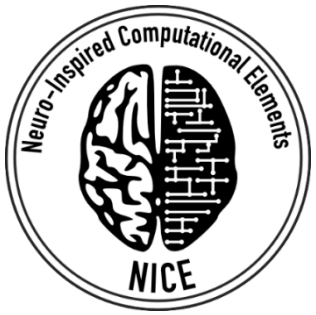



# Bottom-Up and Top-Down Neuromorphic Processor Design:

## *Unveiling Roads to Embedded Cognition*

Charlotte Frenkel

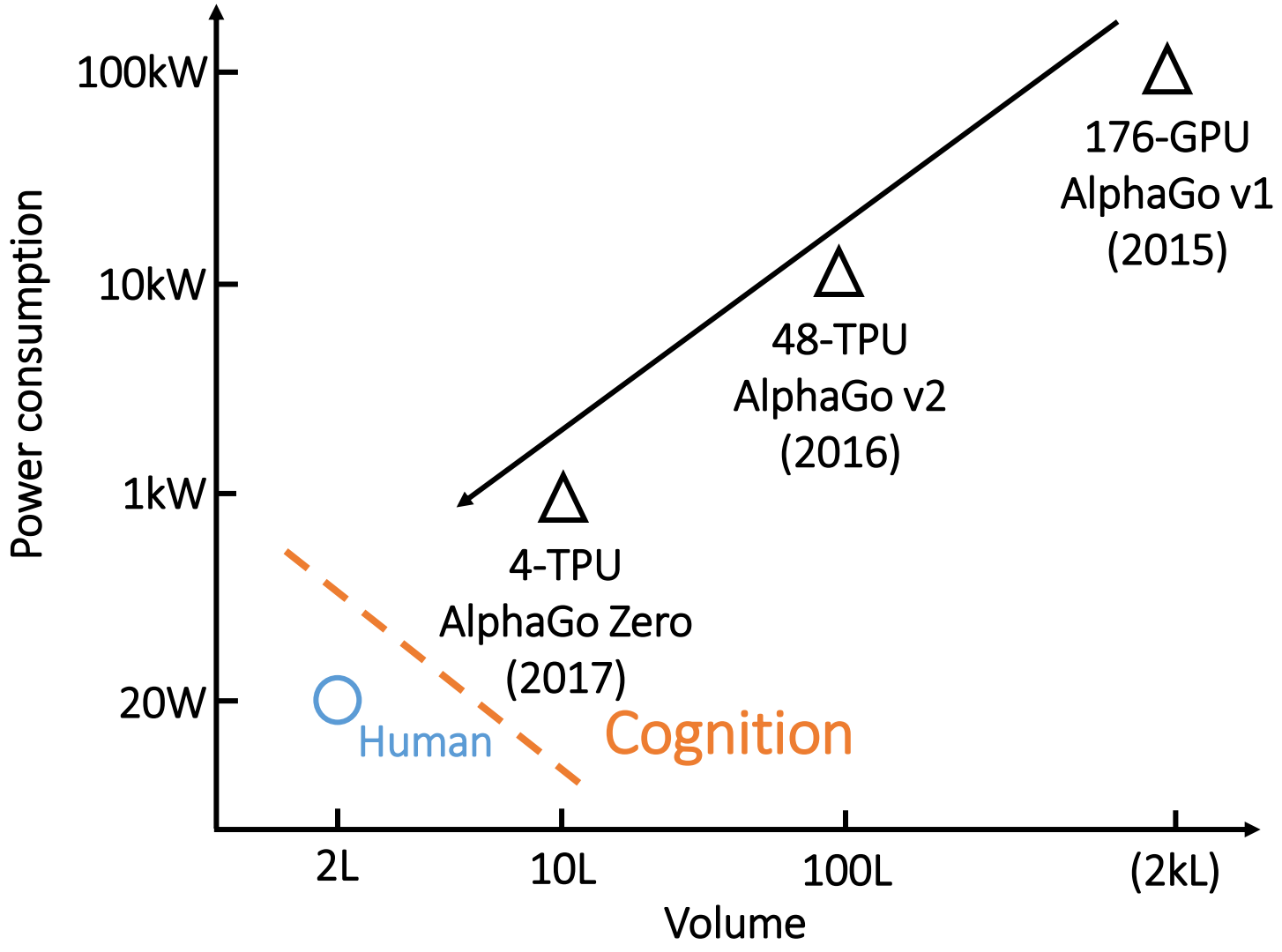
Institute of Neuroinformatics, UZH and ETH Zürich, Switzerland
charlotte@ini.uzh.ch

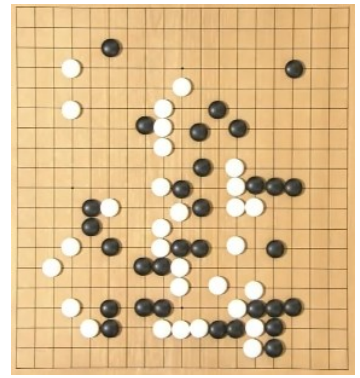
Neuro-Inspired Computational Elements workshop
Virtual, March 16-19, 2021

UCLouvain
fnrs LA LIBERTÉ DE CHERCHER
MEXT www.enmieux.be
FEDER UNION EUROPEENNE
Wallonie
LE FONDS EUROPÉEN DE DÉVELOPPEMENT RÉGIONAL ET LA WALLONIE INVESTISSENT DANS VOTRE AVENIR
University of Zurich[UZH]
ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

# Neuromorphic Engineering – Why?

*Efficiency of artificial intelligence vs. natural intelligence?*

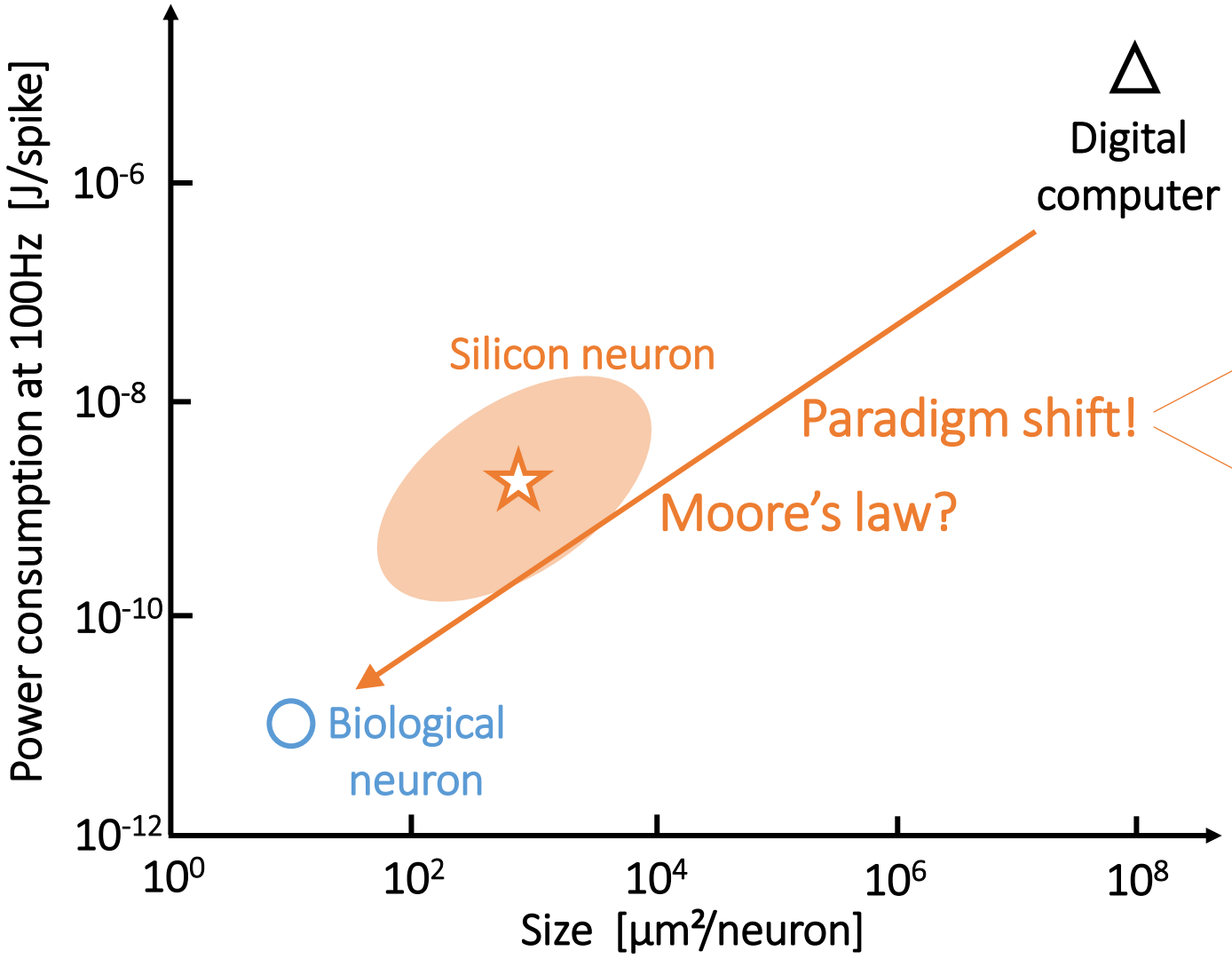


[Silver & Hassabis, *https://deepmind.com/blog/article/alphago-zero-starting-scratch*, 2017]

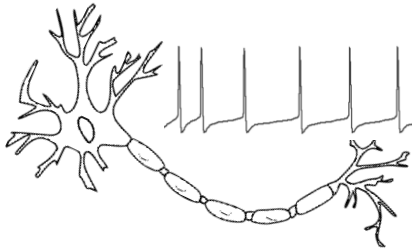
# Neuromorphic Engineering – Why?

*Efficiency of bio-inspired neuromorphic computing?*




Data representation: sparse, event-driven spike trains

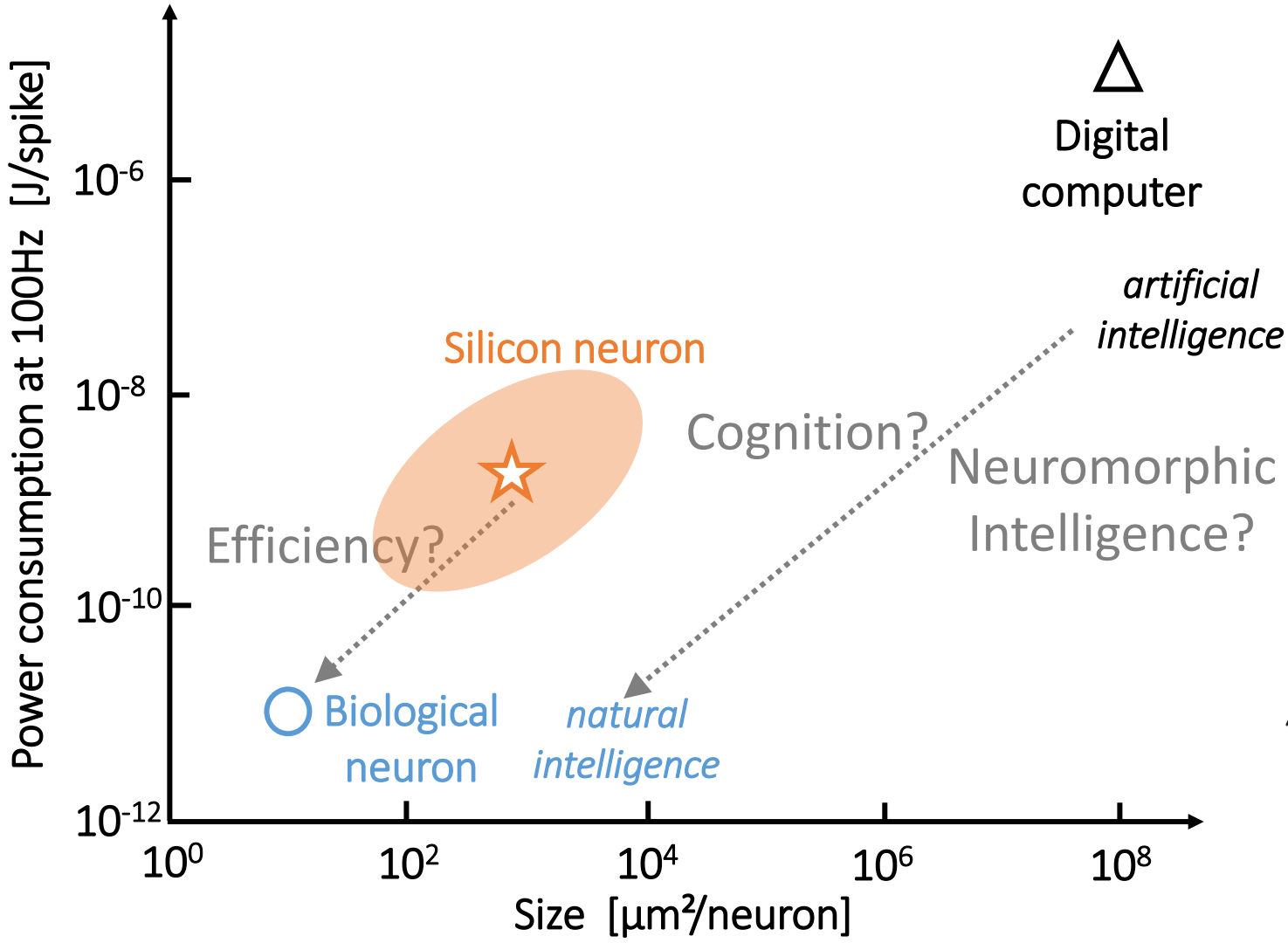
Architecture: distributed processing with co-located neurons and synapses

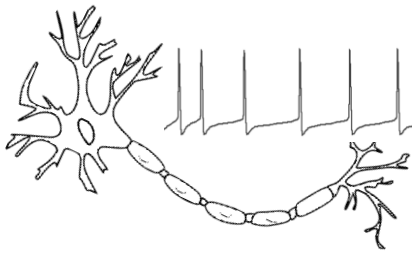


[Poon & Zhou, *Front. Neurosci.*, 2011]

# Neuromorphic Engineering – How?

*A design strategy toward efficiency and cognition?*



[Poon & Zhou, *Front. Neurosci.*, 2011]

# Neuromorphic Engineering – How?

*A design strategy toward efficiency and cognition?*

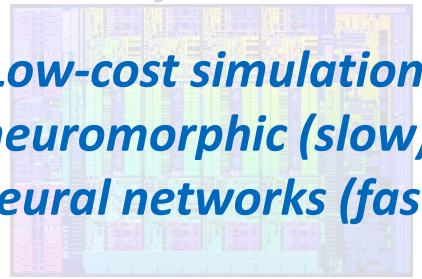
## Subthreshold analog (mixed-signal)

**Biological-time brain emulation and basic research**



ROLLS (ETHZ) DYNAPs (ETHZ) NeuroGrid (Stanford)

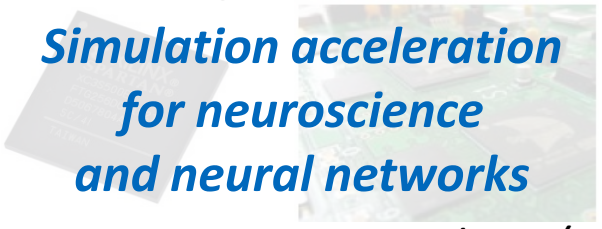
## Software



**Low-cost simulation: neuromorphic (slow), neural networks (fast)**

CPU / GPU

## Dedicated/distributed sim.

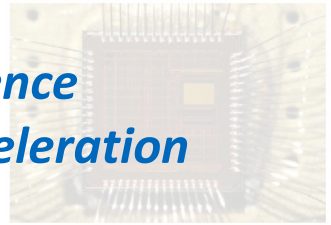


**Simulation acceleration for neuroscience and neural networks**

FPGA SpiNNaker 1/2 (Manchester, TUD)

