

Methods

General:

In each of the architectures, we will be deploying ResNet-50, an image recognition Convolution Neural Network (CNN) model. The model was evaluated across different hardware platforms to compare performance and deployment efficiency.

CPU:

ResNet-50 was deployed on 4th Gen Intel® Xeon® CPU using both OpenVINO and PyTorch to compare inference performance. This allows evaluation of latency, throughput, and overall efficiency.

GPU:

The experiments were conducted using the Google Colab environment, which provides access to GPUs (including the A100) hosted on a remote server farm.

Architecture Comparison

CPU:

The CPU platform used is the 4th Gen Intel® Xeon® Scalable processor (Sapphire Rapids), deployed through an Amazon EC2 c7i.xlarge instance. This CPU is selected because it is optimized for AI workloads and it is fully compatible with frameworks like OpenVINO, PyTorch, and TensorFlow,

GPU:

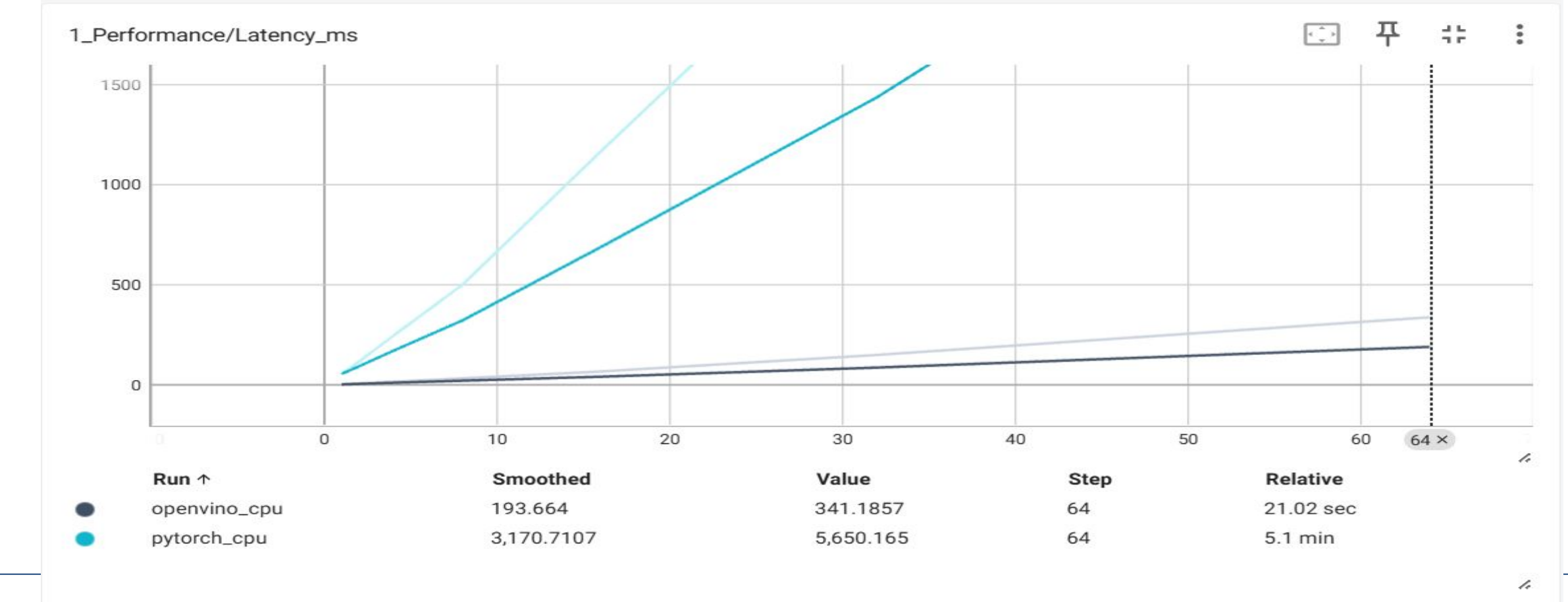
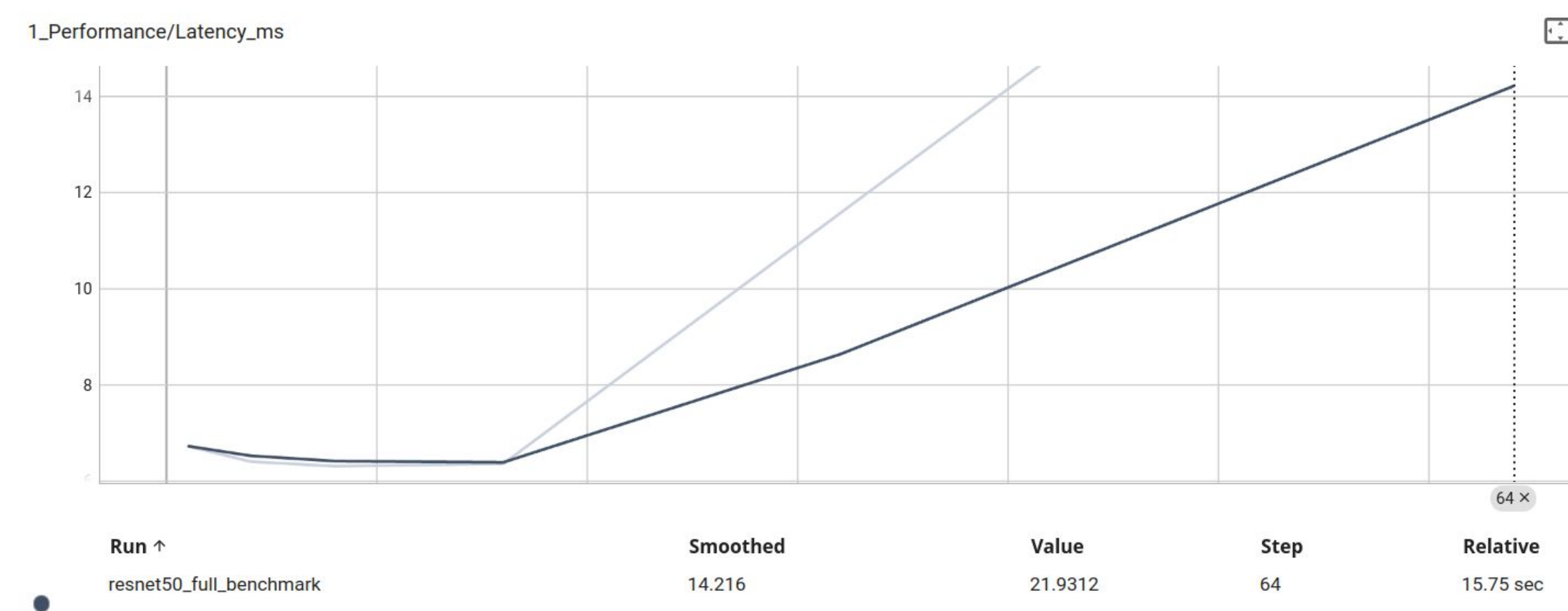
The GPU that is being deployed is NVIDIA's A100. The reason as to why this one is chosen is because the A100 is the industry standard. The A100 has one of the higher performance to cost ratios, scalability for data centers, and is unrestricted for data centers (compared to the RTX series).

