Thank you, tinyML Strategic Partners, for committing to take tinyML to the next Level, together.
Executive Strategic Partners
Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more: www.arm.com/techtalks
EDGE IMPULSE

The Leading Development Platform for Edge ML

dgeimpulse.com
Advancing AI research to make efficient AI ubiquitous

- **Power efficiency**: Model design, compression, quantization, algorithms, efficient hardware, software tool
- **Personalization**: Continuous learning, contextual, always-on, privacy-preserved, distributed learning
- **Efficient learning**: Robust learning through minimal data, unsupervised learning, on-device learning

**Perception**: Object detection, speech recognition, contextual fusion

**Reasoning**: Scene understanding, language understanding, behavior prediction

**Action**: Reinforcement learning for decision making

A platform to scale AI across the industry

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Accelerate Your Edge Compute

Making Edge AI A Reality

www.syntiant.com
Platinum Strategic Partners
Fastest Video Analytics Solutions on Arm CPUs
High-Value or Safety-Critical Use Cases?

For your most important projects, use

TinyML software that covers the full engineering lifecycle: Reality AI Tools®

- AutoML for non-visual sensing based on advanced signal processing math
- Hardware design analytics
- Explanation of TinyML models in terms of underlying physics
- Automated Data Readiness assessment

https://reality.ai/  @SensorAI  info@reality.ai
Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.
Sony Semiconductor Solutions Corporation
Gold Strategic Partners
Witness potential made possible at analog.com.

Where what if becomes what is.
FOTA HUB

Making Over-the-Air Firmware and ML models Updates Simple and Accessible!

- Securely update your IoT devices regardless of their Hardware Platform (Silicon) Provider and physical location.
- Unlock TinyML business value through OTA Firmware and ML models update.
- Pay-as you-go

www.fotahub.com
contact@fotahub.com
TOGETHER, WE ACCELERATE THE BREAKTHROUGHS THAT ADVANCE OUR WORLD

www.nxp.com/ai
Deploy TinyML into the Real World - Plug and Play ML

Sensors:
- modulated and ready-to-use sensors to simplify the setup process
- support 500+ grove modules

Wio Terminal:
- completed AI platform integrated with a 2.4" LCD Screen, onboard IMU (LIS3DHTR), microphone, buzzer, microSD card slot, light sensor, infrared emitter (IR 940nm)

Codecraft:
- no code Programming platform to Get Started With TinML
- supports Arduino, Python, C or JavaScript etc.

Edge Impulse:
- to optimize data utilization and enable deploy a machine learning model faster than ever

TensorFlow Lite:
- to easily train low memory usage machine learning models

Wio Terminal
- Motion/Gesture/Speech/Smell/Sports
- Barcode/Face/Image

Applications
- AI powered thermal camera for safe camping
- Azure IoT Squirrel Feeder

Sense
Train
Inference
The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

Analytics Toolkit Suite

AutoML

Data Collection

Test & Validation

Code Generation

Model Building

Version Control and Model Management

sensiml.com/tinyML
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com
SynSense builds sensing and inference hardware for ultra-low-power (sub-mW) embedded, mobile and edge devices. We design systems for real-time always-on smart sensing, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai
Silver Strategic Partners
Join Growing tinyML Communities:

![Meetup](https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/)

*tinyML - Enabling ultra-low Power ML at the Edge*

11.5k members in 46 Groups in 37 Countries

![LinkedIn](https://www.linkedin.com/groups/13694488/)

*The tinyML Community*

3.1k members & 9.3k followers

https://www.linkedin.com/groups/13694488/
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)
www.youtube.com/tinyML
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Dr. Kaiyuan Yang is currently an Assistant Professor of ECE at Rice University, USA. He received his B.S. degree in Electronic Engineering from Tsinghua University, China, in 2012, and his Ph.D. degree in Electrical Engineering from the University of Michigan, Ann Arbor, in 2017. His research interests include digital and mixed-signal circuit and system design for secure and intelligent microsystems, bioelectronics, and hardware security. Dr. Yang is a recipient of the 2022 National Science Foundation (NSF) CAREER award, 2016 IEEE SSCS Predoctoral Achievement Award, and multiple best paper awards from premier conferences in various fields, including 2021 IEEE Custom Integrated Circuit Conference (CICC), 2016 IEEE Symposium on Security and Privacy (Oakland), 2015 IEEE International Symposium on Circuits and Systems (ISCAS), and the Best Student Paper Award finalist at 2022 RFIC and 2019 CICC.
Weier Wan

