Neuromorphic Computing

a computer systems perspective

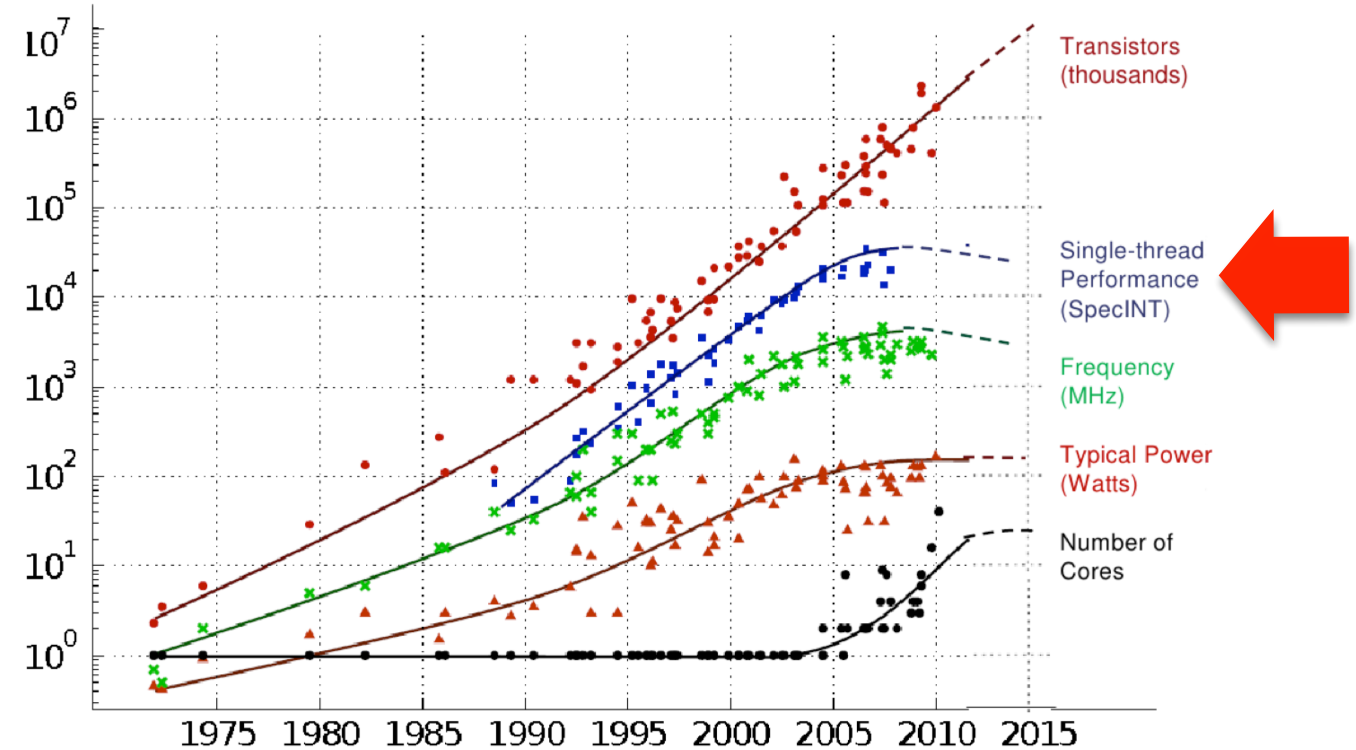
Rajit Manohar
Computer Systems Lab
Yale University
`http://csl.yale.edu/`
`http://avlsi.csl.yale.edu/`

Yale

AVLSI

The context: microelectronics scaling

- It's been a great ride…
  - ❖ … but sequential programs don't speed up each year like they used to in the "good old days."
- Computation demand is growing!
  - ❖ Massive amounts of data being collected by cheap, ubiquitous sensors.
  - ❖ ~ 1.5B smartphones (with cameras) shipped in 2017.*
  - ❖ ~ 0.75B monthly active users on Instagram in 2017.*
  - ❖ Modern machine learning depends on massive amounts of data.



