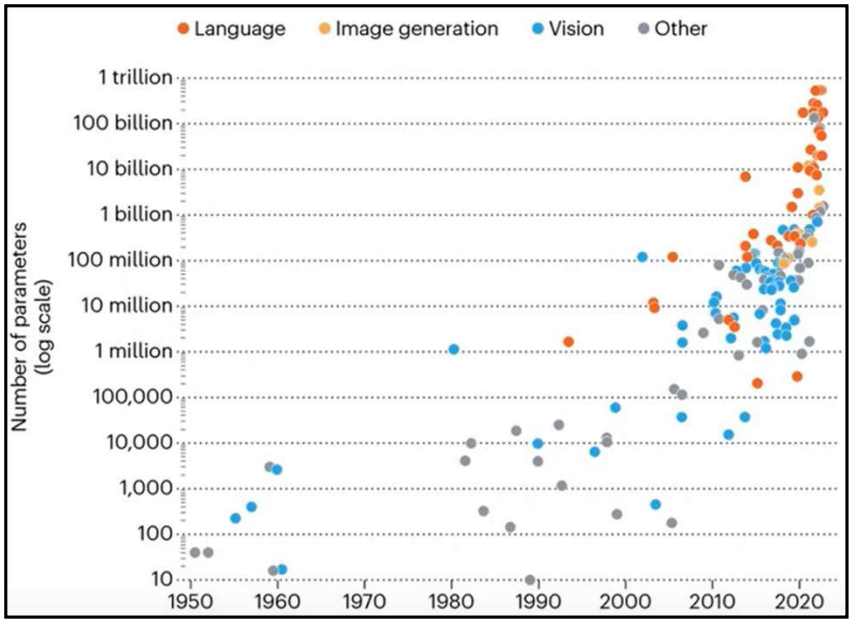
Davide Bertozzi

University of Manchester (UK)

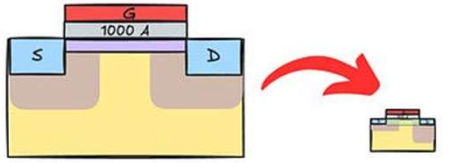
AI is Power Bound

- There is a drive toward bigger AI models

[Nature 2023]



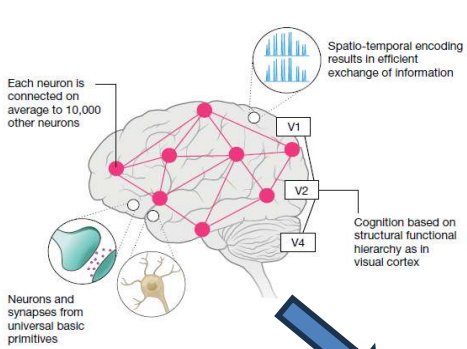
Roughly 10x every 2 years

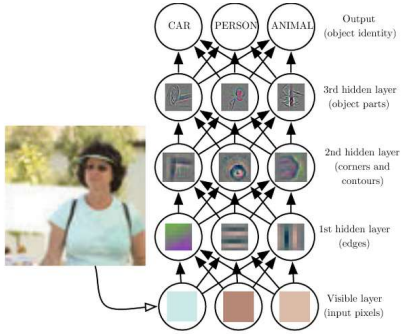


Energy Efficiency from Scaling $\left(\frac{1}{\text{Power}\cdot\text{Time}}\right)$

10x every 12 years => SUSTAINABILITY ISSUE

Time to revisit the AI foundations?





Less and less resemblance

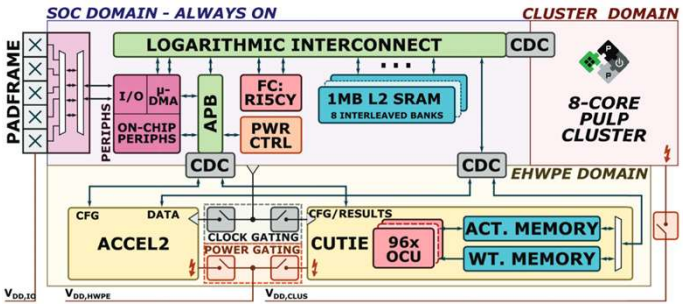
- But AI systems are power capped!

Datacenter PMAX < 150 MW

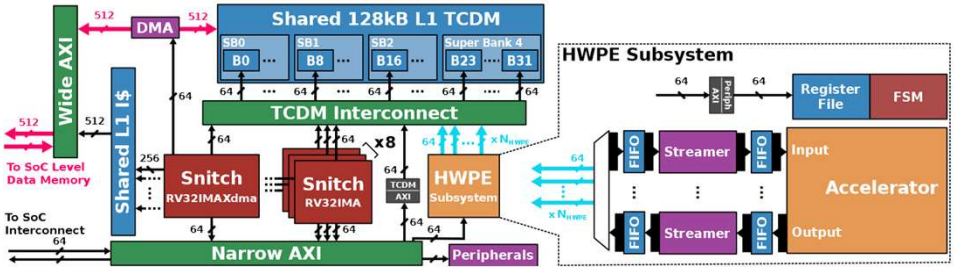
On-car computing PMAX < 1.5KW

Edge Computing

TinyML is actually pushing the limits of on-device machine learning

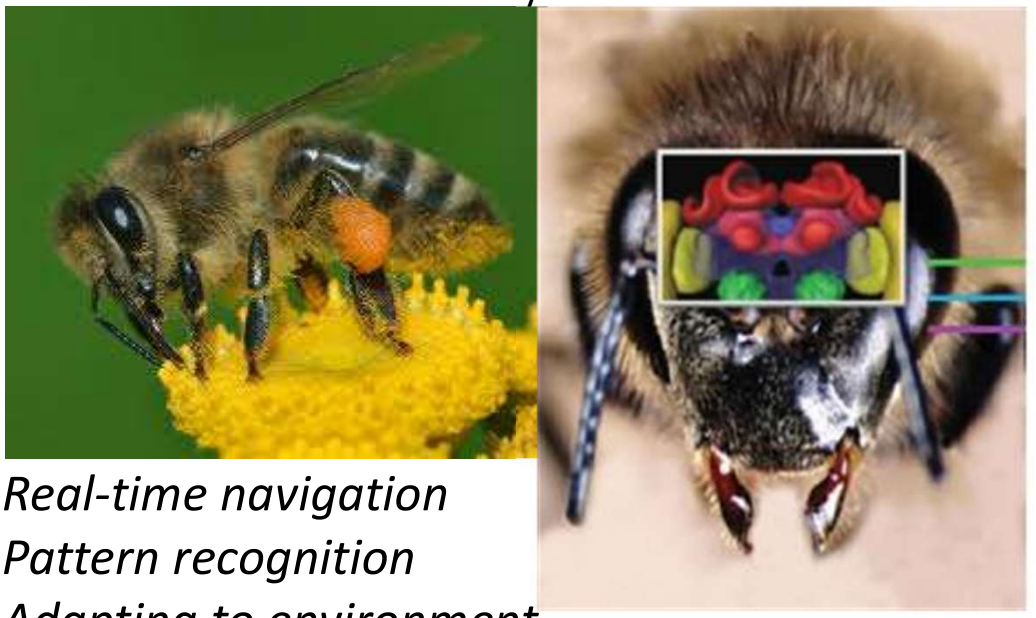


ETH/UniBO, TNN accelerator @12mW (2022)



ETH/UniBO, accelerator for attention-based model@52mW

Honey bees



Real-time navigation
Pattern recognition
Adapting to environment

~mWs for basic image classification tasks

1 million neurons
Power budget: 10 uW

Biological efficiency is still a long way off

Back to Where it all Started

- The brain has something very special about energy efficiency..

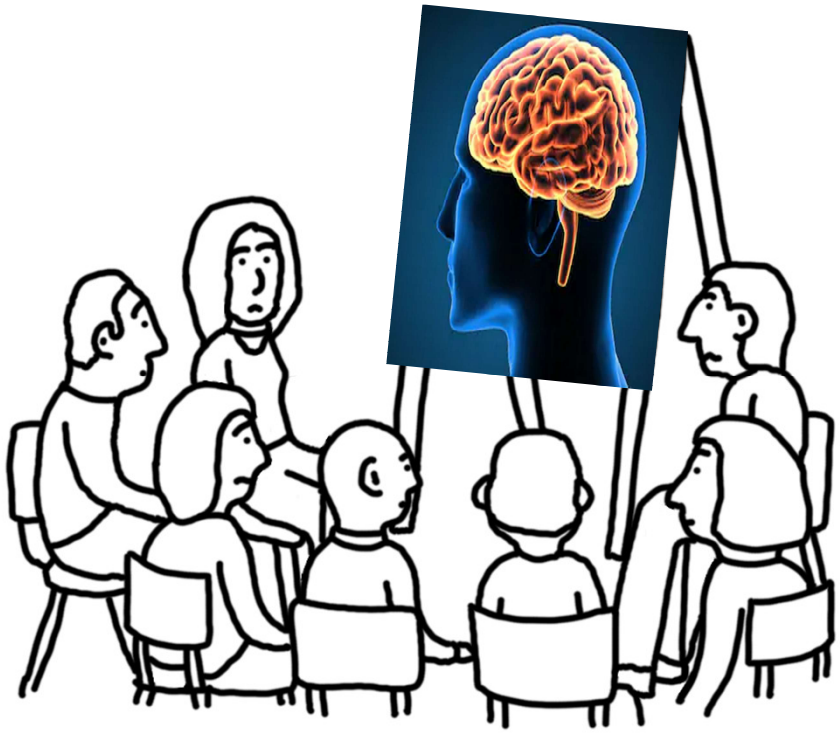
12-20W

- 1 kg
- 80B neurons
- 100T connections

250W

- *Cox, D. D. & Dean, T. Neural networks and neuroscience-inspired computer vision. Curr. Biol. 24, R921–R929 (2014)*
- *Mink, J.W., Blumenschine, R.J., and Adams, D.B. (1981). Ratio of central nervous system to body metabolism in vertebrates: its constancy and functional basis. Am. J. Physiol. 241, 203–212*

Brain Inspiration



Taking biological brains as a guide appears as a natural research direction toward more energy-efficient forms of computing

The Silent Drift

Global drive to cut energy use in edge devices

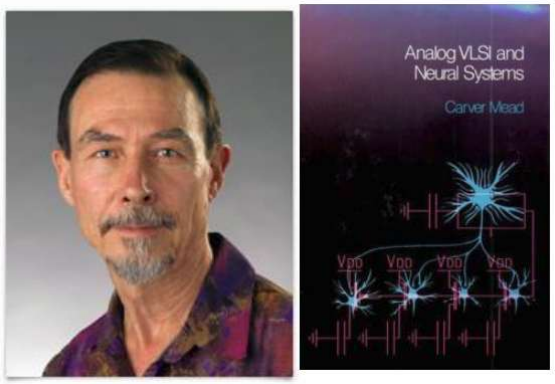
- Minimizing access to centralized memories
- In-memory computing architectures
- Always-on wake-up controllers
- Weight and activation quantization
- Approximate computing.

.....

