ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

NextRAN-AI

ETH- Huawei Sweden, kickoff

Marco Bertuletti mbertuletti@iis.ee.ethz.ch

Yichao Zhang yiczhang@iis.ee.ethz.ch

Mahdi Abdollahpour mahdi.abdollahpout@unibo.it

Alessandro Vanelli-Coralli avanelli@iis.ee.ethz.ch

Luca Benini lbenini@iis.ee.ethz.ch

PULP Platform

Open Source Hardware, the way it should be!

@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

Outline

- Presentation of previous work on the **TeraPool** project
- Presentation of work-packages for Y1
- First steps and research directions for WP1.Y1

Presentation of the previous work on the TeraPool project

Lightweight cores with specialized ISA

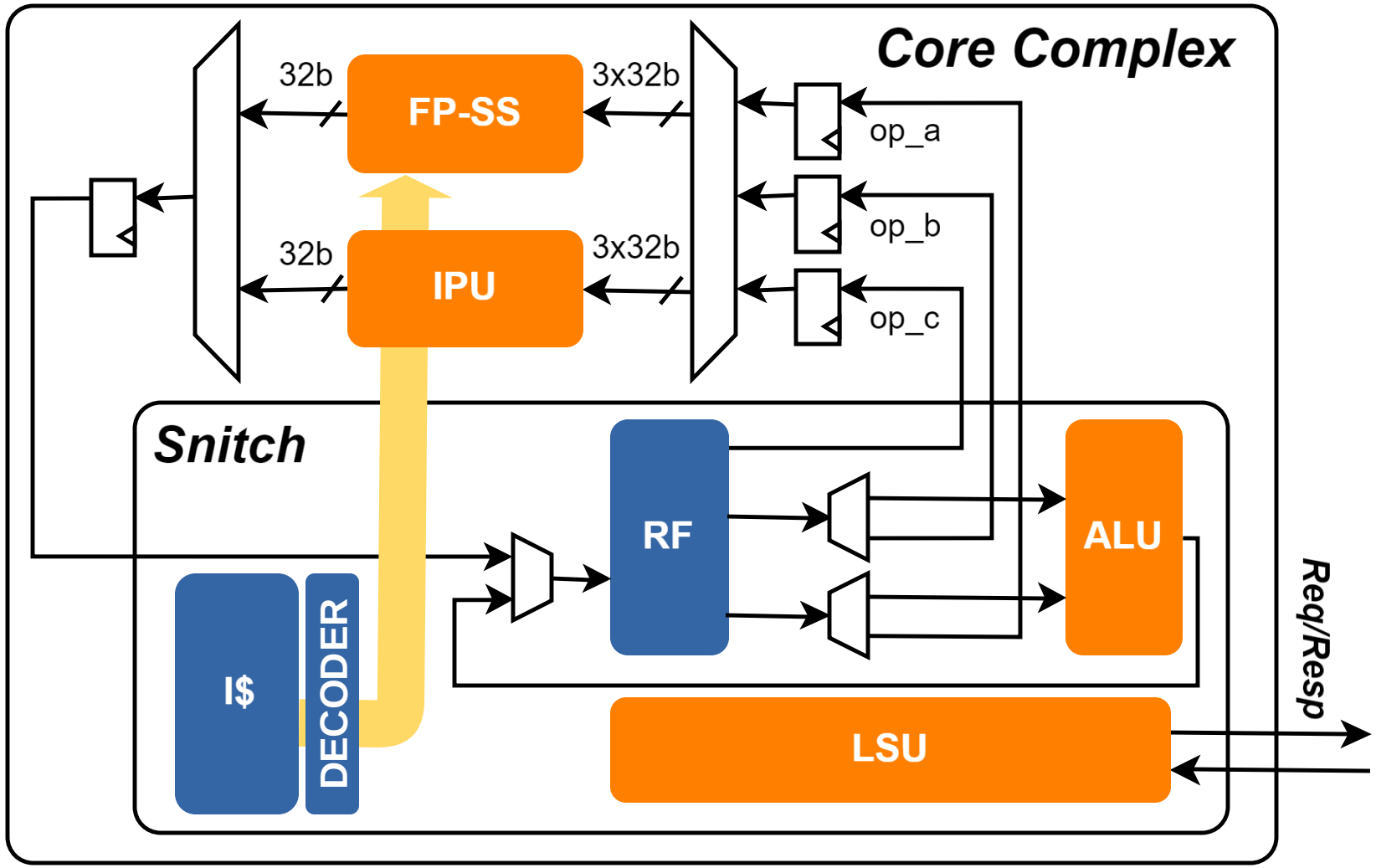
Snitch core:

