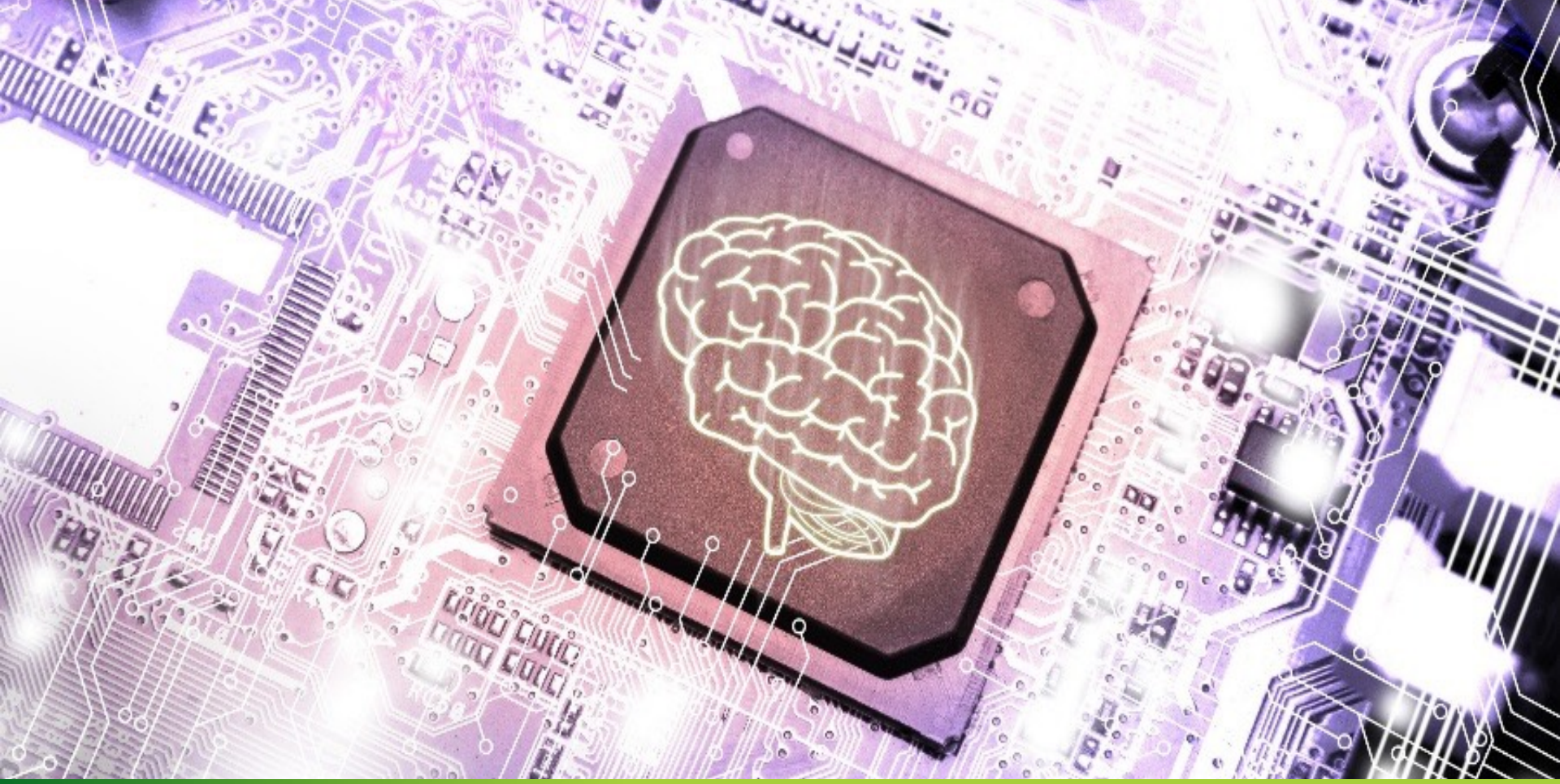


LES CHALLENGES POSÉS PAR LES SYSTÈMES DE CALCUL POUR LES APPLICATIONS CYBER-PHYSIQUES COGNITIVES

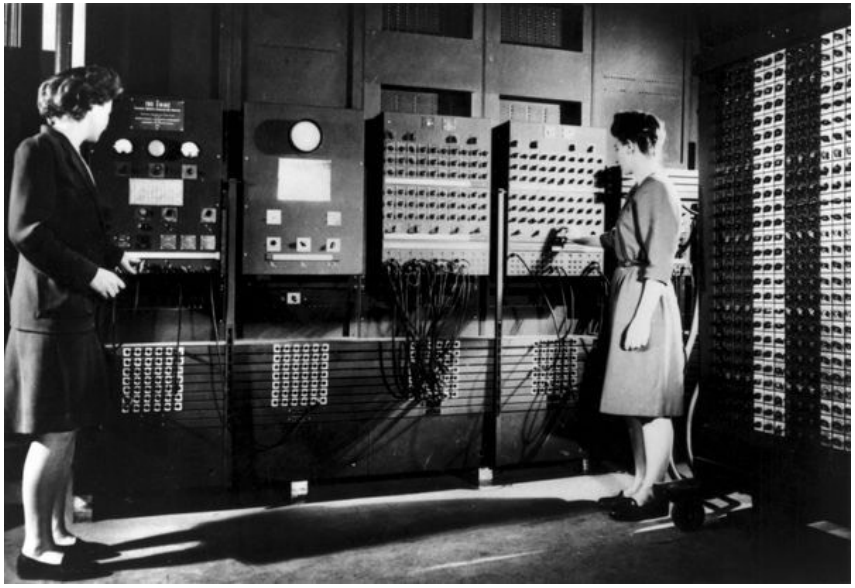
Marc Duranton
CEA Fellow
Commissariat à l'énergie atomique et aux énergies alternatives

Séminaire Aristote : Réinventer l'informatique
6 décembre 2018

“The best way to predict the future is to invent it.”

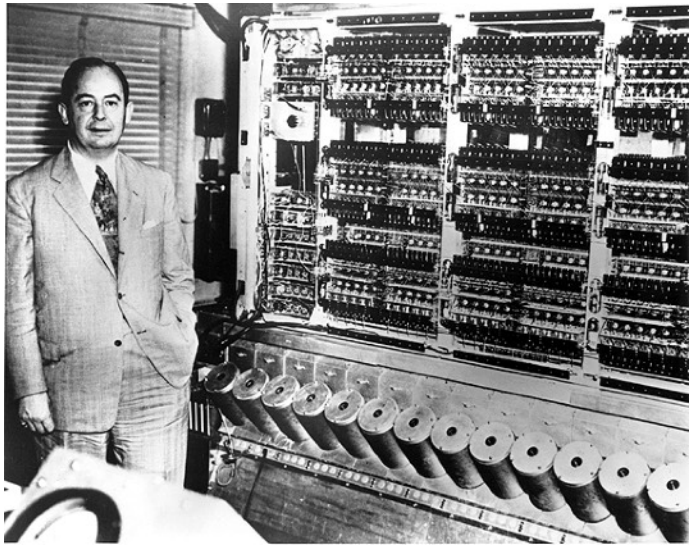
Alan Kay

LOOKING BACK… *1 COMPUTER FOR THE WHOLE PLANET*



ENIAC (Electronic Numerical Integrator And Computer), built between 1943 and 1945.
ENIAC contained 20,000 vacuum tubes,
7200 crystal diodes, 1500 relays. It weighed more than 27 t, was roughly 2.4 m × 0.9 m × 30 m in size, occupied 167 m^2 and consumed 150 kW of electricity.

From Wikipedia



EDVAC was delivered in 1949.Functionally, EDVAC was a binary serial computer with automatic addition, subtraction, multiplication, programmed division and automatic checking with an ultrasonic serial memory capacity of 1,000 44-bit words. EDVAC's average addition time was 864 microseconds and its average multiplication time was 2,900 microseconds.

LOOKING BACK… *1 COMPUTER PER (MAJOR) COUNTRY*

"I think there is a world market for maybe five computers."

Thomas Watson, president of IBM, 1943

LOOKING BACK… *1 COMPUTER PER HOUSE*

"There is no reason anyone would want a computer in their home."

Ken Olsen, founder of Digital Equipment Corporation, 1977

The Altair 8800 by MITS – 1974 - Intel 8080 CPU.
"1024 word" memory board populated with 256 bytes.
The BASIC language was announced in July 1975
and it required one or two 4096 word memory boards

IBM PC – 1981- Intel 8088 CPU.
Basic configuration 16K RAM.

From Wikipedia

LOOKING BACK… *1 ~~COMPUTER~~ SMARTPHONE PER PERSON*

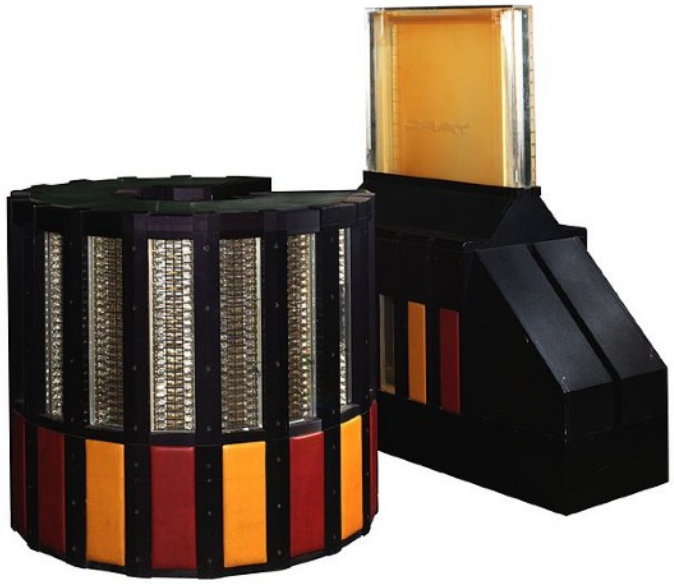
iPhone, introduced **June 29, 2007**;
Samsung 32-bit RISC ARM
Underclocked to 412 MHz
128 MB eDRAM Storage
4, 8 or 16 GB flash memory

From Wikipedia

EVOLUTION OF SOCIETY

Remember: the iPhone was introduced just 11 years ago…

Exponential increase of performances in 33 years



Cray 2 – 1985
2 GFLOPS (2×10^9 FLOPS)

X 100 000 000

Summit – 2018
200 PFLOPS (2×10^{17} FLOPS)

Peta = 10^{15} = million of milliard

Exponential increase of performances in 33 years

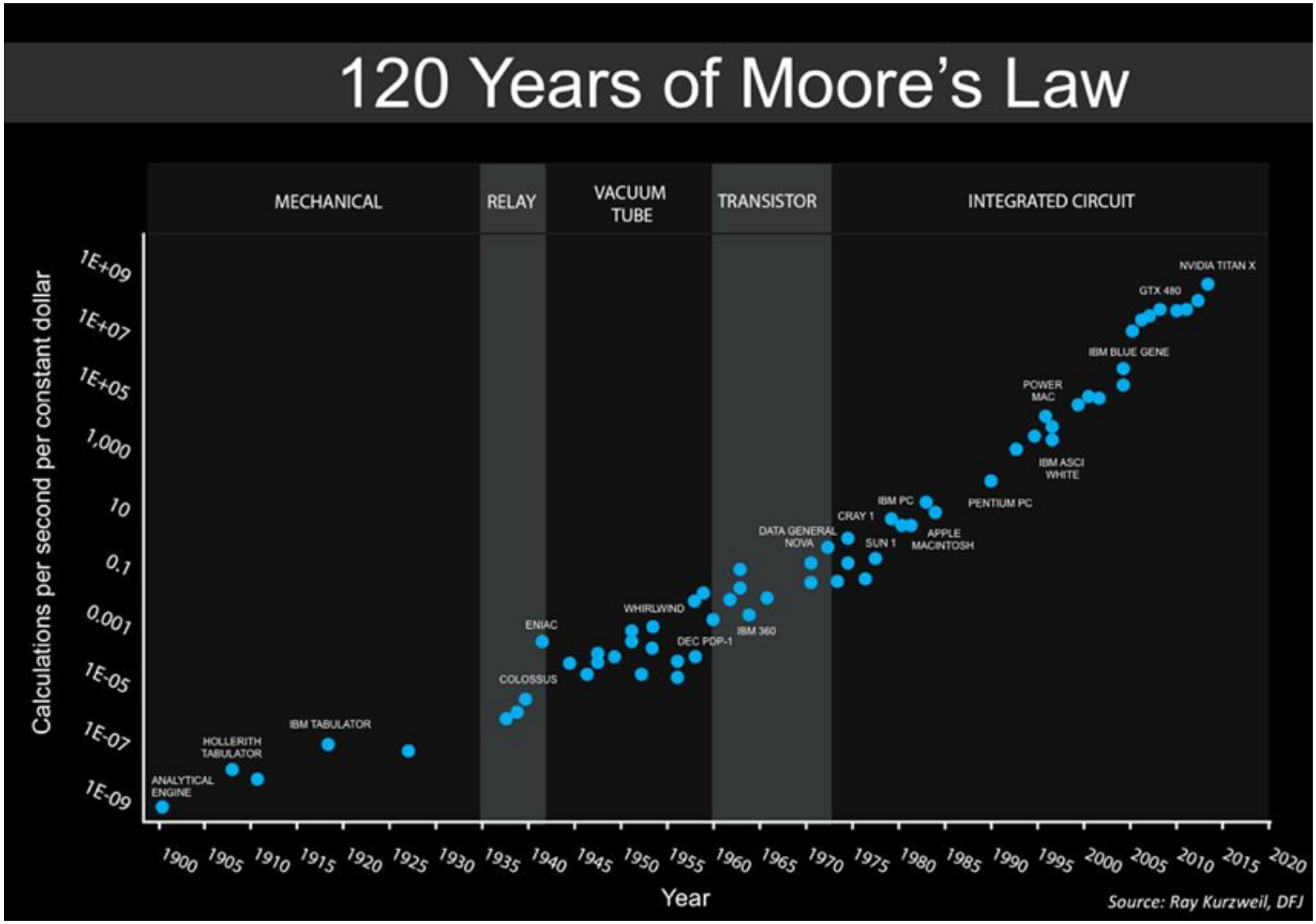
Production car of 1985
Lamborghini Countach 5000QV
Max speed 300 Km/h

X 100 000 000

27 times the speed of light
Warp 3 ?
Star Trek Enterprise
(Year: about 2290)

Peta = 10^{15} = million of milliard

PROGRESS OF COMPUTING TECHNOLOGY: CALCULATIONS PER SECOND AND PER DOLLAR



And after CMOS?

