

## Background

Introducing a cutting-edge hardware accelerator meticulously designed to enhance Convolutional Neural Network (CNN) computations and streamline inference processes. These accelerators play a vital role in expediting the execution of deep learning models, specifically in applications like image recognition, natural language processing, and computer vision. The overarching objective is to significantly reduce the computational load and energy consumption associated with CNN inference, rendering them well-suited for deployment in resource-constrained environments like edge devices and embedded systems.

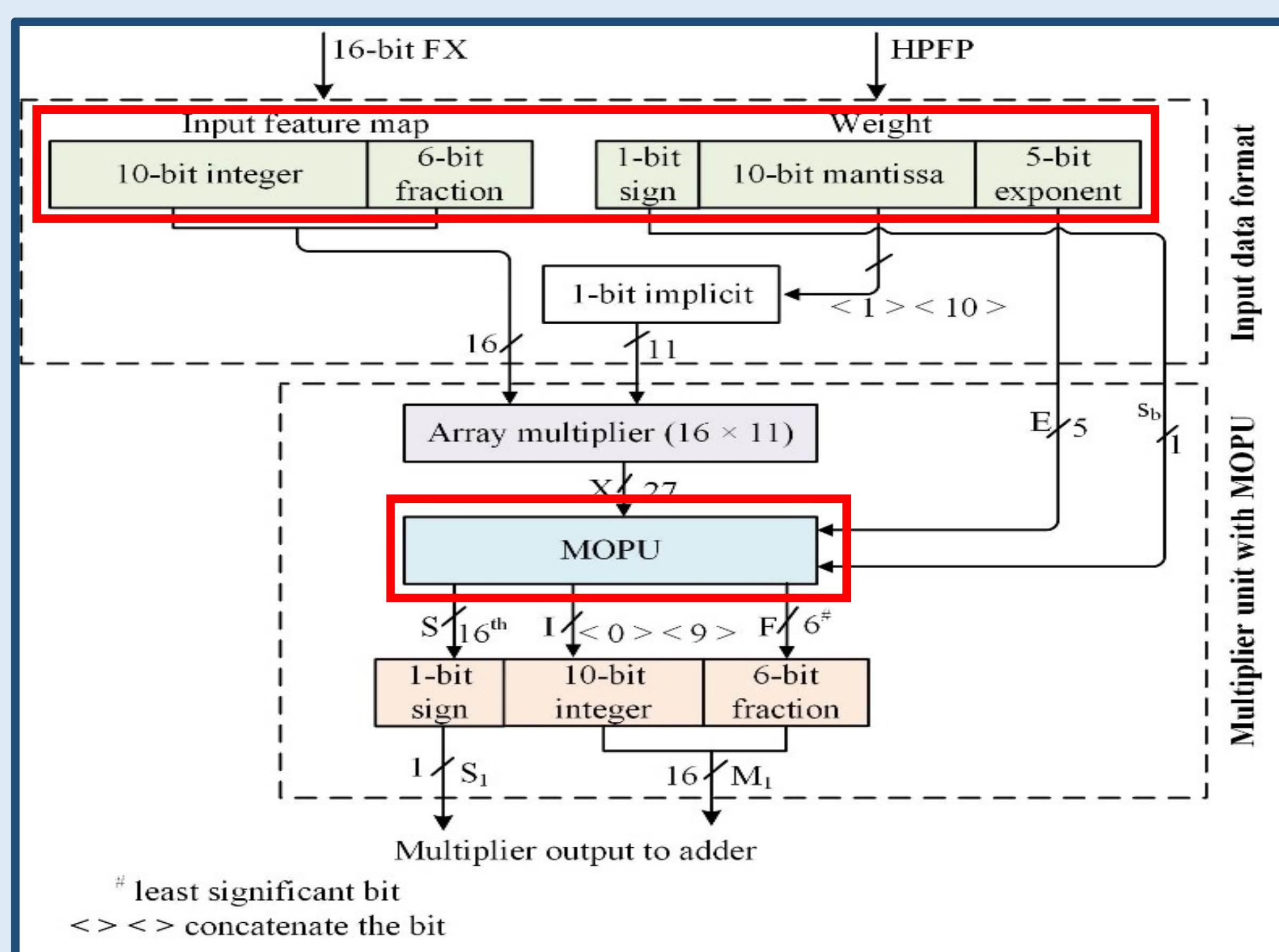
The research improvements in the CNN inference engine have been categorized into:

- ❖ Suitable data format representation
- ❖ Designing the optimized arithmetic units (multiplier and adder) and processor array
- ❖ Data flow scheduling between memory and processor array
- ❖ Skipping the sparse data computation

## Proposed Methodologies

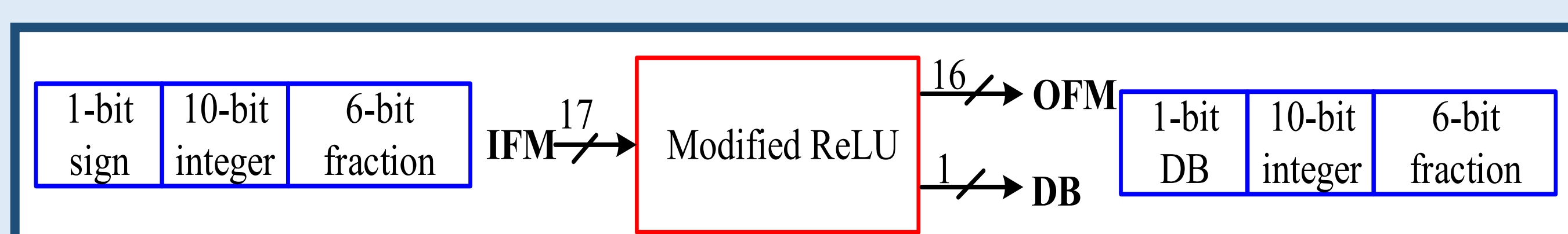
### 1. Design of Convolution Operator for CNN Inference Engines

- ❖ The Inference Engines require an optimum data format for representing weights and input feature maps (IFMs) to reduce the hardware complexity of the convolution operator (CO).
- ❖ 16-bit Fix/Float: IFMs and weights are represented as 16-bit fixed-point (10-bit integer, 6-bit fraction) and Half-precision floating-point (HPFP). It could be more beneficial in terms of both accuracy and cost.
- ❖ A 16-bit Fix/Float 2x1 CO architecture has a Fix/Float array multiplier (16x11) with a multiplication operation processing unit (MOPU) and an Adder/subtractor with a sign processing unit (SPU). A simple algorithm is proposed for correcting the radix position using a multiplication operation processing unit (MOPU).



