Neuro-Symbolic Computing Architectures and Circuits for Embodied Intelligence

Arijit Raychowdhury

Steve W. Chaddick School Chiar and Professor
School of Electrical and Computer Engineering
Georgia Institute of Technology

arijit.raychowdhury@ece.gatech.edu

Embedded Systems Week (ESWEEK), Oct. 2, 2024

Outline

- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- Challenges and Conclusions

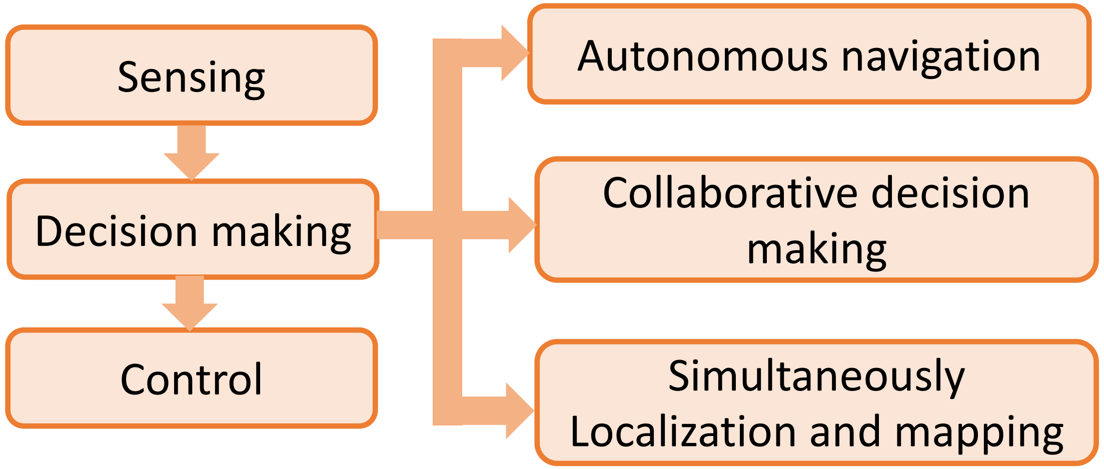
Outline

- **Motivation**
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- Challenges and Conclusions

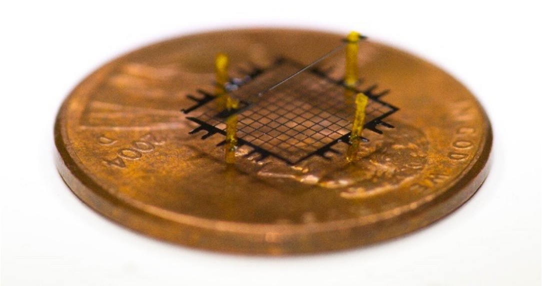
EI and Micro-Robotics

Palm-sized Drones

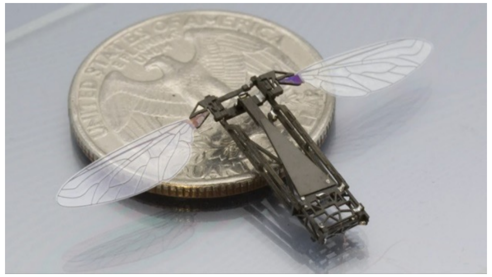
Intelligent Autonomous Cars



Jasmine microrobots



Berkeley Microrobots



Harvard Bee Microrobots

Georgia Tech Microrobot

Neuro-Symbolic Computing

Towards Cognitive and Trustworthy Embodied AI Systems

Neural

Recognition
Flexibility
Scalability

Symbolic

Explainability
Knowledge
Data Efficient

- Neural Components:
  - Bio-inspired: neuromorphic
  - CNN-inspired: non-neuromorphic

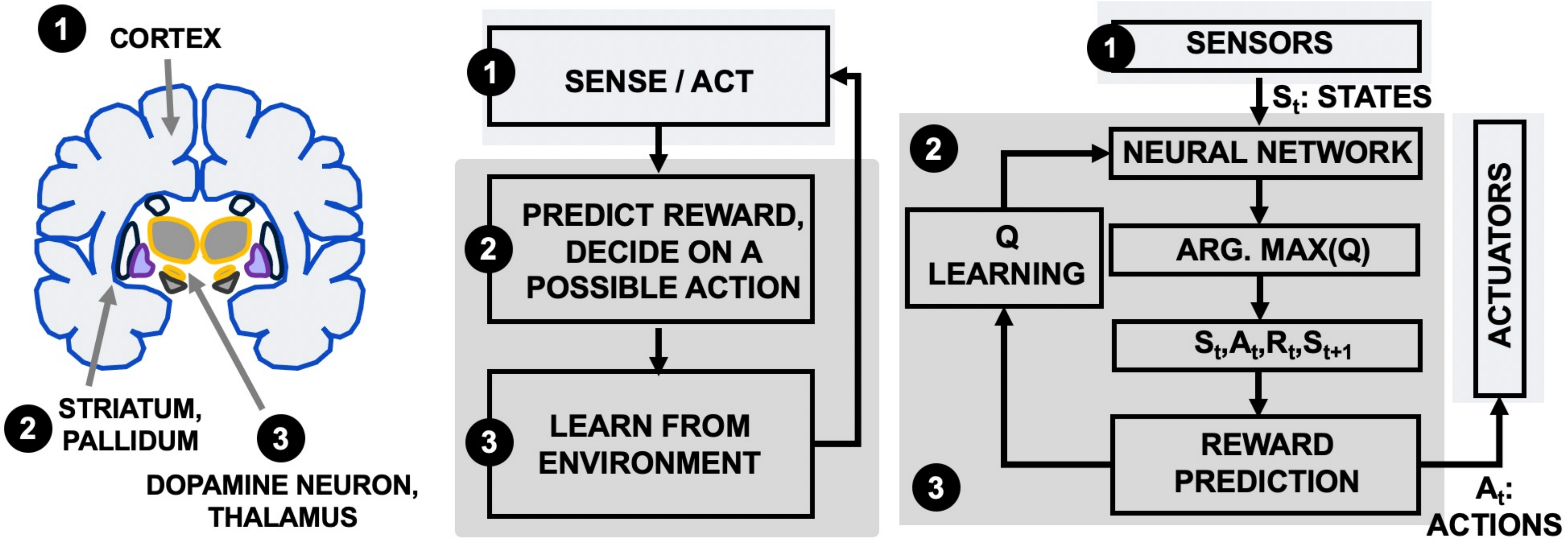
Outline

- Motivation
- **Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence**
  - Reinforcement Learning on the Edge Robotics
  - Swarm Intelligence on the Edge Robotics
  - Neuro-inspired SLAM for Edge Robotics
  - Hybrid Architecture for Target Tracking
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- Challenges and Conclusions

Outline

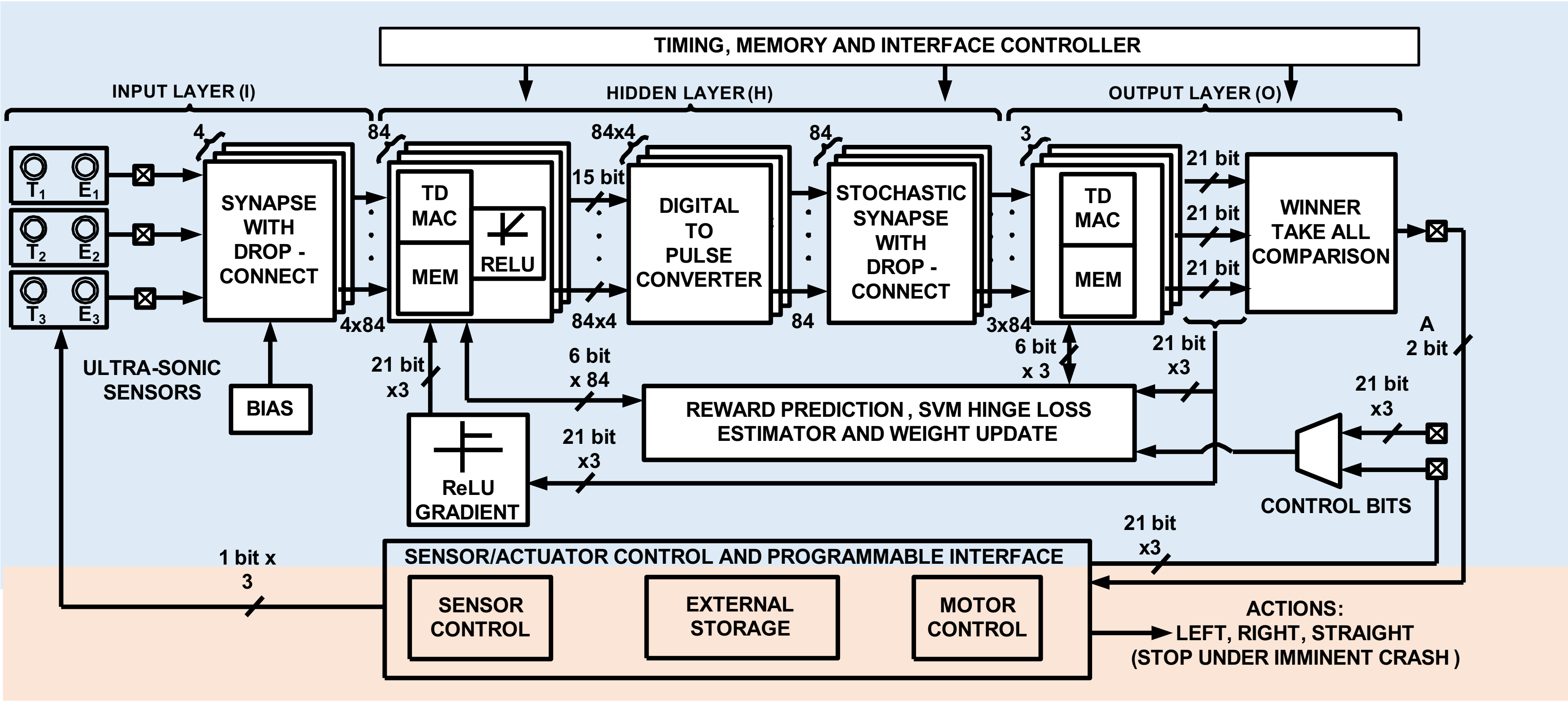
- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - **Reinforcement Learning on the Edge Robotics**
  - Swarm Intelligence on the Edge Robotics
  - Neuro-inspired SLAM for Edge Robotics
  - Hybrid Architecture for Target Tracking
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- Challenges and Conclusions

Providing Autonomy to Edge Devices



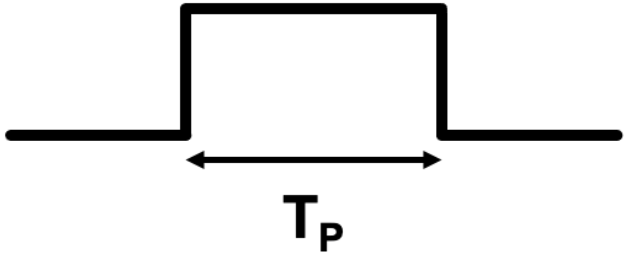
- Reinforcement Learning can maximize a set reward through exploration of the state-space and taking actions.
- A neural network maps the state-space to the action space optimally.

