Development of DNN trigger for Belle II experiment

Shuangshuang Zhang, Qi-Dong Zhou

Institute of frontier and interdisciplinary science, Shandong University

12 Dec. 2024

CEPC TDAQ meeting

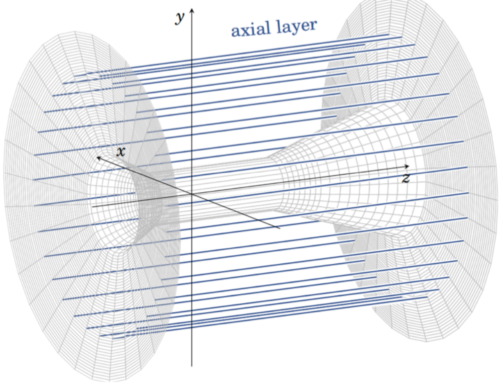
Outline

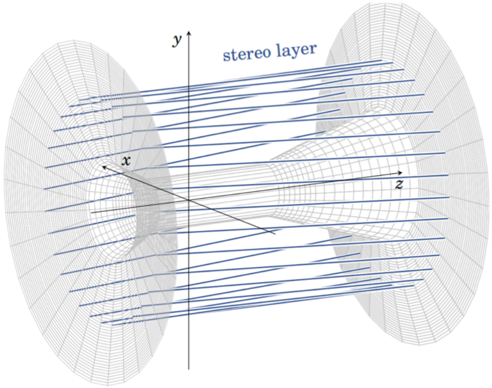
- Motivation
- Deep Neural Network (DNN) model
- Development of DNN model with python
- High Level Syntheis (HLS) with Vitis
- Testing IP cores in Vivado

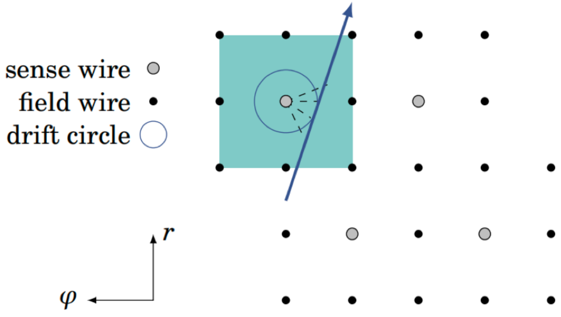
Motivation

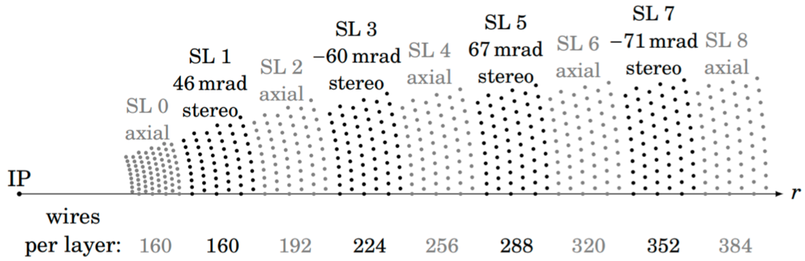
- Study the CDC DNN trigger and master the implementation process of DNN in FPGA
- Optimize the model and observe changes in performance
- PLAN: Investigating the possiblity of using more input variables in DNN models, such as ADC signals and momentum
- PLAN: Development of DNN model to Versal ACAP

Central drift chamber



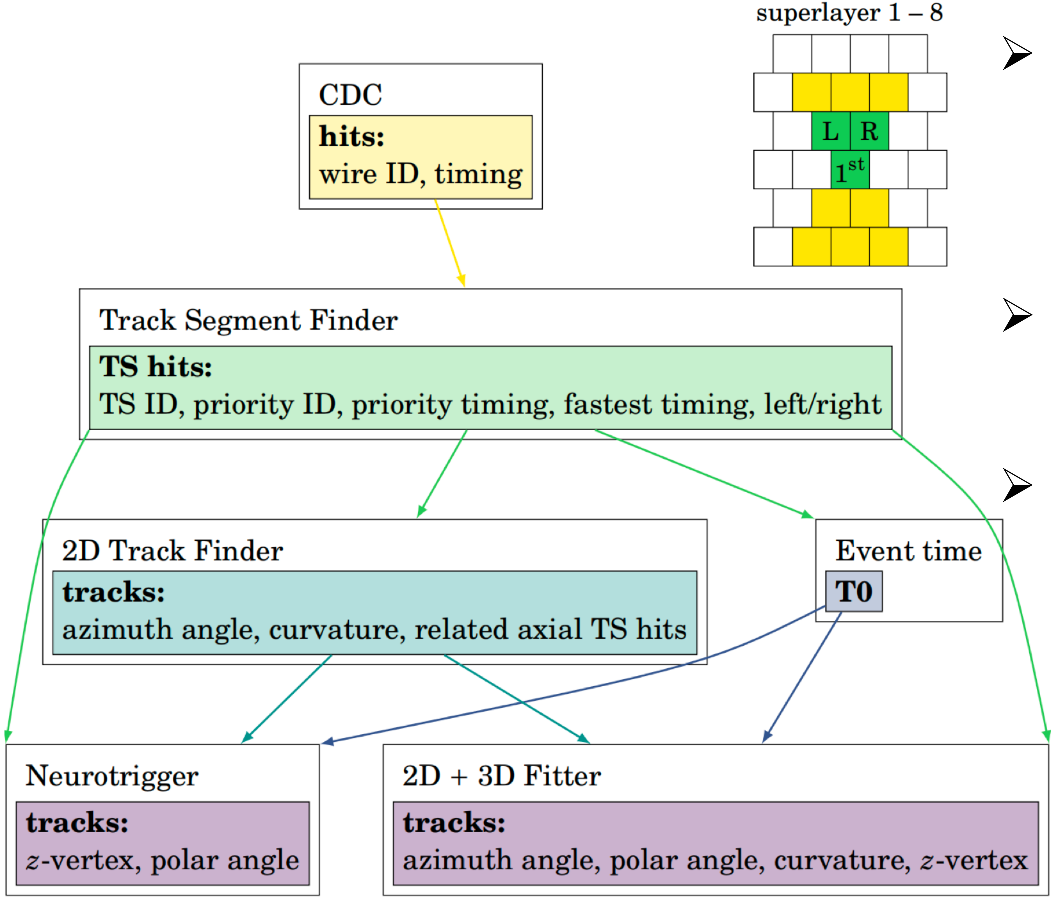






- A drift cell of the CDC is formed of one sense wire and eight field wires
- A hit on an axial wire provides coordinates in the transverse plane(2D)(Φ,r)
- By combining axial and stereo hits, a three dimensional track can be reconstructed(3D)

