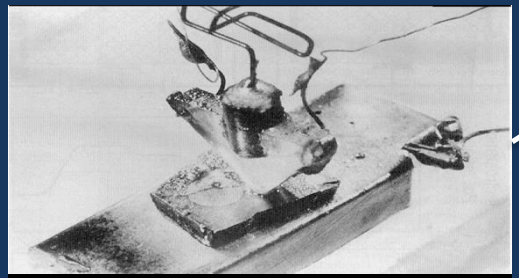
The Future of Microelectronics

Research Trends and Career Opportunities

Amitabh Yadav

Ph.D. Candidate, Integrated Neurotechnologies Laboratory (INL)
École Polytechnique Fédérale de Lausanne (EPFL).
Genève, Switzerland.



First Transistor, Bell Labs, 1947.



The Cerebras WSE-3 Turbo (WSE-3T) is the largest AI chip ever built, measuring 46,225 mm² and containing 4 trillion transistors. It delivers 250 petaflops of AI compute through 900,000 AI-optimized, (2026).

Amitabh Yadav

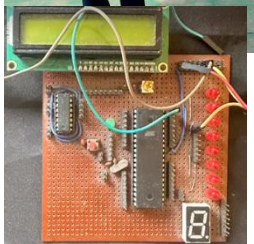
B.Tech. (2013-17)
Electronics Engg.
UPES Dehradun.

M.Sc. (2017-19)
Computer Engg.
TU Delft.

PreDoc (2019-21)
Physics, UC Berkeley.

Ph.D. (2023-28)
Microelectronics, EPFL.



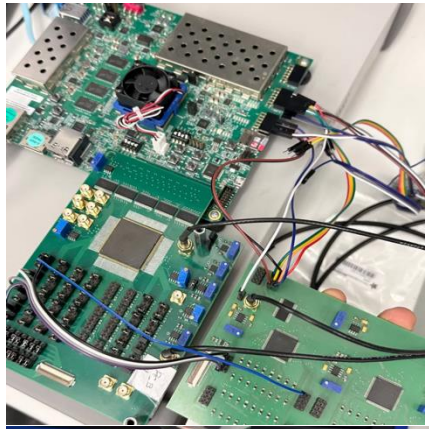


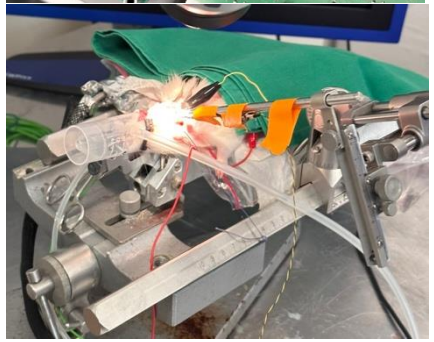












How do you navigate the complex academic landscape?

Caution: This is a dense presentation[1]

[1]Richard Hamming, “You and Your Research”, Bell Communications Research Colloquium Seminar, 1986.

intention is to mark the 'important' stuff, but I end up creating this mess instead.

A programmable two-qubit quantum processor in silicon

Case study: 1D Ising with IBMQ

Quantum 322w

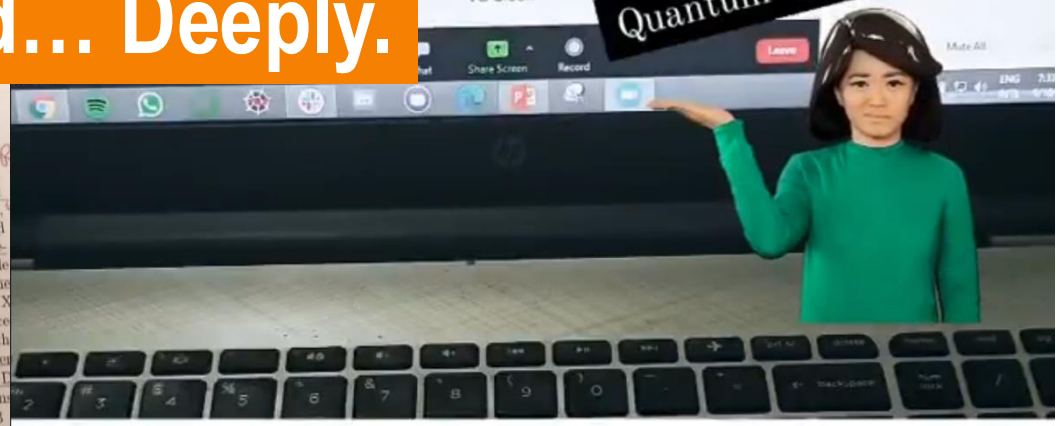
- One of our goals: perform the full chain on a real QC
  - Demonstrate the ability to perform e.g. dynamics simulations beyond the NISQ depth limit
- Planning demonstration with 2-spin 1D Ising
- Current challenge:
  Learning the uncomputing circuit with sufficient accuracy
  - Gradient-based learning: need < 1% determination of $|\langle 0|U(\theta)|\psi\rangle|^2$, $|\langle 0|U(\theta+\frac{\pi}{2})|\psi\rangle|^2$, $|\langle 0|U(\theta-\frac{\pi}{2})|\psi\rangle|^2$
  - Plan to try gradient-free minimization (e.g. COBYLA)

Quantum Chemistry

Machine Learning

Quantum Information

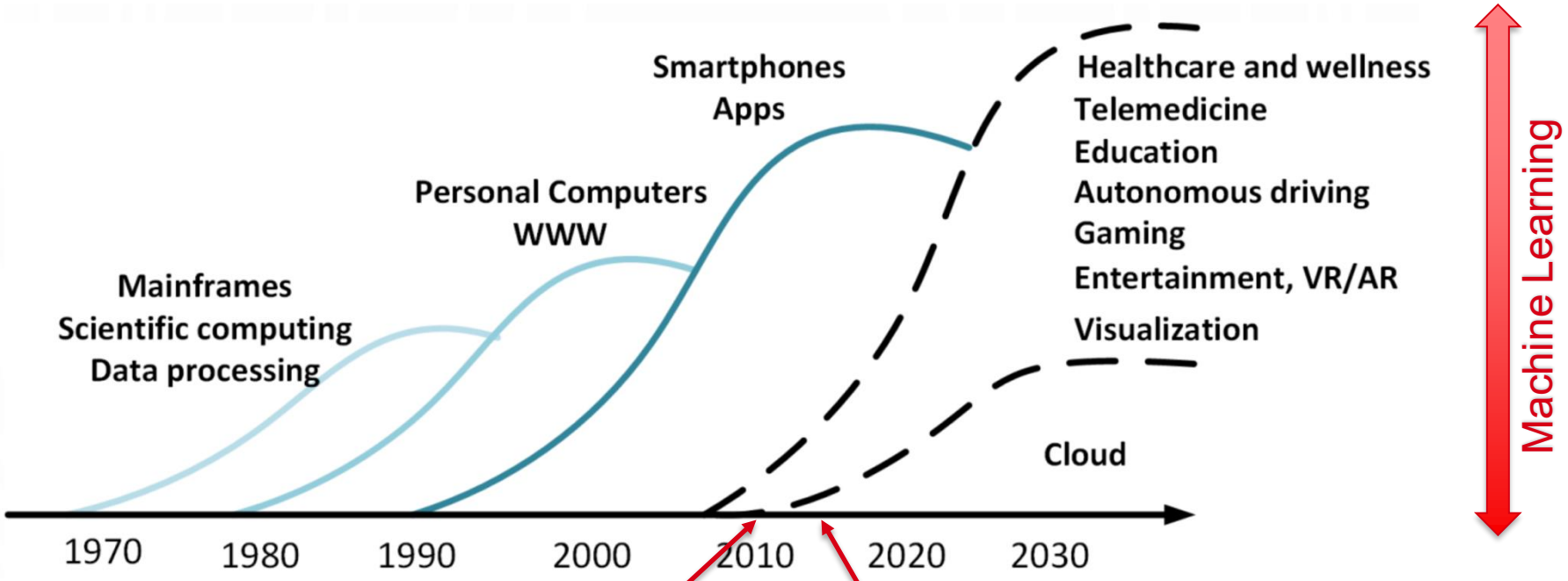
Read... Deeply.



Sometimes it feels, besides the department heads, nobody understands the complete presentation.

Integrated Circuit Design

Diverse Applications of Microelectronics



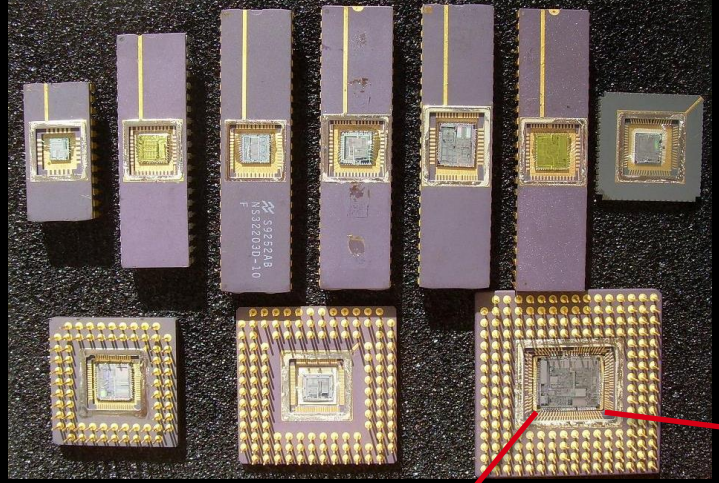
Source: Sophia Shao, EECS151, UC Berkeley.

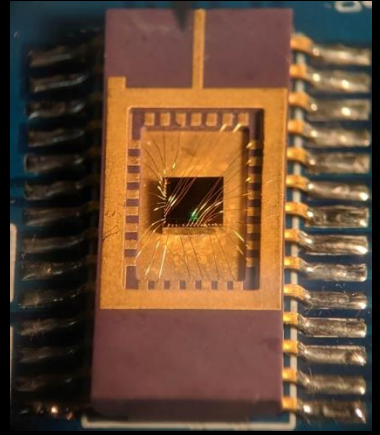
AlexNet
https://dl.acm.org/doi/10.5555/2999134.2999257

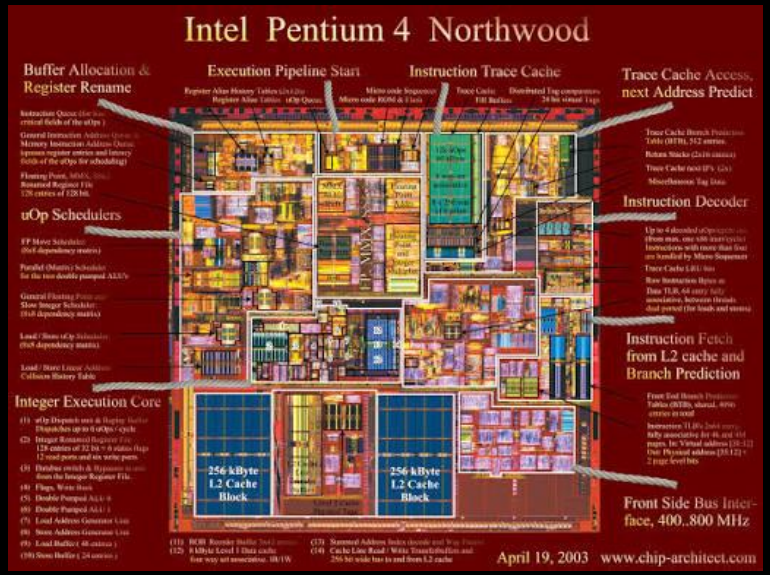
Attention is all you need!
https://arxiv.org/abs/1706.03762

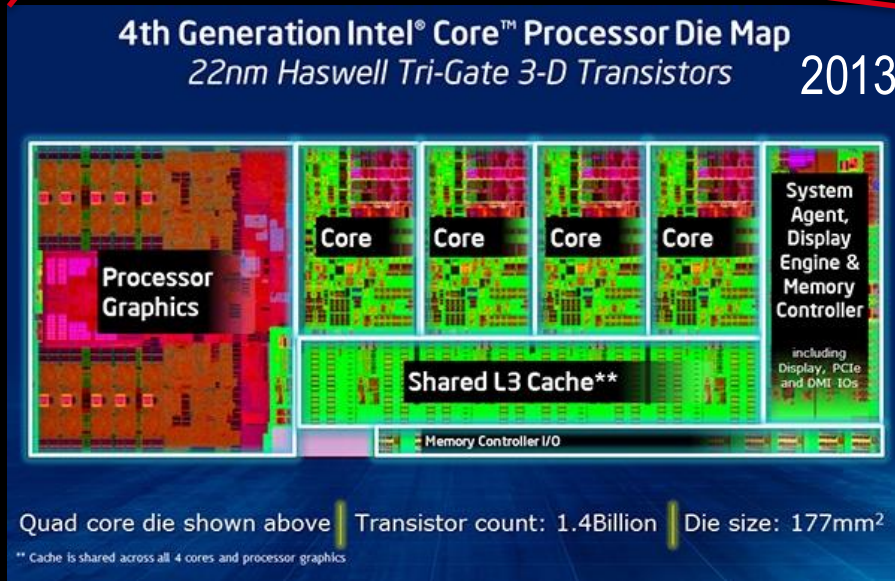
CPU Era


intel
PENTIUM PRO
intel
pentium
AMD-K5

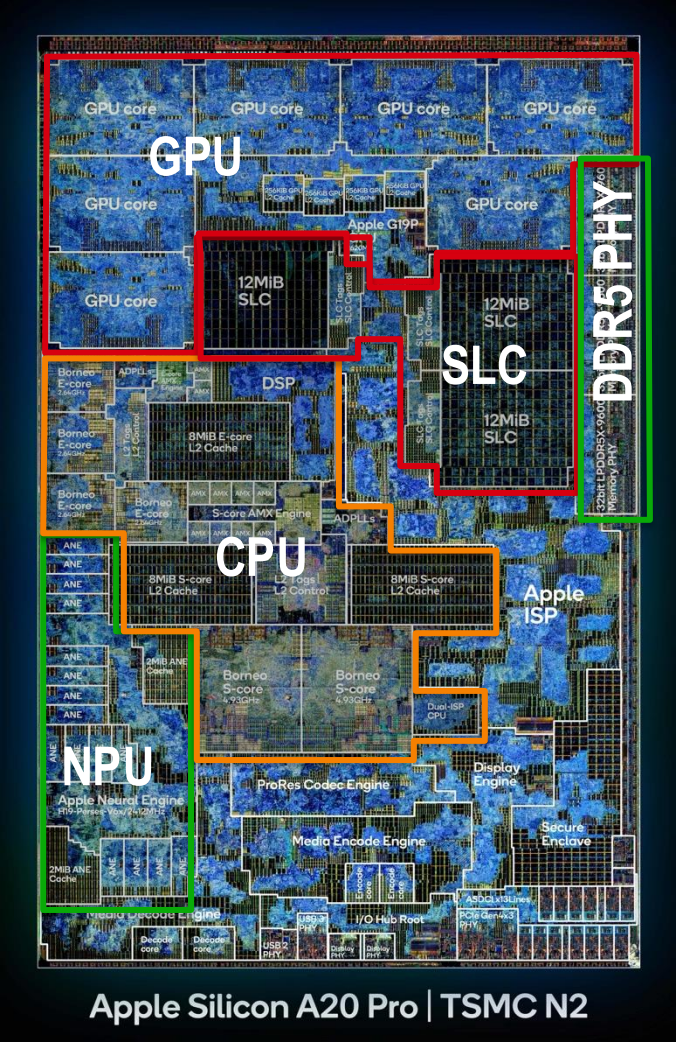





Intel Pentium 4 Northwood
Buffer Allocation & Register Rename
Execution Pipeline Start
Instruction Trace Cache
Trace Cache Access, next Address Predict
uOp Schedulers
Instruction Decoder
Instruction Fetch from L2 cache and Branch Prediction
Integer Execution Core
256 kByte L2 Cache Block
256 kByte L2 Cache Block
Front Side Bus Interface, 400..800 MHz
April 19, 2003
www.chip-architect.com


4th Generation Intel® Core™ Processor Die Map
22nm Haswell Tri-Gate 3-D Transistors
2013
Processor Graphics
Core
Core
Core
Core
System Agent, Display Engine & Memory Controller
including Display, PCIe and DMI IOs
Shared L3 Cache**
Memory Controller I/O
Quad core die shown above
Transistor count: 1.4Billion
Die size: 177mm²
** Cache is shared across all 4 cores and processor graphics

A20 Pro System-on-Chip (SoC) (2026)

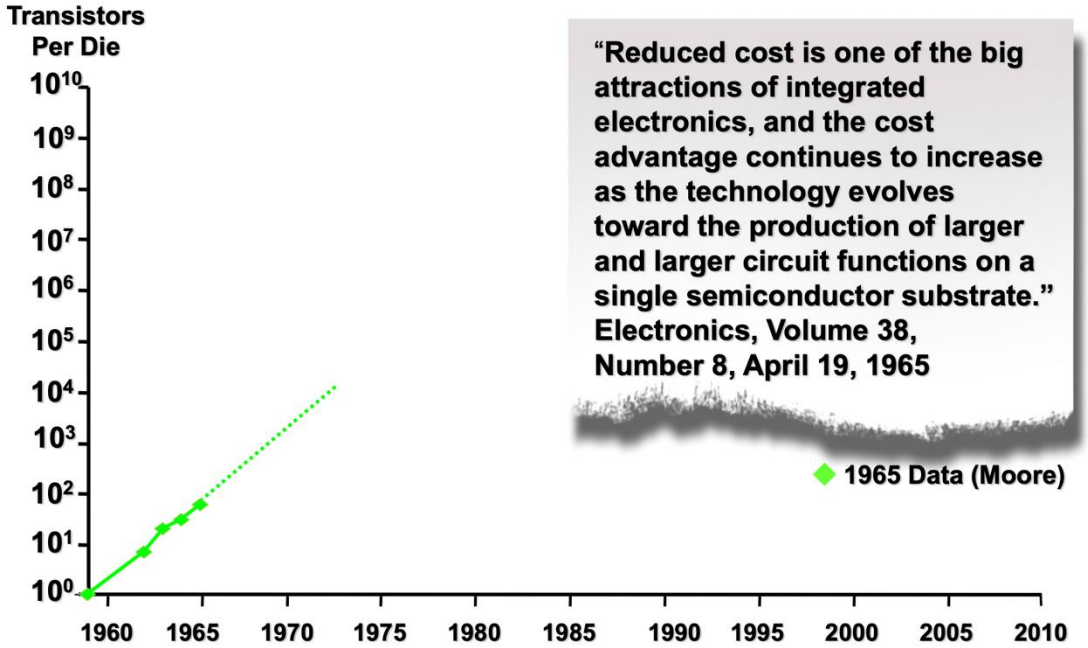


BLOCK	AREA MM²	SHARE OF DIE
CPU	19.05	19.3%
P-core cluster = CPU – E-core cluster	12.58	12.7%
E-core cluster	6.47	6.5%
GPU	21.26	21.5%
NPU	8.54	8.6%
ISP not measured yet	–	
Modem not measured yet	–	
DDR PHY	3.99	4.0%
SLC	9.89	10.0%
Other (IO, interconnect, unlabeled)	36.07	36.5%
Whole die 8 × 12.35 mm	**98.80**	100%

