tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

June 26 - 28, 2023

TINY ML

www.tinyML.org

Should tinyML Processors be Multi-core?

Marian Verhelst (marian.verhelst@kuleuven.be)

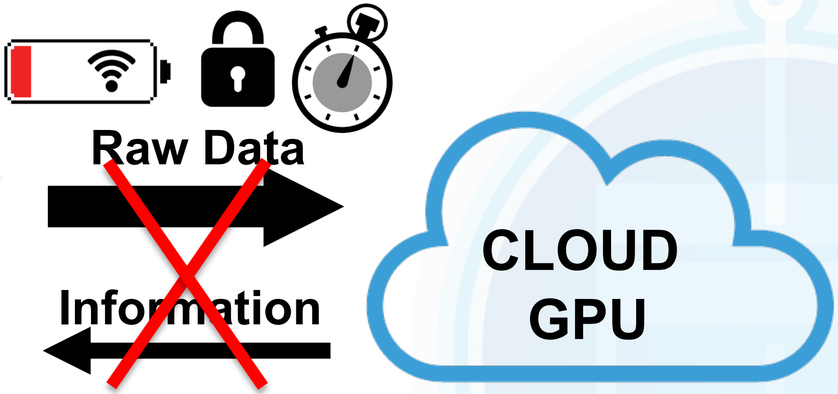
KU LEUVEN

imec

KU LEUVEN micas

Making extreme edge (ExE) devices smart…

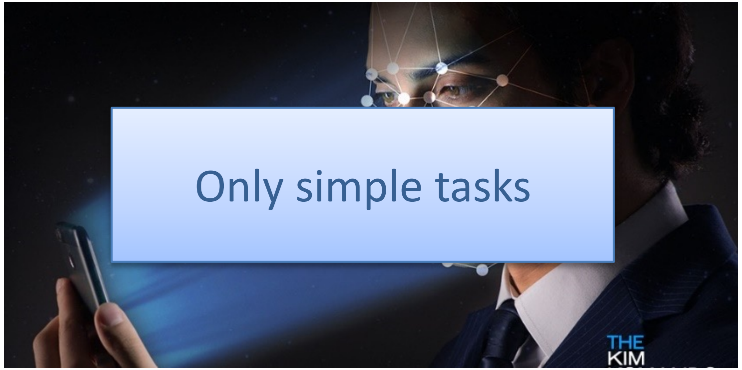
ExE systems = wearables, implantables, smart speaker, drones, cars, …

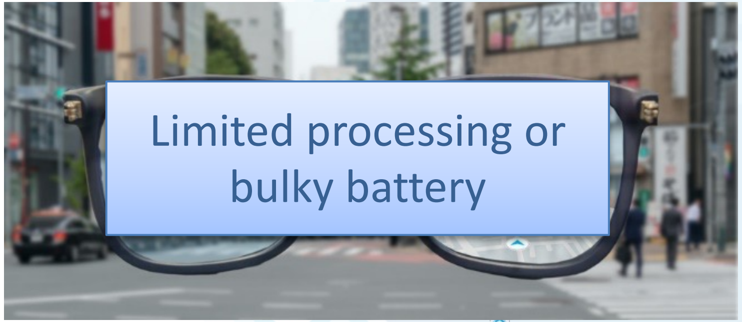


www.tinyml.org

Embedded machine learning
at the extreme edge

Deep neural networks are everywhere in our edge devices… Are they?





Deep neural networks are everywhere in our edge devices… Are they?

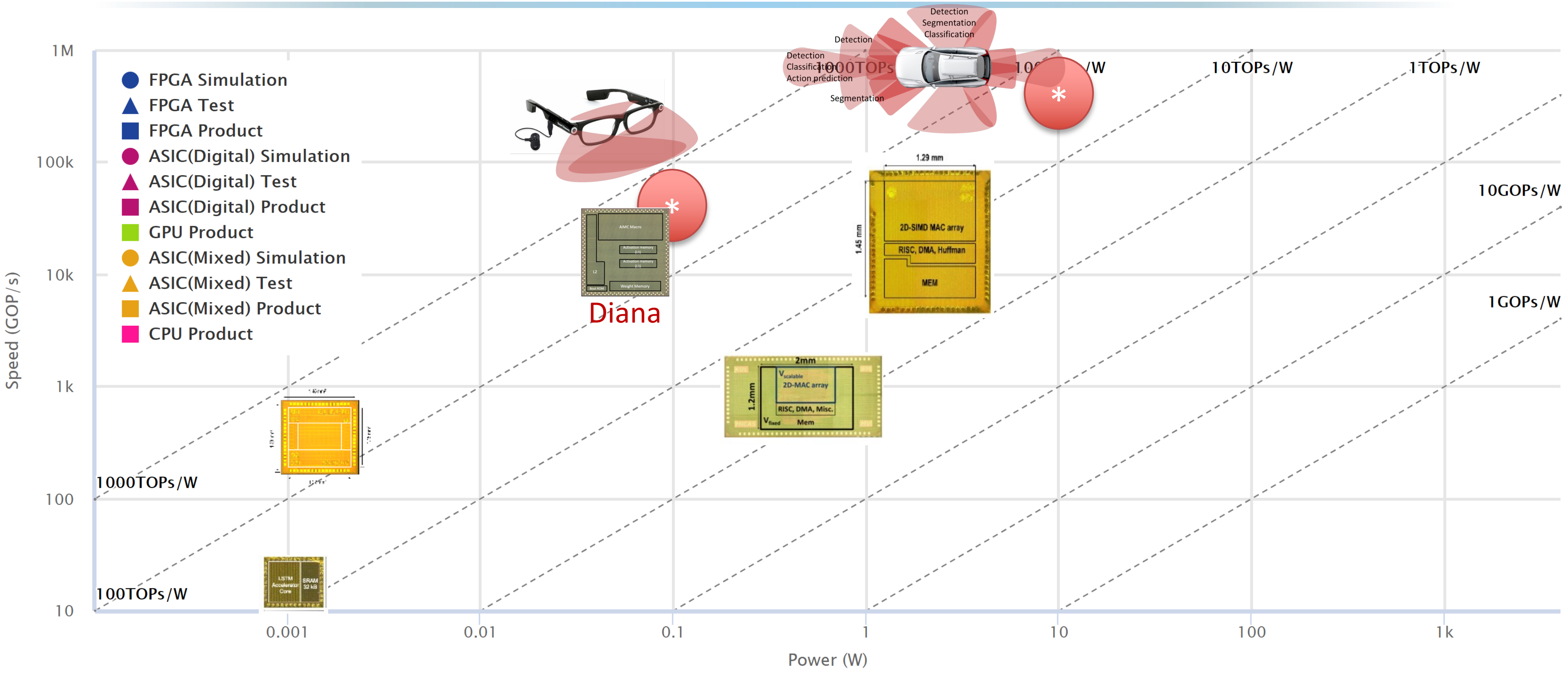
6 full HD cameras @30fps
10Watt, ResNet-50/frame (under est.!)

➔ 1TOPs/frame, **300 TOPs**
➔ 30TOPs/Watt

Stereo HD + eye tracking camera @30fps
100mWatt, ResNet-50/frame (under est.!)

➔ 400GOPs/frame, 30 TOPs
➔ **300TOPs/Watt**

Neural network processors: state-of-the-art



KU LEUVEN

micas

Overview

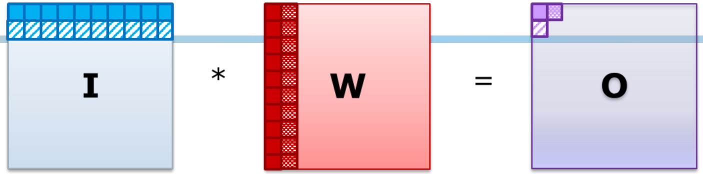
- Peak vs workload performance
  - Dependency on precision
  - Dependency on dataflows
- Motivation for heterogeneous systems
- The Diana system
- Heterogeneous scheduling with ZigZag
- The future?

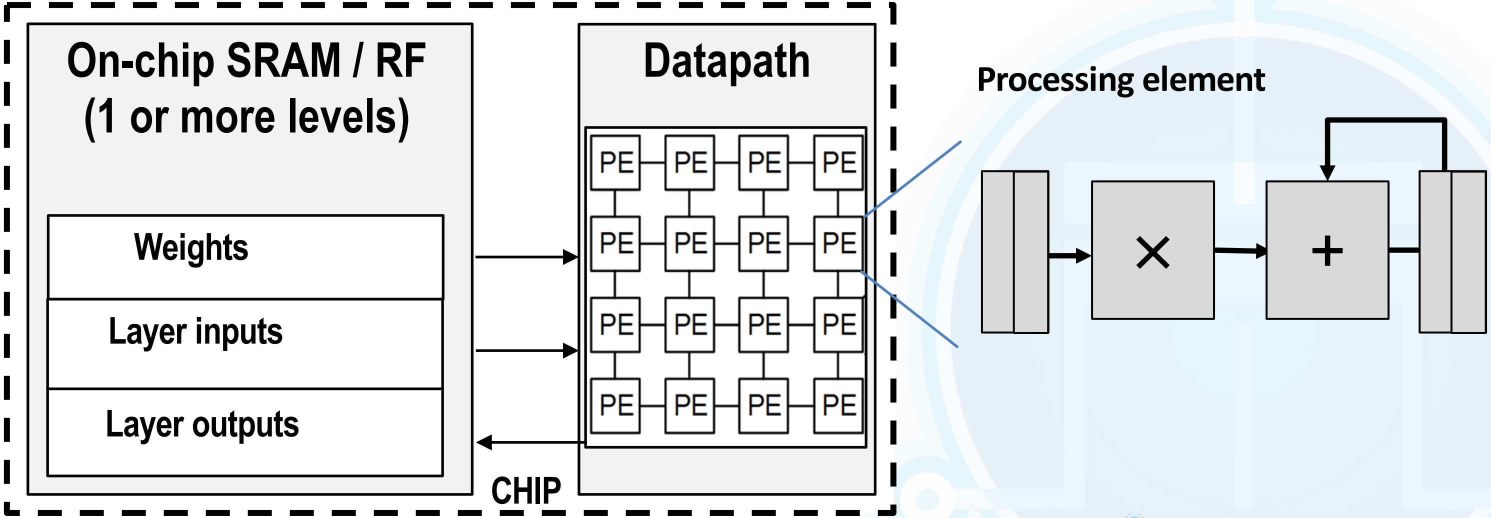
Overview

- Peak vs workload performance
  - Dependency on precision
  - Dependency on dataflows
- Motivation for heterogeneous systems
- The Diana system
- Heterogeneous scheduling with ZigZag
- The future?

A typical Neural (co)processor unit (NPU)

1 neural network layer

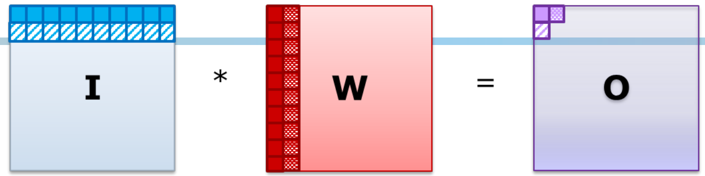




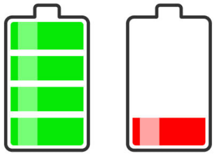
KU LEUVEN

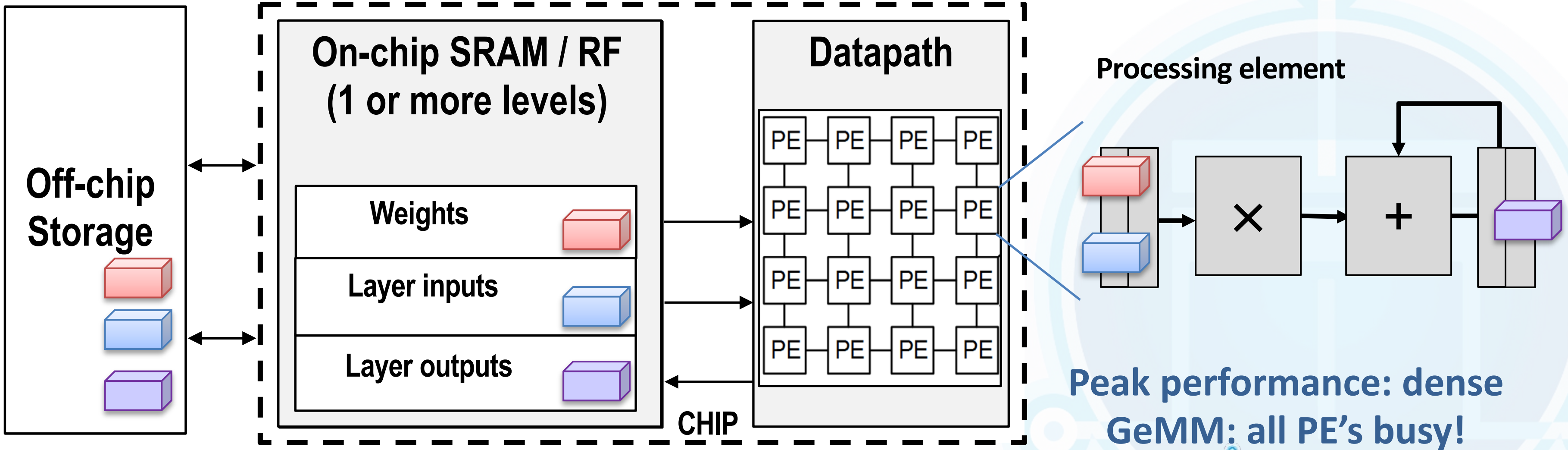
micas

A typical Neural (co)processor unit (NPU)



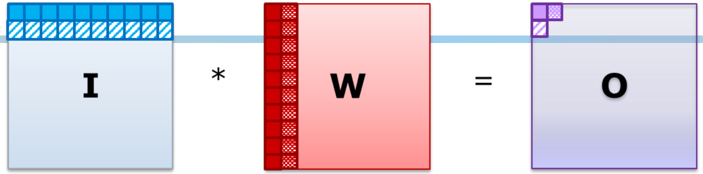
= *Energy per IO transfer* + *Energy per memory read/write* + *Energy per MUL + ADD*



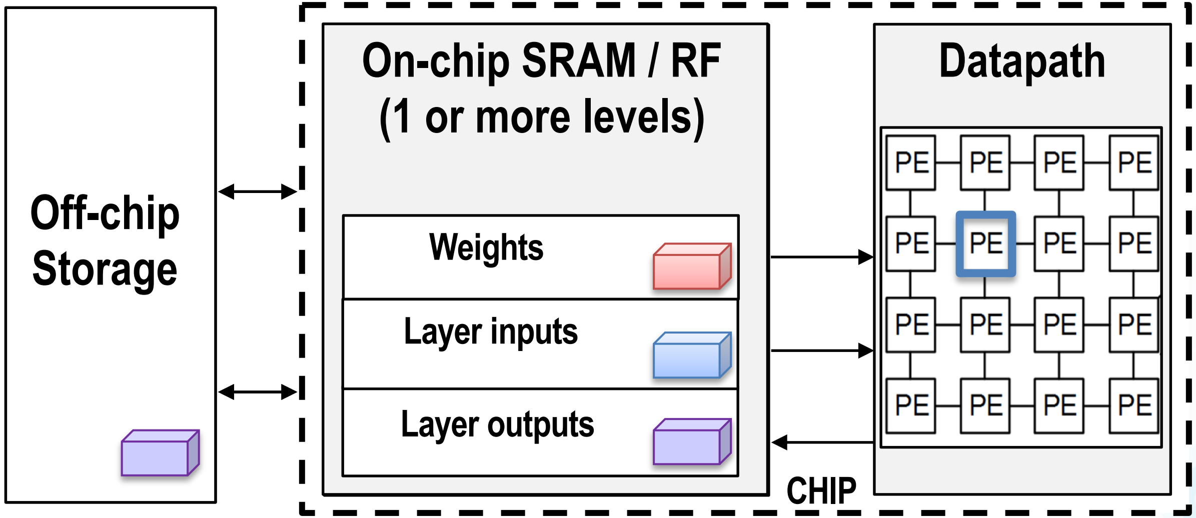


Peak performance: dense GeMM: all PE's busy!

