Custom Acceleration with FPGAs

CS5222 – Project 2

SHEN JIAMIN (A0209166A)

Your report in `.pdf` format with concise answers to the questions asked in Parts 1 and 2. In addition, the report should contain a short (couple paragraphs) description of your optimized implementation for Part 3 (what you changed about the hardware, or classifier, or both).

1 MATRIX MULTIPLICATION PIPELINE OPTIMIZATION IN HLS

1.1 Understanding the baseline matrix multiply (background)

1.2 B. Pipelining in HLS (8 marks)

Report

(1) the design latency in cycles,
(2) the overall device utilization (as Total per Resource),
(3) the number of floating point adders and multipliers (you can find this information under the Instance section of the synthesis report) and
(4) the Initiation Interval of the loops you pipelined.

1.3 C. Increasing Pipeline Parallelism by Repartitioning Memories (8 marks)

Report

(1) the design latency in cycles,
(2) the overall device utilization (as Total per Resource),
(3) the number of floating point adders and multipliers (you can find this information under the Instance section of the synthesis report) and
(4) the Initiation Interval of the loops you pipelined.

1.4 D. Amortizing Iteration Latency with Batching (8 marks)

Report

(1) the design latency in cycles, and
(2) the overall device utilization (as Total per Resource).

1.5 E. Extending Batch Size with Tiling (8 marks)

Report

(1) the design latency in cycles, and
(2) the overall device utilization (as Total per Resource).

Author's address: Shen Jiamin (A0209166A), shen_jiamin@u.nus.edu.

1.6 F. Hardware compilation and FPGA testing on the PYNQ (8 marks)

Report

(1) the measured speedup and
(2) measured classification accuracy.

2 PART 2: FIXED-POINT OPTIMIZATIONS (30 MARKS)

(1) the fixed-point validation accuracy reported by mnist.py after you've tweaked the SCALE factor.
(2) the design latency in cycles
(3) the overall device utilization (as Total per Resource).
(4) your measured system speedup over the fixed-point CPU implementation
(5) your measured classification accuracy on the 8k MNIST test sample
(6) how many multipliers are instantiated in your desing?
(7) report the initiation interval of the matrix multiplication loop that you pipelined
(8) given the number of multipliers in your design and input throughput via the AXI port, is the design bandwidth- or compute-limited?

3 PART 3: OPEN-ENDED DESIGN OPTIMIZATION (30 MARKS)