## Above-threshold analog (mixed-signal)

**Neuroscience simulation acceleration**



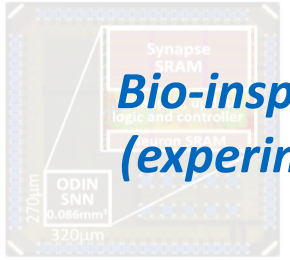
BrainScaleS 1/2 (Heidelberg)

## Large-scale full-custom digital designs

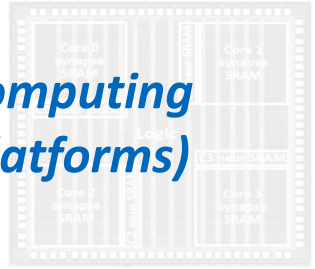
**Cognitive computing**

TrueNorth (IBM) Loihi (Intel)

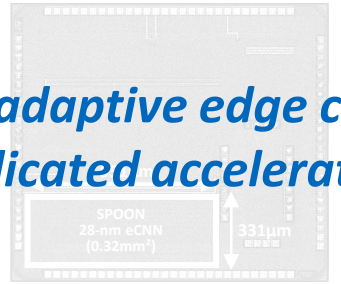
## Small-scale full-custom digital designs



**Bio-inspired edge computing (experimentation platforms)**



ODIN (UCLouvain) MorphIC (UCLouvain)



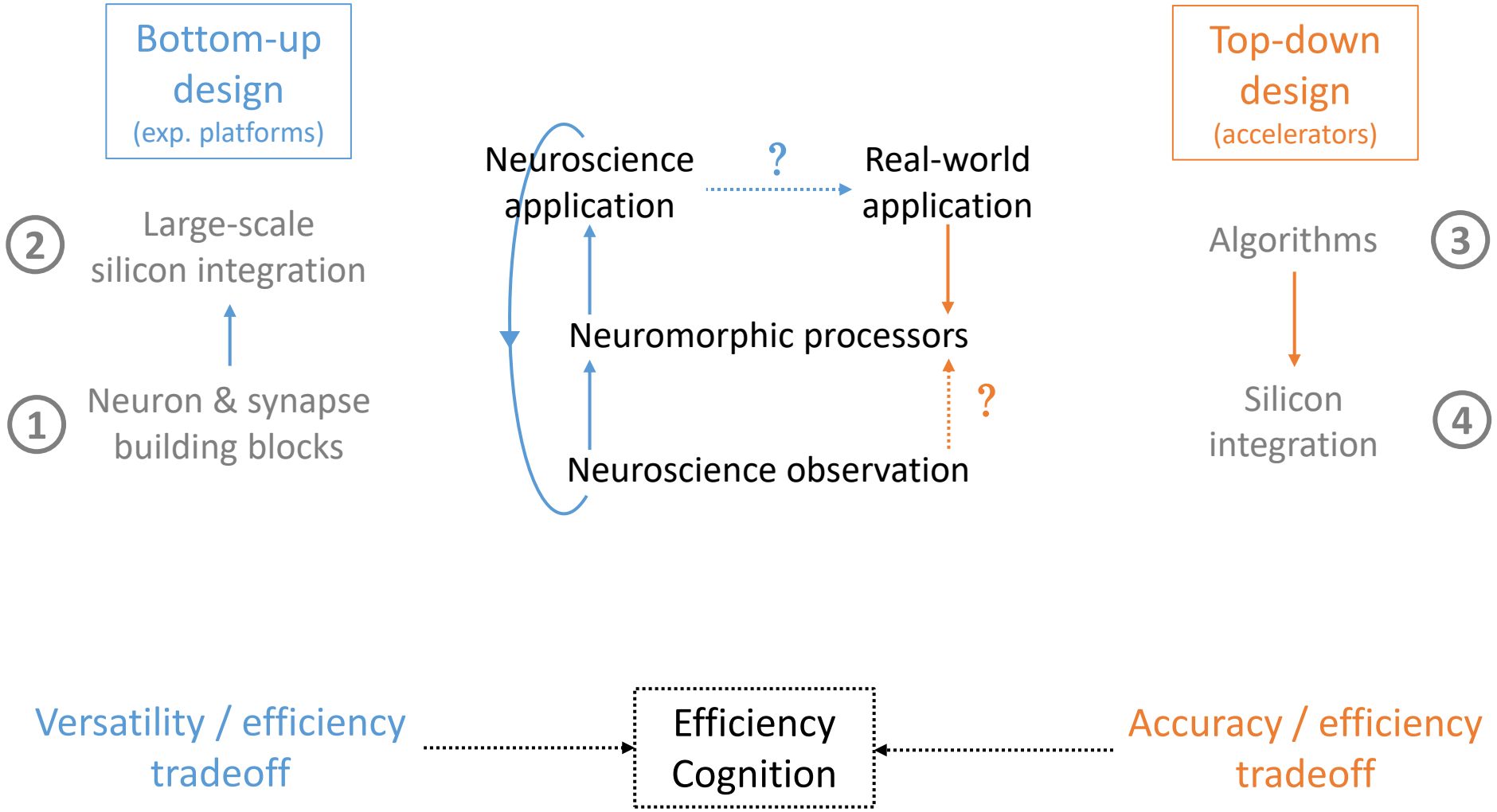
**Low-cost adaptive edge computing (dedicated accelerators)**

SPOON (UCLouvain)

*See also:*
*[Seo, CICC'11]*
*[Knag, JSSC'15]*
*[Park, ISSCC'19]*

# Neuromorphic Engineering – How?

*Unveiling roads to embedded cognition*



# Outline

Part I – Bottom-up neuromorphic design

- Building blocks
- Integration

Part II – Top-down neuromorphic design

- Algorithms
- Integration

Conclusion and perspectives

# Outline

Part I – Bottom-up neuromorphic design

- Building blocks

  Neurons and synapses as adaptive processing and memory elements

  [Frenkel, *ISCAS*, 2017]
  [Frenkel, *BioCAS*, 2017]

- Integration

Part II – Top-down neuromorphic design

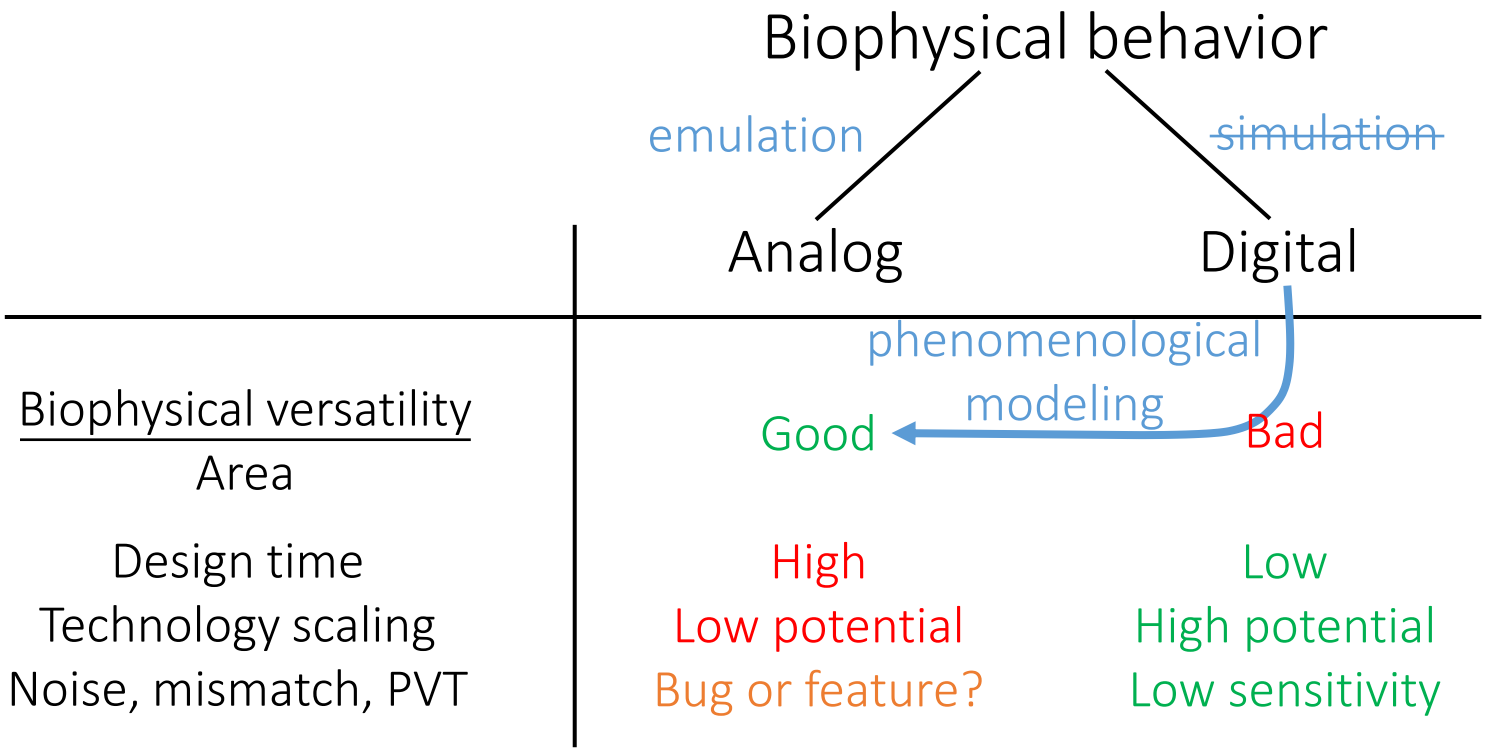
- Algorithms
- Integration

Conclusion and perspectives

# Design strategy

*Analog or digital?*

→ *perspectives*




How can we make the best of both worlds?

# Design strategy

*What should we aim for and phenomenologically implement?*

## Neurons

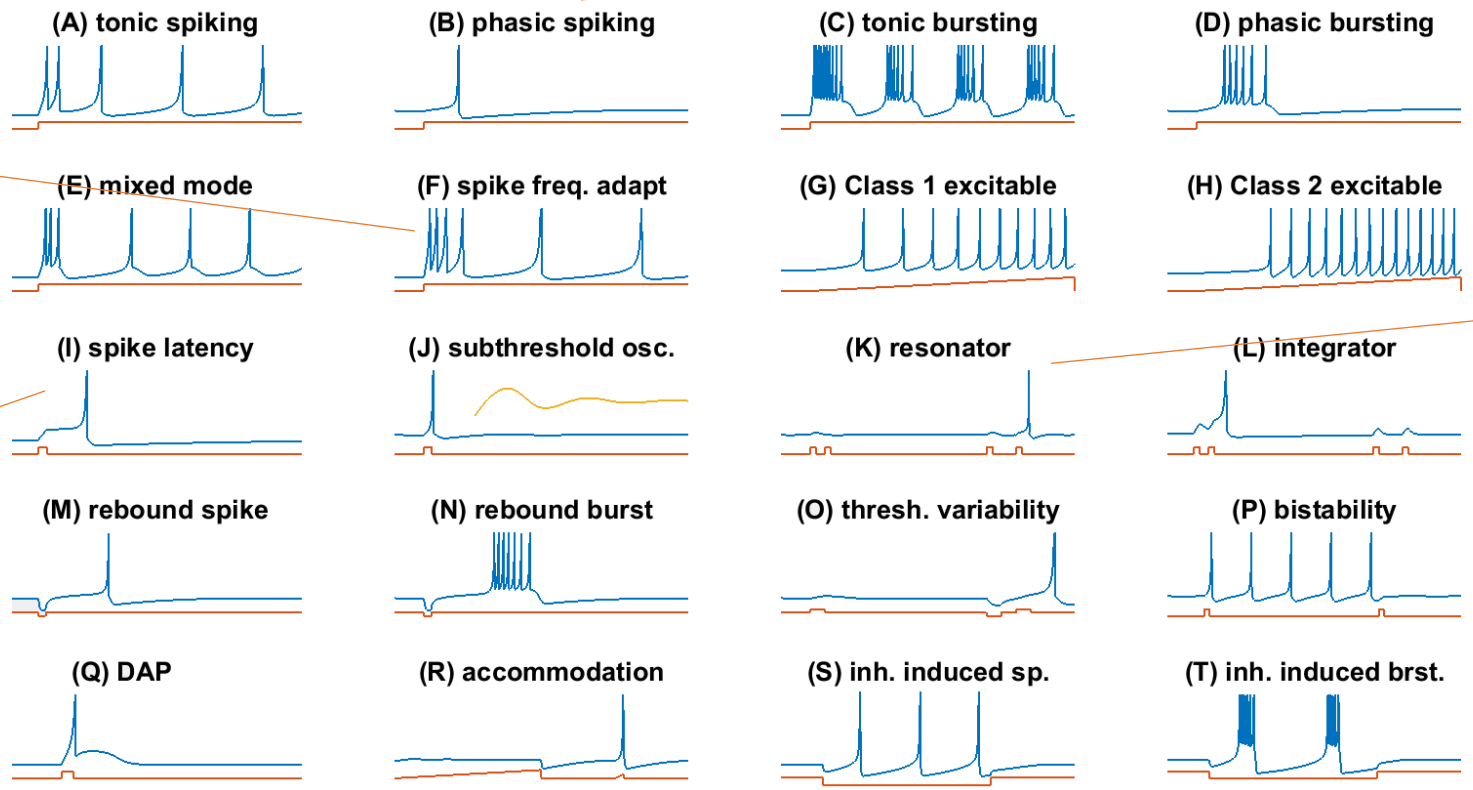
- 20 Izhikevich behaviors of cortical spiking neurons

Useful for time-to-first-spike encodings

Introduce competition for unsupervised learning in winner-take-all networks [Kreiser, BioCAS'17]

Discrimination of specific frequencies

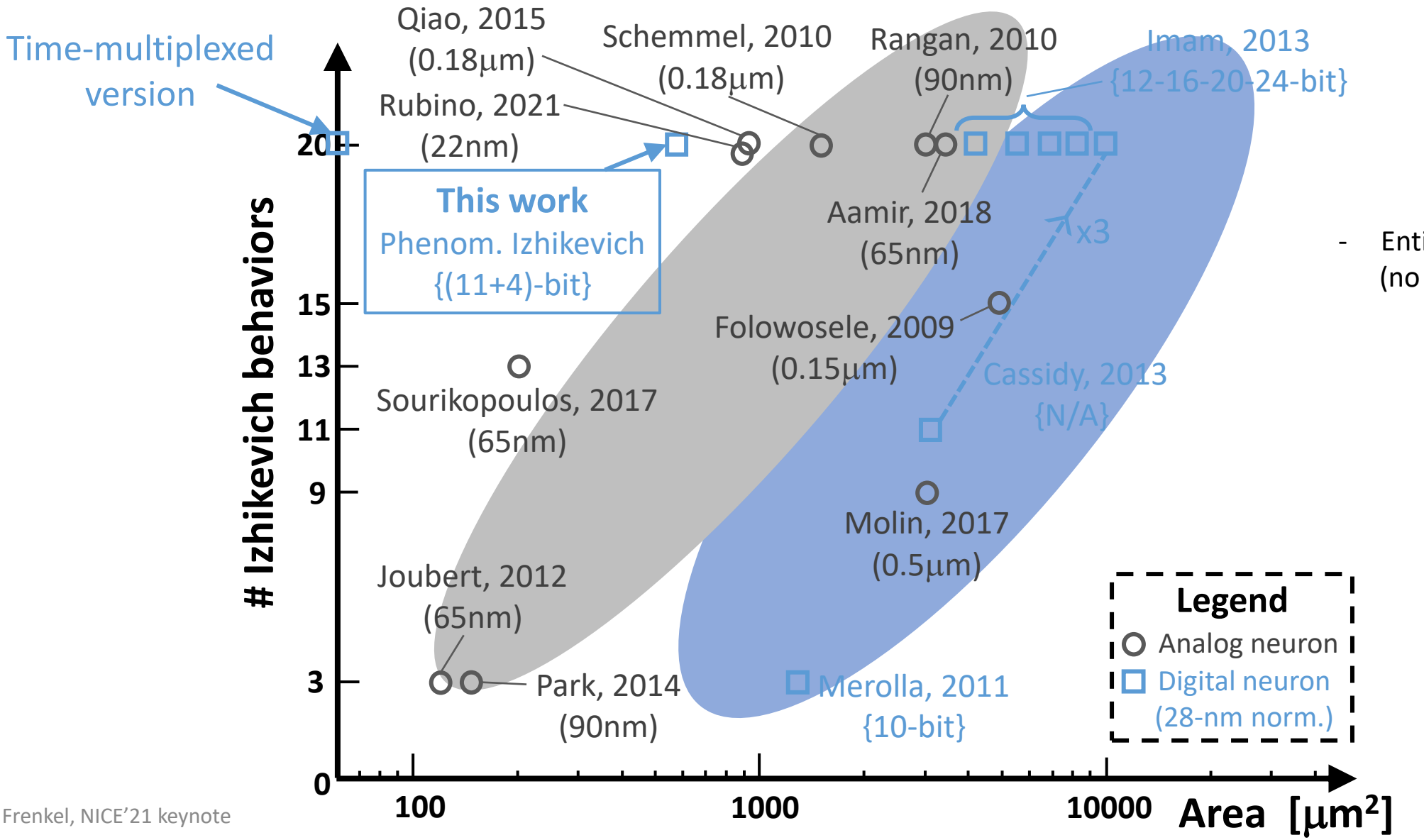
Stereo sound source localization [Schoepe, BioCAS'19]




