ICOS International Cooperation On Semiconductors

Challenges in Advanced Computing and Functionalities International Cooperation on Semiconductors

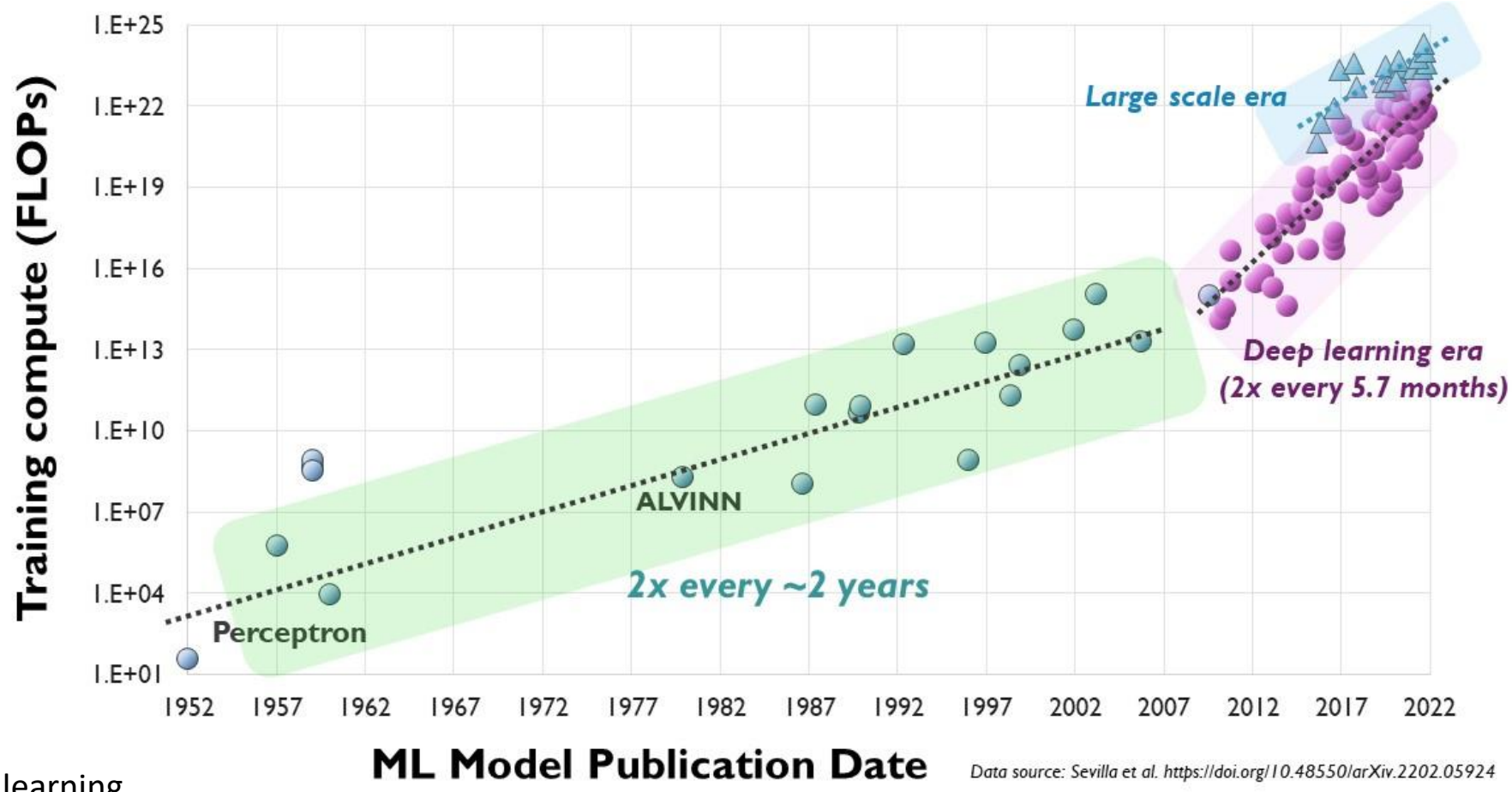
Future technologies in Advanced Computation

Nadine Collaert (imec) and Olivier Faynot (CEA-LETI)

Outline

- Introduction: trends and challenges
- Computing roadmap:
  - CMOS device architecture
  - 2D materials for FEOL
  - New materials for BEOL
  - Lithography
  - CMOS 2.0
- Memory technologies
- Beyond Von Neumann disruptive approaches:
  - Near or in-memory computing
  - Quantum computing
- Heterogeneous integration: from chiplets to functional backside
- Analysis: EU and non-EU actors
- Conclusions

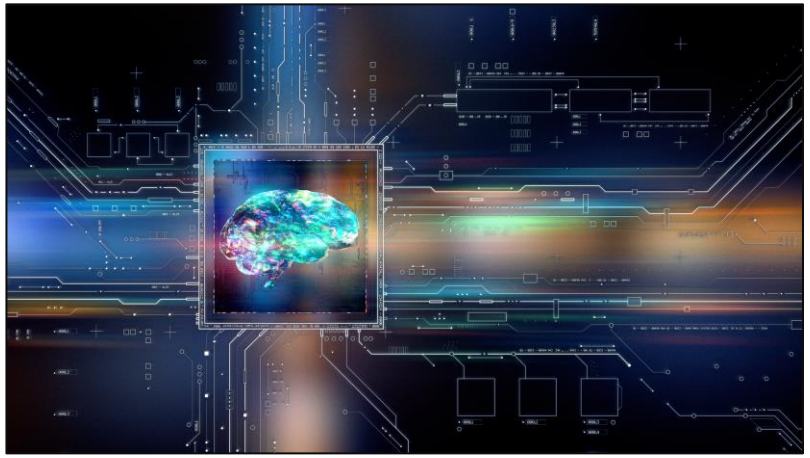
Compute needs for ML continue to grow



ML=machine learning

Diversity of Applications and Workloads

GPUs for Training



High throughput parallel compute
Very high memory bandwidth
Very high GPU-GPU bandwidth

AR/VR

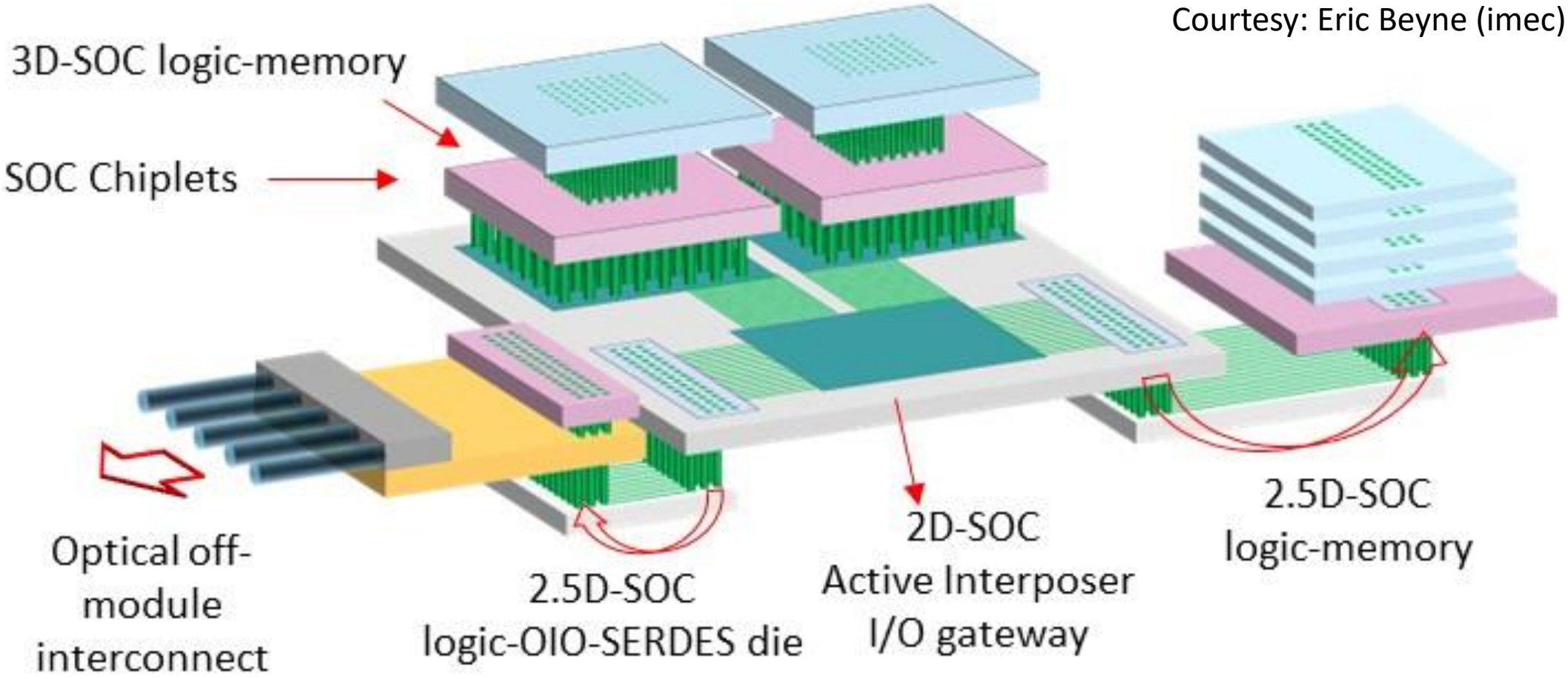
Low power
Ultra low latency
High memory bandwidth
Small form factor

Autonomous driving

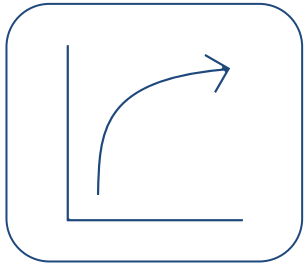
Multi-sensor fusion
Distributed real-time computation
Reliable and explainable AI

Future systems are heterogeneous

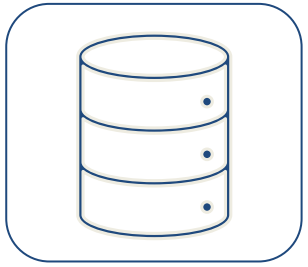
Courtesy: Eric Beyne (imec)



