algorithms: (a) data-flow diagram, and (b) computations required in support vector machine (SVM), template matching (TM), k -nearest neighbor (k -NN), and matched filter (MF) algorithms.

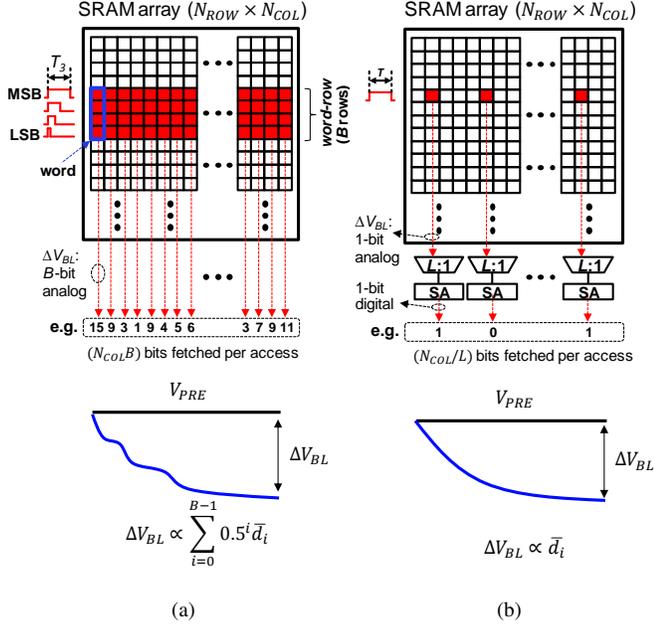


Fig. 3: Memory access pattern and bitline swing (ΔV_{BL}) in: (a) DIMA, and (b) the conventional architecture, assuming $B = 4$ and $L = 4$. Bitcells marked in RED are accessed simultaneously.

A. DIMA Overview

DIMA (Fig. 1(a)) based on a $N_{row} \times N_{col}$ BCA consists of four sequentially executed stages: a) the **multi-row functional read** (FR) stage fetches an N -dimensional data vector D consisting of B -bit elements D , stored in a column-major format, by reading B rows (*word-row*) per BL precharge (read cycle). This stage includes the precharge circuitry, the FR WL drivers and the BCA; b) the **BL processing** (BLP) stage computes scalar distances (SDs) between the N elements D of D and the N elements P of another N -dimensional pattern vector P . The N_{col} BLP blocks operate in parallel in a single-instruction multiple data (SIMD) manner; c) **cross BL processing** (CBLP) aggregates the SDs across the N_{col} analog BLP block outputs to obtain a vector distance (VD); and d) the **ADC and residual digital logic** (RDL) for realizing a thresholding/decision function $f(\cdot)$ and other miscellaneous functions.

The DIMA is well-matched to the data-flow (Fig. 2(a)) intrinsic to commonly encountered ML algorithms (Fig. 2(b)). Note: the VD is a dot product (DP) if the SD is the product DP , and it is a Manhattan distance (MD) if the SD is the absolute difference $|D - P|$. Both types of VDs are commonly encountered in ML algorithms and tend to dominate their complexity.

Both DIMA (Fig. 1(a)) and a conventional digital (SRAM+digital MAC) architecture (Fig. 1(b)) employ identical BCAs to store D and an input buffer to store a streamed pattern/template P . However, DIMA stores the B bits of

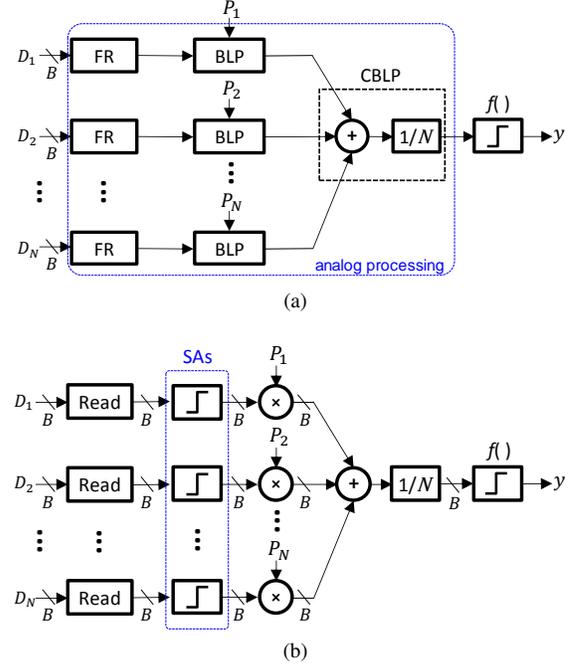


Fig. 4: Simplified data-flow in: (a) DIMA, and (b) the conventional architecture.

the scalar D in a column-major format (Fig. 3(a)) vs. row-major used in the digital architecture (Fig. 3(b)). The FR generates a BL voltage drop ΔV_{BL} as a function of the B bits per column [21]. On the other hand, the SRAM in the digital architecture requires a $L : 1$ column mux ratio (typical $L = 4$ to 32) to accommodate large area sense amplifiers (SAs) as shown in Fig. 2(b), which limits the number of bits per access to N_{col}/L in standard SRAM compared to $N_{col}B$ in FR. Moreover, unlike the digital architecture, DIMA's N_{col} BLPs and the CBLP are pitch-matched to the BCA avoiding the need to fully read out data (see Table I and Fig. 3 for a comparison summary).

Though DIMA needs much fewer precharge cycles to read the same number of bits leading to both energy and throughput gains, it relaxes the fidelity/accuracy of its reads and computations. However, this loss is easily absorbed by the *system-level robustness inherent in DIMA's functional data-flow*. A comparison of DIMA's functional data-flow (Fig. 4(a)) with that of the conventional architecture (Fig. 4(b)) shows that: (a) the FR stage combined with the column-major format enables DIMA to assign significantly (exponentially) more weight (pulse width [21] or pulse amplitude [27]) to MSBs vs. LSBs (**significance-weighted bit protection**), and (b) DIMA makes its first and only hard decision (thresholding) in the final RDL stage thereby enabling the CBLP to enhance the signal-to-noise ratio (SNR) by averaging out the impact of cross-column variations in the analog BLP output $V_B(D, P)$ by a factor of $\sqrt{N_{col}}$ (**delayed decision** and **SNR boosting**). The conventional digital architecture lacks these features and thus needs to rely on the SA making accurate decisions

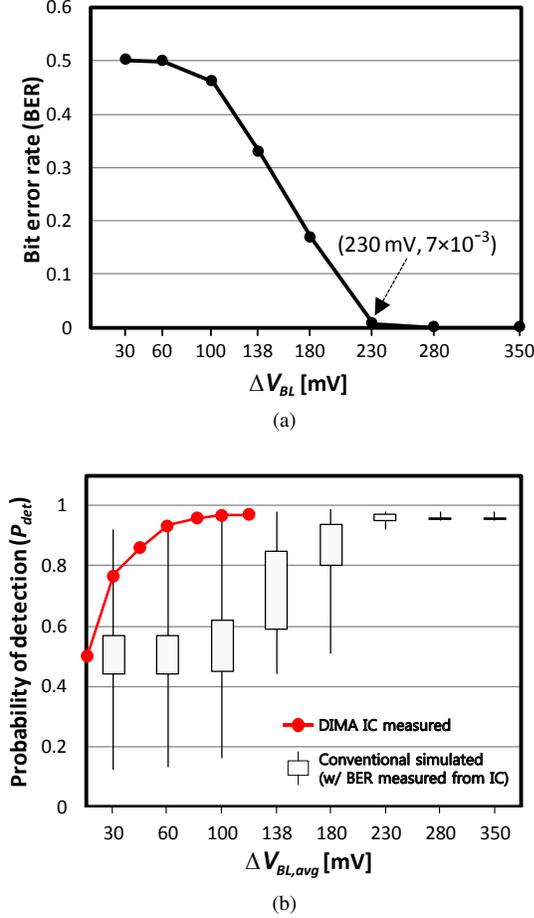


Fig. 5: Impact of reducing BL swing in the conventional architecture: (a) measured SA bit error-rate (BER) during normal SRAM read operation, and (b) the decision accuracy P_{det} of face detection task using SVM for the conventional system (simulated with SA bit errors injected based on measured BER from Fig. 5(a)) and DIMA (measured). Here, $\Delta V_{BL,avg}$ is the average ΔV_{BL} per bit.

to prevent MSB errors in addition to the decision in the thresholding function $f(\cdot)$.

In fact, even for a 1-b architecture, one can show that DIMA's decision accuracy will be higher than that of the conventional architecture under identical total noise variance. For example, in Fig. 4, with $D_i = D$ as a Bernoulli random variable with parameter $1/2$, $P_i = 1$, and a total noise variance of σ_n^2 in the analog chain of DIMA and the READ operation of the conventional architecture with odd N , one can show that:

$$\begin{aligned}
 p_{e,DIMA} &= Q\left(\frac{\sqrt{N}}{2\sigma_n}\right) \leq \\
 p_{e,CONV} &= \sum_{i=(N+1)/2}^N \binom{N}{i} (1-\epsilon)^{N-i} \epsilon^i \quad (1)
 \end{aligned}$$

TABLE I: DIMA vs. conventional architecture (with $N_{col} \times N_{row}$ BCA).

