Final

project presentation

Using HLS to accelerate Machine Learning algorithm

Presenters:
Zhixin TANG (z5142033)
Shuyi WANG (z5190769)

Contents

01 • Project overview

02 • Design

03 • Results

04 • Conclusion

PART ONE

- Project overview

01

Aims

Accelerating a k-nearest neighbours (KNN) algorithm in Machine Learning using high-level synthesis by a given dataset.

- Design the algorithm and performance evaluation
- Use HLS for acceleration

Challenges

- Algorithm improvement
    - accuracy comparison
- Handling different values of K
- Large computational intensity
    - limitations on acceleration

PART TWO

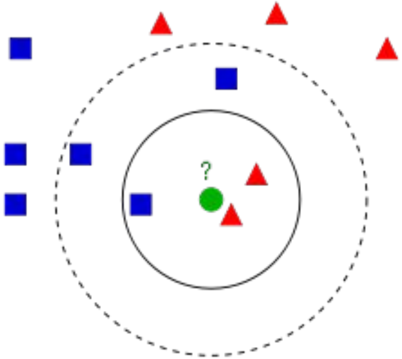
- Design

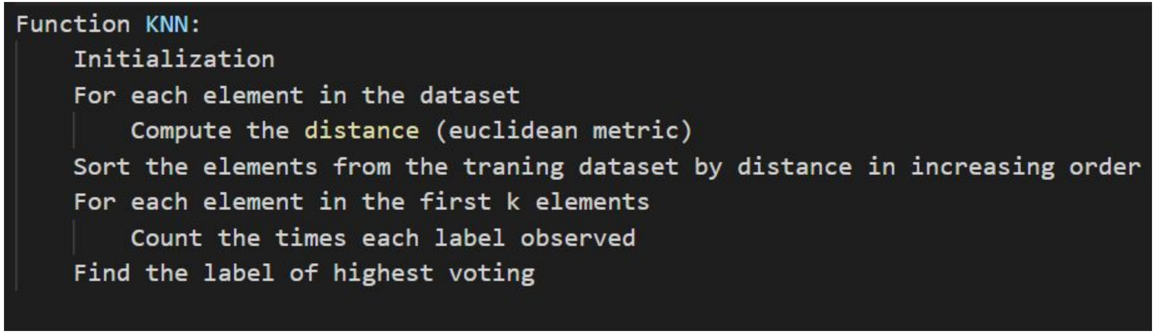
02

KNN

KNN is used for classification and regression

- Given training set, compute the accuracy of the testing set
- Fit the algorithm to improve accuracy





```
Function KNN:
    Initialization
    For each element in the dataset
        Compute the distance (euclidean metric)
    Sort the elements from the traning dataset by distance in increasing order
    For each element in the first k elements
        Count the times each label observed
    Find the label of highest voting
```

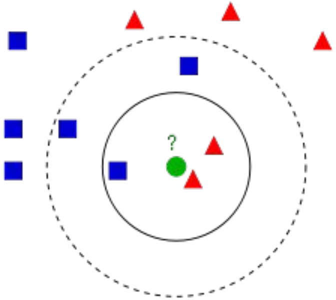
C++ code structure

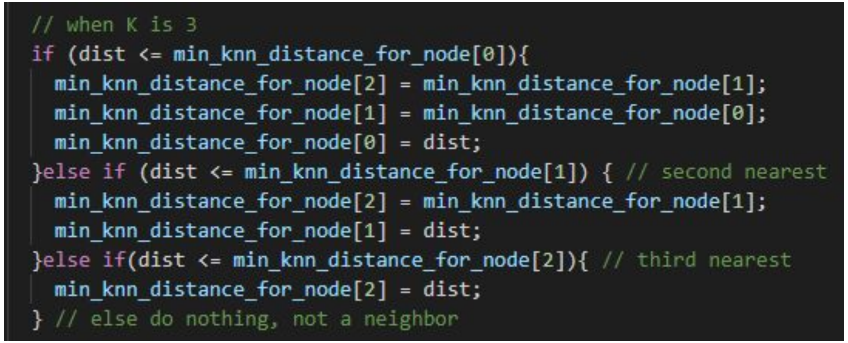
KNN Algorithm:

1: Distance Calculation

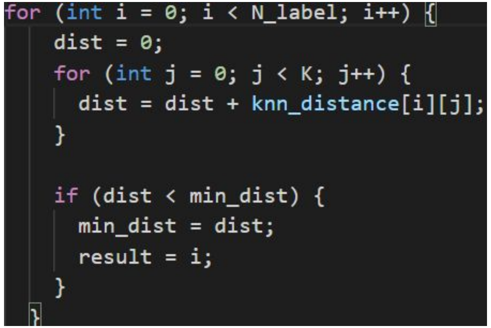
2: K-Nearest Neighbors Finder

3: Class Determination





```
// when K is 3
if (dist <= min_knn_distance_for_node[0]){
  min_knn_distance_for_node[2] = min_knn_distance_for_node[1];
  min_knn_distance_for_node[1] = min_knn_distance_for_node[0];
  min_knn_distance_for_node[0] = dist;
}else if (dist <= min_knn_distance_for_node[1]) { // second nearest
  min_knn_distance_for_node[2] = min_knn_distance_for_node[1];
  min_knn_distance_for_node[1] = dist;
}else if(dist <= min_knn_distance_for_node[2]){ // third nearest
  min_knn_distance_for_node[2] = dist;
} // else do nothing, not a neighbor
```

