ECE 260B Final Project Report - Winter 2024

Team: CoMingtOShort

Deliverables Paths

Step	Path
1	/home/linux/ieng6/ee260bwi24/hojin/ECE260B_FinalPr oject/Grace/step1/deliverable
2	/home/linux/ieng6/ee260bwi24/hojin/ECE260B_FinalPr oject/Grace/step2/deliverable
3	/home/linux/ieng6/ee260bwi24/hojin/ECE260B_FinalPr oject/Grace/step3/deliverable
4	/home/linux/ieng6/ee260bwi24/hojin/ECE260B_FinalPr oject/Grace/step4/deliverable
5	/home/linux/ieng6/ee260bwi24/hojin/ECE260B_FinalPr oject/Grace/step5/deliverable

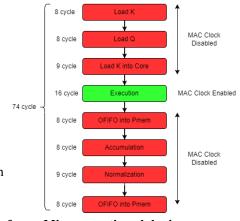
Introduction

In the context of expanding demands for machine learning capabilities, enhancing the adaptability, performance, and energy consumption of ML accelerators is essential. The need for higher performance per watt becomes important for scalability. To address this, our designed accelerator aims to achieve higher throughput and energy efficiency compared to traditional scalar processing architectures. Our design process started with a single-core design. After we verified the functionality and efficacy of the architecture, we moved onto a dual-core design for larger computations. Finally, we analyzed the power, timing and area reports of this design and implemented multiple optimization techniques in order to improve the frequency, minimize the power consumption, and minimize the area of our design.

Alpha1 - Clock Gating

To optimize power consumption, clock gating techniques were strategically implemented in our design. Analysis of the testbench revealed that the MAC Array submodule is actively utilized solely during the execution stage. Therefore, we implemented clock gating logic for the MAC Array, effectively disabling the clock signal to this submodule when its functionality is not required. This approach ensures that the MAC Array remains in an non-switching state and does not consume dynamic power during the remaining operational stages. As a result, clock gating ensures power efficiency without compromising functional correctness.

To quantify the impact of clock gating on power optimization, we conducted a comparative analysis across three design configurations: the synthesis power, the pnr power, and the power with clock-gated vcd. As we can see from the below table, both synthesis and pnr power are both relatively high at above 600mW. However, a significant improvement was achieved after implementing clock gated VCD, which demonstrated a remarkable 73.97% reduction in total



power consumption, consuming only 200.45 mW during the measured operational stages. As Professor Mingu mentioned during our poster presentation, the VCD power calculation excluded the data loading stages to reflect realistic power consumption.

Units (mW)	Internal Power	Switching Power	Leakage Power	Total Power	Power reduction
Synthesis	96.9033	671.0560	2.0138	769.9726	N/A
PNR	353.80515673	247.43905255	3.44975380	604.69396316	21.5%
Clk-Gated VCD	155.92022002	40.99203742	3.53522366	200.44747981	73.97%

Table 2 - Power consumption table (excluding loading phases)

Alpha2 - Dual-Port Psum Memory

Observing the operation workflow, several opportunities to improve our throughput were identified. Originally, during the division process of normalization, psum needed to be read from pmem (Psum Memory) and then written back to ofifo after normalization. Another stage is required to move the result from ofifo to pmem. To improve the throughput and increase the parallelism, we decided to enhance the psum memory architecture by converting it to a dual-port design, enabling write and read operations at the same time. This is done by providing two separate address pointers as inputs for these two operations. Now during normalization, data is read and the result is written to psum memory in the same cycles. As a result, the total clock cycle needed for the two operations was reduced by 8 cycles. After implementing both the changes, we verified the functionality of the design and proved that this improved throughput by over 10%. The throughput data was measured by counting the total cycles needed for the K*Q operation (including normalization).

Stage	Cycles
load K	8
load Q	8
load k into core	9
Execution	16
Move OFIFO to Pmem	8
Accumulation	8
Normalization	9
Move OFIFO to Pmem	8→0

Alpha3 - Multi-Cycle Path

In order to improve the frequency performance, based on observation of the timing report, it is suggested that the time critical path is in the division stage during the normalization process. However, we cannot apply pipelining to it since it is a single operation. Thus we applied a three-cycle multicycle path to the division paths in order to achieve a 1GHz clock frequency.