Time-Based Design for Online RL



- Time-domain mixed-signal multiply-and-accumulate unit.
- Bio-mimetic and takes advantages of inherent sparsity in the network.

Processing with Time-Encoded Pulses



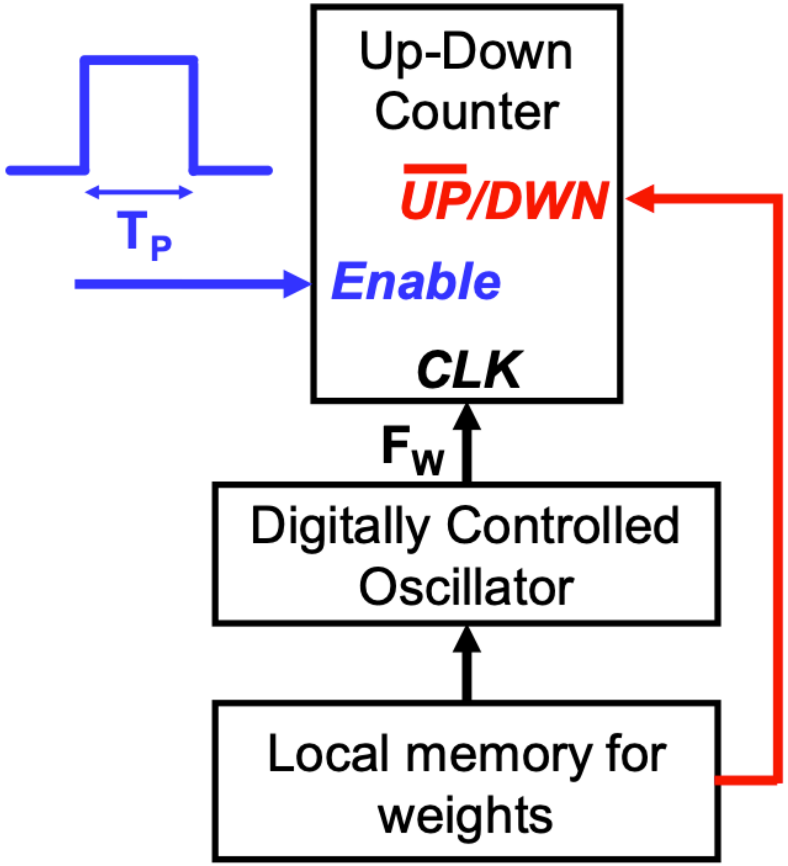
$$T_P = X.\ T_0$$

Operand Reference Time Constant

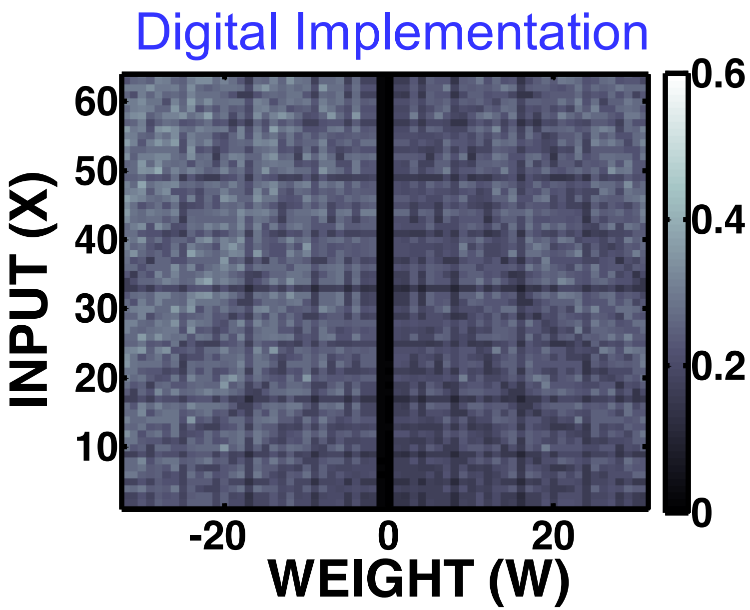
$$F_W = W.\ F_0$$

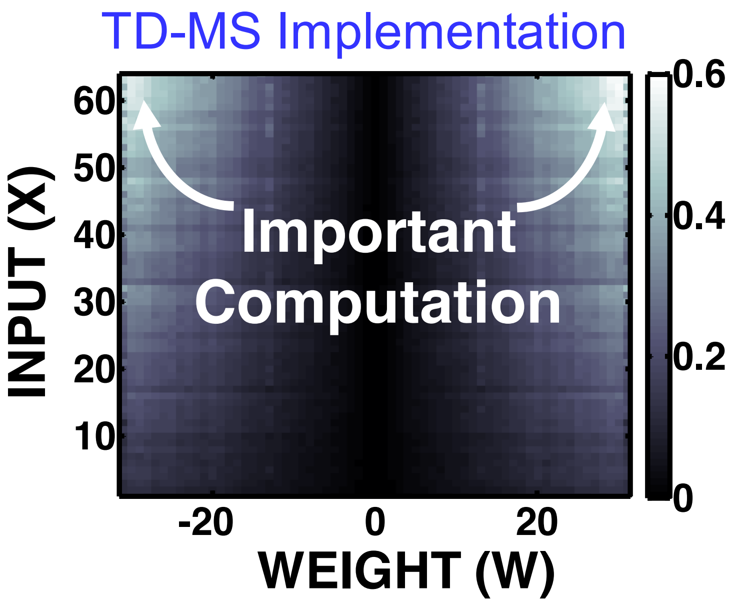
$$Y = T_P\ .\ F_W = X.W.(T_0.F_0)$$

$$Y = T_0.F_0 \sum X_i.W_i$$



Energy Efficiency of Time-Domain Processing

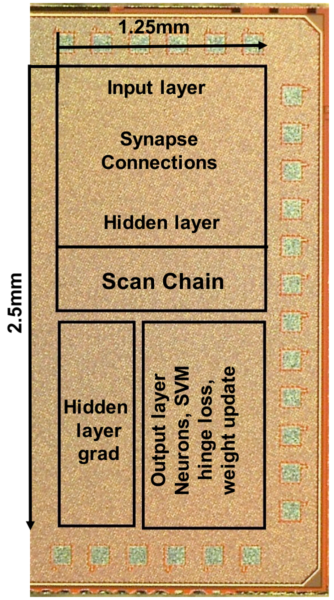




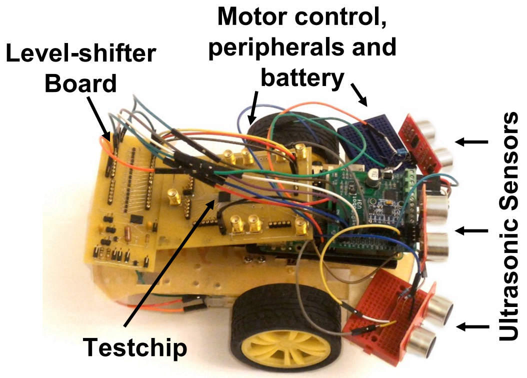
TD-MS: time-domain mixed-signal

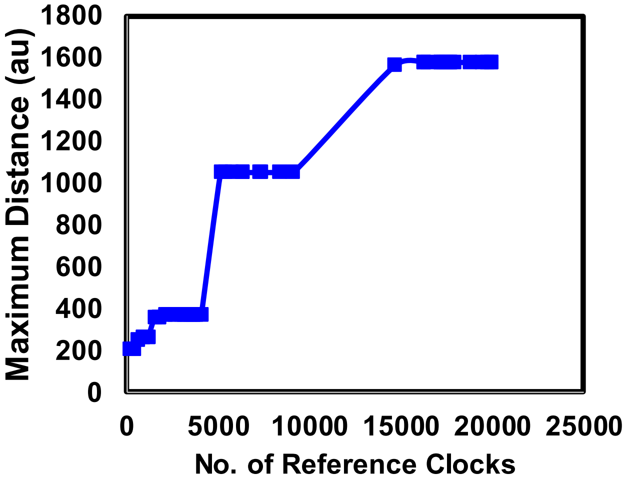
- Number of switching events (and hence, <u>energy/op</u>) in TD neuron <u>is proportional to</u> the value of the operands (and hence, <u>the importance of the computation)</u>
- Bio-mimetic and takes advantage of inherent sparsity in the network
- An average of 42% reduction in energy/op
- 45% lower area, 47% lower interconnect power and 16% lower leakage

