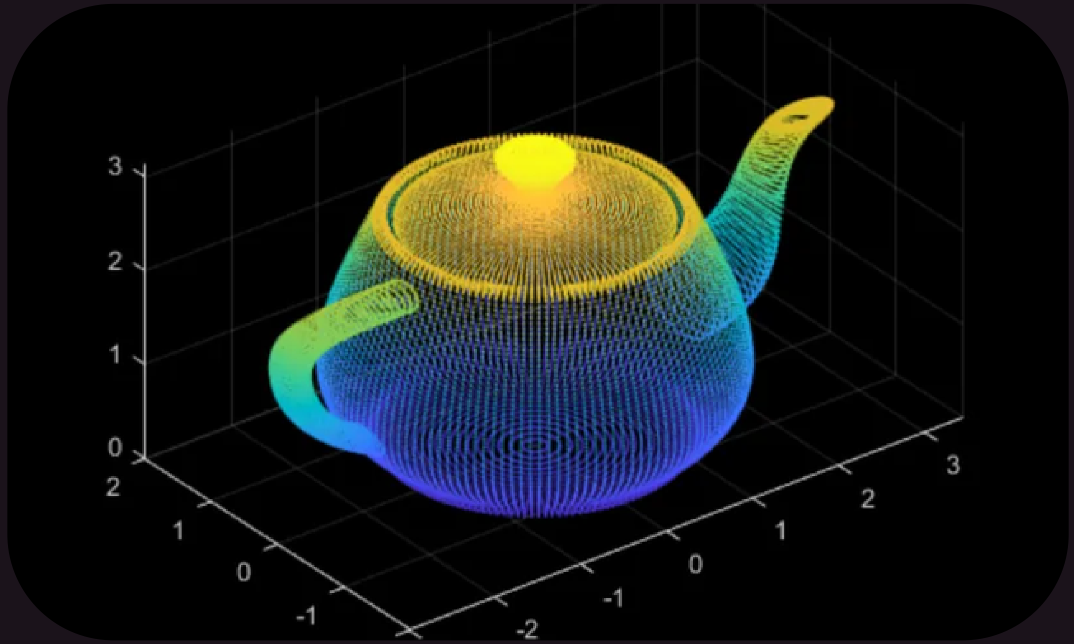
3D POINT CLOUDS ON EMBEDDED PLATFORMS

Presented by: Amur Pal - Junaid Ali - Sannan Zia

Advisor: Dr. Faisal Shafait
Co-advisor: Dr. Adnan Ul Hassan

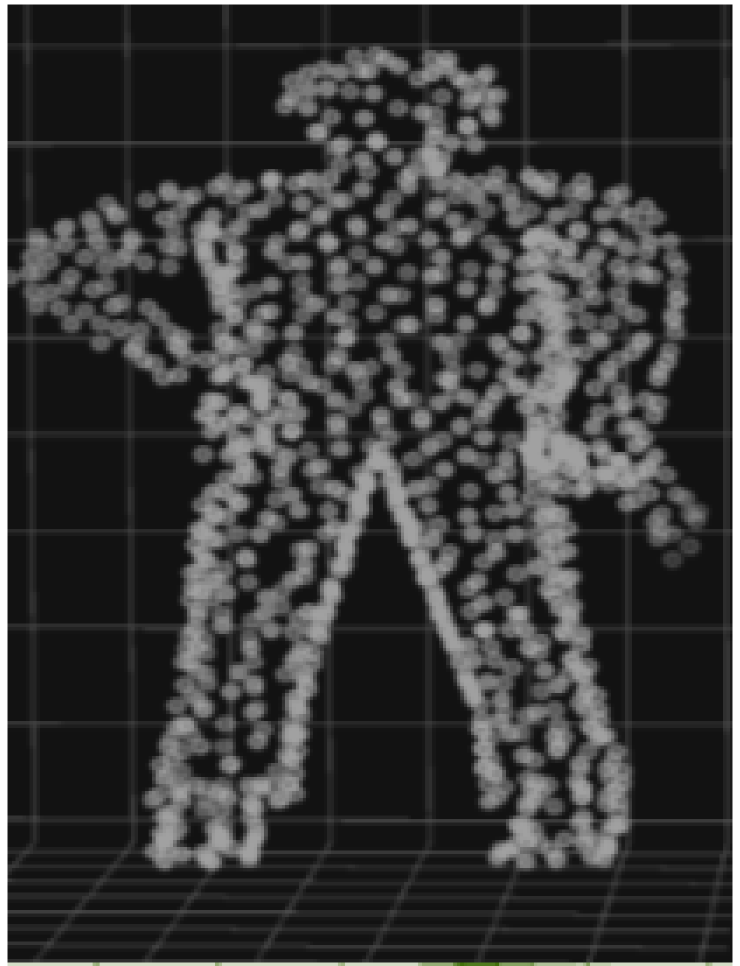
May 17th, 2023

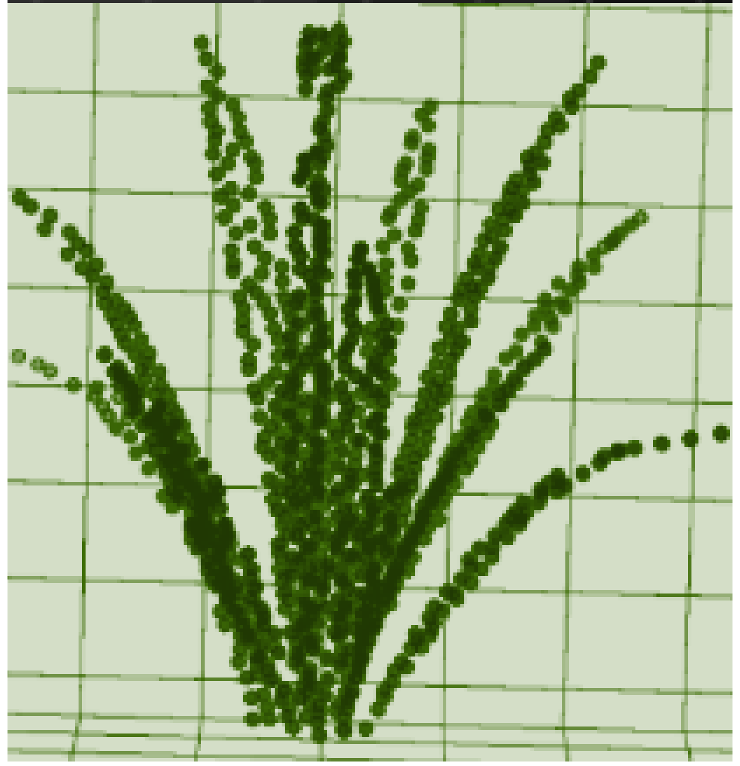
Problem Statement



IMPLEMENT A FRAMEWORK FOR EFFICIENT IMPLEMENTATION OF POINT CLOUD DEEP LEARNING **MODELS ON FPGA**

- BUILDING A HIGH-LEVEL SYNTHESIS LIBRARY WITH THE SUPPORTED ALGORITHMS
