IEEE Xplore® Digital Library

ISCAS 2020

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

CAS

A Convolutional Neural Network Accelerator Architecture with Fine-Granular Mixed Precision Configurability

Xian Zhou , Li Zhang , Chuliang Guo , Xunzhao Yin , Cheng Zhuo

College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China

2020 IEEE International Symposium on Circuits and Systems Virtual,
October 10-21, 2020

浙江大學
Zhejiang University

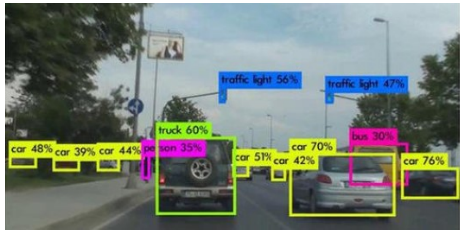
Outline

- Motivation
- Architecture and Dataflow
- Optimization and Tradeoff
- Experimental Results

CNN in Mobile Applications

CNN has been widely deployed in various deep learning domains

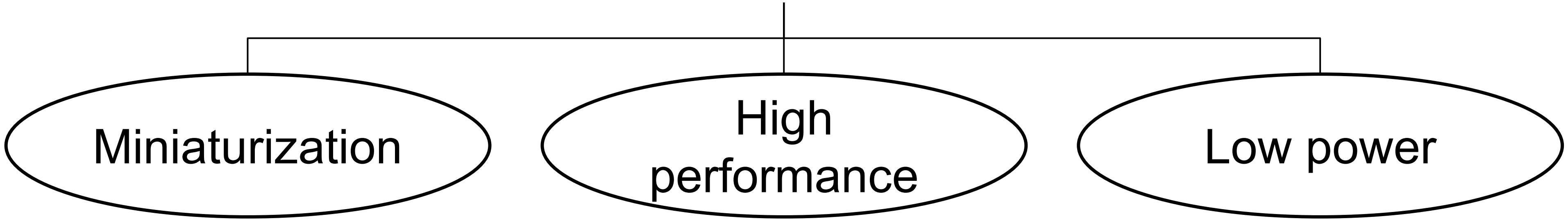
Traffic



Industry

Monitor

Characteristics



Miniaturization

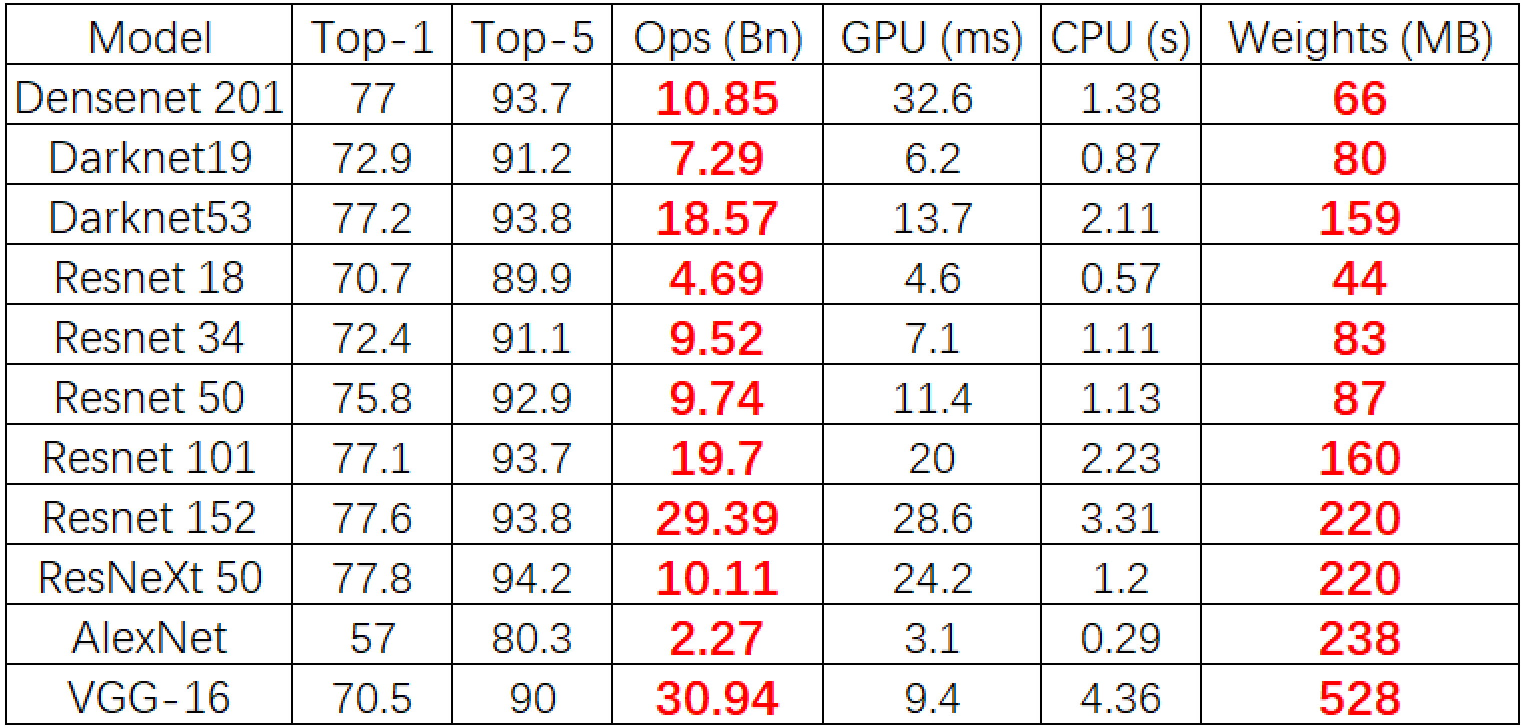
High performance

Low power

CNN hardware accelerator become a popular solution

Difficulties in CNN Deployment

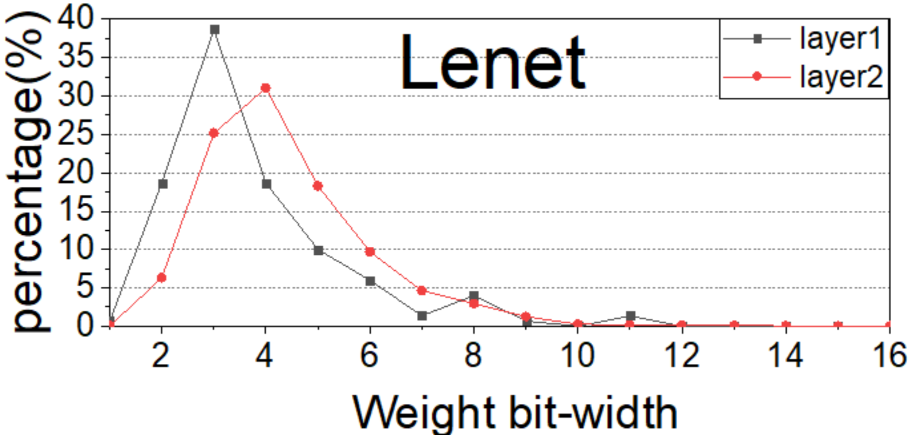
CNN algorithm complexity continues growing and significant arithmetic and storage consumption

