Custom Acceleration with FPGAs

CS5222 – Project 2

SHEN JIAMIN (A0209166A)

In this project, I'm going to port the lab to **PYNQ 2.7** and **Vivado/Vitis 2020.2**. The experiment is done on ASUS RS500-E8-PS4 V2, with operating system Ubuntu 20.04.4 LTS (GNU/Linux 5.4.0-100-generic x86_64).

1 MATRIX MULTIPLICATION PIPELINE OPTIMIZATION IN HLS

1.1 A. Understanding the baseline matrix multiply (background)

The report generated by HLS (as in Figure 1) shows that some pipelining has already been done automatically by Vitis. In order to prepare for the next part, I disabled the pipelining.

```
--- hls.tcl        2022-03-03 21:17:24.651417872 +0800
+++ hls_nopipe.tcl        2022-03-03 21:33:53.435003340 +0800
@@ -7,6 +7,7 @@
 open_solution "solution0" -flow_target vivado
 set_part {xc7z020clg484-1}
 create_clock -period 10 -name default
+config_compile -pipeline_loops 0
 csim_design -clean
 csynth_design
 close_project
```

The new report is as Figure 2. It turns out that the overall performance is a little bit worse than documented. This is because every iteration in L3 loop takes 11 cycles and thus 2816 cycles in total to perform a single inner product.

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.

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

(a) Performance Estimates

```
* Summary:
+---------+---------+----------+----------+-------+-------+---------+
|  Latency (cycles) |  Latency (absolute) |    Interval   | Pipeline|
|   min   |   max   |    min   |    max   |  min  |  max  |   Type  |
+---------+---------+----------+----------+-------+-------+---------+
|    85160|    85160|  1.236 ms|  1.236 ms|  85161|  85161|     none|
+---------+---------+----------+----------+-------+-------+---------+

+ Detail:
    * Instance:
    N/A

    * Loop:
    +----------------------+---------+---------+----------+-----------+-----------+------+----------+
    |                      |  Latency (cycles) | Iteration|  Initiation Interval  | Trip |          |
    |       Loop Name      |   min   |   max   |  Latency |  achieved |   target  | Count| Pipelined|
    +----------------------+---------+---------+----------+-----------+-----------+------+----------+
    |- LOAD_OFF_1          |        5|        5|         1|          1|          1|     5|       yes|
    |- LOAD_W_1_LOAD_W_2   |     1280|     1280|         2|          1|          1|  1280|       yes|
    |- LOAD_I_1_LOAD_I_2   |     1024|     1024|         2|          1|          1|  1024|       yes|
    |- L1_L2               |    82800|    82800|      1035|          -|          -|    80|        no|
    | + L3                 |     1031|     1031|        12|          4|          1|   256|       yes|
    |- STORE_O_1_STORE_O_2 |       42|       42|         4|          1|          1|    40|       yes|
    +----------------------+---------+---------+----------+-----------+-----------+------+----------+
```

(b) Utilization Estimates

```
* Summary:
+-----------------+---------+-----+--------+-------+-----+
|       Name      | BRAM_18K| DSP |   FF   |  LUT  | URAM|
+-----------------+---------+-----+--------+-------+-----+
|DSP              |        -|    -|       -|      -|    -|
|Expression       |        -|    -|       0|    712|    -|
|FIFO             |        -|    -|       -|      -|    -|
|Instance         |        0|    5|     384|    751|    -|
|Memory           |       13|    -|      64|      5|    -|
|Multiplexer      |        -|    -|       -|    494|    -|
|Register         |        -|    -|     703|     96|    -|
+-----------------+---------+-----+--------+-------+-----+
|Total            |       13|    5|    1151|   2058|    0|
+-----------------+---------+-----+--------+-------+-----+
|Available        |      280|  220|  106400|  53200|    0|
+-----------------+---------+-----+--------+-------+-----+
|Utilization (%)  |        4|    2|       1|      3|    0|
+-----------------+---------+-----+--------+-------+-----+

+ Detail:
    * Instance:
    +---------------------------------+---------------------------------+---------+----+-----+-----+-----+
    |             Instance            |              Module             | BRAM_18K| DSP|  FF | LUT | URAM|
    +---------------------------------+---------------------------------+---------+----+-----+-----+-----+
    |CONTROL_BUS_s_axi_U              |CONTROL_BUS_s_axi                |        0|   0|   36|   40|    0|
    |fadd_32ns_32ns_32_5_full_dsp_1_U1|fadd_32ns_32ns_32_5_full_dsp_1   |        0|   2|  205|  390|    0|
    |fmul_32ns_32ns_32_4_max_dsp_1_U2 |fmul_32ns_32ns_32_4_max_dsp_1    |        0|   3|  143|  321|    0|
    +---------------------------------+---------------------------------+---------+----+-----+-----+-----+
    |Total                            |                                 |        0|   5|  384|  751|    0|
    +---------------------------------+---------------------------------+---------+----+-----+-----+-----+
```

