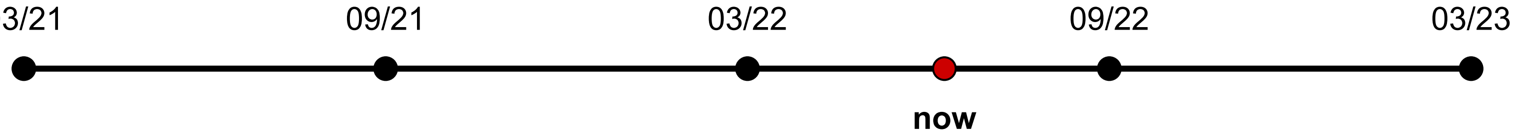
# Project Status Report

20/07/2022

Matteo Perotti
Matheus Cavalcante
Nils Wistoff
Gianmarco Ottavi

Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

# Timeline



# WP1 Workload analysis and benchmarking (70%)

- Select target sensors and **key algorithms** - develop a **benchmark suite** with performance targets for key application kernels used in multi-dimensional sensor processing and fusion for consumer devices.

# WP1 Workload analysis and benchmarking (70%)

- `AraV1` benchmarked only with `axpy`, `fmatmul`, `fconv2d`
    - Reductions?
    - Strided/Idx memory ops?
    - Masks?
    - ALU/FPU cooperation?
    - More complex permutations?

# WP1 Workload analysis and benchmarking (70%)

- `AraV1` benchmarked only with `axpy`, `fmatmul`, `fconv2d`
  - Reductions?
  - Strided/Idx memory ops?
  - Masks?
  - ALU/FPU cooperation?
  - More complex permutations?

- New benchmark suite rationale for `AraV2`
  - Diversified programs
  - Stress different datapaths
  - Show Ara merits, and its limitations

# WP1 Workload analysis and benchmarking (70%)

- Select target sensors and **key algorithms** - develop a **benchmark suite** with performance targets for key application kernels used in multi-dimensional sensor processing and fusion for consumer devices.

1. Focus on new **kernels** that can **stress** the **different DPs** of the architecture
2. Do not over-optimize them. Find a **reasonable** performance + bottlenecks **analysis**

# WP1 Workload analysis and benchmarking (70%)

- Linear algebra:
  - [i,f]matmul - workhorse, b2b vmacc
  - [i,f]conv2d - vslides
  - jacobi2d - misaligned accesses
  - axpy - memory bound
  - spmv - indexed mem ops
  - [i,f]dotp - [i,f]reductions
- Machine Learning:
  - dropout - memory bound
  - softmax - fpred, fpdivisions
  - roi_align - different mem approaches
- DSP:
  - FFT- segmented, masked permutations
  - DWT - segmented or strided
- Others:
  - fp- cos, log, exp
  - memcpy, strncmp, strncpy
  - pathfinder

# WP1 Workload analysis and benchmarking (70%)

- What’s next?
  - Finish Softmax, fdotp, strncmp/len
  - If not too tight: other 2 simple kernels
  - One sensor algorithm suggested by Huawei

- Report of the bottlenecks, with analysis of each program and execution

# WP 2: Microarchitecture design and exploration of instruction extensions (80%)

- Moving **from ARAv1 develop** the **new micro-architecture** supporting the **standard ISA** and evaluate the **impact** in HW cost and performance benefits of **instruction extensions** targeted to the workloads in WP1.

# WP 2: Microarchitecture design and exploration of instruction extensions (80%)

- Moving **from ARAv1 develop** the **new micro-architecture** supporting the **standard ISA** and evaluate the **impact** in HW cost and performance benefits of **instruction extensions** targeted to the workloads in WP1.

1. RVV1.0 with only some missing instructions
   a. Paper published (ASAP2022)
2. Improved system with new interface and memory coherence/ordering
3. Transprecision FPU (16alt and 8 bit precisions)?

# WP 2: Microarchitecture design and exploration of instruction extensions (80%)

- What’s next?

- From bottlenecks analysis
  Suggest arch modification and/or ISA extensions if needed

# WP 3: Microarchitectural optimization and design implementation (60%)

- **Moving from** version **0.1** the goal is to **develop** an **optimized microarchitecture**, design with a target frequency in the **GHz range** and support for the **ISA extension** explored in WP2. The design will emphasize **scalability to large lane count** and high energy efficiency. PPA of the resulting design will be assessed

# WP 3: Microarchitectural optimization and design implementation (60%)

- **Moving from** version **0.1** the goal is to **develop** an **optimized microarchitecture**, design with a target frequency in the **GHz range** and support for the **ISA extension** explored in WP2. The design will emphasize **scalability to large lane count** and high energy efficiency. PPA of the resulting design will be assessed

1. Optimizing uArch + Bug fixes
2. Design in GHz range
   a. Paper published (ASAP2022)
3. Scalability to large lane count
4. ISA extension: transprecision FPU

# WP 3: Microarchitectural optimization and design implementation (60%)

- What’s next?
- Scalability to large lane count
- Transprecision FPU

# WP 4: Design space exploration, software environment, open source release (60%)

- The main goal of this WP is to **verify** the design of VP4, provide **software support** (intrinsics, optimized libraries, preliminary support for vectorization). An **open-source release** of HW and SW will be prepared with **documentation**.

1. Code with LLVM intrinsics: exploring intrinsics limitations
2. **Verification**
   a. +Student project next semester
3. **Open-Source release** + Documentation

# WP 4: Design space exploration, software environment, open source release (60%)

- What’s next?

- Get more insights on the LLVM intrinsics

- Improve verification
  - Constrained-random instruction sequences

- Documentation

# Final Deliverable

- Project report
- Open-Source release

# Update on Ara

- ✓ Present paper during ASAP2022
- ✓ Implement floating-point `exp`, `cos`, `log` benchmarks
- ➢ Implement `softmax`
- ➢ Finish porting of FP-reductions