POMMEL: Exploring Off-Chip Memory Energy & Power Consumption in Convolutional Neural Network Accelerators

Alexander Montgomerie-Corcoran and Christos-Savvas Bouganis

intelligent Digital Systems Lab

Dept. of Electrical and Electronic Engineering

www.imperial.ac.uk/idsl

AlexMontgomerie/pommel

Motivation

What is the motivation for this work?

Ability to understand the ***impact of memory power*** *consumption early on in the design process*

Rapidly evaluate *the power consumption of any given* ***CNN accelerator, memory or network***

Explore the ***impact of coding schemes*** *on the power consumption*

Contribution

What does POMMEL do?

This tool estimates memory subsystem power consumption for a given ***memory type, accelerator*** *and* ***network***

What do I need to run the tool?

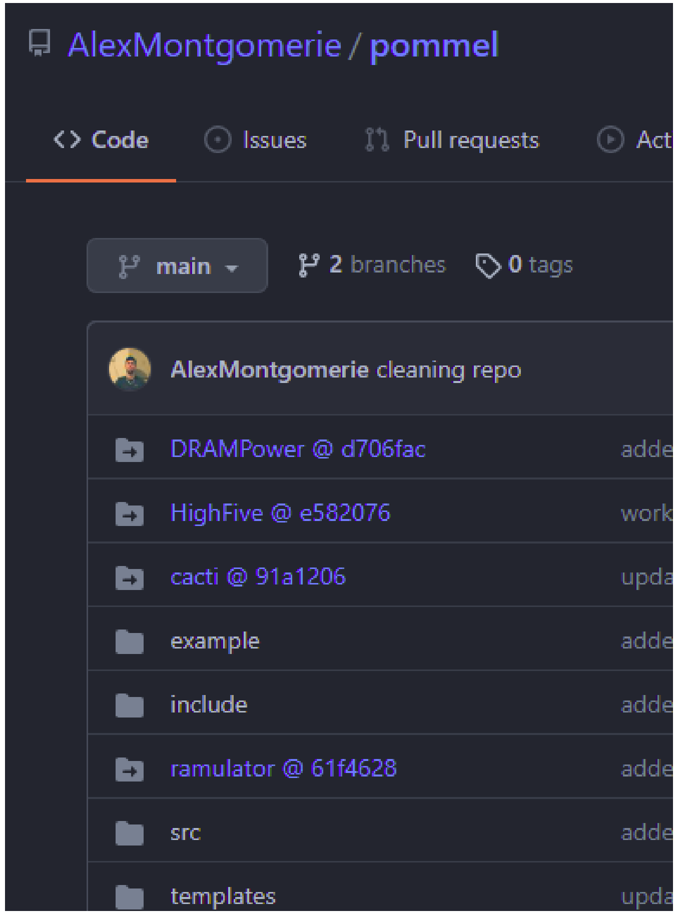
Only requires ***three*** *high-level* ***configuration files*** *to run*

What does the tool produce?

Produces a report with a ***breakdown of power consumption*** *for the memory subsystem*

Can I use the tool?

It has been open-sourced on github



AlexMontgomerie / pommel

AlexMontgomerie cleaning repo

DRAMPower @ d706fac	adde
HighFive @ e582076	work
cacti @ 91a1206	upda
example	adde
include	adde
ramulator @ 61f4628	adde
src	adde
templates	upda

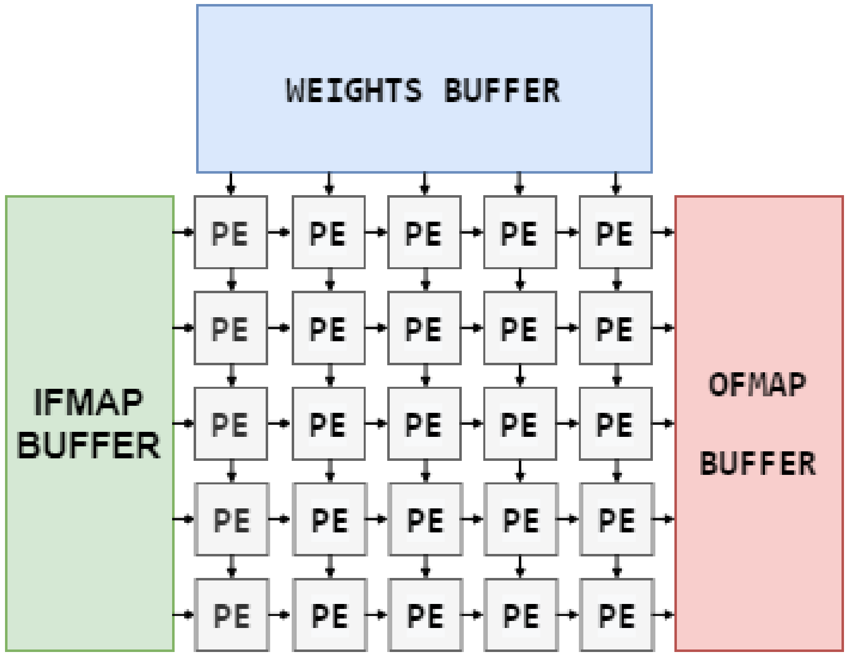
Background

- Convolutional Neural Network Accelerators
- CNN Accelerator Memory Subsystem
- Power Consumption in DRAM

Background: CNN Accelerators

What does a CNN accelerator do?

