FPL2019
@Barcelona

Tokyo Tech

FPGA-based Training Accelerator Utilizing Sparseness of Convolutional Neural Network

Hiroki Nakahara, Youki Sada, Masayuki Shimoda,
Akira Jinguji, Shimpei Sato
Tokyo Institute of Technology, JP

Challenges in DL Training

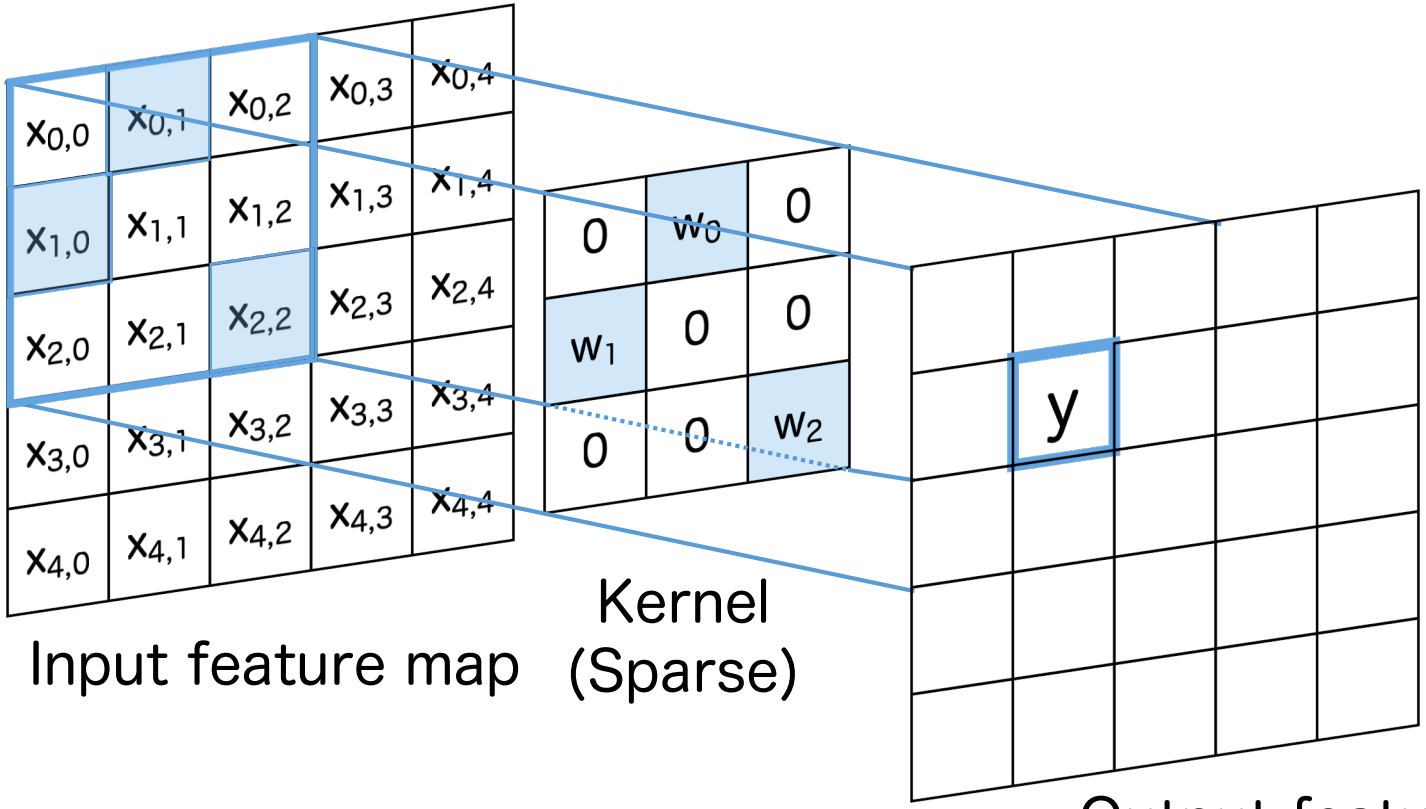
- High-speed
  - 1.2 min@2,048 GPUs
- Low-power consumption

