Implementing the TensorFlow Model on FPGA

Yunpeng Men, Andrei Kazantsev,
Ramesh Karuppusamy, Michael Kramer

Max Planck Institute for Radio Astronomy

Outline

- Introduction
- Workflows
- Implementation on FPGA
- Test results

Introduction

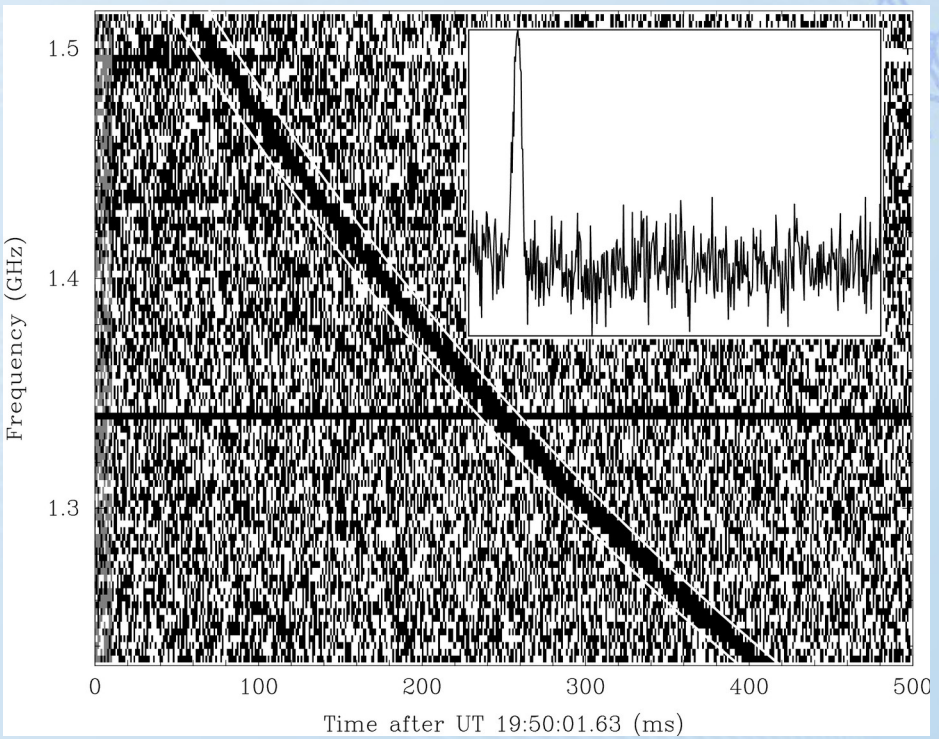