"Trick" 1: Reduced precision in ML processors



= *Energy per IO transfer* / **P** + *Energy per memory read/write* / **P** + *Energy per MUL + ADD* / $\mathbf{P^{1.5}}$



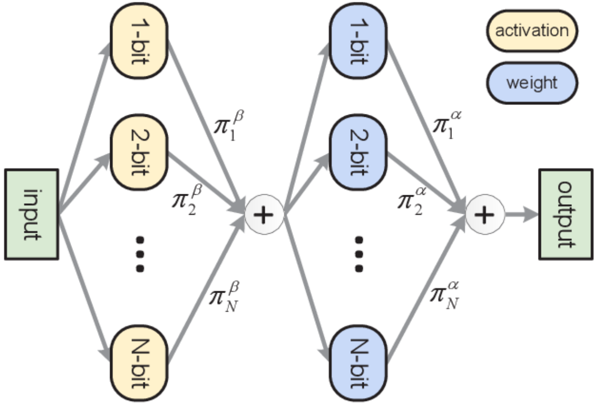
- Compute at 1-2-4 bit precision
- Exploit **precision** to reduce **memory & compute energy**

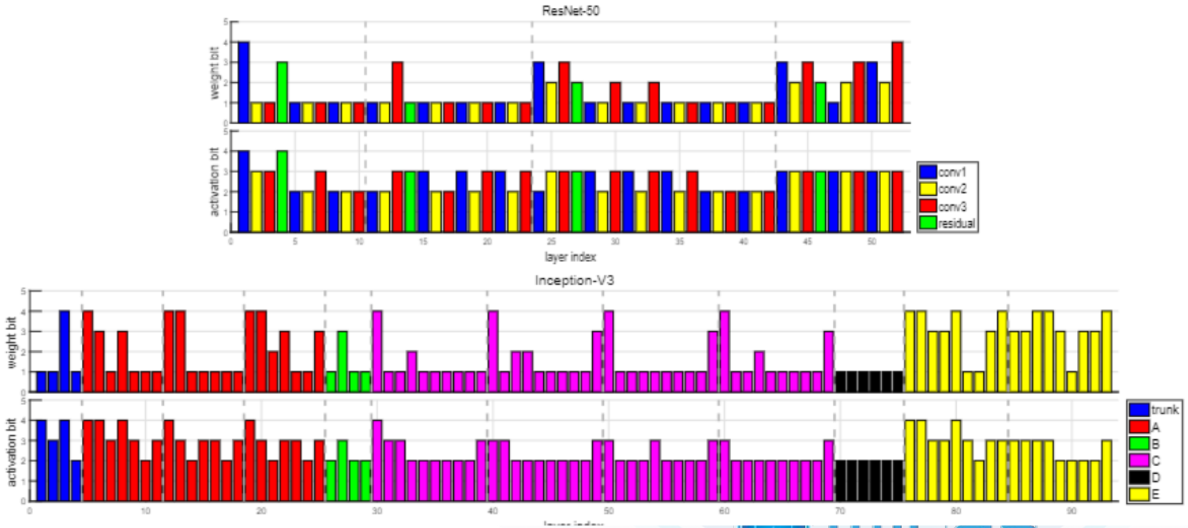
Peak performance ~ 1/P!

"Trick" 1: Reduced precision in ML processors

- Active field of research in algorithmic community
- Promising results!
- Hardware support needed

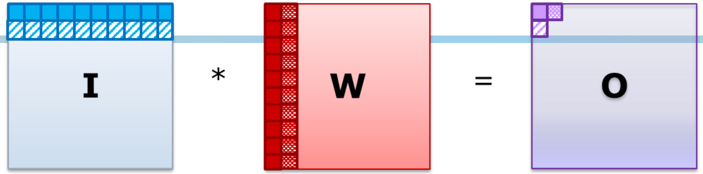
Differential architecture search



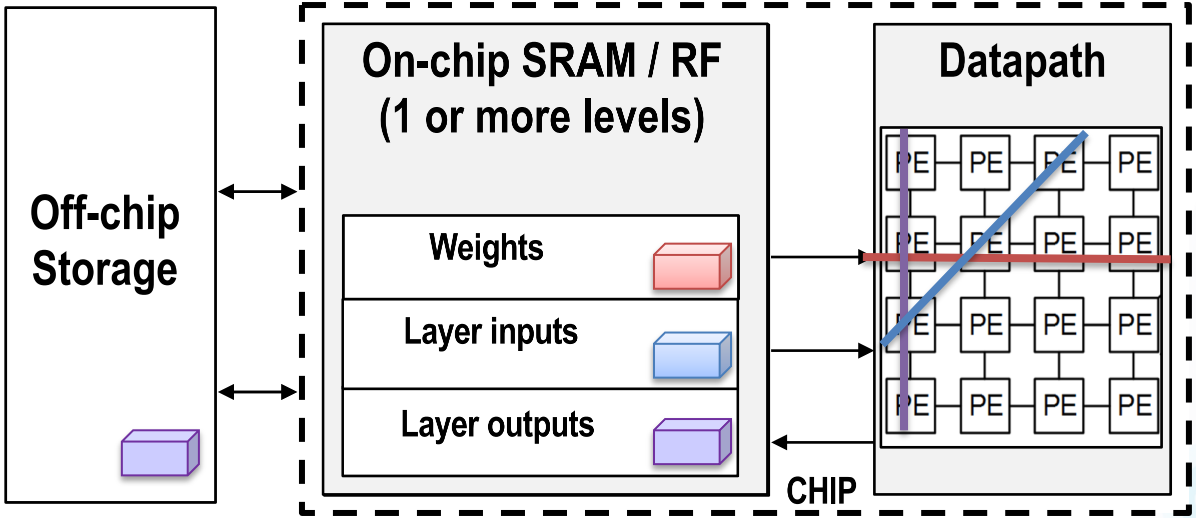


Z. Cai, N. Vasconcelos. "Rethinking Differentiable Search for Mixed-Precision Neural Networks." *CVPR*, 2020 2346-2355.

“Trick” 2: Data reuse in ML processors



= *Energy per IO transfer* **/N’** + *Energy per memory read/write* **/N** + *Energy per MUL + ADD*



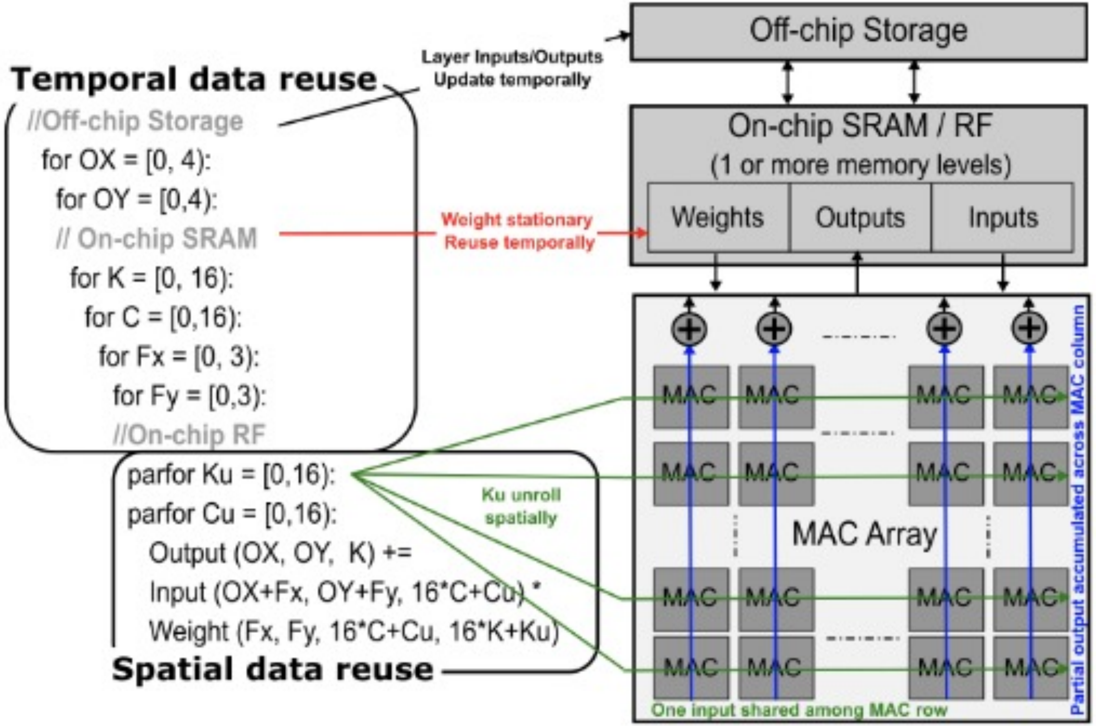
- Remember: every W & I used multiple times, and O accumulated!
- Exploit **data reuse** to reduce **memory energy**

Peak performance:
all data dimensions reused!

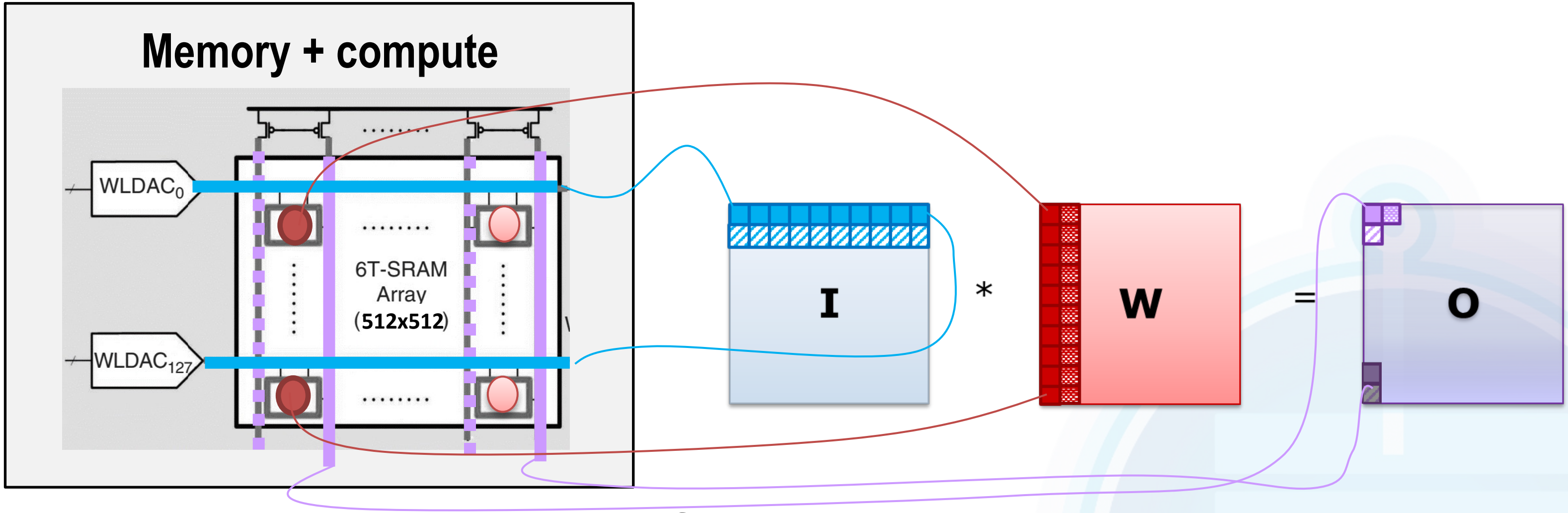
KU LEUVEN

Data reuse in ML workloads

- Convolutions and GeMM allow significant data reuse:

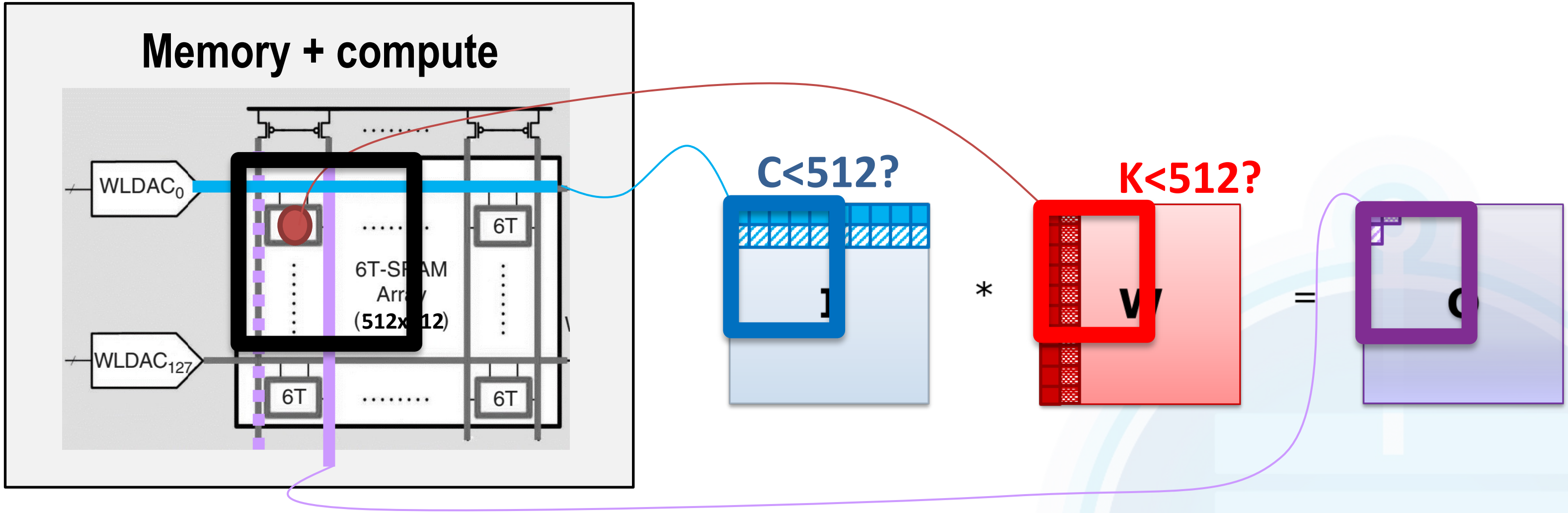


Highest peak performance? “Trick” 1 & 2!
Analog In-memory Compute: data reuse and low precision!



