IEEE/ACM ICCAD 2024 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN 43rd Edition

TU Delft
Delft University of Technology

Hardware-Aware Quantization for Accurate Memristor-Based Neural Networks

Sumit Diware, Mohammad Amin Yaldagard & Rajendra Bishnoi

Computer Engineering

Faculty of Electrical Engineering, Mathematics & Computer Science,
Delft University of Technology (TU-Delft), The Netherlands

19th May 2025

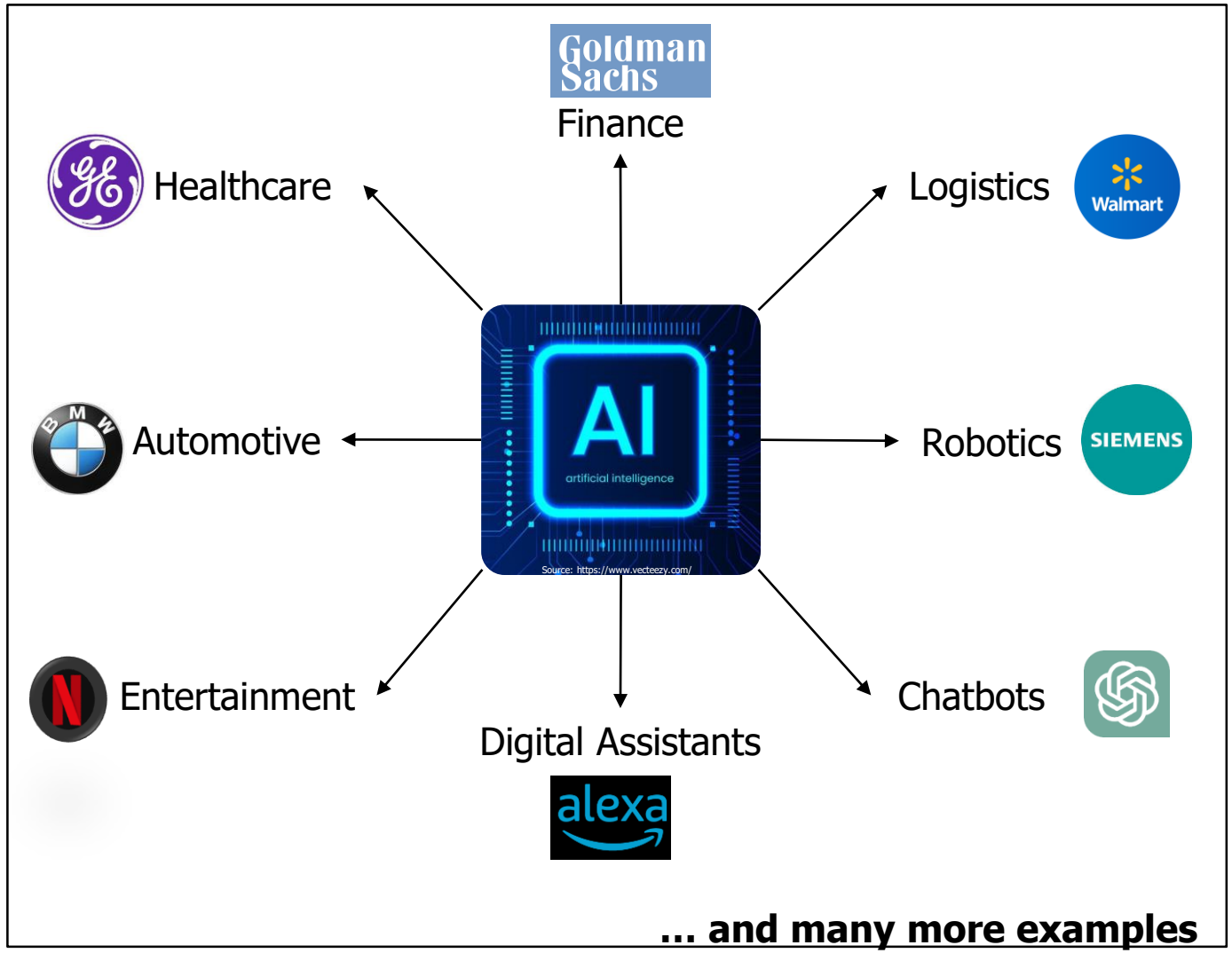
Email: S.S.Diware@tudelft.nl

Agenda

- Introduction
- Computation-In-Memory (CIM) for edge-AI
- Challenge & related-works
- Proposed methodology
- Results
  - Simulation
  - Chip Prototype
- Conclusions

Artificial Intelligence (AI)

- Systems that can perform cognitive tasks



UN trade & development
UNCTAD

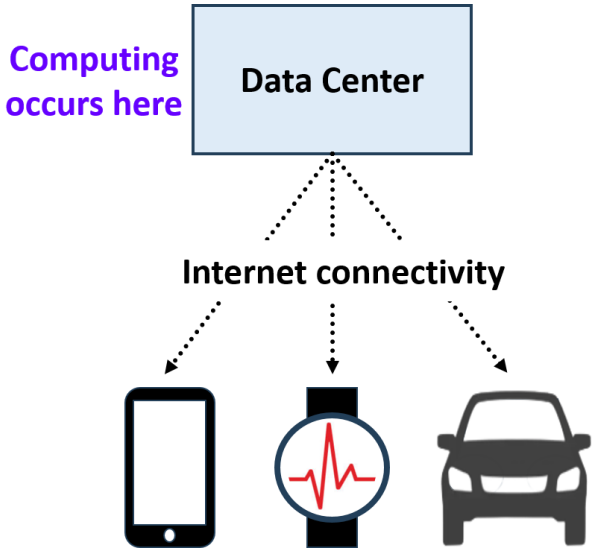
Technology and innovation report 2025

AI market is projected to reach $4.8 trillion by 2033 – almost the size of Germany's economy

