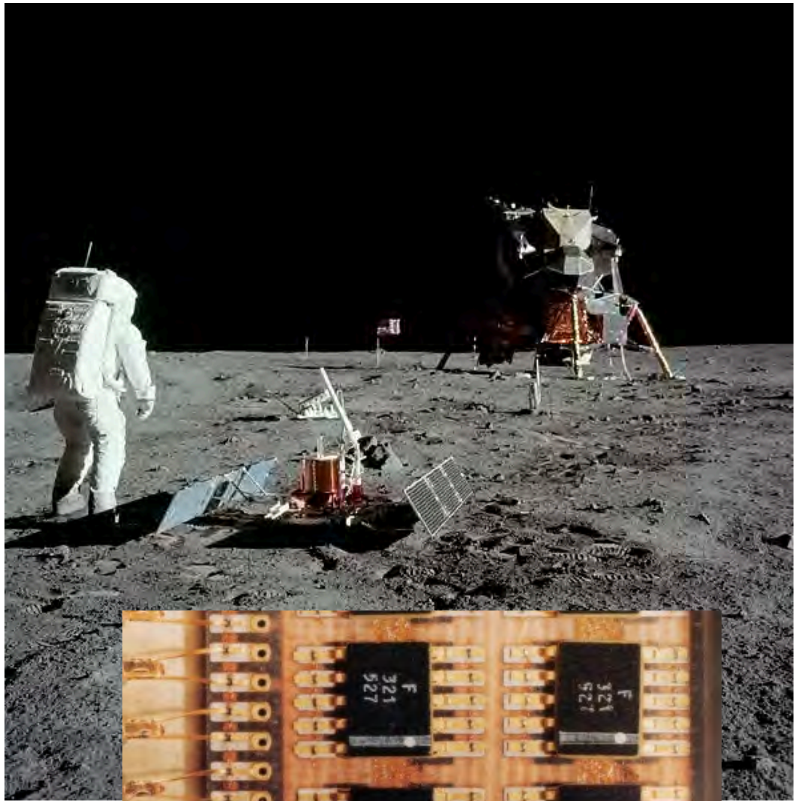
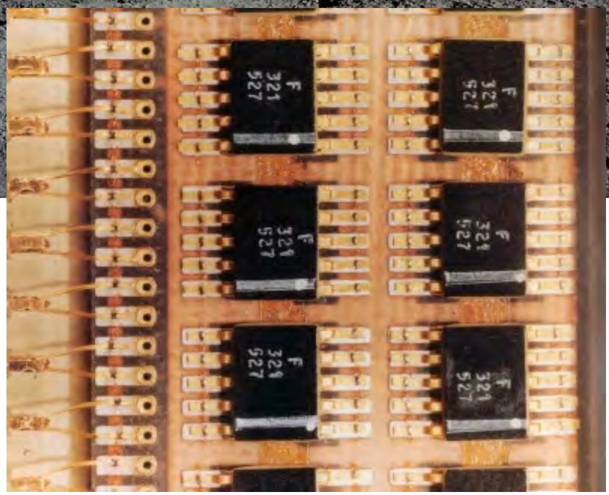
Fin de l'ère 'Logique binaire-architecture Von Neumann-CMOS': tapage ou véritable nouveau monde ?

Alain Cappy

IEMN-IRCICA

50 years ago…. (July 21, 1969)



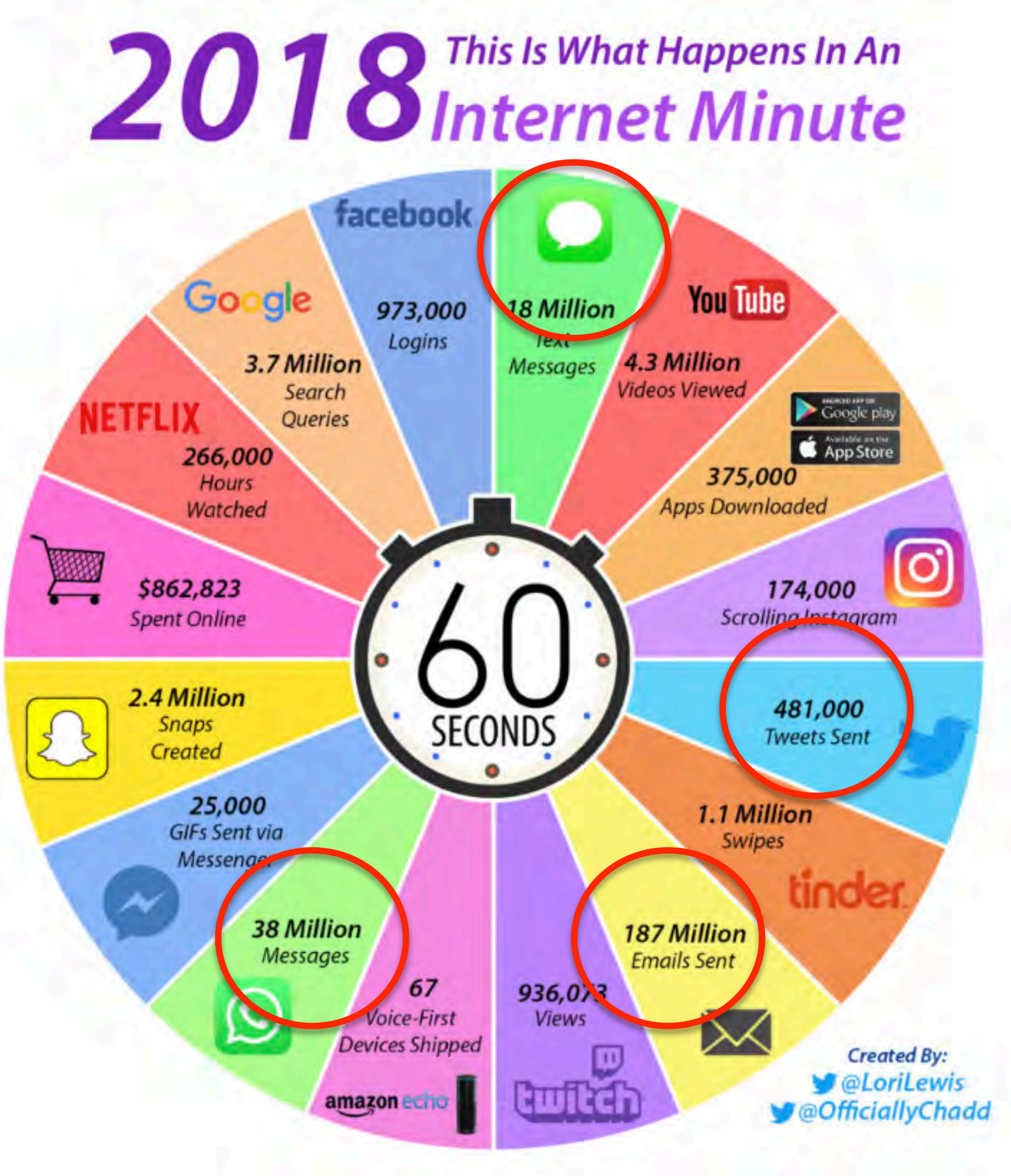


The Apollo flight computer

- ✓ 2800 ICs, each with dual three-input NOR gates
- ✓ ICs implemented using resistor–transistor logic (RTL)
- ✓ 2048 words of erasable magnetic-core memory and 36 kilowords of read-only memory.
- ✓ Clock frequency: 80 kHz

- ✓ 55W and 32kg

Nowadays : the Data Deluge !



Architecture for Information Processing

- The success of present processing machines is due to three factors:
  - Boole algebra and associate algorithms
  - Turing machine/ Von Neumann architecture
  - CMOS with scalability properties

