CNNs on Multi-Core 2-D Systolic Array with Pruning and Huffman Coding

UC San Diego

284 little group

Zhen Bian, Hanyi Chen, Mingyu Liu, Cheng Qian, Xin Zhao

VGGnet with quantization-aware training

	VGG16
Accuracy(CIFAR 10)	92.140%
Quantization error	0.0006

Mapping on FPGA (Cyclone IV GX)

OPs	128
Frequency	132.82 MHz
Dynamic Power	35.72mW
GOPs/s	17
GOPs/w	0.00358
Logic Elements	22126

Fig. 1) IP Hierarchy

Alpha 1. Pruning on Hardware

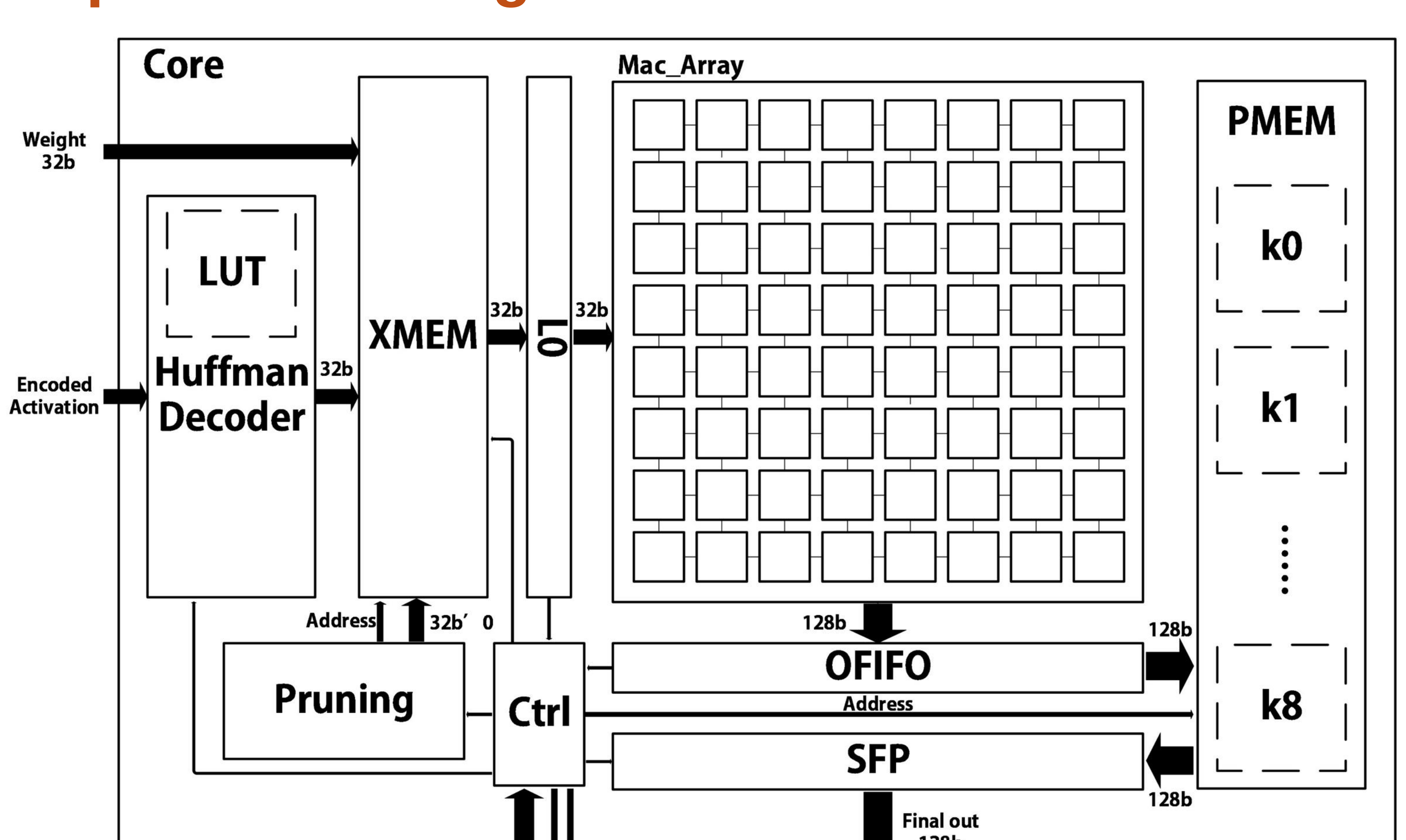

Fig. 2) Circuit

Alpha 2. Huffman Coding

	Huffman coding
Before compression	1152 bits
After compression	422 bits
Compression rate	0.3663