- Merge the memory and compute functions
  ➔ ***bring compute to the data, instead of data to the compute!***
- Energy benefits from a.) data reuse; b.) low precision analog compute
- “Analog In-memory compute” (AiMC) ➔ **<1fJ/op**
- E.g. 512x512 size array

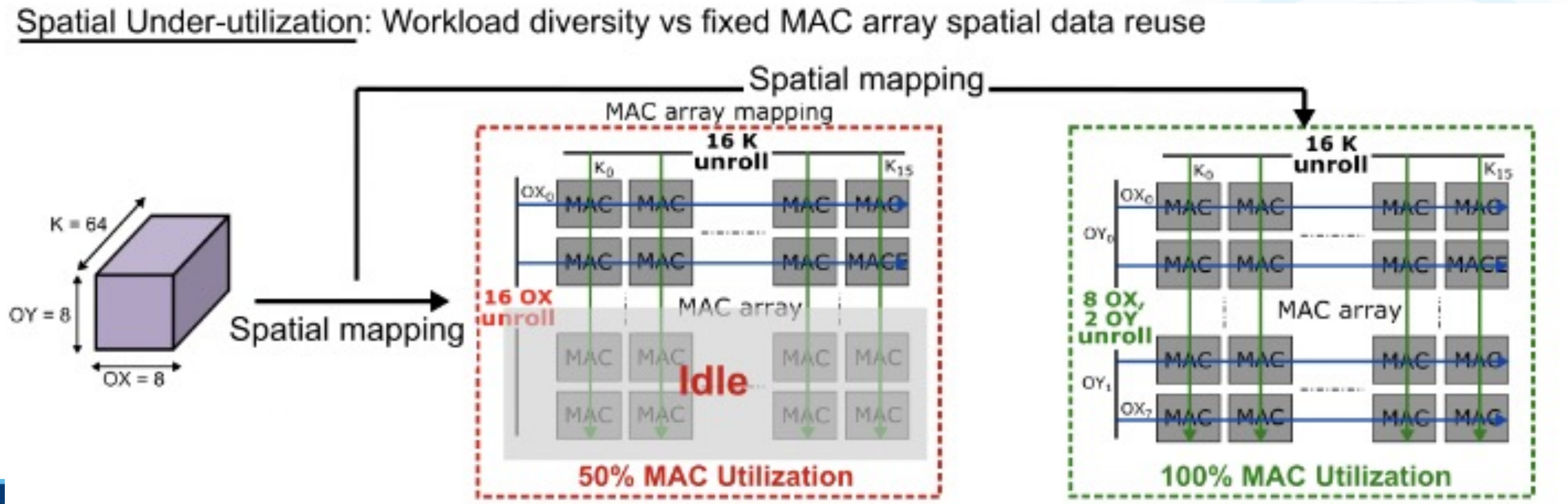
Analog In-memory Compute: Under-utilization



- But… utilization costs!
  - Low data flow flexibility
  - Only matrices with dimensions aligned with memory array efficiently used
- E.g. 512x512 size array ➔ waste of utilization / power / …

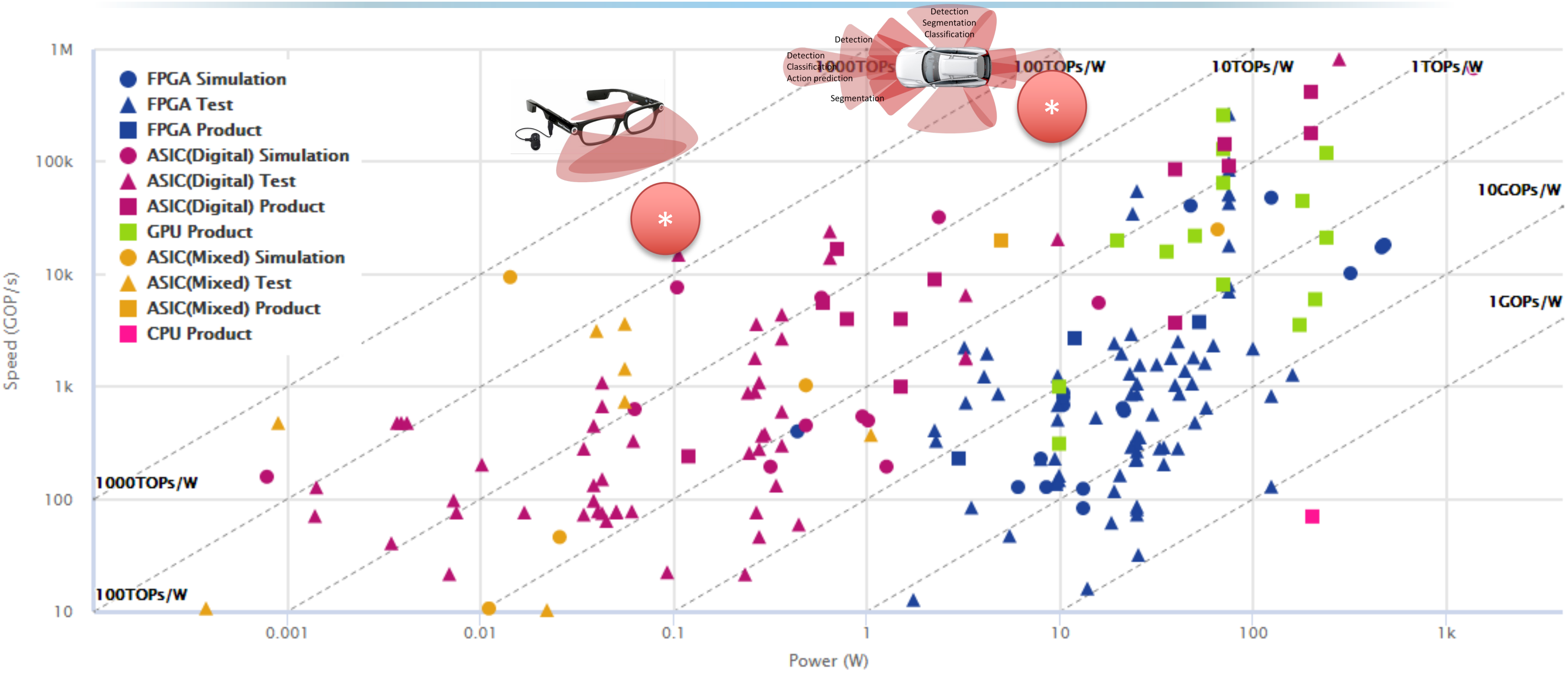
Efficiency for actual ML workloads

- In-memory compute enable to exploit low precision and massive parallelism
- But…:
  - But data reuse opportunities are layer dimension dependent
  - What about non GeMM layers? DW layers? FC layers? …?

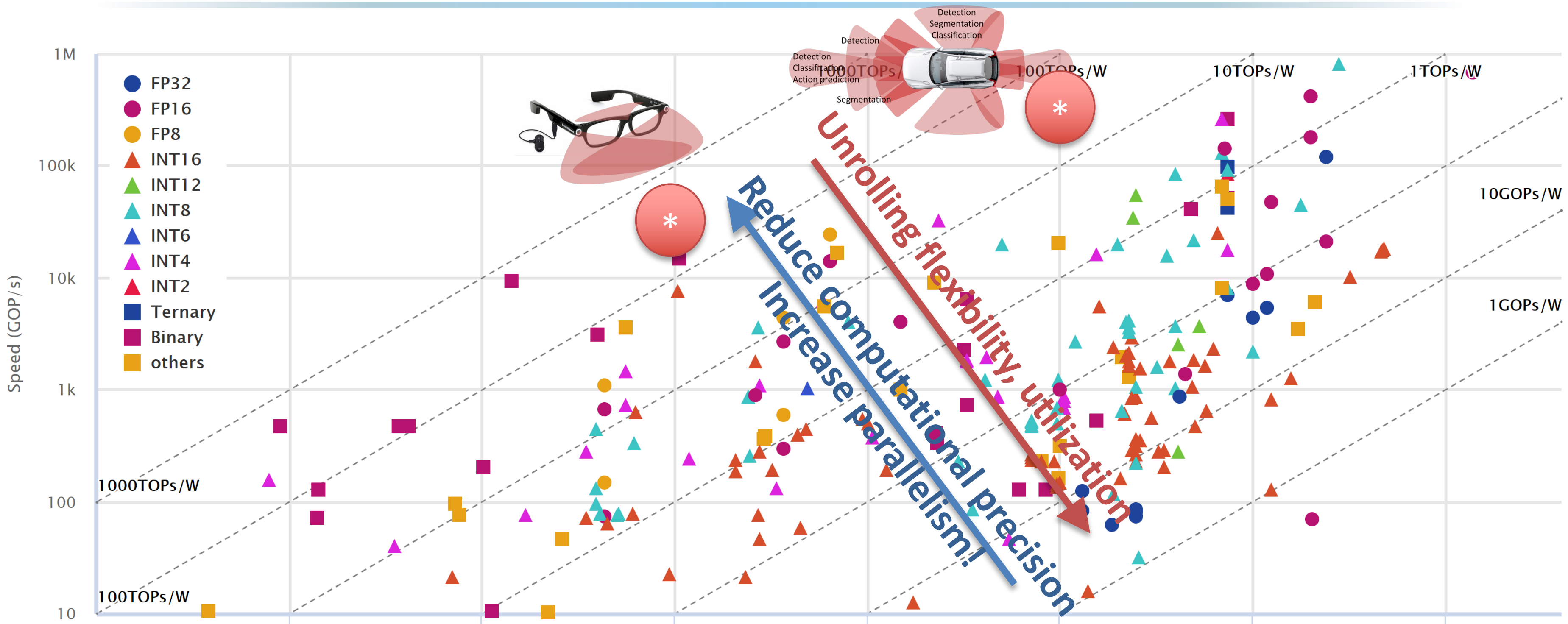


KU LEUVEN

Neural network processors: state-of-the-art



Neural network processors: state-of-the-art



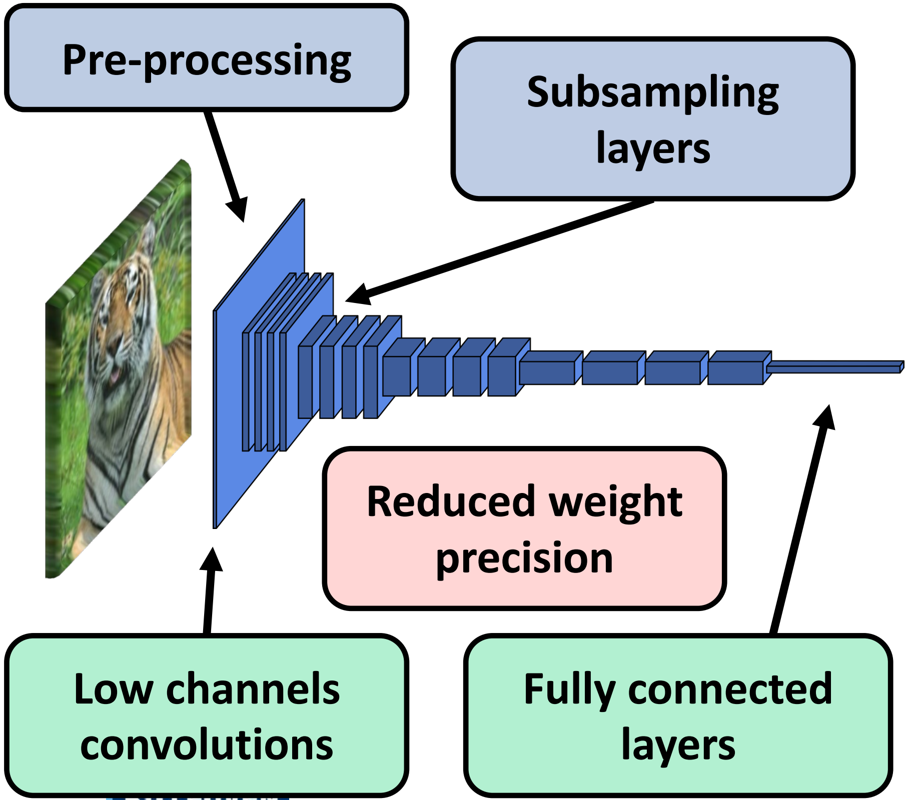
Peak performance != workload performance

KU LEUVEN micas

Overview

- Peak vs workload performance
  - Dependency on precision
  - Dependency on dataflows
- **Motivation for heterogeneous systems**
- The Diana system
- Heterogeneous scheduling with ZigZag
- The future?