- Single-stage Single-issue
- Outstanding Load&Stores

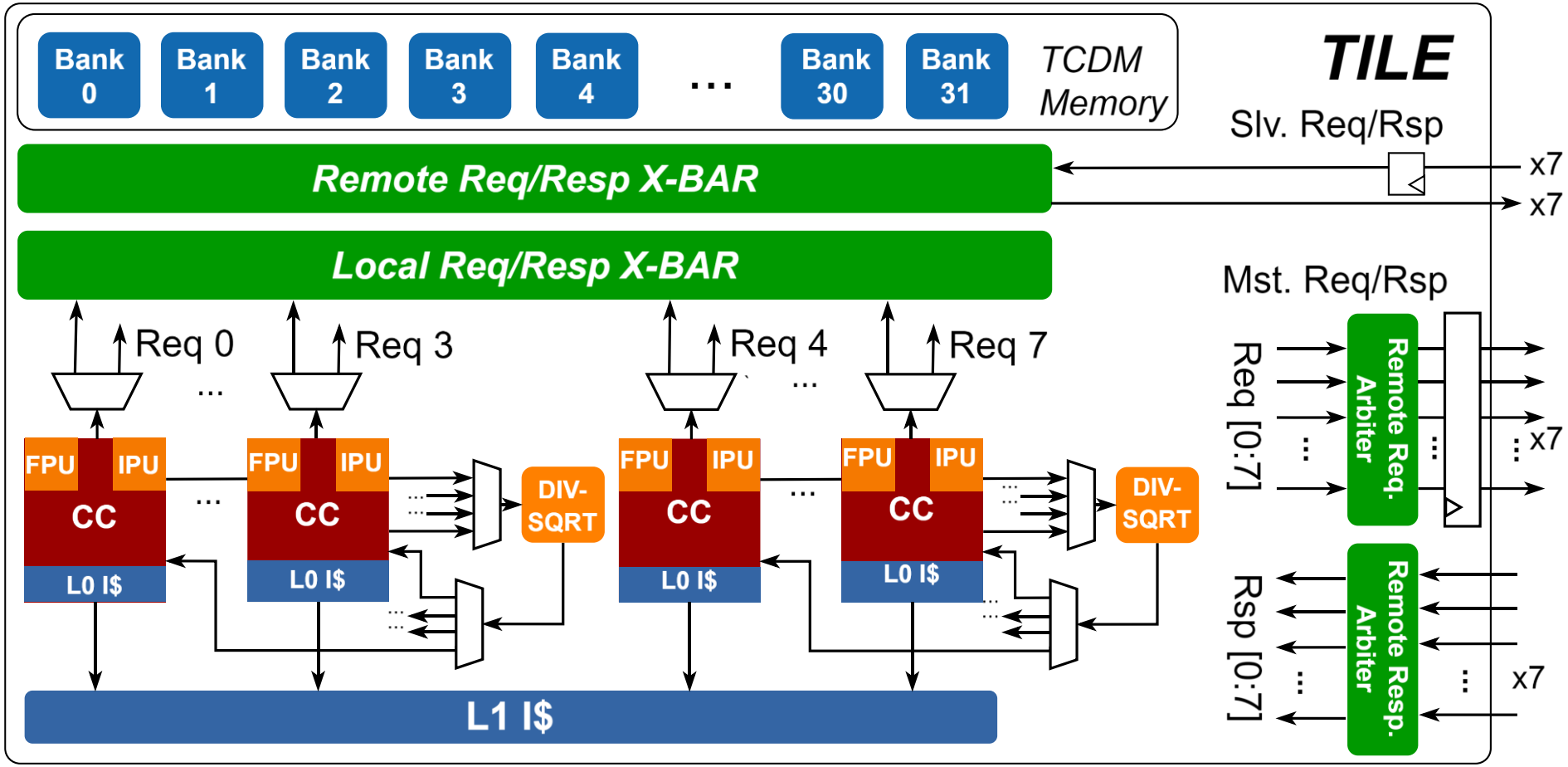
(hides latency of mem. ops.)

