Multi-Precision Deep Neural Network Accelerator on FPGAs

Negar Neda[1], Salim Ullah[2], Azam Ghanbari[1], Hoda Mahdiani[1], Mehdi Modarressi[1], and Akash Kumar[2]

[1]University of Tehran, Tehran, Iran
[2]Dresden University of Technology, Dresden, Germany

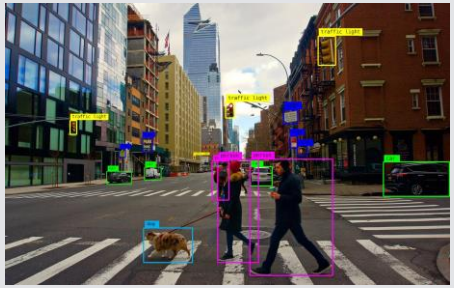
ASP DAC - 2022

Outline

- Introduction
- Proposed Architecture
- Experimental Results
- Conclusion

Introduction

Deep Learning



Deep Neural Networks

- **Features:**
  - Robust against approximation
  - Different minimum required bit-width across variant neural networks

Quantization	AlexNet Accuracy
16-bit fixed point	76.12
8-bit fixed point	76.10
4-bit fixed point	54.45
4-bit layer 3,4; 8-bit the rest	75.91

Motivation

- **Need for multi-precision design:**
  - Adjust the precision of the computation to each DNN layer
  - Multi-tenant accelerators
    - Adjust the precision of the accelerator to the model
  - Accuracy-energy trade-off
    - Temporarily give up accuracy to meet energy constraints
  - Training support
    - Support training and inference simultaneously

Motivation

- **Why FPGA?**
  - Flexible platform to support future Neural Networks
  - Inefficiency of ASIC designs on FPGAs
  - Fast time to market
  - LUT-based design:
    - Almost 5% of the FPGA area are DSP blocks
    - More than 60% of the FPGA are LUTs

Multi-precision Accelerator

- Reduce computation bit-width with negligible accuracy loss
- Design an FPGA-based multi-precision multiplier with adaptable data bit-width at run-time
  - Support of 16, 8, and 4-bit computation
  - Reduce power consumption
  - Increase throughput

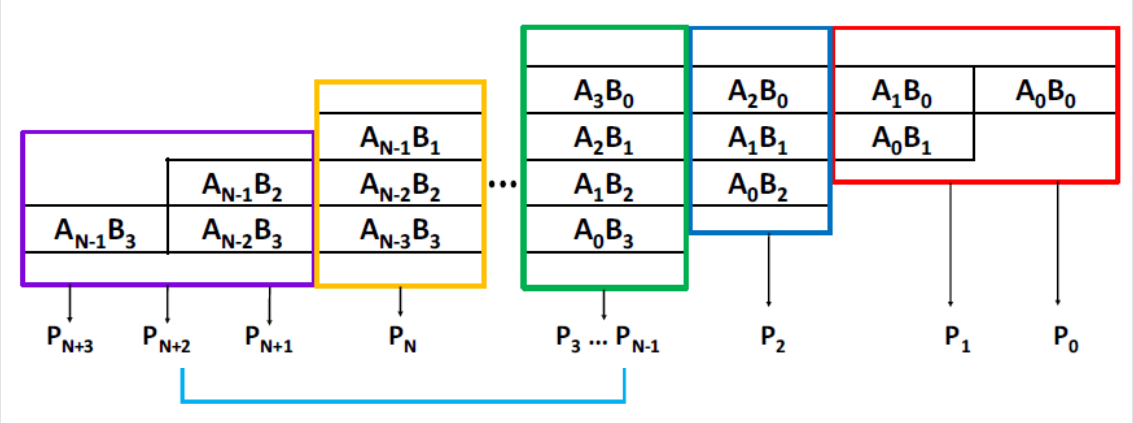
Proposed Achitecture

Proposed Arch.: 4-bit Mult

- 4×4-bit Multiplier as the base of fusible design
  - Recursively build 8 × 8 and 16 × 16 multiplier
- Use two 4 × 4 multipliers based on FPGA LUT
  - Area Optimized