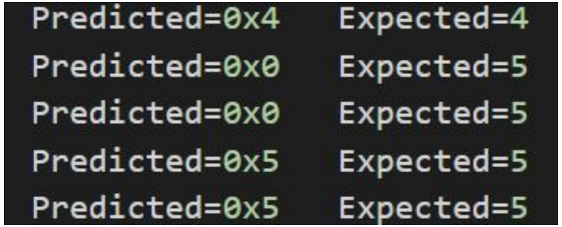


```
for (int i = 0; i < N_label; i++) {
    dist = 0;
    for (int j = 0; j < K; j++) {
      dist = dist + knn_distance[i][j];
    }

    if (dist < min_dist) {
      min_dist = dist;
      result = i;
    }
}
```

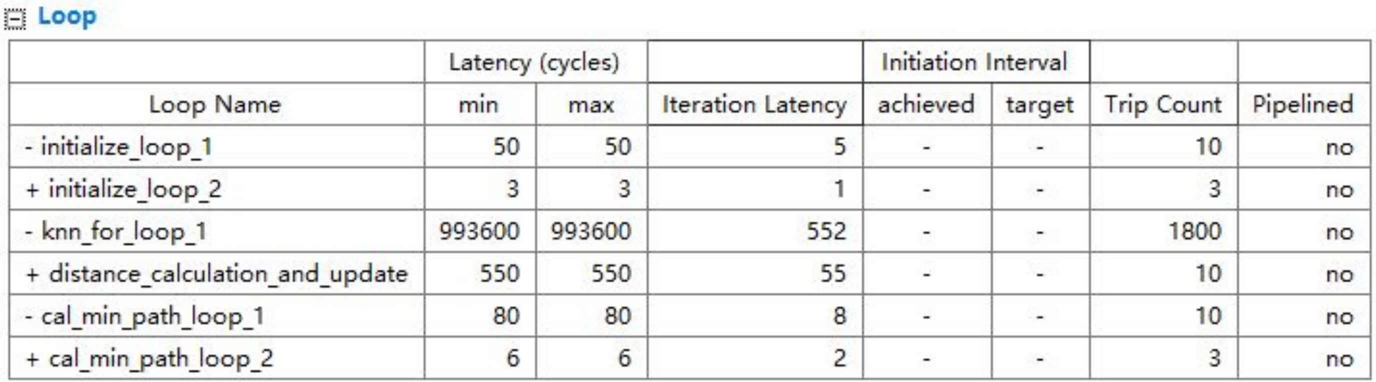
Testbench

- Read a dataset labeled by numbers
- Produce output to compare
- Accuracy = correct prediction / total prediction
  - around 90% in our results



```
Predicted=0x4   Expected=4
Predicted=0x0   Expected=5
Predicted=0x0   Expected=5
Predicted=0x5   Expected=5
Predicted=0x5   Expected=5
```