THE END OF ~~MOORE'S LAW~~ DENNARD SCALING

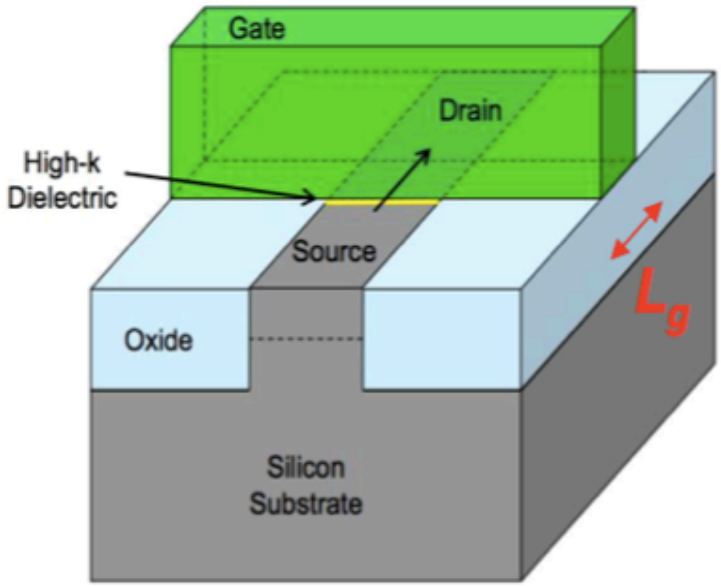
Cf. talk of Alain Cappy

Parameter (scale factor = a)	Classic Scaling	Current Scaling
Dimensions	1/a	1/a
Voltage	1/a	1
Current	1/a	1/a
Capacitance	1/a	>1/a
Power/Circuit	$1/a^2$	1/a
Power Density	1	a
Delay/Circuit	1/a	~1

Source: Krisztián Flautner "From niche to mainstream: can critical systems make the transition?"

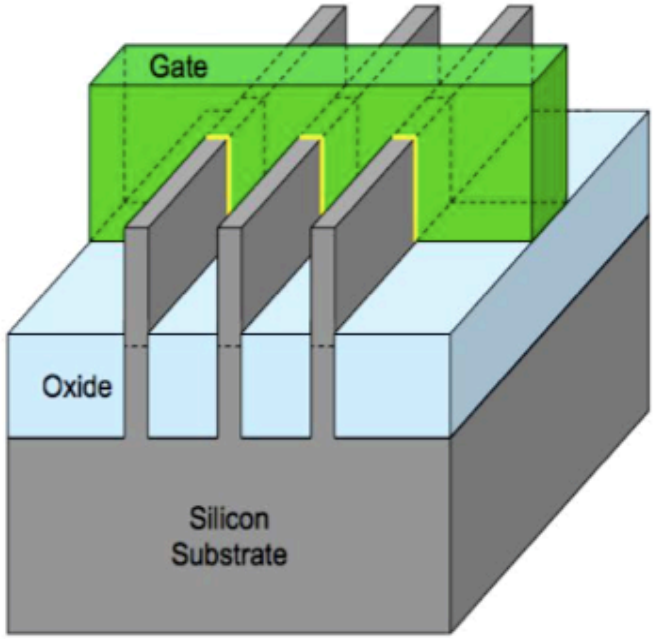
Technology evolution

Transistor 2D



Current control difficult when L_g< 20nm

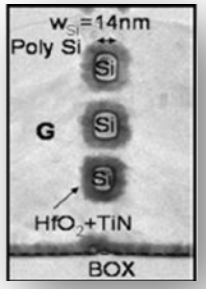
Transistor « Tri-gate » 3D

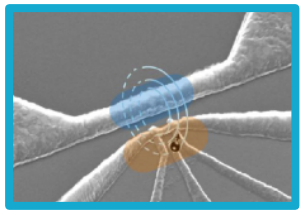


Better control (2 sides + multiple gates)

Technology evolution

Silicon Quantum bits



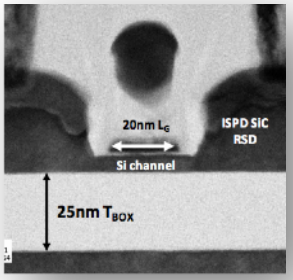


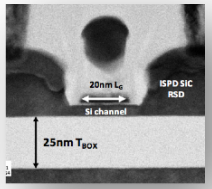
FDSOI

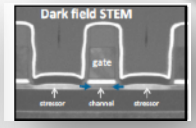
22FD

12FD

Next Gen







Non planar / trigate / stacked Nanowires

FinFET

28nm

10nm

2018

5nm

2017

14nm

7nm

Disruptive scaling

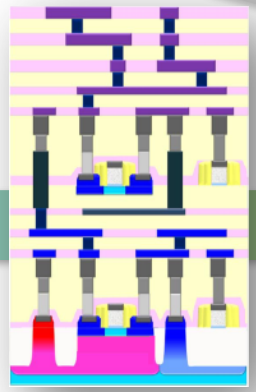
Hybrid logic

Steep slope devices

Mechanical switches

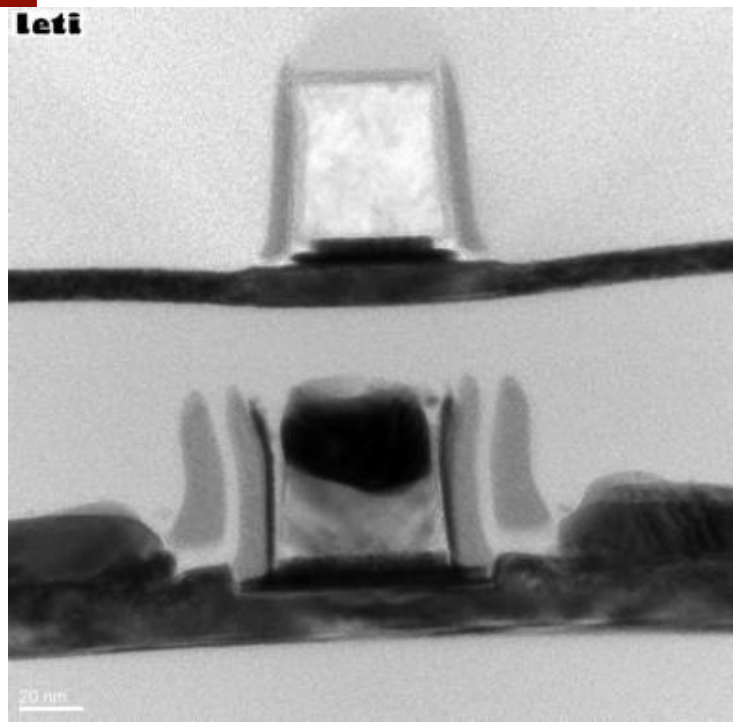
Si Quantum bits

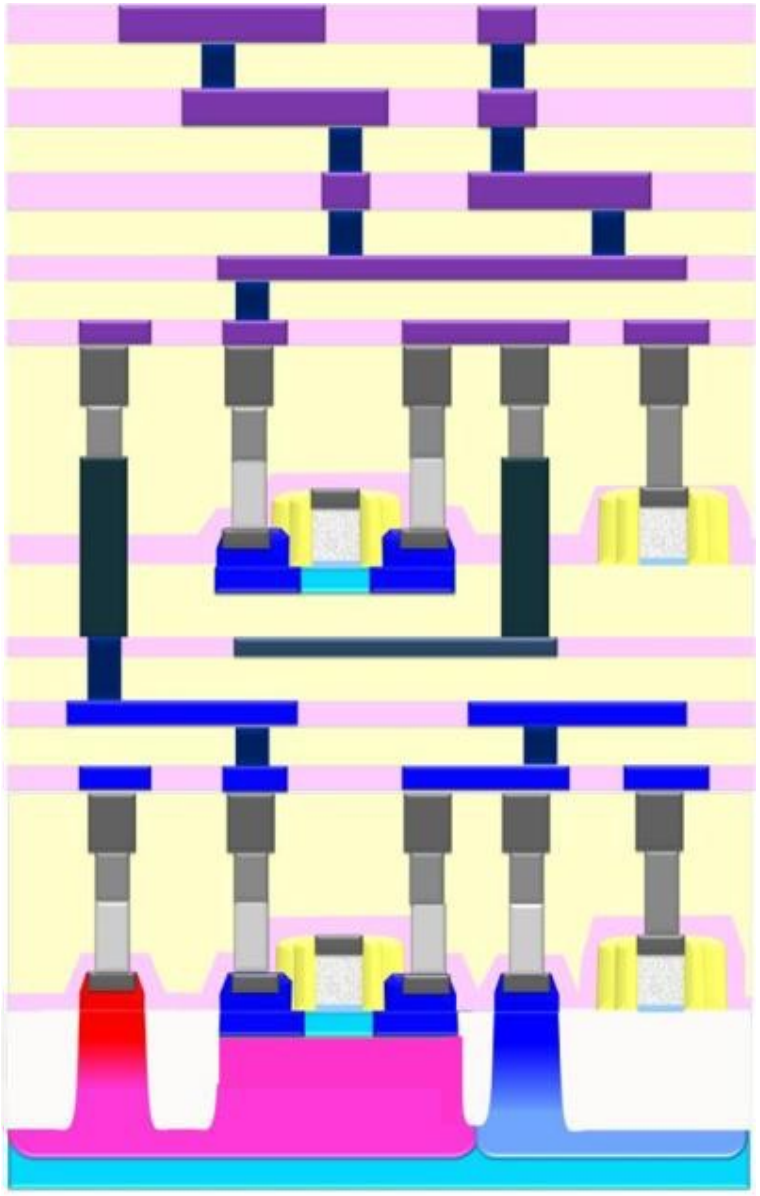
Alternative to scaling and diversification



Monolithic 3D for 3D VLSI

M3D PRINCIPLE





CMOS/CMOS: 14nm vs 2D:
Area gain=55%
Perf gain = 23%
Power gain = 12%

LETI, DAC 2014 QUALCOMM

LOOKING FORWARD…

What will be after the smartphone?

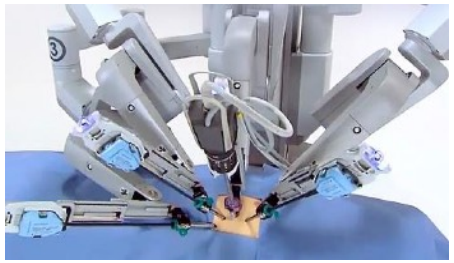
…

LOOKING FORWARD… HER (THE MOVIE)

Multiple “computers” closely linked (or implanted?) with the individual through an “**intelligent** interface”

Entering in
Human and machine collaboration era





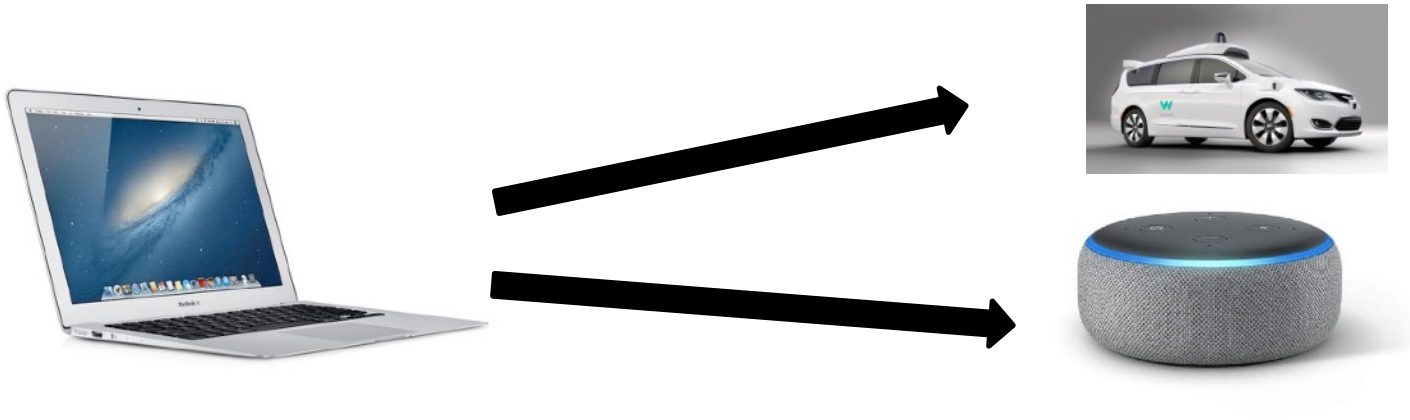



ENABLED BY ARTIFICIAL INTELLIGENCE
(AND DEEP LEARNING)

ENABLED BY ARTIFICIAL INTELLIGENCE (AND DEEP LEARNING)