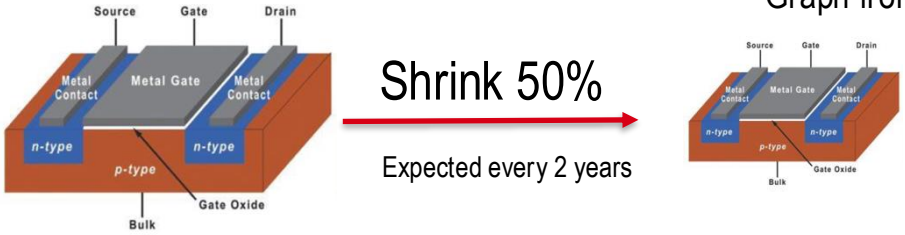
Source: X @HiEq_2005

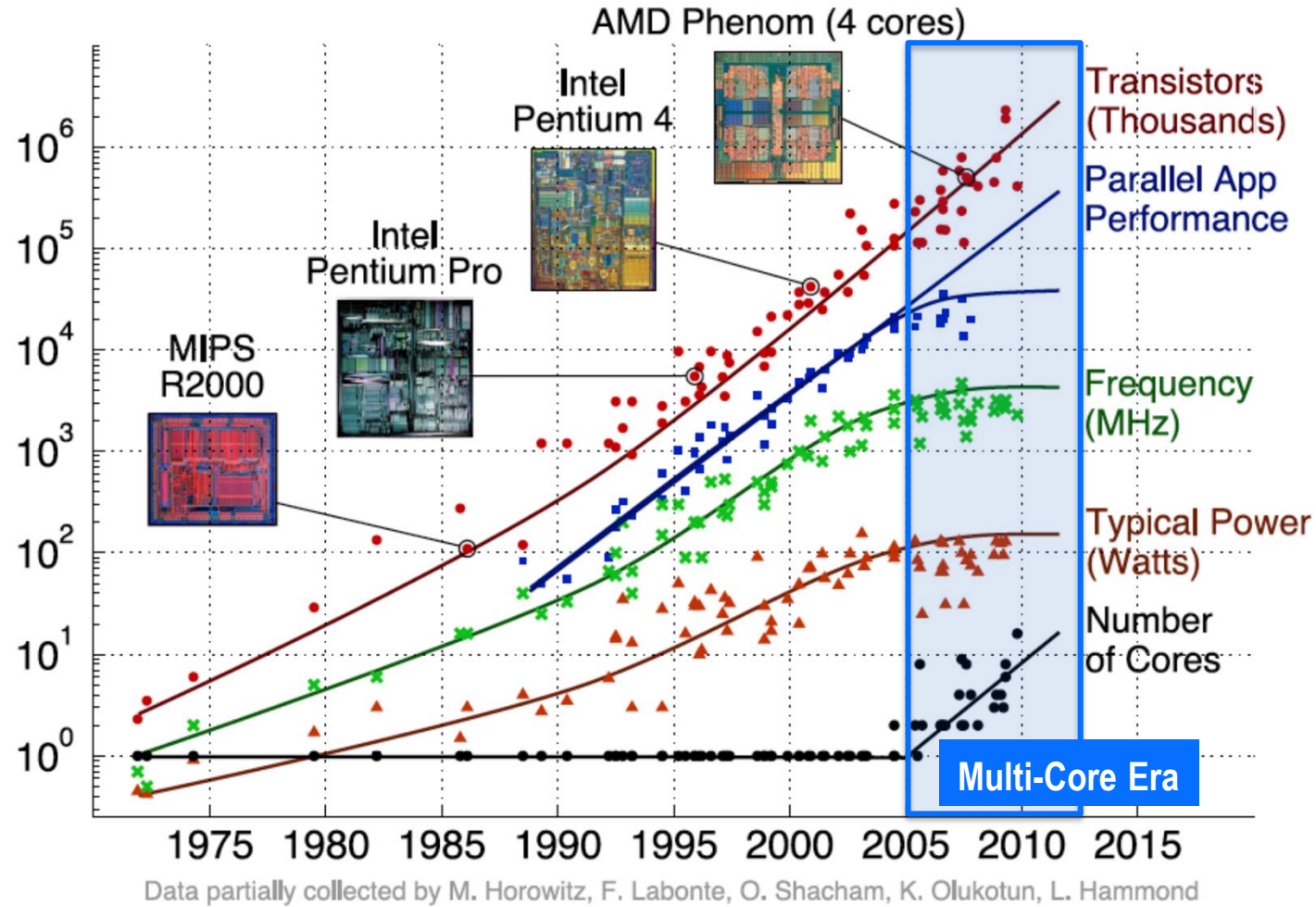
Moore's Law, Dennard's Law… Scaling Laws

- In 1965, Gordon Moore noted that the number of transistors on a chip doubled every 12 months.
- Semiconductor technology will double its effectiveness every ~~12~~ 18 months!



Graph from S.Chou, ISSCC'2005





R. H. Dennard *et al.*,
IEEE J. Solid State Circuits, (1974).

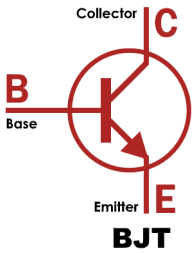
Dennard's Law (1974)

As transistors get smaller, both voltage and current scale (downward) with length, so overall power density stays constant.

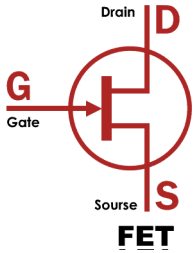
$$P = CV^2f$$

P=power
C=capacitance
V=voltage
f= frequency

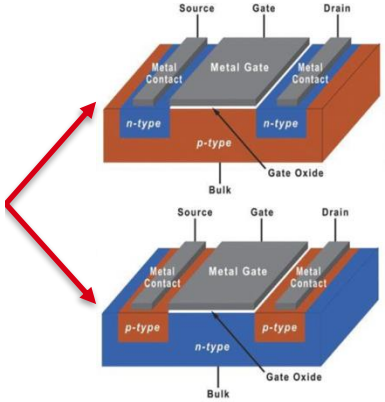
As transistors shrank, engineers lowered **Voltage**. Because voltage is squared in the formula, dropping it created huge power savings.



BJT

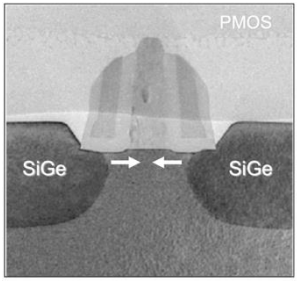


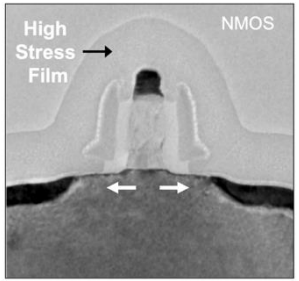
FET



nMOS

pMOS





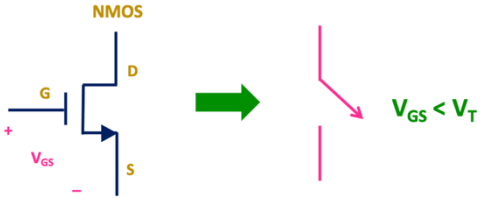
CMOS Inverter

$V_o = \sim V_{in}$ (Boolean NOT operation)

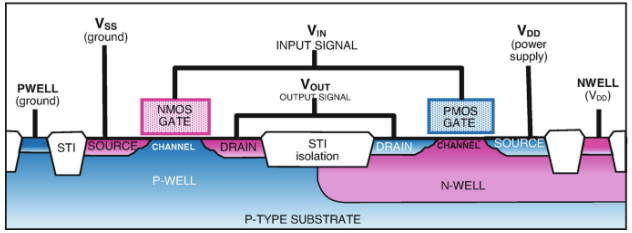
Planar CMOS!

Combine them together, it's CMOS!

Transistor as a switch

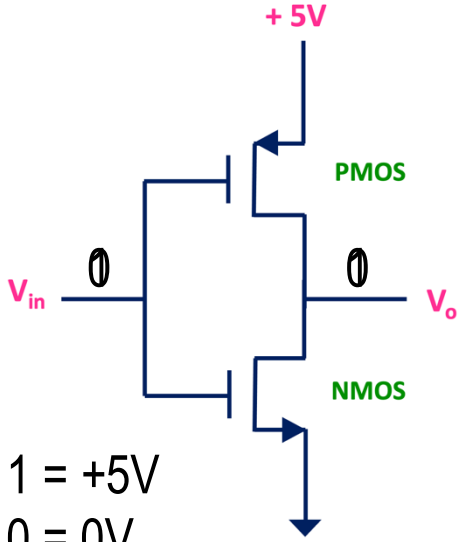






CMOS was combinable, because together pMOS+nMOS did boolean algebra!

Boolean Algebra = Mathematics on Computer.



1 = +5V

0 = 0V

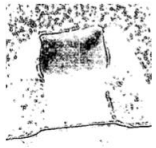
Transistors are Changing

Dennard Scaling ended here.

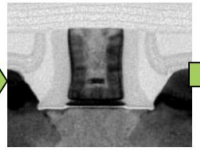
Enabling Moore's law when there is no Dennard Scaling to help.

65/55 nm | 45/40 nm | 32/28nm | 22/20nm | 16/14nm | 10nm | 7nm | 5nm

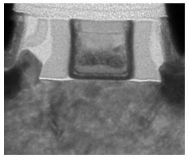
Bulk



SiO_2/SiN Strain



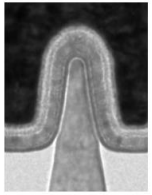
Intel, IEDM'07 HK/MG Strain



Intel, IEDM'09

TSMC, Samsung

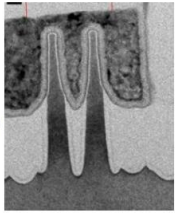
FinFET



Intel, IEDM'12



Intel, VLSI'14

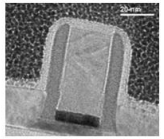


Intel, IEDM'17



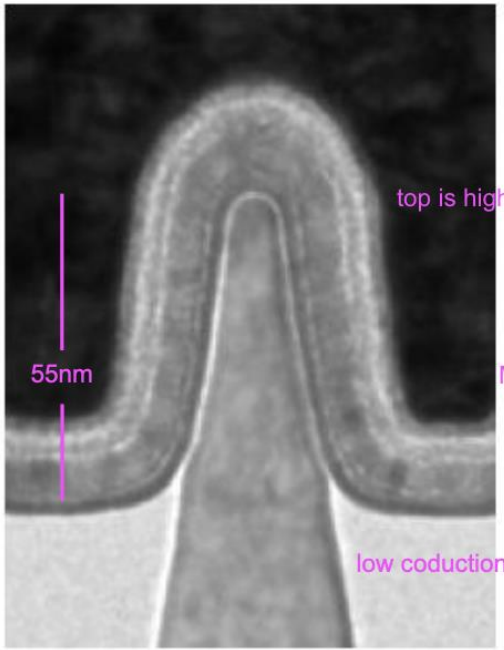
TSMC 7nm (hexus.net)

FDSOI



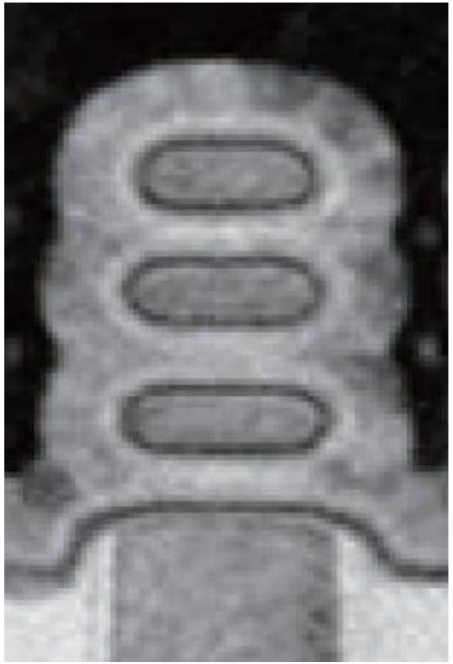
ST, VLSI'12

2012



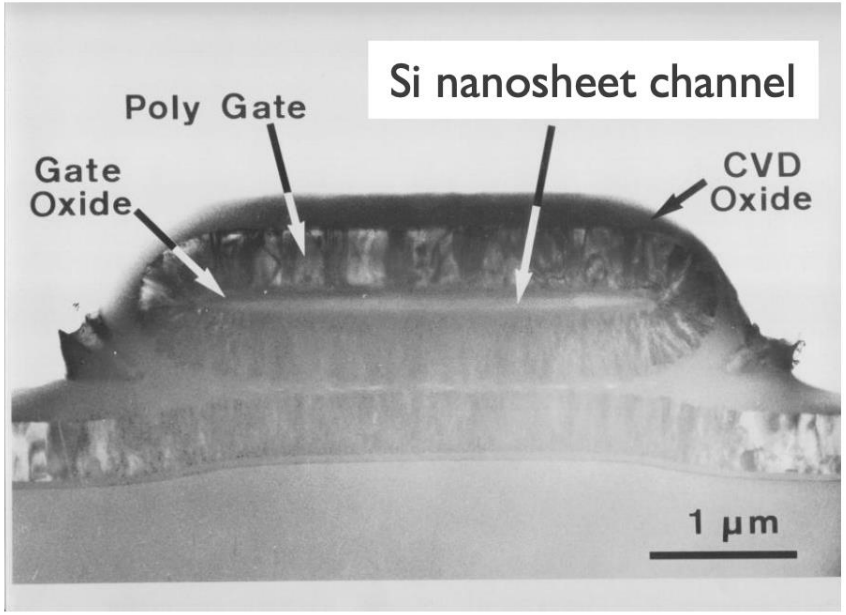
FinFET

2022

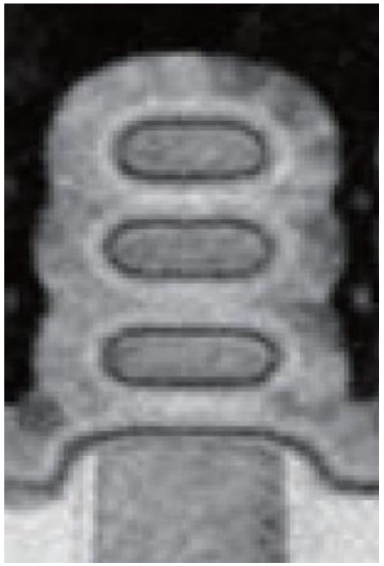


Nanosheet or GAA

Nanosheet device in '90s



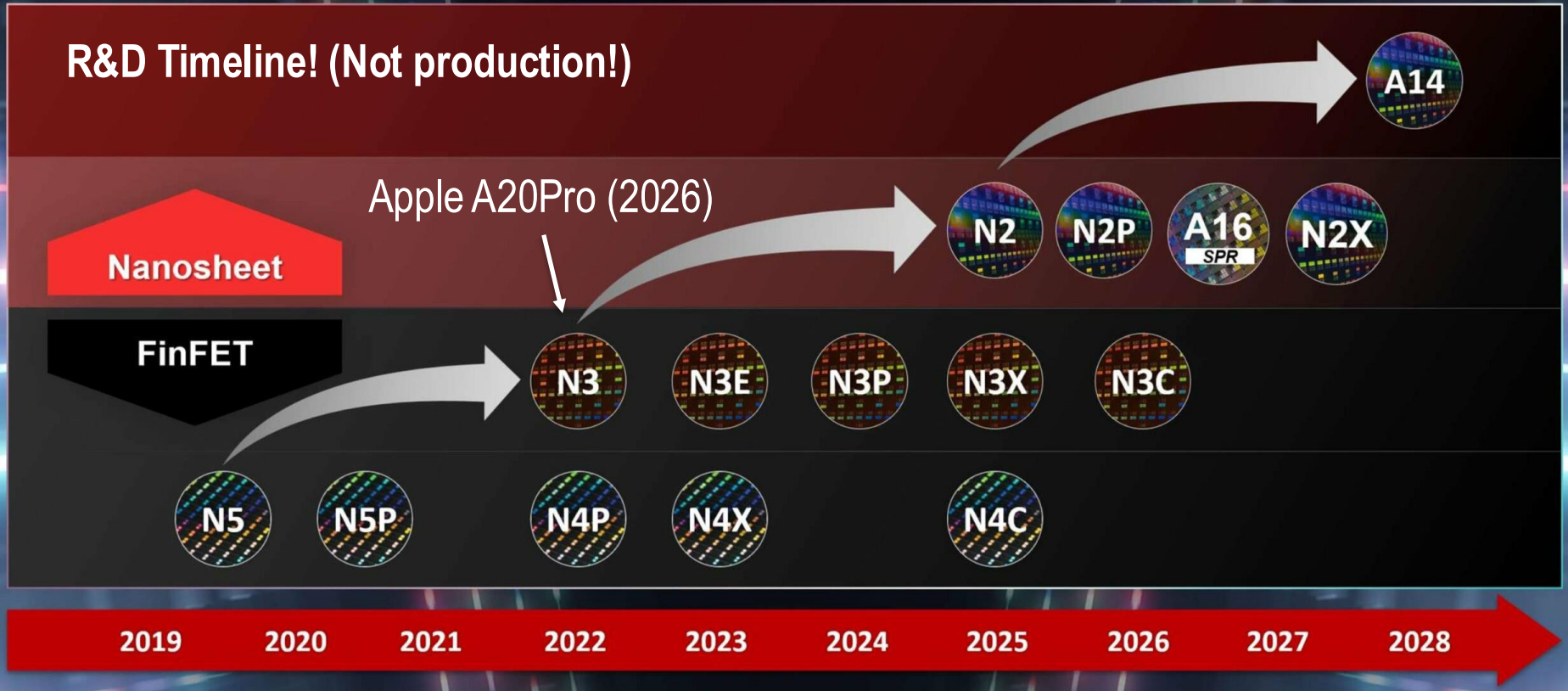
J. P. Colinge (imec) et al. IEDM 1990



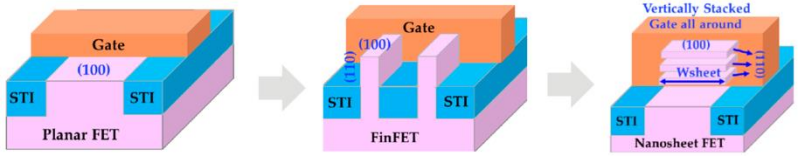
Today

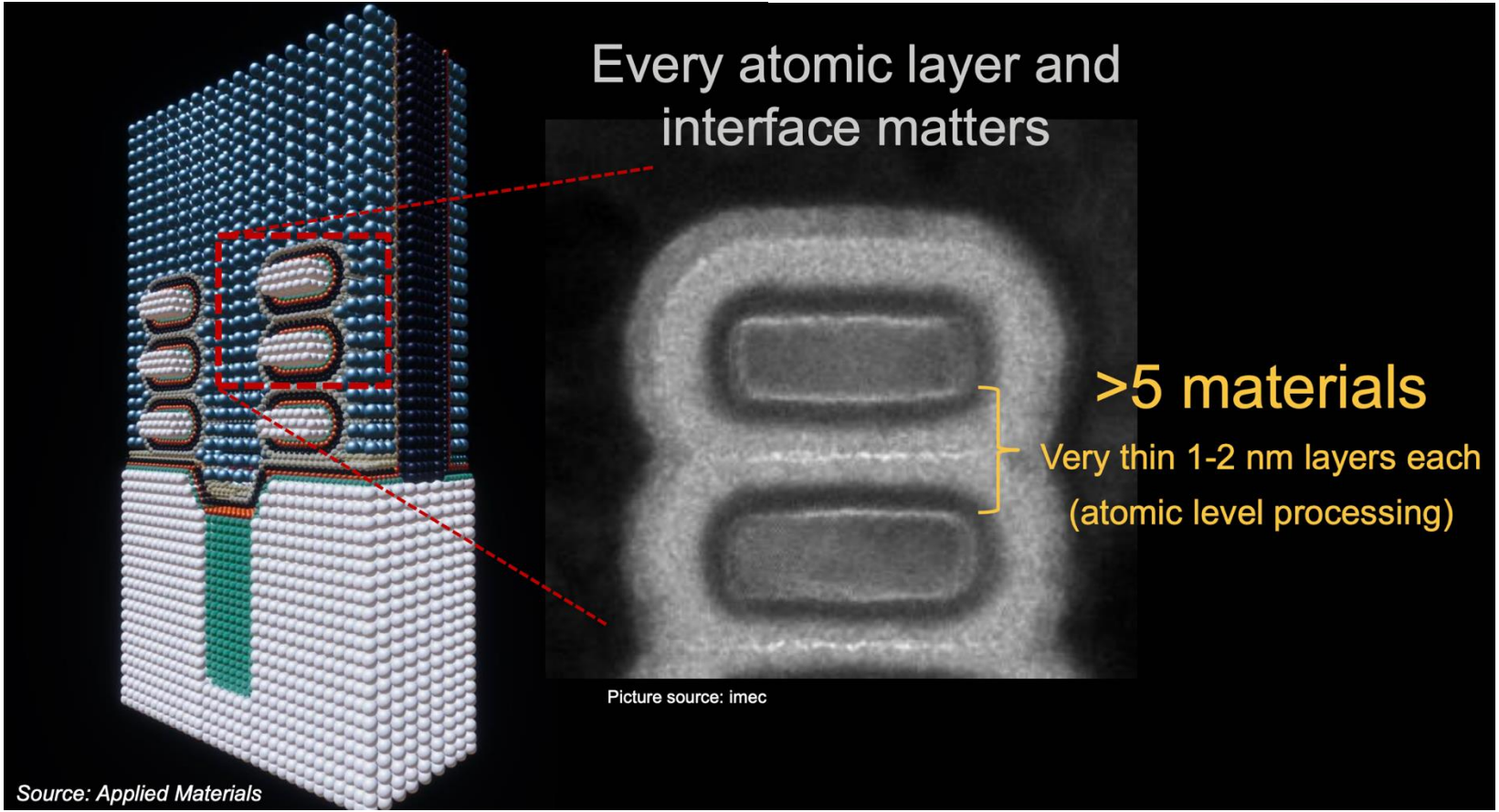
It took more than 30 years from concept to production
→ Building a roadmap is key for this industry!

