Circuits and System Technologies for Energy-Efficient Edge Robotics

Zishen Wan, Ashwin Lele, Arijit Raychowdhury

School of Electrical and Computer Engineering
Georgia Institute of Technology

(Work done by many people)

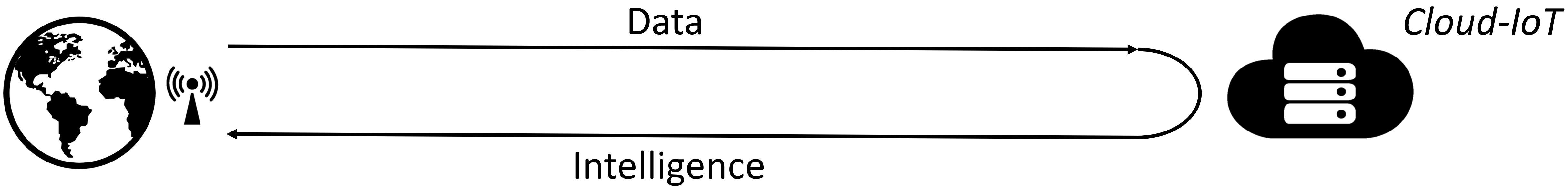
Outline

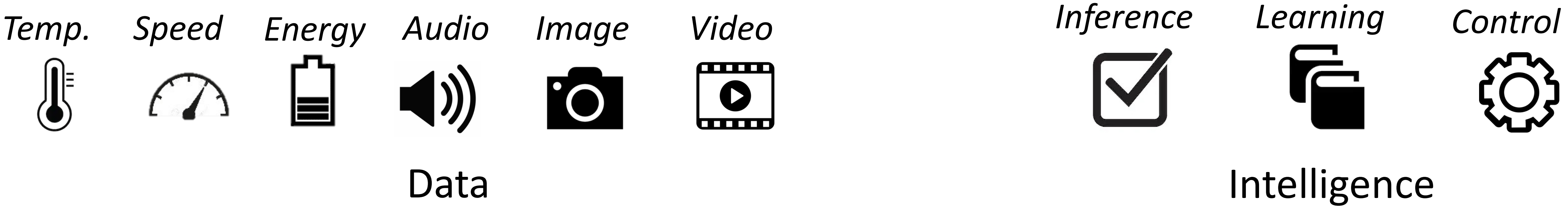
- Motivation
- Reinforcement Learning on the Edge
- Swarm Intelligence on the Edge
- Neuro-inspired SLAM on the Edge
- Challenges and Conclusions

Outline

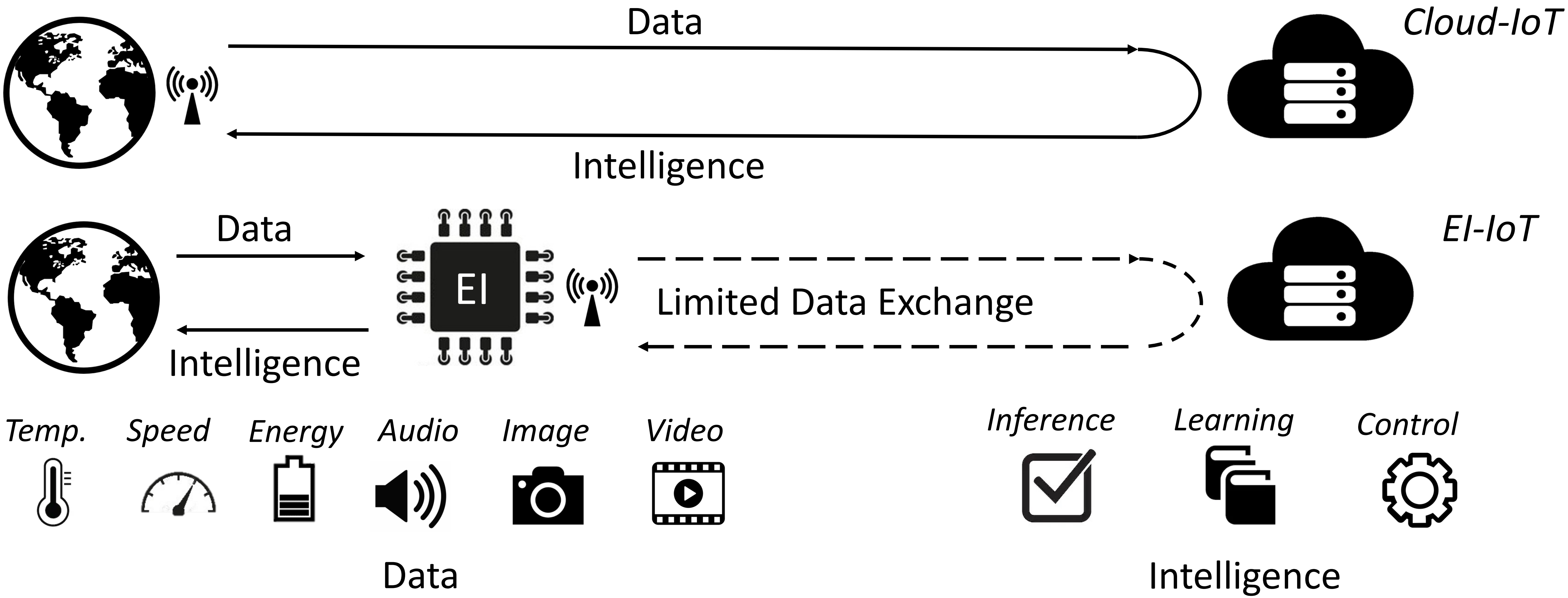
- **Motivation**
- Reinforcement Learning on the Edge
- Swarm Intelligence on the Edge
- Neuro-inspired SLAM on the Edge
- Challenges and Conclusions

Intelligence at the Edge of the Cloud

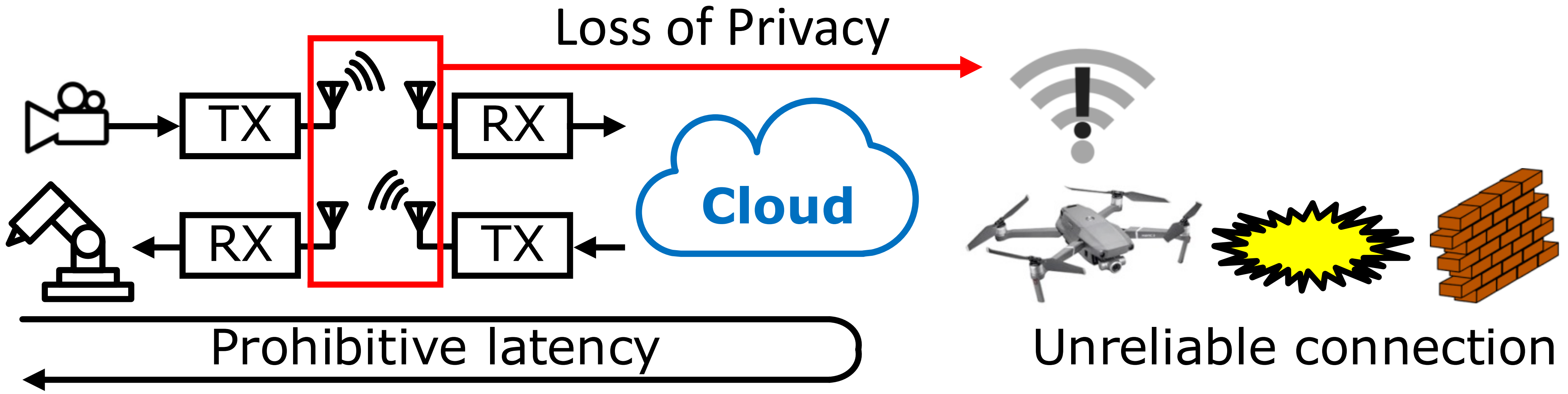




Intelligence at the Edge of the Cloud

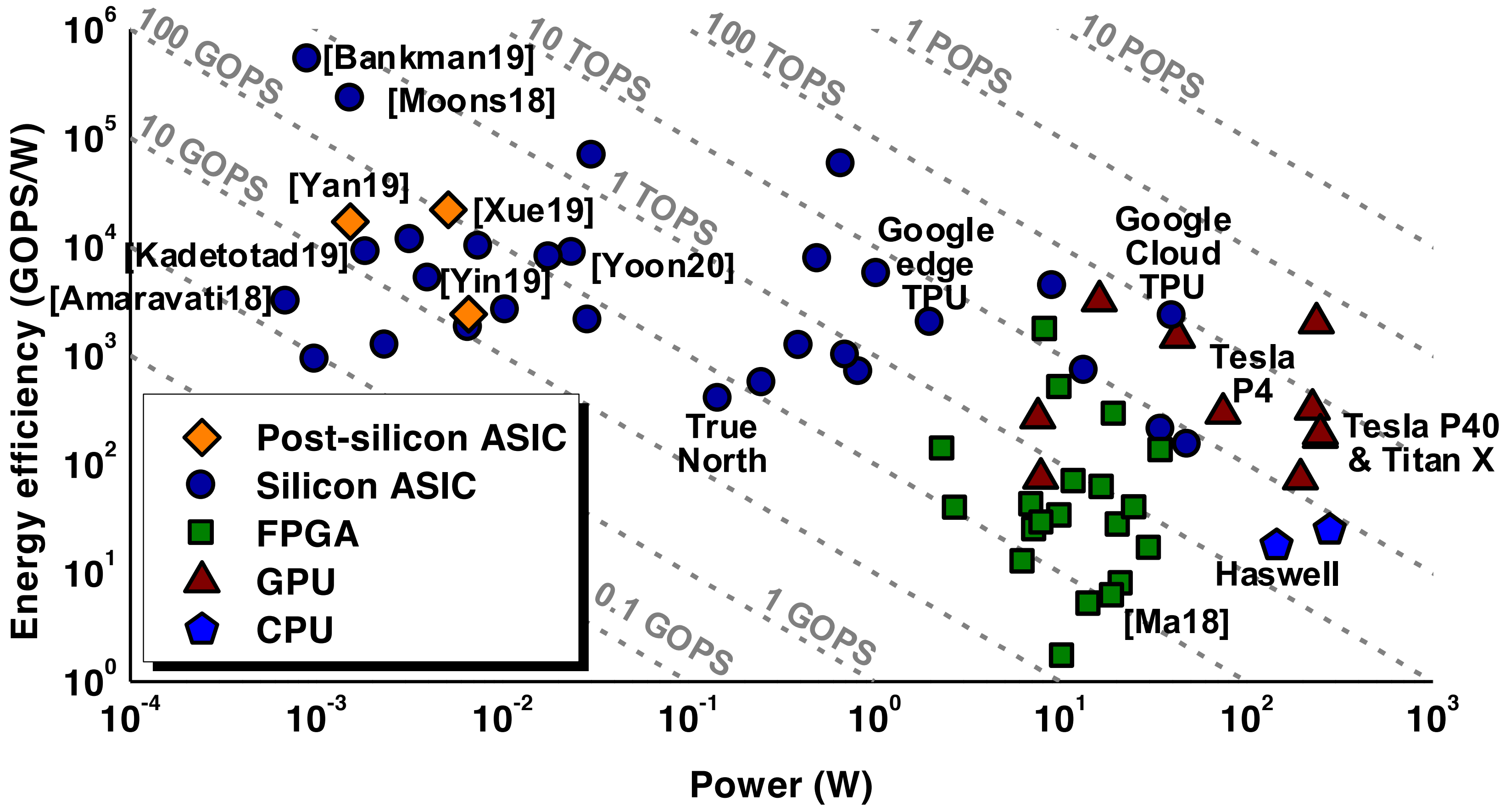


Importance of EI for Autonomous Systems



- Large latency
- Lack of *always-on* communication link
- GPS-denied environments
- Limited privacy for reconnaissance and security related mission

