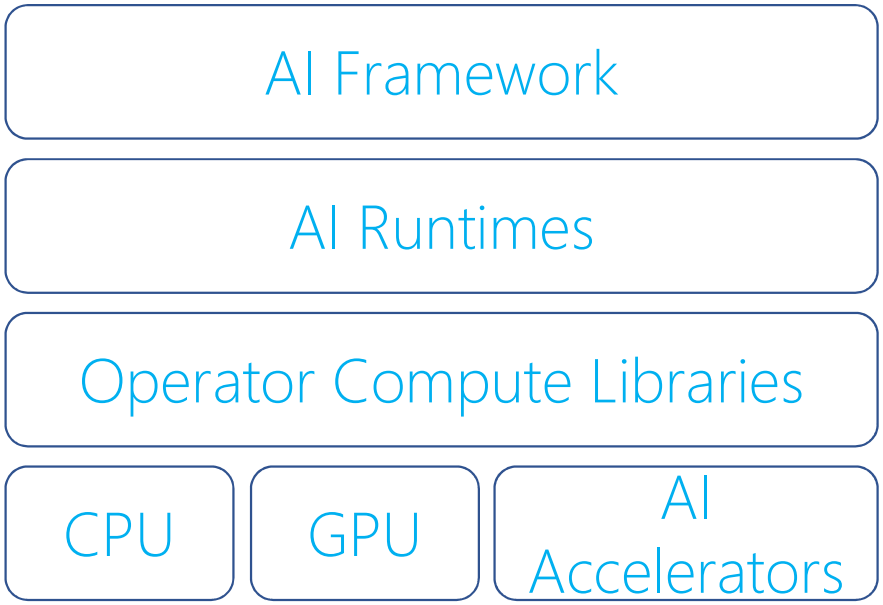
Beyond IPS: Towards a wholistic Measure of Machine Learning Performance

Beyond IPS

The views expressed in this discussion are those of the people delivering the talk. – We are not speaking for our employers.

AI Software and Hardware



AI Landscape

AI is now in core of the platforms

- Apple - CoreML
- Microsoft - WindowsML
- Google – MLKit, Android NN Runtime
- Intel – DLBoost

Goal:

1. Enable development of no frills AI.
2. Bring model from anywhere and run it on any platform

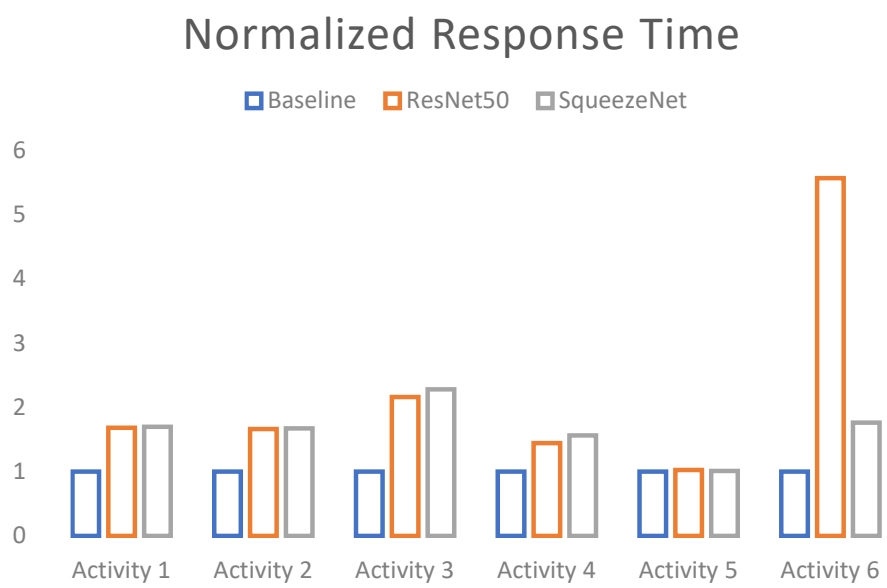
With so many options how do you compare solutions?