Fast Radio Burst

Bright ~ 50 mJy to ~100 Jy

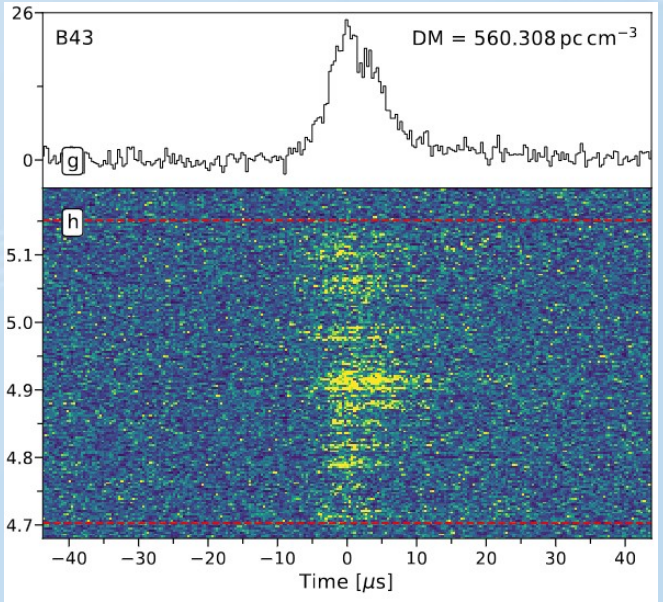
Short duration ~ 1 us to ~ 100 ms

Extragalactic origin

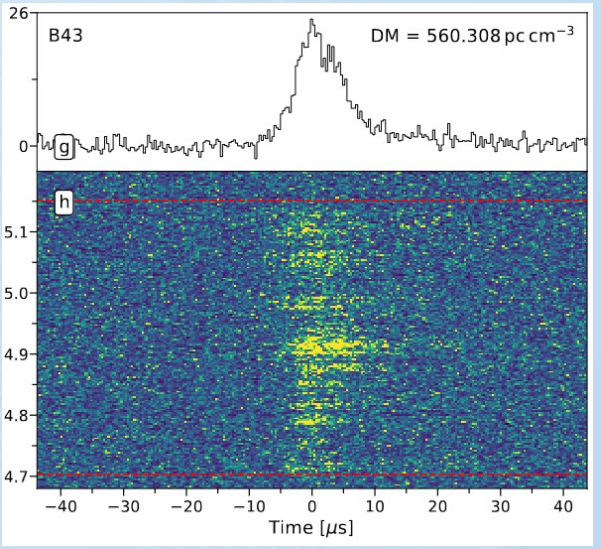
The Parkes 64m Radio Telescope



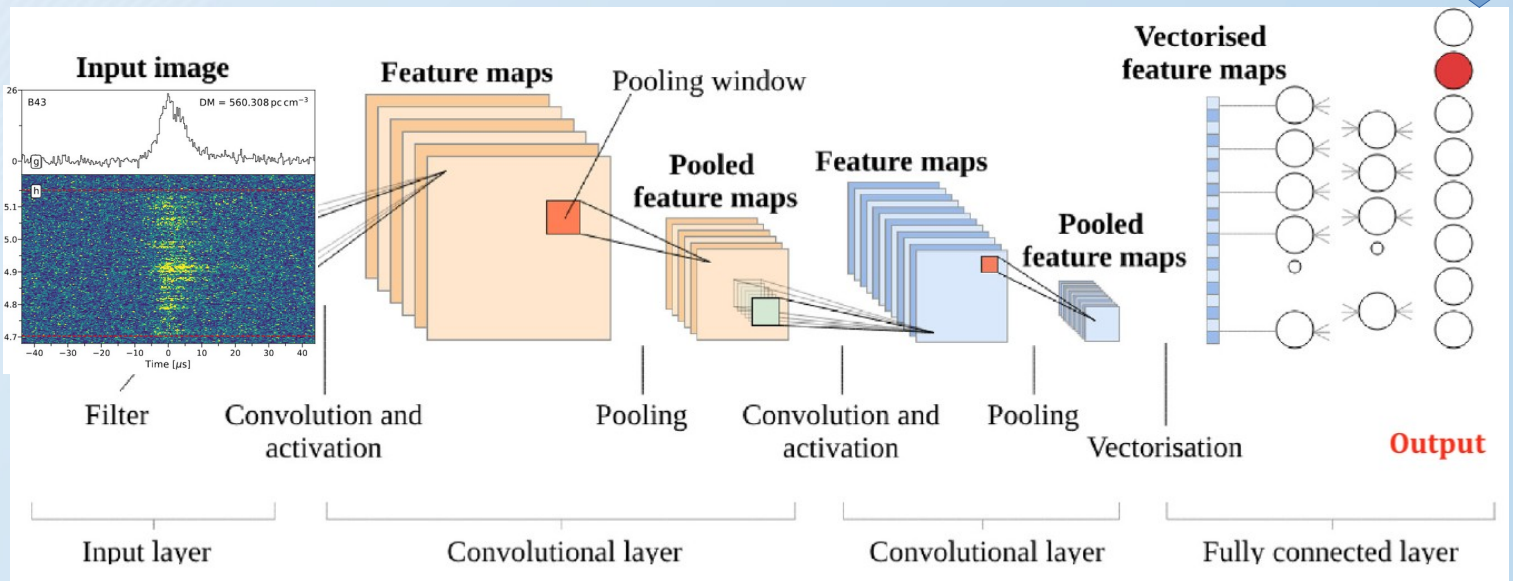
Lorimer Burst (Lorimer et al. 2007)