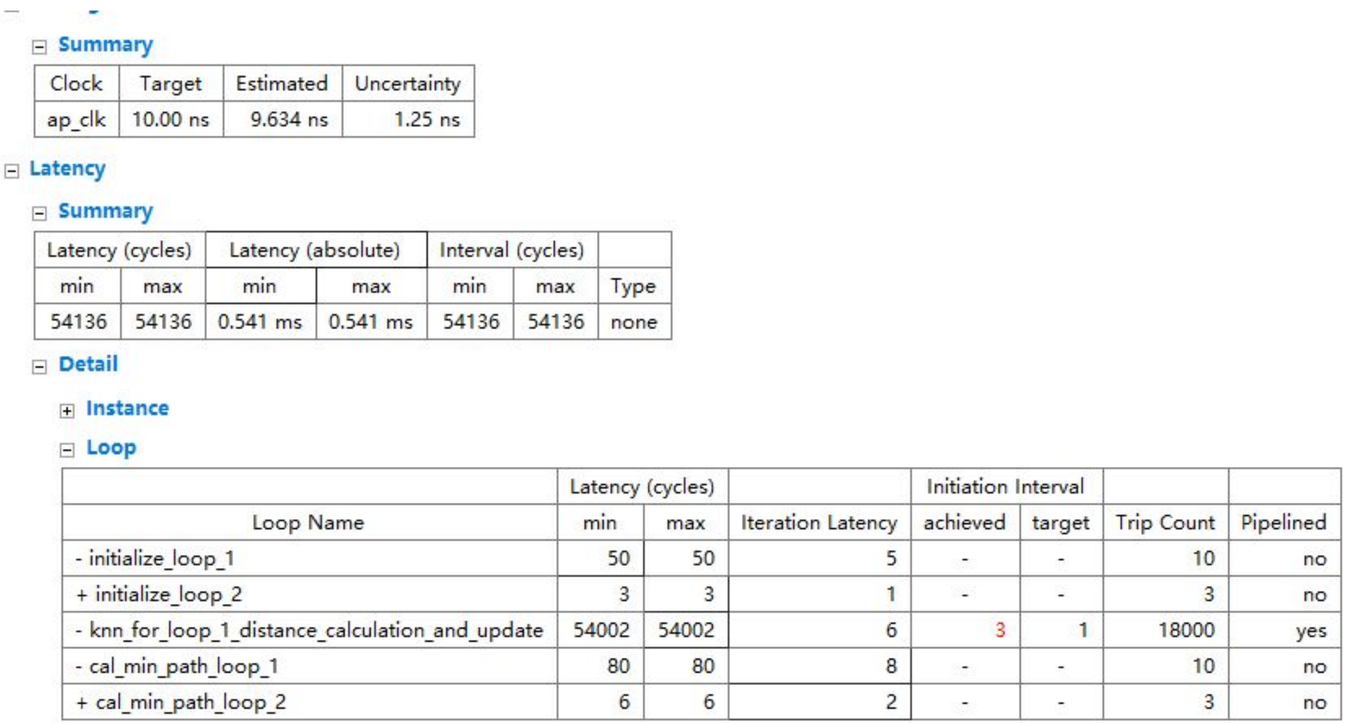
Baseline synthesis



Loop

	Latency (cycles)			Initiation Interval			
Loop Name	min	max	Iteration Latency	achieved	target	Trip Count	Pipelined
- initialize_loop_1	50	50	5	-	-	10	no
+ initialize_loop_2	3	3	1	-	-	3	no
- knn_for_loop_1	993600	993600	552	-	-	1800	no
+ distance_calculation_and_update	550	550	55	-	-	10	no
- cal_min_path_loop_1	80	80	8	-	-	10	no
+ cal_min_path_loop_2	6	6	2	-	-	3	no

Pipelining Distance Calculation Loop



⊟ **Summary**

Clock	Target	Estimated	Uncertainty
ap_clk	10.00 ns	9.634 ns	1.25 ns

⊟ **Latency**

⊟ **Summary**

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
54136	54136	0.541 ms	0.541 ms	54136	54136	none

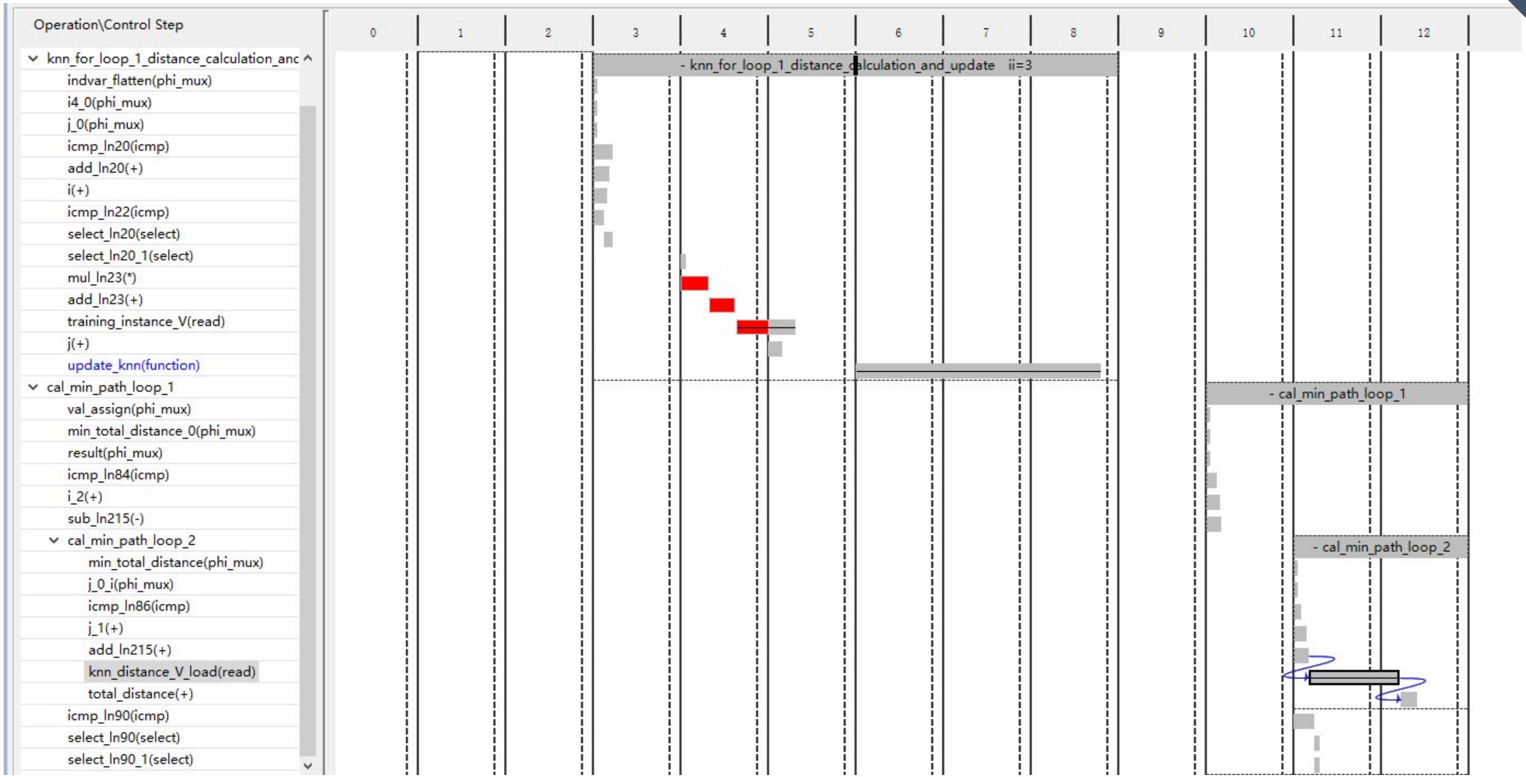
⊟ **Detail**

⊞ **Instance**

⊟ **Loop**

	Latency (cycles)			Initiation Interval			
Loop Name	min	max	Iteration Latency	achieved	target	Trip Count	Pipelined
- initialize_loop_1	50	50	5	-	-	10	no
+ initialize_loop_2	3	3	1	-	-	3	no
- knn_for_loop_1_distance_calculation_and_update	54002	54002	6	3	1	18000	yes
- cal_min_path_loop_1	80	80	8	-	-	10	no
+ cal_min_path_loop_2	6	6	2	-	-	3	no