TSMC Industry Leading Advanced Logic Technology



CMOS Scaling Trend





Scaling driven by a technology roadmap

2018	2020	2023	2025	2027	2029	2031	2033	2035	2037	2039
N7	N5	N3	N2	A14	A10	A7	A5	A3	A2	sub-A2

DEVICE AND MATERIAL INNOVATIONS

FinFET → **NanoSheet** → **CFET** → **2DFET, ?, …**

	Introduction	Scaled	Extended	Introduction	Scaled	Extended	Introduction	Scaled	Extended	Introduction	Scaled
											
Metal Tracks	6	6	6	5	5	5	4	<4	<4	<4	<4

CONTINUED DIMENSIONAL SCALING

0.33NA EUV → 0.55NA EUV → 0.75NA EUV ?

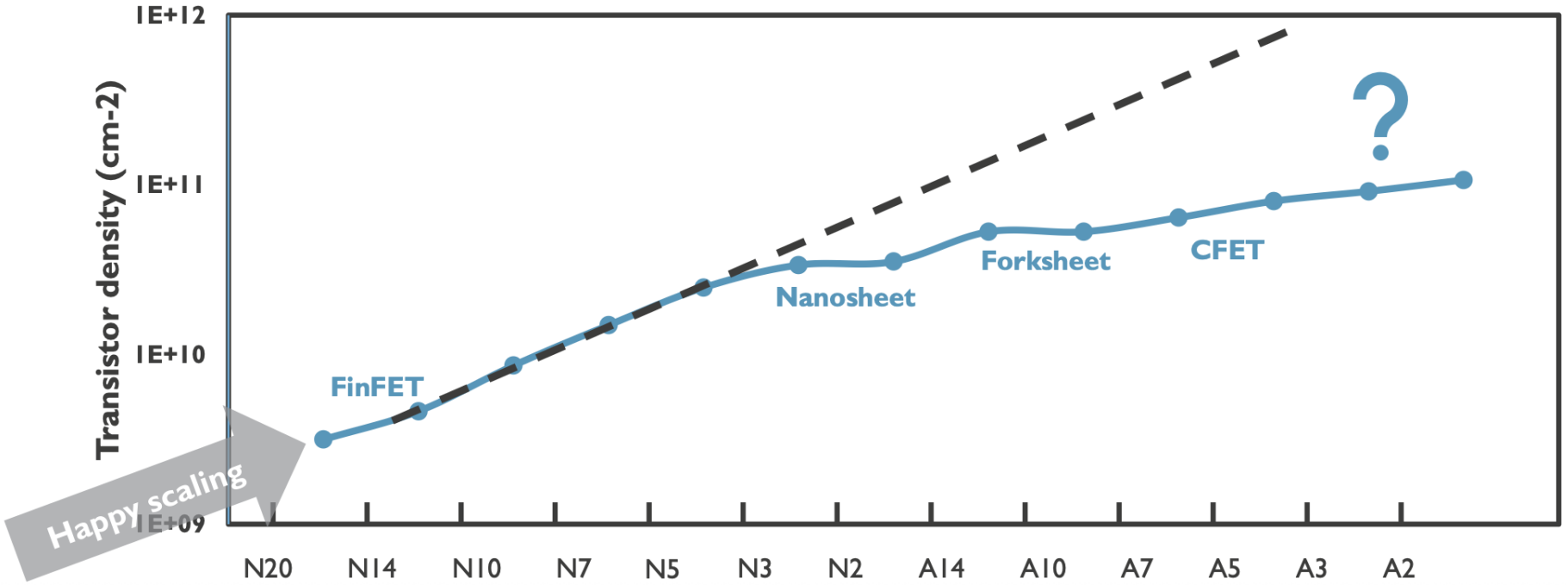
Metal Pitch [nm]	40	28	23	22	21	18	18-16	18-14	16-12	16-12	14-10

CHIP INTERCONNECT ARCHITECTURE

Back-side Power - Global interconnects → Back-side Power - Local signal lines

										…
Back-side Metal Pitch estimate [nm]	-	-	160-65	115-45	90-42	40-34	40-28	tbd	tbd	

imec



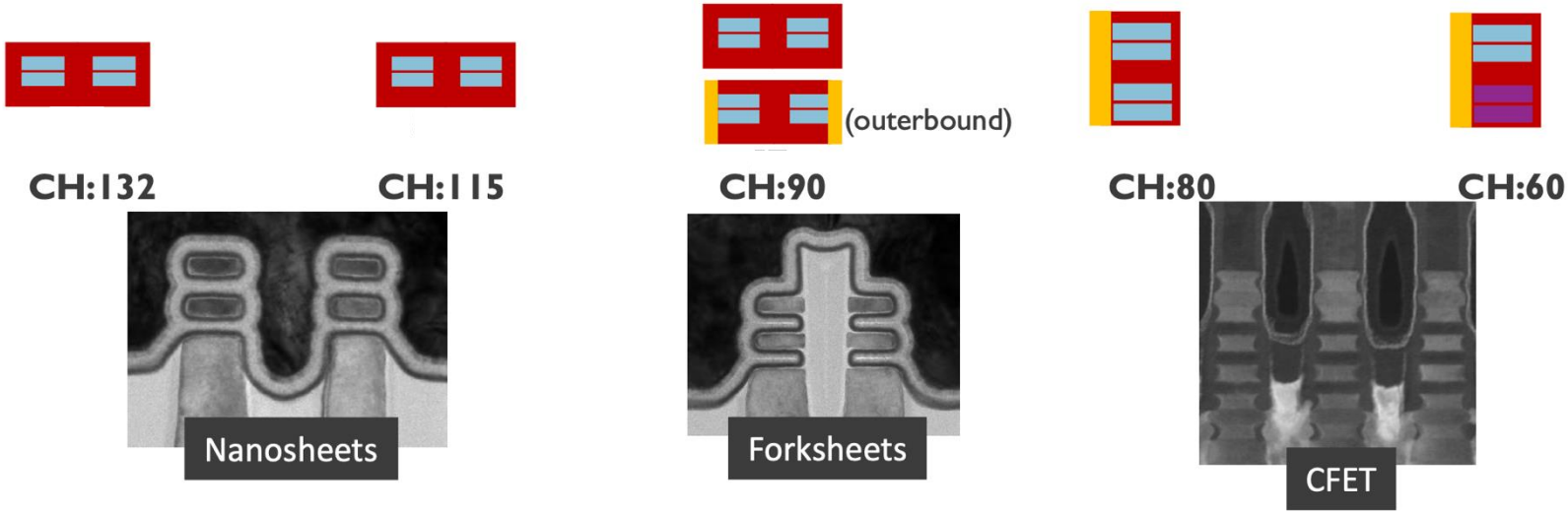
Beyond N20, device scaling is mostly about architecture optimization

CFET drives the future of logic scaling

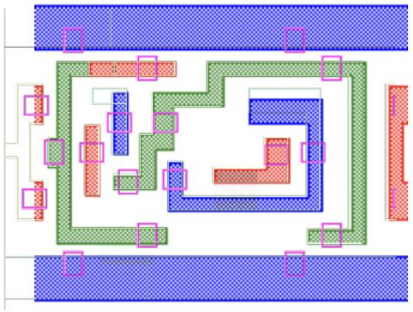
PP: poly pitch (nm)
MP: dense metal pitch (nm)
CH: cell height (nm)

Industry Timeline

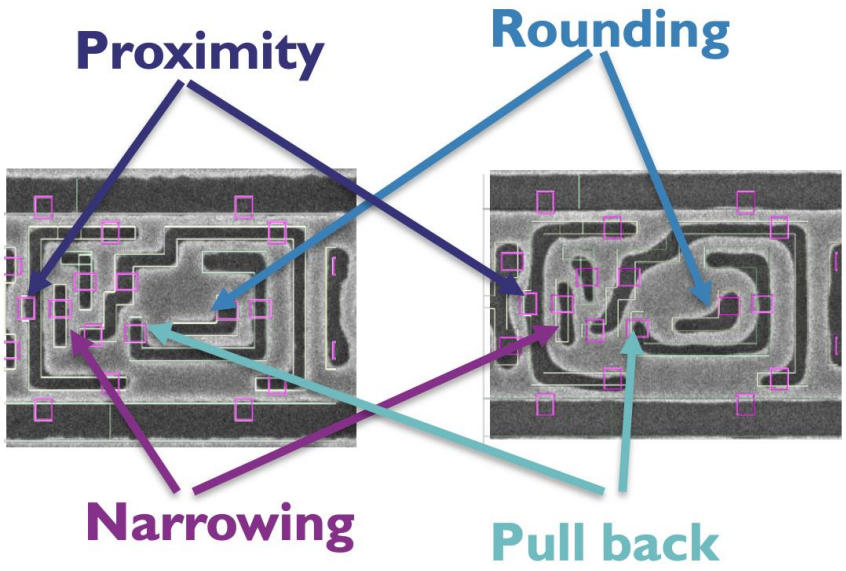
2025	2028	2030	2033	2035...
2nm PP:48, MP: 24-26	A14 PP: 45, MP: 20-22	A10 PP: 40-42, MP: 18-20	A7 PP: 40-42, MP: 16-18	A5 / A3 PP: 38-40, MP: 12-16



Enabling <100nm cell height requires device architecture innovations



Imec N10 layout with LELELE M1
(64nm CPP – 48nm MP)



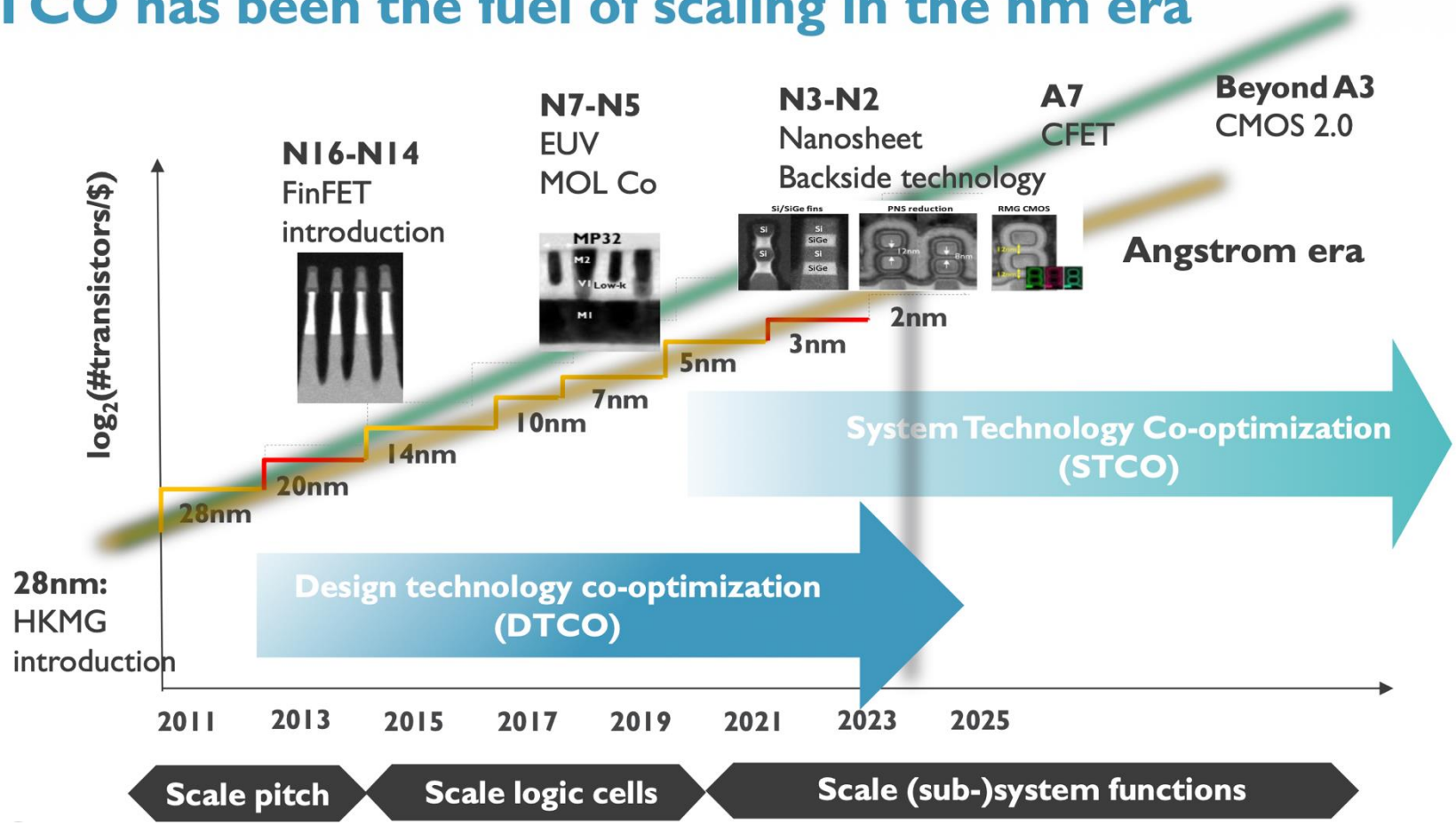
You draw this...

You expect this...

You get this...

The only way to resolve these scaling issues is to understand the circuit requirements to make the right adjustments in technology

DTCO has been the fuel of scaling in the nm era

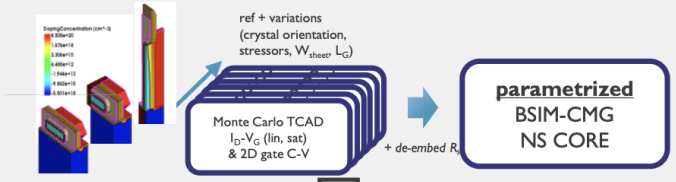


How modern ICs are designed – the core of DTCO infrastructure

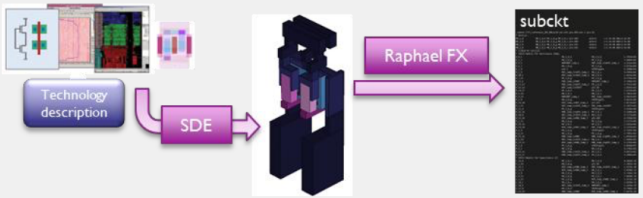
Design Enablement

Device modeling

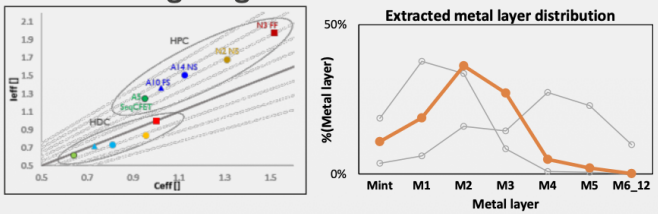
TCAD-based **model** (calibration to exp and publications)



Synopsys Raphael FX for extrinsic **parasitic extraction**

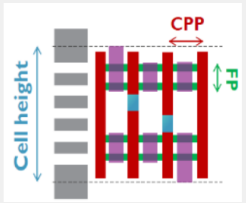


Targeting based on enhanced-RO



Std cell Lib / SRAM

Defining **cell templates** according to litho rules



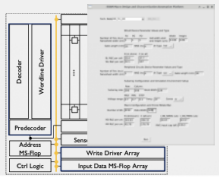
Techno options translated to mini-libs (~8 cells)



Full "virtual" Library
(~140 cells)

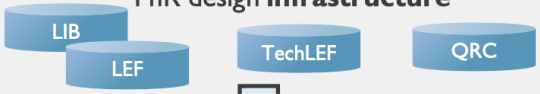


Memory compiler



Physical Design

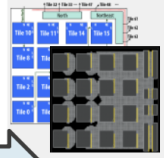
PnR design **infrastructure**



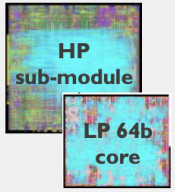
Various **design benchmarks**



Mempool many core distributed (RISC-V)



