Hardware Sensitivity Analysis for Deep Learning Training Workloads

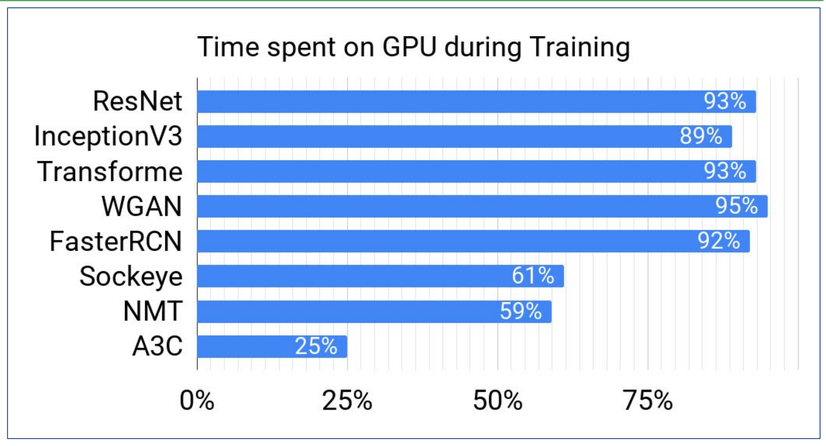
Anand Jayarajan, Alexandra Fedorova
University of British Columbia

How to speed up DL training???

- Deep learning model training is compute intensive and could take from hours to months to finish.
- Can more expensive GPUs always speed up the training?
- How does other hardware components affect the performance of DL training?
- Where are the bottlenecks?

The Need for Diversity

Deep learning models vary widely in their hardware requirements



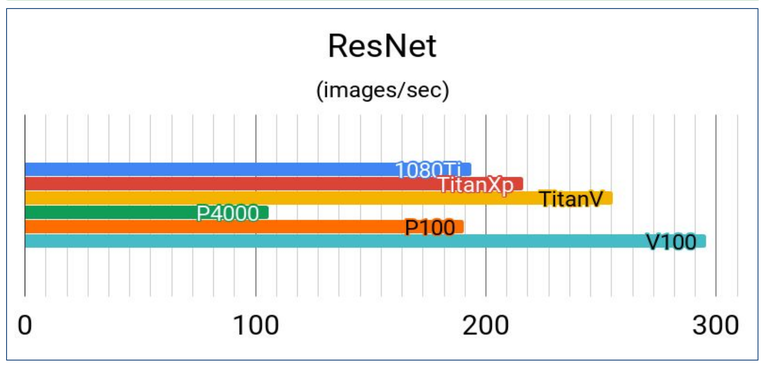
Applications and Models

Applications	Models	Dataset	# of layers	Dominant layer
Image Classification	ResNet, Inception	ImageNet	50 (152 max) 22	CONV
Machine Translation	Sockeye/NMT, Transformer	IWSLT15	5 12	LSTM Attention
Object Detection	Faster RCNN	Pascal VOC	101	CONV
Unsupervised Learning	WGAN	ImageNet-small	101+101	CONV
Reinforcement Learning	A3C	Atari 2600	4	CONV

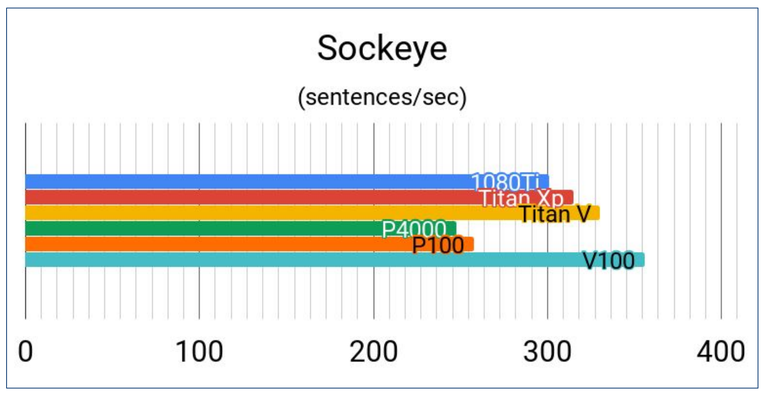
Hardwares

Specs	TITAN Xp	TITAN V	1080 Ti	P4000	P100	V100	Vega Frontier
Peak FP32 TFLOPS	12.1	12.2	11.3	5.2	9.3	15.7	13.1
Memory (GB)	12	12	11	8	16	16	16
Memory Bandwidth (GB/s)	547.7	652.8	484	243	732	900	484

