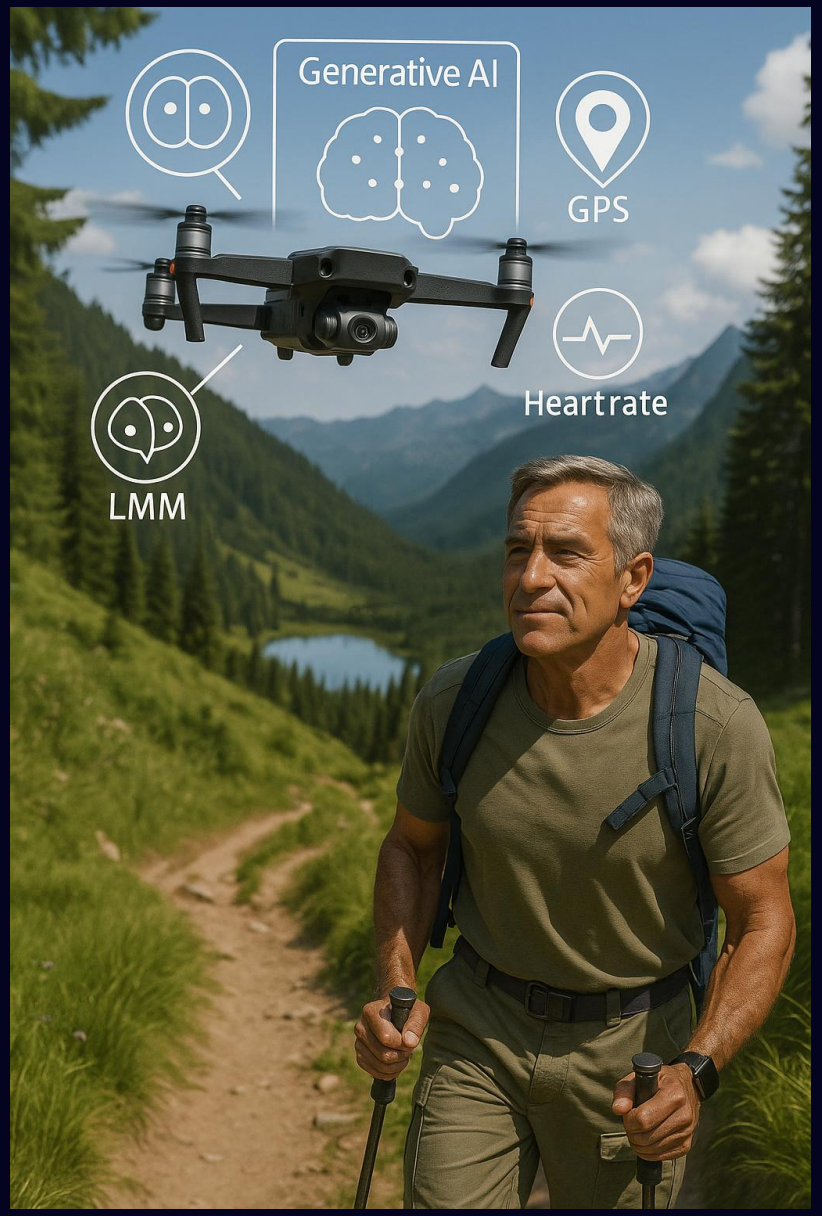
Edge AI:
Closer, Safer, Smarter

by Prof. Amrouch
Chair of AI Processor Design


AI
38°C

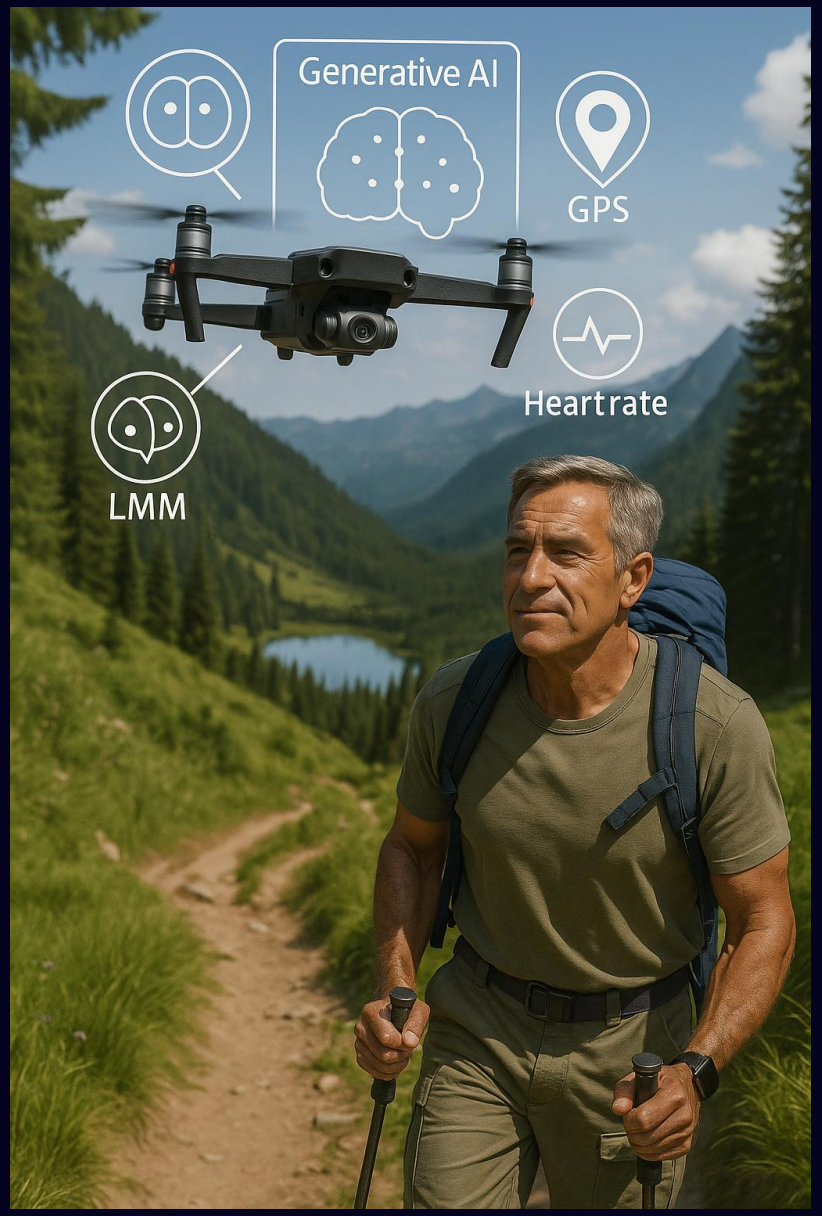

AI


AI


Generative AI
GPS
Heartrate
LMM


AI

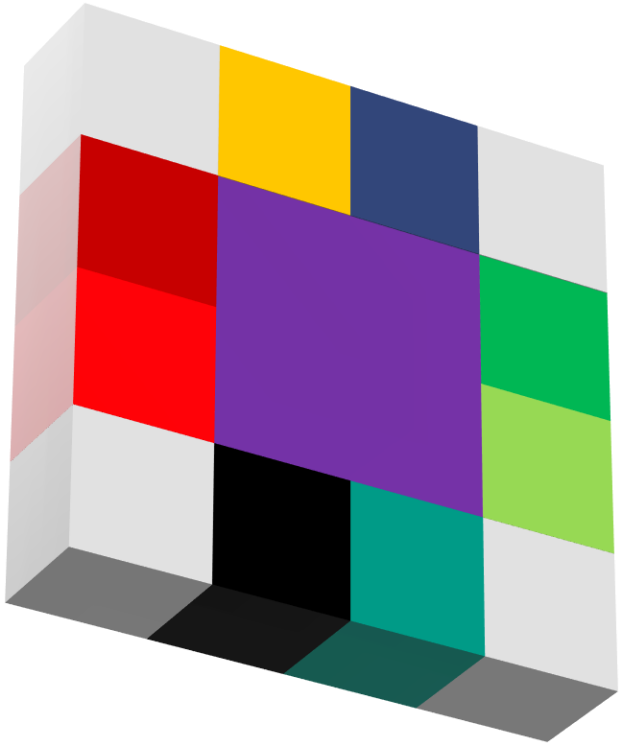

AI
38 °C



We Fuel the Cloud with Our Data

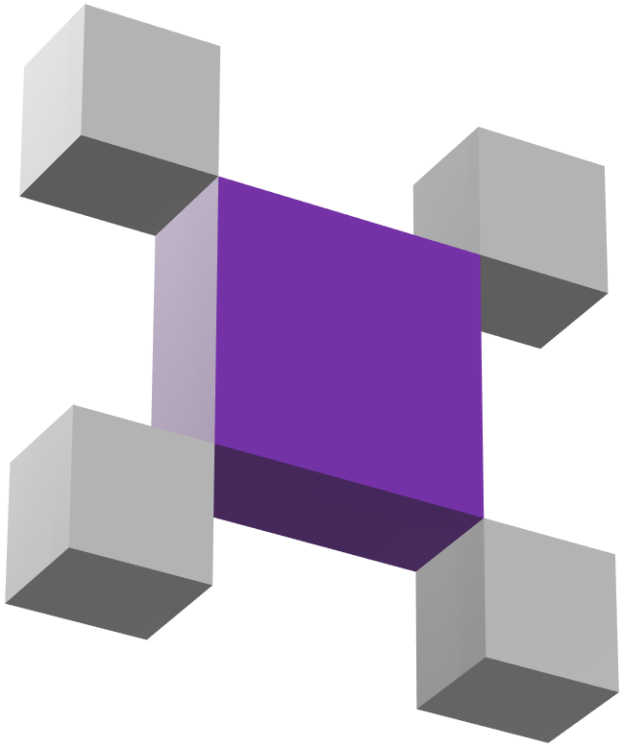
Data on Cloud — Serious Privacy Concern

Data on Cloud — Serious Privacy Concern

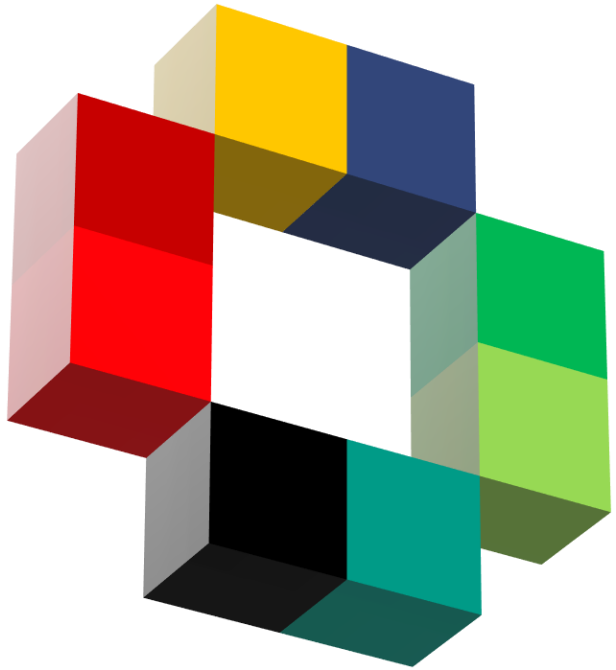


AI is split between the end-user and cloud

Cloud must not have the Full Picture!



