MANCHESTER 1824
The University of Manchester

APT

Fine-grained energy and performance profiling of ConvNets on ARM mobile platforms

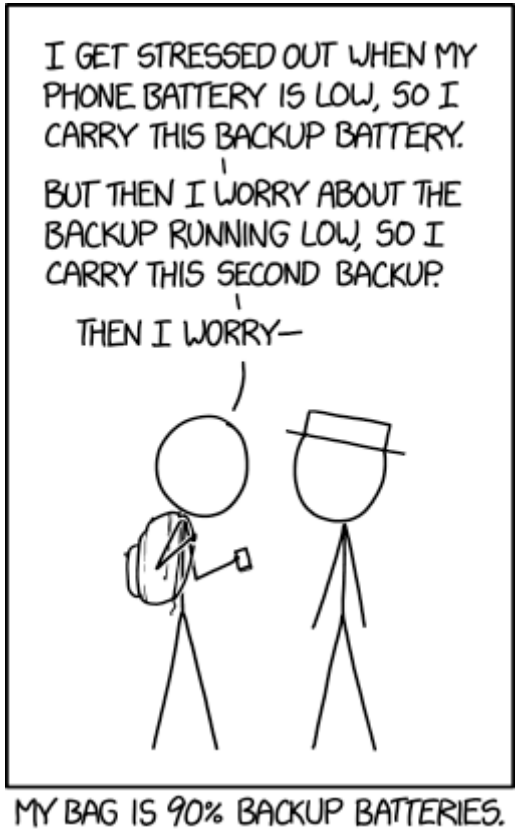
a.k.a
SyNERGY

1st International Workshop on Energy Efficient Data Mining and Knowledge Discovery ECML-PKDD

Crefeda Faviola Rodrigues – PhD student
Graham Riley & Mikel Luján
Advanced processors technology group – University of Manchester, UK

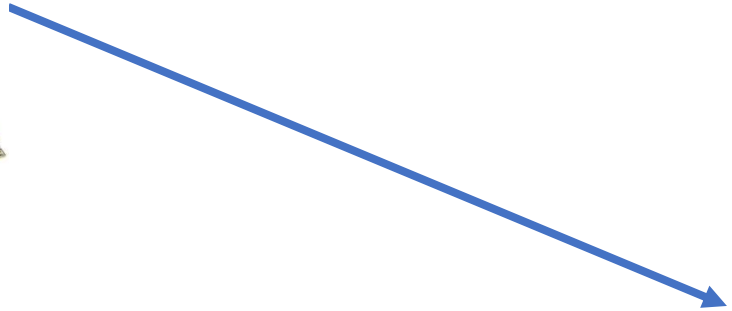
arm

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING


I GET STRESSED OUT WHEN MY PHONE BATTERY IS LOW, SO I CARRY THIS BACKUP BATTERY.
BUT THEN I WORRY ABOUT THE BACKUP RUNNING LOW, SO I CARRY THIS SECOND BACKUP.
THEN I WORRY—
MY BAG IS 90% BACKUP BATTERIES.

Scale of the problem

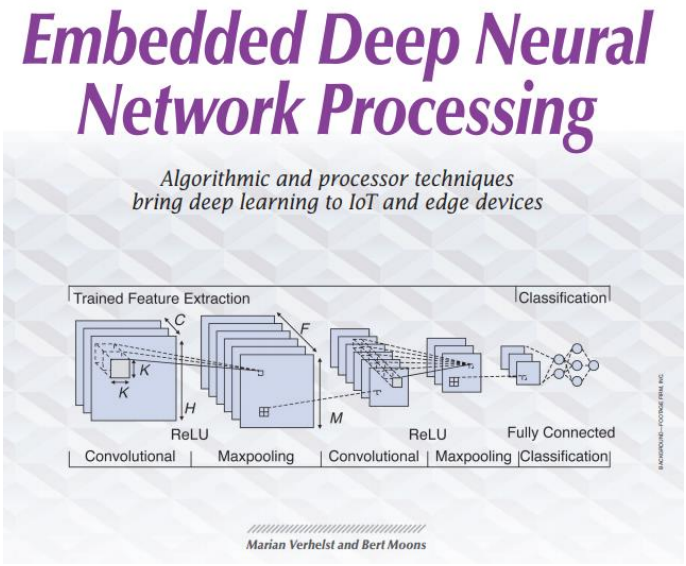
Embedded systems

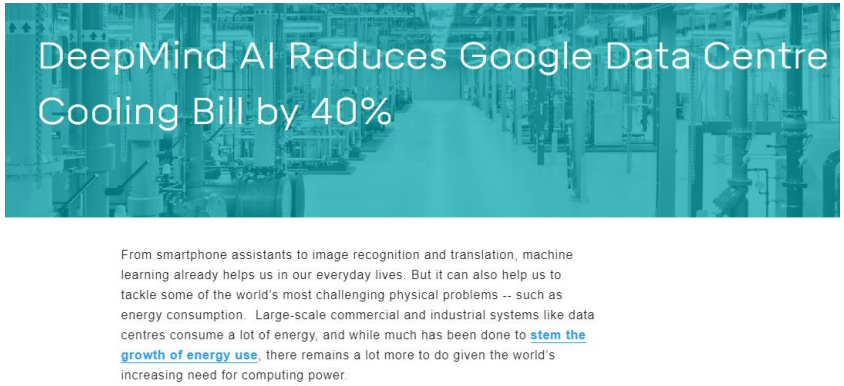


Datacenters

1) Improving energy-efficiency of ML?

Example: Energy optimizations





2) Use ML to improve/predict energy efficiency?

Example: Google's datacenters

Improving energy efficiency of ML?

Compression

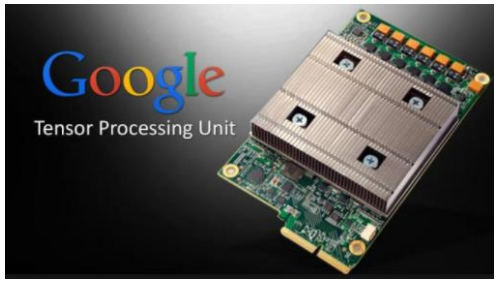
Compact models

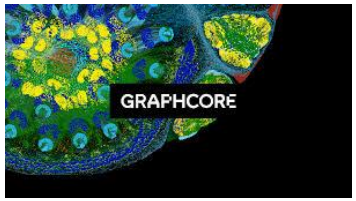
FFT, winograd, im2col

Quantization

Lower precision

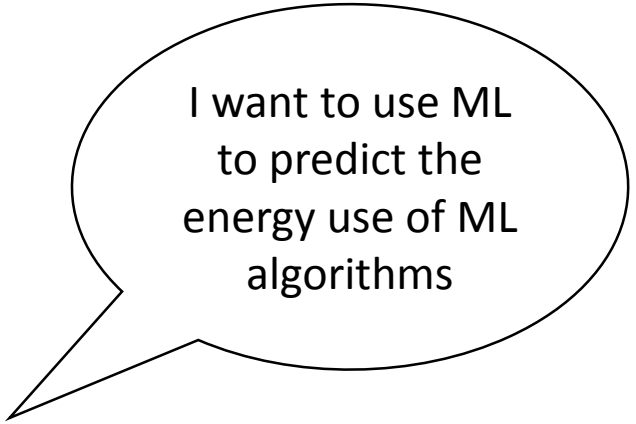
Vendor specific acceleration libraries





My area of research

1) Improving energy-efficiency of ML?



