A 10.60 μW 150 GOPS Mixed-Bit-Width Sparse CNN Accelerator for Life-Threatening Ventricular Arrhythmia Detection

Yifan Qin[1], Zhenge Jia[1], Zheyu Yan[1], Jay Mok[2], Manto Yung[2], Yu Liu[2], Xuejiao Liu[2], Wujie Wen[3], Luhong Liang[2], Kwang-Ting Tim Cheng[2], Xiaobo Sharon Hu[1], Yiyu Shi[1]

[1]University of Notre Dame, [2]Hong Kong University of Science and Technology/AI Chip Center for Emerging Smart System, [3]North Carolina State University

Speaker: Prof. Yiyu Shi

University of Notre Dame

UNIVERSITY OF NOTRE DAME

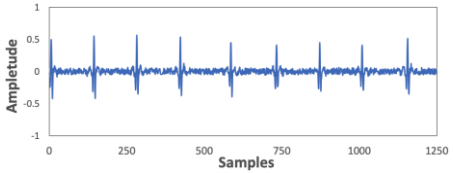
ASP-DAC 2025

ACCESS
AI Chip Center for Emerging Smart Systems
智能晶片與系統研發中心

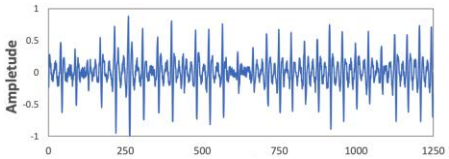
- **Background**
- **Demonstration**
- **Design Highlights**

Background

- Ventricular arrhythmia (VA)



Healthy



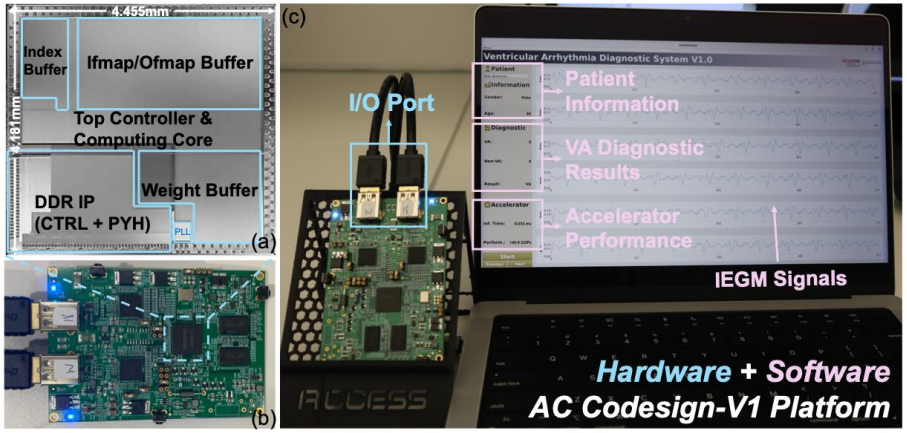
VA-detected

- On-device detection

Data acquisition and detection

Low energy, small chip area, high throughput and speed

Demonstration

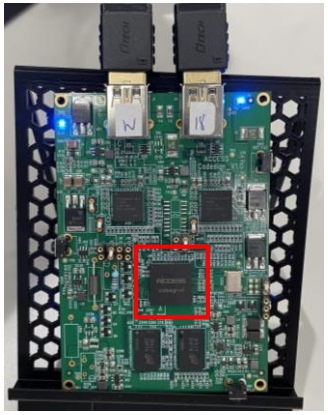


On-Board Testing:

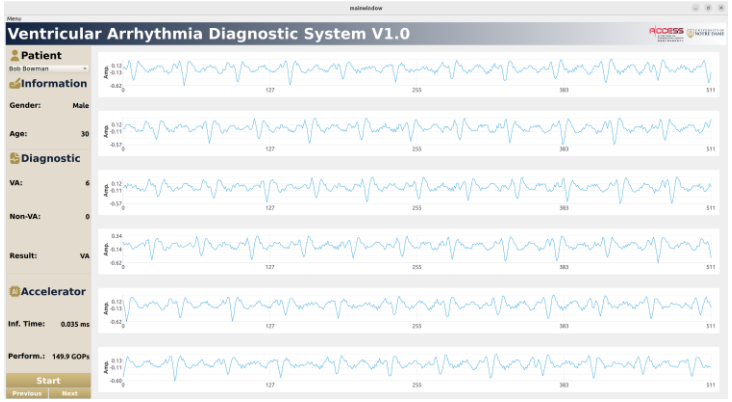
CNN sparsity:	50%
Diagnostic accuracy:	99.95%
Power:	10.60 μW
Throughput:	150 GOPS
Power density:	0.57 μW/mm^2