Data collected by: M. Horowitz, F. Labonte, O. Shacham, K. Olokutun, C. Batten; extrapolations by C. Moore

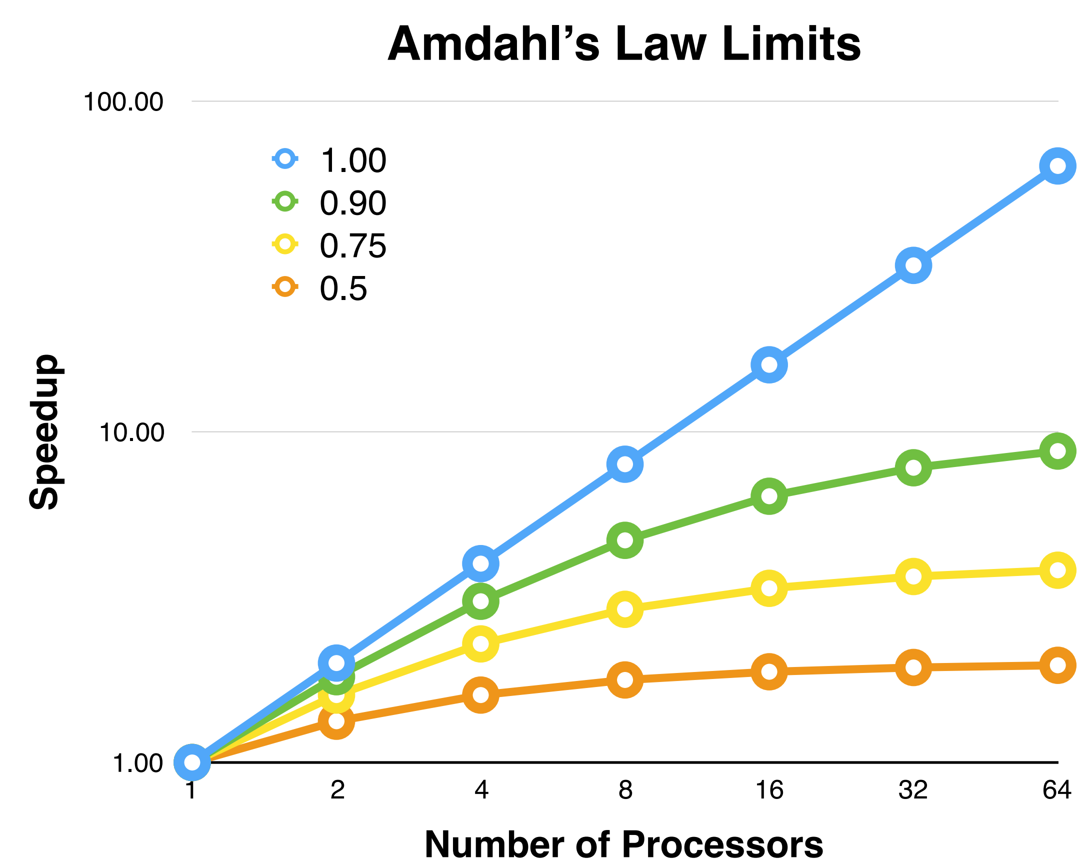
Yale

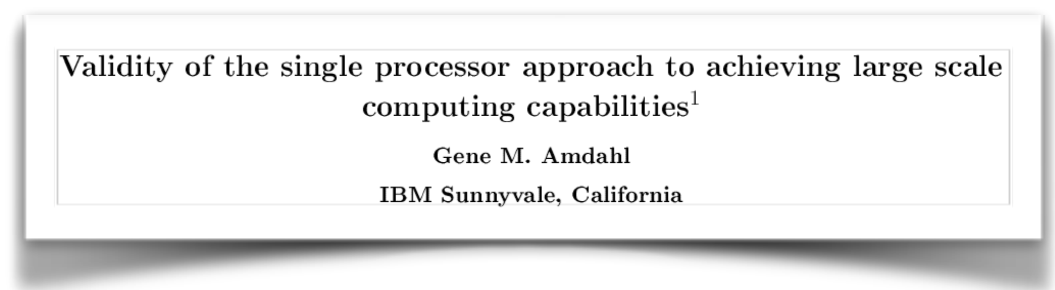
**Kleiner-Perkins 2018 Internet Trends*

AVLSI

Parallelism to the rescue?

