Power Efficiency Analysis of a Deep Learning Workload on an IBM "Minsky" Platform

M. D. Mazuecos Pérez[1] N. G. Seiler[1] C. S. Bederián[1,2]
N. Wolovick[1] A. J. Vega[3]

FaMAF, Universidad Nacional de Córdoba

CONICET

IBM T. J. Watson Research Center

September 27, 2018 – CARLA 2018

The way to CARLA

Platform

Measurements

Conclusions

Backup

The grant


FAMAF
UNC
IBM
2 DE SEPTIEMBRE 16
GRUPO DE INVESTIGACIÓN GPGPU COMPUTING
(FaMAF) DE LA UNIVERSIDAD NACIONAL DE CÓRDOBA
"FACULTY AWARD" POR USD 40.000
Universidad Nacional de Córdoba
www.unc.edu.ar

The trip


POWER8
NVIDIA TESLA
Qumulo
IBM Power Systems S822LC for HPC
Only Architecture with NVIDIA NVLink from CPU:GPU
IBM

The previous

Proceedings of the 51st Hawaii International Conference on System Sciences | 2018

Performance Characterization of State-Of-The-Art Deep Learning Workloads on an IBM "Minsky" Platform

Mauricio Guignard
FAMAF, Universidad Nacional de Córdoba
mguignard311@famaf.unc.edu.ar

Marcelo Schild
FAMAF, Universidad Nacional de Córdoba
marceloschild90@gmail.com

Carlos S. Bederián
FAMAF, Universidad Nacional de Córdoba
CONICET
bc@famaf.unc.edu.ar

Nicolás Wolovick
FAMAF, Universidad Nacional de Córdoba
nicolasw@famaf.unc.edu.ar

Augusto J. Vega
IBM T. J. Watson Research Center
ajvega@us.ibm.com

Abstract

Deep learning algorithms are known to demand significant computing horsepower, in particular when it comes to training these models. The capability of developing new algorithms and improving the existing ones is in part determined by the speed at which these models can be trained and tested. One alternative to attain significant performance gains is through hardware acceleration. However, deep learning has evolved into a large variety of models, including but not limited to fully-connected, convolutional, recurrent and memory networks. Therefore, it appears difficult that a single solution can provide effective acceleration for this entire deep learning ecosystem.

This work presents detailed characterization results of a set of archetypal state-of-the-art deep learning workloads on a last-generation IBM POWER8 system with NVIDIA Tesla P100 GPUs and NVLink interconnects. The goal is to identify the performance bottlenecks (i.e. the accelerable portions) to provide a thorough study that can guide the design of prospective acceleration platforms in a more effective manner. In addition, we analyze the role of the GPU (as one particular type of acceleration engine) and its effectiveness as a function of the size of the problem.

This research was developed, in part, with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

1. Introduction

The current success of deep learning techniques for machine learning is directly related to three complementary trends: the progress in new algorithms, the availability of big amounts of labeled data and the increasing computational power. Improving one of these areas usually demands improvements in the others. In particular, it has been noticed that research productivity is inversely proportional to the turnaround time of a deep learning experiment. While a few days is considered as tolerable, weeks are considered as "progress stalls" and experiments that take about a month are simply not worth running [1].

The huge computational demand from existing deep learning methods is driving a variety of new hardware solutions that emerge as deep learning application platforms. In recent months, platforms like Google's tensor processing unit (TPU) [2], NVIDIA's DGX-1 [3] and IBM's "Minsky" [4] have been announced or released, to mention just a few relevant examples. Hardware design has started to be shaped according to the needs of deep learning models with performance improvements that range from 10 to 100 times over conventional computing systems. As a result, previously intractable research problems turned into overnight jobs, opening up new types of learning algorithms and research opportunities.