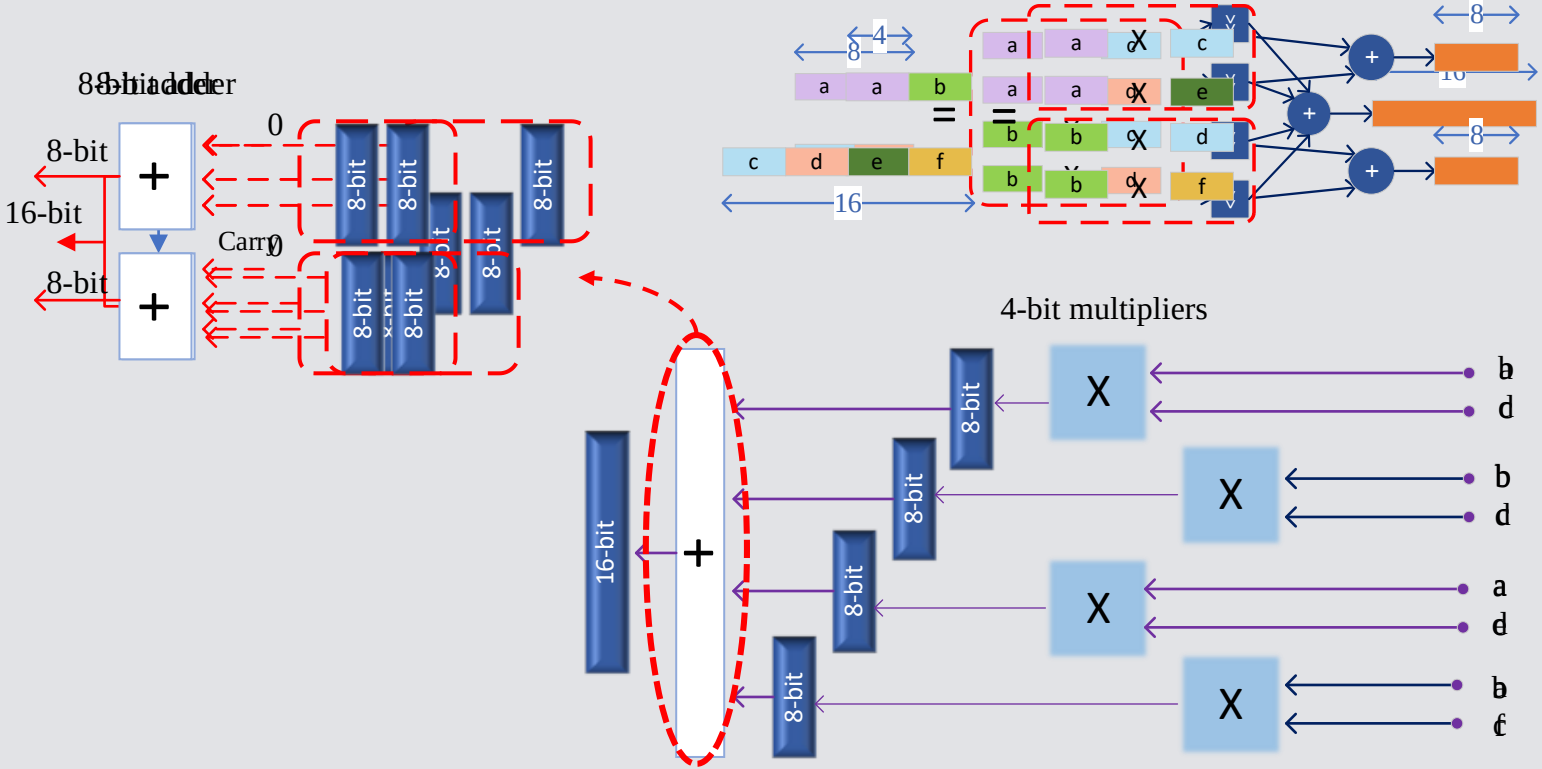
    (0.24% accuracy loss)
  - Approx3

    (12.33% accuracy loss)