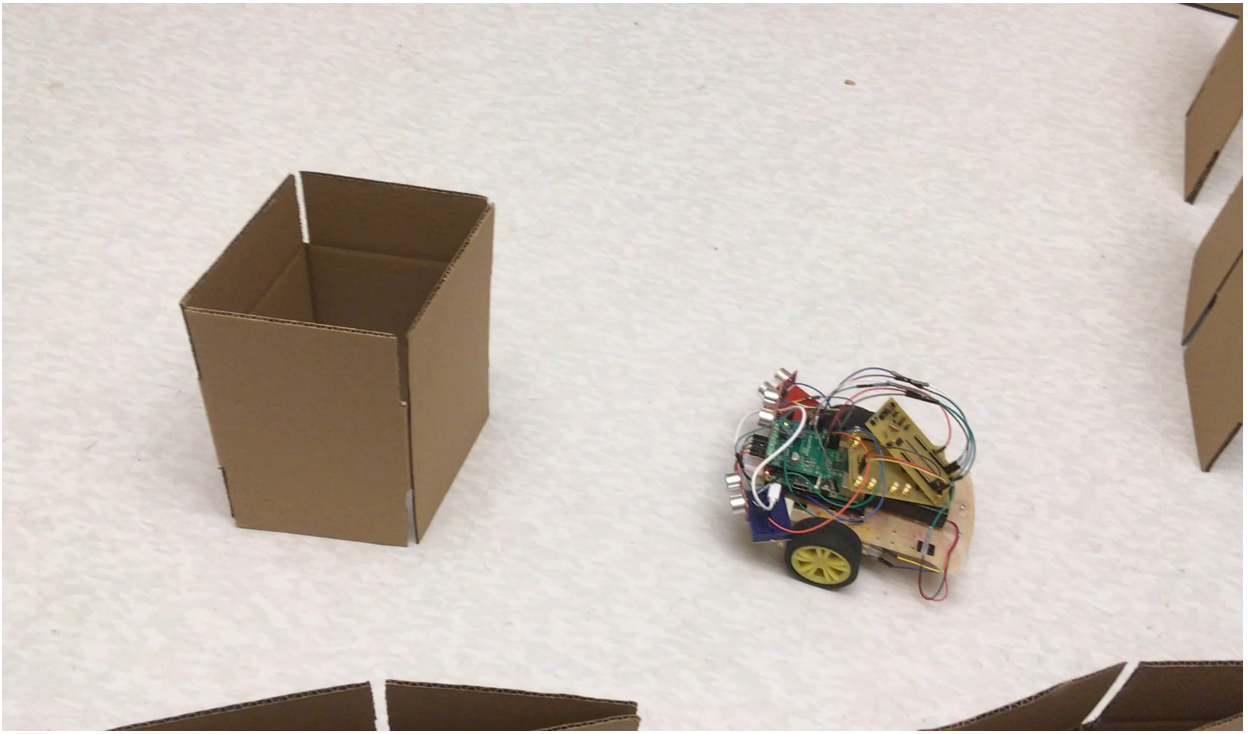
Reinforcement Learning Chip in Action



55nm 1P8M CMOS	
1.2*2.5mm	
QFN package	







Anvesha Amravati et al., **ISSCC** 2018
Anvesha Amravati et al., **JSSC** 2019

Outline

- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Reinforcement Learning on the Edge Robotics
  - **Swarm Intelligence on the Edge Robotics**
  - Neuro-inspired SLAM for Edge Robotics
  - Hybrid Architecture for Target Tracking
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- Challenges and Conclusions

Collaborative Intelligence in Swarms

Applications

Model-Free

Multi-robot patrolling

Multi-robot predator-prey

linear operation / nonlinear activation

Physical-Model-Based

Obstacle/collision avoidance

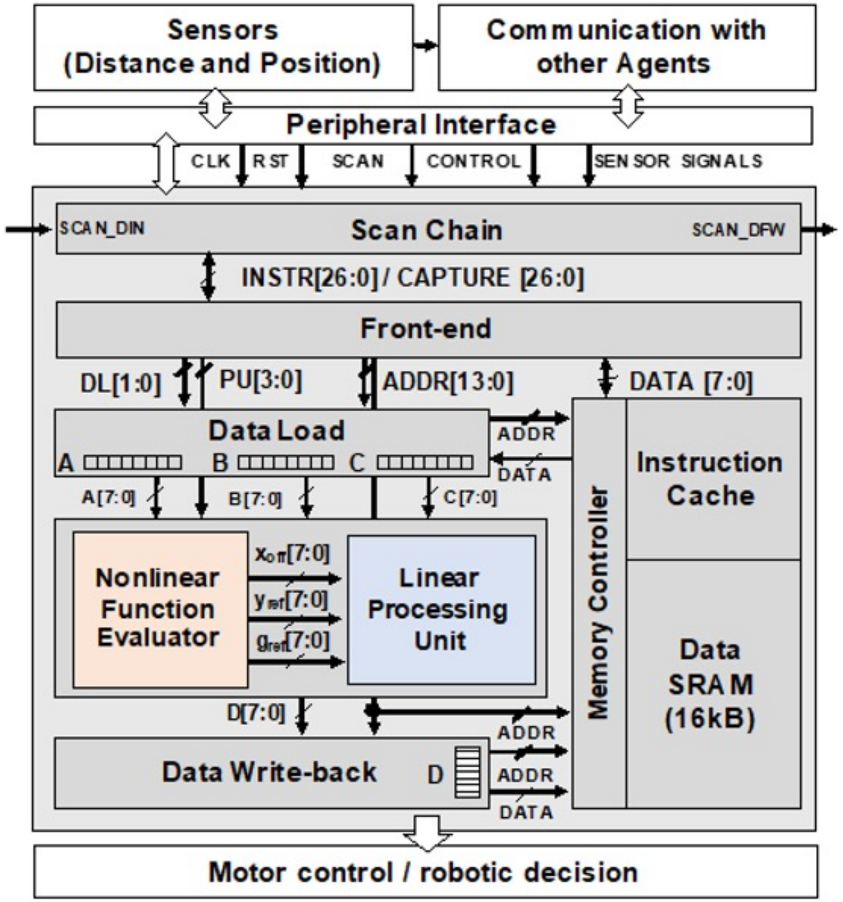
Pattern-formation

nonlinear function / linear operation

Algorithms

Algorithm	Algorithm Type	Application Support	Mathematical Structure	Nonlinear Functions	Linear Operations
Cooperative reinforcement learning	Model-Free (Neural Network based)	1. Multi-robot predator-prey [9] 2. Multi-robot patrolling [10]	$ReLU(\sum x_i w_i)$	ReLU	x, +, $\sum$
		3. Cooperative exploration [11]	$tanh(\sum x_i w_i)$	tanh	
Potential field approach	Model-based	4. Path planning [12] 5. Collision avoidance [12]	$\sum x_i \cos(y_{id})$	cosine	x, +, -, $\sum$
		6. Pattern-formation [13]	$\sum x_i tanh(\frac{\sqrt{y^2 - y_1^2}}{\zeta})$	tanh, reciprocal, square, sqrt	

System Architecture

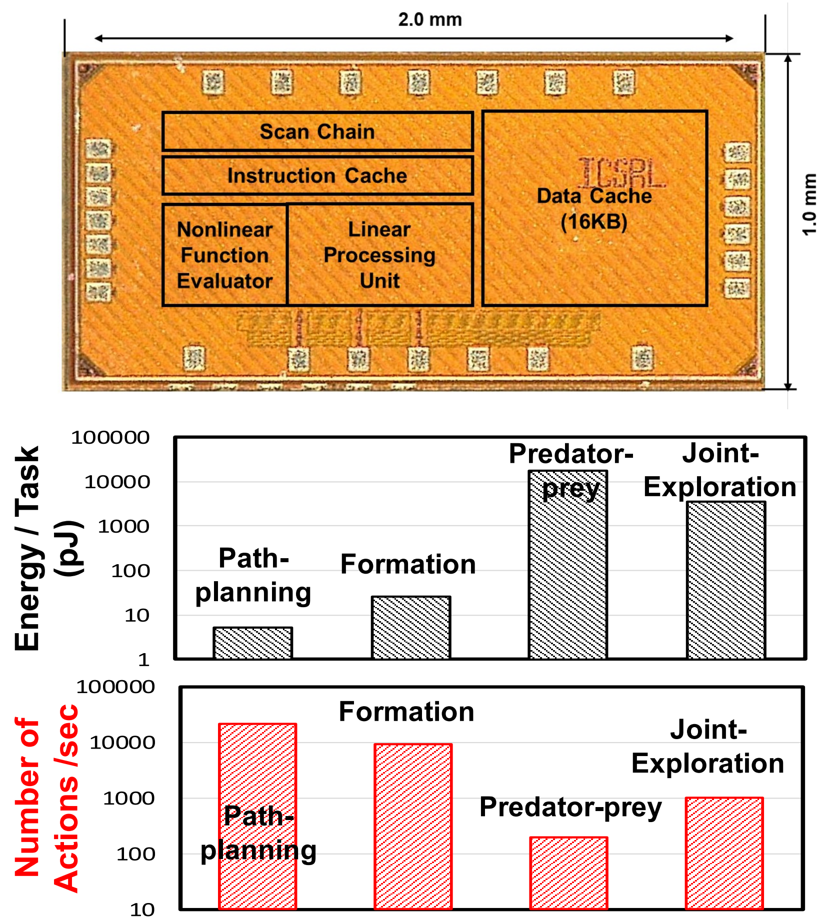


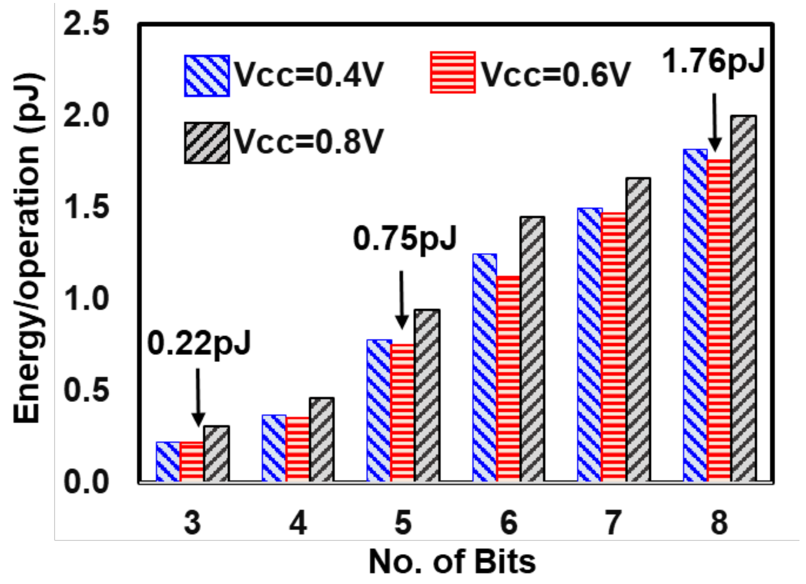
No. Bits	TD-MS		HDMS	
	Average	Worst	Average	Worst
3	0.10	0.49	0.16	0.52
4	0.14	0.56	0.19	0.61
5	0.28	0.72	0.29	0.74
6	0.64	1.74	0.69	0.94
7	2.21	3.86	0.70	1.02
8	5.82	9.32	0.69	1.27

Energy/MAC (Normalized to Digital)

- Increasing swarm size requires higher bit-precision
- Time-domain mixed-signal MAC design for low bit-precision
- Digital MAC design for high bit-precision