FPGA:

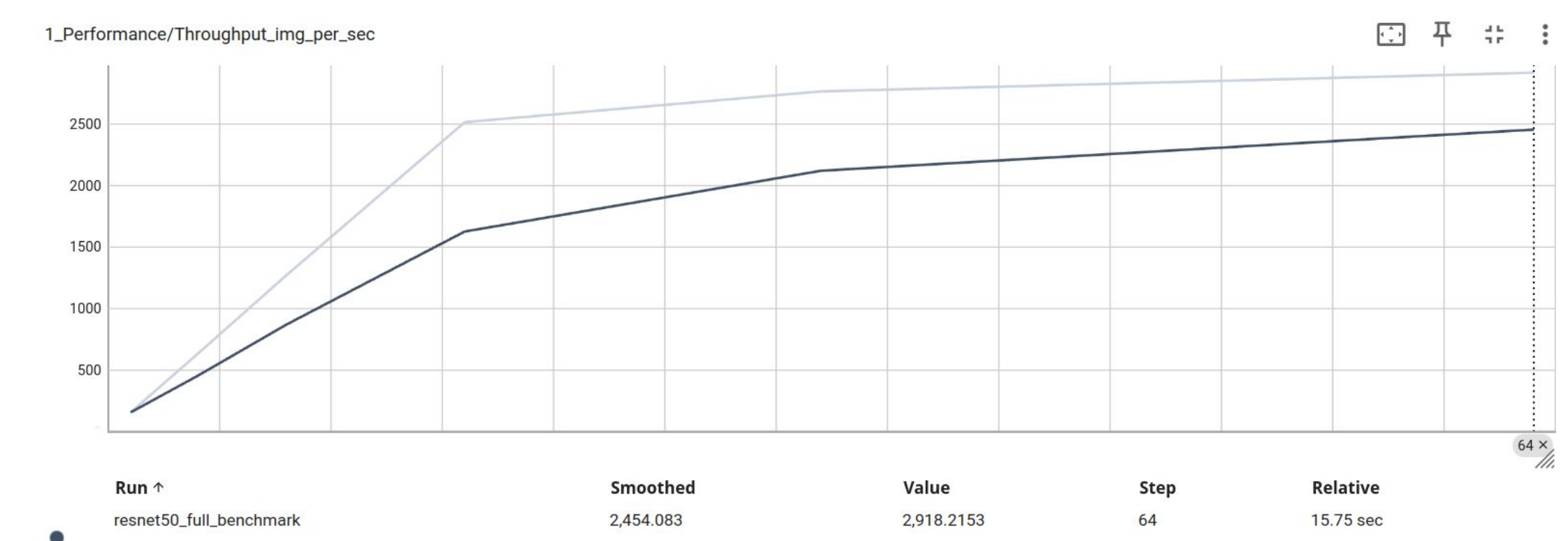
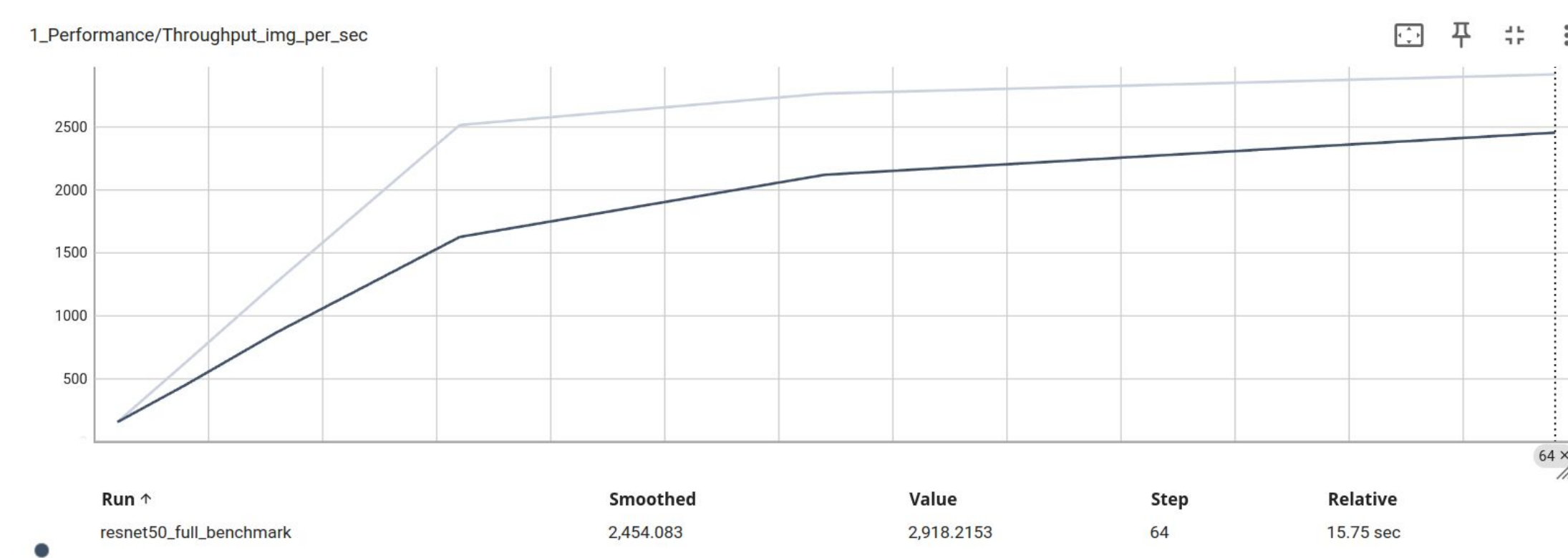
The FPGA devices that are being deployed is the Agilex 5 and Agilex 7. The Agilex 3 is meant more for small-scale AI deployment while Agilex 7 has scalability for data centers.

