ESSCIRC/ESSDERC 2023

SiNANO-ICOS Workshop

"European Strengths and Gaps in Emerging Semiconductor Technologies"

Advanced computation: review of main EU and international activities and technologies

Olivier Faynot and Nadine Collaert

CEA-Leti, IMEC

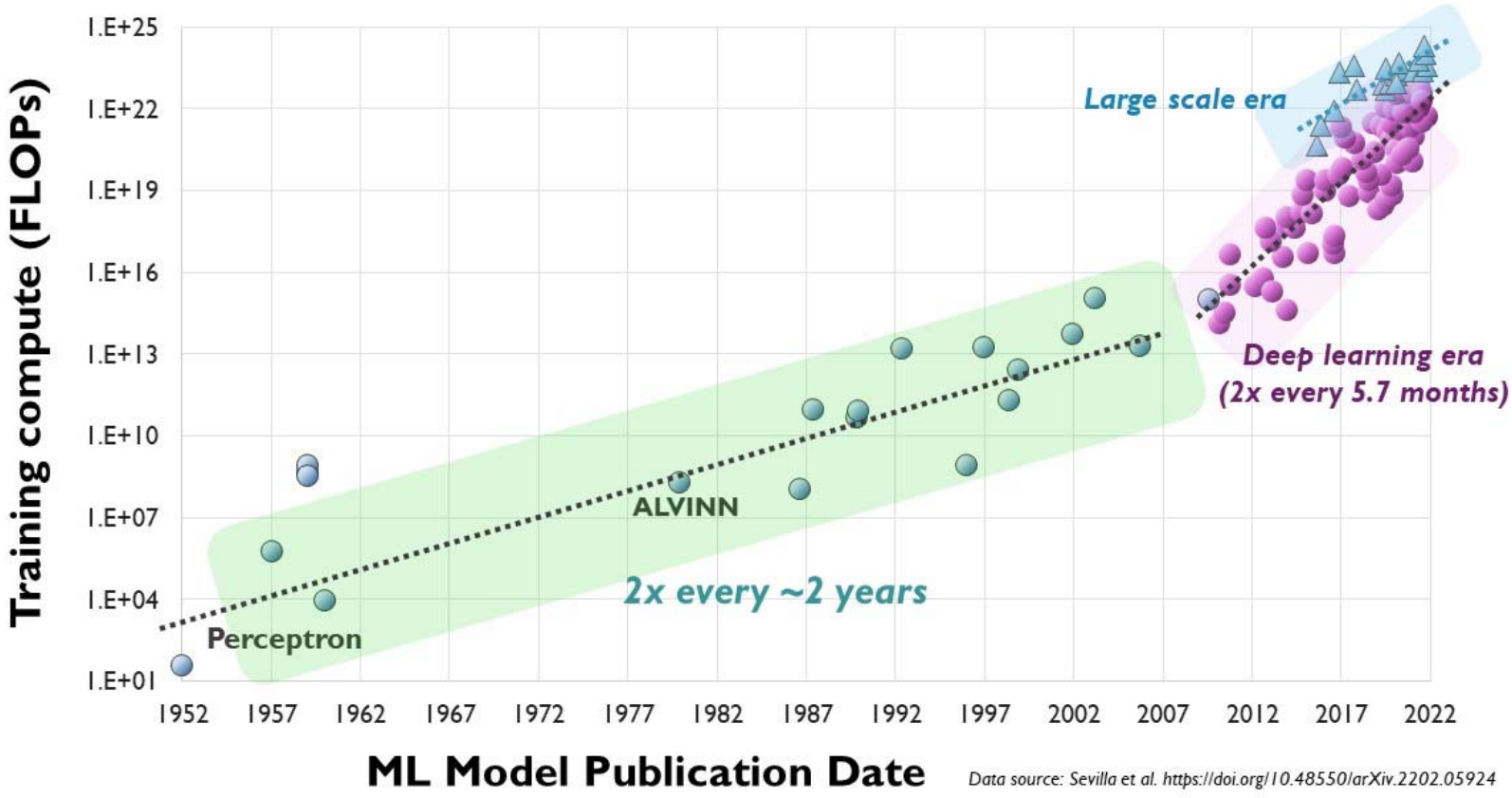
olivier.faynot@cea.fr, nadine.collaert@imec.be

Lisbon, September 11, 2023

Outline

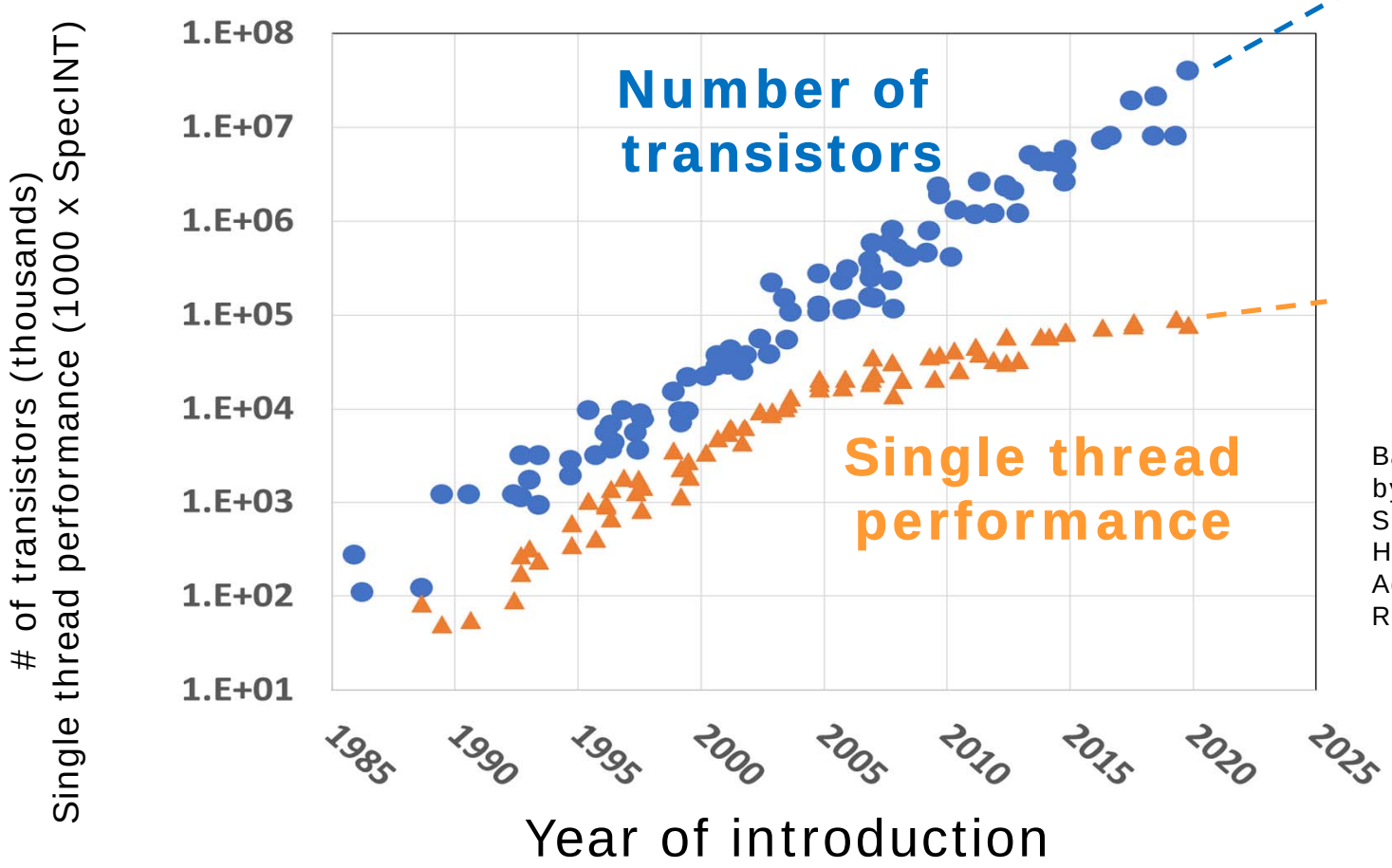
- Introduction: trends and challenges
- Computing roadmap:
  - CMOS device architecture
  - 2D materials for FEOL
  - New materials for BEOL
  - Lithography
- Memory technologies
- Beyond Von Neumann disruptive approaches:
  - Near or in-memory computing
  - Quantum computing
- Heterogenous integration: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