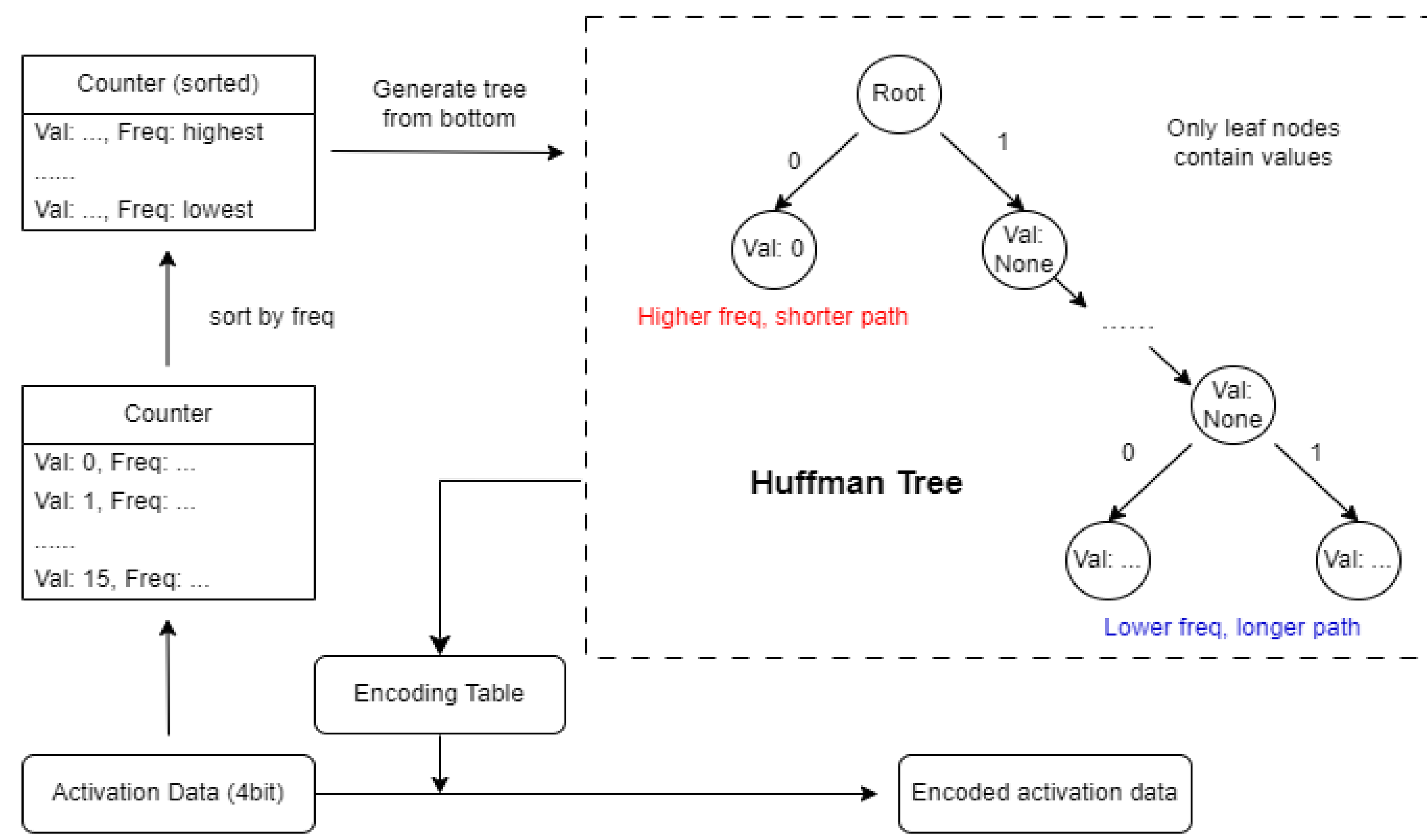

Fig. 3) Procedure of Huffman Coding

Alpha 3. Resnet with quantization-aware training

	Resnet20	VGG16
Accuracy	89.080%	92.140%
Quantization error	0.0398	0.0006

- Modified Resnet has **less accuracy**
- Modified Resnet has **larger psum recover loss**

Alpha 4. Pruning on VGG16

	Accuracy
80% sparsity, unstructured, only mapped layer	**89%**
80% sparsity, unstructured, all layers	**88%**
40% sparsity, structured, only mapped layer	**73%**

- Unstructured pruning can achieve **higher sparsity**
- Sparsity has a significant impact **on the layer we mapped**

