

Tailored Intelligence: The Art of Customized Processors for AI Acceleration

by Hussam Amrouch
Chair of AI Processor Design



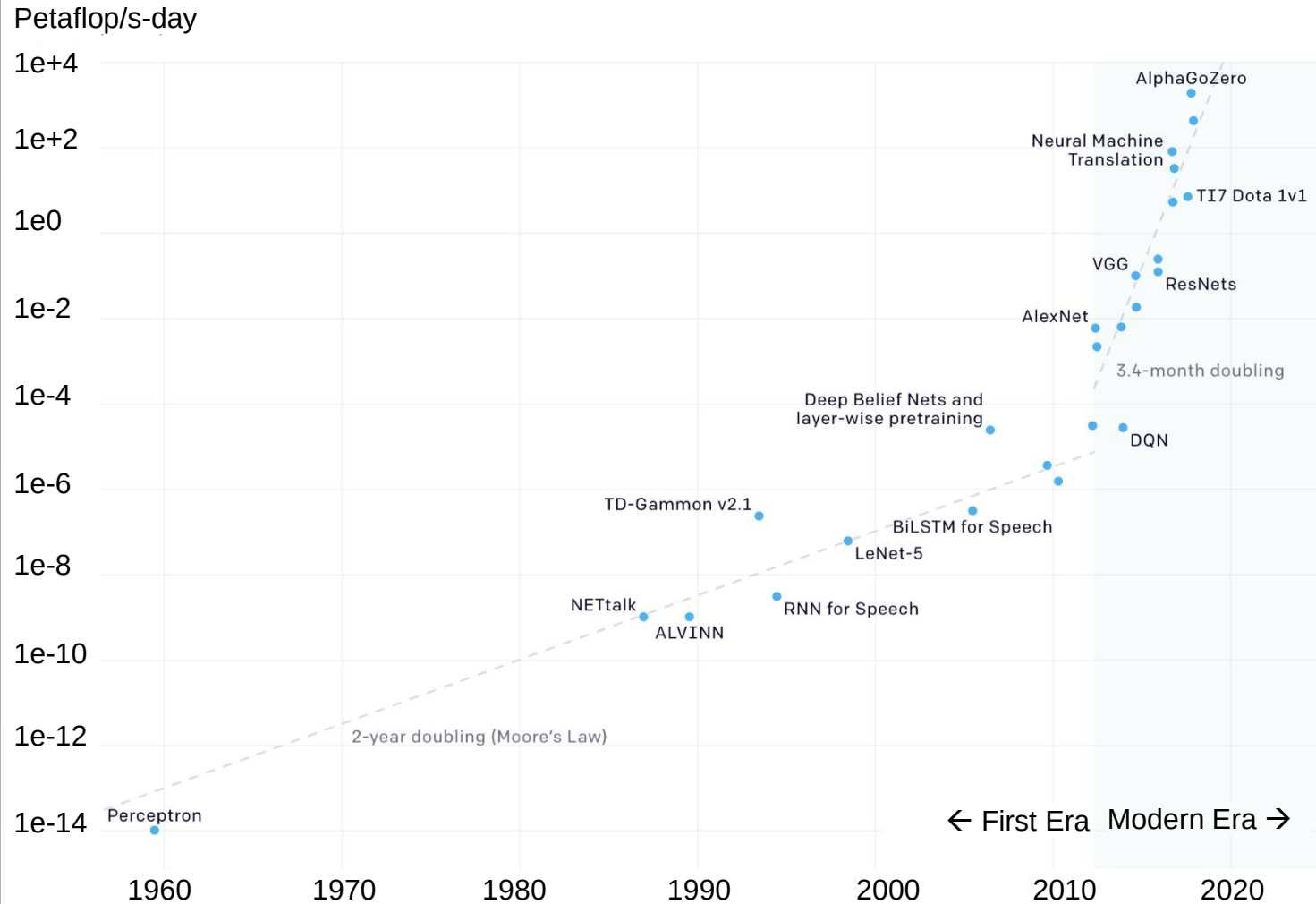
From Tradition to Innovation: The Art of AI Chips!



Journey with Edge AI



The Next Revolution: AI

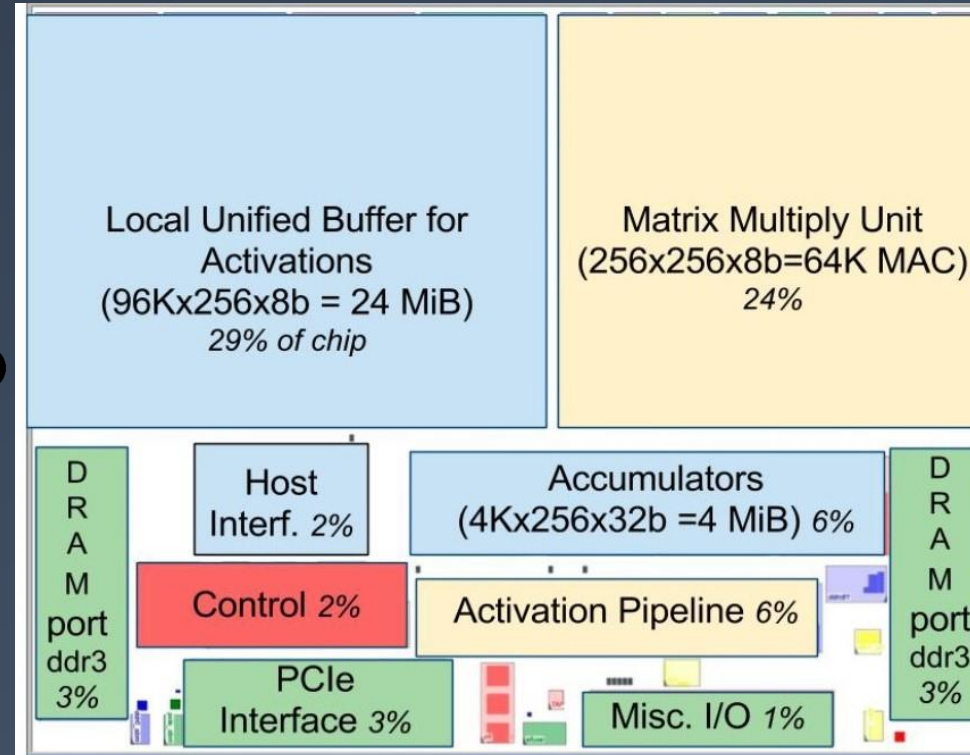
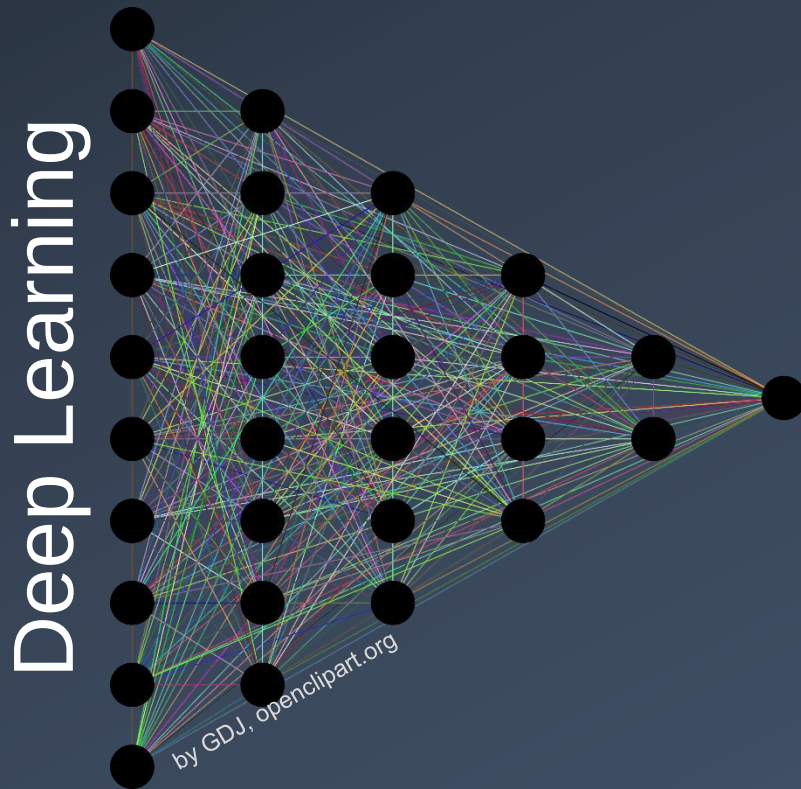


**Computing
demand**

**3.4 months
doubling!**

Source: <https://openai.com>

The Next Revolution: AI



Training BERT
DNN Google
TPUv3: 1.8min $\approx$

2048 GPUs +
512 CPUs

AI Chip: Google TPUv1 [ISCA'17]

src: <https://venturebeat.com/2020/07/29/google-claims-its-new-tpus-are-2-7-times-faster-than-the-previous-generation>

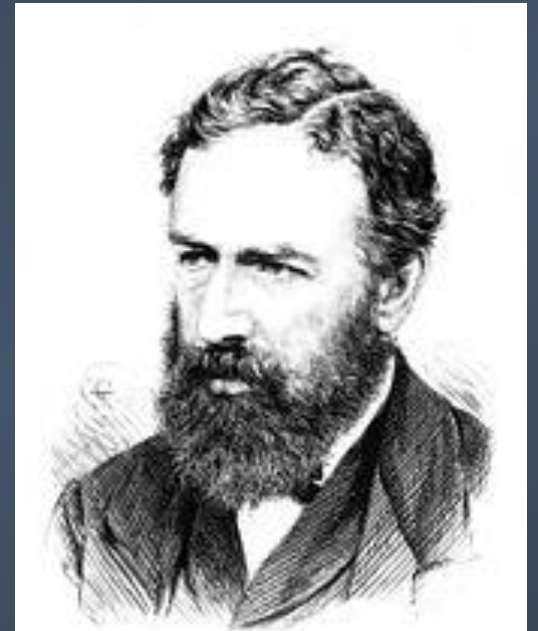
Could Efficiency be Dangerous?

Let's go back to 1865...

