Neuromorphic Computing

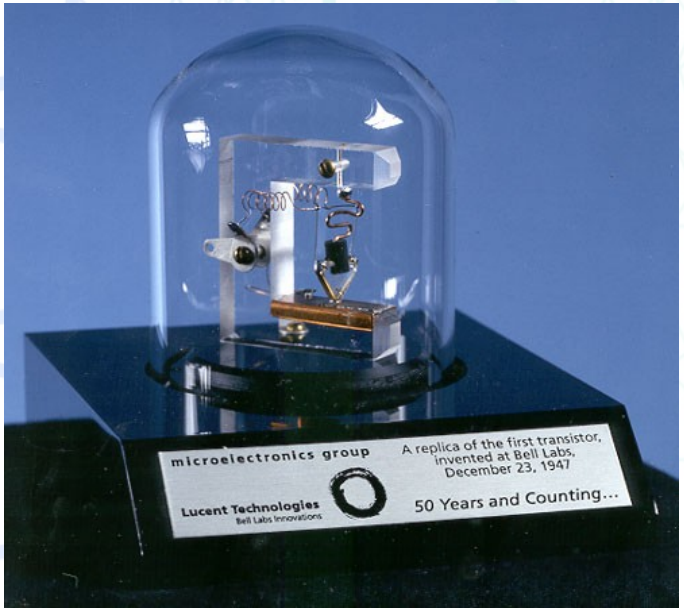
Dr. Charles Kind

Computational Neuroscience Group, SCEEMS, Bristol University

October 2022

Integrated circuits

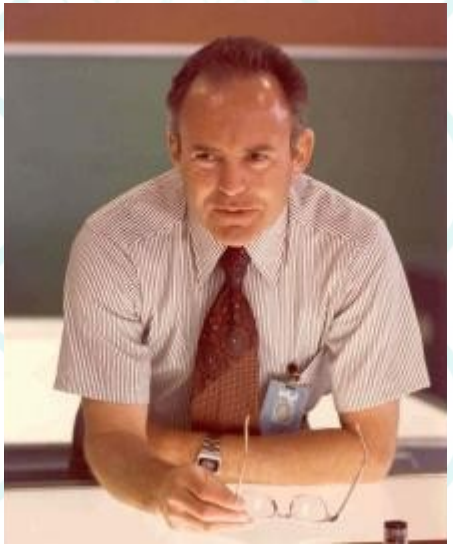
- Transistors
  - an amplifier or a two state switch
  - (CMOS) complementary metal oxide semiconductor
  - As of 2011 > 99% of all integrated circuit (IC) chips made this way
  - Core i7-8700K has ~ 3 billion



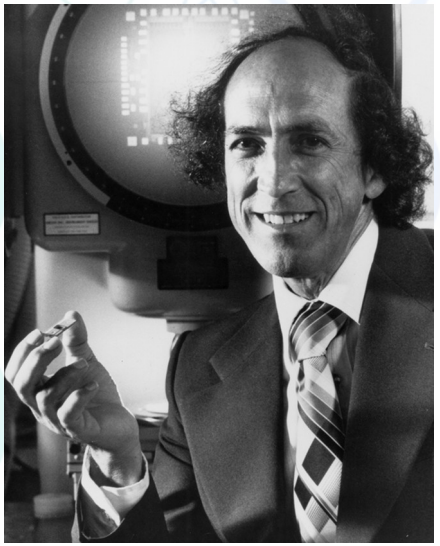
A replica of the first working transistor,
Lucent Technologies 1997

Integrated circuits, it’s the LAWS!

- Moore’s ‘law’: The number of transistors in an integrated circuit doubles every two years
  - Gordon Moore co-founder of Intel, Electronics magazine, 1965.
- Dennard’s scaling law: As transistor density doubles, chip performance improves while its power density remains the same.
  - Robert H. Dennard et al., IEEE Journal of Solid-State Circuits, 1974.
- Together they imply performance per watt doubles approximately every 18 months (Koomey's law)
- Options to increase performance
  - More transistors
  - Faster speeds (clock rate)
  - Multiple cores
  - Something else ???



Moore



Dennard

Integrated circuits, what really happened.

- Comparing computing power with the VAX 11/780 (1970)
- From 1986-2006 growth doubled
- Intel Xeon 3.2 GHz (2003) 6000 times more powerful than the VAX 11/780
- Hooray … but wait …
- Since 2015, growth in computing power is estimated at an annual rate of 3.5%
- But whhhyyyyyyy!!!

