Upstream technologies

Marc Duranton, JF Lavignon

The EXDCI-2 project has received funding from the European Unions Horizon 2020 research and innovation programmed under grant agreement no. 800957.

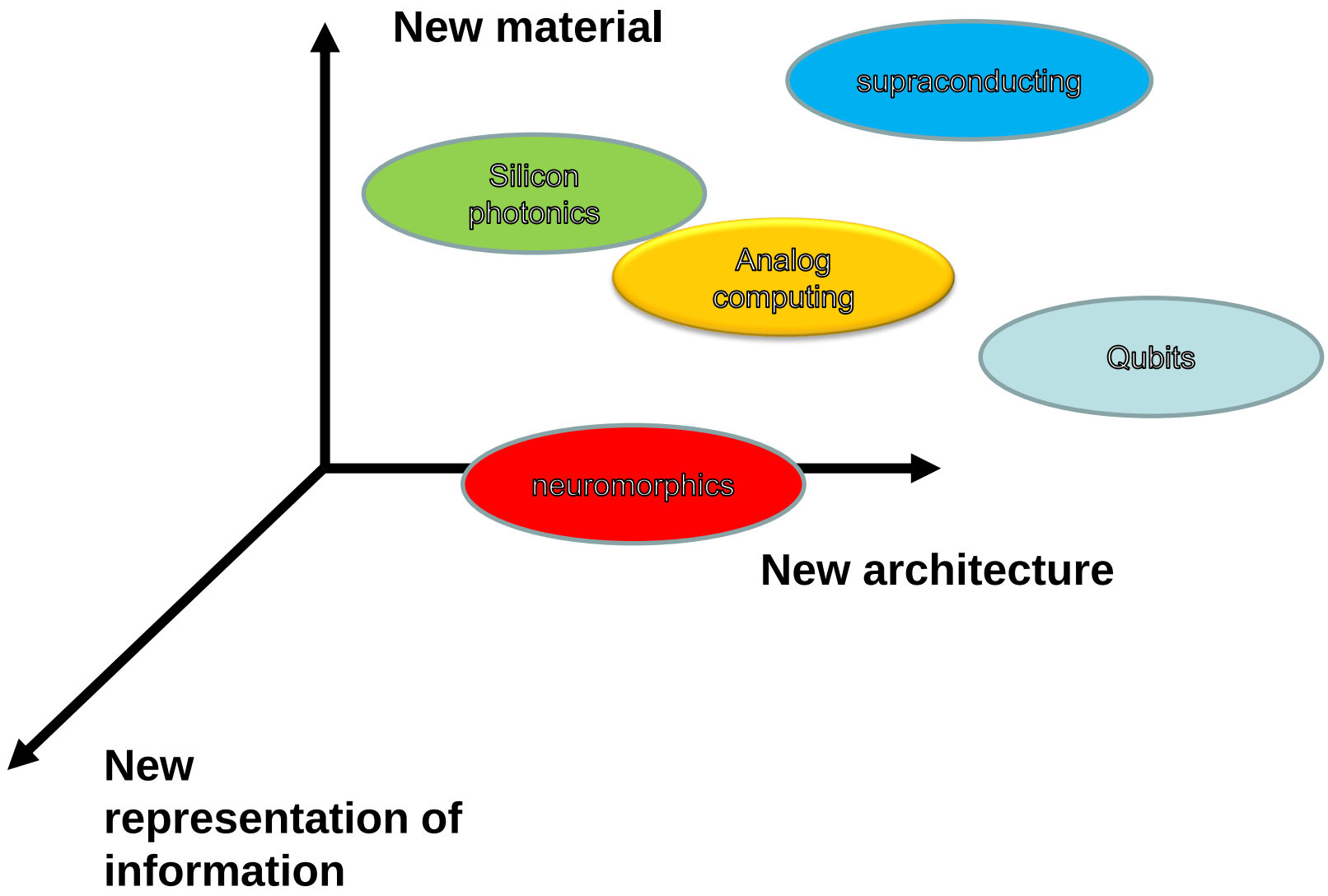
Introduction

- HPC walls
  - Frequency
  - Memory
  - Scaling
- Research ecosystem in Europe
  - Photonics : photonics21
  - Electronics: ECSEL JU AENEAS
  - 3 RTO : CEA/LETI, Fraunhofer, IMEC
- Discussion around workshops
  - Nov 2018
  - Nov 2019

Workshop with photonics and electronics experts

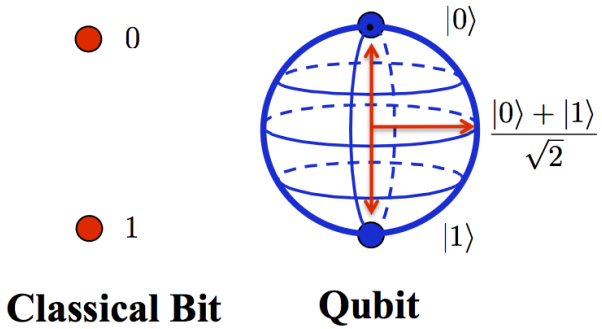
- Active discussions on
  - What are the most relevant technologies and/or new architectures for future HPC/edge systems;
  - How to accelerate the uptake of these technologies/architectures;
  - How Europe can develop a value chain for these new approaches and get a strong position.
- As results
  - Set a list of promising technologies/architectures relevant for future HPC/edge systems and meaningful to develop in Europe;
  - Some indications of what will be required for them to emerge;
  - Be in position to write short but credible "science fiction success story" for some of these high potential technologies.

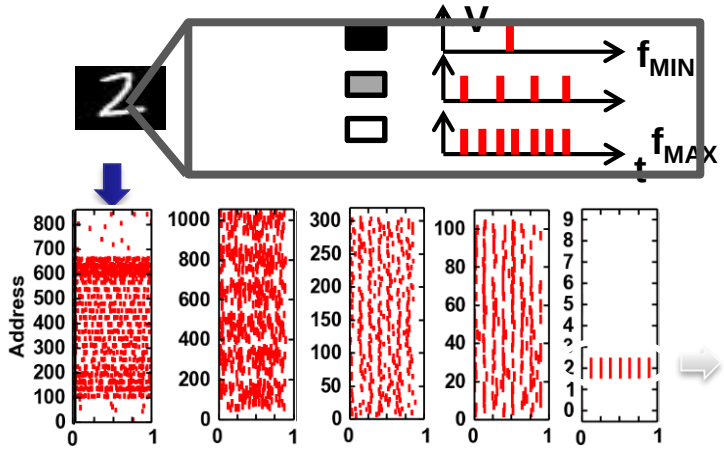
We need to reinvent

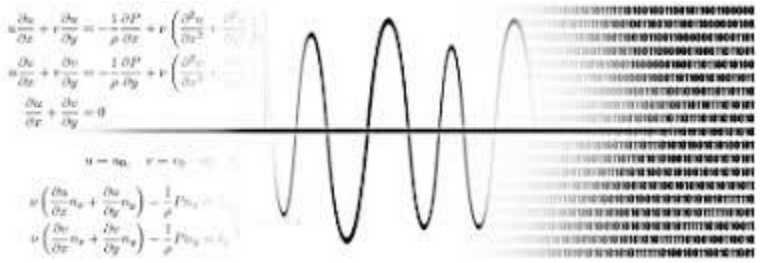


New way to represent information

- Trade-off precision/cost of compute ie precision 64b/32b/16b/8b
- Different data representation ie spikes
- Qbits
- Analog coding of information

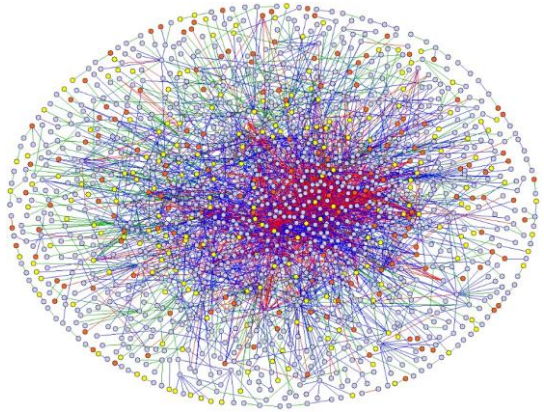


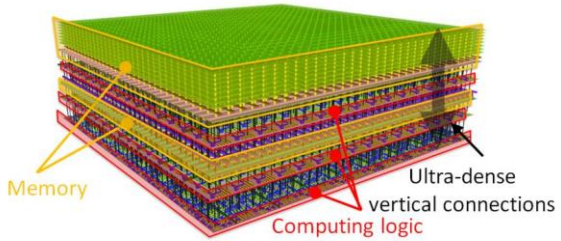




New architectures

- Processor in memory
- Data flow
- Neuromorphic
- Graph computing
- Simulated annealing
- Quantum annealing
- Quantum computing





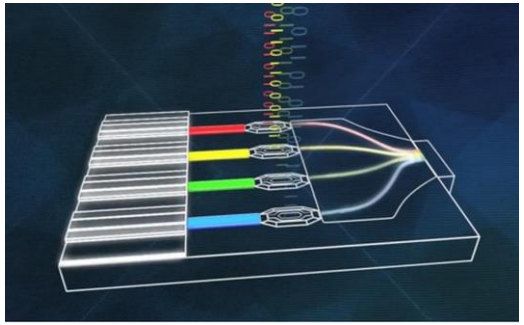
Tensor Processing Unit (TPU)

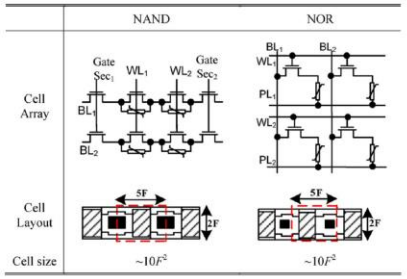
- **30-80x** TOPS/watt vs. 2015 CPUs and GPUs.
- 8 GiB DRAM.
- 8-bit fixed point.
- 256x256 MAC unit.
- Support for data reordering, matrix multiply, activation, pooling, and normalization.

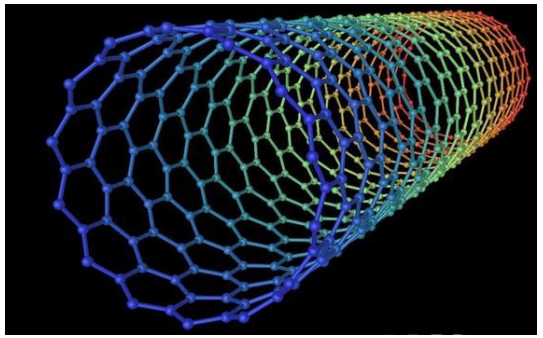
Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

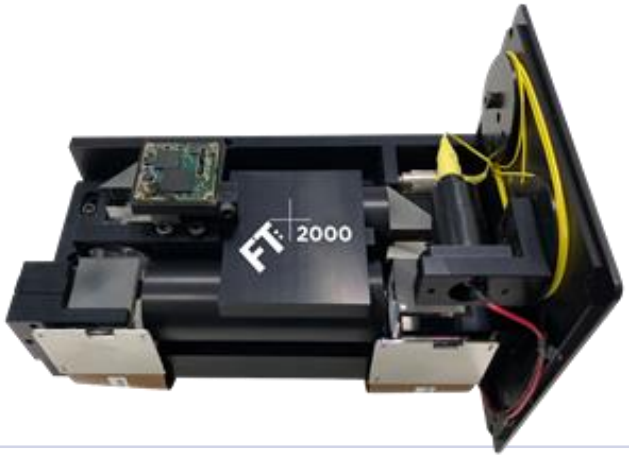
Hybrid CMOS – new material

- NVM
- Silicon photonics
- Memristive technologies
- New materials
- Analog computing