- Some **algorithms** just aren't parallel
  - ❖ "Unfortunately, for most interesting algorithms, […] no architecture is scalable […]" -- Agarwal et al. (CACM 1991)
- But maybe we're going about this the wrong way…
- Physical systems, by their very nature, are massively parallel.
- ***Can we build computing systems inspired by physical ones?***




Validity of the single processor approach to achieving large scale computing capabilities[1]

Gene M. Amdahl

IBM Sunnyvale, California

1967

Neuromorphic computing*

- Philosophical motivation
  - ❖ Understand thought, consciousness
- Biological motivation
  - ❖ Understand the brain through engineering
- Computational motivation
  - ❖ Real-time vision, speech, pattern recognition, …

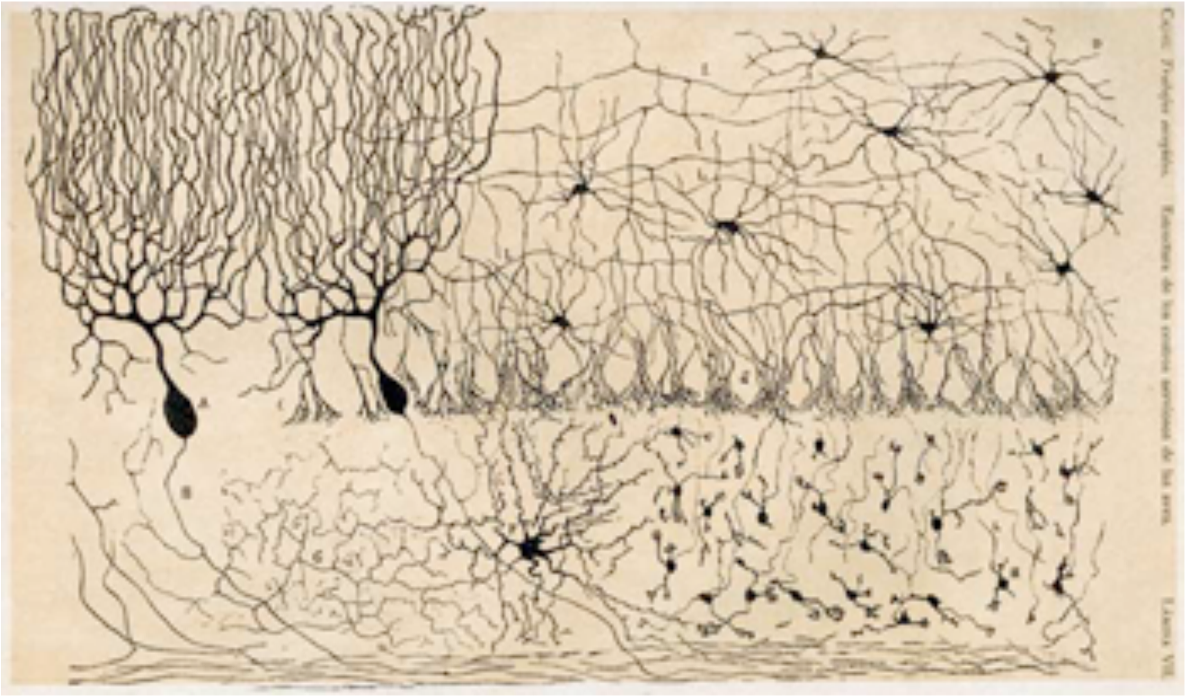
“Neuro” = neural

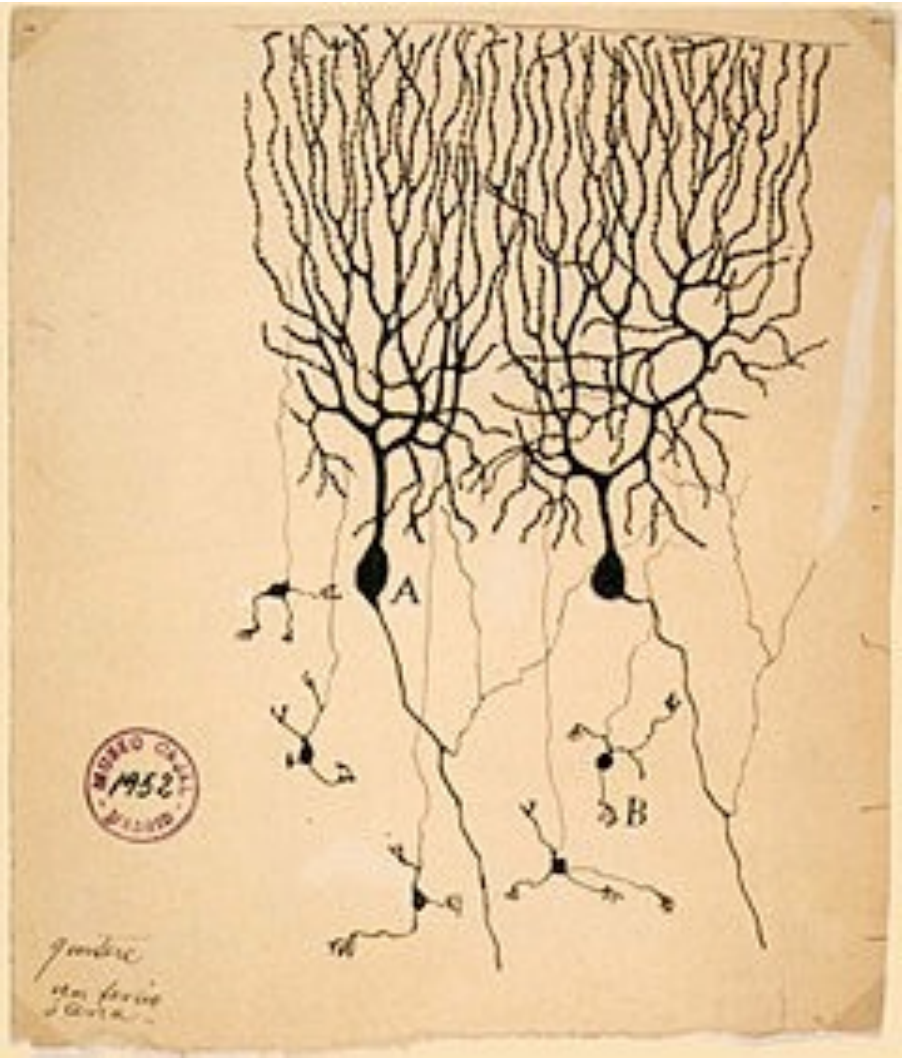
“-morphic” = “having the shape, form, or structure”

**term coined by Carver Mead*

Neuromorphic systems

- Neurons: nodes in the network
- Axons: out-going links
- Dendrites: in-coming links
- Axons connect to dendrites at synapses





Ramón y Cajal, (1852-1934)

