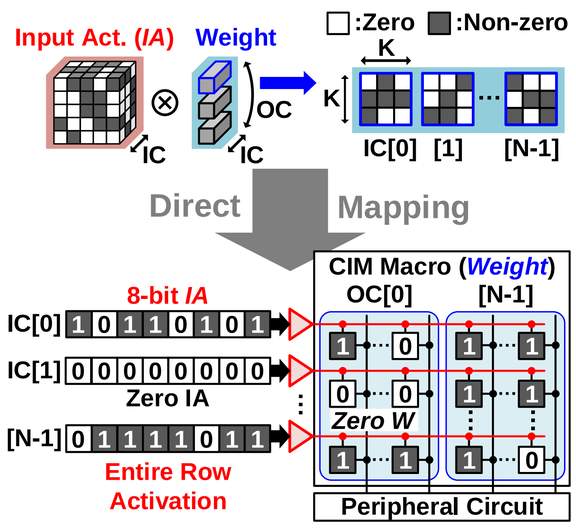
IDEC Chip Design Contest

A Sparsity-Aware Analog-Digital Hybrid eDRAM CIM by Effective Row Activation

Seungbin Kim[1], Hoichang Jeong[1], and Kyuho Lee[2]
[1]Department of Electrical Engineering, UNIST, Republic of Korea
[2]Department of Electrical and Electronic Engineering, Yonsei University, Republic of Korea