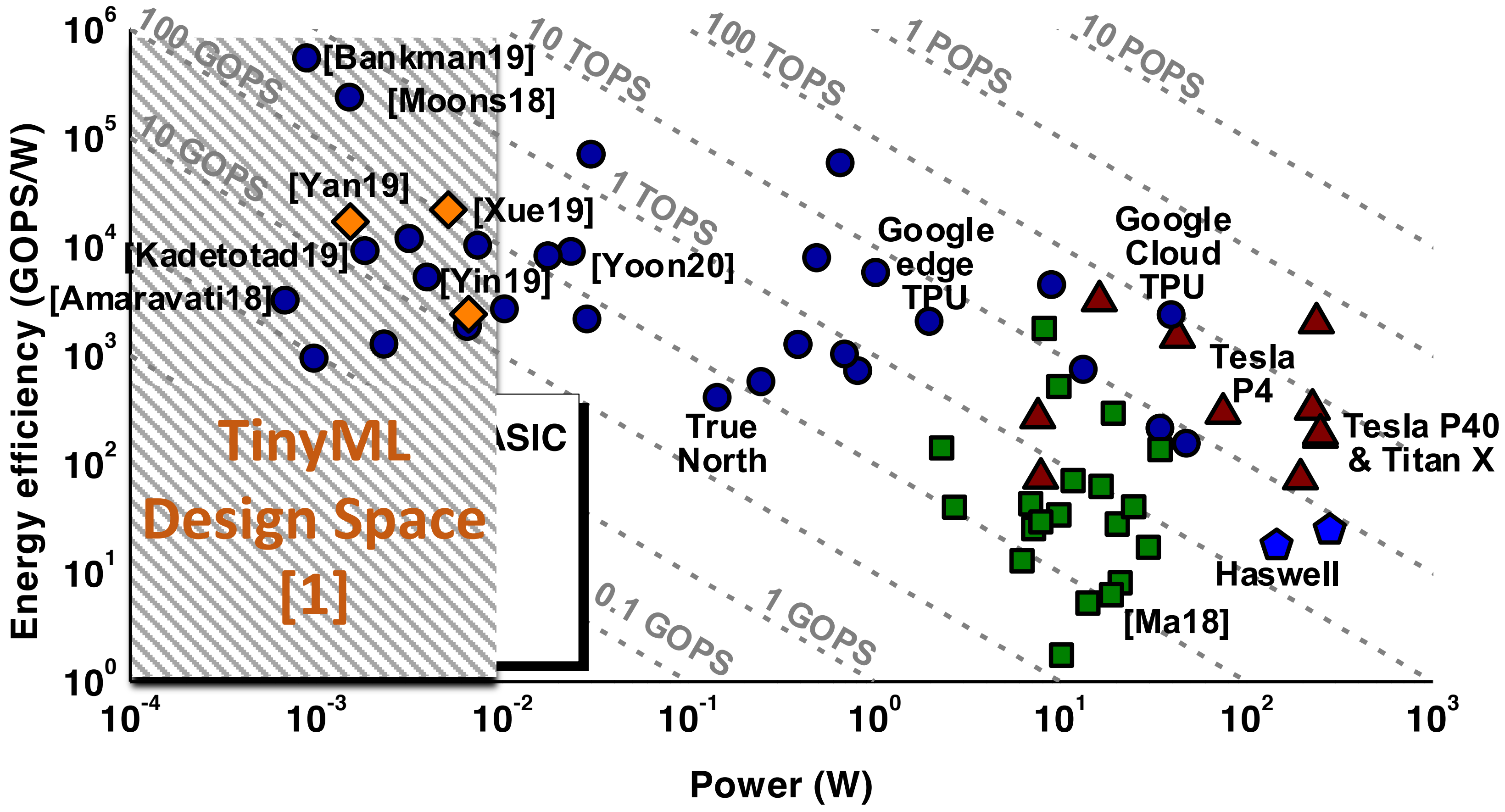
Power-Performance Design Space



- Fully programmable
  - Low efficiency and high power
- High efficiency
  - Low configurability
- Solution for tinyML?

☐ Trade-off between programmability and efficiency

Approaches of TinyML Startups



- Startups in the area
  - Mythic
  - Wave computing
  - Syntiant
  - Eta Compute
  - XNOR.ai
  - ….
- ULP processors
  - Cortex
  - MIPS
  - GPUs
  - …

☐ Ultra low power consumption at target performance

EI and Micro-Robotics

Palm-sized Drones

Intelligent Autonomous Cars

Sensing

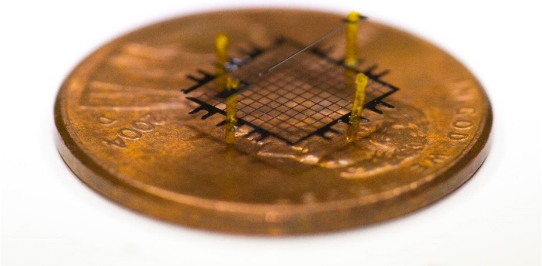
↓

Decision Making

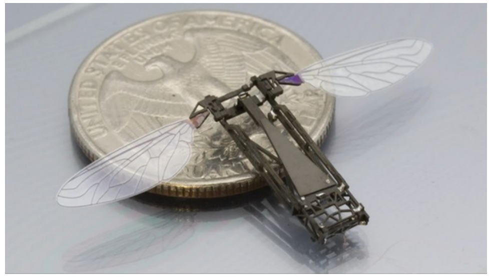
↓

Control

Jasmine microrobots



Berkeley Microrobots

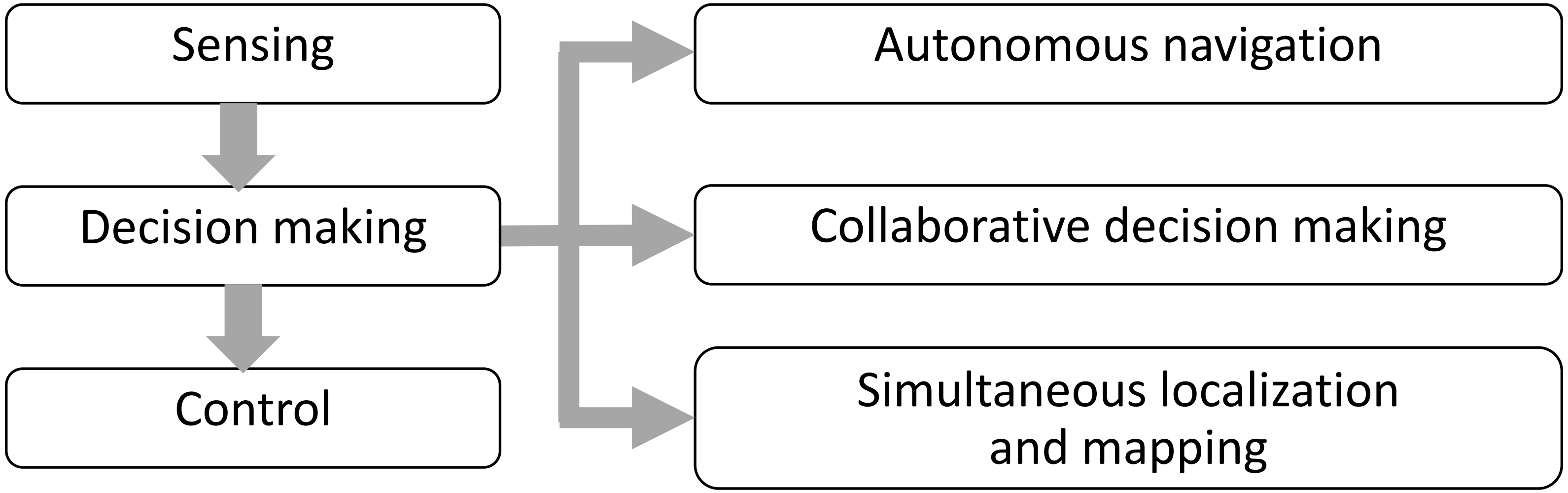


Harvard Bee Microrobots

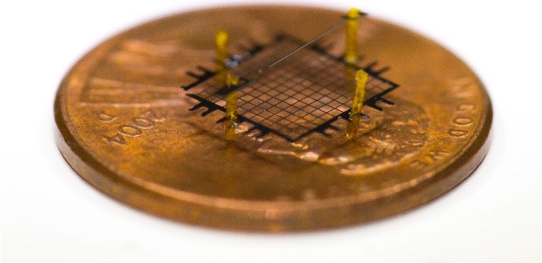
Georgia Tech Microrobot

EI and Micro-Robotics

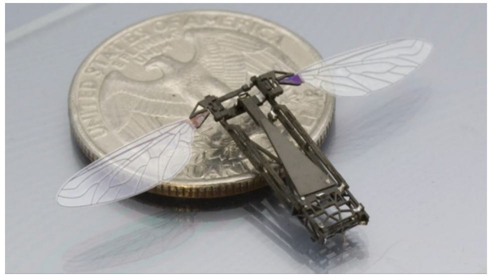
Palm-sized Drones [2]



Jasmine microrobots [4]



Berkeley Microrobots [5]



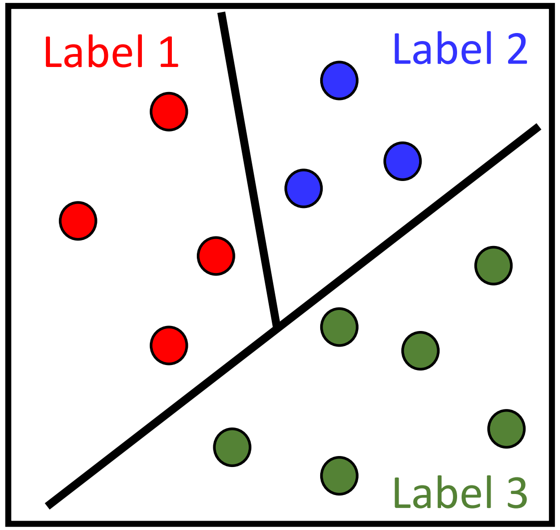
Harvard Bee Microrobots [6]

Georgia Tech Microrobot [7]

Outline

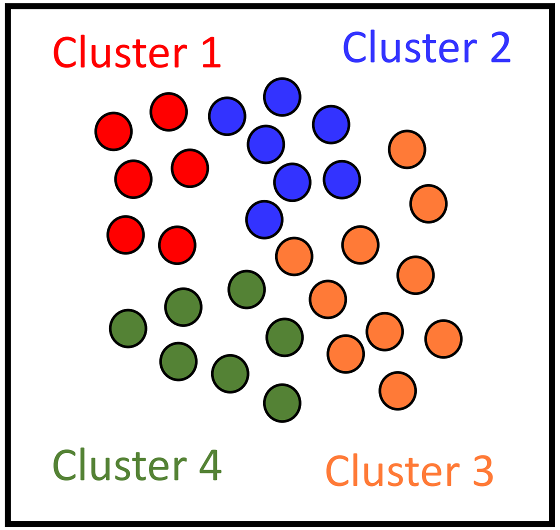
- Motivation
- **Reinforcement Learning on the Edge**
- Swarm Intelligence on the Edge
- Neuro-inspired SLAM on the Edge
- Challenges and Conclusions

Learning with Streaming Data



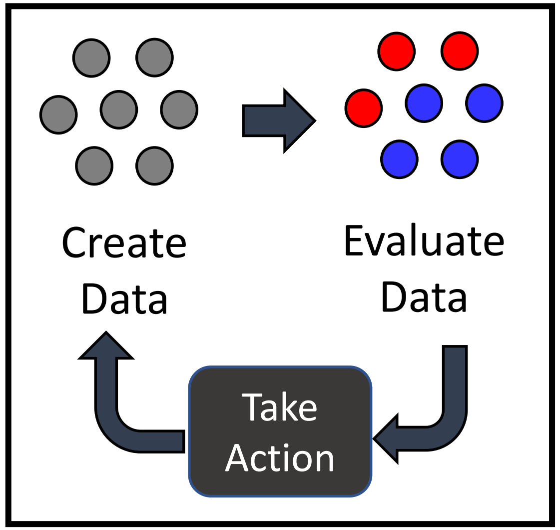
Supervised Learning

- Learning known patterns over labeled data
- Expert supervision required
- Enjoys large success with Deep Neural Networks



