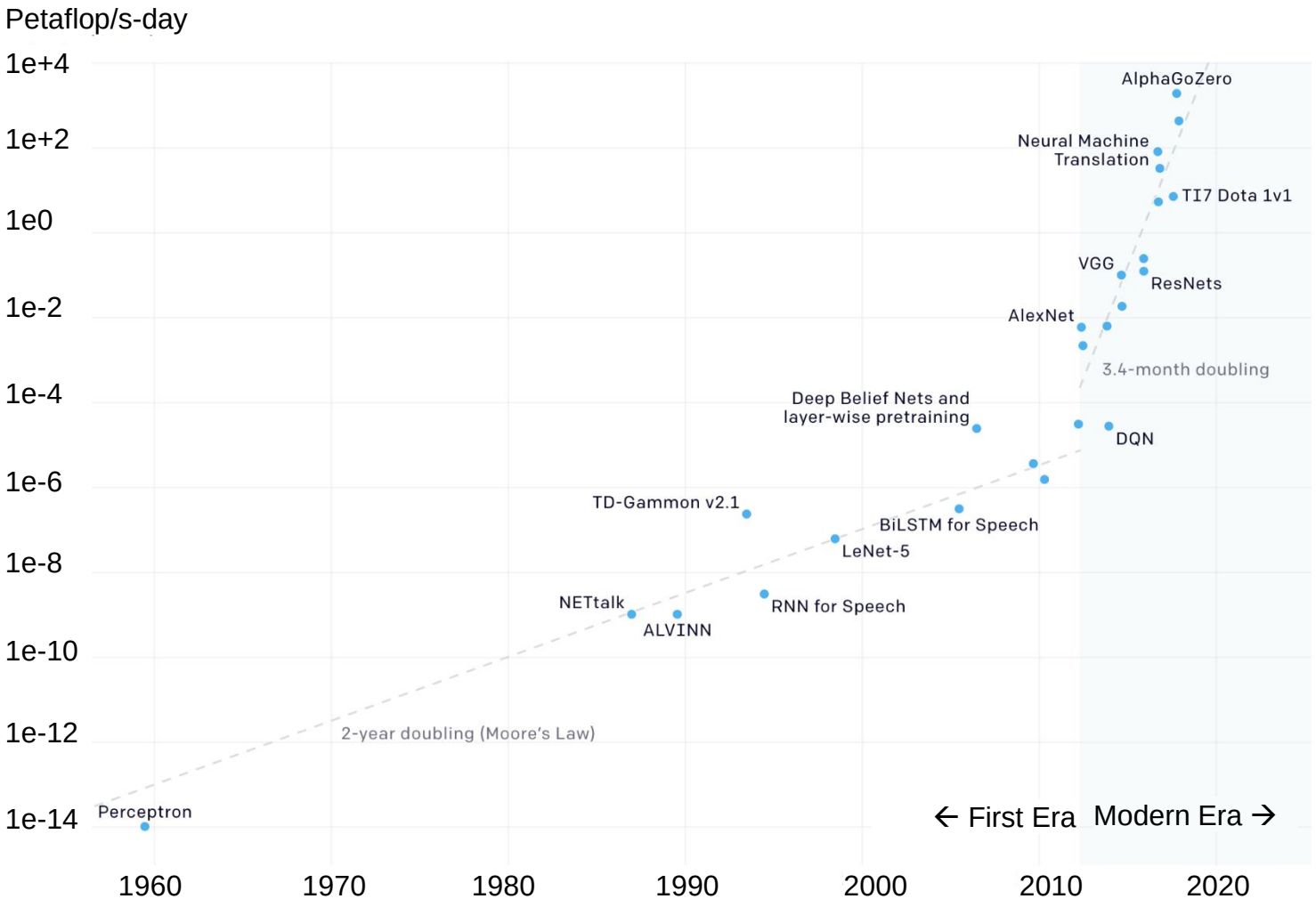
# Challenges in the Era of AI Chips and Advance Technologies

