

SONY



Sense the Wonder

Privacy-preserving fall detection at the edge using Sony IMX636 event-based vision sensor and Intel Loihi 2 neuromorphic processor

**Lyes Khacef, Philipp Weidel, Susumu Hogyoku, Harry Liu, Claire Alexandra Bräuer, Shunsuke Koshino,
Takeshi Oyakawa, Vincent Parret, Yoshitaka Miyatani, Mike Davies, Mathis Richter**

Sony® Advanced Visual Sensing, Sony Semiconductor Solutions Corporation

Intel® Labs, Intel Corporation



SONY



Why fall detection using neuromorphic technologies?

Sony Advanced Visual Sensing AG

Copyright 2026 Sony Semiconductor Solutions Corporation

Why fall detection?

The world population is rapidly growing with an increase in life expectancy, particularly in developed countries.

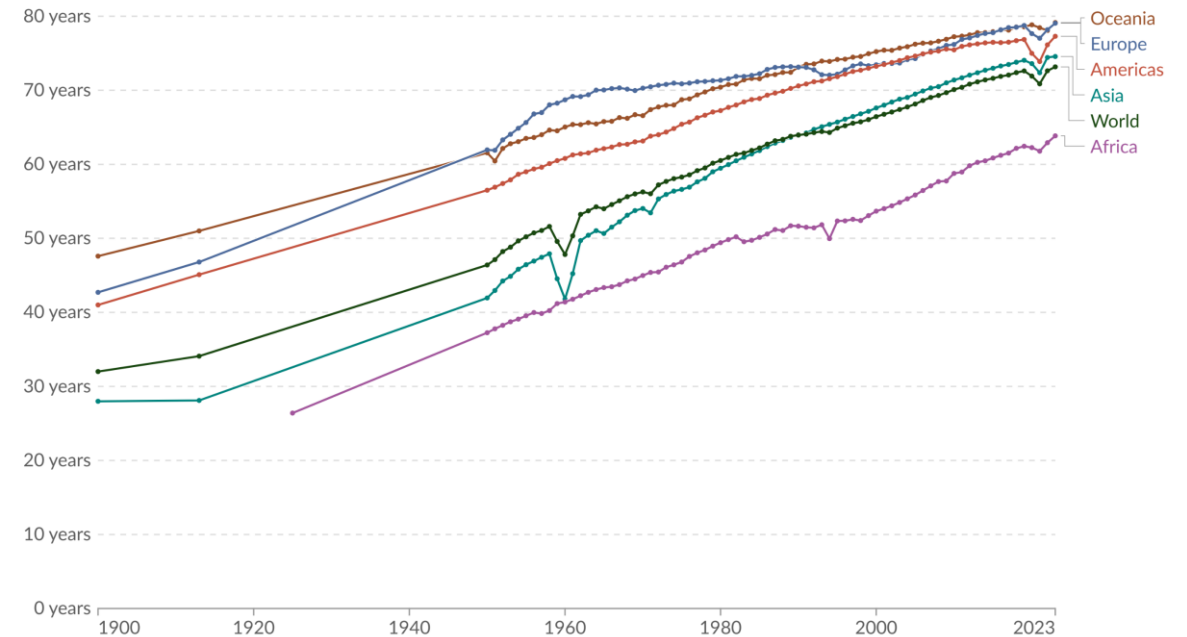
→ Elderly care is a global concern which is gaining attention:

- Falls are becoming a major public health problem for the elderly.
- Falls are the second most common cause of unintentional injury-related deaths worldwide, after road traffic accidents [World Health Organization].

→ **Growing need for fall detection systems in the healthcare industry (e.g., hospitals and nursing-care facilities).**

Life expectancy

Period life expectancy¹ is the number of years the average person born in a certain year would live if they experienced the same chances of dying at each age as people did that year.



Data source: Riley (2005); Zijdemann et al. (2015); HMD (2025); UN WPP (2024)

OurWorldinData.org/life-expectancy | CC BY

1. **Period life expectancy** Period life expectancy is a metric that summarizes death rates across all age groups in one particular year. For a given year, it represents the average lifespan for a hypothetical group of people, if they experienced the same age-specific death rates throughout their whole lives as the age-specific death rates seen in that particular year.

Learn more in our articles:

- [Life expectancy – what does this actually mean?](#)
- [Period versus cohort measures: what's the difference?](#)

Why fall detection *using neuromorphic technologies?*

Fall detection systems use two main types of sensors: **body-worn inertial sensors** and **vision sensors**.

Body-worn inertial sensors

(smartphone, smartwatches, etc.)

✓ Measure body motion independently of environmental factors.

✗ Continuous wear and regular recharging limit their practicality, particularly for elderly people.

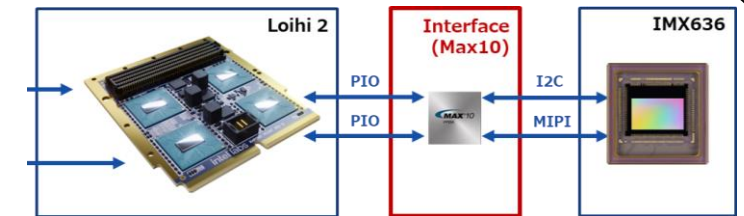
Vision sensors

(RGB, infra-red, depth cameras)

✓ Non-intrusive for the user and provide continuous passive monitoring without depending on user compliance.

✓ Combined with DL models, vision-based solutions can be highly accurate and adaptable to different environments.

✗ Lack strategies to effectively ensure data privacy, which prevents real-life deployment.

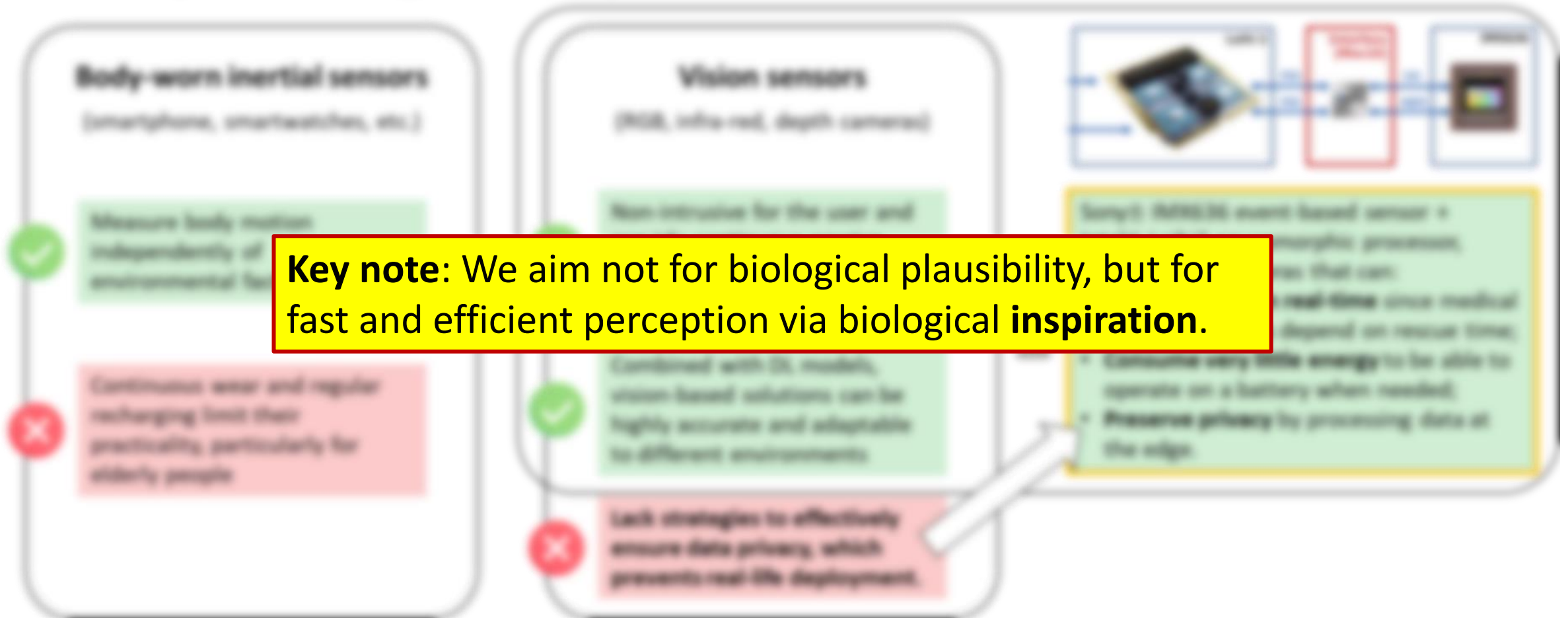


Sony® IMX636 event-based sensor + Intel® Loihi2 neuromorphic processor, toward smart cameras that can:

- **Recognize falls in real-time** since medical outcomes of falls depend on rescue time;
- **Consume very little energy** to be able to operate on a battery when needed;
- **Preserve privacy** by processing data at the edge.

Why fall detection using neuromorphic technologies?

Fall detection systems use two main types of sensors: **body-worn inertial sensors** and **vision sensors**.