65nm Test-Chip and Measured Results

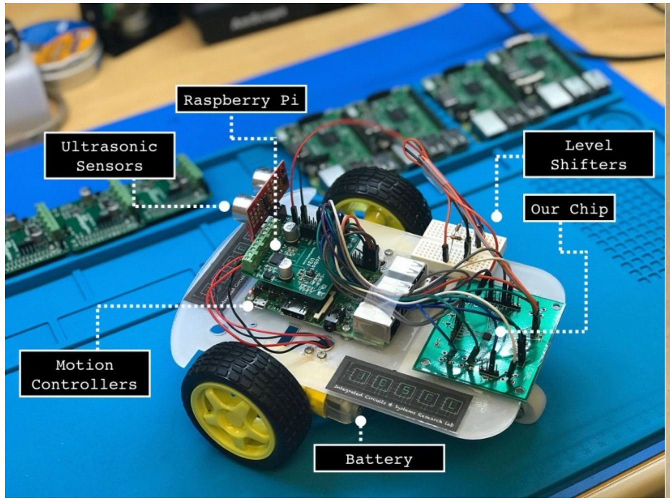


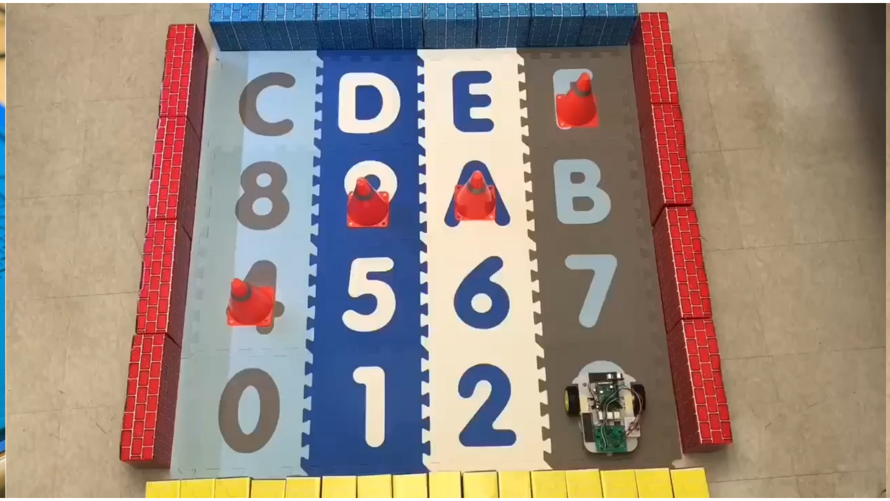


Energy/MAC for varying bit-width

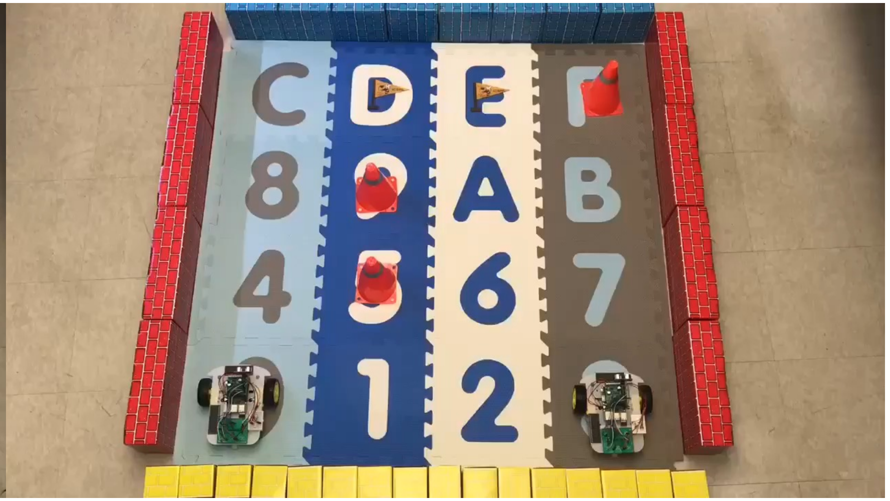
- 0.22-1.76 pJ/operation at 0.6V
- Maximum arithmetic energy efficiency 9.1 TOPS/W @ 3b, 0.6V, 1.1 TOPS/W @8b, 0.6V

Swarm Intelligence in Action





Exploration 16X real time



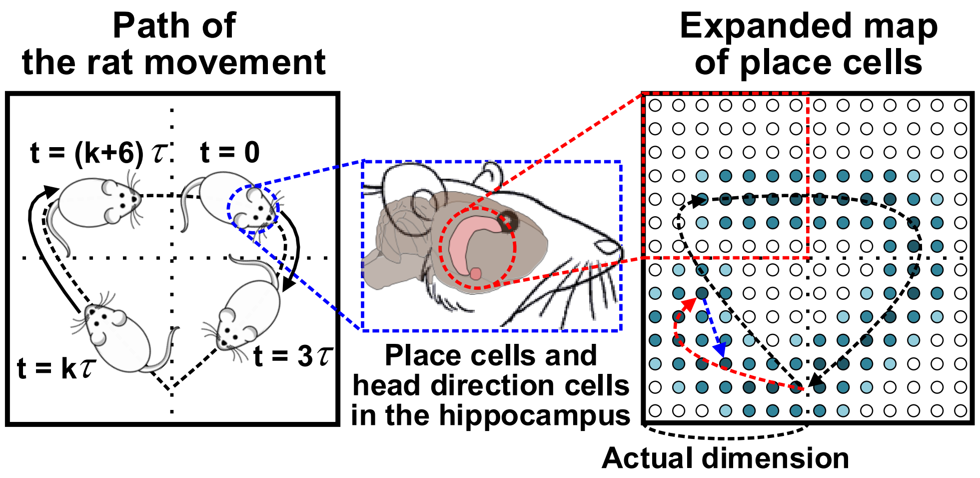
Collaborative RL in real time

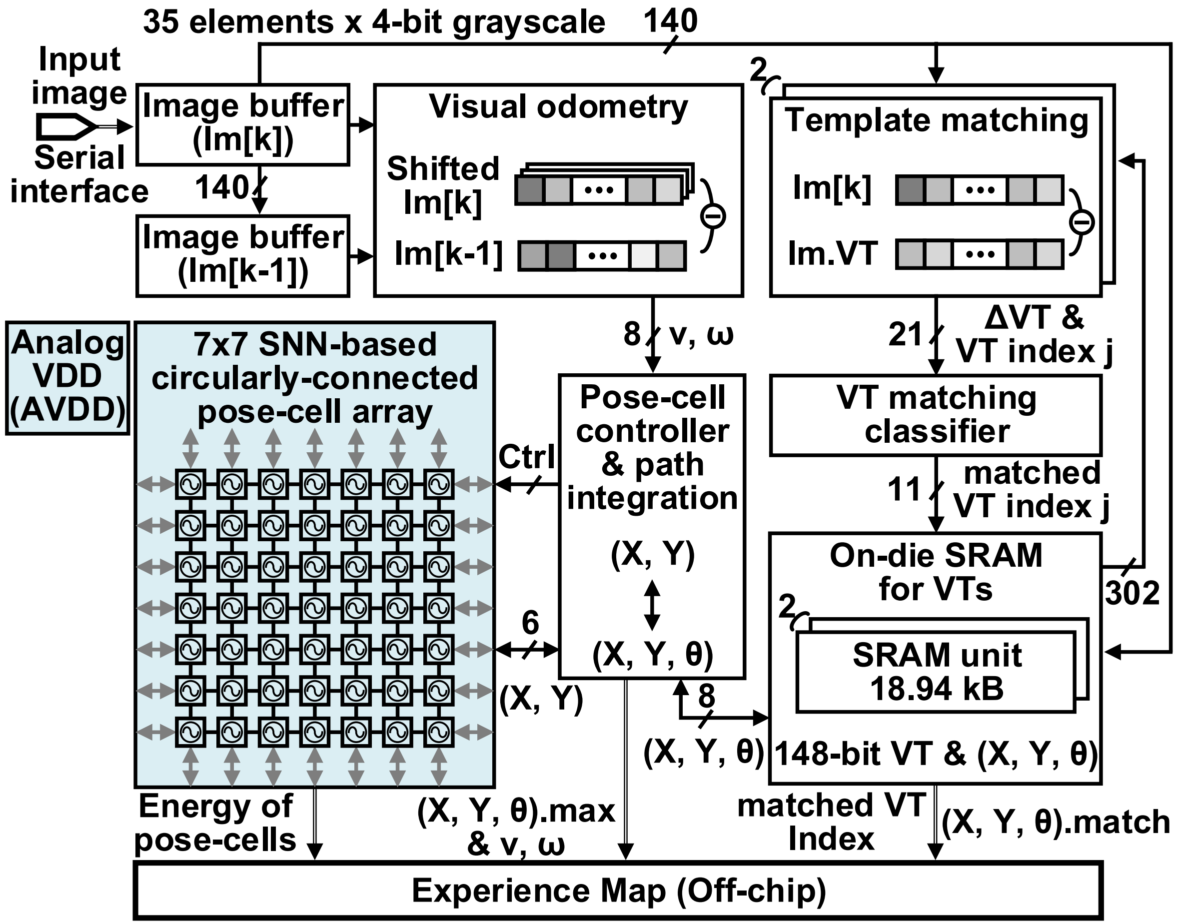
Ningyuan Cao et al., **ISSCC** 2018
Ningyuan Cao et al., **JSSC** 2019

Outline

- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Reinforcement Learning on the Edge Robotics
  - Swarm Intelligence on the Edge Robotics
  - **Neuro-Inspired SLAM for Edge Robotics**
  - Hybrid Architecture for Target Tracking
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- Challenges and Conclusions