Source: https://unctad.org/publication/technology-and-innovation-report-2025

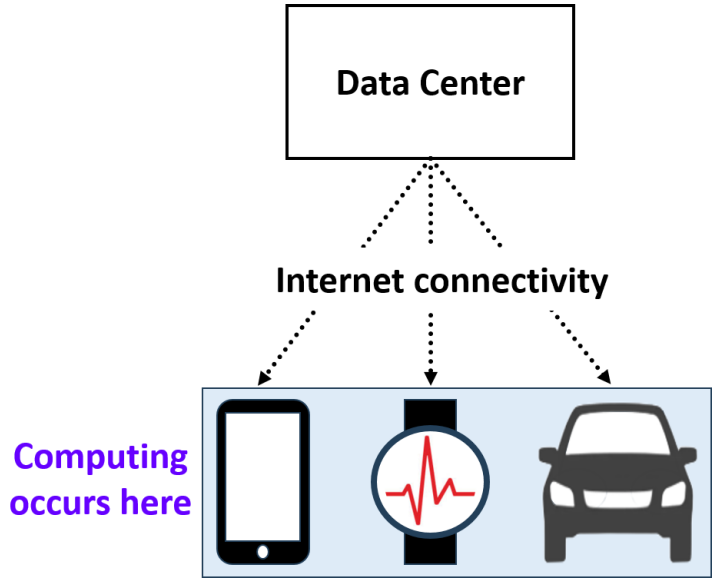
AI Computing Paradigms

Cloud-AI



- Fast response
- Low network costs
- Data privacy & security
- Service reliability

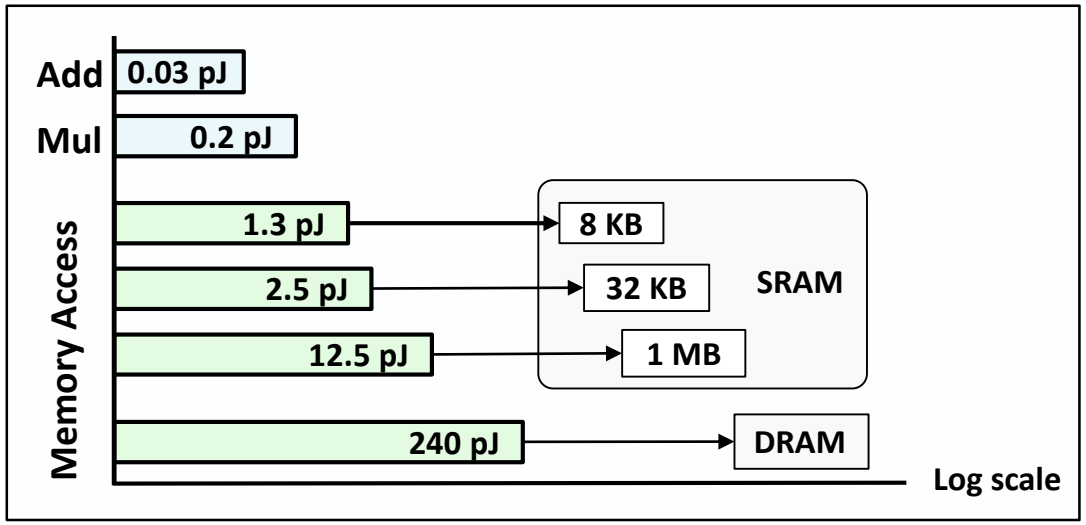
Edge-AI



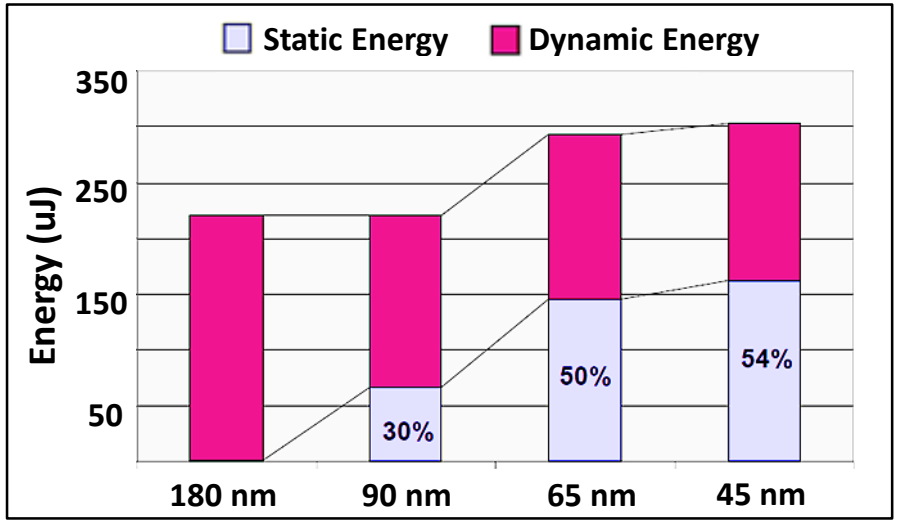
Growing preference shift towards edge-AI

Edge-AI using Conventional Hardware

- The von Neumann architecture
  - Data transfer bottleneck
- Conventional memory technologies
  - Standby energy & scalability issues



Energy for 8-bit arithmetic in 45 nm node [Meng-MNANO'2023]



SRAM energy consumption trend [Goudarzi-HiPEAC'2008]

