From Eyewear to Silicon: RISC-V for Low-Power AI in Next-Generation Smart Glasses

Marco Fariselli - Sr. SW Engineer at EssilorLuxottica
May 2026

The «Holy Grail» of Wearables

the best place to be, closer to all body sense

Smart Eyewear Lab

19 July 2022

Official Announcement of the Partnership with Politecnico di Milano

October 2023

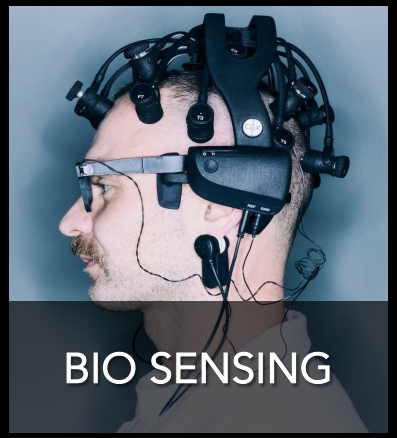
casted the first stone

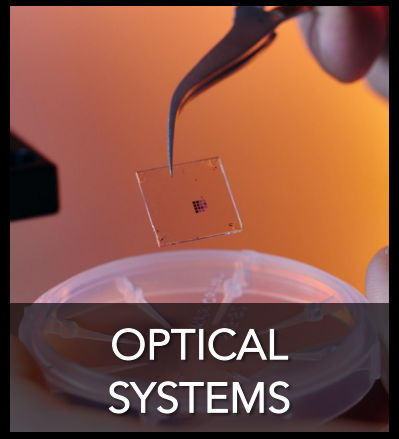
May 2026

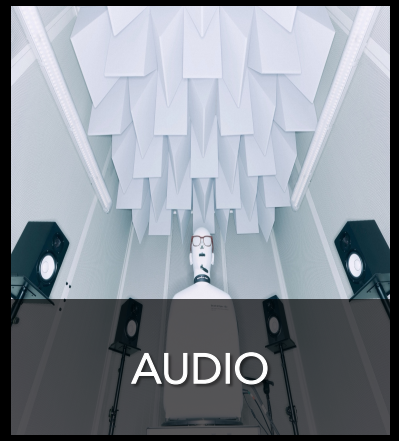
~100 researchers

Hardware Components & System Integration Technology

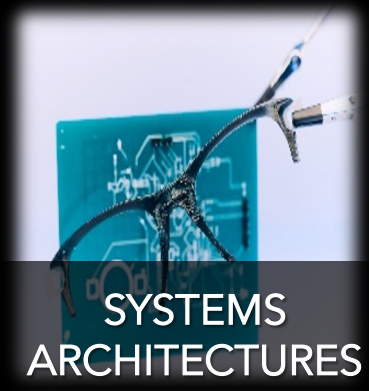
RESEARCH PLATFORMS







CROSS SERVICE PLATFORMS

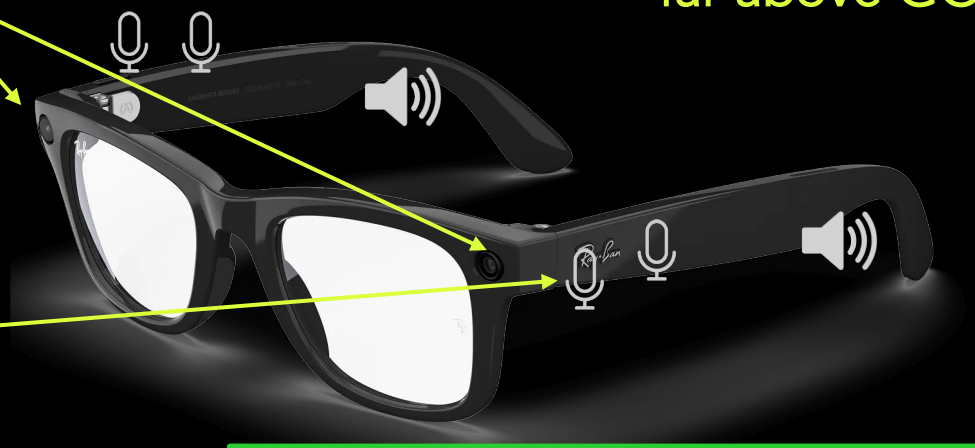


A lot of things... in one device

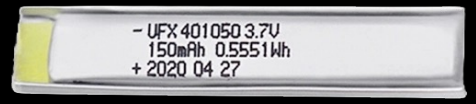
Video:
- Context understanding
- Eye-tracking
- ...

Audio:
- Multi-microphones
- Hearing aid
- ...

AI:
- Assistant
- Online translation
- ...



Computation requirements are **far above GOP/s**



With 150mAh battery operating 16 hours means a system consuming **35mW on average**

Today's MCUs are not enough...
NEED TO BOOST ENERGY EFFICIENCY

A Continuously Evolving Compute Challenge

- Smart glasses workloads evolve faster than silicon lifecycle
- We need both flexibility and energy efficiency

DSP

CNN/RNN

DSP

CNN/RNN

DSP

LLMs

CNN/RNN

DSP

?