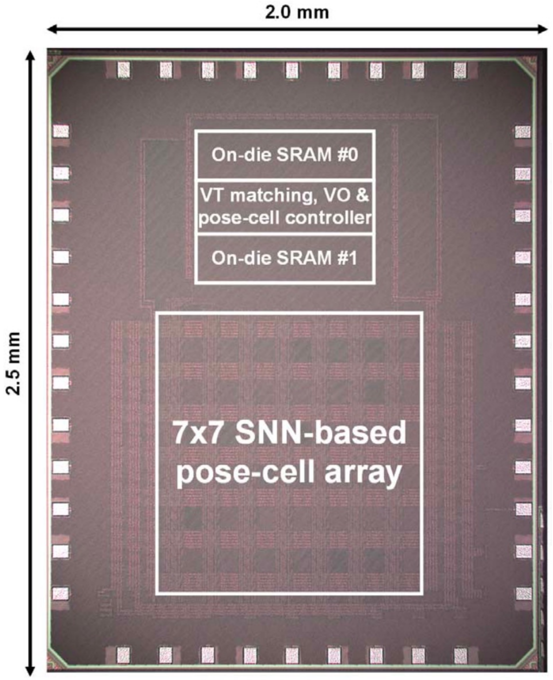
Spatial Cognition in the Rodent Brain



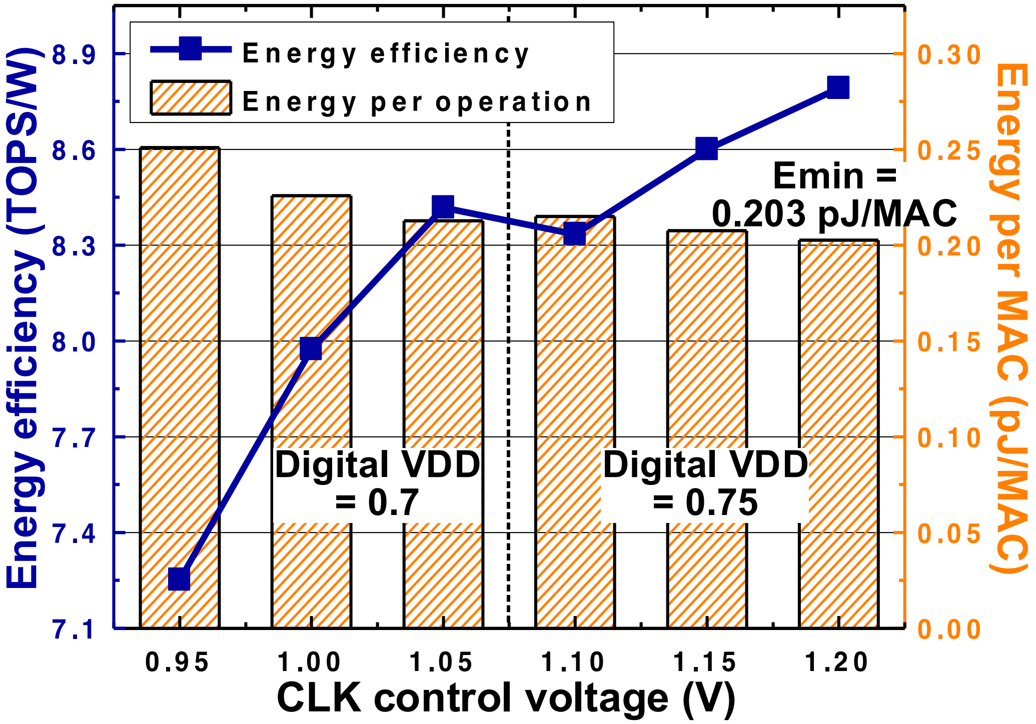


- SLAM in edge-robotics requires power-efficient circuit solutions
- Biological approaches can solve SLAM with extreme energy efficiencies
- Neuromorphic vision-based SLAM algorithm is a promising solution

Measured Results on 65nm Test-chip

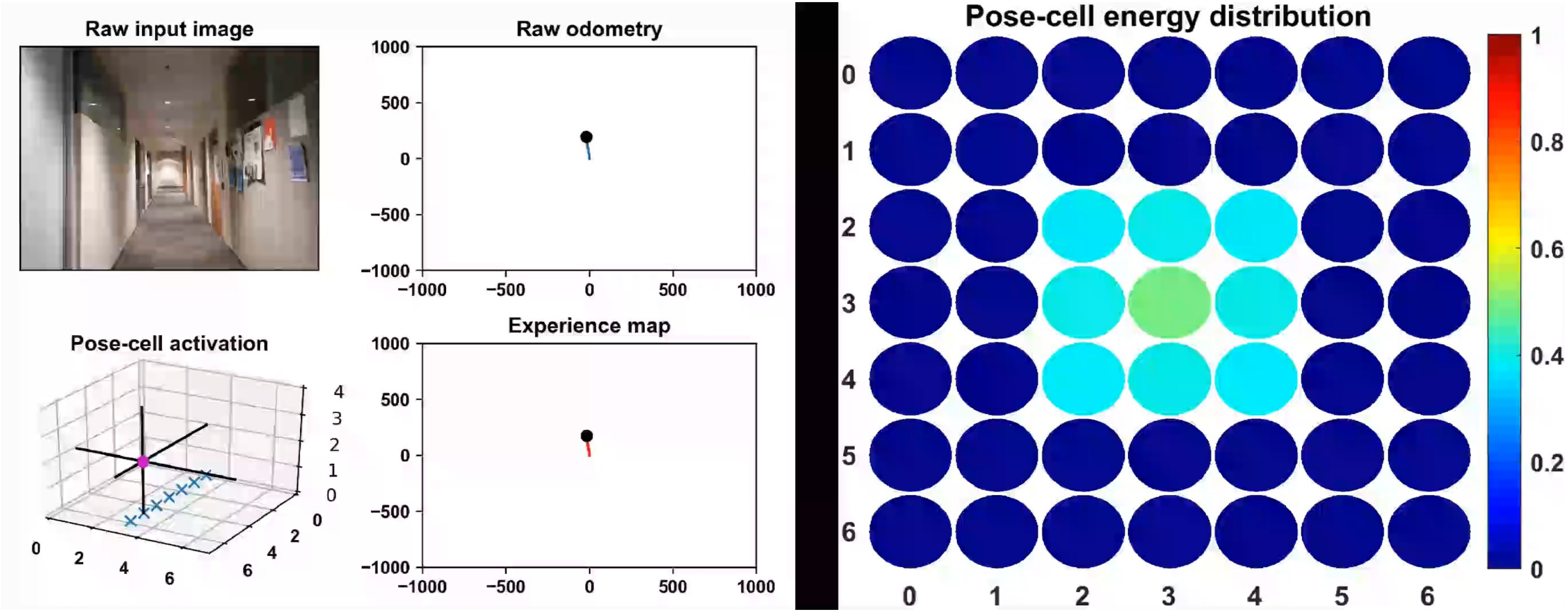


Technology	65 nm 1P9M CMOS
Die area	2.0 mm x 2.5 mm
On-chip memory	37.9 kB
Frequency	78.22-130.8 MHz
Digital VDD	0.7-0.75 V
Analog VDD	0.95-1.2 V
I/O VDD	2.5 V
Power	17.27-23.82 mW
Energy efficiency	7.25-8.79 TOPS/W
Package	QFN48



- 0.203-0.251 pJ/MAC at 0.95-1.2V
- Arithmetic energy efficiency (8.79 TOPS/W @ 4b, 1.2V), (7.25 TOPS/W @ 4b, 0.95V)

NeuroSLAM Operation in Action



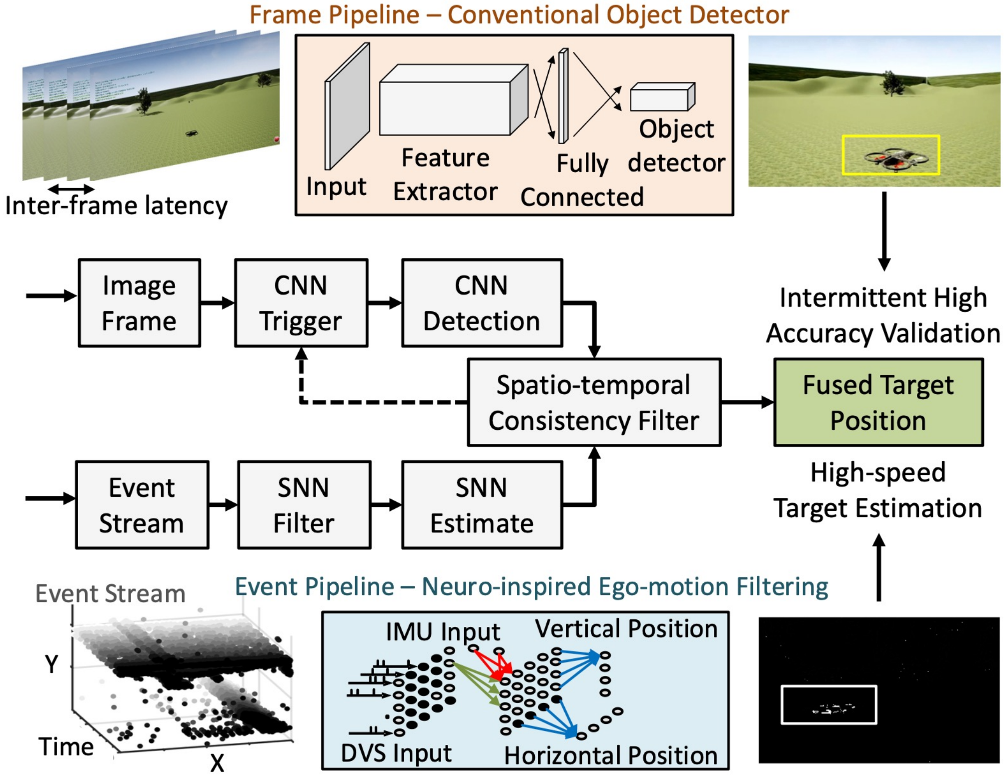
- SLAM operation and pose-cell energy distribution over input frames

