Extended Abstracts of the 2018 International Conference on Solid State Devices and Materials, Tokyo, 2018, pp545-546

Technology-Driven Emerging Computational Models and Systems

Srivatsa Rangachar Srinivasa, Nagadastagiri Reddy C, Nicholas Jao, Akshay Krishna Ramanathan, John Sampson and Vijaykrishnan Narayanan

The Pennsylvania State University
E-mail : {sxr5403, nrc53, naj5075, axr499, sampson}@psu.edu, vijay@cse.psu.edu

Abstract

The landscape of device innovations is rapidly changing not only influenced by CMOS scaling nearing physical limits, but also due to the quest for new computational models that go beyond tradition Von Neumann architectures. This paper will illustrate the influence of emerging devices, memories and interconnect innovations at the system level. It will also highlight coordinated device-circuit-system interactions that have enabled promising new systems for analytics, machine learning, and internet-of-things.

1. Introduction

The looming end of the roadmap for CMOS scaling and the gargantuan rise of machine learning promises an exciting era for semiconductors. The access to a large amount of labeled data and the computational capacity offered by Moore's law has triggered unprecedented interest in machine learning innovations. The computational needs continue to grow to support processing of even larger datasets and design of more complex intelligent systems. To support this growth even as Moore's law nears its end requires radical shifts in our computational circuits and models.

This paper explores two complementary circuit techniques that leverage process and device technology innovations. The first one focuses on in-memory computing. One of the significant challenges of current Von Neumann architectures is the ability to move data from a separate memory into the processing unit. The memory wall problem is especially a concern when operating with deep neural networks that have significant amount of intermediate data and configuration storage requirements. The ability to compute in-situ in the memory reduces the need to transfer data and mitigates the power and performance overheads of data movement. We present an approach using a monolithic 3D technology to integrate logic operations on a SRAM buffer and the use of a cross-point memory to support multi row write for database applications. As a corollary, we have explored the integration of distributed memory elements closely with logic to support instant-on/instant-off processors [1].

The second computational paradigm focuses on physics-based intrinsic computing. Our prior efforts have focused on leveraging behavior of an IMT-based Transistor for neuron behavior and the use of a weakly coupled system for image analytics [2]. In this work, we illustrate the use of weakly coupled IMT-based oscillators for corner detection and indicate how the algorithms should adapt to tap the intrinsic computational capabilities.

2. In-Memory Computations enabled by Monolithic 3D integration technology.

Monolithic 3D integration (M3D-IC) [3] is an emerging technology that overcomes integration and connectivity limitations of TSV based integration. Through high density interconnects (M3D via), storage nodes of the SRAM cells are directly accessible which further paves way for novel In-Memory computation support. Several published works enable bitwise Boolean operations at the cell level granularity [3] [4]. Arithmetic operations can also be performed by combination of computing at the cell level and array level [5].

Convolution operation is one of the most common and a primitive computation employed by various machine learning algorithms including Convolutional Neural Networks (CNN). Convolution operation has two parts to it represented by equation 1. First one is the multiplication of weights and the feature vectors and the second part is adding the bias.

$$Convolution = \sum_i WiXi + b \qquad (1)$$

Fig. 1 shows the computation of convolution by a standard Von-Neumann computer and an In-Memory processor enabled by M3D-IC. Due to very high number of computations involved, output of the first step needs to be stored and then retrieved from the memory for adding the bias. The steps involved in computing is shown in fig. 1(a). Frequent data movement in and out of the memory will cause computation slow down and huge energy consumption. While In-memory computational support enabled by the M3D-IC helps in reducing the data traffic through the highly parasitic interconnects. Fig. 1(b) pictorially represents the 3D-SRAM memory structure and the computational steps. Once the output of step

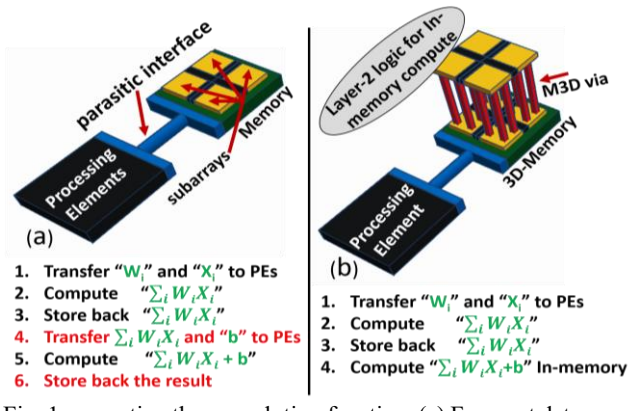

Fig. 1 computing the convolution function. (a) Frequent data movement in and out of the memory. (b) In-memory computations enabled by M3D-IC technology reduces frequent data movement out of the memory.

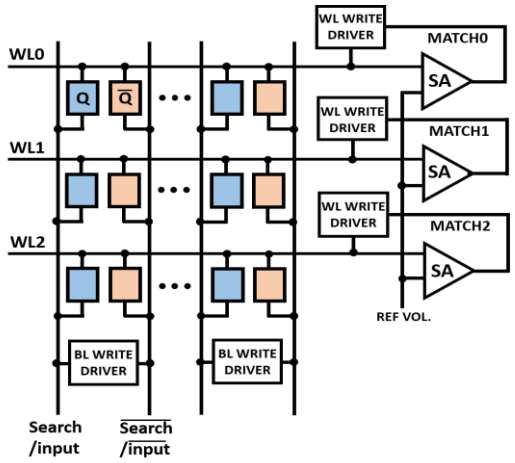

Fig. 2 Schematic representation of conditional Multi Write.

