OmniCIM: A Sparsity-Aware Computing-in-Memory based Processor for Accelerating Arbitrary Quantized Neural Networks

Jianxun Yang[1], Yuyao Kong[2], Yiqi Wang[1], Zhao Zhang[1], Jing Zhou[1],

Zhuangzhi Liu[1], Yonggang Liu[1], Chenfu Guo[1], Te Hu[1], Congcong Li[1],

Leibo Liu[1], Jun Yang[2], Shaojun Wei[1], Shouyi Yin[1]

[1]Tsinghua University, Beijing, China

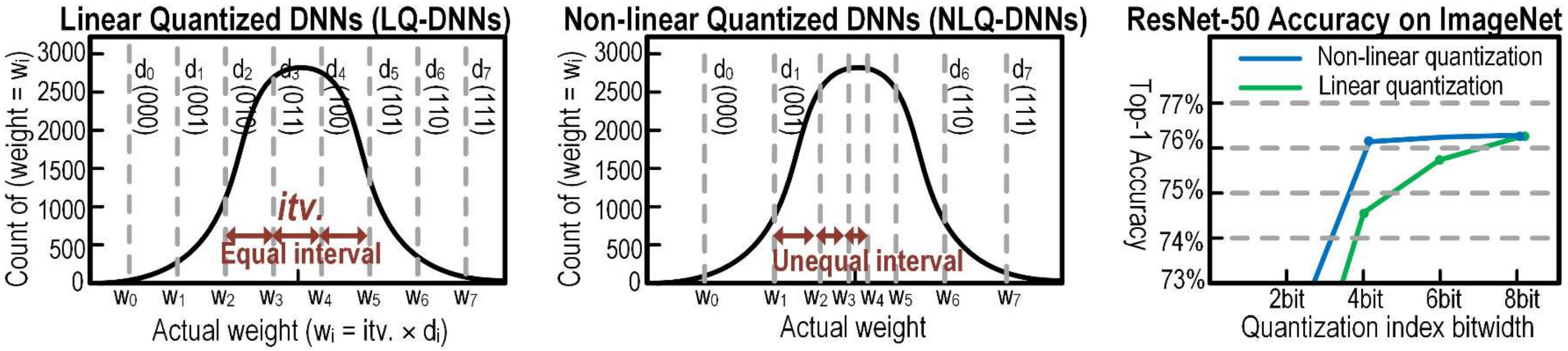
[2]Southeast University, Nanjing, China

Outline

- **Background and Challenges**
- **Proposed Processor**
  - Time-Domain Computing-in-Memory (CIM) to reduce computation power
  - Bit-sparsity-aware pulse computation to reduce computation amount
  - Predictor to early-stop computation
- **Measurement Results**

Background

Non-linear quantized DNNs have higher accuracy than linear quantized DNNs



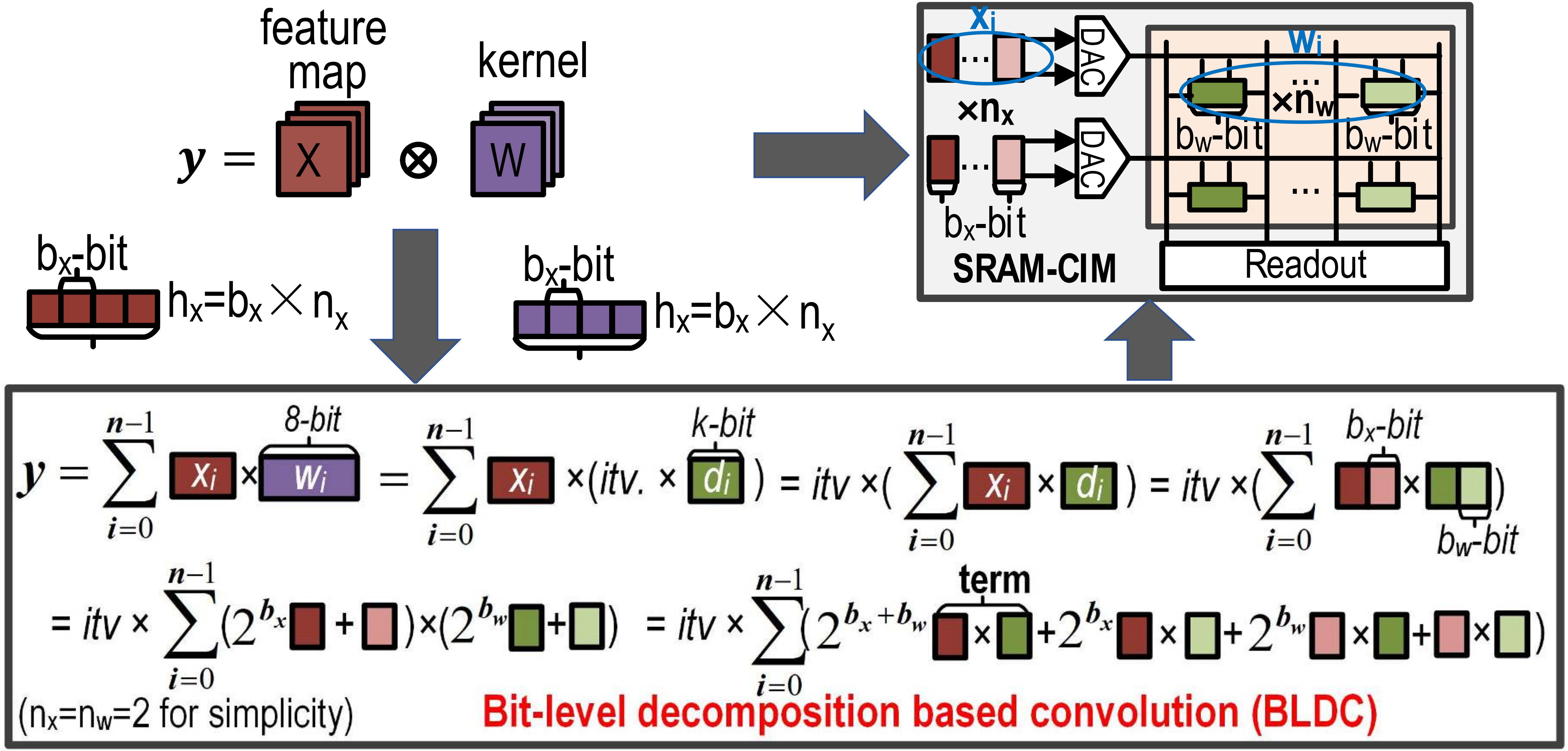
k-bit quantization:

w_i: High-bitwidth quantized actual weight (8b)

di: Low-bitwidth quantization index (1~8b)

Background

Conventional convolution computation for LQ-DNNs on CIM-based accelerators



Challenge 1: High Computation Complexity

Use BLDC to accelerate NLQ-DNNs

High-precision actual weights have to be used for computation of NLQ-DNNs

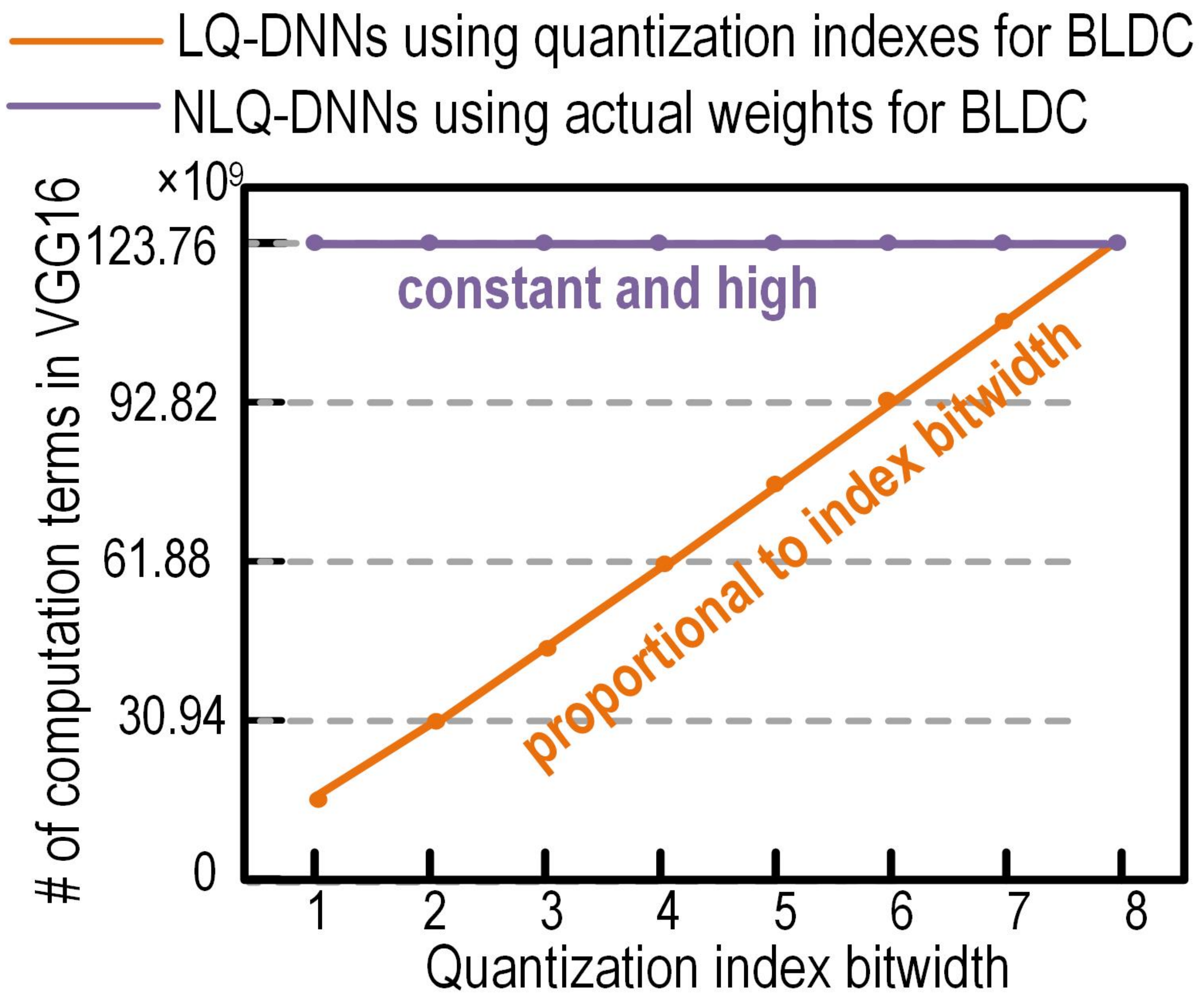
3 Challenges

Quantization Type	Quantization interval	Convolution	BDDC	Decomposed computation term
Linear	Equal	$itv.\times(\sum_{i=0}^{n-1} x_i \times \overbrace{d_i}^{k\text{-bit}})$	$itv.\times(\sum_{i=0}^{n-1} \times \overbrace{\ }^{b_w\text{-bit}})$	× , × , × , × More terms
Non-linear	Unequal	$\sum_{i=0}^{n-1} x_i \times \overbrace{W_i}^{8\text{-bit}}$	$\sum_{i=0}^{n-1} (\overbrace{\ }^{b_x\text{-bit}} \times \overbrace{\ }^{b_w\text{-bit}})$	×, ×, ×, ×, ×, ×, ×, ×