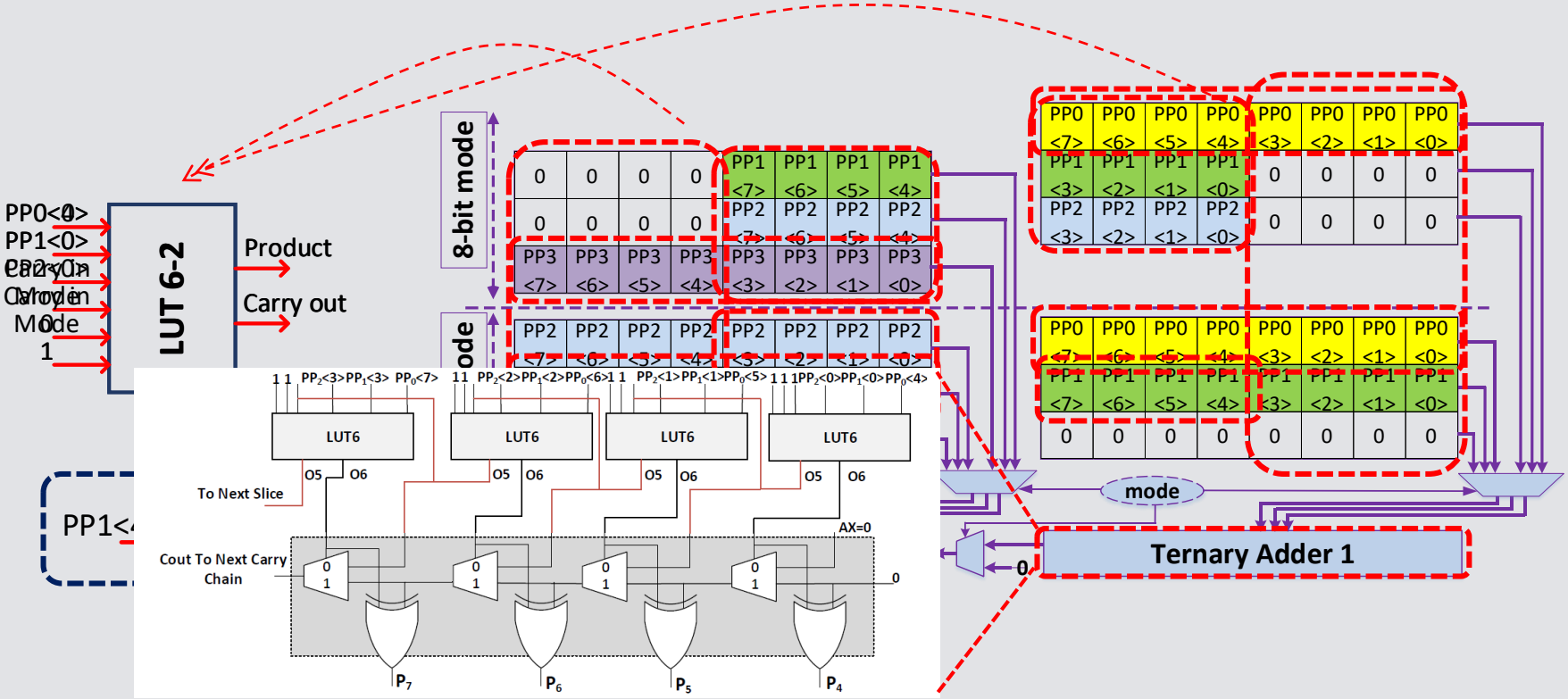


[S. Ullah, DAC 2018]