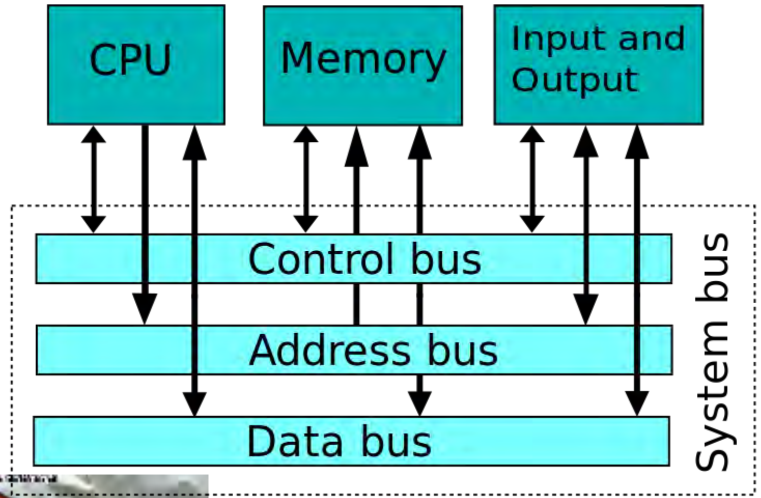
∧	0	1
0	0	0
1	0	1

∨	0	1
0	0	1
1	1	1

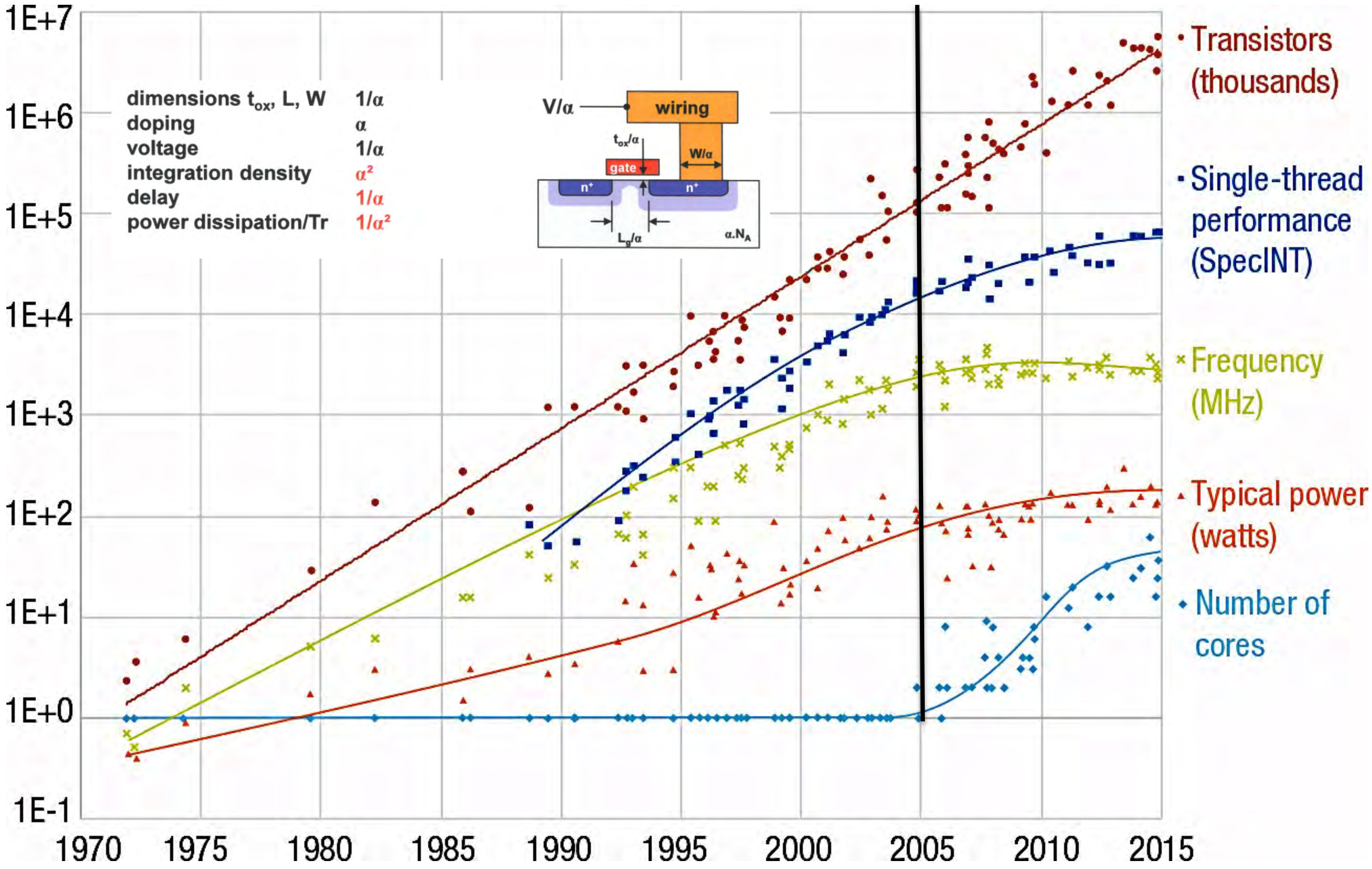
→	0	1
0	1	1
1	0	1

Figure 1. Truth tables

Figure 2. Logic gates

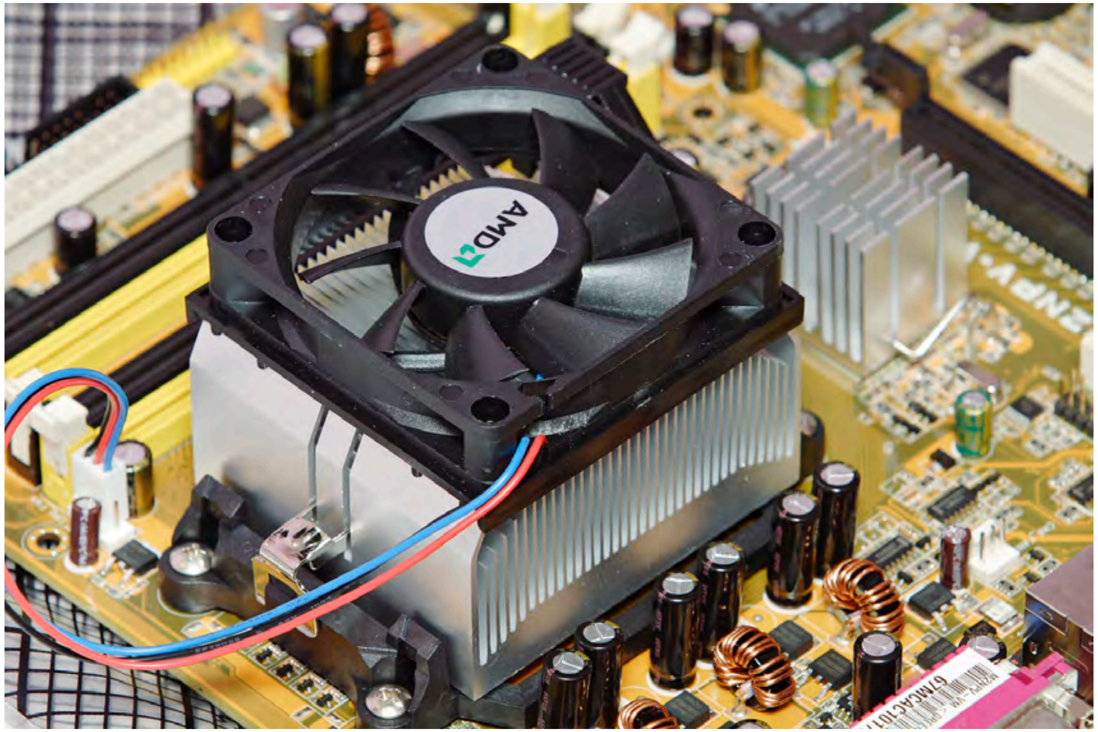


40 Years of Microprocessor trend data



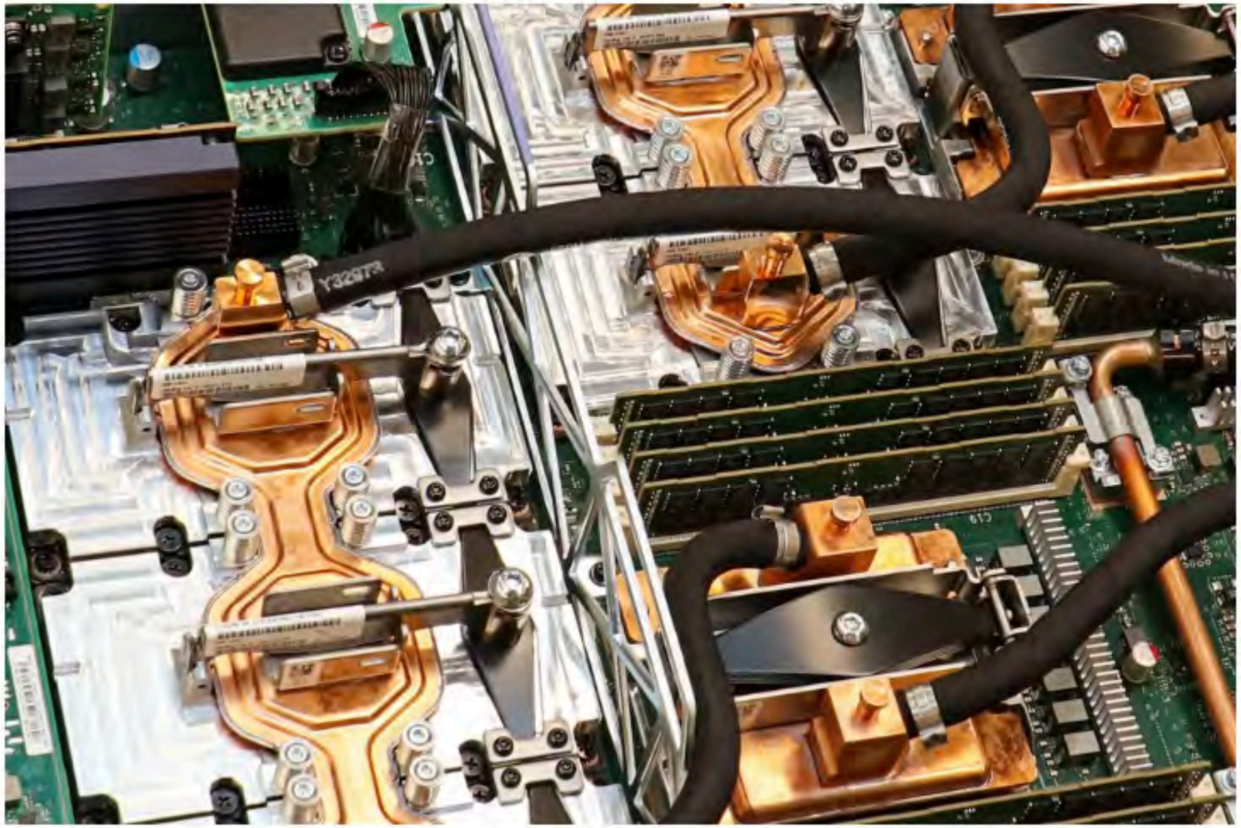
The Main Limitation: Air Fan Cooling

$P_{max} < 100\ W/cm^2$



$100 W/cm^2$ = 10 times a hotplate !

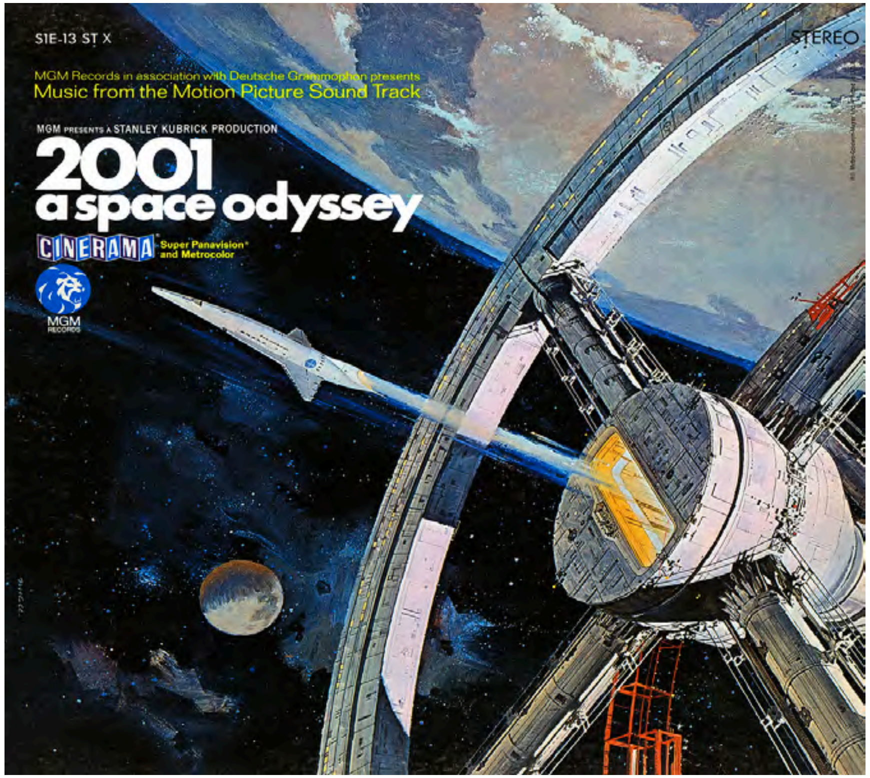
SUMMIT Supercomputer (IBM)



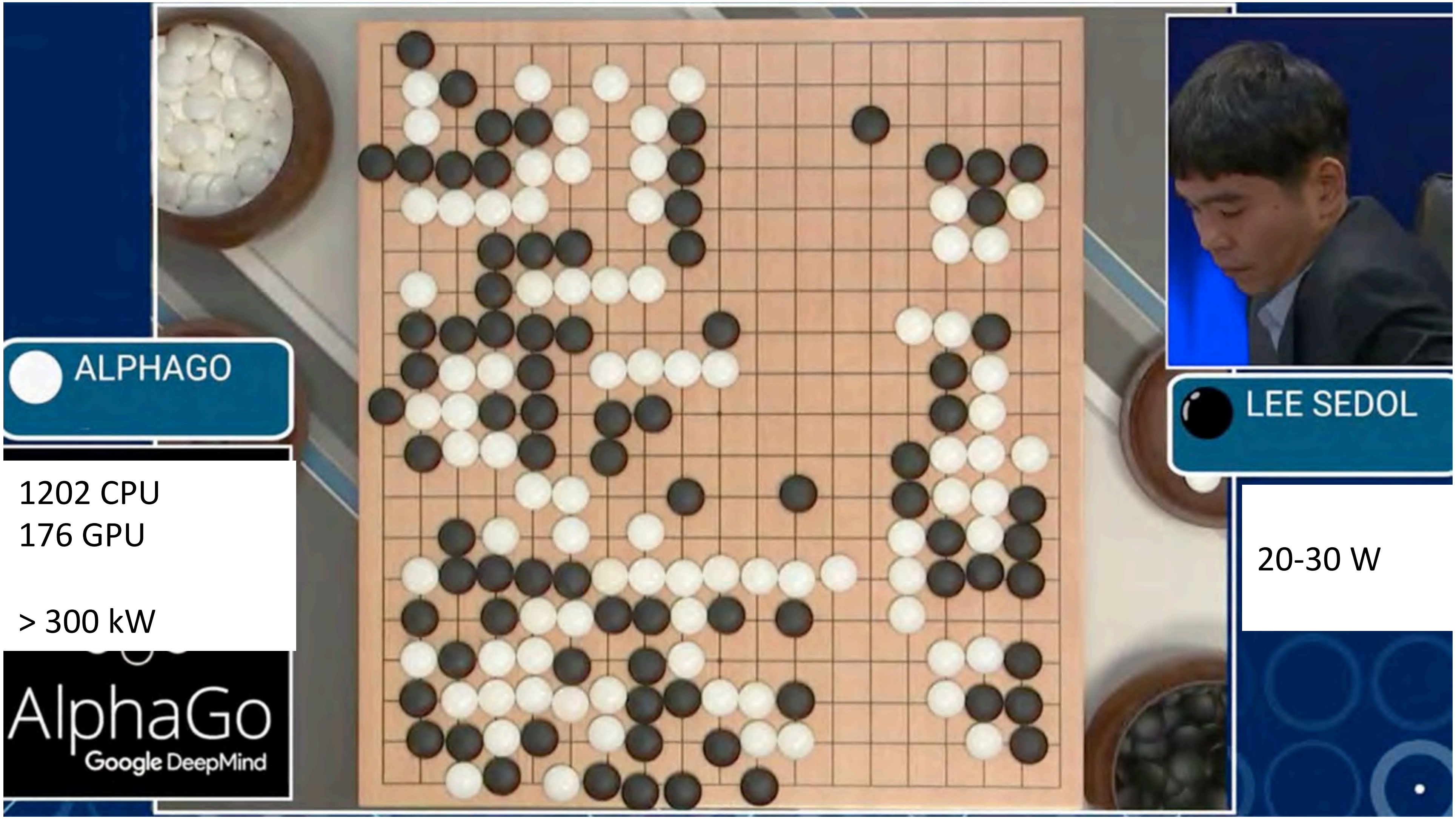
Tubes carry water to the Summit supercomputer's chips so they can run faster without overheating.