However, significantly higher levels of performance and power efficiency are necessary in computationally constrained environments, like mobile applications and the Internet of Things (IoT) — unmanned aerial vehicles (*drones*), driverless cars, and "wearable" devices,

URI: http://hdl.handle.net/10125/50591
ISBN: 978-0-9981331-1-9

HICSS

The way to CARLA

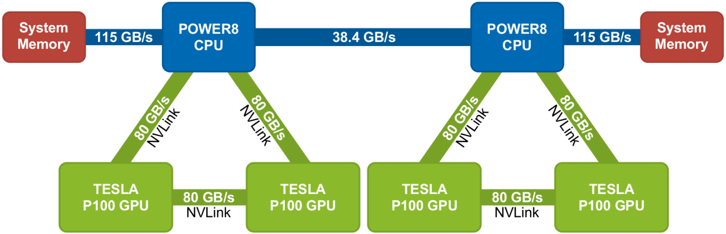
Platform

Measurements

Conclusions

Backup

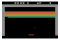
IBM S822LC, “Minsky”



- CPU 2×10 cores, SMT$\{1,2,4,8\}$
- GPU 4×5.3 TFLOPS

ML Workload

Fathom
Reference Workloads for
Modern Deep Learning

- **AlexNet** DNN, IMAGENET.
- **Autoenc** variational autoencoder, feature extraction.
- **DeepQ** deep reinforcement learning, play Stella games .
- **Memnet** end-to-end memory network, Q&A.
- **Residual** residual networks, IMAGENET.
- **Seq2Seq** recurrent neural network, language translation.
- **Speech** recurrent neural network, Baidu Research speech recognition.
- **VGG** 19 layers convolutional network, IMAGENET.