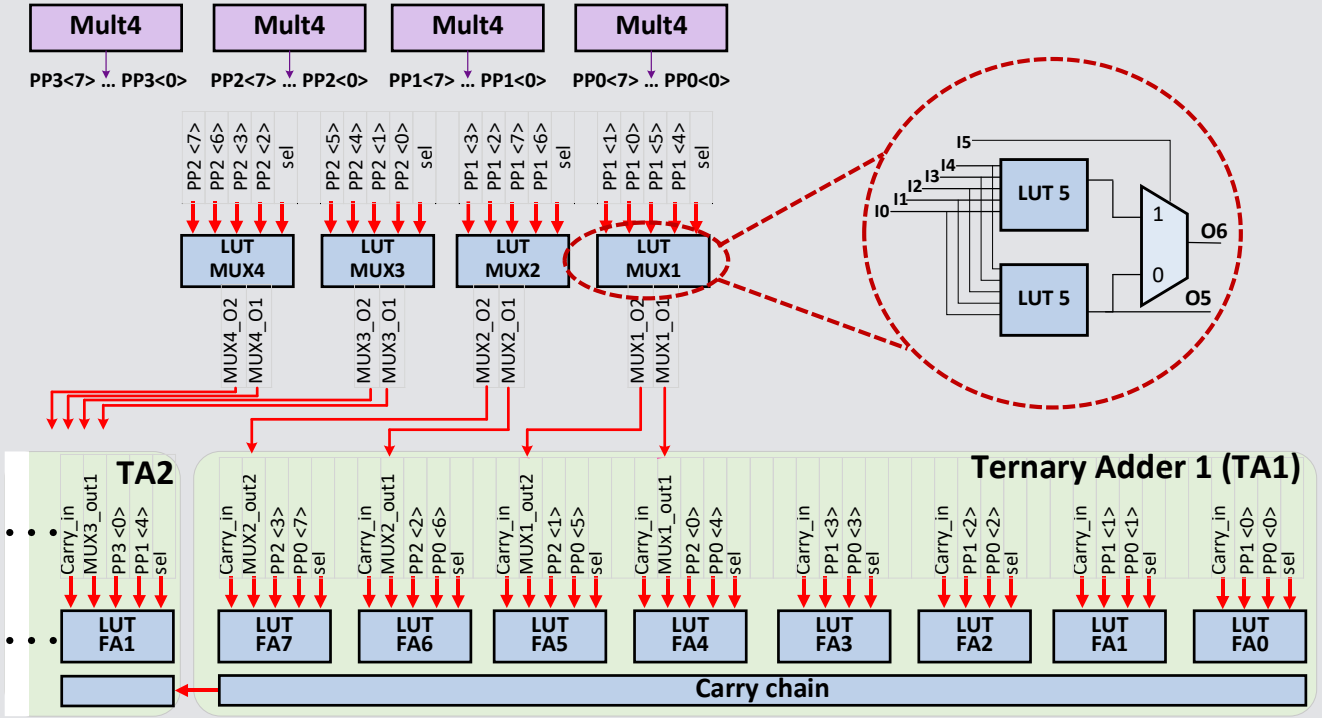
Proposed Arch.: 8 & 4-bit



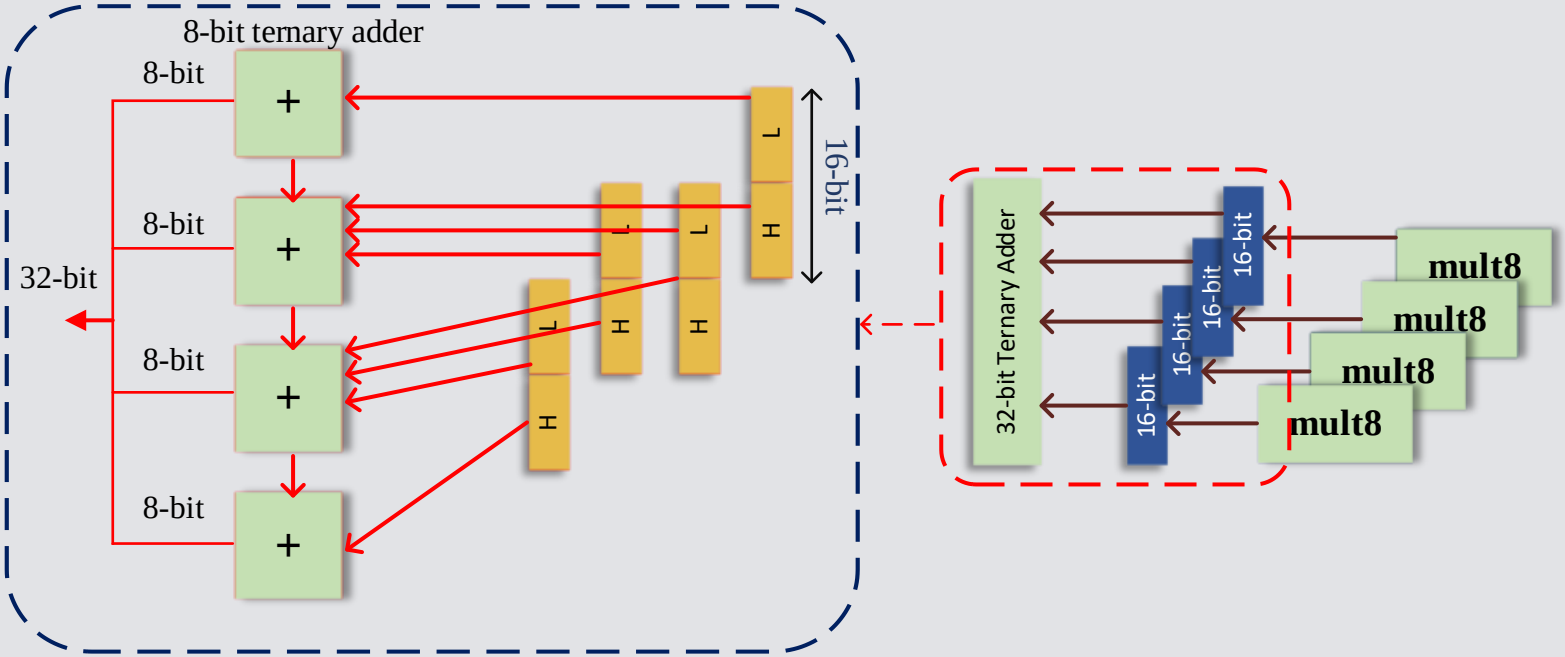
Proposed Arch.: Level One Adder