- DEMONSTRATING NOVEL HARDWARE RESULTS WITH SOTA 3D POINT CLOUD MODEL POINTMLPELITE
- DEMONSTRATING FPGA ACCELERATION OVER CPU & GPU





State of the art:
PointMLPElite

Models: PointMLP, PointMLPElite

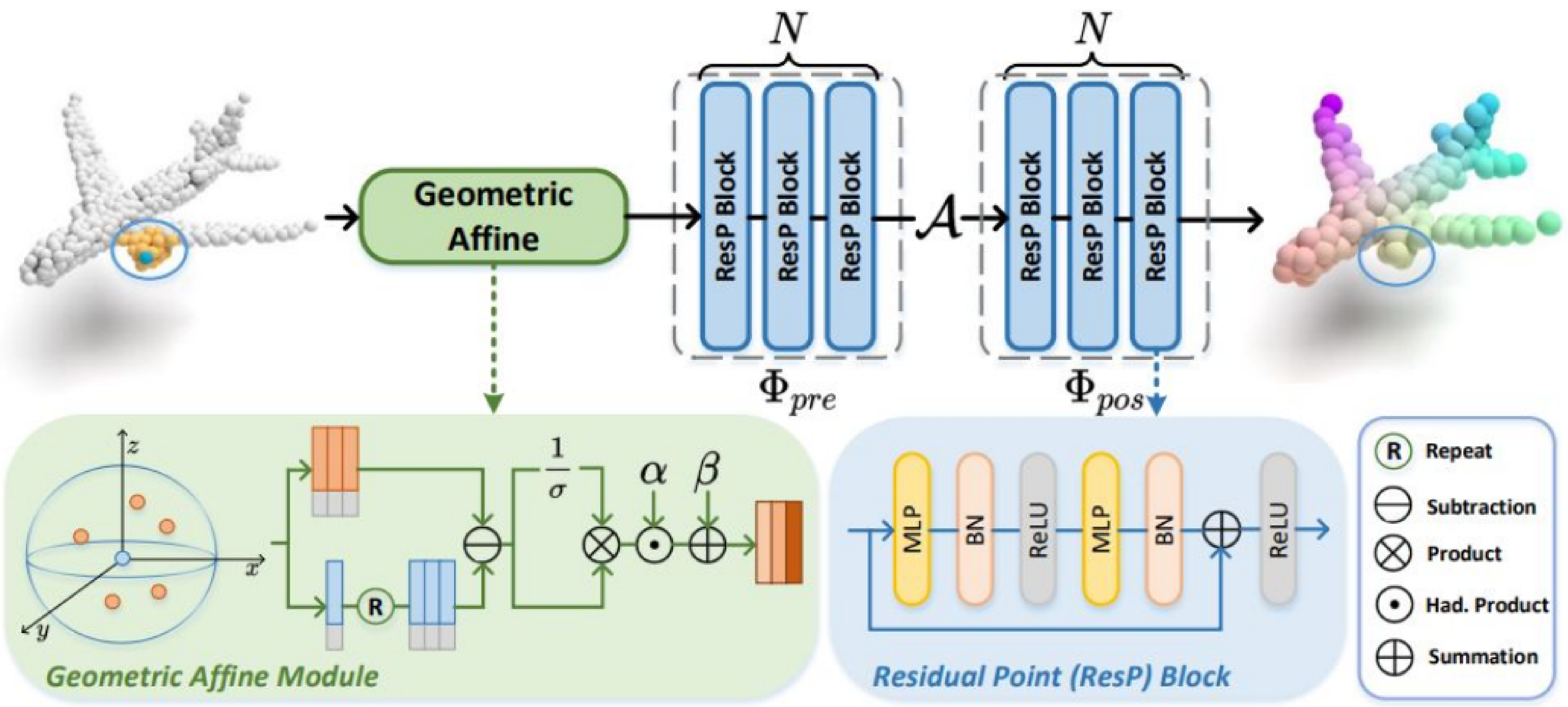
Task: Classification, Segmentation

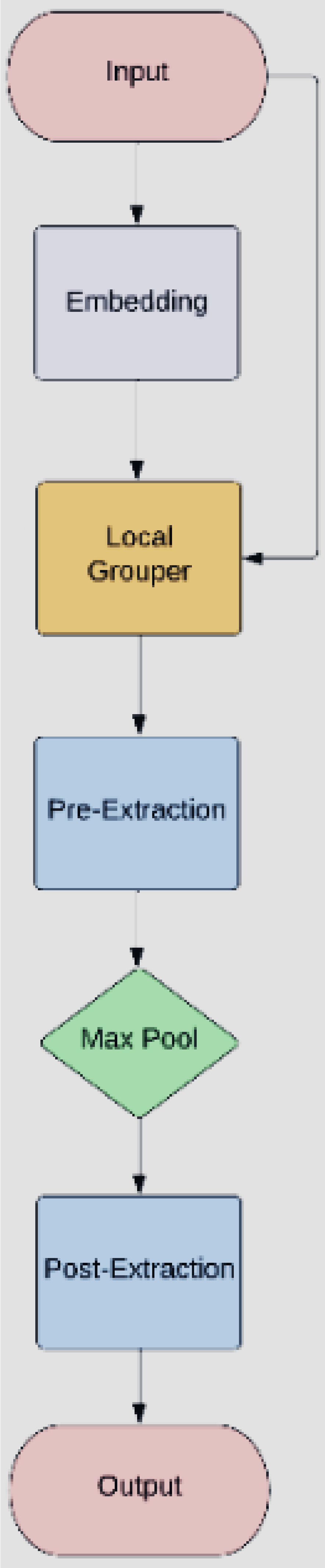
Dataset: ModelNet40

Model	Accuracy (%)	Parameters (M)	Layers	Size (MBs)	Throughput (samples/sec)
PointMLP	94.1 0.5%	12.6 18x	40	50.49 18.4x	112
PointMLPElite	93.6	0.68	25	2.744	176