Training with ResNet-50 on ImageNet

	Processor	DL Library	Time	Accuracy
He et al. [1]	Tesla P100 × 8	Caffe	29 hours	75.3 %
Goyal et al. [2]	Tesla P100 × 256	Caffe2	1 hour	76.3 %
Smith et al. [3]	full TPU Pod	TensorFlow	30 mins	76.1 %
Akiba et al. [4]	Tesla P100 × 1,024	Chainer	15 mins	74.9 %
Jia et al. [5]	Tesla P40 × 2,048	TensorFlow	6.6 mins	75.8 %
Ying et al. [6]	TPU v3 × 1,024	TensorFlow	1.8 mins	75.2 %
Mikami et al. [7]	Tesla V100 × 3,456	NNL	2.0 mins	75.29 %
This work	**Tesla V100 × 2,048**	**MXNet**	**1.2 mins**	**75.08%**

TSUBAME-KFC (TSUBAME Kepler Fluid Cooling)

Sparse Weight Convolution



Input feature map

Kernel (Sparse)

Output feature map

skip

$X_{0,1} \times W_0$

skip

$X_{1,0} \times W_1$

skip

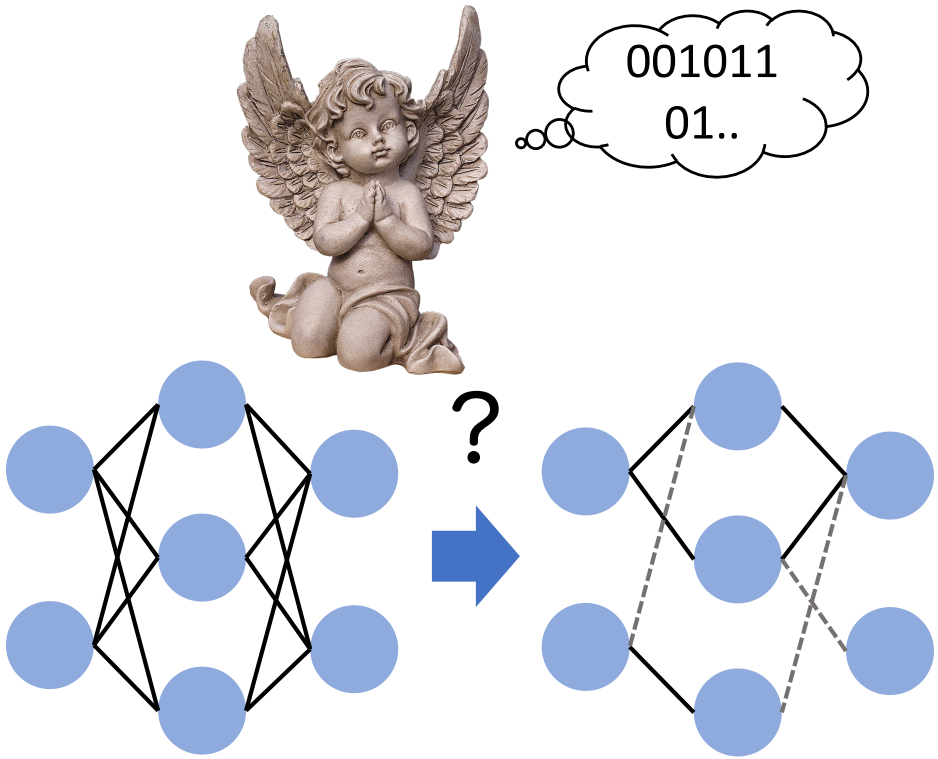
+) $X_{2,2} \times W_2$

y

ρ: Threshold

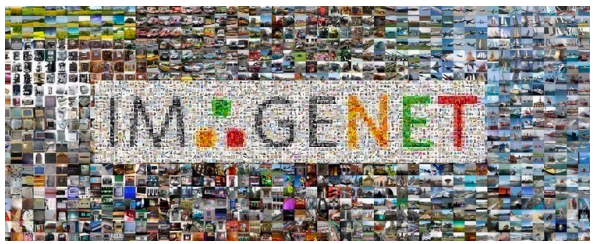
Training of Sparseness CNN

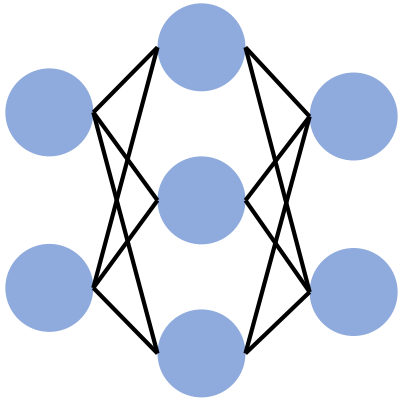
- Initial weight
  - Lottery ticket assumption
- Special hardware



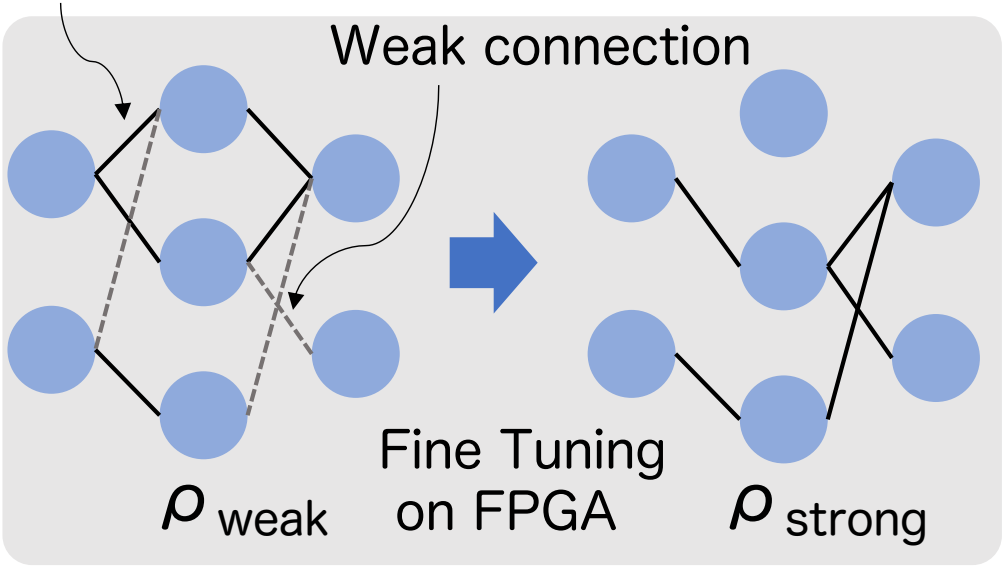
Fine-Tuning for a Sparse CNN

- Use pre-trained model (sparse weight) by ImageNet
  - Retain strong connection for recognition accuracy