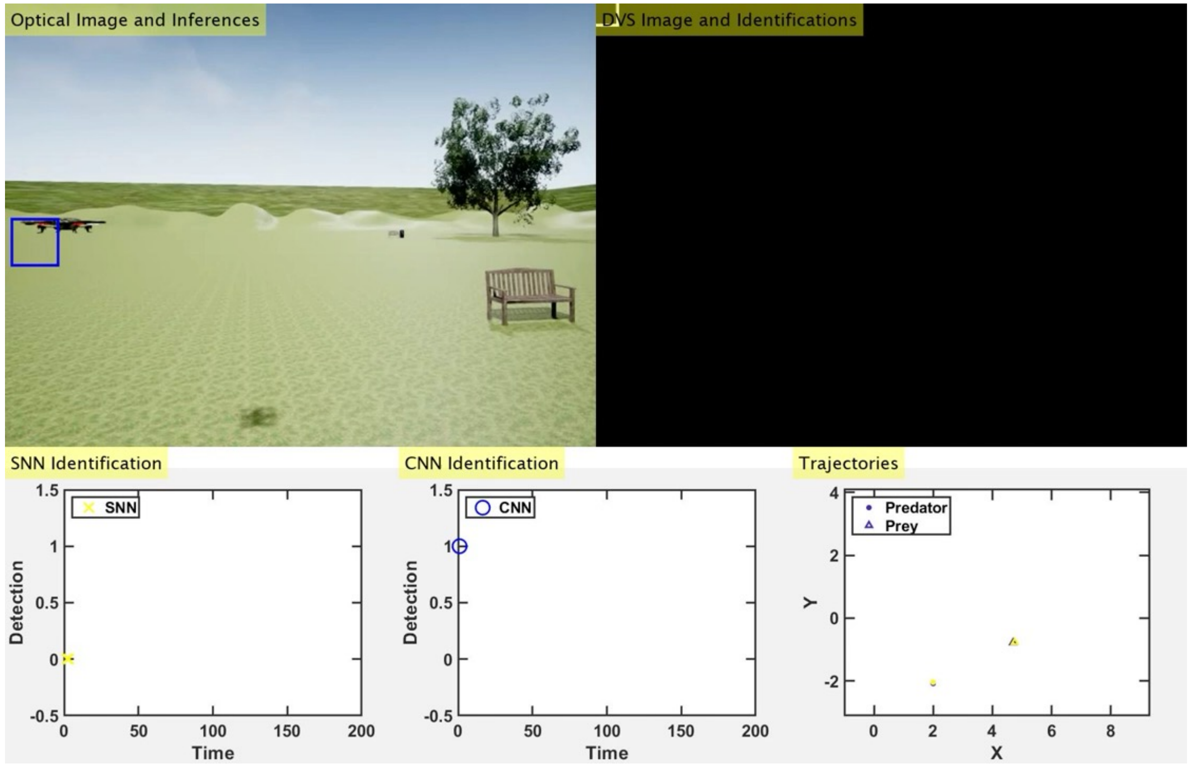
Jong-Hyeok Yoon et al., **ISSCC** 2020
Jong-Hyeok Yoon et al., **JSSC** 2020

Outline

- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Reinforcement Learning on the Edge Robotics
  - Swarm Intelligence on the Edge Robotics
  - Neuro-inspired SLAM for Edge Robotics
  - **Hybrid Architecture for Target Tracking**
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
- Challenges and Conclusions

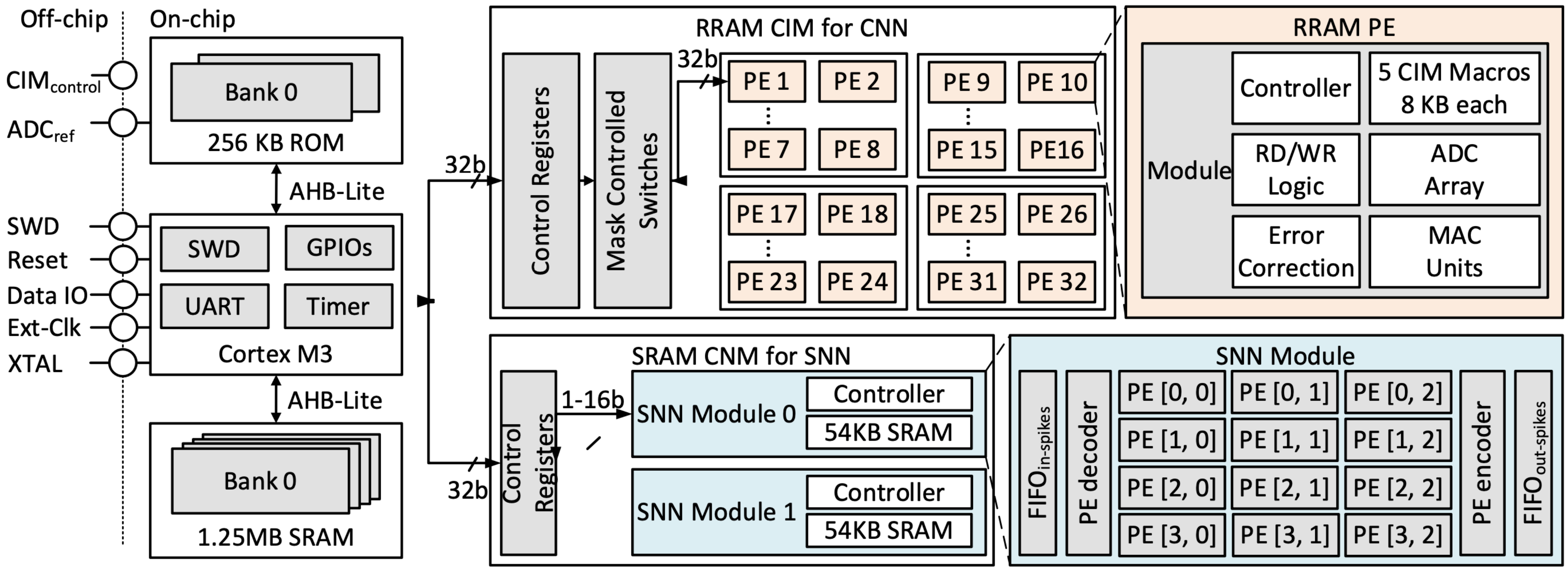
Hybrid SNN/CNN for Target Tracking





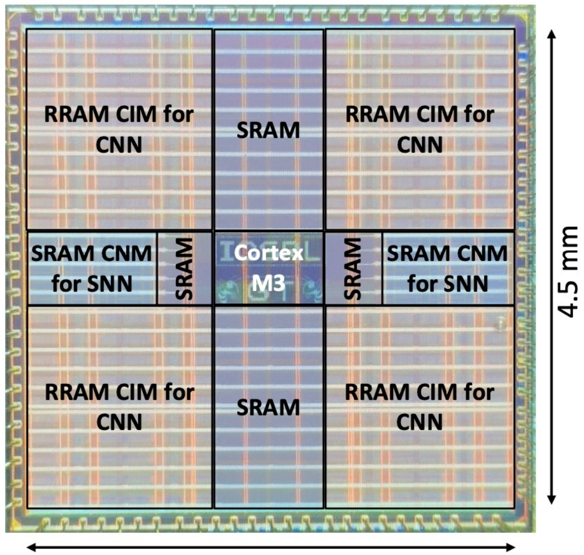
- CNNs are constrained by high latency, while SNNs are constrained by low accuracy
- Hybrid CNN/SNN algorithm shows potential to achieve low latency with high accuracy

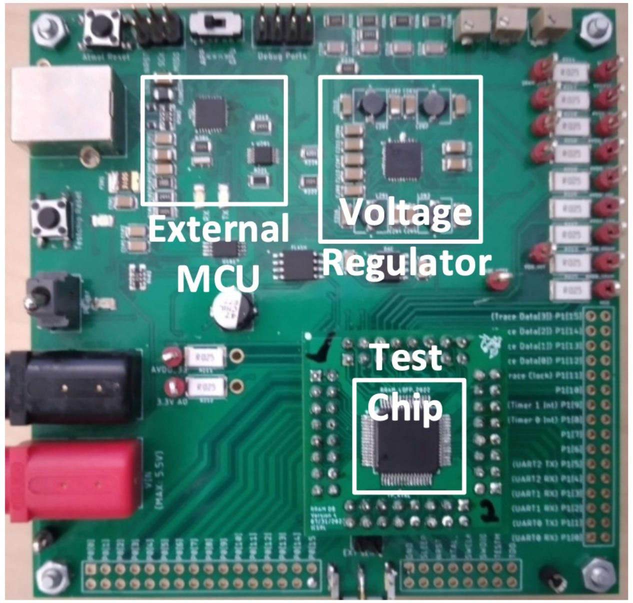
System Architecture



- Heterogenous programmable domain-specific accelerator architecture
- RRAM-based compute-in-memory for CNN, SRAM-based compute-near-memory for SNN

Chip Prototype





Technology	40 nm ULP TSMC	Microprocessor	Cortex M3
Chip Size	4.5 mm x 4.5 mm	Number of IO	62
Package	QFN 64	Communication	UART
On-chip RRAM	1.25 MB	Voltages Levels	7
On chip SRAM	1.25 MB	IO Supply	3.3 V
Max Clock (Hz)	100 MHz	Core Supply	0.9 V

Peak TOPS	14.74
Peak TOPS/W	73.53
SNN Throughput	11.1 Mevents/ sec
SNN + CIM_{Off}	4.6 mW
BER w/o ECC	$7x10^{-3}$
BER with ECC	$4.1x10^{-8}$

Muya Chang et al., **ISSCC** 2023
Ashwin Lele et al., **JSSC** 2023

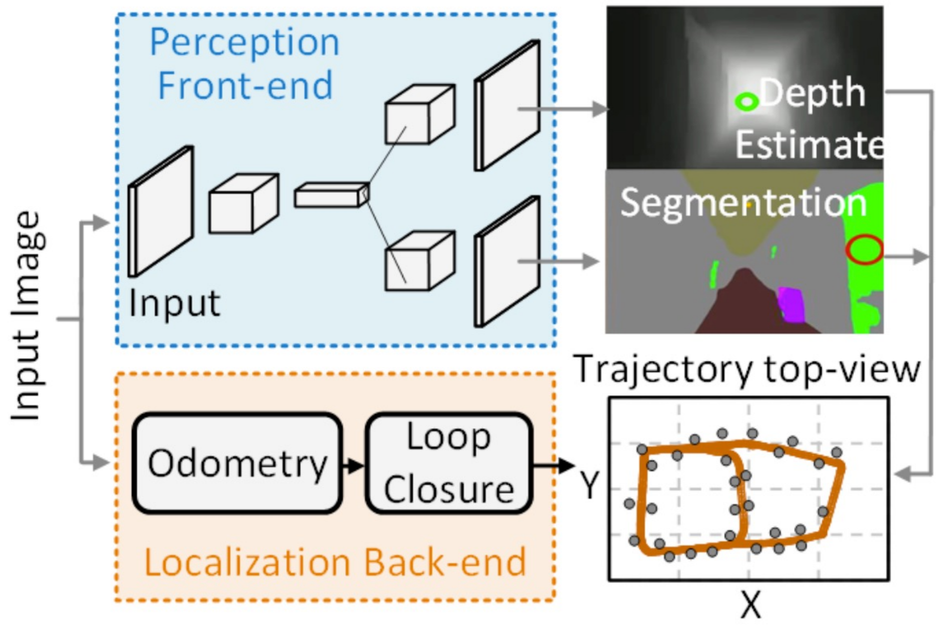
Outline

- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Reinforcement Learning on the Edge Robotics
  - Swarm Intelligence on the Edge Robotics
  - Neuro-inspired SLAM for Edge Robotics
  - Hybrid Architecture for Target Tracking
- **CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence**
  - Neuro-Symbolic Robotic Surveillance SoC
  - Neuro-Symbolic Workload Characterization and VSA architecture
- Challenges and Conclusions