Compute needs for ML continue to grow



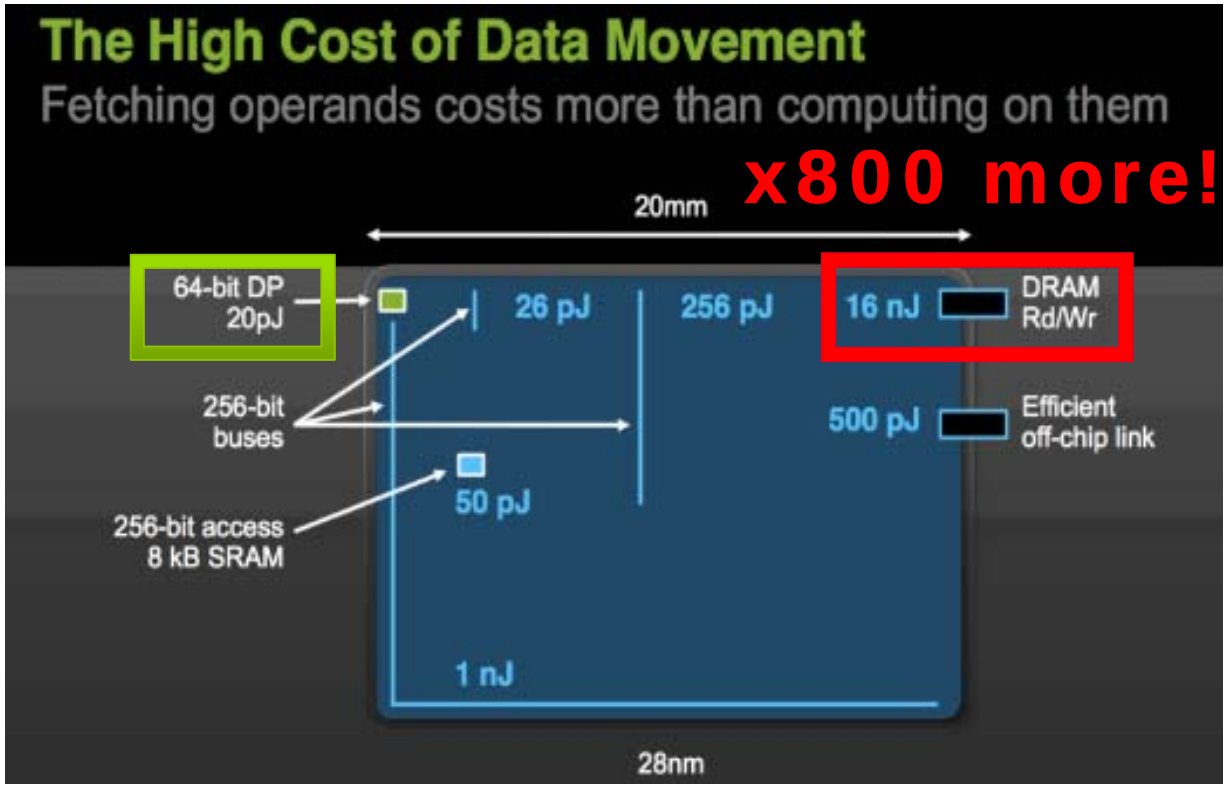
ML = machine learning

Slowdown of Single Thread Performance in CPUs



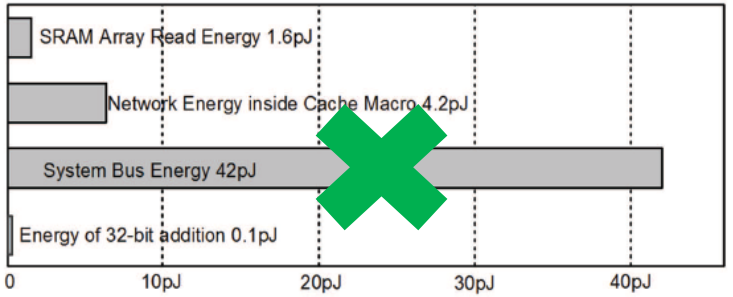
Based on original data plotted by M. Horowitz, F. Labonte, O. Shachan, K. Olukotun, L. Hammond, C. Batten. Additional data compiled by K. Rupp

The cost of moving data



Bill Dally, “To ExaScale and Beyond”, 2010

[J. Wang – ISSCC’19]

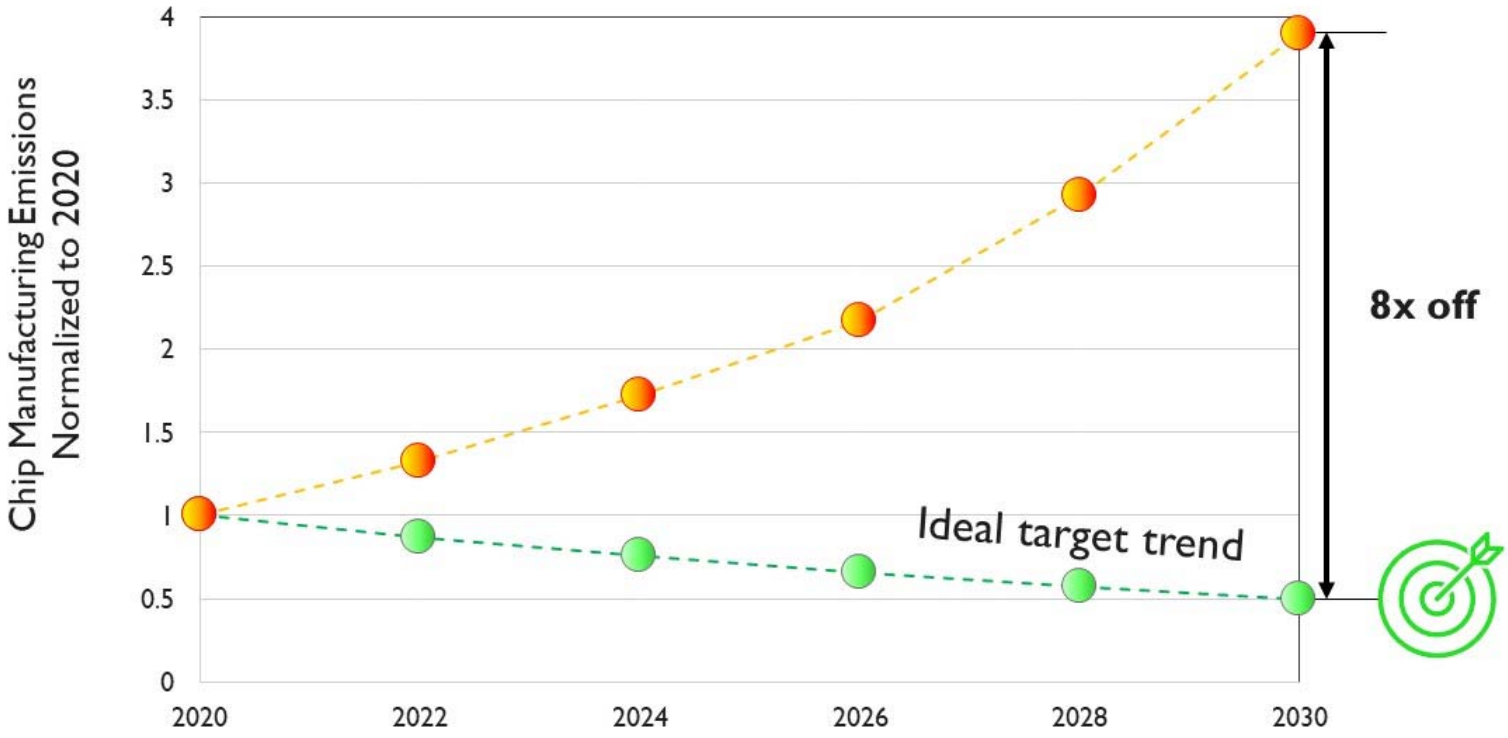


~90% of energy is in **data transfer**
→ IMC could lead **8x reduction**

Operation energy is negligible

Memory access and control energies dominate

Carbon Emissions Estimate of Logic Manufacturing: "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.*

The required gain in energy efficiency

>1000x by 2030

- CMOS scaling
- Memory technologies
- Disruptive Computing
- Chiplet & 3D System

