

From Algorithms to Devices: Neuromorphic Computing for Energy- Efficient Autonomy

Kaushik Roy

kaushik@purdue.edu



Biological system



Despite tremendous progress in AI, 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



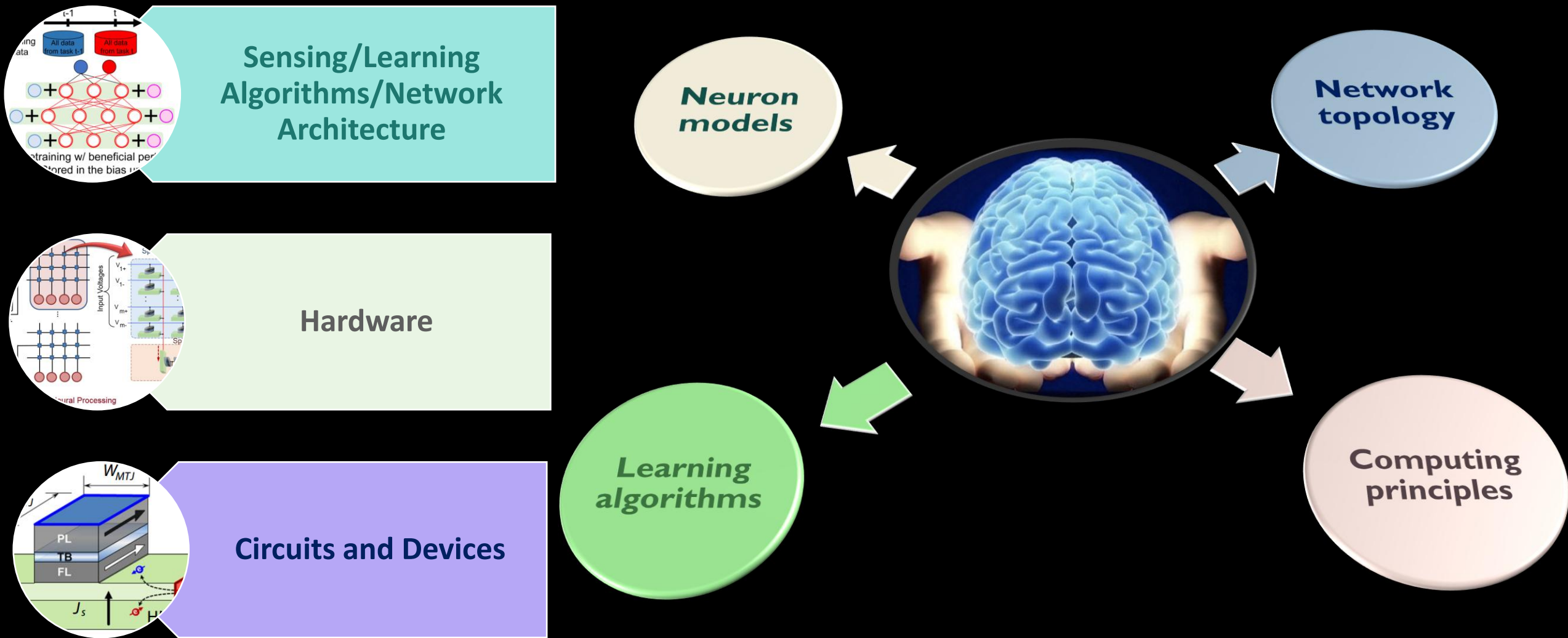
Flying monkey UAV,
UPenn
~1-2W compute

VS

Fruit fly
~uW compute



Can We Take Cues from the Brain?



Example cross-layer design: drone navigation



Tasks

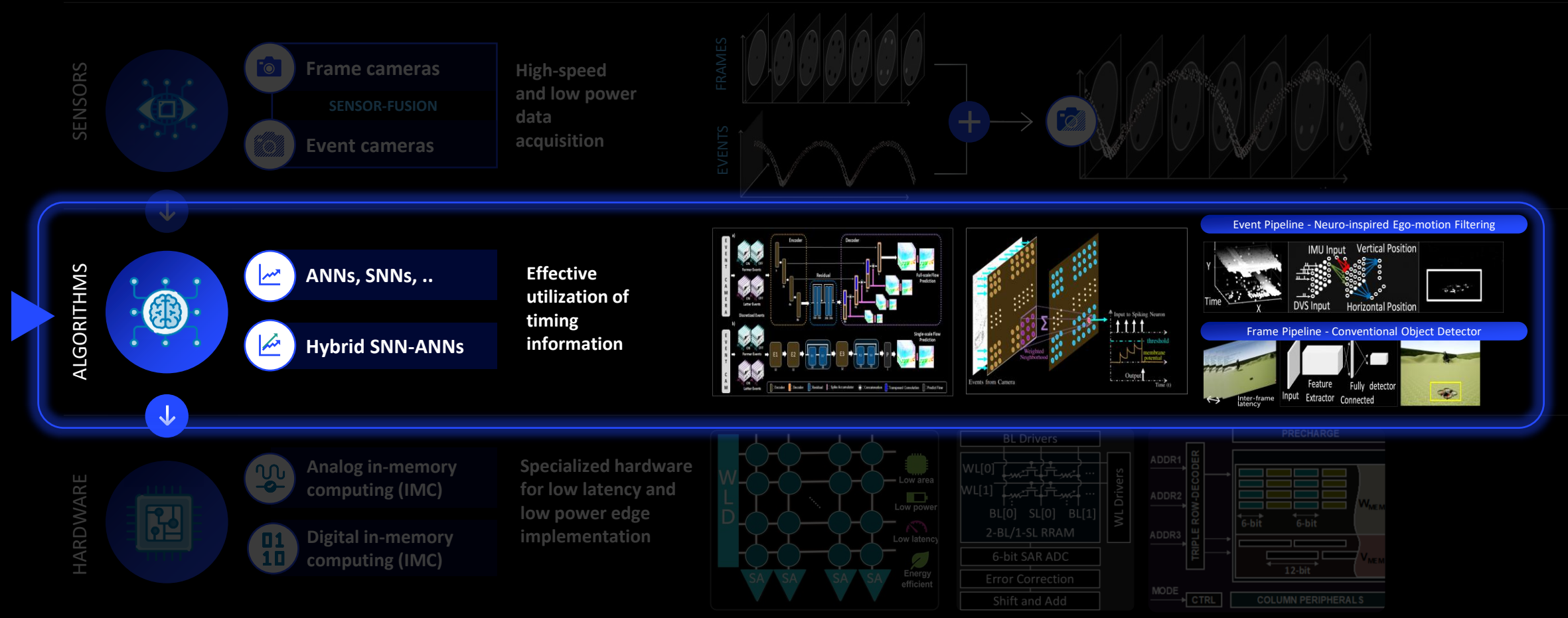
PERCEPTION

- Optical flow & depth estimation
- Segmentation
- Object detection and tracking

ACTION

- Localization and path planning
- Reinforcement learning and control

Example cross-layer design: drone navigation

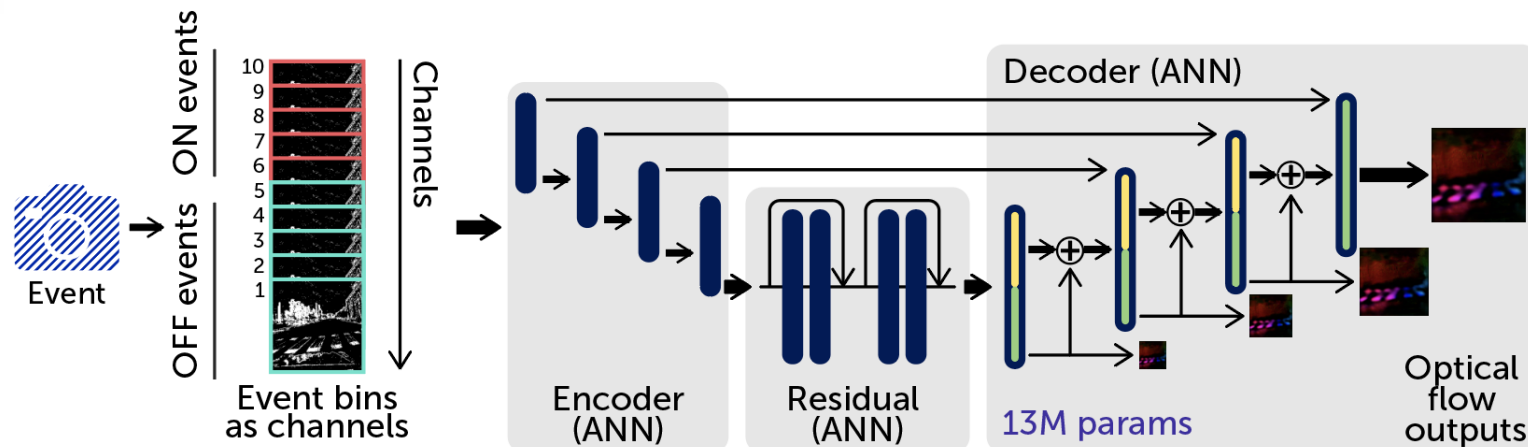


Comparing architectures for optical flow estimation

Average endpoint error (AEE) & energy evaluation on MVSEC dataset

- Spiking layer
- ANN layer
- Output accumulator
- Concatenation
- Transposed convolution
- Predict flow

A Fully-ANN model



BASELINE

- Spatial only
- NO Temporal Information handling

AVERAGE AEE: 4.77

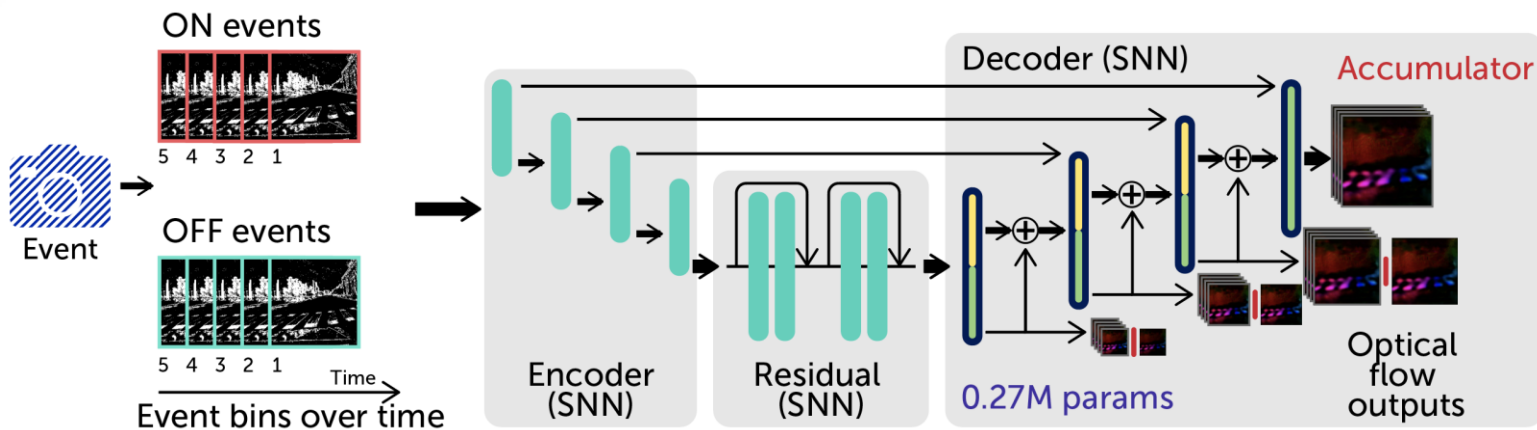
AVERAGE INFERENCE
ENERGY: 24.54 mJ

Comparing architectures for optical flow estimation

Average endpoint error (AEE) & energy evaluation on MVSEC dataset

- Spiking layer
- ANN layer
- Output accumulator
- Concatenation
- Transposed convolution
- Predict flow

B Fully-SNN model



Fully-Adaptive SNN with Learnable Neuronal Dynamics for Effective Temporal Processing

AVERAGE AEE: 4.30

LOWER AEE: 10%

AVERAGE INFERENCE ENERGY:
2.40 mJ

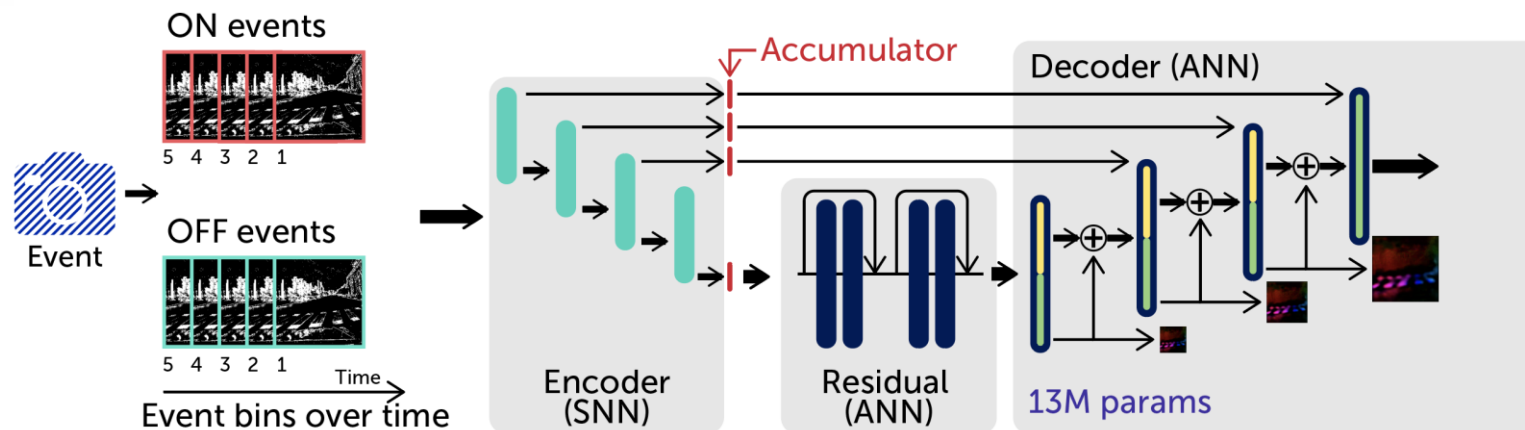
LOWER ENERGY: 10.2X

Comparing architectures for optical flow estimation

Average endpoint error (AEE) & energy evaluation on MVSEC dataset

- Spiking layer
- ANN layer
- Output accumulator
- Concatenation
- Transposed convolution
- Predict flow

C SNN-ANN hybrid



Encoder SNN with Fixed Neuronal Dynamics for Temporal Processing Residual & Decoder ANN for Spatial Processing

AVERAGE AEE: 3.72

LOWER AEE: 22%

AVERAGE INFERENCE ENERGY:
20.46 mJ

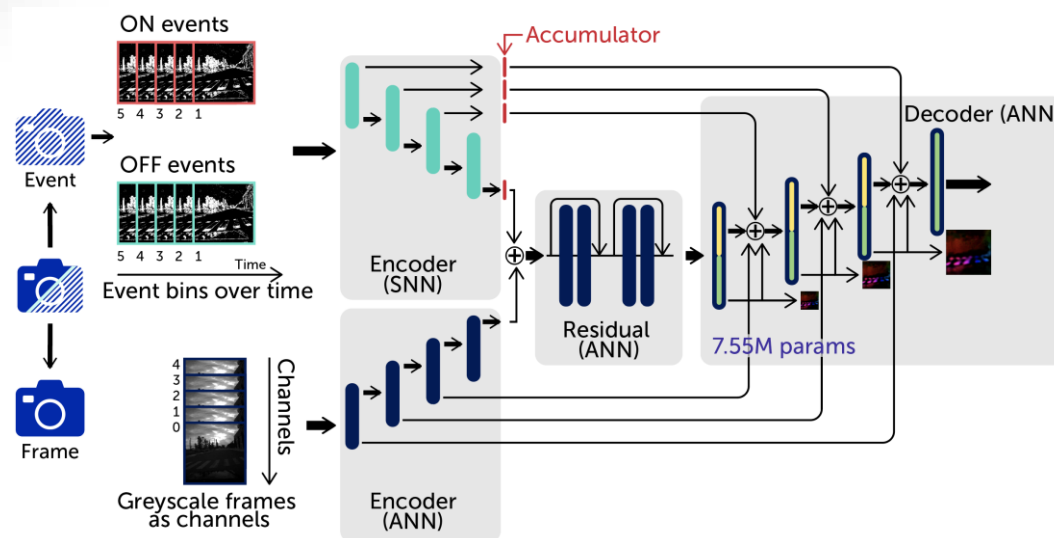
LOWER ENERGY: 1.21X

Comparing architectures for optical flow estimation

Average endpoint error (AEE) & energy evaluation on MVSEC dataset

- Spiking layer
- ANN layer
- | Output accumulator
- $\oplus$ Concatenation
- | Transposed convolution
- | Predict flow

D Sensor fused SNN-ANN hybrid



Events and frames as Input

- SNN for Events
- ANN for Frames

AVERAGE AEE: 2.90

LOWER AEE: 39%

AVERAGE INFERENCE ENERGY:
13.11 mJ

LOWER ENERGY: 1.88X

der SNN with Fixed
onal Dynamics for Temporal
ssing Residual & Decoder ANN
patial Processing

AVERAGE AEE: 3.72

OWER AEE: 22%

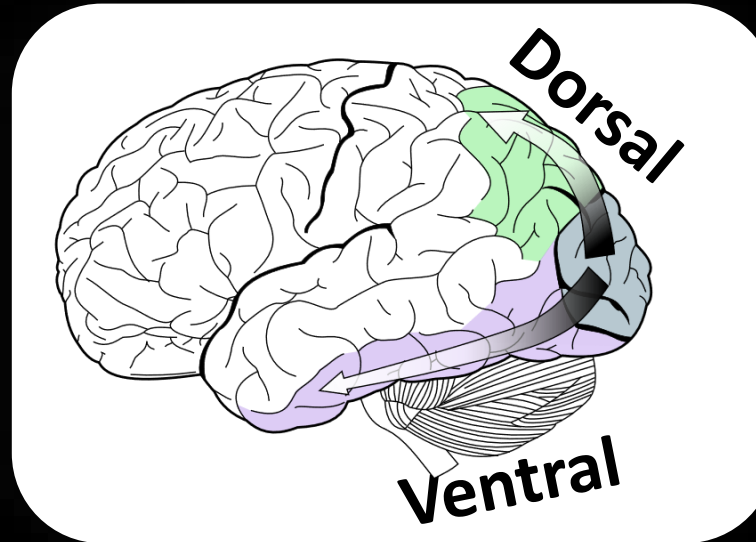
AVERAGE INFERENCE ENERGY:
46 mJ

OWER ENERGY: 1.21X

Object Detection and Tracking: Two Stream Hypothesis

○ Ventral (*what*) stream

- Recognition functionality
- High spatial frequency
- Slow processing
- High details
- Visual input focused at fovea
- Object-centered



○ Dorsal (*where*) stream

- Guiding functionality
- High temporal frequency
- Fast processing
- Low details
- Visual input across retina
- Viewer-centered



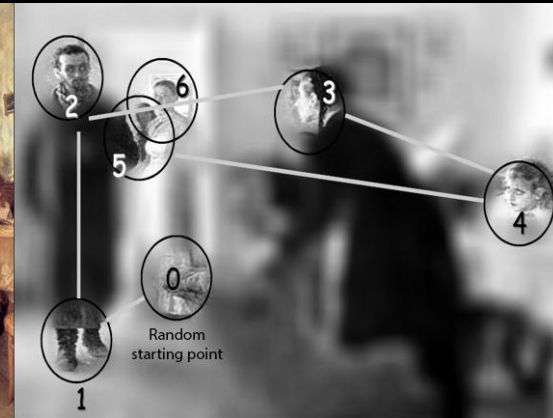
(a) $f_0 = 50$



(b) $f_0 = 100$



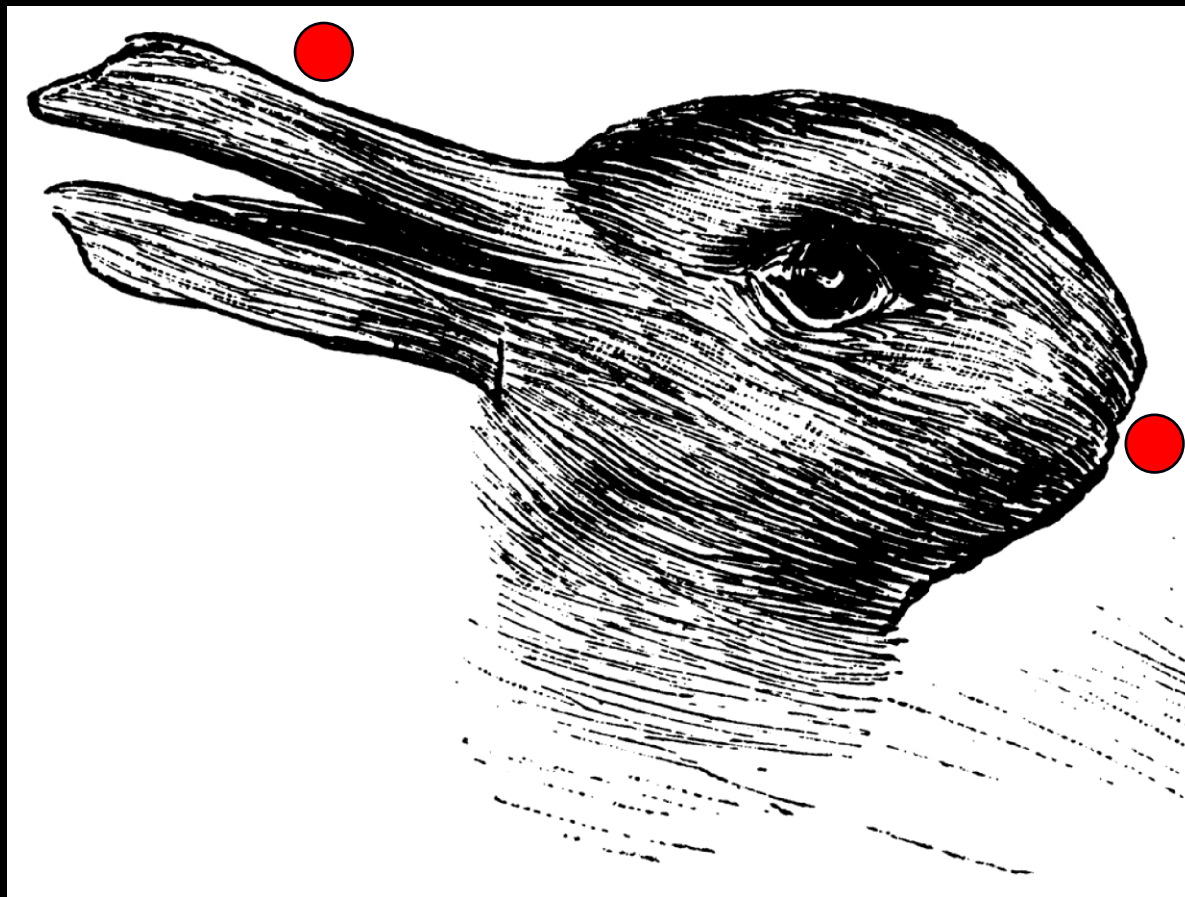
(c) $f_0 = 150$



Optical Illusion

Duck-Rabbit optical illusions: at any moment human vision perceives...

... or duck



... either rabbit

Foveation & Saccades for Optical Illusion

- Explored alignment of our method with human perception using a variation of the optical illusion



Cross-layer design: sensors, algorithms, hardware

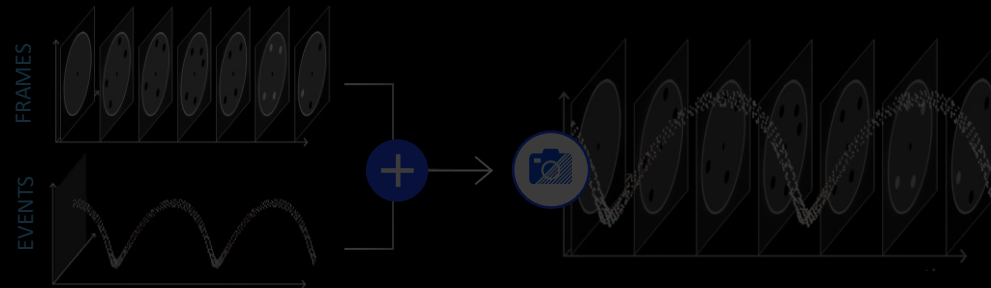


SENSORS

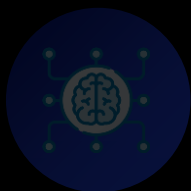


- Frame cameras
- SENSOR-FUSION
- Event cameras

High-speed
and low power
data
acquisition

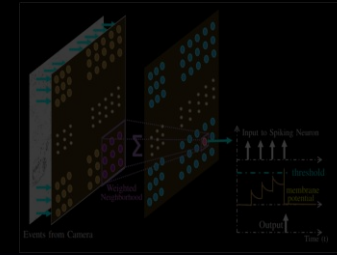
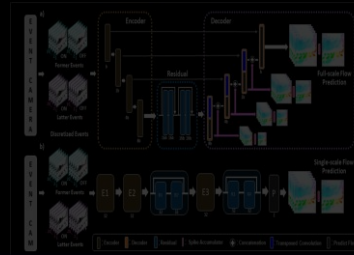


ALGORITHMS

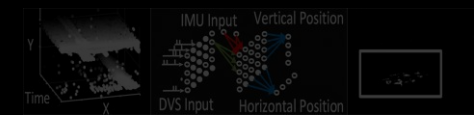


- Spiking networks
- Hybrid SNN-ANNs

Effective
utilization of
timing
information



Event Pipeline - Neuro-inspired Ego-motion Filtering



Frame Pipeline - Conventional Object Detector

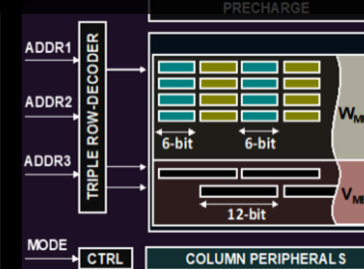
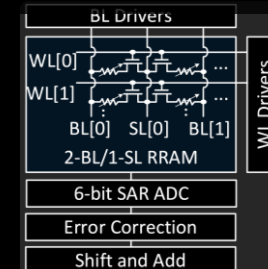
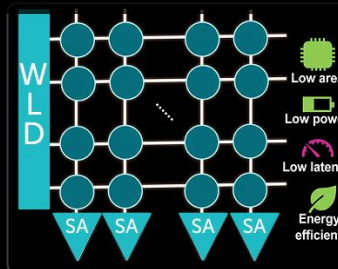


HARDWARE

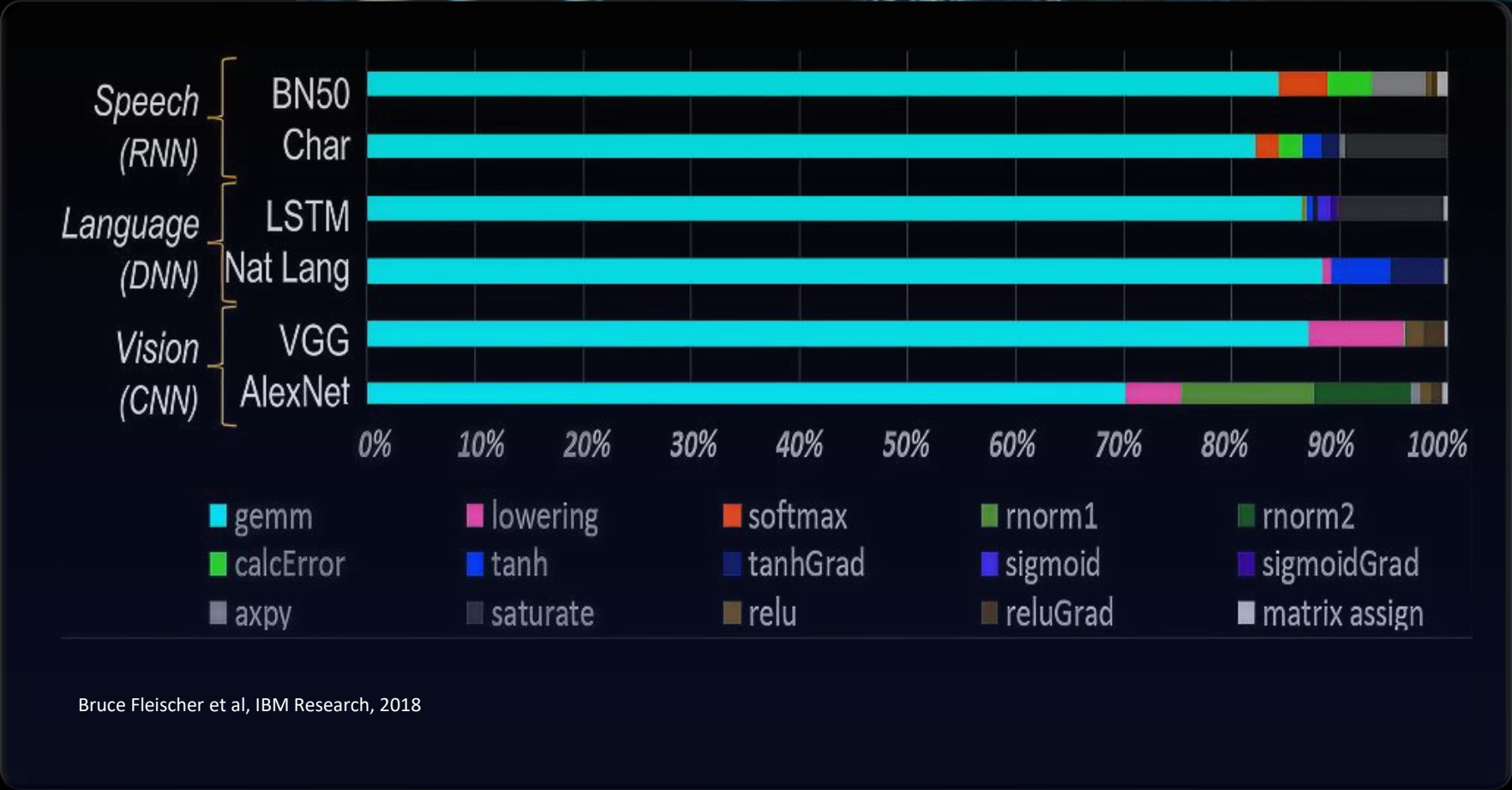


- Analog in-memory computing (AIMC)
- Digital in-memory computing (DIMC)

Specialized hardware
for low latency and
low power edge
implementation



Hardware fabrics



Hardware fabrics



Circuits and architectures that can efficiently implement the algorithms (possibly embody computing principles from the brain)

- GPUs, systolic arrays,...
- Quantization, sparsity
- Memory architecture
- Near-/in-memory computing (data movement cost)
- Approximate (quantization..) and stochastic hardware
- Neuromorphic devices and interconnects
- Event-driven computation



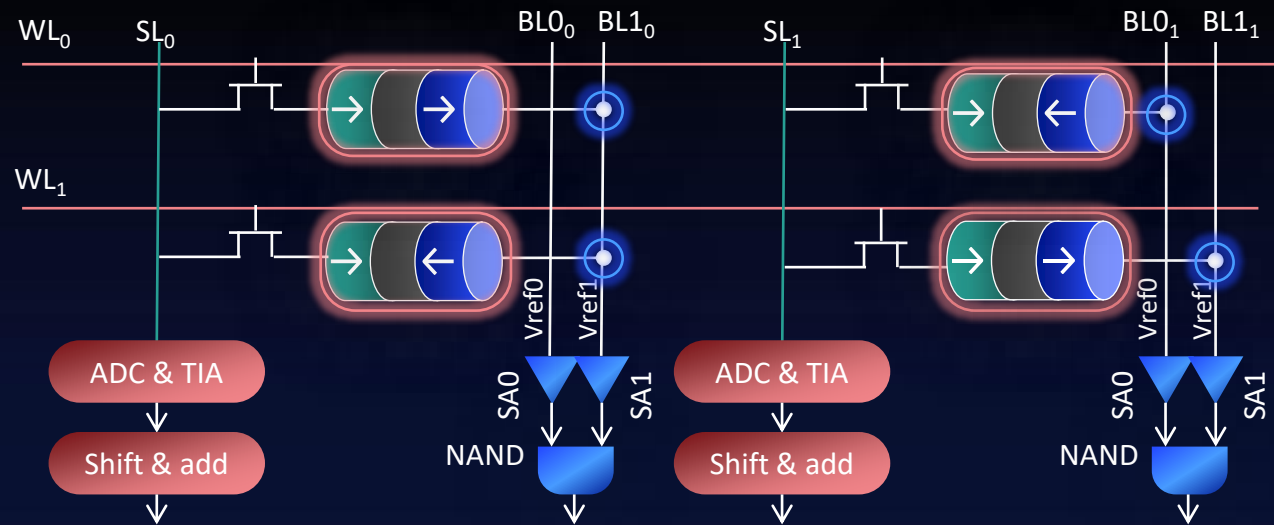
Bruce Fleischer et al, IBM Research, 2018

EXPONENTIAL
Softmax Tanh Sigmoid

Compute In Memory

Multiverse of MRAM Compute in Memory

ROM overlaying RAM (in MRAM technology)



RAM data	
1(P)	0(AP)
0(AP)	1(P)

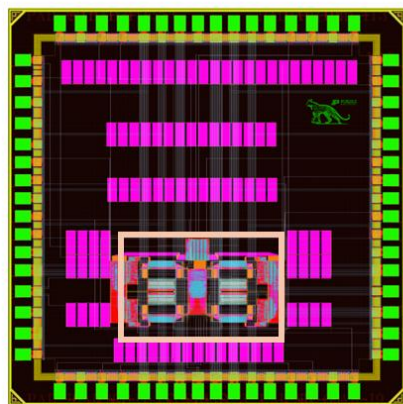
ROM data	
1(BL1)	0(BL0)
1(BL1)	1(BL1)

→ ROM behavior supported through 1 extra bitline (NAND and Sense Amplifiers supporting peripherals)

→ ROM data storage achieved through physical connections to either bitline

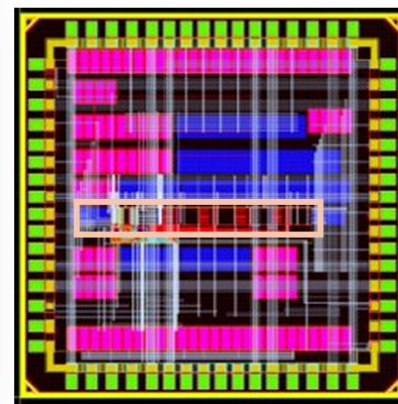
→ RAM data stored in the MRAM (P state vs AP state)

CiMs Fabricated in FEOL GF 22FDX + BEOL (IMEC MTJs)



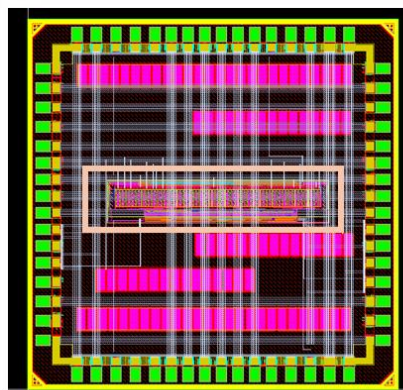
3T-2R STT-MRAM CiM

- Energy Efficient
- Outlier aware



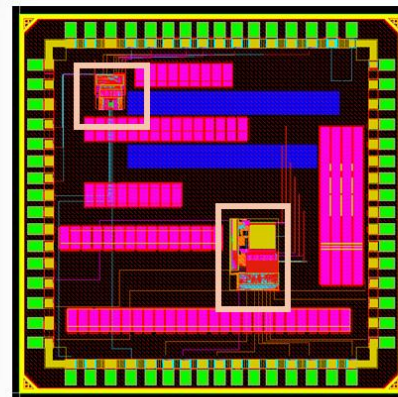
1T-1R STT-MRAM CiM

- Exact MVMs (lower word-line activation)
- Smallest & foundry standard bit-cell



5T-2R STT-MRAM Digital CiM

- Exact MVMs
- Low latency due to per-column adder tree

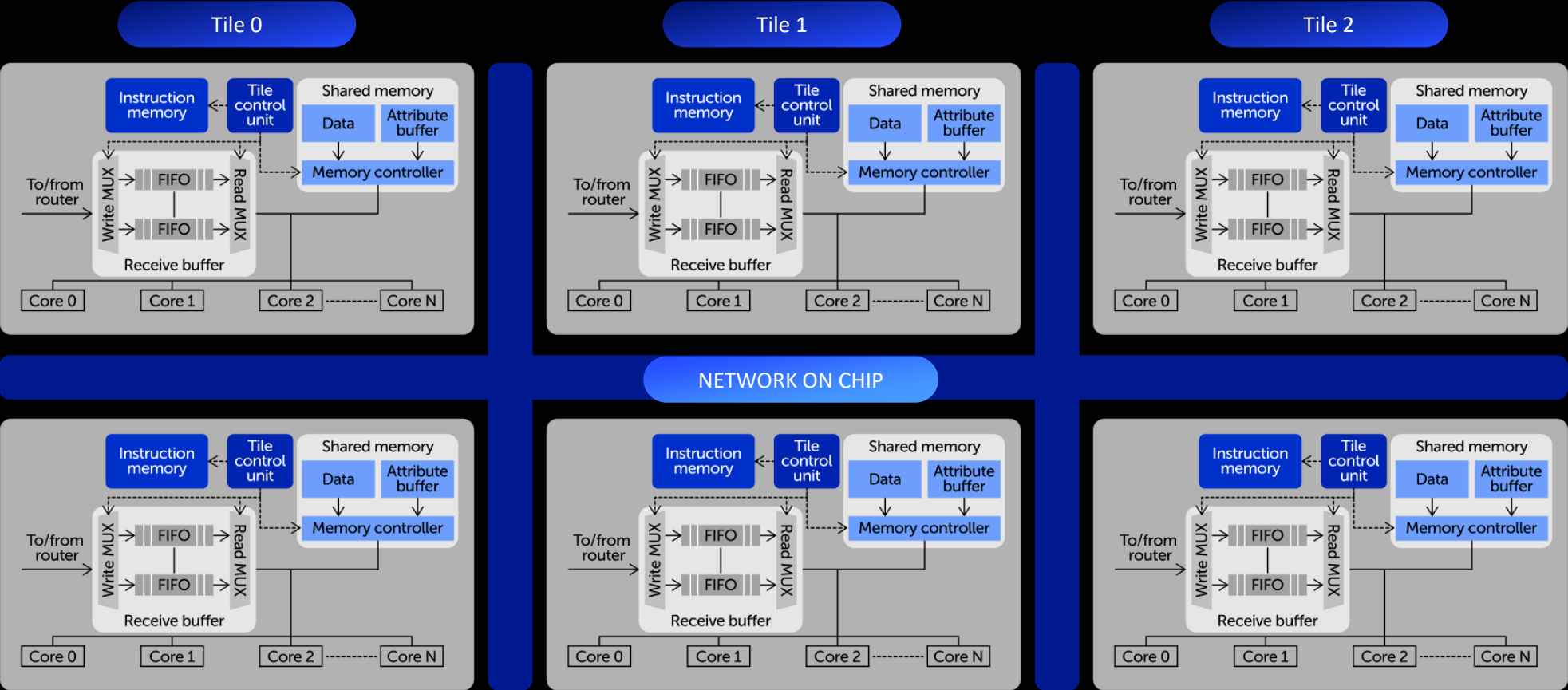


2T-2R ADC-Less CiM + RNG

- ADC-Less CiM
- MRAM based RNG for probabilistic computing

Hardware architecture

PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning Inference



- ML INFERENCE
- Massively parallel accelerator
- ↓
- Amenable to data-level parallelism
- ↓
- Highly efficient

Vision based navigation



SENSORS

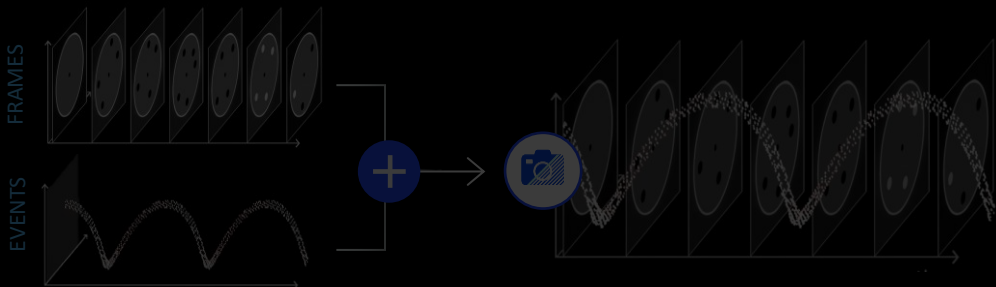


Frame cameras

SENSOR-FUSION

Event cameras

High-speed
and low power
data
acquisition



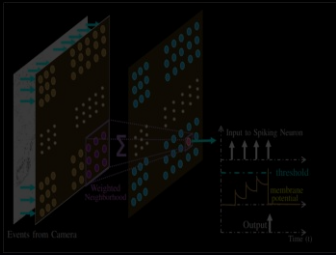
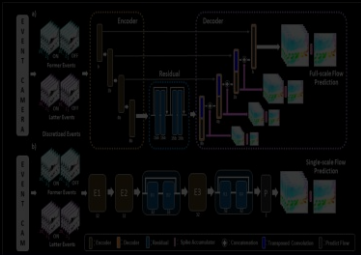
ALGORITHMS



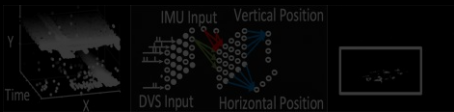
Spiking networks

Hybrid SNN-ANNs

Effective
utilization of
timing
information



Event Pipeline - Neuro-inspired Ego-motion Filtering



Frame Pipeline - Conventional Object Detector



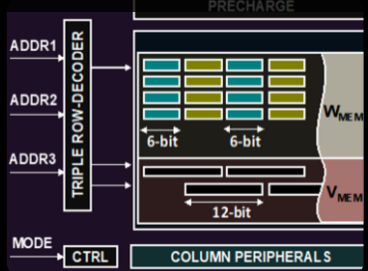
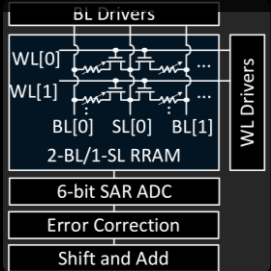
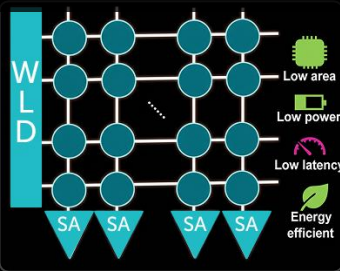
HARDWARE



Analog in-memory
computing (AIMC)

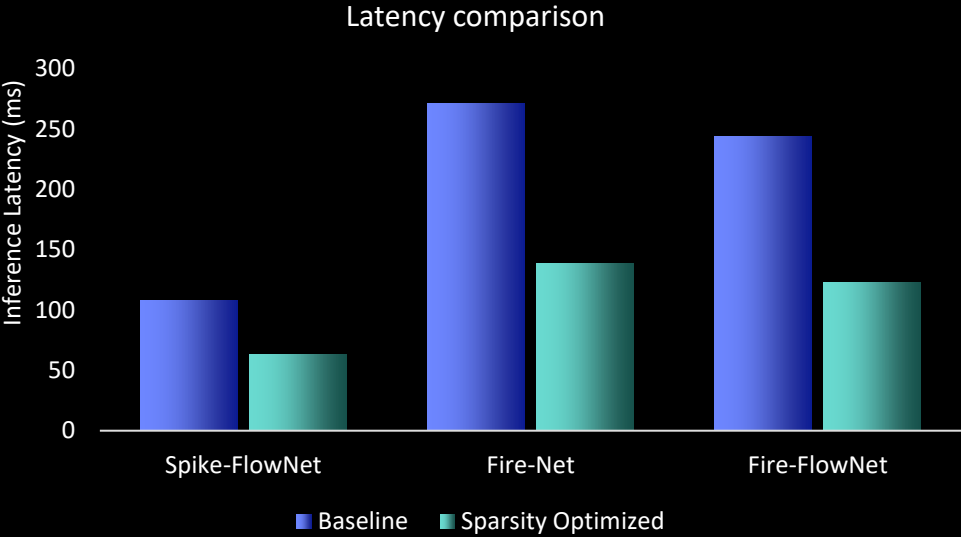
Digital in-memory
computing (DIMC)

Specialized hardware
for low latency and
low power edge
implementation



Hardware architecture: design and implementation

GPU-based baseline



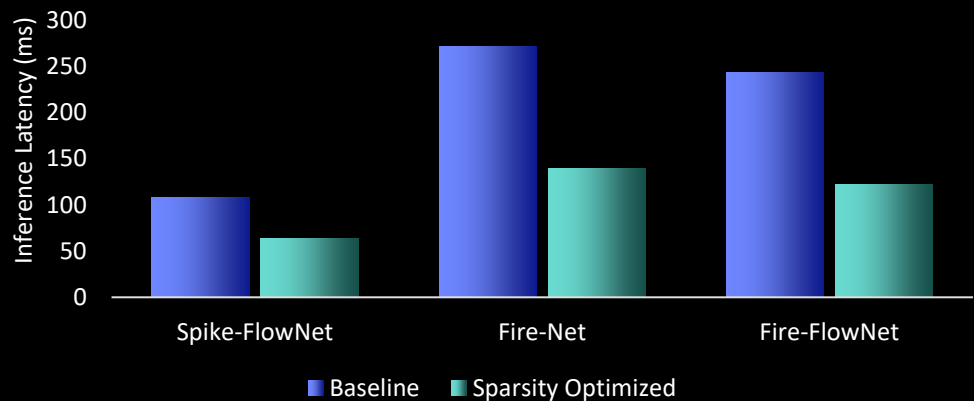
Hardware architecture: design and implementation

GPU-based baseline

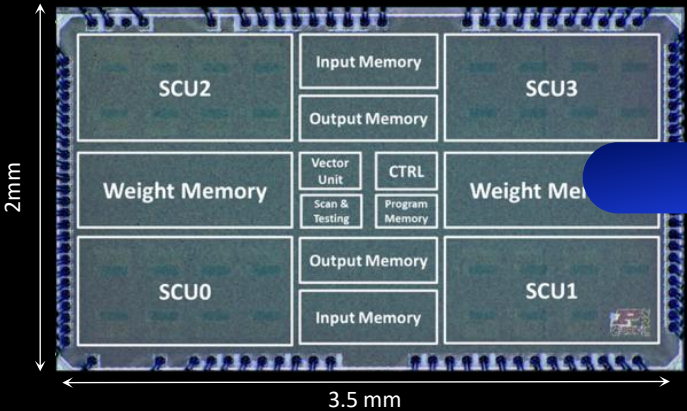
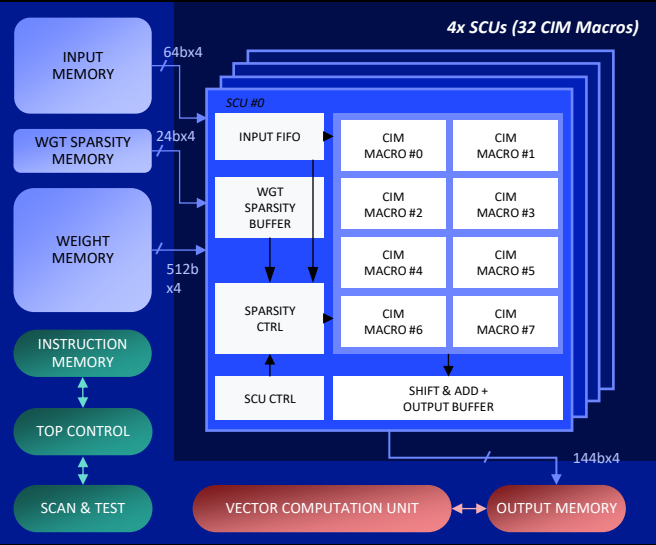


BASELINE
High latency due to inefficient handling of SNNs on GPUs.

Latency comparison



Adaptive-SNR sparsity-aware CiM



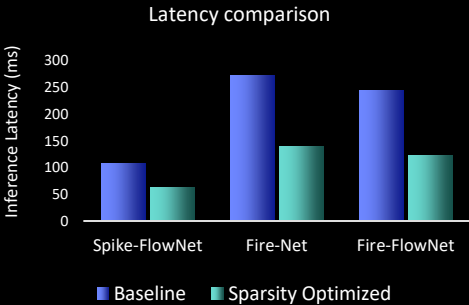
Chip layout

Hardware architecture: design and implementation

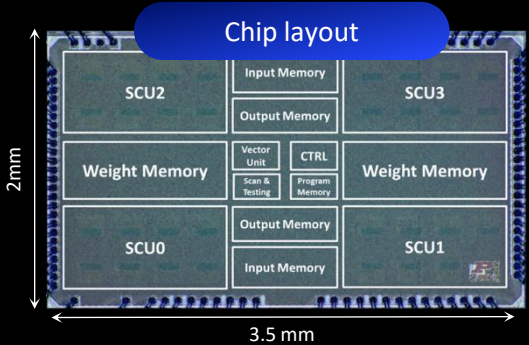
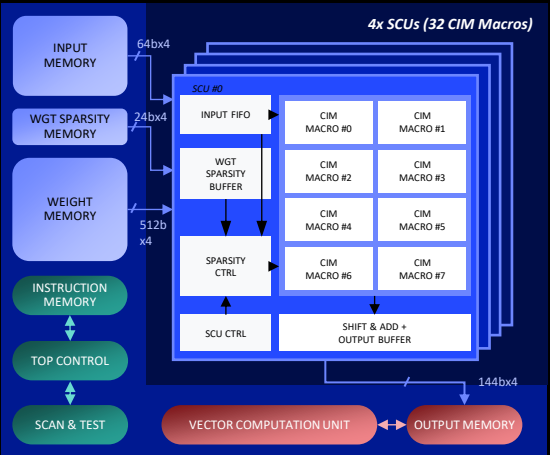
GPU-based baseline



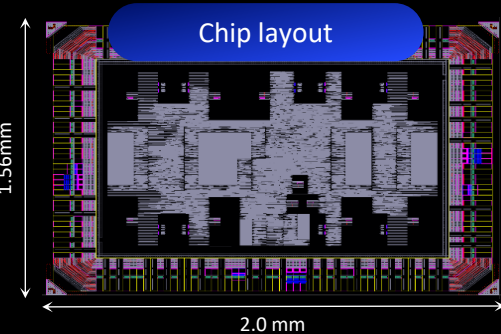
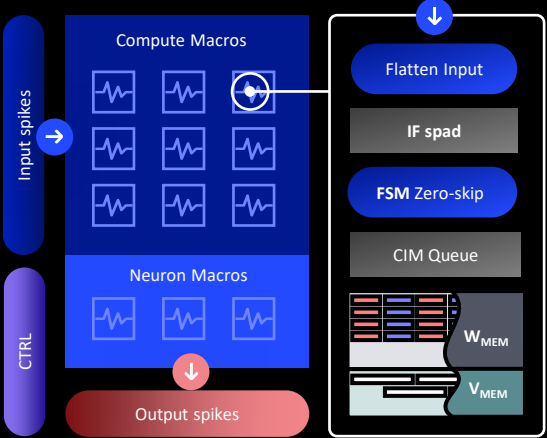
BASELINE
High latency due to inefficient handling of SNNs on GPUs.



Adaptive-SNR sparsity-aware CiM



Spiking neural network (SNN) accelerator

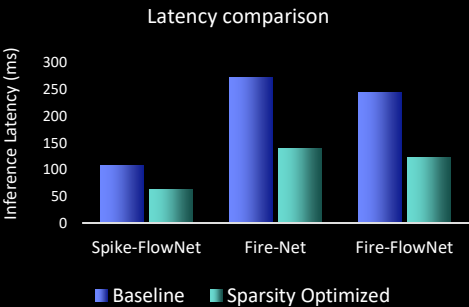


Hardware architecture: design and implementation

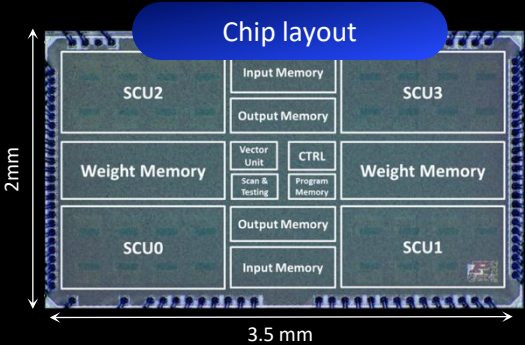
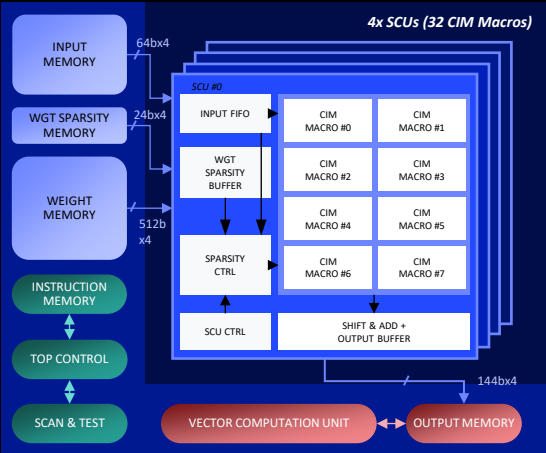
GPU-based baseline



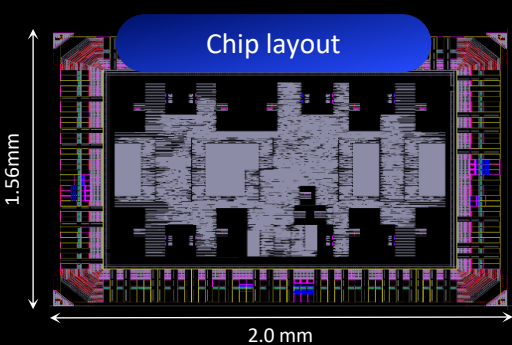
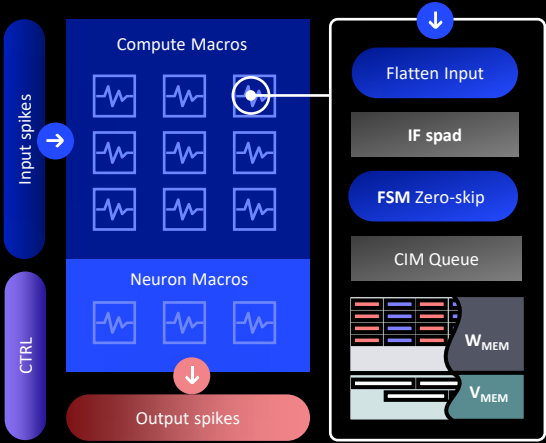
BASELINE
High latency due to inefficient handling of SNNs on GPUs.



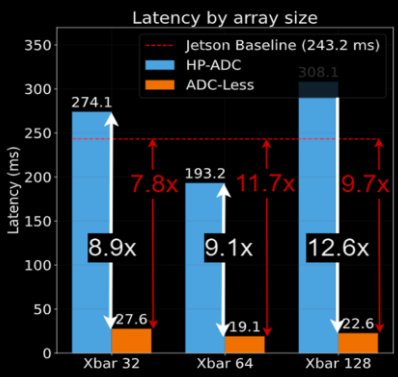
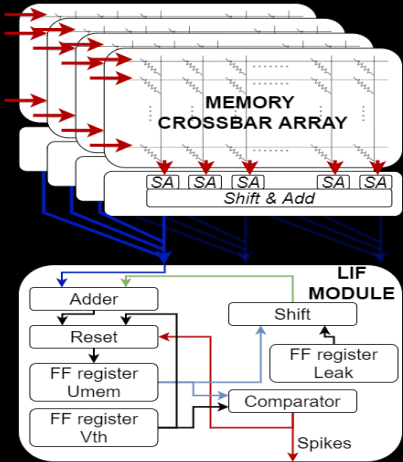
Adaptive-SNR sparsity-aware CiM



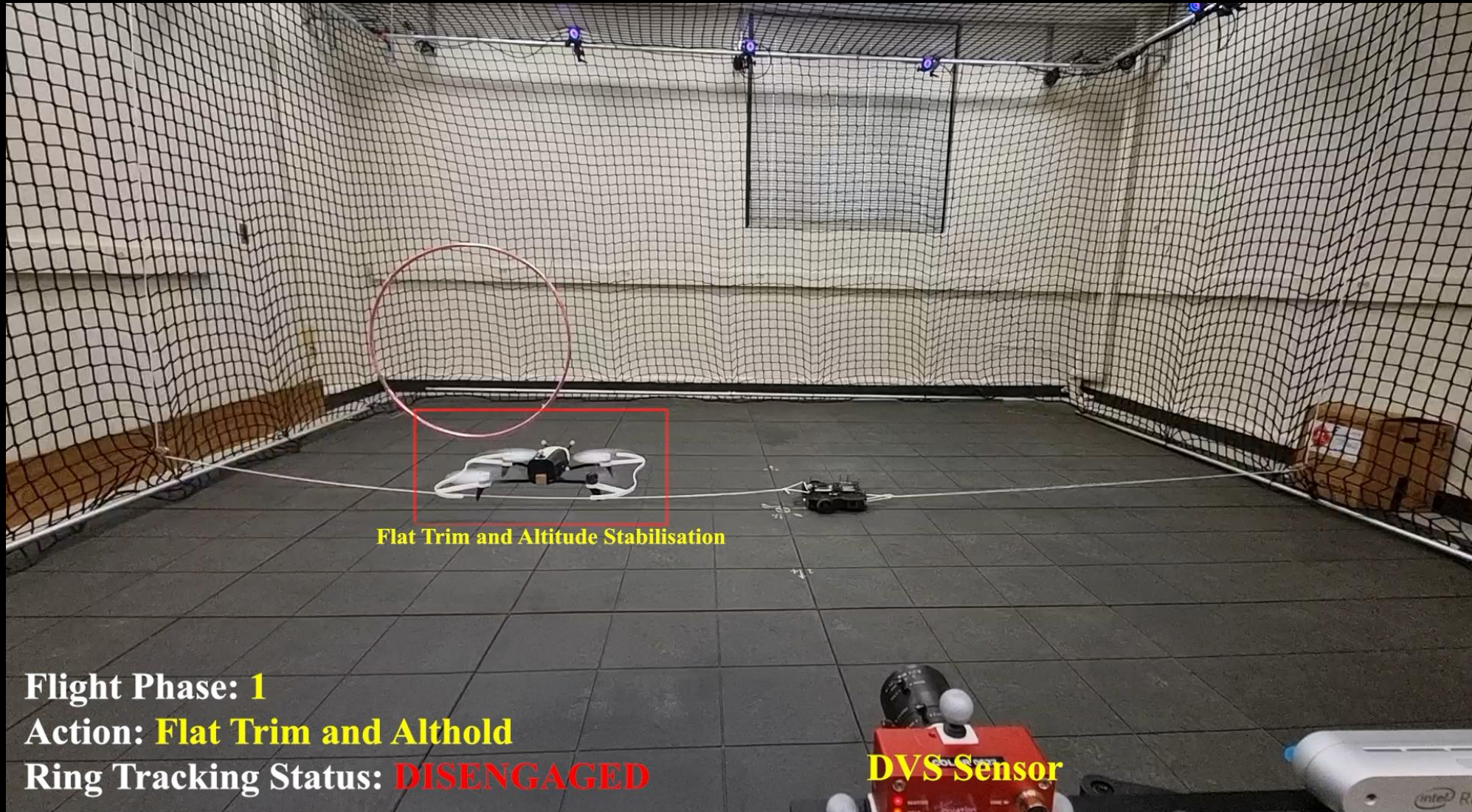
Spiking neural network (SNN) accelerator



HW/SW co-design with ADC-Less IMC



Vision based Navigation





Acknowledgement

Director: C-BRIC JUMP Center
Institute of Chips and AI

CocoSys: JUMP2.0 Center



Vannevar Bush
Faculty Fellow

