Performance Evaluation of TPUs and FPGAs for Deep Neural Network Inference

Deepak Kumar Athur, Computer Engineering
Mentor: Dr. Aman Arora, Assistant Professor
Ira A. Fulton Schools of Engineering

OBJECTIVE

- Deep Neural Networks have become ubiquitous in our lives, with applications in computer vision, robotics, text-to-speech, etc.
- Several hardware platforms are available for deep neural network inference. They include FPGAs (Field Programmable Gate Arrays), Graphic Processing Units (GPUs) and Tensor Processing Units (TPUs).
- There is a lack of quantitative performance comparison when it comes to FPGAs and TPUs. Such a comparison can help identify the tradeoffs between these platforms enabling informed platform choices for different application scenarios.
- The main objective of this project is to compare FPGA and TPU in terms of performance (Inferences/second), energy, energy cost per inference, and flexibility.

METHODOLOGY

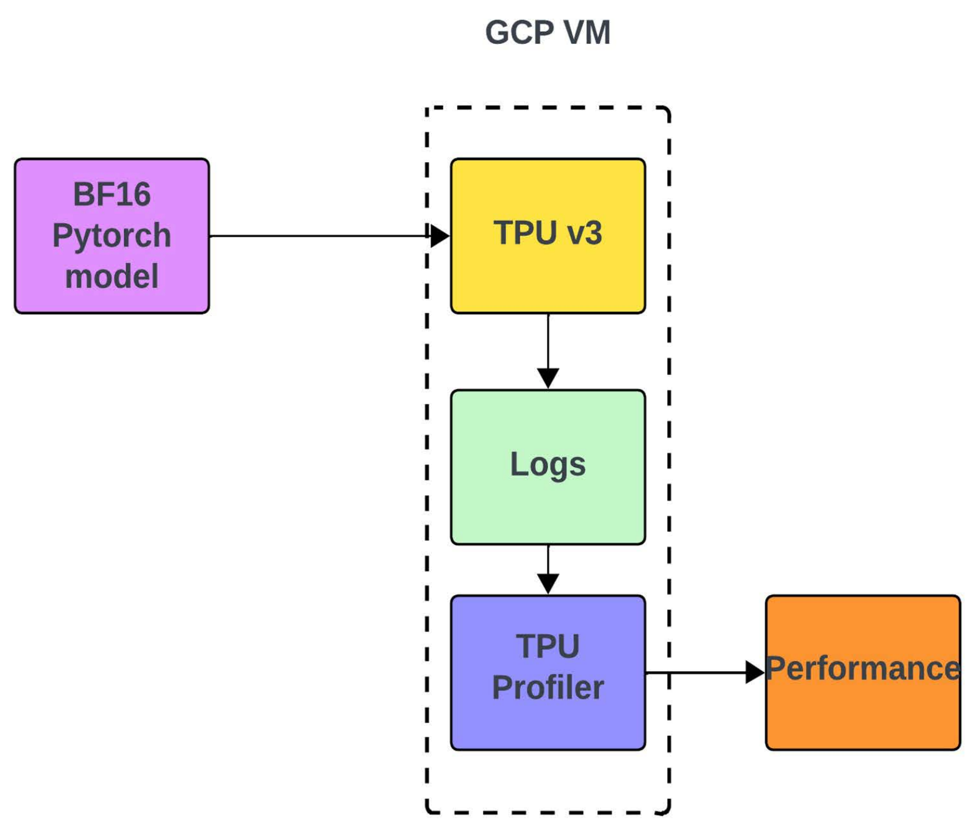
Device Selection

Device	Name	Tech Node	Power	On Chip Memory	Peak TOPS
Cloud	TPU v3	16 nm	220 W	32	123
	U55c	16 nm	115 W	43	26
Edge	KV 260	16 nm	15 W	26.6	3.3
	Edge TPU	-	2 W	28	4