<<The field is thus naturally trending toward key properties of biological neural processing systems: processing and memory co-location, event-driven processing, low-precision computation with binary spike encoding,…>>

Charlotte Frenkel, David Bol, Giacomo Indiveri :
Bottom-Up and Top-Down Approaches for the Design of Neuromorphic Processing Systems:
Tradeoffs and Synergies Between Natural and Artificial Intelligence. Proc. IEEE 111(6): 623-652 (2023)

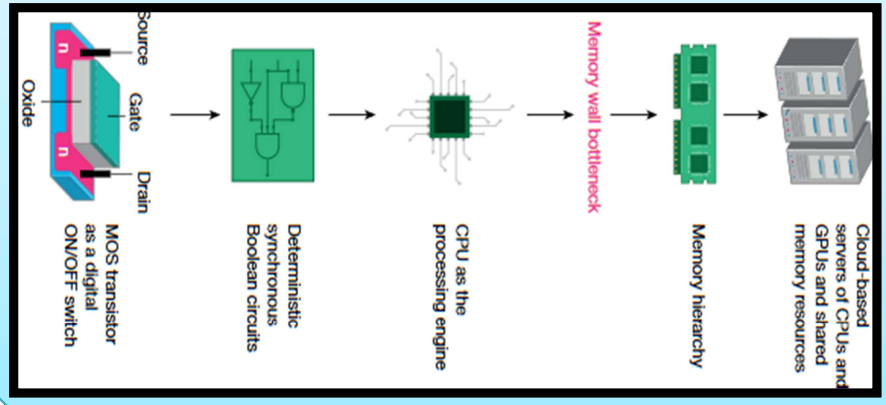
Working More Like the Brain



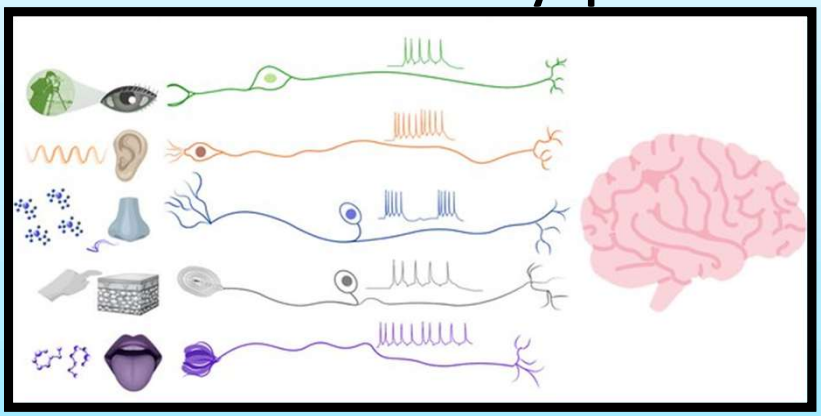
Carver Mead, 1989
Direct emulation of the brain ion channel dynamics could be performed by the MOS transistor operated in the subthreshold regime

Neuromorphic Engineering: paradigm shifts

Distributed computation that co-locates processing and memory



Data encoding both in space and time with all-or-none binary spike events



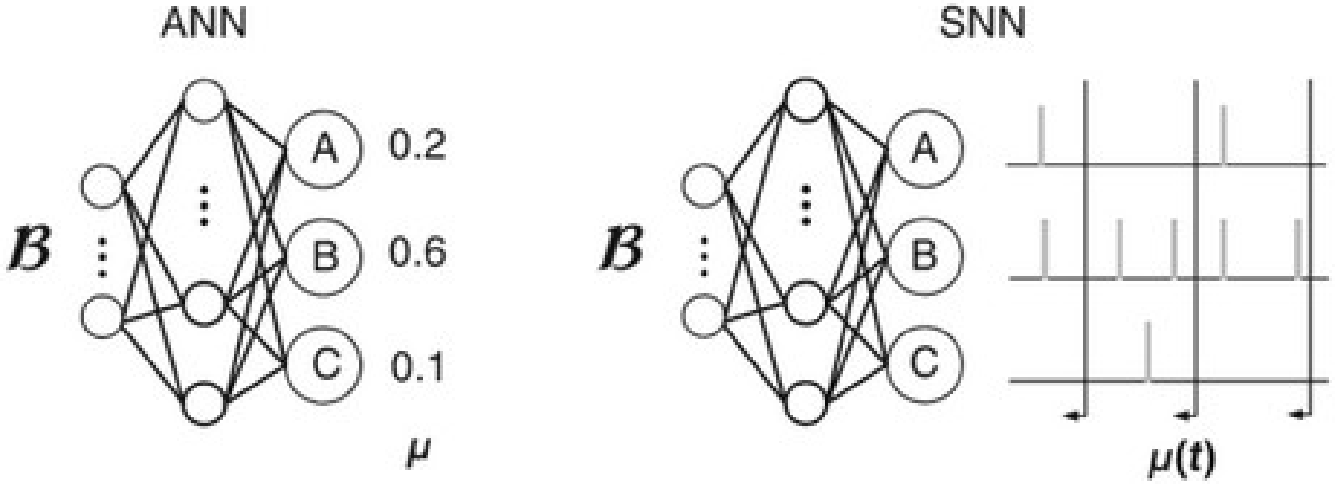
Often not fully attained in actual neuromorphic hardware..

Bio-Inspired Model of Computation

Spiking neural networks:

- Computations are triggered by events
- Energy efficiency associated with the sparsity of events in time and space

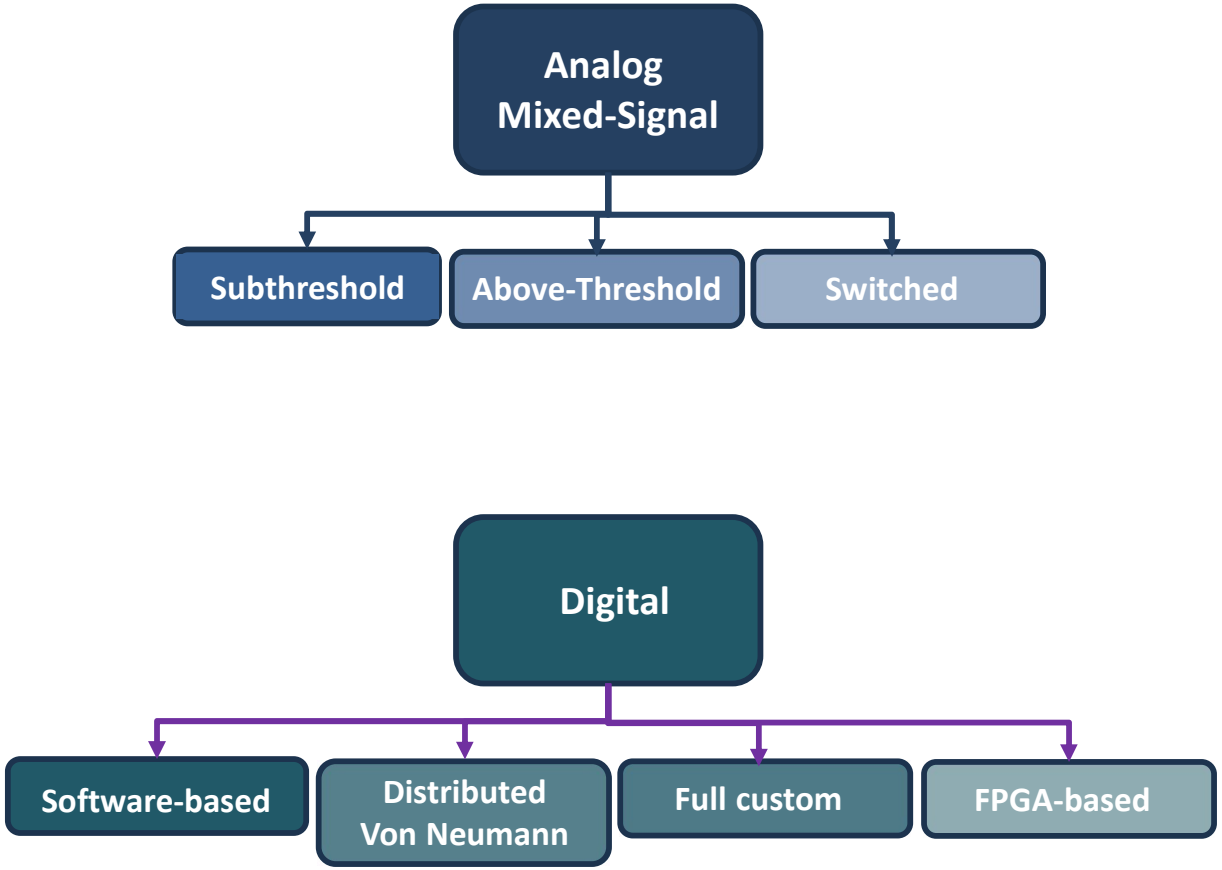
Sparse event-triggered processing toward reduced power consumption



Source:
Nan Zheng,
Pinaki Mazumder

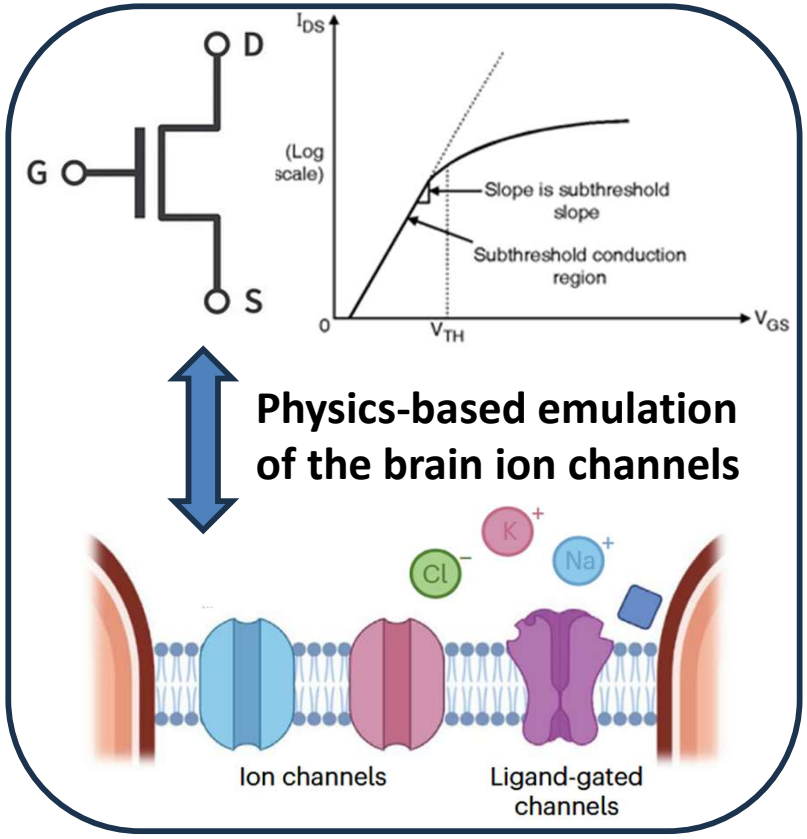
How to support this model of computation in hardware?

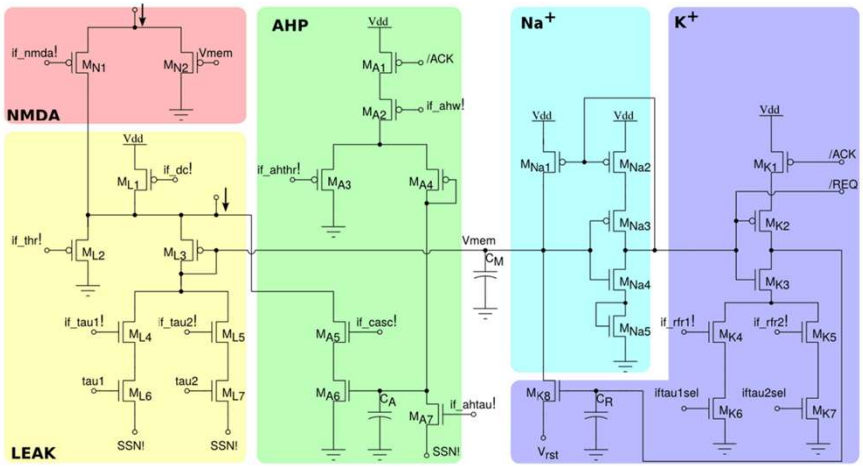
A Plethora of Circuit Design Styles



Circuit Design Styles - Analog

An emulation approach directly grounded in the physics of the silicon substrate



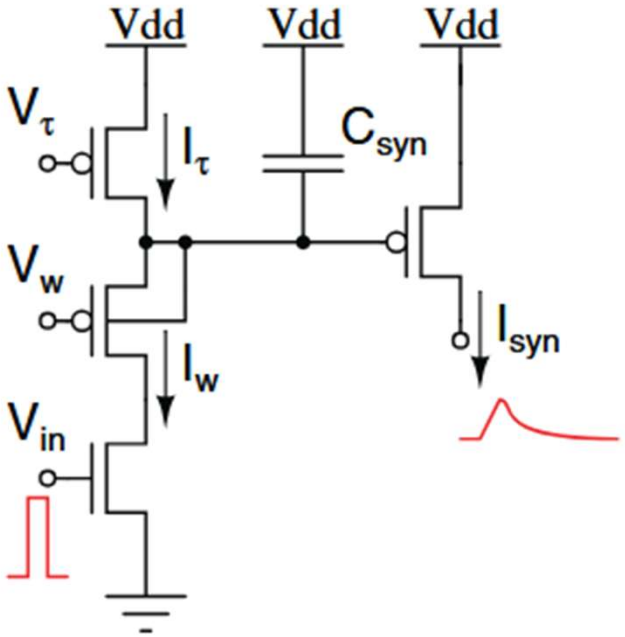


Silicon neuron

Potentially: compact and ultra-low power devices that lie close to brain bio-physics