Dr. Weier Wan is currently leading the software-hardware co-design and is a founding member at Aizip, a Silicon Valley startup providing TinyML solutions. He received his Ph.D. degree in electrical engineering from Stanford University in 2022, where he worked on designing efficient AI hardware system to enable intelligence at the edge. His research work has been published in top journals and conferences, including Nature, International Solid-State Circuits Conference (ISSCC), and Symposium on VLSI Technology and Circuits. He is the first author of a monumental work published in Nature this year, titled “A compute-in-memory chip based on resistive random-access memory”. Previously, he received his master’s degree in electrical engineering from Stanford University in 2018 and his bachelor’s degree in physics, electrical engineering and computer sciences from University of California, Berkeley in 2015.
Processing-In-Memory for Efficient AI Inference at the Edge

Kaiyuan Yang
Assistant Professor

Weier Wan
Head of Software-Hardware Co-design
Battery Powered Smart Devices
Memory Wall in Data-Centric Computing

Data movement >> Arithmetic operations

Operation:  | Energy (pJ) |
-----------|------------|
8b Add     | 0.03       |
16b Add    | 0.05       |
32b Add    | 0.1        |
16b FP Add | 0.4        |
32b FP Add | 0.9        |
8b Mult    | 0.2        |
32b Mult   | 3.1        |
16b FP Mult| 1.1        |
32b FP Mult| 3.7        |
32b SRAM Read (8KB) | 5 |
32b DRAM Read | 640 |

M. Horowitz, “Computing’s Energy Problem: (and what we can do about it)” ISSCC 2014
Problem Definition: Matrix-Vector Multiplication (MVM)

- Inference of deep neural networks

- Multiply-and-Accumulate (MAC) to obtain each filter in CNN:

\[ Y_K = \sum_{i=1}^{R \times R \times C} W_{K,i} X_i \]

Y: Output feature map
X: Input feature map
R: Filter kernel size
C: # of channels
K: K\textsuperscript{th} position of the output
Promises of Process-In-Memory (PIM)

Von Neumann Architecture

\[ E_{VN} = ME_{WL} + M(E_{BL} + E_{SA}) \]
\[ T_{VN} = M(T_{Array} + T_{SA}) \]
Promises of Process-In-Memory (PIM)

Von Neumann Architecture

\[ E_{VN} = ME_{WL} + M(E_{BL} + E_{SA}) \]
\[ T_{VN} = M(T_{Array} + T_{SA}) \]

Process-In-Memory

\[ E_{IMC} = ME_{WL} + E_{BL} + E_{ADC} \]
\[ T_{IMC} = T_{Arr} + T_{ADC} \]

PIM amortizes read energy, reduces compute energy, increases bandwidth!
Outline of the talk

SRAM Based PIM

**CAP-RAM**
- SRAM Read/Write
- ADC
- DAC
- ADC CLK & SAR Logic
- 330μm

**DCT-RAM**
- 576x128 SRAM
- 512x128 SRAM
- 441μm

RRAM Based PIM

**NeuRRAM**
- 48 core
- 3M RRAMs
- 12K neurons

PIM SW/HW Co-Design

- Silicon IP
- PIM-net
- SoC Architecture
- PIM aware training

Chip Summary
- ADC
- Others
- DCC
- Area Breakdown
- Array
- 57.7%
- 576x128
- 8T SRAM
- DCC
- ADC
- One ADC per 8T SRAM
- 431 μM
- 526 μM

SoC Architecture

PIM aware training
Outline of the talk

**SRAM Based PIM**

**CAP-RAM**

**DCT-RAM**

**RRAM Based PIM**

**NeuRRAM**

48 core
3M RRAMs
12K neurons

**PIM SW/HW Co-Design**

- Silicon IP
- PIM-net
- SoC Architecture
- PIM aware training
Motivation of SRAM PIM