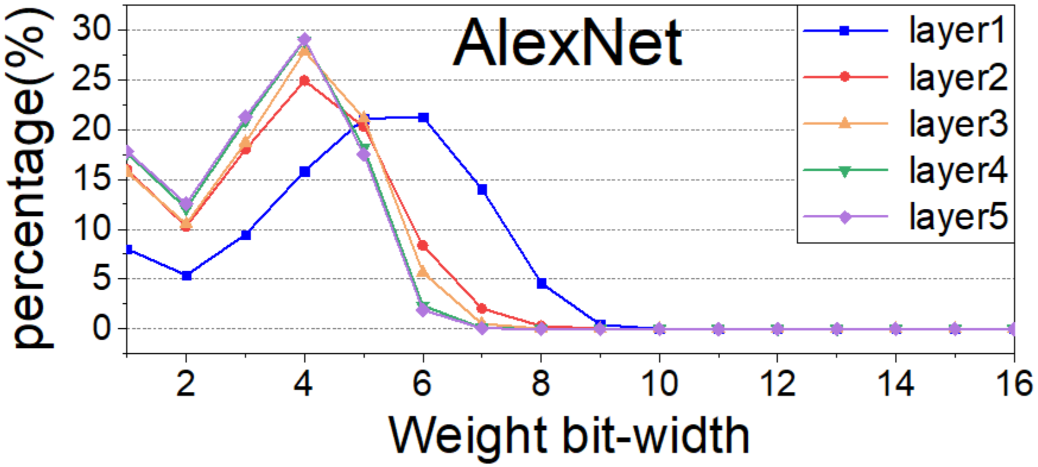


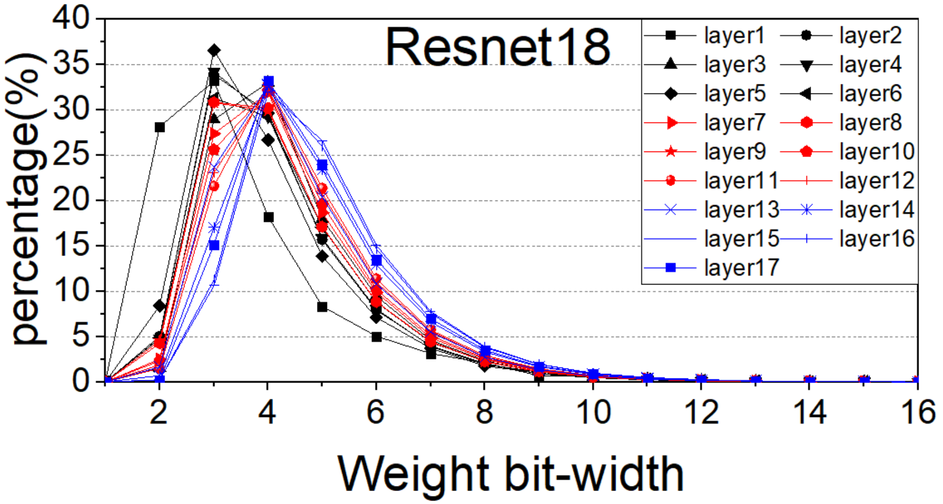
Model	Top-1	Top-5	Ops (Bn)	GPU (ms)	CPU (s)	Weights (MB)
Densenet 201	77	93.7	**10.85**	32.6	1.38	**66**
Darknet19	72.9	91.2	**7.29**	6.2	0.87	**80**
Darknet53	77.2	93.8	**18.57**	13.7	2.11	**159**
Resnet 18	70.7	89.9	**4.69**	4.6	0.57	**44**
Resnet 34	72.4	91.1	**9.52**	7.1	1.11	**83**
Resnet 50	75.8	92.9	**9.74**	11.4	1.13	**87**
Resnet 101	77.1	93.7	**19.7**	20	2.23	**160**
Resnet 152	77.6	93.8	**29.39**	28.6	3.31	**220**
ResNeXt 50	77.8	94.2	**10.11**	24.2	1.2	**220**
AlexNet	57	80.3	**2.27**	3.1	0.29	**238**
VGG-16	70.5	90	**30.94**	9.4	4.36	**528**

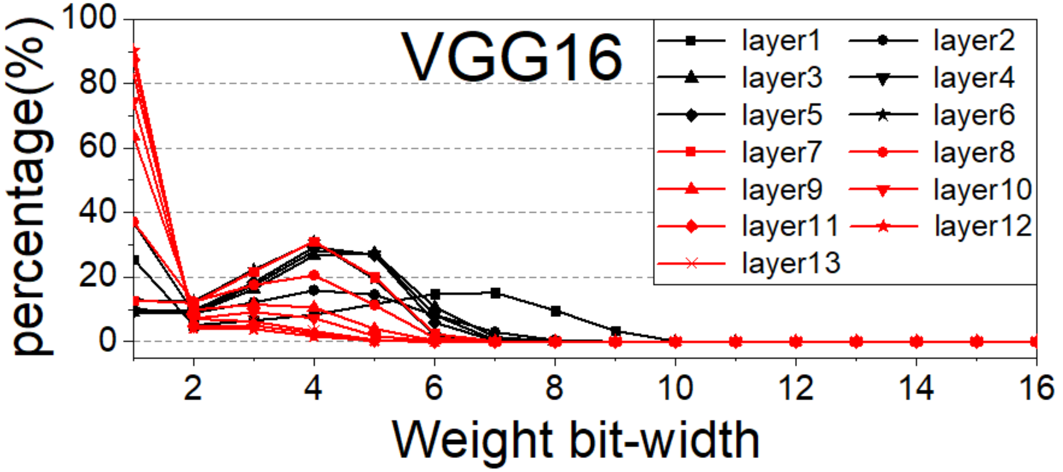
High consumption and limited memory size in hardware deployment

Weight Distribution





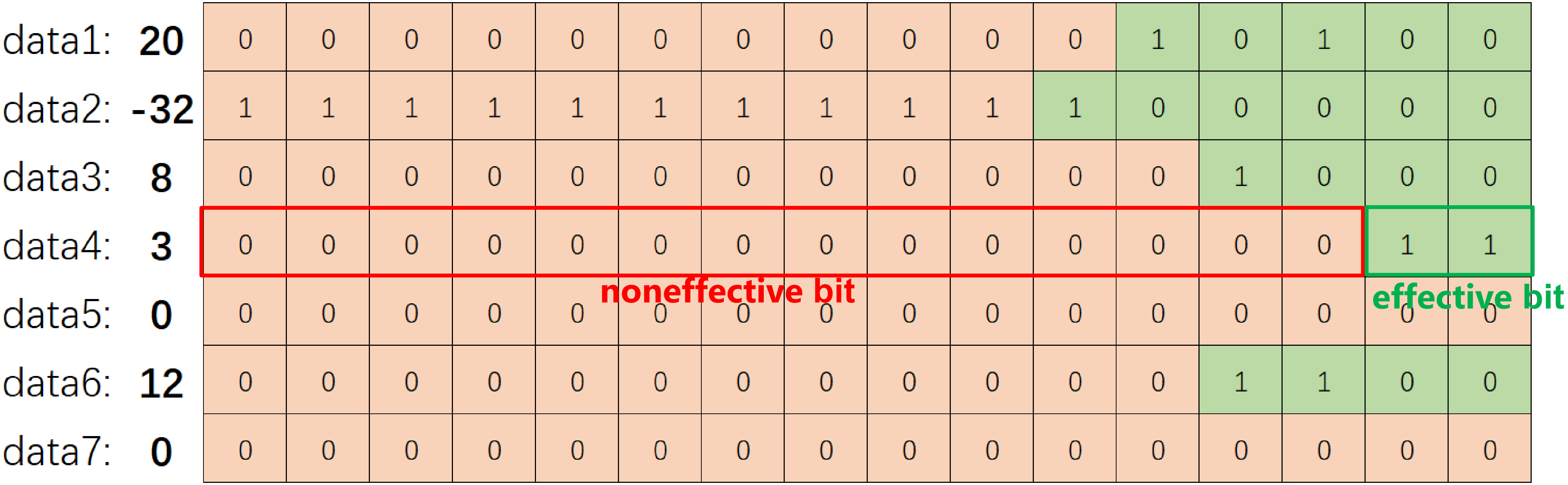




Most weights with a low effective bit-width and few with a high effective bit-width