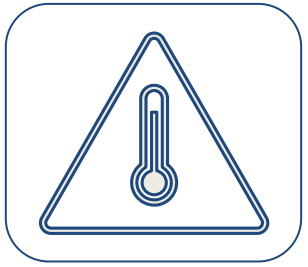
Challenges For Future Compute Systems



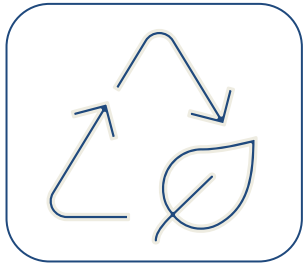
CMOS and DRAM PPAC



Memory Wall



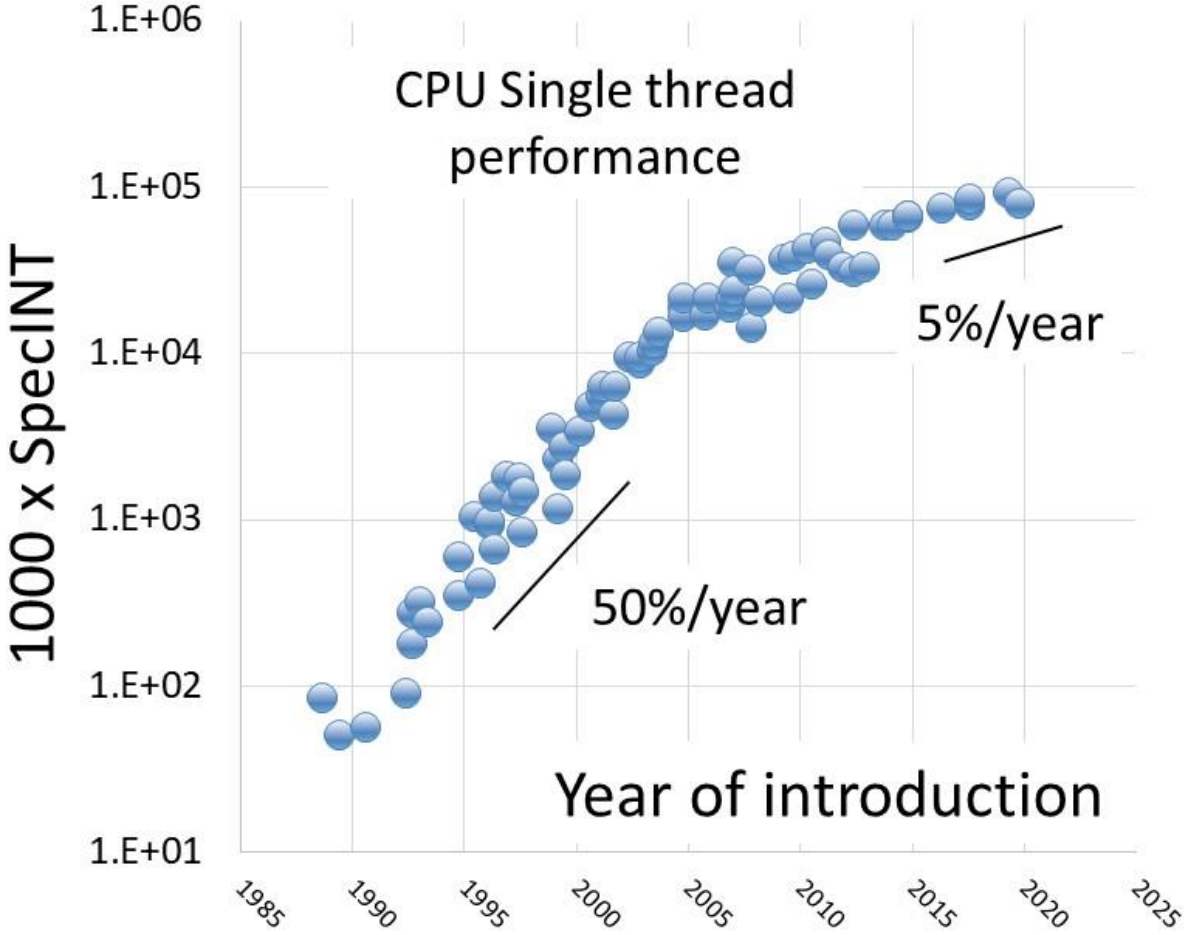
Power Wall

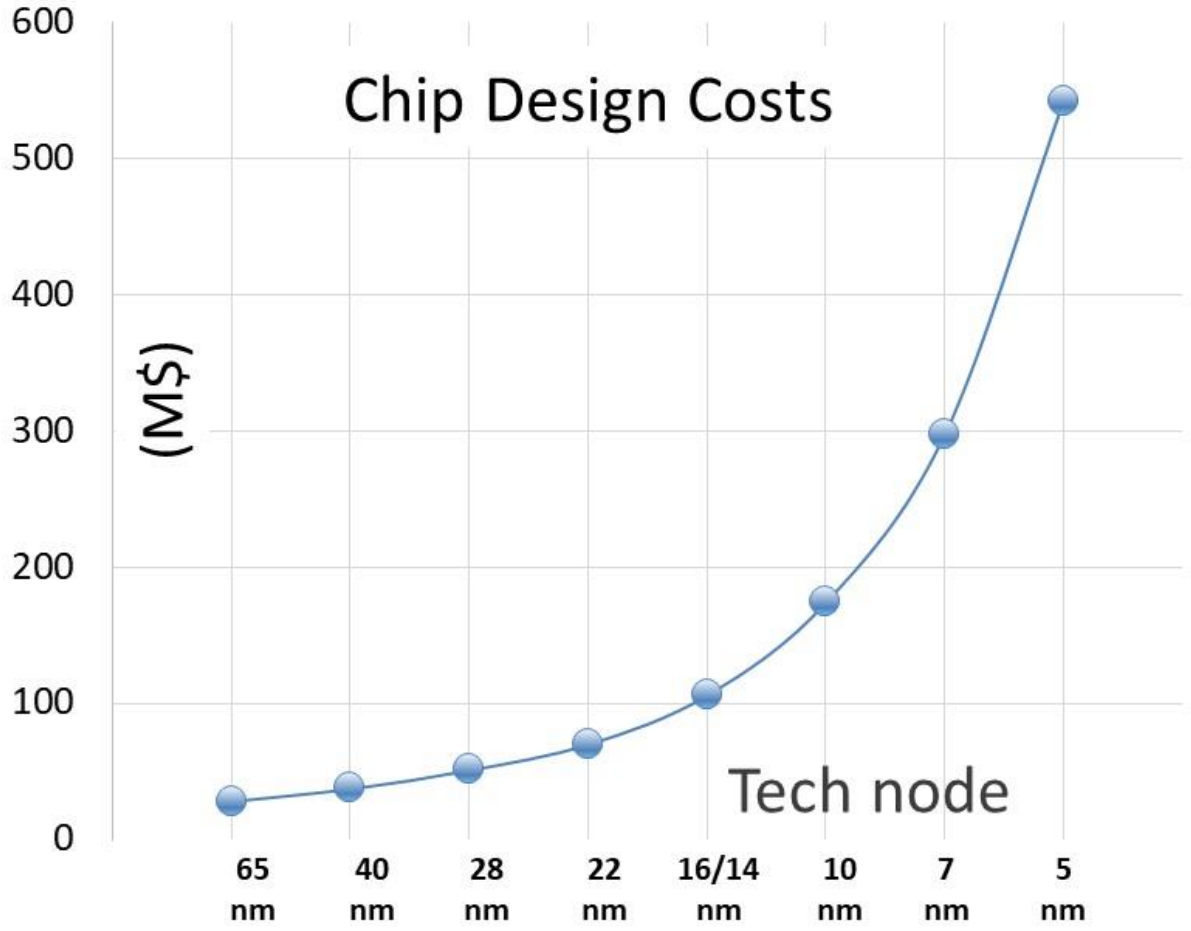


Sustainable Manufacturing

PPAC=Power-Performance-Area-Cost

Slowdown in System Performance and Increasing Costs

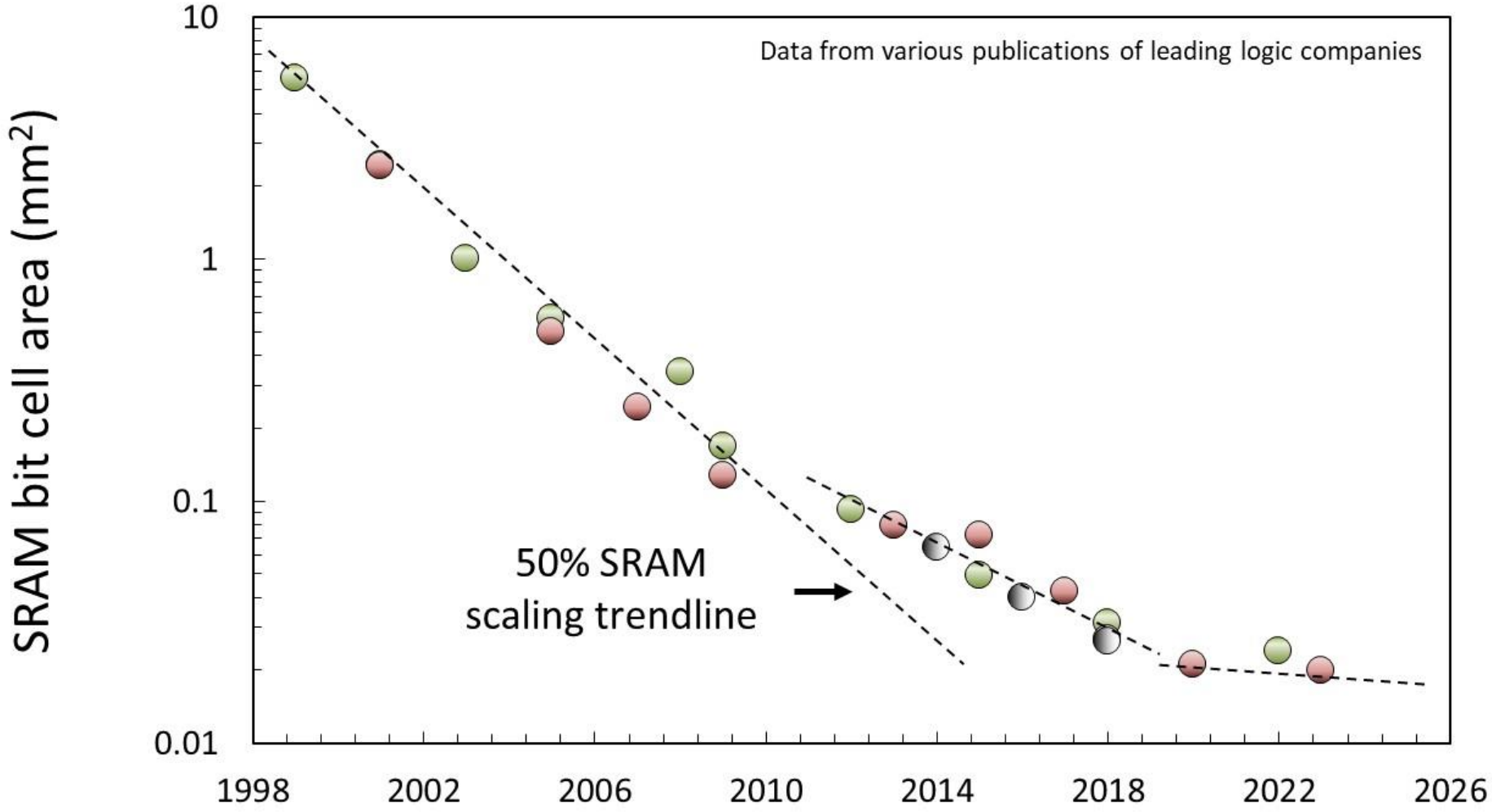




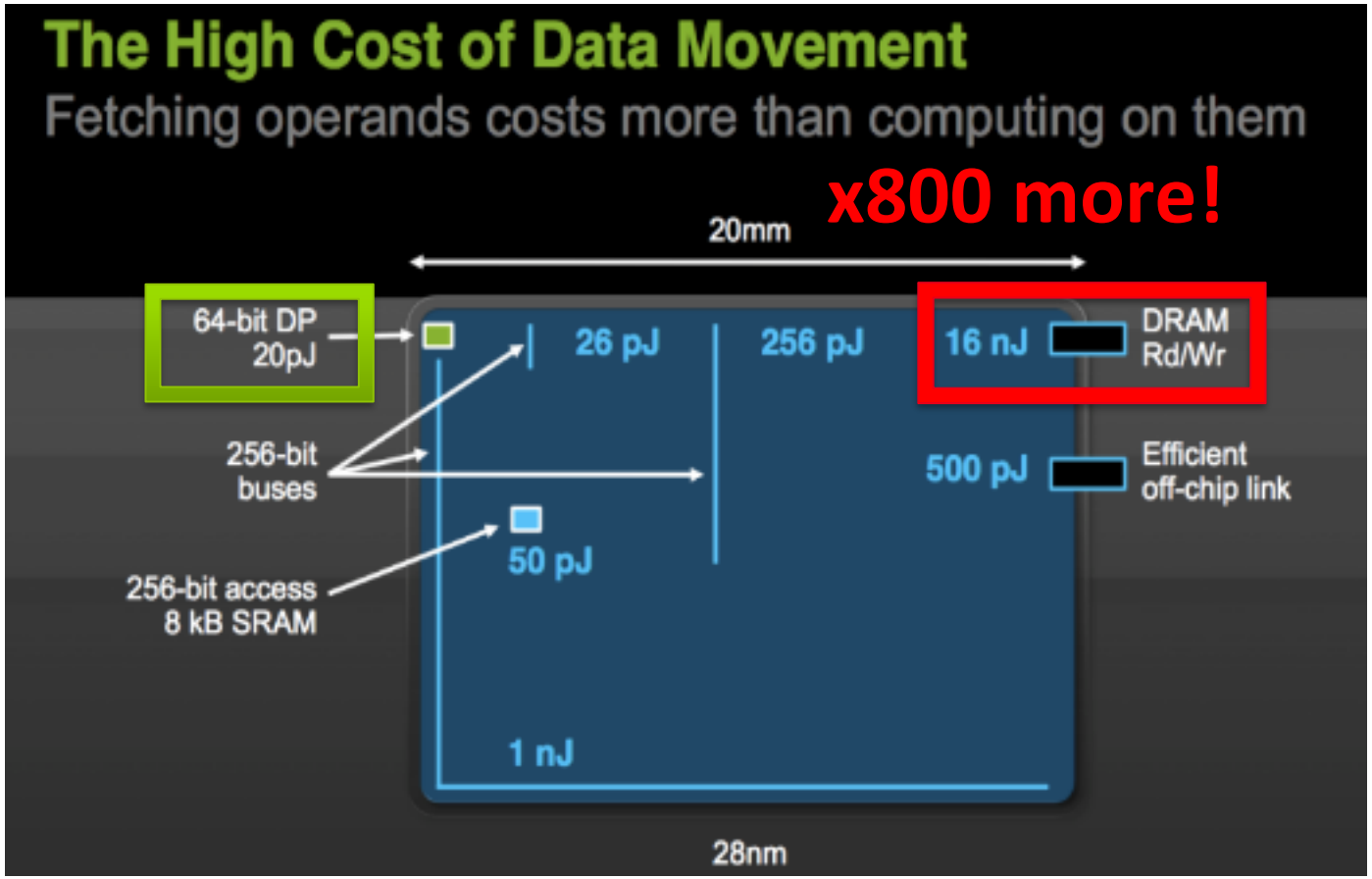
Based on original data plotted by M. Horowitz, F. Labonte, O. Shachan, K. Olukotun, L. Hammond, C. Batten. Additional data compiled by K. Rupp

Source: AI Chips and why they matter", S. Khan and A. Mann, 2020

SRAM scaling has slowed down

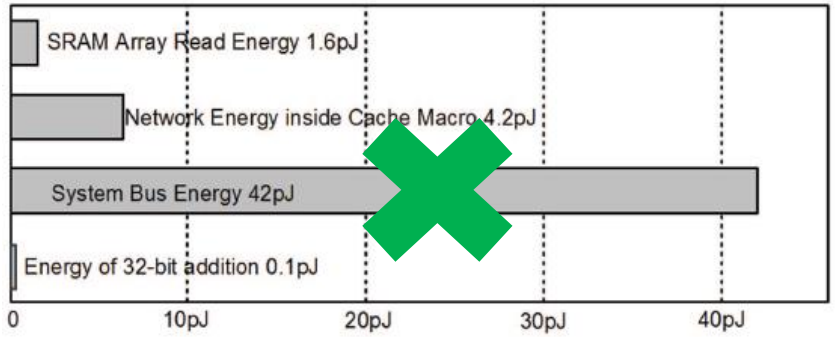


The cost of moving data



Bill Dally, "To ExaScale and Beyond", 2010

[J. Wang – ISSCC'19]