LLMs

CNN/RNN

DSP

2025

2027

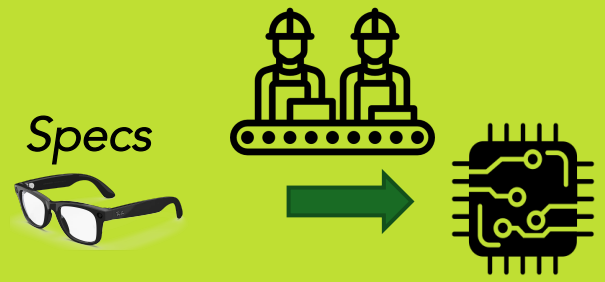
2029

2031

2+1 Silicon design paths...

Full custom:
- Ultra-specialized system design
- Build ecosystem from scratch

Specs

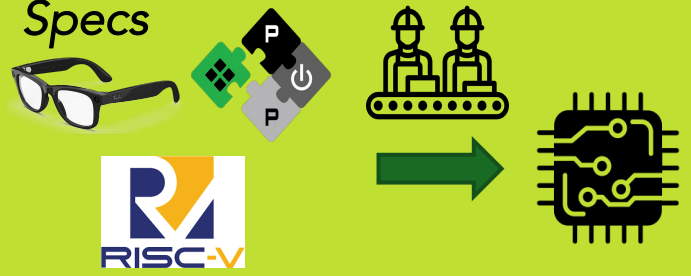


En. Efficiency

Time-to-market

Open source specialization:
- Adapt to specific needs
- Optimize the sinergy
- Reuse ecosystem

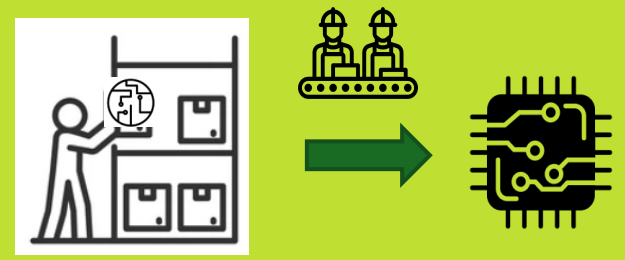
Specs



En. Efficiency

Time-to-market

Off-the-shelf:
- Integration of pre-designed components
- Low interoperability between the actors

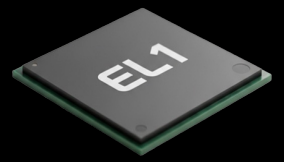


En. Efficiency

Time-to-market

How did we start?

100+ Researchers
10+ years of development
20+ silicon prototypes



RISC-V

- **Extendible**
- **Modular**

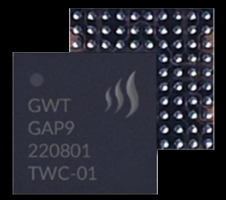
PULP Platform

Chips-IT

OPEN SOURCE APPROACH to keep the community active and EL1 at the state of the art

EssilorLuxottica

Adaptation in the context of smart low-power systems



Industrialized in 2019

ISA Extension example: RV32IMCXpulp

```
char a[], b[], d[];
for (i = 0; i < 100; i++)
    d[i] = a[i] + b[i];
```

Baseline

```
mv    x5, 0
mv    x4, 100
Lstart:
  lb    x2, 0(x10)
  lb    x3, 0(x11)
  addi  x10,x10, 1
  addi  x11,x11, 1
  add   x2, x3, x2
  sb    x2, 0(x12)
  addi  x12,x12, 1
  addi  x4, x4, -1
bne     x4, x5, Lstart
```

11 cycles/output

Auto-incr load/store

```
mv    x5, 0
mv    x4, 100
Lstart:
  lb    x2, 0(x10!)
  lb    x3, 0(x11!)
  addi  x4, x4, -1
  add   x2, x3, x2
  sb    x2, 0(x12!)
bne     x4, x5, Lstart
```

8 cycles/output

HW Loop

```
lp.setupi 100, Lend
  lb    x2, 0(x10!)
  lb    x3, 0(x11!)
  add   x2, x3, x2
Lend:   sb x2, 0(x12!)
```