Attribute	Conventional	DIMA
data storage pattern	row major	column major
column mux ratio	$L : 1$	$1 : 1$
fetched words per access	$N_{col}/(LB)$	N_{col}
BL swing/LSB (ΔV_{BL})	250–300 mV	5–30 mV
# of rows per access	1	B
WL driver	fixed pulse width	pulse width/amp modulated

where $\epsilon = Q(1/2\sigma_n)$, $p_{e,DIMA}$, and $p_{e,CONV}$ are the error probabilities of DIMA and the conventional architecture, respectively, for any σ_n . Furthermore, the gap between $p_{e,DIMA}$ and $p_{e,CONV}$ is maximum in the low-SNR regime. This result is a direct consequence of DIMA's delayed decision and SNR boosting properties which the conventional system lacks. Therefore, DIMA is intrinsically superior in accuracy to the conventional architecture in the low-SNR regime. We validated this observation via measurements on the DIMA prototype IC as shown in Fig. 5. A face detection task using SVM is implemented via the configurations described in Section III-C. Figure 5(b) shows that the conventional system suffers from drastic accuracy degradation in the low-SNR regime, i.e., under reduced ΔV_{BL} , due to the SA bit errors in the MSBs. On the other hand, DIMA achieves much higher accuracy due to above mentioned system-level robustness inherent in DIMA's functional data-flow.

The following sections describe each processing stage and related design issues in detail.

B. Multi-row Functional READ (FR)

The FR stage generates a BL voltage drop $\Delta V_{BL}(D)$ proportional to the weighted sum $D = \sum_{i=0}^{B-1} 2^i d_i$ of column-major stored data $\{d_0, d_1, \dots, d_{B-1}\}$ (see Fig. 6(a)) via a simultaneous application of binary-weighted pulse widths (PWM) $T_i \propto 2^i$ ($i \in [0, B-1]$) that discharges the BL capacitance C_{BL} via B bitcell discharge paths (Fig. 6(b)) [21] as follows:

$$\begin{aligned}
 \Delta V_{BL}(D) &= \frac{V_{PRE}}{R_{BL}C_{BL}} T_0 \sum_{i=0}^{B-1} 2^i \bar{d}_i \\
 &= \Delta V_{lsb} \sum_{i=0}^{B-1} 2^i \bar{d}_i = \Delta V_{lsb} \bar{D} \quad (2)
 \end{aligned}$$

where $\Delta V_{lsb} = \frac{V_{PRE}T_0}{R_{BL}C_{BL}}$, V_{PRE} is BL precharge voltage level, and \bar{D} is the decimal value of the one's complement of D . Equation (2) holds if the following four conditions apply: 1) $T_i \ll R_i C_{BL}$ (R_i is the i^{th} bitcell's discharge path

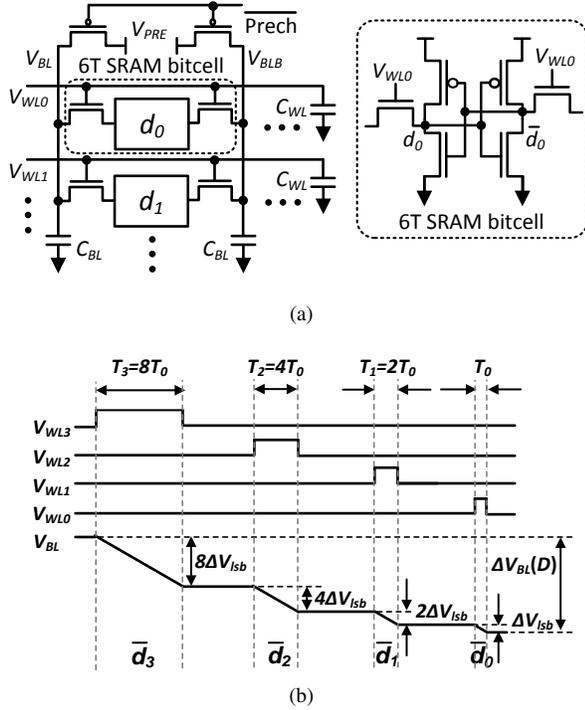


Fig. 6: FR using pulse-width modulated (PWM) WL access pulses [21]: (a) column structure and bitcell, and (b) idealized waveforms during a 4-b word ($D = 0000b'$) read-out. The WL pulses can be overlapped but are shown as being sequentially applied to enhance clarity.

resistance), 2) $T_i = 2^i T_0$, 3) $R_i = R_{BL}$ (no variation across rows, R_{BL} is nominal resistance of the BL discharge path via the access and pull-down transistors of the enabled bitcell), and 4) R_{BL} is a constant over V_{BL} . A similar expression for $\Delta V_{BLB}(D)$ can be obtained by replacing \bar{d}_i with d_i in (2). Thus, the FR stage converts the stored data D into $\Delta V_{BL}(D)$ and $\Delta V_{BLB}(D)$, i.e., a digital-to-analog conversion.

The FR stage can also realize simple SD functions such as the addition and subtraction of two B -bit words (D and P) stored different rows but in the same column. For example, from (2), $D + P$ is obtained by applying FR to word-rows containing D and P to obtain:

$$\Delta V_{BL}(D, P) = \Delta V_{Isb} \sum_{i=0}^{B-1} 2^i (\bar{d}_i + \bar{p}_i) = \Delta V_{Isb} (\bar{D} + \bar{P}) \quad (3)$$

Similarly, subtraction $D - P$ can be realized by storing \bar{P} (one's complement of P) in the same column as D . Subtraction will be discussed in a later section in more detail.

1) FR Design Issues

The ΔV_{BL} generated by the FR is subject to the following circuit non-idealities which make it difficult to meet the four conditions described earlier: 1) spatial transistor threshold voltage (V_t) variations caused by random dopant fluctuations [29] (Condition 3), 2) BL voltage-dependence of the discharge path (access and pull-down transistors in the bitcell)

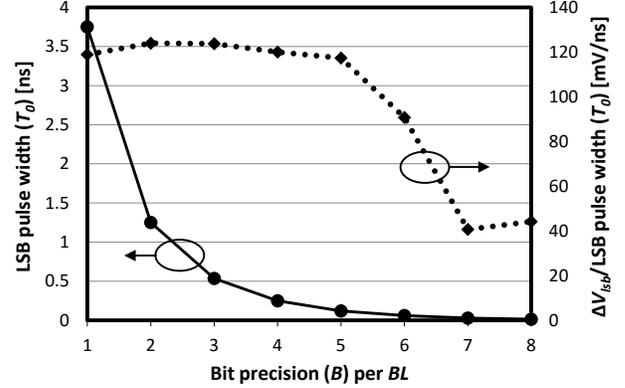


Fig. 7: FR accuracy vs. bit precision (B) per BL obtained via post-layout simulations.

resistance R_{BL} (see (2)) (Condition 1 and 4), and 3) the finite transition (rise and fall) times of the PWM WL pulses (Condition 2). These non-idealities can be alleviated to some extent via appropriate choice of the LSB WL pulse widths T_0 and the WL pulse amplitude V_{WL} . The aggregation process in the CBLP is effective in alleviating V_t variations across columns.

The choice of V_{WL} is governed by the need to keep it sufficiently low to alleviate the BL voltage dependence of R_{BL} . Doing so ensures that NMOS access transistor does not transition from saturation into triode region thereby satisfying Condition 4. This has the additional benefit that R_{BL} is increased thereby making it easier to satisfy the overarching Condition 1 ($T_i \ll R_{BL} C_{BL}$). However, V_{WL} needs to be sufficiently large (lower bounded) so that variations in R_i are reduced, i.e., Condition 3 ($R_i = R_{BL}$) is approximated well, and simultaneously upper bounded to avoid destructive read operation, e.g., $V_{WL} < 0.8V_{PRE}$.

Similarly, the choice of T_0 is lower bounded ($T_0 > T_{min}$) so that the rise (T_r) and fall (T_f) times of WL are a small fraction of T_0 , e.g., $T_r + T_f < 0.5T_{min}$, and hence Condition 2 ($T_i = 2^i T_0$) can be met. That and Condition 1 implies

$$2(T_r + T_f) < T_0 \ll 2^{1-B} R_{BL} C_{BL}. \quad (4)$$

Hence, for the prototype IC, we chose $V_{WL} = 0.65V$, $V_{PRE} = 1V$, and $T_0 = 250$ ps.

2) FR Design Techniques

We present two FR design techniques to overcome the design issues described in section II-B.

