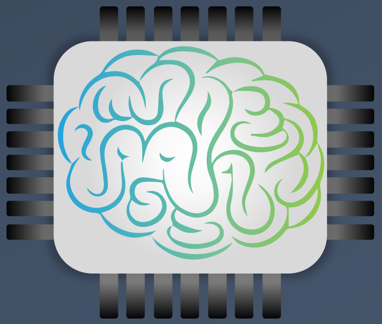
Frontiers in Edge AI with RISC-V: Hyperdimensional Computing vs. Quantized Neural Networks

by Hussam Amrouch
Chair of AI Processor Design

Technical University of Munich

Journey with Edge AI



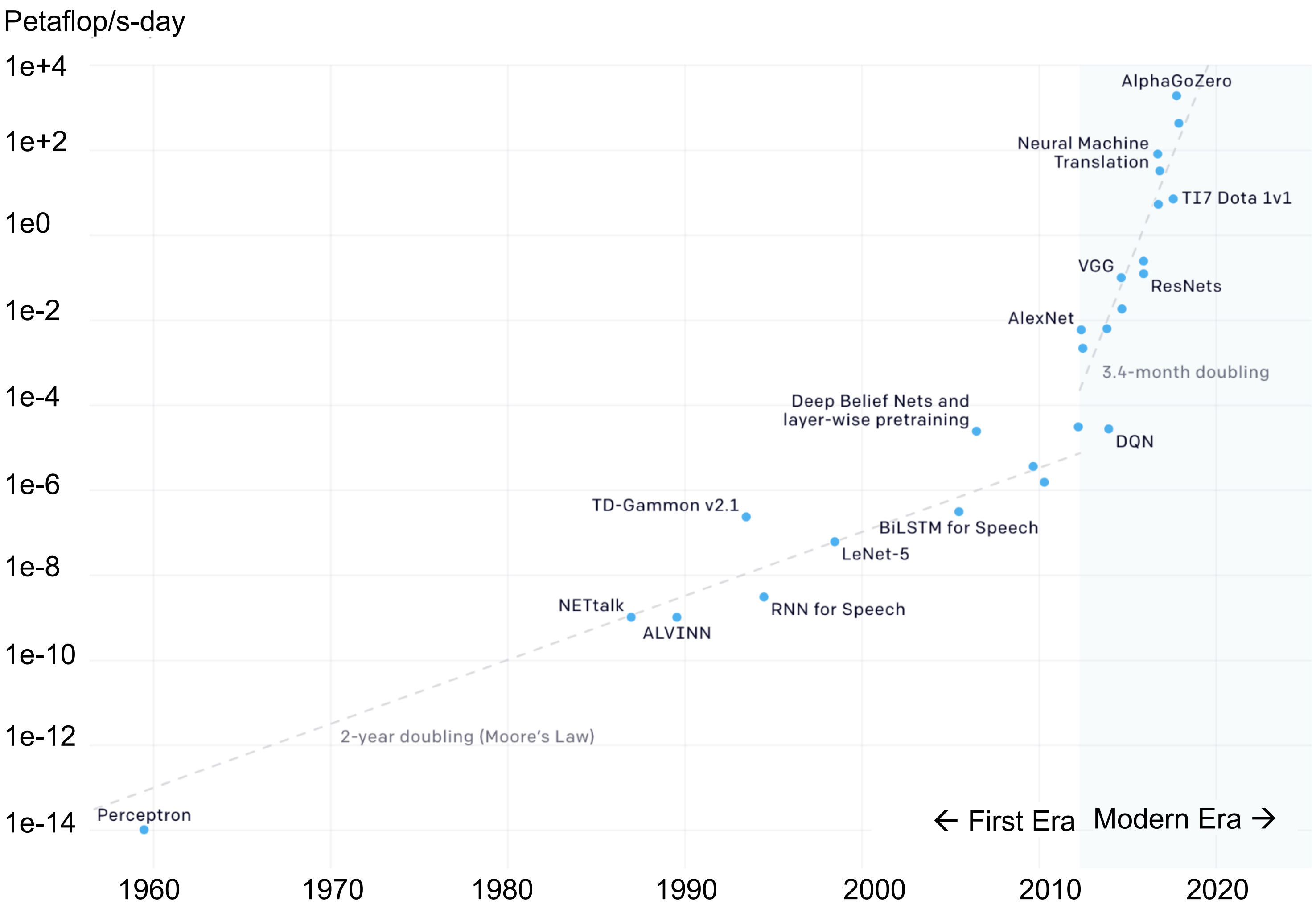
Edge AI
In-Memory
Computing

Domain
Specific
CPU

RISC-V

Hyperdimensional
Computing

The Next Revolution: AI



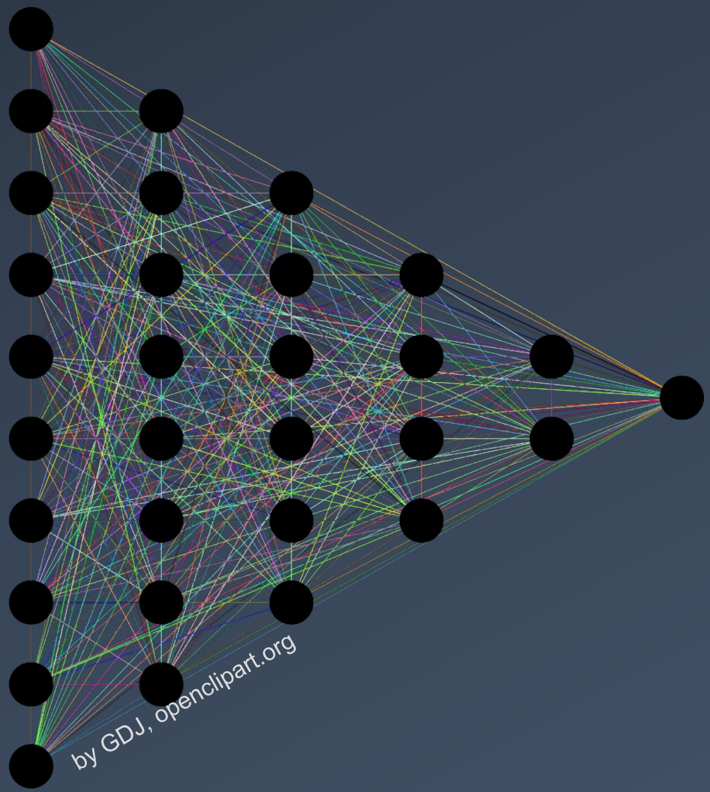
Computing demand

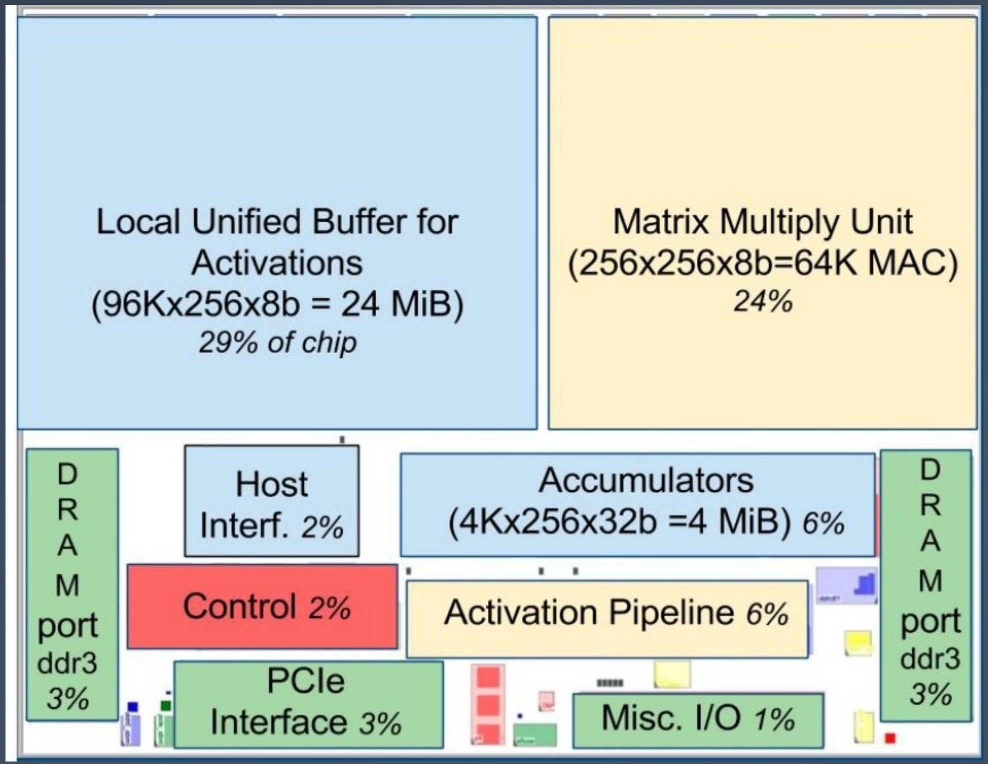
3.4 months doubling!

Source: https://openai.com

The Next Revolution: AI

Deep Learning





AI Chip: Google TPUv1 [ISCA'17]

Training BERT DNN Google TPUv3: 1.8min ≈

2048 GPUs + 512 CPUs

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

The Upcoming Jevons Paradox

Increase in AI Hardware Efficiency → **Cost of DNN Training drops** → **More Companies are adopting AI**

2030: 13% of Total CO_2

More and more data centers

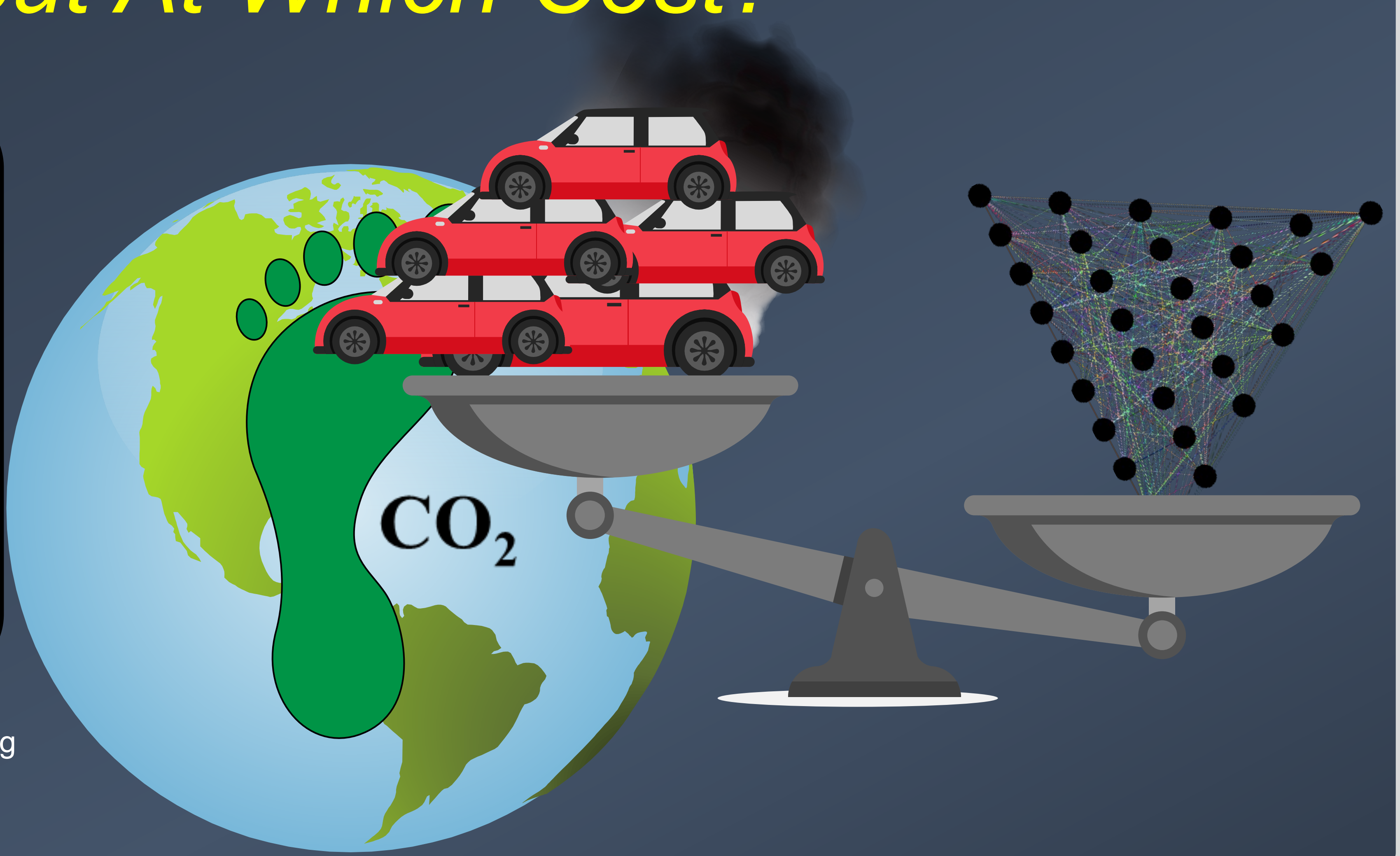
sources: IEEE Spectrum (2019), Nature (2020)