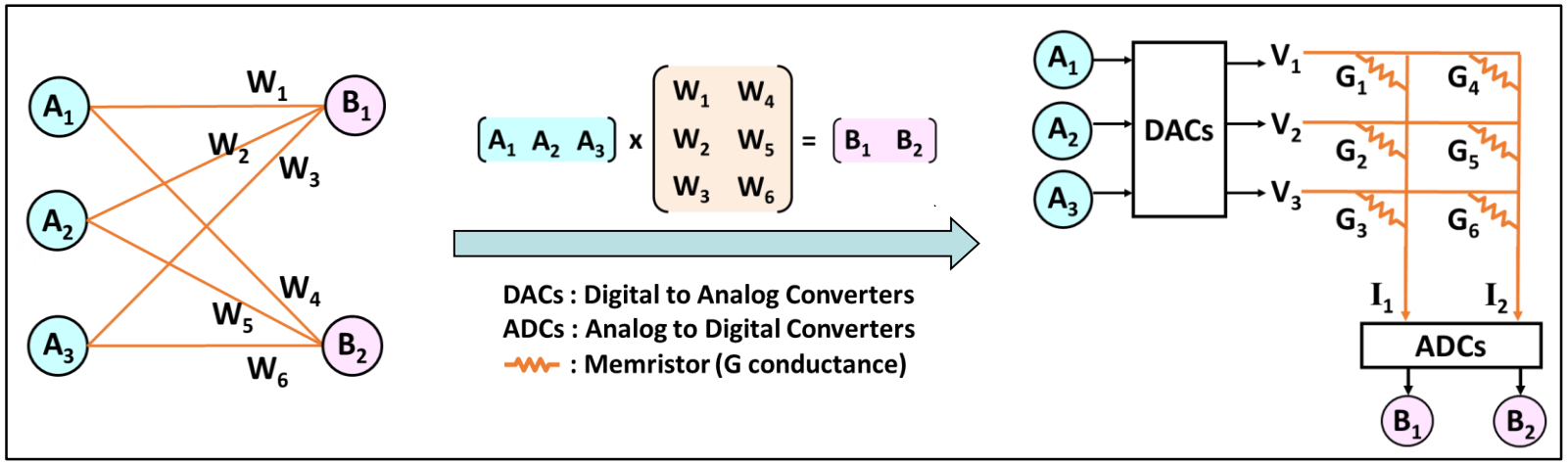
Conventional hardware is not suited for edge-AI

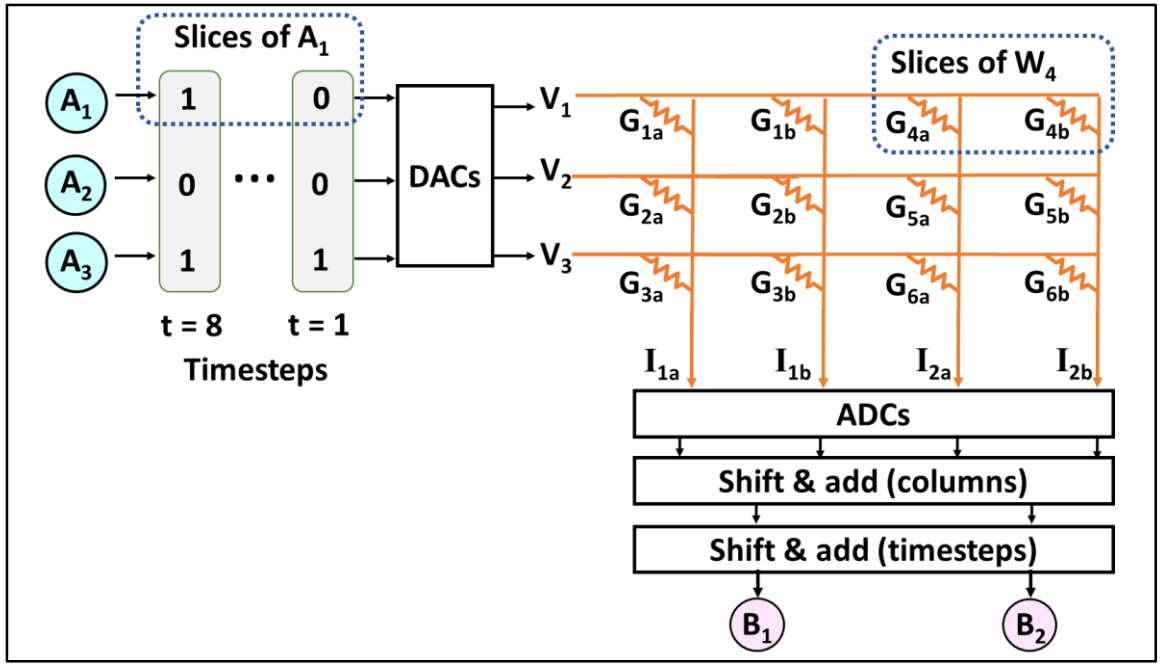
Agenda

- Motivation
- Computation-In-Memory (CIM) for edge-AI
- Challenge & related-works
- Proposed methodology
- Results
  - Simulation
  - Chip Prototype
- Conclusions

Computation-In-Memory (CIM)

- Emerging paradigm
  - In-place computations
  - Emerging memory devices
    - Known as memristors
- Key benefits
  - Energy-efficiency
  - Small area footprint
  - Brain-like architecture
- Bit-slicing
  - ADC/DAC limitation
  - Memristor bits < weight bits

