ELearn: Edge Learning Processor with Bidirectional Speculation and Sparsity & Mixed-Precision aware Dataflow Parallelism Reconfiguration

Fengbin Tu[1], Weiwei Wu[1], Yang Wang[1], Hongjiang Chen[1], Feng Xiong[1], Man Shi[1], Ning Li[1], Jinyi Deng[1], Tianbao Chen[2], Leibo Liu[1], Shaojun Wei[1], Shouyi Yin[1]

[1]Institute of Microelectronics, Tsinghua University, Beijing, China

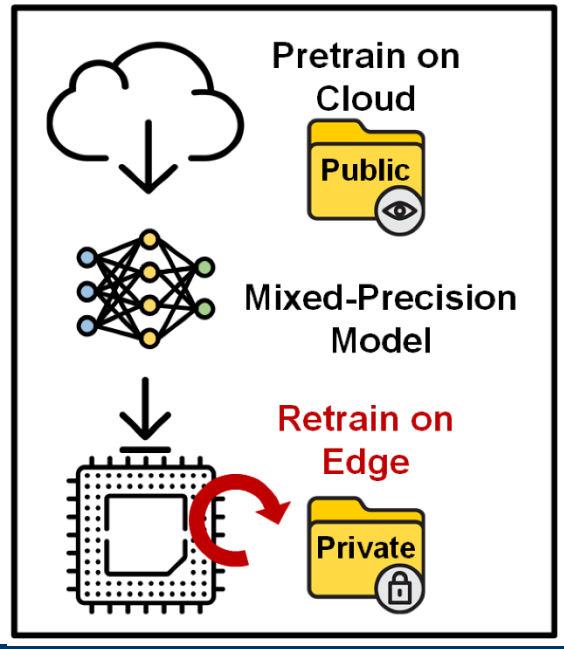
[2]TsingMicro Tech, Beijing, China

Abstract

To achieve higher accuracy on local devices, the necessity of retraining a DNN model on edge increases. This paper proposes a sparsity and mixed-precision aware edge learning processor ELearn. With bidirectional speculation for runtime data sparsity in training's feedforward and backpropagation passes, ELearn reduces 63.3% computation and 85.5% memory access. A parallelism reconfigurable engine and a runtime parallelism optimizer are designed to match the time-varying workload parallelism caused by sparsity and mixed-precision. Our techniques achieves 3.60x higher energy efficiency and 1.44x higher PE utilization. ELearn is fabricated in 28nm CMOS and achieves energy efficiency of 3.8-to-172.8 TOPS/W.

Edge Learning

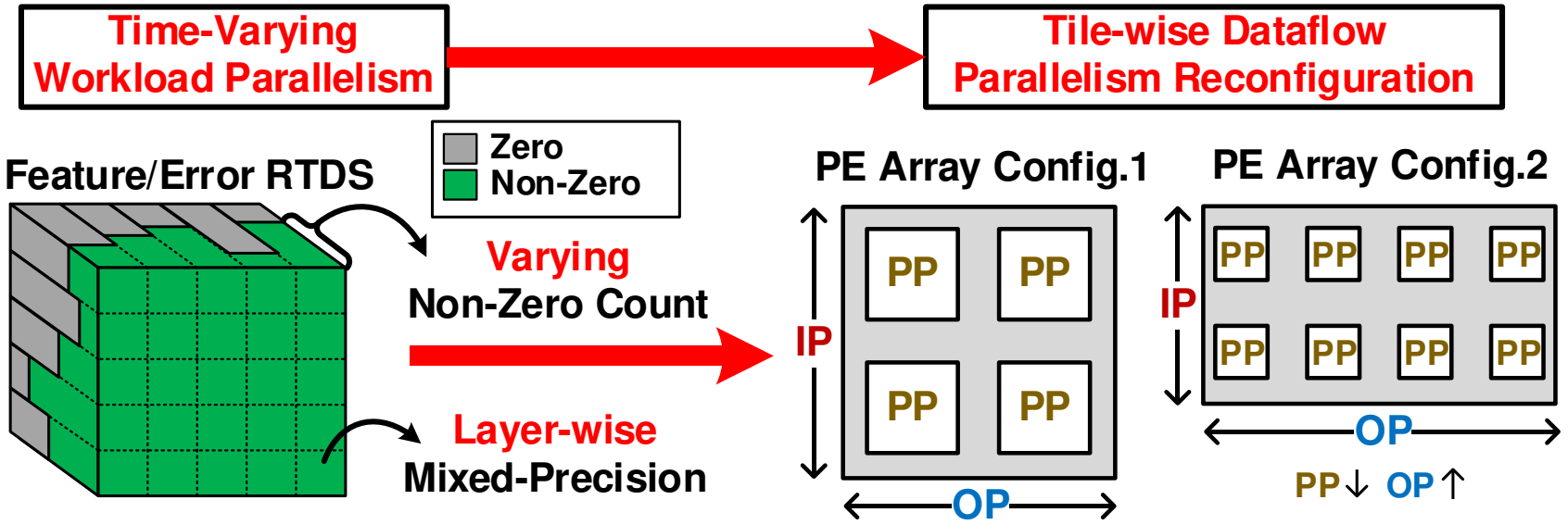
- We usually pretrain a mixed-precision DNN model on cloud and then deploy it on edge devices. This would cause accuracy degradation in complex local environments. However, sending local data to the cloud for retraining incurs high transmission cost and privacy issues.
- Retraining mixed-precision models directly on edge is essential to adapt DNNs to local environments.