Weight storage accuracy and reliability

In-SRAM Computing in 7nm

Accurate weight storage

Technology readiness and scalability

Smaller periphery

V. Joshi, et al., Nat Comm 2020

Q. Dong, et al., ISSCC 2020

J. Lee, et al., VLSI 2021

W. Li, et al., CICC 2021
Prior Arts: Current-Domain PIM

\[ V_{SUM} = \frac{1}{C_{BL}} \sum_{i=1}^{N} I_i t_i = \sum_{i=1}^{N} W_i \tilde{I}_i t_i \]

- \( W_i \): Binary value in the SRAM cell
- \( \tilde{I}_i \): Modulated by gate voltage \( V_i \)
- \( t_i \): Controlled by input pulse width

**Compatible with standard 6T/8T cells**

**Transistor nonideality: nonlinearity and process variation**

**Restricted Parallelism**
Prior Arts: Charge-Domain PIM

A. Biswas et al., ISSCC 2018
H. Jia et al., JSSC 2020

Simulated linearity

\[ V_{\text{SUM}} = \frac{1}{NC_{BL}} \sum_{i=1}^{N} V_i = \sum_{i=1}^{N} W_i \tilde{V}_i \]

\( W_i \): Binary value in the SRAM cell
\( \tilde{V}_i \): Input voltage from DAC

Capacitor variation

\[ \sigma/V_{\text{DD}} \]

\( \sigma \): Standard deviation
\( V_{\text{DD}} \): Power supply voltage

\[ R^2 = 0.9999 \]

H. Valavi, et al., JSSC 2019

No transistor non-idealities

Relatively complex multi-step computing steps, mostly requiring custom memory cells and computing circuits
Prior Arts: Charge-Domain PIM – Bottom Plate Driving

One-Phase Computing

Example: PIM macro in 28 nm

Energy breakdown

Area breakdown

No transistor non-idealities
Small cell area: full parallelism
Require always-on analog buffers: power and area overheads
Additional Challenges of In-SRAM PIM Macro

• Energy and Area of Drivers

![Diagram showing DAC Driver, Memory Row, and Input Driver](image)

**OP-Amp for Charge Domain MAC**

- **500 MHz Charge-domain MAC**
  - \( E_{PCH} = C_{V} V_{S}^2 \)
  - \( = 180 \times 0.6^2 = 65 \text{ fJ} \)
  - \( I_D = \text{Slew Rate} \times C_{V} \)
  - \( = 0.6 \times 180 = 108 \mu \text{A} \)
  - \( E_D = I_D V_{DD} t_{IMC} \)
  - \( = 108 \times 1.2 \times 2 = 259 \text{ fJ} \)

![Diagram showing DAC Driver, Memory Row, and Input Driver](image)

**OP-Amp for SAR ADCs**

- **100 MHz 8-bit SAR ADC**
  - \( E_{ADC} = 3.56 \text{ pJ} \)
  - \( I_{D1} = 0.6 \times 256 = 154 \mu \text{A} \)
  - \( E_{D1} = I_{D1} V_{DD} t_{sample} \)
  - \( = 154 \times 1.2 \times 1 = 0.19 \mu \text{J} \)
  - \( I_{D2} = \frac{0.6 V}{0.5 \mu \text{s}} \times 256 = 307 \mu \text{A} \)
  - \( E_{D2} = I_{D2} V_{DD} t_{SAR} \)
  - \( = 307 \times 1.2 \times 8 = 2.9 \mu \text{J} \)
Wishlist and Trade-offs in Designing SRAM PIM Circuits

- Parallelism (turning on more rows at a time)
  - ADC energy amortization
  - higher throughput

- Storage Density
  - Compact memory cells and smaller computing/ADC circuits

- Computing Accuracy
  - Suppressing non-linearity, variations, noise
  - Considering information loss due to ADCs through SW/HW co-design

- Energy Efficiency
  - Analog driver free
  - Area- and energy-efficient ADCs
CAP-RAM: Charge-Domain Computing in 6T SRAM