2) Use ML to improve/predict energy efficiency?

MANCHESTER 1824
The University of Manchester

APT

Energy measurement

MANCHESTER 1824
The University of Manchester

APT

Why don’t we measure energy?

Different measurement ways

Voltmeter or ammeter

- ✓ **P = I * V**
- ✓ **E = P * t**

on-board

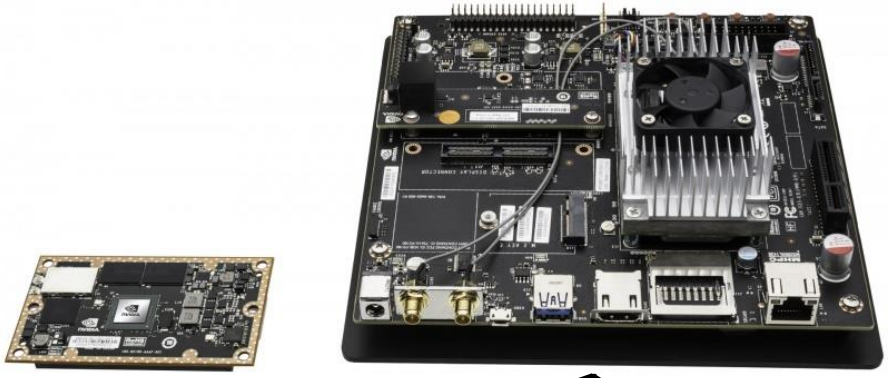
Power sensor chips

Power pins

* Measuring Power Consumption for DragonBoard™ 410c based on the Qualcomm® Snapdragon™ 410E processor

Example development boards

Jetson TX1 (Quad core ARM A57 + CUDA Maxwell GPU)

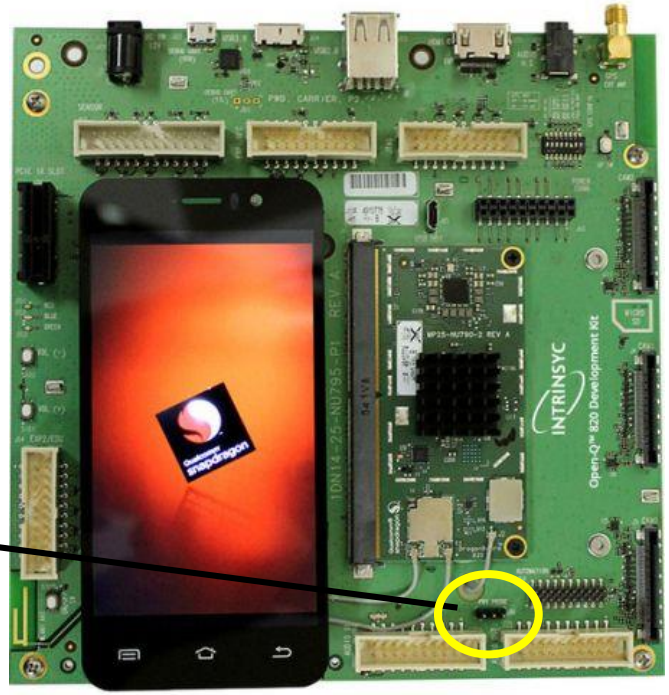


CPU

GPU

System (processor + memory …)

Snapdragon 820 (Quad core Kryo CPU + Adreno GPU)

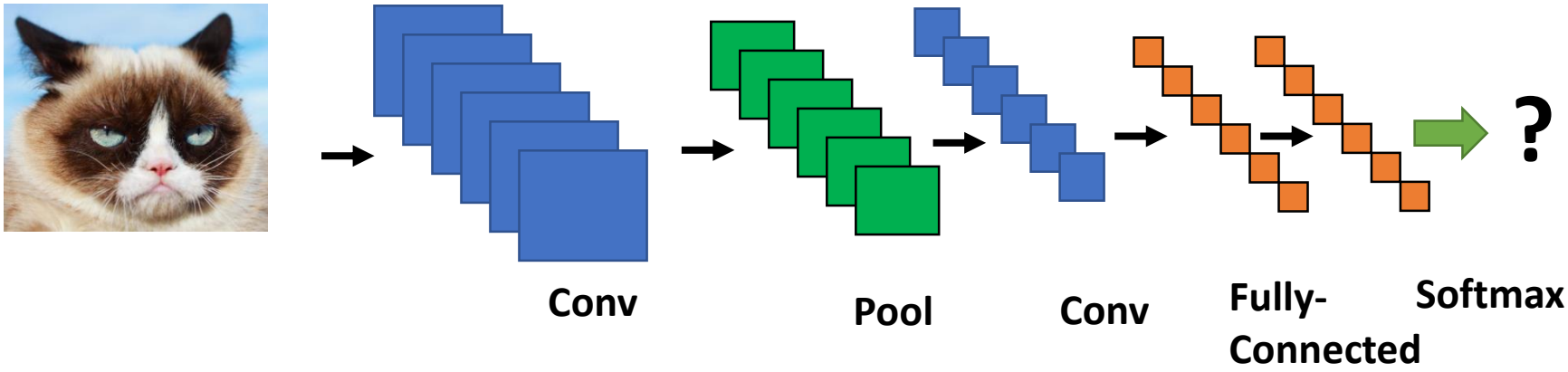


Can we find a consistent point to measure energy on different systems?

