



Krembil
Relentless.



UHN Toronto General
Toronto Western
Princess Margaret
Toronto Rehab
Michener Institute
COURAGE LIVES HERE



UNIVERSITY OF
TORONTO

Neuromorphic Computing for Accelerating AI Models: SpiNNaker and Loihi Platforms

Idir Mellal, Ph.D. Eng

Neural Systems and Brain Signal Processing Lab (NSBSPL)

Overview

- Biological Neurons implementation on Hardware devices can **accelerate** the computations
- **Complexity** and **high parallelism** of biological neurons **complicate** the HW implementations
- **Neuromorphic computing**: the end of the Van Neuman architecture
- Few examples of neuromorphic platforms: **SpiNNaker**, **Loihi** and **TrueNorth**
- **ASIC** and **FPGA**: One purpose and two different ways to **emulate** complex architectures

Plan

- Introduction
- Neuromorphic Computing
- From Biology to Electronic!
- Introduction to Neuromorphic computing
- Neuromorphic Platforms: SpiNNaker and Loihi
- Why Neuromorphic for Accelerating AI models?

Introduction

- Hardware Implementation can **speed up the computations** and **increase performances** of a complex architectures
- **FPGA** (Field Programmable Gate Array) are a popular devices used to **prototype** and **emulate digitally** mathematical and logical models for different purposes: industry, control, security, AI, and Neuromorphic
- **Neuromorphic computing** allows us to perform **high parallel** and **real-time** computing with **non-conventional** Van Neuman machine

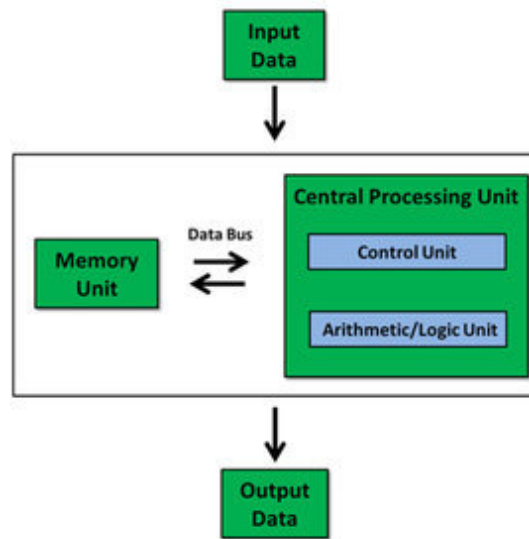
Neuromorphic Computing

- Neuromorphic devices represent an attempt to **mimic aspects of the brain's architecture and dynamics** with the aim of replicating its hallmark functional capabilities in terms of computational power, robust learning and energy efficiency
- It offers a **realistic emulation** of the neuronal membrane dynamics using electronic circuits or simulated using specialized digital systems
- Some **applications**: speech recognition, character recognition, grammar modeling, noise modeling, as well as the generation and prediction of chaotic time series
- Neuromorphic chips, unlike conventional processor **are energy efficient and fully parallelized**
- Resolve the **von Neuman** bottleneck by **collocating** the processor and the memory

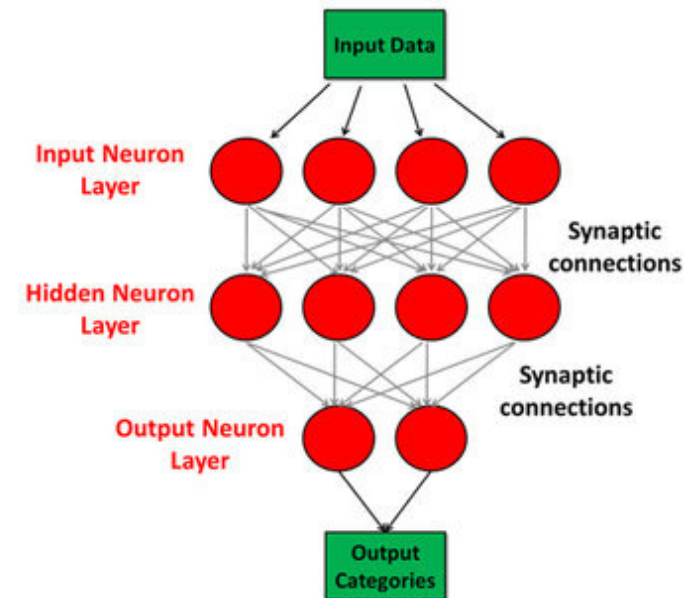
Neuromorphic Computing

- Limitations of the tradition Von-Neuman Architecture
- Neuromorphic architecture breakthrough computation architecture