New device architectures to extend scaling

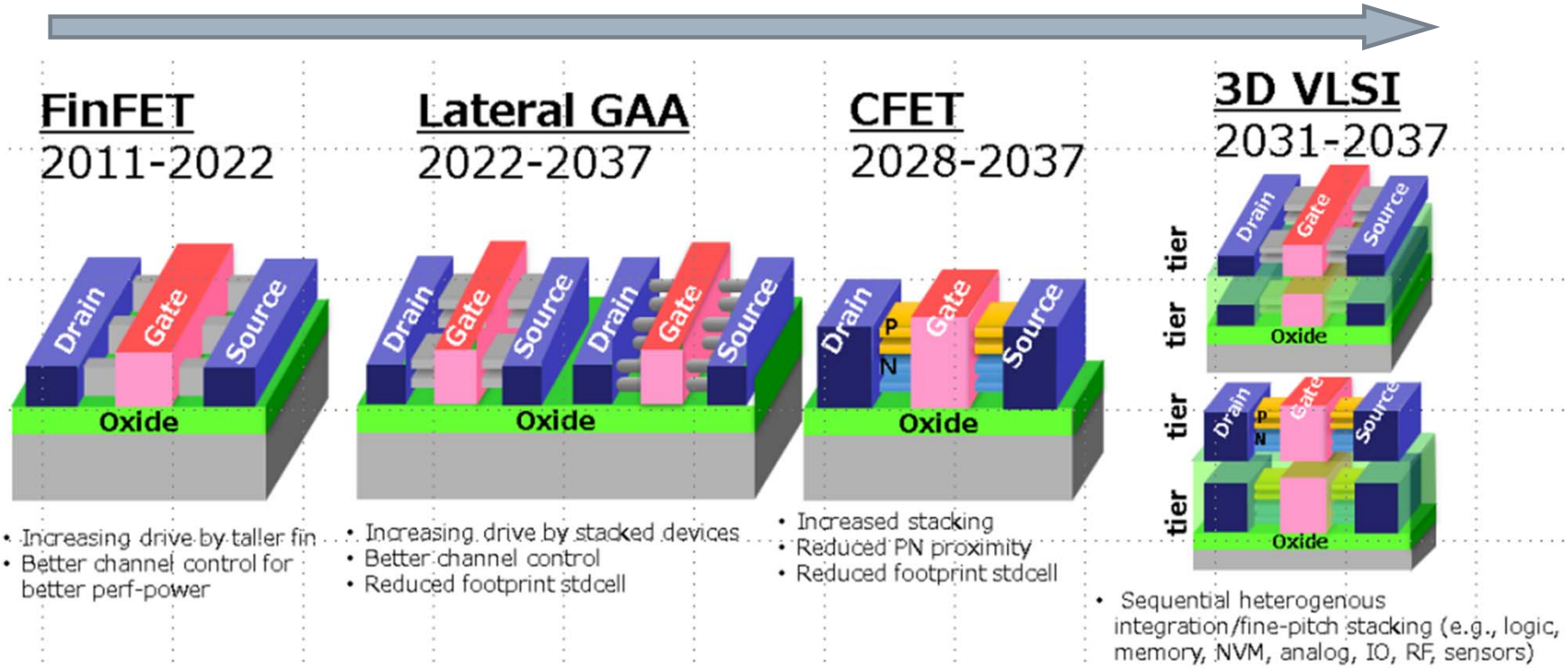
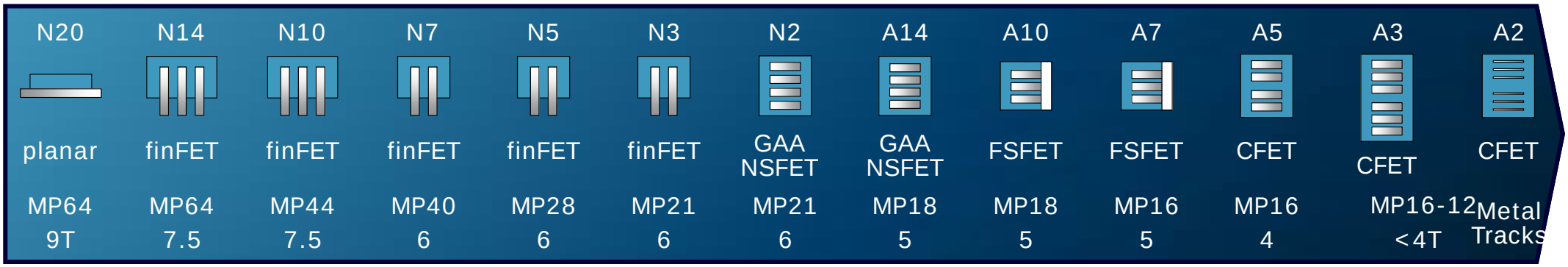
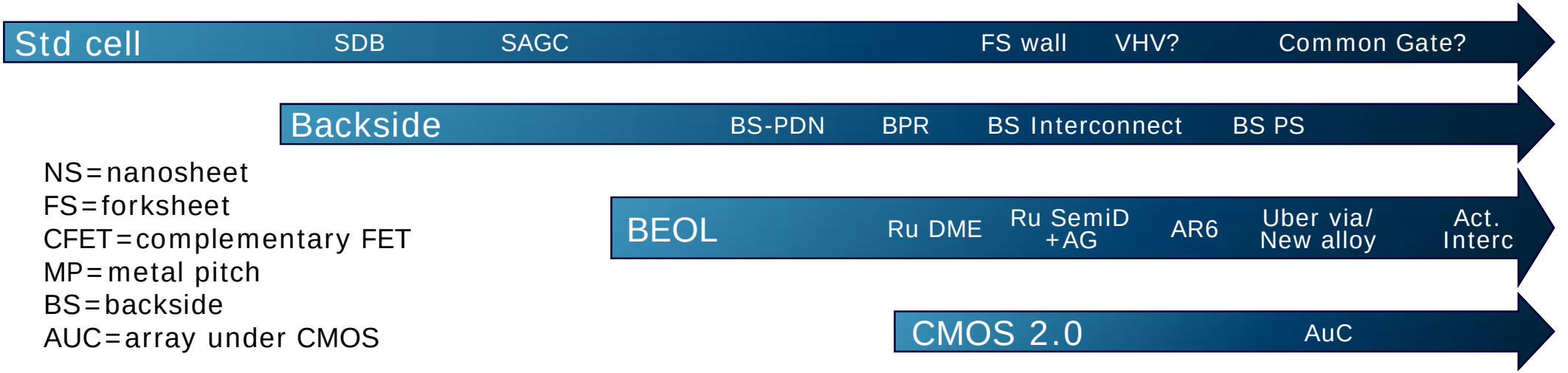


Figure MM-4 Evolution of device architectures in the IRDS More Moore roadmap

Source: 2023 IRDS Roadmap

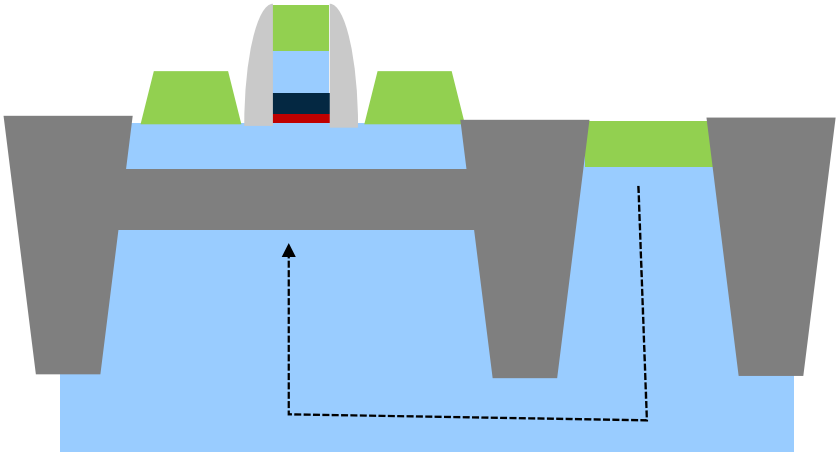
Logic scaling roadmap



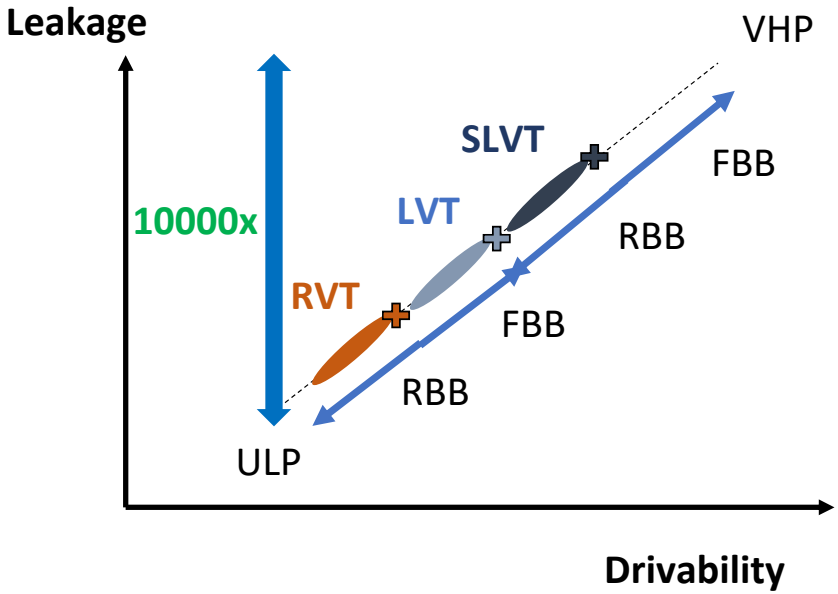


NS=nanosheet
FS=forksheet
CFET=complementary FET
MP=metal pitch
BS=backside
AUC=array under CMOS