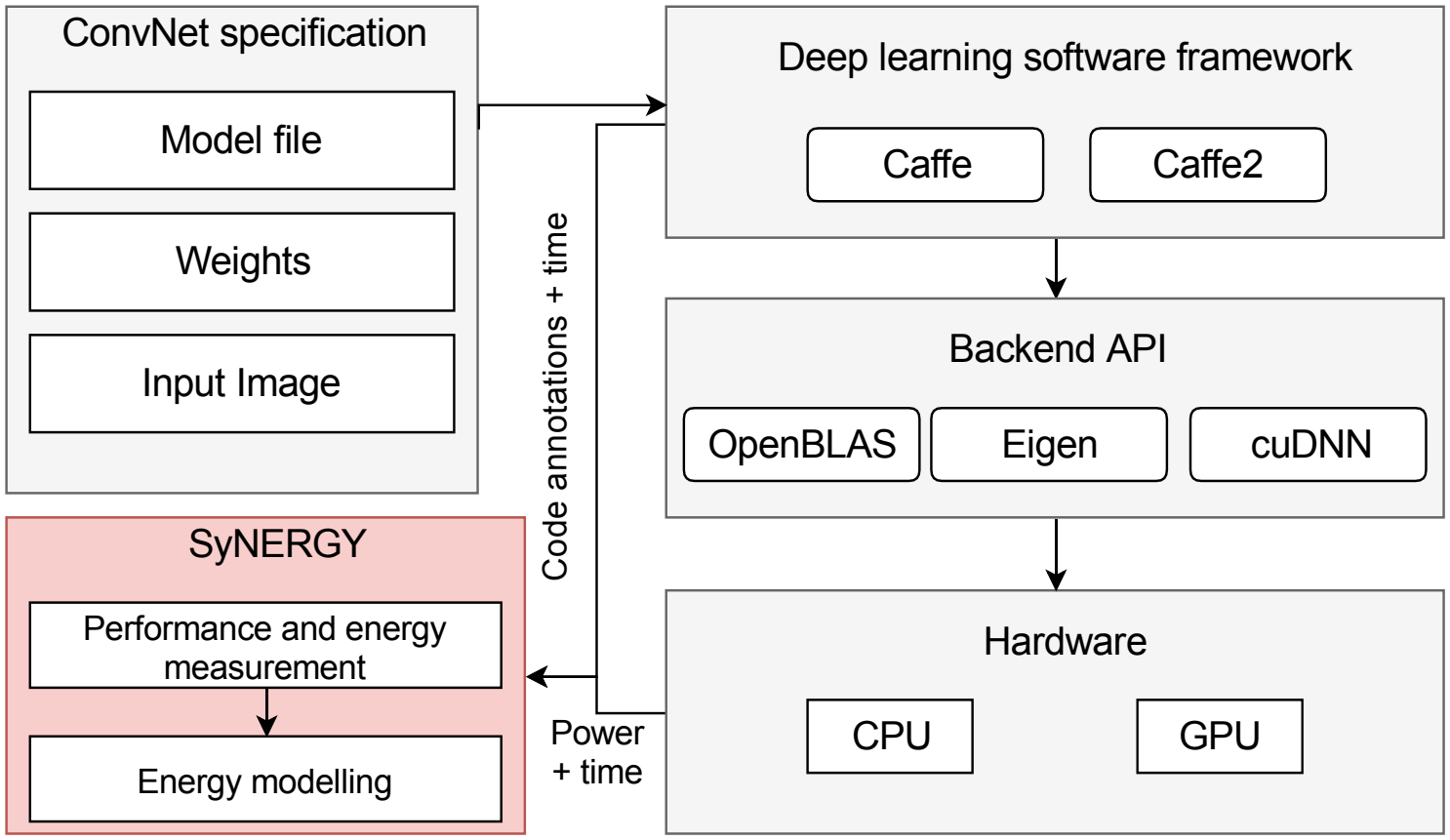
Motivation

- Power measurements are **difficult** to obtain: ammemeter, power sensors …
- Most of the work focuses on **execution time** on **desktop/ server** CPUs and GPUs
- **Lack** of support for **energy** measurements in current **deep learning frameworks**: Caffe2, Tensorflow and others
- Lack of evaluations with energy as a **metric**

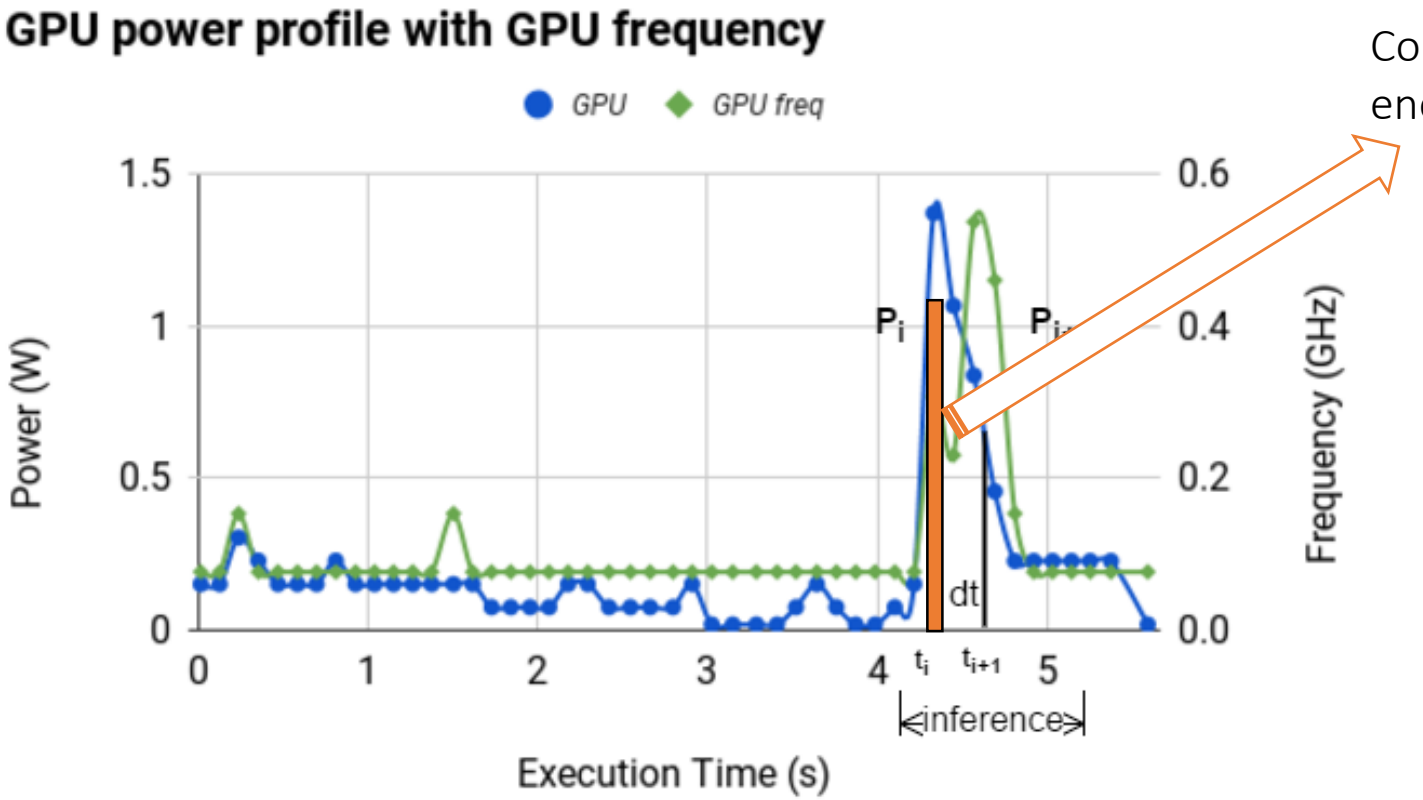
Convolutional Neural Network



SyNERGY

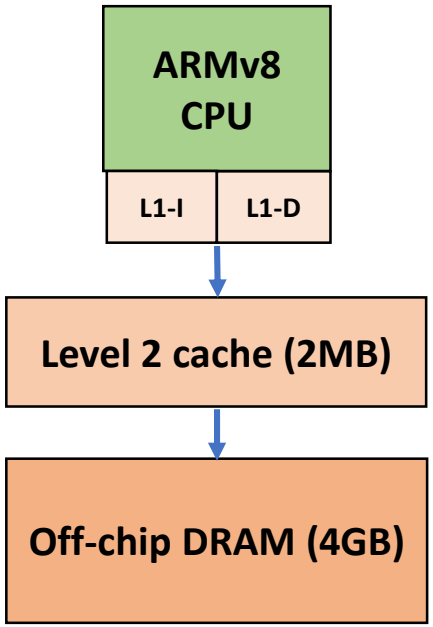

ConvNet specification
Model file
Weights
Input Image
Deep learning software framework
Caffe
Caffe2
Backend API
OpenBLAS
Eigen
cuDNN
Hardware
CPU
GPU
Code annotations + time
Power + time
SyNERGY
Performance and energy measurement
Energy modelling

Example power profile



Conv layer energy ?