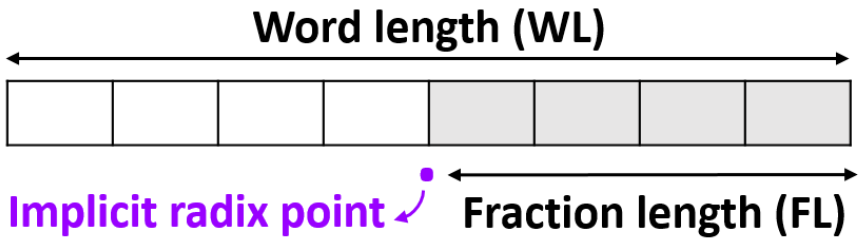




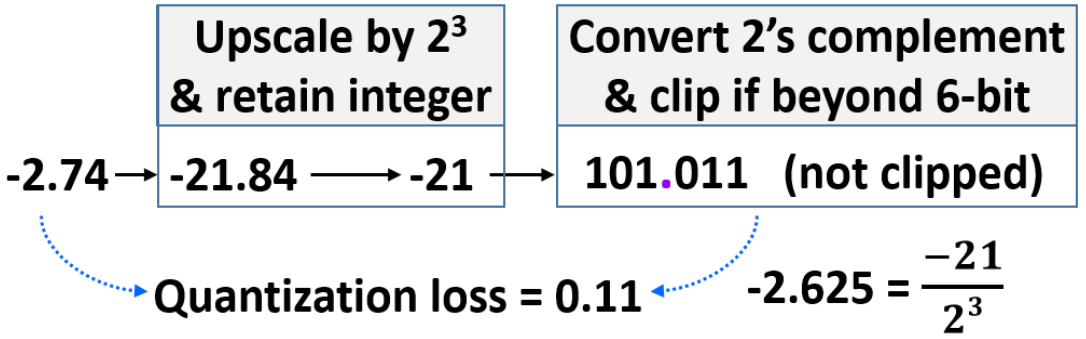
Quantization in CIM

- CIM typically uses fixed-point quantization
  - Fixed-point number = integer with implicit scaling factor
  - Further energy-savings due to integer post-processing

Fixed-point structure

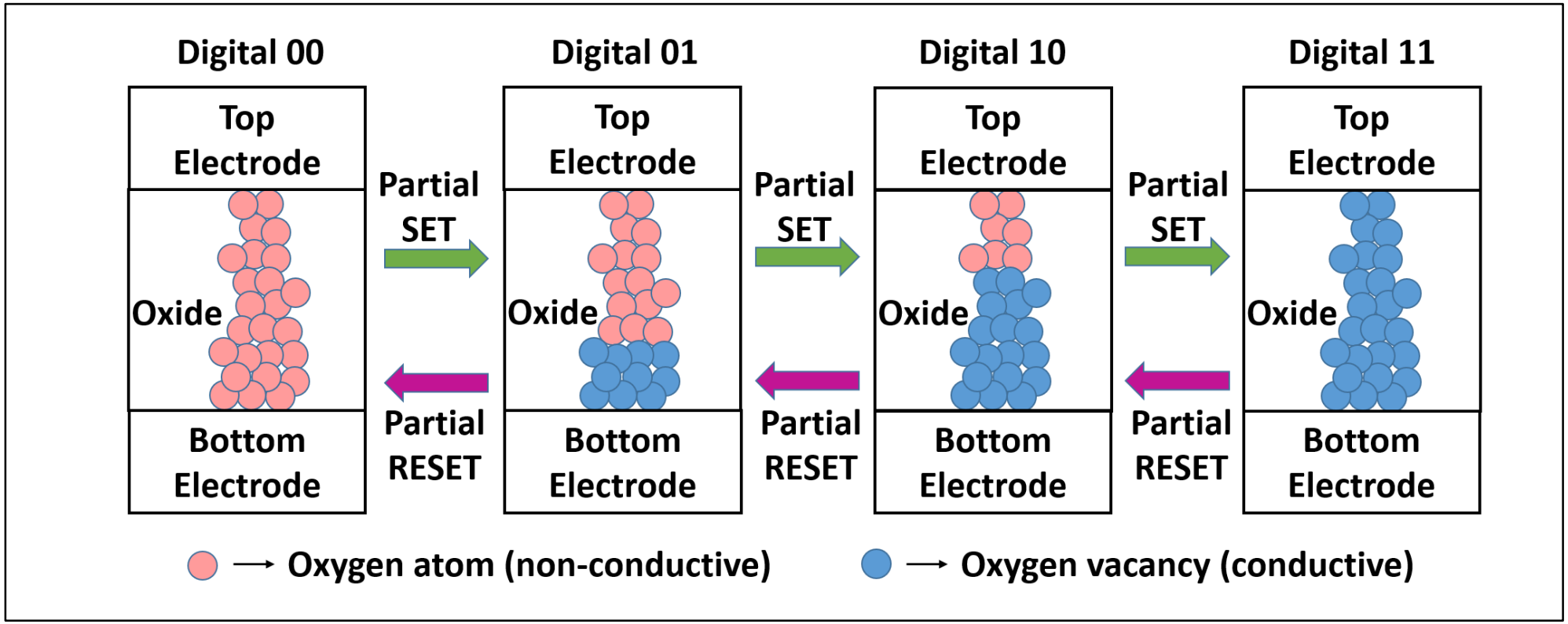


Quantization example (WL=6, FL=3)



Memristor Device Technology

- Resistive random access memory (RRAM)
  - Data stored as oxide conductance
  - Multi-bit storage capability

