In-Memory Computing for AI: Hope or Hype?

by Hussam Amrouch
Chair of AI Processor Design

In-memory Computing for AI: Hope or Hype?

Confession ....

This talk is neither to advertise in-memory computing

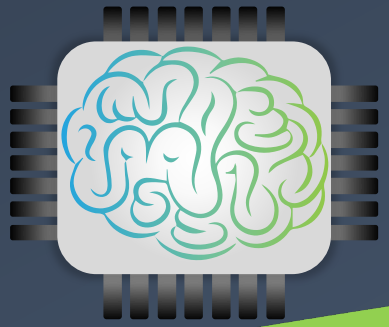
In-memory Computing for AI: Hope or Hype?

But instead… is to share

Fundamental Challenges for in-memory computing… and where the focus should be

It is a Journey…

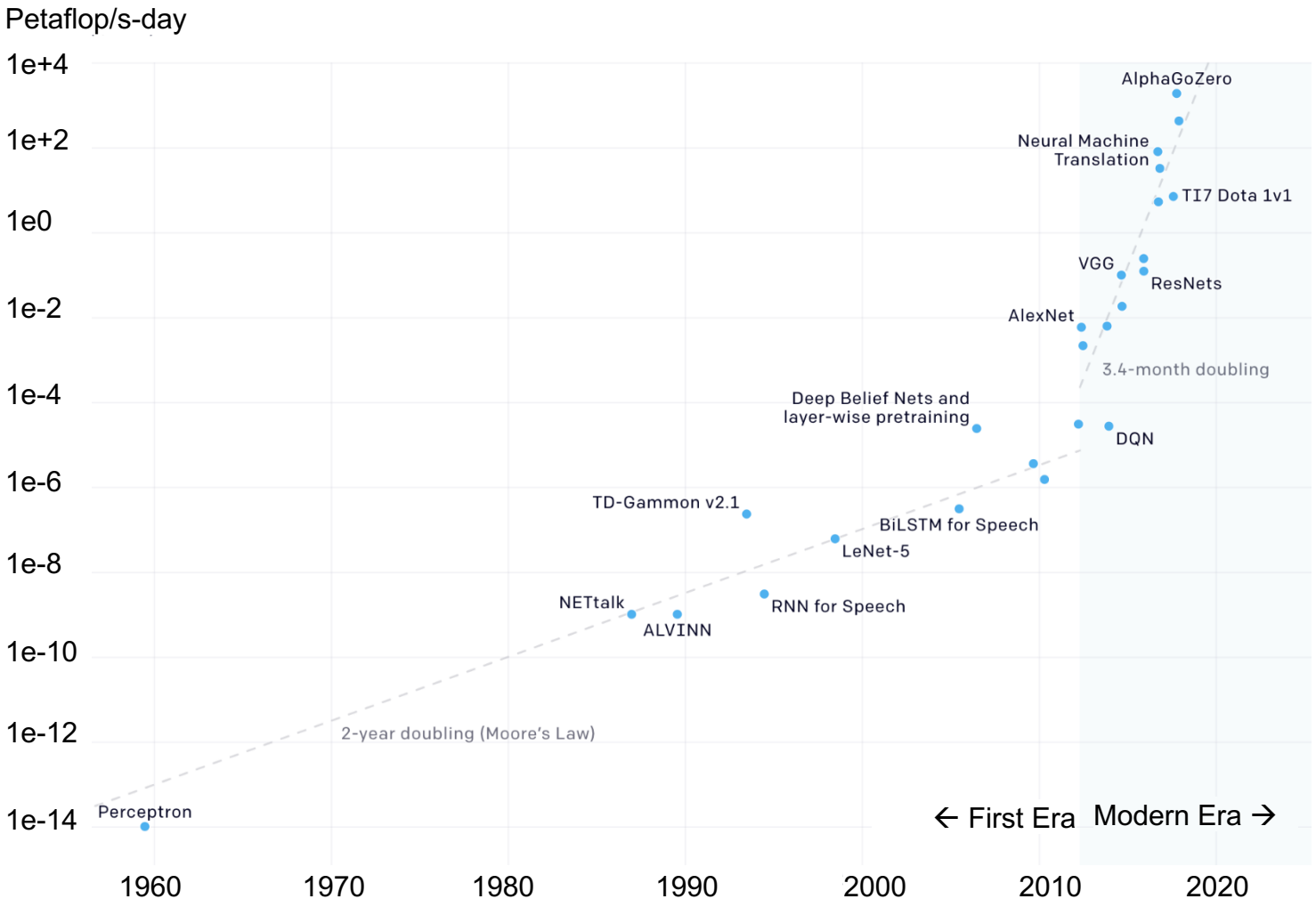
Novel Technology



Novel Chip Design

Novel Brain-inspired Computing

AI : The Next Revolution



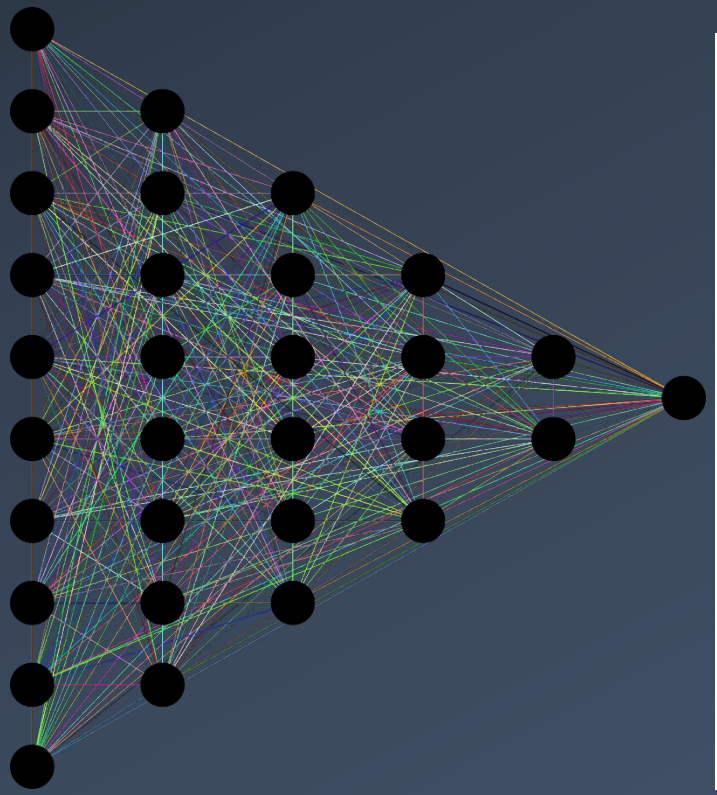
Computing demand

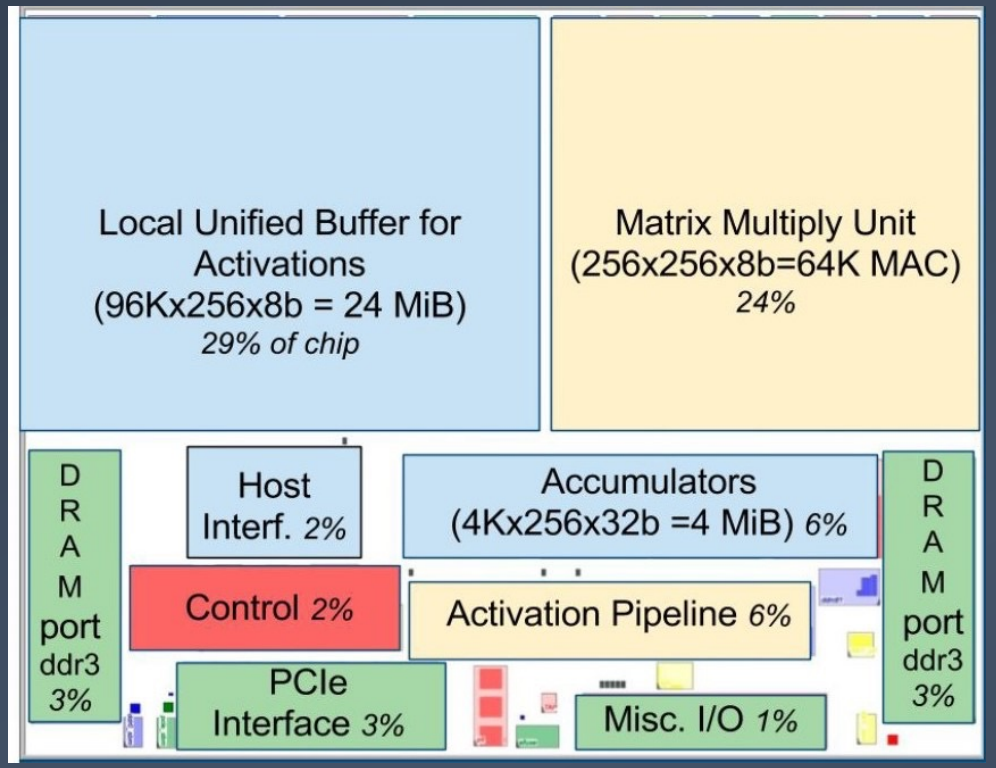
3.4 months doubling!

Source: https://openai.com

It's possible because Efficient AI Chips

Deep Learning





AI Chip: **Google TPUv1** [ISCA'17]

Complex DNN on one TPUv3: 1.8min ≈

2048 GPUs + 512 CPUs

pictures sources: by GDJ, openclipart.org and https://venturebeat.com/2020/07/29/google-claims-its-new-tpus-are-2-7-times-faster-than-the-previous-generation

It’s possible because Efficient AI Chips

BUT….

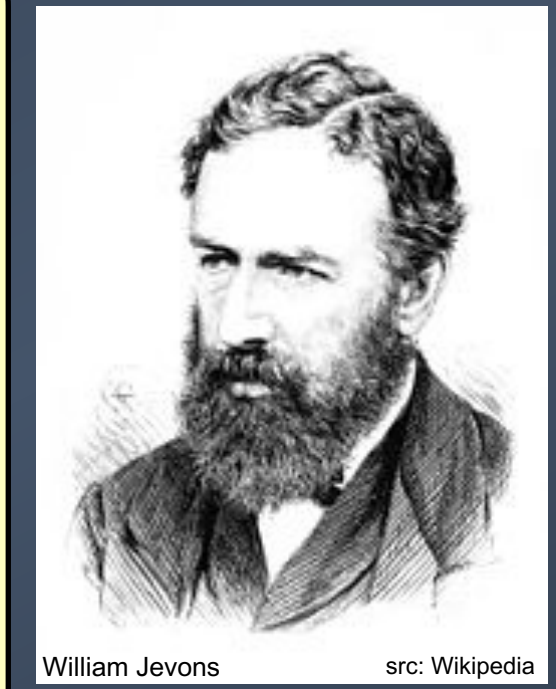
More Efficiency is really Good?

Lesson from Previous Revolution?
Let's go back to 1865…

Jevons Paradox

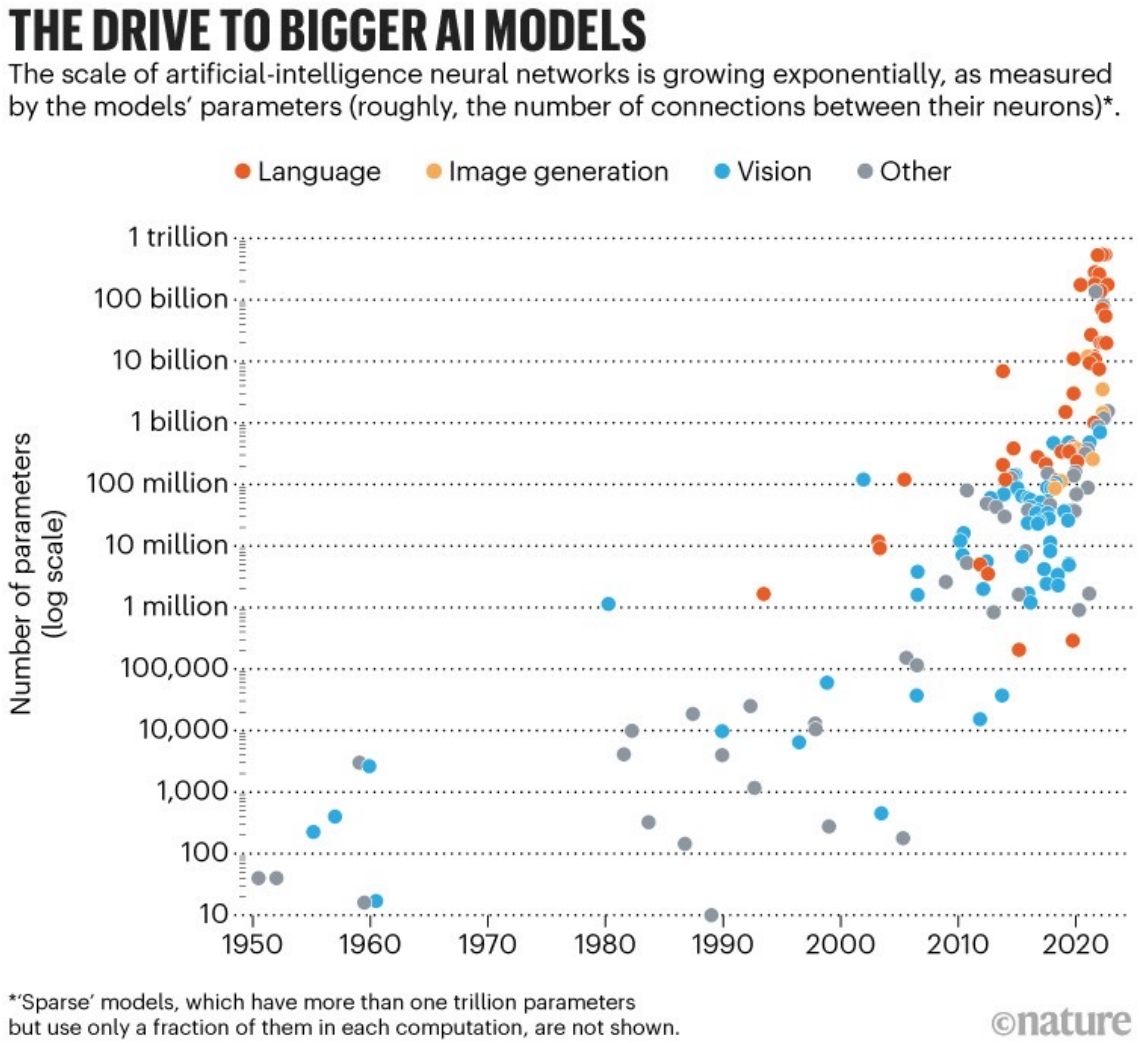
When technology increases the efficiency, the consumption rises.

→ ***Gain from efficiency will backfire*!**



William Jevons src: Wikipedia

AI Acceleration and Efficiency Paradox



More Efficient HW for AI

→ Larger and larger AI models

→ Memory Bottleneck!

Energy Crisis in AI Hardware

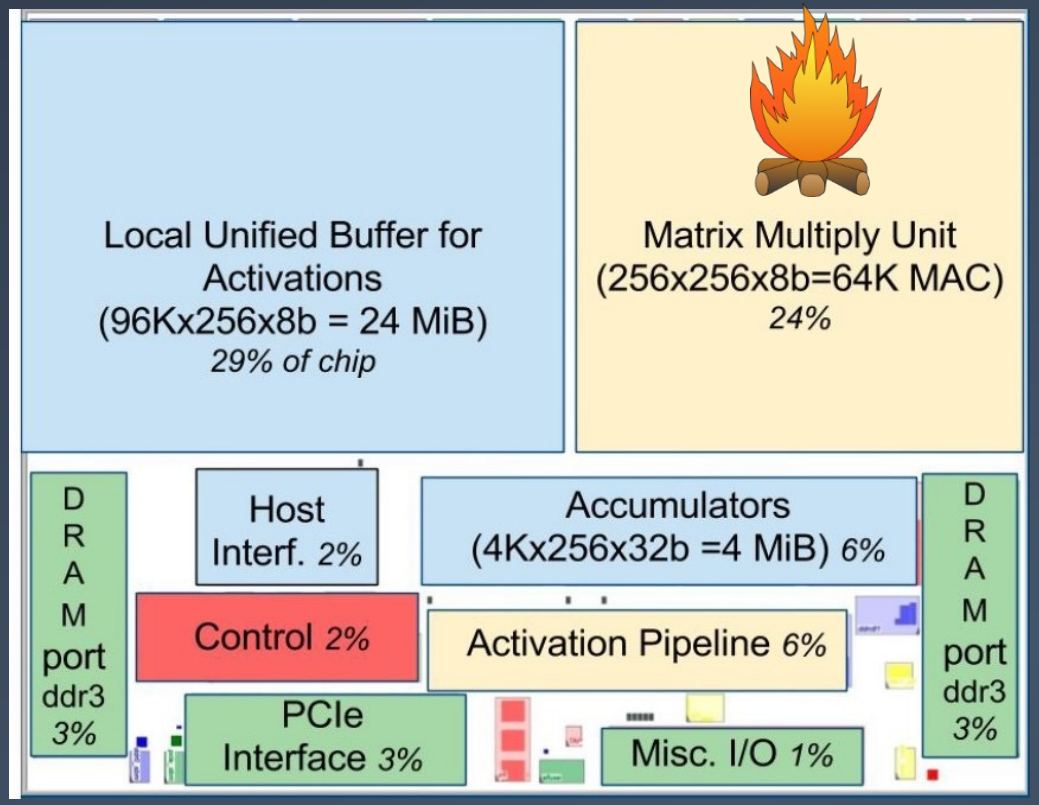
Insufficient On-Chip Memory

↓

Massive Data to Move

↓

von-Neumann Bottleneck



AI Chip: Google TPU [ISCA'17]

Massive Computation

↓

Excessive Heat

↓

Expensive Cooling

Where is the Problem?

