An Energy-Efficient Unified Deep Neural Network Accelerator with Fully-Variable Weight Precision for Mobile Deep Learning Applications

Jinmook Lee (jinmooklee@kaist.ac.kr), Changhyeon Kim, Sanghoon Kang, Dongjoo Shin, Sangyeob Kim, and Hoi-Jun Yoo

KAIST

Motivations

Requirement 1

- **Needs to support variable weight precisions**

A. Differnet DNNs have different optimal precision
B. Desired weight precision depends on target accuracy

Requirement 2

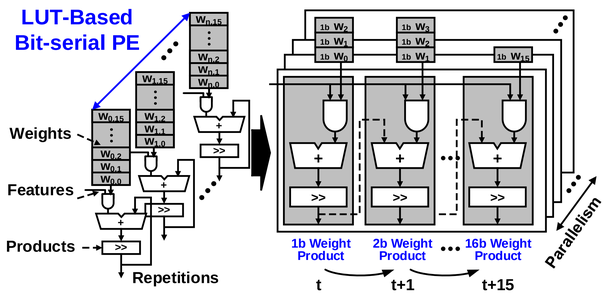
- **To unify DNN Core arch. for higher utilization**

A. Prior work "DNPU" has heterogeneous cores, it is not suitable for various DNN workloads. (Low Util.)

This Work - UNPU

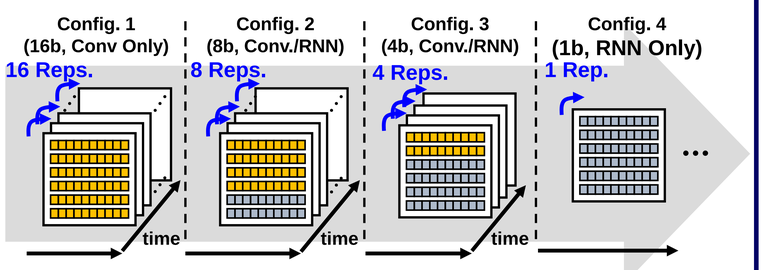
1. LUT-based Bit-serial PE (Fully variable 1b to 16b)

- Enhances energy-efficiency by 53 % than fxp arithematic (1b)

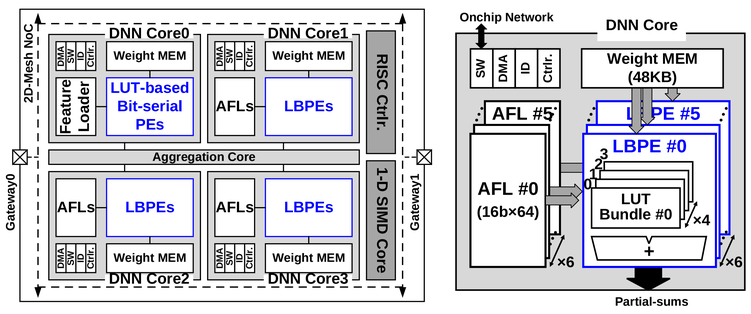


2. Unified Datapath for CNN/RNN/FC with Fmap reuse

- Support fully-variable (1bit ~ 16bit) weight precision
- Support versatile CNN & RNN workload combinations