Sub-ranged Read: Realizing a highly linear FR stage when $B > 4$ bits is challenging because the constraint $2(T_r + T_f) < T_0 \ll 2^{1-B} R_{BL} C_{BL}$ is hard to meet when driving high WL capacitance (e.g., 200 fF) with a row pitch-matched WL driver. Figure 7 shows the relationship between bit precision (B), the LSB pulse width T_0 , and the BL swing per LSB (ΔV_{Isb}), via post-layout simulations. Equation (4) indicates that the upper bound on T_0 reduces and approaches the lower bound as B increases. Due to the finite rise and

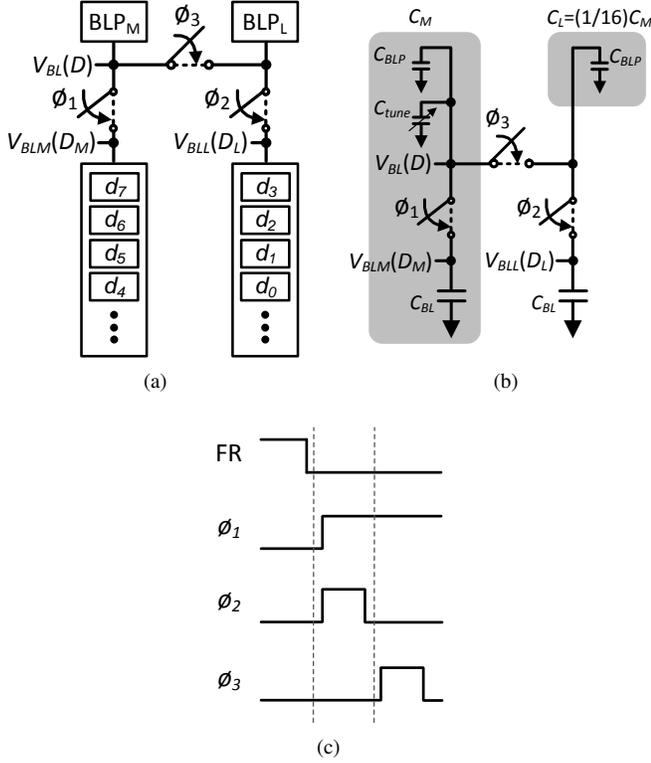


Fig. 8: Sub-ranged read with $B = 8$ (BLB is not shown for simplicity): (a) BL pair structure (two neighboring bit-cell columns), (b) equivalent capacitance model [28], where $D_M = 8d_7 + 4d_6 + 2d_5 + d_4$ and $D_L = 8d_3 + 4d_2 + 2d_1 + d_0$, and (c) timing diagram.

fall times, the lower bound in (4) becomes difficult to meet resulting in the ratio $\Delta V_{lsb}/T_0$ varying with B when it needs to remain constant ($= V_{PRE}/R_{BL}C_{BL}$) as defined by (2). Figure 7 shows that the ratio $\Delta V_{lsb}/T_0$ is constant until $B = 5$. Therefore, we restrict the FR per column to $B = 4$.

The sub-ranged read technique enables FR with higher bit precision (e.g., $B = 8$) by storing the $B/2$ bits representing the MSB word D_M and $B/2$ bits representing the LSB word D_L in adjacent columns (BLM and BLL) of the BCA (see Fig. 8(a)). The MSB and LSB BL capacitances ratio $C_M : C_L$ is set to $2^{\frac{B}{2}} : 1$ via a tuning capacitor C_{tune} , where $C_M = C_{BL} + C_{BLP} + C_{tune}$ and $C_L = C_{BLP}$ in Fig. 8(b). Three switches $\phi_{1,2,3}$ are used in specific sequence (Fig. 8(c)) to charge-share BLM and BLL , generating a BL voltage drop:

$$\begin{aligned} \Delta V_{BL}(D) &= \frac{C_M \Delta V_{BLM}(D_M) + C_L \Delta V_{BLL}(D_L)}{C_M + C_L} \\ &= \frac{2^{\frac{B}{2}} \Delta V_{BLM}(D_M) + \Delta V_{BLL}(D_L)}{2^{\frac{B}{2}} + 1} \end{aligned} \quad (5)$$

FR Replica BCA: Computing the difference $D - P$ requires \bar{P} to be stored in the same column but a different word-row as D as shown in (3). As P is a streamed in data,

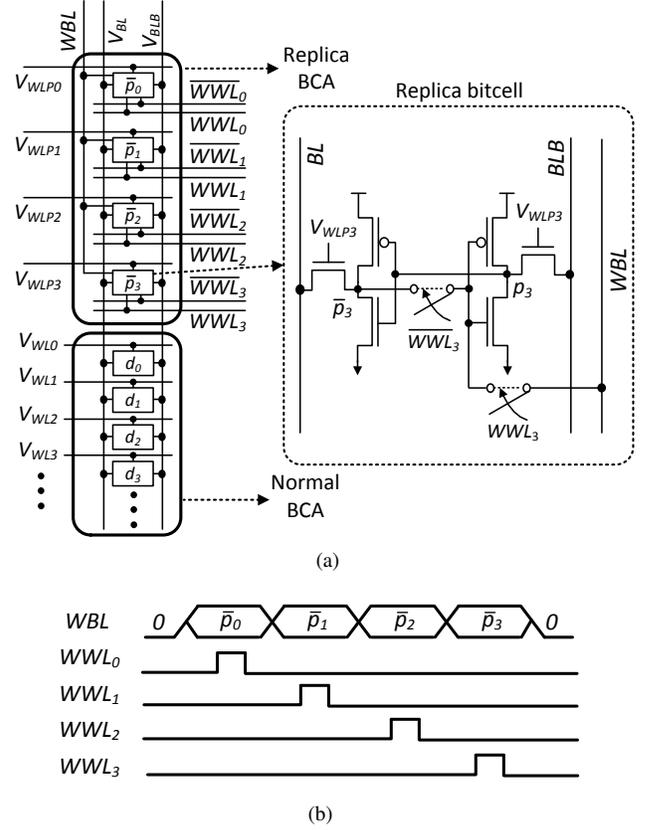


Fig. 9: Replica BCA: (a) bitcell column ($B = 4$), and (b) timing diagram for replica BCA writing.

e.g., a template in TM, storing \bar{P} in the same BCA as D will require repeated SRAM write operations which incur large energy and delay costs as these require full BL swing. The replica BCA (Fig. 9(a)) solves this problem by enabling fast and energy-efficient writes of \bar{P} via a separate write BL (WBL) and WL (WWL). The \bar{P} is written into the replica BCA in a bit-serial manner via the WBL (Fig. 9(b)). Subsequently, the FR process is applied to the regular and replica BCA simultaneously. The layout of replica bitcell needs to be similar to regular bitcell to have the same discharge strength except the WBL and WWL circuitry.

After applying these techniques, Monte Carlo post-layout simulations show that $\Delta V_{BL}(D)$ exhibits a 12% variation (σ/μ) for typical values of $V_{WL} = 0.65$ V, $V_{PRE} = 1$ V, $T_0 = 250$ ps, $N = 128$, $B = 8$, and $N_{row} = 512$. The measured integral non-linearity (INL) of FR was found to be less than 0.87 LSB. Measured results in section IV show that this level of accuracy from FR stage is sufficient for the data sets being considered in this paper.

C. Bitline Processing (BLP) and Cross Bitline Processing (CBLP)

The N_{col} BLP blocks in Fig. 1(a) accept two operands: 1) the bitline voltage drop $\Delta V_{BL}(D)$ generated via the

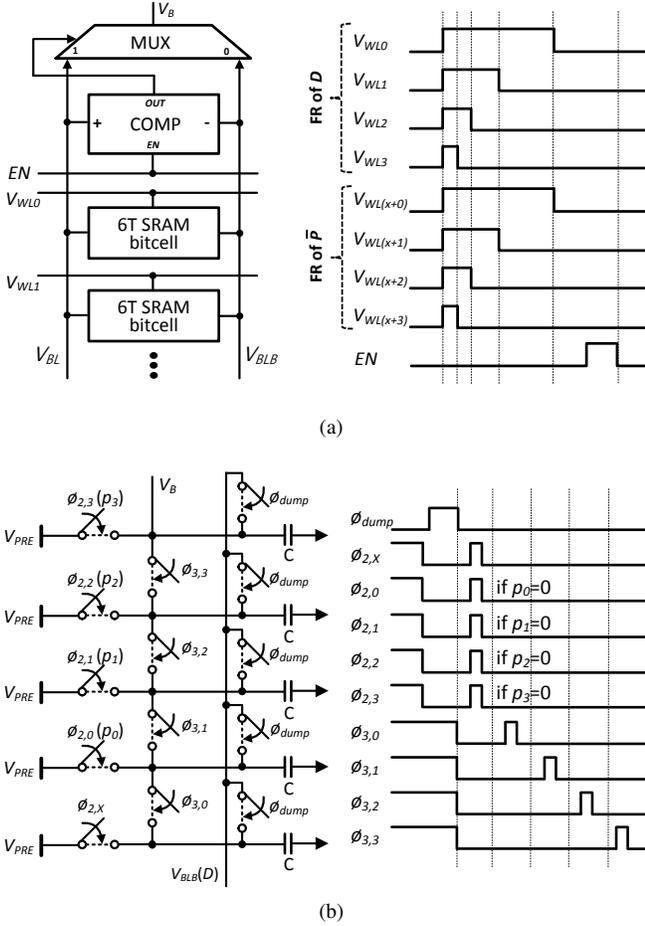


Fig. 10: BLP implementation: (a) absolute difference, where D and \bar{P} are stored in the same column [21], and (b) charge redistribution-based multiplication with $B = 4$ [23].

FR stage, and 2) a word P to generate an output voltage $V_B(D, P)$. The implementation of absolute difference $|D - P|$ [21] and multiplication DP [23] in the BLP is described next, as these are commonly used in ML algorithms.