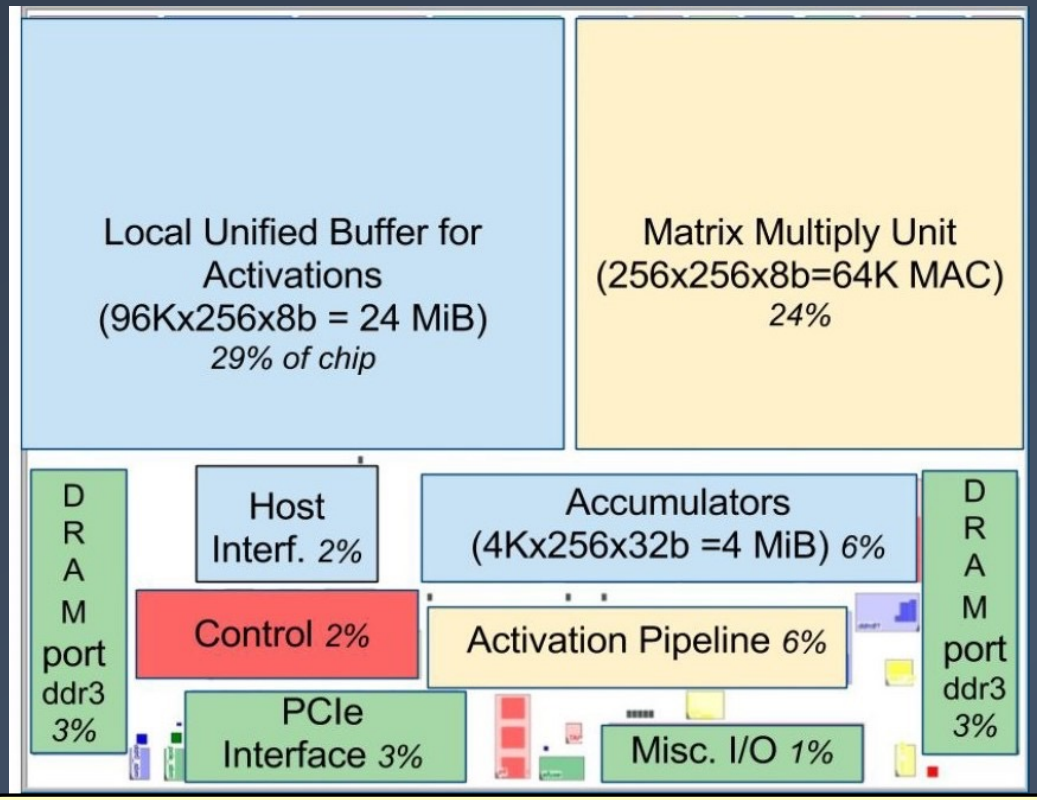
Insufficient On-Chip Memory

↓

Massive Data to Move

↓

Von-Neumann Bottleneck



Massive Energy Cost

Massive Computation

↓

Excessive Heat

↓

Expensive Cooling

The Energy Crisis that AI Bring

Microsoft is going nuclear to power its AI ambitions

/ Microsoft is looking at next-generation nuclear reactors to power its data centers and AI, according to a new job listing for someone to lead the way.

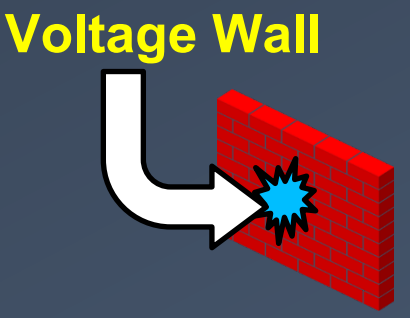
By **Justine Calma**, a science reporter covering the environment, climate, and energy with a decade of experience. She is also the host of the Hell or High Water podcast.

Sep 26, 2023, 4:32 PM GMT+2 | 35 Comments / 35 New

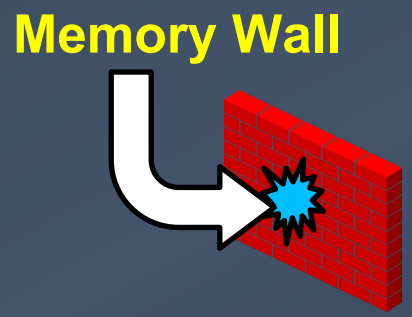
src: https://www.theverge.com/2023/9/26/23889956/microsoft-next-generation-nuclear-energy-smr-job-hiring

But….What is the Root?

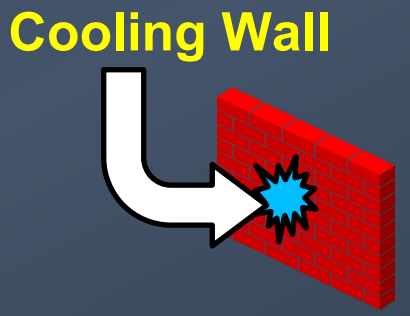
- Voltage: Reaching its Fundamental Limit



- Memory: Massive Data in DNNs

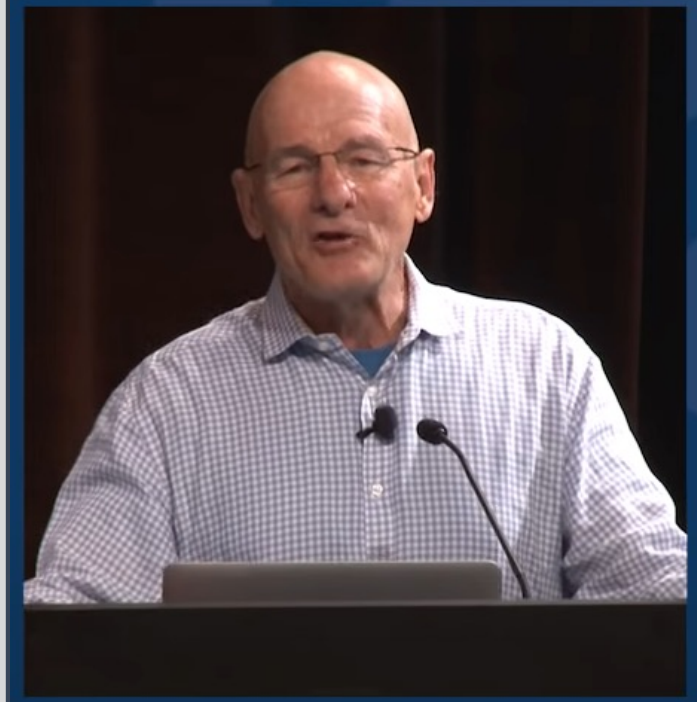


- Cooling: Inherently Inefficient

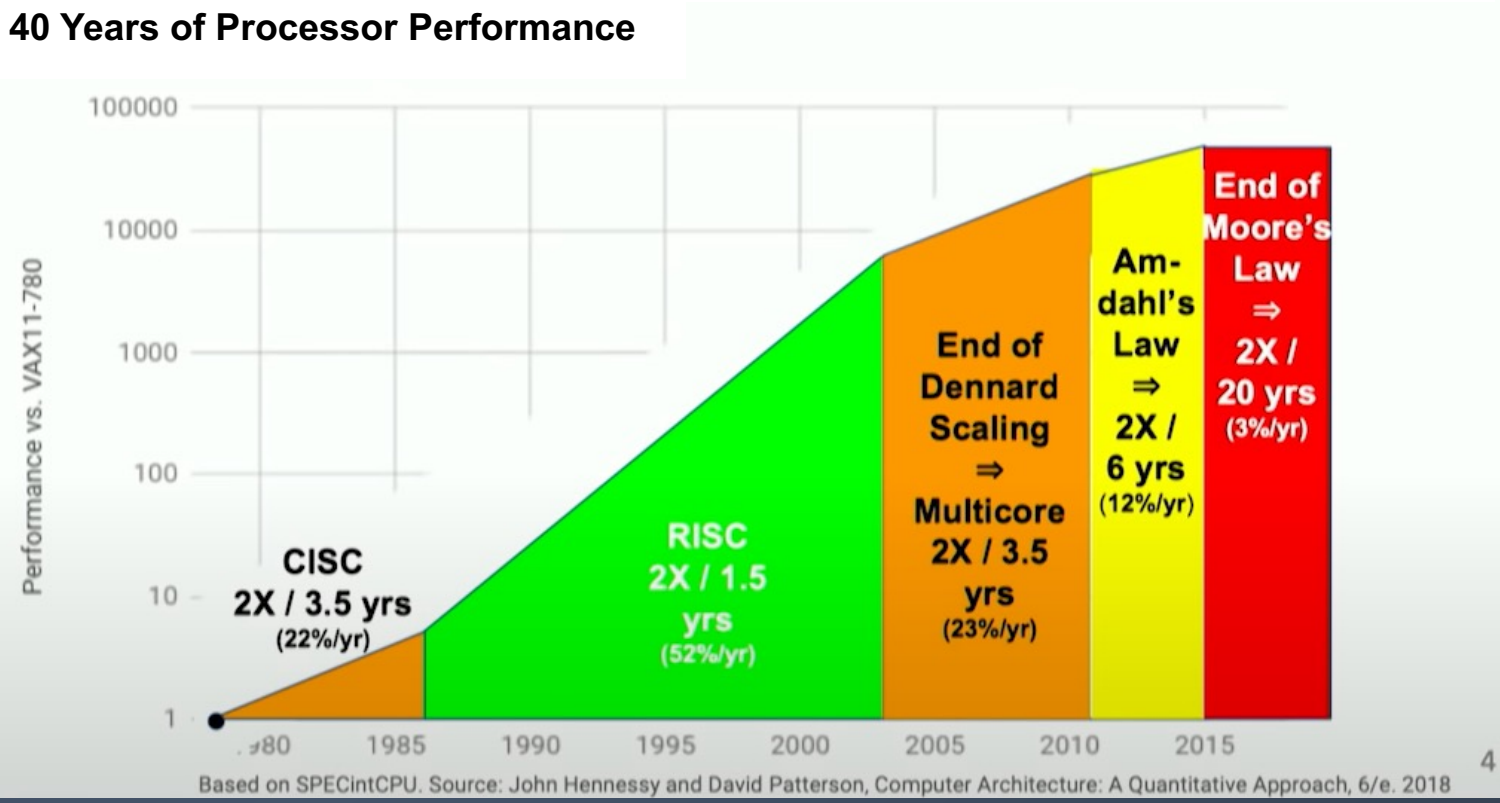


Performance: End of the Line…

David Patterson
UC Berkeley, Google

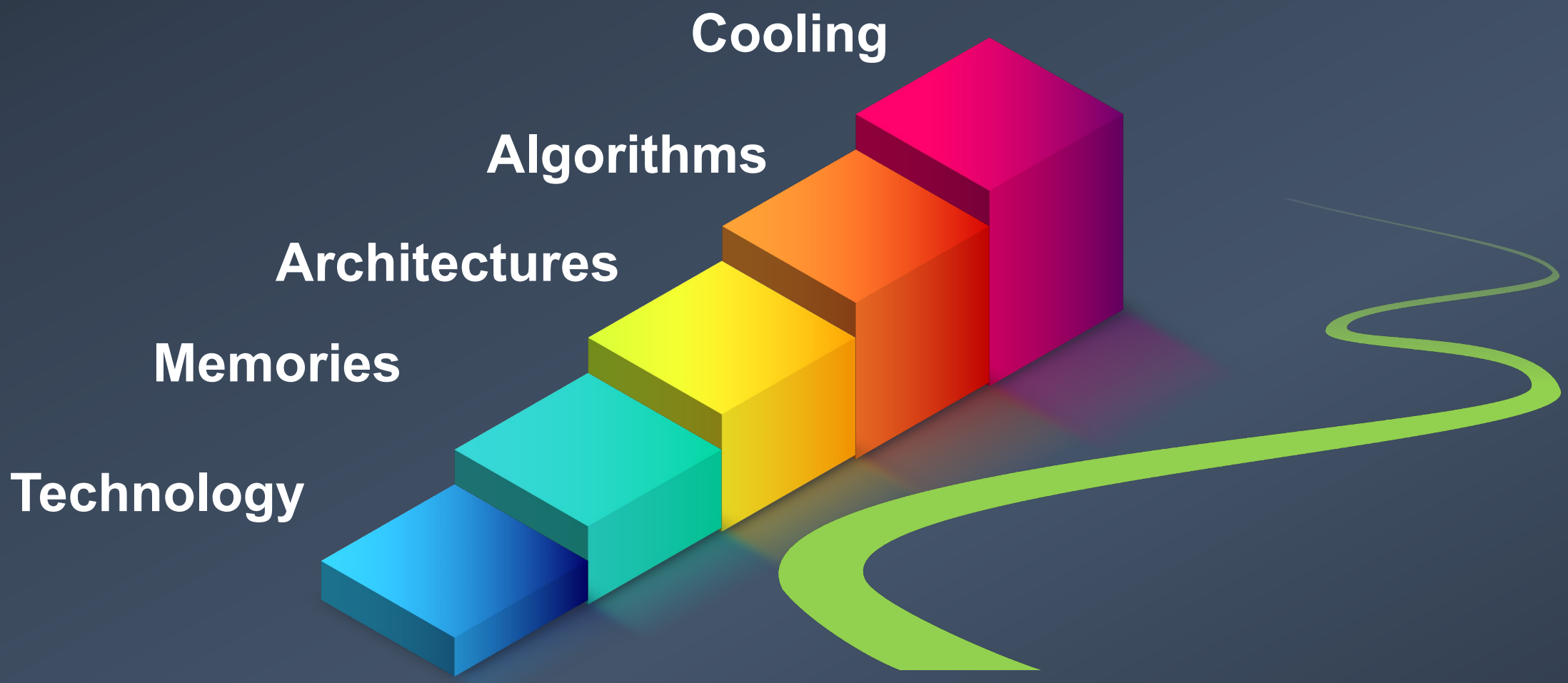


End of Growth of Performance?



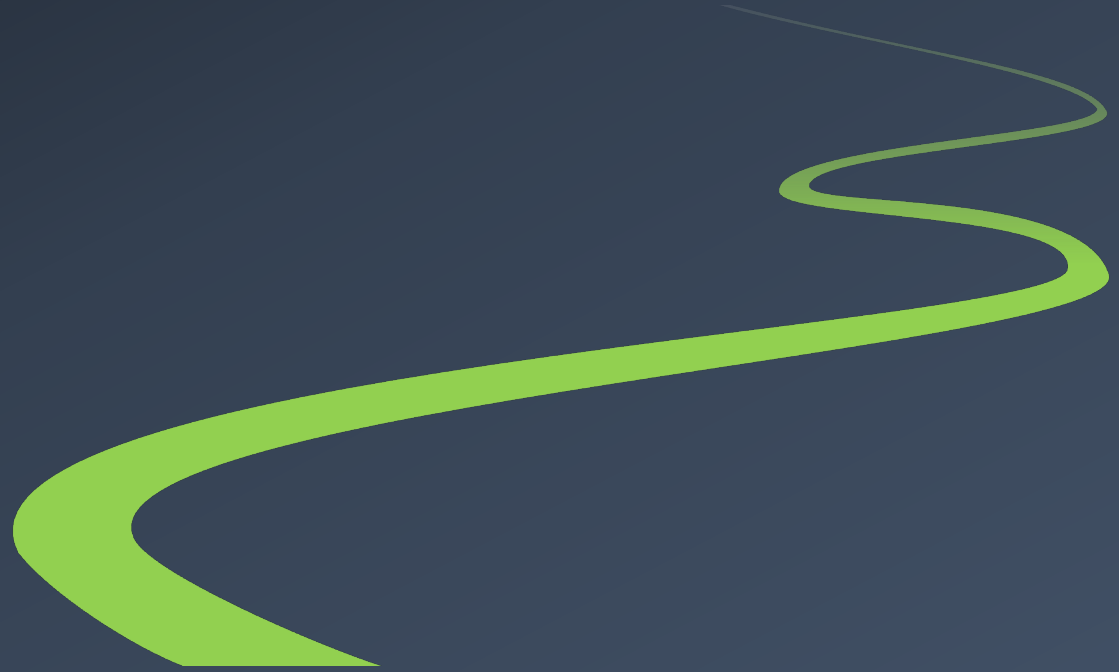
src: https://www.youtube.com/watch?v=FSwKCL8A9JQ&t=2163s

Automotive Industry needs Innovations in

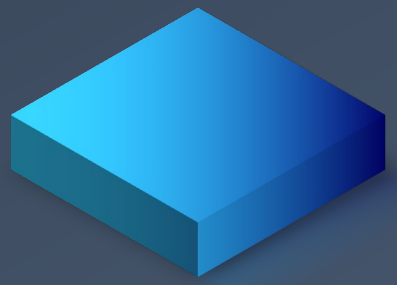


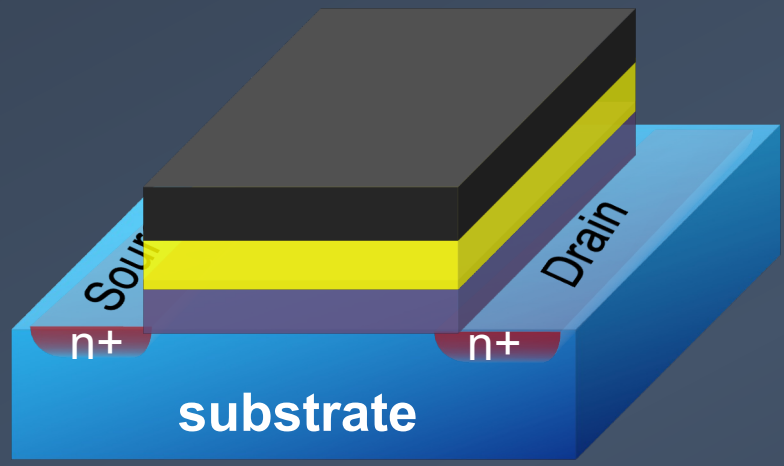
ack: Creative Venus

What's Wrong in Advanced Nodes (7nm, 5nm, 3nm...)