Jevons Paradox

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

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



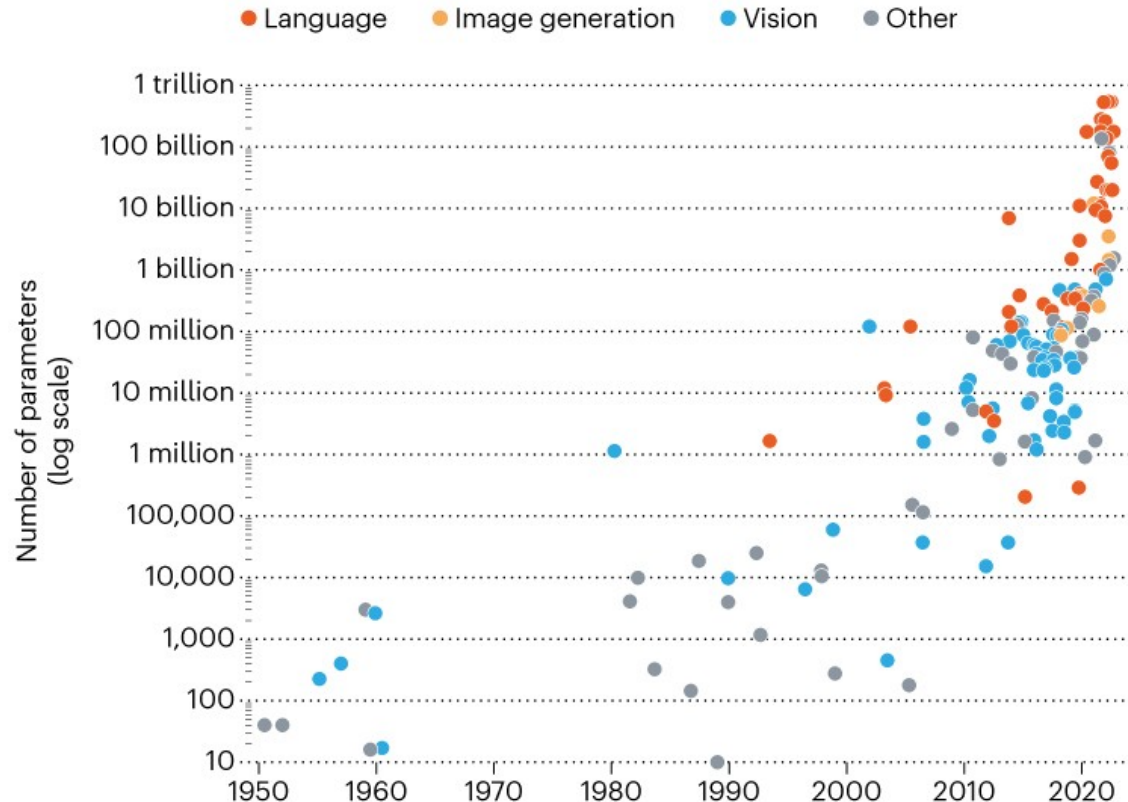
William Jevons

src: Wikipedia

AI Acceleration and Efficiency Paradox

THE DRIVE TO BIGGER AI MODELS

The scale of artificial-intelligence neural networks is growing exponentially, as measured by the models' parameters (roughly, the number of connections between their neurons)*.



*'Sparse' models, which have more than one trillion parameters but use only a fraction of them in each computation, are not shown.

©nature

More Efficient HW for AI

→ Larger and larger AI models

→ Memory Bottleneck!

Energy Crisis in AI Chips!

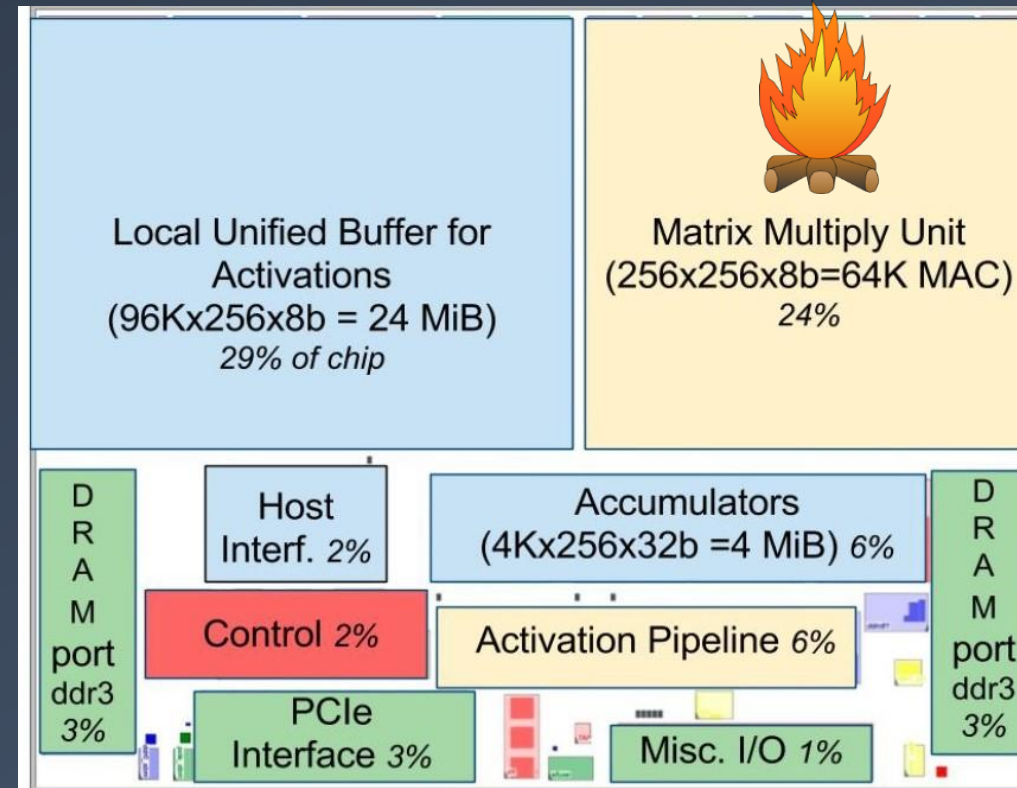
Insufficient
On-Chip Memory



Massive Data
to Move



von-Neumann
Bottleneck



AI Chip: Google TPU [ISCA'17]

Massive
Computation



Excessive
Heat



Expensive
Cooling

Where is the Problem?

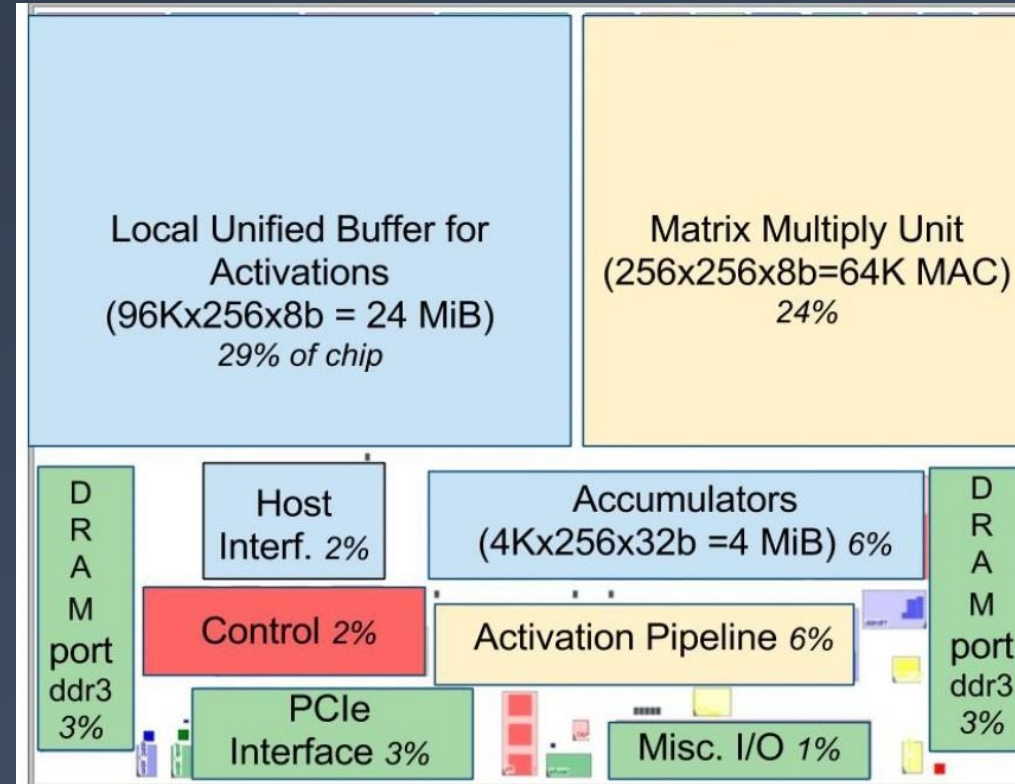
Insufficient
On-Chip Memory



Massive Data
to Move



Memory
Wall



Massive
Computation



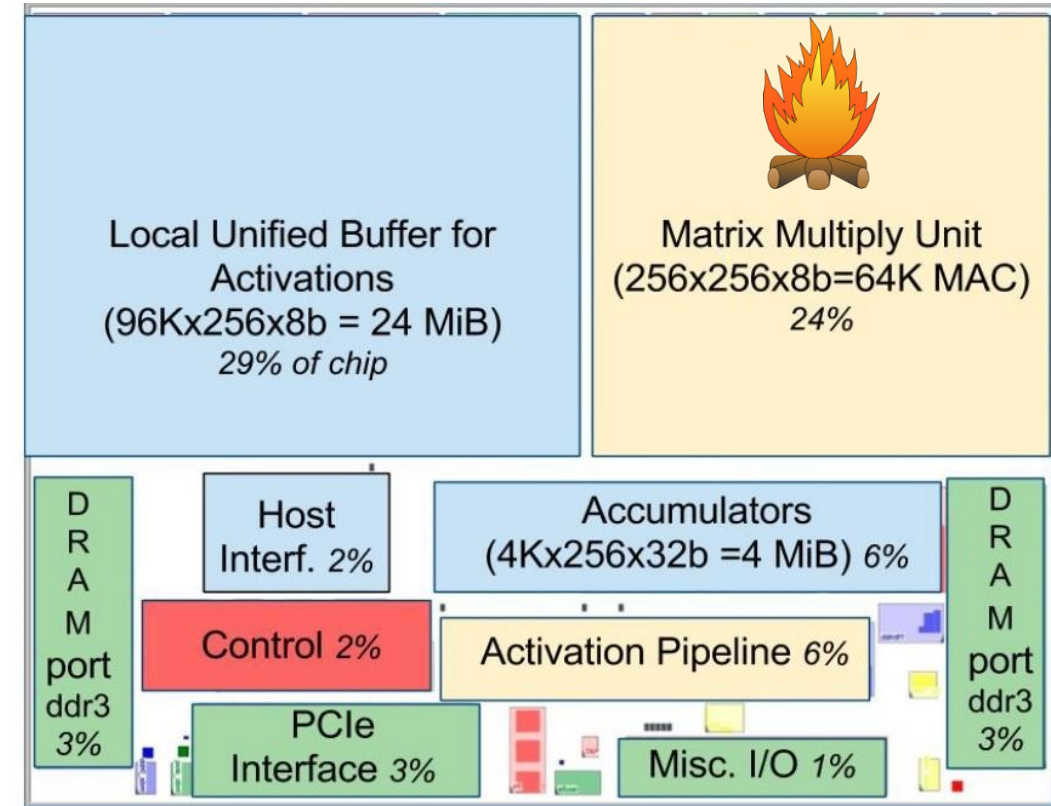
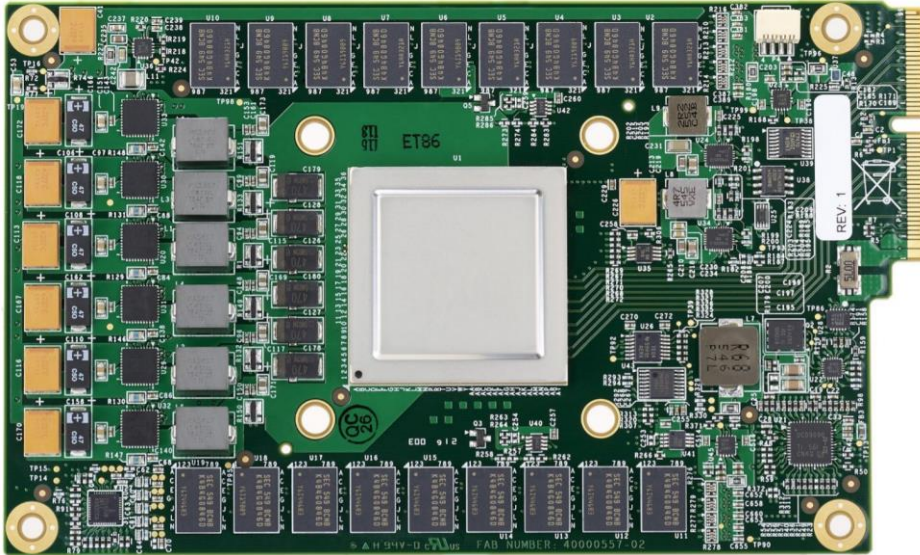
Excessive
Heat