Real-time pipeline

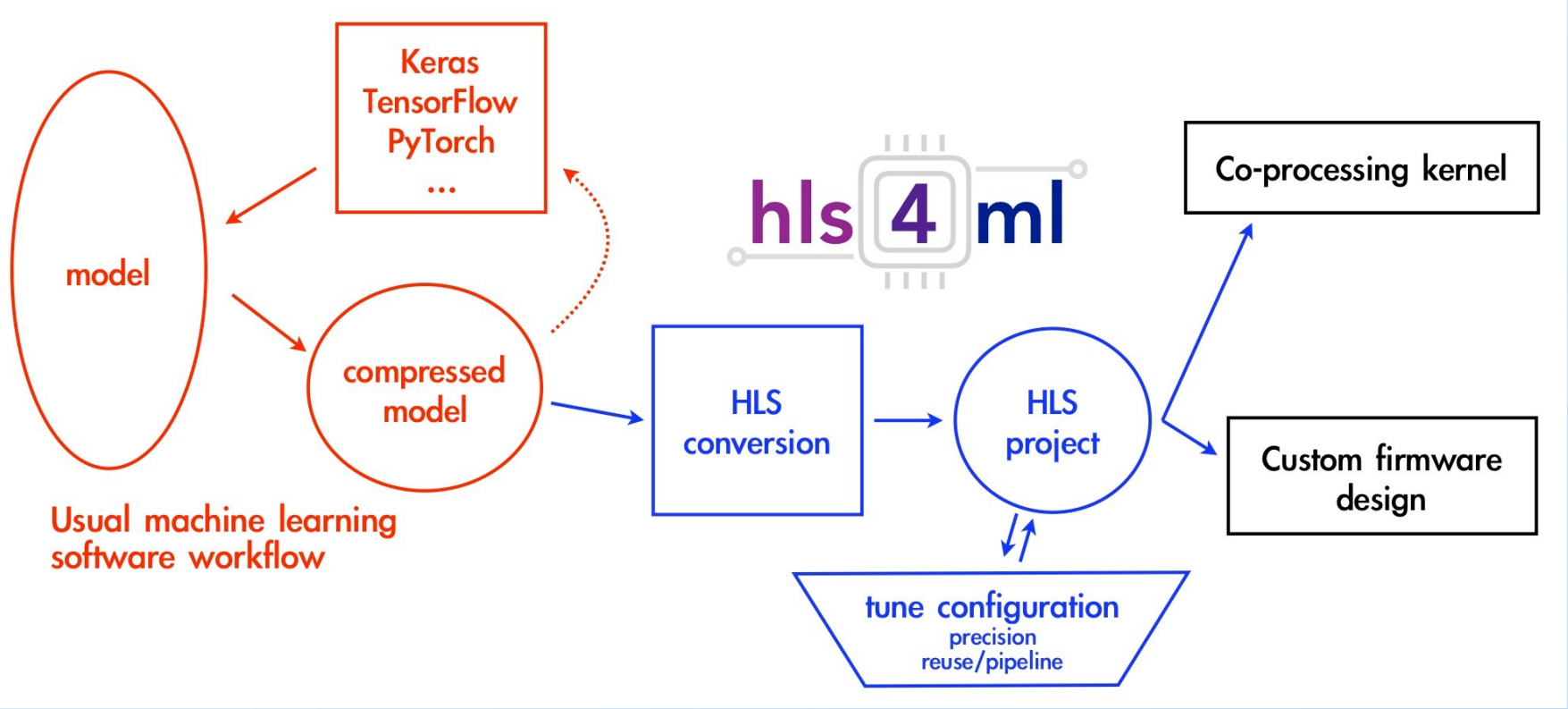


trigger

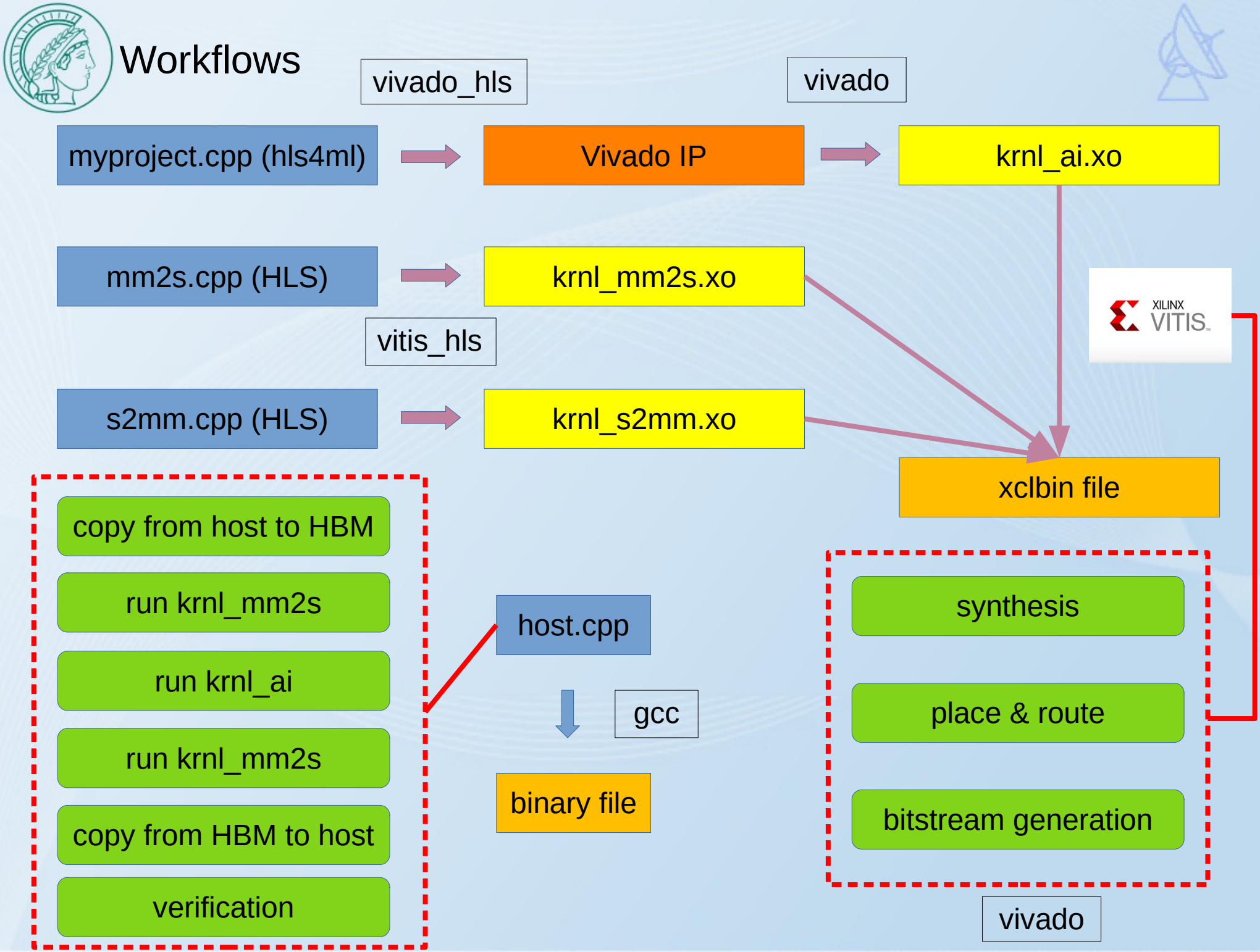


HLS4ML

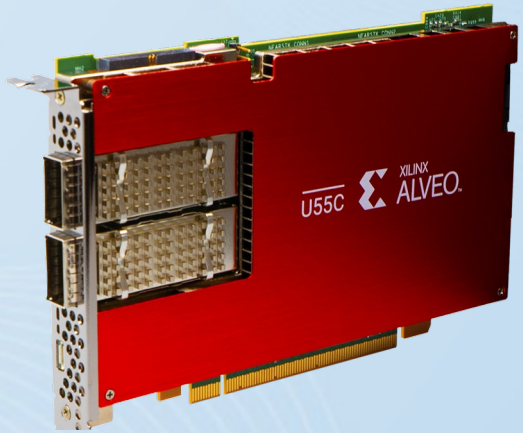
https://github.com/fastmachinelearning/hls4ml



Workflows


vivado_hls
vivado
myproject.cpp (hls4ml)
Vivado IP
krnl_ai.xo
mm2s.cpp (HLS)
krnl_mm2s.xo
vitis_hls
XILINX VITIS
s2mm.cpp (HLS)
krnl_s2mm.xo
xclbin file
copy from host to HBM
run krnl_mm2s
run krnl_ai
run krnl_mm2s
copy from HBM to host
verification
host.cpp
gcc
binary file
synthesis
place & route
bitstream generation
vivado

Xilinx Alveo Card



Specification	U55C
Product SKU	A-U55C-P00G-PQ-G
Total electrical card load[1]	150W
Thermal design power (TDP)[2]	115W
Thermal cooling solution	Passive
Weight	519g
Form factor	Full height, half length
Network interface	2 x QSFP28
PCIe interface[3]	Gen3 x16, 2 x Gen4 x8
HBM2 total capacity	16 GB
HBM2 bandwidth	460 GB/s
Look-up tables (LUTs)	1,304K
Registers	2,607K
DSP slices	9,024
Maximum distributed RAM	36.7 Mb
36 Kb block RAM	70.9 Mb
288 Kb UltraRAM	960 (270 Mb)
GTY transceivers	24
Qualified for deployment	Yes

Pre-Implementation metrics

tensorflow model

Performance Estimates

Timing

Summary

Clock	Target	Estimated	Uncertainty
ap_clk	5.00 ns	13.773 ns	0.62 ns

Latency

Summary

Latency (cycles)		Latency (absolute)		Interval (cycles)		
min	max	min	max	min	max	Type
71281	71624	0.982 ms	0.987 ms	42337	71345	dataflow