It takes up an eighth of an acre -- the size of two tennis courts. Its peak energy consumption is about 15 megawatts, enough to power more than 7,000 homes.

50 years ago.... (September 27, 1968)



Lee Sedol vs Alpha Go



1202 CPU
176 GPU

> 300 kW

20-30 W

The Autonomous Car

- On board processing power 50 to 100 processors ≈ 2-5 kW

+ Off board power (Vehicule to Vehicule, Vehicule to infra…) estimated at 5kWh per Gigabytes of data

An Old Fashion Driven Car

- 20-30 W of power only, needed for the driver…

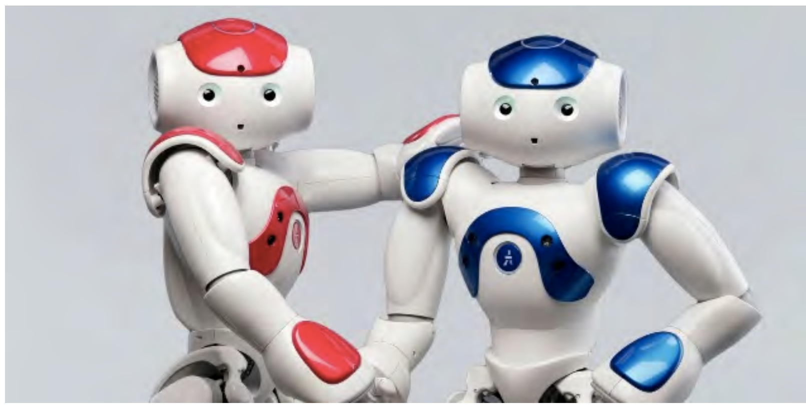
End of Moore's Law: What Next ?

What are the needs ?

- High computing power

- New fonctionnalities. More ‘intelligence’. Processing of natural data (image, sound…..)

- Ultra-low power dissipation
- High energy efficiency

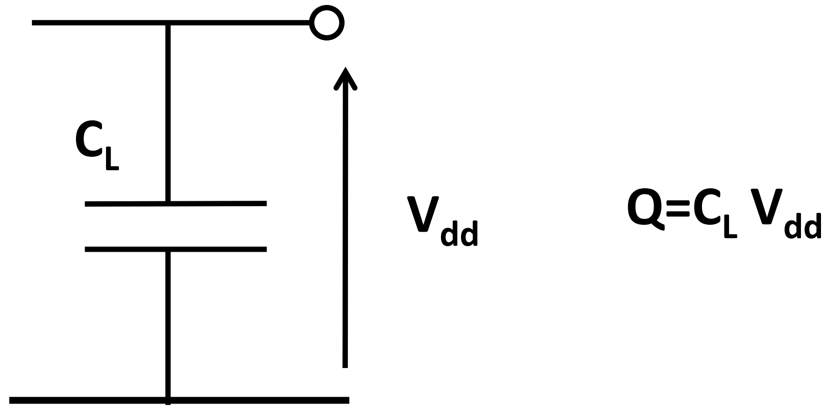


Questions

- Why such a large wasted energy-power?
  - CMOS ?
  - Von Neumann architecture ?
  - Binary logic ?

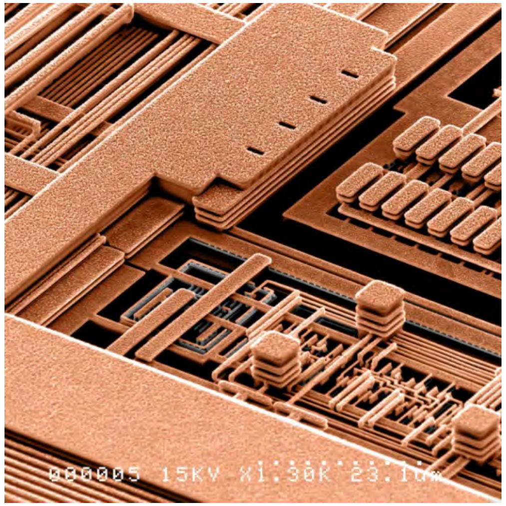
CMOS Power Dissipation

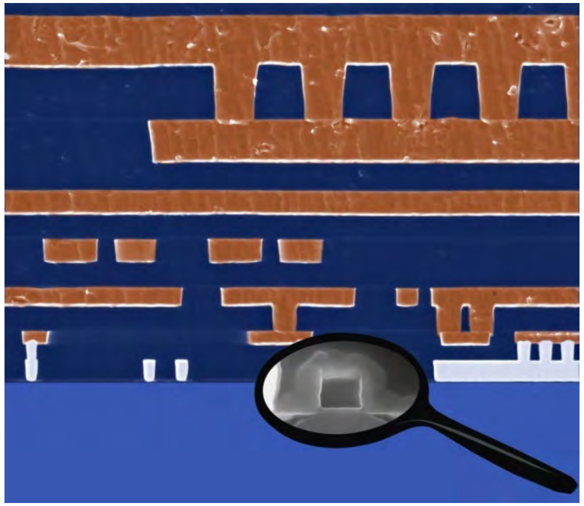
- Dynamics power : the state variable is a charge (or a voltage at a capacitance terminals). Information processing needs charge and discharge of capacitances

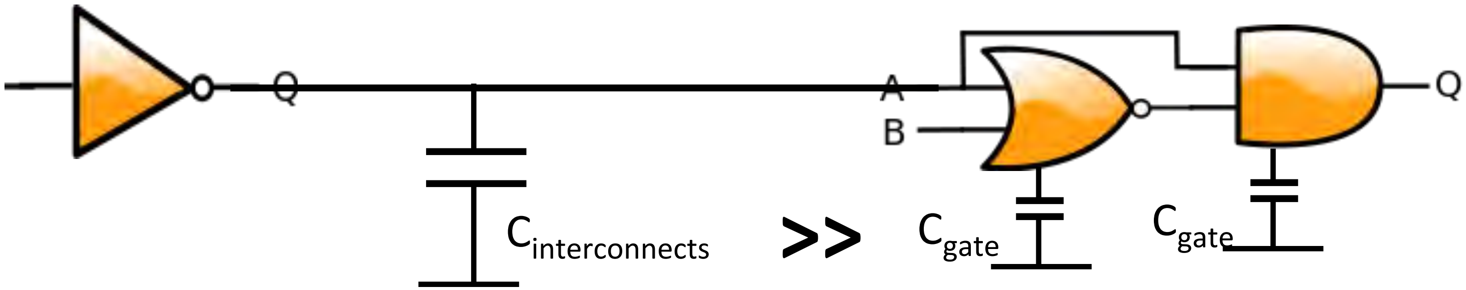


- ½ $C_L V_{dd}^2$ is lost at each transition (for constant V_{dd})
- **Two key parameters : C_L and V_{dd}**

Interconnects and gate capacitances







Fetching operands costs more than computing on them !

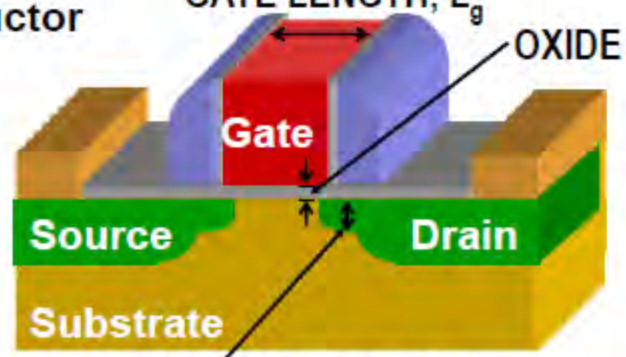
iRCica

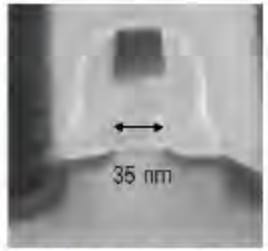
The MOSFET

Metal-Oxide-Semiconductor Field-Effect Transistor:

Desired characteristics:

- **High ON current**
- **Low OFF current**

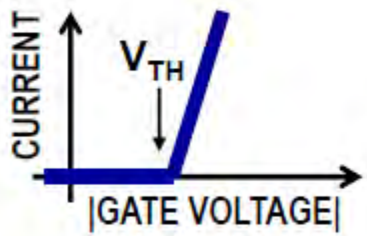




M. Bohr, Intel Developer Forum, September 2004

- Current flowing between the SOURCE and DRAIN is controlled by the voltage on the GATE electrode
- **"N-channel" & "P-channel" MOSFETs operate in a complementary manner**

  "CMOS" = Complementary MOS



CMOS Power Dissipation

- Three reasons for the large power dissipation:
  - The state variable is a charge. Moving charges provides energy dissipation
  - In Von Neumann architecture, memory and information processing are separated. Large number of memory access to perform complex calculation
  - MOS transistors cannot be scaled anymore. The bias voltage cannot be significantly smaller than 1 V.

The post Von Neumann-CMOS Era

Two objectives: better energy efficiency and new functionalities

- Evolutionary concepts:
(new switch, state variable other than charge),….. but binary logic

- Revolutionary concepts
(liquid state, quantum, neuromorphic, …

PERSPECTIVE

It's Time to Reinvent the Transistor!

Thomas N. Theis* and Paul M. Solomon

26 MARCH 2010 VOL 327 **SCIENCE**

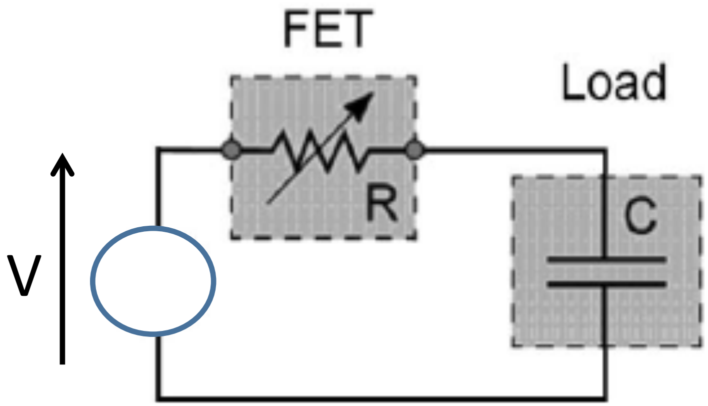
Evolutionary Concepts: Possible Ways

- Improved Von Neuman (Adiabatic logic, computing in memory)
- Improved switching device (new transistor, other technology)
- New state variable (New hardware)
- New information processing paradigm (quantum computing, neuromorphic)

Post ‘Von Neumann- CMOS’ Era

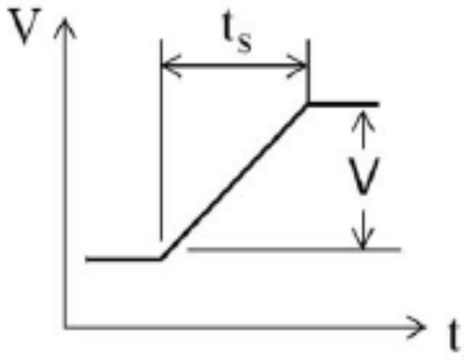
	Binary	**Von Neumann**	**CMOS**
Adiabatic logic, Computing in memory	✔	✔	✔
New switches	✔	✔	TFET, SET…
New state variables	✔	✔	Dipole, molecule, spin….
Quantum Computing	Quantum state	Quantum Turing machine	Photons, ions, atoms, RMN, superconducting
Neuromorphic	Spikes	ANNs, learning	✔, memristors

Capacitance charging at constant voltage: an energetic disaster!