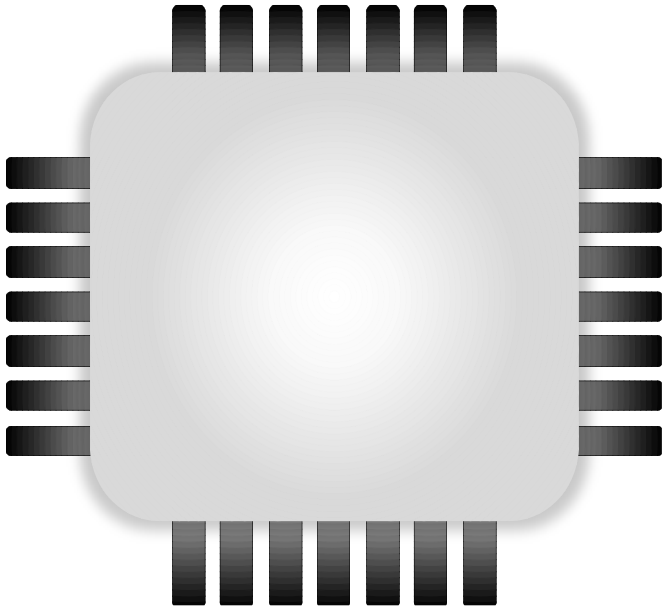
AI Cloud Outside



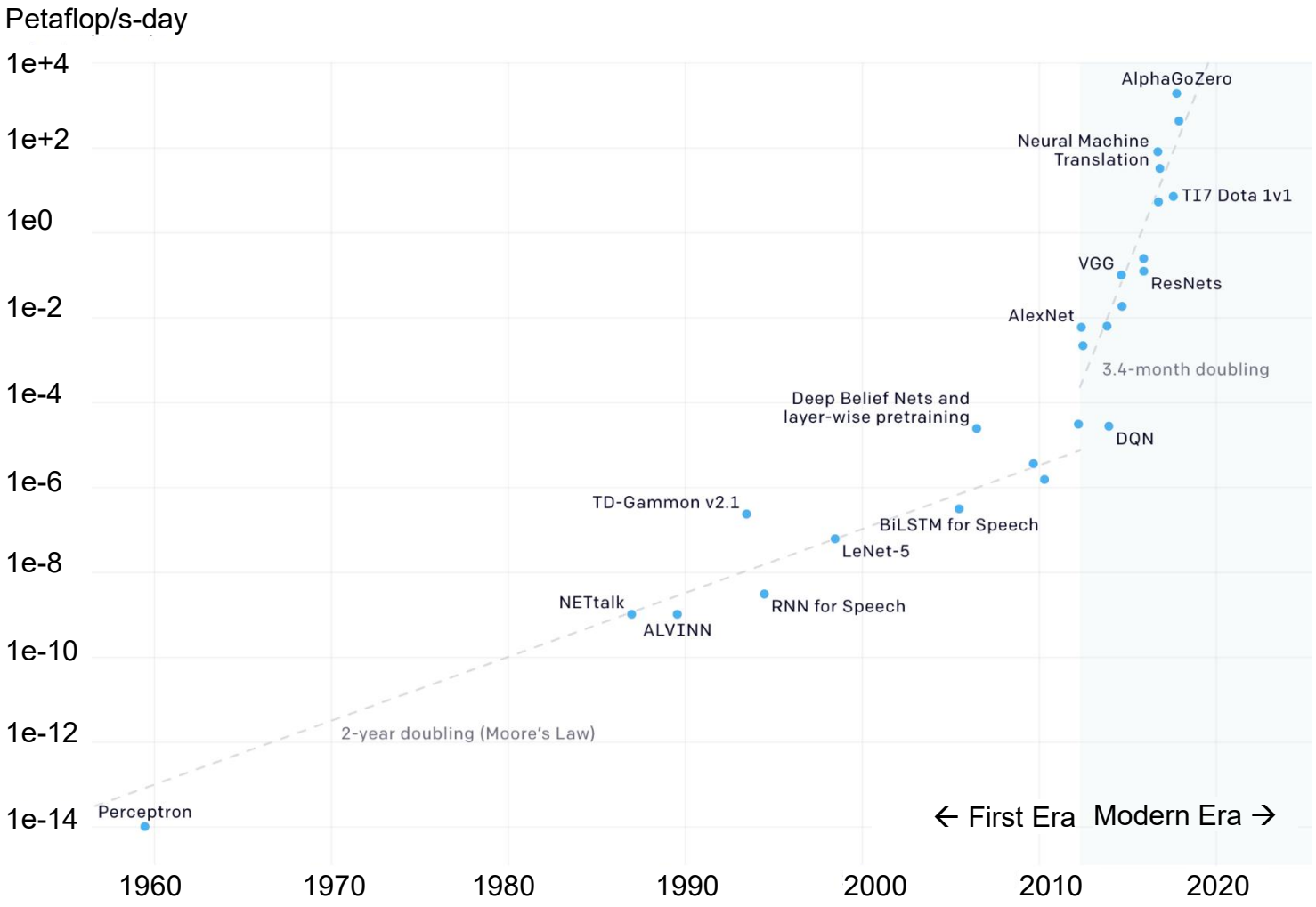
Edge AI with the User

Edge AI Chip for Training

It necessitates → Advanced Technology + Novel Algorithm + Novel Architecture



Classical AI → Massive Computing Demands

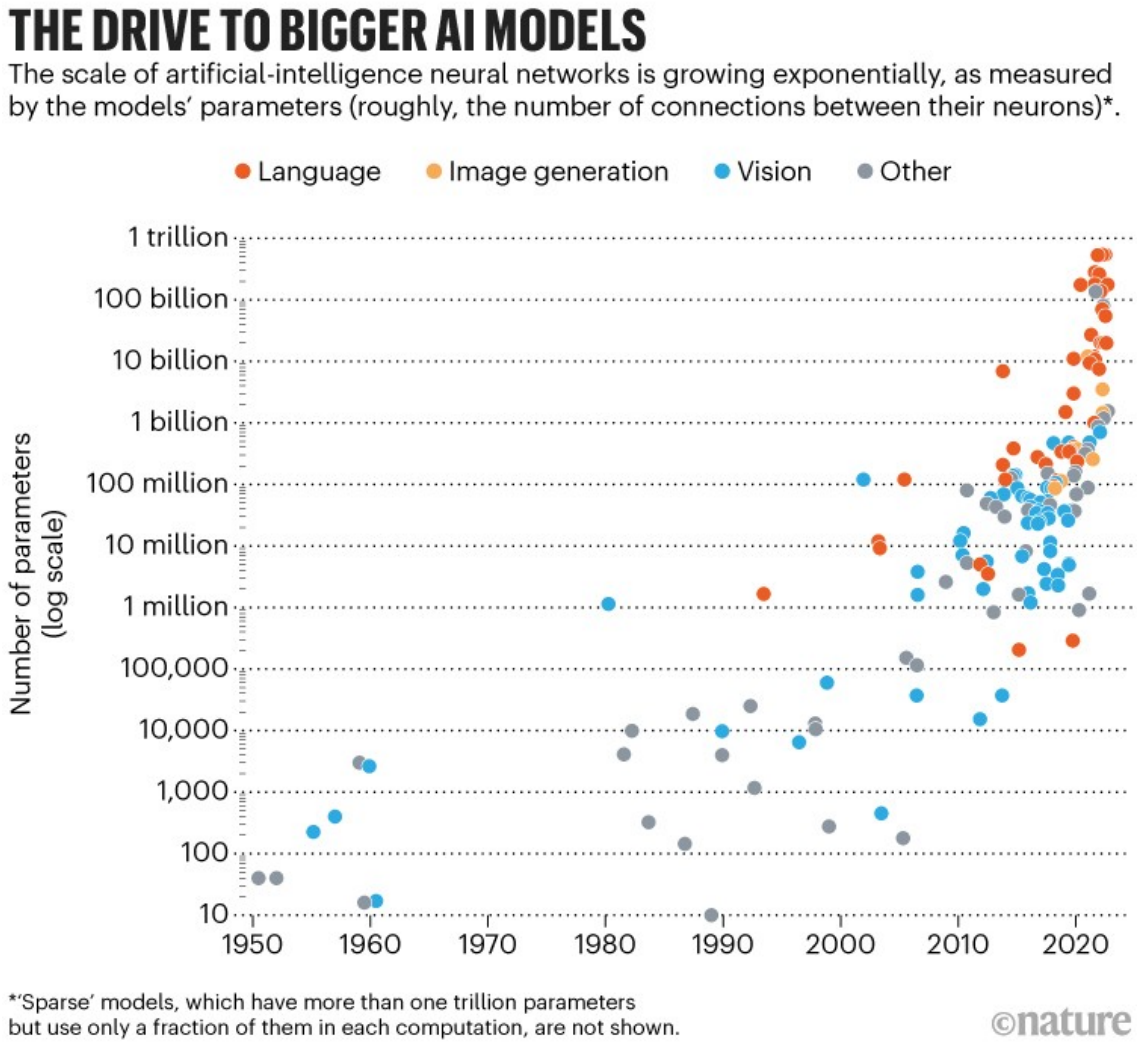


Computing demand

3.4 months doubling!