by Hussam Amrouch
Chair of AI Processor Design

# AI : The Next Revolution



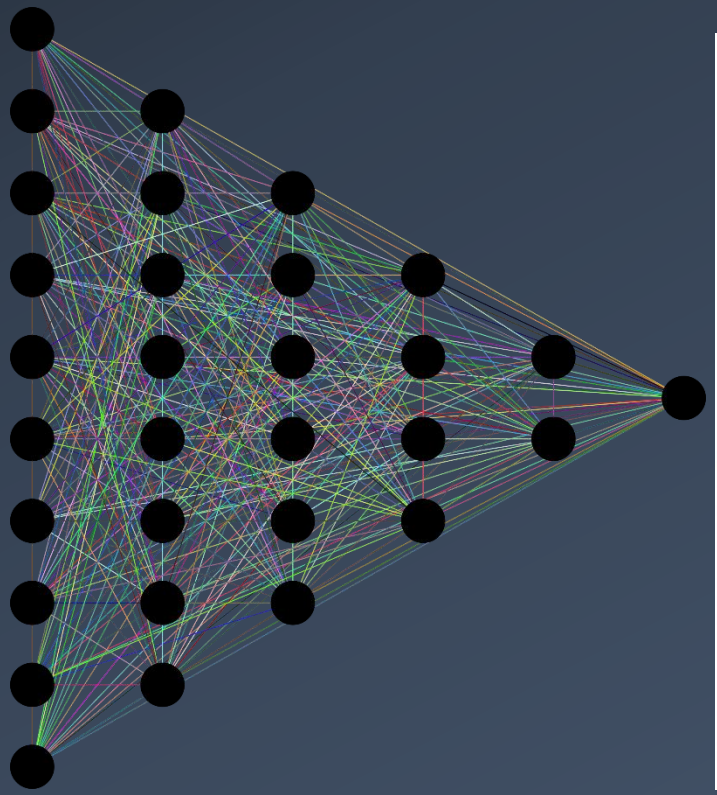

**Computing demand**

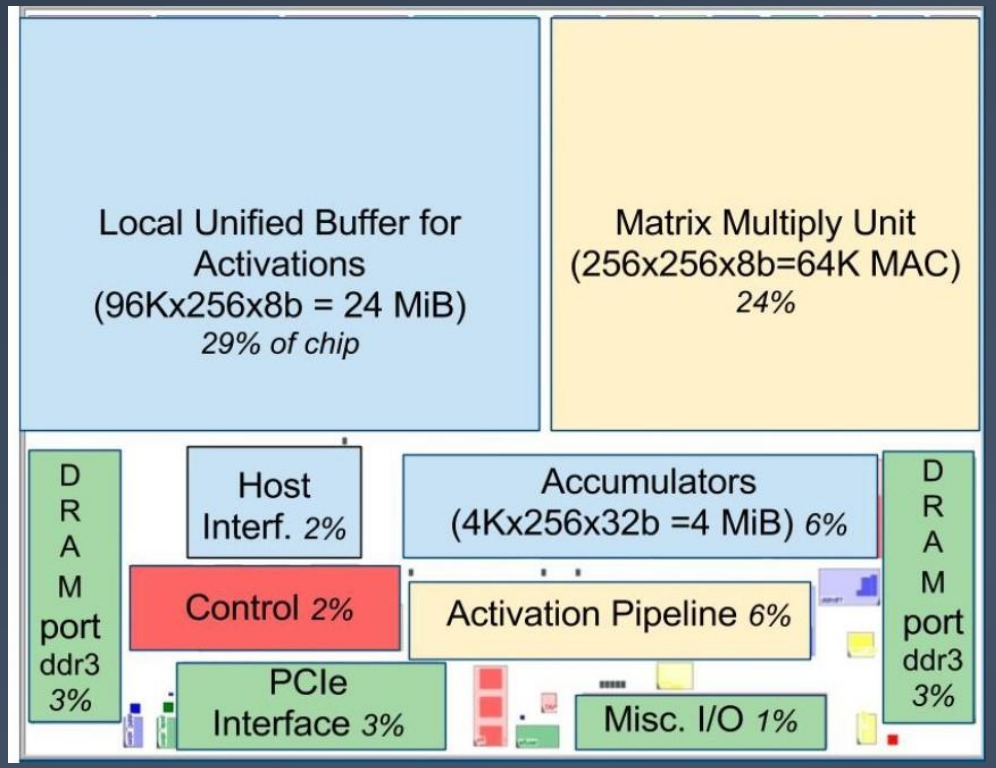
**3.4 months doubling!**

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….**

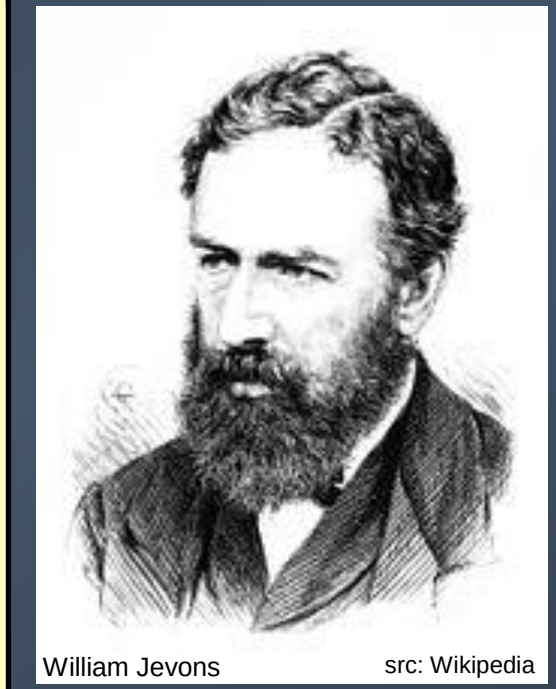
**More Efficiency is really Good?**

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

**Jevons Paradox**

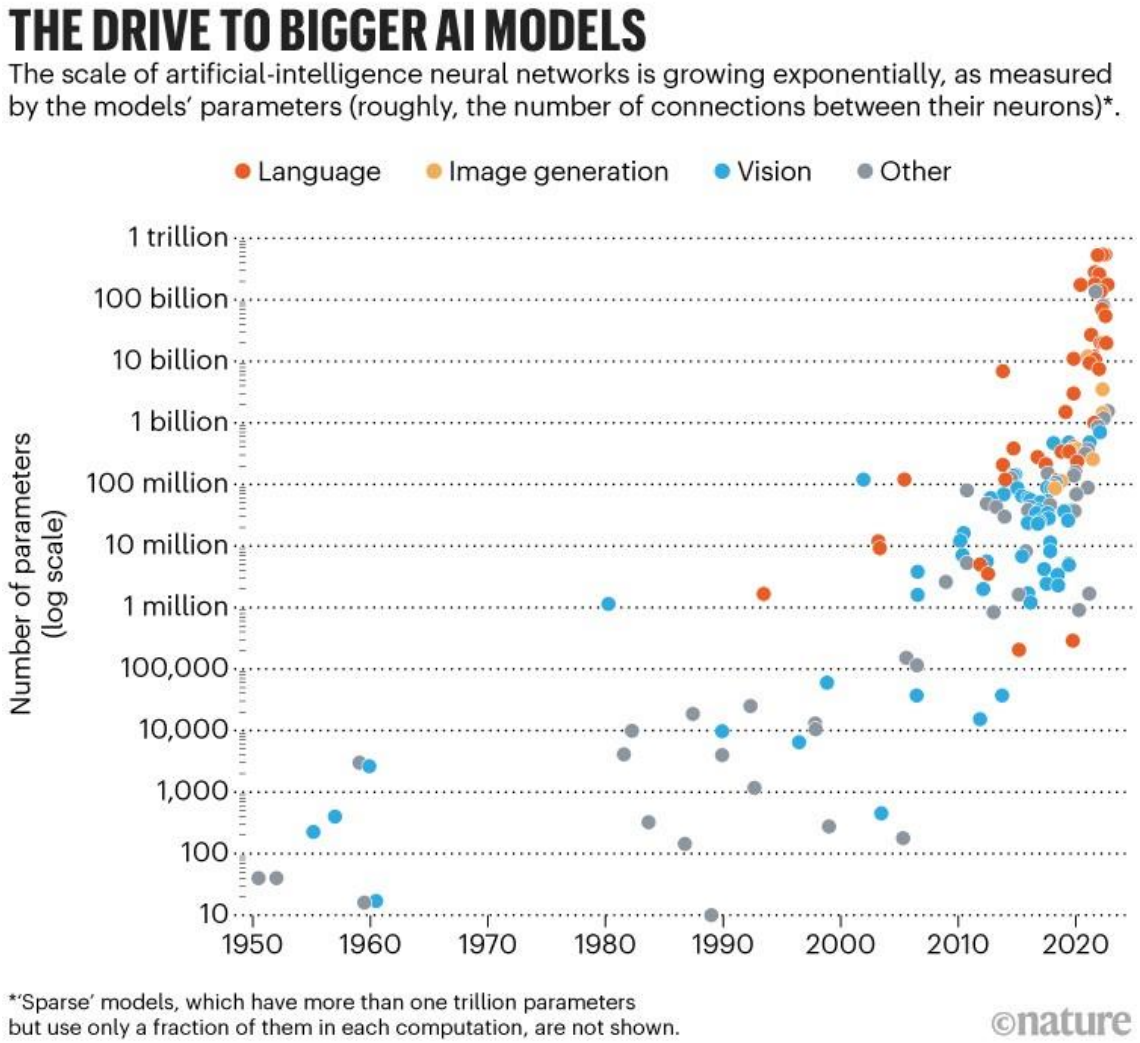
**When technology increases the efficiency, the consumption rises**

→ ***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!

→ **Significant Efficiency Loss**

# Energy Crisis in AI Hardware

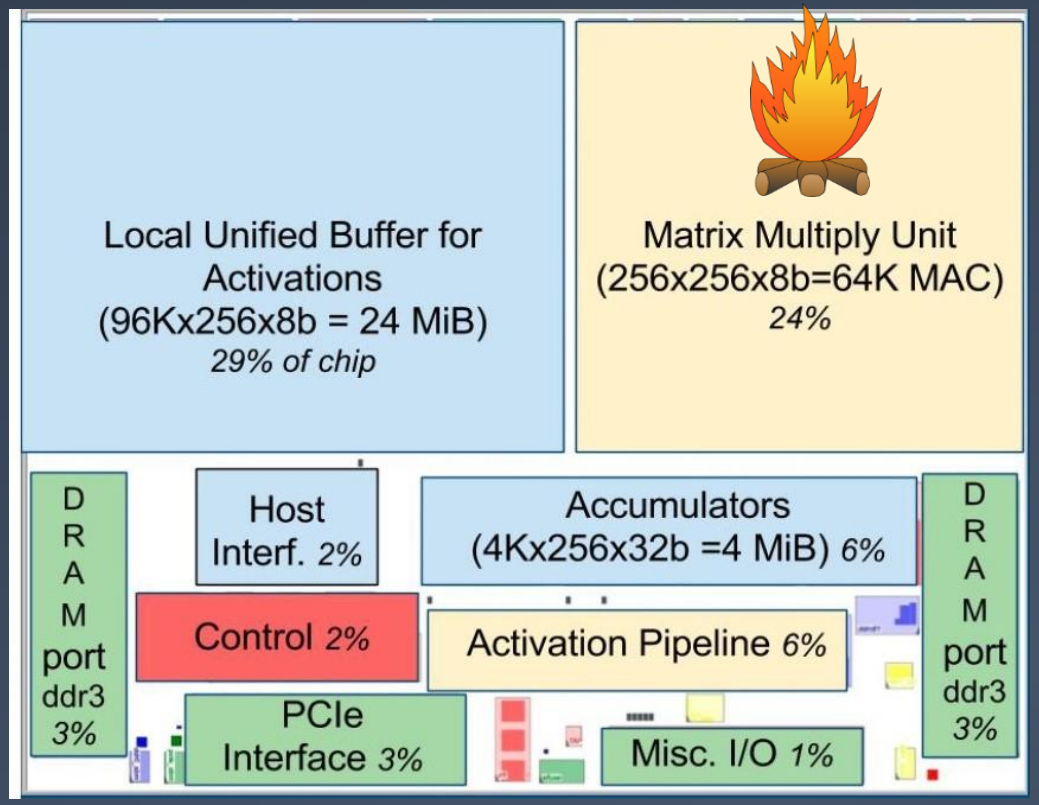
**Insufficient On-Chip Memory**

↓

**Massive Data to Move**

↓

**von-Neumann Bottleneck**




AI Chip: Google TPU [ISCA'17]

