INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD
R&D SHOWCASE 2021
Technology, Social Impact

TreeNet: An Overlay Architecture for Vector Reduction using a Network of Trees

Abstract

- High-throughput reduction circuits complements FPGA based accelerator designs for machine learning and Image processing.
- TreeNet is a tree based FPGA overlay design, that can reduce arbitrary length vectors, on the fly, in a data-parallel fashion.

Design Overview

- Each vector is split into partitions, whose sizes are in powers of 2.
- Same sized partitions from different vectors are grouped and laid out over a binary reduction tree in the decreasing order.
- The per-partition reduced values inside the tree are added by shifting the per level outputs.

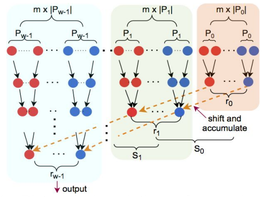


Figure 1: Pictorial description of our algorithm processing M vectors of equal but arbitrary length N.

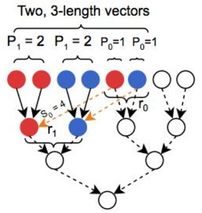


Figure 2: The figure shows the same 8 sized Tree Network used to reduce two vectors of 3-length each.

Hardware Architecture

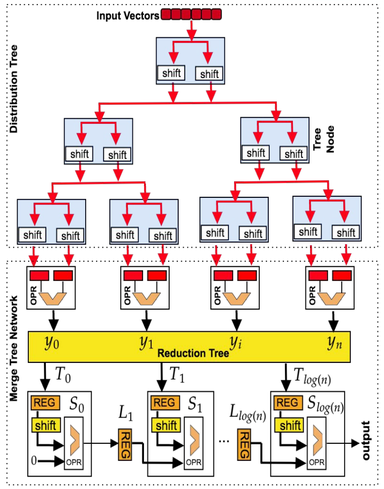


Figure 3: The overall architecture of TreeNet

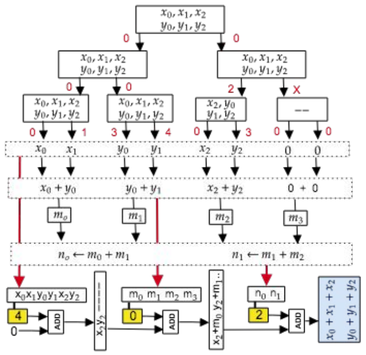
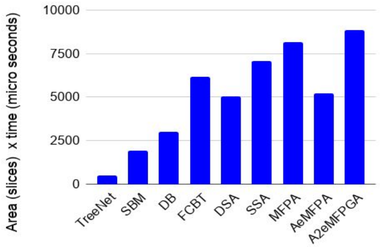


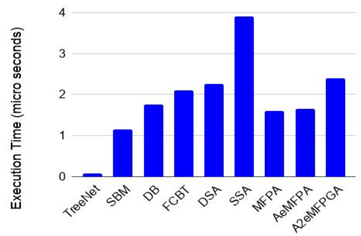
Figure 4: How TreeNet reduces 2 3 length vectors X and Y in parallel. The shift values are shown over the edges of the distribution tree

- TreeNet hardware, Figure 3, consists of two pipelined tree structures, each of height log N, here N is the vector length.
- The distribution tree interleaves the vector partitions using combinatorial shifters and lays them over the reduction tree.
- The reduction tree, reduces the per-vector partitions into single values. Which are later merged using a sequence of log N, shift-and-accumulate stages.

Experimental Results

- We synthesized TreeNet on a Xilinx Virtex-7-690t FPGA, connected to an Intel Core-i5 CPU running at 3.0 GHz.
- The TreeNet hardware achieves a 20x speed-up and a 3.9x improvement in the area-time product compared to the state-of-the-art.





	Throughput GOps/Sec	Performance Density GOps/KLUTs	Precision Fixed-Point	Frequency MHz
AlexNet	1200	5.38	16 bits=8,8	166
VGG-16	1025	4.60	16 bits=8,8	166
	Execution Kilo Cycles (EC)	Memory Kilo Cycles (MC)	Theoretical Kilo Cycles (TC)	Pipeline Rate (EC+MC)/TC
AlexNet	1096	112	856	1.45
VGG-16	7046	431	5149	1.44

We test the adaptivity of TreeNet by using it to process the AlexNet and VGG-16 CNNs with a tree size of 1024.

Authors: Ziaul Choudhury, Shashwat Shrivastav, Shashwat Khandelwal, Anish Gulati and Suresh Purini

Research Center Name: Computer Systems Group