5 cycles/output

Packed-SIMD

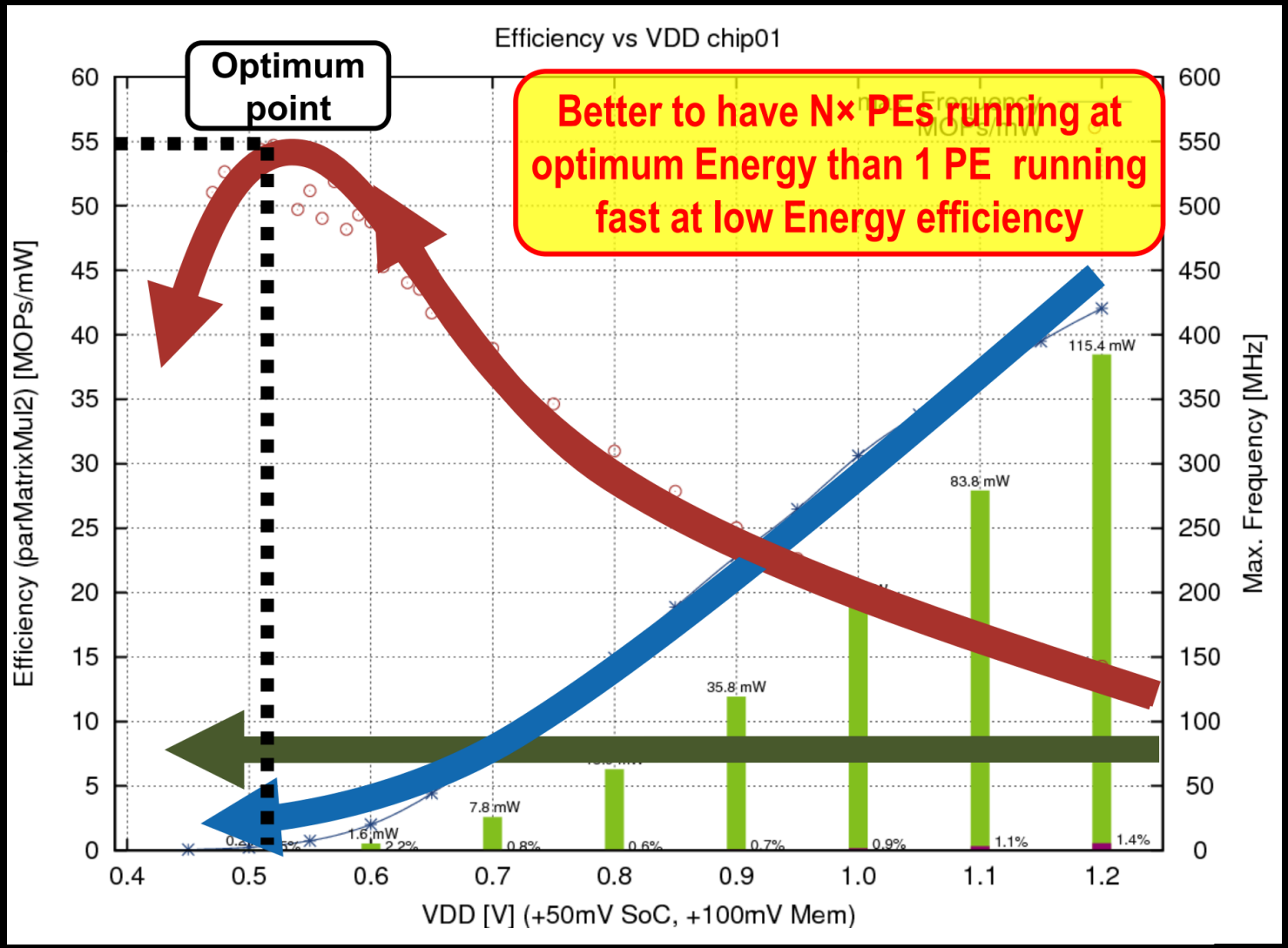
```
lp.setupi 25, Lend
  lw   x2, 0(x10!)
  lw   x3, 0(x11!)
  pv.add.b x2, x3, x2
Lend: sw x2, 0(x12!)
```

1,25 cycles/output

Multi-core to increase energy efficiency

$$P = P_{leak} + \alpha \cdot C_L \cdot f \cdot V_{dd}^2$$

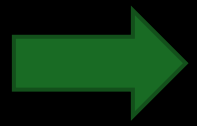
- DSP/ML workloads scale well with parallelism
- Reduce Vdd to increase overall efficiency
- More work done per Joule
- …until leakage effects start to dominate
- Parallelization overheads (sync, fork, …) minimized thanks to RISC-V

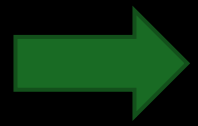


Heterogeneous computing: HW Acceleration

- Using 22FDX tech, NT@0.6V, High utilization, minimal IO & overhead

Energy-efficient RV Core → 20 pJ/Op (8bit)

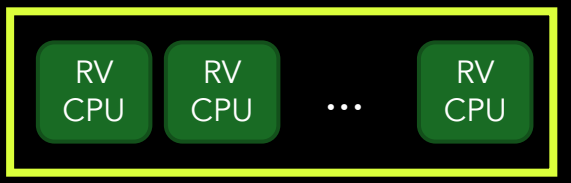
ISA-Extended RV multi-Core → 1-2 pJ/Op (8bit) (10x)  XPULP/GAP9

Specialized Data Path → 0.1-0.2 pJ/Op (8bit) (10x)  NE16/SFU

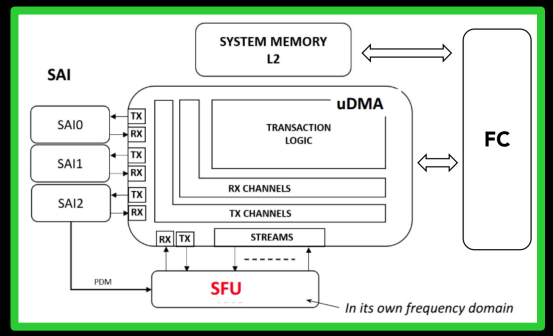
Heterogeneous computing: HW Acceleration

- Example of smart-glasses workload: **Nuance Audio (smart glasses for hearing aid)**
- **i.e. Always-on DSP + AI:**
  - DSP floating point or high precision
  - Low-latency filters
  - AI denoising