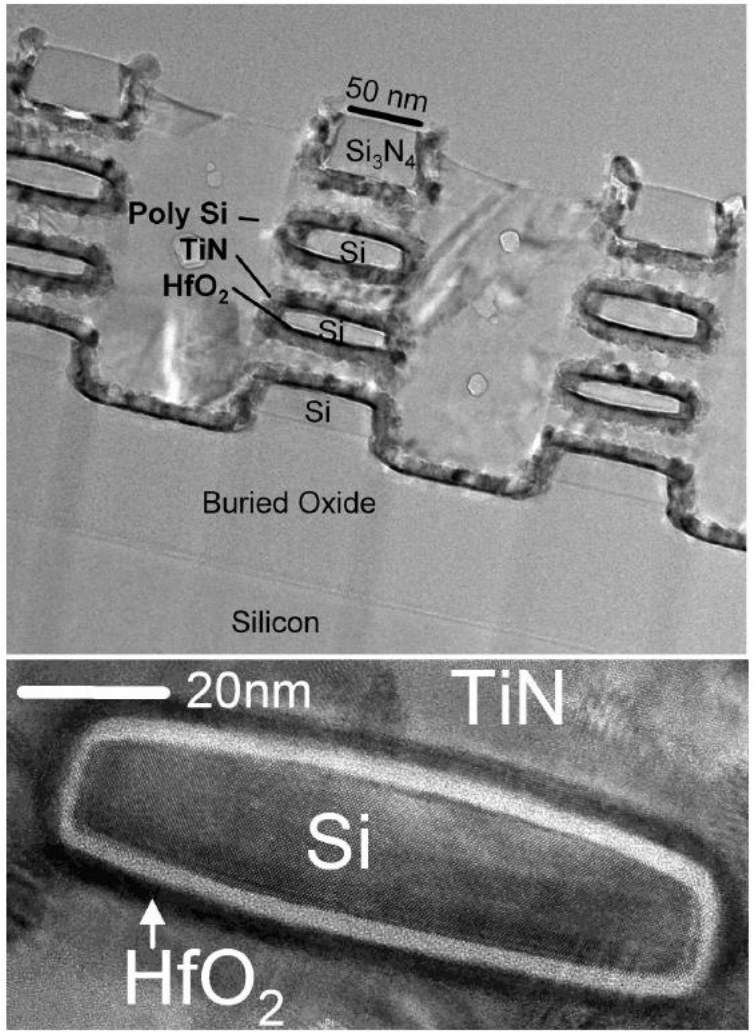




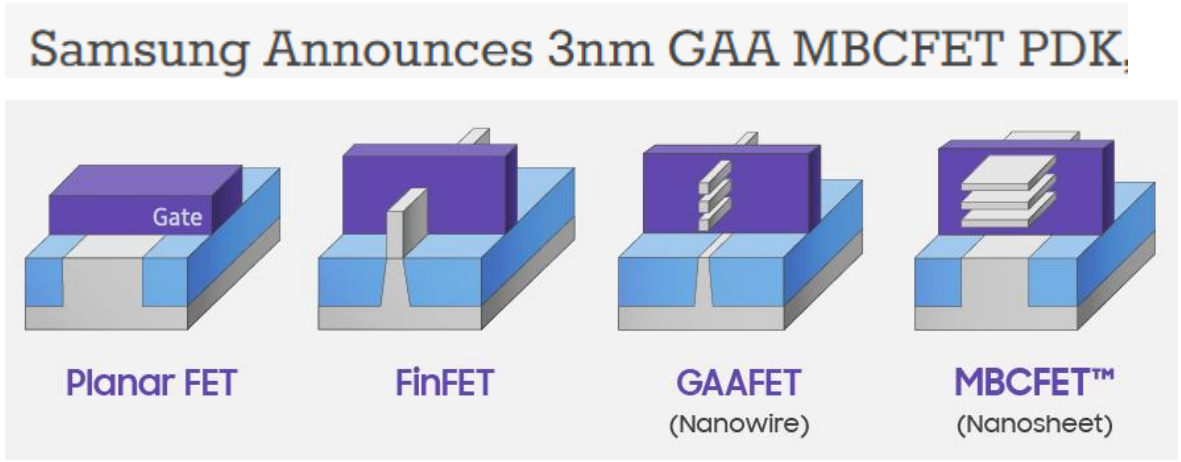
Workshop sessions

- Silicon process
- AI oriented and neuromorphic architectures
- New technologies and PIM
- Silicon photonics and analog computing
- Transversal challenges, wrap up and next steps

NanoSheets will be in production for 3nm node

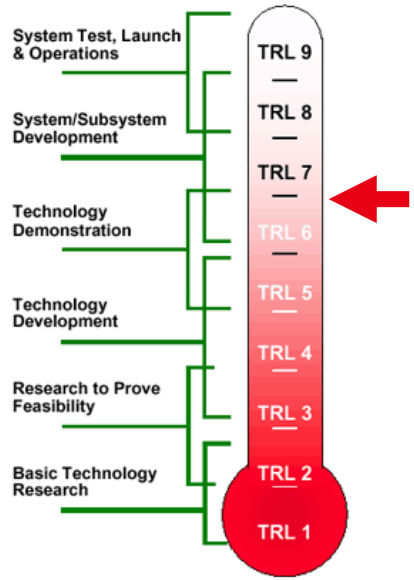


LETI, Dec. 2006



Samsung, May 2019

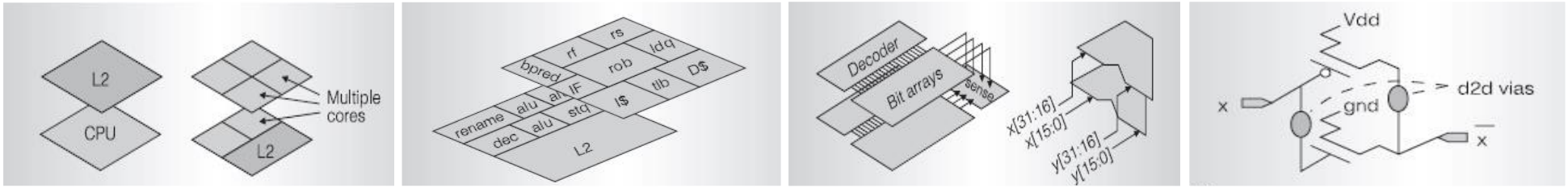
- Relevance: low-power 3nm transistor
- Timeline: already exist
- Value chain: not in Europe today



Two 3D VLSI complementary approaches @ LETI

3D integration scheme

1-Entire core	2-Logic bloc	3-Logic gates	4-Transistors

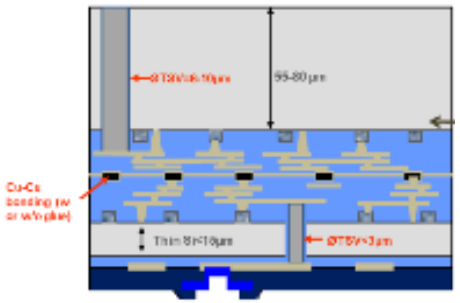


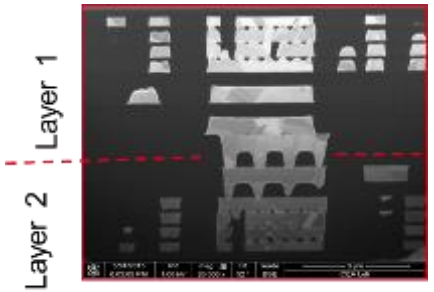
Granularity scale

3D Parallel

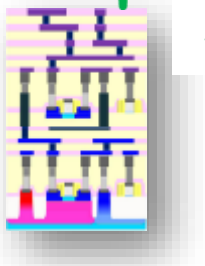
3D Sequential

Partitioning granularity





Cu-Cu bonding

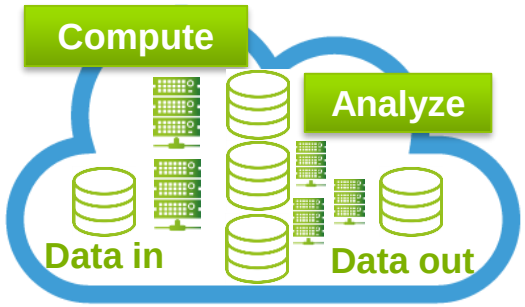


CoolCube™

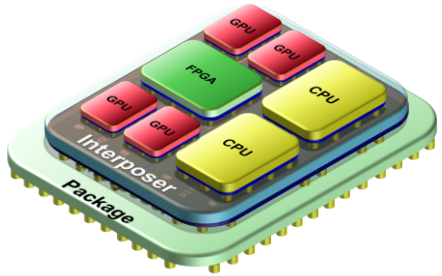
Interconnect pitch

100µm | 10µm | 1µm | 100nm | 10nm

TAKE AWAY MESSAGES



- A **major** evolution of High Performance processor architecture required to cope with **data deluge and energy efficiency**
  - 3D technologies to provide heterogeneous integration
  - Many available products: HBM, 2.5D interposers



- **Chiplet Partitioning & Active Interposer**
  - More & more WW interest
  - CHIPLET: enabler for Yield, cost control, heterogenity, Genericity, Specialisation
  - ACTIVE interposer: enabler for Smart functions **Interconnect + Power Management + SoC infrastructure**

From IP provider …
… to chiplet provider !

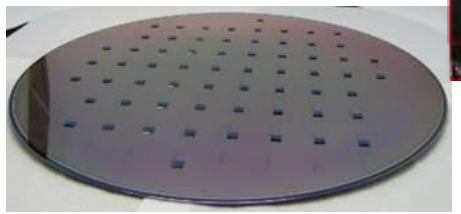
- **Proof of Concept achieved**
  - 96 core demonstrator
  - FDSOI 28nm **energy efficient** chiplet + 65nm **Smart** Interposer
  - Scalable concept

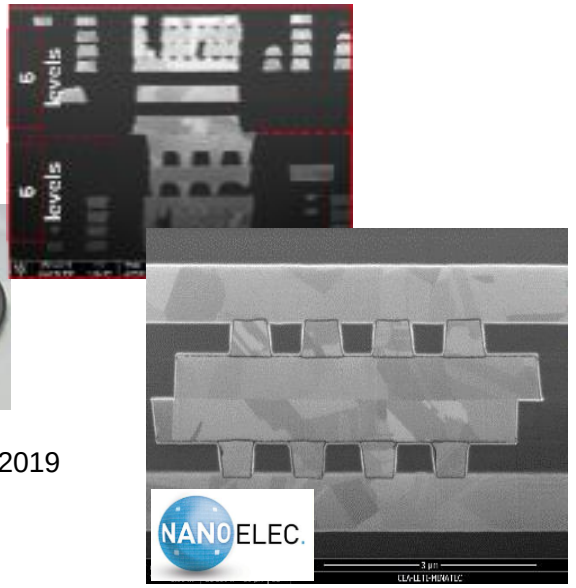
Terabit/s/mm2 achievable

IntAct performances to be published at ISSCC

ON-GOING NEXT STEPS

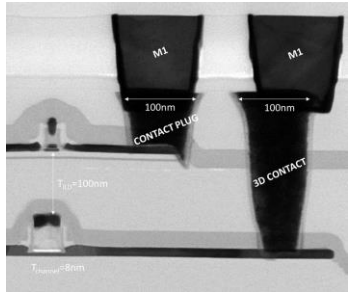
- **D2W direct hybrid bonding**
  - Reduced chip-to-chip pitch : 3-5µm foreseen
  - Reduced chiplet-to-chiplet gap
  - Better thermal coupling & reliability





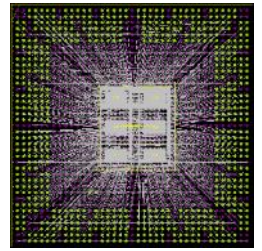
P. Metzger & al, Minapad 2019
A. Jouve & al, 3DIC 2019

- **Towards ultimate pitch thanks to 3D Sequential**
  - Coolcube™ CEA concept
  - Combination of sequential / parallel techologies for the best trade-off performances / cost



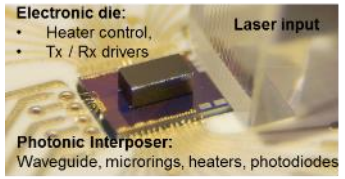
- **Partitionning & CAD aspects**
  - Co-design between chiplet ⇔ interposer ⇔ package mandatory
  - Assembly Design Kit + more CAD automation is required