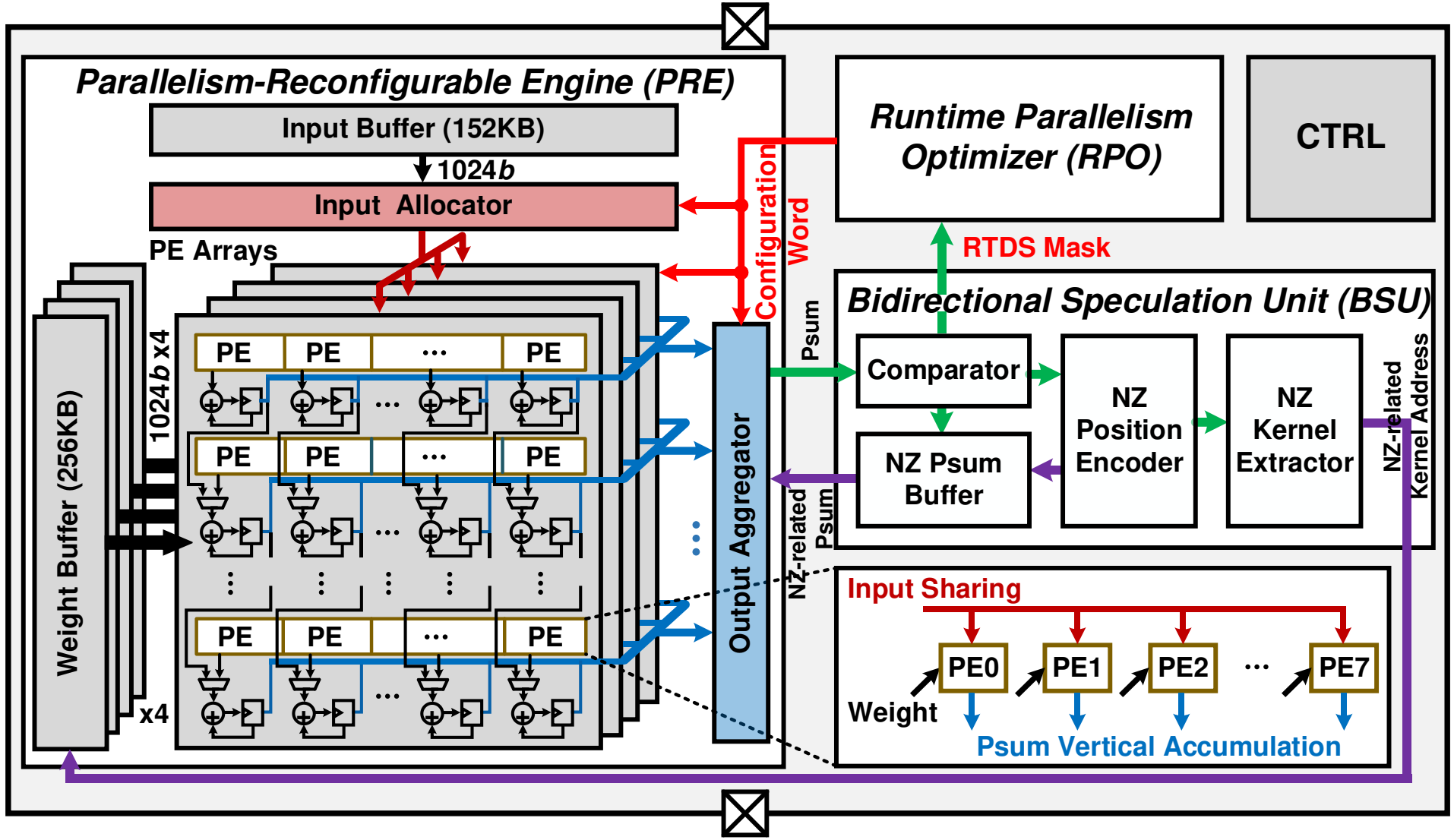
Time-Varying Workload Parallelism

- We exploit training's runtime data sparsity (RTDS) and mixed-precision computation for efficient edge learning.
- RTDS and mixed-precision causes time-varying workload parallelism.
- Dynamic hardware reconfigurability to match the runtime changes.

  (IP: Input Parallelism, OP: Output Parallelism, PP: Precision Parallelism).



