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.
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

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IO Dynamic Power</strong></td>
<td>Power consumed by load capacitances</td>
</tr>
<tr>
<td><strong>Termination Power</strong></td>
<td>Power consumed within the IO terminations</td>
</tr>
<tr>
<td><strong>Interconnect Power</strong></td>
<td>Power dissipated along the DRAM to accelerator bus</td>
</tr>
<tr>
<td><strong>PHY Power</strong></td>
<td>Power from other components present in the memory subsystem</td>
</tr>
</tbody>
</table>

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

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**

<table>
<thead>
<tr>
<th>DDR3</th>
<th>Static Power</th>
<th>Bandwidth Coefficient</th>
<th>Activity Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>768.4</td>
<td>249.0</td>
<td>5.3</td>
</tr>
<tr>
<td>Actual</td>
<td>594.0</td>
<td>390.5</td>
<td>5.3</td>
</tr>
</tbody>
</table>

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?

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

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

Findings:


• 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

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**

*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*
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