POINTMLPELITE – TOPOLOGY

OVERVIEW & COMPONENTS



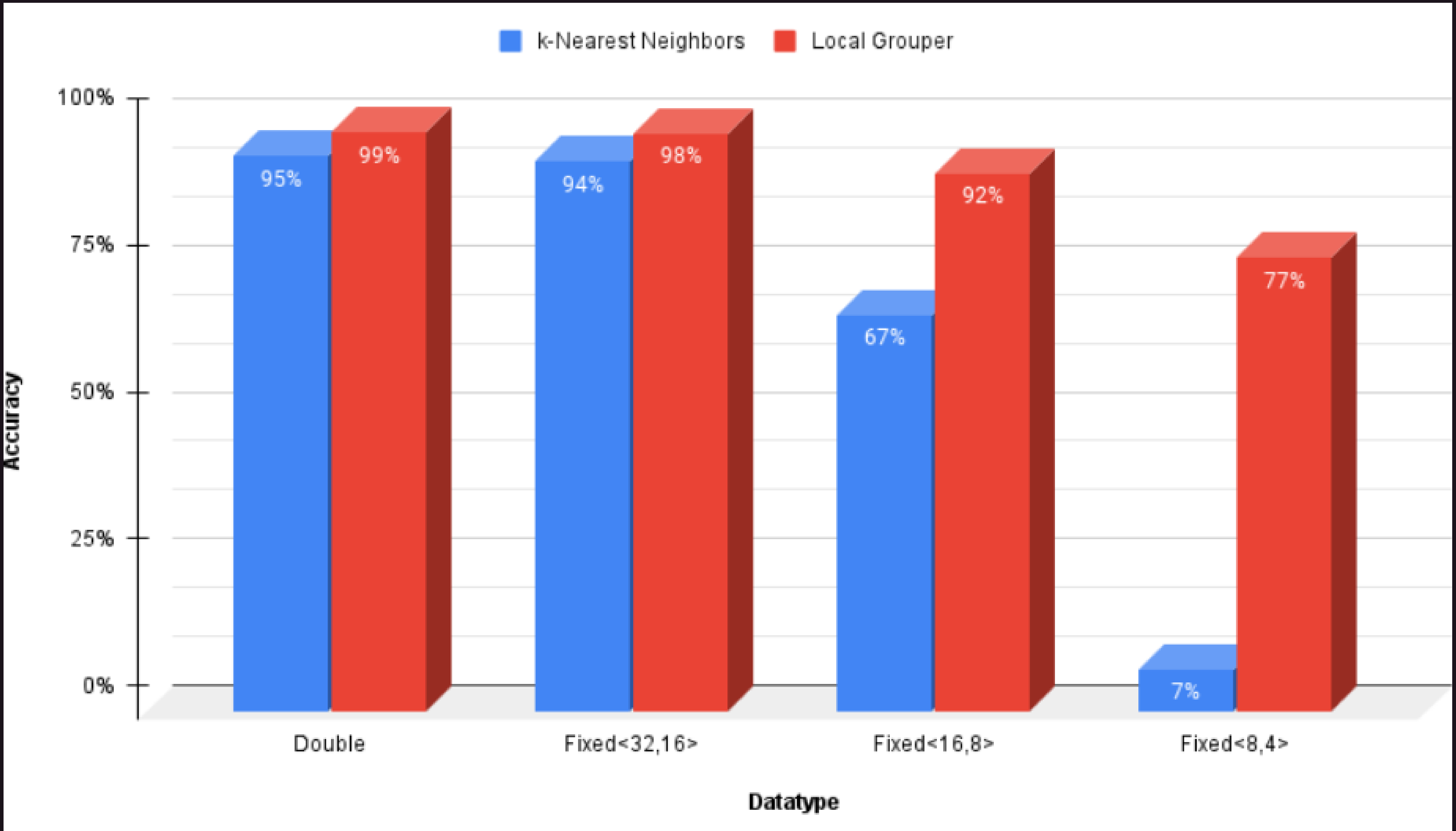


METHODOLOGY: SOFTWARE EXPLORATION

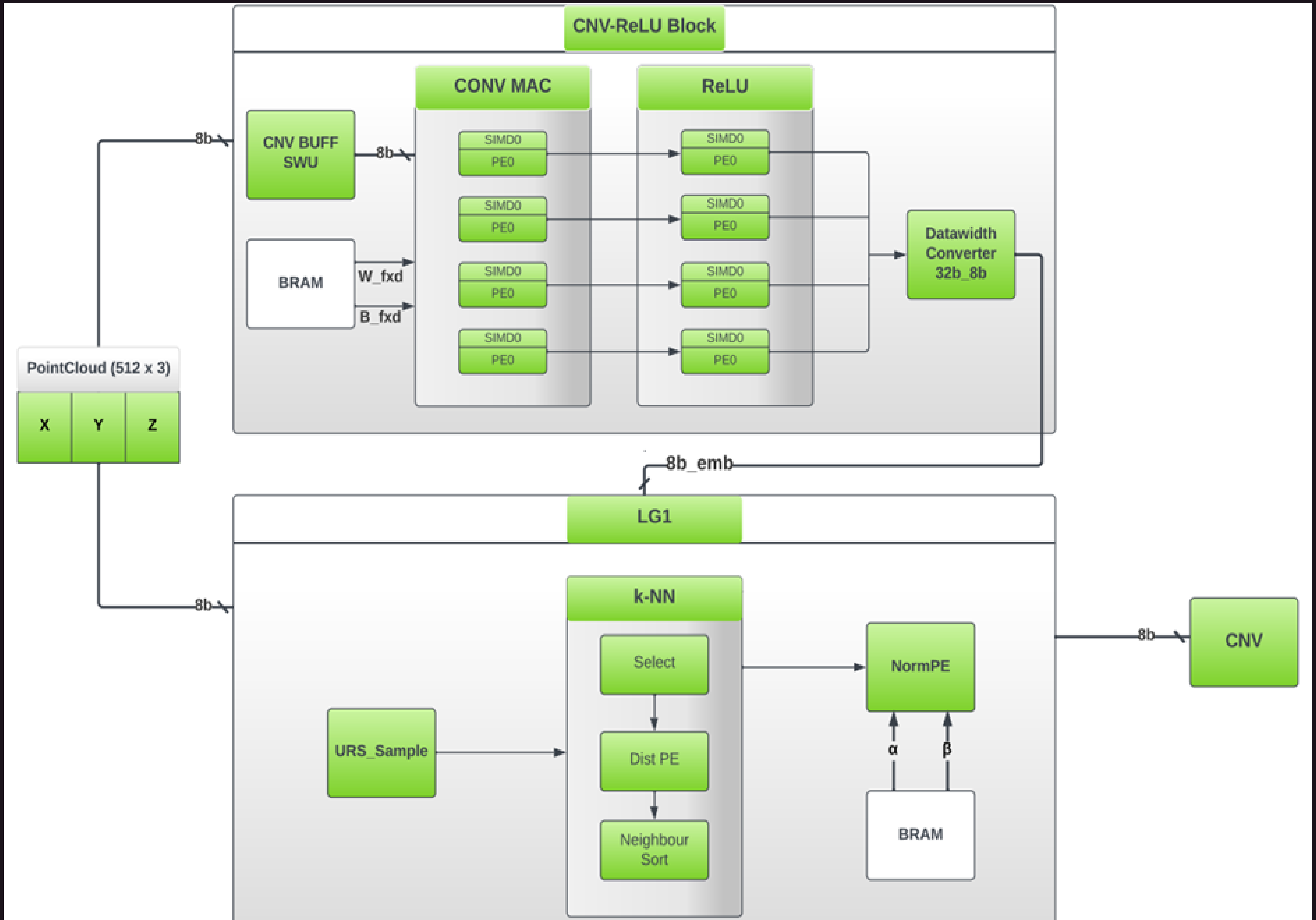
Precision	Geometric Parameters	Accuracy (%)	Size (KBs)	Precision	Geometric Parameters	Input Points	Accuracy (%)
FP32	α = FP32, β = FP32	93.07	2810	FP32	α = FP32, β = FP32	1024	93.07
W8A8	α = FP32, β = FP32	93.00	702.5	Fused W8A8	α = Q1.7, β = None	1024	92.75
W4A4	α = FP32, β = FP32	91.04	351.2	Fused W4A4	α = Q1.7, β = None	1024	92.34
Fused FP32	α = Q1.7, β = None	93.07	2,806.75	Fused W8A4	α = Q1.7, β = None	1024	92.18
Fused W8A8	α = Q1.7, β = None	93.23	702.0	Fused W6A2	α = Q1.7, β = None	1024	92.02
Fused W4A4	α = Q1.7, β = None	91.49	351.5	Fused W8A2	α = Q1.7, β = None	1024	88.57
Fused W8A4	α = Q1.7, β = None	92.79	702.0	Fused W8A8	α = Q1.7, β = None	512	91.98
Fused W6A2	α = Q1.7, β = None	90.70	527	Fused W8A8	α = Q1.7, β = None	256	91.69
Fused W8A2	α = Q1.7, β = None	90.64	702.0	Fused W8A8	α = Q1.7, β = None	128	74.07