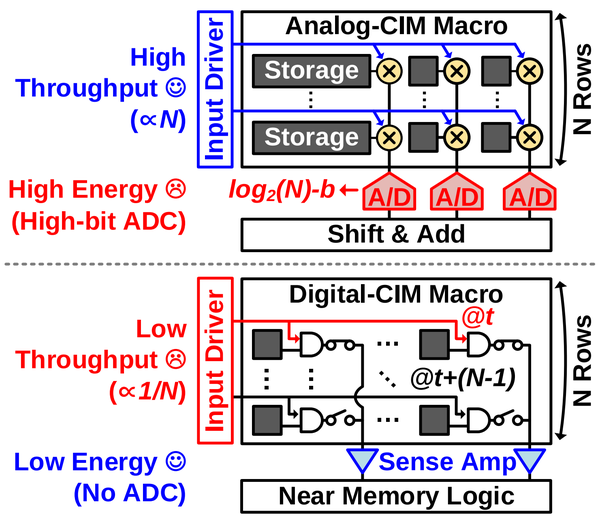
Design Challenges of Previous CIMs

< Waste of Energy Consumption by Zero Computation>



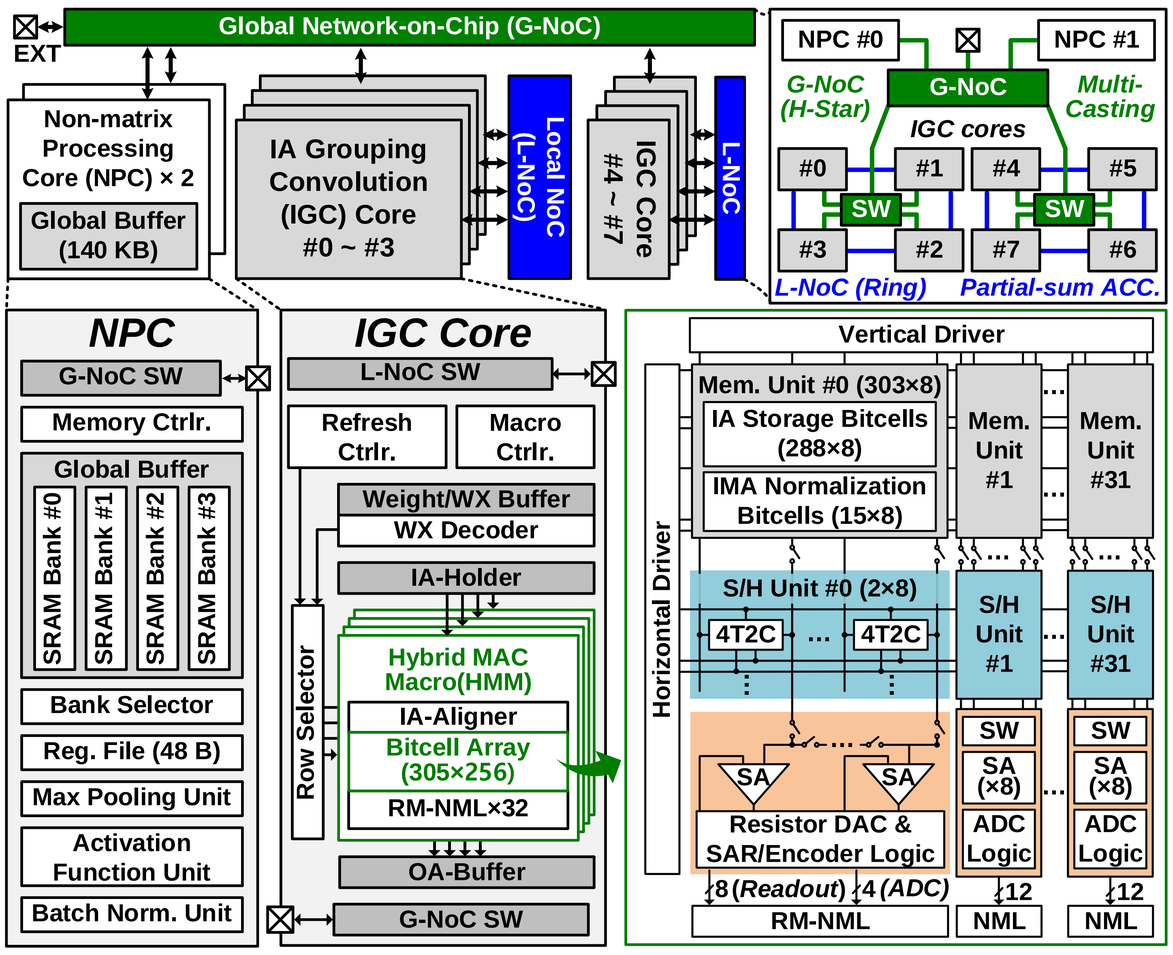
- Input Activations (IA) and Weights in DNN exhibit ***Tremendous Sparsity*** with random pattern.
- ***Direct Mapping*** of convolution into the CIM Macro and ***Entire Row Activation*** for MAC operations disables the handling of random sparsity.
- A significant portion of energy is wasted in CIM macro for calculating ***83% of zero computation***.

< Distinct Trade-off of Analog CIM and Digital CIM >



- ***A-CIMs*** ensure high throughput by activating entire rows simultaneously.☺
- ***A-CIMs*** require high-resolution ADCs to maintain computational accuracy, leading to enormous power usage. ☹
- ***D-CIMs*** consume less power without utilizing ADC. ☺
- ***D-CIMs*** suffer from low throughput due to sequential row-by-row operations. ☹

