Re-Thinking Computing with Neuro-Inspired Learning: Algorithms to Hardware

Kaushik Roy

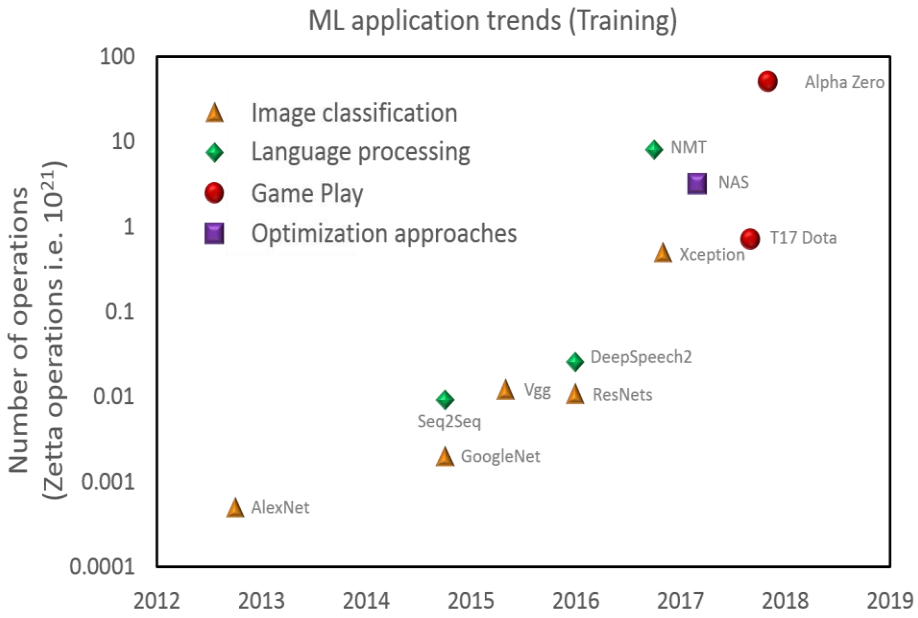
kaushik@purdue.edu

Elmore School of Electrical and Computer Engineering

Purdue University

PURDUE UNIVERSITY

Motivation

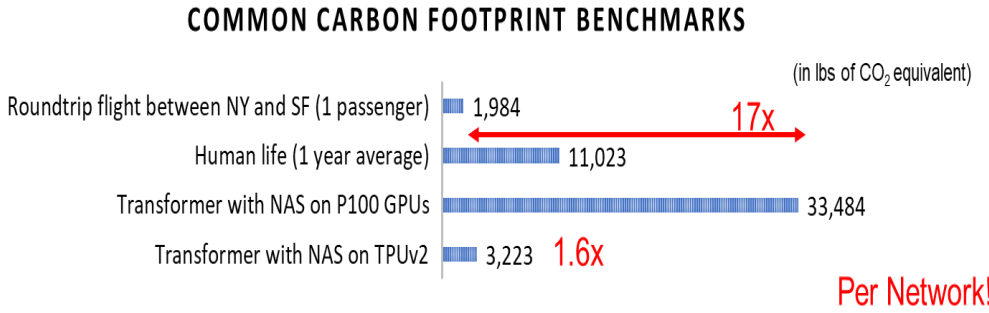


Google Edge TPU

Retinanet DNN* on a smart glass

Performance	
Frames/sec	13.3
Battery Life	
Energy/op	0.5 pJ/op
Energy/frame	0.15 J/frame
Time-to-die (2.1WH)	**64 mins**

*300 GOPs/inference



Where do the in-efficiencies come from?

Algorithms Sensors/Hardware Architecture Circuits and Devices

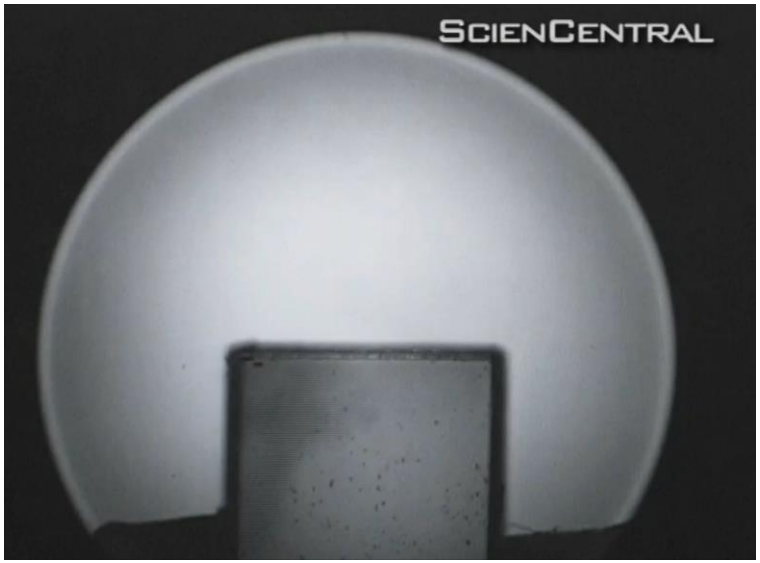
Comparison with Biological Systems

- Biological systems still possess a level of functionality that is unmatched in artificial systems
- Consider a reactive behavior of a Fruit fly (~100K neurons)
  - Fly fast while avoiding obstacles in cluttered environments
  - Dodge dynamic obstacles and active attacks

Flying monkey UAV, UPenn
~1-2W compute

VS

Fruit fly
~uW compute



Dickinson's lab Caltech

The Big Picture

*Enable autonomous intelligent systems by **improving the compute efficiency and robustness of cognitive tasks** through cross-layer innovations from algorithms to hardware*

Exemplary application driver: Autonomous drones

Goal:
Vision-based navigation (localization, mapping, odometry, obstacle avoidance, path planning)

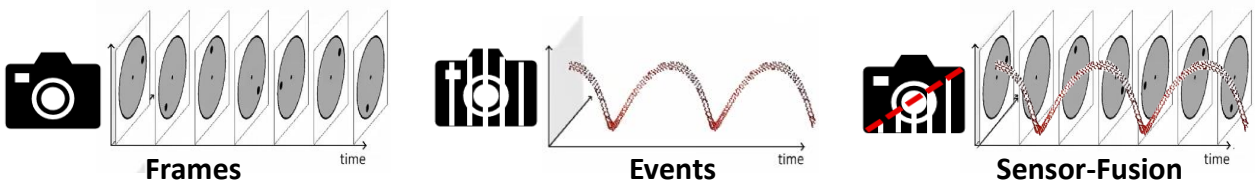
Compute (Hardware) efficiency (Need > 100X improvement in energy-delay over current state-of-the-art*)

Diverse learning modalities (Supervised, reinforcement, explainable, spike-based hybrid learning)

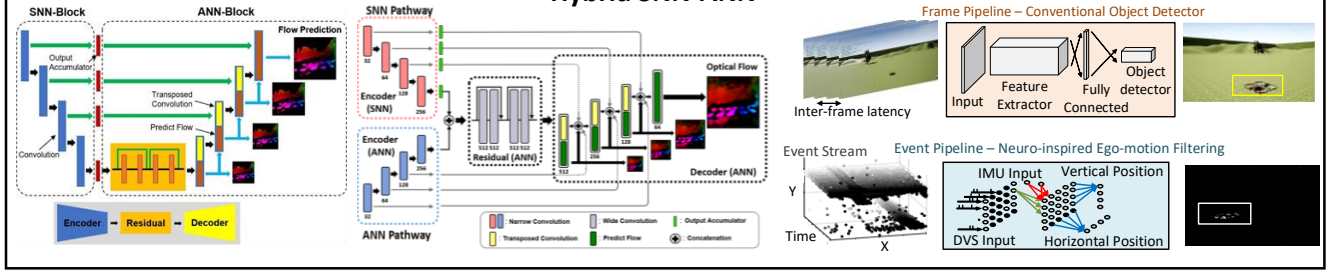
Robustness

* Based on cumulative TOPS required for visual SLAM, depth map generation and object detection @ 30fps in under 0.5W

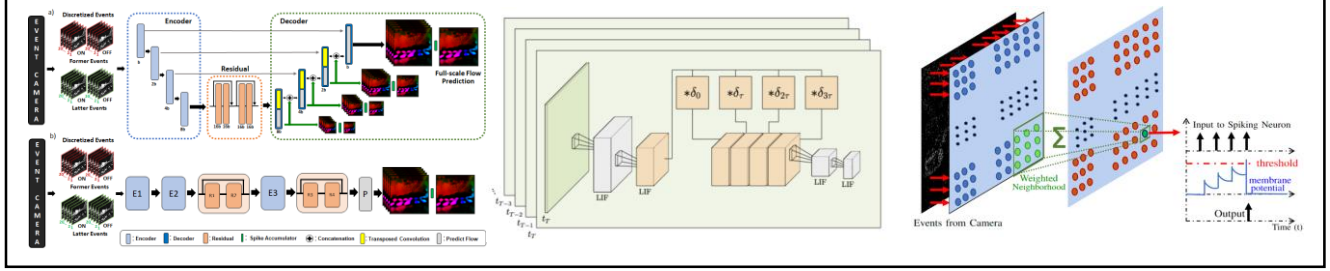
Cross-Layer Design: Sensors, Algorithms, Hardware



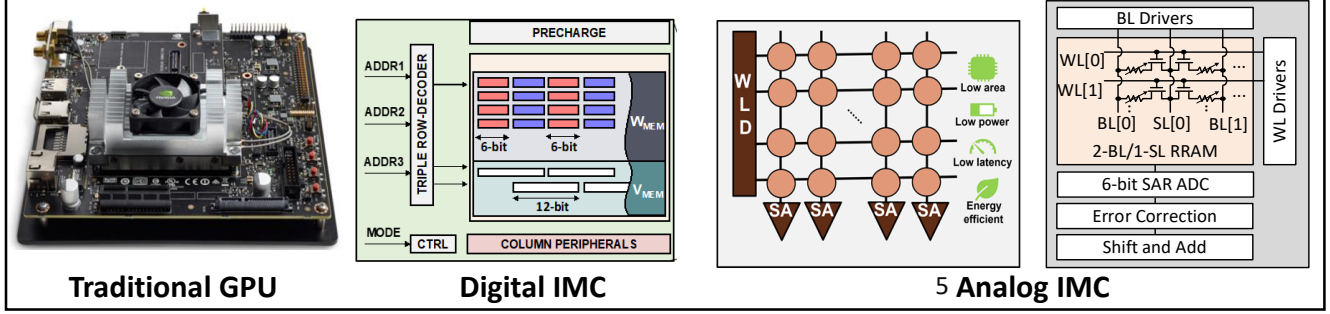
Hybrid SNN-ANN



Lightweight SNNs



Hardware Accelerators



Traditional GPU | Digital IMC | Analog IMC

- Optical Flow / Depth
- Segmentation
- Object Detection
- Tracking
- Localization

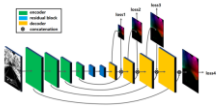
Modalities…..

NETWORK ARCHITECTURES

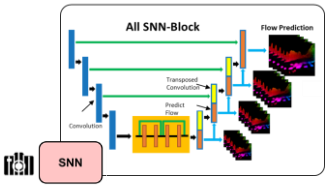
HARDWARE COMPUTATIONAL EFFICIENCY

TRAINING COMPLEXITY

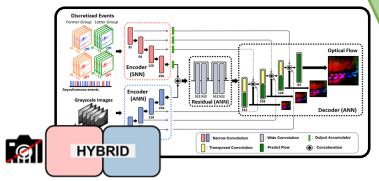
ANN Architectures



SNN Architectures



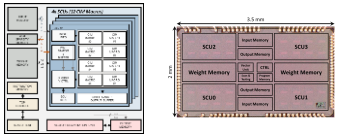
Hybrid Architectures



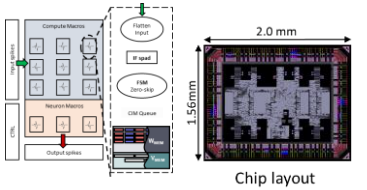
Current GPUs

NVIDIA Jetson TX-2

Adaptive-SNR Sparsity aware CiM

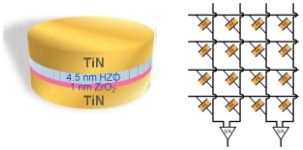


SNN Accelerator

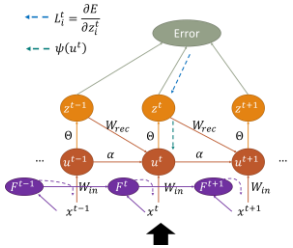


Chip layout

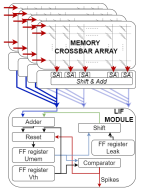
Emerging Devices



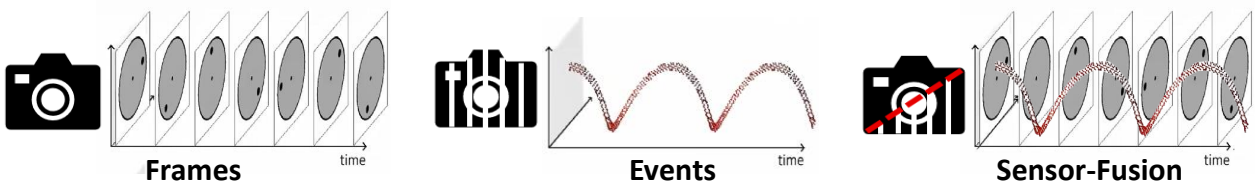
On-Chip Learning



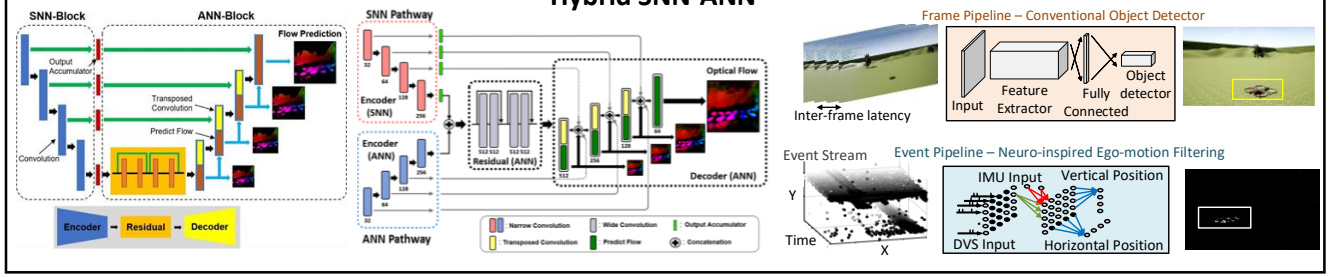
HW/SW codesign with ADC-Less IMC



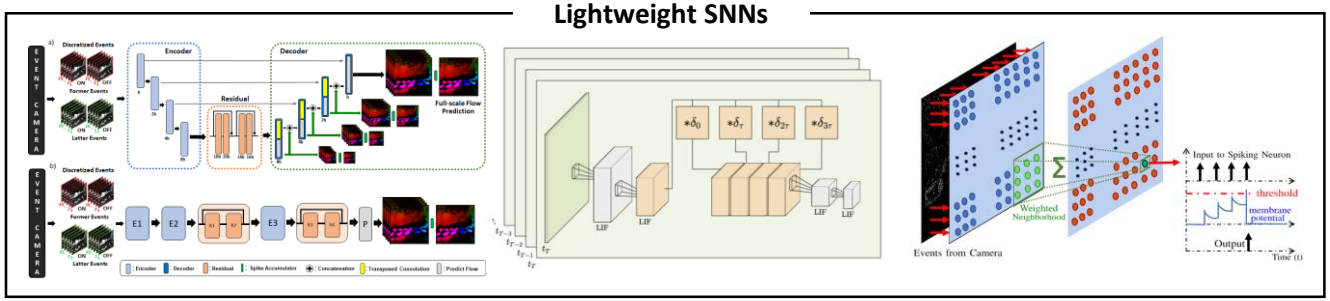
Cross-Layer Design: Sensors, Algorithms, Hardware



