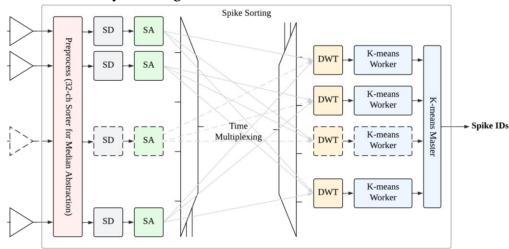
11020EE521800 Application Acceleration with High-Level Synthesis

Final Project – Spike Sorting Acceleration and Beamforming Acceleration

Team Member: 109061564 馬婕芸 110061551 呂易縉 110061540 賴聖耘

I. Spike Sorting Acceleration – Introduction

Spike sorting (SS) is the grouping of action potential (AP) into clusters based on the similarity of their shapes, and it is crucial step to extract information from extracellular recordings. With an emerging generation of high-density microelectrode arrays (MEAs) capable of recording spiking activity from thousands of neurons, reliable and scalable spike detection and analysis are needed. Historically, **spike sorting algorithm including spike detection (SD)**, **spike alignment (SA)**, **feature extraction (FE) and classification**. The whole SS system diagram is shown below.



Among these steps, the bottleneck is feature extraction and classification, discrete wavelet transform (DWT) and k-means clustering (K-means) in our system, since it needs massive computation resources. As a result, we **propose to design an end-to-end acceleration application for these two steps using Vitis tool**. As for SD and SA, we use MATLAB to do simulation and generate input data for DWT kernel. In this way, we can explore:

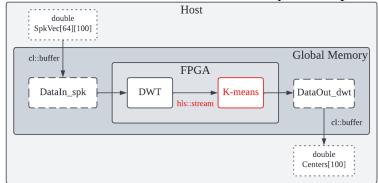
- i. Whether the proposed system is feasible or not?
- ii. Is K-means algorithm provided by Vitis library scalable or not?
- iii. How many resources will each kernel take?
- iv. What is the throughput of whole the SS system?

II. Spike Sorting Acceleration – SS System Description

The proposed system for DWT and K-means is shown below. SpkVec[][] is spike data, and it is preprocessed (detected and aligned) by MATLAB. Each spike has 64 points sampled at 32kHz, and each point is double. The input will then be buffered in global memory and serve as input for a DWT kernel. DWT kernel will do 4-layer convolution with an 8-point sym-4 wavelet shown below.

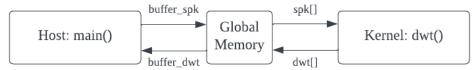
```
double HiD[8] = \{-0.0322, -0.0126, 0.0992, 0.2979, -0.8037, 0.4976, 0.0296, -0.0758\}; double LoD[8] = \{-0.0758, -0.0296, 0.4976, 0.8037, 0.2979, -0.0992, -0.0126, 0.0322\};
```

The result is 55-point DWT coefficient, and each point is double. And they will be streamed into K-means kernel. Its output is trained center of each clusters, and it will be copied in global memory again. In our system, we assumed there are 100 clusters. As a result, the system output is Centers[].

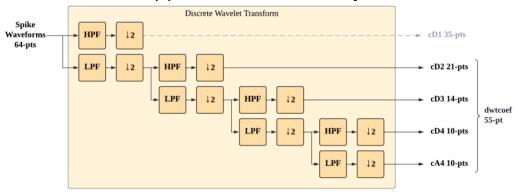


However, the whole system is not successfully built, since K-means is not applicable for spike data set. Also, the interface of each sub-function in Vitis K-means kernel is hard to modify. Alternatively, we built end-to-end acceleration application for DWT and K-means kernel respectively in order to make sure each kernel is feasible.

III. Spike Sorting Acceleration – DWT System Description



The system including host program, buffers for I/O (I: spk, O: dwt) in global memory and dwt function. The dwt function scheme is using lifting scheme shown below. Since each layer is doing the same thing, the whole function can be pipelined to enhance the latency.