Accelerates the convolution layers in CNN models

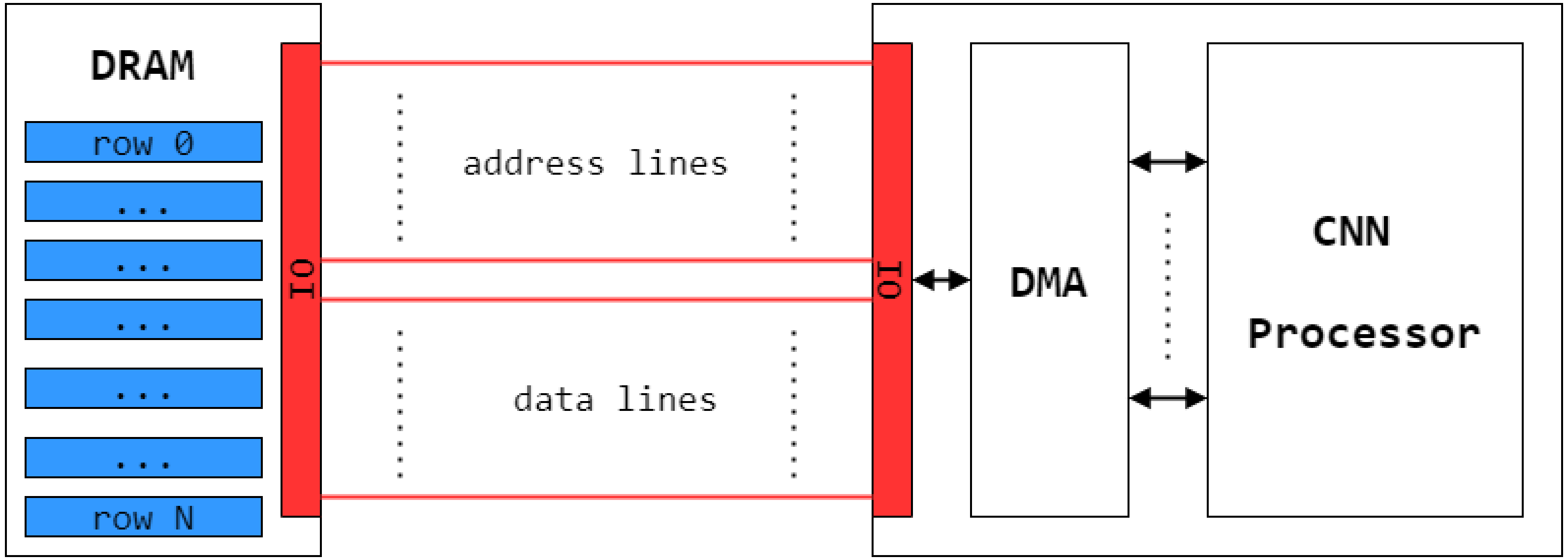


- Systolic Arrays (**SA**) are the most common type of accelerator architecture:
  - **TPU** [1]
  - **EYERISS** [2]
- Processing Element **(PE)** performs MAC operations for **computing kernel dot products**
- SA accelerators have three main on-chip SRAM buffers:
  - **IFMAP**: input feature-map
  - **WEIGHTS**: convolution parameters
  - **OFMAP**: output feature-map

Background: Memory Subsystem in CNN Accelerators

What is the memory subsystem?

Off-chip memory used to store feature-maps and weights



Feature-maps are typically **100x larger than weights,** and experience **computational bottlenecks**