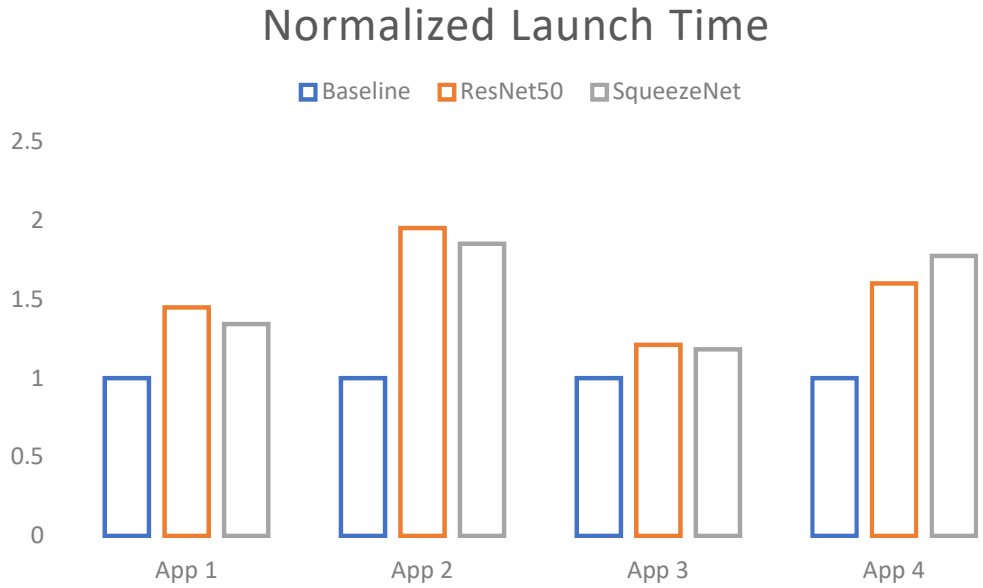
Common Metric: Inferences/sec

Pitfalls of Common Metric

Scenarios	Practical Considerations
Latency Sensitive application	Start up Time
Always On Edge Application	Performance/Watt
Distributed Application	MultiNode Scaling Efficiency
Heterogeneous Application	Multi Device Scaling Efficiency
Hybrid AI Application	Container Performance
Home Surveillance Application	Security
Gaming Application	CNN, RNN, Reinforcement Learning
Mobile Application	Lower GFLOPS, Memory=Precision, Sparcity

Impact of Inference





Target workloads are generic productivity tasks. Do not have permission to provide specifics.

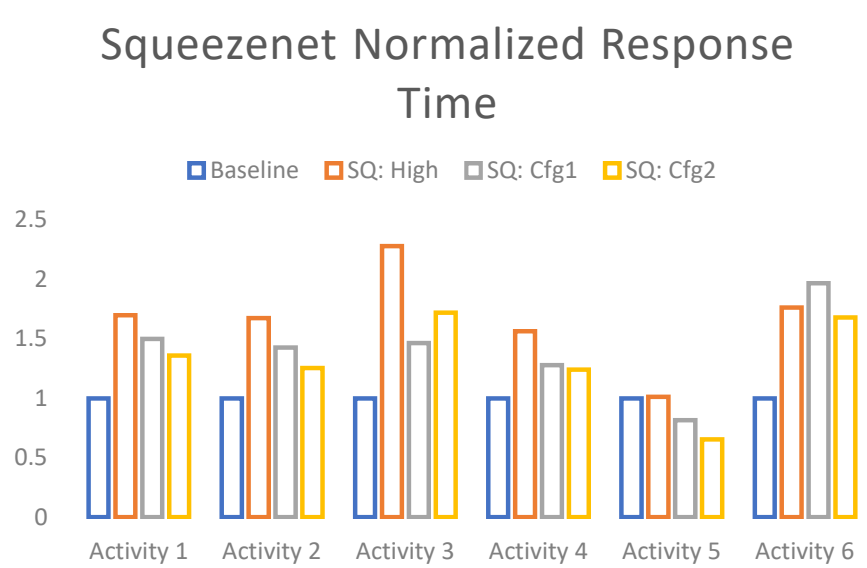
Squeezenet and Resnet50 represent the family of topologies. These are HW optimized versions of the versions described by their authors. No claims made on accuracy of their implementation.