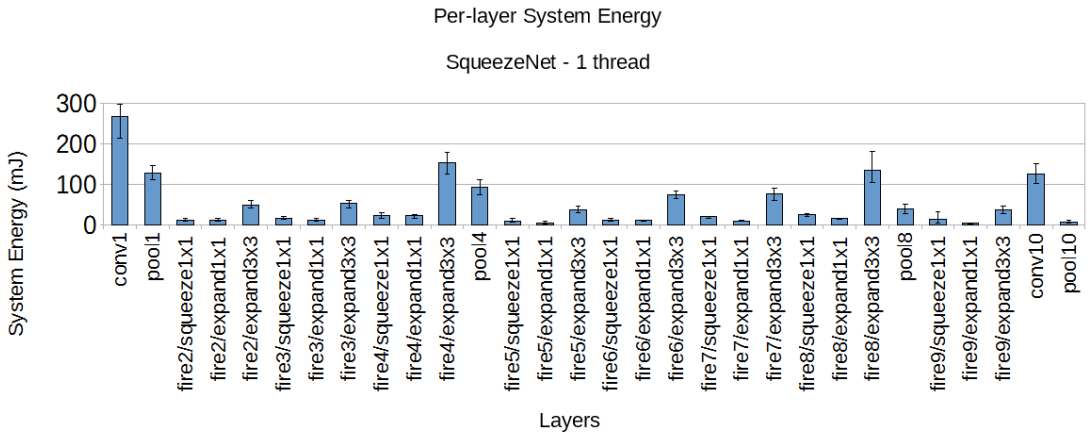
Power measurement environment



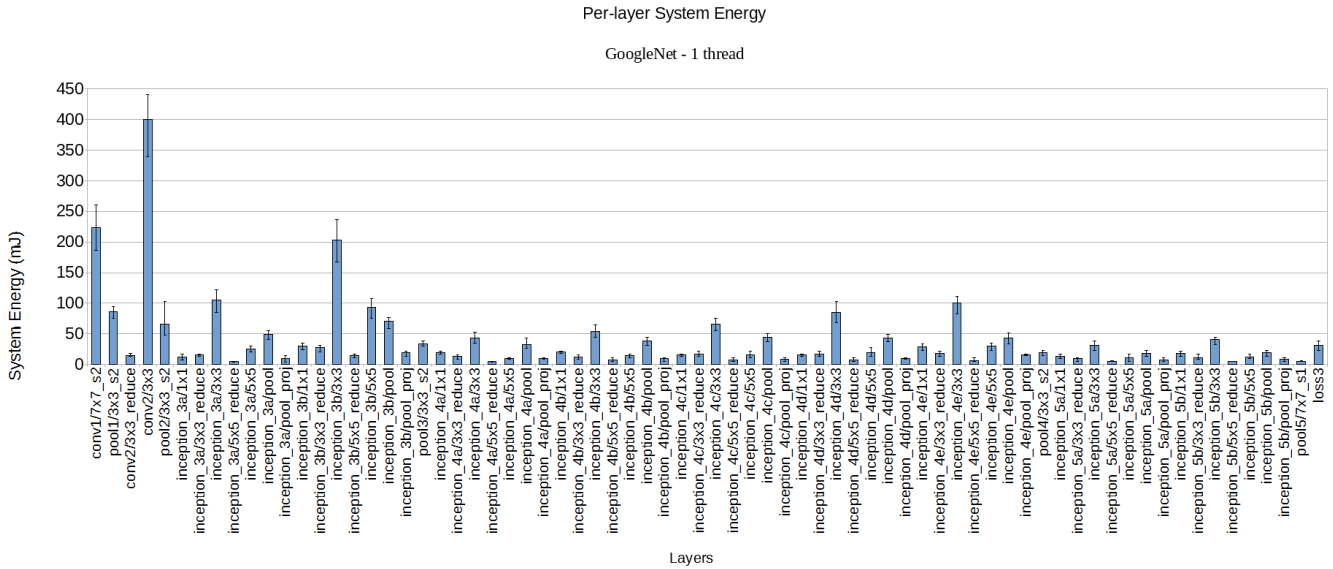
Jetson TX1 + OpenBLAS

- ✓ Single-threaded
- ✓ Single image inference
- ✓ No other applications running on the system
- ✓ Interactive governor for power management

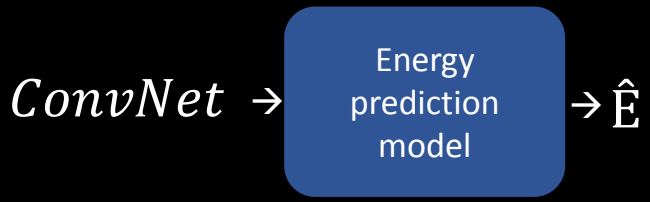
Per-layer system energy measurement



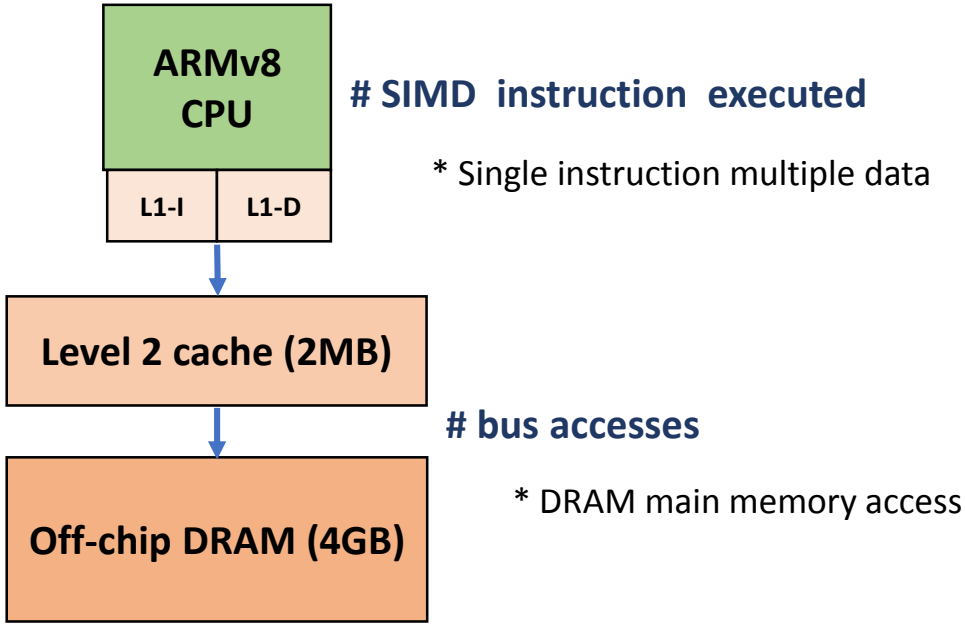
- ✓ 82% and 77% - Conv layers
- ✓ 17% and 21% - Pooling layers
- ✓ 1-2% - Other layers, For example, (fc layer in GoogleNet 1.1%)



