ESSERC 2024

SiNANO-ICOS-INPACE Workshop

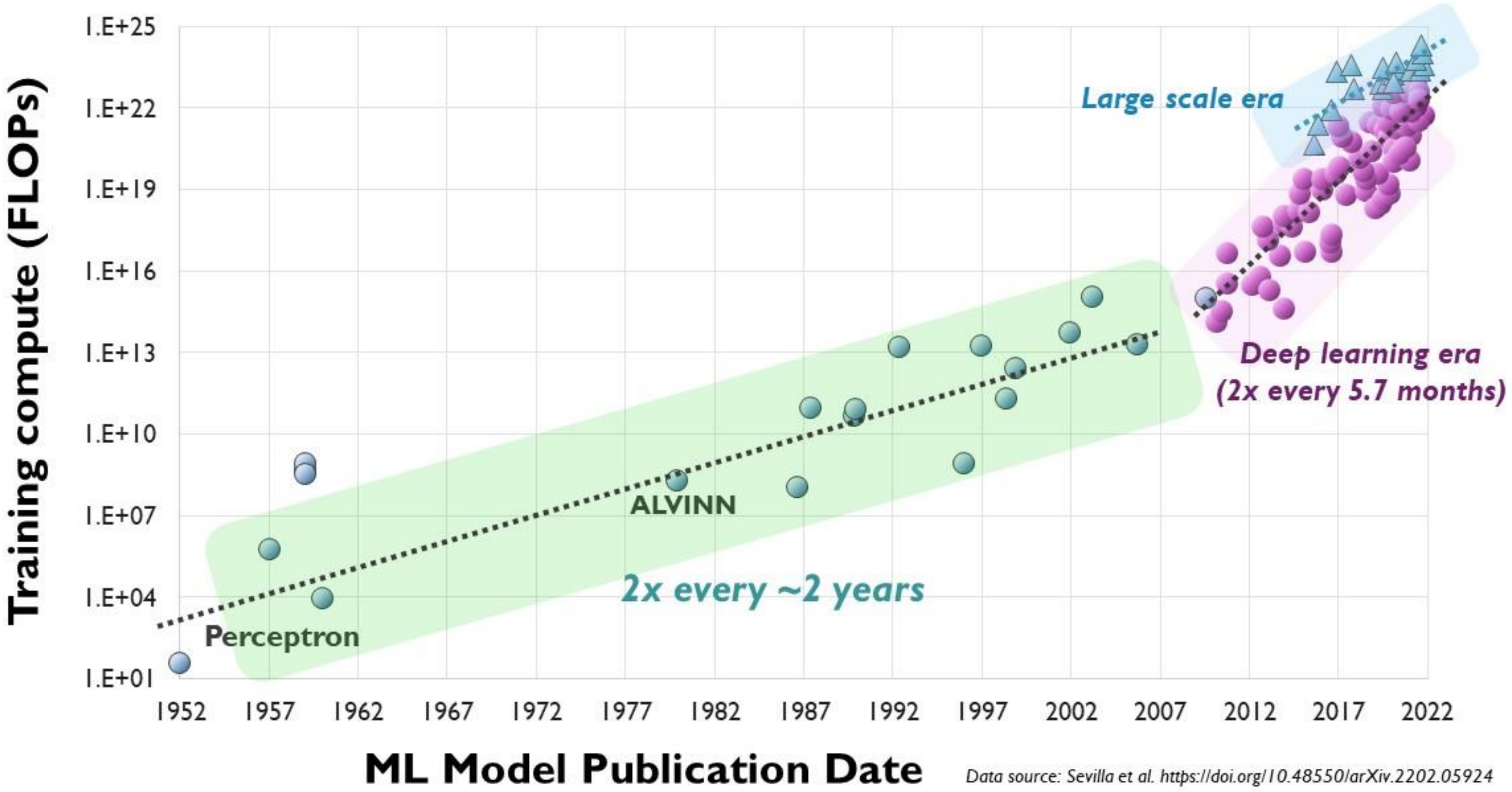
"Emerging technologies in Advanced Computation, Advanced Functionalities, Ground-breaking Technologies: Impact on International Cooperation"

New device architectures for advanced compute

Nadine Collaert
Imec, Belgium
collaert@imec.be

Leuven, September 9, 2024

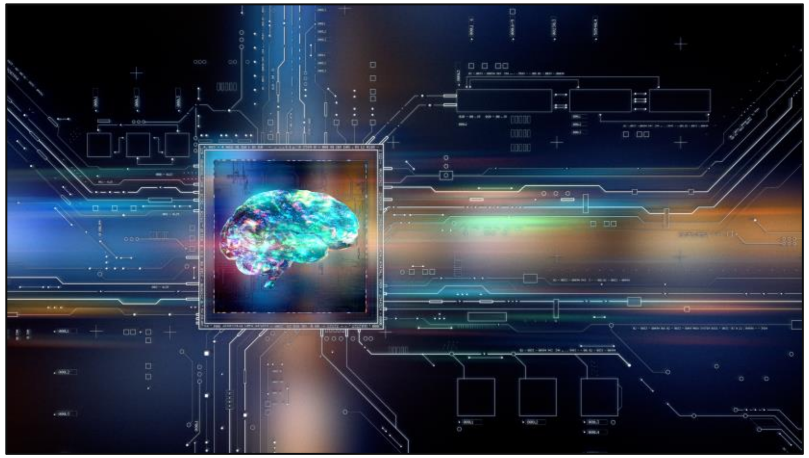
Compute needs for Machine Learning (ML) continue to grow



