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

Complex DNN on one TPUv3:
1.8min ≈ 2048 GPUs + 512 CPUs

Deep Learning

AI Chip: Google TPUv1 [ISCA’17]

Local Unified Buffer for Activations (96Kx256x8b = 24 MiB) 29% of chip
Matrix Multiply Unit (256x256x8b=64K MAC) 24%

Host Interf. 2% Accumulators (4Kx256x32b=4 MiB) 6%
Control 2% Activation Pipeline 6%
PCIe Interface 3% Misc. I/O 1%

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!
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)*.

More Efficient HW for AI

→ Larger and larger AI models

→ Memory Bottleneck!

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

Insufficient On-Chip Memory

Massive Data to Move

von-Neumann Bottleneck

Massive Computation

Excessive Heat

Expensive Cooling

AI Chip: Google TPU [ISCA’17]

Local Unified Buffer for Activations
(96Kx256x8b = 24 MiB)
29% of chip

Matrix Multiply Unit
(256x256x8b=64K MAC)
24%

von-Neumann Bottleneck

Massive Data to Move

Insufficient On-Chip Memory

Excessive Heat

Expensive Cooling

Massive Computation

by H. Amrouch @ TUM Event Hardware
for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design,
Technical University of Munich
Where is the Problem?

Insufficient On-Chip Memory

Massive Data to Move

Von-Neumann Bottleneck

Massive Computation

Excessive Heat

Expensive Cooling

Massive Energy Cost

Local Unified Buffer for Activations (96Kx256x8b = 24 MiB) 29% of chip

Matrix Multiply Unit (256x256x8b=64K MAC) 24%

Host Interf. 2%

Accumulators (4Kx256x32b =4 MiB) 6%

Control 2%

Activation Pipeline 6%

PCIe Interface 3%

Misc. I/O 1%

by H. Amrouch @ TUM Event Hardware for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design, Technical University of Munich
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
But….What is the Root?

- **Voltage:** Reaching its Fundamental Limit
- **Memory:** Massive Data in DNNs
- **Cooling:** Inherently Inefficient
End of Growth of Performance?

40 Years of Processor Performance

- CISC: 2X / 3.5 yrs (22%/yr)
- RISC: 2X / 1.5 yrs (52%/yr)
- Multicore: 2X / 6 yrs (12%/yr)
- Dennard Scaling: 2X / 3.5 yrs (23%/yr)
- End of Moore's Law: 2X / 20 yrs (3%/yr)
- End of Amdahl's Law: 2X / 20 yrs (3%/yr)

By H. Amrouch @ TUM Event Hardware for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design, Technical University of Munich
Automotive Industry needs Innovations in

- Technology
- Memories
- Architectures
- Algorithms
- Cooling
What’s Wrong in Advanced Nodes (7nm, 5nm, 3nm...)

Technology
Reliability is BIG Killer!

Self-heating in 7nm FinFET

Lattice Temperature (K)

Self-heating in 7nm Nanosheet

Self-heating in 14nm Nanowire

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

by H. Amrouch @ TUM Event Hardware for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design, Technical University of Munich
Can we Suppress the Fundamental Limit?

Yes, if $C_{\text{ins}}$ 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)

Impact of NCFET on Google TPU

Replacing CMOS with NCFET

- Not just energy saving
- BUT

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


Design-Time Variation

Run-Time Variation
Many Fundamental Challenges

Reliability Modeling is ESSENTIAL

Technology/Circuit Reliability Modeling

...and HW/SW Co-design is the KEY!

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**

by H. Amrouch @ TUM Event Hardware for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design, Technical University of Munich
Deep Learning is REALLY Power Hungry!

Why not alternative algorithm to Deep Learning?

![Power Density and Temperature Chart](chart.png)

- Power Density [W/cm²]
  - Nominal (0.7V)
  - Turbo-Boost (0.8V)
- Temperature
  - $T_{crit} = 105°C$
- On-Chip Temperature [°C]

by H. Amrouch @ TUM Event Hardware
for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design, Technical University of Munich
Brain-Inspired Hyperdimensional Computing

Example: Language classification

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

\[
\begin{align*}
a &= [10110000010000110101] \\
b &= [10100011011010000001] \\
\vdots \\
! &= [10101111000111100101]
\end{align*}
\]

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

“Hi” $\rightarrow$ [H] XOR [Rotate(i)]
Brain-Inspired Hyperdimensional Computing

Training Text:

```
[0010010000111110001]
[01110100110101111111]
[10100100010111100010]
... 
[10100100010111100010]
```

To be, or not to be

Count 1's

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

Majority gate

```
[1010011000110011001]
```

Entire language is just one Hyper Vector
In-Memory Hyperdimensional Computing

The row with the smallest hamming distance \(\rightarrow\) best match

Why NOT Classical Deep Learning!

Beyond-CMOS Technology + Beyond-von-Neumann

→ HW Errors are inevitable!

Robust AI Algorithms against Errors is a MUST
On-Chip Cooling

Algorithm

Architecture

Memory

Technology

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

Intelligent Cooling: Superlattice Thermoelectric

Traditional Cooling
Cooling the entire chip ➔ Inefficient

Cooling Wall

AI Chip: Google TPU [ISCA'17]
Intelligent Cooling: Superlattice Thermoelectric

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


Intelligent Cooling: Superlattice Thermoelectric

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

On-chip cooling: localized and On-demand $\rightarrow$ VERY Efficient

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

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

We developed the First Printed Thermal Sensors Array for Processor Chips

We developed the First Printed Thermal Sensors Array for Processor Chips

11 x 11 = 121 Thermal Sensors

21 x 21 = 441/cm² Thermal Sensors

Ack: Heidelberg InnovationLab, U. Lemmer, KIT
Intelligent Temperature Sensing

We developed the First Printed Thermal Sensors Array for Processor Chips.

\[ 21 \times 21 = 441 \text{ cm}^2 \]

\[ 11 \times 11 = 121 \text{ Thermal Sensors} \]

\[ \text{Full Thermal Map at Run-time is a Key} \]

\[ \text{Measured Temperature} \quad \text{Predicted Temperature} \]

by H. Amrouch @ TUM Event Hardware
for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design,
Technical University of Munich
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
- Conventional Deep Learning is Hungry → Novel Algorithms
- Temperature is KILLER → On-Chip Cooling
Funding Agencies

- Bundesministerium für Bildung und Forschung
- Office of Naval Research
- ADVANTEST
- GS-IMTR
- DAAD
- DFG

by H. Amrouch @ TUM Event Hardware for AI, E-mail: amrouch@tum.de

Chair of AI Processor Design,
Technical University of Munich
On the Brink of a new Era in AI Hardware

Device Physics

Computer Science