Weight Storage Pattern in A Single Precision System

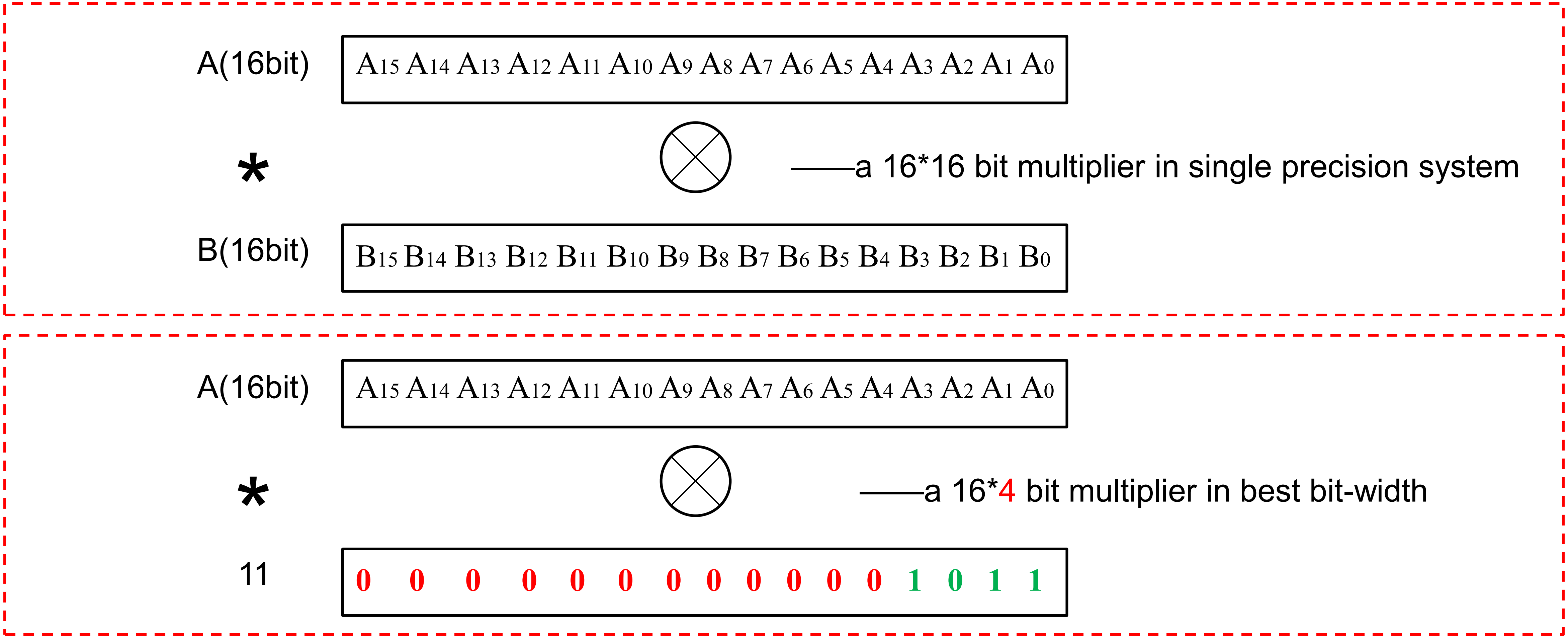
Format of 16-bit fixed-point data in Memory



data1: **20**	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
data2: **-32**	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
data3: **8**	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
data4: **3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
data5: **0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
data6: **12**	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
data7: **0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Too many noneffective bit makes a significant waste of storage resources!

Operation in A Single Precision System



A(16bit) A_{15} A_{14} A_{13} A_{12} A_{11} A_{10} A_9 A_8 A_7 A_6 A_5 A_4 A_3 A_2 A_1 A_0

* ⊗ ——a 16*16 bit multiplier in single precision system

B(16bit) B_{15} B_{14} B_{13} B_{12} B_{11} B_{10} B_9 B_8 B_7 B_6 B_5 B_4 B_3 B_2 B_1 B_0

A(16bit) A_{15} A_{14} A_{13} A_{12} A_{11} A_{10} A_9 A_8 A_7 A_6 A_5 A_4 A_3 A_2 A_1 A_0

* ⊗ ——a 16*4 bit multiplier in best bit-width

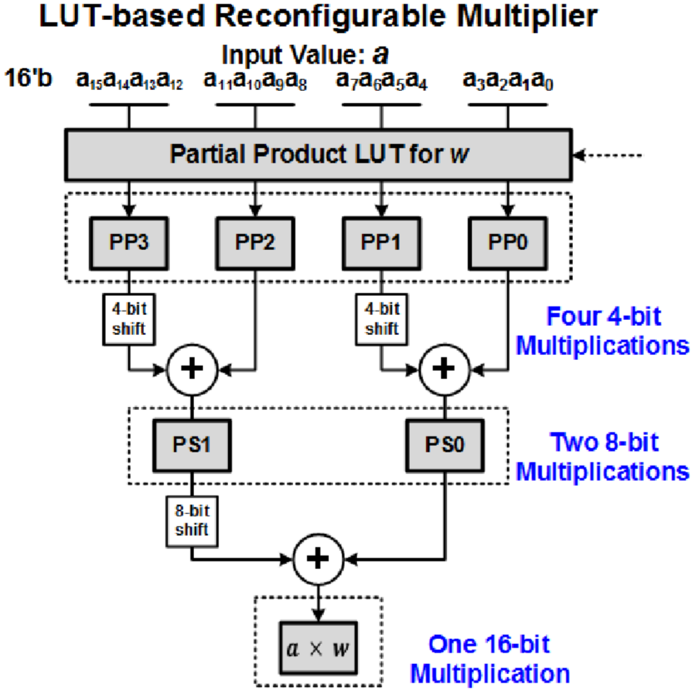
11 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1

All data is processed using the multiplier with the largest bit-width

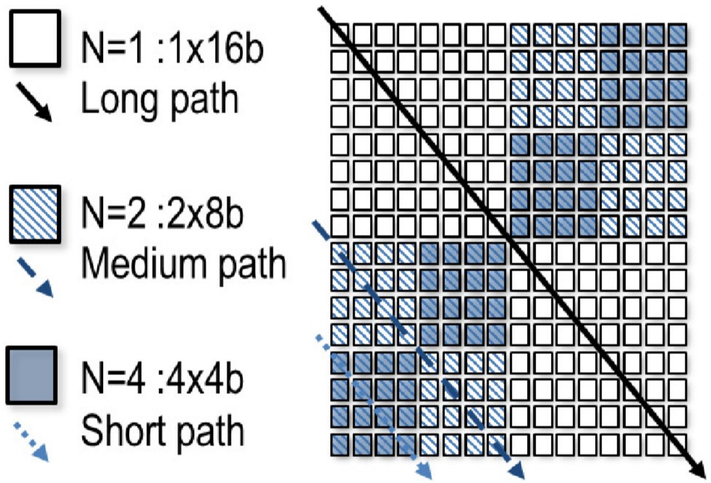
A significant waste of operation resources!

