Hardware Accelerated Learning

USC University of Southern California

USC Viterbi School of Engineering

Ming Hsieh Department of Electrical Engineering

A Highly Parallel FPGA Implementation of Sparse Neural Network Training

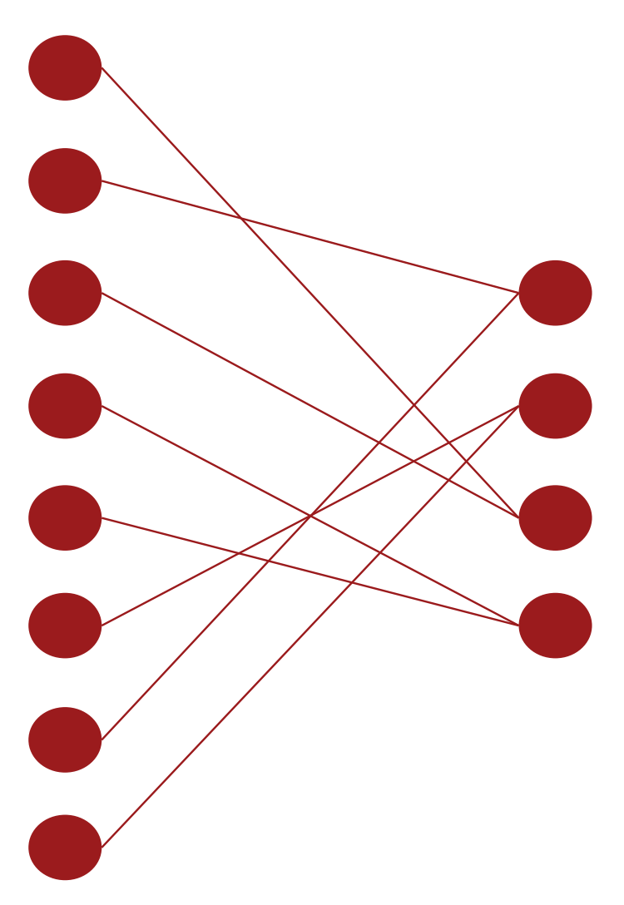
Sourya Dey, Diandian Chen, Zongyang Li, Souvik Kundu, Kuan-Wen Huang, Keith Chugg, Peter Beerel, Hardware Accelerated Learning group, USC

Motivation & Introduction

Neural networks too big to be trained on-chip
Cloud resources are costly

Our Solution: **Pre-defined sparsity**
Reduces edges, hardware friendly
Fixed in-, out-degree of each node