[Izhikevich, IEEE Trans. NN, 2004]

# Proposed phenomenological digital neuron

*Tackling the versatility/efficiency tradeoff*




**Key features:**

- Entirely event-driven (no time-stepped integration)

# Design strategy

*What should we aim for and phenomenologically implement?*

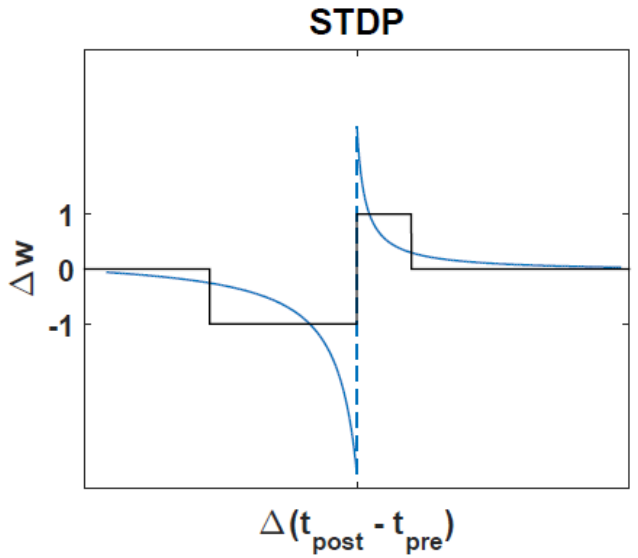
→ *perspectives*

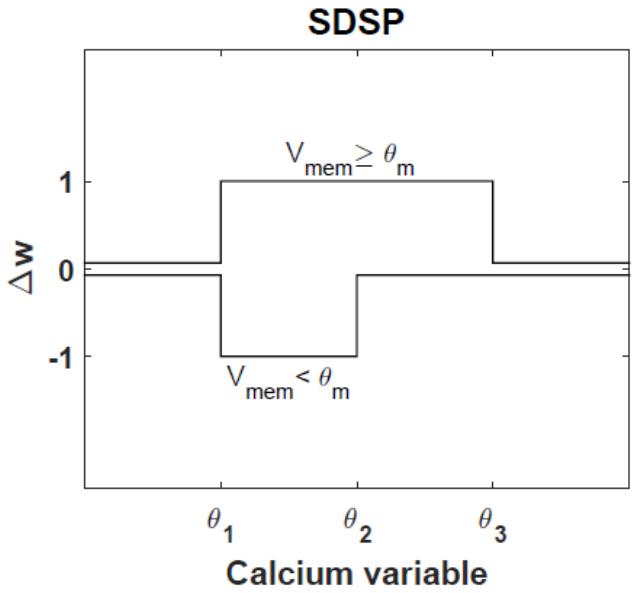
## Neurons

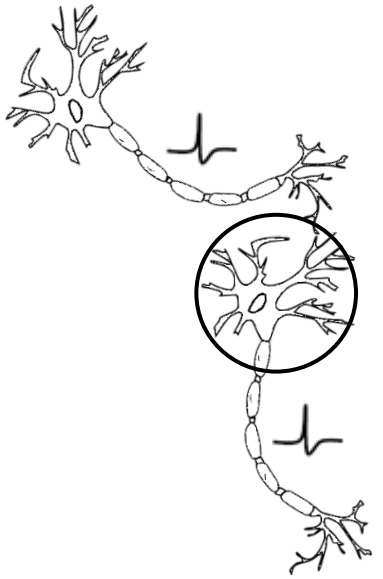
- 20 Izhikevich behaviors of cortical spiking neurons

## Synapses

- Spike-based online learning



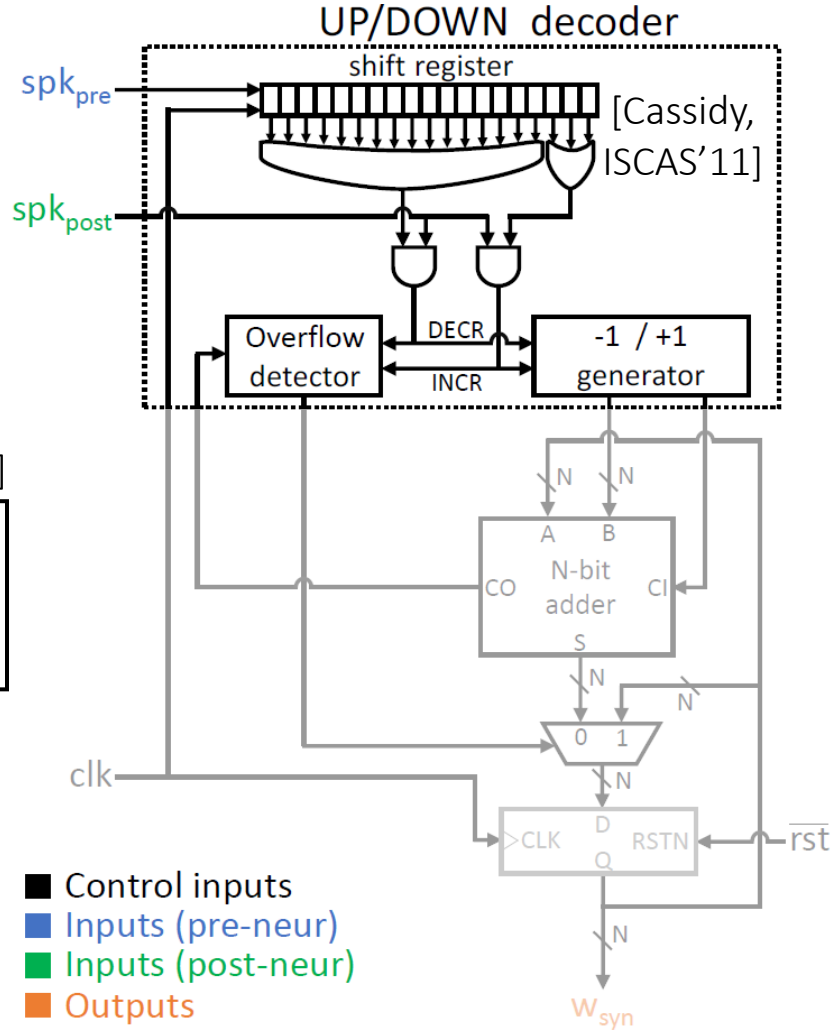
# Proposed digital synapse

*Tackling the versatility/efficiency tradeoff*

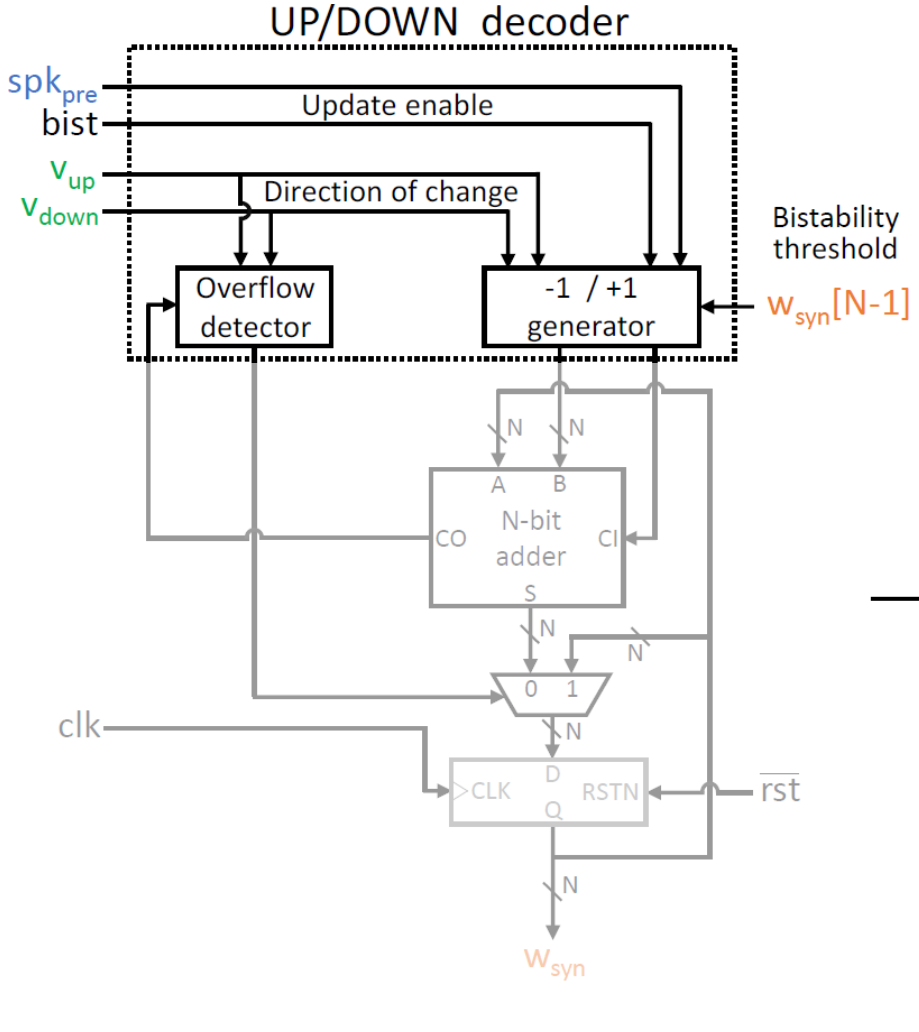
## Key challenge – Fan-in = 100-10000 synapses/neuron

[Seo, CICC'11]

Custom dual-port SRAM




STDP




SDSP

Foundry single-port SRAM

Record density

# Outline

Part I – Bottom-up neuromorphic design

- Building blocks
- Integration

Proposed neuromorphic experimentation platforms

[Frenkel, *Trans. BioCAS*, 2019a]
[Frenkel, *Trans. BioCAS*, 2019b]