Outline

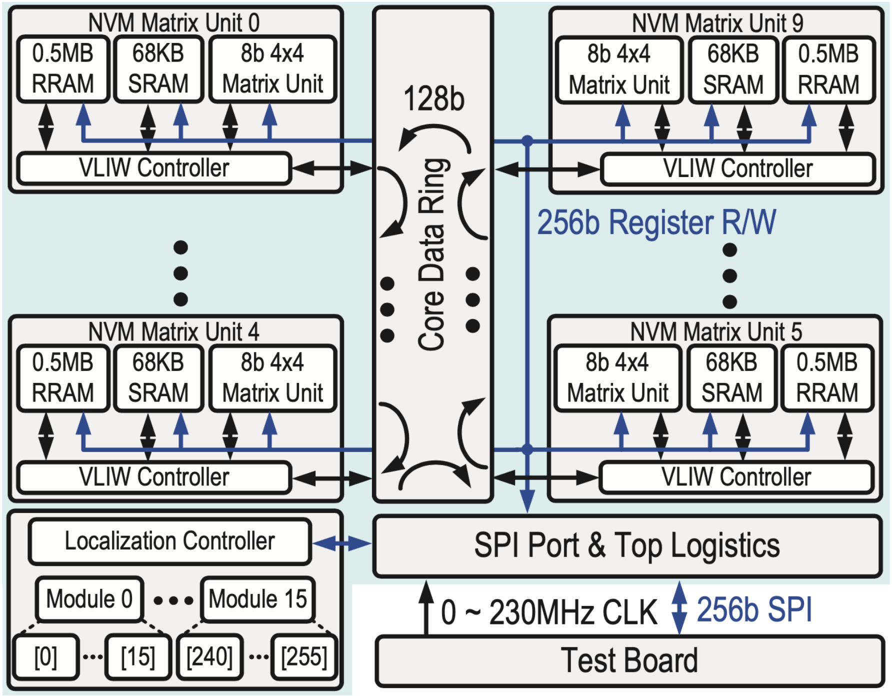
- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Reinforcement Learning on the Edge Robotics
  - Swarm Intelligence on the Edge Robotics
  - Neuro-inspired SLAM for Edge Robotics
  - Hybrid Architecture for Target Tracking
- **CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence**
  - **Neuro-Symbolic Robotic Surveillance SoC**
  - Neuro-Symbolic Workload Characterization and VSA architecture
- Challenges and Conclusions

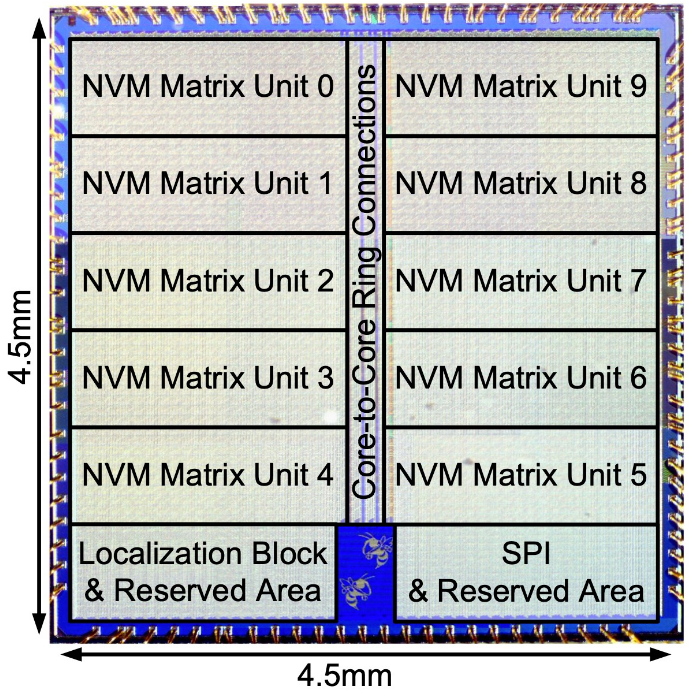
Neuro-symbolic for Robot Surveillance

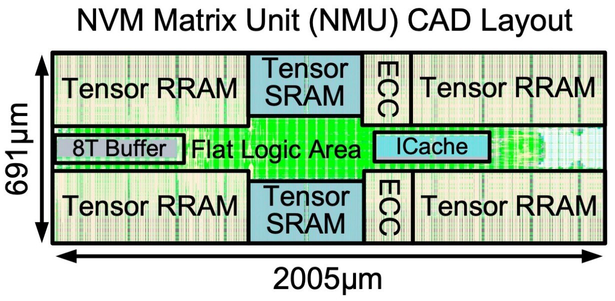


- ☐ **Perception (CNN):** Autonomous steering with obstacle avoidance:
  - Depth estimation: avoiding obstacles
  - Segmentation: identifying objects of interest for mapping
- ☐ **Localization:** Placing identified object/locations onto 2D map.

40nm VLIW/RRAM Integrated System-on-Chip







Technology	40nm CMOS with RRAM
Die Area	20.25mm^2
Voltage	**0.8 ~ 1.1V VDD/1.5 – 4.0V Write**
Frequency	**80 ~ 210MHz (NMUs)**
Memory	5MB RRAM/760KB SRAM
Sleep Mode	**110µW @ 500mV w/ Retention**

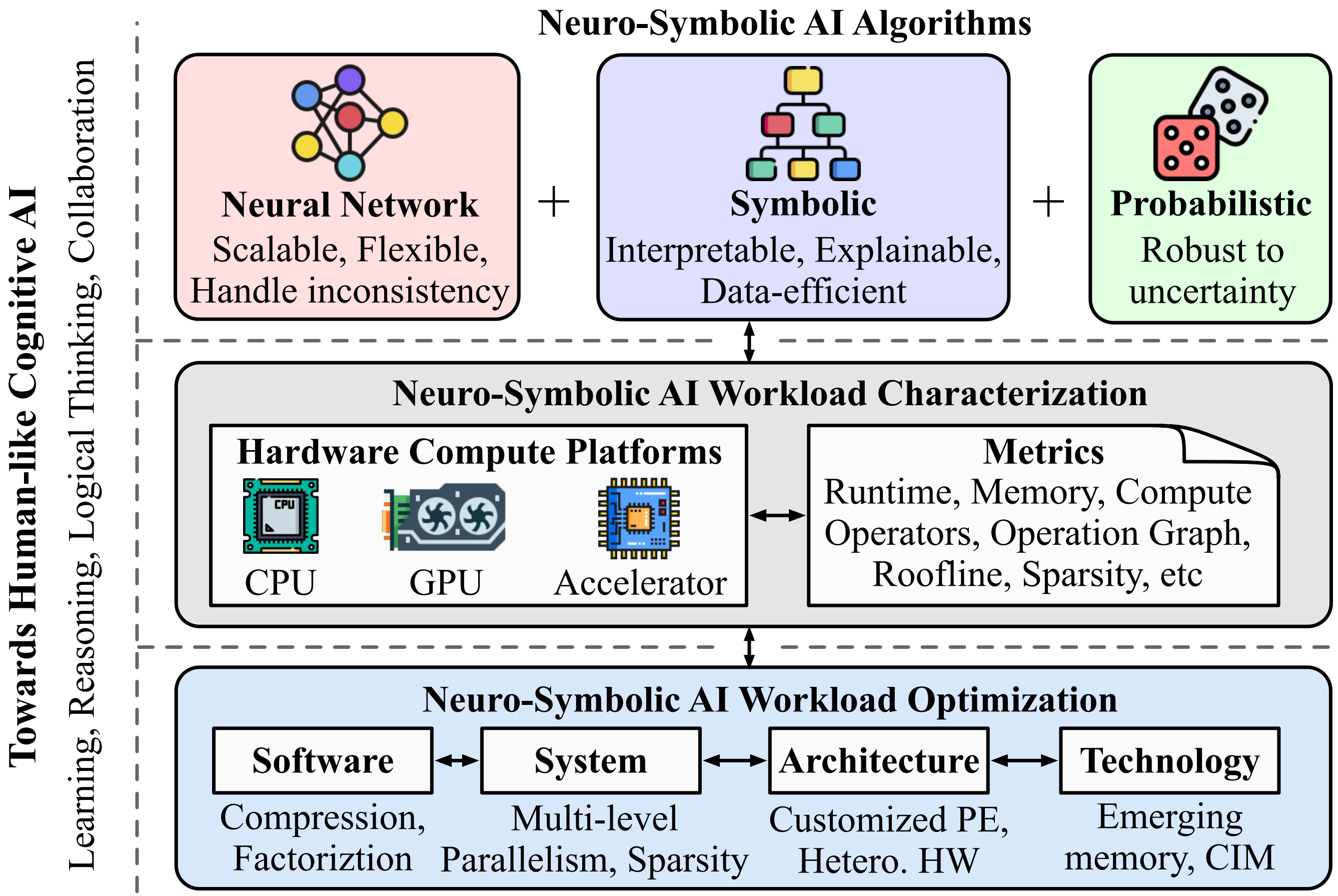
- Architecture: 10 VLIW-controlled NVM matrix units + localization block
- Technology: 760KB SRAM, 5MB RRAM with 2.07Mb/mm^2 and 0.256pJ/b

Samual Spetalnick et al., **ISSCC** 2024, **JSSC** 2024

Outline

- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Reinforcement Learning on the Edge Robotics
  - Swarm Intelligence on the Edge Robotics
  - Neuro-inspired SLAM for Edge Robotics
  - Hybrid Architecture for Target Tracking
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Neuro-Symbolic Robotic Surveillance SoC
  - **Neuro-Symbolic Workload Characterization and VSA architecture**
- Challenges and Conclusions