AI is reshaping the Future of Humankind

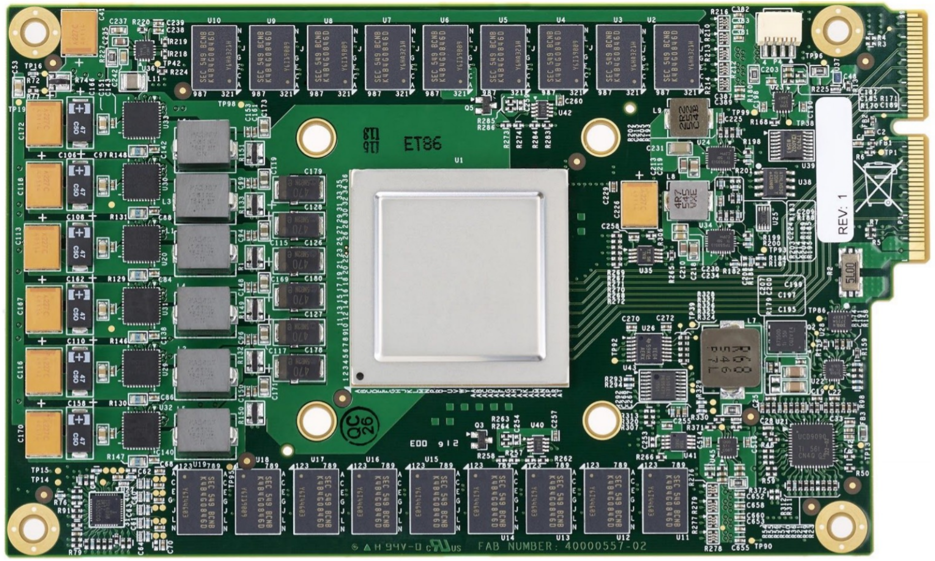
But At Which Cost?

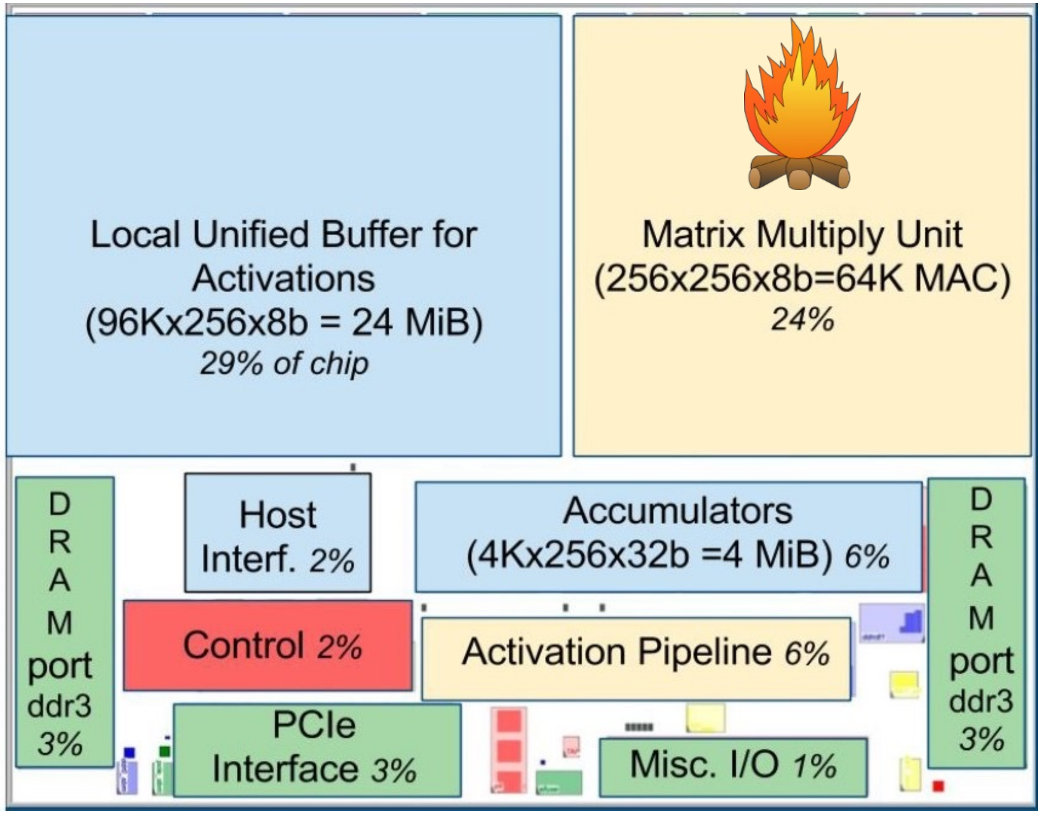
Training a single AI model emits carbon > 5x cars in their lifetimes

src: Emma Strubell, et al. "Energy and Policy Considerations for Deep Learning in NLP" in 57th ACL, 2019.



Deep Learning is REALLY Power Hungary!

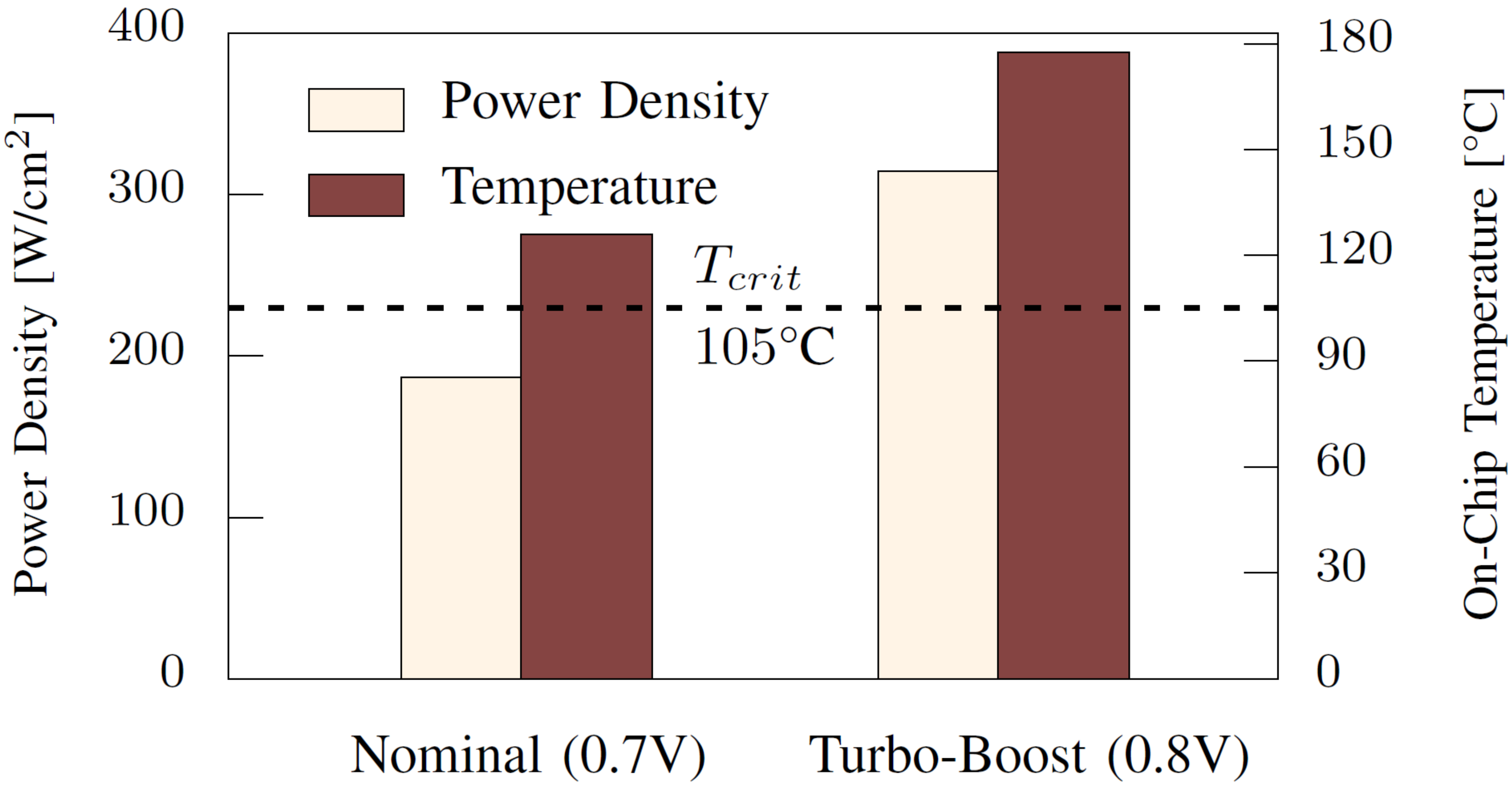




Google TPU [ISCA'17]

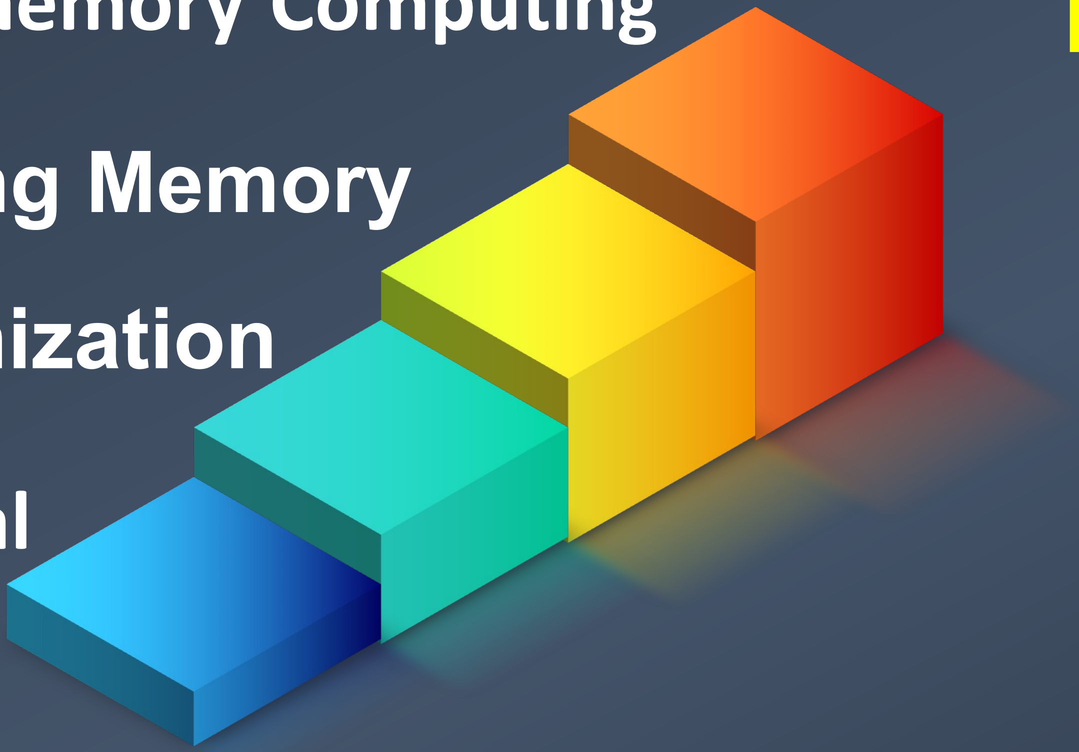
Deep Learning is REALLY Power Hungry!

Why not alternative algorithm to Deep Learning?