Expensive
Cooling

Massive Energy
Cost

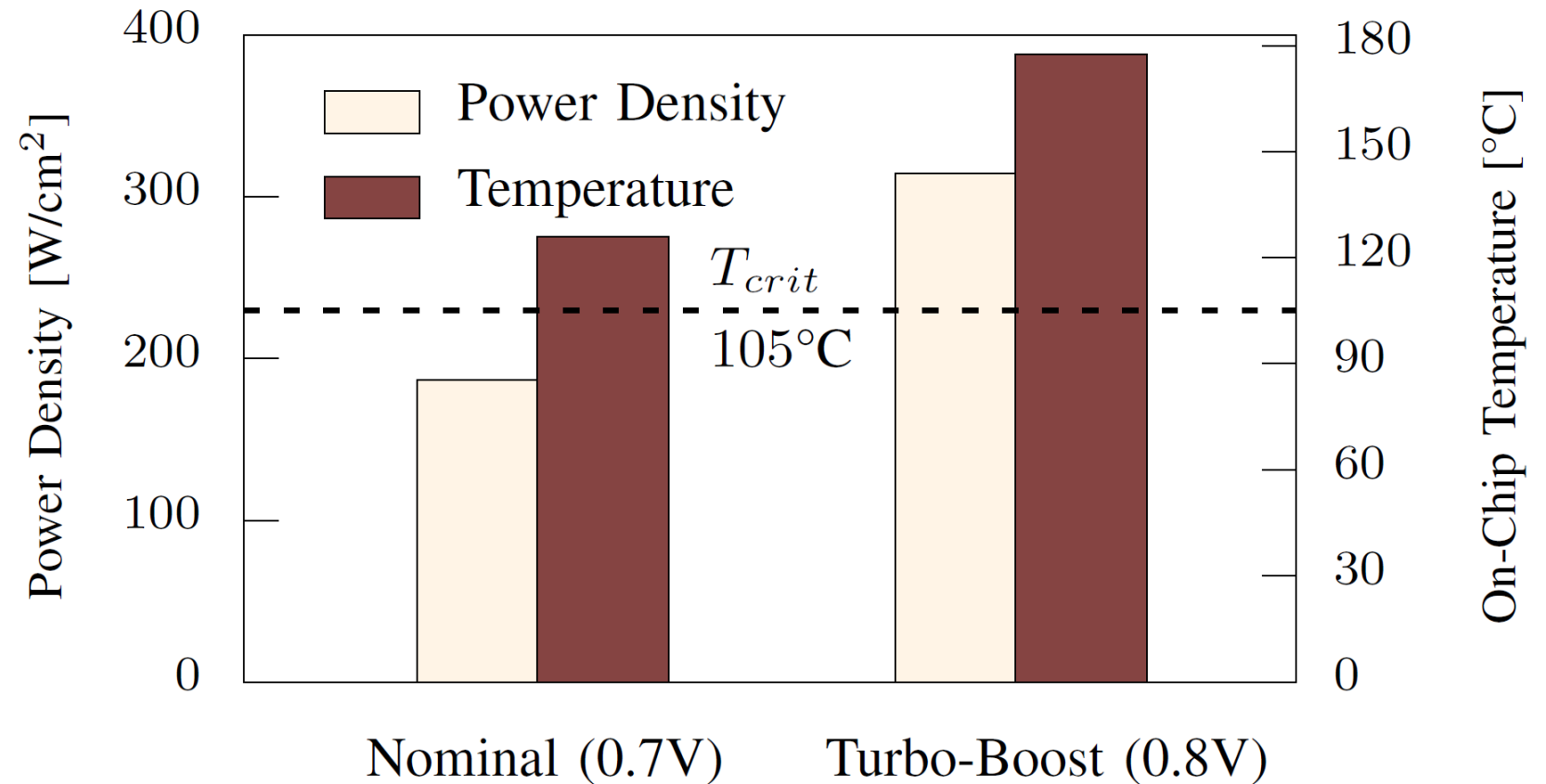
Deep Learning is REALLY Power Hungary!



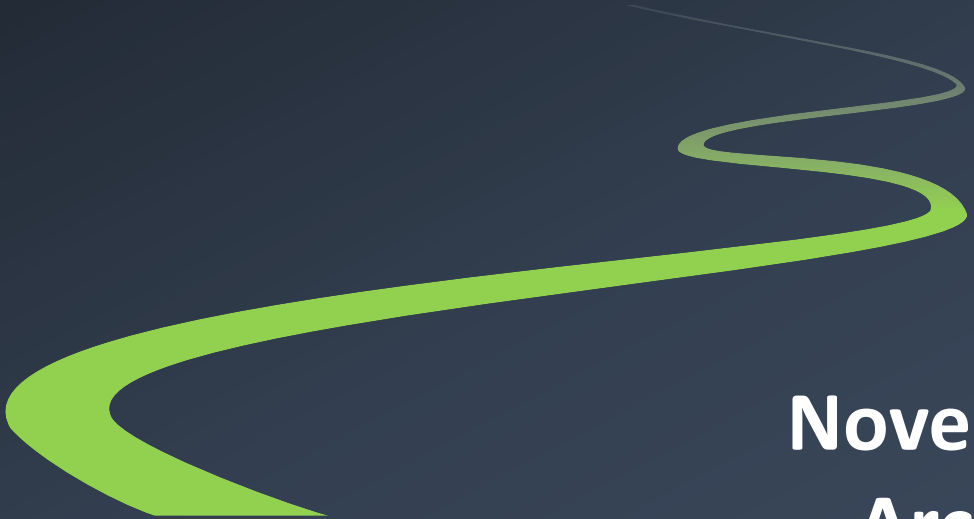
Google TPU [ISCA'17]

Deep Learning is REALLY Power Hungry!

Why not alternative algorithm to Deep Learning?



Amrouch [TCAD'20]



**Novel Computing
Architectures**

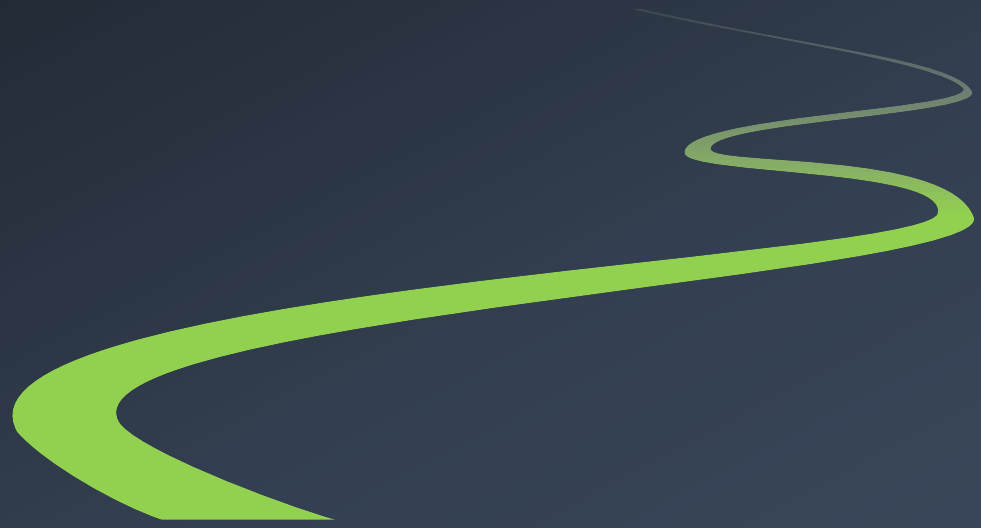
Novel Memories

Novel Processors

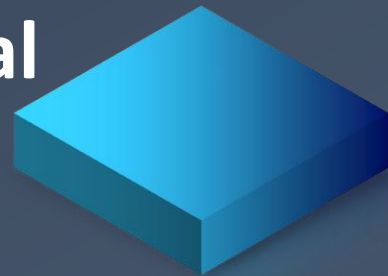
Novel Algorithms



**Ultra-Efficient Chips
Highly Tailed for AI**



Hyperdimensional Computing



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

Training Text:

[00100100000111110001]

[01110100110101111111]

[10100100010111100010]

⋮

[10100100010111100010]

Count 1's

[3, 6, 10, 9, 13, 4, 19, .. 70]

To be, or not to be

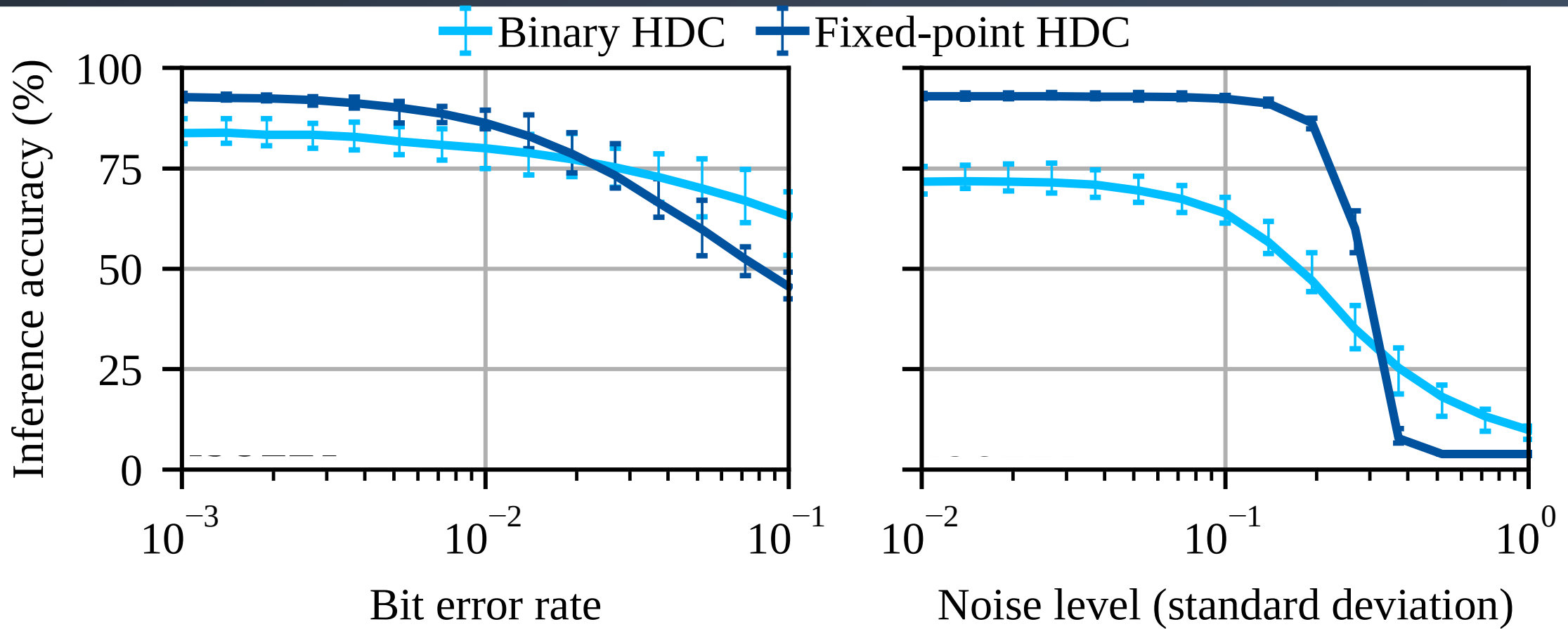
.....