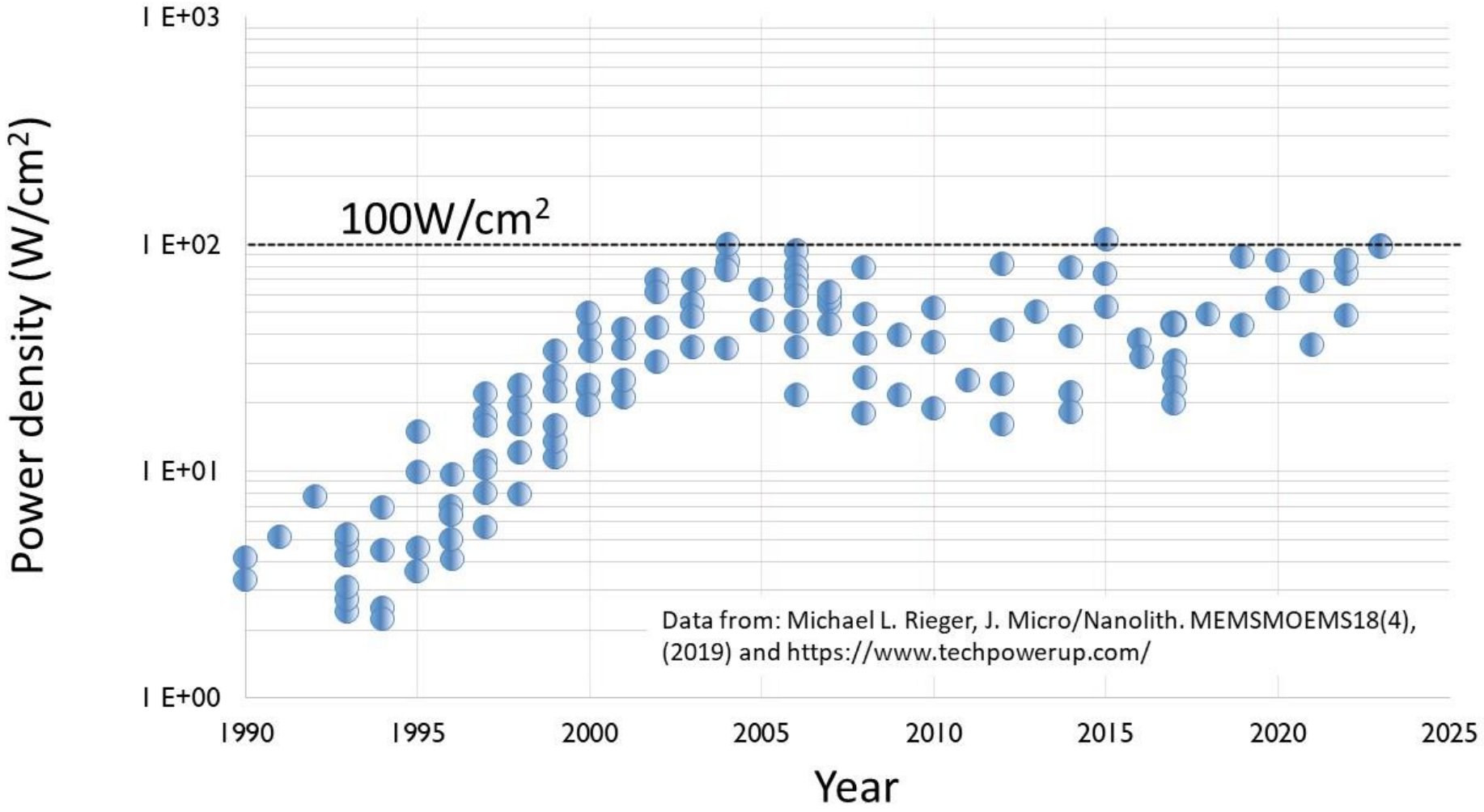


~90% of energy is in **data transfer**
→ IMC could lead **8x reduction**

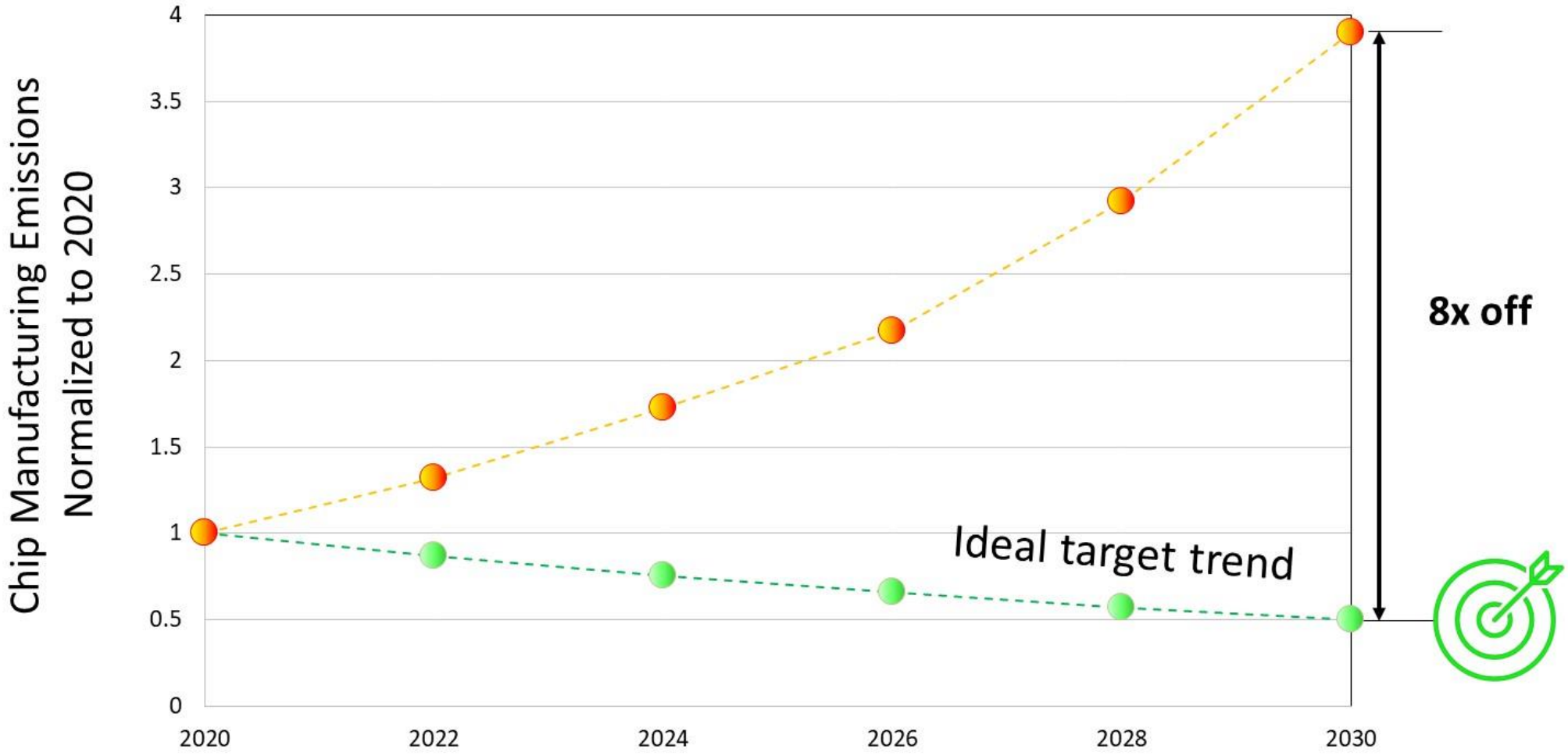
Operation energy is negligible

Memory access and control energies dominate

Chip Cooling Technologies Limit Power Density of Chips

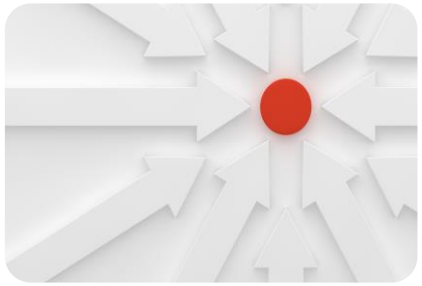


Carbon Emissions: “Do Nothing” Scenario



Constant electricity mix (0.49 kCO2eq/kWh), Abatement and GHG global warming potential according to IPCC assumed for the years 2020-2030. Volume technology mix from IBS “Foundry Market Trends and Strategic Implications” Vol 30, N 12 , Dec 2021. Logic nodes only. *imec.netzero emissions estimate of imec process nodes representative of foundry nodes.

How to Enable Sustainable Manufacturing



Adopt 50% emissions reduction this decade as target

Must-do for fabs: green power, best abatement

Research new materials: F-gas and PFAS alternatives

The required gain in energy efficiency

>1000x
by
2030

CMOS scaling

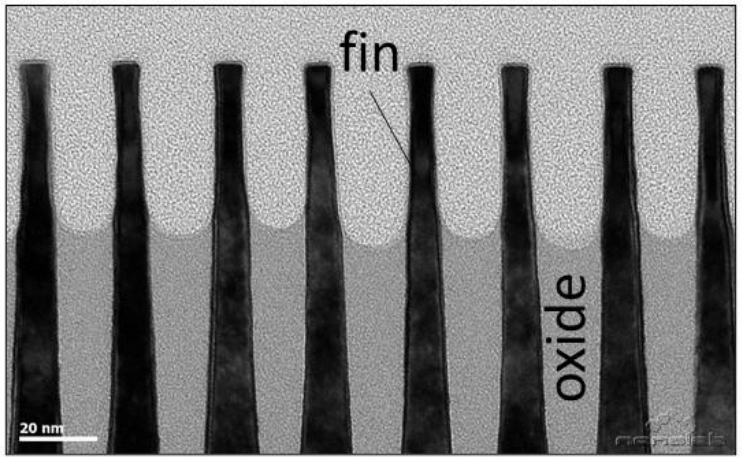
Memory technologies

Disruptive Computing

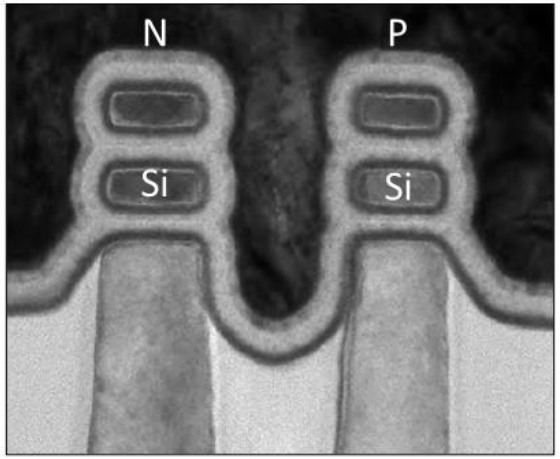
Chiplet & 3D System

New device architectures to extend scaling

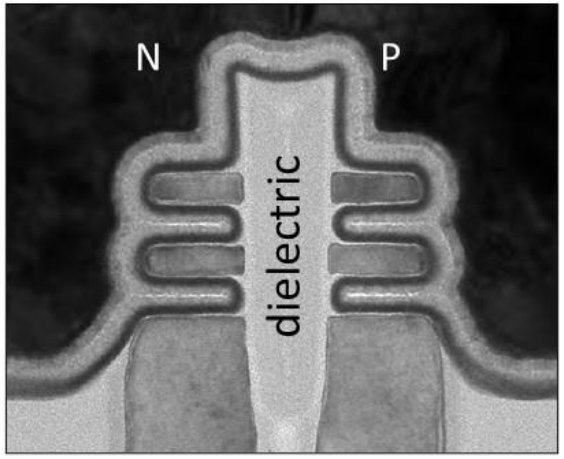
FINFET



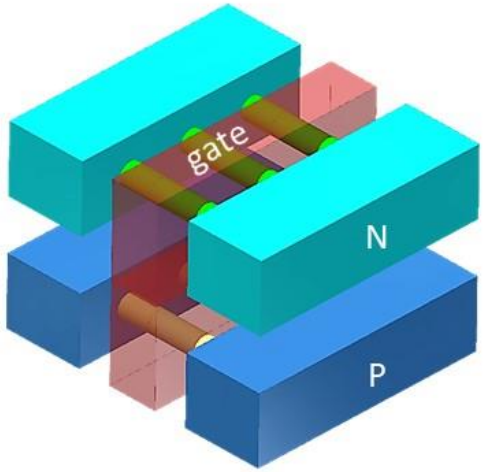
NANOSHEETS



FORKSHEET



CFET

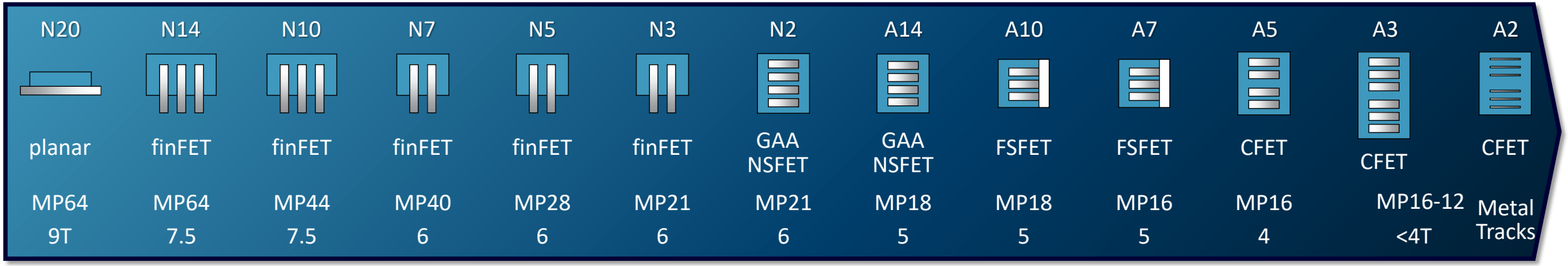


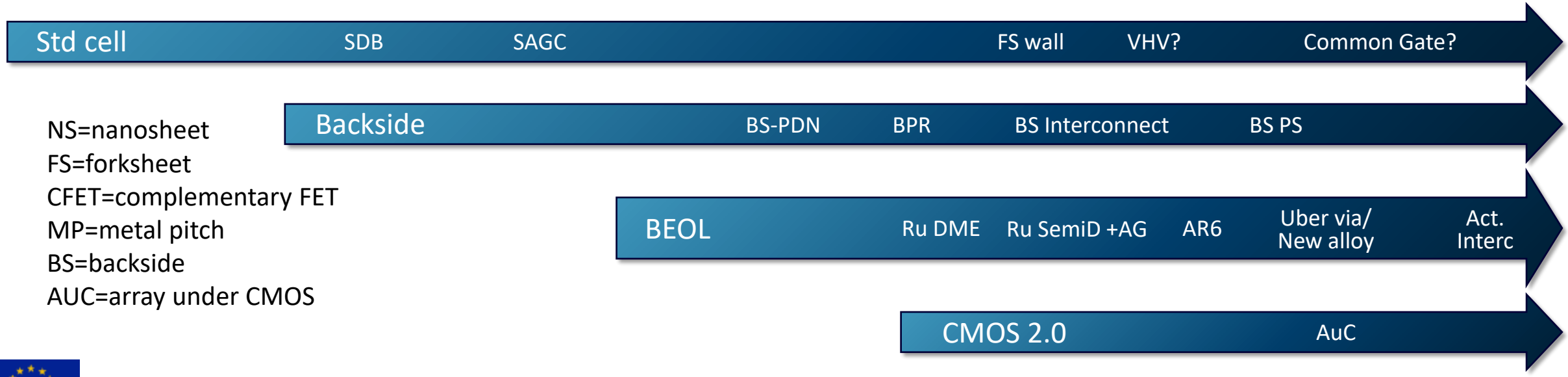
N14-N3 → N2-A14 → A10-A7 → A5-A2

N. Collaert, "Advancements in IC Technologies: A look toward the future," in IEEE Solid-State Circuits Magazine, vol. 15, no. 3, pp. 80-86, 2023, doi: 10.1109/MSSC.2023.3280433.

Logic scaling roadmap

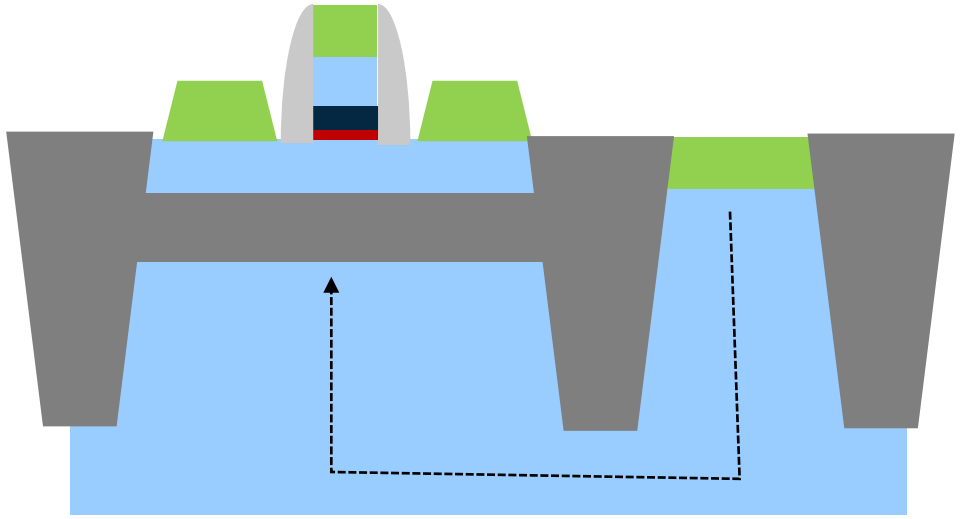
A. Veloso, ICOS workshop, April 2023.



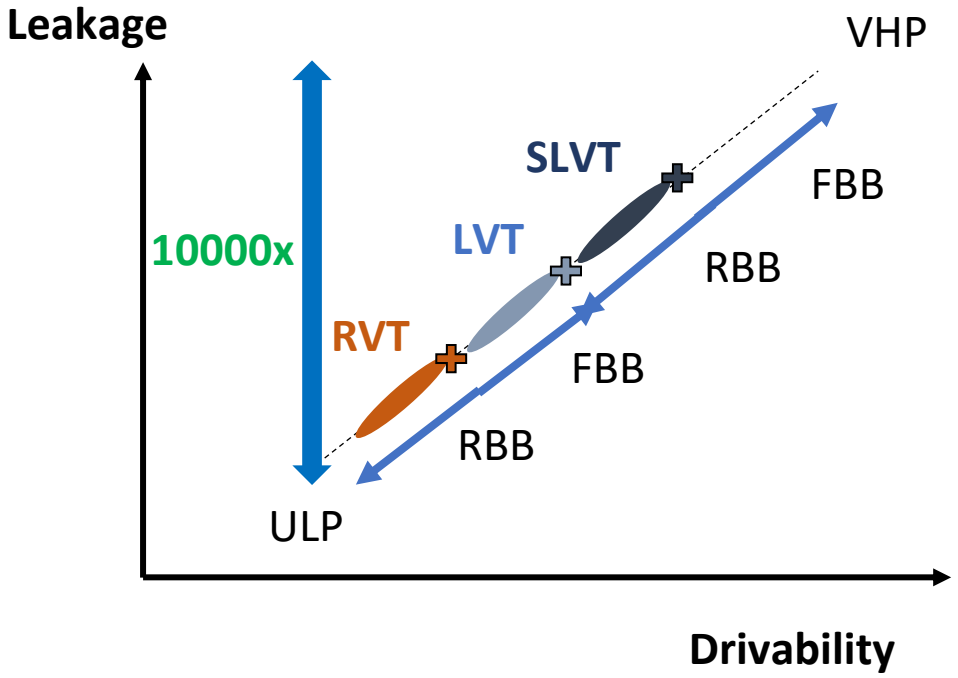


NS=nanosheet
FS=forksheet
CFET=complementary FET
MP=metal pitch
BS=backside
AUC=array under CMOS