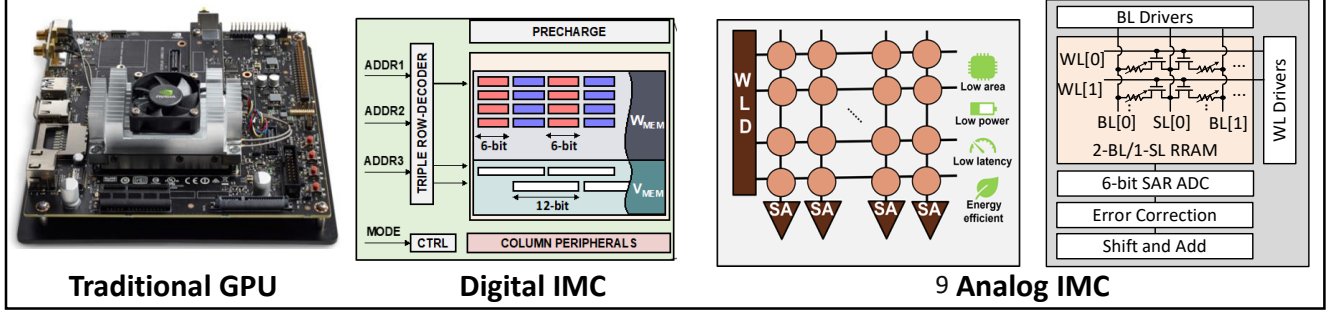
Hybrid SNN-ANN



Lightweight SNNs



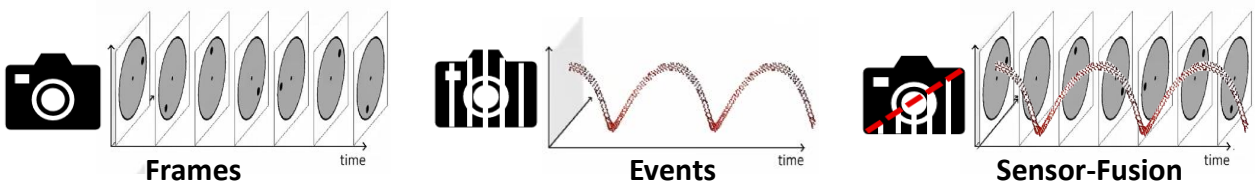
Hardware Accelerators

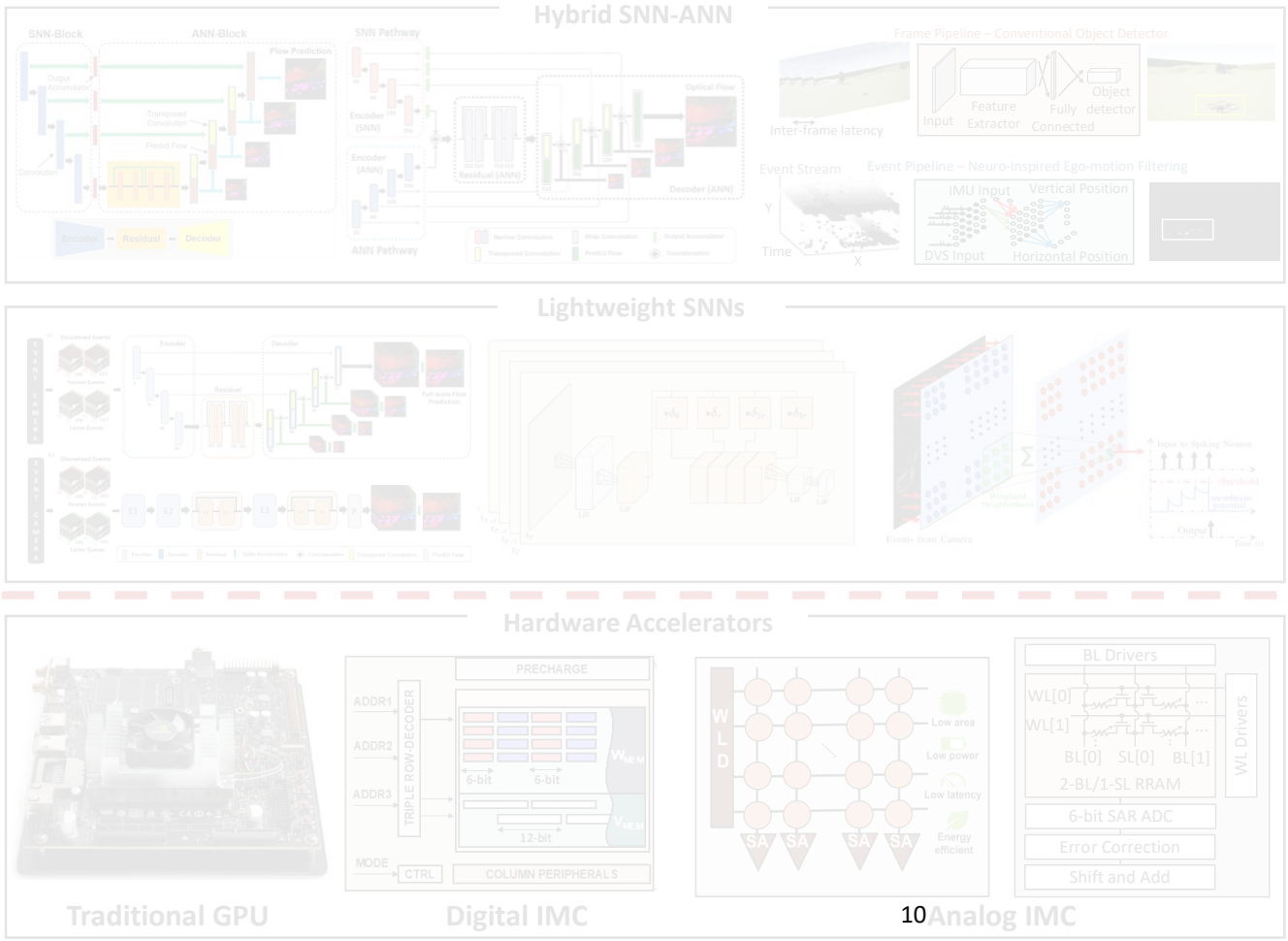


Traditional GPU | Digital IMC | Analog IMC

- Optical Flow / Depth
- Segmentation
- Object Detection
- Tracking
- Localization

Cross-Layer Design: Sensors, Algorithms, Hardware

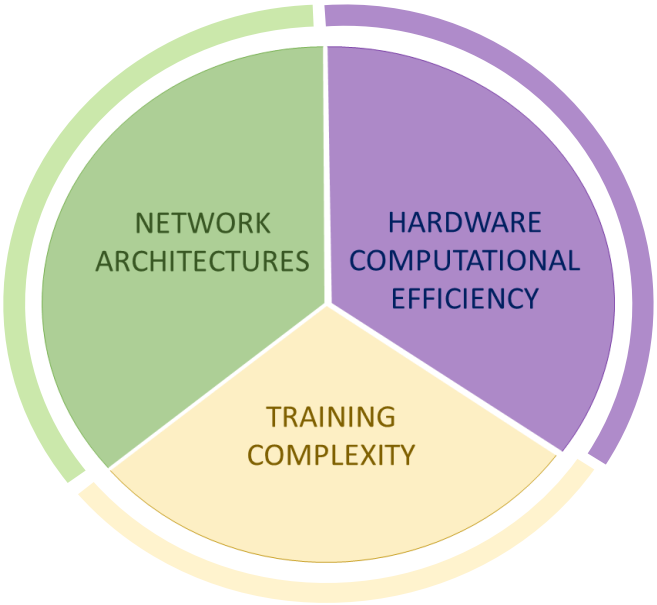




- Optical Flow / Depth
- Segmentation
- Object Detection
- Tracking
- Localization

10

Tradeoffs: Starting with Sensors….



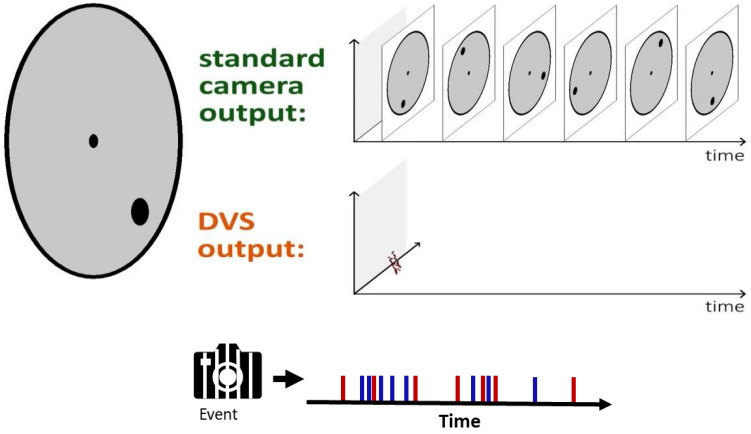
Memory

Accuracy

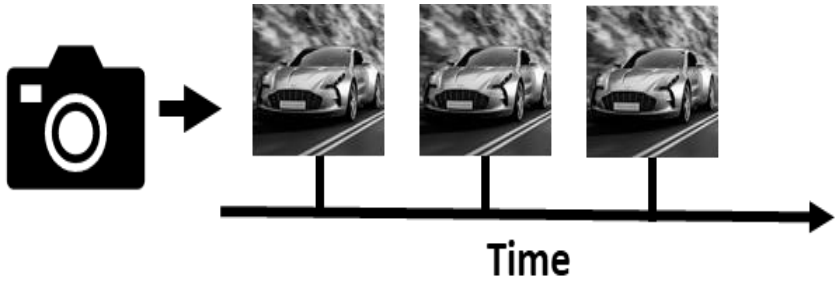
The choice of sensors dictates the choice of any of the modalities

Latency

Energy



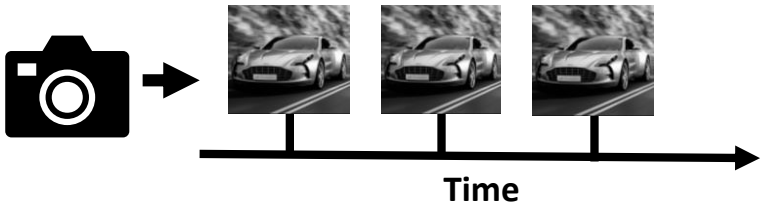
Event-based Cameras

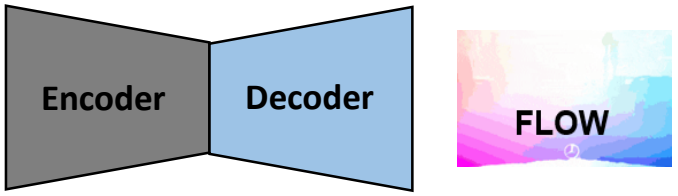


Frame-based Cameras

Frame vs Event-based Cameras

Frame-based Cameras



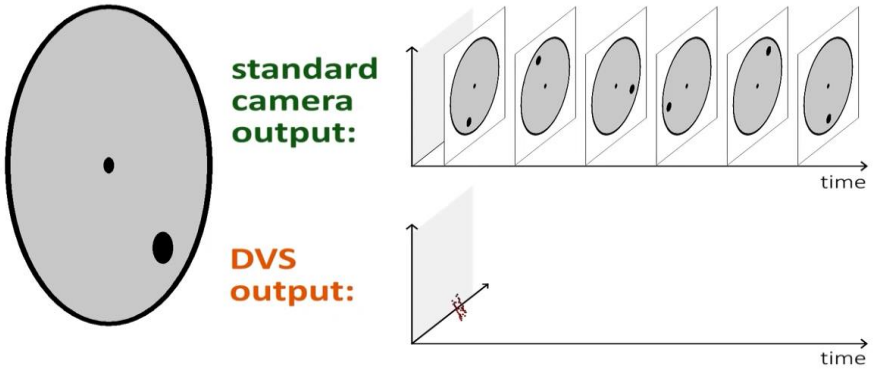


Analog Neural Network (ANN)

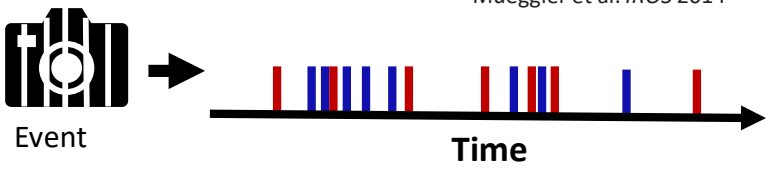
Motion Blur

HDR

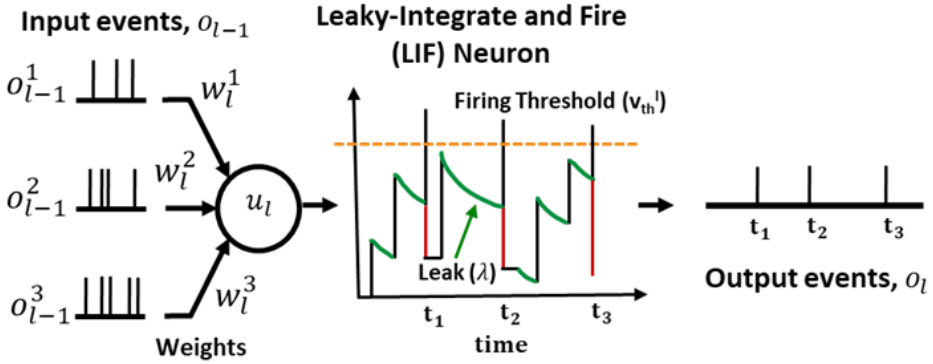
Event-based Cameras



Mueggler et al. *IROS* 2014



Spiking Neural Network (SNN)



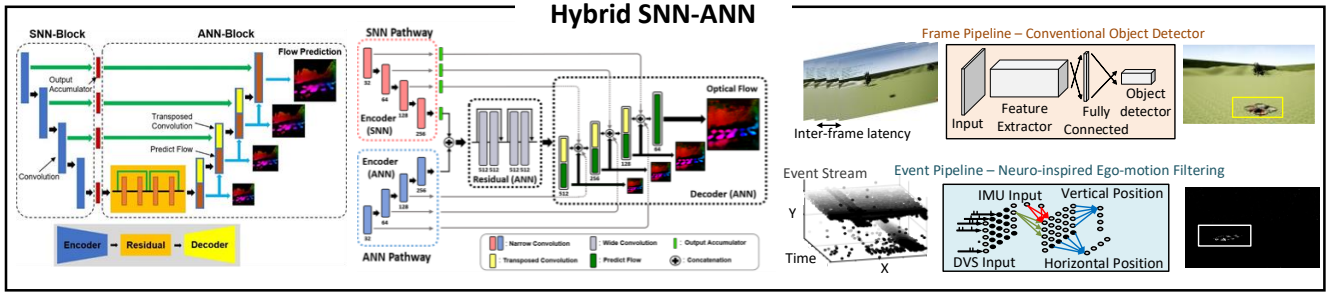
Asynchronous events naturally fit with SNNs

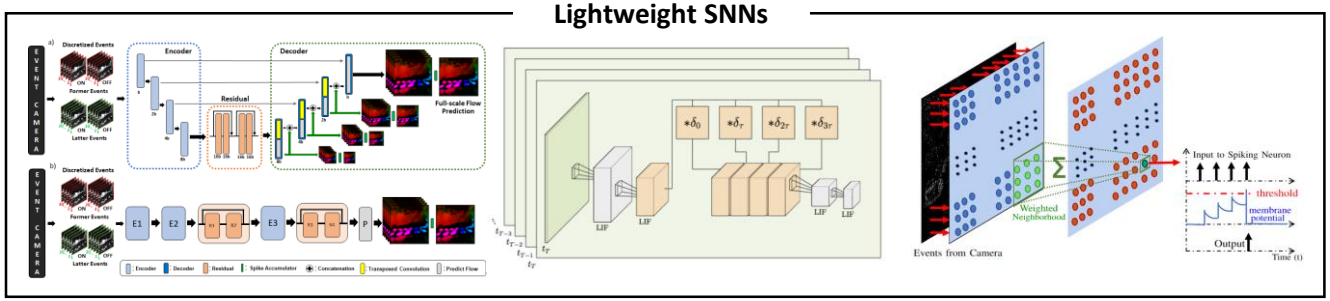
High temporal resolution

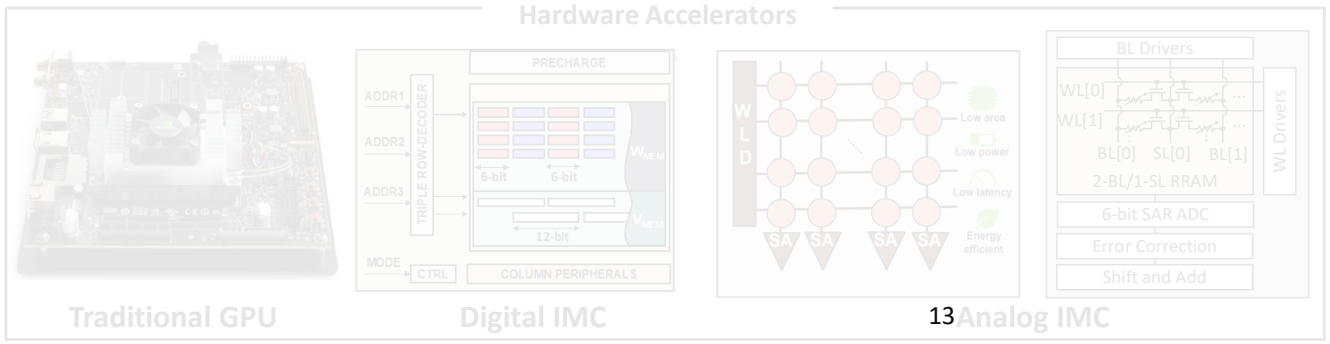
High dynamic range

Low-Power

Cross-Layer Design: Sensors, Algorithms, Hardware







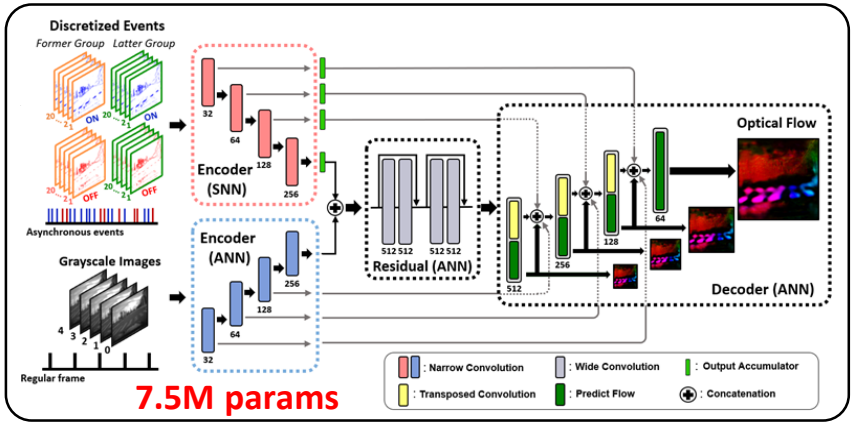
- Optical Flow / Depth
- Segmentation
- Object Detection
- Tracking
- Localization

Our Approaches

SNN

Fusion-FlowNet: Sensor-fusion

HYBRID



7.5M params

Event inputs

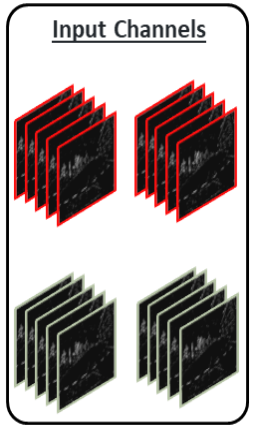
- Data at high **temporal** but low spatial resolution

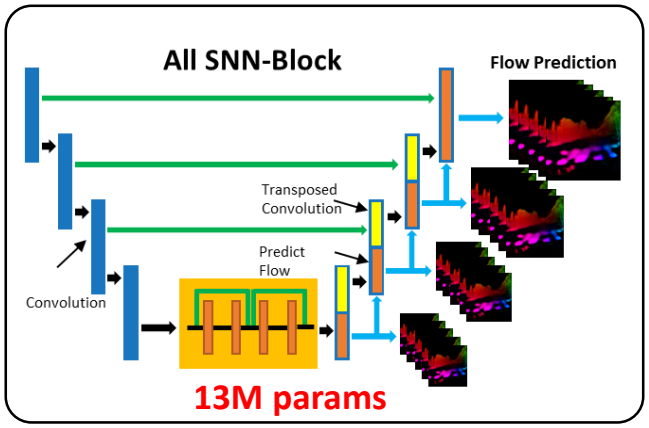
Frame inputs

- Data at high **spatial but low temporal** resolution

Combined inputs for a better flow estimation

Adaptive-FlowNet: Fully Spiking Architecture





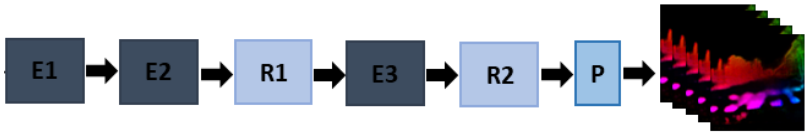
13M params

Model with all spiking layers

- Directly **compatible** with **event inputs**
- Capture **temporal information**
- Combat **vanishing gradient** with adaptive spiking neuronal model