Unsupervised Learning

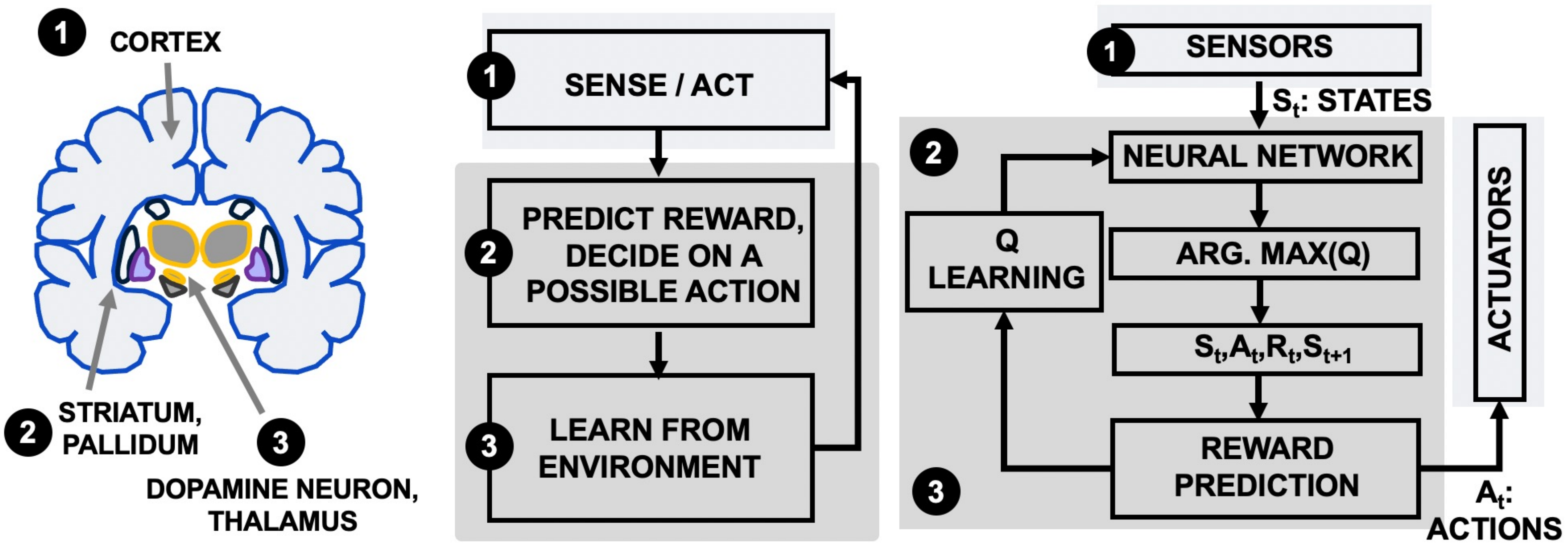
- Learning unknown patterns over unlabeled data
- No supervision required
- Creates clusters on high-dimensional data



Reinforcement Learning

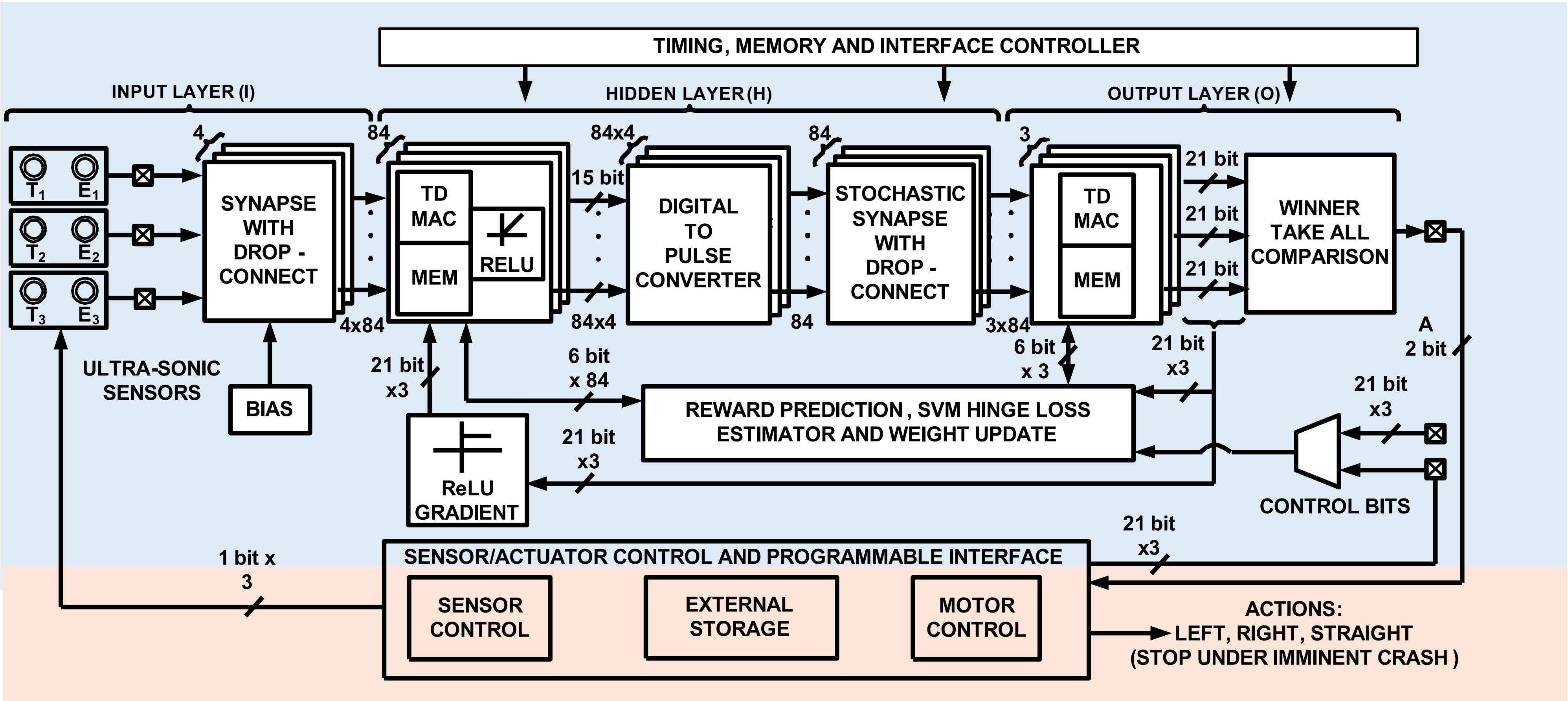
- Generating data through exploration
- Gathering and exploiting knowledge
- Fully autonomous [8]

Providing Autonomy to Edge Devices



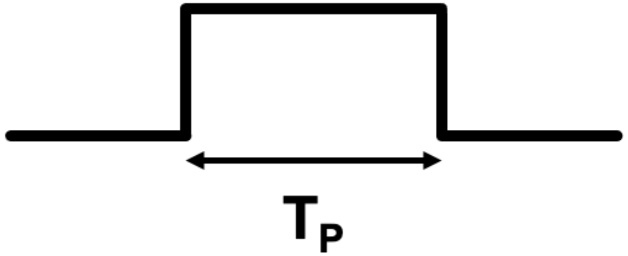
- Reinforcement Learning can maximize a set reward through exploration of the state-space and taking actions.
- A neural network maps the state-space to the action space optimally.

Time-Based Design for Online RL



[ISSCC18, JSSC19]

Processing with Time-Encoded Pulses



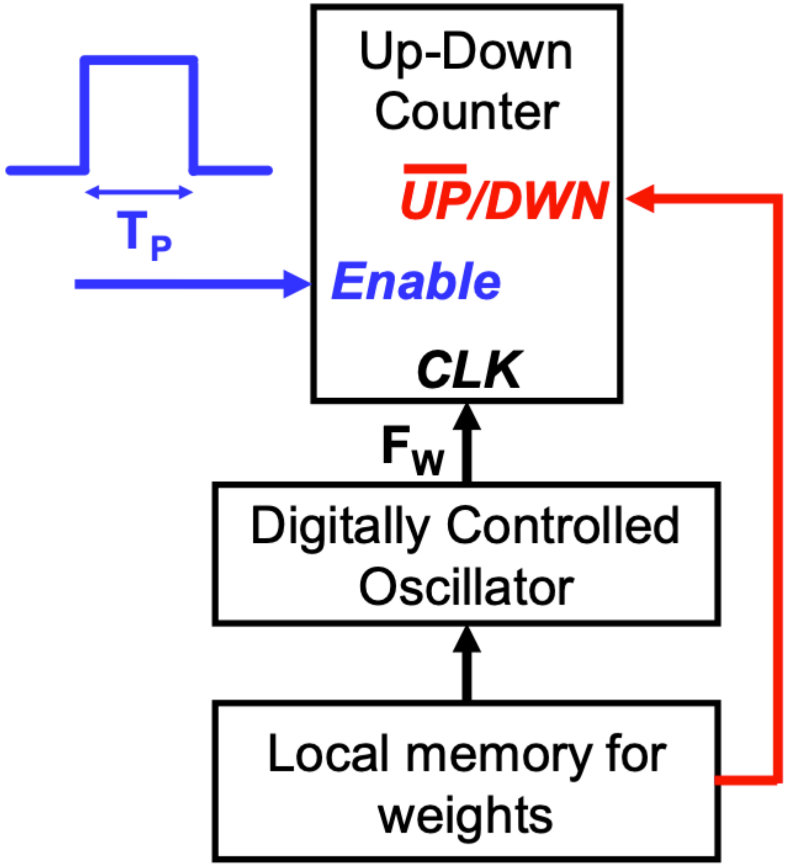
$$T_P = X.\ T_0$$

Operand   Reference Time Constant

$$F_W = W.\ F_0$$

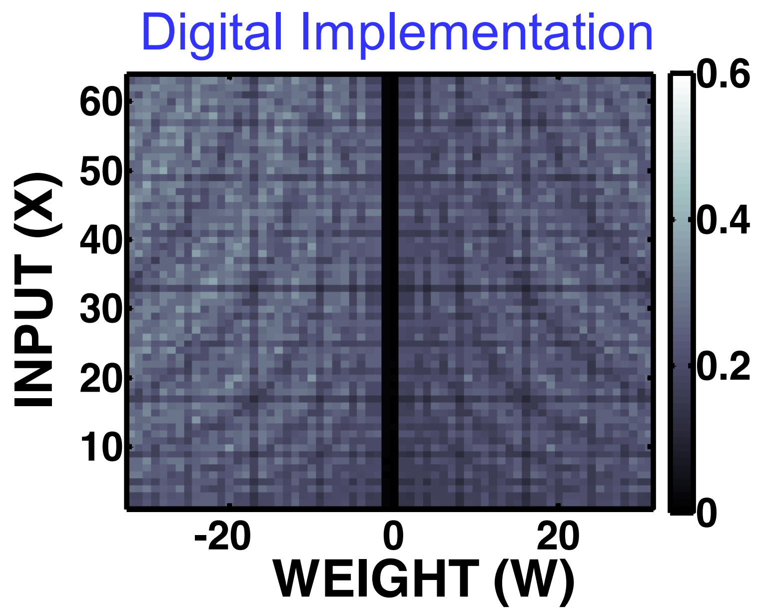
$$Y = T_P\ .\ F_W = X.W.(T_0.F_0)$$

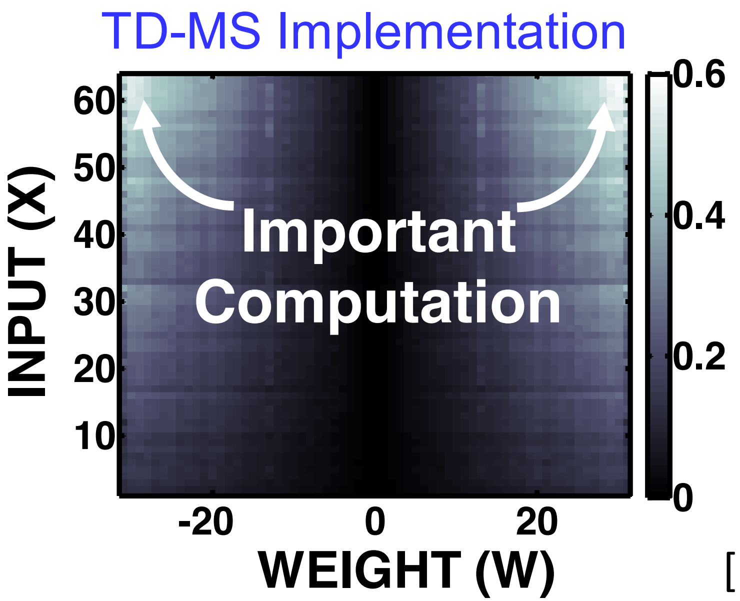
$$Y = T_0.F_0\ \Sigma\ X_i.W_i$$



[ISSCC18, JSSC19]

