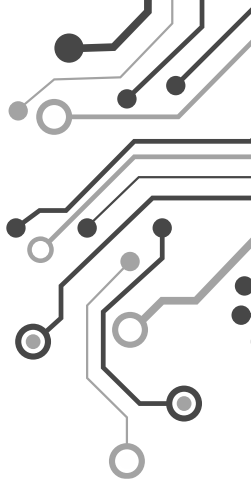
THE UNIVERSITY of EDINBURGH

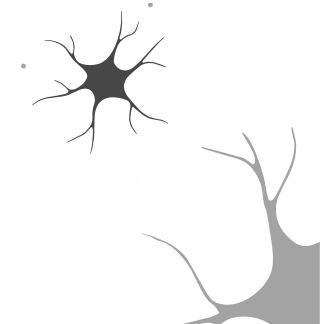


37th IEEE International Conference on Application-specific Systems, Architectures and Processors
September 2-4, 2026, London, UK

BenDi: An Energy-Efficient Quasi-Stochastic Systolic Architecture for Edge Bioelectronics

Bochen Ye, Yihan Pan, Shady Agwa, Themis Prodromakis

Centre for Electronics Frontiers, Institute for Integrated Micro and Nano Systems, School of Engineering, University of Edinburgh



Outline

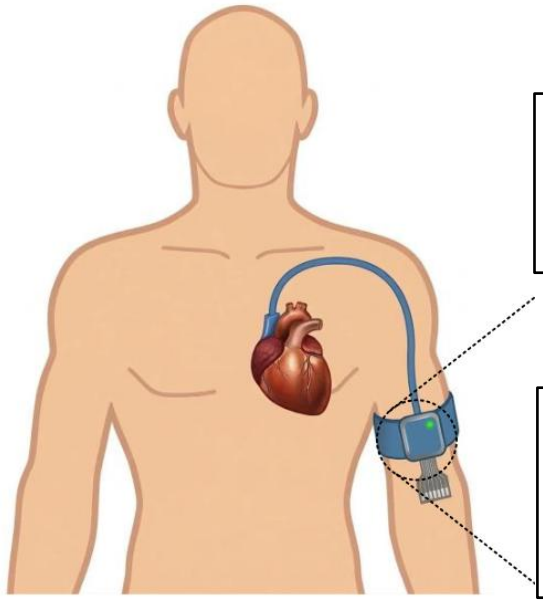
01
Motivation

02
BenDi Architecture

03
Results

04
Conclusion

Long-term Biomedical Monitoring on the Edge



Application

Sleep Apnea

Arrhythmia

Workloads

MAC-intensive CNN

HW Requirement

Long battery life

High portability

Real-time processing

HW Targets

Low energy/power

Small silicon area

Low latency

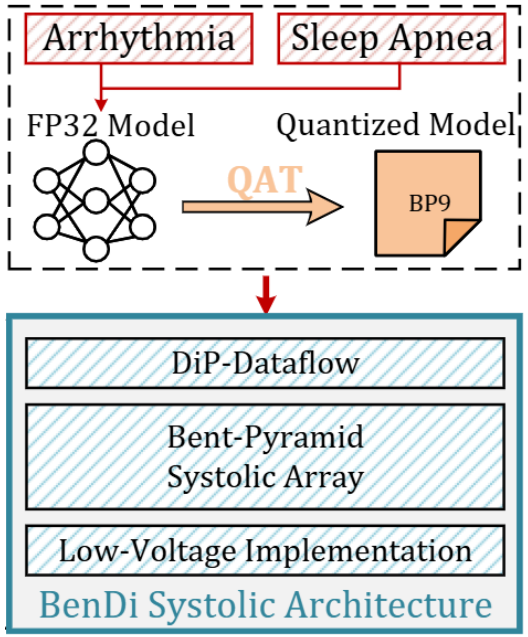
Bottleneck

Multipliers are expensive

Existing computing paradigms

Paradigm	Determinism	Hardware Cost	Latency
Binary	Deterministic ✓	High ✗	Low ✓
Stochastic	Non-deterministic ✗	Very Low ✓	High ✗
(Quasi-Stochastic) **Our Work**	**Deterministic** ✓	**Low** ✓	**Low** ✓

Software-Hardware Co-design Overview



Software level: Hardware-aware QAT (Recovery accuracy loss)