**Massive Computation**

↓

**Excessive Heat**

↓

**Expensive Cooling**

# Where is the Problem?

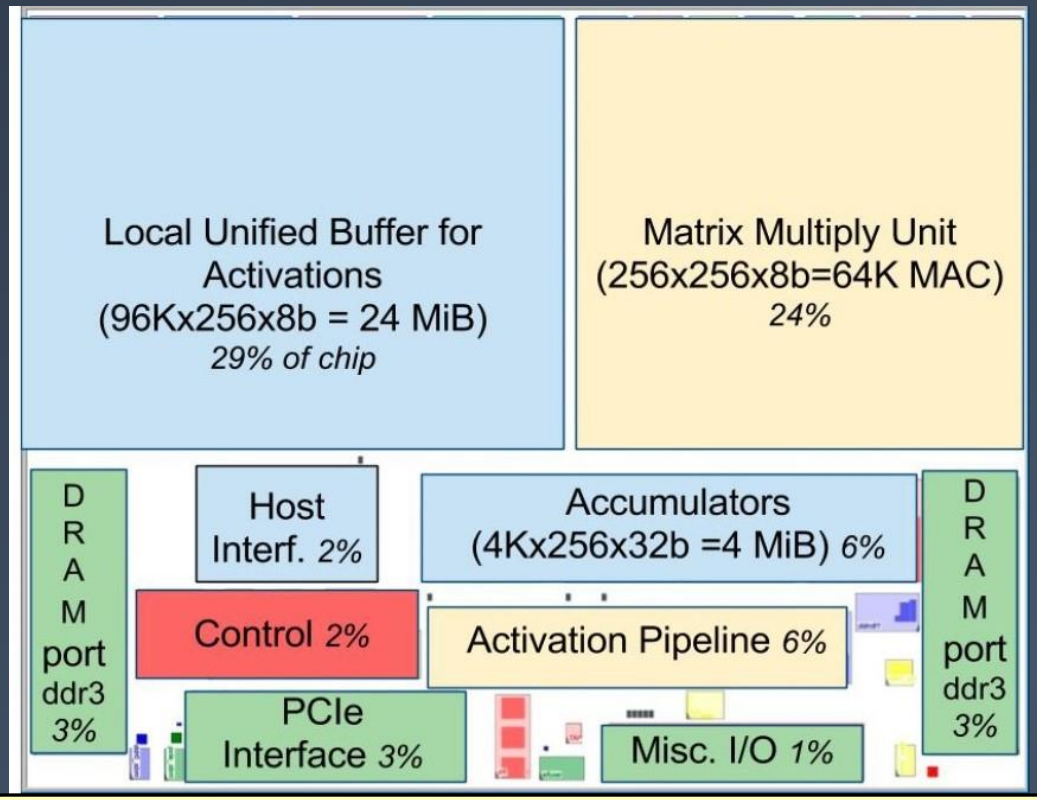
**Insufficient On-Chip Memory**

↓

**Massive Data to Move**

↓

**Von-Neumann Bottleneck**




**Massive Energy Cost**

**Massive Computation**

↓

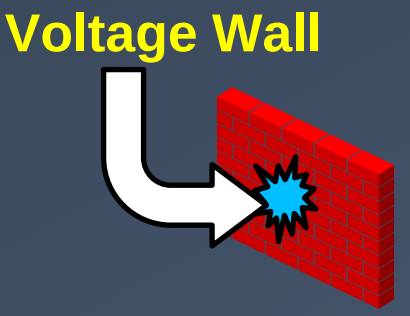
**Excessive Heat**

↓

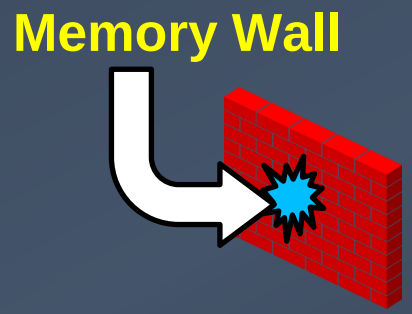
**Expensive Cooling**

# But….What is the Root?

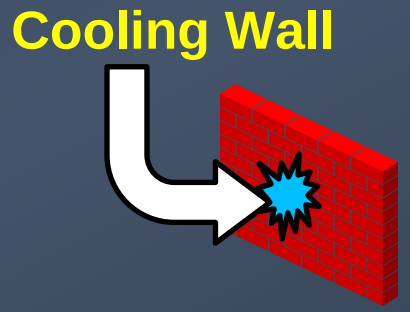
- Voltage: Reaching its Fundamental Limit