Benchmarks

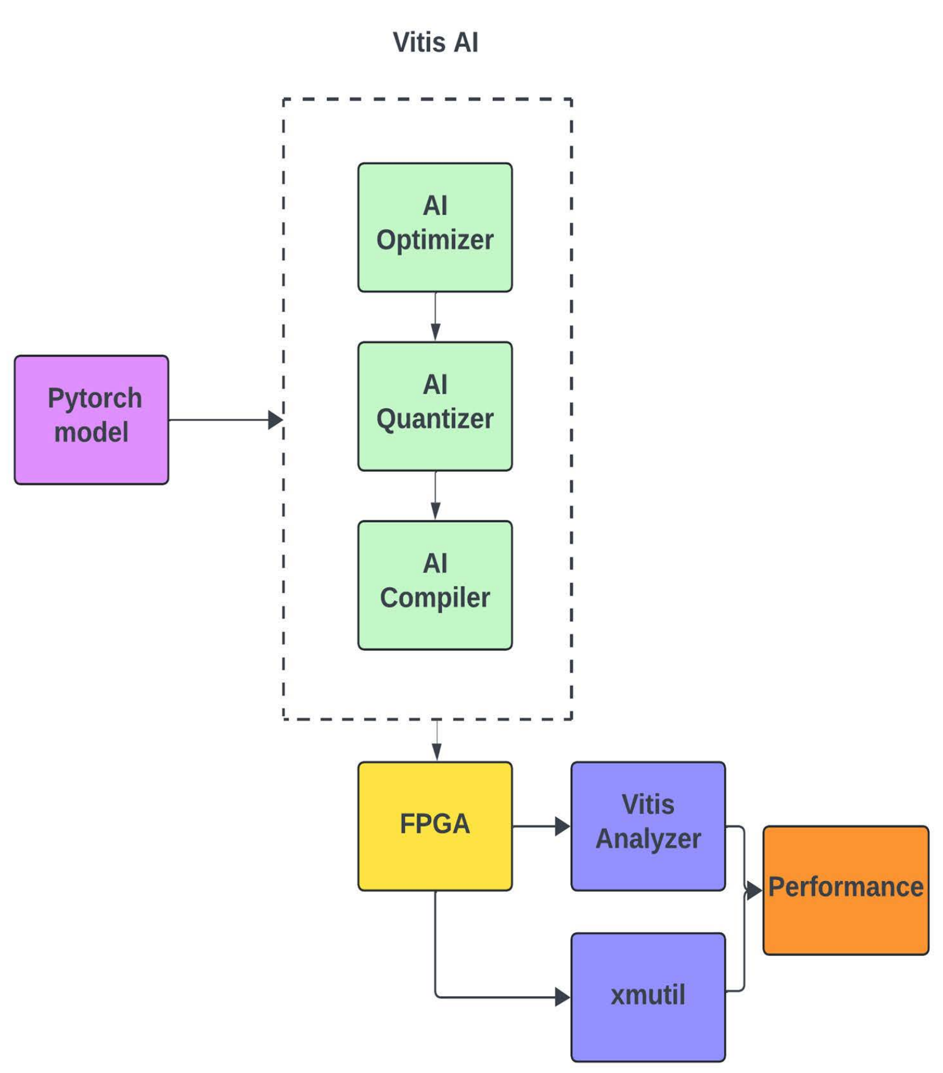
Resnet50
Bert small
Mobilenet_v2
Vision Transformer
VGG16, VGG19
Inception V4

TPU Implementation



- Pytorch AI models from torchvision library and hugging face were evaluated on TPU v3.
- This was done using Virtual Machines (VM) from Google Cloud Platform (GCP) that had TPU compute units.
- The profile of a workload run was captured and analyzed in TPU profiler.
- Memory and compute utilization were analyzed for perf ormance.