VAX 11/780

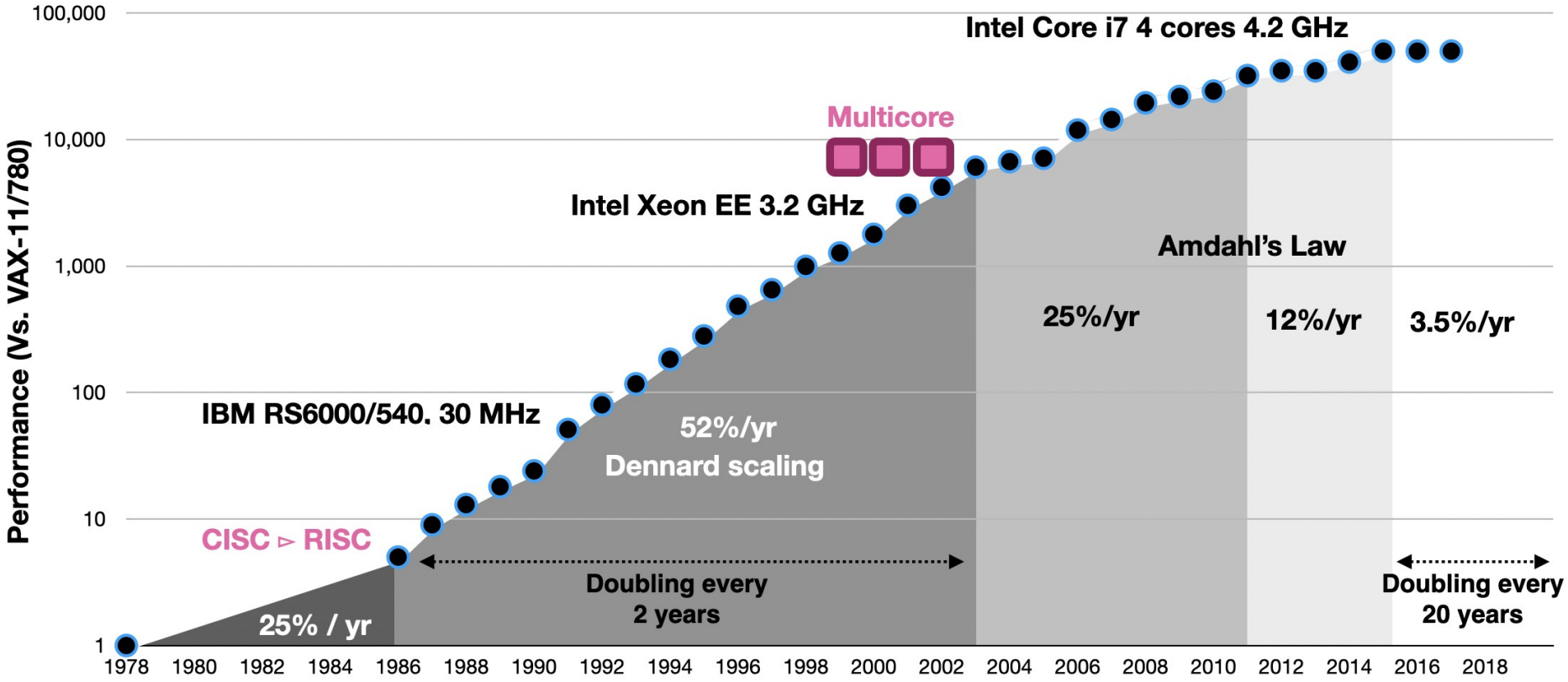
Dell Poweredge Server from 2003

Data from John L Hennessy and David A Patterson. Computer Architecture:

A Quantitative Approach. Elsevier, 2017

Integrated circuits, what really happened.