Fire-FlowNet: Lightweight Architecture for the Edge

SNN

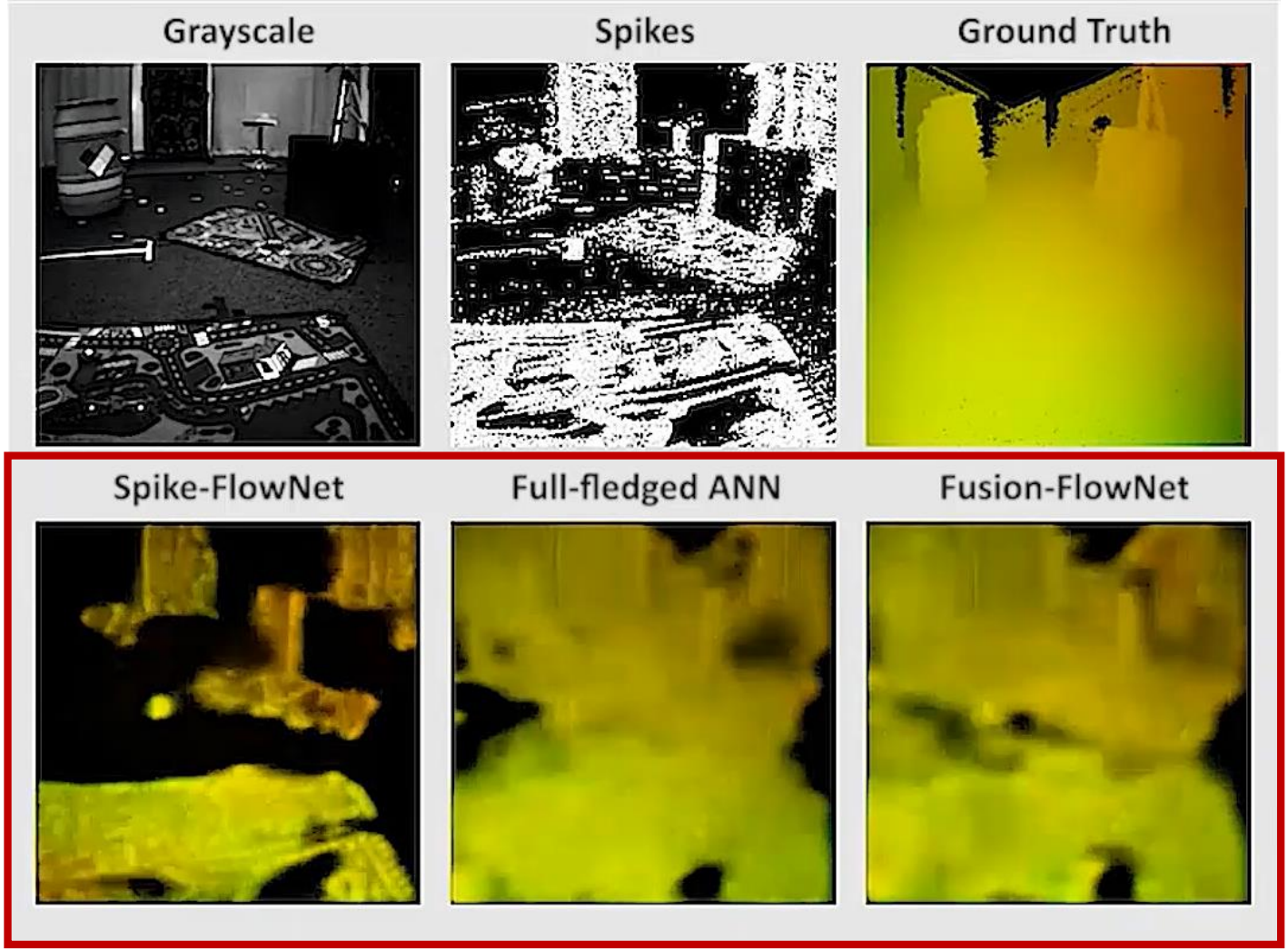


60K params

- Highly **energy efficient** and **low latency** implementation
- Suitable for **fast inference** on the resource constrained edge

C. Lee, et. al., Spike-FlowNet:..., ECCV 2020
C. Lee, A. Kosta, and K. Roy., Fusion-FlowNet:..., ICRA 2022

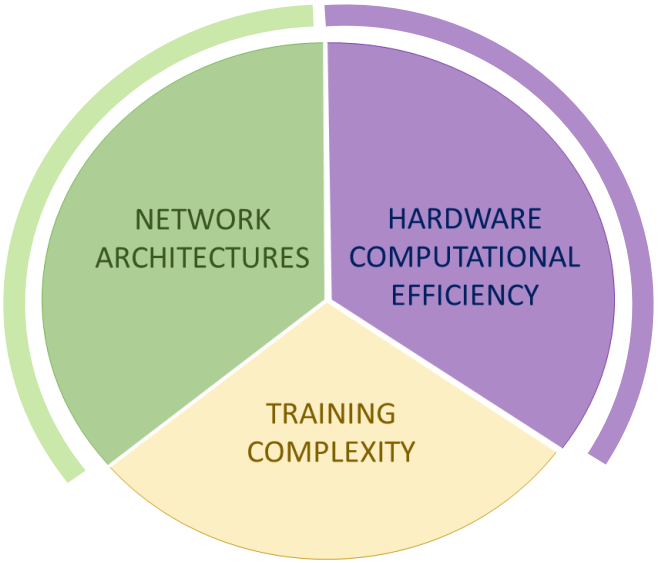
Visualization – Flow Comparison on MVSEC



Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar and Kostas Daniilidis . "The multivehicle stereo event camera dataset: An event camera dataset for 3D perception." IEEE Robotics and Automation Letters 3.3 (2018): 2032-2039.

Chankyu Lee, Adarsh Kosta and Kaushik Roy. "Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks." *European Conference on Computer Vision*. Springer, Cham, 2020.

Tradeoffs

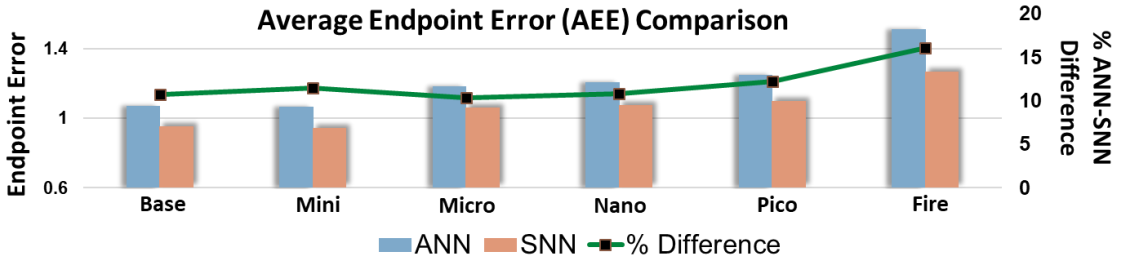


Accuracy

Different network architectures differ in Accuracy and Energy Consumption

Latency

Energy



- Adaptive SNNs consistently outperform similarly sized ANNs
- As the model size reduces the performance difference between SNN and ANN models increases

Computational Efficiency

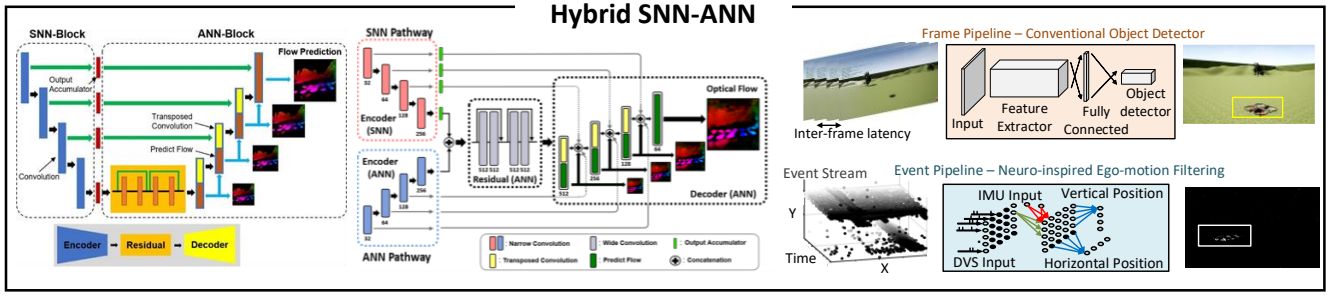
Architecture	#Params (M)	#Params Improvement	#ANN OPS ($x10^9$)	Spike Activity (%)	#SNN OPS ($x10^6$)	Energy (mJ)	Energy Improvement
Base-ANN	13.046	1x				24.54	1x
Base-SNN	13.04	1x				4.66	5x
Mini-SNN	3.41	3.8x				1.75	14x
Micro-SNN	0.93	14x				1.05	23.4
Nano-SNN	0.27	48x				**0.48**	**51x**
FireFlow-SNN	**0.057**	**142x**				0.95	26x

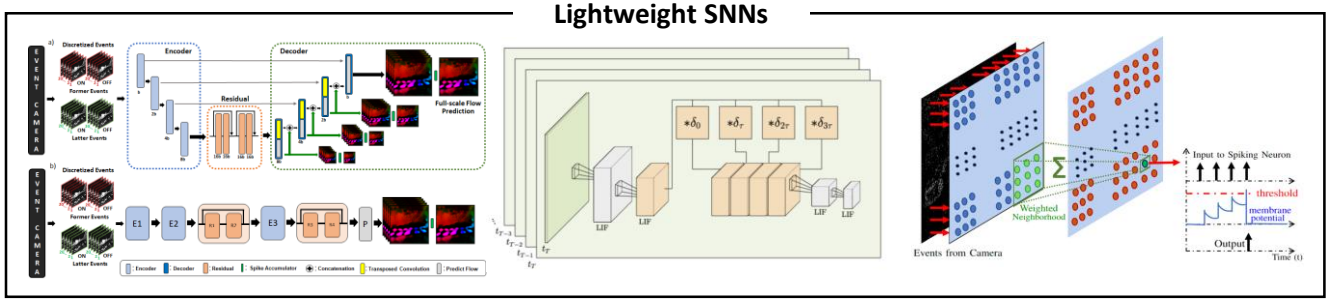
Nano-SNN

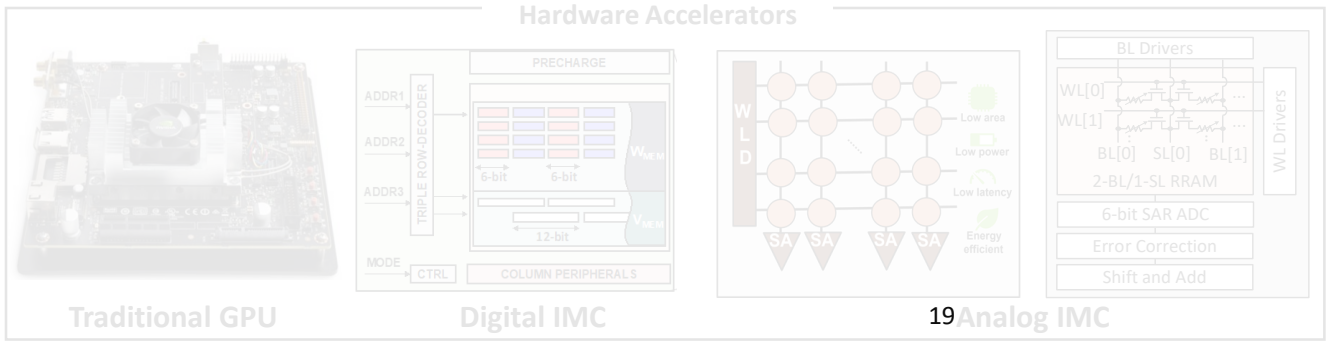
48x Lower Parameters
51x Lower Compute Energy
Similar error as Base-ANN

Kosta, Roy, ICRA 2023

Cross-Layer Design: Sensors, Algorithms, Hardware



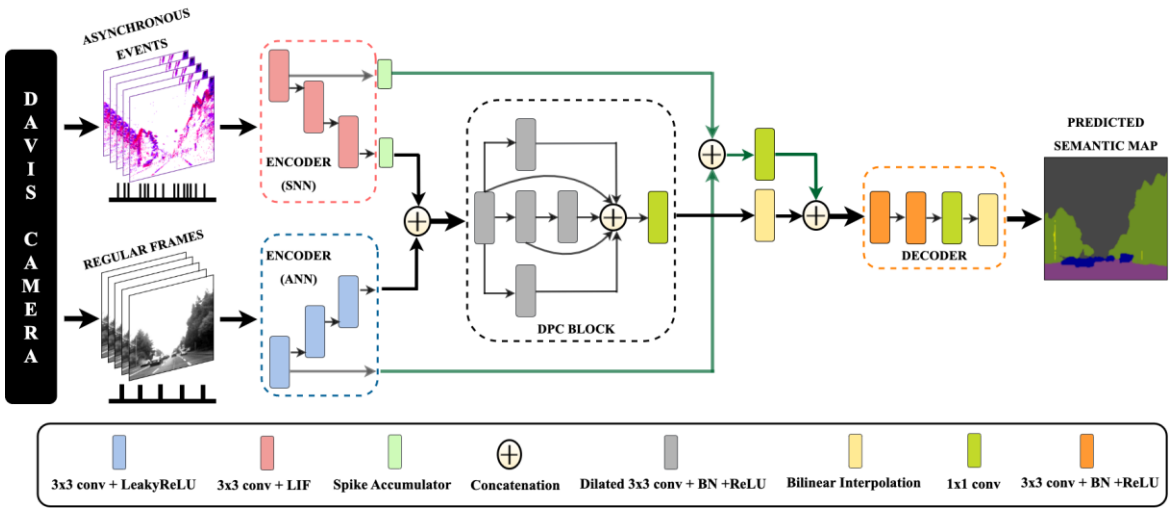


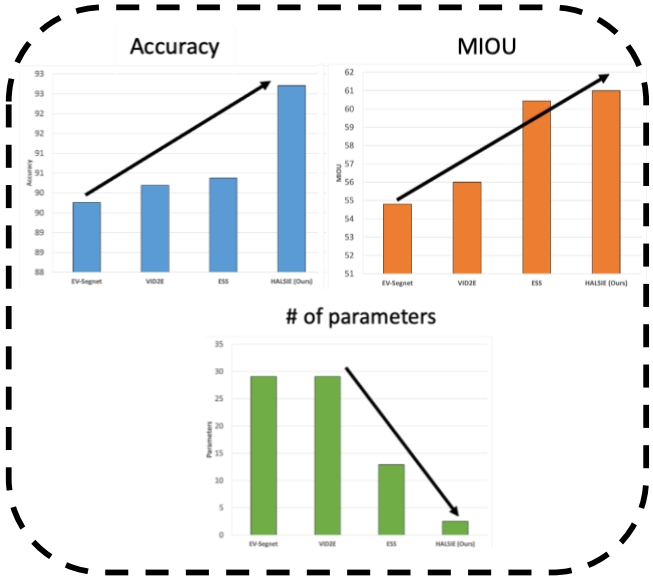


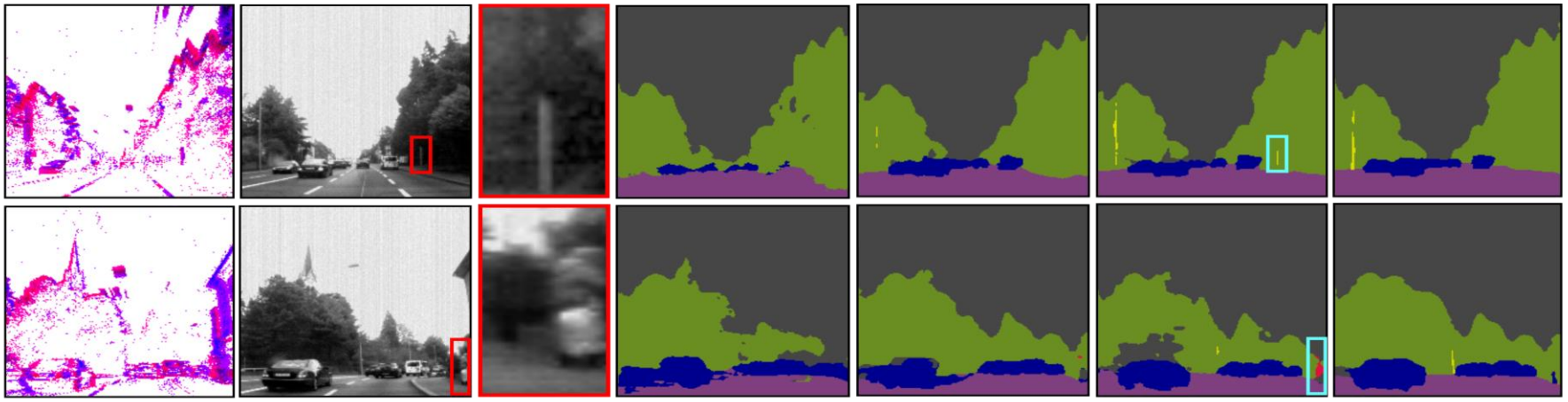
- Optical Flow / Depth
- Segmentation
- Object Detection
- Tracking
- Localization

19

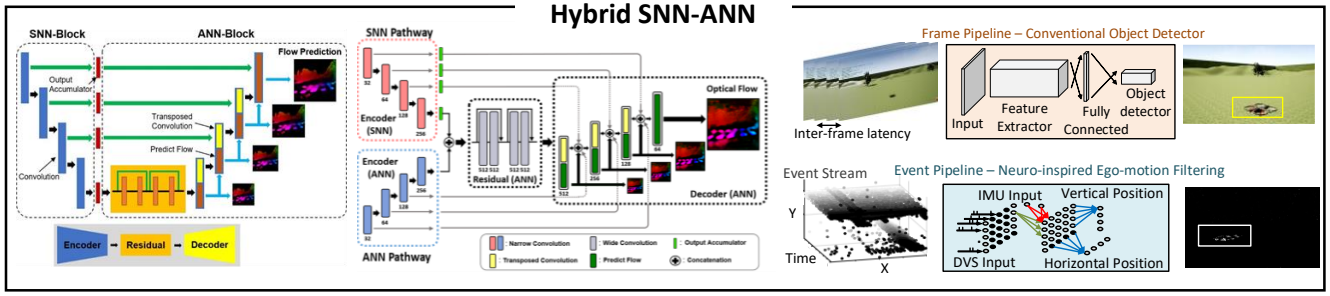
HALSIE: Hybrid Approach to Learning Segmentation by Simultaneously Exploiting Image and Event Modalities