Entire language is
*just one Hyper
Vector*

Majority gate

[10100110001100111001]

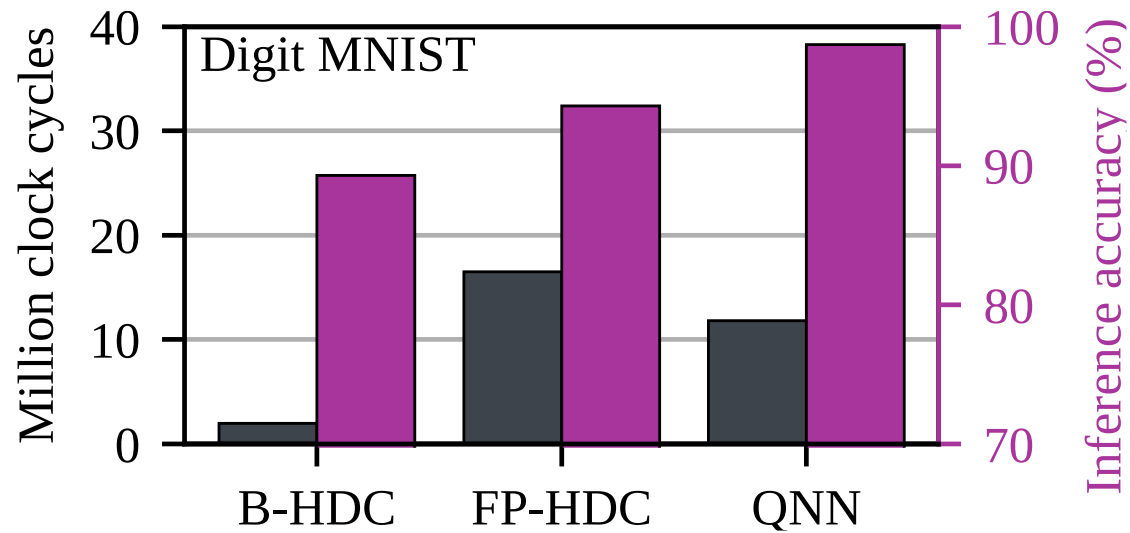
Robustness against HW Errors and Noise



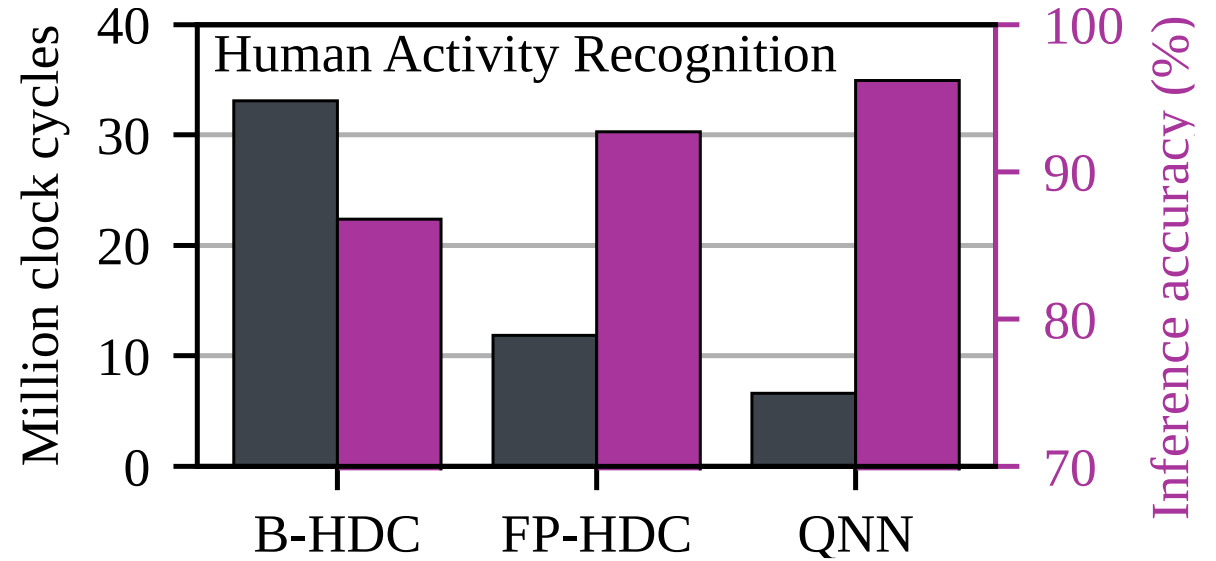
Errors injected in the
underlying HW operations