Utilization Estimates

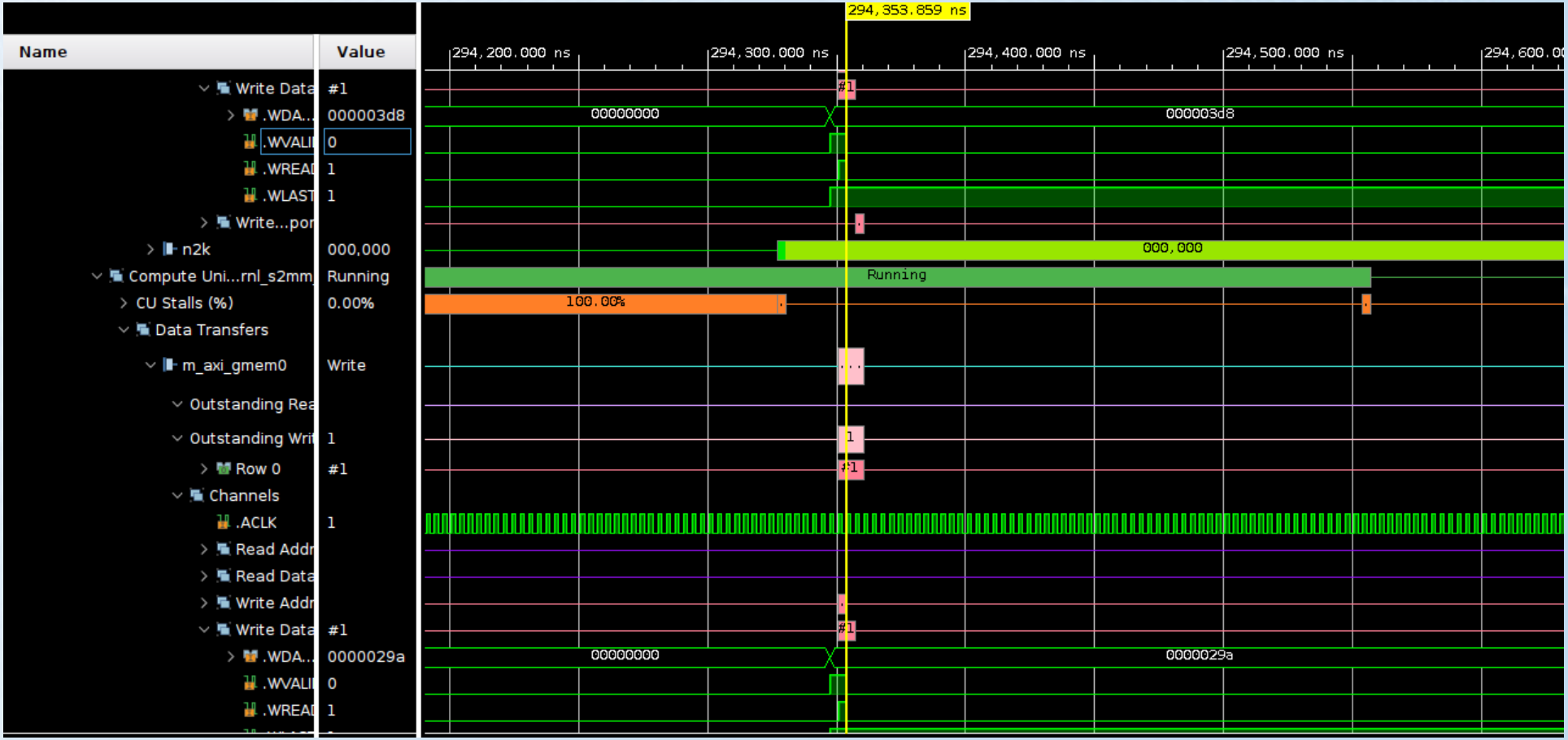
Summary

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	2	-
FIFO	76	-	3820	8632	-
Instance	102	649	61768	201532	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	-	-
Register	-	-	-	-	-
Total	178	649	65588	210166	0
Available	4032	9024	2607360	1303680	960
Available SLR	1344	3008	869120	434560	320
Utilization (%)	4	7	2	16	0
Utilization SLR (%)	13	21	7	48	0

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_19 (Conv2D)           (None, 30, 30, 4)         40

conv2d_20 (Conv2D)           (None, 28, 28, 4)         148

conv2d_21 (Conv2D)           (None, 24, 24, 4)         404

max_pooling2d_4 (MaxPooling  (None, 12, 12, 4)         0
2D)

conv2d_22 (Conv2D)           (None, 10, 10, 8)         296

conv2d_23 (Conv2D)           (None, 8, 8, 8)           584

conv2d_24 (Conv2D)           (None, 6, 6, 8)           584

max_pooling2d_5 (MaxPooling  (None, 3, 3, 8)           0
2D)