Data source: Sevilla et al. https://doi.org/10.48550/arXiv.2202.05924

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

Challenges for future compute systems

CMOS and DRAM PPAC

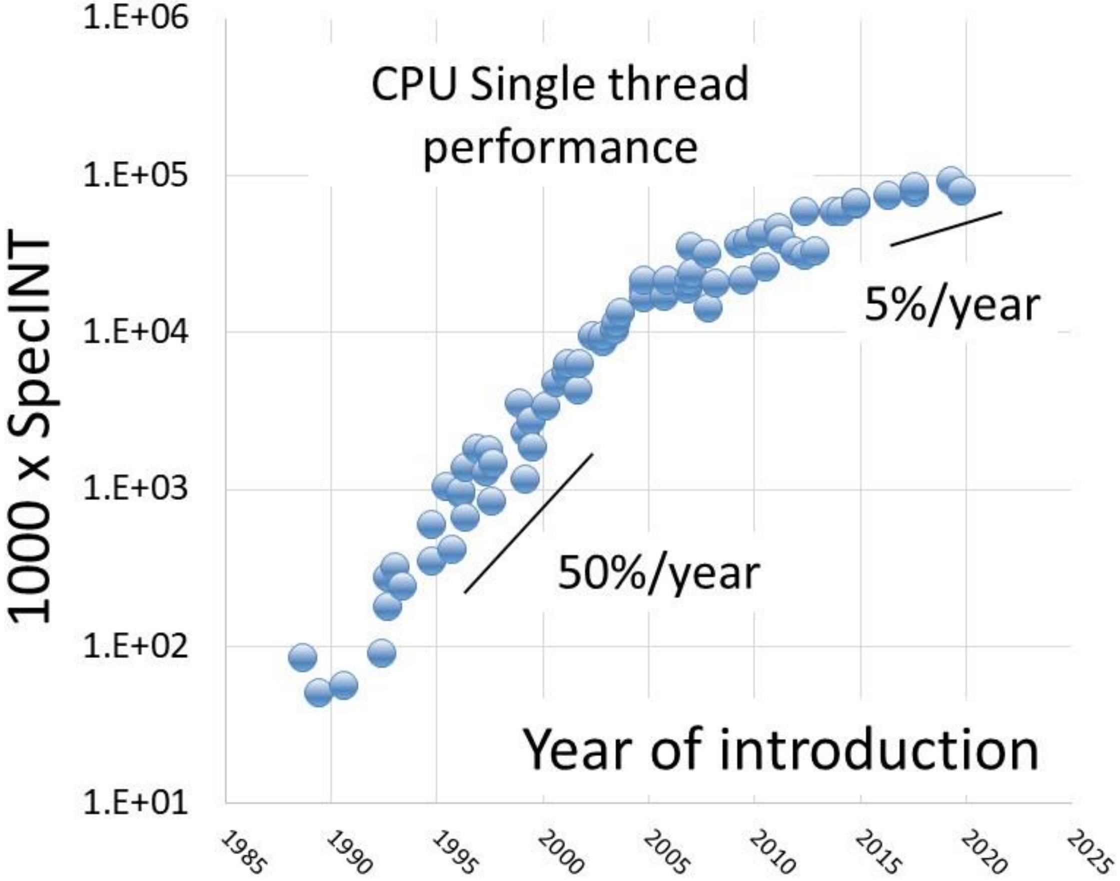
Memory Wall

Power Wall

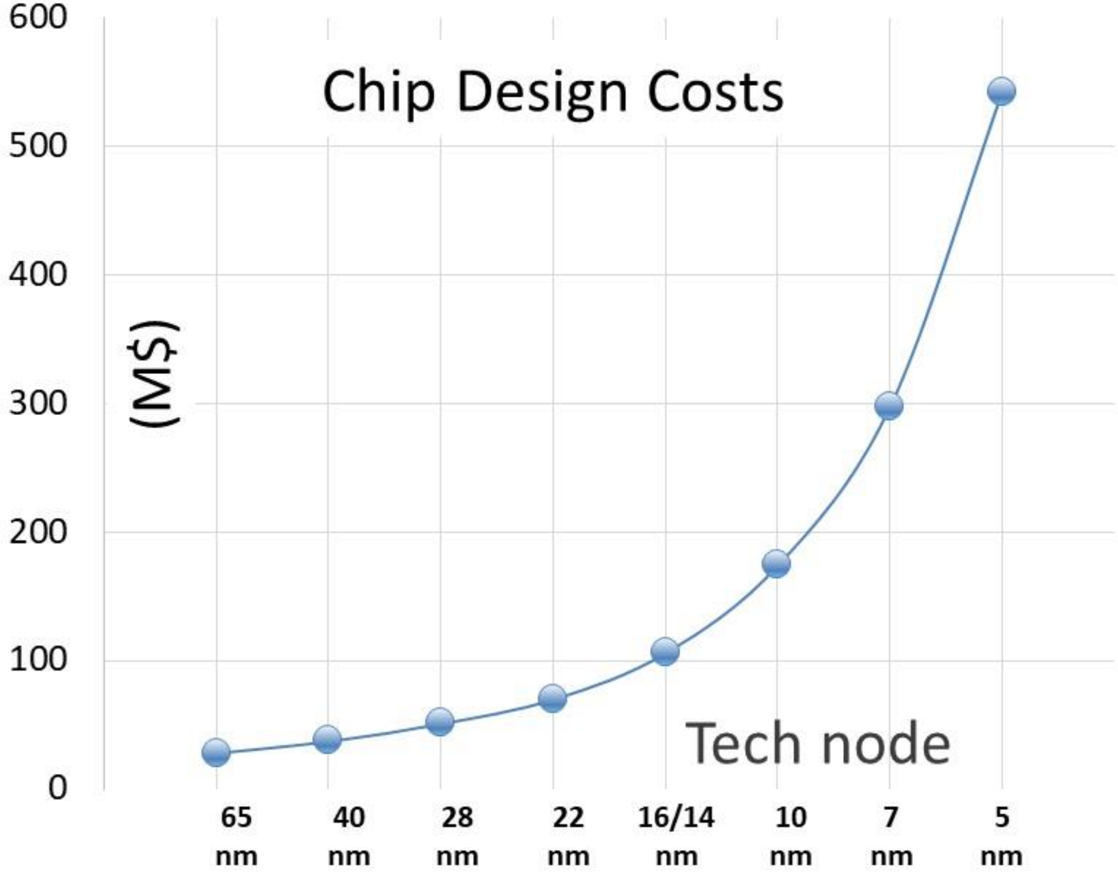
Sustainable Manufacturing

PPAC=Power-Performance-Area-Cost

Slowdown in system performance and increasing costs

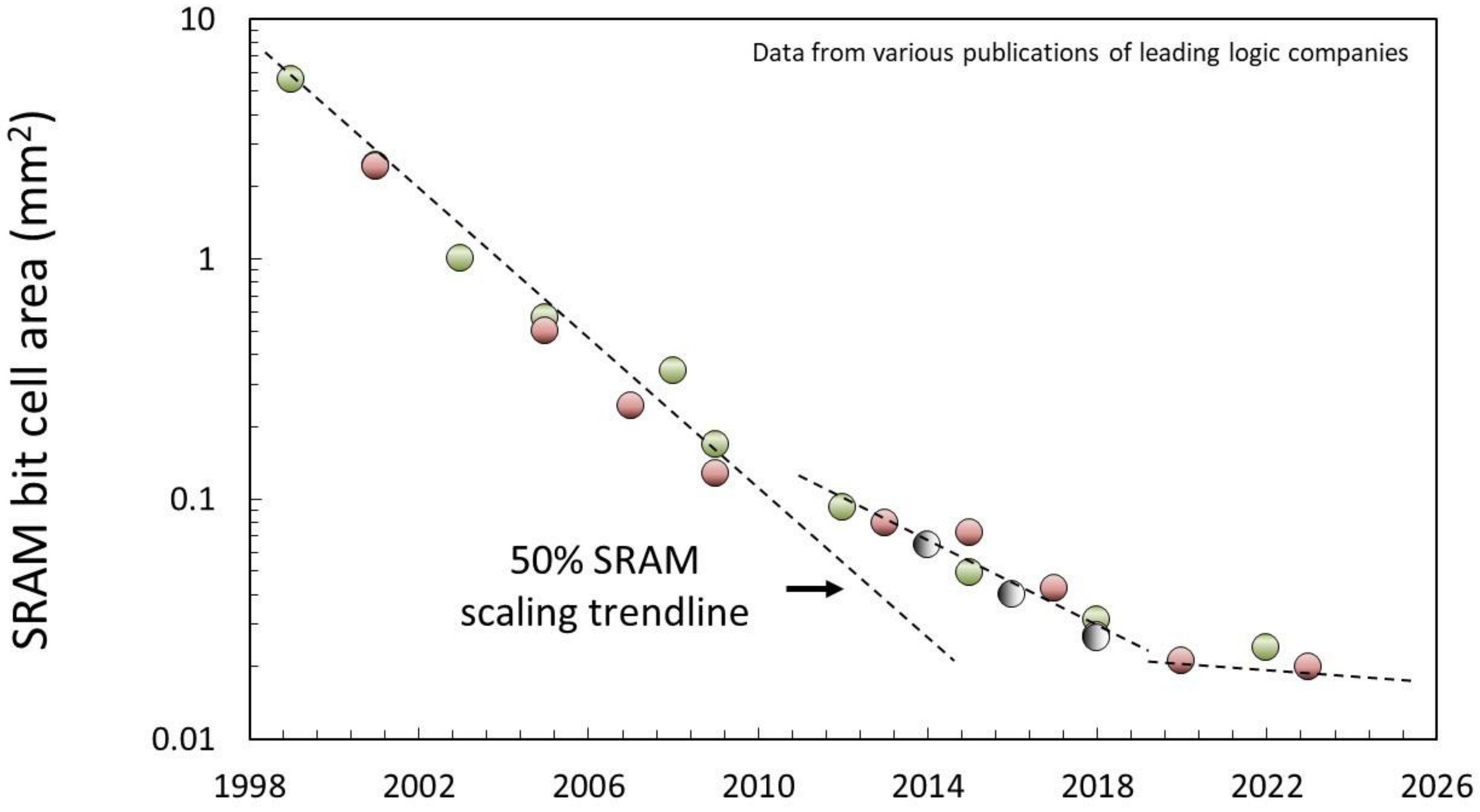


Based on original data plotted by M. Horowitz, F. Labonte, O. Shachan, K. Olukotun, L. Hammond, C. Batten. Additional data compiled by K. Rupp