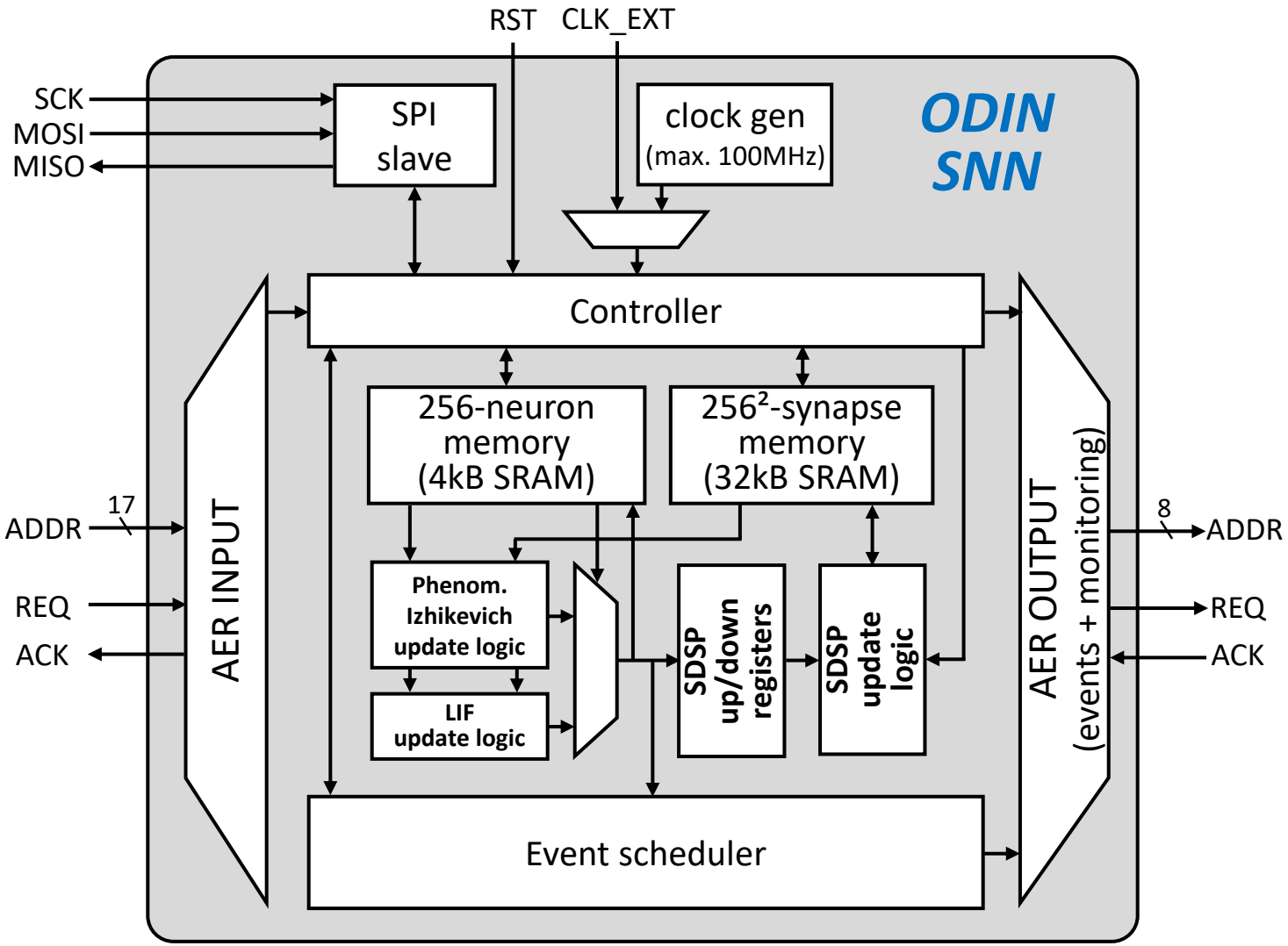
Part II – Top-down neuromorphic design

- Algorithms
- Integration

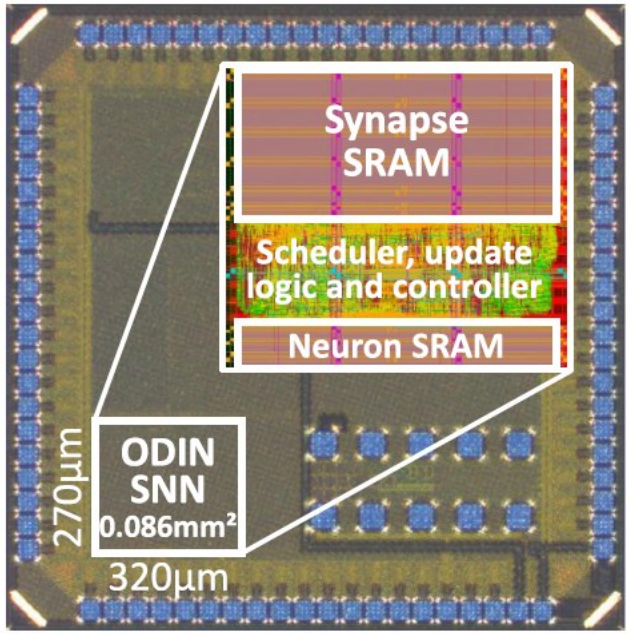
Conclusion and perspectives

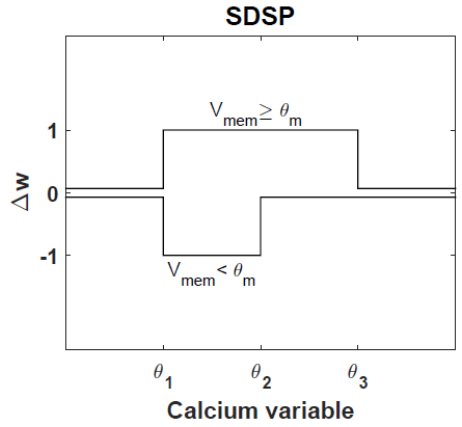
# Architecture of ODIN

ODIN – A 256-neuron 64k-synapse Online-learning Digital Neurosynaptic core



# ODIN – Chip microphotograph and specifications

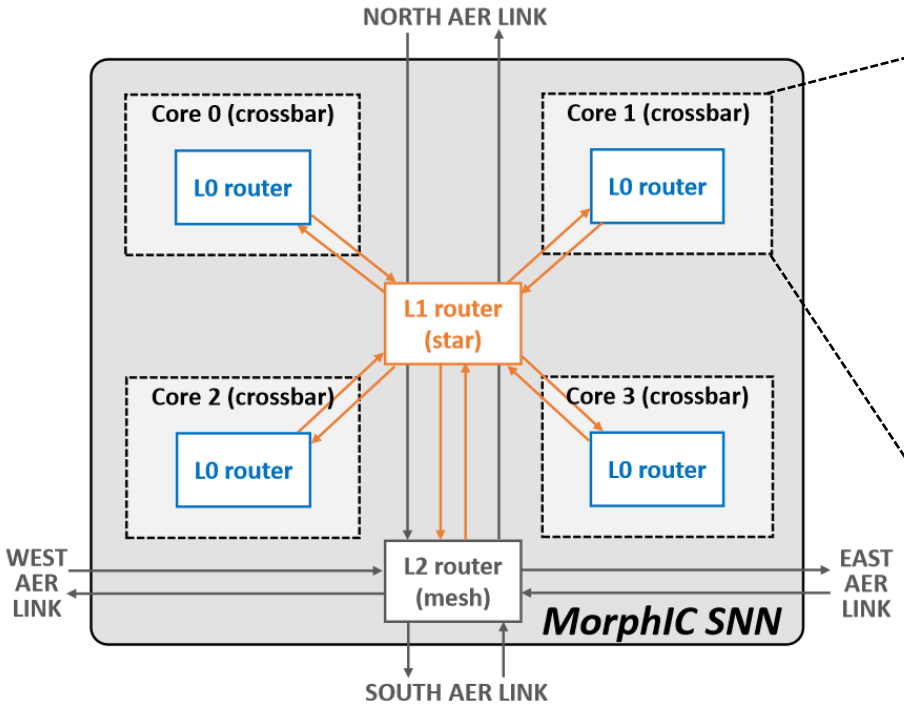


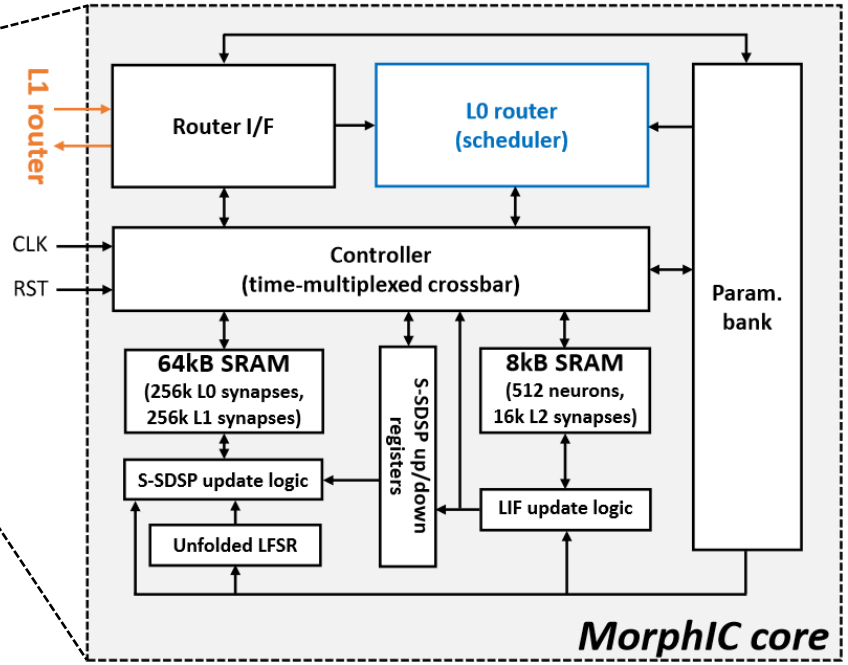

| | |
|---|---|
| Technology | 28nm FDSOI |
| Implementation | Digital |
| Area | 0.086mm² |
| # neurons | 256 |
| # synapses | 64k |
| # Izhikevich behav. | 20 |
| Online learning | SDSP, (3+1)-bit weight |
| Time constant | Biological to accelerated |
| Supply voltage | 0.55V – 1.0V |
| Leakage power ($P_{leak}$) | 27.3µW @0.55V |
| Idle power ($P_{idle}$) | 1.78µW/MHz @0.55V |
| Incr. energy/SOP ($E_{SOP}$) | 8.43pJ @0.55V |
| Global energy/SOP ($E_{tot.SOP}$) | >12.7pJ @0.55V |
| Routing flexibility/efficiency | ☹ (AER) |
| Fan-in | 256 |
| Fan-out | 256 |

# Architecture of MorphIC

## Chip-level architecture




## Core architecture

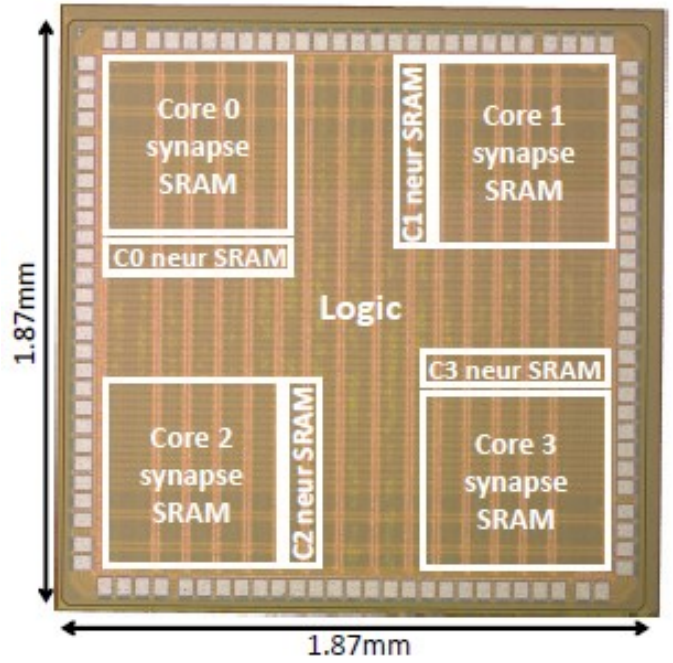



| | |
|---|---|
| Neurons/core | 512 |
| Synapses/core | 528k |

| | |
|---|---|
| Fan-in | 1k |
| Fan-out | 2k |

Stochastic SDSP (S-SDSP) on binary synapses

# MorphIC – Chip microphotograph and specifications




| | |
|---|---|
| Technology | 65nm LP CMOS |
| Implementation | Digital |
| Area | 3.5mm² (incl. pads)<br>2.86mm² (excl. pads) |
| Number of cores | 4 |
| Total # neurons (type) | 2048 (LIF) |
| Total # synapses (hier.) | 1M (L0), 1M (L1), 64k (L2) |
| Fan-in (hier.) | 512 (L0), 512 (L1), 32 (L2) |
| Fan-out (hier.) | 512 (L0), 3x512 (L1), 4 (L2) |
| Online learning | Stochastic SDSP, 1-bit weight |
| Time constant | Biological to accelerated |
| Supply voltage | 0.8V – 1.2V |
| Max. clock frequency | 55MHz (0.8V) – 210MHz (1.2V) |
| Leakage power ($P_{leak}$) | 45µW @0.8V |
| Idle power ($P_{idle}$) | 41.3µW/MHz @0.8V |
| Energy/SOP ($E_{SOP}$) | 30pJ @0.8V |

# Comparison with SoA experimentation platforms

| Author | Schemmel [30] | Benjamin [32] | Qiao [27] | Moradi [29] | Park [26] | Mayr [28] | Painkras [31] | Seo [25] | Akopyan [33] | Davies [34] | **Frenkel** | **Frenkel** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Publication | ISCAS, 2010 | PIEEE, 2014 | Front. NS, 2015 | TBioCAS, 2017 | BioCAS, 2014 | TBCAS, 2016 | JSSC, 2013 | CICC, 2011 | TCAD, 2015 | IEEE Micro, 2018 | TBCAS, 2019a | TBCAS, 2019b |
| Chip name | HICANN | Neurogrid | ROLLS | DYNAPs | IFAT | – | SpiNNaker | – | TrueNorth | Loihi | **ODIN** | **MorphIC** |
| Implementation | Mixed-signal | Mixed-signal | Mixed-signal | Mixed-signal | Mixed-signal | Mixed-signal | Digital | Digital | Digital | Digital | Digital | Digital |
| Technology | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 90nm | 28nm | 0.13$\mu$m | 45nm SOI | 28nm | 14nm FinFET | 28nm FDSOI | 65nm LP |
| # cores$^{\diamond}$ | 1 | 16 | 1 | 4 | 32 | 1 | 18 | 1 | 4096 | 128 | 1 | 4 |
| Neurosynaptic core area [mm$^2$] | 49 | 168 | 51.4 | 7.5 | 0.31 | 0.36 | 3.75 | 0.8 | 0.095 | 0.4 | 0.086 | 0.715 |
| # Izhikevich behaviors$^{\dagger}$ | (20) | N/A | (20) | (20) | 3 | 3 | Programmable | 3 | 11 (3 neur: 20) | (6) | 20 | 3 |
| # neurons per core | 512 | 64k | 256 | 256 | 2k | 64 | max. 1000$^{\circ}$ | 256 | 256 | max. 1024 | 256 | 512 |
| Synaptic weight storage | 4-bit (SRAM) | Off-chip | Capacitor | 12-bit (CAM) | Off-chip | 4-bit (SRAM) | Off-chip | 1-bit (SRAM) | 1-bit (SRAM) | 1- to 9-bit (SRAM) | (3+1)-bit (SRAM) | 1-bit (SRAM) |
| Embedded online learning | STDP | No | SDSP | No | No | SDSP | Programmable | S-STDP | No | Programmable | SDSP | S-SDSP |
| # synapses per core | 112k | – | 128k | 16k | – | 8k | – | 64k | 64k | 1M to 114k (1-9 bits) | 64k | 528k |
| Time constant | Accelerated | Biological | Biological | Biological | Biological | Bio. to accel. | Bio. to accel. | Biological | Biological | N/A | Bio. to accel. | Bio. to accel. |
| Flexibility routing | Medium | Medium | Low | Medium | Medium | Low | High | Low | Medium | High | Low | Medium |
| Flexibility learning | Low | – | Low | Low | – | Low | – | Low | – | High | Low | Low |
| Neuron core density [neur/mm$^2$]$^{*}$ raw | 10.5 | 390 | 5 | 34 | 6.5k | 178 | max. 267$^{\circ}$ | 320 | 2.6k | max. 2.5k | 3.0k | 716 |
| Neuron core density [neur/mm$^2$]$^{*}$ norm. | – | – | – | – | – | – | max. 5.8k | 826 | 2.6k | max. 1k | 3.0k | 3.9k |
| Synapse core density [syn/mm$^2$]$^{*}$ raw | 2.3k | – | 2.5k | 2.1k | – | 22.2k | – | 80k | 674k | 2.5M to 282k | 741k | 738k |
| Synapse core density [syn/mm$^2$]$^{*}$ norm. | – | | – | – | | – | | 207k | 674k | 1M to 113k | 741k | 4M |
| Supply voltage | 1.8V | 3.0V | 1.8V | 1.3V-1.8V | 1.2V | 0.75V, 1.0V | 1.2V | 0.53V-1.0V | 0.7V-1.05V | 0.5V-1.25V | 0.55V-1.0V | 0.8V-1.2V |
| Energy per SOP$^{\ddagger}$ raw | N/A | (941pJ)$^{\blacktriangle}$ | >77fJ$^{\triangle}$ | 134fJ$^{\triangle}$/30pJ$^{\blacktriangle}$ (1.3V) | 22pJ$^{\blacktriangle}$ | >850pJ$^{\blacktriangle}$ | >11.3nJ$^{\triangle}$/26.6nJ$^{\blacktriangle}$ | N/A | 26pJ$^{\blacktriangle}$ (0.775V) | >23.6pJ$^{\triangle}$ (0.75V) | 8.4pJ$^{\triangle}$/12.7pJ$^{\blacktriangle}$ (0.55V) | 30pJ$^{\triangle}$/51pJ$^{\blacktriangle}$ (0.8V) |
| Energy per SOP$^{\ddagger}$ norm. | | – | – | – | – | – | >2.4nJ$^{\triangle}$/5.7nJ$^{\blacktriangle}$ | | 26pJ$^{\blacktriangle}$ | (66.1pJ$^{\triangle}$) | 8.4pJ$^{\triangle}$/12.7pJ$^{\blacktriangle}$ | 12.9pJ$^{\triangle}$/22pJ$^{\blacktriangle}$ |

$^{\diamond}$ When chips are composed of several neurosynaptic cores, we report the density data associated to a single core. Care should be taken that, depending on the core definition in the different chips, routing resources might be included (all single-core designs, IFAT, TrueNorth, Loihi and MorphIC) or excluded (Neurogrid, DYNAPs and SpiNNaker). As opposed to the other reported designs, we consider the full Neurogrid system, which is composed of 16 NeuroCore chips, each one considered as a core; routing resources are off-chip. For DYNAPs and SpiNNaker, sharing routing overhead among cores would lead to 28-% and 37-% density penalties compared to the reported results, respectively. The HICANN chip can be considered as a core of the BrainScaleS wafer-scale system. Pad area is excluded from all reported designs.

$^{\dagger}$ By its similarity with the Izhikevich neuron model, the AdExp neuron model is believed to reach the 20 Izhikevich behaviors [76], but it has not been demonstrated in HICANN, ROLLS and DYNAPs. The neuron model of TrueNorth can reach 11 behaviors per neuron and 20 by combining three neurons together [85]. The neuron model of Loihi is based on a LIF model to which threshold adaptation is added: the neuron should therefore reach 6 Izhikevich behaviors, although it has not been demonstrated.

$^{\circ}$ Experiment 1 reported in Table III from [31] is considered as a best-case neuron density: 1000 simple LIF neuron models are implemented per core, each firing at a low frequency.

