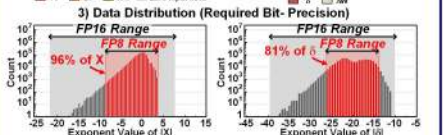
LNPU: An Energy-Efficient Deep-Neural-Network Training Processor with Fine-Grained Mixed Precision

Jinsu Lee (jinsulee@kaist.ac.kr), Juhyoung Lee, Donghyeon Han, Jinmook Lee, Gwangtae Park and Hoi-Jun Yoo

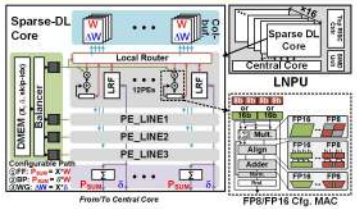
KAIST

DNN Training Analysis

1) ~43% of Operation is Zero Input MAC Operation
2) X and δ occupy ~78% of External Memory Access
3) FP8 Cover ~80% of overall data @ ResNet-18



Overall Architecture

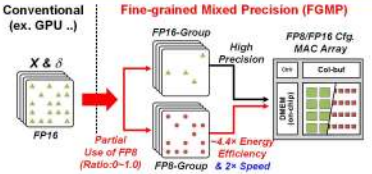


1) Fine-Grained Mixed Precision of float16 / float8
- Improve Performance & Lower EMA

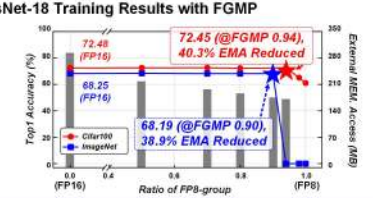
2) Sparse DL Core
- Skip Zero Input MAC Op. caused by FGMP

3) Central Core
- Global Mem. Manage for Irregular Core latency

Proposed Fine-Grained Mixed Precision

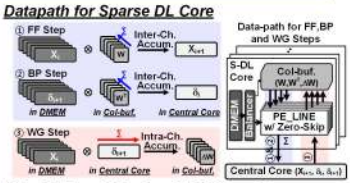


ResNet-18 Training Results with FGMP

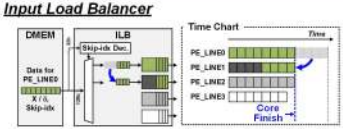


Sparse DL Core & Central Core

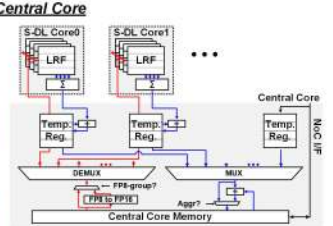
Datapath for Sparse DL Core



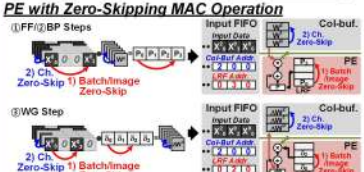
Input Load Balancer



Central Core



PE with Zero-Skipping MAC Operation



Implementation Results

Chip Photo & Summary

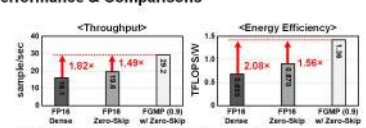
Specifications	
Technology	65nm 1P8M CMOS
Die Area	4mm × 4mm (16mm²)
SRAM	372 KB
Supply Voltage	0.78V ~ 1.1V
Frequency	~ 200MHz
Data Type	FP8, FP16
Power Consumption	43.1mW @ 0.78V, 50MHz 367mW @ 1.1V, 200MHz

Power Efficiency (TFLOPS/W)		FP16	FP8
	50MHz, @0.78V	1.74	3.48
		-15.6*	-29.3*
	200MHz, @1.1V	0.817	1.63
		-7.32*	-11.9*

*Effective TFLOPS/W with 90% Input Sparsity

Autonomous Drone System w/ LUPU

Performance & Comparisons



VGG16 Conv-layer Inference	JSSC'17 [5]	JSSC'17 [6]	JSSC'18 [2]	LNPU
Frame Rate (fps)	0.7	1.67	18.3	32.8
Power (mW)	234	26	297	367
Energy-Eff. (TOPS/W)	0.092	2.9	1.91	2.74
bit-precision	INT16	N/A	INT8	FP8