Need for heterogeneity

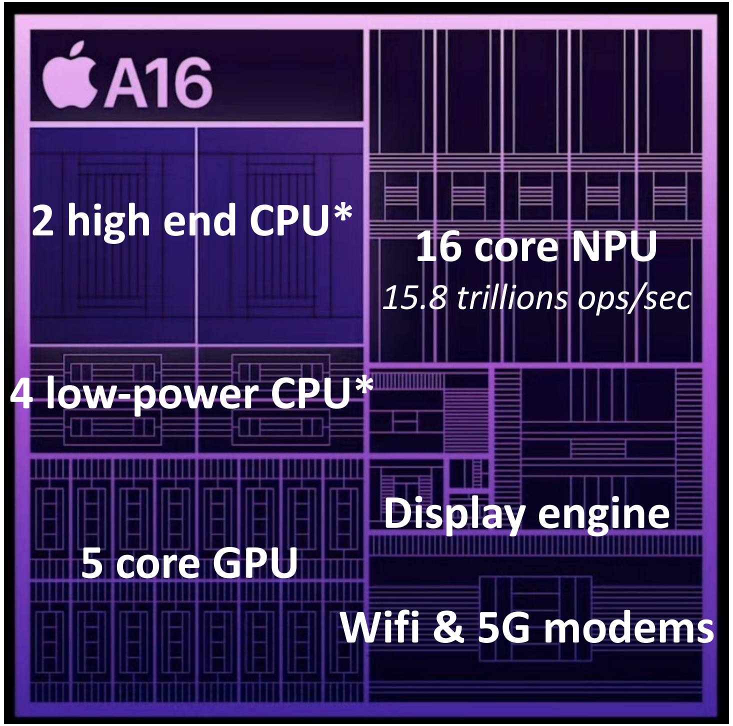


- **Exploit D/AiMC high energy efficiency**
  - For layers ok with **lower precision**
  - For layers with good parallelism for **high utilization**
- **BUT have alternative accelerator(s) for other layers!**

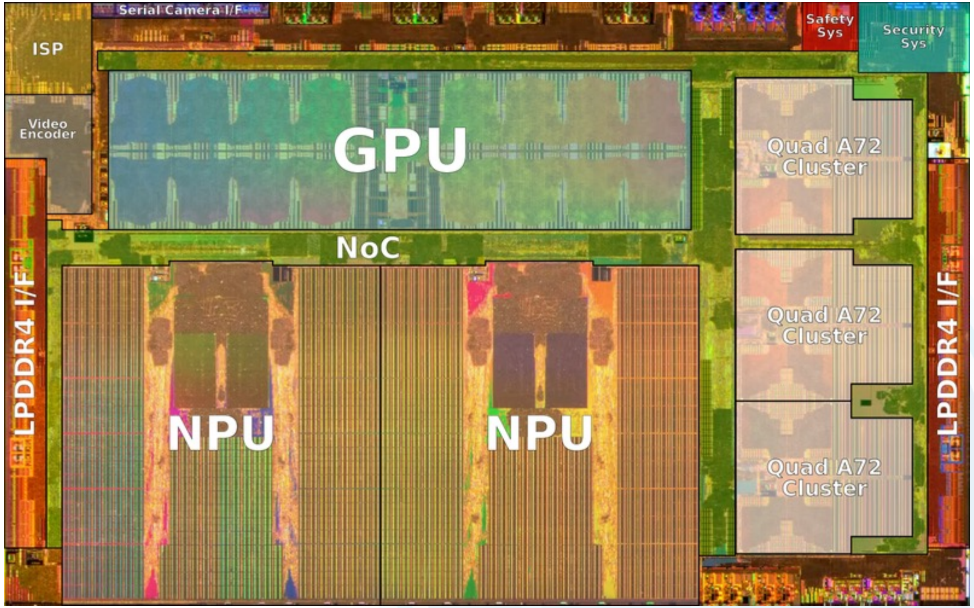
➔ **Heterogeneous systems**

Flexibility AND efficiency? → Heterogeneous systems!

Examples in the "edge"



Apple A16

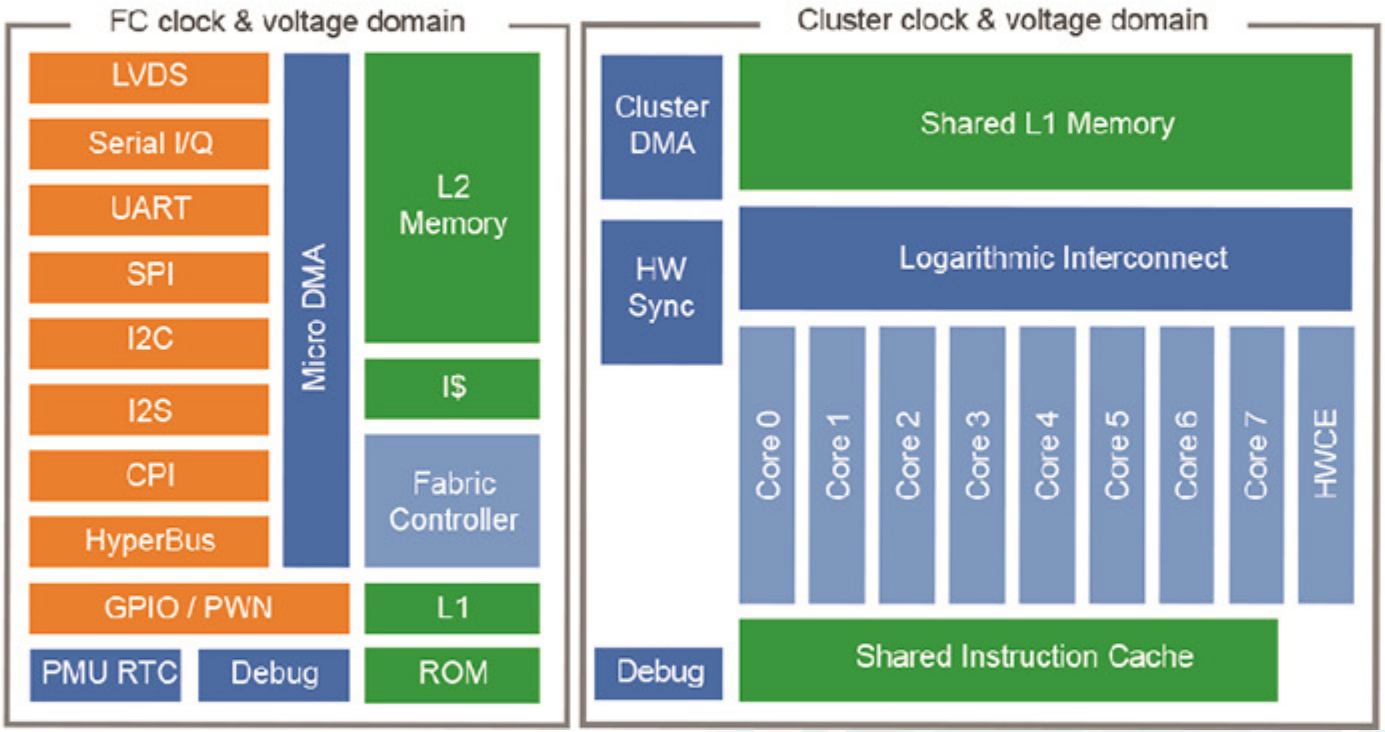


Tesla FSD

But not low power…

Low power for the extreme edge (tinyML)?

GREENWAVES TECHNOLOGIES

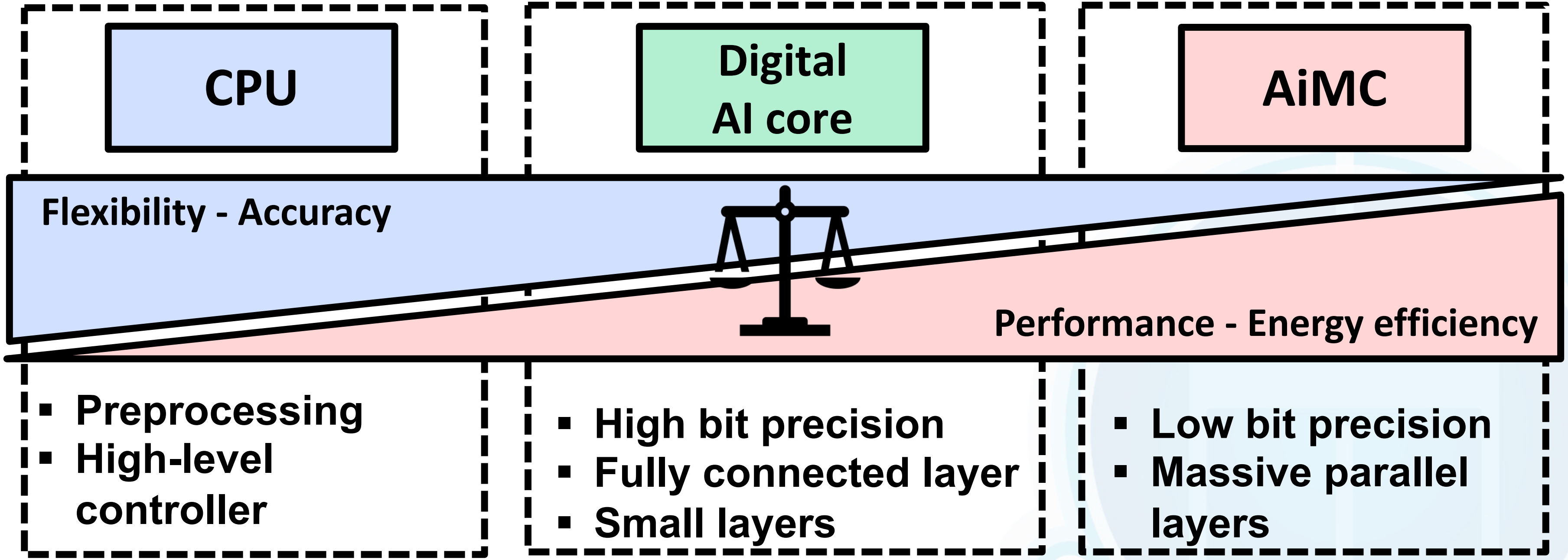


- But lack of heterogeneity… (CPU heavy)

Overview

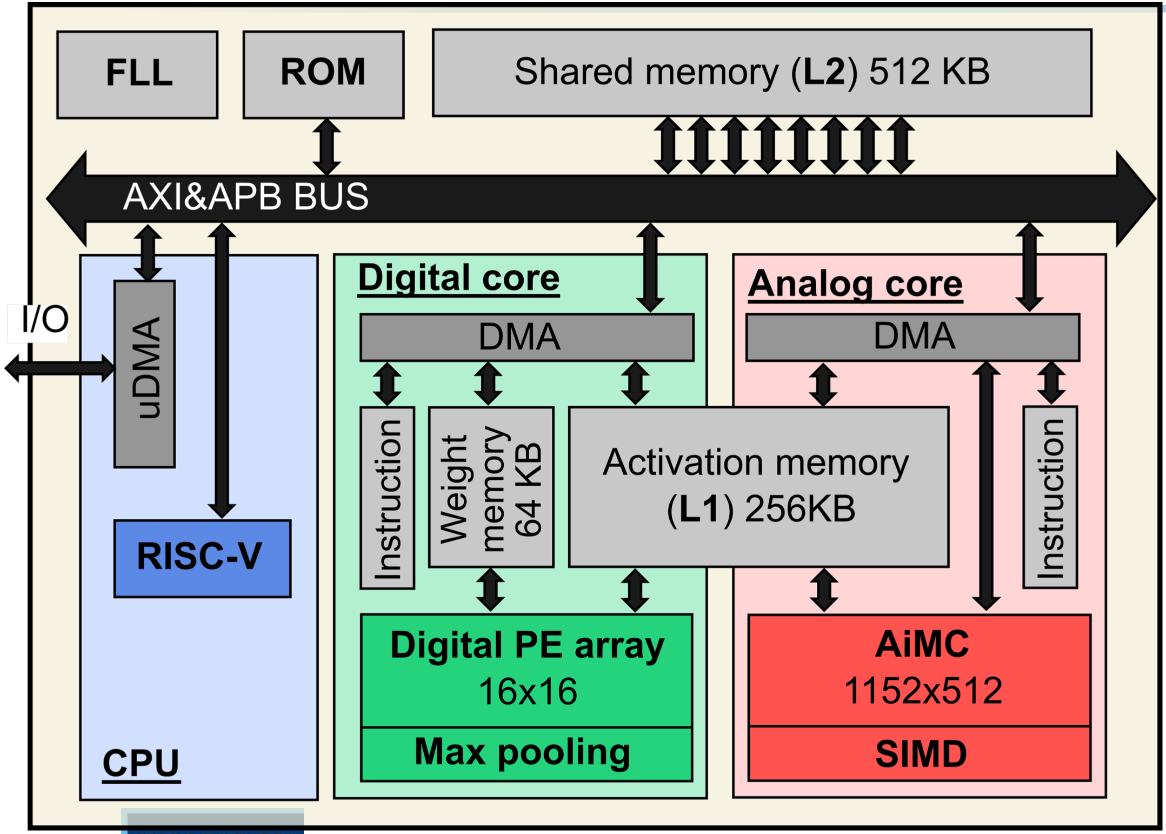
- Peak vs workload performance
  - Dependency on precision
  - Dependency on dataflows
- Motivation for heterogeneous systems
- **The Diana system**
- Heterogeneous scheduling with ZigZag
- The future?