Errors injected in the
input data → noise in samples

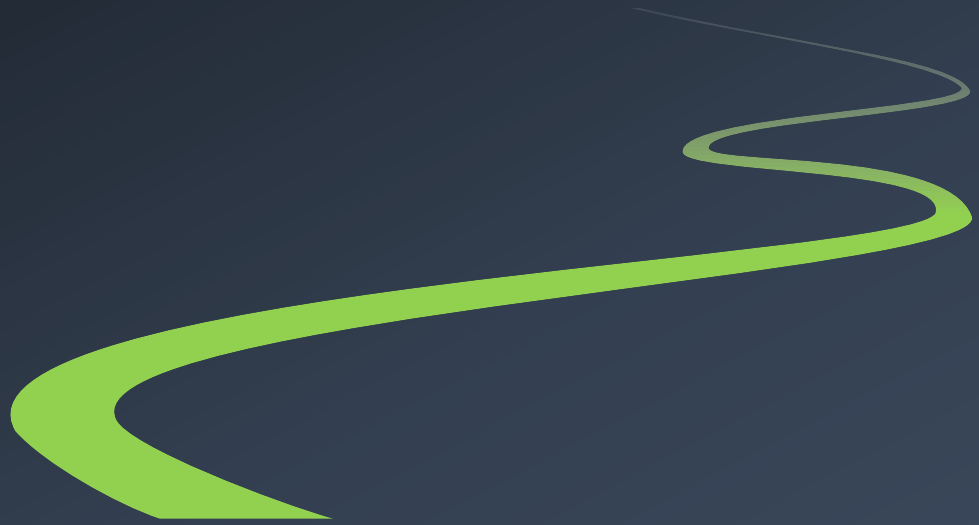
HDC vs. QNN: Performance / Accuracy



Binary HDC has a superior speed in image classification

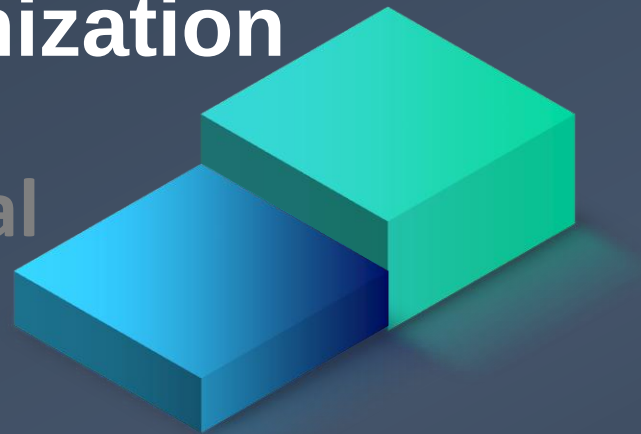


Both QNN, HDC employ MACs, but QNN is faster than Fix-point HDC

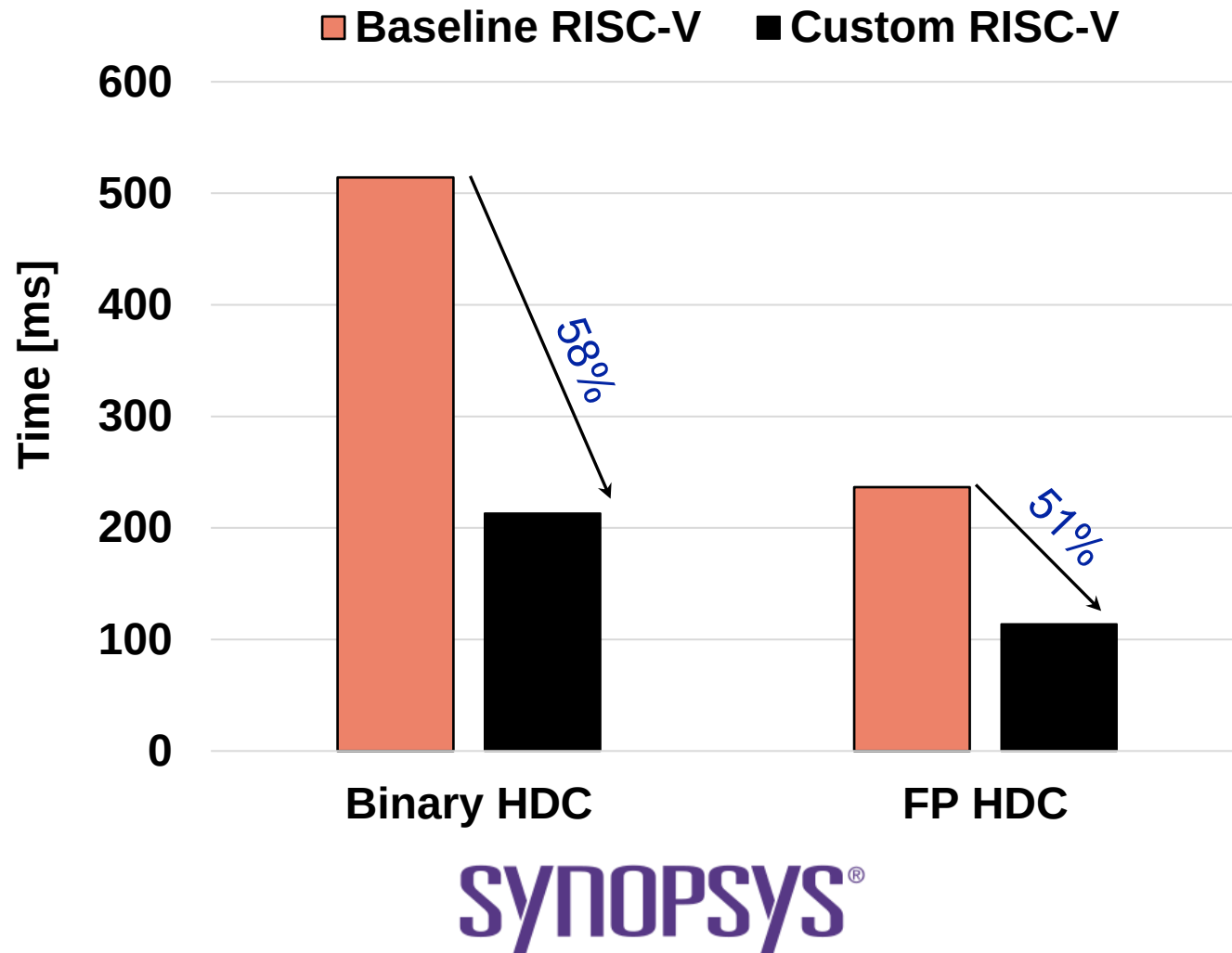


RISC-V Customization

Hyperdimensional
Computing



RISC-V Customization for Edge AI: Training



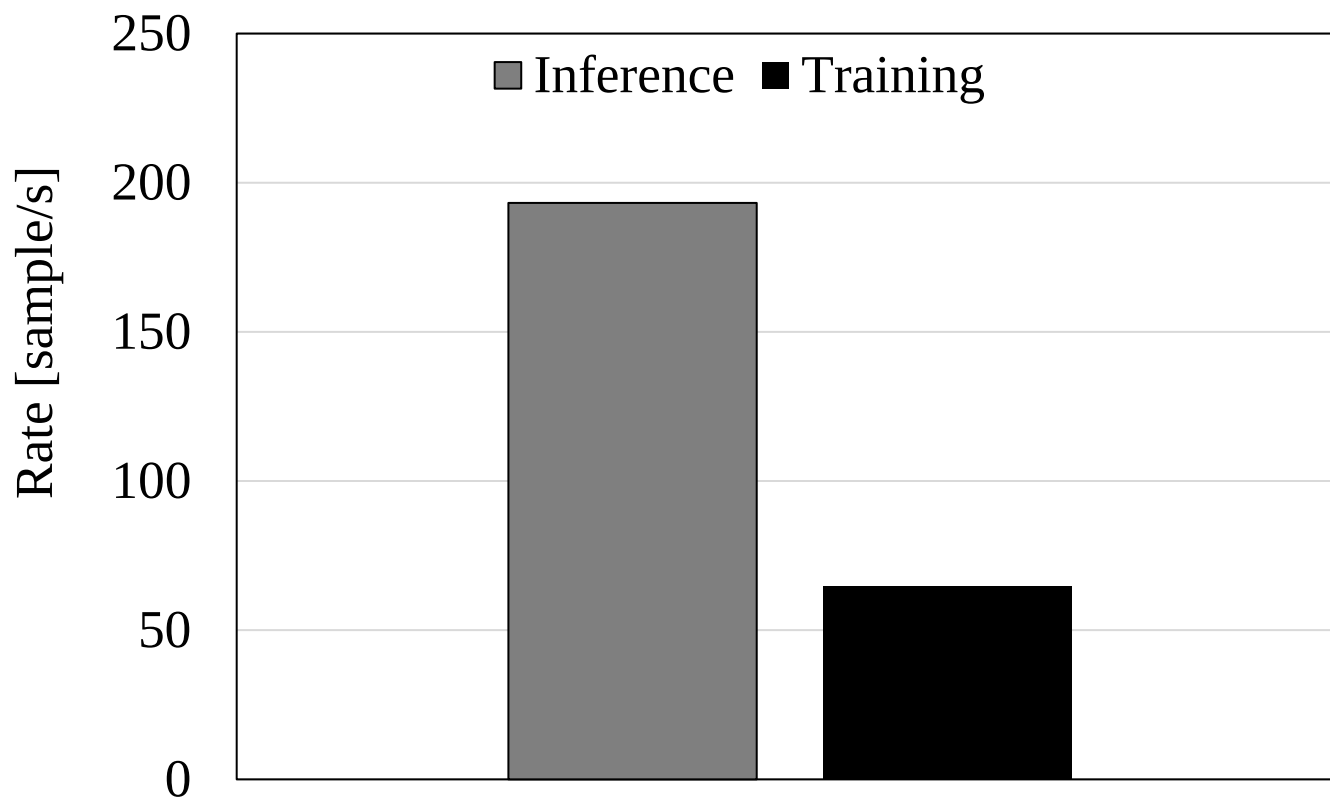
Human activity dataset
with ~7500 samples

Fully trained in less
than 2 seconds!