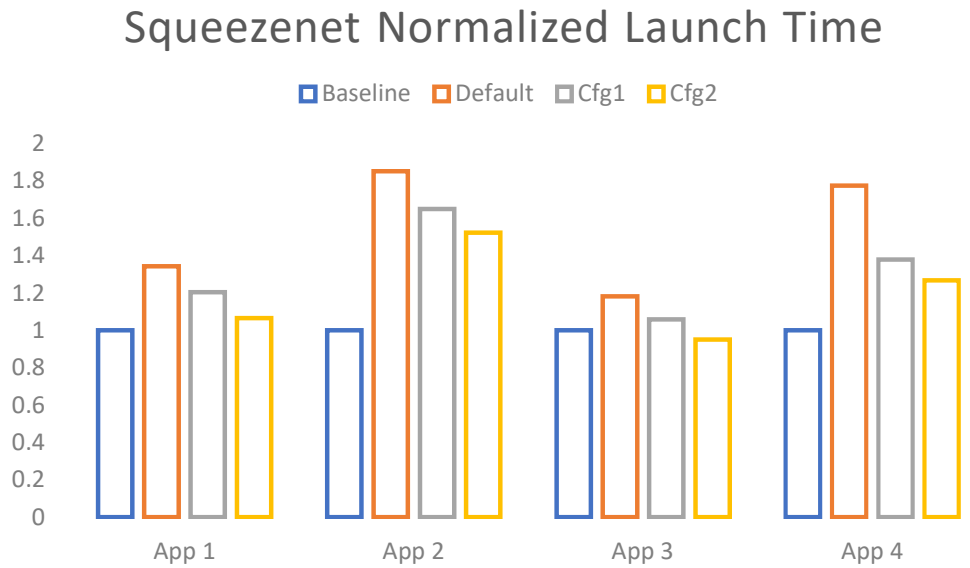
There are many way an AI Solution can do inference.

1. An AI solution can processes a graph using a ML graph compiler and implement optimization techniques and generate a very efficient code with very different compute and memory profile.

2. This impacts how devices, cores their ISA is used. It dictates how memory peaks and translates to how overall energy is utilized.

