AN ENERGY-EFFICIENT VLSI ARCHITECTURE FOR PATTERN RECOGNITION VIA DEEP EMBEDDING OF COMPUTATION IN SRAM

Mingu Kang[⋆]*, Min-Sun Keel*[⋆]*, Naresh R. Shanbhag*[⋆]*, Sean Eilert*[†]*, and Ken Curewitz*[†]

[⋆]Dept. Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
[†]Micron Technology, Inc

ABSTRACT

In this paper, we propose the concept of *compute memory*, where computation is deeply embedded into the memory (SRAM). This deep embedding enables multi-row read access and analog signal processing. Compute memory exploits the relaxed precision and linearity requirements of pattern recognition applications. System-level simulations incorporating various deterministic errors from analog signal chain demonstrates the limited accuracy of analog processing does not significantly degrade the system performance, which means the probability of pattern detection is minimally impacted. The estimated energy saving is 63 % as compared to the conventional system with standard embedded memory and parallel processing architecture, for 256×256 target image.

Index Terms— Compute memory, Pattern recognition, Machine learning, Associative memory, Analog processing

1. INTRODUCTION

Machine learning algorithms are finding use in various applications including embedded-sensor networks and computer vision. Pattern recognition is one of the key kernels for classification in machine learning, and for the multimedia applications such as object detection or speech recognition. Computation in pattern recognition is repetitive, requires frequent memory accesses, and consumes significant energy. In this paper, we consider the most widely used pattern matching algorithm based on the sum of absolute differences (SAD) metric as described below [1]:

$$\begin{aligned}(x_{opt}, y_{opt}) &= \underset{(x,y)}{\arg\min}\, SAD(x,y)\\ &= \underset{(x,y)}{\arg\min} \sum_{i=0}^{M_P-1}\sum_{j=0}^{N_P-1} |D(x+i, y+j) - P(i,j)| \qquad (1)\end{aligned}$$

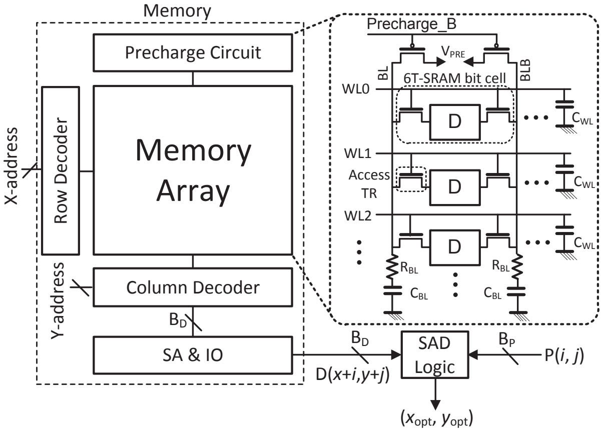


Fig. 1: Architecture of conventional SRAM-based SAD system.

where (x, y) is the pointer address in the original image (D), and (i, j) is an index in the pattern template (P) with a window size of $M_P \times N_P$. The location (x_{opt}, y_{opt}) with the minimum SAD is the best matching point; template-matching pattern recognition is a process to search for this minimum SAD point in a given image.

Figure 1 shows the block diagram of a SAD-based pattern recognition system using a conventional SRAM with 6-transistor (6T) bit cell, where C_{BL}, R_{BL}, C_{WL}, and V_{PRE} are the bit-line (BL) capacitance, BL resistance, world-line (WL) capacitance, and BL precharge voltage, respectively. The numbers of bits per word for the D and P are denoted by B_D, and B_P, respectively. During a row-access, one WL is activated enabling a row of bit cells to discharge the precharged BL voltages according to the bit cell data. The change in BL voltage (ΔV_{BL}) is converted into full-swing logic levels by sense-amplifiers (SA) to be processed by the external SAD logic. Thus, in order to read out one B_D-bit word stored in a row, B_D BLs need to be discharged, which requires B_D BL precharges. This contributes significantly to the SRAM energy consumption. The energy consumption for processing a single word (READ access + SAD logic) in a conventional SRAM-based SAD system can be expressed as

This work was supported by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by SRC and DARPA.

$$\begin{aligned} E_{conv} \quad = \quad & B_D C_{BL} \Delta V_{BL} V_{PRE} + C_{WL} V_{DD}^2 + \\ & B_D E_{SA} + E_{leak} + E_{logic} \end{aligned} \tag{2}$$

where E_{SA}, E_{leak}, and E_{logic} represent the energy consumption due to SA, leakage in bit cell array, and SAD logic, respectively. Energy consumption from other blocks such as address decoders are assumed to be negligible. The first term in (2) is the energy consumption due to BL precharge. This term dominates because C_{BL} is the largest capacitance in an SRAM. The overall energy efficiency can be enhanced by reducing energy in memory, especially in the READ operation.

