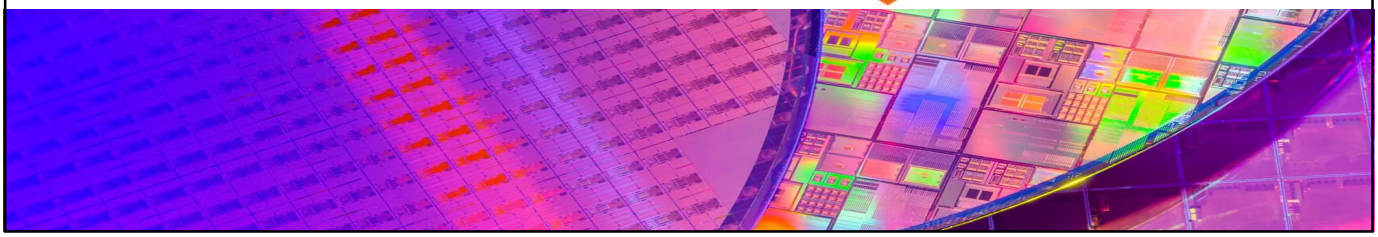
The Need of New Computing Hardware for a Sustainable World

Said Hamdioui
Quantum and Computer Engineering Department
Delft University of Technology & Cognitive-IC
The Netherlands

TU Delft

High Tech NL Semiconductors



1

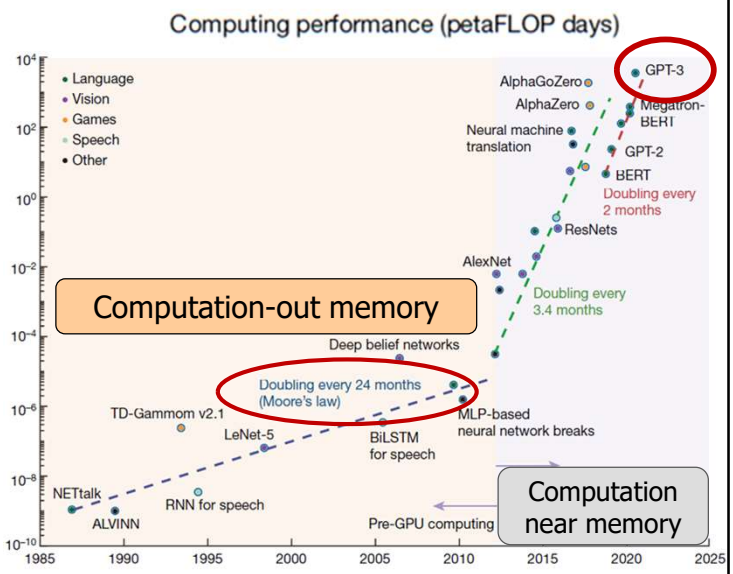
Outline

- Why the need for a new computing chips/ paradigm?
- Why can today's technologies and computers not deliver the needs?
- What are the potential solutions?
- What we do at TU Delft wrt (computing) chip design?
- Conclusion

2

Why the need for a new computing paradigm?

- Demand for computing power
  - Enable new applications/markets (e.g. AI)
  - Flops (e.g., Peta= 10^{15} ops/s)
- Initially driven by Moor's law
  - Uni- and multi-core
  - Tech scaling (~43%)
  - Smart architectures & SW-HW co-design
- From 2010 on
  - GPUs (+HBM)
  - Increase in data size
  - Doubling every 2 to 3.4 months



[Source: K. Boahen, "Dendrocenric learning for synthetic intelligence". Nature 2022]

Demand for FLOPS is increasing rapidly

Why the need for a new computing paradigm?

9.200 GPUs for 2 weeks[1]

1 training: 1.287MWh[2]

Enough for average household for 120 years

Carbon release of 1300 cars in the same period[3]

240 -340 TWh a year[4]

Expected to be 10X more by 2030[5] (~30% ww demand when considering ICT)