Proposed Arch.: Level One Adder



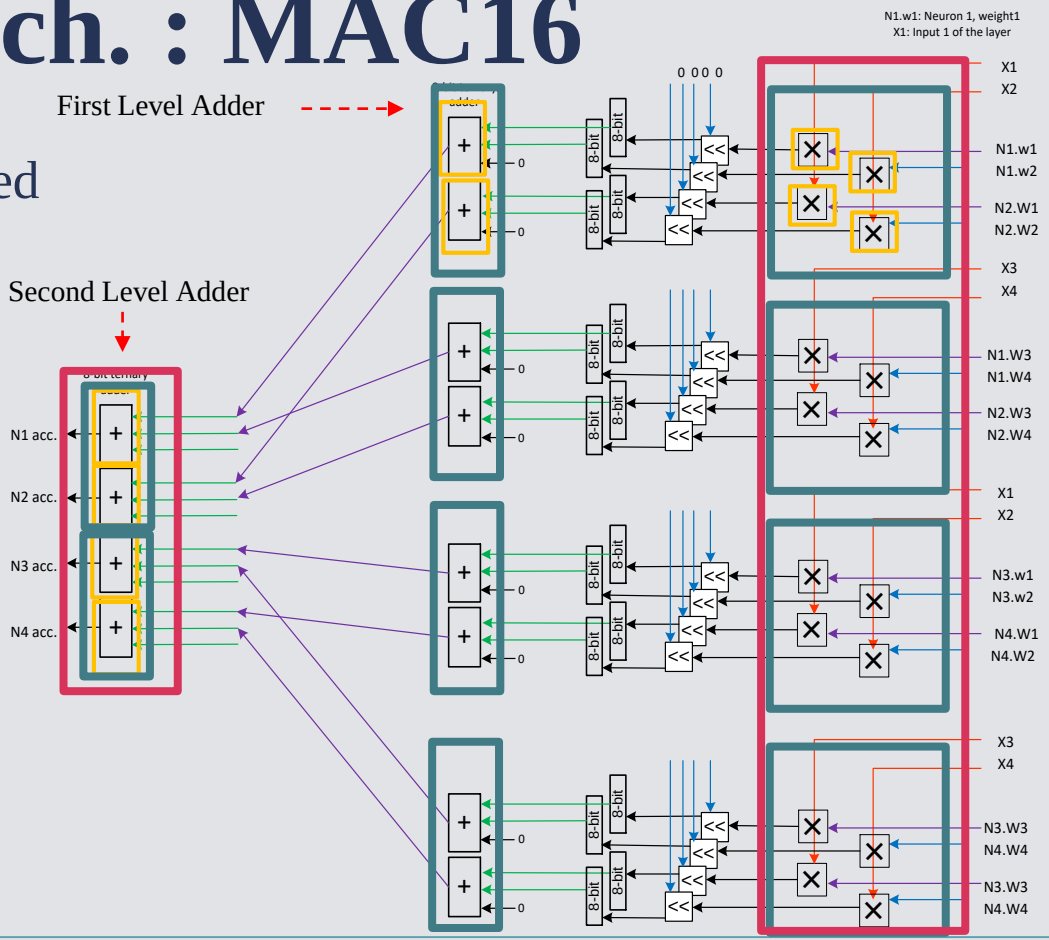
Proposed Arch.: 16-bit



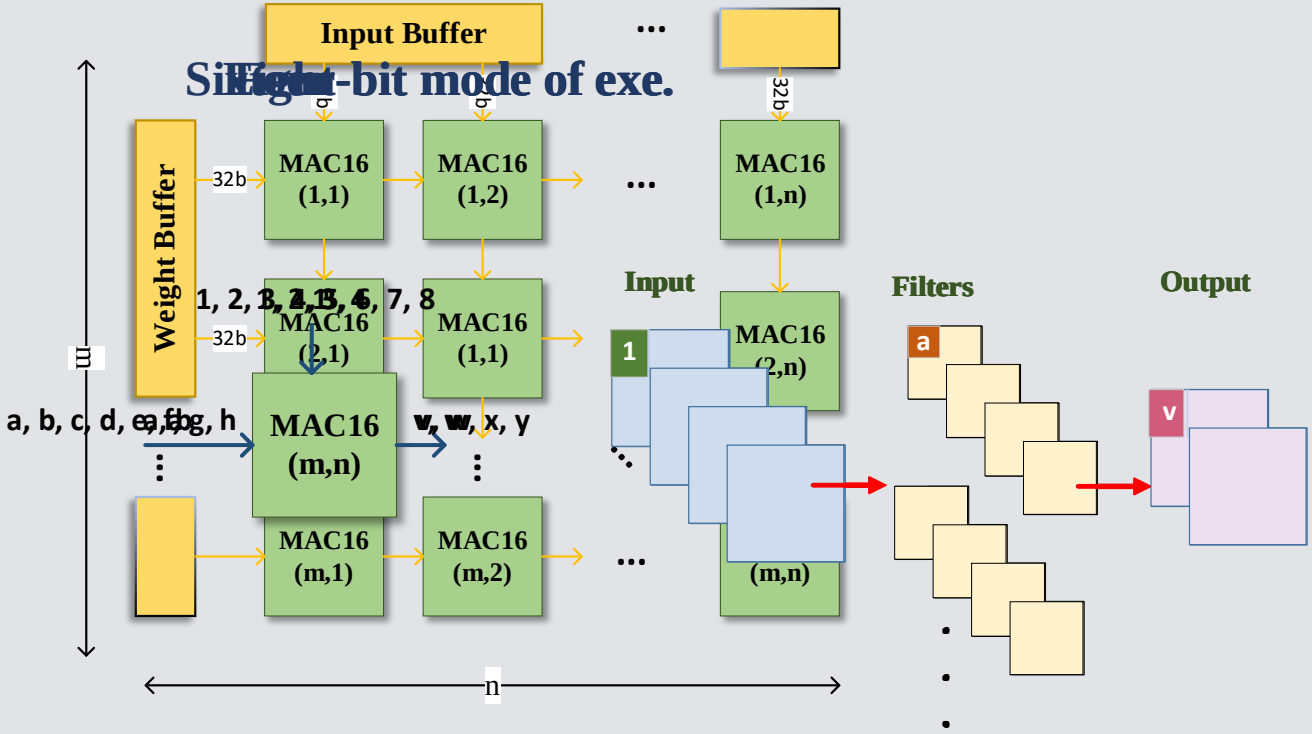
Proposed Arch. : MAC16

Fusible multiplier controlled by a control signal

- Sixteen 4 × 4
  - Calculate four neurons
- Four 8 × 8
  - Calculate two neurons
- One 16 × 16
  - Calculate one neuron



