ESSERC 2024

SiNANO-ICOS-INPACE Workshop

"Emerging technologies in Advanced Computation, Advanced Functionalities, Ground-breaking Technologies: Impact on International Cooperation"

Beyond Von Neumann computing architectures and Heterogeneous integration

Olivier FAYNOT
CEA-Leti, France
olivier.faynot@cea.fr

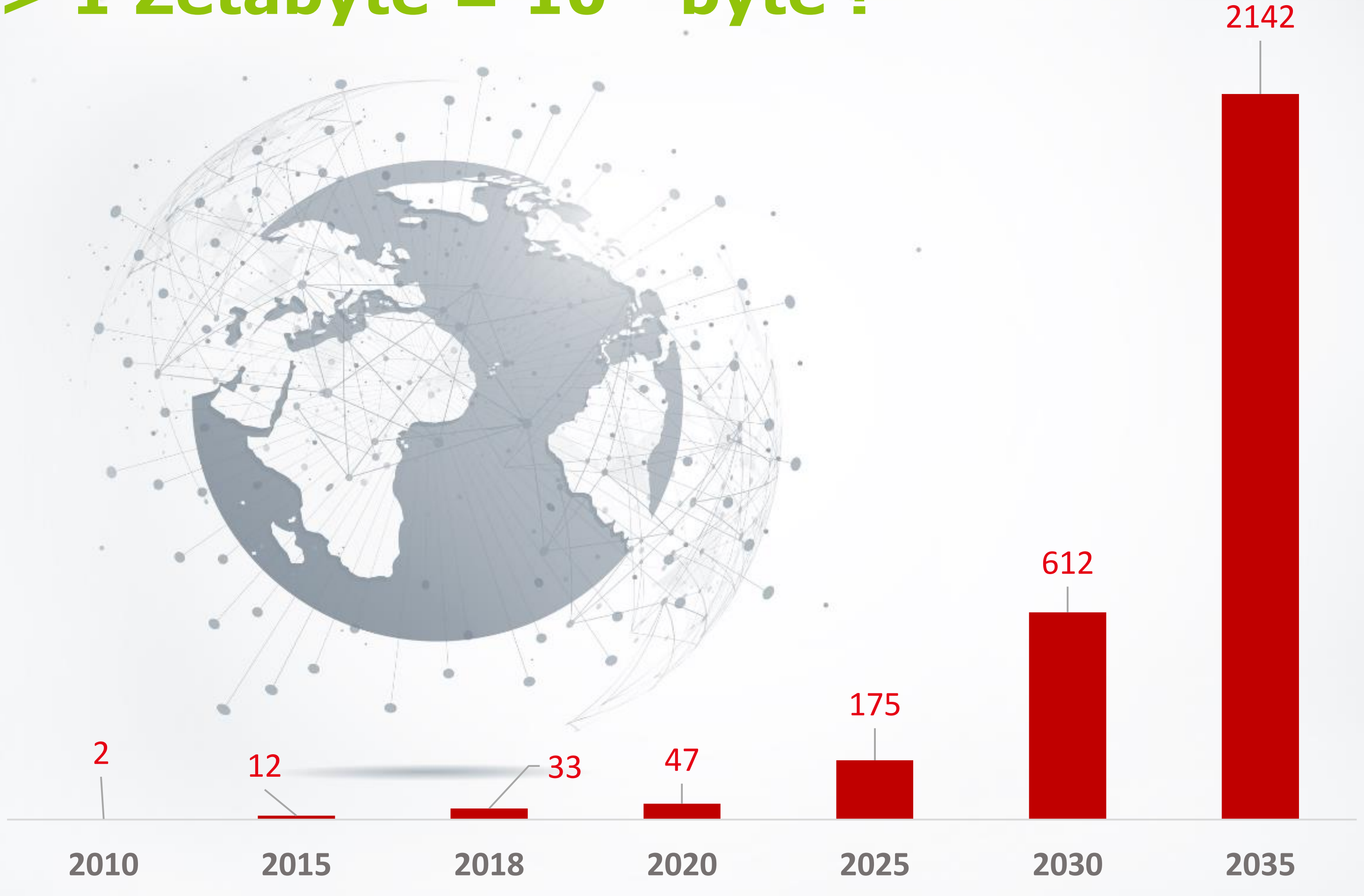
Leuven, September 9, 2024

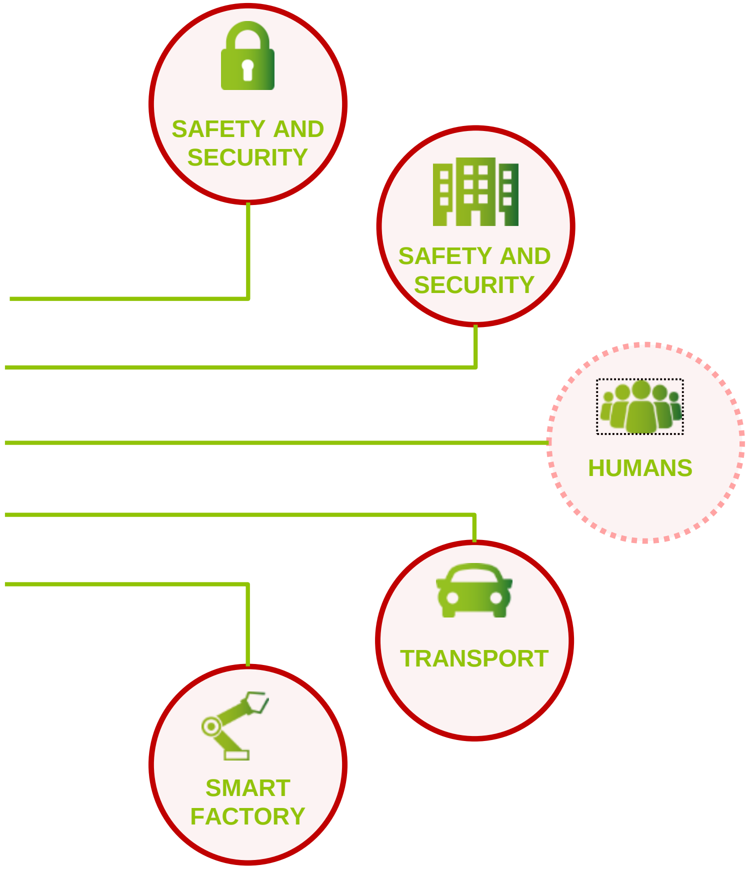
Outline

- Introduction: trends and challenges
- Memory technologies
- Beyond Von Neumann disruptive approaches:
  - Near or in-memory computing
  - Quantum computing
- Heterogeneous integration: from chiplets to functional backside
- EU and non-EU actors
- Conclusions

Global data generation (actual & forecast)

> 1 Zetabyte = 10^{21} byte !





> A true data deluge, not only generated by humans!

The required gain in energy efficiency

>1000x
by
2030

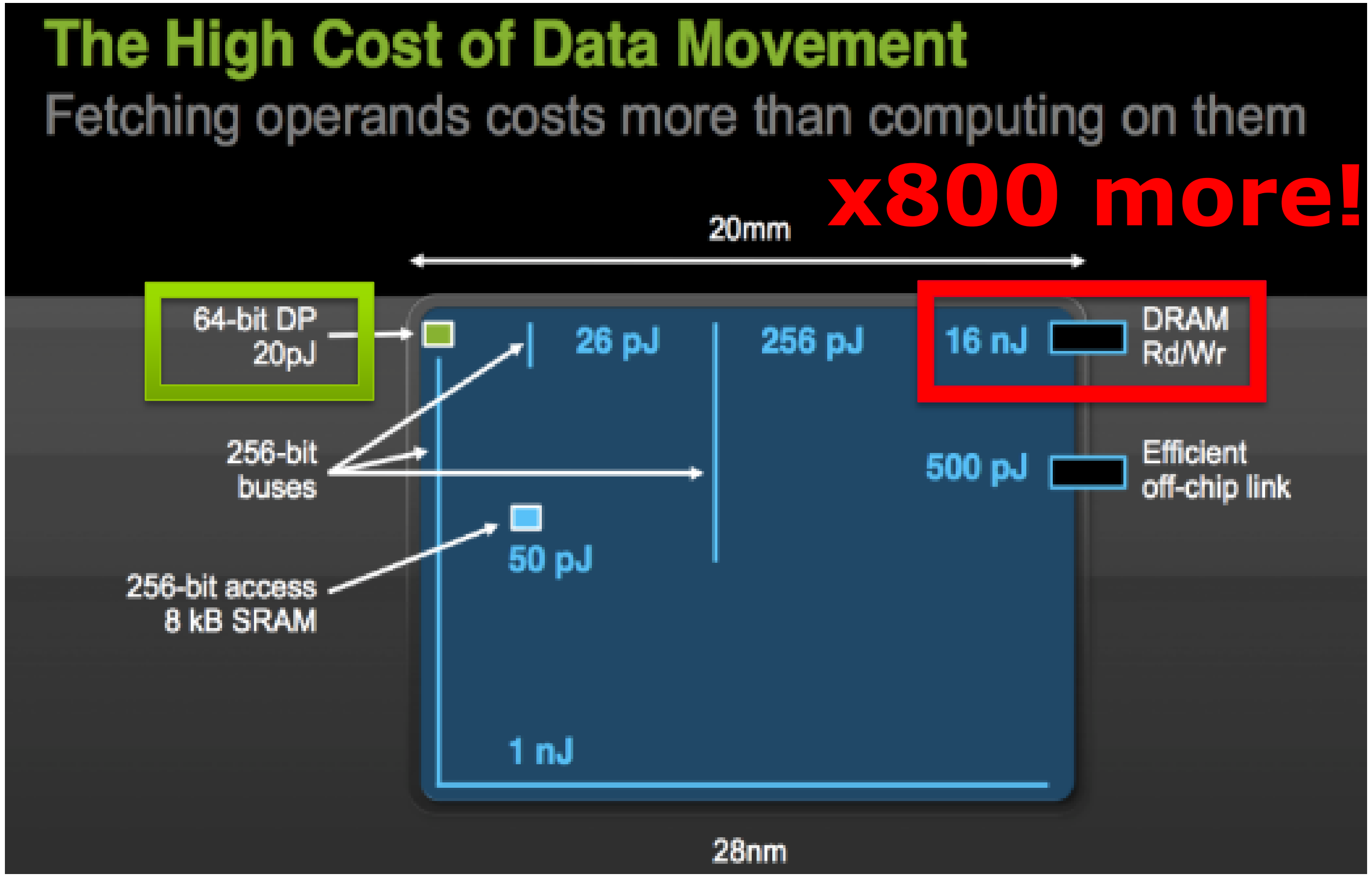
CMOS scaling

Memory technologies

Disruptive Computing

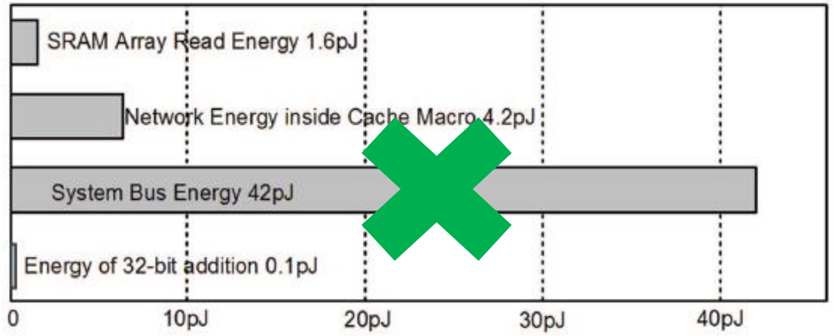
Chiplet & 3D System

The cost of moving data



Bill Dally, "To ExaScale and Beyond", 2010

[J. Wang – ISSCC'19]

