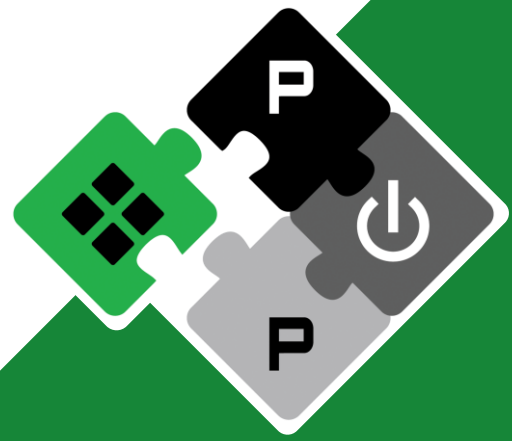
TeraPool-SDR: An 1.89TOPS 1024 RV-Cores 4MiB Shared-L1 Cluster for Next-Generation Open-Source Software-Defined Radios

Integrated Systems Laboratory (ETH Zürich)

Yichao Zhang yiczhang@iis.ee.ethz.ch
Marco Bertuletti mbertuletti@iis.ee.ethz.ch
Samuel Riedel sriedel@iis.ee.ethz.ch
Alessandro Vanelli-Coralli avanelli@iis.ee.ethz.ch
Luca Benini lbenini@iis.ee.ethz.ch