SONY



Sensing and processing technologies

Sony IMX636 → interface → Intel Loihi 2

Sony Advanced Visual Sensing AG

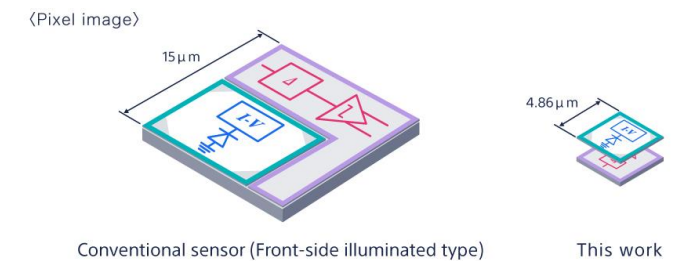
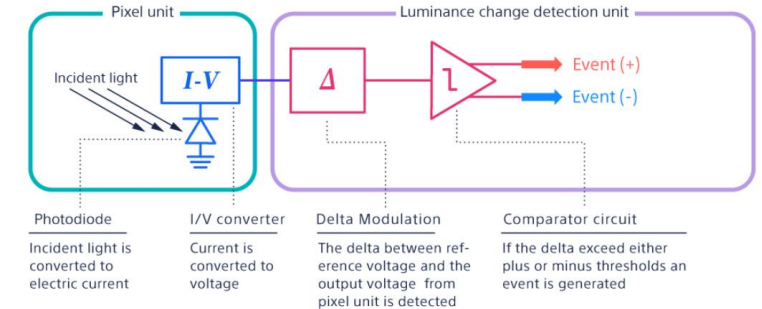
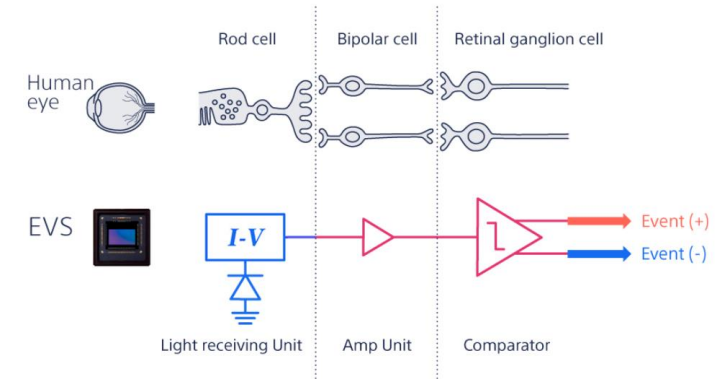
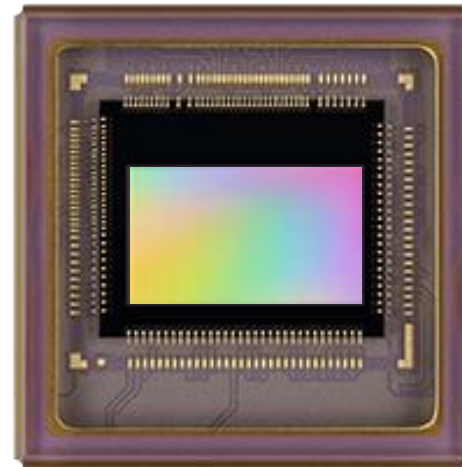
Copyright 2026 Sony Semiconductor Solutions Corporation

Sony IMX636 event-based vision sensor

- Co-developed by Sony Semiconductor Solutions® and Prophesee®.

- IMX636 features:

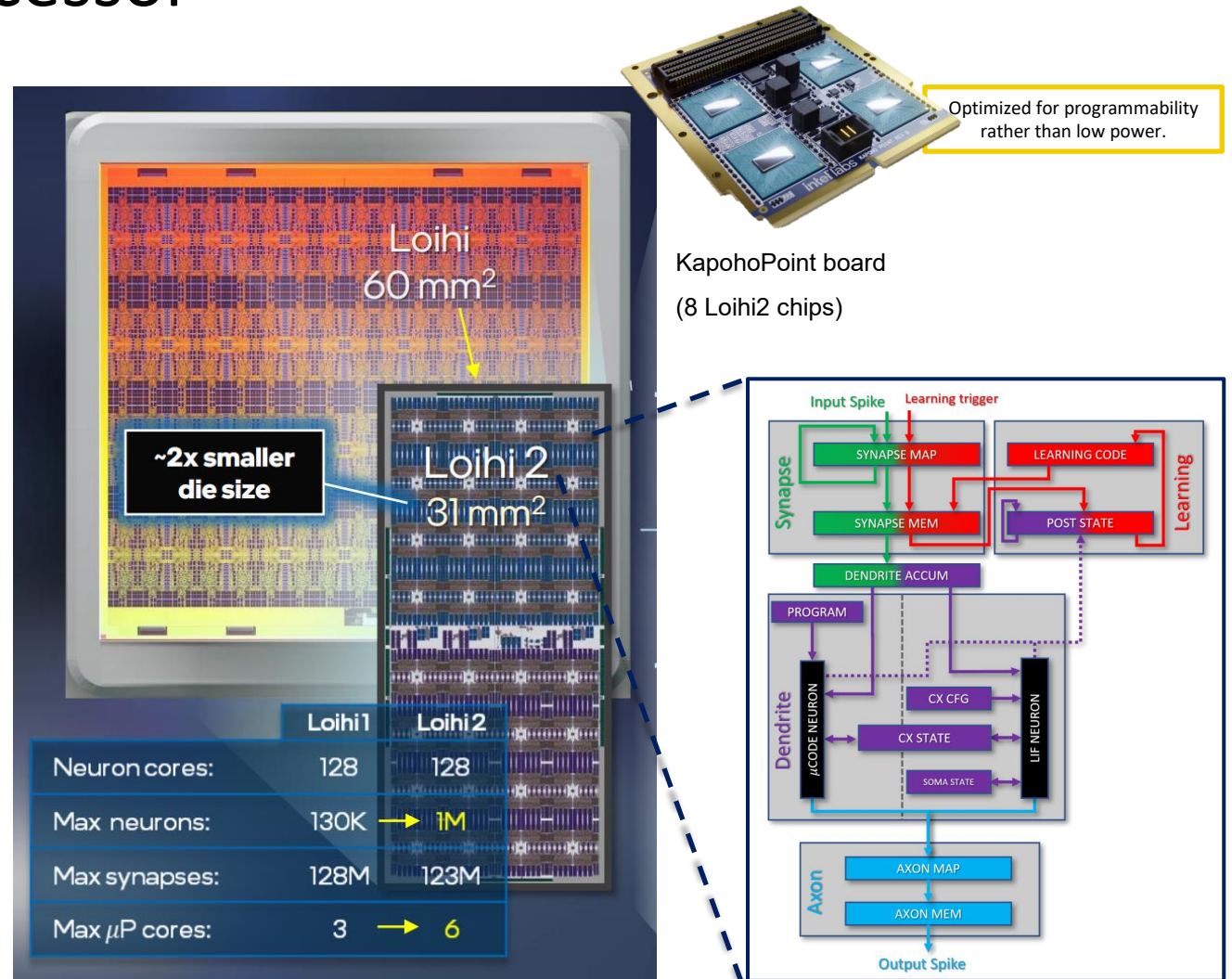
- 1280x720 Pixel CMOS vision sensor
- 4.86µm x 4.86µm event-based pixel
- Monochrome
- **Microsecond temporal resolution**
- **High Dynamic Range (HDR) beyond 120 dB**
- -40°C to +85°C Operational temperature range
- Latency of < 100 us under 1KLux
- Power consumption:
 - **50-70mW typical power consumption**
 - <100mW peak power consumption
- **Sparse output events stream**
- Random Programmable Region of Interest (ROI)
- Event signal processing functions:
 - Anti-Flicker
 - Event Filter
 - Event Rate Control
- 1G event/s (1GEPS) peak output
- I/O interface:
 - MIPI (1 Lane / 2 Lane switching) output
 - SLVS (2 ch / 4 ch switching) output



- <https://www.sony-semicon.com/en/products/is/industry/evs.html>
- <https://docs.Prophesee.ai/stable/hw/sensors/imx636.html>

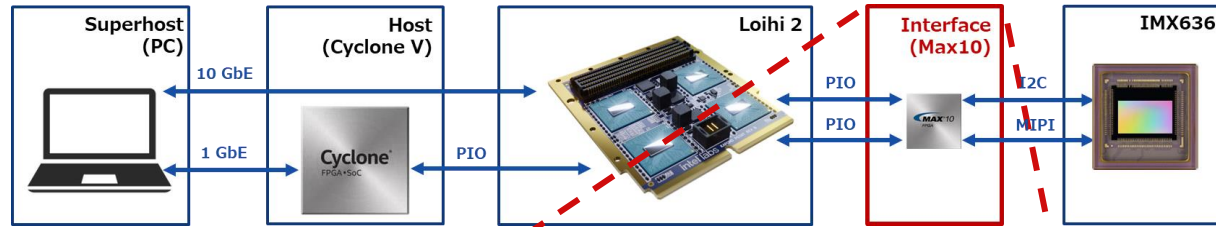
Intel Loihi 2 neuromorphic processor

- Loihi 2 is the second generation of the Intel® Loihi research chip, a fully digital asynchronous chip manufactured using Intel 4 process.
- Loihi 2 includes:
 - 128 near-memory asynchronous processing neuro-cores.
 - 2 networks-on-chip for sparse communication.
 - 6 embedded, synchronous x86 Lakemont cores for management (e.g., network config, data encoding/decoding).
- Loihi uses **asynchronous processing + barrier synchronization** mechanism (i.e., handshaking) to ensure that all neurons compute within the same global algorithmic timestep (**deterministic**).
- Loihi 2 supports sparse (beyond conventional spiking) neural network deployment due to its:
 - **programmable neuro-core** and
 - **graded spike** message-passing.
- Loihi supports **local synaptic plasticity** for onchip/online learning (not used in this work).