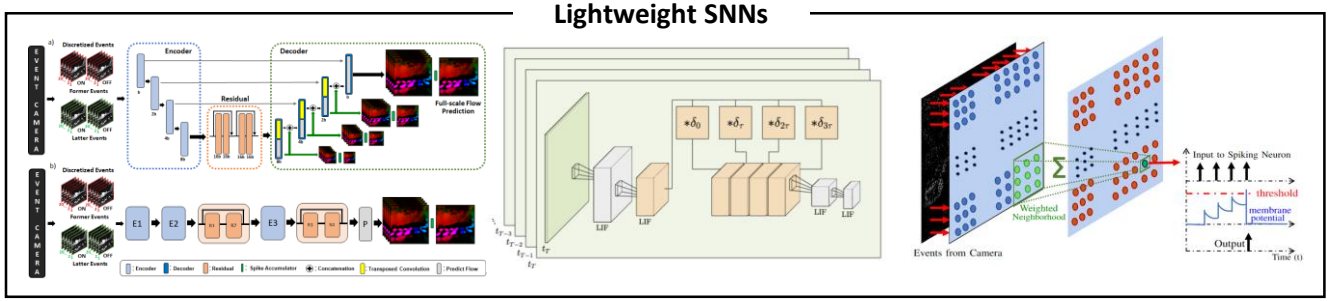


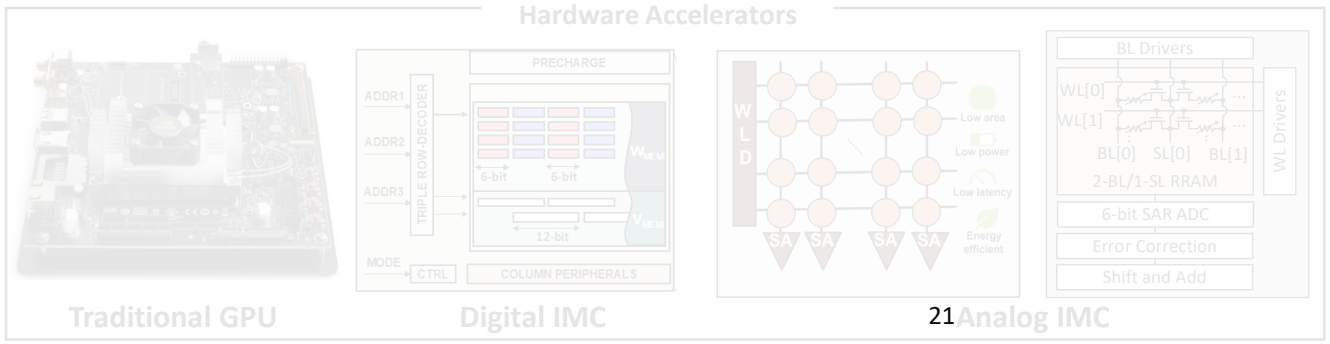




Cross-Layer Design: Sensors, Algorithms, Hardware





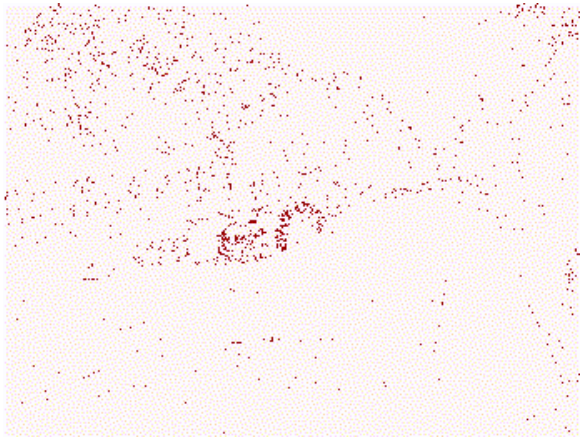


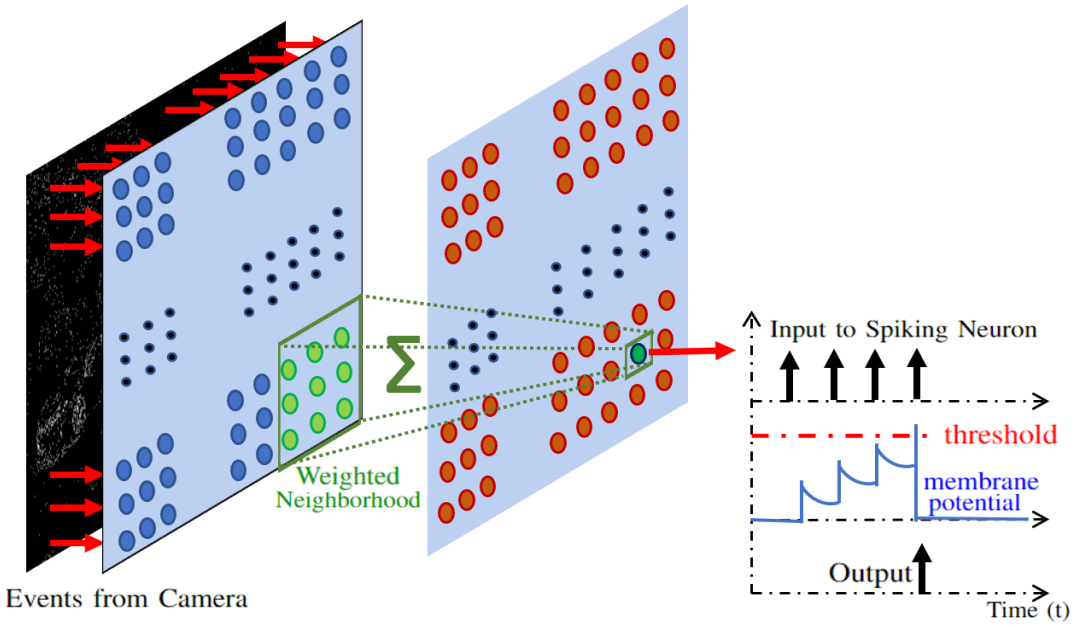
- Optical Flow / Depth
- Segmentation
- Object Detection
- Tracking
- Localization

DOTIE: Detecting Objects through Temporal Isolation of Events

YOLOv3 on Frames

YOLOv3 on Events





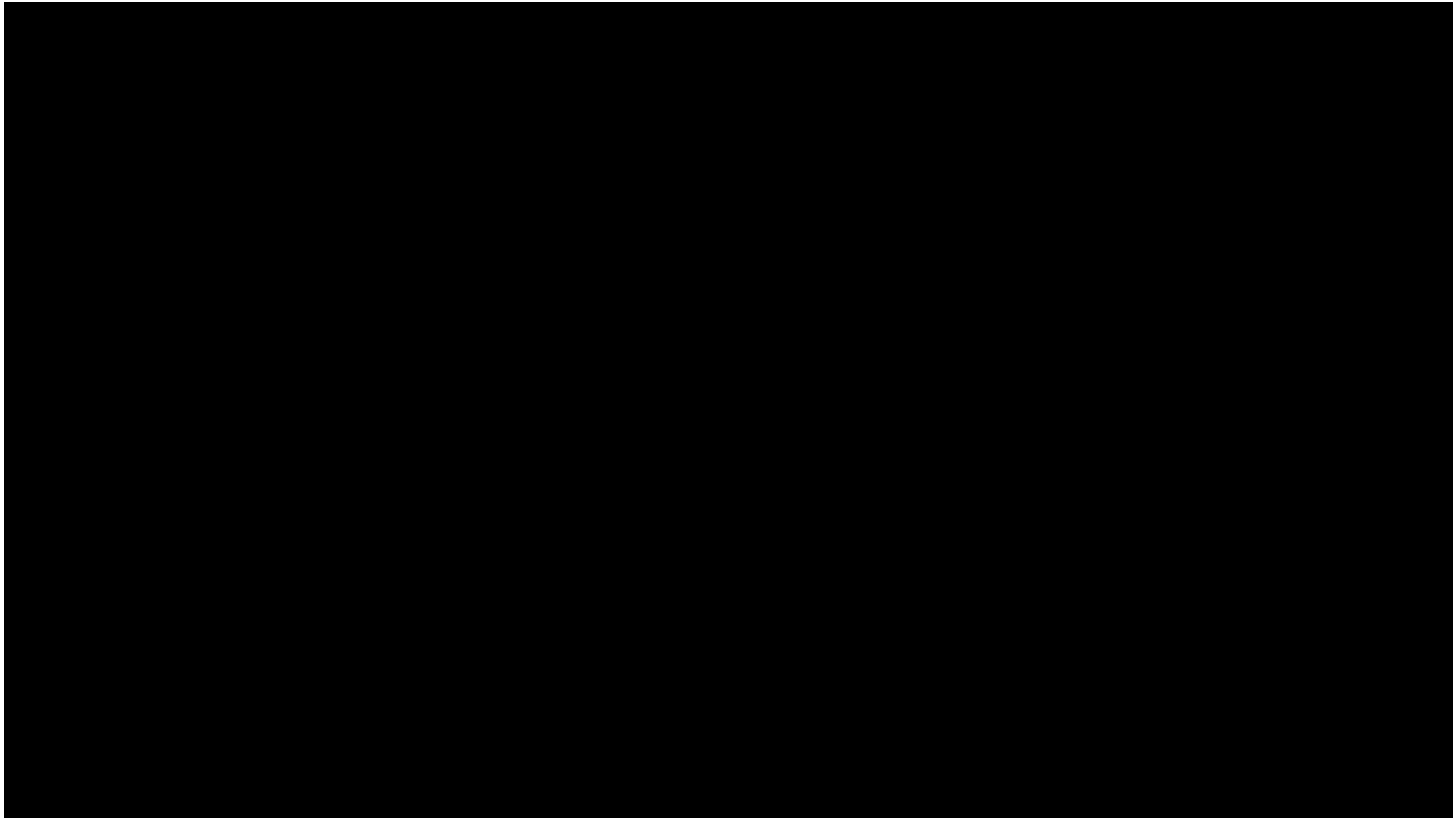
Our single layer network can isolate events corresponding to moving objects and detect objects accurately, with low latency and energy consumption

Events do not contain photometric characteristics such as light intensity and texture

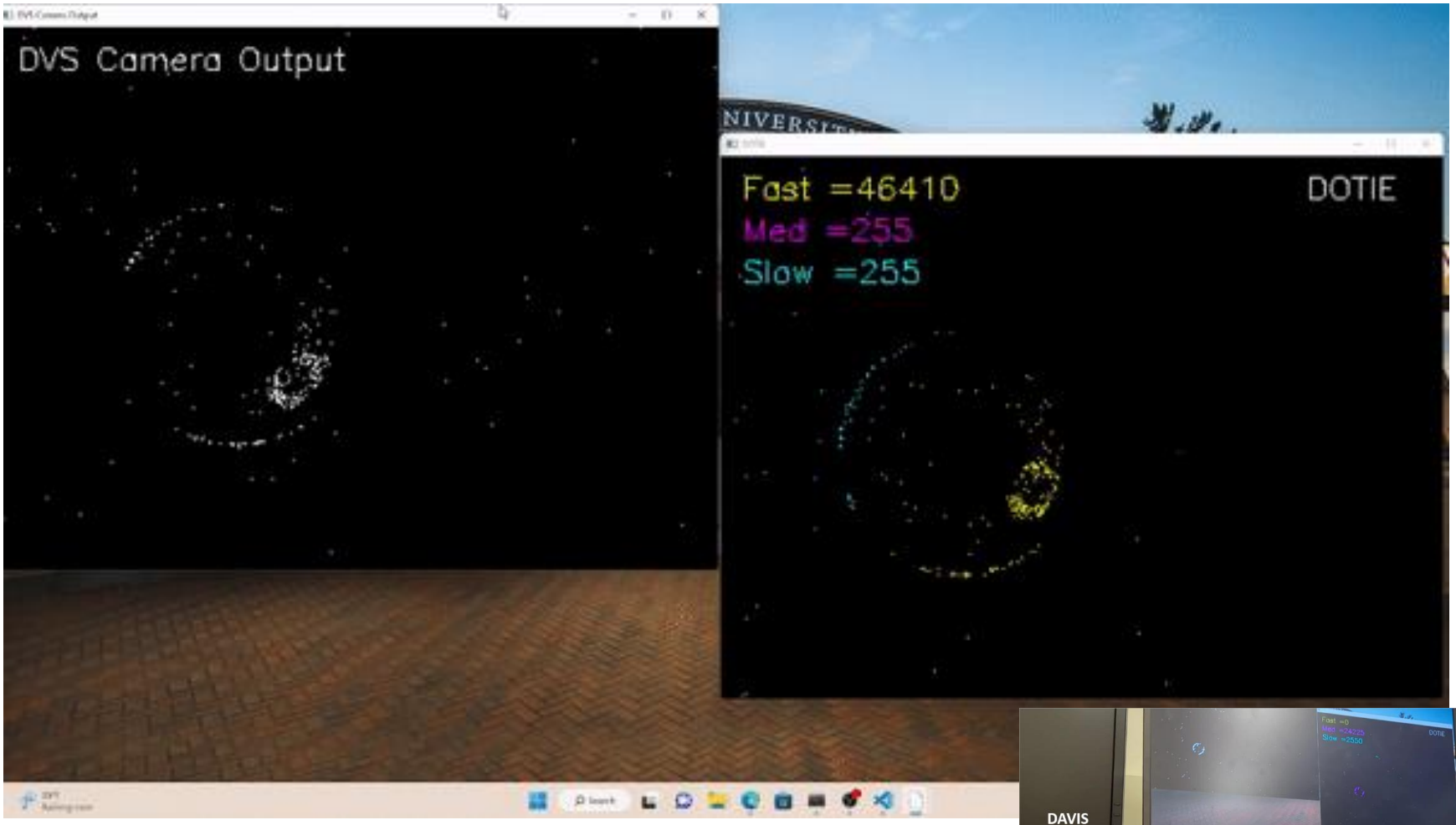
Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).

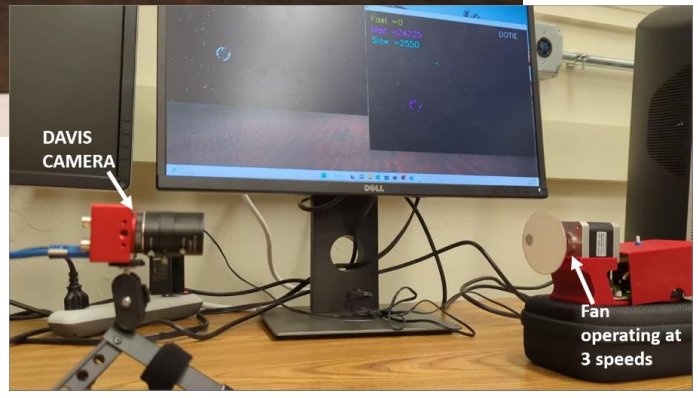
Nagaraj, Liyanagedera, Roy, "DOTIE: …", ICRA 2023

Demonstration of detecting speeds of objects using the DOTIE Algorithm



Demonstration of detecting speeds of objects using the DOTIE Algorithm



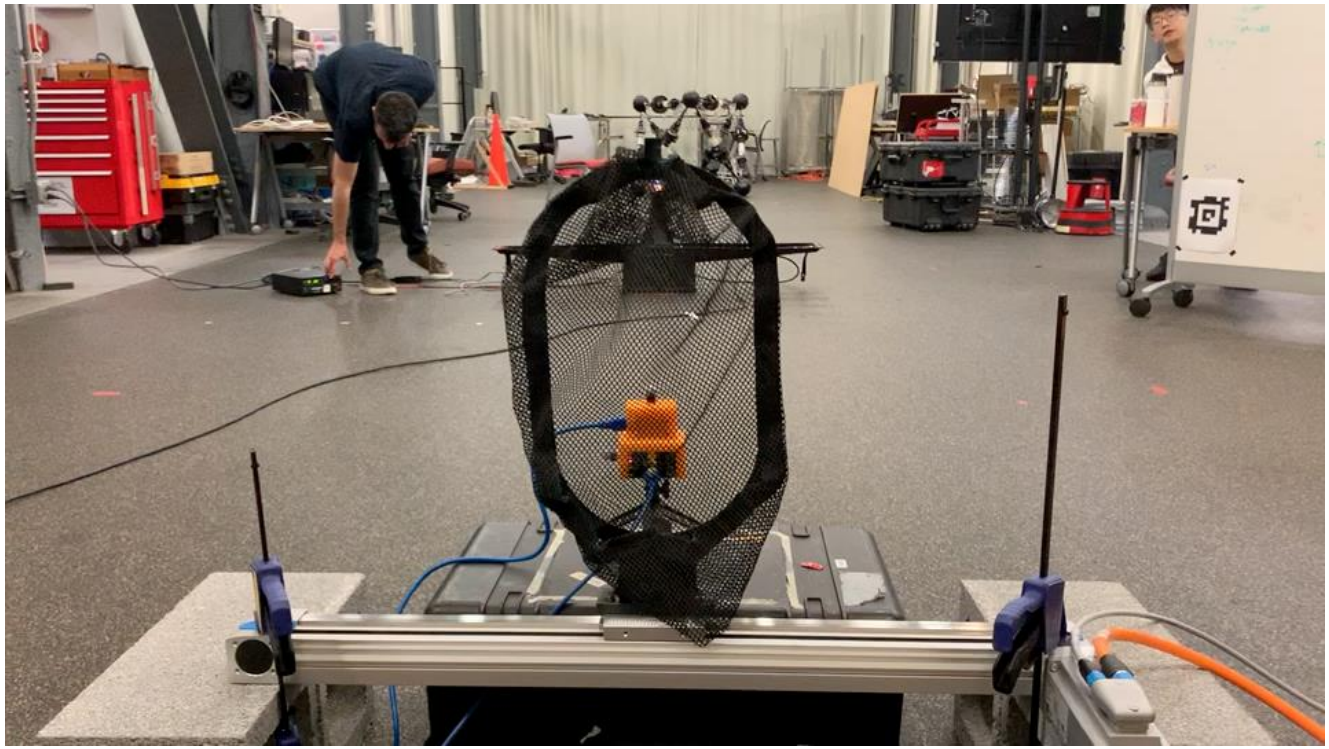


Nagaraj, Liyanagedera, Roy,
"DOTIE: …", ICRA 2023

EV-Catcher: Taking Inspiration from Nature

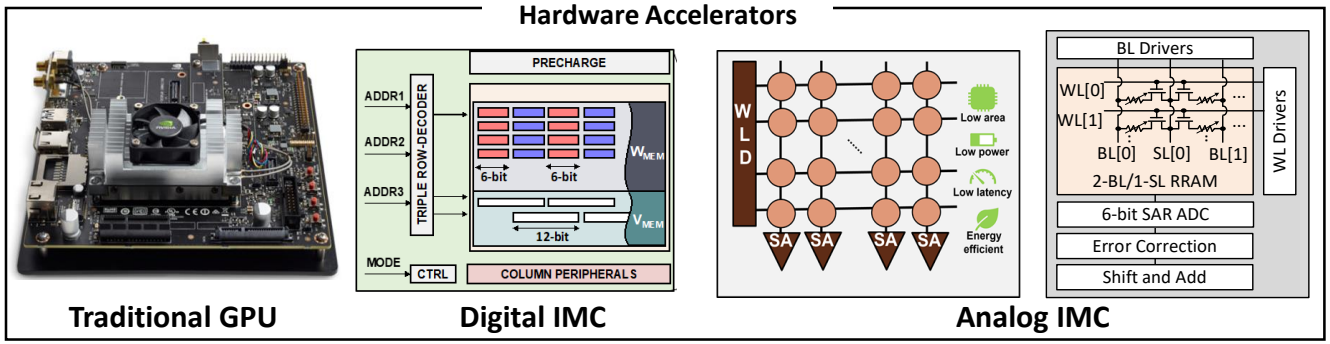
Aim: To minimize ***latency*** in ***reactive*** behaviors such as:

- Interception
- Time to collision



Kostas Daniilidis, UPenn

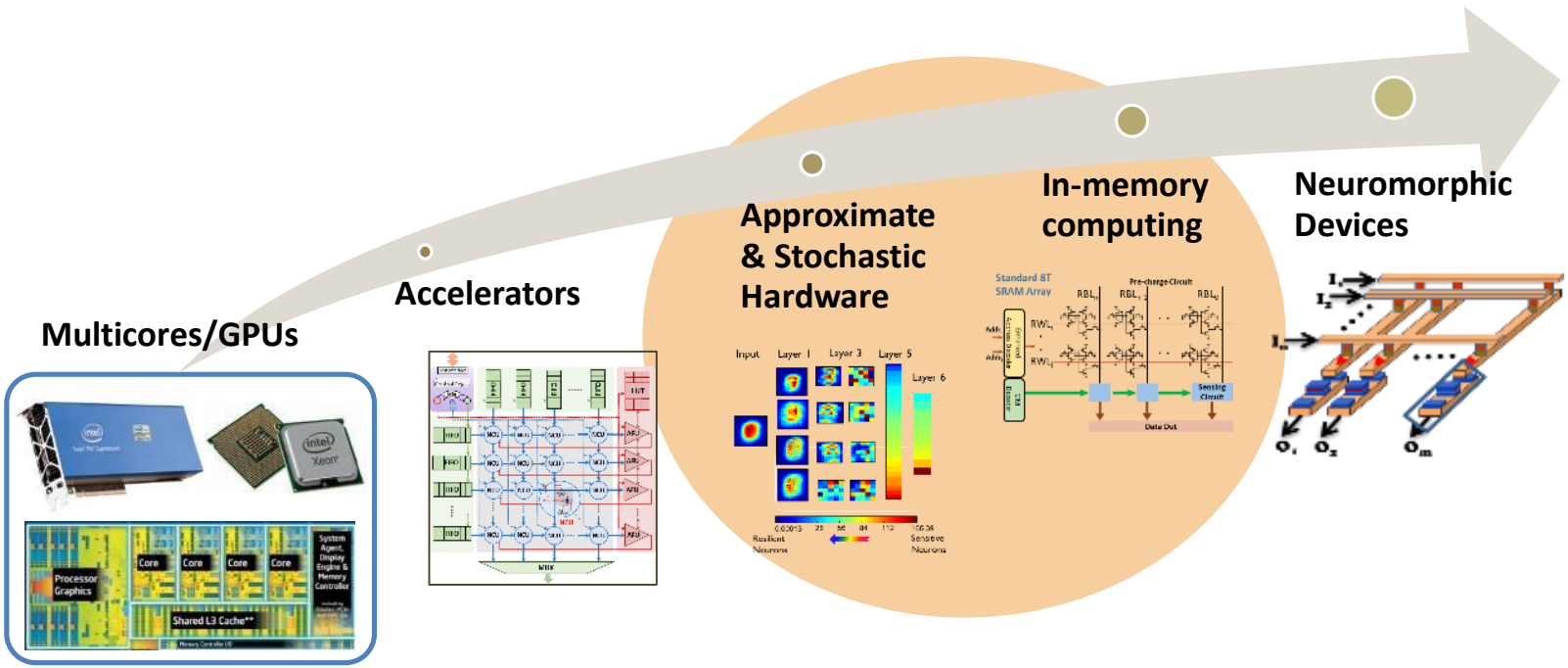
Cross-Layer Design: Sensors, Algorithms, Hardware



- Optical Flow / Depth
- Segmentation
- Object Detection
- Tracking
- Localization

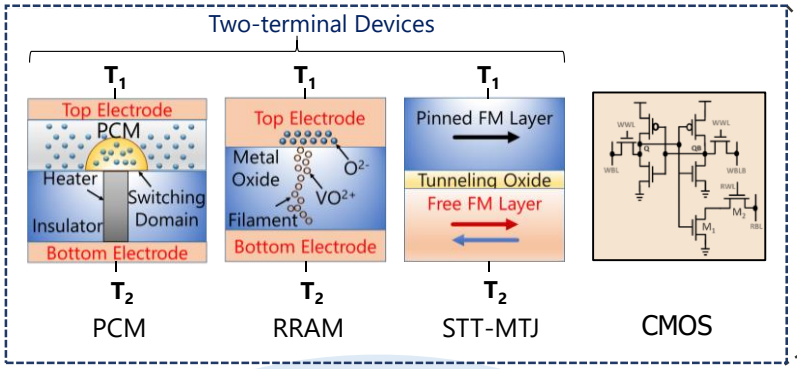
Hardware Architecture

- Circuits and architectures that can efficiently implement the algorithms (SNNs and ANNs): need for hybrid systems
  - Near-/In-Memory Computing for MVMs
  - Approximate and stochastic hardware
  - Neuromorphic devices and interconnects

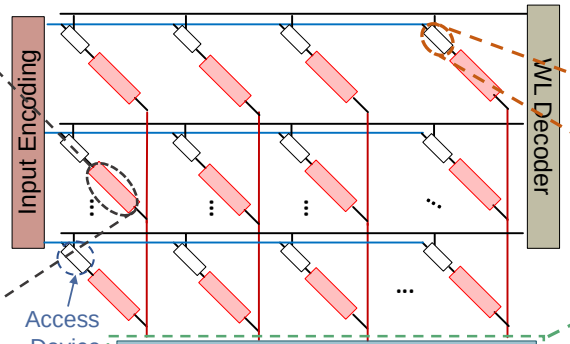


Hardware Architecture: CiM

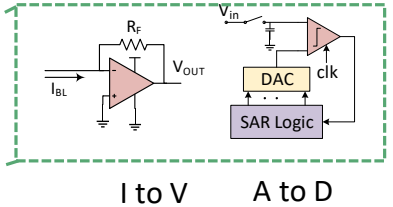
In-Memory Computing Memory Devices

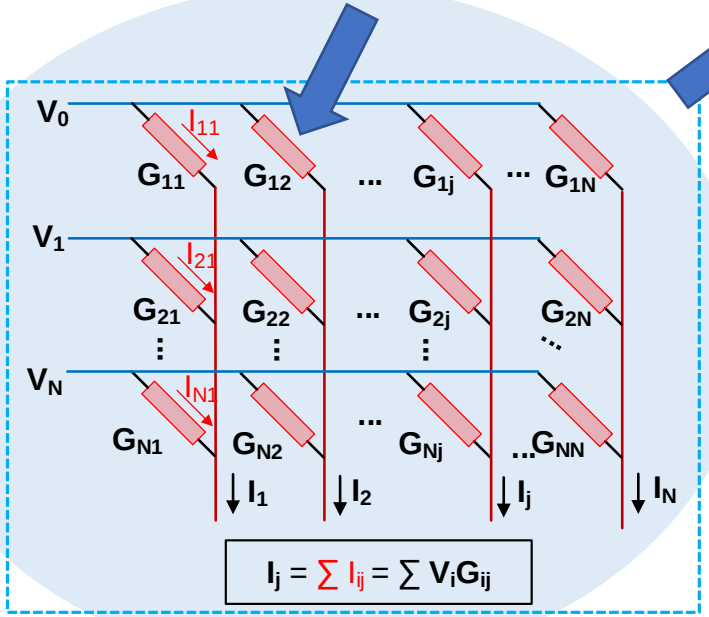


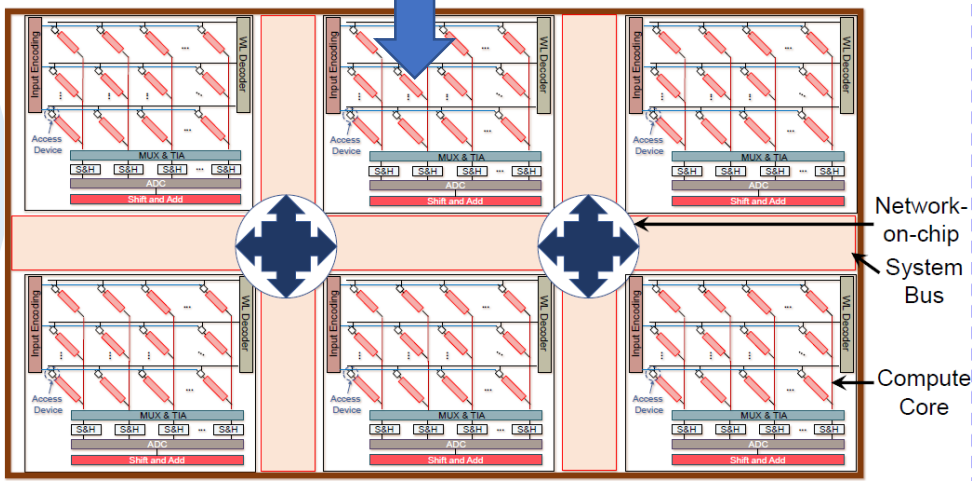
Chakraborty et. al. Resistive Crossbars as Approximate Hardware Building Blocks for Machine Learning: Opportunities and Challenges, Proc. of IEEE, 2020



The Peripherals





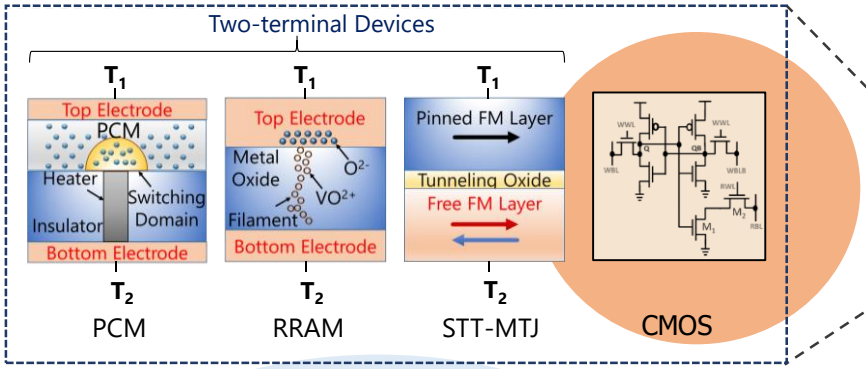


Efficient MVM

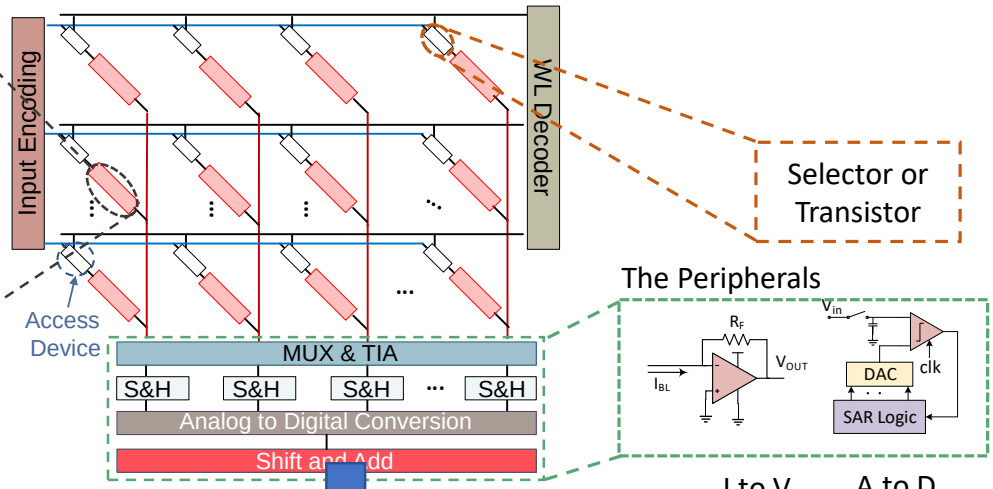
Spatially Distributed Cores

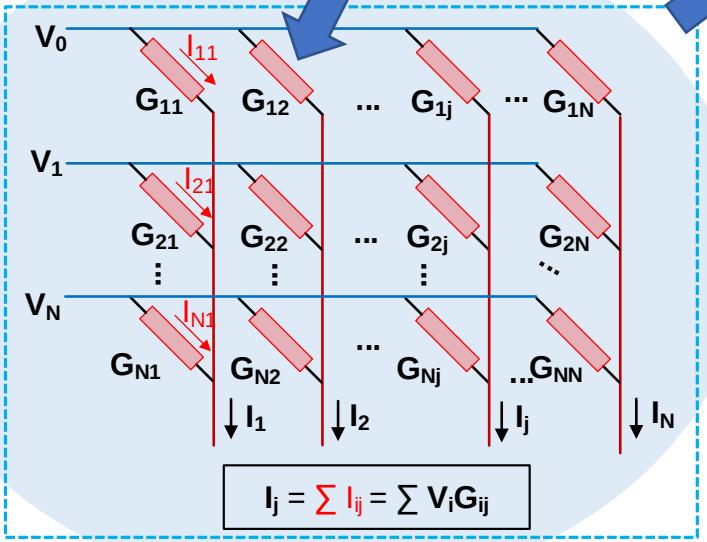
Hardware Architecture: CiM

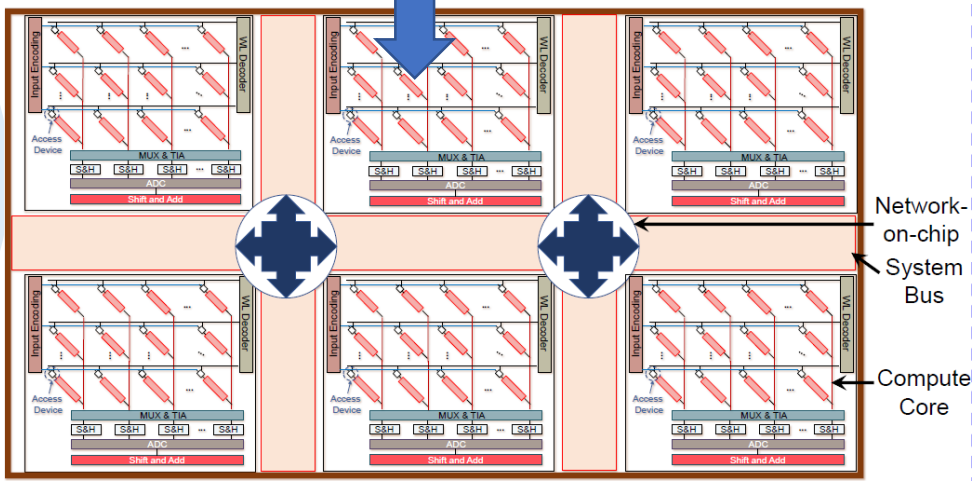
In-Memory Computing Memory Devices



Chakraborty et. al. Resistive Crossbars as Approximate Hardware Building Blocks for Machine Learning: Opportunities and Challenges, Proc. of IEEE, 2020







Efficient MVM

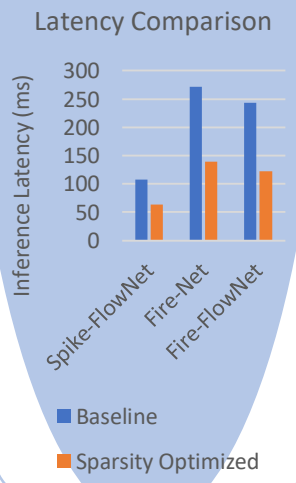
Spatially Distributed Cores

Hardware Implementations

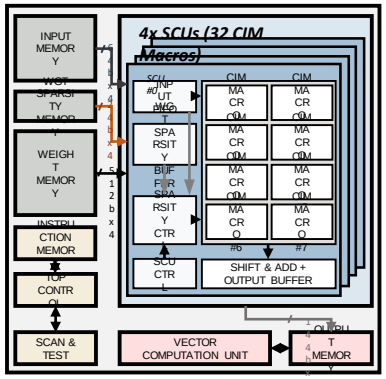
GPU-based baseline

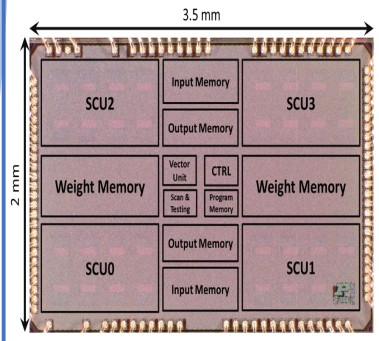
NVIDIA Jetson TX-2

Baseline: High latency due to inefficient handling of SNNs on GPUs.