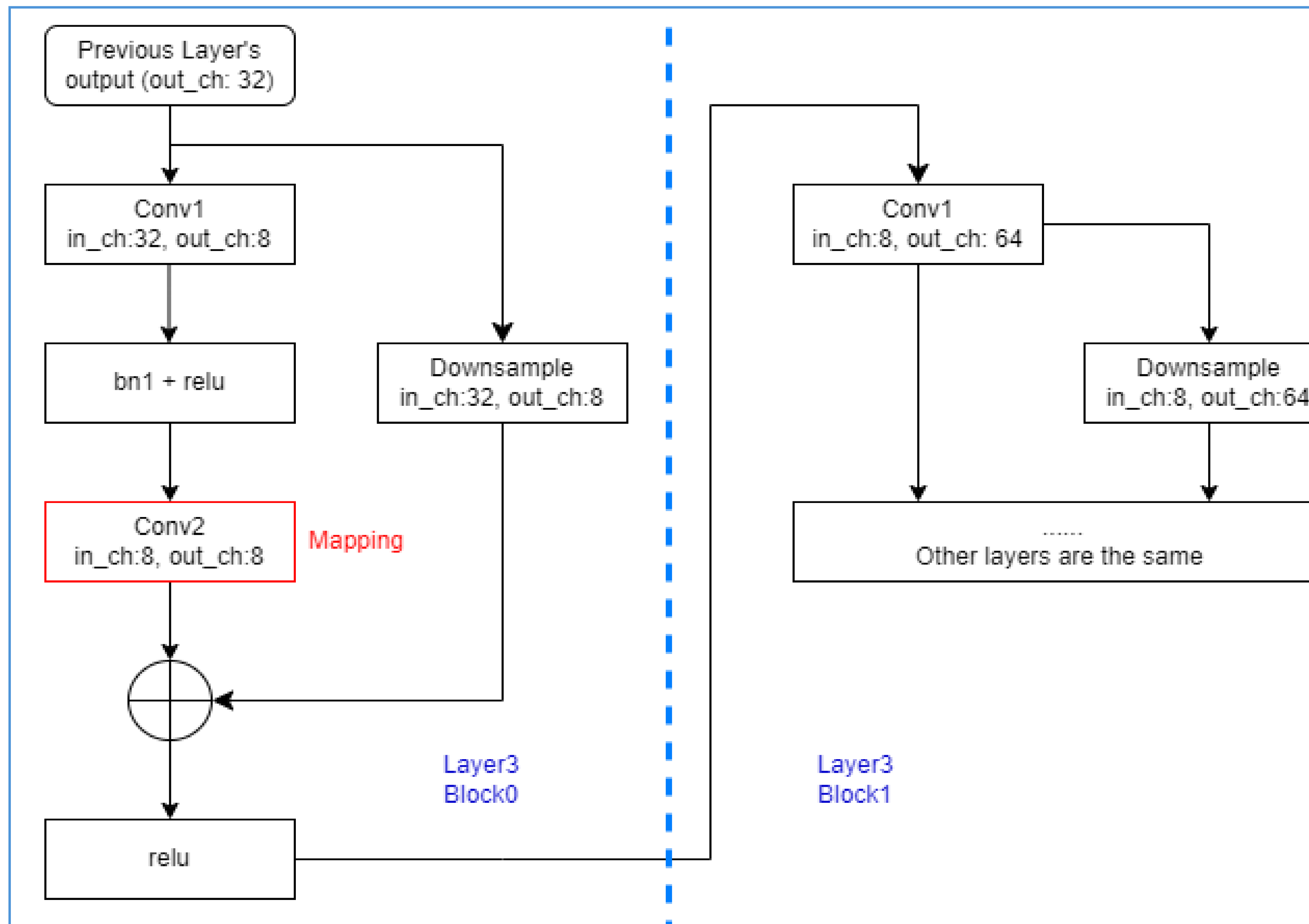

Fig. 4) Modified Resnet Model

Alpha 5. Multi-Core Processor

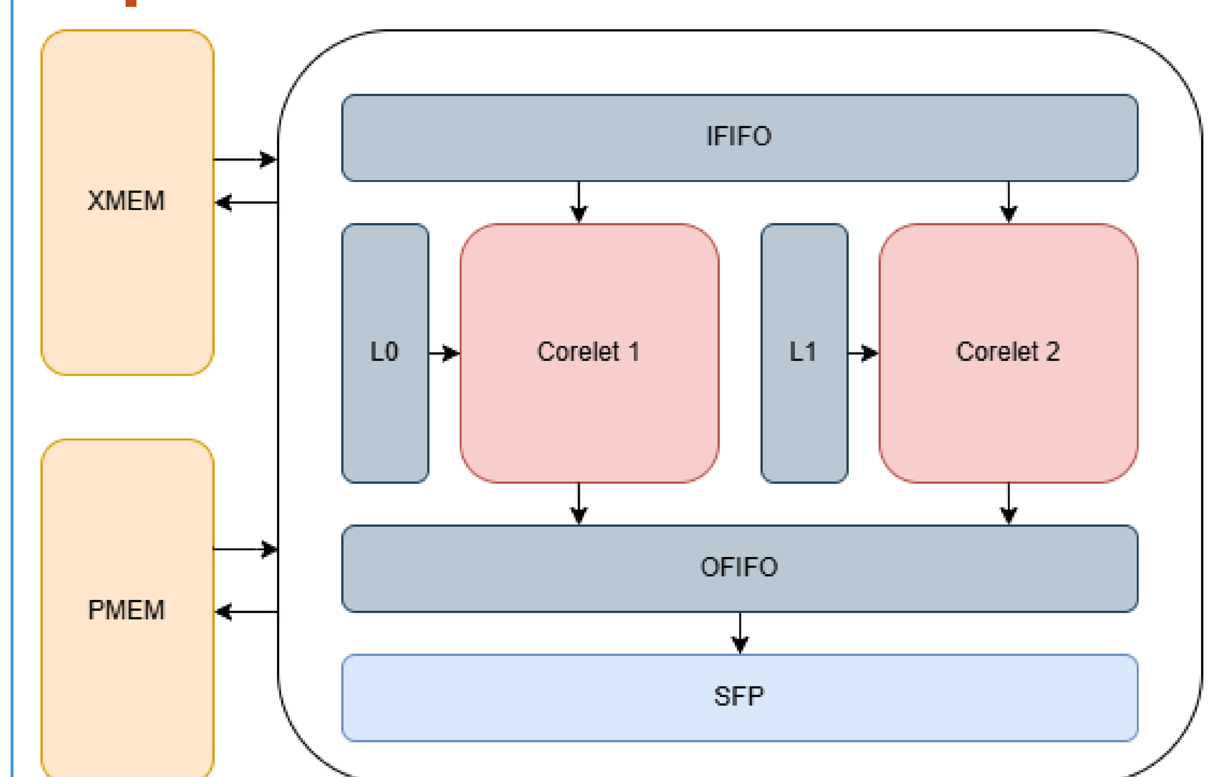