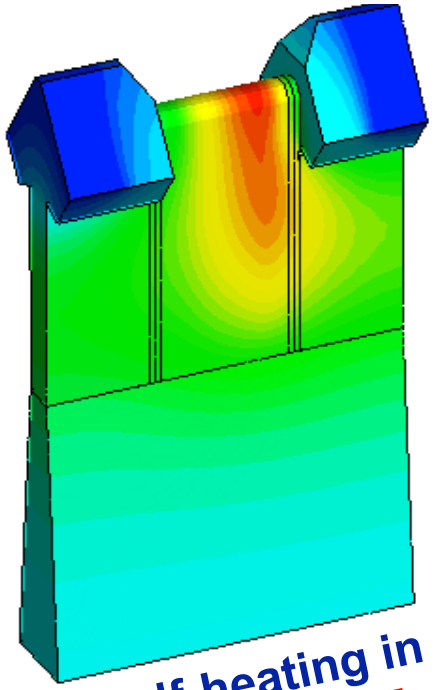


Technology





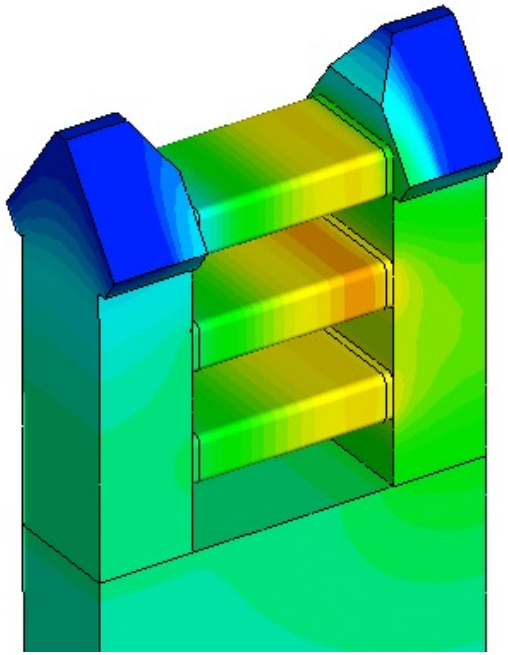
Reliability is BIG Killer!



Lattice Temperature (K)

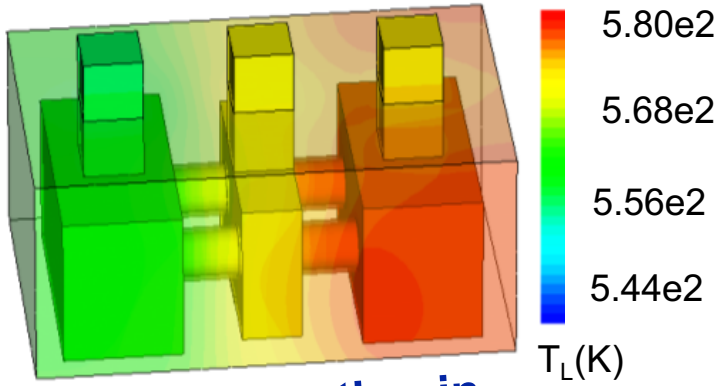
355
345
336
327
318
309
299

Self-heating in
7nm FinFET



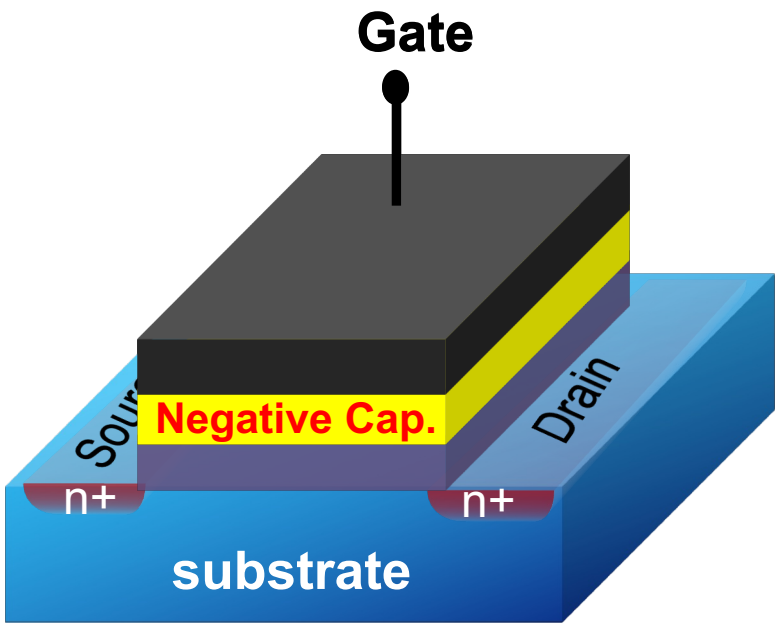
Self-heating in
7nm Nanosheet

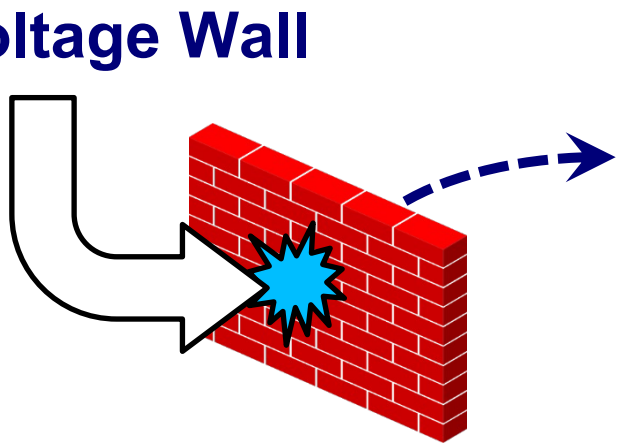
All presented results are validated against measurements from industry (confidential)



Self-heating in
14nm Nanowire

Beyond CMOS





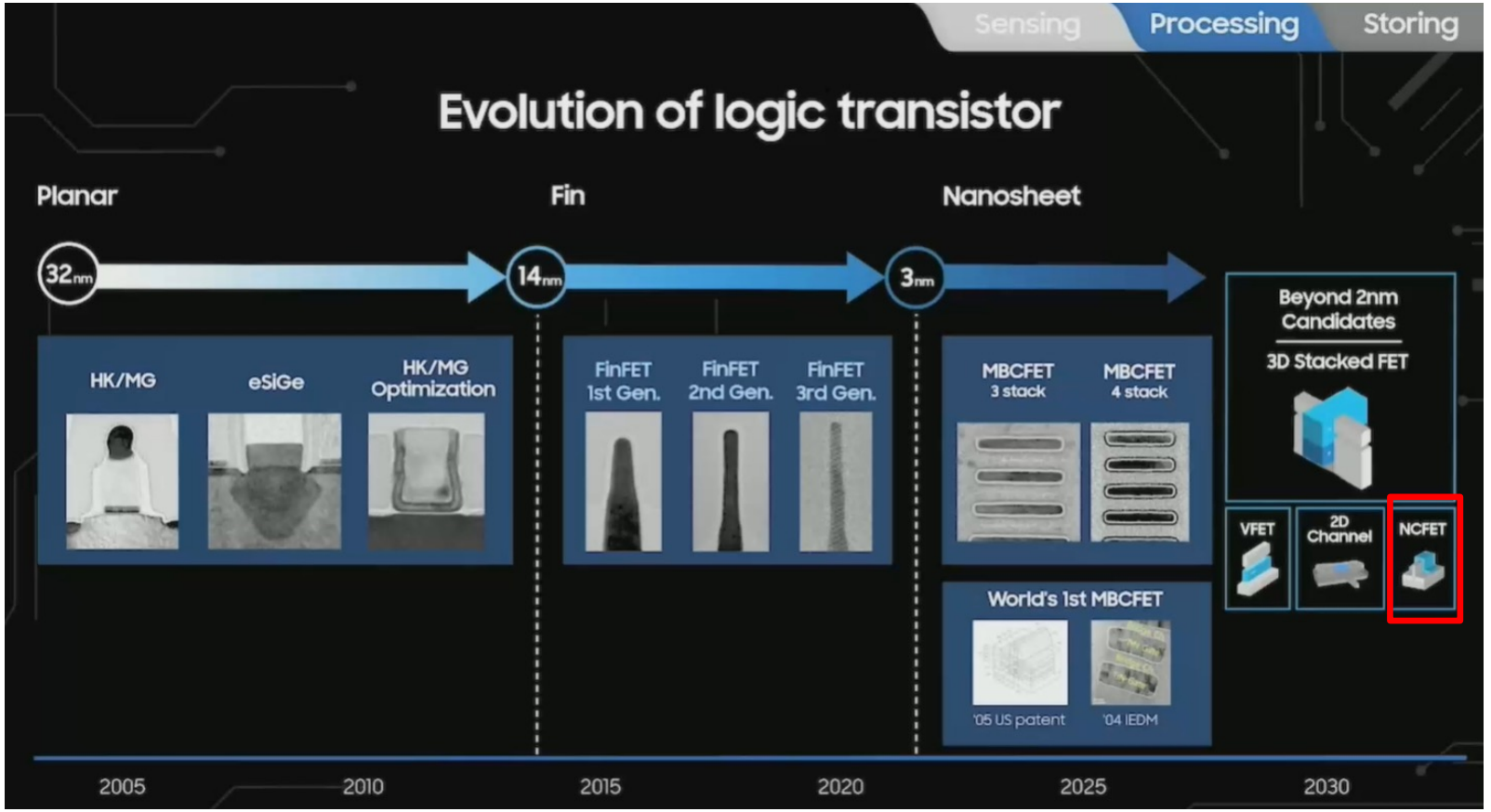
Can we Suppress the Fundamental Limit?

Yes, if C_{ins} is negative!

A. Khan and S. Salahuddin "Negative capacitance in a ferroelectric capacitor", **Nature materials**, 2015

NCFET is in the Roadmap of Samsung

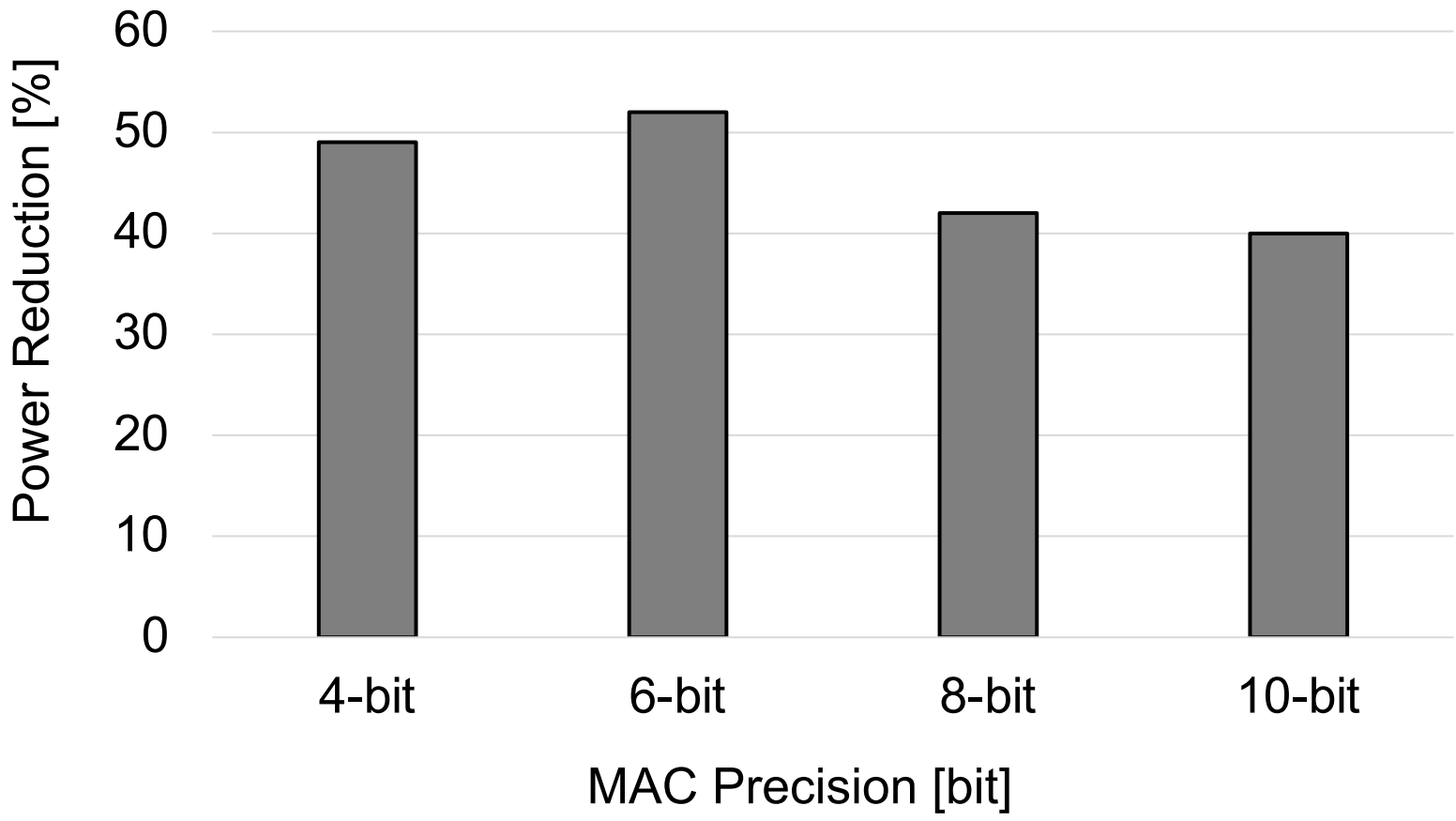
Dr. Kinam Kim, Vice Chairman, Samsung Electronics



Source: Samsung Keynote at IEDM, December 2021

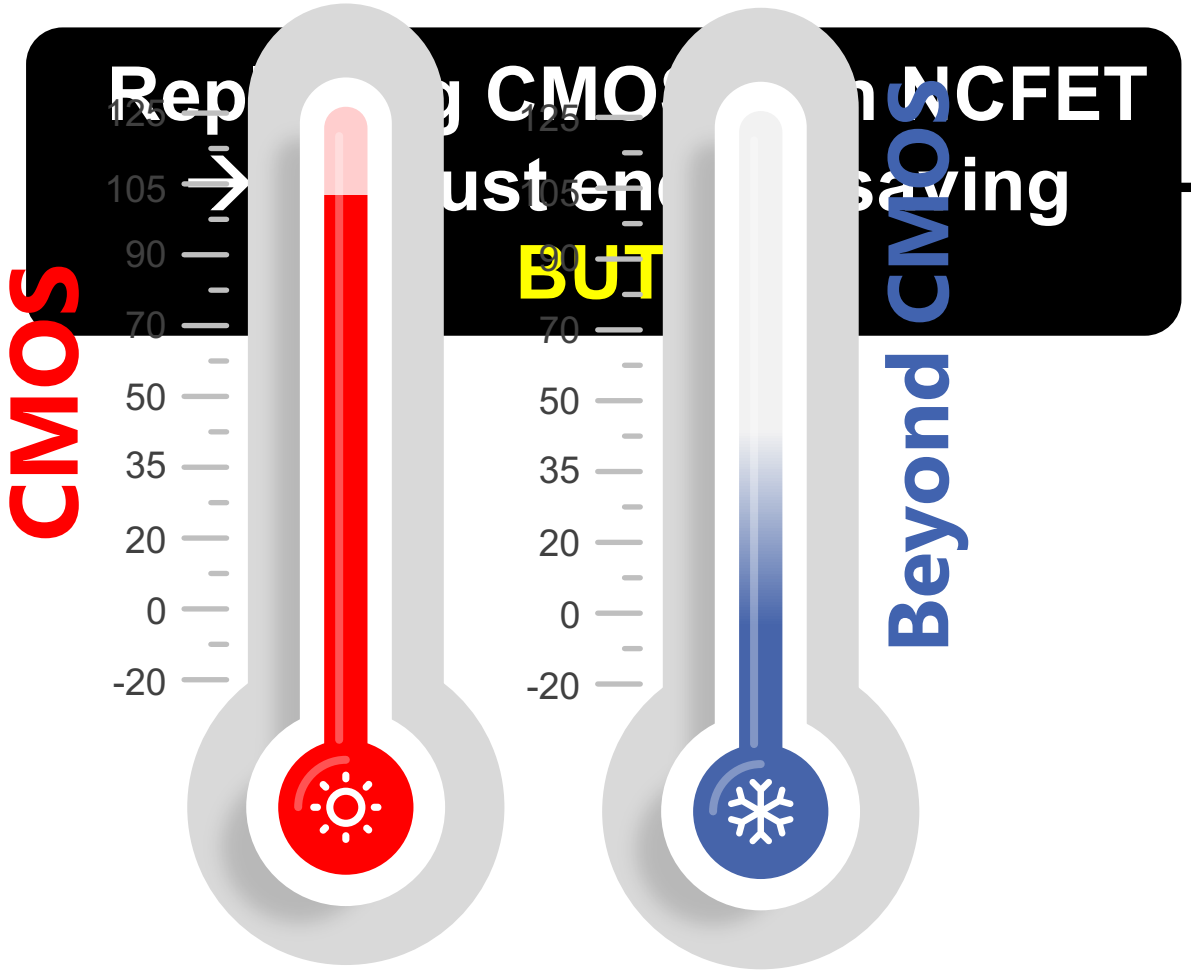
Impact of Negative Capacitance on DNNs

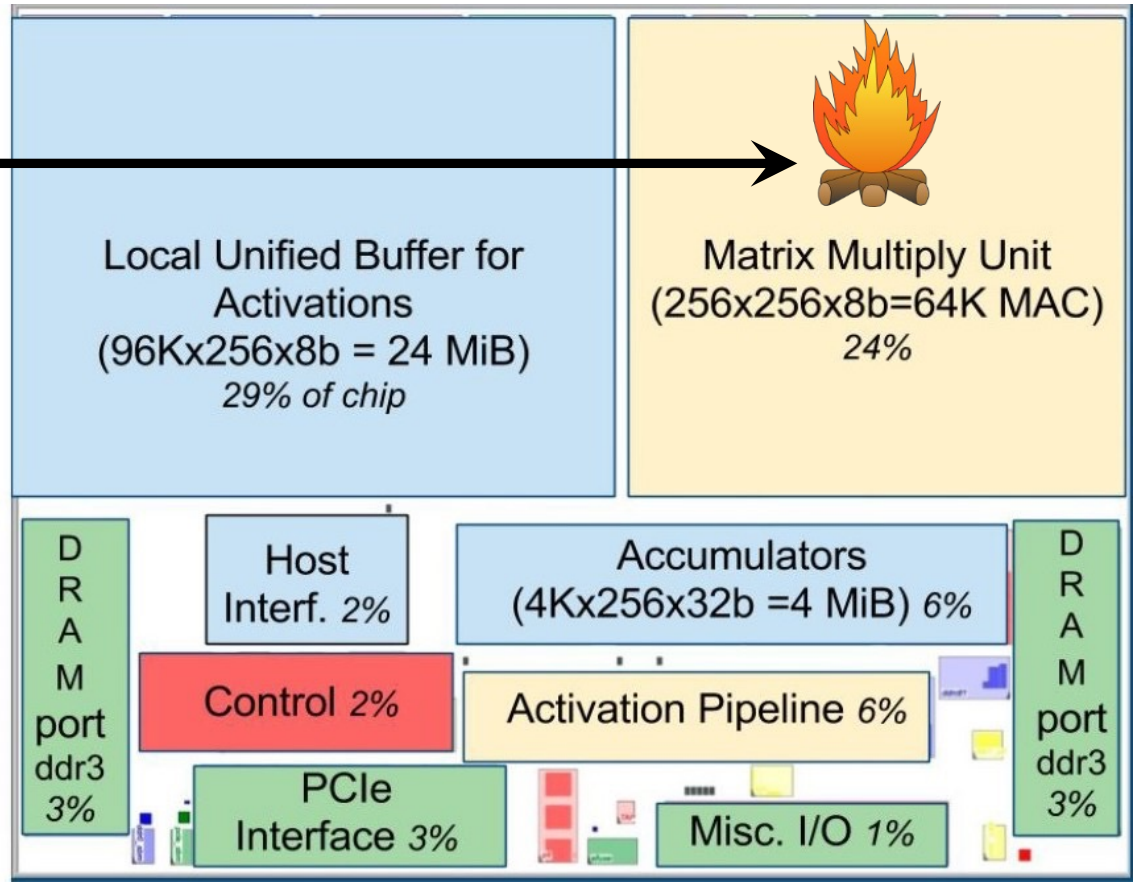
NCFET provides 40% – 50% power reduction (same performance)