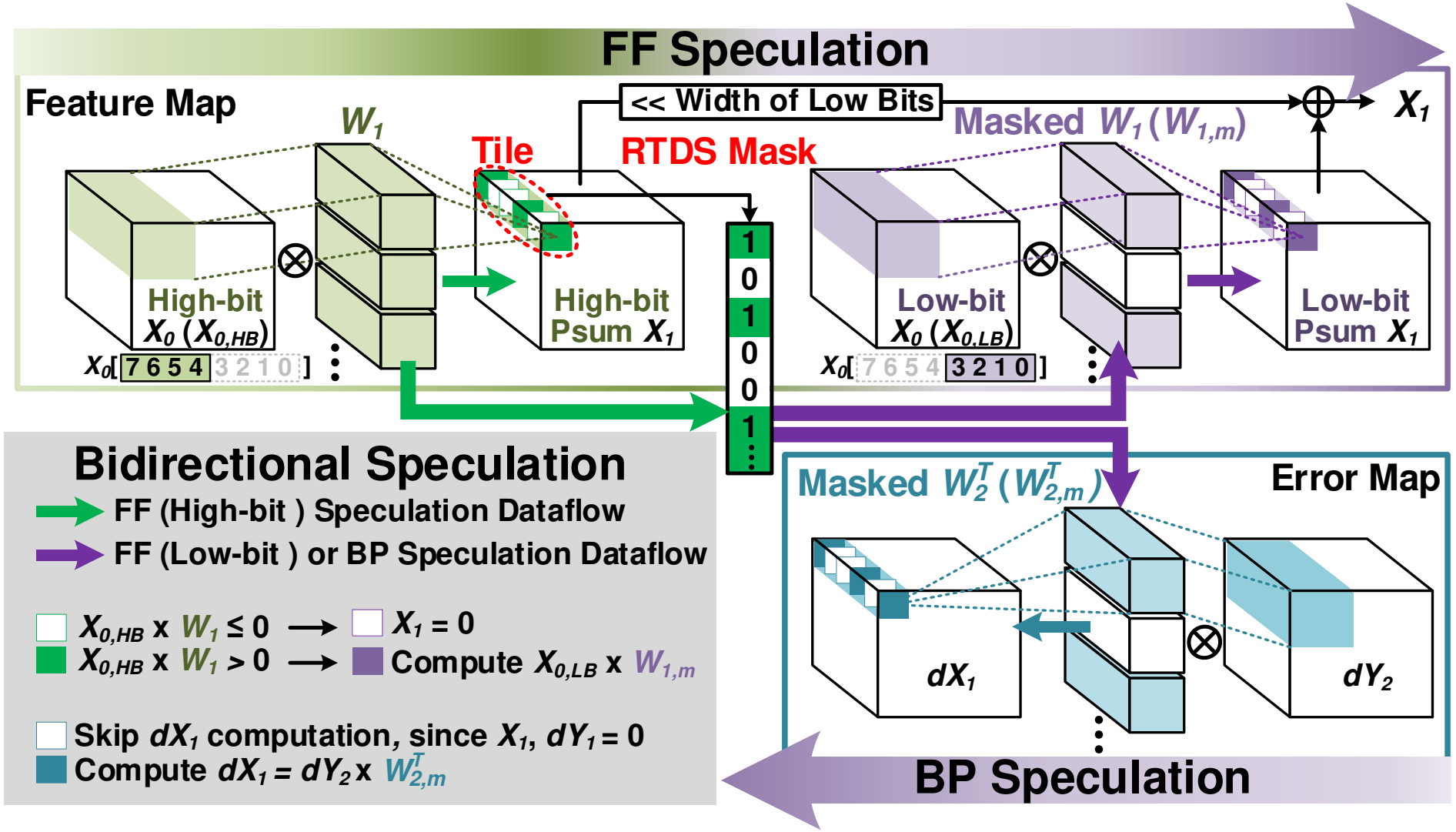
ELearn's Overall Architecture



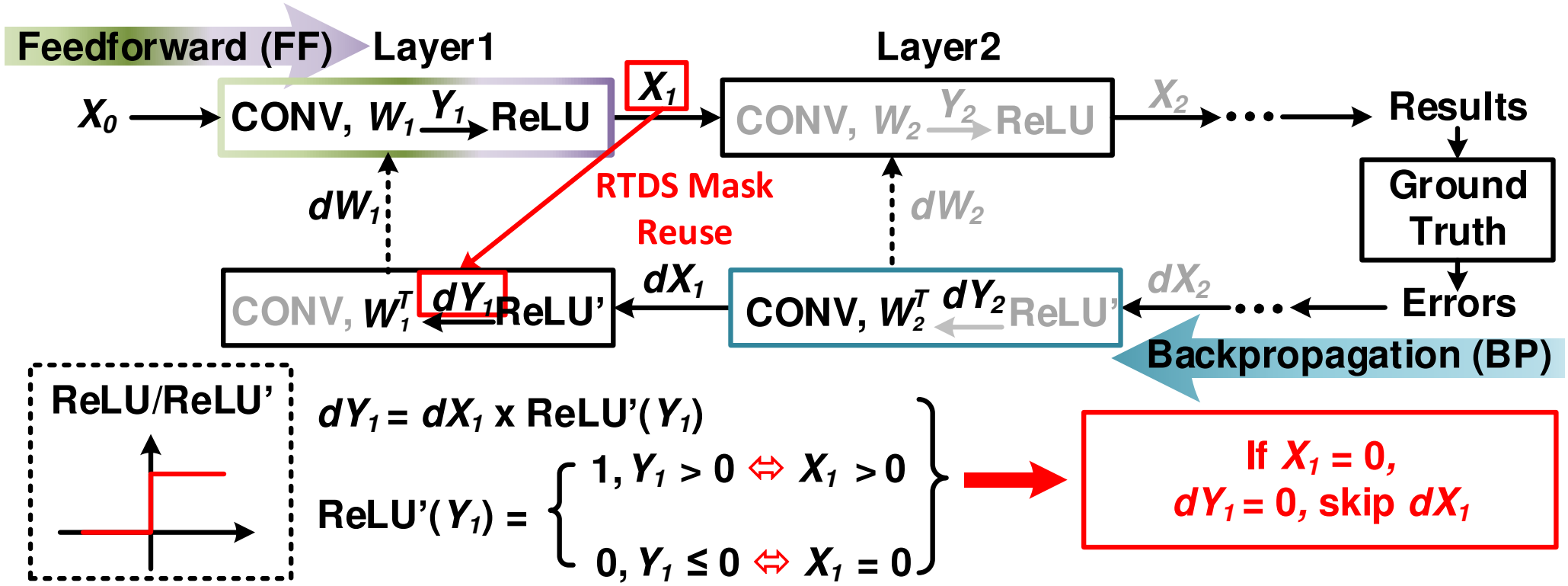
Three Design Features

1. A Bidirectional Speculation Unit (BSU) that predicts RTDS and skips zero-output computation in both FF and BP passes of training.
2. A Parallelism-Reconfigurable Engine (PRE) that enables dataflow with flexible input, output and precision parallelism.
3. A Runtime Parallelism Optimizer (RPO) that dynamically optimizes PRE's dataflow parallelism to match the time-varying workload parallelism caused by RTDS and mixed-precision.