Performance, based on SPEC benchmarks and scaled to newer SPEC versions

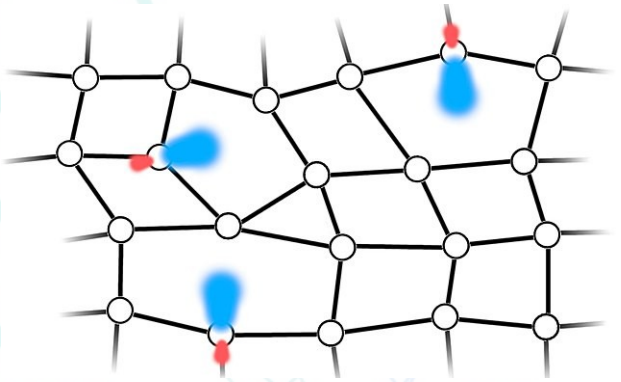


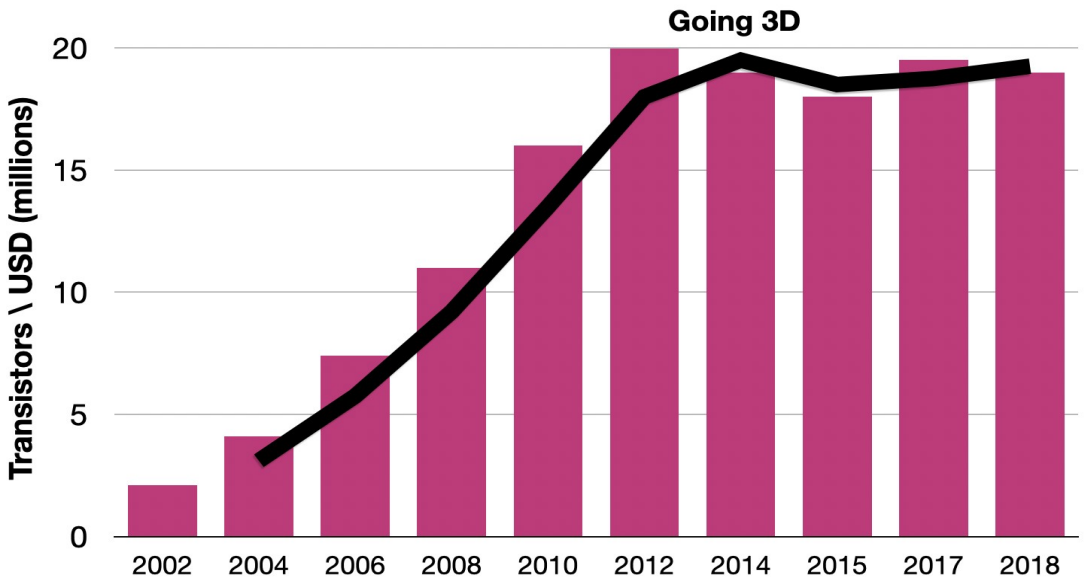
Data from John L Hennessy and David A Patterson. Computer Architecture: A Quantitative Approach. Elsevier, 2017

Integrated circuits, what really happened.

More transistors please

- Dangling bonds (an unsatisfied valence on an immobilized atom) trap electrons, limits transistor size reductions (around 5 nm)
- A partial solution is to move to 3D transistors
  - Increased complexity and cost
- Quantum effects: new and expensive annoyances?



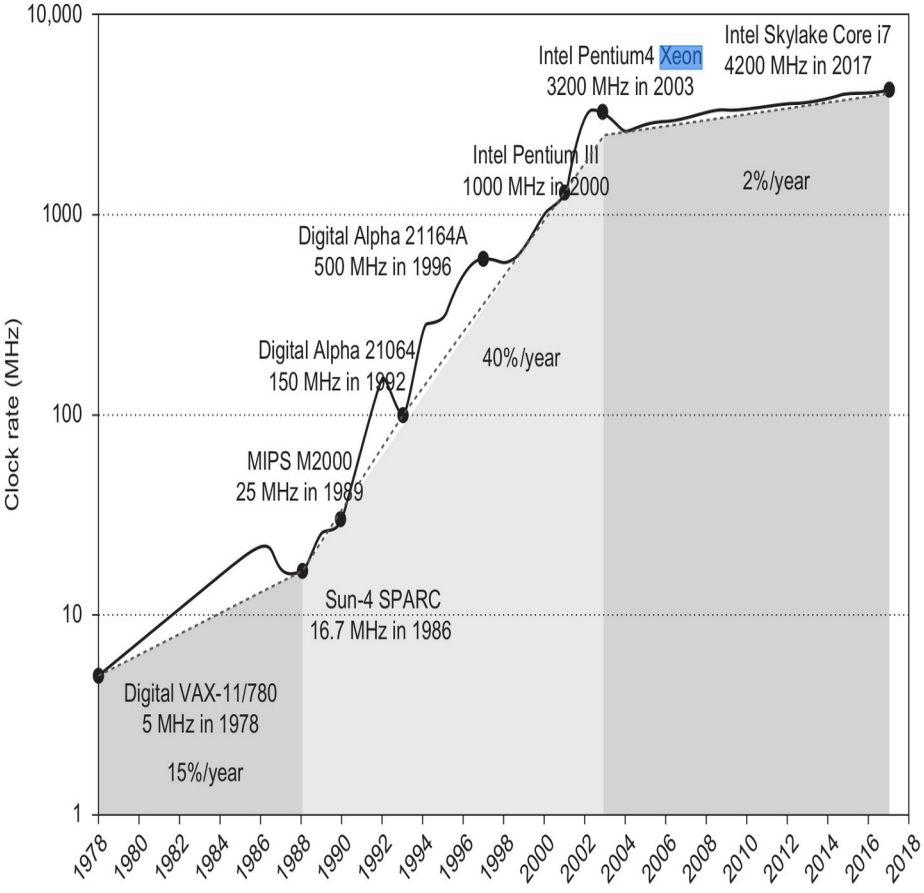


Data from Kwabena Boahen. A Neuromorph's Prospectus. IEEE, 2017

Integrated circuits, what really happened.