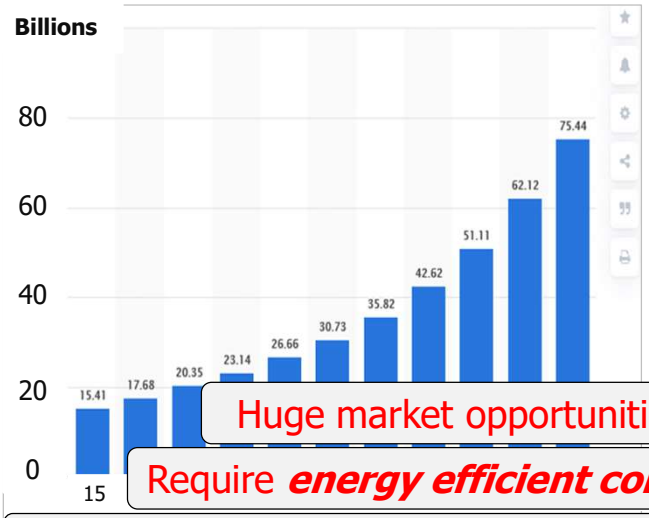
Emits ~ 1.2M tonnes of CO2 a year (0.3% v 2% ww when considering ICT)

Unsustainable growth of energy use!
Unstainable increase of carbon release!

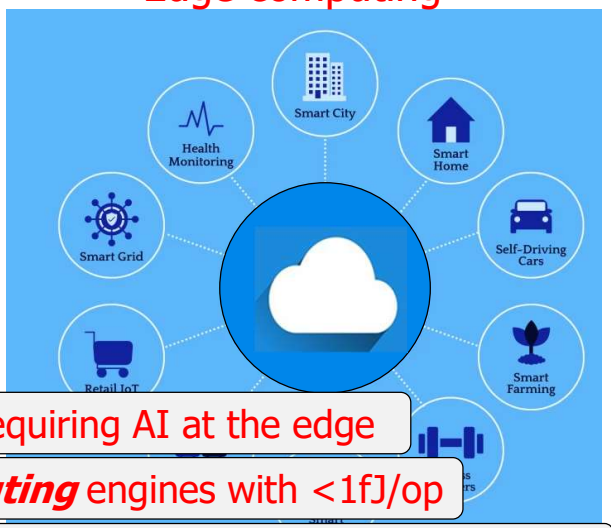
1: Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33,1877–1901
2: D. Patterson, et.al, The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink, arxiv (2022)
3: L.F.W. Anthony, et.al, "tracking and predicting the carbon footprint of training deep learning models". arxiv 03051 (2020)
4. International Energy Agency (*IEA) analysis, 2023*.
5: N. Jones, "How to stop data centres from gobbling up the world's electricity. Nature, (12 September 2018).

Why the need for a new computing paradigm?

IoT connected devices ww



Edge computing



Huge market opportunities requiring AI at the edge

Require ***energy efficient computing*** engines with <1fJ/op

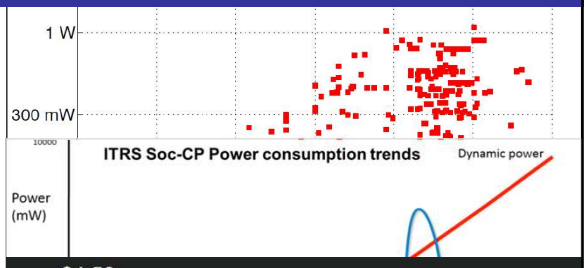
E.g., For GPT-3 to converse on ur phone: each flop @ >15X **faster** and @ >170 X less **energy**

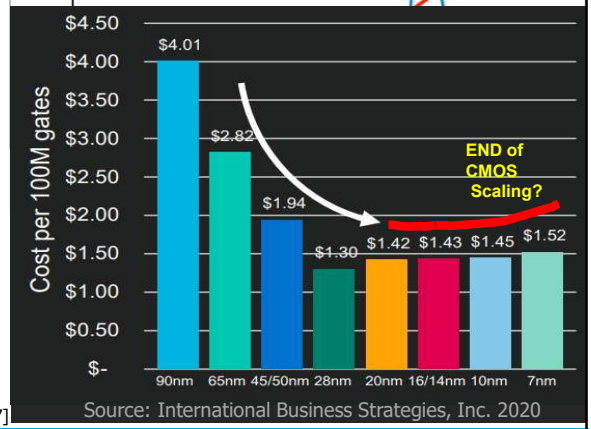
[Source: K. Boahen, "Dendrocenric learning for synthetic intelligence", nature 2022]

Why can today's technologies and computers not deliver the needs?

