Loom: Exploiting Weight and Activation Precisions to Accelerate Convolutional Neural Networks

Electrical and Computer Engineering Department, University of Toronto

Sayeh Sharify, Alberto Delmas Lascorz, Patrick Judd, Andreas Moshovos

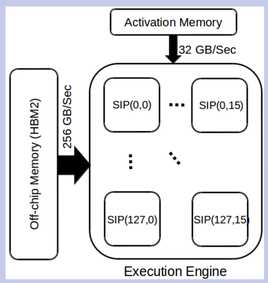
Email: {sayeh, delmasl1, juddpatr, moshovos}@ece.utoronto.ca

Abstract

- Hardware inference accelerator for CNNs.
- Targets area constrained System-on-a-Chip designs.
- Processes both weights and activations bit-serially.
- Weights are supplied via a High Bandwidth Memory v2 (HBM2) interface.
- Lower precision performance gain
  - Convolutional layer: 256/(Pa x Pw)
  - Fully-connected layer: 16/Pw
- Performance and energy efficiency improvements over a state-of-the-art bit-parallel accelerator
  - 2.34x performance improvement
  - 2.23x more energy efficiency
- Enables performance, energy efficiency and accuracy trade-off

Loom Architecture

- Weights
  - From off-chip memory
  - 2048 weights/cycle, 128 filters 16 lanes each
- Input activations
  - From activation memory
  - 256 activations/cycle, 16 windows 16 lanes each
- Serial Inner-Product (SIP) unit
  - Multiplies 16 weights and 16 activations bit-serially
  - Reduces the products to a single output



Loom Architecture

Activation and Weight Precision

Network	Convolutional layers	
	Per Layer Activation Precision in Bits	Network Weight Precision in Bits
	100% Accuracy	
NiN	8-8-8-9-7-8-8-9-9-8-8-8	11
AlexNet	9-8-5-5-7	11
GoogLeNet	10-8-10-9-8-10-9-8-9-10-7	11
VGG_S	7-8-9-7-9	12
VGG_M	7-7-7-8-7	12
VGG_19	12-12-12-11-12-10-11-11-13-12-13-13-13-13-13-13	12

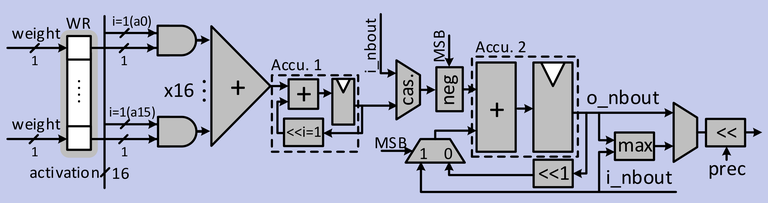
Per layer activation precisions and per network weight precision profiles for the convolutional layers

Network	Fully connected layers	
	Per Layer Weight Precision in Bits	Per Layer Weight Precision in Bits
	100% Accuracy	99% Accuracy
AlexNet	10-9-9	9-8-8
GoogLeNet	7	7
VGG_S	10-9-9	9-9-8
VGG_M	10-8-8	9-8-8
VGG_19	10-9-9	10-9-8

Per layer weight precisions for fully-connected layers

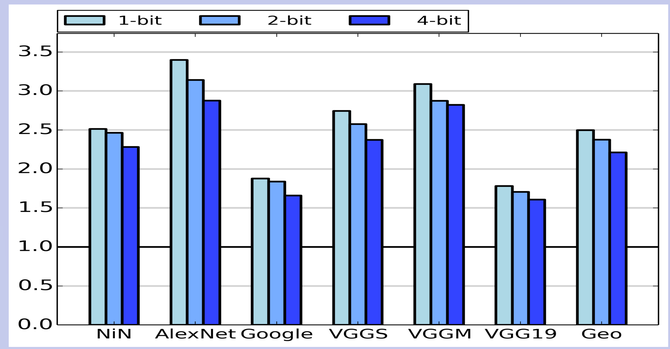
Loom Serial Inner-Product Unit

- Data Inputs:
  - 16 Activations, 1 bit per cycle
  - 16 Weights, 1 bit per cycle
  - Supports cascading for smaller layers
  - Area: 1,840 μm^2

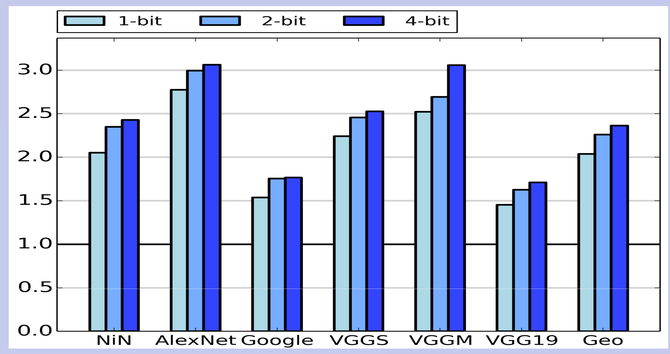


Serial Inner-Product Unit (SIP)

Results



Performance relative to baseline



Energy efficiency relative to baseline

Datapath Layout

- 65nm TSMC
- Clock frequency: 980 MHz
- Area for 1x16 array: 29466.72 μm^2
- Area for Tile:
  - 1-bit/cycle: 2,878,069.88 μm^2
  - 2-bits/cycle: 2,165,571.71 μm^2
  - 4-bits/cycle: 1,450,366.33 μm^2
- Tile Power for 1-bit/Cycle
  - Static: 26.37 mW
  - Dynamic: 986.48 mW