The CDC track trigger



CDC

hits:
wire ID, timing

Track Segment Finder

TS hits:
TS ID, priority ID, priority timing, fastest timing, left/right

2D Track Finder

tracks:
azimuth angle, curvature, related axial TS hits

Event time

T0

Neurotrigger

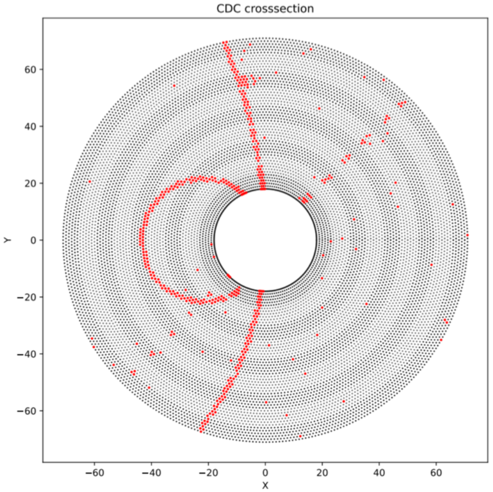
tracks:
z-vertex, polar angle

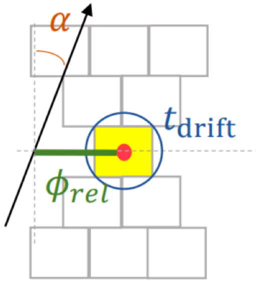
2D + 3D Fitter

tracks:
azimuth angle, polar angle, curvature, z-vertex

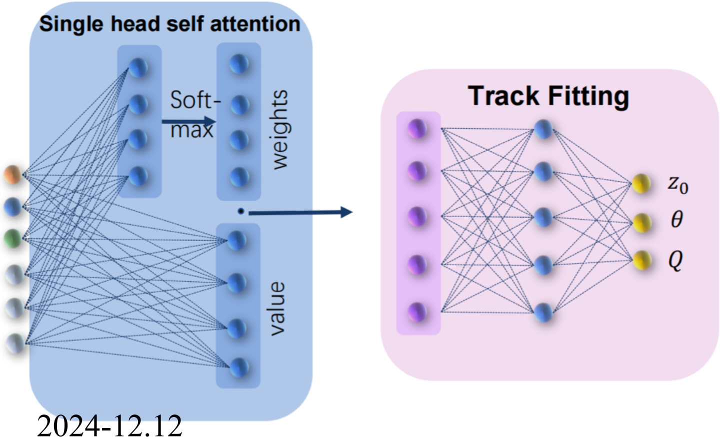
- The Belle II trigger consists of two levels: the first level trigger(hardware) and the high level trigger(software)
- CDC track trigger provides charged track information
- The track finding for the trigger is based on track segments(“TS”)

Deep Neural Network (DNN) model



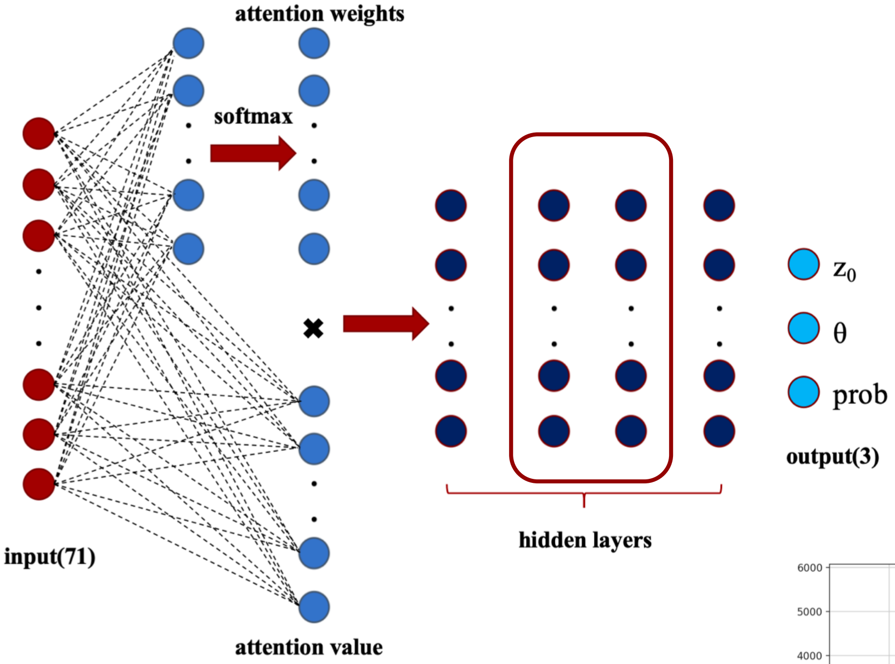


- Extract a track segment from each super layer
- Every track segment (TS) contains a set of α,t_{drift} and Φ



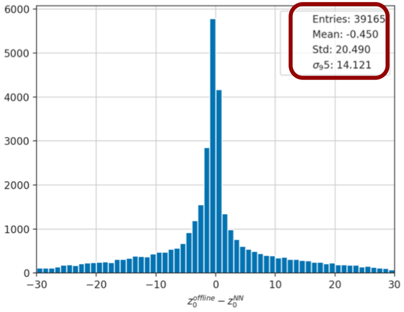
- DNN trigger : 2D track candidates+ Drift time for all other wires in the stereo wire
- Input : α,t_{drift} , phi of priority wire, t_{drift} of all other 10 wires for every TS(3*9+11*4=71)
- Output : z_0,θ_0,classifier output (Q)