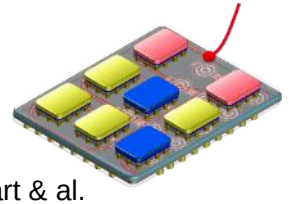
Mentor



XSI tool, Mentor Graphics

- **The next Smart Interposer ?**
  - Photonic Interposer !!!
  - *Convergence of 3D & Photonics*



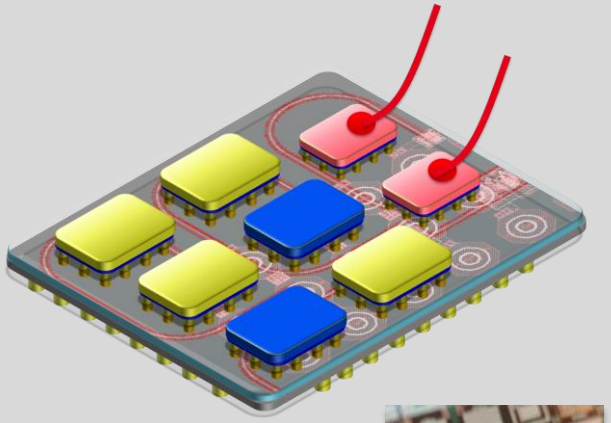


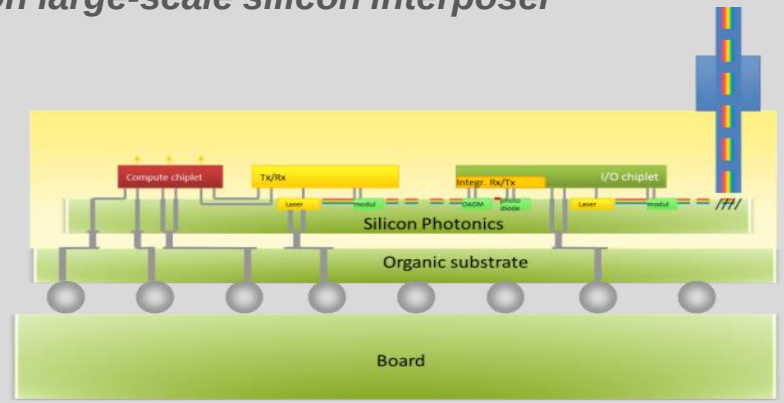
Y. Thonnart & al.
ISSCC'2018

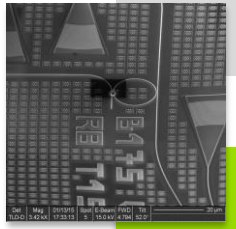
OPTICAL COMMUNICATION ON INTERPOSERS

Key technologies for chip-to-chip photonic communication

Multiprocessor subsystem on large-scale silicon interposer



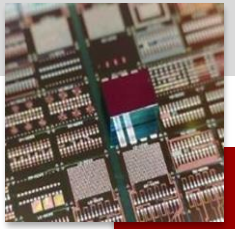




Silicon photonic subsystem

1

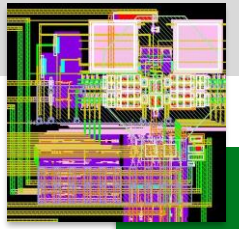
- Full WDM link integration
- Integrated lasers
- Fine-grain thermal control
- Circuit-switched routing
- Optical IO
- Dense integration



Large scale circuit integration

2

- Die assembly on interposer
- Fiber coupling
- TSV & microbump IOs
- Thermal dissipation
- Mechanical stress



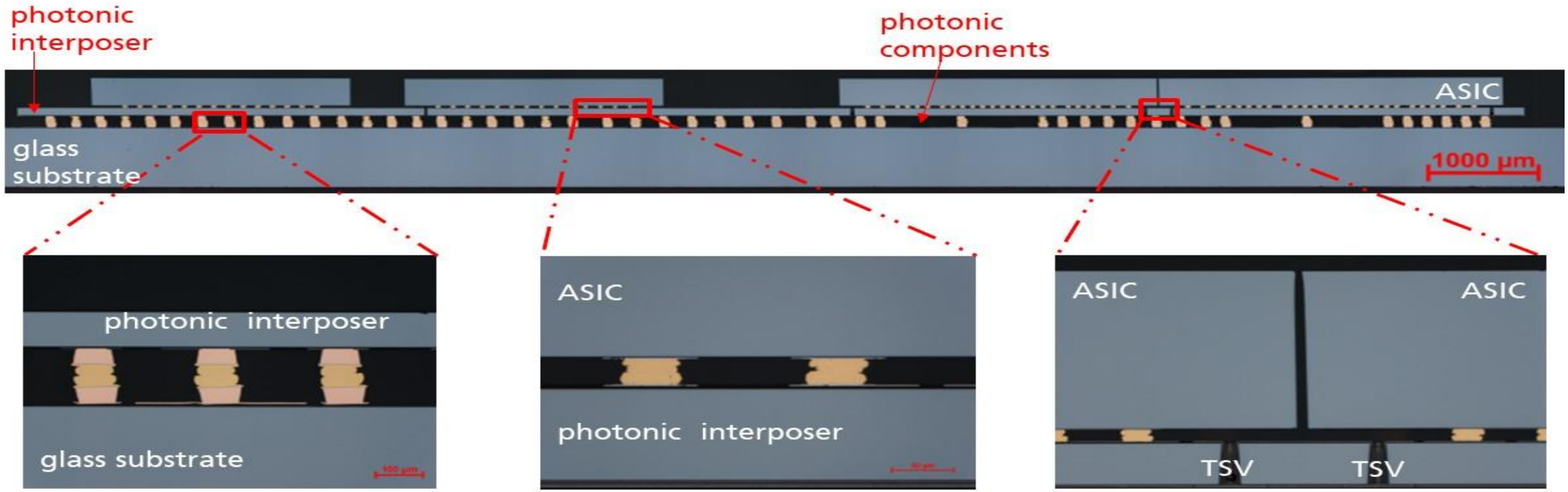
Architecture & circuit design

3

- Optical NoC topology
- Generic E/O chiplet for communication
- Routing, flow-control & arbitration
- Tx/Rx electro-optical drivers
- Autonomous thermal control
- Integration in computing fabric

PhoxTroT: Silicon Photonic Interposer

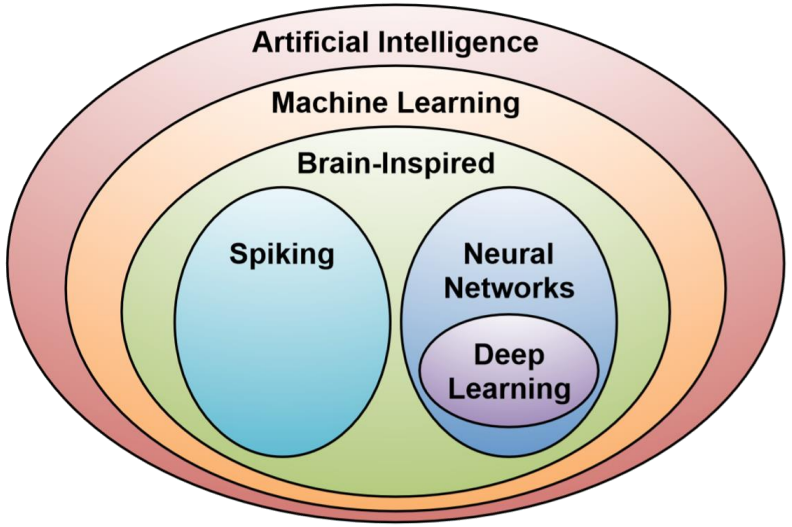
- The objective is to develop an **underlying technology** to enable next generation photonics to overcome these challenges and leverage low-latency and high-bandwidth communication.



Source: www.PhoxTroT.eu

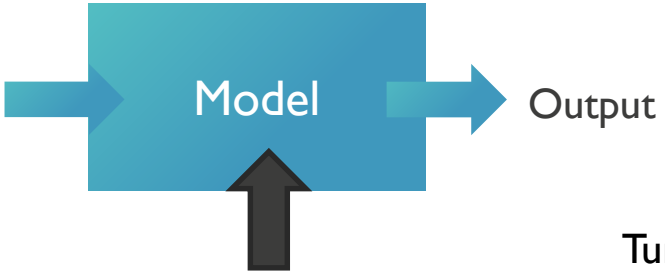
INTELLIGENT HARDWARE?

FROM DEEP NEURAL NETWORKS TO NEUROMORPHIC



V. Sze, et al. "Efficient processing of deep neural networks: A Tutorial and Survey", Proc. of the IEEE, Vol. 105, No. 12, Dec. 2017 [https://arxiv.org/abs/1703.09039]

MACHINE LEARNING



Tuning model parameters based on available data = "learning" without explicit programming

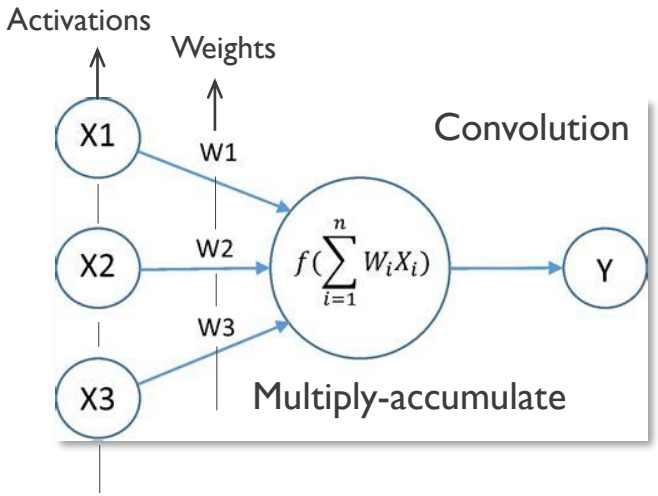
- Pattern Recognition
- Feature Extraction
- …

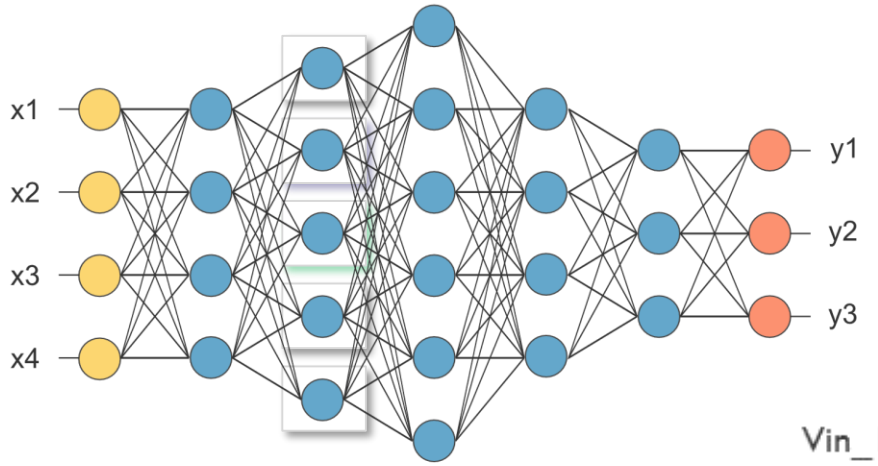
BRAIN-INSPIRED

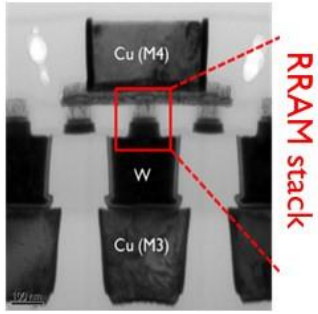
Using artificial neural networks as machine learning model