Von-Neumann architecture



Neuromorphic architecture

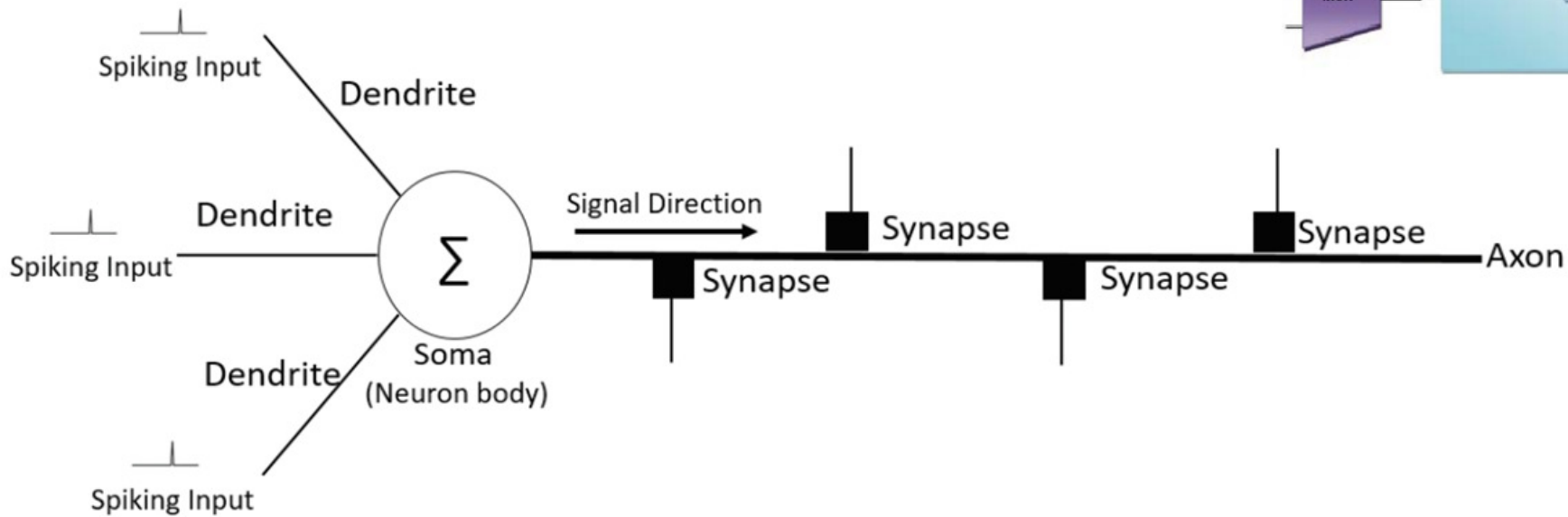
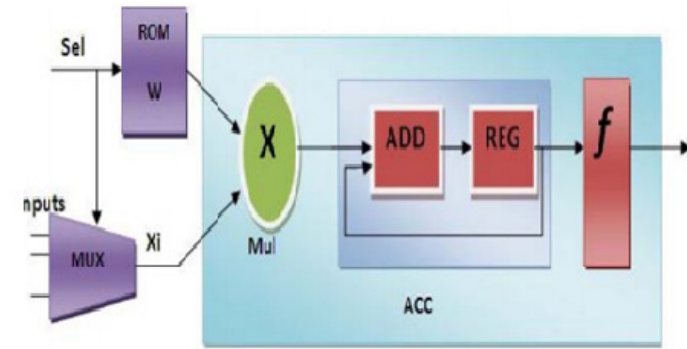
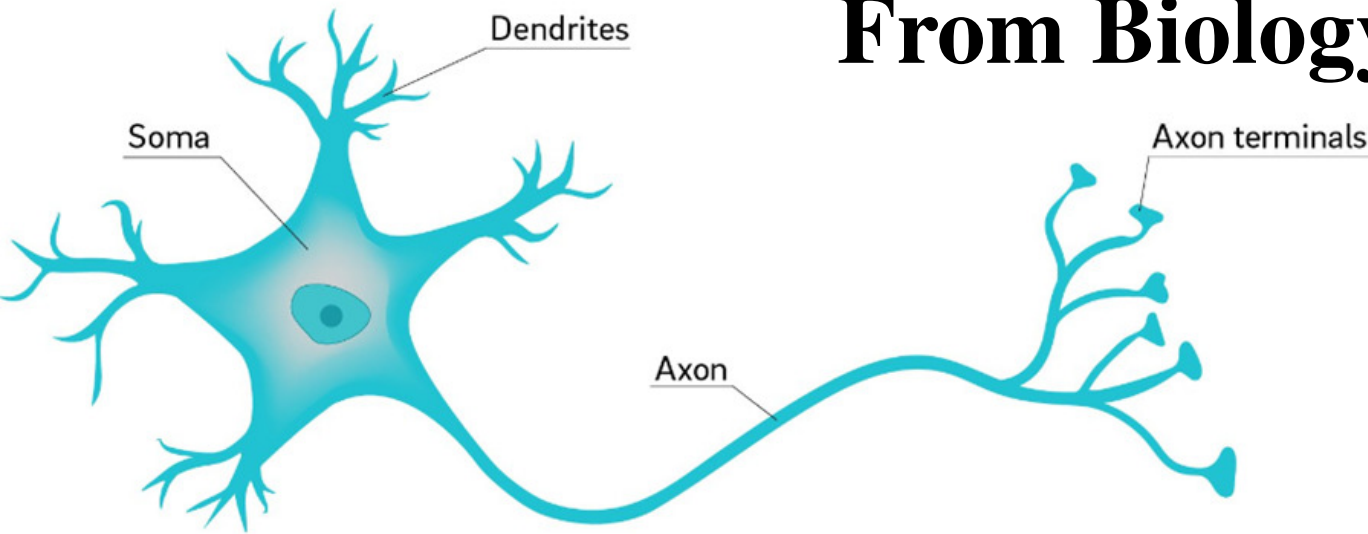


From Biology to Circuit!

- Design an ASIC is **very costly** in time and money because of the complexity of the design flow which require the design of each single transistor and each connection
- ASICs offer **less power** consumption and **higher performances** than **FPGAs**
- Analog platform using **memristive** that can retain a state of internal resistance based on the history of applied voltage and current