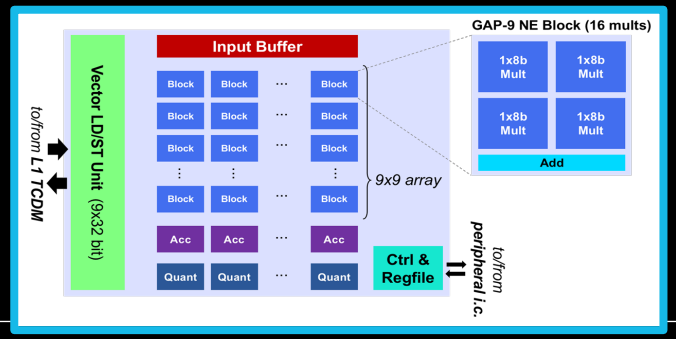
Multi-core RV



SFU



NE16

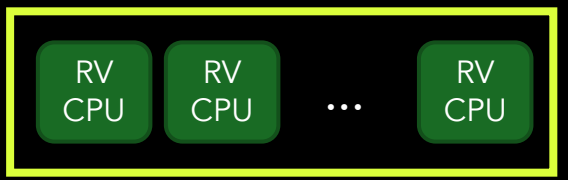


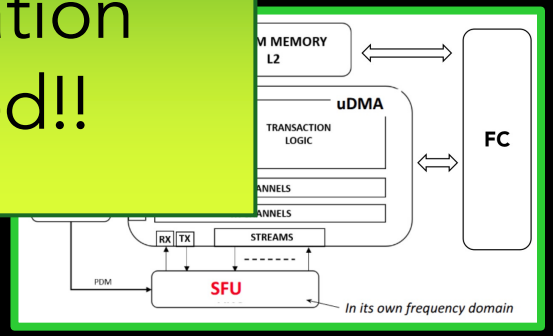
Heterogeneous computing: HW Acceleration

- Example of smart-glasses workload: **Nuance Audio (smart glasses for hearing aid)**
- **i.e. Always-on DSP + AI:**
  - DSP floating point or high precision
  - Low-latency filters
  - AI denoising

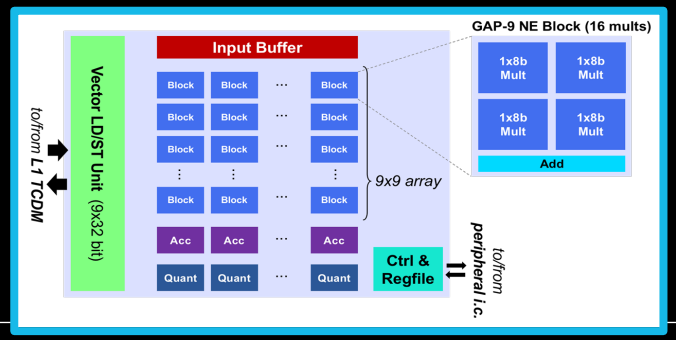
Add specialization where needed!!

