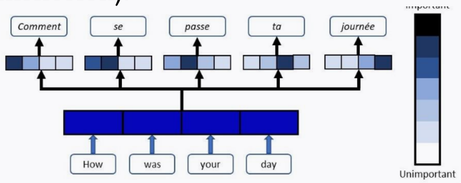
Dual Core Machine Learning Accelerator for Attention Mechanism

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

Chung Lam Tong, Han Zhao, Jiawen Xu, Jaewoo Kim

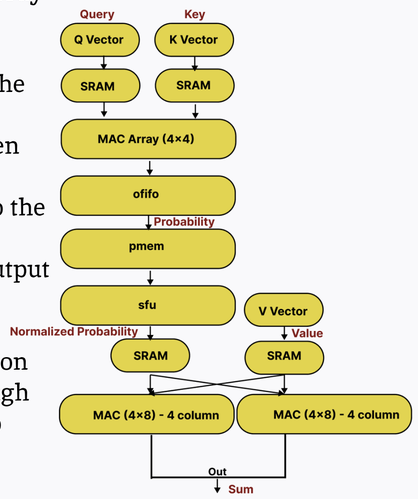
Motivation

Vector Processor is a common ASIC used for Attention Mechanism. The objective of this project is to implement vector processor hardware which able to process V_1 * V_2 each contain total of 8 - 4bit elements. Furthermore, Dual core method is introduced to increase the throughput(larger vector dimsnesion).



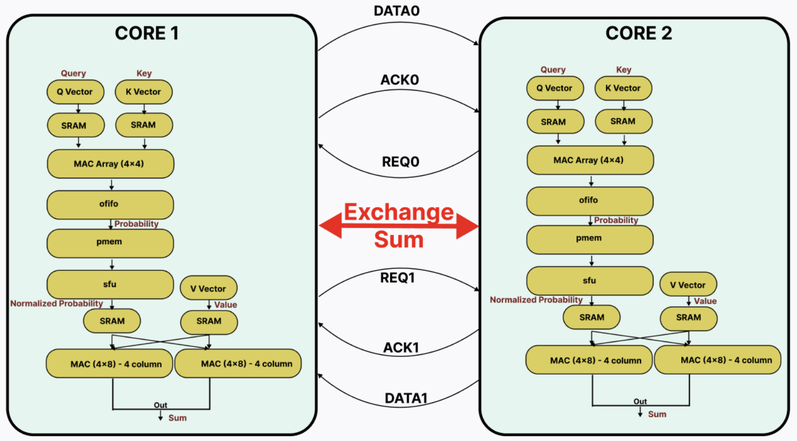
Methodologies(Single Core)

1. Processor loads the vector array into Query and Key
2. The Query and Key vector values are propagated into the 4*4 bit Mac_array
3. Compute dot product between Query and Key
4. Dot product results shift into the normalization stage with normalized probability as output results
5. The normalized probability results from the normalization stage needs to process through the 4*8 bit Mac unit again to multiply with the value and provide final result



Methodologies(Dual Core)

Dual Core method is introduced to increase the throughput (larger vector dimension). The elements of K will be divide into two core (8 elements each) and their sum data will be exchange through handshake medthod with synchronizer.



Methodologies (Pipelining and Parallelism)

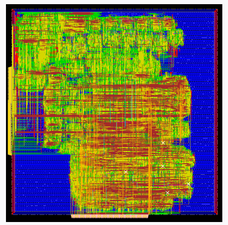
- **Pipelining**: One of the attempt we did for optimization is pipelining. We noticed the contamination delay in mac_8in.v is relatively high because of its complex combination circuit for multiplication. We added a pipeline to reduce the worst time slack.
- **Parallelism:** We applied the parallelism technique in the normalization stage. Instead of normalizing each column individually, we perform the computation of 8 normalization in parallel.

Result(Single Core)

Area (mm^2)	145347.119
Timing (Data Arrival time) (ns)	3.797
Timing (slack) (ns) @ 2 Ghz	-1.837
Total Dynamic Power (mW)	21.0282
Total Leakage Power(mW)	1.2212
Total Power(mW)	22.25

Result(Dual Core)

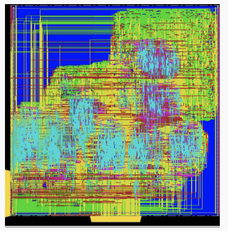
Area(mm^2)	296523.3597
Timing(Data Arrical time) (ns)	3.931
Timing (slack) (ns) @ 2Ghz	-2.023
Total Dynamic Power (mW)	42.7617
Total Leakage Power(mW)	2.4517
Total Power(mW)	45.2134



```
Verification Complete : 4 Viols.  0 Wrngs.

*********End: VERIFY GEOMETRY*********
*** verify geometry (CPU: 0:00:35.5  MEM: -5.0M)

******** End: VERIFY CONNECTIVITY ********
  Verification Complete : 0 Viols.  0 Wrngs.
  (CPU Time: 0:00:04.7  MEM: -0.066M)
```



```
 Verification Complete : 0 Viols.  0 Wrngs.

**********End: VERIFY GEOMETRY**********
 *** verify geometry (CPU: 0:00:35.9  MEM: 176.0M)

******** End: VERIFY CONNECTIVITY ********
  Verification Complete : 0 Viols.  0 Wrngs.
  (CPU Time: 0:00:04.2  MEM: -0.957M)
```

References

[1] M.Kang, “ECE 260B Winter 23 Project description” , la Jolla, 2023