❖ **Massively parallel, asynchronous computation**

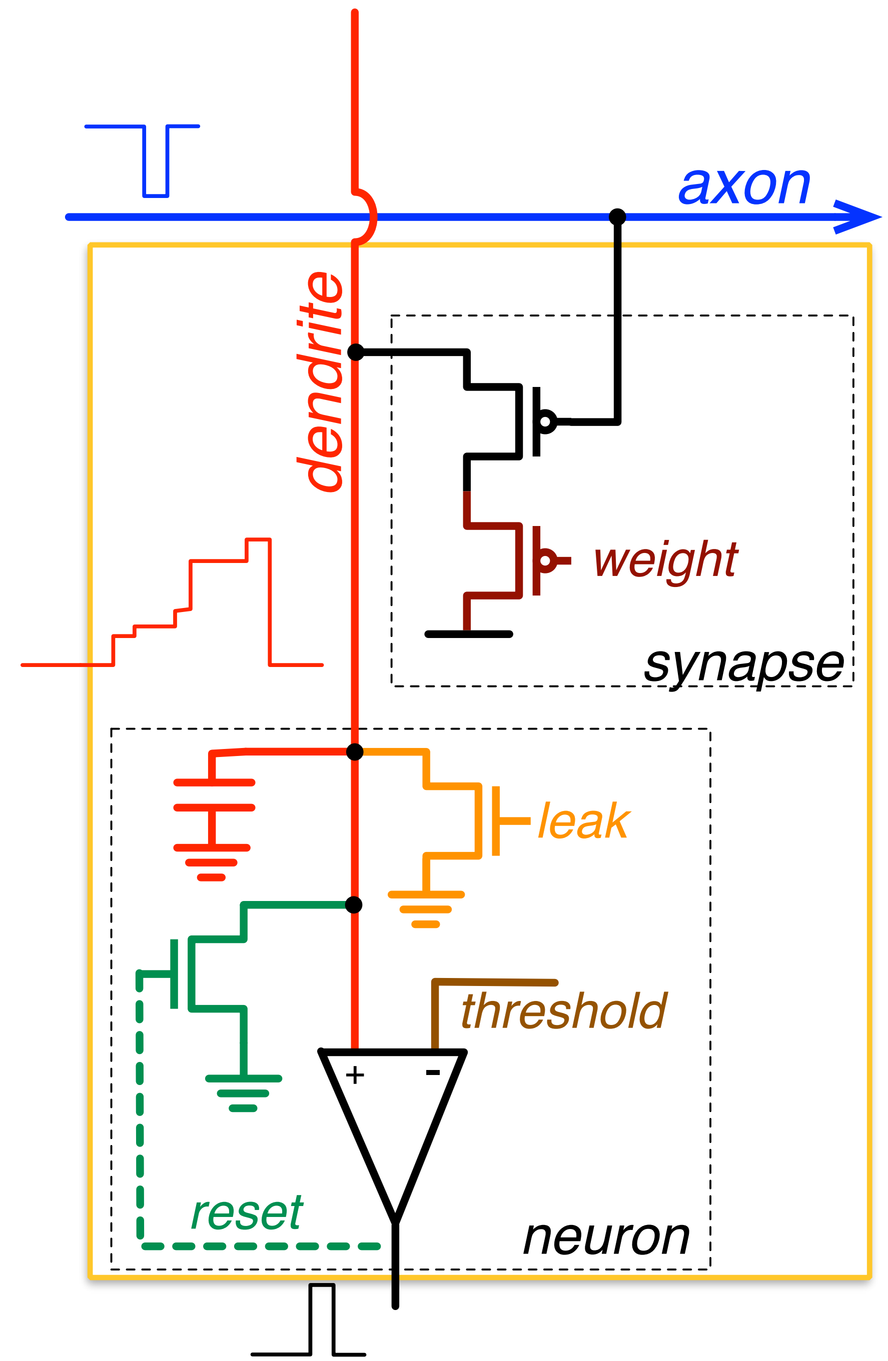
❖ **Many modern success stories (e.g. "deep networks")**

Yale

AVLSI

Neuromorphics 101

- Basic computation
  - ❖ Weighted input spikes are accumulated on a capacitor
  - ❖ The neuron is implemented as a "threshold detector"
  - ❖ On an output spike, the state of the neuron is reset (with a refractory period)
- ~1,000 to 10,000 synapses per neuron
- **Classical approach**
  - ❖ Mixed-signal design: analog neurons and synapse circuits, digital asynchronous communication