ISA-Extended: RV32IMA +

- zfinx & zhinx
- SIMD integer/FP
- Complex FP-dotp



Building the parallel cluster hierarchically from **TILES…**



- 8 Snitches, 4KiB Shared I$, 32KiB **Tightly Coupled Data Memory**
- **1 cycle access** to TCDM, remote access to other Tiles

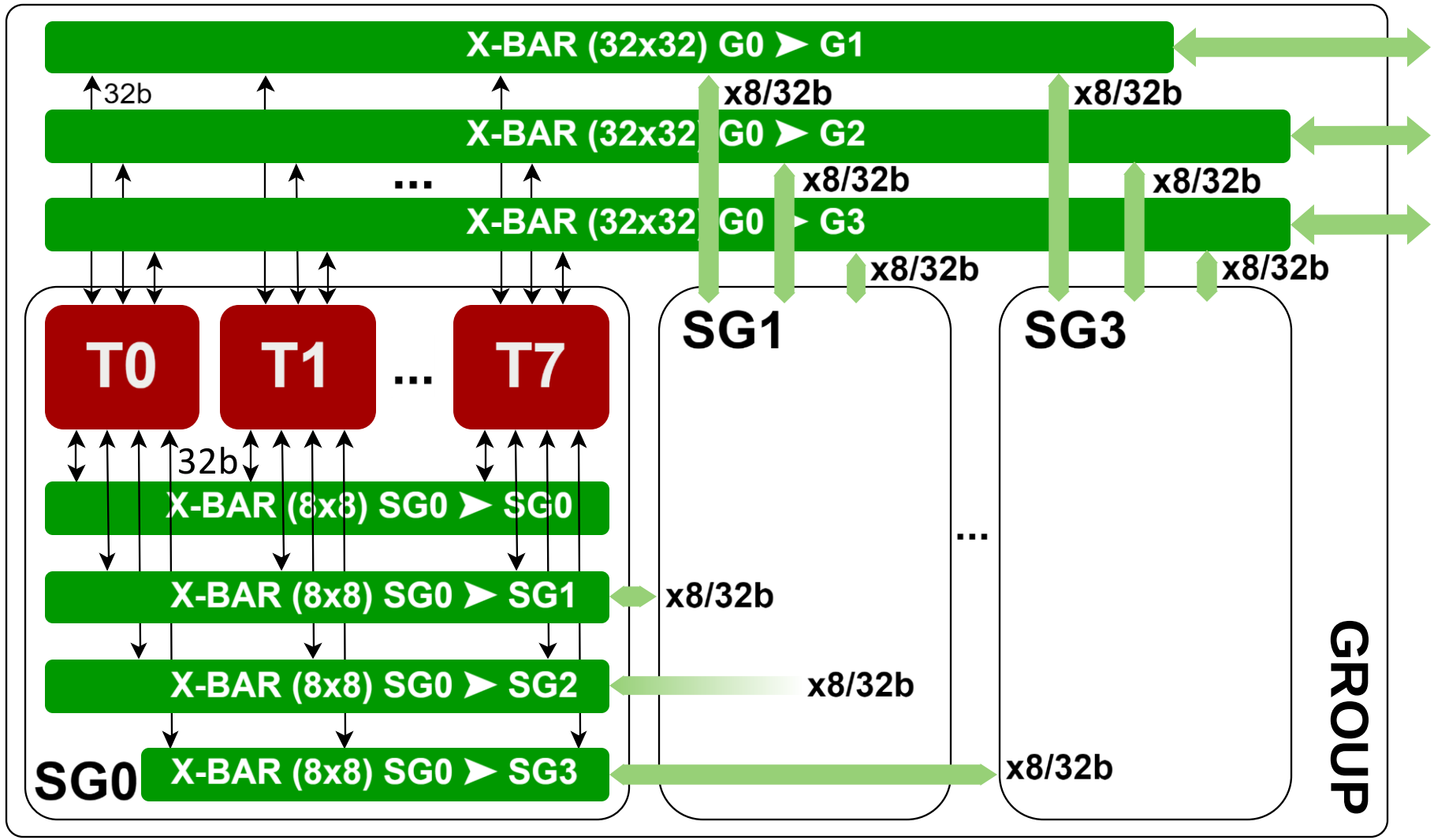
… & High bandwidth X-BAR hierarchical interconnects