Time Constants

Subthreshold analog design allows for the emulation of a large range of neuronal behaviors and synaptic dynamics with few transistors



- Time constants of ~ msecs easily achievable
- Similar to those sensed from the environment or to biological stimuli

Subthreshold design inherently well-suited for real-time and closed-loop processing of natural signals

Analog Circuits – Critical Review

Time constant

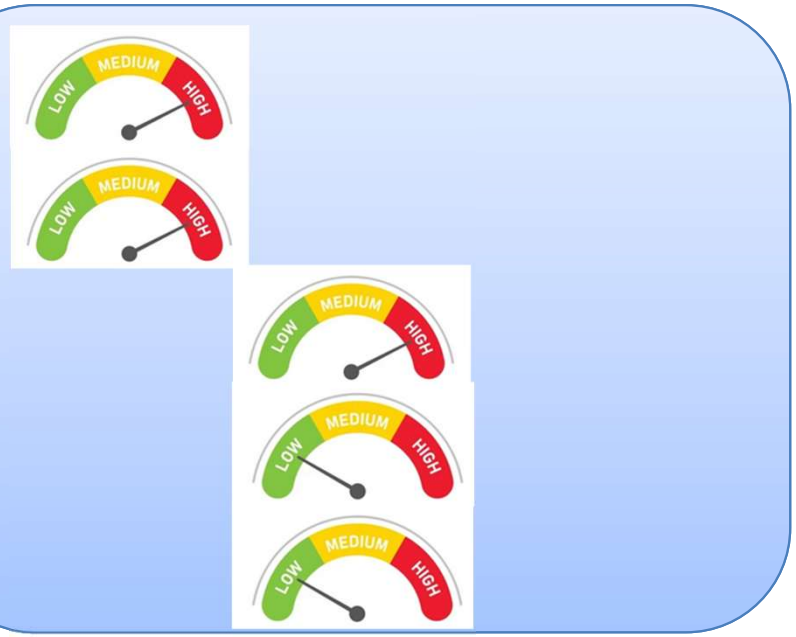
Noise, mismatch, PVT sensitivity

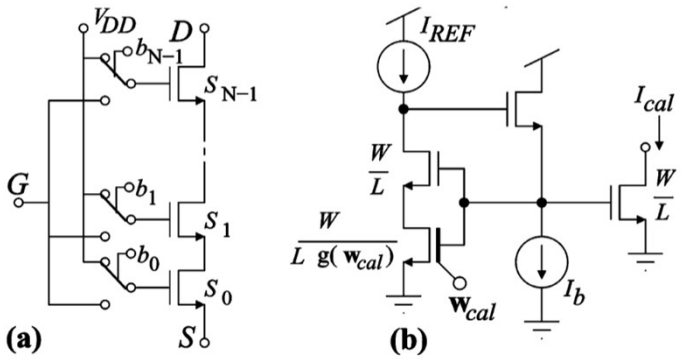
Design time

Technology scaling

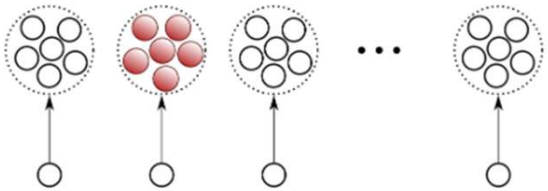
Programmability

Subthreshold

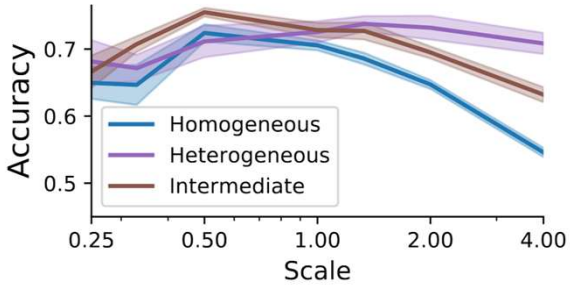




Calibration circuits



Redundancy



Robust learning

- Efficiency is affected.
- However, variability can be exploited for robust learning.

Above Threshold Analog Circuits

Above-threshold (model-based)

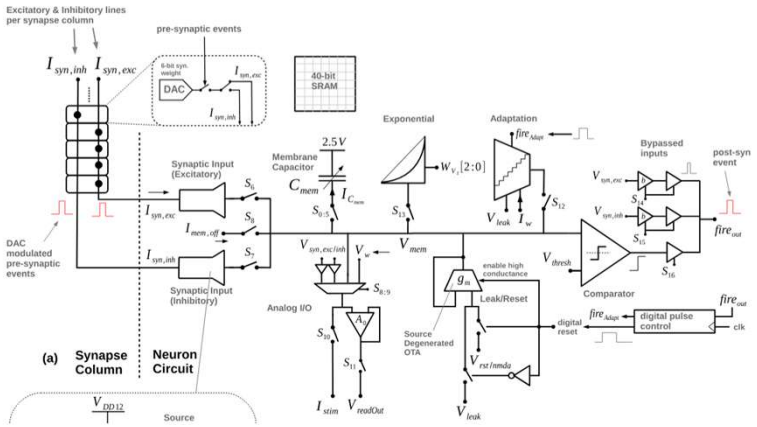
Time constant

Noise, mismatch, PVT sensitivity

Design time

Technology scaling

Programmability

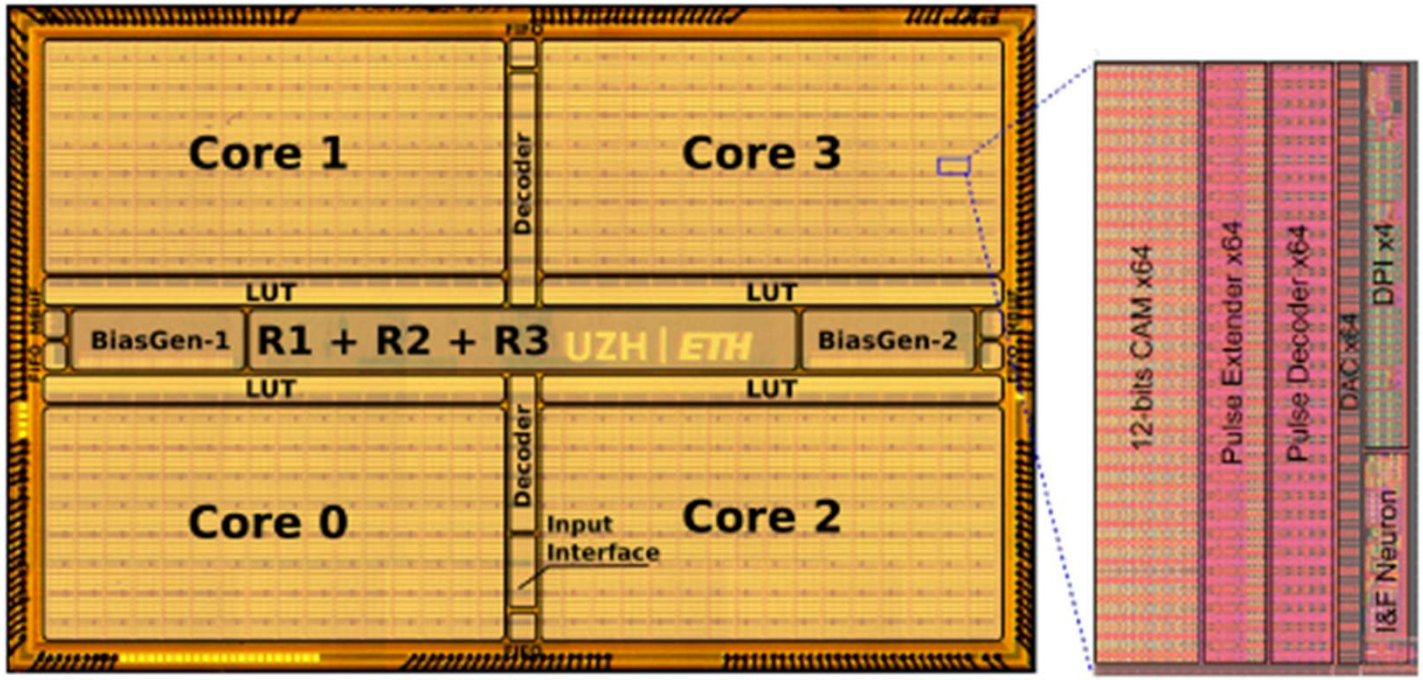


Analog circuits are designed for each term of the equations of the chosen neuron/synapse model

- Above-threshold analog design suited for accelerated-time modelling of biological NNs
- Emulation of neural processes cannot take place anymore at the level of the device physics (drift vs. diffusion)

DYNAP-SE

Multi-core neuromorphic processor
with analog neurocores and asynchronous digital communication

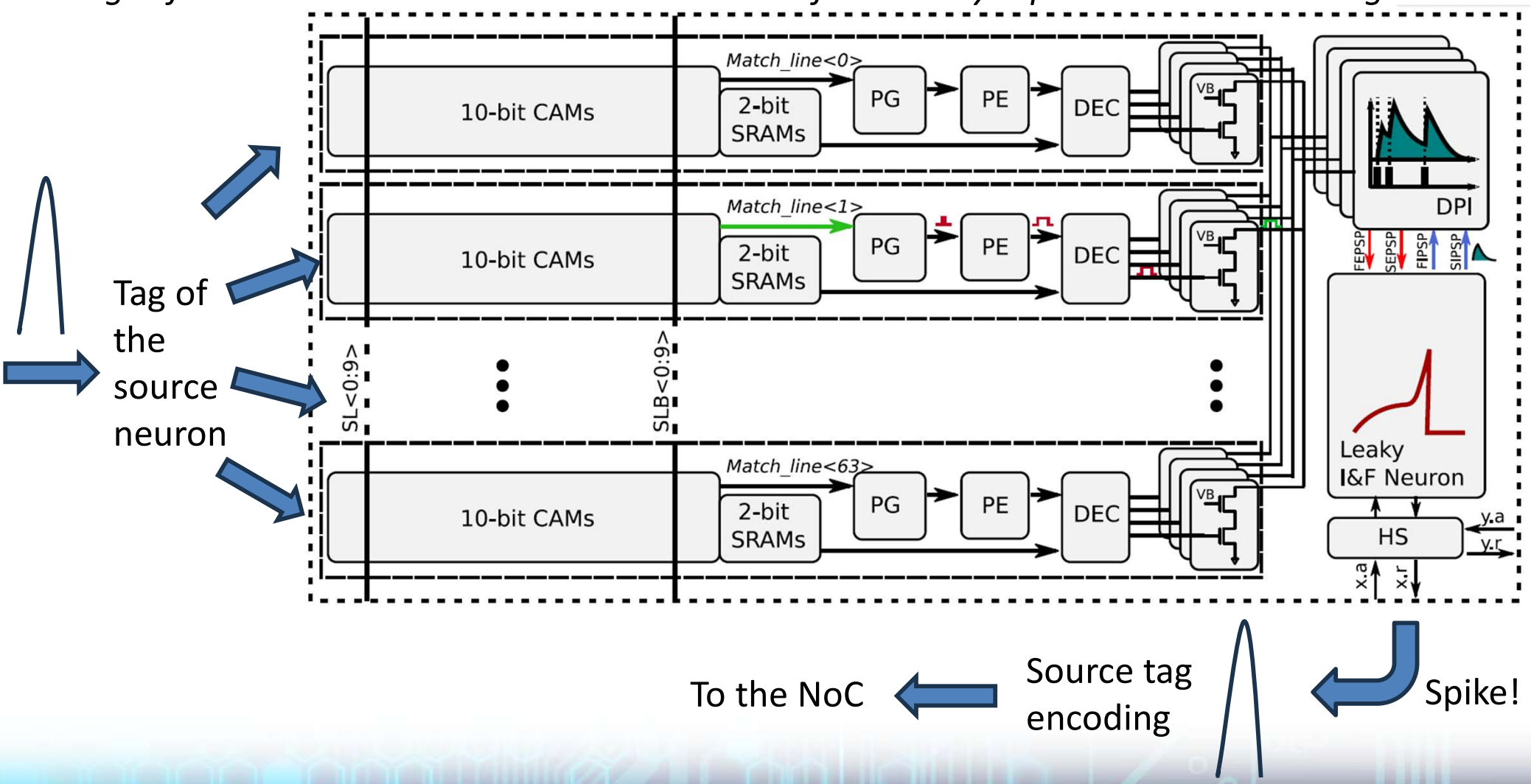


- *0.18um*
- *4 cores, 256 silicon neurons/core*
- *Adaptive-Exponential Integrate and Fire (AdExp-I&F) neuron model*
- *Overall, 1k neurons and 64k synapses*
- *On-chip routing of spikes via R1+R2*
- *Off-chip routing of spikes via R3*
- *Edge computing applications*
- *Typical power: ~5 mW*

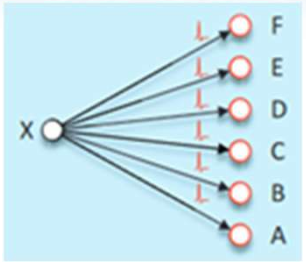
Moradi et al., "A scalable multicore architecture with heterogeneous memorystructures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs),"IEEE Transactions on Biomedical Circuits and Systems, vol. 12, no. 1, pp. 106-122, 2018.