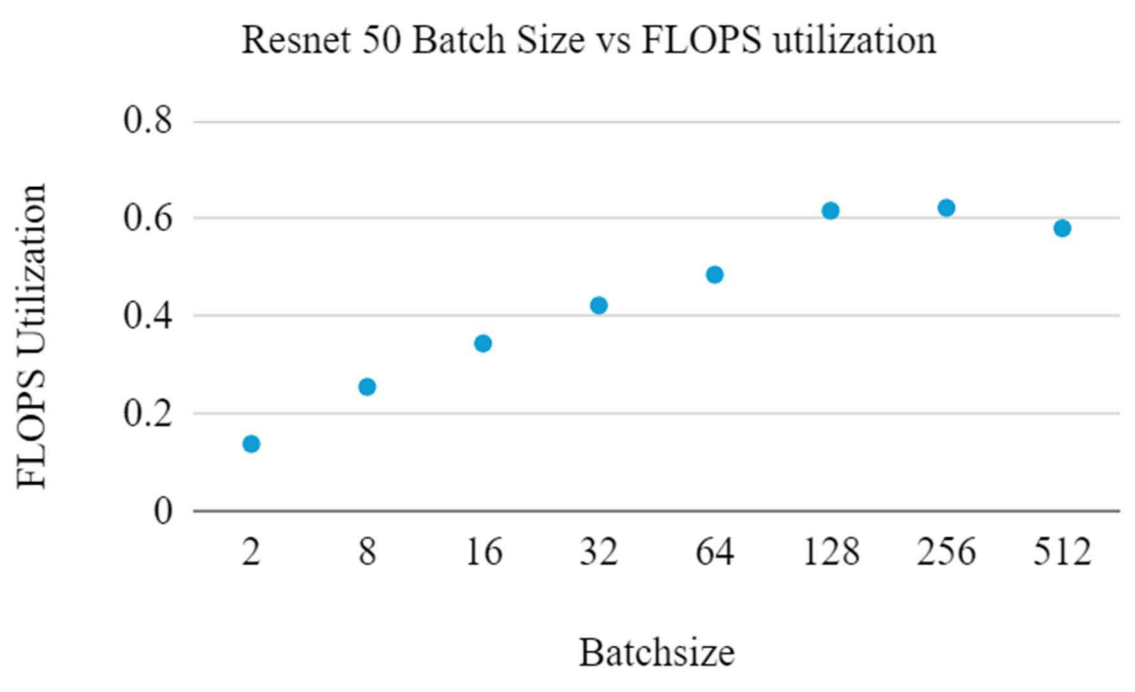
FPGA Implementation



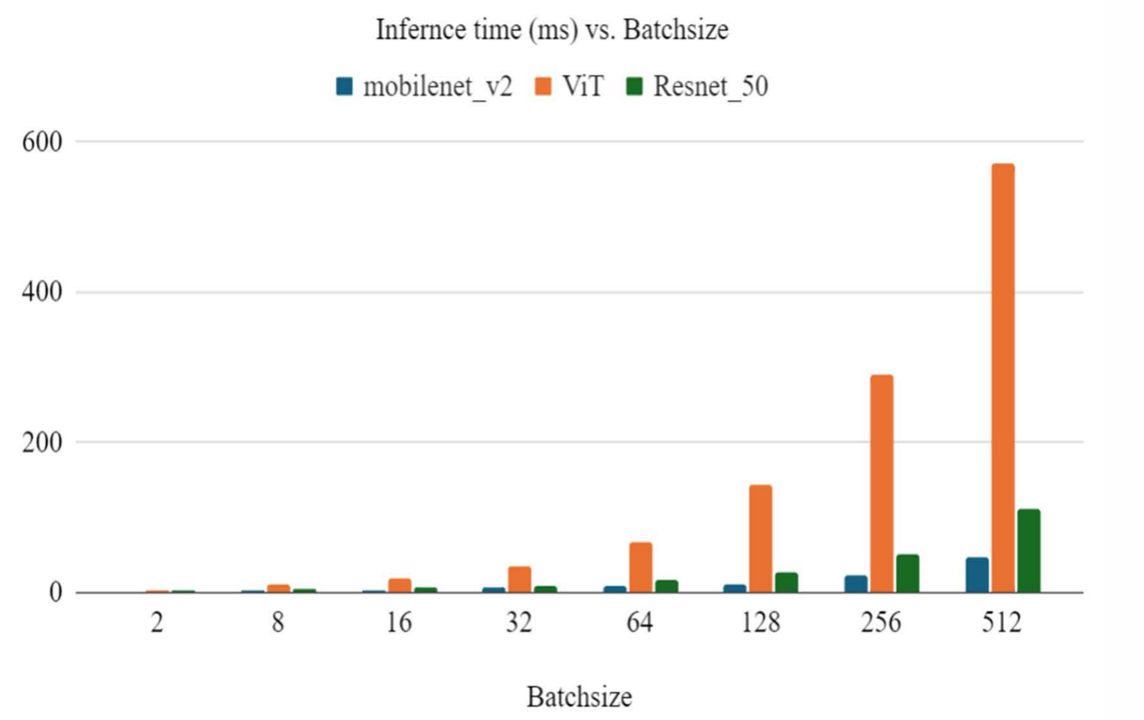
- Vitis AI, an AI inference solution by AMD Xilinx that can support various cloud and edge FPGA platforms, was used to deploy ML models and run inference.
- The Optimizer and Quantizer prune the model and convert floating-point models into fixed-point models which require less memory bandwidth and operations. The Compiler generates the binary/bitstream for the FPGA.
- Vitis analyzer is used to capture the complete AI data pipeline to analyze performance. Xmutil is used for power measurement from the on-chip power IC.