Higher clock speeds please

- For CMOS chips, the primary energy consumption is the switching of transistors, also called dynamic energy.
- The first microprocessors (such as the Intel 4004) consumed less than a watt at around 100 Khz
- 32-bit microprocessors (such as the Intel 80386) used about 2 W at 12-40 Mhz
- Core i7-8700K uses around 100 W (3.7 Ghz) and can peak at 150 W (4.7 Ghz)
- We are at the limits of what can be air cooled and have been for around 15 years



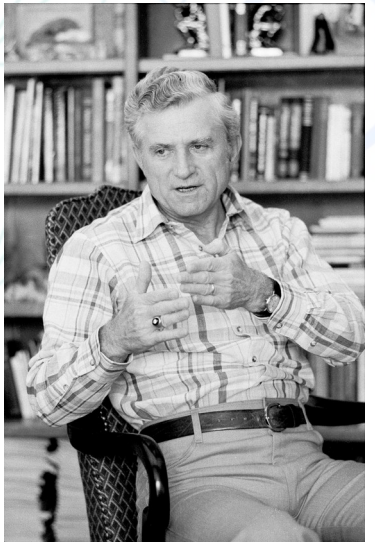
Data from John L Hennessy and David A Patterson. Computer Architecture: A Quantitative Approach. Elsevier, 2017

Integrated circuits, what really happened.

Distributed computing or more cores please

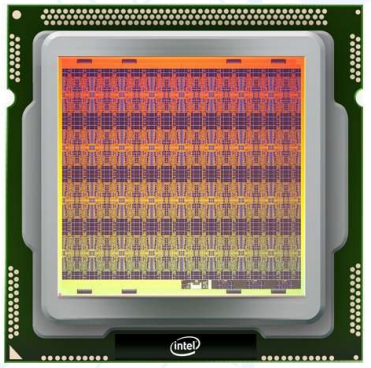
- This has been the predominant driver of increased performance since the failure of the LAWS
- Amdahl's law: the performance improvement to be gained from using some faster mode of execution is limited by the fraction of the time the faster mode can be used
  - Gene Amdahl, 1967
- Suppose 20% of the algorithm cannot be distributed for multicore processing. In that case, the maximum gain in performance is limited to 5x, no matter how many cores are available
- Power consumption and the rise of the GPU
- Von-Neumann bottleneck and the memory wall

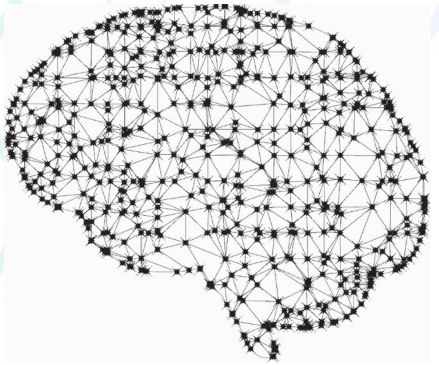
Amdahl



Integrated circuits: Summary

- More transistors
  - Energy and miniaturisation constraints
- Faster processors
  - Energy and heat dissipation constraints
- Distributed computing
  - Can be restricted by applications
  - See above
- General issues
  - Data transfer constraints
  - Power





Computers vs brains

The human brain

- Energy consumed
  - Kwabena Boahen (2017), current through ion channels ~ 20 W
  - Ferris Jabr (2012), using metabolic rates ~ 12 W

- Processing power and memory
  - Eugene Izhikevich (2004), ~ 10^11 neurons ~ 1e18 FLOPS
  - Sandberg and Bostrom (2008) ~ 1e18 bits (~1e8 GB)

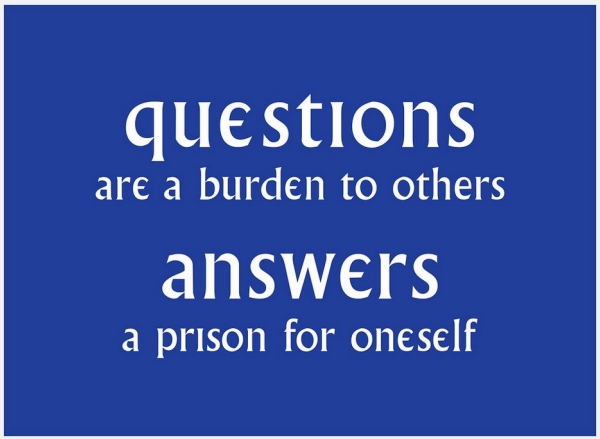
The computer

- Nvidia 3080, 30 TFLOPS (1e12) @ 320 W
- Fugaku, 1e18 FLOPS, 4e16 bits @ 30 MW and $1 Billion

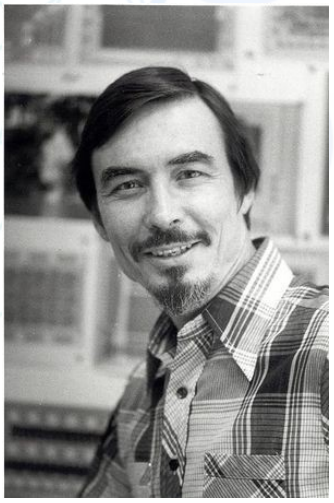
Neuromorphic Computing

The historical inspiration

- What to do about the obvious limits to semiconductor technology?
- Faster to scale old designs to smaller feature sizes than to innovate at the architecture level.
- 1981: Carver Mead, Richard Feynman and John Hopfield join together to teach: “Physics of Computation” at the CalTech
- “the brain is a factor of 1 billion more effective than our present digital technology and a factor of 10 million more efficient than the best digital technology that we can imagine”, Mead, 1990
- Lets make computers like human brains!