If V= constant , E_d=1/2 CV^2

If V(t)



t_s>>RC, E_d= 1/2CV^2*(RC/t_s)

The principle of adiabatic logic is to reduce the power dissipation during switching using a time variable voltage source instead of a DC-voltage source.

New Switch Zoology

- The goal is to reduce V_{dd}
- New transistors: to decrease the threshold voltage V_{th} and to obtain steep switching
    - Tunnel FETs (TFET)
    - Ionization FETS (i-MOS)
    - Exitonic FETs (ExFET)
    - Bilayer spin FETs (BiSFET)
    - Resonant tunneling FETs (RIEFET)
    - Modulation velocity transistors (VMT)
  - New switching devices
    - NEM Switches

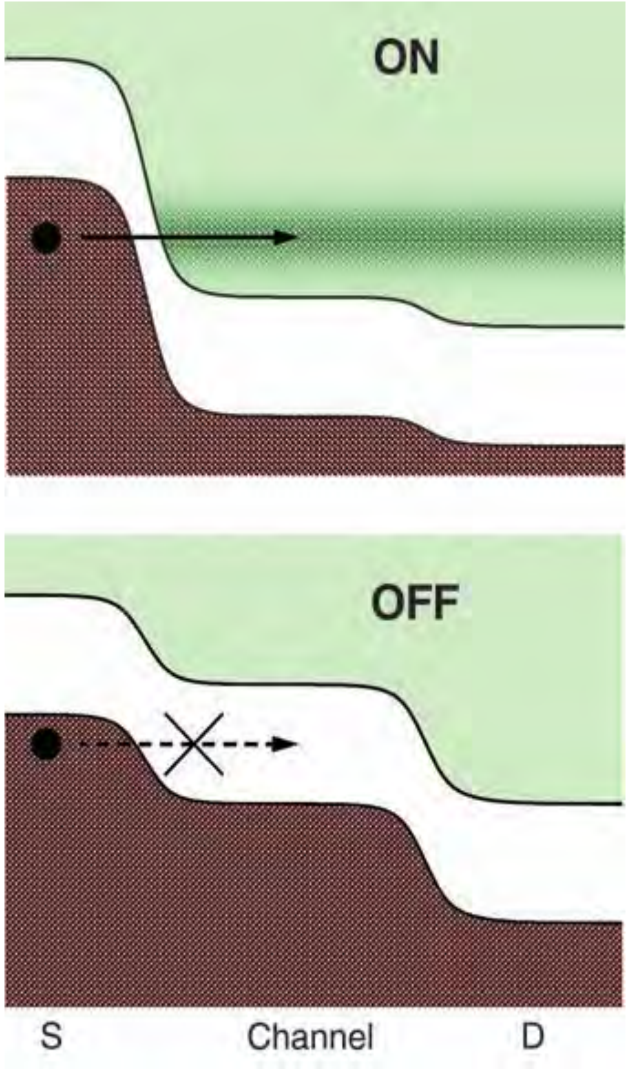
PERSPECTIVE

It's Time to Reinvent the Transistor!

Thomas N. Theis* and Paul M. Solomon

26 MARCH 2010 VOL 327 **SCIENCE**

New transistors



Tunnel FET

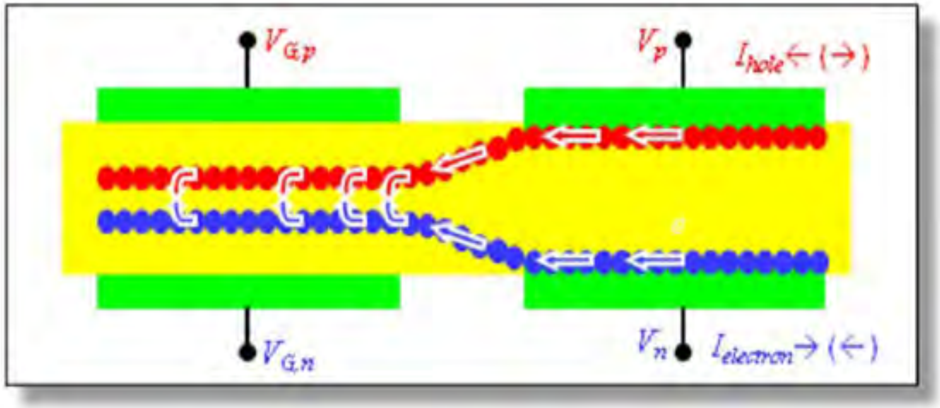


Fig. 7. *BiSFET consists of two graphene layers separated by a tunnel oxide: (a) schematic and (b) expected transistor characteristics.*

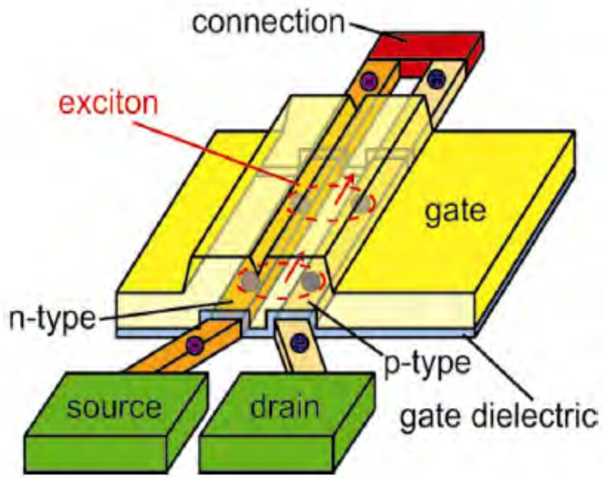
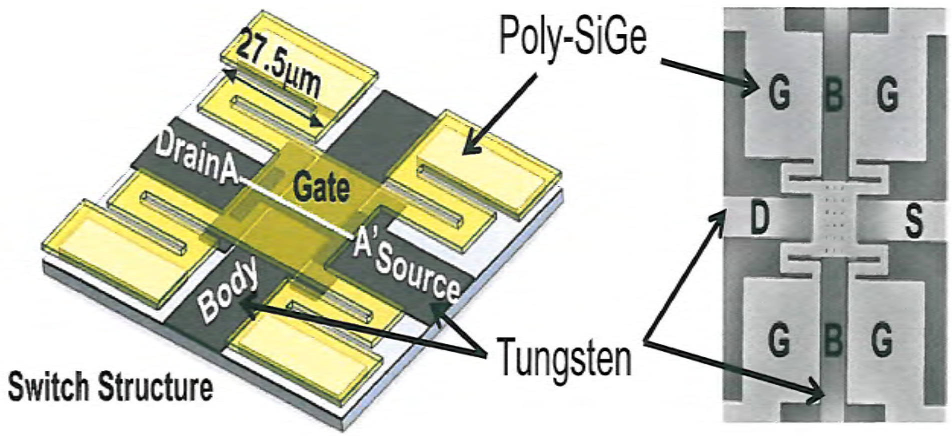
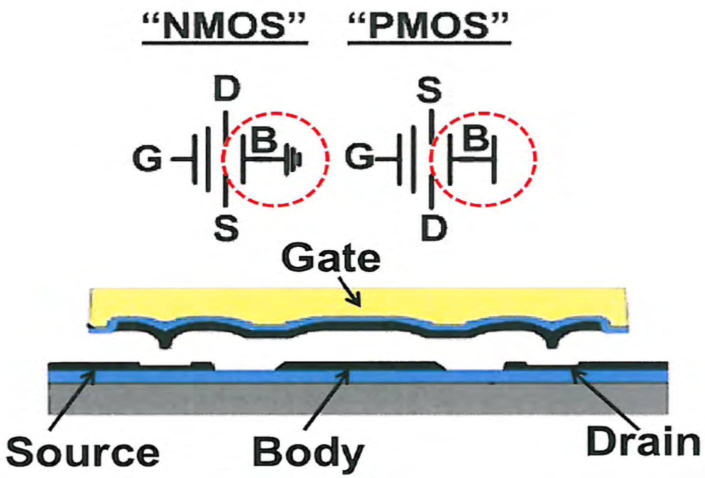
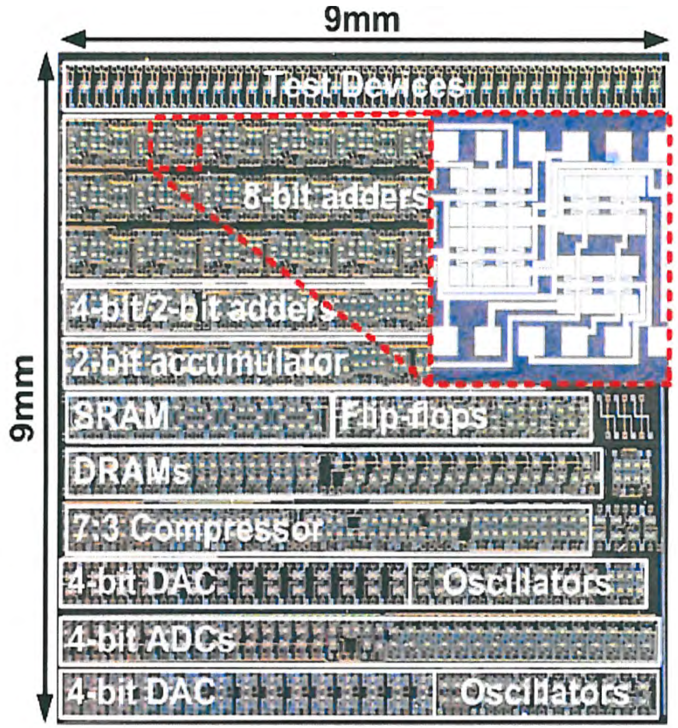


Fig. 8. *Schematic layout of the excitonic field-effect transistor (ExFET) (J. Appenzeller, Purdue University unpublished).*

NEMS



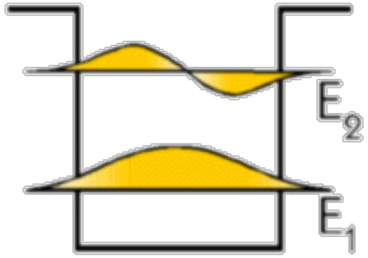


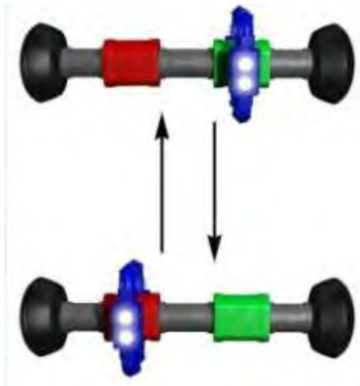


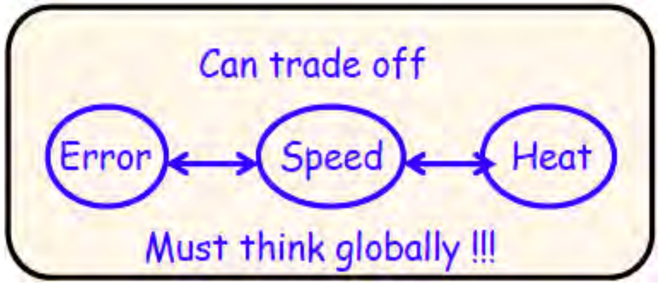
2010 ISSCC Jack Raper award for outstanding technology directions

State variable other than charge

- The problem of power dissipation is partly related to the fact that the state variable is a charge (E=QV). Other state variables can be proposed
  - Spin
  - photons, phonon
  - Quantum state
  - Quantum magnetic flux
  - Mechanical deformation
  - Dipolar orientation
  - Molecular state
  - …….