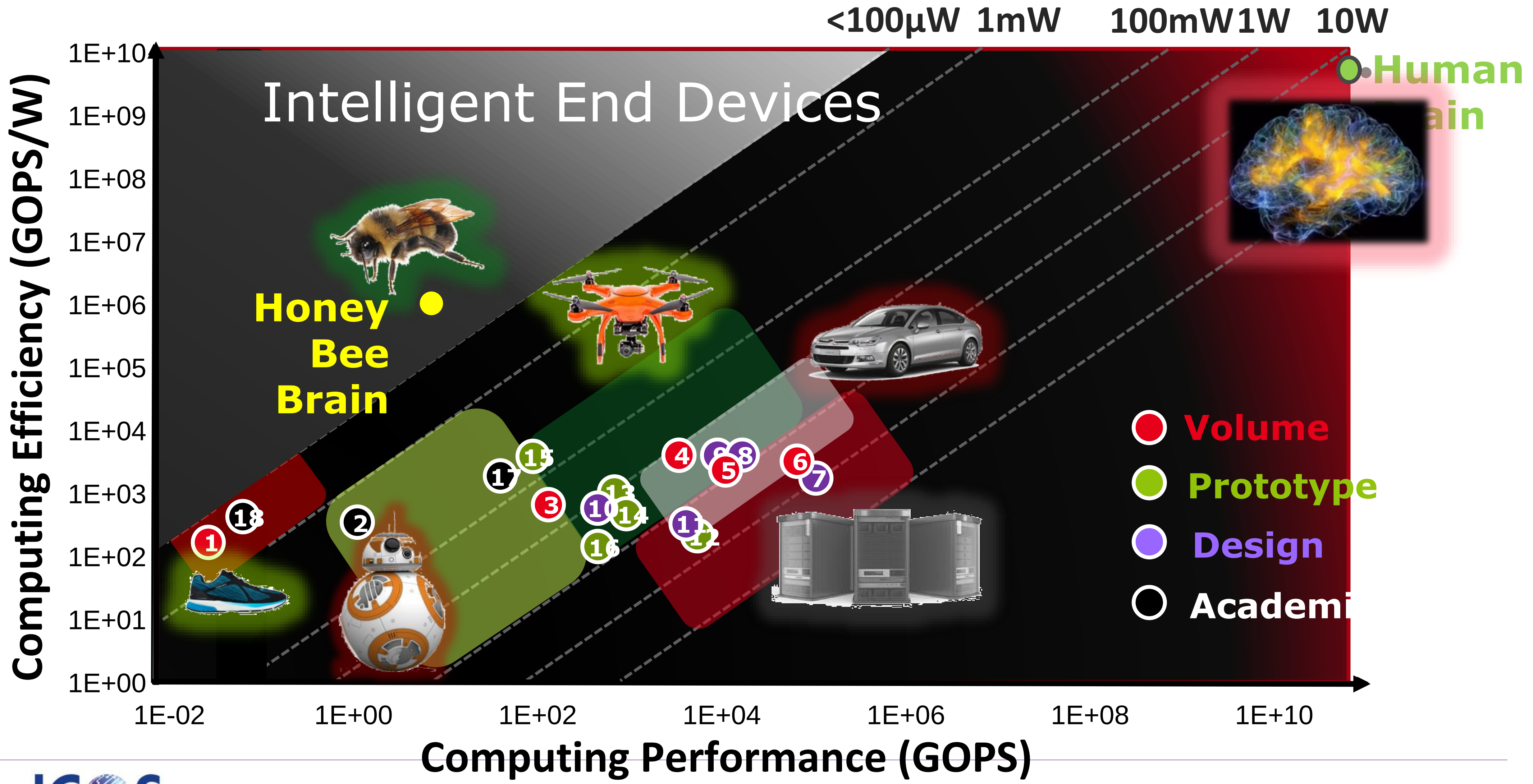


~90% of energy is in **data transfer**
→ IMC could lead **8x reduction**

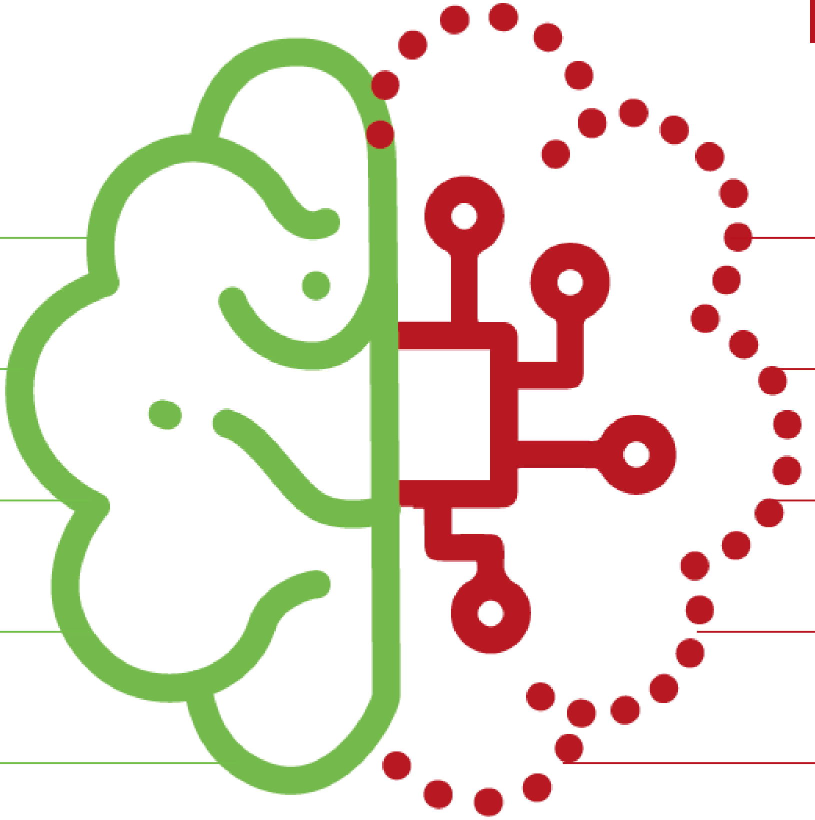
Operation energy is negligible

Memory access and control energies dominate

Energy Efficiency is far from bio Systems



Bio Inspired IC's: How?



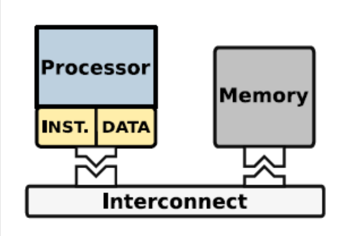
Biology	Technology choices
Asynchronous communication	Spike coding
Plasticity	Re-configurability (routing)
Sensing	Smart sensors
Brain is not flat, it is 3D	Dense 3D integration
Lifelong learning	New technologies coupled with algorithms

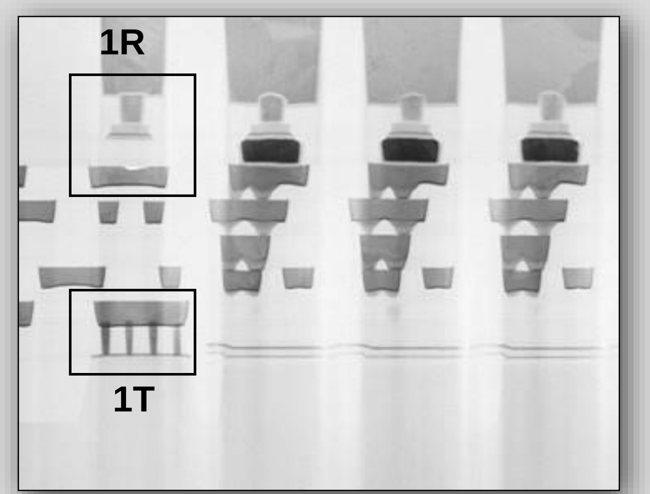
Why Emerging Resistive Memories?

High dense on-chip memory

DRAM access is at least **1500x** more costly than a MAC operation in NN accelerators

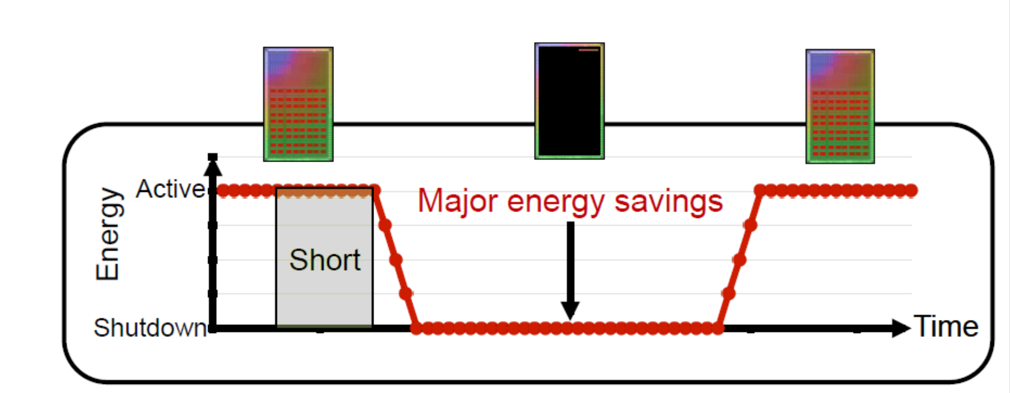
[F. Tu, et al., 2018 ACM/IEEE]





L. Grenouillet et al., 2021

Zero stand-by power thanks to non-volatility



10x better energy efficiency than embedded flash thanks to resistive memories

T. Wu et al., 2019

Emerging Non-Volatile Memories

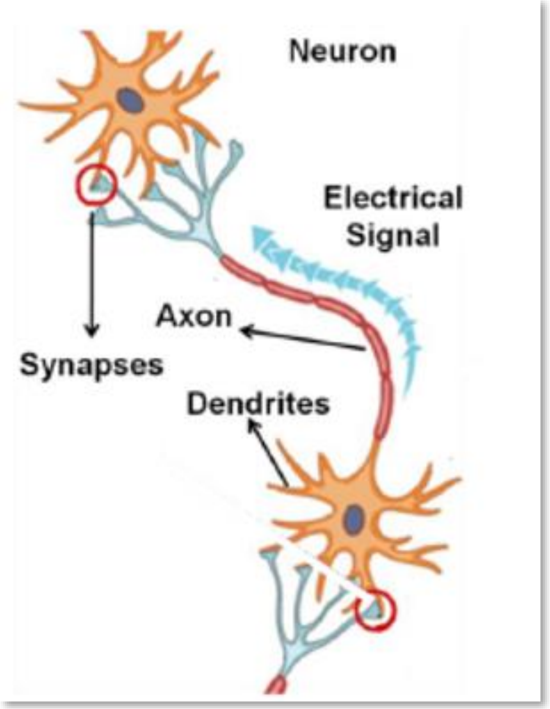
	NOR FLASH	MRAM	PCRAM	OxRAM	FeRAM (PZT)	FeRAM (HfO_2)
Programming power	~200pJ/bit	~20pJ/bit	~300pJ/bit	~100pJ/bit	**~10fJ/bit**	**~10fJ/bit**
Write speed	20 µs	**20 ns**	**10-100 ns**	**10-100 ns**	<100ns	**14ns @ 2.5V (SONY)** **4ns @ 4.8V (LETI)**
Endurance	10^5 - 10^6	**10^6-10^{15}**	10^8	10^5 – 10^6 on 16 kbit	**> 10^{15}**	**> 10^{11} single device** **10^6 – 10^7 on 16 kbit**
Retention	> 125°C	85°C - 165 °C	**165°C**	> 150°C	125°C	125°C
Extra masks	Very high (>10)	Limited (3-5)	Limited (3-5)	**Low (2)**	**Low (2)**	**Low (2)**
Process flow	Complex	Medium	Medium	**Simple**	**Simple**	**Simple**
Multi-Level Cell	Yes	No	**Yes**	**Yes**	No	No
Scalability	Bad	Medium	**High**	**High**	Medium	**Poor (2D)** **High (3D)**

Power Reduction by 10000!