Achieving a similar
accuracy as 8-bit DNN

Ultra-low Power mW

RISC-V Customization for Edge AI: Inference



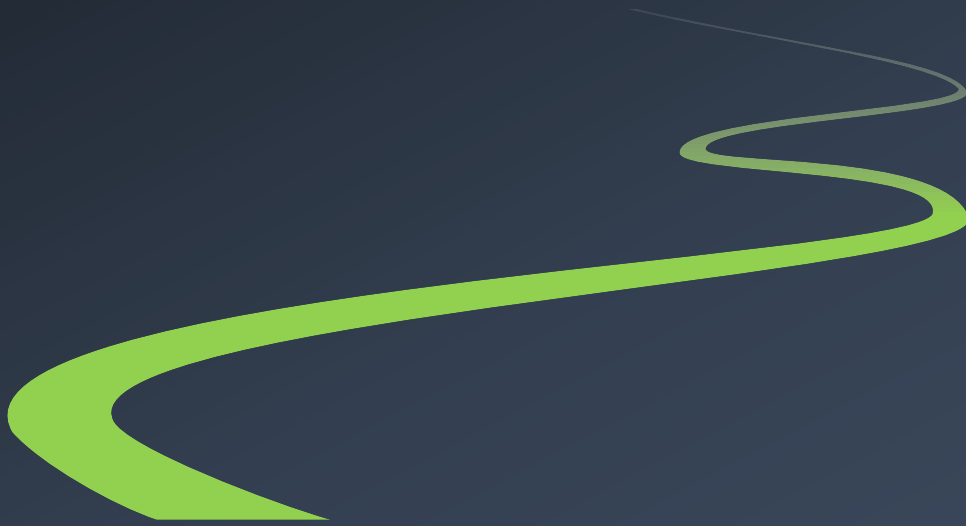
Inference rate reaches

**~200 samples per
second**

Power in merely mW



**Chip Fabrication using
22nm Technology**



Brain-inspired Computing for Edge AI

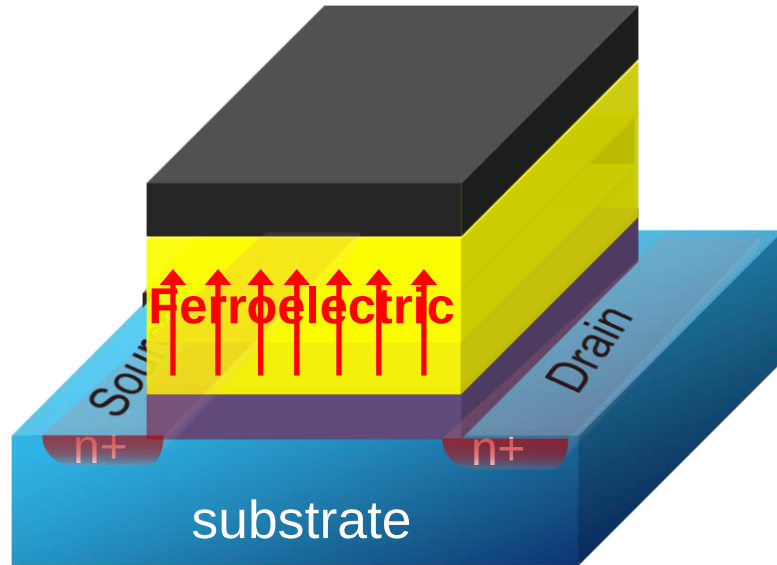
Novel Memory

RISC-V Customization

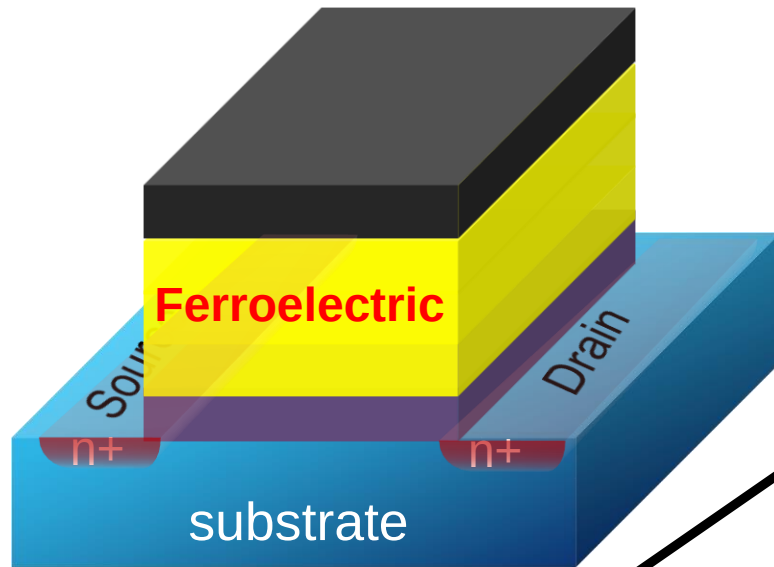
Hyperdimensional
Computing



Novel Advanced Memories

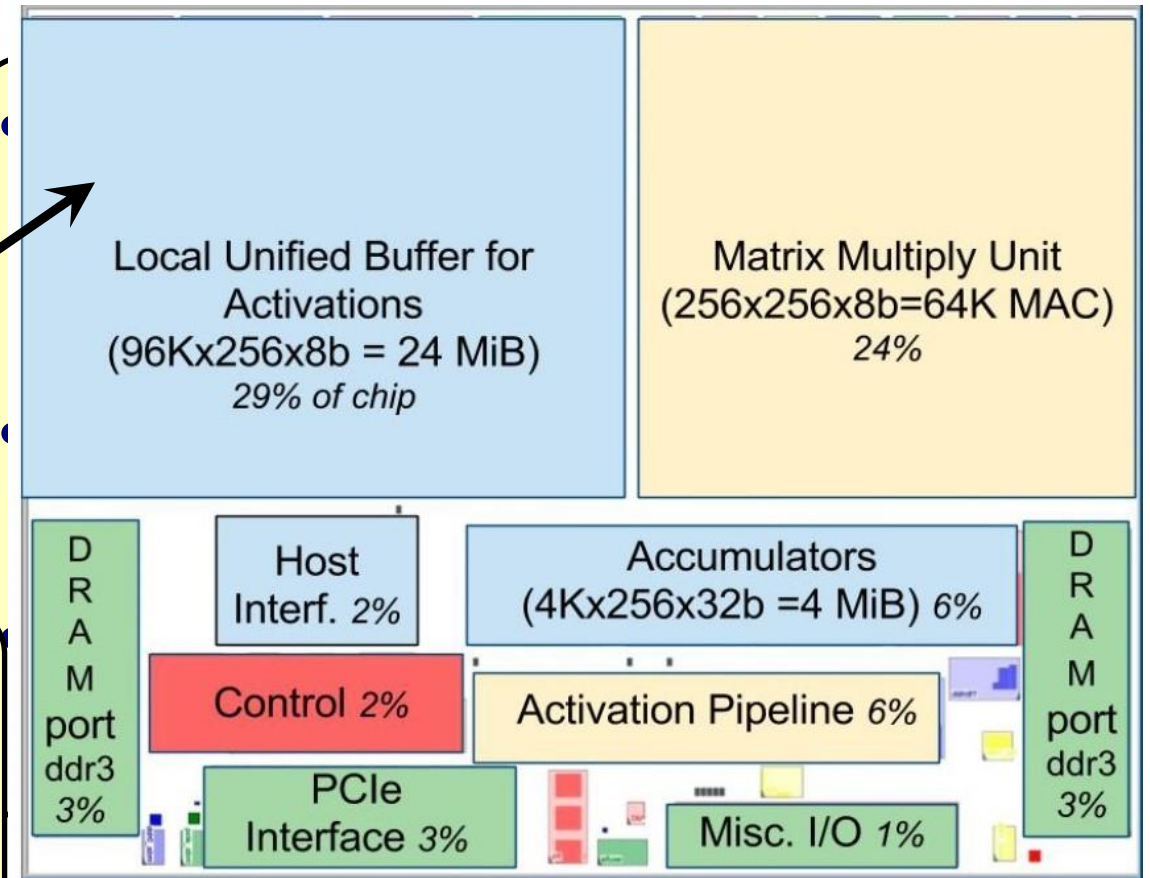


Next Generation of Memory



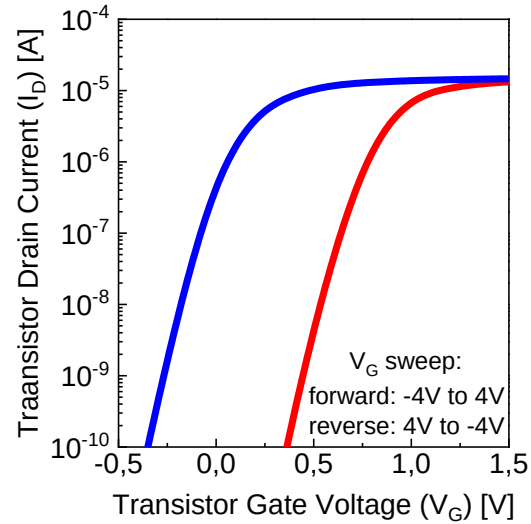
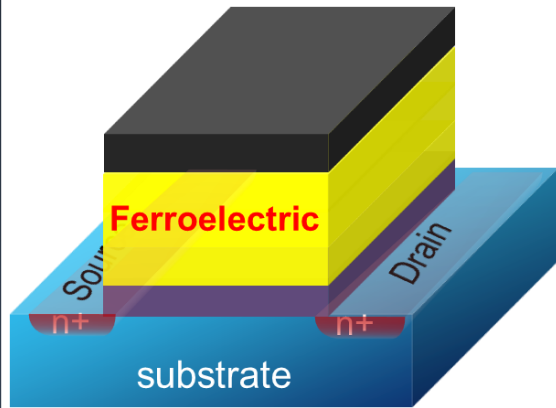
Replacing SRAMs with FeFETs

- Large Power Saving
- Higher Storage Capacity
- Less DRAM Communications

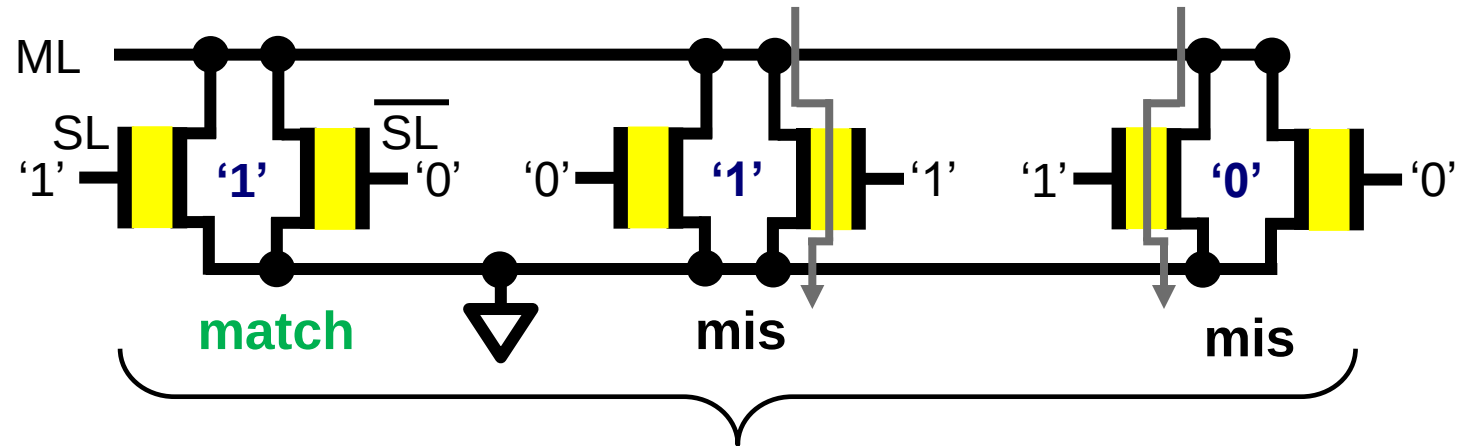


AI Chip: Google TPU [ISCA'17]

In-Memory Computing

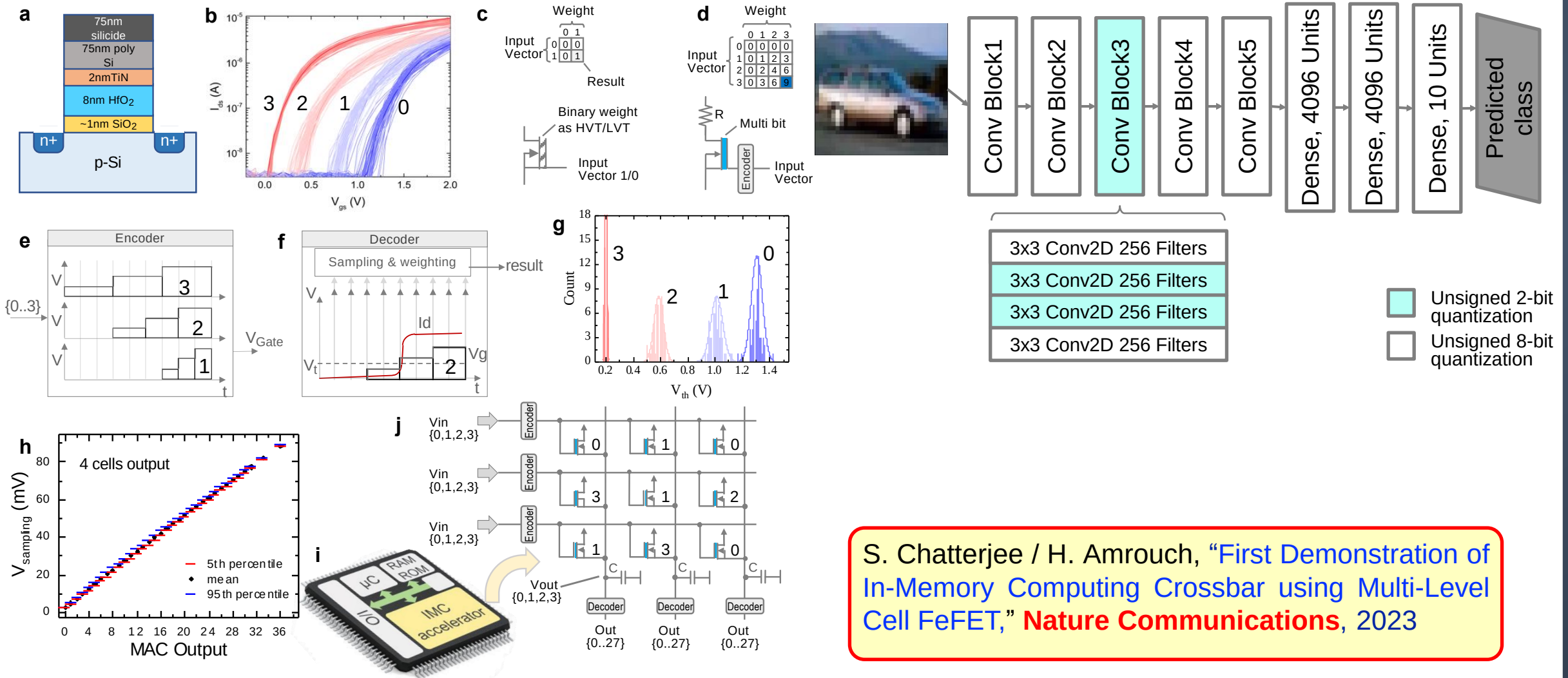


Store '110'
Search '101'

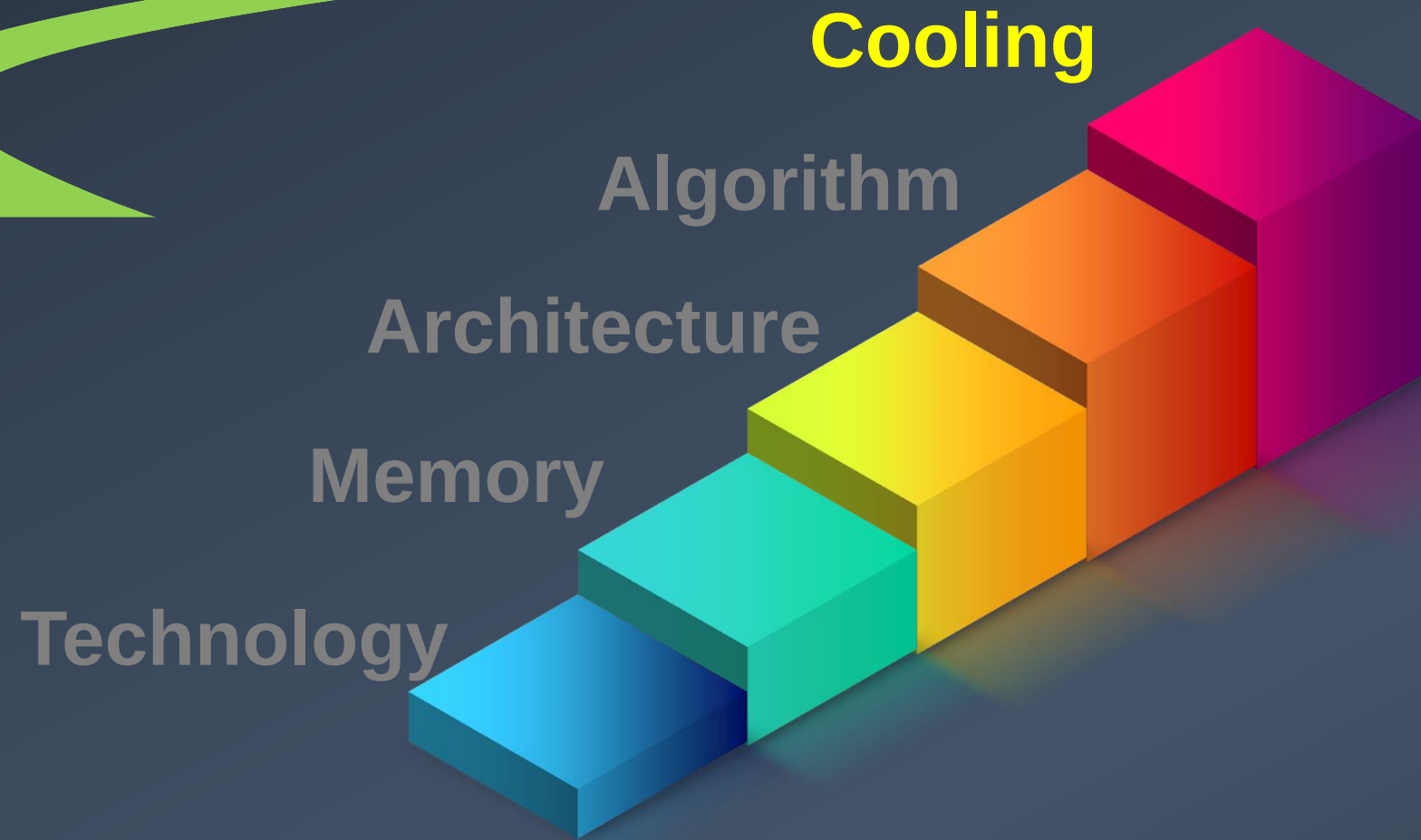
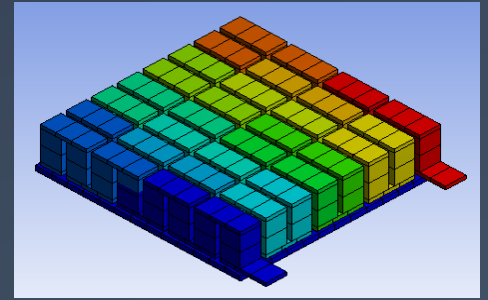