Digital-analog accelerator co-design



DIANA SoC – High Level View

[Ueyoshi, Kodai, et al., "DIANA: An End-to-End Energy-Efficient DIgital and ANAlog Hybrid Neural Network SoC", ISSCC2022]



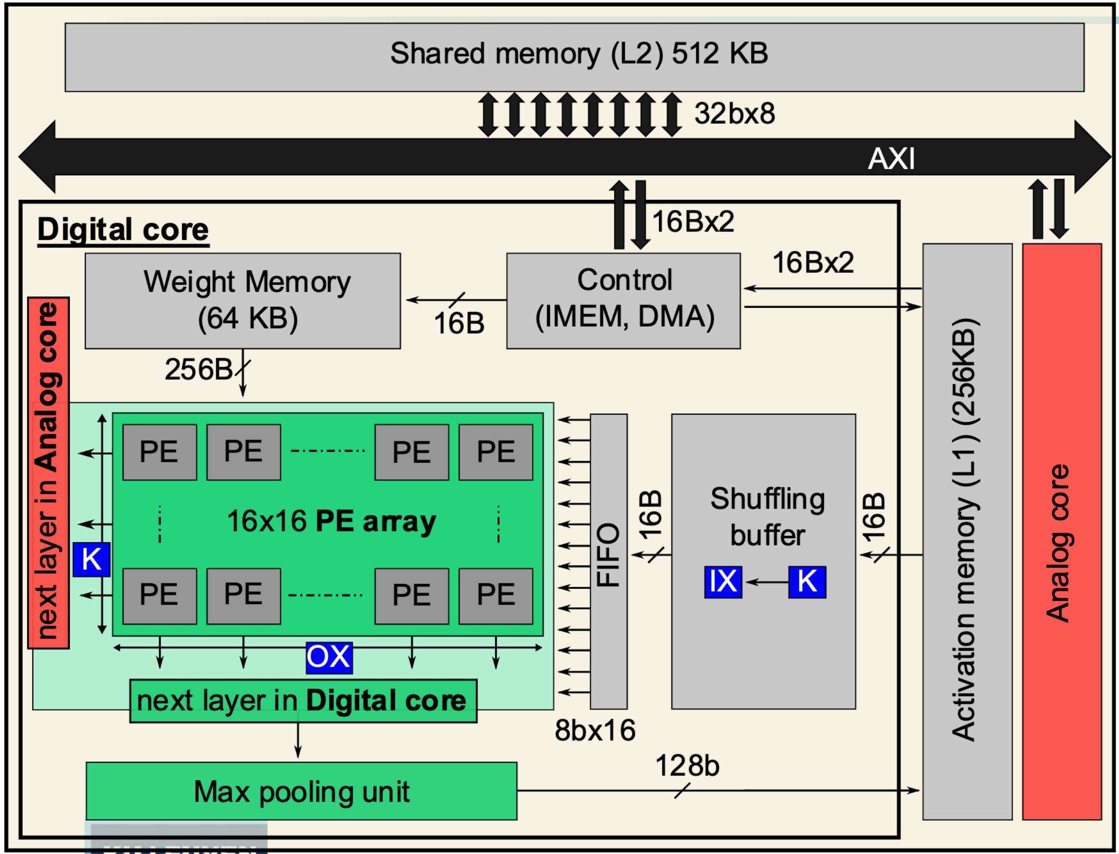
DIANA chip:

- RISC-V CPU*
  - High-level control
  - External I/O
- Digital AI core
  - 16x16 PEs
- Analog AI core
  - AiMC for MVMs
  - SIMD for post process
- Distributed memory hierarchy

*RISC-V CPU and periphery based on PULPissimo platform, ETH

Digital core – Extended flexibility

[Ueyoshi, Kodai, et al., "DIANA: An End-to-End Energy-Efficient DIgital and ANAlog Hybrid Neural Network SoC", ISSCC2022]



3 different levels of flexibility

Computation flexibility

- 16x16 **PE array**
  - 2, 4 and 8-bit precision

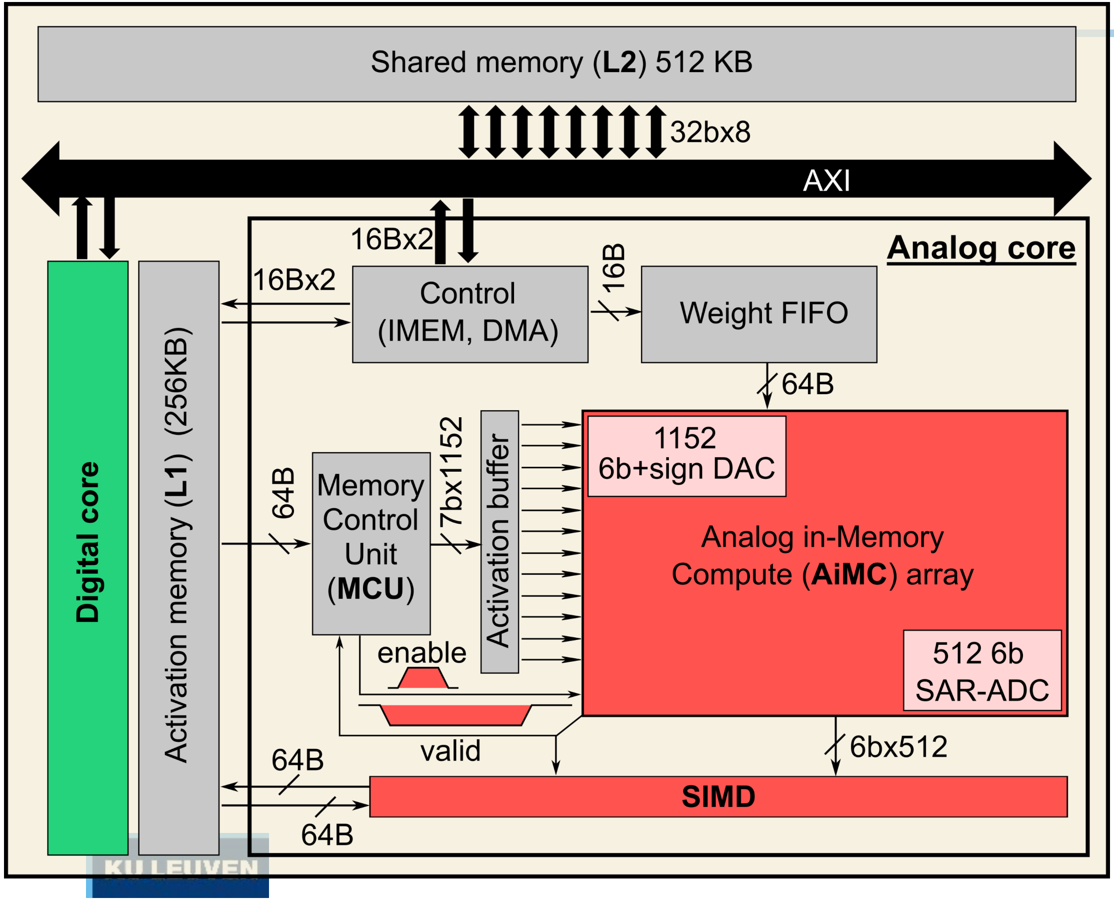
Operation flexibility

- Convolutional layers
- Fully connected layers
- Element-wise operations
- Max pooling

Dataflow flexibility

Analog core – Computing Units

[Ueyoshi, Kodai, et al., “DIANA: An End-to-End Energy-Efficient DIgital and ANAlog Hybrid Neural Network SoC”, ISSCC2022]

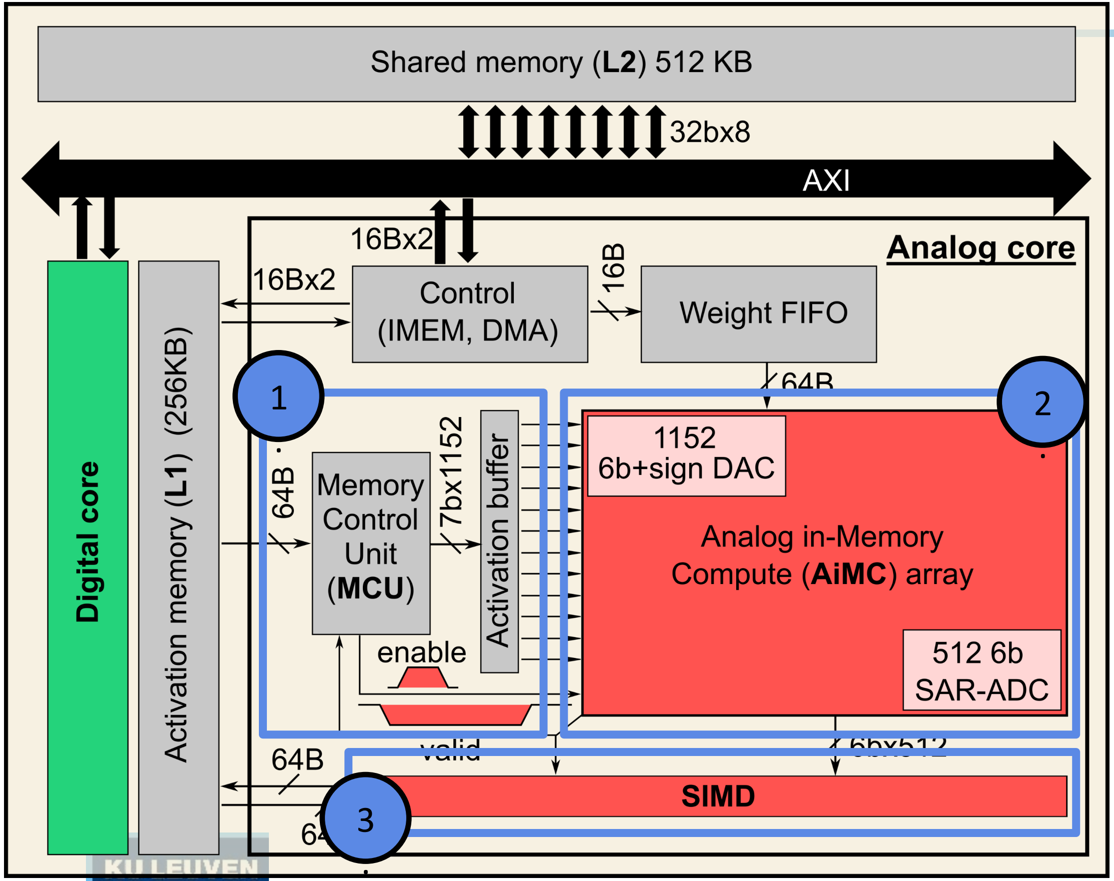


Computation units:

- **AiMC array** SRAM-based [6] for Matrix-Vector-Multiplications
  - 1152 7-bit input DACs
  - 512 6-bit output ADCs
  - 590k compute cells (ternary weights)

    Half a million MAC/cc!
- **SIMD** for post-processing
  - 64 parallel computing units
  - 6 stages

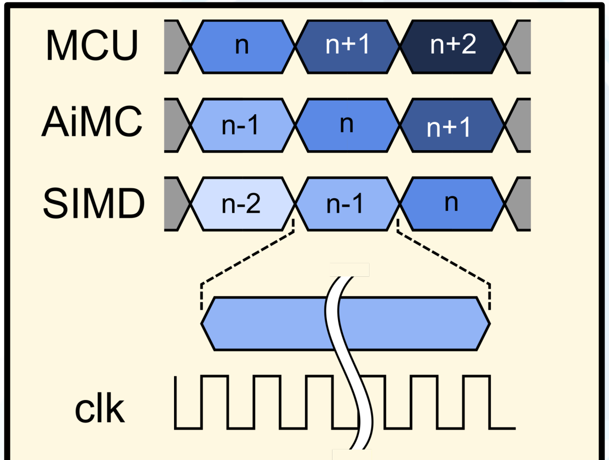
Analog core – Pipeline



Three processing stages:

1. **MCU** → Input fetch stage
2. **AiMC** → Compute stage
3. **SIMD** → Post-processing stage

AiMC macro always in use



macro pipeline, several clock cycle

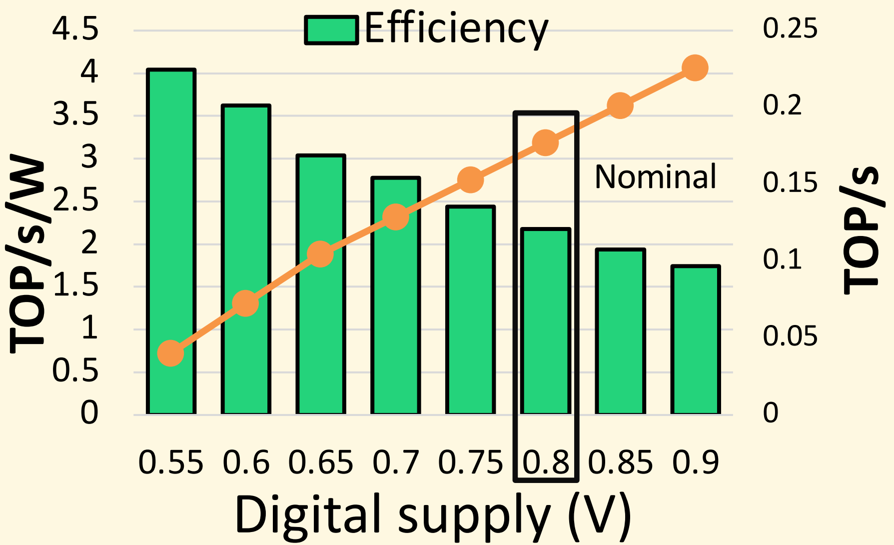
Measured results – Peak numbers

*Analog supply @ 0.8V (nominal)

DIANA SoC - Digital core working

8b weight - actv., 32b accumulation

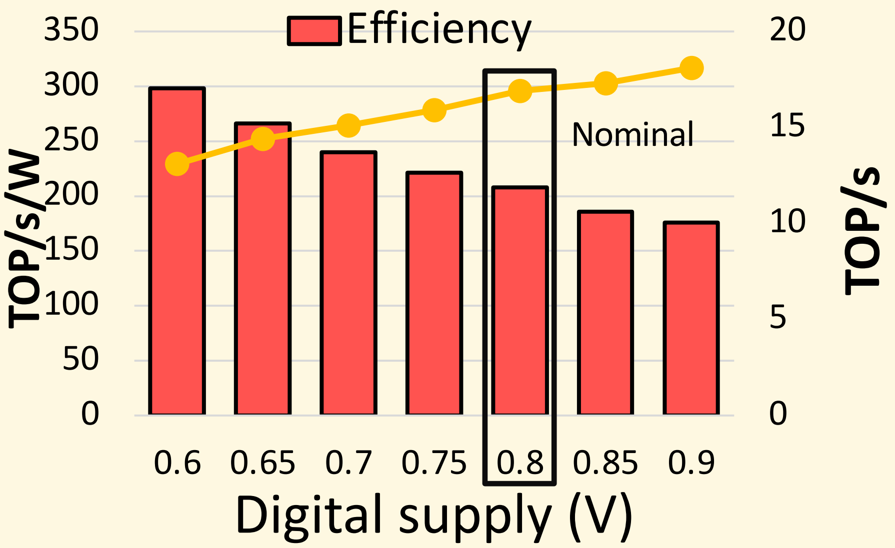
Peak efficiency/performance



DIANA SoC - Analog core* working

7b DAC, 6b ADC, {-1,0,1} weights

Peak efficiency/performance



Measured results – Peak numbers

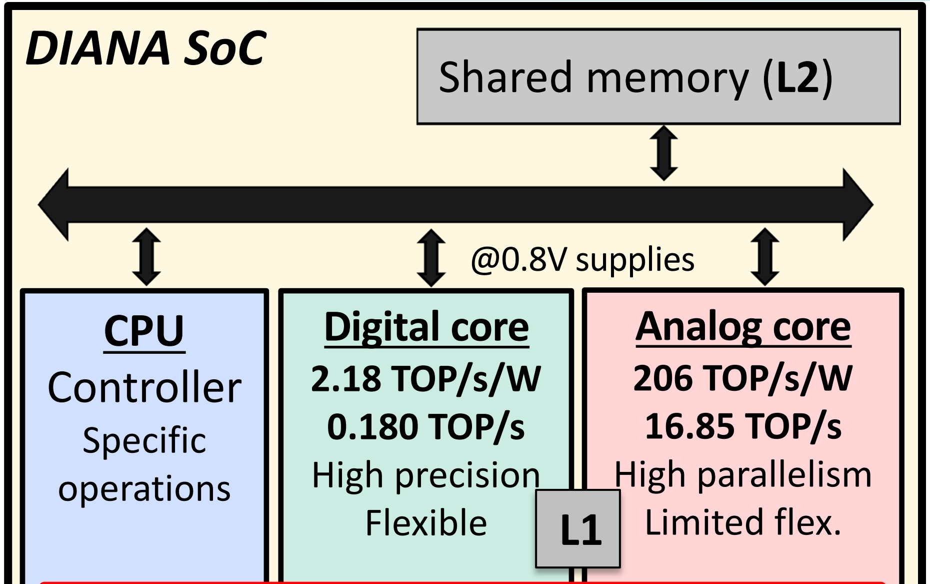
2 orders of magnitude difference between cores
(8b digital – 2b analog weights)