Higher computation complexity

Challenge 2: Poor Adaptability to DNNs

The same computation complexity for different-precision NLQ-DNNs

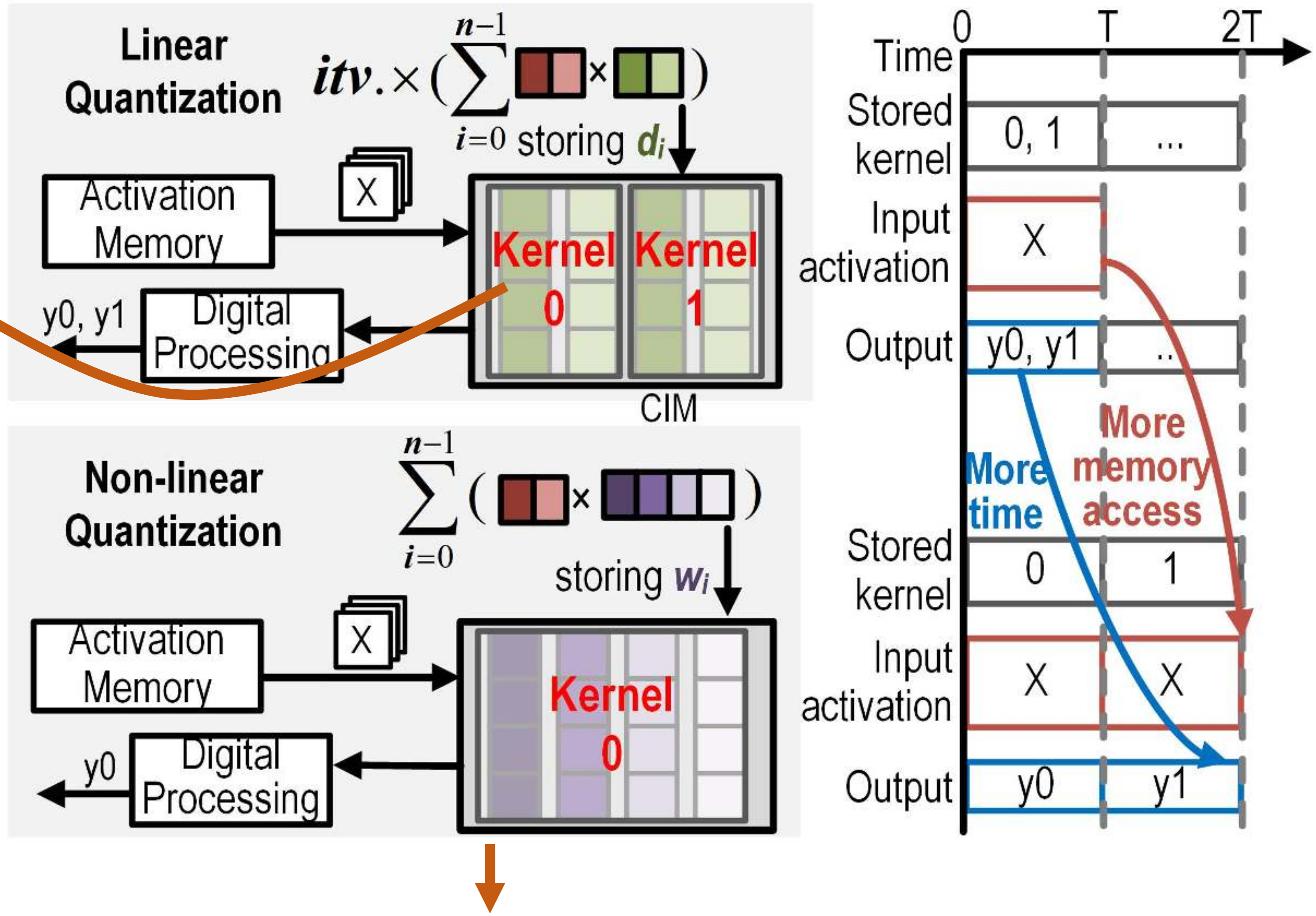


Assume bit-segment size of activation (b_x) and weight (b_w) is 8 and 1

Challenge 3: High Memory Access and Latency

High kernel parallelism

Actual Weights have to be used in NLQ-DNNs

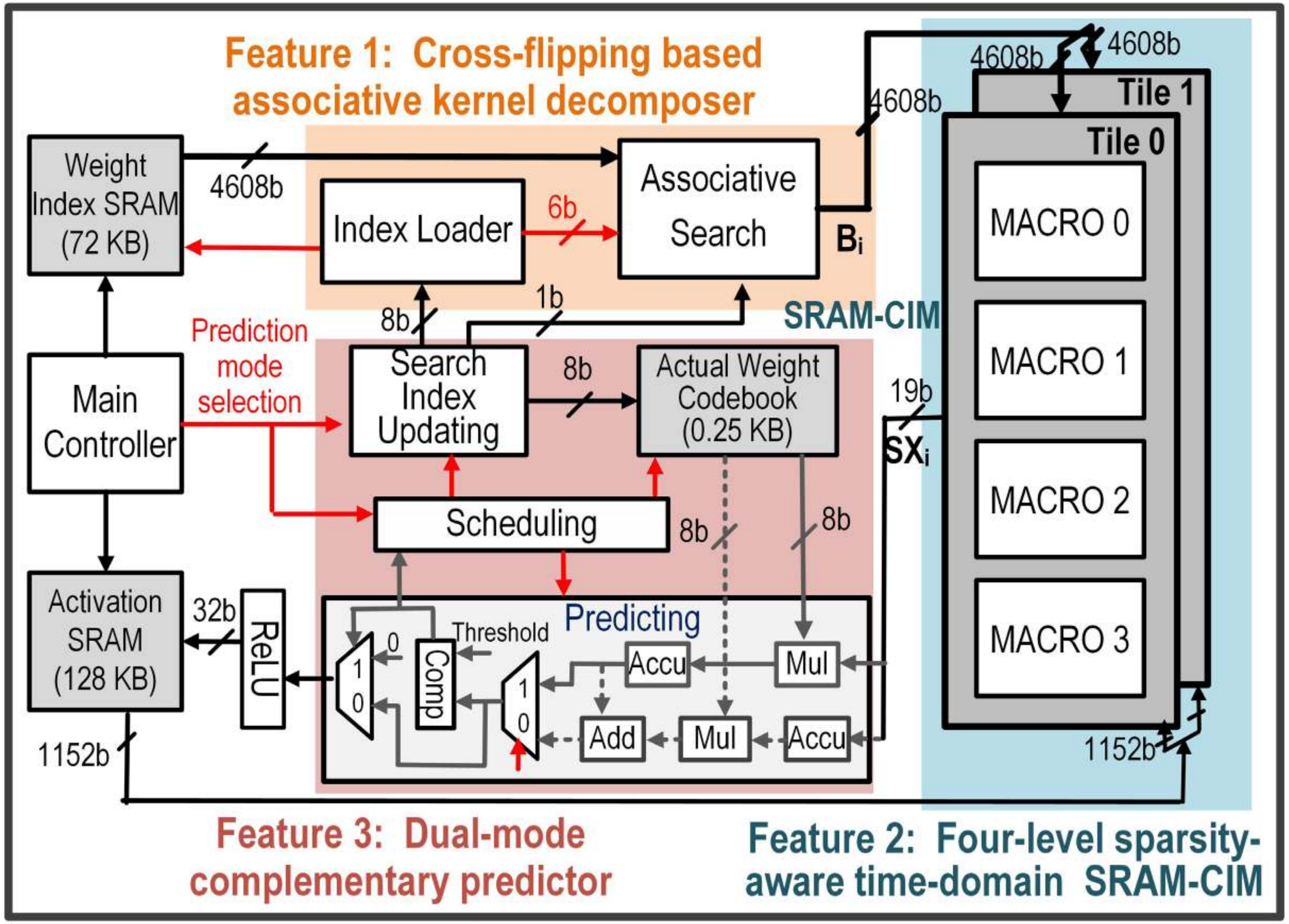