Energy Efficiency of Time-Domain Processing



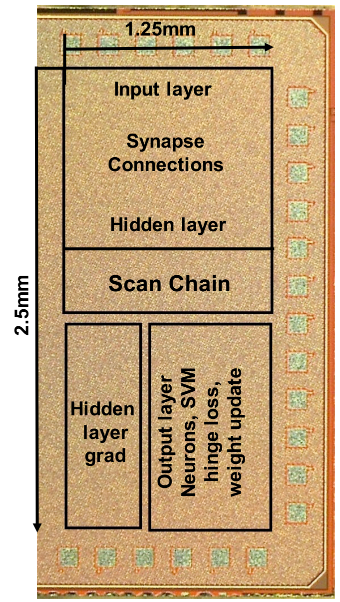


TD-MS: time-domain mixed-signal

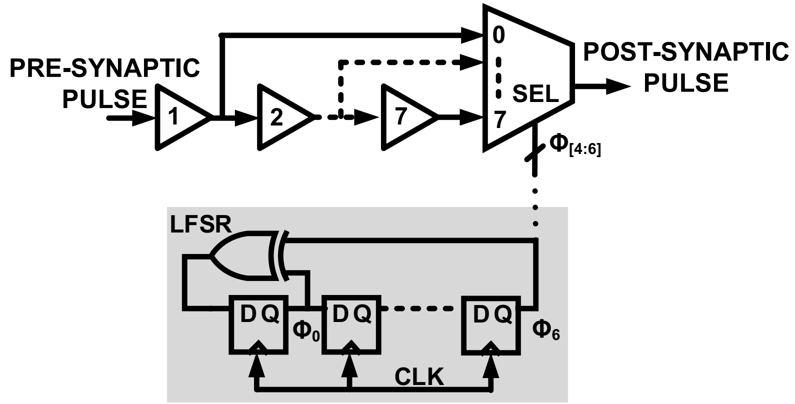
[ISSCC18, JSSC19]

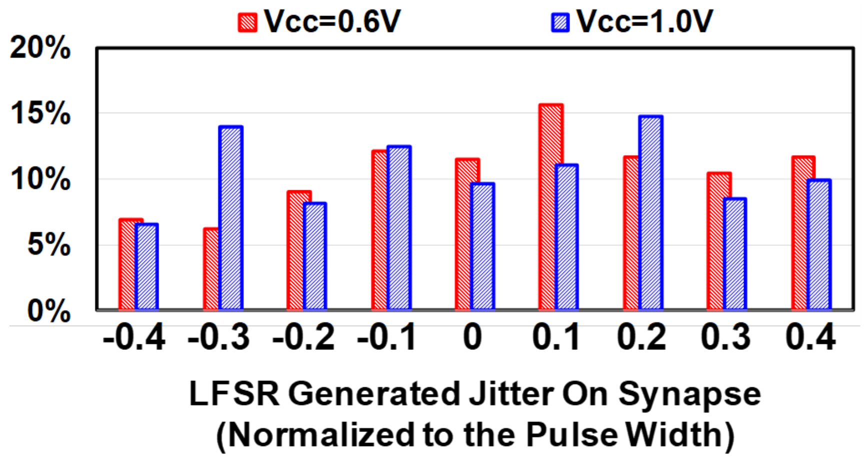
- Number of switching events (and hence, energy/op) in TD neuron is proportional to the value of the operands (and hence, the importance of the computation)
- Bio-mimetic and takes advantage of inherent sparsity in the network
- An average of 42% reduction in energy/op
- 45% lower area, 47% lower interconnect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

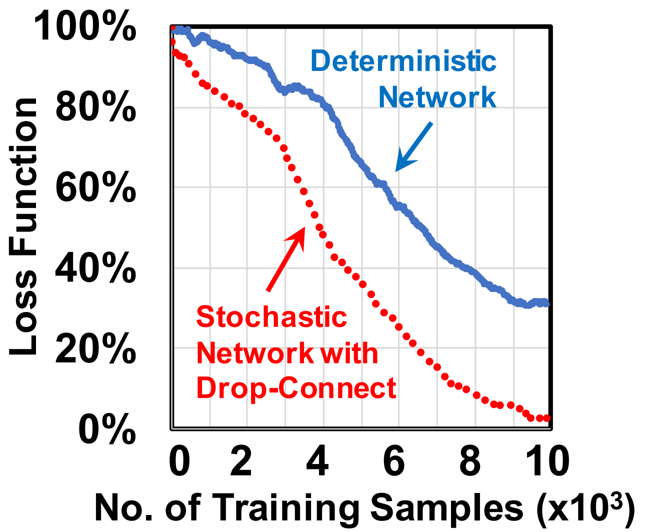
Enabling Regularization via Stochasticity



55nm 1P8M CMOS
1.2*2.5mm
QFN package

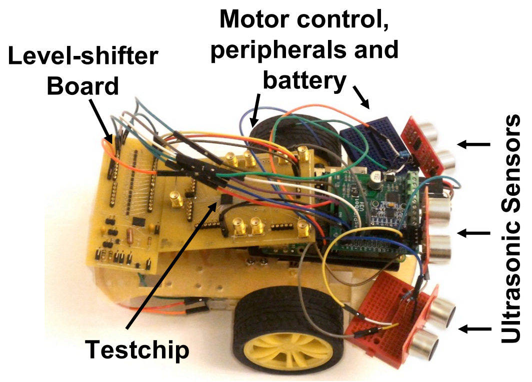


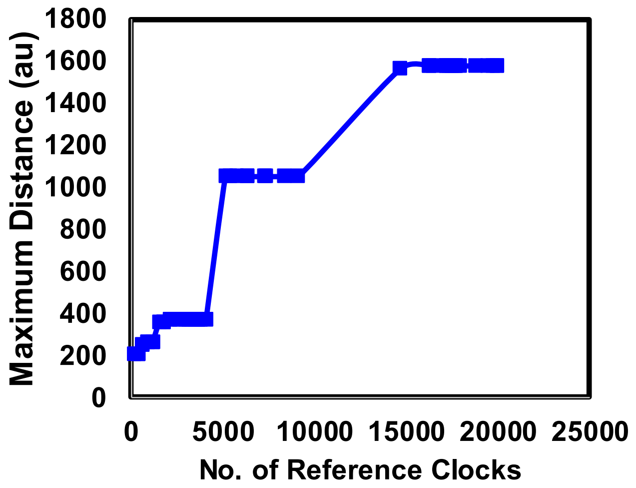


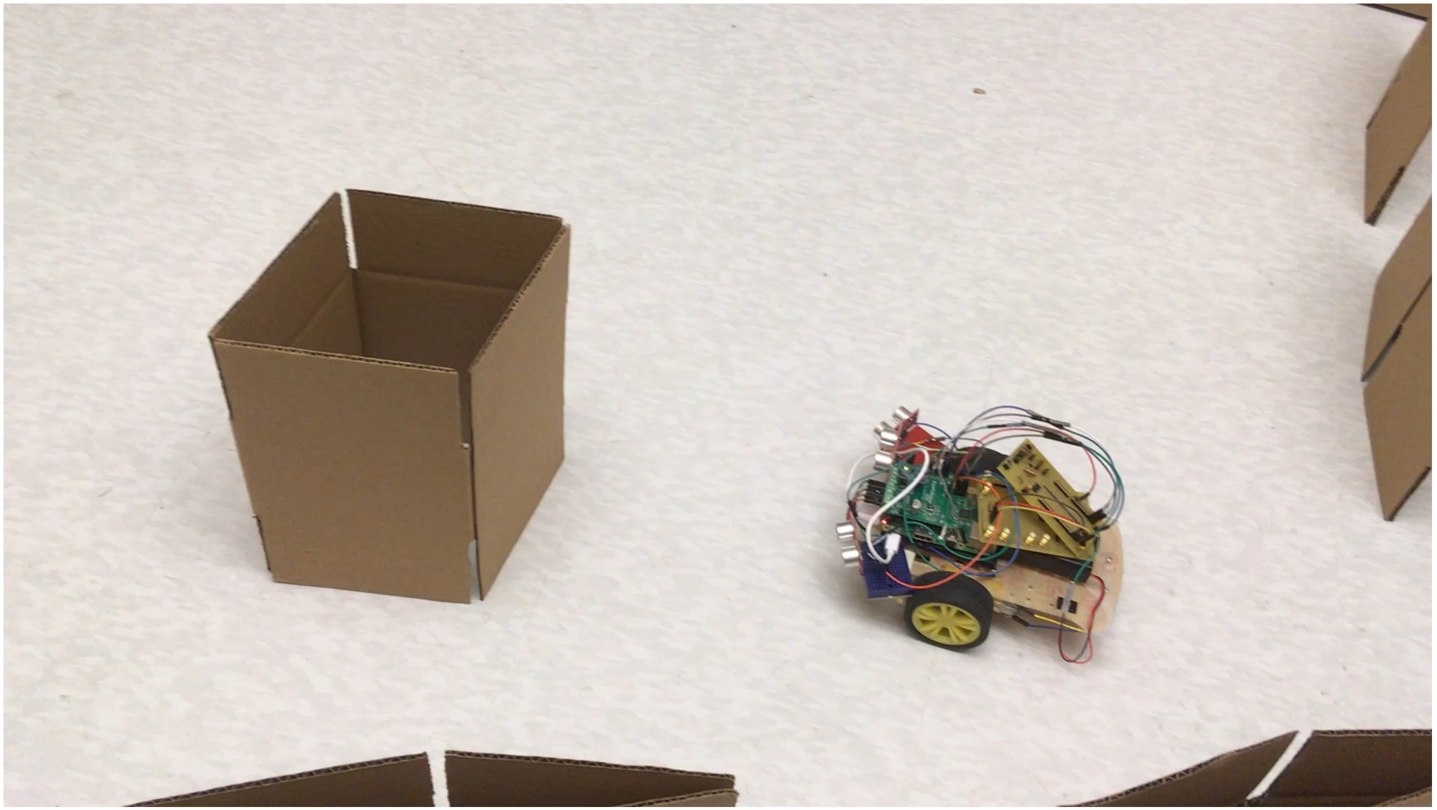


- ❑ Measured stochasticity of ±40% for a mean pulse width is measured
- ❑ Stochasticity in the synaptic weights allow the system to achieve convergence faster for a prototypical robotic application

RL Chip in Action







[ISSCC18, JSSC19]

Outline

- Motivation
- Reinforcement Learning on the Edge
- **Swarm Intelligence on the Edge**
- Neuro-inspired SLAM on the Edge
- 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