S. Salamin, G. Zervakis, F. Klemme, H. Kattan, Y. Chauhan, and H. Amrouch, "Impact of NCFET Technology on Eliminating the Cooling Cost and Boosting the Efficiency of Google TPU" **IEEE Transactions on Computers (TC), 2021**

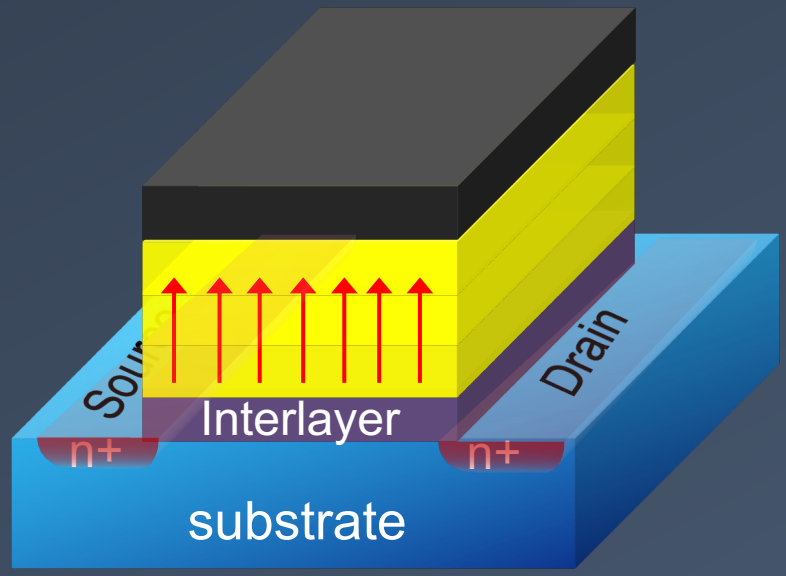
Impact of NCFET on Google TPU





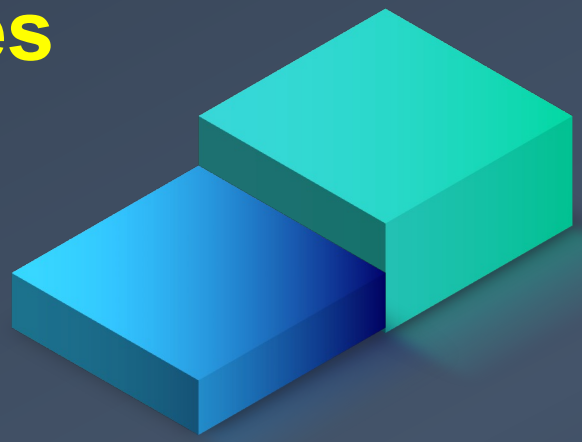
AI Chip: Google TPU

S. Salamin, G. Zervakis, F. Klemme, H. Kattan, Y. Chauhan, and H. Amrouch, "Impact of NCFET Technology on Eliminating the Cooling Cost and Boosting the Efficiency of Google TPU" **IEEE Transactions on Computers (TC), 2021**

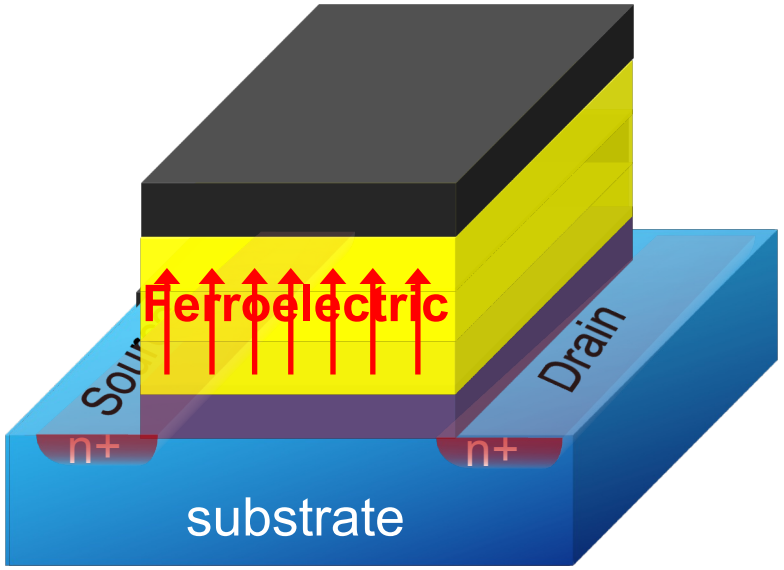


Memories

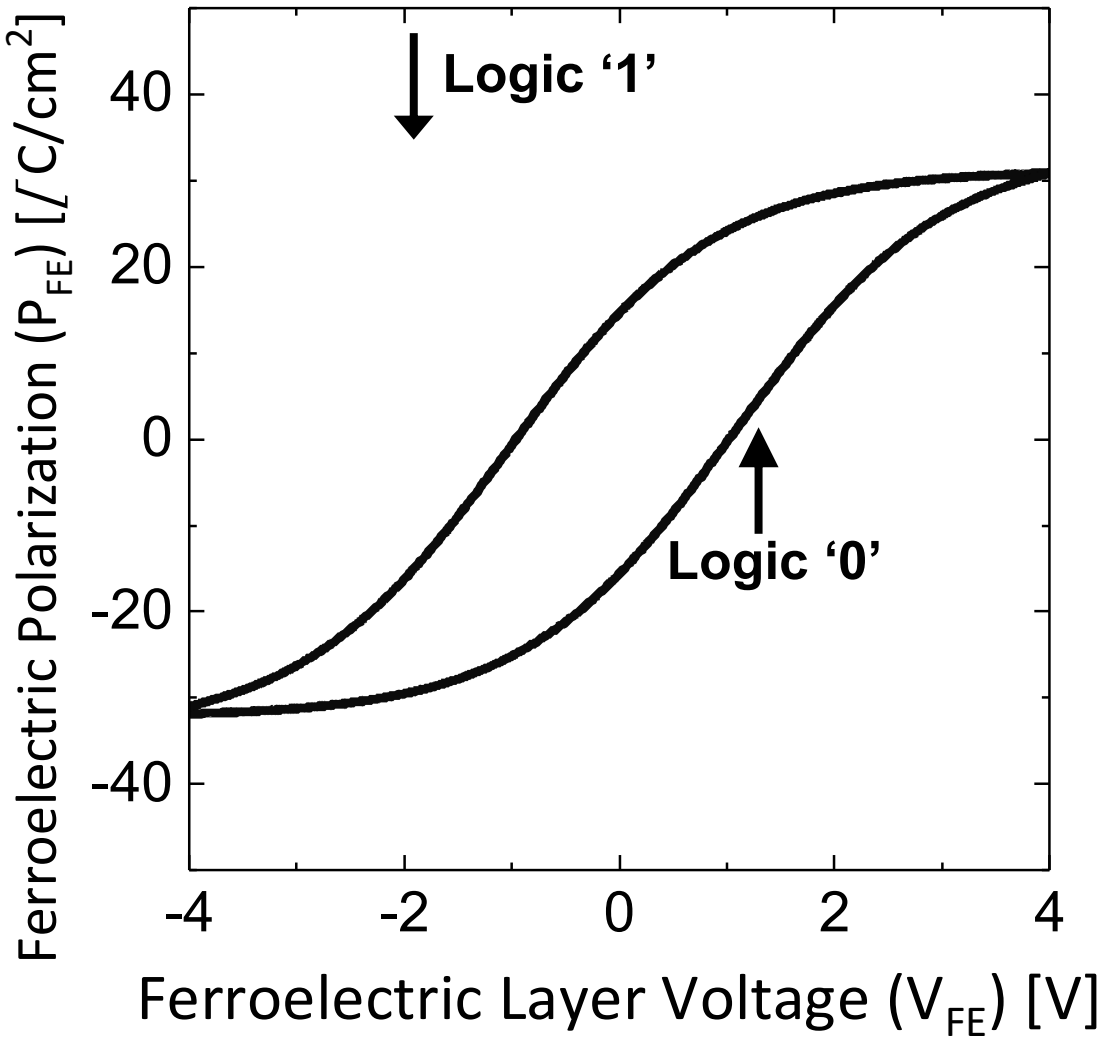
Technology



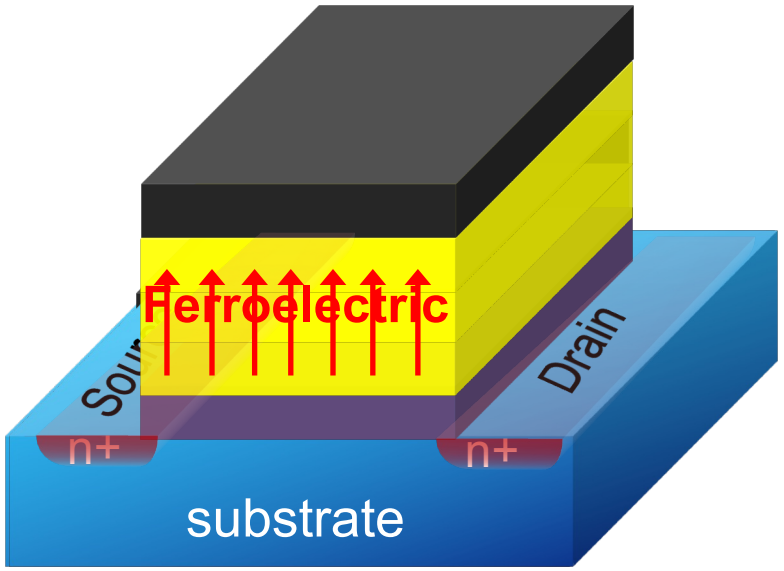
From NCFET to FeFET



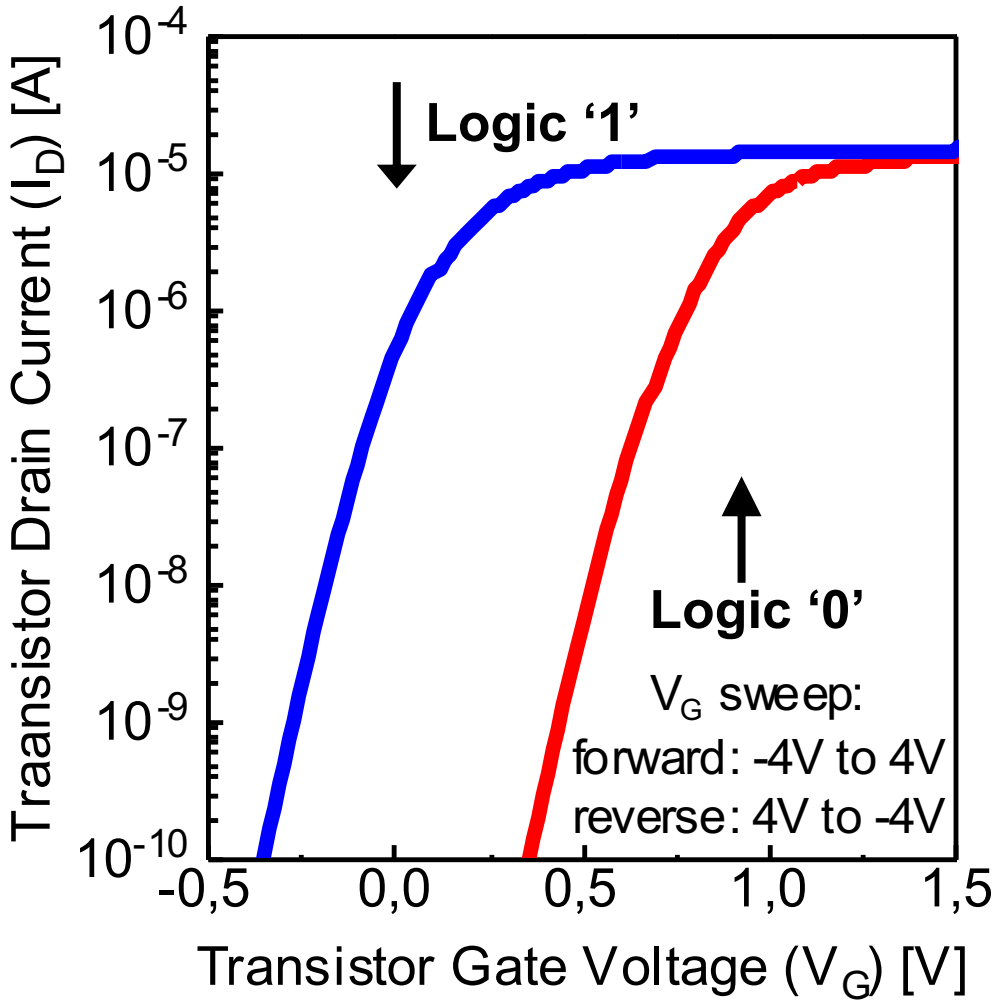
When a Bug turns into a Feature!



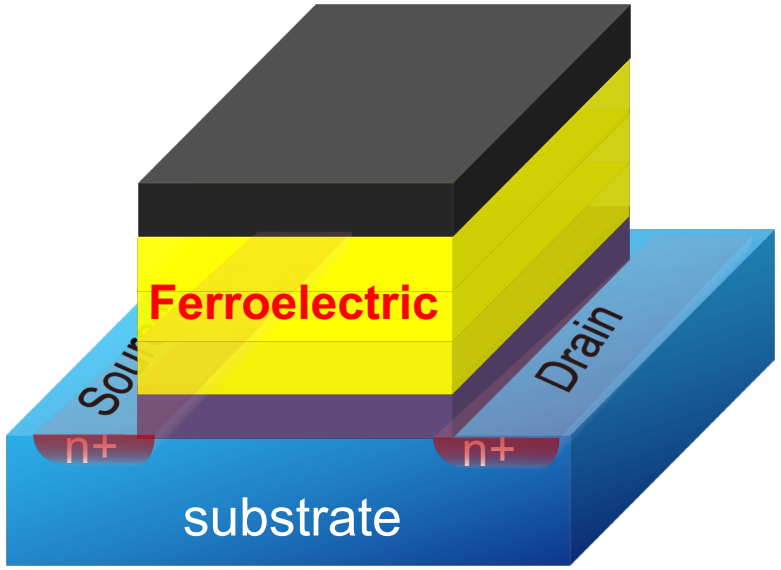
From NCFET to FeFET



When a Bug turns into a Feature!