Hardware Implementation

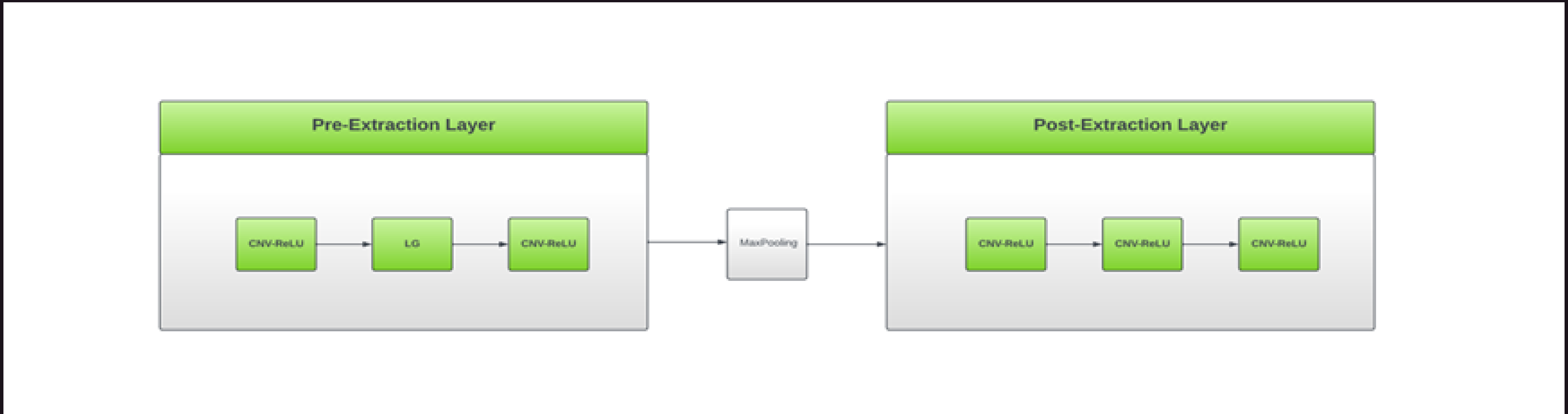
Local Grouper



Architecture



Architecture



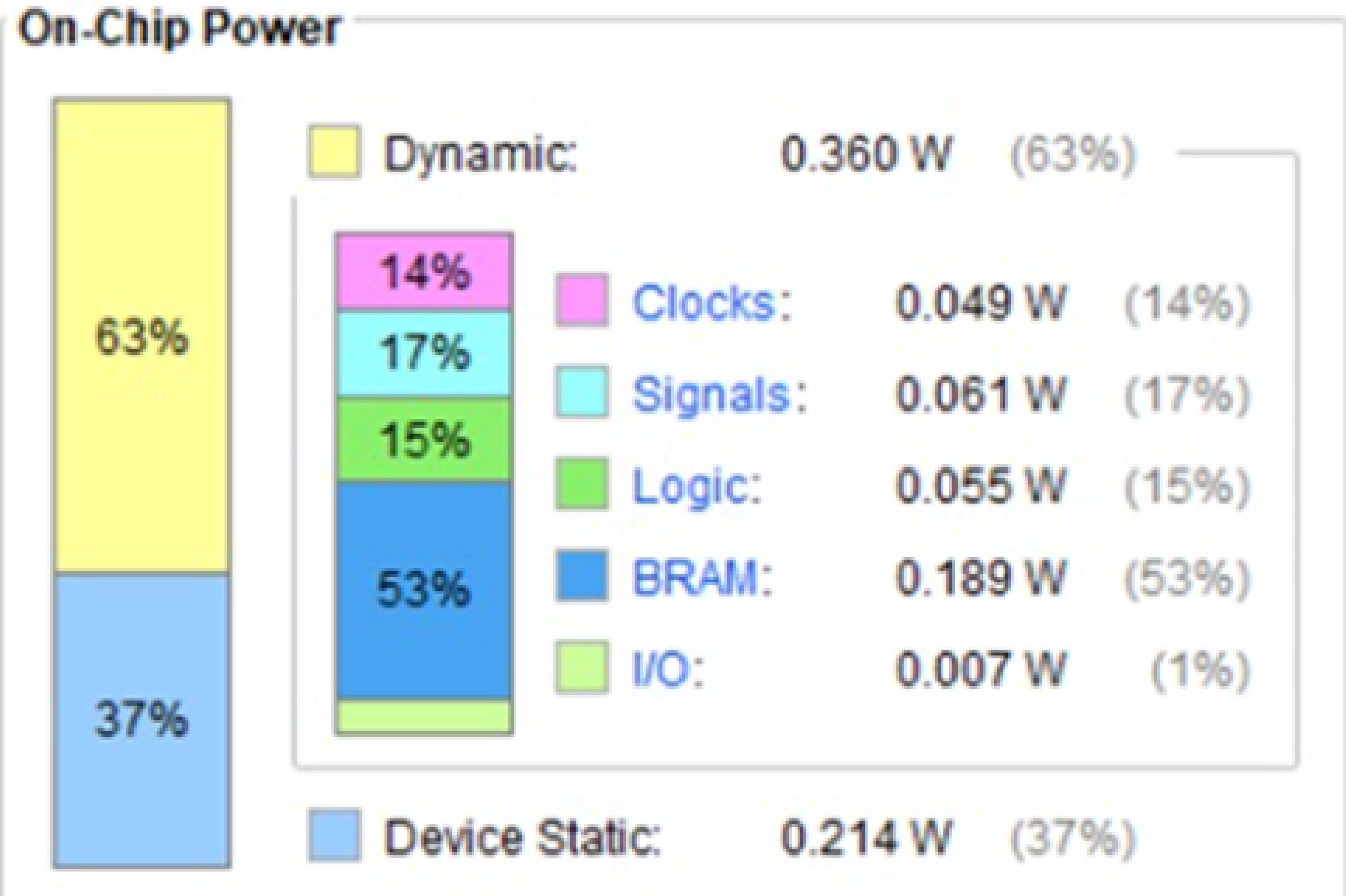
Block Design

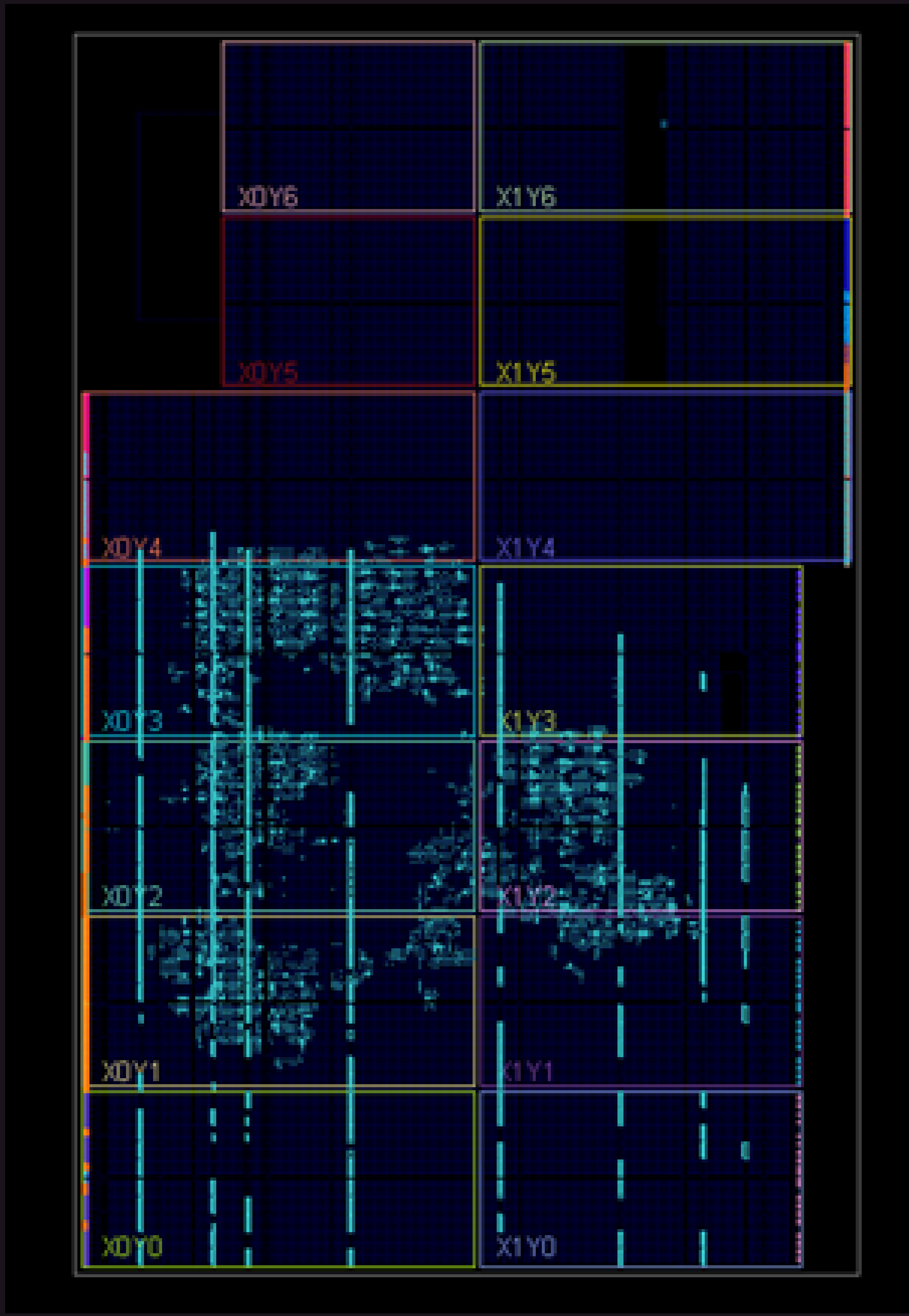
Hardware Results

Power & Placement