FD-SOI Technology

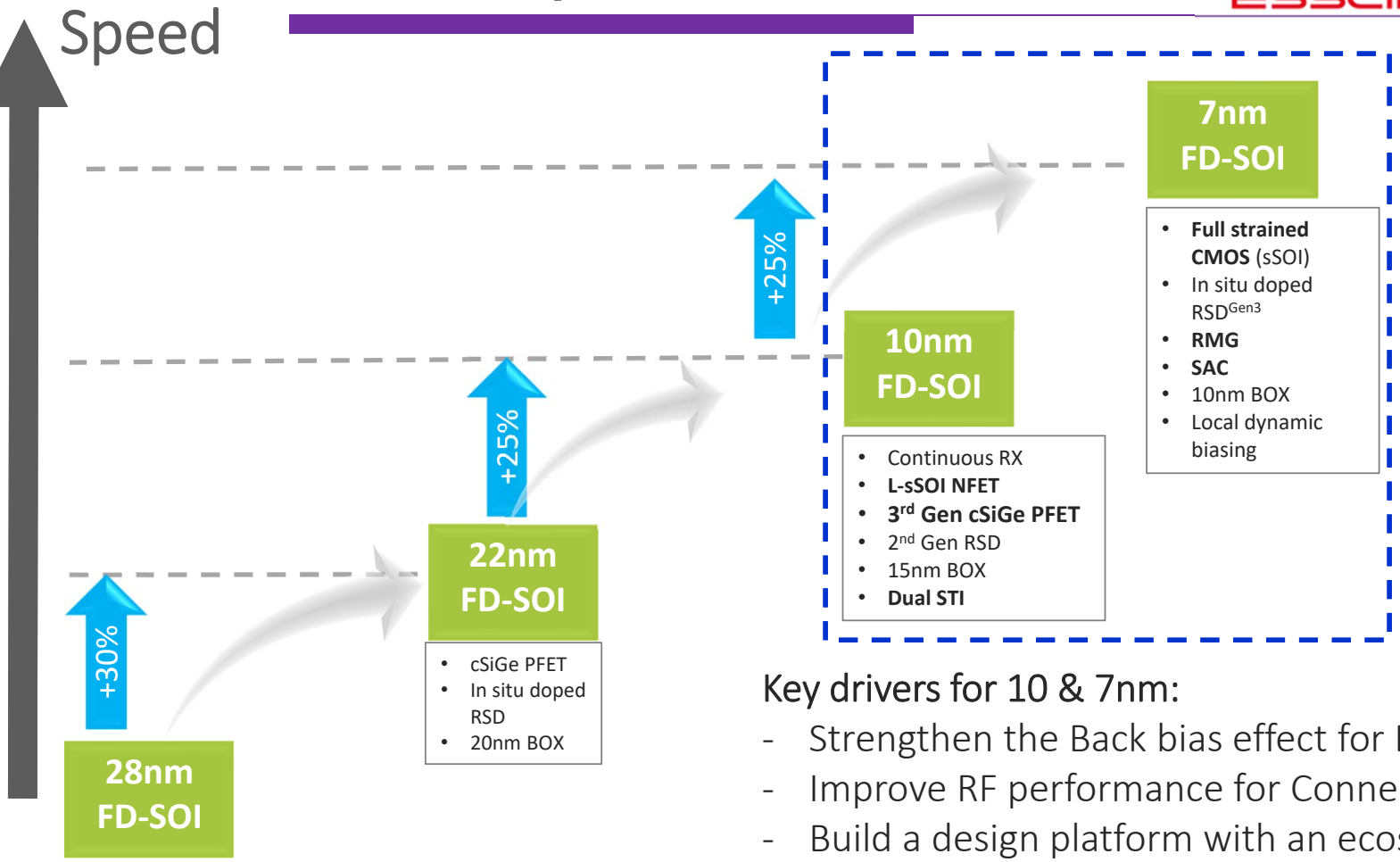


	22FDX	14nm FInFET	28nm Bulk	45nm PDSOI
f_T n-FET [GHz]	347	314	310	296
f_{max} n-FET [GHz]	371	180	161	342
f_T p-FET [GHz]	242 275 (mmWave)	285	185	-
f_{max} p-FET [GHz]	288 299 (mmWave)	140	104	-



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

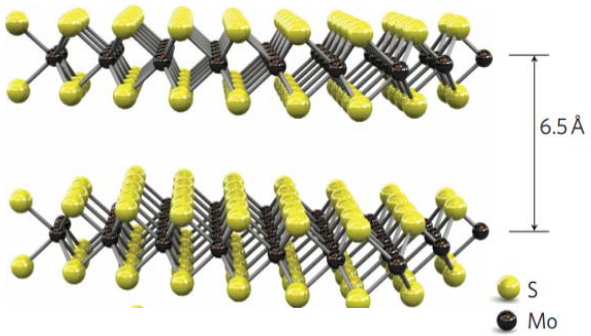
FD-SOI Roadmap



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

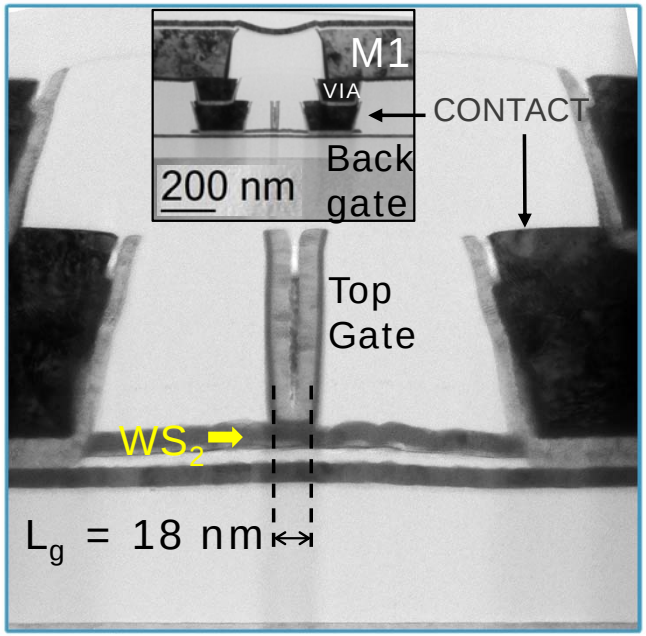
2D Atomic Channels: Next generation logic devices



MoS_2

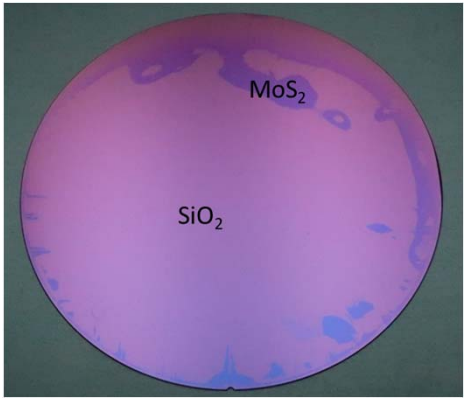
Monolayer channel thickness enables gate length scaling while keeping high performance

300mm Flow



I. Asselberghs et al, IEDM 2020

Layer Transfer:

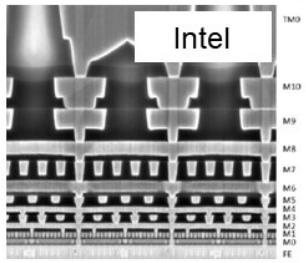


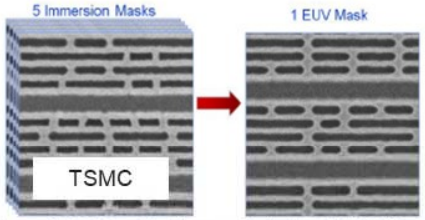
CEA-Leti, un-published

Industry BEOL trends

~36-40nm BEOL pitch

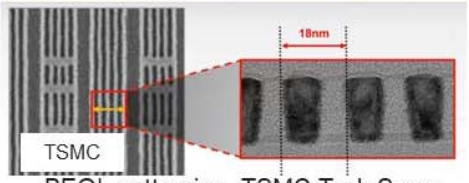
10nm technology, C. Auth et al. IEDM 2017





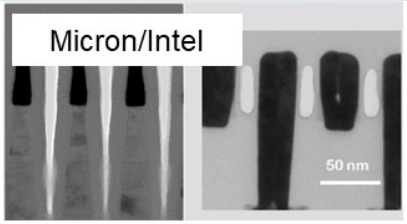
5nm CMOS, G. Yeap et al. IEDM2019

<20nm pitch

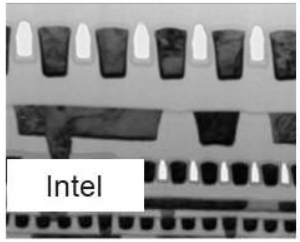


BEOL patterning, TSMC Tech Symp. 2020

AGs in memory & logic

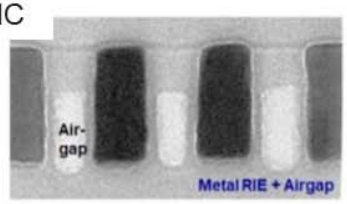


25nm NAND using air gaps K. Prall/K. Parat IEDM2010



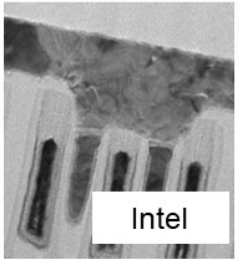
K. Fisher et al., 14nm, IITC2015

TSMC

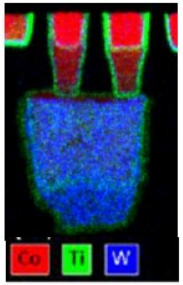


Y.J. Mii et al. Keynote, VLS022

Co interconnects



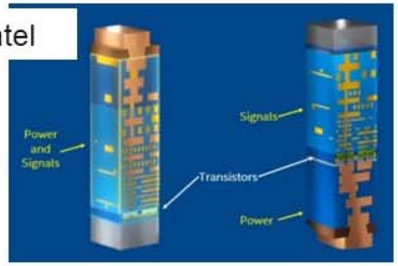
Self- aligned contact, C. Auth et al. IEDM 2017