Advanced VN Technologies

during the last decades

- 1000’s of scientists worldwide
- 1000’s of journal papers
- 1000’s of conference presentations

No solution found !

Evolution or Revolution ?

- Improved Von Neumann (Adiabatic logic, computing in memory)
- Improved switching device (new transistor, other technology)
- New state variable ([illegible] hardware)

→ New information processing paradigm

New Information Processing Paradigms

- Non binary coding : analog, quantum, spikes…
- At least two possible approaches…..
  - Quantum computing
  - Neuromorphic computing

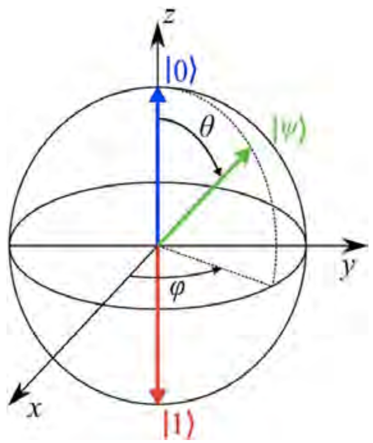
- Multidisciplinary approach mandatory: physicist, information theory, computer architecture, neuroscientists ….
- Proof of concepts needed at system level.

Quantum Information Processing

- Qbits

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

$$\alpha, \beta \in \mathbb{C} \qquad |\alpha|^2 + |\beta|^2 = 1$$



- The outcomes of measurements on quantum systems are random.
- Quantum systems evolve according to linear equations. Solutions obey a superposition principle: linear combinations of solutions are still solutions.

Quantum Parallelism

- 50 coin tosses, heads or tails

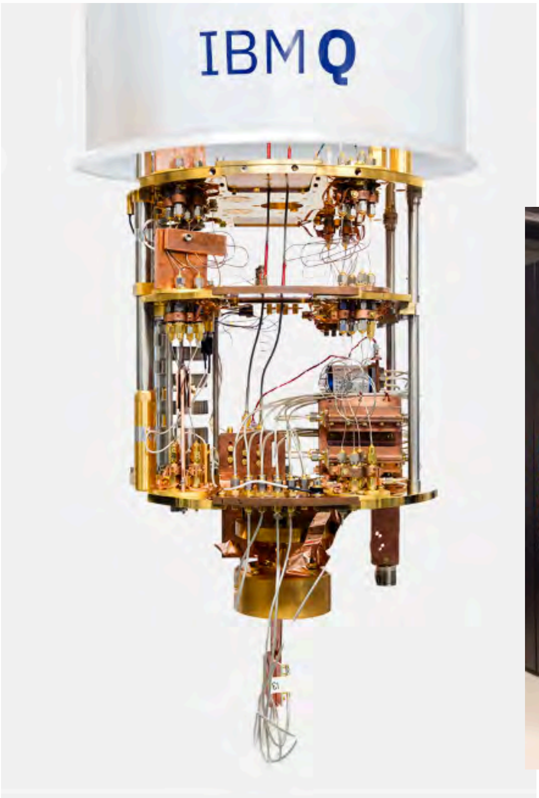
→ Hundred terabytes of data storage

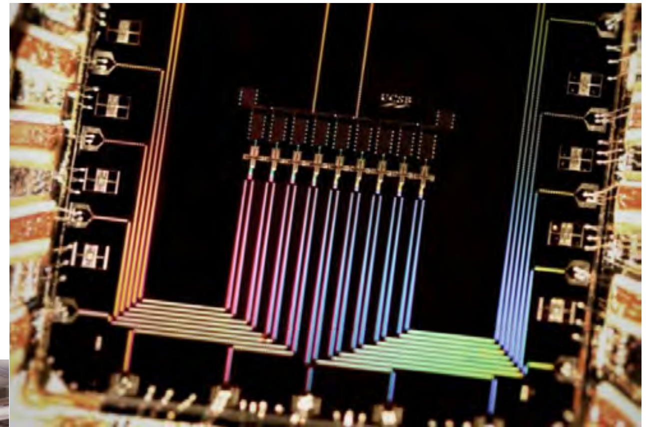
- 50 Qbits can store the same information

Impressive Recent Progress

IBM Raises the Bar with a 50-Qubit Quantum Computer

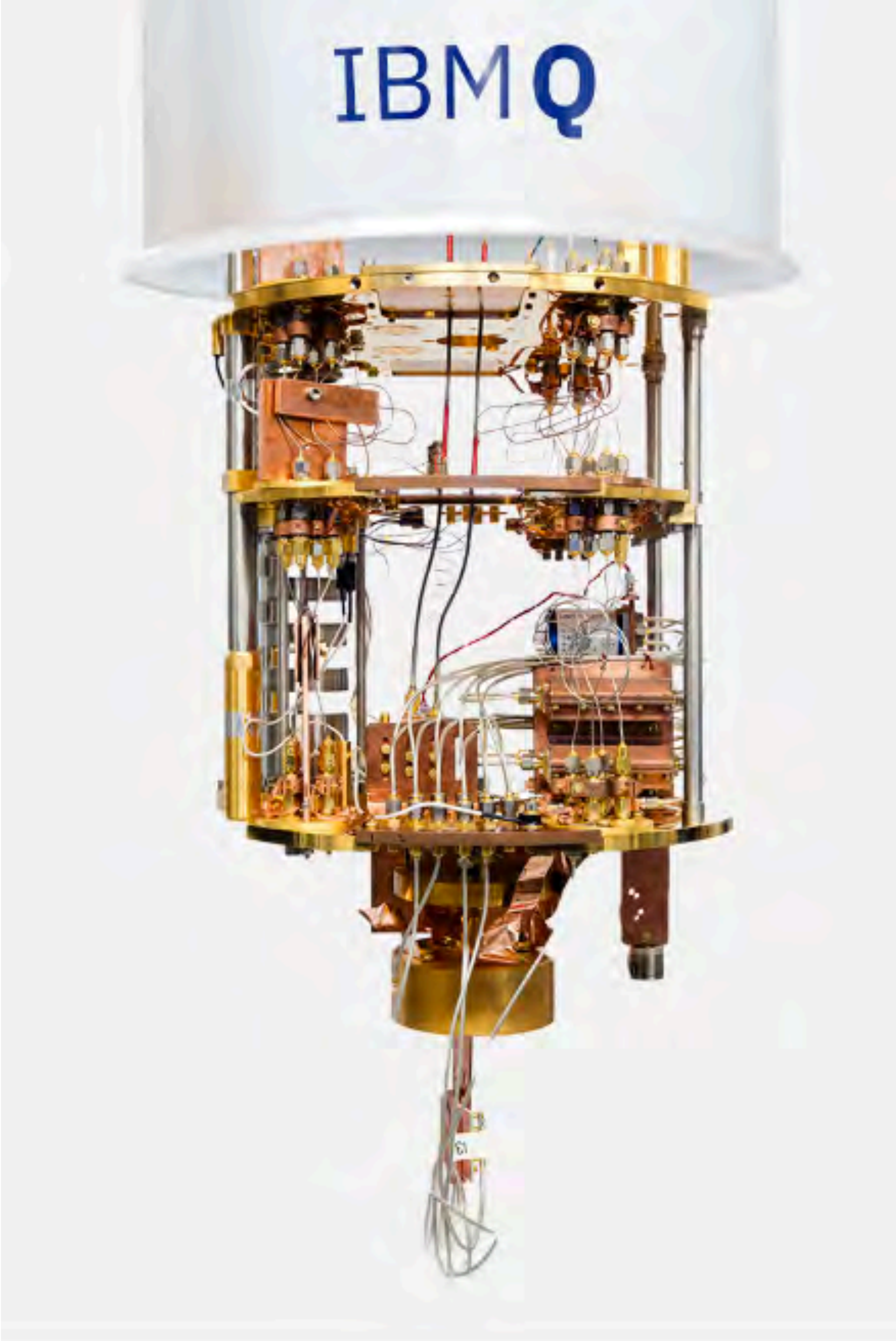
Google quantum computer test shows breakthrough is within reac





D:wave

The Quantum Computing Company™

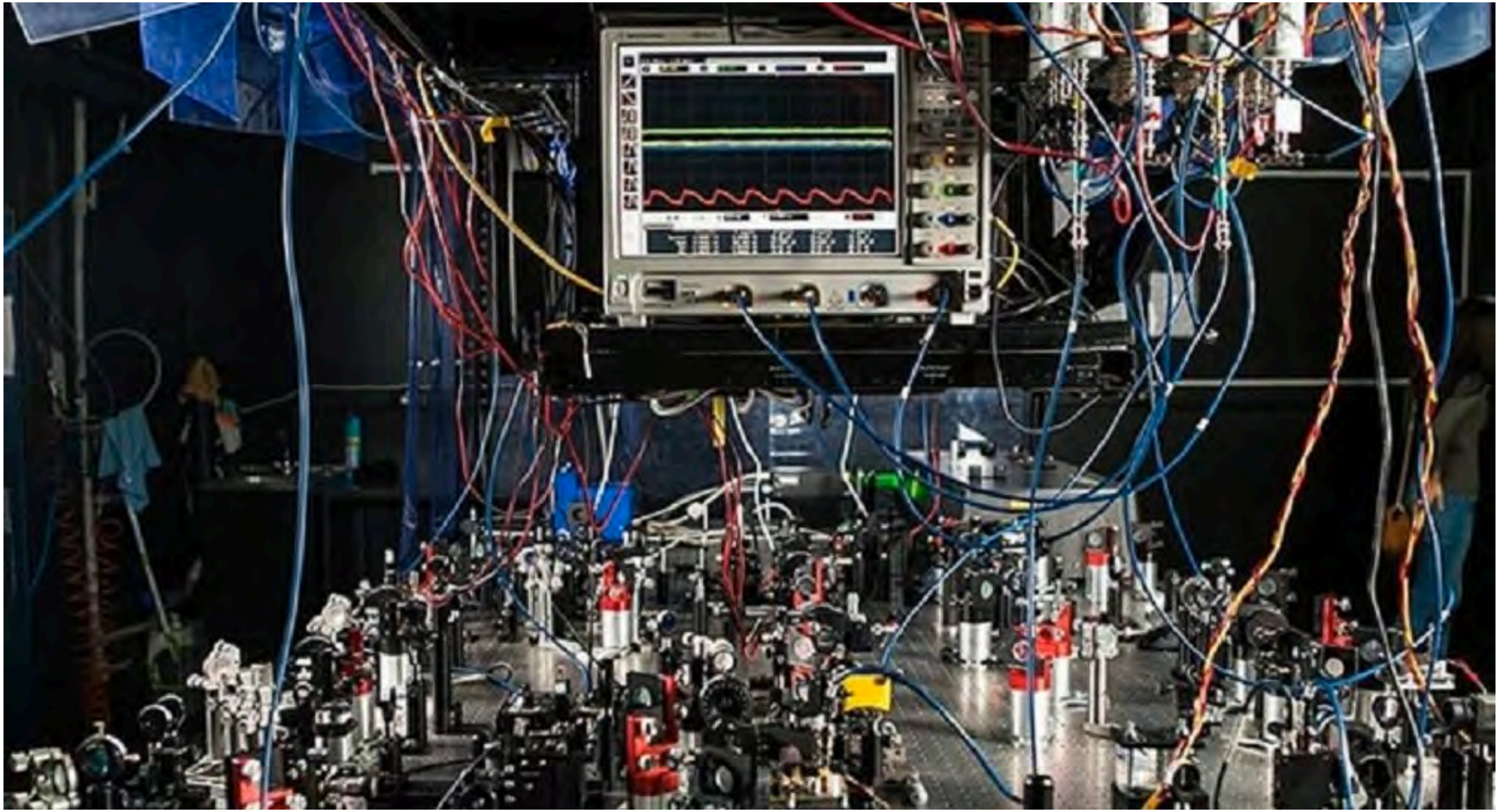

IBM Q

But…

- Several orders of magnitude slower than expected
- No quantum algorithm library
- Quantum state coherence time limited (100 µS only)
- Operation temperature often close to 0°K (-273 °C) !