IV. Spike Sorting Acceleration – K-means System Description

K-means clustering algorithm is unsupervised. The step-by-step details is shown below:

Algorithm 1 k-means algorithm

- 1: Specify the number k of clusters to assign.
- 2: Randomly initialize k centroids.
- 3: **repeat**
- 4: **expectation:** Assign each point to its closest centroid.
- 5: **maximization:** Compute the new centroid (mean) of each cluster.
- 6: until The centroid positions do not change.

Thanks to Vitis Libarary, it provides the HLS implementation of k-means clustering algorithm. Originally, it supports IRIS dataset (150 samples) with 4 features clustered into 3 clusters. The data format is shown below:

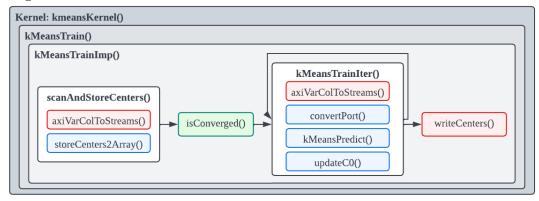
```
float irisVec[150][5] = {

{5.1, 3.5, 1.4, 0.2, 0}, {4.9, 3.0, 1.4, 0.2, 0}, {4.7, 3.2, 1.3, 0.2, 0}, {4.6, 3.1, 1.5, 0.2, 0}, {5.0, 3.6, 1.4, 0.2, 0}, {5.4, 3.9, 1.7, 0.4, 0}, {4.6, 3.4, 1.4, 0.3, 0}, {5.0, 3.4, 1.5, 0.2, 0}, {4.4, 2.9, 1.4, 0.2, 0}, {4.9, 3.1, 1.5, 0.1, 0}, {5.4, 3.7, 1.5, 0.2, 0}, {4.8, 3.4, 1.6, 0.2, 0}, {4.8, 3.0, 1.4, 0.1, 0}, {4.3, 3.0, 1.1, 0.1, 0}, {5.8, 4.0, 1.2, 0.2, 0}, {5.7, 4.4, 1.5, 0.4, 0}, {5.4, 3.9, 1.3, 0.4, 0}, {5.1, 3.5, 1.4, 0.3, 0}, {5.7, 3.8, 1.7, 0.3, 0}, {5.1, 3.8, 1.5, 0.3, 0}, {5.4, 3.4, 1.7, 0.2, 0}, {5.1, 3.7, 1.5, 0.4, 0}, {4.6, 3.6, 1.0, 0.2, 0}, {5.1, 3.3, 1.7, 0.5, 0}, {4.8, 3.4, 1.9, 0.2, 0}, {5.0, 3.0, 1.6, 0.2, 0}, {5.0, 3.4, 1.6, 0.4, 0}, {5.2, 3.5, 1.5, 0.2, 0}, {5.2, 3.4, 1.4, 0.2, 0}, {4.7, 3.2, 1.6, 0.2, 0}, {4.8, 3.1, 1.6, 0.2, 0}, {5.4, 3.4, 1.5, 0.4, 0},
```

In 32-channel SS system, the dwt coefficient has 55 features. Also, based on academic papers, each channel usually has $3\sim4$ clusters. As a result, we set cluster number = 100. And the data format is also shown below:

Here is the k-means system diagram. The k-means system including host program, DDR and kmeanskernel function. The interface is quite complicated. The blue blocks use HLS stream to communicate, and red blocks communicate with external DDR. The input of this function is data samples

with features, and the output is the centers of each clusters. We can then know which cluster a new data sample belongs to.