Amrouch [TCAD'20]

Brain-inspired Computing for Edge AI



Hyperdimensional Computing

RISC-V Customization

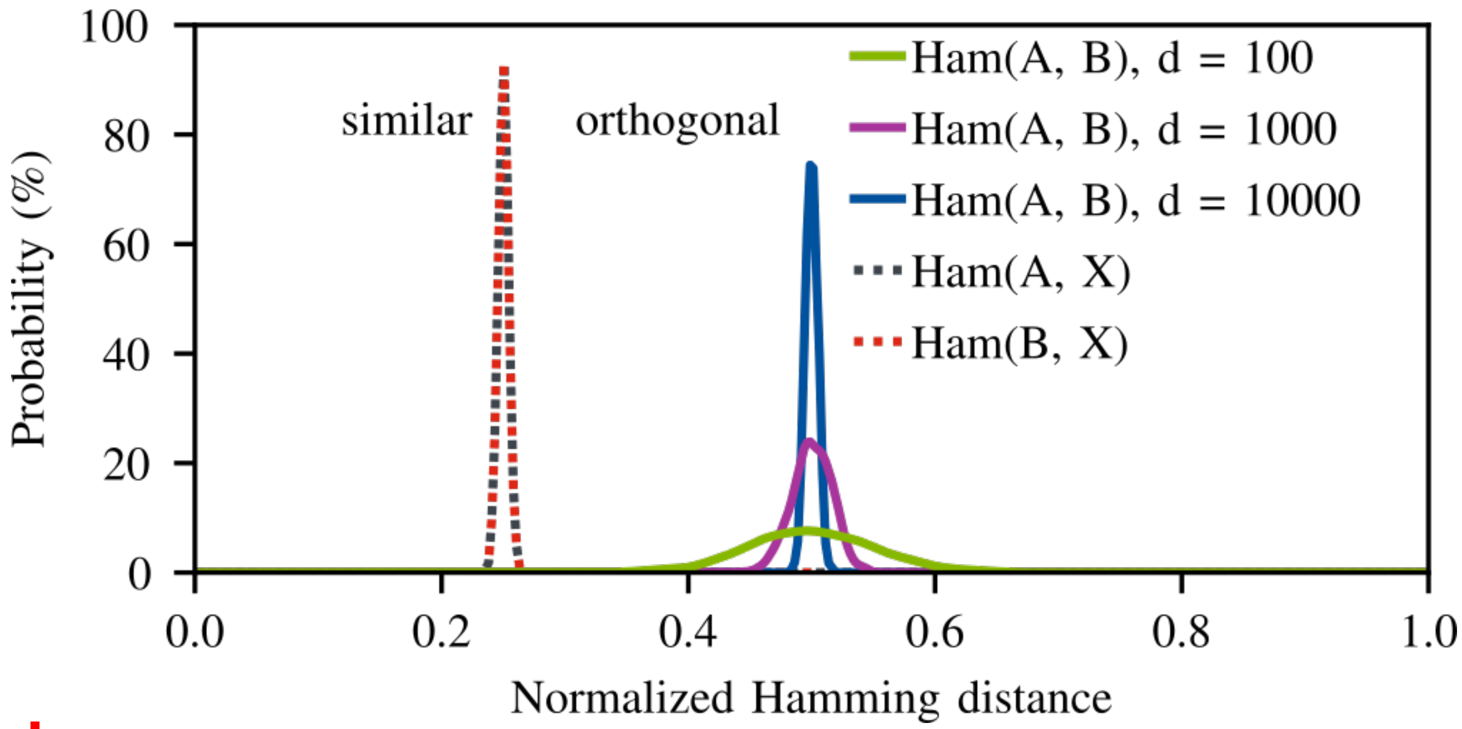
Emerging Memory

Hyperdimensional in-memory Computing

Hyperdimensional Computing

Brain-Inspired Hyperdimensional Computing

- Large vectors, e.g., 10000 elements
- Randomness is a feature not a bug
- Simple Operations
  - Permutation
  - Binding
  - Bundeling



Similarity is the Core Principle

Brain-Inspired Hyperdimensional Computing

1. Prepare: Encode real-world data into hyperspace

Random vectors representing basic real-world values → Store in *Item Memory (IM)* → Combine item vectors to encode complex real-world data into a vector

2. Learn: Train the model

Encode training data into one vector per class → Store label ↔ vector in *Associative Memory (AM)*

3. Inference: Recognize unknown data

Encode test data into a vector → Find most similar vector in AM → Inferred class

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: **To be, or not to be**

....

[00100100000111110001]

[01110100110101111111]

[10100100010111100010]

⋮

[10100100010111100010]

Count 1's

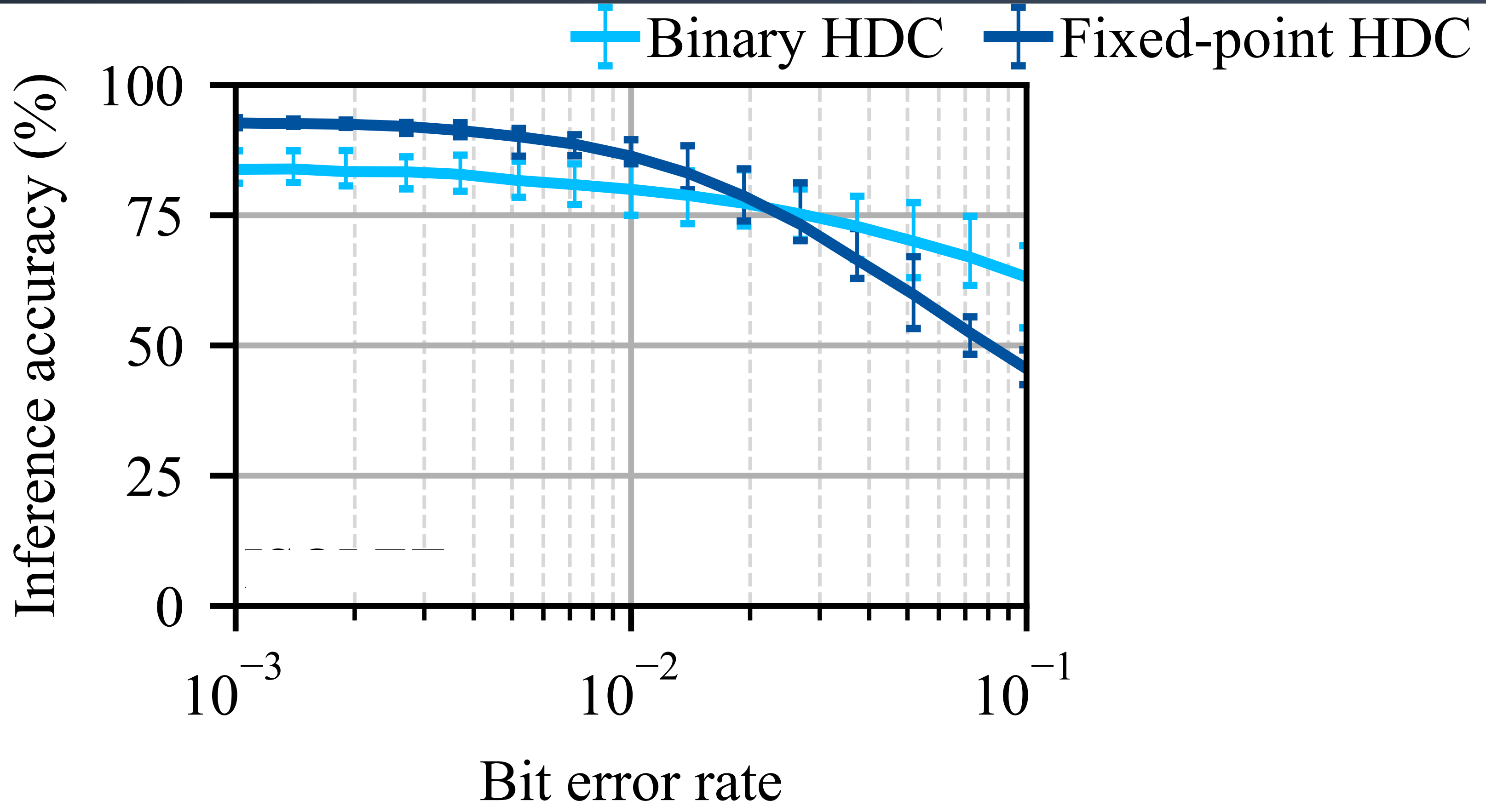
[3,6,10,9,13,4,19,..70]

Majority gate

[10100110001100111001]

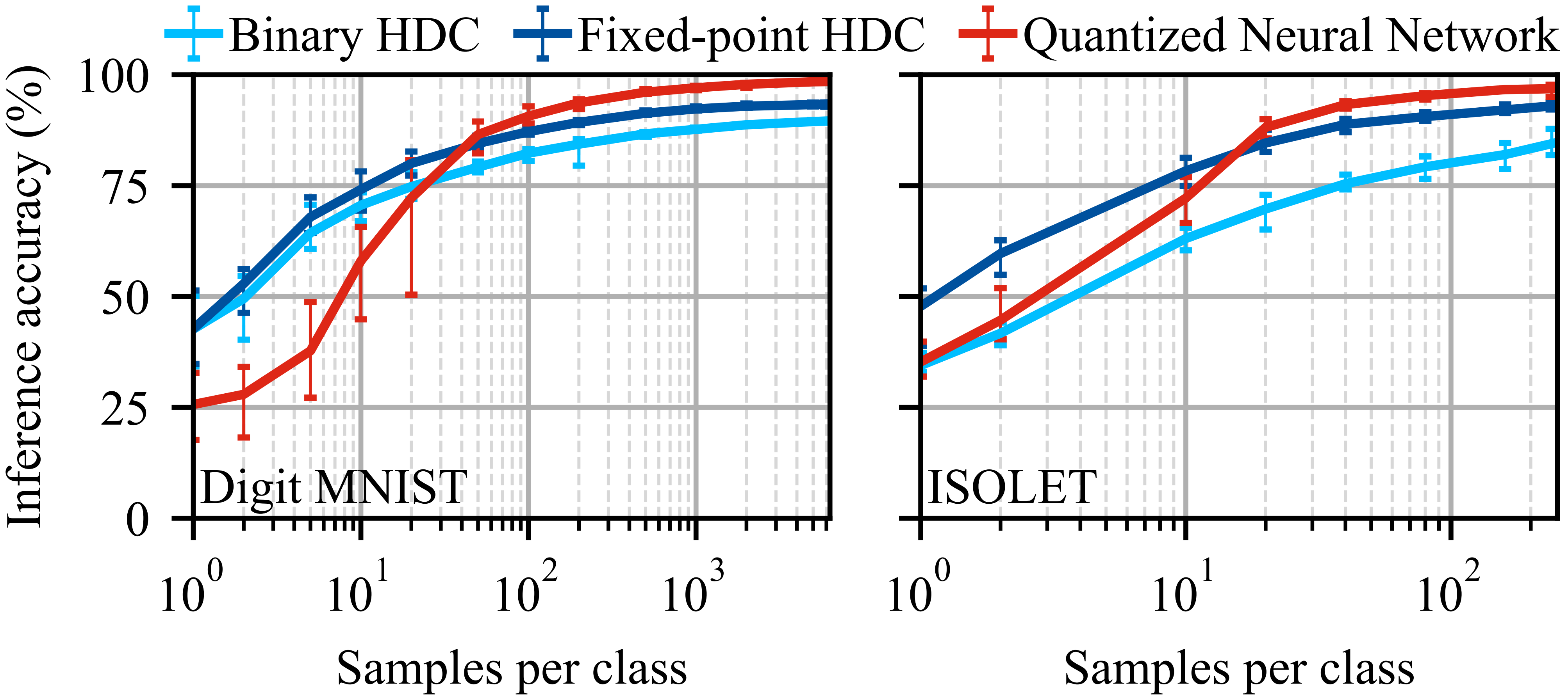
Entire language is *just one Hyper Vector*

Robustness against HW Errors and Noise



Errors injected in the underlying HW operations

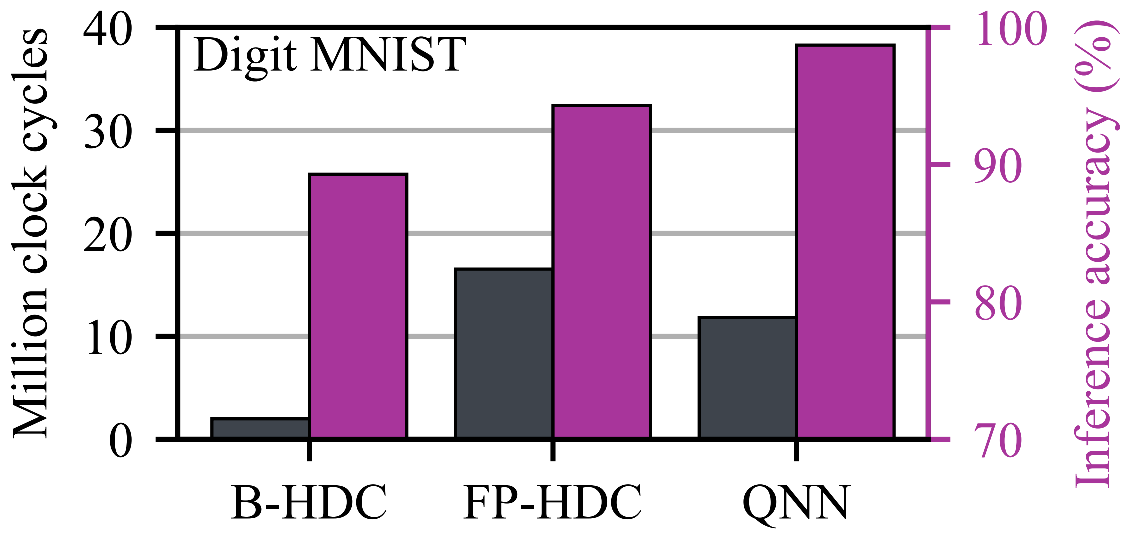
HDC vs. QNNs: Learning from Little Data!