Low kernel parallelism

Proposed Processor for LQ- and NLQ-DNNs

- Predictable Convolution Computation

- Time-Domain CIM based Processor



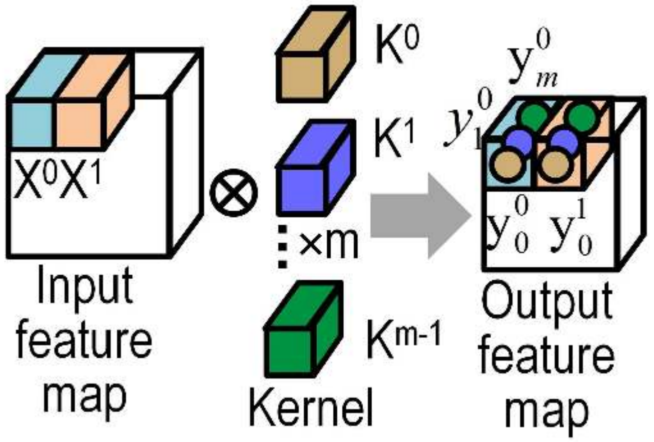
Dataflow

$$y = W\cdot X = (-2)\times 2 + 2\times 1 + 1\times 2 + (-1)\times 2 + (-2)\times 4 = \underbrace{2\times 1 + 1\times 2 + (-2)\times(4+2)}_{=-8<0} \; \underbrace{+\,(-1)\times 2}_{=-2<0} = -10 \longrightarrow \mathrm{ReLU} = 0$$