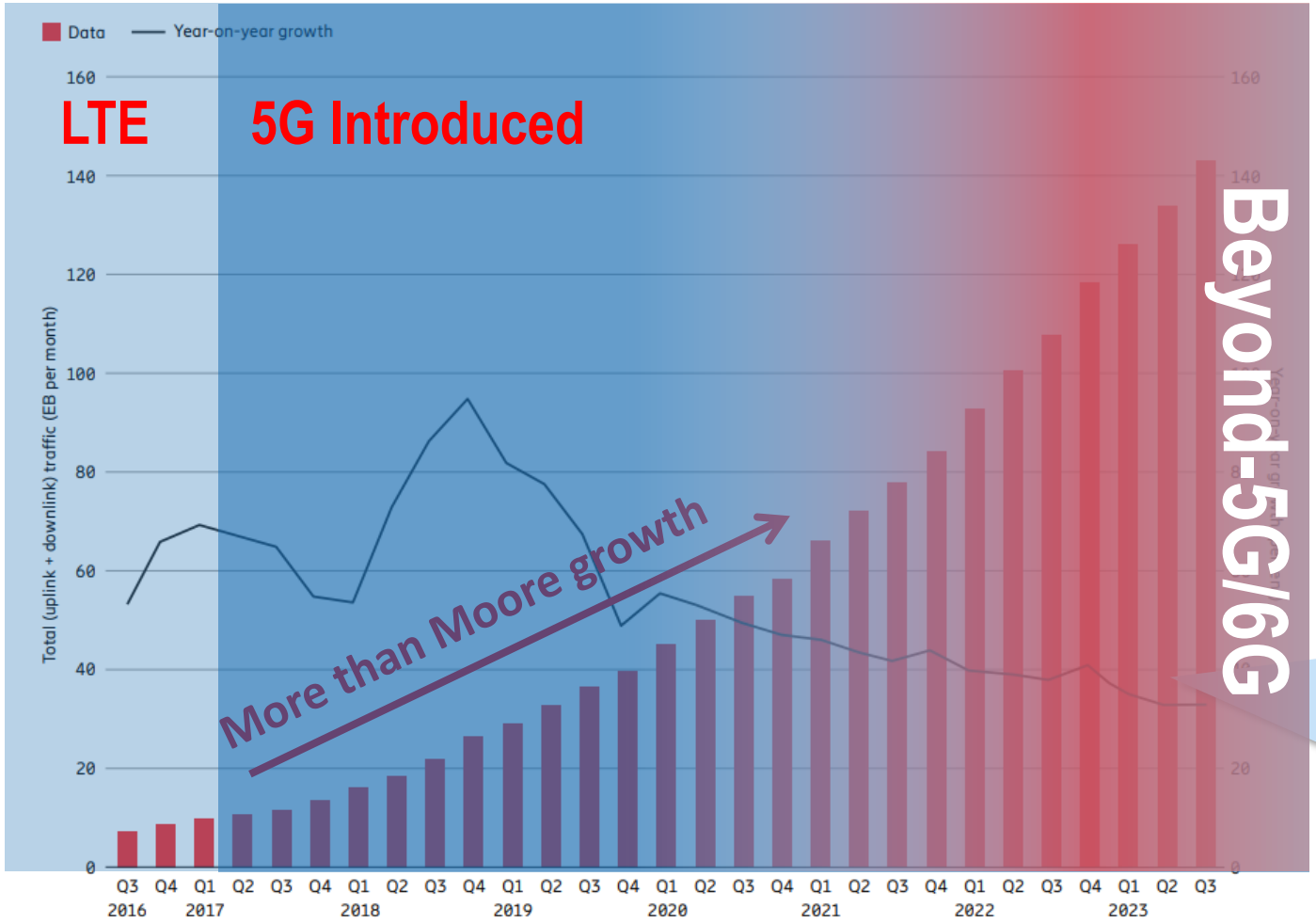


@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

PULP Platform
Open Source Hardware, the way it should be!

Network data traffic grows at Moore's rate

Mobile Network data traffic, Ericsson Mobility report November 2023



Increasing subscriptions

Increasing average data volume per subscription (mainly due to **video streaming**)

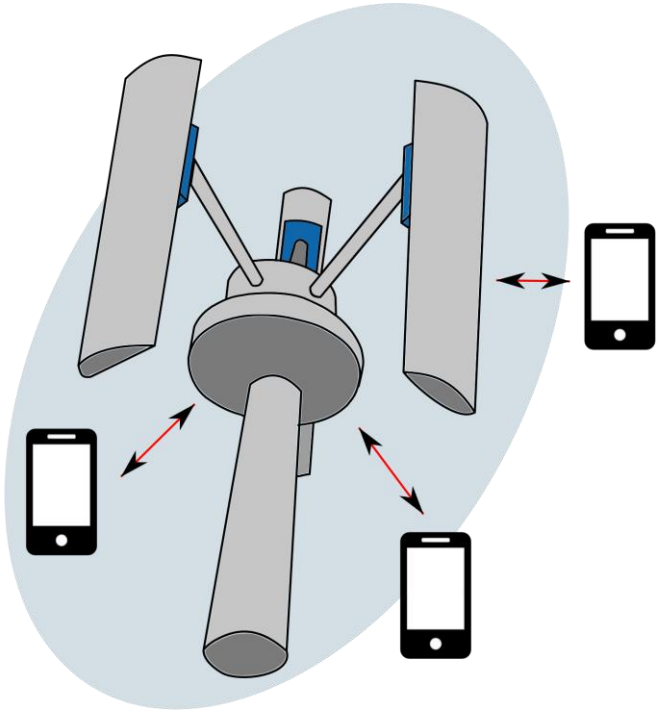
Growth in percent is falling, but data traffic will keep growing until 2029, driven by **X-reality**