Fig. 5) Multi-Core 2D Systolic Array Structure

- **Less** Latency
- **Higher** Freq
- **Smaller** MEM

Alpha 6. Accumulation while Execution

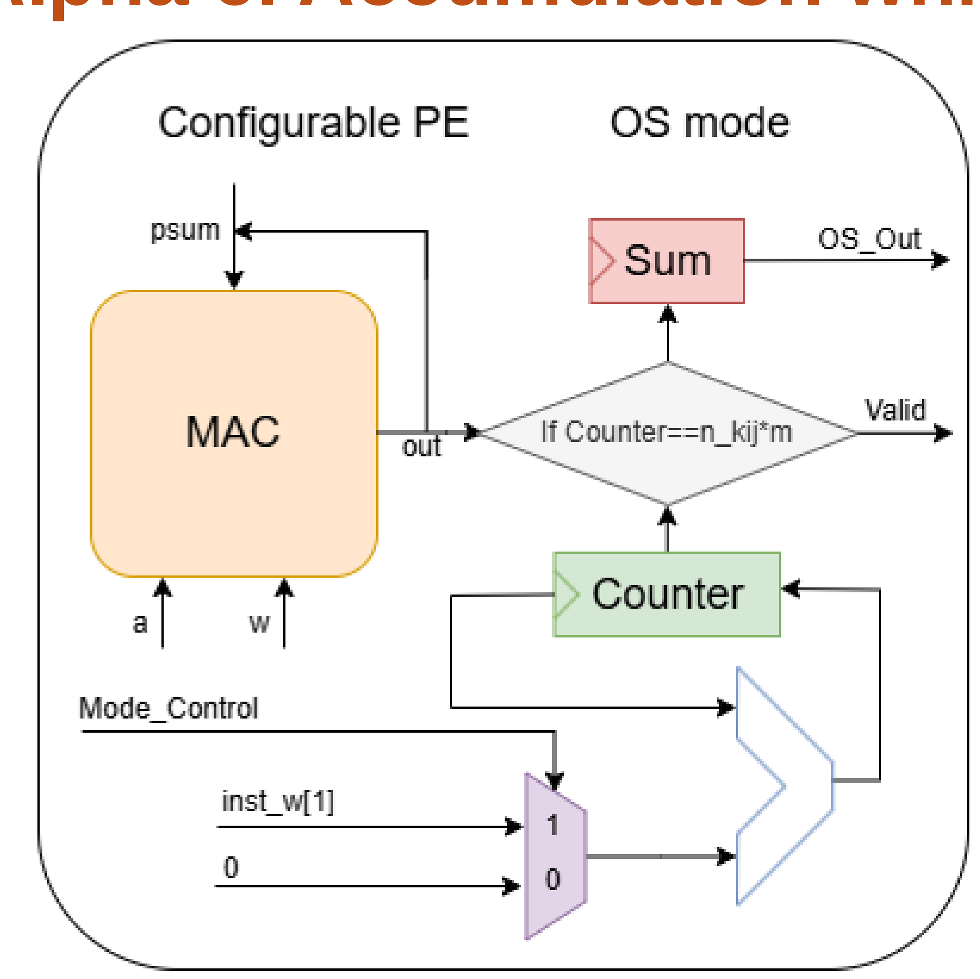

Fig. 6) Configurable PE Structure

- Be re-designing the PE, SFP structure can be **Saved**
- Reducing the chance to access MEM, **Less** Latency, **Higher** Freq