Source: https://openai.com

Classical AI Algorithms → Massive Memory Demands



More Efficient HW for AI

→ Larger and larger AI models

→ Memory Bottleneck!

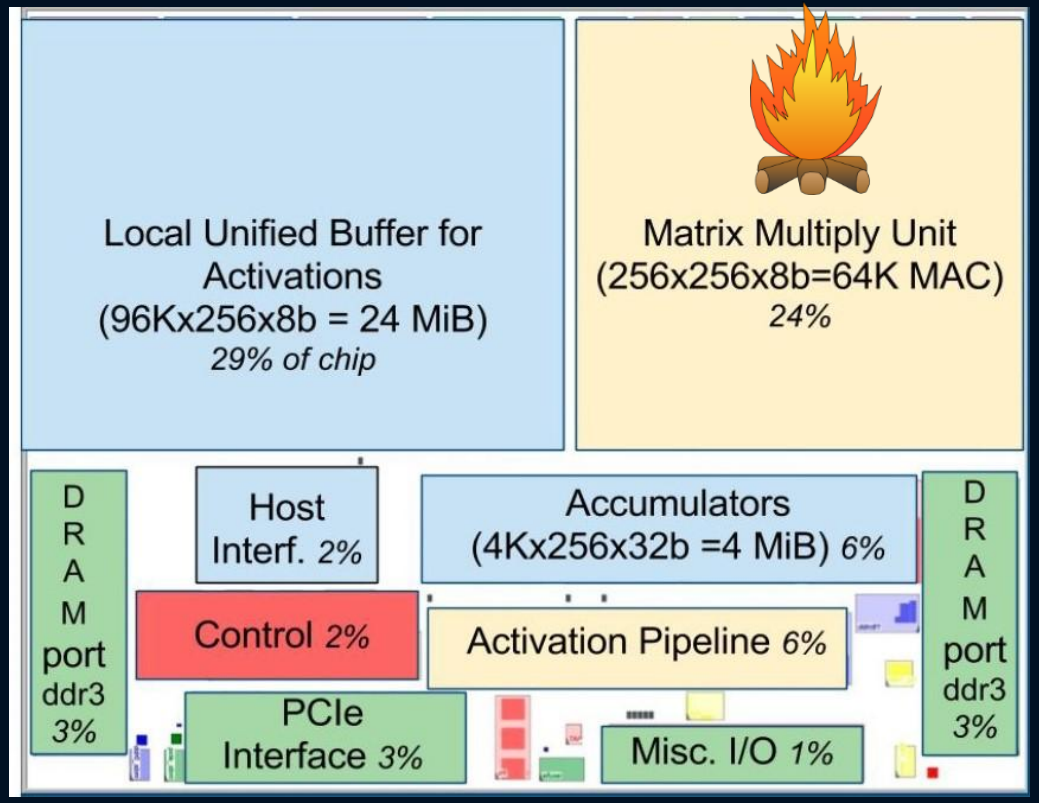
→ **Significant Efficiency Loss**

Massive Temperature!

Insufficient On-Chip Memory

↓

Massive Data to Move



AI Chip: Google TPU [ISCA'17]

