MANCHESTER 1824
The University of Manchester

APT

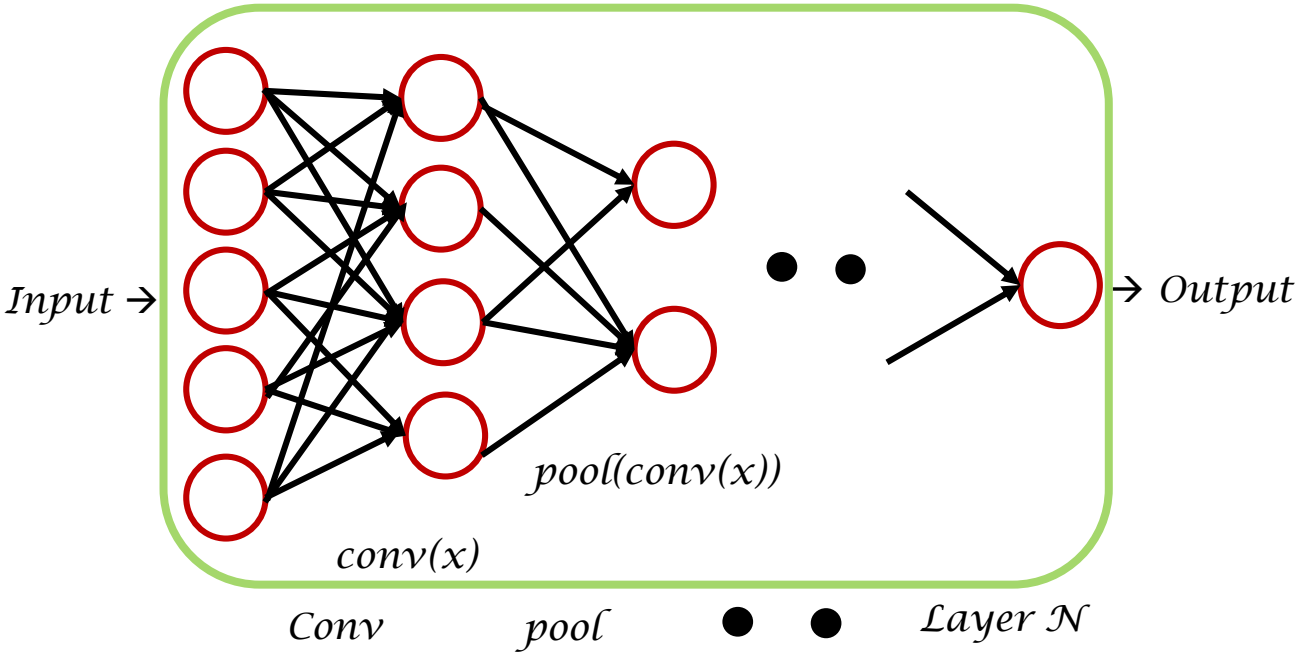
SyNERGY: An energy measurement and prediction framework for ConvNets on Jetson TX1

Int'l Conf on Parallel and Distributed Processing Techniques and Applications -(PDPTA -18)

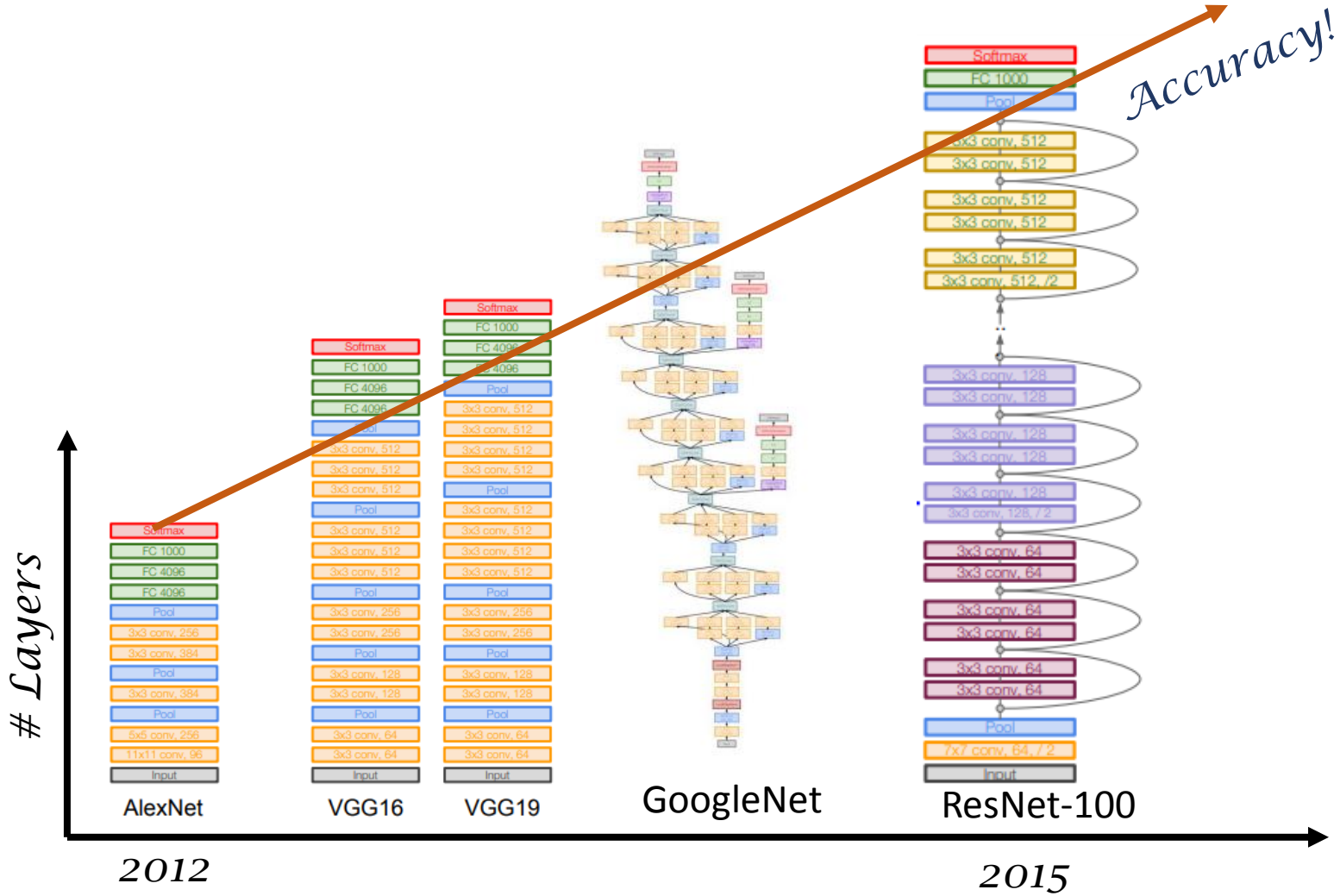
Crefeda Faviola Rodrigues – PhD student
Graham Riley & Mikel Luján
Advanced processors technology group – University of Manchester