PROGRESS

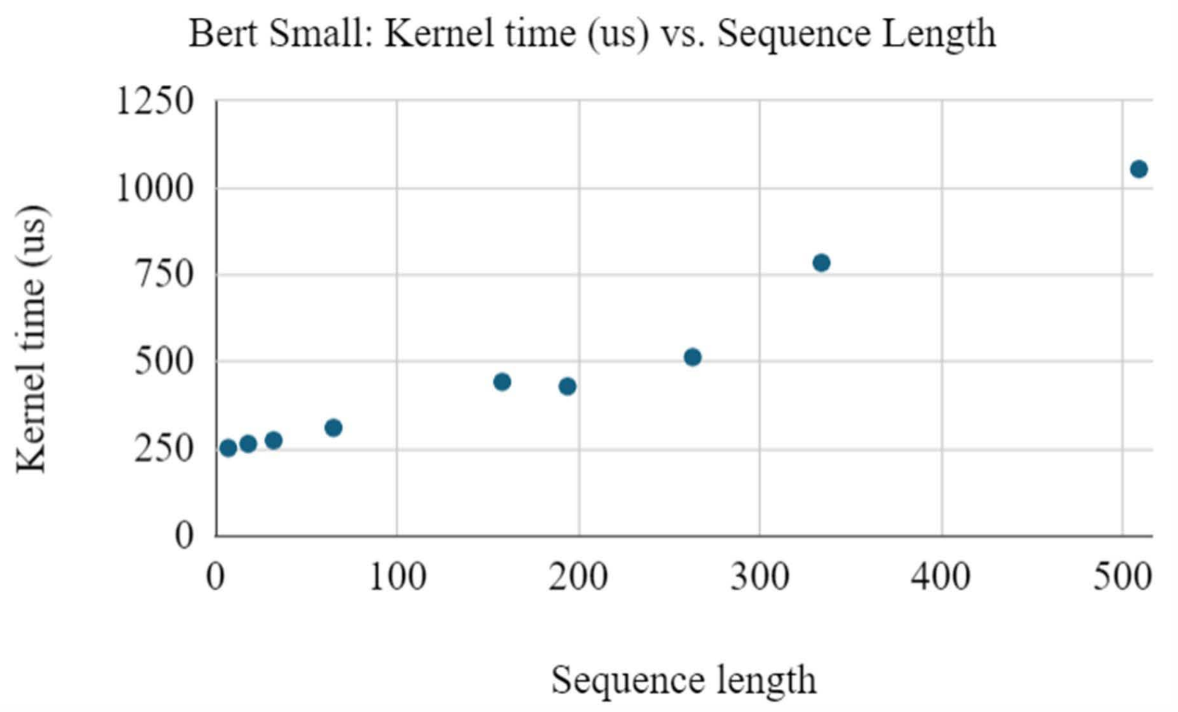
Cloud



For Resnet50, the matrix multiplier utilization plateaus and starts decreasing after a particular batch input size