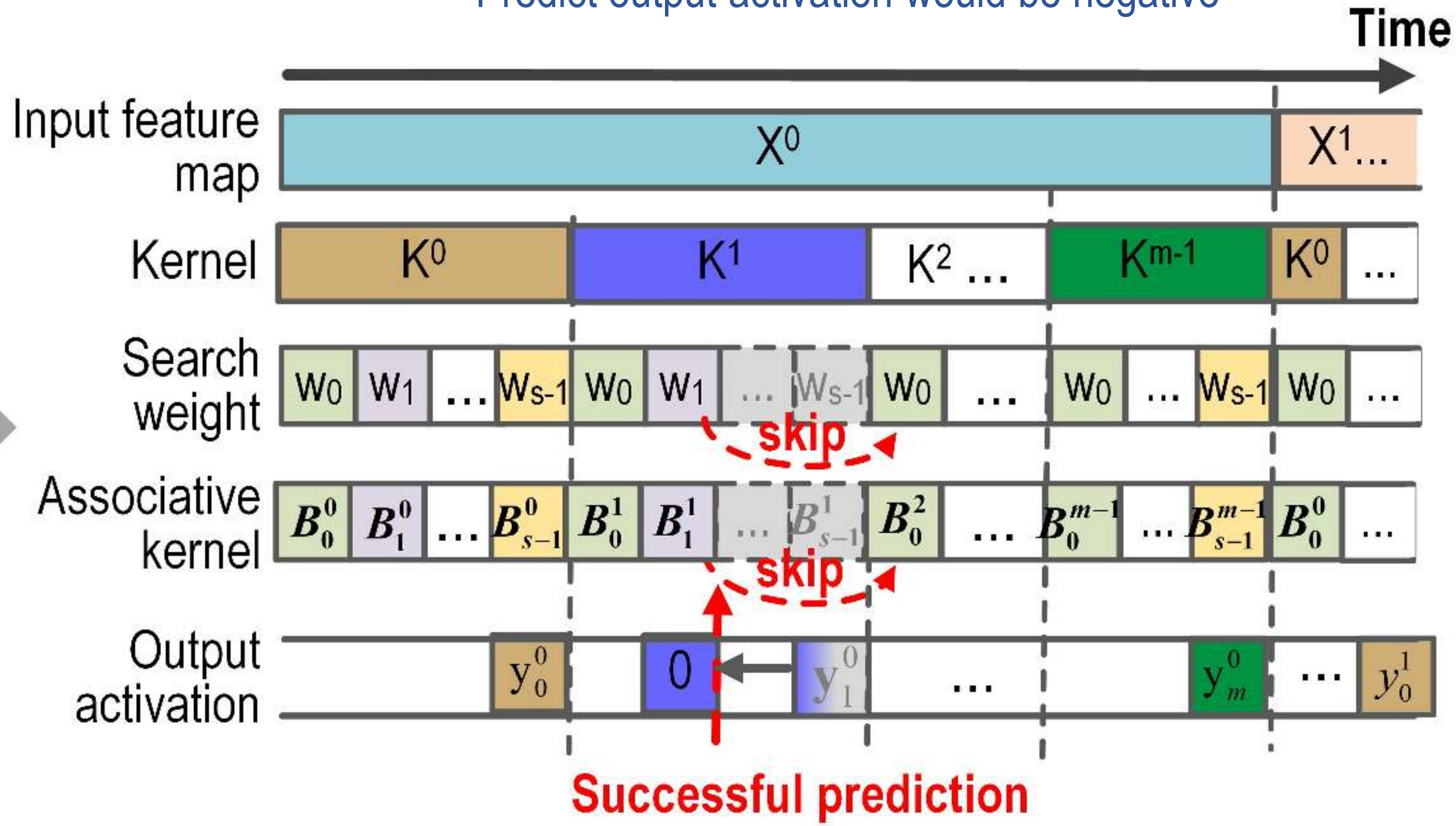
Early-stop

Predict output activation would be negative

Kernel Size	# of weights in one kernel	Macros for one kernel	Computed kernels concurrently
Large	2305~4608	4	1
Medium	1153~2304	2	2
Small	1~1152	1	4



