Flash Attention on FPGA

Aakarsh A
Department of Computer Science and Automation
Indian Institute of Science

May 30, 2025

Overview

1. e^{-x} Implementation & simulations
2. Phase-2 Implementation & simulations
3. Future Work

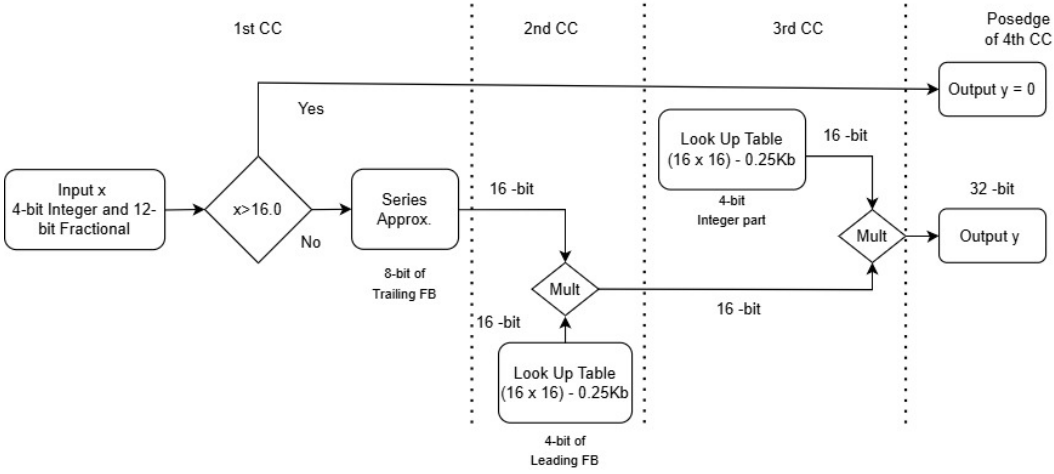
Why Hybrid Approach?

Our goal: Hardware acceleration.

- Faster Memory Access (Infer LUTs) at the expense of hardware (DSPs).
- Minimize Latency and maximize operation frequency.

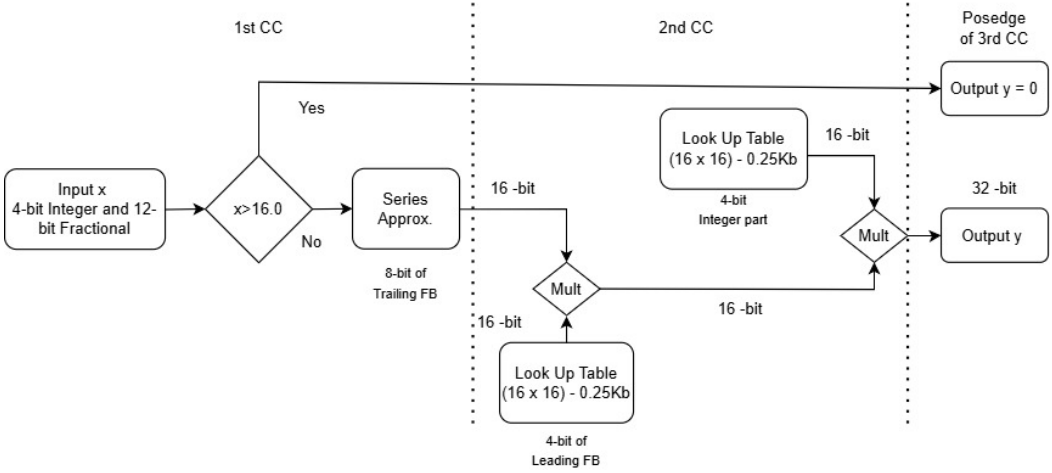
Hybrid Approach

Pipelined Dual LUT & Series Approximation: 1



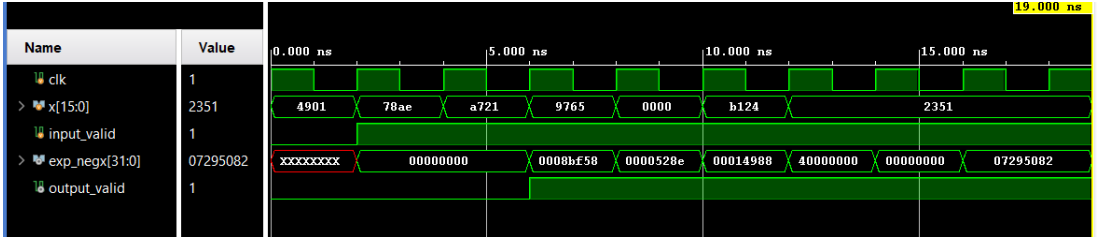
Latency is 3 clock cycles, Throughput is 1 and Higher Operating frequency possible.
DSPs used: 2