1 is stored in the memory, bias can be added to it to obtain the final result in parallel across the subarrays. This offers less data movement and computations in parallel and thereby providing energy efficiency.

3. Conditional Multi Write (CMW) in cross point NVMs

We have designed a cross point based Phase Change Memory (PCM) capable of performing conditional multi write operations (CMW). CMW searches for a particular data within the memory and replaces all its occurrences with the desired data on the matched rows. To search for the required data, we configure the peripherals to achieve the CAM functionality. Search voltages are asserted on the bitlines, wordlines are kept floating and ref.voltages are set according to the match voltage. Matched data is responsible for charging the wordlines. SA outputs a HIGH voltage if the wordline voltage is greater than the ref.voltage. SA output is fed into the WL WRITE DRIVER (fig. 2) which intern determines whether or not the wordline is biased for write mode. While, BL WRITE DRIVER prepares the write data. Once multiple rows are asserted in this way, operation is completed. Many of the search and replace queries in database workloads will benefit from this logic.

4. FAST Corner Detection using Coupled Oscillators

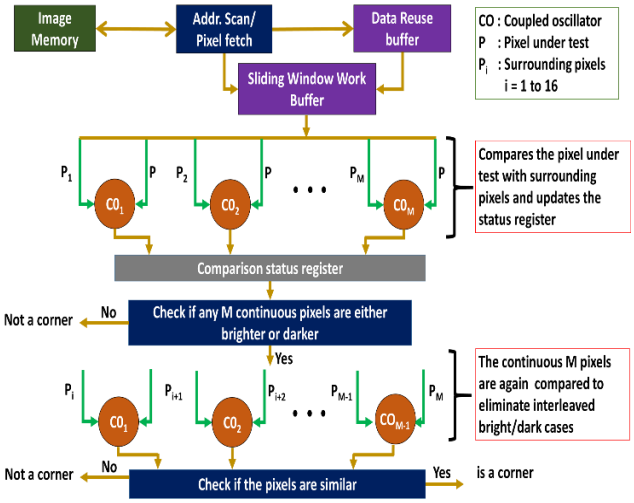

Fig. 3 Dataflow for FAST corner detection using coupled oscillators

Features from Accelerated Segment Test (FAST) corner detection algorithm compares a pixel with its surrounding 16 pixels on a Bresenham circle of radius 3. If the pixel is either darker or brighter than the N contiguous pixels on the circle, it will be marked as a corner. It involves systematic parallel comparisons uniform across all the pixels which makes it a favorable candidate for hardware acceleration. Coupled oscillator is one of the well-known paradigms in non-Boolean computing platforms and has been used for several image processing applications [6] [7]. Coupled oscillator exhibits tunable resistive and capacitive coupling with the input voltages and can be leveraged to find the input voltage difference [6]. We use this principle to map the comparison operation in FAST algorithm to coupled oscillator based design. Fig. 3 shows the overall dataflow design of the mapping. Two stages of comparison operations are needed to identify whether a pixel is a corner. In the first stage, pixel under test will be compared with the surrounding pixels to identify if it's either brighter or darker than N contiguous pixels (this comparison gives only the magnitude of difference but not the direction). In the second stage, to avoid the interleaved bright-dark cases, N contiguous pixels identified from the first stage will be compared with each other to identify whether they are all similar. We used the experimental graphs modelling the behavior of coupled oscillator [6] to evaluate the algorithm and compared with the reference algorithm. Albeit approximate computing, coupled oscillator based design shows performance comparable to the reference algorithm.

Acknowledgements

This work is supported in part by NSF Expeditions in Computing CCF-1317560 and by SRC JUMP center for Research on Intelligent Storage and Processing-in-memory along with CBRIC.

References

[1] K. Ma *et al.*, "Nonvolatile Processor Architectures: Efficient, Reliable Progress with Unstable Power," in *IEEE Micro*, vol. 36, no. 3, pp. 72-83, May-June 2016.

[2] M. Jerry *et al.*, "Phase transition oxide neuron for spiking neural networks," *2016 74th Annual Device Research Conference (DRC)*, Newark, DE, 2016, pp. 1-2.

[3] F. K. Hsueh *et al.*, "TSV-free FinFET-based Monolithic 3D+-IC with computing-in-memory SRAM cell for intelligent IoT devices," *2017 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp. 12.6.1-12.6.4.

[4] Srivatsa et al, "A Monolithic-3D SRAM Design with Enhanced Robustness and In-Memory Computation Support", ISLPED 2018 [Accepted].

[5] N. Jao *et al.*, "Harnessing Emerging Technology for Compute-In-Memory Support", *2018 ISVLSI* [Accepted].

[6] N. Shukla *et al.*, "Pairwise coupled hybrid vanadium dioxide-MOSFET (HVFET) oscillators for non-boolean associative computing," *2014 IEEE International Electron Devices Meeting*, San Francisco, CA, 2014, pp. 28.7.1-28.7.4.

[7] N. Shukla *et al.*, "Ultra low power coupled oscillator arrays for computer vision applications," *2016 IEEE Symposium on VLSI Technology*, Honolulu, HI, 2016, pp. 1-2.