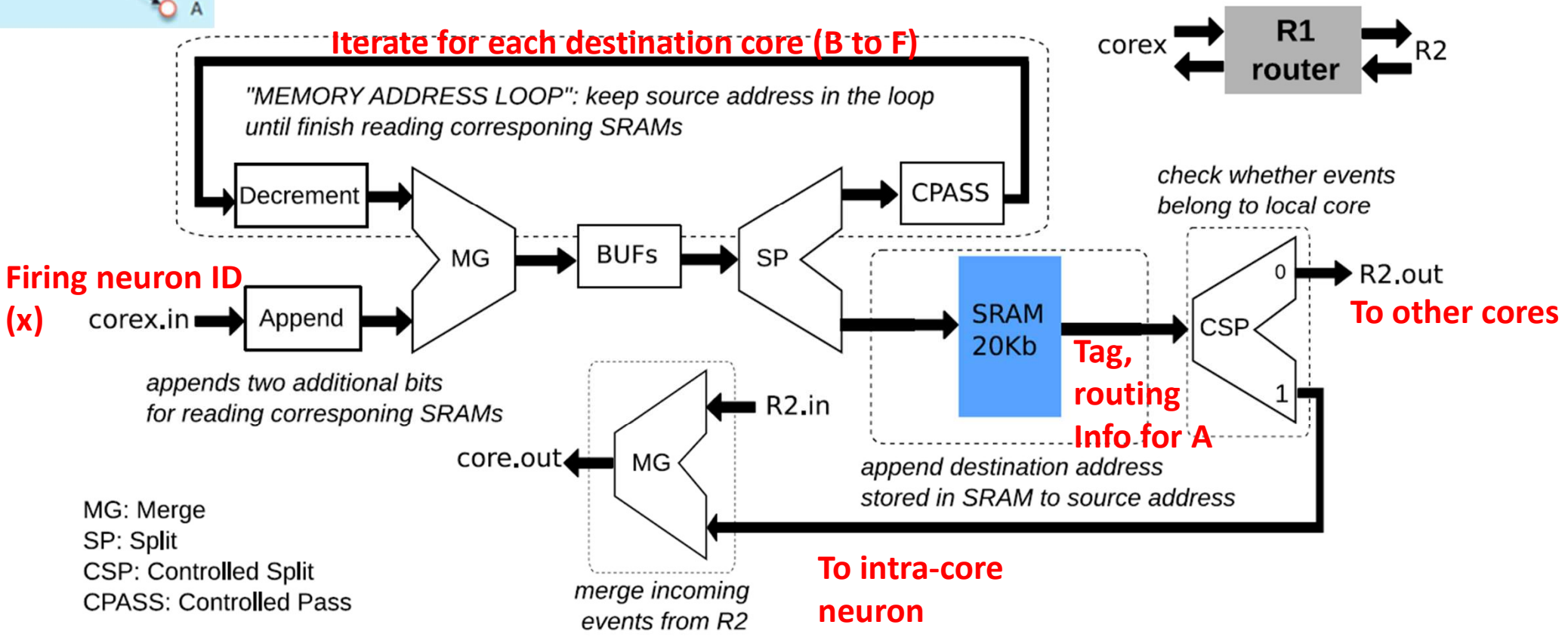
DYNAP-SE – Silicon Neuron

- Circuit architecture of one of the 256 computing primitives per core



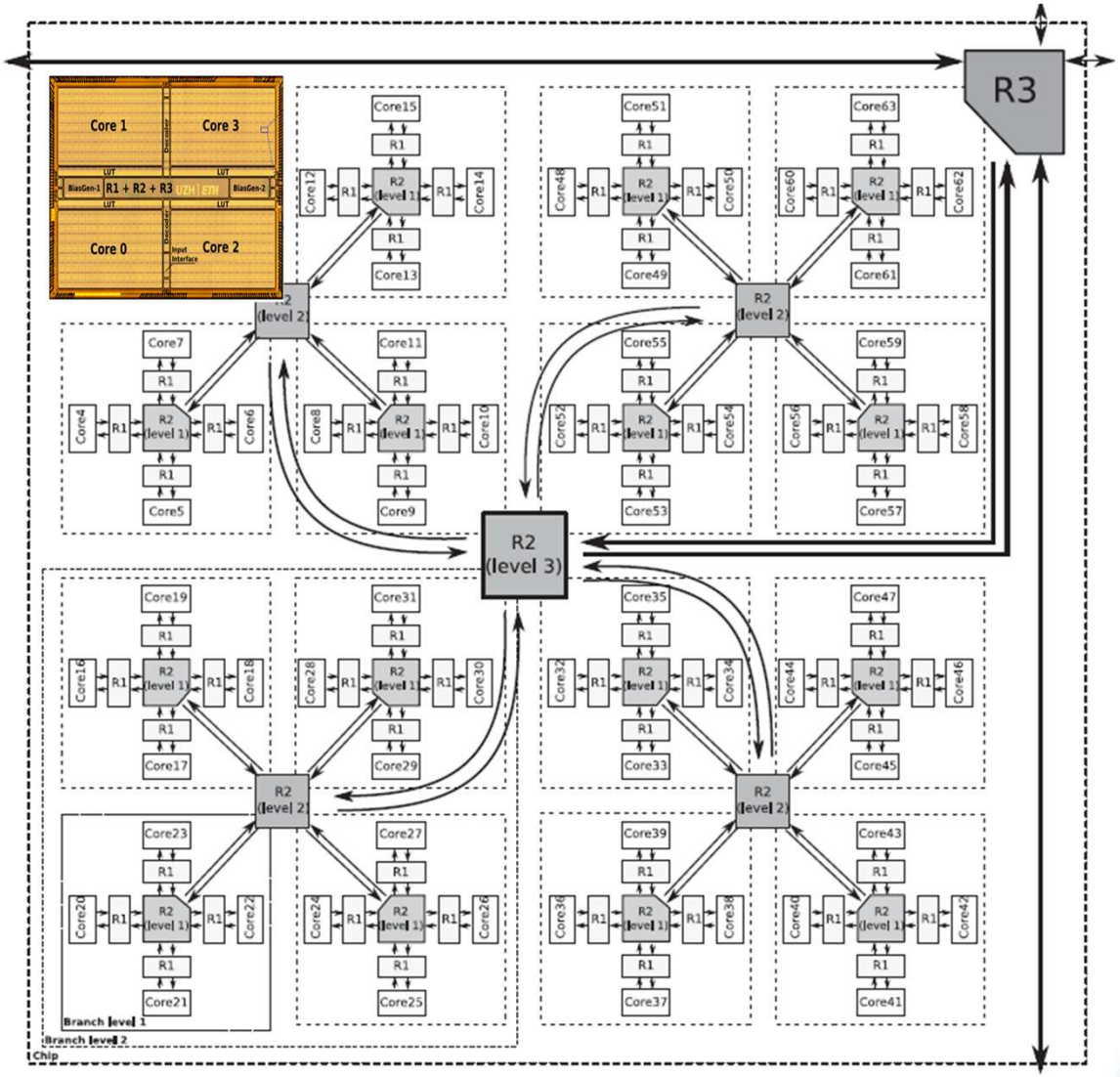
DYNAP-SE – Router R1





- Designed with 4-phase QDI asynchronous logic.
- No support for multicast: packet replication!

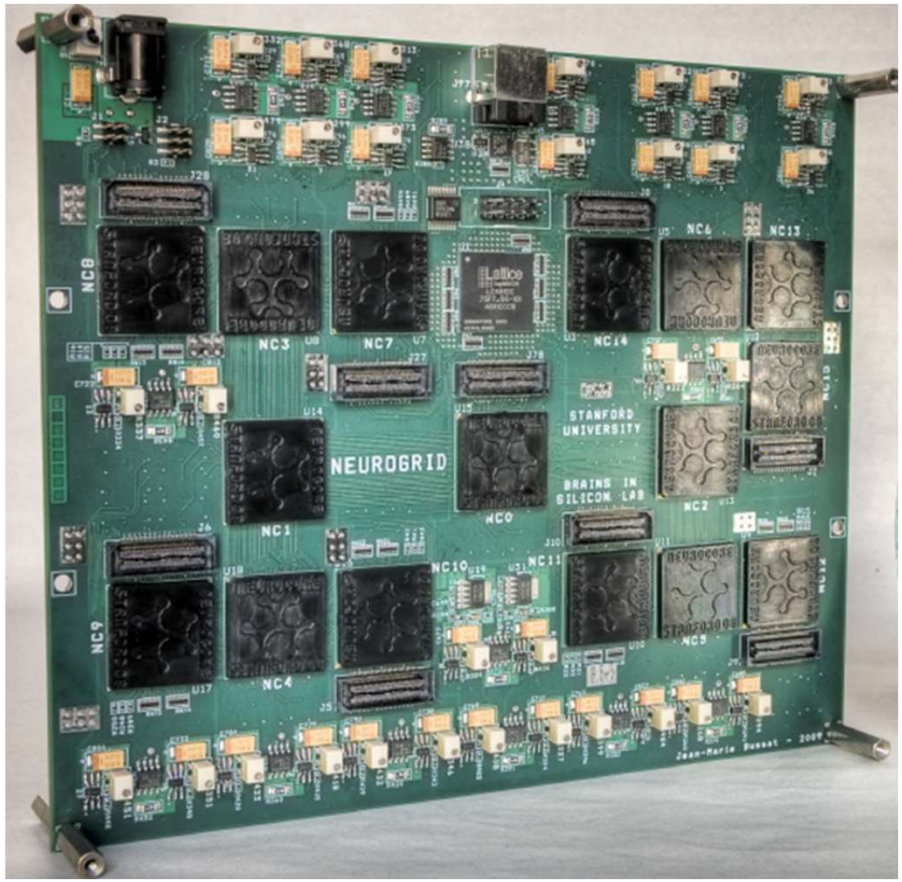
DYNAP-SE – Scalability



How are address events sent to different cores?