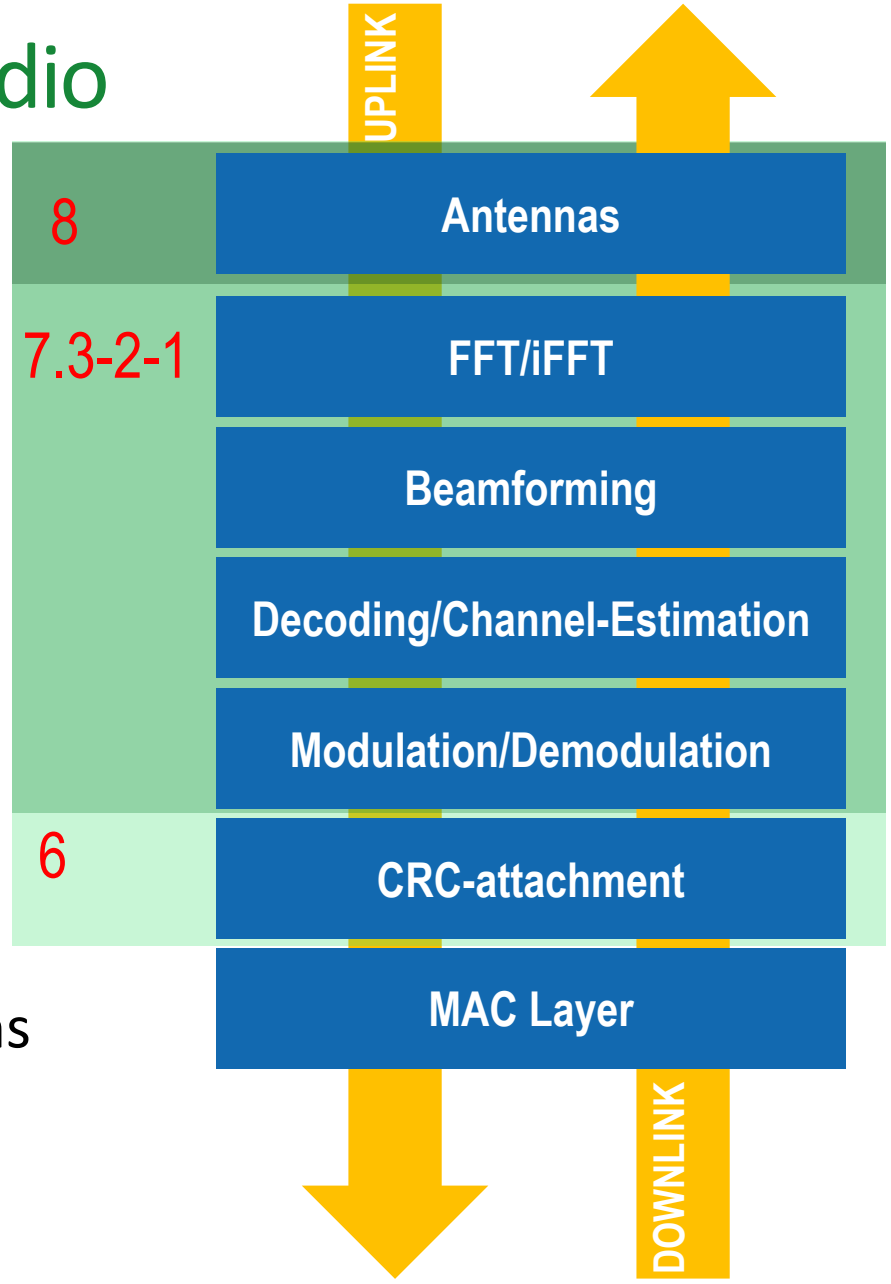
Software Defined Radio

- Reduce **time to market**
- Adaptation to evolving standard (LTE → 4G → 5G → 6G in <10 years)
- Increase **return on investments** (programmable components have longer lifetimes)

Platform	PHY processors	ISA	Multi-Core	5G-split
EdgeQ S-series	TXU processor ARM Neoverse	RISCV ARM	✓ ✓	6-7.2
Picocom/PC802	Ceva XC12 1280-bit Scalar-processor cluster	RISCV RISCV	X ✓	7.2
Marvell/Octeon10	ARM Neoverse DSP processors + Acc.	ARM NA	✓	7.X
Qualcomm X100	NA	NA	NA	7.X
Nokia ReefShark	DSP processors + Acc.	NA	NA	7.X
Ericsson Silicon	NA	NA	NA	7.X

Software Defined Radio





5G-split
6-7.2
7.2X
7.X
7.X
7.X
7.X

Split defines which operations are executed in the network base-stations

Software Defined Radio

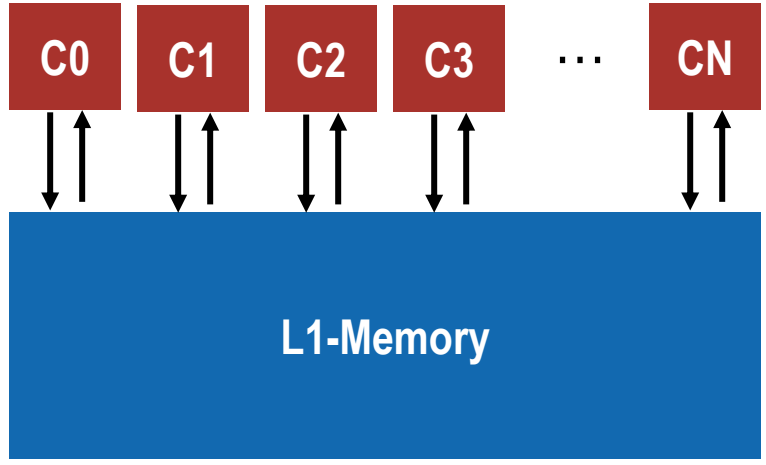
- VLIW/Vector Processors/MultiCore are dominant architectures
- They address the large scale B5G computations