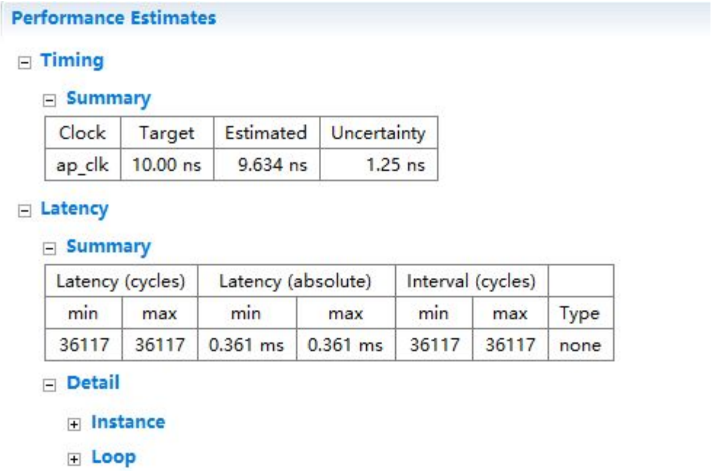
Pipelining Distance Calculation Loop



Operation\Control Step

- knn_for_loop_1_distance_calculation_anc
  - indvar_flatten(phi_mux)
  - i4_0(phi_mux)
  - j_0(phi_mux)
  - icmp_ln20(icmp)
  - add_ln20(+)
  - i(+)
  - icmp_ln22(icmp)
  - select_ln20(select)
  - select_ln20_1(select)
  - mul_ln23(*)
  - add_ln23(+)
  - training_instance_V(read)
  - j(+)
  - update_knn(function)
- cal_min_path_loop_1
  - val_assign(phi_mux)
  - min_total_distance_0(phi_mux)
  - result(phi_mux)
  - icmp_ln84(icmp)
  - i_2(+)
  - sub_ln215(-)
  - cal_min_path_loop_2
    - min_total_distance(phi_mux)
    - j_0_i(phi_mux)
    - icmp_ln86(icmp)
    - j_1(+)
    - add_ln215(+)
    - knn_distance_V_load(read)
    - total_distance(+)
  - icmp_ln90(icmp)
  - select_ln90(select)
  - select_ln90_1(select)

Array_partition



Performance Estimates

⊟ Timing

⊟ Summary

Clock	Target	Estimated	Uncertainty
ap_clk	10.00 ns	9.634 ns	1.25 ns

⊟ Latency

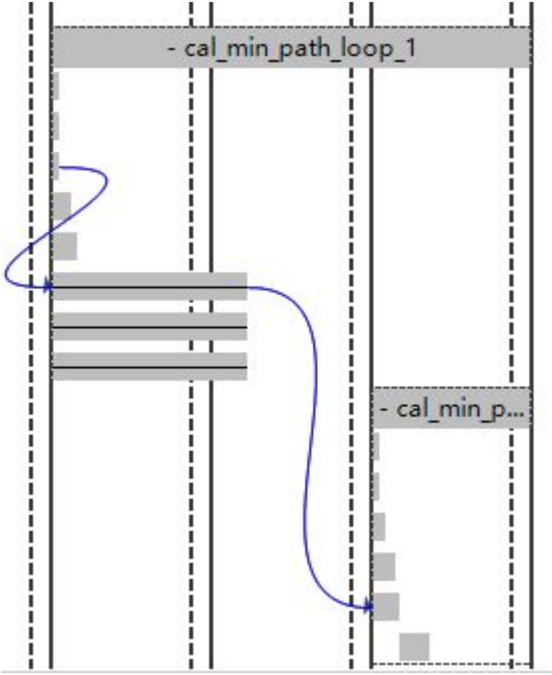
⊟ Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
36117	36117	0.361 ms	0.361 ms	36117	36117	none