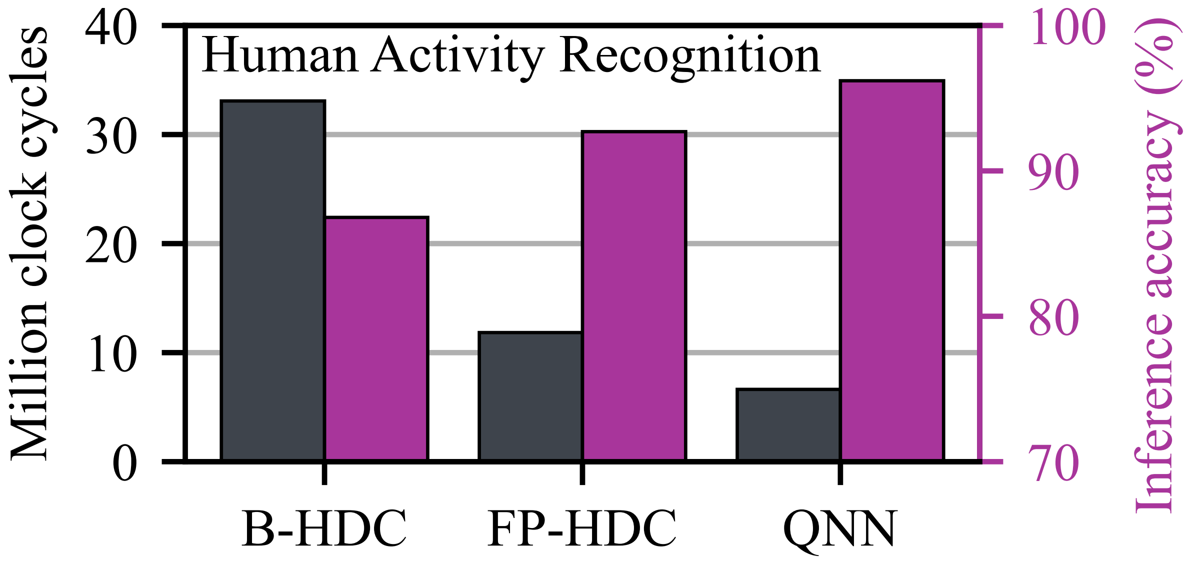


HDC learns from little samples

Fix-point HDC : similar accuracy to QNN

HDC vs. QNN: Performance / Accuracy





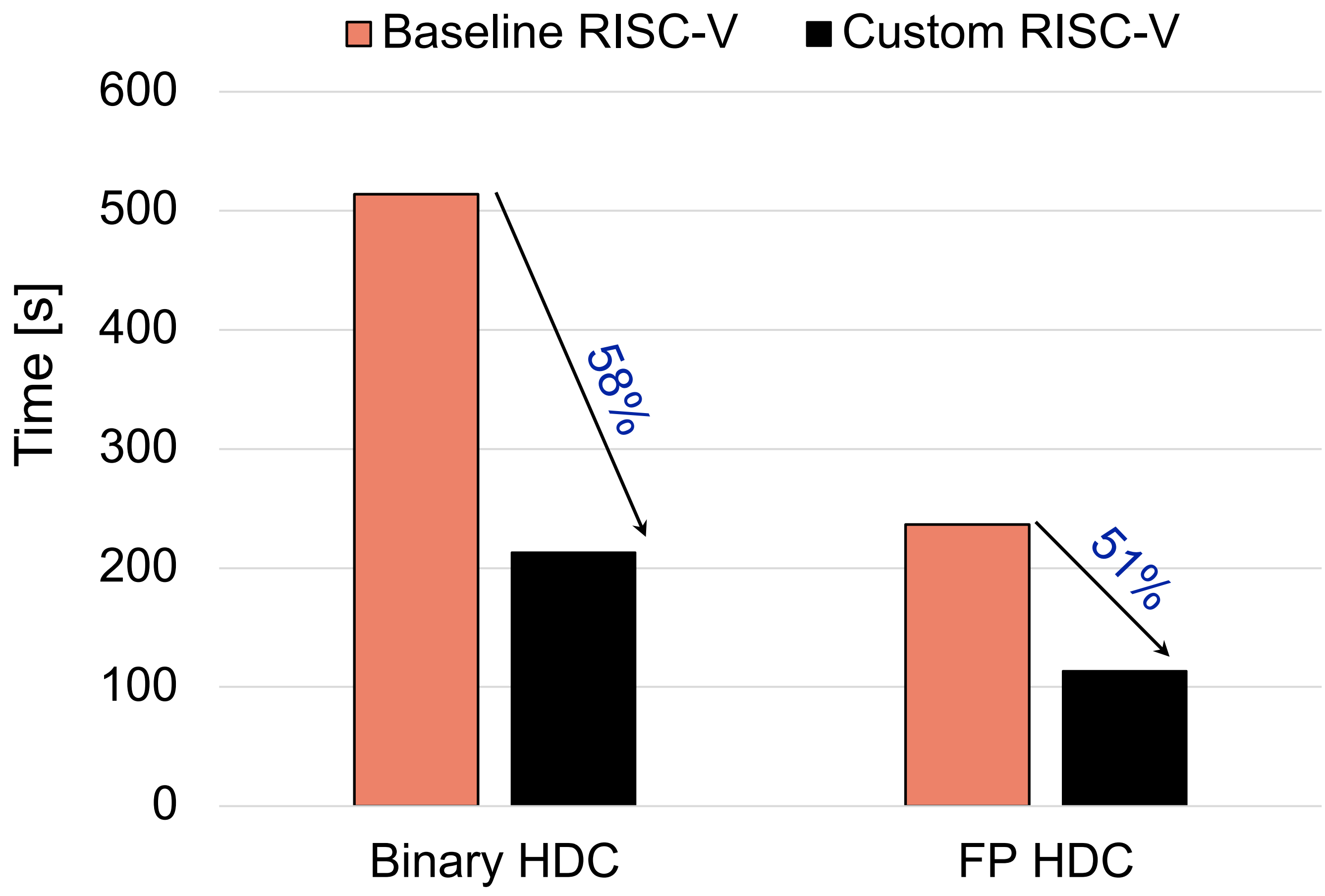
Binary HDC has a superior speed in image classification

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



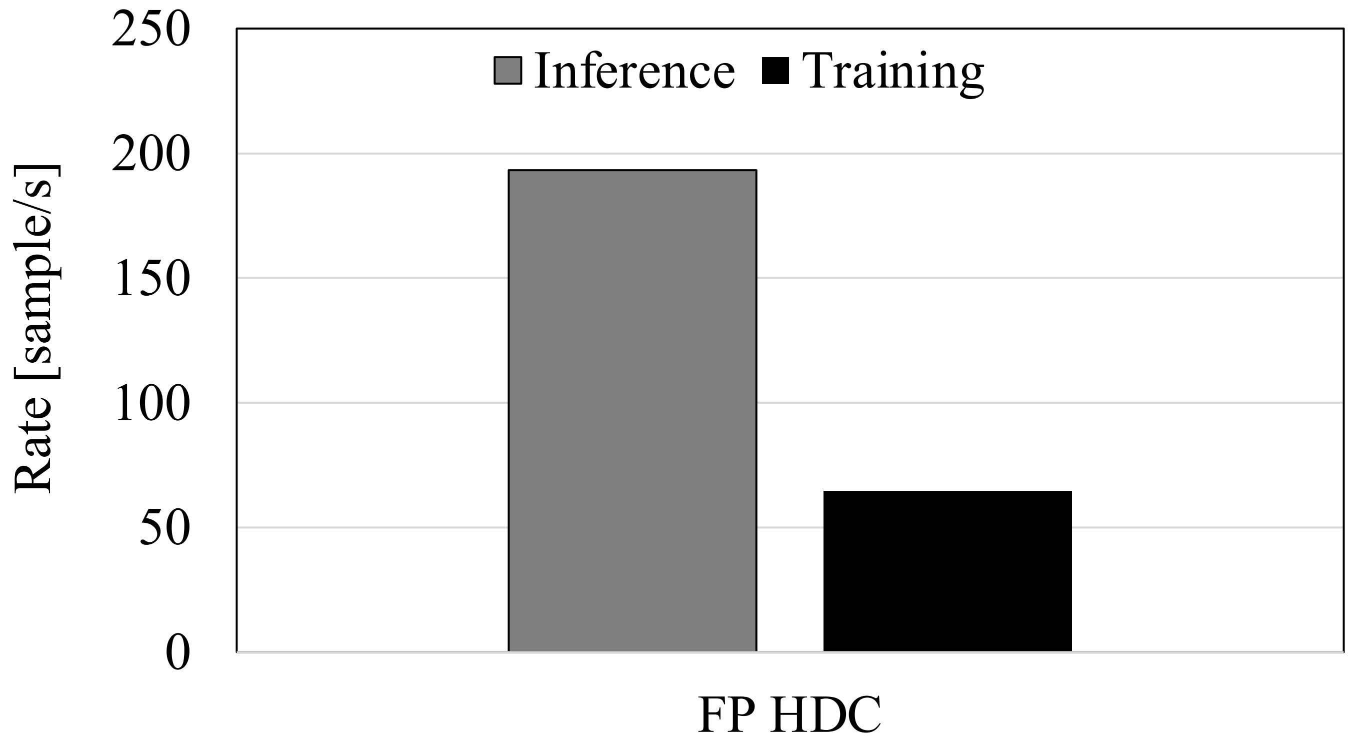
SYNOPSYS® ASIP Designer

Human activity dataset with ~7500 samples

Fully trained in less than 2 seconds!

Achieving a similar accuracy as QNN

RISC-V Customization for Edge AI: Training



Inference rate reaches
~200 samples per second

SYNOPSYS® ASIP Designer

HDC: Inevitable Memory Bottleneck

HDC relies on large vectors with > 1000 dimensions

→ Von Neumann architecture and memory bottleneck

In our analysis: Loading the vectors > 30% of cycles !