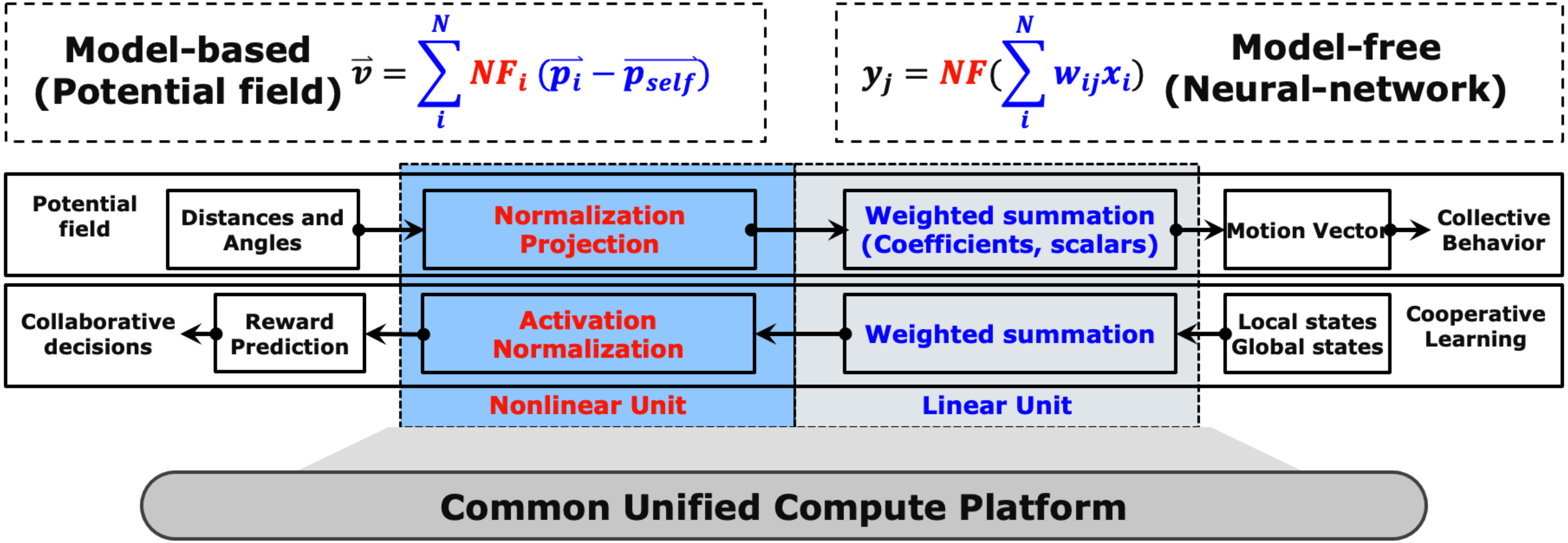
Pattern-formation

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	

A Common Platform to Support Swarm EI



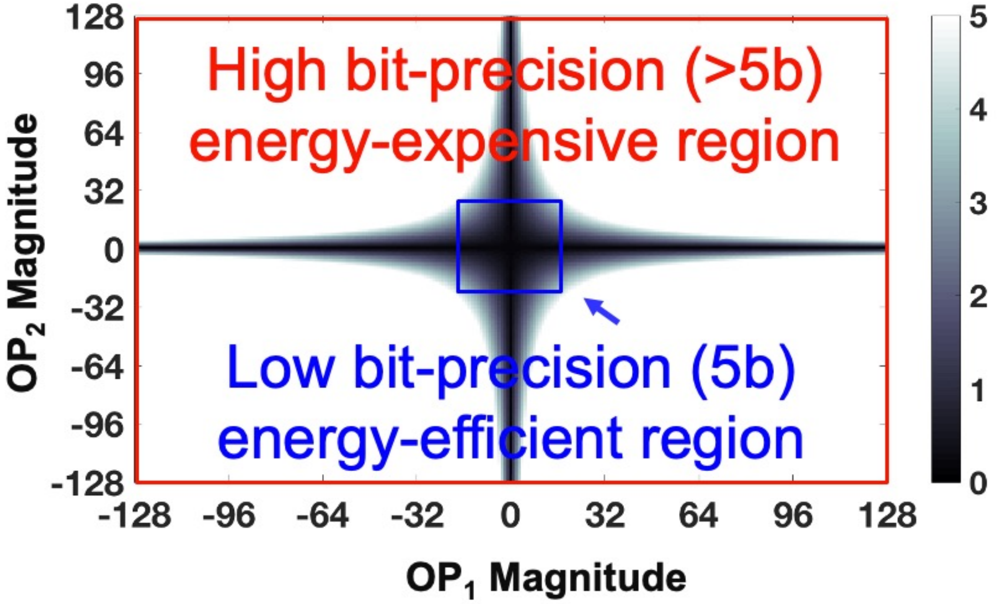
- Unified compute platform with dedicated nonlinear / linear unit for both model-based and learning-based swarm applications.

[ISSCC19, JSSC19]

Swarm Size vs Bit-Width Requirement

Swarm size	Algorithms			
	Path-planning	Formation	Predator-prey	Cooperative exploration
2	3	3	5	4
5	4	4	7	4
10	5	5	7	5
15	5	6	8	5
20	6	7	8	6

Required bit-precision vs. Swarm size

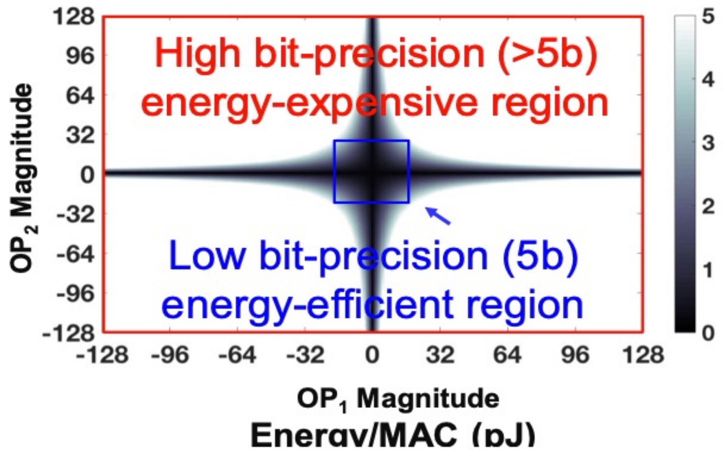


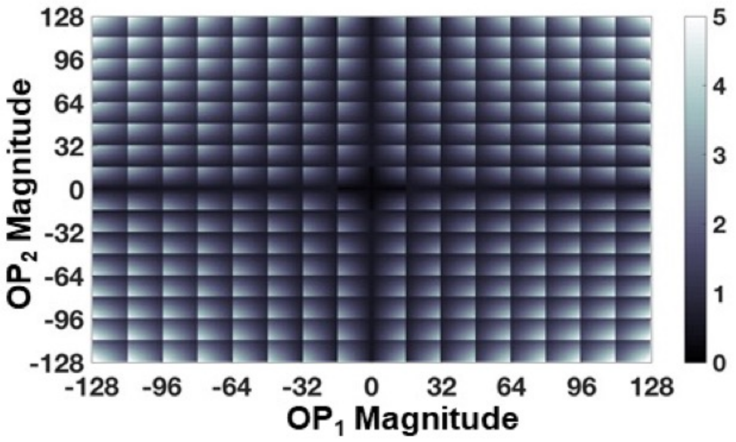
Energy/MAC (pJ)

- ❑ Increasing swarm size requires higher bit-precision
- ❑ Energy efficiency of TD-MS decreases at higher bit-precisions

[ISSCC19, JSSC19]

From TD-MS to Hybrid Designs





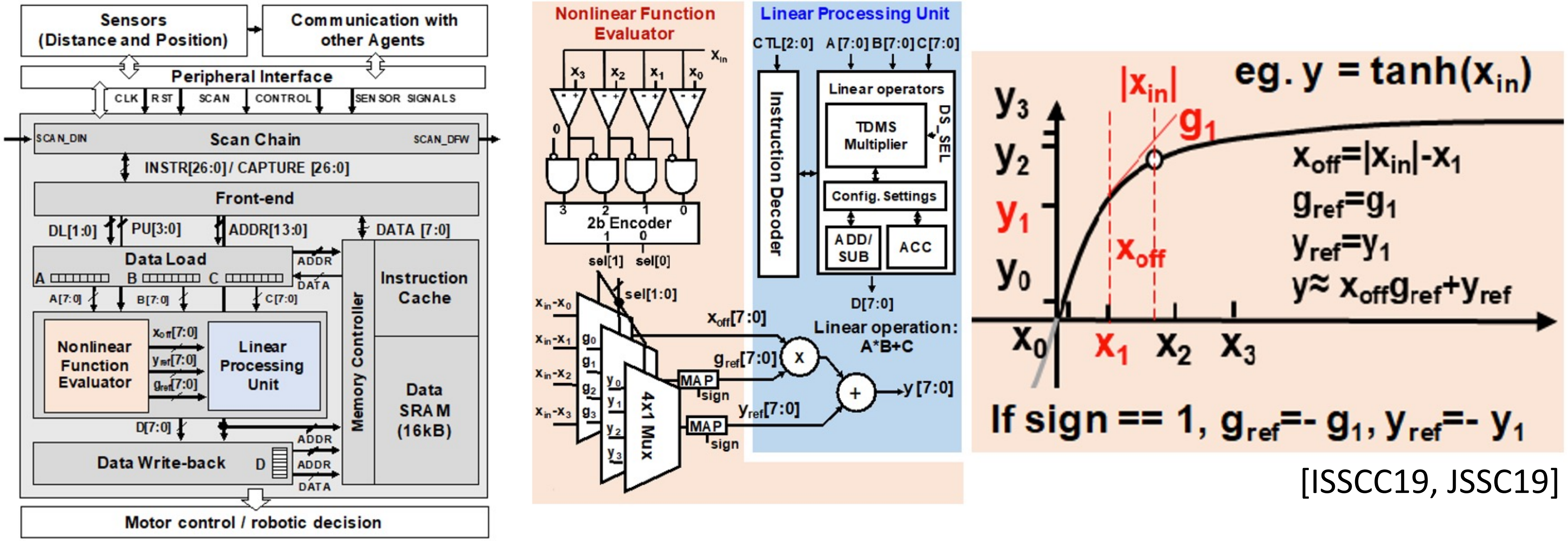
No. Bits	TD-MS		HDMS	
	Average	Worst	Average	Worst
3	0.10	0.49	0.16	0.52
4	0.14	0.56	0.19	0.61
5	0.28	0.72	0.29	0.74
6	0.64	1.74	0.69	0.94
7	2.21	3.86	0.70	1.02
8	5.82	9.32	0.69	1.27

Energy/MAC (Normalized to Digital)

- ❑ Analog techniques are energy-efficient for low bit-widths
- ❑ Smarter designs are required when bit-widths need to scale
- ❑ The break-even point between digital and analog compute is around 5-6 bits

[ISSCC19, JSSC19]

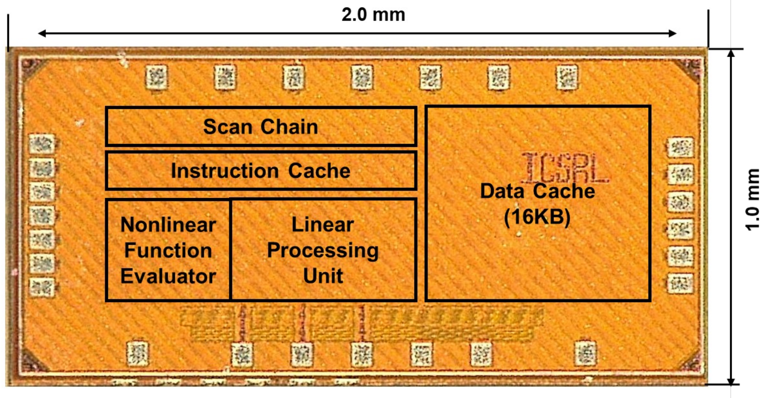
System Architecture

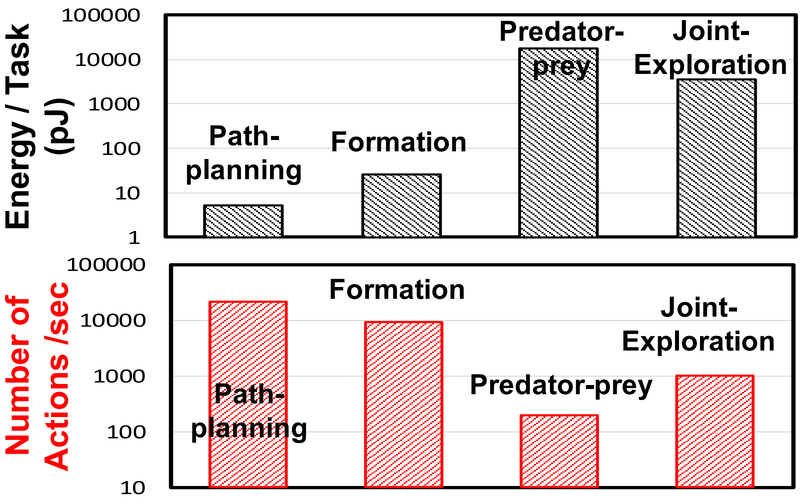


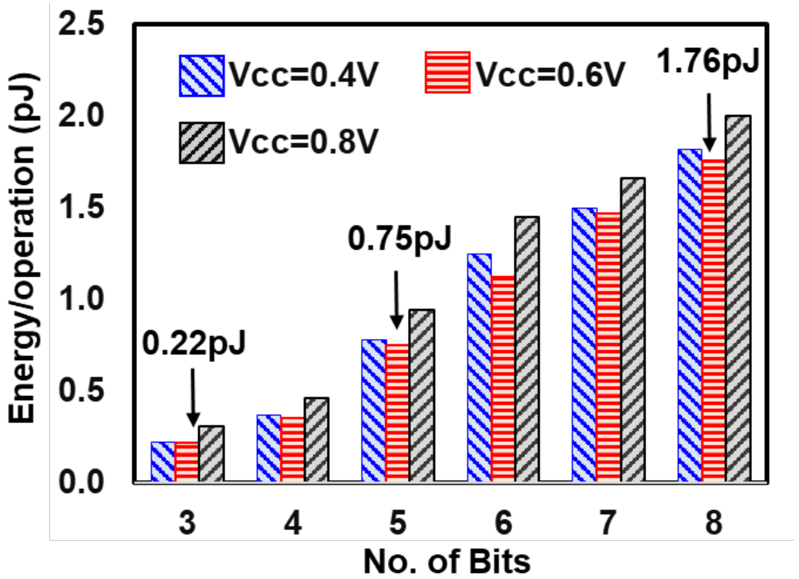
[ISSCC19, JSSC19]