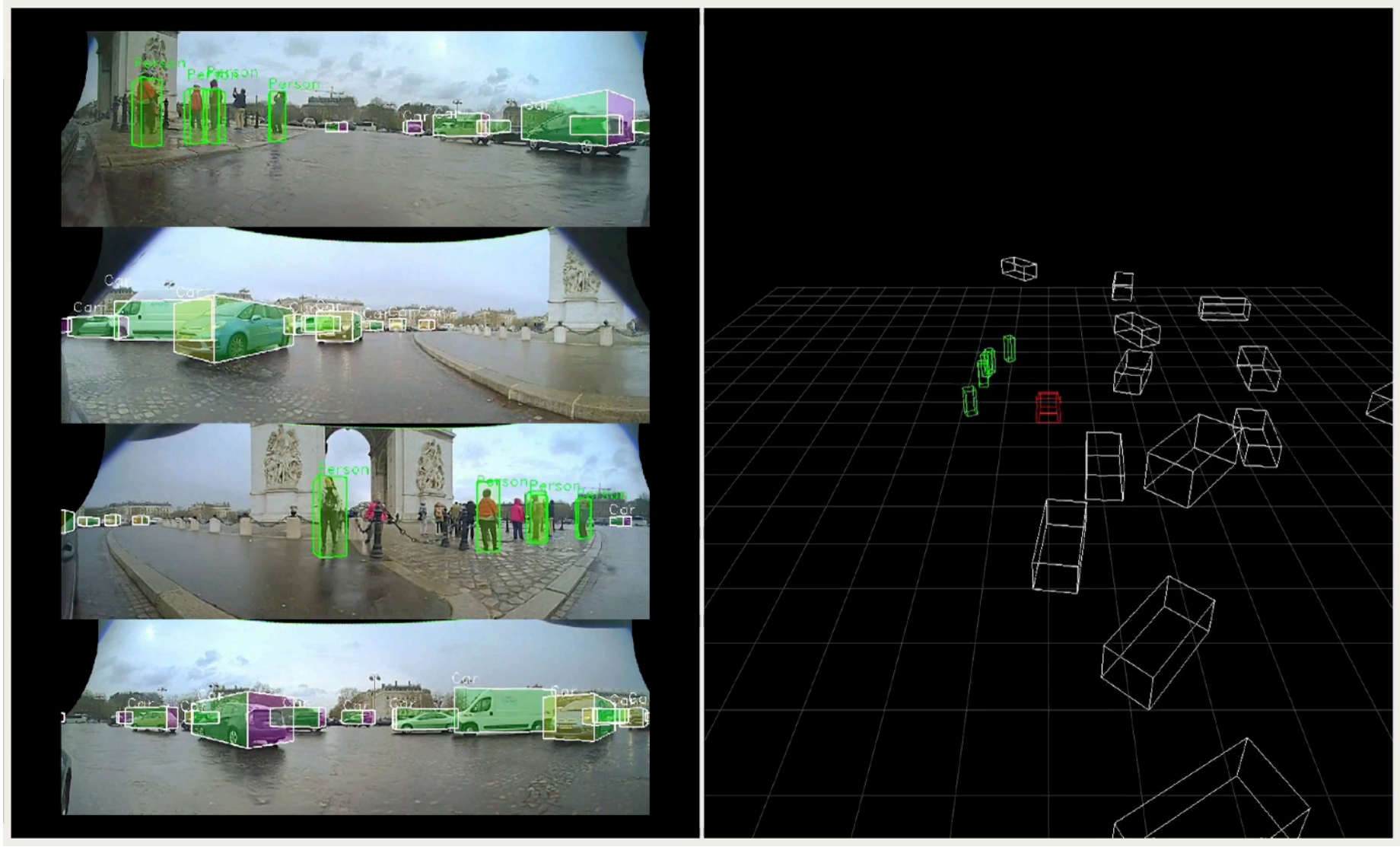
- ***Artificial Intelligence* is changing the man-machine interaction – natural interfaces, ”intelligent” behavior**
  - Image and situation understanding
  - Voice recognition and synthesis
  - ***Unstructured data pattern recognition, direct interfacing with the world***
  - *Creating the bridge between cyber and real world:* ***Enabling true Cyber Physical Systems***
  - …decision taking…

- ***Computer are not anymore a “PC”***
  - They get input from the real world with sensors, not anymore with keyboards
  - ***They are everywhere, morph in our environment***



DEEP MANTA

MANY-TASK DEEP NEURAL NETWORK FOR VISUAL OBJECT RECOGNITION



BUT COMPUTING SYSTEMS WERE NOT DESIGNED FOR CPS SYSTEMS

In nearly all hardware and software of computing systems:

- Time is abstracted or even not present at all
  - Very few programming languages can express time or timing constraints
- All is done to have the best average performance, not predictable performances
  - Caches, out of order execution, branch prediction, speculative execution,…
  - (Hidden) compiler optimization, call to (time) unspecified libraries
- Energy is also left out of scope
  - This can have impact on data movement, optimizations
- Interaction with external world are second priorities vs. computation
  - Done with interrupts (introduced as an ***optimization***, eliminating unproductive waiting time in polling loops) which were design to be ***exceptional events***…
- Etc.

We need new computing paradigms more suited for Cyber AND Physical Systems

Cloud and data centers are not the answer to everything…

Embedded intelligence needs local high-end computing

System should be autonomous to make good decisions in all conditions and ***in time***

And **should not consume** most **power** of an electric car!

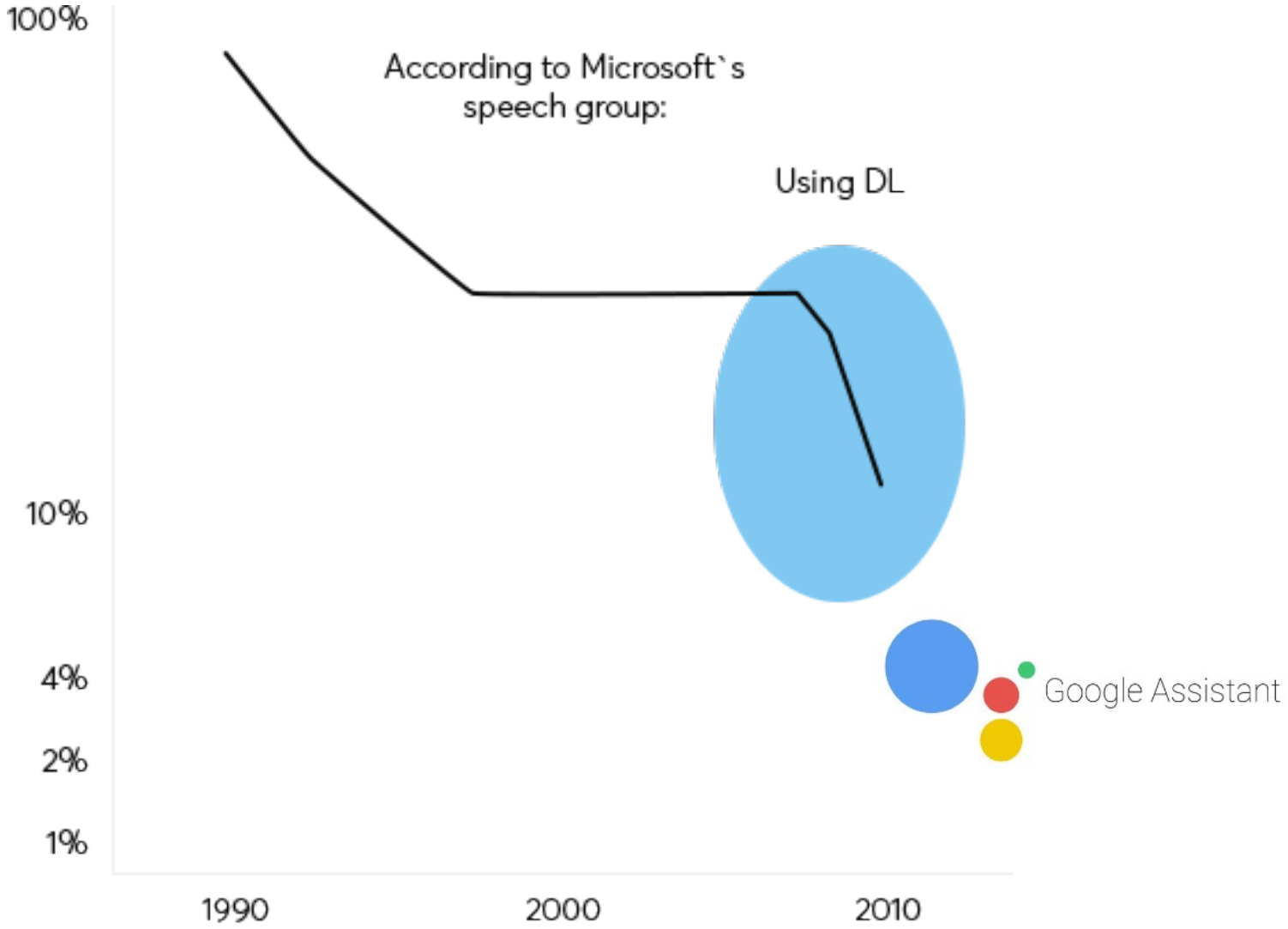
ONE ASPECT OF AI: PERSONAL ASSISTANTS....

Google Assistant

Apple Siri

Amazon Alexa
with Zigbee

DEEP LEARNING AND VOICE RECOGNITION

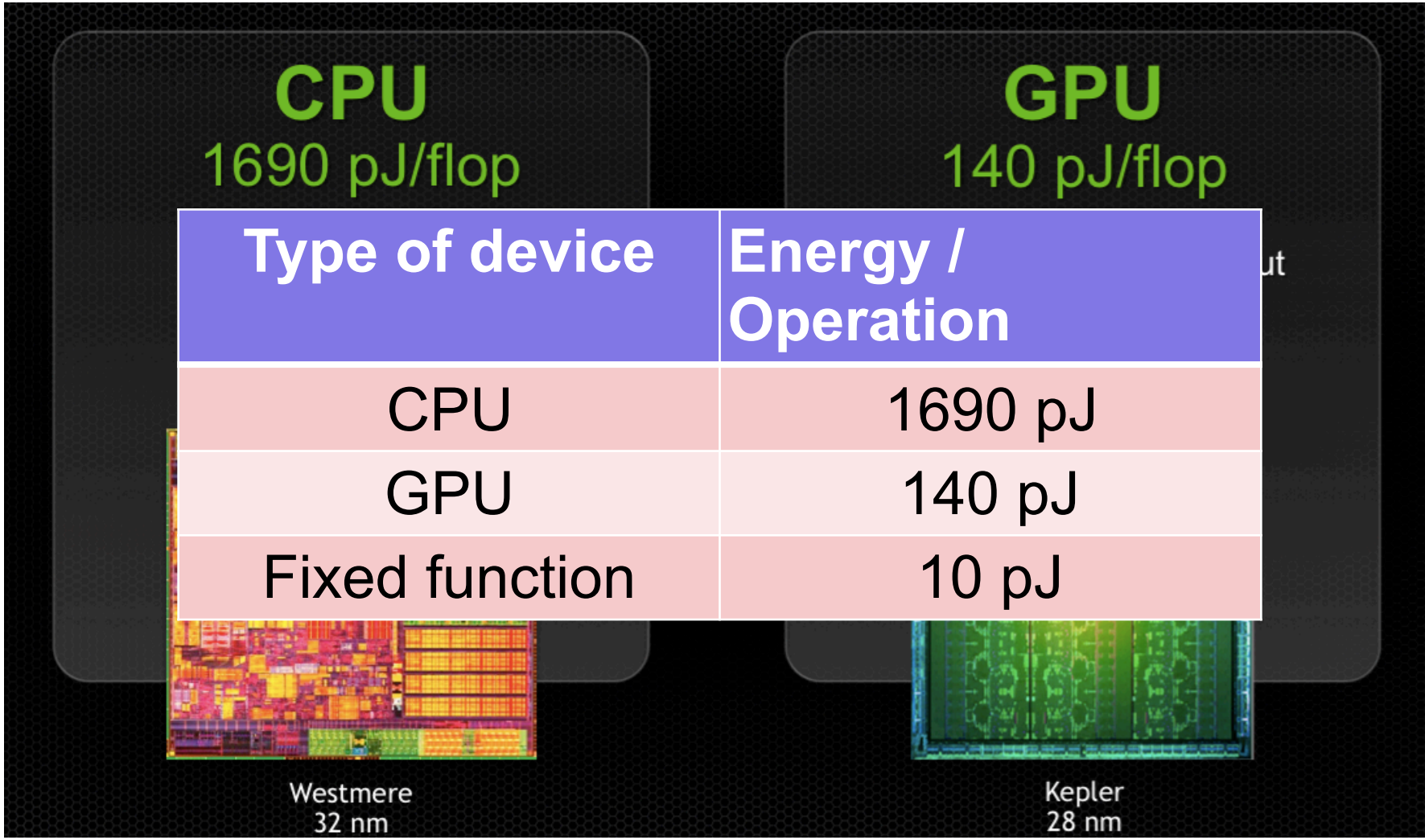


Deep Learning in **Speech Recognition**

DEEP LEARNING AND VOICE RECOGNITION

" The need for TPUs really emerged about six years ago, when we started using computationally expensive deep learning models in more and more places throughout our products. The computational expense of using these models had us worried. If we considered a scenario where people **use Google voice** search for just **three minutes a day** and we ran deep neural nets for our speech recognition system on the processing units we were using, **we would have had to double the number of Google data centers**!"