Patrick McGoohan, 1967



Carver Mead

Neuromorphic Computing

Goals

- Scalable architecture designed to run brain like computations
- Replace virtual neurons/synapses with physical, analog devices
- ‘In memory’ devices
- Reduce power consumption
- Enable ‘edge computing’
- Sit in the gap between high accuracy, high energy classical computing and low accuracy, low energy neural computation
- Become standard CPU/hardware extension like SIMD SSE
- Learn more about how the human brain works

Neuromorphic Computing

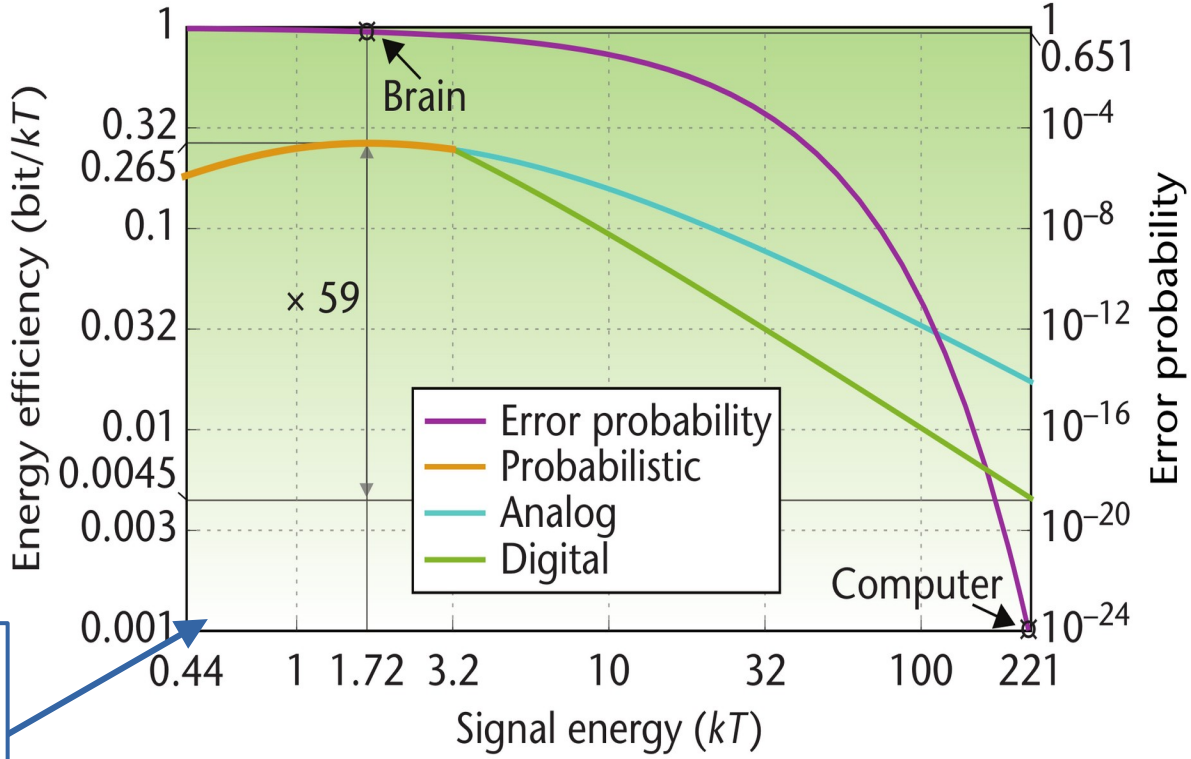
Power vs accuracy

- Computer's digital computations cannot tolerate errors and pay a high energy cost to minimise it.
- By using analog computation, which degrades gracefully, brains tolerate errors

Claude Shannon (1948), provides us with a method to analyse communication energy efficiency with respect to errors.