Absolute Difference: The BLP computes $|D - P| = \max(D - P, P - D)$ by employing $\Delta V_{BL}(D, \bar{P}) = \Delta V_{lsb}(\bar{D} + P)$ from (3) and exploiting the intrinsically differential structure of the SRAM bitcell to evaluate $\max(V_{BL}, V_{BLB})$ via a BL compare-select (Fig. 10(a)) to obtain

$$\max(V_{BL}, V_{BLB}) = V_{PRE} - (2^B - 1)\Delta V_{lsb} + \Delta V_{lsb}\max(P - D, D - P) \quad (6)$$

Thus, applying FR to D and \bar{P} simultaneously results in V_{BL} and V_{BLB} being proportional to $P - D$ and $D - P$, respectively. The local analog BL compare-select block (Fig. 10(a)) provides the maximum of V_{BL} and V_{BLB} , and hence the absolute difference $|D - P|$.

Multiplication: Figure 10(b) shows a charge redistribution-based mixed-signal multiplier [23] with

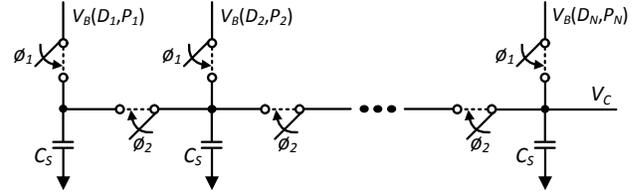


Fig. 11: CBLP implementation.

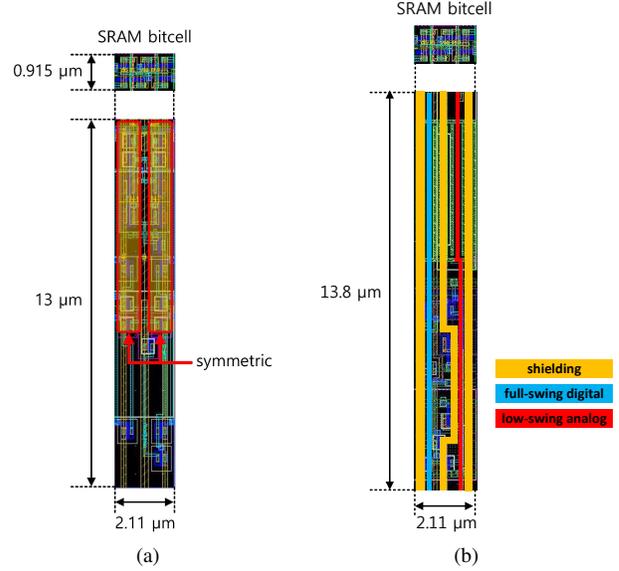


Fig. 12: Pitch-matched layouts of BLP blocks relative to a SRAM bitcell: (a) analog comparator, and (b) a part (1/5) of charge redistribution-based multiplier.

inputs $\Delta V_{BLB}(D)$ (FR stage output) and an externally provided B -bit digital word P , whose bits p_i control the $\phi_{2,i}$ switches. The timing diagram in Fig. 10(b) describes the multiplier operation. The multiplier output voltage V_B is given by:

$$\begin{aligned} V_B(D, P) &= V_B(DP) = V_{PRE} - (0.5)^B P \Delta V_{BLB}(D) \\ &= V_{PRE} - (0.5)^B \Delta V_{lsb} DP \end{aligned} \quad (7)$$

Thus, voltage drop $\Delta V_B(DP) = V_{PRE} - V_B(DP) \propto DP$ represents the product of D and P . Note that the multiplier employs unit size (25 fF) capacitors rather than binary weighted ones as in [30] due to stringent column pitch-match constraints on the BLP.

The cross BL processor (CBLP) (Fig. 11) samples the output voltage ($V_B(D, P)$) of the BLP on the BL -wise sampling capacitors C_S at each column by pulsing the ϕ_1 switches. Next, the ϕ_2 switches are pulsed to generate the CBLP output V_C in one step.

1) BLP and CBLP Design Issues: Though the input offset of the comparator in Fig. 10(a) is the dominant source of non-ideality in computing $|D - P|$, it affects the BLP output $V_B = \max(V_{BL}, V_{BLB})$ minimally as V_{BL} and V_{BLB} are

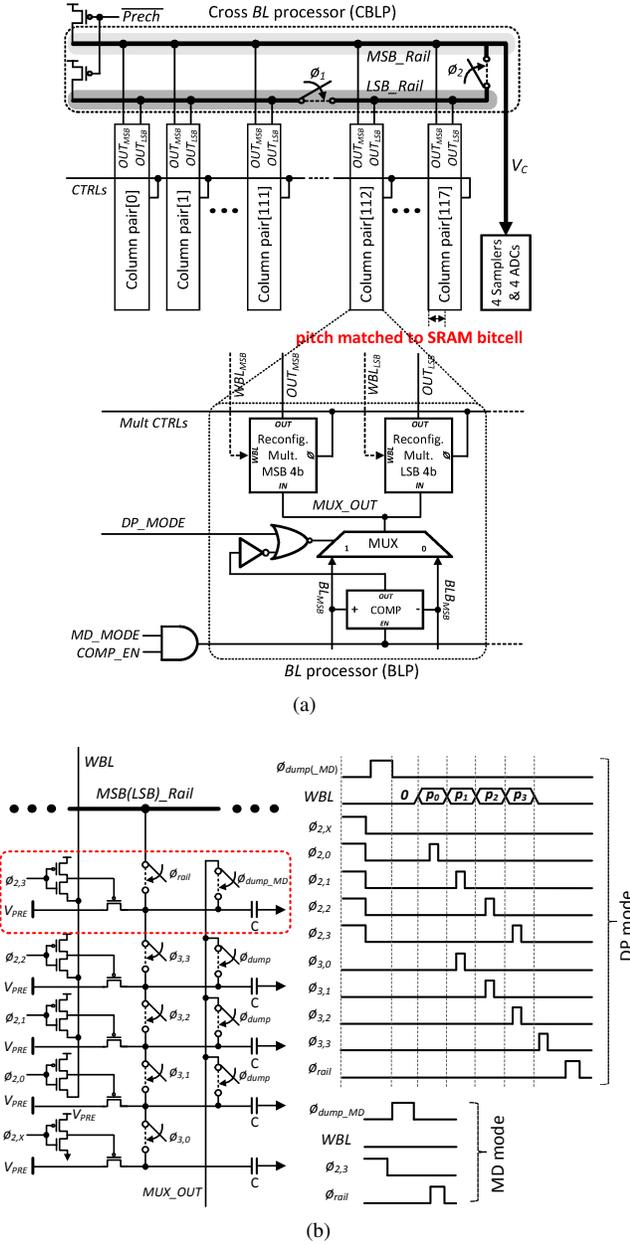


Fig. 13: BLP and CBLP implementations for reconfiguration: (a) overall structure, and (b) reconfigurable charge-based multiplier for 4-b MSB (or LSB) and its enable signals (MD mode uses only the circuits in the dotted RED box).

close to each other when offset matters. Additionally, the error in V_B being uncorrelated across the columns gets averaged out further by the CBLP. The comparator layout in Fig. 12(a) is constrained to be symmetric to minimize the offset. Monte Carlo post-layout simulations in a 65 nm CMOS process indicates that the input offset has a standard deviation of $\sigma = 10$ mV.

The charge redistribution circuits in computing DP in the BLP (see Fig. 10(b)) and summation in the CBLP (see Fig. 11) suffer from multiple noise sources: 1) charge-

injection noise, 2) thermal noise, and 3) coupling noise. The 8-b ADC (to convert V_C to digital value) with an input dynamic range of 300 mV results in a target voltage resolution per LSB $V_{res} \approx 1$ mV. Simplified thermal noise considerations ($\sqrt{KT/C} < 0.5V_{res}$) lead to the requirement of $C > 17$ fF at $T = 300$ K. Therefore, we choose $C_S = 25$ fF in order to minimize the effect of thermal noise and charge injection. These random noise sources are further minimized by the averaging effect of the CBLP stage. Coupling noise was alleviated by shielding low-swing analog nodes from the digital full-swing lines (see Fig. 12(b)) by using dummy metal lines.

2) *BLP and CBLP Design Techniques*: The BLP block needs to be reconfigurable in order to compute the absolute difference (MD mode) or the scalar product (DP mode) functionality under pitch-matching constraints.

Circuit sharing with reconfiguration: Figure 13(a) shows the BLP (and CBLP) circuitry to support the DP and MD modes. In both modes, the capacitor C (red box in Fig. 13(b)) is shared with CBLP to realize capacitor C_S in Fig. 11. Furthermore, in the DP mode, the comparator needed by the MD mode, is bypassed and mux always chooses BLB .

Sub-ranged processing: The charge-based multiplier in Fig. 10(b) employs unit capacitors to meet the column pitch constraints necessitating sequential processing of multiplicand bits (p_i) and thereby limiting the throughput. Sub-ranged multiplication alleviates this problem by employing two 4-b MSB and LSB multipliers operating in parallel (Fig. 13(a)) which generate their output voltages V_{BS} on the MSB_Rail and LSB_Rail , respectively, when $\phi_1 = 1$. A weighted sum of the two voltages is achieved by setting $\phi_1 = 0$ and $\phi_2 = 1$ so that the voltage on MSB_Rail is weighed $16\times$ more than the one on LSB_Rail .