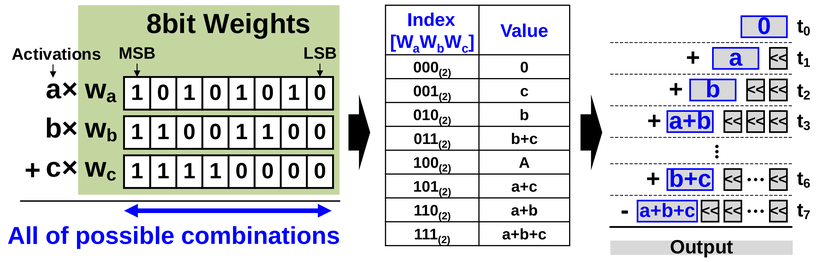
Overall Architecture of UNPU



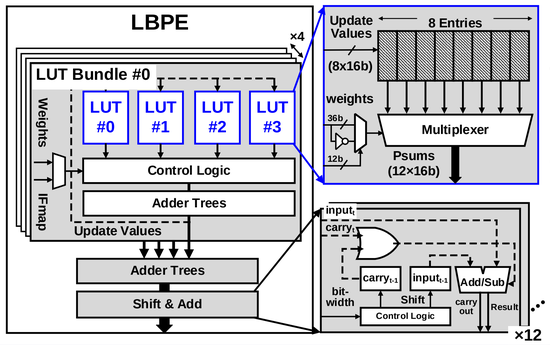
- Parallel Bit-serial MAC array (18,432 Bit-serial MACs/cycle)

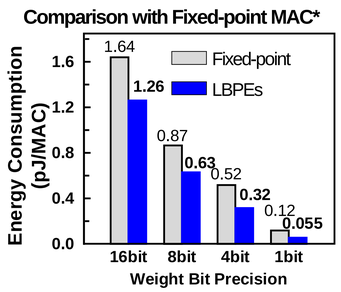
Table-based Bit-serial Processing

Concept: Makes partial-sum table for given input activations



Architecture: LBPE accumulate partial-sums from LUTs





*Samsung 65nm process, 200MHz, 1.2V, 576MACs/(Ncycles) (N=1,4,8,16)