Pipelined Dual LUT & Series Approximation: 2



Latency is 2 clock cycles, Throughput is 1 and Lower Operating frequency than previous method. DSPs used: 2

Simulation Results

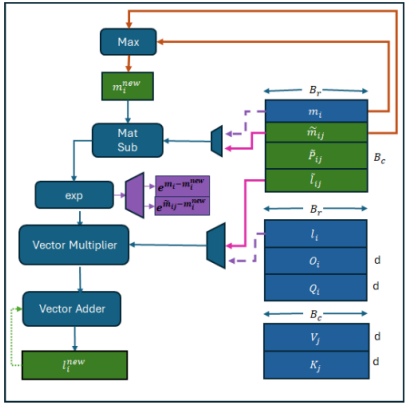


- Output is **32** bits: 2 Integer Bits and 30 Fractional bits
- Achieved a maximum error of less than 0.0005 (0.05 % error in values).
- The previous implementation using $2^{\frac{-x}{\ln 2}}$ was replaced due to higher latency when pipelined and their difficulty for realizing smaller fractions.

Just a look at Phase-2 block diagram

Phase 2 : Calculation of m_i^{new} and l_i^{new}

- $m_i^{new} = \max(m_i, \tilde{m}_{ij})$,
- $l_i^{new} = e^{m_i - m_i^{new}} l_i + e^{\tilde{m}_{ij} - m_i^{new}} \tilde{l}_{ij}$



RTL Components

- Negative Exponential Engine
- BRAM/URAM Access and Storage
- Timing and Control Unit
- Vector Multiplier
- Vector Adder
- Subtractor
- Comparator

Deeper dive: 1

Process & Constraints

- For a single tile, the practical values of B_r and B_c are in the range [1K, 10K].
- We obtain [B_r x 1] vector from Phase-1 for m_{ij} and l_{ij} through the Dual-port RAM, which has W-width bits and D-depth places.
- The parameter W of inferred RAM limits our simultaneous access to the memory in a single clock cycle. Thus, for the computation of the vector [B_r x 1], we require

$$B_r \cdot \frac{\text{datawidth}}{W} \tag{1}$$

additional cycles to compute the full vector, and the following operation can be pipelined with throughput $= 1$.

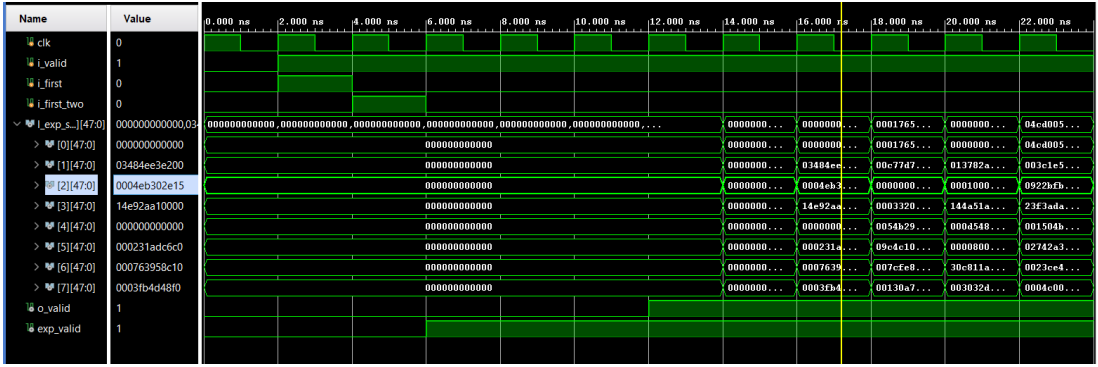
Deeper dive: 2

Timing Performance & Resource utilization

- Latency: $1 + 2 + 1 + 1 + 1 = 6$ clock cycles (includes Latency due to BRAM access, Subtraction, Exponential Calculation & final product) for first W bits output.
- BRAM - Use the full Width and Depth of BRAM block - number depends on the size of tile.
- DSPs inferred for Phase-2

$$\frac{6 \times W}{\text{datawidth}} \tag{2}$$

Simulation Results



- Output is **48** bits which can approximatied to 16-bits depending on requirements.
- Output tested for dummy values where $B_r = 128$, where in each burst access, 16 elements are accessed, and this is done for 8 clock cycles.

Future Work

1. Fine-tuning Phase-2 module for different parameters.
2. RTL implementation of Phase-3 with integration with Phase-2.
3. Verification of Phase-2 on the Alveo V80 FPGA.
4. Any other work?

THE END

Feedback & Improvement Ideas?