conv2d_25 (Conv2D)           (None, 1, 1, 16)          1168

flatten_1 (Flatten)          (None, 16)                0

dense_2 (Dense)              (None, 64)                1088

dense_3 (Dense)              (None, 4)                 260

=================================================================
Total params: 4,572
Trainable params: 4,572
Non-trainable params: 0
```

Hardware emulation

- Tensorflow output: (3.3878484e-06, 9.6108353e-01, 6.5116578e-01, 5.2037290e-08)
- Emulation output: (0, 9.609375e-01, 6.50390625e-01, 0)



Post-Inplementation Metrics

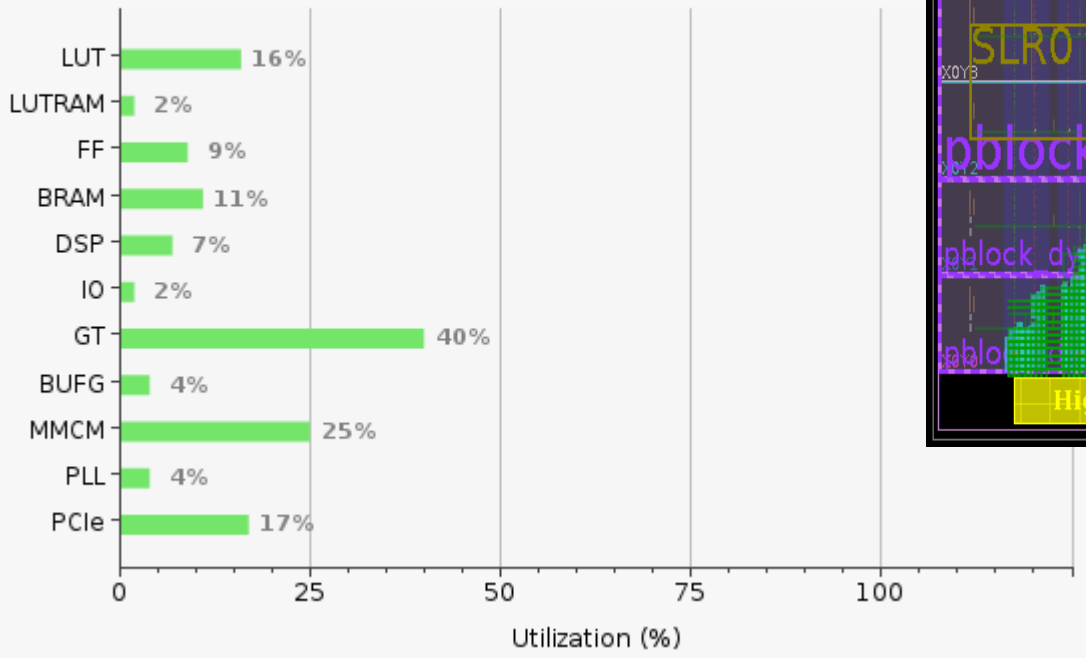
Power — **Summary** | On-Ch

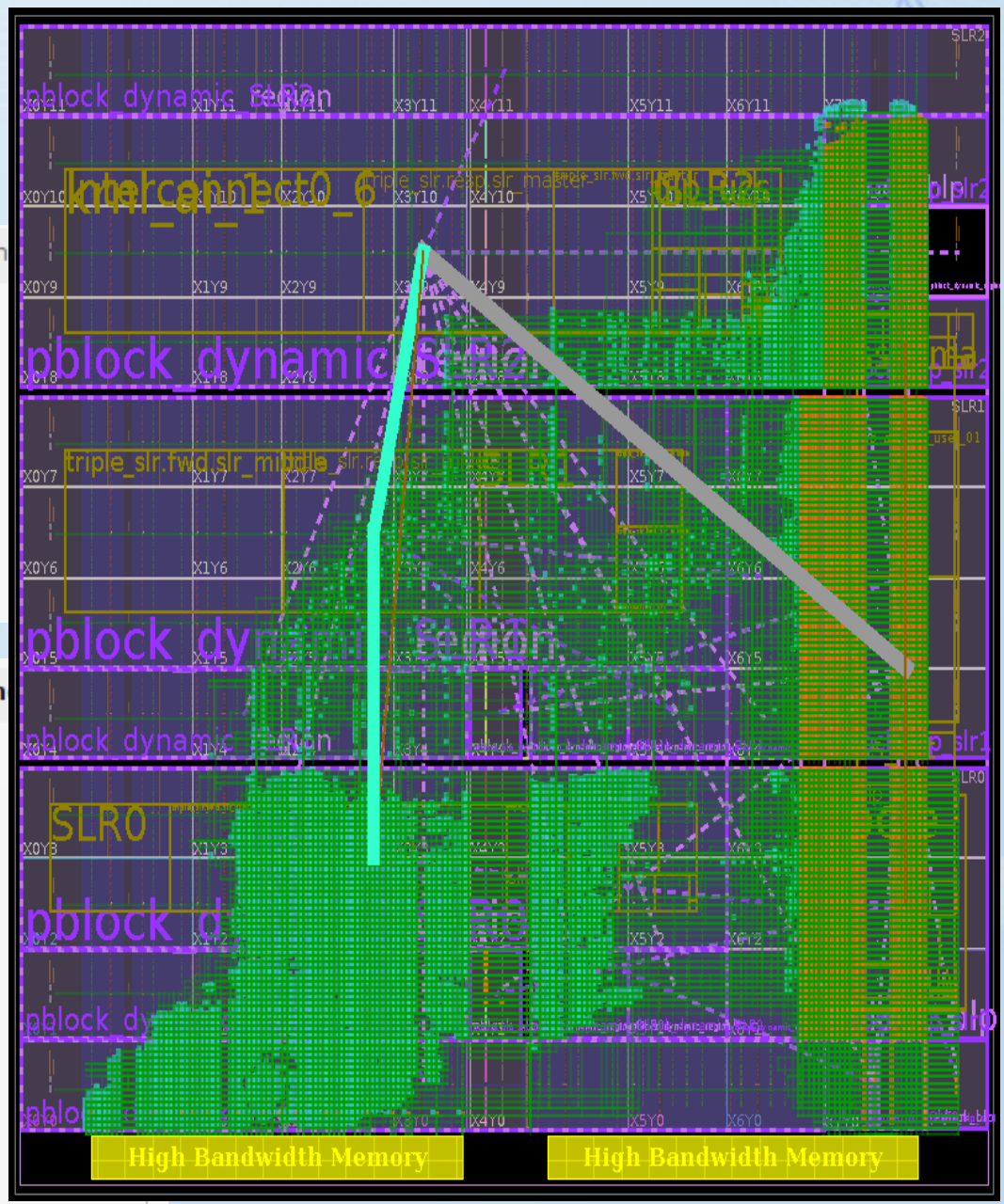
Total On-Chip Power:	**19.77 W**