[https://cloudplatform.googleblog.com/2017/04/quantifying-the-performance-of-the-TPU-our-first-machine-learning-chip.html]



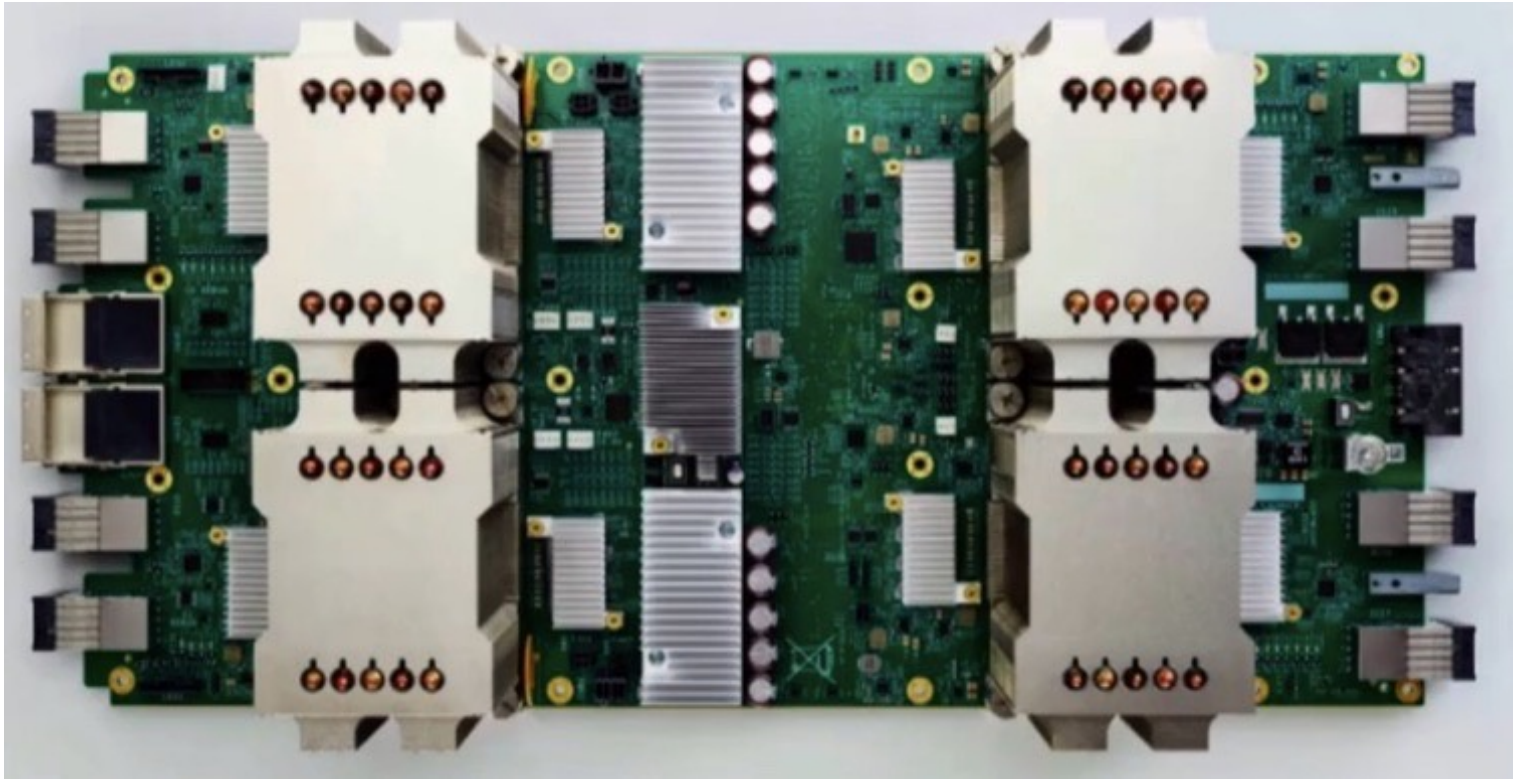
Type of device	Energy / Operation
CPU	1690 pJ
GPU	140 pJ
Fixed function	10 pJ

Source from Bill Dally (nVidia) « Challenges for Future Computing Systems » HiPEAC conference 2015

26

2017: GOOGLE'S CUSTOMIZED HARDWARE…

… required to increase energy efficiency
with **accuracy adapted to the use (e.g. float 16)**



Google's TPU2 : training and inference in a **180 teraflops$_{16}$** board
(over 200W per TPU2 chip according to the size of the heat sink)

2017: GOOGLE'S CUSTOMIZED TPU HARDWARE…

… required to increase energy efficiency
with accuracy adapted to the use (e.g. float 16)

Google's TPU2 : 11.5 $petaflops_{16}$ of machine learning number crunching
(and guessing about 400+ KW…, 100+ $GFlops_{16}/W$)

From Google

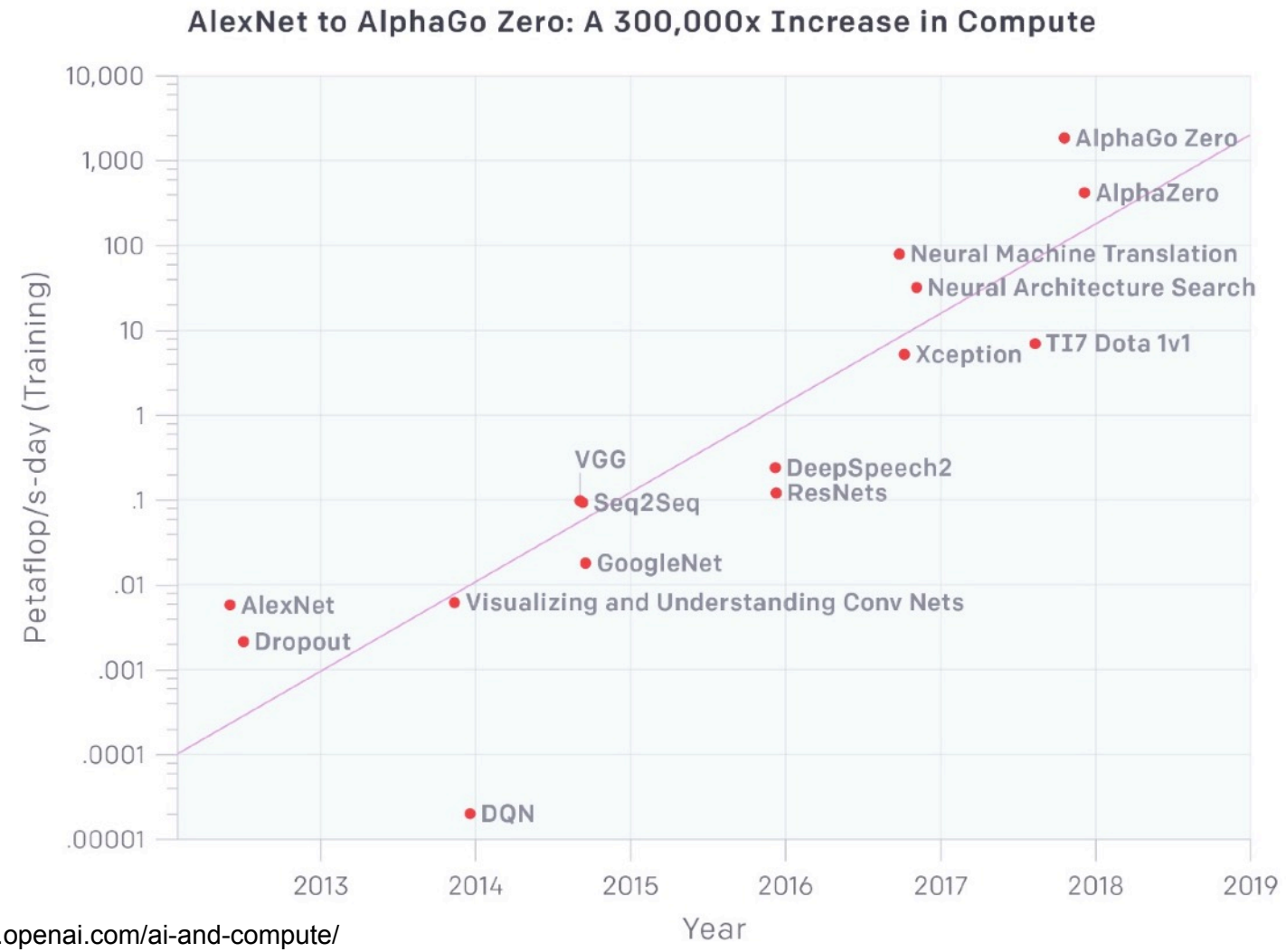
Peta = 10^{15} = million of milliard

ALPHAGO ZERO: SELF-PLAYING TO LEARN

From doi:10.1038/nature24270 (Received 07 April 2017)

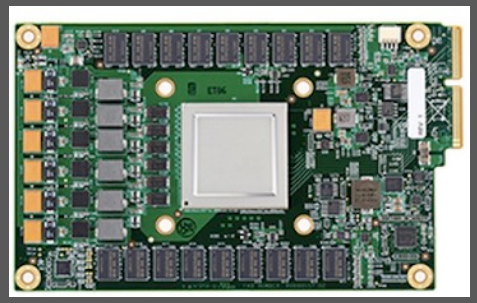
EXPONENTIAL INCREASE OF COMPUTING POWER FOR AI TRAINING

"Since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time
(by comparison, Moore's Law had an 18-month doubling period)*"



* https://blog.openai.com/ai-and-compute/

ALPHAZERO: COMPUTING RESOURCES



X 5000 = 200 KW*

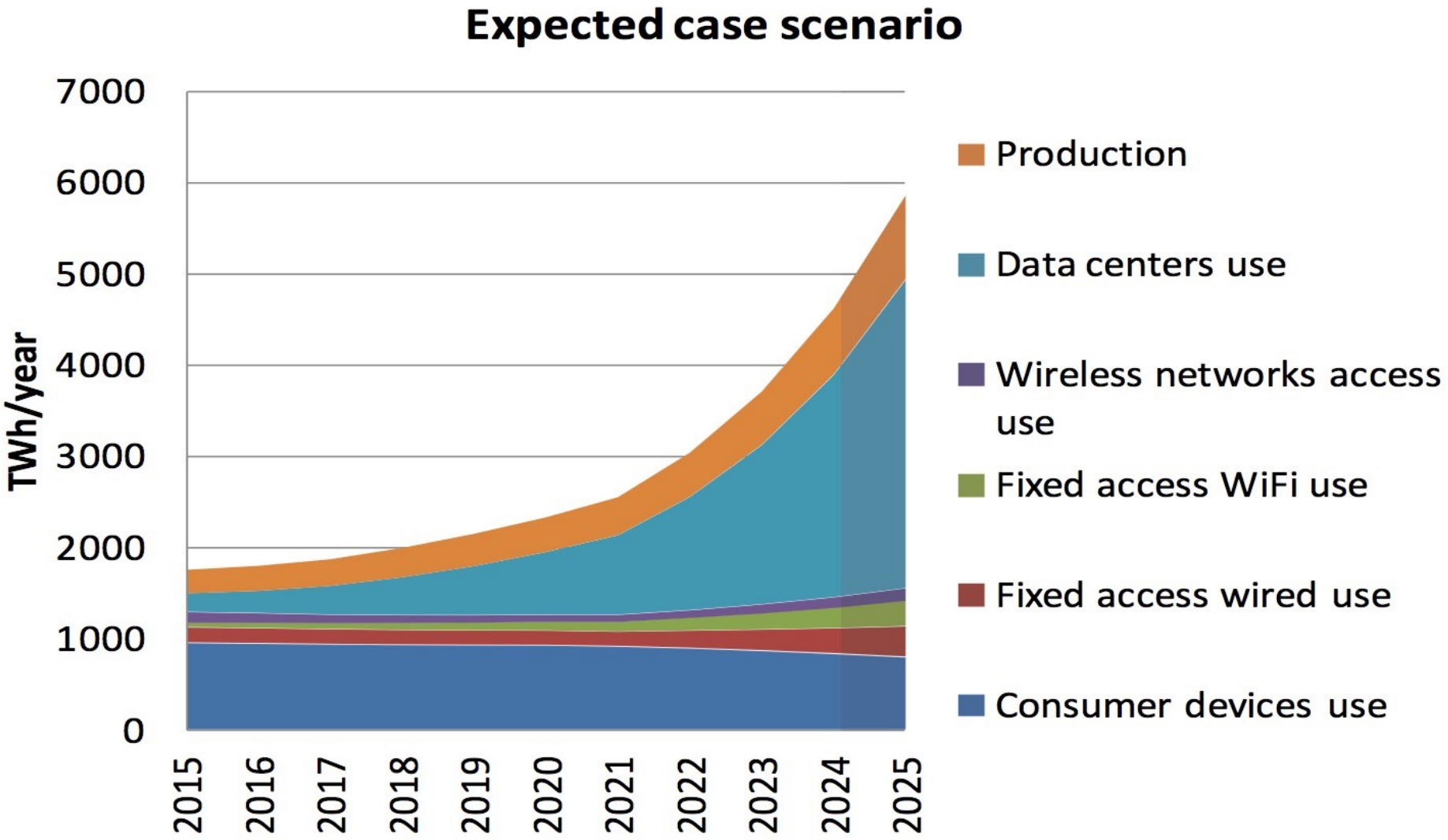
X 40 days…

Peta = 10^{15} = million of milliard

* https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

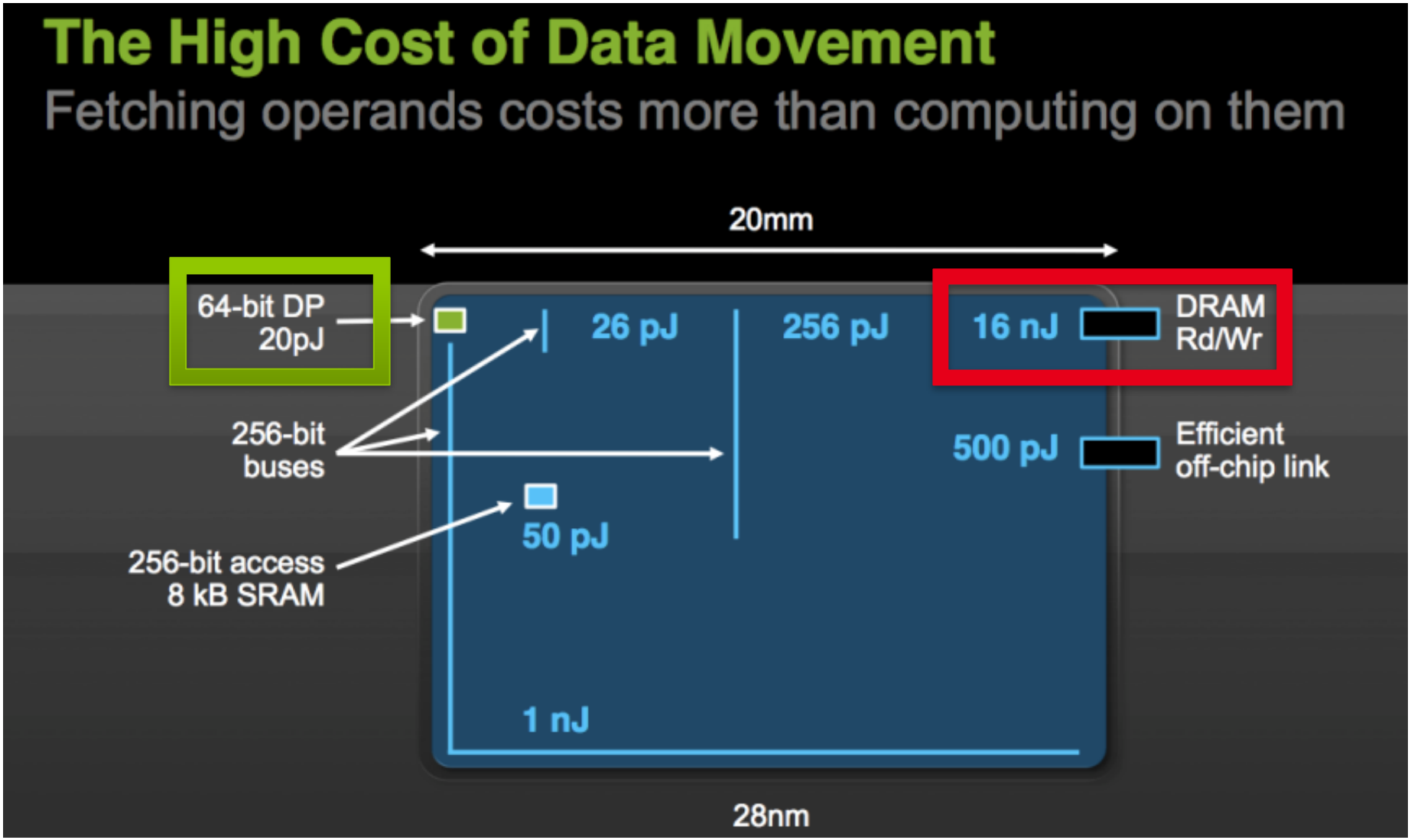

HOUSTON
WE HAVE A PROBLEM...

