IDEC Chip Design Contest

High-Performance and Energy-Efficient Computing-In-CAM Architecture for Binary Neural Network Acceleration

Fatia Uftiani Putri and Yeongkyo Seo

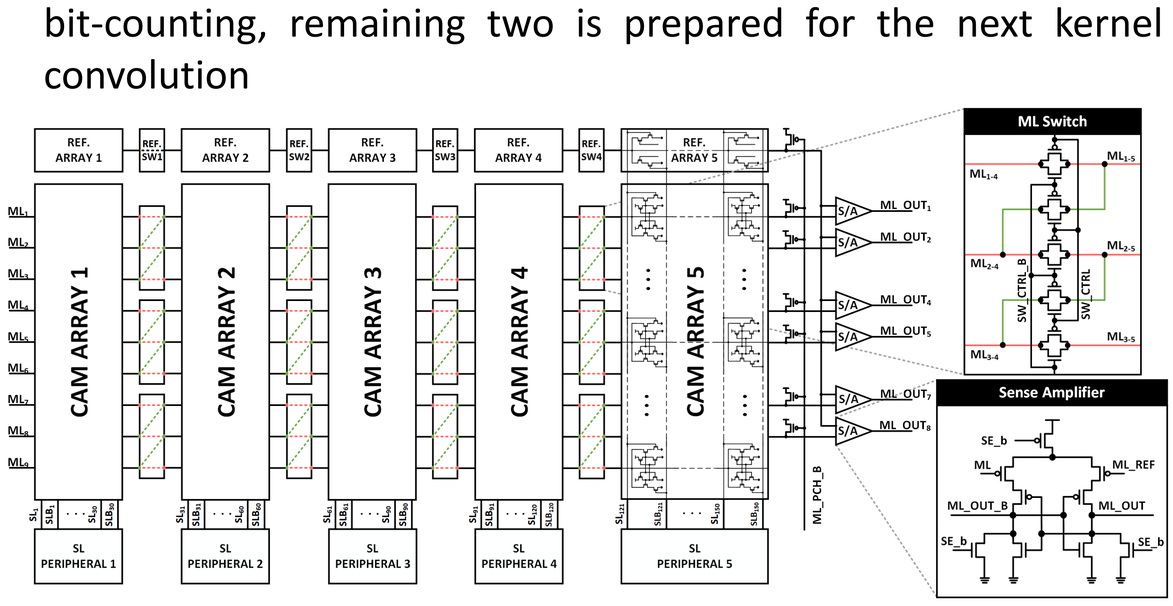
Dept. of Electrical and Computer Engineering, Inha University

▪ MOTIVATION

- Binary neural networks (BNNs), alternative to CNNs, utilize CAM memory for BNN accelerator, called CAM-BNN
- The downward sliding (reordering) phase in BNN-CAM is a time-consuming process in terms of the number of cycles
- Proposing a new method employs multi-bitline techniques and a single-cycle reordering circuit to enhance the performance and power efficiency

▪ BNN-CAM ARRAY STRUCTURE

- Implemented for accelerating 2nd convolution layer of LeNet-5 using MNIST dataset
- The proposed architecture consists of five banks with 9 × 30 CAM
- Reference array acts as the activation function
- ML switches connect rows between banks
- Computing four non-overlapping kernels in parallel to achieve 4x higher performance in XNOR bit-counting operations
- Only 6 ML peripheral circuits in last bank, four for parallel XNOR bit-counting, remaining two is prepared for the next kernel convolution



<Proposed circuit configuration related to XNOR bit-counting operation>

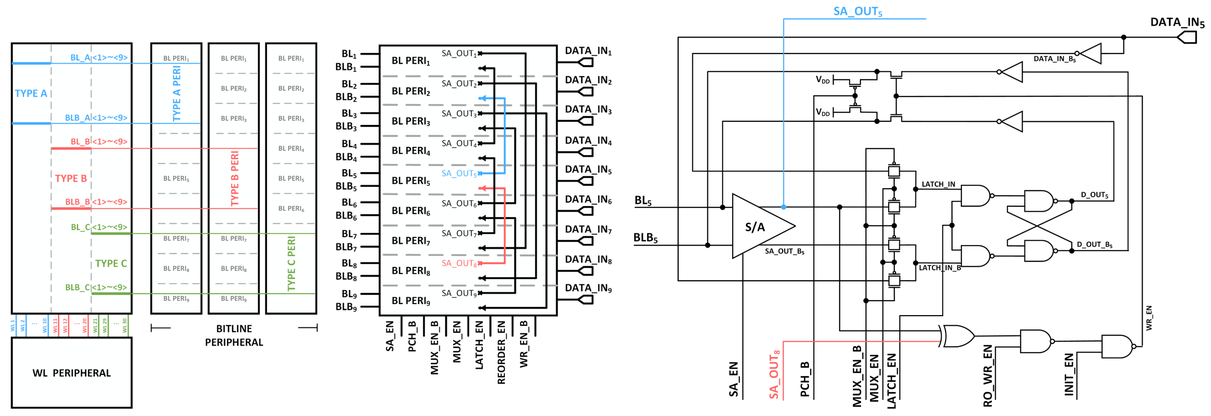
▪ IMPLEMENTED TECHNIQUE

MULTI-BITLINE FOR PARALLEL REORDERING

- The fixed number of cycles and ordering patterns required for write and read operations provide significant advantages for multi-bitline implementation
- The triple bitline is the most number of bitlines that can be applied without additional layout penalty area
- The triple bitline implementation reduced the initialization cycle from 30 cycles to 10 cycles, and the reordering phase cycle from 12 cycles to 4 cycles