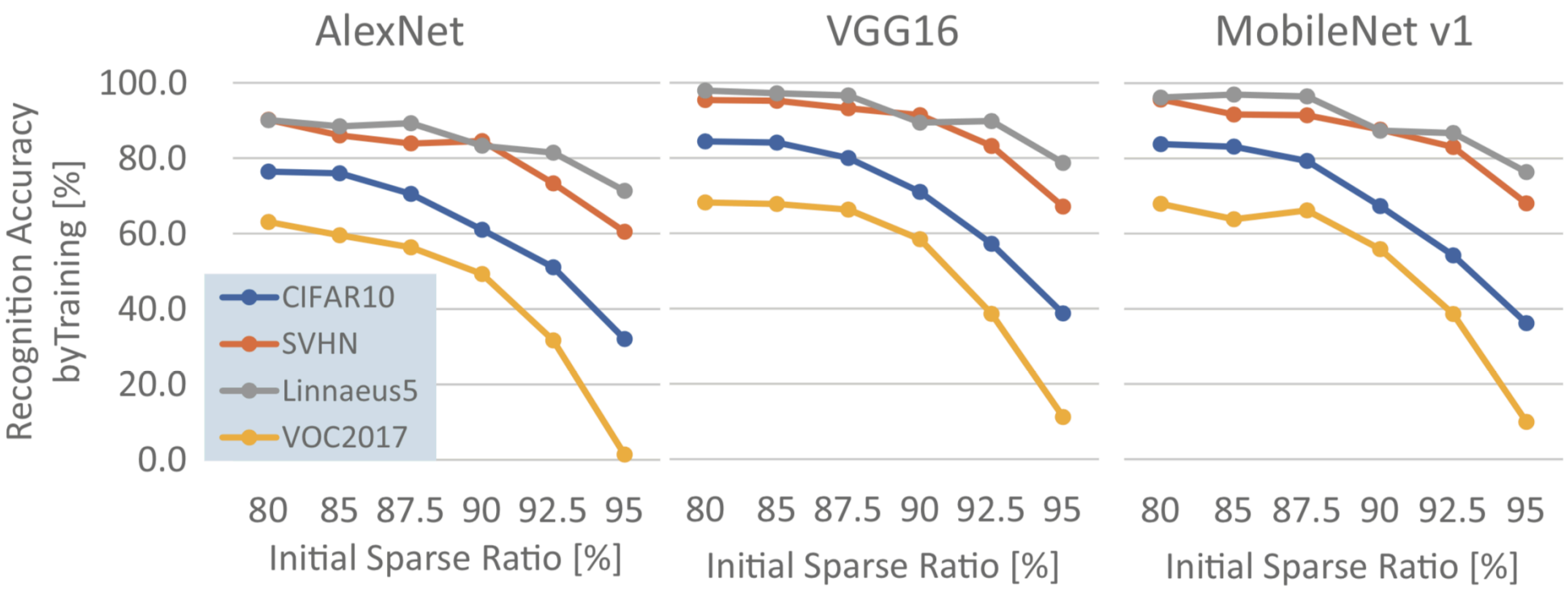


Dense CNN

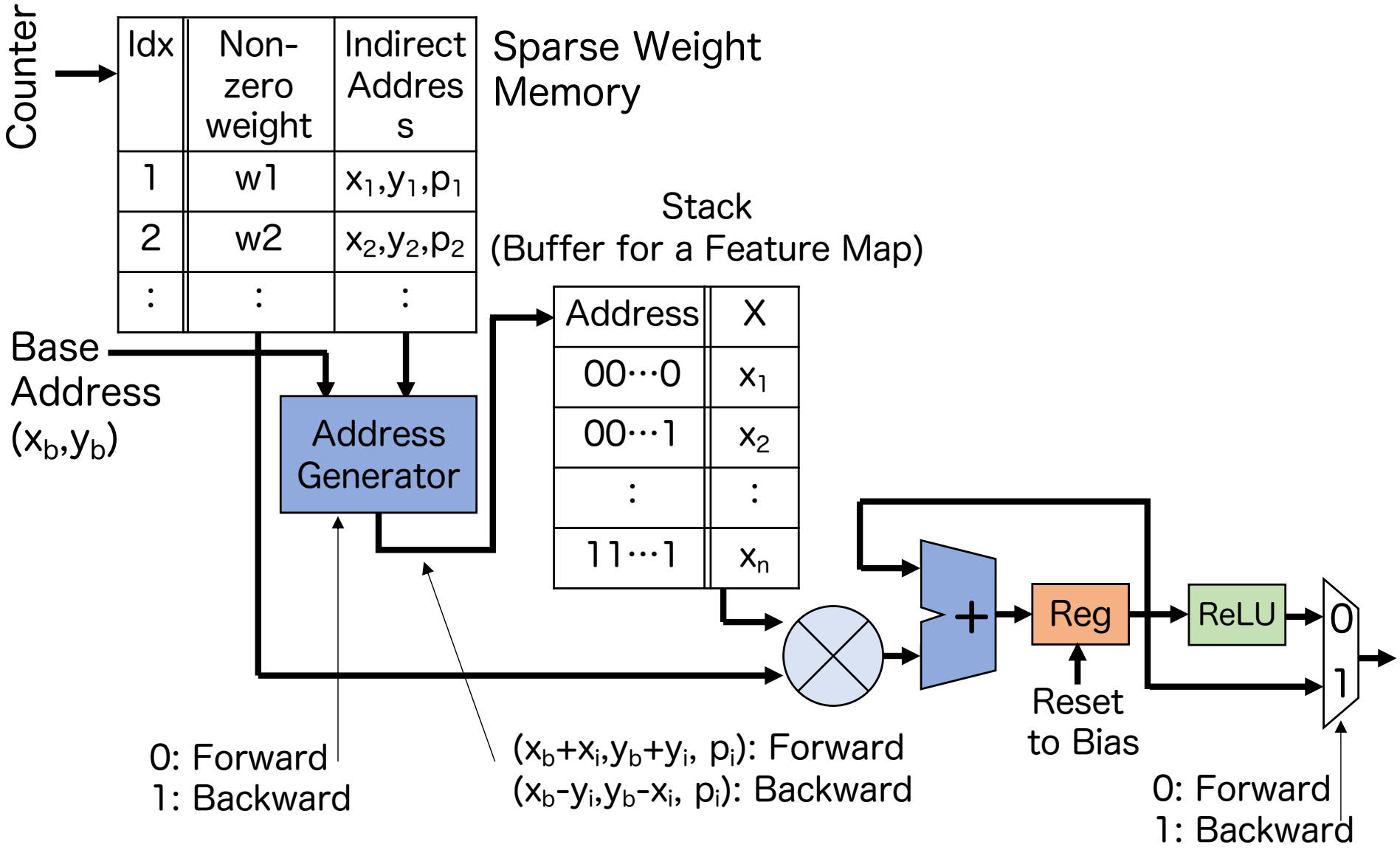


Sparseness vs. Accuracy

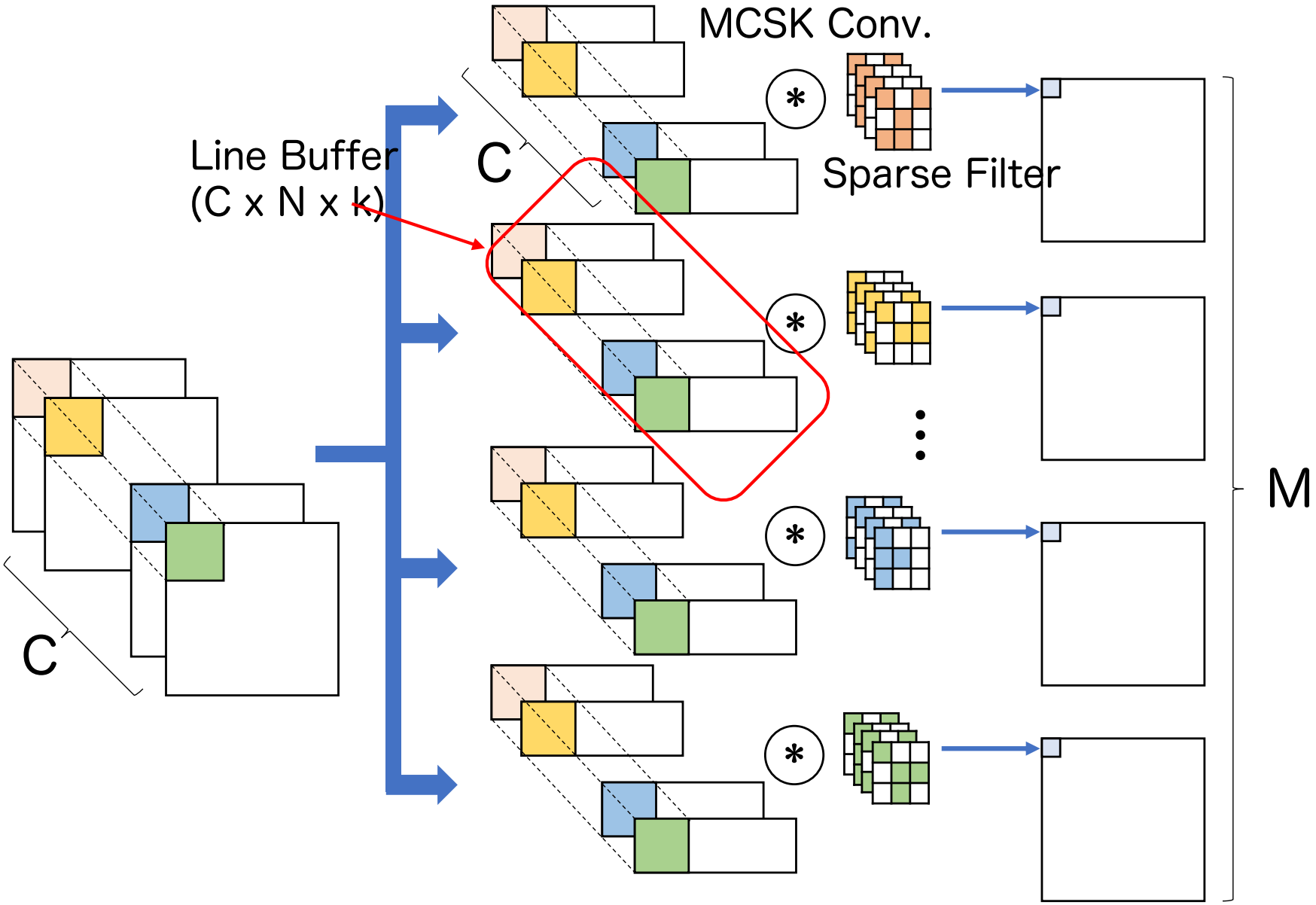
- 85% of weight can be pruned initially



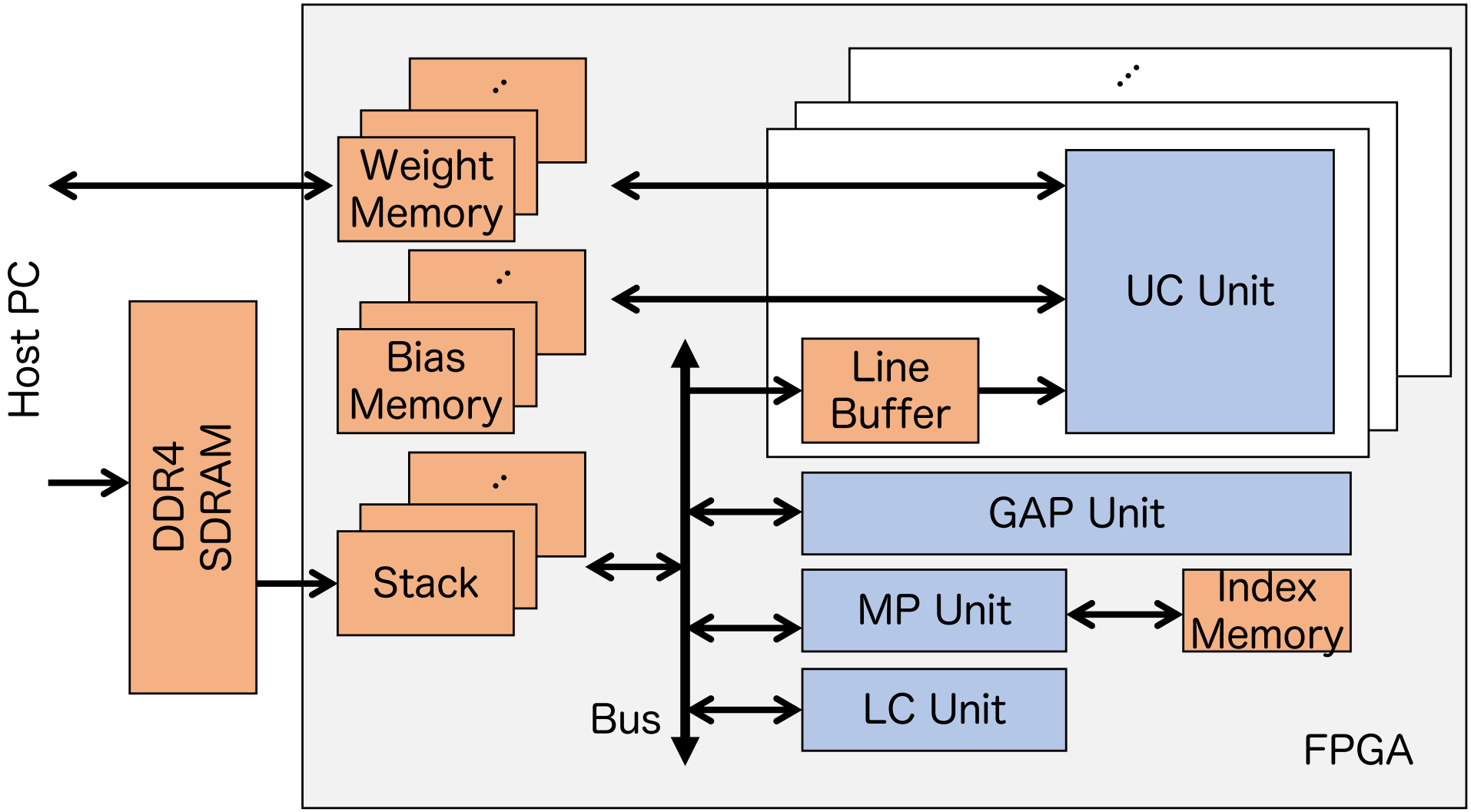
Universal Convolution (UC) Unit



Parallel MCSK Convolution



Overall Architecture



Results

Resource Consumption

Module	LUTs	FFs	BRAMs	URAMs	DSPs
Total	934,381	370,299	3,806	960	1,106

Training Time (Batch size=32, Epoch=100)

Dataset	CNN	Sparse Ratio [%]	GPU [sec]	FPGA [sec]
CIFAR-10	AlexNet	92.1	2,548	615
SVHN	AlexNet	91.0	3,672	875
Linnaeus5	AlexNet	93.7	1,482	372
VOC2017	AlexNet	94.3	2,697	680
CIFAR-10	VGG16	93.4	4,178	1,025
SVHN	VGG16	95.4	6,121	1,435
Linnaeus5	VGG16	93.3	2,430	612
VOC2017	VGG16	92.5	4,458	1,098
CIFAR-10	MobileNetv1	89.2	8,352	2,052
SVHN	MobileNetv1	89.8	12,058	2,871
Linnaeus5	MobileNetv1	90.1	4,902	1,223
VOC2017	MobileNetv1	88.3	8,944	2,184

FPGA: VCU1525
1,182K LUTs
2,364K FFs
6,840 DSPs
4,216 BRAMs
960 URAMs

GPU: RTX2080Ti