- A unified processor that can provide scalability as well as support for model-based and learning-based tasks
- It provides high efficiency across a wide range of program and environmental settings

65nm Test-Chip and Measured Results





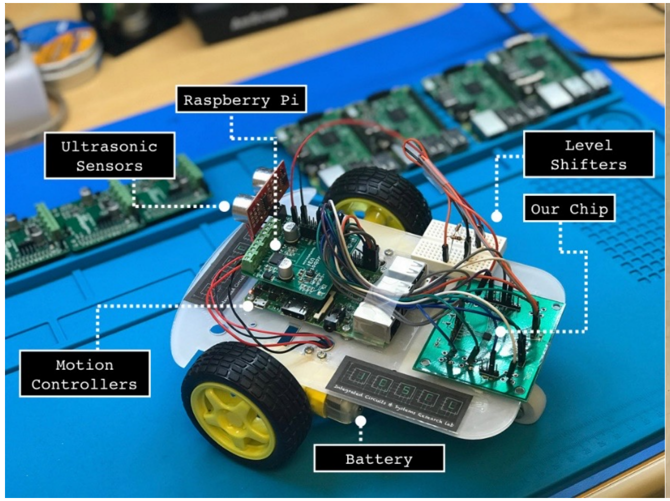


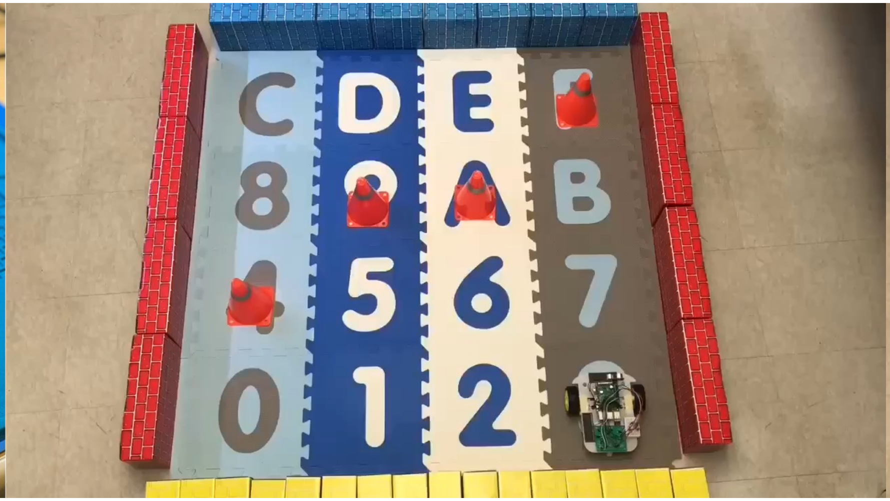
Energy/MAC for varying bit-width

- ❑ 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

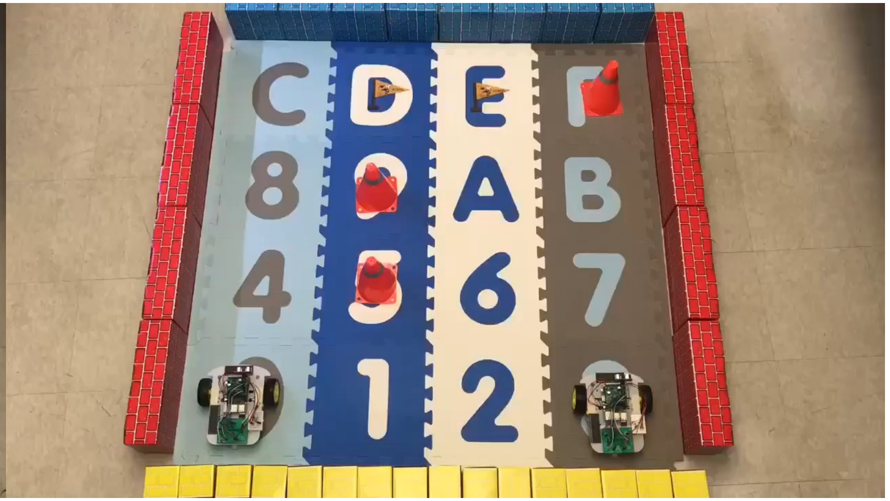
[ISSCC19, JSSC19]

Swarm Intelligence in Action





Exploration 16X real time



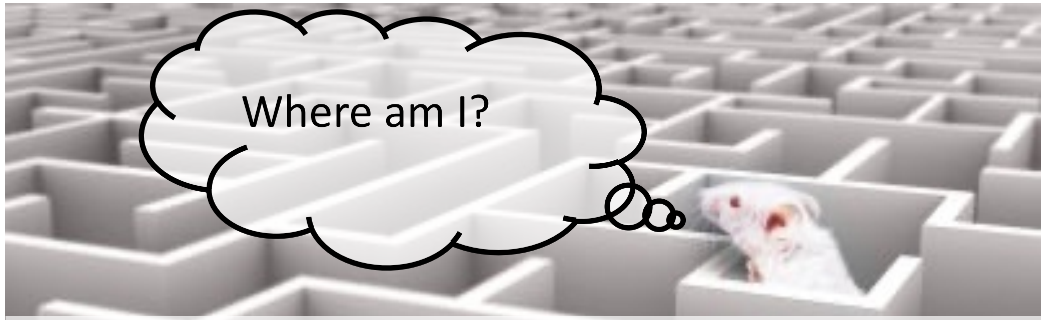
Collaborative RL in real time

[ISSCC19, JSSC19]

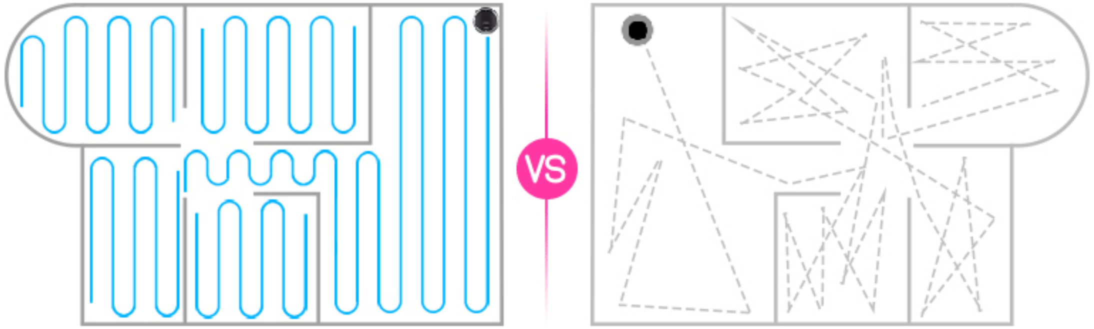
Outline

- Motivation
- Reinforcement Learning on the Edge
- Swarm Intelligence on the Edge
- **Neuro-inspired SLAM on the Edge**
- Challenges and Conclusions

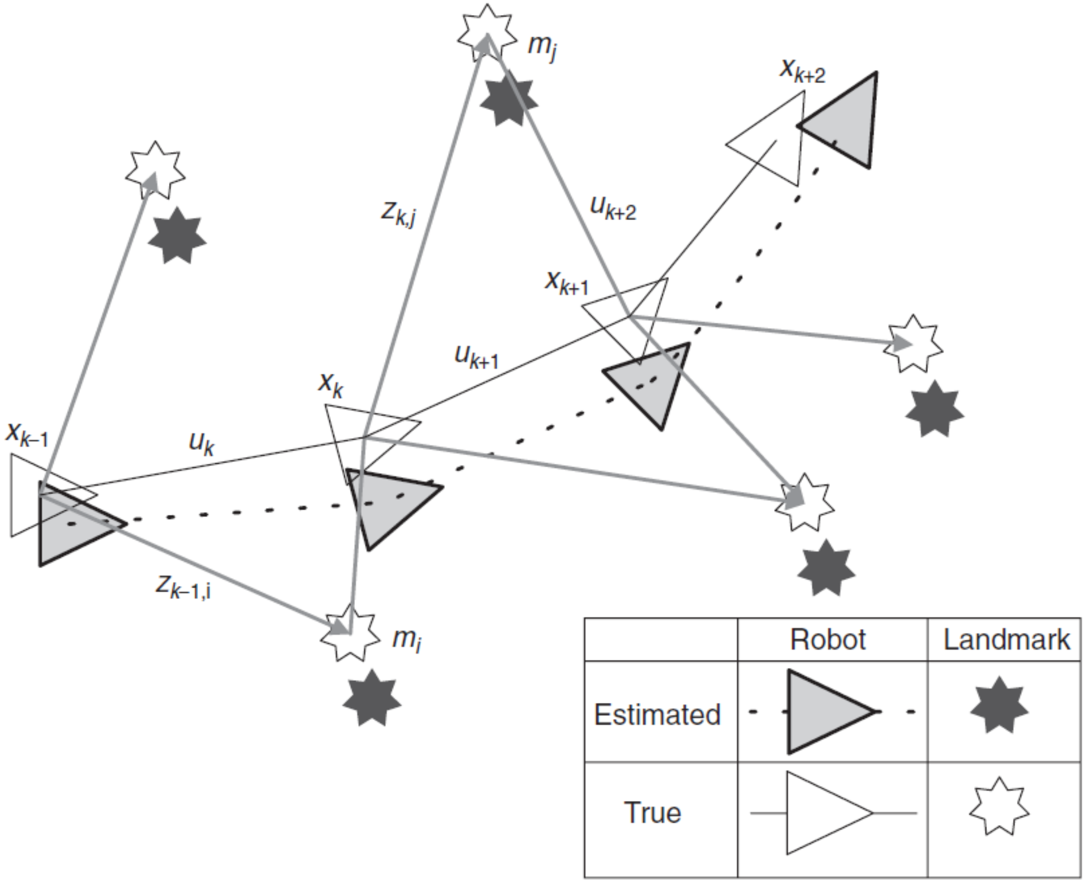
Simultaneous Localization and Mapping (SLAM)



Exploration in unknown environments while identifying one's own position



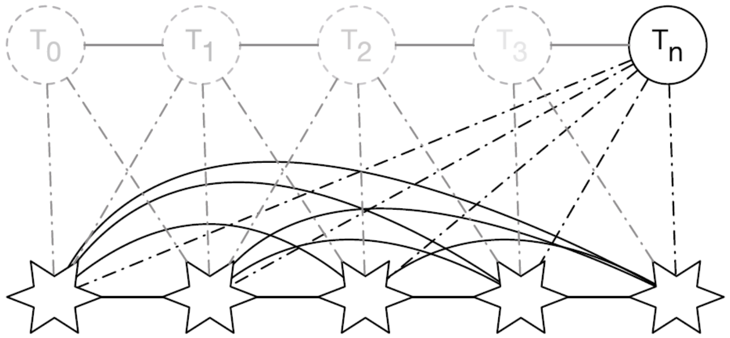
Path planning considering the result of SLAM



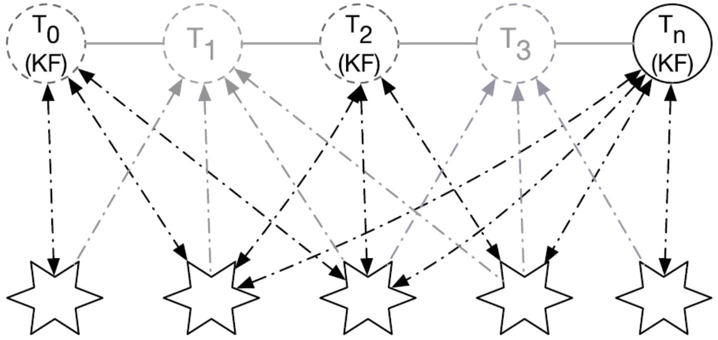
Landmark-based pose estimation

SLAM Algorithms Towards Edge-AI

	Probabilistic SLAM	Keyframe-based SLAM	NeuroSLAM
Algorithm of data association	Direct method, feature-based method (SIFT, SURF, CNN, etc.)		Direct method with maxpooled images & SNN-based pose-cell activities
Sensor	Monocular, stereo, RGB-D camera, etc.		Monocular camera
Odometry	Visual odometry, inertia sensor		Visual odometry
Map maintenance	Every frame	Keyframe	VT-matched frame
Application	High-performance AR, VR, UAVs		Ultra low power Microrobotics