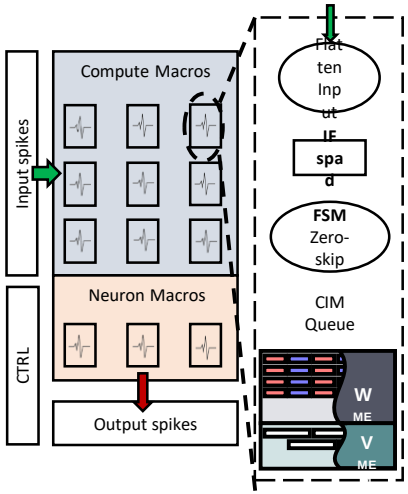
Adaptive-SNR Sparsity-Aware CiM

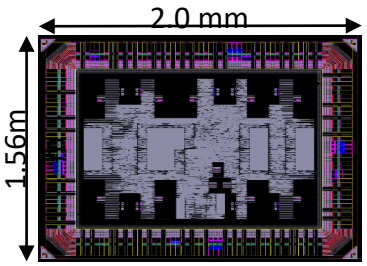




Chip layout

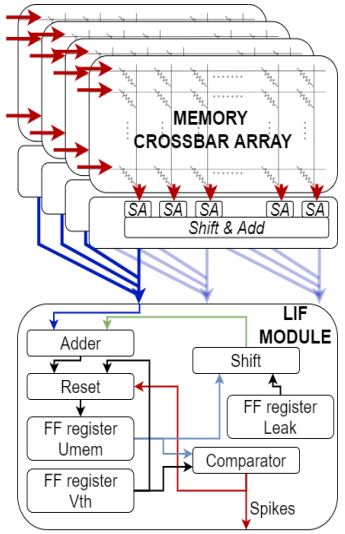
Spiking Neural Network (SNN) Accelerator

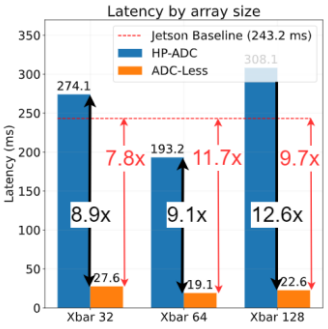




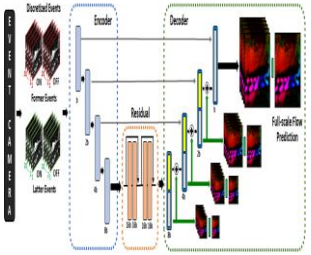
Chip layout

HW/SW co-design with ADC-Less IMC

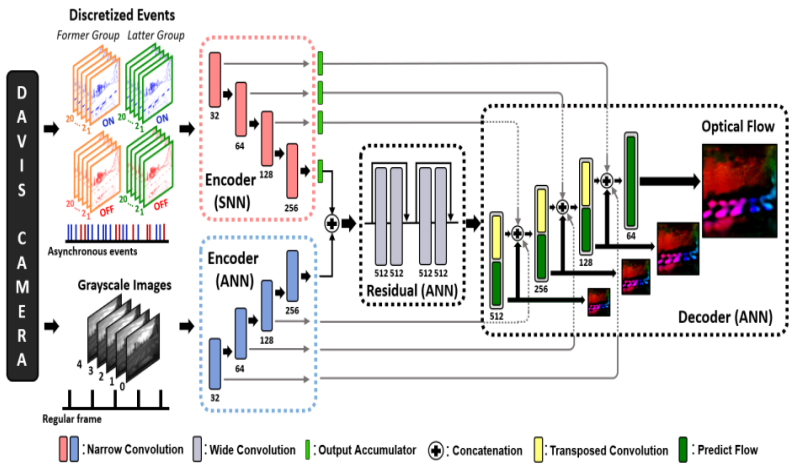




Intel Loihi



Hardware Architecture: Jetson TX-2



NVIDIA Jetson TX-2

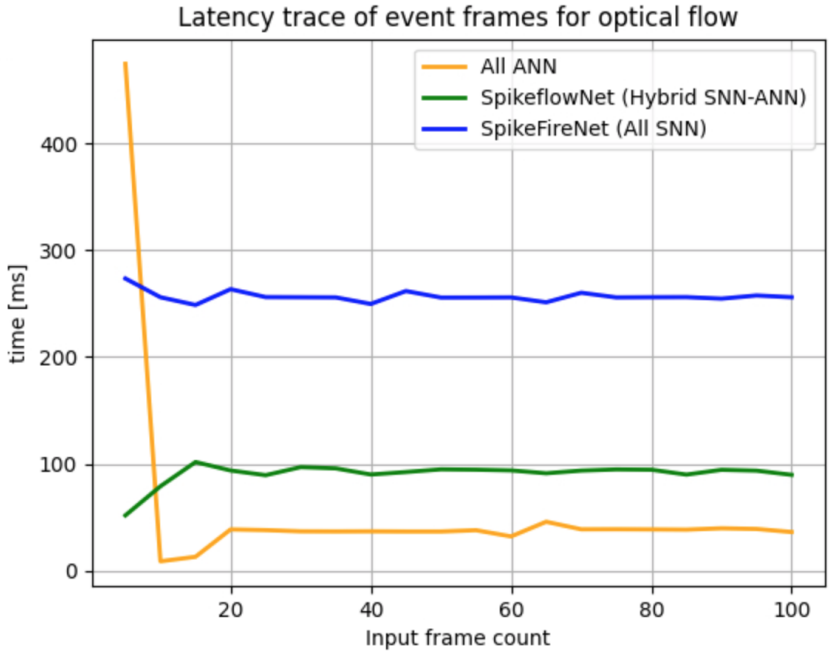
Example optical flow prediction

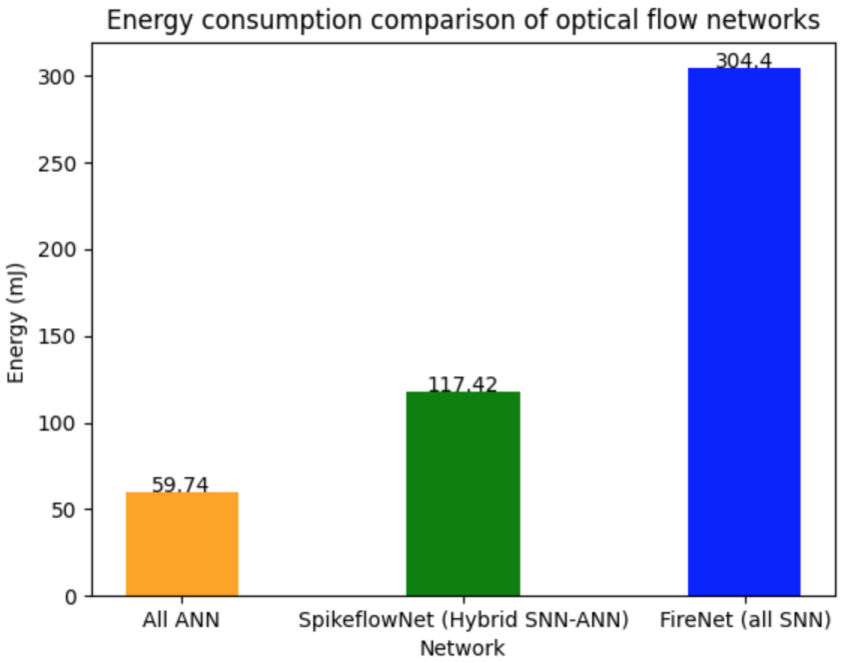
Gray image

Spike image

Flow prediction

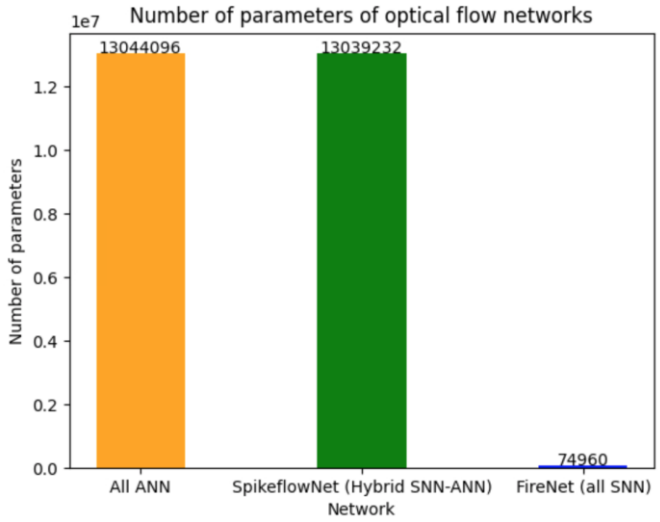
Performance estimation of optical flow networks





Observations

- Networks mapped to GPU cores in Jetson TX2 at max frequency
- Low model complexity doesn't translate to better inference performance!

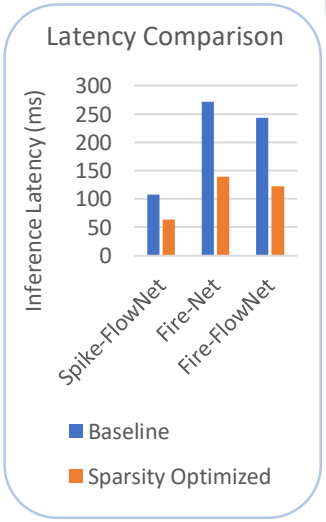


Hardware Implementations

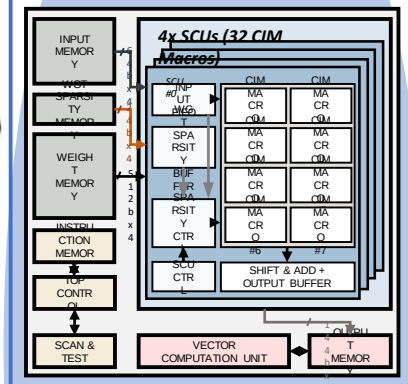
GPU-based baseline

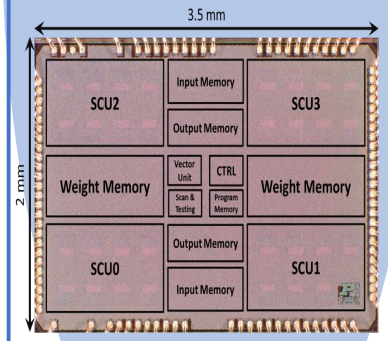
NVIDIA Jetson TX-2

Baseline: High latency due to inefficient handling of SNNs on GPUs.



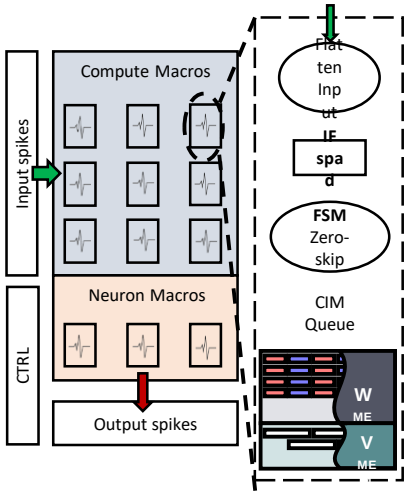
Adaptive-SNR Sparsity-Aware CiM

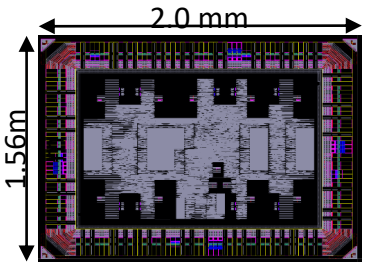




Chip layout

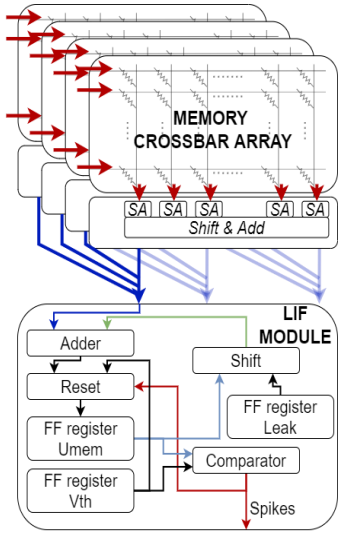
Spiking Neural Network (SNN) Accelerator

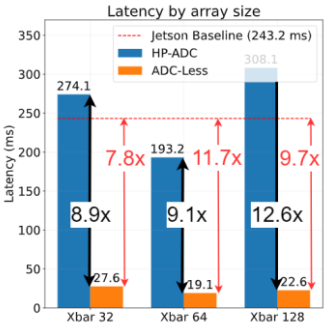




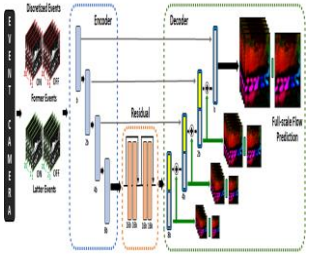
Chip layout

HW/SW co-design with ADC-Less IMC

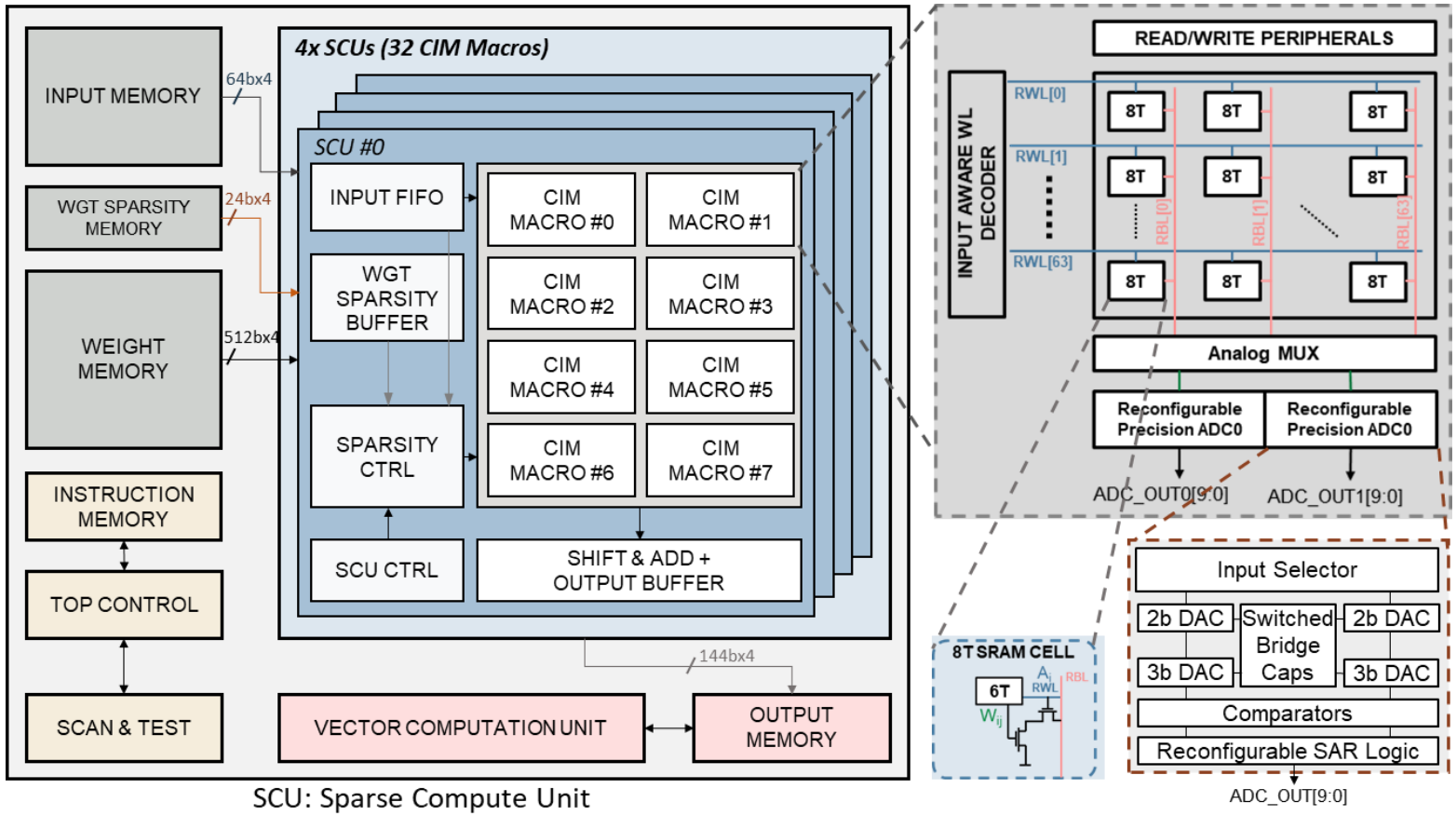




Intel Loihi



Energy Efficient DNN: Adaptive-SNR Sparsity-Aware CiM Core with Load Balancing Support

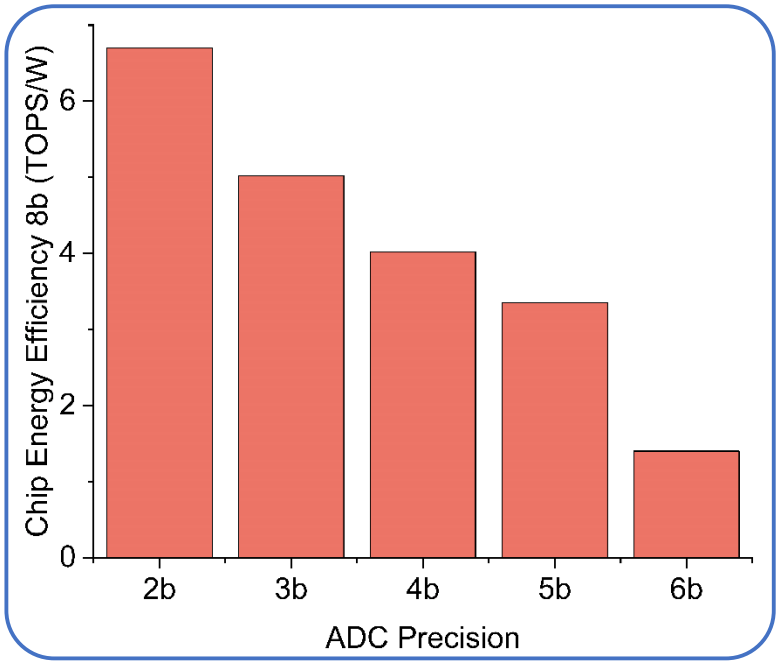


SCU: Sparse Compute Unit

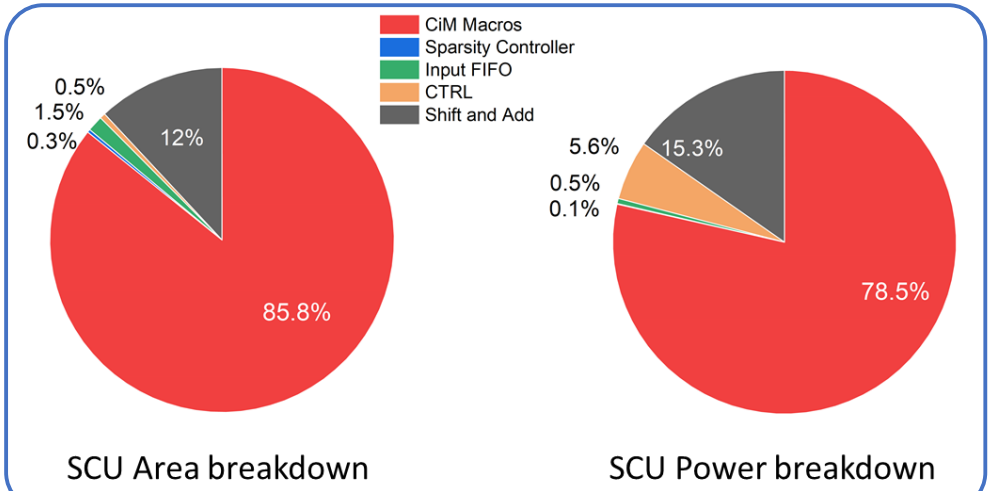
- Hierarchical Microarchitecture with Sparsity-aware Bit-Serial Compute Units and reconfigurable ADC
- Row Gating based on SNR requirements of DNN workloads
- On-chip row and column re-arrangement hardware support for load balancing

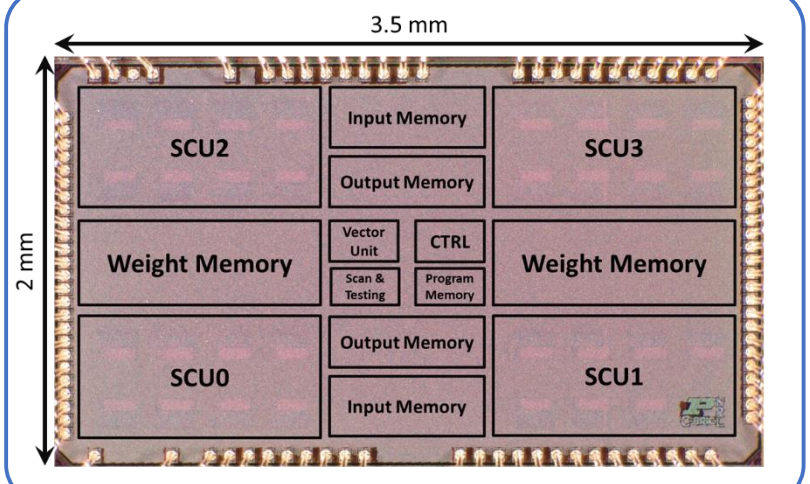
M. Ali, "A 35.5-127.2 tops/w dynamic sparsity-aware reconfigurable-precision compute-in-memory sram macro for machine learning", IEEE Solid-State Circuits Letters, 2021

Chip Results & Summary



Chip Summary	
Technology (nm)	65
Voltage (V)	1.2
Frequency (MHz)	100.806
Input/Weight Precision	4b/4b, 4b/8b, 8b/4b, 8b/8b
Output Precision	18b/22b
Total Area (mm^2)	7
CiM Macro Area (mm^2)	0.036
Total Digital SRAM (KB)	90.2
CiM SRAM (KB)	16
Performance (1b/1b operation)	117-552 GOPs
CIFAR-10 Accuracy (Resnet-20)	91.8%
Chip Energy Efficiency (8b/8b operation)	1.4-6.7 TOPs/W



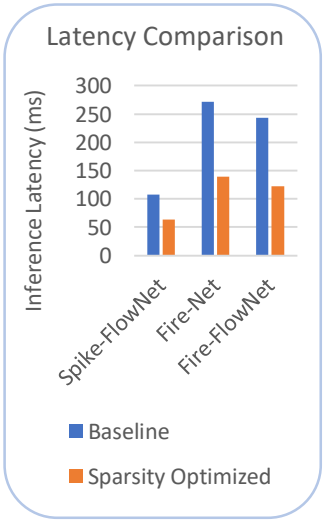


Hardware Implementations

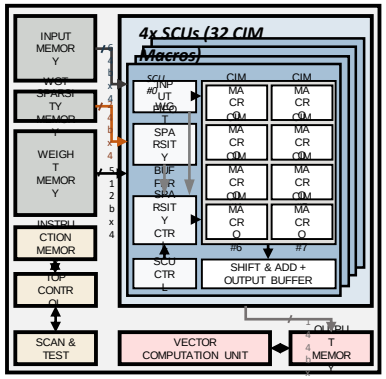
GPU-based baseline

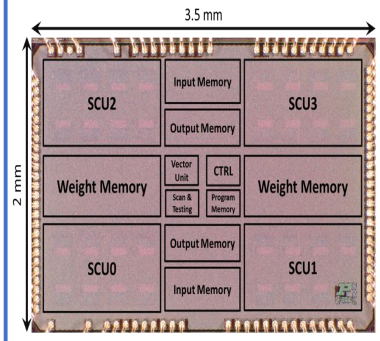
NVIDIA Jetson TX-2

Baseline: High latency due to inefficient handling of SNNs on GPUs.