Energy prediction

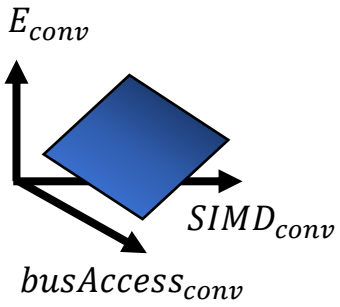


Performance counters

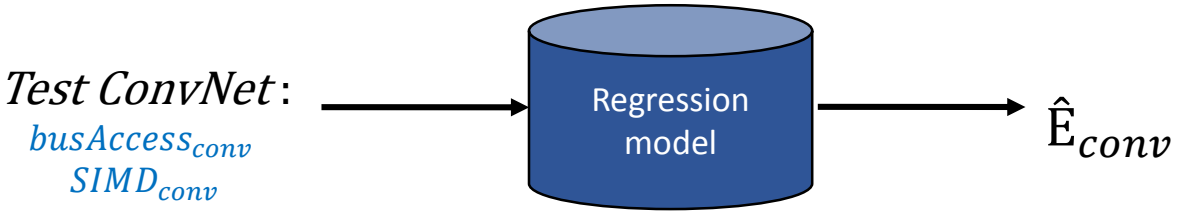


"Measure the number of SIMD instructions and bus access that take place in all the Conv Layers"

Regression-based prediction

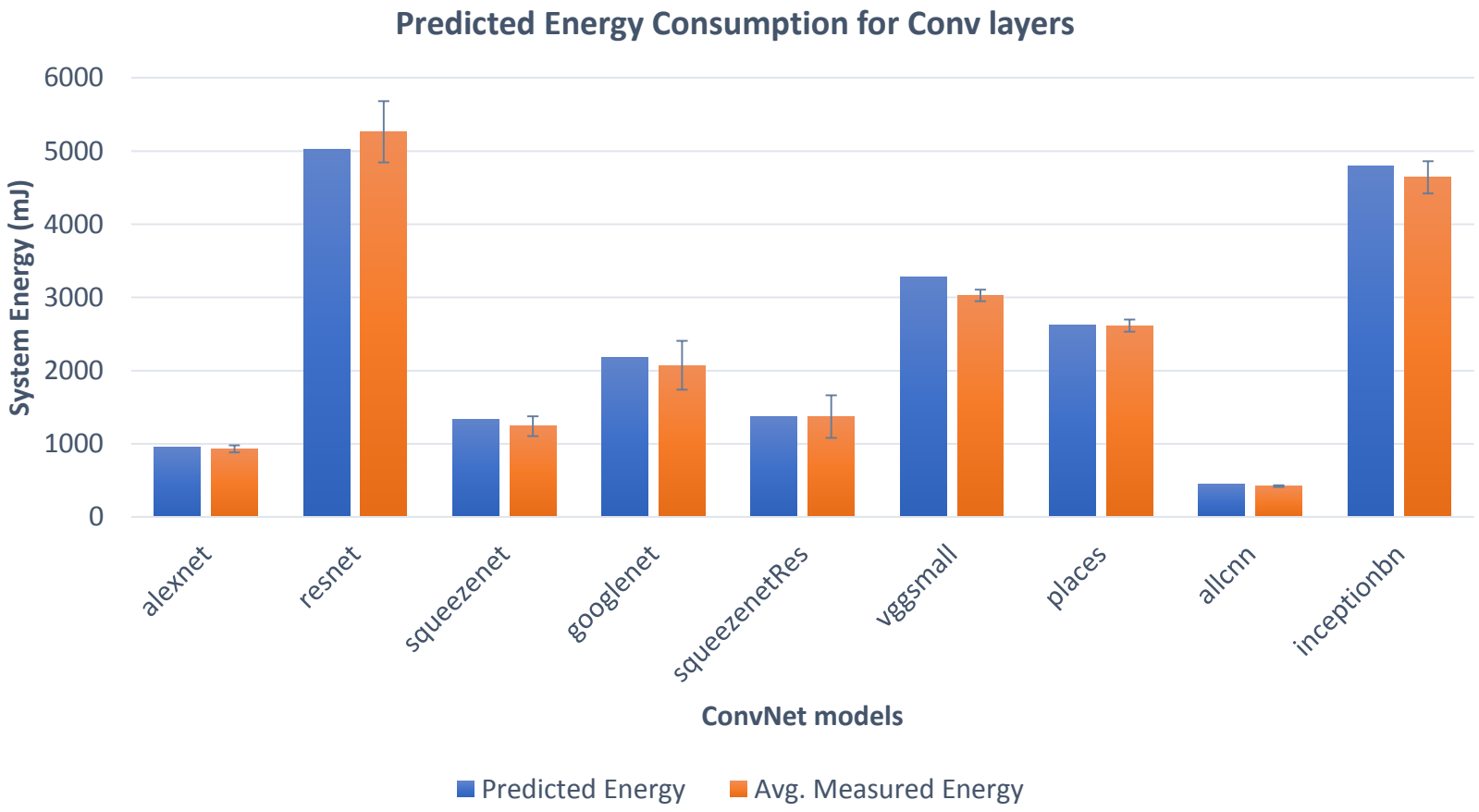


$$E_{conv} = x_1 \times busAccess_{conv} + x_2 \times SIMD_{conv}$$



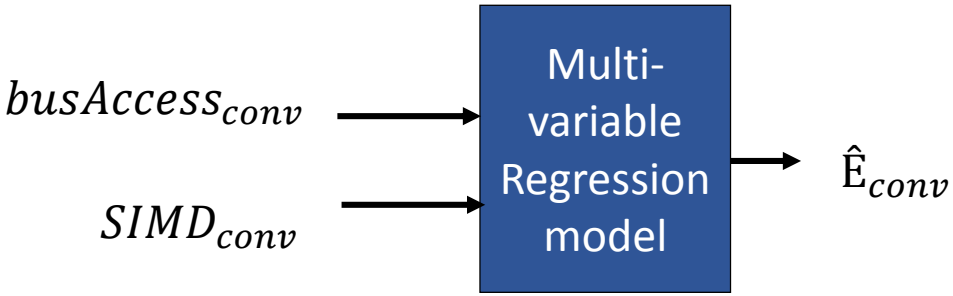
$$\hat{E}_{conv} = x_1 \times busAccess_{conv} + x_2 \times SIMD_{conv}$$