Hierarchical Design:

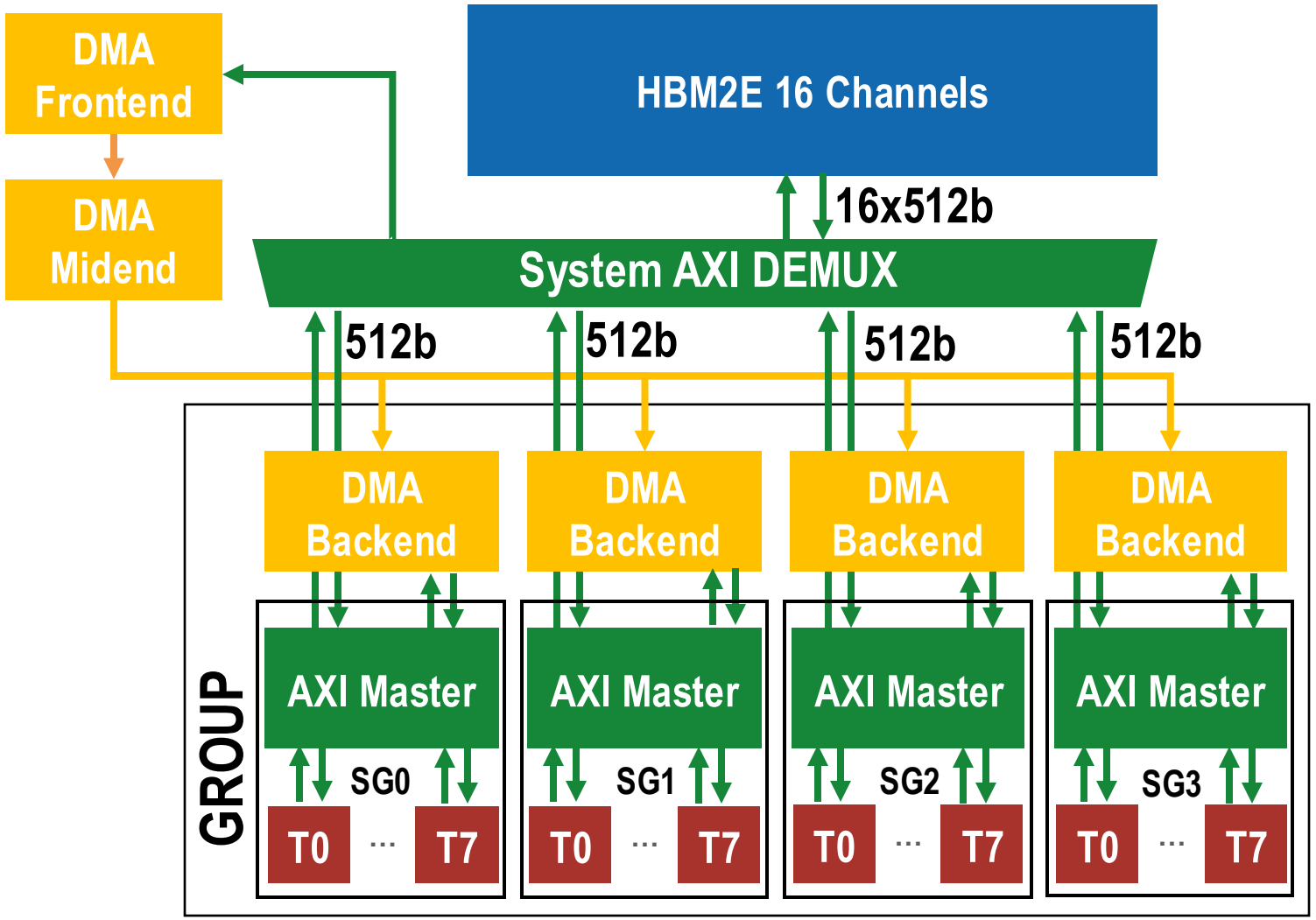
- **8x cores/**Tile
- **8xTiles/**SubGroup
- **4xSubGroups/**Group
- **4xGroup/**Cluster