Fig. 1. HLS Report in default condition

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

Report

(1) the design latency in cycles, and

(a) Performance Estimates

```
* Summary:
+---------+---------+----------+----------+--------+--------+---------+
|  Latency (cycles) |  Latency (absolute) |     Interval    | Pipeline|
|   min   |   max   |    min   |    max   |   min  |   max  |   Type  |
+---------+---------+----------+----------+--------+--------+---------+
|   228022|   228022|  2.280 ms|  2.280 ms|  228023|  228023|     none|
+---------+---------+----------+----------+--------+--------+---------+

+ Detail:
    * Instance:
    N/A

    * Loop:
    +--------------+---------+---------+----------+-----------+-----------+------+----------+
    |              |  Latency (cycles) | Iteration|  Initiation Interval  | Trip |          |
    |   Loop Name  |   min   |   max   |  Latency |  achieved |   target  | Count| Pipelined|
    +--------------+---------+---------+----------+-----------+-----------+------+----------+
    |- LOAD_OFF_1  |        5|        5|         1|          -|          -|     5|        no|
    |- LOAD_W_1    |     1300|     1300|       130|          -|          -|    10|        no|
    | + LOAD_W_2   |      128|      128|         1|          -|          -|   128|        no|
    |- LOAD_I_1    |     1040|     1040|       130|          -|          -|     8|        no|
    | + LOAD_I_2   |      128|      128|         1|          -|          -|   128|        no|
    |- L1          |   225536|   225536|     28192|          -|          -|     8|        no|
    | + L2         |    28190|    28190|      2819|          -|          -|    10|        no|
    |  ++ L3       |     2816|     2816|        11|          -|          -|   256|        no|
    |- STORE_O_1   |      136|      136|        17|          -|          -|     8|        no|
    | + STORE_O_2  |       15|       15|         3|          -|          -|     5|        no|
    +--------------+---------+---------+----------+-----------+-----------+------+----------+
```

(b) Utilization Estimates

```
* Summary:
+-----------------+---------+-----+--------+-------+-----+
|       Name      | BRAM_18K| DSP |   FF   |  LUT  | URAM|
+-----------------+---------+-----+--------+-------+-----+
|DSP              |        -|    -|       -|      -|    -|
|Expression       |        -|    -|       0|    503|    -|
|FIFO             |        -|    -|       -|      -|    -|
|Instance         |        0|    5|     384|    751|    -|
|Memory           |       14|    -|      64|      5|    -|
|Multiplexer      |        -|    -|       -|    376|    -|
|Register         |        -|    -|     369|      -|    -|
+-----------------+---------+-----+--------+-------+-----+
|Total            |       14|    5|     817|   1635|    0|
+-----------------+---------+-----+--------+-------+-----+
|Available        |      280|  220|  106400|  53200|    0|
+-----------------+---------+-----+--------+-------+-----+
|Utilization (%)  |        5|    2|      ~0|      3|    0|
+-----------------+---------+-----+--------+-------+-----+

+ Detail:
    * Instance:
    +-----------------------------------+-------------------------------+---------+----+-----+-----+-----+
    |              Instance             |             Module            | BRAM_18K| DSP|  FF | LUT | URAM|
    +-----------------------------------+-------------------------------+---------+----+-----+-----+-----+
    |CONTROL_BUS_s_axi_U                |CONTROL_BUS_s_axi              |        0|   0|   36|   40|    0|
    |fadd_32ns_32ns_32_5_full_dsp_1_U1  |fadd_32ns_32ns_32_5_full_dsp_1 |        0|   2|  205|  390|    0|
    |fmul_32ns_32ns_32_4_max_dsp_1_U2   |fmul_32ns_32ns_32_4_max_dsp_1  |        0|   3|  143|  321|    0|
    +-----------------------------------+-------------------------------+---------+----+-----+-----+-----+
    |Total                              |                               |        0|   5|  384|  751|    0|
    +-----------------------------------+-------------------------------+---------+----+-----+-----+-----+
```

Fig. 2. HLS Report with pipelining explicitly disabled

(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).

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)

Vitis High-Level Synthesis User Guide