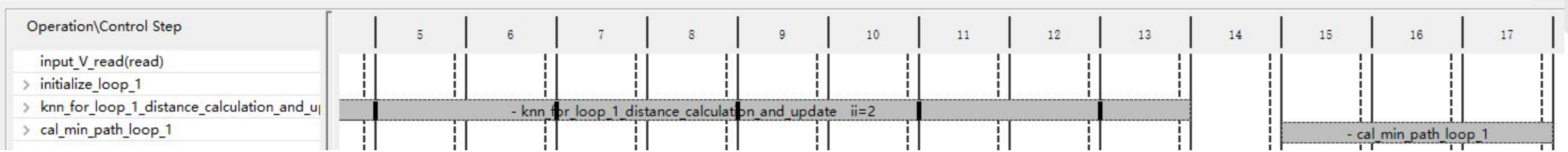
⊟ Detail

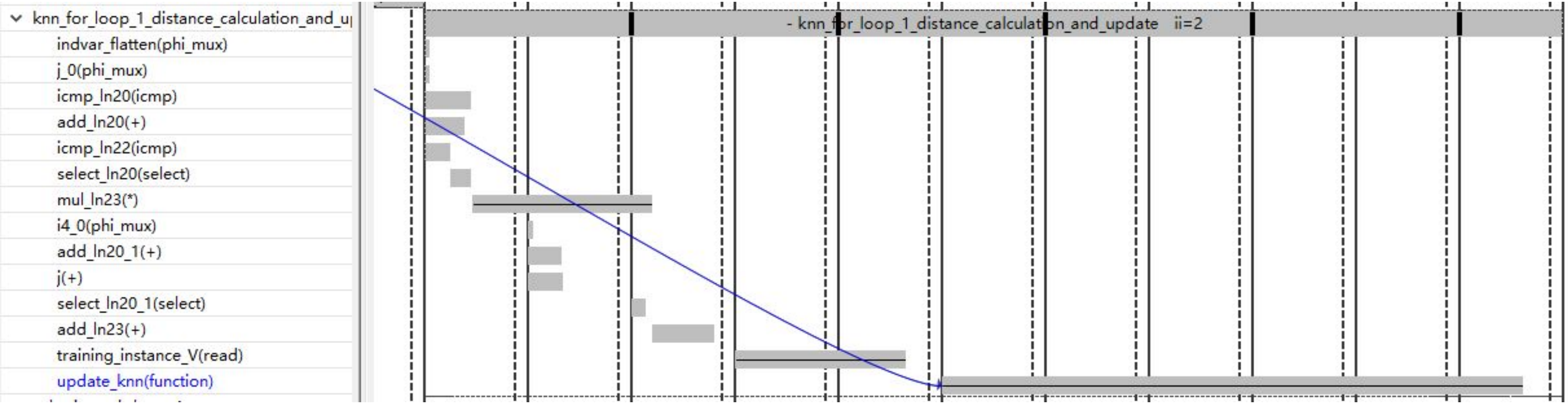
⊞ Instance

⊞ Loop



Shorten clock_period



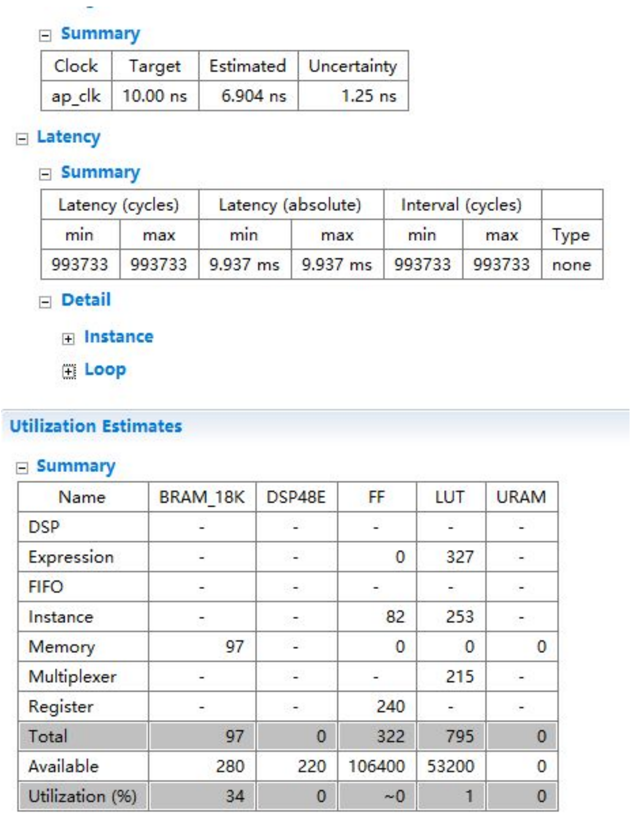


PART THREE

- Results

03

Baseline synthesis



⊟ Summary

Clock	Target	Estimated	Uncertainty
ap_clk	10.00 ns	6.904 ns	1.25 ns

⊟ Latency

⊟ Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
993733	993733	9.937 ms	9.937 ms	993733	993733	none