Related Work

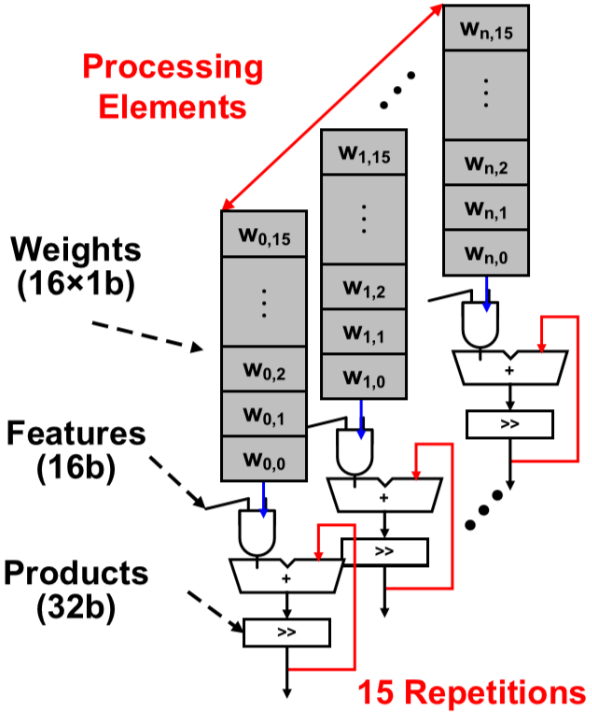
DNPU(ISSCC2017)


LUT-based Reconfigurable Multiplier
Input Value: a
16'b
$a_{15}a_{14}a_{13}a_{12}$
$a_{11}a_{10}a_9a_8$
$a_7a_6a_5a_4$
$a_3a_2a_1a_0$
Partial Product LUT for w
PP3
PP2
PP1
PP0
4-bit shift
4-bit shift
Four 4-bit Multiplications
PS1
PS0
Two 8-bit Multiplications
8-bit shift
$a \times w$
One 16-bit Multiplication

ENVISION(ISSCC2017)

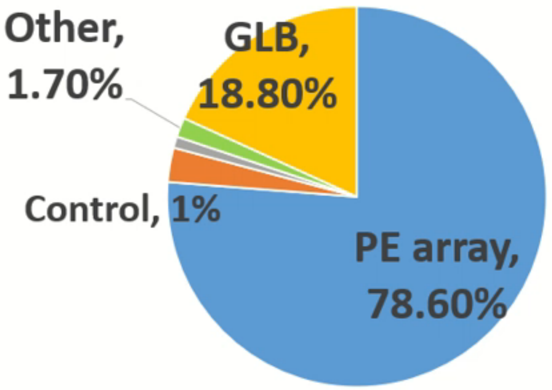


UNPU(JSSC2019)

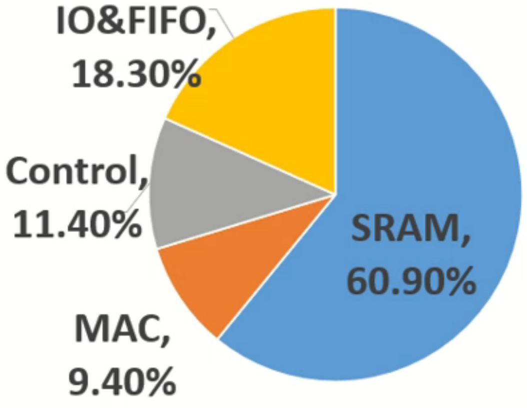


The prior work do **not** support mixed precision computing within the same layer

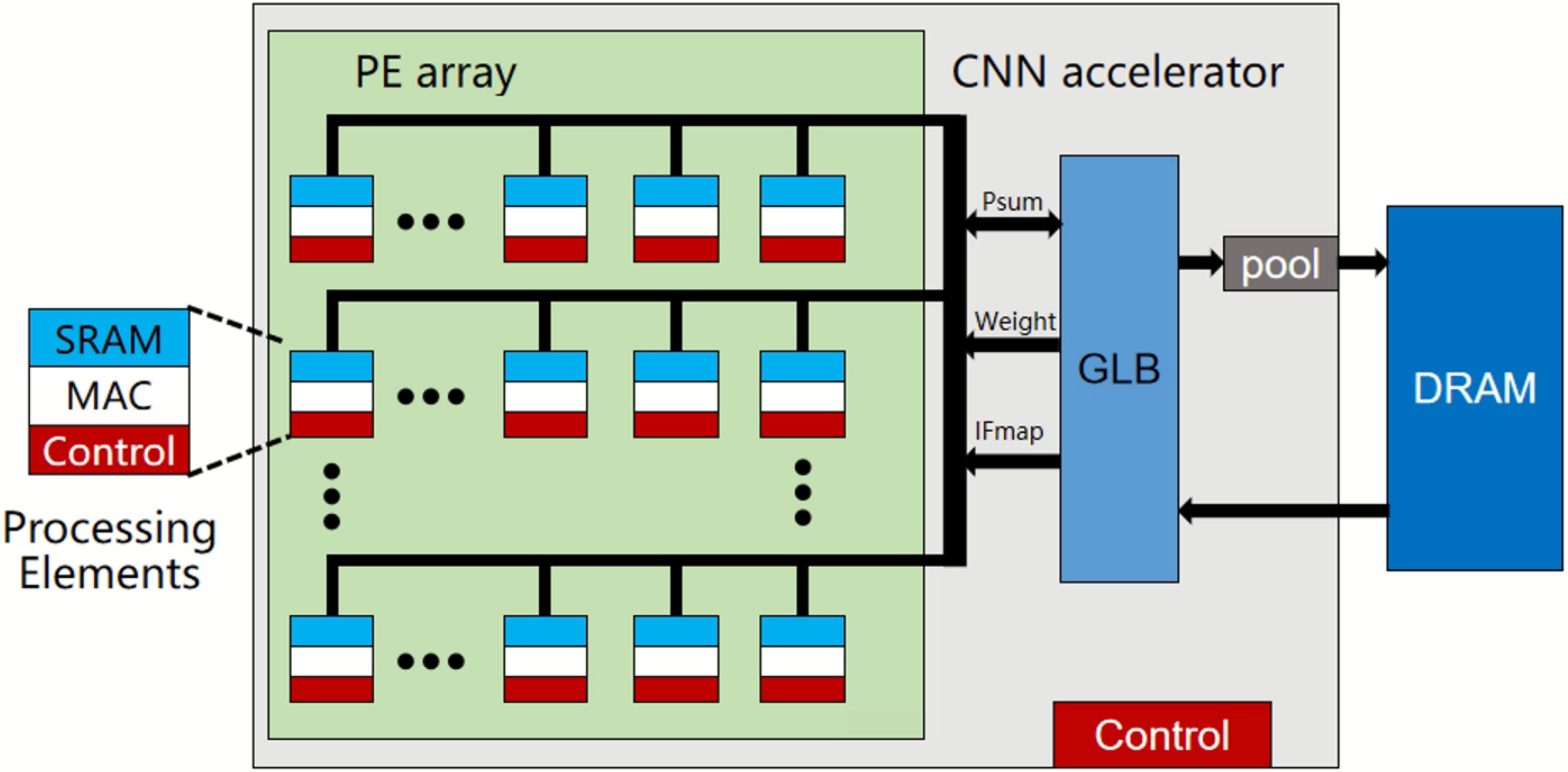
Commonly-Used CNN Accelerators Architecture