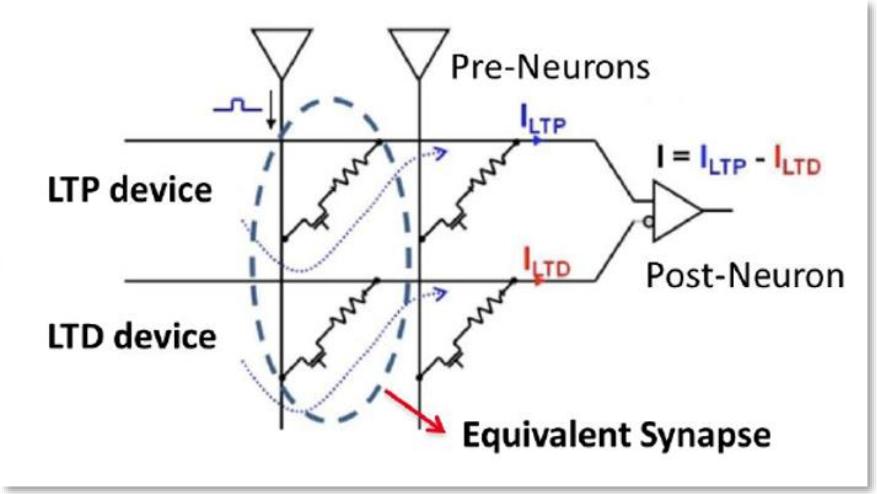
Memory activity focus on embedded NVM for NOR flash replacement

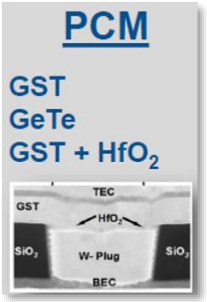
Neuromorphic based RRAM circuit

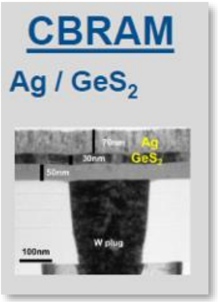
M. Suri et al, IEDM 2011.

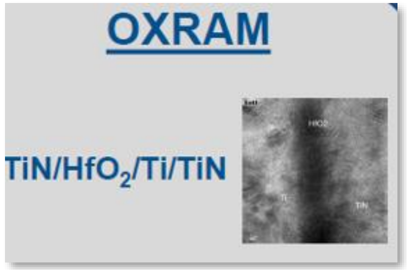


2 PCRAM Example:

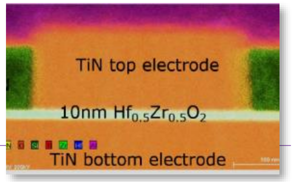




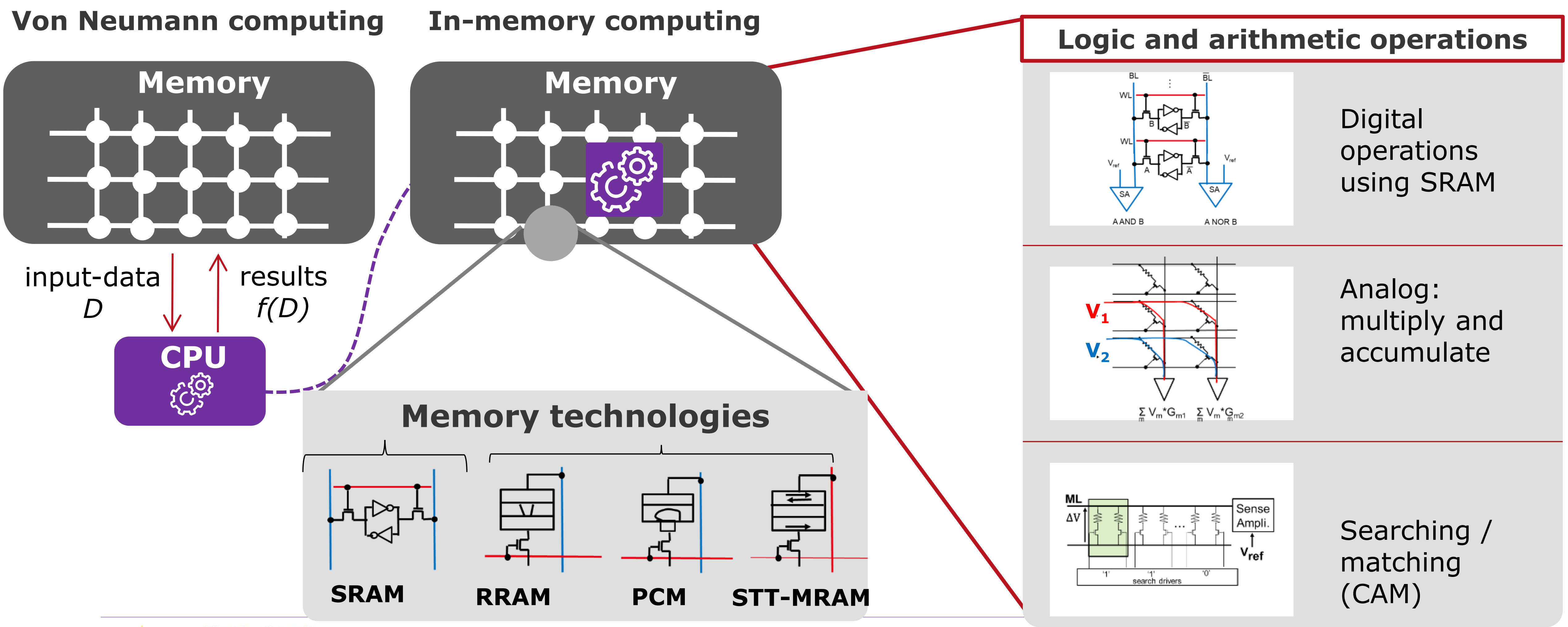




FeRAM

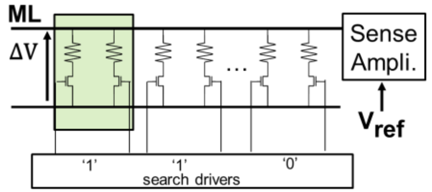


In memory computing

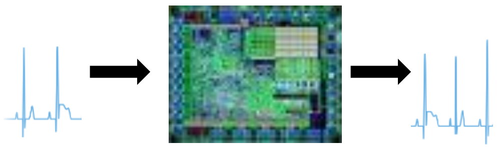


Near- & in-memory computing

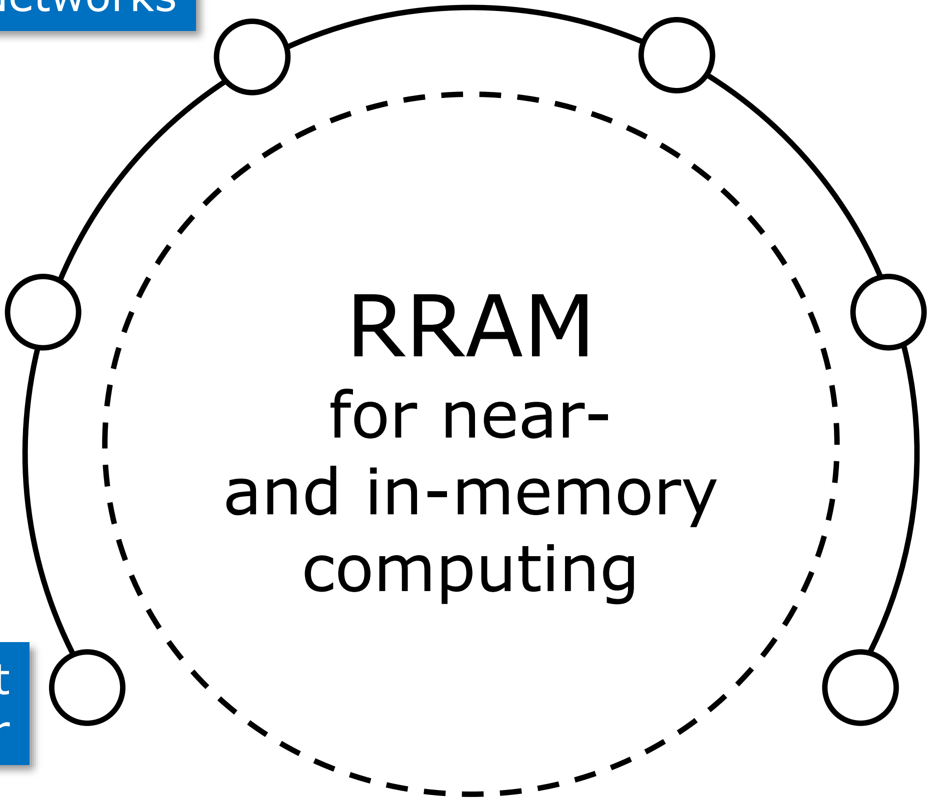
Bayesian
Neural Networks



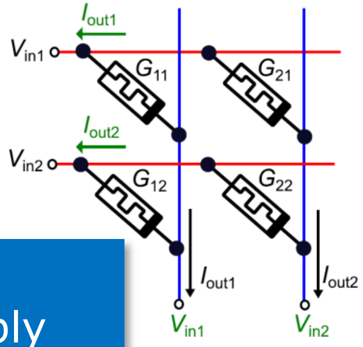
Search/match
operations (CAM)



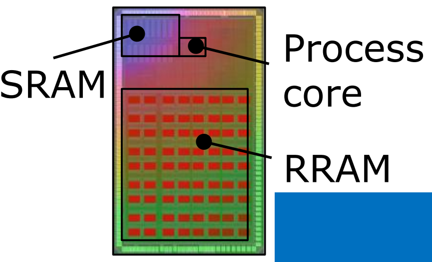
Spiking
Neural Network



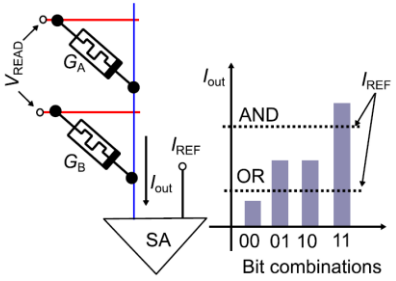
$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$