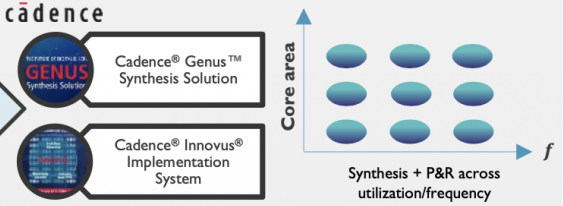
arm cores



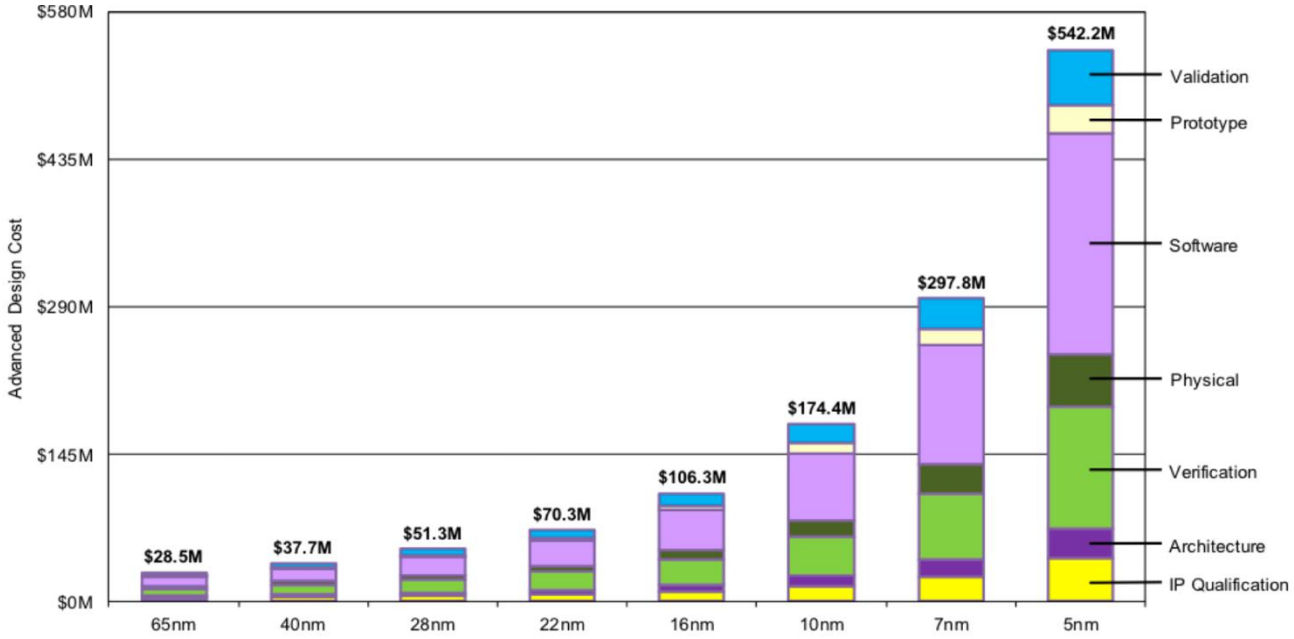
AI core mockup built internally



PPA evaluation driven by DTCO

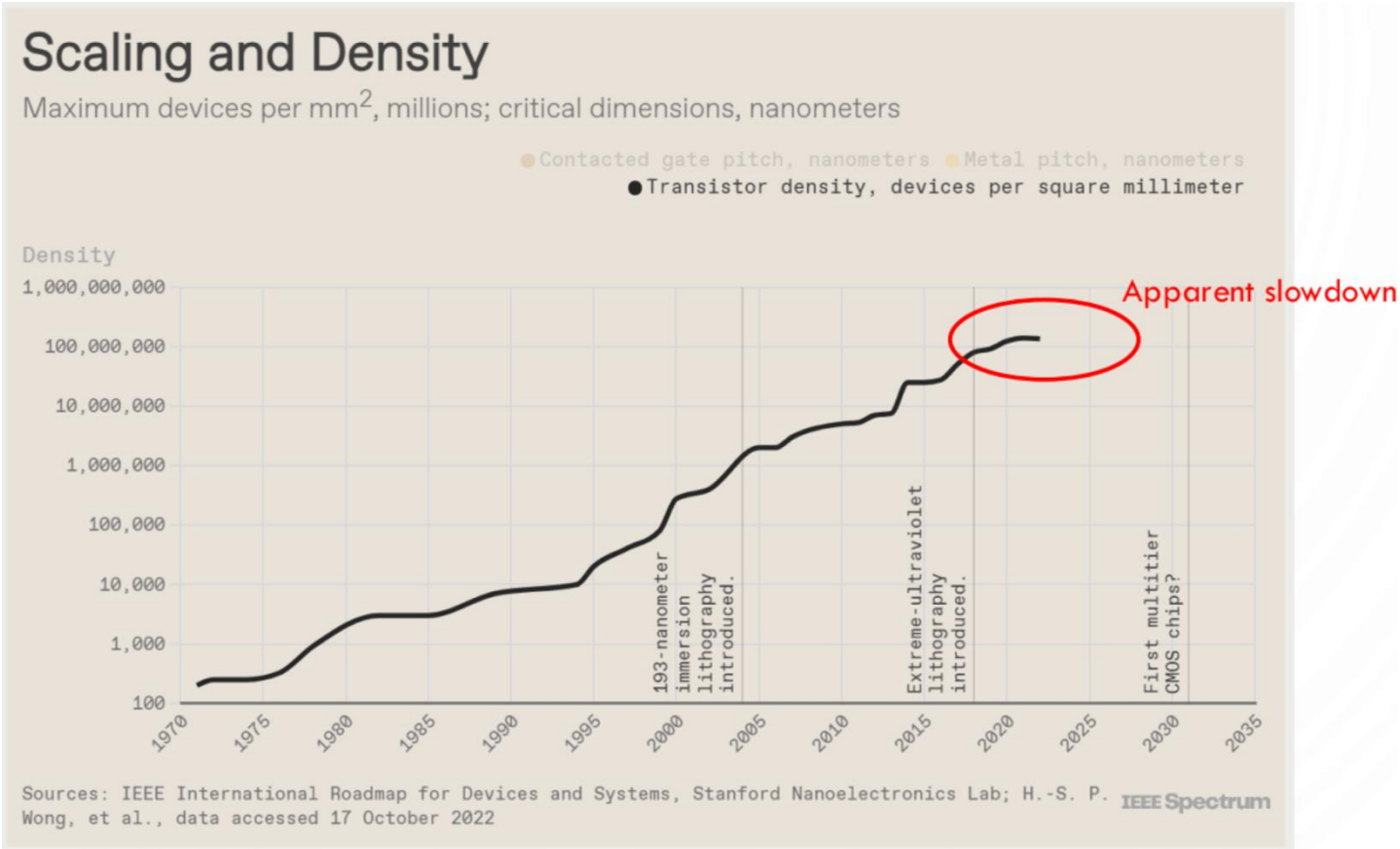


Cost Of Developing New Integrated Circuits

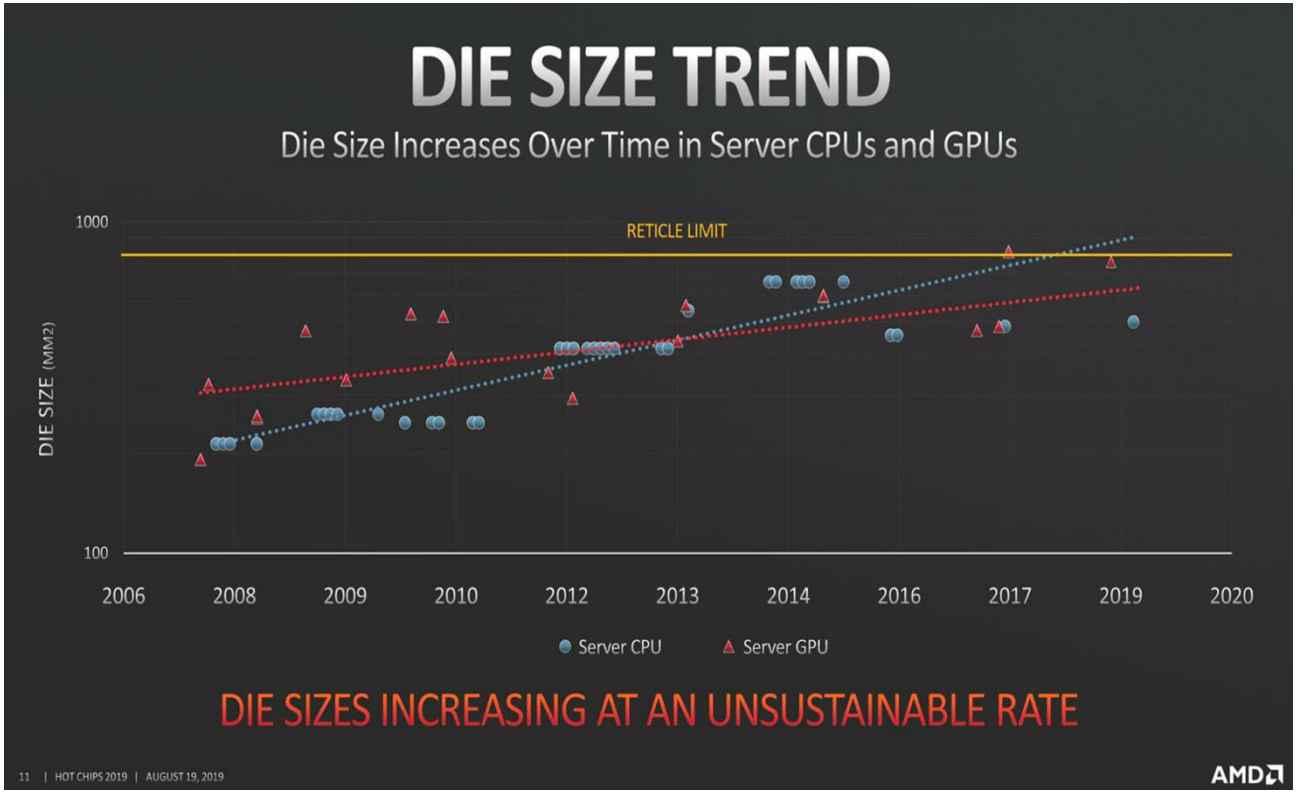


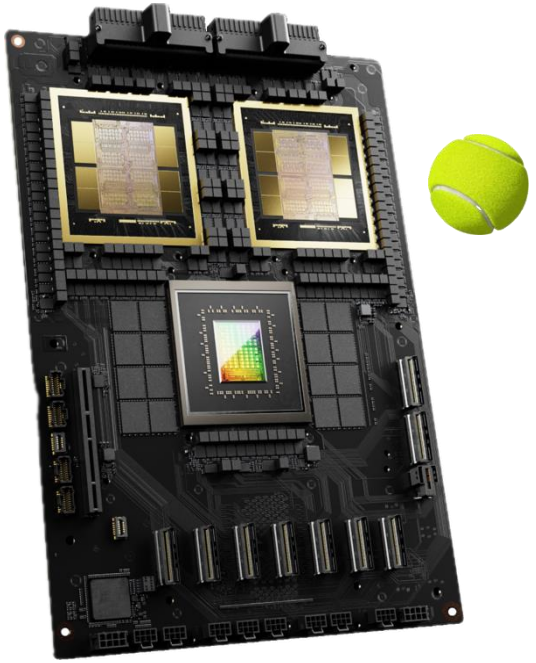
Cost of maintaining reliability!

Moore's Law considered ended sometime ~2021



To increase functionality and performance, silicon chip (called, die) sizes have been increasing.





Problem?
Chances of defects ⬆
Yield ⬇

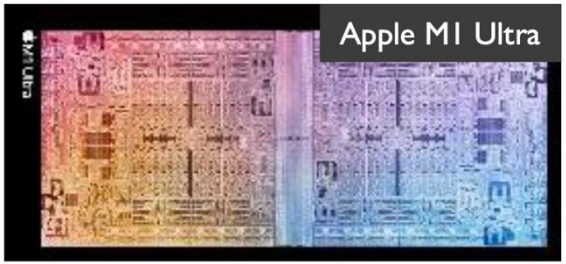
STCO: the era of smart disintegration

TECHNOLOGY

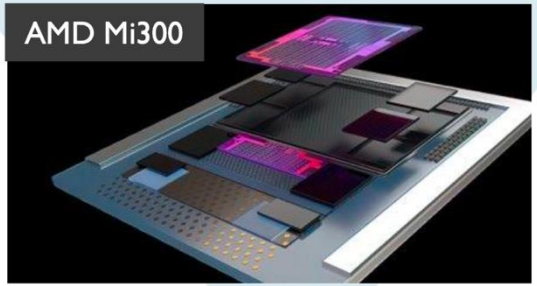
Advanced interconnect **solutions offering SoC-level connectivity performance**

APPLICATION

Diversity of Applications **requiring ever increasing compute and memory resources**



HP Mobile



HPC



AI/GPU

https://dl.acm.org/doi/10.1145/3154484
https://dl.acm.org/doi/full/10.1145/3563946

And even more applications to come...

AR/VR



6G Wireless



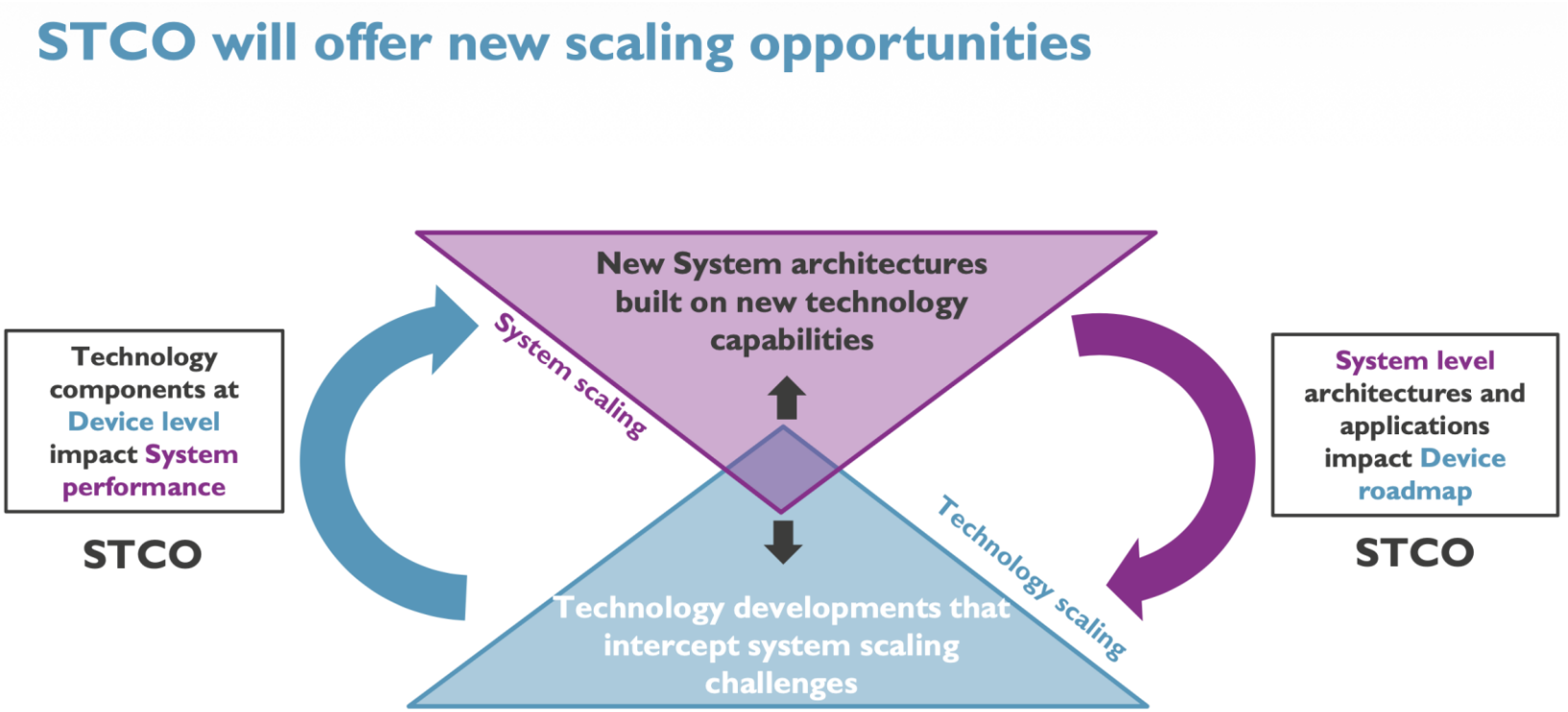
Automotive



A general purpose technology platform will end up in a *hardware lottery (S. Hooker)*

https://arxiv.org/abs/2009.06489

Future lies in System Technology Co-Optimisation (STCO)



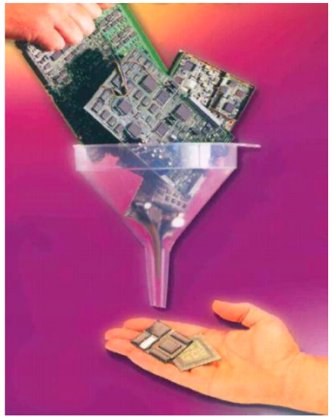
Continuing Microelectronic System Scaling

Monolithic CMOS single chip SOC system

Scaling continues through technology and device innovations

Challenges:
- Increasing complexity (cost)
- System performance limited by
  - access to data: Memory Wall
  - heat removal: Power Wall

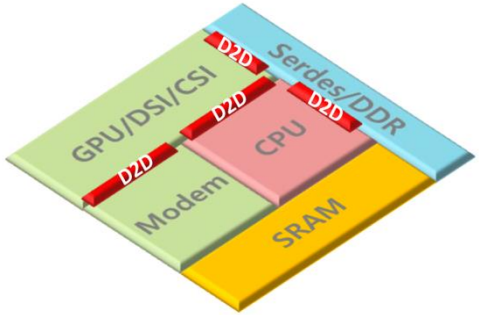
Solutions:
- Multi-Chip Heterogeneous Integration
- Functional System Partitioning
- High density 3D interconnect Technology



Migration to Chiplets

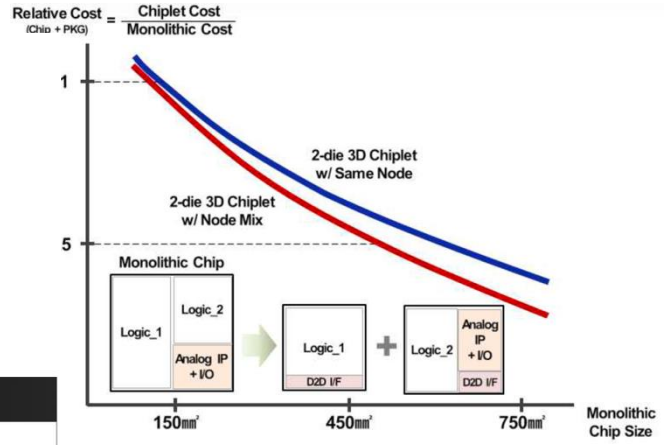


VS



Monolithic Design	Chiplet Design
• Large single die area • Low process yield • High cost due to low yield • Many different functions implemented in one die	• System split to chiplets each with manageable die size • High process yield • Low cost due to high yield • Each function can be designed in optimal process. • Reusable chiplets

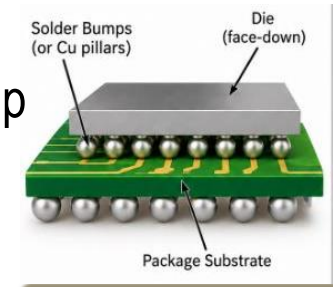
Reference) S-H. You, et al., IEDM 2020

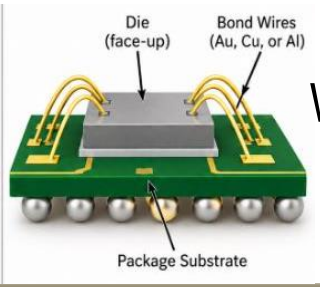


K. Kim. IEDM'21

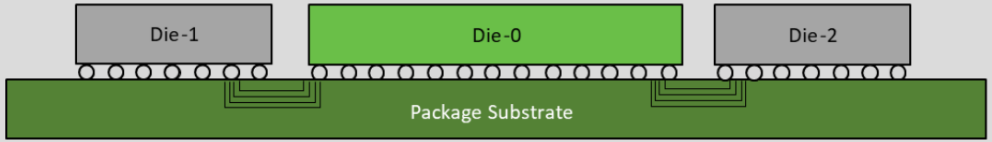
2D and 2.5D Integration

Flip-Chip





Wire-Bond

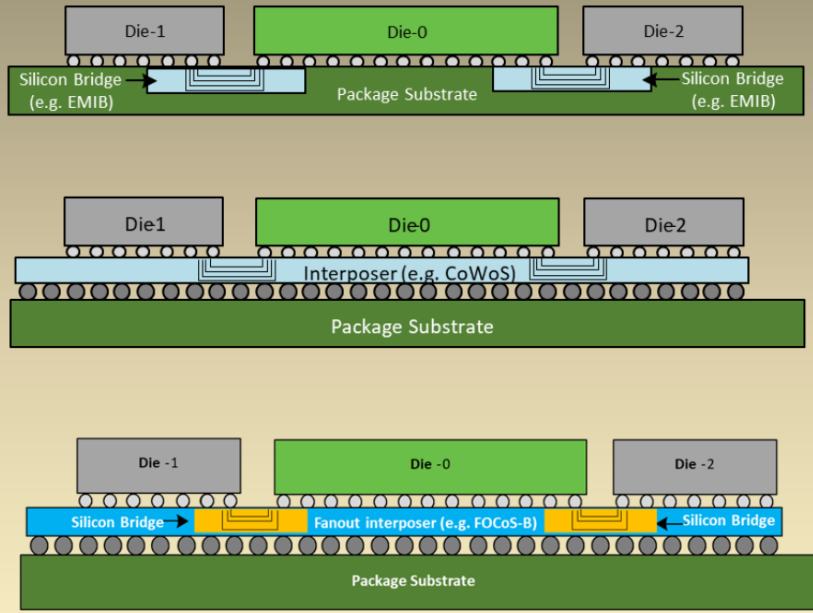


(Standard Package)