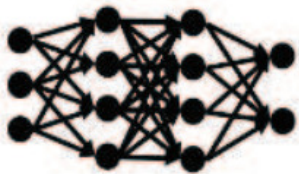
From Biology to Circuit



Comparison between brain, von Neumann, and neuromorphic computing architectures

Neuromorphic Computing

Learning Algorithms



Neural Networks

Spiking Signals

Artificial Neurons and Synapses



Human Brain

Unknown



Brain Computing System

Spiking Signals

Biological Neurons & Synapses

Neurons
• Synapses

Algorithm

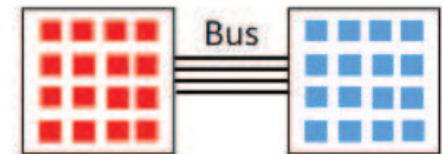
Architecture

Encoding Scheme

Devices

Digital Computer

Programs/Logic

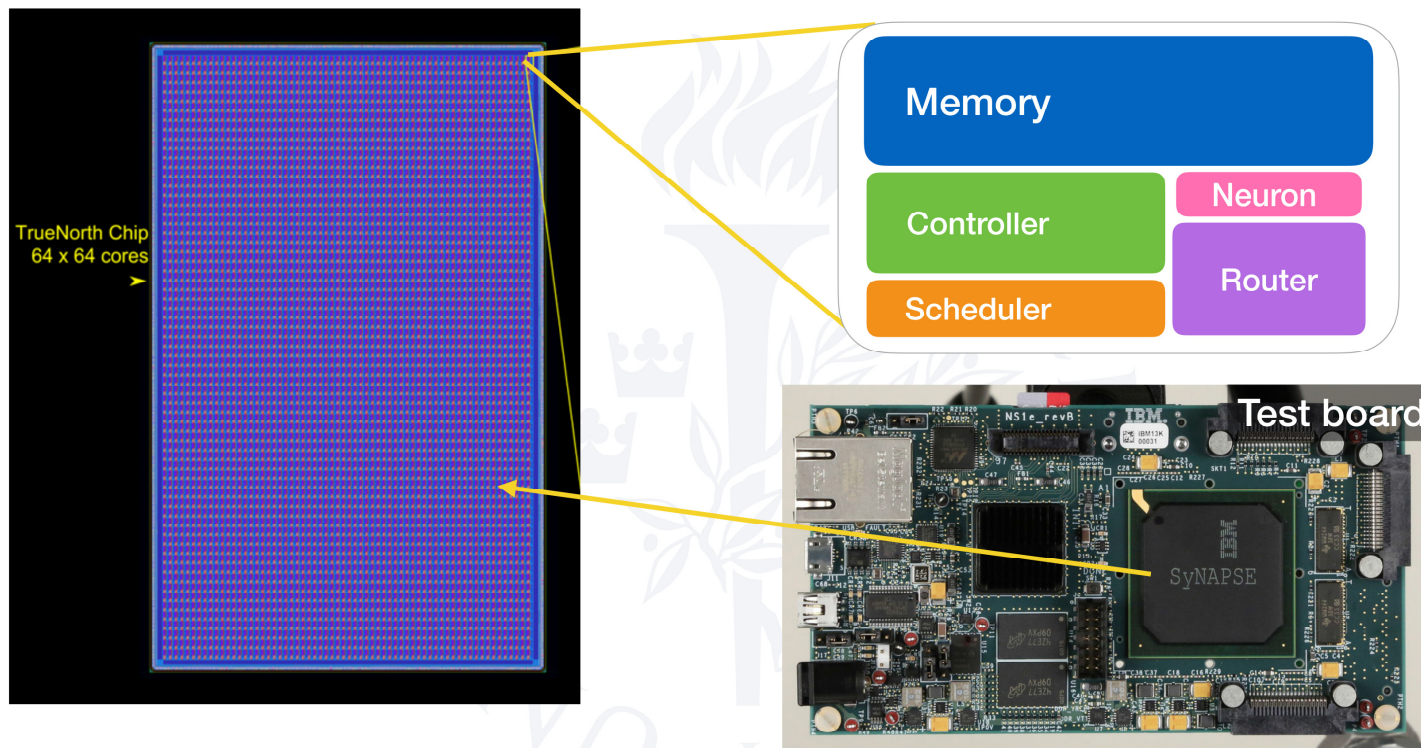


CPU
Memory
Von Neumann Architecture

Binary Signals

CPUs (Logic Gates, etc.),
Memory (SRAM, etc.)

Processing Units and Memory are collocated in Neuromorphic platforms

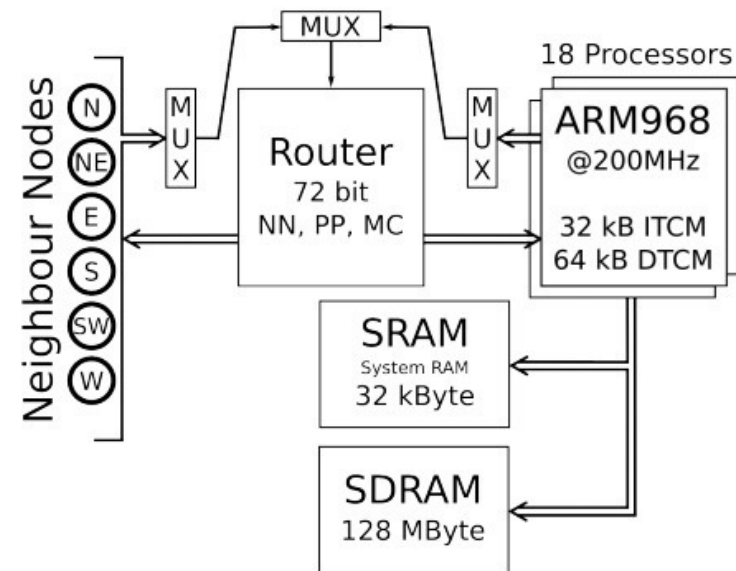


SpiNNaker: Human Brain Project Neuromorphic Platform

- SpiNNaker: SpiNNaker 2 will consist of 10 M ARM cores distributed across 70.000 Chips in 10 server racks.



The SpiNNaker Board: A building block of a SpiNNaker machine, containing 48 chips for a total of 864 ARM processors.




SpiNNaker Architecture: The schematics of a SpiNNaker Chip with processors, router and shared memory.

You can access documentation and tutorials related to HBP at:

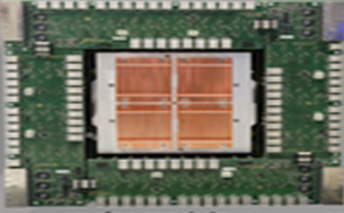
<https://www.humanbrainproject.eu/en/silicon-brains/>

2020-03-26


The BrainScaleS neuromorphic physical model system



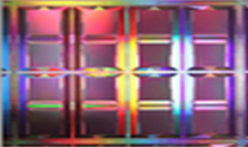
20 wafer modules
3.932.160 neurons
880.803.840 synapses



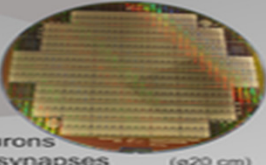
wafer module (50 cm x 50 cm)



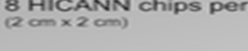
components of a wafer module




48 reticles per wafer
196.608 neurons
44.040.192 synapses




(\approx 20 cm)




8 HICANN chips per reticle
(2 cm x 2 cm)




512 neurons
114.688 synapses per HICANN chip
(0.5 cm x 1 cm)



1 plastic synapse
(10 μ m x 10 μ m)