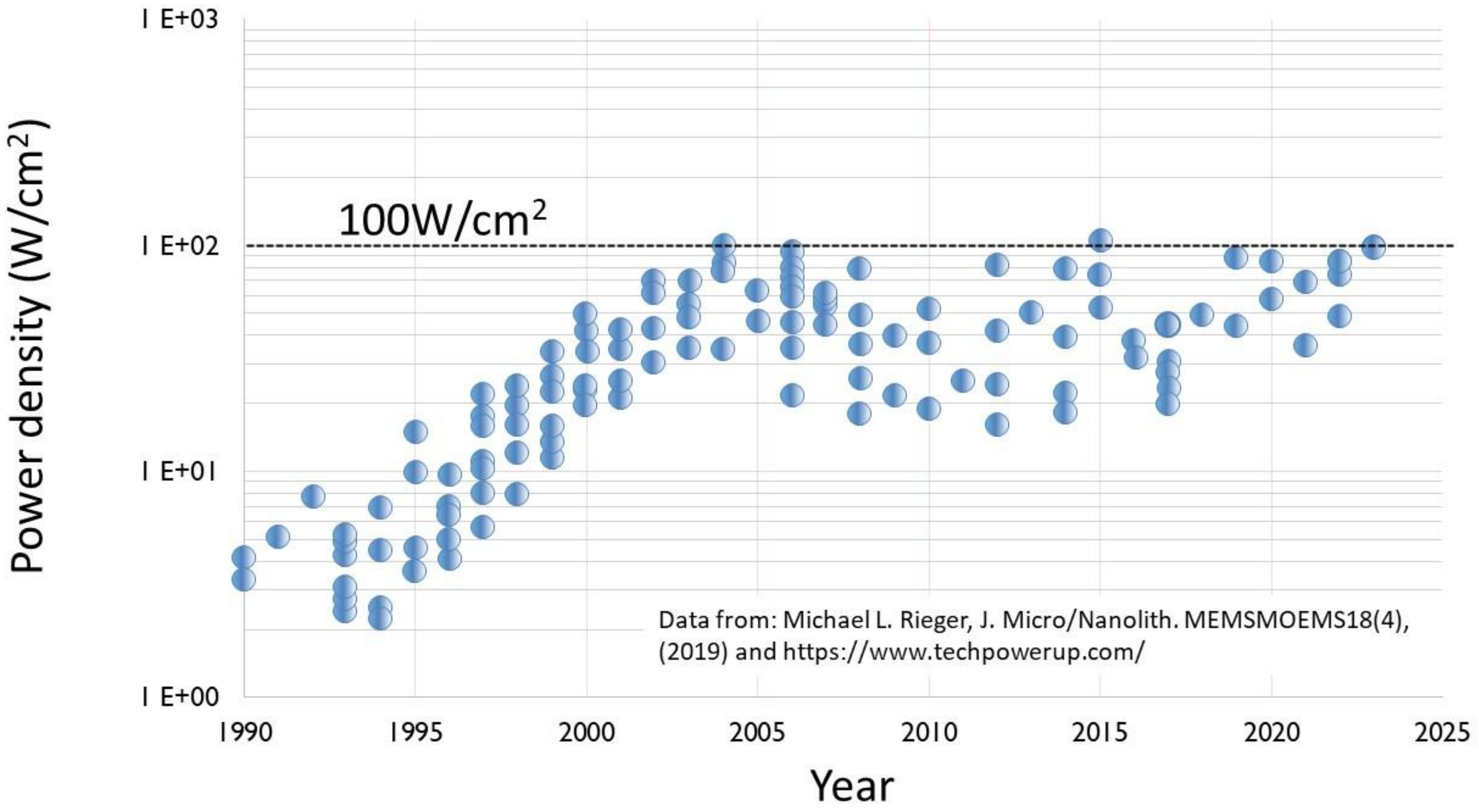


Source: AI Chips and why they matter", S. Khan and A. Mann, 2020

Slowdown of SRAM scaling

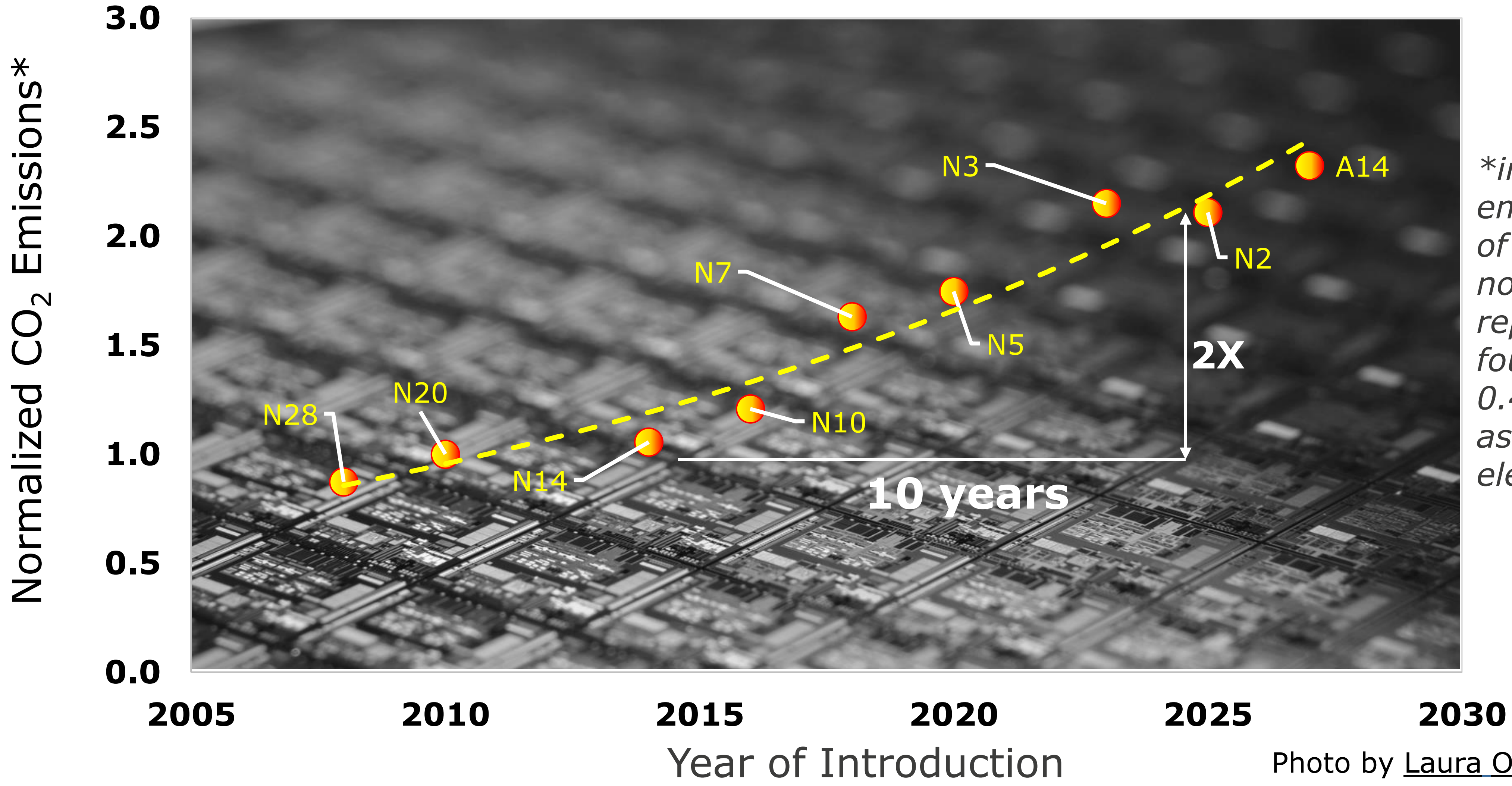


Chip Cooling Technologies Limit Power Density of Chips



Data from: Michael L. Rieger, J. Micro/Nanolith. MEMSMOEMS18(4), (2019) and https://www.techpowerup.com/

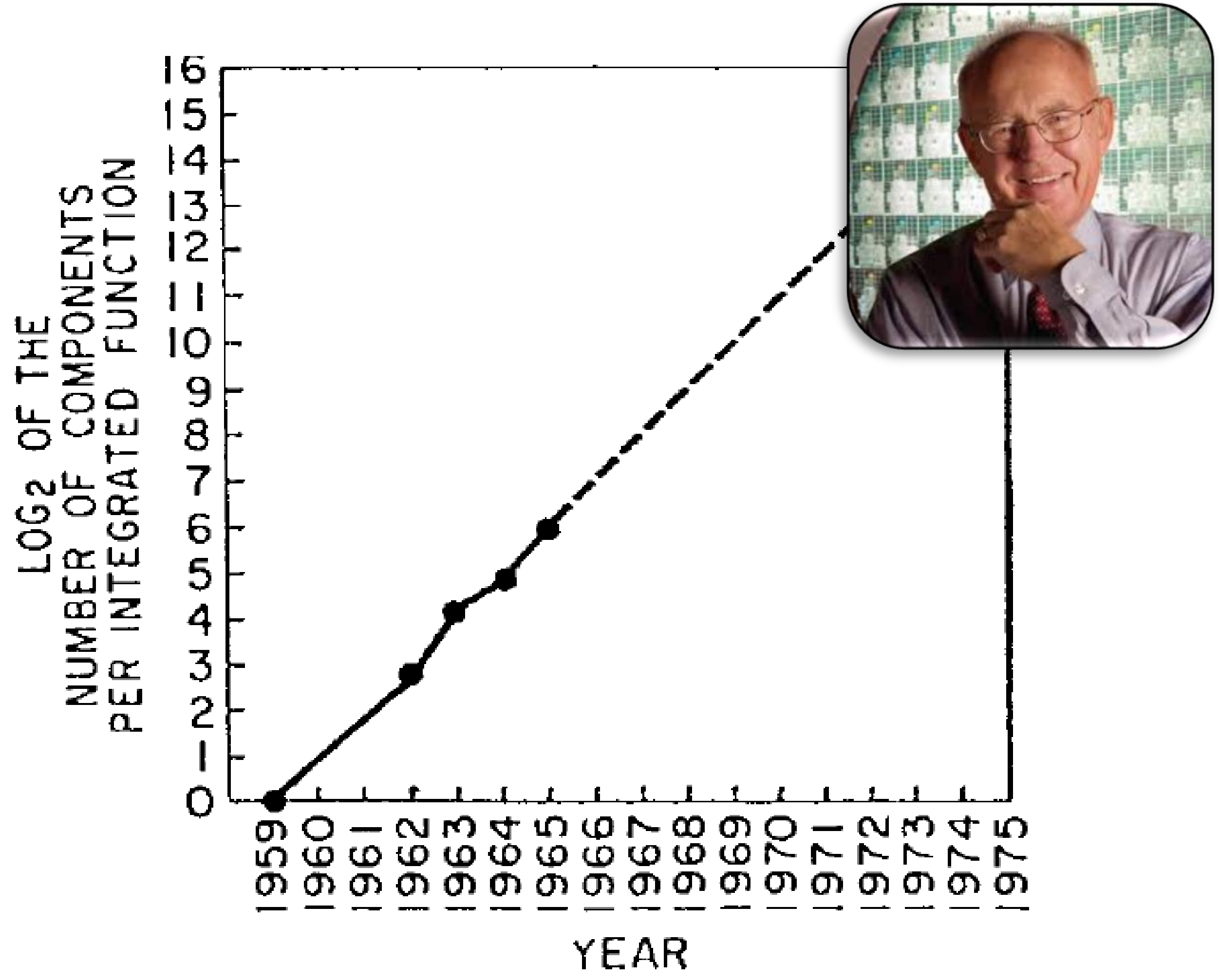
Increased Carbon Emissions for Recent Technology Nodes



imec.netzero: emissions estimate of imec process nodes representative of foundry nodes. 0.49kgCO_2eq/kWh assumption for electricity

Photo by Laura Ockel on Unsplash

The economics of scaling



" Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years...." (1965)