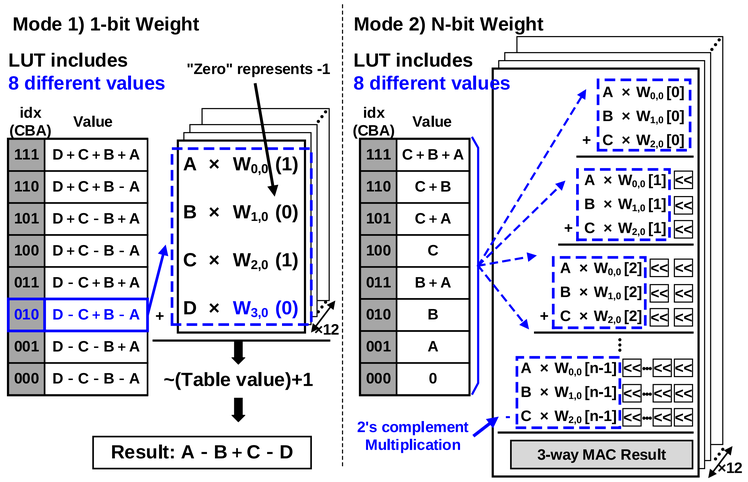
Two Modes of LBPE

A. Multi-bit mode (2, 3, ,,,. 16bit)

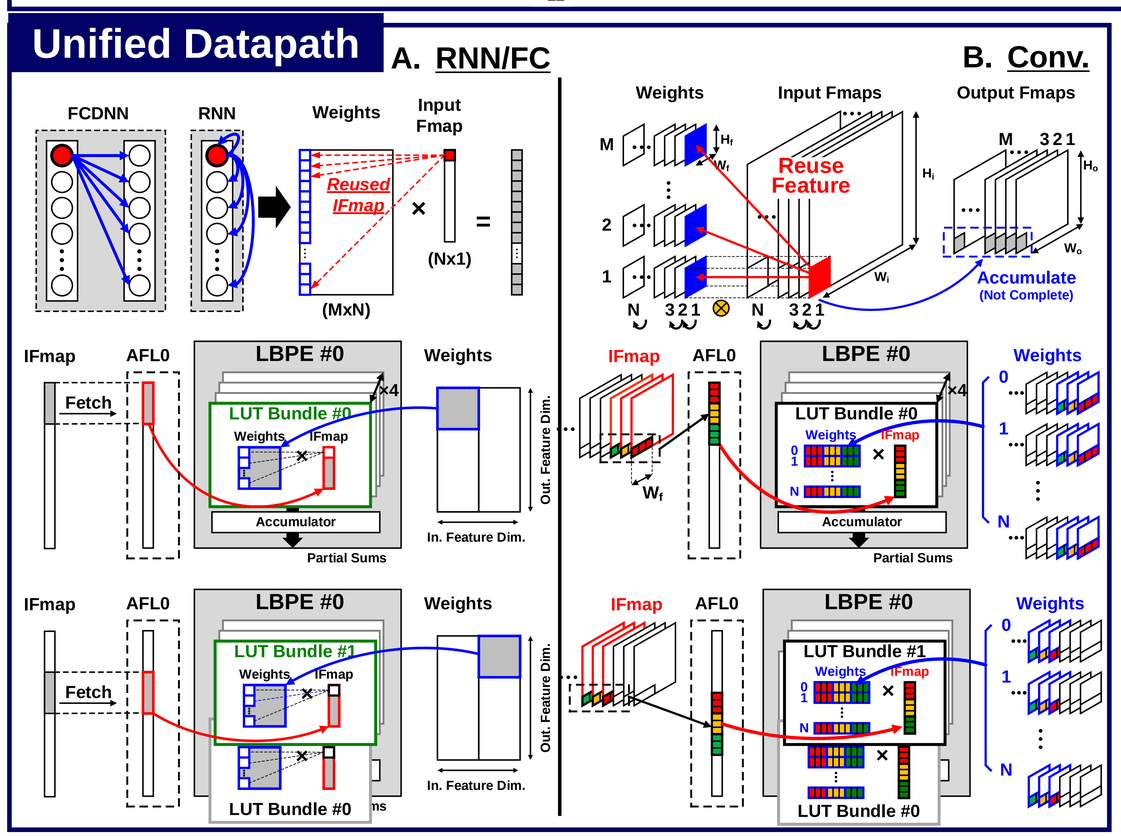
- a LUT performs 3-way MAC

B. 1-bit mode (weight is either 0 or 1)

- a LUT performs 4-way MAC using 2's complement
- if weight of $D(W_{3,0})$ is -1, negates the table values



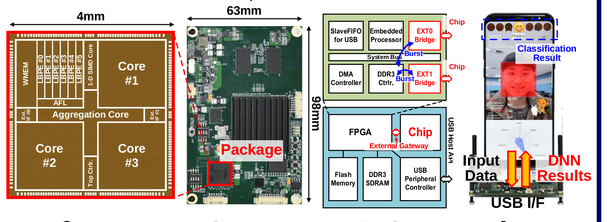
Unified Datapath



Implementation Results

Demonstration System

- UNPU gets input data from Android device, accelerates DNNs, and transfer DNN results.



Performance Summary & Comparisons

- Peak 7,372 GOPS (1b) for CNN/RNN/FC
- Peak Power efficiency of 50.6 TOPS/W (1bit)

AlexNet Benchmark	ISSCC'17 [1]	ISSCC'17 [2]	This Work		
Frame Rate [fps]	177	47	**19.7**	**346**	**158**
Power Eff. [TOPS/W]	4.2	1.42	**4.3**	**1.59**	**27.3**
Power [mW]	63	44	**6.1**	**290**	**7.7**
Bit Precision	8	N/A	**8**	**8**	**1**

VGG-16 Benchmark	ISSCC'17 [2]	JSSC'17 [7]	This Work	
Frame Rate [fps]	1.67	0.7	**0.97**	**18.3**
Power Eff. [TOPS/W]	2.0	0.092	**4.7**	**1.91**
Power [mW]	26	236	**6.4**	**297**
Bit Precision	N/A	16	**8**	**8**

	ISSCC'17 [1]	ISSCC'18 [8]	This Work
Technology	65nm	40nm	**65nm**
Area [mm^2]	16	121.6	**16**
PE Precision	4, 8, 16	1 - 4	**1 - 16**
Performance [GOPS]	1,200(4b)	7,490(1b) 1,960(4b)	**7,372(1b)** **1,382(4b)**
Power Eff. [TOPS/W]	8.1 (4b)	2.27 (1b) 0.59 (4b)	**50.6 (1b)** **11.6 (4b)**
Area Eff. [GOPS/mm^2]	75 (4b)	61.6 (1b) 16.2 (4b)	**461 (1b)** **86.4 (4b)**
Power [mW]	35 - 279	3300	**3.2 - 297**