Energy prediction – performance counters

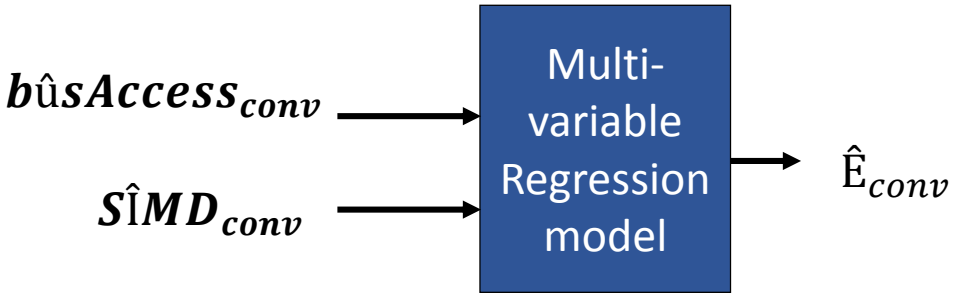


Avg. Relative Test Error = 5.72 ± 5.2 %

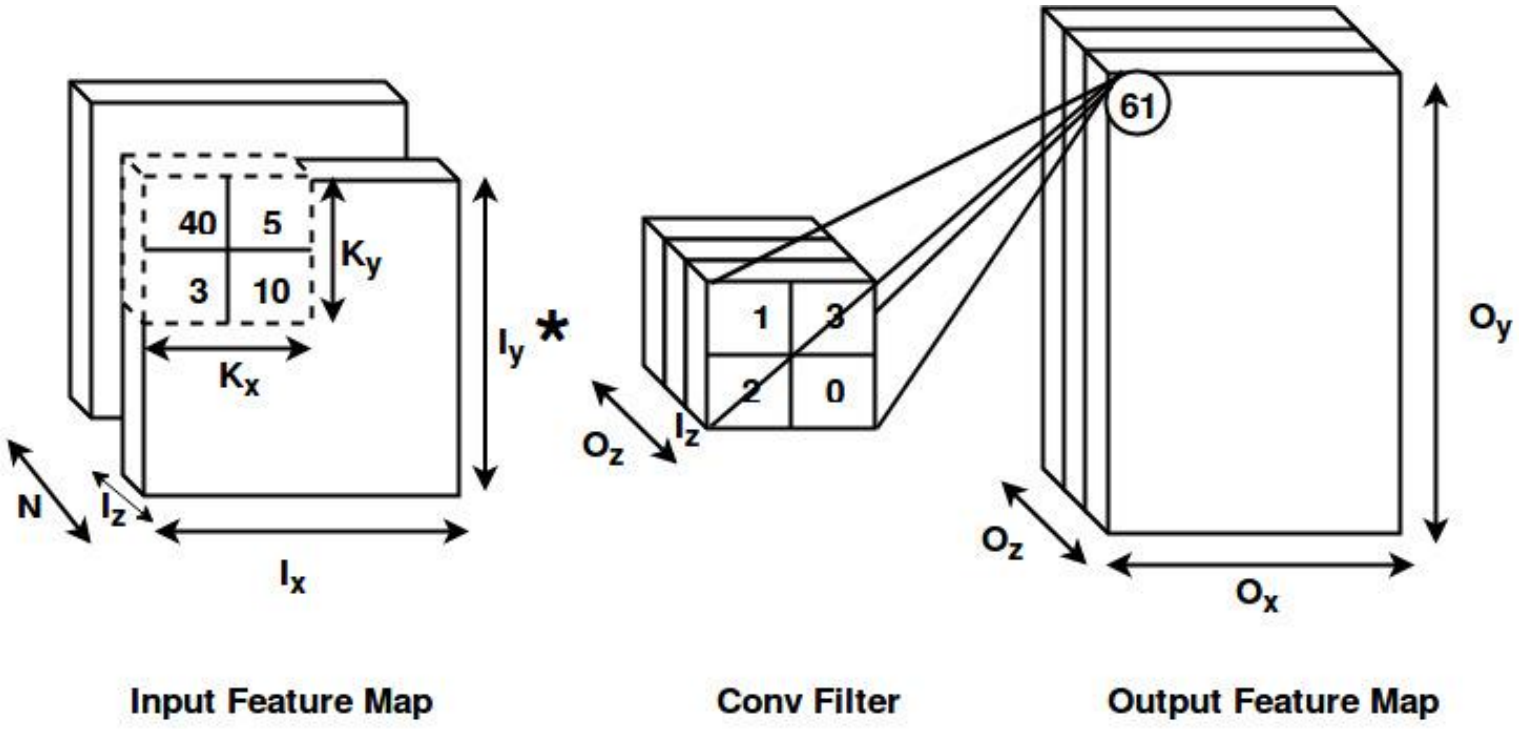
I don’t want to measure



Predict SIMD and bus accesses



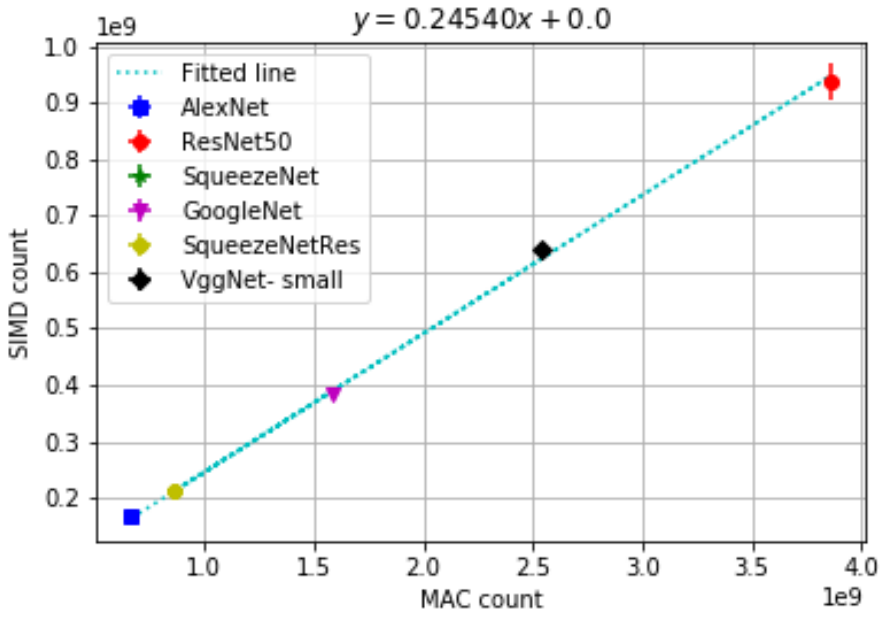
Multiply accumulate (MAC) count



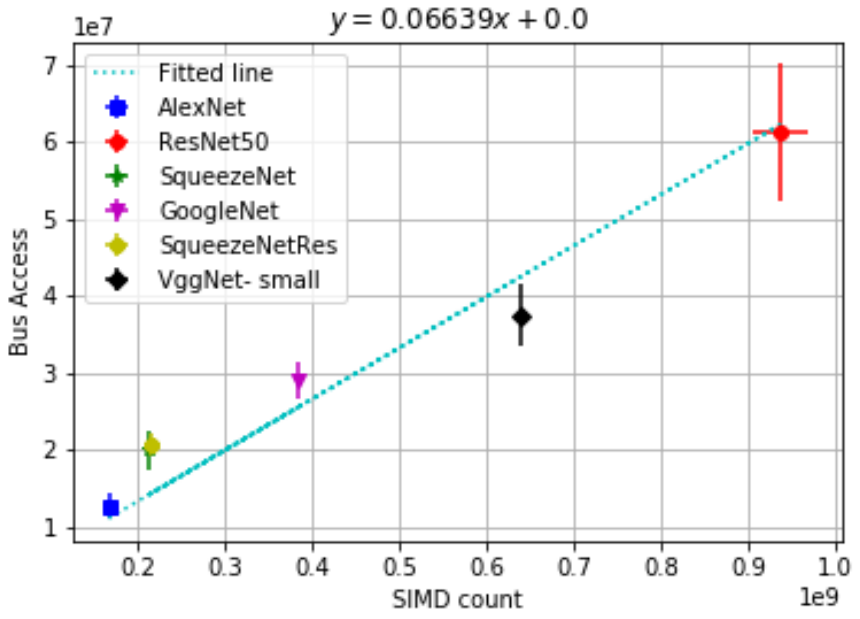
$$MAC_{conv} = O_x * O_y * O_z * K_x * K_y * I_z$$