In this paper, we propose the concept of compute memory (CM), where computation (SAD computation) is deeply embedded into the memory array (SRAM) for aggressive energy efficiency and parallel processing. This deep embedding enables: 1) multi-row READ access, and 2) analog signal processing. CM exploits the relaxed precision and linearity requirements of pattern recognition applications.

Prior work in integrating memory and computation includes [2] and [3], where associative memory is employed to find the vector with minimum Hamming distance from reference data only for the limited application of 1-D vector comparison. Several processor-in-memory architectures have been proposed such as the computational RAM (C·RAM) [4], smart memories [5], and intelligent RAM (IRAM) [6]. These approaches maintain the separation between a low-swing/low-SNR memory array and a high-swing/high-SNR logic. *Kerneltron* eliminates the boundary by embedding processor into memory cells [7], where analog processing is used in the limited portion of computation. Thus, it suffers from frequent analog-to-digital conversion.

2. PROPOSED APPROACH: COMPUTE MEMORY

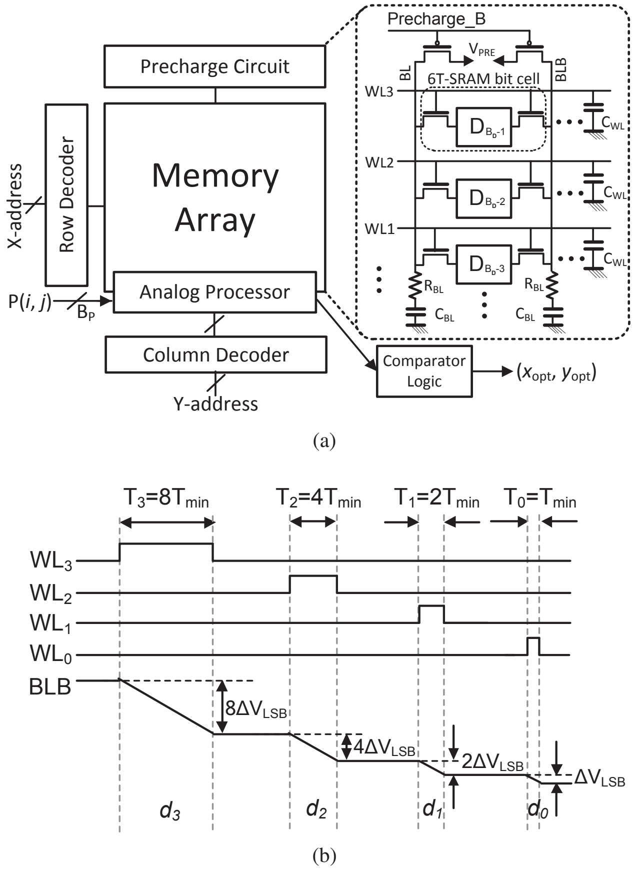
2.1. The Compute Memory Architecture

The CM illustrated in Fig. 2(a) employs two concepts:

1) multi-row READ: vertically-stored B_D bits are read out per single precharge employing width-modulated WL pulse. This generates a ΔV_{BL} equal to a binary-weighted sum of the data word D.

2) analog SAD computation: ΔV_{BL} and ΔV_{BLB} representing D are processed along with the template pattern P in the discrete-time analog domain to compute the SAD in (1). Additionally, a simple sequencer is employed to slide the template P across the memory array storing D.

The CM achieves superior energy-efficiency over conventional systems (Fig. 1) because the multi-row READ process reduces the number of precharge operations, and the analog SAD computation is a low-swing operation. However, both features 1) and 2) result in non-linearity and low-SNR computation. Fortuitously, pattern recognition has relaxed linearity and precision requirements. The CM exploits this feature. As the processor is directly embedded into memory, off-chip data transfer from memory to processor is not required. Thus, the system throughput is not limited by the number of I/O ports or bus width.



(a)

(b)

Fig. 2: The proposed compute memory: (a) architecture, and (b) timing diagram with pulse width modulation (PWM) (for 4-bit word read-out, D = 1111b').

In conventional system, parallel computation of the SAD can be supported by reading data from multiple banks of memory at the same time. On top of that, CM enables parallel computing even in one bank. Because data transfer is not required between processor and memory, SAD can be computed at different locations of array in parallel with multiple windows of template pattern.