Throughput Analysis

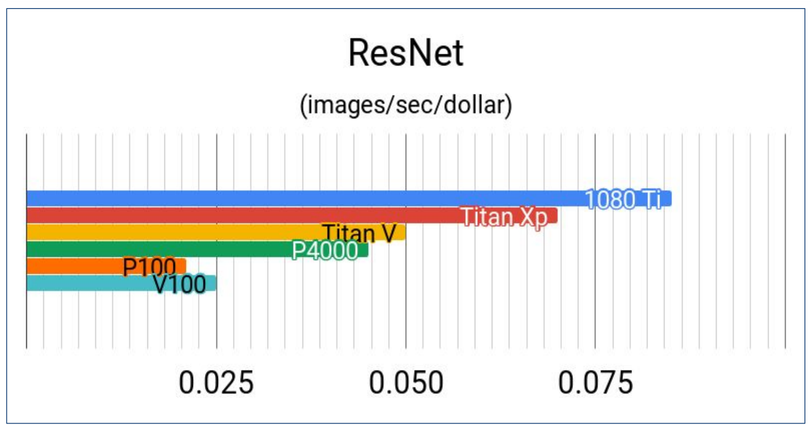


Convolution networks scale well on GPUs



RNN models don't scale well on GPUs

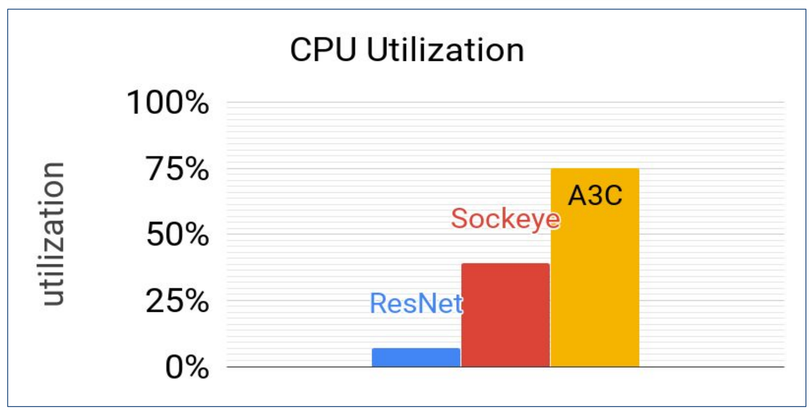
Cost Analysis



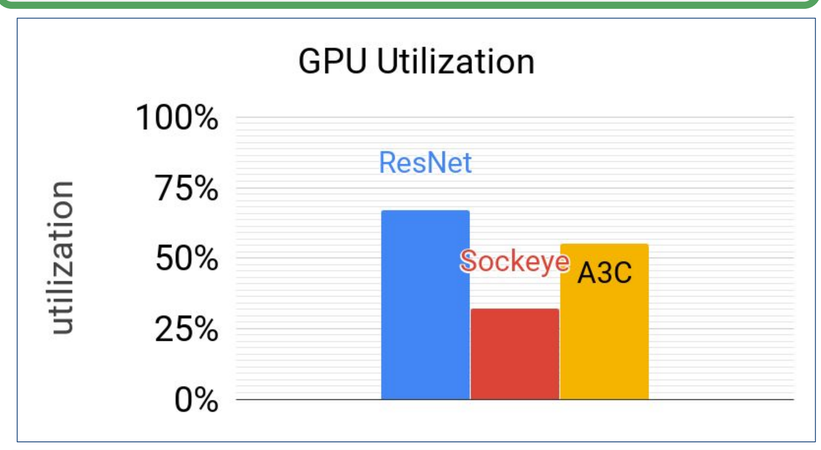
Gaming GPUs are more cost effective

- All experiments are conducted on a common CPU configuration.
- Measurements are verified by training the model until it reaches reported accuracy.
- Cost analysis is done by considering the cost of GPU and CPU.
- GPU/CPU utilization, cycle stalls are calculated based on the measurements reported by *nvprof* profiler.
- CPU measurements are taken using *Intel VTune*.

CPU/GPU Utilization



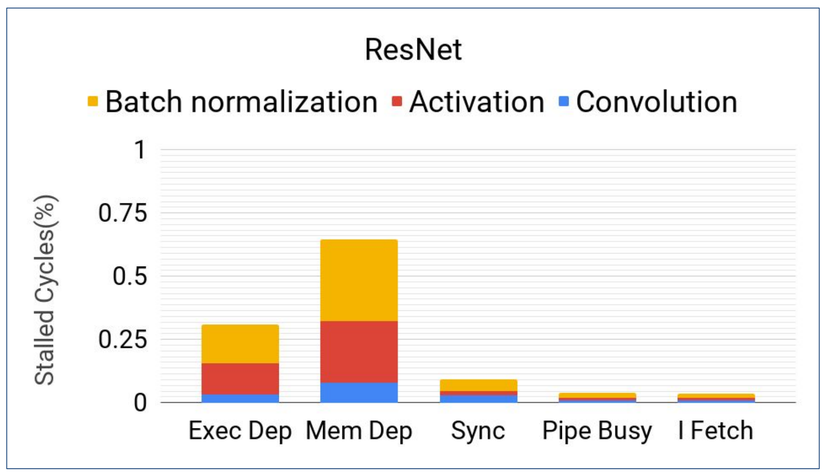
RNN and RL models have high dependency on CPU