- ☐ **P**OWER
- ☐ **P**ERFORMANCE
- ☐ **A**REA
- ☐ **C**OST
- ☐ **E**NVIRONMENTAL COST

New device architectures to fuel the roadmap

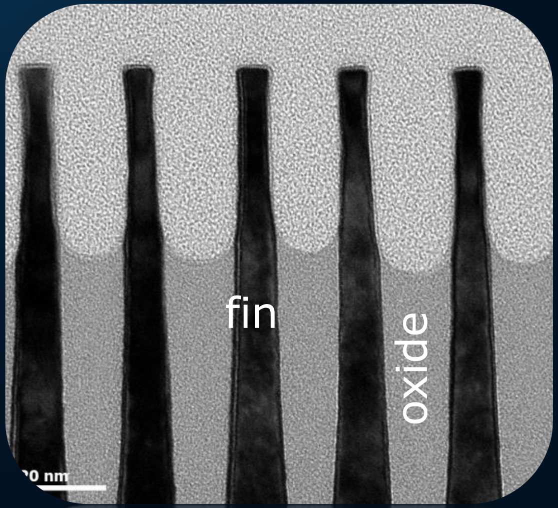
14-3nm

2nm-14Å

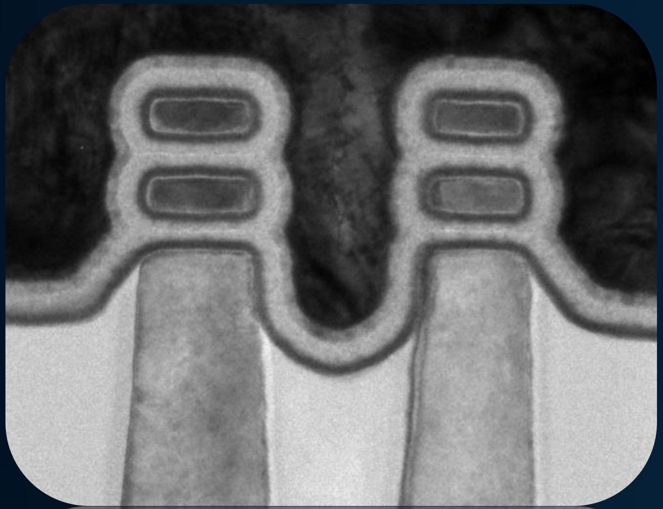
10-7Å

5-2Å

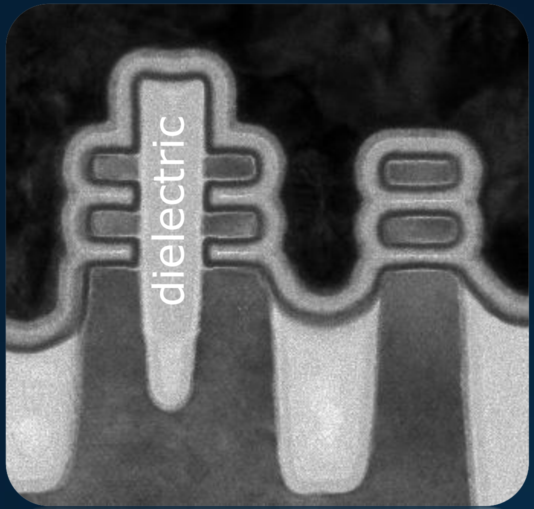
FinFET



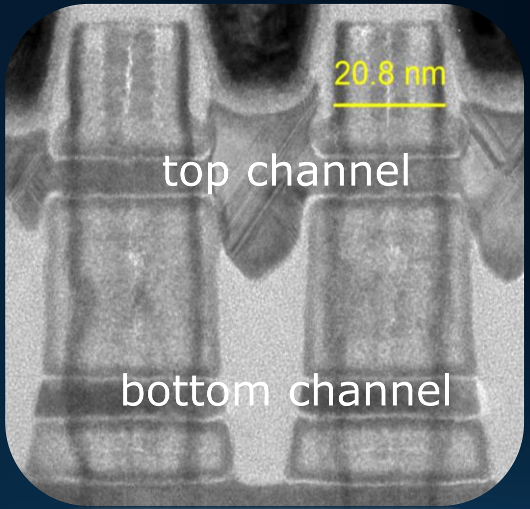
Nanosheets



Forksheets

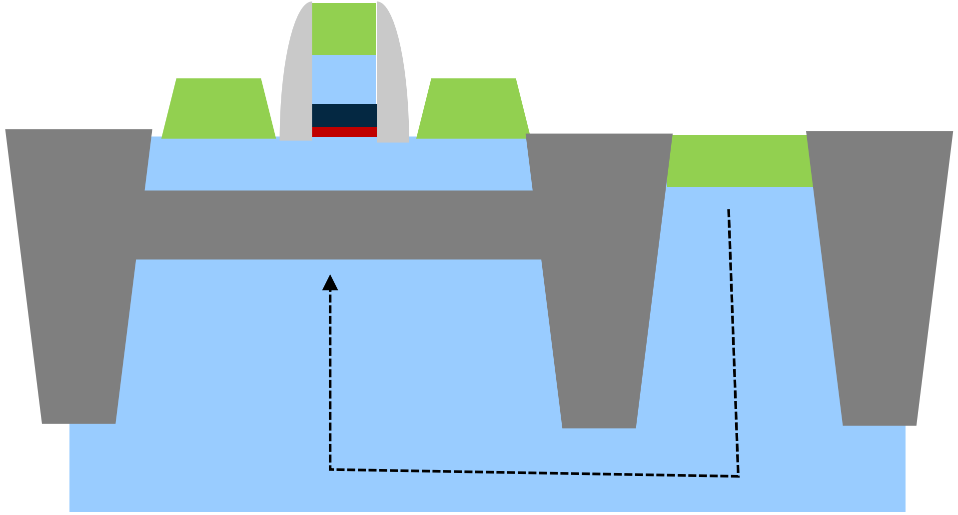


CFET

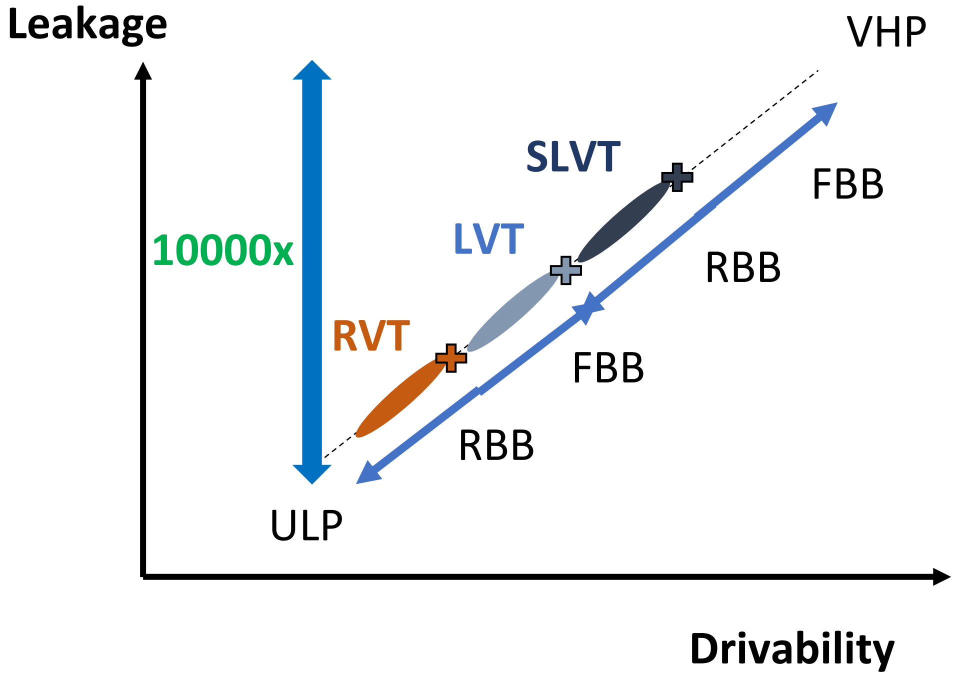


N. Collaert, "Advancements in IC Technologies: A look toward the future," in IEEE Solid-State Circuits Magazine, vol. 15, no. 3, pp. 80-86, 2023.

FD-SOI technology

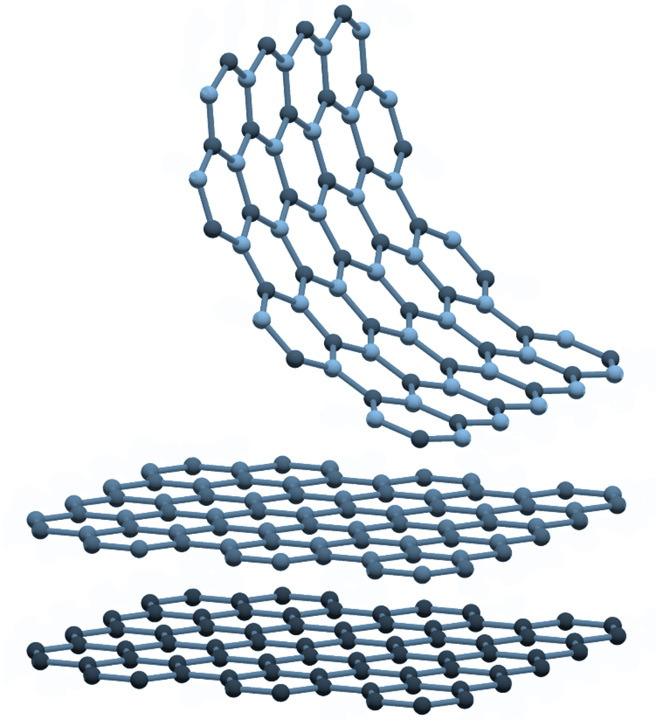


	22FDX	14nm FInFET	28nm Bulk	45nm PDSOI
f_T n-FET [GHz]	347	314	310	296
f_{max} n-FET [GHz]	371	180	161	342
f_T p-FET [GHz]	242 275 (mmWave)	285	185	-
f_{max} p-FET [GHz]	288 299 (mmWave)	140	104	-

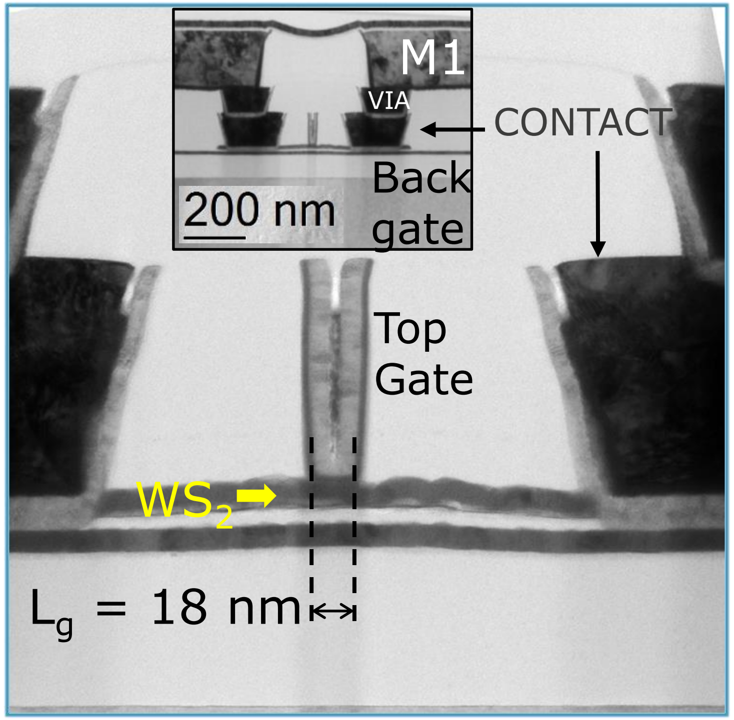


RBB: Reverse Body Bias
FBB: Forward Body Bias
ULP: Ultra-Low Power
VHP: Very High Performance