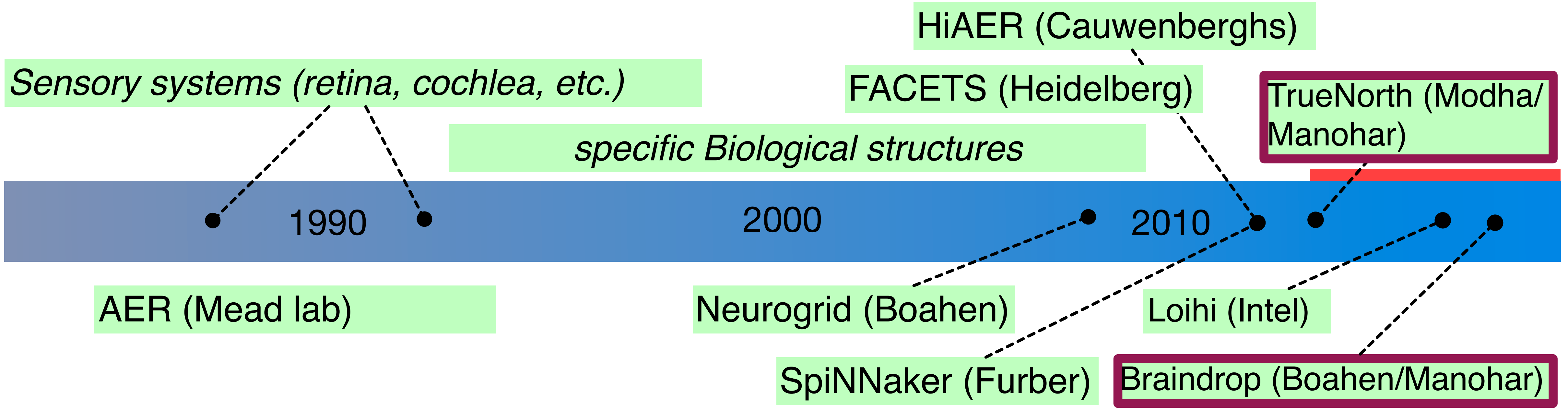


Yale

AVLSI

A bit of recent history…

- Since the mid 1980
  - ❖ specialized sensory systems
  - ❖ specialized neural circuits
- Today: “general-purpose” architectures