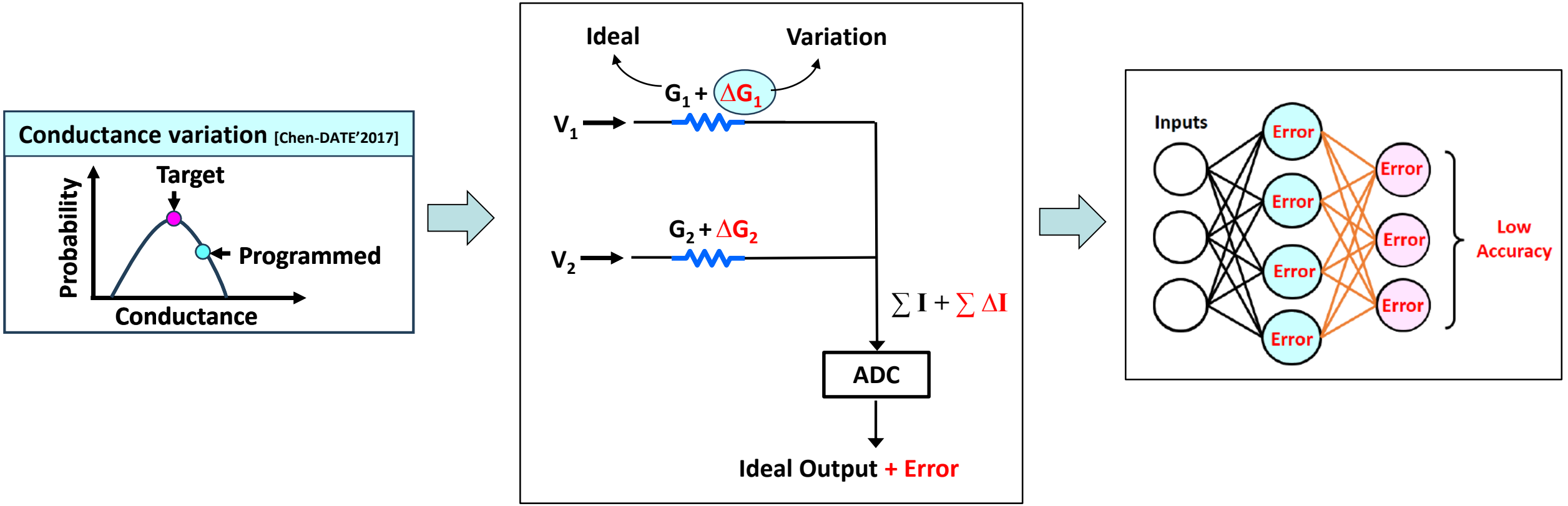


Agenda

- Motivation
- Computation-In-Memory (CIM) for edge-AI
- Challenge & related works
- Proposed methodology
- Results
  - Simulation
  - Chip Prototype
- Conclusions

Challenge: Conductance Variation

- Deviation from ideal resistive behavior



Related Works

- On-chip training [Nandakumar-FrontiersNS'20, Li-NatureCom'2018]
  - Not scalable, energy and endurance issues
- Off-chip training [Charan-JXCDC'20, Jiang-TC'21, Antolini-JESTCS'23]
  - Not scalable, target error tolerance
- Characterization-based mapping [Song-TCAD'21, Chen-DATE'17]
  - Not scalable, restrictive
- Hardware compensation [He-ASPDAC'23, Chang-NSR'22, Milo-IRPS'21]
  - Hardware overhead

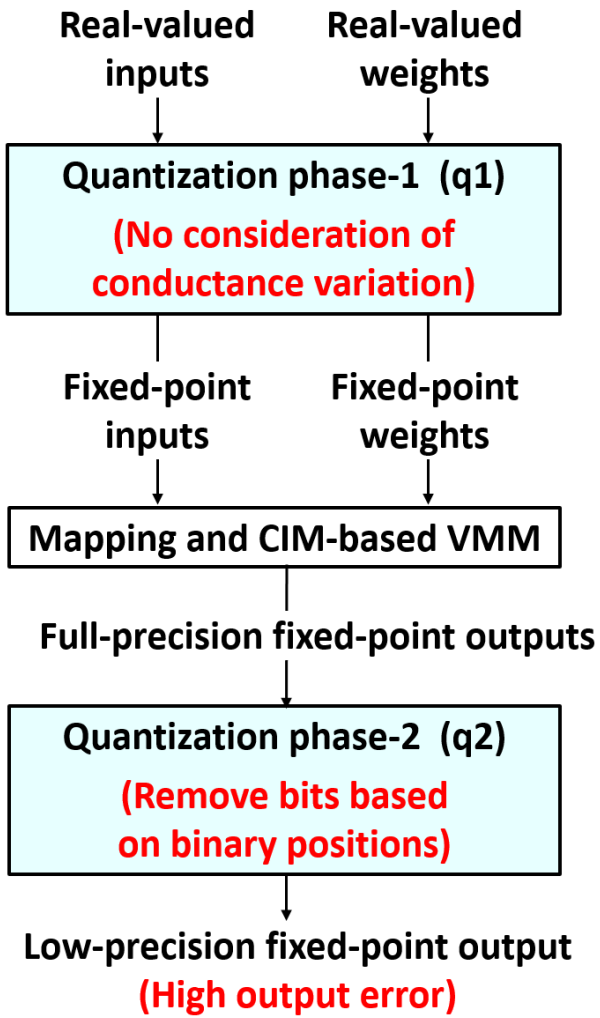
Not scalable, target error tolerance, incur hardware overhead