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

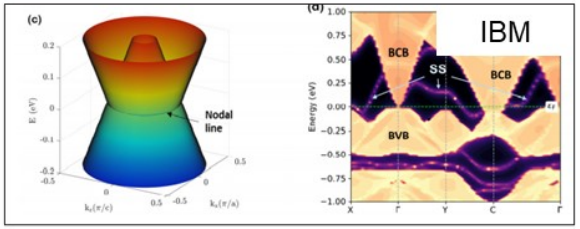
Backside Power Delivery

Intel

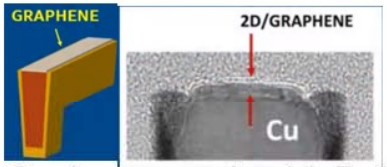


M.C. Mayberry, Keynote IITC 2020

New materials on the horizon

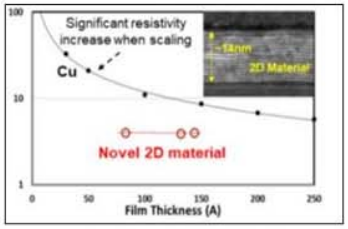


Topological semi-metal, C.T. Chen et al. IEDM2020



Intel

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



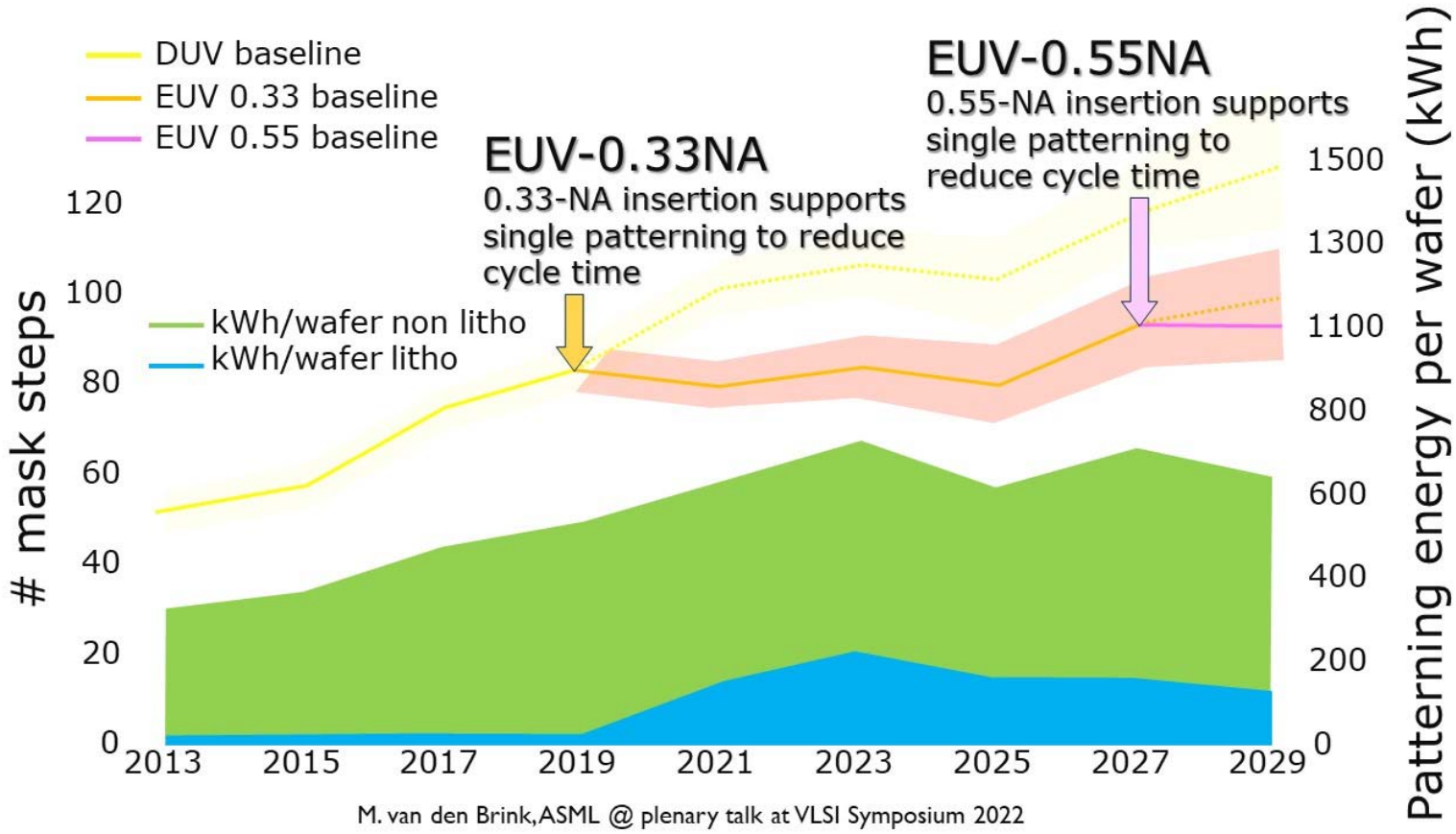
TSMC

New conductor, Y.J Mii Keynote VLSI2022

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

Courtesy: Zsolt Tokei (imec)

EUV lithography key enabler for scaling



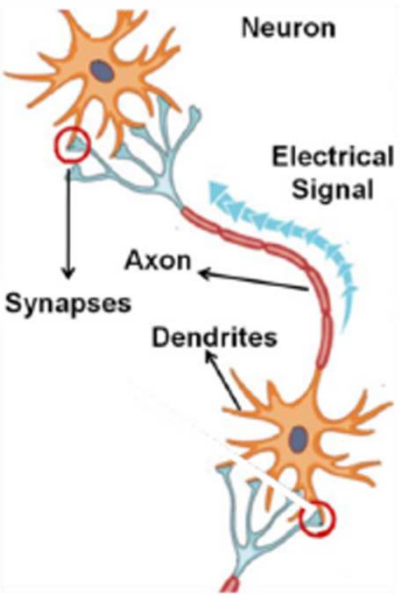
Emerging Non-Volatile Memories

	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)**

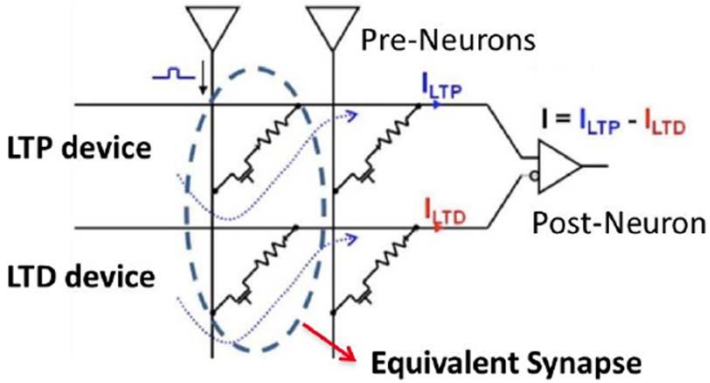
Memory activity focus on embedded NVM for NOR flash replacement