Each function description is shown below:

- ♦ kmeansKernel() is k-means clustering kernel called by host program.
- ♦ kmeansTrain() is k-means clustering function.
- ♦ kMeansTrainImp() is main HLS implementation of k-means clustering.
 - scanAndStoreCenters() scans data from DDR and store initial centers in local memory.
 - axiVarColToStreams() scans data (DWT coefficients) from DDR.
 - storeCenters2Array() stores centers to local memory.
 - isCoverage() decides whether there is any clusters should be merged.
 - kmeansTrainIter() is main k-means clustering function in each iteration.
 - axiVarColToStreams() scans data (DWT coefficients) from DDR.
 - convertPort() changes input data into various streams.
 - kMeansPredict() decides which cluster a new data sample belongs.
 - updateC0() update each cluster's center.
 - writeCenters() writes the centers back into DDR.

V. Spike Sorting Acceleration – Results & Discussion

- 1. DWT system:
 - i. Hardware Target: Resources

We can find that a single DWT kernel only takes up 3.76% LUT, 0.6% BRAM and 3.06% DSP in FPGA. The U50 platform may not suitable for up to 32 channels, since $100/3.76 \approx 27$.

Name	LUT	LUTAsMem	REG	BRAM	URAM	DSP
Platform	99966	8343	123023	178	0	4
∨ User Budget	770050	393673	1620337	1166	640	5936
Used Resources	28496	3751	33888	8	0	182
Unused Resources	741554	389922	1586449	1158	640	5754
v dwtkernel (1)	28496	3751	33888	8	0	182
dwtkernel_1	28496	3751	33888	8	0	182

ii. Hardware Target: Latency

We can find that the dwtkernel takes 4.070 us to generate data. The latency is much shorter than software simulation. After pipelining, the throughput can reach II=1.

Compute Unit		Module Name	Start Interval	Best (cycles)	Avg (cycles)	Worst (cycles)	Best (absolute)
dwtkernel_1	dwtkernel	dwtkernel_Pipeline_VITIS_LOOP_45_1	87	87	87	87	0.290 us
dwtkernel_1	dwtkernel	dwtkernel_Pipeline_VITIS_LOOP_61_2	253	253	253	253	0.843 us
dwtkernel_1	dwtkernel	dwtkernel_Pipeline_VITIS_LOOP_82_3	9	9	9	9	29.997 ns
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 99 4	80	80	80	80	0.267 us
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 124 5	9	9	9	9	29.997 ns
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 140 6	74	74	74	74	0.247 us
dwtkernel_1	dwtkernel	dwtkernel_Pipeline_VITIS_LOOP_165_7	9	9	9	9	29.997 ns
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 181 8	71	71	71	71	0.237 us
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 201 9	92	92	92	92	0.307 us
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 202 10	85	85	85	85	0.283 us
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 203 11	81	81	81	81	0.270 us
dwtkernel 1	dwtkernel	dwtkernel Pipeline VITIS LOOP 204 12	81	81	81	81	0.270 us
dwtkernel 1	dwtkernel	dwtkernel	1222	1221	1221	1221	4.070 us

iii. Hardware Target: Co-simulation

The co-simulation is successfully in target hardware stage.

Found Platform
Platform Name: Xilinx
INFO: Reading /users/course/20225/HLS
Loading: '/users/course/20225/HLS
hw: 0.00761115/ sw: 0.00761115
hw: -1.0289/ sw: -1.0289
hw: 1.40291/ sw: 1.40291
hw: -0.251538/ sw: -0.251538
hw: -0.00551868/ sw: -0.00551868
hw: 0.00084852/ sw: 0.00084852
hw: 0.358819/ sw: 0.358819
hw: -1.36879/ sw: -1.36879
hw: 1.18532/ sw: -1.3622907
hw: 1.16798/ sw: 1.6798
hw: -1.30458/ sw: -1.30458

2. K-means system:

i. Hardware Target: Resources

We can find that a single k-means kernel only takes up 21% LUT, 2% BRAM and 1% DSP in FPGA. The U50 platform may not suitable for up to 32 channels, since $100/21 \approx 5$. Compared to DWT system, K-means takes up much larger resources. The reason maybe the function is pipelined and paralleled to each high throughput. However, SS system does not need this characteristic.