Accelerator*

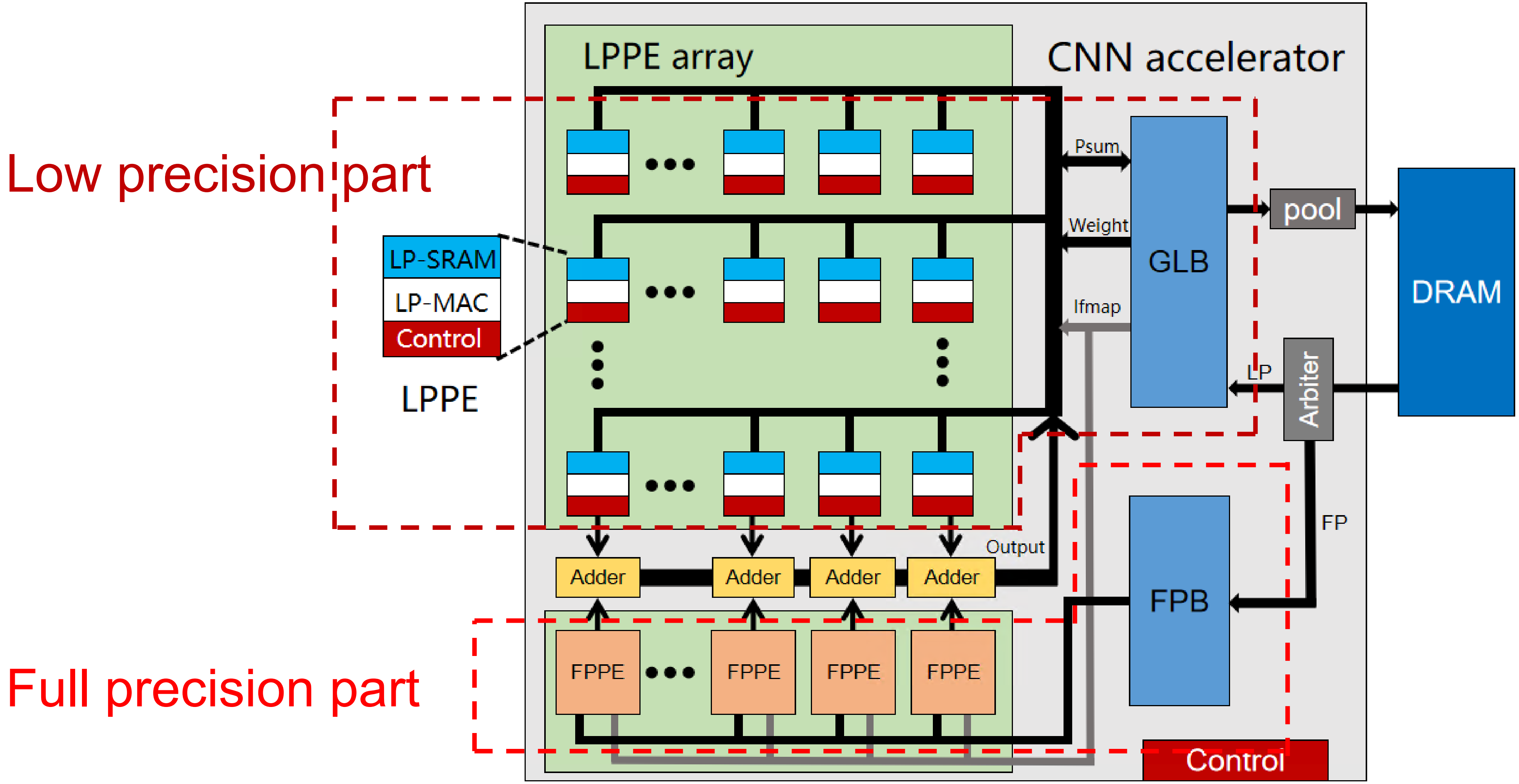


PE*

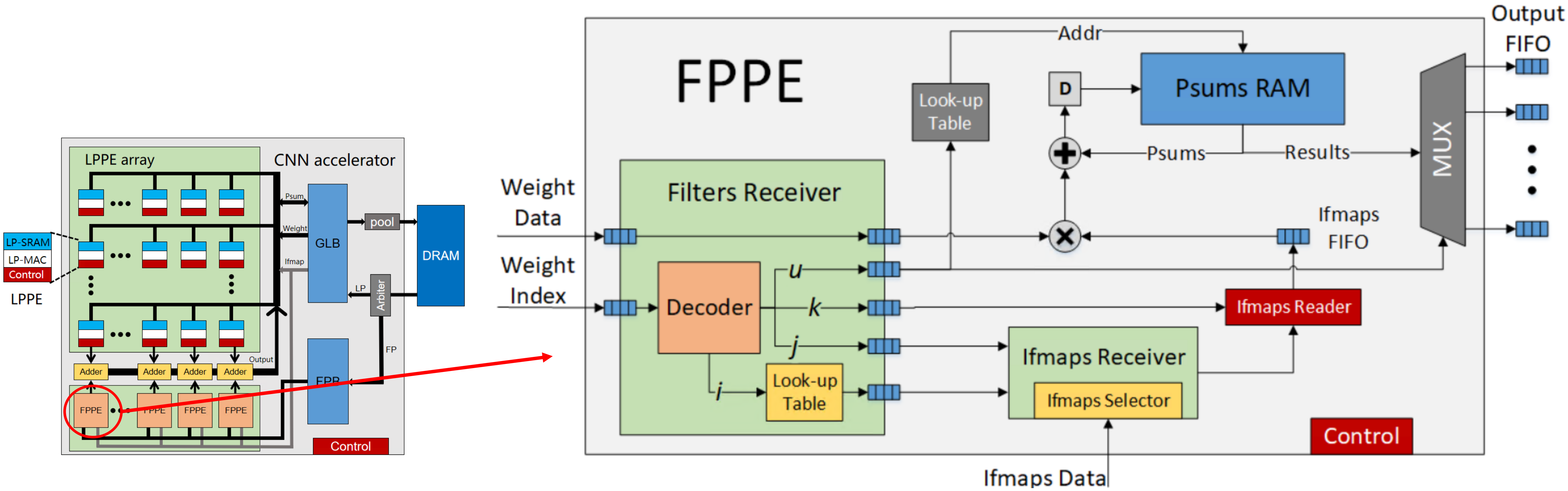


*Data from Y. H. Chen, etc, "Eyeriss: An energy efficient reconfigurable accelerator for deep convolutional neural networks," IEEE J. Solid-State Circuits