Deep learning,
matrix-vector multiply

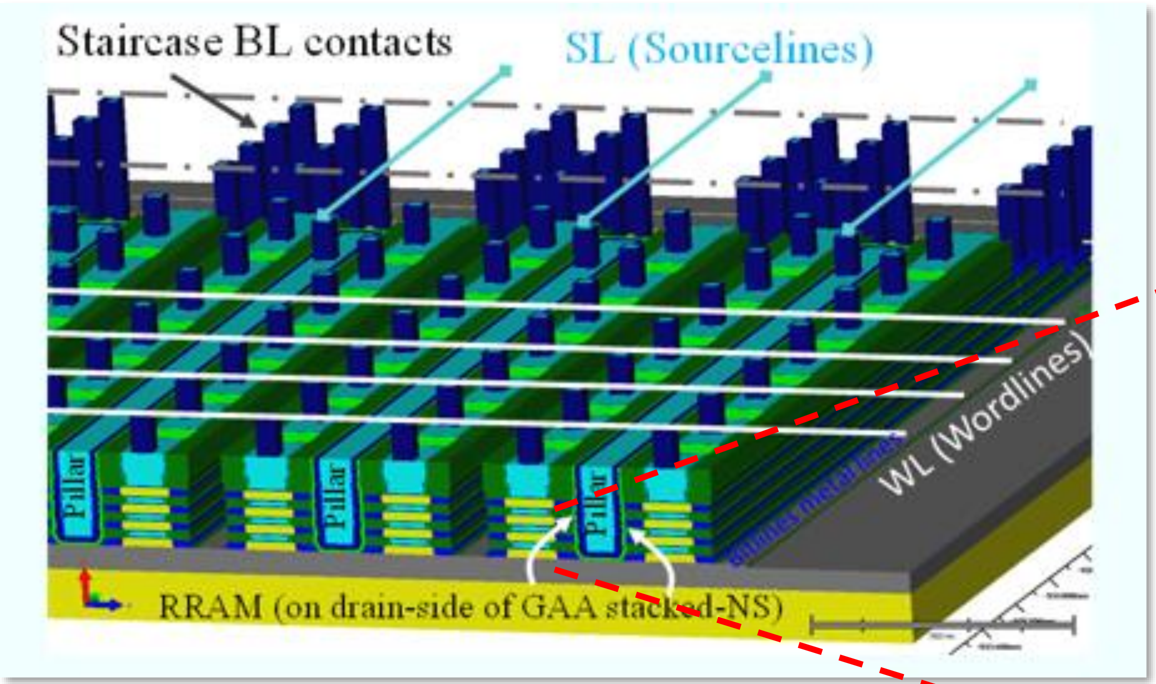


Energy-efficient
NN accelerator



In-memory
logic

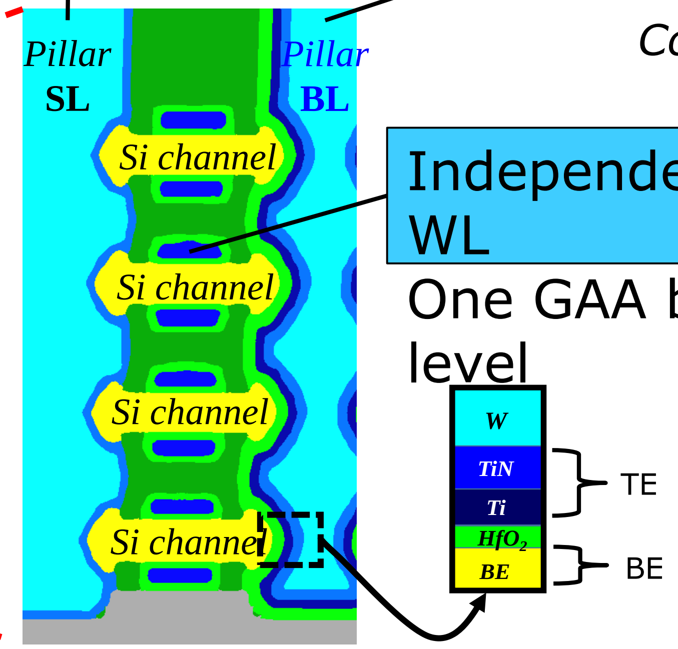
Towards In Memory Computing



Common SL for all GAA stacked-NS

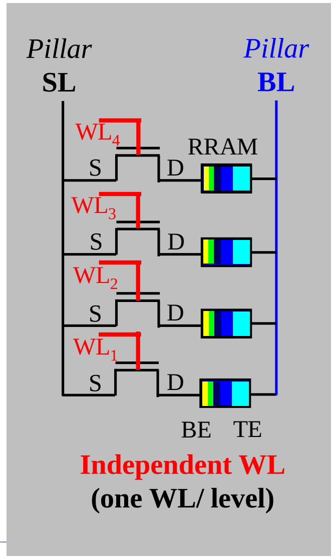
Common BL for all GAA stacked-NS

Courtesy of S. Barraud,



Independent WL

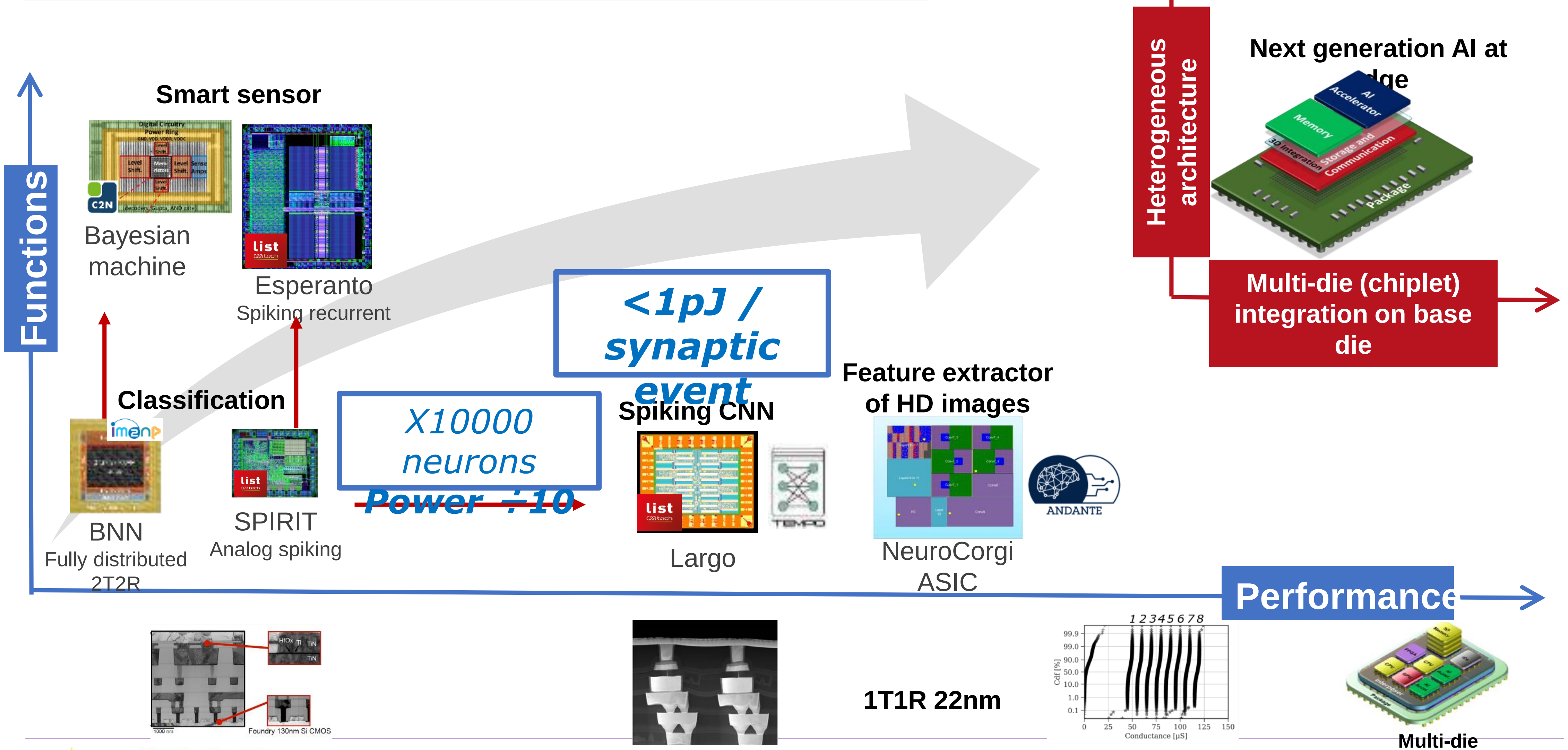
One GAA by level



Enabler for Hyperdimensional computing

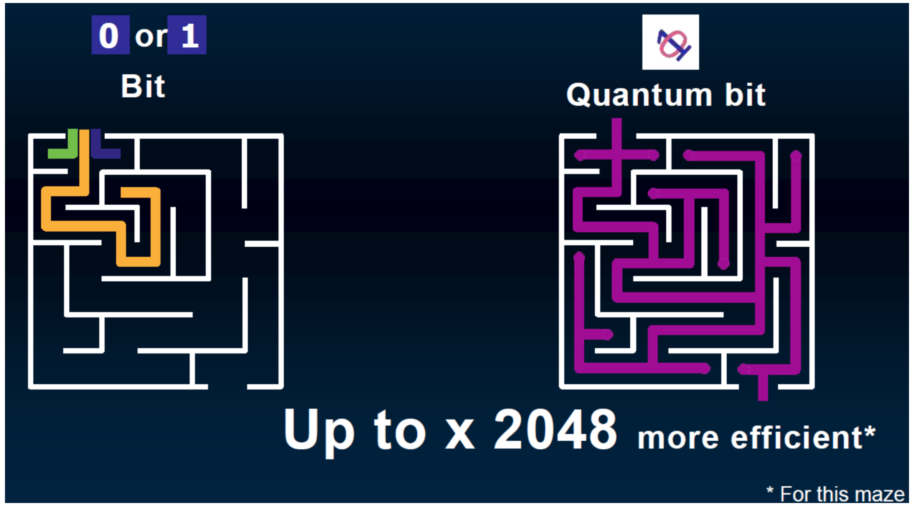
A possible Edge IA Roadmap

Courtesy of E. Vianello, Leti



From bits to q-bits

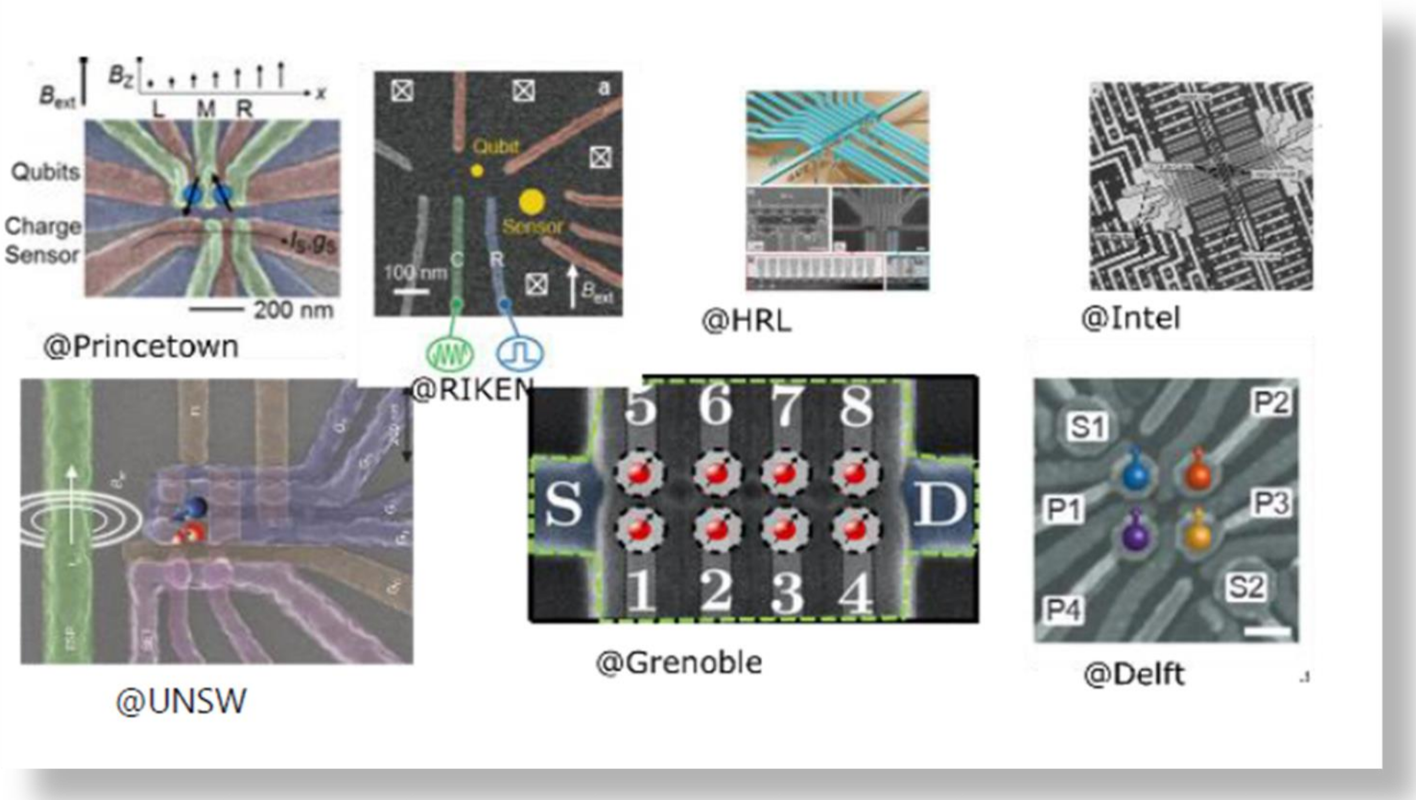
Quantum Physics to compute

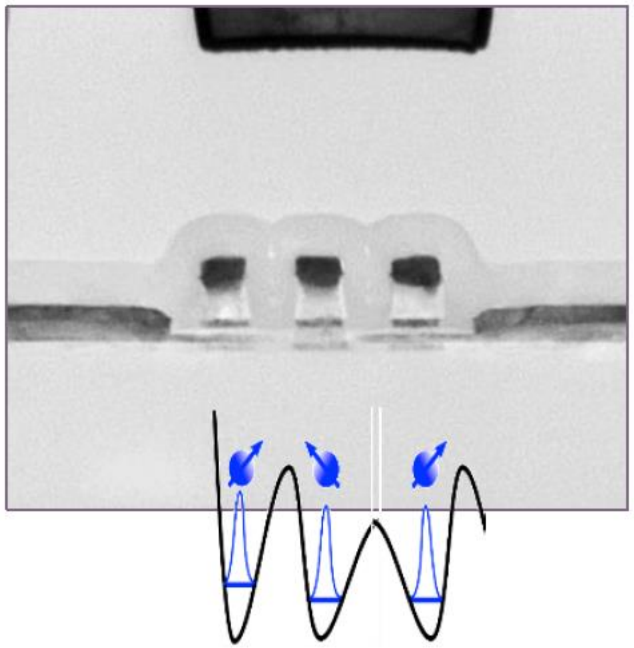


	Superconductor	Si spin qubit	Trapped ion	Photon
Size*	(100µm)²	(100nm)²	(1mm)²	~(100µm)²
1qubit fidelity	99.96%	99.93%	99.98%	
2qubit fidelity	~99.3%	>99%	99.9%	50% (measurement) 98% (gates)
