**NATIONAL CHENG KUNG UNIVERSITY**

**College of Electrical Engineering and Computer Science**

**DEPARTMENT OF ELECTRICAL ENGINEERING**

**VLSI System Design (Graduate Level)**

**Fall 2021**

**Summary of Final Project**

**Please don’t just write yes/no if there need more details,** **and use single-sided printing**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Simulate at SoC(yes/no)** | | | **Not yet** | | | | | |
| **Upload 5-min media(yes/no)** | | | **yes** | | | | | |
| **Basic** | | | | | | | | |
| **MCU** | **Pipeline** | | **Stage** | | **Max working Freq.** | | | **Data Width** |
| **5 stage** | | **77 MHz** | | | **32 bits** |
| **Number of Instructions** | | **45** | | | | | |
| **Realized Cache Specification** | | **L1 size：1KB**  **Write policy：write through、write around**  **(e.g. L1 size, associative, read/write policy)** | | | | | |
| **Cache Hit Rate of each program** | | **Inst data**  **Prog0: 0.992 0.648**  **Prog1: 1.000 0.932**  **Tpu\_bias: 0.995 0.013**  **Tpu\_input0 0.996 0.042**  **Tpu\_input0\_s1 as above**  **Tpu\_input0\_s2 as above**  **Tpu\_input1 as above**  **Tpu\_input1\_s1 as above**  **Tpu\_input1\_s2 as above**  **Tpu\_moving\_q as above**  **Tpu\_output\_q 0.996 0.068**  **Tpu\_output\_raw 0.097 0.042**  **Tpu\_weight 0.995 0.041**  **Tpu\_pw 0.998 0.799**  **Tpu\_layer0 0.992 0.831** | | | | | |
| **List of Realized Forwarding in Types and Stages** | | **Read-after-write data hazard implemented in EX-EX, MEM-EX, WB-EX stages.**  **(e.g. which kind)** | | | | | |
| **Realized Performance Counters (IPC) of each program** | | **Just like requirements in CSR in hw4，counting cycle and number of instruction** | | | | | |
| **Interrupt mechanism** | | **增加PLIC，達到處理多個interrupt(sensor, DMA TPU)的目的，以memory-mapped方式enable需要的interrupt source，PLIC做相關的判斷後，傳送interrupt訊號給CPU，並且CPU可以load PLIC的位子，得到現在interrupt的來源** | | | | | |
| **Memory** | **On-chip memory**  **(Total size <= 320KB)** | | **IM** | | | **DM** | | |
| **64KiB** | | | **64KiB** | | |
| **Off-chip memory** | | **SDRAM** | | | **ROM** | | |
| **32MiB** | | | **16KiB** | | |
| **ASPU** | **Max working Freq.** | | **77MHz** | | | | | |
| **Processing speed (throughput or… )** | | * **Maximum 5.4GOPS of int8 MACC.** * **727.4 MOPs/s for 3x3 convolution and 669 MOPs/s for point-wise convolution.** | | | | | |
| **Realized Specification of Functionalities in details** | | * **8x9 Systolic array to provide 72 peak MACC ops/cycle** * **Input fetcher supports 1-to-9 input feature map expansion** * **Two sets of output SRAMs to support 1-cycle partial sum generation with SRAM taking 1 cycle for read and write respectively.** * **2 sets of input SRAMs to provide ping-pong mechanism.** * **To support ping-pong mechanism, independent input and computation controls are implemented.** * **To perform requantization from 24-bit-integer partial sum to 8-bit-integer feature map, by reusing one of input SRAM as look-up table.** | | | | | |
| **Comparison with other works if any** | | **D** | | | | | |
| **BUS** | **Specify Memory and I/O mapping** | | **Slave** | **Start address** | | | **End address** | |
| **ROM** | **0x0000\_0000** | | | **0x0000\_3FFF** | |
| **IM** | **0x0001\_0000** | | | **0x0001\_FFFF** | |
| **DM** | **0x0002\_0000** | | | **0x0002\_FFFF** | |
| **SCTRL** | **0x1000\_0000** | | | **0x1000\_03FF** | |
| **DRAM** | **0x2000\_0000** | | | **0x21FF\_FFFF** | |
| **DMA** | **0x3000\_0000** | | | **0x3000\_0FFF** | |
| **TPU** | **0x3000\_1000** | | | **0x3000\_100F** | |
| **PLIC** | **0x4000\_0000** | | | **0x4000\_3FFF** | |
| **Implemented Features of AXI Bus, Level of Realization, Operating Frequency,**  **Outstanding number** | | **Read/write從req到resp皆是2個cycle可以完成**  **Level：RTL**  **Frequency：同CPU**  **Outstanding：1** | | | | | |