4	4				
Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-		- -	 -	-
Expression	-1	-	0	240	-
FIFO	-1	-	-	-	-
Instance	60	64	60712	185161	0
Memory	-	-	1	77	10
Multiplexer		-	-	1240	-
Register	-	-	707		-
Total	60		61420		
Available SLR	1344		871680	435840	320
Utilization SLR (%)	4	2	7	42	
Available	2688	5952	1743360		
Utilization (%)	2	1	3	21	1

ii. Hardware Target: Latency

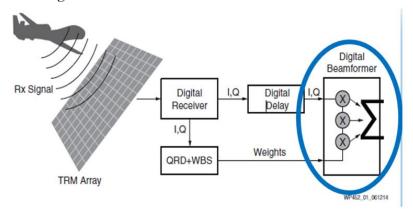
Latency Inform	mation			
Compute Unit	Kernel Name	Module Name	Start Interval	Best (cycles)
kmeansKernel	kmeansKernel	entry_proc	0	0
kmeansKernel	kmeansKernel	readRaw_512_32_64_Pipeline_READ_RAW	600003	600003
kmeansKernel	kmeansKernel	readRaw_512_32_64_s	4 ~ 600078	4
kmeansKernel	kmeansKernel	cageShitRight_512_64_s	1	0
kmeansKernel	kmeansKernel	varSplit_512_64_Pipeline_LOOP1_LOOP2	600005	600005
kmeansKernel	kmeansKernel	varSplit_512_64_s	600009	600009
kmeansKernel	kmeansKernel	axiVarColToStreams_32_512_64_s	600010 ~ 600079	600083
kmeansKernel	kmeansKernel	dupStrm_64_8_7_Pipeline_VITIS_LOOP_184_1	70002	70002
kmeansKernel	kmeansKernel	dupStrm_64_8_7_Pipeline_VITIS_LOOP_194_3	10	10
kmeansKernel	kmeansKernel	dupStrm_64_8_7_s	70016	70016
kmeansKernel	kmeansKernel	stream_n_to_one_read_64_8_Pipeline_VITIS_LOOP_111_2	undef	undef
kmeansKernel	kmeansKernel	stream_n_to_one_read_64_8_Pipeline_VITIS_LOOP_127_4	12	12
kmeansKernel	kmeansKernel	stream_n_to_one_read_64_8_s	undef	undef
kmeansKernel	kmeansKernel	stream_n_to_one_collect_64_64_8_Pipeline_VITIS_LOOP_174_1	undef	undef
kmeansKernel	kmeansKernel	stream_n_to_one_collect_64_64_8_s	undef	undef
kmeansKernel	kmeansKernel	stream_n_to_one_distribute_64_64_8_Pipeline_VITIS_LOOP_233_1	undef	undef
kmeansKernel	kmeansKernel	stream_n_to_one_distribute_64_64_8_Pipeline_VITIS_LOOP_252_2	10	10
kmeansKernel	kmeansKernel	stream_n_to_one_distribute_64_64_8_s	undef	undef
kmeansKernel	kmeansKernel	stream_n_to_one_round_robin_64_64_8_s	undef	undef
kmeansKernel	kmeansKernel	streamNToOne_64_64_8_s	undef	undef
kmeansKernel	kmeansKernel	split_64_1_55_Pipeline_VITIS_LOOP_163_1	550002	550002
kmeansKernel	kmeansKernel	split_64_1_55_s	550004	550004
kmeansKernel	kmeansKernel	storeCenters2Array_double_55_100_715_8_1_Pipeline_VITIS_LOOP_191_1	5725	5725
kmeansKernel	kmeansKernel	storeCenters2Array_double_55_100_715_8_1_s	5727	5727
kmeansKernel	kmeansKernel	scanAndStoreCenters_double_55_100_715_8_1_s	undef	undef
kmeansKernel	kmeansKernel	isConverged_double_55_100_715_8_1_Pipeline_LOOP1_LOOP2_LOOP3	45787	45787
		~		

iii. Hardware Target: Co-simulation

The co-simulation is failed in target hardware stage.