Core **@ 0.8V**	**Efficiency (TOP/s/W)**	**Performance (TOP/s)**
Digital	2.18	0.177
Analog	208	16.9

Overview

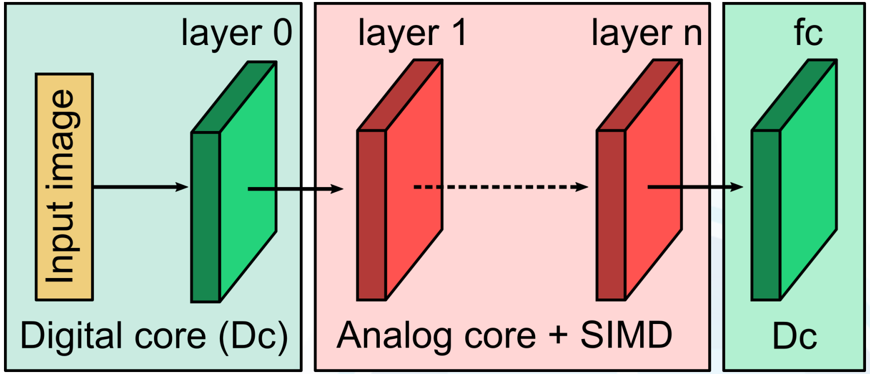
- Peak vs workload performance
  - Dependency on precision
  - Dependency on dataflows
- Motivation for heterogeneous systems
- The Diana system
- **Heterogeneous scheduling with ZigZag&Stream**
- The future?

Hybrid execution – Pipelining



Different accelerators optimized for **different workloads** to boost **system level performance**

End-to-end mapping



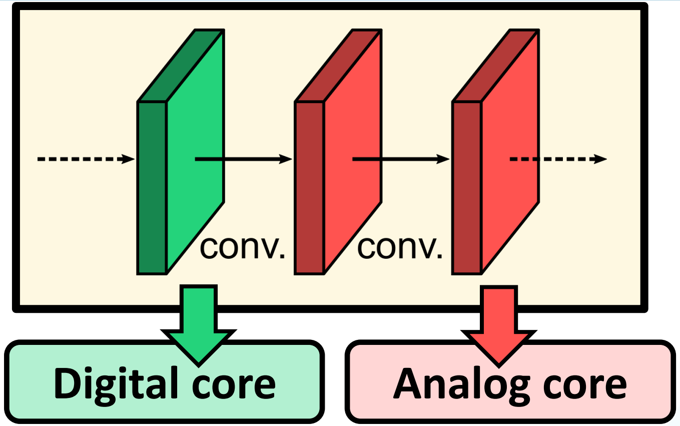
1.) Streaming operation

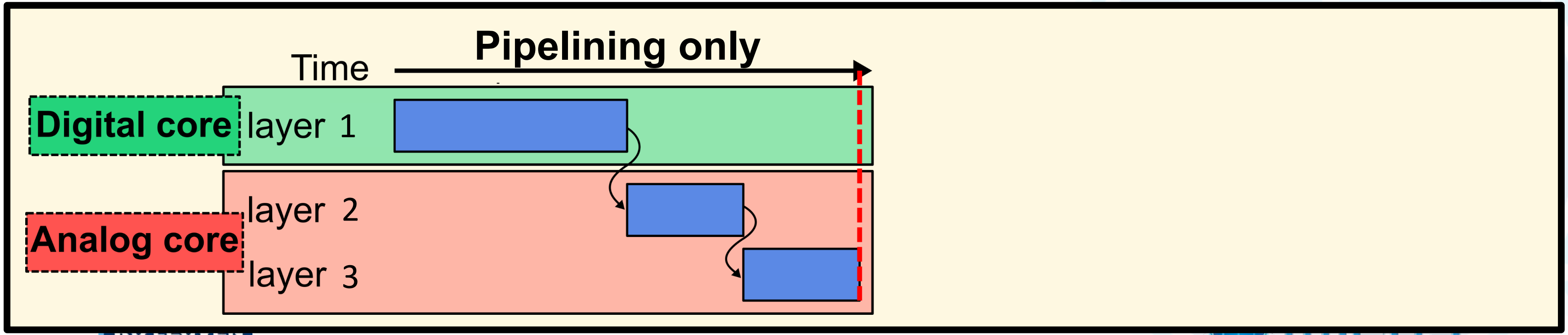
➔ Efficient data sharing

2.) Scheduling of which layer on which core?

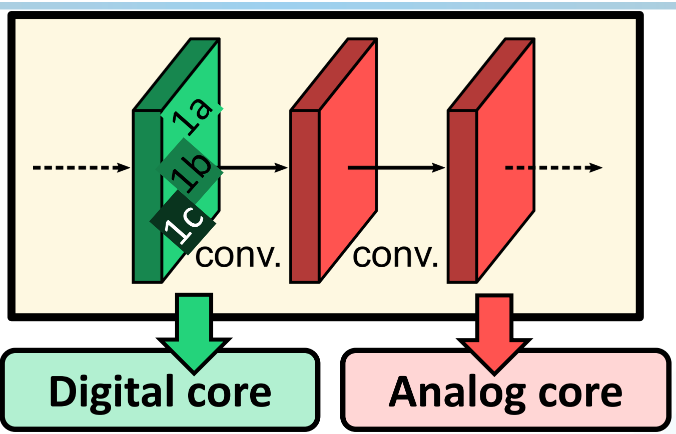
➔ ZigZag!

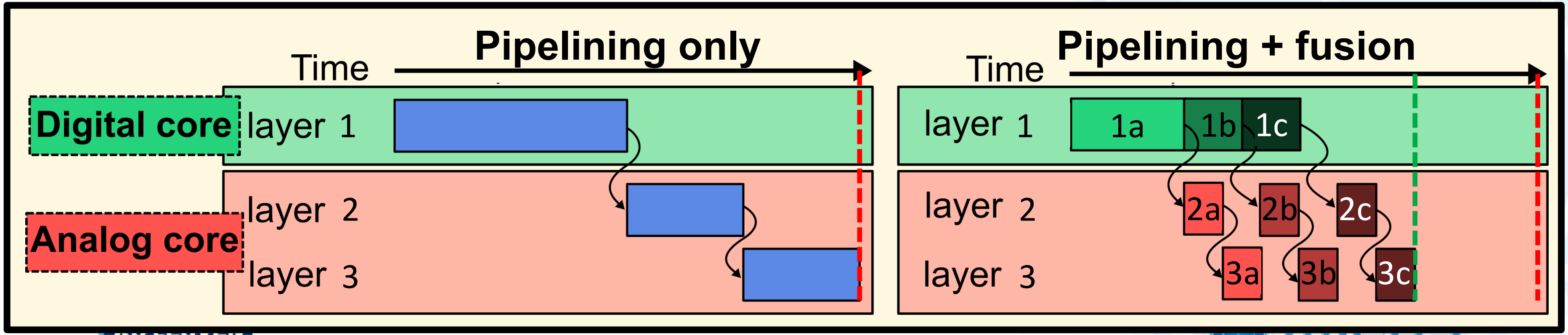
Hybrid execution – Layer fusion





Hybrid execution – Layer fusion

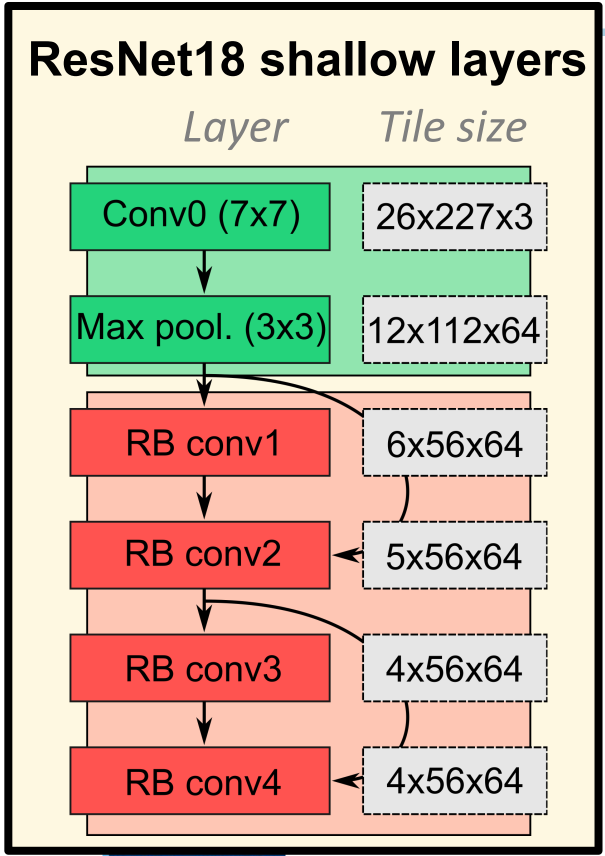


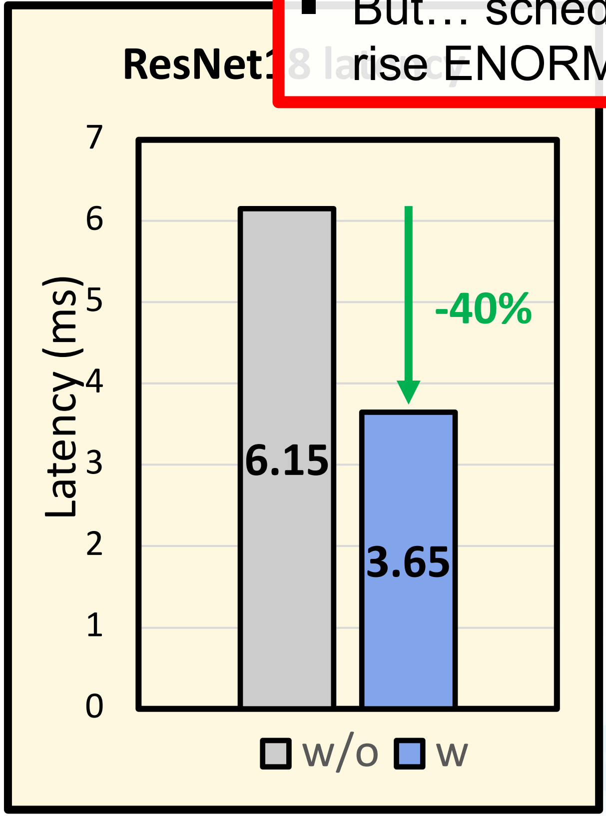


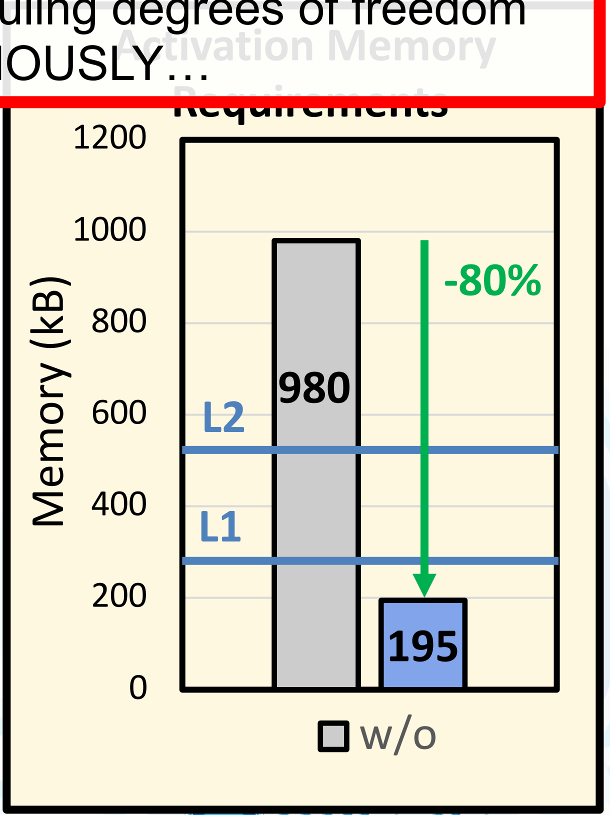
Layer fusion benefit on ResNet18

- Speeds up end-to-end execution
- Reduces memory requirements

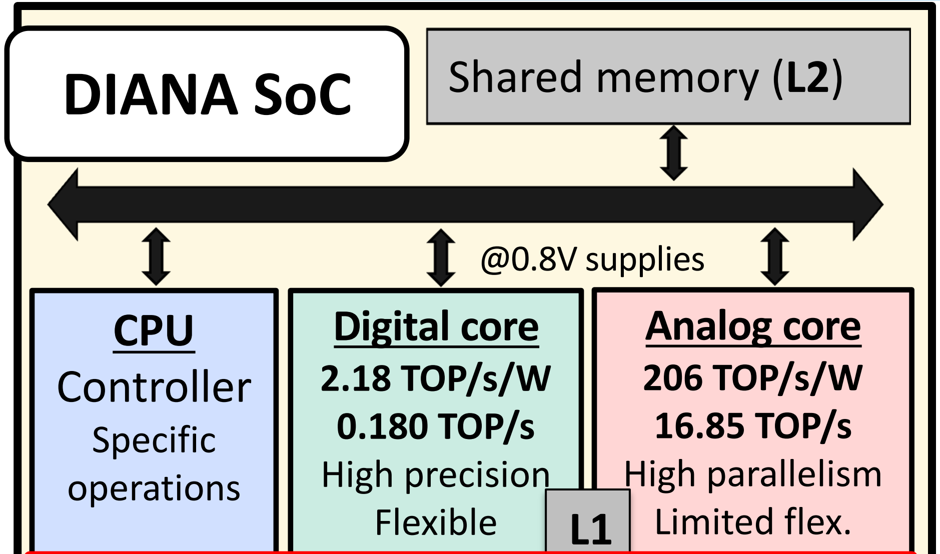
- But… scheduling degrees of freedom rise ENORMOUSLY…