Platform	PHY processors	ISA	Multi-Core	5G-split
EdgeQ S-series	TXU processor ARM Neoverse	RISCV ARM	✓ ✓	6-7.2
Picocom/PC802	Ceva XC12 1280-bit Scalar-processor cluster	RISCV RISCV	X ✓	7.2
Marvell/Octeon10	ARM Neoverse DSP processors + Acc.	ARM NA	✓	7.X
Qualcomm X100	NA	NA	NA	7.X
Nokia ReefShark	DSP processors + Acc.	NA	NA	7.X
Ericsson Silicon	NA	NA	NA	7.X

BIG workload → **MANY** cores & **BIG** memory

Shared memory parallel cluster:

- Parallel execution + final synchronization
- Streamlined programming model



Must scale-up to accomodate 5G workloads

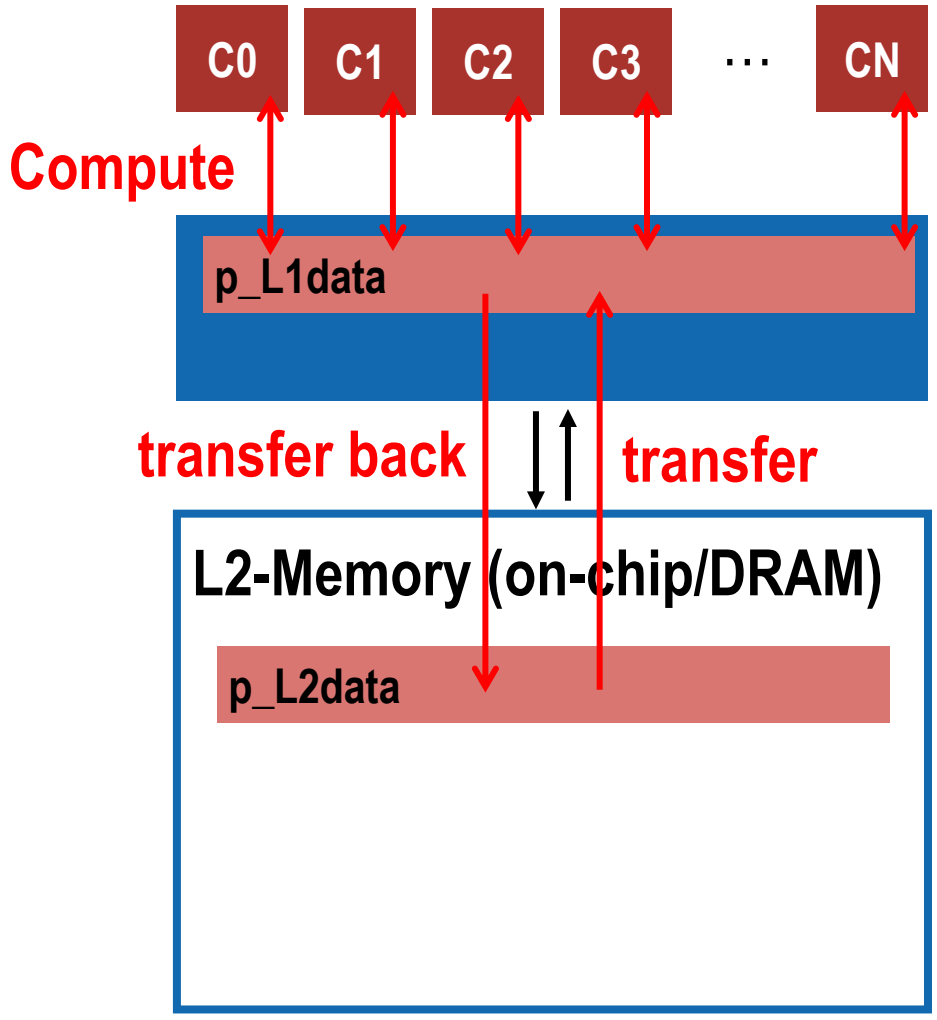
BIG workload → **MANY** cores & **BIG** memory