D. ADC and Residual Digital Logic (RDL)

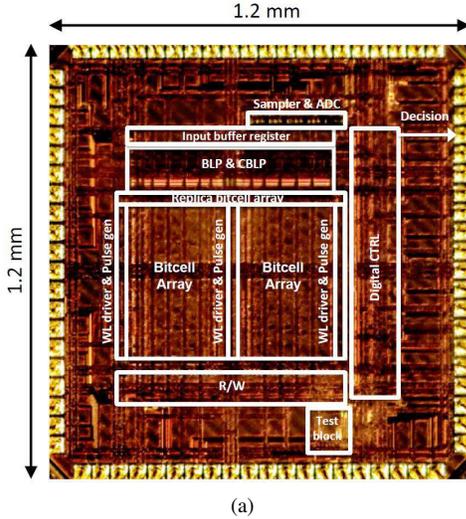
The ADC digitizes the CBLP output V_C for further processing by the RDL, which implements slicing/thresholding functions such as min, max, sign, sigmoid, and majority vote. The ADC and RDL need to process one scalar value (V_C) generated from a massively parallel (> 128) SD processing in the BLP. Thus, the energy overhead of ADC and RDL is negligible.

III. PROTOTYPE IC ARCHITECTURE AND ALGORITHM MAPPING

This section describes the multi-functional DIMA prototype IC (Fig. 14) designed in a 65 nm CMOS process and packaged in a 88-pin QFN.

A. Architecture

The chip architecture in Fig. 15 comprises a DIMA core (CORE), a digital controller (CTRL), and an input register to stream in the operand P . The CORE includes a 512×256 BCA, the conventional SRAM read/write circuitry,



Technology	65 nm CMOS
Die size	1.2 mm × 1.2 mm
CTRL power	4.3 mW @1 GHz
SRAM capacity	16 KB (512 × 256)
Bitcell dimension	2.11 × 0.92 μm ²
Supply voltage	CORE: 1.0 V, CTRL: 0.85 V

(b)

Fig. 14: Prototype IC: (a) die micrograph, and (b) chip summary.

the BLP and CBLP, and four 8-b single-slope ADCs [31]. The RDL is embedded in the CTRL. The SRAM bitcell was custom-designed following standard layout design rules as the memory compiler did not allow modifications of the peripheral circuitry. As a result, the horizontal and vertical dimensions of bitcell were approximately $1.7\times$ larger than typical foundry-provided bitcells [32]. The column muxing ratio with $L = 4$ was chosen to maximize the throughput for the standard SRAM read.

An 8-b precision is chosen for D and P in order to maintain almost the same accuracy as floating point [2], [33], [34]. Based on the design principles in section II-B, parameter values $V_{WL} = 0.65$ V and $T_0 \approx 250$ ps were chosen resulting in the longest PWM-WL pulse width $T_3 < 0.4R_{BL}C_{BL}$ to ensure sufficient linearity and avoid destructive read. Serially provided reconfiguration word $RCFG$ initializes the local controllers in the CTRL.

The circuitry for normal read and write operations (R/W block in Fig. 14(a)) occupies 14% of the CORE, which includes SA, write driver, and column decoder. On the other hand, the total area overhead due to DIMA circuitry was found to be 19% of the CORE area. Specifically, the charge-based multiplier (Fig. 13(a)) for DP mode takes 10% whereas

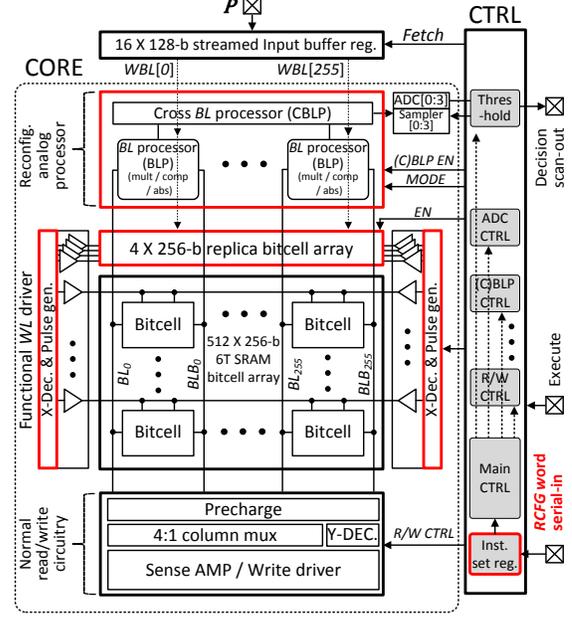


Fig. 15: Multi-functional DIMA prototype IC architecture.

the analog comparator, mux, and replica BCA (Fig. 13(b)) for MD mode takes 9%. The CBLP does not require additional area as it shares its circuit with BLP blocks. Supporting FR does not incur additional area penalty as it only requires enabling multiple, but pre-existing WL drivers.

B. Timing

The chip operations are sequenced via the CTRL which operates with a master 1 GHz CLK thereby providing a 1 ns time resolution to generate control signals for the CORE. Self-timed control [35], [36] can improve the throughput of both normal read and DIMA operations but we chose synchronous design for simplicity.

The timing diagram in Fig. 16(a) describes the series of ten events that occur during a *word-row period*, i.e., when processing a single word-row of $B = 8$ bits through the FR, BLP, and CBLP stages. The first event in both MD and DP modes is the BL precharge. Next, the FR, BLP, and CBLP stages are sequentially executed to generate the corresponding outputs V_{BL} , V_B , and V_C , respectively. One difference between the two modes - the MD mode requires transferring P from the input buffer into replica BCA before initiating FR. On the other hand, the DP mode needs to make this transfer of P to the mixed-signal multiplier before initiating BLP and requires additional delay in the CBLP stage to support sub-ranged processing. The last event samples the CBLP output to generate the input voltage for the ADC. Each event requires an integer number of CLK cycles which are estimated via post-layout simulations. The prototype IC provides tunability to allow additional CLK cycles for each stage in order to accommodate deviations from the nominal process corner.

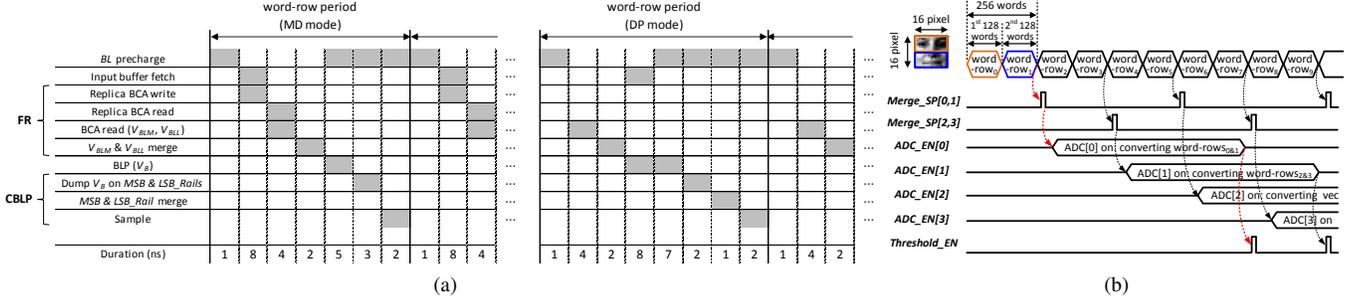


Fig. 16: DIMA timing for: (a) a single word-row, and (b) multiple word-rows (dotted RED line shows a single thread to process 256 words).

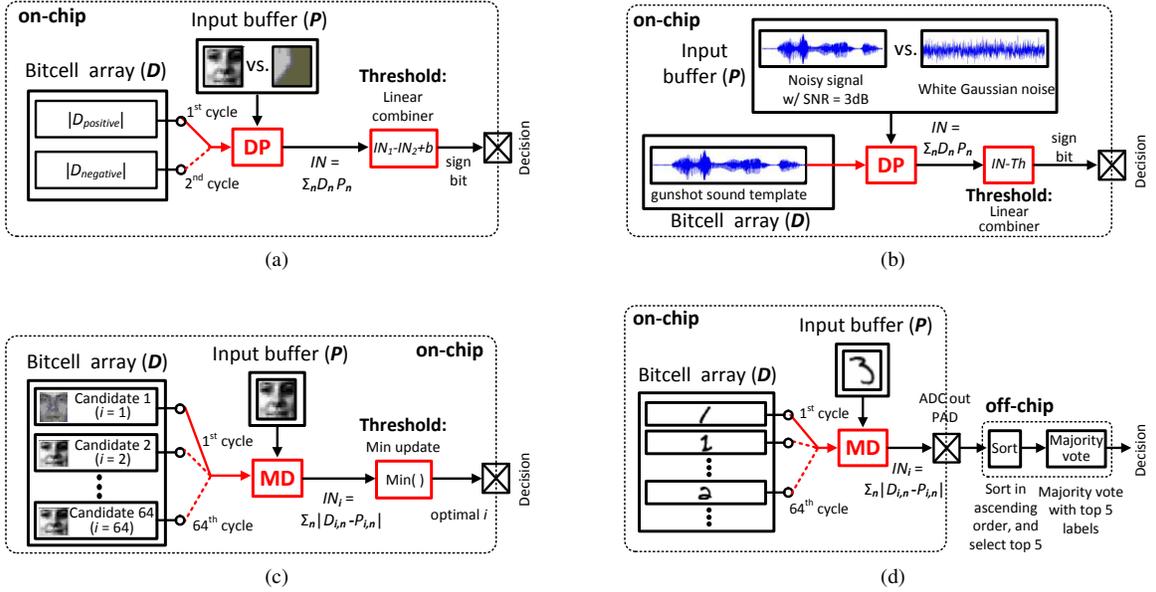


Fig. 17: Four inference tasks mapped on the prototype IC: (a) SVM for face detection, (b) MF for event detection, (c) TM for face recognition, and (d) k -NN for handwritten number recognition.