Flexible kernel mapping to achieve 100% CIM utilization

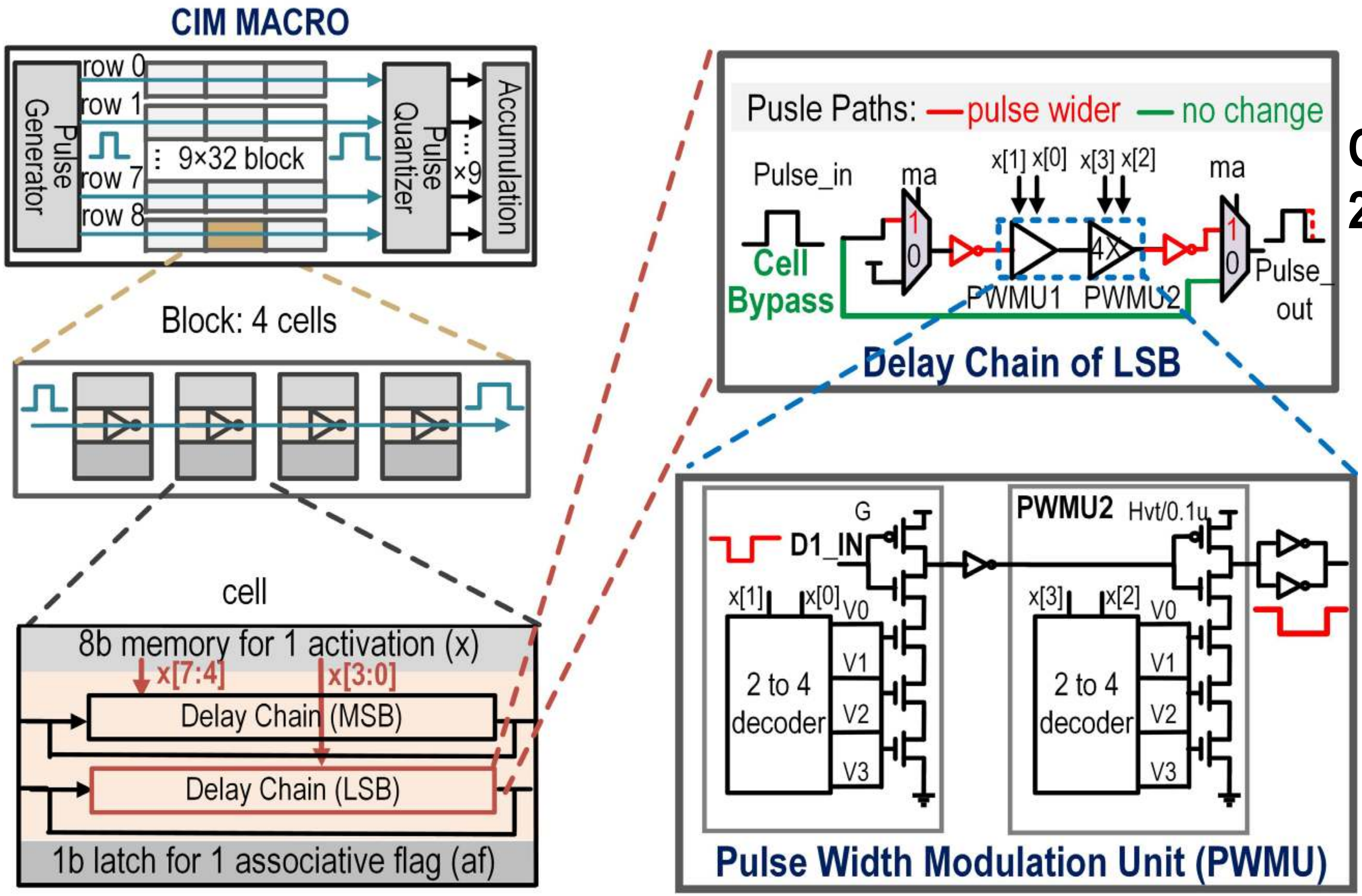


Input Stationary to reuse activations

Time-Domain CIM

9×128 cells

One PWMU for 2b activation

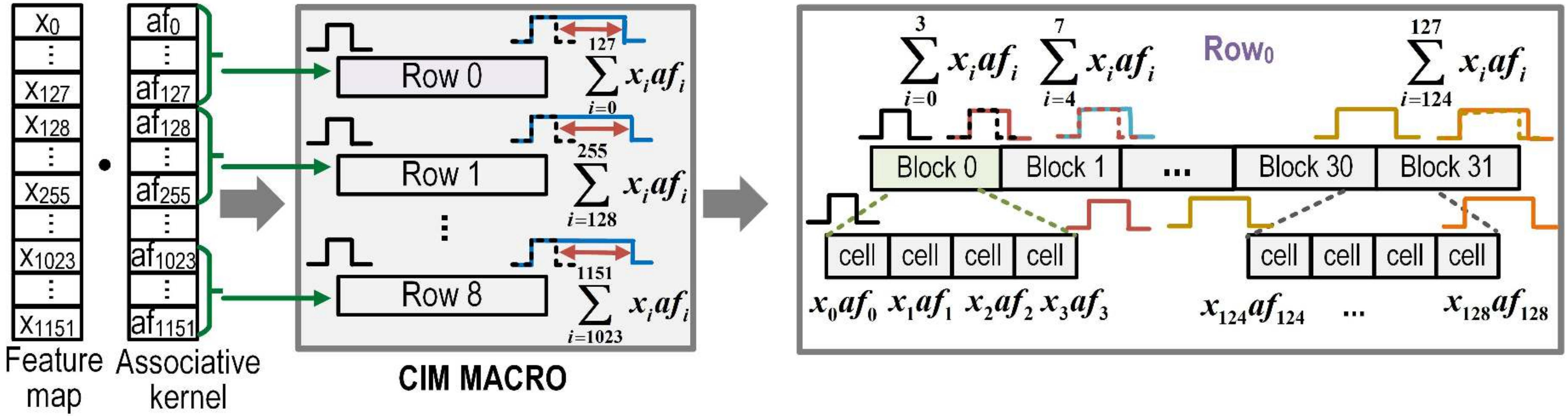


Two delay chain for 8b activation

Four-level voltage modulates pulse width

Time-Domain CIM

1152 MACs are computed in one CIM MACRO

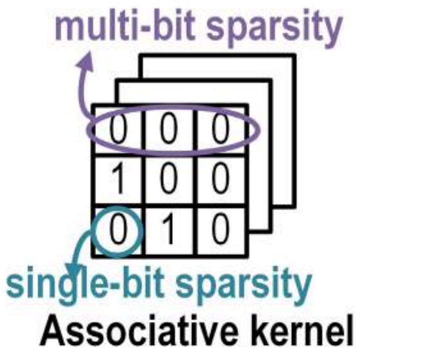


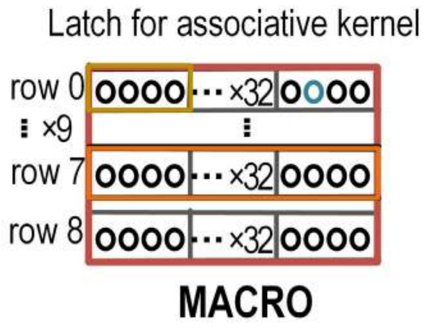
Computation mapping in CIM MACRO

Bit-Sparsity-Aware Pulse Computation

Single-bit and Multi-bit Sparsity

Four-level bypass to reduce power and latency





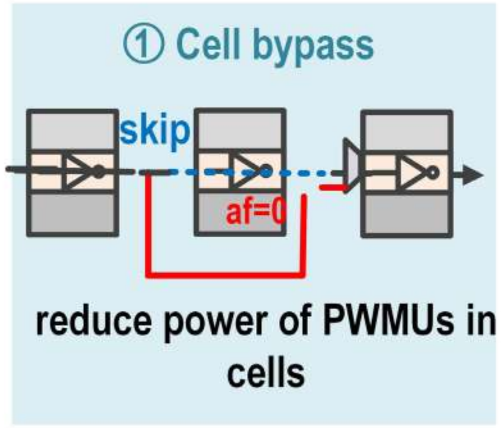
single-bit sparsity

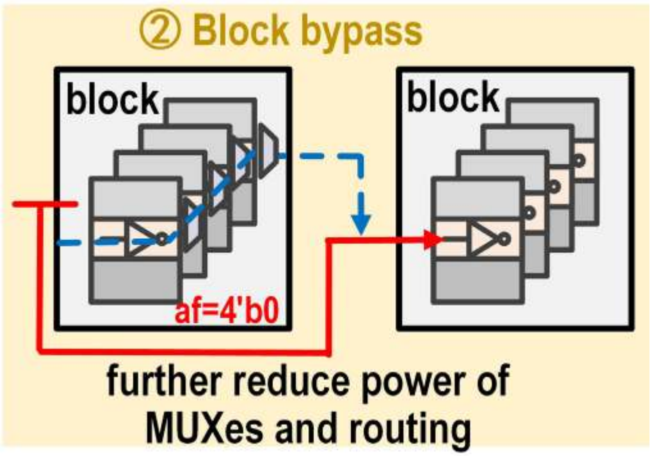
1b'0: Cell Bypass

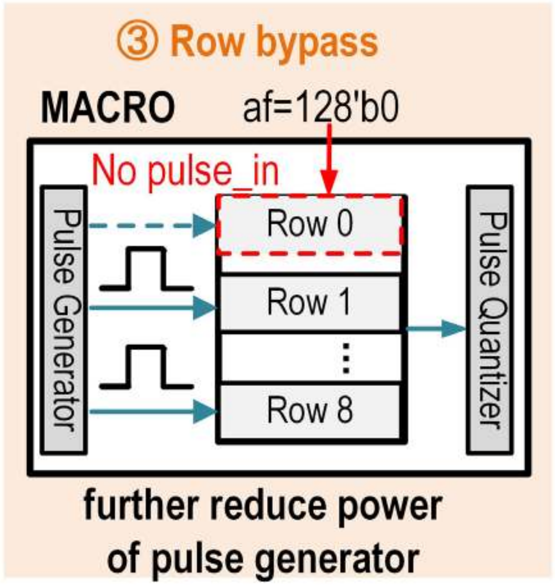
4b'0: Block Bypass

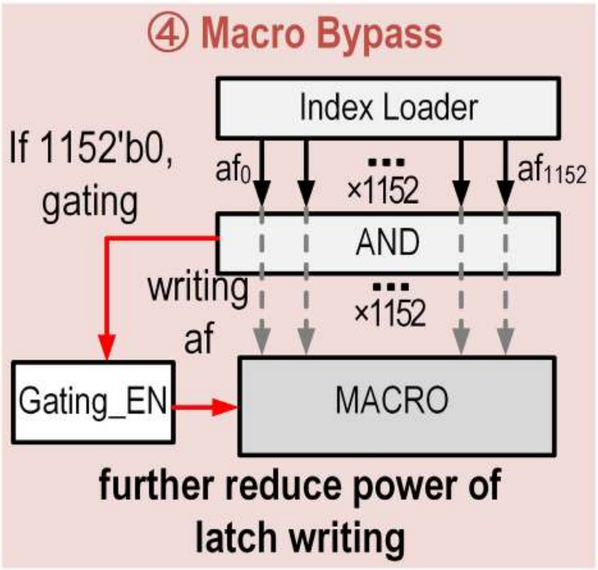
128b'0: Row Bypass

1152b'0: Macro Bypass