Power analysis from Implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

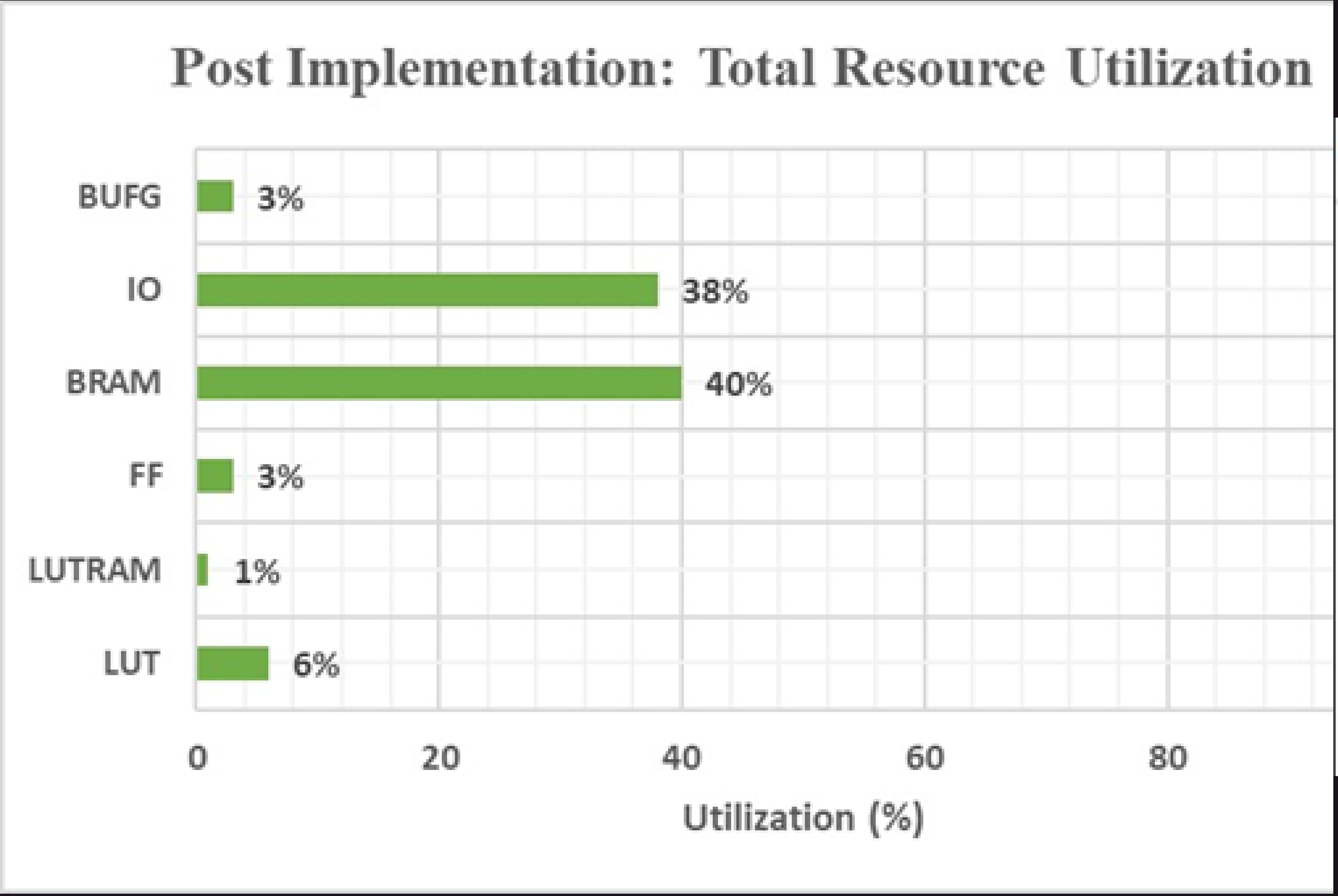
Total On-Chip Power:	**0.575 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**26.0°C**
Thermal Margin:	59.0°C (32.5 W)
Effective ϑJA:	1.8°C/W
Power supplied to off-chip devices:	0 W