Figure 16(b) shows timing diagram for processing 256-dimensional D and P . Each word-row consists of 128 8-b words (though $N_{col} = 256$, the sub-ranged read results in a 128 dimensional vector) generate the CBLP output V_C . Two word-rows are processed consecutively and their outputs V_C s are sampled and charge-shared (*Merge_SP* step) to aggregate 256 SD results. This is followed by using 1-of-4 ADCs to digitize the analog V_C into an 8-b word. The single-slope ADC conversion takes 140 CLK cycles for both MD and DP modes, which is approximately 5.6 MD word-row periods. However, this slow conversion rate is not an issue as four ADCs operate in parallel. The ADC output is further processed in the RDL block to realize the thresholding operation (*Threshold_EN* step).

C. Algorithm and Application Mapping

Four tasks (face detection using SVM, gun shot detection using MF, face recognition using TM, and handwritten digit

recognition using k -NN) (see Fig. 17) were mapped on to the prototype IC. These tasks cover both binary and multi-class (4-class and 64-class) scenarios, requiring both MD and DP modes of operations, and processing of both image and sound data sets [37]–[39] as summarized in Table II. Table III defines the set of operations per stage that can be chosen using the *RCFG* word. In this process, the prototype IC is able to realize the four different algorithms as shown in Table IV.

The MF (Fig. 17(b)) creates the decision right after single DP processing and thresholding. On the other hand, SVM requires signed coefficients. Thus, the absolute values of positive and negative coefficients are stored in the separate rows (Fig. 17(a)). The rows are processed in consecutive cycles, and then compared in the RDL stage to obtain the sign of the DP. TM (Fig. 17(c)) and k -NN (Fig. 17(d)) make decisions after comparison (to find minimum) or majority voting across multiple candidates. All the data sets are

TABLE II: Data sets used for chip measurements [37]–[39].

	Task	# of classes	Algorithm	Data set	Remarks (P : query input, D : data stored in array)
1	Face detection	2	SVM	MIT CBCL data set	- 100 query inputs tested - D : feature extractor and classifier combined 23 X 22 8-b coefficient - P : 23 X 22 8-b pixel image (face / non-face)
2	Event (gun shot) detection	2	MF	Gun shot Sound	- 100 query inputs tested - D : gun shot mono sound data with 256 8-b words - $P1$: gun shot sound contaminated by AWGN with 3 dB SNR or $P2$: Only AWGN with equal power of "signal + AWGN" in $P1$
3	Face recognition	64	TM	MIT CBCL data set	- 64 query inputs tested (due to array size limit) - 16 X 16 8-b pixel image for P and D - D : 64 candidate faces, P : one of the 64 candidate faces in D
4	Hand-written number recognition	4	k -NN	MNIST data set	- 100 query inputs tested - 4 classes from "0" to "3" (due to array size limit) - 16 X 16 8-b pixel image for P and D , D : 16 images per class, P : image from 4 classes

TABLE III: Multi-functions in each processing stage.

Stage	configurations
FR	① Normal read ② Digital to analog conversion ③ Scalar ADD or SUBT
BLP	① Scalar MULT ② BL -wise sampling ③ Absolute value
CBLP	① Aggregation ② Weighted aggregation
RDL	① MIN or MAX ② Linear combination ③ Send outside chip

TABLE IV: Configurations of each stage to enable four algorithms (the operations corresponding to numbers are described in Table III).

Mode	Algorithm	FR	BLP	CBLP	RDL
DP	SVM	②	①, ②	②	②
	MF	②	①, ②	②	②
MD	k -NN	②, ③	②, ③	①	③
	TM	②, ③	②, ③	①	①

processed fully on-chip except for k -NN, where the last step of majority voting was done off-chip.

IV. MEASURED RESULTS

This section describes measured results from the prototype IC in terms of its energy, delay, and accuracy, both at the stage level and at the inference task level.

A. Accuracy of FR

The measured results of sub-ranged FR of 8-b word D is shown in Fig. 18(a). The ΔV_{BL} generated by FR for all

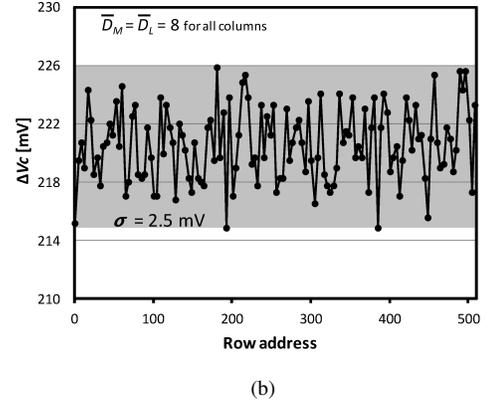
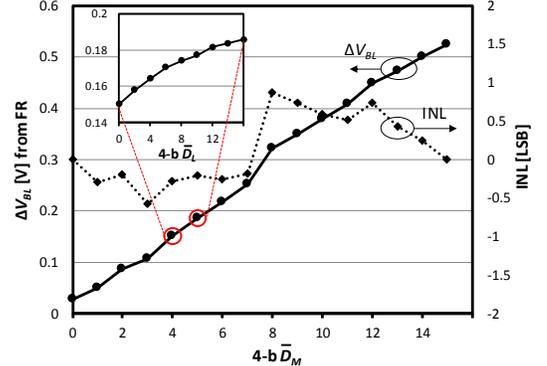


Fig. 18: Measured FR accuracy of 8-b D : (a) BL voltage drop ΔV_{BL} with sub-ranged read, and (b) the impact of spatial variations on CBLP output voltage drop ΔV_C , which is generated by storing identical ($D_M = D_L = 7$) across the entire BCA and accessing the word-rows via FR, then aggregating via the CBLP.

256 values of D was measured at the output of the column mux along the normal SRAM read path. The sudden jump when D transits from 7 to 8 is due to the large change in the average portion of transition time in the WL pulses. The overall integral non-linearity (INL) was found to be less than 0.87 LSB.

The variation in ΔV_{BL} was measured in Fig. 18(b), where

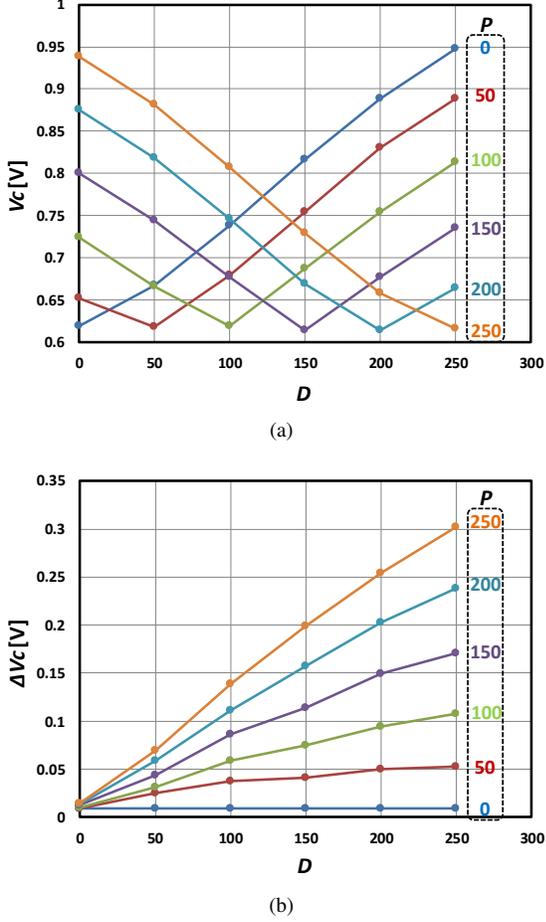


Fig. 19: Measured CORE analog output with 8-b operands D and P in the: (a) MD mode ($\sum |D_i - P_i| \propto V_C$), and (b) the DP mode ($\sum D_i P_i \propto \Delta V_C$), where the same data D and P are stored in all the columns.

the stored data $D_M = D_L = 7$ is chosen to generate the worst case (maximum) variation in ΔV_{BL} , as in this case, the BL is discharged by a single SRAM bitcell without the averaging effect. Figure 18(b) shows that the variation in ΔV_{BL} has a standard deviation $\sigma = 2.5$ mV ($\sigma/\mu = 1.1\%$), which is only 0.6% of the dynamic range (410 mV) of CBLP output V_C in this test mode. This small deviation arises because of the averaging effect during the aggregation of 128 BLP outputs in the CBLP stage as described in section II-A. We show that these errors result in negligible impact on the inference accuracy in the following sections.