Standard Package: 2D – cost effective, longer distance

Advanced Package: 2.5D – power-efficient, high bandwidth density

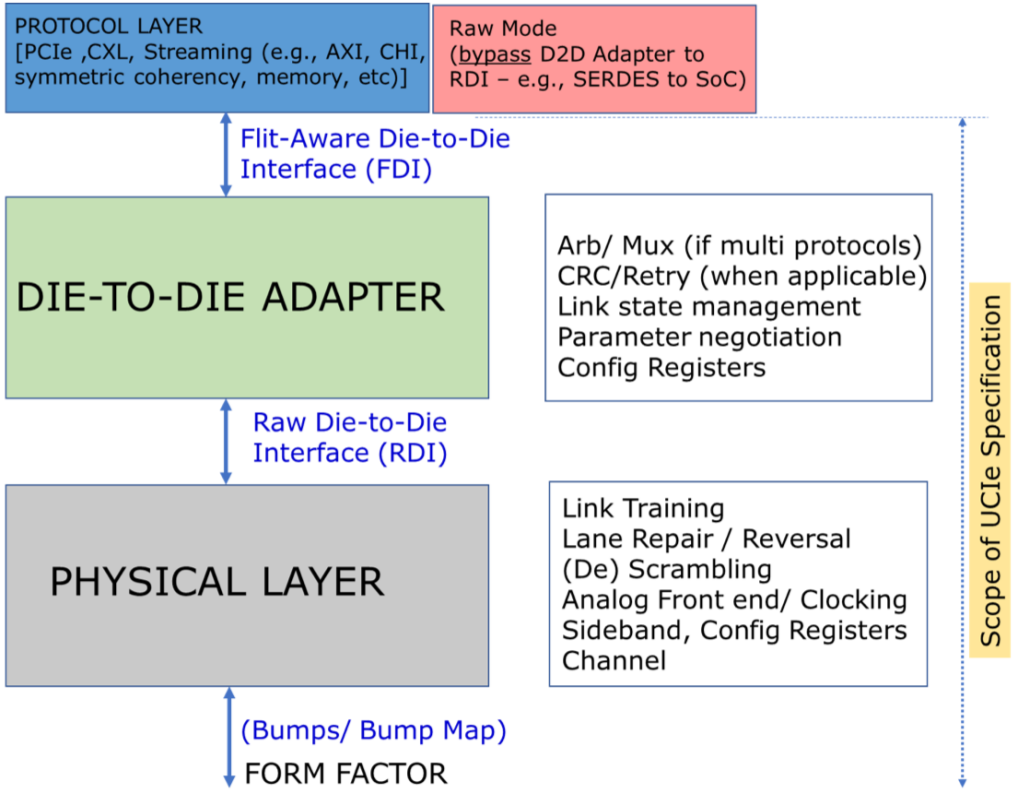
Dies can be manufactured anywhere and assembled anywhere – can mix 2D and 2.5D in same package – Flexibility for SoC designer



(Multiple Advanced Package Choices)

UCIe: A protocol for Inter-Chiplet Communication

- **Layered Approach**
- **Physical Layer:** Die-to-Die I/O
- **Die to Die Adapter:** Reliable delivery
  - Support for multiple protocols
  - bypassed in raw mode
- **Protocol:** CXL/PCIe and Streaming
  - CXL/PCIe for volume attach and plug-and-play
    - Common use cases: I/O attach, Memory, Accelerator.
  - Streaming for other protocols
    - Scale-up (e.g., CPU/ GP-GPU/Switch from smaller dies)
    - Protocol can be anything (e.g., AXI/CHI/SFI/CPI/ etc)



AMD AI accelerator the "Instinct MI300"

MI300a:

- 3 CPU chiplets ("CCD"s - compute complex die)
- 6 Accelerator chiplets (XCDs)
- 5nm TSMC – 2.2 Ghz peak clock

On top of:

- 4 active interposer input-output dies (IODs)
- 6nm TSMC

Connected to:

- 8 HBM3 high-bandwidth DRAM stacks (Samsung) (128 or192 GB – 5.2 GHz clock – 5.3 TBps)

On top of:

- A very large Si interposer

On top of:

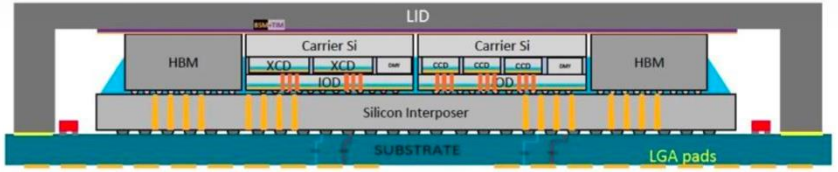
- An LGA Organic pakage substrate

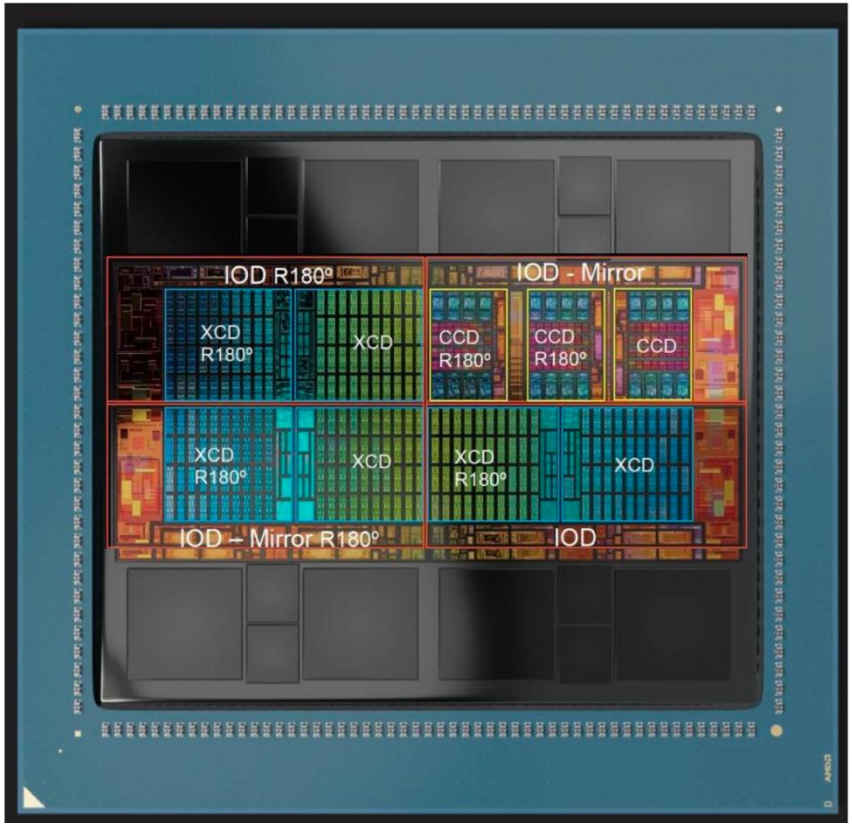
MI300x: accelerator-only system 8 XCDs, no CCD

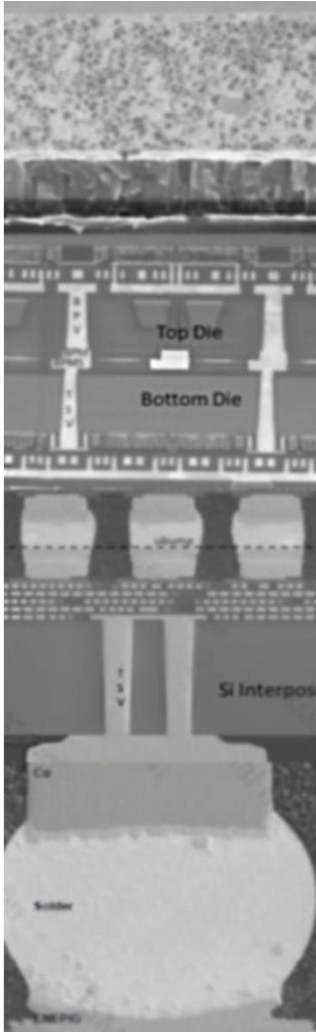
153 Billion transistors

750 W board

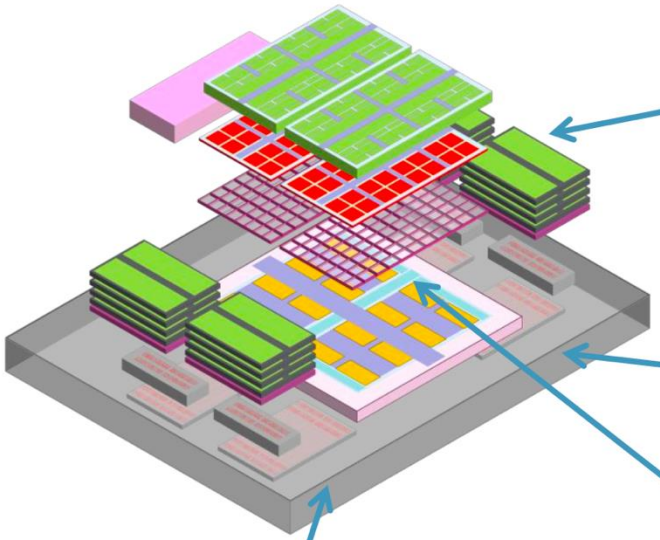
The MI300, will power the "El Capitan" subpercomputer (HP enterprise @ Lawrence Livermore Labs 2024)



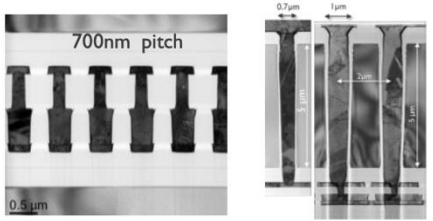




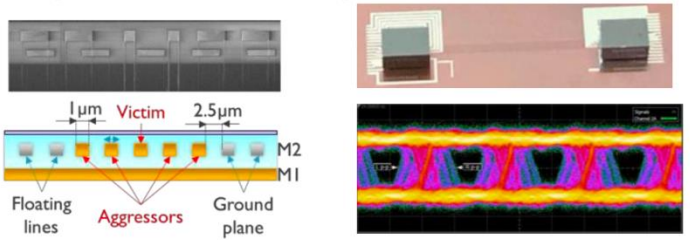
3DIC is an STCO technology driving SoC partitioning



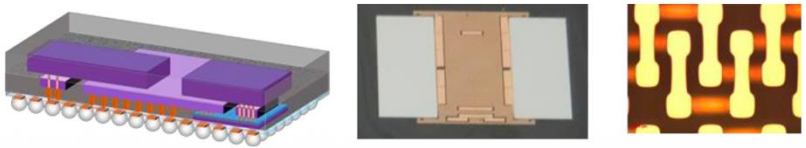
High density W2W bonding



High BW Si interposer/bridge

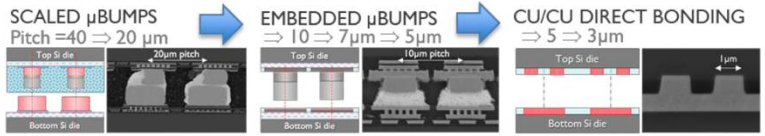


Advanced packaging technology



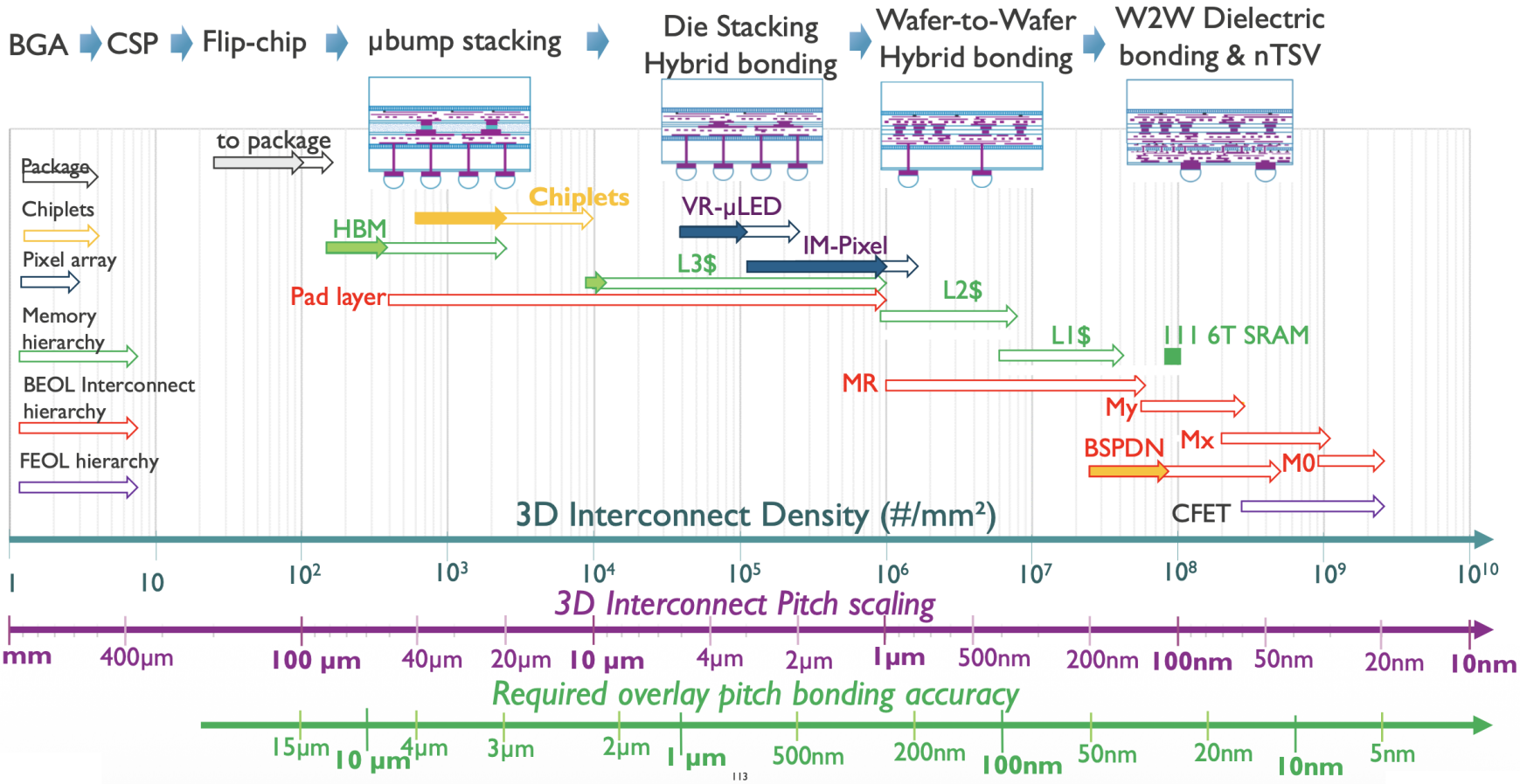
Reconstructed wafer – FOWLP – fine pitch RDL

D2W chiplets on (active) interposer



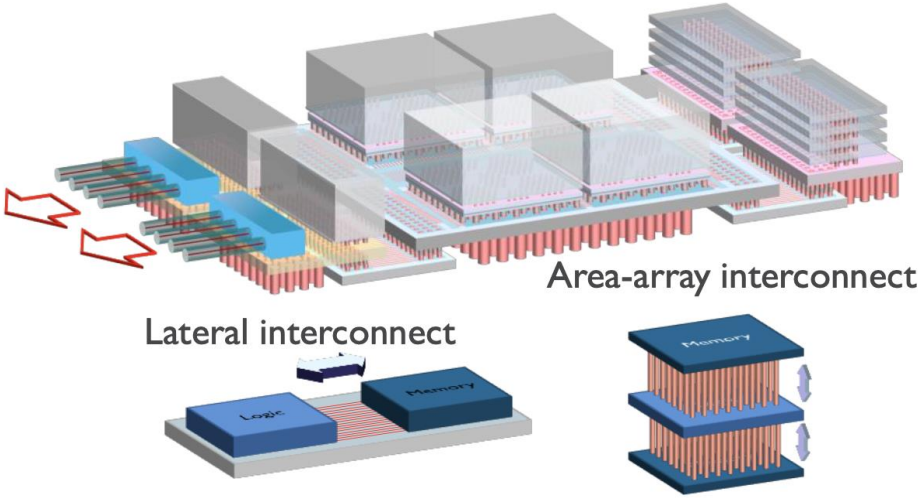
Scaling die-to-wafer interconnect

The 3D Interconnect density roadmap

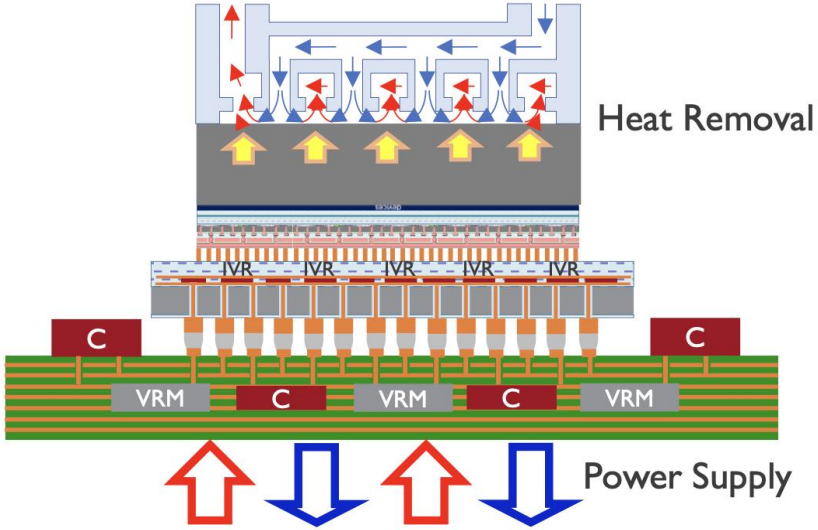


3D System Integration Technology opportunities

Drivers: "Memory-wall"



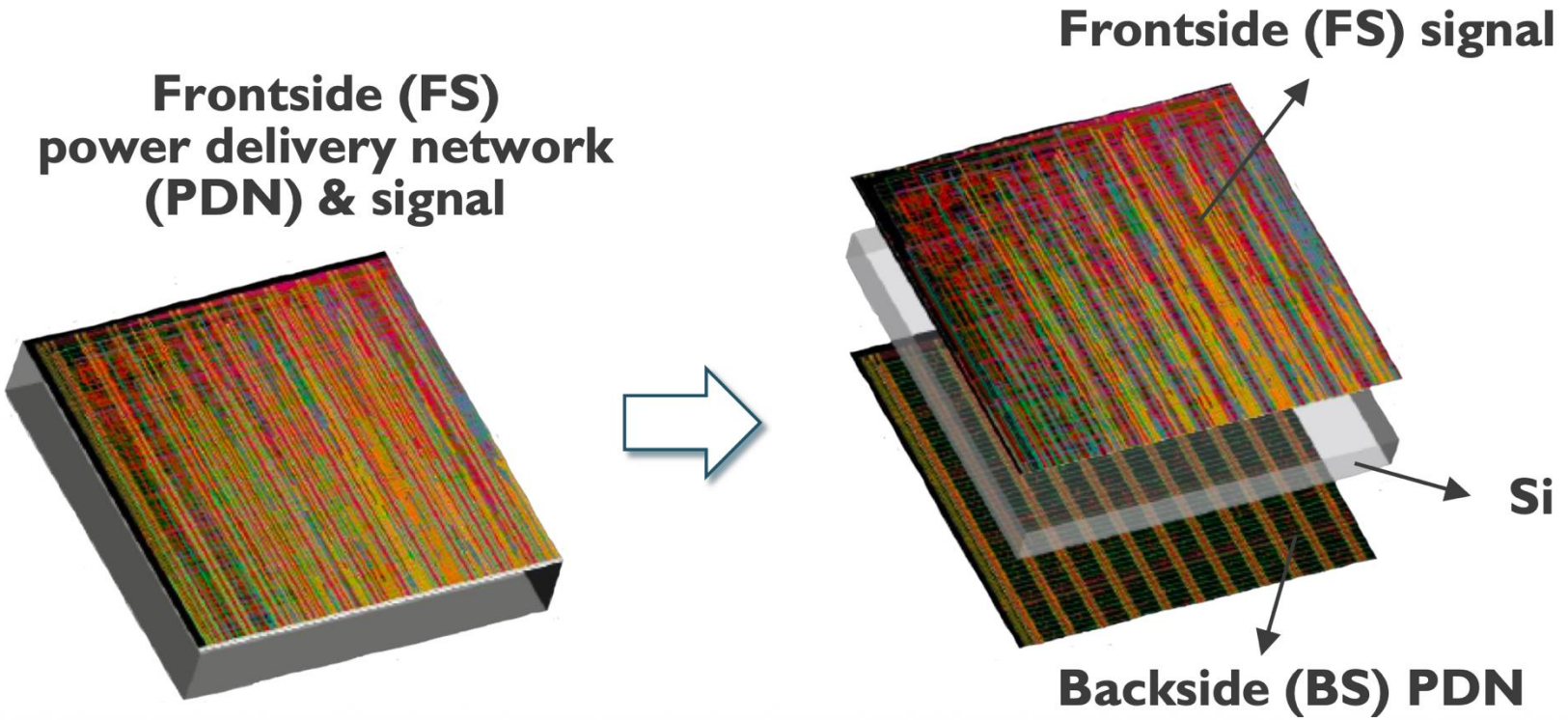
"Power-wall"