Brain: Analog signals, high energy efficiency at low precision

Computer: Digital signals, high energy efficiency at high precision



Kwabena Boahen (2017)

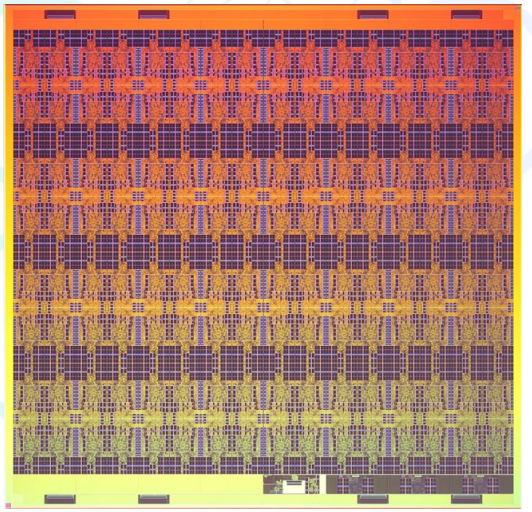
A signal with energy E conveys $b = \frac{1}{2}\log_2(1 + E/kT)$ bits of information with an energy efficiency of $b/(E + kT)$ bits per joule.

The Neuromorphic Computing Roadmap

The main academic approach

1) Implement computation with analog circuits (elements) to consume close to the theoretical minimum energy.
2) Implement communication with asynchronous digital circuits to be robust to transistors that shut off intermittently.
3) Distribute a computation across a pool of (silicon) neurons to be robust to transistors that shut off intermittently or permanently.
4) Communicate spikes from pool to pool at a rate that scales linearly with the number of neurons per pool.
5) Encode continuous signals in these spike trains with precision that scales linearly with the number of neurons per pool.

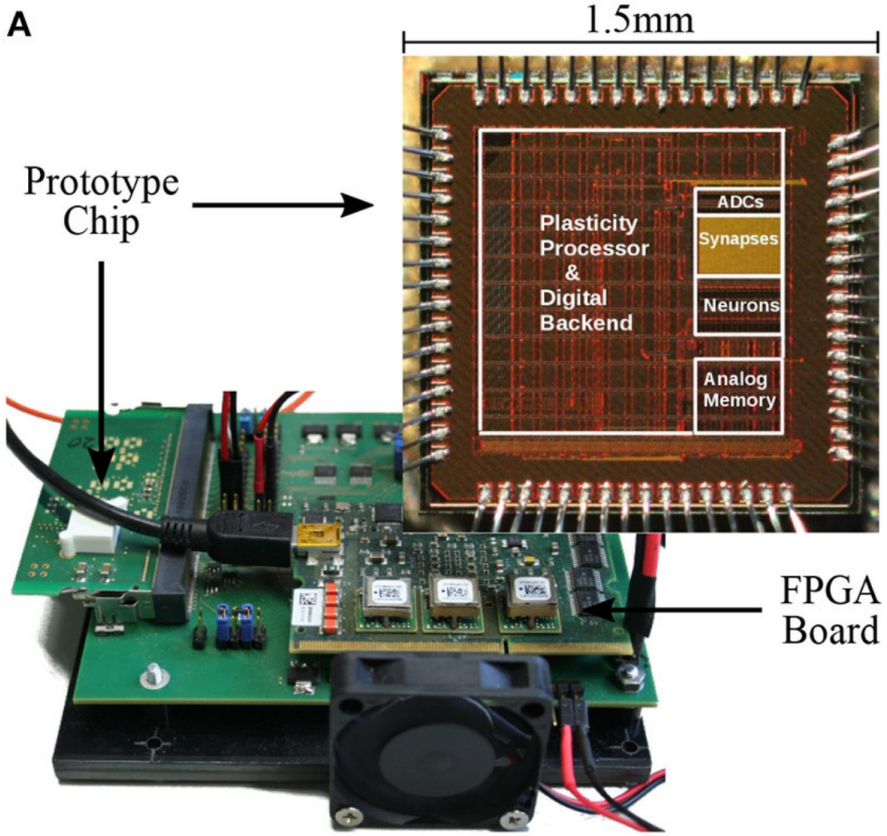
Intel Loihi chips on an Intel Nahuku board. 4194304 neurons and 4160000000 synapses