- One R1 router per core
- Hierarchical communication scheme
  - Low latency
  - Scalability
- R3 router to connect to other chips
- Only small part implemented
- A new generation of DYNAP-SE is coming (NeoCorAl)

Neurogrid

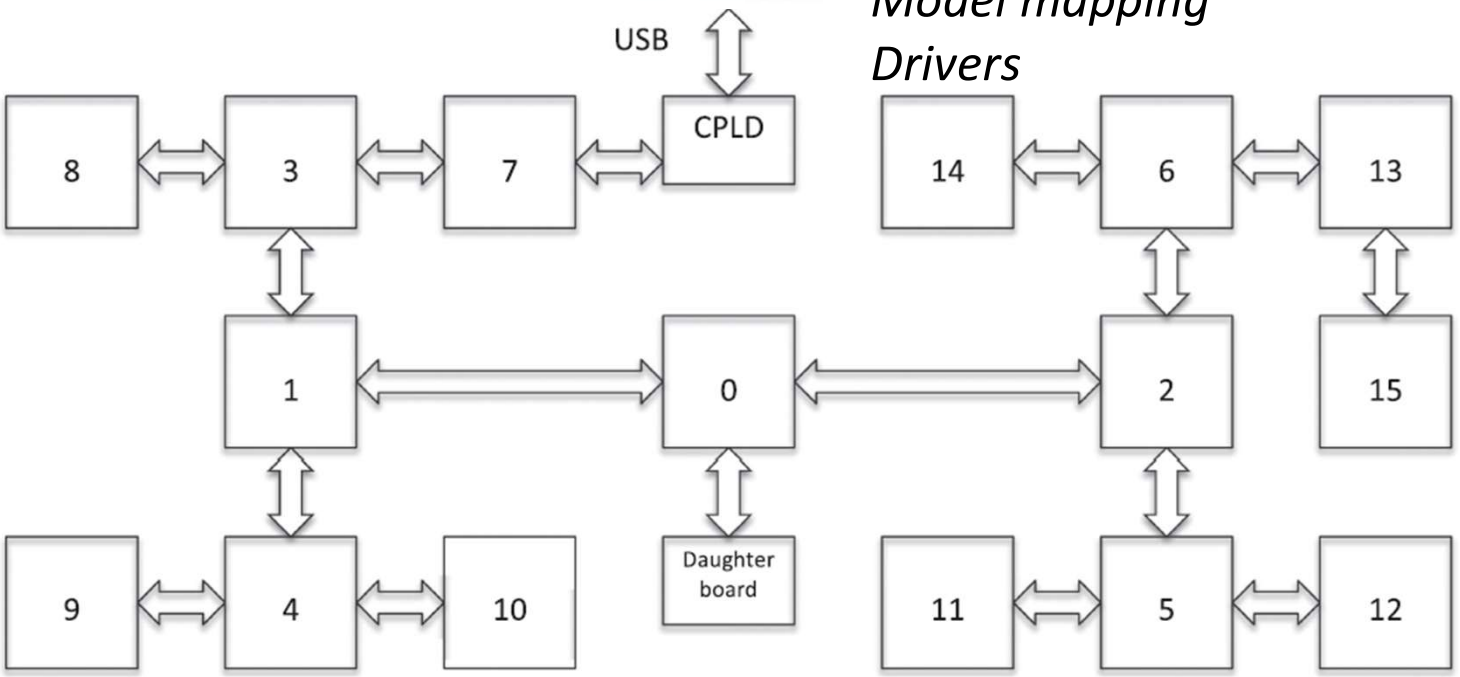


0.18um
16 analog neurocore chips,
64k silicon neurons/chip
Overall: 1M neurons, 6 billion synapses
Off-chip multicast-enabled networking
Brain emulation applications (e.g., control of prosthetic limbs)
Power: ~ 3 Watts

B. V. Benjamin *et al*., "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," in *Proceedings of the IEEE*, vol. 102, no. 5, pp. 699-716, May 2014.

Neurogrid Structure

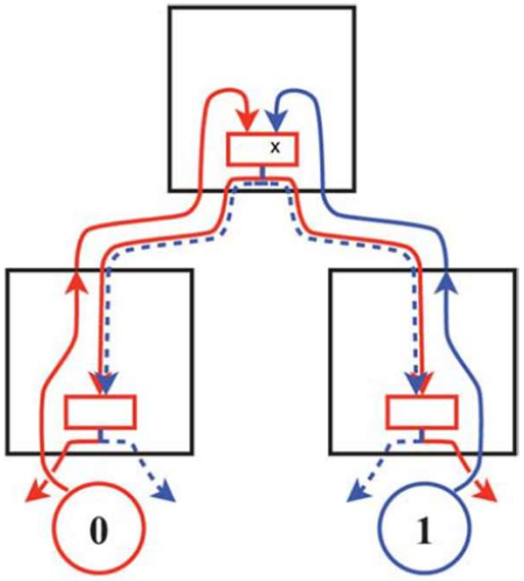
Model definition
Visualization and runtime interaction
Model mapping
Drivers



- Binary tree where each Neurocore chip (e.g. chip 1) connects to one parent (0) and two child chips (3 and 4).
- Binary tree used for spike notifications

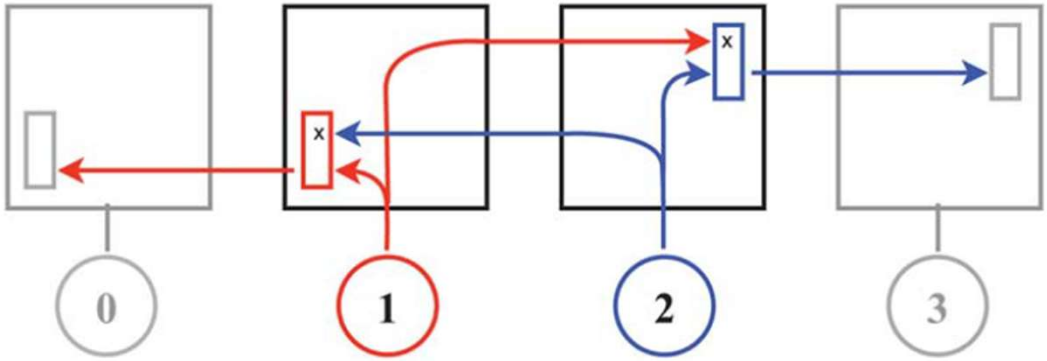
Communication Fabric

- Binary tree topology with a routing algorithm that supports local broadcast (i.e., multicast) through branching
  - Prone to deadlock



Branching enabled only in the descending phase (i.e., after merging)

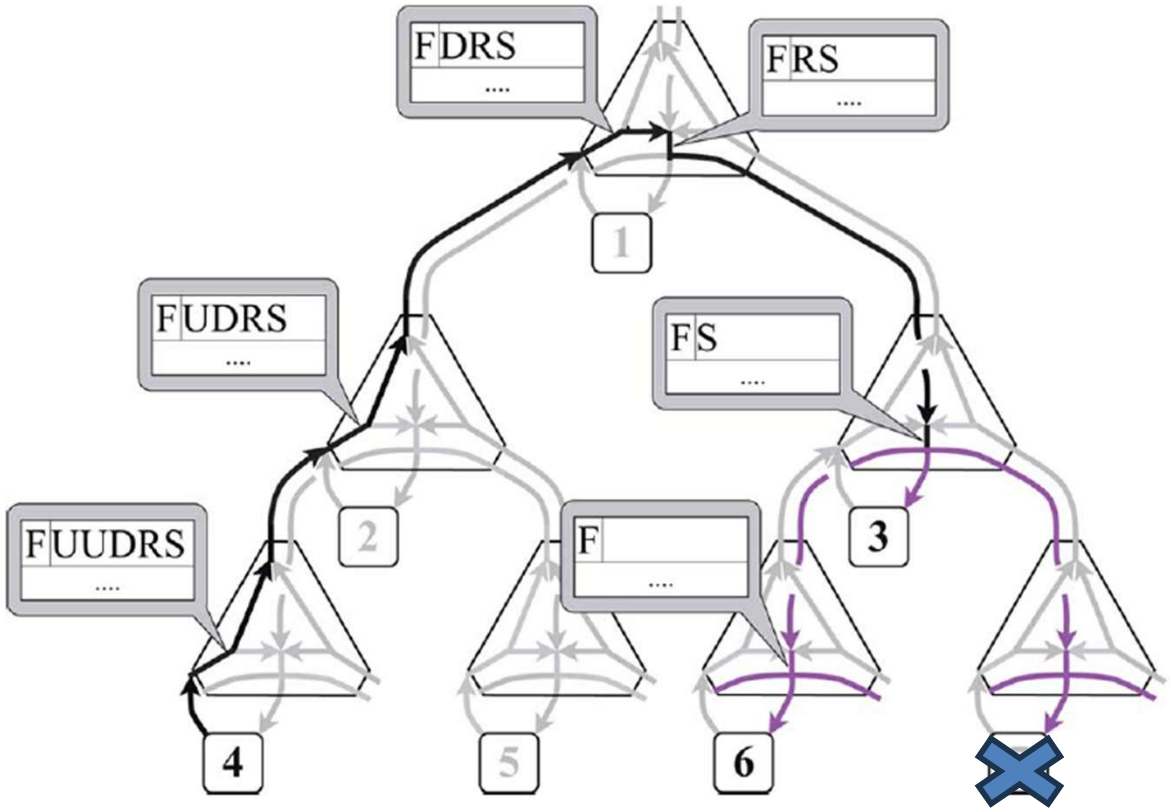
1 locks West vs. 2 locks East



Branching precedes merging

- Neurogrid avoids deadlock by reversing the order in which branching and merging occur.

Routing Algorithm



Unintended destinations filter packets out

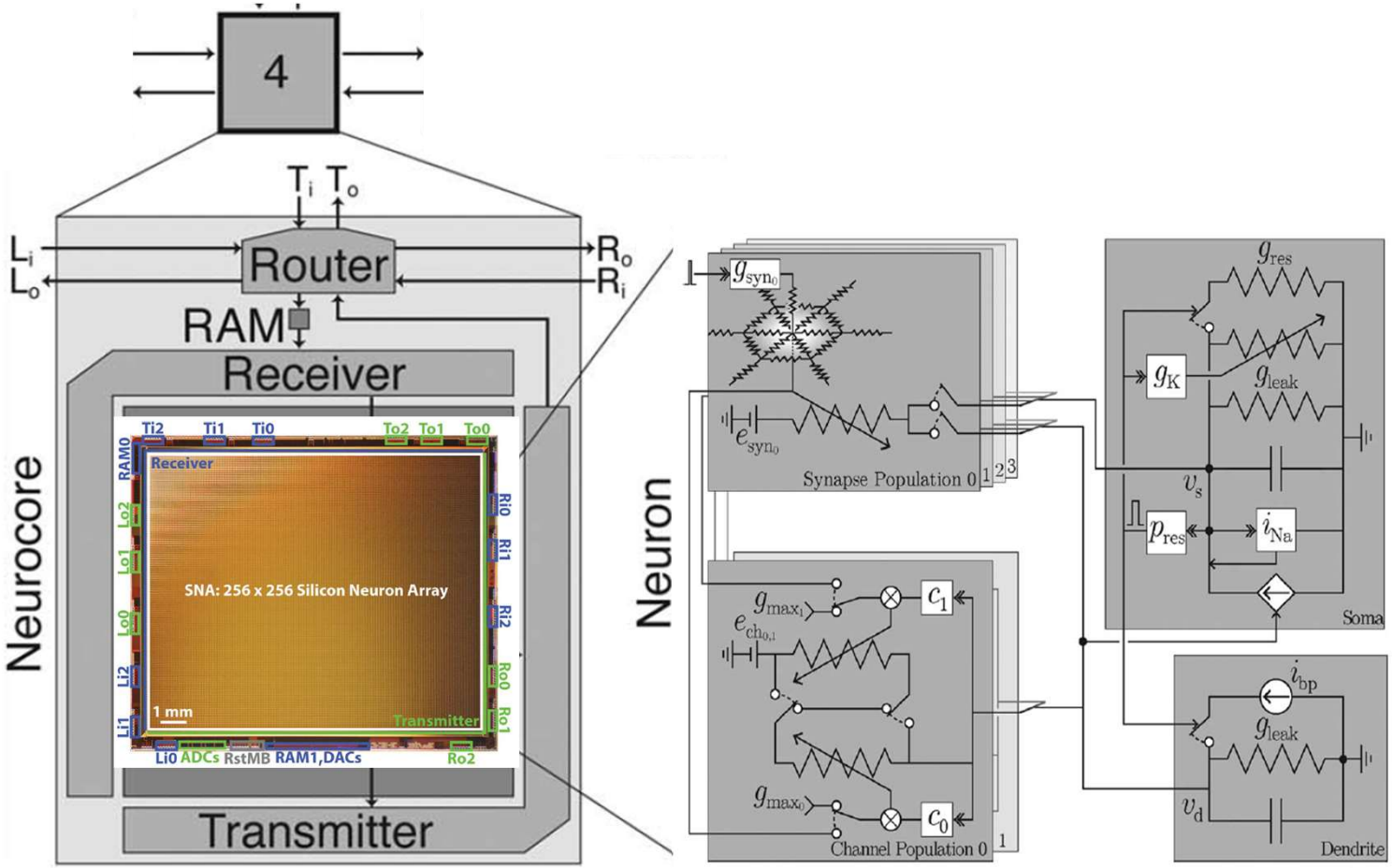
3-phase routing:

- **Ascending point-to-point routing** to the lowest common ancestor of source (4) and all destinations (3,6)
- **Descending point-to-point routing** to the lowest common ancester of all destinations
- **Local broadcast**