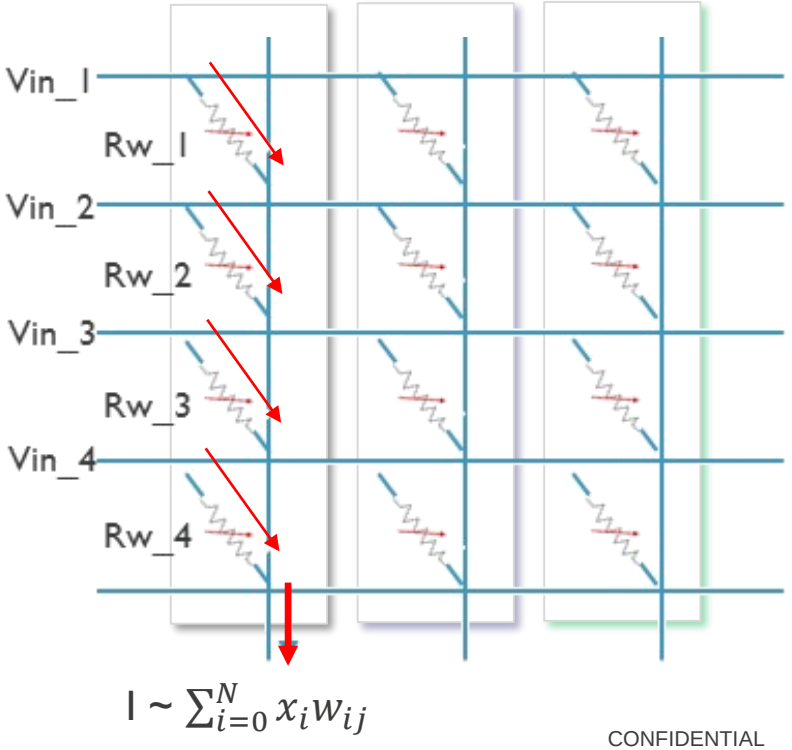
- Multilayer perceptron
- Convolutional Neural Networks (CNN)
- Long Short Term Memory (LSTM)
- Spiking Neural Networks (SNN)
- Hierarchical Temporal Memory (HTM)
- …

ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML
SUPPORTED BY NEW MEMORY TECHNOLOGY









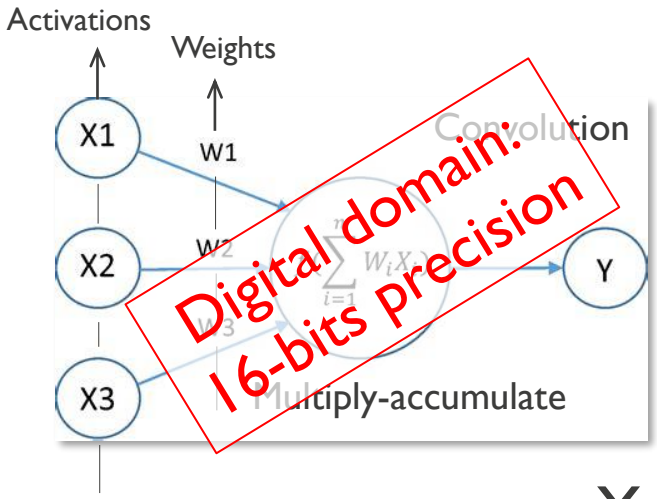
- Use memory array for massive parallel analog implementation of multiply-accumulate operations in DNN layer
- Memory array stores weights *and* implements a logic function (MAC) in analog fashion

  → compute-in-memory
  → computational memory
  → neuromorphic computing

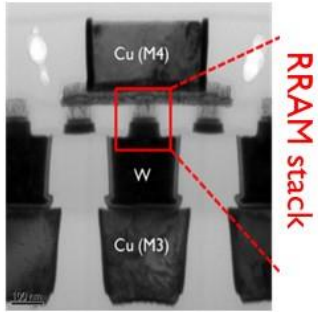
imec

ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

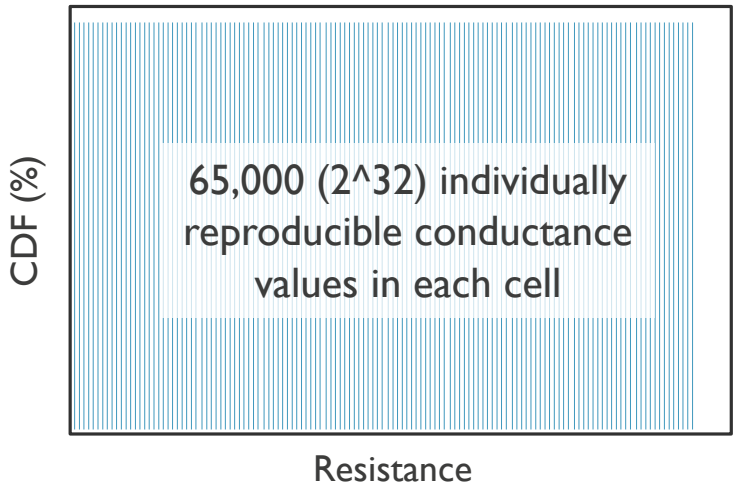
CHALLENGES …



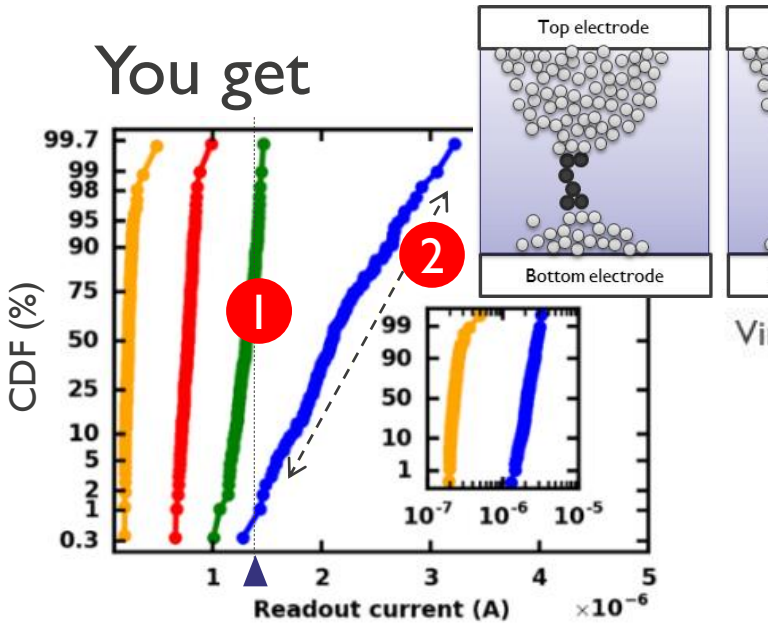
1. Memristor (programmable resistor) stores weight value W_I as an analog quantity: conductance Rw_I
2. Activation X_I applied as analog voltage Vin_I
3. Ohm's law: memristor cell current ~ $X_I.W_I$
4. Kirchoff's law: bit-line current ~ $\sum X_i.W_i$