Dataflow level: Diagonal-Input & Permutated weight-stationary (DiP) dataflow[1] (Eliminate synchronization buffer for less power & area)

Computing level: BP9 Quasi-stochastic computing (Simplified multiplication)

Circuit level: Low voltage and frequency in 22nm node (Reduce power)

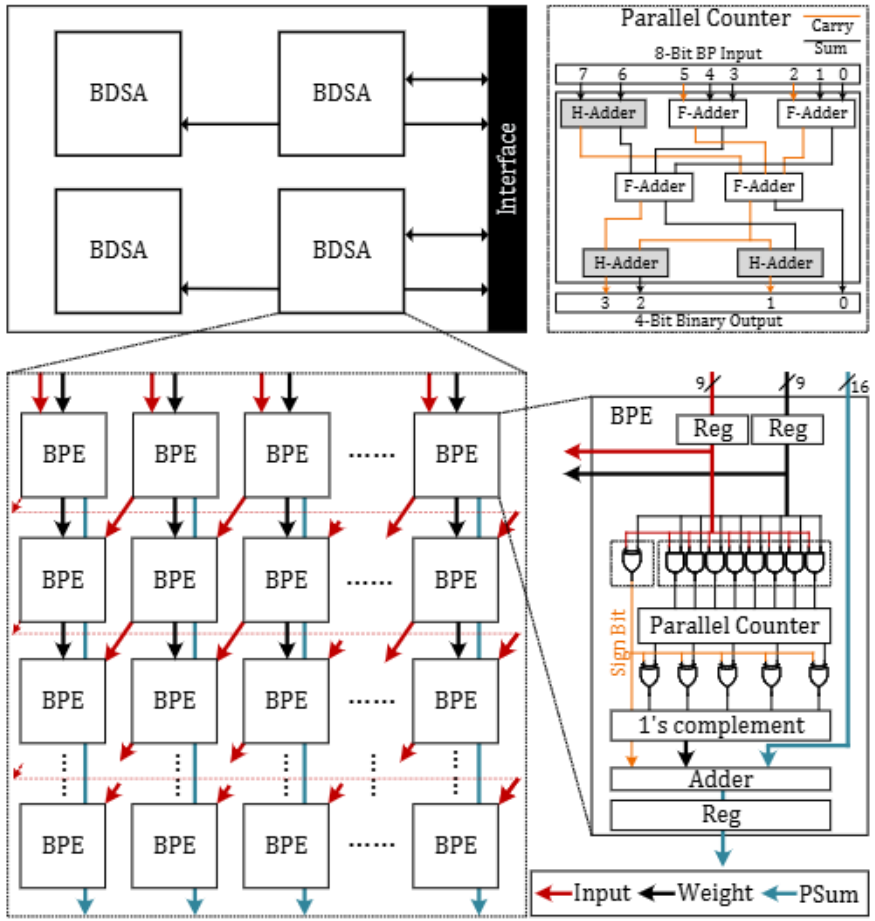
Multi-level co-design for energy- and area-efficient edge AI

[1] A. J. Abdelmaksoud, S. Agwa, and T. Prodromakis, "Dip: A scalable,energy-efficient systolic array for matrix multiplication acceleration,"IEEE Transactions on Circuits and Systems I: Regular Paper

Hardware Architecture and Mapping

BenDi Hardware Architecture:

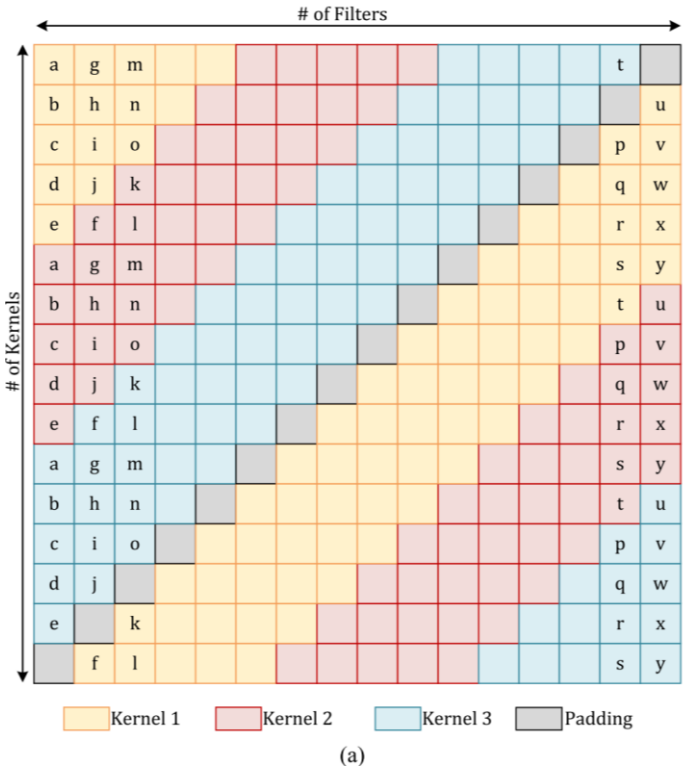
- Each BenDi Systolic Array (BDSA) has **16x16** BP9 processing elements (BPEs).
- BPE use **AND** gates to perform multiplication.

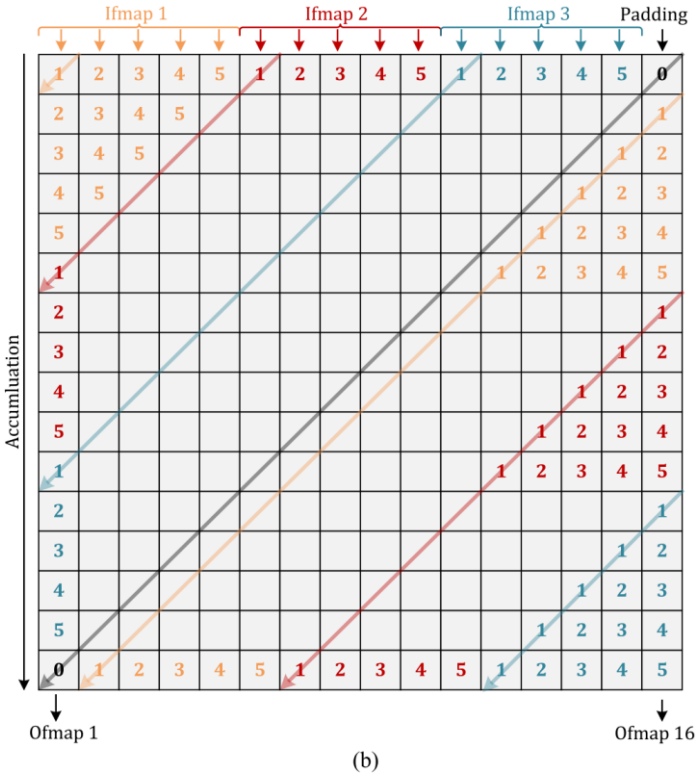


CNN Mapping on BenDi Architecture:

(a) For **weights**, kernels within the same filter are arranged in a single column and permutated by row numbers vertically.

(b) For **activation** mapping, the activations are diagonally distributed across the array and each column produces one output feature map.





Evaluation and Results

- Accuracy Comparsion:

TABLE II
ACCURACY COMPARISON OF QUANTIZATION OVER THREE DATASETS

Methods	MNIST	MIT-BIH		Apnea-ECG	
	ACC	Macro-F1	ACC	Macro-F1	ACC
FP32	98.84%	0.90	97.89%	0.871	87.59%
PTQ-INT8	98.63%	0.824	96.57%	0.754	77.28%
PTQ-BP9	97.45%	0.701	91.66%	0.771	77.14%
QAT-BP9	-	**0.854**	**96.89%**	**0.829**	**84.31%**
		-0.046	-1%	-0.042	-3.28%

- Layer-wise Utilization, Latency & Energy:

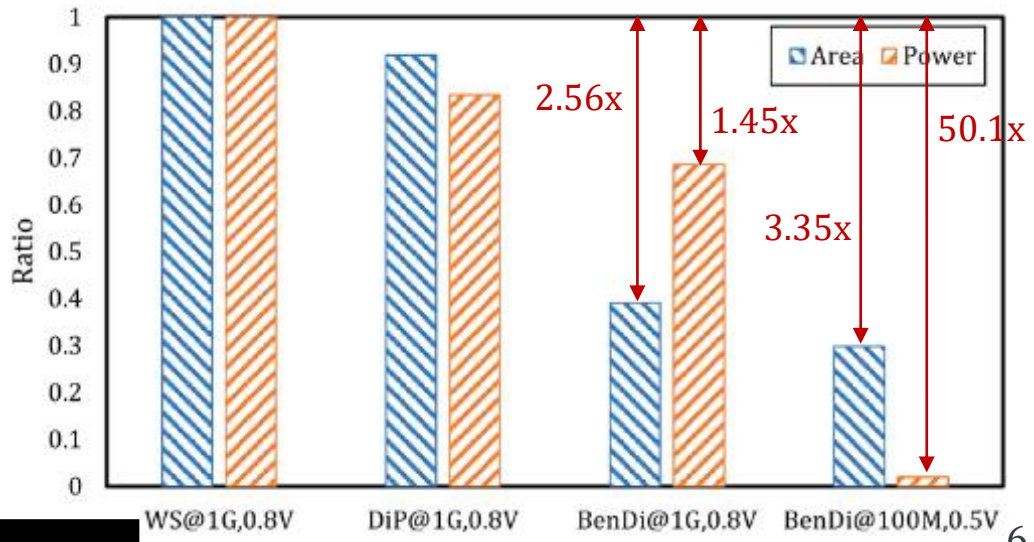
TABLE III
THE PERFORMANCE OF BENDI FOR MIT-BIH AND APNEA-ECG BENCHMARKS, AT 0.5V AND 100MHZ.

	MIT-BIH			Apnea-ECG		
	Util.	Latency	Energy	Util.	Latency	Energy
Conv1	25%	1.28	1.64	50%	4.8	12.31
Conv2	100%	3.36	17.23	100%	9.6	49.24
FC1	100%	4	20.51	100%	7.84	40.21
FC2	50%	0.48	2.46	50%	0.48	1.23
Total	**86.84%**	**9.12**	**41.84**	**88.38%**	**22.72**	**102.99**

Note: Latency is measured in μs. Energy is measured in nJ.

@ 1 GHz, 0.8 V
2.56x smaller area
1.45x lower power
Vs. WS

@ 100 MHz, 0.5 V
3.35x smaller area
50.1x lower power
5x higher energy-efficiency
Vs. WS



Conclusion

Comparison with SOTA work:

24x

BenDi improving the area efficiency by 24.33x , compared to [11].

62x

BenDi improving energy efficiency by 62.86x, compared to [11].

TABLE IV
HARDWARE COMPARISON WITH OTHER BIOELECTRONICS IMPLEMENTATIONS.

	[11]	[12]	**BenDi**
Paradigm	Binary	Binary	**Bent-Pyramid**
Model	1D-CNN	BNN	1D-CNN
Dataset	MIT-BIH	MIT-BIH	MIT-BIH/Apnea-ECG
Data Width	16-bit	1-bit	**9-bit**
Accuracy(%)	98.4	98.78	**96.89**/84.31
Frequency (Hz)	100M	100K	**100M**
Process (nm)	55	55	**22**
Voltage (V)	0.85	1.2	**0.5**
Area (mm^2)	1.39 (0.305*)	0.43 (0.204*)	**0.085**
Power (mW)	5.14 (3.212*)	0.00784 (0.00482*)	**5.1299**
Throughput (GOPS)	2.04	-	**204.8**
Area Eff. (TOPS/mm^2)	0.001 (0.099*)	-	**2.409**
Energy Eff. (TOPS/W)	0.397 (0.635*)	-	**39.92**

*: Technology scaling to 22nm according to DeepScaleTool [13].

Conclusion:

- Quasi-stochastic computing for efficient CNN inference
  BP9 enables deterministic computation with low-cost multiplication.
- Multi-level software–hardware optimization
  Hardware-aware QAT + DiP dataflow + BP9 computing + low-voltage implementation.
- Targeting edge bioelectronics
  Demonstrated on arrhythmia and sleep-apnea detection.

Thanks !

Contacts:

Bochen Ye
bochen.ye@ed.ac.uk

Yihan Pan
yihan.pan@ed.ac.uk

Shady Agwa
shady.agwa@manchester.ac.uk

Themis Prodromakis
t.prodromakis @ed.ac.uk