$^{*}$ Neuron (resp. synapse) core densities are computed by dividing the number of neurons (resp. synapses) per neurosynaptic core by the neurosynaptic core area. Regarding the synapse core density, Neurogrid, IFAT and SpiNNaker use an off-chip memory to store synaptic data. As the synapse core density cannot be extracted when off-chip resources are involved, no synapse core density values are reported for these chips. Values normalized to a 28-nm CMOS technology node are provided for digital designs using the node factor, at the exception of the 14-nm FinFET node of Loihi for which Intel data from [120] has been used.

$^{\ddagger}$ The synaptic operation energy measurements reported for the different chips do not follow a standardized measurement process. There are two main categories for energy measurements in neuromorphic chips. On the one hand, incremental values (denoted with $^{\triangle}$) describe the amount of energy paid per each additional SOP computation, they are measured by subtracting the leakage and idle power consumption of the chip, as in Eq. (2.2), although the exact power contributions taken into account in the SOP energy vary across chips. On the other hand, global values (denoted with $^{\blacktriangle}$) are obtained by dividing the total chip power consumption by the SOP rate, as in Eq. (2.3). Values normalized to a 28-nm CMOS technology node are provided for digital designs using the node factor, including for the 14-nm FinFET node of Loihi in the absence of reliable data for power normalization in [120]. The conditions under which all of these measurements have been done can be found hereafter. For Neurogrid, a SOP energy of 941pJ is reported for a network of 16 Neurocore chips (1M neurons, 8B synapses, 413k spikes/s): it is a board-level measurement, no chip-level measurement is provided [32]. For ROLLS, the measured SOP energy of 77fJ is reported in [163], it accounts for a point-to-point synaptic input event and includes the contribution of weight adaptation and digital-to-analog conversion, it represents a lower bound as it does not account for synaptic event broadcasting. For DYNAPs, the measured SOP energy of 134fJ at 1.3V is also reported in [163] while the global SOP energy of 30pJ can be estimated from [29] using the measured 800-$\mu$W power consumption with all 1k neurons spiking at 100Hz with 25% connectivity (26.2MSOP/s), excluding the synaptic input currents. For IFAT, the SOP energy of 22pJ is extracted by measuring the chip power consumption when operated at the peak rate of 73M synaptic events/s [26]. In the chip of Mayr *et al.*, the SOP energy of 850pJ represents a lower bound extracted from the chip power consumption, estimated by considering the synaptic weights at half their dynamic at maximum operating frequency [28]. For SpiNNaker, an incremental SOP energy of 11.3nJ is measured in [164], a global SOP energy of 26.6nJ at the maximum SOP rate of 16.56MSOP/s can be estimated by taking into account the leakage and idle power; both values represent a lower bound as the energy cost of neuron updates is not included. For TrueNorth, the measured SOP energy of 26pJ at 0.775V is reported in [165], it is extracted by measuring the chip power consumption when all neurons fire at 20Hz with 128 active synapses. For Loihi, a minimum SOP energy of 23.6pJ at 0.75V is extracted from pre-silicon SDF and SPICE simulations, in accordance with early post-silicon characterization [34]; it represents a lower bound as it includes only the contribution of the synaptic operation, without taking into account the cost of neuron update and learning engine update. For ODIN and MorphIC, the detailed measurement process is described in Sections 2.2.2 and 2.3.2, respectively.

# Comparison with SoA experimentation platforms

| | | Mixed-signal | | | | Digital | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Author | | Schemmel | Benjamin | Qiao | Moradi | Painkras | Akopyan | Davies | **Frenkel** | **Frenkel** |
| Publication | | ISCAS, 2010 | PIEEE, 2014 | Front. NS, 2015 | TBioCAS, 2017 | JSSC, 2013 | TCAD, 2015 | IEEE Micro, 2018 | TBCAS, 2019a | TBCAS, 2019b |
| Chip name | | HICANN | Neurogrid | ROLLS | DYNAPs | SpiNNaker | TrueNorth | Loihi | **ODIN** | **MorphIC** |
| Implementation | | Mixed-signal | Mixed-signal | Mixed-signal | Mixed-signal | Digital | Digital | Digital | Digital | Digital |
| Technology | | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 0.13$\mu$m | 28nm | 14nm FinFET | 28nm FDSOI | 65nm LP |
| # cores | | 1 | 16 | 1 | 4 | 18 | 4096 | 128 | 1 | 4 |
| Neurosynaptic core area [mm$^2$] | | 49 | 168 | 51.4 | 7.5 | 3.75 | 0.095 | 0.4 | 0.086 | 0.715 |
| # Izhikevich behaviors | | (20) | N/A | (20) | (20) | Programmable | 11 (3 neur: 20) | (6) | 20 | 3 |
| # neurons per core | | 512 | 64k | 256 | 256 | max. 1000 | 256 | max. 1024 | 256 | 512 |
| Synaptic weight storage | | 4-bit (SRAM) | Off-chip | Capacitor | 12-bit (CAM) | Off-chip | 1-bit (SRAM) | 1- to 9-bit (SRAM) | (3+1)-bit (SRAM) | 1-bit (SRAM) |
| Embedded online learning | | STDP | No | SDSP | No | Programmable | No | Programmable | SDSP | S-SDSP |
| # synapses per core | | 112k | – | 128k | 16k | – | 64k | 1M to 114k (1-9 bits) | 64k | 528k |
| Time constant | | Accelerated | Biological | Biological | Biological | Bio. to accel. | Biological | N/A | Bio. to accel. | Bio. to accel. |
| Flexibility | routing | Medium | Medium | Low | Medium | High | Medium | High | Low | Medium |
| | learning | Low | – | Low | Low | – | – | High | Low | Low |
| Neuron core density [neur/mm$^2$] | raw | 10.5 | 390 | 5 | 34 | max. 267 | 2.6k | max. 2.5k | 3.0k | 716 |
| | norm. | – | – | – | – | max. 5.8k | 2.6k | max. 1k | 3.0k | 3.9k |
| Synapse core density [syn/mm$^2$] | raw | 2.3k | – | 2.5k | 2.1k | – | 674k | 2.5M to 282k | 741k | 738k |
| | norm. | – | – | – | – | | 674k | 1M to 113k | 741k | 4M |
| Supply voltage | | 1.8V | 3.0V | 1.8V | 1.3V-1.8V | 1.2V | 0.7V-1.05V | 0.5V-1.25V | 0.55V-1.0V | 0.8V-1.2V |
| Energy per SOP | raw | N/A | (941pJ)$^▲$ | >77fJ$^Δ$ | 134fJ$^Δ$/30pJ$^▲$ (1.3V) | >11.3nJ$^Δ$/26.6nJ$^▲$ | 26pJ$^▲$ (0.775V) | >23.6pJ$^Δ$ (0.75V) | 8.4pJ$^Δ$/12.7pJ$^▲$ (0.55V) | 30pJ$^Δ$/51pJ$^▲$ (0.8V) |
| | norm. | | – | – | – | >2.4nJ$^Δ$/5.7nJ$^▲$ | 26pJ$^▲$ | (66.1pJ$^Δ$) | 8.4pJ$^Δ$/12.7pJ$^▲$ | 12.9pJ$^Δ$/22pJ$^▲$ |

Most direct comparison: IBM TrueNorth core vs. ODIN (same technology node, same number of neurons and synapses per neurosynaptic core, same area).

| | ODIN | TrueNorth | |
|---|---|---|---|
| Synapses | ✔ 4-bit **with** learning | 1-bit **without** learning | ✘ |
| Neurons | ✔ 20 Izh. beh. | 11 Izh. beh. | ✘ |
| Energy/SOP | ✔ 12.7pJ @0.55V | 26pJ @0.775V | ✘ |
| Connectivity | ✘ AER | large-scale mesh | ✔ → MorphIC |

# Comparison with SoA experimentation platforms

| | | Mixed-signal | | | | Digital | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Author | | Schemmel | Benjamin | Qiao | Moradi | Painkras | Akopyan | Davies | **Frenkel** | **Frenkel** |
| Publication | | ISCAS, 2010 | PIEEE, 2014 | Front. NS, 2015 | TBioCAS, 2017 | JSSC, 2013 | TCAD, 2015 | IEEE Micro, 2018 | TBCAS, 2019a | TBCAS, 2019b |
| Chip name | | HICANN | Neurogrid | ROLLS | DYNAPs | SpiNNaker | TrueNorth | Loihi | **ODIN** | **MorphIC** |
| Implementation | | Mixed-signal | Mixed-signal | Mixed-signal | Mixed-signal | Digital | Digital | Digital | Digital | Digital |
| Technology | | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 0.13$\mu$m | 28nm | 14nm FinFET | 28nm FDSOI | 65nm LP |
| # cores | | 1 | 16 | 1 | 4 | 18 | 4096 | 128 | 1 | 4 |
| Neurosynaptic core area [mm$^2$] | | 49 | 168 | 51.4 | 7.5 | 3.75 | 0.095 | 0.4 | 0.086 | 0.715 |
| # Izhikevich behaviors | | (20) | N/A | (20) | (20) | Programmable | 11 (3 neur: 20) | (6) | 20 | 3 |
| # neurons per core | | 512 | 64k | 256 | 256 | max. 1000 | 256 | max. 1024 | 256 | 512 |
| Synaptic weight storage | | 4-bit (SRAM) | Off-chip | Capacitor | 12-bit (CAM) | Off-chip | 1-bit (SRAM) | 1- to 9-bit (SRAM) | (3+1)-bit (SRAM) | 1-bit (SRAM) |
| Embedded online learning | | STDP | No | SDSP | No | Programmable | No | Programmable | SDSP | S-SDSP |
| # synapses per core | | 112k | – | 128k | 16k | – | 64k | 1M to 114k (1-9 bits) | 64k | 528k |
| Time constant | | Accelerated | Biological | Biological | Biological | Bio. to accel. | Biological | N/A | Bio. to accel. | Bio. to accel. |
| Flexibility | routing | Medium | Medium | Low | Medium | High | Medium | High | Low | Medium |
| | learning | Low | – | Low | Low | – | – | High | Low | Low |
| Neuron core density [neur/mm$^2$] | raw | 10.5 | 390 | 5 | 34 | max. 267 | 2.6k | max. 2.5k | 3.0k | 716 |
| | norm. | – | – | – | – | max. 5.8k | 2.6k | max. 1k | 3.0k | 3.9k |
| Synapse core density [syn/mm$^2$] | raw | 2.3k | – | 2.5k | 2.1k | – | 674k | 2.5M to 282k | 741k | 738k |
| | norm. | – | – | – | – | | 674k | 1M to 113k | 741k | 4M |
| Supply voltage | | 1.8V | 3.0V | 1.8V | 1.3V-1.8V | 1.2V | 0.7V-1.05V | 0.5V-1.25V | 0.55V-1.0V | 0.8V-1.2V |
| Energy per SOP | raw | N/A | (941pJ)$^▲$ | >77fJ$^Δ$ | 134fJ$^Δ$/30pJ$^▲$ (1.3V) | >11.3nJ$^Δ$/26.6nJ$^▲$ | 26pJ$^▲$ (0.775V) | >23.6pJ$^Δ$ (0.75V) | 8.4pJ$^Δ$/12.7pJ$^▲$ (0.55V) | 30pJ$^Δ$/51pJ$^▲$ (0.8V) |
| | norm. | | – | – | – | >2.4nJ$^Δ$/5.7nJ$^▲$ | 26pJ$^▲$ | (66.1pJ$^Δ$) | 8.4pJ$^Δ$/12.7pJ$^▲$ | 12.9pJ$^Δ$/22pJ$^▲$ |

## Area

ODIN and MorphIC have the highest neuron and synapse densities among all SNNs with embedded synaptic weight storage

# Comparison with SoA experimentation platforms

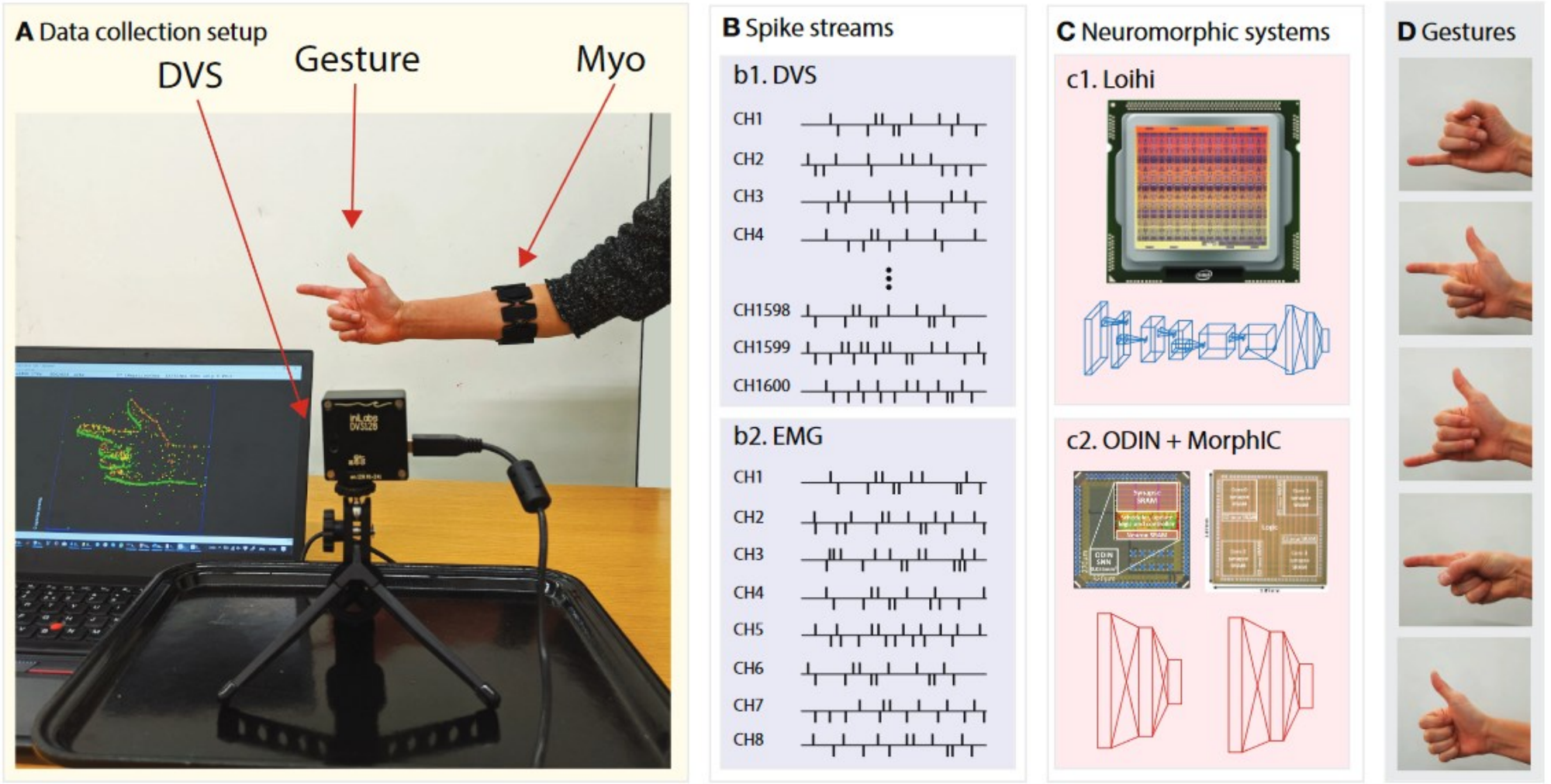
| | | Mixed-signal | | | | Digital | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Author | | Schemmel | Benjamin | Qiao | Moradi | Painkras | Akopyan | Davies | **Frenkel** | **Frenkel** |
| Publication | | ISCAS, 2010 | PIEEE, 2014 | Front. NS, 2015 | TBioCAS, 2017 | JSSC, 2013 | TCAD, 2015 | IEEE Micro, 2018 | TBCAS, 2019a | TBCAS, 2019b |
| Chip name | | HICANN | Neurogrid | ROLLS | DYNAPs | SpiNNaker | TrueNorth | Loihi | **ODIN** | **MorphIC** |
| Implementation | | Mixed-signal | Mixed-signal | Mixed-signal | Mixed-signal | Digital | Digital | Digital | Digital | Digital |
| Technology | | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 0.18$\mu$m | 0.13$\mu$m | 28nm | 14nm FinFET | 28nm FDSOI | 65nm LP |
| # cores | | 1 | 16 | 1 | 4 | 18 | 4096 | 128 | 1 | 4 |
| Neurosynaptic core area [mm$^2$] | | 49 | 168 | 51.4 | 7.5 | 3.75 | 0.095 | 0.4 | 0.086 | 0.715 |
| # Izhikevich behaviors | | (20) | N/A | (20) | (20) | Programmable | 11 (3 neur: 20) | (6) | 20 | 3 |
| # neurons per core | | 512 | 64k | 256 | 256 | max. 1000 | 256 | max. 1024 | 256 | 512 |
| Synaptic weight storage | | 4-bit (SRAM) | Off-chip | Capacitor | 12-bit (CAM) | Off-chip | 1-bit (SRAM) | 1- to 9-bit (SRAM) | (3+1)-bit (SRAM) | 1-bit (SRAM) |
| Embedded online learning | | STDP | No | SDSP | No | Programmable | No | Programmable | SDSP | S-SDSP |
| # synapses per core | | 112k | – | 128k | 16k | – | 64k | 1M to 114k (1-9 bits) | 64k | 528k |
| Time constant | | Accelerated | Biological | Biological | Biological | Bio. to accel. | Biological | N/A | Bio. to accel. | Bio. to accel. |
| Flexibility | routing | Medium | Medium | Low | Medium | High | Medium | High | Low | Medium |
| | learning | Low | – | Low | Low | – | – | High | Low | Low |
| Neuron core density [neur/mm$^2$] | raw | 10.5 | 390 | 5 | 34 | max. 267 | 2.6k | max. 2.5k | 3.0k | 716 |
| | norm. | – | – | – | – | max. 5.8k | 2.6k | max. 1k | 3.0k | 3.9k |
| Synapse core density [syn/mm$^2$] | raw | 2.3k | – | 2.5k | 2.1k | – | 674k | 2.5M to 282k | 741k | 738k |
| | norm. | – | – | – | – | | 674k | 1M to 113k | 741k | 4M |
| Supply voltage | | 1.8V | 3.0V | 1.8V | 1.3V-1.8V | 1.2V | 0.7V-1.05V | 0.5V-1.25V | 0.55V-1.0V | 0.8V-1.2V |
| Energy per SOP | raw | N/A | (941pJ)$^{\blacktriangle}$ | >77fJ$^{\Delta}$ | 134fJ$^{\Delta}$/30pJ$^{\blacktriangle}$ (1.3V) | >11.3nJ$^{\Delta}$/26.6nJ$^{\blacktriangle}$ | 26pJ$^{\blacktriangle}$ (0.775V) | >23.6pJ$^{\Delta}$ (0.75V) | 8.4pJ$^{\Delta}$/12.7pJ$^{\blacktriangle}$ (0.55V) | 30pJ$^{\Delta}$/51pJ$^{\blacktriangle}$ (0.8V) |
| | norm. | | – | – | – | >2.4nJ$^{\Delta}$/5.7nJ$^{\blacktriangle}$ | 26pJ$^{\blacktriangle}$ | (66.1pJ$^{\Delta}$) | 8.4pJ$^{\Delta}$/12.7pJ$^{\blacktriangle}$ | 12.9pJ$^{\Delta}$/22pJ$^{\blacktriangle}$ |