Mixed Precision CNN Accelerators Architecture

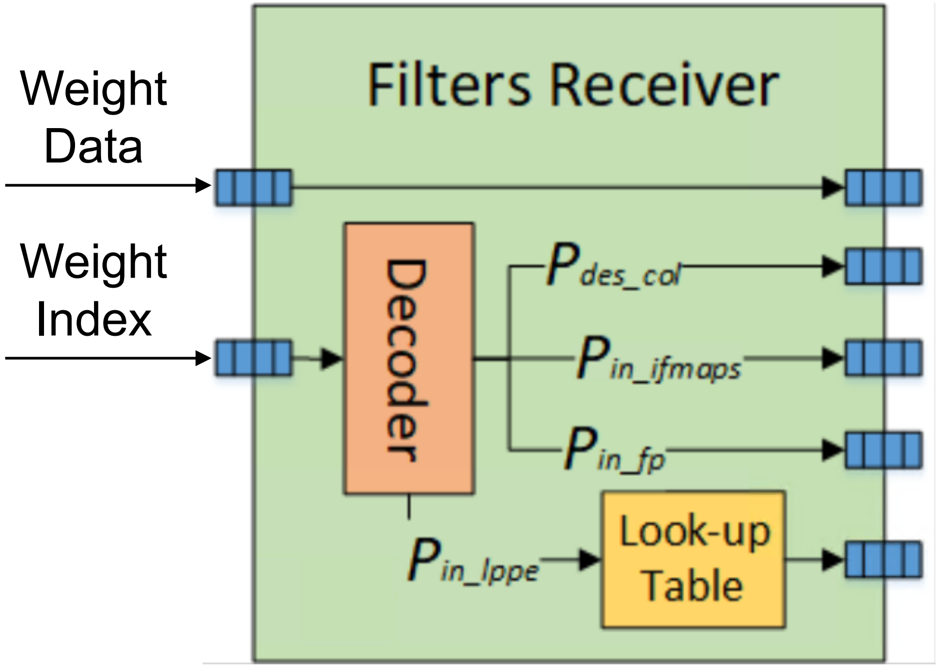


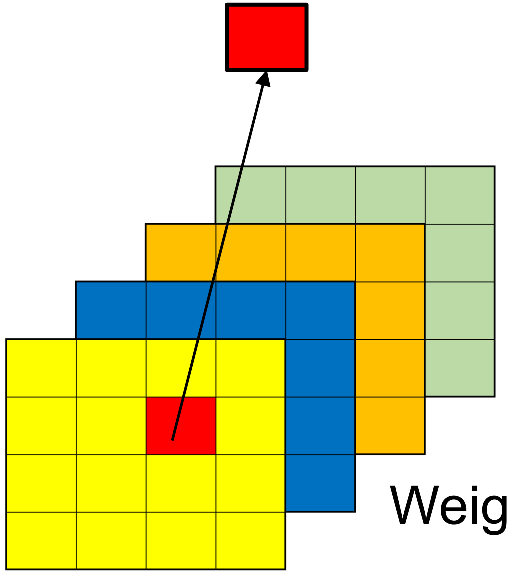
Structure of The Proposed FPPE



Decoder in FPPE

Decoder translates weight position into FPPE parameter by PE and CNN shape





Weights plane

Weight position

Channel:1

Row:2

Col:3

decoder →

FPPE parameter

P_{des_col}

P_{in_ifmap}

P_{in_fp}

P_{in_lppe}

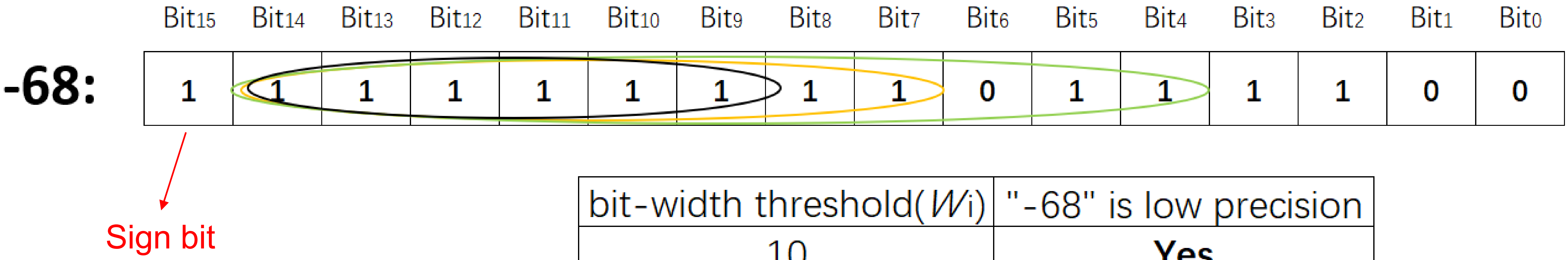
Arbiter Logic

Assume that a 16-bit signed fixed-point weight is fetched as $\{W_{15}, W_{14} \ldots W_1, W_0\}$

For a given bit-width threshold W_i :

low precision weights : $\{W_{14}, W_{13} \ldots W_{i-1}\} = (16 - i)\{W_{15}\}$

full precision weights : $\{W_{14}, W_{13} \ldots W_{i-1}\} \neq (16 - i)\{W_{15}\}$



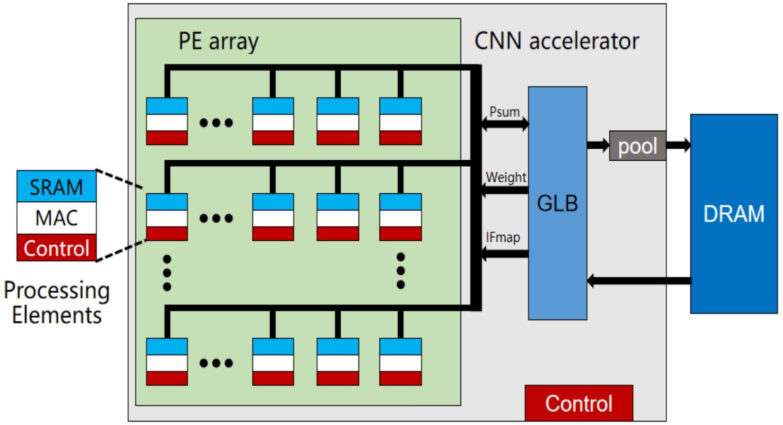
	Bit$_{15}$	Bit$_{14}$	Bit$_{13}$	Bit$_{12}$	Bit$_{11}$	Bit$_{10}$	Bit$_{9}$	Bit$_{8}$	Bit$_{7}$	Bit$_{6}$	Bit$_{5}$	Bit$_{4}$	Bit$_{3}$	Bit$_{2}$	Bit$_{1}$	Bit$_{0}$
-68:	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0

Sign bit

bit-width threshold(Wi)	"-68" is low precision
10	**Yes**
8	**Yes**
5	**No**

Comparison of Weight Storage Size

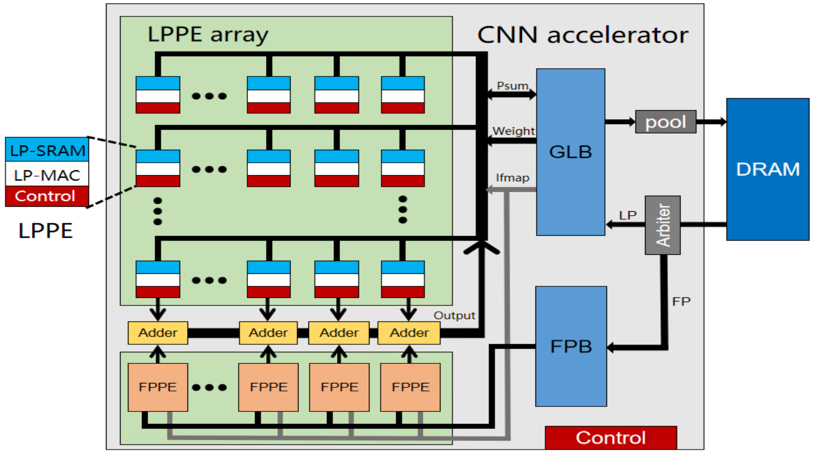
Weight storage for **single** precision architecture



$$S_{PE} = N \times W \times k$$
$$S_{GLB} = M \times W$$

VS

Weight storage for **mixed** precision architecture



Full precision part

$$S_{FPPE}^{'} = N \times k \times \left(1 - p(W_i)\right) \times W$$
$$S_{FPB}^{'} = M \times (1 - p(W_i)) \times W$$

Low precision part

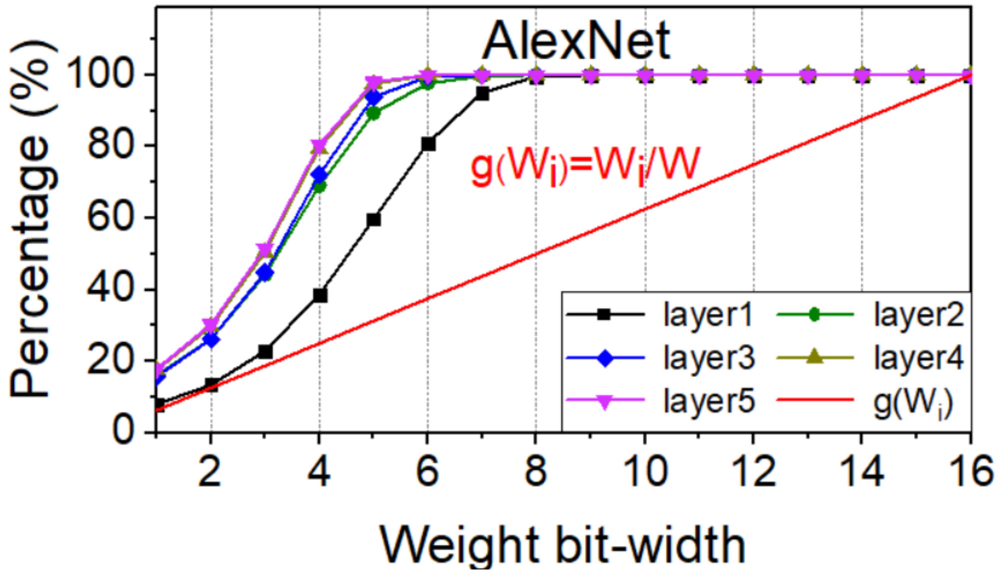
$$S_{LPPE}^{'} = N \times k \times W_i$$
$$S_{GLB}^{'} = M \times W_i$$