⊟ Detail

⊞ Instance

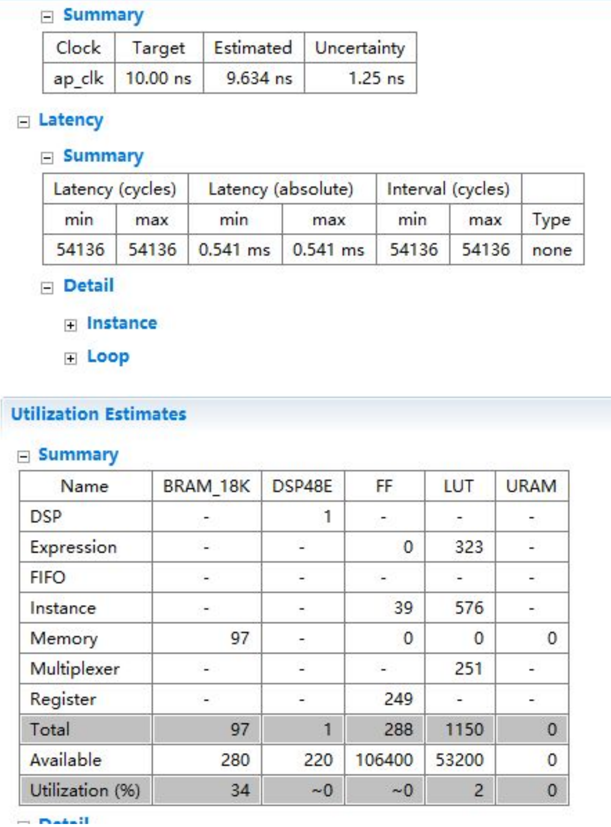
⊞ Loop

Utilization Estimates

⊟ Summary

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	327	-
FIFO	-	-	-	-	-
Instance	-	-	82	253	-
Memory	97	-	0	0	0
Multiplexer	-	-	-	215	-
Register	-	-	240	-	-
Total	97	0	322	795	0
Available	280	220	106400	53200	0
Utilization (%)	34	0	~0	1	0

Pipelining Distance Calculation



⊟ Summary

Clock	Target	Estimated	Uncertainty
ap_clk	10.00 ns	9.634 ns	1.25 ns

⊟ Latency

⊟ Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
54136	54136	0.541 ms	0.541 ms	54136	54136	none

⊟ Detail

⊞ Instance

⊞ Loop

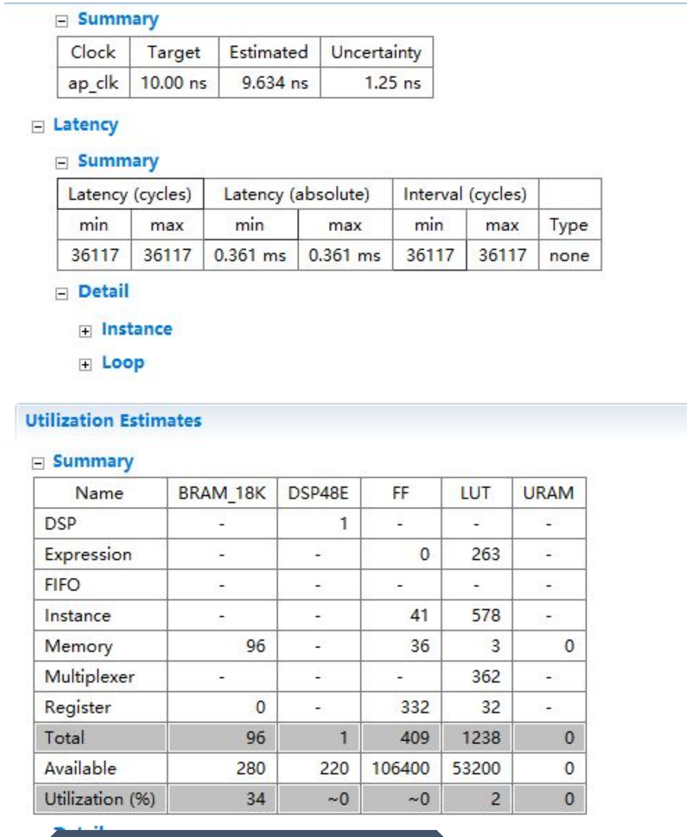
Utilization Estimates

⊟ Summary

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	1	-	-	-
Expression	-	-	0	323	-
FIFO	-	-	-	-	-
Instance	-	-	39	576	-
Memory	97	-	0	0	0
Multiplexer	-	-	-	251	-
Register	-	-	249	-	-
Total	97	1	288	1150	0
Available	280	220	106400	53200	0
Utilization (%)	34	~0	~0	2	0

⊟ Detail

Array_Partition knn_distance array



⊟ Summary

Clock	Target	Estimated	Uncertainty
ap_clk	10.00 ns	9.634 ns	1.25 ns

⊟ Latency

⊟ Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
36117	36117	0.361 ms	0.361 ms	36117	36117	none