Multi-core RV



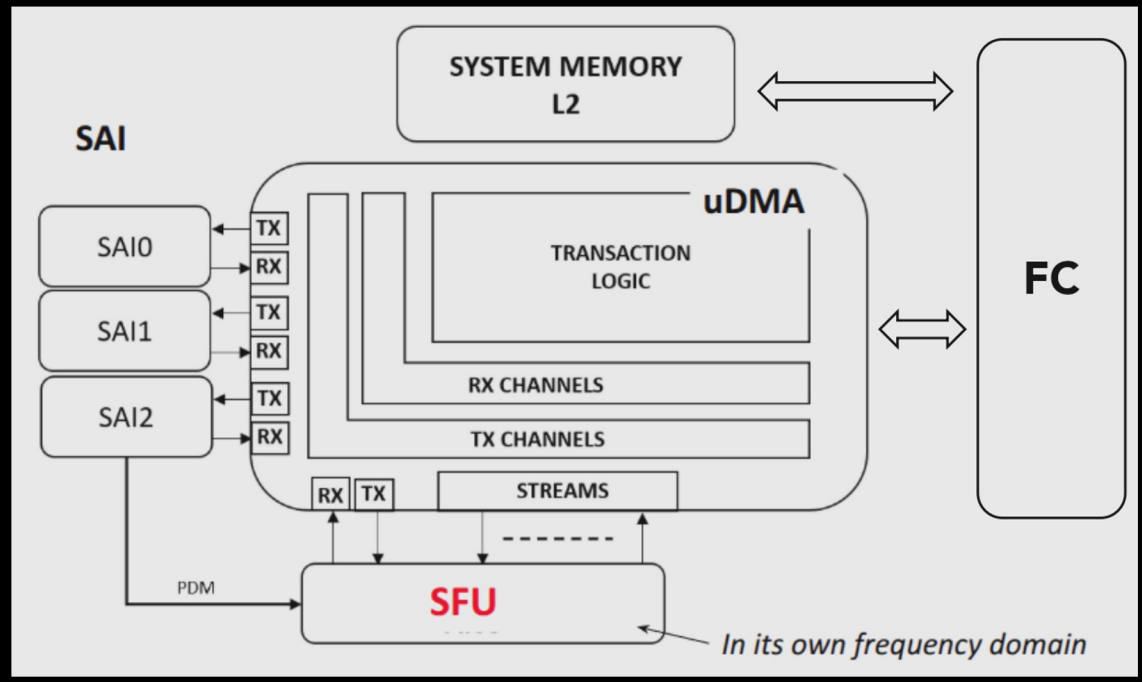


NE16

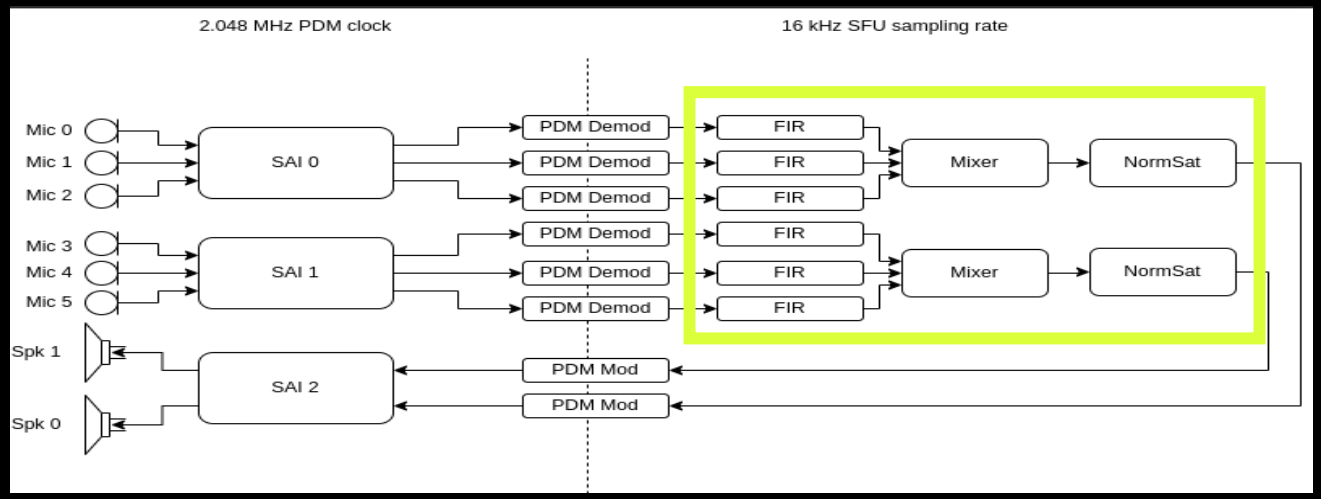


Always-on audio filtering: Smart Filtering Unit

- Stream processor for composable **filters**
- Tightly-coupled with Audio peripherals to minimize data traffic
- Re-programmable data path
  - **Instructions:** a graph of filters, splitter/joiners, sample rate converters, etc
  - **In/Outs:** memory and SAI/PDM and other peripherals
- **Examples:**
  - Single channel ANC @5mW
  - Multi channel beamformer
  - ...