2.2. Multi-row READ

Let the data D in (1) be represented by a B_D-bit vector $\mathbf{d} = \{d_0, d_1, ..., d_{B_D-1}\}$. Then, $D = \Sigma_{m=0}^{B_D-1} 2^m d_m$ is the decimal value of D. In a multi-row READ operation:

(a) $\mathbf{d}$ is stored in one column.

(b) B_D WLs are activated per precharge.

(c) the WL access pulse width T_m for $d_m \propto 2^m, m \in [0, B_D - 1]$, i.e., gate control signals of the access transistors are pulse-width modulated.

By ensuring that $T_m \ll RC$ time constant of the BL and BLB, we can show that the change in the BL and BL voltage at the end of the multi-row READ process is given by

$$\Delta V_{BL}(\mathbf{d}) = \frac{V_{PRE}}{R_{BL}C_{BL}} T_{min} \sum_{m=0}^{B_D-1} 2^m \overline{d_m} \qquad (3)$$

$$\Delta V_{BLB}(\mathbf{d}) = \frac{V_{PRE}}{R_{BL}C_{BL}} T_{min} \sum_{m=0}^{B_D-1} 2^m d_m \qquad (4)$$

where T_{min} is the LSB pulse width. Figure 2(b) shows how a 4-bit word, 1111b' is read out on BLB via binary-weighted WL pulse widths T_0 to T_3.

2.3. Analog SAD Processing

The SAD computation requires the evaluation of absolute difference (AD), $|D-P|$ (see (1)). The AD can be written as:

$$\begin{aligned} |D-P| &= max(D-P, P-D) \\ &= max(D+\overline{P}+1, P+\overline{D}+1) \\ &\Rightarrow max(D+\overline{P}, P+\overline{D}) \end{aligned} \qquad (5)$$

where $\overline{D}$ and $\overline{P}$ are the 1's complement of D and P, respectively. Note: $\overline{D}$ and $\overline{P}$ are automatically available due to the complementary nature of the SRAM bit cell.

The template pattern P is stored with the polarity opposite to that of D. Thus, a multi-row READ of D followed by a multi-row READ of P results in ΔV_{BL} and ΔV_{BLB} being proportional to $P+\overline{D}$ and $D+\overline{P}$, respectively.

Local BL comparators provide the maximum of ΔV_{BL} and ΔV_{BLB}, and hence the AD, $|D-P|$. The output of each BL comparator determines one of the terms in (5). These outputs are summed up via a capacitive network using a charge transfer mechanism to generate the SAD. A global comparator keeps track of the minimum SAD to generate the final output (x_{opt}, y_{opt}) in (1).

2.4. Behavioral Models with Circuit Non-idealities

The analog-intensive CM operation is subject to a number of circuit-level non-idealities. Dominant among these are:

(a) non-linearity of the multi-row READ process. This is caused by voltage-dependent discharge path resistance, R.

(b) local transistor threshold voltage V_{th}-mismatch across bit cells caused by random dopant fluctuations.

Figure 3 shows the behavior of multi-row READ on BLB vs. $D+\overline{P}$. The non-linearity of the multi-row READ process is represented by the integral non-linearity (INL). Figure 3 shows that the dynamic range of ΔV_{BL} is limited to 0.9 V in order to obtain INL within ±5 LSB. Even though this non-linearity degrades P_{det}, the effect is mitigated by the maximum operation in (5), by which the local comparator always selects less distorted one between BL and BLB. Fitting a 4th order polynomial to the curve in Fig. 3 results in:

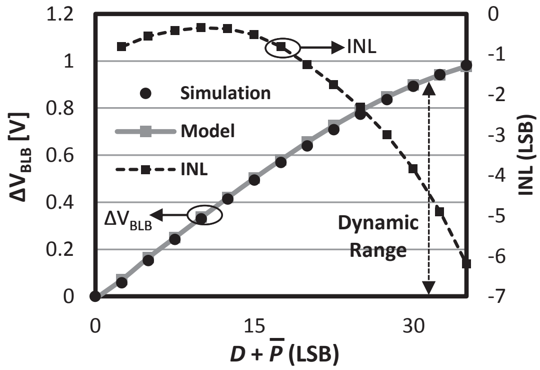


Fig. 3: ΔV_{BLB} and INL vs. $D+\bar{P}$ during multi-row READ for $B_D = 4$ with $V_{PRE} = 1.1V$.