Neuro-Symbolic AI Workload Characterization



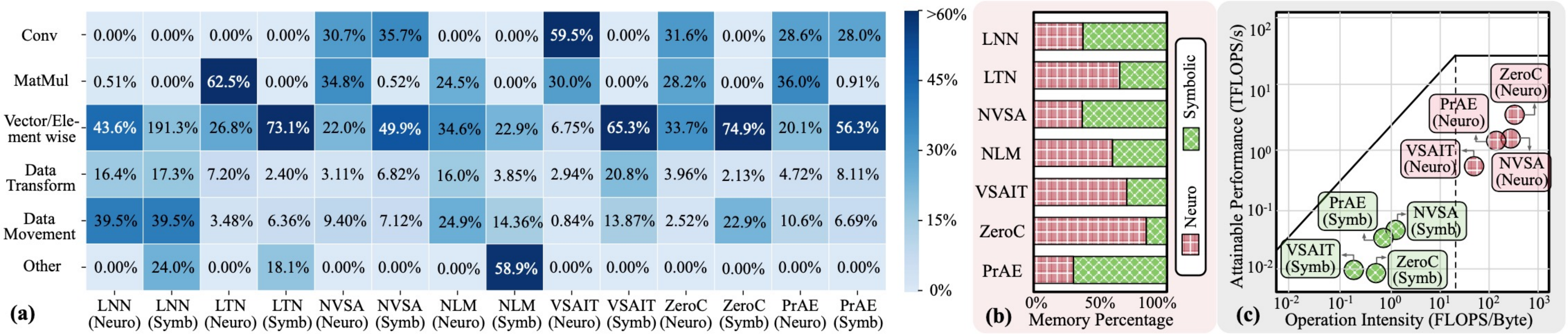
- **System 1:** thinking fast (neuro)
- **System 2:** thinking slow (symbolic)

- **Characterize** neuro-symbolic workloads
- **Identify** potential inefficiency reasons
- **Optimize** neuro-symbolic system via SW/HW co-design

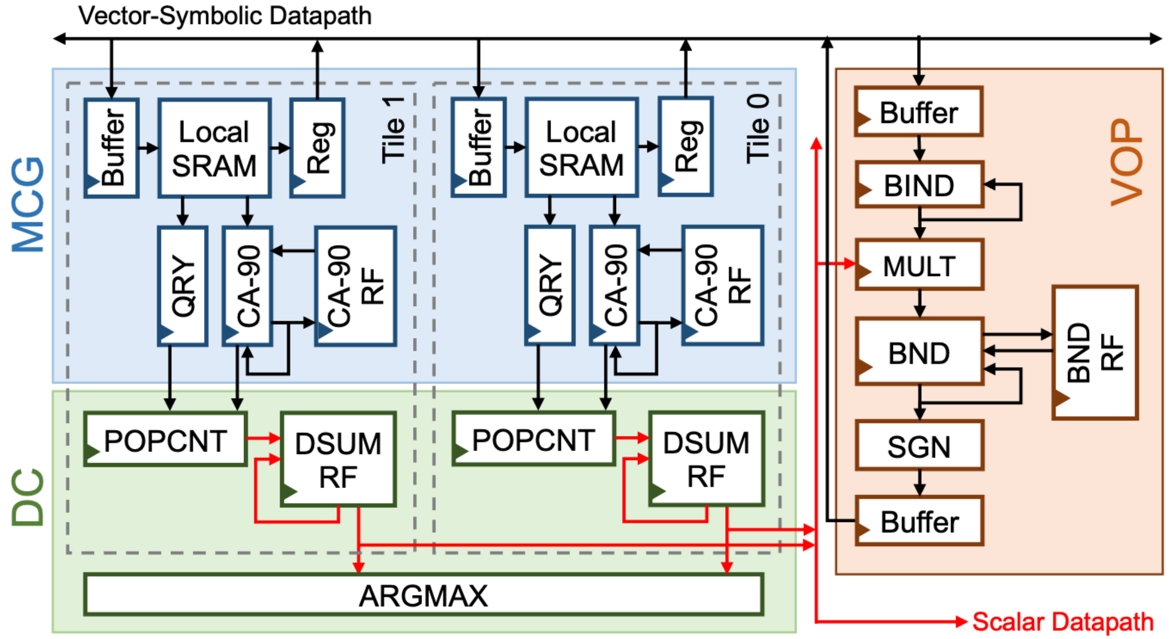
Zishen Wan et al., **ISPASS** 2024

Profiling and Arch Support for Neuro-Symbolic

- **Goal**: understand compute/memory characteristics of neuro-symbolic workloads
- **Key Idea**: profile neuro-symbolic workloads on heterog. CPU/GPU systems
- **Key Takeaways**:
  - Operator: symbolic is dominated by vector/element tensor and logical ops
  - Latency: symbolic is inefficient on CPU/GPU
  - System: neuro is compute-bounded, symbolic is memory-bounded; complex control



SW/HW Co-Design for Vector-Symbolic Arch



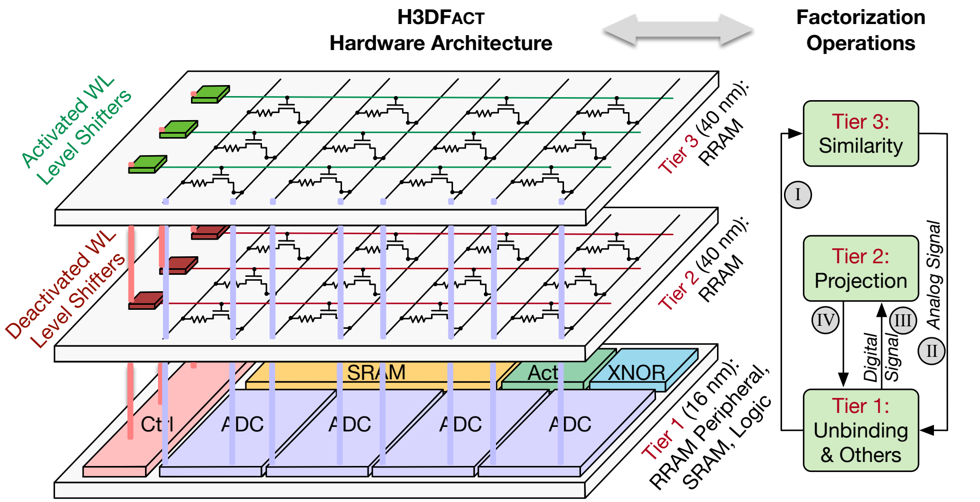
Workload	Layer	Application
MULT	Perception	Multi-modal learning and Inference [61]
TREE	Reasoning	Tree encoding and search [53]
FACT		Factorization of data sets [54]
REACT	Control	Motor learning and recall [62]

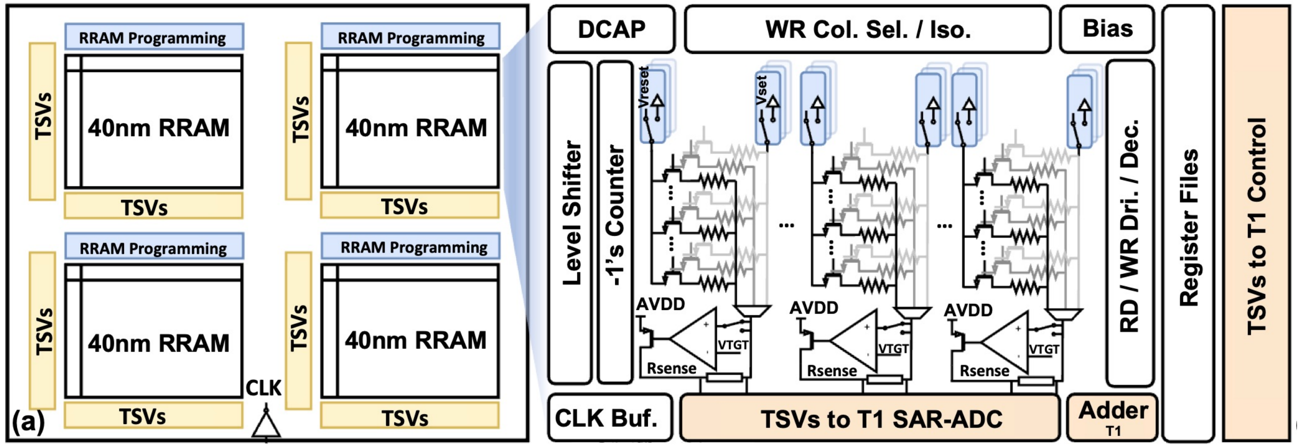
- Multi-tile hardware and dataflow for vector-symbolic architecture (VSA)
- Applicable to various VSA workloads and applications

Zishen Wan et al., **TCASAI** 2024
Mohamed Ibrahim et al., **DATE** 2024

Heterogeneous 3D CIM for Neuro-Symbolic

- **Goal**: Efficient & scalable factorization of holographic sensory representation
- **Key Idea**:
  - <u>Algorithm</u>: High-dimensional holographic vector-based factorization solver
  - <u>Hardware</u>: Heterogeneous 3D-CIM architecture; Improve factorization accuracy and convergence with intrinsic hardware stochasticity





Zishen Wan et al., **DATE** 2024 (SRC TECHCON)

Outline

- Motivation
- Bio-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Reinforcement Learning on the Edge Robotics
  - Swarm Intelligence on the Edge Robotics
  - Neuro-inspired SLAM for Edge Robotics
  - Hybrid Architecture for Target Tracking
- CNN-Inspired Neuro-Symbolic Computing for Embodied Intelligence
  - Neuro-Symbolic Robotic Surveillance SoC
  - **Neuro-Symbolic Workload Characterization and VSA architecture**
- **Challenges and Conclusions**

Conclusion

- Next generation of autonomy will be all-pervasive and ubiquitous
- Autonomy requires sensing, decision making, learning from actions and actuation.
- TinyML in micro-robotics will enable exciting new features in remote sensing, reconnaissance and disaster relief.
- Analog and mixed-signal compute can be augmented with digital techniques for seamless scalability of bit-precision.
- Smart algorithms need to be married to smart hardware design to enable intelligence at high energy efficiency.
- Golden age for hardware design…!!

Neuro-Symbolic Computing Architectures and Circuits for Embodied Intelligence

Arijit Raychowdhury

Steve W. Chaddick School Chiar and Professor
School of Electrical and Computer Engineering
Georgia Institute of Technology

arijit.raychowdhury@ece.gatech.edu

Embedded Systems Week (ESWEEK), Oct. 2, 2024