SINGLE-CYCLE REORDERING WITH OVERWRITE TERMINATION

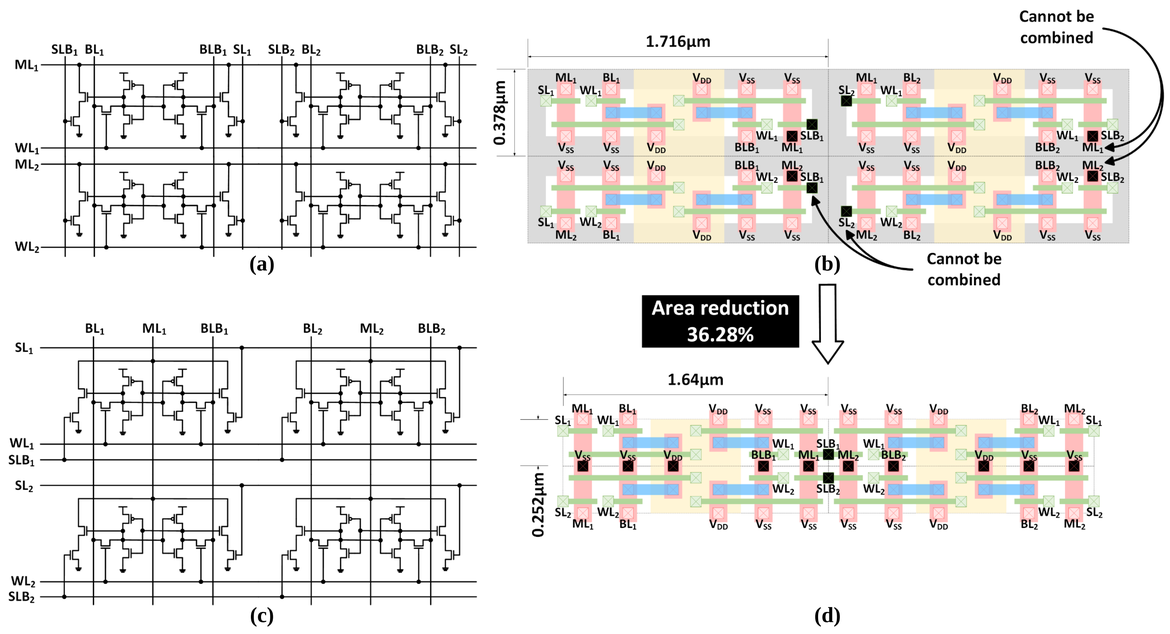
- The stored data from the target row is compared to the data located three rows below in a single cycle
- Only need to activate WL once to perform both read and write operations in reordering phase
- An XOR gate is employed to determine whether the overwrite termination technique should be enabled



(a) (b) (c)

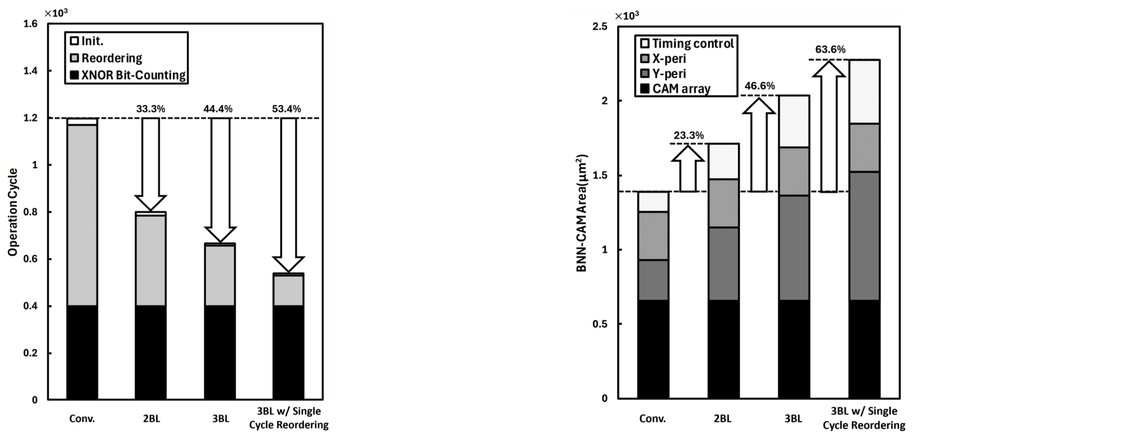
<(a) Triple bitline technique. Single-cycle reordering (b) configuration and (c) related bitline peripheral>

▪ HIGH DENSITY BNN – CAM CELL



<(a) Schematic and (b) layout of CPU-use 2×2 CAM. (c) Schematic and (d) layout of 2×2 BNN-CAM.>

▪ RESULTS COMPARISON



	Total Cycle	Energy(pJ)	Area(µm²)	FoM(nJ.m²)	Percentage
Conventional	1,198	1,124.8	1,390.1	1.87	100%
Double Bitline	799	978	1,713.5	1.34	71.7%
Triple Bitline	666	945.2	2,037.2	1.28	68.5%
Triple Bitline w/ Single Cycle Reordering	558	916.4	2,274.6	1.16	62%

*Due to IR drop, the slew rate of the chip did not work properly.
The layout design of the power supply part is planned to be changed.

인하대학교 INHA UNIVERSITY

Department of Electrical and Computer Engineering

The chip fabrication and EDA tool were supported by the IC Design Education Center (IDEC), South Korea.

반도체설계교육센터 IC DESIGN EDUCATION CENTER