Convolutional Neural Network

Convolutional Neural Network



State-of-the-art ConvNet models



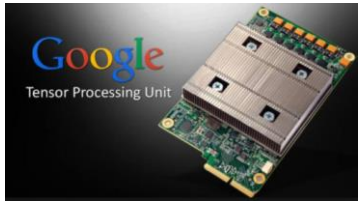
Current focus on energy optimizations

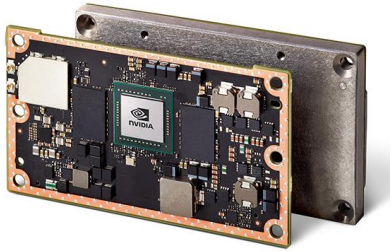
Under review as a conference paper at ICLR 2017

SQUEEZENET: ALEXNET-LEVEL ACCURACY WITH 50X FEWER PARAMETERS AND <0.5MB MODEL SIZE

MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko
Weijun Wang Tobias Weyand Marco Andreetto Hartwig Adam





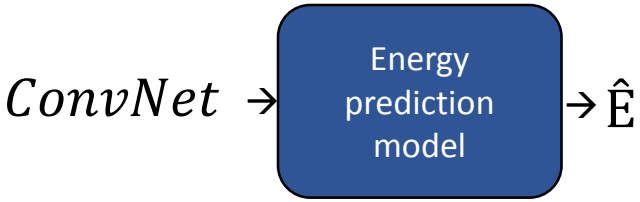
Published as a conference paper at ICLR 2016

DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING, TRAINED QUANTIZATION AND HUFFMAN CODING

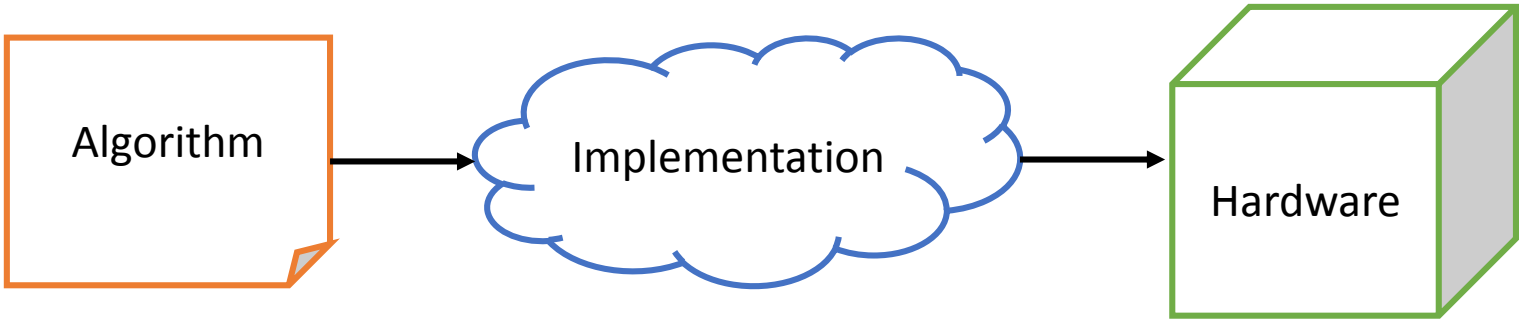
Motivation

- No support for energy measurements in current deep learning frameworks
- Power measurements are difficult to obtain: power meter, power sensors …
- Few studies evaluating models with energy as a metric