Agenda

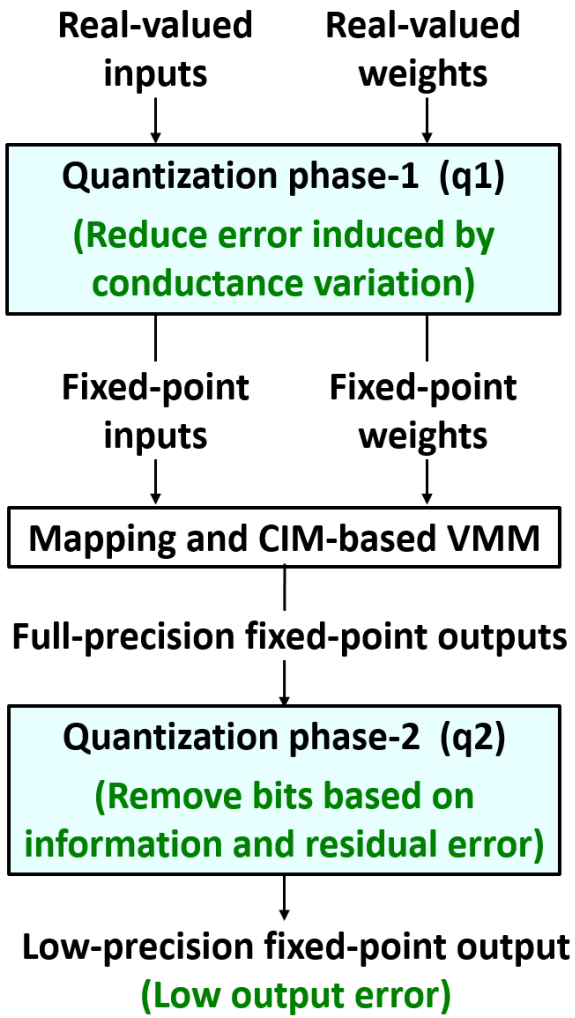
- Motivation
- Computation-In-Memory (CIM) for edge-AI
- Challenge & related works
- Proposed approach
- Results
  - Simulation
  - Chip Prototype
- Conclusions

Overview

Conventional



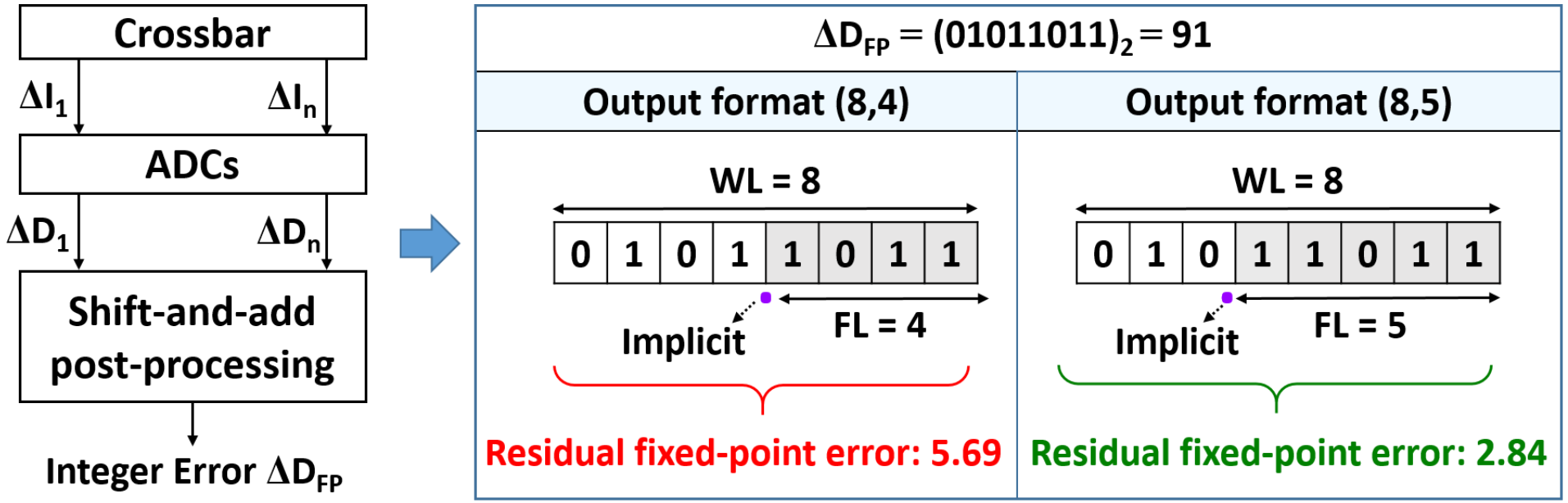
Proposed



Quantization Phase-1

- High output fraction length suppresses conductance variation-induced errors

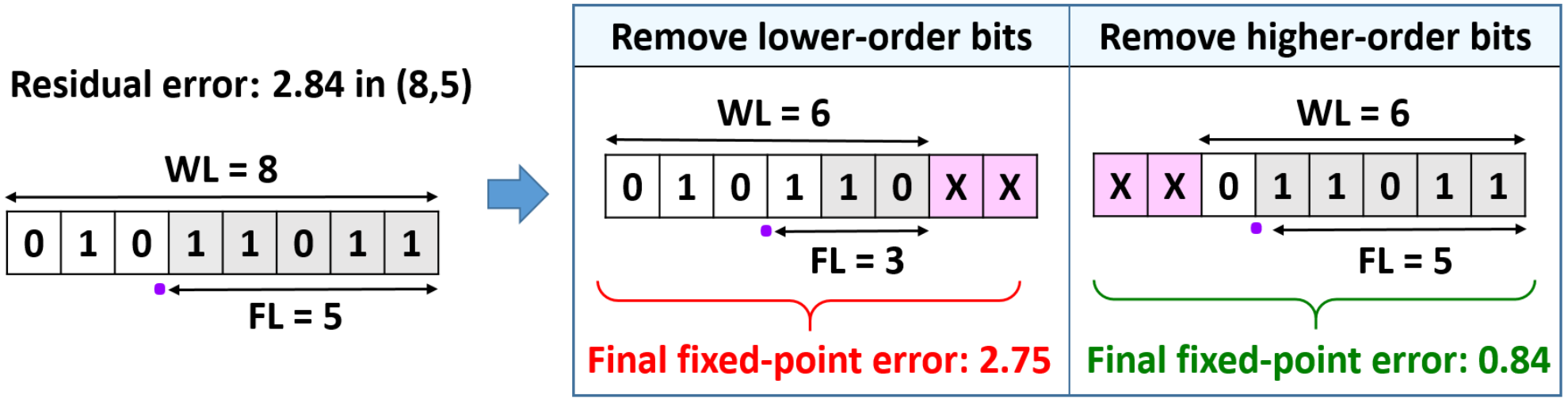
Example



Quantization Phase-2

- Residual error can persist after phase-1

- Discard bits with low information content but high error content
  - These may not always be lower order bits
  - Example: Higher order bits that just exist as sign extension