Hierarchical interconnects

→ Cores in any Tile can access the TCDM of other Tiles (**7, 9, 11 cycles**)

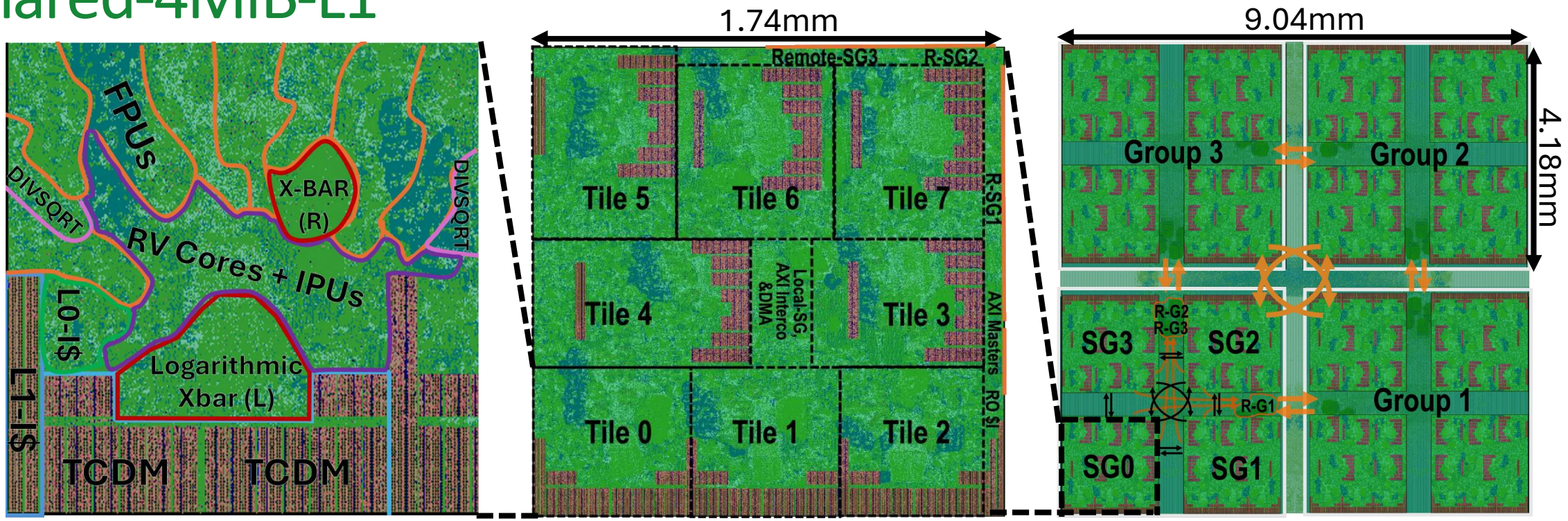


Connected to **HBM** for up to **900Gbps uplinks**



- A hierarchical AXI interconnect allows access to main-memory
- Each core can program the **DMA-Frontend** through AXI-writes
- The **Backend** initiates transfers from DRAM to each SubGroup

TeraPool: Physically-feasible, 1024 FP RISC-V Cores Shared-4MiB-L1



Methodology:

- GlobalFoundries' 12P+ FinFET
- Synopsys' FusionCompiler 2022.03
- Synopsys' PrimeTime 2022.03
- WC: SS/0.72V/125C ;TT: TT/0.80V/25C

