Hardware Acceleration of Softmax Function using FPGA

Objective:

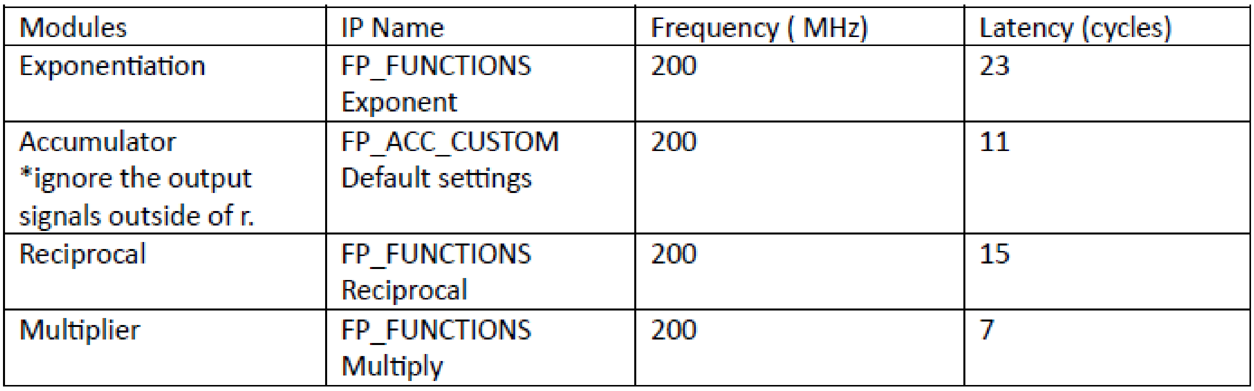
To accelerate the function by implementing pipelining and parallelization using Quartus IP.

This is done by first implementing the base circuit for single cycle calculation of the function. Acceleration is then done by introducing RAMs as buffers/pipeline registers. Then modules are parallelized to further boost performance.

Benchmarking is done at the end to evaluate the optimal amount of parallelization to other parameters (latency, power consumption, area etc.). Basically, comparing the performance of non-parallelized circuit with circuit with parallelized modules.

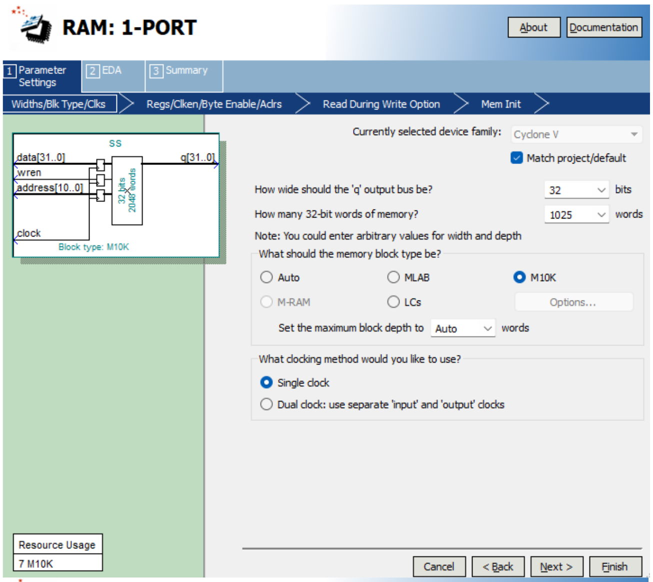
System Specifications:

- Input:
  32-bit IEEE754 Floating Point.
- Software :
  Synthesis ➞ Intel Quartus Prime Lite edition 20.1
  Simulation ➞ ModelSim
- Hardware:
  Cyclone V.
- IP Module Specifications (adhere to this):



Modules	IP Name	Frequency (MHz)	Latency (cycles)
Exponentiation	FP_FUNCTIONS Exponent	200	23
Accumulator *ignore the output signals outside of r.	FP_ACC_CUSTOM Default settings	200	11
Reciprocal	FP_FUNCTIONS Reciprocal	200	15
Multiplier	FP_FUNCTIONS Multiply	200	7

RAMS are specified as below. Feel free to use manipulate these settings(other bits and words), or use other RAM IPs if there are better ways to do so.



Currently selected device family: Cyclone V

☑ Match project/default

How wide should the 'q' output bus be? 32 bits

How many 32-bit words of memory? 1025 words

Note: You could enter arbitrary values for width and depth

What should the memory block type be?