MAC to SIMD and *SIMD to Bus accesses*

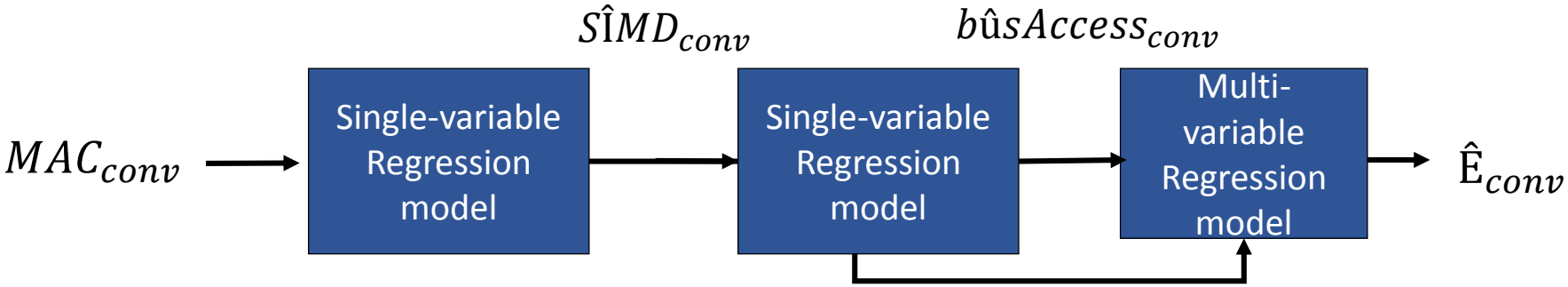
MAC to SIMD



SIMD to Bus accesses

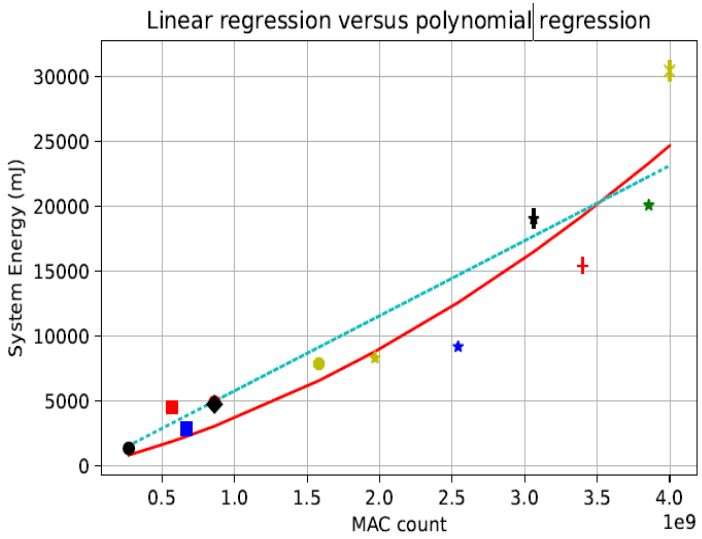


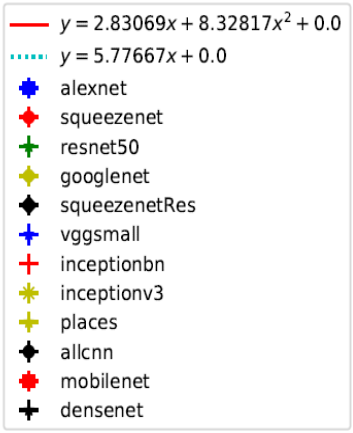
MAC to Energy relationship?



$$\hat{E}_{conv} = x_1 \times MAC_{conv} \, ?$$

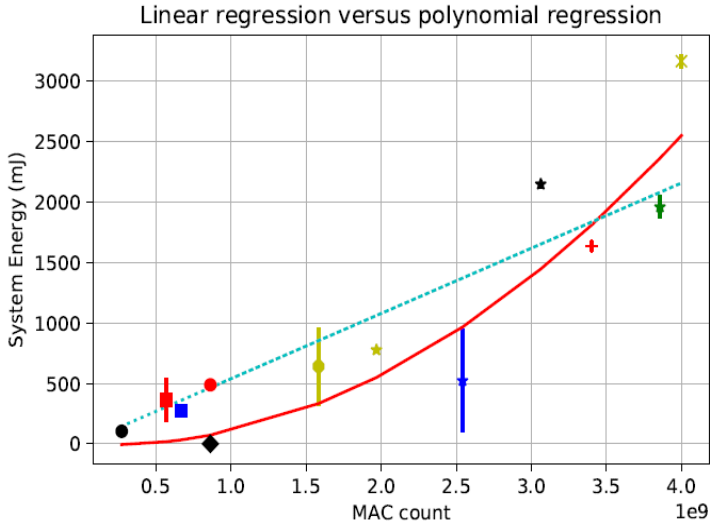
MAC to Energy

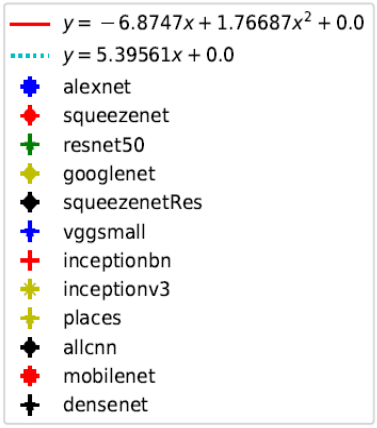




Jetson TX1- Eigen library

Snapdragon 820 - Eigen library





Possible future directions

- ✓ Evaluate the energy consumption of neural networks
- ✓ Build energy consumption models
  - Underlying hardware
  - Software implementation
  - Understand the neural network models
- ✓ Explore system's research to optimize for energy consumption
  - ✓ Power management techniques

Any questions?

crefeda.rodrigues@postgrad.manchester.ac.uk

The University of Manchester

Github repo: https://github.com/Crefeda/SyNERGY

References:

1. Rodrigues, Crefeda Faviola, Graham Riley, and Mikel Luján. "**SyNERGY: An energy measurement and prediction framework for convolutional neural networks on Jetson TX1**" Int'l Conf on Parallel and Distributed Processing Techniques and Applications -(PDPTA'18), 2018 CSREA Press,United States of America **(late September)**
2. Rodrigues, Crefeda Faviola, Graham Riley, and Mikel Luján. "Fine-grained energy and performance profiling for deep convolutional neural networks (early arXiv draft) [https://arxiv.org/pdf/1803.11151.pdf]
3. Rodrigues, Crefeda Faviola, Graham Riley, and Mikel Luján. "Fine-grained energy profiling for deep convolutional neural networks on the Jetson TX1." Workload Characterization (IISWC), 2017 IEEE International Symposium on. IEEE, 2017.