2D materials to fuel the roadmap

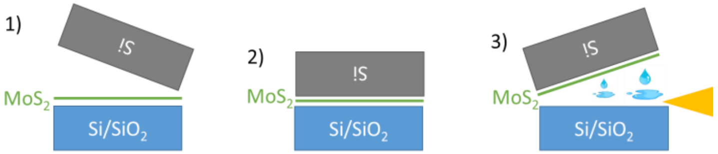


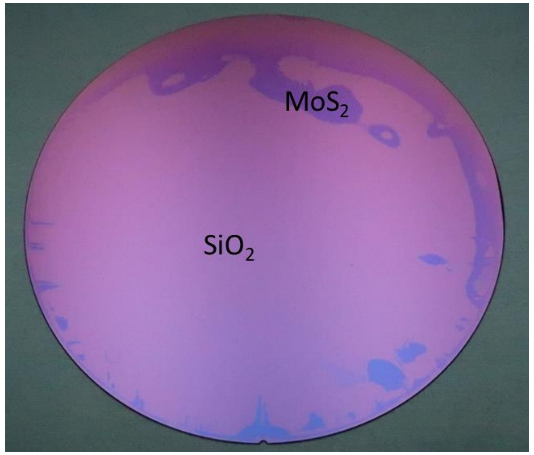
300mm Flow



I. Asselberghs et al, IEDM 2020

Layer Transfer:

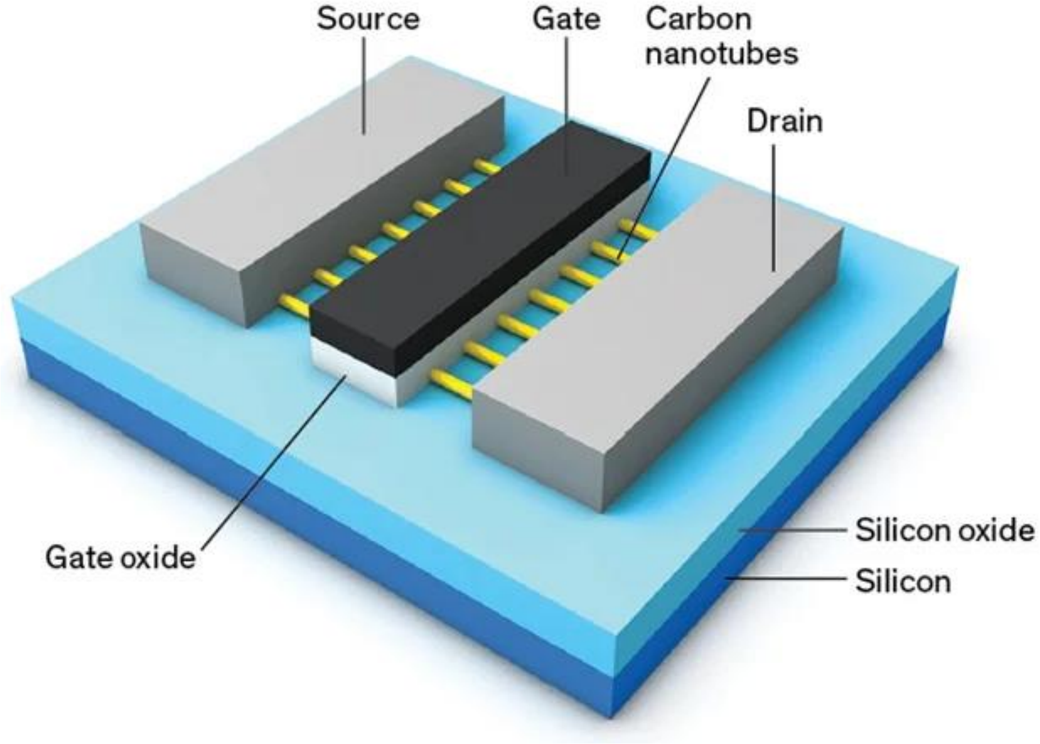




CEA-Leti, un-published

New materials to fuel the roadmap

Carbon nanotubes



https://spectrum.ieee.org/how-well-put-a-carbon-nanotube-computer-in-your-hand (2016)

Advantages:

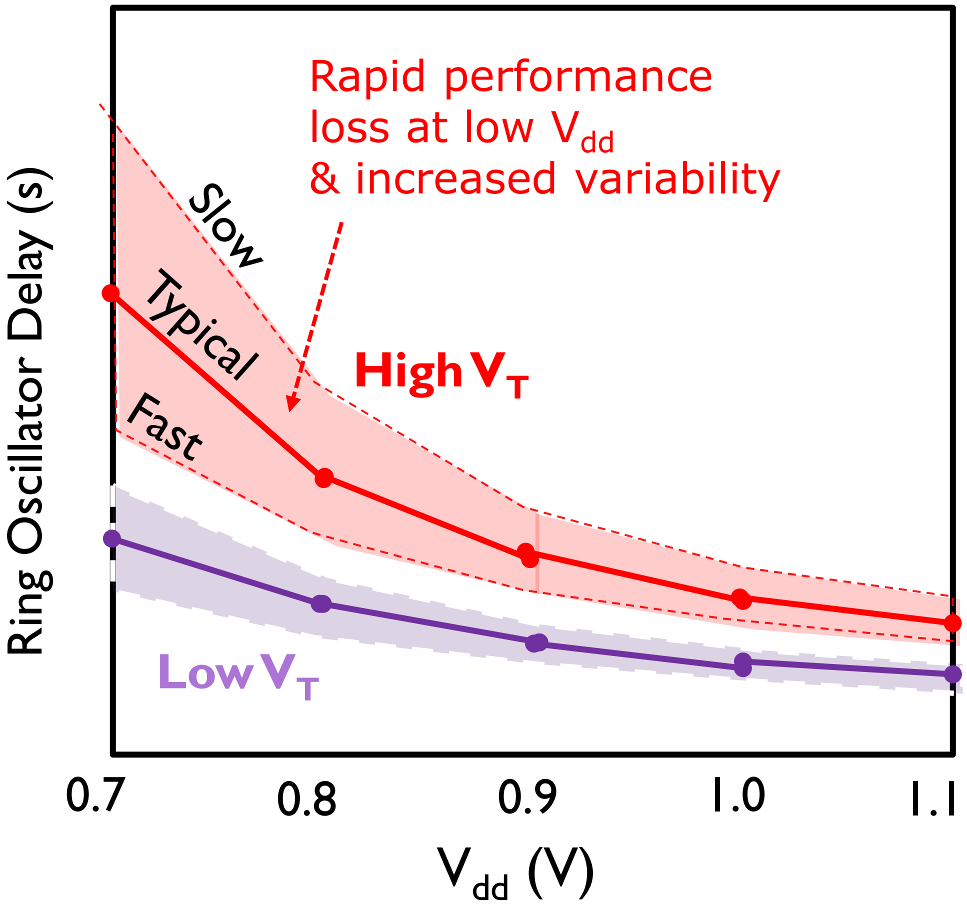
☐ High mobility
☐ Reduced power consumption
☐ Scalability
☐ Thermal stability
☐ Diversity of applications

Challenges:

☐ Purity and uniformity
☐ Large scale cost-effective production
☐ Co-integration with existing technology
☐ Contact resistance
☐ Environmental stability