B. Accuracy of CORE Output

We characterize the accuracy of the CORE output that includes FR, BLP and CBLP stages as shown in Fig. 19. The measured error magnitudes at V_C (from ideal linear trend) in the DP and MD modes are < 18 mV and < 28 mV with the mean of 4 mV and 8 mV, respectively, over all the combinations of (D, P). Though these errors are significantly larger than the chosen target resolution $V_{res} = 1$ mV, these

TABLE V: Error magnitude (compared to dynamic range) of each function per stage (*obtained from Monte Carlo simulations).

Error magnitude	FR	BLP	
		DP	MD
Deterministic	max: 5.8% (mean: 2.6%)	max: 6% (mean: 2.1%)	max: 7.5% (mean: 2.5%)
Random (σ/μ)*	max: 12.9%	max: 2.8%	max: 3.2%

errors do not affect the accuracy of inference tasks with sufficient decision margins as will be discussed next. In addition, it is possible to train the engine in presence of these errors to obtain circuit-optimized hyperparameters.

Table V summarizes the accuracy of each function per stage in terms of deterministic and random error contributions. Deterministic error contributions arise from the inherent non-linearity of the proposed circuits whereas random error contributions arise from process variations. As it is difficult to isolate the BLP stage from the FR stage, the deterministic error of BLP is obtained by comparing: 1) the estimated BLP output with the measured non-linear FR curve (in Fig. 18(a)) assuming ideal BLP operations, and 2) measured BLP output in Fig. 19, which includes both FR and BLP non-linearities. On the other hand, the random errors are obtained via Monte Carlo simulations, as it is difficult to extract the variations across a statistically significant number of BLP units. Each function generates a deterministic error less than 3% of the output dynamic range on average whereas the FR stage creates a dominant random error exhibiting $\sigma/\mu = 12.9\%$ variations. Therefore, the variation at CBLP output V_C can be estimated to be 1.1% ($= 12.9\%/\sqrt{128}$), which matches very well with the measured variation in Fig. 18(b).

C. Energy, Delay, and Accuracy

We consider the energy consumption of the CORE block only because its energy scales up with the number of banks and the BCA size. In contrast, the energy of the CTRL block is amortized over the number of banks and the BCA size. Introducing the reconfigurability does not incur additional energy penalty as the circuitry for unselected functions are disabled (and bypassed). The disabled circuitry adds negligible leakage energy, which is an order-of-magnitude smaller than that of the BCA.

We measured the CORE decision energy and decision accuracy for SVM (face detection, binary class) and TM (face recognition, 64-class) tasks in Fig. 20. The CORE decision energy was normalized by the number of 8-b data words processed per decision to obtain the energy-per-word as a function of BL swing per LSB $\Delta V_{lsb} = \Delta V_{BL}(\bar{D}_M = 15)/15$ in Fig. 20(a). Figure 20(a) indicates that CORE energy reduces at a rate of 0.2 pJ (0.4 pJ) per 20 mV for binary or DP mode (64-class or MD mode) task. The greater

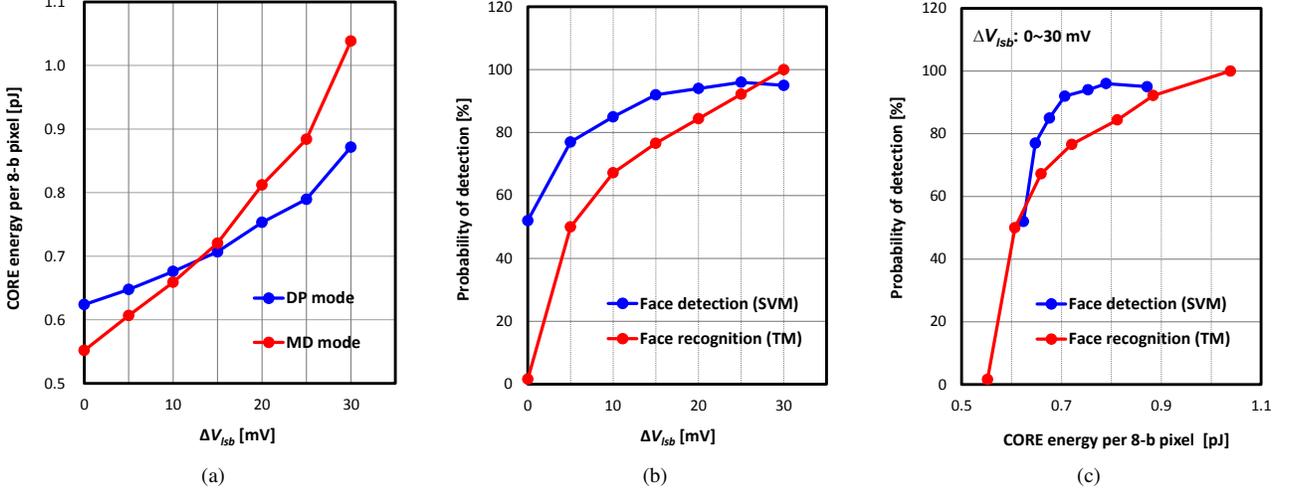


Fig. 20: Measured trade-offs between: (a) CORE energy vs. BL swing per LSB (ΔV_{lsb}), (b) decision accuracy (P_{det}) vs. ΔV_{lsb} , and (c) P_{det} vs. CORE energy.

TABLE VI: Comparison of energy efficiency, delay, accuracy with prior art.

	Process (nm)	# of algorithms	Memory size	Input bit precision	Decision throughput (Decisions/s)	Decision energy (pJ/decision)	Decision EDP (fJ-s)	Accuracy
This work	65 CMOS	4 (SVM, MF, k -NN, TM)	SRAM 512 X 256-b	D : 8b P : 8b	SVM: 9.3M	446	0.05	95 %
					MF: 18.5M	223	0.01	100 %
					TM: 312.5K	16.9K	54.0	100 %
					KNN: 312.5K	16.9K	54.0	92 %
8-b digital* (REF)	65 CMOS	synthesized dedicated processor per algorithm	SRAM 512 X 256-b	D : 8b P : 8b	SVM: 1.7M	4.5K	2.6	96 %
					MF: 3.4M	2.2K	0.6	100 %
					TM: 54.3K	93.0K	1715.3	100 %
					KNN: 54.3K	93.0K	1715.3	90 %
[2] [†]	14 Tri-gate	1 (k -NN)	128 byte	D : 8b, P : 8b	21.5M	3.4K	0.2	Not reported
[26]**	130 CMOS	1 (Adaboost)	SRAM 128 X 128-b	D : 1b, P : 5b	50M	633.4	0.01	90 %

* memory energy and delay measured from prototype IC, but those of digital computation obtained from post-layout simulations;

[†] single function with SRAM memory access cost not included;

** single function with 1b weight vector

slope of the energy vs. ΔV_{lsb} plot for MD mode and its higher energy consumption for $\Delta V_{lsb} > 15$ mV is because MD mode uses the replica BCA which causes additional voltage drop on the BL during FR.

The accuracy of the inference task is measured by the probability of detection (P_{det}) obtained by normalizing the number of queries correctly classified by the total number of queries. Figure 20(b) shows that the binary task is more robust than the 64-class task at the same ΔV_{lsb} . Furthermore, the binary and the 64-class task achieve $> 90\%$ detection accuracy for $\Delta V_{lsb} > 15$ mV and $\Delta V_{lsb} > 25$ mV, respectively. Figure 20(c) plots the detection accuracy against the CORE energy per 8-b pixel and shows the accuracy and energy trade-off.

Next, we compare the DIMA prototype with a conventional 8-b digital reference architecture (REF). REF is a 2-stage pipelined architecture comprising an SRAM of the same size as the one in the DIMA prototype, and a digital block

synthesized separately for realizing an SVM (DP mode) and a TM (MD mode). The energy and delay of the digital block in REF was estimated from post-layout simulations. The energy and delay of the SRAM in REF was measured from the DIMA prototype in the normal read mode. Figure 21 shows the energy breakdown for REF, DIMA (post layout simulations of the DIMA), and the DIMA prototype IC. The measured energy savings in the DP and MD modes are $10\times$ and $5.3\times$, respectively, due to small swing FR, BLP, and CBLP. Furthermore, the DIMA energy estimates obtained from post layout simulations are close to that obtained from measurements.