Hamming distance of 2

Very Efficient In-Memory Computing

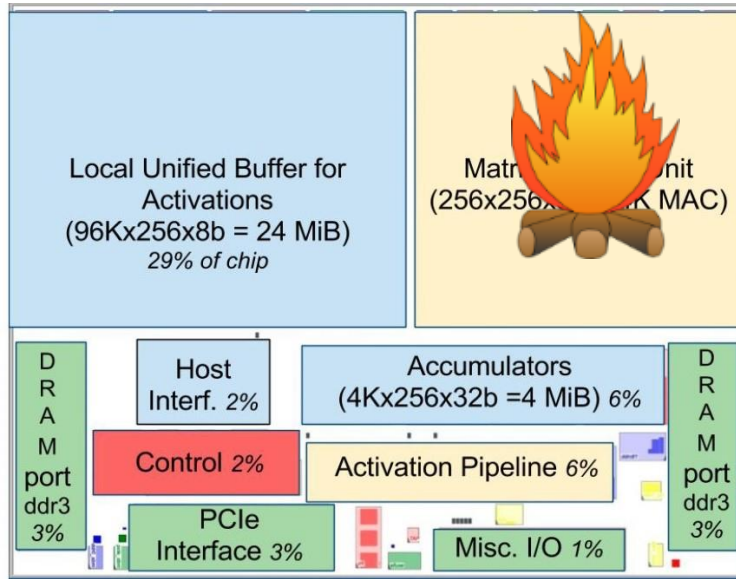


S. Chatterjee / H. Amrouch, "First Demonstration of In-Memory Computing Crossbar using Multi-Level Cell FeFET," **Nature Communications**, 2023



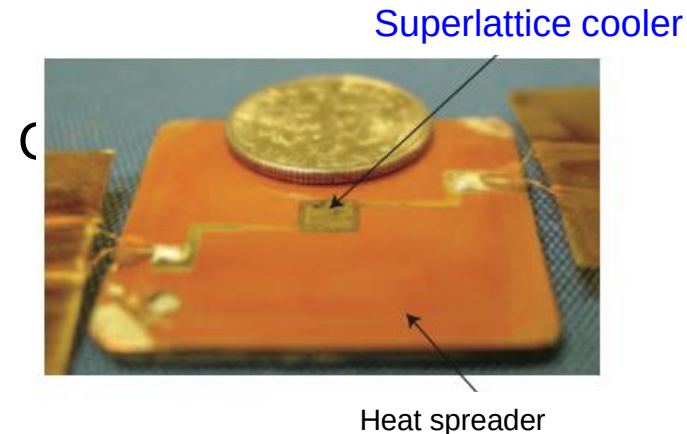
**On-Chip
Cooling**

Intelligent Cooling: Superlattice Thermoelectric



AI Chip: Google TPU [ISCA'17]

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

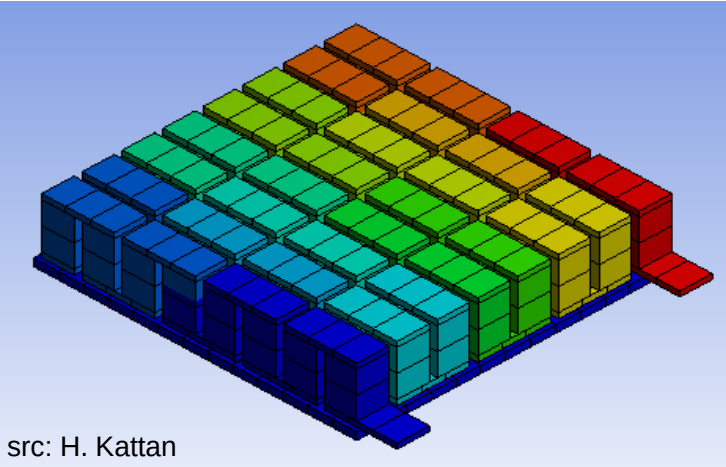


Bulman, G., Barletta, P., Lewis, J. et al. Superlattice-based thin-film thermoelectric modules with high cooling fluxes. *Nature Communication*, 2016

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

Intelligent Cooling: BEOL Thermoelectric

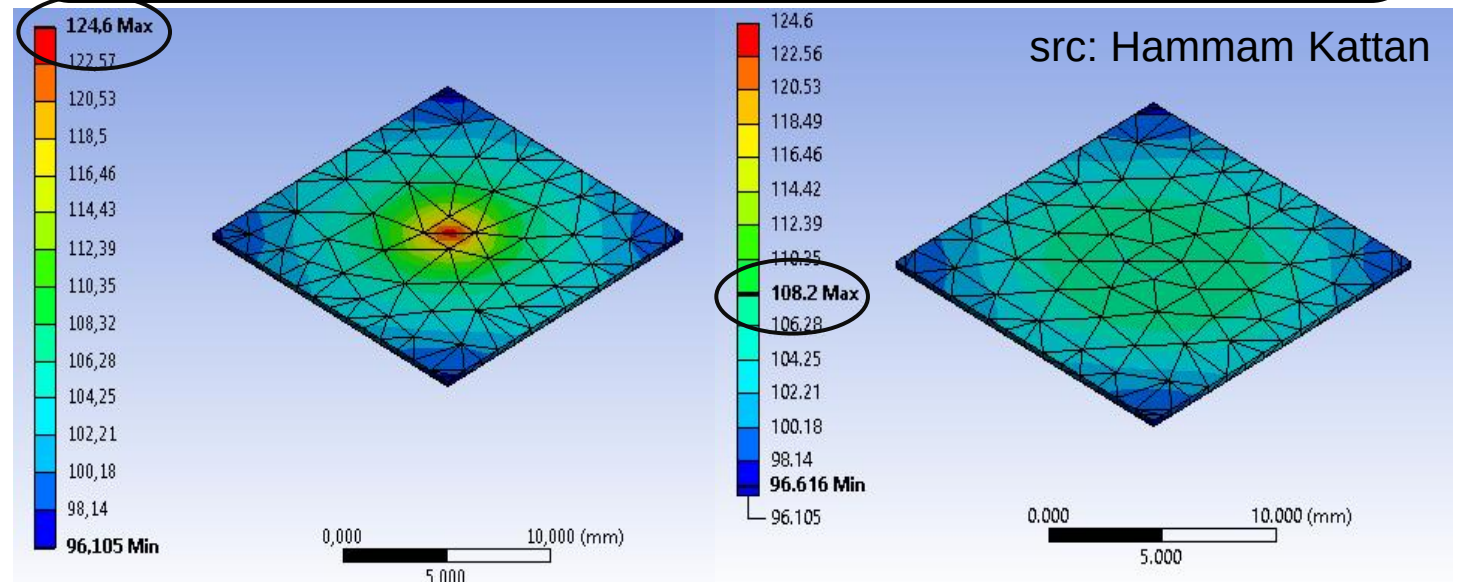
Thermoelectric:
Peltier's effect $Power \rightarrow \Delta T$



ANSYS®

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

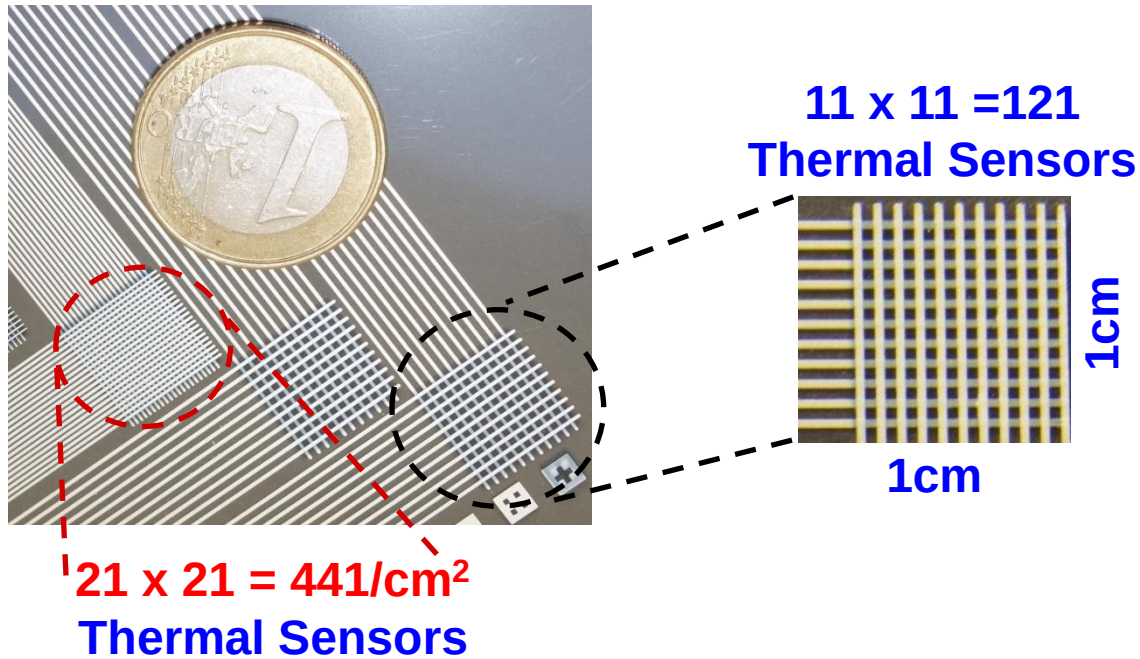
**On-chip cooling:
localized and On-demand
→ VERY Efficient**