Development of DNN model with python

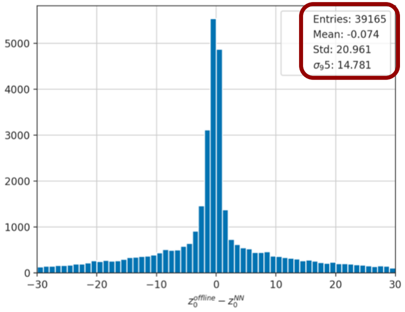


- Using pytorch lib for model building and training
- Add two hidden layers
- Modify learning rates

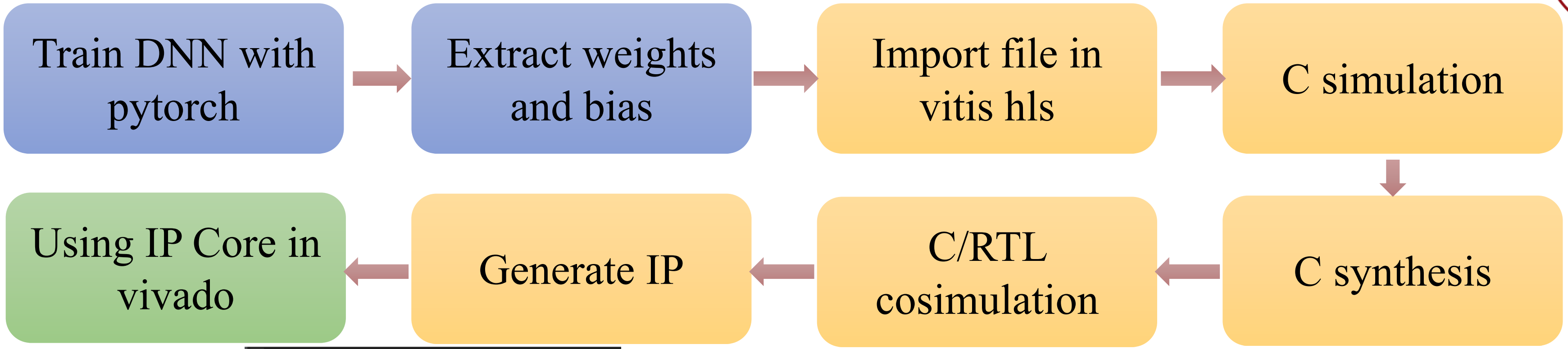
Initial



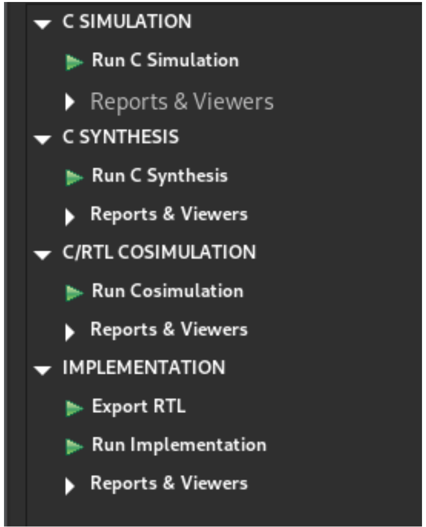
Development



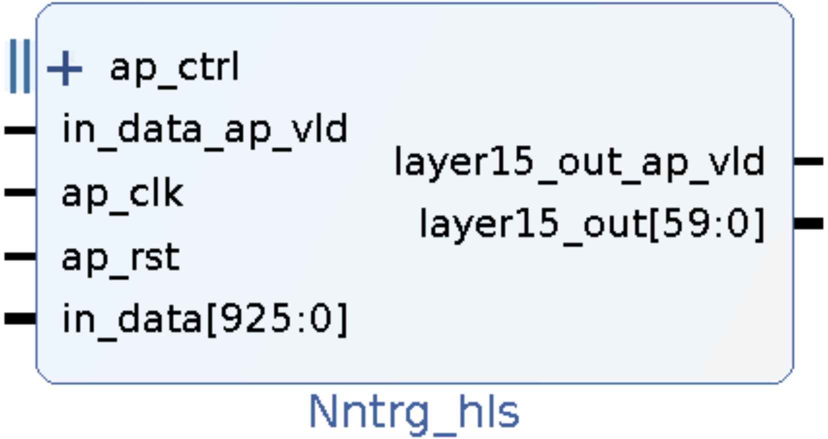
Workflow of DNN model development on FPGA



Train DNN with pytorch → Extract weights and bias → Import file in vitis hls → C simulation → C synthesis → C/RTL cosimulation → Generate IP → Using IP Core in vivado



- C SIMULATION
  - Run C Simulation
  - Reports & Viewers
- C SYNTHESIS
  - Run C Synthesis
  - Reports & Viewers
- C/RTL COSIMULATION
  - Run Cosimulation
  - Reports & Viewers
- IMPLEMENTATION
  - Export RTL
  - Run Implementation
  - Reports & Viewers

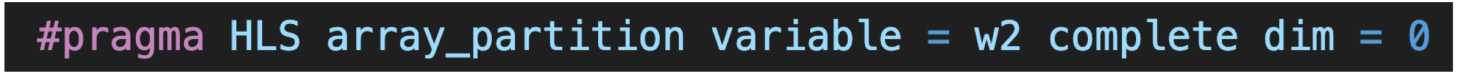


Development flow with Vitis HLS

A typical flow using HLS has the following steps:

- Convert python based model to C/C++ model with hls4ml (Optional)
- Write the algorithm using C/C++ with a target architecture in mind
- Verify the functionality at the behavioral level
- Generate the RTL based model
- Verify the functionality of the generated RTL model

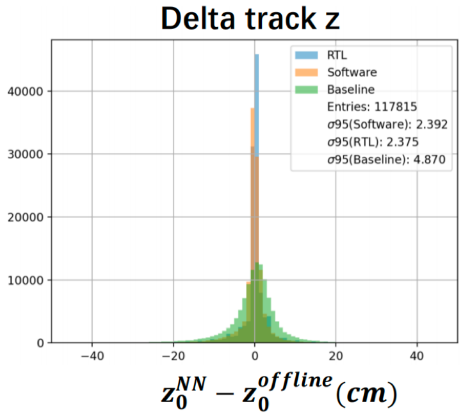
Optimizing performance: “#pragma HLS”