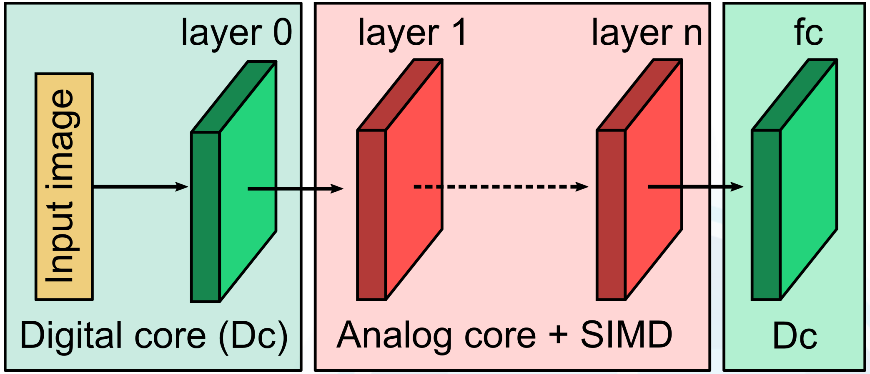


Hybrid execution – Pipelining



Different accelerators optimized for **different workloads** to boost **system level performance**

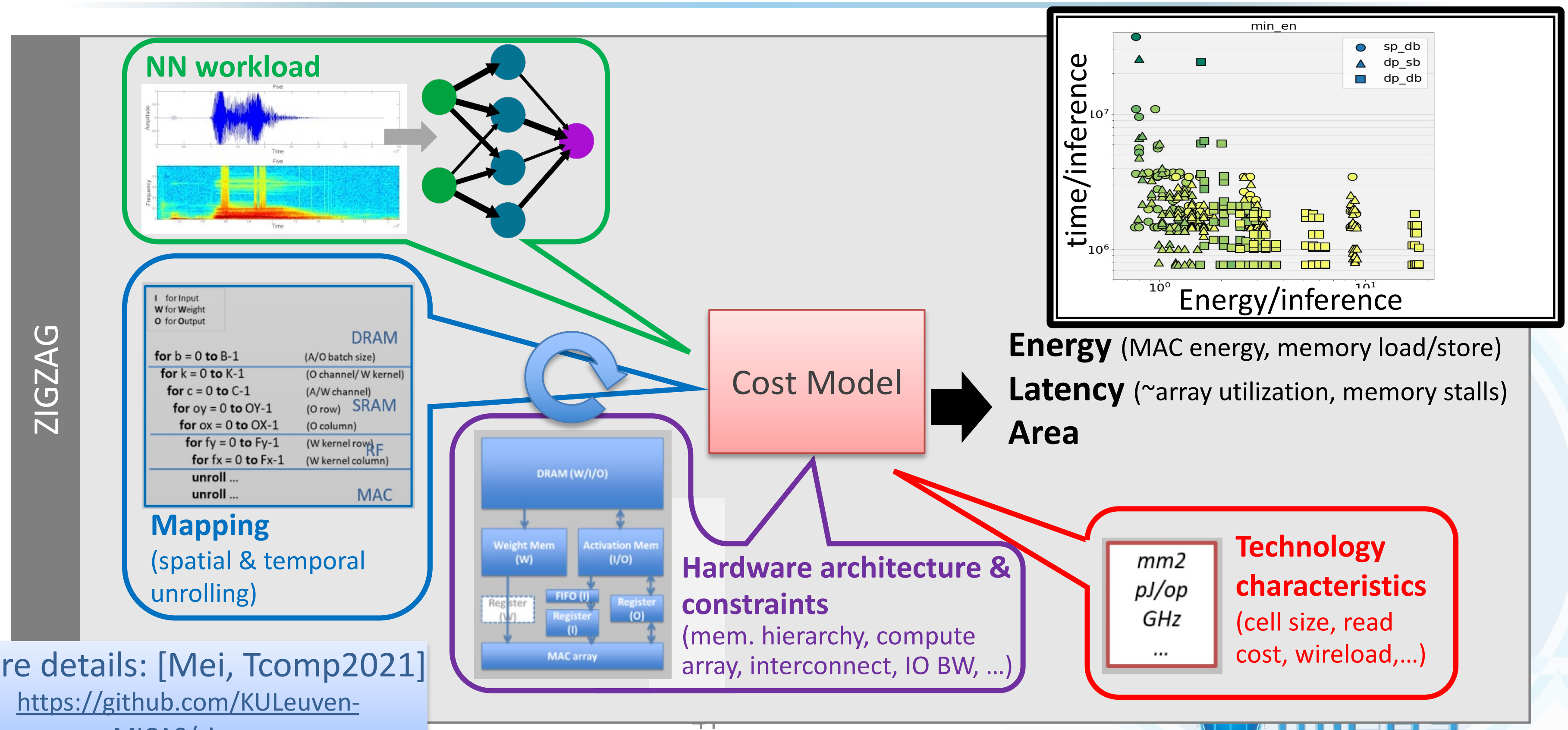
End-to-end mapping



1.) Streaming operation
➔ Efficient data sharing

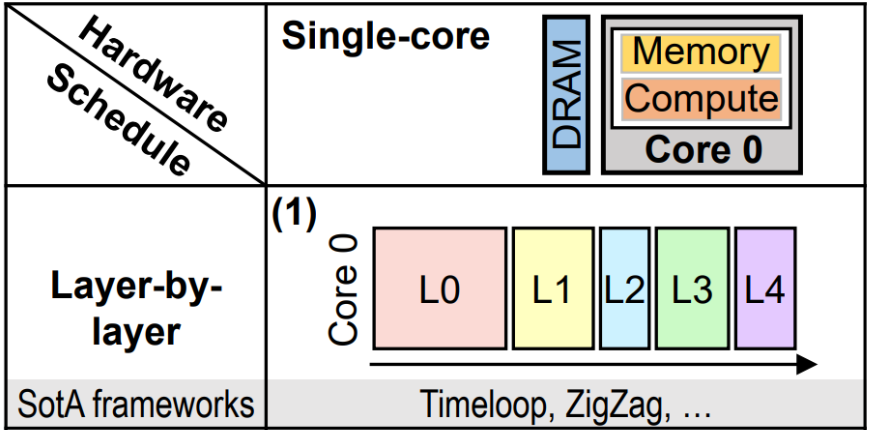
2.) Scheduling of which layer tile on which core at what moment?
➔ ZigZag!

Optimizing DNN embedded processing stack with ZigZag



More details: [Mei, Tcomp2021]
https://github.com/KULeuven-MICAS/zigzag

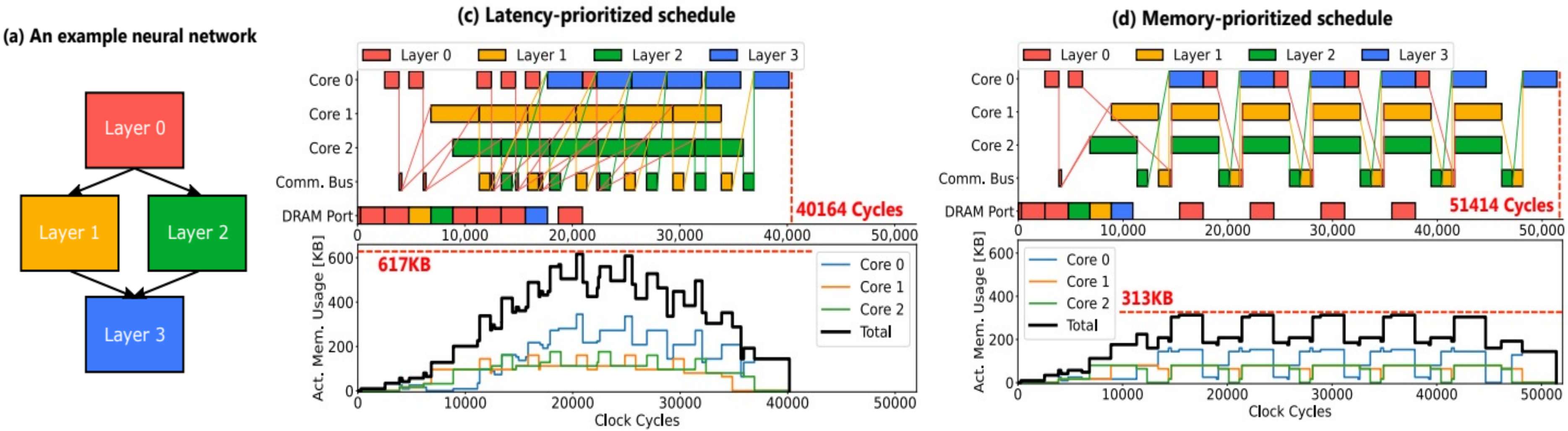
Stream: ZigZag extension to layer fusion and multi-core



[arXiv:2212.10612]

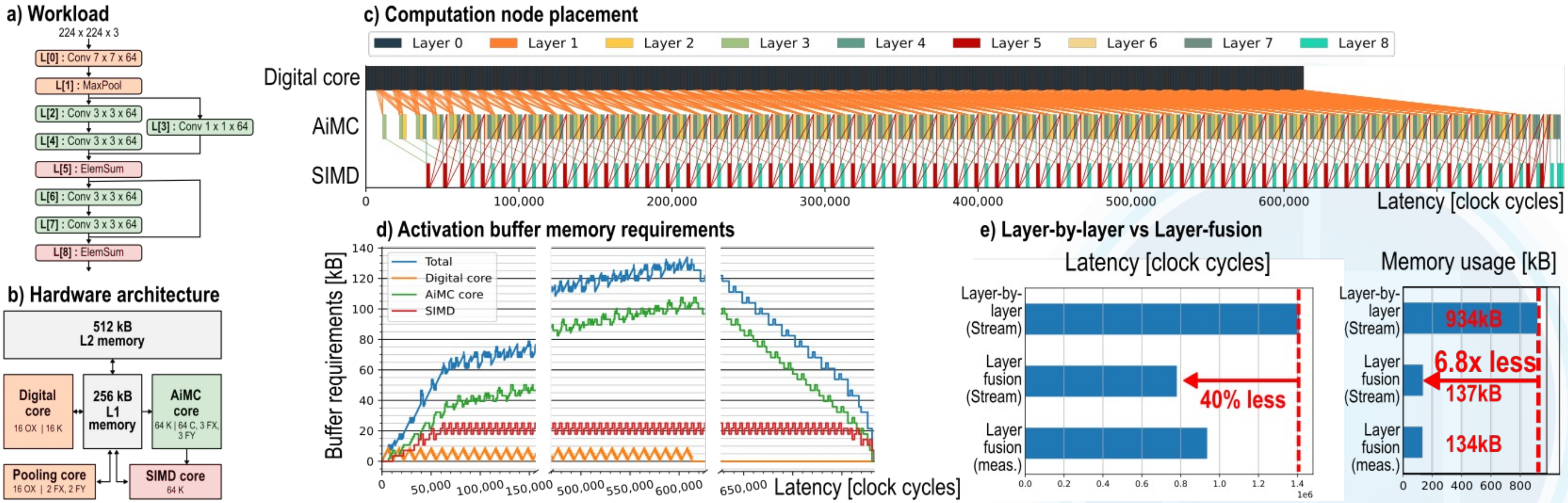
Power of ZigZag exploration: Scheduling optimization: latency or memory

- Optimize unrolling, temporal schedule, tile size, (core allocation),…



[arXiv:2212.10612]

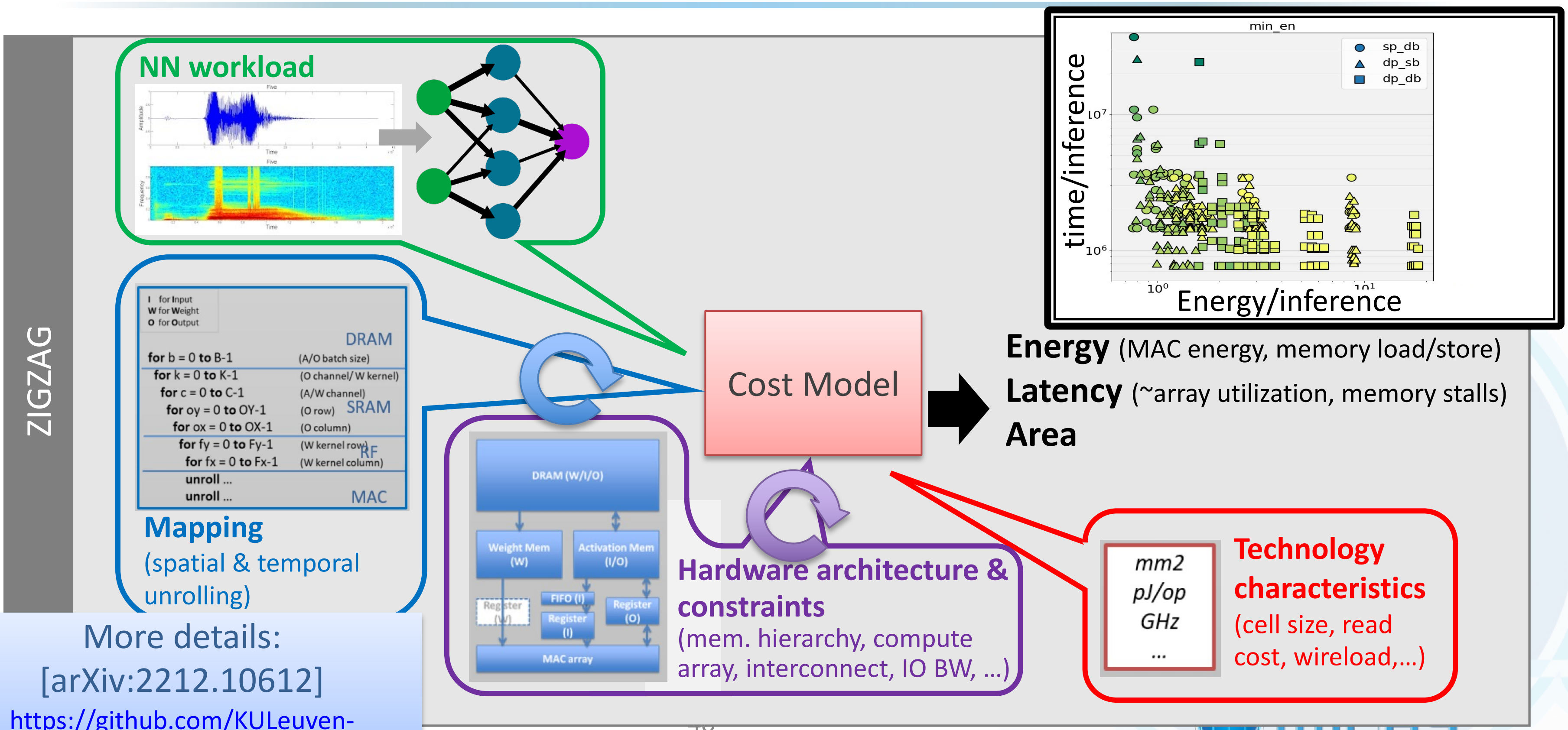
Scheduling optimization for Diana with Stream