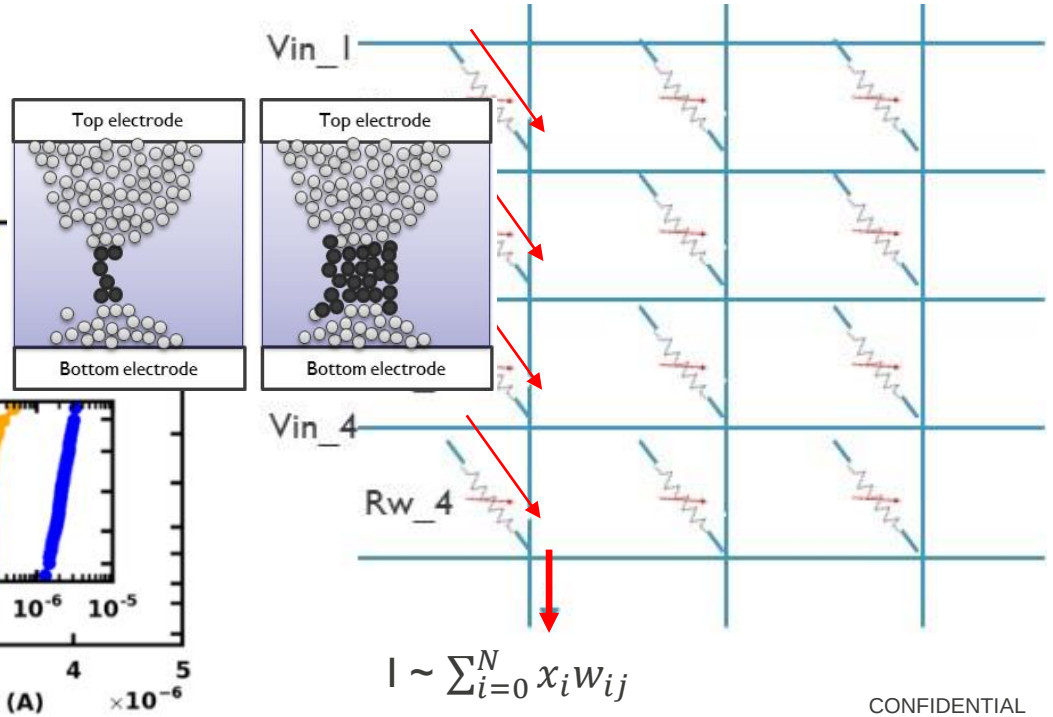


You want

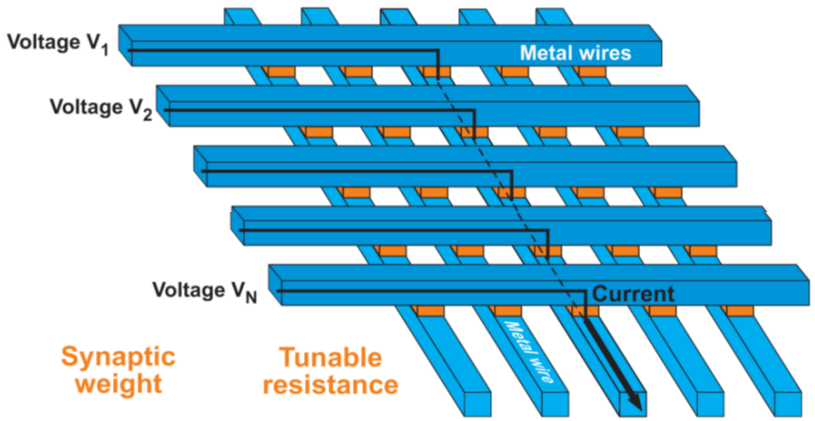


You get

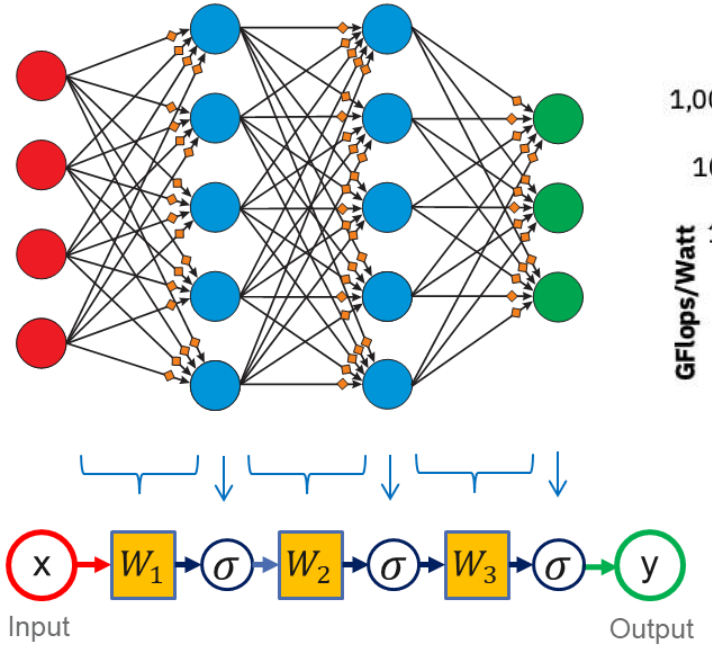




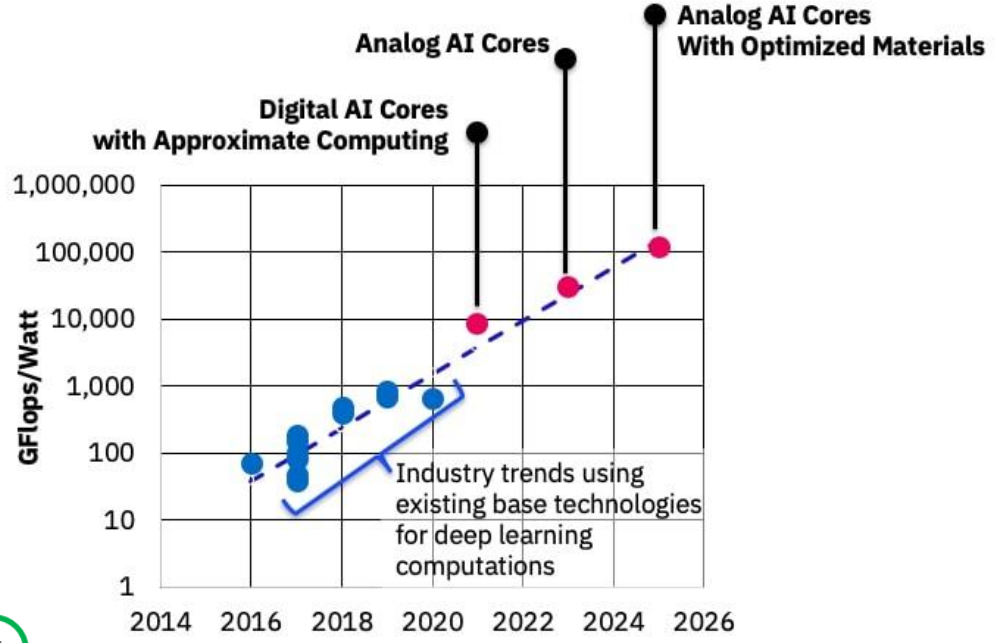
AI Technology roadmap



Analog synaptic processing



Neural Network architecture



Compute performance efficiency

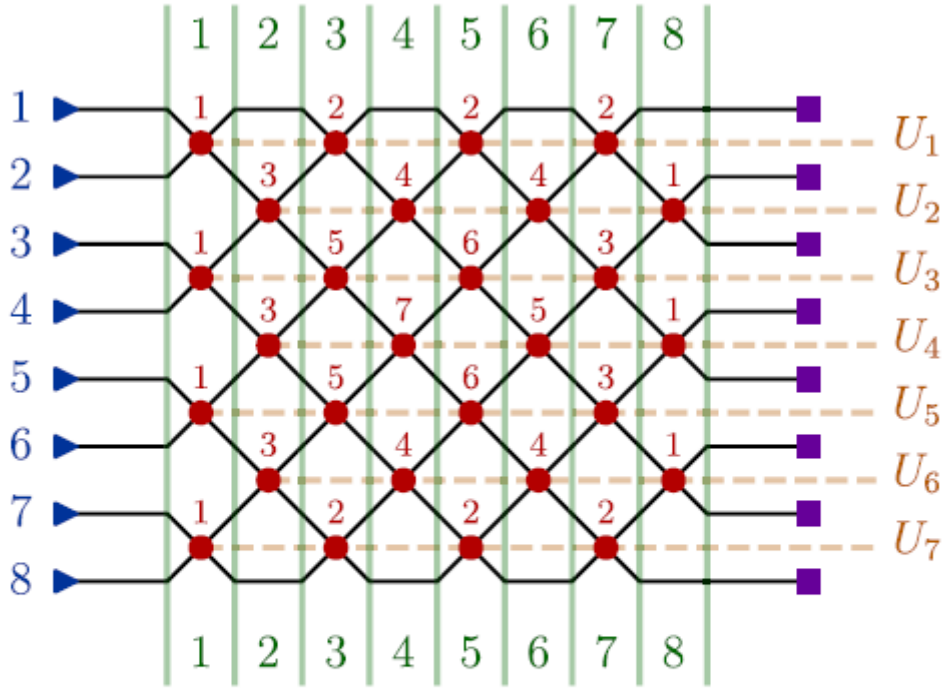
New memristive devices required

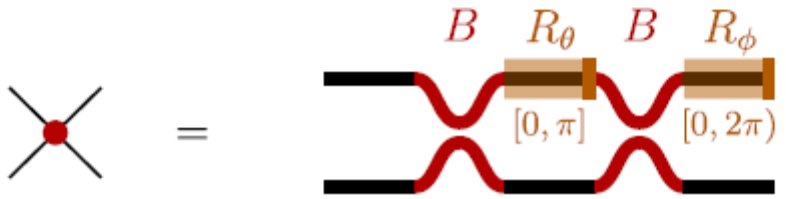
	Inference	Training
Resistance	1-100 MΩ	1-100 MΩ
# Levels	100	1000
Weight set / update	To desired level	Symmetric

Bert Jan Offrein – ofb@zurich.ibm.com

Application example : Universal Multiport Interferometers

- Implementation of any linear transformation between multiple channels
  - Factorization of any *N*×*N* unitary matrix into a sequence of 2x2 unitary transformations
- composed of a regular mesh of beam splitters and phase shifters
- straightforward fabrication using integrated photonic architectures and ready scalability





Sunil Pai et al, "Matrix Optimization on Universal Unitary Photonic Devices," in PHYSICAL REVIEW APPLIED 11, 064044 (2019)
William R. Clements et al, "An Optimal Design for Universal Multiport Interferometers," in Optica Vol. 3, Issue 12, pp. 1460-1465 (2016)
Reck et al, "Experimental realization of any discrete unitary operator," in Phys. Rev. Lett. 73, 58 (1994)

BIRD'S EYE VIEW OF THE TEMPO PROJECT

ECSEL 2018 (*)
(19 partners, 33 ME budget)

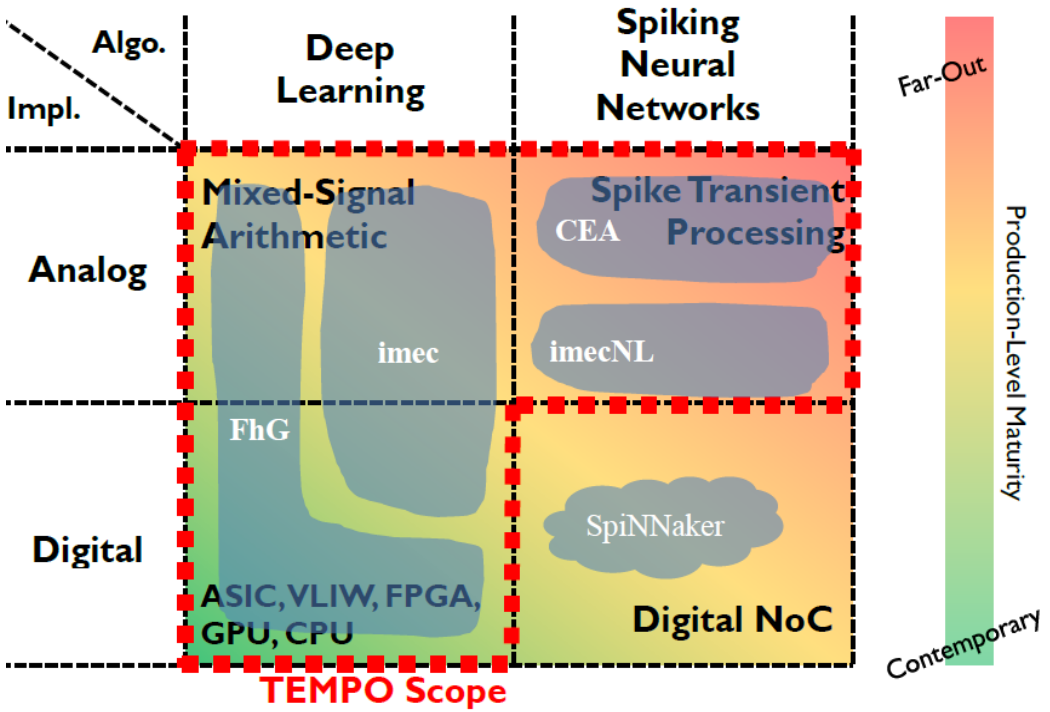
Innovative Design : CNN and SNN

Different Memories

- MRAM
- FeRAM
- PCRAM
- OxRAM

Silicon Technology

- 300mm Silicon wafers
- 22 and 28nm FDSOI technology

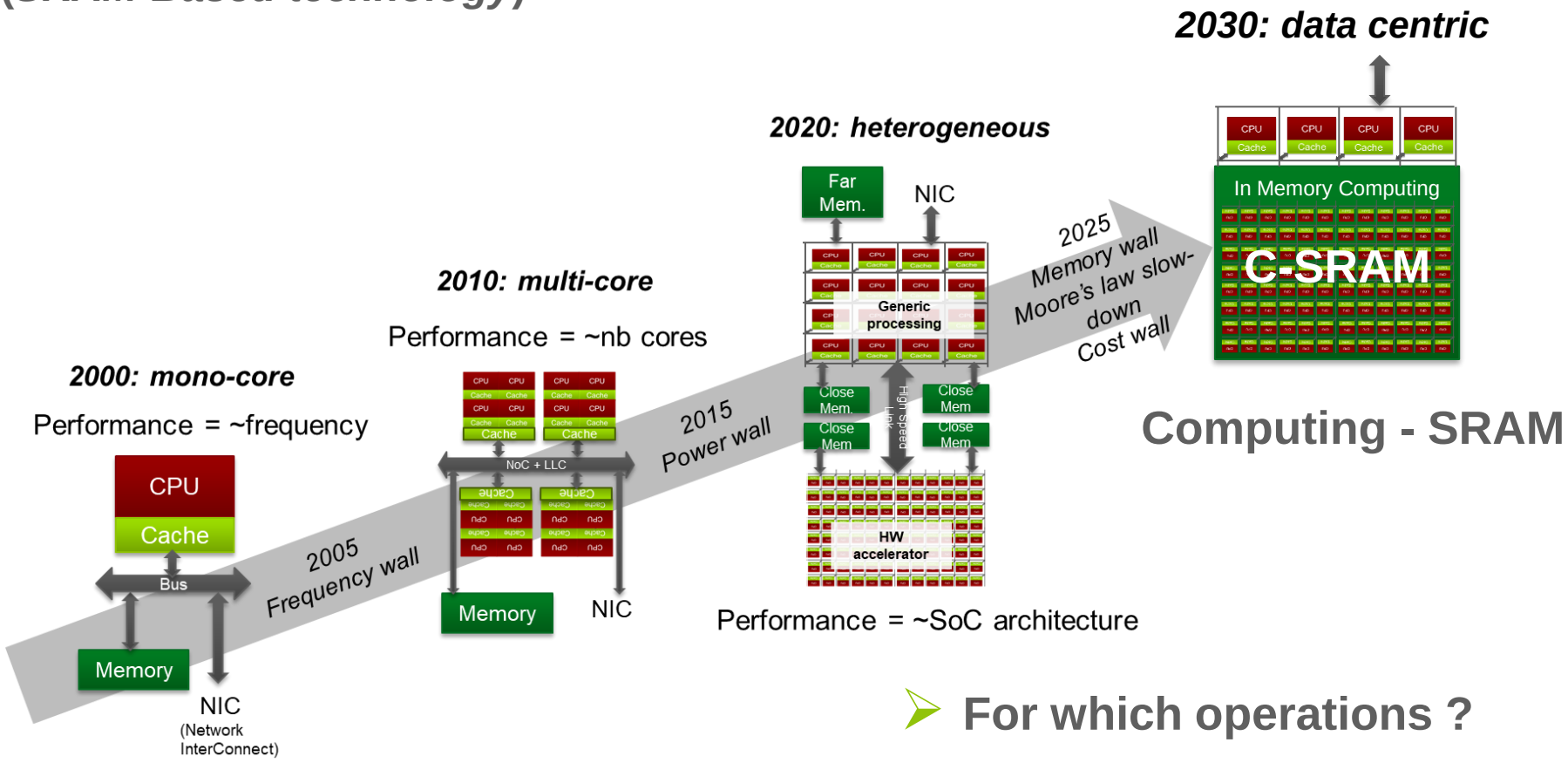


Next EU proposal: "ANDANTE" (ECSEL 2019)

(*) The TEMPO project has received funding from the Electronic Components and Systems for European Leadership Joint Undertaking under grant agreement No 826655. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, Switzerland, The Netherlands.