Speed**	12-400 ns	~1 µs	100 µs	1 ms
Variability	3%	0.1%-0.5%	0.01%	0.5%
T° of operation	15mK	1K	10K	4K/10K
Entangled qubits	433 (IBM)	3 (TU) (6 - QuTech)	32 (IonQ)	70 (Pan-China)

From bits to q-bits

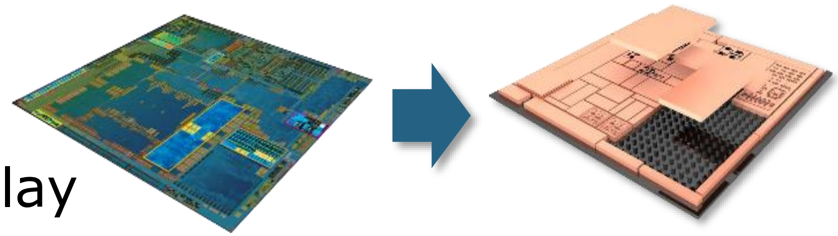
- Quantum Physics to compute



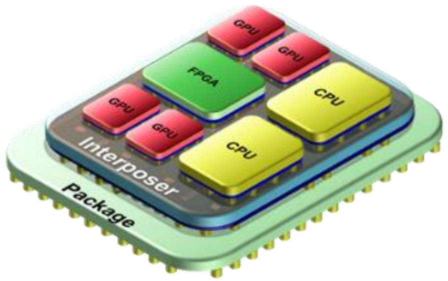


Chiplet approach: Heterogenous IC design

- Interposer & chiplets
  Interconnects performance → R.C delay
  Exceeding latency & bandwidth limits
  Cost/form factor advantages

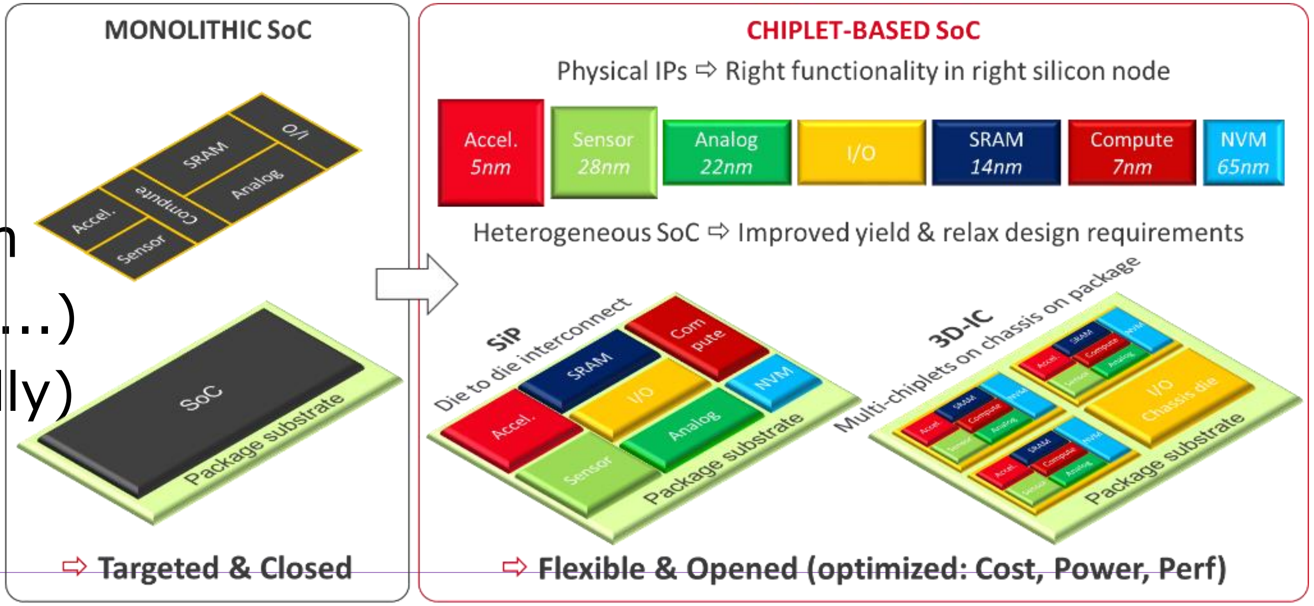


The end of "all for the SoC" paradigm (image from DARPA)

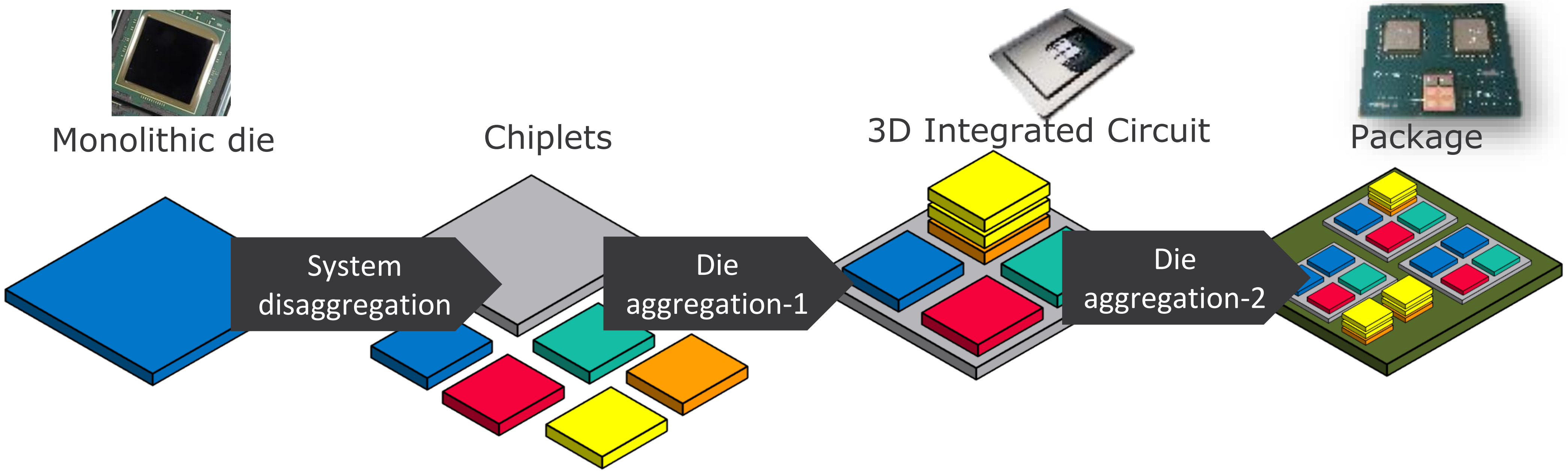


Chiplet topology on interposer

- Appropriate partitioning
- Heterogeneous IC design
  Optimized technology for each function
  Specialization by app.: CPU, GPU, AI (...)
  Standardization (coming soon, hopefully)