- Scaling wall
  - Past: constant **field**
    - 2x density & higher frequency @ same power/chip
    - Freq: 1/0.7 = **1.43; P=** C x V^2 x f = **0.5**
  - Recent: slowed constant **voltage**
    - Freq= constant; **P=1**
    - **Leakage & reliability**
- Cost wall
  - Complex manufacturing
  - Expenses fabs
  - Low yield

Less/no economical benefit
Need for new device technologies





Source: International Business Strategies, Inc. 2020

[Ref, S. Hamdioui, et. al, DATE 2017]

Why can today's technologies and computers not deliver the needs?

1. Memory Wall
- Limited bandwidth/ Communication bottleneck
- Stored program principle

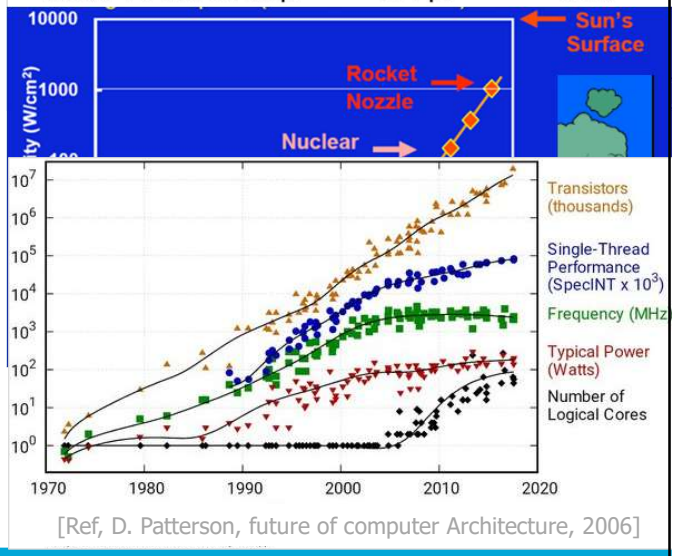
2. Power Wall
- Practical power limit for cooling; Dark Silicon
- Dominated by com & memory

3. ILP Wall
- Insufficient parallelism at instr. level
- Programmability Complexity & overhead

=> Reduced / Saturated performance
- Enhancement based on on chip memory
- **Requires LD & ST**: killers of overall perf.

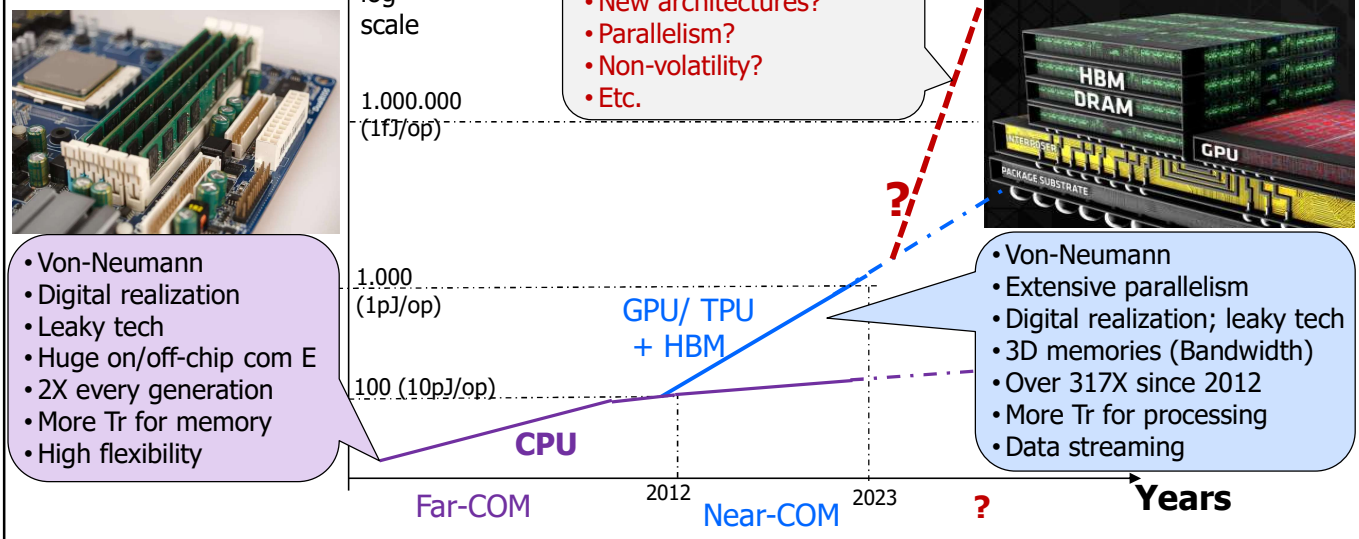
Need of new architectures?