FD-SOI Technology

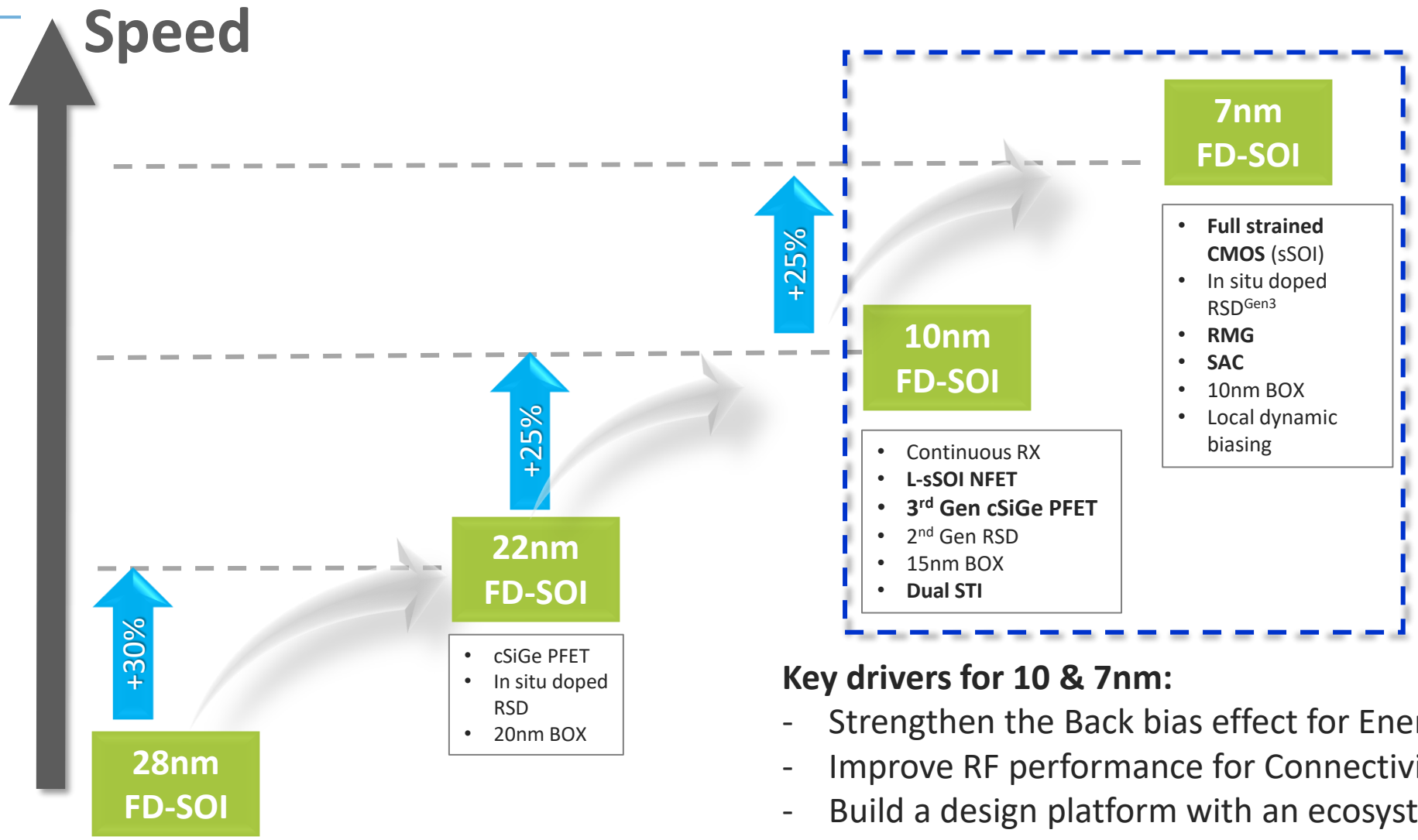


	22FDX	14nm FInFET	28nm Bulk	45nm PDSOI
f_T n-FET [GHz]	347	314	310	296
f_{max} n-FET [GHz]	371	180	161	342
f_T p-FET [GHz]	242 275 (mmWave)	285	185	-
f_{max} p-FET [GHz]	288 299 (mmWave)	140	104	-



RBB: Reverse Body Bias
FBB: Forward Body Bias
ULP: Ultra-Low Power
VHP: Very High Performance

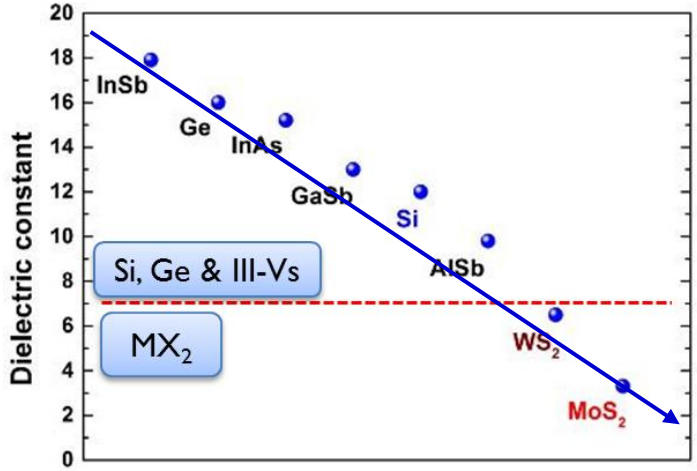
FD-SOI Technology Roadmap

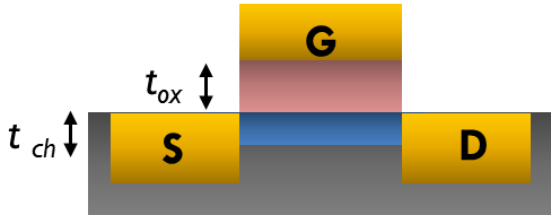


Key drivers for 10 & 7nm:

- Strengthen the Back bias effect for Energy saving
- Improve RF performance for Connectivity
- Build a design platform with an ecosystem

2D Atomic Channels: Next generation logic devices



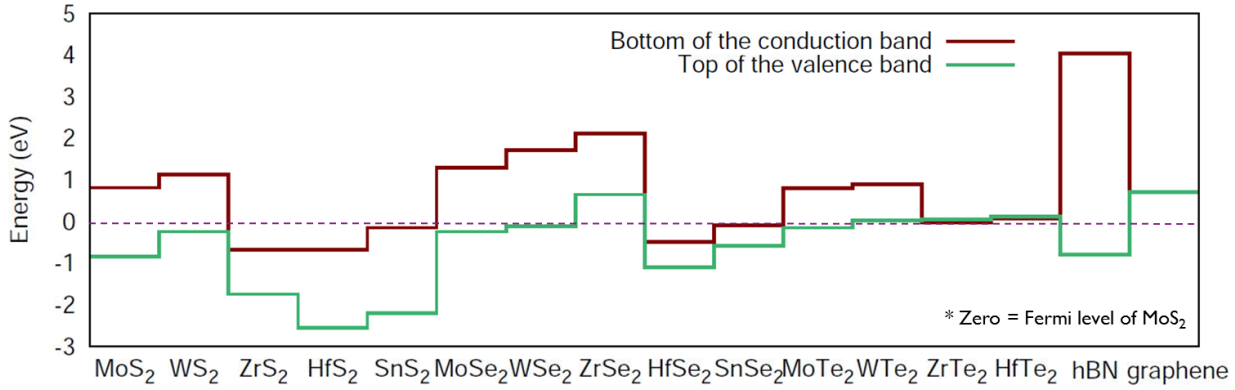


Characteristic length of short channel FETs:

$$\lambda = \sqrt{\frac{\epsilon_{ch}}{\epsilon_{ox}} t_{ch}.tox}$$

Expect reduced short channel effects in planar devices

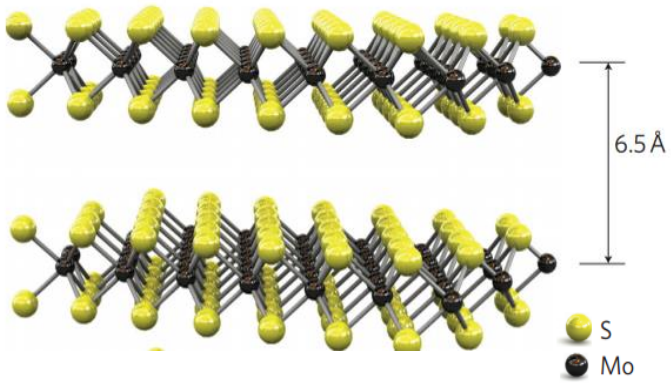
Ultra-thin materials



Choice of bandgaps and band alignment

No/few dangling bonds at interfaces

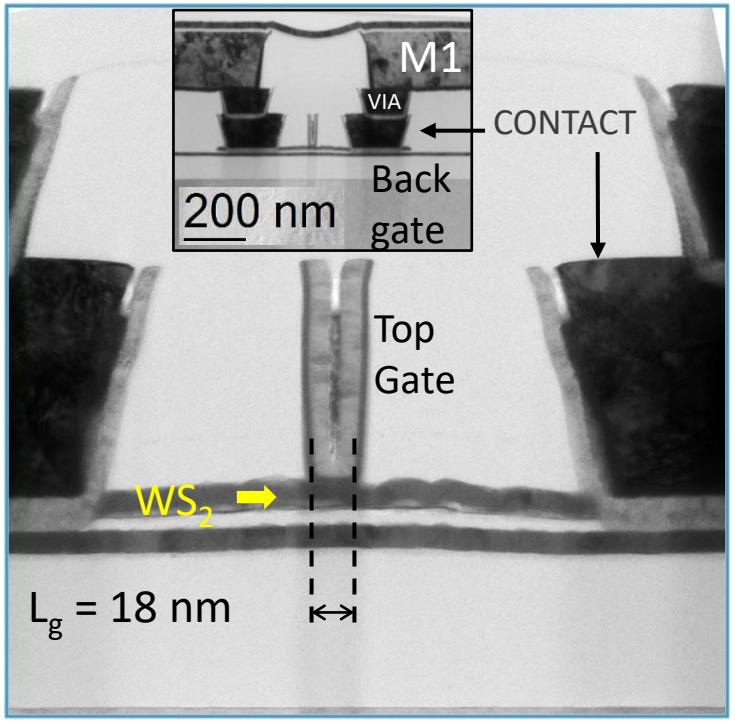
2D Atomic Channels: Next generation logic devices



MoS_2

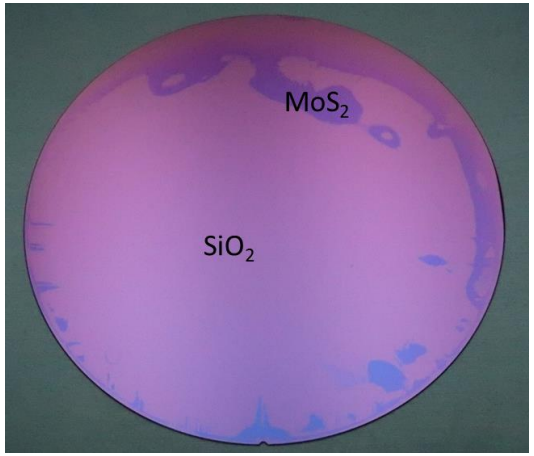
Monolayer channel thickness enables gate length scaling while keeping high performance

300mm Flow



I. Asselberghs et al, IEDM 2020

Layer Transfer



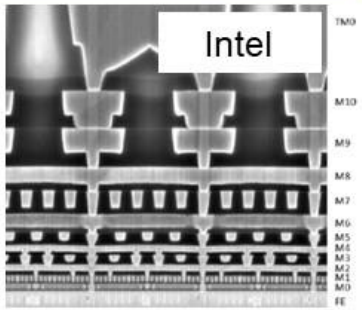
CEA-Leti, unpublished

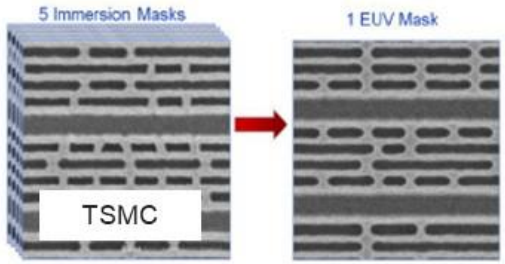
Industry BEOL trends

Courtesy: Zsolt Tokei (imec)

~36-40nm BEOL pitch

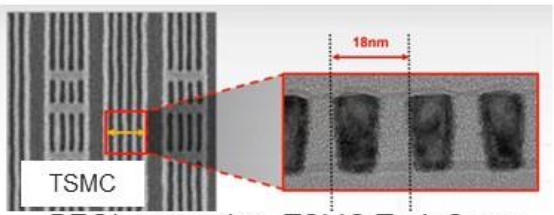
10nm technology, C. Auth et al. IEDM 2017





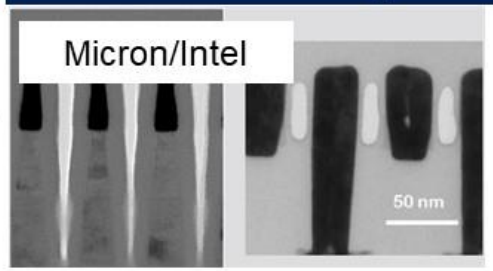
5nm CMOS, G. Yeap et al. IEDM2019

<20nm pitch

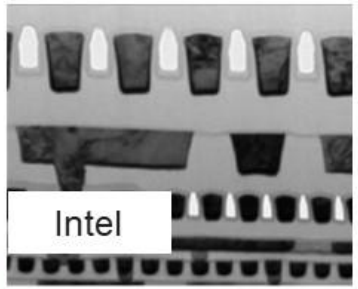


BEOL patterning, TSMC Tech Symp. 2020

AGs in memory & logic

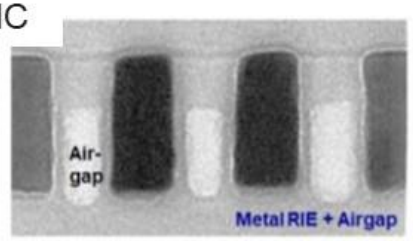


25nm NAND using air gaps K. Prall/K. Parat IEDM2010



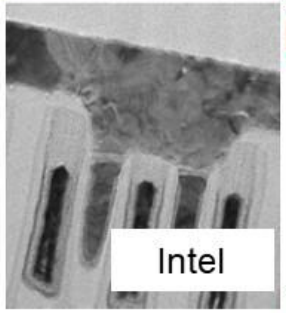
K. Fisher et al., 14nm, IITC2015

TSMC

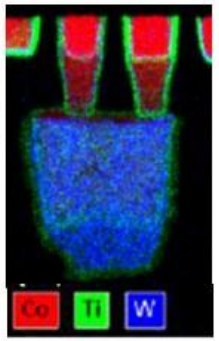


Y.J. Mii et al. Keynote, VLS022

Co interconnects

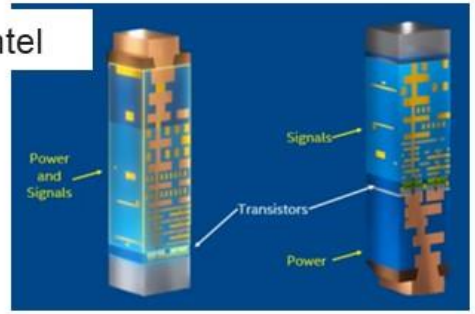


Self- aligned contact, C. Auth et al. IEDM 2017



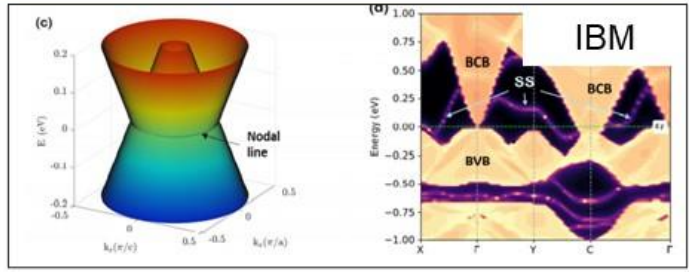
Reliability of Co, F. Griggo et al. IRPS 2018