Audio graph use case: Beamforming

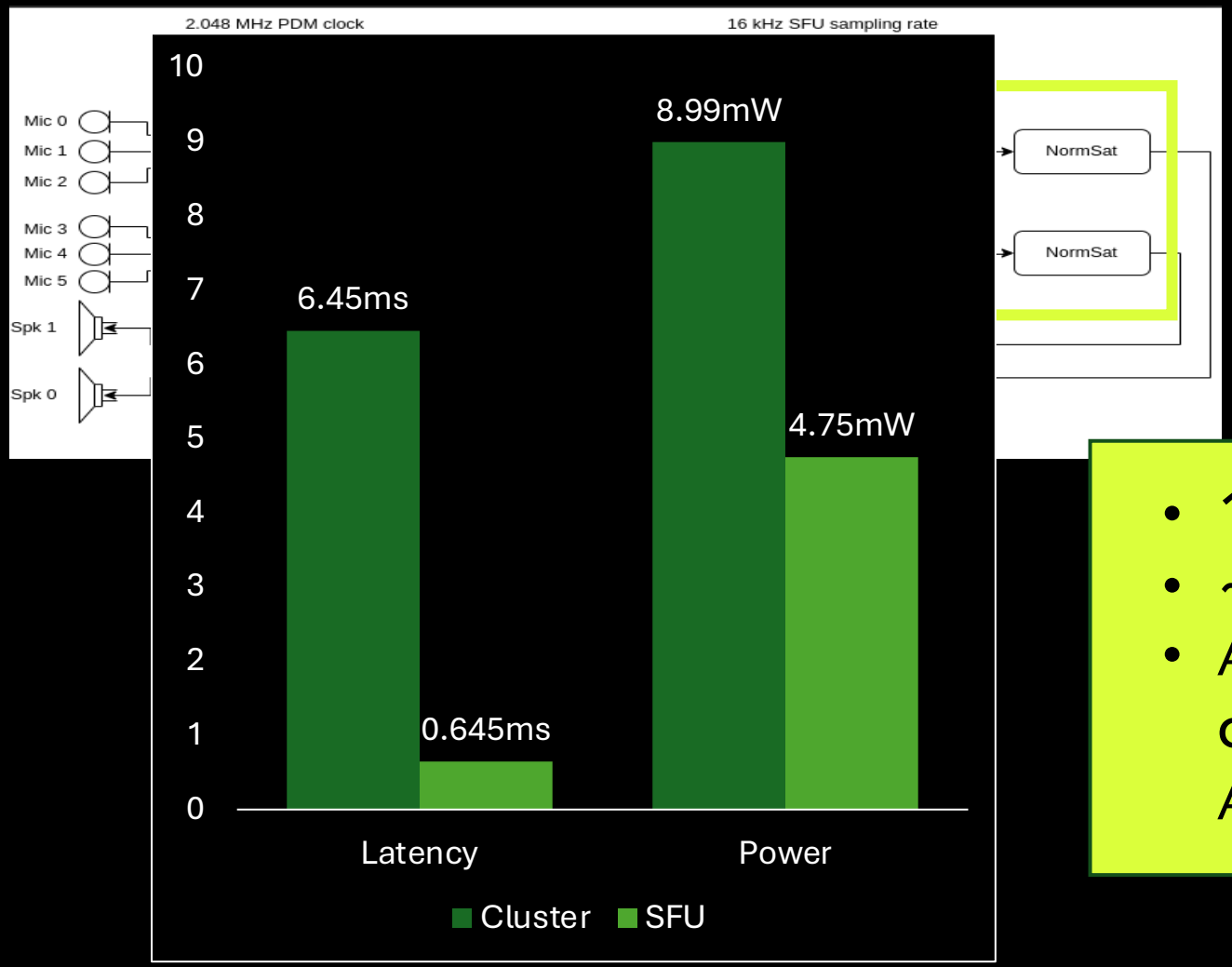


6 mics Beamformer:

- 6 PDM Demod
- 2 PDM Mod
- 6 FIR filters + Mixer + NormSat

Tested

Audio graph use case: Beamforming



6 mics Beamformer:

- 6 PDM Demod
- 2 PDM Mod

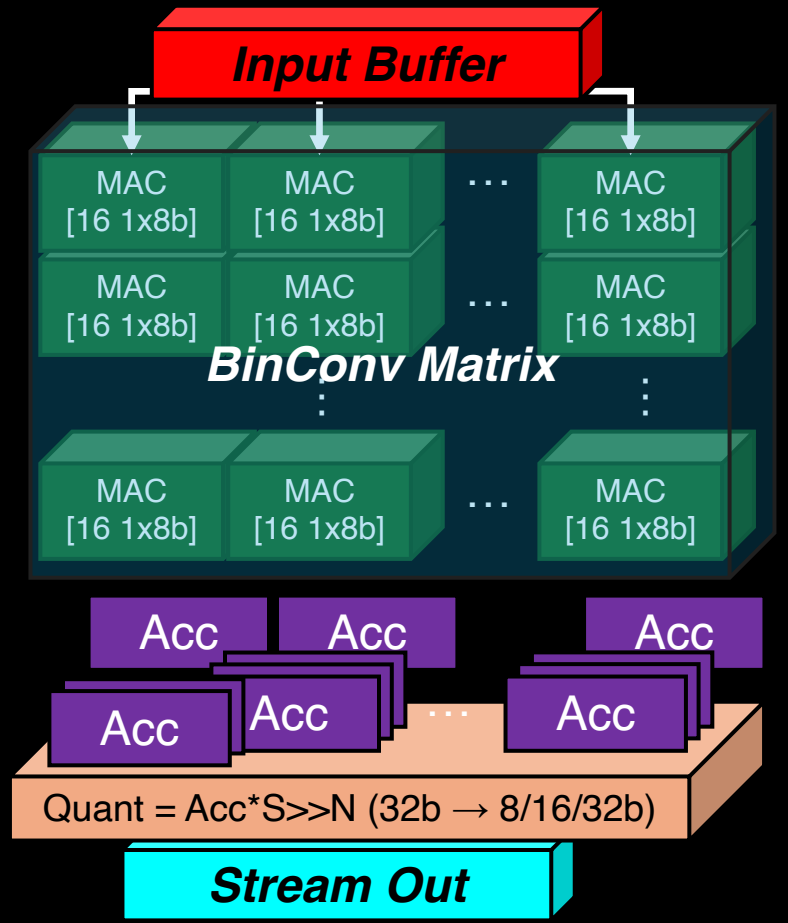
Tested

- 6 FIR filters + Mixer + NormSat

- **10x** less latency
- **~2x** less power
- Ability to run much more **complex graphs**, e.g. adaptive ANC