Operation (8-bit operand)	Energy/Op (45 nm)	Cost (vs. ALU)
ALU operation	0.05 pJ	1X
Move 10 mm on-chip	2.4 pJ	50X

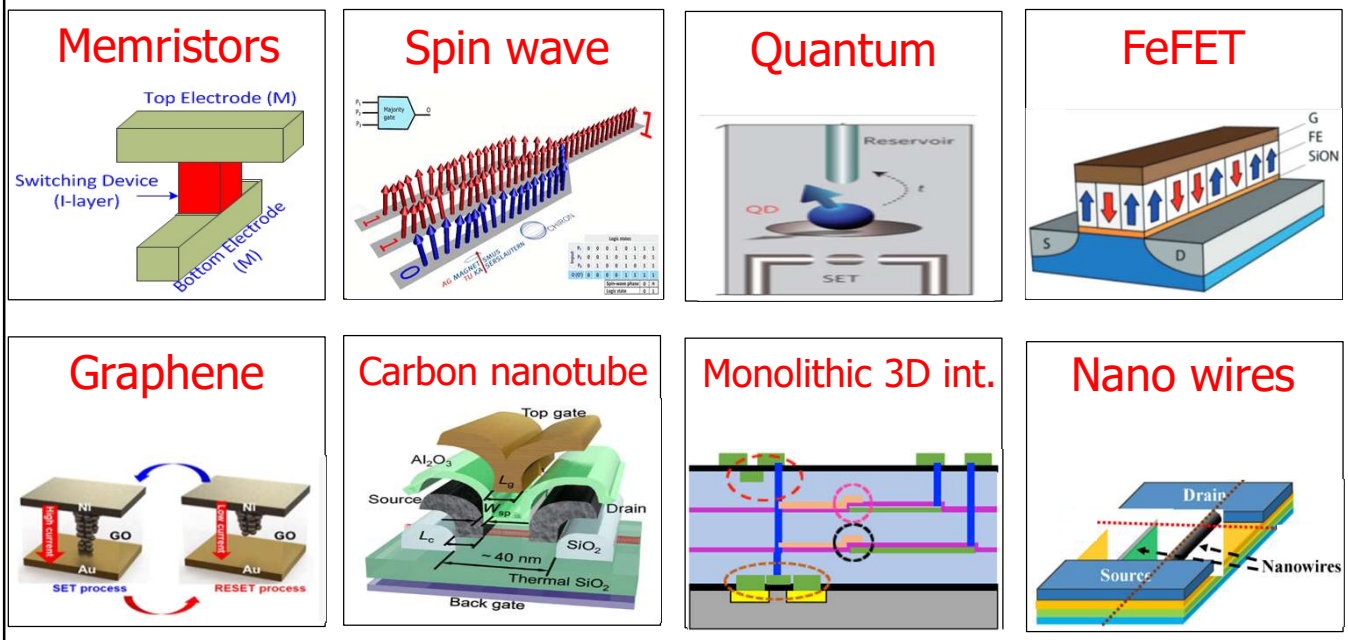


[Ref, D. Patterson, future of computer Architecture, 2006]

What are the potential solutions?



What are the potential solutions? Device Tech examples

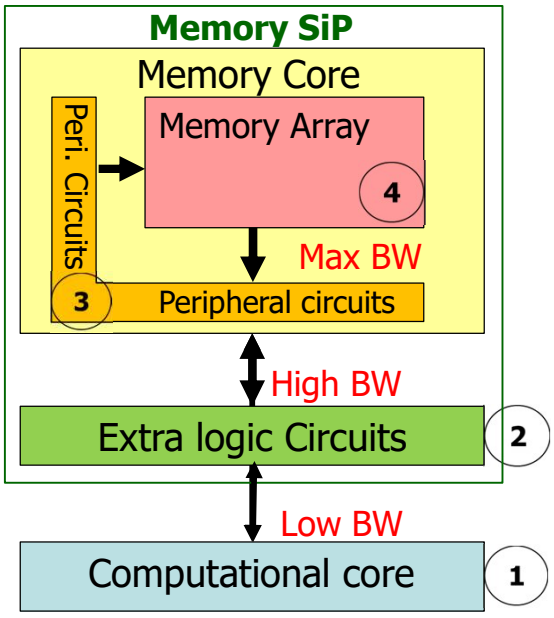


9

What are the potential solutions? Architectures (memory centric view)