Chiplets: the new IC design paradigm



Architecture

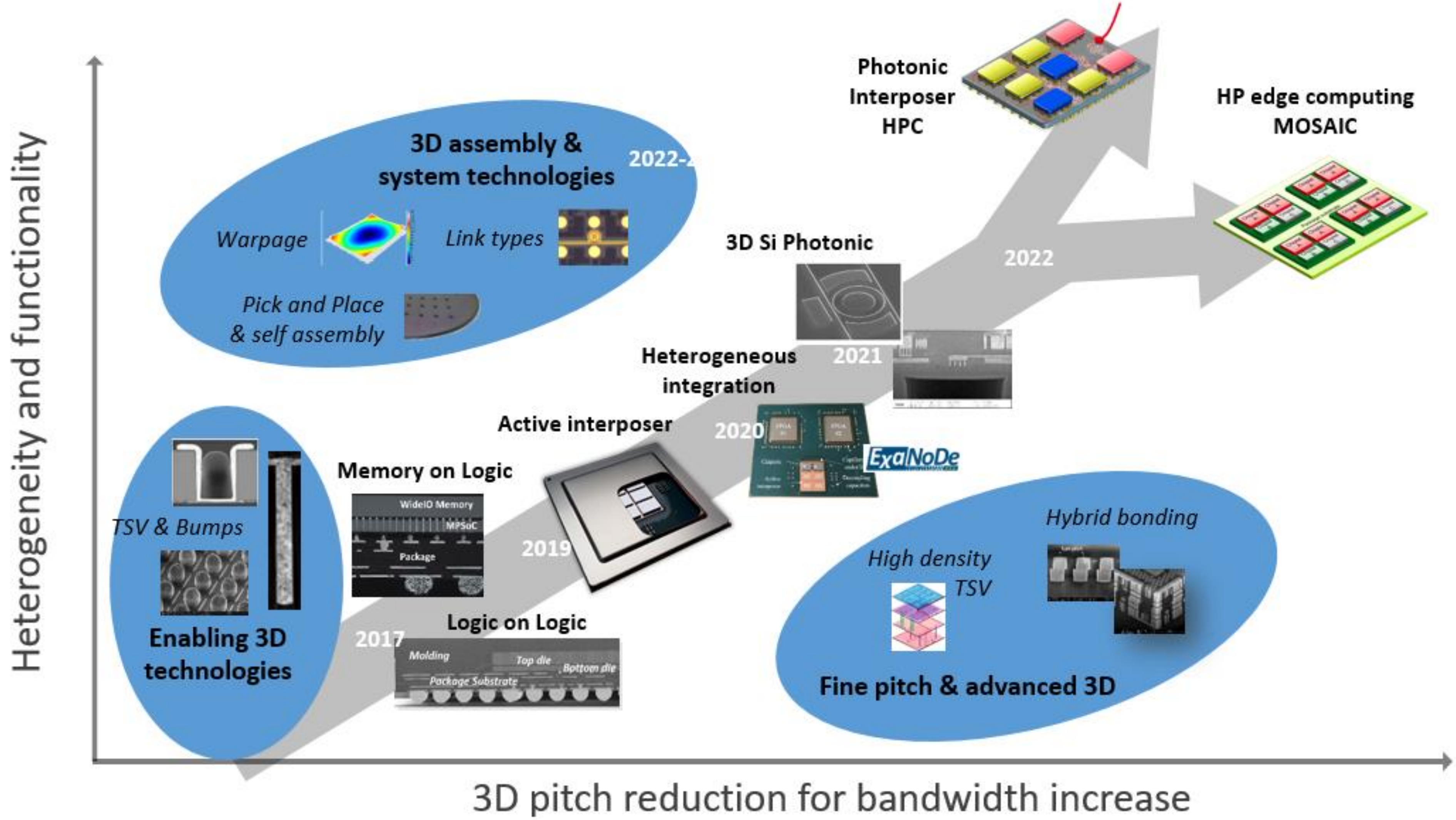
Design
Design flow
Wafer-level integration

Design
Design flow
Package-level integration

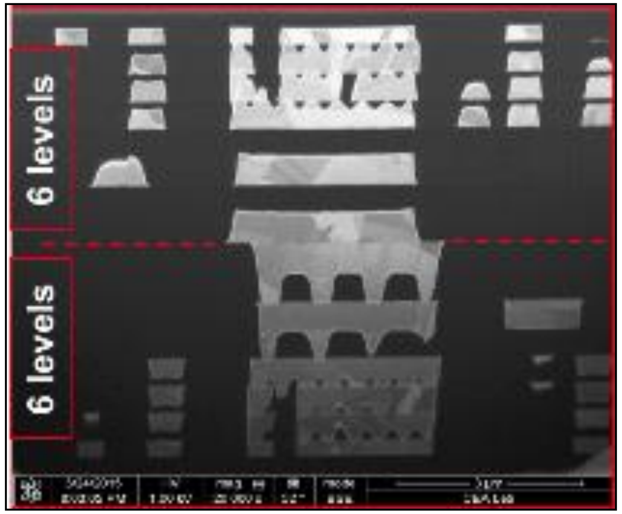
System Technology Co-optimization

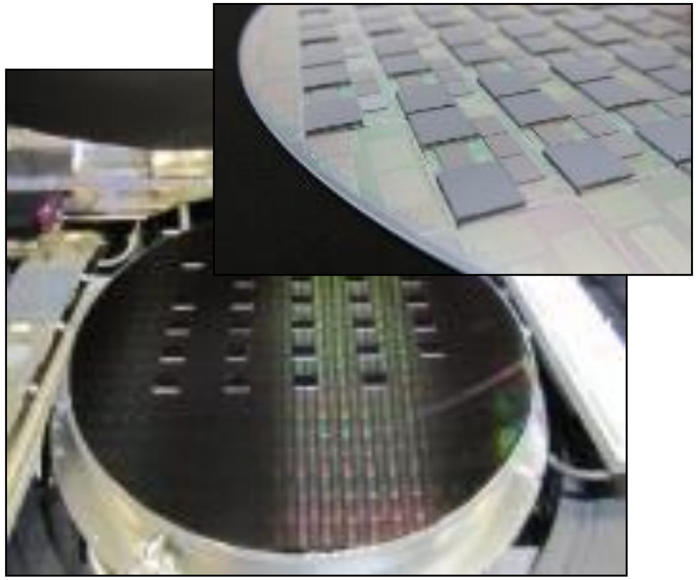
Up to 100x gain on Power Efficiency with 3D

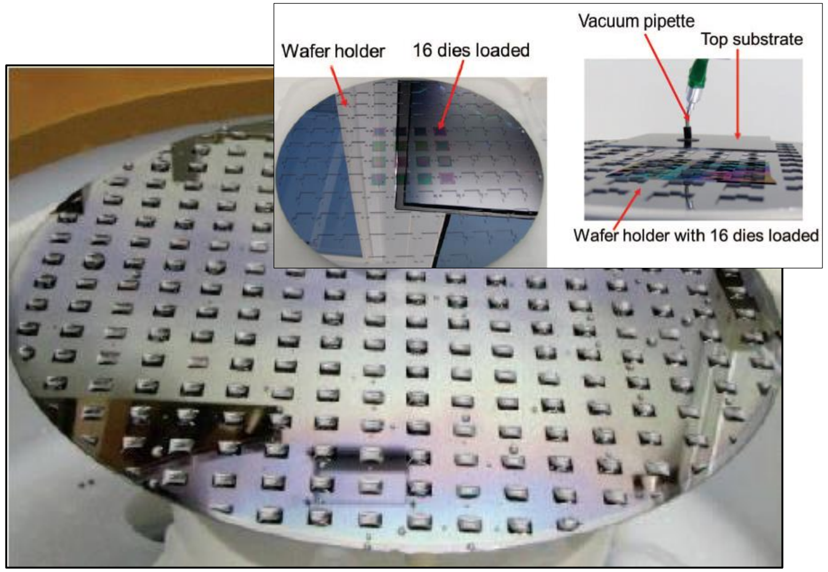
3D Tool Box for Chiplet integration



Hybrid Bonding Solutions





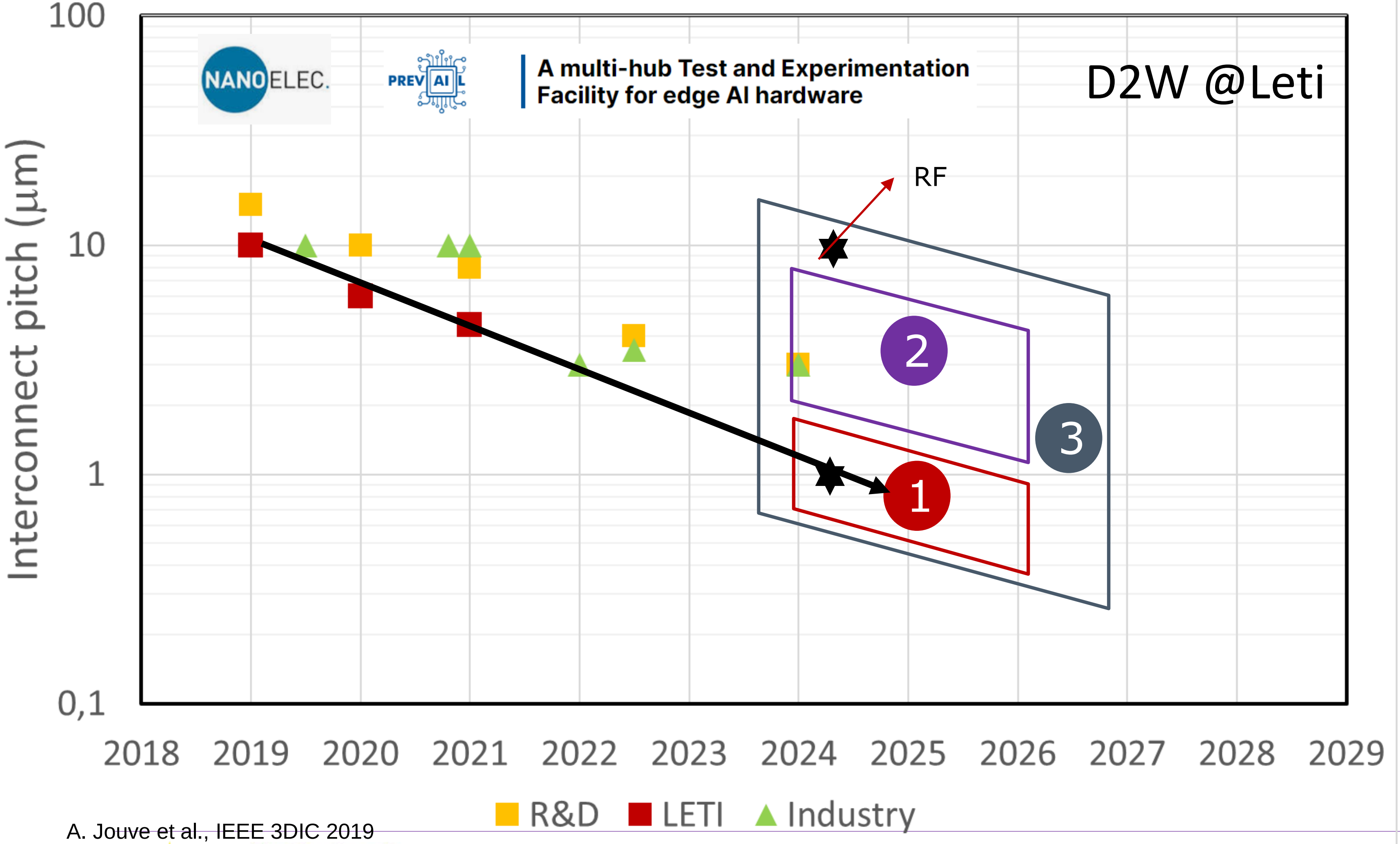


- Direct bonding of metal and dielectric
- Down to 1 micron pitch interconnects

- Wafer-to-wafer (W2W) or Die-to-wafer (D2W) technologies
- High heterogeneity allowed by D2W