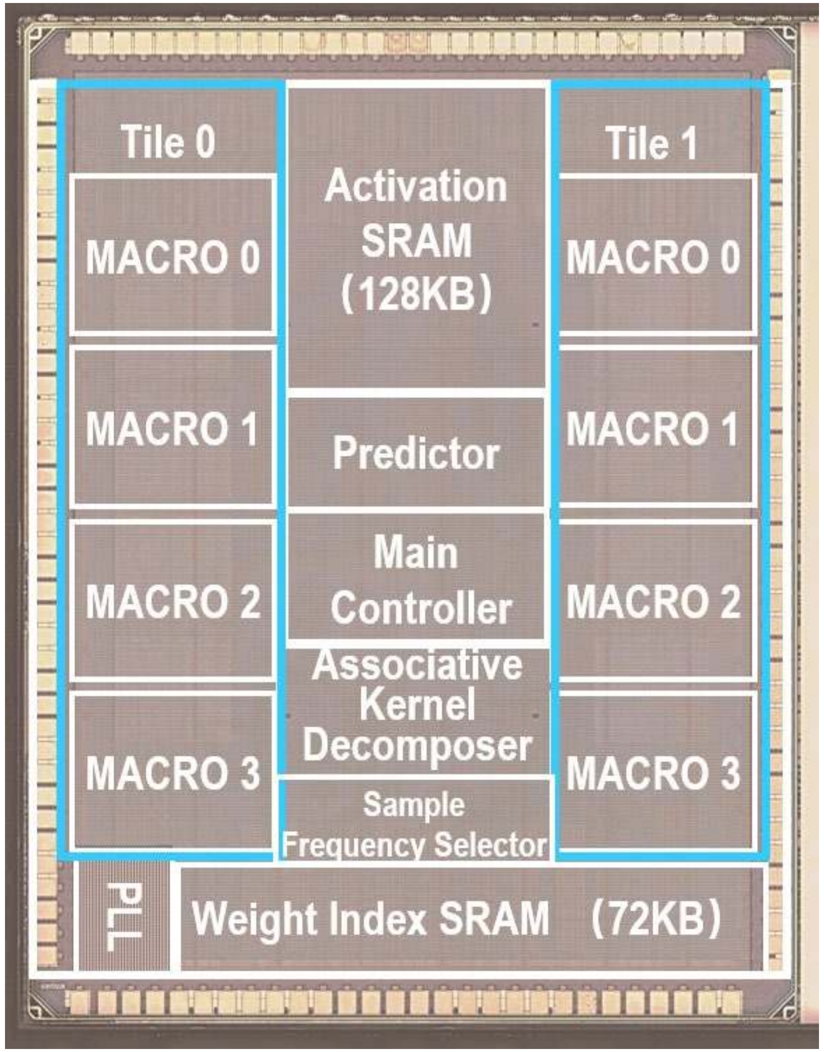






Measurement Results

Support 1-8b LQ- and NLQ-DNNs



2.4-152.7 TOPS/W peak energy efficiency

CHIP SUMMARY	
Process (nm)	28
Supply Voltage (V)	0.55-1.05
Frequency (MHz)	50-170
Die Size (mm)	2.35*3.07
Weight Precision	1-8 b
Quantization Type	Linear & Non-linear
Power (mW)	4.3-33.8
Peak Energy Efficiency (TOPS/W)	2.4-152.7

VGG16 Accuracy on ImageNet (8b activation)

Quantization Type	Linear				Non-Linear			
Quantization Bitwidth	1b	2b	4b	8b	1b	2b	4b	8b
Top-1 Accuracy (%)	65.1	67.5	70.1	71.9	67.3	69.8	71.5	72.1

Measurement Results

Improvement of energy efficiency: 2.59x for LQ-DNNs, 2.15x for NLQ-DNNs

Reference		ISSCC 2018 [4]	ISSCC 2018 [5]	ASSCC 2016 [5]	ISSCC 2019 [7]	ISSCC 2019 [8]	This work	
Tech. [nm]		65	65	65	28	55	28	
Circuit		Voltage	CIM	Time	Digital	Voltage	Digital + CIM	
Die Area [mm^2]		0.067	1.44	3.61	2.7	0.037	7.21	
Supply Voltage [V]		0.9-1.2	0.65-1	-	0.6-1.1	-	0.55-1.05	