X 14.23 power density compare with SOTA

ICD
250 Hz sample rate



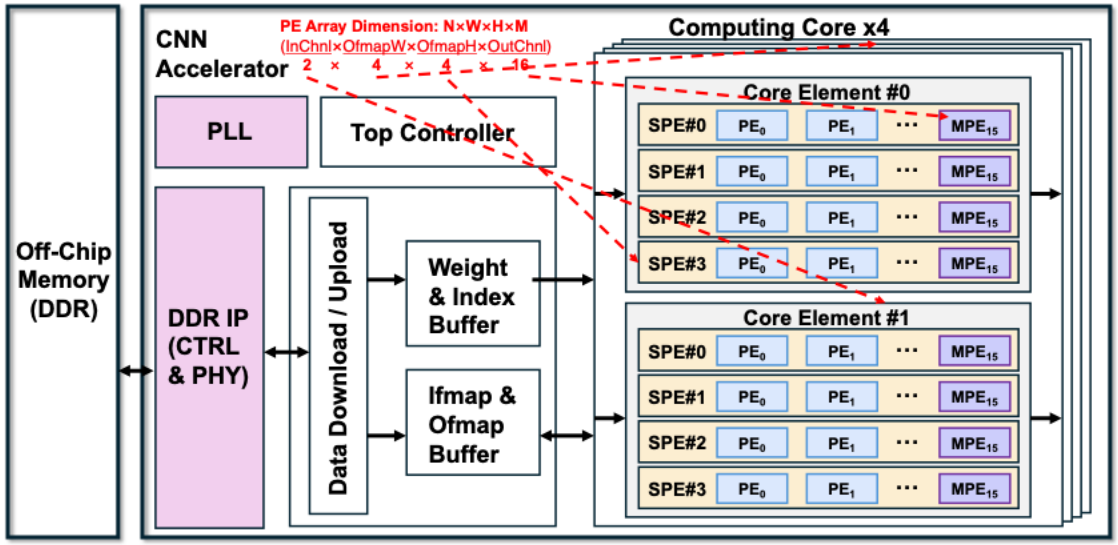
Accelerator
0.035 ms/sample
Quantized CNN



UI for user
6 samples vote for diagnostic result

Design Highlights

- Chip Structure



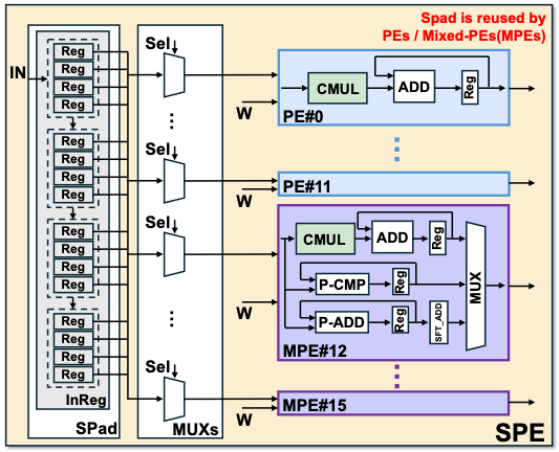
- **Parallel Computing**

2 input channels:
2 Core Elements
16 output channels:
16 PEs per Sparse PE(SPE)
4@4 output feature map W@H:
4 Computing Cores
4 SPEs per Core Element

- Details
  - 512 PEs in total
  - 12 PEs + 4 Mixed-PEs (MPEs) in one SPE
  - Mixed- PE (MPEs): Conv + Pooling OP.
  - 16@4@4 output computing in parallel

Design Highlights

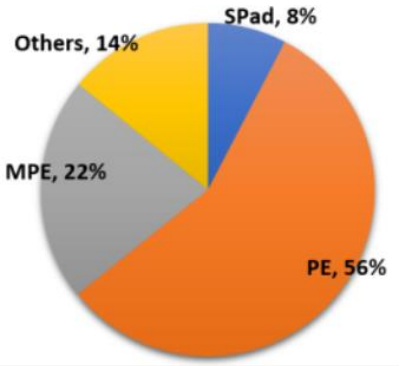
- Sparse Process Element (SPE)



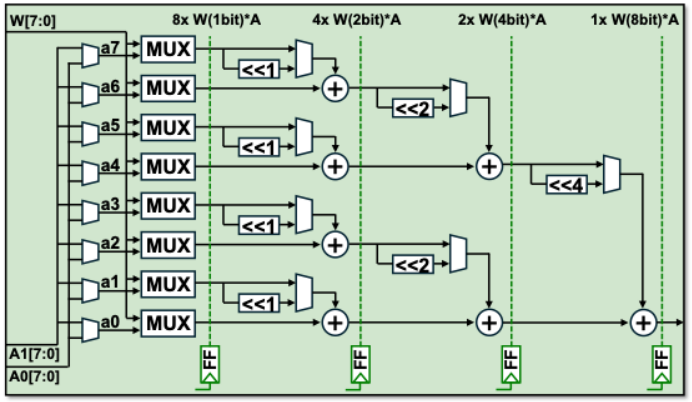
- Single scratch pad (SPad) shares in SPE

Area
Power
Computation density

SPE Area Breakdown



- Mixed-bit Signed Reconfigurable Multiplier (CMUL)



- Reconfigurable 8/4/2/1-bit multiplication

Compression
Sparsity

Thank you !

Further questions please reach out: yshi4 / yqin3 [at] nd.edu