- Average cell area is only 30% larger than 6T
- Only one cell in the cluster is activated at a time
- 4-bit input sampled on the MOM capacitor
- Charge-sharing accumulation

Charge-Injection SAR ADC

- Long channel length device behaves like capacitors (K. Choo et al., ISSCC 2016)
- Reuse local MOM cap as input cap (no sampling)
- Small area (429 um²)
CAP-RAM Macro in 65nm LP CMOS

- Superior storage density
- Excellent accuracy
- High compute density
- High energy efficiency
- Scalable configurations
CAP-RAM: Computing Accuracy

Simulated current domain IMC (ideal ADC)

Measured results
CAP-RAM: Computing Linearity and Noise

- INL within $\pm2.5/-2.8$ LSB
- Average RMS error under random noise is 0.35 LSB (bounded by 0.5 LSB)
DCT-RAM: Driver-Free Fully-Parallel Capacitive PIM

576×128 8T SRAM PIM with row-parallel driver-free computing and column-parallel TD-ADCs

Z. Chen et al., “DCT-RAM: A Driver-Free Process-In-Memory 8T SRAM Macro with Multi-Bit Charge-Domain Computation and Time-Domain Quantization,” in 2022 IEEE CICC.
Idea 1: Driver-Free Charge-Domain Computing (DCC)

Two-Phase Operation: Sample and Sharing

- No transistor non-idealities
- Shared switch enables simple cell design
- No analog buffers: **sampled** input and **constant** DAC output impedance
Idea 1: Driver-Free Charge-Domain Computing (DCC)

Two-Phase Operation: Sample and Sharing

- Constant impedance
- No transistor non-idealities
- Shared switch enables simple cell design
- No analog buffers: sampled input and constant DAC output impedance

65 nm LP Logic Process

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Area</td>
<td>1.50 μm² (355 F²)</td>
</tr>
<tr>
<td>MOM Capacitance</td>
<td>0.9 fF</td>
</tr>
</tbody>
</table>
DCC Phase 1: Pre-Charge

- Pre-charge input lines (IN) output lines (OL) and local MOM caps
Dual input lines ensure constant impedance.
Current DAC requires no analog buffer.
S5 Reduces settling time.
DCC Phase 3: Multiplication

- Cells with logic ‘1’ will keep $V_{DAC}$
- Cells with logic ‘0’ will be pulled up to $V_{DD}$
DCC Phase 4: Accumulation

- Charge is shared on bottom plates
- Local MOM capacitors are further reused as the input capacitor for ADCs
Idea2: Time-Domain ADC (TD-ADC) Array

Driver-Free

Highly Digital

VTC

Flash TDC

From Array

V_{TH} \rightarrow

TD-ADC

>: TD-ADC

Fast and efficient for moderate-resolution ADCs

No reference drivers required

Highly digital and scalable
Idea2: Time-Domain ADC (TD-ADC) Array

- Fast and efficient for moderate-resolution ADCs
- No reference drivers required
- Highly digital and scalable
- Area of DFFs increases exponentially
Idea2: Time-Domain ADC (TD-ADC) Array

- Fast and efficient for moderate-resolution ADCs
- No reference drivers required
- Highly digital, scalable and reduced area
Idea2: Time-Domain ADC (TD-ADC) Array

- Fast and efficient for moderate-resolution ADCs
- No reference drivers required
- Highly digital, scalable and reduced area
- RO with small phase noise is area- and power-hungry
Idea 2: Time-Domain ADC (TD-ADC) Array

- Fast and efficient for moderate-resolution ADCs
- No reference drivers required
- Highly digital, scalable and optimal area
- Shared RO for less area, power and small phase noise
Idea 2: Time-Domain ADC (TD-ADC) Array

- Shared RO
  - Reduced area
  - Upsized stage buffers (less variation)