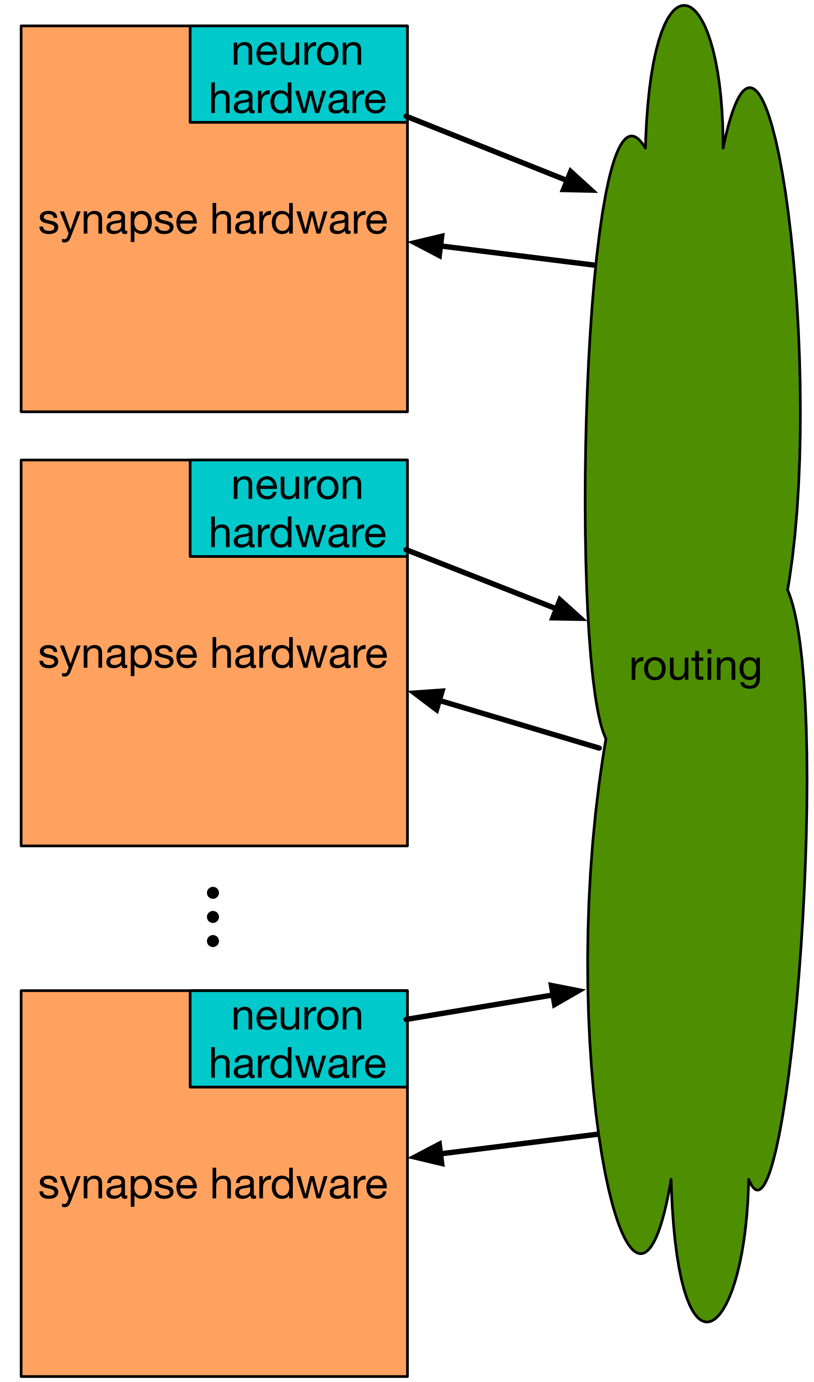


Yale

AVLSI

General purpose neuromorphic systems

- Core components
  - ❖ Set of neurons + synapses from the network being modeled mapped to hardware
  - ❖ Synapses can be made “superposable”
  - ❖ Routing network handles spike communication between hardware elements
- Time-multiplexing
  - ❖ Common hardware for computation
  - ❖ Per-neuron/per-synapse *state*

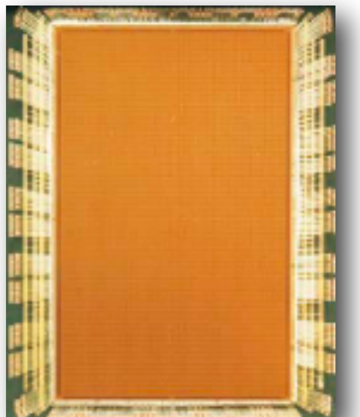


Yale

AVLSI

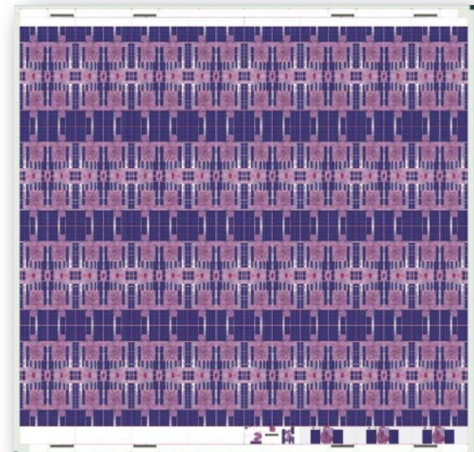
Current state-of-the-art

2014



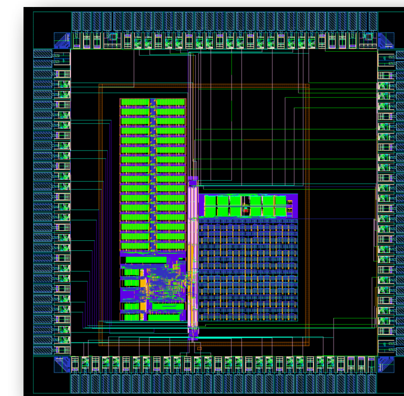
- IBM/Cornell "TrueNorth" chip
  - ❖ ~25 pJ/synaptic operation
  - ❖ 65mW for 1M neurons, 256M synapses
- 28nm technology
- QDI + bundled data asynchronous digital logic