$$D' = \sum_{i=1}^{4} c_i D^i; \qquad \overline{P}' = -D' + \sum_{i=1}^{4} c_i (D+\overline{P})^i \qquad (6)$$

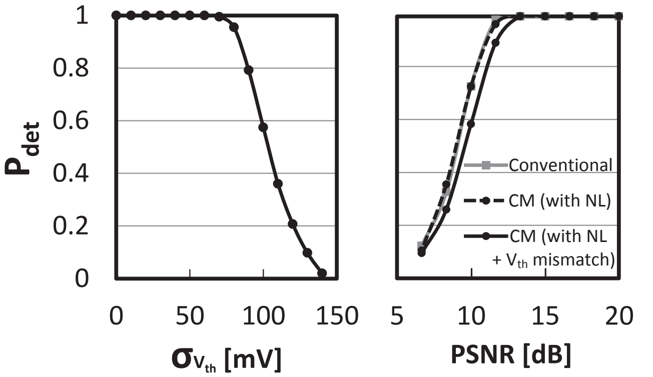
where D' and $\overline{P}'$ are the distorted version of D and $\overline{P}$, respectively, and $\overline{P}'$ depends on D because $\overline{P}$ is read after D. c_i's are the fitting parameters, which depend upon the process technology including V_{th}, carrier mobility, saturated carrier velocity, and channel length modulation parameter. In Fig. 3, $c_1 = 1$, $c_2 = 0.0111$, $c_3 = -0.0005$, and $c_4 = 4.05 \times 10^{-6}$. The impact of V_{th}-mismatch is modeled as:

$$\begin{aligned} \widehat{I}_{DS} &= K_n(V_{DD} - \widehat{V}_{th})^\alpha, \;\; \widehat{V}_{th} \sim N(V_{th}, \sigma^2_{Vth}) \\ \beta_m &= \left[\frac{1-\widehat{V}_{th}/V_{DD}}{1-V_{th}/V_{DD}}\right]^\alpha \\ \widehat{D} &= \sum_{m=0}^{B_D-1} 2^m d_m \beta_m, \;\; \widehat{\overline{P}} = \sum_{m=0}^{B_D-1} 2^m \overline{p_m} \beta_m \end{aligned} \qquad (7)$$

where K_n is transconductance factor of transistor. $\widehat{I}_{DS}$ expresses the modified drain current due to V_{th} variation which is modeled as a Gaussian random variable ($\widehat{V}_{th}$) with variance σ^2_{Vth}, and β_m is a random variable, which is a function of $\widehat{V}_{th}$. It is assumed that all the access transistors of SRAM bit cells are operating in the saturation region, which sets $\alpha = 1.2$ [8]. Thus, V_{th}-mismatch results in scaling of the binary values d_m and p_m by the scalar β_m.

Substituting (7) in (6), the result in (1) provides the following system model which captures both circuit non-idealities.

$$(x_{opt}, y_{opt}) = \arg\min_{(x,y)} \sum_{i=0}^{M_P-1} \sum_{j=0}^{N_P-1} |\widehat{D}'(x+i, y+j) + \widehat{\overline{P}}'(i,j)| \qquad (8)$$



(a) (b)

Fig. 4: Probability of detection (P_{det}): (a) with local mismatch of transistors' threshold voltage, (b) with pixel noise.

This model is employed in section 3 to study the impact of circuit non-idealities on the probability of detection (P_{det}).

2.5. Energy Models

The energy consumption of the CM for SAD processing a word is given by

$$\begin{aligned} E_{CM} \;=\;& C_{BL}(\Delta V_{BL\_CM} + \Delta V_{BLB\_CM})V_{PRE} + \\ & C_{WL}V_{DD}^2 + E_{comp} + E_{leak\_CM} + E_{add} \quad (9) \end{aligned}$$

where E_{comp} and E_{add} are the energies consumed by the local comparators and the analog adder, respectively. The energy consumption from global comparator is negligible. Note: the first term in (9) is much smaller than that in (2), because the term, $\Delta V_{BL\_CM} + \Delta V_{BLB\_CM} < 3 \times \Delta V_{BL}$, while $B_D \geq 8$. This reduces the energy consumption from BL precharge significantly. Moreover, only one comparator is used per word in CM, whereas B_D SAs are required per word in the conventional memory. The term, E_{add} is much smaller than E_{logic} in (2) because the switched-capacitor adder consumes energy only in switch control logic and buffers.

3. SIMULATION RESULTS

In this section, system simulations using (8) are performed assuming 4 banks of 512×256 bit CM in which 8-bit gray scale image of size 256×256 is pre-stored. The 8-bit word is implemented by subranging into two 4-bit words. The standard test image "Lena" is used for D and a 16×16 block of pixels around the right eye is used for P.