## Power

ODIN has the lowest energy per synaptic event among all digital SNNs,
MorphIC keeps a competitive energy efficiency.
They outperform subthreshold analog SNNs in accelerated time,
but not for biological-time processing.

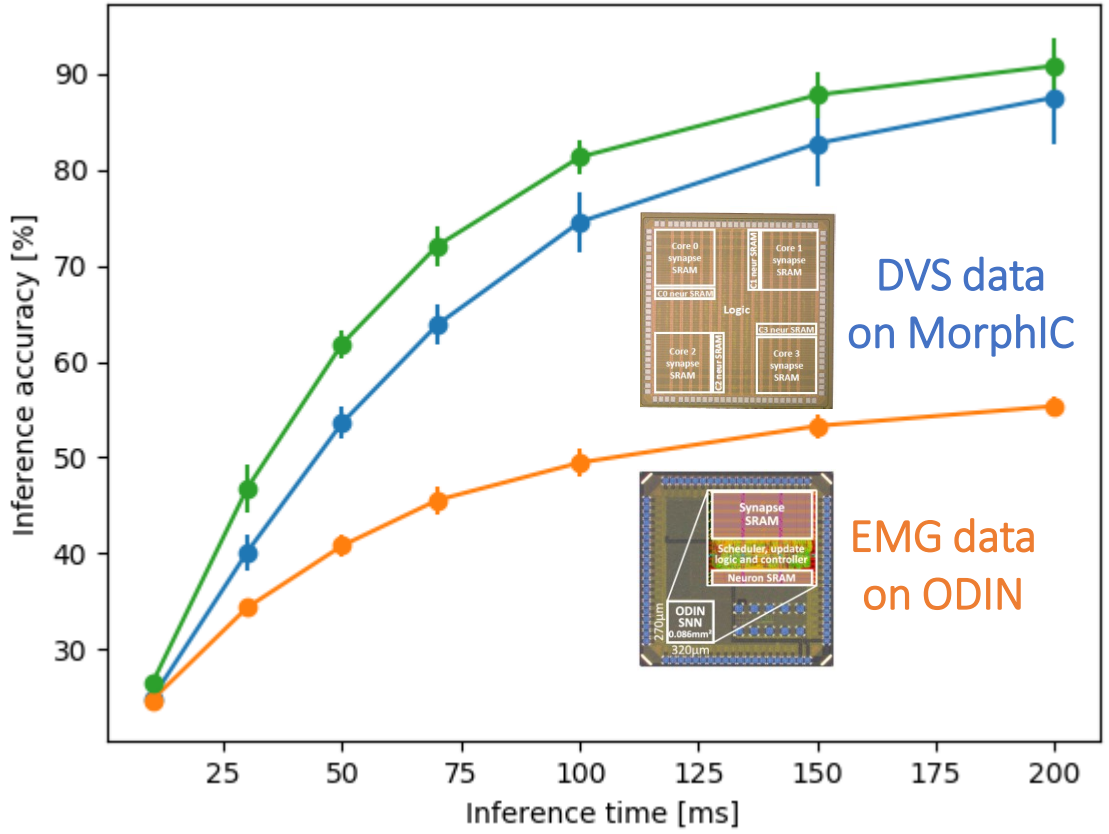
# Results on the spiking EMG/DVS sensor fusion benchmark

[Ceolini, Frenkel, Shrestha et al., *Front. Neurosci.*, 2020]



# Results on the spiking EMG/DVS sensor fusion benchmark

[Ceolini, Frenkel, Shrestha et al., *Front. Neurosci.*, 2020]




Sensor fusion

| | | |
|---|---|---|
| ODIN+MorphIC | 89.4% / | 37.4µJ |
| Loihi | 96% / | 1105µJ |
| Software | 95.4% / | 32100µJ |

Accuracy / Energy tradeoff

Neuromorphic designs are more efficient than GPUs, as would be expected from dedicated hardware. But are they more efficient than conventional accelerators?

→ *perspectives*

*See the ODIN and MorphIC papers for more benchmarking, incl. online- and offline-trained MNIST.*

# Outline

Part I – Bottom-up neuromorphic design

- Building blocks
- Integration

Part II – Top-down neuromorphic design

- Algorithms

  Minimizing the training cost of neural networks for adaptive edge computing

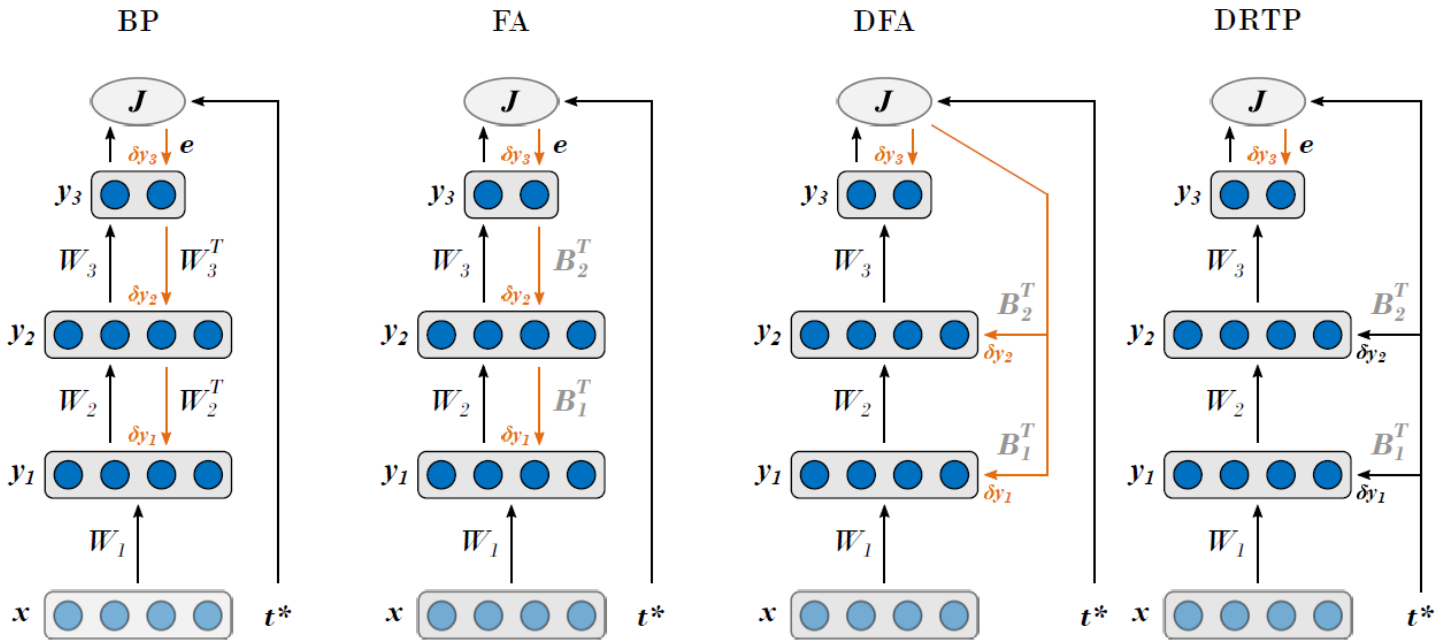
  [Frenkel & Lefebvre, *Front. Neurosci.*, 2021]

- Integration

Conclusion and perspectives

# Learning without feedback

*Releasing the weight transport and update locking of backprop*




| | $\delta y_k$ | $\frac{\partial J}{\partial y_k} = W_{k+1}^T \delta z_{k+1}$ | $B_k^T \delta z_{k+1}$ | $B_k^T e$ | $B_k^T t^*$ |
|---|---|---|---|---|---|
| Weight-transport-free | | × | ✓ | ✓ | ✓ |
| Update-unlocked | | × | × | × | ✓ |

Feedforward local training

↘ Computational and memory cost ↘

# Direct Random Target Projection (DRTP)

*Ideal use cases?*

## Adaptive edge computing

- Very low power and area overheads can be expected for an on-chip implementation.
- Datasets representative of the complexity associated to autonomous smart sensors: MNIST or CIFAR-10.

→ We'll verify these claims *in silico*.

Disclaimer: whether DRTP scales to ImageNET is probably **not** the right question. ☺

## Neuroscience

DRTP could come in line with recent findings in cortical areas that reveal the existence of output-independent target signals in the dendritic instructive pathways of intermediate-layer neurons.

[Magee & Grienberger, *Annual Review of Neuroscience*, 2020]

# Outline

Part I – Bottom-up neuromorphic design

- Building blocks
- Integration

Part II – Top-down neuromorphic design

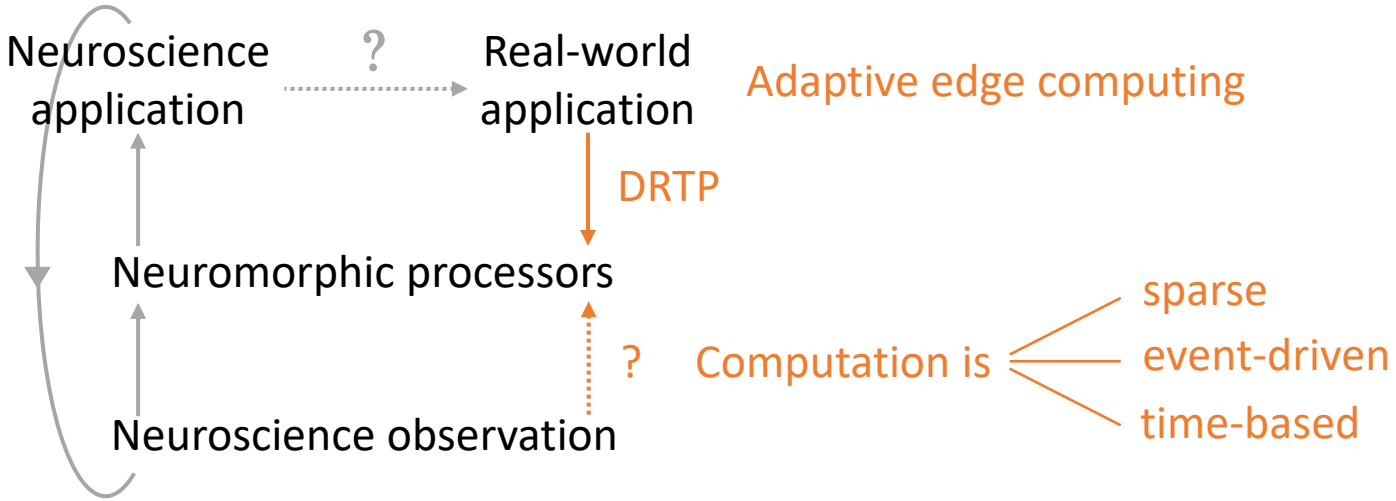
- Algorithms
- Integration

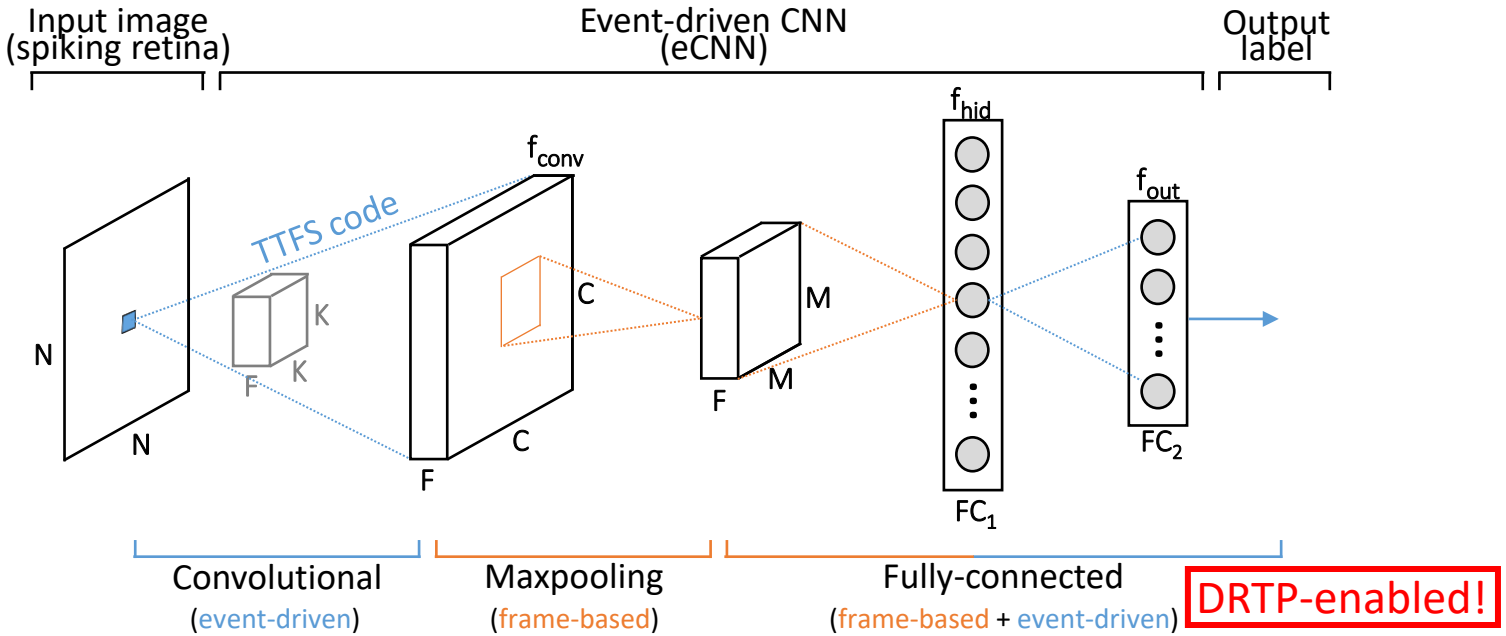
Neuromorphic accelerators [Frenkel, *ISCAS*, 2020]