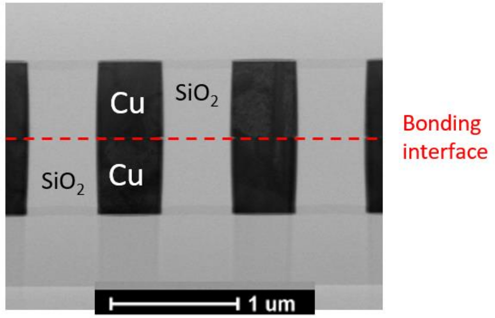
- Collective D2W approaches
- Self-assembly for high precision & high throughput

Hybrid bonding pitch roadmap



A. Jouve et al., IEEE 3DIC 2019
E. Bourjot et al., ESTC 2022

Pitch down to 1µm



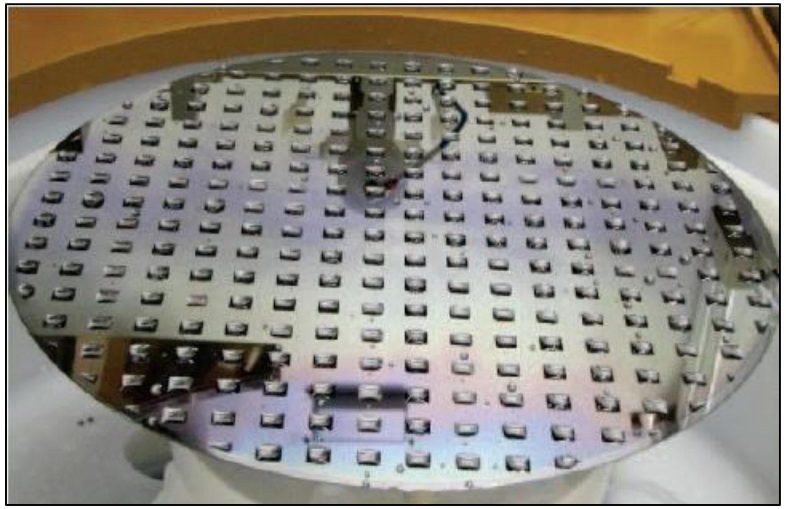
TEM on W2W bonding at LETI

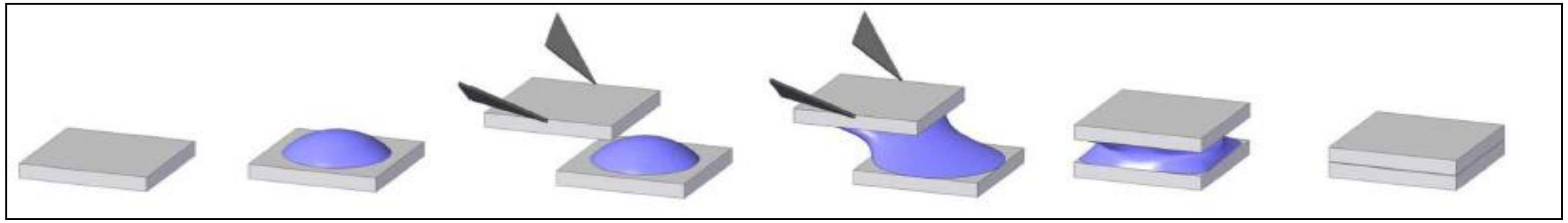
1 **Pitch reduction** (towards < 1µm)
Interconnect / bandwidth density
Towards 3DIC

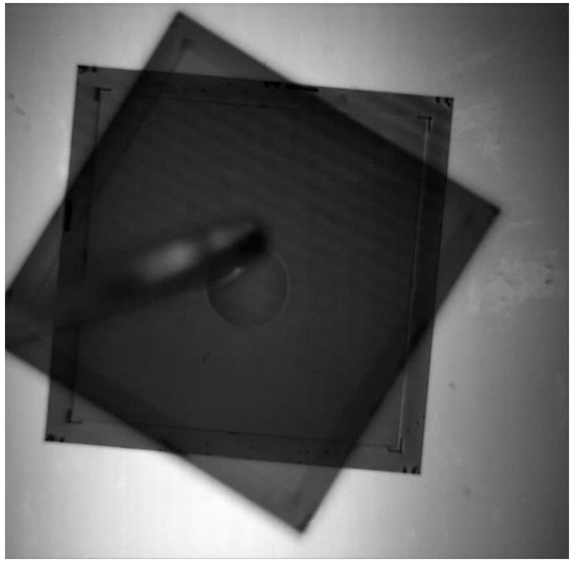
2 **New capabilities** for heterogeneity
- Temp. reduction (150-300° C)
- Self-assembly (precision, throughput)
- New materials (III-V, supraconductors)

3 **Architectures & Demonstrators**
- Multistacking with TSV
- Smart Imagers & displays
- Edge AI, based on chiplets
- RF mmW, incl. III-V