The problem:



From “Total Consumer Power Consumption Forecast”, Anders S.G. Andrae, October 2017

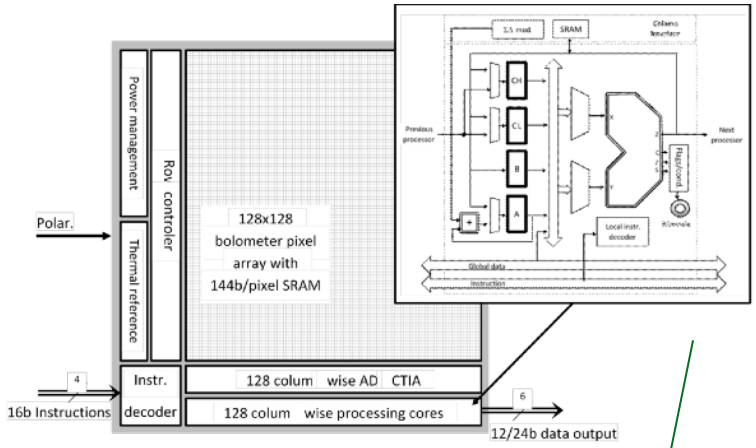
COST OF MOVING DATA



Source: Bill Dally, « To ExaScale and Beyond »
www.nvidia.com/content/PDF/sc_2010/theater/Dally_SC10.pdf

IR sensor with built-in 1st level image processing

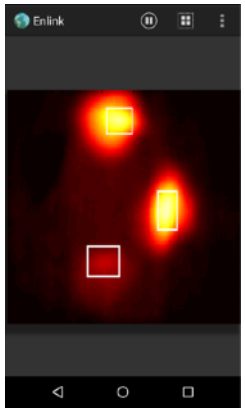
- Solution for presence sensing
- 128 processing DSP cores
- Read-out IC with image correction & processing capabilities
  - 128x128 pixel array for µ-bolometer detector
  - 128 column-wise ΔΣ ADCs + SIMD RISC processors
  - 12 frames of pixel-memory

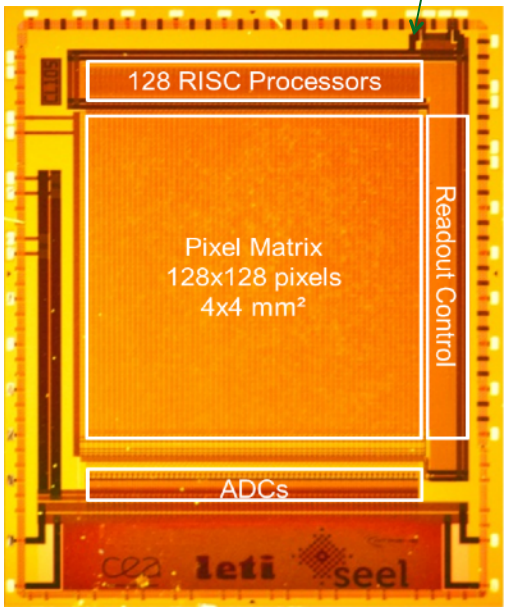


- ***Huge gain in power consumption*** compared to standard system for presence detection: ***power/ pixel reduced by over 90% compared to best in class low power IR product***
  (7,03 µW/pixel → 0,54 µW/pixel)

- **Dedicated algorithms for**
  - Presence sensing
  - Localization
  - People counting

- **Privacy respected**
  - Pre-processing on the sensor
  - Video is not transmitted

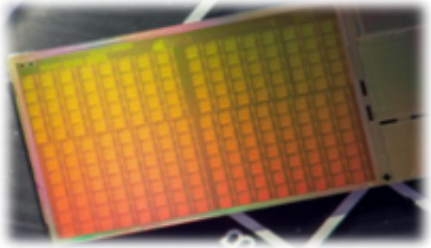




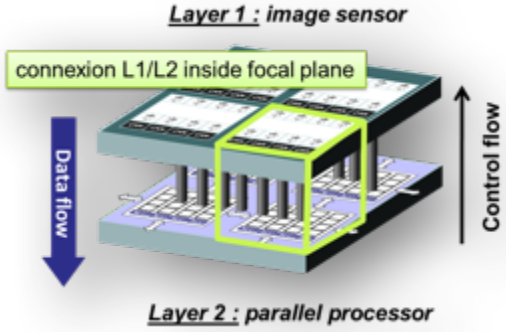
NEW COMPUTING PARADIGMS AND DEEP LEARNING

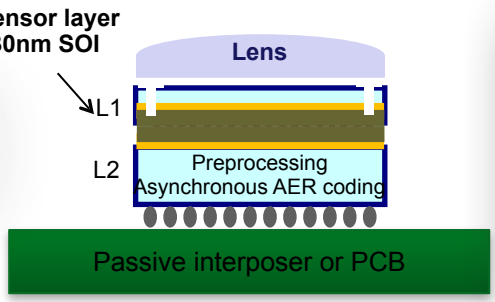
3D stacked retina with Spiking Neural Networks

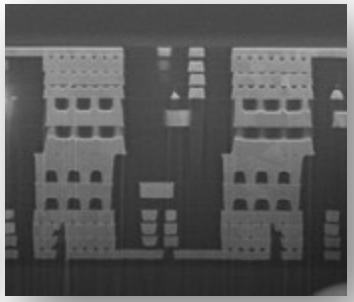
Goal: to meet the performances and flexibility of the human eye for image analysis, inspection of defaults, detection of problems, … Making an "*intelligent retina*"



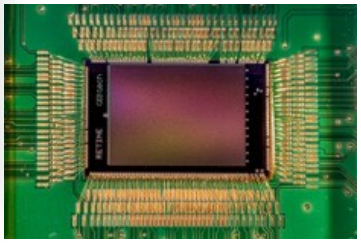
Processor array die

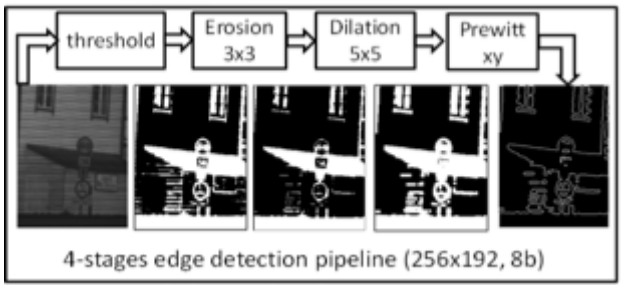




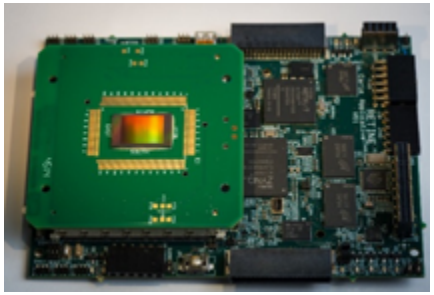


Retine Chip
ALTIS 130nm, CuCu bonding





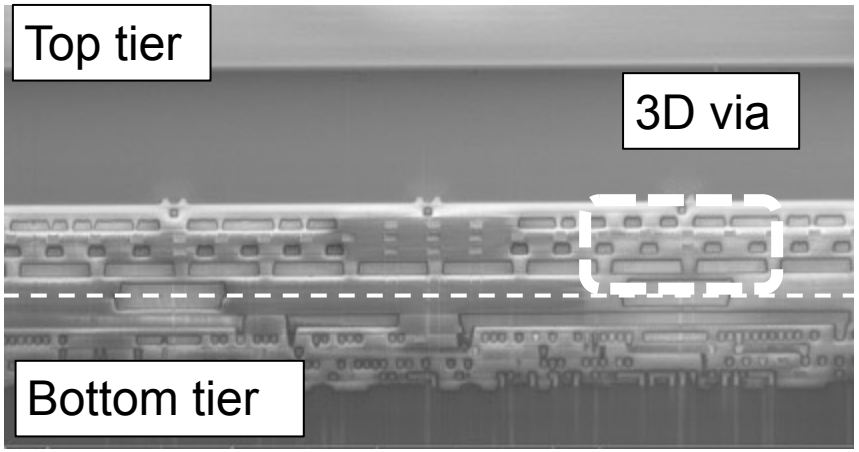
Circuit demonstrator "Retine"
L1@130 nm / L2@130 nm
IC size : 160 mm²
- Sensor : 192x256 @ 5500 fps / 768x1024 @ 60 fps
- Processing : 72 GOPS (192 SIMD processors)



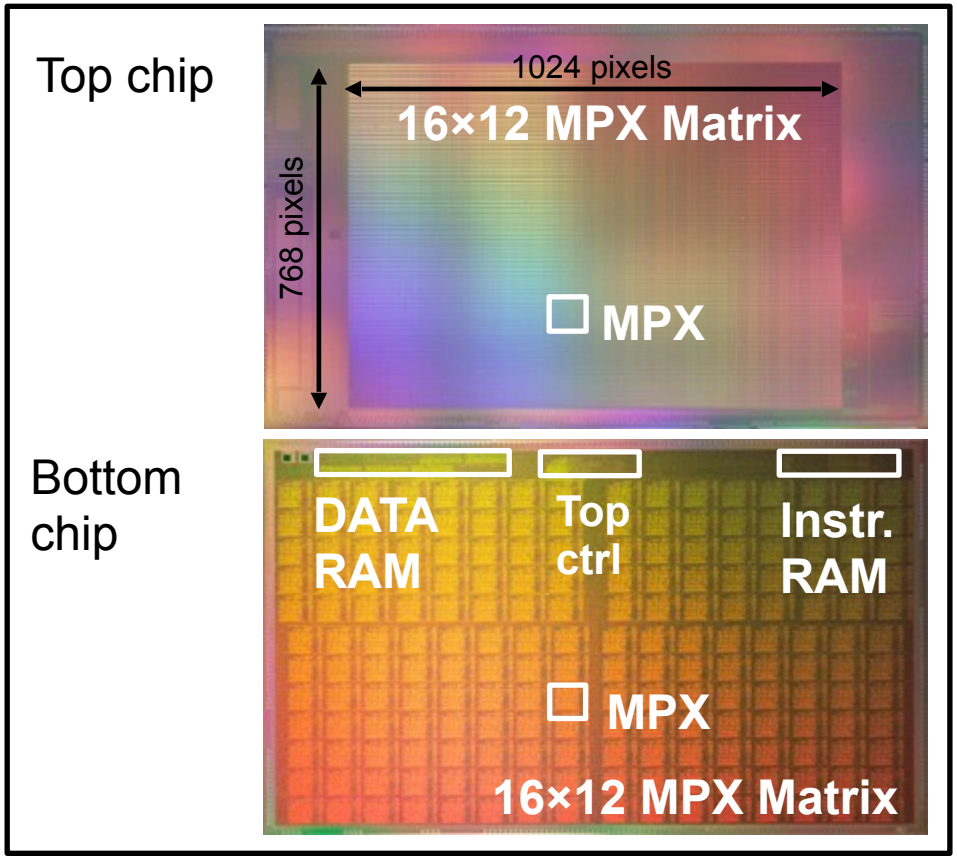
➔ x100 computing power, x10 energy efficiency, /15 processing latency

3D stacked retina with Spiking Neural Networks

113mm²
130nm/130nm
1024x768 (192x256 binned)
3D vias (signals): 6500



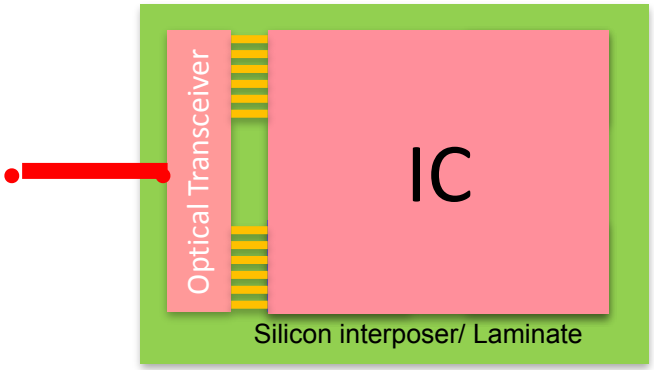
Direct bonding cross section

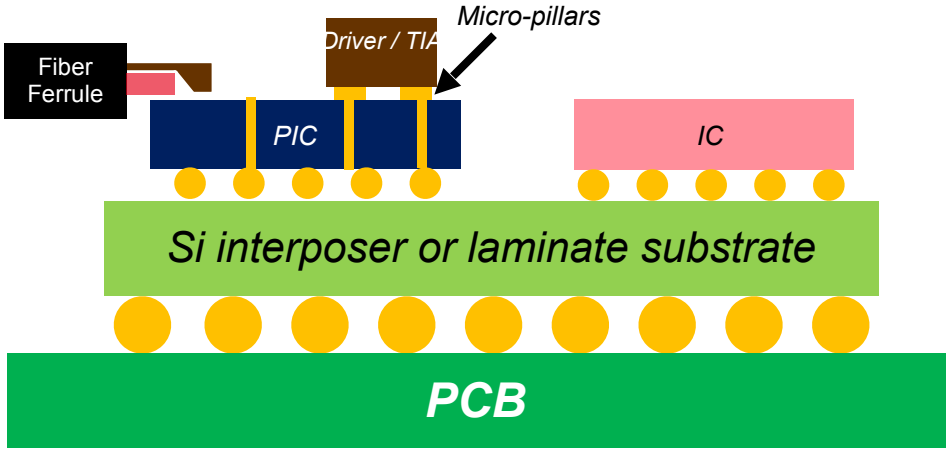


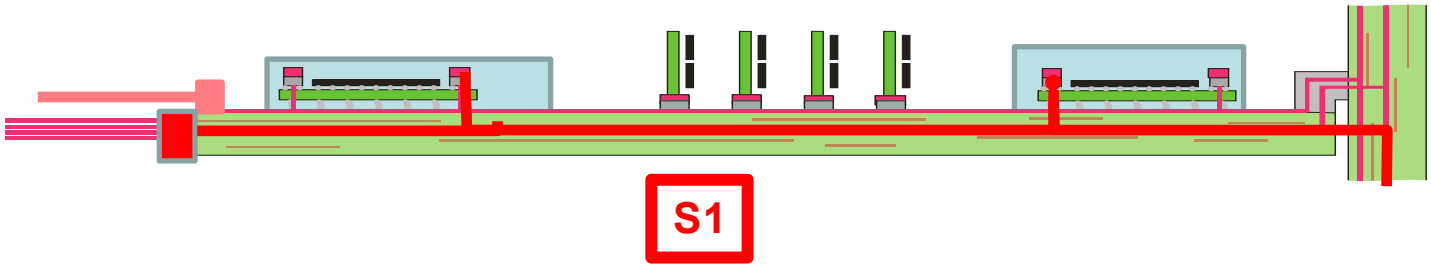
→ **x100 computing power, x10 energy efficiency, /15 processing latency**