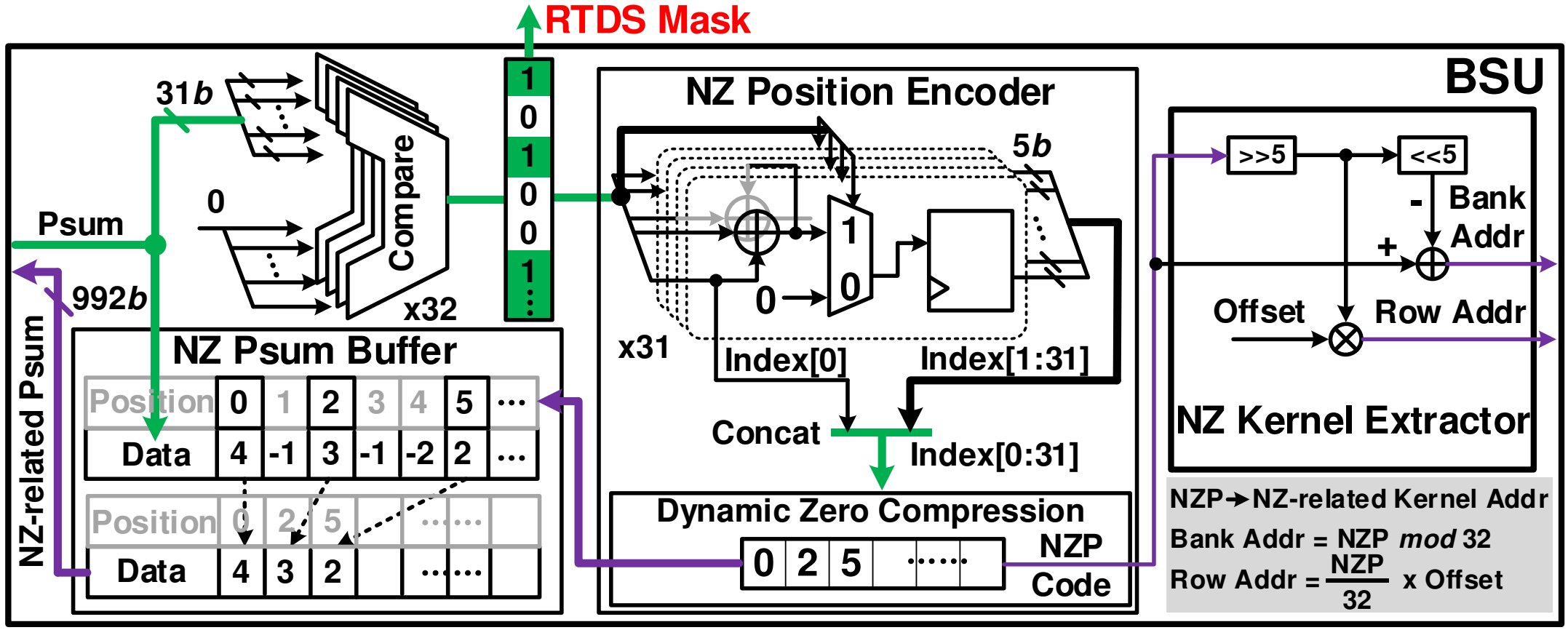
Bidirectional Speculation Mechanism



Same Zero Distribution in X and dY

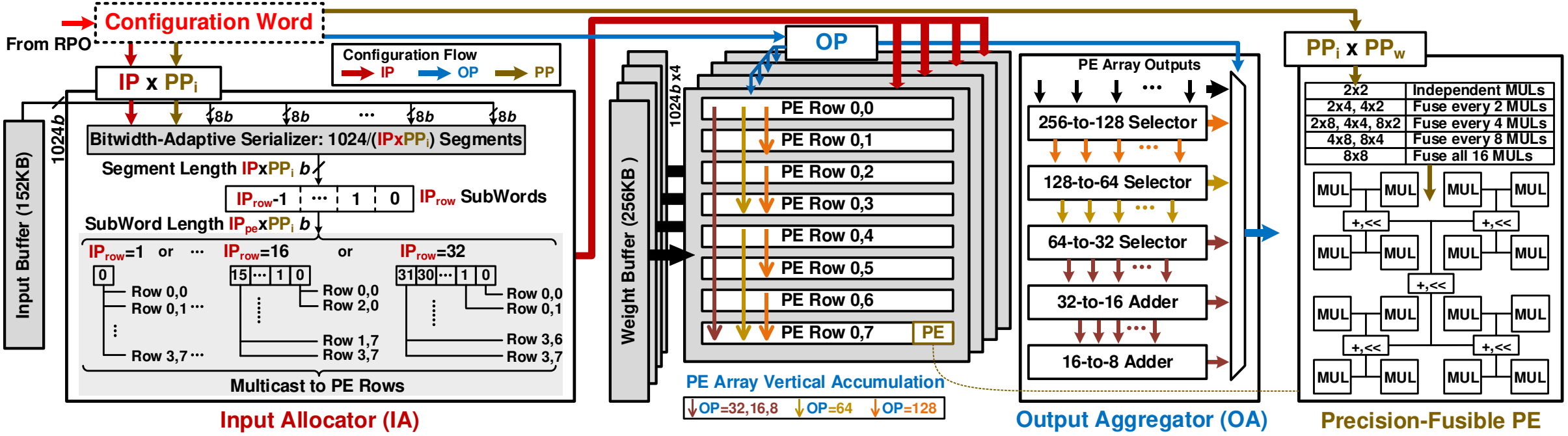


Bidirectional Speculation Unit (BSU)



Parallelism-Reconfigurable Engine (PRE)

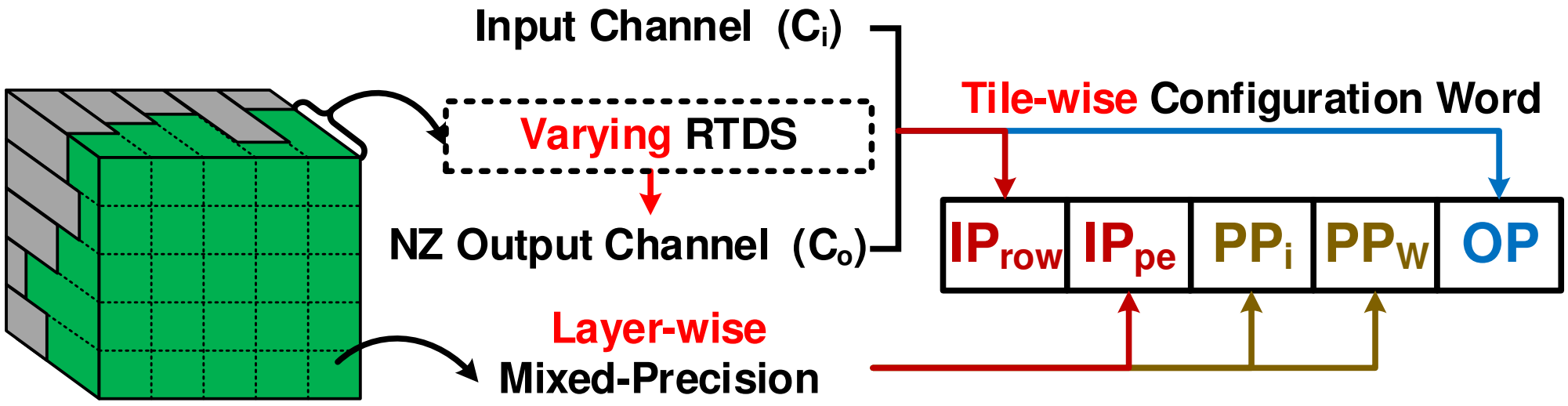
- IP: The input allocator (IA) multicasts input segments to PEs.
- OP: The output aggregator (OA) merges partial sums.
- PP: The PE supports INT2/4/8 input-weight multiplication.