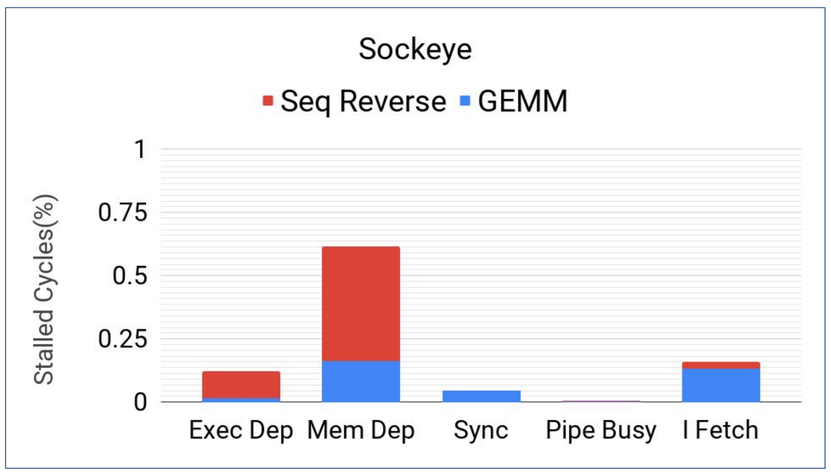


RNN models are very inefficient in utilizing GPUs

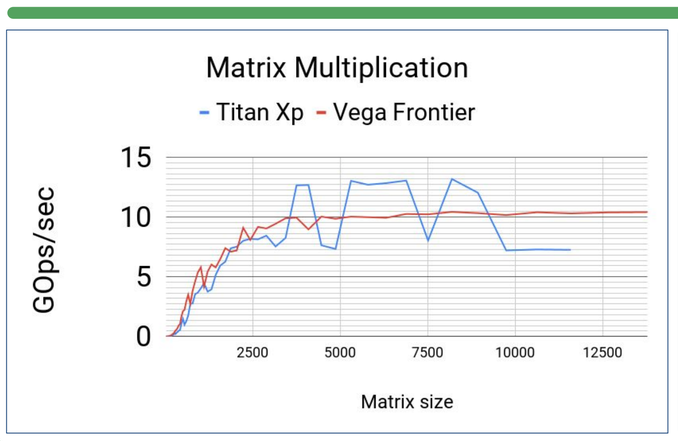
Source of inefficiencies

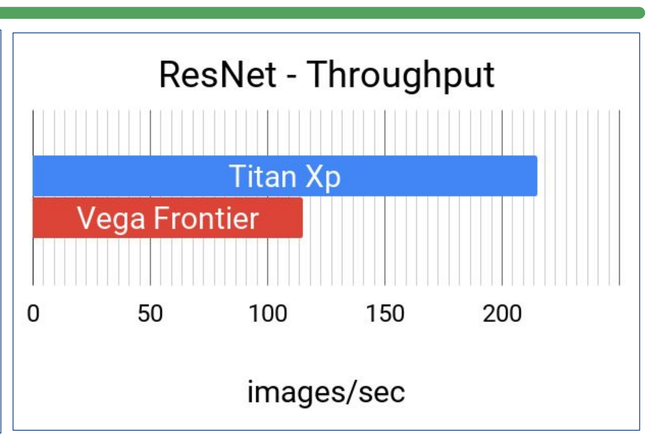
Stalls	Reasons
Execution dependency	Input required by the next instruction is not available
Memory dependency	Memory resources are unavailable
Synchronization	Warp blocked at __syncthreads() call
Pipe busy	Compute pipeline is busy
Instruction fetch	Instruction fetch buffer is empty





Nvidia vs AMD





Key Prior Works

- Hongyu et al. TBD: Benchmarking and Analyzing Deep Neural Network Training. arXiv preprint arXiv:1803.06905
- Adolf et al. Fathom: Reference workloads for modern deep learning methods. Workload Characterization (IISWC), 2016
- Abadi et al. TensorFlow: A System for Large-Scale Machine Learning. OSDI, 2016
- Chen et al. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv, 2015.
- Yu et al. An introduction to computational networks and the computational network toolkit. Microsoft Technical Report, 2014.

This research has been funded in part or completely by the Computing Hardware for Emerging Intelligent Sensory Applications (COHESA) project. COHESA is financed under the National Sciences and Engineering Research Council of Canada (NSERC) Strategic Networks grant number NETGP485577-15.