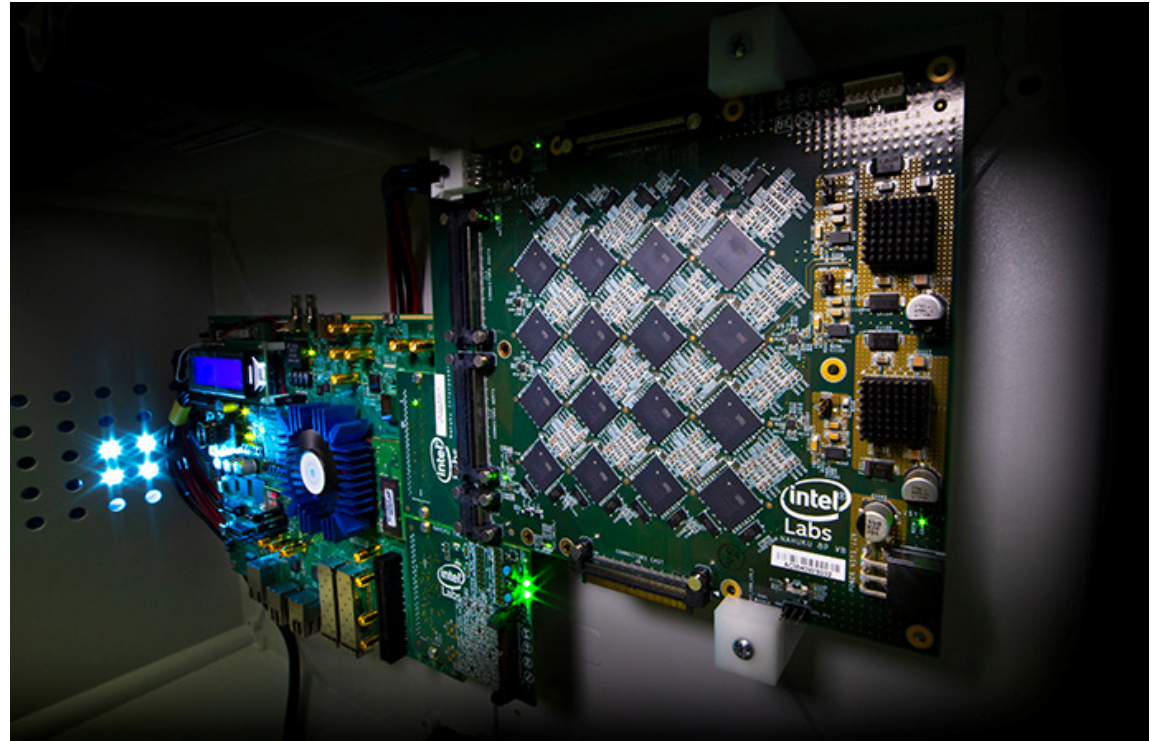
2 neurons
(150 μ m x 20 μ m)



info@neuromorphic.eu

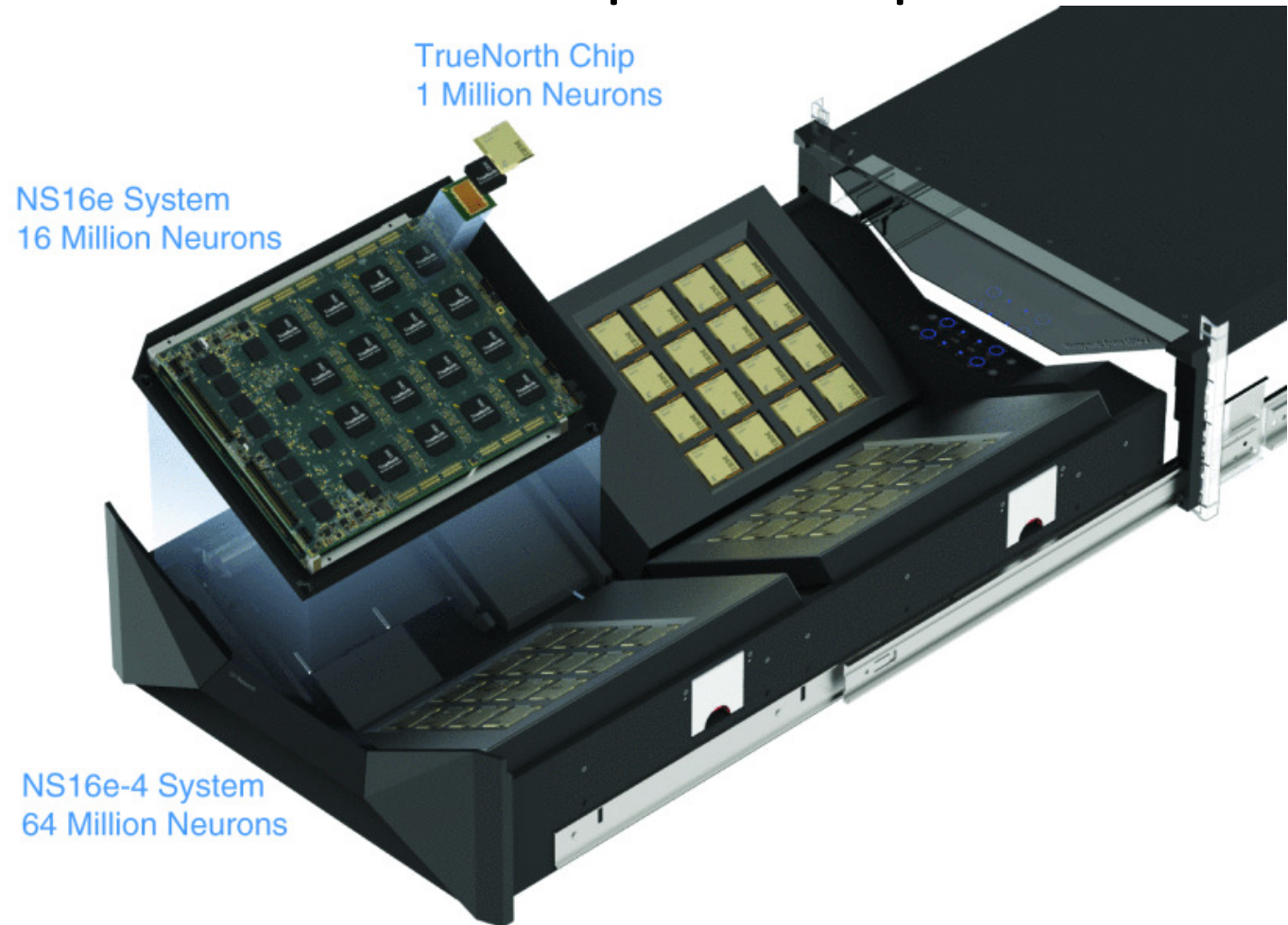
Loihi: HP Neuromorphic Platform

- **Loihi** is the last development of Neuromorphic technology. It was proposed by Intel in 2018.
- It was updated in July 2019 to become **Pohoiki Beach Platform** and including 64 **Loihi board**. It can implement more than 8M neurons (8.3)!