CEA PROPOSAL

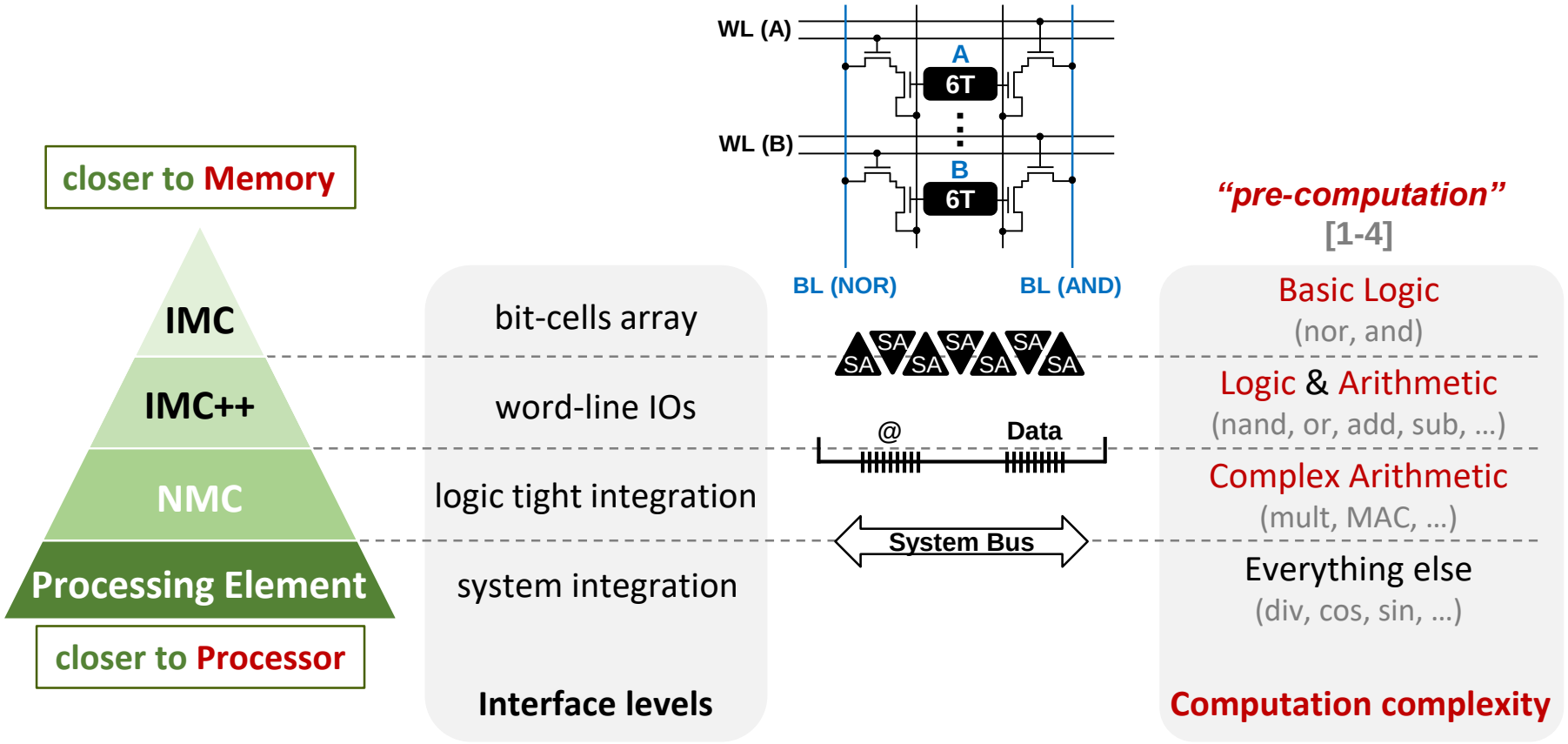
Concept (and CEA proposal): Bring computation closer to the memory (*SRAM-Based technology*)



Computing - SRAM

➢ **For which operations ?**

IN-MEMORY OR NEAR-MEMORY COMPUTING?



Source: R.Gauchi, VLSI-SoC 2019

[1] K. C. Akyel, DRC², 2016
[2] S. Aga, Compute Caches, 2017
[3] Y. Zhang, Recryptor, 2018
[4] A. Agrawal, X-SRAM, 2018

C-SRAM -> SOME RESULTS / 2 APPLICATIONS

- **AES - Advanced Encryption Standard application**

Cryptography	Scalar vs. C-SRAM
Clock Cycle	**x84**
Energy (nJ)	**x47**

- **Frame Difference application**

	Clock Cycle		Energy
Image size	**Scalar vs. CSRAM**	**SIMD vs. CSRAM**	**Scalar vs. CSRAM**
4x4	**x32**	x3.9	**x18**
VGA	**x6614**	x260	**x 32**

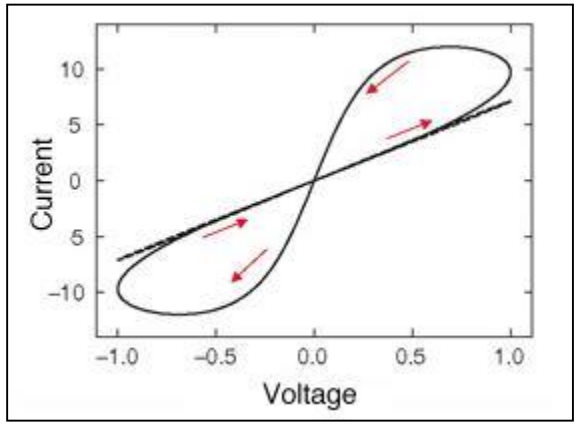
Source: CEA Leti

WHAT IS A MEMRISTOR OR A MEMRISTIVE SYSTEM?

- **Memristors or better memristive devices as common roof**
  - „If it‘s pinched it‘s a memristor“

Leon Chua, If it’s pinched it’s a memristor, IOP Publishing, Semiconductor Science and Technology, 29 (2014), 104001 (42pp)

L. Chua, S. Kang, Memristive devices and systems, Proc. IEEE, 64 (2), 209-223, (1976)



J. Walker, Memristors and the Future
http://www.nobeliefs.com/memristor.htm

Dynamical system

$$y = h(x, u, t)$$
$$\dot{x} = f(x, u, t)$$

↓

Memristive system

$$y = h(x, u, t) \cdot u$$
$$\dot{x} = f(x, u, t)$$

←

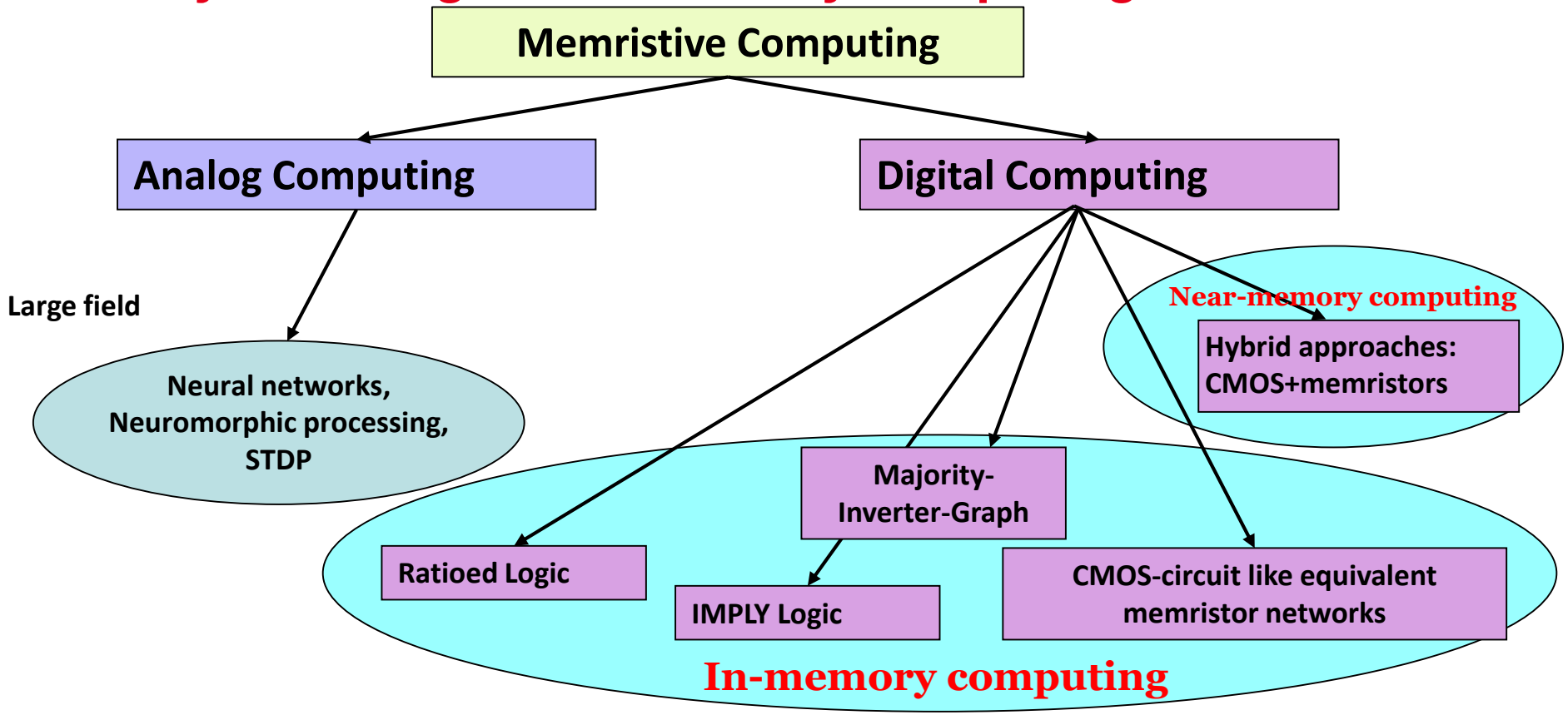
Memristor

$$y = R(q) \cdot I$$
$$\dot{q} = I$$

$$y = G(\Phi) \cdot V$$
$$\dot{\Phi} = V$$

←

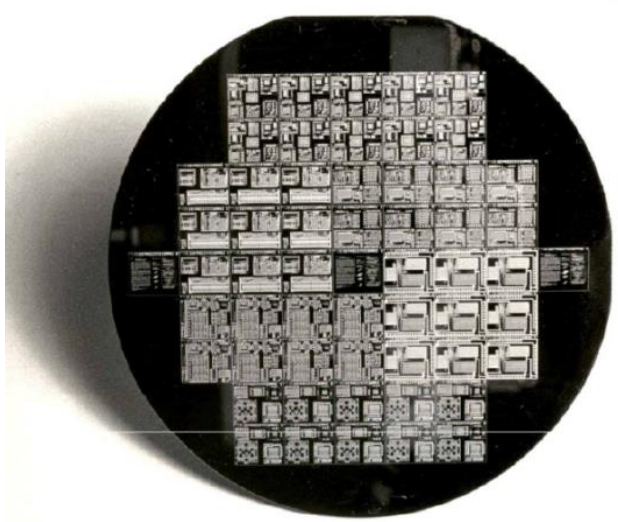
- More visionary in-storage or in-memory computing