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

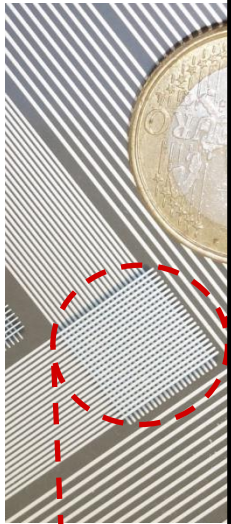
Intelligent Temperature Sensing

First Printed Thermal Sensors Array for Processor Chips

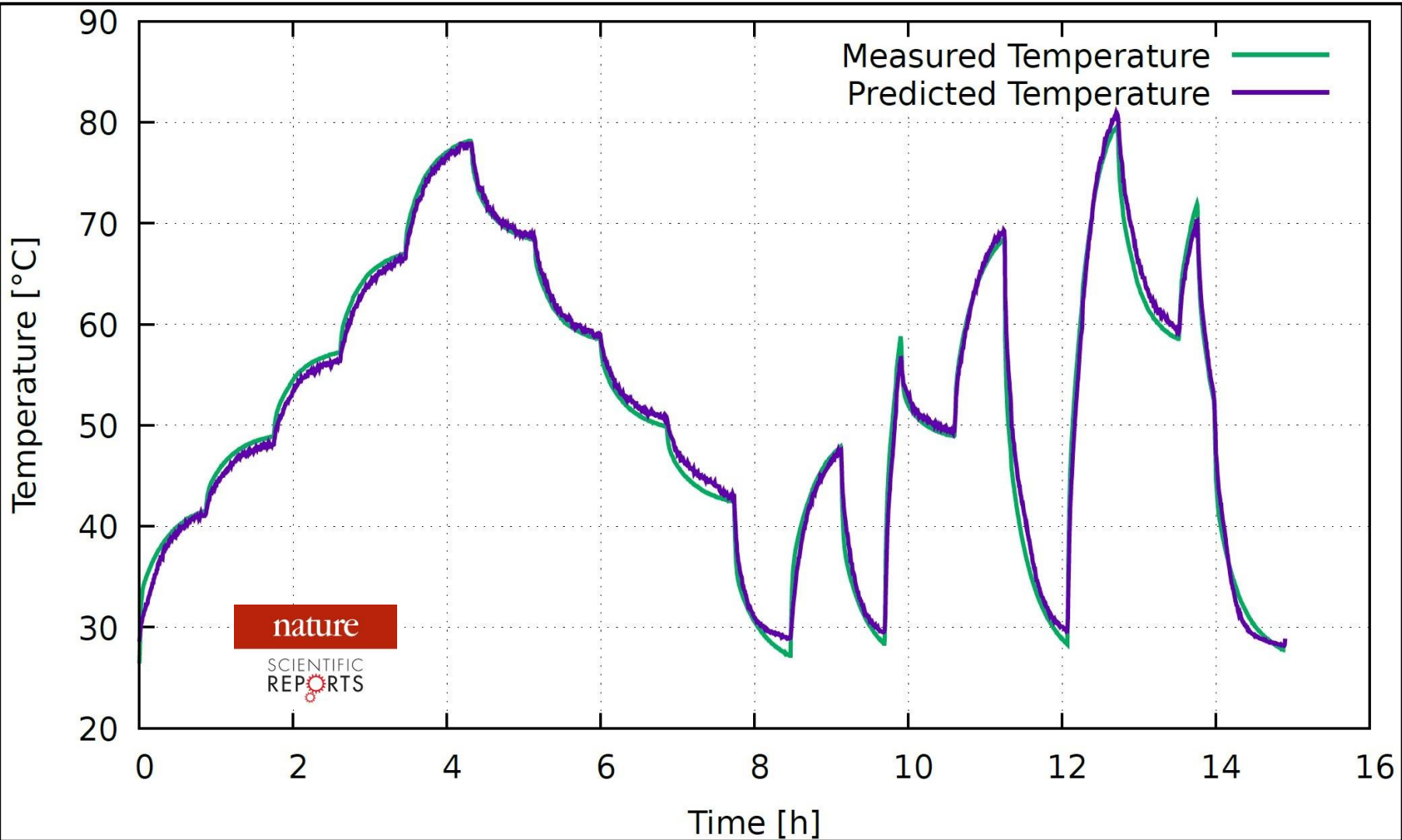


Ack: Heidelberg InnovationLab, U. Lemmer, KIT

Intelligent Temperature Sensing

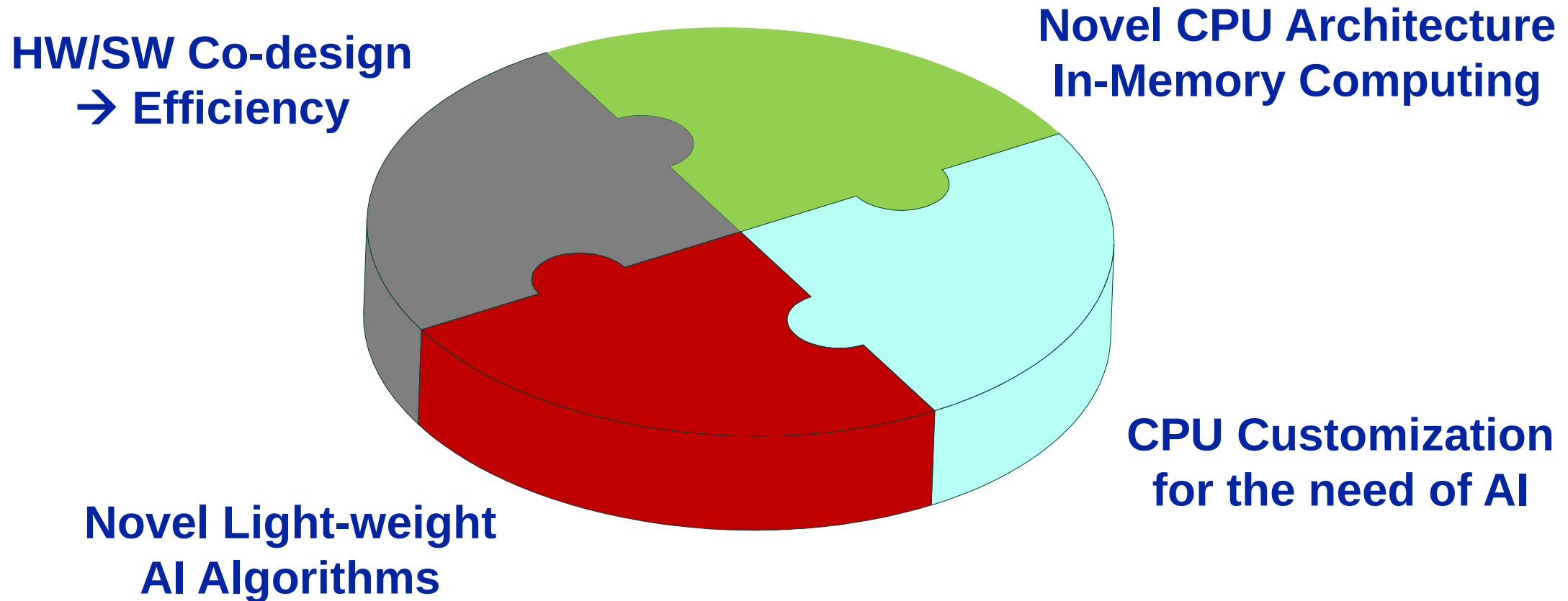


21 x 21
Thermal



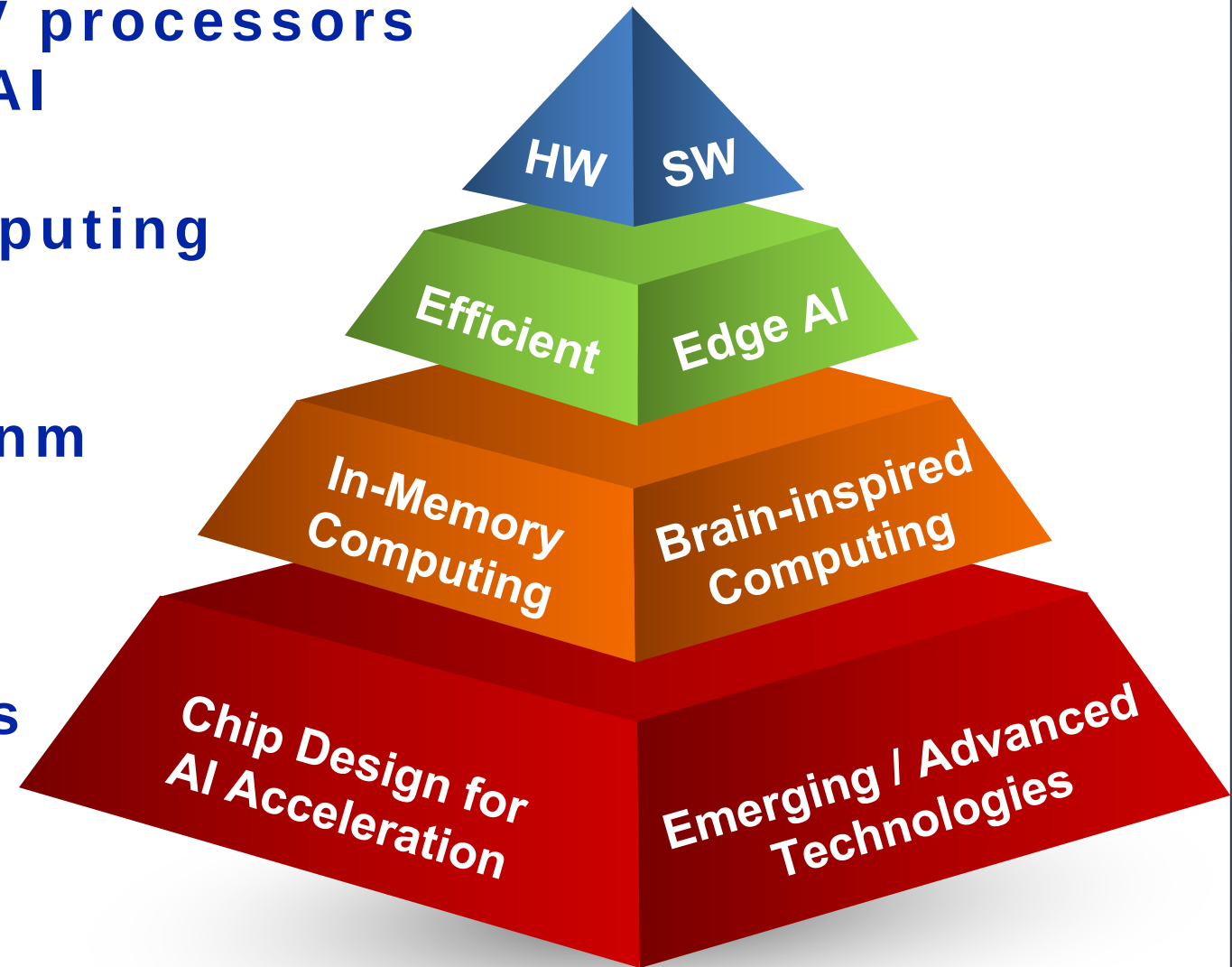
onLab, U. Lemmer, KIT

AI Chips: All about Customization!



TUM: AI Processor Design

- Highly-customized RISC-V processors for ultra-low power edge AI
- Advanced in-memory computing for AI acceleration
- Chip design in 22nm & 16nm
- Reliability analysis in Foundry 3nm technologies



Research Partners

SYNOPSYS®



ADVANTEST®

Berkeley
UNIVERSITY OF CALIFORNIA

On the Brink of a new Era in AI Chips



**Novel AI
Algorithms**

**Processor
Customization**

Technical
University
of Munich