REDUCING COMMUNICATIONS: OFF-CHIP PHOTONICS

Photonics: cost in sending information, nearly ***nothing in transmission***







Off board: AOC, optical modules

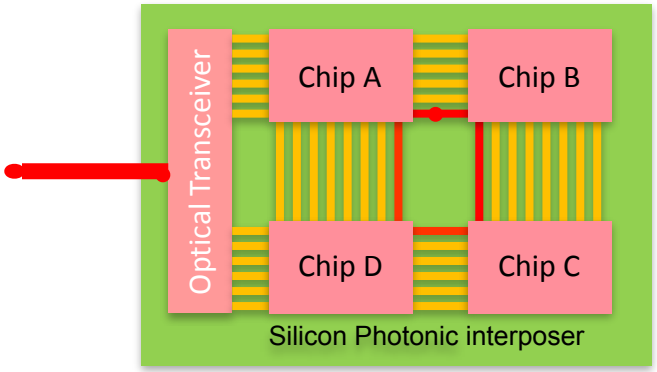
Off chip: Optical I/O

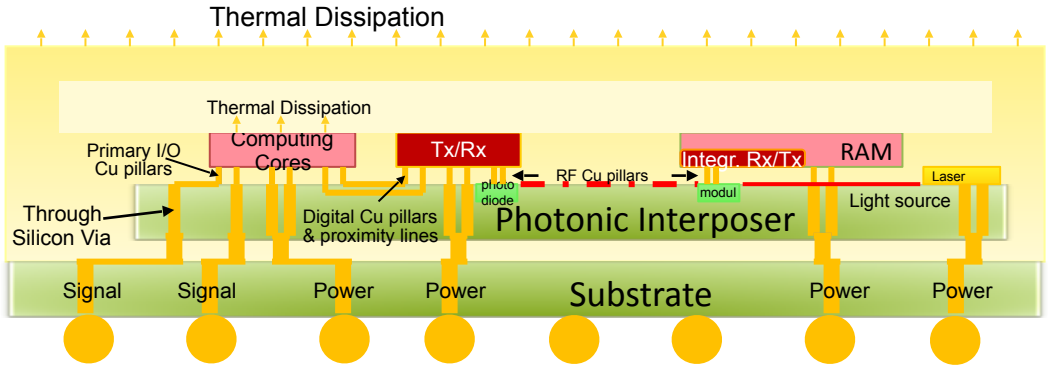
Time

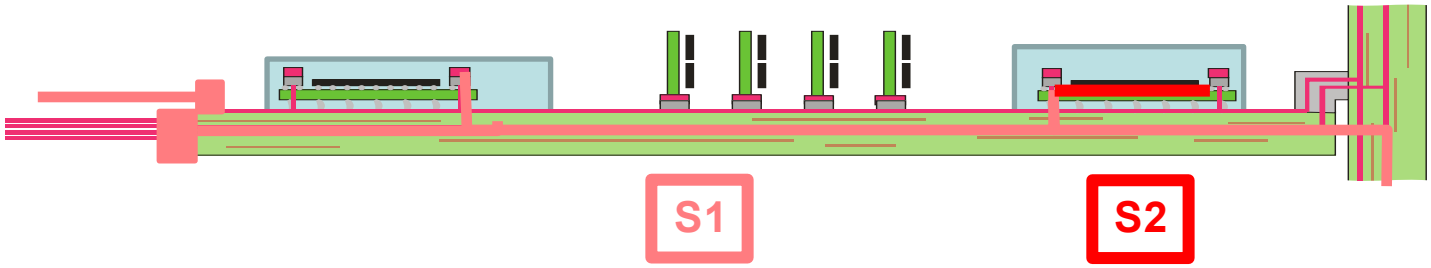
Now

Step 1

REDUCING COMMUNICATIONS: IN-PACKAGE PHOTONICS







Off board: AOC, optical modules

Off chip: Optical I/O

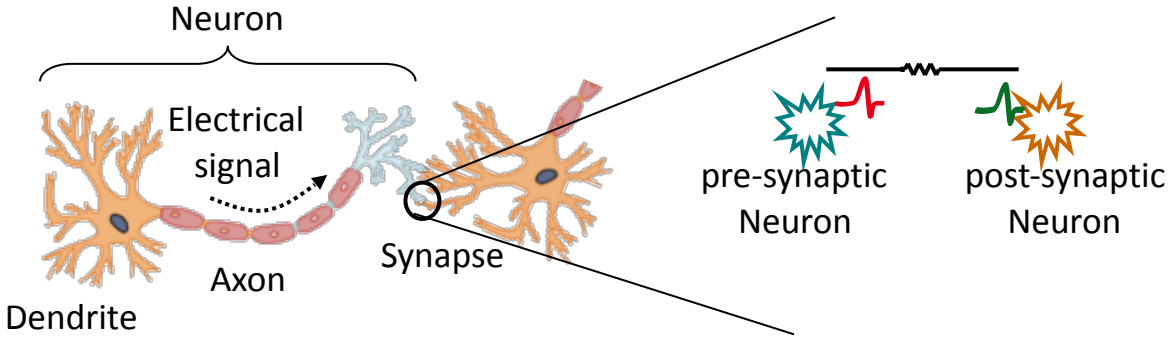
Optical network in package

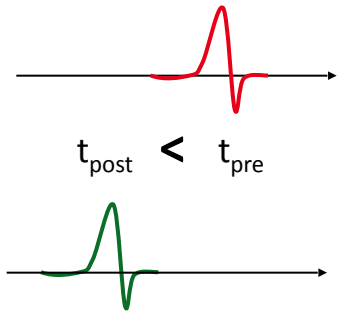
Now

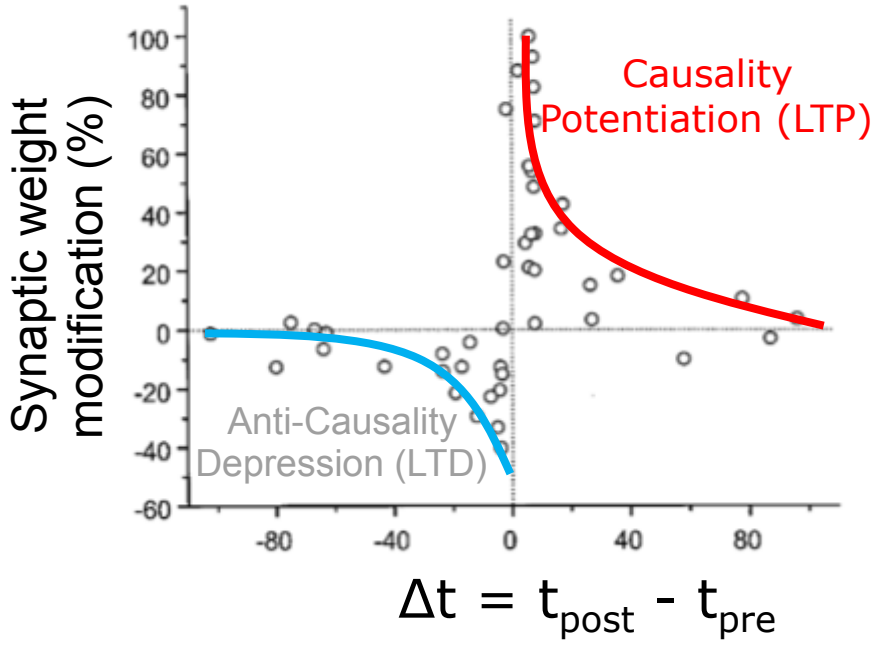
Step 1

Step 2

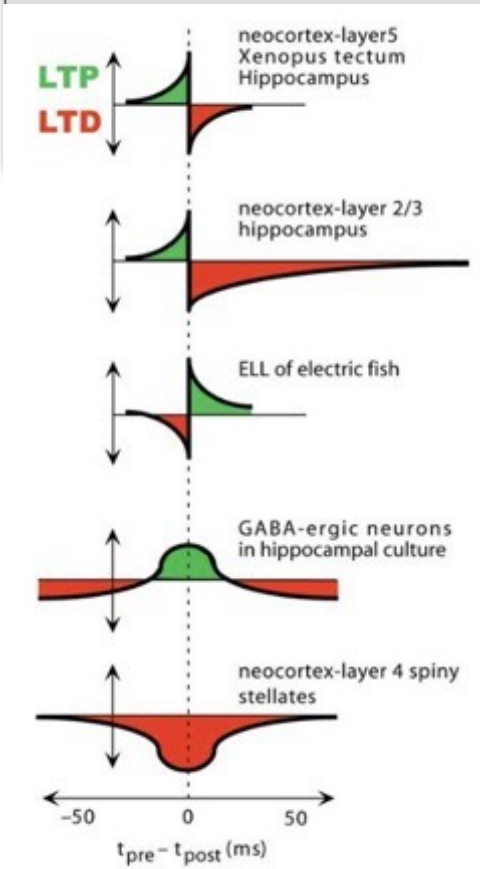
CODING INFORMATION DIFFERENTLY, ENABLING STDP (SPIKE TIMING DEPENDENT PLASTICITY)



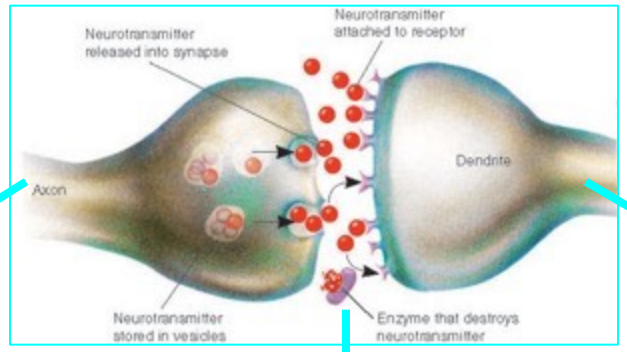




STDP = correlation detector



INVESTIGATION OF RRAM AS SYNAPSES
UNSUPERVISED LEARNING (INFORMATION CODED BY SPIKES)



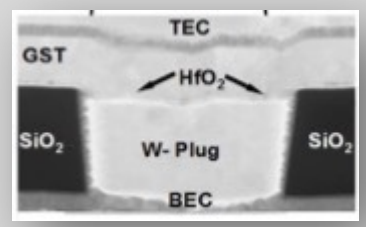
Thermal effect

Electrochemical effect

Electronic effect
oxygen vacancies

PCM

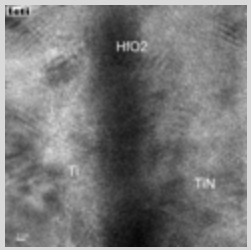
GST
GeTe
GST + HfO_2



M.Suri, et. al, IEDM 2011
M.Suri, et. al, IMW 2012 , JAP 2012
O.Bichler et al. IEEE TED 2012
M.Suri et al., EPCOS 2013
D.Garbin et al., IEEE Nano 2013

OXRAM

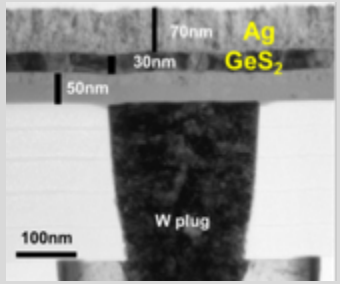
$TiN/HfO_2/Ti/TiN$



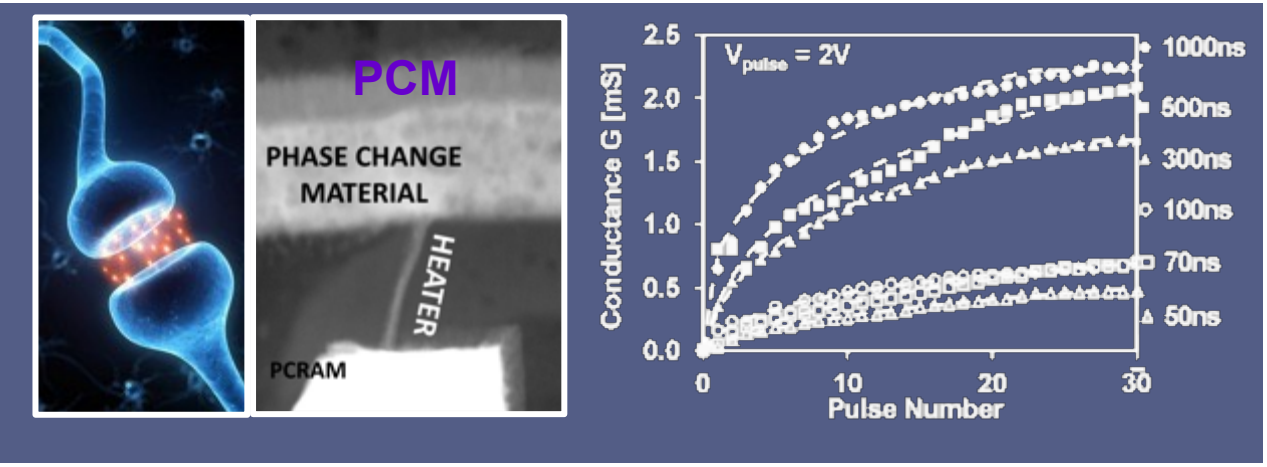
D.Garbin et al. IEDM 2014
D.Garbin et al., IEEE TED 2015