Energy prediction models



Factors to consider

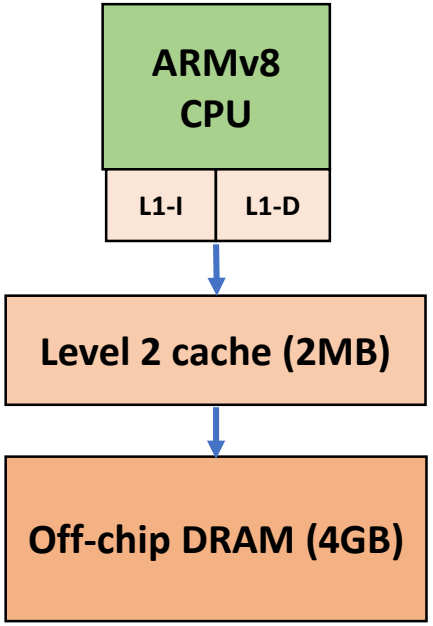


- Overall energy
- **Fine-grained energy**

- **Caffe/Caffe2– OpenBLAS**/ATLAS/EIGEN
- Tensorflow
- Torch

- **Jetson TX1**
- Snapdragon 820

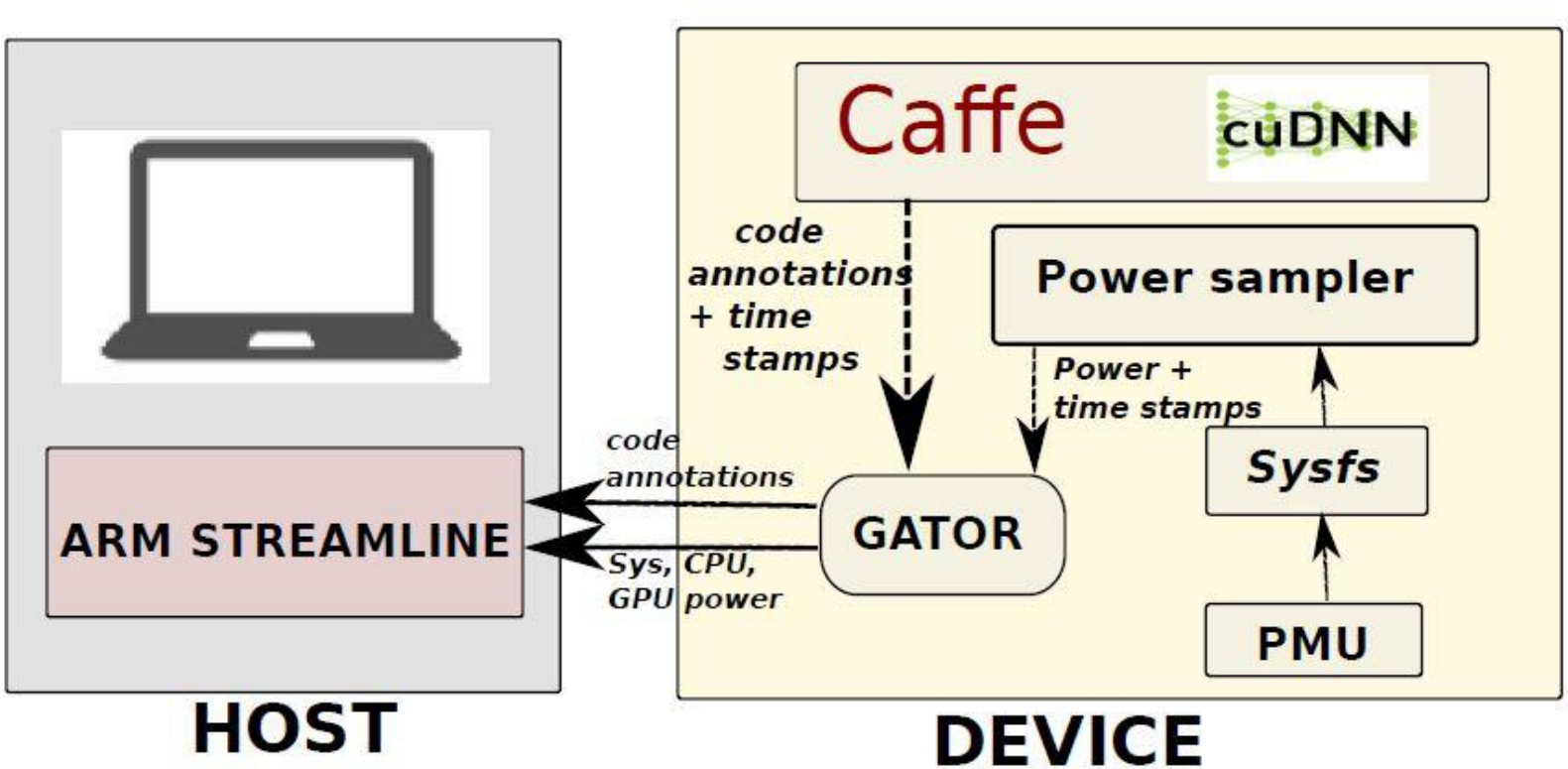
Jetson TX1 - context



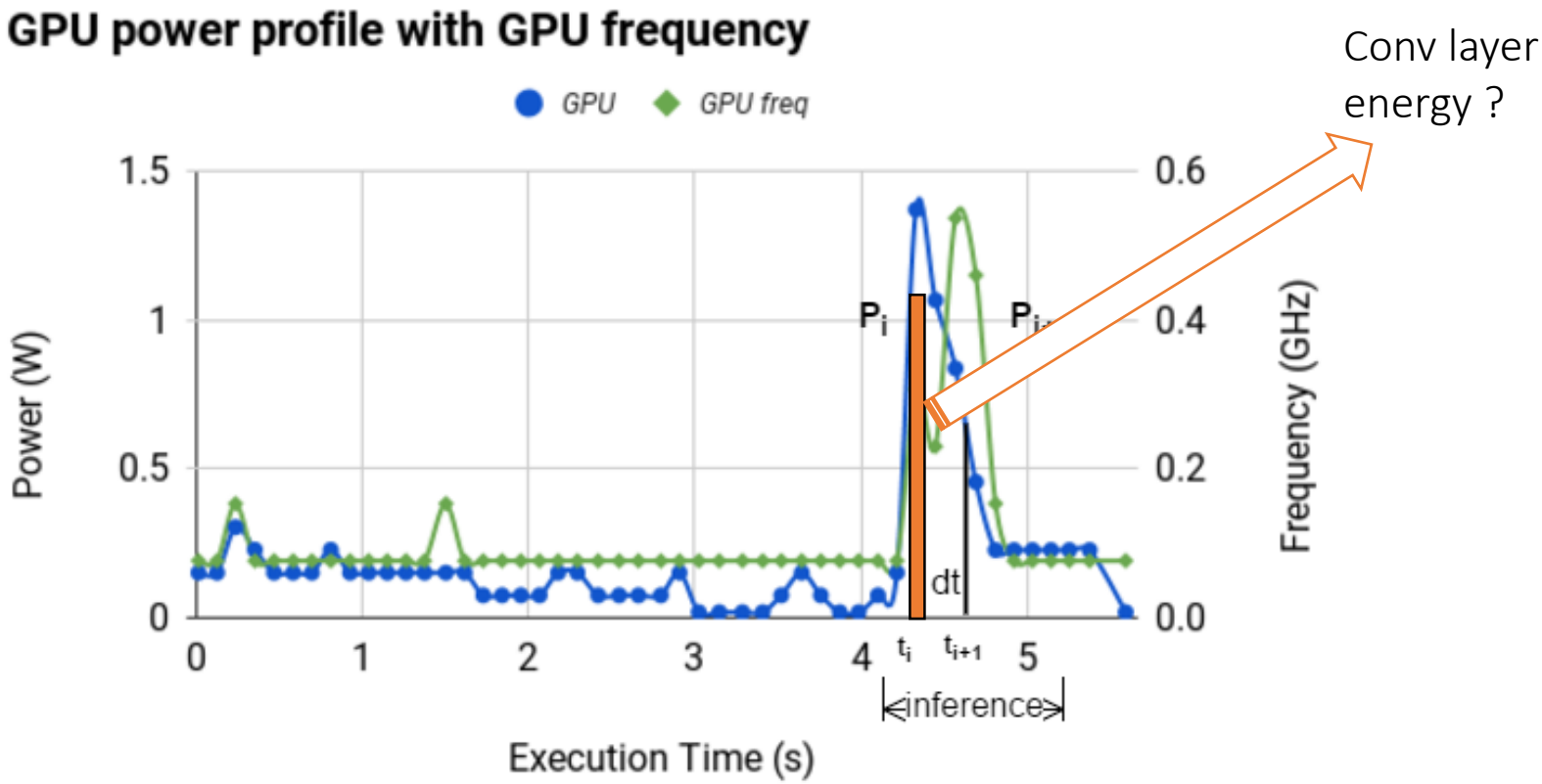
- ✓ Single-threaded
- ✓ Conv layers
- ✓ 1 image inference
- ✓ No other applications running on the system
- ✓ Interactive governor for power management