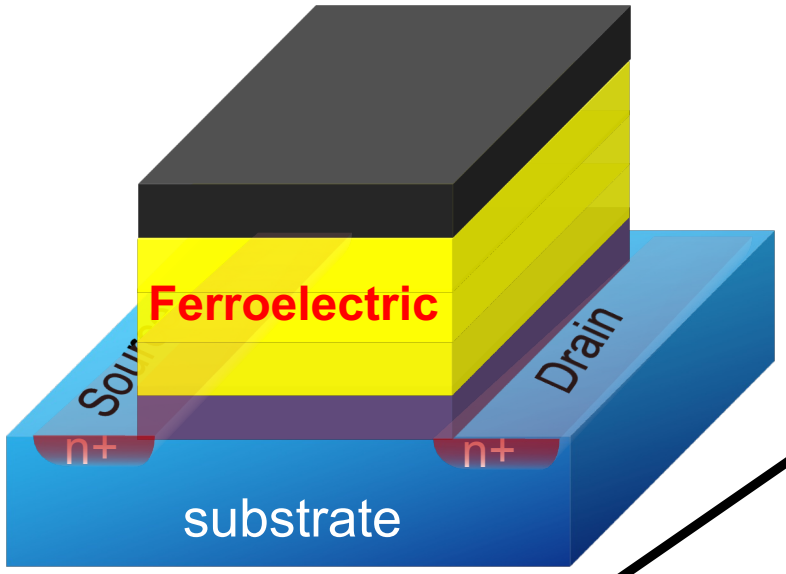


FeFET: Emerging Memory



- **Ultra Dense Memory (Single Transistor)**
- **Ultra Low-power**
- **Fully compatible with CMOS fabrication**

FeFET: Emerging Memory

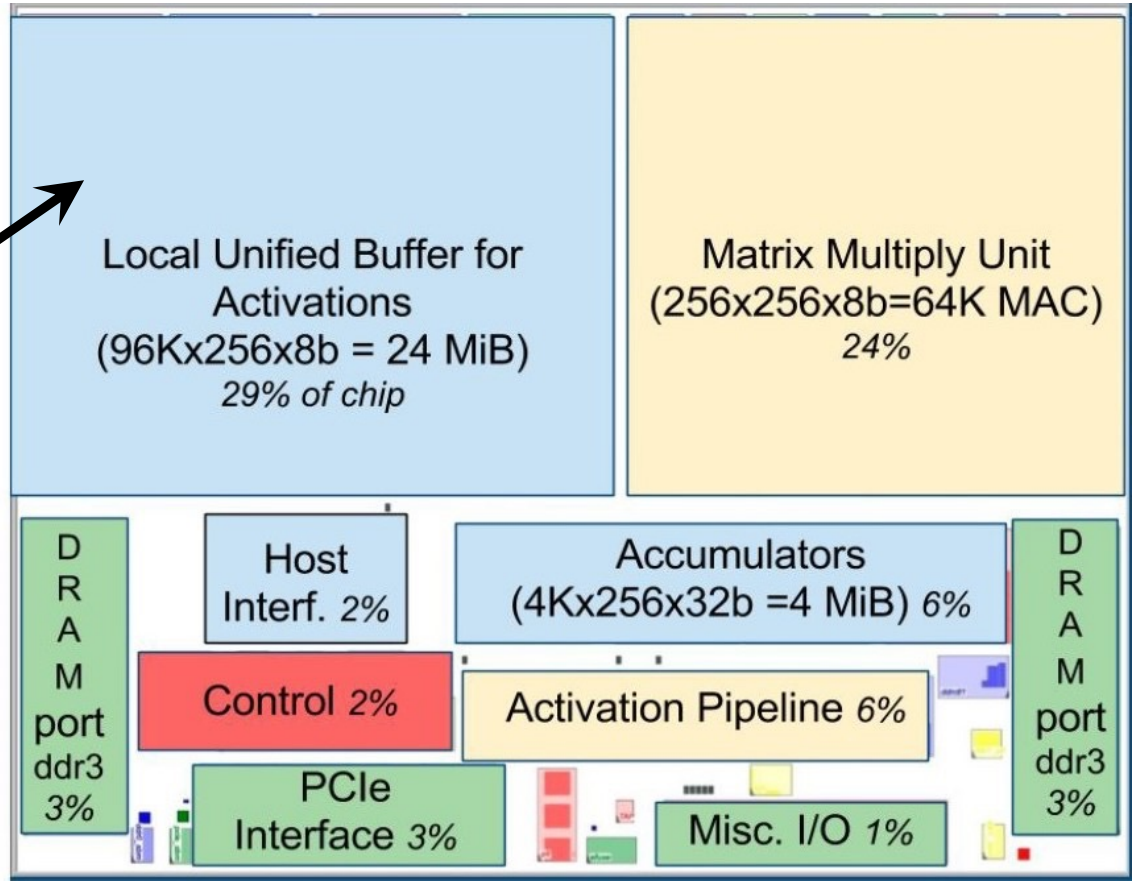


Replacing SRAMs with FeFETs

→ **Large Power Saving**

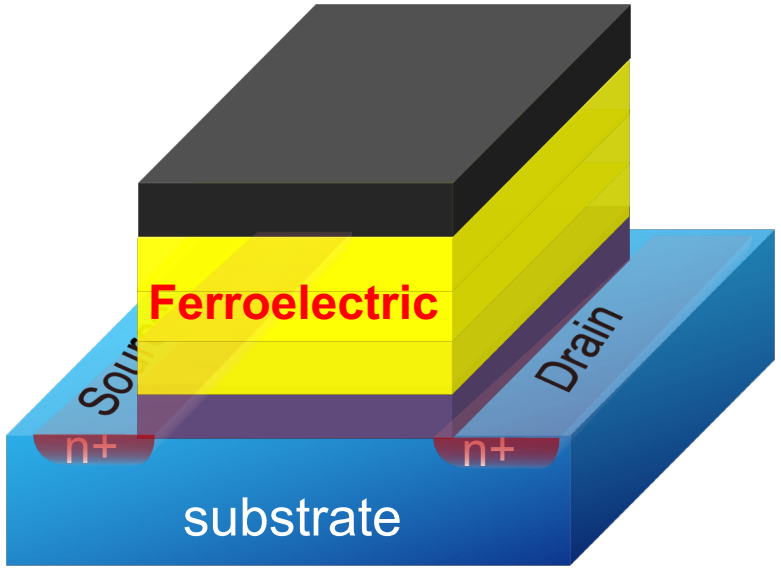
→ **Higher Storage Capacity**

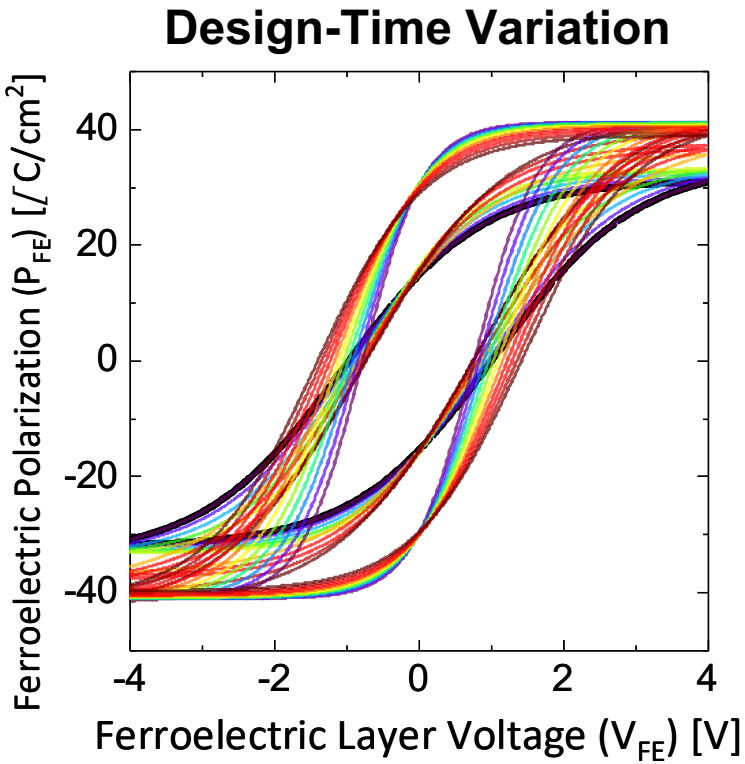
→ **Less DRAM Communications**

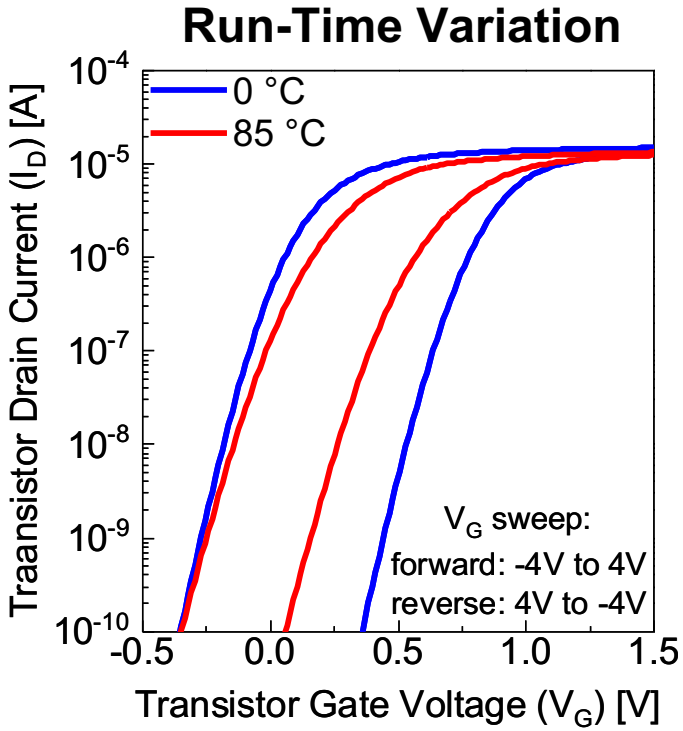


AI Chip: Google TPU [ISCA'17]

Many Fundamental Challenges



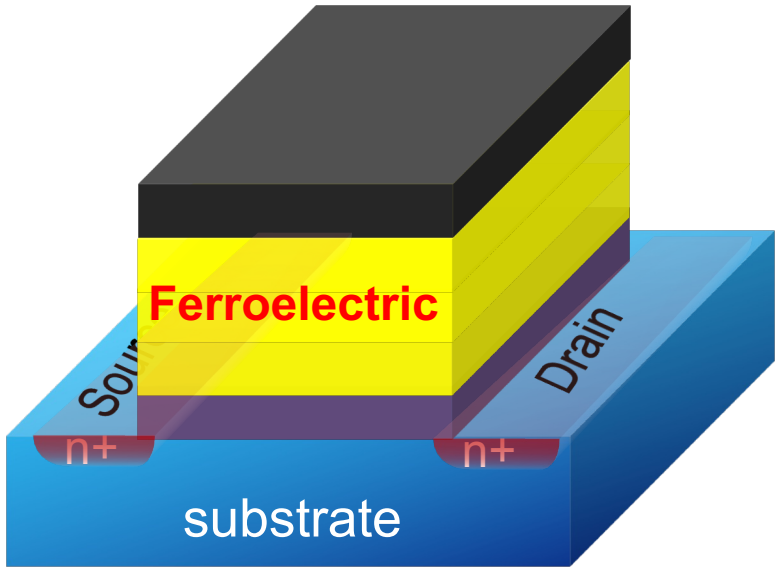


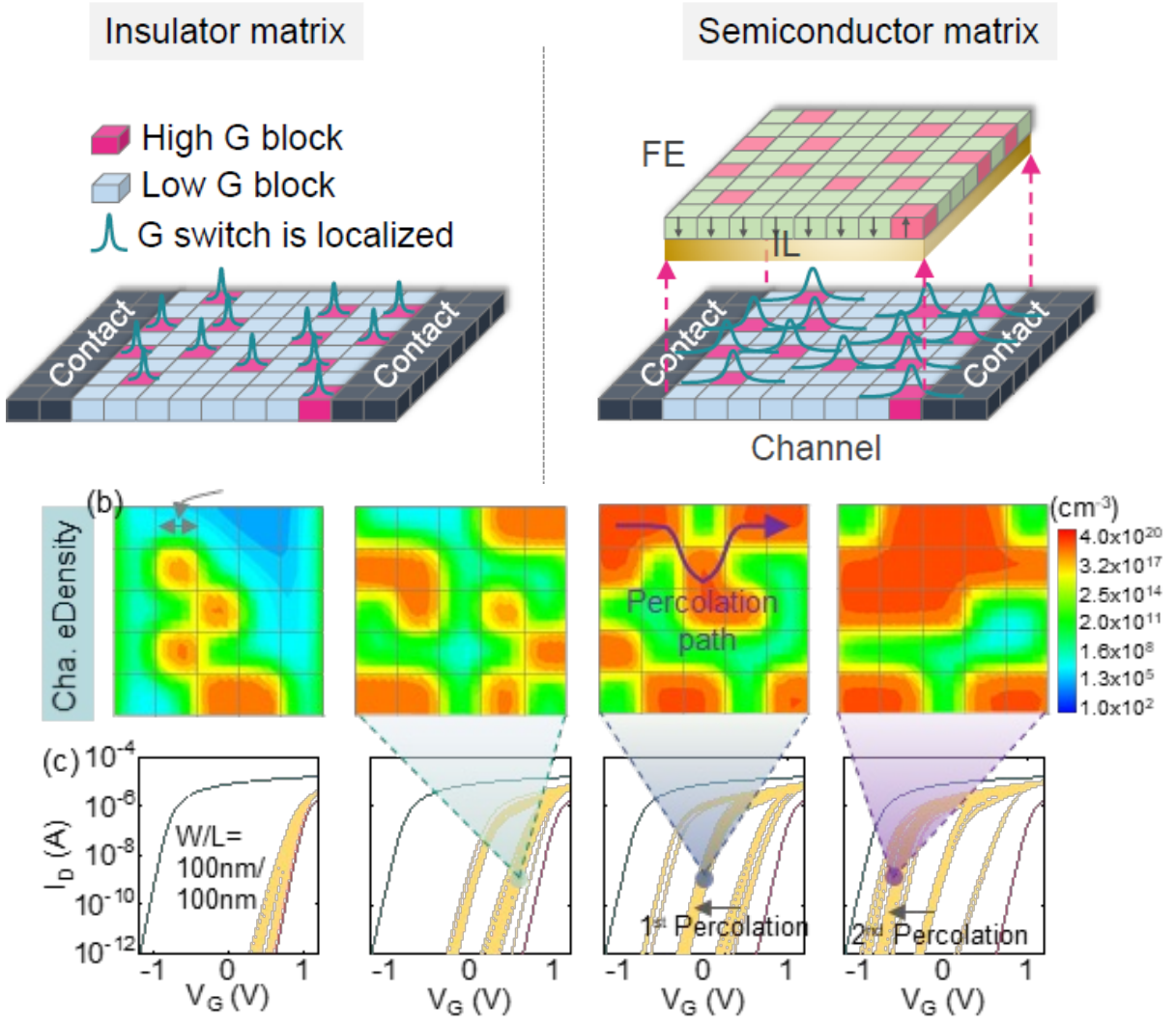


H. Amrouch, et al. "Impact of Extrinsic Variation Sources on the Device-to-Device Variation in Ferroelectric FET", **IEEE 58th International Reliability Physics Symposium (IRPS'20), 2020**

H. Amrouch, et al. "Temperature Dependence and Temperature-Aware Sensing in Ferroelectric FET", **IEEE 58th International Reliability Physics Symposium (IRPS'20), 2020**

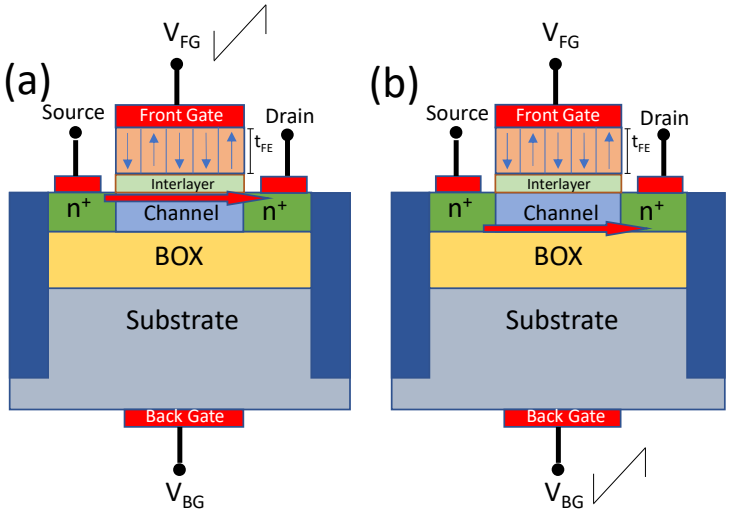
Many Fundamental Challenges



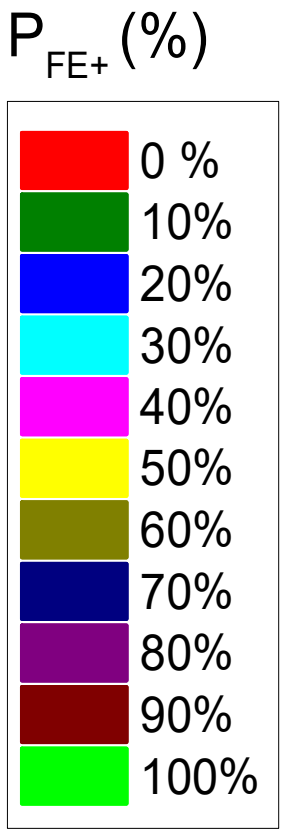


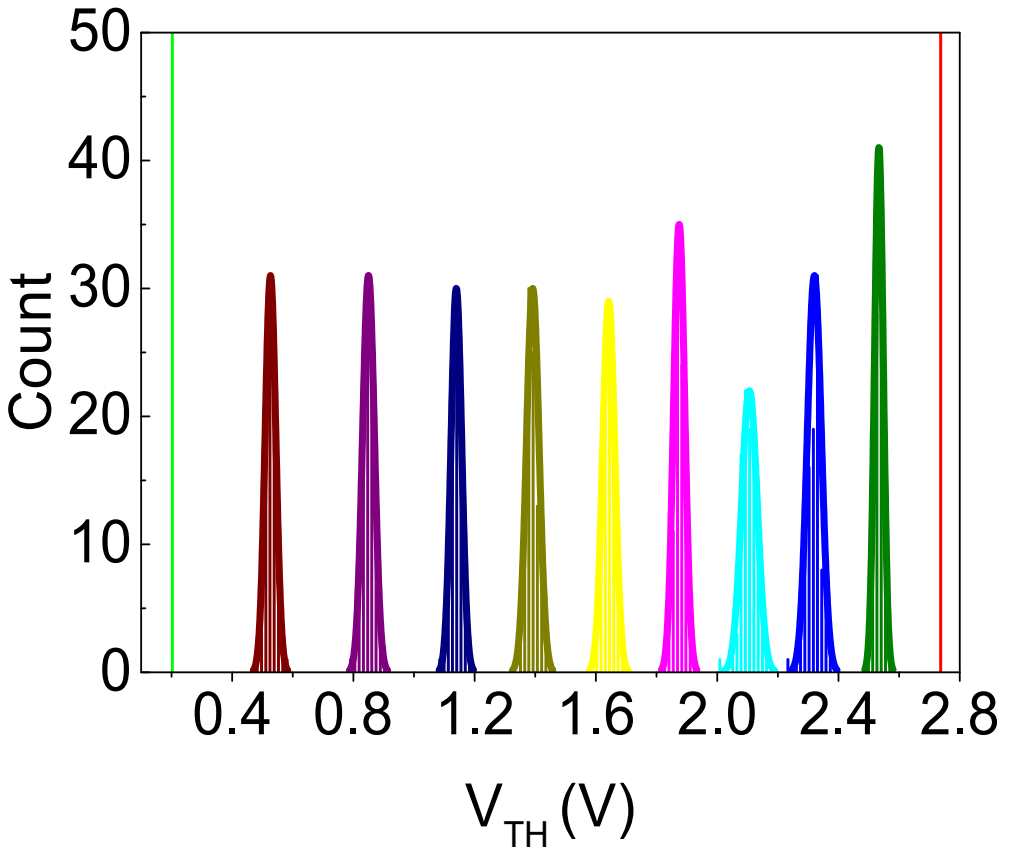
Kai. Ni / H. Amrouch "On the Channel Percolation in Ferroelectric FET Towards Proper Analog States Engineering." **In 67th IEEE International Electron Devices Meeting (IEDM'21), 2021**