Train neural networks on FPGAs

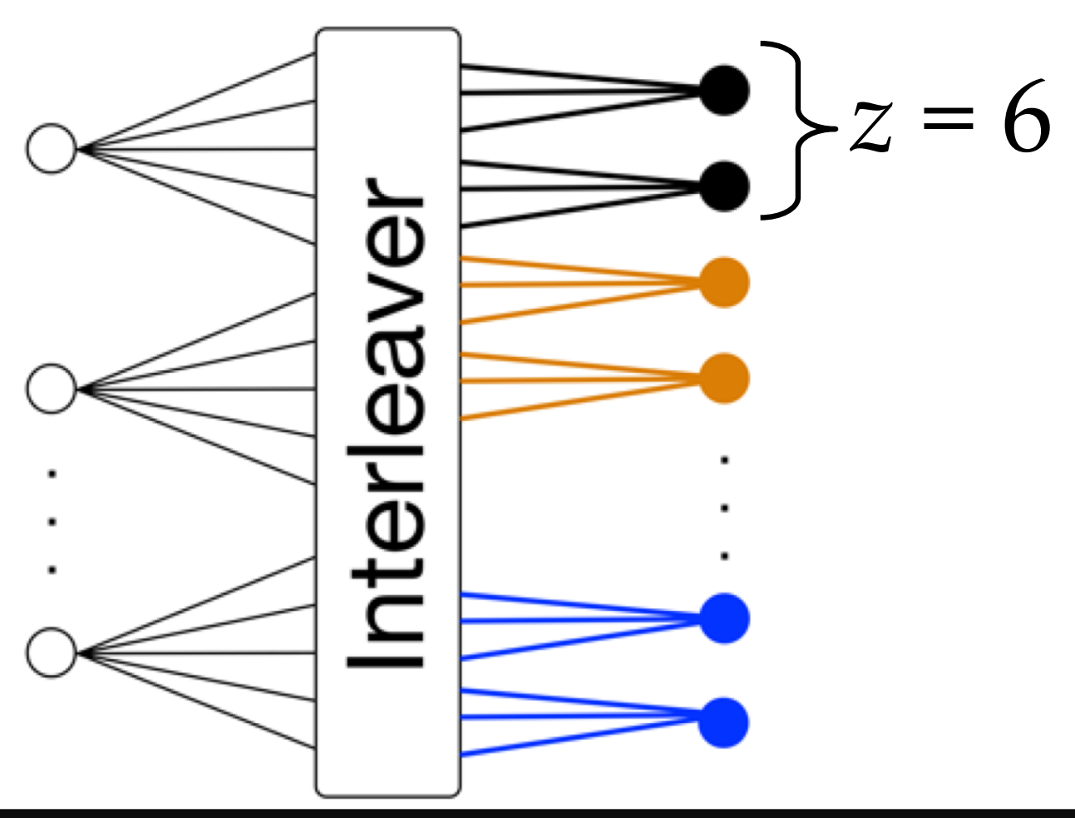


Methodology

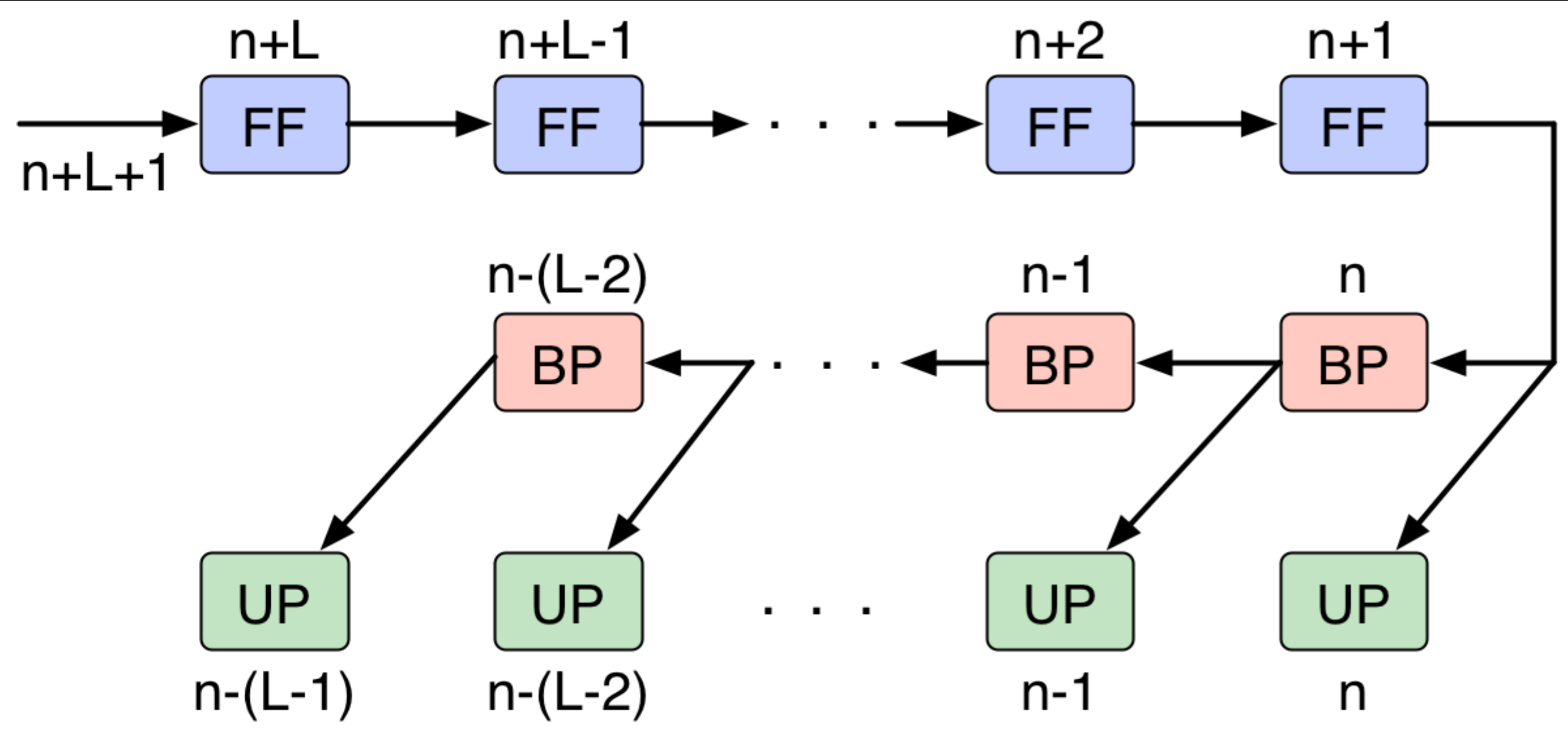
3 operations:
- ➢ Feedforward (FF)
- ➢ Backpropagate (BP)
- ➢ Update (UP)