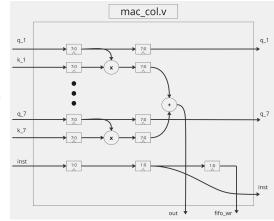
WNS before Multi-Cycle Path	WNS after Multi-Cycle Path
-1.716 ns	-0.025 ns

As we can see in the above table, after the Multi-Cycle path is applied, we obtained a worst negative slack of -0.025ns, compared to -1.716ns before. Most importantly, the new critical path is no longer the division path. That means we can further improve our operation frequency by applying other techniques in other areas of the chip. Therefore, this technique not only improved our performance, but also enabled us for further optimization opportunities.

Alpha4 - Pipelining

After applying the multi cycle pathing to the division operation in the normalization process. The critical path change to the MAC performed for each column. The primary reason for applying pipelining to the design is reducing the propagation delay from the registers through both the multiplier and addition. The pipelining occurs between the multiplication and addition stages while also pipelining the q and instruction values to prevent the values from being passed to the next column early. We saw after applying the pipelining that the critical path goes back to the divider so we decided not to further pipeline the additional stages as those are now off the critical path. We successfully verified with the testbench the functionality of the design.

WNS before Pipelining	WNS after Pipelining
(with Multi-Cycle Path)	(with Multi-Cycle Path)
-0.025ns	0 ns



Alpha5 - Flattened pnr

To further optimize our design's area cost, we decided to flatten the pnr and achieved a significant area saving.

Chip area using hierarchical pnr	Chip area using flattened pnr	Area reduction
1856400.00 μm²	699731.04 μm²	62.3 %

Alpha6 - DRC check

We have used eco techniques to achieve 0 DRC violations in our designs. Please see the right screenshot from final design pnr.

```
VERIFY DRC ..... Sub-Area : 25 of 25
VERIFY DRC ..... Sub-Area : 25 complete 0 Viols.

Verification Complete : 0 Viols.

*** End Verify DRC (CPU: 0:00:54.6 ELAPSED TIME: 55.00 MEM: 74.6M) ***
```

Conclusion

As discussed in the above sections, multiple design optimization techniques were implemented in the design. We have explored many techniques to improve various aspects of the design metrics from improving the throughput, power consumption, to frequency performance. During the project, we learned to balance the trade-off between latency and throughput, power, area and performance. However, by combining multiple different techniques, we are able to systematically improve both performance and reduce power in our final iteration. At the end, we finally met the 1GHz clock frequency with significant power reduction and throughput enhancement. As a team, we learned a valuable lesson regarding the VLSI design flow from RTL to GDS. Future improvements to the design include changing the architecture to a 2D systolic array where the data K and Q values are loaded simultaneously. Additionally, using pipelining to couple stages together execution and memory loading operations can also help performance.