KU LEUVEN

micas

Optimizing DNN embedded processing stack with Stream



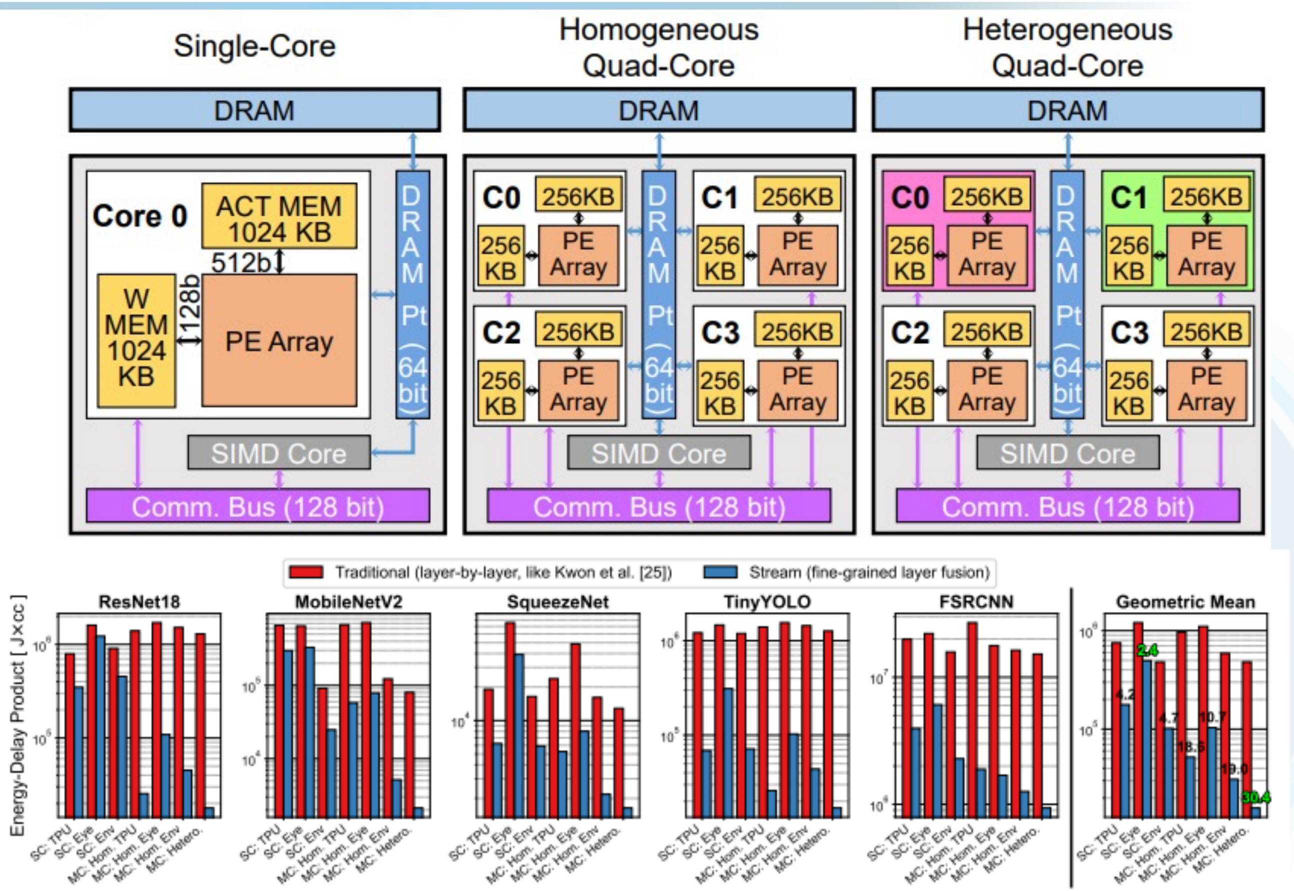
More details:
[arXiv:2212.10612]
https://github.com/KULeuven-MICAS/stream

[arXiv:2212.10612]

Design space exploration

Huge design space!
- AI core sizes
- Memory volume
- Interconnect scheme
- …

No optimal design for all networks ➔ heterogeneity helps!



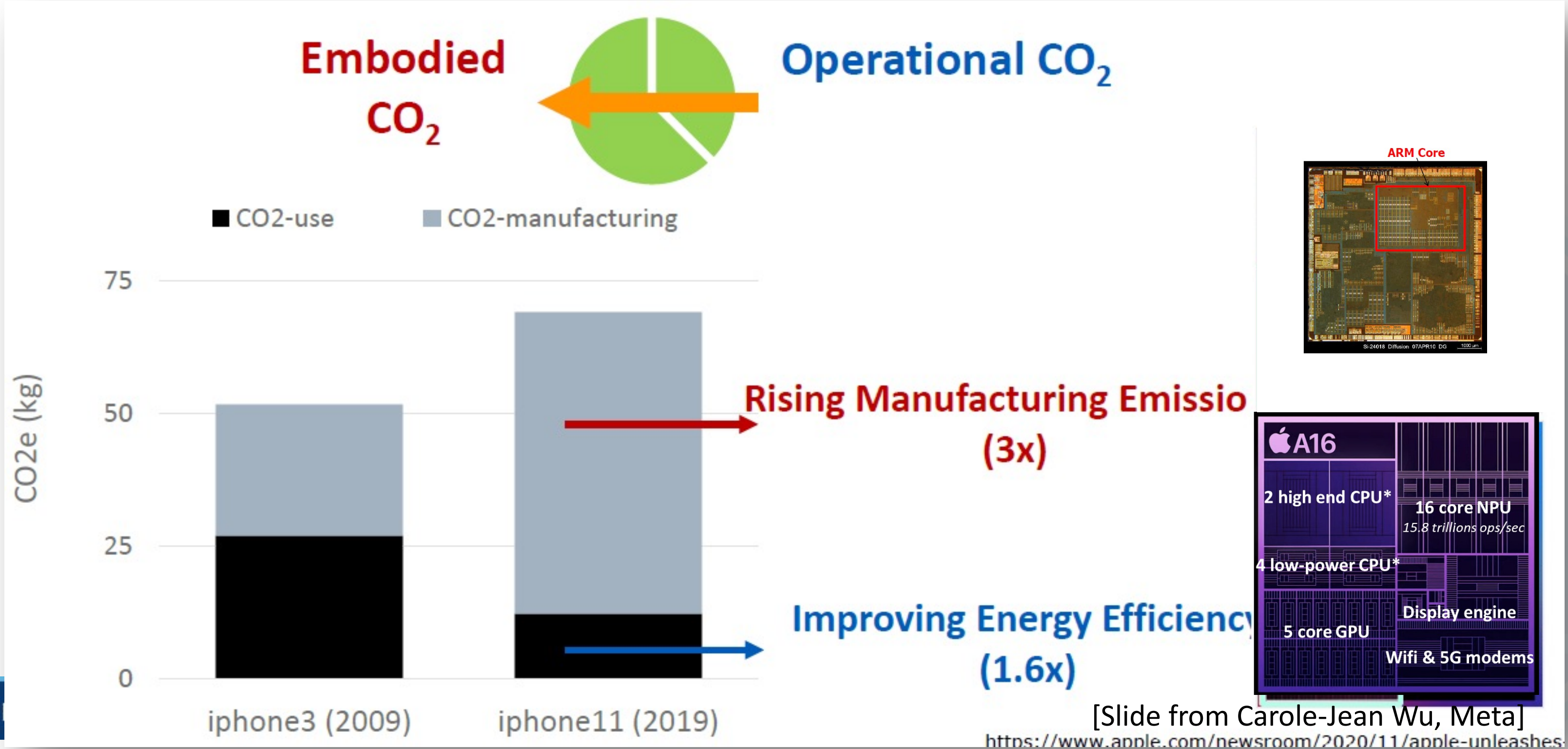
KU LEUVEN

Overview

- Peak vs workload performance
  - Dependency on precision
  - Dependency on dataflows
- Motivation for heterogeneous systems
- The Diana system
- Heterogeneous scheduling with ZigZag&Stream
- **The future?**

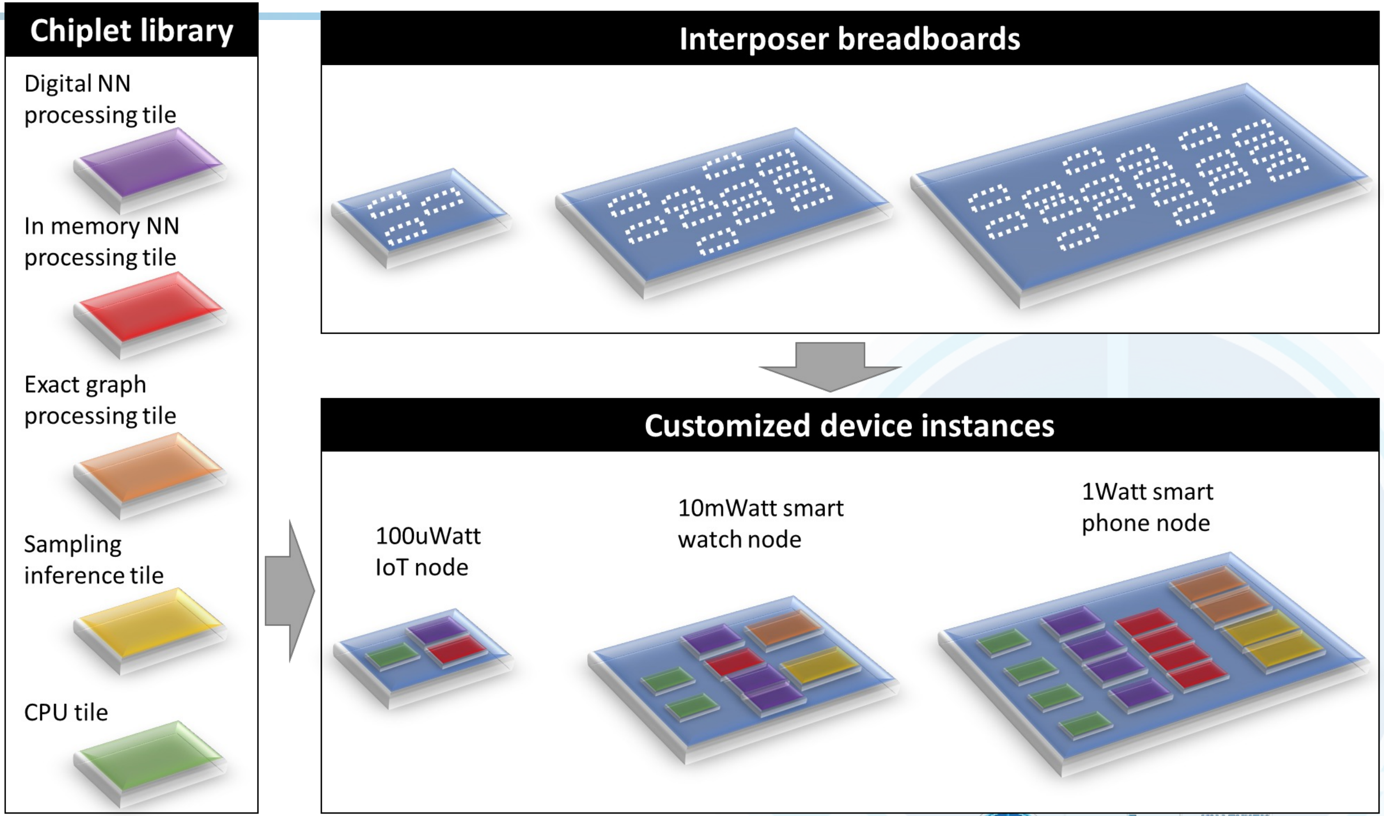
Carbon footprint

- More and more accelerators on chip? No!



Future?: More modular max-and-match with chiplets

- Supporting more workloads
- Rapid prototyping
- Customization
- Lower carbon footprint



Conclusion

- Specialization can bring significant efficiency gains
- **<u>Yet</u>** loss of flexibility while thriving in terms of peak performance

➔ Heterogeneity to have efficiency/customizability across workloads

- Towards heterogeneous, multi-core AI processing platforms
  – Rapidly customizable to algorithmic workloads
  – Supported by customizable multi-accelerator compilers
  – Large challenges ahead!

- References:

  – Ueyoshi, Kodai, et al., "DIANA: An End-to-End Energy-Efficient DIgital and ANAlog Hybrid Neural Network SoC", In 2022 IEEE International Solid-State Circuits Conference (ISSCC), IEEE, 2022.
  – Mei, Linyan, et al. "ZigZag: Enlarging joint architecture-mapping design space exploration for DNN accelerators." IEEE Transactions on Computers 70.8 (2021): 1160-1174.
  – Symons, Arneet al., "Towards Heterogeneous Multi-core Accelerators Exploiting Fine-grained Scheduling of Layer-Fused Deep Neural Networks", arXiv:2212.10612.

KU LEUVEN

imec

KU LEUVEN micas

Copyright Notice

TINY ML EMEA
AMSTERDAM, NL
June 26-28,2023

www.tinyml.org