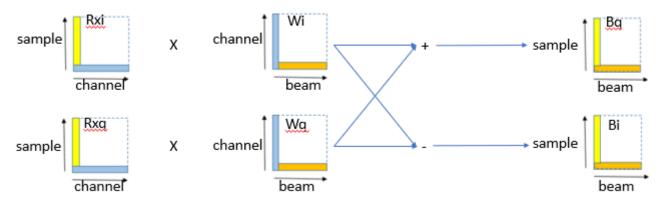
VI. Beamforming Acceleration – Introduction

Beamforming is a signal processing technique used in sensor arrays for directional signal transmission or reception. This is achieved by combining elements in an antenna array in such a way that signals at particular angles experience constructive interference while others experience destructive interference. Beamforming can be used for radio or sound waves. It has found numerous applications in radar, sonar, seismology, wireless communications, radio astronomy, acoustics and biomedicine. We will focus on the circled in the diagram below.



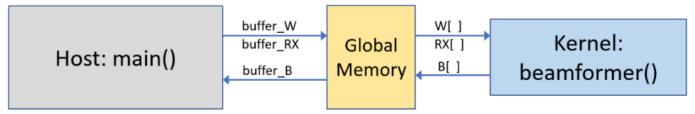
VII. Beamforming Acceleration – Function Description

The complex sensor data (RX_I and RX_Q) is delivered to the Beamformer module as a 2D array [SAMPLES][CHANNELS]. Similarly, the adaptive weights (W_I & W_Q) are also stored as a 2D array [BEAMS][CHANNELS]. Output Beam data (beams_i & beams_q) is computed for each sample [BEAMS][Sample]. The matrix calculation process is shown below. The number of SAMPLES represents the Pulse Repetition Interval (PRI). We need to achieve a PRI < 200 uS for WP452 application .



VIII. Beamforming Acceleration – System Description

The overall application structure is represented in this block diagram with all the Beamformer calculations done by the Kernel C++ code in 'beamformer()'.

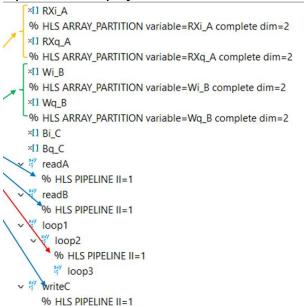


The dataflow between the Host and Kernel. The API commands in the HOST code will be spotlighted that control the data movement for steps 4-7 in the Execution model shown here.



IX. Beamforming Acceleration – HLS directive

In order to achieve higher performance, we tried lots of different pragmas. The array_partition directive is the most important because it can remove memory bottlenecks. We have attempt to use pipeline for read and write, also attempt to use unroll for matrix calulation. In the end, we use array_partition, and pipeline II=1 for three kind of loops. The directive we use and the latency&utilization is shown below. It is OK that utilization exceed 100% because we use pynq-z2 to test, but we will use x50 platform to deploy our hardware.



tency (cycles)	Latency (ab	solute)	Interv	al (cycl	es)
nin max	min	max	min	ma	x Typ
5088 55088	1.653 ms 1.	653 m:	5508	9 550	89 r
∃ Summary					
Name	BRAM_18K	DSP	FF	LUT	URAM
DSP		-	70	-	-
Expression	120	-	0	6	
FIFO		-		-	-
Instance	8	192	7197	7056	
Memory	288	-	0	0	C
Multiplexer	3.0	-	-	2154	
Register		-	404	-	-
Total	296	192	7601	9216	
Available	280	220 1	06400	53200	
Utilization (%)	105	87	7	17	