RTDS & Mixed-Precision aware Reconfiguration

Tile-wise configuration word

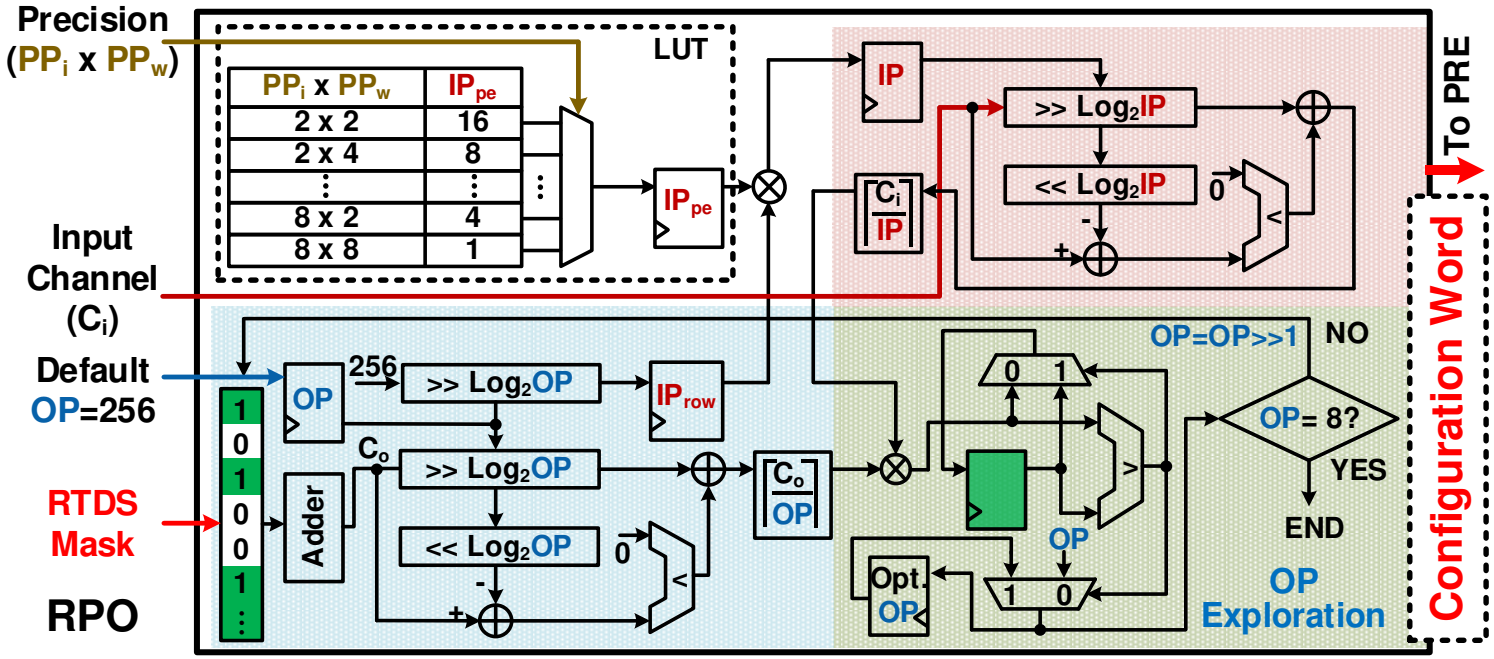
- IP: Number of input data simultaneously fed to the PEs.
- OP: Number of output data concurrently produced by the PEs.
- PP: Product of input and weight precision.



Runtime Parallelism Optimizer (RPO)

RPO finds out the optimal IP-OP pair to maximize PE utilization, based on the input channel count C_i, NZ output channel count C_o, and current precision PP.

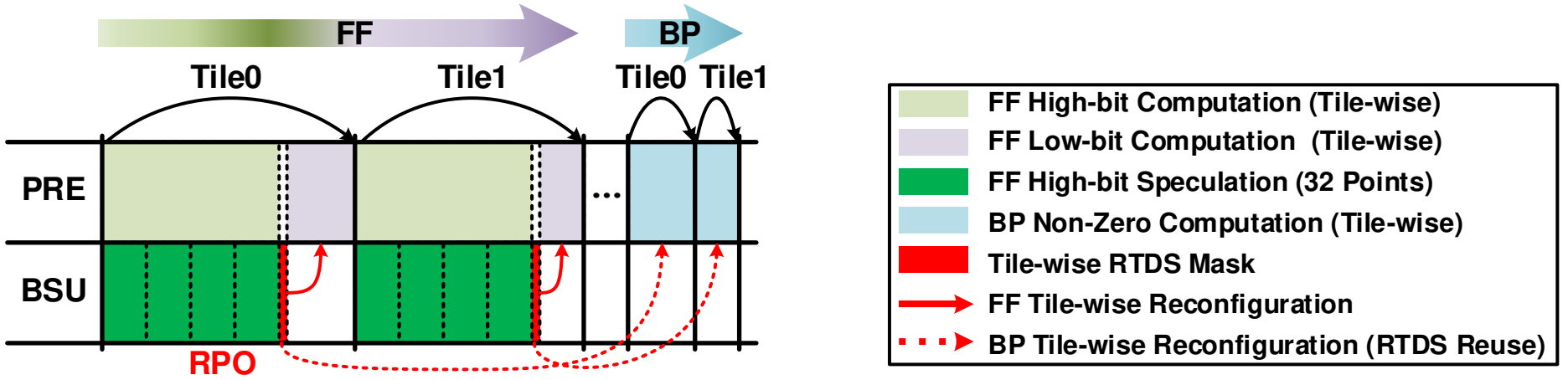
$$PE\ Utilization = \frac{C_i \times C_o}{\left\lceil \frac{C_i}{IP} \right\rceil \times \left\lceil \frac{C_o}{OP} \right\rceil \times IP \times OP}$$