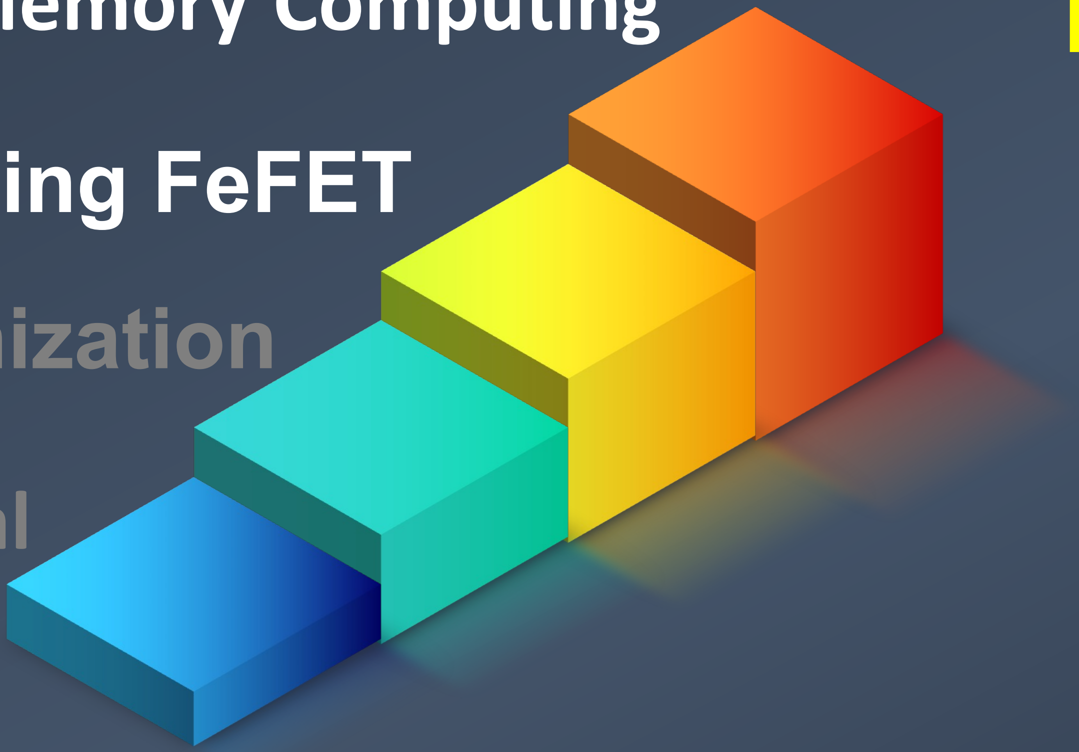
Brain-inspired Computing for Edge AI

Hyperdimensional In-Memory Computing

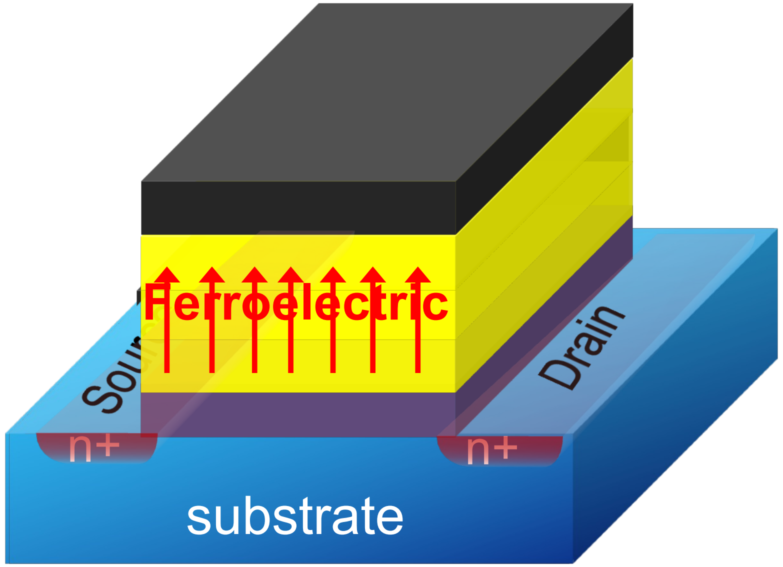
Emerging FeFET

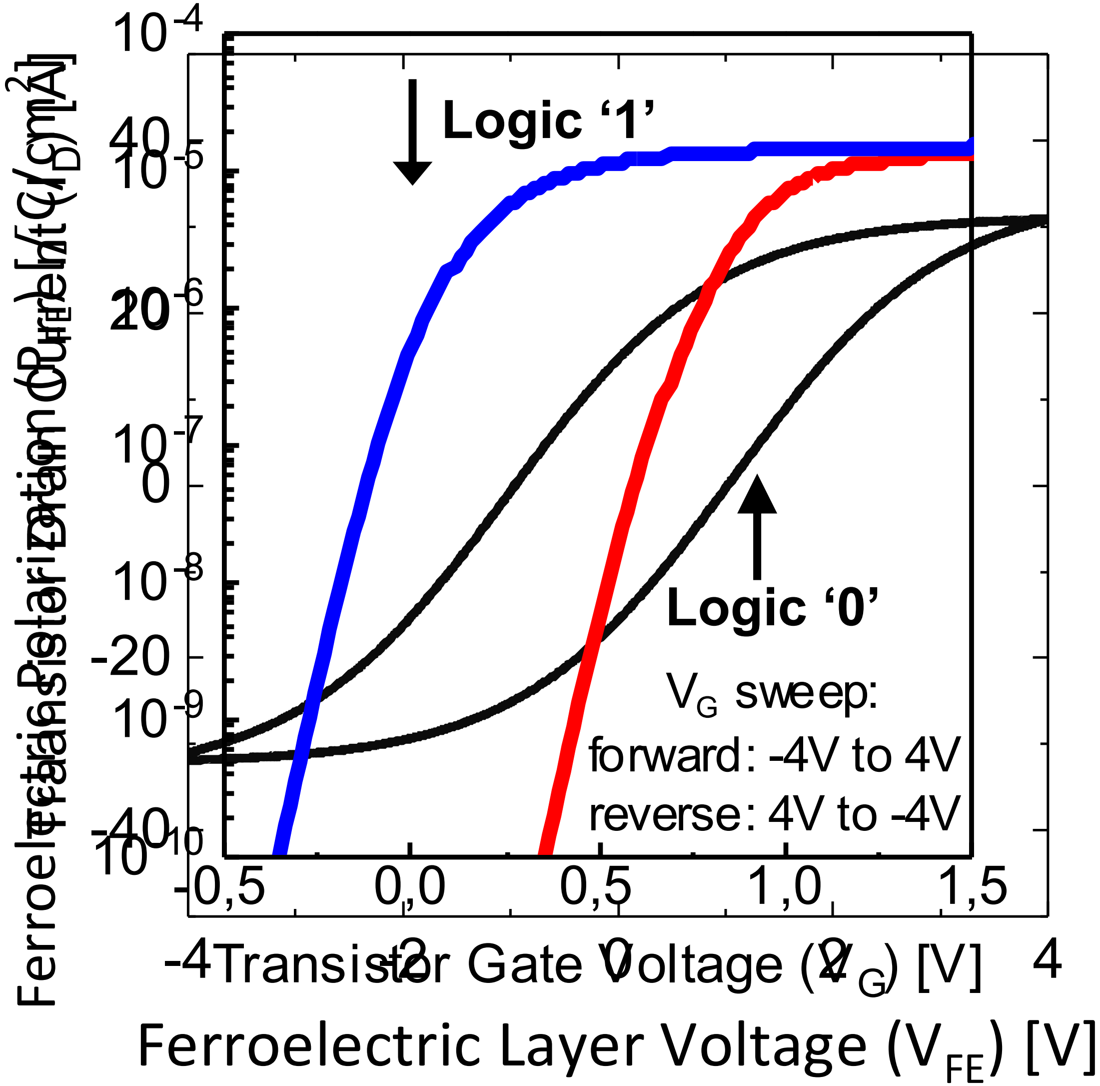
RISC-V Customization

Hyperdimensional Computing

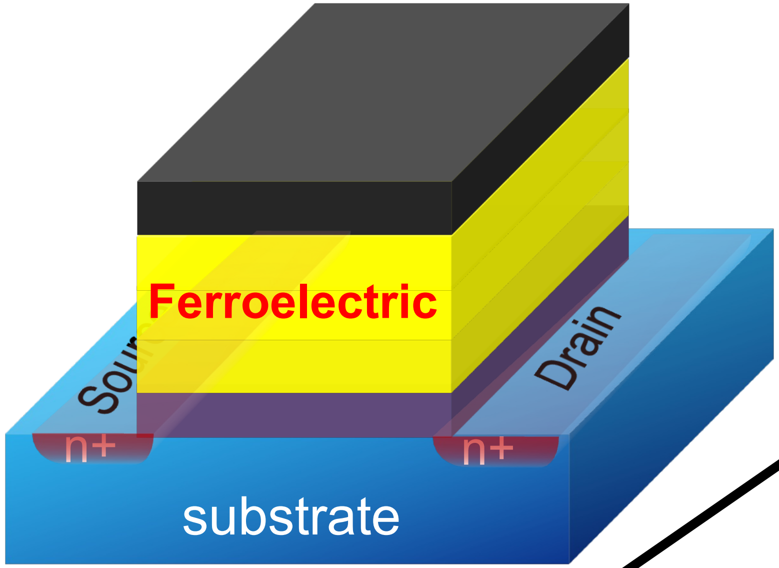


From FET to FeFET





FeFET: Emerging Memory

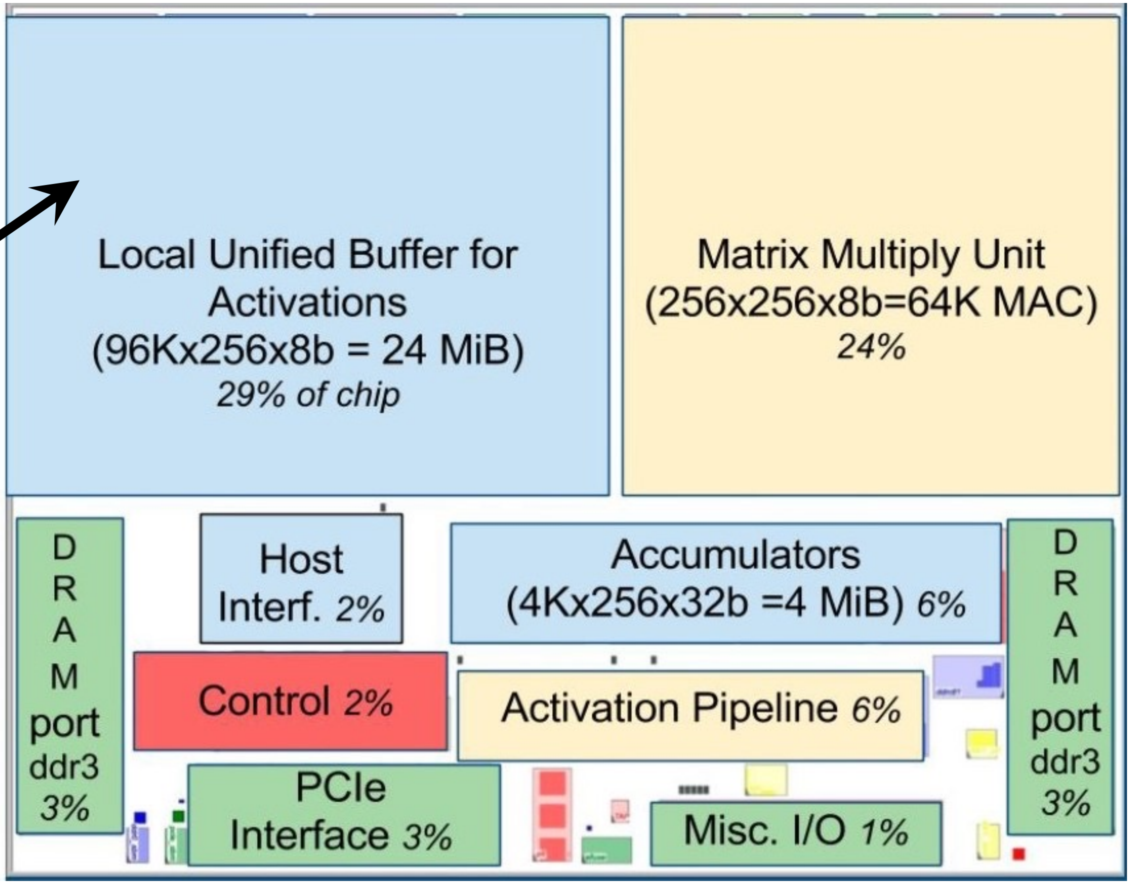


Replacing SRAMs with FeFETs

→ **Large Power Saving**

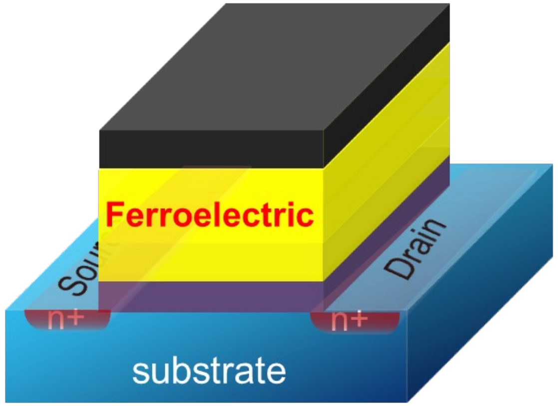
→ **Higher Storage Capacity**

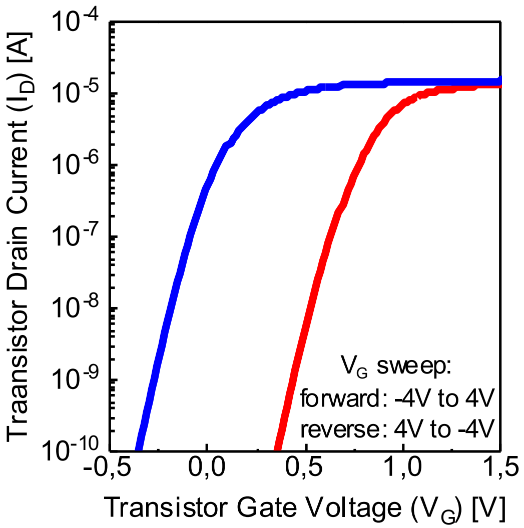
→ **Less DRAM Communications**

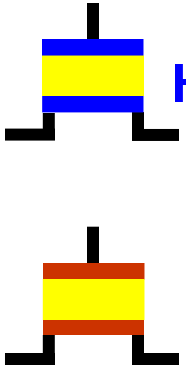


AI Chip: Google TPU [ISCA'17]