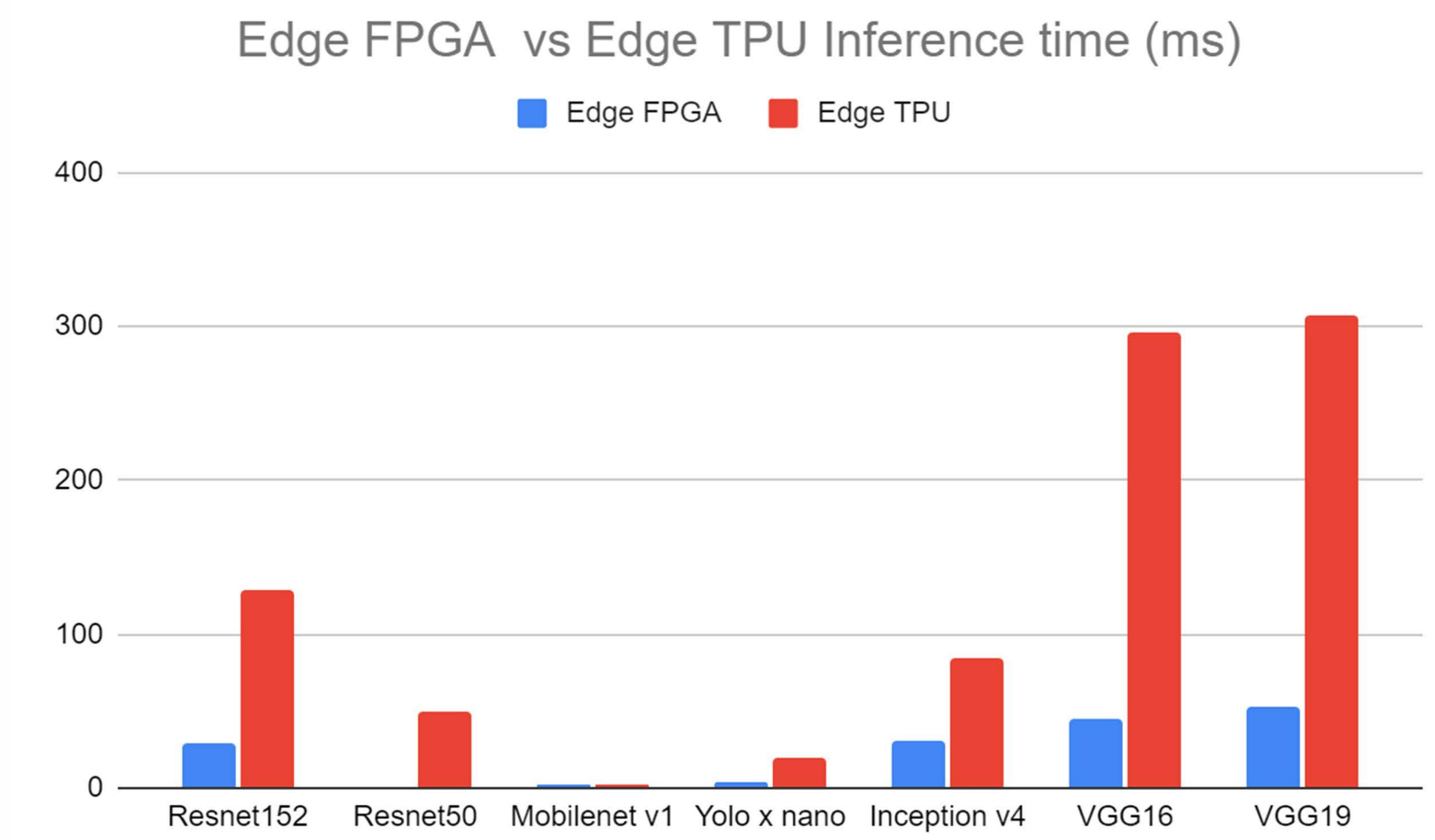
For all the models, an increase in batch size, inference time increases. Throughput also increases with batchsize.



When it comes to transformers like Bert, the inference time grows linearly (quadratic expected) with sequence length.

Execution of benchmarks on Cloud FPGA is in progress.

Edge



The Edge TPU is approximately 80% percent slower than Edge FPGA on average. Hardware utilization and power consumption results need to be analyzed for a more realistic comparison.

CHALLENGES FACED

- Xilinx FPGA available in Microsoft Azure had set up issues and hence couldn't use Vitis AI. Xilinx FPGA available in AMD research cluster (HACC) could not be used because Vitis AI needs container support. We have purchased our own FPGA now.
- No means to measure Cloud TPU power.
- Normalization (similar devices, same benchmarks) for a valid comparison has been difficult.

FUTURE RESEARCH

- Extend the performance comparison for cloud FPGA and cloud TPU
- Customize the FPGA design to a specific model to improve performance.

ACKNOWLEDGEMENT

I thank Dr. Aman Arora for his unwavering guidance, invaluable insights, and tireless support throughout this research project.