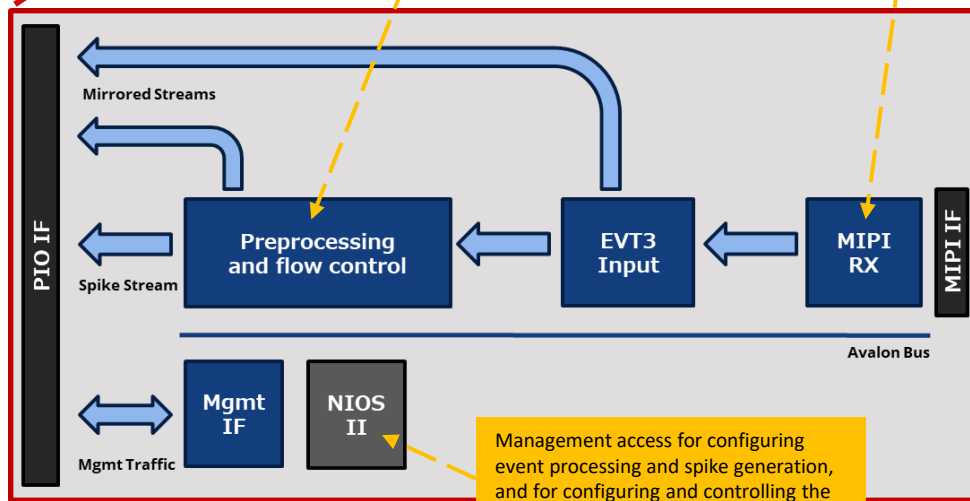
• <https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html>

FPGA-based direct interface



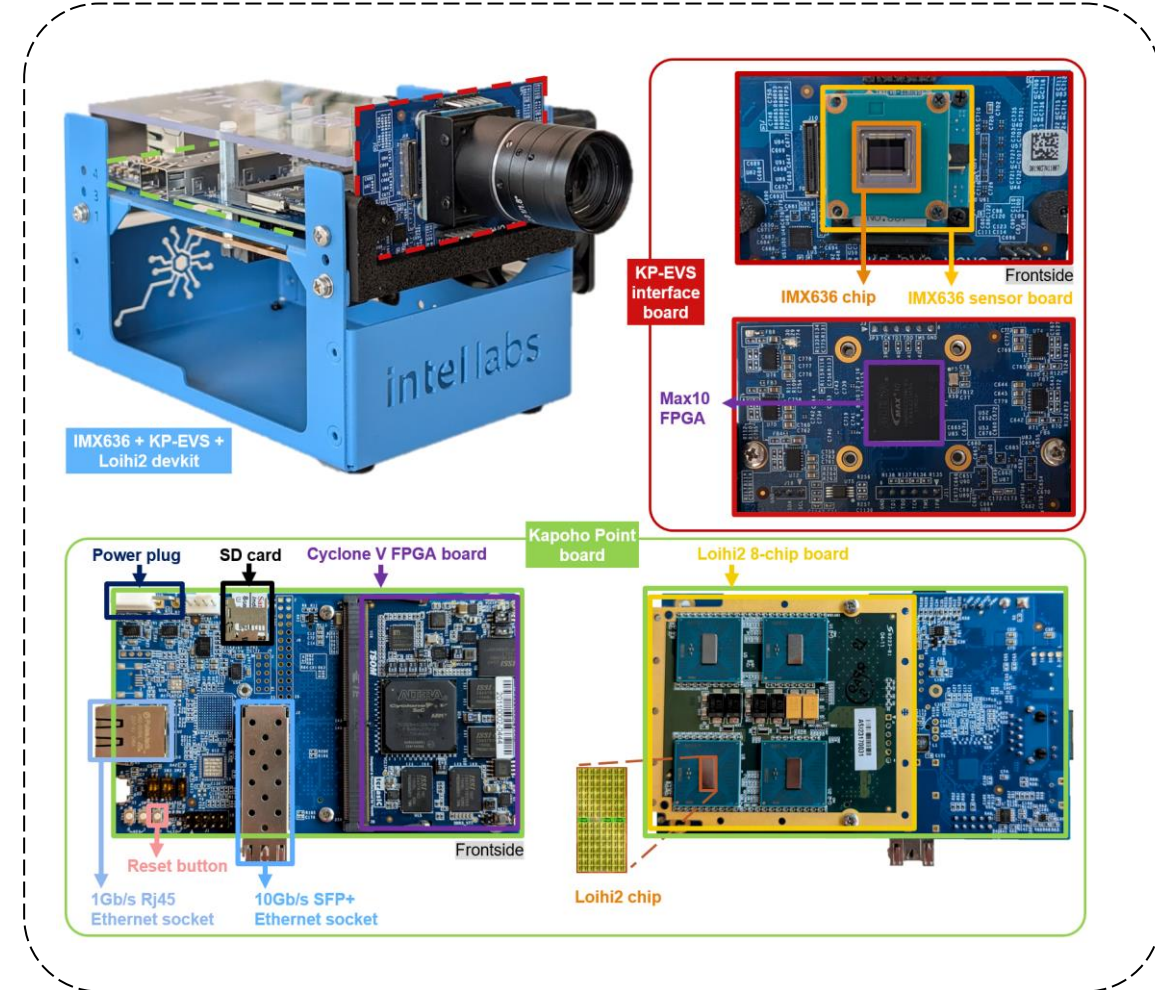
- Spatial down-sampling (640x640 -> 160x160)
- Mapping each event to an algorithmic timestep of e.g., 20ms duration, and up to 4 neuro cores to spike to.
- Synchronizes with Loihi 2 and forwards the spikes via PIO at the required timestep.

MIPI CSI-2 (2 lanes at 600 Mbps)



Management access for configuring event processing and spike generation, and for configuring and controlling the IMX636 sensor through I2C.

Hardware system pipeline with details of Max10 FPGA interface.



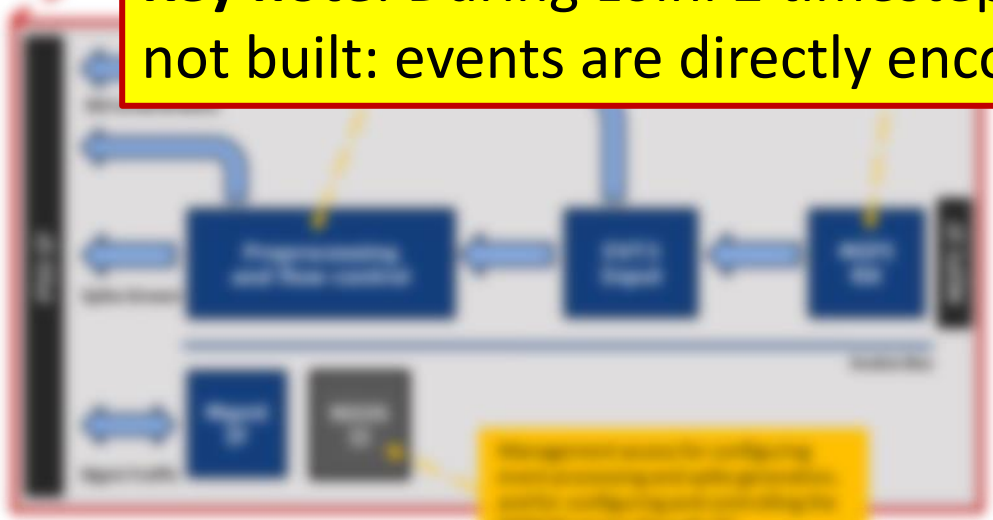
Hardware system overview showing the whole system (top left), the KP-EVS interface board (top right), the host board with IO and power interfaces (bottom left), and a Kapoho Point (KP) board with 8 Loihi 2 chips (only one chip is used in our work) (bottom right)

FPGA-based direct interface

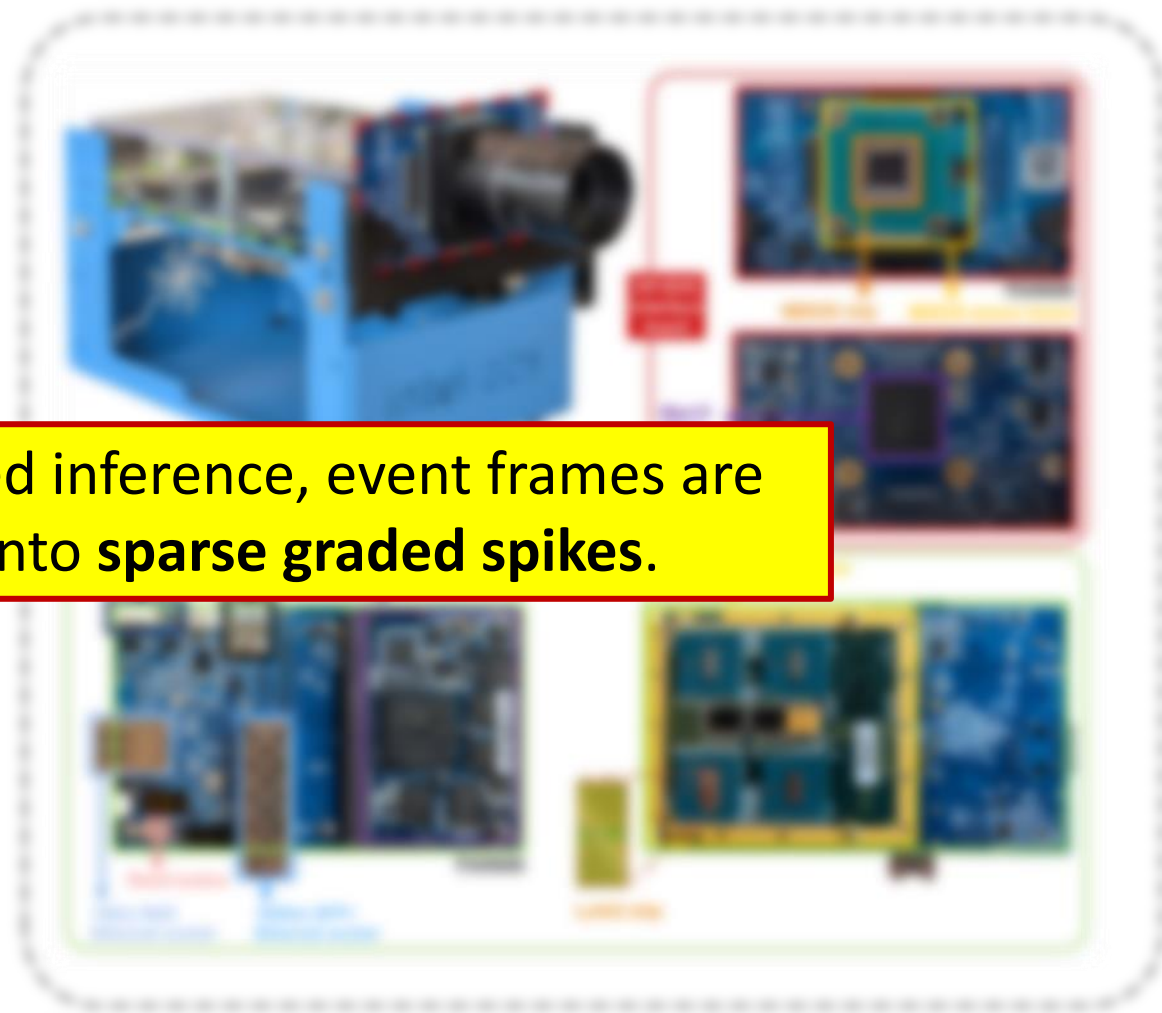


During inference, the event frames are not built: events are directly encoded into sparse graded spikes.

Key note: During Loihi 2 timestep-based inference, event frames are not built: events are directly encoded into **sparse graded spikes**.



Hardware system setup with details of the direct interface



Hardware system overview showing the whole system (top left), the direct interface board (top right), the front board with GPU and sensor interface (bottom left), and a separate front GPU board with a GPU chip (bottom right) (chip is used to run work) (bottom right)

SONY



Sense the Wonder

Hardware-aware algorithmic exploration and benchmarking

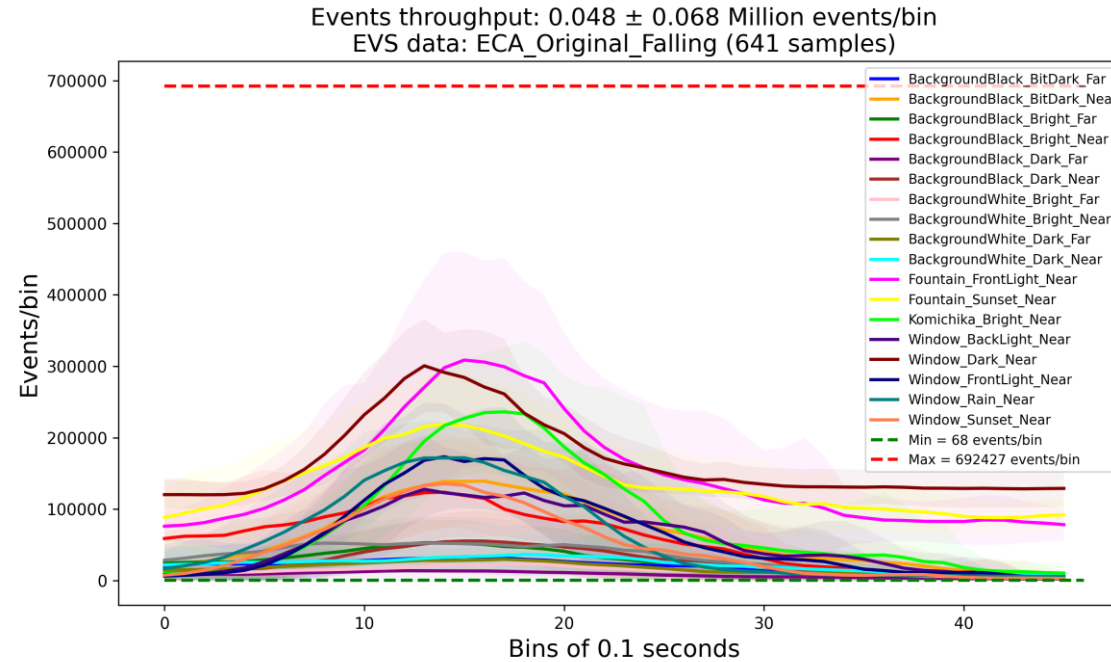
Dataset, NN architectures, neural models, training and inference

Sony Advanced Visual Sensing AG

Copyright 2026 Sony Semiconductor Solutions Corporation

Fall detection internal dataset

- Filming was done in 5 locations with different
 - backgrounds,
 - light conditions (source/brightness),
 - distance.
- 14 classes, about 600 videos per class (total of 8888 videos).
- Data are cropped to 640x640.
- Max throughput of the ECA_Original cropped (640 x 640) data:
 - ~9 Million events/s (in class "Walking")
 - It is a reduction of ~20% compared to the full resolution (1280 x 720) data.
- The average peak throughput is ~4 Millions events/s (in class "Walking").

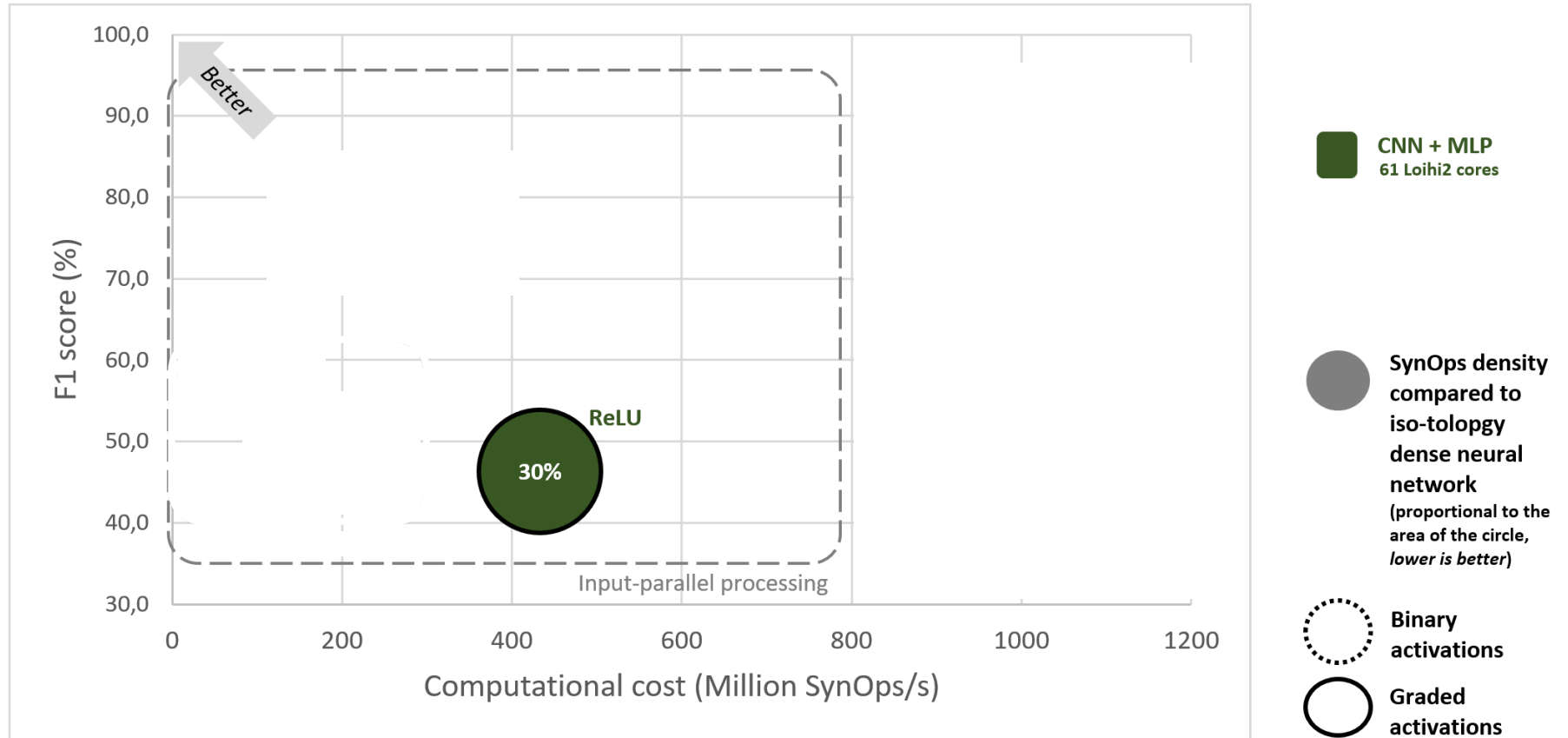


Class	Files
Seated	658
Standing	638
Sit Down	639
Stand Up	629
Lying	643
Pick Up	619
Walking	639
No person	635
Touching head	634
Touching back	633
Touching torso	633
Touching neck	621
Coughing	626
Falling	641
Total	8888

- Examples:



Algorithmic-level evaluation



- **Baseline model is a ReLU-based CNN model** → F1 score of ~46%, due to the stateless processing where only individual event frames are perceived.
- The sparsity of ReLU-based models is exploited by the asynchronous processing of Loihi2 (i.e., proportional dynamic energy consumption).

Spiking neurons: **Stateful** neuron models with **sparse** outputs

For all neuron models, let us consider the activation z at the algorithmic time step t as follows:

$$z[t] = \sum_{i=1}^n w_i \times x_i[t] + b$$

Sigma-Delta (SD) neurons consist of two main units:

- **Delta encoder** in the axon (output): If the activation difference (e.g., using ReLU) with the previous timestep is higher than a threshold, it emits the that difference in the form a of a graded spike.
- **Sigma decoder** in the dendrite (input): Accumulates sparse inputs to restore the original information.

$$a[t] = \begin{cases} z[t] & \text{if } z[t] \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\Delta a[t] = a[t] - a[t-1] + r[t-1]$$

$$y[t] = \begin{cases} \Delta a[t] & \text{if } \Delta a[t] \geq \vartheta \\ 0 & \text{otherwise} \end{cases}$$

$$r[t] = \Delta a[t] - y[t]$$

Leaky-Integrate and Fire (LIF) neurons have two main states:

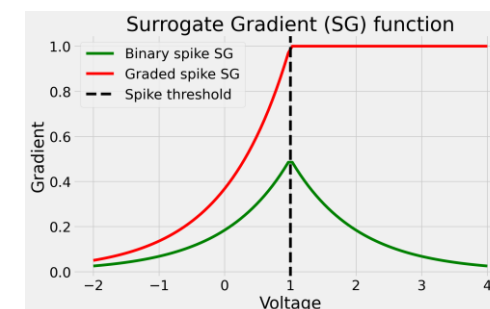
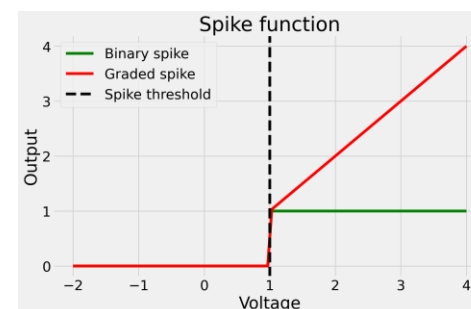
- **Current state**: Accumulates the weighted input spikes with a decay.
- **Voltage state**: Accumulates the current state with a decay.

When the voltage exceeds a threshold, the neuron emits a binary spike (i.e., 1) or a graded spike (voltage value) and resets the voltage to zero.

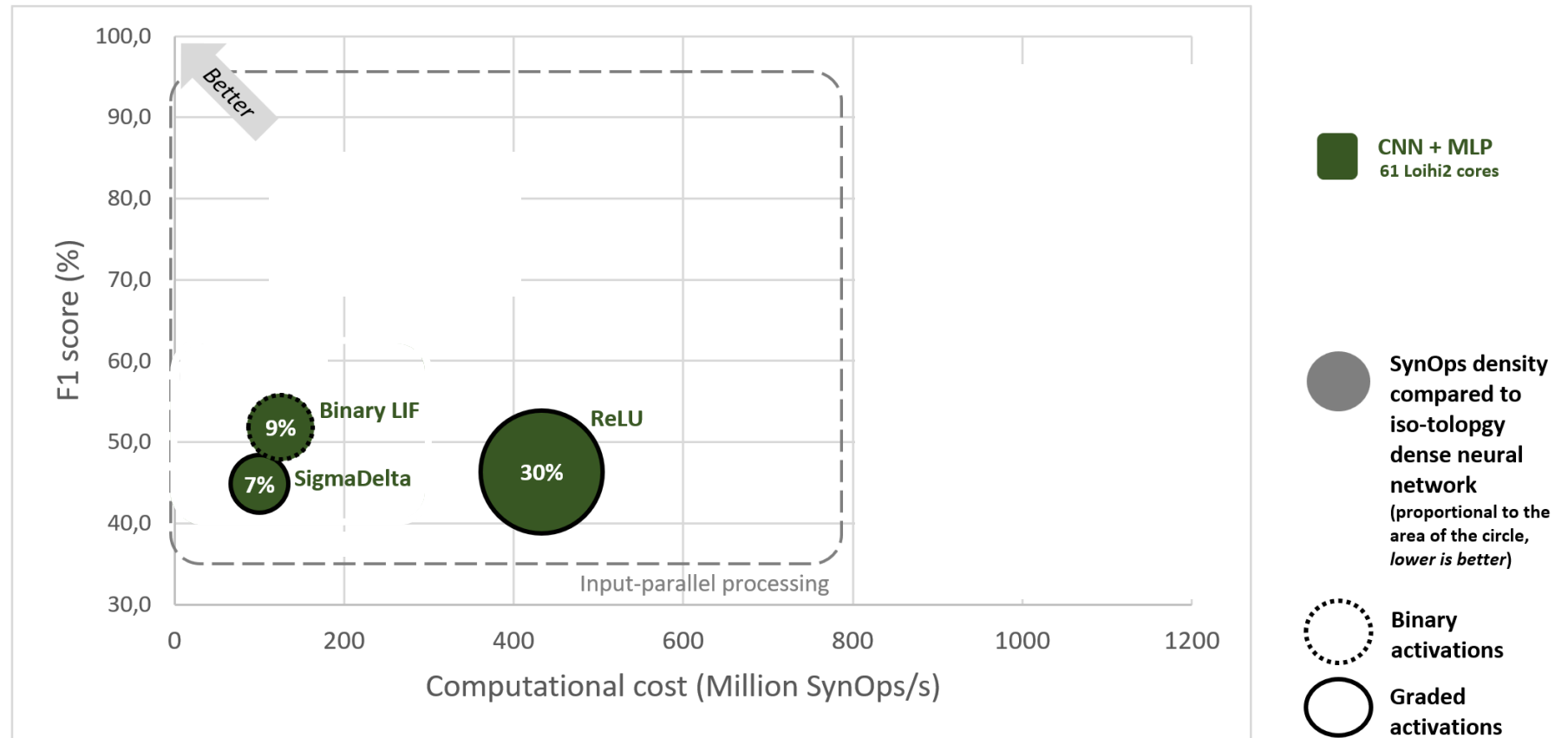
$$i[t] = \alpha \times i[t-1] + z[t]$$

$$u[t] = \beta \times u[t-1] \times (1 - \mathcal{H}(u[t-1] - \vartheta)) + i[t]$$

$$y[t] = \begin{cases} u[t] & \text{if } u[t] \geq \vartheta \text{ and graded spikes} \\ 1 & \text{if } u[t] \geq \vartheta \text{ and binary spikes} \\ 0 & \text{otherwise} \end{cases}$$

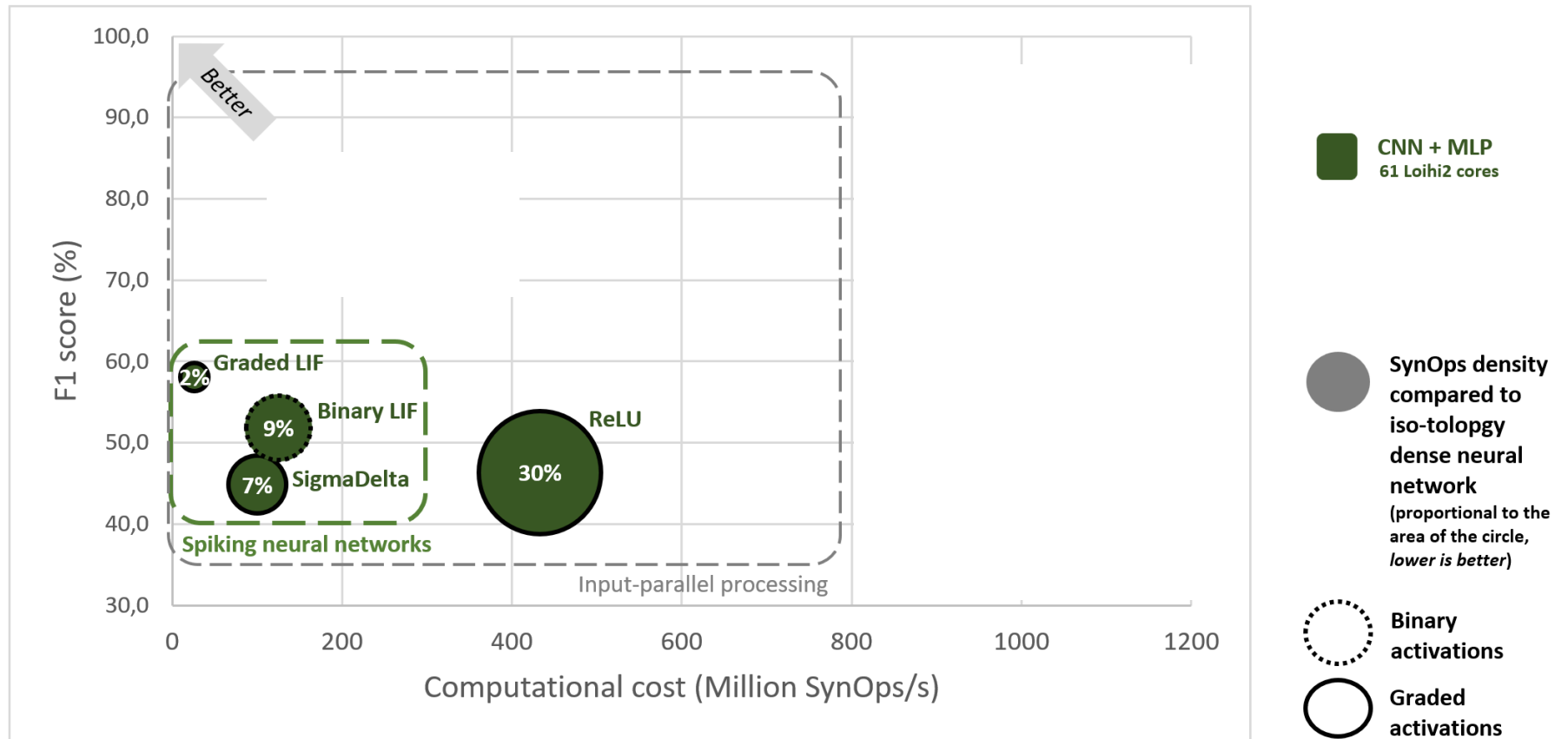


Algorithmic-level evaluation



- Replacing ReLU by SigmaDelta → F1 score slightly decreases by ~1% (states are used after the original ReLU activation to sparsify the activations only without temporal feature extraction), while SynOps/s decrease of ~4x.
- Replacing SigmaDelta by binary LIF → F1 score increases by ~7% (thanks to the temporal feature extraction), while SynOps/s increases by ~25%.

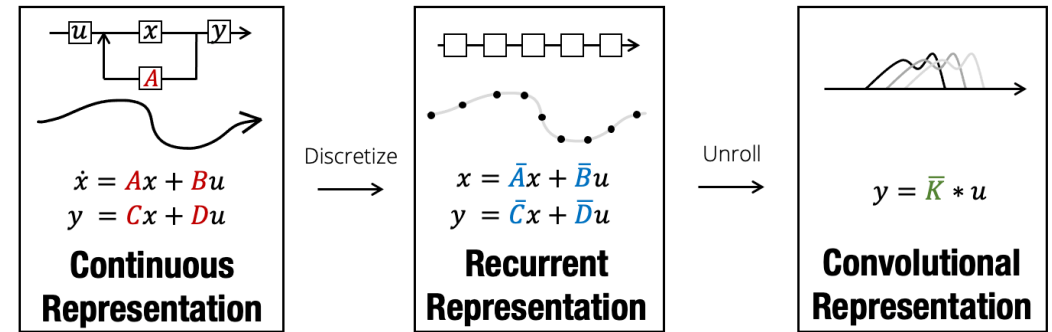
Algorithmic-level evaluation



- Replacing binary LIF by graded LIF → F1 score increases by ~6% to reach ~60%, while SynOps/s decreases by ~5x (26 M SynOps/s).
- The multiplication overhead of a graded SynOp is not significant when taking into account the local SRAM memory access cost.
- Compared to an iso-topology model with dense processing, the graded LIF-based SNN has a higher memory footprint, but it increases F1 score by ~12% and reduces SynOps/s by 55x.

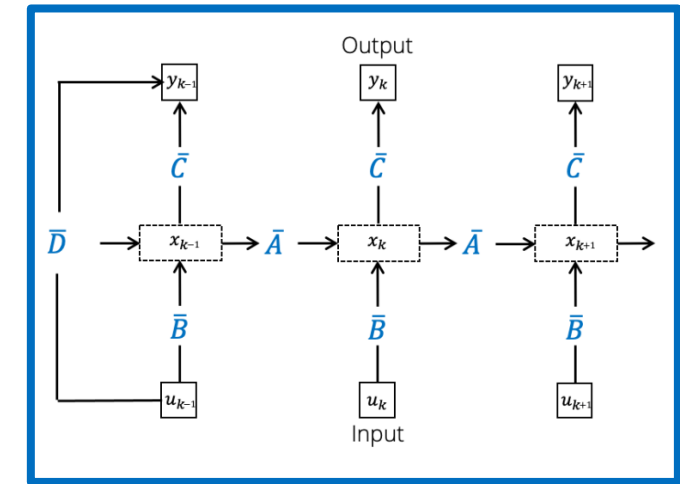
State space model: S4D

- **State Space Models** are a promising alternative to Transformers for sequence modeling:
 - **In their convolutional form:** They allow for highly parallel training on GPUs.
 - Fast training for stored data.
 - **In their recurrent form:** They do not suffer from the quadratic scaling of compute cost of the attention mechanism.
 - Linear scaling, efficient inference for streaming data.
 - SSM dynamics is similar to the sub-threshold dynamics of spiking neurons.
- The S4D model is a variant of the Structured State Space for Sequence Modeling (S4) architecture, where a diagonal state space matrix is used to simplify the kernel computation.

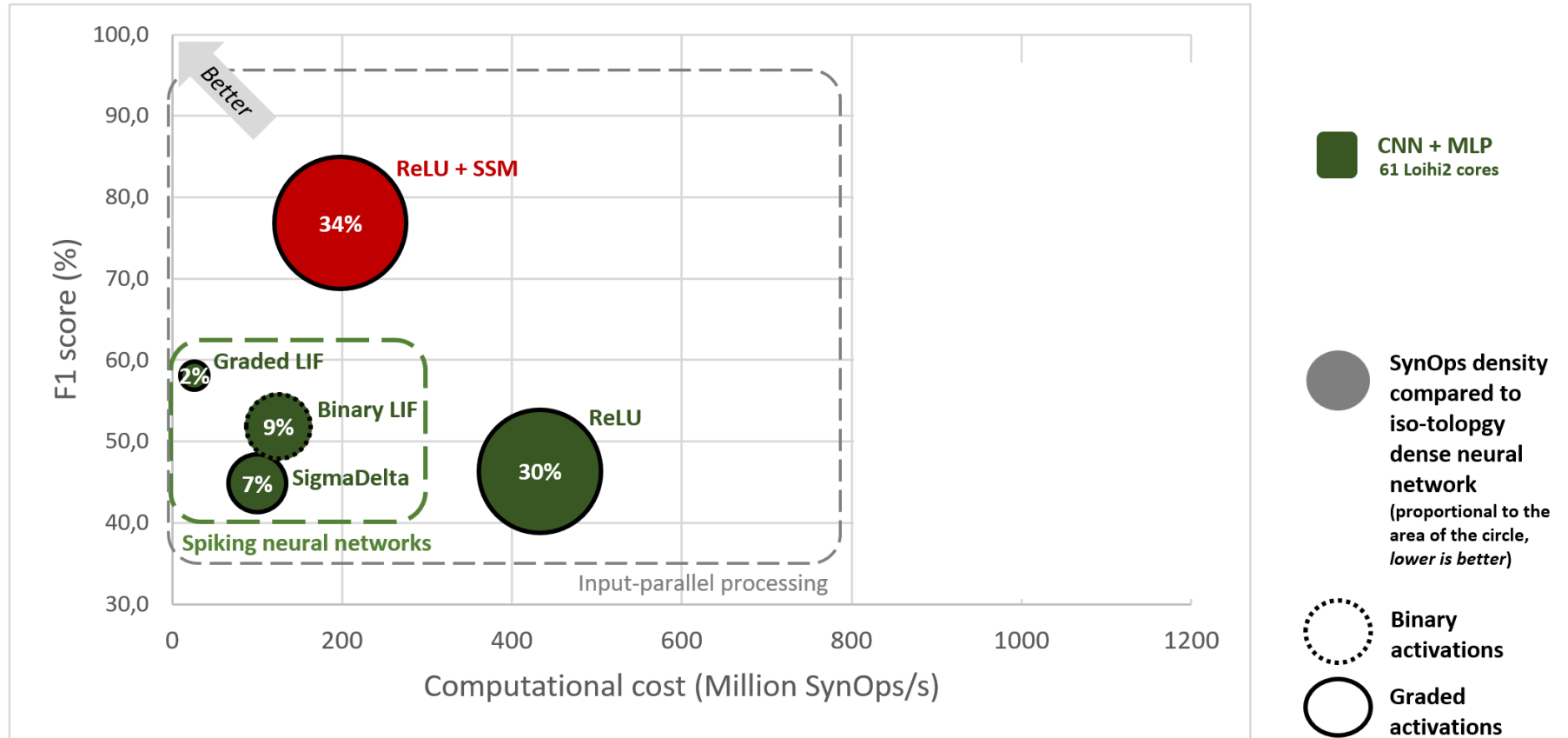


$$s[t] = a \times s[t - 1] + b \times z[t]$$

$$y[t] = c \times s[t]$$



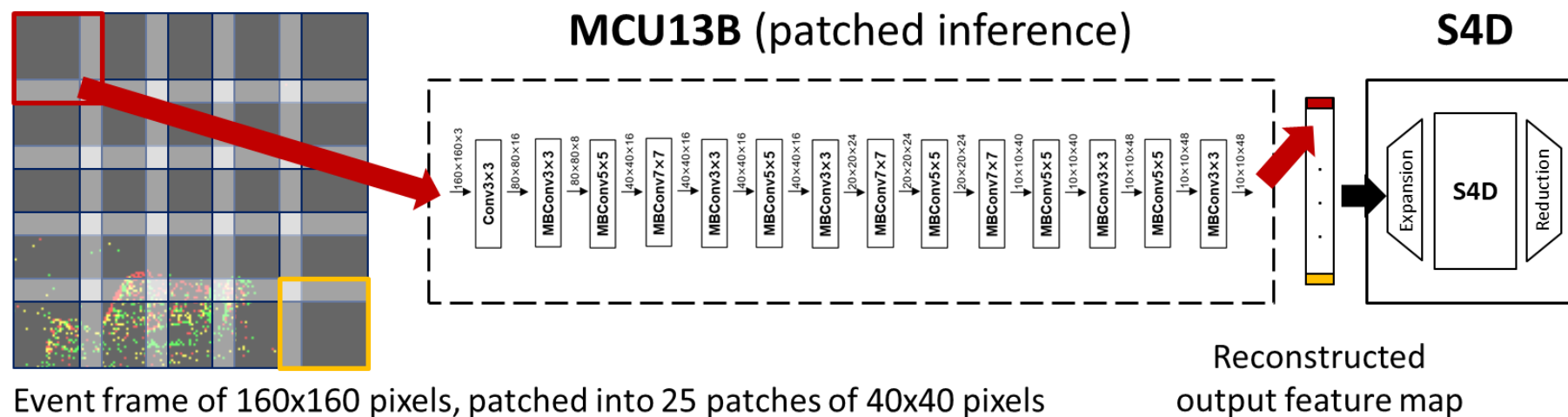
Algorithmic-level evaluation



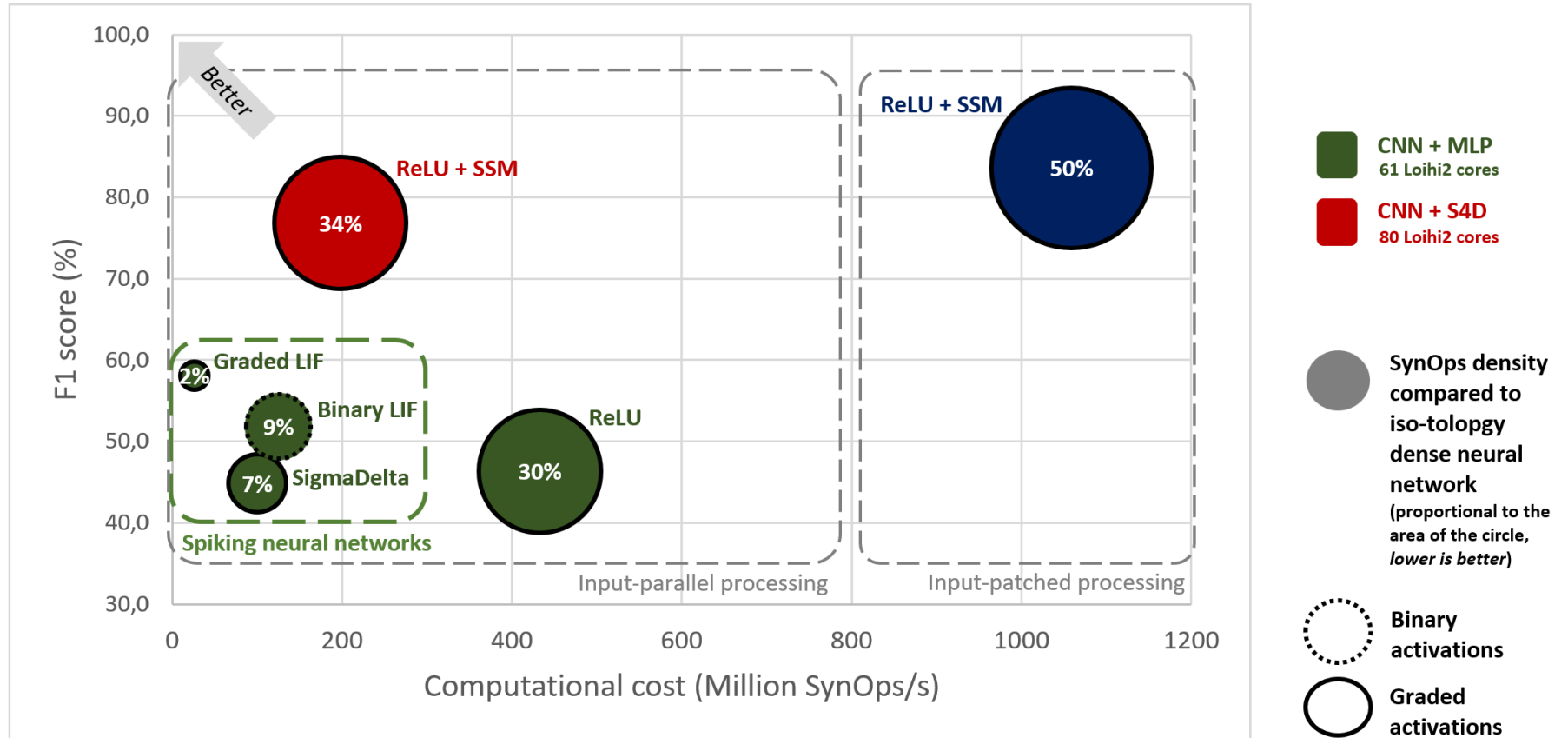
- Replacing the MLP by an S4D layer, decoupling spatial from temporal feature extraction → F1 score significantly increases by ~19% at a cost of ~200 M SynOps/s.
- The CNN+S4D model sparsity can be improved (with minimal cost in F1 score) by wrapping ReLU activations with SigmaDelta neurons.

MCUNet: Patched inference on Loihi 2

- **MCUNet-in2 architecture:** MB{expansion ratio} {kernel size}x{kernel size}
- **Key characteristics of a MobileNet Block (MB)**
 - Depthwise separable convolution
 - Linear bottlenecks and inverted residuals
 - Non-linear activation (e.g., ReLU6)
- **Limitation:**
 - Due to the large number of neurons, we need ~10 Loihi 2 chips to fit the entire network.
 - However, since the network is stateless, we do not need to keep the neurons states for consecutive time steps.
 - Therefore, **patched inference** can be used:



Algorithmic-level evaluation



- Replacing the CNN by the MCU13B → F1 score increases by ~7% thanks to the enhanced spatial feature extraction, while SynOps/s increase by ~5x.
- The patched inference degrades the F1 score by about ~2%, but allows the model to fit in a single Loihi 2 chip.

System-level evaluation on Loihi 2

F1 score:

- Each of the three models in the pareto front of F1 score vs. computational cost is deployed in a single Loihi2 chip. The F1 score on the Loihi2 chip is the same as the GPU-based simulations.

Input to output latency (mainly using fall-through processing):

- CNN-based models: 2 ms (max throughput of 500 Hz).
- MCUN13B+S4D model: 40 ms (max throughput of 25 Hz).

Power consumption:

- The static power is roughly proportional to the number of used cores, with about 0.75-0.95 mW/core depending on the used resources.
- Dynamic power is about 8-12 pW/(SynOp/s), showing that dynamic energy consumption is indeed proportional to the number of SynOps thanks to the asynchronous processing of Loihi 2.
- Nevertheless, in the current research chip, static power is dominant. It reduces the computational efficiency impact of the CNN+MLP with graded LIFs (46.3 mW for 26 M SynOps/s) compared to the MCU13B+S4D (88.9 mW for 1059 M SynOps/s).

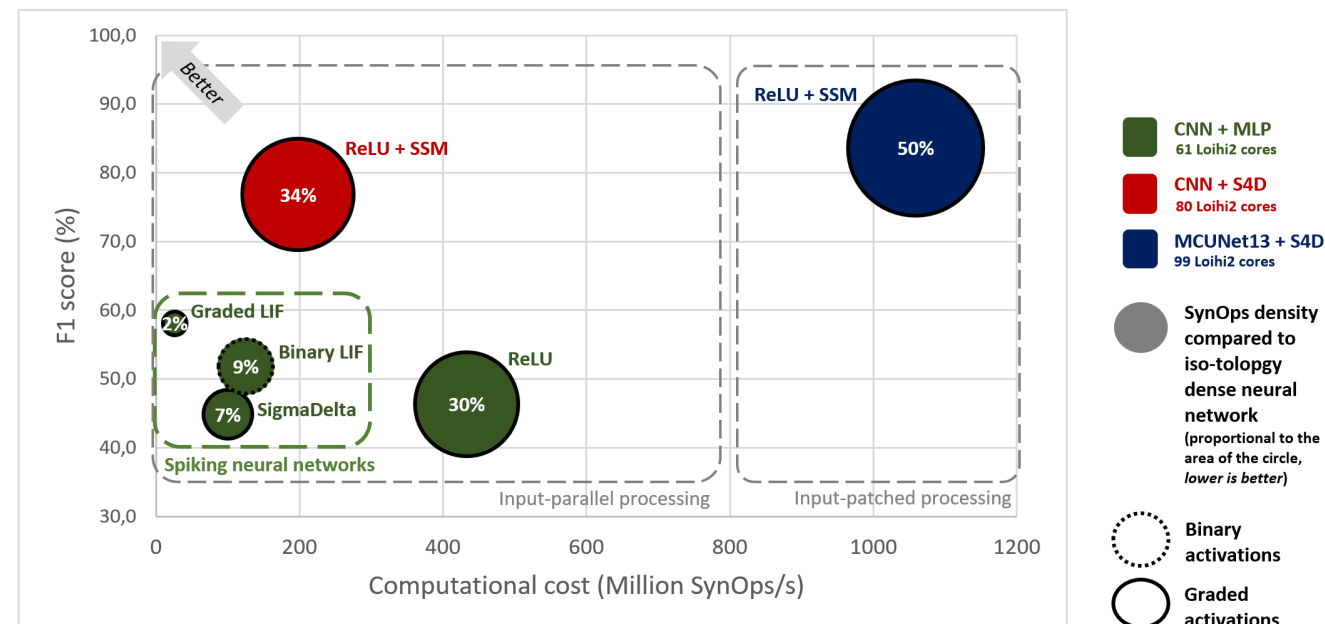


TABLE II
SYSTEM-LEVEL BENCHMARKING RESULTS OF THE PARETO-FRONT SOLUTIONS.

Pareto front model	Loihi2 cores	Power (mW)		
		Static	Dynamic	Total
CNN + MLP [Graded LIF]	61	46.0	0.3	46.3
CNN + S4D	80	76.0	1.5	77.5
MCU13B + S4D	99	80.7	8.2	88.9

SONY



Sense the Wonder

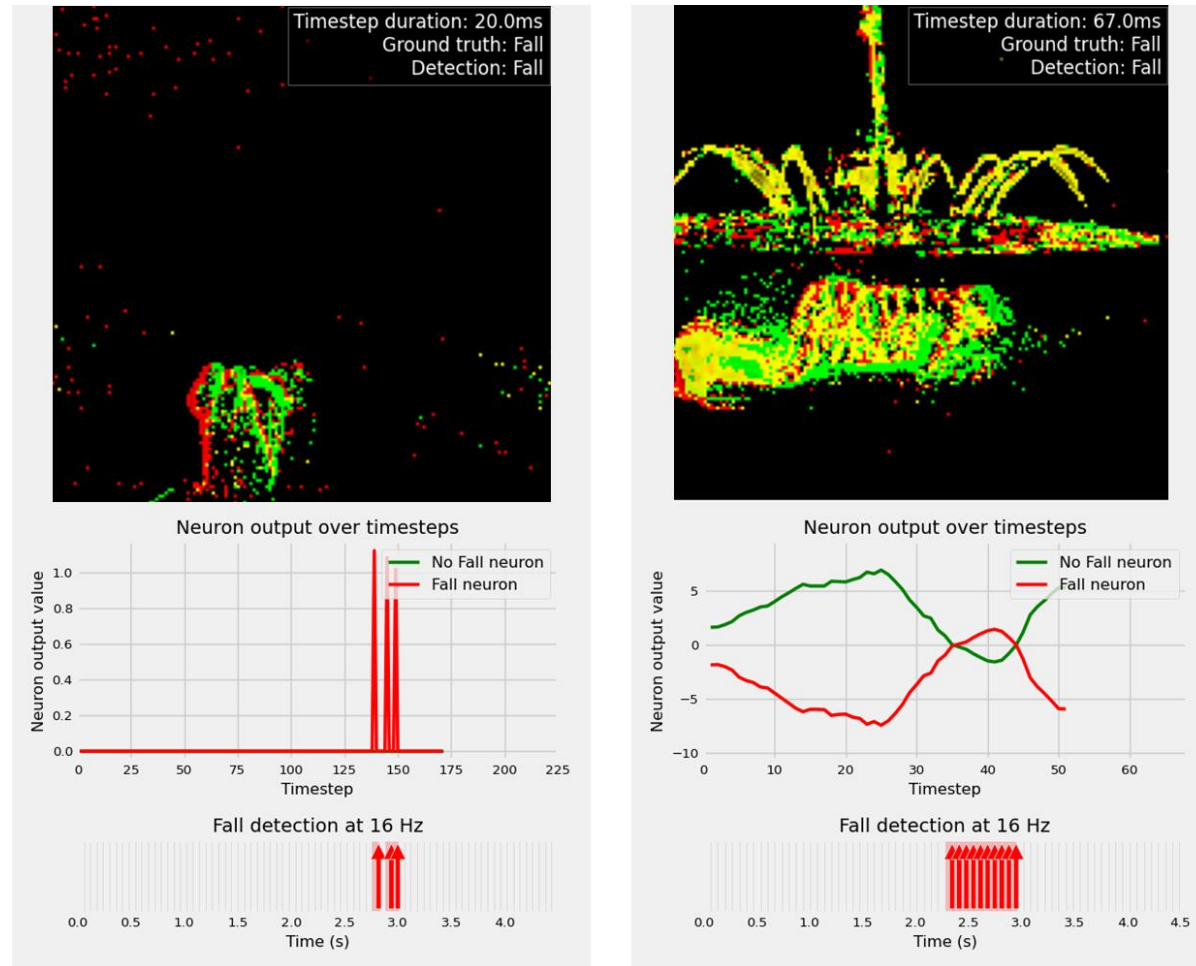
Fall detection visualization

CNN with graded LIF spiking neurons

Sony Advanced Visual Sensing AG

Copyright 2026 Sony Semiconductor Solutions Corporation

Visualization example *(videos are not included in this PDF version)*



Events at 3.4s, logits and detection of CNN+MLP with graded LIF (*left*) and MCU13B+S4D (*right*).

SONY



Conclusion and research directions

Sony Advanced Visual Sensing AG

Copyright 2026 Sony Semiconductor Solutions Corporation

Conclusion and research directions

Our approach:

Start from the application requirements and target system-level performance (sparsity-aware sensors, processors and algorithms).

Main insights:

- Dedicated (asynchronous) processing with a direct interface realizes EVS gains (low-latency and low-power) at the system level to enable real-time, always-on and privacy-preserving perception systems.
- The graded LIF-based SNN is the most compute-efficient model in the CNN+MLP architecture, reaching ~60 F1 score with:
 - ~5x less SynOps compared to binary LIF-based SNN.
 - ~50x less SynOps compared to a stateless dense ANN.→ What matters in SNNs in the sparsity of spikes.
- The MCUNet+S4D is the most accurate model, reaching ~85% F1 score with a SynOps sparsity of 2x. Patched inference enables the deployment of such SOTA architecture in a single Loihi 2 chip.

Next steps:

- Fully embedded processing is necessary for privacy-preserving fall detection, but it may not be sufficient at the application level → We need on-chip anonymization of the event data that would be transferred to the cloud in the case of a fall.
- Always-on perception at the edge serves as a filter of raw sensory data by only sending out some data when a pattern of interest is detected → It drastically reduces the amount of data that need to be transferred, processed and stored in the cloud and thus reduces the overall energy consumption of the application ecosystem.
- Neuromorphic technologies can enable this transition from cloud-centric to edge-distributed real-time processing.

SONY

SONY is a registered trademark of Sony Group Corporation.

Names of Sony products and services are the registered trademarks and/or trademarks of Sony Group Corporation or its affiliates.

Other company names and product names are registered trademarks and/or trademarks of the respective companies.