Backside Power Delivery

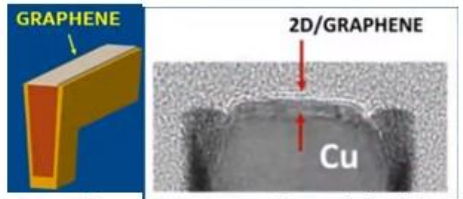


M.C. Mayberry, Keynote IITC 2020

New materials on the horizon

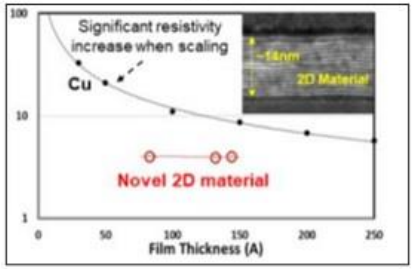


Topological semi-metal, C.T. Chen et al. IEDM2020



Intel

Graphene capped metal, R. Chau et al. Keynote IEDM2019

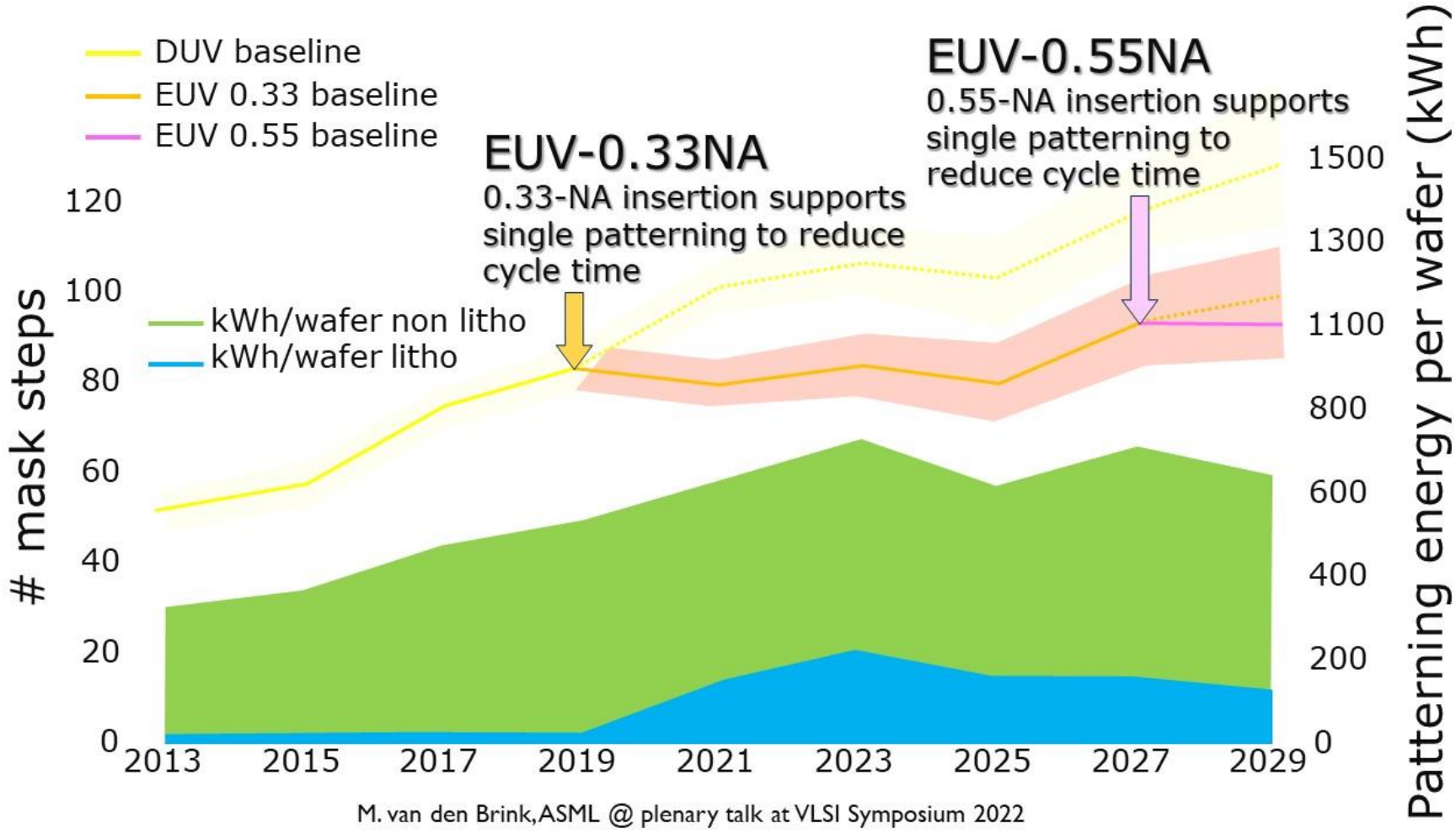


TSMC

New conductor, Y.J Mii Keynote VLSI2022

Pitch scaling, airgaps both in memory and logic, new materials

EUV lithography key enabler for scaling



M. van den Brink, ASML @ plenary talk at VLSI Symposium 2022

Today's Scaling Challenges Drive the Need for CMOS 2.0

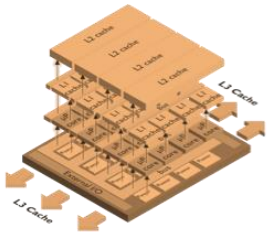
CMOS Today

SRAM scaling?
I/O?
HP vs LP?
Power delivery?
Workloads?

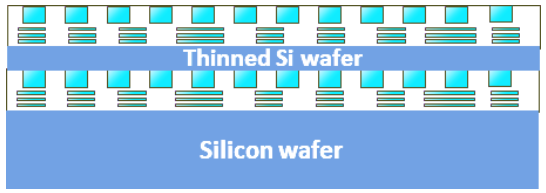
CMOS 2.0

Separate, optimize
re-assemble

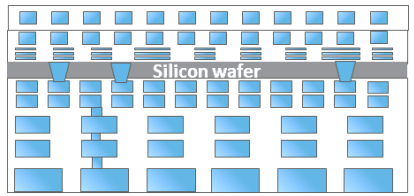
3D-SoC



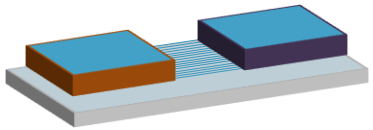
Sequential 3D



Backside Process



Active Interposer



S. Samavedam, ITF @ IEDM 2023

CMOS 2.0 Vision brings heterogeneity inside the SoC

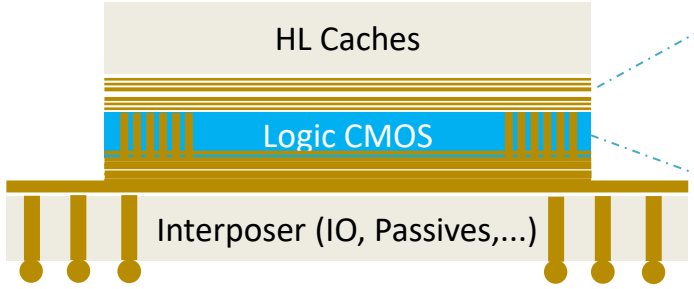
International Cooperation On Semiconductors

Past

CMOS 1.0
Homogeneous Platform

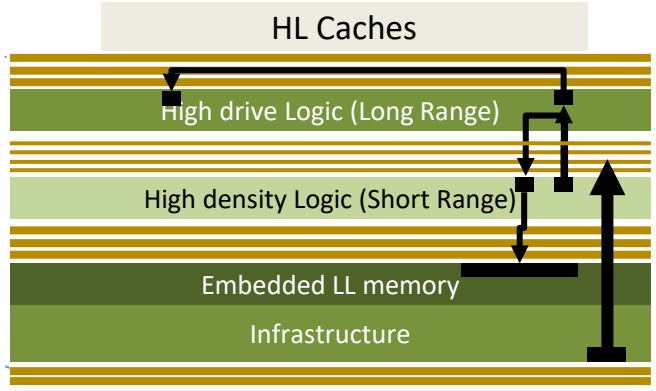
Present

Heterogeneous Systems



Future

CMOS 2.0
Heterogeneous Logic Platform



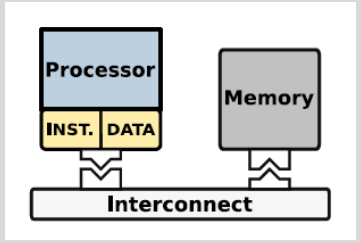
POWER WALL	Multi-core HW accelerators	Backside PDN	Functional Backside
BANDWIDTH WALL	Complex Cache /Memory hierarchy	Active Interposer	Active Interconnect
MEMORY WALL		Hybrid Bonding	Array-Under-CMOS Large embedded memory

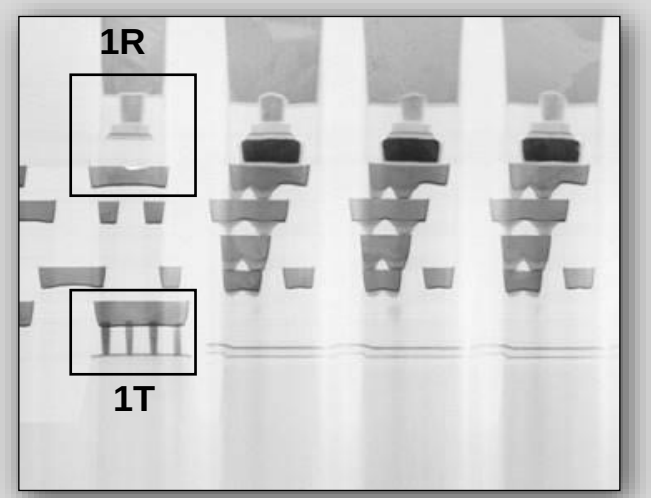
Why Emerging Resistive Memories?

High dense on-chip memory

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

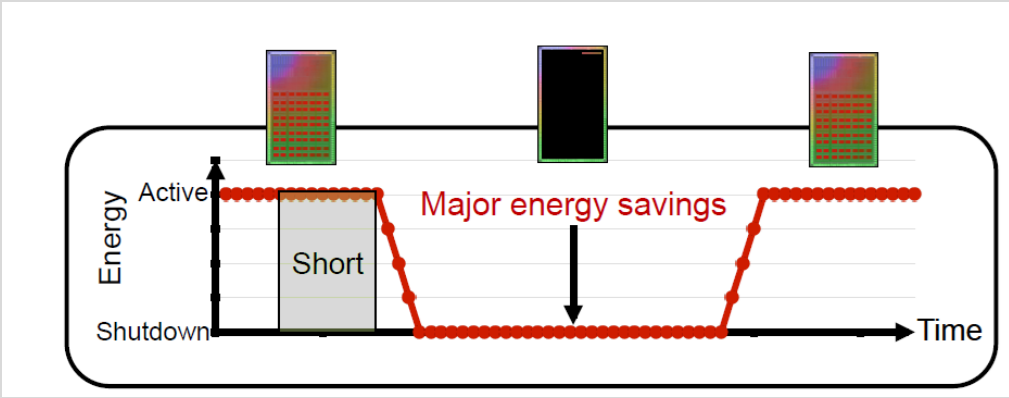
[F. Tu, et al., 2018 ACM/IEEE]





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

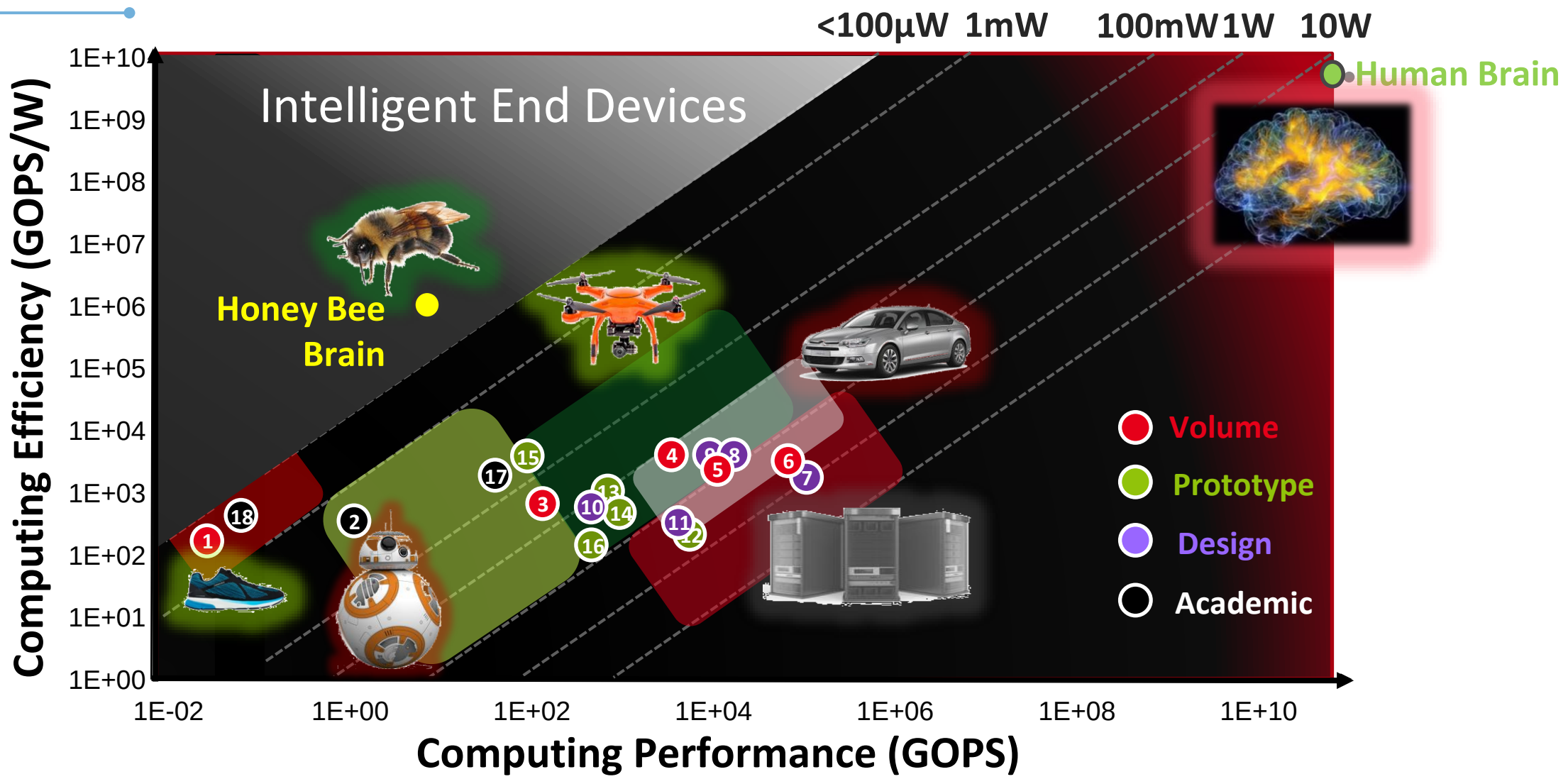
Emerging Non-Volatile Memories

	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!

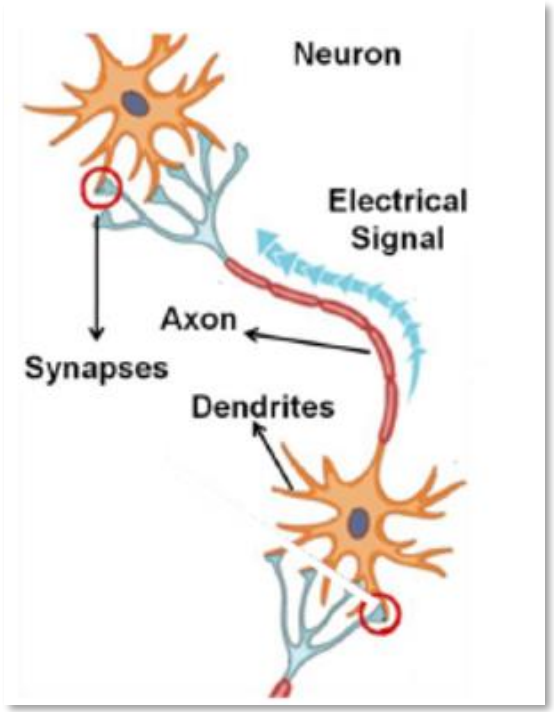
Memory activity focus on embedded NVM for NOR flash replacement

Energy Efficiency is far from biological Systems

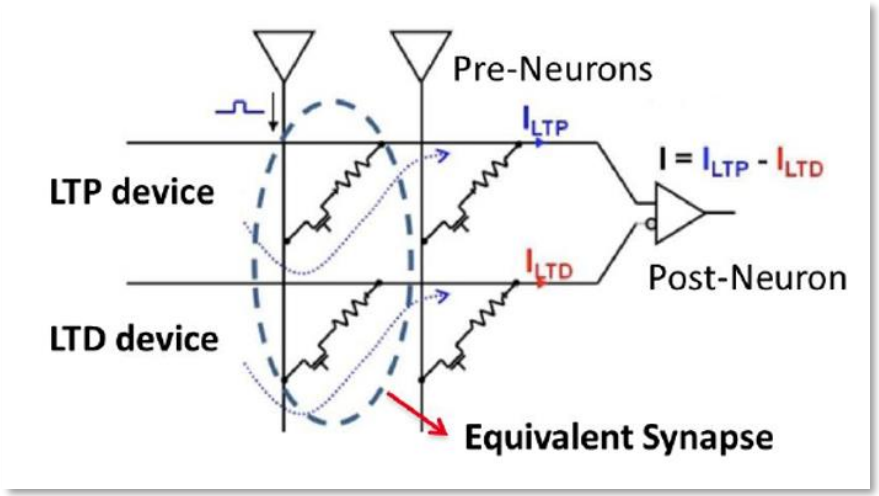


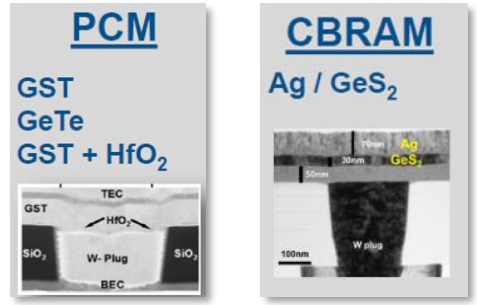
Neuromorphic based RRAM circuit

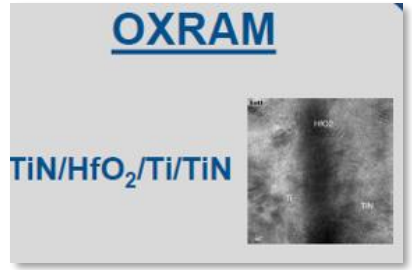
M. Suri et al, IEDM 2011.