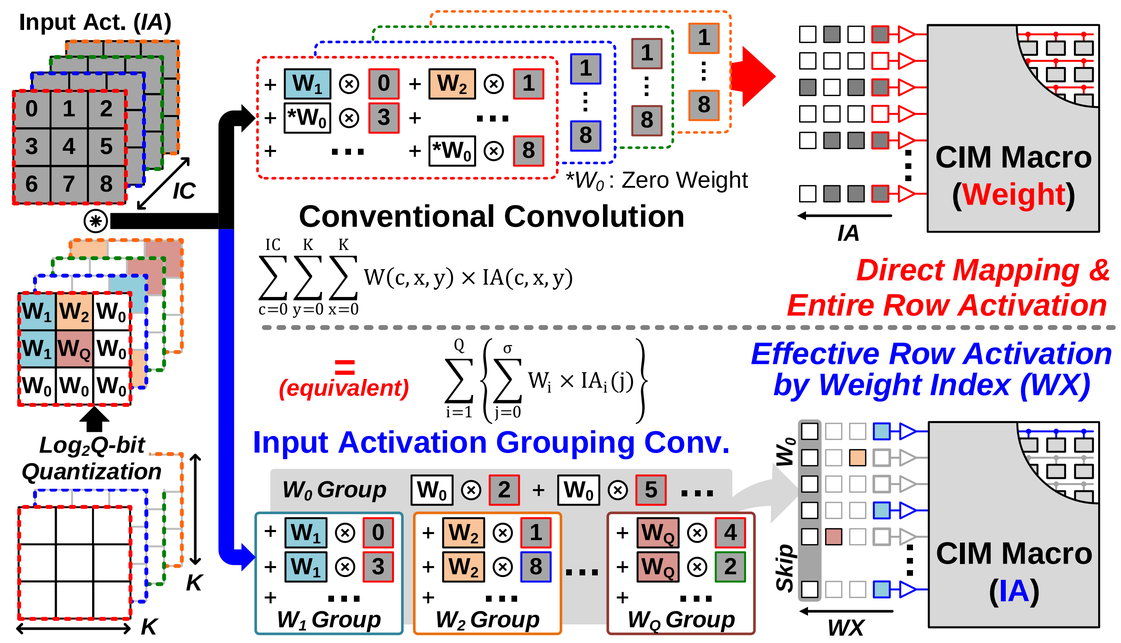
Overall Architecture



- ***Hierarchical network-on-chips*** consists of global (***G-NoC***) and local (***L-NoC***).
  G-NoC: Multicasting data from global buffers to 8 IGC cores w/ H-Star architecture
  L-NoC: Facilitating partialsum accumulation from 4 IGC cores w/ Ring architecture
- ***Hybrid MAC macro*** is consists of 305×256 eDRAM bitcell array.
  288 rows for storing IAs and perform either in-memory accumulation (IMA) in analog domain or readout in digital domain
  15 rows for maintaining computation linearity during IMA
  2 rows for sample and hold (S/H) the IMA result and connected to ADC