3.1. Detection Probability

Figure 4(a) shows the P_{det} with local V_{th} mismatch (σ_{Vth}) and 10-mV offsets in local and global comparators. P_{det} is not degraded until $\sigma_{Vth} = 80\,mV$. Because $\sigma_{Vth} < 40\,mV$ in 45-nm process technology [9], V_{th} mismatch is successfully rejected by the inherent averaging effect of SAD. Next, the input SNR is degraded by adding Gaussian noise (σ_η) for performance comparison of CM with conventional system. Figure 4(b) plots P_{det} for 1) conventional system, 2) CM with the non-linearity (NL) model from (6), and 3) CM with the NL, 10-mV offsets in comparators, and V_{th}-mismatch model from (8), where σ_{Vth} = 20 mV considering the size of access transistor [9]. From the graphs of the first and the second configurations, it is concluded that non-linearity affects P_{det} negligibly. The discrepancy between the second and the third configurations by V_{th} mismatch and comparator offset appears when peak signal-to-noise ratio (PSNR) < 12 dB. Typical PSNR values are > 20 dB. Thus, it is concluded that the non-idealities from the CM is effectively compensated by the inherent error resiliency in SAD.

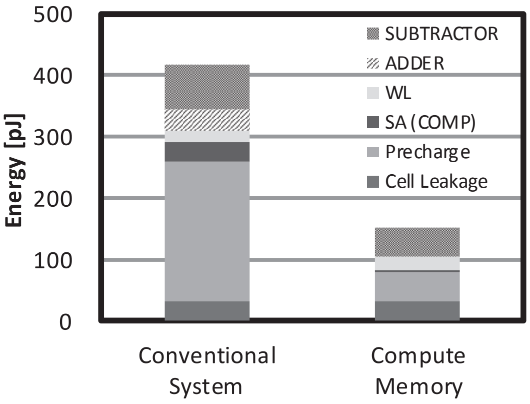


Fig. 5: Estimated energy savings.

3.2. Energy Savings

Based on energy model in (9), estimated energy consumption is plotted in Fig. 5. About 63 % energy saving in overall system is achieved with minimal impact on P_{det} (No degradation of P_{det} when PSNR > 12 dB). This significant energy savings are achieved mostly by reduced BL precharge and analog summation as indicated by (9). Even though it is not explicitly described in (2) and (9), in conventional systems, frequent data transfer from memory to external processor requires extra energy to drive parasitic capacitance loading from I/O ports and routing. In this sense, even more energy saving is expected in a practical implementation.

4. CONCLUSION

The compute memory (CM) embeds computation seamlessly into memory. Though this paper focused on SRAMs, other memory topologies can be considered. Programmable CM architecture can also be explored.

5. REFERENCES

[1] M. Mori and K. Kashino, “Fast template matching based on normalized cross correlation using adaptive block partitioning and initial threshold estimation,” in *IEEE Int. Symp. on Multimedia (ISM)*, 2010, pp. 196–203.

[2] H. J. Mattausch, T. Gyohten, Y. Soda, and T. Koide, “Compact associative-memory architecture with fully parallel search capability for the minimum Hamming distance,” *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 218–227, Feb. 2002.

[3] Y. Oike, M. Ikeda, and K. Asada, “A high-speed and low-voltage associative co-processor with exact Hamming/Manhattan-distance estimation using word-parallel and hierarchical search architecture,” *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1383–1387, Aug. 2004.

[4] D. G. Elliott, M. Stumm, W. M. Snelgrove, C. Cojocaru, and R. McKenzie, “Computational RAM: Implementing processors in memory,” *IEEE Des. Test. Comput.*, vol. 16, no. 1, pp. 32–41, 1999.

[5] K. Mai, T. Paaske, N. Jayasena, R. Ho, W. J. Dally, and M. Horowitz, “Smart memories: A modular reconfigurable architecture,” in *ACM Proc. Int. Symp. Comput. Arch.*, 2000, pp. 161–171.

[6] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, “Intelligent RAM (IRAM): Chips that remember and compute,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 1997, pp. 224–225.

[7] R. Genov and G. Cauwenberghs, “Kerneltron: support vector,” *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1426–1434, 2003.

[8] T. Sakurai and A. R. Newton, “Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas,” *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, 1990.

[9] K. J. Kuhn, “Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale cmos,” in *IEEE Int. Electron Devices Meeting (IEDM)*, 2007, pp. 471–474.