Hardware Results

Timing & Resources



Design Timing Summary

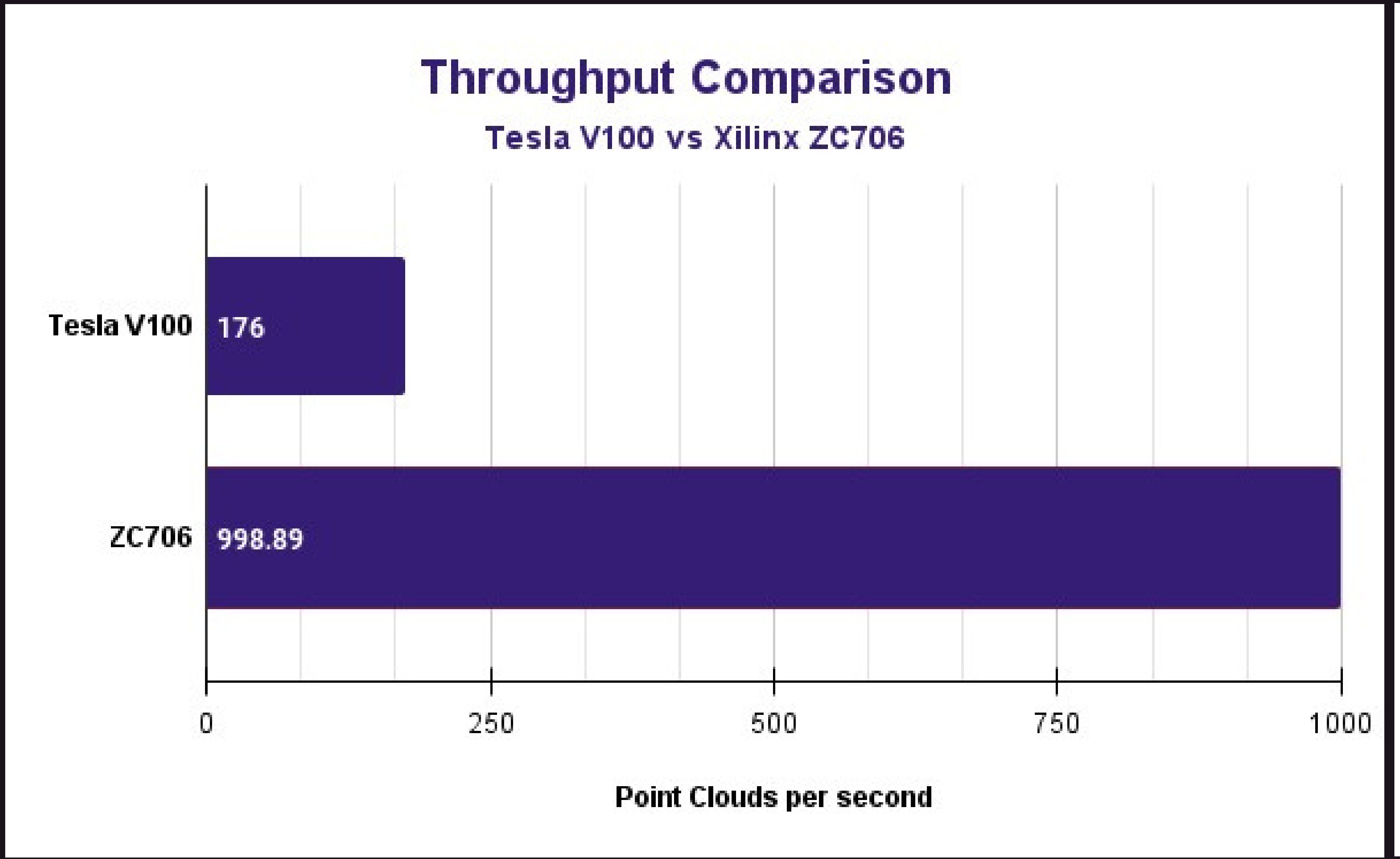
Setup

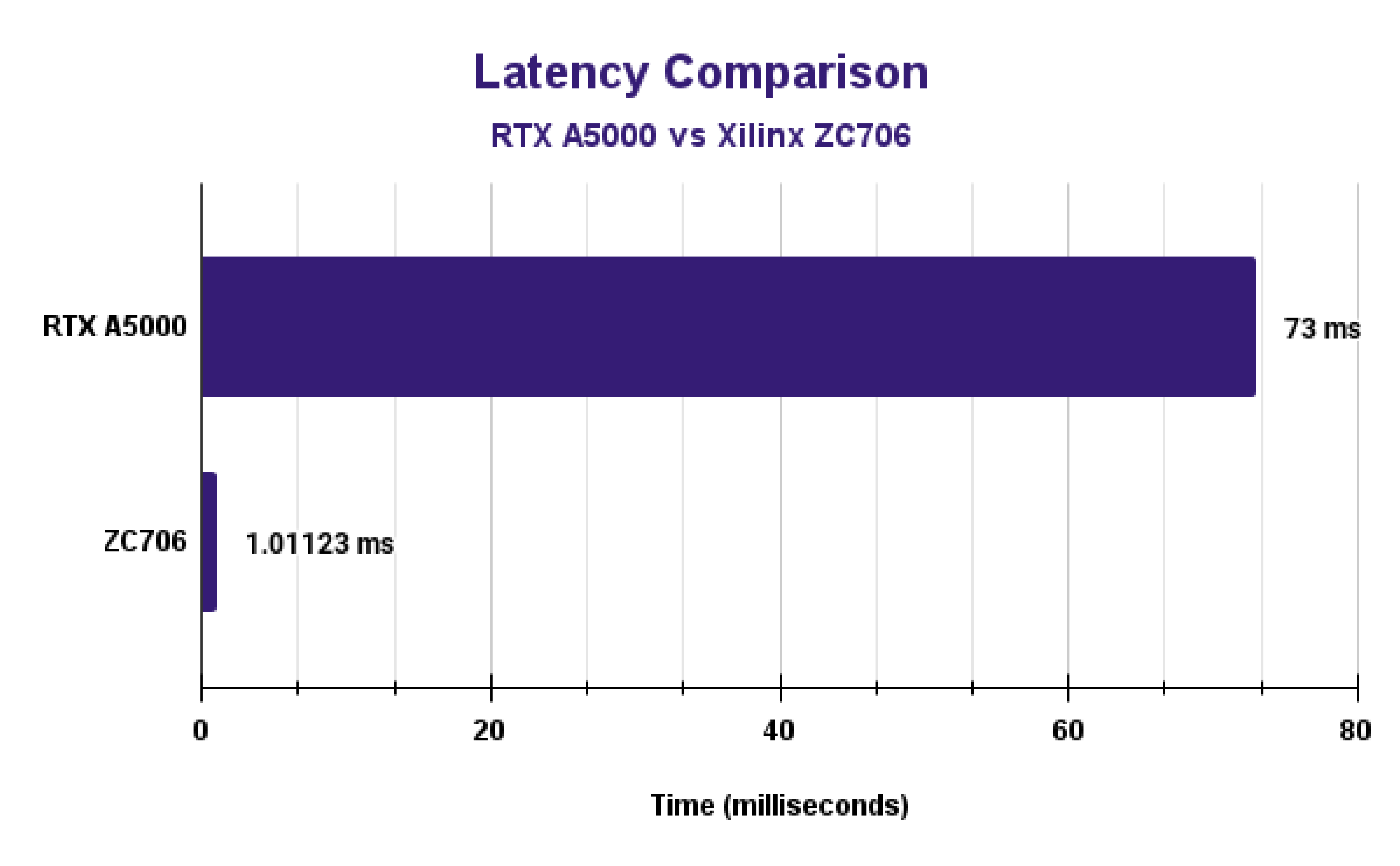
Worst Negative Slack (WNS):	2.377 ns
Total Negative Slack (TNS):	0.000 ns
Number of Failing Endpoints:	0
Total Number of Endpoints:	25867