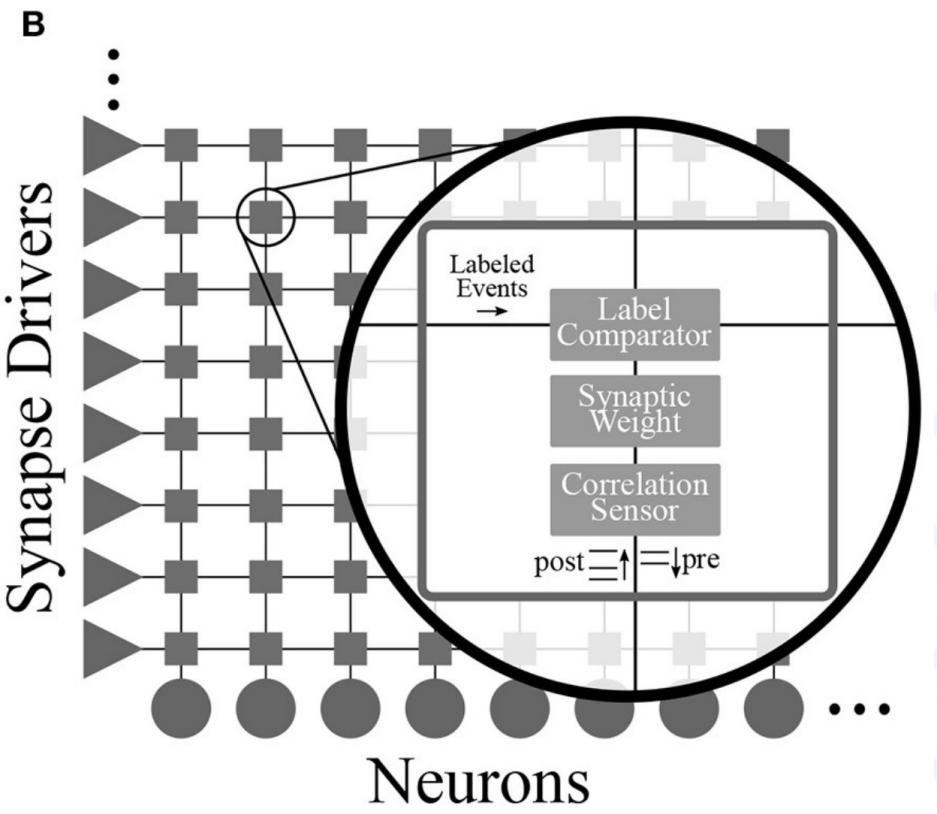


Loihi die, 2184 neurons per mm²

Kwabena Boahen. A Neuromorph's Prospectus. IEEE, 2017

Neuromorphic Computing, new directions





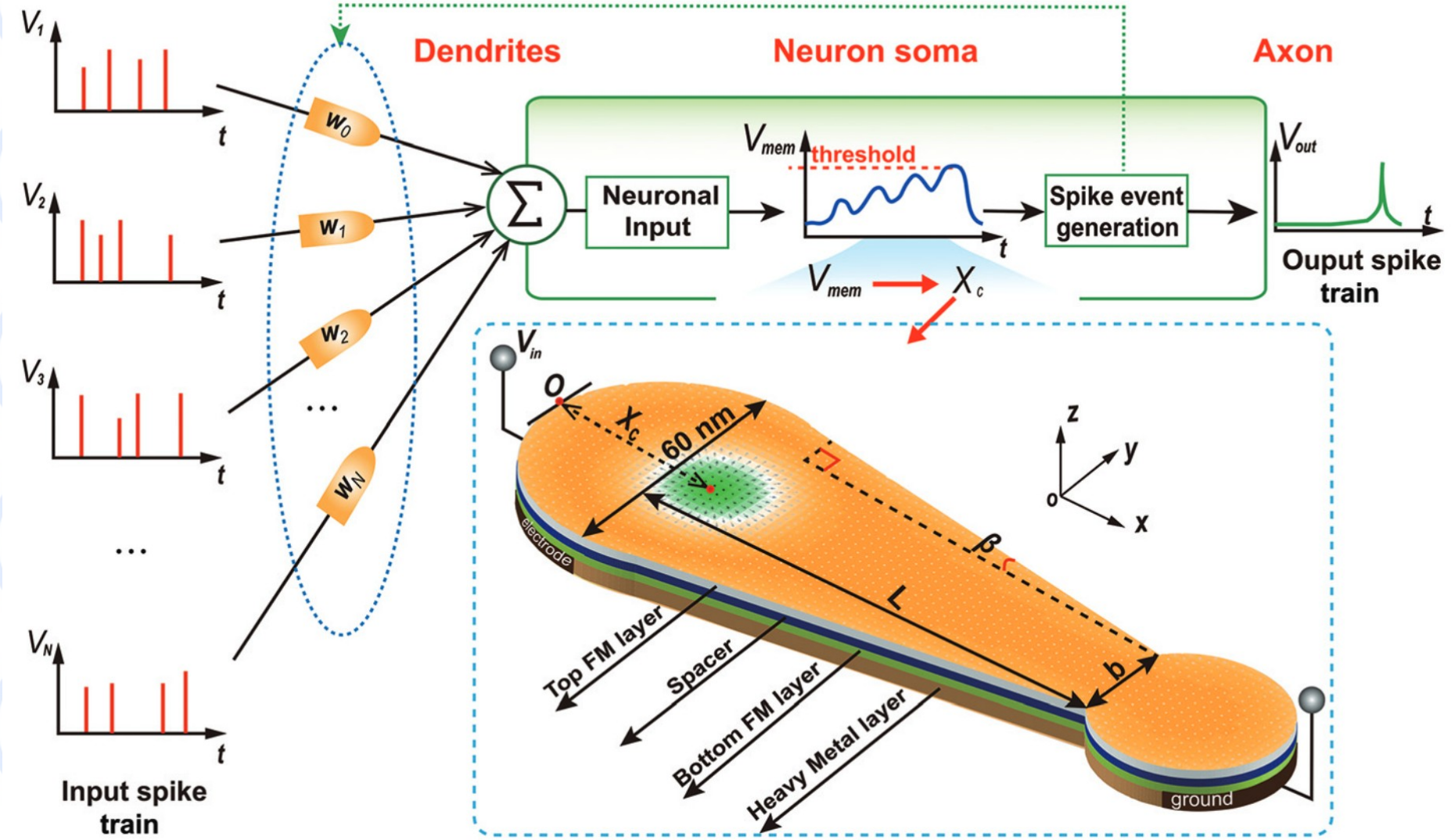
Brainscales-2 chip

- CMOS based analogue neurons, Leaky Integrate-and-Fire (LIF)