Background: IO Power of the Memory Subsystem

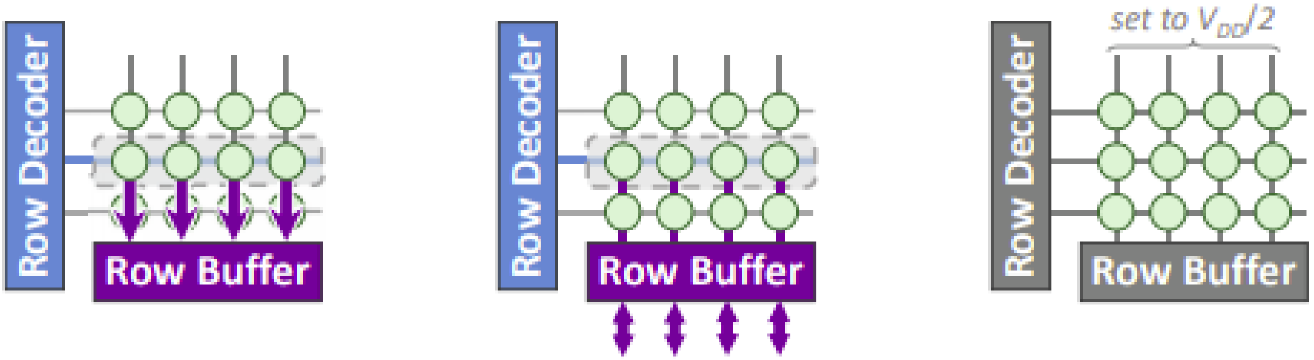
IO Power Components

IO Dynamic Power	Power consumed by load capacitances
Termination Power	Power consumed within the IO terminations
Interconnect Power	Power dissipated along the DRAM to accelerator bus
PHY Power	Power from other components present in the memory subsystem

*Contains both **Static** and **Dynamic** power that vary based on **bandwidth** and **activity***

Background: DRAM Power of the Memory Subsystem

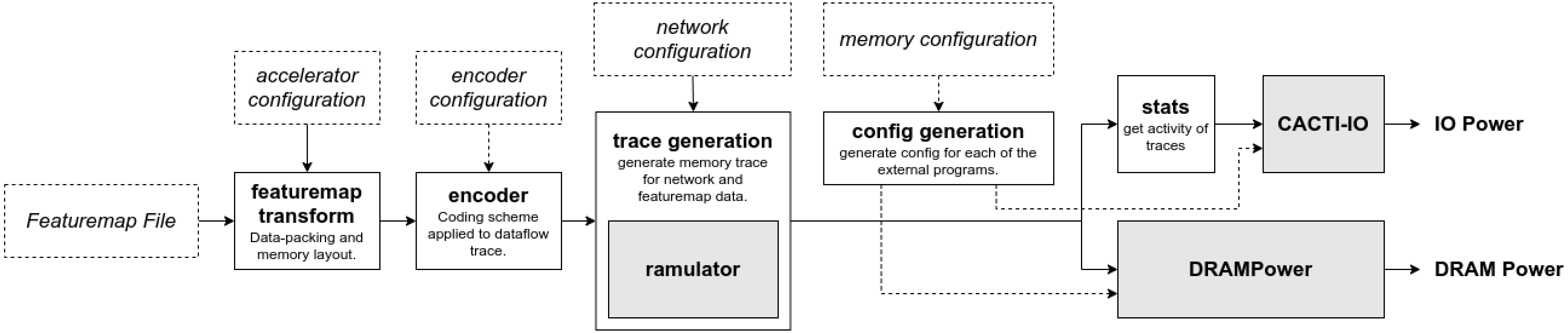
- Power consumed by cells within the DRAM chip
- Different commands (ACT, PRE, READ, WRITE) consume different amounts of energy
- More power consumed when actively reading and writing than when idle
- (Mostly) consumes **static** power



(a) Activate (b) Read/write (c) Precharge

From *What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study [3]*

POMMEL Framework



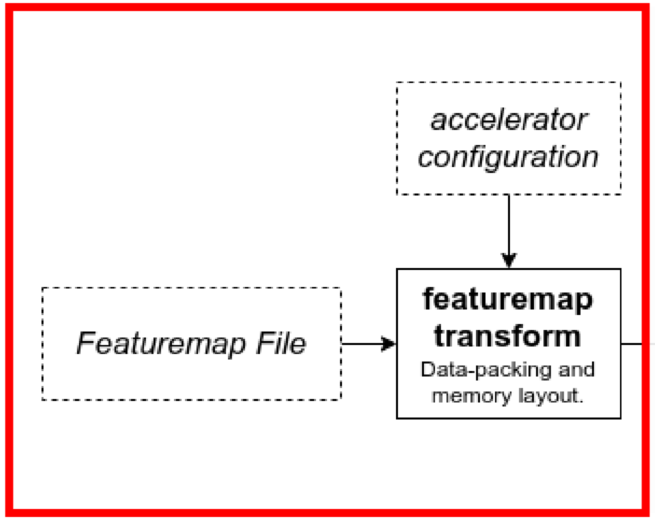
Inputs:

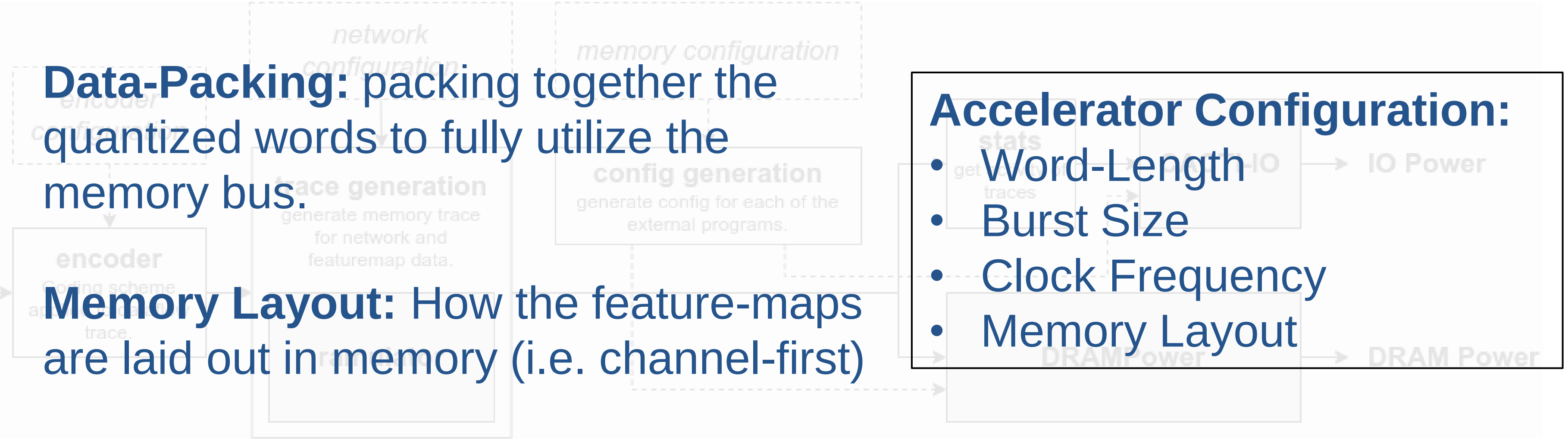
- Feature-map file
- Configuration files:
  - Accelerator
  - Network
  - Memory

Outputs:

- IO Power
- DRAM Power

POMMEL Framework: Feature-map Transform





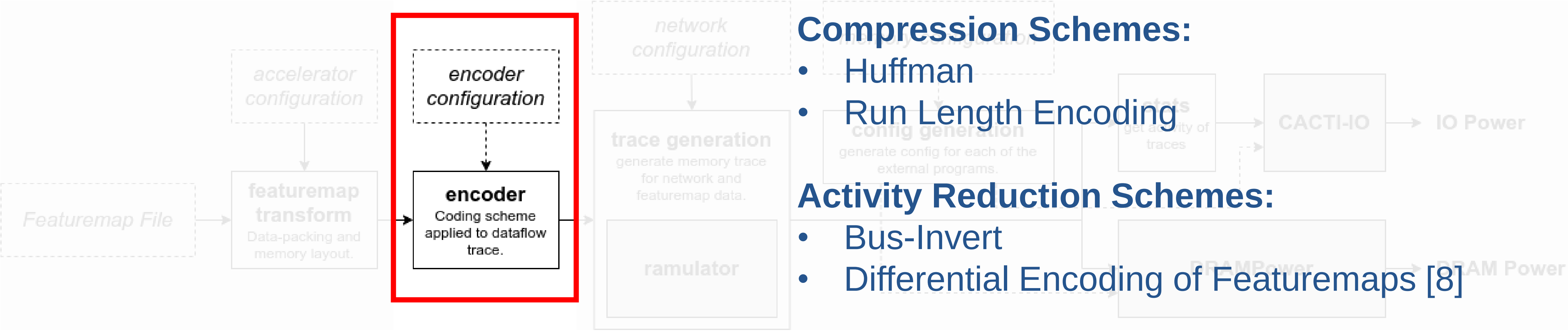
Data-Packing: packing together the quantized words to fully utilize the memory bus.

Memory Layout: How the feature-maps are laid out in memory (i.e. channel-first)

Accelerator Configuration:

- Word-Length
- Burst Size
- Clock Frequency
- Memory Layout

POMMEL Framework: Encoding (Optional)



Compression Schemes:

- Huffman
- Run Length Encoding

Activity Reduction Schemes:

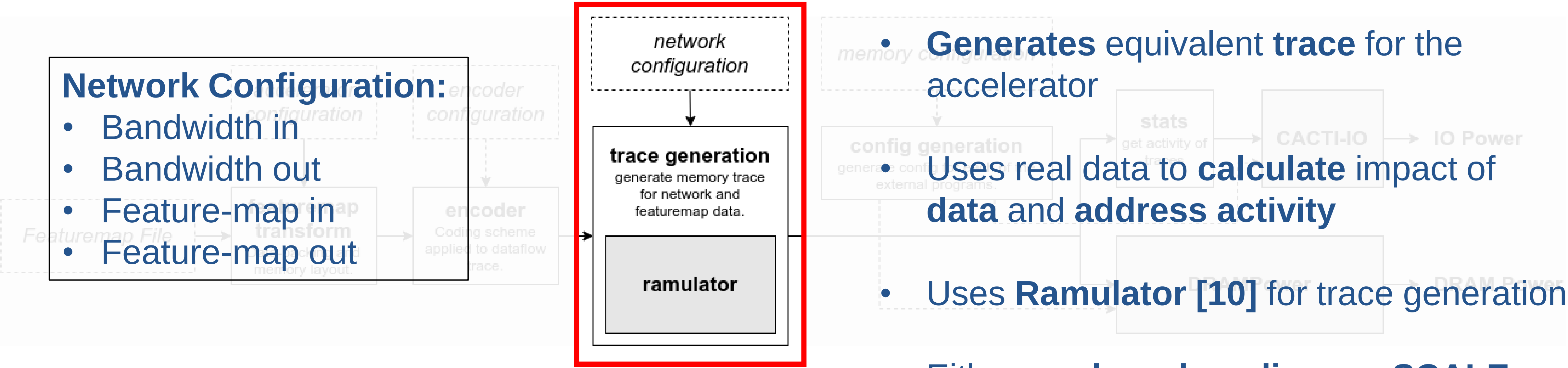
- Bus-Invert
- Differential Encoding of Featuremaps [8]

Ability to add custom encoders

POMMEL Framework: Trace Generation

Network Configuration:

- Bandwidth in
- Bandwidth out
- Feature-map in
- Feature-map out

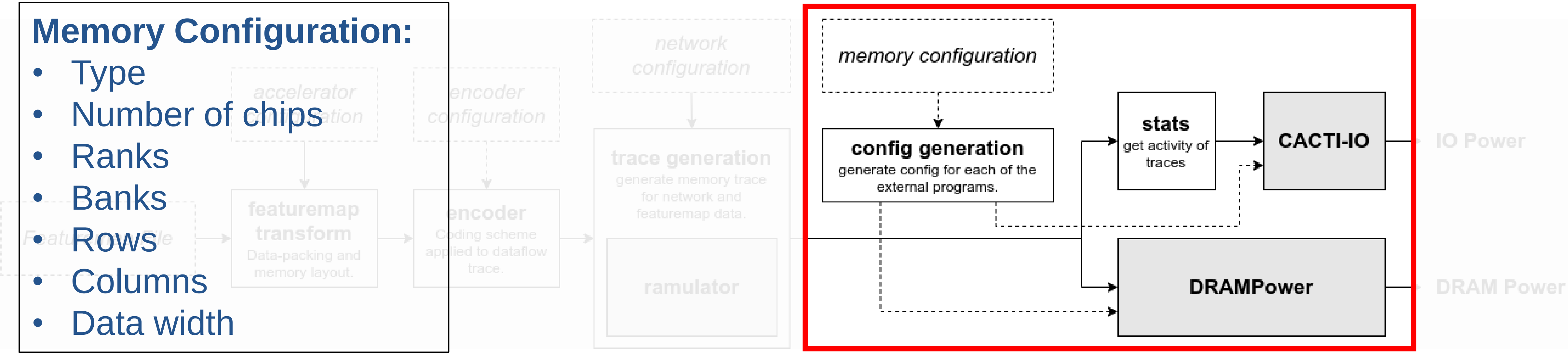


- **Generates** equivalent **trace** for the accelerator
- Uses real data to **calculate** impact of **data** and **address activity**
- Uses **Ramulator [10]** for trace generation
- Either use **board readings** or **SCALE-Sim [11] estimates**

POMMEL Framework: Power Estimation

Memory Configuration:

- Type
- Number of chips
- Ranks
- Banks
- Rows
- Columns
- Data width



- Generates configurations for power estimation tools
- Uses trace files as well as bandwidth and activity statistics

- **CACTI-IO [9]**: memory interface
- **DRAMPower [3]**: DRAM cells

Evaluation

- Accuracy of the tool
- Characteristics of different memory types
- Accelerator Investigation
- Encoder Investigation