Energy Measurements

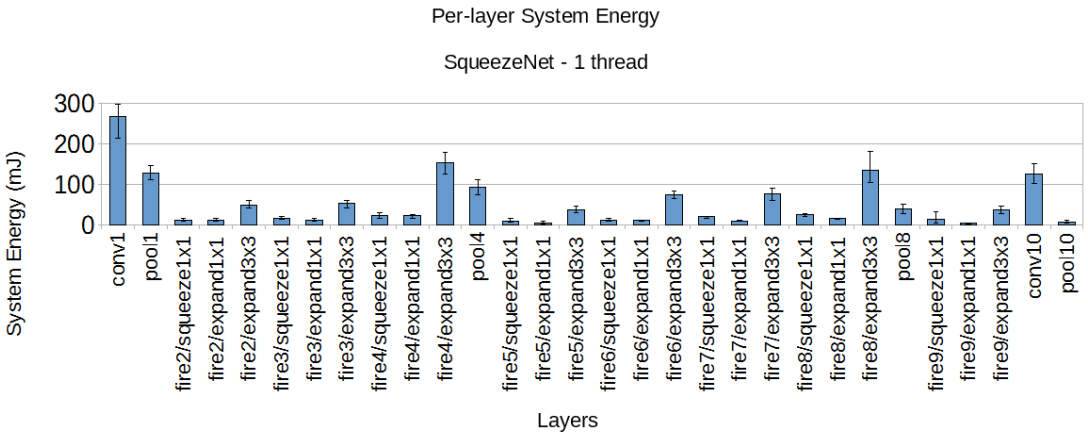
Energy measurement framework



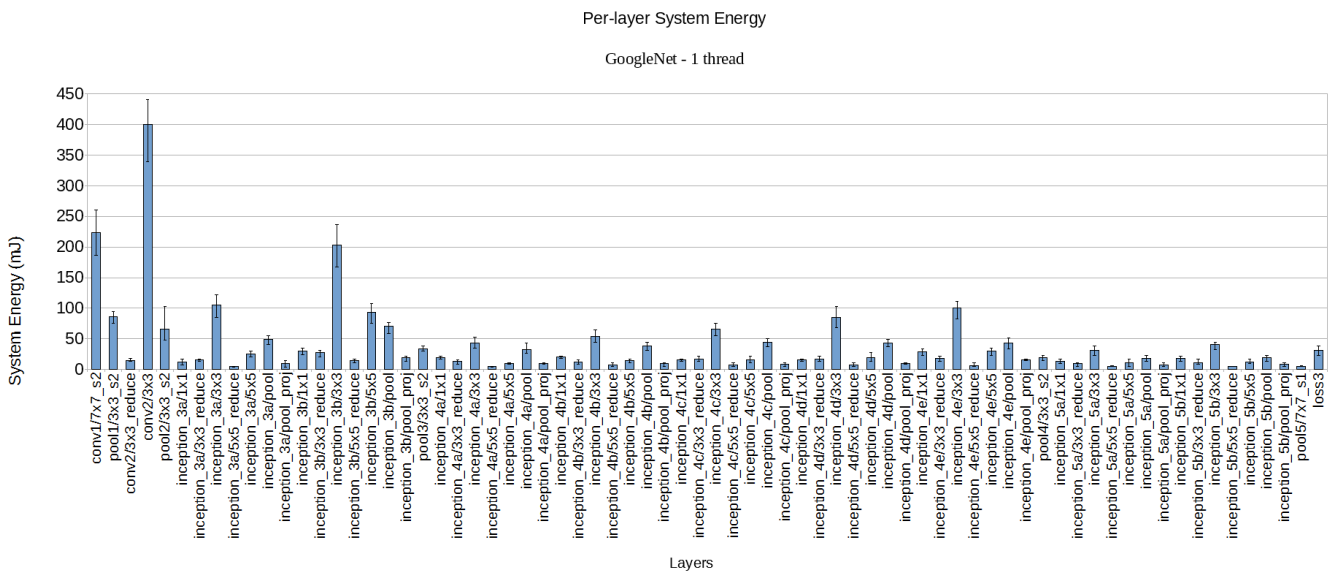
Example power profile



Per-layer energy measurement

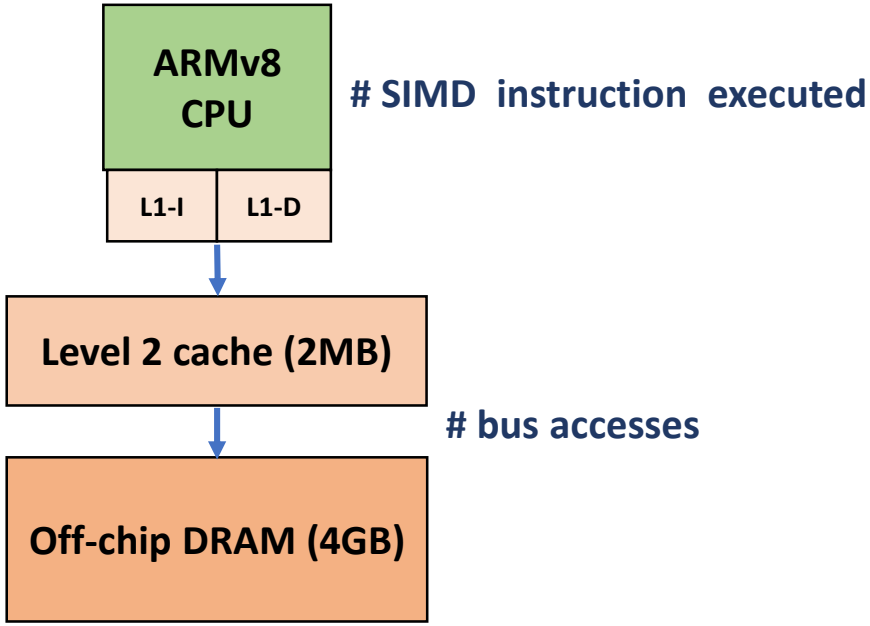


- ✓ 82% to 77% - Conv layers
- ✓ 17% to 21% - Pooling layers
- ✓ 1-2% - Other layers, For example, (fc layer in GoogleNet 1.1%)