○ Auto ○ MLAB ● M10K

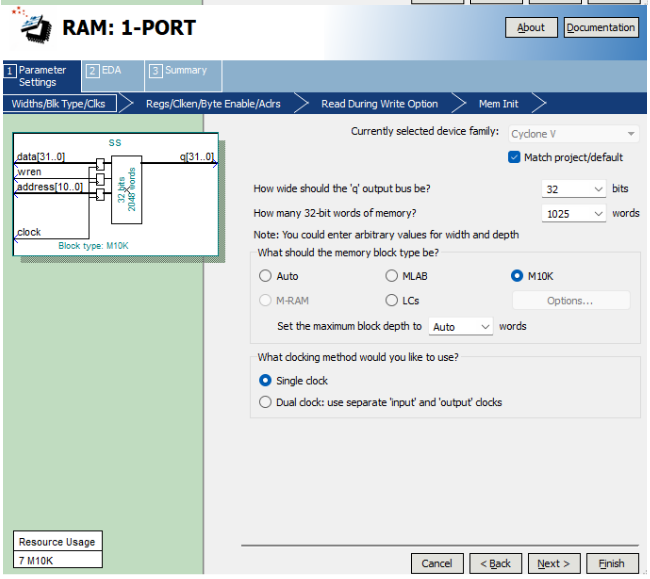
○ M-RAM ○ LCs

Set the maximum block depth to Auto words

What clocking method would you like to use?

● Single clock

○ Dual clock: use separate 'input' and 'output' clocks


Currently selected device family: Cyclone V

☑ Match project/default

How wide should the 'q' output bus be? 32 bits

How many 32-bit words of memory? 1025 words

Note: You could enter arbitrary values for width and depth

What should the memory block type be?

○ Auto ○ MLAB ● M10K

○ M-RAM ○ LCs

Set the maximum block depth to Auto words

What clocking method would you like to use?

● Single clock

○ Dual clock: use separate 'input' and 'output' clocks

Project Scope:

1. Design Phase
   - **Current architecture analysis:**
     Baseline architecture is the high-level block diagram for the Softmax accelerator.
     Please verify if the intended output is correct, and if not please modify wherever possible.

   - **Control Unit FSM design:**
     To ensure smooth operation and correct function of the circuit after pipelining and parallelization, a control unit is required to correctly address the timing of address generation, data input/output from RAMs.

2. Implementation Phase
   - Softmax core implementation using Quartus IP with parallelization and pipeline integration.
   After analyzing (and making corrections if needed) the baseline circuit, please implement pipelining using RAMs and other modules if needed. For parallelization, parallelize the exponentation module and reciprocal module*)

   - Control unit development :
   A control unit to :
   1) Control the signals to ensure correct read/write for RAMs from the output of parallelized modules.
   2) Ensure smooth transition of data so that the output is correct (I am unclear about this, I need your insights whether if its possible or not)
   3) A control unit to ensure smooth operation in accordance to the addition of parallelized modules (Again, I am unclear about this and I need your insights)

3. Testing phase
   -Functional Verification:
   Conduct functionality of the baseline, pipelined and parallelized circuit through testbenches.

   -Timing Analysis:
   Conduct timing analysis. I will leave this to your expertise.

   -Performance Benchmarking :
   Conduct resource utilization analysis of the circuit in terms of latency, power consumption, area, etc. according to the amount of parallelization. For example an analysis of non parallelized circuit to only 2 parallelized modules to 3 parallelized modules. If possible, until 10 parallelized modules.

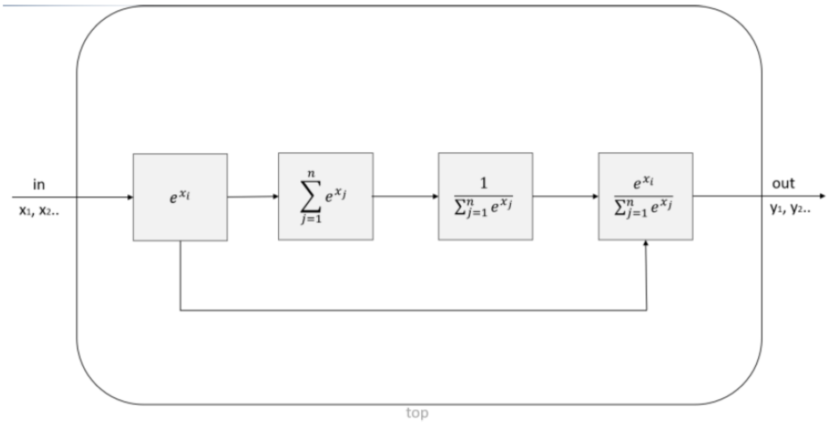
4. Documentation
   - **Technical Documentation:**
     Please include description of the working of circuit through waveform analysis. Also include performance report, description of the FSM working and modifications made. Other than that, I will leave it to your expertise.

Expected outcome:

Something similar to

https://github.com/maomran/softmax?tab=readme-ov-file

Where there is a control unit to ensure correct flow of data in according to the generated parallelized modules of exponentiation and reciprocal modules.



in
x₁, x₂..
e^{x_i}
$\sum_{j=1}^{n} e^{x_j}$
$\frac{1}{\sum_{j=1}^{n} e^{x_j}}$
$\frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$
out
y₁, y₂..
top

Picture 1. Baseline circuit