Appendix

Step1. Single core RTL, Synthesis & PNR

Waveform vcd from gate-level sim

Gate-level sim console output



Step2. Output normalization

• Waveform vcd from behavioral sim snapshot



Testbench output

```
##### read normalized for Q*K output from pmem ######
Result Matched! Memory read from PSUM mem address:
Result Matched! Memory read from PSUM mem address:
Result Matched! Memory read from PSUM mem address:
                                                                                       0 data: 0001fffffaffffe0001fffff0ffff1fffebfffe4, expect: 0001fffffaffffe0001fffff0fffffffffe4ffe4
                                                                                                    0001affff6fffdc0000cffffeffff0fffecffff5, expect: 0001affff6fffdc0000cffffeffff0fffecffff5
000130000d000190001fffff20000affffffffe5, expect: 000130000d000190001fffff20000affffffffe5
                                                                                           data:
Result Matched! Memory read from PSUM mem address:
Result Matched! Memory read from PSUM mem address:
                                                                                                     00010fffdb00008003efffeeffff8fff3fffe7, expect: 0001affffdfffd00004fffd2ffff4fffe2fffee, expect:
                                                                                                                                                                                         00010fffdb000080003efffeeffff8ffff3fffe7
                                                                                                                                                                                         000laffffdfffd00004fffd2ffff4fffe2fffee
                                                                                           data:
Result Matched! Memory read from PSUM mem address:
Result Matched! Memory read from PSUM mem address:
                                                                                                    0002a000010001a000440000700007ffff8fffd3, expect: 0002a000010001a00044000070007ffff8fffd3
ffff70001800031ffffd000000000b00017ffff6, expect: ffff70001800031ffffd00000000b00017ffff6
                                                                                           data:
                                                                                           data:
Result Matched! Memory read from PSUM mem address:
                                                                                           data:
                                                                                                     fffeb0000d0002afffff100000001a000180000c.expect: fffeb0000d0002afffff100000001a000180000c
###### read V*N output from pmem #######
Result Matched! Memory read from PSUM mem address: 0
data: 0000b0003200006000606fff7dfffc800056
Result Matched! Memory read from PSUM mem address: 1
                                                                                           expect: 0000b0003e0003c0006e00066fff7dfffc800056
                                                                                           expect: 000450002b0005300001000130005d000840001b
            000450002b0005300001000130005d000840001b
Result Matched! Memory read from PSUM mem address: 2 data: fff1fff61fffd1fffb9fffdffff9dff73fffaa Result Matched! Memory read from PSUM mem address: 3 data: fff32fff36ffeebfff7dfffa0fff50fff3dfff51
                                                                                           expect: fff1ffffe1fffd1fffb9fffdffff9dfff73fffaa
                                                                                           expect: fff32fff36ffeebfff7dfffa0fff50fff3dfff51
 Result Matched! Memory read from PSUM mem address:
data: fffa5fffadfffd5fffa7fffb0000ldfffeffffc5
                                                                                           expect: fffa5fffadfffd5fffa7fffb0000ldfffeffffc5
 data:
Result Matched! Memory read from PSUM mem address:
data: ffeelfff49ffeecfff3dfff4afff0bfff2afff2d
Result Matched! Memory read from PSUM mem address:
data: 0001bfffbbfffd2fffedffff50008300052fffee
                                                                                           expect: ffee1fff49ffeecfff3dfff4afff0bfff2afff2d
                                                                                          expect: 0001bfffbbfffd2fffedffff50008300052fffee
 Result Matched! Memory read from PSUM mem addres
data: fff9ffffdcfff84fffd1fff90fff61fff42fffa7
                                                                                           expect: fff9ffffdcfff84fffd1fff90fff61fff42fffa7
```

Step3. Hierarchical Synthesis of core

• Waveform vcd from gate-level sim snapshot



Console output

```
###### read normalized for Q*K output from pmem ######
Result Matched! Memory read from PSUM mem address: 0 da
Result Matched! Memory read from PSUM mem address: 1 da
Result Matched! Memory read from PSUM mem address: 2 da
Result Matched! Memory read from PSUM mem address: 3 da
Result Matched! Memory read from PSUM mem address: 4 da
Result Matched! Memory read from PSUM mem address: 5 da
Result Matched! Memory read from PSUM mem address: 6 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 7 da
Result Matched! Memory read from PSUM mem address: 8 da
Result Matched! Memory read from PSUM mem address: 9 da
Result Matched! Memory read from PSUM mem address: 9 da
Result Matched! Memory read from PSUM mem address: 9 da
Result Matched! Memory read from PSUM mem address: 9 da
Result Matched! Memory read from PSUM mem address: 9 da
Result Matched! Memory read from PSUM mem address: 9 da
Result Matched! Memory read from PSUM mem address: 9 da
Result Matched! Result Memory read from PSUM mem address: 9 da
Result Matched! Result Memory read from PSUM mem address: 9 da
Result Matched! Result Memory read from PSUM mem address: 9 da
Result Matched! Result Memory read from PSUM mem address: 9 da
Result Matched! Result Memory read from PSUM memory read from P
                                                                                                                                                                                                                                                                                                                                                                                            0001fffffafffe0001fffff0fff1fffebfffe4, expect:
0001affff6fffd0000cffffefff0fffecffff5, expect:
000130000d000190001fffff20000afffffffe5, expect:
                                                                                                                                                                                                                                                                                                                                          0 data:
1 data:
2 data:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            0001fffffaffffe0001fffff0ffff1fffebfffe4
                                                                                                                                                                                                                                                                                                                                                                                              0001affff6fffd0000cffff6fff0fff6cffff5,
000130000d000190001fffff20000afffffffff6,
00010fff6db00008003efffeefff8ffffffffff.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            0001affff6fffd0000cffff6ffff0fffeeffff5
000130000d000190001fffff20000afffffffff6
00010ffffdb00008003efffeefff8fffffffff6
0001affffdfffd00004fffd2ffff4fffe2fffee
                                                                                                                                                                                                                                                                                                                                                         data:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          expect:
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        expect:
expect:
                                                                                                                                                                                                                                                                                                                                                        data:
                                                                                                                                                                                                                                                                                                                                                                                              0002a000010001a000440000700007ffff68ff63,
ffff70001800031ffffd00000000b00017ffff6,
fffeb0000d0002affff100000001a000180000c,
                                                                                                                                                                                                                                                                                                                                                        data
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            0002a000010001a000440000700007ffff8fffd3
ffff70001800031ffffd000000000b00017ffff6
                                                                                                                                                                                                                                                                                                                                                        data:
   ##### read V*N output from pmem ######
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    56
```