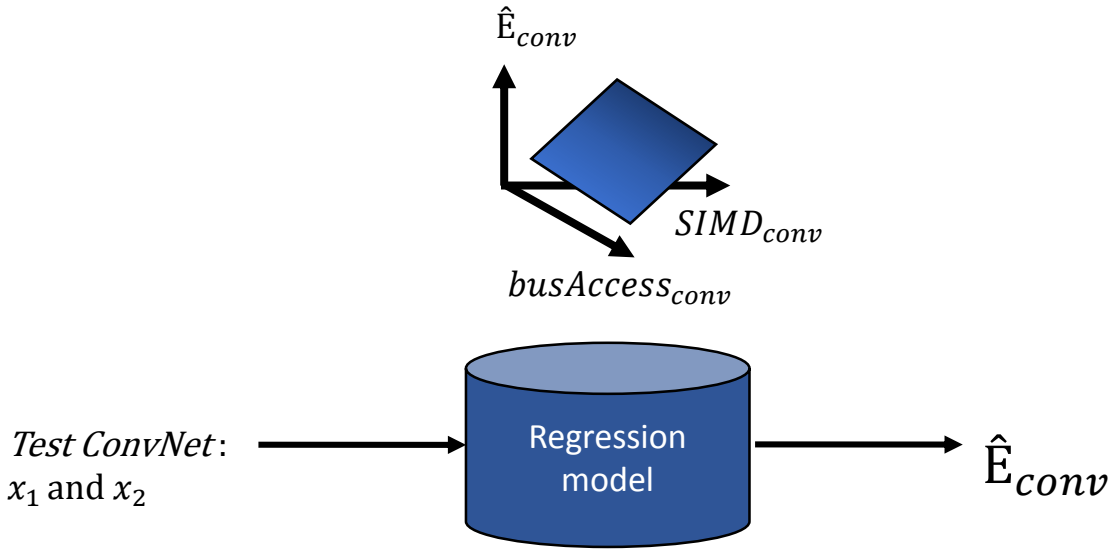


Energy Predictions

Performance counters

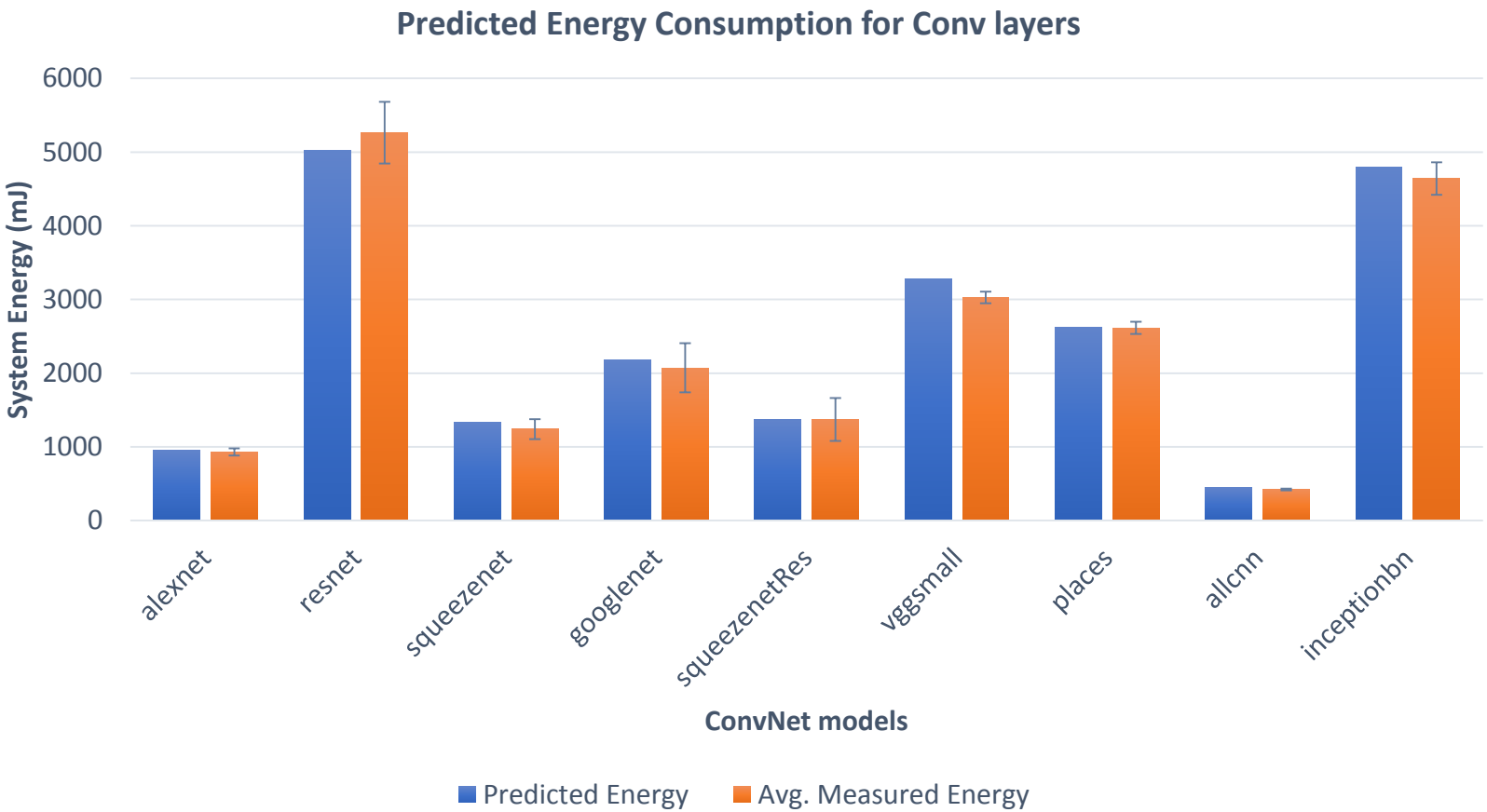


Regression-based prediction



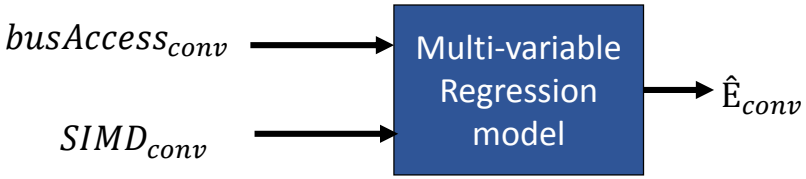
$$\hat{E}_{conv} = x_1 \times busAccess_{conv} + x_2 \times SIMD_{conv}$$