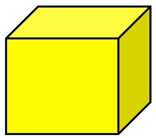
- Quantum computers often resemble fundamental physics experiments !

This is a quantum computer .....



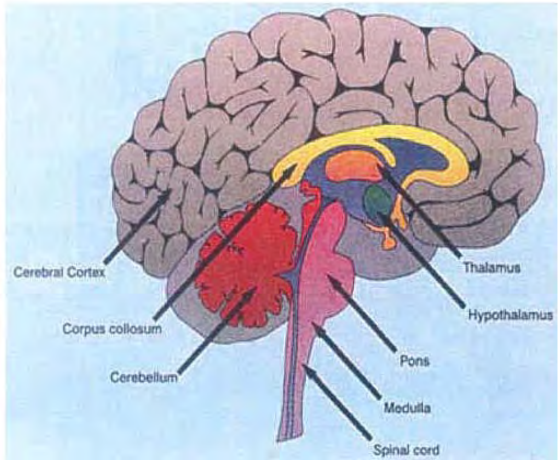
Neuromorphic architectures

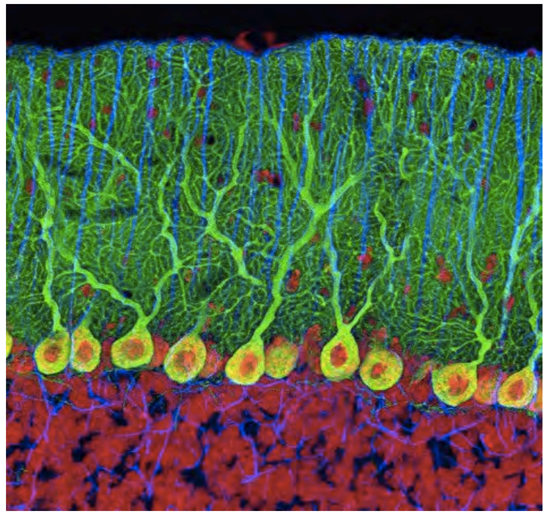
- The most complex information processing system known to date.
- 2% of the human weight but 20 % of the consumed energy
- In human, $\sim 10^{11}$ neurons and $\sim 10^{4}$ synapses/neurons



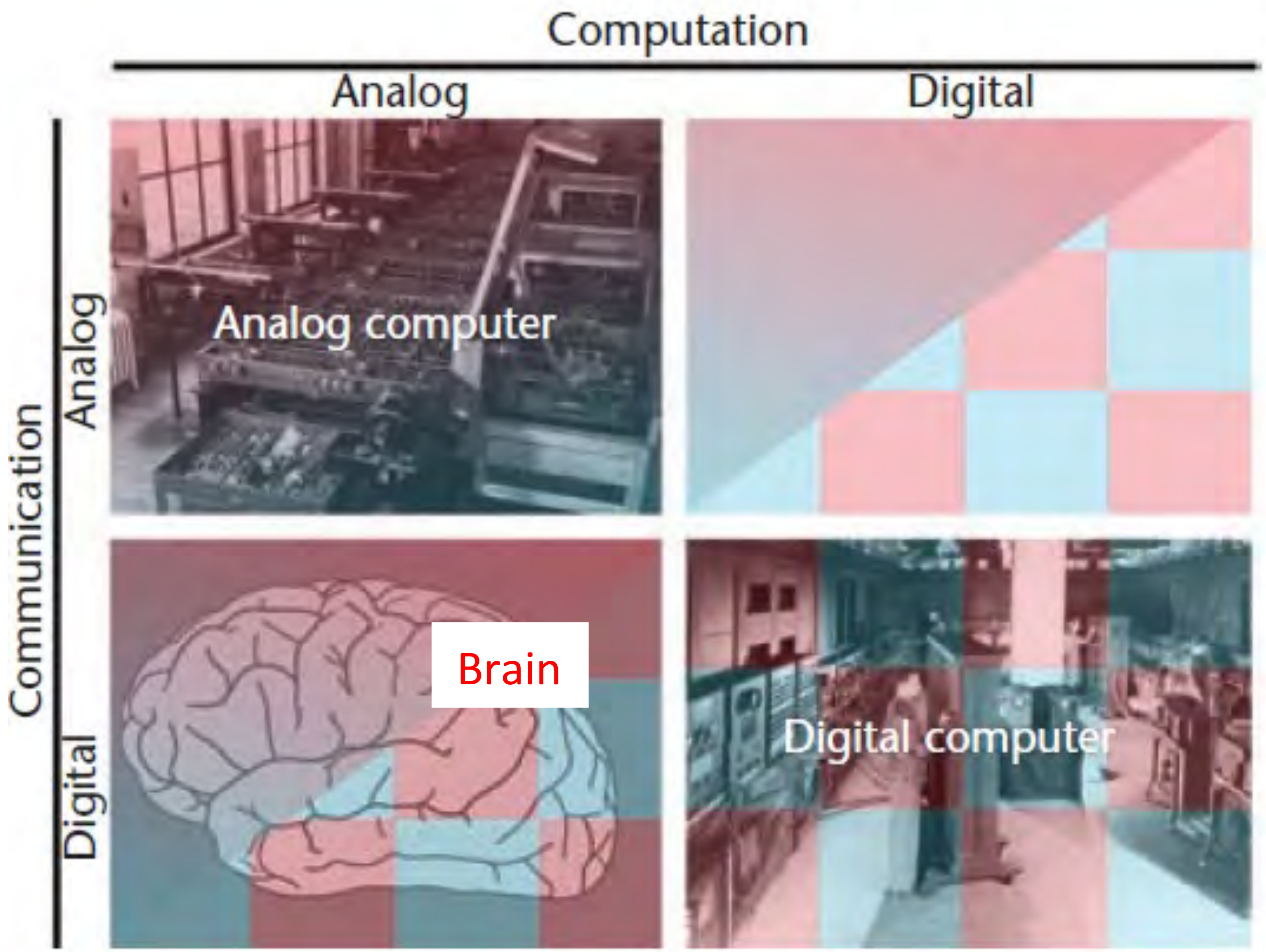
1 mm^3 cortex

5×10^4 to 10^5 neurons
3×10^8 to 7×10^8 synapses
4 km of axon
0,5 km of dendrites





Computing and Communication



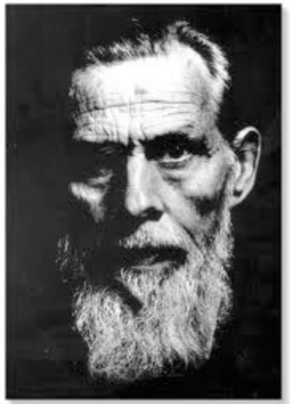
Kwabena Boahen

A Neuromorph's Prospectus

Computing in Science & Engineering
Copublished by the IEEE CS and the AIP
March/April 2017

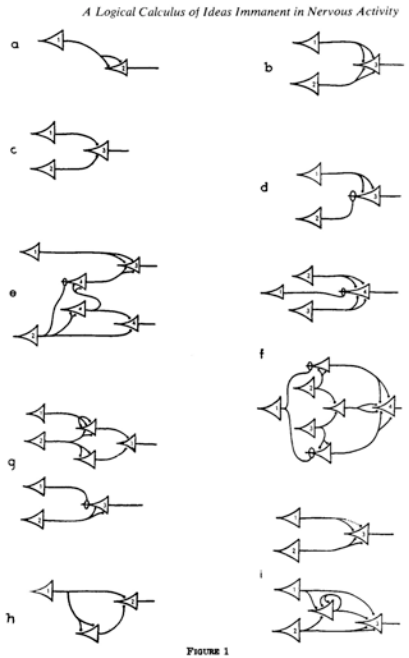
Bioinspired Information Processing : the Pioneers

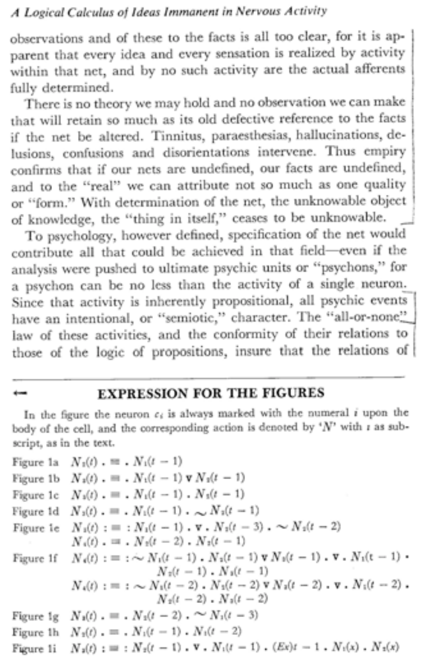
[1] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” Bull. Math. Biophysics, no. 5, pp. 115-133, **1943**.



Warren McCulloch

Walter Pitts





A Logical Calculus of Ideas Immanent in Nervous Activity

observations and of these to the facts is all too clear, for it is apparent that every idea and every sensation is realized by activity within that net, and by no such activity are the actual afferents fully determined.

There is no theory we may hold and no observation we can make that will retain so much as its old defective reference to the facts if the net be altered. Tinnitus, paraesthesias, hallucinations, delusions, confusions and disorientations intervene. Thus empiry confirms that if our nets are undefined, our facts are undefined, and to the “real” we can attribute not so much as one quality or “form.” With determination of the net, the unknowable object of knowledge, the “thing in itself,” ceases to be unknowable.

To psychology, however defined, specification of the net would contribute all that could be achieved in that field—even if the analysis were pushed to ultimate psychic units or “psychons,” for a psychon can be no less than the activity of a single neuron. Since that activity is inherently propositional, all psychic events have an intentional, or “semiotic,” character. The “all-or-none” law of these activities, and the conformity of their relations to those of the logic of propositions, insure that the relations of

EXPRESSION FOR THE FIGURES

In the figure the neuron c_i is always marked with the numeral i upon the body of the cell, and the corresponding action is denoted by 'N' with i as subscript, as in the text.

Figure 1a $N_2(t) . \equiv . N_1(t-1)$

Figure 1b $N_3(t) . \equiv . N_1(t-1) \vee N_2(t-1)$

Figure 1c $N_3(t) . \equiv . N_1(t-1) . N_2(t-1)$

Figure 1d $N_3(t) . \equiv . N_1(t-1) . \sim N_2(t-1)$

Figure 1e $N_3(t) : \equiv : N_1(t-1) . \vee . N_2(t-3) . \sim N_2(t-2)$
$N_4(t) . \equiv . N_2(t-2) . N_2(t-1)$

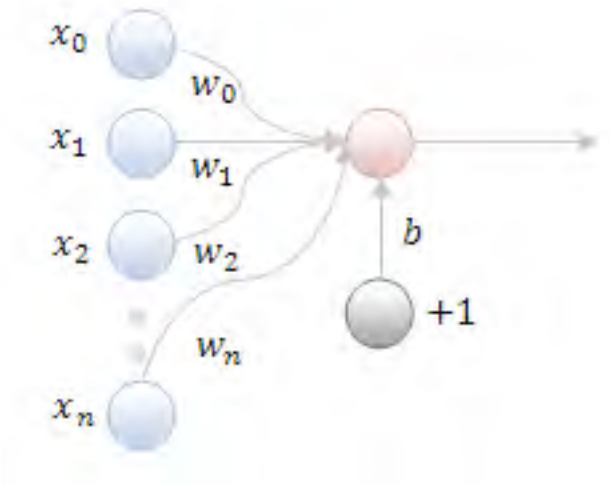
Figure 1f $N_4(t) : \equiv : \sim N_1(t-1) . N_2(t-1) \vee N_3(t-1) . \vee . N_1(t-1) . N_2(t-1) . N_3(t-1)$
$N_4(t) : \equiv : \sim N_1(t-2) . N_2(t-2) \vee N_3(t-2) . \vee . N_1(t-2) . N_2(t-2) . N_3(t-2)$