} All use weighted junction edges

- ✓ Process z edges in 1 clock cycle
- ✓ 1 **block cycle** = Total clock cycles to process all edges in any junction
- ✓ Ideal throughput = (Block cycle)$^{-1}$

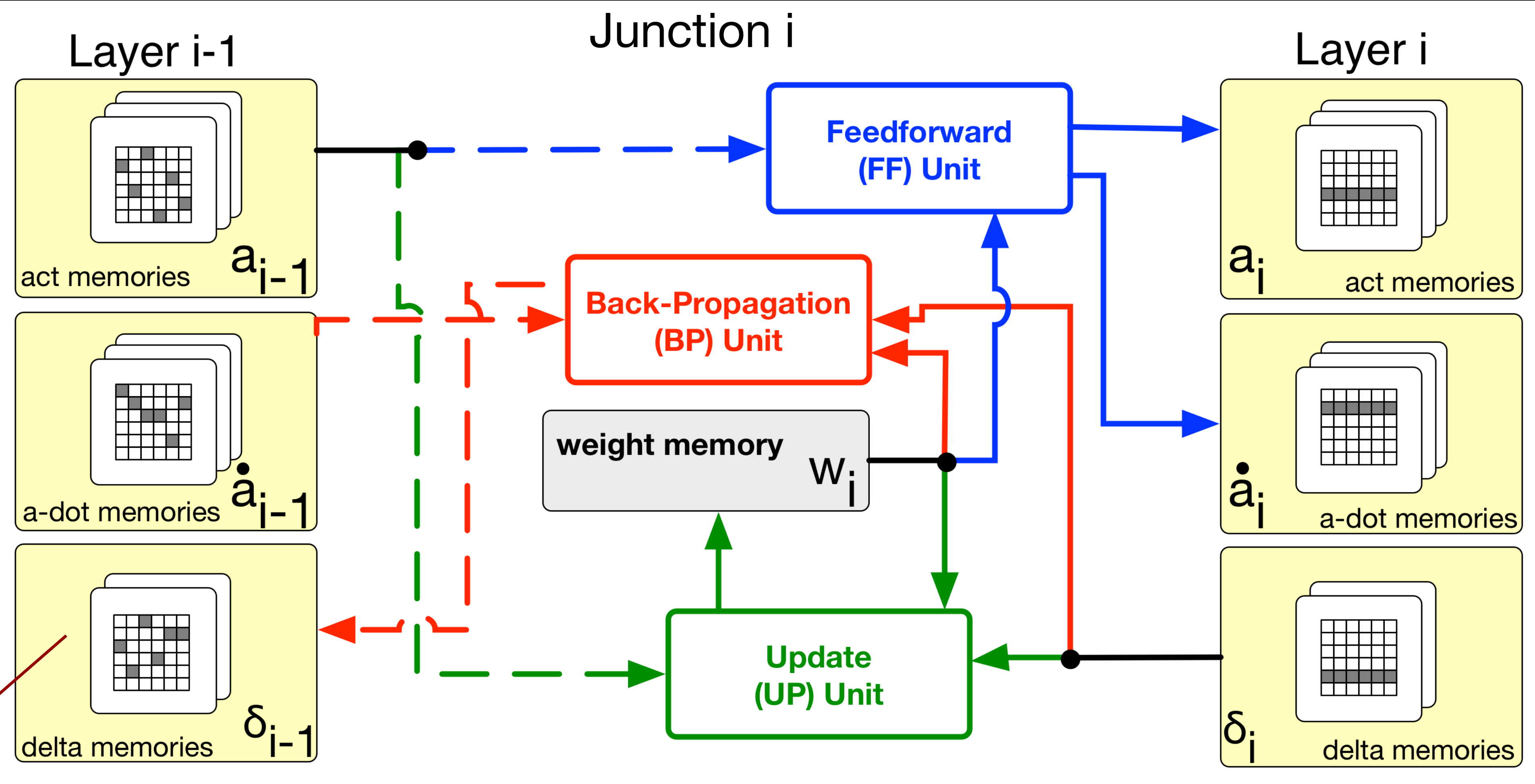


Hardware Acceleration – Parallelism and Pipelining



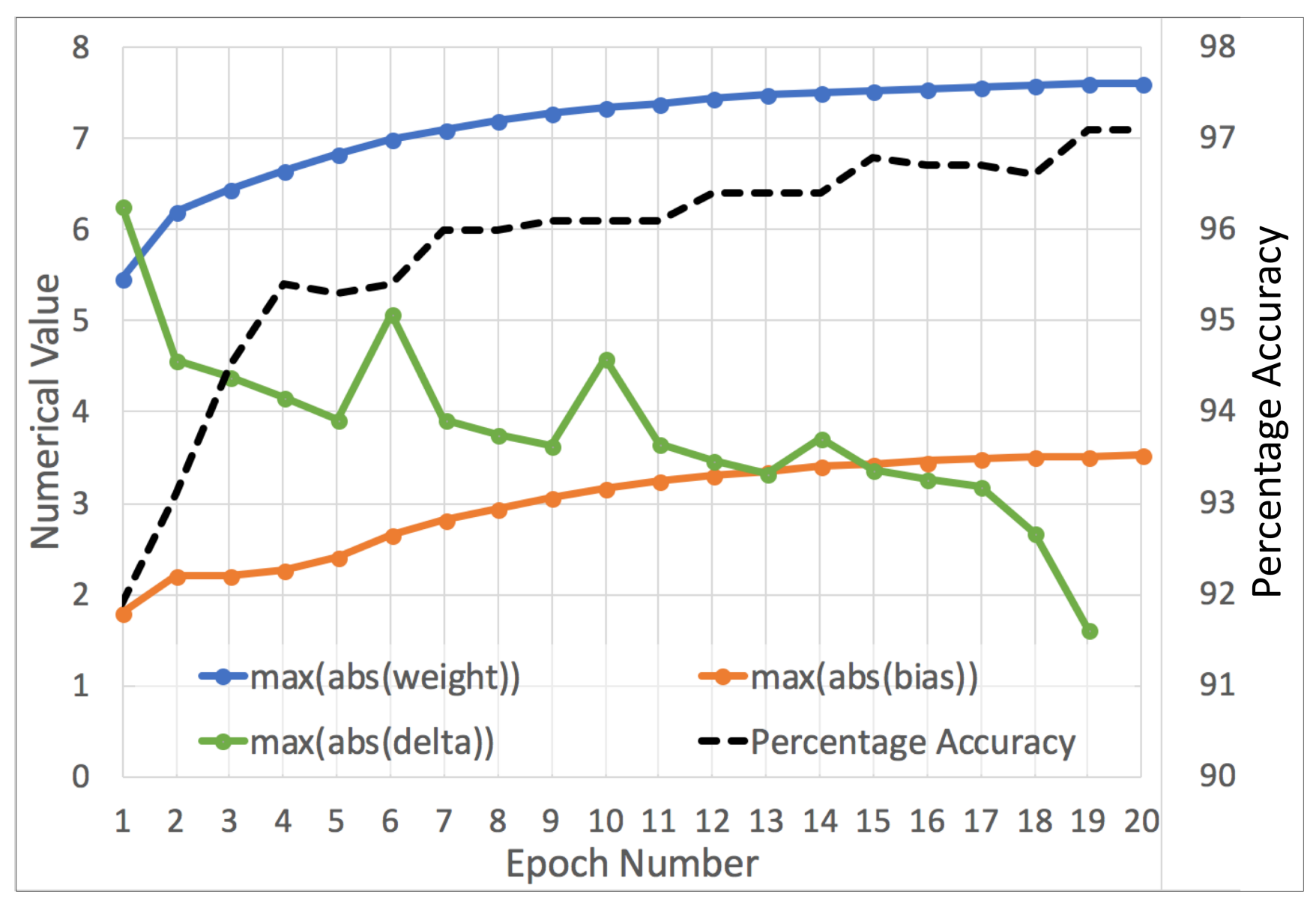
Junction pipelining: Different input samples processed together across all junctions