⊟ Detail

⊞ Instance

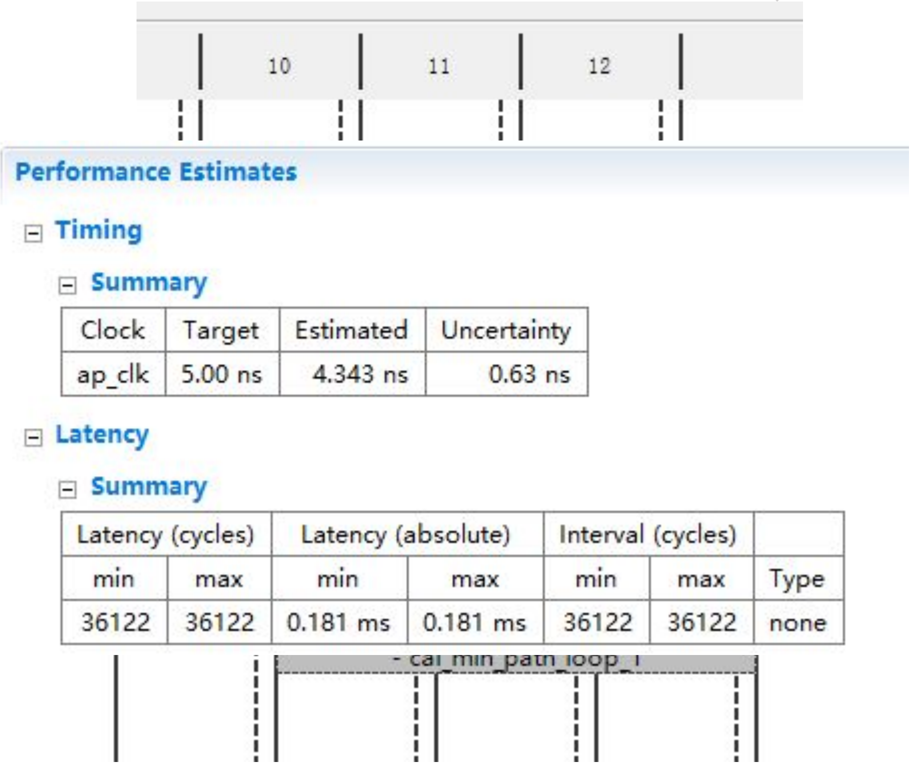
⊞ Loop

Utilization Estimates

⊟ Summary

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	1	-	-	-
Expression	-	-	0	263	-
FIFO	-	-	-	-	-
Instance	-	-	41	578	-
Memory	96	-	36	3	0
Multiplexer	-	-	-	362	-
Register	0	-	332	32	-
Total	96	1	409	1238	0
Available	280	220	106400	53200	0
Utilization (%)	34	~0	~0	2	0

Shorten clock_period



Performance Estimates

⊟ Timing

⊟ Summary

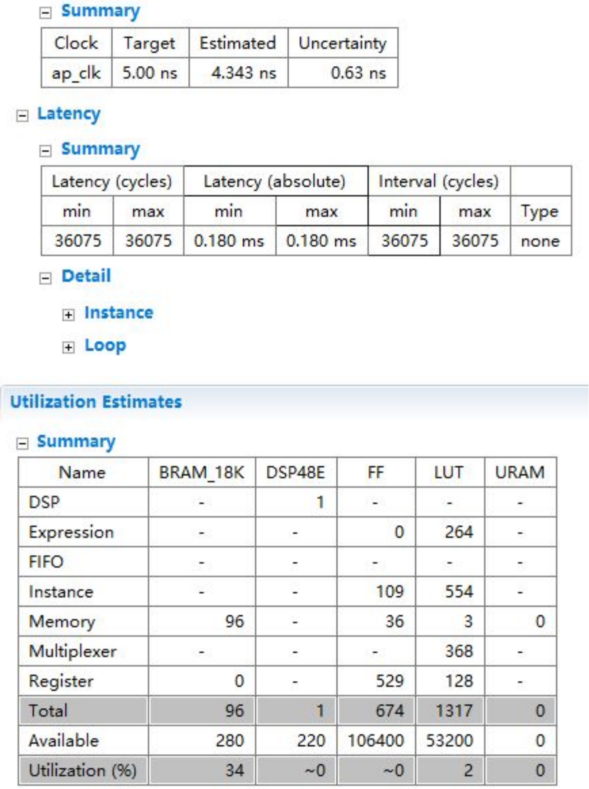
Clock	Target	Estimated	Uncertainty
ap_clk	5.00 ns	4.343 ns	0.63 ns

⊟ Latency

⊟ Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
36122	36122	0.181 ms	0.181 ms	36122	36122	none

Pipelining Classifiaction Determination



Summary

Clock	Target	Estimated	Uncertainty
ap_clk	5.00 ns	4.343 ns	0.63 ns

Latency

Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
36075	36075	0.180 ms	0.180 ms	36075	36075	none

Detail

Instance

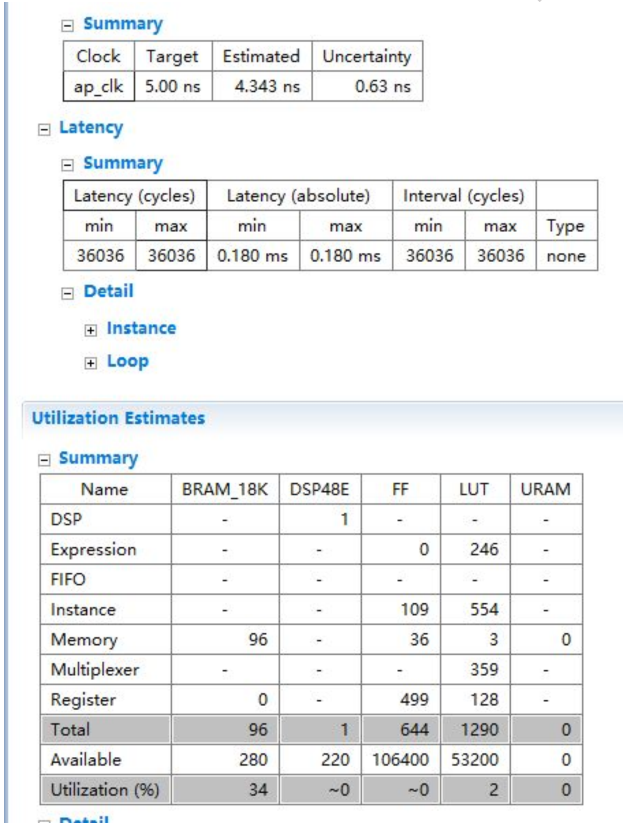
Loop

Utilization Estimates

Summary

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	1	-	-	-
Expression	-	-	0	264	-
FIFO	-	-	-	-	-
Instance	-	-	109	554	-
Memory	96	-	36	3	0
Multiplexer	-	-	-	368	-
Register	0	-	529	128	-
Total	96	1	674	1317	0
Available	280	220	106400	53200	0
Utilization (%)	34	~0	~0	2	0