Result Matched! Memory read from PSUM mem address: data: 0000b0003e0003c0006e00066fff7dfffc800056		ect: 0000b0003e0003c0006e00066fff7dfffc800056
		ecc: 000000003e0003c0006e000661117d111c600036
Result Matched! Memory read from PSUM mem address:		
data: 000450002b0005300001000130005d000840001b		ect: 000450002b0005300001000130005d000840001b
Result Matched! Memory read from PSUM mem address:		
data: fff1ffffe1fffd1fffb9fffdffff9dfff73fffaa	exp	ect: fff1ffffe1fffd1fffb9fffdffff9dfff73fffaa
Result Matched! Memory read from PSUM mem address:		
data: fff32fff36ffeebfff7dfffa0fff50fff3dfff51	exp	ect: fff32fff36ffeebfff7dfffa0fff50fff3dfff51
Result Matched! Memory read from PSUM mem address:	4	
data: fffa5fffadfffd5fffa7fffb00001dfffeffffc5	exp	ect: fffa5fffadfffd5fffa7fffb00001dfffeffffc5
Result Matched! Memory read from PSUM mem address:		
data: ffee1fff49ffeecfff3dfff4afff0bfff2afff2d		ect: ffee1fff49ffeecfff3dfff4afff0bfff2afff2d
Result Matched! Memory read from PSUM mem address:		
data: 0001bfffbbfffd2fffedffff50008300052fffee		ect: 0001bfffbbfffd2fffedffff50008300052fffee
Result Matched! Memory read from PSUM mem address:	7 0/1	ecc. Goodbiilabiilabiilaaliilaaciila
data: fff9ffffdcfff84fffd1fff90fff61fff42fffa7	′	ect: fff9ffffdcfff84fffd1fff90fff61fff42fffa7
######################################		ecc: fili9fiffdcfff04fffdffff90fff61fff42fffa/
##############No error. Good Job!##############	****	

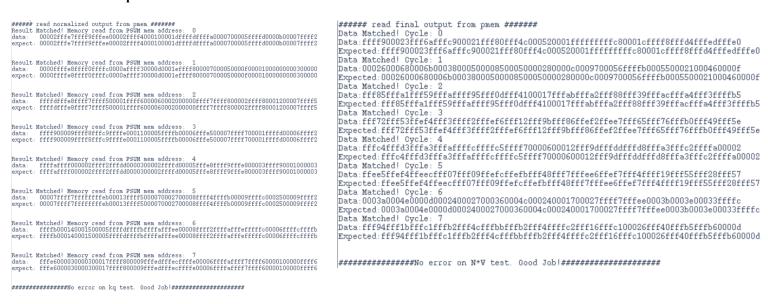
Team Members: Connor Kuczynski, Cory Huynh, Grace Jin, Mudi Huang

Step4. Hierarchical Synthesis of dual core

Waveform vcd from gate-level sim snapshot



Console output

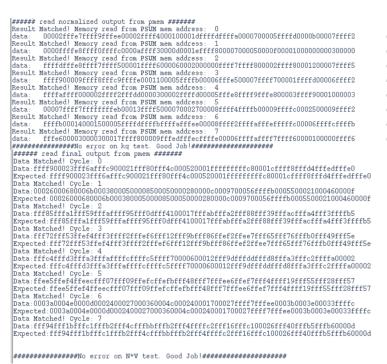


Step5. Optimization & PNR

Waveform vcd from gate-level sim snapshot



Console output



 Team Members: Connor Kuczynski, Cory Huynh, Grace Jin, Mudi Huang