Quantization Algorithm and Hardware Design

- Proposed quantization algorithm
  - Constraints: architecture details and accuracy threshold
  - Tunes the fraction length of weights
  - Derives the quantization parameters
    - Phase-1: output fraction length
    - Phase-2: distribution of discarded higher and lower order bits

- Hardware design considerations
  - Phase-1: no hardware modifications
  - Phase-2: configurable truncation logic
    - Simple registers and multiplexing logic
    - Can be adapted for any workload

Agenda

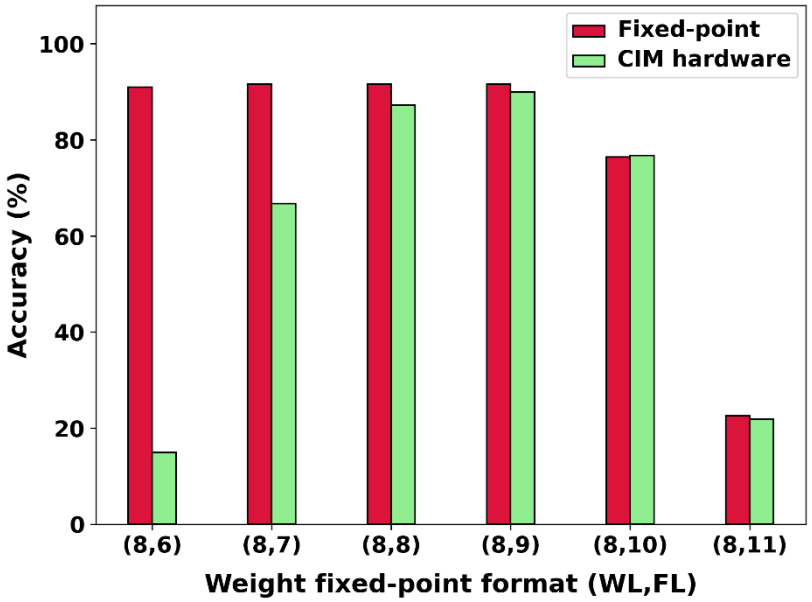
- Motivation
- Computation-In-Memory (CIM) for edge-AI
- Challenge & related works
- Proposed methodology
- Results
  - Simulation
  - Chip Prototype
- Conclusions

Simulation: Setup

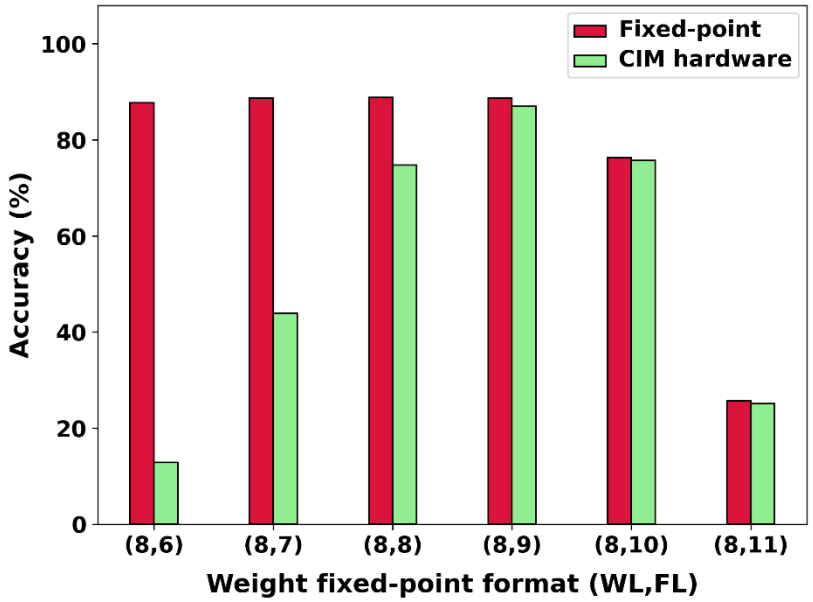
Parameters	Specification/Source
CIM hardware simulation	• Architecture: [ISAAC-ISCA'16] • 8-bit weights, 2-bit memristors • Variation data: [Prakash-PSR'16]
Benchmarks	• Modified Alexnet + SVHN dataset • Modified VGG + CIFAR-10 dataset • Modified ResNet + CIFAR-100 dataset
Evaluation toolchain	• Software training: PyTorch • Hardware inference: In-house CIM simulator • ASIC synthesis: Cadence Genus (TSMC 40nm)