- Memory: Massive Data in DNNs

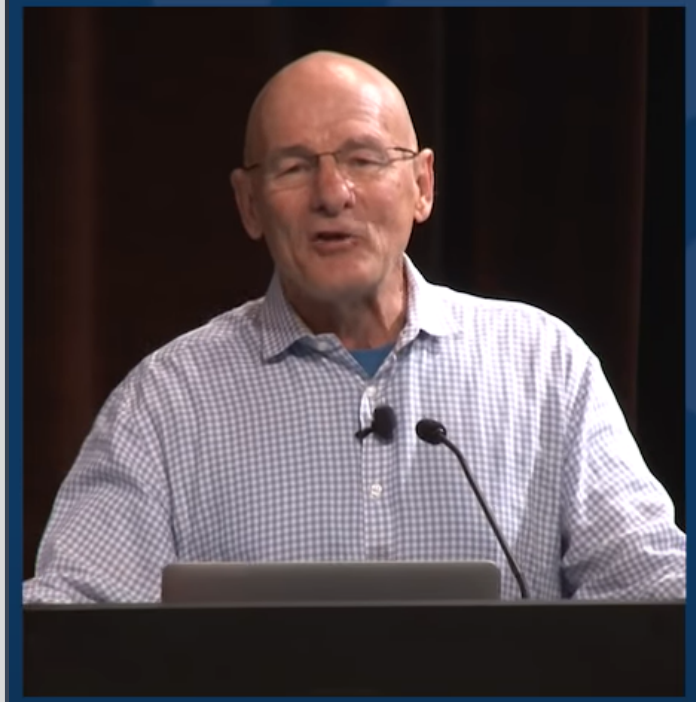



- Cooling: Inherently Inefficient

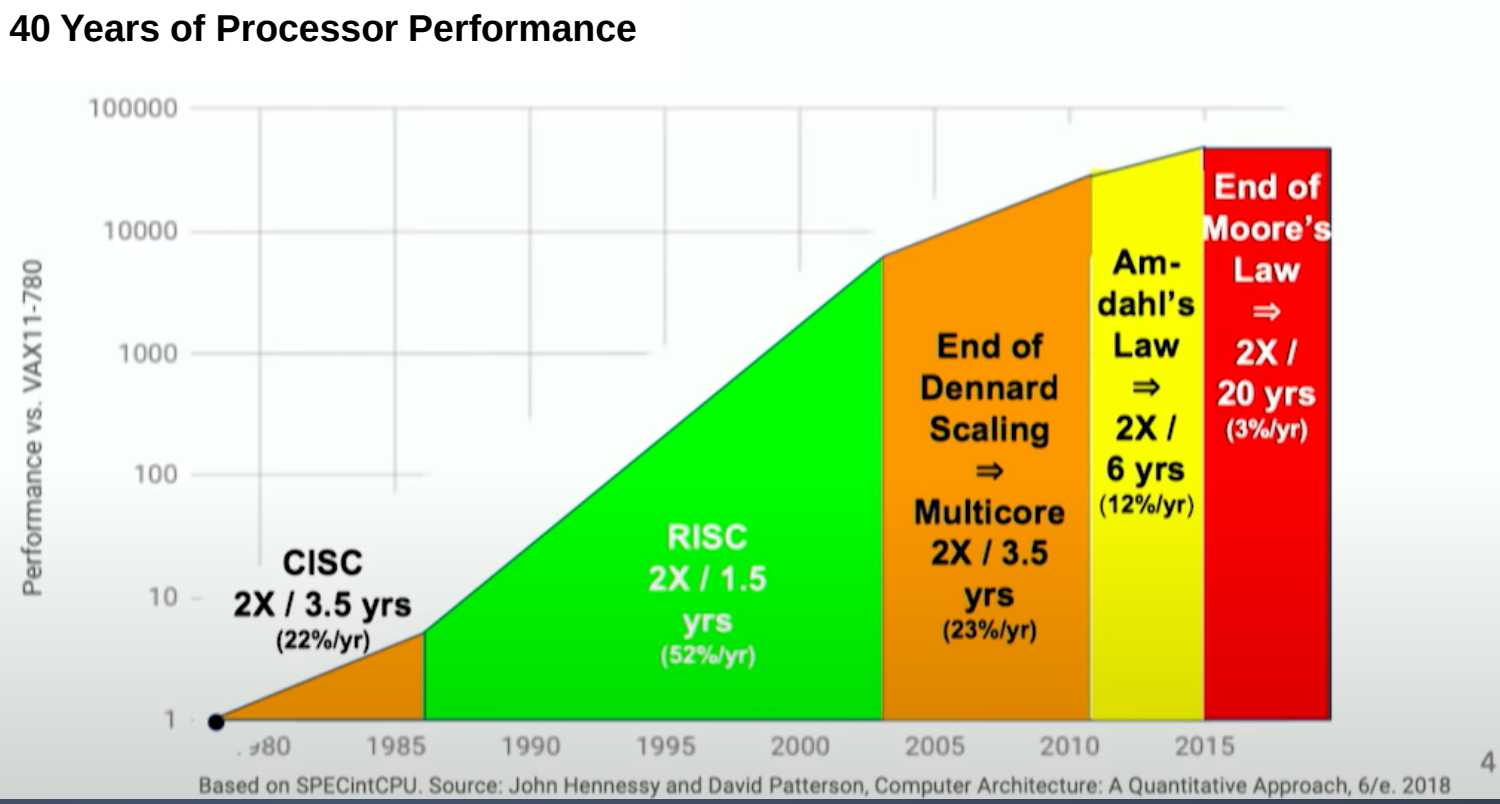


# Performance: End of the Line…

**David Patterson**
**UC Berkeley, Google**

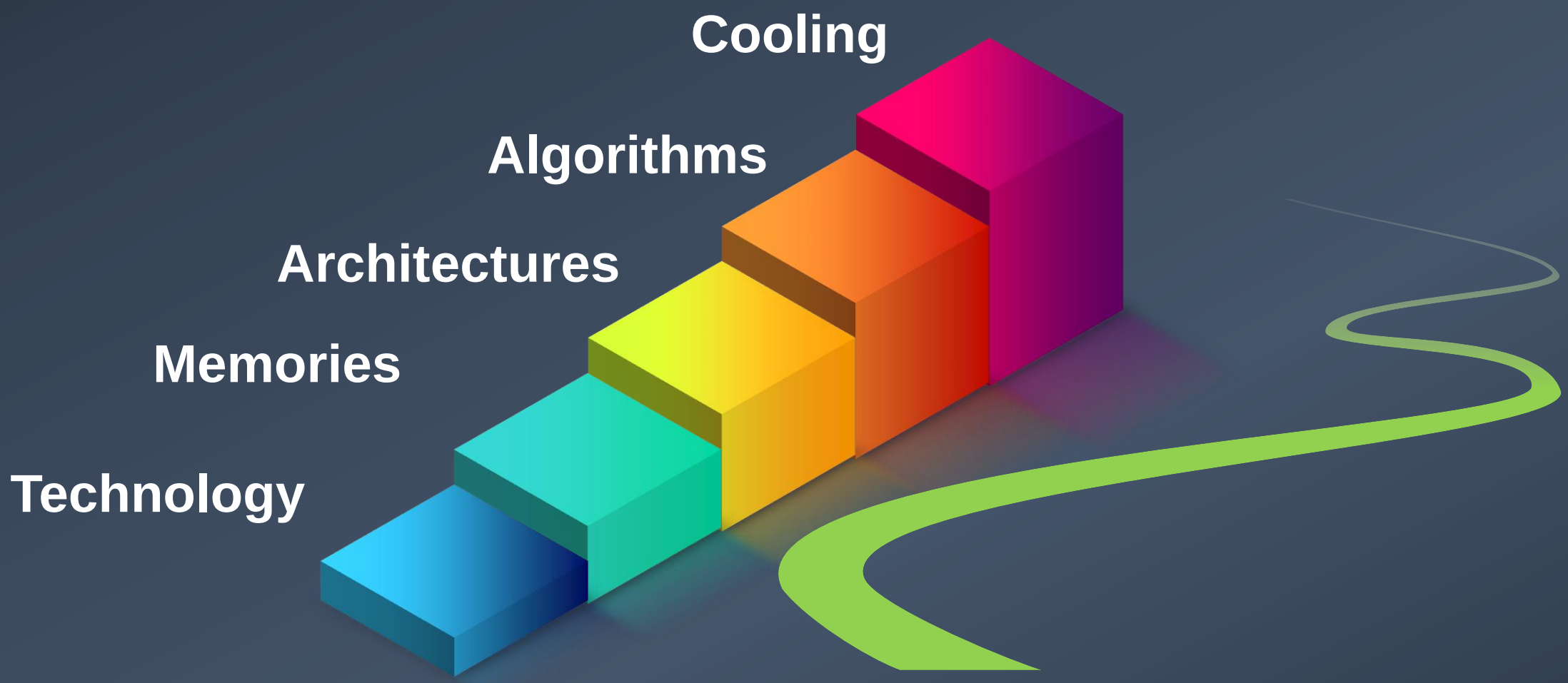


## End of Growth of Performance?



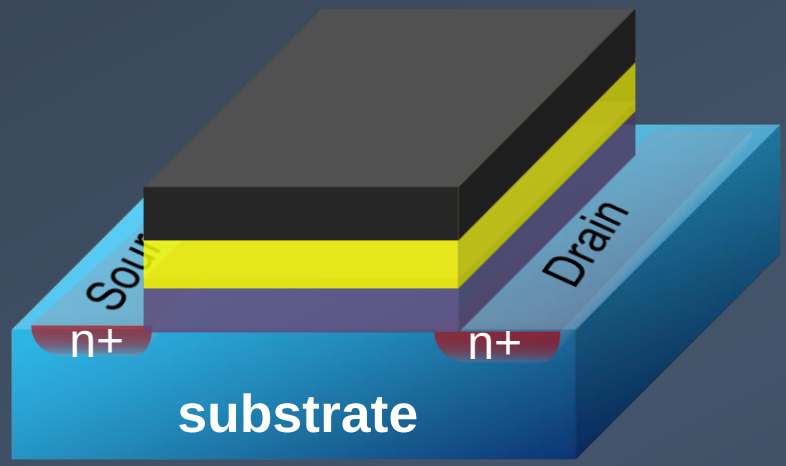

src: https://www.youtube.com/watch?v=FSwKCL8A9JQ&t=2163s

# Industry needs Innovations in

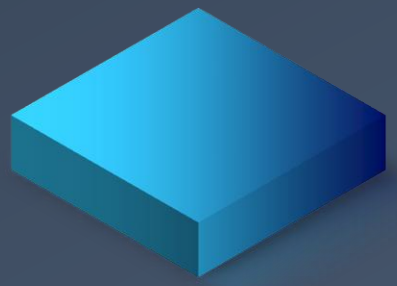



ack: Creative Venus

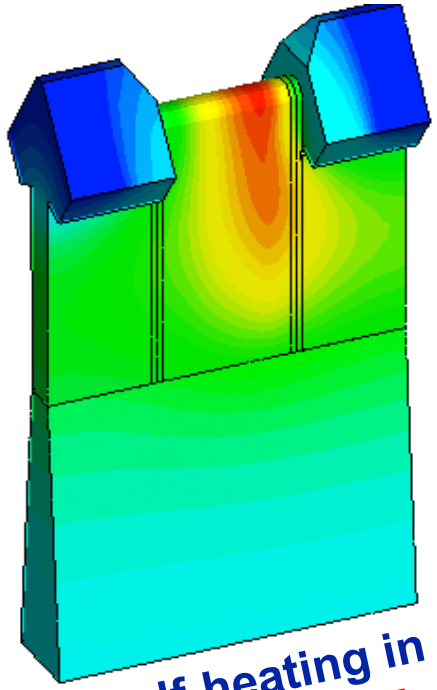
# What's Wrong in Advanced Nodes (5nm, 3nm, 2nm...)




Technology



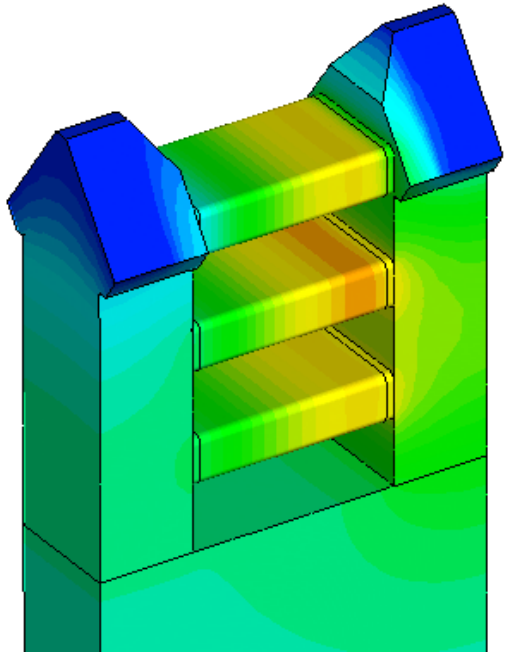
# Reliability is BIG Killer!