Reliability Modeling is ESSENTIAL



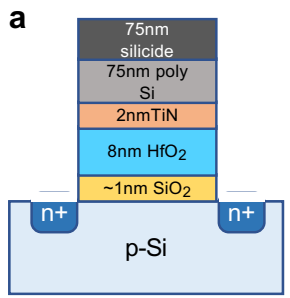
Technology/Circuit Reliability Modeling

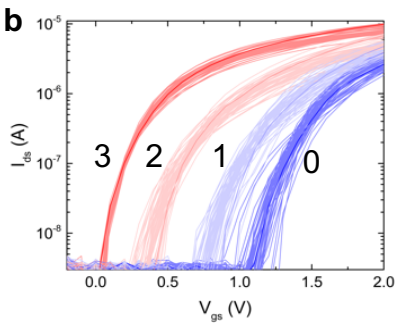


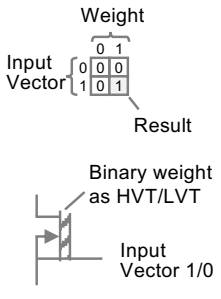


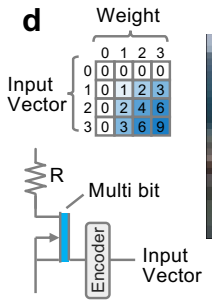
S. Chatterjee / H. Amrouch. "Comprehensive Variability Analysis in Dual-Port FeFET for Reliable Multi-Level-Cell Storage," **in IEEE Transactions on Electron Devices (T-ED), 2022**

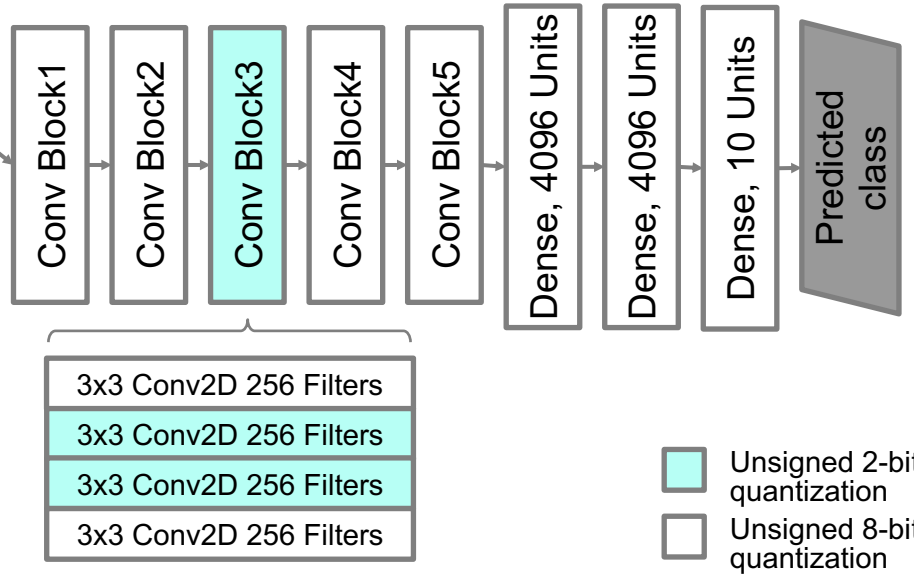
…and HW/SW Co-design is the KEY!

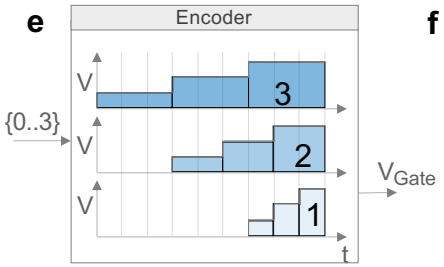


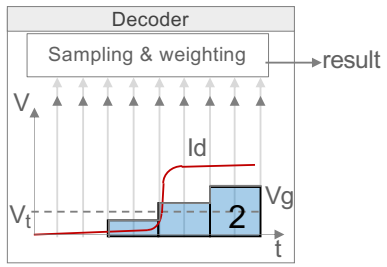


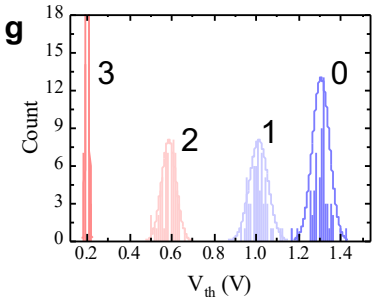


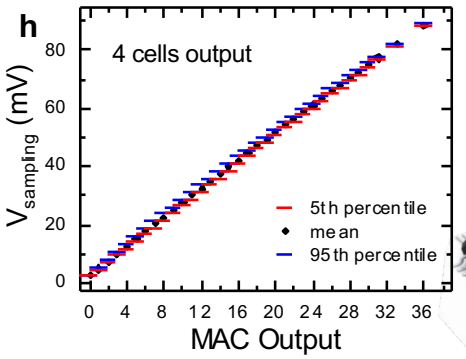


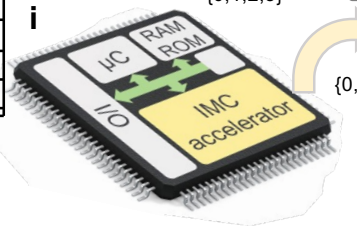


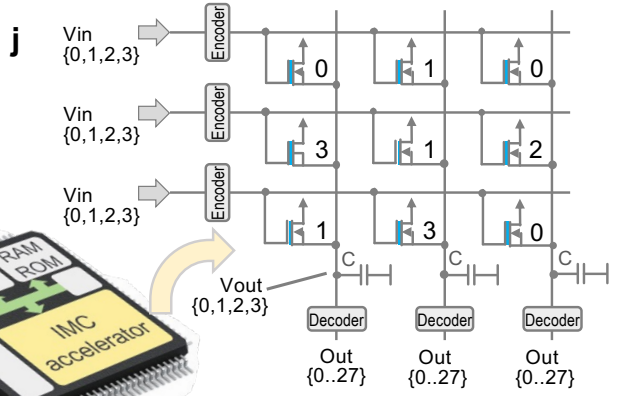




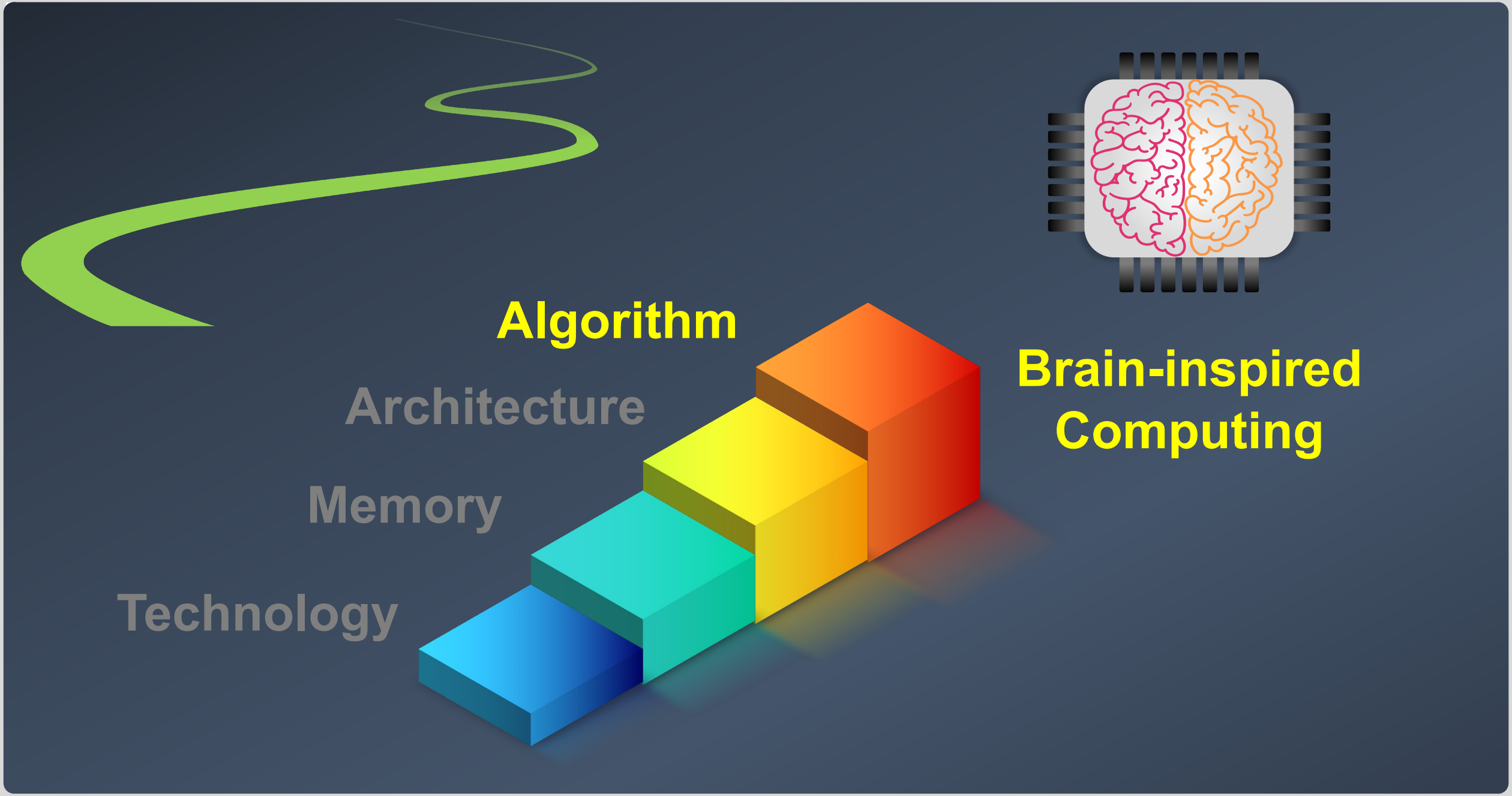




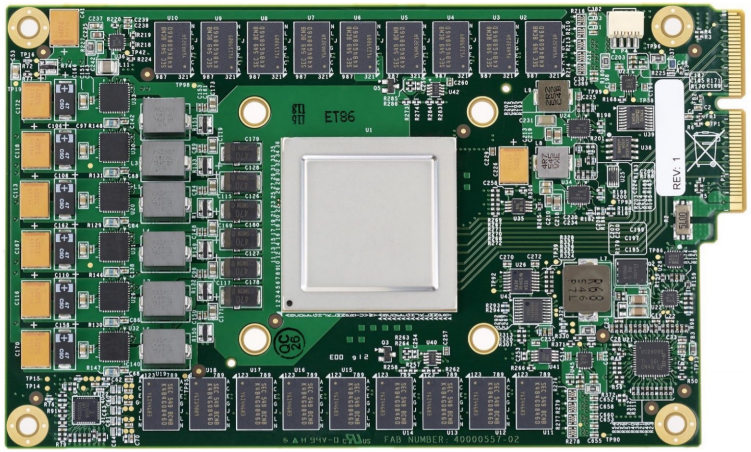




S. Taha / H. Amrouch. "First Demonstration of In-Memory Computing Crossbar using Multi-level Cell FeFET", **Nature Communications**, 2023. In collaboration with BOSCH


Algorithm
Architecture
Memory
Technology
Brain-inspired
Computing

Deep Learning is REALLY Power Hungary!

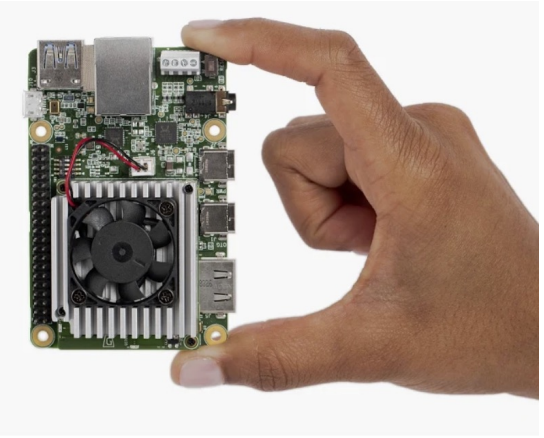


Google TPU [ISCA'17]

Datacenters

Samsung Exynos 9 [samsung.com]

Mobile Devices

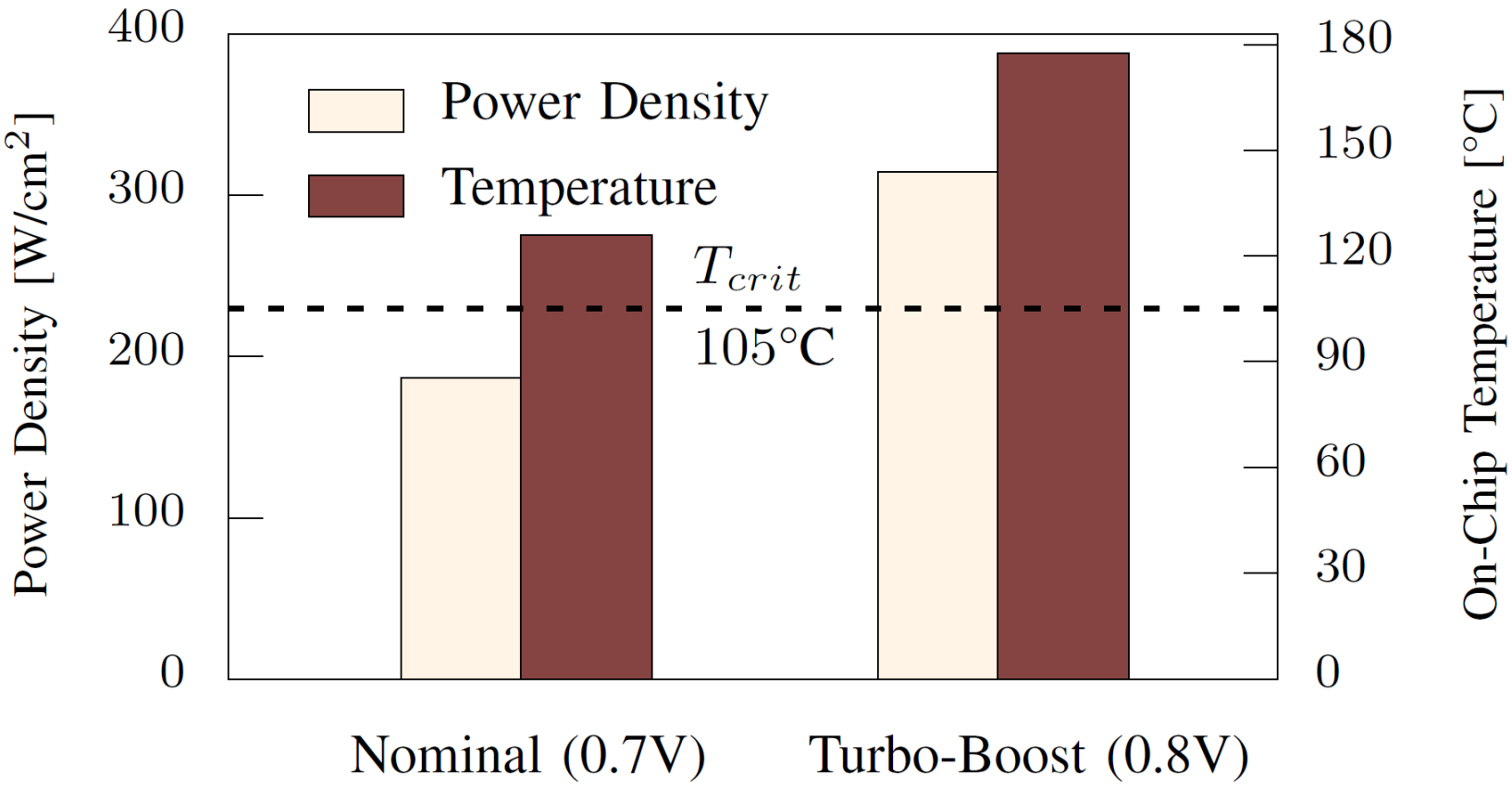


Google EDGE TPU [coral.ai]

Edge-Computing

Deep Learning is REALLY Power Hungry!

Why not alternative algorithm to Deep Learning?