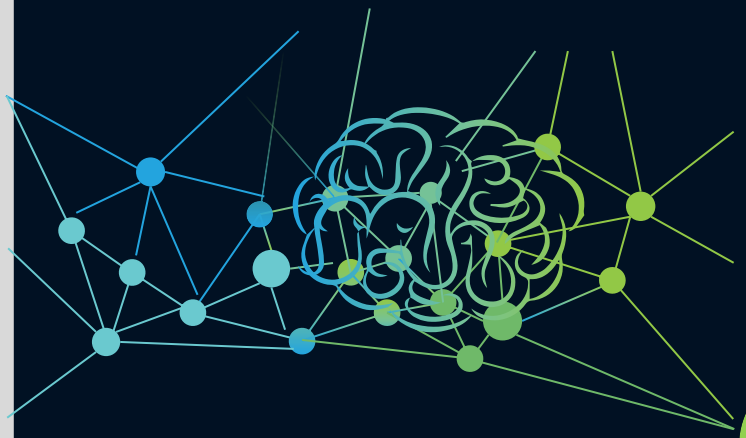
Massive Computation

↓

Very High Power Density

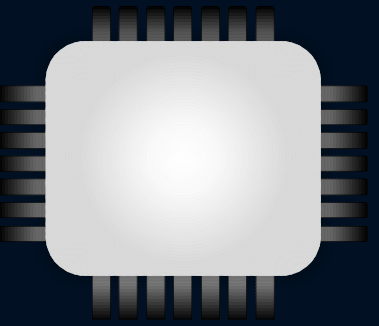
Excessive Temperature

DNN Alternative for Training at the Edge-AI



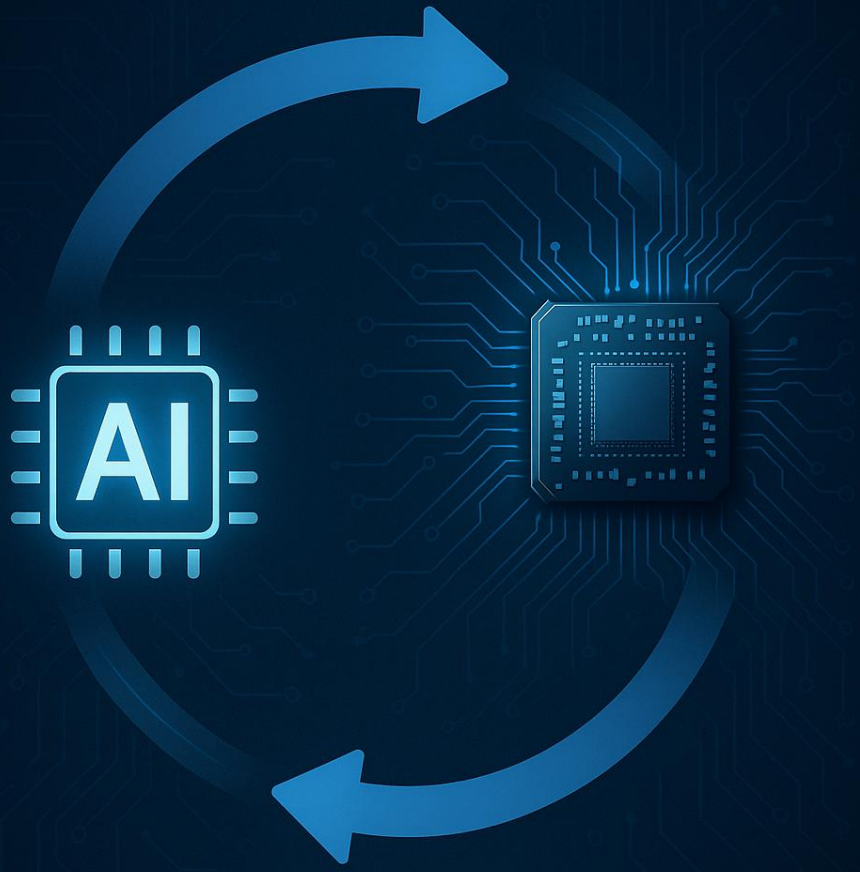
Brain-Inspired Computing

Domain Specific CPU



Edge AI Training TSMC 7nm

Alternative for Deep Neural Networks



Alternative for von-Neumann Architecture

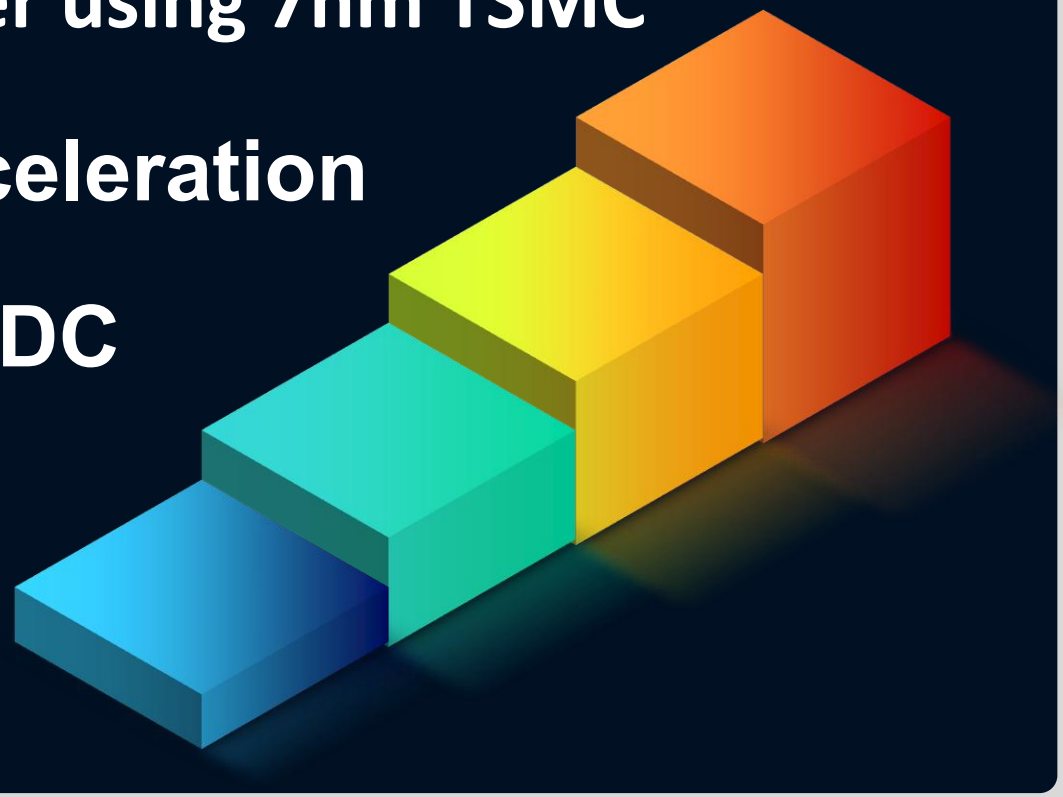
Novel Algorithm-Architecture Co-Design for Edge AI

Performance and Power using 7nm TSMC

Multi-Core RISC-V for HDC Acceleration

RISC-V Customization for HDC

Brain-Inspired Hyperdimensional Computing (HDC)



Brain-Inspired Hyperdimensional Computing

Example: Language classification

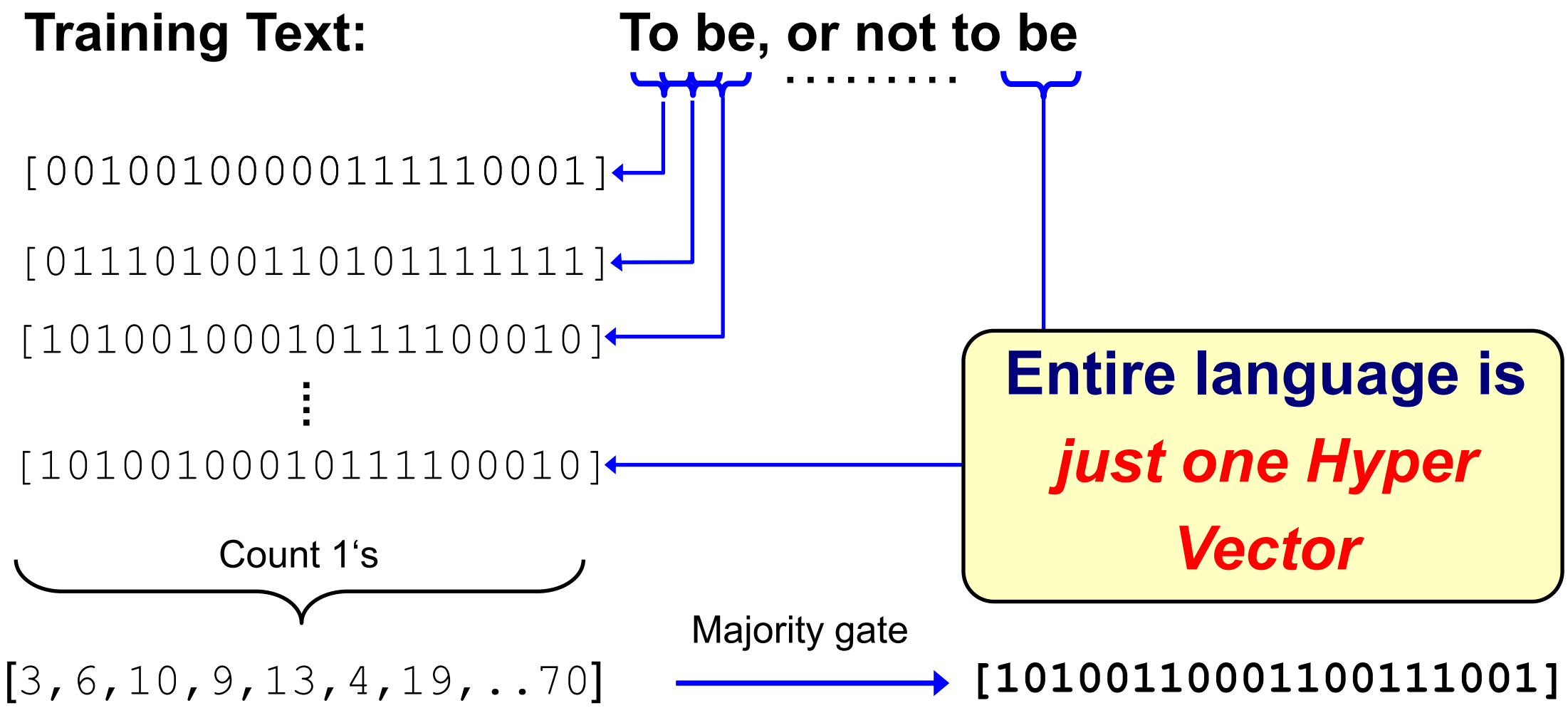
(1) Assign a random vector: VERY large (10k bits)