Energy prediction – performance counters

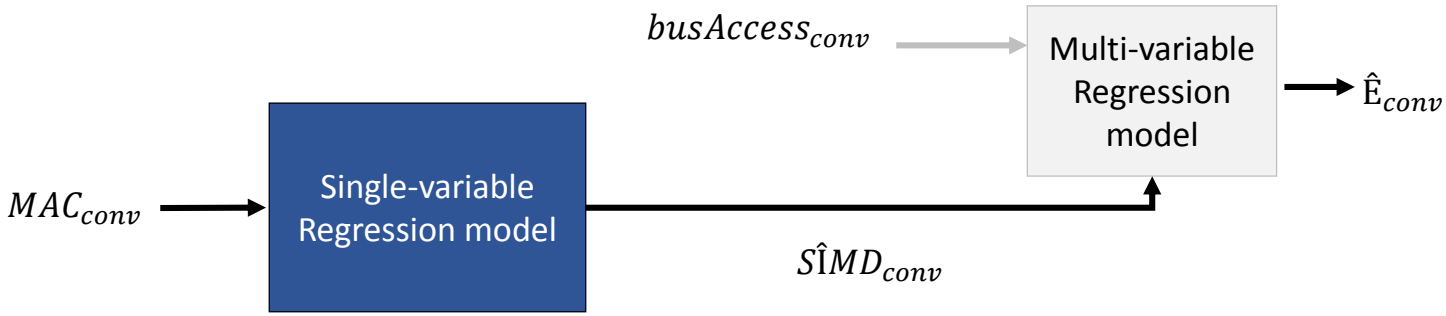


Avg. Relative Test Error = 5.72 ± 5.2 %

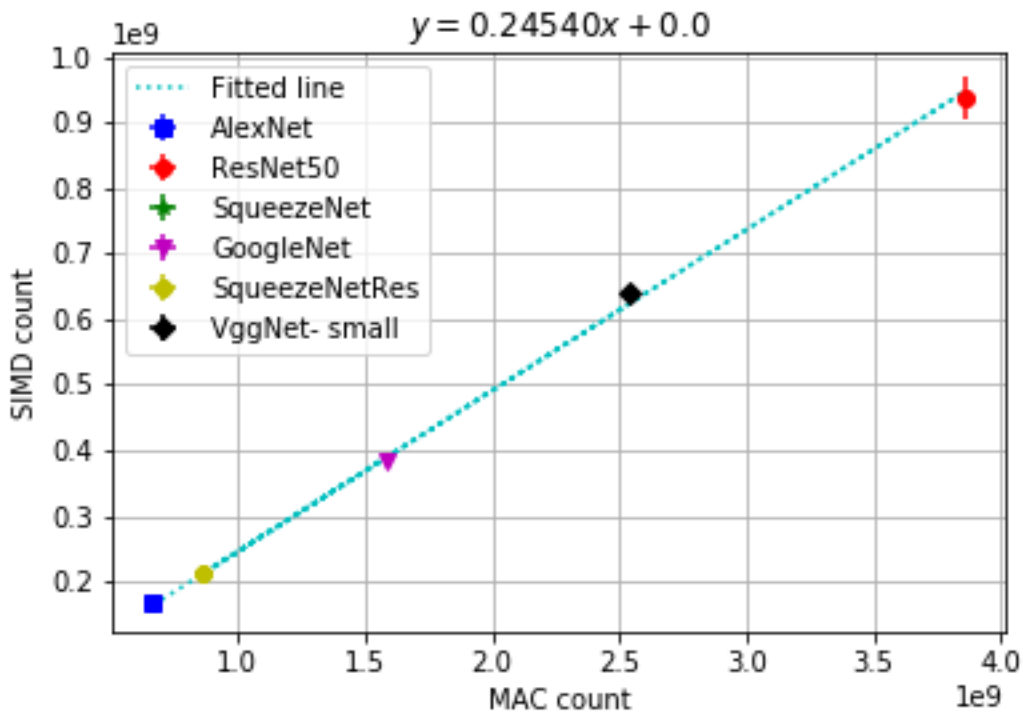
I don't want to measure



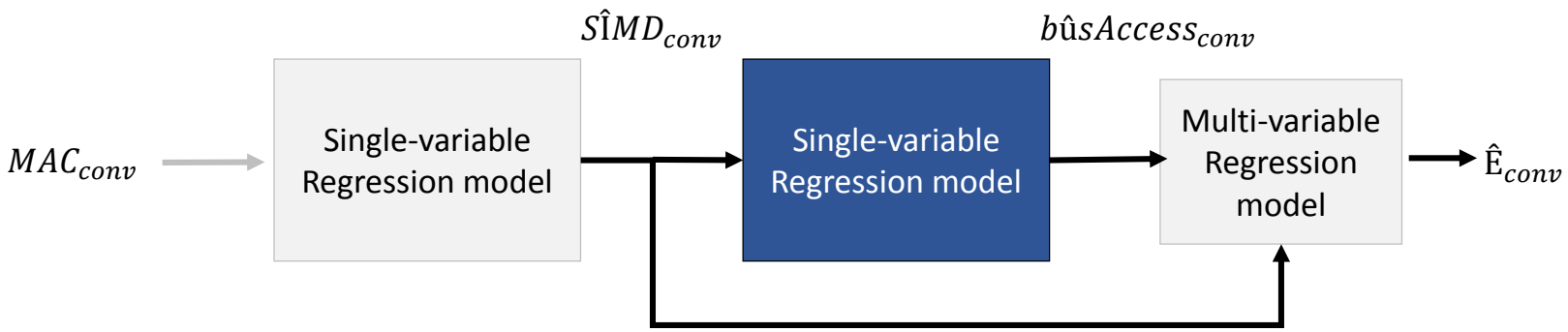
Predict SIMD



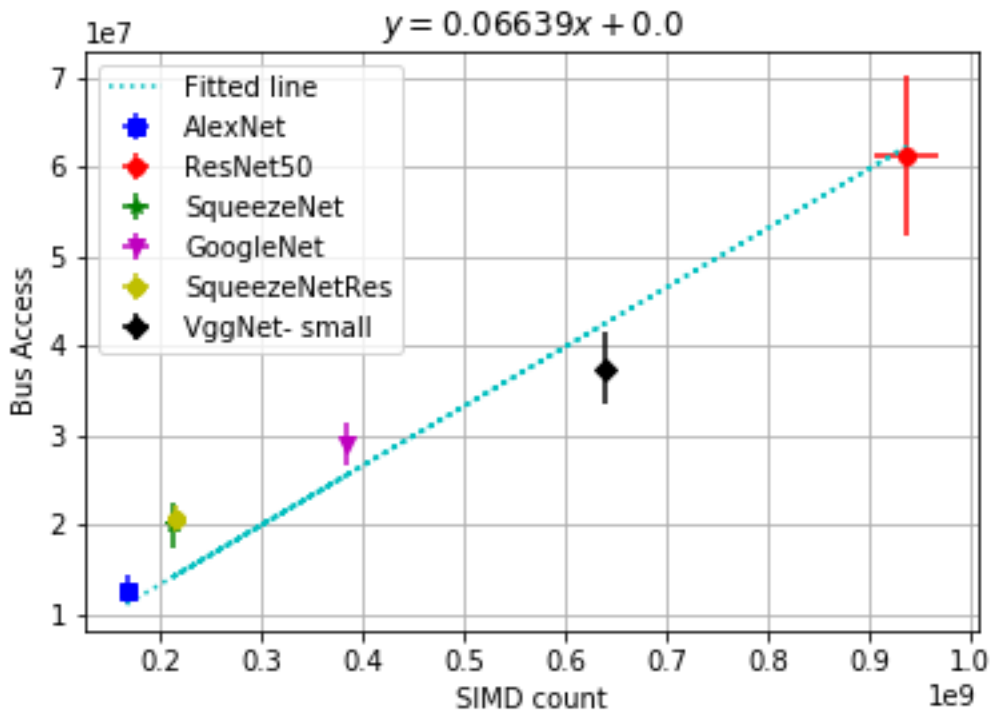
MAC to SIMD relationship



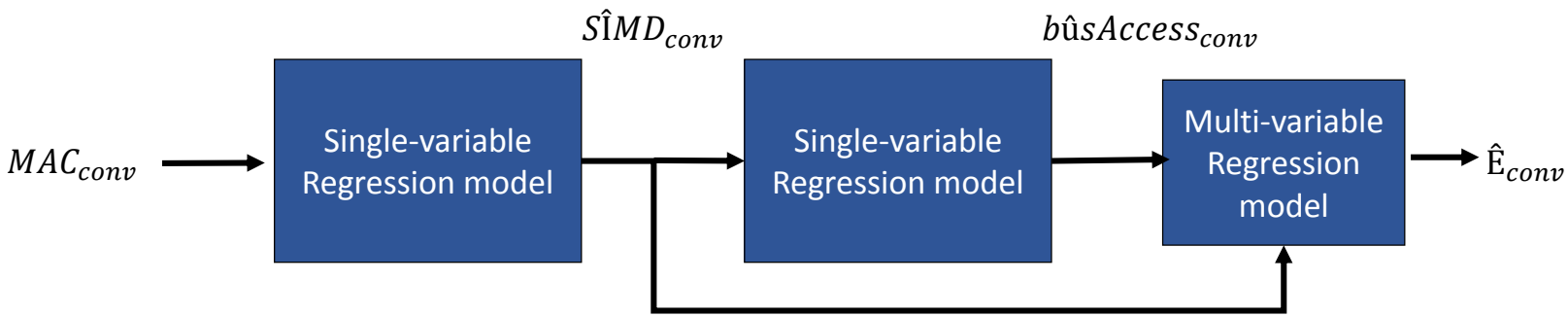
Predict bus access



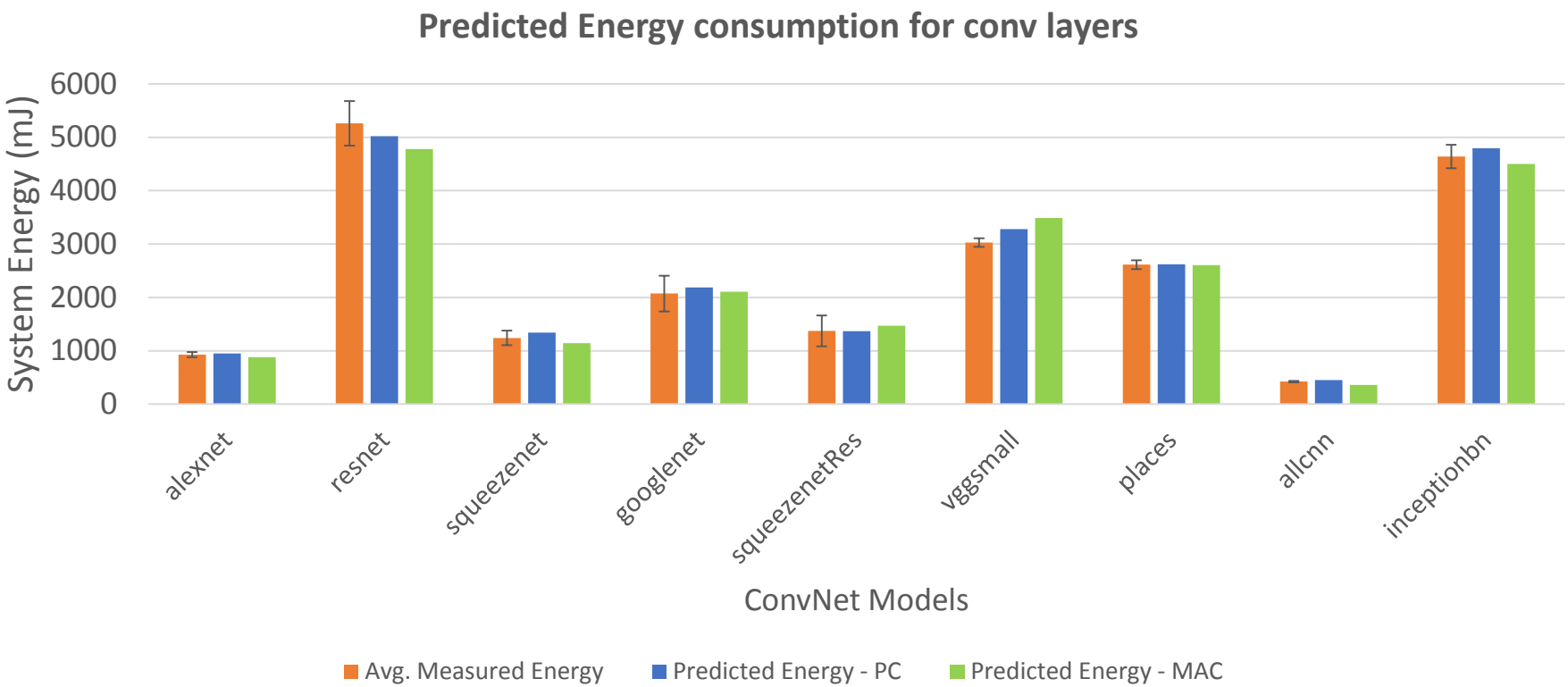