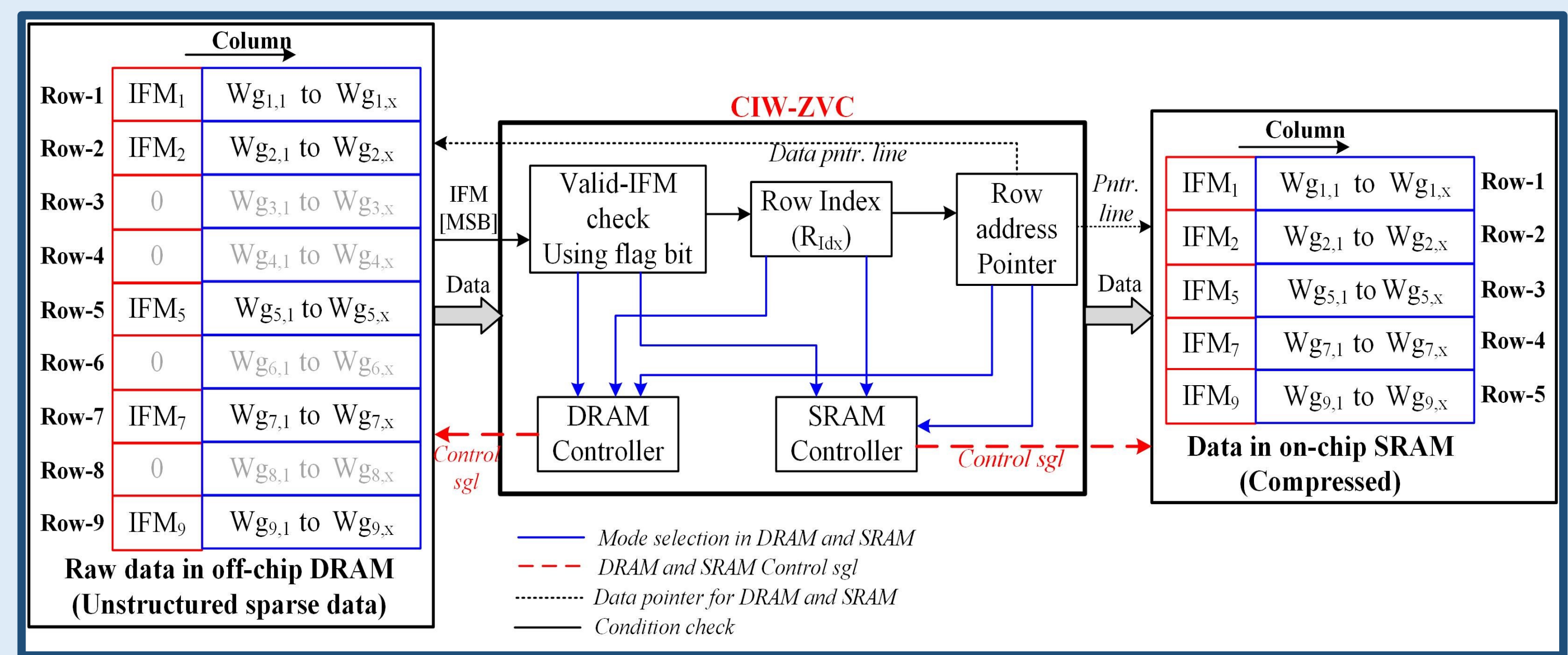
### 2. Design of Optimized and efficient Processor for Convolutional Layer

- ❖ The sparse-based CNN accelerator compresses the insignificant inputs (input feature maps & weights), gating/skipping the inefficient computations in the CNN.
- ❖ A Modified Rectified Linear Unit (M-RELU) is proposed to identify zero sparse in the input feature maps, setting a Detection-Bit (DB) based on the condition if activation maps are less than or equal to 0.
- ❖ These DBs decide the mode of effective operations in 3x3 Compressed Processing Element (CPE) with zero-detect-skip control units. The 20-CPEs convolution array architecture is implemented using 3x3.



### 3. An Energy Efficient Processor Array and Memory Controller for Fully Connected Layer of CNN-based Inference Engines

- ❖ The dedicated hardware accelerator of 256 CO array processes the sparse (IFM and weight) based fully connected layers of the CNN models.
- ❖ A memory controller unit with the proposed Combined IFM and Weight-Zero Valued Compression (CIW-ZVC) accomplishes compress and skip the zero-valued IFMs and their associated weights.
- ❖ This optimizes the data movement rate between the off-chip DRAM and on-chip SRAM by 3.3 times per layer in the case of VGG-16.