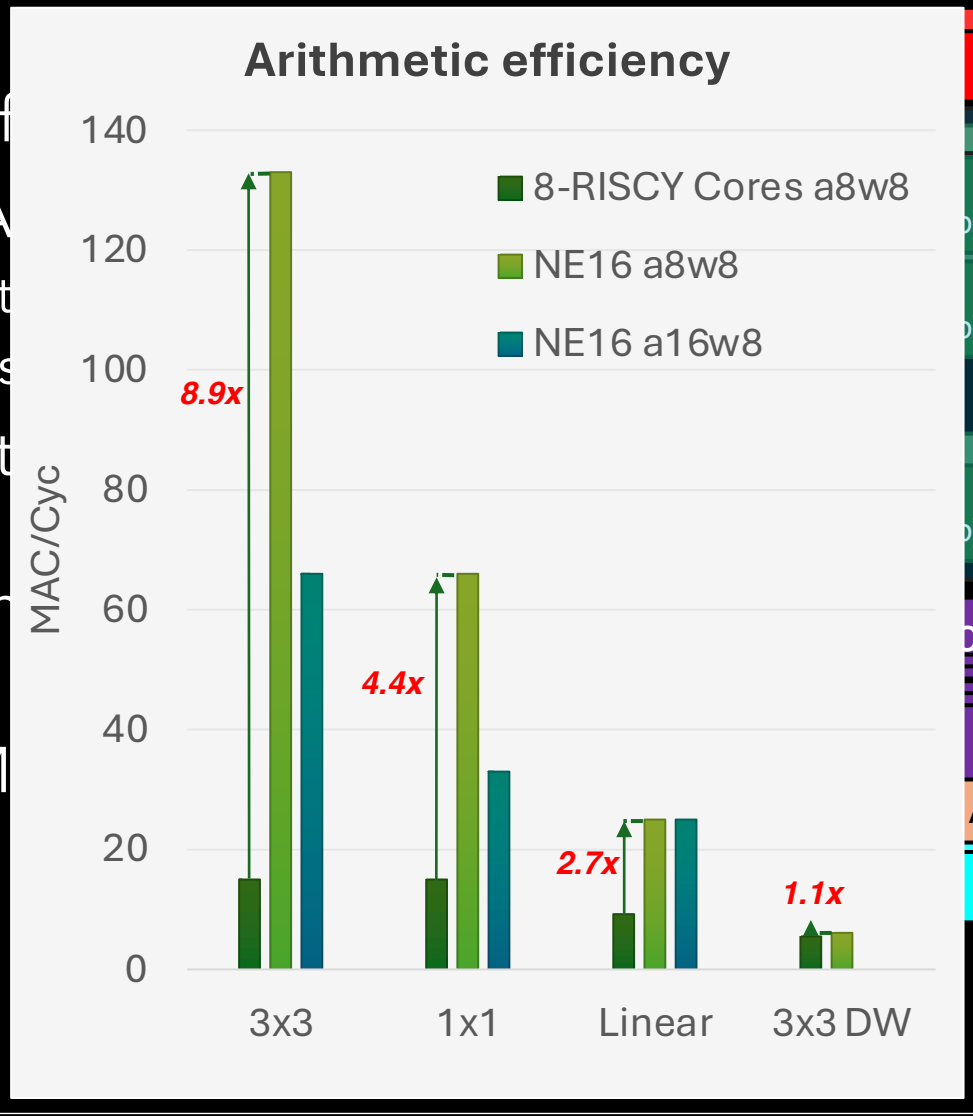
AI Acceleration: NE16

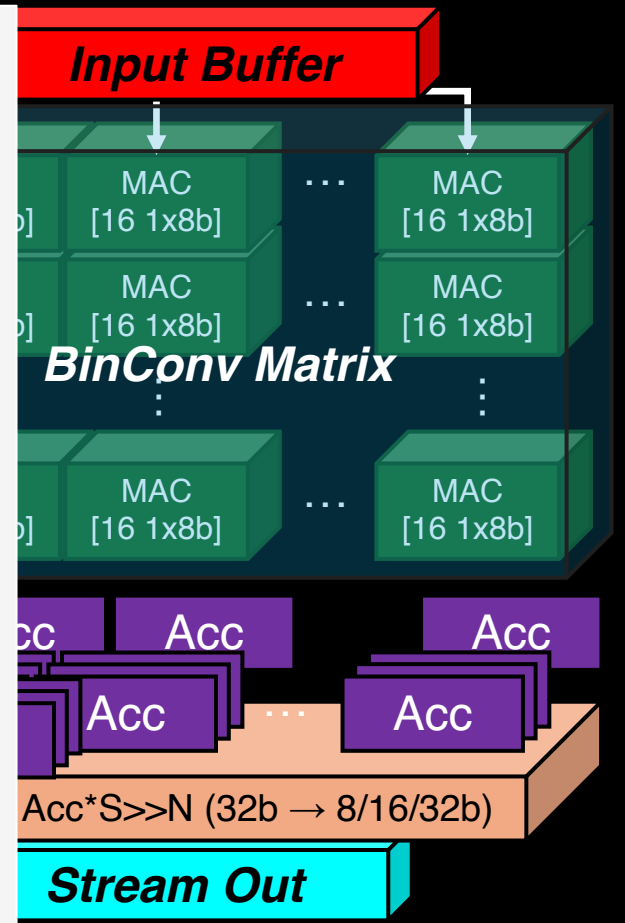
- Integer arithmetic for convolutions/GEMMS
- Bit-serial: 1x8b MAC units
  - Accelerate sub-byte arithmetic even further
  - 2,3,...,8 bit weights x 8/16 bits activations
- Reconfigurable data path for multiple usage:
  - 3x3 convolutions (CNN)
  - 3x3 Depthwise
  - MatMul/GEMM (Transformers)
  - Linear (RNN)



AI Acceleration: NE16

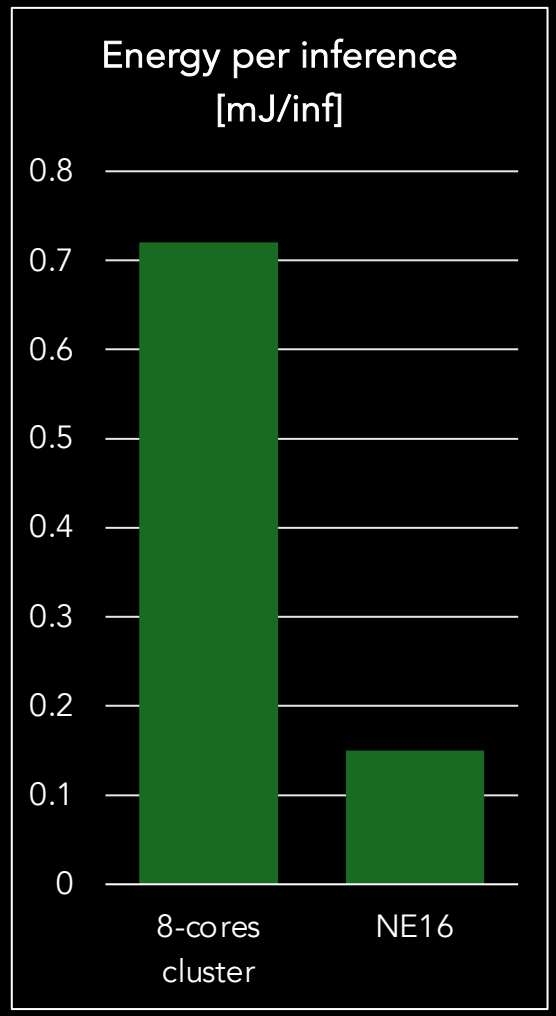
- Integer arithmetic f
- Bit-serial: 1x8b MA
  - Accelerate sub-byt
  - 2,3,...,8 bit weights
- Reconfigurable dat
  usage:
  - 3x3 convolution
  - 3x3 Depthwise
  - MatMul/GEMM
  - Linear (RNN)