Evaluation: Accuracy of the Tool

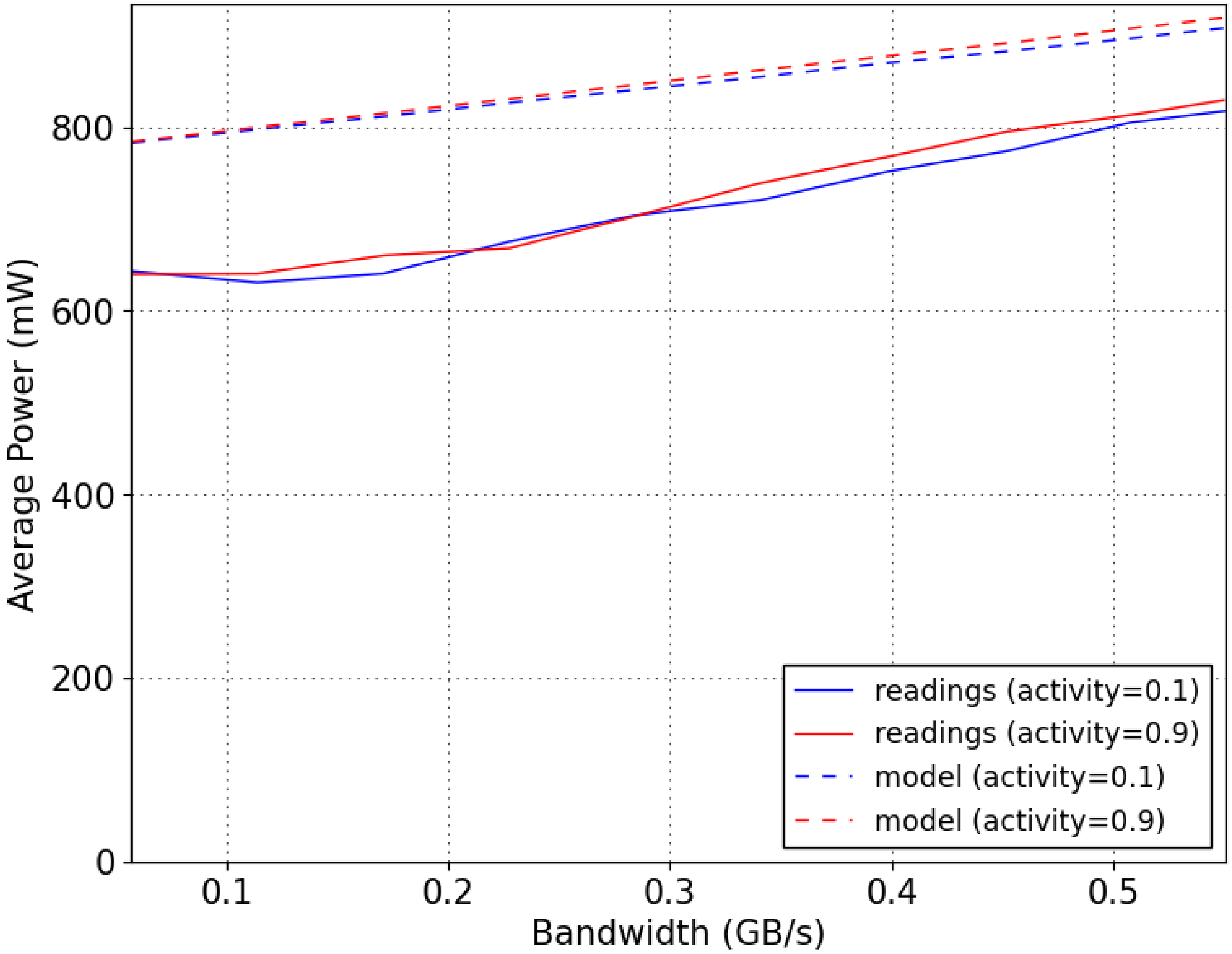


Fig. 1: Comparison of estimated and actual power readings for DDR3 memory

DDR3	Static Power	Bandwidth Coefficient	Activity Coefficient
Model	768.4	249.0	5.3
Actual	594.0	390.5	5.3

Tab. 2: Comparison of estimated and actual static power and bandwidth and activity coefficients

- **Similar bandwidth** and **activity** coefficients
- **Static power** is **much lower** in reality

Evaluation: Power Coefficients of Memory Types

What are the characteristics for different types of memory?

DRAM Type	Static Power (mW)	Bandwidth Coefficient (GB/s/mW)	Activity Coefficient (GT/s/mW)
DDR3	768.4	253.7	5.3
DDR3L	268.0	230.7	4.5
DDR4	151.7	171.5	3.1
LP-DDR2	288.5	142.9	20.3
LP-DDR3	157.3	144.1	13.7

Tab. 1: Comparison of estimated static power and bandwidth and activity coefficients for different types of memory

Evaluation: Accelerator Power Comparison

Findings:

- Investigation for **EYERISS** [1],**SCNN** [4] and **ShiDianNao** [5] Accelerators (using SCALE-SIM [11])
- Shows **dominance of static power** in all systems
- Performance has a significant impact on total energy