Brain-Inspired Hyperdimensional Computing

Example: Language classification

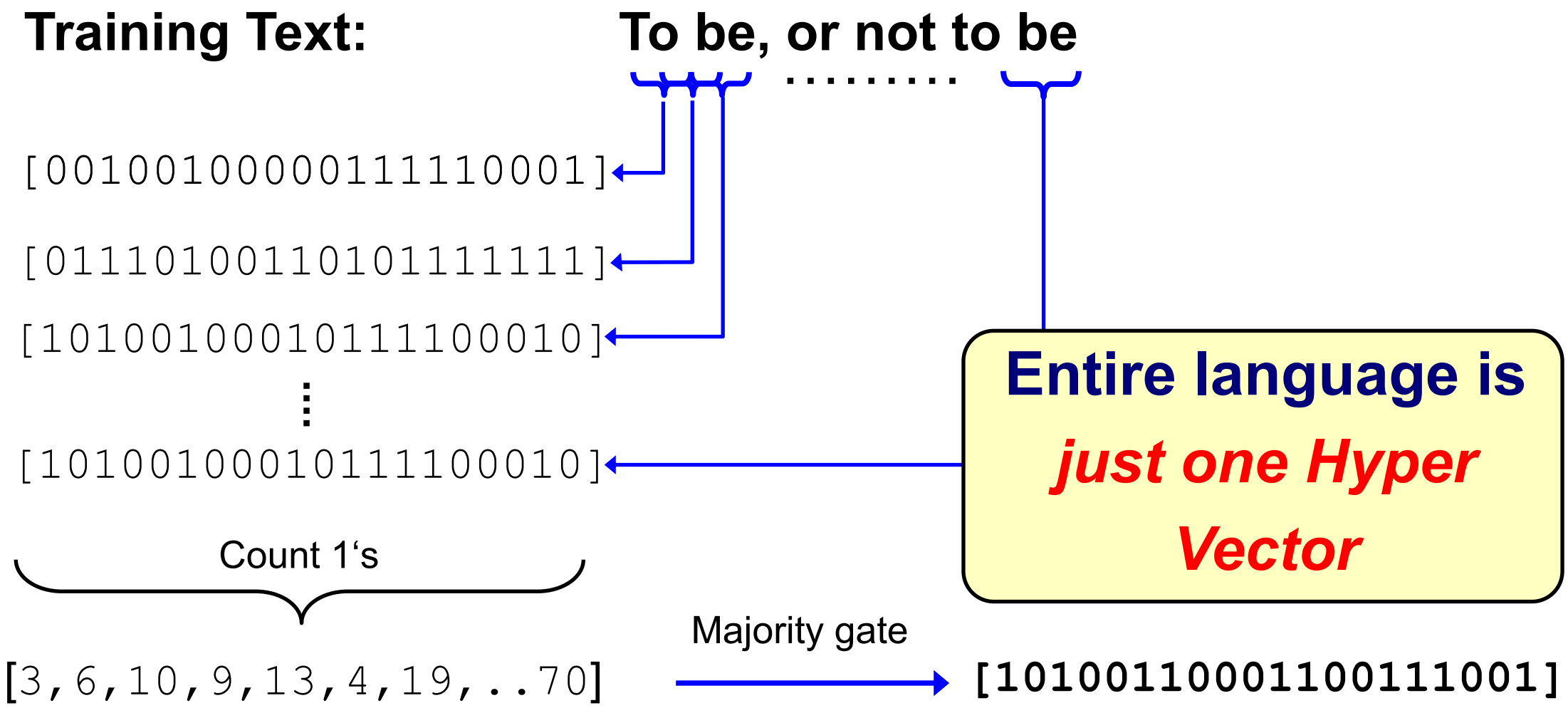
(1) Assign a random vector: VERY large (10k bits)

```
a=[10110000010000110101]
b=[10100011011010000001]
          ⋮
!=[10101111000111100101]
```

(2) Encoding with N-Grams using two simple operations: **XOR**, **Rotate**

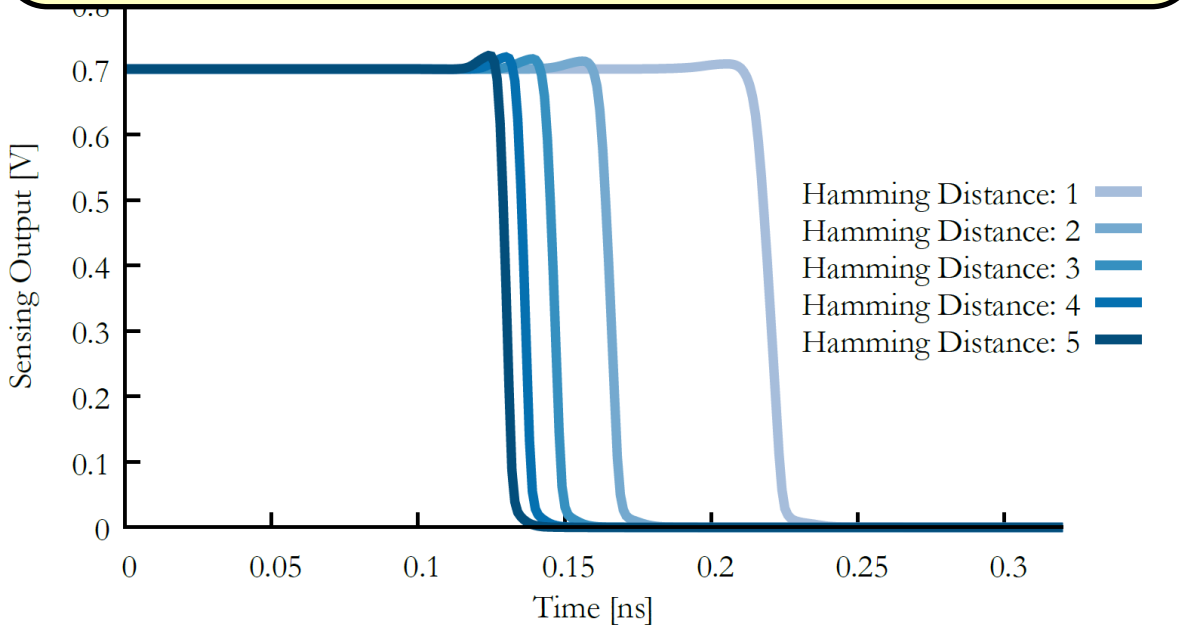
“Hi” → [H] XOR [Rotate(i)]

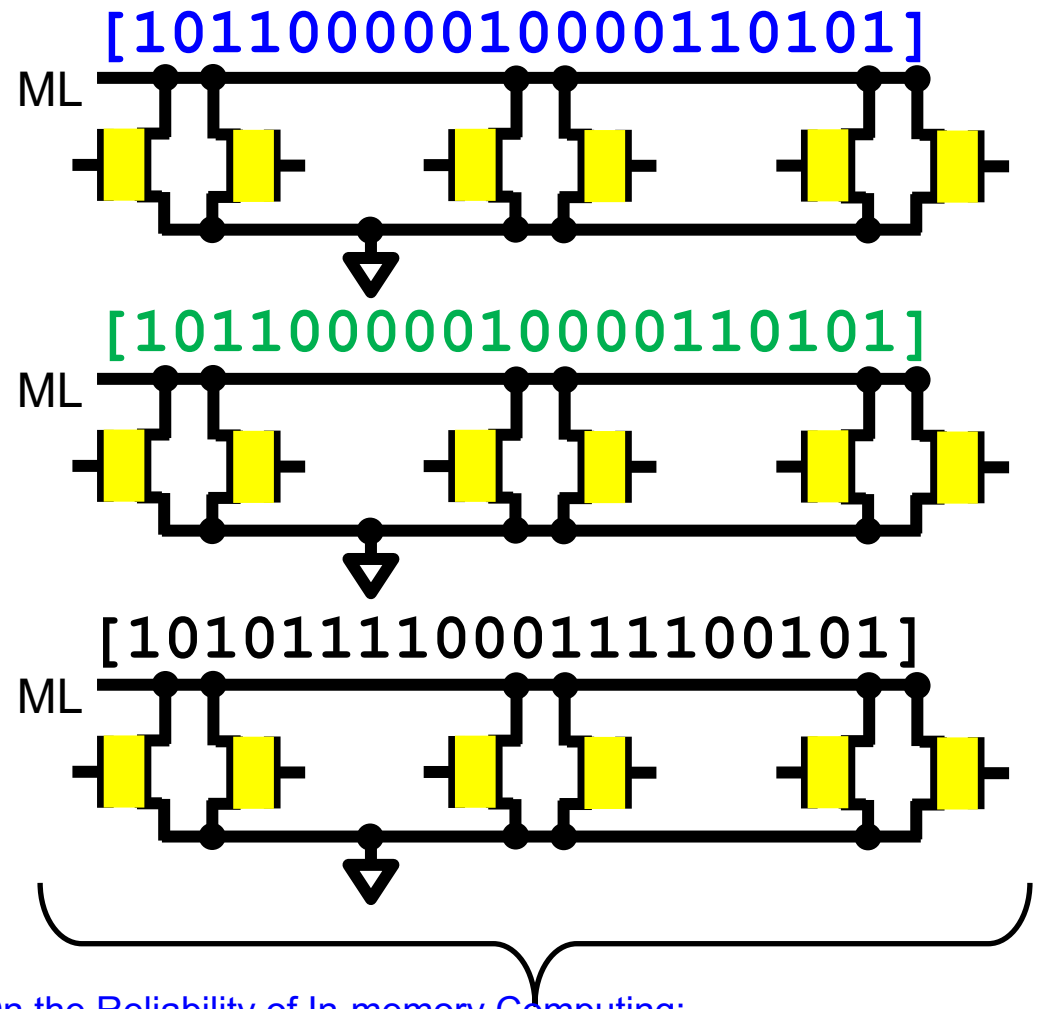
Brain-Inspired Hyperdimensional Computing



In-Memory Hyperdimensional Computing

The row with the smallest hamming distance → *best match*





Search = [00101011000100100011]

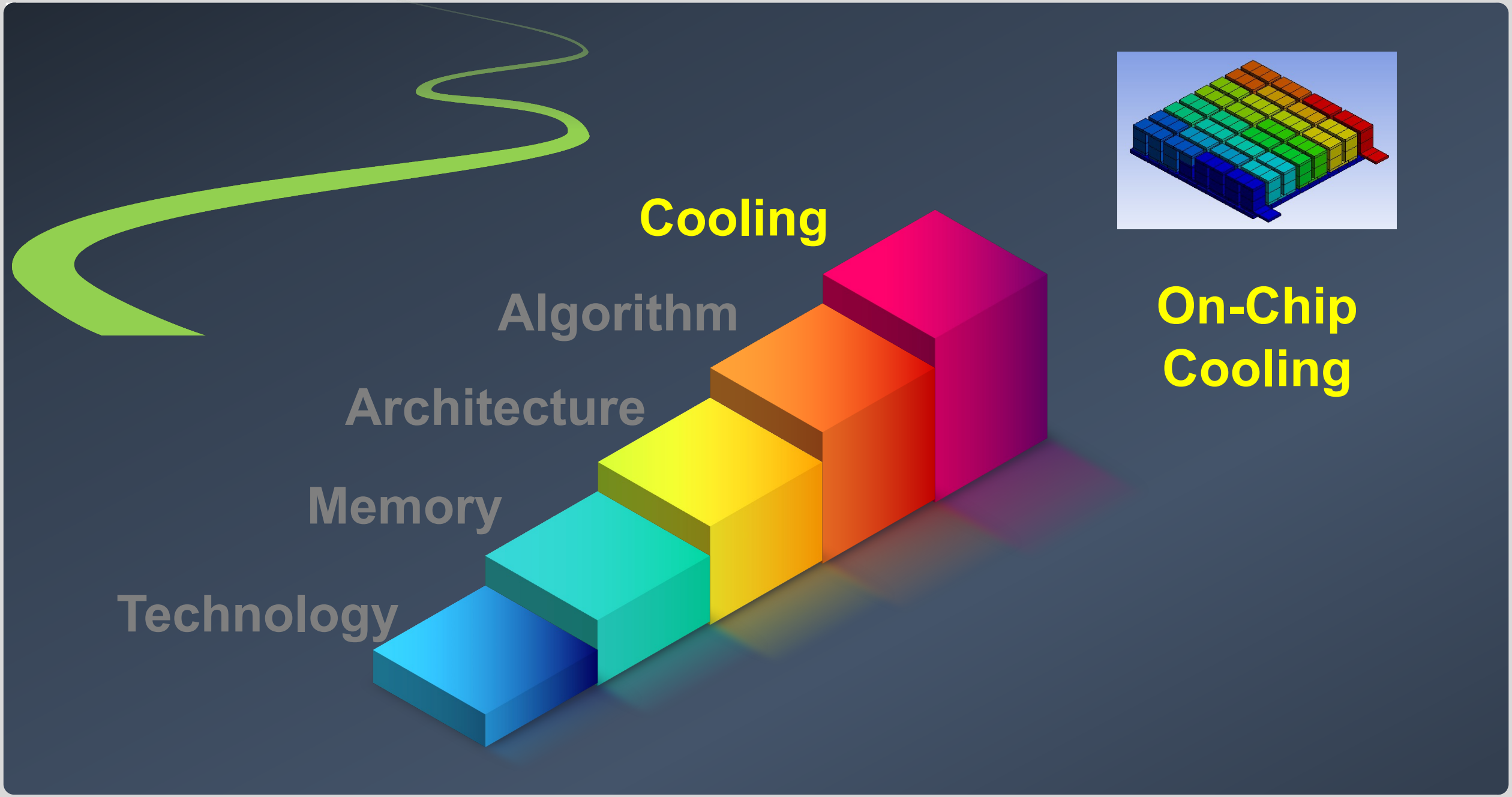
S. Thomann, C. Li, C. Zhuo, O. Prakash, X. Yin, X. S. Hu, and H. Amrouch, "On the Reliability of In-memory Computing: Impact of Temperature on Ferroelectric TCAM," **IEEE VLSI Test Symposium (VTS'21), 2021 (Best Paper Nomination)**

Why NOT Classical Deep Learning!

Beyond-CMOS Technology + Beyond-von-Neumann

→ HW Errors are inevitable!

Robust AI Algorithms against Errors is a MUST


Cooling
Algorithm
Architecture
Memory
Technology
On-Chip
Cooling

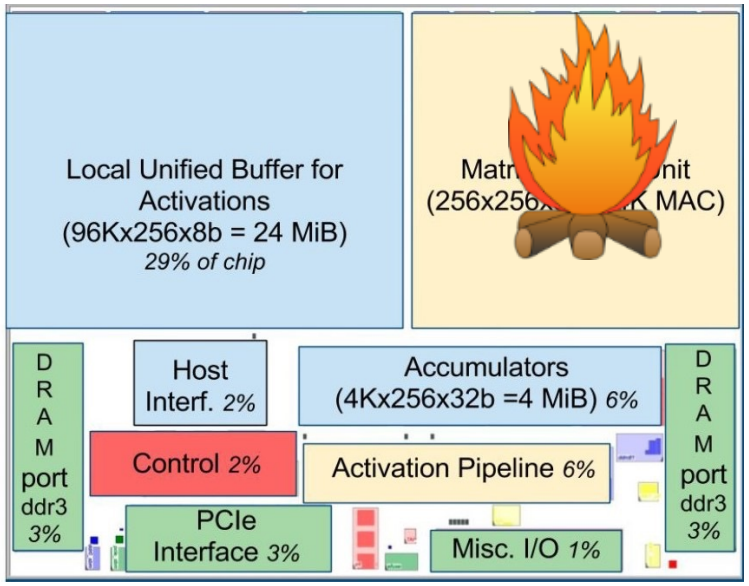
Intelligent Cooling: Why Now?

Google TPU…

“These chips are so powerful, that for the first time we've had to introduce liquid cooling in our data centers”, Google CEO Sundar Pichai in 2020

Source: https://www.datacenterdynamics.com/en/news/googles-latest-machine-learning-chip-to-use-liquid-cooling/

Intelligent Cooling: Superlattice Thermoelectric



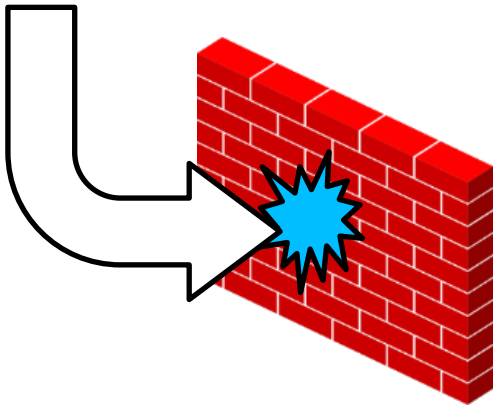
AI Chip: Google TPU [ISCA'17]

Traditional Cooling

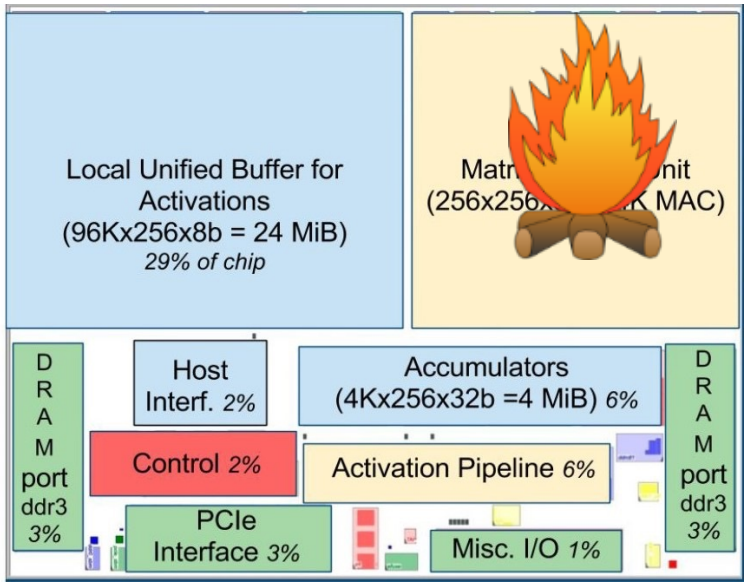
Cooling the entire chip

→ *Inefficient*

Cooling Wall



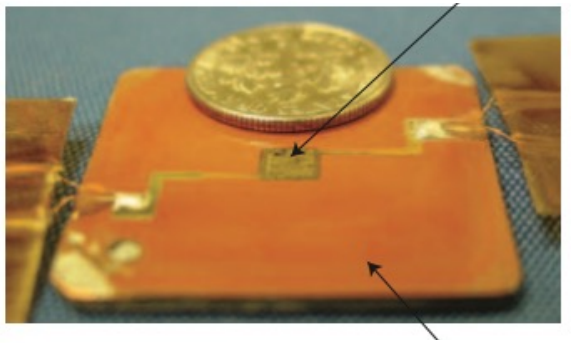
Intelligent Cooling: Superlattice Thermoelectric



AI Chip: Google TPU [ISCA'17]