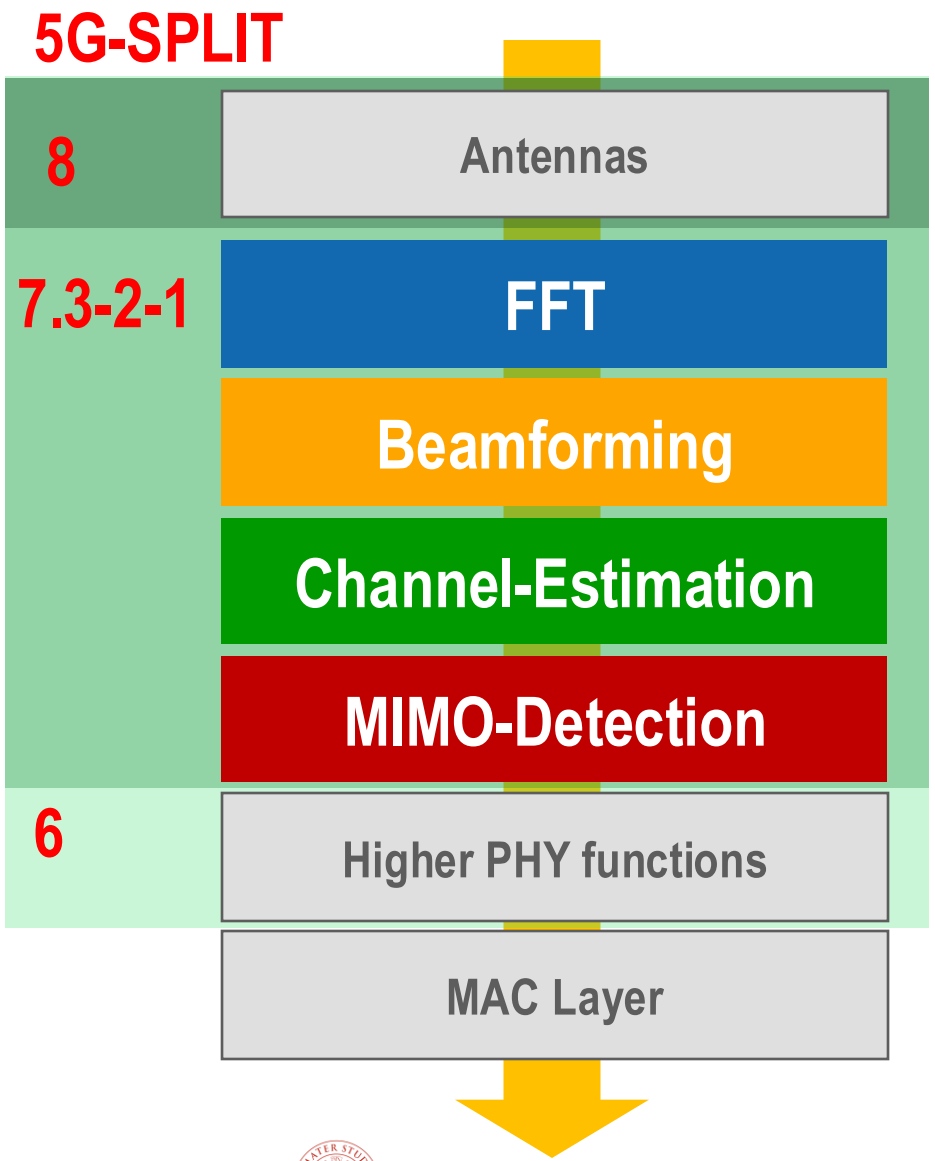
Area:

- Subgroup: 1.74 x 1.74 mm^2 (58% utilization)
- Group: 4.2 x 4.2 mm^2
- Cluster: 9 x 9mm^2