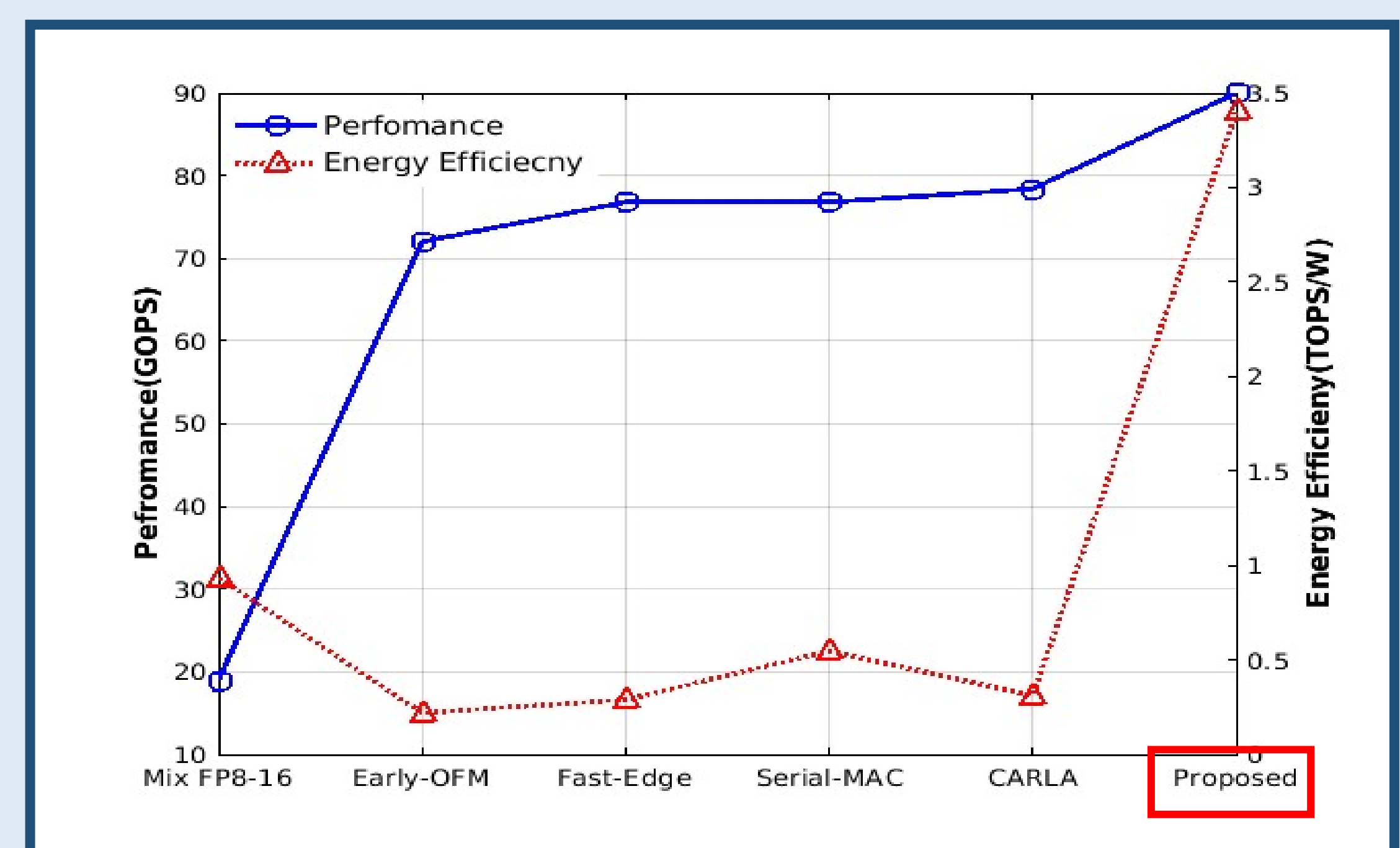


## Results

### Classification Accuracy (IMAGENET Dataset)

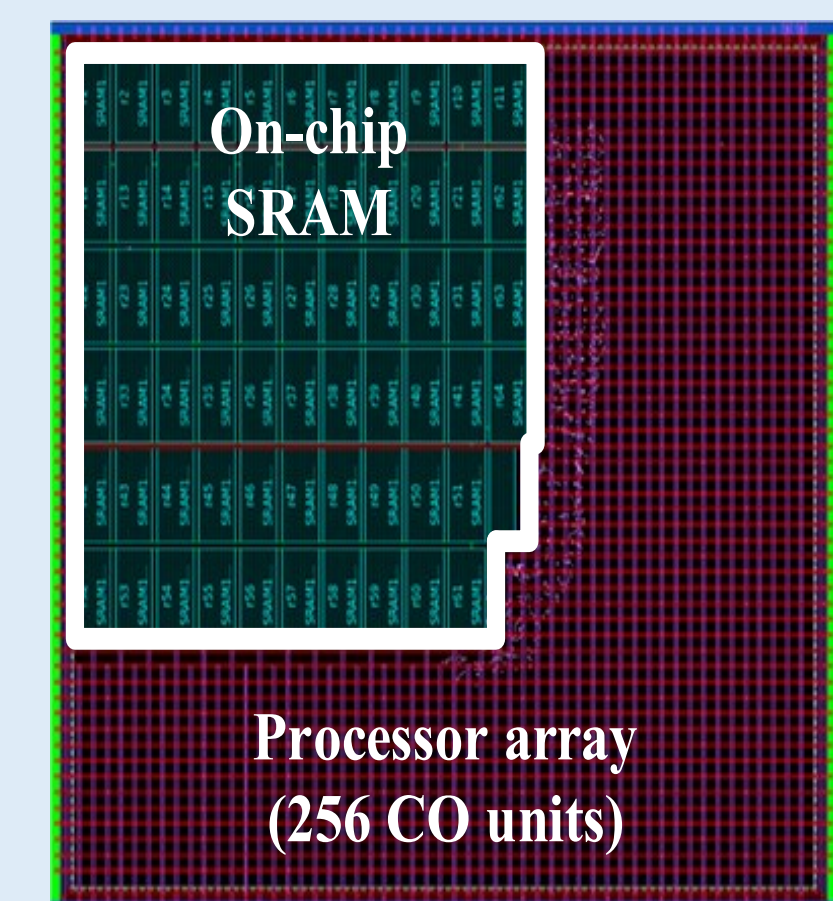
Data format		AlexNet (%)	VGG-16 (%)	VGG-19 (%)
Weight	IFM	100	100	100
SPFP	SPFP	98.95	98.70	99.46
16-bit FX	16-bit FX	79.8	88.32	-
S-Float (4,4)	S-Float (4,5)	79.83	88.22	-
12-bit Dual-mode FX	11-bit SSE	97	97.50	96.5
<b>Proposed HPFP</b>	<b>Proposed 16-bit FX</b>	<b>97.53</b>	<b>97.41</b>	<b>97.33</b>

### Hardware Metrics Vs CNN Accelerators



### Proposed Architecture Characteristics and Core Layout

Technology	FinFET-14nm, 125°C, RVT
Number of CO/MACs	256
Core Area (mm <sup>2</sup> )	0.624
Logic area (2-NAND gates)	157.7K
Frequency (MHz)	200-500
Voltage (V)	0.8
On-chip SRAMs (KB)	137
Power (mW)	7.2-16.8
Performance (GOPS)	102.4-256



## Inferences

- ❖ The proposed 2x1 CO implementation with different data formats reveals that the proposed data format is accurate enough and consumes 22% less area and consumed 17.98% less power than HPFP.
- ❖ A 20 - Compressed Processing Element (CPE) array using an improved zero-detect-skip controller attains the performance of 90 GOPS and an energy efficiency of 3.42 TOPS/W. The CPE with improved control strategy enhanced the performance by a factor of 2.45 while consuming 8.8 times less energy on average than the state-of-the-art CNN accelerators.