Lattice Temperature (K)

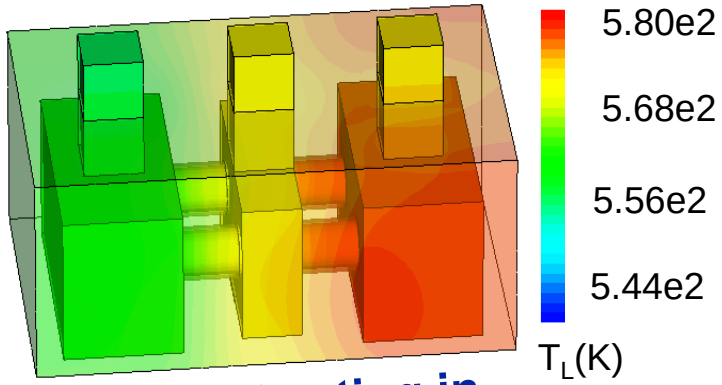
355
345
336
327
318
309
299

Self-heating in
7nm FinFET



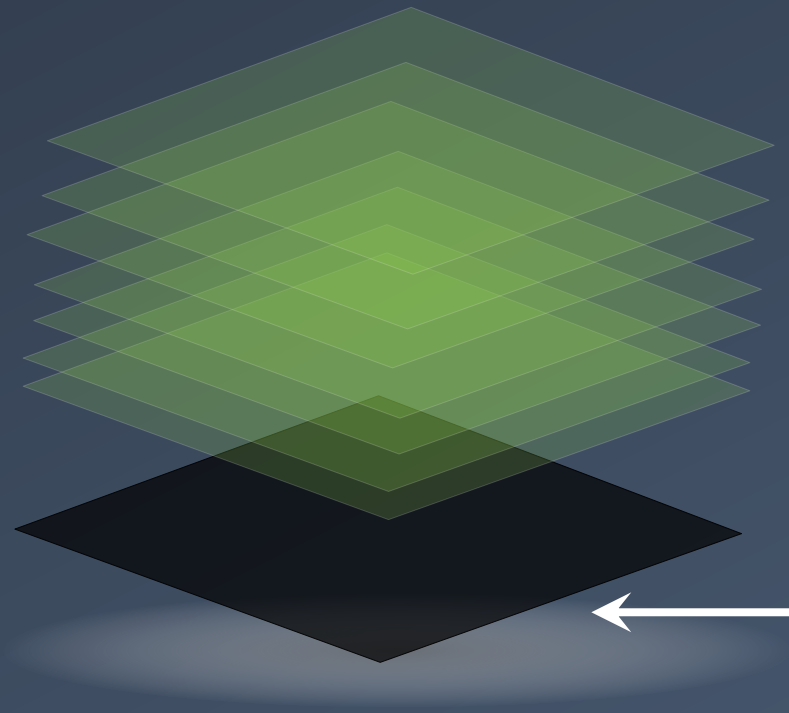
Self-heating in
7nm Nanosheet

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




Self-heating in
14nm Nanowire

# 3D AI Chips Architecture

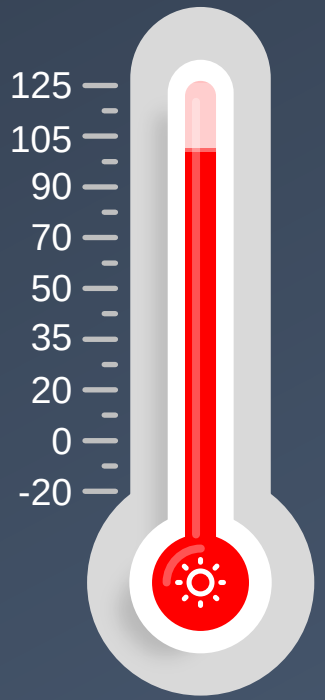


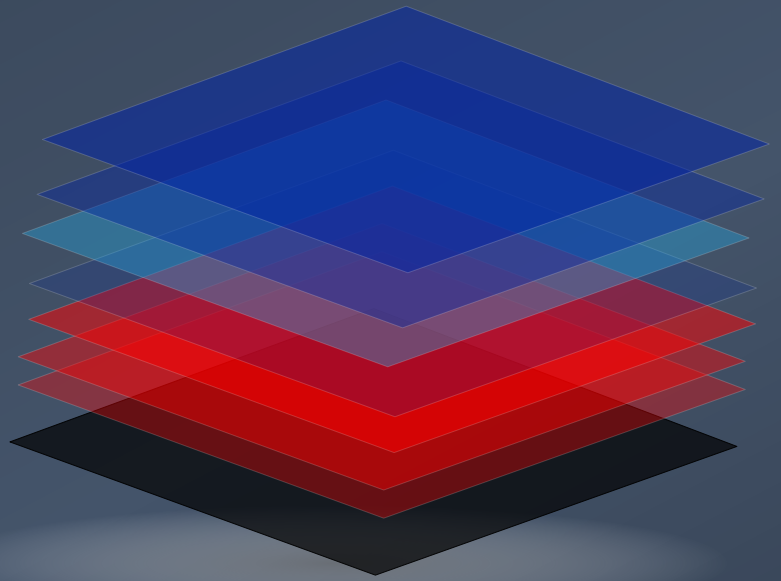
Different layer of memory

High-Performance
Logic CMOS

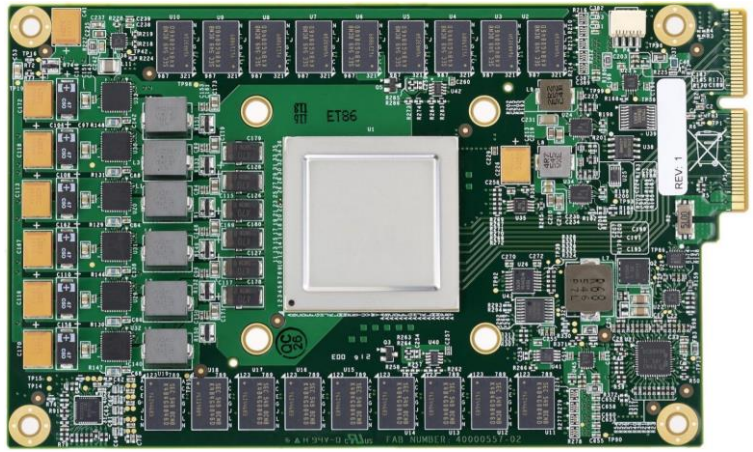
# Temperature: The Unseen Enemy for 3D

**Monolithic 3D**
**Neuromorphic**



# Deep Learning is REALLY Power Hungary!

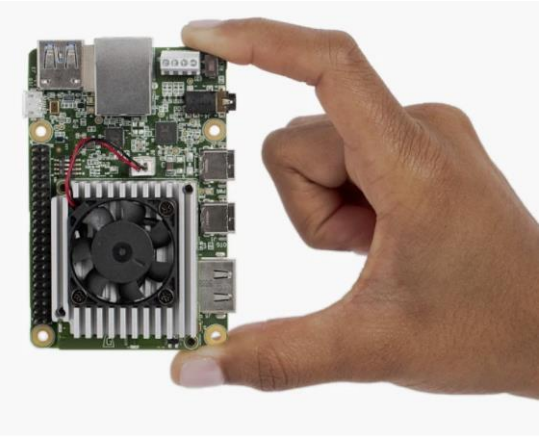


Google TPU [ISCA'17]

**Datacenters**


Samsung Exynos 9 [samsung.com]