Junction Temperature:	**69.8 °C**
Thermal Margin:	30.2 °C (32.7 W)
Effective ϑJA:	0.8 °C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Low

Implemented Power Report

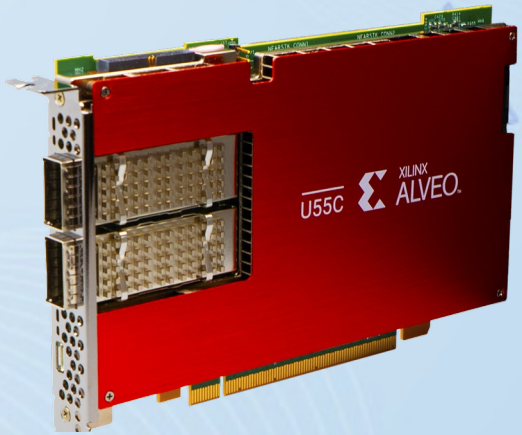
Utilization — Post-Synthesis | **Post-Implem**

Graph





Hardware test



The results are consistent!

```
ypmen@fpgdev:~/alveo/punch$ ./host -x krnl_ai.xclbin
Open the device0
Load the xclbin krnl_ai.xclbin
Allocate Buffer in Global Memory
synchronize input buffer data to device global memory
Getting Results...
verify data...
0
0.960938
0.650391
0
Test passed!
```

- Tensorflow output: (3.3878484e-06, 9.6108353e-01, 6.5116578e-01, 5.2037290e-08)
- Emulation output: (0, 9.60938e-01, 6.50391e-01, 0)

Thank you!