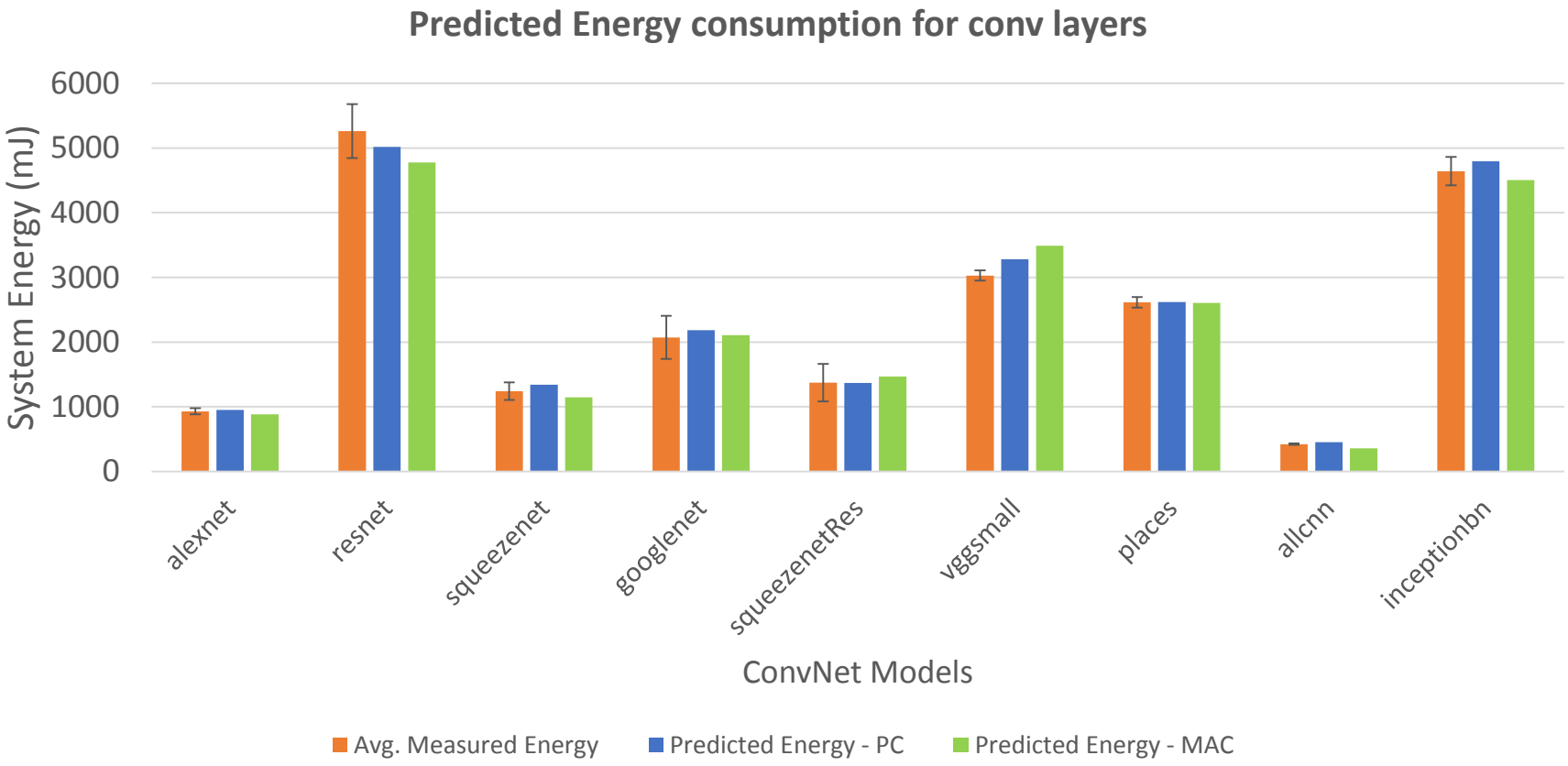
Part-funded by arm is-enes

MANCHESTER 1824
The University of Manchester

APT

Back up slides

Final energy prediction



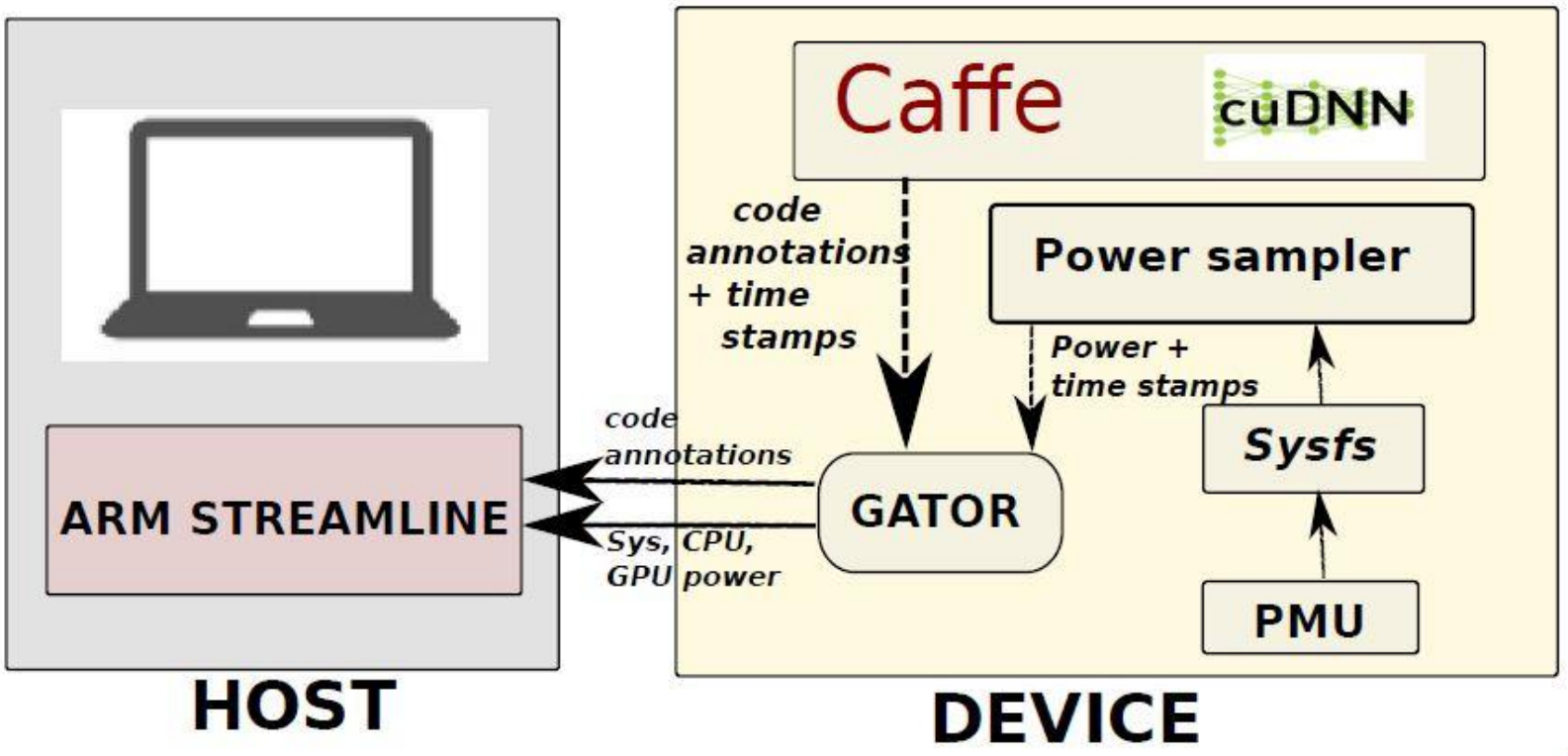
Avg. Relative Test Error = 7.08 ± 6.0 %

Previous result: Avg. Relative Test Error = 5.72 ± 5.2 %

Results

Conv layers	Linear regression (%)	Polynomial regression (%)
TX1 Eigen	74 +/- 6	77 +/- 7
Snapdragon 820	68 +/- 14	73 +/- 11

Energy measurement – Jetson TX1



**https://developer.arm.com/products/software-development-tools/ds-5-development-studio/streamline*