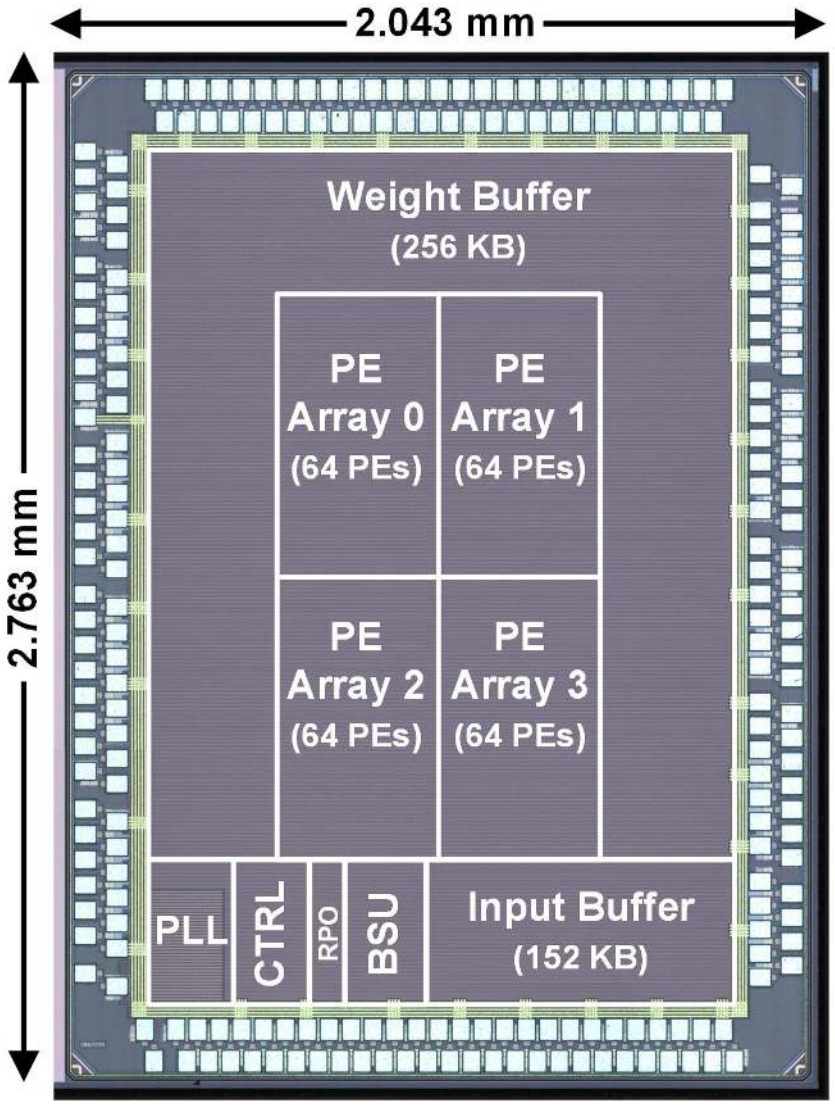
ELearn's Reconfiguration Time Chart

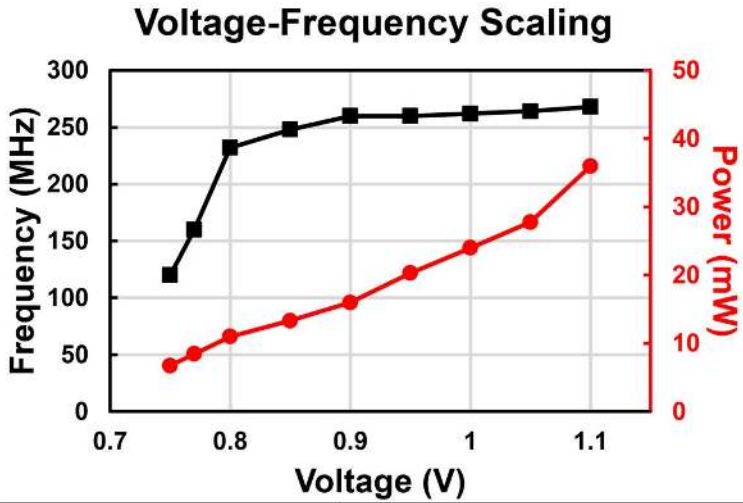
Reconfigure once for each tile.

- FF: RPO generates a configuration word for PRE's low-bit computation.
- BP: The previously FF configurations are directly reused for each tile.