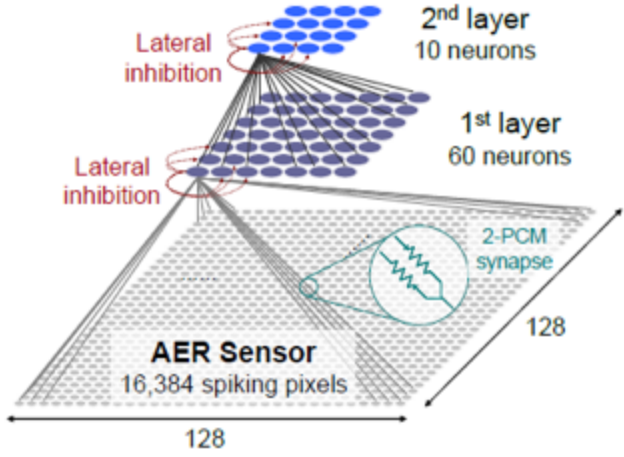
CBRAM

Ag / GeS_2



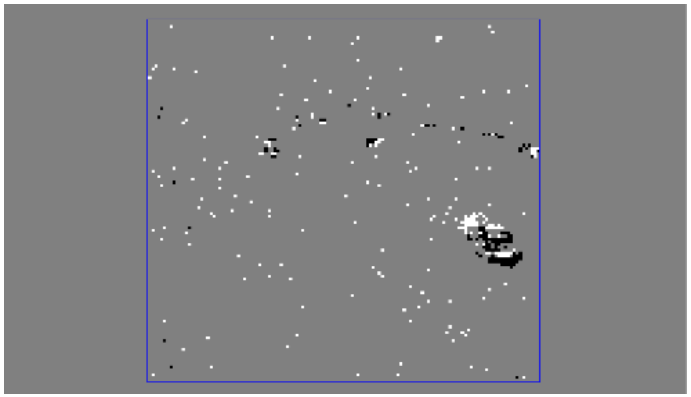
PCM-SYNAPSES



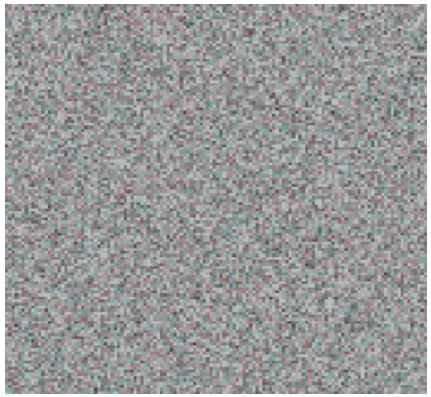


M. Suri, IEDM 2011

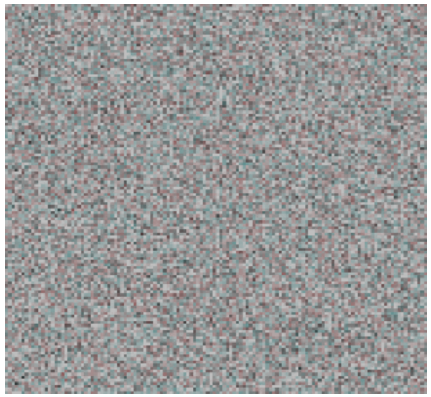
N2D2



Recorded Stimuli



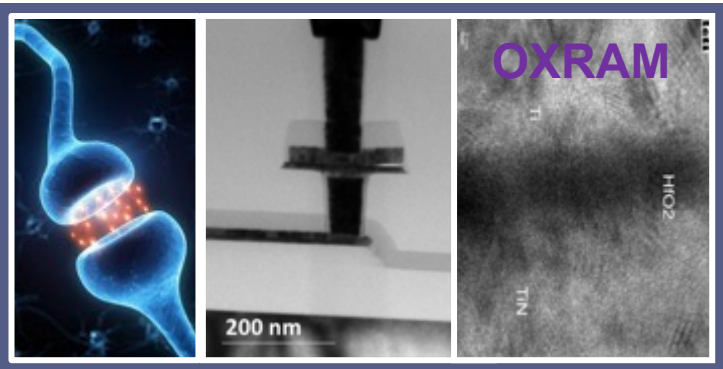
Neuron-4th lane



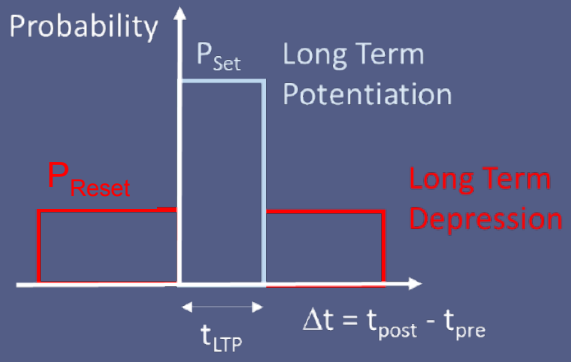
Neuron-5th lane

92% avg detection rate 112µW

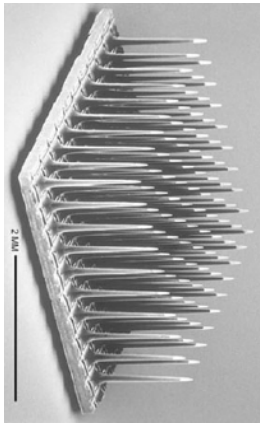
OXRAM-SYNAPSES

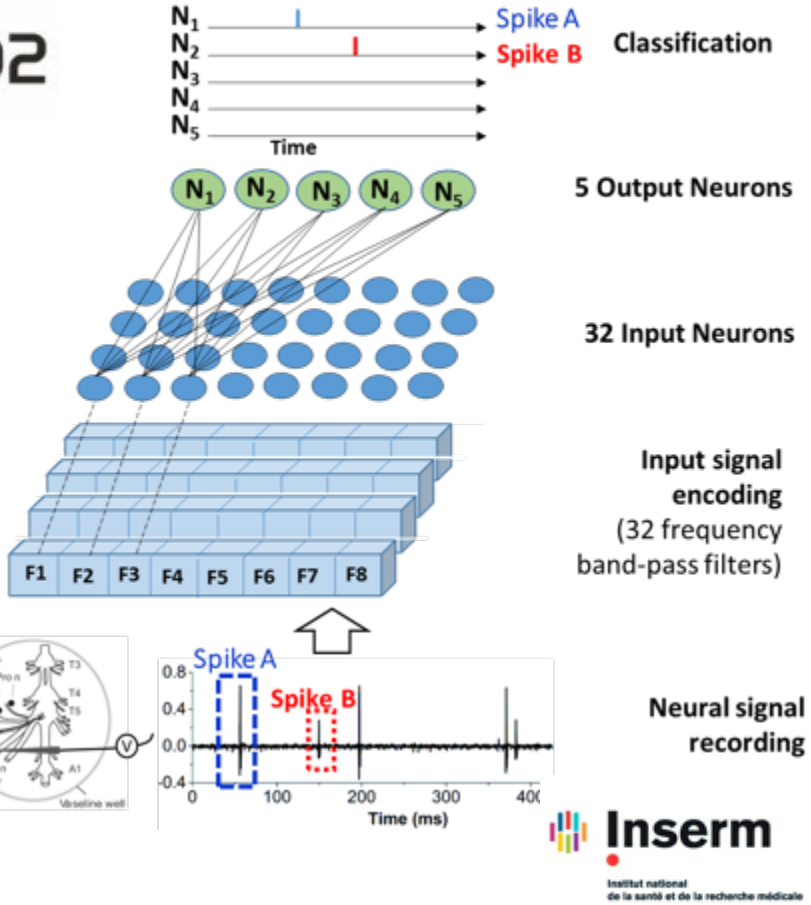


Spike-Timing-Dependent Plasticity



N2D2

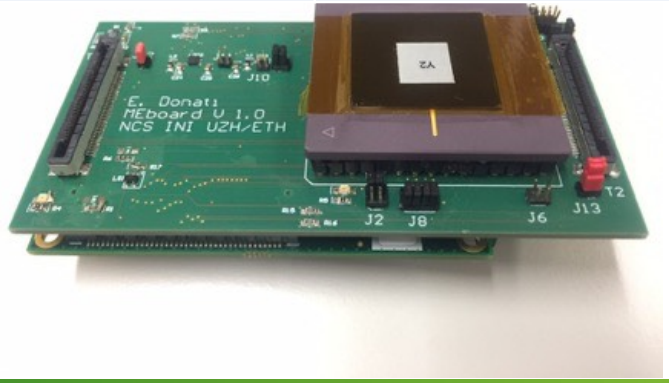




Inserm
Institut national de la santé et de la recherche médicale

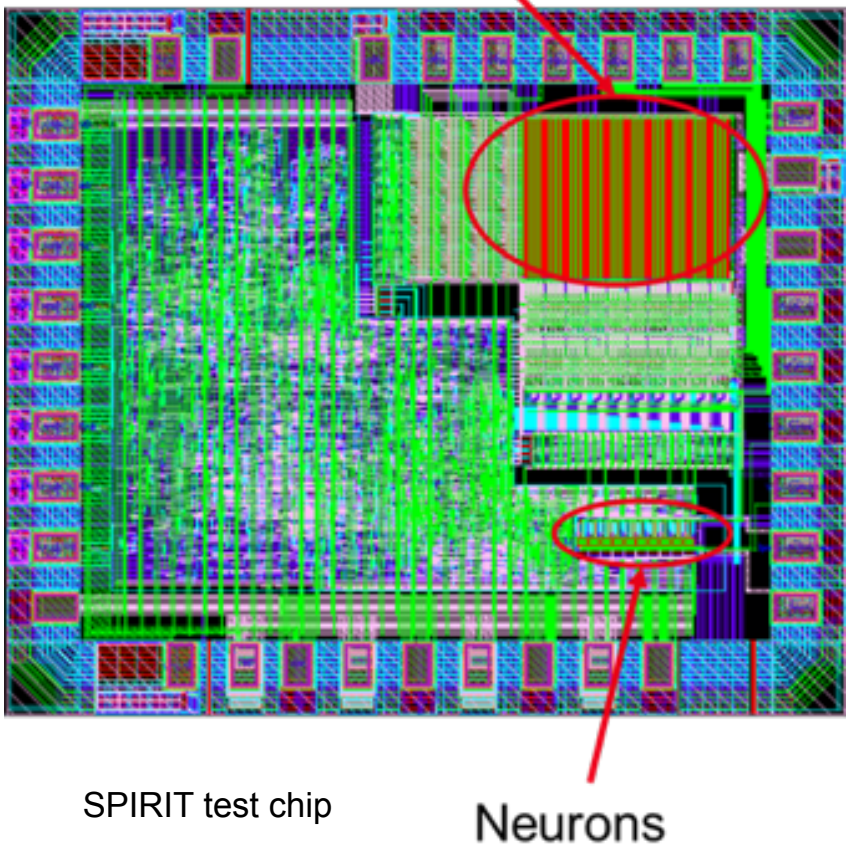
BENCHMARK

	IBM TrueNorth	Intel Loichi	DynapSEL
Technology	28nm CMOS	14 nm CMOS	28 nm FDSOI
Supply Voltage	0.7-1.05 V	0.5-1.25 V	0.73-1 V
Design Type	Digital	Digital	Mixed-signal
Neurons per core	256	Max 1k	256
Core Area	0.094 mm^2	0.4 mm^2	0.36 mm^2
Computation	Time multiplexing	Time multiplexing	**Parallel processing**
Fan In/Out	256/256	16/4k	**2k/8k**
On-line Learning	No	Programmable	**STDP**
Synaptic Operation / Second / Watt	46 GSOPS/W		**300 GSOPS/W**
Energy per synaptic operation	26 pJ	23.6 pJ	**<2 pJ**



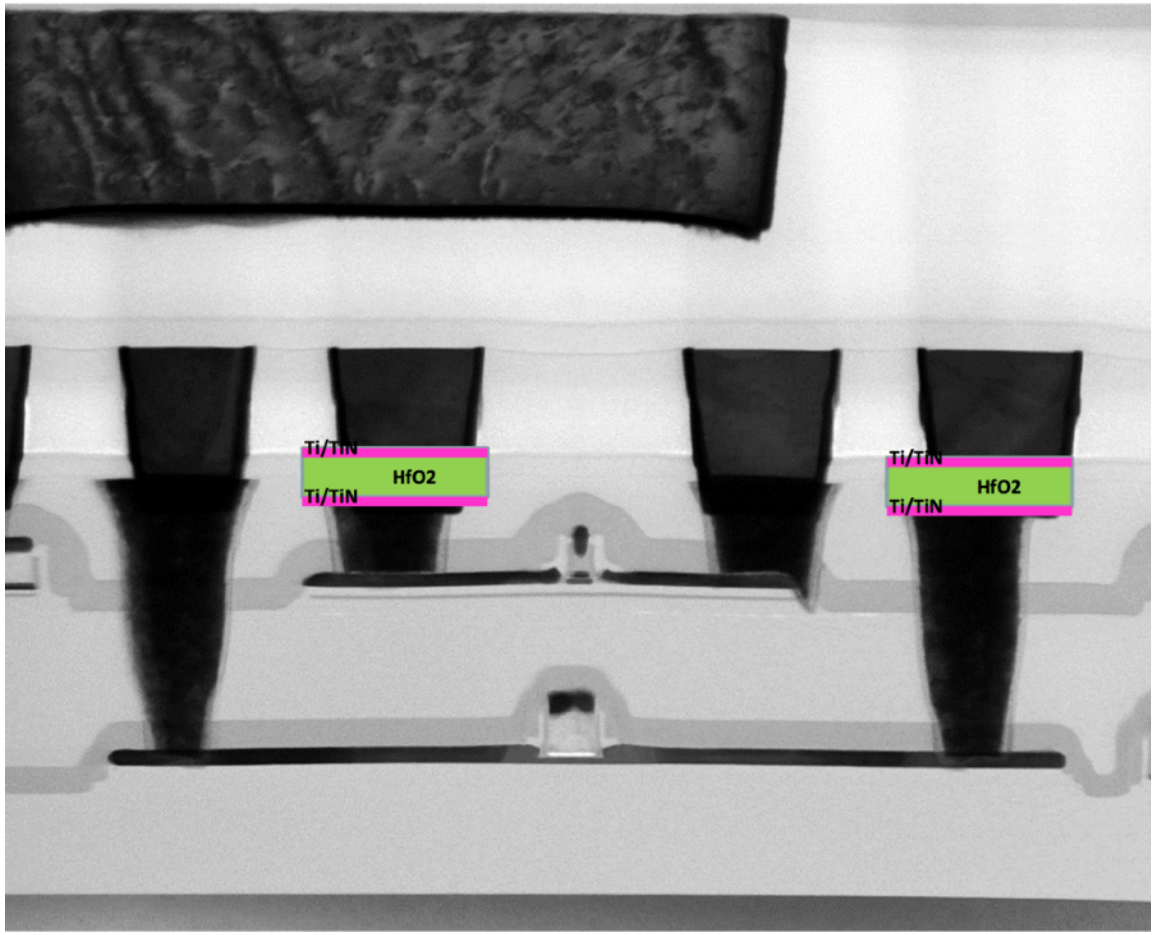
NEW COMPUTING PARADIGMS AND DEEP LEARNING
SPIKING NEURAL NETWORK WITH OXRAM

- **Test vehicle for spiking neural networks in 130nm CMOS with OxRAM elements between Metal 4 and Metal 5 of the back-end is done at CEA LETI.**
- **Area is 1,8mm². It contains 10 neurons and 1440 synapses, (11,5k OxRAMs)**
- **It can run MNIST (Characters recognition)**



