HSLU Hochschule Luzern

Technik & Architektur

Master of Science in Engineering
Specialization: Electrical Engineering

Master of Science in Engineering – Master Thesis

Energy-Efficient Edge-AI with Low-Precision Posit Arithmetic

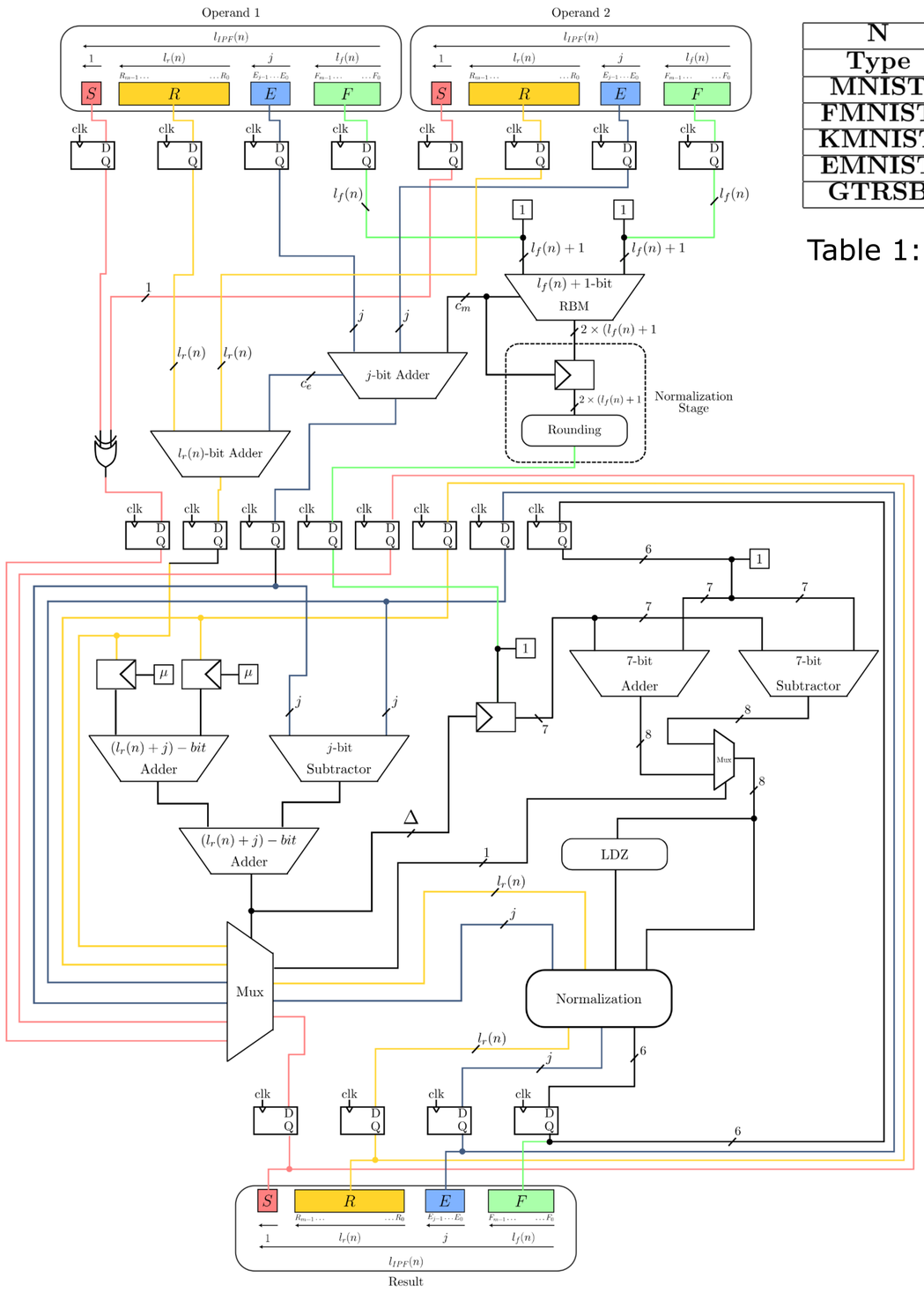

Figure 1: RTL diagram of the Posit Fused-Multiply-Accumulate Unit

N	1		2		4		8	
Type	$P_{sm}\langle 6,2\rangle$	$Q_{tc}\langle 5.6\rangle$	$P_{sm}\langle 6,2\rangle$	$Q_{tc}\langle 5.6\rangle$	$P_{sm}\langle 6,2\rangle$	$Q_{tc}\langle 5.6\rangle$	$P_{sm}\langle 6,2\rangle$	$Q_{tc}\langle 5.6\rangle$
MNIST	29.45	19.15	50.15	30.15	75.75	31.1	160.30	48.95
FMNIST	17.35	27.25	43.45	32.15	72.95	38.15	145.45	64.15
KMNIST	34.35	27.45	53.95	31.8	85.05	39.6	168.95	65.65
EMNIST	29.65	33.75	62.15	41.95	92.15	46.50	178.25	73.9
GTRSB	36.25	30.02	66.55	44.55	108.55	47.00	210.95	77.05

Table 1: Measured Power [mW] with Posit ($P_{sm}\langle 6,2\rangle$) and Fixed-point ($Q_{tc}\langle 5.6\rangle$)