ELearn's Chip Micrograph and Summary

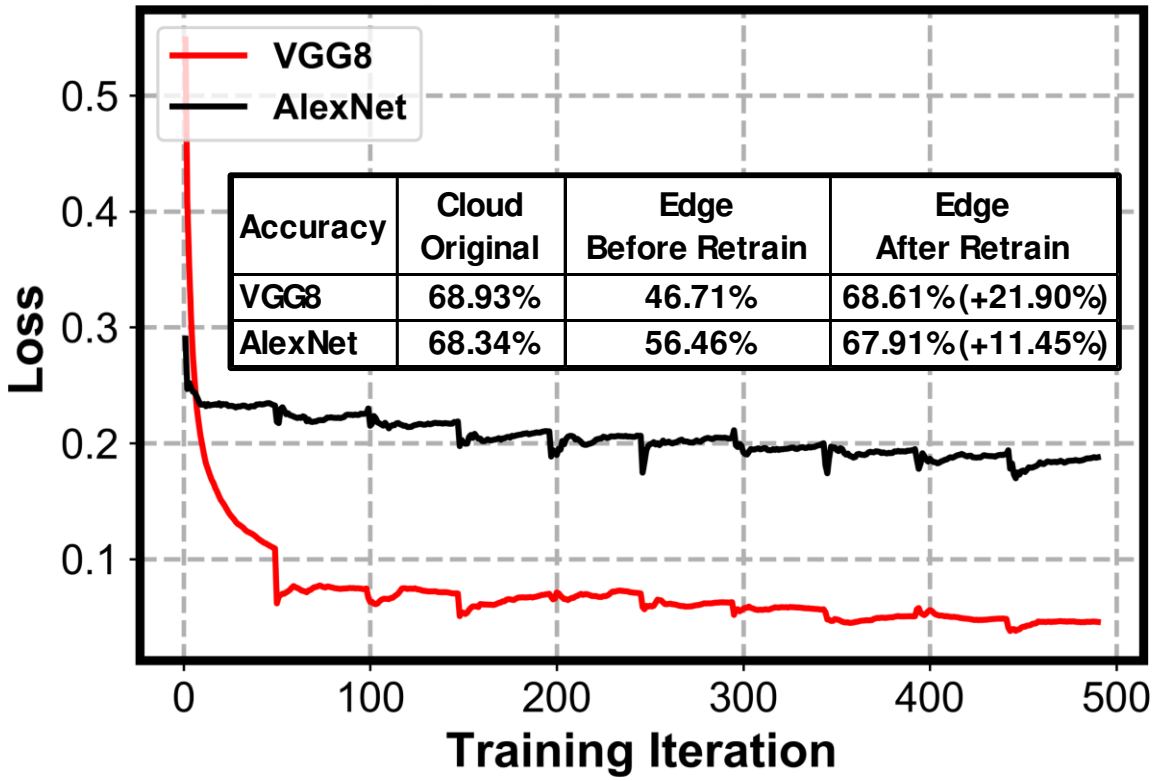




Technology	28nm 1P8M CMOS	
Die Area	2.763mm×2.043mm	
SRAM	416KB	
Supply Voltage	0.75–1.1V	
Frequency	120–268MHz	
Data Precision	Input & Weight: INT2/4/8	
Power	6.75mW @0.75V, 120MHz	
	36mW @1.1V, 268MHz	
Peak Performance	0.137TOPS @INT8x8	
	2.195TOPS @INT2x2	
Energy Efficiency (INT8x8-2x2)	0.75V, 120MHz	9.1–145.6TOPS/W
	0.80V, 232MHz	10.8–172.8TOPS/W
	1.1V, 268MHz	3.8–61.0TOPS/W

Retraining Results on Private CIFAR100

ELearn improves Top1 accuracy of VGG8 and AlexNet by 21.90% and 11.45% after retraining 500 iterations on the private CIFAR100 dataset.



Accuracy	Cloud Original	Edge Before Retrain	Edge After Retrain
VGG8	68.93%	46.71%	68.61%(+21.90%)
AlexNet	68.34%	56.46%	67.91%(+11.45%)