**Mobile Devices**

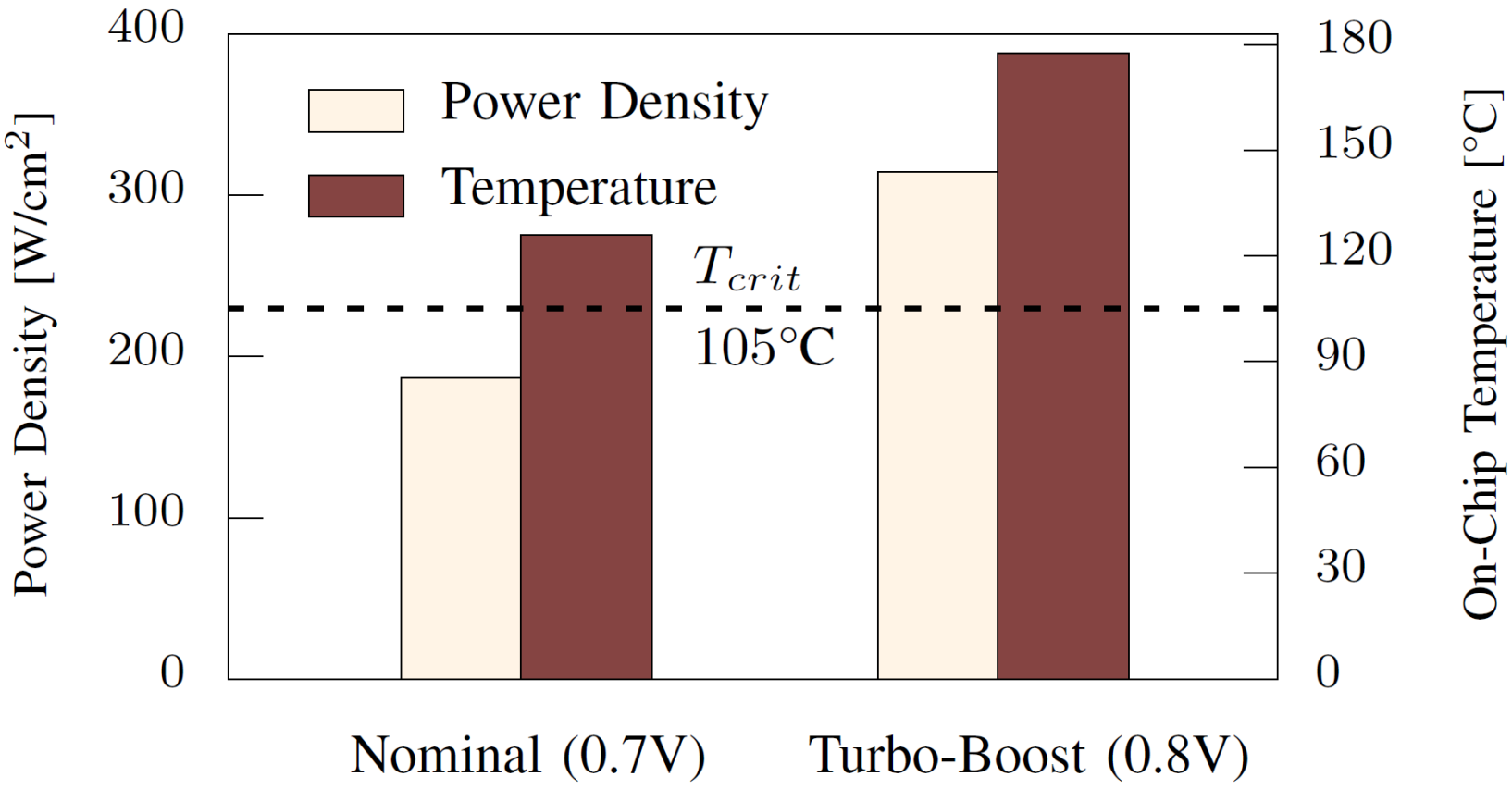


Google EDGE TPU [coral.ai]

**Edge-Computing**

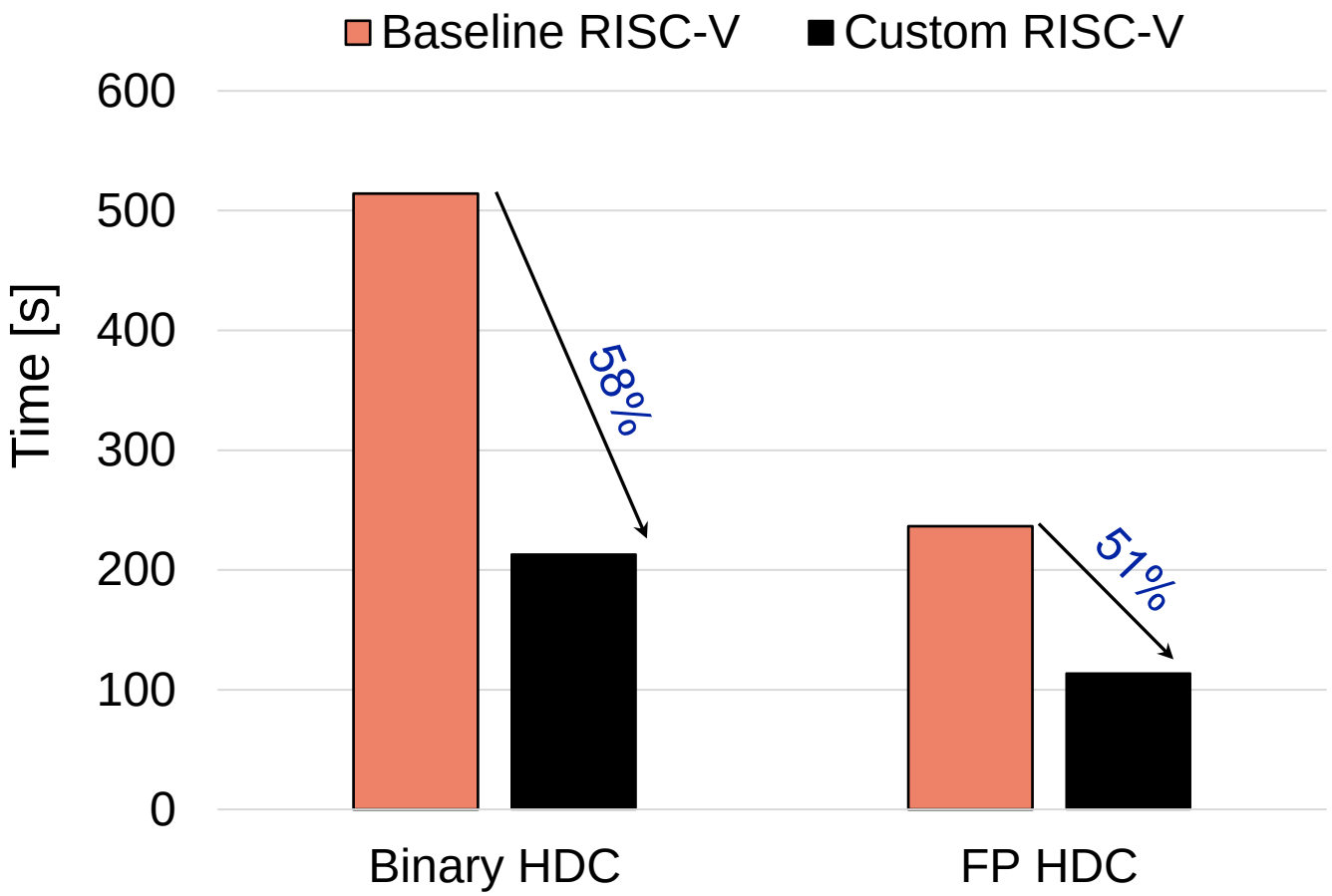
# Deep Learning is REALLY Power Hungry!