- Reused local MOM capacitor
  - No S&H and input analog buffers

- Column parallel local ADC
  - Only 24% of column area

- Self-referenced structure
  - No reference buffers
  - Tunable gain

- Safe stop of counters
DCT-RAM Prototype in 65nm LP CMOS

**Area Breakdown**
- Array: 57.7%
- DCC: 8.4%
- ADC: 23.4%
- Others: 10.6%

**Chip Summary**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>65 nm</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>0.8-1.0V</td>
</tr>
<tr>
<td>Clock Frequency</td>
<td>26-40 MHz</td>
</tr>
<tr>
<td>Active Area</td>
<td>0.23 mm²</td>
</tr>
<tr>
<td>Memory Capacity</td>
<td>576x128</td>
</tr>
<tr>
<td>Cell Structure</td>
<td>8T1C</td>
</tr>
<tr>
<td>Cell Area</td>
<td>1.5 μm² (355 F²)</td>
</tr>
</tbody>
</table>
Measurements: PIM Linearity of One Column

• The standard deviation of PIM errors (including ADC) is 0.69 LSB
Measurements: Linearity Across Multiple Columns

- Gain is tunable by tuning the discharging current of VTC
- Variation of the output across different columns is 0.59 LSB/1.19 LSB when Gain = 1 or 2
Measurements: System Accuracy

- Inference accuracy changes with different configurations
- Achieve 69.5% and 94.0% on CIFAR-10 and CIFAR-100

* 4-bit activation/4-bit weight standard ResNet-20
† 4-bit activation/4-bit VGG-like network with 11 layers (H. Jia, JSSC 2021)
## Comparison with State-of-the-Arts

<table>
<thead>
<tr>
<th></th>
<th>Proposed</th>
<th>Chen, JSSCC’21</th>
<th>Jia, VLSI’21</th>
<th>Jiang, JSSC’20</th>
<th>Si, ISSCC’20</th>
<th>Dong, ISSCC’20</th>
<th>Kim, JSSC’21</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>65 nm</td>
<td>65 nm</td>
<td>28 nm</td>
<td>65 nm</td>
<td>28 nm</td>
<td>7 nm</td>
<td>65 nm</td>
</tr>
<tr>
<td><strong>Memory Capacity</strong></td>
<td>576x128</td>
<td>512x128</td>
<td>1152x256</td>
<td>256x64</td>
<td>512x128</td>
<td>64x64</td>
<td>128x128</td>
</tr>
<tr>
<td><strong>Analog MAC Precision</strong></td>
<td>4-8</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>ADC Resolution</strong></td>
<td>4-8</td>
<td>7</td>
<td>8</td>
<td>11 levels</td>
<td>5</td>
<td>4</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Cell Structure</strong></td>
<td>8T-1MOM</td>
<td>6.6T (Shared)</td>
<td>10T-1MOM</td>
<td>8T-1MOS</td>
<td>6.375T(Shared)</td>
<td>8T</td>
<td>6T + Digital</td>
</tr>
<tr>
<td><strong>Driver Free</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Computing Mechanism</strong></td>
<td>Charge sharing</td>
<td>Charge sharing</td>
<td>Charge sharing</td>
<td>Coupling</td>
<td>Current</td>
<td>Current</td>
<td>Digital</td>
</tr>
<tr>
<td><strong>Computing Parallelism</strong></td>
<td>576 rows</td>
<td>64 rows</td>
<td>1152 rows</td>
<td>256 rows</td>
<td>32 rows</td>
<td>64 rows</td>
<td>128 rows</td>
</tr>
<tr>
<td><strong>Throughput (GOPS)</strong></td>
<td><strong>3894-6036</strong></td>
<td>573.4</td>
<td>1229</td>
<td>1638</td>
<td>999</td>
<td>372.4-455.1</td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td><strong>bOPs</strong></td>
<td>15576-24144</td>
<td>4587.2</td>
<td>6144</td>
<td>1638</td>
<td>1998</td>
<td>5958-7281</td>
</tr>
<tr>
<td><strong>Compute Density (TOPS/mm²)</strong></td>
<td>OPs</td>
<td>16.9-26.4</td>
<td>3.4</td>
<td>2.4</td>
<td>20.4</td>
<td>N/A</td>
<td>116-142</td>
</tr>
<tr>
<td></td>
<td>bOPs</td>
<td>67.7-105.0</td>
<td>27.2</td>
<td>12</td>
<td>20.4</td>
<td>N/A</td>
<td>1856-2272</td>
</tr>
<tr>
<td><strong>Energy Efficiency (TOPS/W)</strong></td>
<td>OPs</td>
<td>227-322</td>
<td>49.4</td>
<td>1159</td>
<td>671.5</td>
<td>47.85-68.44</td>
<td>189.3-610.5</td>
</tr>
<tr>
<td></td>
<td>bOPs</td>
<td>908-1288</td>
<td>319.2</td>
<td>5796</td>
<td>671.5</td>
<td>766-1069</td>
<td>3029-9768</td>
</tr>
<tr>
<td><strong>CNN Model</strong></td>
<td></td>
<td>ResNet20 a:4b, w:4b</td>
<td>VGG11 a:4b, w:4b</td>
<td>ResNet20 a:4b, w:4b</td>
<td>VGG14 a:5b, w:5b</td>
<td>VGG11 a:2b, w:2b</td>
<td>ResNet20 a:4b, w:8b</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td></td>
<td>88.3-90.0%</td>
<td>90.5-94.0%</td>
<td>89.0%</td>
<td>91.1%</td>
<td>85.5%</td>
<td>91.9%</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>N/A</td>
<td>60.1-69.5%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>67.6%</td>
</tr>
</tbody>
</table>