In-Memory Computing using FeFETs

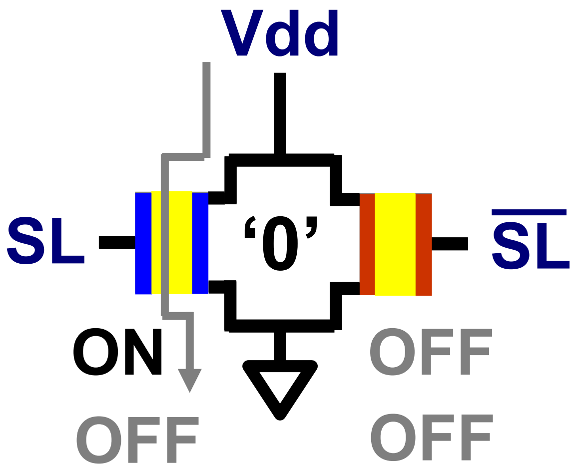






High current → '1'

Low current → '0'

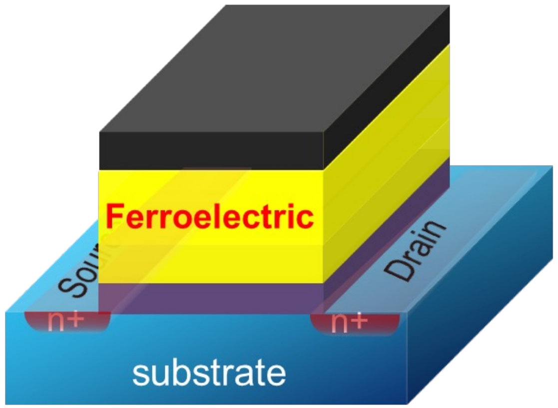


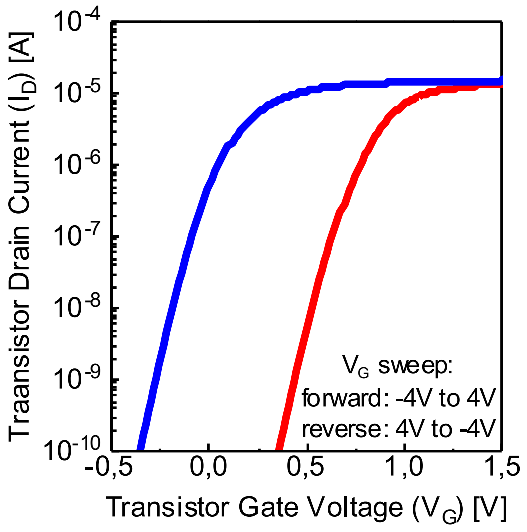
Sensing the current
→ *Match* or *Mismatch*

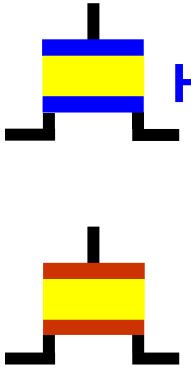
Search '1' → **Discharge**

Search '0' → **No Discharge**

In-Memory Computing using FeFETs





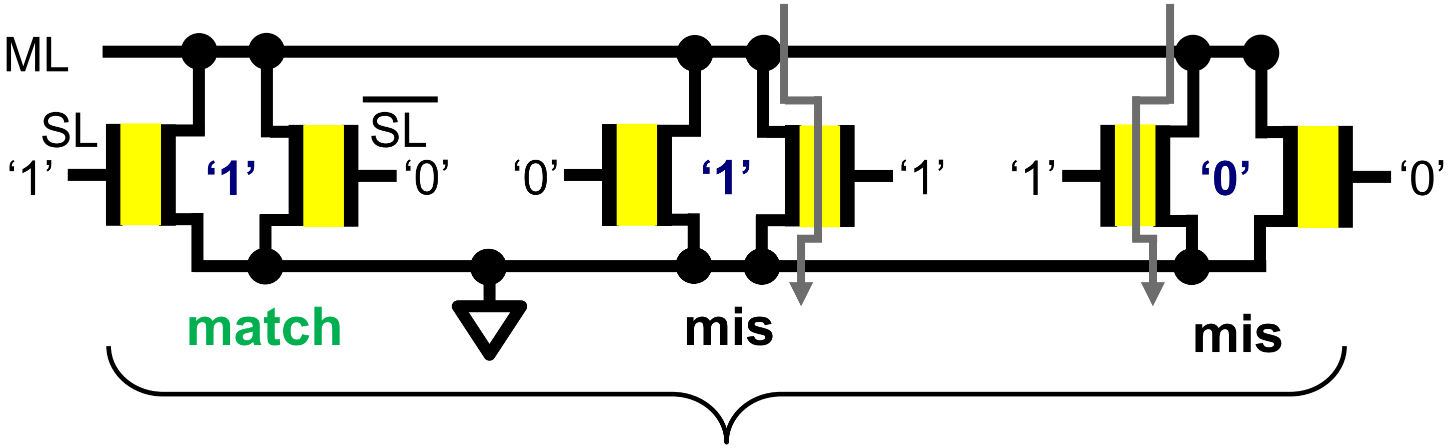


High current → '1'

Low current → '0'

Store '110'

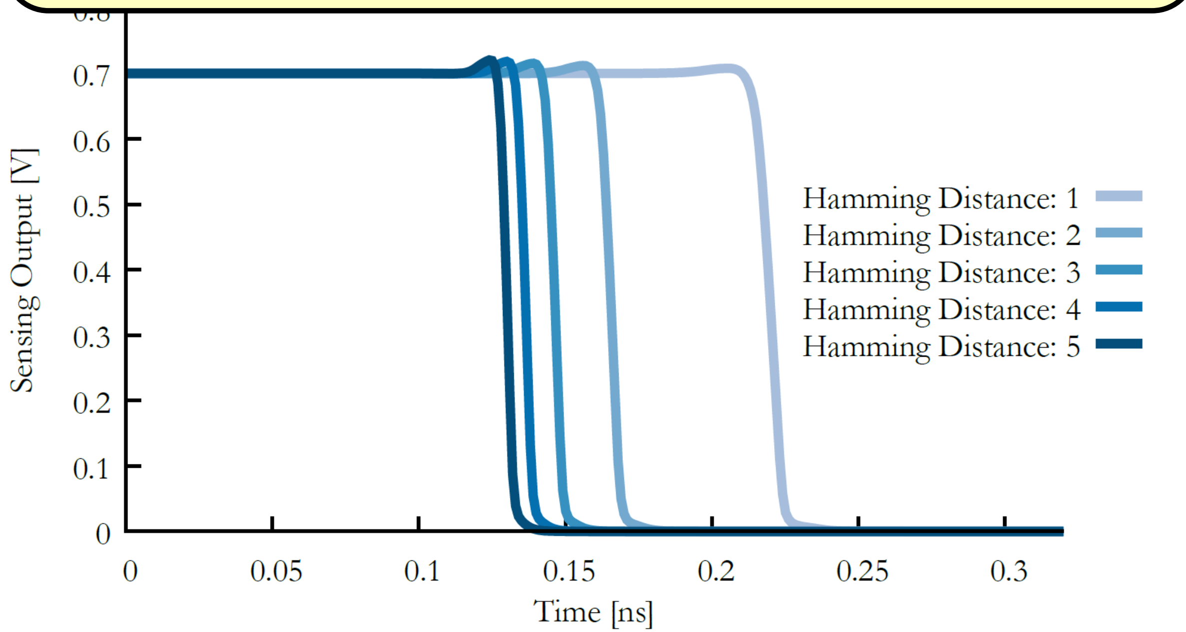
Search '101'