Impact of Threading when optimizing AI

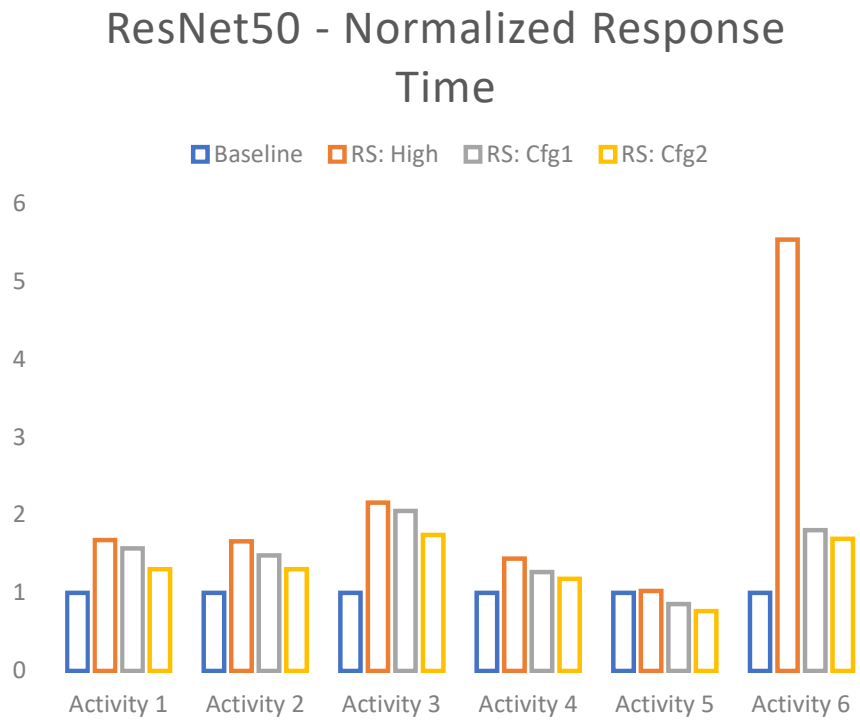


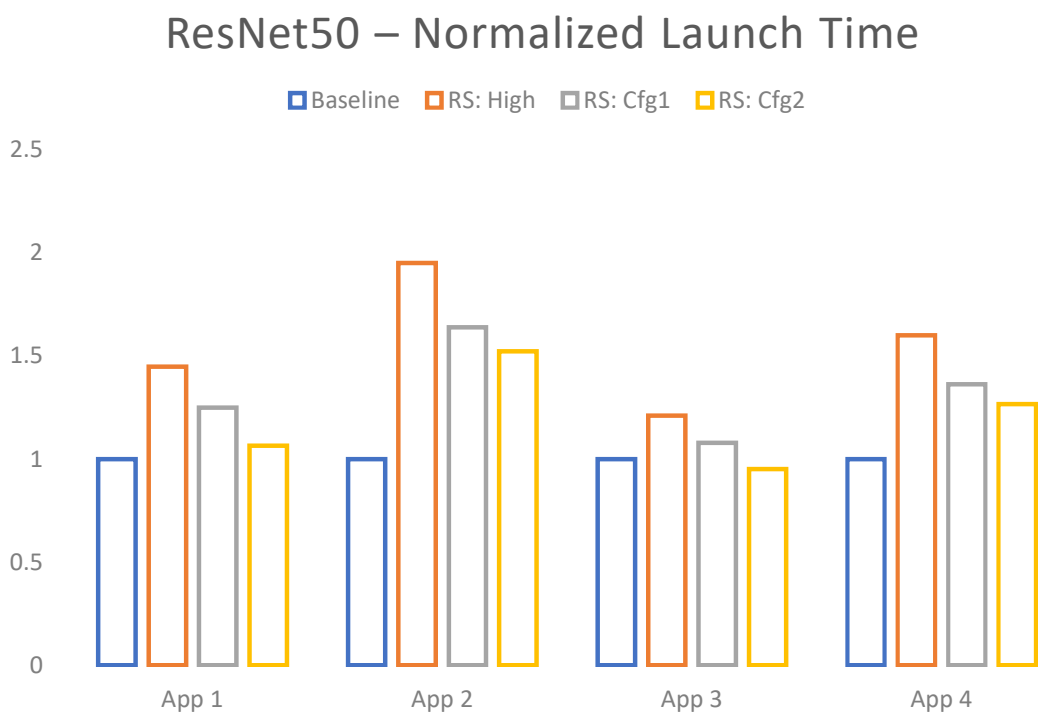


Target workloads are generic productivity tasks. Do not have permission to provide specifics.

Squeezenet and Resnet50 represent the family of topologies. These are HW optimized versions of the versions described by their authors. No claims made on accuracy of their implementation.

Impact of Threading when optimizing AI





Target workloads are generic productivity tasks. Do not have permission to provide specifics.

Squeezenet and Resnet50 represent the family of topologies. These are HW optimized versions of the versions described by their authors. No claims made on accuracy of their implementation.

Next Steps

- Gaming Workloads
- Browsing Workloads.
- RNNs
- Quantization
- GPU and AI Accelerators

Backpage