Neuromorphic based RRAM circuit

M. Suri et al, IEDM 2011.

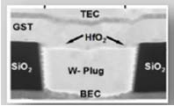


2 PCRAM Example:



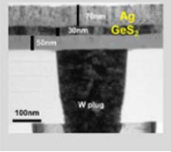
PCM

GST
GeTe
GST + HfO_2



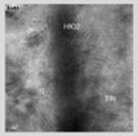
CBRAM

Ag / GeS_2

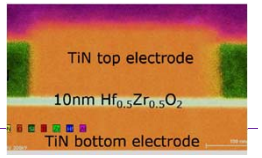


OXRAM

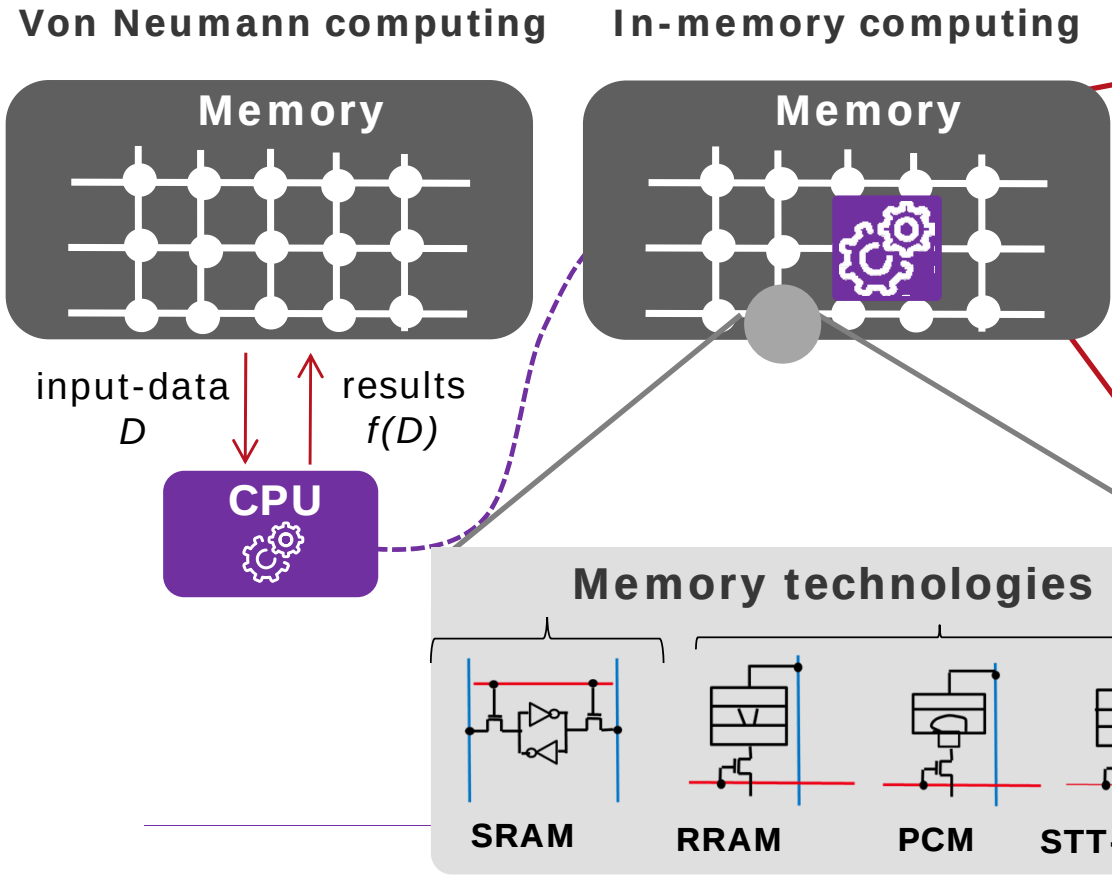
$TiN/HfO_2/Ti/TiN$



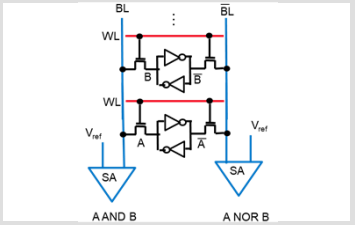
FeRAM



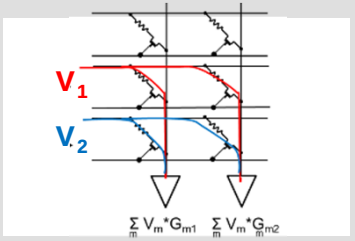
In memory computing



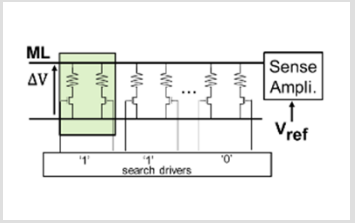
Logic and arithmetic operations



Digital operations using SRAM



Analog: multiply and accumulate

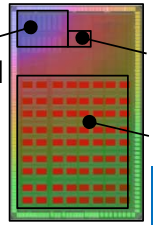


Searching / matching (CAM)

Near- & in-memory computing

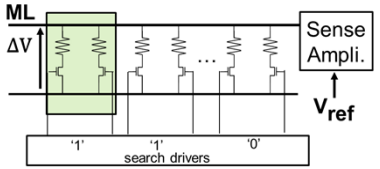
Bayesian Neural Networks

Spiking Neural Network



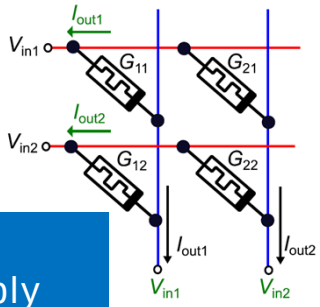
Energy-efficient NN accelerator

RRAM for near- and in-memory computing

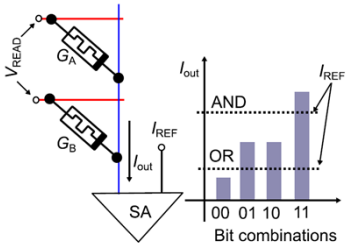


Search/match operations (CAM)

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



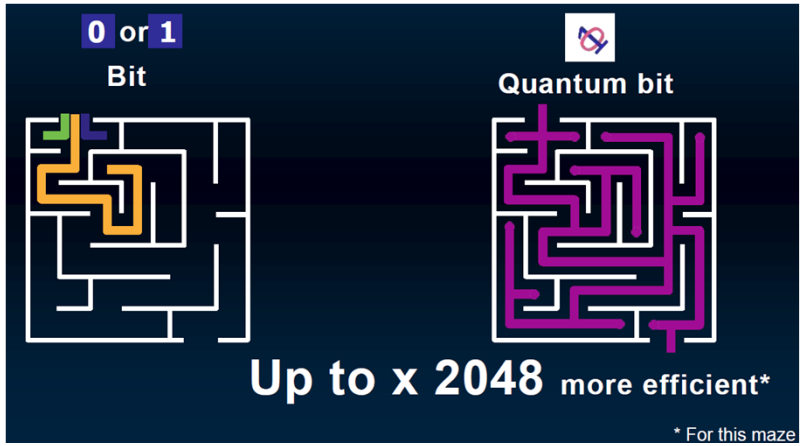
Deep learning, matrix-vector multiply



In-memory logic

From bits to q-bits

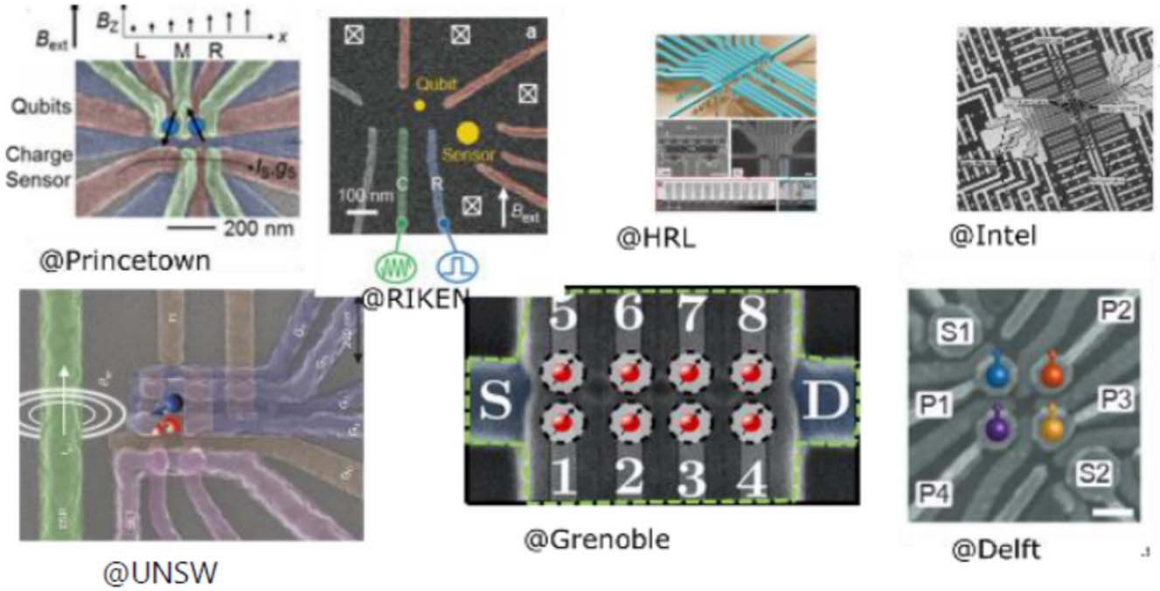
Quantum Physics to compute

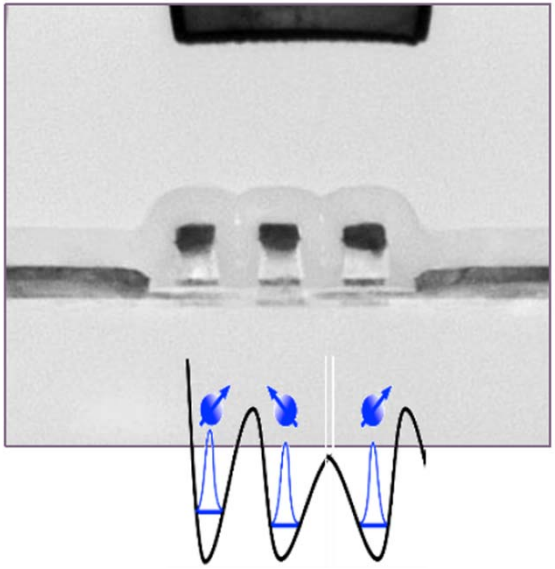


	Superconductor	Si spin qubit	Trapped ion	Photon
Size*	(100µm)²	(100nm)²	(1mm)²	~(100µm)²
1qubit fidelity	99.96%	99.93%	99.98%	
2qubit fidelity	~99.3%	>99%	99.9%	50% (measurement) 98% (gates)
Speed**	12-400 ns	~1 µs	100 µs	1 ms
Variability	3%	0.1%-0.5%	0.01%	0.5%
T° of operation	15mK	1K	10K	4K/10K
Entangled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