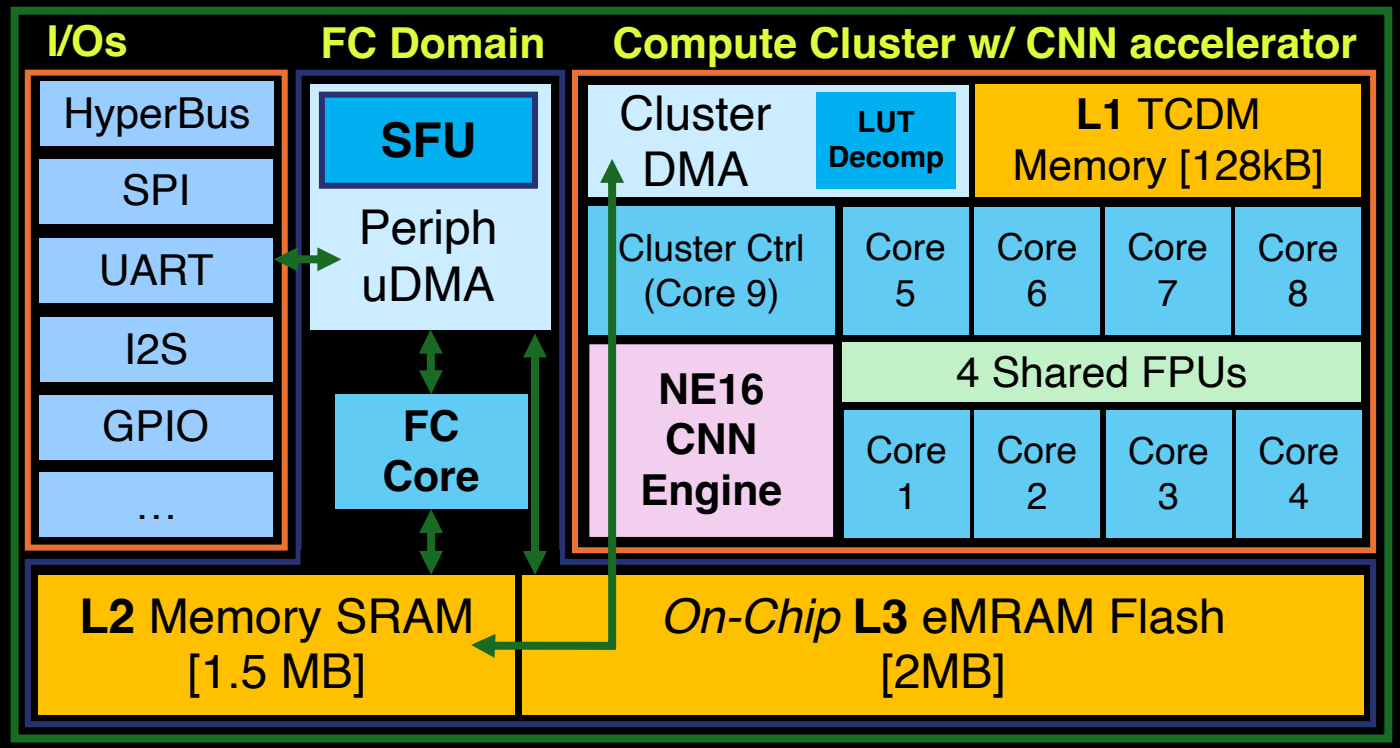


AI use case: object detection on smart glasses

- TinyissimoYOLO [Moosmann et al. https://arxiv.org/abs/2307.05999]
  - Fully quantized yolo-like object detection
  - <1M params, <57 MMAC
- Multi-core RISC-V vs NE16 -> ~5x less energy in a complex end-to-end use case



System Integration Is the Real Differentiator



The real benefits vs off-the-shelf approach come from the efficient interaction between the actors

- Local memory for data reusage
- Tightly coupling between AI engine and multi-core RISC-V
- Low-latency processor close to data streams
- Multi DMAs to move data efficiently

Conclusions

- Smart glasses as the next main smart wearable system
- Application-driven silicon architecture is the key for energy efficiency
- RISC-V offers the best trade-off for energy efficiency and flexibility in a continously evolving compute world
- Synergy between general purpose and dedicated HW ips


SMART
EYEWEAR
LAB
EssilorLuxottica
POLITECNICO MILANO 1863