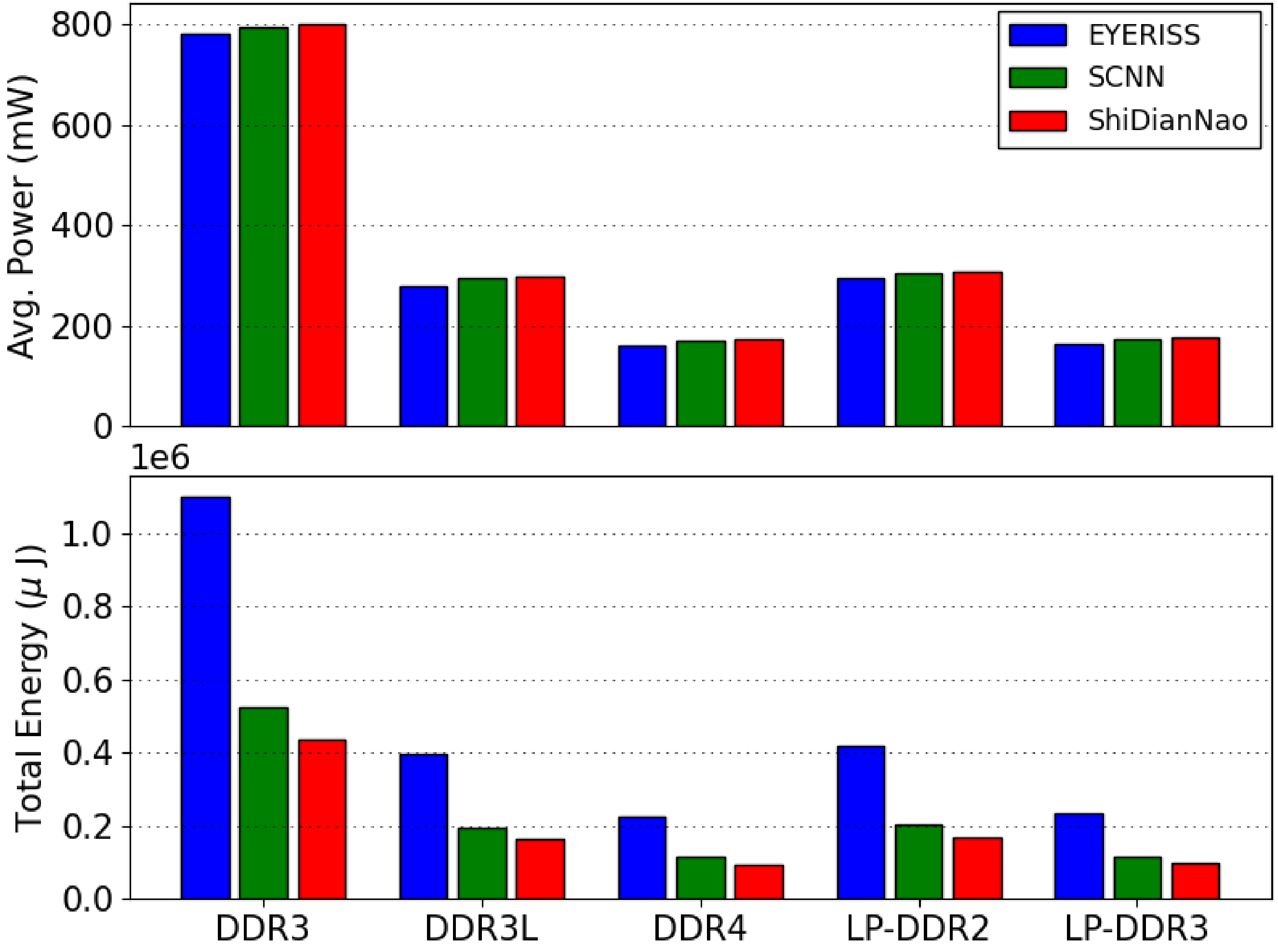


Fig. 2: Comparison of Energy and Power for different research accelerators and different memories running ResNet18 [6]