- Memory lookups avoided while branching
- Simple and short packet header
- **Filtering energy can be minimized through ad-hoc mapping frameworks**

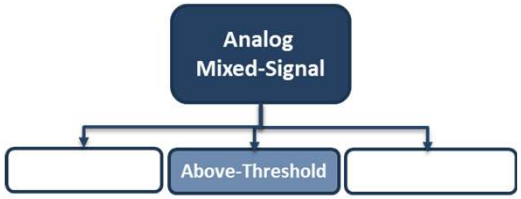
Neurogrid - Neurocore

Neurogrid uses subthreshold analogue circuits to model neuron and synapse dynamics in biological real time, with digital spike communication

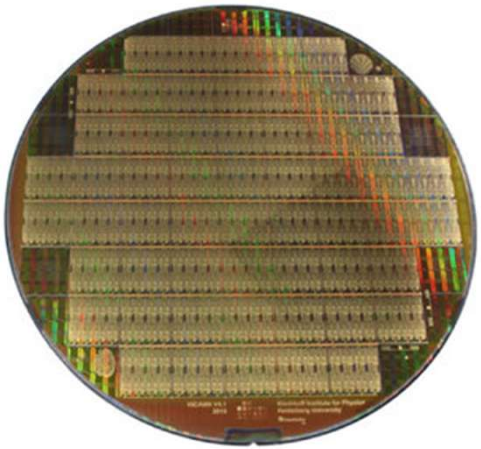


Multi-compartment neuron model that simulates ion channel dynamics to capture rich neuronal behaviors

BrainScaleS



It follows a model-based approach to neurosynaptic dynamics for time-accelerated neuroscience simulations



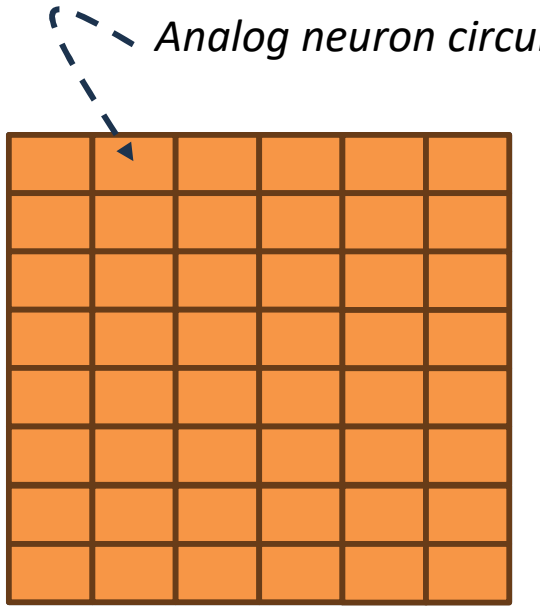
Wafer-scale integration
0.18um
352 chips/wafer,
180k neurons
40M synapses
Accelerated neuroscience simulation
(up to 10000 faster than real time)
1.3W (chip)
500W (board)

20 wafer modules organized in 5 racks
4M neurons, 1B synapses
10kW

Currently developing the second generation

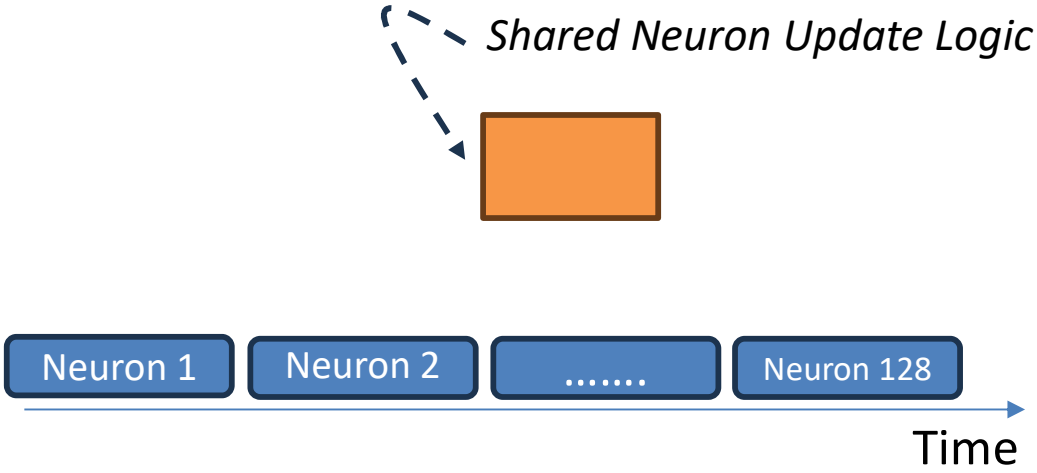
Circuit Design Styles - Digital

They “simulate” (not emulate) neural processes by relying on circuit implementations that lie far from physics but take advantage of the speed of the silicon substrate



Analog: spatial redundancy

Time constants in the msec range are achieved by tuning physical parameters.

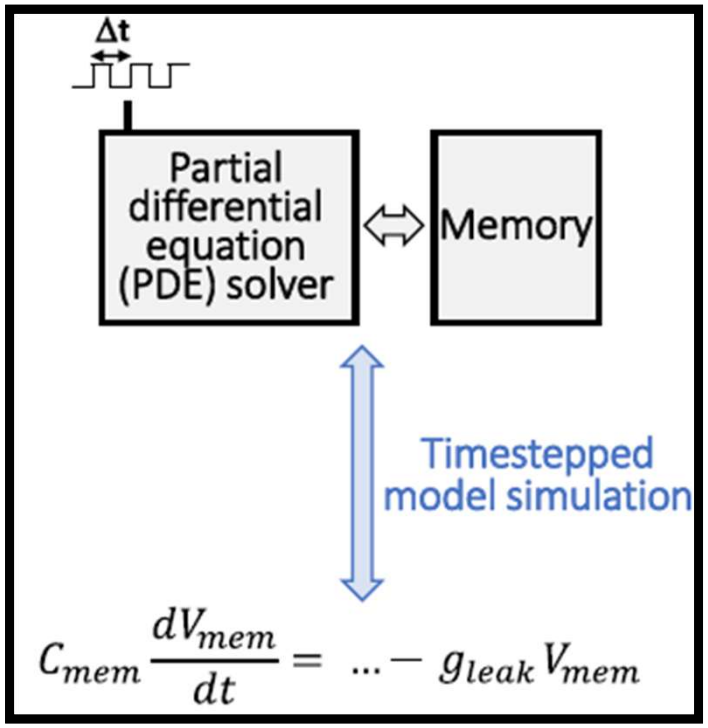


Digital: time-division multiplexing

The high clock speed allows time-division multiplexing of many neurons while maintaining biologically-realistic update rates

Solver-Based Digital Design

PDE solvers numerically solve the chosen neuron mathematical model



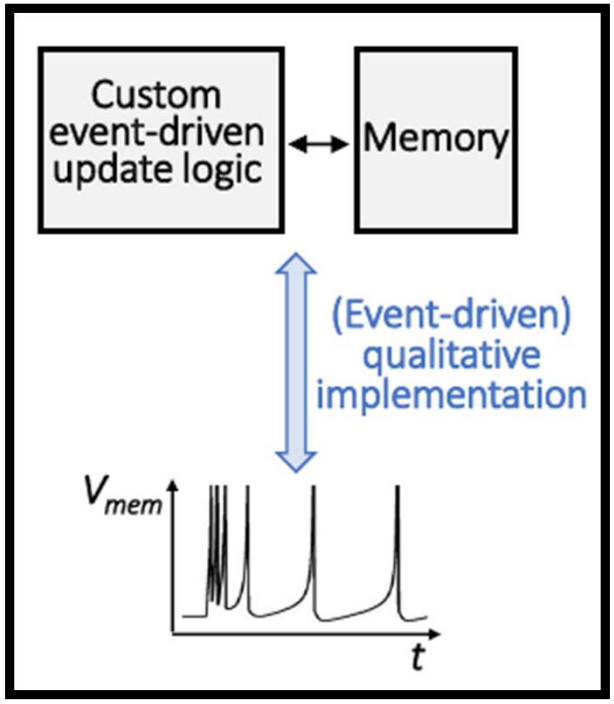
This comes at the cost of:

- *time discretization*
- *The PDE solver must fetch/update model state from/to memory at each timestep Δt.*
  - ✓ *Even without relevant activity in the network*

The extensive data movement hardly matches the low-power event-driven neuromorphic approach

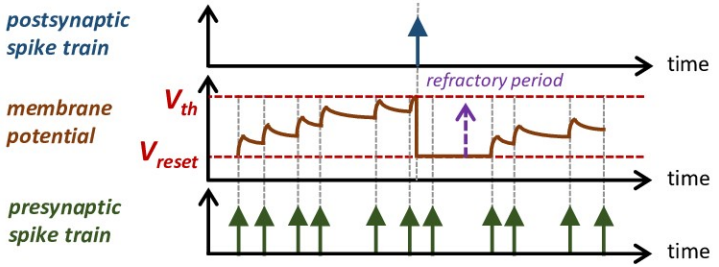
Phenomenological Digital Design

Aims at carrying out updates only where and when relevant in the neural network



Two main approaches:

- *The detail level of biophysical modeling can be reduced and the model simplified*
  - ✓ *E.g., Leaky integrate-and-fire neuron model*



- *Low-cost reproduction of firing behaviour without using model equations (e.g., through custom update logic)*

Relaxed model requirements enable event-driven state updates, thus strongly reducing data movement overhead

Circuit Design Styles - Digital

Rely on circuit implementations (that lie far from physics) to simulate neural processes

Time constant

Noise, mismatch, PVT sensitivity

Design time

Technology scaling

Programmability

Solver-based and Event-driven

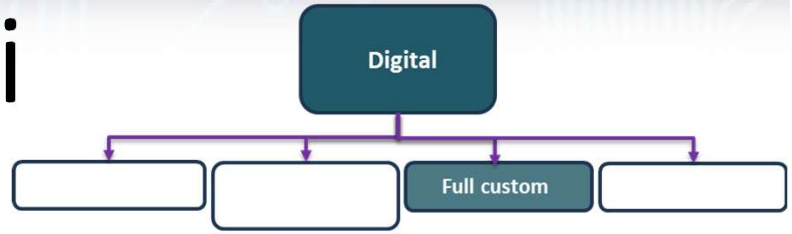
- Predictable behaviour and one-to-one correspondence with the simulation software
- Competitive area and power efficiencies
- Flexibility

The Clock: a Digital Bottleneck

- A significant source of overhead is the clock tree (up to 20%–45% of the total power consumption)
- Asynchronous digital circuits avoid this clock tree overhead and ideally support the event-driven nature of spike-based processing
  - Maximal activity gating during idle periods
- Asynchronous logic is a widespread choice in the most impactful platforms, especially for spike communication
- Despite the benefits, its adoption is sporadic and does not constantly grow in time
  - Poor CAD tool support
  - Overly conservative design styles come with large area/power overhead
  - Less conservative approaches are gaining momentum that achieve major overall cost benefits – will they take off?

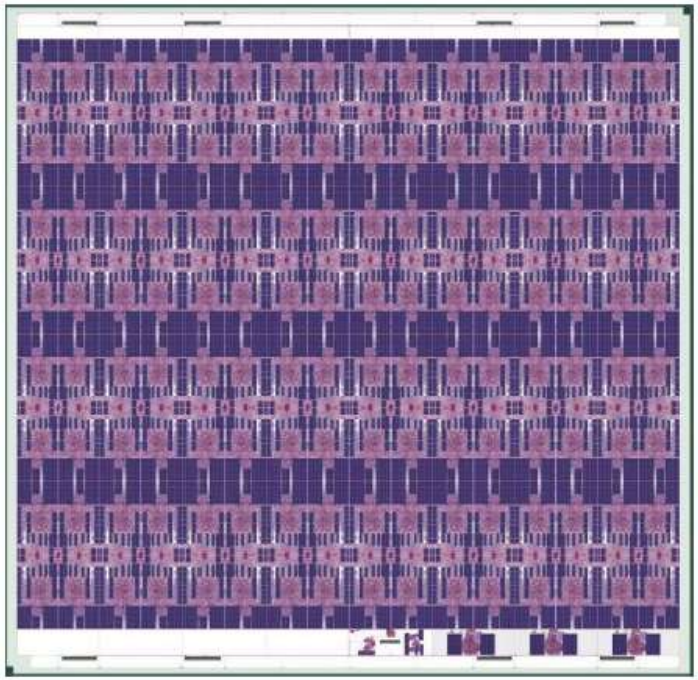
S. M. Nowick and M. Singh, "Asynchronous design - part 1: Overview and recent advances," IEEE Des. Test, vol. 32, no. 3, pp. 5–18, Jun. 2015.