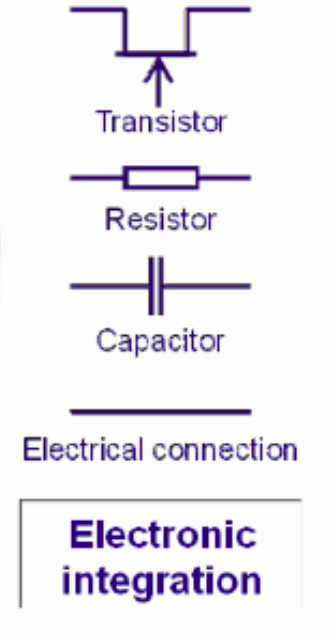


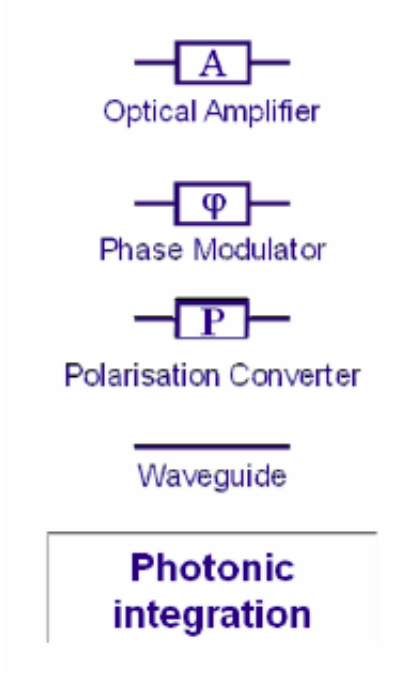
Adopt Foundry Model from Electronic ICs to InP PICs

Like Electronics: Make Building Blocks, Separate Design from Process

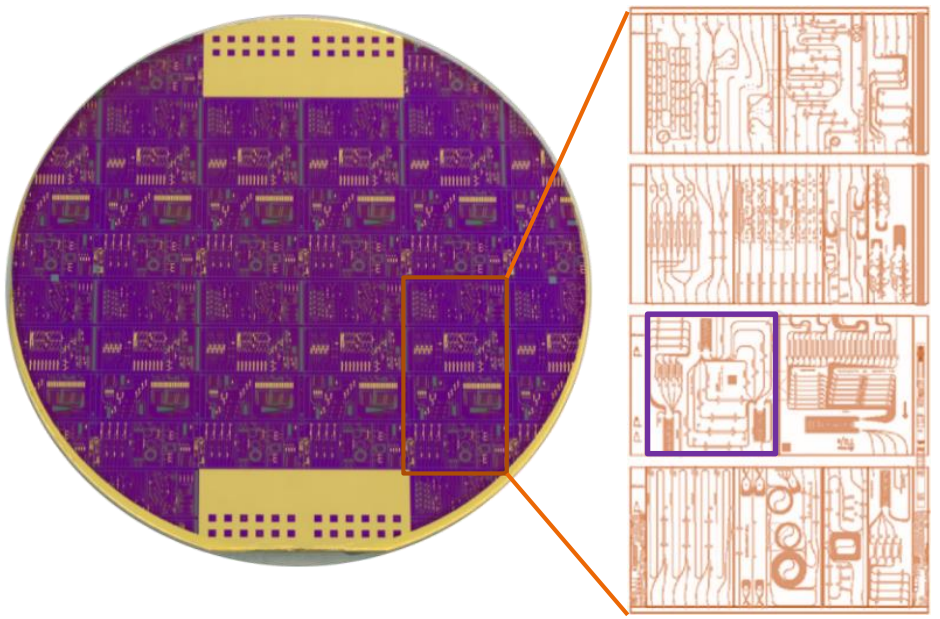
Silicon ICs ~1979





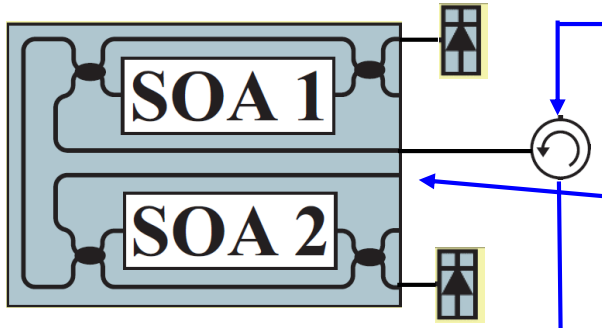


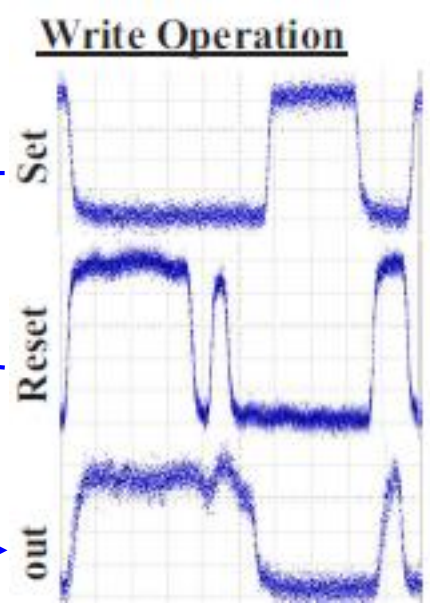
InP Photonic ICs ~2014

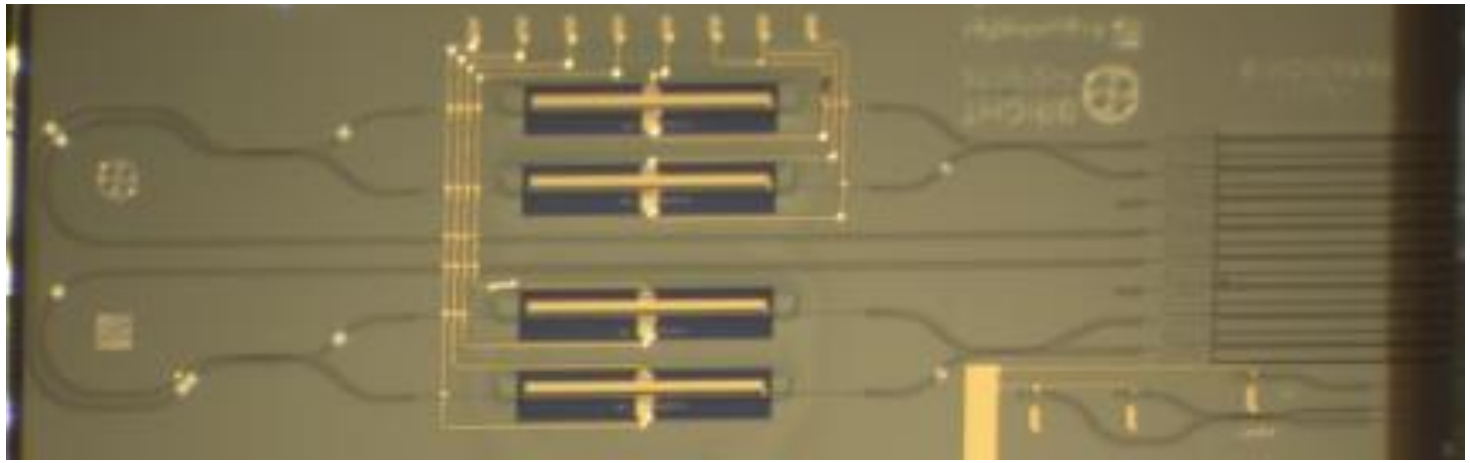


All-Optical Memory

Flip-Flops Based on Photons, not Electrons







2 mm

6 mm

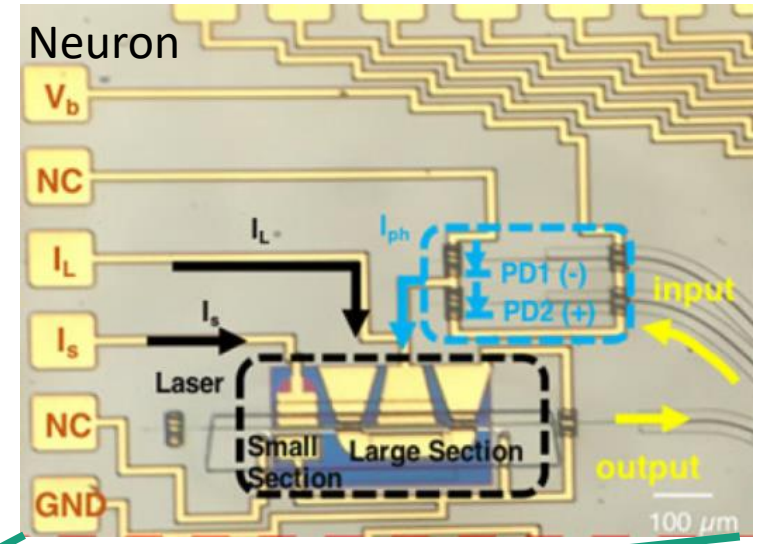
BRIGHT Integrated PHOTONICS

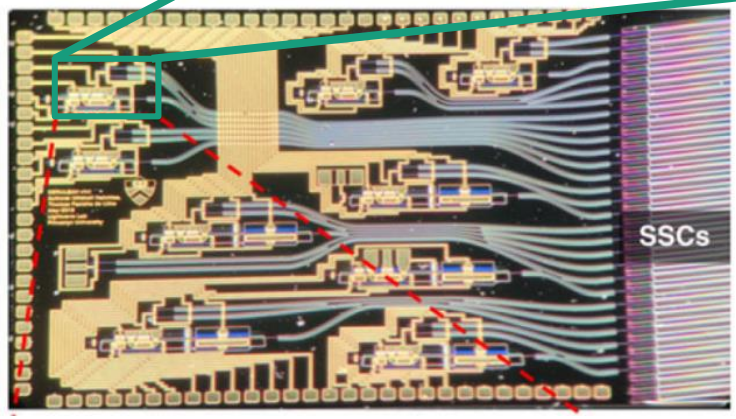
ARISTOTLE UNIVERSITY OF THESSALONIKI

G. Mourgias-Alexandris, et al., "All-optical 10Gb/s ternary-CAM cell for routing look-up table applications," *Opt. Express,* Mar. 2018.