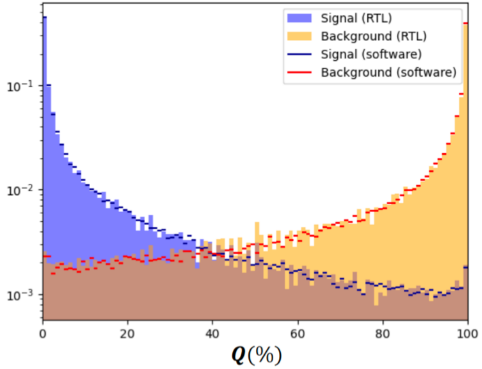


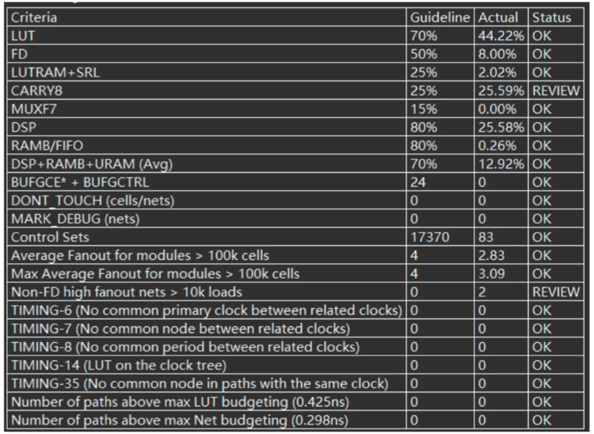
```
#pragma HLS array_partition variable = w2 complete dim = 0
```

- Generate multiple small memories
- Increases the amount of read and write ports for the storage

Performance

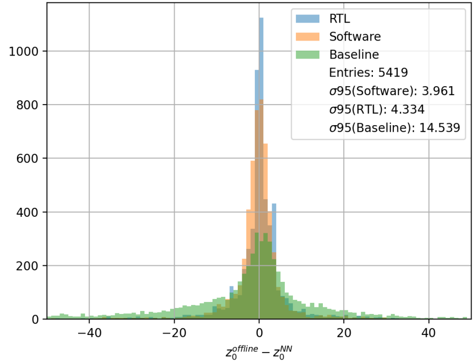


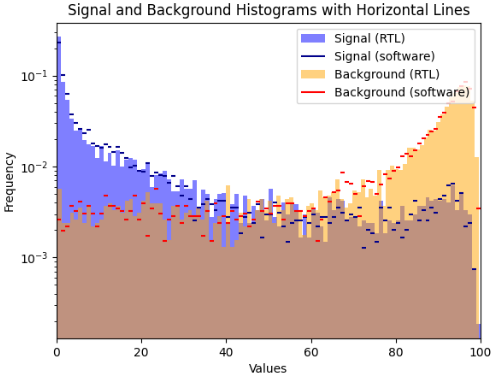


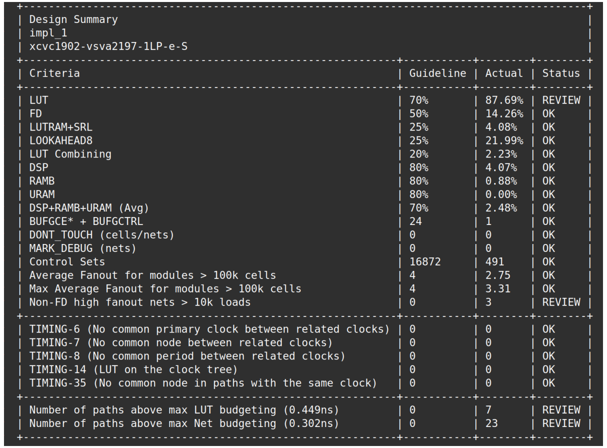


Criteria	Guideline	Actual	Status
LUT	70%	44.22%	OK
FD	50%	8.00%	OK
LUTRAM+SRL	25%	2.02%	OK
CARRY8	25%	25.59%	REVIEW
MUXF7	15%	0.00%	OK
DSP	80%	25.58%	OK
RAMB/FIFO	80%	0.26%	OK
DSP+RAMB+URAM (Avg)	70%	12.92%	OK
BUFGCE* + BUFGCTRL	24	0	OK
DONT_TOUCH (cells/nets)	0	0	OK
MARK_DEBUG (nets)	0	0	OK
Control Sets	17370	83	OK
Average Fanout for modules > 100k cells	4	2.83	OK
Max Average Fanout for modules > 100k cells	4	3.09	OK
Non-FD high fanout nets > 10k loads	0	2	REVIEW
TIMING-6 (No common primary clock between related clocks)	0	0	OK
TIMING-7 (No common node between related clocks)	0	0	OK
TIMING-8 (No common period between related clocks)	0	0	OK
TIMING-14 (LUT on the clock tree)	0	0	OK
TIMING-35 (No common node in paths with the same clock)	0	0	OK
Number of paths above max LUT budgeting (0.425ns)	0	0	OK
Number of paths above max Net budgeting (0.298ns)	0	0	OK