Intel's Loihi



It is a neuromorphic manycore processor with on-chip learning capability

Intel's 14-nm process
Fully digital, fully asynchronous
2.07 billion transistors
33 MB of SRAM
(16 MB of synaptic memory,
2.5MB of routing tables)
60 mm2
Extendable in 4 planar directions



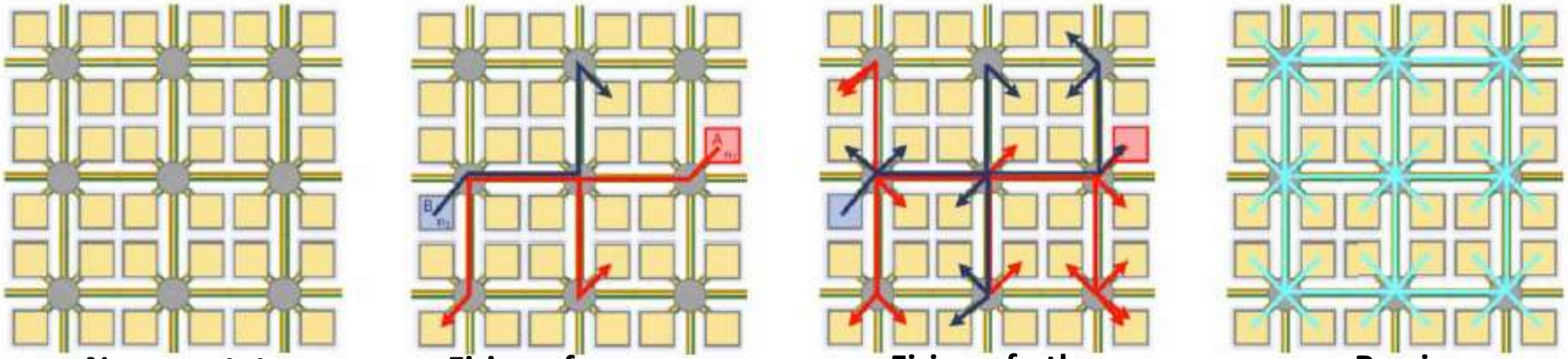
128 neuromorphic cores
3 x86 control cores
Overall, 128k neurons and
128M synapses
Cognitive computing apps.
~1.5W

M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," in IEEE Micro, vol. 38, no. 1, pp. 82-99, January/February 2018

Mesh Operation

- Even though Loihi uses event-driven SNN simulation, the latter is still organized in discrete timesteps
  - Regular scheduling of events
  - Only “useful” work performed at each timestep

At each algorithmic time step, state variables are updated in a time-multiplexed way



Neuron state update | **Firing of neurons n1 and n2** | **Firing of other neurons from same cores** | **Barrier synchronization**

The network supports only unicast transactions;

multicast implemented through packet replication

Memory Requirements

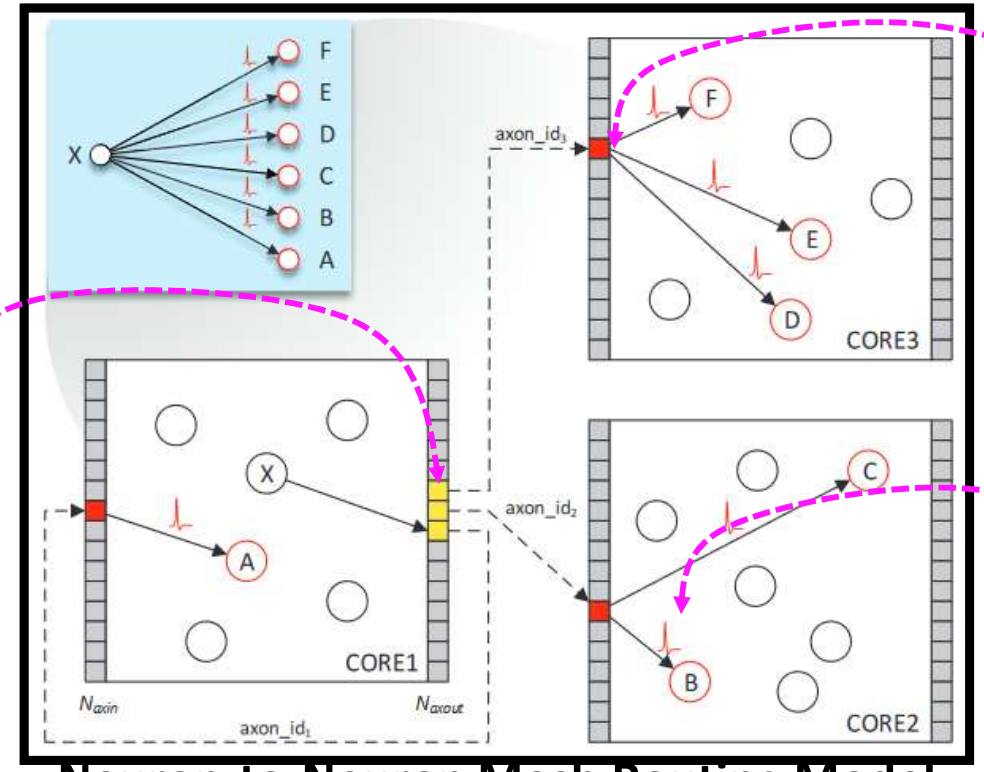
Limited on-chip storage resources pose network connectivity constraints

Neurons are mapped to cores based on each core's resource limits

=> Connectivity is programmed accordingly, if compatible with available storage resources

List of core-to-core edges

< 4096 entries



Neuron-to-Neuron Mesh Routing Model

Synaptic List

< 4096 entries

Synaptic memory

< 128 kB

In Loihi, the synaptic list size and the synaptic memory per core turn out to be the most limiting constraints

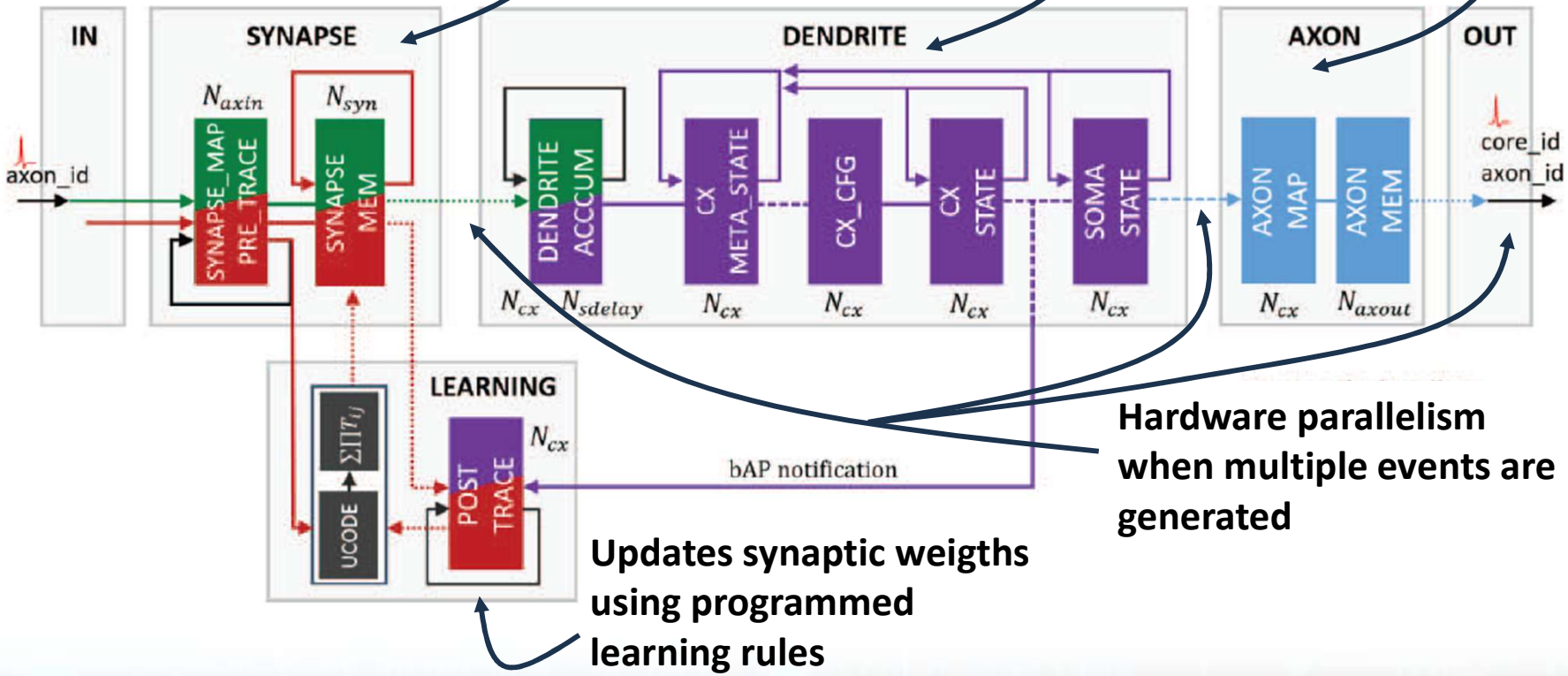
Loihi – State Update logic

- The neuromorphic core is **an asynchronous pipeline**
- It implements the **LIF neuron model (hardwired), with some flexibility**
- The 4 sections operate independently with minimal synch., at their own local rates

Processes incoming spikes and reads out synaptic weights from memory

Updates state variables of neurons in the core

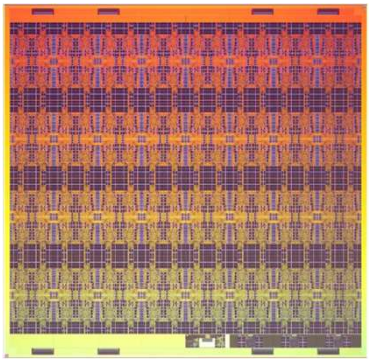
Generates spike messages for all fanout cores of firing neurons



Hardware parallelism when multiple events are generated

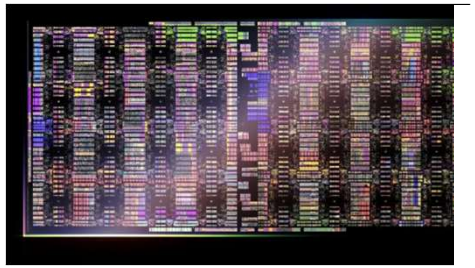
Updates synaptic weigths using programmed learning rules

Loihi Evolution



14 nm
128 neurocores
128k neurons
128M synapses
Digital, fully asynchronous
Cognitive computing
1.5W

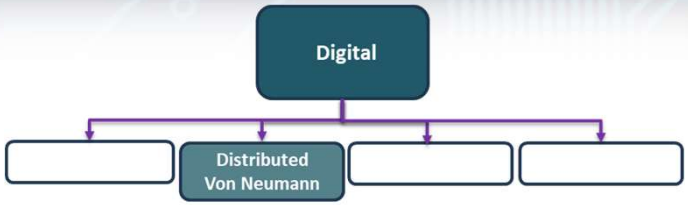
Pohoiki Springs
768 chips
100M neurons
100B synapses
500W



Loihi-2
7 nm
1M neurons
120M synapses
1W

1152 Loihi-2
1.15B neurons
128B synapses
2.6kW

SpiNNaker



Massively parallel digital supercomputer to model large-scale spiking neural networks