SIMD to bus access relationship



MAC to Energy relationship?



Final energy prediction



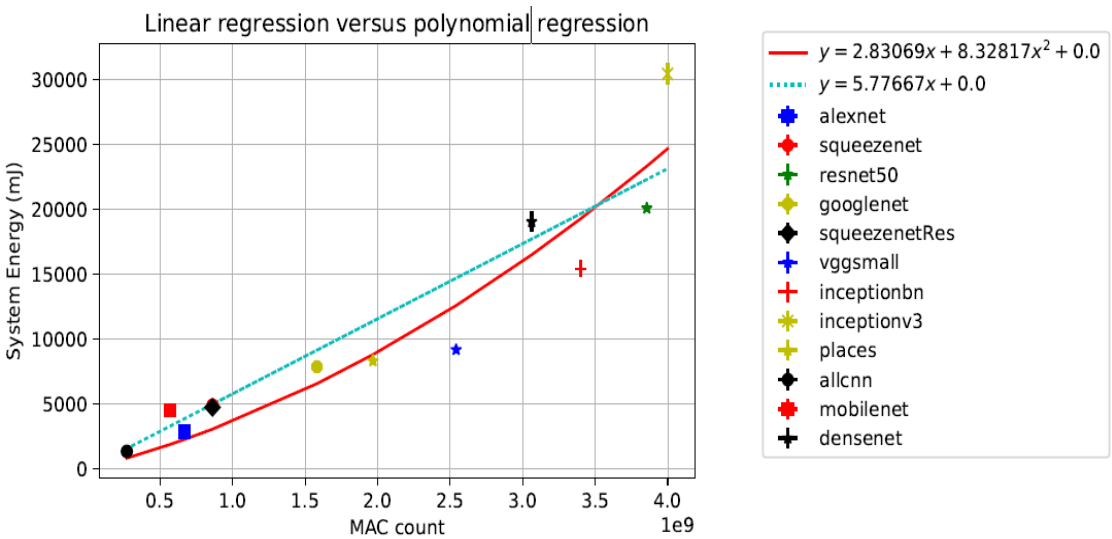
Avg. Relative Test Error = 7.08 ± 6.0 %

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

In progress work

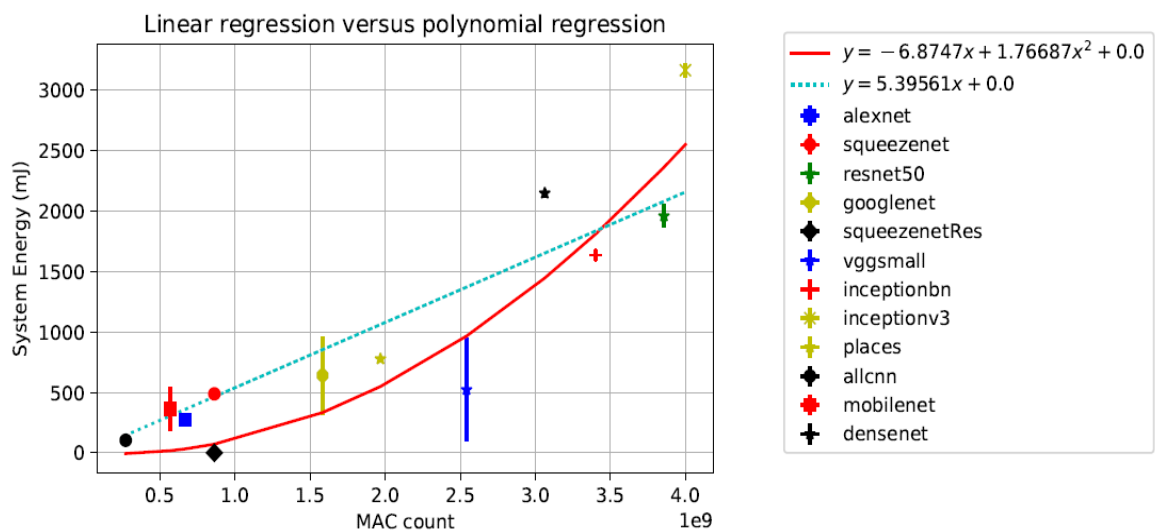
- ✓ New Caffe: **Caffe2**
- ✓ Other library backends e.g **Eigen library**
  - Jetson TX1
- ✓ Other platforms e.g. **Snapdragon 820**
  - Eigen library
- ✓ Non-linear energy prediction models e.g Polynomial regression
- Full ConvNet predictions e.g Pooling layer, fc layer

Extended work – Conv layers



Jetson TX1- Eigen library

Snapdragon 820 - Eigen library



Any questions?

crefeda.rodrigues@postgrad.manchester.ac.uk

The University of Manchester

References:

- 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