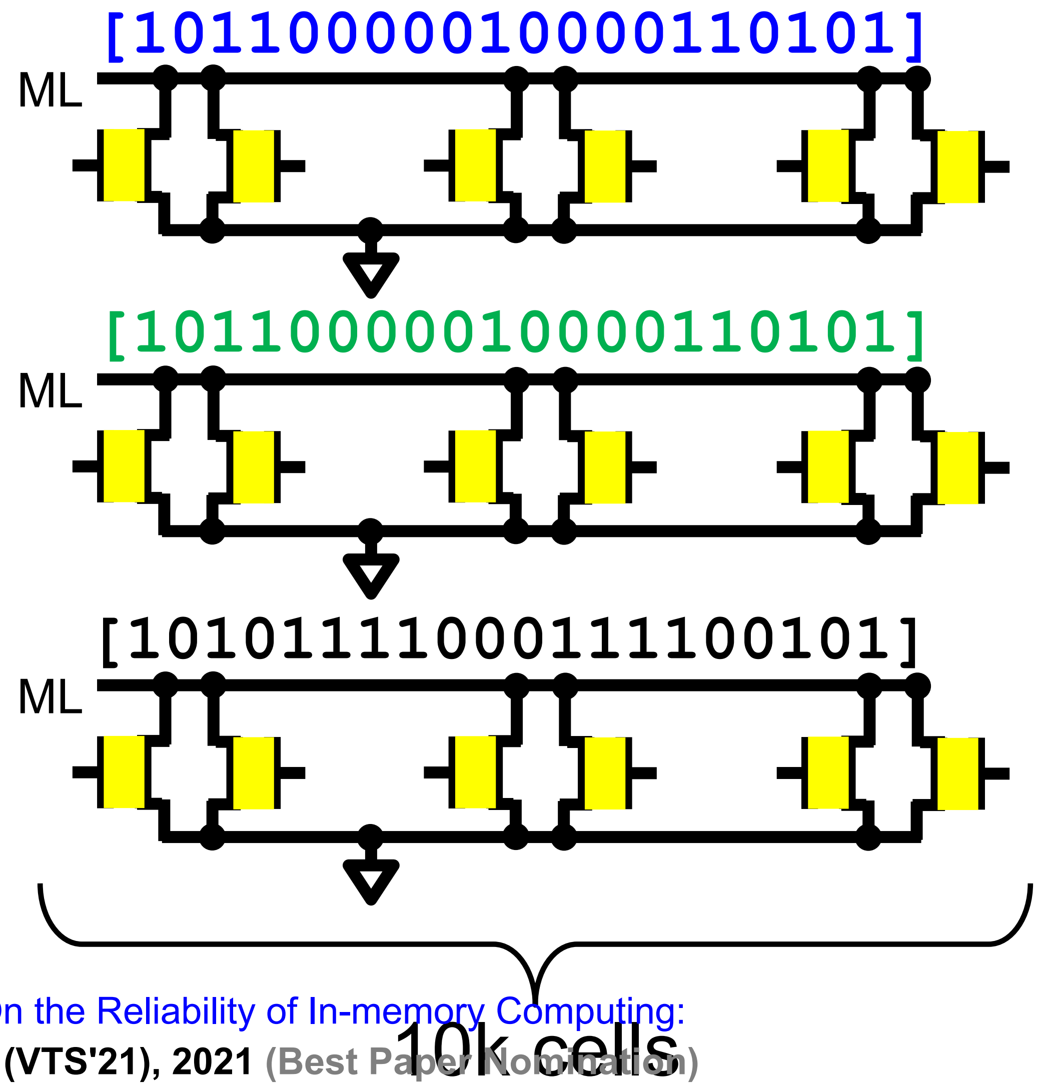


Hamming distance of 2

In-Memory Hyperdimensional Computing

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

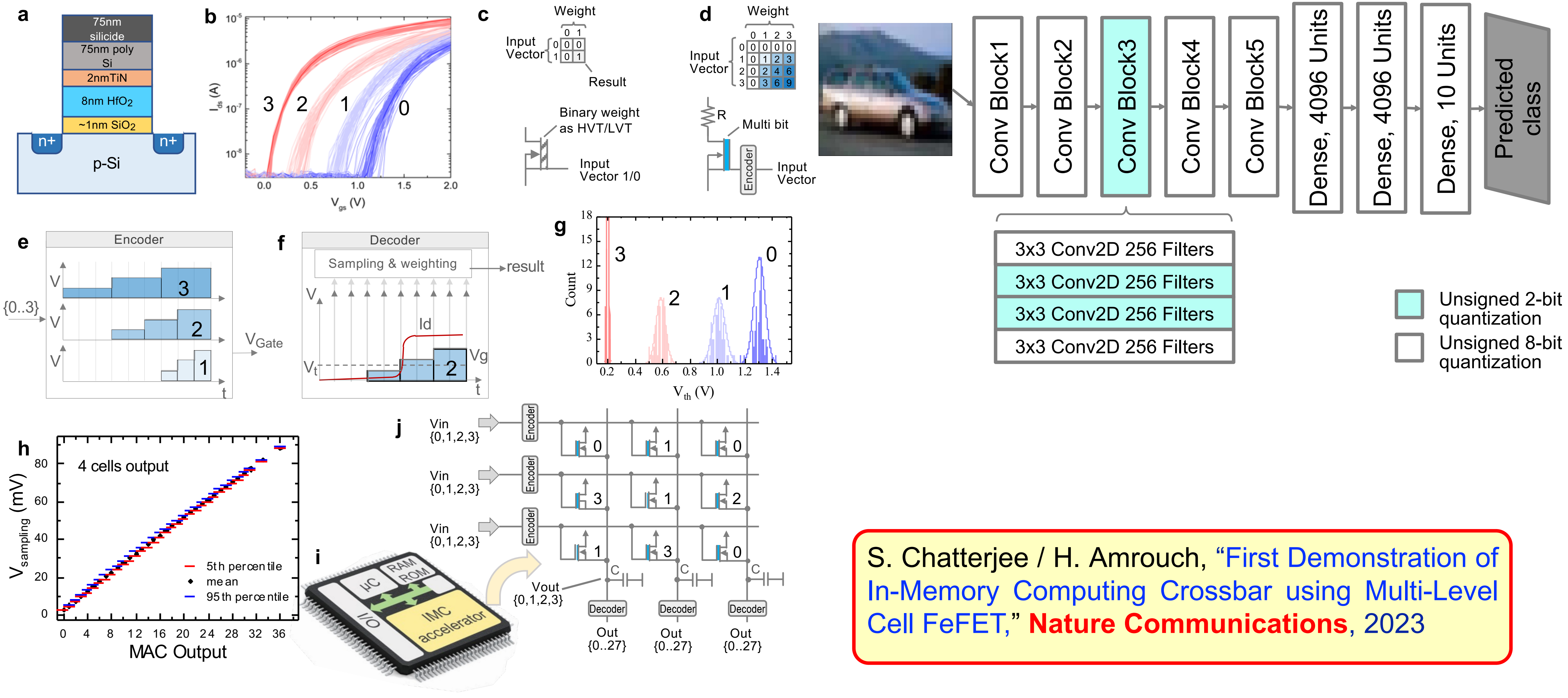




Search = [001010110001001001011]

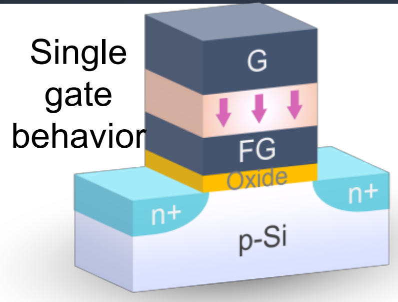
S. Thomann, C. Li, C. Zhuo, O. Prakash, X. Yin, X. S. Hu, and H. Amrouch, "On the Reliability of In-memory Computing: Impact of Temperature on Ferroelectric TCAM," **IEEE VLSI Test Symposium (VTS'21), 2021** **(Best Paper Nomination)**

Very Efficient MAC using FeFET Crossbar

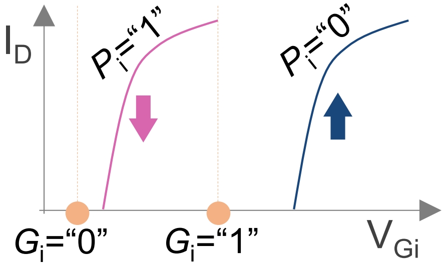


S. Chatterjee / H. Amrouch, "First Demonstration of In-Memory Computing Crossbar using Multi-Level Cell FeFET," **Nature Communications**, 2023

From In-Memory → In-Transistor Computing

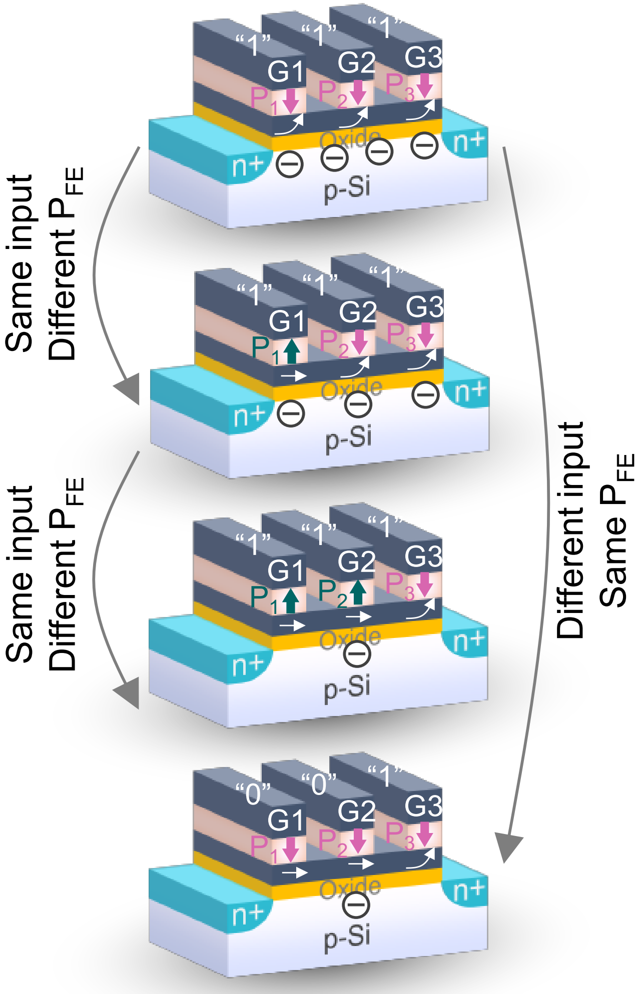


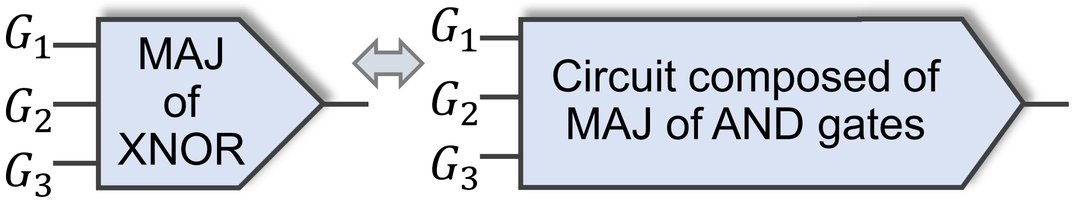
Each gate i performs AND operation: $G_i \land P_i$