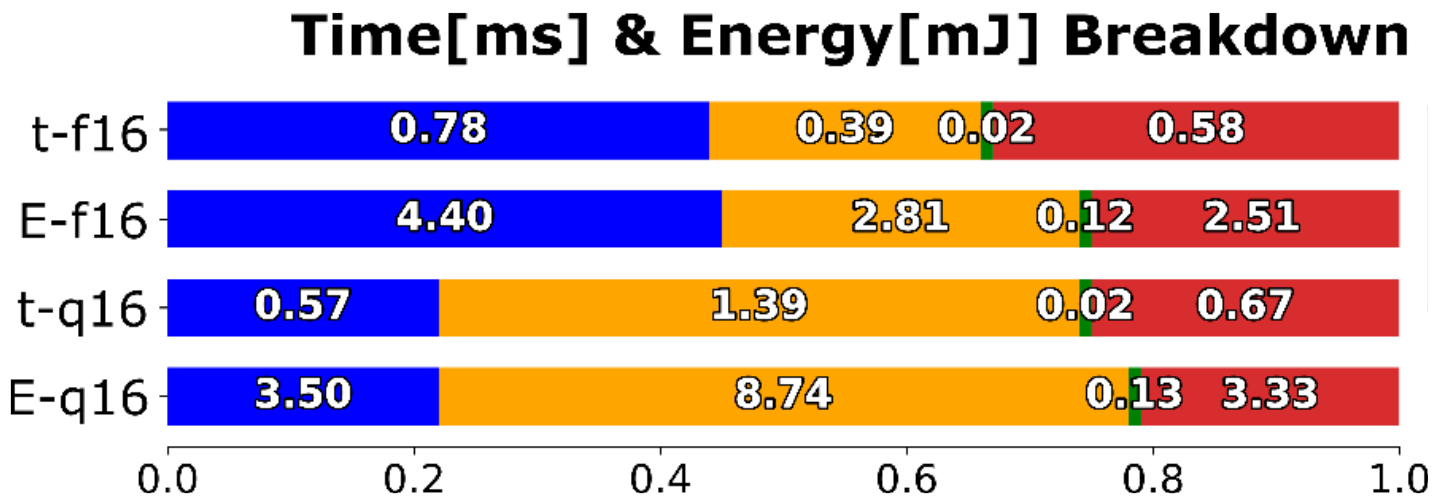
Performance:

- 730MHz @WC, 910MHz @TT

TeraPool accelerates Lower-PHY tasks in-line



- **Physical Uplink Shared Channel**
  - 273 Resource Blocks, 50MHz, 15kHz sub-carrier spacing
  - 64RX, 4TX, 32 Beams
  - 1.77ms (floating-point) 2.5ms (fixed-point) latency,
  - 5.5W average power-consumption



Presentation of work-packages for Y1

Project plan

YEAR 1	
WP1.Y1	**Hardware and Architecture -** SoA Analysis, requirement analysis , architectural specification, initial design
WP2.Y1	**AI Models -** Exploration, selection, initial benchmarking
WP3.Y1	**Software and mapping -** SoA Analysis, requirements. Codesign specification and initial design

YEAR 2	
WP1.Y2	**Hardware and Architecture -** Design, optimization and characterization
WP2.Y2	**AI Models -** Benchmarking and tuning of models. End-to-end heterogeneous (ML + non-ML functions needed) workload benchmarking extension
WP3.Y2	**Software and mapping -** Development and optimization of software stack for heterogeneous workloads

- M1.1.Y1 (M6) SoA analysis + benchmarking report
- M1.2.Y1 (M12) Open-source HW, SW (preliminary)
- M1.1.Y2 (M24) Open-source HW, SW (final)
- M1.2Y2 (M24) Open-Source HW, SW documentation (final) + benchmarking report

WP1.Y1: Hardware and Architecture

- **Classification of AI models for baseband processing:**
  - Function (CSI, CSI feedback, beam management, …);
  - Model type (RNN, CNN, transformer, …);
  - Memory footprint and compute complexity.
- **First assessment of the hardware platform requirements to meet the latency-throughput constraints.**

WP2.Y1: Exploration, selection, initial benchmarking

Moving from the SoA analysis (Y1.WP1)…

- **Selection of the models that optimize the telecommunication performance:**
  - Quality of service KPIs (BER, MSE vs. input SNR),
  - B5G specifications (bandwidth, number of users, and number of base station antennas).
- **Initial benchmarking of the model**
- **Possible extension to the B5G use-case if the desired KPI are not matched.**