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

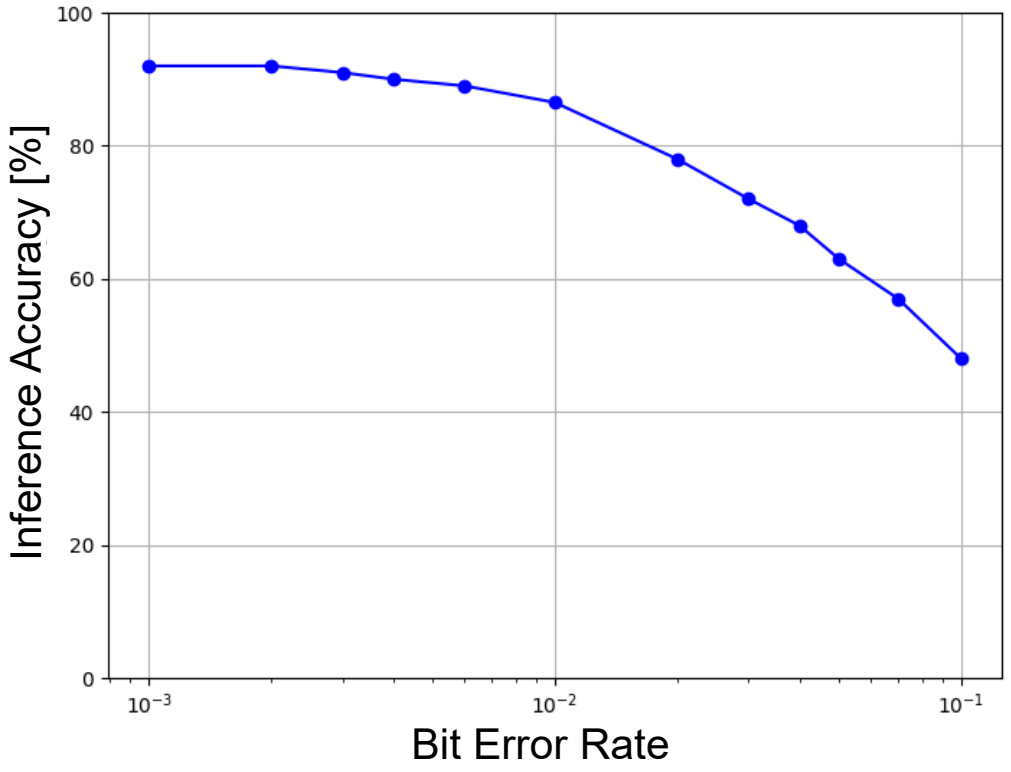
(2) Encoding with N-Grams using two simple operations: **XOR**, **Rotate**

“Hi” → [H] XOR [Rotate(i)]

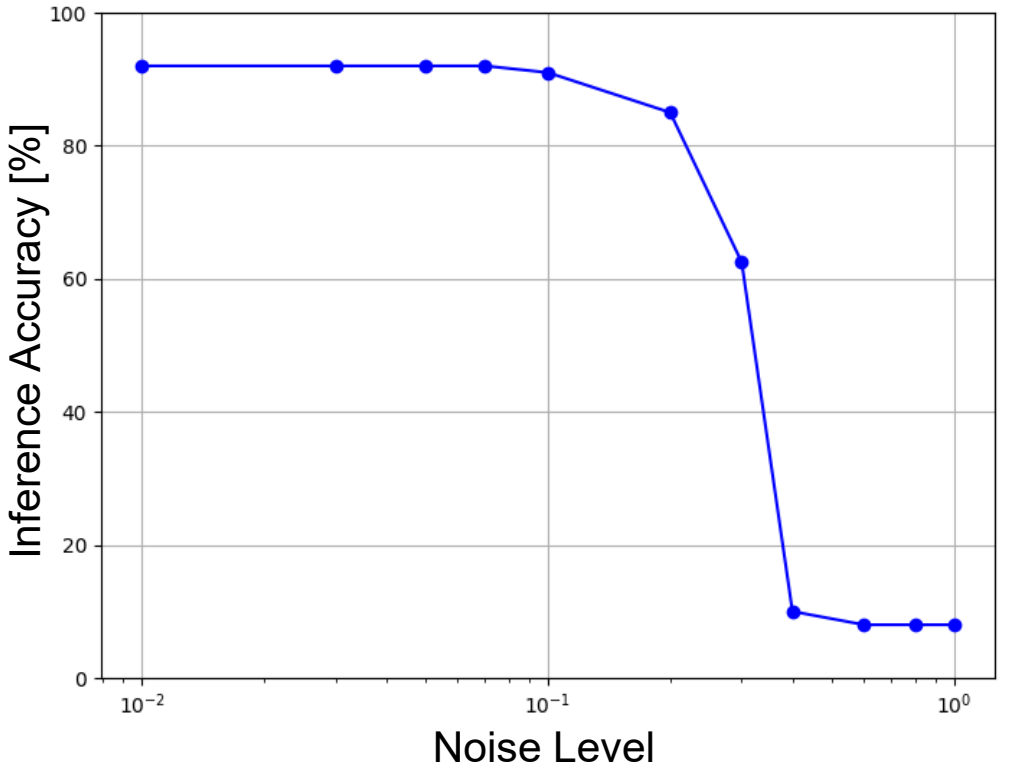
Brain-Inspired Hyperdimensional Computing



Robustness against HW Errors and Noise



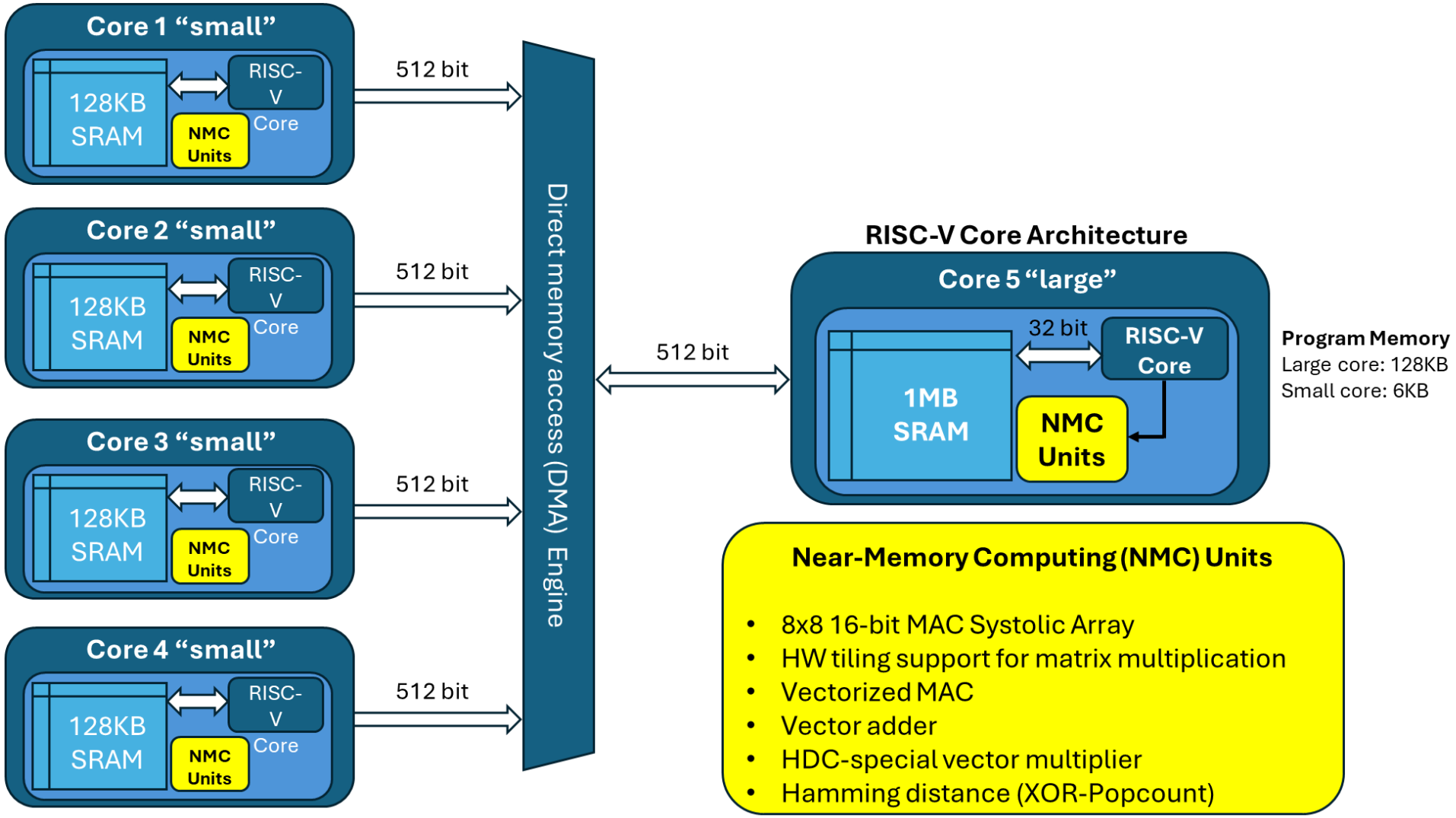
Errors injected in the underlying HW operations