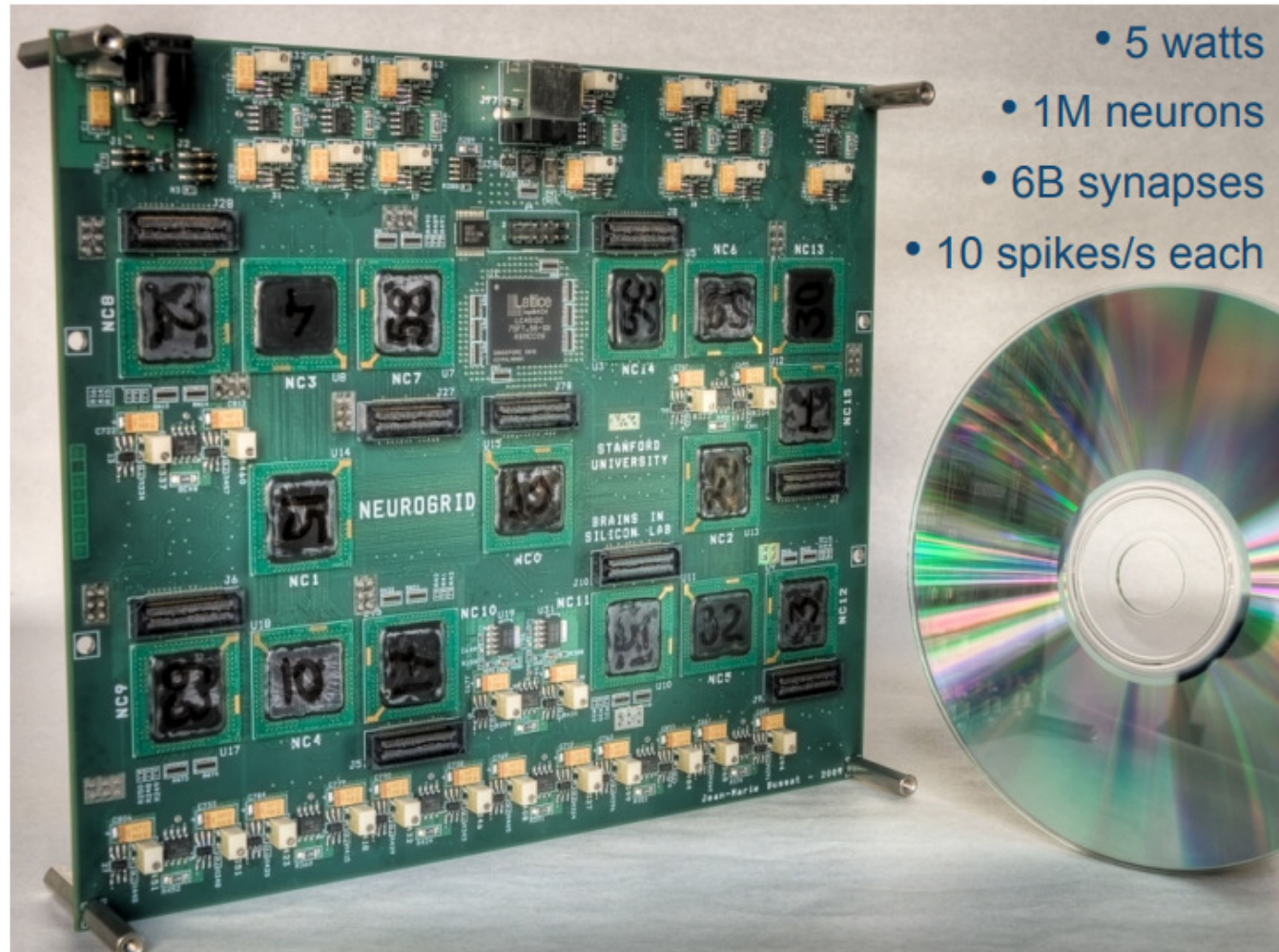
TrueNorth: IBM Neuromorphic Chip

- The NS16e-4 is one of the largest neurosynaptic computer built to date, totaling **64 million neurons** and **16 billion synapses**. At only **70 W** the system's computational energy efficiency is unprecedented, on the order of $\sim 10^{11}$ synaptic operations per second/W



NEUROGRID

- Developed by Stanford University
- It emulates biological neurons
- Uses a analog computations and digital communication module
- 16 chips of 256x256 array



Why Neuromorphic for Accelerating AI Models?