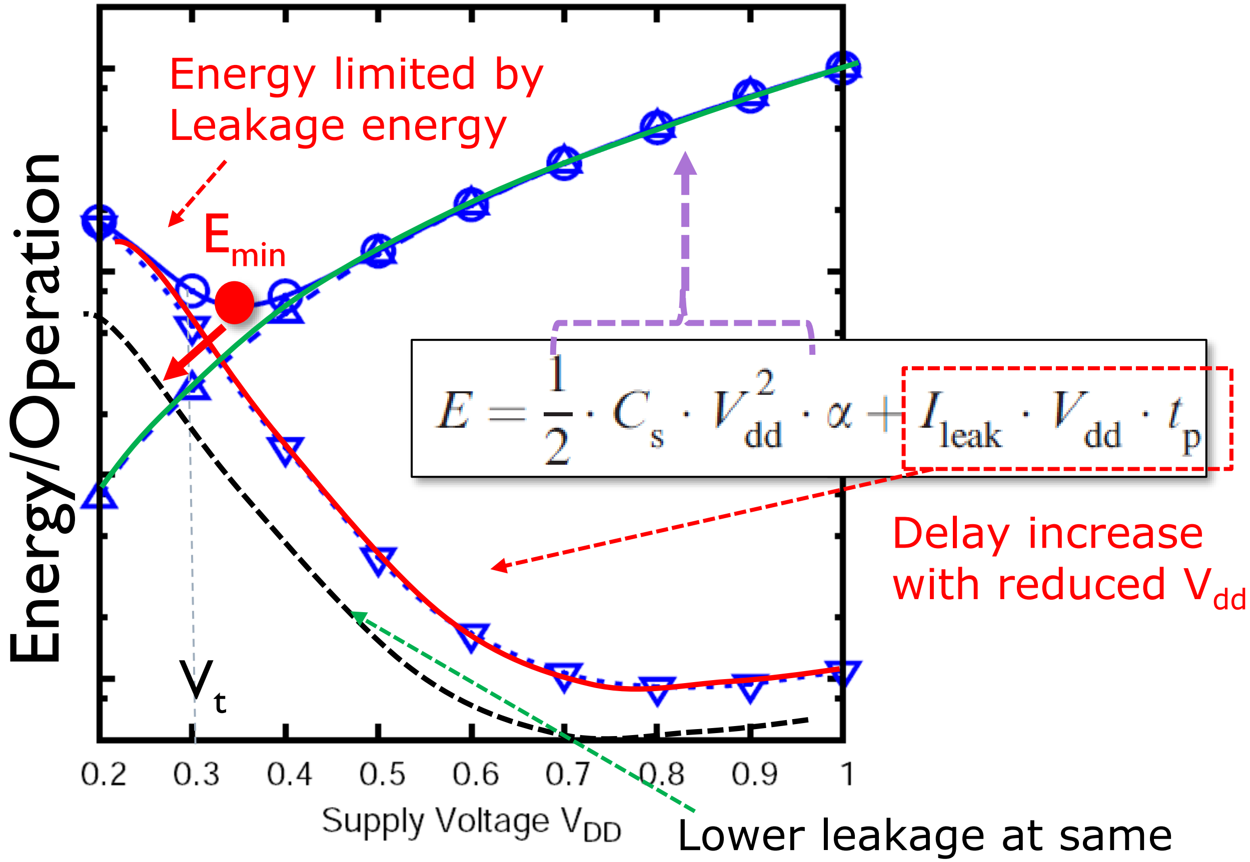
The continued search for ultimate low power switch

Low-V_{dd} Circuit Performance



Rapid performance loss at low V_{dd} & increased variability

Near-Threshold Operation



Energy limited by Leakage energy

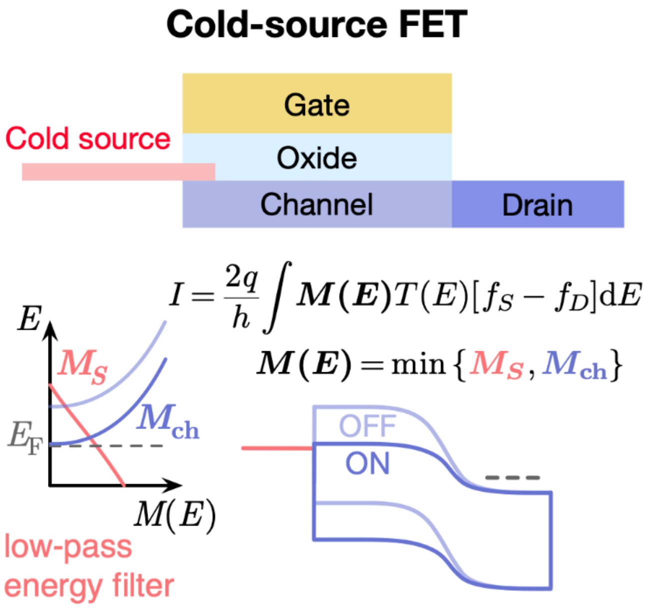
$$E = \frac{1}{2} \cdot C_s \cdot V_{dd}^2 \cdot \alpha + \boxed{I_{leak} \cdot V_{dd} \cdot t_p}$$

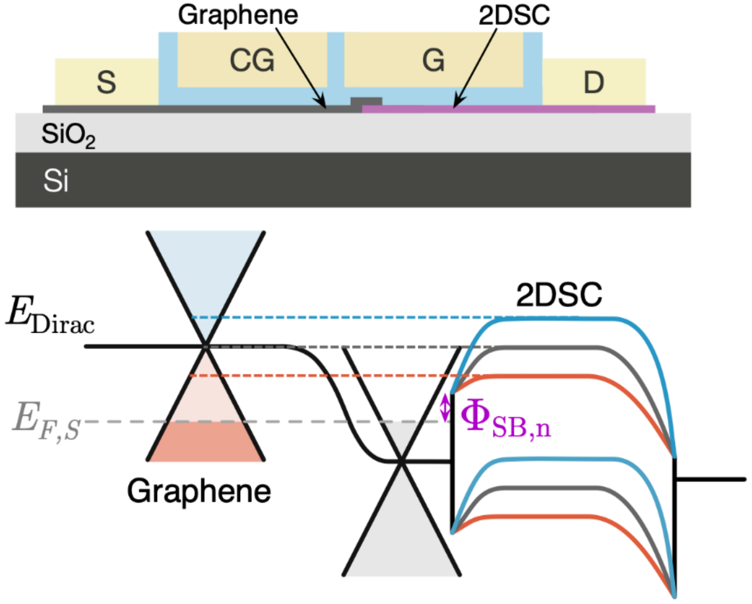
Delay increase with reduced V_{dd}

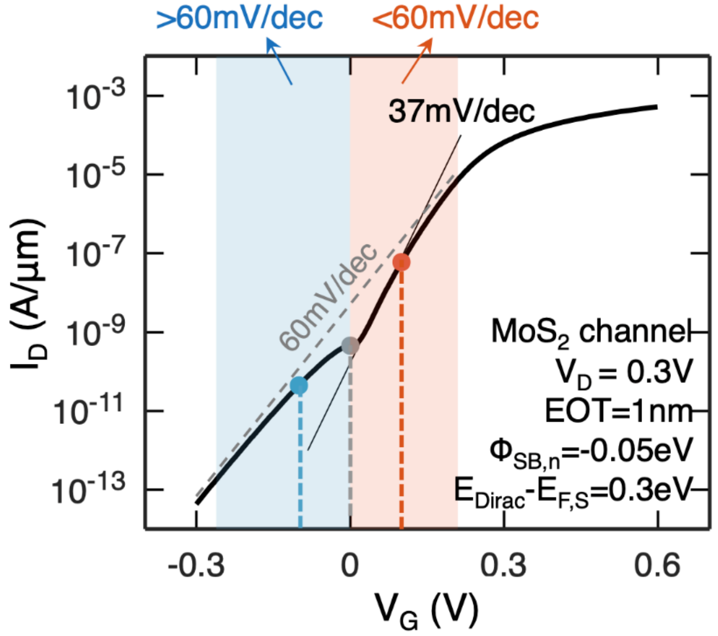
Lower leakage at same perf. → Lower E_{min}