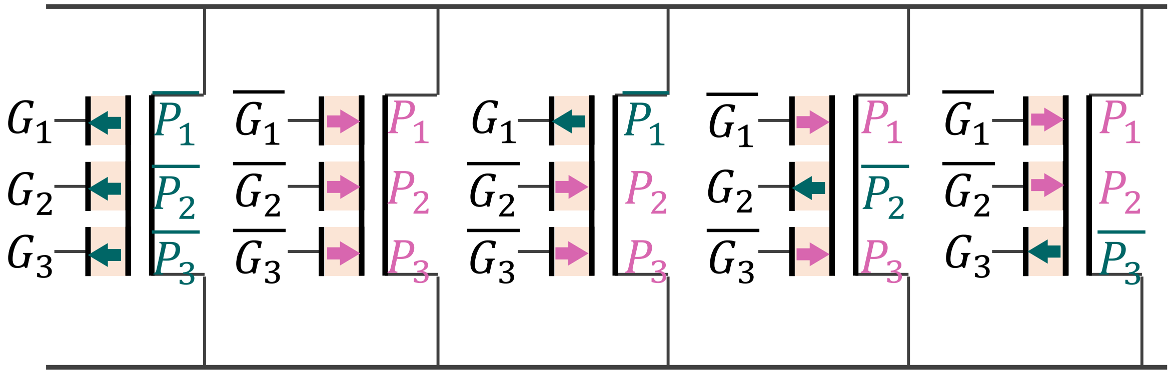


Contribution of gate i to V_{FG}

G_i \ P_i	"1" ↓	"0" ↑
"1" (e.g., 1 V)	High	Low
"0" (e.g., -1 V)	Low	Low

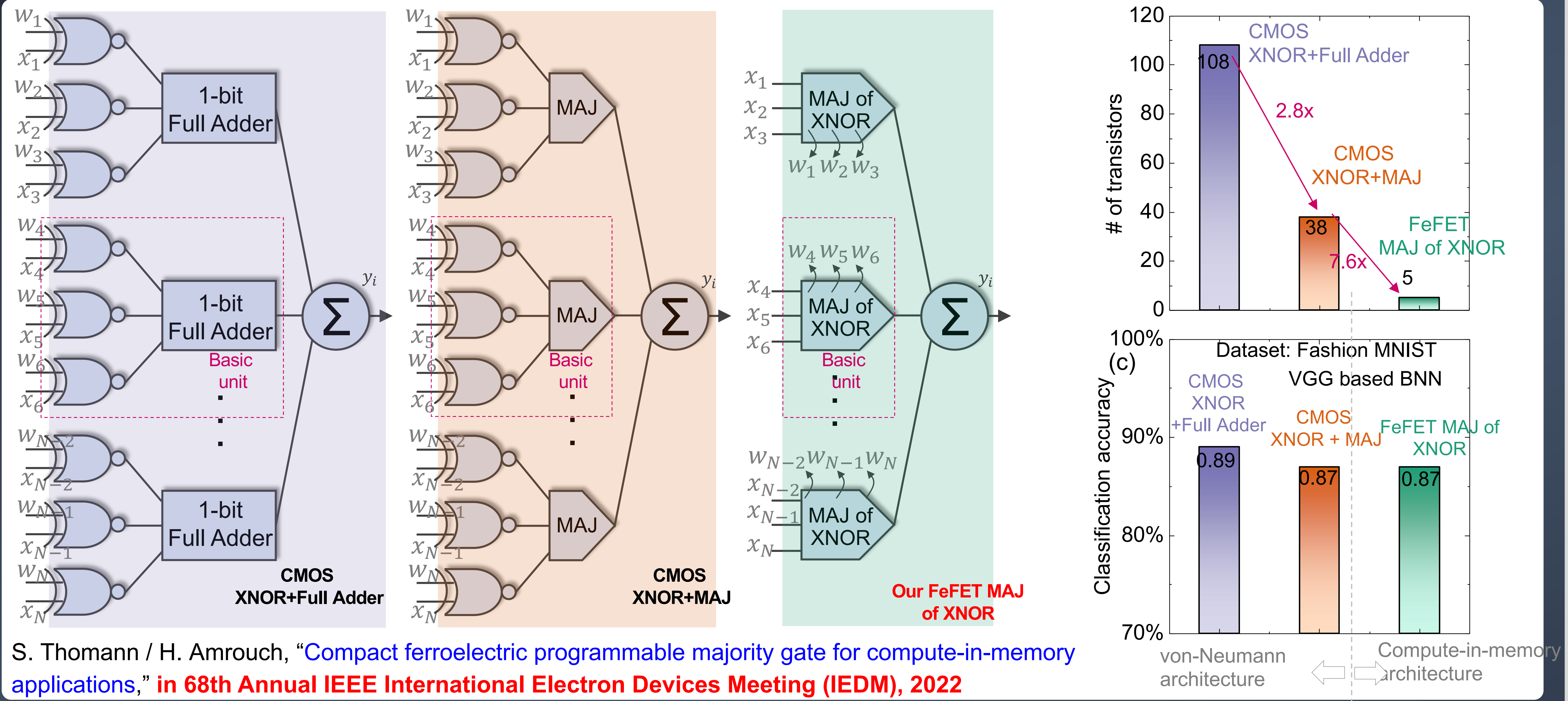






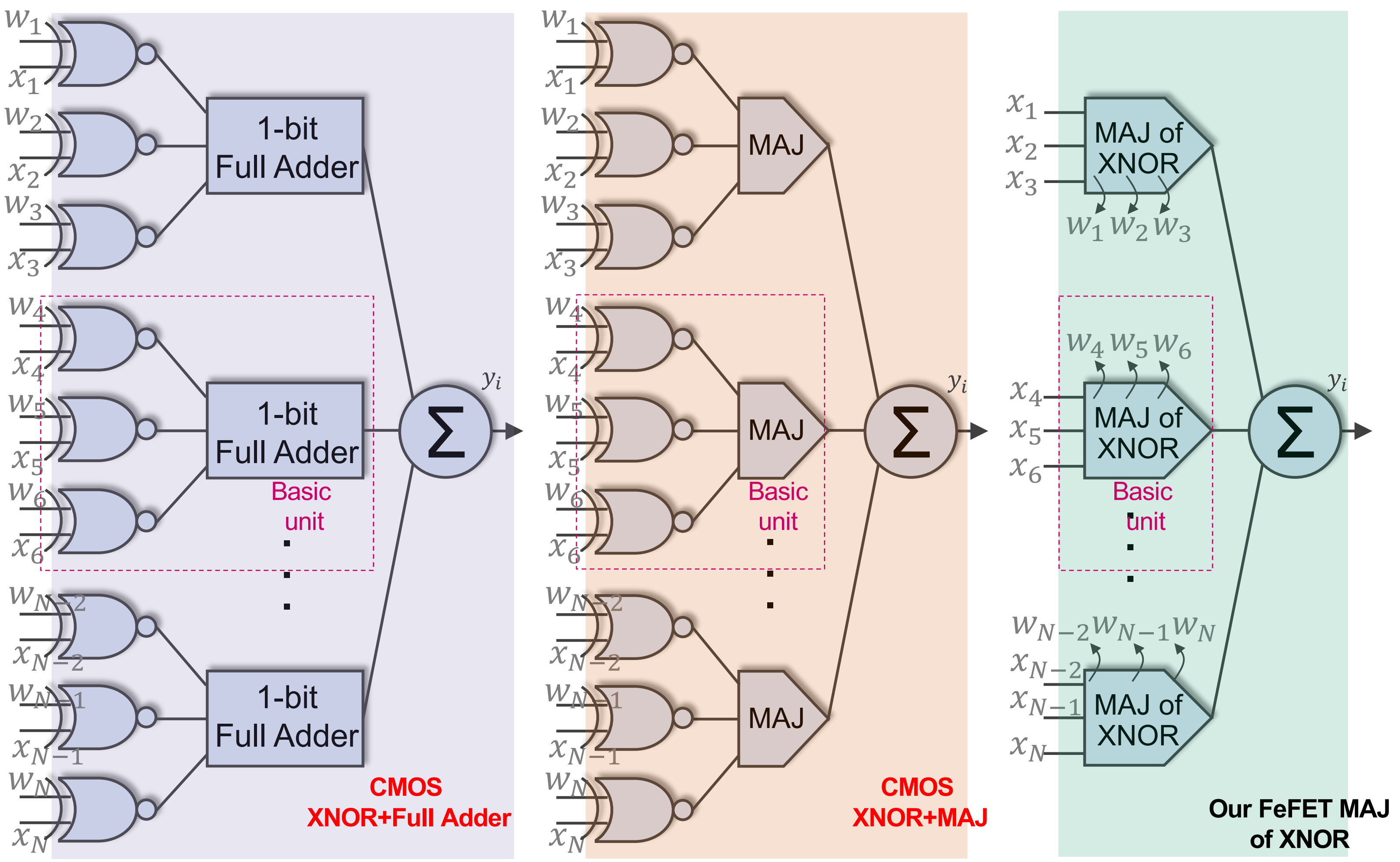
S. Thomann / H. Amrouch, "Compact ferroelectric programmable majority gate for compute-in-memory applications," **in 68th Annual IEEE International Electron Devices Meeting (IEDM), 2022**

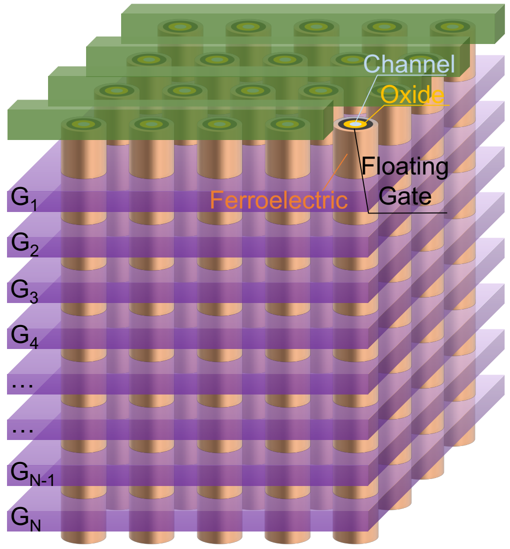
From In-Memory → In-Transistor Computing

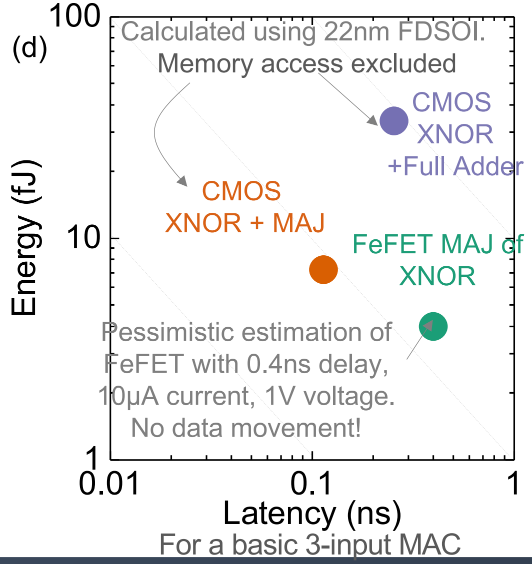


S. Thomann / H. Amrouch, "Compact ferroelectric programmable majority gate for compute-in-memory applications," **in 68th Annual IEEE International Electron Devices Meeting (IEDM), 2022**

From In-Memory → In-Transistor Computing





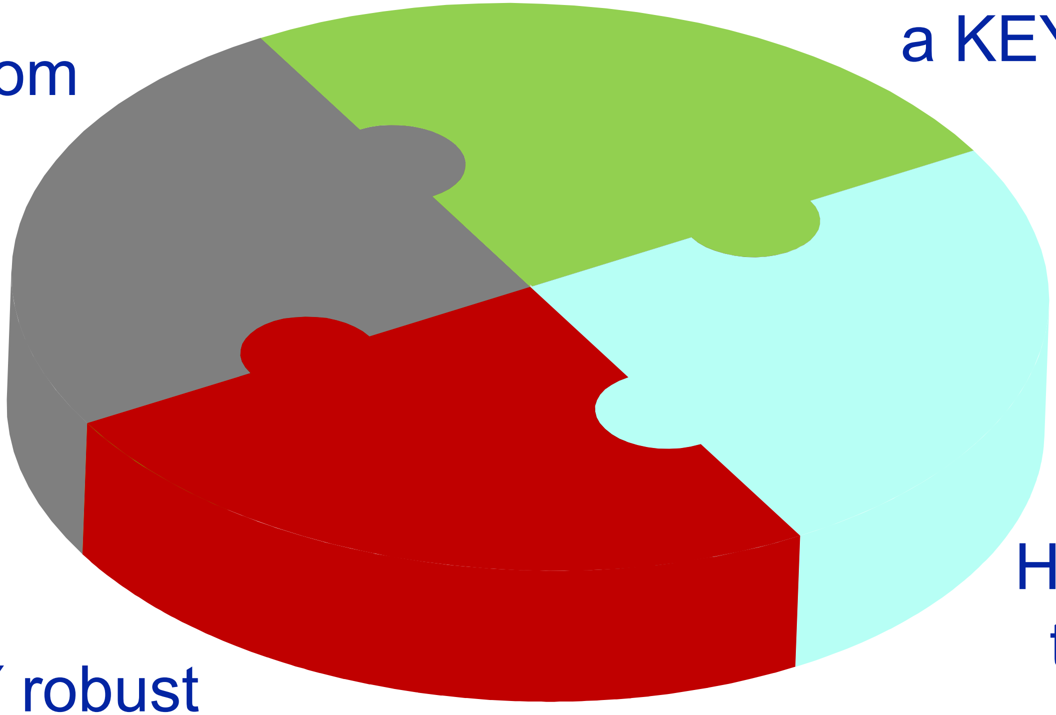


S. Thomann / H. Amrouch, "Compact ferroelectric programmable majority gate for compute-in-memory applications," **in 68th Annual IEEE International Electron Devices Meeting (IEDM), 2022**

HDC Computing … Hope or Hype?

HDC can learn from
little data

In-Memory Computing is also
a KEY for very efficient HDC



HDC is VERY robust
against errors

HDC enables training on
the edge BUT RISC-V
customization
is the KEY

Without them, nothing would be possible

Device / Circuit

Shubham Kumar (PhD)

Swetaki Chatterjee (PhD)

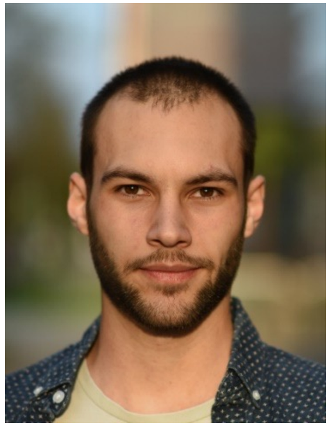
Device / Circuit

Shivendra Parihar (PhD)

Dr. Victor van Santen

Digital Design

Shubham Kumar (PhD)



Simon Thomann (PhD)

Deep Learning

Paul Genssler (PhD)

Rodion Novkin (PhD)

Acknowledgement

SYNOPSYS®

GLOBALFOUNDRIES

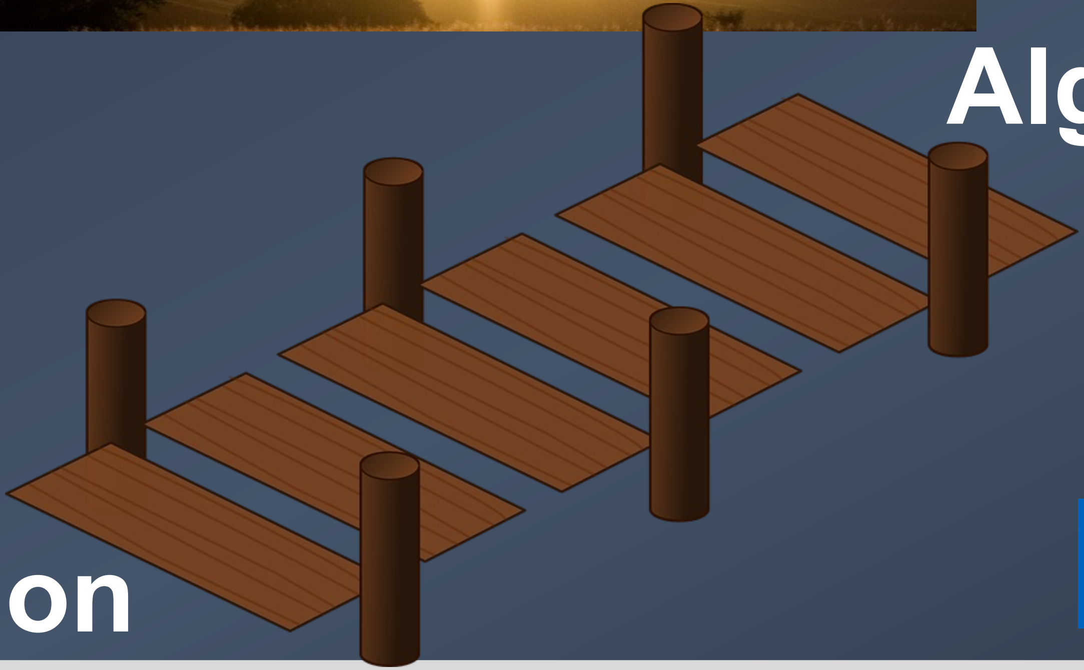
BOSCH

Berkeley
UNIVERSITY OF CALIFORNIA

ADVANTEST®

On the Brink of a new Era in Edge AI

Novel AI Algorithms



RISC-V Customization

Technical University of Munich TUM