Investigation on the Power Consumption Models for ML Accelerator

By Xuesi Chen, ECE '22

Abstract

ML accelerators are specialized hardwares that are designed to process machine learning applications. The main components of a typical Machine Learning (ML) accelerator is a matrix of Multiply-Accumulate (MAC) Unit, also called Processing Element (PE) that focus on processing matrix multiply and accumulation, which is often referred as the systolic array. To supply the systolic array with enough memory for computation, memory buffers are placed near the systolic array on the ML accelerator. The goal of this paper is to summarize the power consumption of the MAC/PE unit on the systolic array and estimate the power consumption of SRAM buffers on three different ML accelerators: ML accelerators mapped to Field Programmable Gate Array (FPGA), and Application-specific Integrated Circuit(ASIC) based ML accelerators.

Introduction

The current ML accelerator model we use suffers from limited performance data with no energy and power simulation results regarding each part of the accelerator. Therefore, in order to accurately predict the power and energy use of each PE units and SRAM buffers on chip, as well as knowing how much area they occupy on a chip, we need to collect power and area information for each components of ML accelerators in use. The paper focuses on summarizing data of few different ML accelerators, each including power and area information with different semiconductor technologies and modeling platforms. In the rest of the report, I will be summarizing the takeaway messages from each of the model and provide equations on how to estimate the power consumption of the overall ML accelerator using the power consumption of individual PE units per operation and conclude the paper with a power and area models our group adopts for simulating the power consumption for ML accelerators.

Motivation

SCALE-Sim is the baseline ML accelerator framework we are relying on to study the power and performance behavior of a ML accelerators. Though SCALE-Sim does a detailed job at capturing data movement, it lacks the the necessary power and area information we need to build a hotspot mitigation tool [4] [3]. Finding an existing research that has the area and energy specification accurately modeled for PE units and SRAM buffers is crucial to the success of our research. Below are the summaries of papers I found helpful for gathering and calculating the energy and area data for ML accelerators.

ML Accelerator Power Consumption on Different Models

FPGA Model

FPGAs are common semiconductor devices for synthesize ML accelerators. The configurable logic blocks connected via programmable interconnects on the chip allows flexible designs of MAC units. The paper Low Power Datapath Architecture for Multiply - Accumulate(MAC) Unit synthesized the power and area of a single MAC unit using ISE Xilinx Version 14.7. Table I and II shows synthesized MAC area, delay, and power results based on conditions, such as MAC with or without compressors, MAC units with or without the adders, etc. The data shows us a general trend of how bit representations and different adders affect the area and power of MAC units. However, given it is modeled on FPGA, the area and power and especially delay is probably higher than other models given that FPGA nodes are usually older and bigger [5].

TABLE I. SYNTHESIS AND COMPARISON RESULTS OF MAC UNIT IN TERMS OF AREA, DELAY, POWER.

Multiplier		Area(um^2)	Delay(ns)	Power(nw)
MAC without compressor	Baugh_8bit	1264	30.953	44747.364
	Baugh_16bit	5147	57.985	267166.003
	Baugh_32bit	20556	111.433	1684137.231
Proposed MAC with compressor	Baugh_8bit	1259	30.578	38478.422
	Baugh_16bit	5141	57.110	241560.170
	Baugh_32bit	20550	108.978	1584558.026

TABLE II. SYNTHESIS AND COMPARISON RESULTS OF MAC UNIT WITH DIFFERENT ADDERS.

Multiplier with Adders		Area(um^2)	Delay(ns)	Power(nw)
MAC without compressor	RCA	1249	30.730	44197.016
	CSA	1255	33.369	44346.959
	Brent - Kung	1240	33.200	44197.016
Proposed MAC with compressor	RCA	1240	30.502	38182.612
	CSA	1246	34.484	38311.156
	Brent - Kung	1249	32.937	44318.720

[5]

ASIC Model

ASIC is a piece of hardware that is design for a particular use, in this case accelerating ML workloads, rather than for general-purpose. Some of the ML accelerator in ASIC form are Google TPU and MIT Eyeriss. The paper Thermal-Aware Design Space Exploration of 3-D Systolic ML Accelerators does a system-level design space exploration between 3D memory and 2D memory ASIC ML accelerators. It considers a power, performance and thermal trade off between 2D and 3D systolic ML accelerator designs. This paper includes how PE power, SRAM power and DRAM power is calculated for systolic ML accelerators, which is so crucial to the design we are doing [2].