- ❖ The absence of the complex on-chip controller, weight-based zero-gated and 16-bit Fix/Float COs reduces overall power consumption. The proposed 256-CO with CIW-ZVC on-chip processor array has energy efficiency and area efficiency of 15.2 (TOPS/W) and 410 (GOPS/mm<sup>2</sup>) respectively when the performance is 256 (GOPS). The performance is 6 times more, which leads to energy efficiency 6 times and area efficiency 7 times compared to the existing designs.

## Acknowledgment

Financial Support: Council of Scientific and Industrial Research (CSIR), Government of India, through the Senior Research Fellow award under Grant 09/844(0104)/2020- EMR.I.

## Publications

- Deepika, S. and Arunachalam, V., 2021. Analysis & Design of Convolution Operator for High Speed and High Accuracy Convolutional Neural Network-Based Inference Engines. *IEEE Transactions on Computers*, 71(2), pp.390-396.
- Deepika, S. and Arunachalam, V., 2023. Design of high performance and energy-efficient convolution array for convolution neural network-based image inference engine. *Engineering Applications of Artificial Intelligence*, 126.
- Deepika, S., Venkatesan, A. and Novo, D., 2023, January. Analysis of Optimum 3-Dimensional Array and Fast Data Movement for Efficient Memory Computation in Convolutional Neural Network Models. In *International Conference on Computer, Communication, and Signal Processing* (pp. 94-108). Cham: Springer Nature Switzerland