Evaluation: Comparison of Coding Schemes

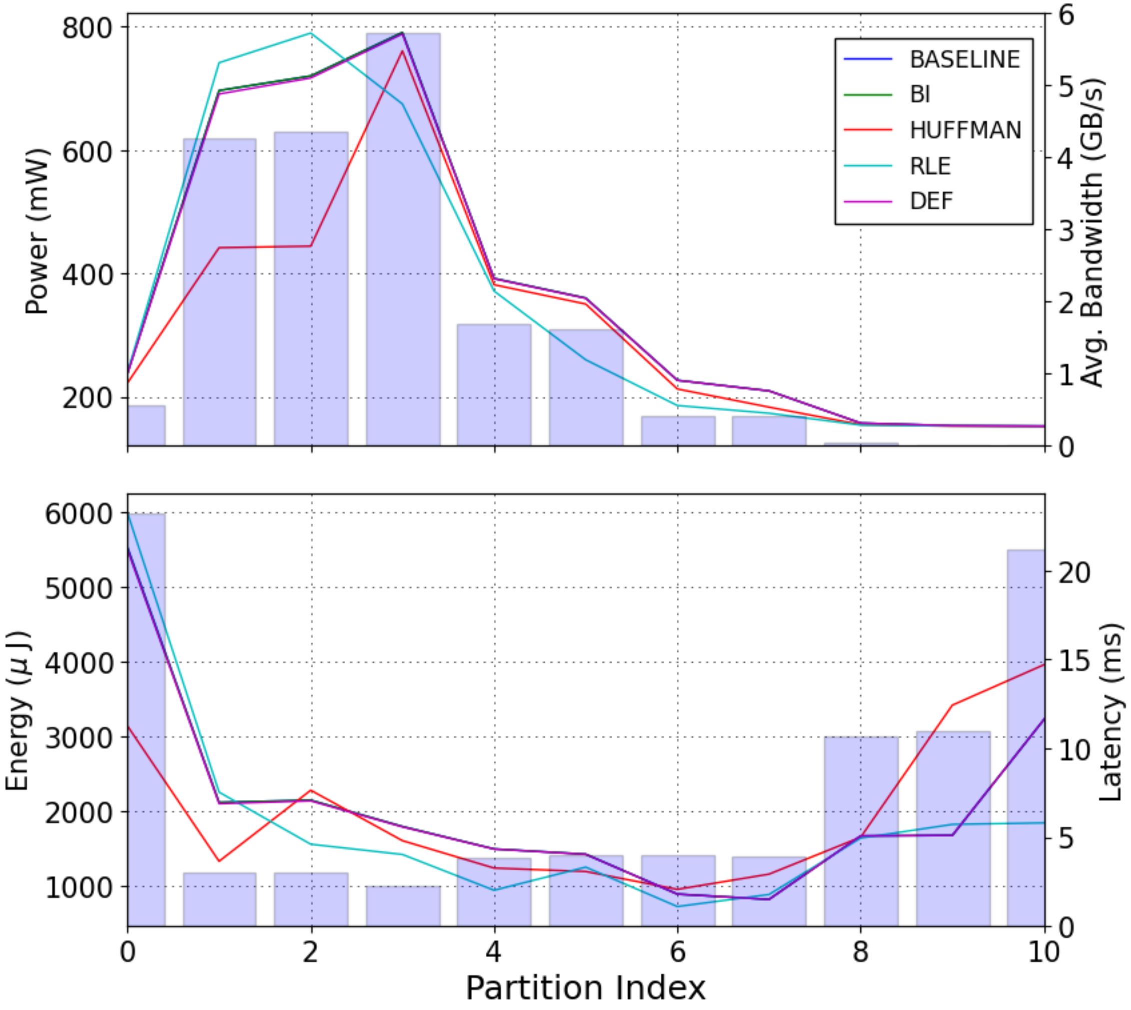


Fig. 3: Comparison of Energy, Power, Bandwidth and Latency per layer for different coding schemes on a TPU-like accelerator running VGG11 [7] on DDR4

Findings:

- **Huffman and RLE** have the **most impact** on power and energy
- **Activity reduction** schemes have **no noticeable impact** on power and energy
- **Power reduction** is only realised in **high-bandwidth layers** of the network
- **Low-bandwidth layers** typically have the **largest impact on total energy** usage

Conclusion

- Presented a new open-source tool for evaluating the memory power consumption for CNN accelerator systems
- It can evaluate power for a given network, accelerator and type of memory
- The accuracy of the tool is shown to be acceptable
- It can be used to investigate power optimization techniques at a high level

Thank you for listening!

AlexMontgomerie/pommel   am9215@ic.ac.uk

References

1. N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), 2017

2. Y. Chen, T. Krishna, J. S. Emer and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," in IEEE Journal of Solid-State Circuits, vol. 52, no. 1, pp. 127-138, Jan. 2017

3. Saugata Ghose, Abdullah Giray Yaglikçi, Raghav Gupta, Donghyuk Lee, Kais Kudrolli, William X. Liu, Hasan Hassan, Kevin K. Chang, Niladrish Chatterjee, Aditya Agrawal, Mike O'Connor, and Onur Mutlu. 2018. What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study. Proc. ACM Meas. Anal. Comput. Syst. 2, 3, Article 38 (December 2018

4. Angshuman Parashar, Minsoo Rhu, Anurag Mukkara, Antonio Puglielli, Rangharajan Venkatesan, Brucek Khailany, Joel Emer, Stephen W. Keckler, & William J. Dally. (2017). SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks.

5. Z. Du et al., "ShiDianNao: Shifting vision processing closer to the sensor," 2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA), 2015

6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. (2015). Deep Residual Learning for Image Recognition.

7. S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015

8. A. Montgomerie-Corcoran and C. Savvas-Bouganis, "DEF: Differential Encoding of Featuremaps for Low Power Convolutional Neural Network Accelerators," 2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC), 2021

9. N. P. Jouppi, A. B. Kahng, N. Muralimanohar and V. Srinivas, "CACTI-IO: CACTI with off-chip power-area-timing models," 2012 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2012

10. Y. Kim, W. Yang and O. Mutlu, "Ramulator: A Fast and Extensible DRAM Simulator," in IEEE Computer Architecture Letters, vol. 15, no. 1, pp. 45-49, 1 Jan.-June 2016

11. Ananda Samajdar and Yuhao Zhu and Paul N. Whatmough and Matthew Mattina and Tushar Krishna (2018). SCALE-Sim: Systolic CNN Accelerator. CoRR