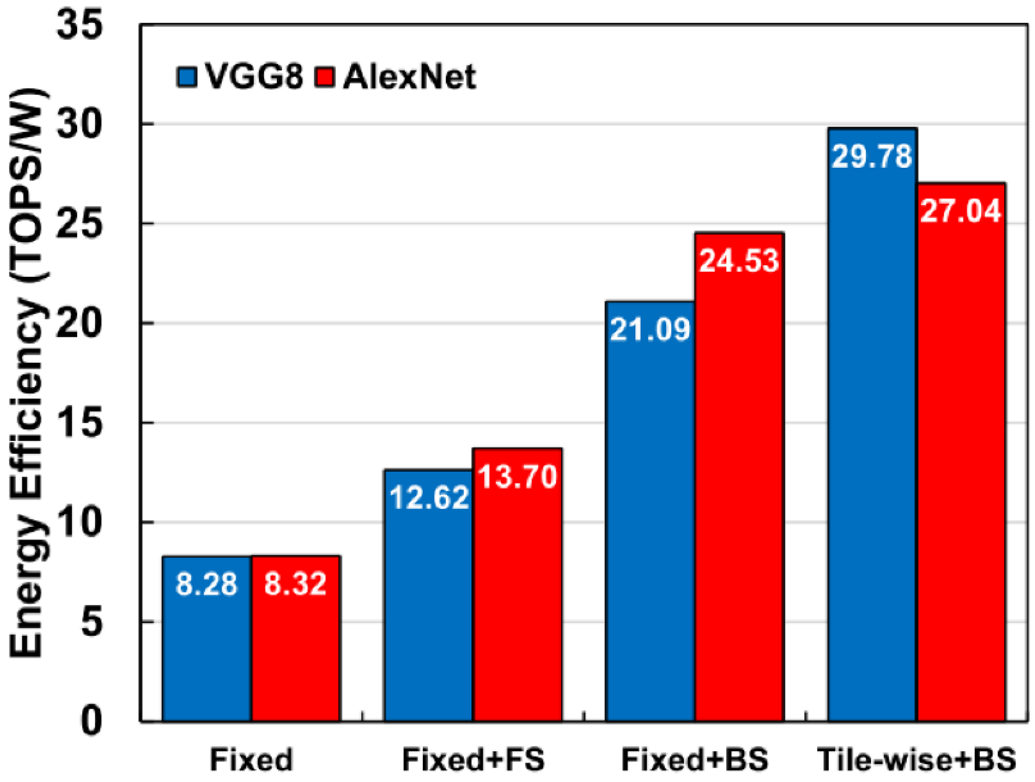
Experiment Setup

Private Dataset: 5000+1000
Batch Size: 100
500 Iterations = 10 Epoches

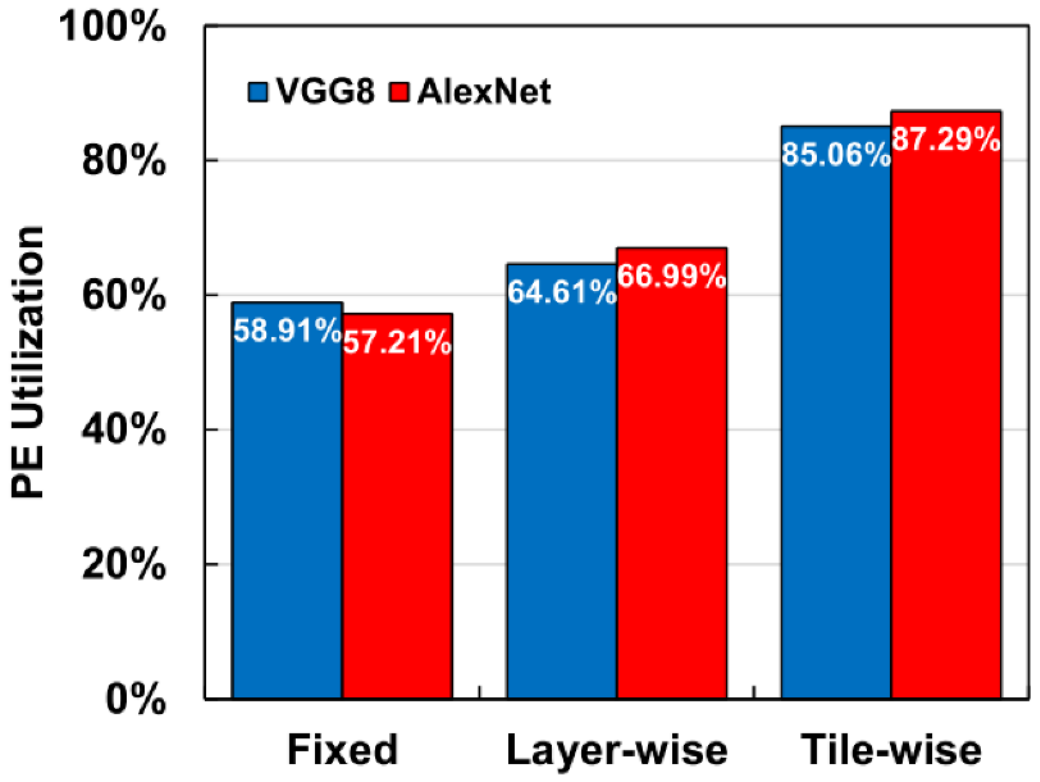
Weight Precision:
VGG8, INT2-4-8-8-4-8-4-8
AlexNet, INT8-4-8-4-8-8-8-8

Feature, Error Precision: INT8
Gradient Precision: INT16

Energy Efficiency and PE Utilization Analysis



(a) Energy efficiency.



(b) PE utilization.

Comparison with State-of-the-art DNN Training Processors

	VLSI'19 [1]	ISSCC'19 [2]	ASSCC'19 [3]	ASSCC'19 [4]	VLSI'18 [5]	ISSCC'20 [6]	This Work
Sparsity Support	×	√	×	×	×	√	√
Dynamic Reconfig.	×	×	×	×	×	×	√
Technology	65nm	65nm	65nm	40nm	14nm	65nm	28nm
Die Area (mm^2)	5.76	16	10.24	5	9	32.4	5.64
Data Precision	INT13/16	FP8/16	INT8	INT5/10, FP16	INT2/3, FP16	FP8/16	INT2/4/8
Frequency (MHz)	50-200	50-200	5-160	82-232	750-1500	25-200	120-268
Supply Voltage (V)	0.78-1.1	0.78-1.1	0.63-1.0	0.6-0.9	0.58-0.9	0.70-1.1	0.75-1.1
Power Consumption (mW)	36.8 @0.78V, 50MHz	43.1 @0.78V, 50MHz	9.55 @0.63V, 5MHz	18.7 @0.6V, 82MHz (INFER, INT5/10)	-	58 @0.70V, 25MHz	6.75 @0.75V, 120MHz
	252.4 @1.1V, 200MHz	367 @1.1V, 200MHz	120.5 @ 1.0 V, 80MHz	64.5 @0.6V, 82MHz (TRAIN, FP16)	-	647 @1.1V, 200MHz	36 @1.1V, 268MHz
Peak Performance (TOPS or TFLOPS)	0.155 @INT13/16	>0.3 @FP16	0.13 @INT8	-	1.5 @FP16	14.03 @FP16	0.137 @INT8 [1]
		>0.6 @FP8			12 @INT3	24.13 @FP8	2.195 @INT2 [1]
					24 @INT2		
Energy Efficiency (TOPS/W or TFLOPS/W)	1.32 @INT13/16	15.6 @FP16	1.03 @INT8	2.25 (INFER, Norm. to INT16)	-	1.81-75.68 @FP16	32.9 @INT8 [2]
		25.3 @FP8		0.65 (TRAIN, Norm. to INT16)		3.62-135.10 @FP8	130.6 @INT4 [2]
							172.8 @INT2 [2]

1) @1.1V, 268MHz; 2) @0.8V, 232MHz, VGG8 retraining.

References

[1] D. Han, J. Lee, J. Lee, and H. Yoo, “A 1.32 tops/w energy efficient deep neural network learning processor with direct feedback alignment based heterogeneous core architecture,” in VLSI, 2019.

[2] J. Lee, J. Lee, D. Han, J. Lee, G. Park, and H.-J. Yoo, “lnpu: A 25.3 tflops/w sparse deep-neural-network learning processor with fine-grained mixed precision of fp8-fp16,” in ISSCC, 2019.

[3] S. Choi, J. Sim, M. Kang, Y. Choi, H. Kim, and L.-S. Kim, “A 47.4j/epoch trainable deep convolutional neural network accelerator for in-situ personalization on smart devices,” in ASSCC, 2019.

[4] C.-H. Lu, Y.-C. Wu, and C.-H. Yang, “A 2.25 tops/w fully-integrated deep cnn learning processor with on-chip training,” in ASSCC, 2019.

[5] B. Fleischer, S. Shukla, M. Ziegler, J. Silberman, J. Oh, V. Srinivasan, J. Choi, S. Mueller, A. Agrawal, T. Babinsky, N. Cao, C. Chen, P. Chuang, T. Fox, G. Gristede, M. Guillorn, H. Haynie, M. Klaiber, D. Lee, S. Lo, G. Maier, M. Scheuermann, S. Venkataramani, C. Vezyrtzis, N. Wang, F. Yee, C. Zhou, P. Lu, B. Curran, L. Chang, and K. Gopalakrishnan, “A scalable multi- teraops deep learning processor core for ai trainina and inference,” in VLSI, 2018.

[6] S. Kang, D. Han, J. Lee, D. Im, S. Kim, and H.-J. Yoo, “Ganpu: A 135tflops/w multi-dnn training processor for gans with speculative dual-sparsity exploitation,” in ISSCC, 2020.