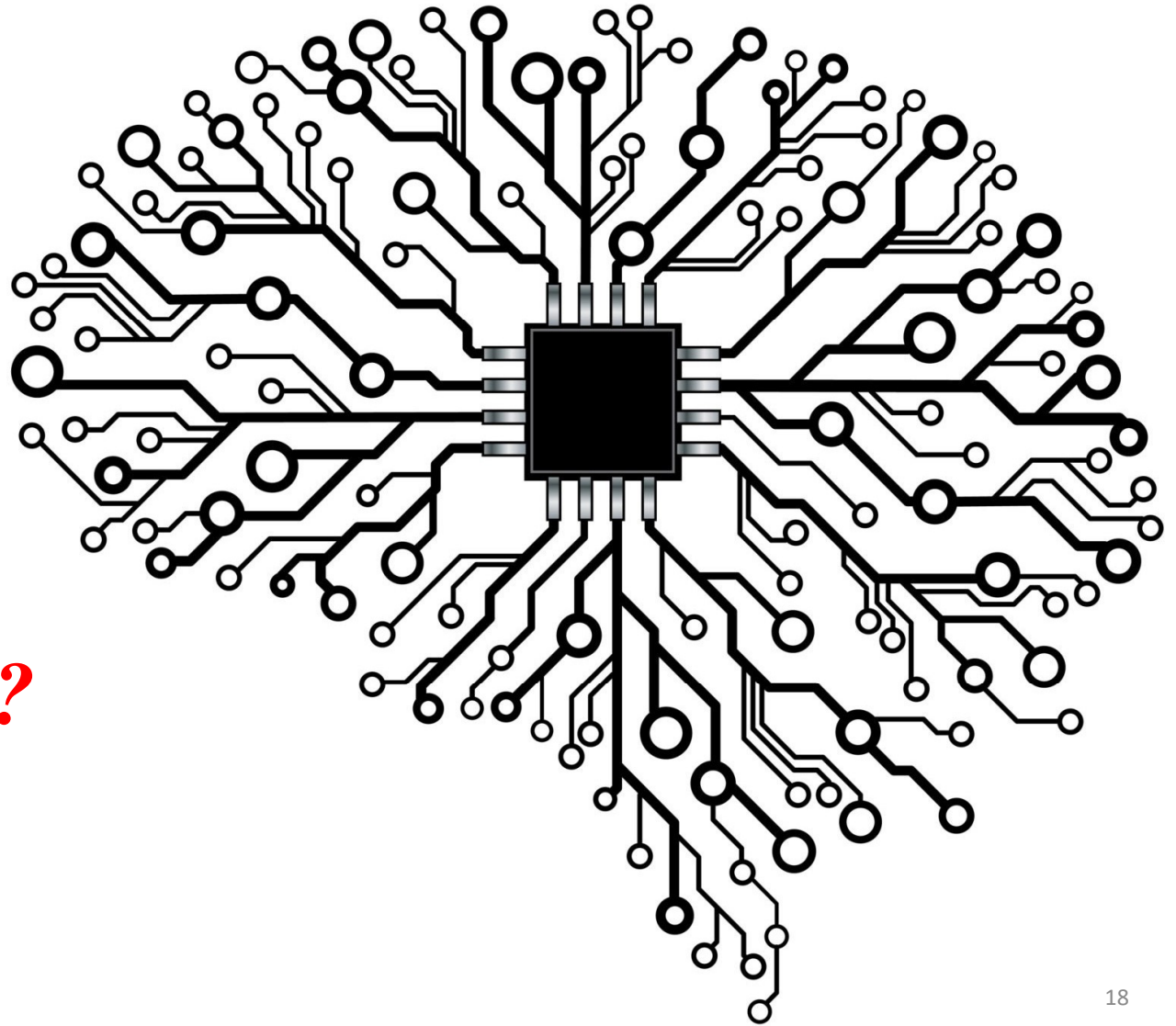
Neuromorphic Computing offers :

- Implementation Efficiency by ensuring:
 - High throughput
 - Low energy consumption
 - Real-Time processing
- Flexibility of the implementations: Ability to run different model under different constraints without being restricted to one model.
 - Variation in layers
 - Algorithm optimization
 - Online training

References:

- [1] F. Perez-Peña, M. A. Cifredo-Chacon, and A. Quiros-Olozabal, “Digital neuromorphic real-time platform,” *Neurocomputing*, Sep. 2019.
- [2] C. D. Schuman *et al.*, “A Survey of Neuromorphic Computing and Neural Networks in Hardware,” *ArXiv170506963 Cs*, May 2017.
- [3] T. Wunderlich *et al.*, “Demonstrating Advantages of Neuromorphic Computation: A Pilot Study,” *Front. Neurosci.*, vol. 13, 2019.
- [4] H. An, K. Bai, and Y. Yi, “The Roadmap to Realize Memristive Three-Dimensional Neuromorphic Computing System,” *Adv. Memristor Neural Netw. - Model. Appl.*, Oct. 2018.
- [5] S. J. van Albada *et al.*, “Performance Comparison of the Digital Neuromorphic Hardware SpiNNaker and the Neural Network Simulation Software NEST for a Full-Scale Cortical Microcircuit Model,” *Front. Neurosci.*, vol. 12, May 2018.
- [6] R. A. Nawrocki, R. M. Voyles, and S. E. Shaheen, “A Mini Review of Neuromorphic Architectures and Implementations,” *IEEE Trans. Electron Devices*, vol. 63, no. 10, pp. 3819–3829, Oct. 2016.
- [7] B. Rajendran, A. Sebastian, M. Schmuker, N. Srinivasa, and E. Eleftheriou, “Low-Power Neuromorphic Hardware for Signal Processing Applications,” *ArXiv190103690 Cs*, Jan. 2019.

Thank you



Questions?