Virtex Ultra
XCVU160





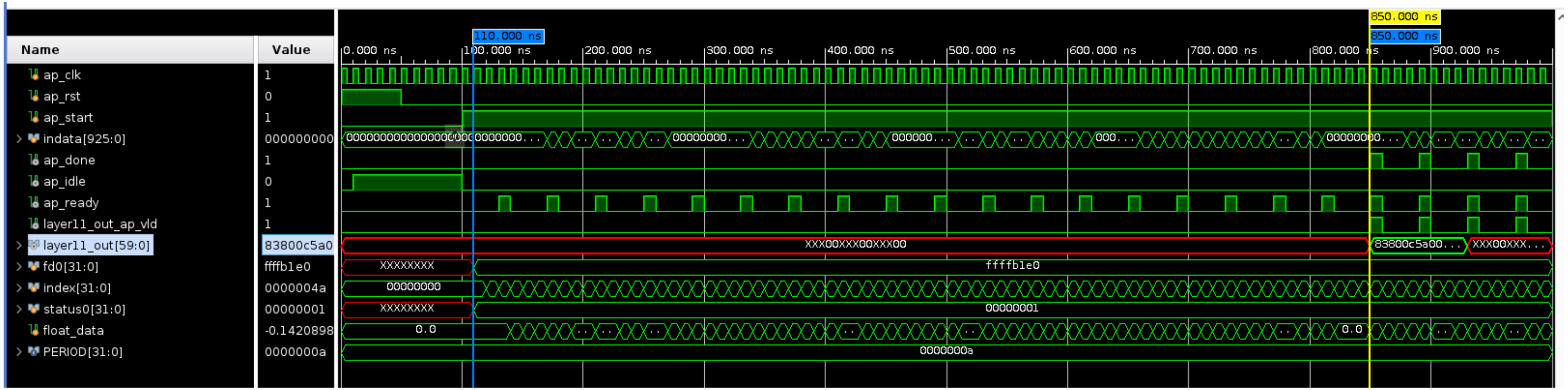


```
+-------------------------------------------------------------------------------------+
| Design Summary                                                                      |
| impl_1                                                                              |
| xcvc1902-vsva2197-1LP-e-S                                                           |
+-------------------------------------------------------+-----------+--------+--------+
| Criteria                                              | Guideline | Actual | Status |
+-------------------------------------------------------+-----------+--------+--------+
| LUT                                                   | 70%       | 87.69% | REVIEW |
| FD                                                    | 50%       | 14.26% | OK     |
| LUTRAM+SRL                                            | 25%       | 4.08%  | OK     |
| LOOKAHEAD8                                            | 25%       | 21.99% | OK     |
| LUT Combining                                         | 20%       | 2.23%  | OK     |
| DSP                                                   | 80%       | 4.07%  | OK     |
| RAMB                                                  | 80%       | 0.88%  | OK     |
| URAM                                                  | 80%       | 0.00%  | OK     |
| DSP+RAMB+URAM (Avg)                                   | 70%       | 2.48%  | OK     |
| BUFGCE* + BUFGCTRL                                    | 24        | 1      | OK     |
| DONT_TOUCH (cells/nets)                               | 0         | 0      | OK     |
| MARK_DEBUG (nets)                                     | 0         | 0      | OK     |
| Control Sets                                          | 16872     | 491    | OK     |
| Average Fanout for modules > 100k cells               | 4         | 2.75   | OK     |
| Max Average Fanout for modules > 100k cells           | 4         | 3.31   | OK     |
| Non-FD high fanout nets > 10k loads                   | 0         | 3      | REVIEW |
+-------------------------------------------------------+-----------+--------+--------+
| TIMING-6 (No common primary clock between related clocks) | 0     | 0      | OK     |
| TIMING-7 (No common node between related clocks)      | 0         | 0      | OK     |
| TIMING-8 (No common period between related clocks)    | 0         | 0      | OK     |
| TIMING-14 (LUT on the clock tree)                     | 0         | 0      | OK     |
| TIMING-35 (No common node in paths with the same clock) | 0       | 0      | OK     |
+-------------------------------------------------------+-----------+--------+--------+
| Number of paths above max LUT budgeting (0.449ns)     | 0         | 7      | REVIEW |
| Number of paths above max Net budgeting (0.302ns)     | 0         | 23     | REVIEW |
+-------------------------------------------------------+-----------+--------+--------+
```

Versal
AI Core
XCVC1902

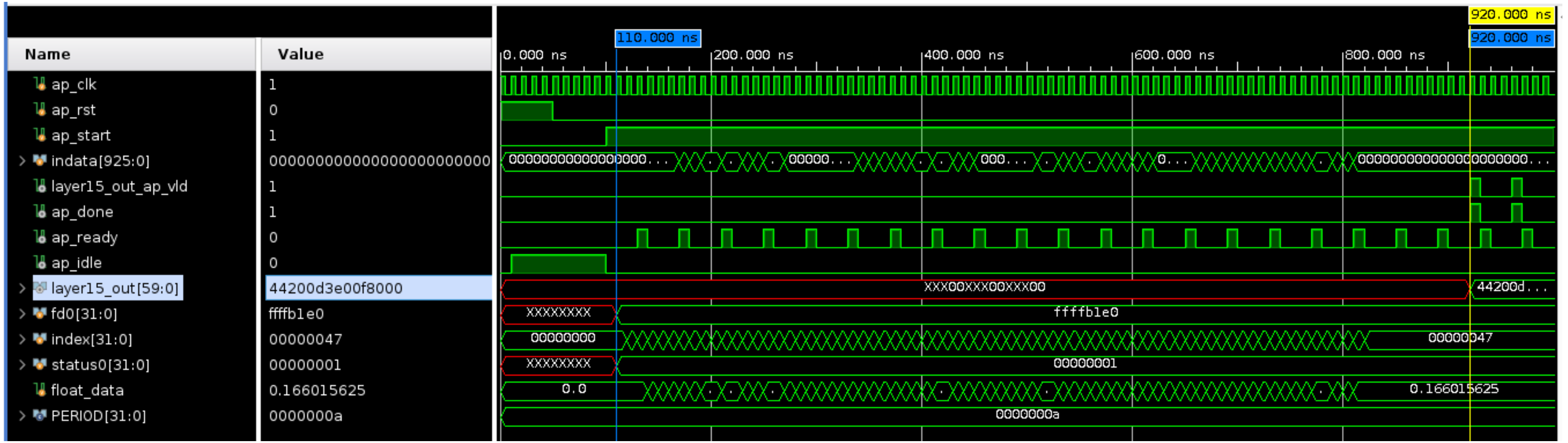
baseline(NN)/software(DNN)/RTL(HLS)

Latency(testbench)