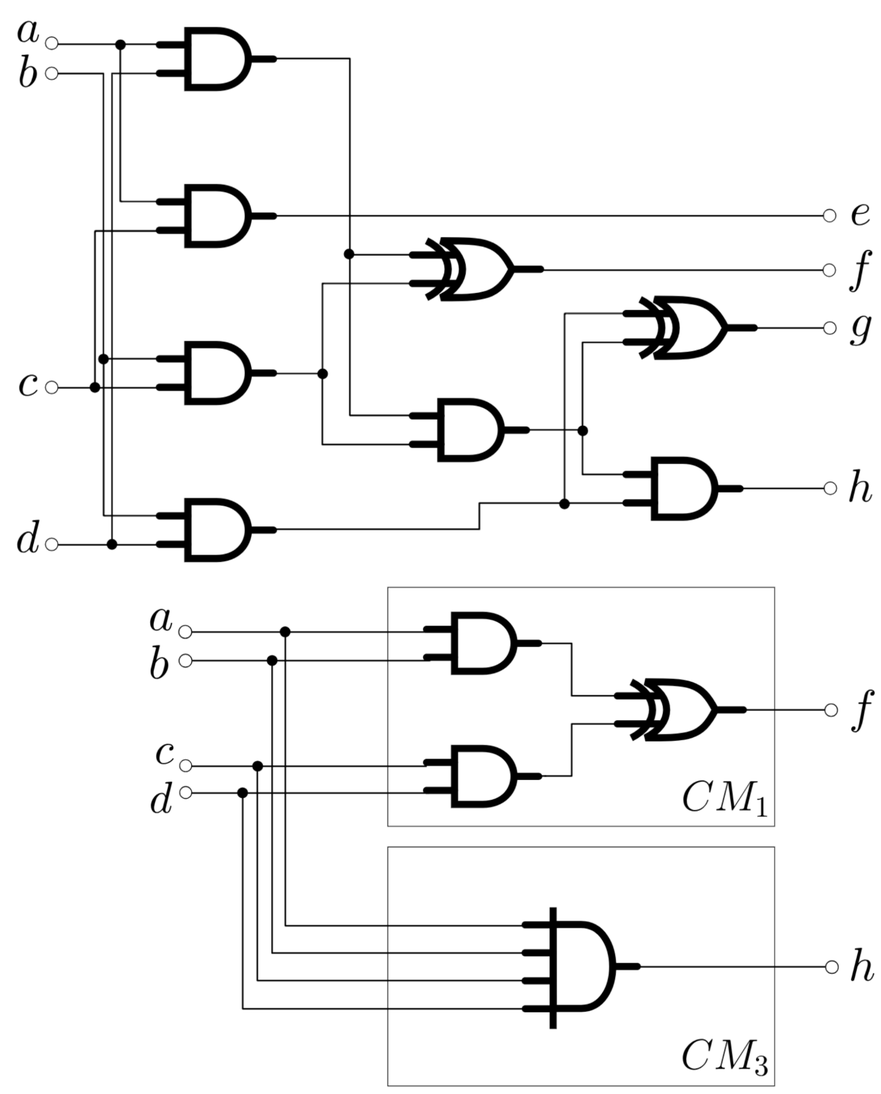

Figure 2: Regular multiplier (top) vs. RBM (bottom)

Project Task

In modern computational systems - especially for neural network inference - the choice of arithmetic format strongly influences performance and accuracy. Traditional floating- and fixed-point representations often struggle with the wide dynamic range and small values common in deep learning. Posit arithmetic addresses these challenges by flexibly allocating bits to the exponent, improving the representation of tiny numbers and reducing quantization error. Although not inherently more hardware-efficient, posit delivers higher precision where it matters, yielding more accurate inference results. Initial studies showed that N-bit posit matches the accuracy of 2*N-bit fixed-point formats and offers substantial power savings in ultra-low-precision multiplier designs. These promising findings suggest that replacing traditional FPGA DSP slices and CNN accelerators with posit units could cut power consumption without accuracy loss. This master thesis validated those predictions through FPGA power measurements of low-precision posit arithmetic designed at gate-level.

Concept

Recursive Bitwise Multiplication (RBM) is a novel ultra-low bitwidth mantissa multiplication optimization that computes each result bit with a dedicated Cell Multiplier (CM) block, avoiding partial sums and carry-over bits. An RBM circuit chains multiple CMs, omitting the hidden and first result bits to boost efficiency compared to traditional multipliers. For accumulation, the posit hardware was reimplemented in Python to evaluate fractional bit-widths from 3 to 6 bits under CNN inference. Testing revealed that a 6-bit fractional extension preserves full-precision accuracy while minimizing resource usage.

Results

The performed comparison evaluates the performance of posit and fixed-point representations across various datasets and levels of parallelization (N) in a closed CNN accelerator design. The CNN accelerator used for this measurement was built on the concept of a systolic array, adapted to efficiently leverage the computational capabilities of Processing Elements (PEs). The results show that fixed-point representation significantly outperforms Posit in overall accelerator power consumption. The circuit overhead required by Posit to handle a high dynamic range in Fused-Multiply-Accumulate (FMA) operations outweighs any multiplier savings.

Table 1 shows the net power values [mW] measured on an FPGA platform for the CNN accelerator using Posit and fixed-point arithmetic at various parallelism levels N.

Student:
Matej Hrvat

Advisor:
Prof. Dr. Jürgen Wassner

Expert:
Dr. Marc Wegmüller

FH Zentralschweiz

MSE | MASTER OF SCIENCE IN ENGINEERING