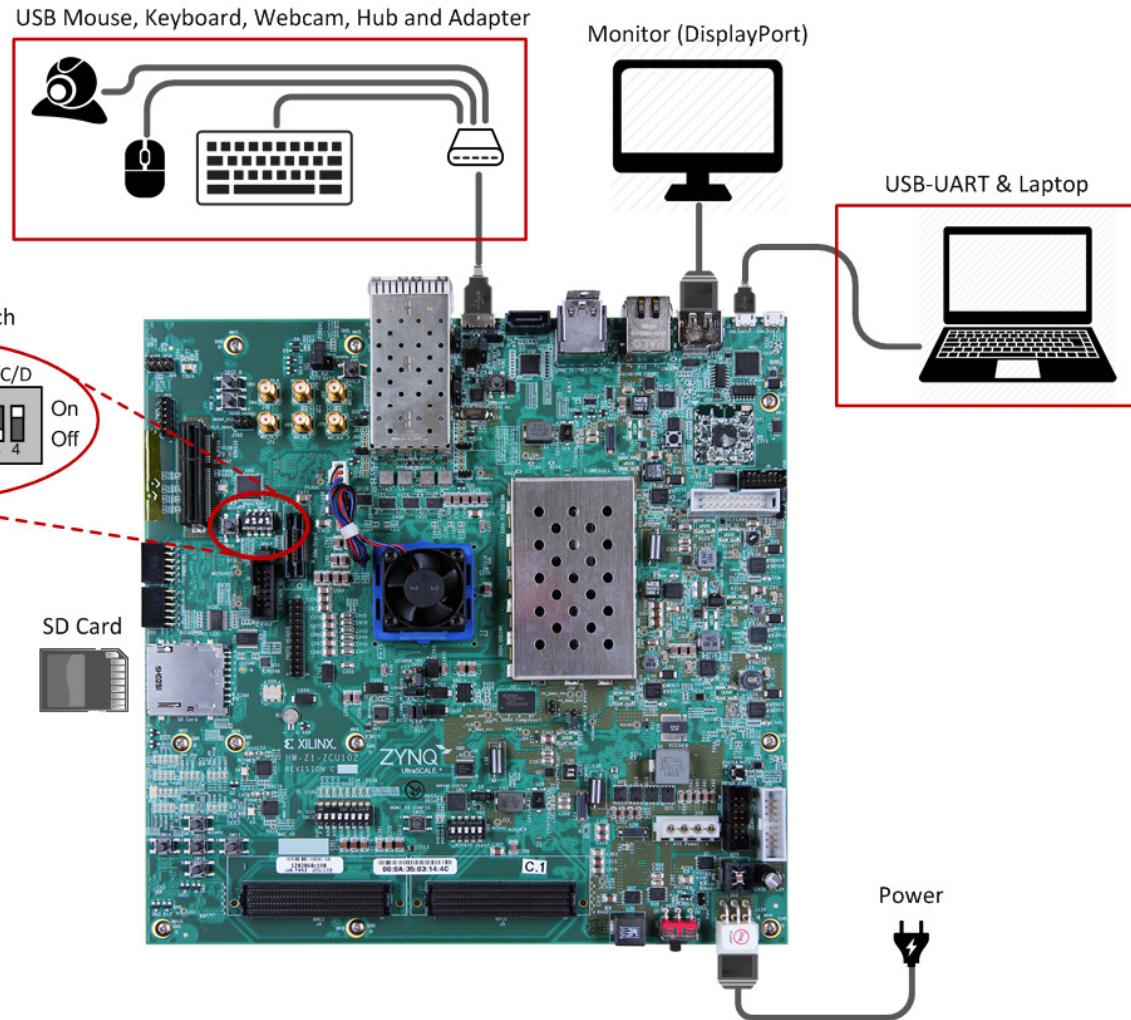
Challenges:

We can categorize on three types:

- Computational challenges:
- Memory challenges:
- Accuracy challenges:

Xilinx ZCU FPGA

We will use the Xilinx Zynq UltraScale+ MPSoC ZCU102 Evaluation Kit to implement our neuronal Model and interface it with experimental data



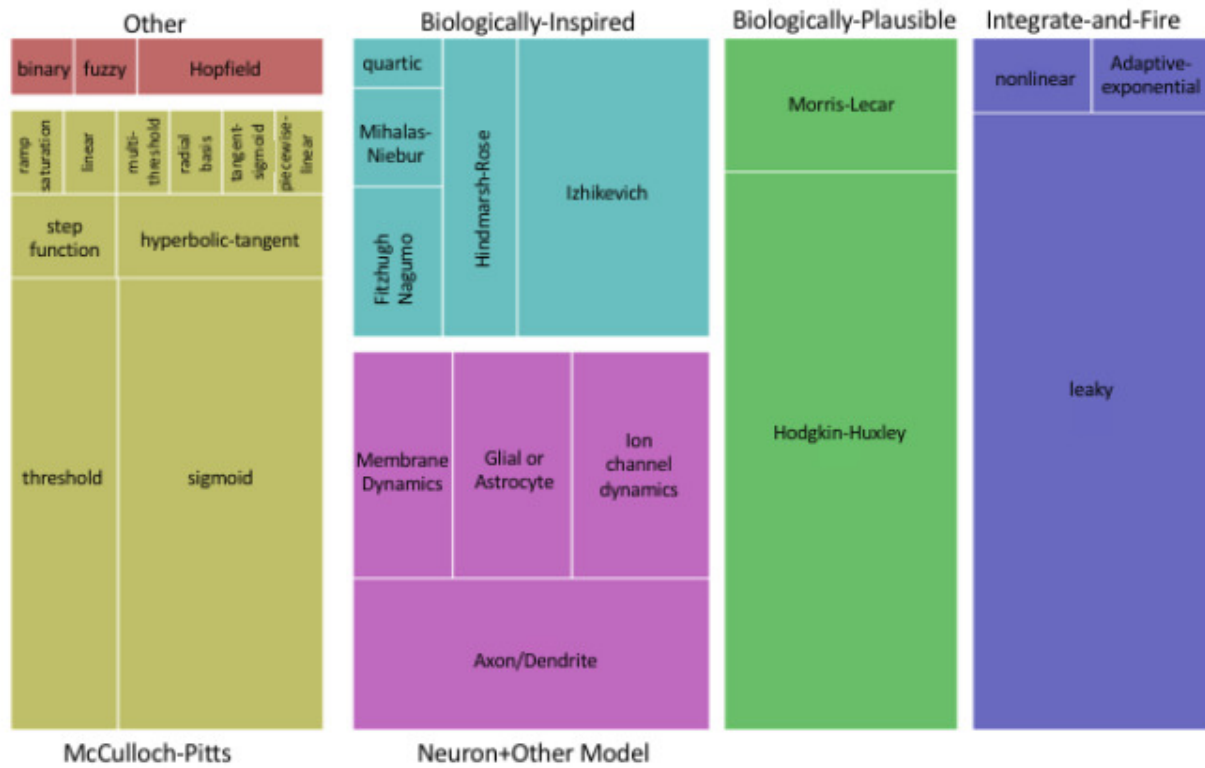
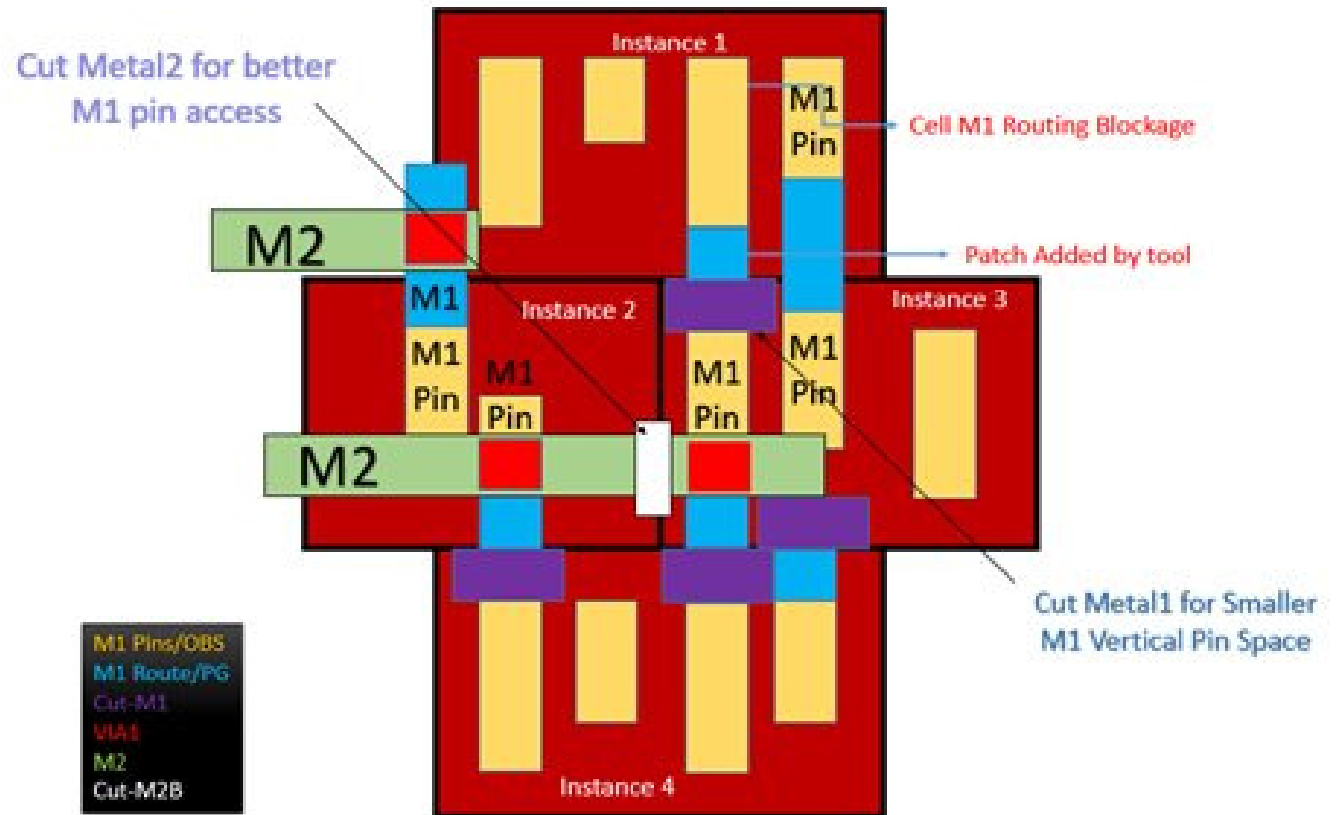