A million mobile phone processors

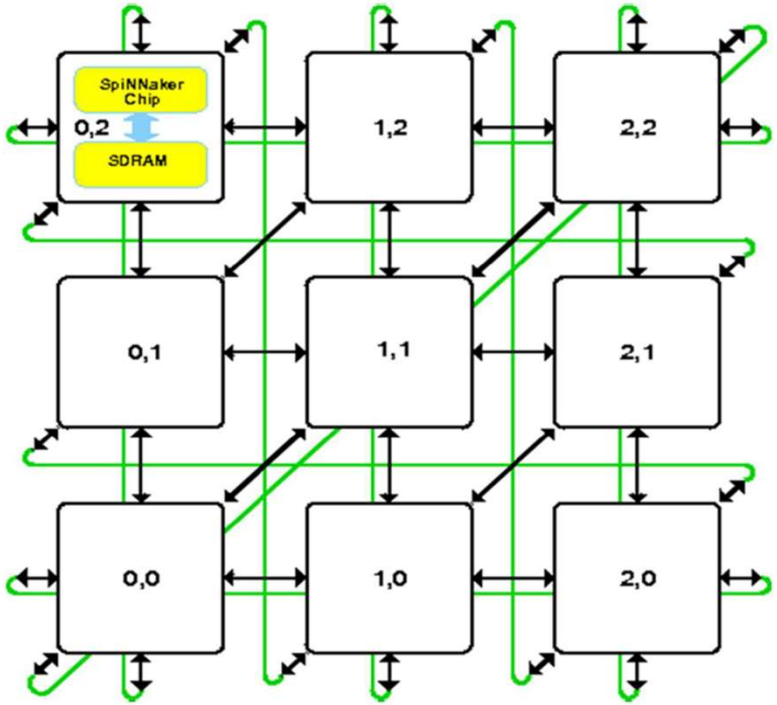
1 million ARM cores, 460M neurons, 460B synapses, 50kW (1% of human brain)

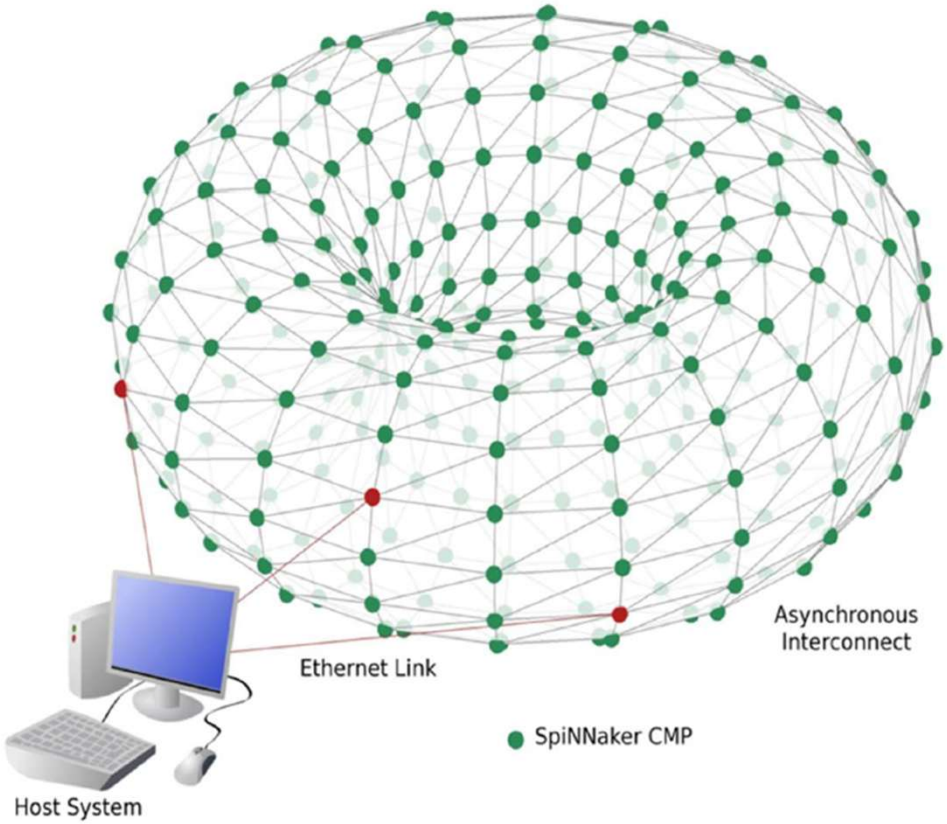
Differences with respect to a conventional supercomputer

Small integer cores from mobile/embedded designs, unlike the 'fat' cores in supercomputers

Interconnect tailored to the static multicast of small packets, unlike point-to-point dynamic routing

SpiNNaker Architecture





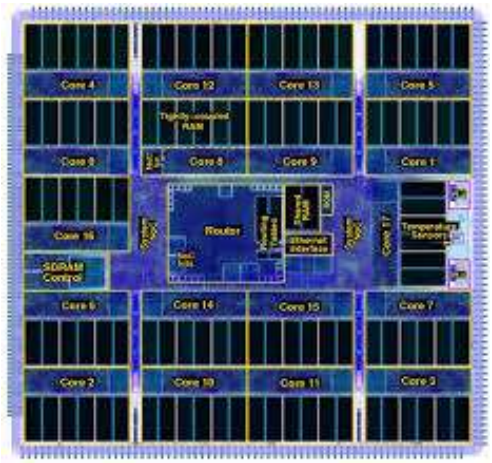
Compute nodes with
1 SpiNNaker chip + DRAM chip (co-packaged)

6 connection links per node provide rich connectivity

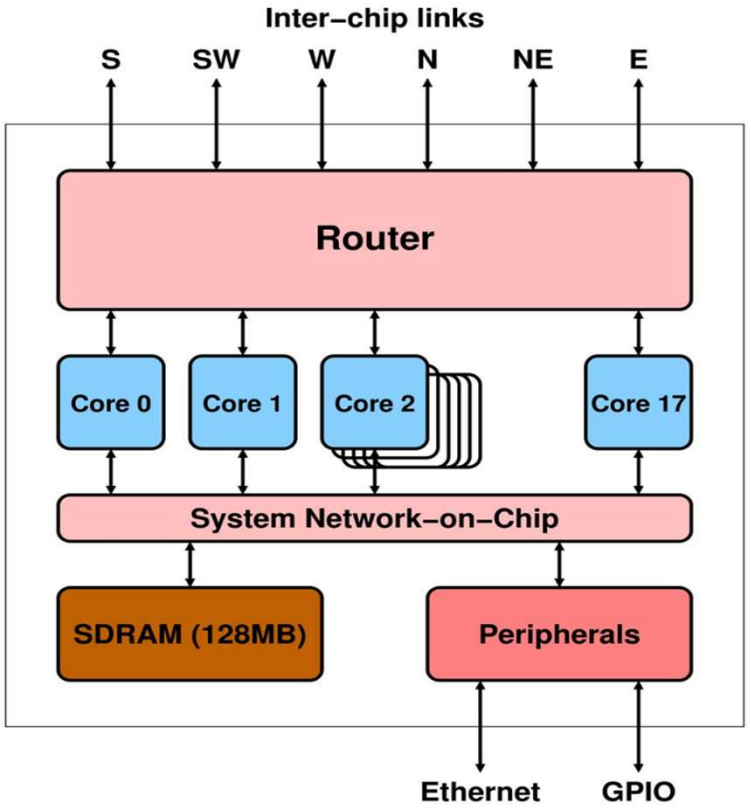
- Low latency, suitable for multicast, path diversity,..

Furber S. et al., «Overview of the SpiNNaker system architecture», IEEE Trans. Comput. 62(12), 2454–67, 2013

SpiNNaker1 Chip



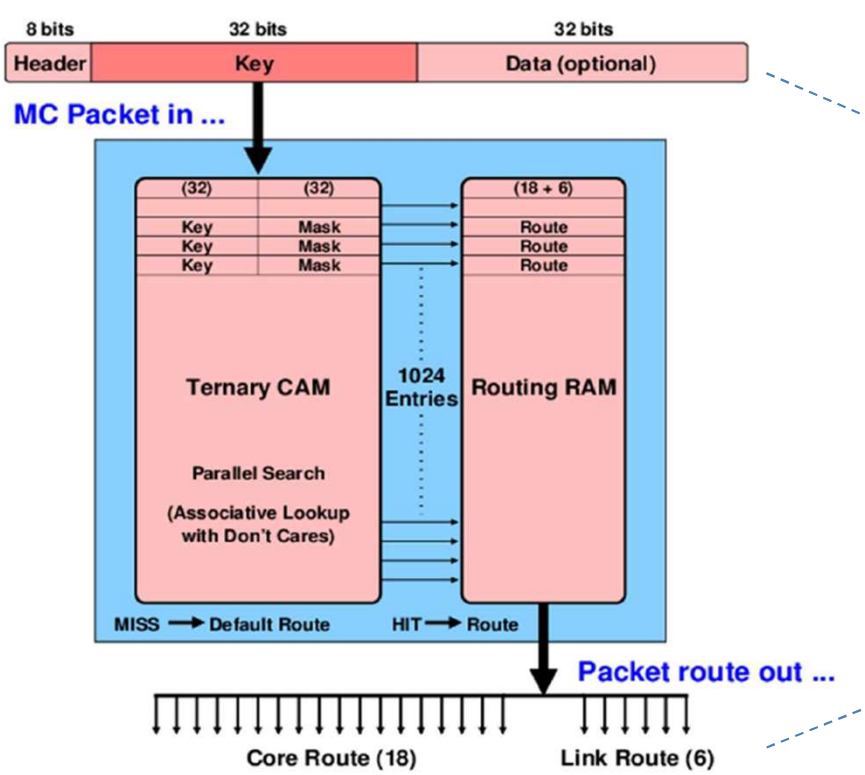
- 18 ARM9 cores
  - 200 MHz
  - 32-bit integer core
  - 130nm (Spinnaker2 in 22m)
- Each with 32kB of TC instruction memory and 64kB of TC data memory
- 16k neurons
- 16M synapses
- Neuroscience applications
- *~1W*

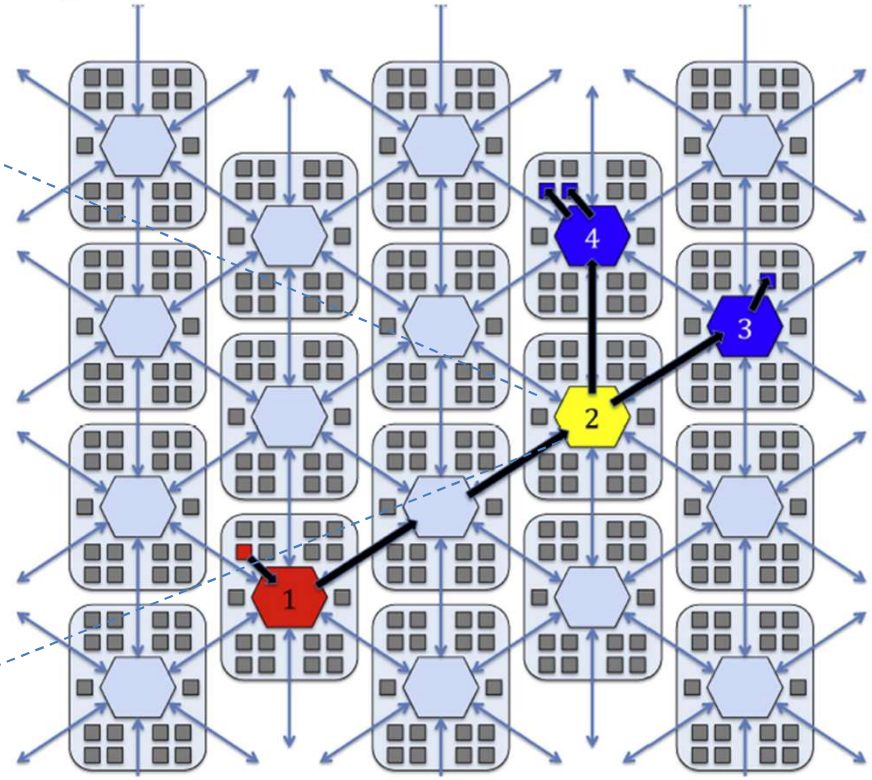


- Router: synchronous, with asynchronous wrapper for self-timed chip-to-chip links
  - Contains routing tables
- System NoC is asynchronous
- Cores are synch. Islands (GALS)

Multicast Router

- Accepts packets from 6 off-chip links and from the 18 on-chip cores
- Each neuron fires to a statically-defined multicast routing tree, which is programmed in the distributed routing tables of the routers





- The router is highly sophisticated: table compression, several routing engines, fault-tolerance,..

Conclusions

- AI systems at all scales are far from biological efficiency
- Neuromorphic computing (SNNs + custom-tailored hardware) is a promising way of thinking out of the box
- A plethora of circuit design styles
  - Communication is always digital
- Designing the integration scheme between processing and memory defines the resulting trade-offs