Experiment Results

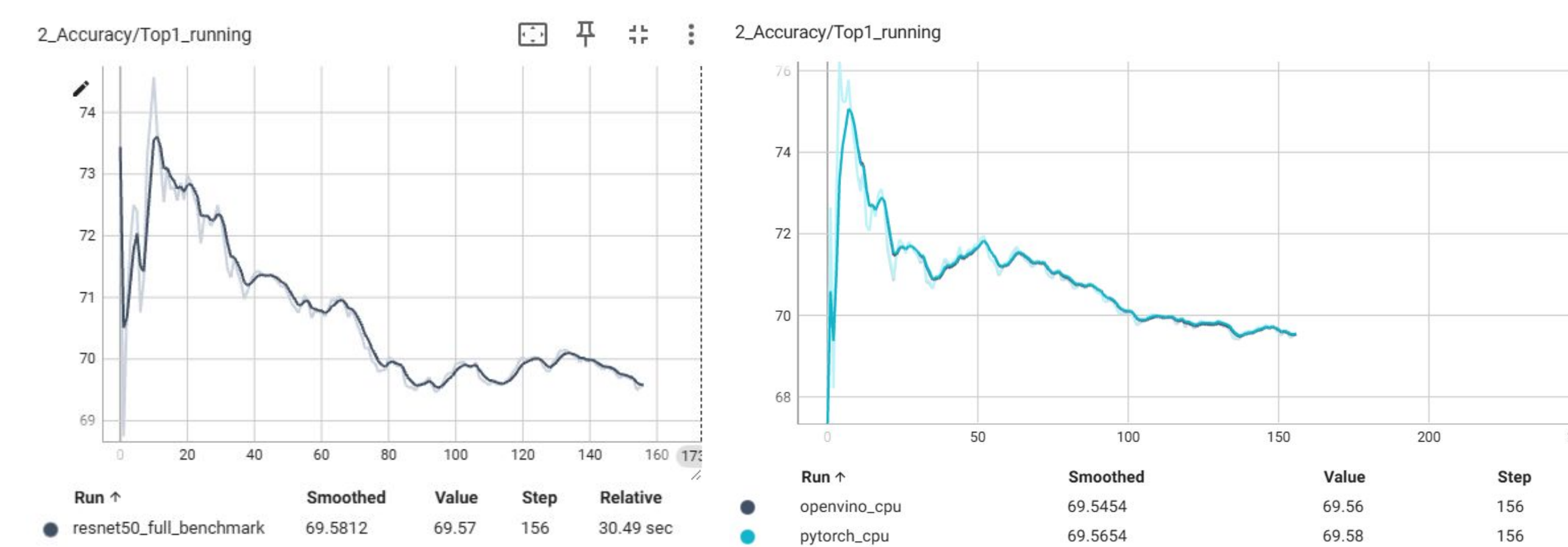
Latency: X-Axis = Batch Size, Y - Axis = Time (ms)



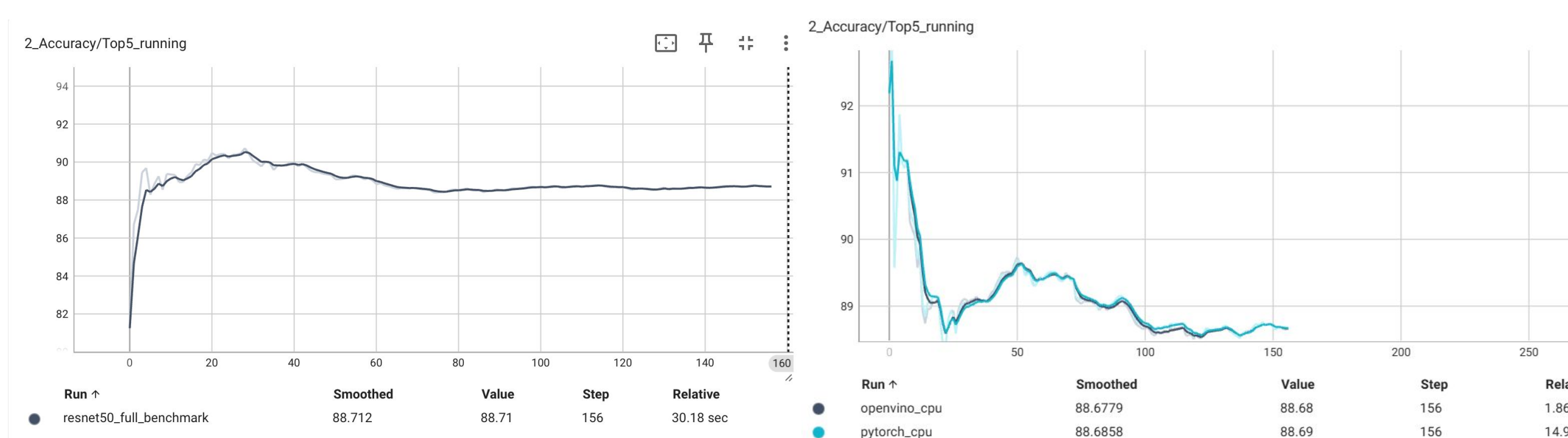
Throughput: X-Axis = Batch Size, Y - Axis = Throughput (images/second)



Top 1 Accuracy (%)



Top 5 Accuracy (%)



Benchmarking Metrics

- Latency: The time it takes to do one inference
- Throughput: The amount of images processed in one second
- Top 1: The model's most confident response to what the image is
- Top 5: The model's top 5 responses to what the image is
- Energy efficiency
- Memory Utilization

Future Implications

- FPGA inference will likely perform better in latency and throughput
- FPGA inference will likely be less accurate than the other devices
- Further optimization techniques will likely increase the FPGA performance with CNN models

References