Max Frequency [MHz]		6.7	1000	-	475	-	170	
Weight Precision		1b	1b	1b	Arbitrary	1-4b	1-8b	
Activation Precision		7b	8b	1b	Arbitrary	1-5b	8b	
Peak Throughput [GOPS]		10.7	-	-	32.7	-	246.3	
Benchmark		LeNet-5	SVM	LeNet-5	Addition	MNIST	VGG16[4]	AlexNet
Energy Efficiency [TOPS/W]	Linear[1]	28.1 @ (1,7)	3.13 @ (1,8)	48.20 @ (1,1)	5.27 @ (8,8)	72.1 @ (2,1) 18.37 @ (5,2)	63.64 @ (1,8) 27.17 @ (2,8) 9.65 @ (4,8)	56.71 @ (1,8) 32.34 @ (2,8) 8.96 @ (4,8)
	Linear[2] Normalized	24.6 @ (1,8)	3.13@ (1,8)	6.03@ (1,8)	5.27 @ (8,8)	18.03 @ (1,8) 4.51 @ (4,8)		
	Non-linear[3]	3.07 @ (*,8)	0.39 @ (*,8)	0.75@ (*,8)	5.27 @ (*,8)	2.87 @ (*,8)	65.89 @ (1,8) 11.32 @ (4,8)	58.37 @ (1,8) 9.98 @ (4,8)

[4] A. Biswas, et al., ISSCC, 2018. [5] S. K. Gonugondla, et al., ISSCC, 2018.
[6] A. Sayal, et al., ISSCC, 2019. [7] J. Wang, et al., ISSCC, 2019. [8] X. Si, et al., ISSCC, 2019.

Thank you!