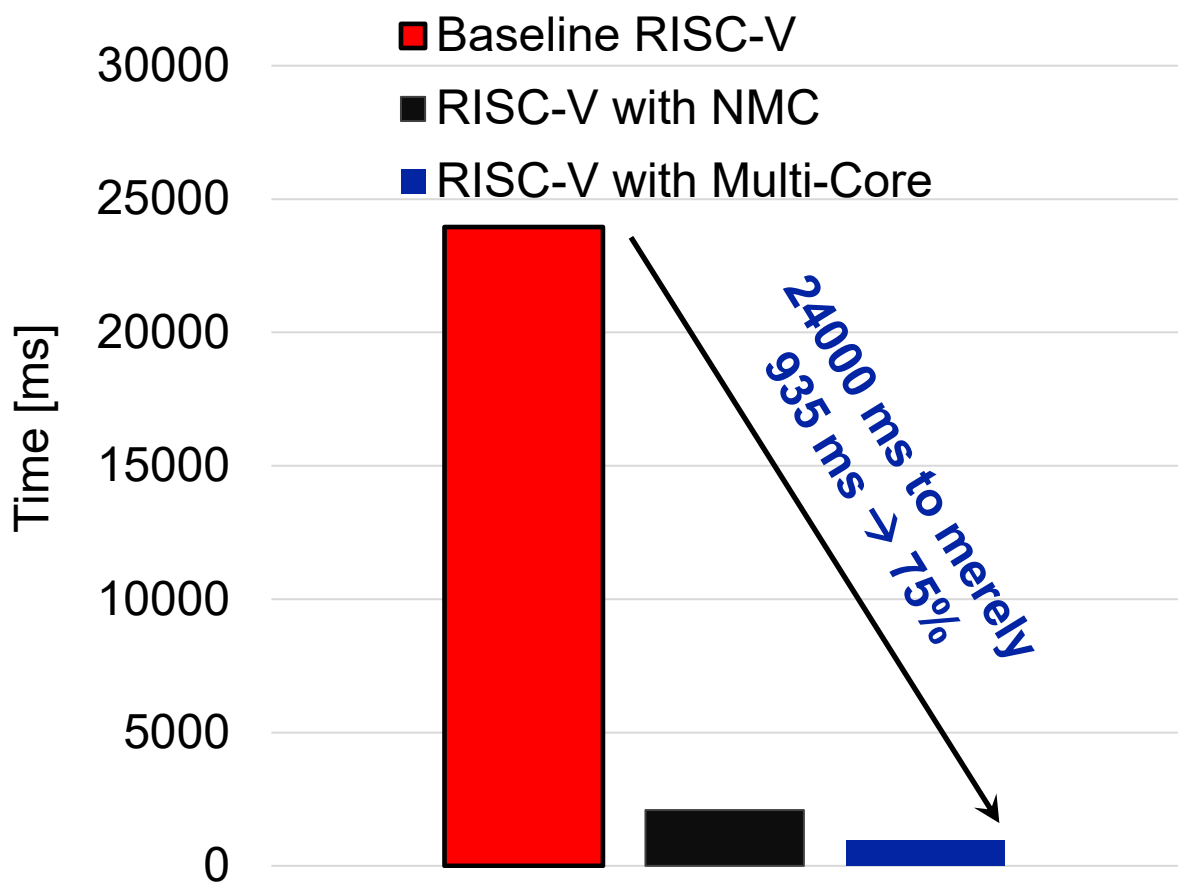
Errors injected in the input data → noise in samples

Edge-AI Processor: From Tradition to Customization

RISC-V: CPU Customization and Architecture



RISC-V Customization for Edge AI: Training

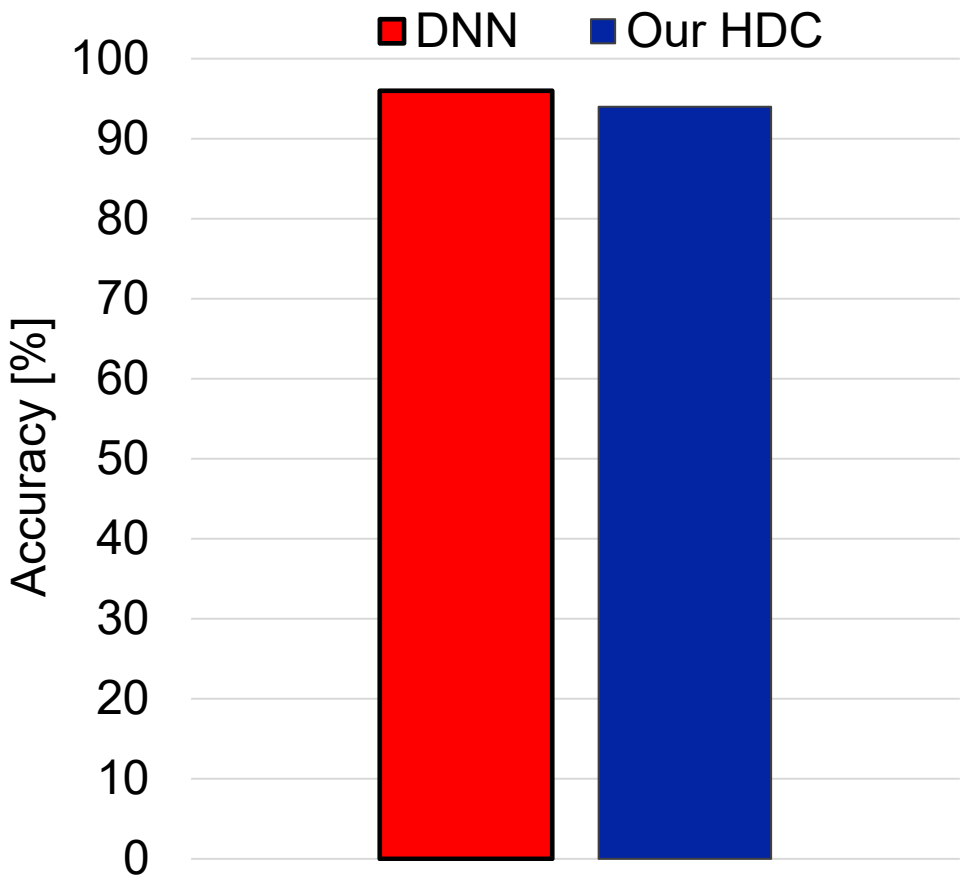


Human activity dataset

1000 Samples for Training

- **Training in < 1s**
- **Low Power: 23mW (full chip including IO power)**
- **Frequency: 200MHz**
- **Energy: 21.5 µJ**
- **Voltage: can be scaled to 0.5V → further improvements in power**