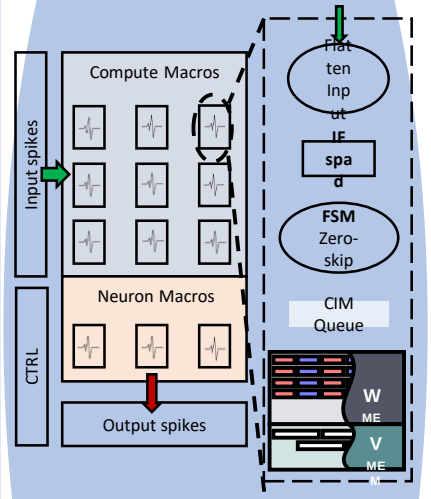
Adaptive-SNR Sparsity-Aware CiM

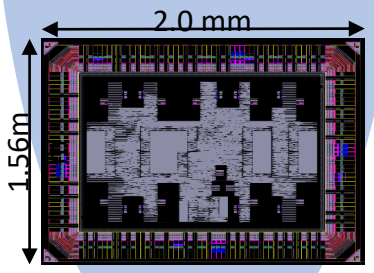




Chip layout

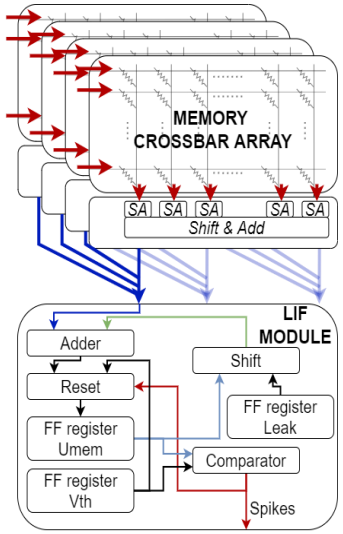
Spiking Neural Network (SNN) Accelerator

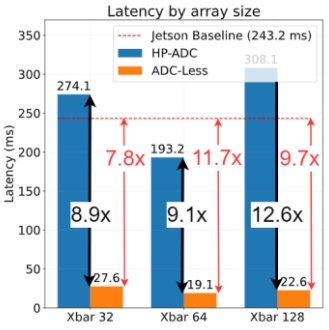




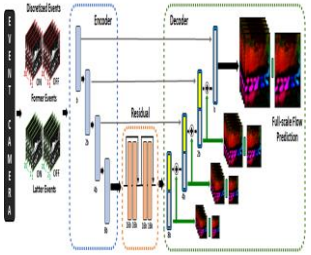
Chip layout

HW/SW co-design with ADC-Less IMC





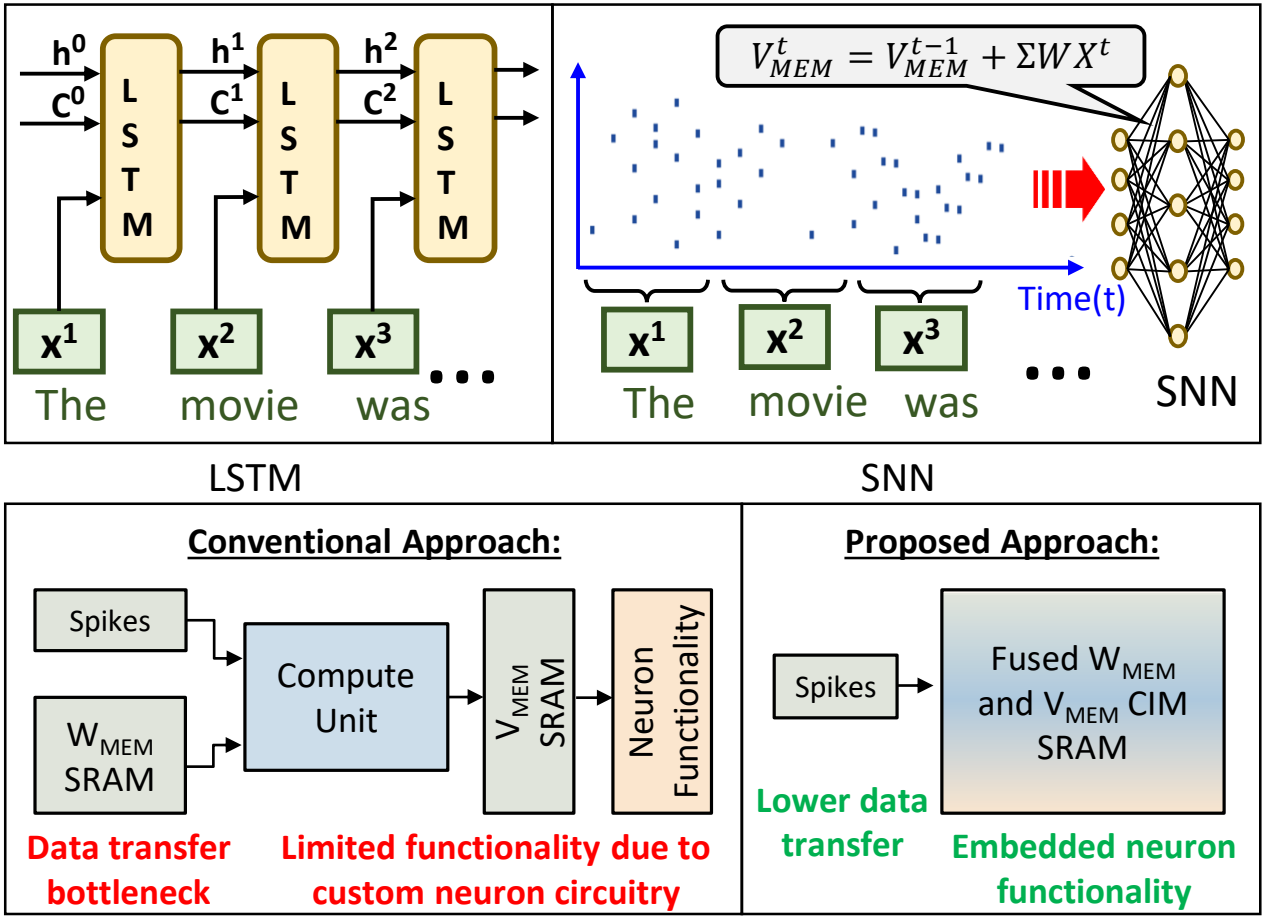
Intel Loihi



65 nm Spiking Neural Network (SNN) Accelerator based on in-memory Processing: Suitable with DVS Camera

- Motivation (DVS input…)

- Spiking Neural Networks (SNNs) can perform sequential learning tasks efficiently using spike-based membrane potential (Vmem) accumulation over several timesteps.
- However, the movement of Vmems creates additional memory accesses making data-transfer a bottleneck.
- Additionally, the sparsity in binary spike inputs can be leveraged for efficiency.



A. Agrawal et. al., "IMPULSE: A 65nm Digital Compute-in-Memory Macro with Fused Weights and Membrane Potential for Spike-based Sequential Learning Tasks", IEEE Solid-State Circuits Letters, 2021

65 nm Spiking Neural Network (SNN) Accelerator based on in-memory Processing

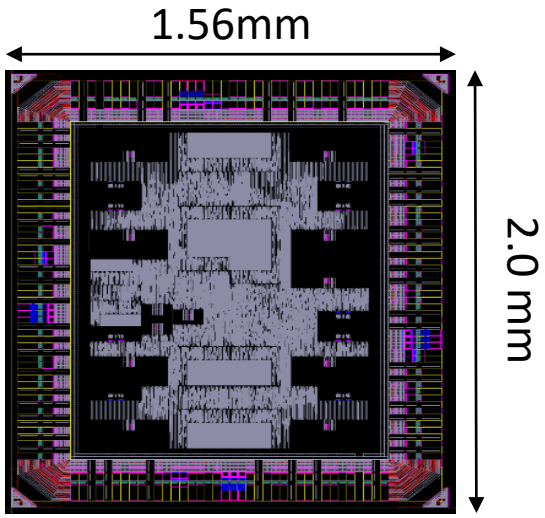
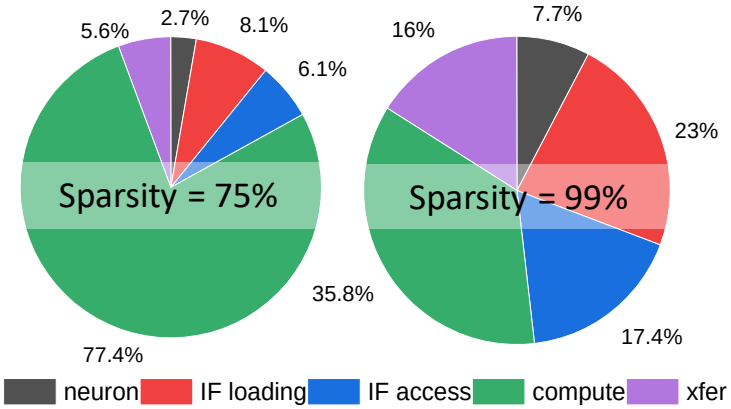
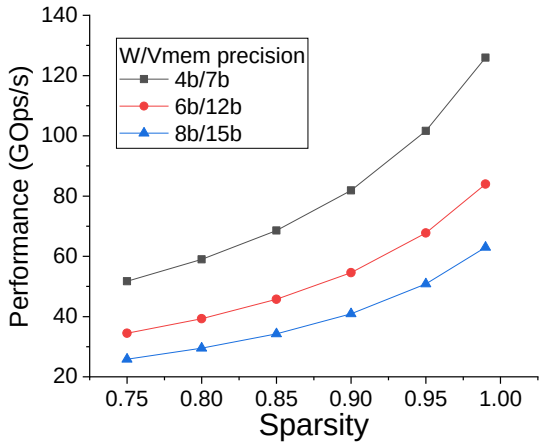


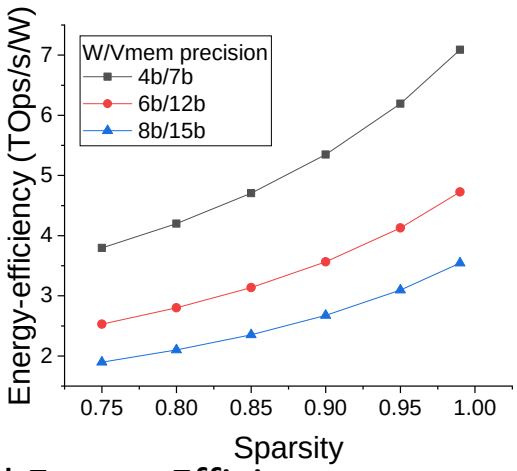
Fig. 1: Chip layout
Area = 3.12 mm^2

Results



Component-wise energy distribution at 75% and 99% input spike sparsity. Energy dominated by CIM even at very high sparsity.





Performance and Energy Efficiency

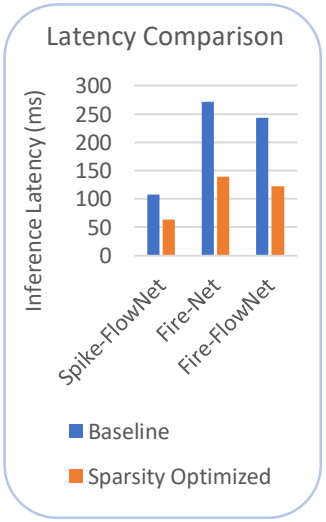
Both the performance and energy efficiency increase with increasing sparsity (zero-skipping) and decreasing bit-precision (more parallelism).

Hardware Implementations

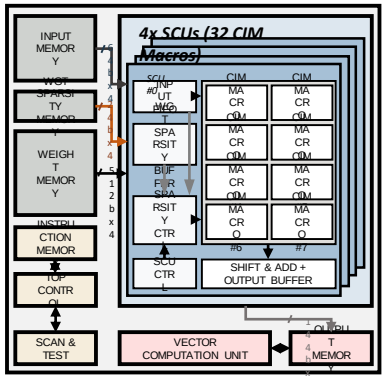
GPU-based baseline

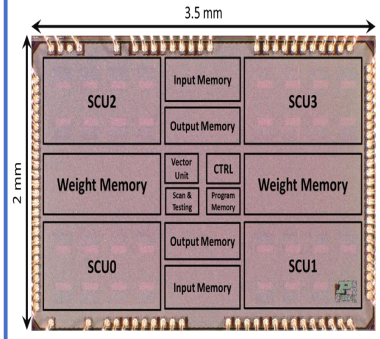
NVIDIA Jetson TX-2

Baseline: High latency due to inefficient handling of SNNs on GPUs.



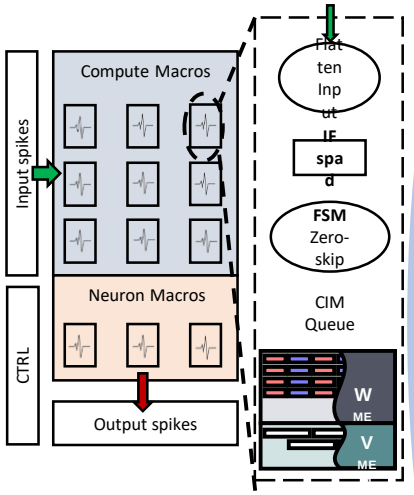
Adaptive-SNR Sparsity-Aware CiM

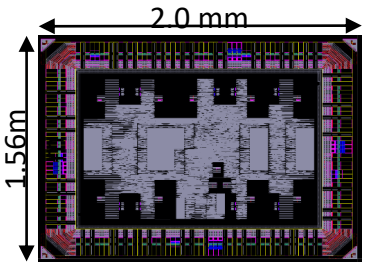




Chip layout

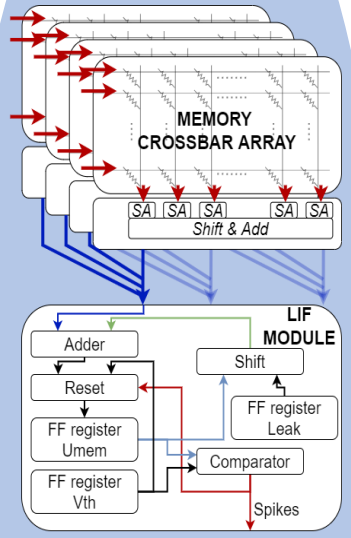
Spiking Neural Network (SNN) Accelerator

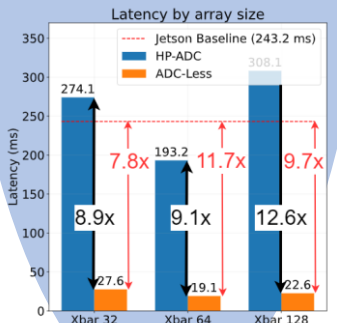




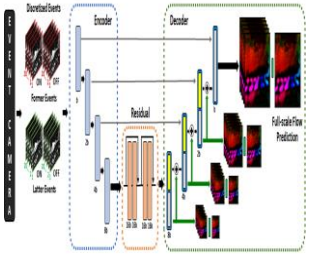
Chip layout

HW/SW co-design with ADC-Less IMC



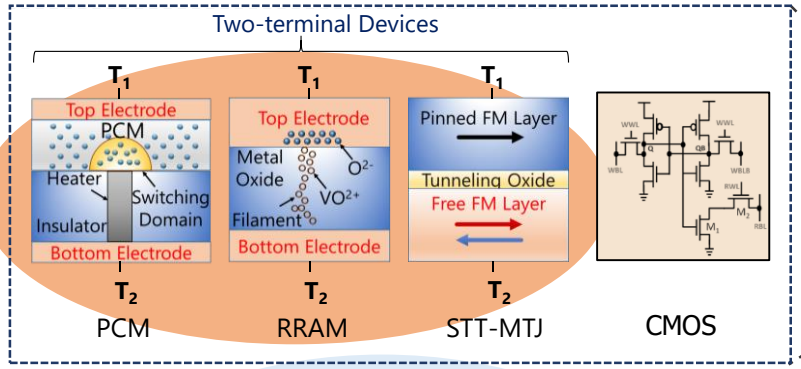


Intel Loihi

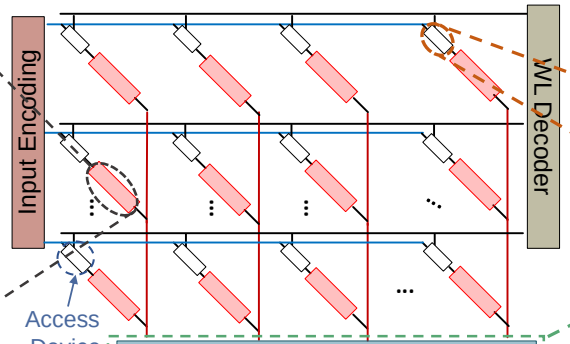


Hardware Architecture: CiM

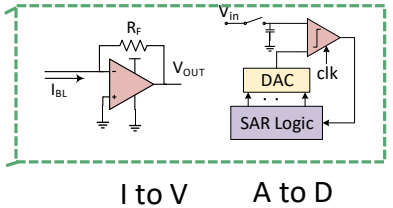
In-Memory Computing Memory Devices



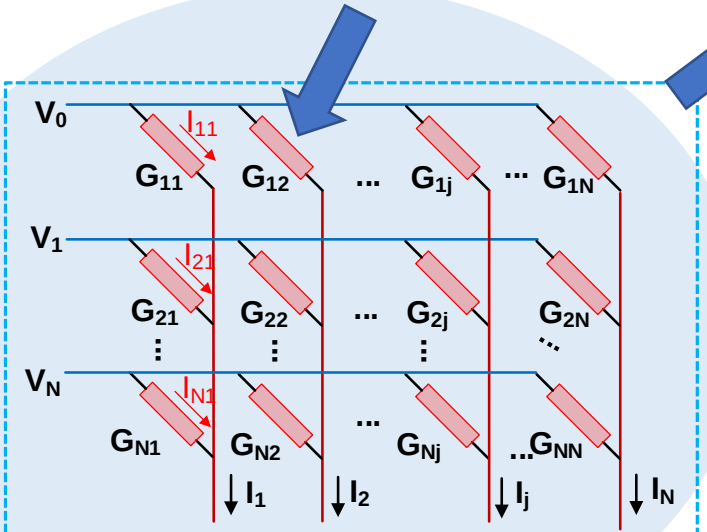
Chakraborty et. al. Resistive Crossbars as Approximate Hardware Building Blocks for Machine Learning: Opportunities and Challenges, Proc. of IEEE, 2020

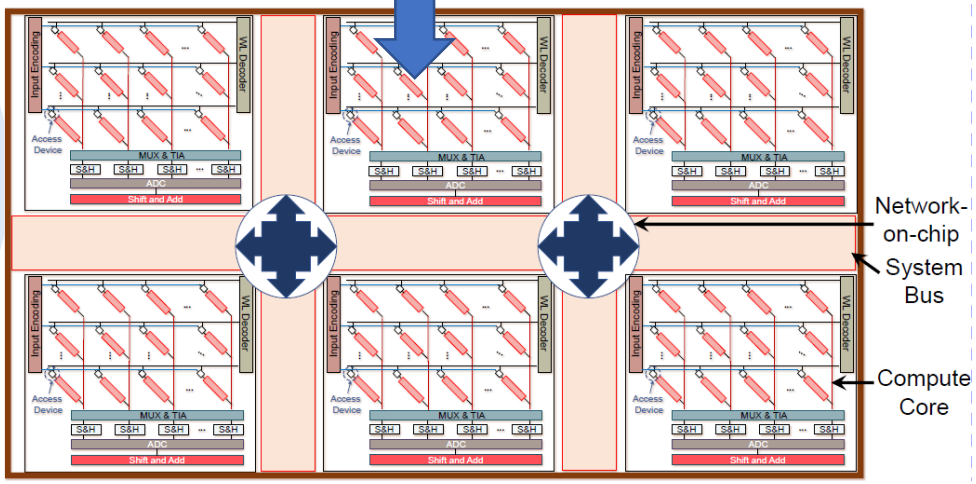


The Peripherals



I to V A to D



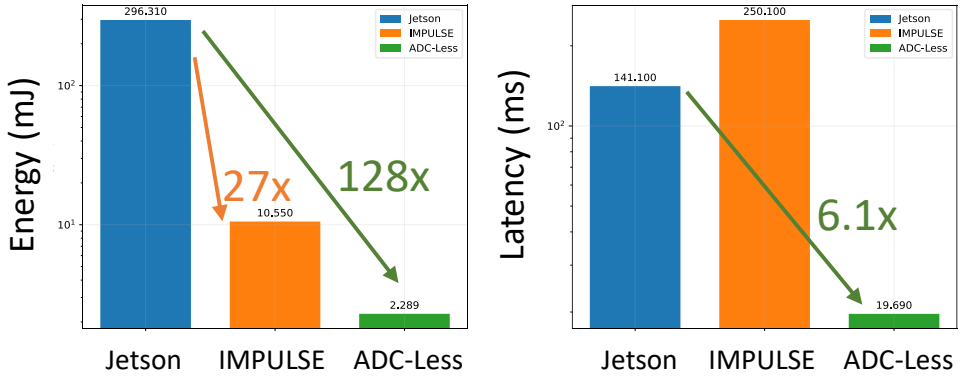


Efficient MVM

Spatially Distributed Cores

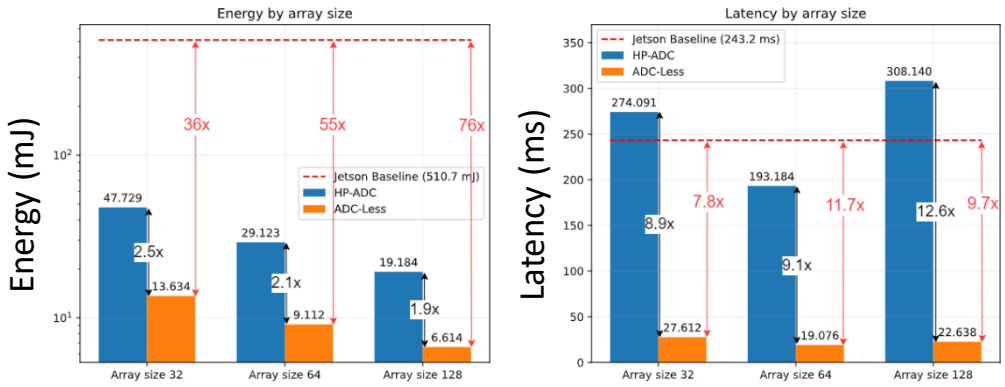
Energy and latency of the IMC architectures

Cross-architecture Comparison: FireFlowNet



Both IMC architectures (IMPULSE and ADC-Less) require less energy consumption compared to the Jetson platform.
In addition, the ADC-Less can improve the latency enabling real-time inference.

ADC-Less IMC energy and latency analysis: Spike-FlowNet

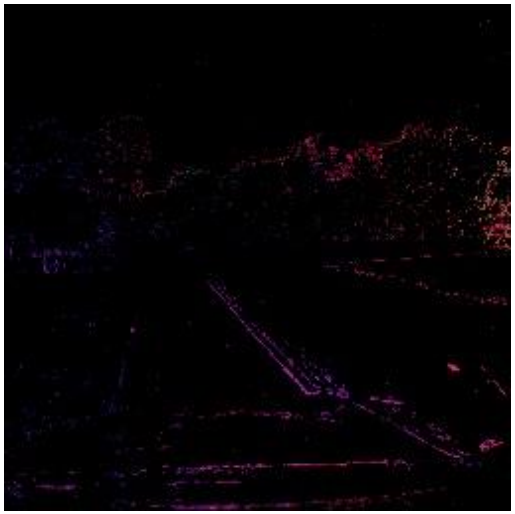


36-76x less energy consumption and **7.8–12x faster** than Jetson platform.

1.9-2.5x less energy consumption and **8.9–12.6x faster** than a conventional HP-ADC IMC.

Hardware-aware training for the ADC-Less IMC

Optical flow prediction of the Fully-Spiking FlowNet (FS-FN) during the Hardware-aware training.



Full precision

ADC-Less training

Performance on MVSEC dataset (dt=1) [AEE lower is better]

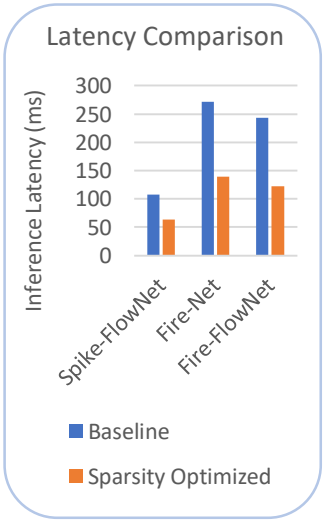
Model	IN1 – AEE	IN2 – AEE	IN3 - AEE	Training
FSFN	**0.82**	**1.21**	**1.07**	Full-precision
FSFN	0.88	1.39	1.18	ADC-Less training
Spike-FlowNet	0.84	1.28	1.11	Full-precision

Hardware Implementations

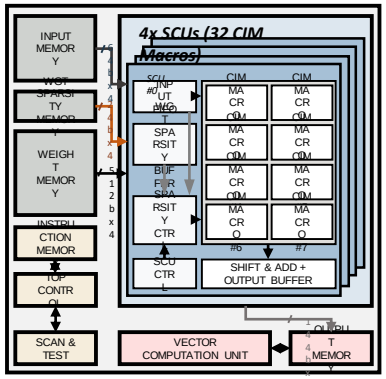
GPU-based baseline

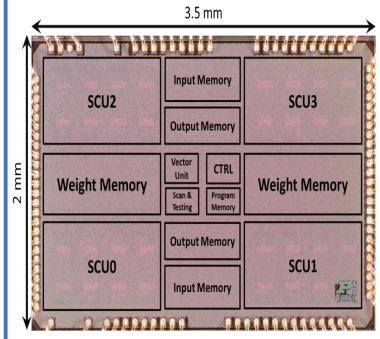
NVIDIA Jetson TX-2

Baseline: High latency due to inefficient handling of SNNs on GPUs.



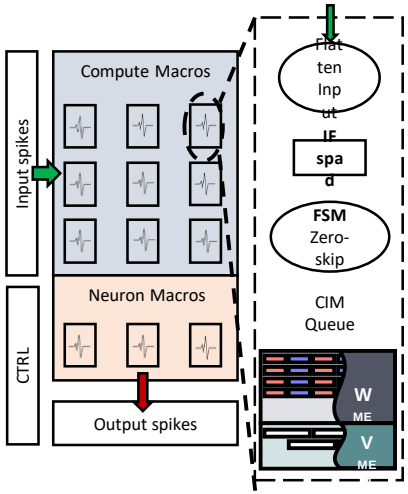
Adaptive-SNR Sparsity-Aware CiM

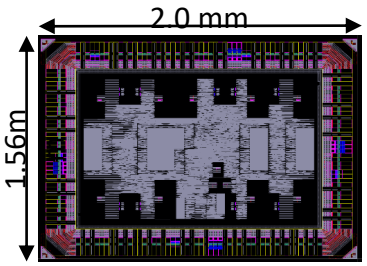




Chip layout

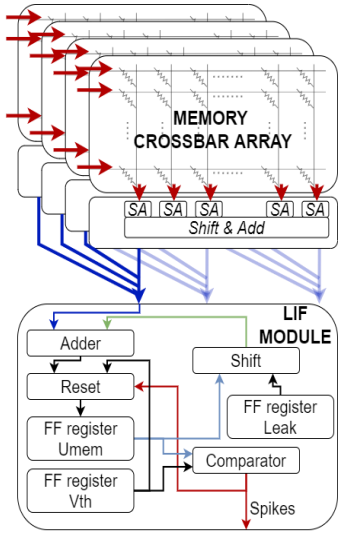
Spiking Neural Network (SNN) Accelerator

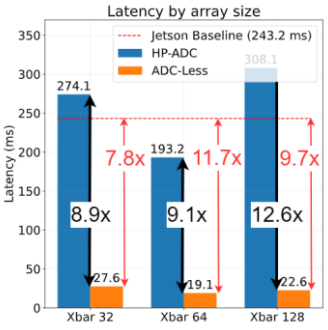




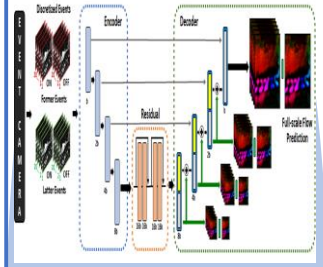
Chip layout

HW/SW co-design with ADC-Less IMC



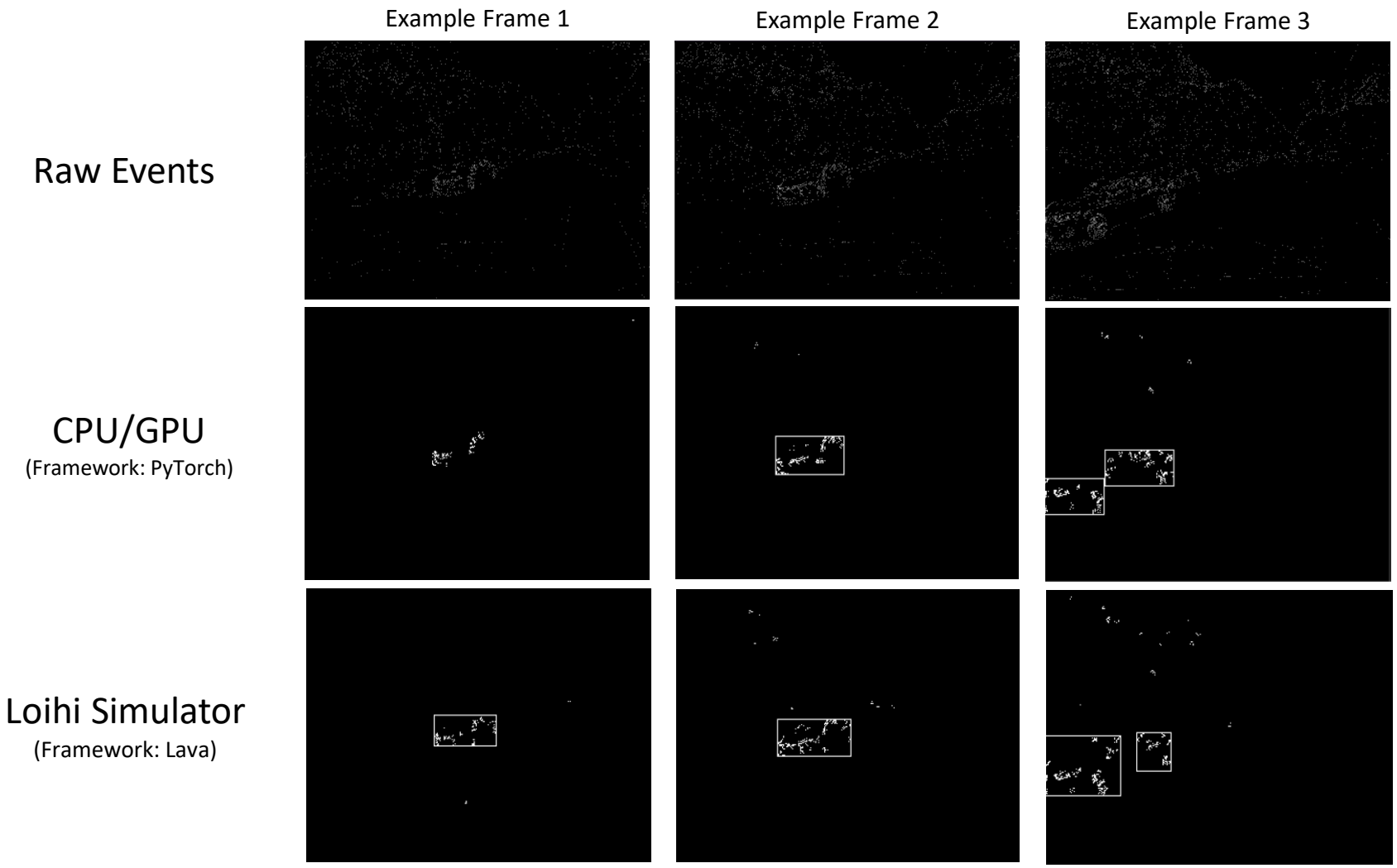


Intel Loihi

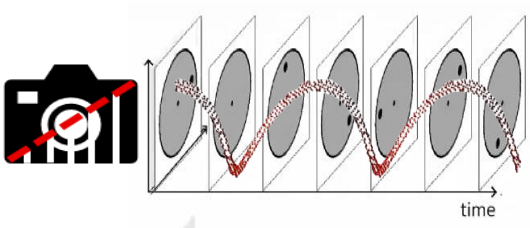


Results of Implementing DOTIE on *Loihi*

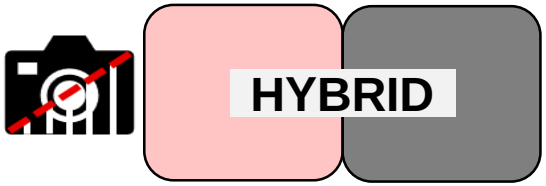
- Intersection over union (IoU) between object detection bounding boxes of **~81%** from single car events dataset



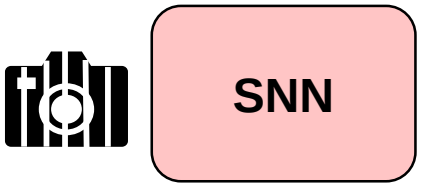
Key Takeaways – Sensors and Algorithms



Sensor-fusion of Frame and Event data exploits their **complementary benefits** improving overall performance



Hybrid SNN-ANN models **naturally handle event data** while preserving **performance benefits** and **ease of training** of ANNs

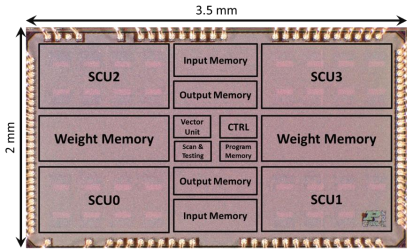


Fully-Spiking Architectures better capture timing information and lead to **lightweight models** suitable for the edge

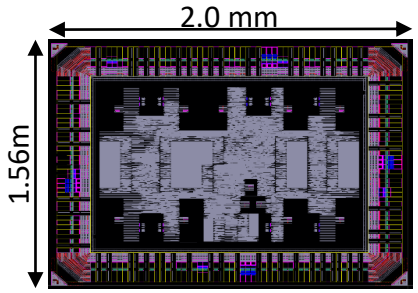
Next-generation datasets will enable movement of simulated and example problems into real world data

These techniques improve the current **state-of-the art,** both in terms of **accuracy** and **efficiency** on several tasks for **vision-based autonomous navigation**

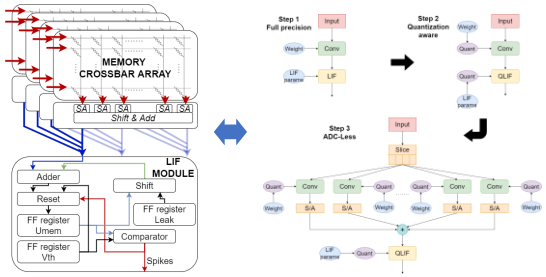
Key Takeaways – Efficient Hardware Platforms



Adapting to SNR and exploiting **sparsity in a workload** can **significantly improve the overall performance** of IMC architectures**.**



Specialized hardware accelerators for Spiking Neutral Networks focused on reducing the membrane potential overhead can **give better performance and energy benefits**.



Hardware/Software co-design approaches can lead to **energy-efficient implementations** based on co-optimization processes.

Hardware architectures and design techniques enables the deployment of **energy efficient vision-based autonomous navigation**