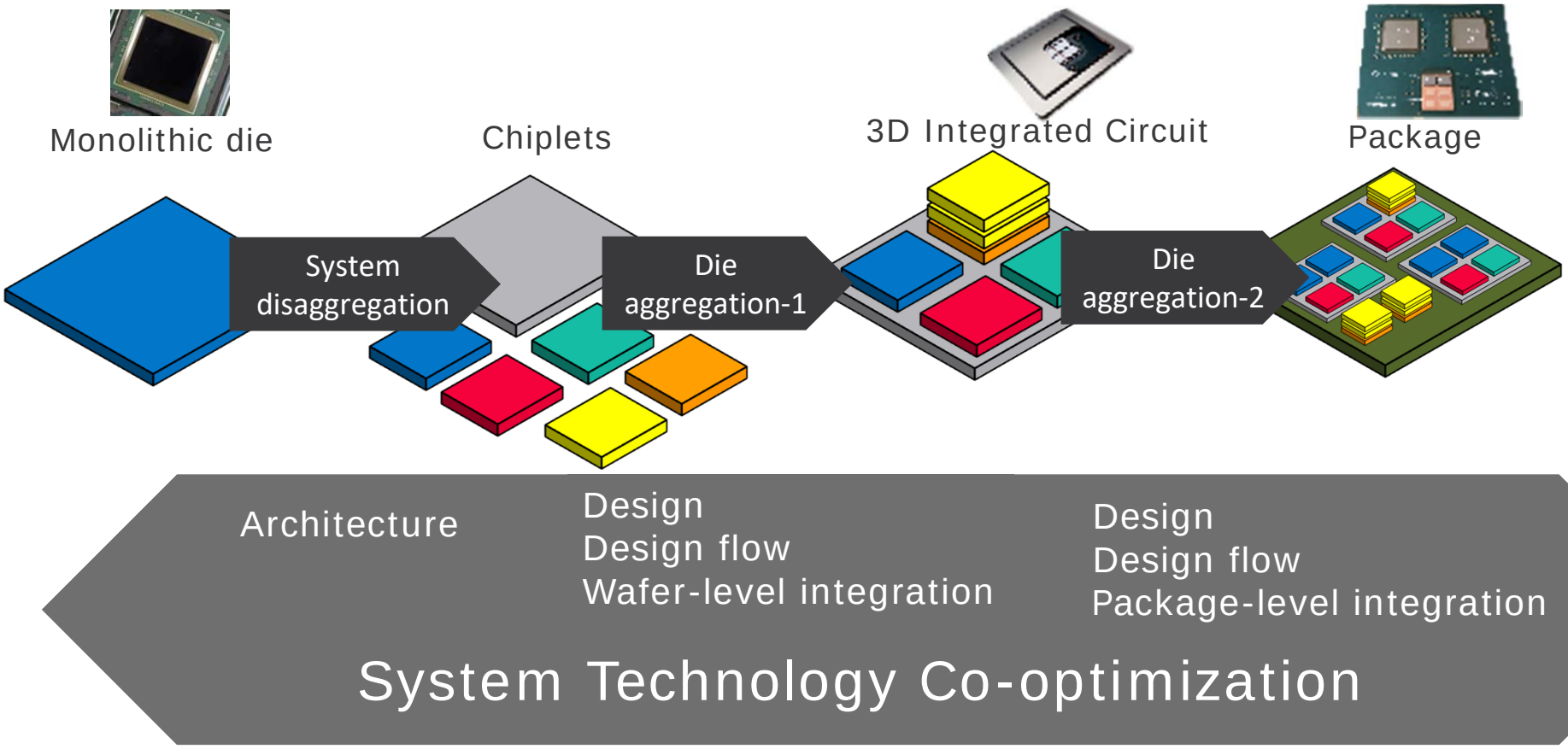
From bits to q-bits

- Quantum Physics to compute



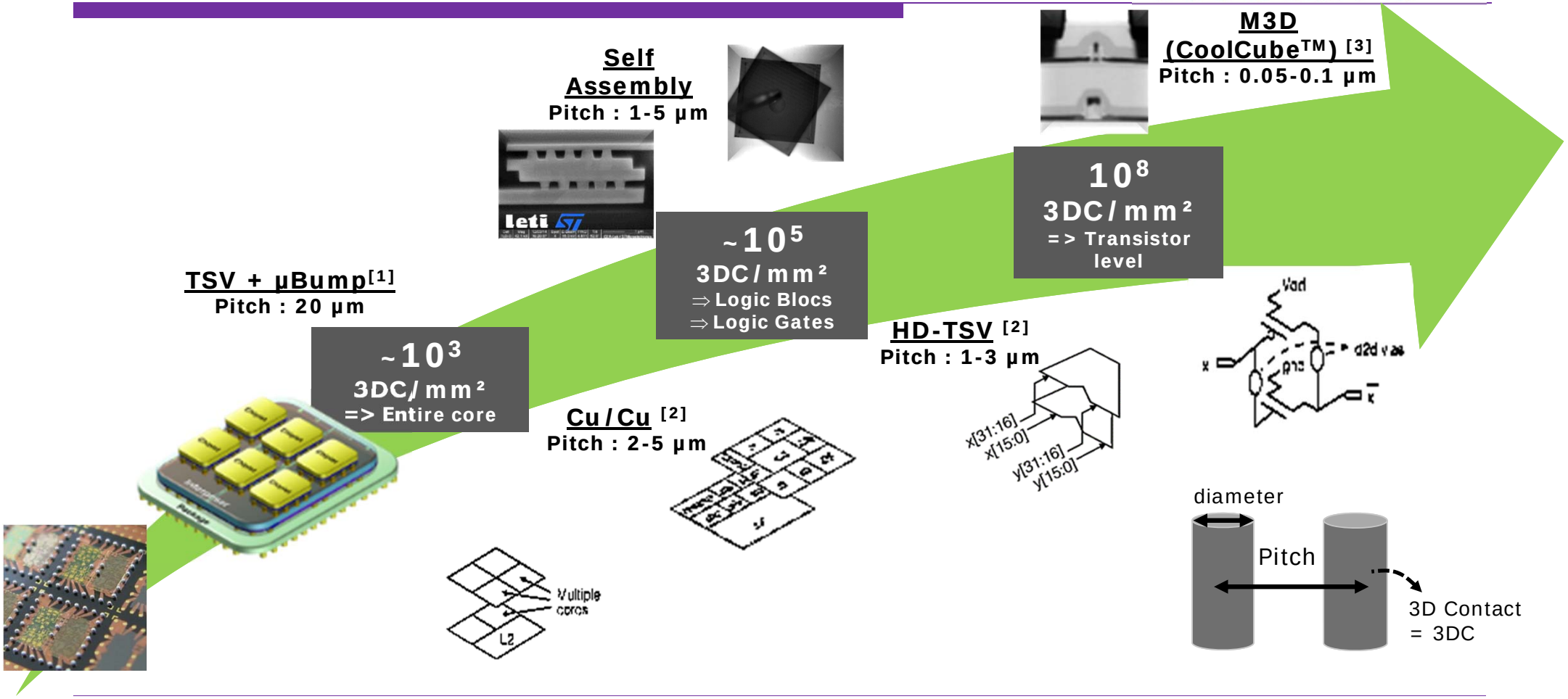


Chiplets: the new IC design paradigm



Up to 100x gain on Power Efficiency with 3D

3D: from packaging to monolithic

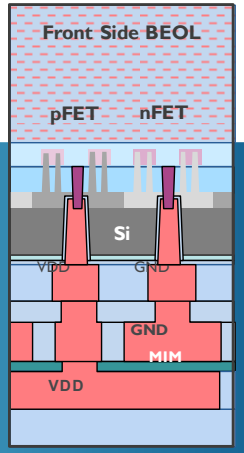


[1] Cheramy, S., et al. “Advanced Silicon Interposer”, C2MI Workshop, 2015
[2] Patti, B., “Implementing 2.5D and 3D Devices”, In AIDA workshop in Roma, 2013
[3] Batude, P., et al. "3DVLSI with CoolCube process: An alternative path to scaling ." VLSI technology symposium 2015