RISC-V Customization for Edge AI: Training



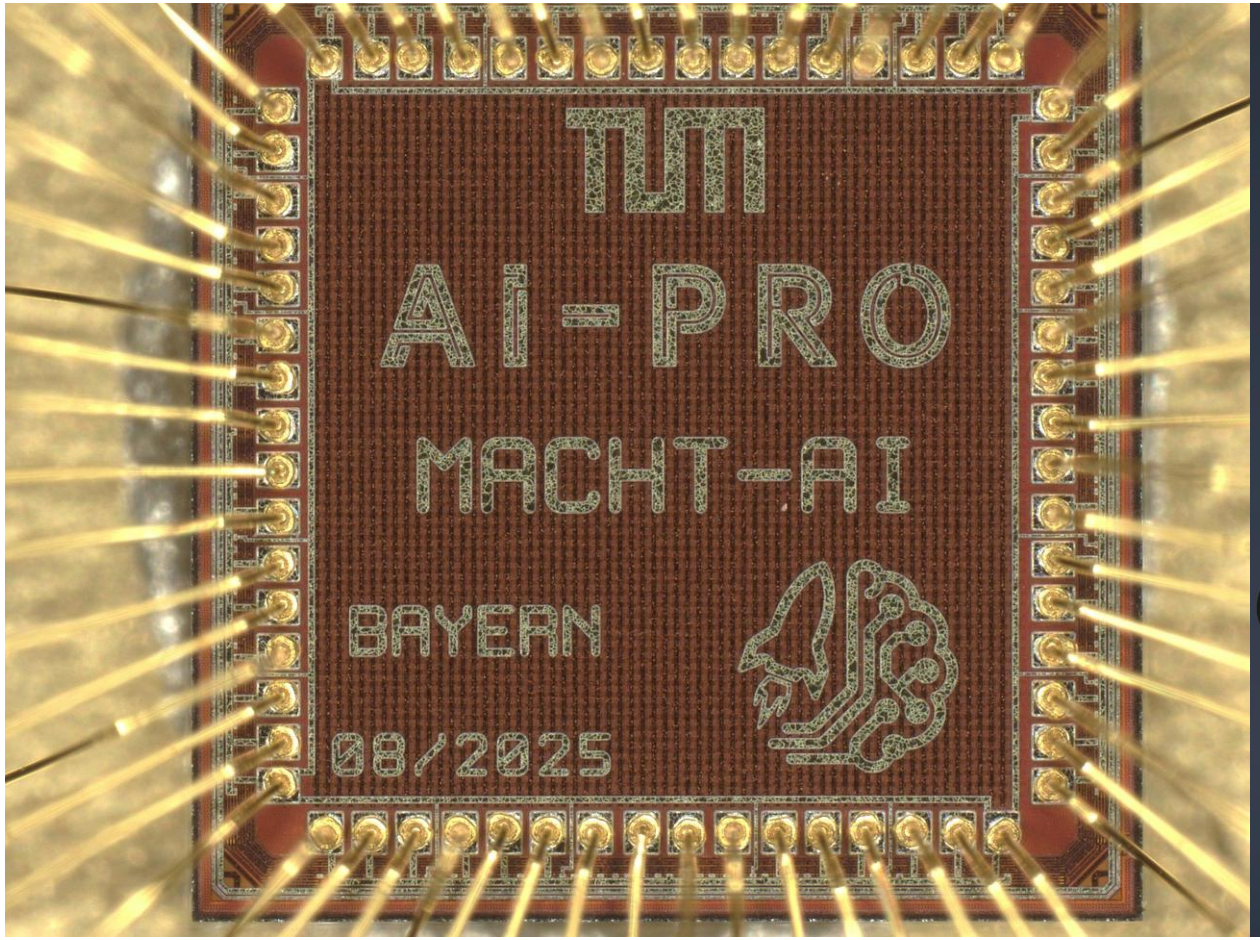
Human activity dataset

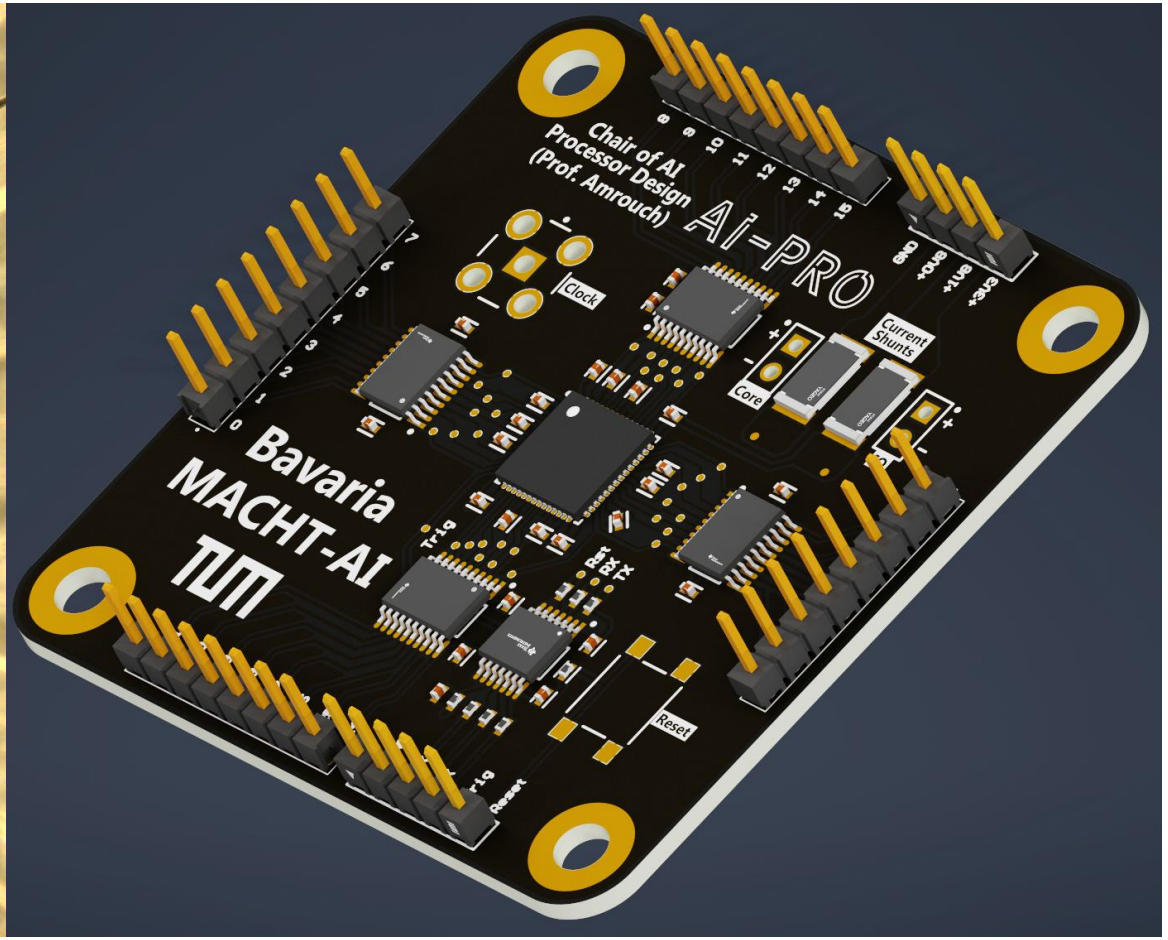
1000 Samples for Training

Similar Accuracy to exiting DNN

- **Our HDC Accuracy: 94%**
- **State-of-the-art DNN: 95% → but far from training at the edge**

First Chip from Uni. in EU in TSMC 7nm!