Proposed Arch.: MpDNN



Evaluation

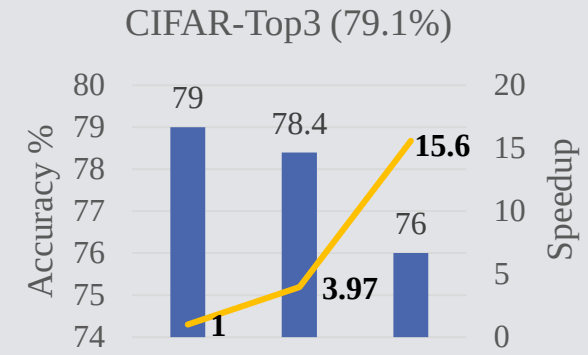
Evaluation

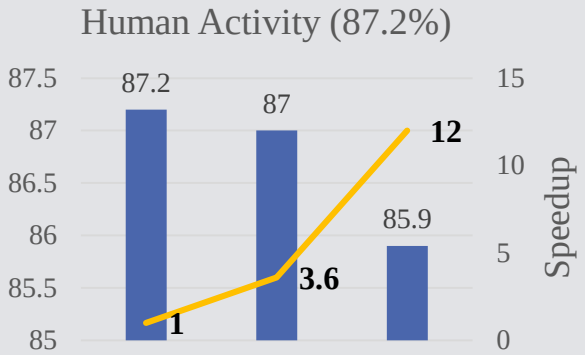
- Implemented the design on Zynq7Z020
- Used MLP & CNNs to evaluate the accuracy and throughput of MpDNN
- MLP
  - CIFAR-Top3, IRIS, Heart Disease, Human Activity
- CNN
  - LeNet (Mnist dataset), AlexNet (ImageNet dataset)
- Two 4-bit multipliers as the base multiplier
  - Area Optimized
  - Approx3

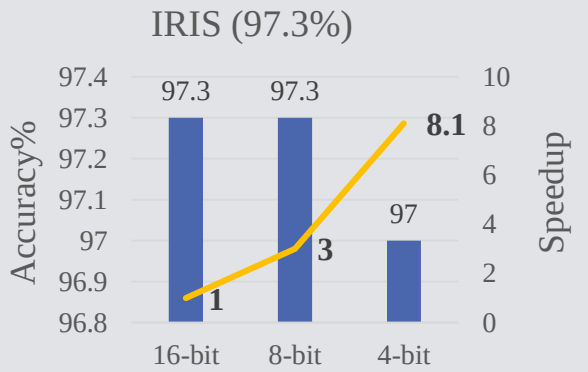
Multipliers Analysis

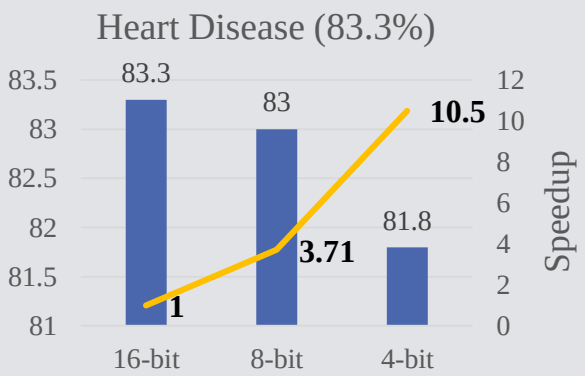
Multipliers	#LUT		Delay (ns)	
	16-bit	8-bit	16-bit	8-bit
mpDNN_AO	324	74	11.55	7.27
mpDNN_Approx3	261	59	10.1	6.7
Area Optimized	251	65	9.54	6.18
Approx3	195	49	8.59	5.6
Xilinx-IP	293	72	7.9	6