Clash Freedom: Each memory accessed at most once in a cycle



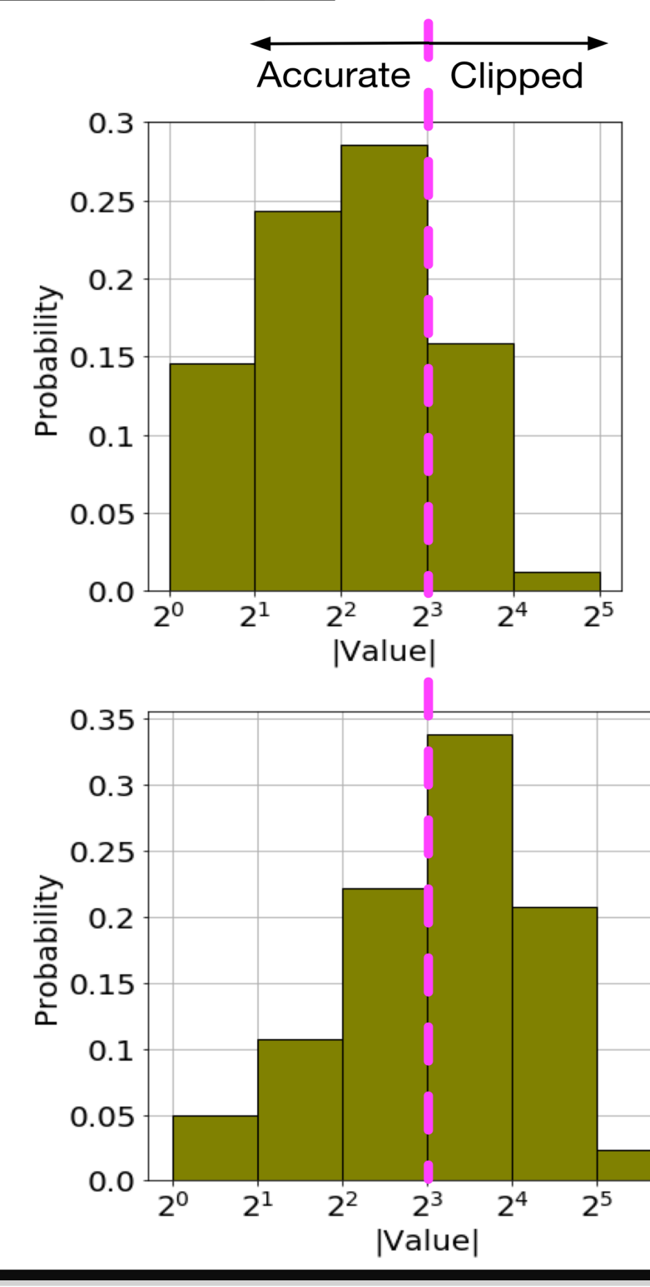
Operational Parallelization: FF, BP, UP together inside a junction

Bit Width Studies



Network parameters stay within 8 => Use fixed point with 3 integer bits

Histograms of weight-activation dot product in junction 1.



(a) Sparse
Mean = 4.6
Stddev = 3.8
% Clipped = 17

(b) FC
Mean = 11
Stddev = 8.2
% Clipped = 57

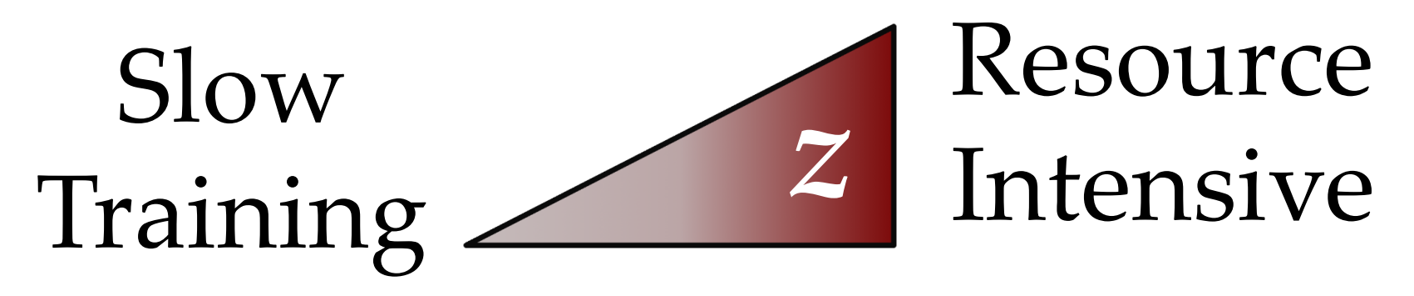
Pre-defined sparse networks have less errors due to finite bit-width effects.

FPGA Implementation – MNIST Training and Inference on Artix-7

Junction Number	1	2
Left Neurons	1024	64
Right Neurons	64	32
Out-degree	4	16
Weights	4096	1024
In-degree	64	32
z	**128**	**32**
Block Cycle	32	32
Density	6.25%	50%

Overall Density	7.576%
Fixed Point Bit Width	12
Clock Frequency	15 MHz
Block Cycle Duration	2.27 μs
Accuracy (after 14 training epochs)	96.5%

Reconfigurability



Ongoing Research:
- ➢ Increased pipelining to improve speed
- ➢ Memory bandwidth management for bigger networks

Contact Information: souryade@usc.edu, chugg@usc.edu, pabeerel@usc.edu

This work is partly supported by National Science Foundation, USA, Grant #1763747.

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering