Hardware Sostenible

Hardware Neuromórfico

Escuela AIHUB,

4 Julio 2022, Palma de Mallorca

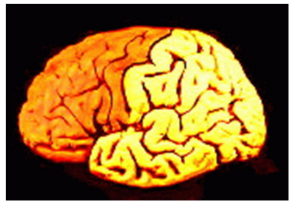
Teresa Serrano-Gotarredona

- **Neuromorphic Systems: Introduction**

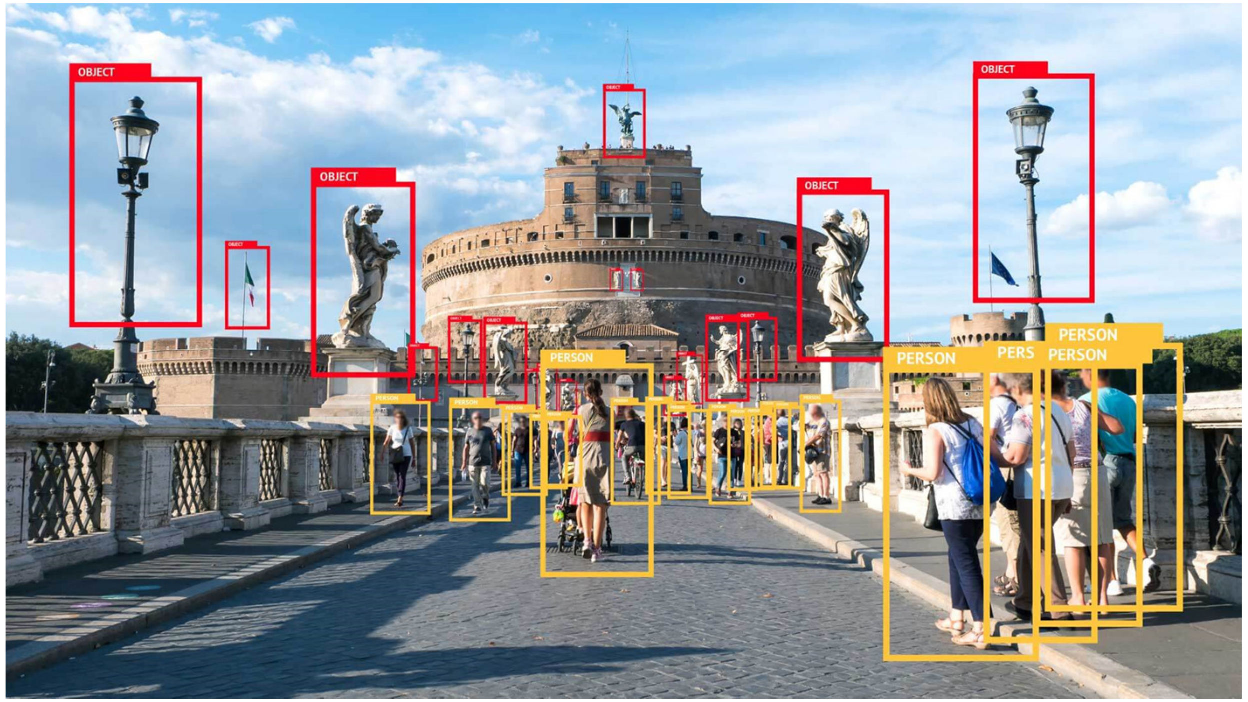
Engineering Neuromorphic Systems:

- Massive Connectivity: Address Event Representation
- Neuromorphic Vision Sensors
- Neuromorphic Processors: Spiking ConvNets
- Memristors: dense memory, on-line learning

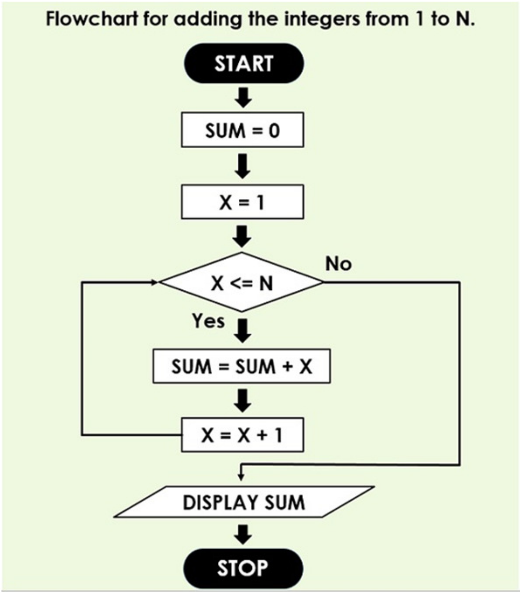
Neuromorphic Systems: Introduction



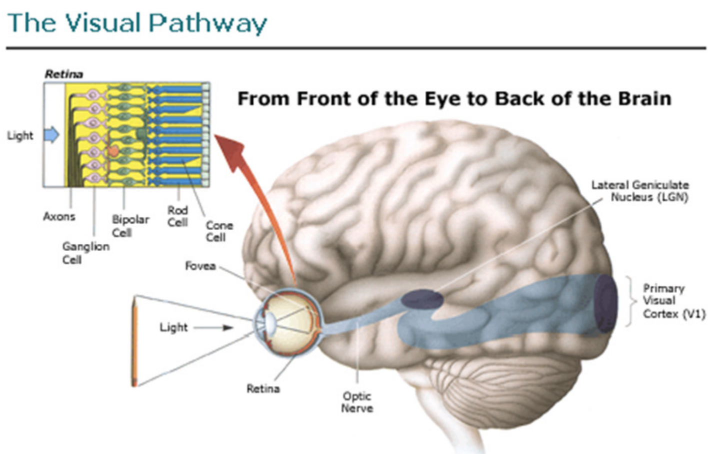
Smart Systems

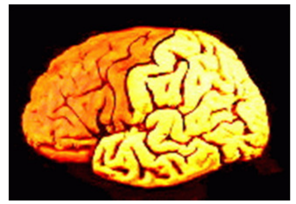


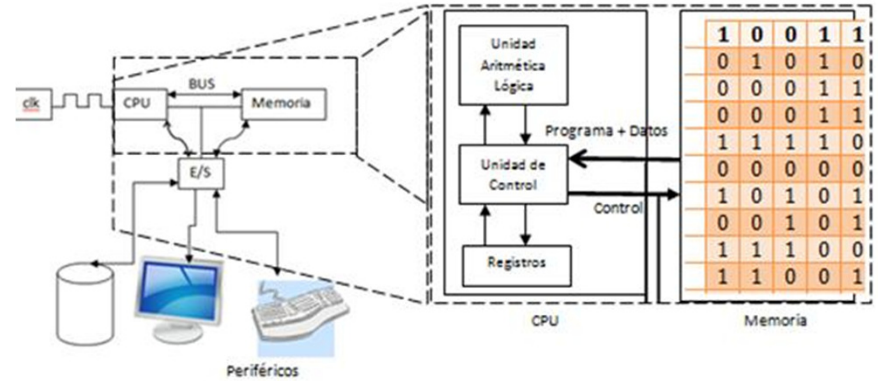
Conventional Computing



Neuromorphic Systems: Introduction



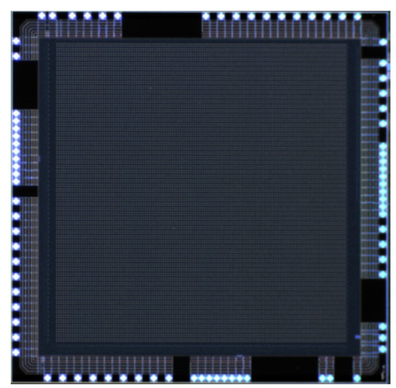




Smart Systems

- Highly parallel (10^{10}-10^{11} neurons)
- Massively interconnectd (10^{14} synapsis)
- Low frequency component
- Highly noisy and variable components
- Adaptive components
- Communication by electrical spikes
- Asynchronous processing
- Low power consumption: 20-25W

Artificial Intelligence

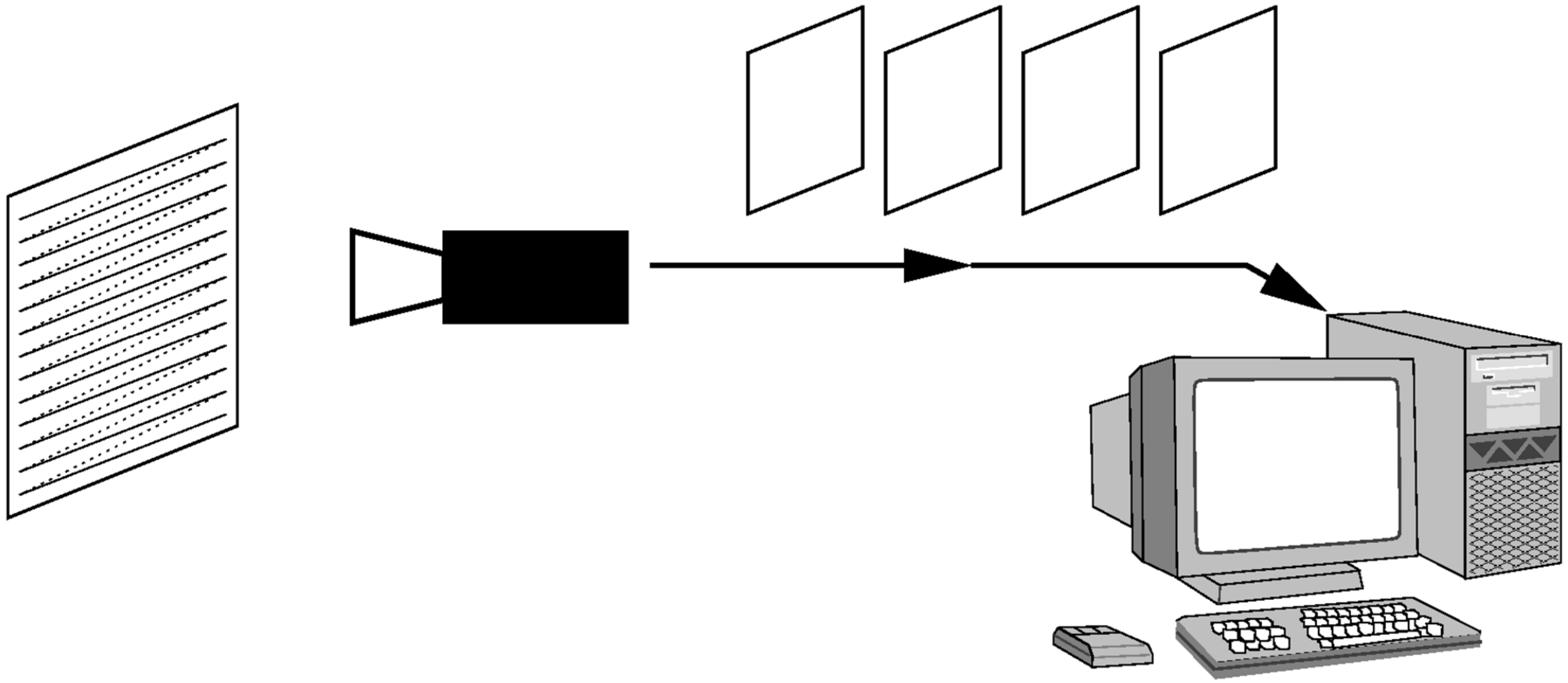


Conventional Computing

- Separated memory-CPU (Von Neumann)
- Sequential processing
- High frequency computing
- High frequency communication
- High precision computing
- Time sampled information
- Synchronous computation
- High power consumption

Neuromorphic Engineering
Neuromorphic Sensors and Processors

Conventional vision processing: frame-based



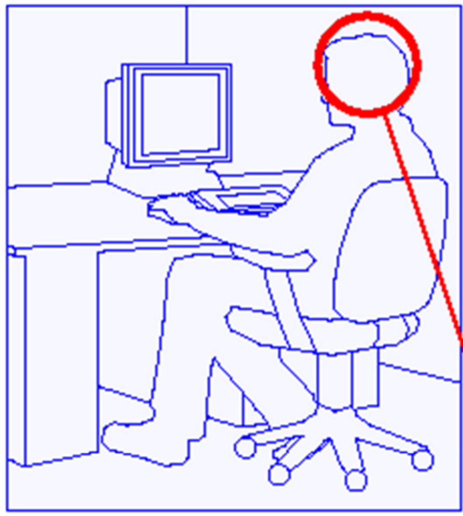
FRAME based conventional vision processing
- Frame-based acquisition
- Localization of ROIs
- Feature Extraction Stages
- Feature Combination Stages
- Classification/Decision Stages

Sequential

Synchronous

Biological Vision Systems

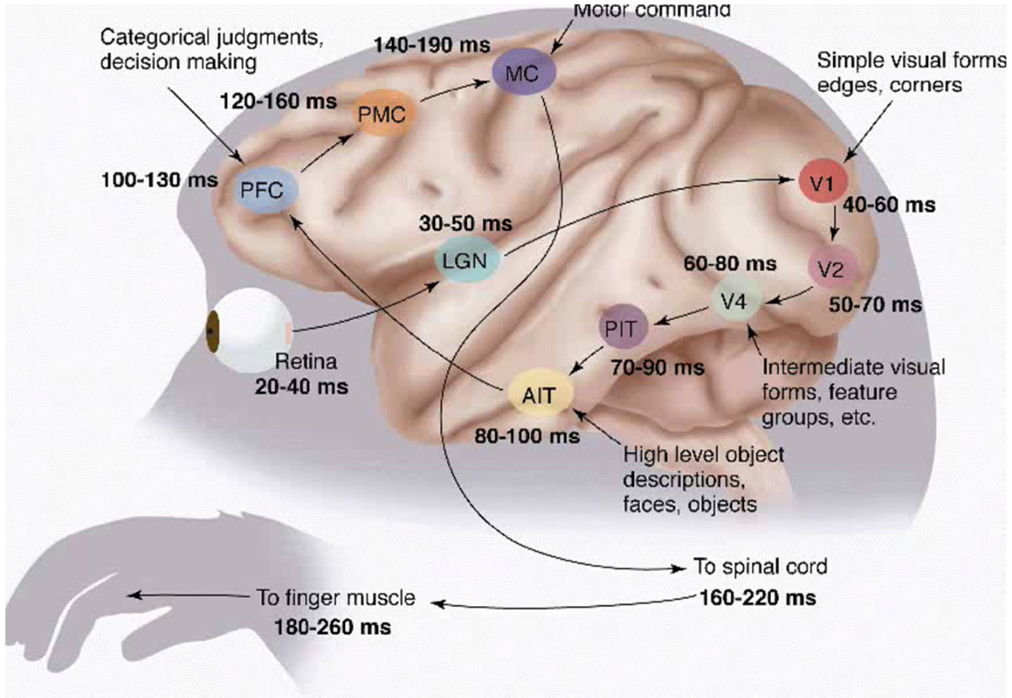
Biology

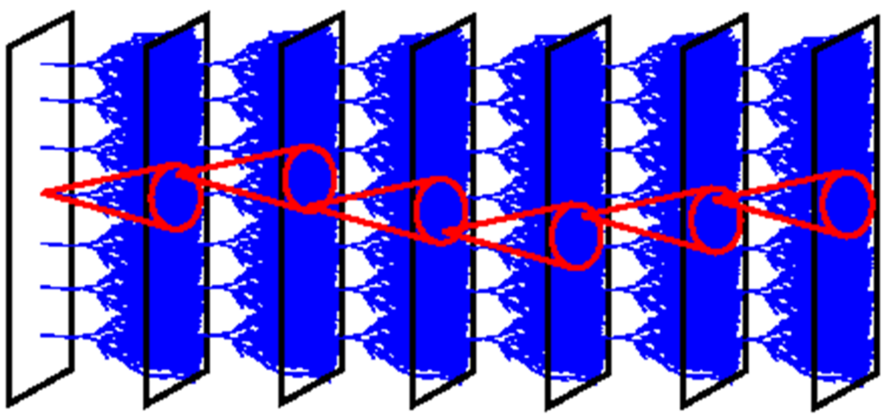


Recognition Delay < 150ms

- feedforward
- 1 spike/neuron

Simon Thorpe
Nature 1996

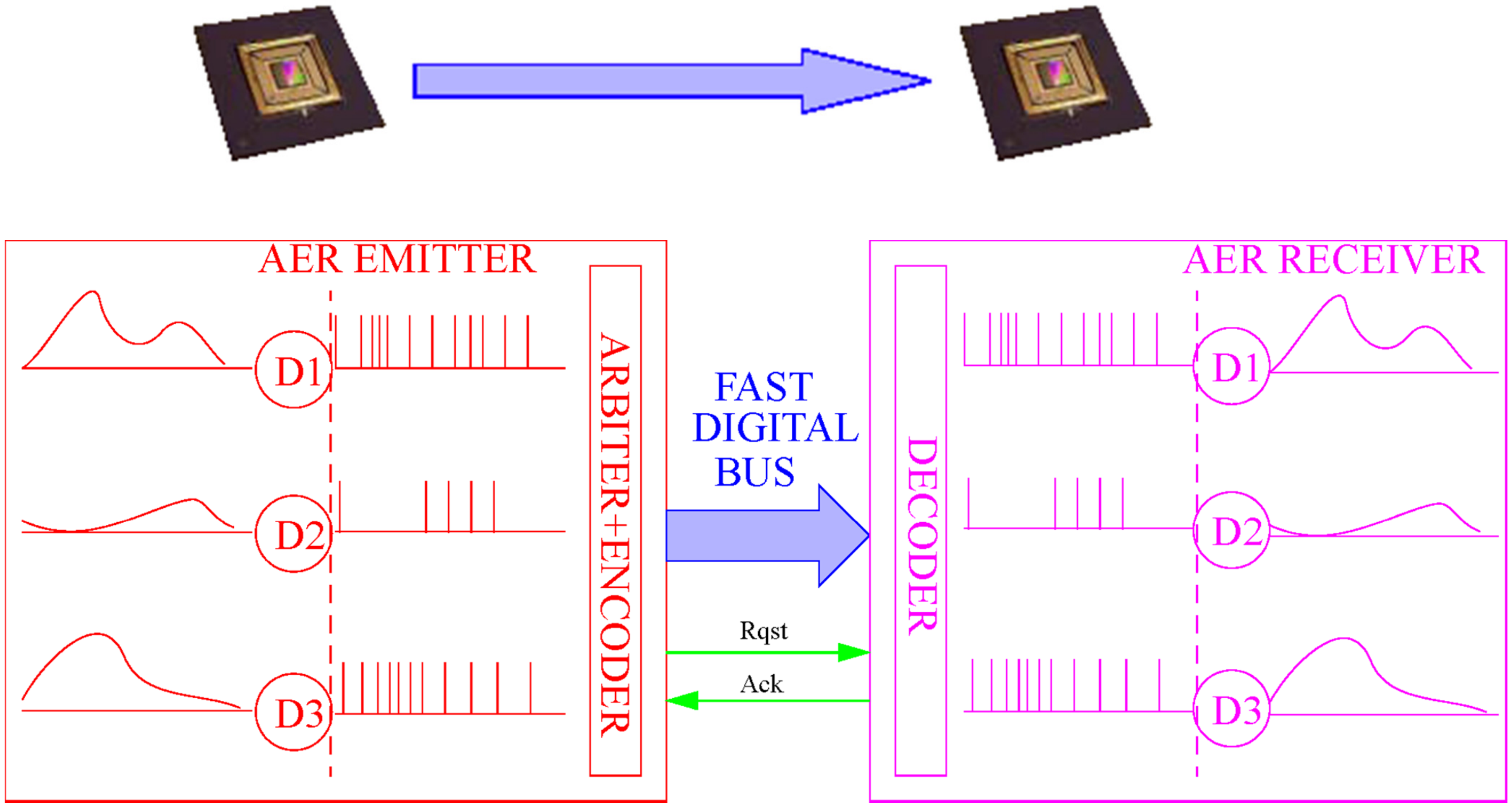




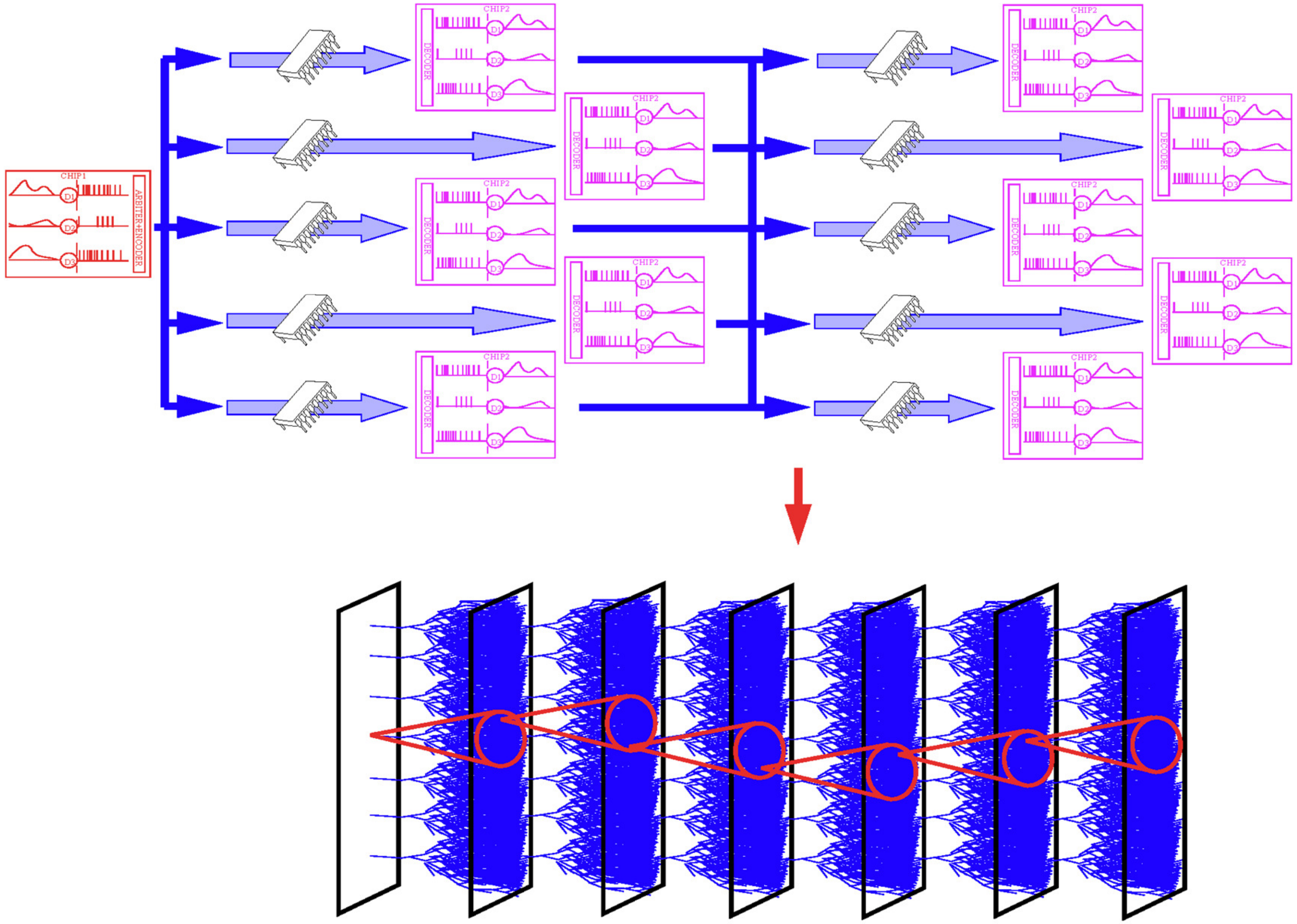
Parallel

Asynchronous

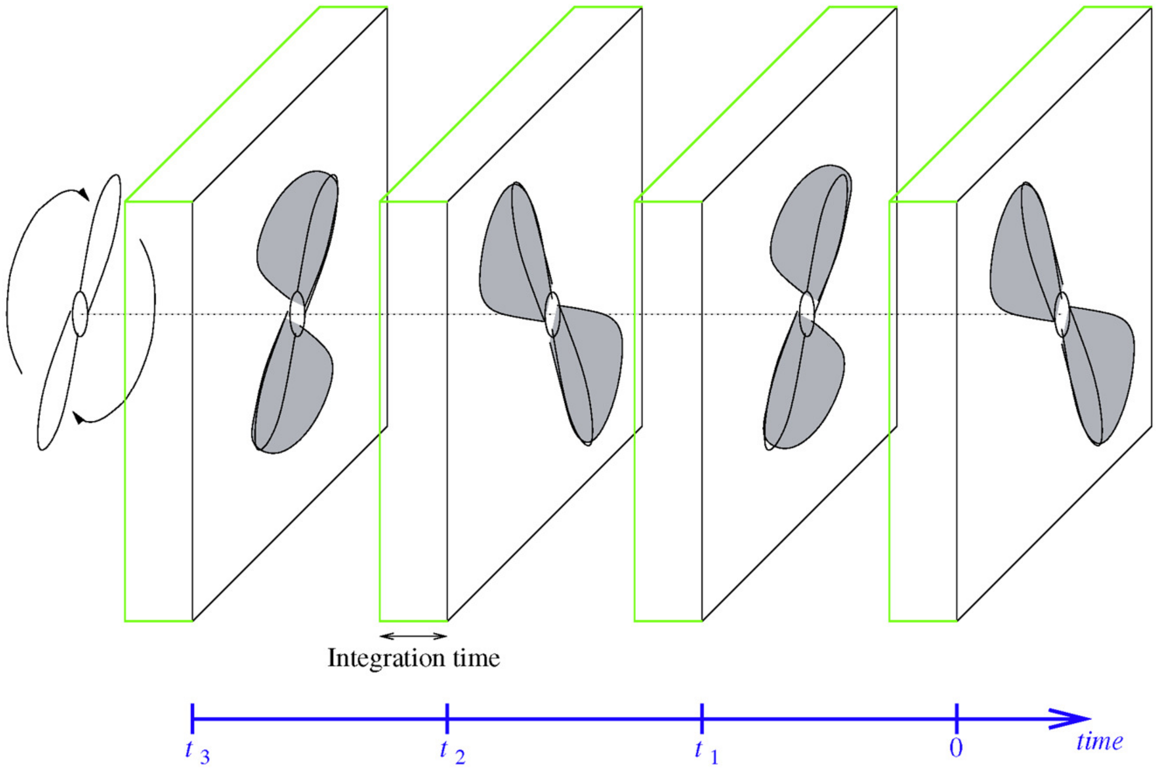
Massive Connectivity: Address Event Representation

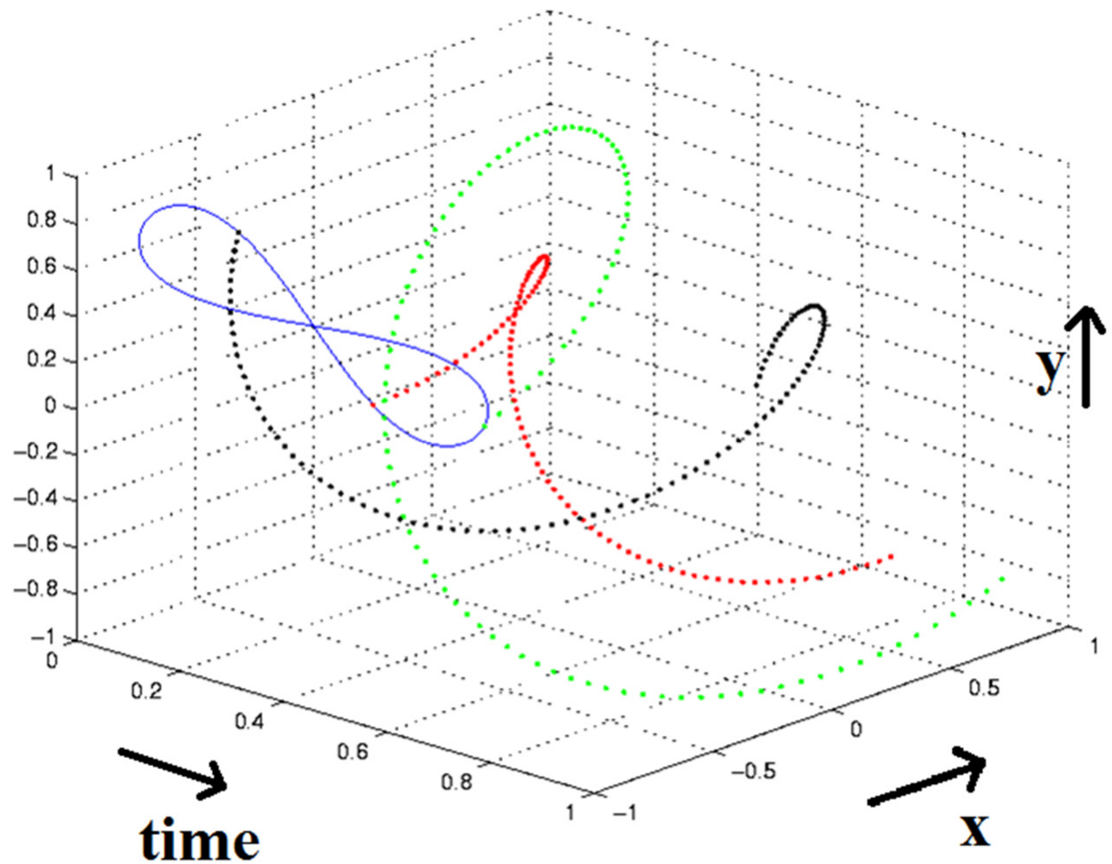


Massive Connectivity: Address Event Representation

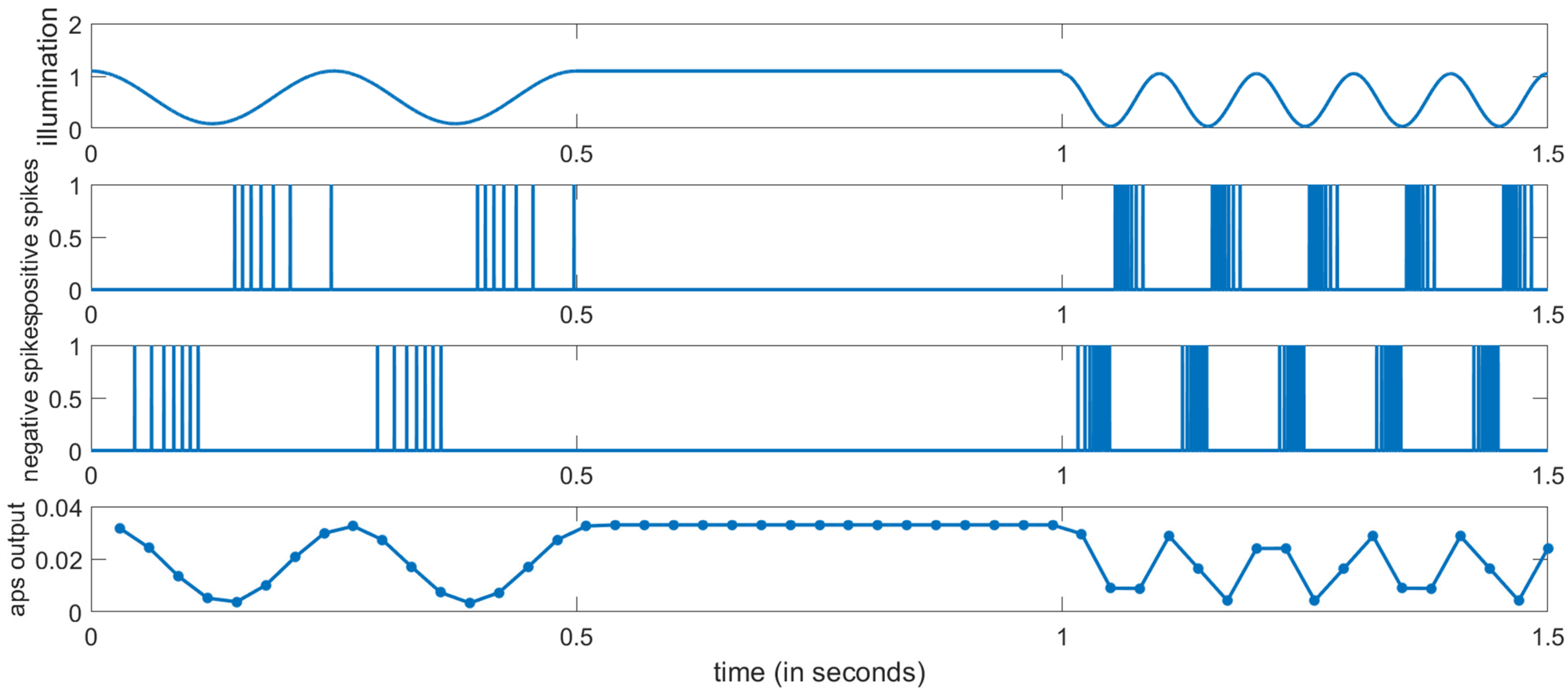


Neuromorphic Vision Sensors vs Frame-based Vision Sensors



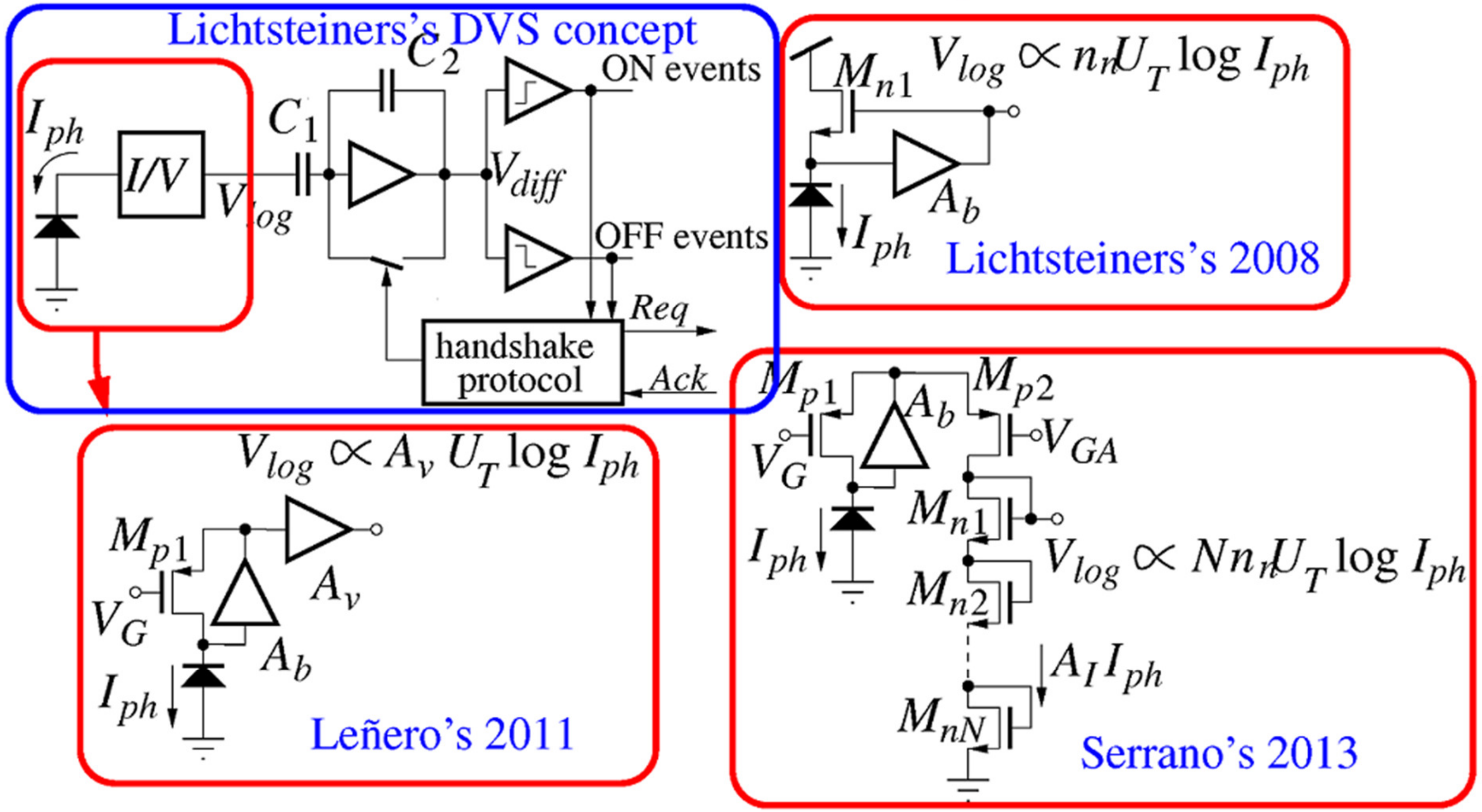


Temporal contrast: Dynamic Vision Sensor



Dynamic Vision Sensor

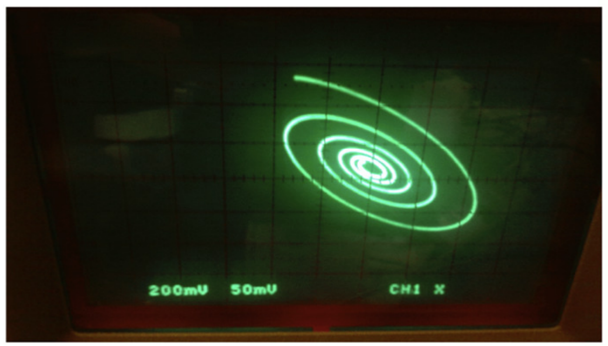
Improved Contrast Sensitivity DVS

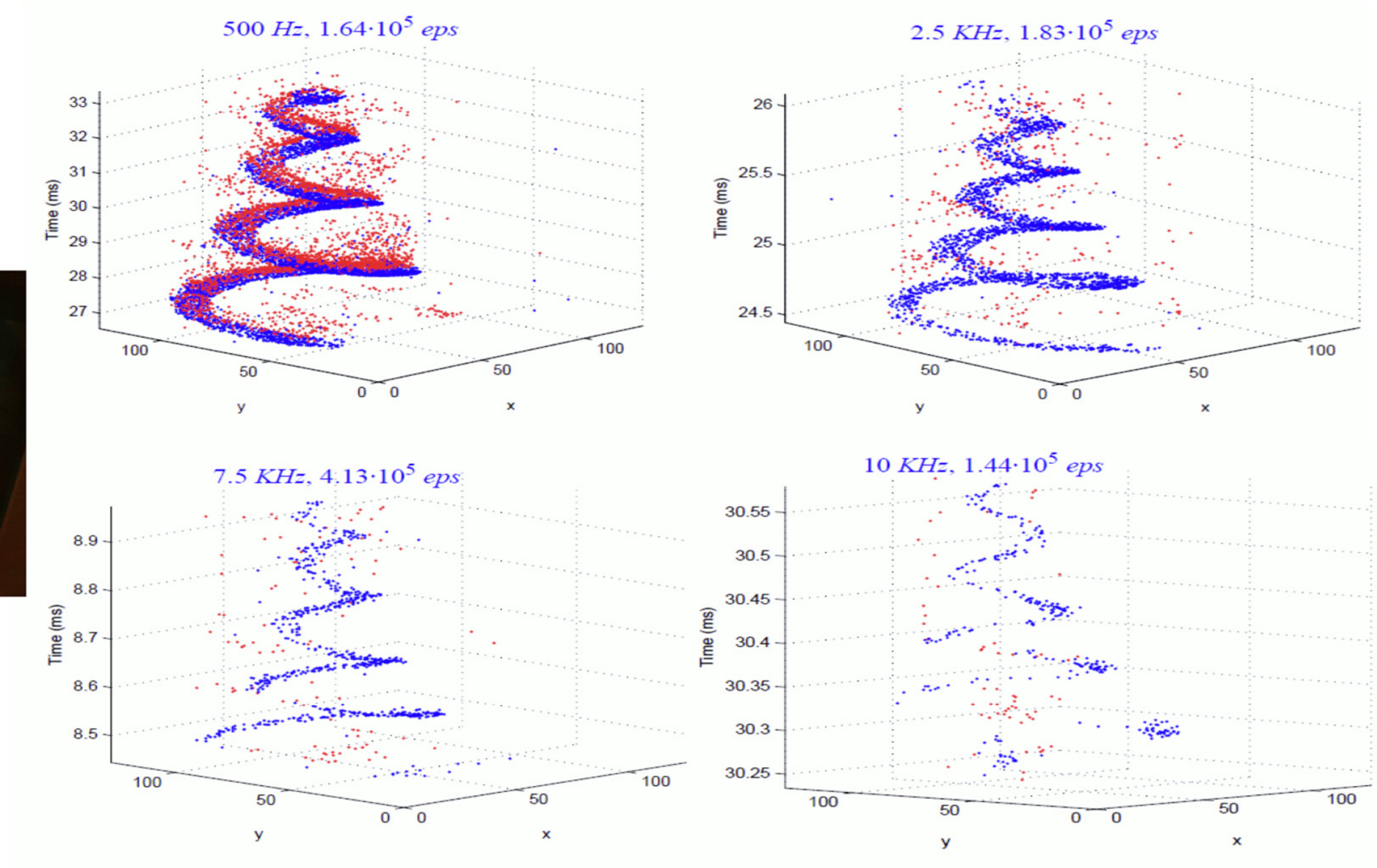


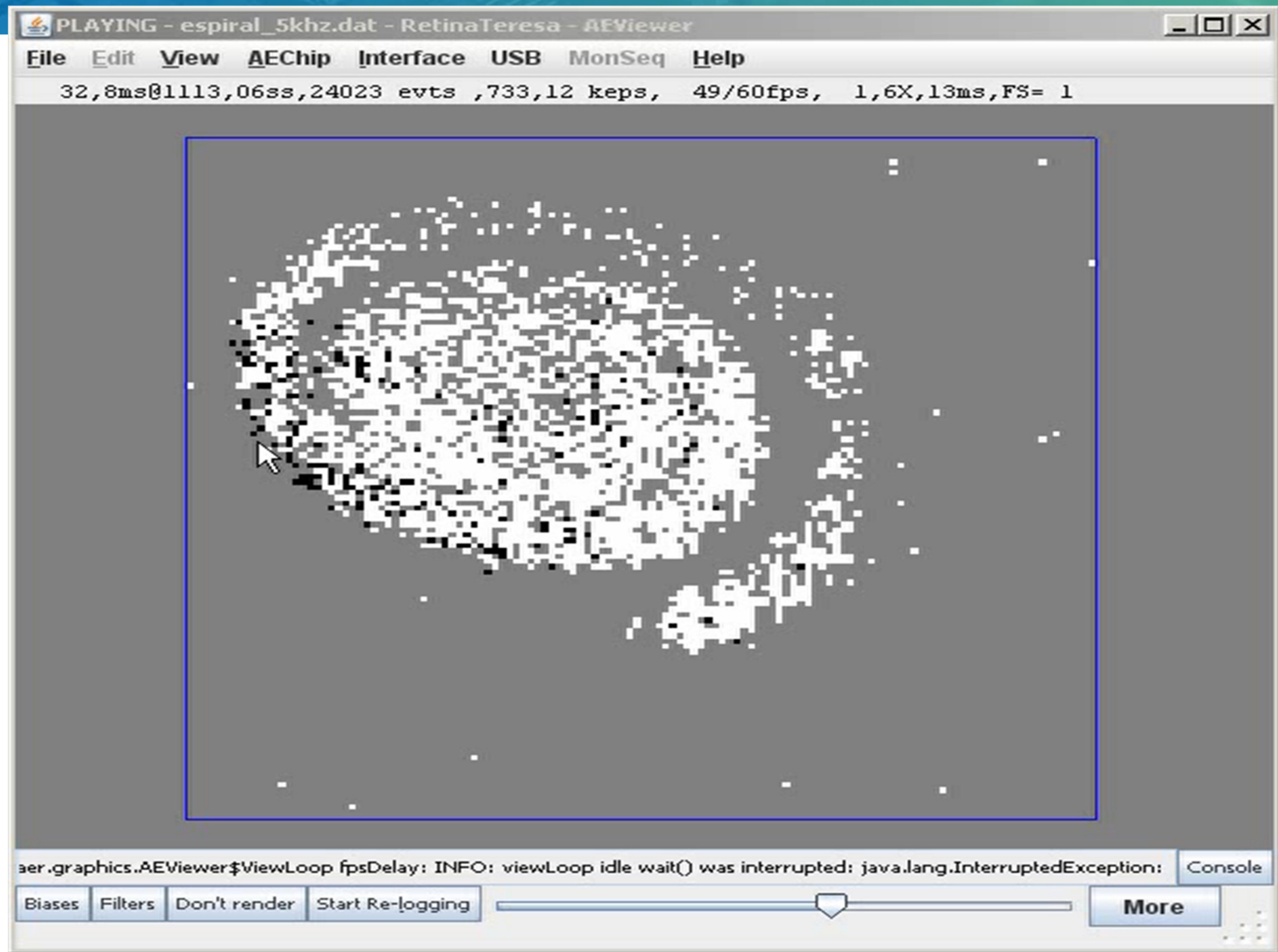
J. A. Leñero-Bardallo, T. Serrano-Gotarredona and B. Linares-Barranco, "A 3.6μs latency asynchronous frame-free event-driven dynamic-vision-sensor," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, June 2011.

T. Serrano-Gotarredona and B. Linares-Barranco, "A 128x128 1.5% contrast sensitivity 0.9% FPN 3μs latency 4mW asynchronous frame-free event-driven dynamic-vision-sensor using transimpedance preamplifiers," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, March 2013.

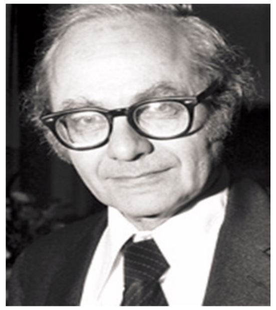
IMSE Motion Retina

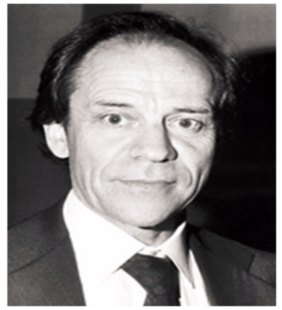




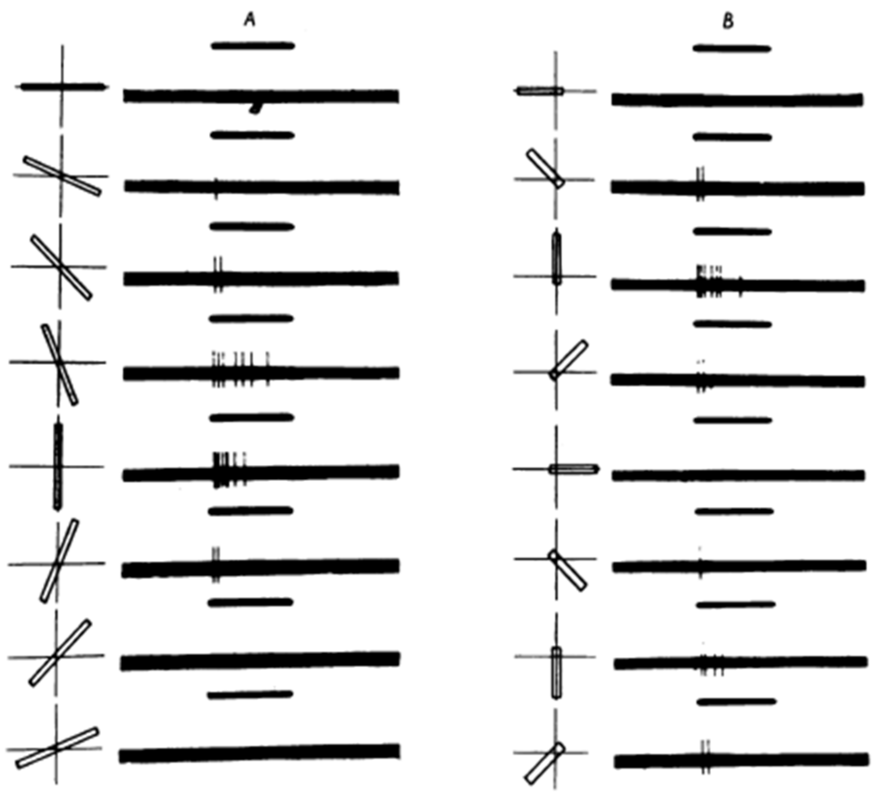

PLAYING - espiral_5khz.dat - RetinaTeresa - AEViewer
File
Edit
View
AEChip
Interface
USB
MonSeq
Help
32,8ms@1113,06ss,24023 evts ,733,12 keps, 49/60fps, 1,6X,13ms,FS= 1
aer.graphics.AEViewer$ViewLoop fpsDelay: INFO: viewLoop idle wait() was interrupted: java.lang.InterruptedException:
Console
Biases
Filters
Don't render
Start Re-logging
More

Spiking ConvNets: Biological Vision Processing





Hubel & Wiesel Nobel Price 1981



J. Physiol. (1959) 148, 574–591

RECEPTIVE FIELDS OF SINGLE NEURONES IN THE CAT'S STRIATE CORTEX

By D. H. HUBEL* and T. N. WIESEL*

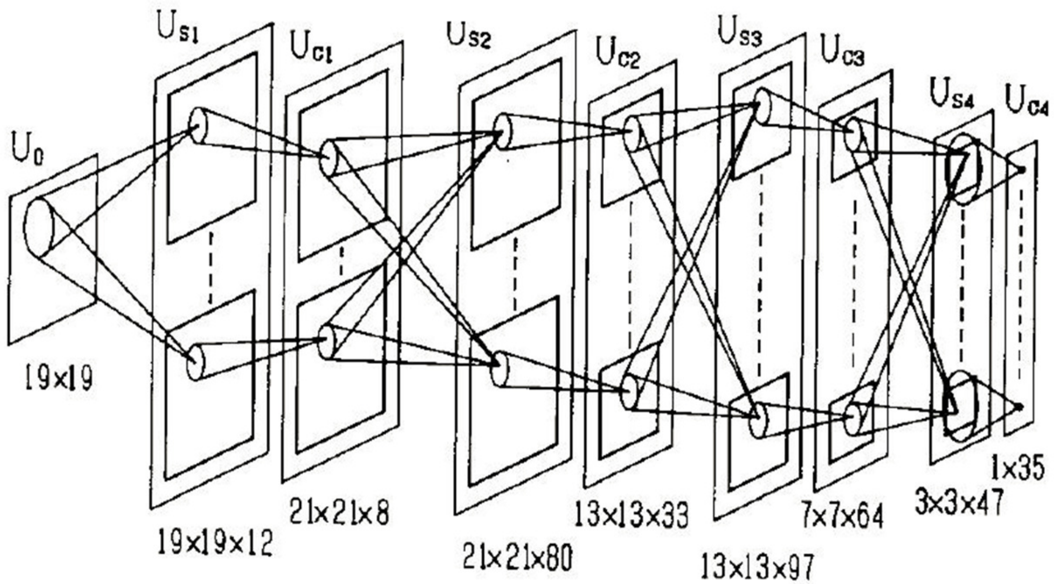
From the Wilmer Institute, The Johns Hopkins Hospital and University, Baltimore, Maryland, U.S.A.

(*Received* 22 *April* 1959)

In the central nervous system the visual pathway from retina to striate cortex provides an opportunity to observe and compare single unit responses at several distinct levels. Patterns of light stimuli most effective in influencing units at one level may no longer be the most effective at the next. From differences in responses at successive stages in the pathway one may hope to gain some understanding of the part each stage plays in visual perception.

By shining small spots of light on the light-adapted cat retina Kuffler (1953) showed that ganglion cells have concentric receptive fields, with an 'on' centre and an 'off' periphery, or vice versa. The 'on' and 'off' areas within a receptive field were found to be mutually antagonistic, and a spot restricted to the centre of the field was more effective than one covering the whole receptive field (Barlow, FitzHugh & Kuffler, 1957). In the freely moving light-adapted cat it was found that the great majority of cortical cells studied gave little or no response to light stimuli covering most of the animal's visual field, whereas small spots shone in a restricted retinal region often evoked brisk responses (Hubel, 1959). A moving spot of light often produced stronger responses than a stationary one, and sometimes a moving spot gave more

Neocognitron - Fukushima 1969



8 layers, 376 convolutions

322 IEEE TRANSACTIONS ON SYSTEMS SCIENCE AND CYBERNETICS, VOL. SSC-5, NO. 4, OCTOBER 1969

Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements

KUNIHIKO FUKUSHIMA, MEMBER, IEEE

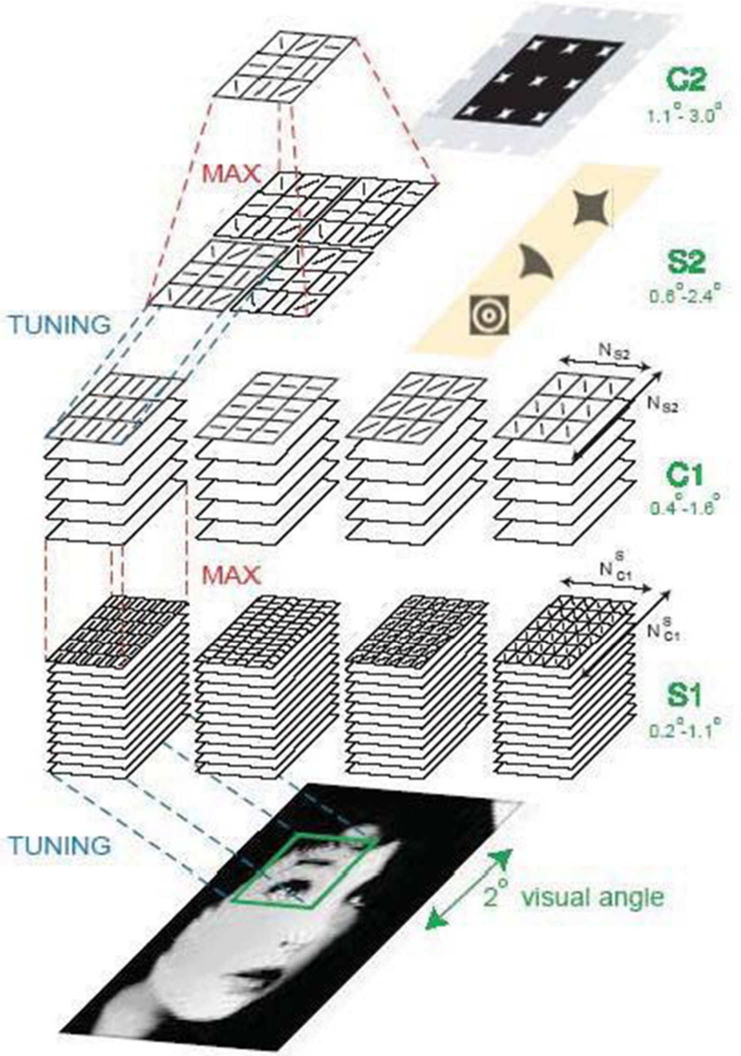
***Abstract*—A new type of visual feature extracting network has been synthesized, and the response of the network has been simulated on a digital computer. This research has been done as a first step towards the realization of a recognizer of handwritten characters. The design of the network was suggested by biological systems, especially, the visual systems of cat and monkey.**

The network is composed of analog threshold elements connected in layers. Each analog threshold element receives inputs from a

Békésy's studies of sensory inhibition [10], Glezer's psychological measurements of the receptive field of the human retina [11], Motokawa's psychophysiological studies of the retinal induction field [12], Watanabe's studies of the eye movement with an eye-marker camera which show how the eye wanders while watching a pattern [13], and Attneave's psychological experiments which investigate at

Serre & Poggio (MIT)

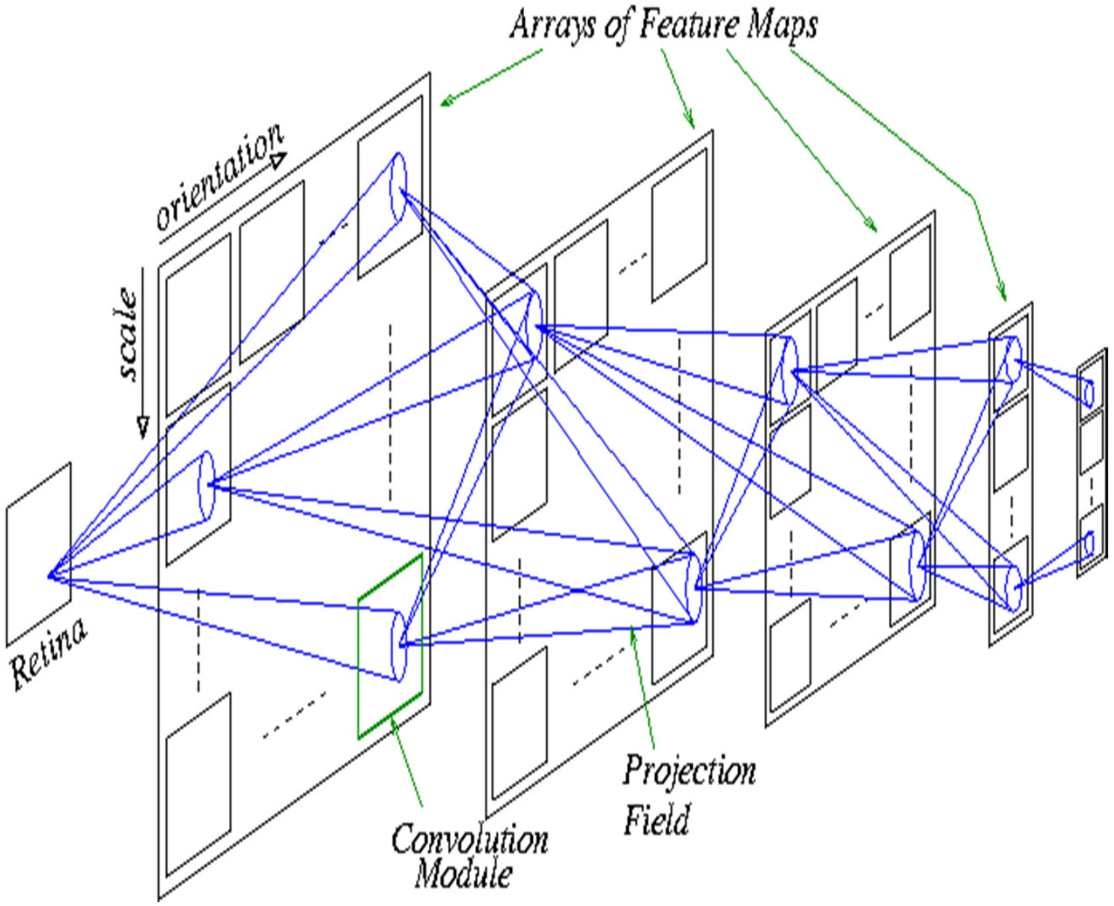
Ventral Stream Model for Immediate Recognition



- projection field processing
- short-range & dense for first layers
- long-range & sparse for later layers
- hard-wired for first layers
- plastic for later layers
- first layers: massive 2D filtering for different angles and scales
- first layers: basic feature extraction
- later layers: grouping of features -> abstractions

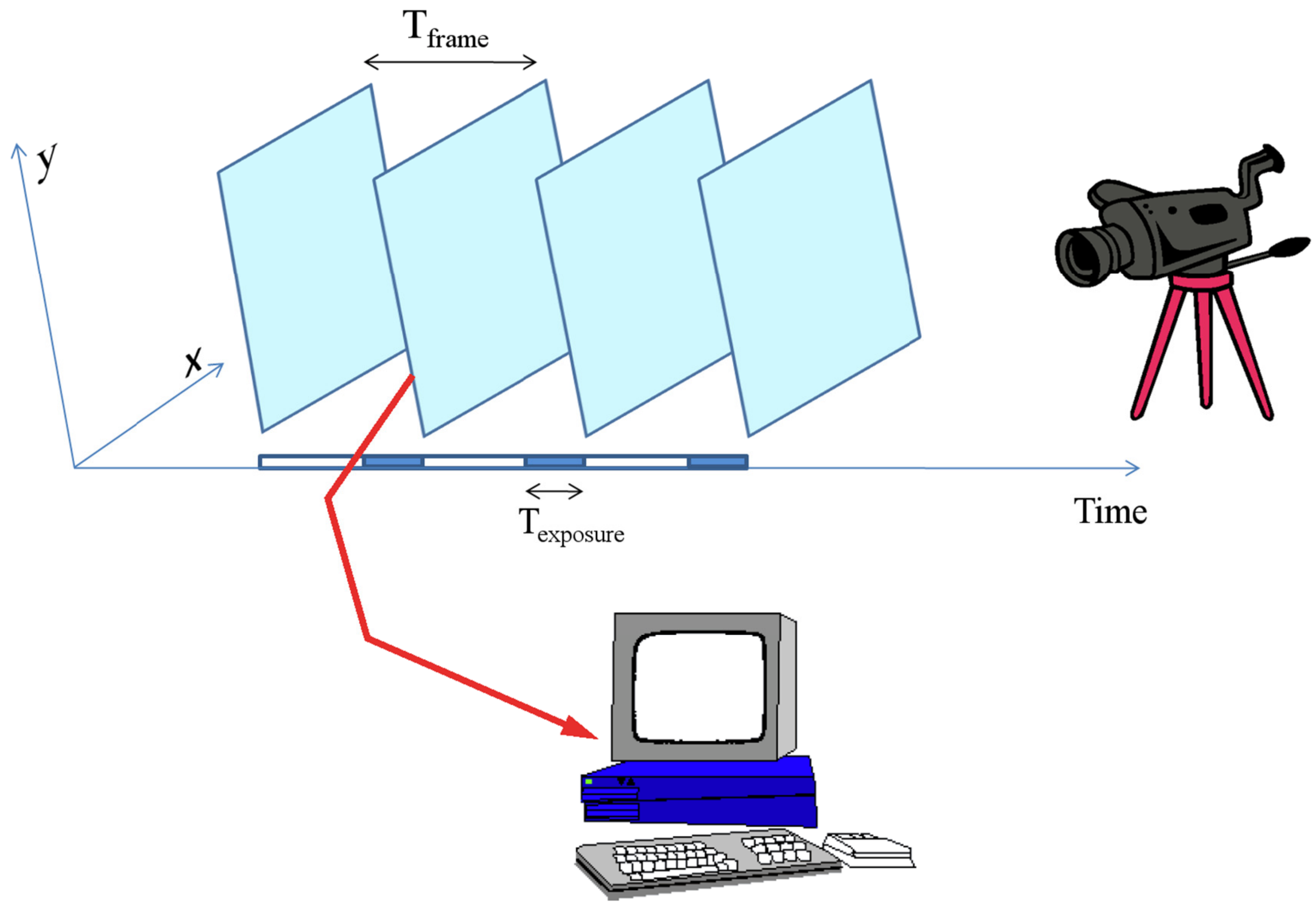
ConvNets Yann LeCun 1989--

- Back propagation learning in ConvNets
- VLSI implementations
- Hand-writen character recognition
- Signature verification
- Localizing object in images
- Word-level recognition
- Speech recognition
- Face detection
- Obstacle avoidance
- Visual guided navigation

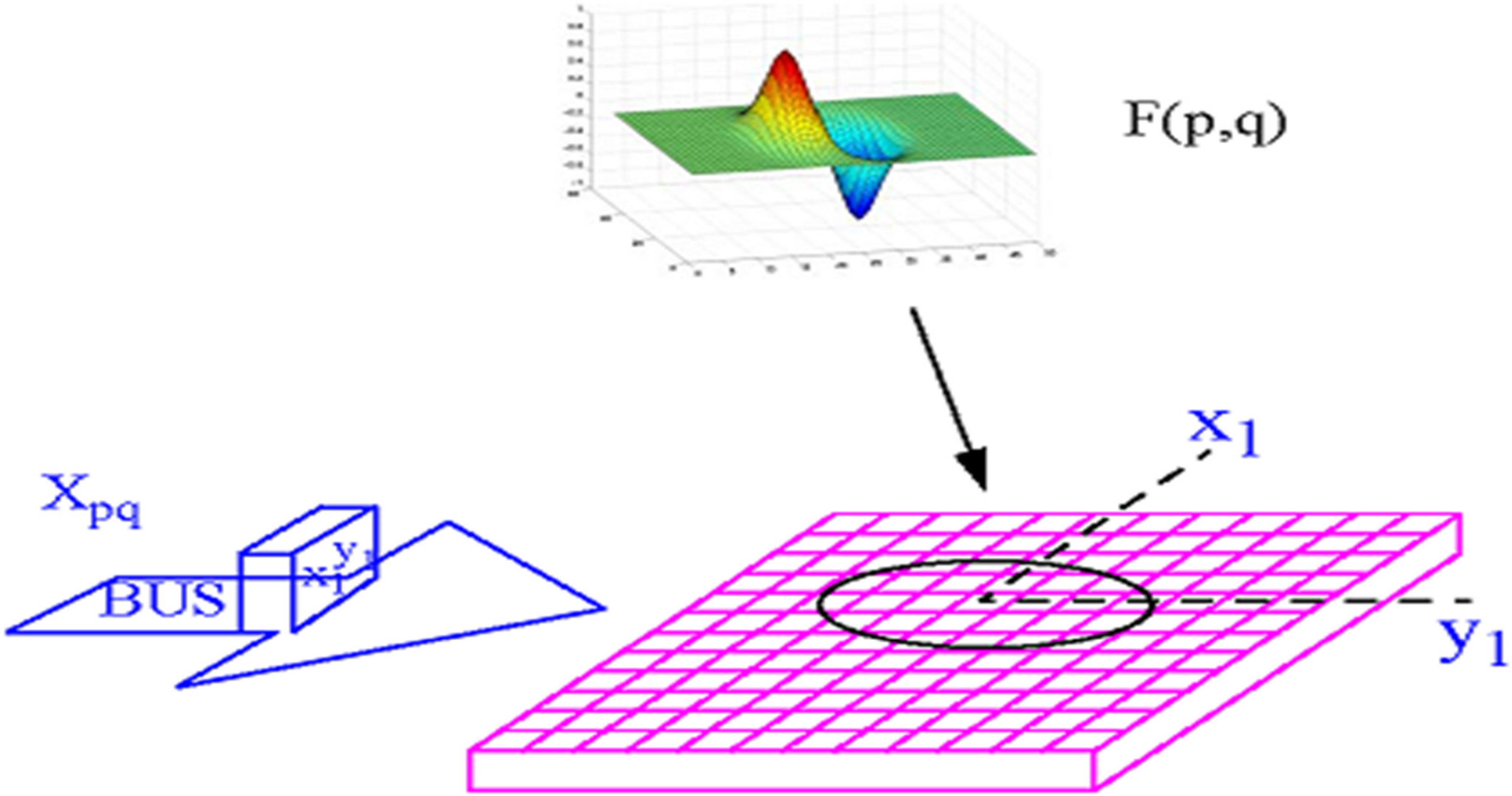


Conventional ConvNets: frame-based

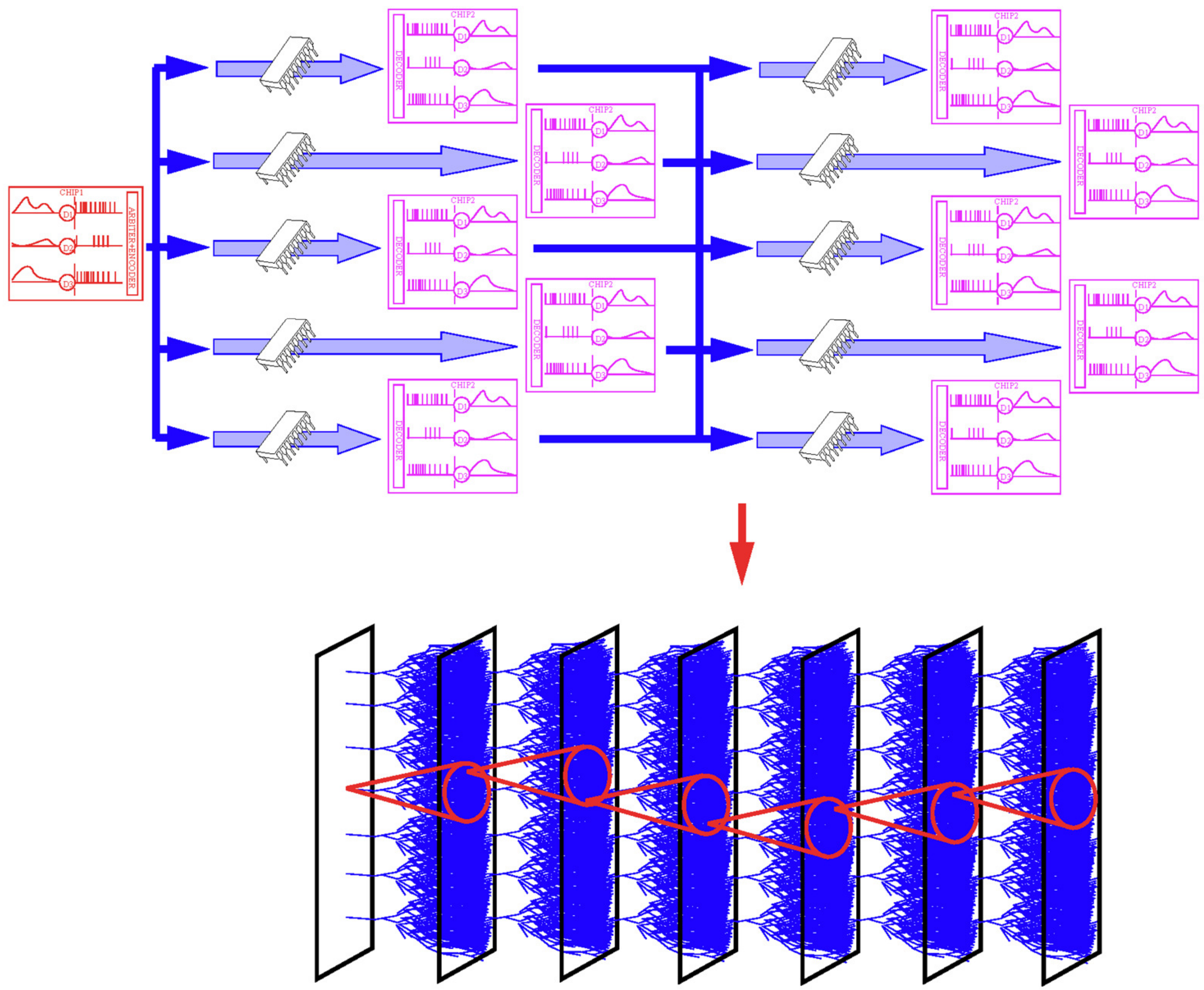
So far: Video Frames & non-Spikes



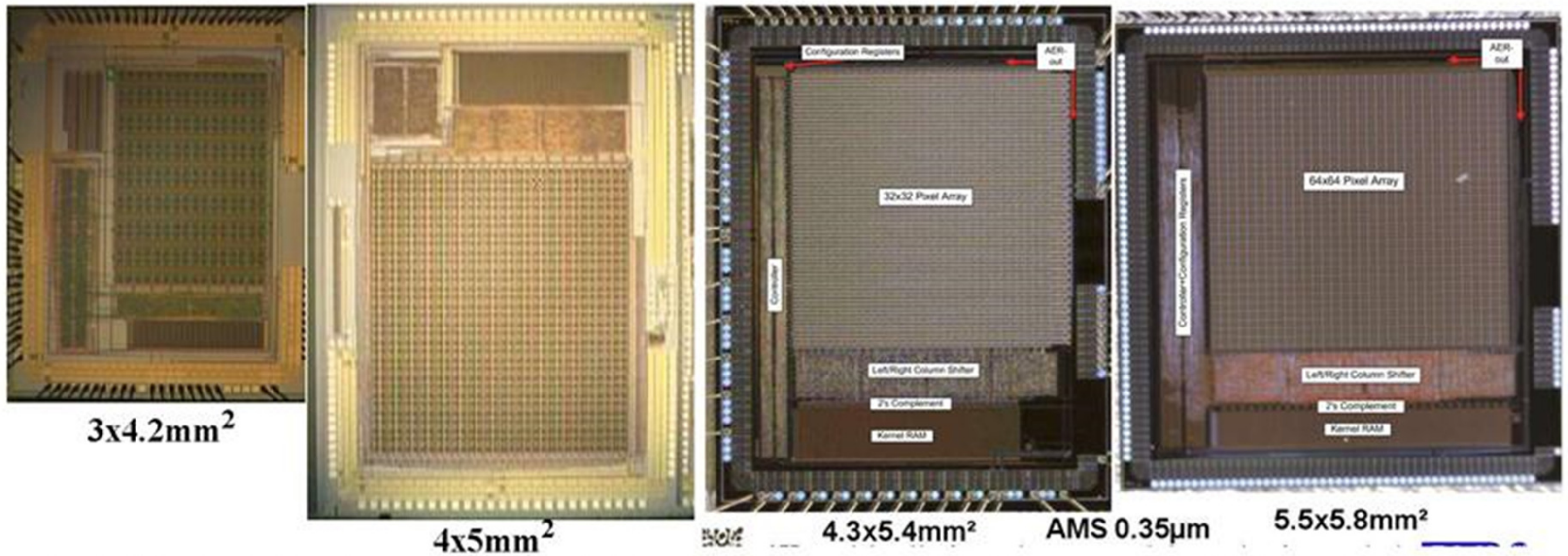
Event-based Convolutions



Matrix of integrators in
the receiver chip



Neuromorphic Convolution Processors



Analog 16x16 Conv Chip | Analog 32x32 Conv Chip | Digital 32x32 Conv Chip | Digital 64x64 Conv Chip

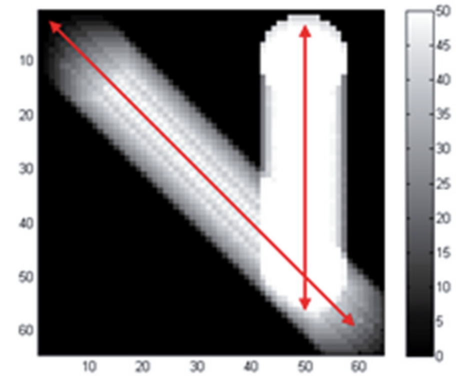
	Technology	Core area (mm^2)	Neurons/core	Max Performance (Gop/s)	Power (mW)	E_{SOP} (pJ)
(Brink2013) FPAA - GaTech	350nm	25	100	0.65e-3	6.5e-3	10
(Park2014) IFAT - UCSD	90nm	139	2000	0.073	1.6	22
(Benjamin2014) NeuroGrid - Stanford Univ.	180nm	170	65636	3.3	3.1	31.2
(Moradi2018) DYNAPs - ETHZ	180nm	7.5	256	0.038	0.8-2.7	17
(Davies2018) Loihi - Intel	14nm	0.4	1024	440	11500	23.6
(Akopyan2015) TrueNorth - IBM	28nm	0.1	256	2.5	65	26.0
(Mayr2016) SC - Dresden Univ.	28nm	0.36	64	0.13	15	850
(Painkras2013) SpiNNaker - Human Brain Project	130nm	3.75	~200	0.086	1000	11300
(Frenkel2019) MorphIC - Delft Univ.	28nm	0.086	256	0.105	0.89	8.4
(Schmitt2017) HICANN – Human Brain Project	180nm	50	512	45	5200	116
(Neckar2019) Braindrop - Stanford Univ.	28nm	0.65	4096	-	-	0.38
(Camuñas2012) Digital Spiking Convolution Processor	350nm	29	4096	3.01	200	83

High-Speed Processing

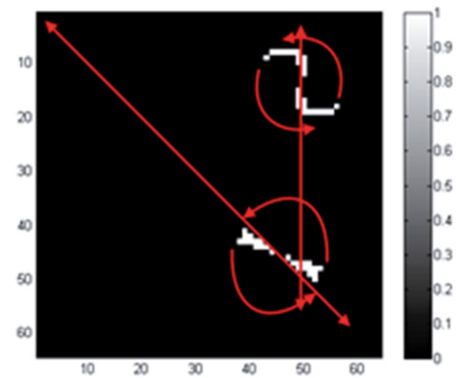
2000 rps

16 pixels diameter

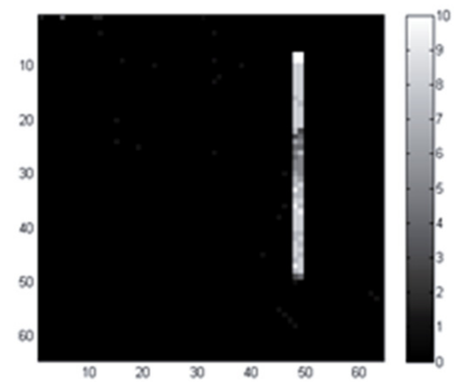
23x23 kernel



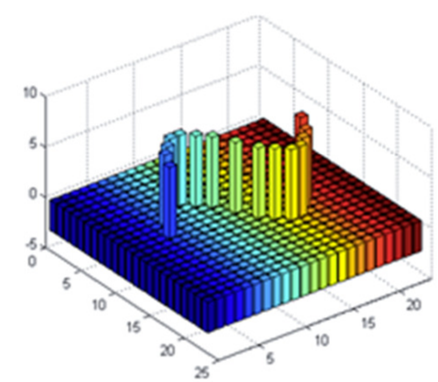
a) Input trajectories of both propellers rotating at a speed of 2000 revolutions per second



b) Short-time reconstruction of both propellers shapes

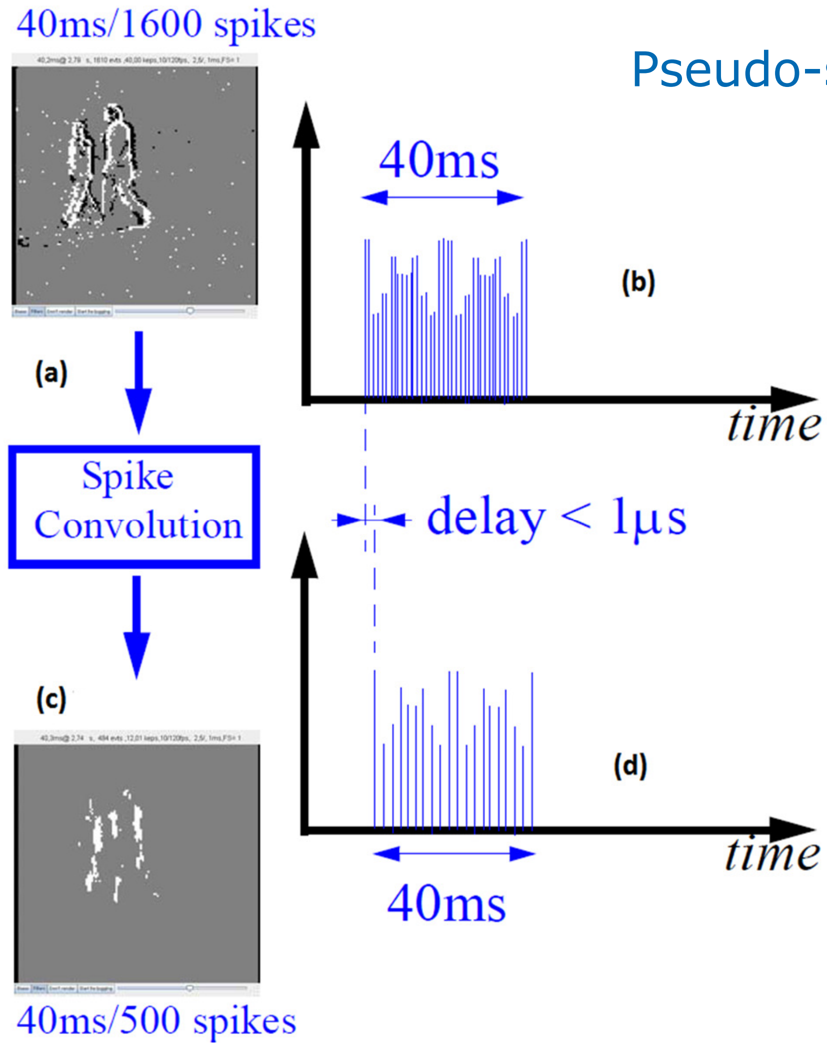


c) Output events located at the center of the S-shape propeller

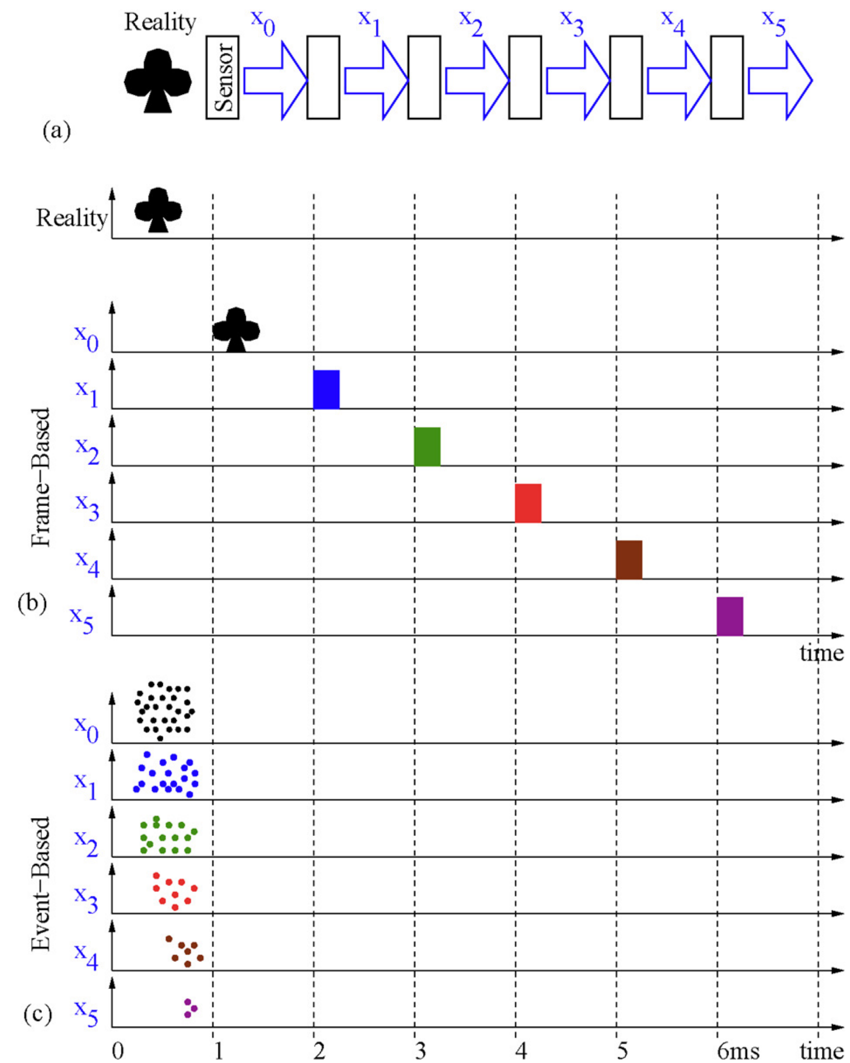


d) Convolution kernel used to discriminate the S-shape propeller

Neuromorphic Processing: Pseudo-simultaneity

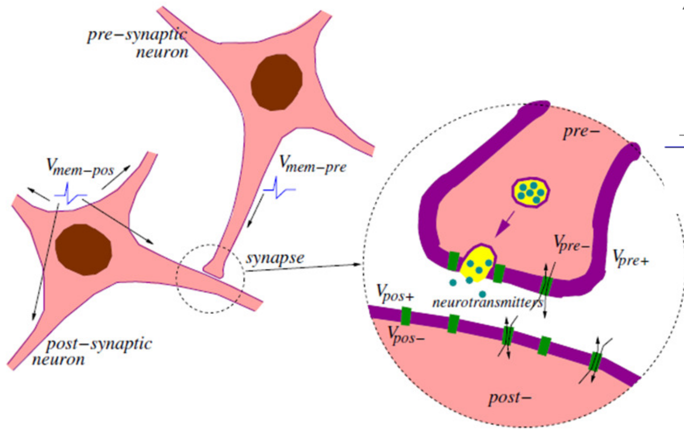


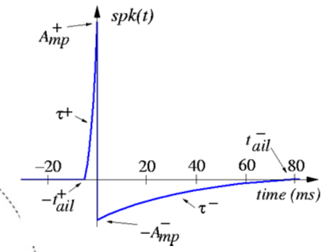
Pseudo-simultaneity



- …, **T. Serrano-Gotarredona**, et al. "Comparison Between Frame-Constrained Fix-Pixel-Value and Frame-Free Spiking-Dynamic-Pixel ConvNets for Visual Processing," Frontiers in Neuromorphic Engineering in Front. in Neuroscience. 6.32.
- …, **T. Serrano-Gotarredona**, et al. A Configurable Event-Driven Convolutional Node with Rate Saturation Mechanism for Modular ConvNet Systems Implementation, Original Research, Front. Neurosci. - Neuromorphic Engineering, 2018.

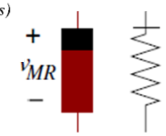
Memristors as synapses: dense memory, on-line learning

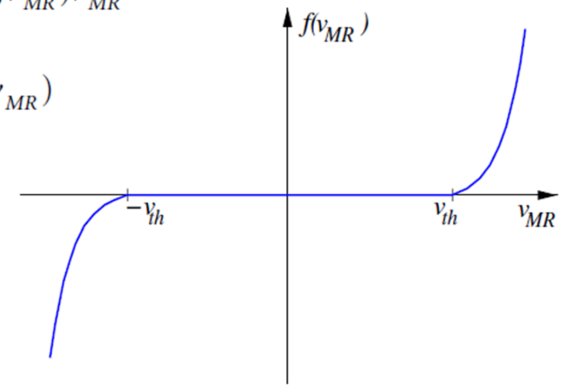


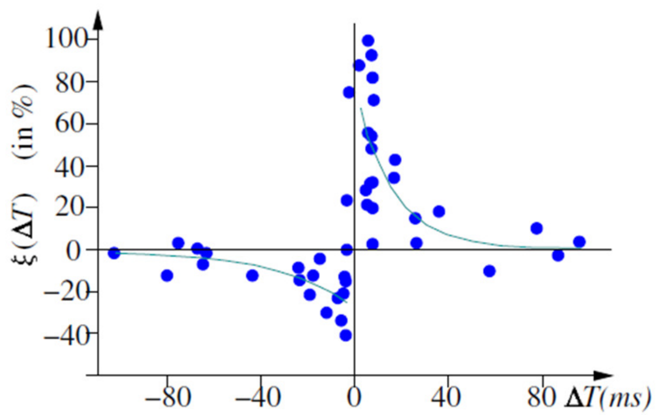


$$i_{MR} = g(w, v_{MR})v_{MR}$$

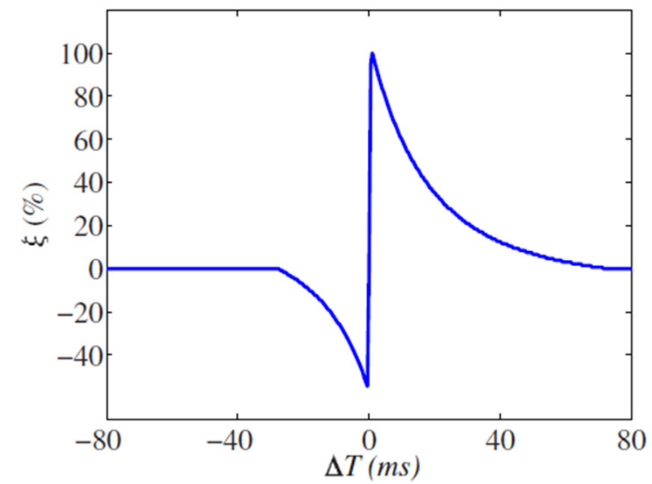
$$\dot{w} = f(w, v_{MR})$$







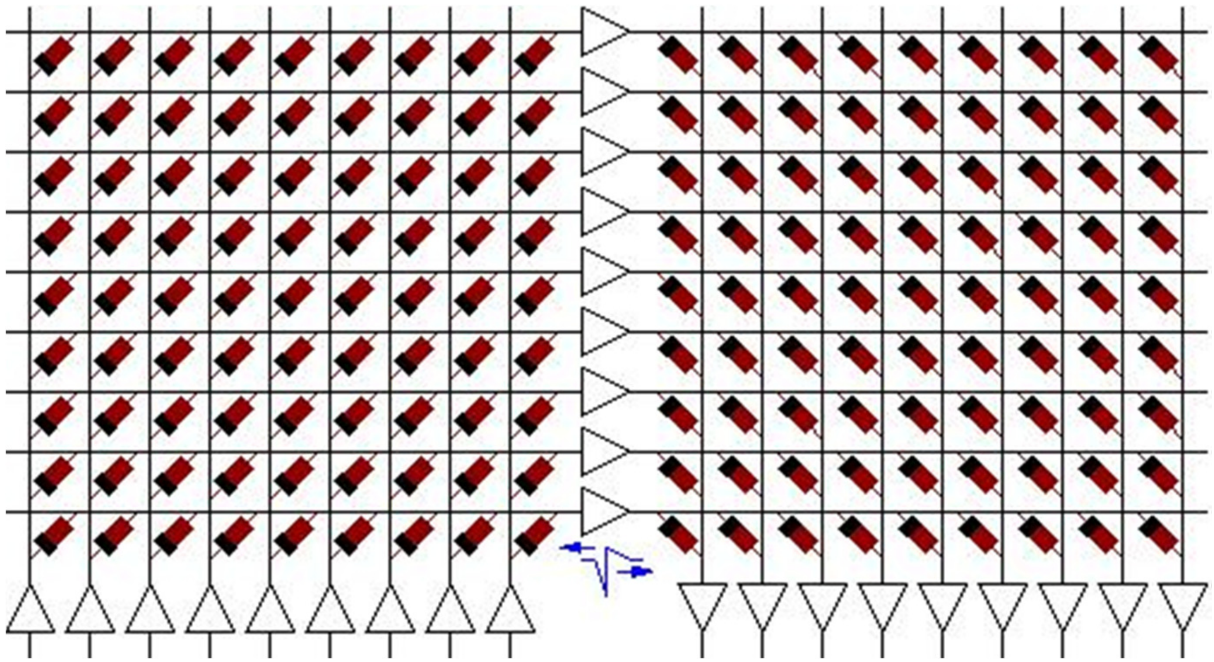
Biological Synapses

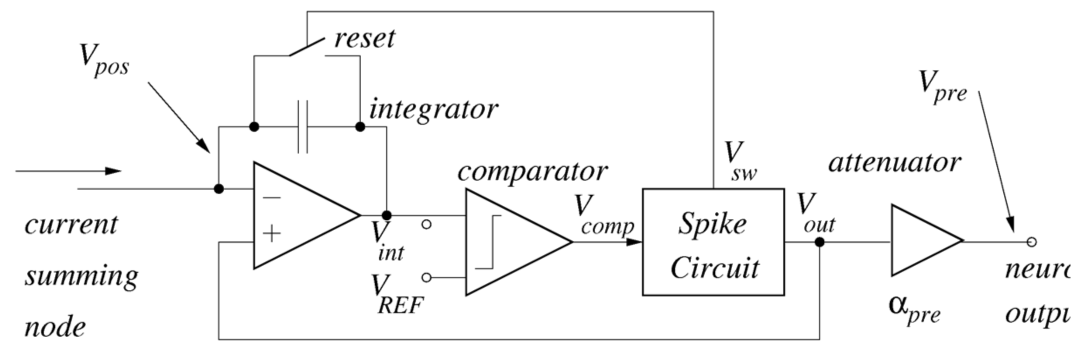


Memristive Synapses

- C. Zamarreño-Ramos, et al., "On Spike-Timing-Dependent-Plasticity, Memristive Devices, and building a Self-Learning Visual Cortex," Fron. in Neuromorphic Engineering. *Front. Neurosci.* **5:**26, 2011. **DOI**: 10.3389/ fnins.2011.00026, 17 March 2011.
- T. Serrano et al. "STDP and STDP Variations with Memristors for Spiking Neuromorphic Learning Systems," Fronti.in Neuromorphic Engineering in Front. in Neuroscience, **7**:02. **DOI:** 10.3389/ fnins.2013.00002, 6 January 2013

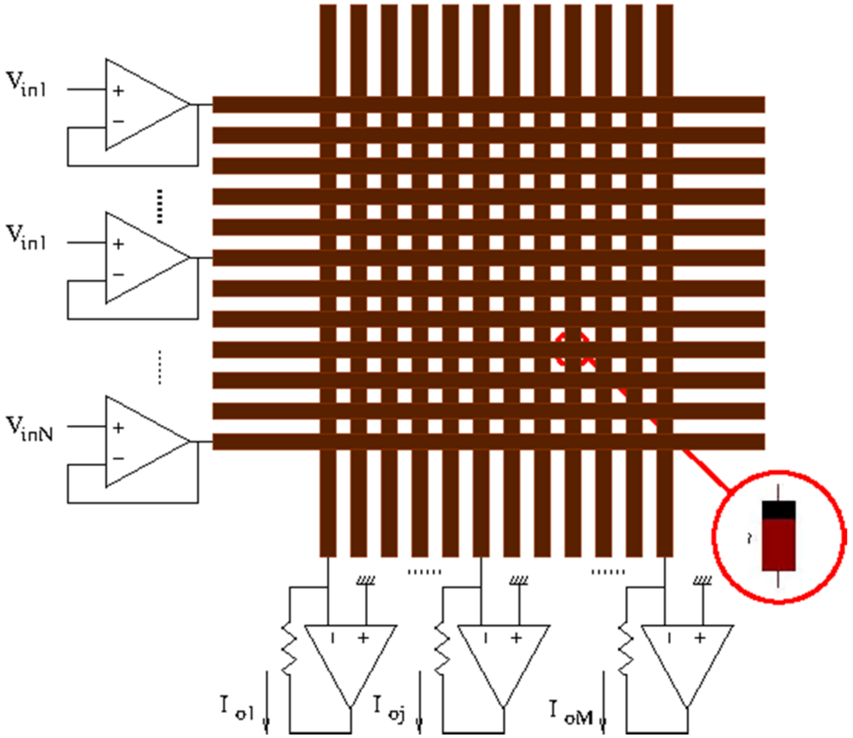
CMOS-memristive Neuromorphic Architectures

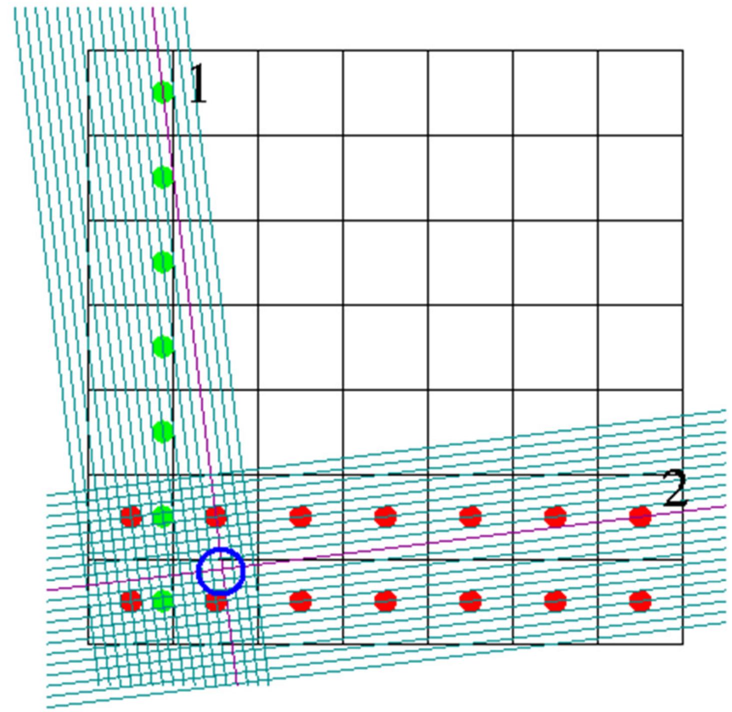


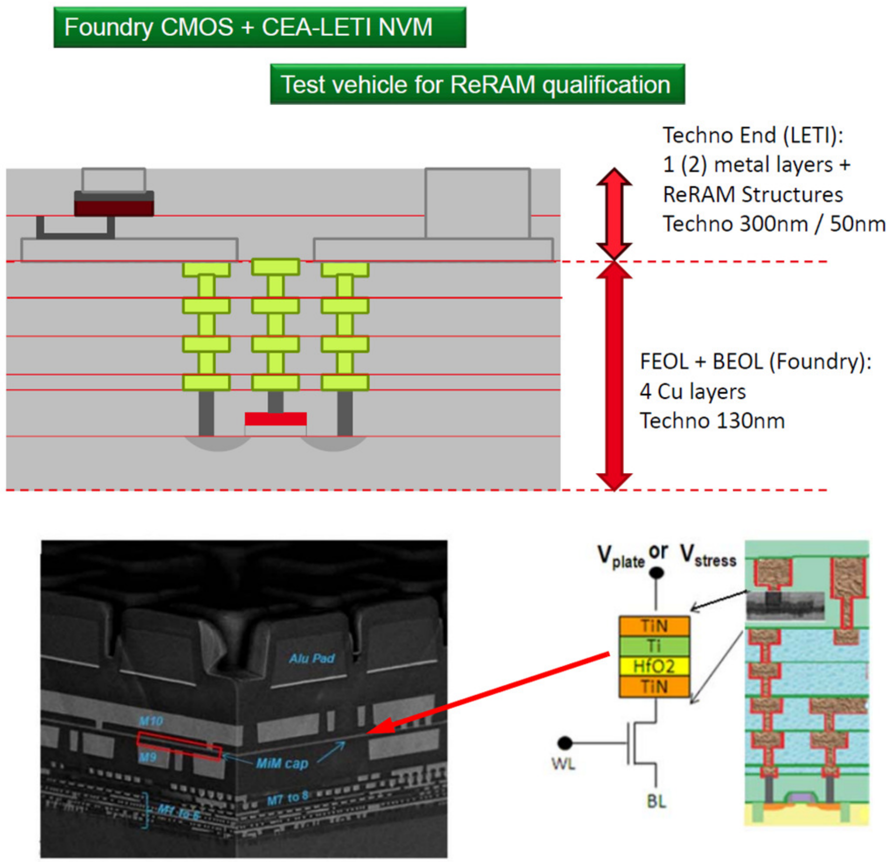


Asynchronous Spiking Computation

CMOS/memristors Processors







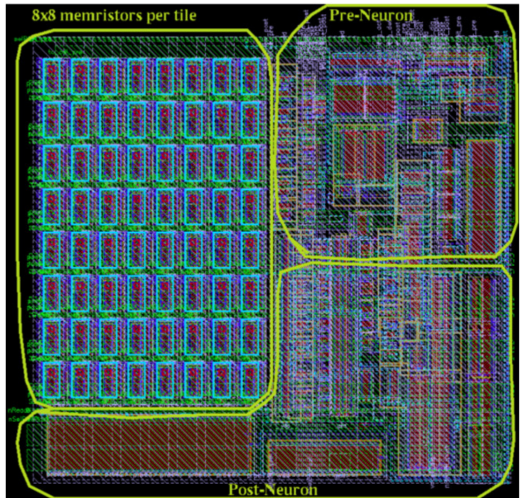
SEM cross section of CMOS 28nm stack including Metal Isulator Metal device

Memory technologies with multi-scale time constants for neuromorphic architectures (MeM-Scales). H2020-ICT-2019-2-871371. January 2020-December 2022

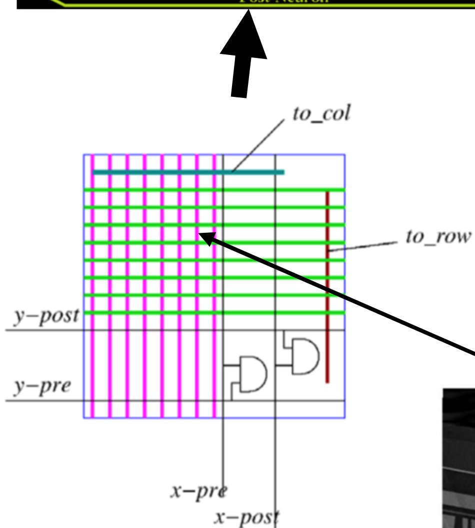
Procesadores Neuromórficos CMOS/memristores

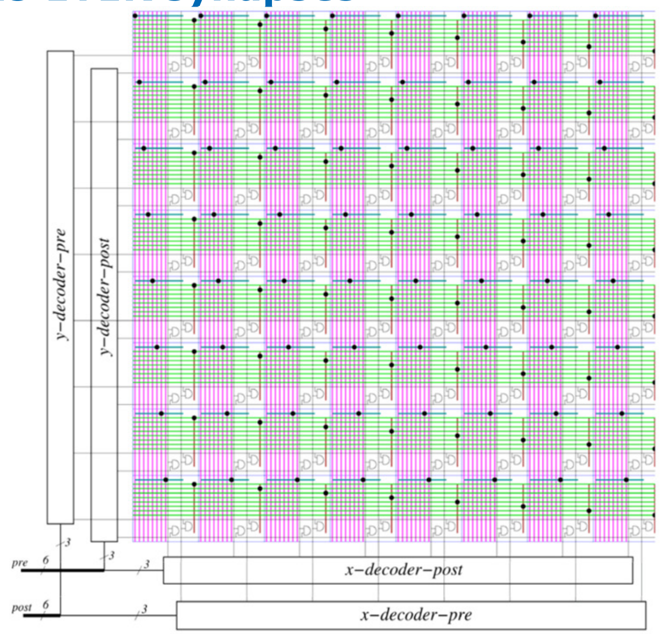
64x64 1T1R CMOL Core with 64-input and 64-output neurons

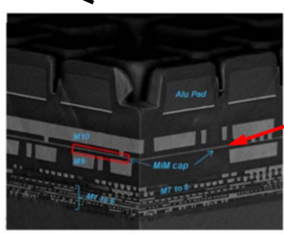
CMOL unit tile



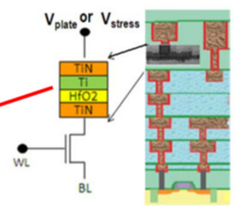
- 130nm CMOS + 1T1R on top (450x415um^2 per core)
- 64x64 (4k) 1T1R analog memristor crossbar
- 64 pre-synaptic, 64 post-synaptic neurons
- 8x8 CMOL tiles arranged in 2D within the chip
- **Each tile** (55x51um^2): 1 pre-neuron, 1 post-neuron, **8x8 1T1R synapses**





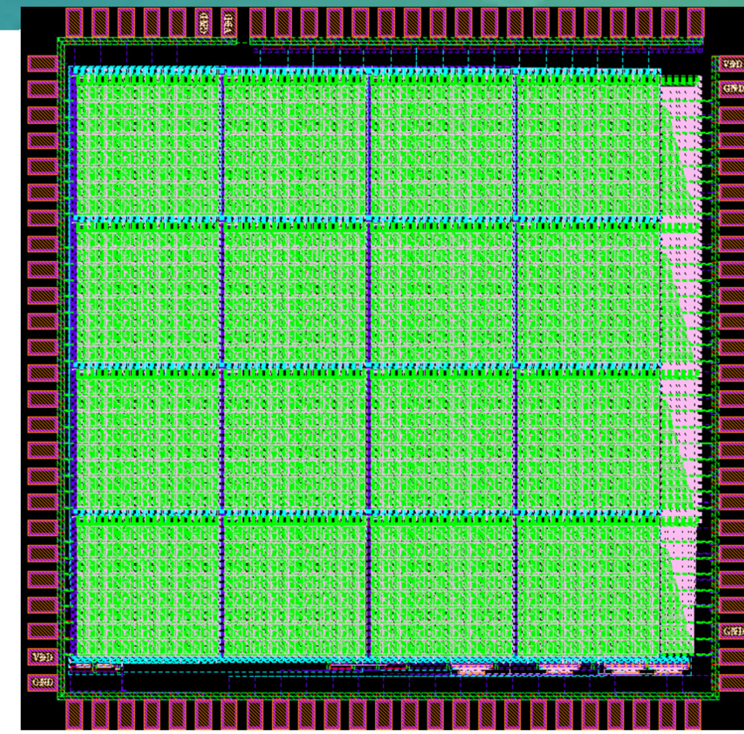


SEM cross section of CMOS 28nm stack including Metal Isulator Metal device



1T1R!!

Under test!



- 16 CMOL tiles, reconfigurable interconnectivity
- Total 64k 1T1R synapses, 5mm^2

Reconfiguration options options:

- 1k input-neurons & 64 output-neurons
- 64 input-neurons & 1k output-neurons
- 256 input-neurons & 256 output-neurons
- 16 independent 64x64 cores

Prospectiva: Si Pitch=100nm
1cm2=10Gsynapses=0,1%cerebro
Equivale a la densidad del cerebro

Conclusions

LOW POWER

Memory-computing colocalization – parallel architectures

Compressive Sensing – sparse coding

Event-based Computing – sparse computation

HIGH SPEED

Asynchronous event-based computing (pseudo-simultaneous input and output at each processing layer)

Parallel computing

Conclusions and Perspectives

OTHER PROS

- Exact spatio-temporal information of the scene is preserved
- The temporal information can also be used to implement biological learning algorithms as Spike-Time-Dependent-Plasticity based in time correlation

CHALLENGES

- Machine Learning Algorithms for conventional sampled information is very well developed
- Explicit time information produces richer information with higher potential but higher complexity (models, learning)

OPORTUNITIES

- Conventional ML systems require high energy and computation resources.
- Low Power Scenarios: autonomous systems (IoT, smart surveillance, remote risky scenarios, …)
- High Speed (drones, car driving, …)
- Real-time biosignal interaction
- Low level sensory processing – high amount of data paralelism
- Memristive-CMOS technology can enable low power in-memory computing systems with on-line learning capability

Institute de Microelectrónica de Sevilla

Thank you !!

terese@imse-cnm.csic.es