The Era of AI Hardware

Myung-Hee Na, Ph.D
Distinguished Engineer, IBM Research
AI: Why Now?

- Exploding & Everywhere
- Rapid Advance of Neural Network
- Accessible & Affordable: Cloud
The path to Broad AI

Explainability & Fairness

Security & Robustness

Learns more from small data

Learns to transfer

AI Hardware

Broad AI is compute intensive!
AI End to End

Data

Algorithms

Compute
“Does AI Have a Hardware Problem?”

Editorial | 17 April 2018
Heterogeneous Systems for AI Workloads

CPU

Memory

Networking

AI Accelerator

OpenCAPI NVLink 2.0 PCIe Gen4
Summit:
World’s Most Powerful Supercomputer

200,000 trillion calculations per second

9216 Power 9 processors

27,648 GPUs
AI Accelerators

Now and Future
Deep Learning Performance/Watt Current Industry Trend

Trend: x 2.5 / year

- **GPUs - Incumbent Deep Learning Accelerators**
  - Thousands of simple cores optimized for Matrix Multiply and Accumulate (e.g. convolution)

- **Explosion of AI hardware start up: Near-term accelerators based on existing CMOS based technologies.**

- **GPU are based on existing CMOS-based technologies**

**How can we progress beyond the GPU?**
Extending the 2.5x/year through 2025

1. Digital AI Cores with Approximate Computing
2. Analog AI Cores
3. Analog AI Cores – Optimized Materials
The Next AI Hardware Platform: 8-bit Precision for DL Training Systems

**FP8 training scheme achieved FP32 baseline accuracies**

**New reduced-precision algorithm for deep learning**

- Break up the accumulation component of deep learning dot-products into *chunks*

---

**8-Bit Training Platform:**
- 16 bit → 8 bit for weight, activation, error and weight gradient.
- 32 bit → 16 bit for update and accumulation

---

N. Wang et al, NeurIPS 2018
Across the Stack Approximate Computing for DL Acceleration

Algorithms/Applications

Compression and Quantization techniques

- Compensated DNN: DAC 2018
  - AdaComp (AAAI 2018), PACT (arxiv)

Programming Model

Reduce Communication load

- Agrawal et al, ICRC 2016

Architecture

Customized dataflow architecture

- Fleischer et al, VLSI 2018
IBM Roadmap in Reduced Precision Scaling

- **8-bit Inference**
- **16-bit Training**

**2012 to 2024**
- **GPU Era for AI:** 1x
- **Specialized ASICs for AI:** > 10x

**2015**
- **16-bit Training**
- **8-bit Inference**

**2018**
- **4-bit Inference ASICs, J.Choi et al, submitted to SysML 2019**
- **2-bit Inference ASICs, J.Choi et al, submitted to SysML 2019**

**2019**
- **Reduced Precision Gupta et al, ICML 2015**
- **8-bit Training ASICs, N.Wang et al, NeurIPS 2018**
- **4-bit Inference ASICs, J.Choi et al**
Analog AI Cores for In-Memory Computation

Eliminate the Von-Neumann Bottleneck

Map DNNs to analog cross-point arrays

NVM materials in array cross-points to store synaptic weights
Transfer Learning with Analog AI Cores NVM Arrays with PCM for Deep Learning

Projecting up to 100x faster while also 280x more energy-efficient

**Transfer Learning:**
using ImageNET to learn CIFAR-10/100

![Diagram of Transfer Learning]

Accuracy (%)
Epoch

https://doi.org/10.1038/s41586-018-0180-5
Optimizing for symmetry & linearity

**New Material Systems**

- Typical PCM material
- Ideal RPU

Invert polarity of pulse

Conductivity (inverse of resistance)

# of voltage pulses

**ECRAM**

Electro-Chemical Random-Access Memory

$$G_{ds} (nS)$$

- Pulse #
- $t_w = 1 \mu s$, $I_G = \pm 100 \mu A$

J. Tiang et al, IEDM 2018, Session 13.1
Thank you
감사합니다