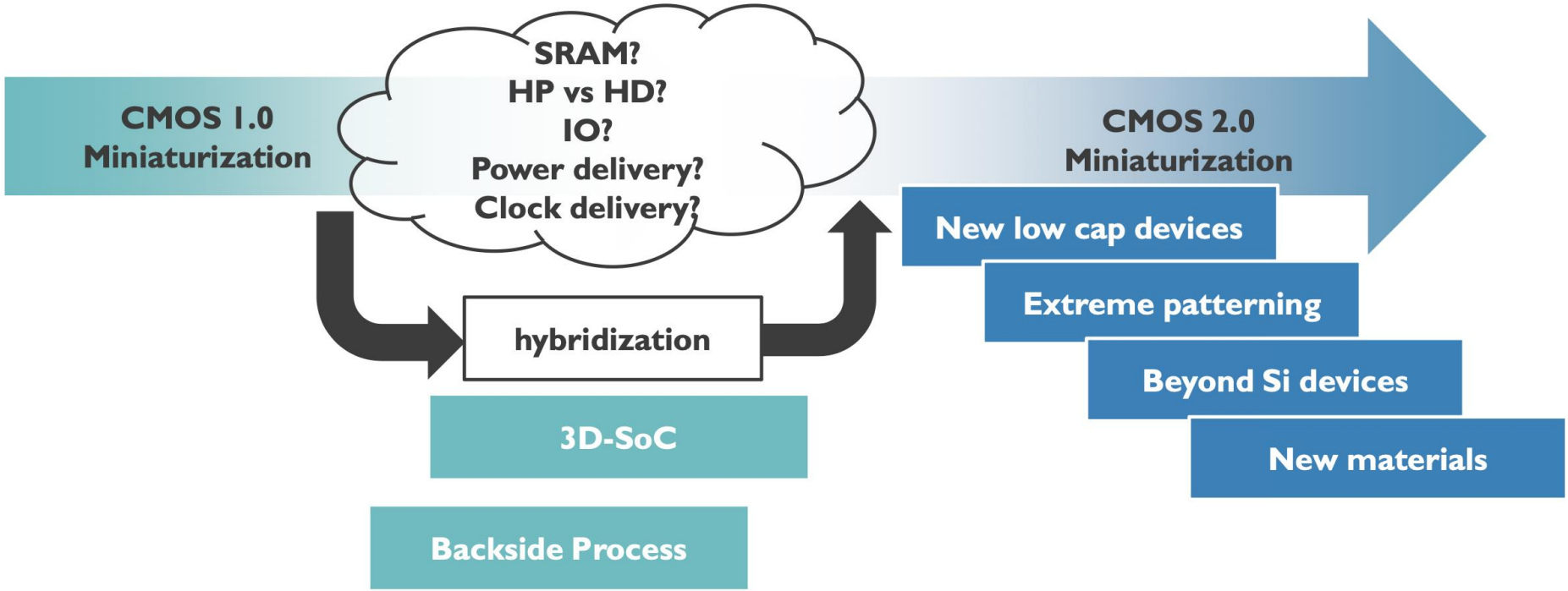
Memory logic partitioning
Need for High Bandwidth & low energy 3D interconnect

High Current delivery:
DC/DC V-conversion
High performance cooling

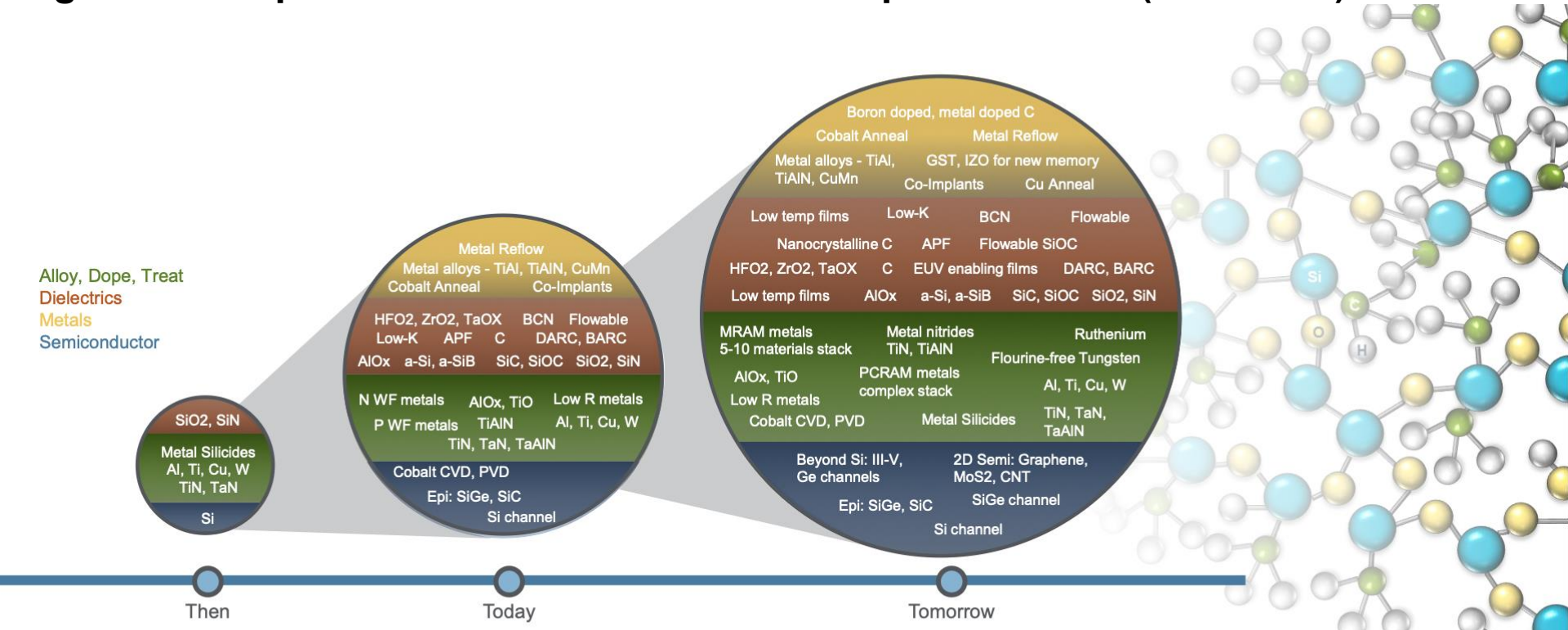
Backside Power Delivery Network (BSPDN): a game changer for on-chip power distribution



General purpose scaling challenges driving CMOS 2.0



Significant Expansion in New Materials for Chip Fabrication (CMOS 2.0)

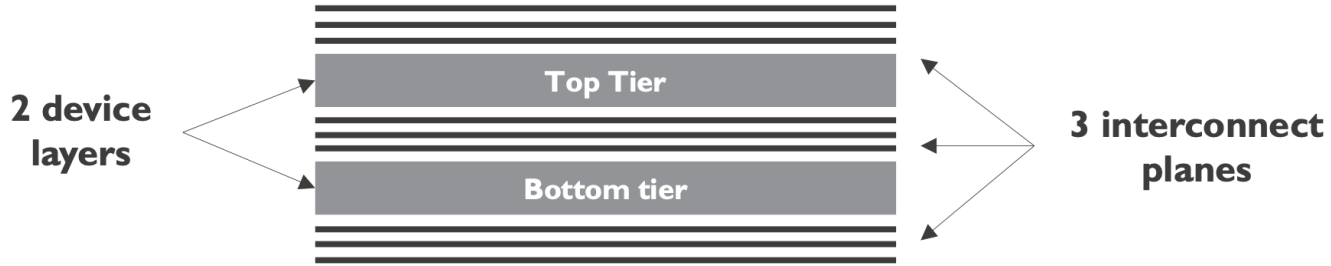


High device reliability and yield presents significant challenges

Source: Tutorials, ESSERC 2025

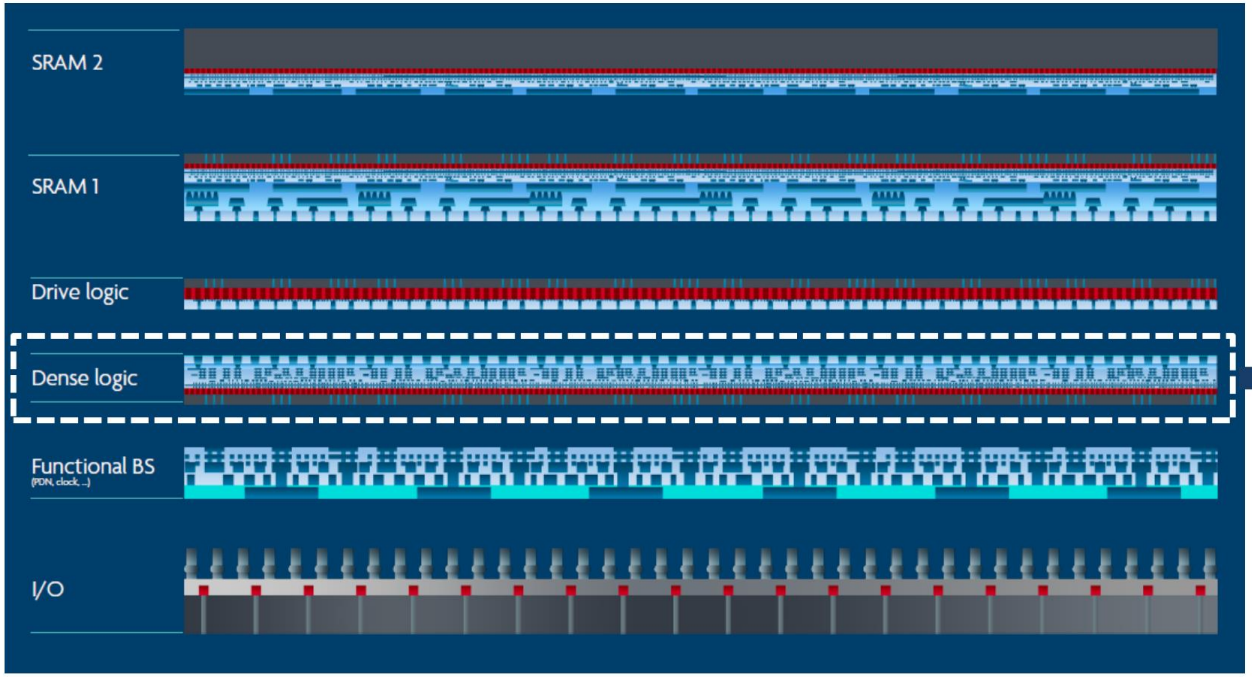
CMOS 2.0 revisits foundations of digital design

General case for a stack of 2 layers:



What is new wrt. 3DIC is it needs a co-design approach across tiers
➔ Requires a partitioning strategy
➔ Need to define tier heterogeneity
➔ The Routing is distributed across 3 interconnect planes
➔ Needs a revision of the infrastructure design (Clock Tree, PDN,...)

Rearchitecting towards a smart partitioning

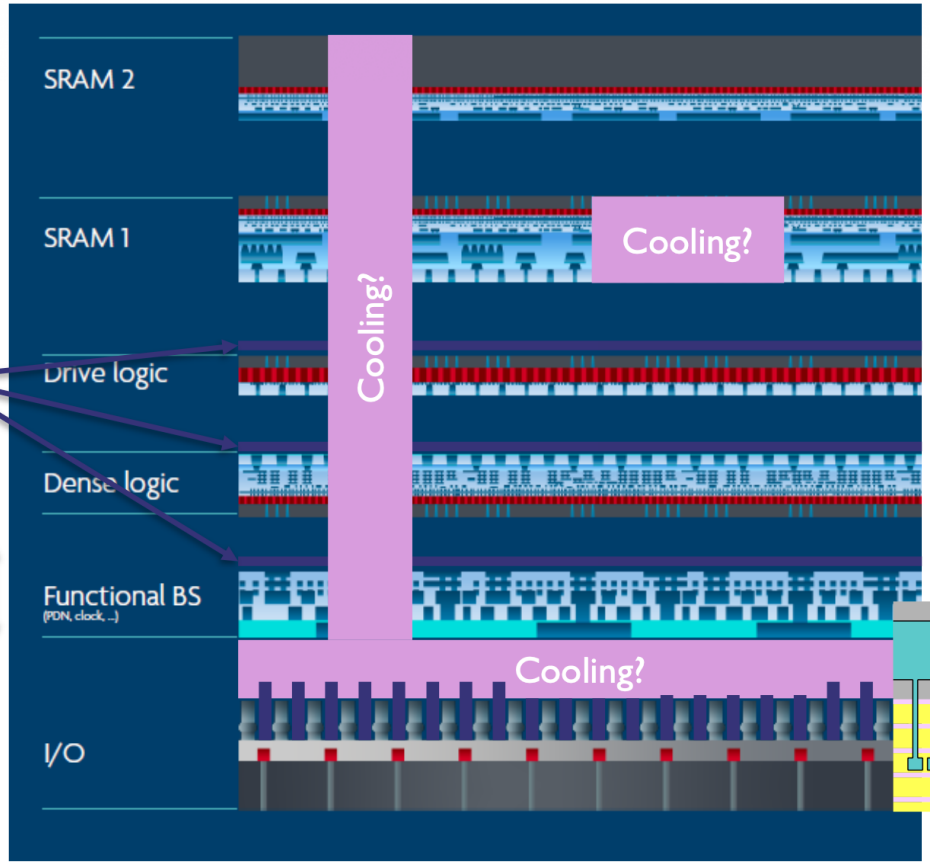


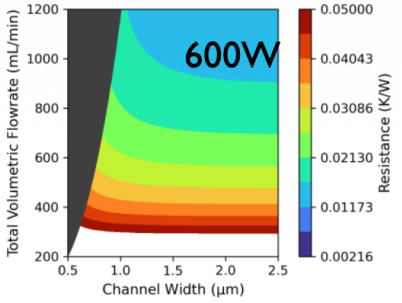
Highest Cost
→ Will drive scaling

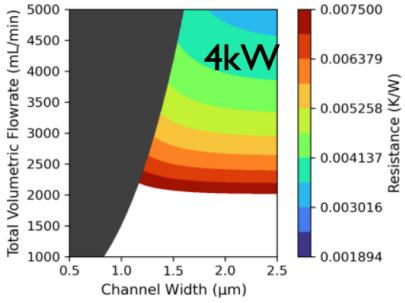
Same 'look & feel' as a classical CMOS platform

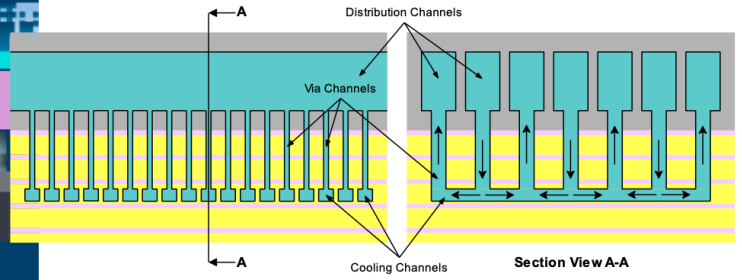
But offering versality for System optimization

It's not (yet) obvious how to cool CMOS2.0

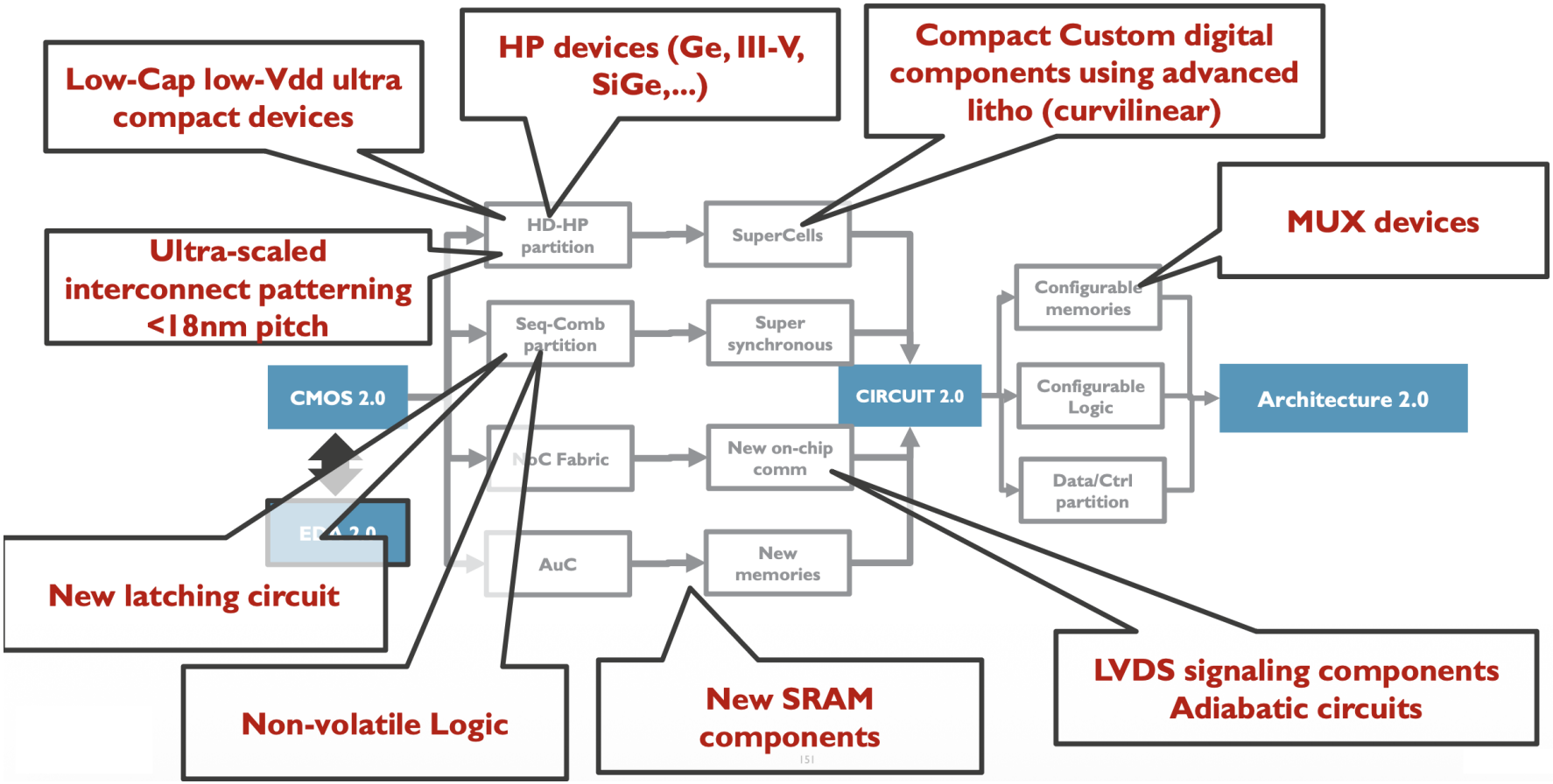






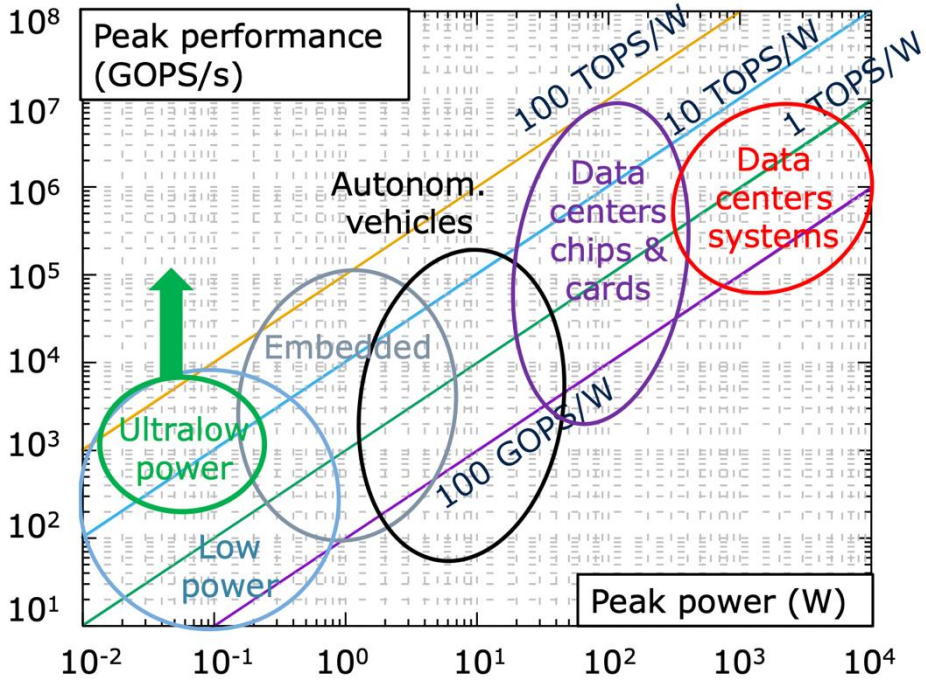


CMOS 2.0 opens the door to new component research



Application Areas

The core idea behind any microelectronics project?