- **COM: Computation-Out-Memory**
  1. Far (COM-F)
  2. Near (COM-N)
- **CIM: Computation-In-Memory**
  3. Periphery (CIM-P)
  4. Array (CIM-A)
- **Hybrid architectures**
- **Status**
  - COM: commercialized & conv technologies
  - CIM: Research, conv & unconv technologies

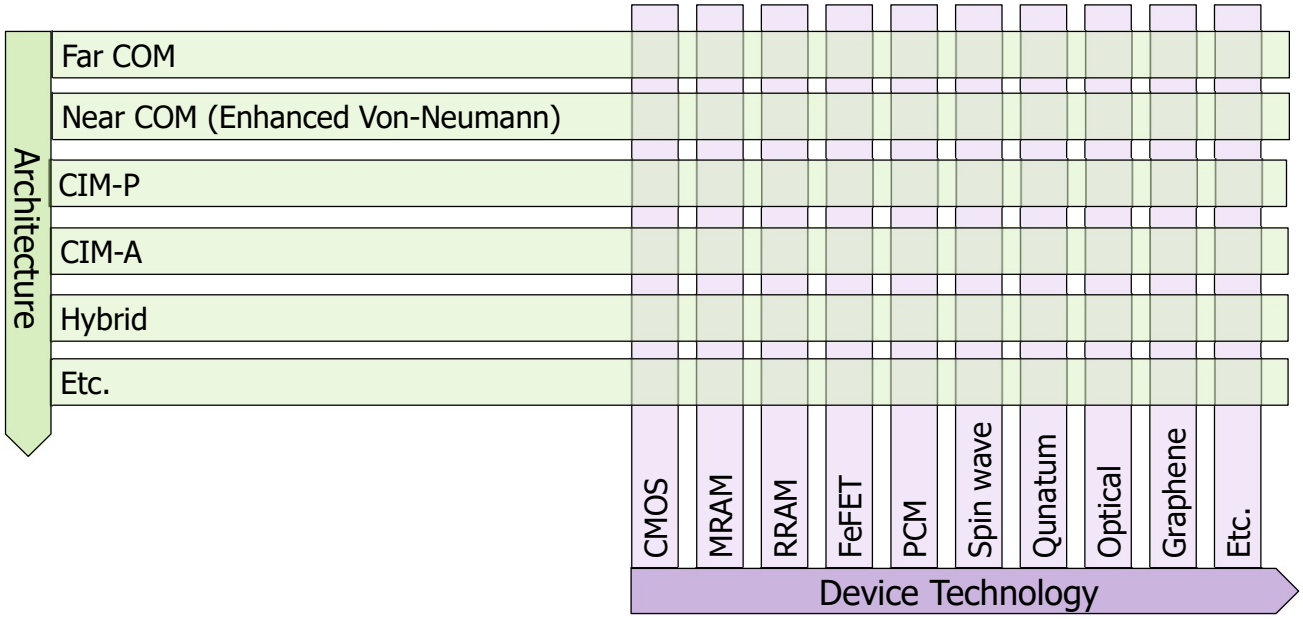
Need CIM to reduce data movement



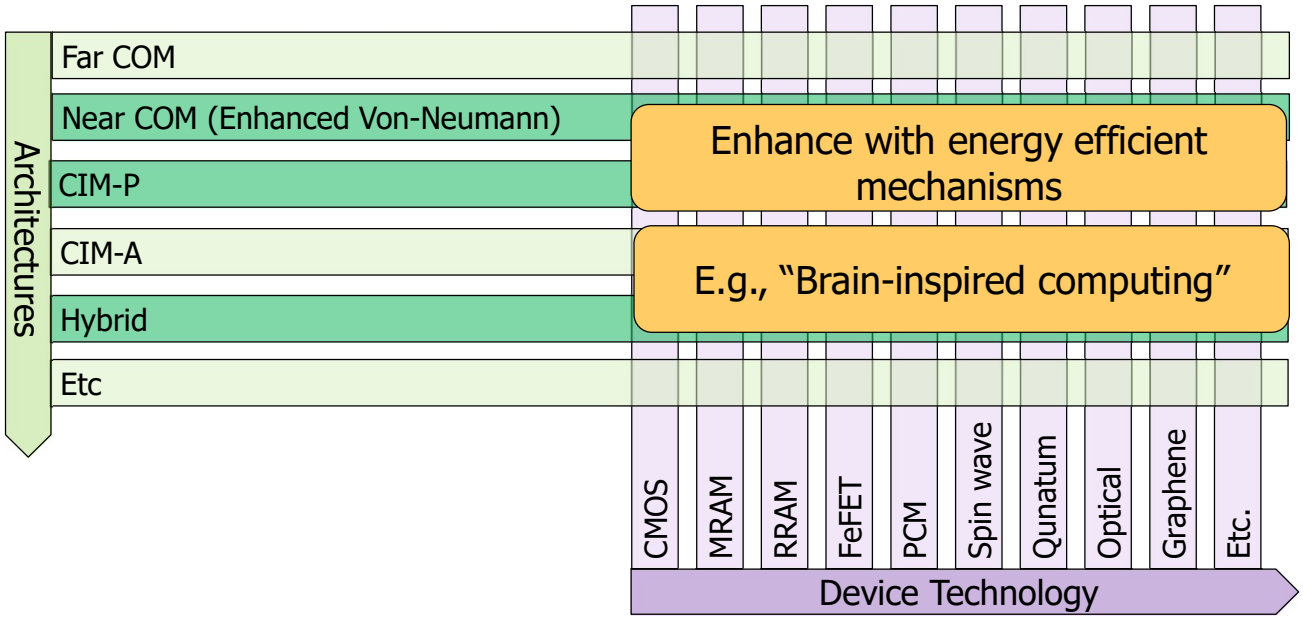
[Source: H.A. Du Nguyen, et. al, "A classification of memory-centric computing" ACM JETC, 16(2), pp.1-26", 2020]

10

What are the potential solutions?