The way to CARLA

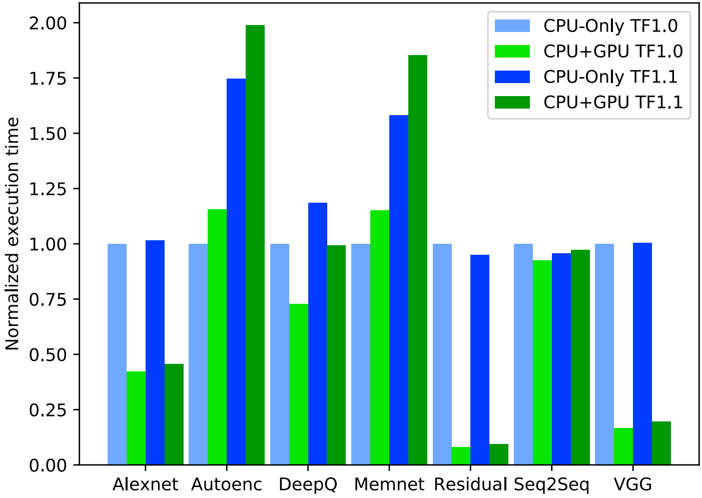
Platform

Measurements

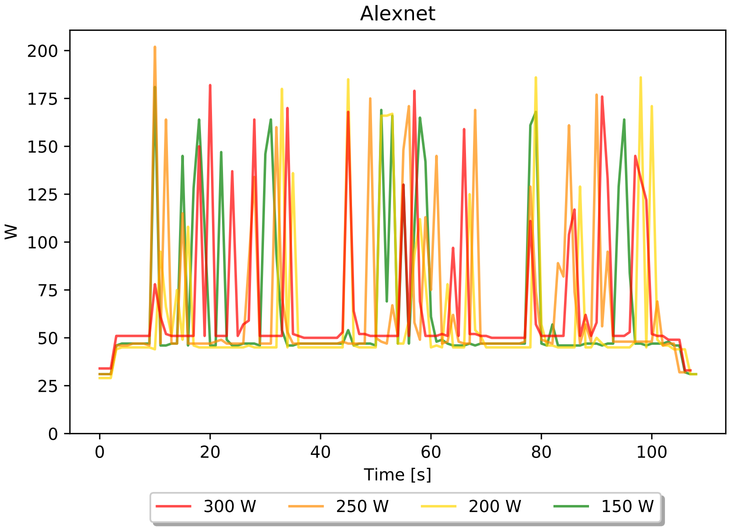
Conclusions

Backup

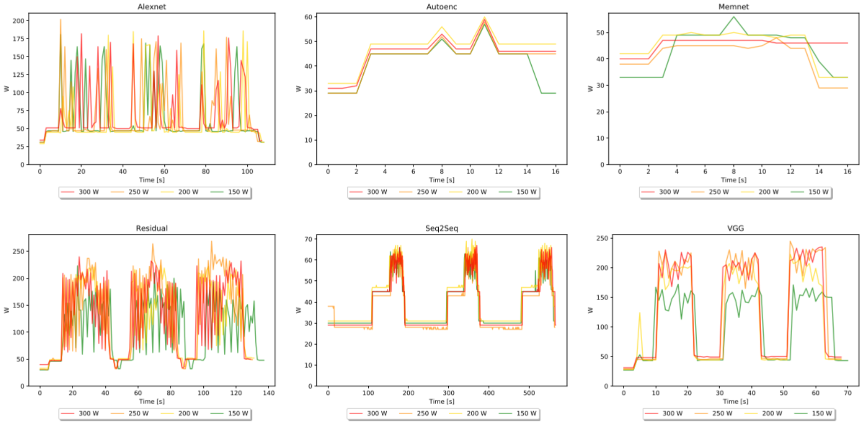
Performance assessment



Power capping, AlexNet power trace



Power capping, power trace



Time, power, energy

Energy-to-solution

$$ETS = \int_0^T P(t)dt$$

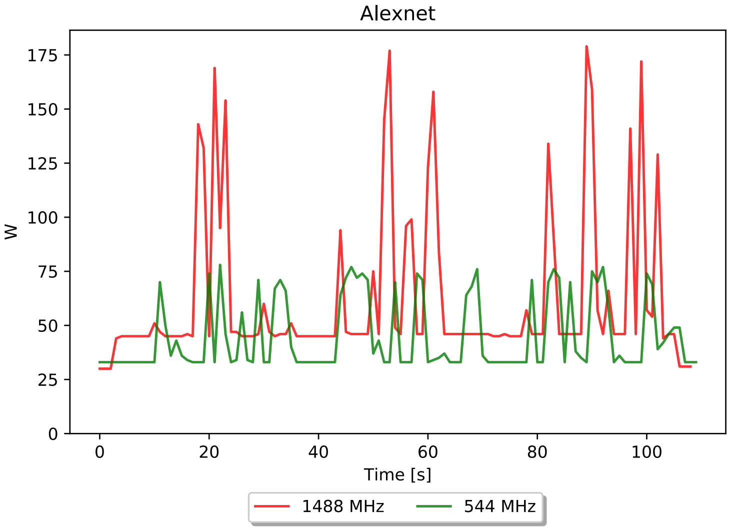
Energy-delay product

$$EDP = ETS \times T$$

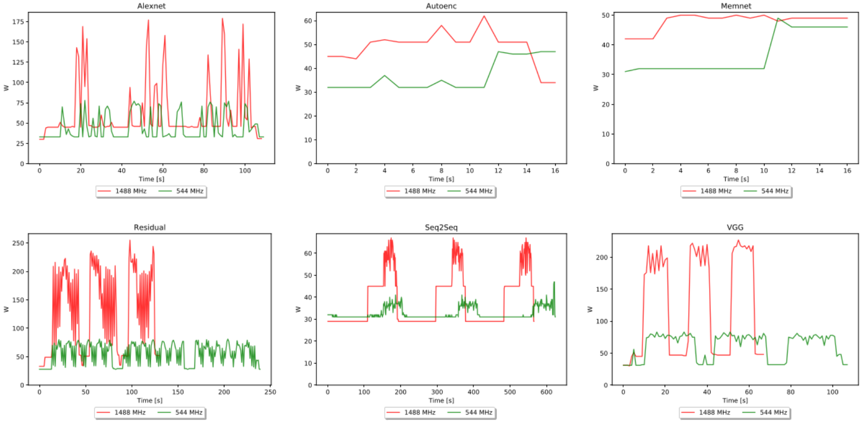
where instantaneous power is

$$P \propto V^2 f$$

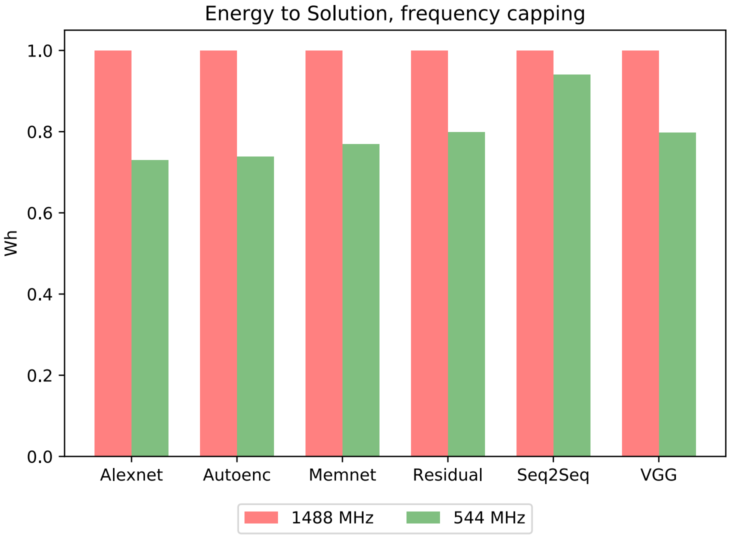
Frequency capping, AlexNet



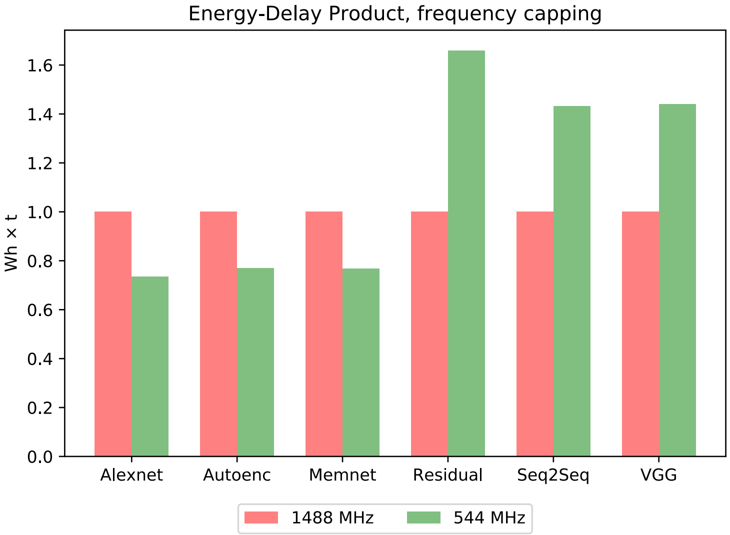
Frequency capping, power trace



Total energy decreased



Half of workloads do not penalize time



The way to CARLA

Platform

Measurements

Conclusions

Backup

Conclusions

- Full-throttle is not the answer.
- Well known for crypto-currency miners, passwords crackers.
- ML workload is not `hashcat`.
- Improve `--power-limit` NVIDIA driver algorithm?

Conclusions

- Full-throttle is not the answer.
- Well known for crypto-currency miners, passwords crackers.
- ML workload is not `hashcat`.
- Improve `--power-limit` NVIDIA driver algorithm?

Will you slow down your Teslas?

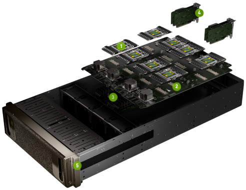
Conclusions

- Full-throttle is not the answer.
- Well known for crypto-currency miners, passwords crackers.
- ML workload is not `hashcat`.
- Improve `--power-limit` NVIDIA driver algorithm?

Will you slow down your Teslas?

Currently

MLPerf



Concentrate on RNN (memory-bound) and DNN (shaders-bound).

The way to CARLA

Platform

Measurements

Conclusions

Backup

Contributions to Fathom

```
 9451f3ed967e1d19ad451d120f9d807bce916cee
Merge pull request #35 from nahuelseiler/master, Porting seq2seq to
tensorflow versions later than 1.x

 f9811bfdcdc620f28575edfb1993bb3b1bd22d27
Merge pull request #27 from Zzzoom/tf-1.0.x, Upgrade to tensorflow 1.0.x
```