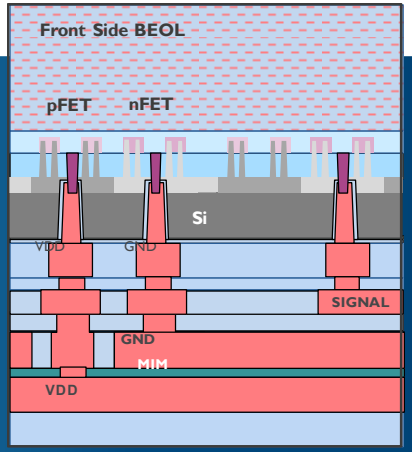
Towards a functional backside

J. Ryckaert, ITF Japan 2022

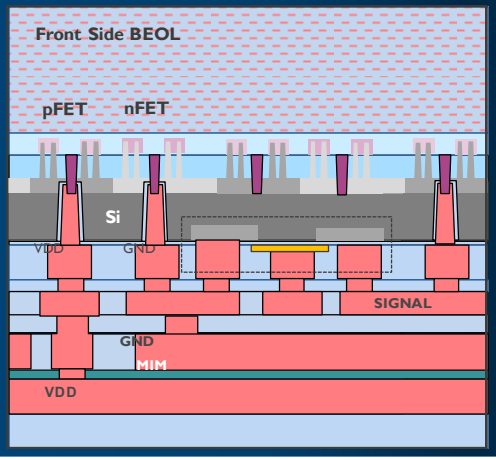
Backside Power Delivery



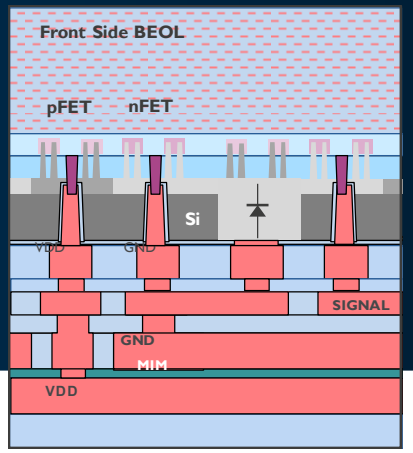
Backside Global Interconnect



Backside Devices

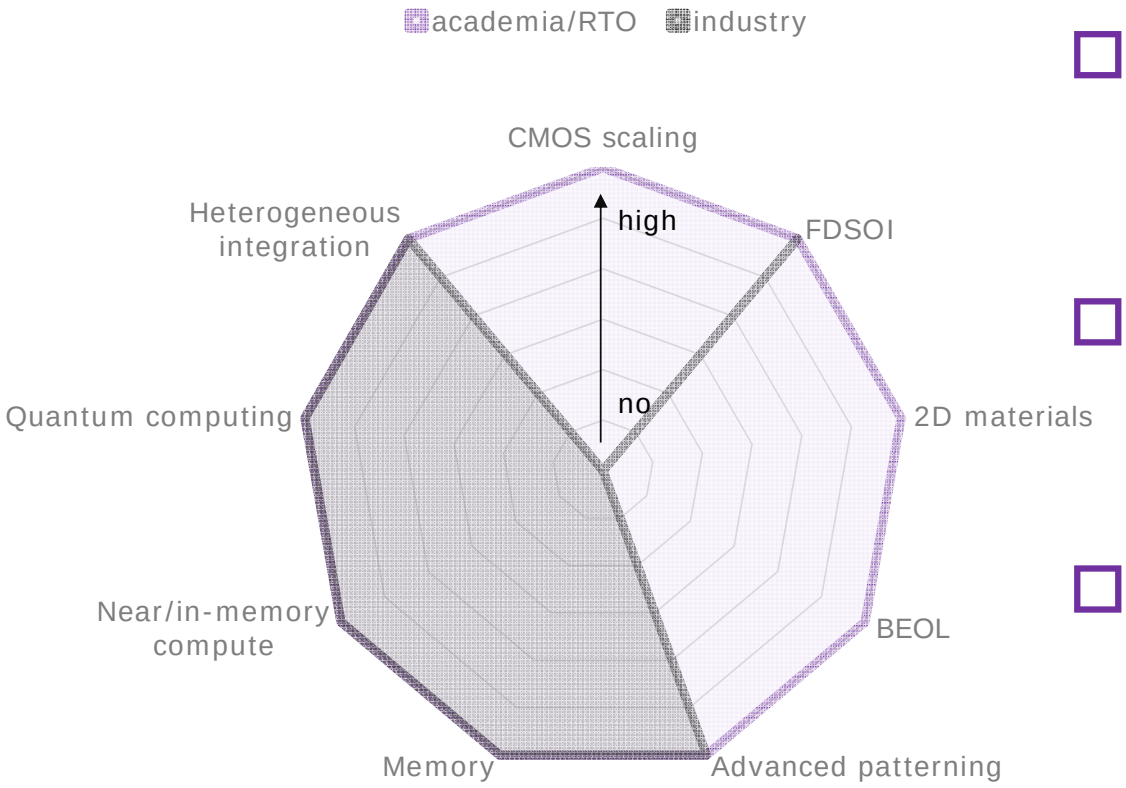


Device Backside Extension



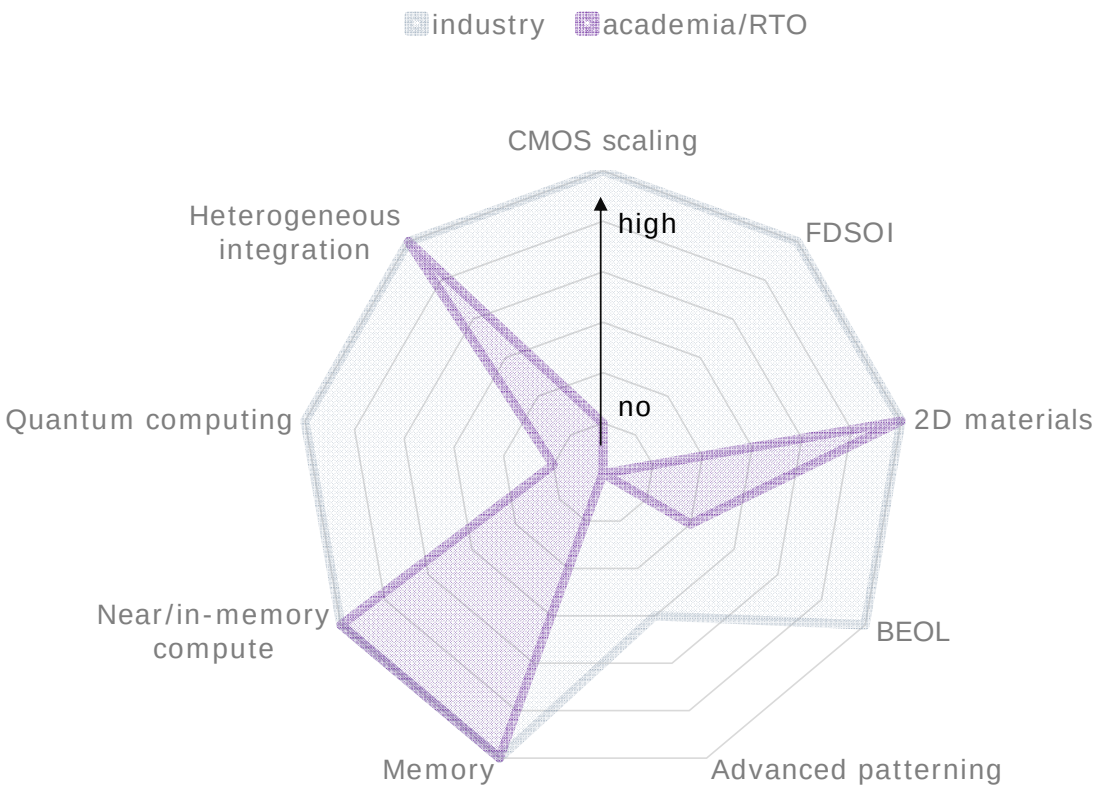
Enhancing system performance by migrating system functions to the backside

EU and non-EU actors - EU



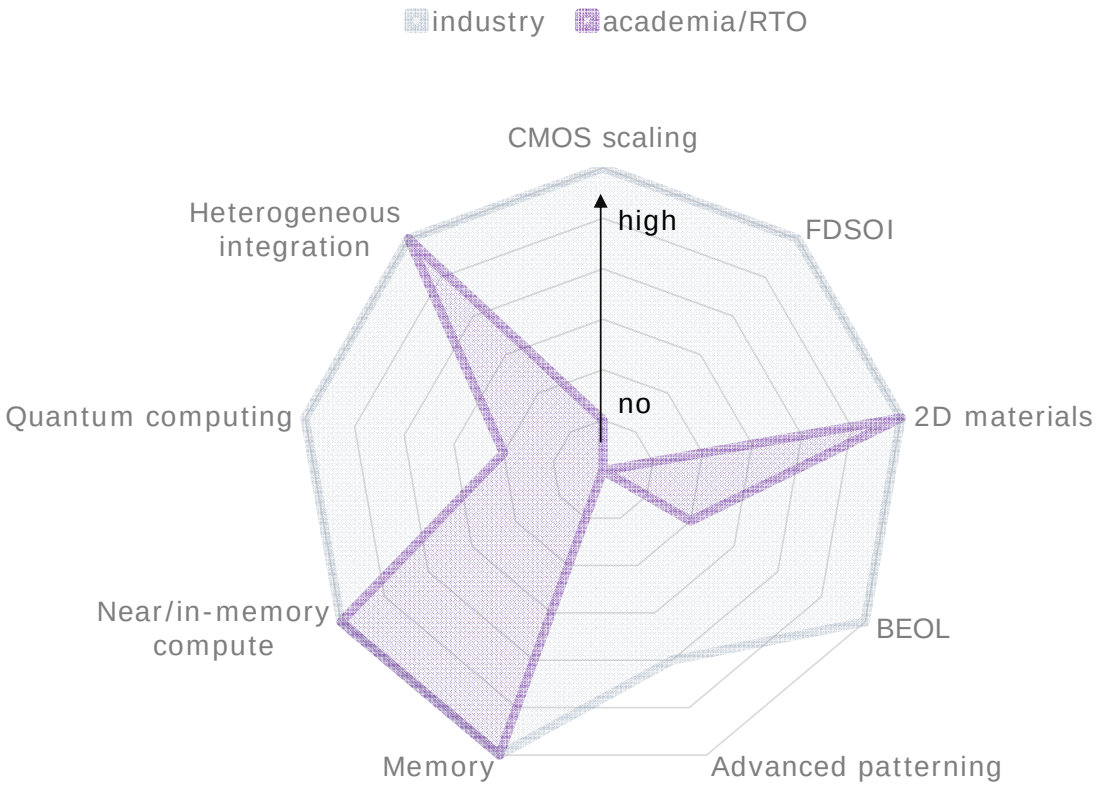
- R&D very strong in all areas of compute
- Unique strong position in EUV lithography
- In general, industrial EU players lacking to take up R&D

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

- Europe is very strong in R&D in all advanced compute areas
- Manufacturing: starting to catch up- Intel and TSMC's initiatives in Germany
- Lack of Fabless companies that drive the needs in terms of advanced compute
- Strong initiatives in EU in start-up company creation. Maybe a solution?