WP3.Y1: Software and Mapping

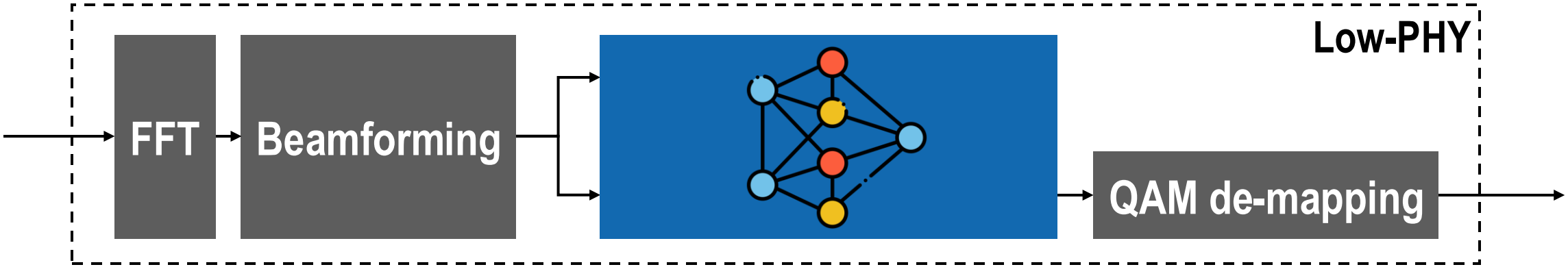
- **Implementation of parallel ML micro-kernels for the TeraPool architecture.**
- **Scheduling of the selected ML model on TeraPool.**
- **Analysis of the hardware KPIs (compute-elements and memory-bandwidth utilization, power consumption) to identify bottlenecks.**

In case the desired latency-throughput is not matched by the current TeraPool architecture we will consider **domain-specialization** and merge the output of **WP3.Y1** with the output of **WP1.Y2 «Design of ML-oriented acceleration»**

First steps and research directions for WP1.Y1

Focus on CSI and full MIMO AI-receivers

- **Channel State Information (CSI)**
  - Influences the performance of the receiver (BER vs SNR)
  - Must be performed **at the edge**, to avoid high-latency data transfer on the fronthaul
  - Compute requirements scale with the MIMO-size (UEs/BW and number of antennas)
- **We target full MIMO receivers → (OFDM, Beamforming, CHE, detection, demapping)**
  - Direct comparison with the work on TeraPool PUSCH
  - Partial **model-driven** and **data-driven** rx, depending on blocks with highest perf. gains



Models currently under study

Name	Processing	NSC	NRXxNTX	Modulation	Model	Gain wrt Conventional receiver @BER10^{-3}
Deep-RX	Ch.Est. + Det.	312	2x1	16QAM	ResNet	2.5 dB *
Deep-RX MIMO	Ch.Est. + Det.	312	16x4	16QAM	ResNet	2.5 dB *
Neural-RX RT	Ch.Est. + Det.	1584	4x2	16QAM	CGNN	1.0 dB *
Neural-RX	Ch.Est. + Det.	1584	4x2	16QAM	CGNN	2.0 dB *
… Extend to more subcarriers, RX, TX for B5G use-cases						

* LS Channel Estimation + LMMSE Detection

Next Steps:

- Evaluate other possible models
- Evaluate computational complexity of the involved operators
- Compare computational complexity with capabilities of TeraPool

Work planned for the next period

- **Deep-learning receivers for the Lower-PHY survey**
  - Telecommunication performance (NTX, NRX, NPRBs, BER/MSE vs SNR)
  - Model type (e.g. Fully-Connected, Convolutional, Transformer, …) and operators involved (e.g. matrix-multiplication, convolution, depthwise-separable convolution, ...)
  - Evaluation of the model complexity -> number of OPS, memory footprint.
- **Performance required for real-time operation and comparison with TeraPool**
  - OPS/s in real-time operation
  - Compare OPS/s and memory footprint with TeraPool performance and L1 capacity

Thank you!