All user specified timing constraints are met.

Hardware Results

Latency, Throughput, Speedup





NATIONAL UNIVERSITY OF SCIENCES & TECHNOLOGY PAKISTAN

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Demonstration

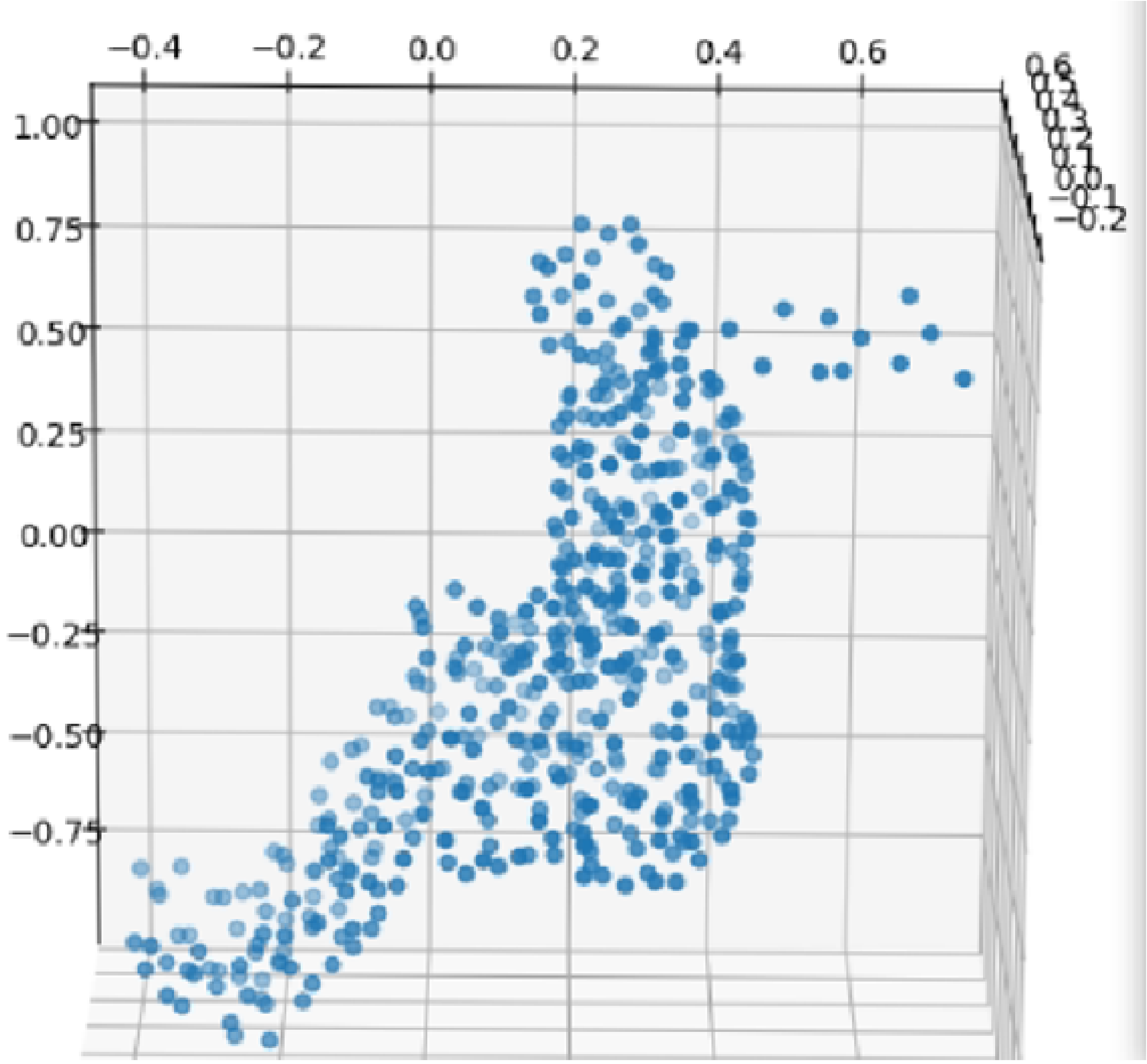
Inference on NVIDIA RTX A5000

```
Class:  person
Average time cost per sample:  0.073 seconds
```

Inference on Xilinx ZC706

```
xilinx@pynq:~/hwCode$ g++ -w dev.cpp PointMLPElite_run.cpp -o PointMLPElite
xilinx@pynq:~/hwCode$ sudo ./PointMLPElite
Latency:    1.01123 ms
Throughput: 988.896 point clouds per second

Class: 31 - person
Probability: 0.968750
```



ModelNet40 sample point cloud

Conclusion

- In this project, we demonstrated that FPGAs are a suitable platform for the acceleration of 3D point cloud deep learning models.
- We also developed an HLS framework for deploying such models on FPGAs, with a set of common layers and algorithms.
- We learned a lot through the challenges, and we thank our supervisors for their help.
- Any questions?

Deliverables Upload Status

Submission Status

Deliverables Submission Status

- Project Report - **Submitted**
- Project Demo **- Submitted**
- Plagiarism Report**- Submitted**
- Project Poster **- Submitted**
- Project Presentation **- Submitted**