What are the potential solutions? TUD



What we do at TUD? Themes

TUDelft
- Bio-signal acquisition
- Computing/ Signal proc.
- **IC Design**
- Micro Fabrication
- **Design for X**

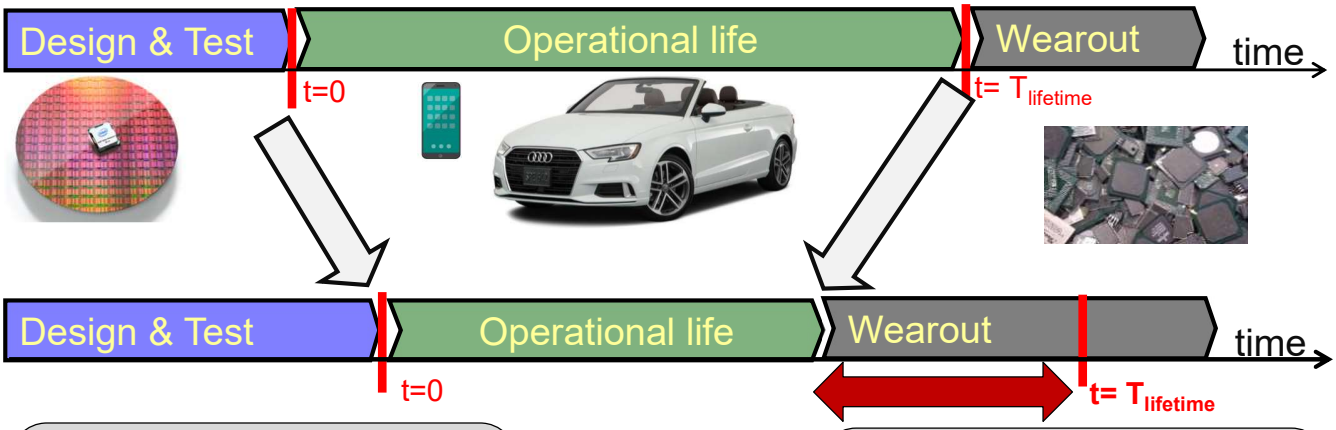
TUDelft
- Sensor design
- Computing/ Signal proc.
- **IC Design**
- Micro Fabrication
- **Design for X**

TUDelft
- Microwave systems
- THz systems
- Computing/ Signal proc.
- **IC Design**
- Micro Fabrication
- **Design for X**

TUDelft
- Design for security
- Design for safety
- Computing/ Signal proc.
- **IC design**
- Micro fabrication
- **Design for X**