X. Beamforming Acceleration – Results & Discussion

1. Hardware Target: Resources

Name	LUT	LUTAsMem	REG	BRAM	URAM	DSP
Platform	11.54%	2.15%	7.10%	13.24%	0.00%	0.07%
∨ User Budget	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Used Resources	1.34%	0.22%	1.37%	13.64%	0.00%	3.23%
Unused Resources	98.66%	99.78%	98.63%	86.36%	100.00%	96.77%
√ beamformer (1)	1.34%	0.22%	1.37%	13.64%	0.00%	3.23%
beamformer_1	1.34%	0.22%	1.37%	13.64%	0.00%	3.23%

2. Hardware Target: Latency

Latency = 80073 clocks @ 300Mhz Latency absolute time = 267 uS

Timing Information (MHz)								
Compute Unit Kernel Nam	Module Name	Target Frequenc	y Estimated Fr	equency				
beamformer_1 beamformer	beamformer_Pipeline_readA	300.300293	411.015198					
beamformer_1 beamformer	beamformer_Pipeline_readB	300.300293	411.015198					
beamformer_1 beamformer	beamformer_Pipeline_loop1_loop2	300.300293	426.985474					
beamformer_1 beamformer	beamformer_Pipeline_writeC	300.300293	411.015198					
beamformer_1 beamformer	beamformer	300.300293	411.015198					
Latency Information								
Latency Information Compute Unit Kernel Nam	Module Name	Start Interval	Best (cycles)	Avg (cycles)	Worst (cycles)	Best (absolute)	Avg (absolute)	Worst (absolute)
Compute Unit Kernel Nam								
Compute Unit Kernel Nam beamformer_1 beamformer	beamformer_Pipeline_readA	80073	80073	80073	80073	0.267 ms	0.267 ms	0.267 ms
Compute Unit Kernel Nam beamformer_1 beamformer beamformer_1 beamformer	beamformer_Pipeline_readA beamformer_Pipeline_readB	80073 169	80073 169	80073 169	80073 169	0.267 ms 0.563 us	0.267 ms 0.563 us	0.267 ms 0.563 us
Computé Unit Kernel Nam beamformer_1 beamformer beamformer_1 beamformer beamformer_1 beamformer	beamformer_Pipeline_readA beamformer_Pipeline_readB beamformer_Pipeline_loop1_loop2	80073 169 7511	80073 169 7511	80073 169 7511	80073 169 7511	0.267 ms 0.563 us 25.034 us	0.267 ms 0.563 us 25.034 us	0.267 ms 0.563 us 25.034 us
Compute Unit Kernel Nam beamformer_1 beamformer beamformer_1 beamformer	beamformer_Pipeline_readA beamformer_Pipeline_readB beamformer_Pipeline_loop1_loop2	80073 169	80073 169	80073 169	80073 169	0.267 ms 0.563 us	0.267 ms 0.563 us	0.267 ms 0.563 us

XI. Work Distribution

- 1. Processed data preparation: 馬婕芸
- 2. C sources preparation (Spike Sorting): 馬婕芸
- 3. C sources preparation (Beamforming): 呂易縉、賴聖耘
- 4. Directives preparation: 馬婕芸、呂易縉、賴聖耘
- 5. HLS simulation: 馬婕芸、呂易縉、賴聖耘
- 6. Report writing: 馬婕芸、呂易縉、賴聖耘
- 7. Presentation: 馬婕芸、呂易縉、賴聖耘

XII. Reference

- 1. Spike Detection: https://ieeexplore.ieee.org/document/6070974
- 2. Spike Alignment: https://ieeexplore.ieee.org/document/1608524/
- 3. Feature Extraction: https://www.sciencedirect.com/science/article/pii/S0165027000002508
- 4. K-means Clustering: https://xilinx.github.io/Vitis_Accel_Examples/2020.2/html/kmeans.html
- 5. KiloSort: https://github.com/cortex-lab/KiloSort
- 6. Beamforming: https://www.xilinx.com/developer/articles/beamforming-acceleration.html

XIII. GitHub: https://github.com/jieyunma/AAHLS_FinalProject