On-chip cooling:
localized and On-demand
→ VERY Efficient

Superlattice cooler



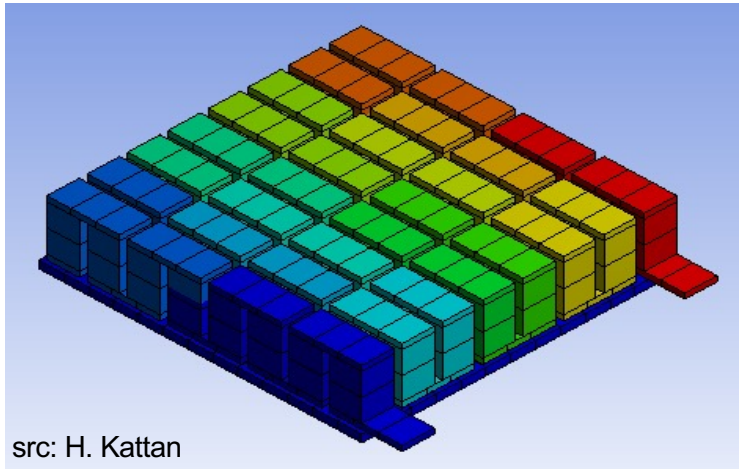
Heat spreader

Bulman, G., Barletta, P., Lewis, J. *et al.* Superlattice-based thin-film thermoelectric modules with high cooling fluxes. ***Nature Communication***, 2016

Chowdhury, et al., "On-chip cooling by superlattice-based thin-film thermoelectrics," ***Nature Nanotechnology***, vol. 4, no. 4, pp. 235-238, 2009

Intelligent Cooling: Superlattice Thermoelectric

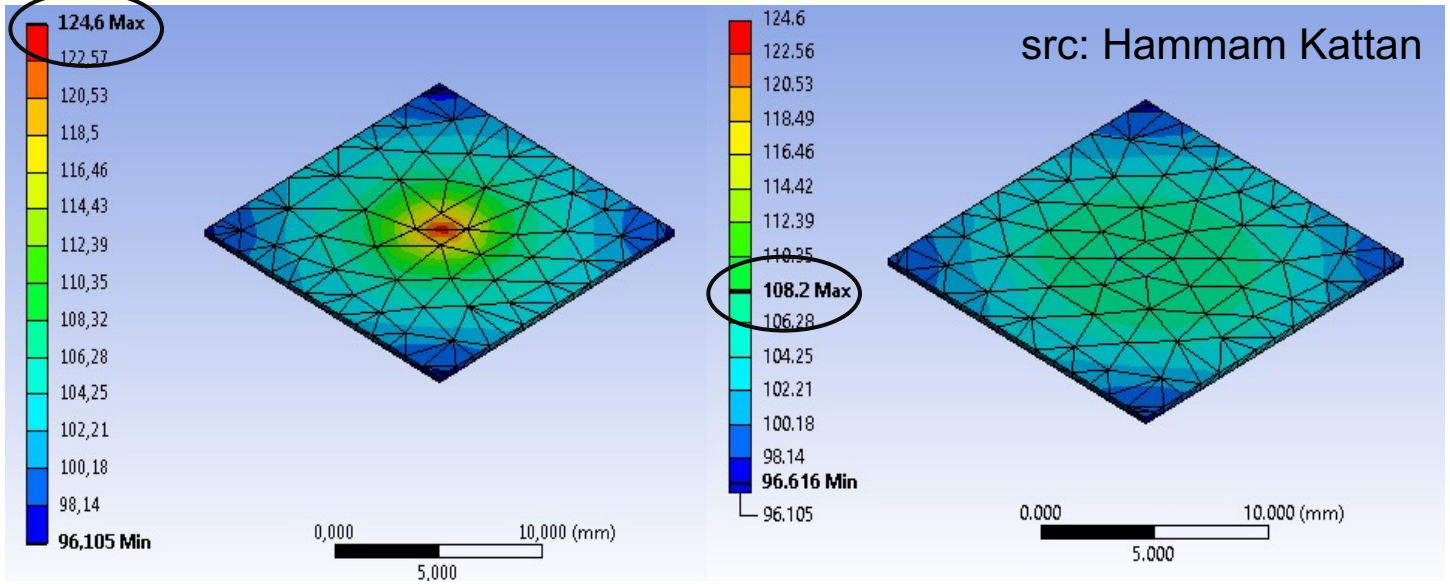
Thermoelectric:
Peltier's effect $Power \rightarrow \Delta T$


src: H. Kattan

ANSYS®

H. Kattan / H. Amrouch "On-demand Mobile CPU Cooling with Thin-Film Thermoelectric Array", IEEE Micro Magazine (**MICRO**), 2021

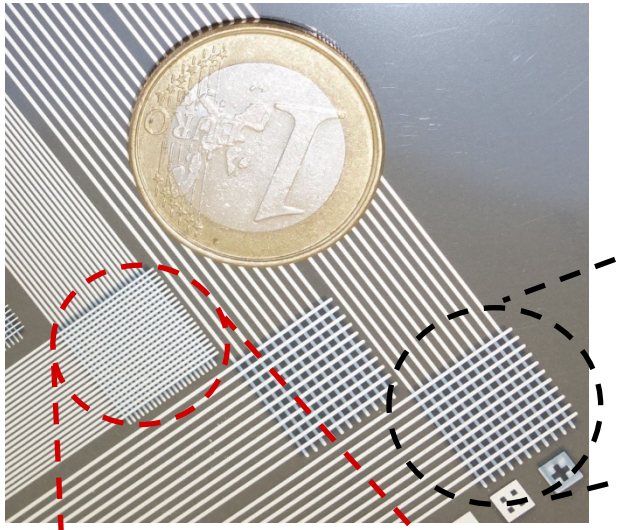
On-chip cooling:
localized and On-demand
→ VERY Efficient



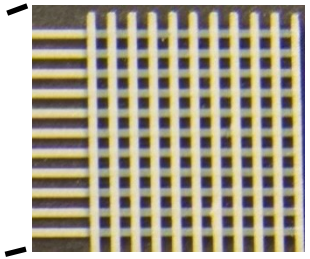
Suppressing Heat Flux (Hot-Spot) of $200\ W/cm^2$

Intelligent Temperature Sensing

We developed the First Printed Thermal Sensors Array for Processor Chips



11 x 11 =121
Thermal Sensors



1cm

1cm

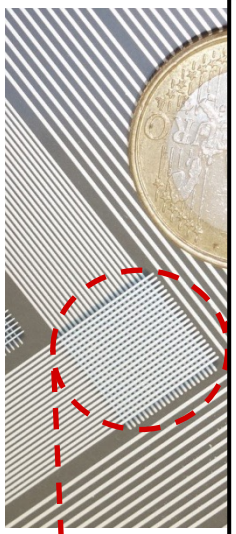
21 x 21 = 441/cm^2
Thermal Sensors

Ack: Heidelberg InnovationLab, U. Lemmer, KIT

Intelligent Temperature Sensing

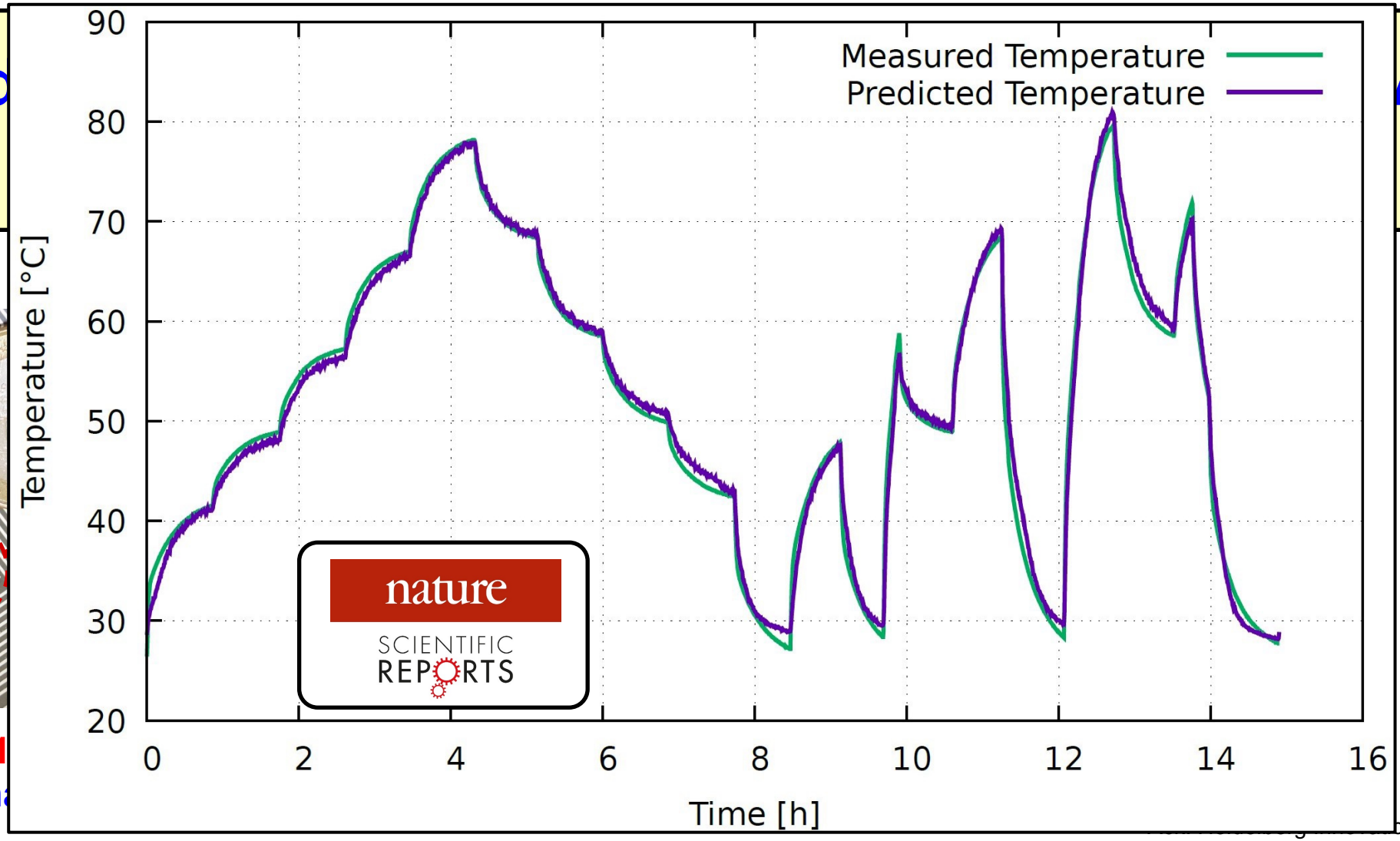
We c

Array



21 x 21

Therma



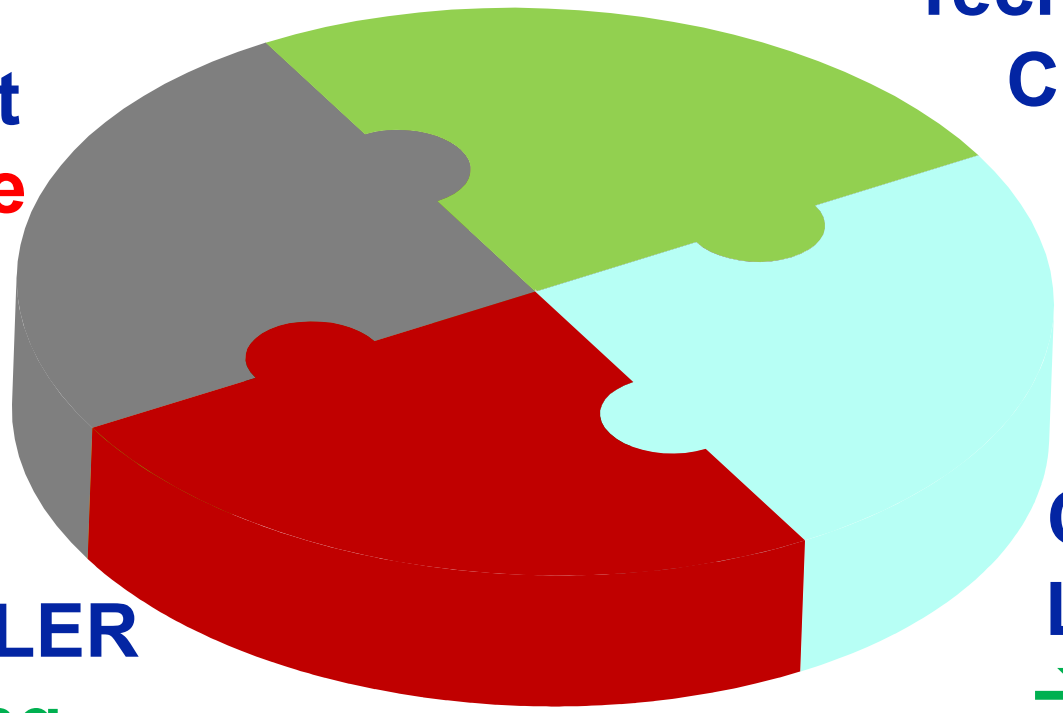
nLab, U. Lemmer, KIT

So What? …. Hope or Hype?

Indeed … Hope but currently lots of Hype ….

In-memory Computing nice but errors are inevitable

Technology/Algorithm Codesign is KEY



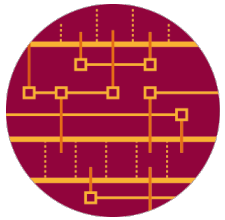
Temperature is KILLER
→On-Chip Cooling

Conventional Deep Learning is Hungry
→ Novel Algorithms

Funding Agencies

Bundesministerium
für Bildung
und Forschung

ADVANTEST®



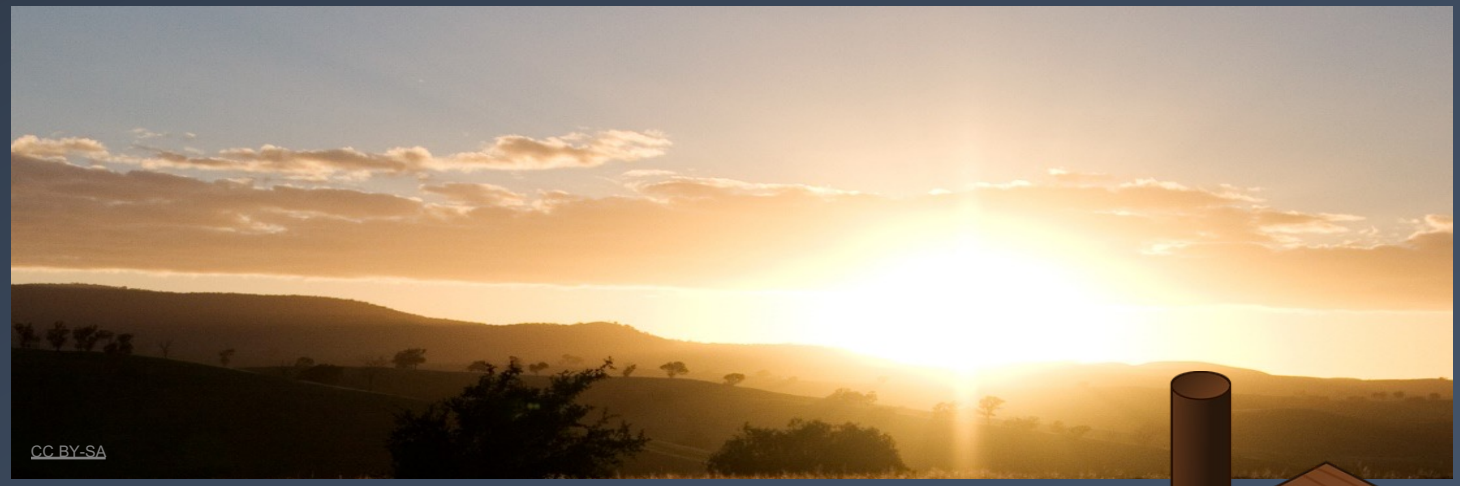
GS-IMTR
Graduate School
Intelligent Methods for Test and Reliability

DAAD
Deutscher Akademischer Austausch Dienst
German Academic Exchange Service

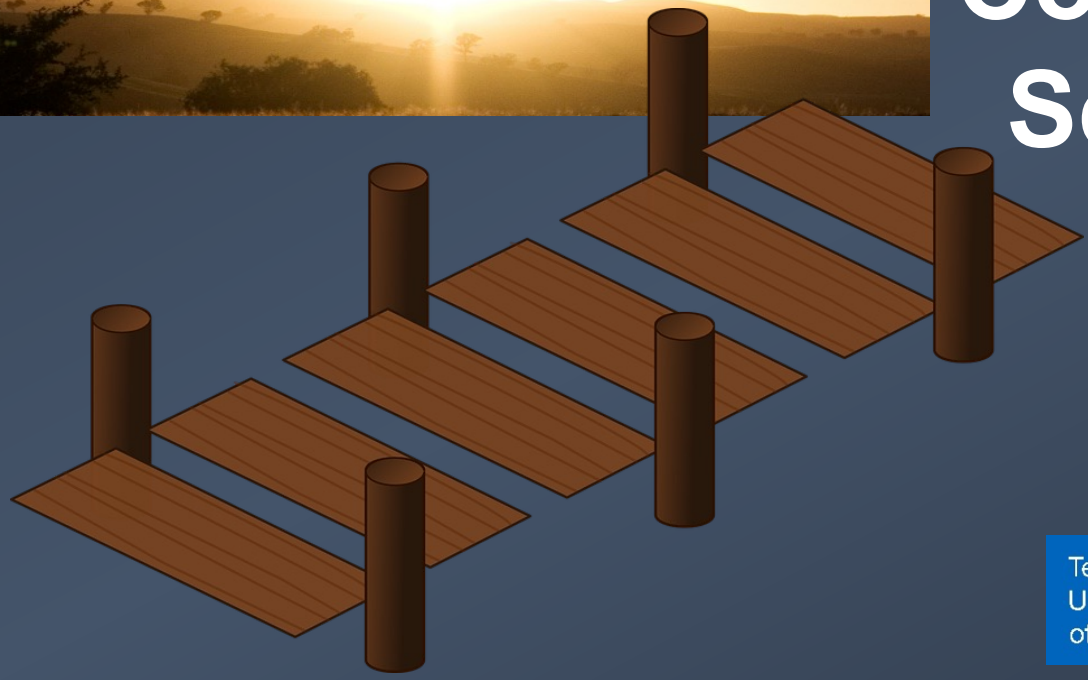
DFG Deutsche Forschungsgemeinschaft

Collaborations

On the Brink of a new Era in AI Hardware



Computer Science



Device Physics

Technical University of Munich