13

What we do at TUD? Design for X



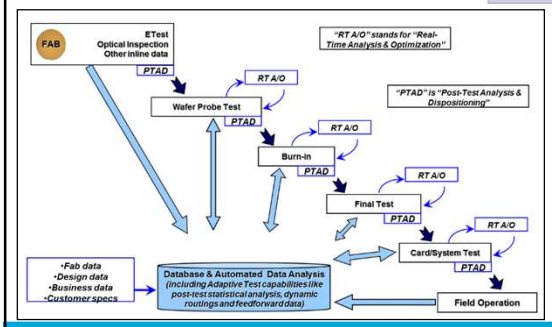
Technology trends & business pressure are redefining the Design, Test & Reliability paradigms

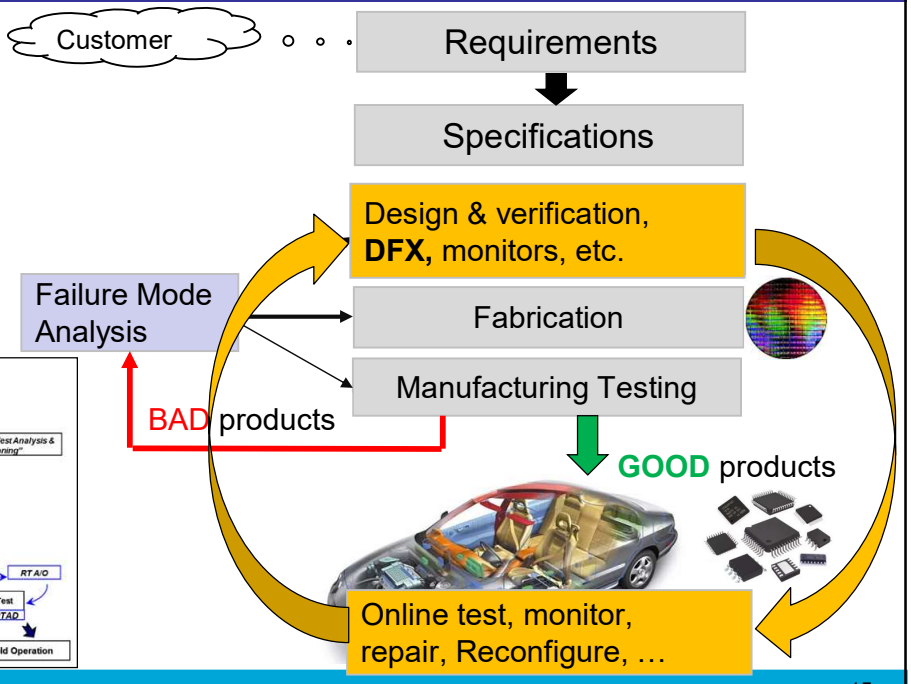
- Reduce **failure rate**
- Extend **lifetime**
- Shorten TTV/ **TTM**

14

What we do at TUD? Design for X

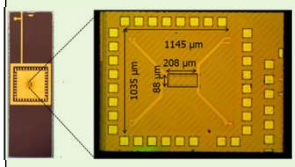
- From idea to shipping
  - Reduce **failure rate**
  - Extend **lifetime**
  - Shorten TTV/ **TTM**





What we do at TUD? Computing

Array-periphery codesign of CIM

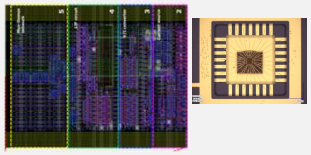


115 TOPS/W

8.7 fJ/ oper.

[Source: A 115.1 TOPS/W, 12.1 TOPS/mm2 Computation-in-Memory engine for Edge AI, AICAS 2023]