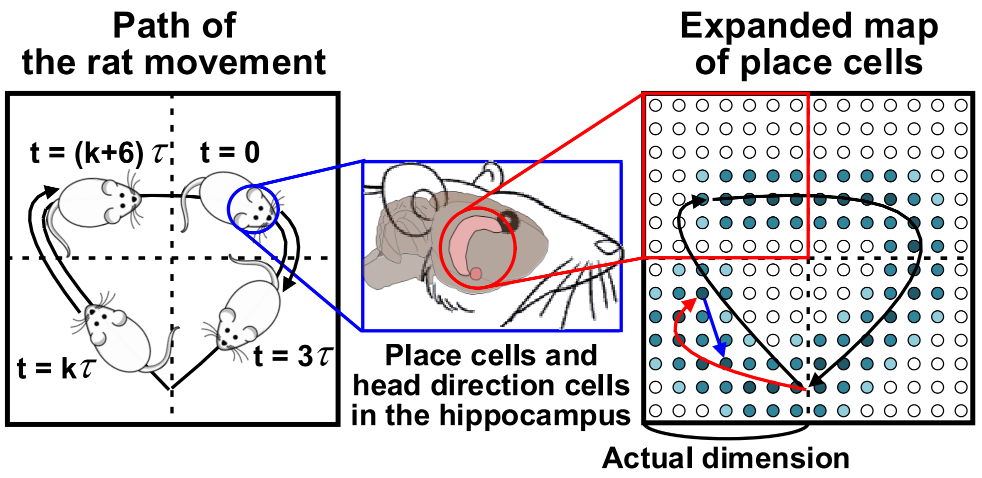
Probablistic SLAM

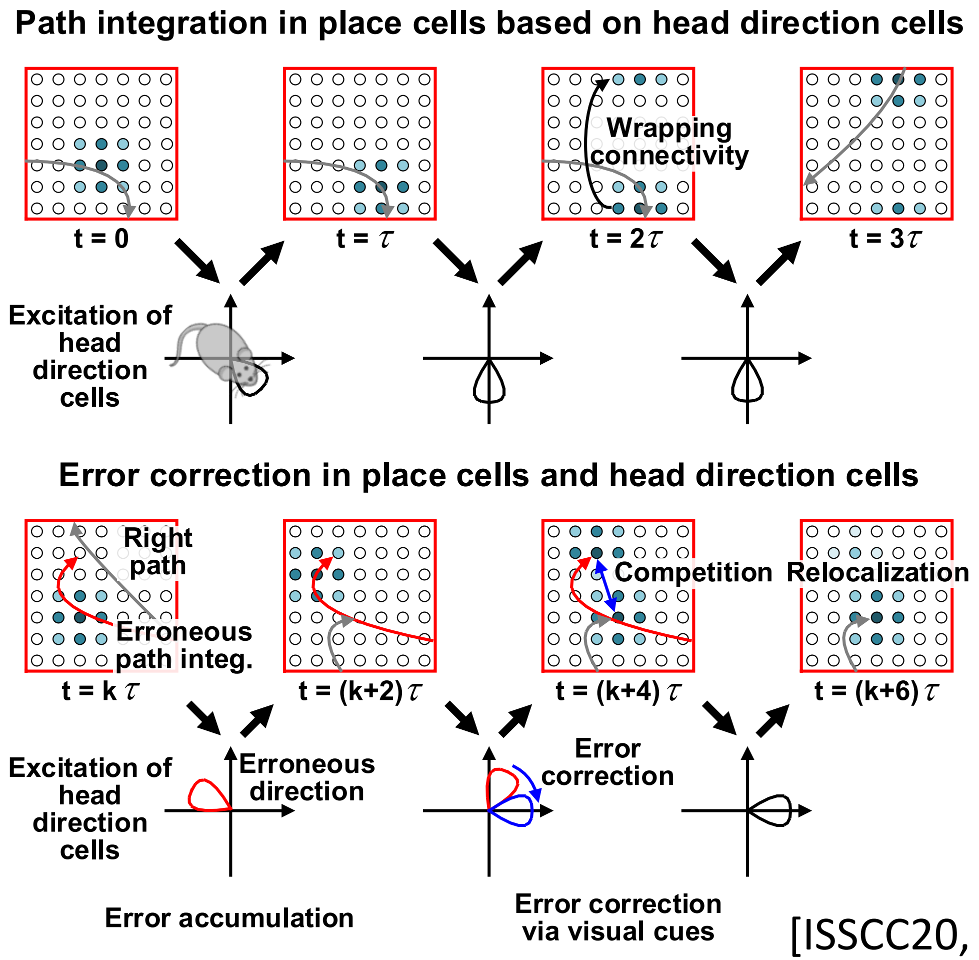


Keyframe-based SLAM

- ☐ Smaller number of computations in a frame
- ☐ Map maintenance in a certain frame, not every frame

Spatial Cognition in the Rodent Brain

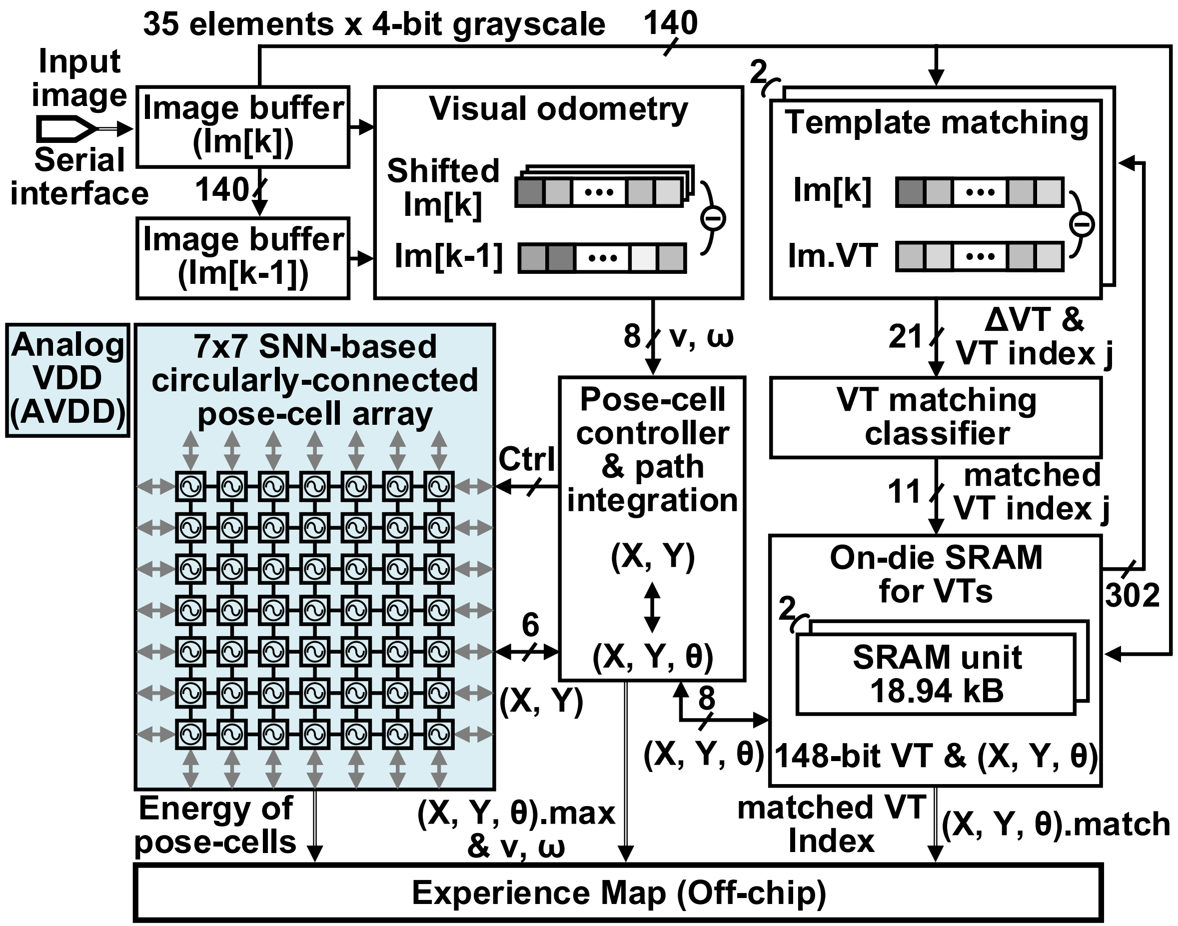




[ISSCC20, JSSC20]

- 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

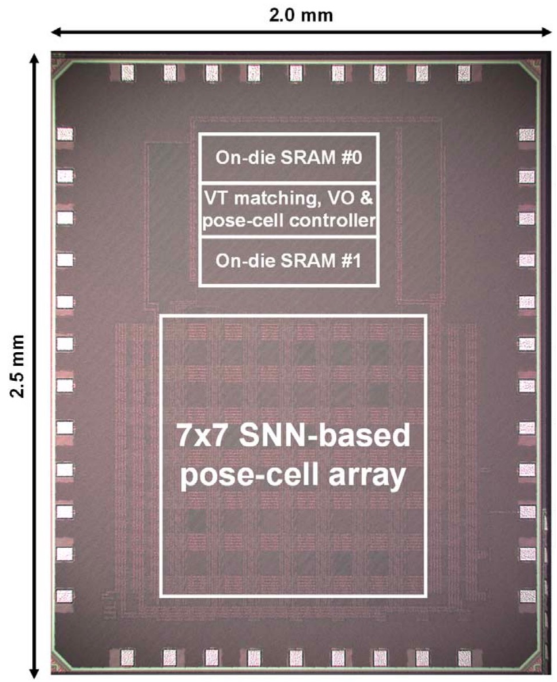
NeuroSLAM Accelerator



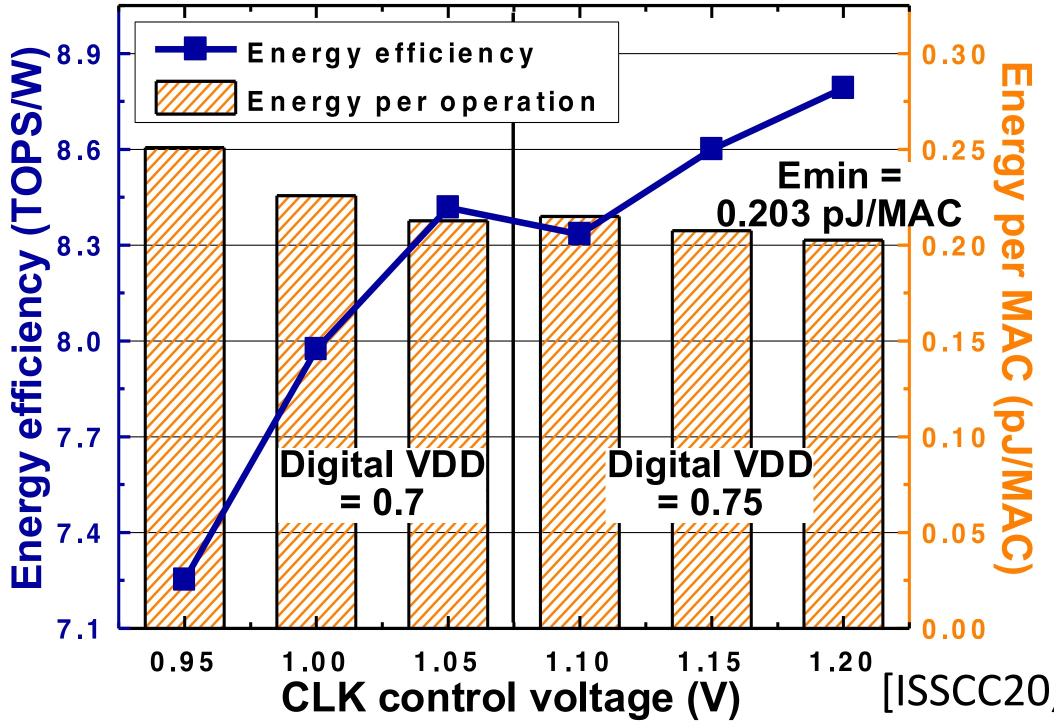
- Mixed-signal oscillator-based NeuroSLAM accelerator
- Spiking neural network-based pose-cell array enables power-efficient SLAM operation
- Competition between visual cues and self-motion allows an autonomous agent to perform loop closure
- This is a continuous time dynamical system implementing a SNN version of RatSLAM

[ISSCC20, JSSC20]