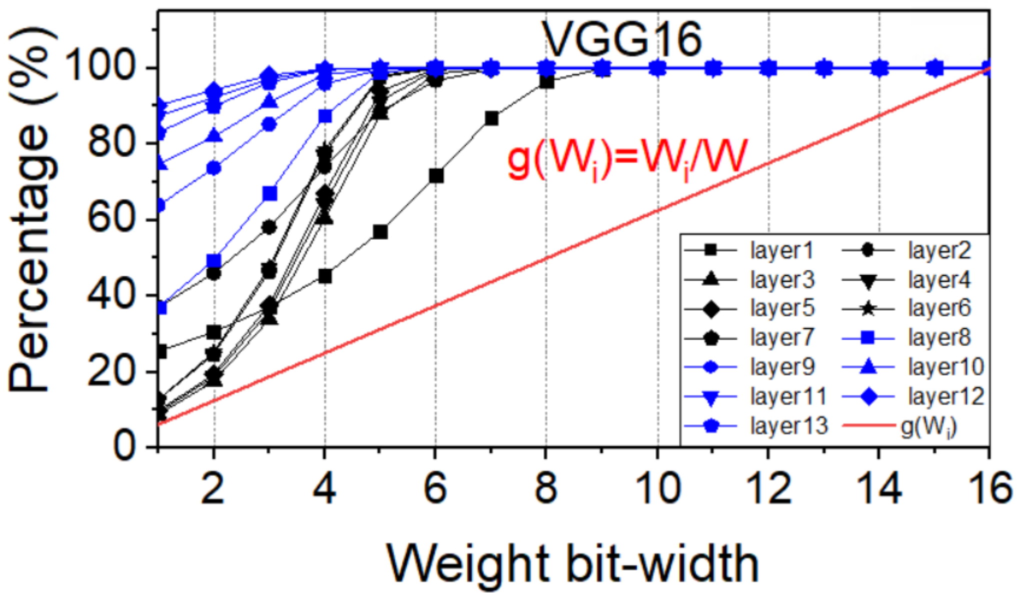
Bit-Width Threshold Selection

Relative storage saving for PE (or GLB) can be approximately calculated by above EQs, and simplified to:

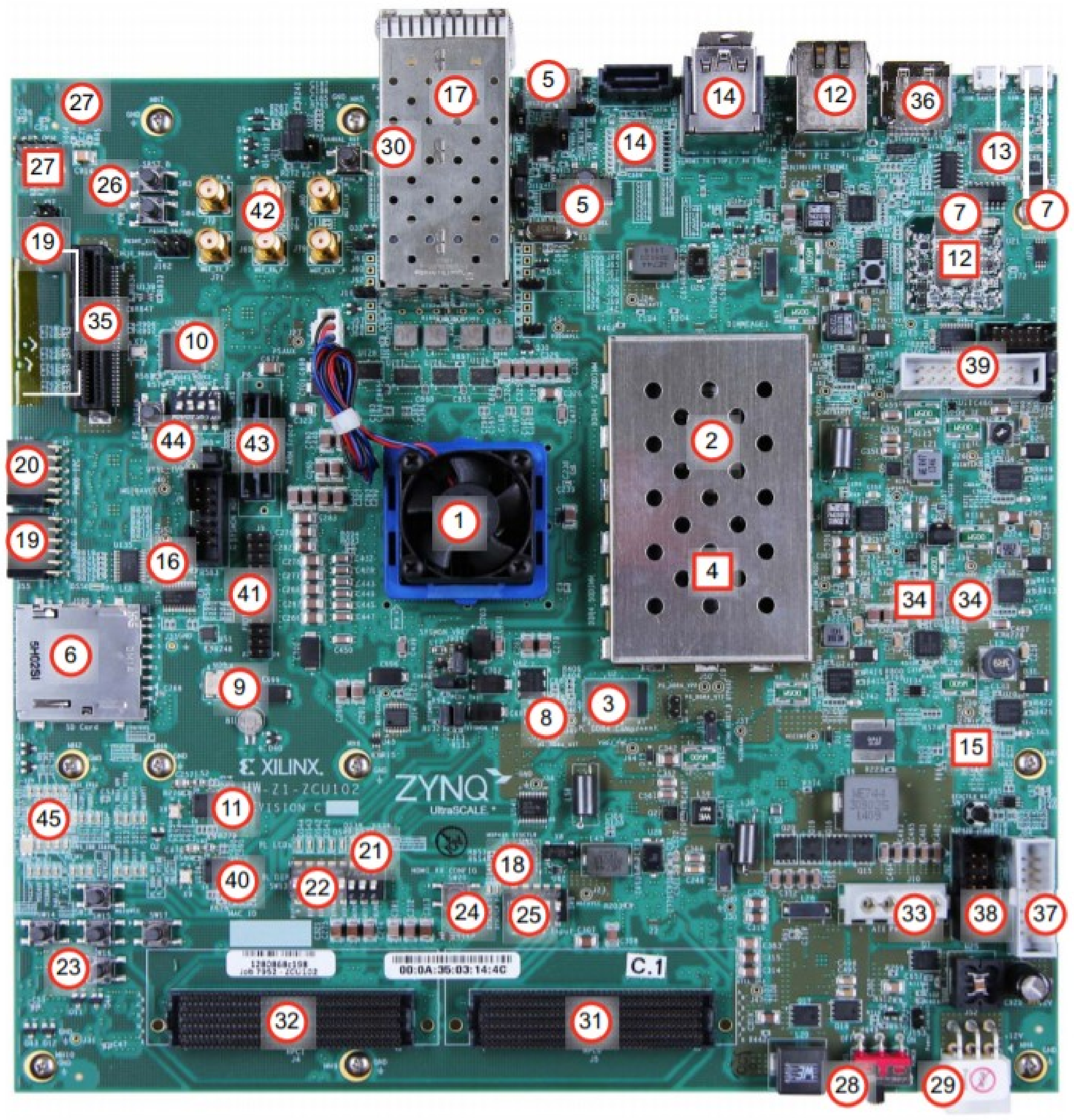
Area saving rate of PE: $\alpha \approx \frac{s_{PE}-s_{PE}^{'}-s_{FP\,PE}^{'}}{s_{PE}} = p(W_i) - \frac{W_i}{W}$

Area saving rate of GLB: $\beta \approx \frac{s_{GLB}-s_{GLB}^{'}-s_{FPB}^{'}}{s_{GLB}} = p(W_i) - \frac{W_i}{W} = \alpha$