| **System** | **Specify** **Cooperation between CPU, Bus, Memory, ASPU and others** | | 1. **CPU would be boosted with program compiled with layer information** 2. **CPU command TPU to prepare for incoming data, e.g. input feature map, bias, weight, zero point, requantize threshold, etc., and then command DMA to move data to TPU.** 3. **If there are multiple layers of input feature map to perform convolution, TPU can utilize ping-pong buffer to receive further incoming input data while computing.** 4. **Later, after accumulation of multiple layers, CPU would command TPU to be prepared for being read by DMA, and later ask DMA to deliver output feature map in TPU back to DRAM.** 5. **Loop back to step 2 until the layer is finished.** | | | | | |
| **Specify Hardware interrupt & Interrupt service routines** | | **Hardware部分在MCU的部分已有說明**  **ISR：透過load PLIC某位子得到interrupt source，根據該src會在ISR執行對應部分的程式，主要是對其他slave做clear的動作**  **(>2 kind, and how they work)** | | | | | |
| **Specify Mechanism for Booting from an external ROM** | | **給DMA src, target, length，以DMA協助搬運boot資料，從DRAM搬至IM和DM** | | | | | |
| **Specify Realized DMA(Direct Memory Access) and Usage** | | **此次Project的DMA因應TPU的需求進行了特殊的設計，如DMA支援Unaligned Memory Access，且支援Discontinuous Address Stride，如此可以Tile的傳輸在一次的Command之內就完成，大量減少CPU的負擔。** | | | | | |
| **Code analysis (Superlint)** | | | **151/16275 = 0.93% 1 - 0.87% = 99.07%**  **(should >99% error free)** | | | | | |
| **System w/ ASPU (yes/no)** | | | **yes** | | | | | |
|  | | | **Synthesis** | | | **APR** | | |
| **clock period** | | | **12.4** | | | **13.2** | | |
| **Power** | | |  | | | **789.4698 mW** | | |
| **Area** | | | **12350777.9** | | | **18385385.76** | | |
| **Verification** | **MCU** | **prog0 pass ratio** | **100%** | | | | | |
| **ASPU** | **# and types of Direct test or constrained random test** | **We use 13 different programs to verify our design, including**   1. **To write weight into weight SRAM on TPU and verify the data after transmission using DMA.** 2. **To write input data into input SRAM1 of TPU without convolution expansion.** 3. **To write input data into input SRAM1 of TPU with stride-1 3x3 sliding window expansion.** 4. **To write input data into input SRAM1 of TPU with stride-2 3x3 sliding window expansion.** 5. **To write input data into input SRAM0 of TPU without convolution expansion.\*** 6. **To write input data into input SRAM0 of TPU with stride-1 3x3 sliding window expansion.\*** 7. **To write input data into input SRAM0 of TPU with stride-2 3x3 sliding window expansion.\*** 8. **To write bias and zero point into corresponding registers in TPU.** 9. **To write quantization thresholds into input SRAM0 of TPU.\*\*** 10. **To read raw output partial sum from TPU in 24-bit format.** 11. **To read quantized output feature map from TPU in 8-bit format.**   **\*\* Input SRAM0 would have different behavior when acting as**  **threshold look-up table.** | | | | | |
| **Specify types, length, operation conditions of benchmarks** | 1. **For testbench 1~11, unit tests are performed to check if all components are working well.** 2. **For testbench 12 and 13, a portion of application is applied as benchmark to examine the correctness of execution of this system.** | | | | | |
| **S**  **Y**  **S**  **T**  **EM** | **prog0 PR**  **simulation time** | **823218 NS** | | | | | |
| **prog1 PR**  **simulation time** | **4627035600 PS** | | | | | |
| **Specify types, length, operation conditions of benchmarks** | 1. **To compute whole layer of 3x3 convolution with 224x224 input feature map in MobileNet V2 (layer 1).** 2. **To compute whole layer of point-wise convolution with 112x112 input feature map in MobileNet V2 (layer 2).**   **\* Input SRAM0 has multiple banks to act as threshold look-up**  **table, thus needs different test suite.** | | | | | |

|  |  |
| --- | --- |
| **Advanced** | |
| **Synthesize AXI bus with burst and fully work with IPs** | **No** |
| **30 more instructions** | **No** |
| **64-bit add/sub, store/load** | **No** |
| **I/O PADs** | **No** |
| **More cache (L2 or L3)** | **No** |
| **dynamic branch prediction** | **No** |
| **CRT for more than two IPs** | **No** |
| **floating-point co-processor** | **No** |
| **Bootable by an operating system** | **No** |
| **Verify with FPGAs, specify FPGA board, what module has been put on the board and how you confirm results** | **No** |
| **Other Properties, please specify** |  |
| **References** |  |