Pipelining initialization



Summary

Clock	Target	Estimated	Uncertainty
ap_clk	5.00 ns	4.343 ns	0.63 ns

Latency

Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
36036	36036	0.180 ms	0.180 ms	36036	36036	none

Detail

Instance

Loop

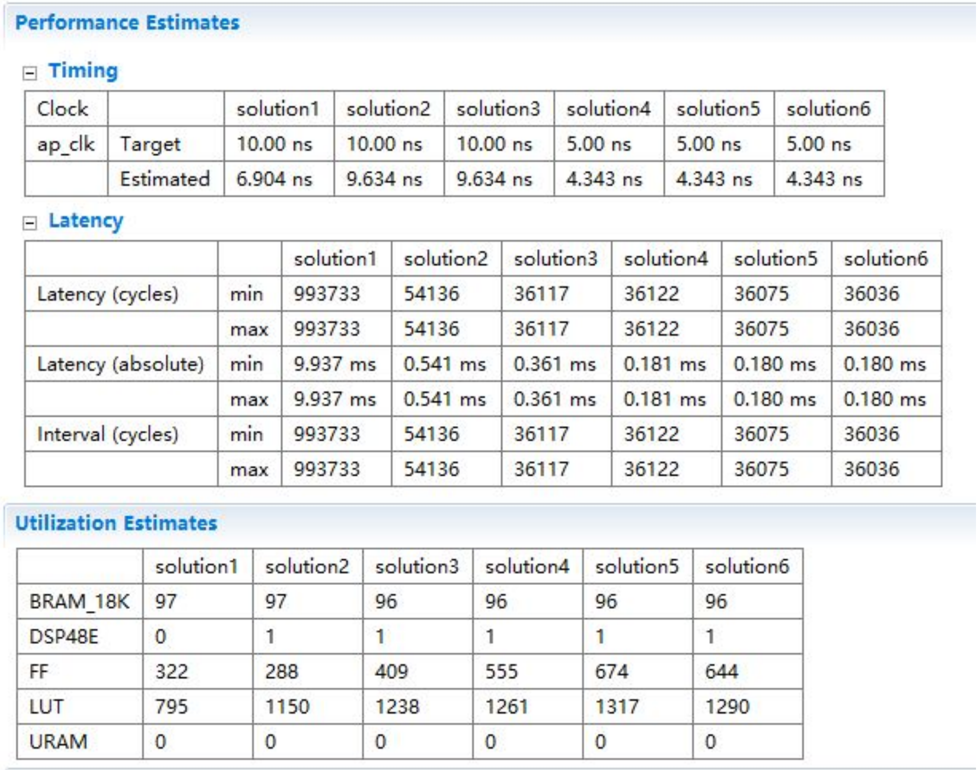
Utilization Estimates

Summary

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	1	-	-	-
Expression	-	-	0	246	-
FIFO	-	-	-	-	-
Instance	-	-	109	554	-
Memory	96	-	36	3	0
Multiplexer	-	-	-	359	-
Register	0	-	499	128	-
Total	96	1	644	1290	0
Available	280	220	106400	53200	0
Utilization (%)	34	~0	~0	2	0

Detail

Disscusion



Performance Estimates

⊟ Timing

Clock		solution1	solution2	solution3	solution4	solution5	solution6
ap_clk	Target	10.00 ns	10.00 ns	10.00 ns	5.00 ns	5.00 ns	5.00 ns
	Estimated	6.904 ns	9.634 ns	9.634 ns	4.343 ns	4.343 ns	4.343 ns

⊟ Latency

		solution1	solution2	solution3	solution4	solution5	solution6
Latency (cycles)	min	993733	54136	36117	36122	36075	36036
	max	993733	54136	36117	36122	36075	36036
Latency (absolute)	min	9.937 ms	0.541 ms	0.361 ms	0.181 ms	0.180 ms	0.180 ms
	max	9.937 ms	0.541 ms	0.361 ms	0.181 ms	0.180 ms	0.180 ms
Interval (cycles)	min	993733	54136	36117	36122	36075	36036
	max	993733	54136	36117	36122	36075	36036

Utilization Estimates

	solution1	solution2	solution3	solution4	solution5	solution6
BRAM_18K	97	97	96	96	96	96
DSP48E	0	1	1	1	1	1
FF	322	288	409	555	674	644
LUT	795	1150	1238	1261	1317	1290
URAM	0	0	0	0	0	0

- **Accelerate 96.37%**

PART FOUR

- Conclusion

04

Conclusion

Work Done:

- **Receive a 90% accuracy at K = 3**
- **Optimization the design**
    - **Pipelining Distance Calculation**
    - **Array_Partition in K Nearest Neighbors Finder**
    - **Pipelining Class Determination**
    - **Shorten Period Time**
- **Accelerate 96.37% comared to baseline synthesis**

Future improvement:

- **Improving Software code of K Nearest Neighbors Finder to handle different K values**
- **Use different size of datasets**

THANKS!