Figure 1g $N_3(t) . \equiv . N_2(t-2) . \sim N_1(t-3)$

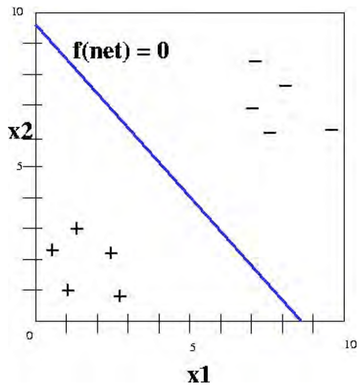
Figure 1h $N_2(t) . \equiv . N_1(t-1) . N_1(t-2)$

Figure 1i $N_3(t) : \equiv : N_2(t-1) . \vee . N_1(t-1) . (Ex)t-1 . N_1(x) . N_2(x)$

A first neuromorphic engine, the perceptron

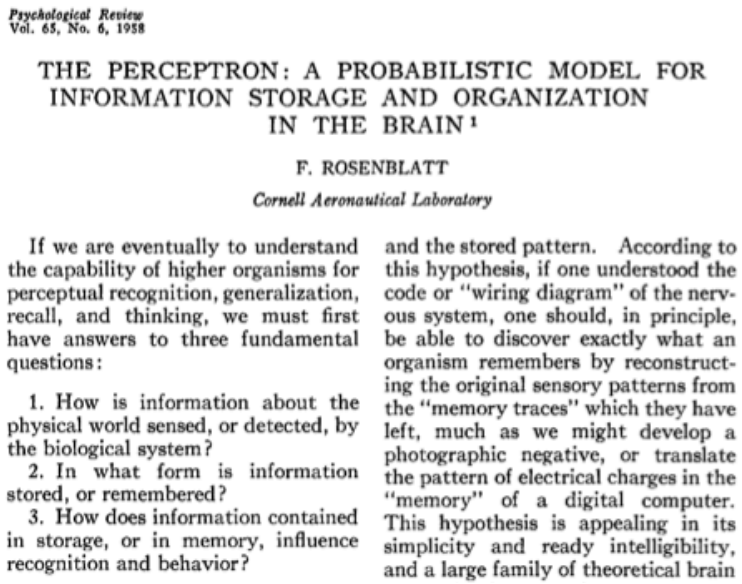


[1] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," Psychological Review, vol. 65, no. 6, pp. 386-408, **1958**.



(Robert Hecht-Nilsen: Neurocomputing, Addison-Wesley, 1990)

Frank Rosenblatt



Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN [1]

F. ROSENBLATT

Cornell Aeronautical Laboratory

If we are eventually to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking, we must first have answers to three fundamental questions:

1. How is information about the physical world sensed, or detected, by the biological system?
2. In what form is information stored, or remembered?
3. How does information contained in storage, or in memory, influence recognition and behavior?

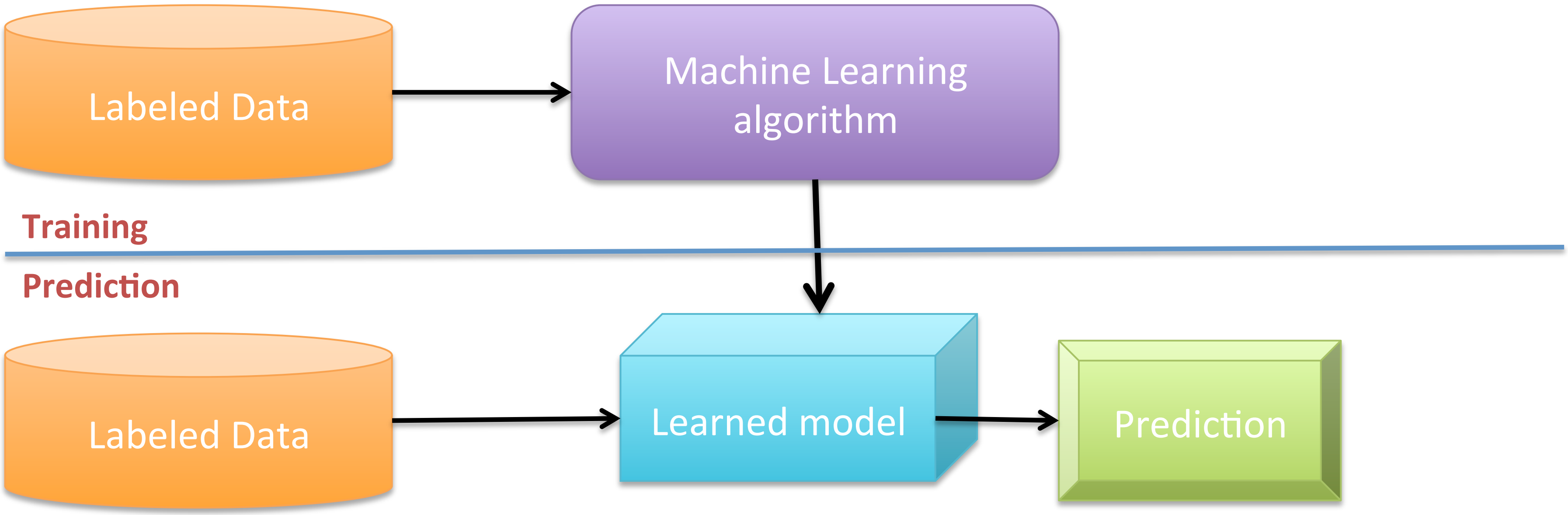
and the stored pattern. According to this hypothesis, if one understood the code or "wiring diagram" of the nervous system, one should, in principle, be able to discover exactly what an organism remembers by reconstructing the original sensory patterns from the "memory traces" which they have left, much as we might develop a photographic negative, or translate the pattern of electrical charges in the "memory" of a digital computer. This hypothesis is appealing in its simplicity and ready intelligibility, and a large family of theoretical brain

John Von Neumann, also thought about neuro-computation and tried, for the first time, to construct a meaningful comparison between brain and computer power.

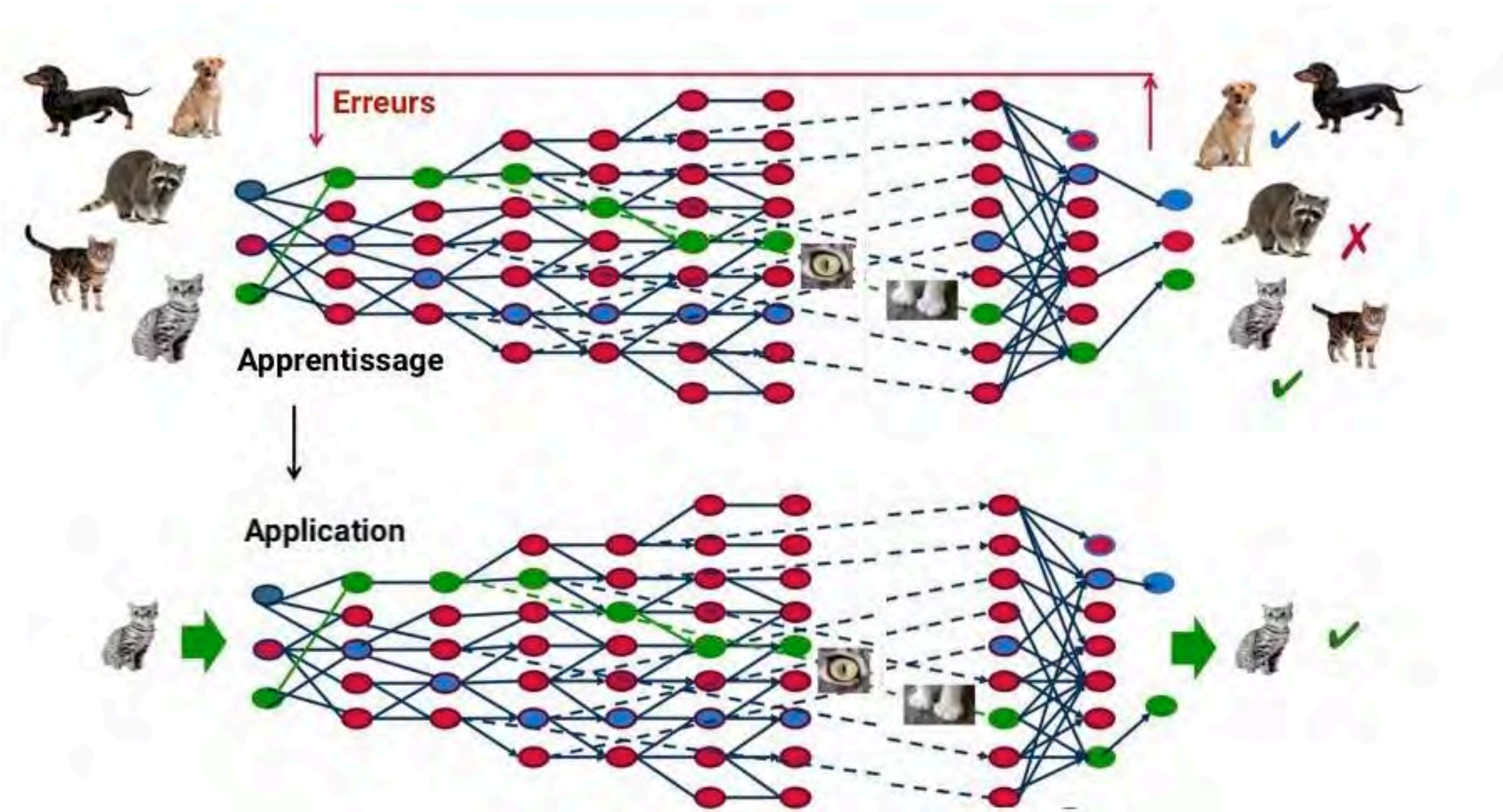
- *"The Language of the Brain is Not the Language of Mathematics … whatever language the central nervous system is using, it is characterised by less logical and arithmetical depth than what we are normally used to … Consequently, there exist here different logical structures from the ones we are ordinarily used to...*

(John Von Neumann, *The Computer and the Brain*, 1958. Published *post-mortem*)