The continued search for ultimate low power switch

2D Dirac-source (cold) FET

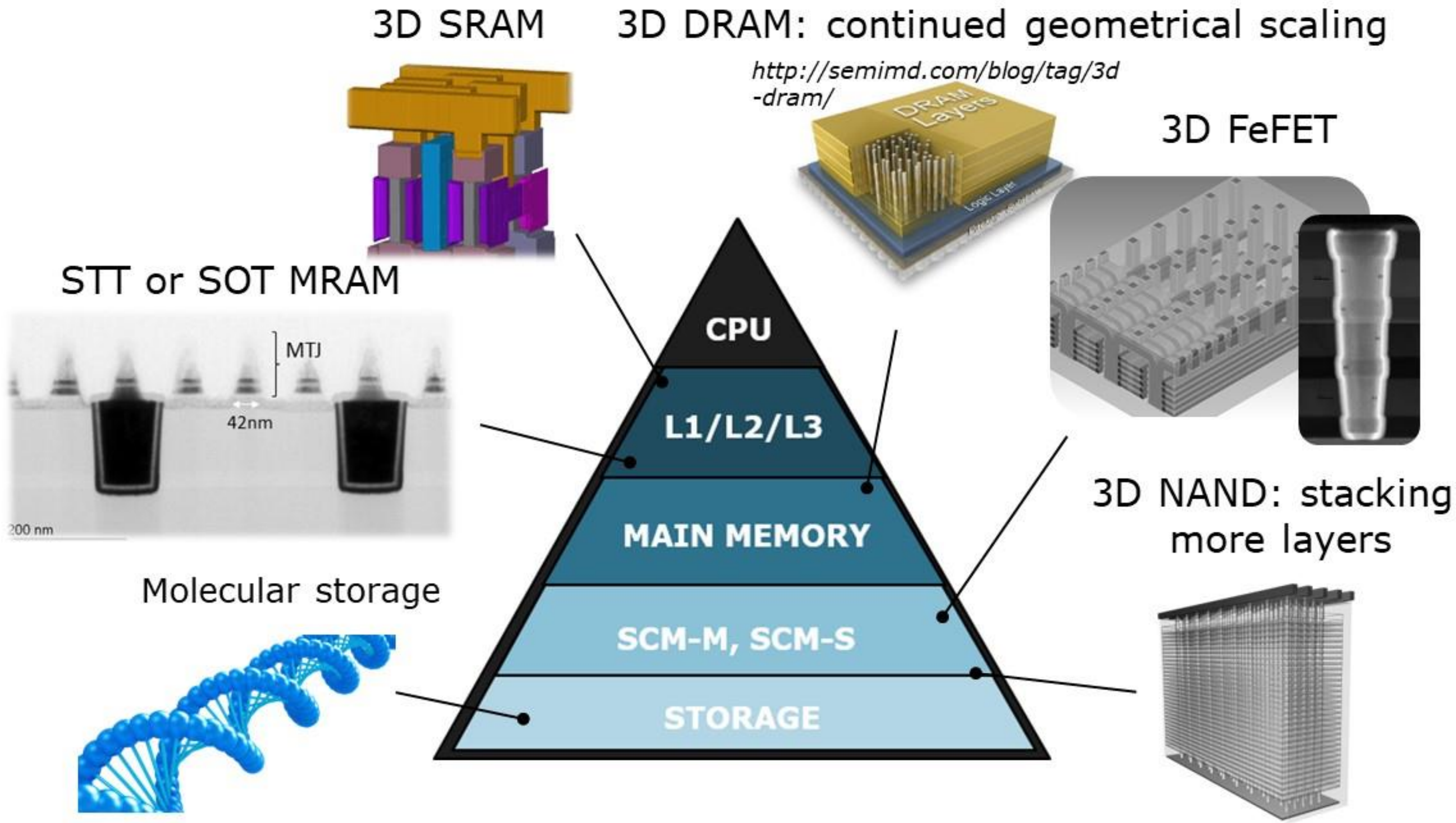






https://www.mit.edu/~pengw/research/csfet/

Increasing demand for storage



3D SRAM

3D DRAM: continued geometrical scaling

http://semimd.com/blog/tag/3d-dram/

3D FeFET

STT or SOT MRAM

3D NAND: stacking more layers

Molecular storage

CPU

L1/L2/L3

MAIN MEMORY

SCM-M, SCM-S

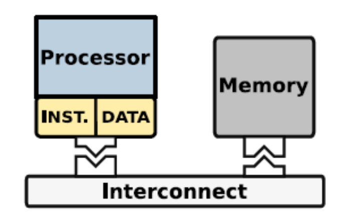
STORAGE

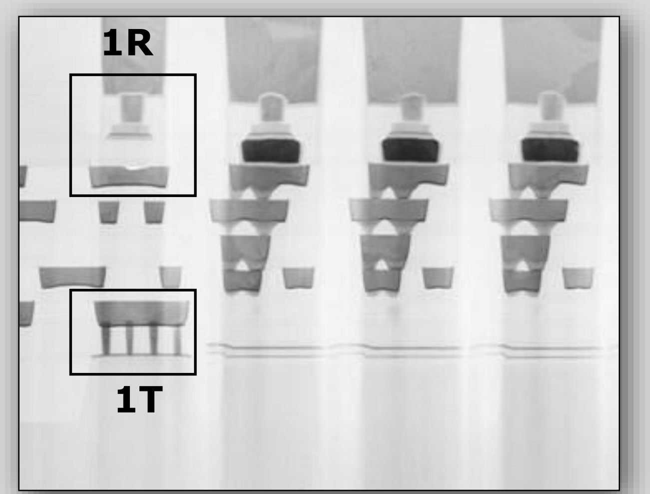
Enabling in & near-memory compute

High dense on-chip memory

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

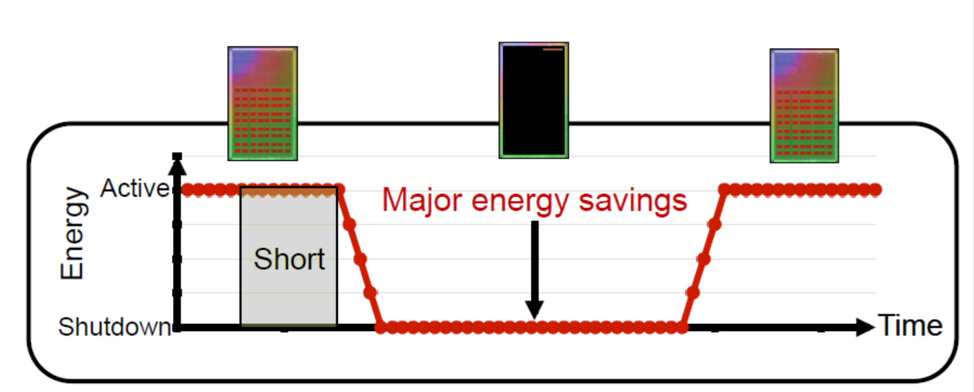
[F. Tu, et al., 2018 ACM/IEEE]





L. Grenouillet et al., 2021

Zero stand-by power thanks to non-volatility



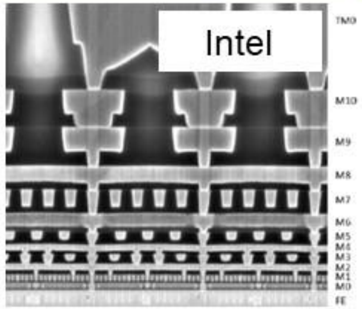
10x better energy efficiency than embedded flash thanks to resistive memories

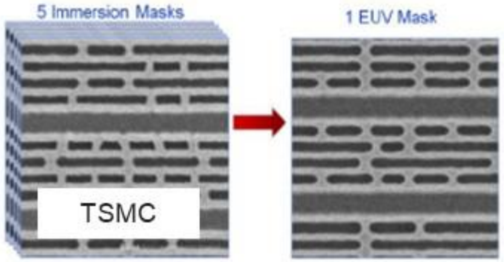
T. Wu et al., 2019

Courtesy: Zsolt Tokei (imec)

~36-40nm BEOL pitch

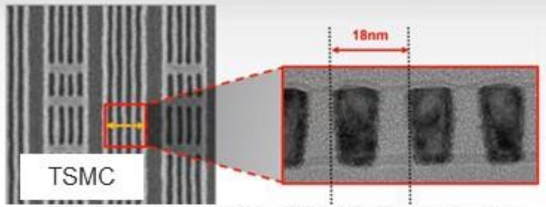
10nm technology, C. Auth et al. IEDM 2017