**Why not alternative algorithm to Deep Learning?**



# AI Chips: From Tradition to Customization

# RISC-V Customization for Edge AI: Training



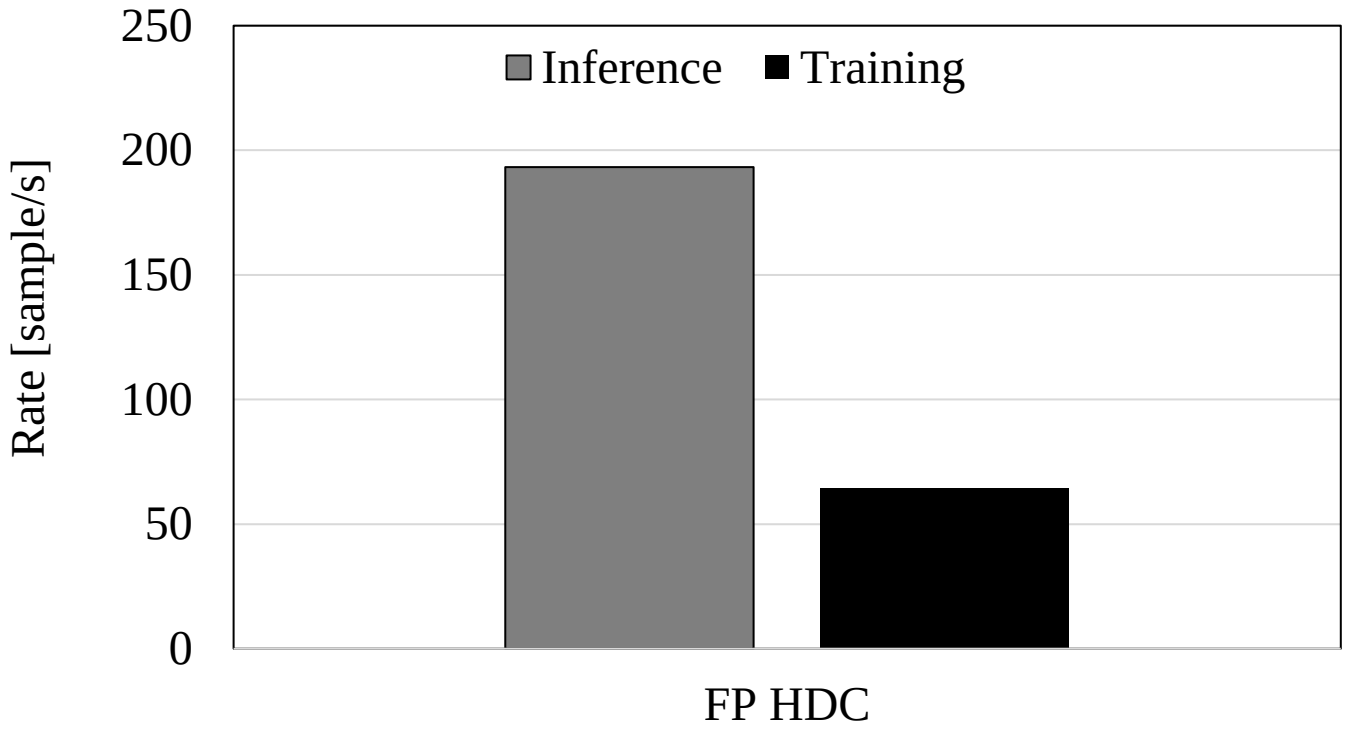

**Human activity dataset with ~7500 samples**

**Training in the order of seconds**

**Achieving a similar accuracy as QNN**

GlobalFoundries™ SYNOPSYS®

# RISC-V Customization for Edge AI: Training




Inference rate reach ~200 samples per second

GlobalFoundries™ SYNOPSYS®

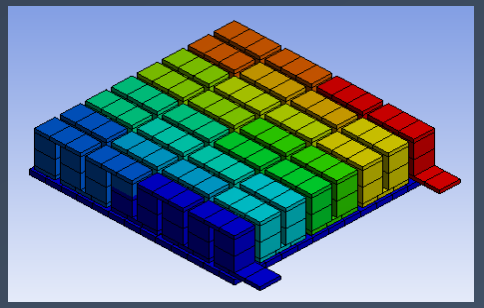
Cooling

Algorithm

Architecture

Memory

Technology



On-Chip Cooling

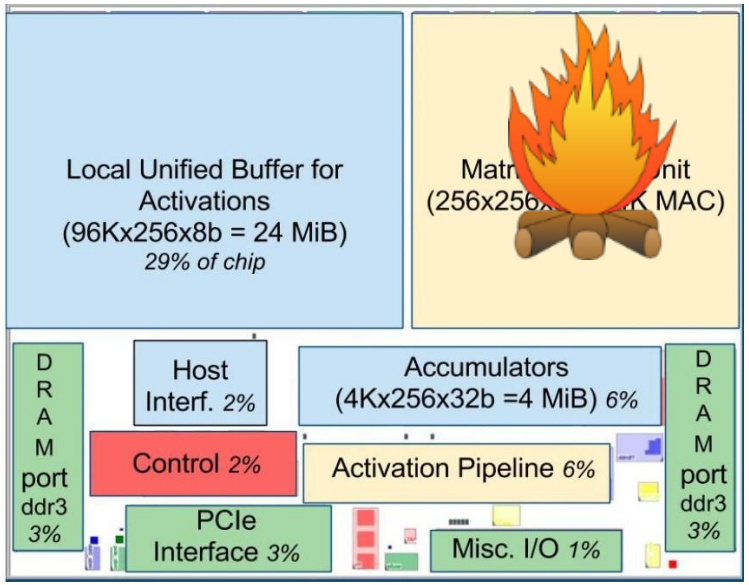
# Intelligent Cooling: Why Now?

***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/

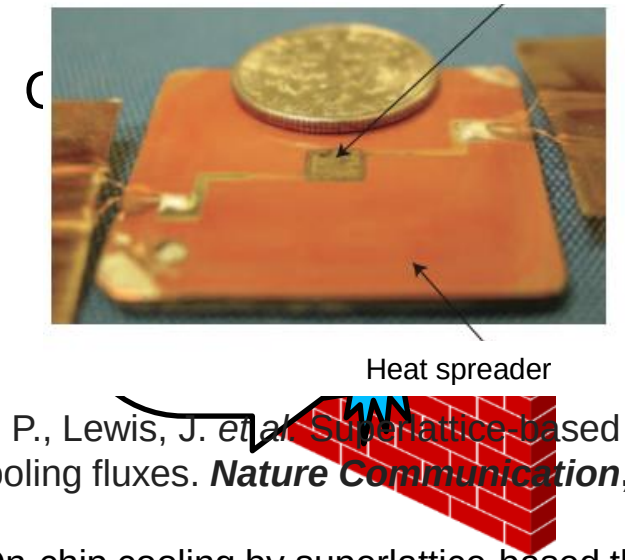
# Intelligent Cooling: Superlattice Thermoelectric




AI Chip: Google TPU [ISCA'17]

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

Superlattice cooler

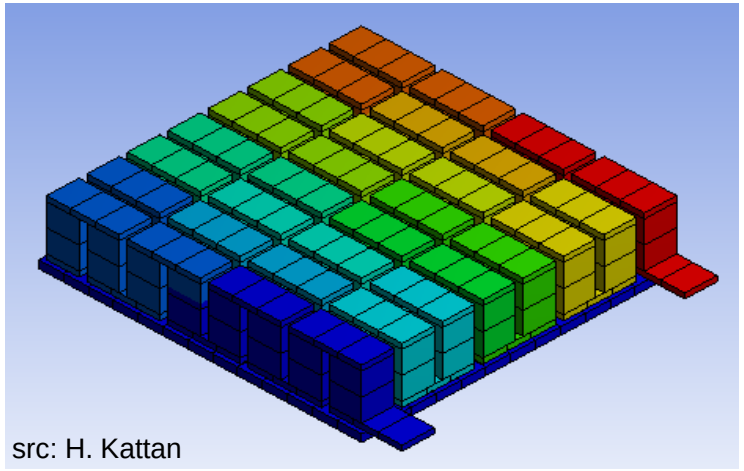



Bulman, G., Barletta, P., Lewis, J. *et al.* Superlattice-based thin-film thermoelectric modules with high cooling fluxes. ***Nature Communication*****,** 2016

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

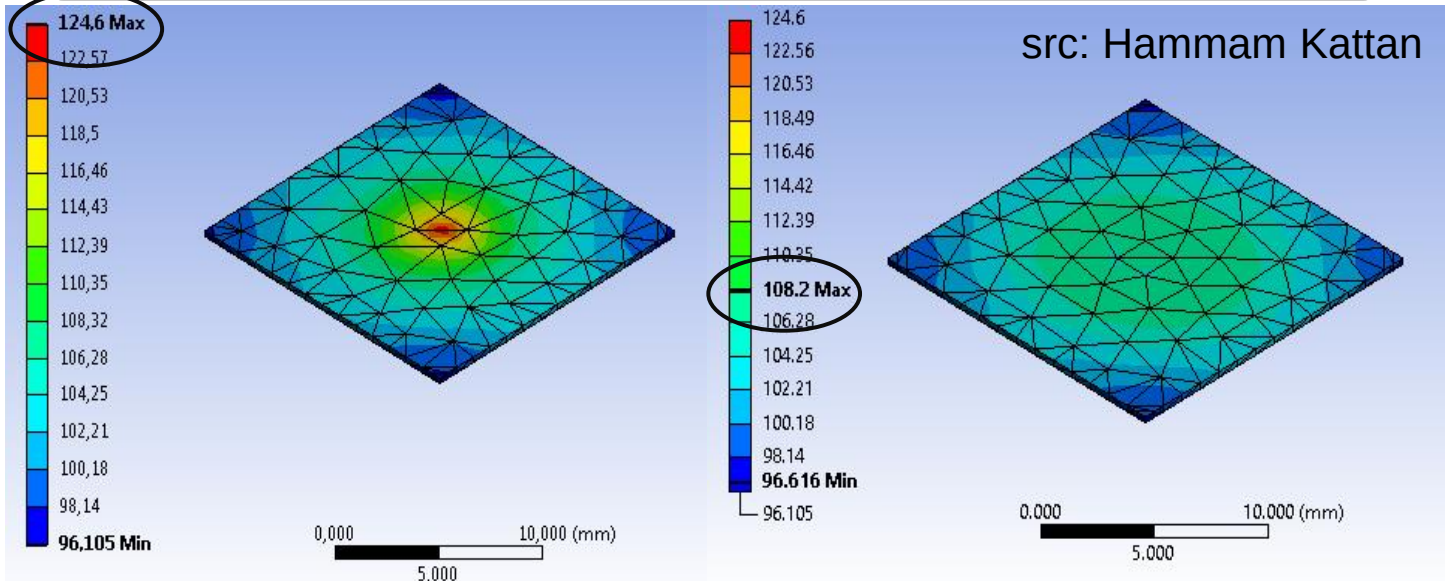
*Thermoelectric:*
*Peltier's effect* $Power \rightarrow \Delta T$