Table VI shows that the DIMA prototype IC achieves negligible ($\leq 1\%$) accuracy degradation for all four tasks as compared to REF. Additionally, DIMA requires $16\times$ fewer read accesses as compared to REF for a fixed data volume. This is due to DIMA's massive parallelism (128 8-b words per access) compared to the normal SRAM mode (only 8 8-b

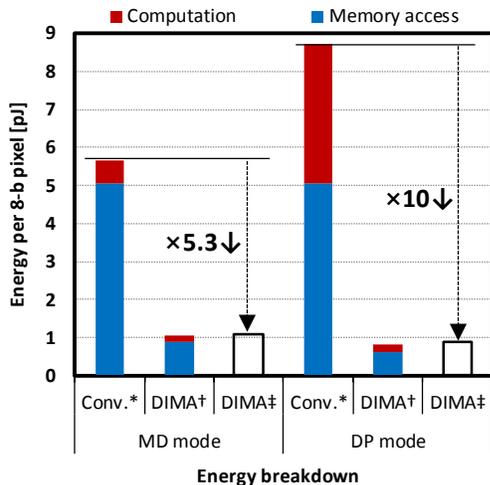


Fig. 21: CORE energy comparison of DIMA ([†]post-layout simulations, [‡]measured) vs. conventional (Conv.) digital architecture (*SRAM energy measured and digital computation energy from post-layout simulations).

words per fetch). Thus, the DIMA prototype IC provides a throughput gain of $5.8\times$ for MD mode tasks (TM and k -NN) and $5.3\times$ for DP mode tasks (SVM and MF). Therefore, the EDP is reduced by $32\times$ and $53\times$ in the MD and DP modes, respectively. As a result, the DIMA prototype IC implements four different algorithms achieving better decision accuracy and comparable energy-delay product (scaled for 65nm) than single function ICs [2], [27] listed in Table VI.

V. CONCLUSIONS AND FUTURE WORK

This paper describes a 65 mV multi-functional DIMA prototype IC, which supports four inference tasks: SVM, TM, k -NN, and MF. Measurement results demonstrate up to $53\times$ EDP reduction as compared to the conventional single-function digital architectures. Extensions to high density memory technologies such as NAND flash and to other ML algorithms can be considered. The multi-functional feature indicates the potential for designing programmable deep in-memory instruction set architecture along with compiler and software support.

ACKNOWLEDGMENTS

This work was supported by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by SRC and DARPA. The authors would like to acknowledge constructive discussions with Sean Eilert, Ken Curewitz, Naveen Verma, Boris Murmann, and Pavan Hanumolu.

REFERENCES

- [1] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.
- [2] H. Kaul, M. A. Anders, S. K. Mathew, G. Chen, S. K. Satpathy, S. K. Hsu, A. Agarwal, and R. K. Krishnamurthy, "A 21.5 M-query-vectors/s 3.37 nJ/vector reconfigurable k-nearest-neighbor accelerator with adaptive precision in 14nm tri-gate CMOS," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 260–261.
- [3] S. Park, K. Bong, D. Shin, J. Lee, S. Choi, and H.-J. Yoo, "A 1.93TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications," in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, 2015, pp. 1–3.
- [4] K. Kim, S. Lee, J.-Y. Kim, M. Kim, and H.-J. Yoo, "A 125 GOPS 583 mW network-on-chip based parallel processor with bio-inspired visual attention engine," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 136–147, 2009.
- [5] J.-Y. Kim, M. Kim, S. Lee, J. Oh, K. Kim, and H.-J. Yoo, "A 201.4 GOPS 496 mW real-time multi-object recognition processor with bio-inspired neural perception engine," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 1, pp. 32–45, 2010.
- [6] J. Oh, G. Kim, B.-G. Nam, and H.-J. Yoo, "A 57 mW 12.5 μ J/Epoch embedded mixed-mode neuro-fuzzy processor for mobile real-time object recognition," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2894–2907, 2013.
- [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 *Solid-State Circuits Conference (ISSCC), 2017 IEEE International*, 2017, pp. 244–245.
- [8] P. N. Whatmough, S. K. Lee, H. Lee, S. Rama, D. Brooks, and G.-Y. Wei, "A 28nm SoC with a 1.2 GHz 568nJ/prediction sparse deep-neural-network engine with >0.1 timing error rate tolerance for IoT applications," in *Solid-State Circuits Conference (ISSCC), 2017 IEEE International*, 2017, pp. 242–243.
- [9] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI," in *Solid-State Circuits Conference (ISSCC), 2017 IEEE International*, 2017, pp. 246–247.
- [10] K. Bong, S. Choi, C. Kim, S. Kang, Y. Kim, and H.-J. Yoo, "A 0.62 mw ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on haar-like face detector," in *Solid-State Circuits Conference (ISSCC), 2017 IEEE International*, 2017, pp. 248–249.
- [11] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, February 2014, pp. 10–14.
- [12] 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 *ACM Sigplan Notices*, vol. 49, no. 4, 2014, pp. 269–284.
- [13] B. Murmann, D. Bankman, E. Chai, D. Miyashita, and L. Yang, "Mixed-signal circuits for embedded machine-learning applications," in *Asilomar Conference on Signals, Systems and Computers*, 2015, pp. 1341–1345.
- [14] Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," in *International Symposium on Computer Architecture (ISCA)*, 2016, pp. 367–379.
- [15] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner *et al.*, "RAZOR: A low-power pipeline based on circuit-level timing speculation," in *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, 2003, pp. 7–18.
- [16] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mV robust Schmitt trigger based subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, 2007.
- [17] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200mV 6T SRAM in 0.13 μ m CMOS," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*. IEEE, 2007, pp. 332–606.
- [18] 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 Journal of Solid-State Circuits*, vol. 50, no. 5, pp. 1310–1323, 2015.

- [19] F. Frustaci, D. Blaauw, D. Sylvester, and M. Alioto, "Approximate SRAMs with dynamic energy-quality management," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 6, pp. 2128–2141, 2016.
- [20] R. Genov and G. Cauwenberghs, "Kerneltron: support vector "machine" in silicon," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1426–1434, 2003.
- [21] M. Kang, M.-S. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 8326–8330.
- [22] N. Shanbhag, M. Kang, and M.-S. Keel, "Compute memory," Jul. 4 2017, US Patent 9,697,877 B2.
- [23] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2015.
- [24] M. Kang, E. P. Kim, M.-S. Keel, and N. R. Shanbhag, "Energy-efficient and high throughput sparse distributed memory architecture," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2015.
- [25] M. Kang and N. R. Shanbhag, "In-memory computing architectures for sparse distributed memory," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 4, pp. 855–863, 2016.
- [26] M. Kang, S. K. Gonugondla, and N. R. Shanbhag, "A 19.4 nJ/decision 364K decisions/s in-memory random forest classifier in 6T SRAM array," in *IEEE European Solid-State Circuits Conf. (ESSCIRC)*, Sept 2017.
- [27] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE Journal of Solid-State Circuits*, 2017.
- [28] M. Kang, S. Gonugondla, A. Patil, and N. Shanbhag, "A 481 pJ/decision 3.4M decision/s multifunctional deep in-memory inference processor using standard 6T SRAM array," *arXiv preprint arXiv:1610.07501*, 2016.
- [29] K. J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*. IEEE, 2007, pp. 471–474.
- [30] 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 *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2016, pp. 21–24.
- [31] Z. Zhou, B. Pain, and E. R. Fossum, "CMOS active pixel sensor with on-chip successive approximation analog-to-digital converter," *IEEE Transactions on Electron Devices*, vol. 44, no. 10, pp. 1759–1763, 1997.
- [32] F. Arnaud, F. Boeuf, F. Salvetti, D. Lenoble, F. Wacquant, C. Regnier, P. Morin, N. Emonet, E. Denis, J. Oberlin *et al.*, "A functional 0.69 μm^2 embedded 6T-SRAM bit cell for 65nm CMOS platform," in *IEEE Symposium on VLSI Technology*, 2003, pp. 65–66.
- [33] D. Silver, A. Huang, and *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–503, 2016.
- [34] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision." in *ICML*, 2015, pp. 1737–1746.
- [35] S. C. Chung, "Circuits and methods of a self-timed high speed SRAM," Nov. 10 2015, US Patent 9,183,897.
- [36] E. Karl, Y. Wang, Y.-G. Ng, Z. Guo, F. Hamzaoglu, M. Meterelliyoz, J. Keane, U. Bhattacharya, K. Zhang, K. Mistry *et al.*, "A 4.6 GHz 162 Mb SRAM design in 22 nm tri-gate CMOS technology with integrated read and write assist circuitry," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 150–158, 2013.
- [37] "Center for biological and computational learning (CBCL) at MIT," 2000, <http://cbcl.mit.edu/software-datasets/index.html>.
- [38] Production Crate, "Gun Shot Sounds," <http://soundcrate.com/gun-related/>.
- [39] Y. LeCun and C. Cortes, "MNIST handwritten digit database," *AT&T Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2010.



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



Sujan K. Gonugondla (S'16) received the B.Tech. and M.Tech. degrees in Electrical Engineering from the Indian Institute of Technology Madras (IIT-M), Chennai, India in 2014. He is currently working towards a Ph.D. in Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, USA. His current research interests are in low-power integrated circuits specifically algorithm-hardware co-design for machine learning systems on resource-constrained platforms.



Ameya Patil (S'15) received the B.Tech. degree in 2014 from the department of Electrical Engineering at the 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 (F'06) 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 interests are in the design of energy-efficient 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 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 Intersymbol Communications, Inc., (acquired in 2007 by Finisar Corporation) a semiconductor start-up that provided DSP-enhanced mixed-signal ICs for electronic dispersion compensation of OC-192 optical links. 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.