*1 OP = 1 multiplication or 1 addition
**bOPs = OPs × input bitwidth × weight bitwidth
Outline of the talk

SRAM Based PIM

- CAP-RAM
- DCT-RAM

RRAM Based PIM

- NeuRRAM
  - 48 core
  - 3M RRAMs
  - 12K neurons

PIM SW/HW Co-Design

- Silicon IP
- PIM-net
- SoC Architecture
- PIM aware training
RRAM: An Emerging Non-Volatile Memory

Resistive Switching Random-Access Memory

“0”
High resistance

“1”
Low resistance

Single-Level-Cell (SLC) commercially available (22 nm); Multi-Level-Cell (MLC) in development

RRAM vs. SRAM

**Dense**

<table>
<thead>
<tr>
<th></th>
<th>SRAM</th>
<th>RRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSMC 40 nm</td>
<td>0.242</td>
<td>0.085</td>
</tr>
<tr>
<td>TSMC 22 nm</td>
<td>0.081</td>
<td>0.026</td>
</tr>
</tbody>
</table>

2.8 X

↑3.1 X

**Analog Programmable**

- Store bigger model on chip
- Non-volatile
- Zero static power consumption
Kirchoff’s Law Based Matrix-Vector Multiplication

\[
\begin{pmatrix}
V_1 & \cdots & V_m
\end{pmatrix}
\begin{pmatrix}
G_{11} & \cdots & G_{1n}
\vdots & \ddots & \vdots
G_{m1} & \cdots & G_{mn}
\end{pmatrix}
= \begin{pmatrix}
I_1 \\
\vdots \\
I_n
\end{pmatrix}
\]

Each RRAM cell represents a multi-bit weight

Note: Output can also be voltage or charge
Lots of Progress, But...

- Few work simultaneously addressed **efficiency**, **versatility**, and **accuracy**
- Results were **partially simulated** (rather than fully measured)
- Demonstrated tasks lacked **complexity** and **diversity**
A compute-in-memory chip based on resistive random-access memory

Weier Wan, Rajkumar Kubendran, Clemens Schaefer, Sukru Burc Eryilmaz, Wengiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, Huaqiang Wu, H.-S. Philip Wong & Gert Cauwenberghs
NeuRRAM: 48-Core RRAM-PIM Chip

Photos by David Baillot/University of California San Diego
RRAM-CMOS Monolithic Integration

- CMOS & M1-M4 fabricated using a commercial foundry 130 nm process
- RRAM & M5 integrated using a research lab process
Challenges for RRAM-PIM Chip Design

**Fundamental Design Trade-off**

**Energy-Efficiency**
Limited by peripheral circuits (e.g. ADCs)

**Reconfigurability**
Re-programming weights is expensive

**Inference Accuracy**
Analog computation susceptible to device & circuit imperfections
Full-Stack Co-Design

Algorithm

System

Architecture

Circuit

Device

48 core
3M RRAMs
12K neurons

W. Wan et al., Nature 2022

ENERGY-EFFICIENCY
Highest among RRAM PIM chips

RECONFIGURABILITY
CNN, MLP, LSTM, RBM, etc.