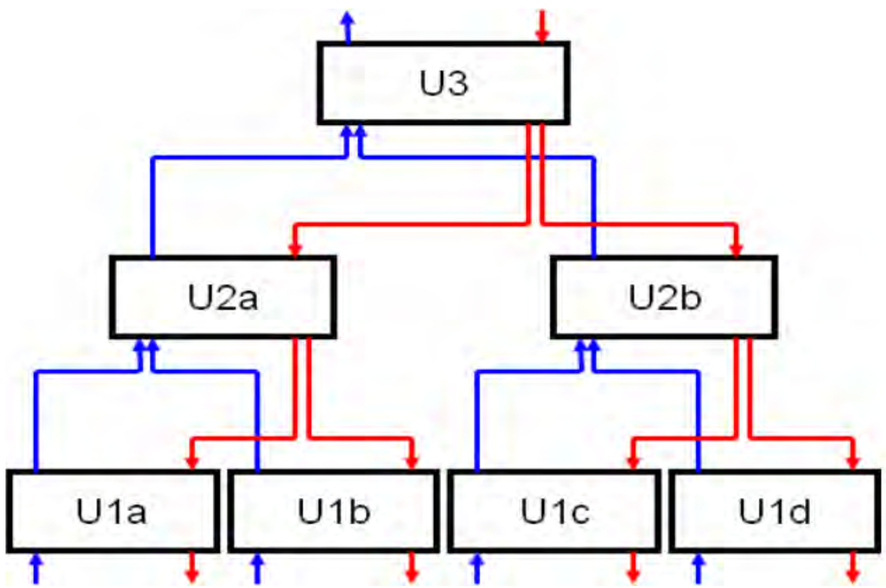
Machine Learning

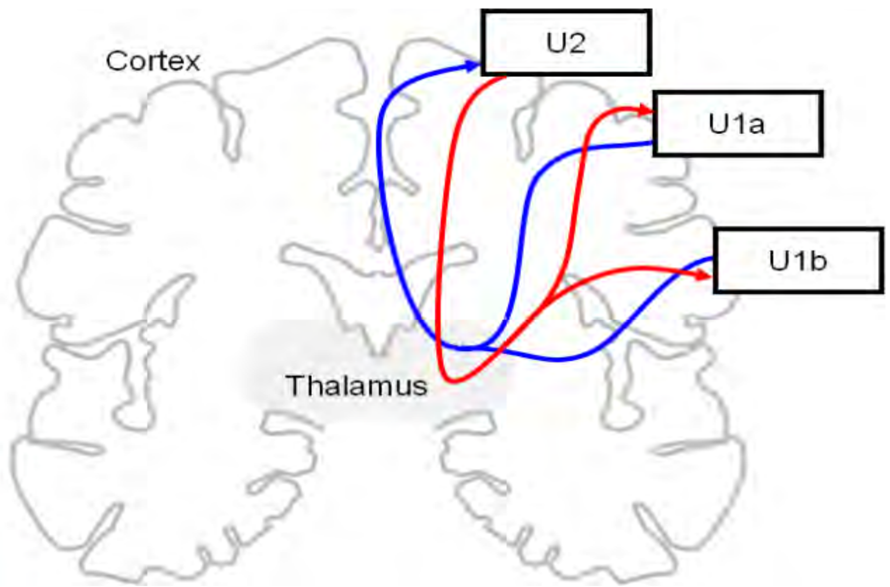


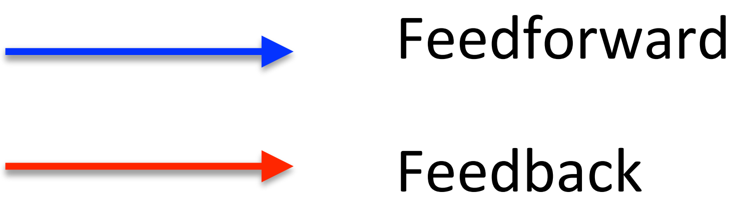
Deep Learning



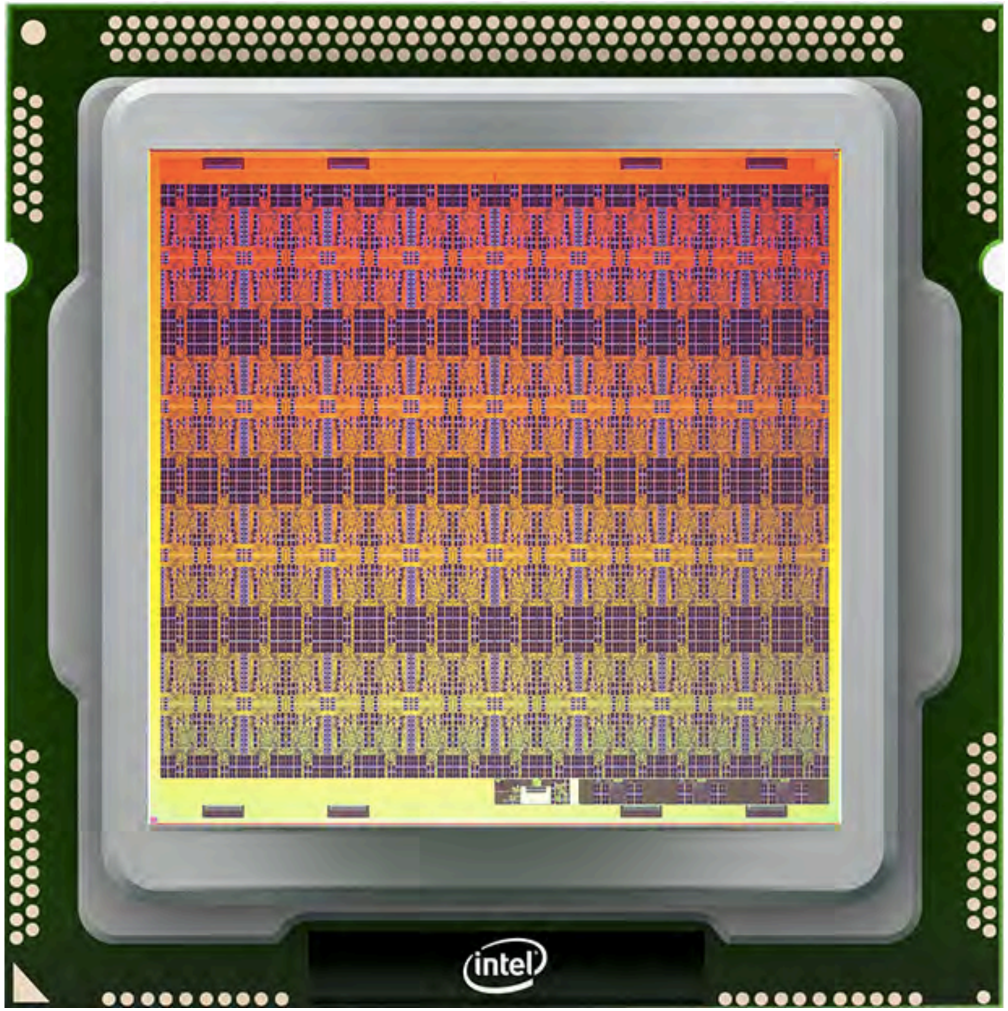
Cortical Neural Network





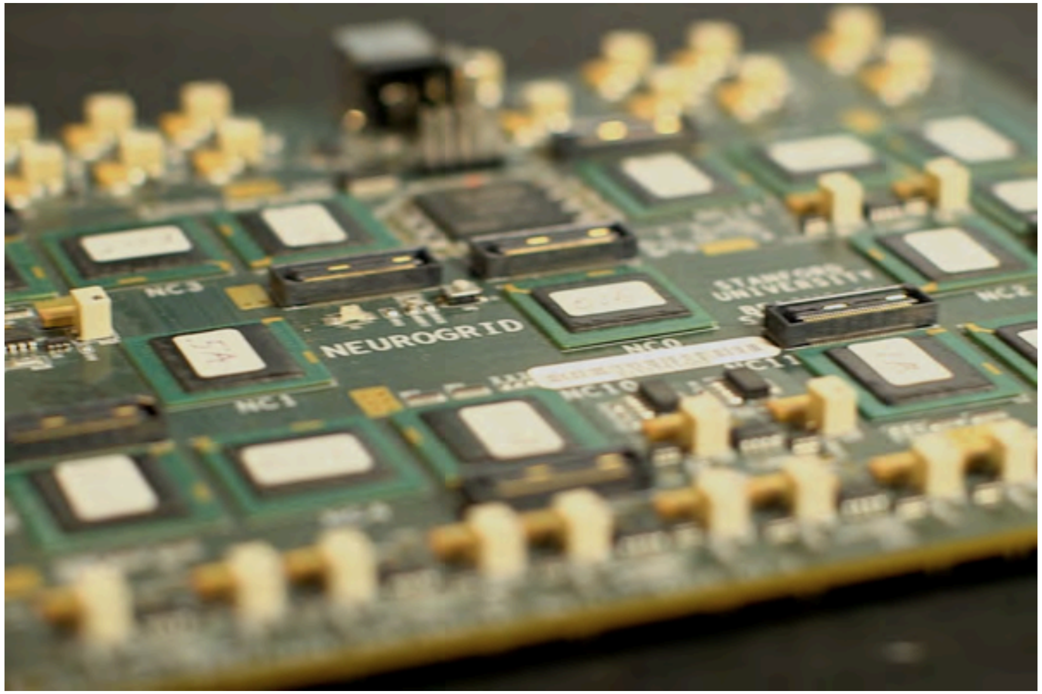


Loihi Intel



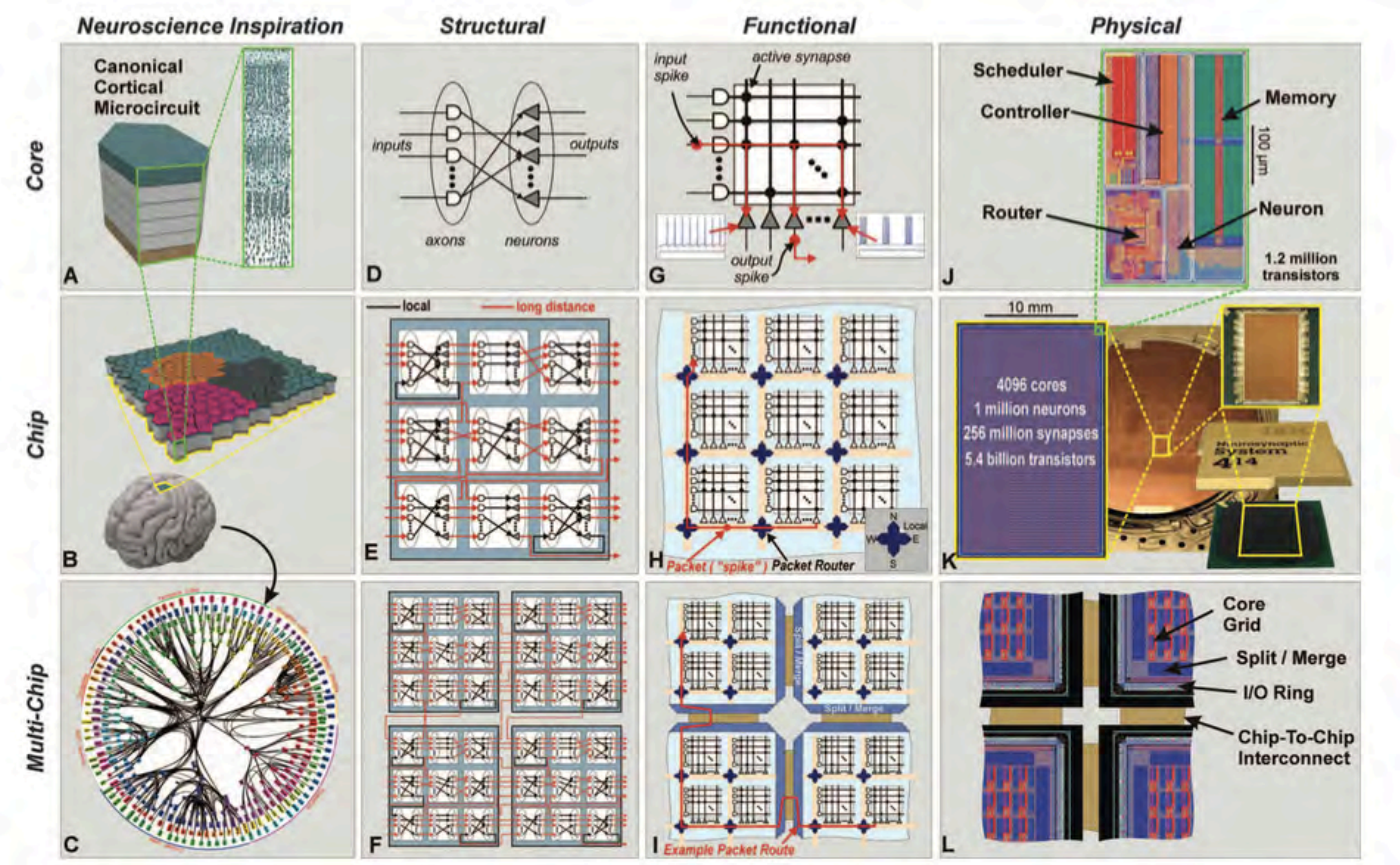
130 000 neurons and 130 M synapses

IBM TrueNorth architecture



1M spiking neurons and 256 M plastic synapses

IBM TrueNorth architecture



1 M neurons and 256 M synapses

Neuromorphic Computing

- Highs and lows during the last decades
- Interest renewal for AI
- Emulation of neural networks on conventional computer
  - Rather simple architectures
  - Energy question not addressed
- Specific hardware
  - Scaling up difficult
  - Interconnects

Conclusion

L’informatique du nouveau monde

VN-CMOS

Merci de votre attention