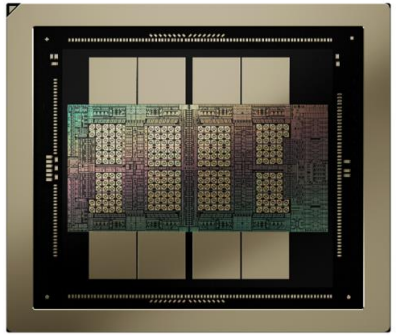
Given a functional description and performance, cost & power constraints, create an implementation using basic building blocks or using a set of primitives, or both.

Microelectronic design is a bit an art as well as a science. The creative spirit is critical in combining primitive elements & other components in new ways to achieve a desired function.

However, unlike art, we have objective measures of a design:
Power, Performance and Cost (Area)

AI Demands the Best of Every Technology

- Massive Compute Horsepower
  - Tensor cores designed for AI compute
  - 50 PFLOPS NVFP4 on Rubin
- High-Bandwidth Memory Technologies
  - 22 TB/s HBM4 Memory Bandwidth
- High-speed I/O
  - 3.6 TB/s NVLink Bandwidth per GPU
- Datacenter Scale Design
  - Innovations in packaging, cooling, interconnect, networking, manufacturing, and reliability



Rubin

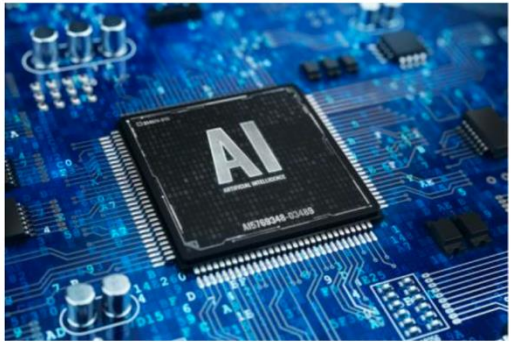


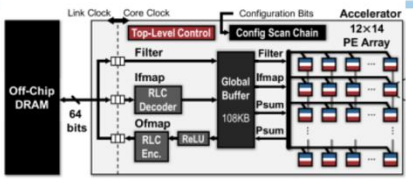
Vera Rubin Pod

Source: ISSCC2026

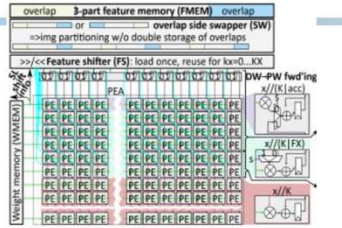
Deep learning Accelerators

Many deep learning accelerators

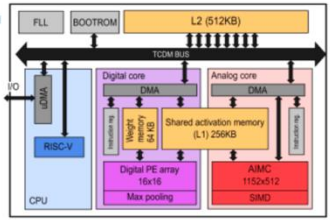




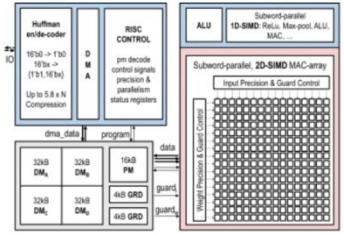
[Eyeriss, 2016]



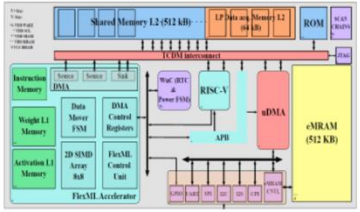
[DepFiN, 2021]



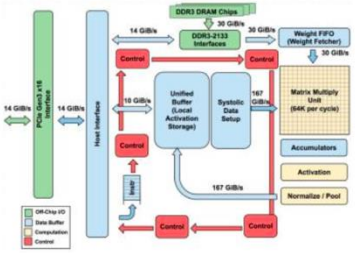
[DIANA, 2022]



[Envision, 2017]



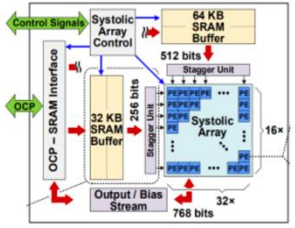
[TinyVers, 2022]



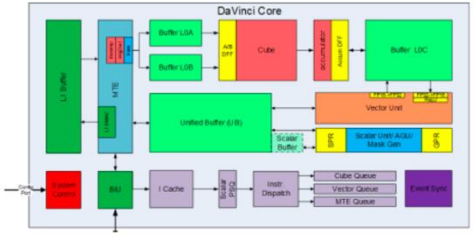
[Google TPU-v1, 2017]

But all using similar tricks!

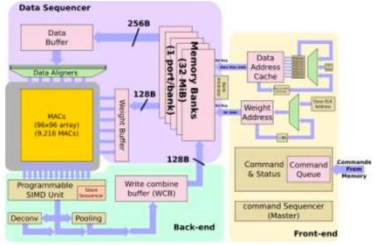
1| Parallelism & spatial reuse
2| Stationarity & temporal reuse
3| Reduced precision
4| Sparsity



[Meta prototype, 2022]



[Huawei Ascend, 2019]



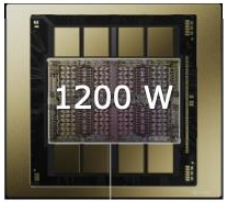
[Tesla NPU, 2019]

Power Scaling at the Rack Level

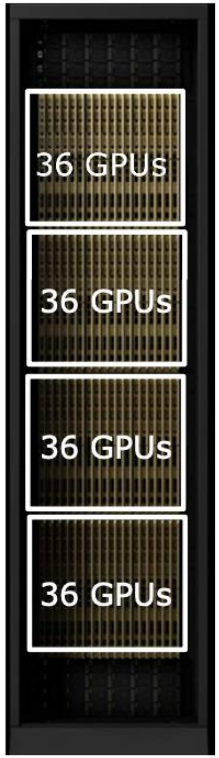
~120 kW Rack Power



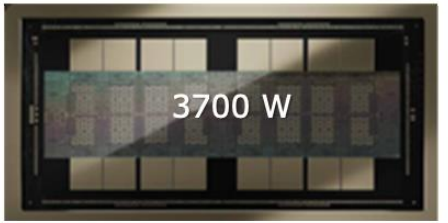
72 x Blackwell GPUs



~700 kW Rack Power

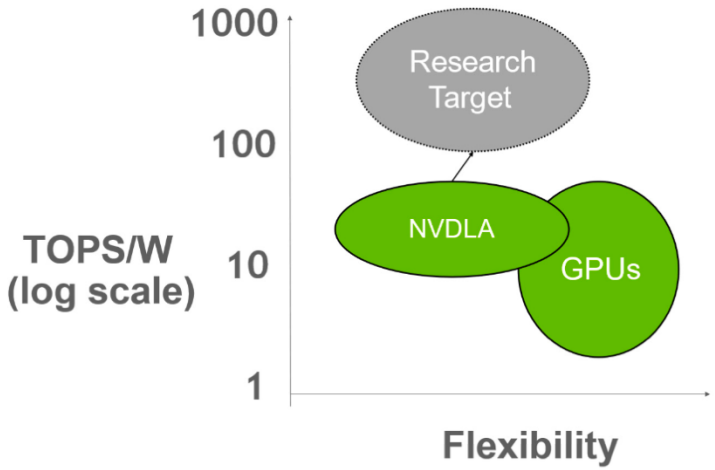


144 x Rubin Ultra GPUs

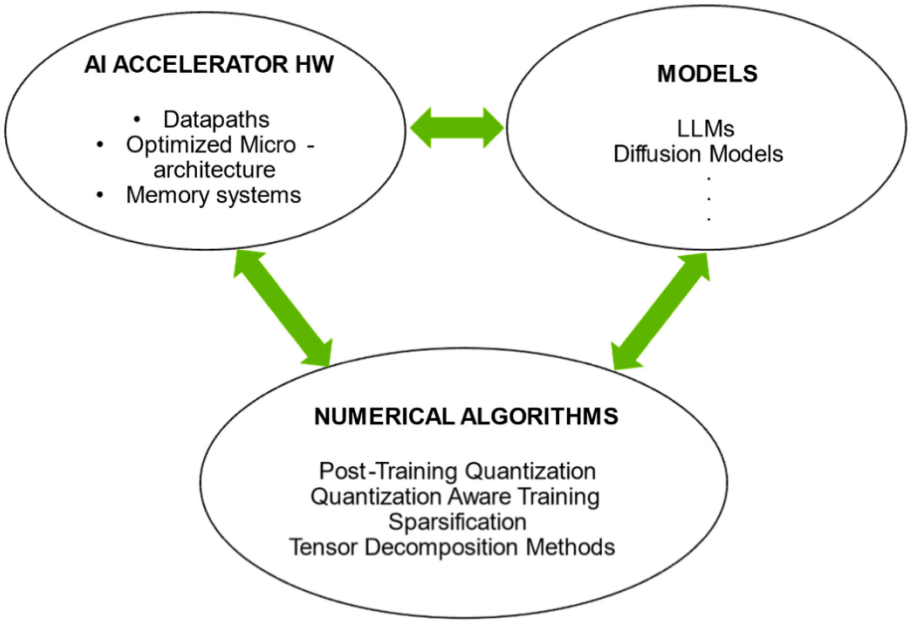


From: GTC March 2025 Keynote with NVIDIA CEO Jensen Huang

Scaling Efficiency to >100 TOPS/W

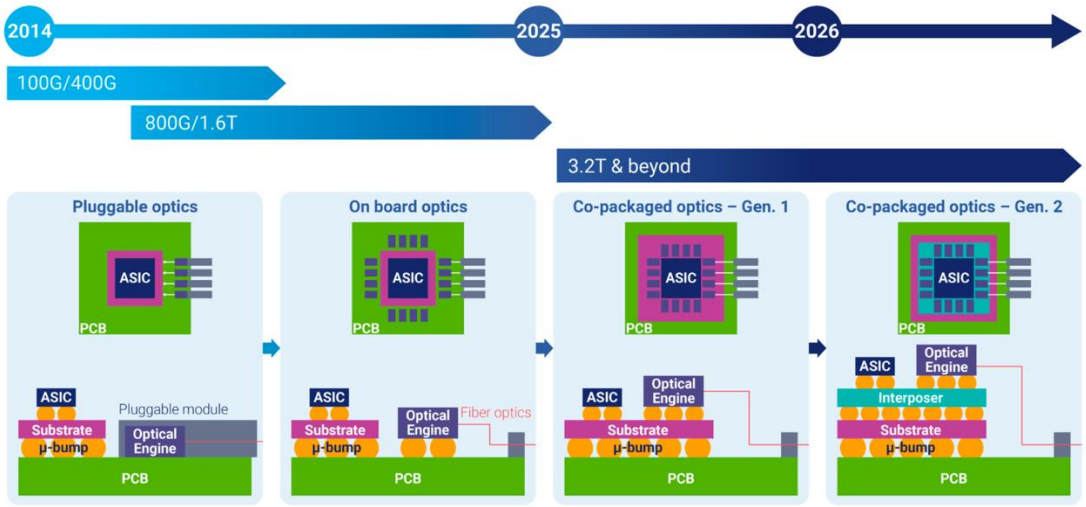


Software / Hardware Co -design

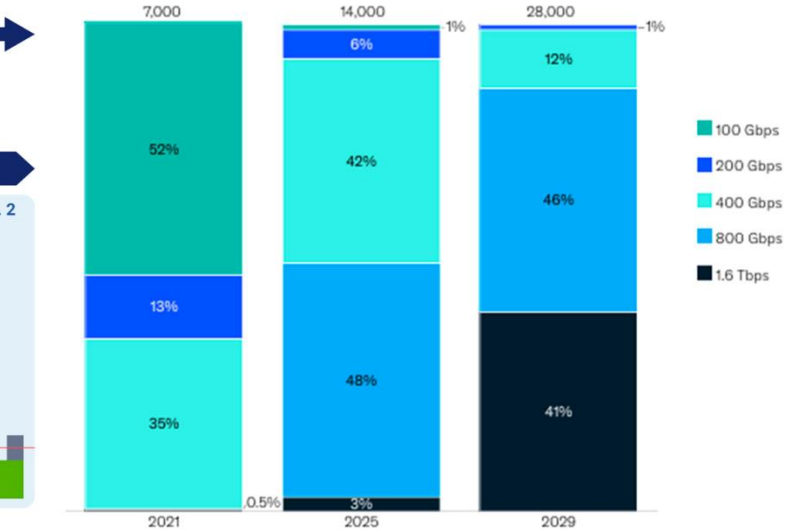


Optical Transceivers and Co-packaged Optics

[Source: IDTechEx Research]

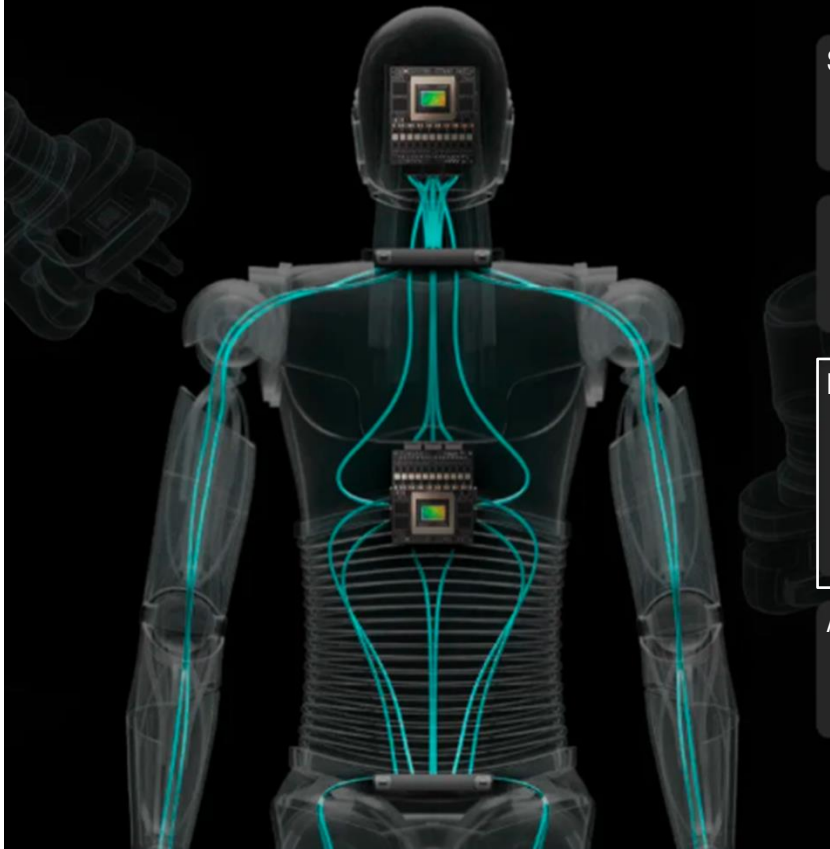


[Source: McKensey & Company]



Robotics Demand Rich Sensors and Low-Latency AI Processing

Multiple AI Models and a Dozen Diverse Sensors



Software & AI — High Level Reasoning [1–5Hz]

- Natural Language Interaction
- Scene Understanding
- Symbolic/ Goal Planning
- Task Planning/ Scheduling

Control Theory — Perception and Planning [30Hz]

- Scene Registration and Self-localization
- Object Recognition and Localization
- Grasp and Motion Planning
- Locomotion

DSP & MCU — Real-Time Control Framework [100Hz–1KHz]

- Model Predictive Control/AI Control Policy
- Velocity Controller
- Position Controller
- Impedance Controller
- Torque Controller

Analog — Hardware Abstraction Layer

- IMU
- Audio and Mic
- Actuators
- Haptics/ Touch Sensors
- Cameras
- Position Sensors