Measured Results on 65nm Test-chip



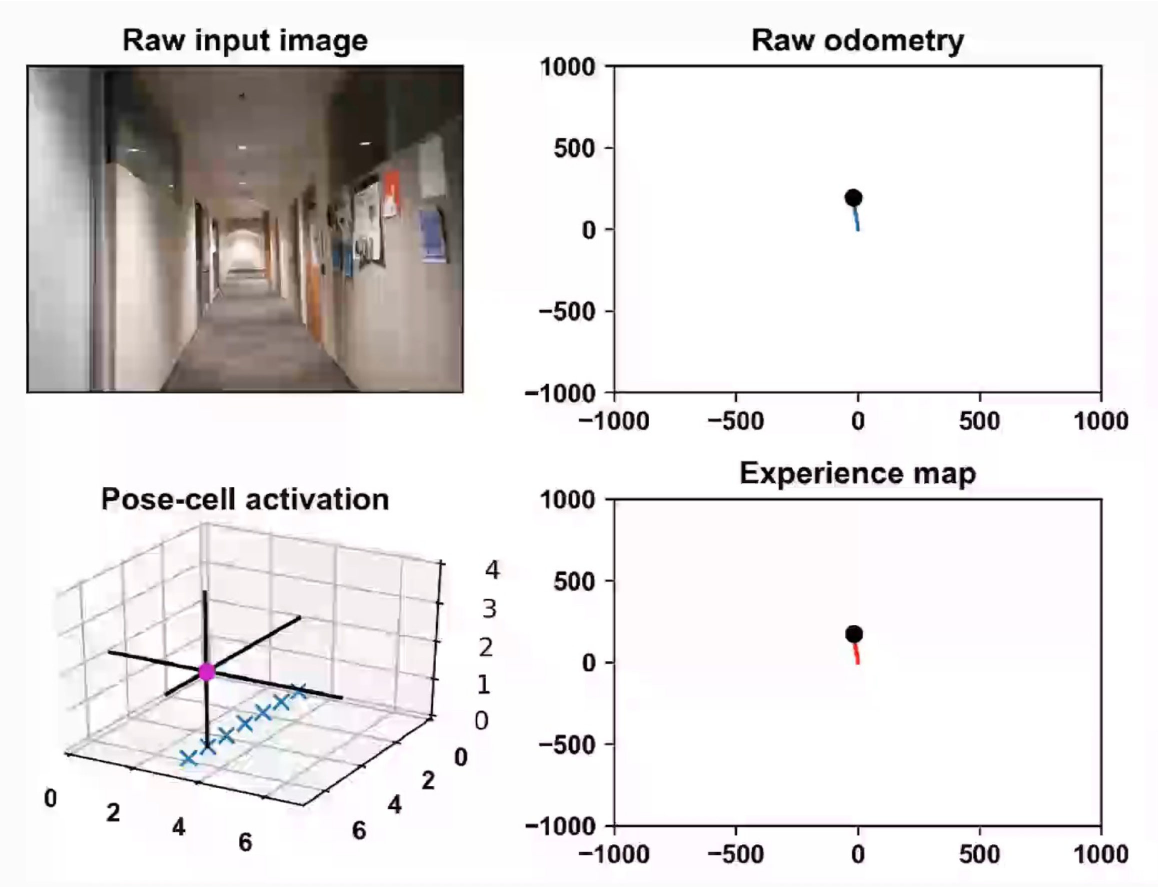
Technology	65 nm 1P9M CMOS
Die area	2.0 mm x 2.5 mm
On-chip memory	37.9 kB
Frequency	78.22-130.8 MHz
Digital VDD	0.7-0.75 V
Analog VDD	0.95-1.2 V
I/O VDD	2.5 V
Power	17.27-23.82 mW
Energy efficiency	7.25-8.79 TOPS/W
Package	QFN48

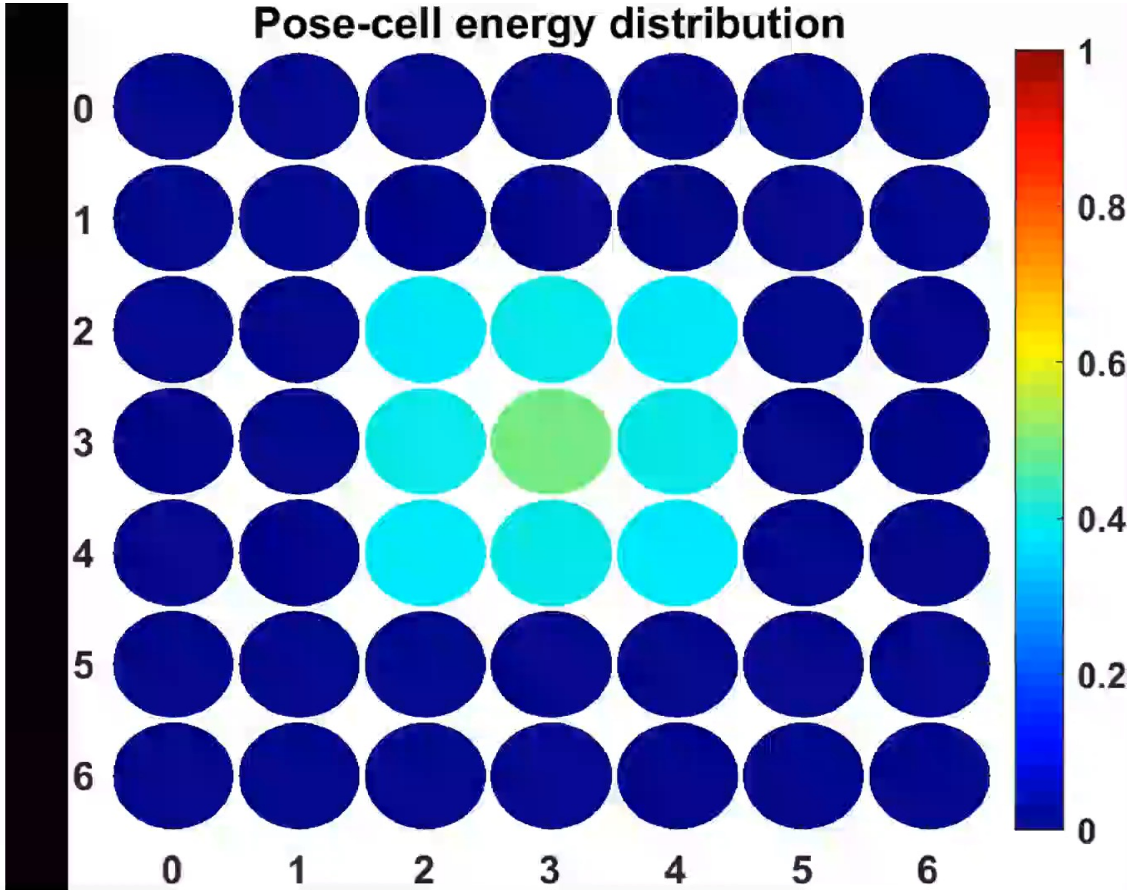


[ISSCC20, JSSC20]

- 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 streaming input frames [ISSCC20, JSSC20]

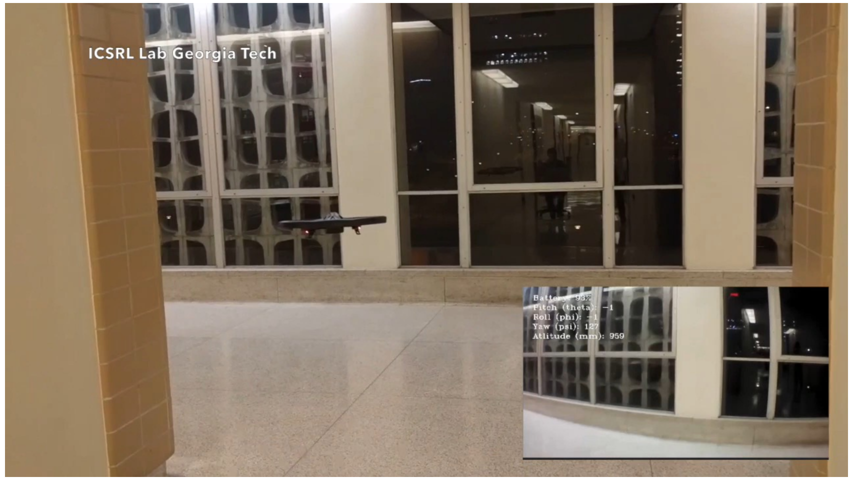
Benchmarking and Software Infrastructure

- Simulation of actual physics of motion and flight
- VR frontend with ML backend
- Rich set of virtual worlds including indoor and outdoor environments
- End-to-end infrastructure from VR to Tensor Flow and python APIs

Transfer Learning

- Trained models can be deployed to the real world
- Limited Training on real world is required
- Enables end-to-end benchmarking

https://github.com/aqeelanwar/DRLwithTL



[DATE19]
[DAC21]

Outline

- Motivation
- Reinforcement Learning on the Edge
- Swarm Intelligence on the Edge
- Neuro-inspired SLAM on the Edge
- **Challenges and Conclusions**

Challenges and Opportunities

- Device and Circuit level:
  - Embedded non-volatile memory (e.g., RRAM, STT-MRAM, FeRAM, etc)
  - 3d integration
- Architecture level:
  - Be adaptive and reconfigurable to various scenarios and applications
- System level:
  - Holistic benchmark and generic hardware platform
- Algorithm level:
  - Lifelong learning, learning with limited data
  - Effective and robust swarm learning

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…!!

References

[DAC21] Z. Wan, A. Anwar, Y. Hsiao, T. Jia, V. Reddi, A. Raychowdhury. "Analyzing and Improving Fault Tolerance of Learning-Based Navigation Systems." In *58th ACM/IEEE Design Automation Conference (DAC)*, pp. 841-846. IEEE, 2021.

[ISSCC20] J. Yoon and A. Raychowdhury, “A 65nm 8.79TOPS/W 23.82mW Mixed-Signal Oscillator-Based NeuroSLAM Accelerator for Applications in Edge Robotic,” in IEEE International Solid-State Circuits Conference, 2020.

[JSSC20] J. Yoon and A. Raychowdhury, “NeuroSLAM: A 65-nm 7.25-to-8.79-TOPS/W Mixed-Signal Oscillator-Based SLAM Accelerator for Edge Robotics,” in *IEEE Journal of Solid-State Circuits*, *56*(1), 66-78, 2020

[ISSCC19] N. Cao, M. Chang, and A. Raychowdhury, “14.1 A 65nm 1.1-to-9.1TOPS/W Hybrid-Digital-Mixed-Signal Computing Platform for Accelerating Model-Based and Model-Free Swarm Robotics,” in 2019 IEEE International Solid- State Circuits Conference, 2019.

[JSSC19] N. Cao, M. Chang, and A. Raychowdhury, “A 65-nm 8-to-3-b 1.0–0.36-V 9.1–1.1-TOPS/W hybrid-digital-mixed-signal computing platform for accelerating swarm robotics,” *IEEE Journal of Solid-State Circuits*, *55*(1), pp.49-59, 2019

References

[JSSC19] A. Amravati, S. Nasir, S. Thangadurai, I. Yoon, A. Raychowdhury, “A 55-nm, 1.0–0.4 v, 1.25-pj/mac time-domain mixed-signal neuromorphic accelerator with stochastic synapses for reinforcement learning in autonomous mobile robots.” in *IEEE Journal of Solid-State Circuits* 54, no. 1 (2018): 75-87, 2019

[DATE19] I. Yoon, A. Anwar, T. Rakshit, A. Raychowdhury. "Transfer and online reinforcement learning in STT-MRAM based embedded systems for autonomous drones." In *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1489-1494. IEEE, 2019.

[ISSCC18] A. Amravati, S. Nasir, S. Thangadurai, I. Yoon, A. Raychowdhury, “A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots,” in IEEE International Solid-State Circuits Conference, 2018

Acknowledgements

Georgia Tech ICSRL members:
- Jong-Hyeok Yoon
- Ningyuan Cao
- Anvesha Amaravati
- Insik Yoon
- Aqeel Anwar
- Saad Bin Nasir

Collaborators:
- Dr. Keith Bowman (Qualcomm)
- Dr. Titash Rakshit (Samsung)
- Dr. Charles Augustine (Intel)
- Dr. Vivek De (Intel)
- Dr. Muhammad Khellah (Intel)
- Dr. Carlos Tokunaga (Intel)

Sponsors:
- Semiconductor Research Corporation
- DARPA
- Center for Brain Inspired Computing
- Qualcomm
- Intel

Thank You

Contact:
Zishen Wan (zishenwan@gatech.edu)
Dr. Arijit Raychowdhury (arijit.raychowdhury@ece.gatech.edu)