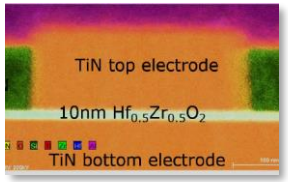
2 PCRAM Example:





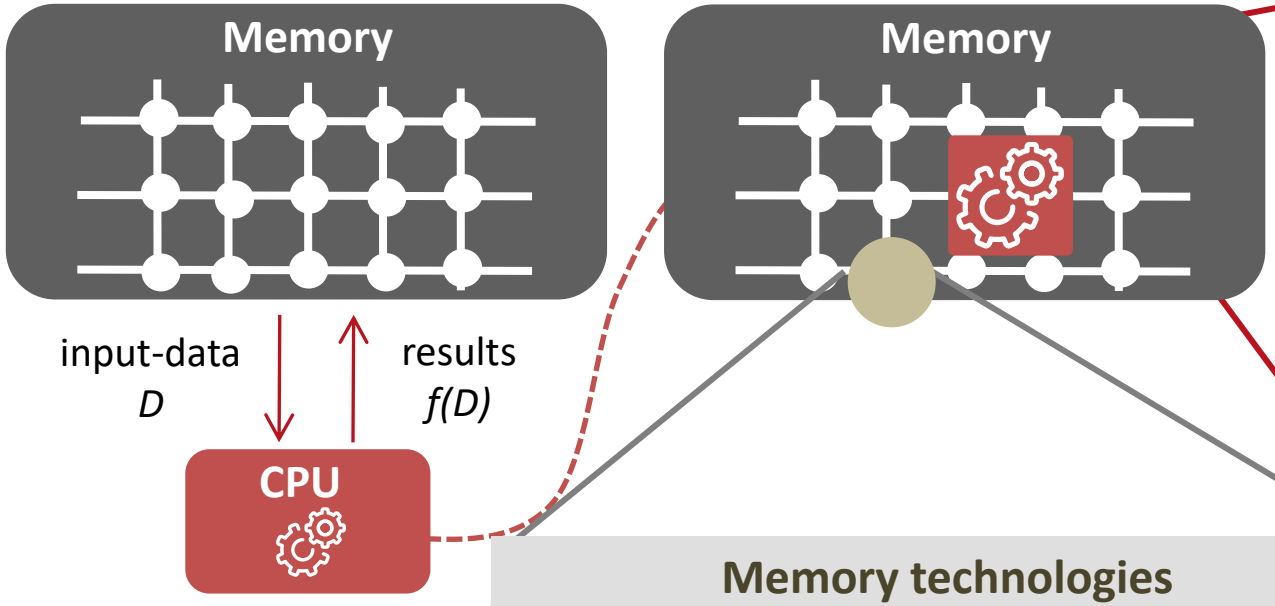


FeRAM



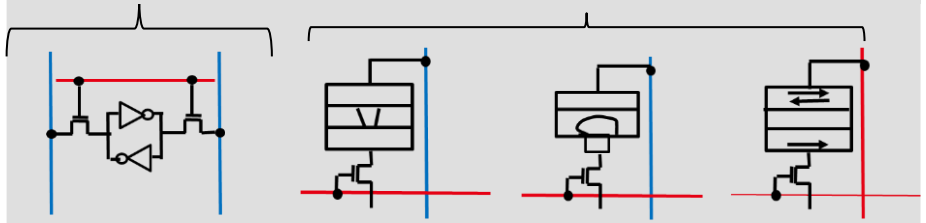
In memory computing

Von Neumann computing

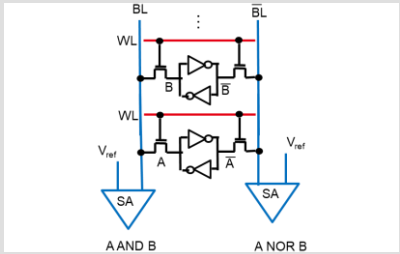


In-memory computing

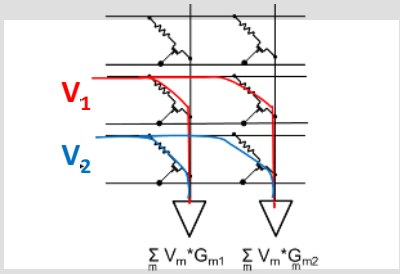
Memory technologies



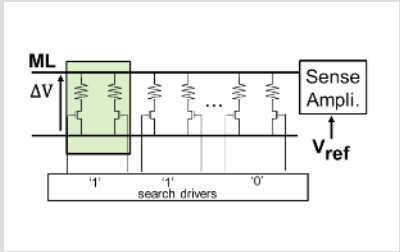
Logic and arithmetic operations



Digital operations using SRAM



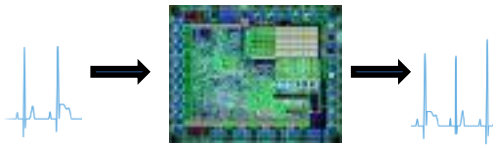
Analog: multiply and accumulate



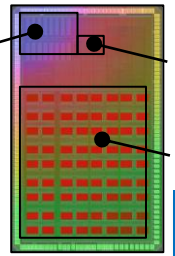
Searching / matching (CAM)

Near- & in-memory computing

Bayesian Neural Networks

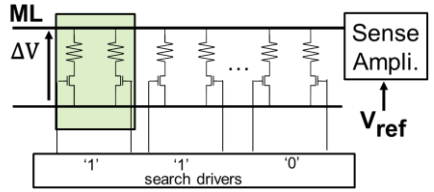


Spiking Neural Network



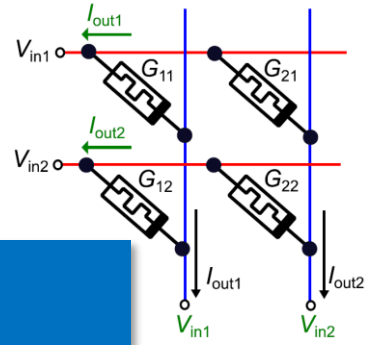
Energy-efficient NN accelerator

RRAM for near- and in-memory computing

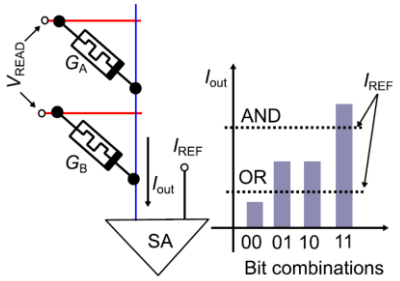


Search/match operations (CAM)