	PE	SRAM	DRAM
Tech. node	14/16 nm	14/16 nm	28 nm
Energy	0.3 pJ	[1.1, 1.5] pJ	120 pJ
Area	525 um2	32502 um2/32 KB	N/A (off-chip)

Figure 1. Energy and Area specs of 14/16nm PE and SRAM[2]

The paper On-Chip Memory Technology Design Space Explorations for Mobile Deep Neural Network Accelerators does a design space exploration of the on-chip memory technologies and co-design for systolic array. The memory technologies evaluated are SRAM, eDRAM, MRAM, and 3 vertical RRAM. Different design are evaluated on models: ResNet-50, MobileNet, and Faster-RCNN. The paper evaluates SCALE-Sim as the baseline evaluation tool, which is also the simulation tool we are using for our research. The biggest pro of the paper is that it provides 14/16nm tech node for PE and SRAM, which is the current industry standard. Cons of this paper is that the PE design data is only limited to 16X16, 24X24, 32X32, which is not applicable to TPU hardware configuration, which has 512x512. The Table III is a summary of energy and area information for each memory components or technologies are shown in Table III [1].

Table III. Energy and Area specs of 14/16nm PE and SRAM with other emerging memory technologies [1]

	PE	SRAM	MRAM	3D VRRAM	eDRAM	DRAM
Tech. node	14/16 nm	14/16 nm	28 nm	28 nm	28 nm	28 nm
Energy	0.3 pJ	[1.1, 1.5] pJ	Read: 4 pJ Write: 14 pJ	Read: 16 pJ Write: 48 pJ	19 pJ	120 pJ
Area	525 μm^2	32502 μm^2/32 KB	0.017 μm^2/bit	0.004 μm^2/bit	0.035 μm^2/bit	N/A
Design space	{16 × 16, 24 × 24, 32 × 32}	Weight/IFMap/OFMap: {32, 64, 128, 256, 1024} KB	MRAM-only (no off-chip DRAM)	VRRAM-only VRRAM + DRAM	eDRAM-only	LPDDR3

Power Consumption of the Whole Chip

For the purpose of best modeling the energy and area information of PE and SRAM on ML accelerators, we choose to use the memory technology configurations provided by On-Chip Memory Technology Design Space Explorations for Mobile Deep Neural Network Accelerators. We believe the 14/16 nm tech node evaluated on SCALE-Sim baseline is the best modeling configuration we can find in existing published documents.

To calculate the power consumption of PE and SRAM buffer, we will be using the equations provided by Thermal-Aware Design Space Exploration of 3-D Systolic ML Accelerators since the fundamental energy consumption theory on ML accelerators are transferable.

$$P_{\text{PE}} = \frac{\sum_{i=1}^{n}(\text{util}(i) * \text{arr\_}h * \text{arr\_}w * e\_\text{mac} * \text{cyc}(i))}{\text{cycles} * \frac{1}{\text{freq}} * 100} \quad (1)$$

$$P_{\text{SRAM}} = \frac{\sum_{i=1}^{n}((\text{srd\_bw}(i) * e\_\text{srd} + \text{swt\_bw} * e\_\text{swt}) * \text{cyc}(i))}{\text{cycles} * \frac{1}{\text{freq}}} \quad (2)$$

$$P_{\text{DRAM}} = \frac{\sum_{i=1}^{n}((d\_\text{if}(i) + d\_\text{filt}(i) + d\_\text{of}(i)) * e\_\text{mem} * \text{cyc}(i))}{\text{cycles} * \frac{1}{\text{freq}}} \quad (3)$$

Figure 2. Equations for modeling PE, SRAM and DRAM power usage.

N is the number of layers in a given neural networks, util(i) correspond to the utilization of the ith layer. arr h and arr w are the height and width of PE array. e mac is the energy consumed by a mac unit per operation. cyc represents the clock cycle. [2]

References

1. Haitong Li et al. "On-Chip Memory Technology Design Space Explorations for Mobile Deep Neural Network Accelerators". *In: Proceedings of the 56th Annual Design Automation Conference 2019*. DAC '19: The 56th Annual Design Automation Conference 2019. Las Vegas NV USA: ACM, June 2, 2019, pp. 1–6. isbn: 978-1-4503-6725-7. doi: 10.1145/3316781.3317874. url: https://dl.acm.org/doi/10.1145/3316781. 3317874.

2. Rahul Mathur et al. "Thermal-Aware Design Space Exploration of 3-D Systolic ML Accelerators". In: *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* 7.1 (June 2021), pp. 70–78. issn: 2329-9231. doi: 10.1109/JXCDC.2021.3092436. url: https://ieeexplore.ieee.org/document/ 9464955/.

3. Ananda Samajdar et al. "A systematic methodology for characterizing scalability of DNN accelerators using SCALE-sim". In: *2020 IEEE International Symposium on Performance Analysis of Systems and Software* (ISPASS). IEEE. 2020, pp. 58–68.

4. Ananda Samajdar et al. "SCALE-Sim: Systolic CNN Accelerator Simulator". In: *arXiv preprint arXiv*:1811.02883 (2018).

5.H R Spoorthi, C P Narendra, and U Chandra Mohan. "Low Power Datapath Architecture for Multiply - Accumulate (MAC) Unit". In: *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*. Bangalore, India: IEEE, May 2019, pp. 391–395. isbn: 978-1-72810-630-4. doi: 10.1109/RTEICT46194.2019.9016717. url: https: // ieeexplore.ieee.org/document/9016717/