- Electronic bus for connections and instructions
- Weights stored in RAM

Wunderlich et al., Frontiers in Neuroscience, 2019

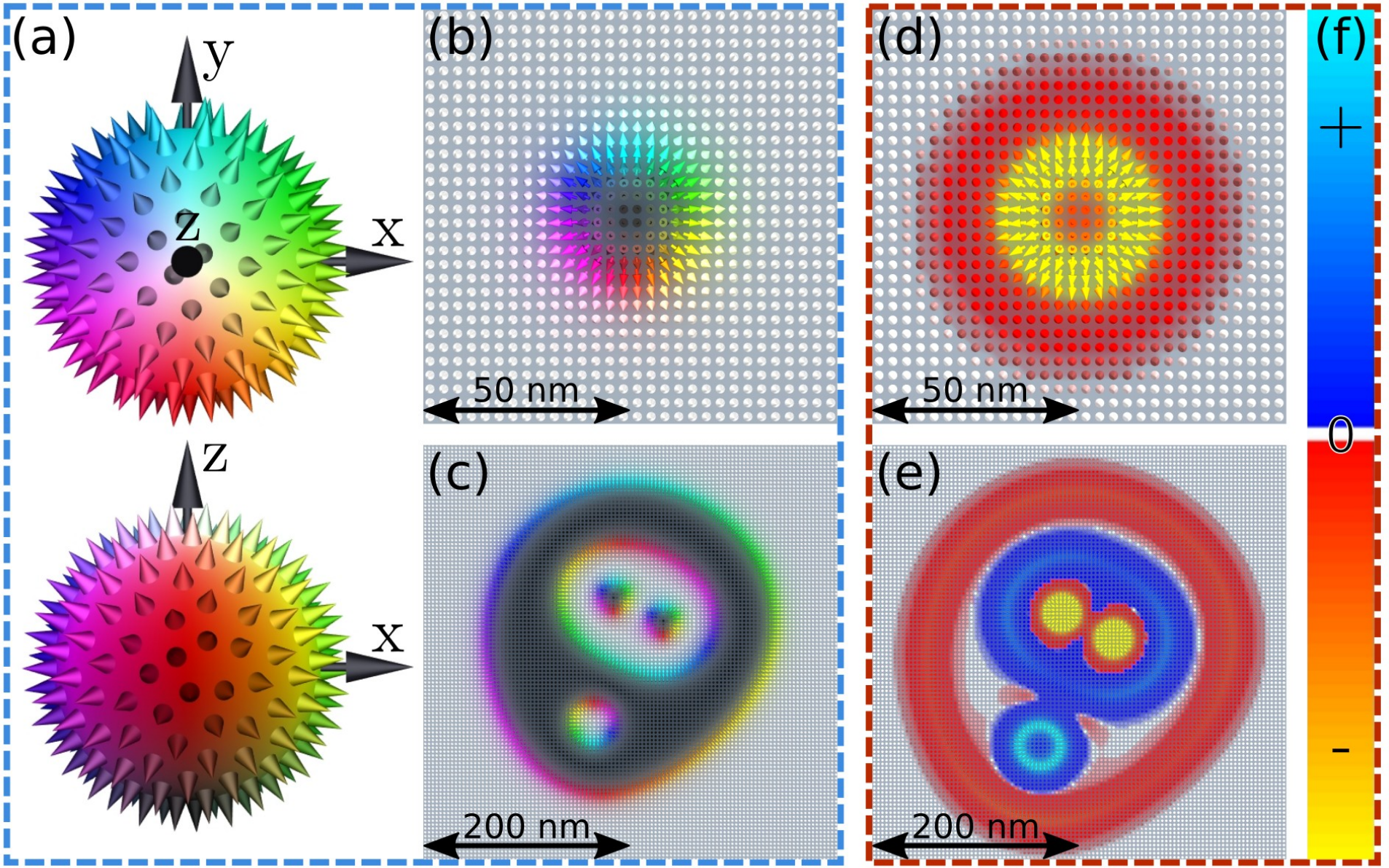
Neuromorphic Computing, new directions

A proposed skyrmion based neuron



Chen et al., Nanoscale, 2017

Neuromorphic Computing, new directions



Me, 2022

Thank you :)