Name	Value
ap_clk	1
ap_rst	0
ap_start	1
indata[925:0]	000000000
ap_done	1
ap_idle	0
ap_ready	1
layer11_out_ap_vld	1
layer11_out[59:0]	83800c5a0
fd0[31:0]	ffffb1e0
index[31:0]	0000004a
status0[31:0]	00000001
float_data	-0.1420898
PERIOD[31:0]	0000000a

initial:
74clock



Name	Value
ap_clk	1
ap_rst	0
ap_start	1
indata[925:0]	000000000000000000000000
layer15_out_ap_vld	1
ap_done	1
ap_ready	0
ap_idle	0
layer15_out[59:0]	44200d3e00f8000
fd0[31:0]	ffffb1e0
index[31:0]	00000047
status0[31:0]	00000001
float_data	0.166015625
PERIOD[31:0]	0000000a

optimize:
81clock

Next plan

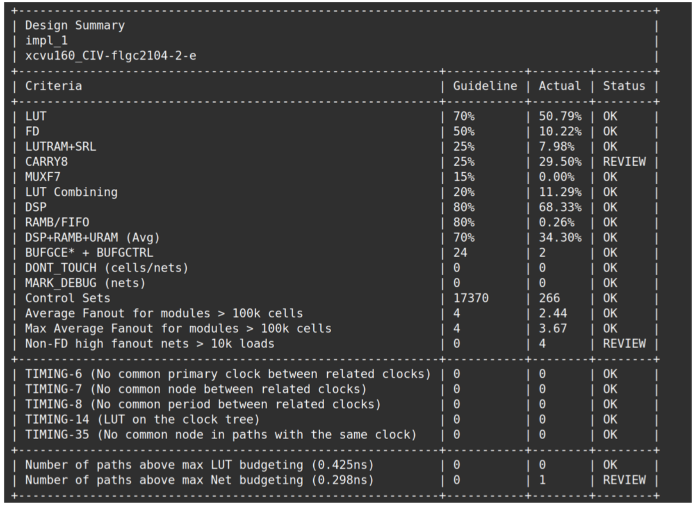
- Further optimize the model and perform pruning
- Investigating the possiblity of using more input variables in DNN models, such as ADC signals and momentum(Hope to receive more suggestions)
  - Check the size of the model scale(Utilization, latency)
  - Predicting the use of the next generation UT5 board
- Deploy the neutral networks to Versal ACAP
  - First step: deploy the DNN model into Versal AI engine
  - Further plan: deploy the NN model on Versal DPU

Thanks for your listening

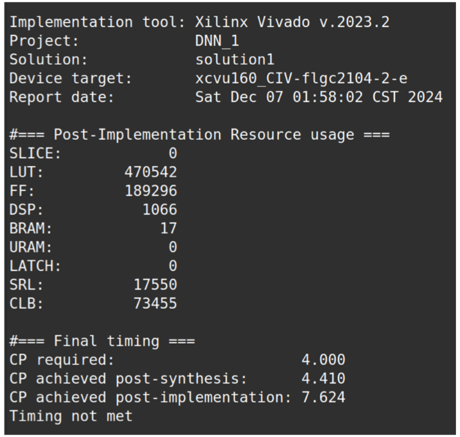
Back up

Virtex UltraScale XCVU160

initial

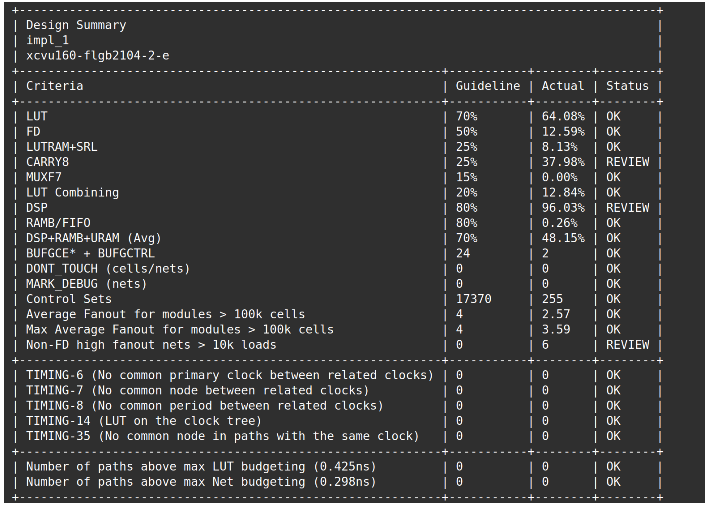


```
+-----------------------------------------------------------------------------------------+
| Design Summary                                                                          |
| impl_1                                                                                  |
| xcvu160_CIV-flgc2104-2-e                                                                |
+-----------------------------------------------------------+-----------+--------+--------+
| Criteria                                                  | Guideline | Actual | Status |
+-----------------------------------------------------------+-----------+--------+--------+
| LUT                                                       | 70%       | 50.79% | OK     |
| FD                                                        | 50%       | 10.22% | OK     |
| LUTRAM+SRL                                                | 25%       | 7.98%  | OK     |
| CARRY8                                                    | 25%       | 29.50% | REVIEW |
| MUXF7                                                     | 15%       | 0.00%  | OK     |
| LUT Combining                                             | 20%       | 11.29% | OK     |
| DSP                                                       | 80%       | 68.33% | OK     |
| RAMB/FIFO                                                 | 80%       | 0.26%  | OK     |
| DSP+RAMB+URAM (Avg)                                       | 70%       | 34.30% | OK     |
| BUFGCE* + BUFGCTRL                                        | 24        | 2      | OK     |
| DONT_TOUCH (cells/nets)                                   | 0         | 0      | OK     |
| MARK_DEBUG (nets)                                         | 0         | 0      | OK     |
| Control Sets                                              | 17370     | 266    | OK     |
| Average Fanout for modules > 100k cells                   | 4         | 2.44   | OK     |
| Max Average Fanout for modules > 100k cells               | 4         | 3.67   | OK     |
| Non-FD high fanout nets > 10k loads                       | 0         | 4      | REVIEW |
+-----------------------------------------------------------+-----------+--------+--------+
| TIMING-6 (No common primary clock between related clocks) | 0         | 0      | OK     |
| TIMING-7 (No common node between related clocks)          | 0         | 0      | OK     |
| TIMING-8 (No common period between related clocks)        | 0         | 0      | OK     |
| TIMING-14 (LUT on the clock tree)                         | 0         | 0      | OK     |
| TIMING-35 (No common node in paths with the same clock)   | 0         | 0      | OK     |
+-----------------------------------------------------------+-----------+--------+--------+
| Number of paths above max LUT budgeting (0.425ns)         | 0         | 0      | OK     |
| Number of paths above max Net budgeting (0.298ns)         | 0         | 1      | REVIEW |
+-----------------------------------------------------------+-----------+--------+--------+
```