All-Optical Neuron for Computing

Breaking von-Neumann Bottlenecks

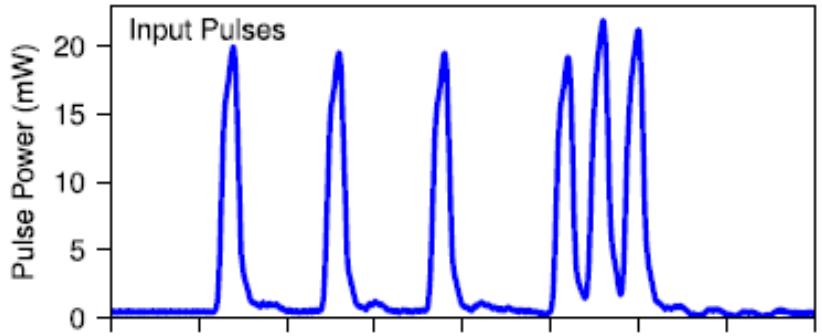


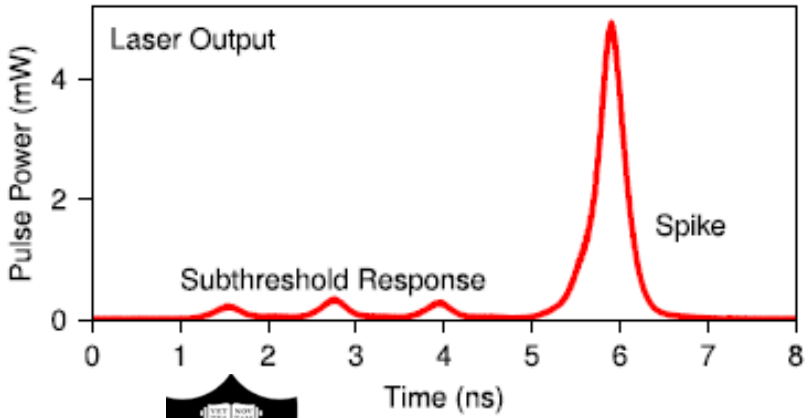


4 mm

6 mm

spikes encode the timing between input pulses





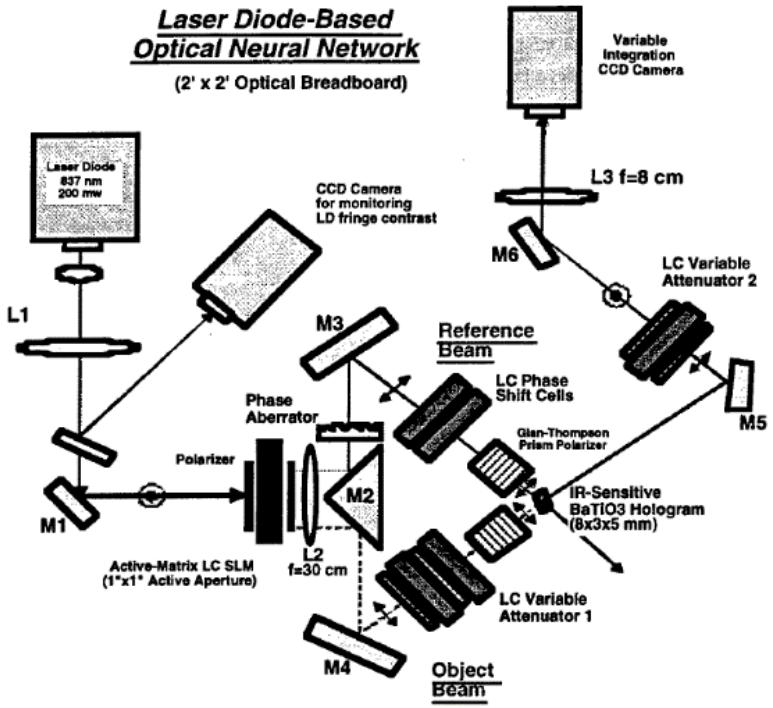
Princeton University

H.-T. Peng et al. “Neuromorphic Photonic Integrated Circuits” *IEEE JSTQE,* 2018

Optical crossbar arrays: Integrated Solution

Concept demonstrated in bulk optics

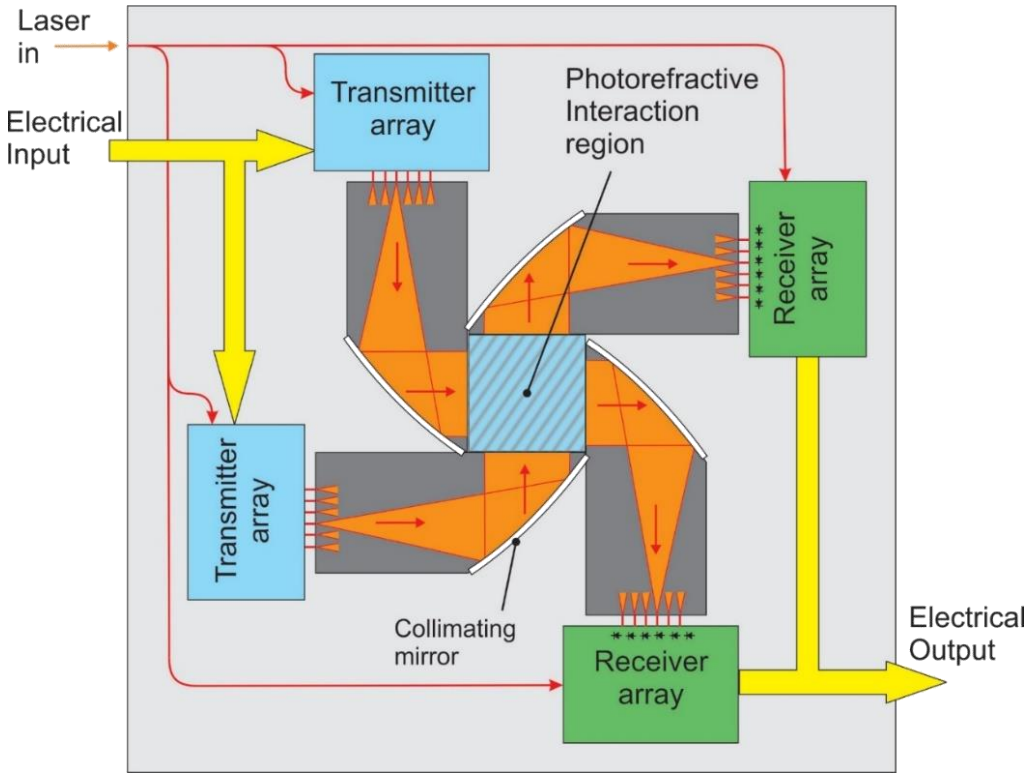
- Backpropagation training of neural networks with hidden layers
- Large setup, slow electro-optics, stability issues



📖 Yuri Owechko and Bernard H. Soffer, "Holographic neurocomputer utilizing laser diode light source“, 1995

Our approach: Miniaturize using Integrated Optics

- Electro-optic conversion and beam shaping optics on a silicon photonics chip
- Memory: Photorefractive thin film on silicon



Bert Jan Offrein – ofb@zurich.ibm.com

The rise of co-processors

A Fourier transform can be obtained thanks to the use of lenses.

An O(n^2) operation becomes an O(1) operation in optics

- Optical Synthetic Aperture Radar Processor
- Optical Correlators for Pattern Recognition
- Joint Transform Correlator, etc…

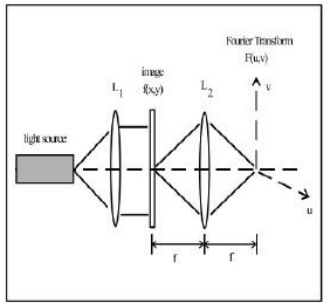
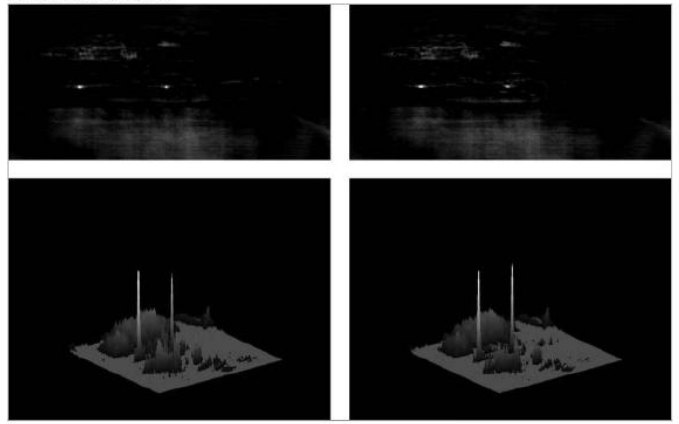


Figure 1: Fourier Transform by a lens. L_1 is the collimating lens, L_2 is the Fourier transform lens, u and v are normalized coordinates in the transform plane.

ent vehicles with complex background and white foreground showing high intensity. Left: objects well ht: one of the cars to be detected is occluded by a jeep

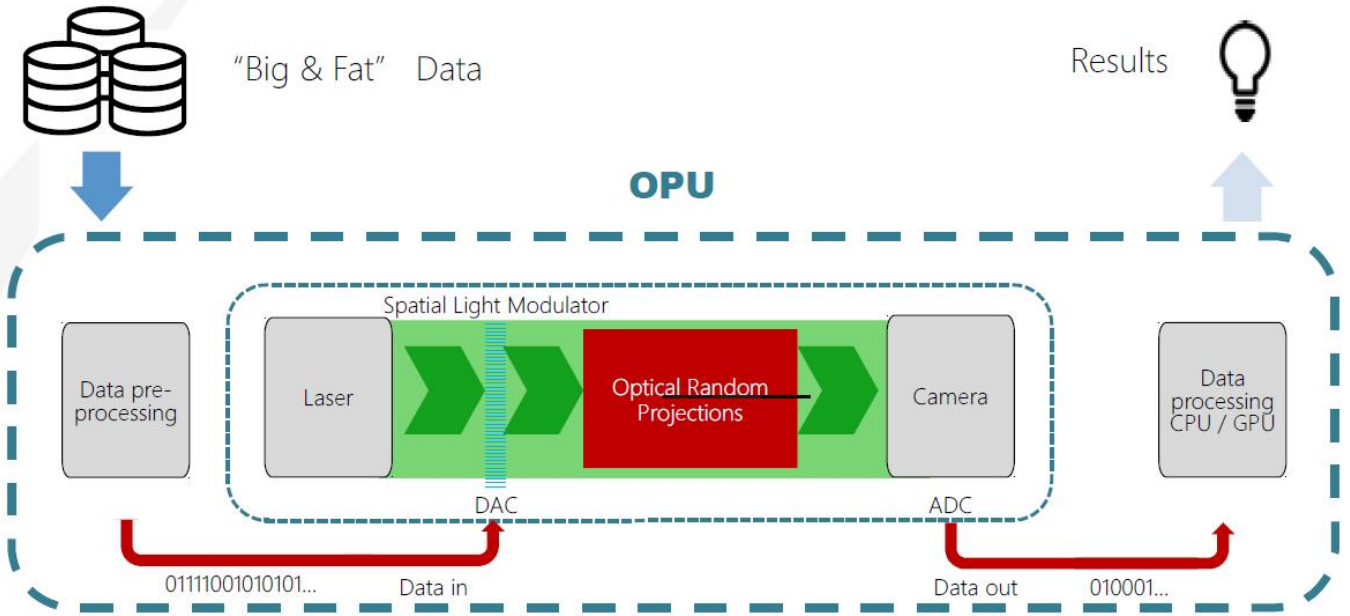
Plate 3 Optical correlations obtained with the images of Plate 2. Top: correlation planes, bottom: three-dimensional plots of the correlation planes. Left: the images with the two cars entirely separated. Right: the rightmost correlation peak corresponding to the occluded car. The white foreground correlation corresponds to the small hill behind the correlation peaks



Source:

- Alain Bergeron, (2000),"Optical correlator for industrial applications, quality control and target tracking", Sensor Review, Vol. 20 Iss 4 pp. 316 – 321
- http://www.phys.unm.edu/msbahae/Optics%20Lab/Fourier%20Optics.pdf
- Fourier Optics, J. W. Goodman, Mcgraw-Hill, 1996

Create and deploy hardware



Science Fiction Success Story SFSC

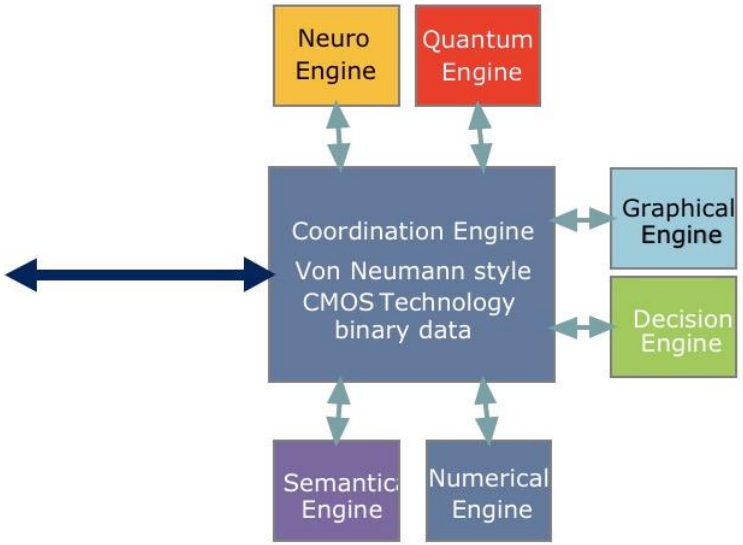
- Description of a future achievement
- Ambitious but realistic
- Description of the innovation
- Quantitative information
- How this has been prepared and achieved: European players involved
- Translation of the advantages into answer to societal challenges: for people directly connected to the field, for the European citizen

What as not work: provider-user research chain

- Some elements that will help you to guide your research, to assess it potential for downstream users, to compare it to state-of-the-art other technological solutions. Just to mention some of the ideas we have in mind:
  - Data sets
  - Benchmarks
  - Small application kernels
  - Communication patterns
- Symmetrically you may have challenges that you would like upstream teams to solve. If you are able to define what you are interested to get, it can help other research teams to focus on your concerns.

Take away

- Future will be diversity



- Integration is a key element

Potential recommendations

- For developing European technologies
  - New long term projects with real co-design but on very little kernels
  - Specification of API at package level

- For application development
  - Modular approach with skeleton of operations that could be accelerated
  - When possible analyze data precision requested by your computation

European Extreme Data & Computing Initiative

Questions ?

The EXDCI-2 project has received funding from the European Unions Horizon 2020 research and innovation programmed under grant agreement no. 800957.