Conclusion and perspectives

# Which bio-inspired elements?

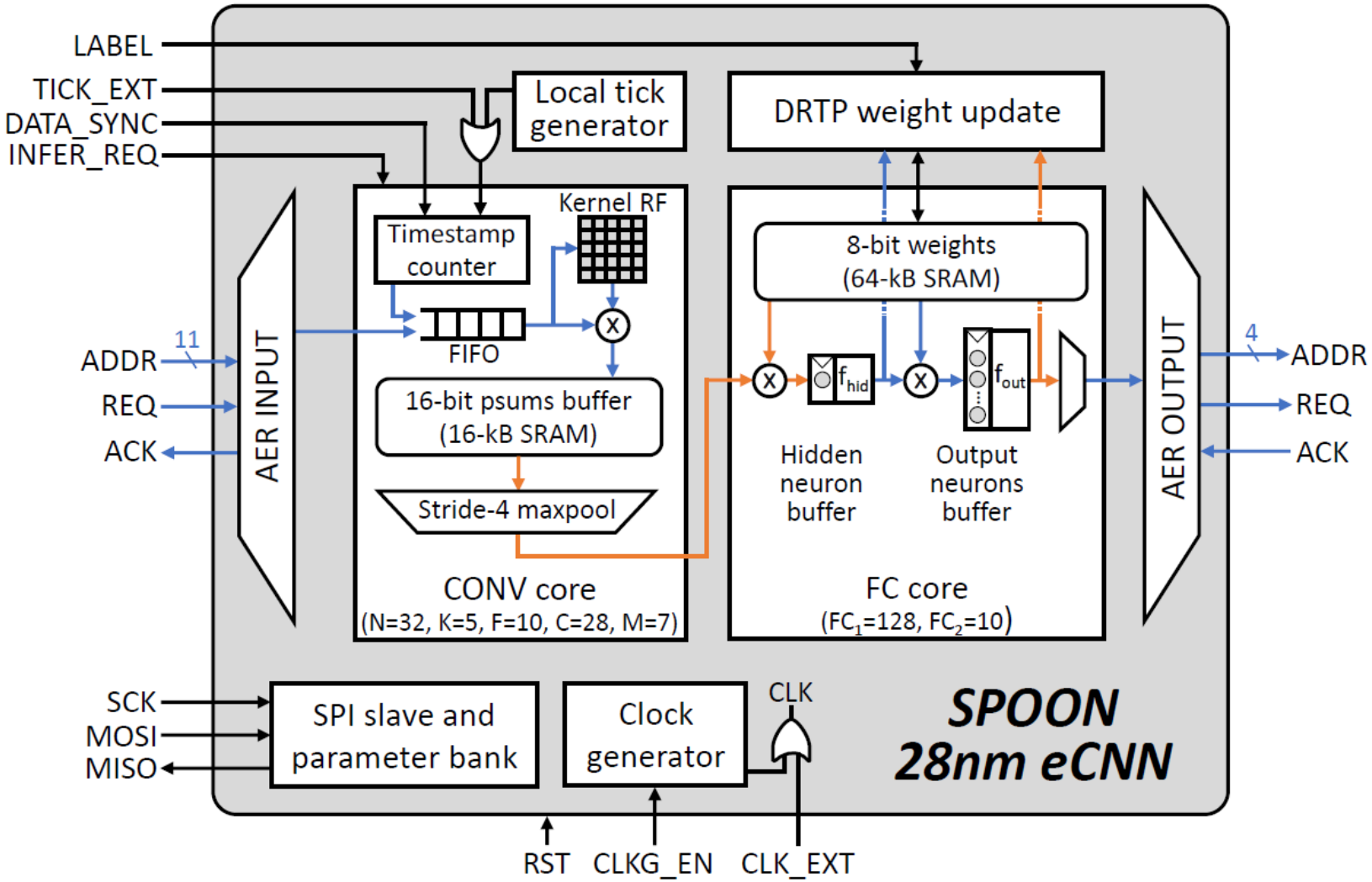
*Taking a step back with the top-down design strategy*



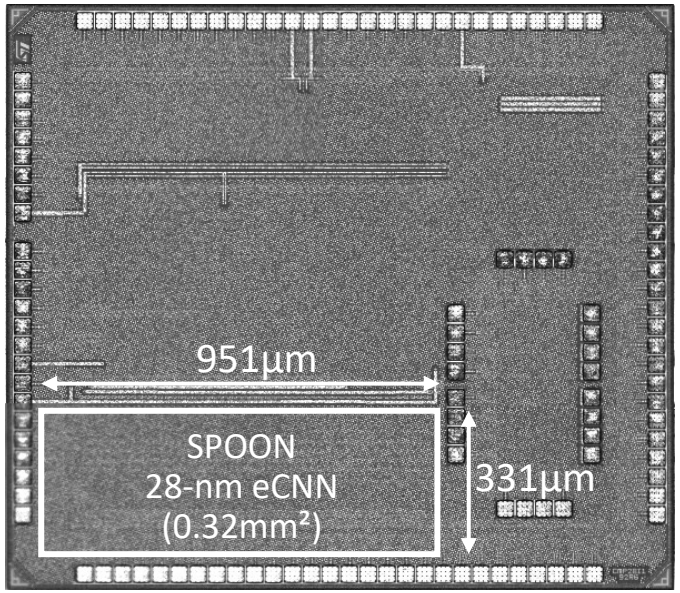
# Architecture of SPOON

SPOON – A Spiking Online-Learning Convolutional Neuromorphic Processor



# SPOON – Chip microphotograph and specifications

→ *perspectives*




*(pre-silicon numbers, not yet updated)*

| | |
|---|---|
| Technology | 28nm FDSOI CMOS |
| Implementation | Digital |
| Area | 0.32mm$^2$ (0.26mm$^2$ excl. rails) |
| Topology | C5×5@10–FC128–FC10 |
| Online learning | Stochastic DRTP, 8-bit weights |
| Time constant | Biological to accelerated |
| Supply voltage | 0.6V – 1.0V |
| Max. clock frequency | 150MHz |
| Leakage power | 61$\mu$W at 0.6V |
| Energy for CONV core | 1.7nJ/event at 0.6V |
| Energy for FC core | 55nJ/inference at 0.6V |
| Online learning overhead | 16.8% in power, 11.8% in area |

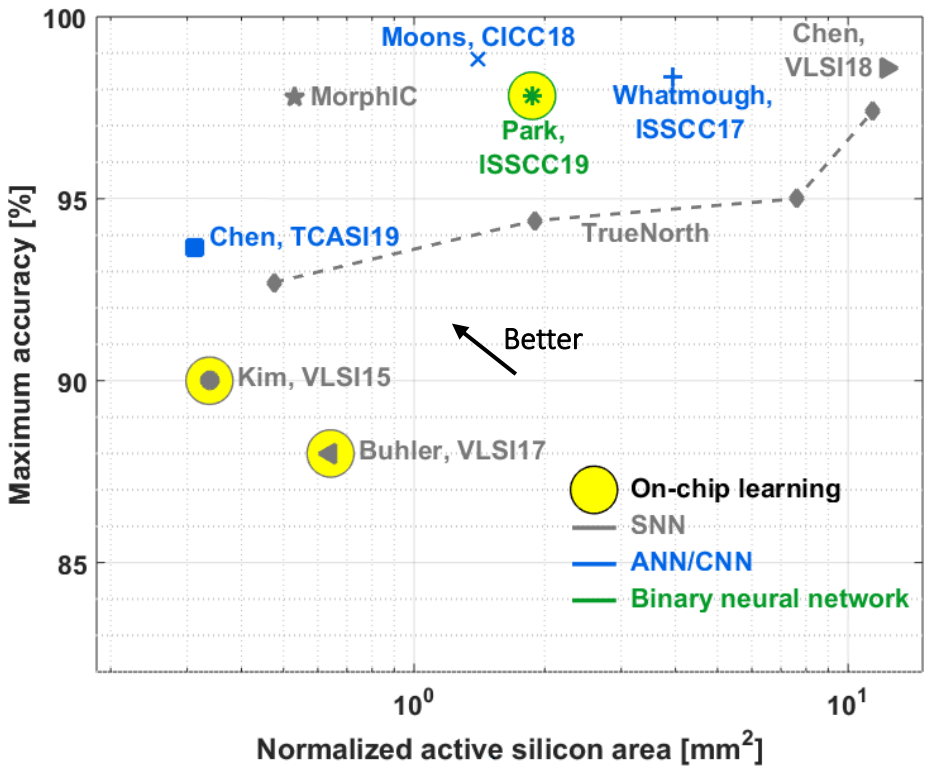
Stay tuned for the journal extension!

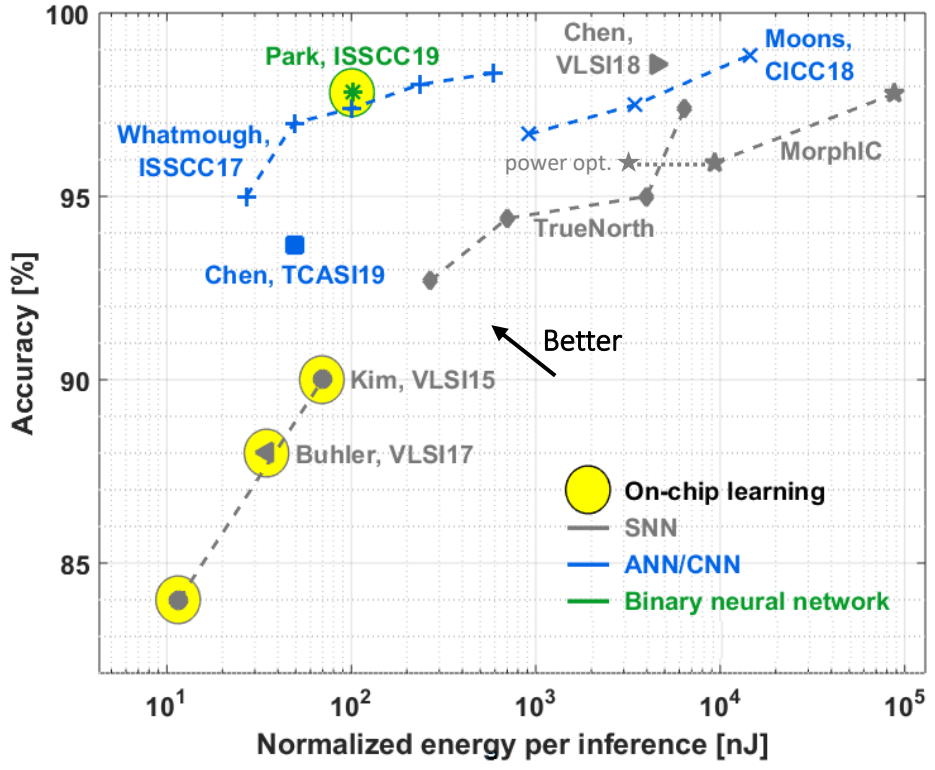
DRTP can be implemented on-chip at a very low cost!

Benchmarking: MNIST and N-MNIST

# SPOON benchmarking

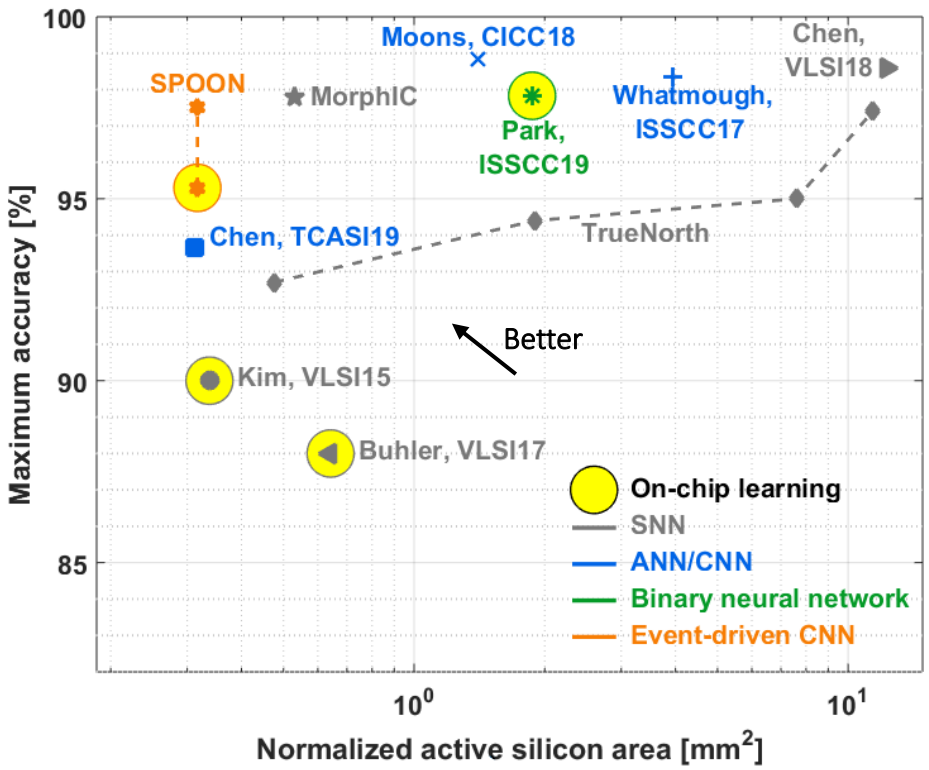
*Against SoA spiking neural networks on MNIST*

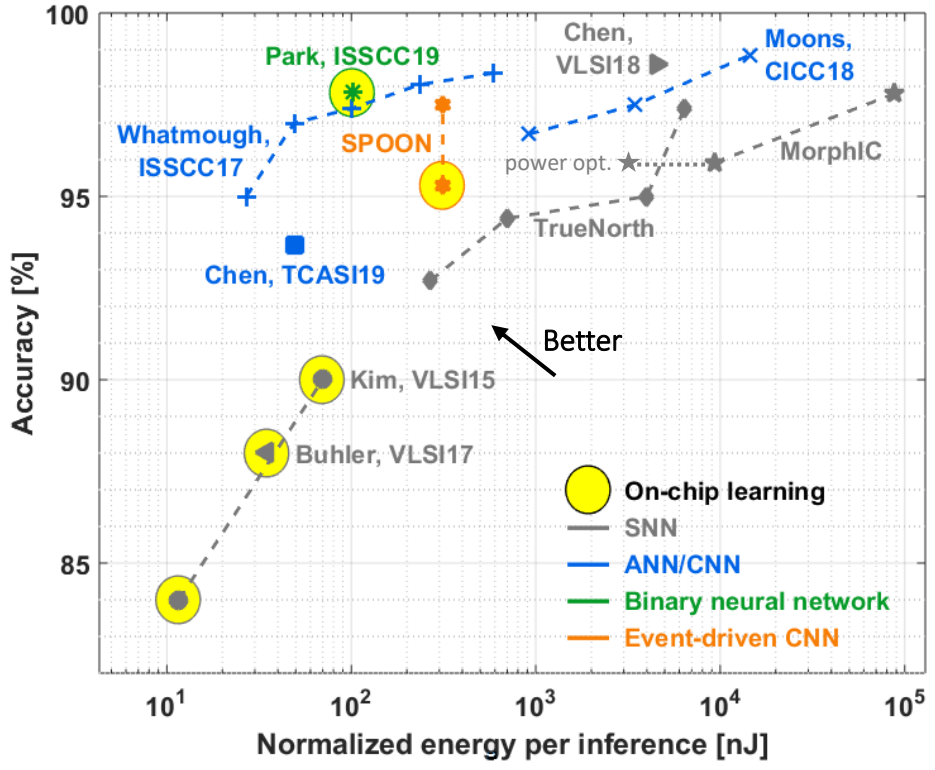


# SPOON benchmarking

*Against SoA spiking neural networks on MNIST*




Only SPOON allows reaching the efficiency of ANN/CNN/BNN accelerators while enabling online learning with event-based sensors.