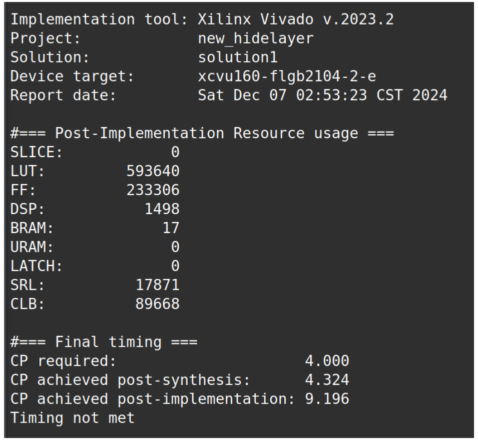


```
Implementation tool: Xilinx Vivado v.2023.2
Project:             DNN_1
Solution:            solution1
Device target:       xcvu160_CIV-flgc2104-2-e
Report date:         Sat Dec 07 01:58:02 CST 2024

#=== Post-Implementation Resource usage ===
SLICE:            0
LUT:         470542
FF:          189296
DSP:           1066
BRAM:            17
URAM:             0
LATCH:            0
SRL:          17550
CLB:          73455

#=== Final timing ===
CP required:                     4.000
CP achieved post-synthesis:      4.410
CP achieved post-implementation: 7.624
Timing not met
```

optimize

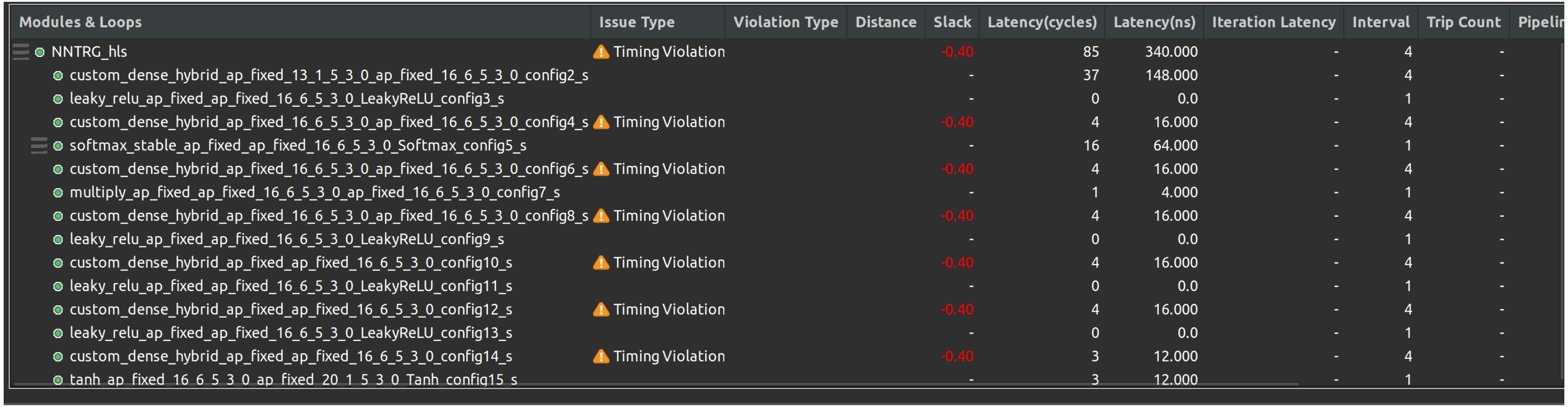


```
+-----------------------------------------------------------------------------------------+
| Design Summary                                                                          |
| impl_1                                                                                  |
| xcvu160-flgb2104-2-e                                                                    |
+-----------------------------------------------------------+-----------+--------+--------+
| Criteria                                                  | Guideline | Actual | Status |
+-----------------------------------------------------------+-----------+--------+--------+
| LUT                                                       | 70%       | 64.08% | OK     |
| FD                                                        | 50%       | 12.59% | OK     |
| LUTRAM+SRL                                                | 25%       | 8.13%  | OK     |
| CARRY8                                                    | 25%       | 37.98% | REVIEW |
| MUXF7                                                     | 15%       | 0.00%  | OK     |
| LUT Combining                                             | 20%       | 12.84% | OK     |
| DSP                                                       | 80%       | 96.03% | REVIEW |
| RAMB/FIFO                                                 | 80%       | 0.26%  | OK     |
| DSP+RAMB+URAM (Avg)                                       | 70%       | 48.15% | OK     |
| BUFGCE* + BUFGCTRL                                        | 24        | 2      | OK     |
| DONT_TOUCH (cells/nets)                                   | 0         | 0      | OK     |
| MARK_DEBUG (nets)                                         | 0         | 0      | OK     |
| Control Sets                                              | 17370     | 255    | OK     |
| Average Fanout for modules > 100k cells                   | 4         | 2.57   | OK     |
| Max Average Fanout for modules > 100k cells               | 4         | 3.59   | OK     |
| Non-FD high fanout nets > 10k loads                       | 0         | 6      | REVIEW |
+-----------------------------------------------------------+-----------+--------+--------+
| TIMING-6 (No common primary clock between related clocks) | 0         | 0      | OK     |
| TIMING-7 (No common node between related clocks)          | 0         | 0      | OK     |
| TIMING-8 (No common period between related clocks)        | 0         | 0      | OK     |
| TIMING-14 (LUT on the clock tree)                         | 0         | 0      | OK     |
| TIMING-35 (No common node in paths with the same clock)   | 0         | 0      | OK     |
+-----------------------------------------------------------+-----------+--------+--------+
| Number of paths above max LUT budgeting (0.425ns)         | 0         | 0      | OK     |
| Number of paths above max Net budgeting (0.298ns)         | 0         | 0      | OK     |
+-----------------------------------------------------------+-----------+--------+--------+
```