Simulation: Neural Network Accuracy

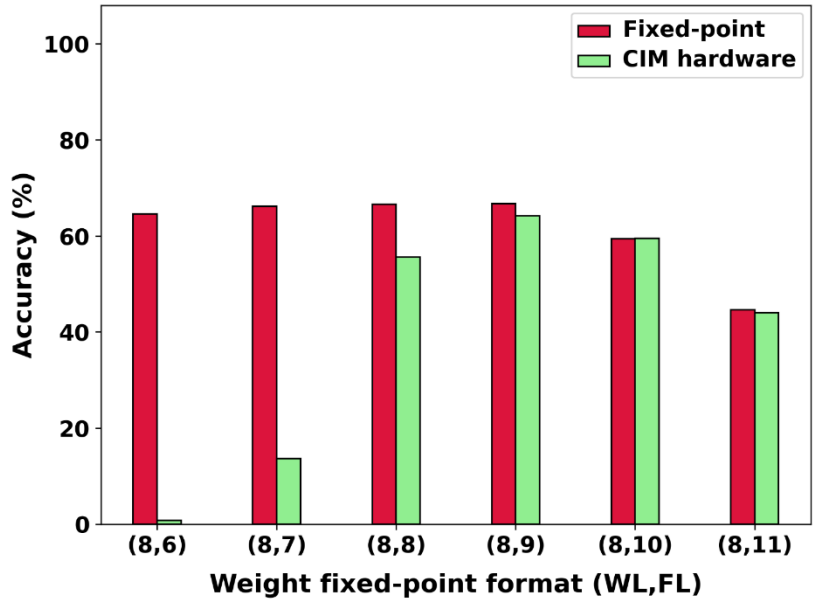
Alexnet + SVHN



VGG + CIFAR-10



ResNet + CIFAR-100



- FL > WL indicates leading implicit zeros in the fraction
- Hardware accuracy peaks at FL = 9 bits
  - FL increase 6 → 9 leads to reduced impact of errors
  - FL increase beyond 9 insufficient integer part precision

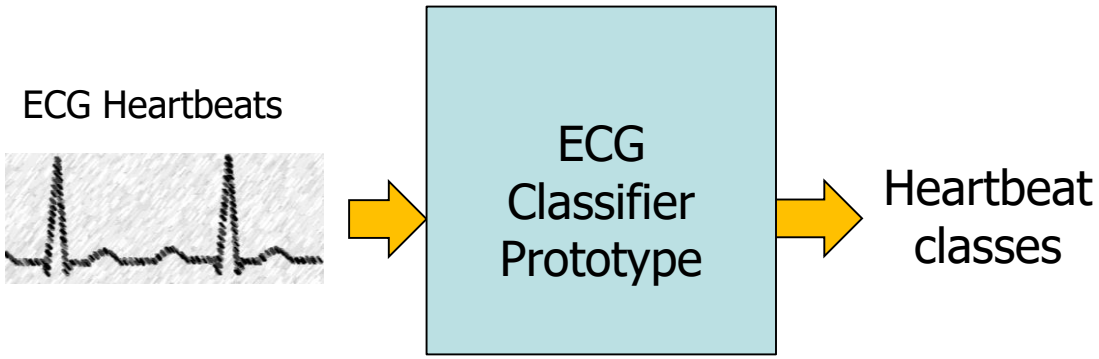
Simulation: Hardware Metrics

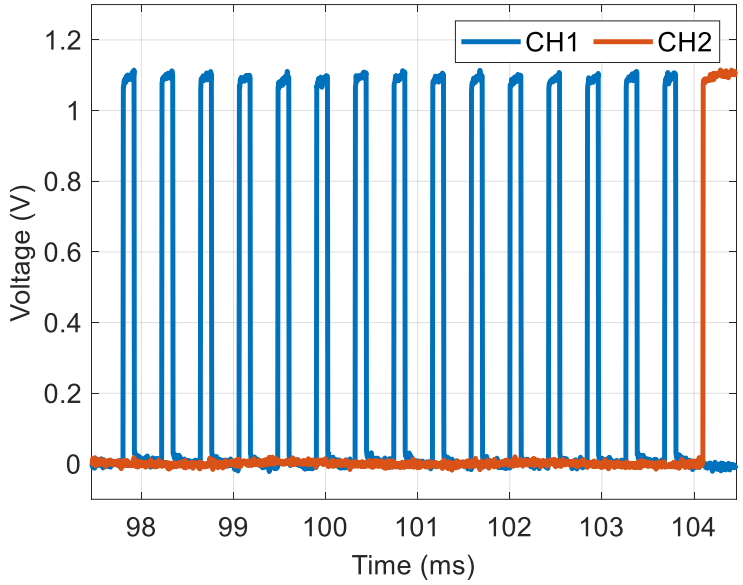
Metric (unit)	Conventional Approach [PANTHER-TC'20, PUMA-ASPLOS'19, ISAAC-ISCA'16]	Proposed Approach
SVHN hardware accuracy (%)	15.94	89.97
Energy (pJ)	3738	3782
Area (μm^2)	21765	23137
Correct operations per unit energy (GOP/J)*	43.7	243.6

$$*\text{Correct operations per unit energy} = \frac{\text{Accuracy} \times \text{Total operations}}{100 \times \text{Energy consumption}}$$

- Up to 5.6× correct operations per unit energy
- Overheads: 1.2% energy and 6.3% area

Chip Prototype





- TSMC 40nm technology
- 2.9 sq. mm. Si area
- 100MHz clock frequency
- 1.1V nominal voltage

Agenda

- Motivation
- Computation-In-Memory (CIM) for edge-AI
- Challenge & related works
- Proposed methodology
- Results
  - Simulation
  - Chip Prototype
- Conclusions

Conclusions

- Computation-in-Memory (CIM) for edge-AI
  - Efficiency beyond von-Neumann computing
- Memristor conductance variation
  - Induces computational errors, impacting the hardware accuracy
- Propose quantization methodology
  - Tunes the output fraction length to suppress the errors
  - Reduces the residual error by discarding bits with high error but less information
- Results
  - 5.6× correct operations per unit energy compared to the conventional approach
  - Deployed the proposed quantization in a chip prototype

Thank You!