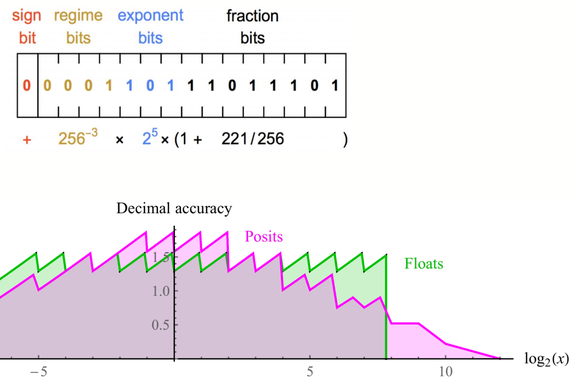
# Accelerating Convolution Neural Network Inference with Posit Computing

Alexander Brown, Electrical Engineering
Mentor: Jeff Zhang, Assistant Professor
School of Electrical, Computer and Energy Engineering

## Background:

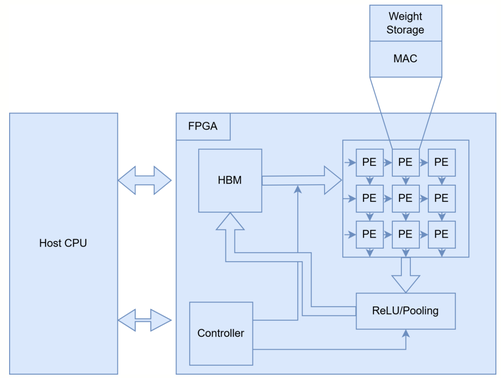
- Dr John Gustafson invented the Posit number system as an alternative to floating point, capable of reducing arithmetic error
- Recently, Posits have been explored for neural network applications, because of their potential to reduce memory usage, power and area, while maintaining accuracy



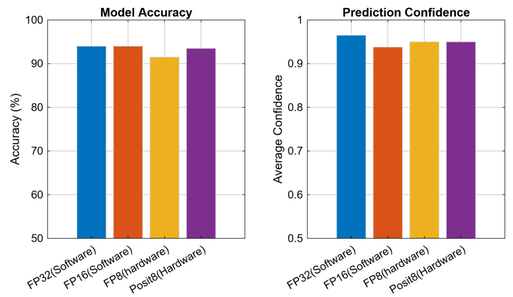

Above: Posit number format decoding and accuracy (from Gustafson and Yonemoto 2017)

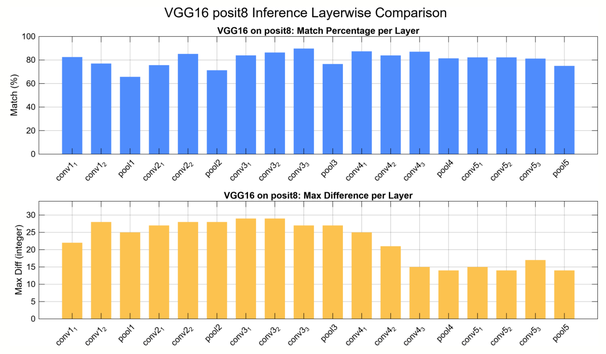
## Objective & Research Question:

The goal of this project is to create an 8bit Posit based accelerator, to improve accuracy in convolution neural network workloads. This project seeks to answer if smaller posit number formats can achieve the same accuracy as floating point?




Above: Flow diagram of the device




Above: accuracy by number system + layer wise accuracy vs FP32

## Methods:

- The project uses custom hardware using SystemVerilog capable of doing convolution and pooling operations using Posit arithmetic.
- Convolution is done with a weight stationary systolic array to do matrix multiplication, and a custom instruction set and address generation
- CNN inference passes were done with the VGG16 model using the CIFAR-10 dataset
- This was compared against a duplicate design using 8 bit floating point for accuracy and throughput

## Results:

- Summary: posit8 can achieve >93% accuracy in image classification on the CIFAR dataset
- Compared to FP32 and FP16, accuracy was only ~0.5% less, while reducing memory usage by ~8x or ~4x respectively
- Compared to FP8 it gets ~2% better accuracy at the same bit depth, showing the better dynamic range
- Overall, very promising results, future research should focus on more difficult models or datasets

FURI

Ira A. Fulton Schools of Engineering
Arizona State University