# Outline

Part I – Bottom-up neuromorphic design

- Building blocks
- Integration

Part II – Top-down neuromorphic design

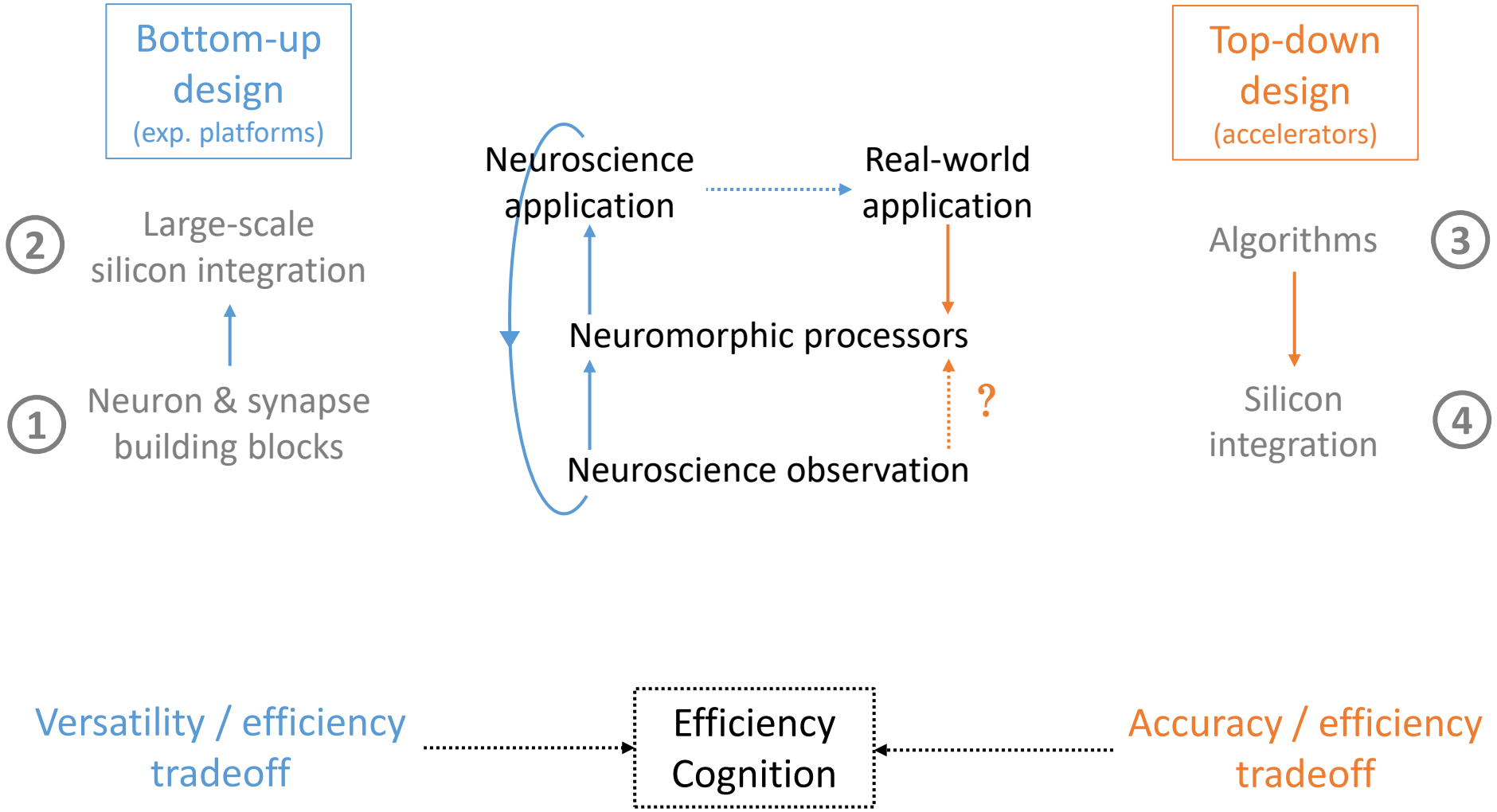
- Algorithms
- Integration

Conclusion and perspectives

Summary of the key messages, next directions

# Neuromorphic Engineering – Key Claims

*Unveiling roads to embedded cognition*



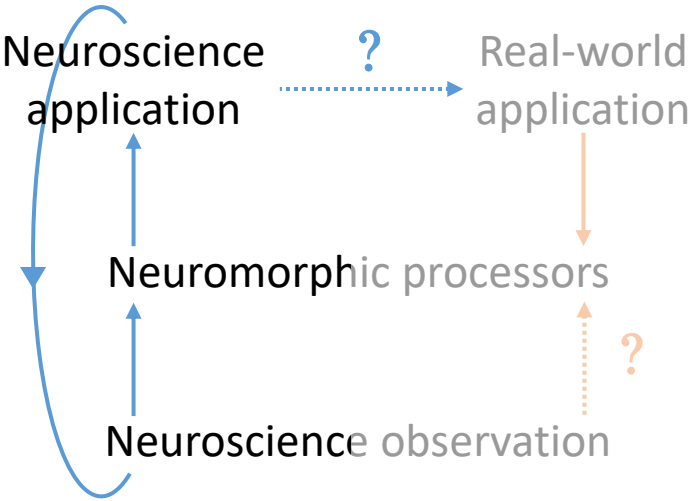
# Neuromorphic Engineering – Key Claims

*Unveiling roads to embedded cognition*

Bottom-up design
(exp. platforms)

② Large-scale silicon integration

①  Neuron & synapse building blocks



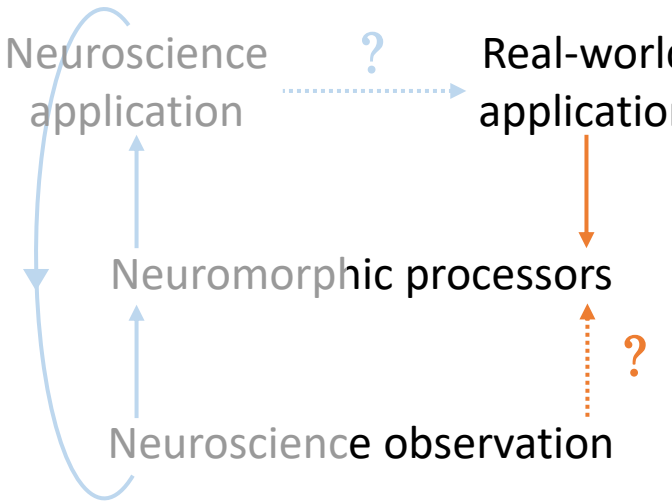

Versatility / efficiency tradeoff

**Claim 1**

Hardware-aware neuroscience model design and selection allows reaching record neuron and synapse densities with low-power operation for large-scale integration *in silico*.

# Neuromorphic Engineering – Key Claims

*Unveiling roads to embedded cognition*




Top-down design
(accelerators)

Algorithms ③

Silicon integration ④

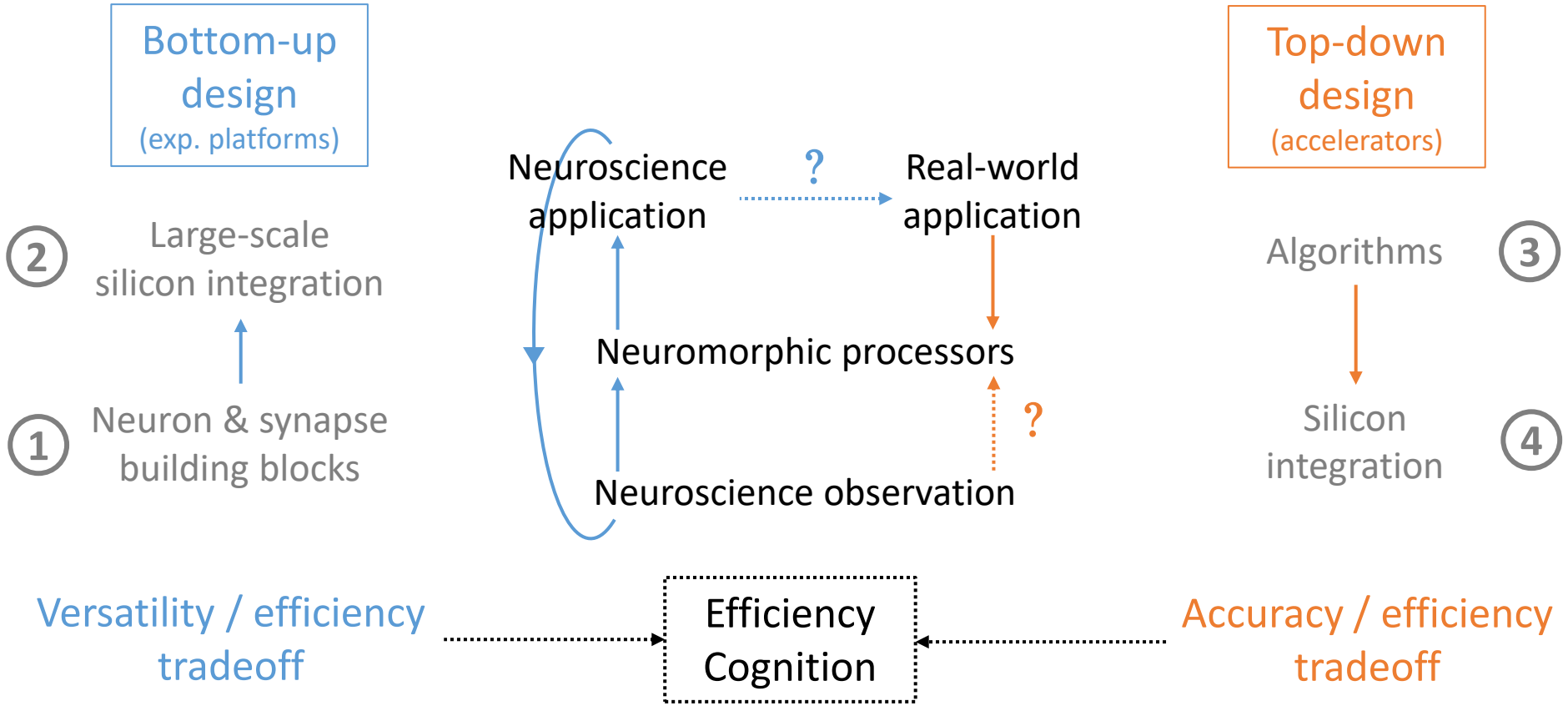
**Claim 2**

Combining event-driven and frame-based processing with weight-transport-free update-unlocked training supports low-cost adaptive edge computing with spike-based sensors.

Accuracy / efficiency tradeoff

# Neuromorphic Engineering – Key Claims

*Unveiling roads to embedded cognition*

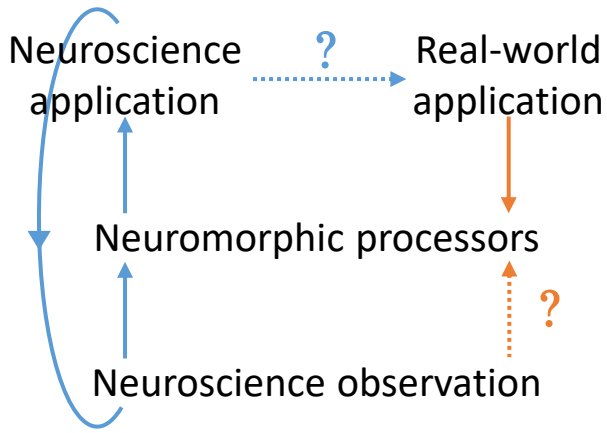



**Claim 3**

Top-down guidance helps pushing bottom-up neuron and synapse integration beyond the purpose of neuroscience experimentation platforms, while bottom-up guidance supports top-down design toward brain reverse-engineering.

# Perspectives

- Neuromorphic engineering and spiking neural networks:
  "Can we make it work?" → "Will it bring a competitive advantage?" (not only against GPUs)
  Need something better than MNIST → Audio (KWS) and bio-signal processing (time, biological-time)
  *[Davies, Nat. Mach. Intel., 2019]*
- Promising avenue: fine-grained mixed-signal design.
- Bottom-up trend: dendrites
- Top-down trend: new wave of training algorithms mapping onto bio-plausible primitives
  [Sacramento, NeurIPS'18]
  [Payeur, bioRxiv, 2020]
  [Bellec, Nat. Comms., 2020]
- Cognition: a case for neuromorphic robots?
  [Man & Damasio, Nat. Mach. Intel., 2019]



# Acknowledgments

PhD → *To be continued…* → Postdoc

Institution:

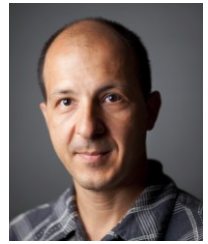
Funding:

Supervisors:

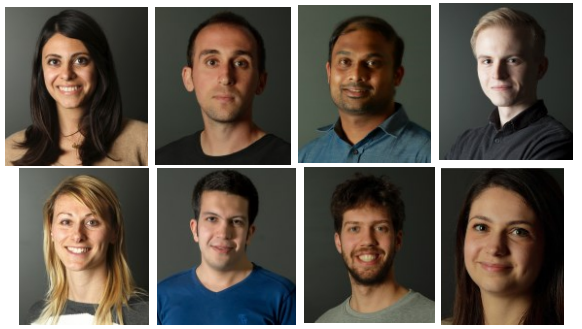

*Profs. David Bol, Jean-Didier Legat*

Key colleagues:

*Martin Lefebvre*



*Prof. Giacomo Indiveri*



# Questions?

@C_Frenkel
cfrenkel
Charlotte-Frenkel
ChFrenkel
charlotte@ini.uzh.ch

## Main references:

- ODIN: [C. Frenkel et al. "MorphIC: A 65-nm 738k-synapse/mm² quad-core binary-weight digital neuromorphic processor with stochastic spike-driven online learning," *IEEE Trans. BioCAS*, 2019]
- MorphIC: [C. Frenkel et al., "A 0.086-mm² 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28nm CMOS," *IEEE Trans. BioCAS*, 2019]
- DRTP: [C. Frenkel, M. Lefebvre et al., "Learning without feedback: Fixed random learning signals allow for feedforward training of deep neural networks," Frontiers in Neuroscience, 2021]
- SPOON: [C. Frenkel et al., "A 28-nm convolutional neuromorphic processor enabling online learning with spike-based retinas," *IEEE ISCAS*, 2020]

*Open-sourced!*
github.com/ChFrenkel/ODIN

*Open-sourced!*
github.com/ChFrenkel/DirectRandomTargetProjection

*Journal extension coming soon*