SPIRIT test chip

3D INTEGRATION COUPLED WITH RRAM FOR SYNAPTIC WEIGHTS



Short term structure

→ RRAM on top level to avoid contamination issue

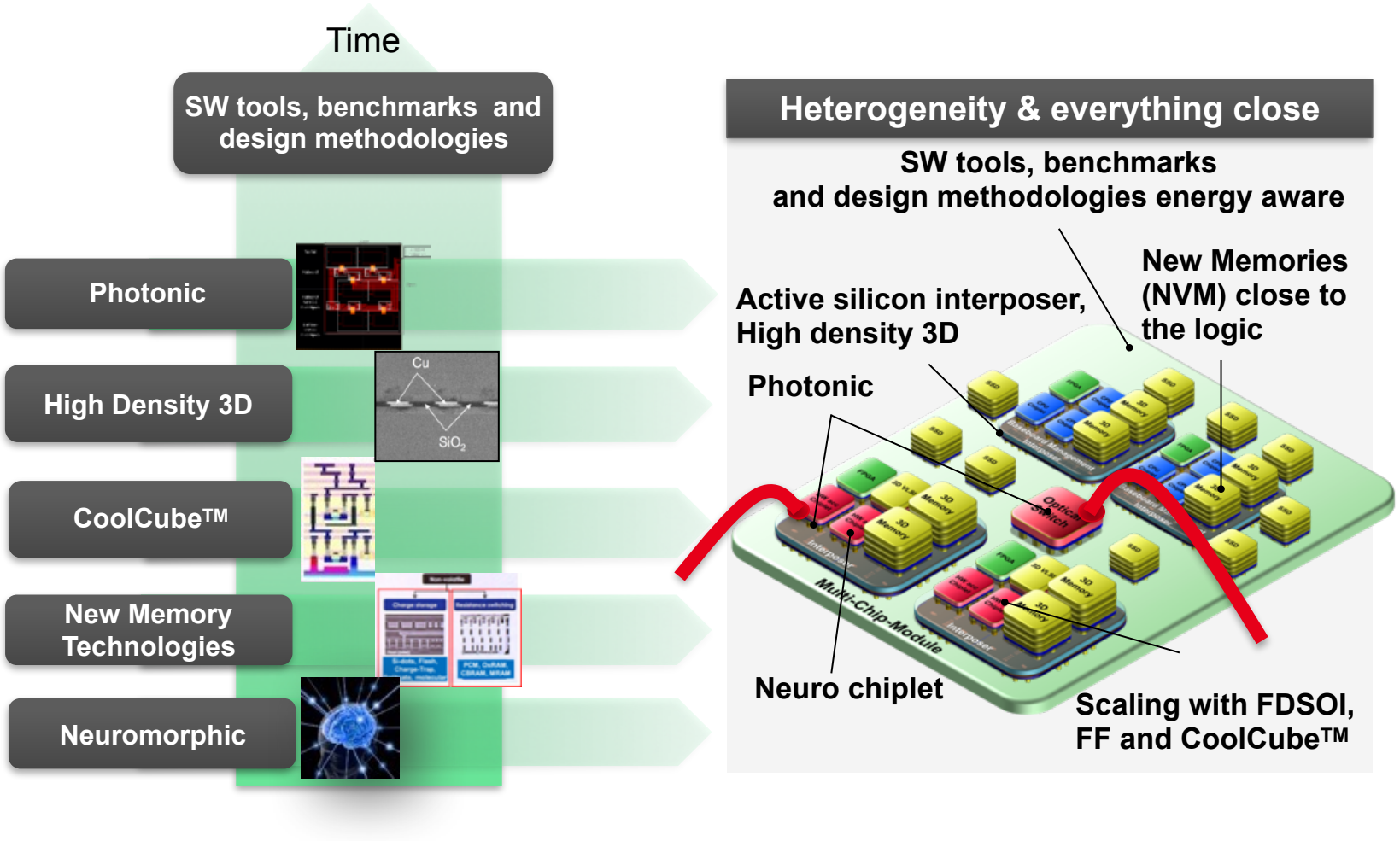
→ Reuse of existing masks plus ebeam to build 1T1R

→ **"Synapses" are integrated in the very fabric of communication**

1 base ebeam required for RRAM definition

RRAM based on HfO_2/Ti/TiN low temp materials (~ 350°C) → no critical problems to integrate on the top level

POTENTIAL SOLUTION FOR COGNITIVE CYBER PHYSICAL SYSTEMS

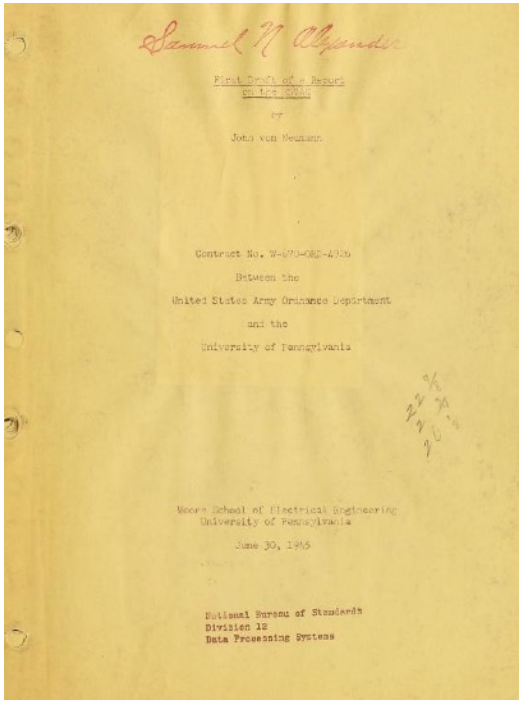


EXPLORE NEW WAYS AS ALTERNATIVE TO SILICON

- **New technologies**
  - Photonics for computing -> talk of Igor Carron
  - Neuromimetic -> talk of Patrick Pirim
  - Statistical –> talk of Pierre Bessière
  - Quantum computing -> all the talks of the afternoon
  - Sub-threshold
  - Printed/flexible electronics
  - Carbon nanotubes
  - Reservoir computing
  - Adiabatic computing
  - MEMS for computing
  - Synthetic biology, blob computing
  - Swarm computing
  - Symbiotic computing
  - Analogue/physic/hybrid computing
  - …

BACK TO THE ORIGIN:
WHAT IS THE TRUE VON NEUMANN ARCHITECTURE?

In “First Draft of a Report on the EDVAC,” the first published description of a stored- program binary computing machine - the modern computer, John von Neumann suggested modelling the computer after Pitts and McCulloch’s neural networks.


Samuel N Alexander

First Draft of a Report
on the EDVAC

by

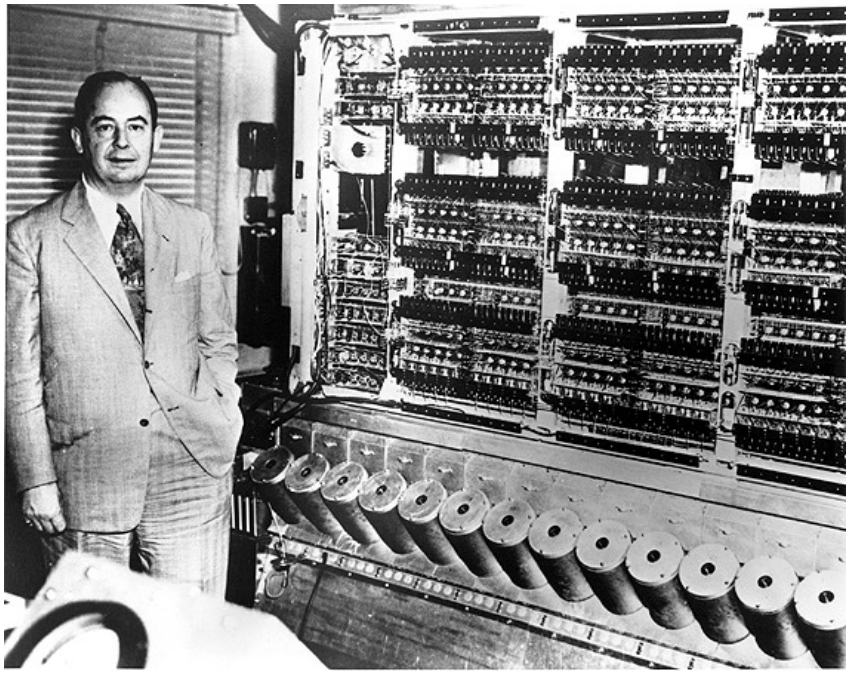
John von Neumann

Contract No. W-670-ORD-4926
Between the
United States Army Ordnance Department
and the
University of Pennsylvania

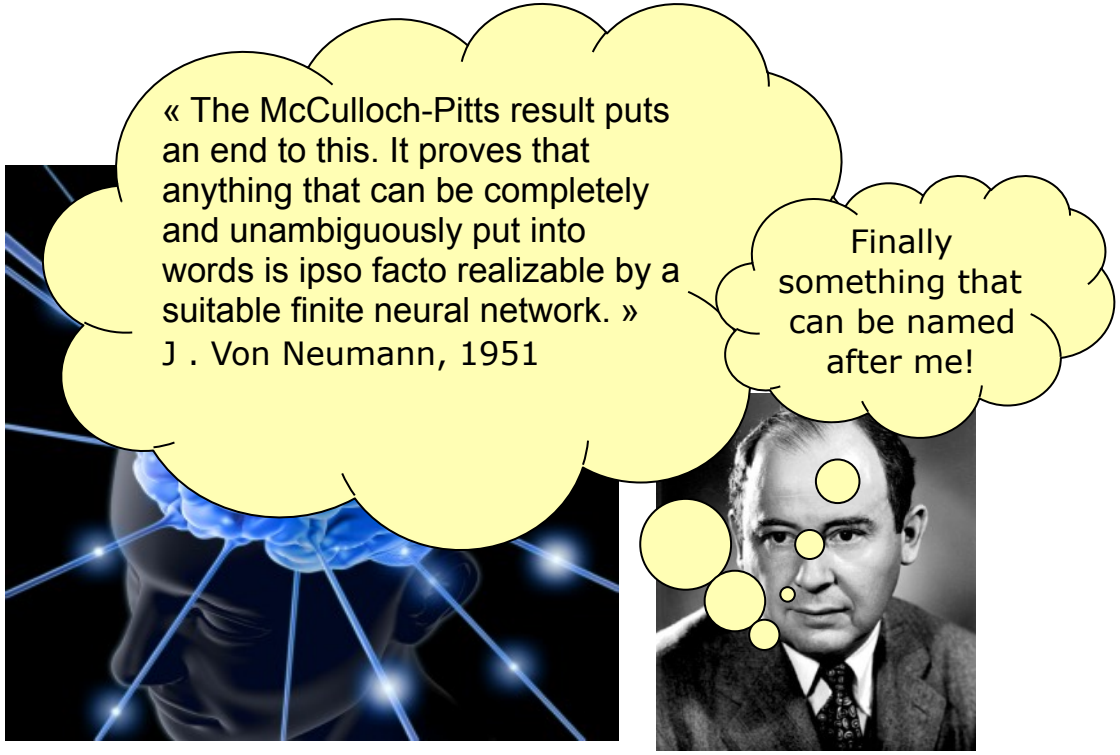
Moore School of Electrical Engineering
University of Pennsylvania

June 30, 1945

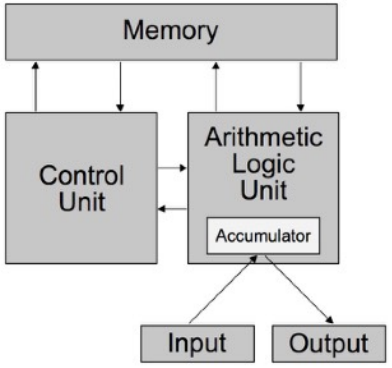
National Bureau of Standards
Division 12
Data Processing Systems



BACK TO THE ORIGIN:
WHAT IS THE TRUE VON NEUMANN ARCHITECTURE?



But technology was not ready in the 50's, leading to realization with sequential processing And to the computer architecture we have now…



CONCLUSION: WE LIVE AN EXCITING TIME!

Thank you for your attention

Special thank you to Denis Dutoit, Christian Gamrat, Carlo Reita for their slides I borrowed.

marc.duranton@cea.fr

leti
Centre de Grenoble
17 rue des Martyrs
38054 Grenoble Cedex

list
Centre de Saclay
Nano-Innov PC 172
91191 Gif sur Yvette Cedex