Fig. 4. A hierarchy of neuron models that have hardware implementations. The size of the boxes corresponds to the number of implementations for that model, and the color of the boxes corresponds to the “family” of neuron models, which are labeled either above or below the group of same-colored boxes.

With **ASIC** the designers should go through each single unit and do verifications and corrections of the technology constraints which is very costly in time



- TrueNorth was able to classify images at **between 1,200 and 2,600** frames per second (fps), while drawing just **25 to 275 milliwatts** of power. That works out to about **6000 fps per watt**, which would allow a low-power device to classify images in real-time from dozens of standard TV video feeds simultaneously. For the sake of comparison, NVIDIA's latest purpose-built inferencing GPU, the Tesla P4, can classify images at about 160 images per second per watt using AlexNet.

LOIH communication

- Spikes are transported between the cores in the chip using packetized messages by an asynchronous network-on-chip (NoC) and allows connecting to 4096 on-chip cores and up to 16,384 chips via hierarchical addressing.
- At nominal operating conditions, Loihi delivers 30 billion synaptic operations per second, consuming about **15 pico Joule per synaptic operation**.

Chip	Technology	Integration density	Key functionality/performance metric
SpiNNaker	ARM968, 130 nm CMOS (next-gen prototypes: ARM M4F, 28 nm CMOS)	Up to 1 K neurons/core, 1 M cores.	Programmable numerical simulations with 72-bit messages, for real-time simulation of spiking networks
TrueNorth	Digital ASIC at 28 nm CMOS	1 M neurons, 256 M Synapses; 1-bit synaptic state to represent a connection, with 4 programmable 9-bit weights per neuron	SNN emulation without on-chip learning; 26 pJ per synaptic operation.
Loihi	Digital ASIC at 14 nm CMOS	130 k neurons, 130 M synapses with variable weight resolution (1-9 bits)	Supports on-chip learning with plasticity rules such as Hebbian, pair-wise, and triplet-STDP, 23.6 pJ per synaptic operation (at nominal operating conditions).
BrainScaleS	Mixed signal waferscale system, 180 nm CMOS (next-gen prototype: 65 nm CMOS)	180 K neurons, 40 M synapses per wafer	$10^3 - 10^4$ fold acceleration of spiking network emulations, with hardware-supported synaptic plasticity. Next-gen prototype: programmable plasticity.
Braindrop	Mixed signal 28 nm CMOS	4096 neurons, 64K programmable weights (with analog circuits that allow realization of all-to-all connectivity)	0.38 pJ per synaptic update, implements the single core of a planned million-neuron chip.
DYNAP-SE	Mixed signal 180 nm CMOS	1024 neurons, 64K synapses (12-bit CAM)	Hybrid analog/digital circuits for emulating synapse and neuron dynamics, 17 pJ per synaptic operation
ODIN	Digital ASIC at 28 nm CMOS	256 neurons, 64K synapses with 3 bit weight and 1 bit to encode learning	12.7 pJ per synaptic operation, implements on-chip spike-driven plasticity