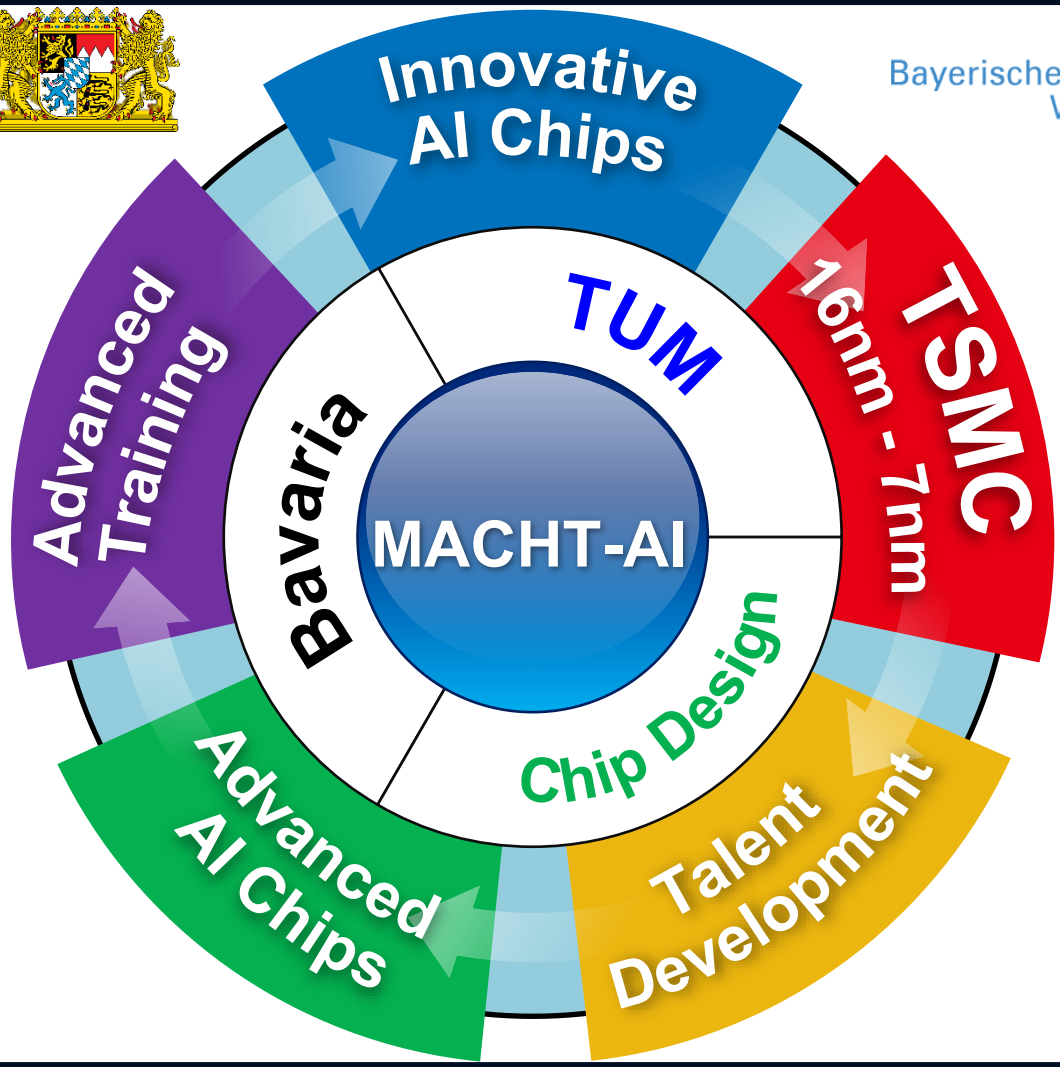


MACHT-AI Center

Munich Advanced-Technology Center for High-Tech AI Chips

Bayerisches Staatsministerium für Wirtschaft, Landesentwicklung und Energie

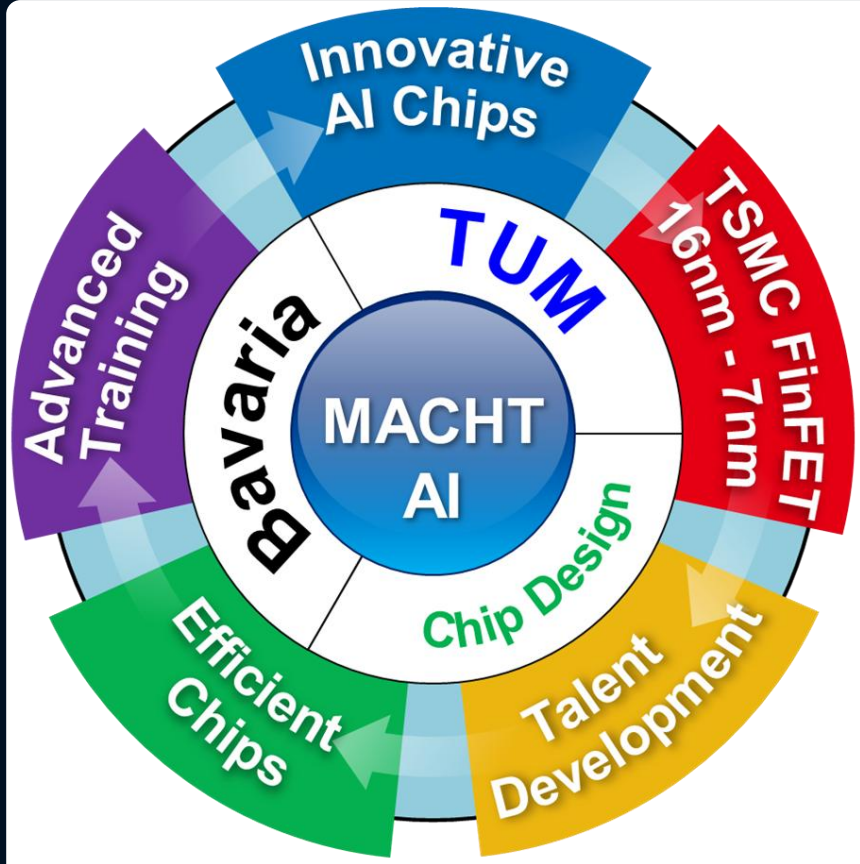
Bayerisches Staatsministerium für Wissenschaft und Kunst



Launched in September 2025

MACHT-AI Center

Munich Advanced-Technology Center for High-Tech AI Chips



TUM Chair of AI Processor Design (Prof. Amrouch)

MACHT-AI: Training and Research Center funded by the **Bavarian State Ministry of Economic** Affairs and **State Ministry of Science.**

Exclusive Focus on 16nm and 7nm TSMC.

Location: Technical University of Munich, Chair of AI Processor Design. Lead: H. Amrouch

Goals:

- **Training on the TSMC's FinFET Technologies**
- **Developing ultra-efficient AI chips in TSMC's 16nm, 7nm**

1st FinFET Training Workshop (TSMC FinFET)

1st Workshop: March 2026

2nd Workshop: July 2026

TSMC 16nm FinFET Academic Design (Digital Design)

Cost: Free

Length: One full week – only in person

Start Date: 09 March 2026

End Date: 13 March 2026

Time: 09:00 to 16:00 each day

Presenters: TUM Chair of AI Processor Design

Teaching Style: The course combines in-depth theoretical foundations in FinFET technology and digital chip design as well as extensive hands-on with practical training.

Attendees: The workshop will host 25-30 participants. In case of high demand, priority will be given to applicants who already have prior experience in chip design and who seek to deepen their expertise in advanced FinFET technology.

Overview: This is a comprehensive and intensive one-week course designed to provide participants with an in-depth training on TSMC 16 nm FinFET technology. The course focuses on digital design, covering the complete workflow from RTL to final tape-out. The training will be offered using the N16ADFP (Academic Design Foster Package) technology.

MACHT-AI: Training Workshops Plan

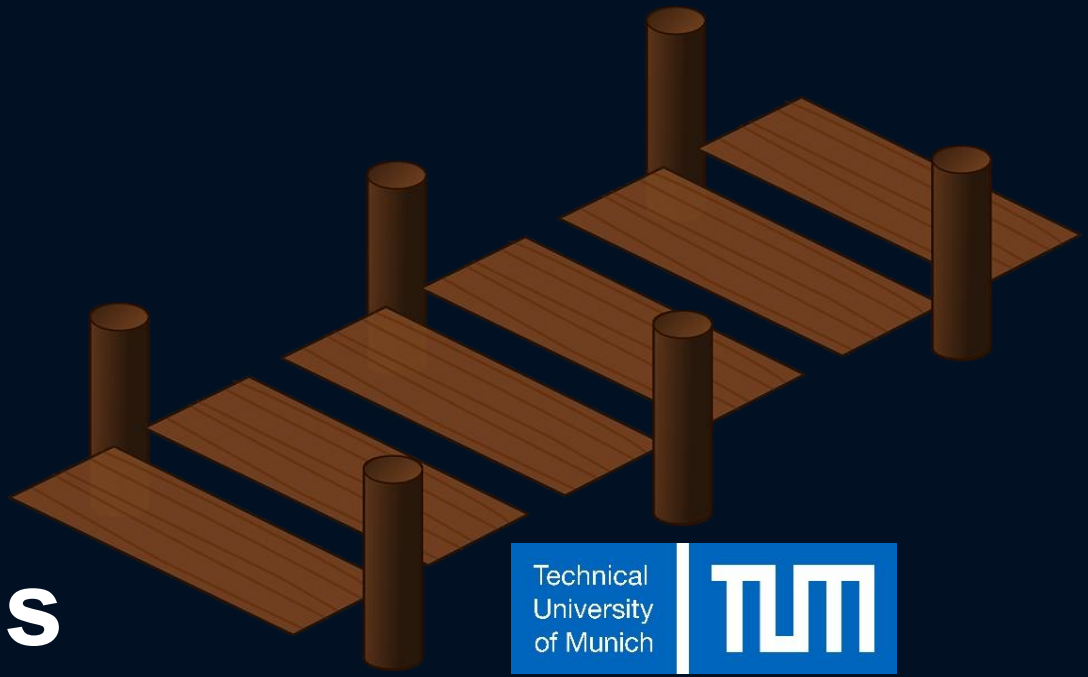
Plan and Vision for 5 Years

- ❑ Train Students > 500 students
- ❑ Tape-out > 10 Tape-out in 7nm and 16nm with priority for 7nm.
- ❑ Chiplet designs brining 16nm and 7nm tape-outs into one AI Chip

Edge AI for Training is Challenging

Processor Customization

Novel AI Algorithms



Advanced Technology

H. Amrouch @ Bavarian Semiconductor Congress, e-mail: amrouch@tum.de

Chair of AI Processor Design, Technical University of Munich