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

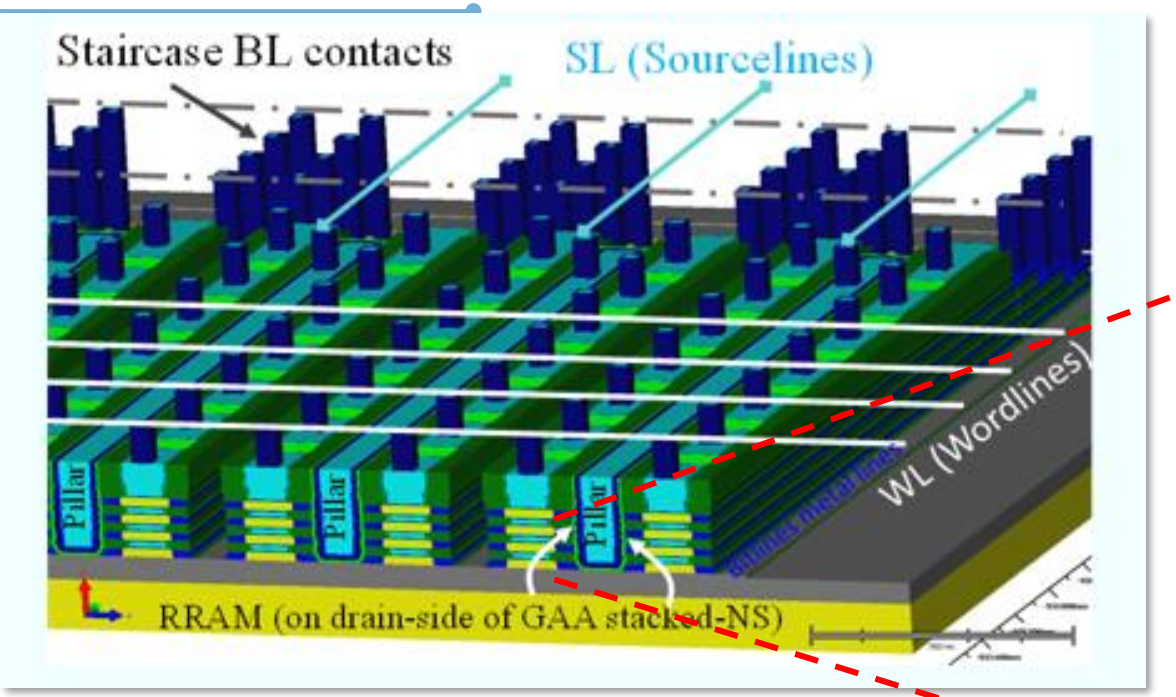


Deep learning, matrix-vector multiply



In-memory logic

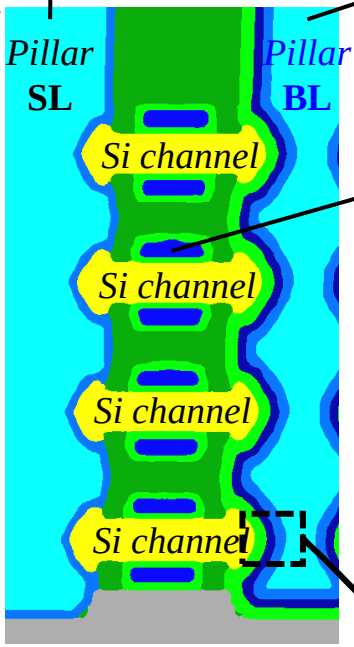
Towards In Memory Computing



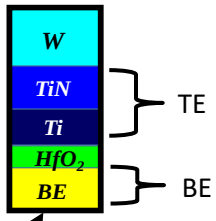
Common SL for all GAA stacked-NS

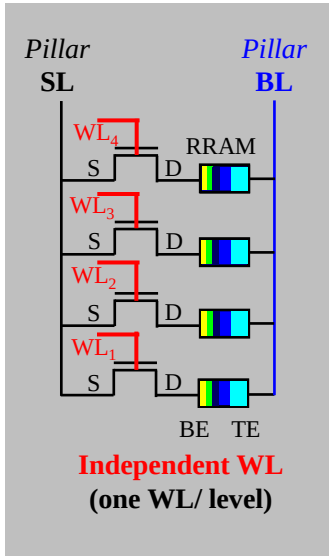
Common BL for all GAA stacked-NS

Courtesy of S. Barraud, Leti



Independent WL
One GAA by level





Enabler for Hyperdimensional computing

A possible Edge IA Roadmap

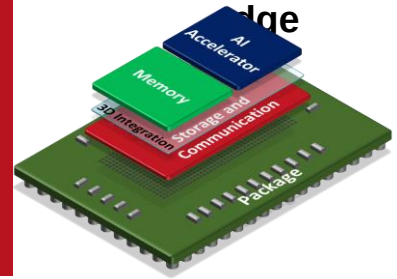
Courtesy of E. Vianello, Leti

Functions

Performance

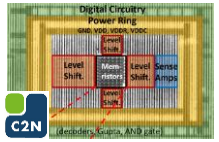
Heterogeneous architecture

Next generation AI at edge

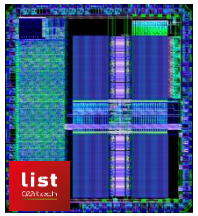


Multi-die (chiplet) integration on base die

Smart sensor



Bayesian machine



Esperanto
Spiking recurrent

Classification

BNN
Fully distributed 2T2R

SPIRIT
Analog spiking

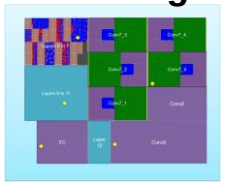
X10000 neurons
Power ÷10

<1pJ / synaptic event

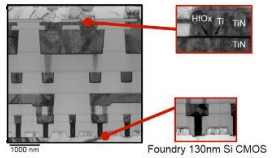
Spiking CNN

Largo

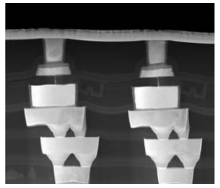
Feature extractor of HD images



NeuroCorgi ASIC

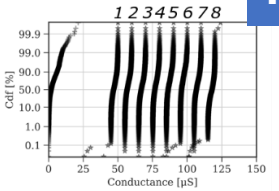


1T1R 130nm

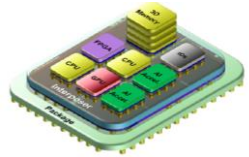


1T1R 28nm

1T1R 22nm



Multi-bit



Multi-die integration

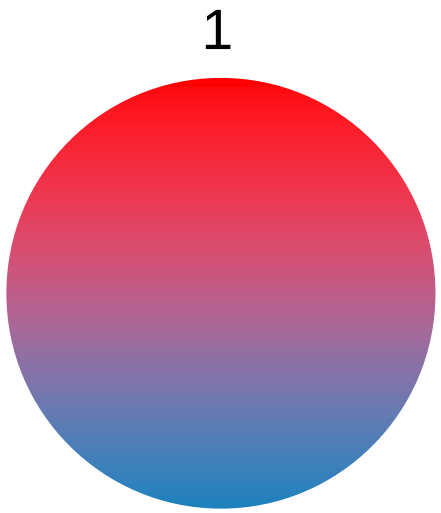
From bits to qubits

Bit

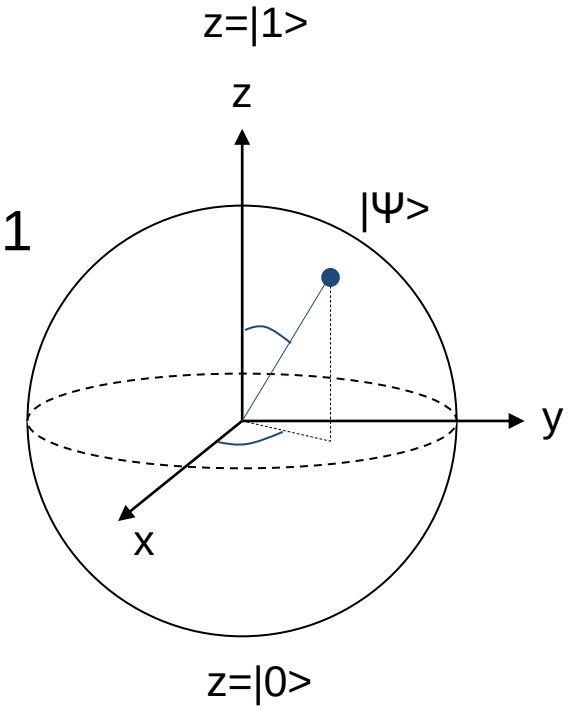
1

0

Qubit



0 AND 1

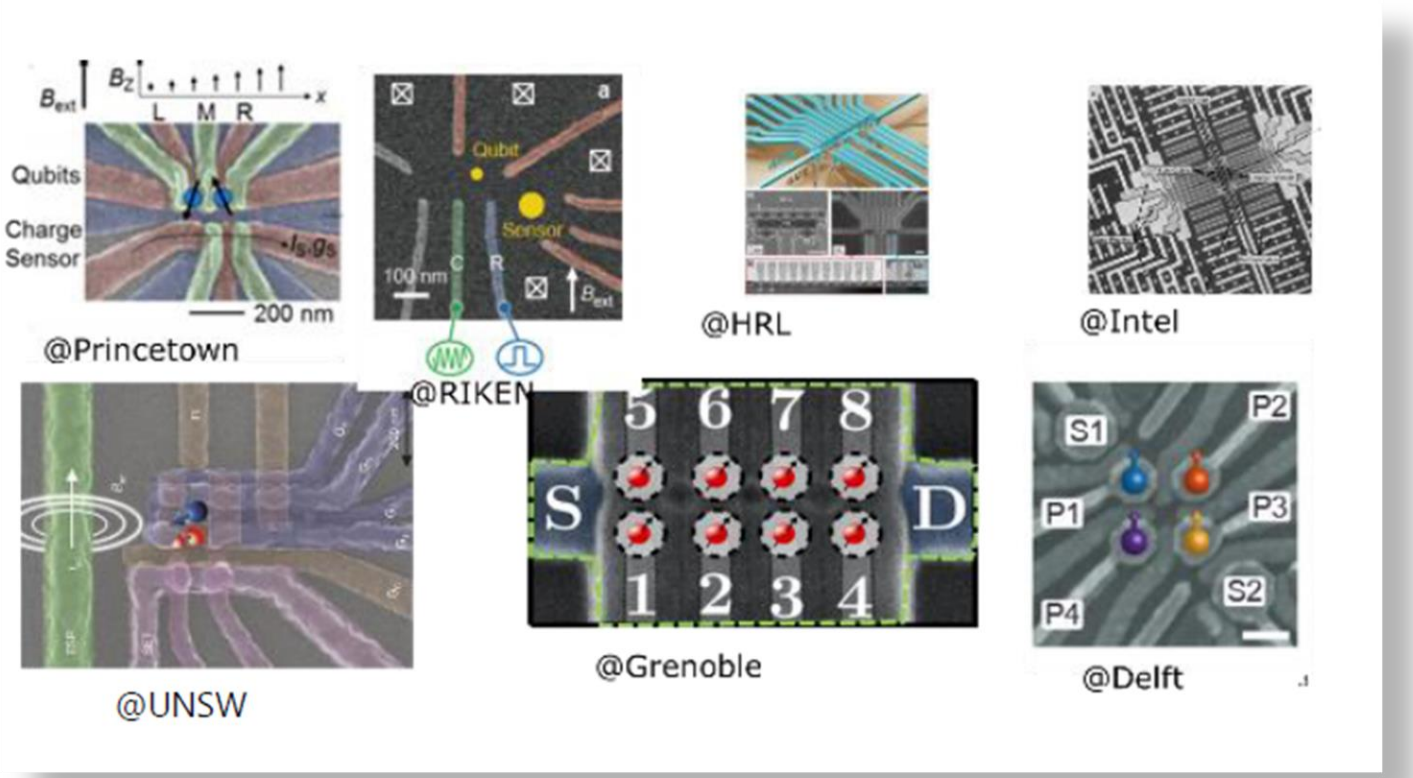


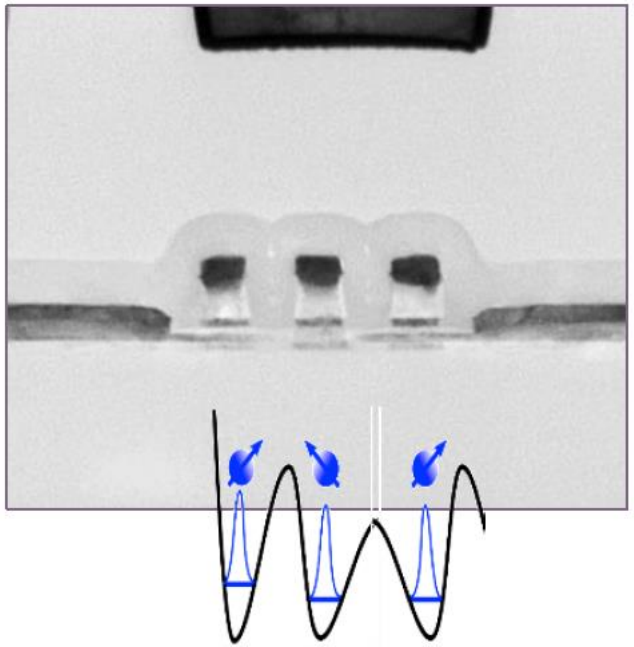
Many flavors of qubits

	Superconductor	Si spin qubit	Trapped ion	Photon
Size*	$(100\mu m)^2$	$(100nm)^2$	$(1mm)^2$	~$(100\mu m)^2$
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)

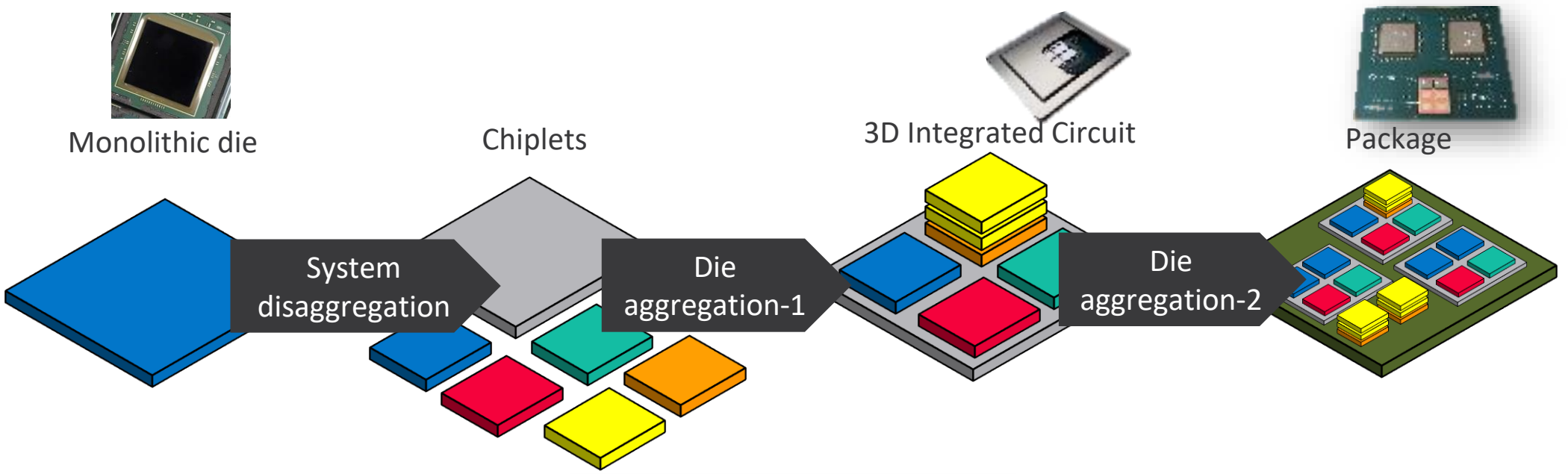
From bits to qubits

- Quantum Physics to compute



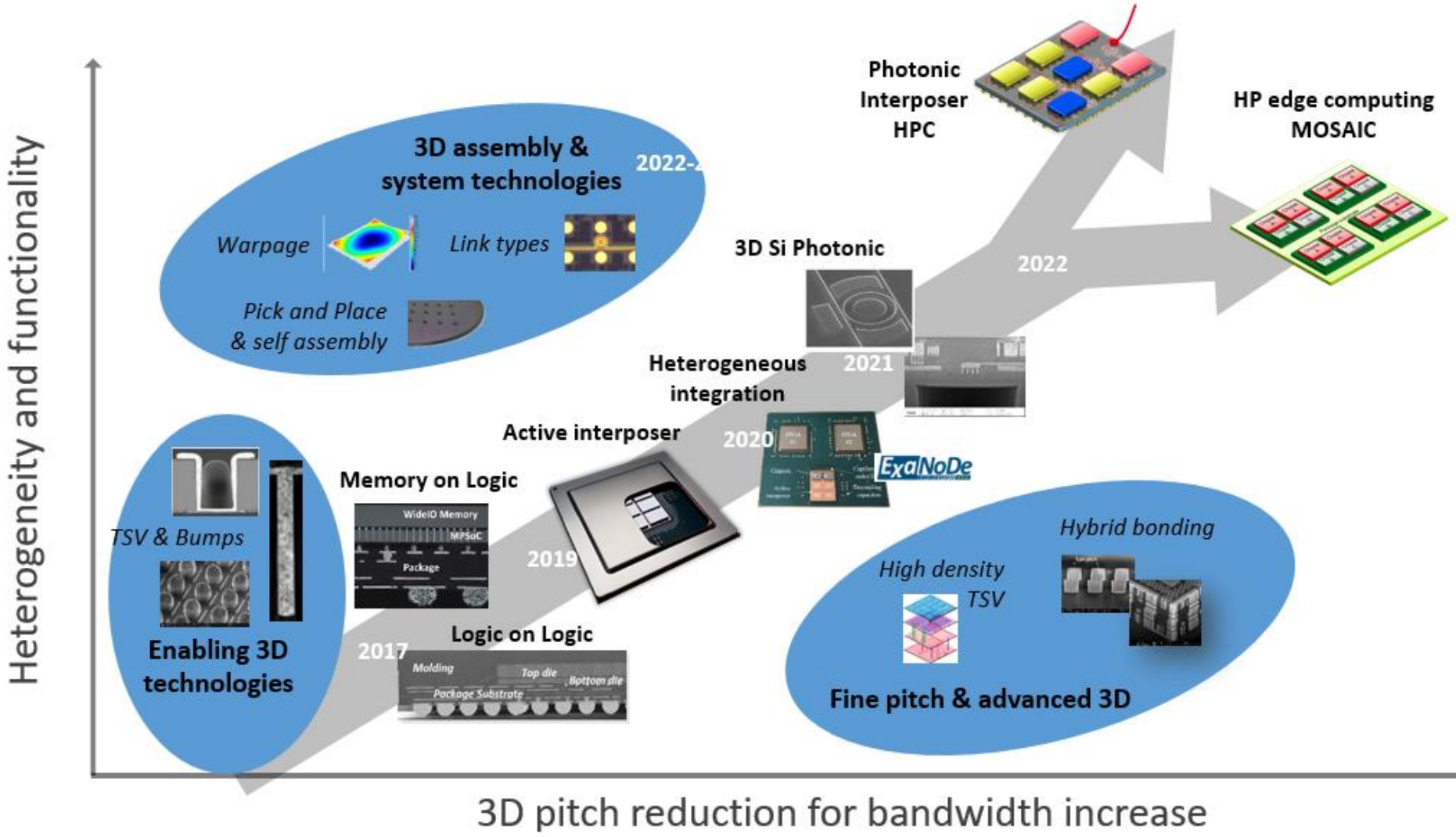


Chiplets: the new IC design paradigm

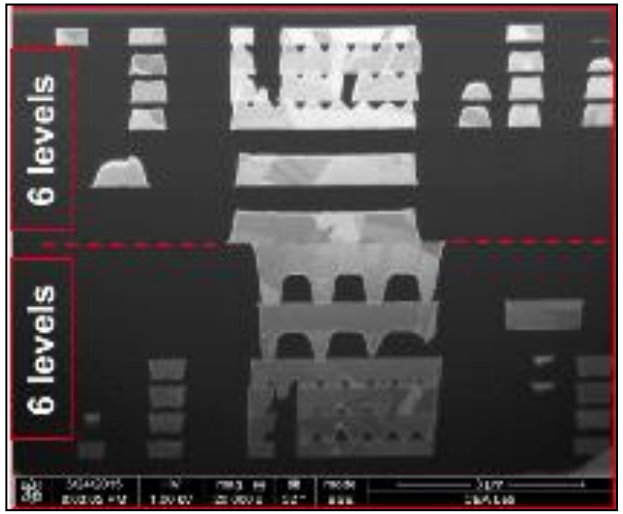


Up to 100x gain on Power Efficiency with 3D

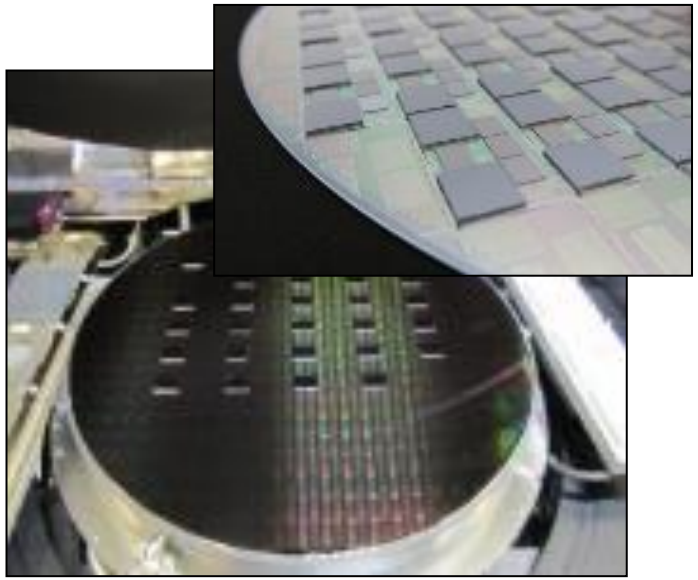
3D Tool Box for chiplet integration



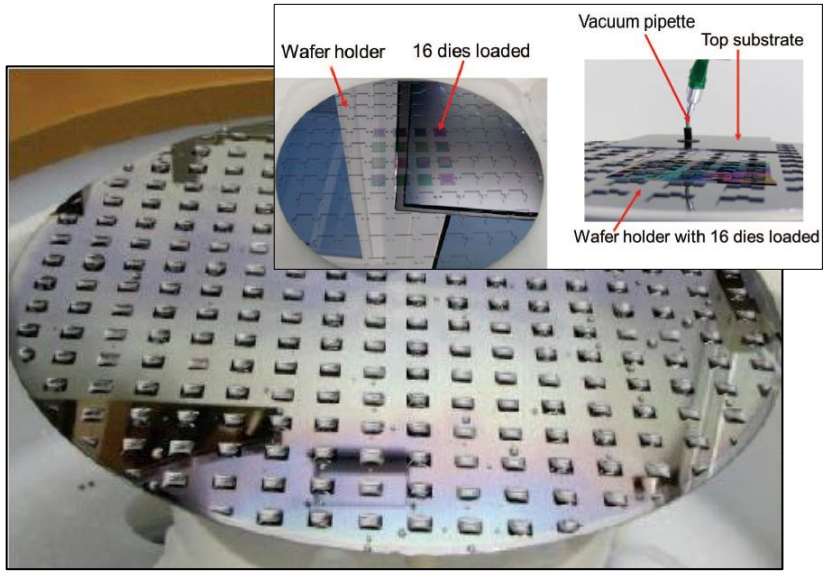
Hybrid Bonding Solutions



- 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

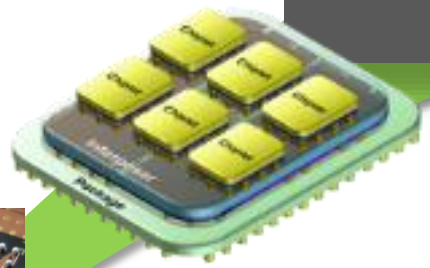


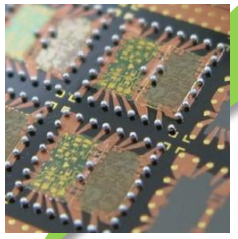
- Collective D2W approaches
- Self-assembly for high precision & high throughput

3D: from packaging to monolithic

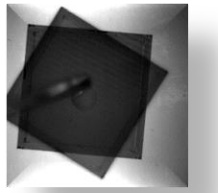
TSV + µBump[1]
Pitch : 20 µm

~10^3 3DC/mm²
=> Entire core



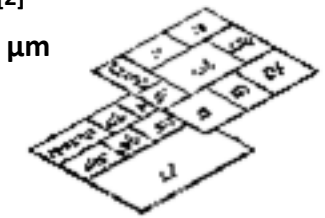


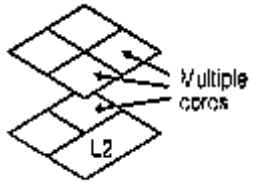
Self Assembly
Pitch : 1-5 µm



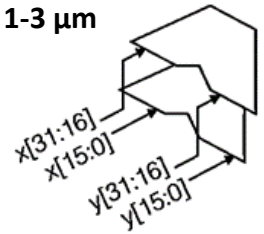
~10^5 3DC/mm²
⇒ Logic Blocs
⇒ Logic Gates