2018



- Intel "Loihi" chip
  - ❖ ~24 pJ/synaptic operation
  - ❖ Integrated on-chip learning support
  - ❖ Microprocessors for management
- 14nm technology
- QDI + bundled data asynchronous digital logic

2019



- Stanford/Yale "Braindrop"
  - ❖ ~0.4 pJ/effective synaptic operation
  - ❖ Support for "NEF" programming model
- 28nm FDSOI
- QDI digital logic, synchronous I/O, and analog circuits for neurons and synapses

Yale

AVLSI

Sampling of applications

- TrueNorth: image recognition
  - ❖ CIFAR-100 dataset
    - near state-of-the-art accuracy*, >1,500 frames/s, 200mW
  - ❖ “Assembly language”: networks of neurons and interconnections
- Loihi: lasso optimization
  - ❖ ~50x lower energy and ~100x lower delay compared to low-power CPU
  - ❖ “Assembly language”: networks of neurons and interconnections
- Braindrop: does not use hand-crafted networks
  - ❖ Assembly language: “neural engineering framework”
  - ❖ Program analog circuits at a higher level of abstraction
  - ❖ Most efficient platform for neural engineering framework

Challenges: design and energy-efficiency

- Biological neural systems
  - ~ 20 **fJ**/synaptic operation
- TrueNorth/Loihi
  - ~ 20 **pJ**/synaptic operation
- How do we close the gap?
  - Many, many proposals (new devices, materials, etc…) for better synapses and neurons
  - Reality
    - ~30-50% power is in spike communication/storage— Amdahl strikes again!
      - Best case: reduce to 7-10 pJ, even after overcoming all the technical obstacles!
  - Many proposals with significantly lower energy reported
    - *… but not for a system, just for small devices/components*

Yale

AVLSI

Challenges: design and energy-efficiency

- All the state-of-the-art solutions include
  - … asynchronous digital communication
  - … and plenty of asynchronous digital computation as well
  - Unsupported by commercial tools!
- Spike communication network
  - Low latency needed, but low bandwidth
  - Asynchronous design makes this easy to support
- We are developing a new open-source flow for asynchronous design
  - DARPA's **E**lectronics **R**esurgence **I**nitiative
  - Goal: to make asynchronous design accessible

Challenges: programmability and algorithms

- How do we best utilize this computation model?
  - ❖ … in a general-purpose framework?
- What's the right "programming language"?
- Current solutions
  - ❖ Use learning/training and artificial neural networks
  - ❖ Use hand-crafted solutions
  - ❖ Time-averaged spike rate is used to represent a value

$$|v - \hat{v}| \leq \epsilon$$

sender receiver

$$\max_{v \in [0,1]} \{\Pr_{\hat{v}}[|v - \hat{v}| > \epsilon]\} \leq \delta$$

ϵ (bits)	Number of "spike slots" needed		
	δ=0.05	δ=0.10	δ=0.25
1	28	20	8
2	176	126	56
3	848	592	288
4	3670	2582	1248
5	15211	10731	5227

Summary

- Neuromorphic systems
  - ❖ Biologically inspired, naturally parallel approach
  - ❖ Various attempts to create programmable platforms

- Biological systems are an existence proof
  - ❖ … we need to better understand *how* they compute

- Challenges
  - ❖ What are efficient ways to compute in this framework?
  - ❖ How do we reduce the cost of communication and storage?
  - ❖ Is there a *different abstraction,* beyond simply emulating Biology?

Yale

AVLSI

Acknowledgments

- Many, many members of the neuromorphic community
  - ❖ Andreas Andreou, Gert Cauwenberghs, Tobi Delbruck, Shih-Chi Liu, …
- Major project collaborators
  - ❖ TrueNorth: Dharmendra Modha, John Arthur, Paul Merolla, …
  - ❖ Braindrop: Kwabena Boahen, Alex Neckar, Sam Folk, Ben Benjamin, …
- Group members
  - ❖ Filipp Akopyan, Nabil Imam, Saber Moradi
- Sponsors
  - ❖ DARPA, ONR, AFRL

Yale

AVLSI