INFEERENCE ACCURACY
Comparable to software model with 4-bit weights
Full-Stack Co-Design

Algorithm
Hardware-aware training & fine-tuning

System
Multi-core parallel operation

Architecture
Transposable neurosynaptic array

Circuit
Voltage-mode neuron with variable precision

Device
Analog programmable RRAM

48 core
3M RRAMs
12K neurons

ENERGY-EFFICIENCY
Highest among RRAM PIM chips

RECONFIGURABILITY
CNN, MLP, LSTM, RBM, etc.

INFERENCCE ACCURACY
Comparable to software model with 4-bit weights

W. Wan et al., Nature 2022
Record Energy-Delay-Product (EDP)

*Measured Across Various Computational Bit-Precisions*

![Graph showing energy-delay-product measured across various computational bit-precisions.]
Software Comparable Accuracy

All Results Obtained From Hardware Measurements
Takeaways

**SRAM Based PIM**

- **CAP-RAM**
  - SRAM Read/Write
  - ADC
  - DAC
  - ADC CLX & SAR Logic

- **DCT-RAM**
  - SRAM Read/Write
  - ADC
  - DAC
  - 576x128 SRAM
  - One ADC per 576x128 SRAM

**RRAM Based PIM**

- **NeuRRAM**
  - 48 core
  - 3M RRAMs
  - 12K neurons

**Chip Summary**

- **ADC** (23.4%)
- **Others** (10.6%)
- **DCC** (8.4%)

**Array Breakdown**

- **Array** (57.7%)
- **8T SRAM**
- **DCC**
- **ADC**

**Takeaways**

- High energy efficiency
- High compute density
- Mature & scalable CMOS process
- Ready for production

- Higher compute density
- Lower static power
- CMOS compatible, SLC @ 22nm
- Promising for the future
Outline of the talk

SRAM Based PIM

RRAM Based PIM

PIM SW/HW Co-Design

Silicon IP

PIM-net

SoC Architecture

PIM aware training

SRAM Based PIM

DCT-RAM

CAP-RAM

ADC

Others (10.6%)

DCC (8.4%)

Area Breakdown

Array 57.7%

8T SRAM

DCC

ADC

One ADC per Column

Read/Write WL Drivers

Controller

431 μM

526 μM

Chip Summary

ADC

355

SRAM

Based PIM

NeuRRAM

48 core

3M RRAMs

12K neurons

SRAM Based PIM

RRAM Based PIM

PIM SW/HW Co-Design

Silicon IP

PIM-net

SoC Architecture

PIM aware training
From Test Chips to PIM Products

Key is full-stack software-hardware co-design

Test Chip

Data
- Robust training
- Model architecture
- SoC architecture
- PIM circuits

Product

Hardware aware software

Software defined hardware
Aizip’s Full-Stack PIM Co-Design Services For IC Companies

PIM-NET DNN Model Family
Architecture optimized to achieve high utilization and accuracy on PIM, spanning vision, audio & time-series

PIM Silicon IP
Foundry CMOS process, silicon-verified, compact, efficient & accurate

PIM-Aware Training Framework
Train DNN models to be resilient to hardware non-idealities (e.g. non-linearities, variations, noises)

SoC Architecture Design
End-to-end DNN support, hybrid analog & digital design, superior system-level efficiency
Complete Support Throughout PIM Product Development

IC Product Development and Distribution

- CAP-RAM
  - Sub-License Agreement
  - License Agreement
  - Aizip
  - Rice University

- IC Design
  - Consulting & Design Service
  - Aizip

- TinyML Models
  - Model Design & Deployment
  - Aizip

- Market Promotion
  - Marketing Support
  - Aizip
Complete Support Throughout PIM Product Development

IC Product Development and Distribution

More on this in the upcoming TinyML Asia forum...
Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org