MpDDN Analysis

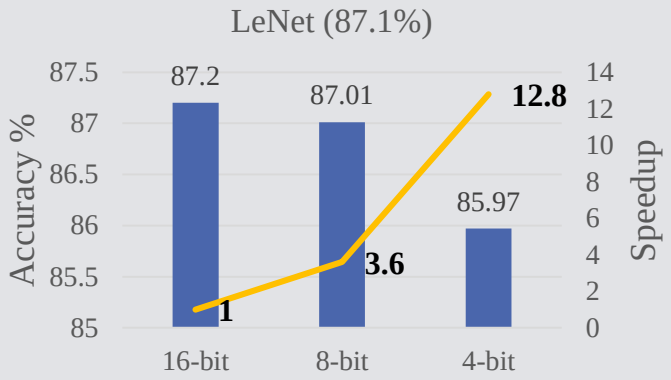


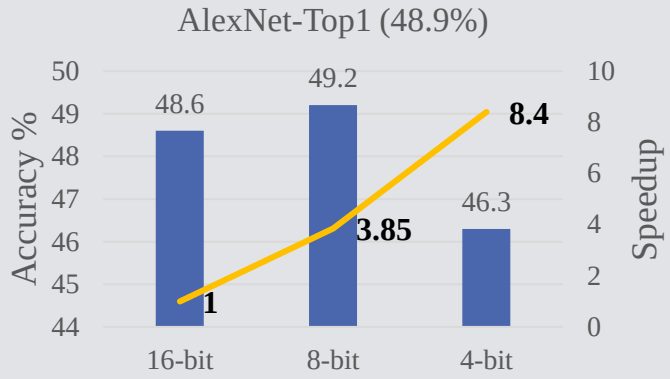


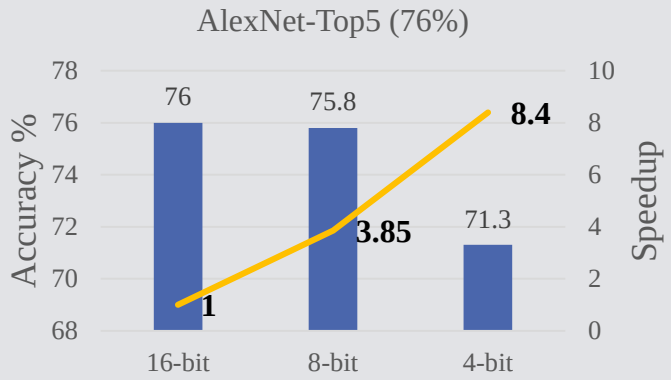




MpDNN Analysis

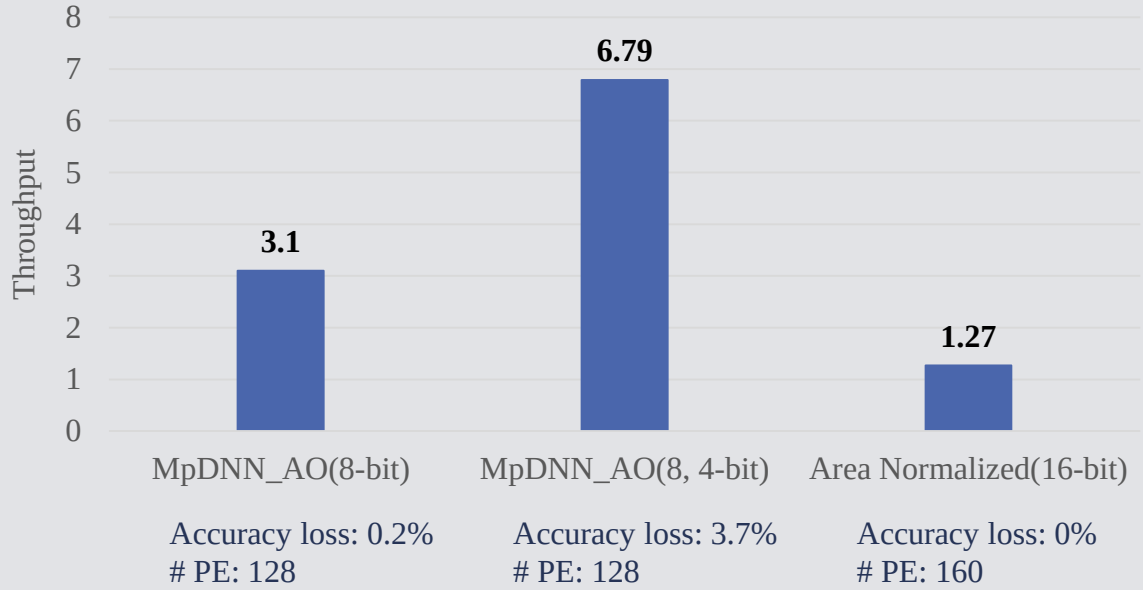






MpDNN_AO Throughput

Compare MpDNN_AO's throughput with 128 area optimized processing elements while implementing AlexNet



MpDNN vs. BitFusion

- Inefficiency of ASIC based design for FPGA

	Critical Path Latency (ns)	#LUT
BitFusion	13.01	337
MpDNN_AO	**10.5**	**313**

Conclusion

- Variable minimum required bit-width in different neural networks & layers of one network.
- Fusible multiplier based on FPGA's LUTs
  - Run various networks with different data bit-width
  - Three mode of execution controlled by a control signal
  - No need to reprogram the FPGA for different bit widths
  - One 16 × 16, four 8 × 8, or sixteen 4 × 4 multiplication
- Throughput 3.2 to 6.75 times more than Area Optimized, with 0.2% to 4.8% accuracy loss.

Thank you for you attention! Questions?

Negar Neda – negar@nyu.edu

ASP-DAC, January 2022