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**

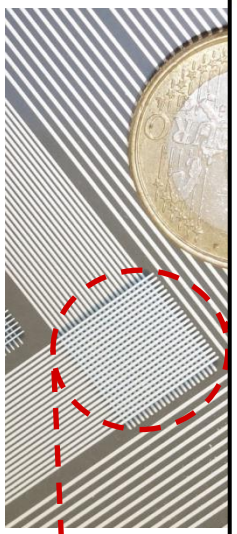



**Suppressing Heat Flux (Hot-Spot) of** $200\ W/cm^2$

# Intelligent Temperature Sensing

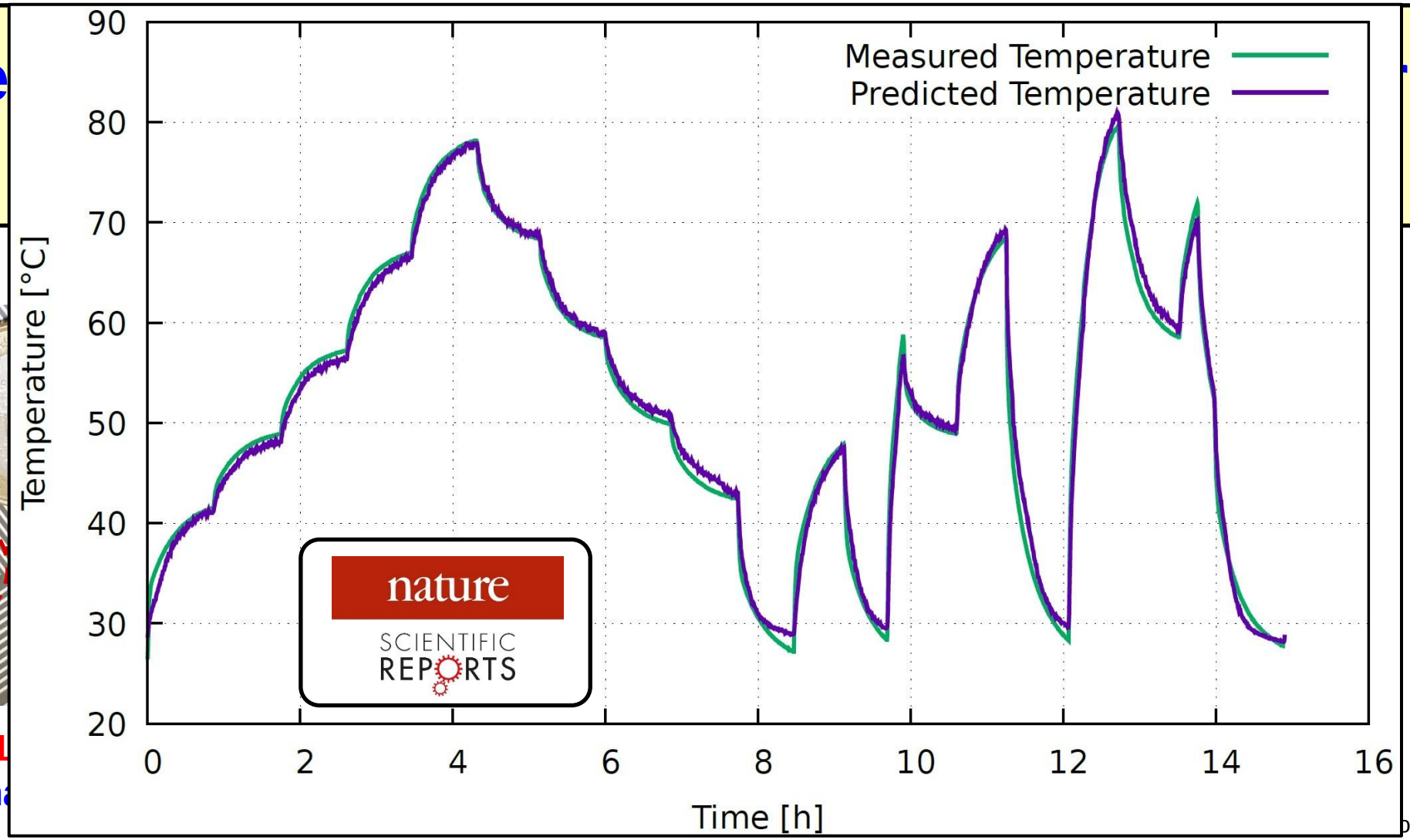
We

mal



21 x 21

Therma




nLab, U. Lemmer, KIT

# Collaborations











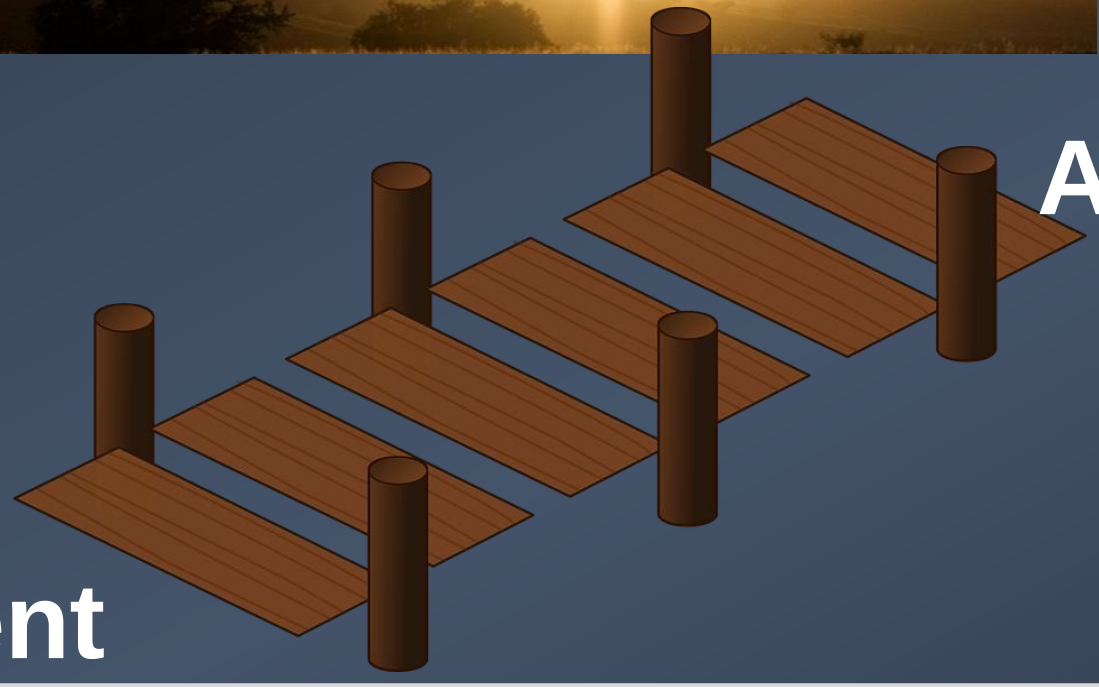



# Team ..without them it’s not possible

1. Dr.-Ing. Paul R. Genssler (post-doc)
2. Dr.-Ing. Victor van Santen (post-doc)
3. Dr. Narendra Dhakad (post-doc)
4. Dr. Anirban Kar (post-doc)
5. Simon Thomann
6. Rodion Novkin
7. Mahdi Benkhelifa
8. Jacky Wei-Ji Chao
9. Albi Mema
10. Matin Yousefzade
11. Sandy Abdelmalek
12. Sufia Shahin
13. Swati Deshwal
14. Konica Rathore
15. Zixu Wang
16. Stefanos Christopoulos
17. Johannes Mutter
18. Tarek Ashraf
19. Munazza Sayed
20. Karthik Pandaram

# On the Brink of a new Era in AI Chips


**AI / ML Algorithms**



**Thermal Management**

Technical University of Munich