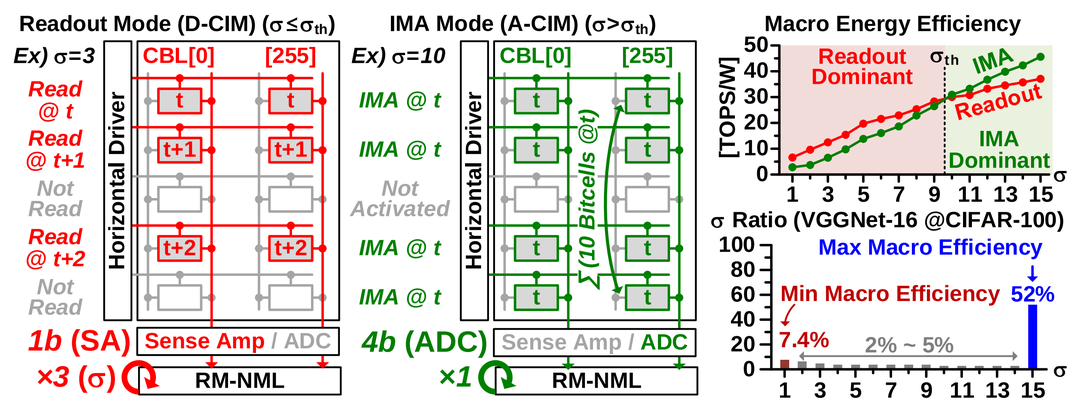
Effective Row Activation w/ Hybrid mode

< Input Activation Grouping Convolution>



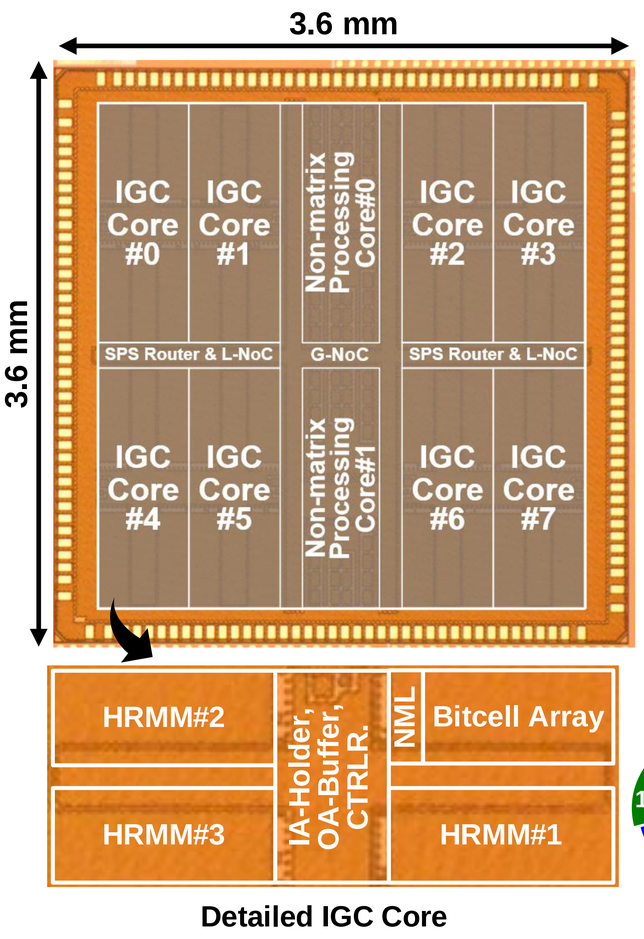
- It ***Eliminates Zero Computations*** by never activating rows that correspond to the W_0 group, ***Retaining Algebraic Equivalence*** to conventional convolution.
- IGC improves ***Non-zero Computation Ratio*** in CIM macro by ***1.93× and 4.59× on VGGNet-16 and ResNet-18.***

< Analog-Digital Hybrid MAC Macro>

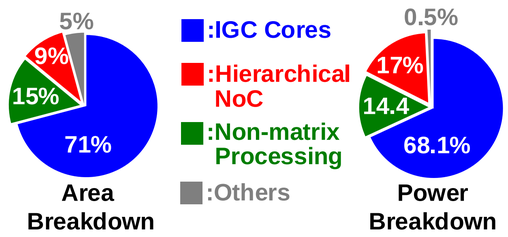


- ***Depending on the number of effective row*** (σ) for Wi group, HMM decided to operate mode of readout or IMA mode to maximize the energy efficiency.
- The proposed processor achieves benchmark energy efficiency of ***24.8 TOPS/W for ResNet-18 and 30.7 TOPS/W for VGGNet-16.***

Chip Photograph & Performance Summary



Technology		28 nm CMOS
Area		12.96 mm^2
Operating Voltage		1.0 V
Frequency		250 MHz
Bit Precision	IA	8-bit
	W	2-bit ~ 8-bit
Cell Structure		4T2C eDRAM
CIM Capacity		305 KB
Peak Macro & System Energy Efficiency		
Macro Energy-η		47.5 TOPS/W
System Energy-η		25.7 TOPS/W
Benchmark Performance		
ResNet-18		
CIFAR-100		ImageNet
24.8 TOPS/W (Accuracy: 74.5%)		21.3 TOPS/W (Accuracy: 68.5%)
VGGNet-16		
CIFAR-100		ImageNet
30.7 TOPS/W (Accuracy: 71.2%)		21.8 TOPS/W (Accuracy: 69.7%)



- 본 연구는 IDEC에서 MPW를 지원받아 수행하였습니다.
- 본 연구는 IDEC에서 EDA Tool를 지원받아 수행하였습니다.

반도체설계교육센터 IC DESIGN EDUCATION CENTER