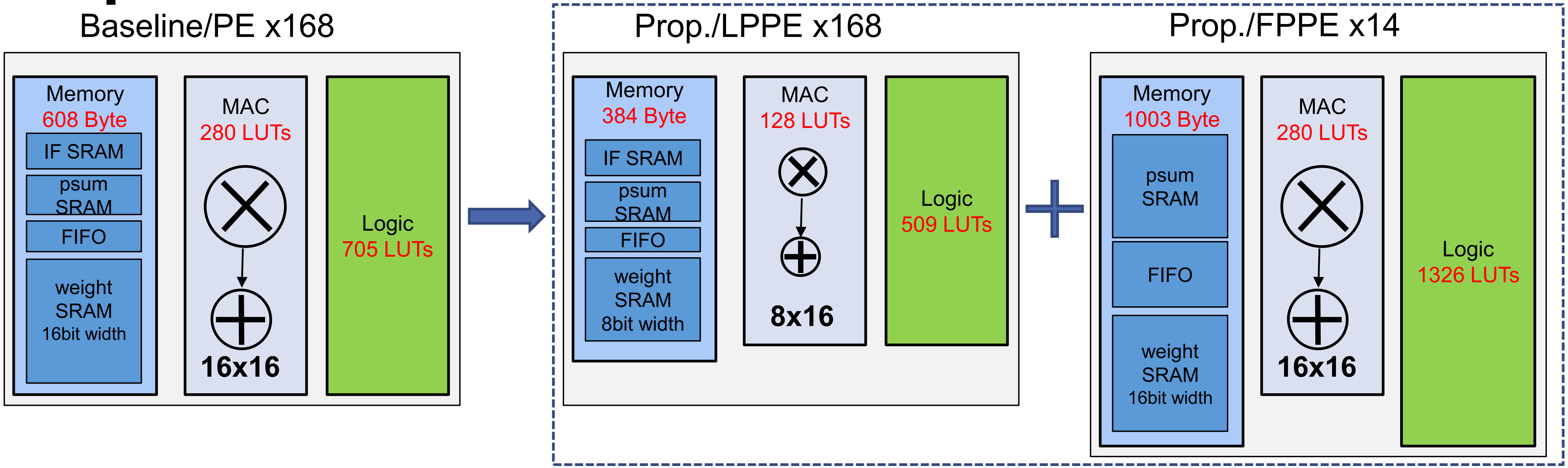
Experimental Setup



-Low power
-Data reuse
-Reconfigurable

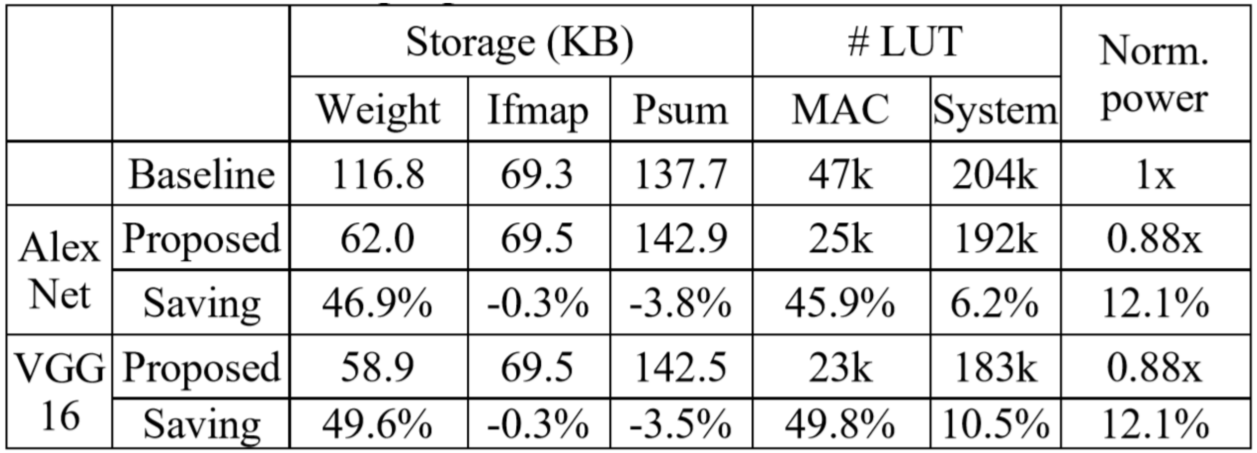
Baseline Accelerator	
Board	ZCU102
Dataflow	Row Stationary
Arithmetic Precision	16-bit fixed point
Clock Rate	200Mhz
GLB storage size	224KB
#PE in a row	14
#PE in a col	12
PE storage size	608B
#LUT used for MAC	280
#LUT used for PE	1313

Experimental Result——Area Reduction

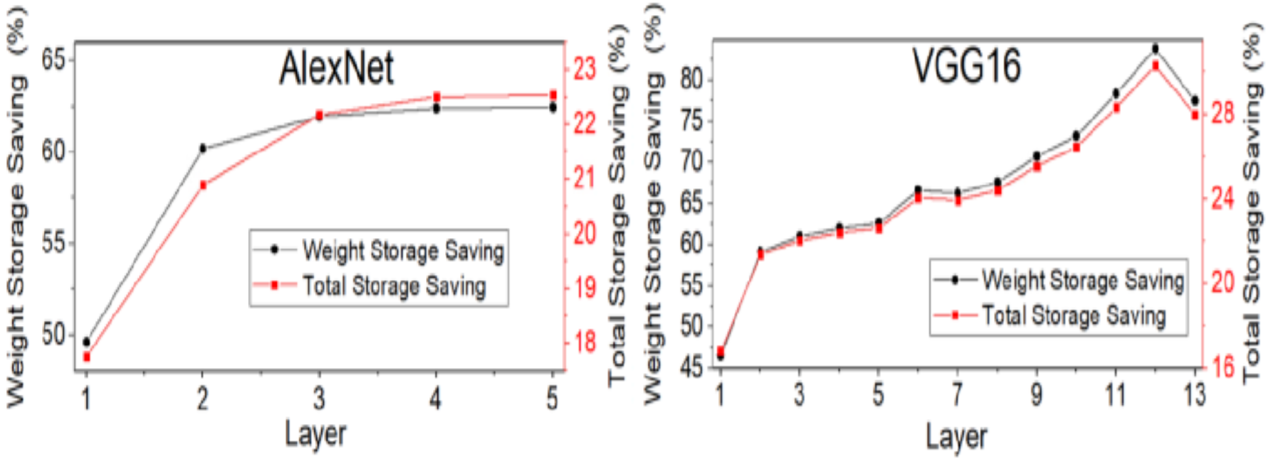


		Baseline/PE	Prop./LPPE	Prop./FPPE
Number		168	168	14
Storage(Byte)		608	384	1003
#LUTs(MAC)		280	128	280
#LUTs(PE)		1313	777	1606
Total	Storage(Byte)	102,144(1x)	78,554(0.77x)	
	#LUTs	220,584(1x)	153,020(0.69x)	

Experimental Result



		Storage (KB)			# LUT		Norm. power
		Weight	Ifmap	Psum	MAC	System	
	Baseline	116.8	69.3	137.7	47k	204k	1x
Alex Net	Proposed	62.0	69.5	142.9	25k	192k	0.88x
	Saving	46.9%	-0.3%	-3.8%	45.9%	6.2%	12.1%
VGG 16	Proposed	58.9	69.5	142.5	23k	183k	0.88x
	Saving	49.6%	-0.3%	-3.5%	49.8%	10.5%	12.1%



(a)

(b)

- Weight storage area reduced by nearly 50%
- Number of LUTs for calculation is reduced by nearly 50%
- Number of LUTs in the system is reduced by 6% to 10%
- Save about 12% power than baseline
- Actual total storage saving in AlexNet and VGG16 is almost 17.8% and 16.8%

Conclusion

- Proposed architecture of PE to simultaneously store and calculate with different employ two separate groups precisions
- Implement the proposed CNN accelerator using an FPGA platform
- Weight storage area in PE and GLB reduced by nearly 50%
- Total storage area reduced by almost 17.8%
- Critical path delay is reduced by almost 28%
- Dynamic power saving by 12.1% without timing penalty

Q & A