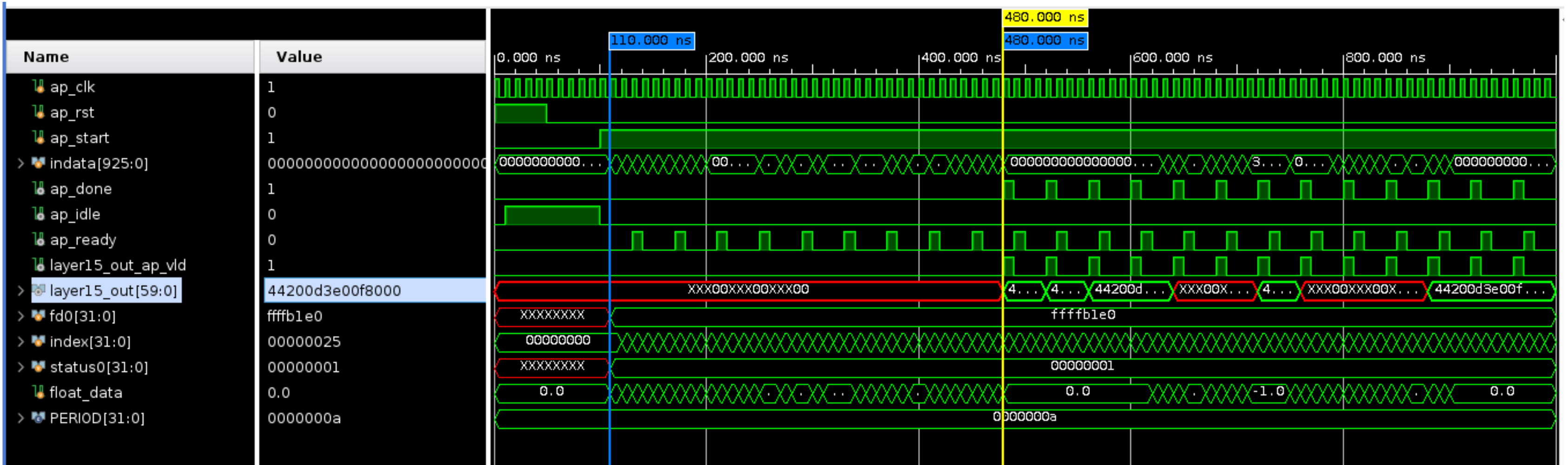


```
Implementation tool: Xilinx Vivado v.2023.2
Project:             new_hidelayer
Solution:            solution1
Device target:       xcvu160-flgb2104-2-e
Report date:         Sat Dec 07 02:53:23 CST 2024

#=== Post-Implementation Resource usage ===
SLICE:            0
LUT:         593640
FF:          233306
DSP:           1498
BRAM:            17
URAM:             0
LATCH:            0
SRL:          17871
CLB:          89668

#=== Final timing ===
CP required:                     4.000
CP achieved post-synthesis:      4.324
CP achieved post-implementation: 9.196
Timing not met
```

Some questions



Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelin
NNTRG_hls	⚠ Timing Violation			-0.40	85	340.000	-	4	-	
custom_dense_hybrid_ap_fixed_13_1_5_3_0_ap_fixed_16_6_5_3_0_config2_s				-	37	148.000	-	4	-	
leaky_relu_ap_fixed_ap_fixed_16_6_5_3_0_LeakyReLU_config3_s				-	0	0.0	-	1	-	
custom_dense_hybrid_ap_fixed_16_6_5_3_0_ap_fixed_16_6_5_3_0_config4_s	⚠ Timing Violation			-0.40	4	16.000	-	4	-	
softmax_stable_ap_fixed_ap_fixed_16_6_5_3_0_Softmax_config5_s				-	16	64.000	-	1	-	
custom_dense_hybrid_ap_fixed_16_6_5_3_0_ap_fixed_16_6_5_3_0_config6_s	⚠ Timing Violation			-0.40	4	16.000	-	4	-	
multiply_ap_fixed_ap_fixed_16_6_5_3_0_ap_fixed_16_6_5_3_0_config7_s				-	1	4.000	-	1	-	
custom_dense_hybrid_ap_fixed_16_6_5_3_0_ap_fixed_16_6_5_3_0_config8_s	⚠ Timing Violation			-0.40	4	16.000	-	4	-	
leaky_relu_ap_fixed_ap_fixed_16_6_5_3_0_LeakyReLU_config9_s				-	0	0.0	-	1	-	
custom_dense_hybrid_ap_fixed_ap_fixed_16_6_5_3_0_config10_s	⚠ Timing Violation			-0.40	4	16.000	-	4	-	
leaky_relu_ap_fixed_ap_fixed_16_6_5_3_0_LeakyReLU_config11_s				-	0	0.0	-	1	-	
custom_dense_hybrid_ap_fixed_ap_fixed_16_6_5_3_0_config12_s	⚠ Timing Violation			-0.40	4	16.000	-	4	-	
leaky_relu_ap_fixed_ap_fixed_16_6_5_3_0_LeakyReLU_config13_s				-	0	0.0	-	1	-	
custom_dense_hybrid_ap_fixed_ap_fixed_16_6_5_3_0_config14_s	⚠ Timing Violation			-0.40	3	12.000	-	4	-	
tanh_ap_fixed_16_6_5_3_0_ap_fixed_20_1_5_3_0_Tanh_config15_s				-	3	12.000	-	1	-	



Name	Value
ap_clk	1
ap_rst	0
ap_start	1
indata[925:0]	000000000000000000000000
ap_done	1
ap_idle	0
ap_ready	0
layer15_out_ap_vld	1
layer15_out[59:0]	44200d3e00f8000
fd0[31:0]	ffffb1e0
index[31:0]	00000025
status0[31:0]	00000001
float_data	0.0
PERIOD[31:0]	0000000a