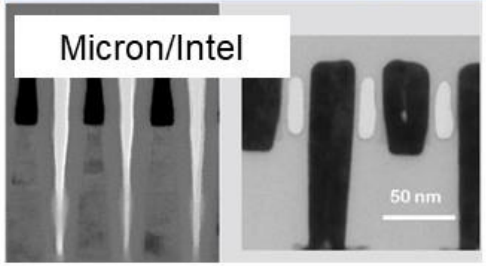
5nm CMOS, G. Yeap et al. IEDM2019

<20nm pitch

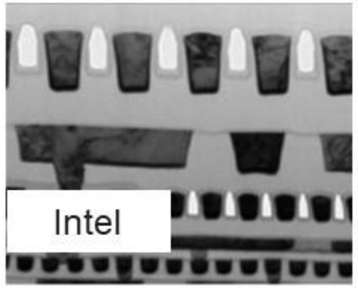


BEOL patterning, TSMC Tech Symp. 2020

AGs in memory & logic

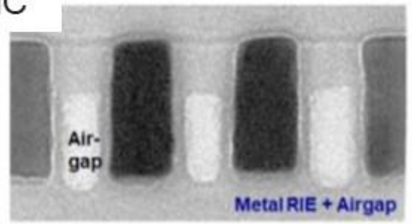


25nm NAND using air gaps K. Prall/K. Parat IEDM2010



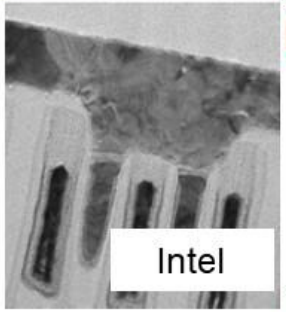
K. Fisher et al., 14nm, IITC2015

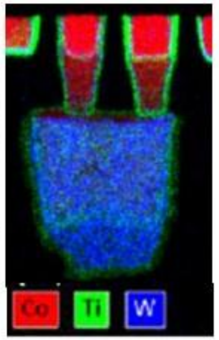
TSMC



Y.J. Mii et al. Keynote, VLS022

Co interconnects



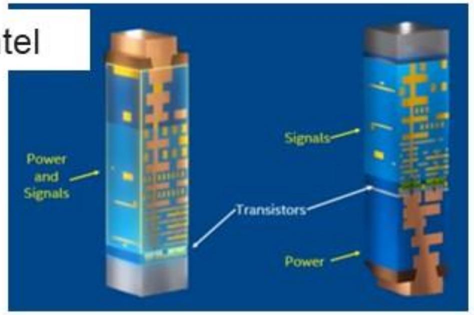


Self- aligned contact, C. Auth et al. IEDM 2017

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

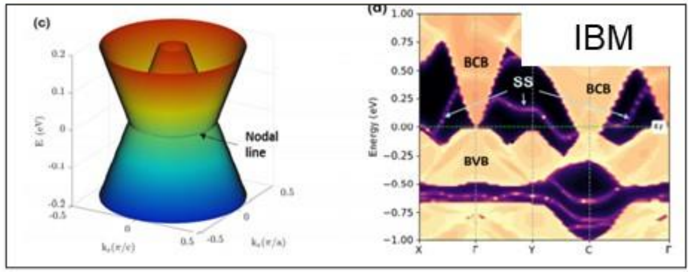
Backside Power Delivery

Intel

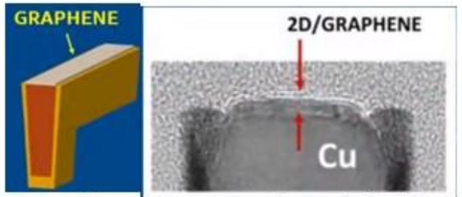


M.C. Mayberry, Keynote IITC 2020

New materials on the horizon

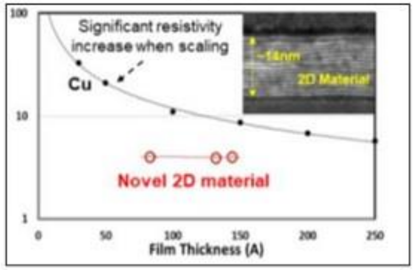


Topological semi-metal, C.T. Chen et al. IEDM2020



Intel

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

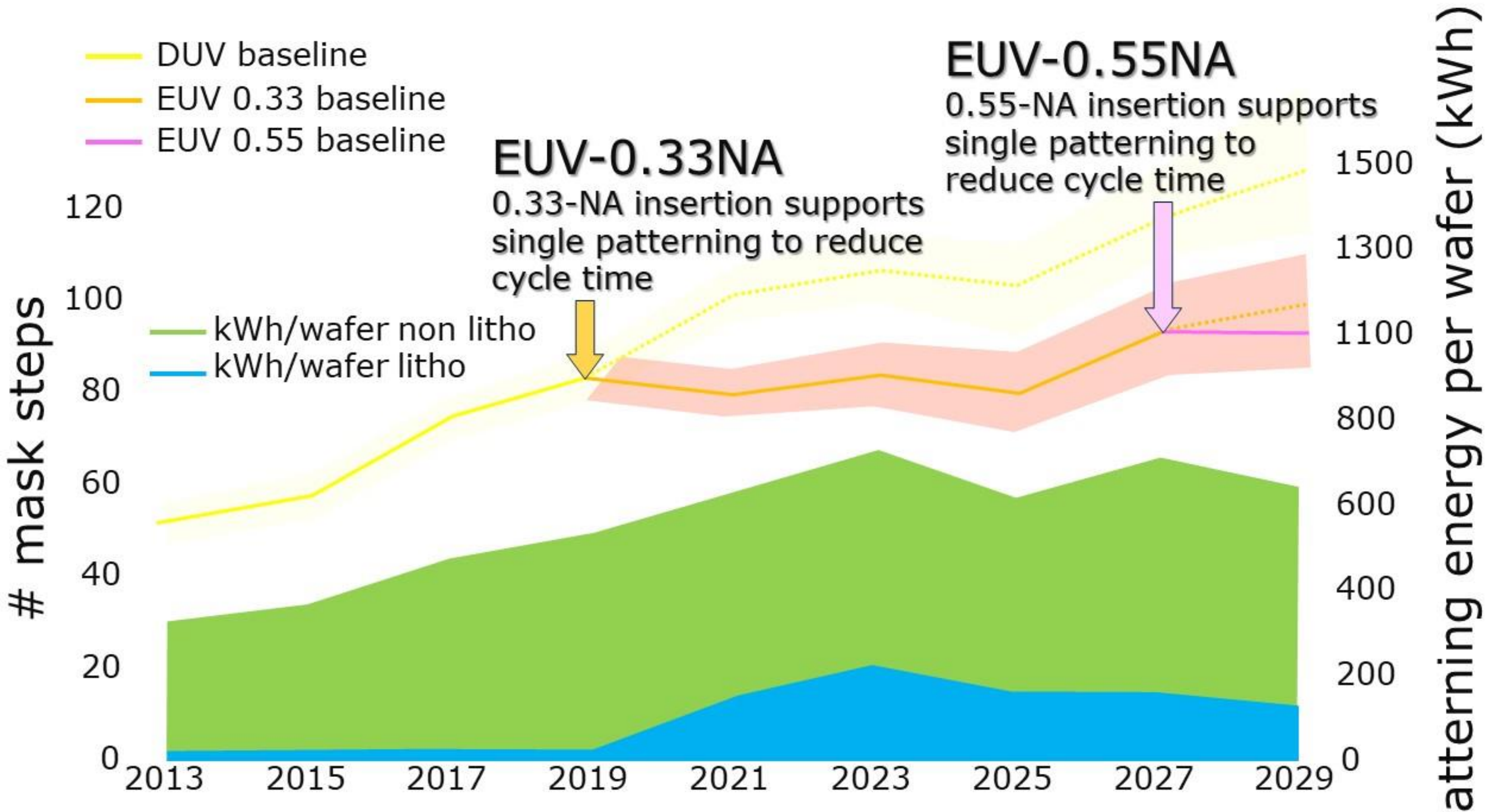


TSMC

New conductor, Y.J Mii Keynote VLSI2022

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

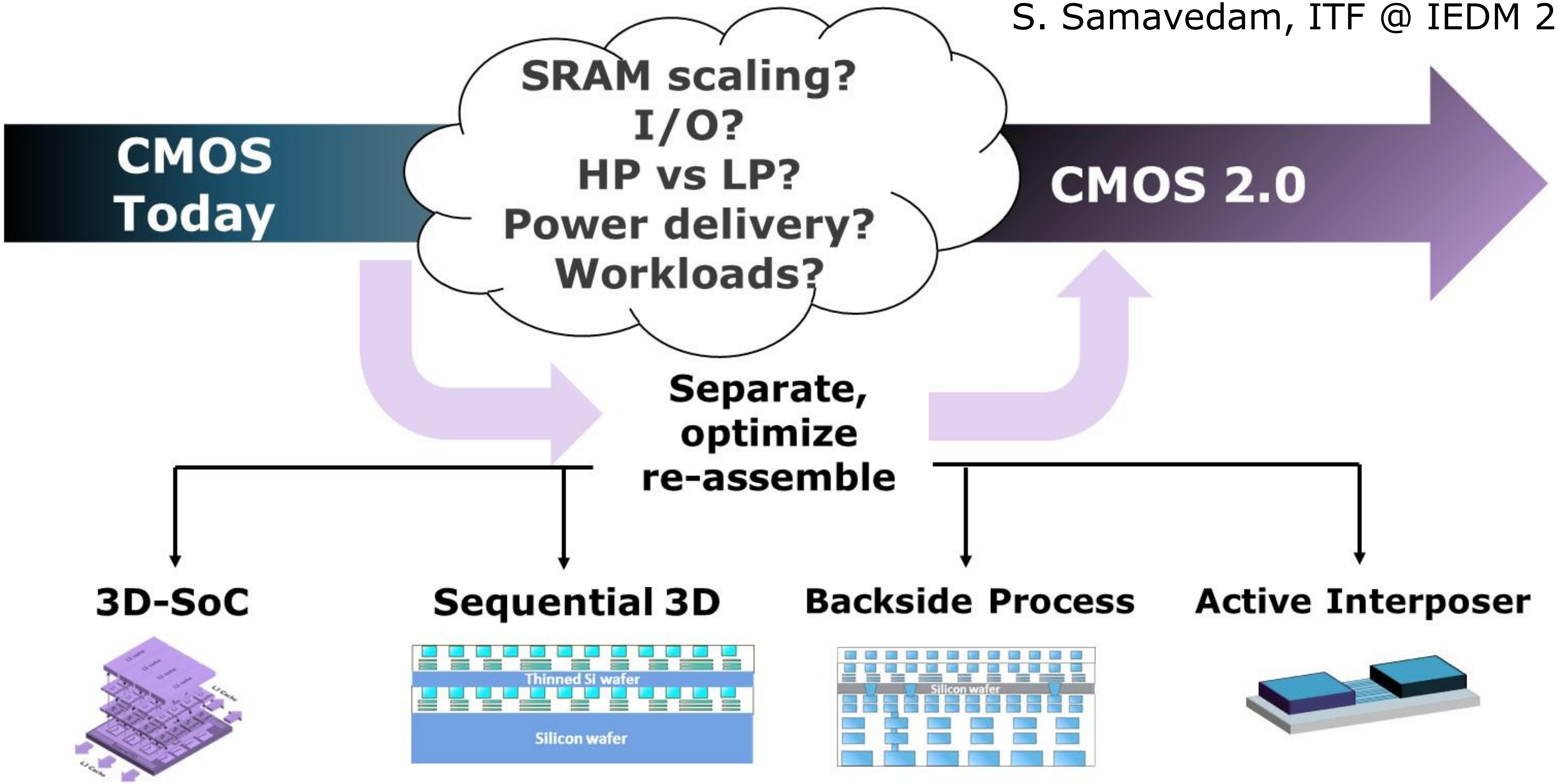
EUV lithography is key enabler



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

Today’s Scaling Challenges Drive the Need for CMOS 2.0

S. Samavedam, ITF @ IEDM 2023



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

CMOS 2.0 Vision brings heterogeneity inside the SoC

Past

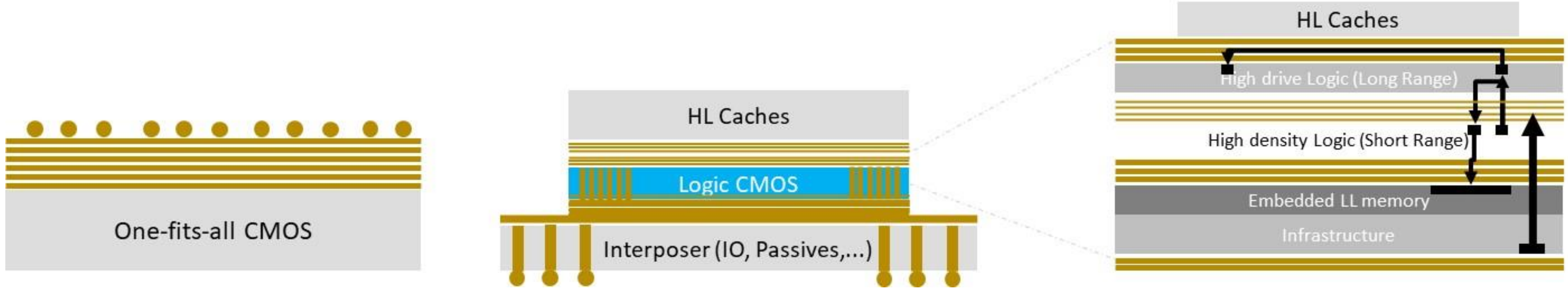
CMOS 1.0
Homogeneous Platform

Present

Heterogeneous Systems

Future

CMOS 2.0
Heterogeneous Logic Platform



One-fits-all CMOS

HL Caches
Logic CMOS
Interposer (IO, Passives,...)

HL Caches
High drive Logic (Long Range)
High density Logic (Short Range)
Embedded LL memory
Infrastructure

Conclusions

- New applications will drive different workloads and technology solutions
- New materials and devices next to novel connectivity solutions and compute architectures will play a key role in compute system scaling
- Sustainability becoming an increasingly important metric for evaluating technology choices