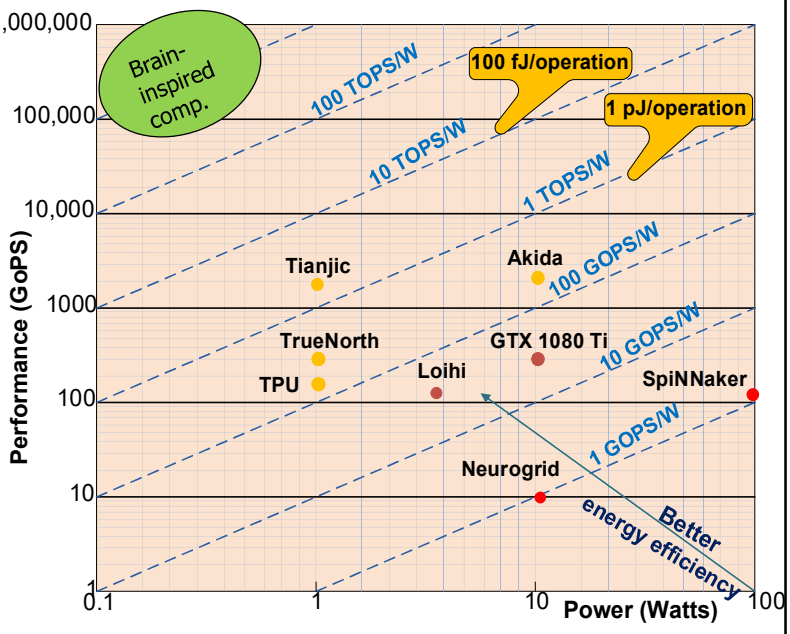
ULP CIM design using CC



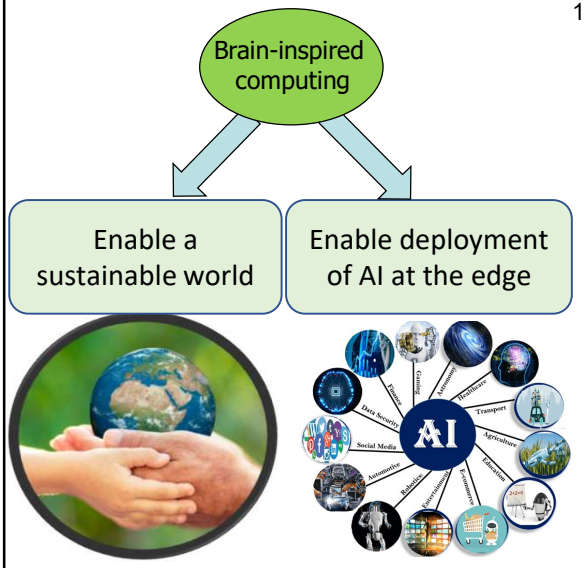
2 POPS/W

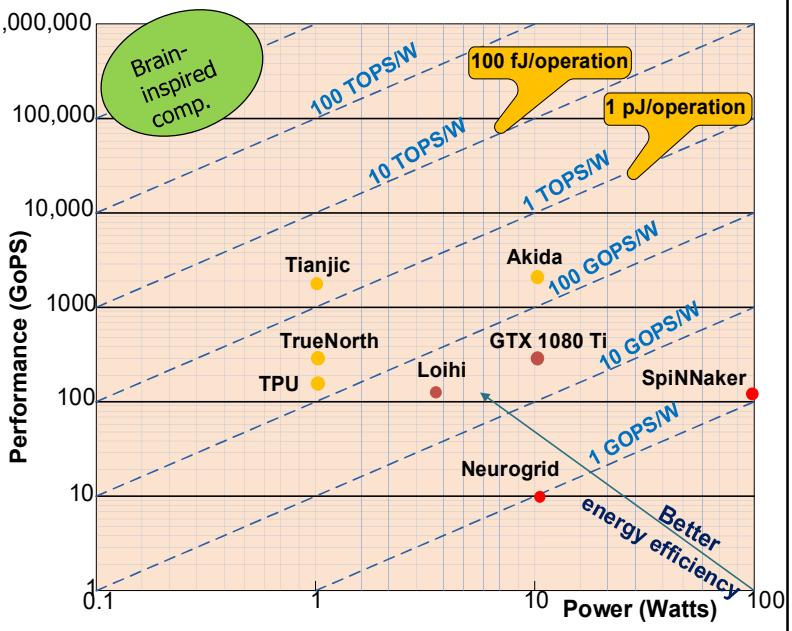
0.5 fJ/ oper.

[Source: ULP Memrisror based CIM microarchitecture using a serial current approach, Patent filed, 2023]



What we do at TUD? Computing





Conclusion

Sustainability, market demands, and reliability call for urgent new energy-efficient and intelligent (computing) chips

Physicists	Chemists	Comp. Scientists
Comp. engineers	Biologists	Neuroscientists
Micro. engineers	Mathematicians	Etc.

Investments!

Research centres!

competence centres!

Thanks

High Tech NL Semiconductors