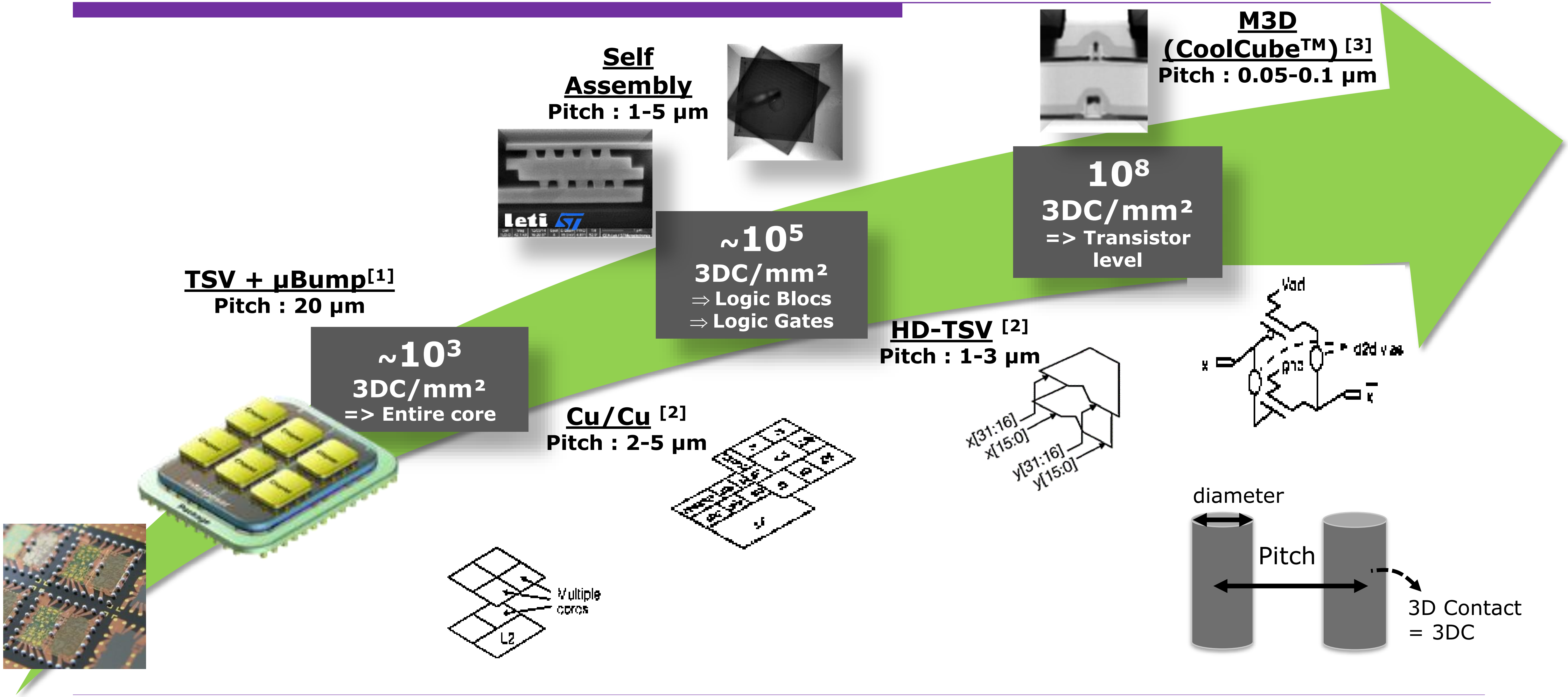
Self Assembly Hybrid Bonding





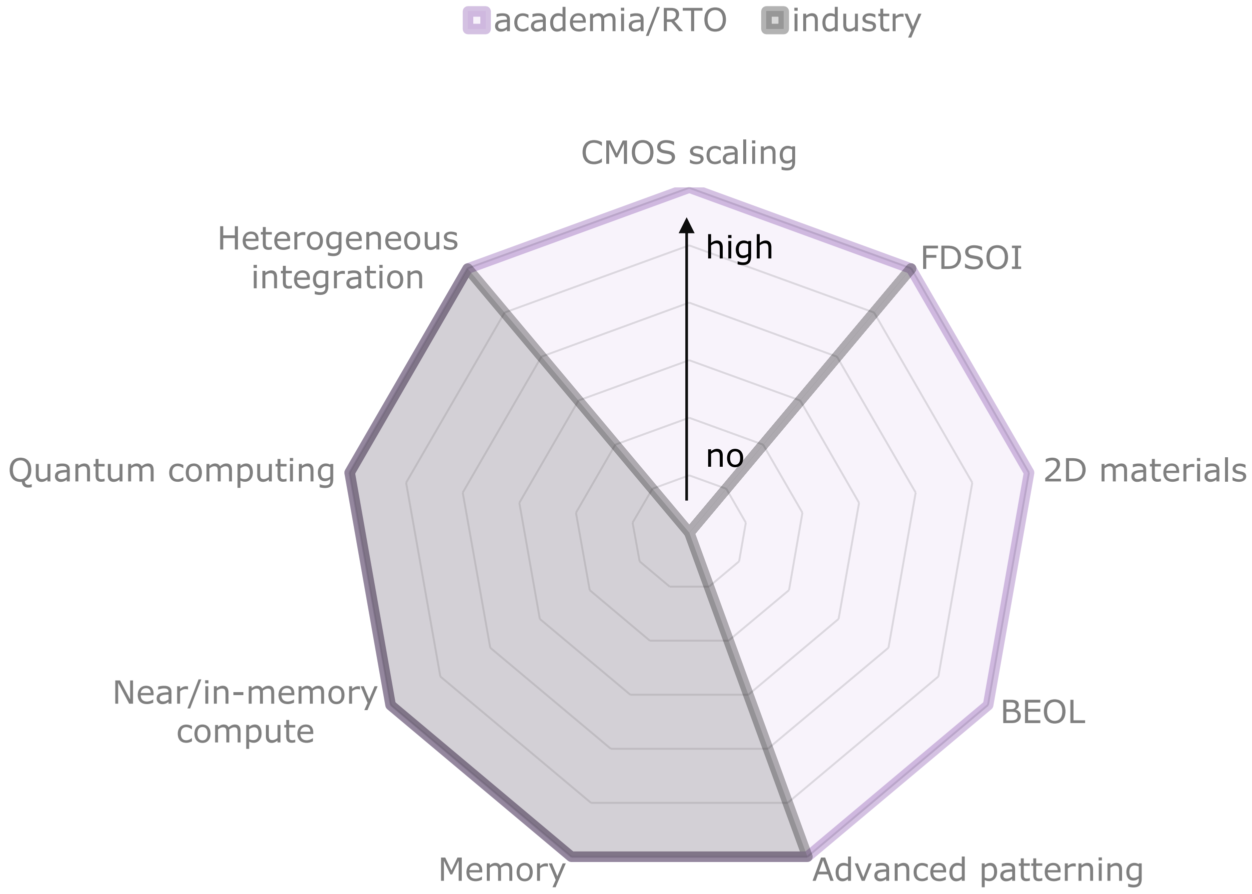


3D: from packaging to monolithic



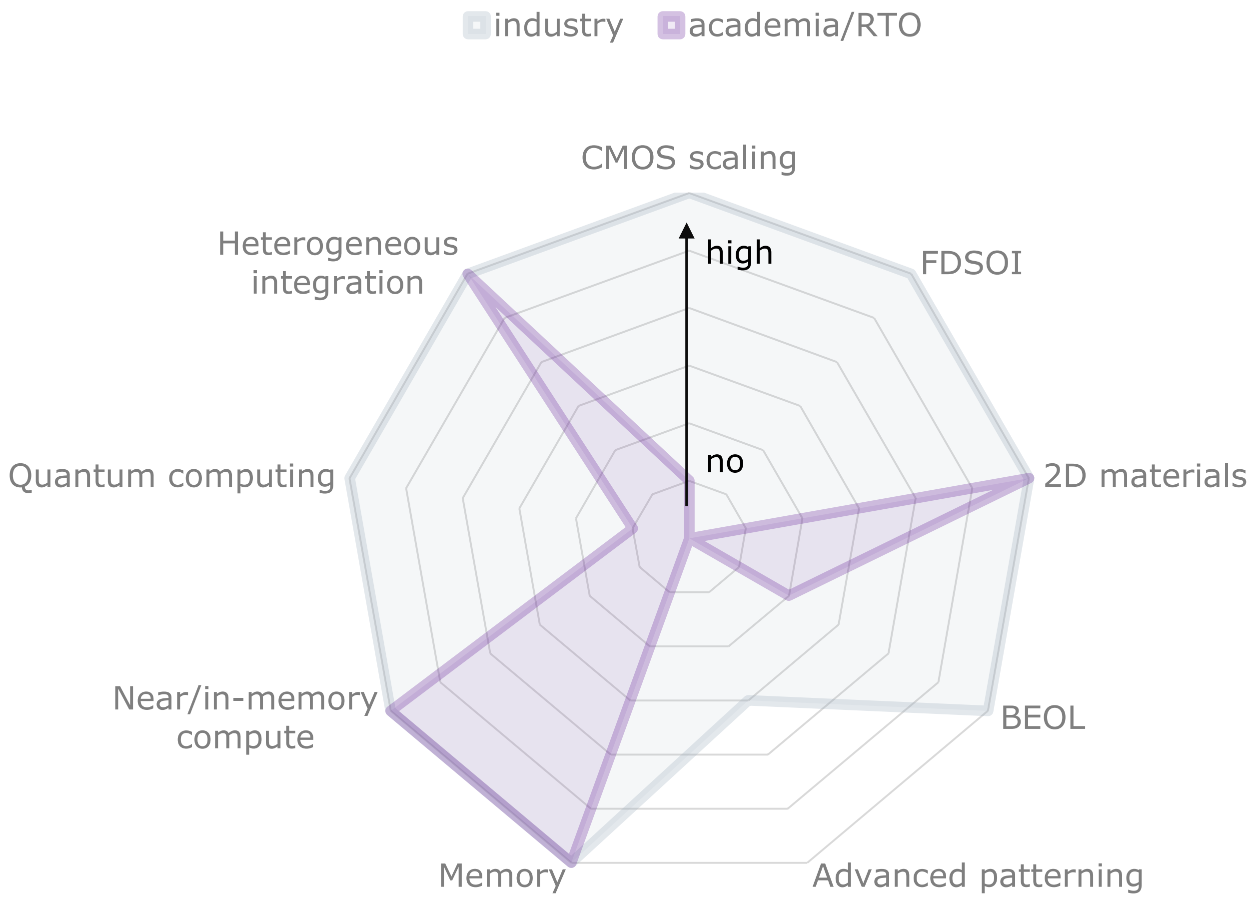
[1] Cheramy, S., et al. "Advanced Silicon Interposer", C2MI Workshop, 2015
[2] Patti, B., "Implementing 2.5D and 3D Devices", In AIDA workshop in Roma, 2013
[3] Batude, P., et al. "3DVLSI with CoolCube process: An alternative path to scaling," VLSI technology symposium 2015

EU and non-EU actors - EU



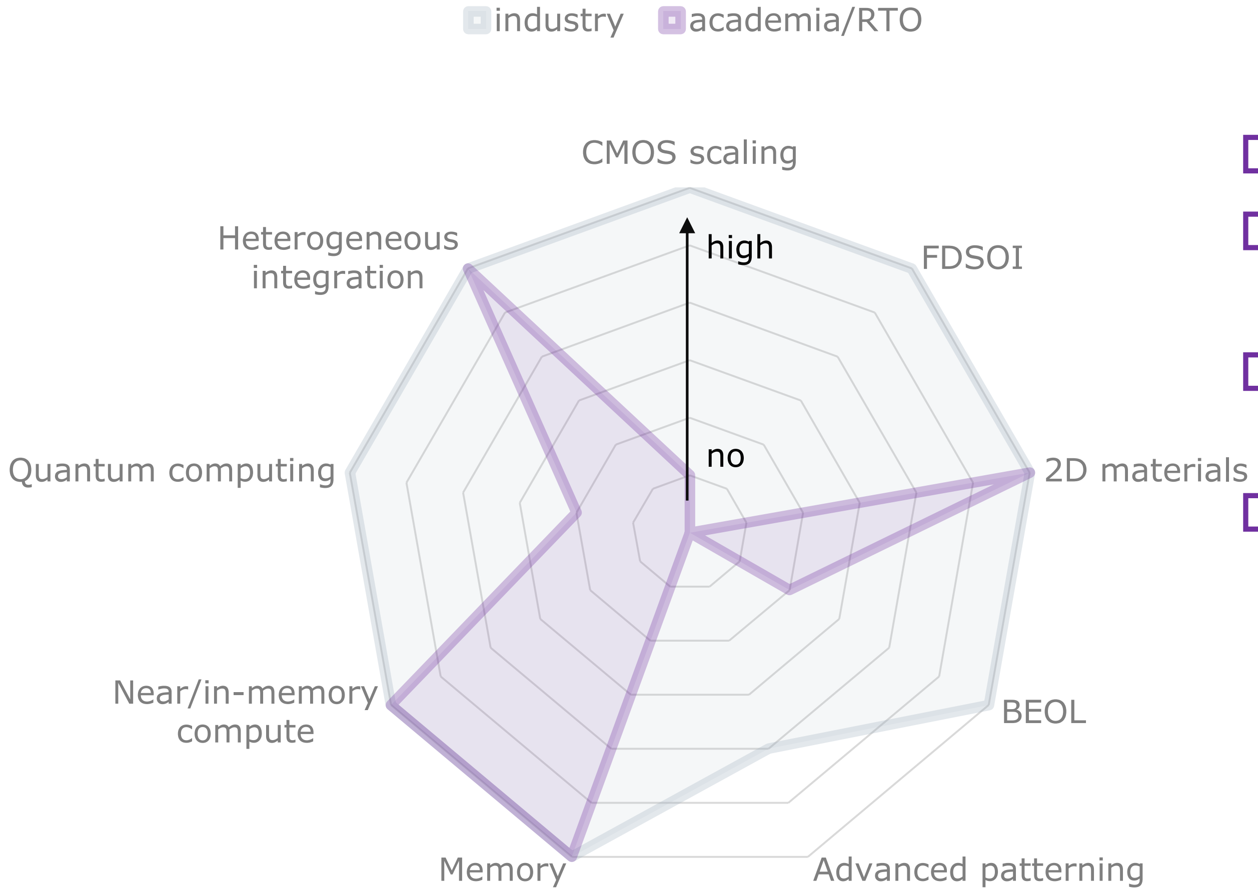
- R&D very strong in all areas of compute
- Unique strong position in EUV lithography
- In general, industrial EU players lacking to take up R&D

EU and non-EU actors - US



- Strong industrial activity in most areas of compute
- Weaker academic activity on traditional logic scaling
- Strong R&D in new materials, heterogeneous integration and memory

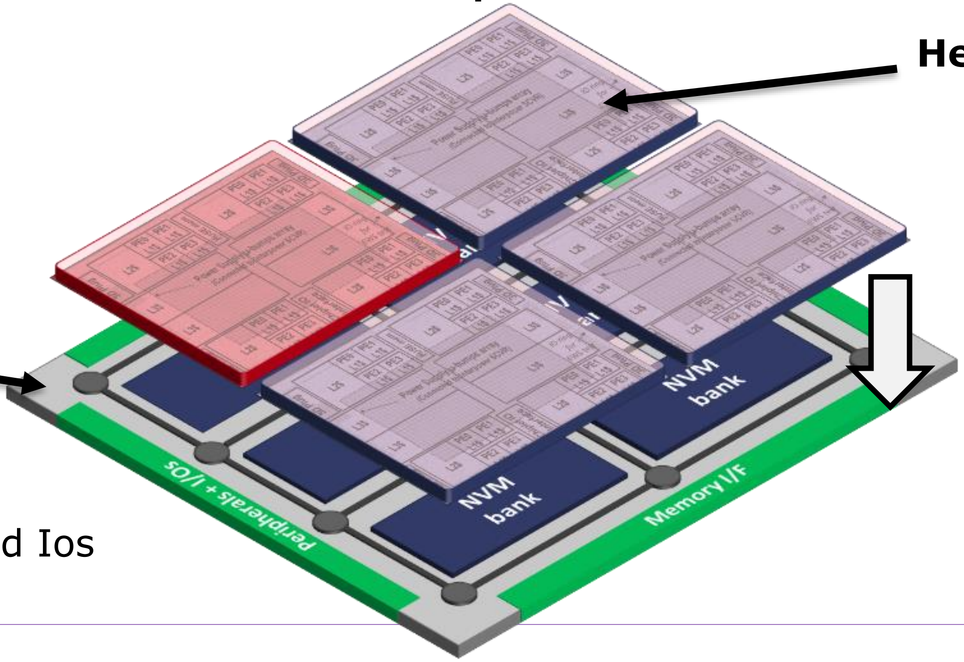
EU and non-EU actors - Asia



- Very similar to US
- Strong industrial activity in most areas of compute
- Weaker academic activity on traditional logic scaling
- Strong R&D in new materials, heterogeneous integration and memory

Summary

- ☐ eNVM: Key feature to store data close to the compute Engine
- ☐ eNVM: An enabler for Neuromorphic designs!
- ☐ 3D technologies: towards Chiplets and Heterogeneous integration



Heterogeneous compute chiplets:

- Advanced node (7nm, 5nm)
- Heterogeneous (size, pitch, node)
- Full digital compute chiplet
- Optional: TSV HD for memory cube
- Face down

Base die:

- Mature node (FD28, GF22FDX,…)
- Face up
- TSV for power delivery and Ios (preferred mid process)

3D assembly:

- Die-to-Wafer
- Hybrid Bonding
- Face-to-Face