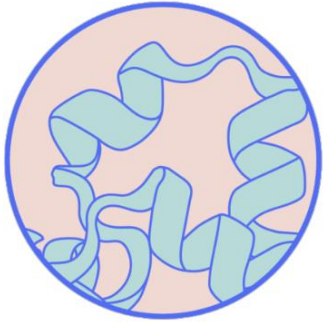
Source: nVIDIA

Quantum Computing

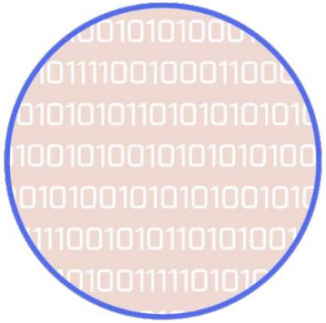
Encryption/decryption



Protein folding



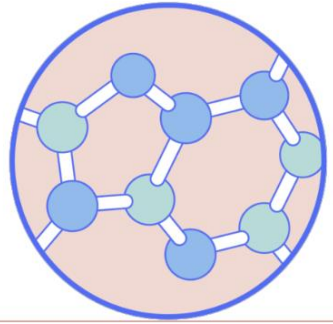
Big data



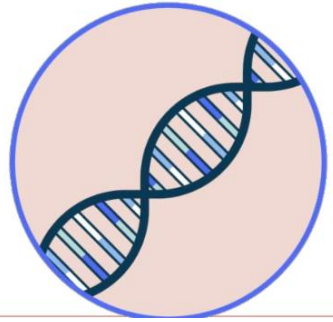
Drug synthesis



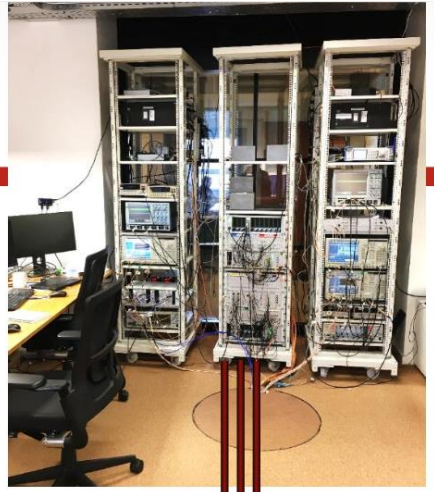
Molecule simulation



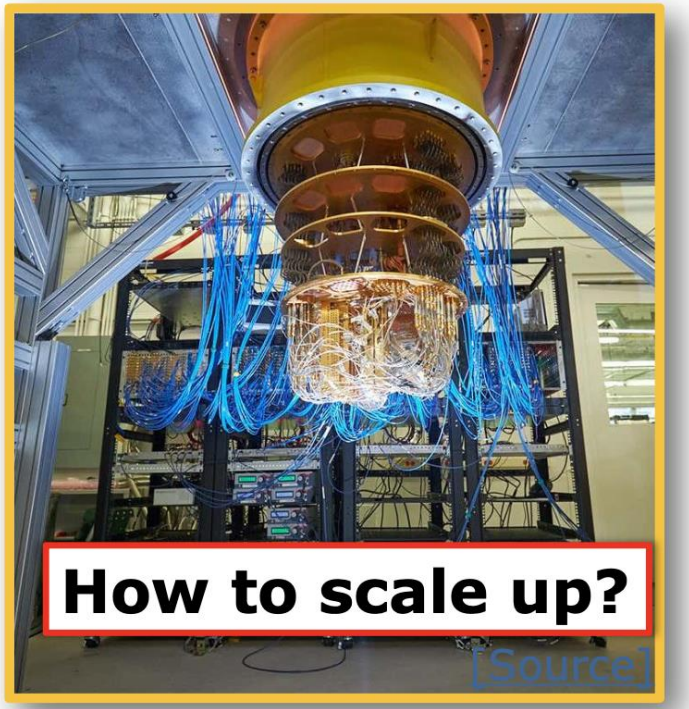
DNA analysis

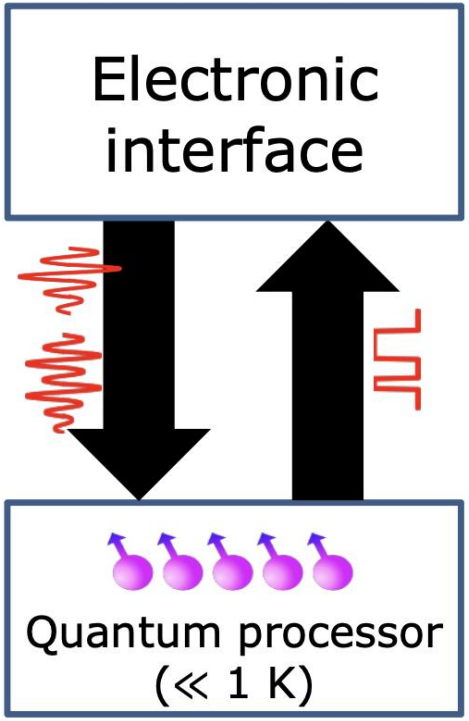


A real-life quantum computer



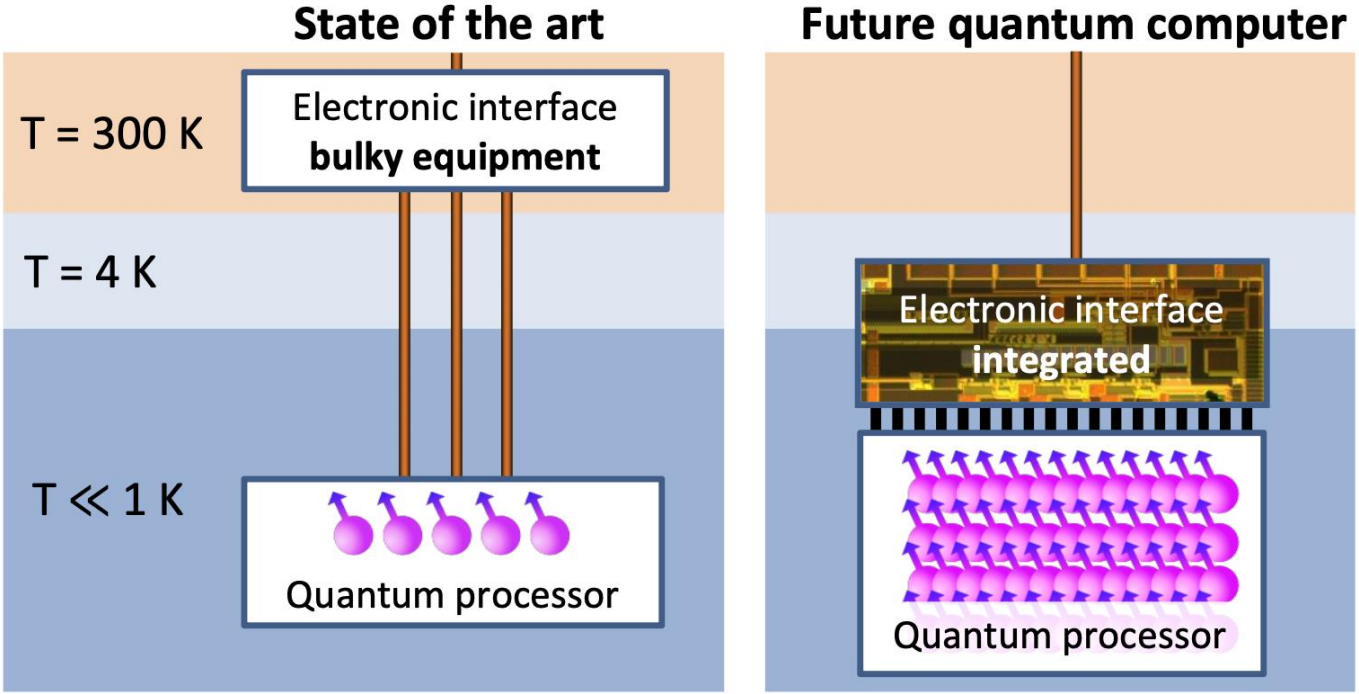






F. Sebastiano, ISSCC2025

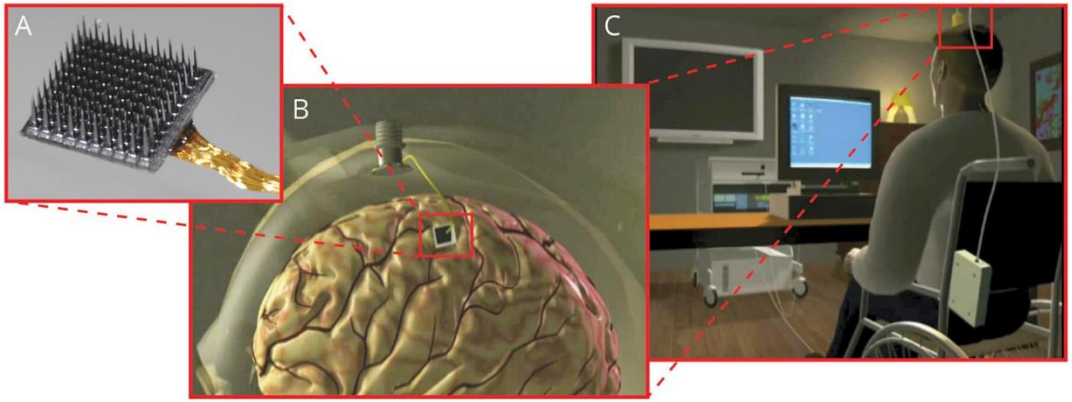
Cryo-CMOS: Towards a large scale quantum computer



F. Sebastiano, ISSCC2025

Brain-Machine Interfaces

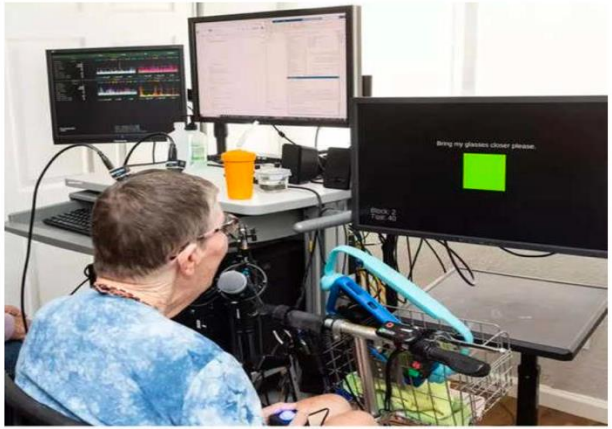
Ultimate Goal: Restoring Movement and Communication after Paralysis



(A) The Utah microelectrode array (Blackrock Neurotech): **10×10 array** on a **4×4-mm** platform
(B) The array is connected through a bundle of wires to a percutaneous **pedestal**
(C) Signals from the pedestal are transmitted through a cable or wirelessly to **external devices** for processing and decoding



Next-Generation BMIs: Efficiency & Intelligence

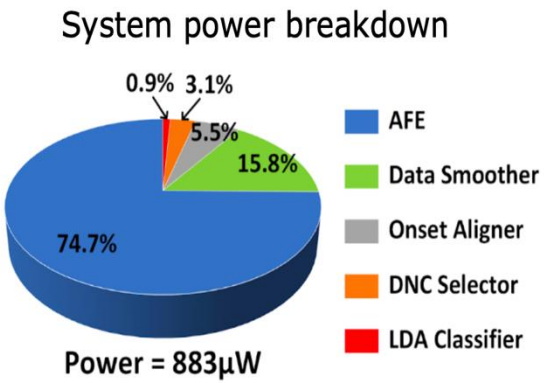




✗ **Limited mobility**
✗ **Physical size**
✗ **Power consumption**

✓ **Efficiency**
✓ **Convenience**
✓ **Intelligence**

Distinctive Neural Codes (DNC) for Brain-to-Text Communication

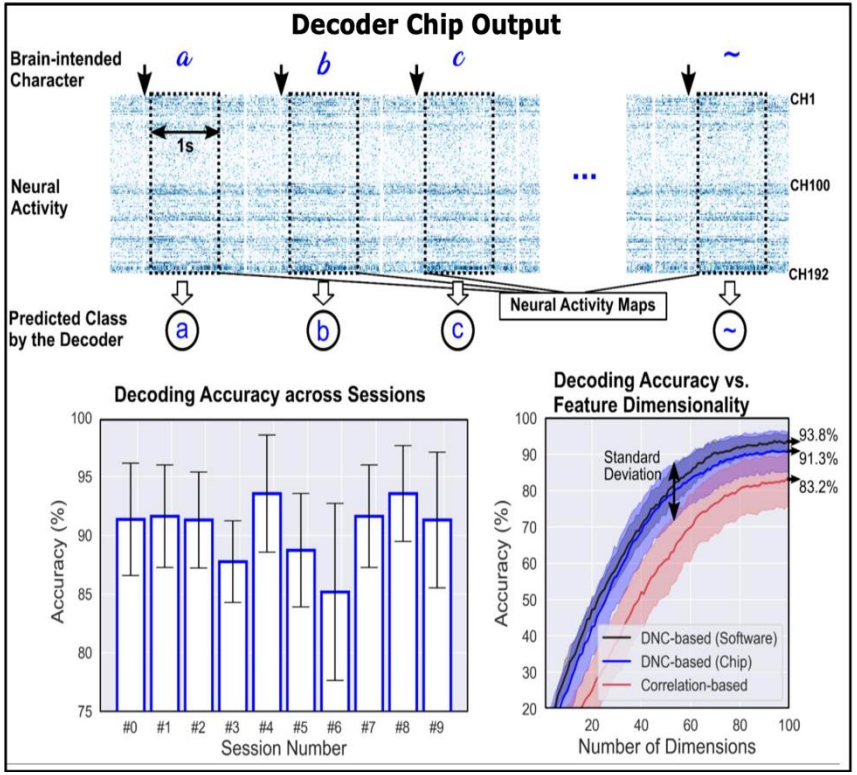


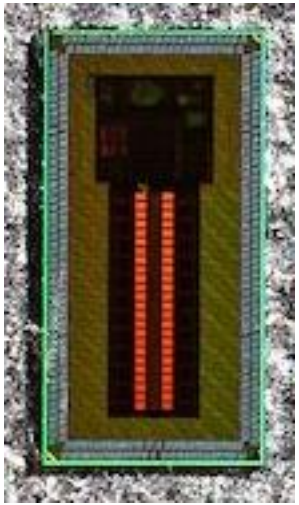
DNC decoder chip: 65nm TSMC, total power: **223µW** (512-Ch.), activity-driven

Decoder area: **0.75mm²**

Measured accuracy: **91.3%**

Speed: **60** Characters per minute





M. A. Shaeri et.al. ISSCC2024

From imaging to sensing diversity

2D **Color imaging**

Traditional 2D imaging either color or monochrome

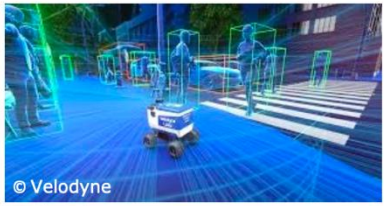




3D **Depth sensing**

Bringing the 3rd dimension to perform 3D sensing





4D **Time, motion sensing**

Bringing the ultra low temporal resolution

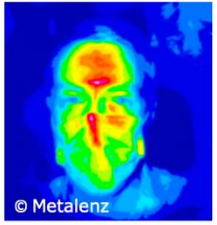




5D **Multispectral, polarization**

Bringing spectral and polarization information beyond the visible





Sensing for Extended Reality

VR

2-6 cameras

AR & MR

4-10 cameras

Spatial computer?

14+cameras?

Tracking cameras x4



Meta Quest 2
~$350

Tracking cameras x4

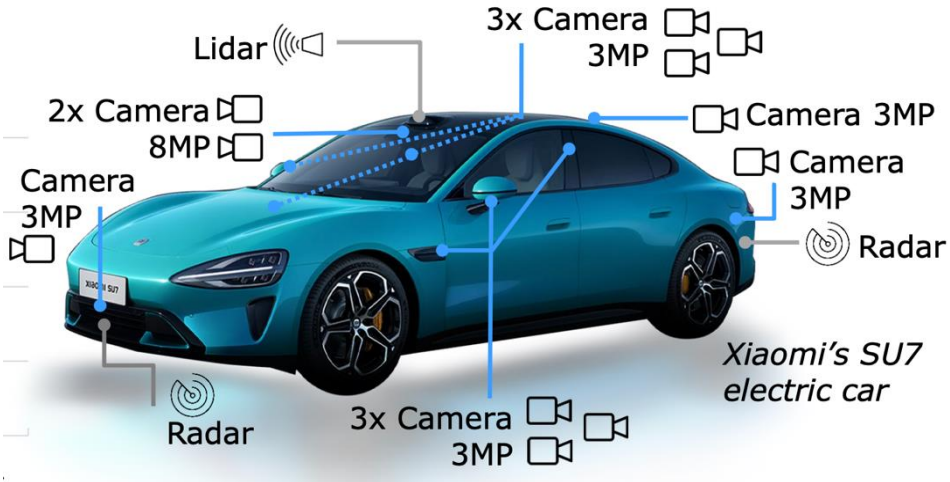
RGB camera

3D camera



Microsoft HoloLens 2
~$3,500

Eye tracking



Xiaomi's SU7 electric car

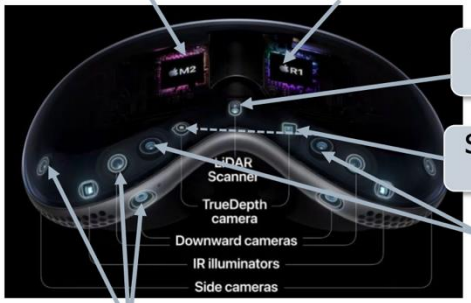
Battery



Apple Vision Pro
~$3,500

M2 APU

R1 co-processor



dToF 3D camera "LiDAR"

Structured light 3D camera

RGB cameras x2

Tracking cameras x6

Eye tracking cameras x4

IoT for Healthcare

Aging Population

- Proportion of elderly people has increased from ~10% to ~20% in the last 50 years [1].



Chronic Diseases Prevalence

- By 2020 they can represent 60% of the global burden of diseases [1].



Increased Healthcare Costs

- Global healthcare spending was increasing at an annual rate of 5.4% in 2017–2019 [2].

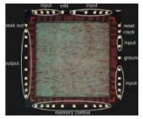


[1] Harris, R. E. "Epidemiology of chronic disease: global perspectives". Jones & Bartlett Learning, 2019.
[2] Deloitte. "2019 Global Health Care Outlook. Shaping the Future", 2019.

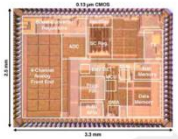
Source: ESL Lab, EPFL



IMEC cardiac patch
(Yazicioglu,2009)



Heart Rate Monitoring
(Massagram, 2010)



Zhang, 2012



Toumaz's Sensium Life
(Wong, 2012)



Corventis's PiiX
(Corventis cardiac systems, 2014)



Holst Centre
(Masse, 2016)



Shimmer3
(shimmerresearch.com, 2017)

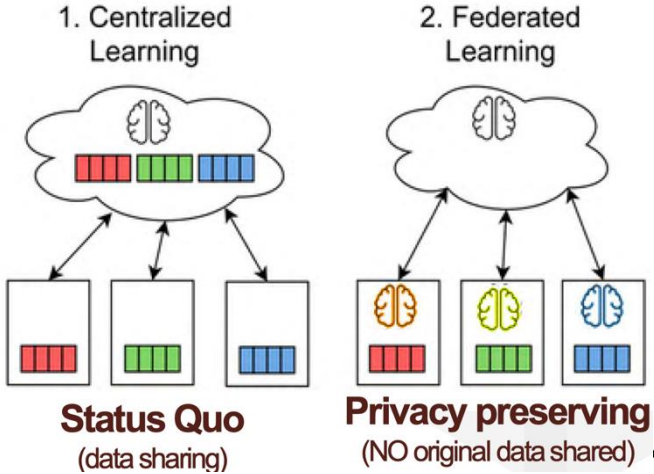


Apple Watch – Series 10
(Apple Inc, 2025)

From raw bio-signal transfer to more complex algorithms (heart rate monitoring, basic filtering, etc.) on single-/multi-input biosignals

- Components
  - ML model:
  - Data:

Future of Wearable Devices



Source: ESL Lab, EPFL

And many more…

1. **CMOS biosensors and lab-on-chip.** DNA sequencing on a chip (Ion Torrent measures pH on CMOS; Oxford Nanopore reads current through a nanopore), point-of-care diagnostics and cell-culture arrays
2. **Ultrasound and medical imaging on a chip.** CMUT/PMUT transducer arrays bonded onto CMOS turn a hospital machine into a phone accessory. How 3D integration enables products, not just faster GPUs.
3. **Image sensors, LiDAR and event cameras.** Stacked CMOS image sensors are already 3D ICs: pixel die on top, logic below. SPAD/LiDAR sensors and event-based (neuromorphic) cameras extend this and connect directly to robotics/physical AI.
4. **Hardware security.** Post-quantum cryptography accelerators, side-channel attacks, PUFs (chip fingerprints) and secure enclaves. Tackling sensitive data with privacy-preserving integrated circuits.
5. **Space and harsh environments.** Radiation-hardened design for satellite constellations plus high-temperature electronics for drilling and engines.
6. **Energy-harvesting IoT and environmental sensing.** Batteryless sensors running on µW from light, vibration or RF. E.g. precision agriculture, structural health monitoring and wildlife tracking.

Conclusion: How to build a career?

1. Read… deeply.
2. Follow ongoing research. It will be difficult, but gets better over time.
    - **Society:** IEEE SSCS (must!), IEEE Computer Society, ACM.
    - **Conferences:** ISSCC, VLSI Symposium, DAC, CICC, ESSERC, ASSCC, HotChips, TCAS, TBioCAS, Nature Electronics, Nature Machine Intelligence, EEE ISCA, IEEE Micro, IEDM etc.
    - **Journals:** JSSC, SSCL, IEEE Communications/Spectrum etc.
    - **AI&ML:** NeurIPS, ICML and ICLR.

All advanced research is based on fundamentals, yes, we still use Maxwell's Equations, Differential Calculus, Linear Equation solving etc. everyday.

3. **When applying to Industry:** The projects you and more importantly, how you solved a problem 'smartly' is the differentiating factor.
4. **Networking:** Meeting people is not enough, communicating your projects, research and knowing what the other person does, matters much more.