Cu/Cu [2]
Pitch : 2-5 µm

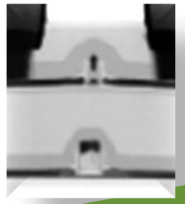




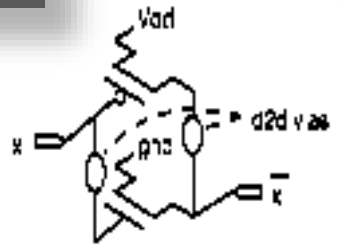
HD-TSV [2]
Pitch : 1-3 µm

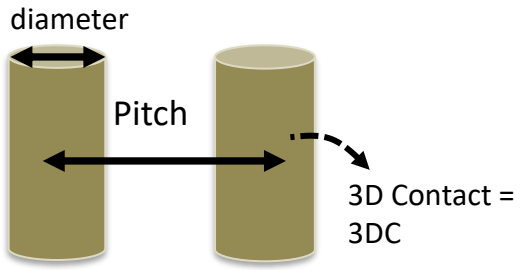


10^8 3DC/mm²
=> Transistor level



M3D (CoolCube™) [3]
Pitch : 0.05-0.1 µm



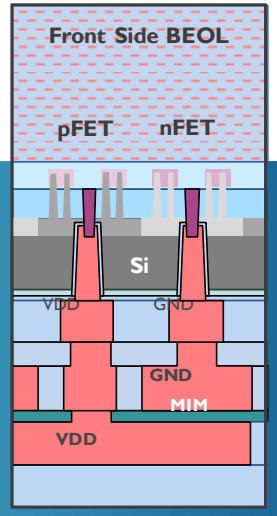


[1] Cheramy, S., et al. "Advanced Silicon Interposer", C2MI Workshop, 2015
[2] Patti, B., "Implementing 2.5D and 3D Devices", In AIDA workshop in Roma, 2013
[3] Batude, P., et al. "3DVLSI with CoolCube process: An alternative path to scaling ." VLSI technology symposium 2015

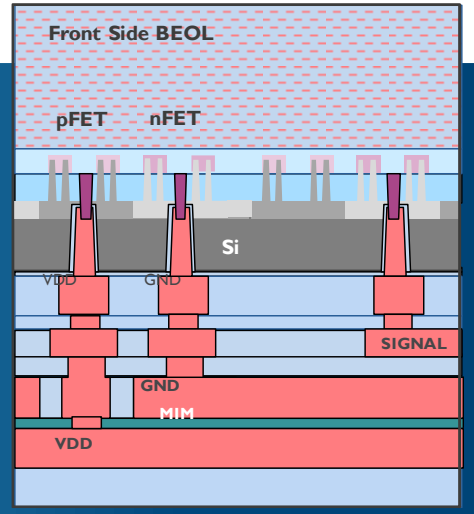
Towards a functional backside

J. Ryckaert, ITF Japan 2022

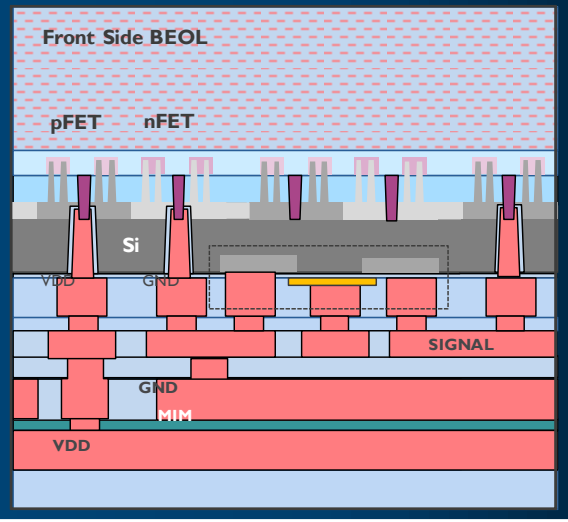
Backside Power Delivery



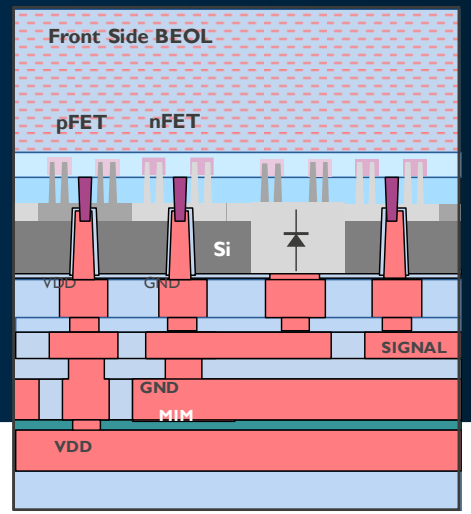
Backside Global Interconnect



Backside Devices

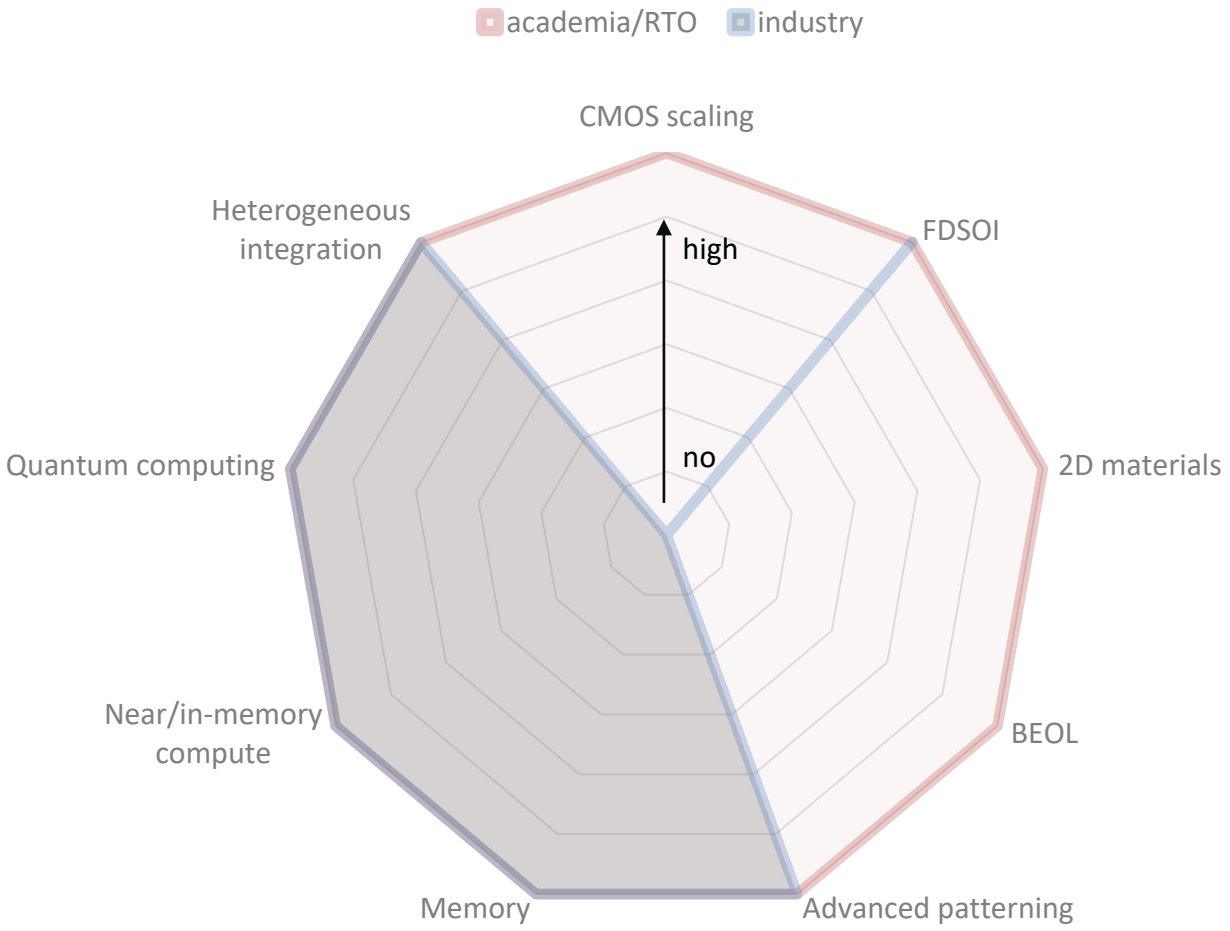


Device Backside Extension



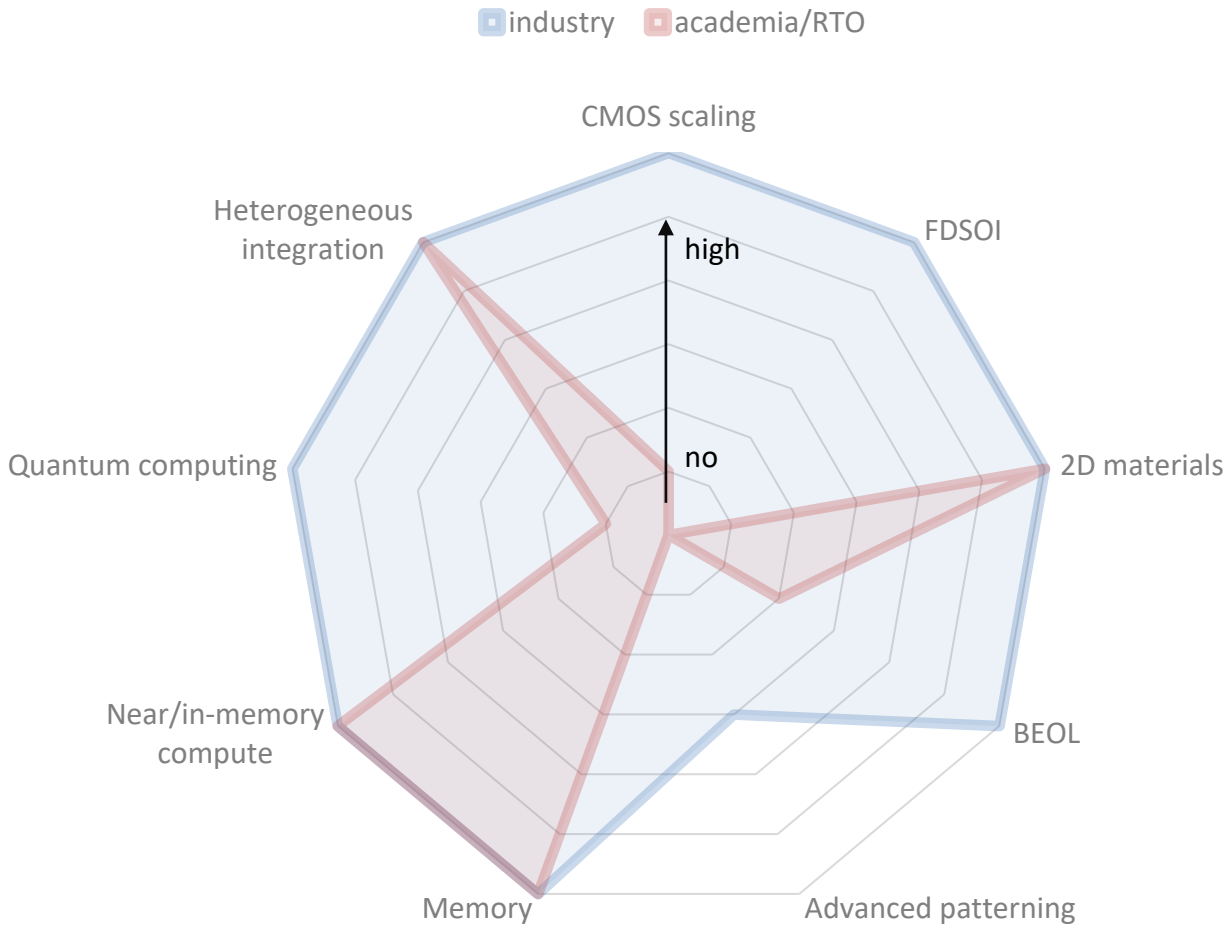
Enhancing system performance by migrating system functions to the backside

EU and non-EU actors - EU



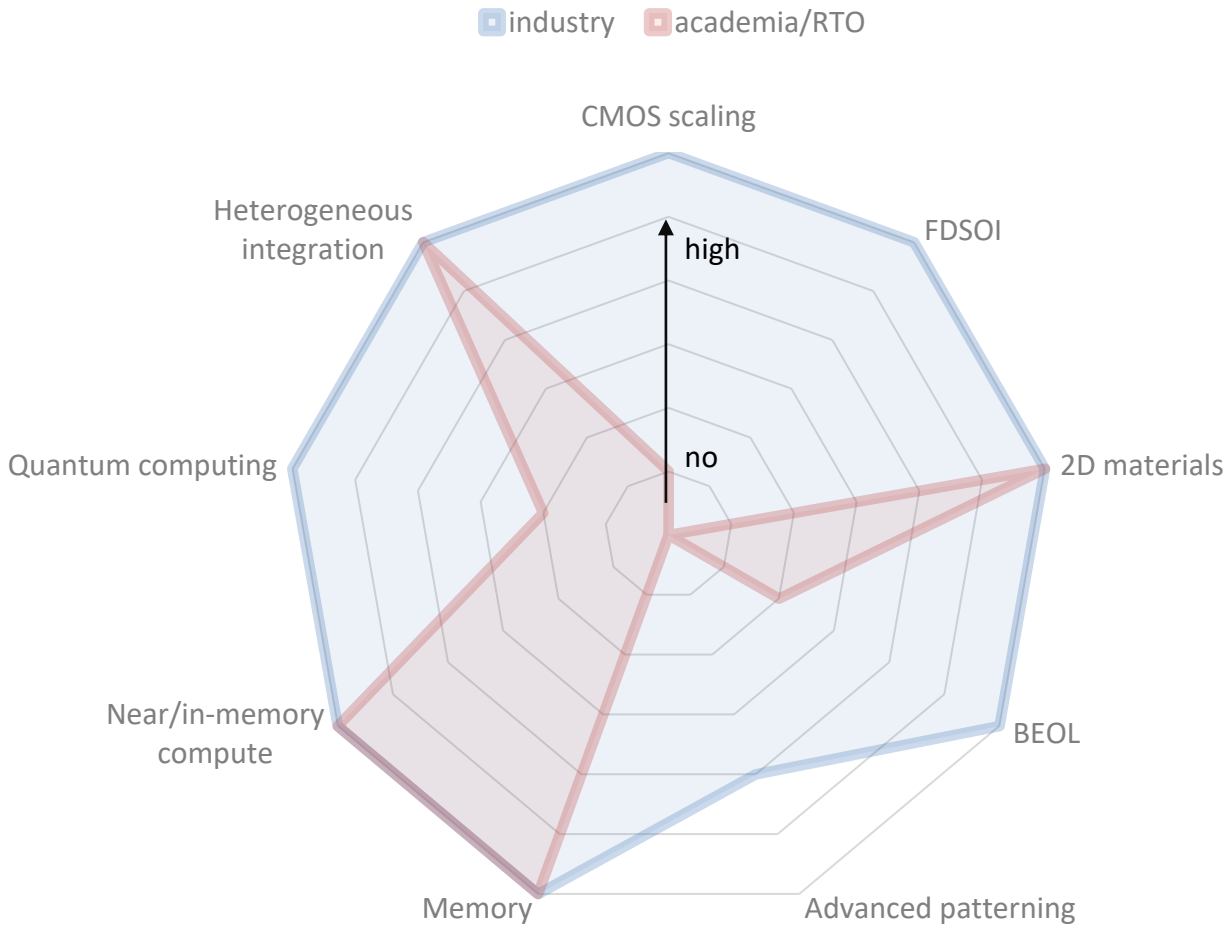
- R&D very strong in all areas of compute
- Unique strong position in EUV lithography
- In general, industrial EU players lacking to take up R&D

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

- Compute needs are growing at an unprecedented speed
- Innovations across different levels needed to enable chip & system performance enhancement
- Sustainability becoming an increasingly important metric for evaluating technology choices
- Europe is very strong in R&D in all advanced compute areas
- Manufacturing: several industry initiatives & pilot line programs
- Lack of Fabless companies that drive the needs in terms of advanced compute
- Strong initiatives in EU in start-up company creation

This project has received funding from the European Union's Horizon Europe research and innovation programme under GA Nº 101092562