Example:

FFT 64/128 antennas, frequency range 275-RB → **0.8/1.6 MiB**

Beamforming, 64/128 antennas, 32 beams → **1.2/2.1 MiB**

If it doesn't fit in L1 we have to **move data from L2**



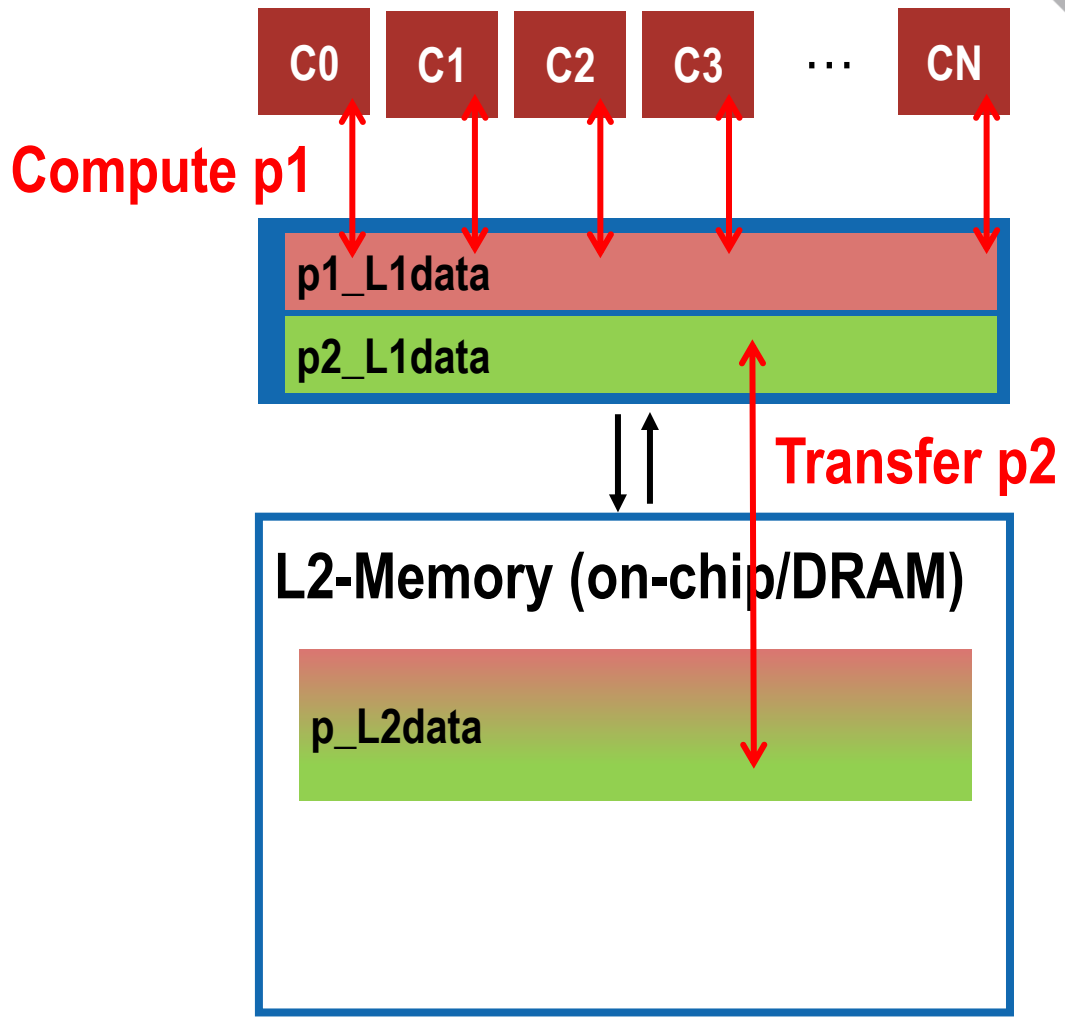
BIG workload → **MANY** cores & **BIG** memory

Example:

FFT 64/128 antennas, frequency range 275-RB → **1.6/0.8 MiB**

Beamforming, 64/128 antennas, 32 beams → **2.1/1.2 MiB**

We can **double buffer** the computation to hide **transfer latency**, but 2X **L1** space is required!



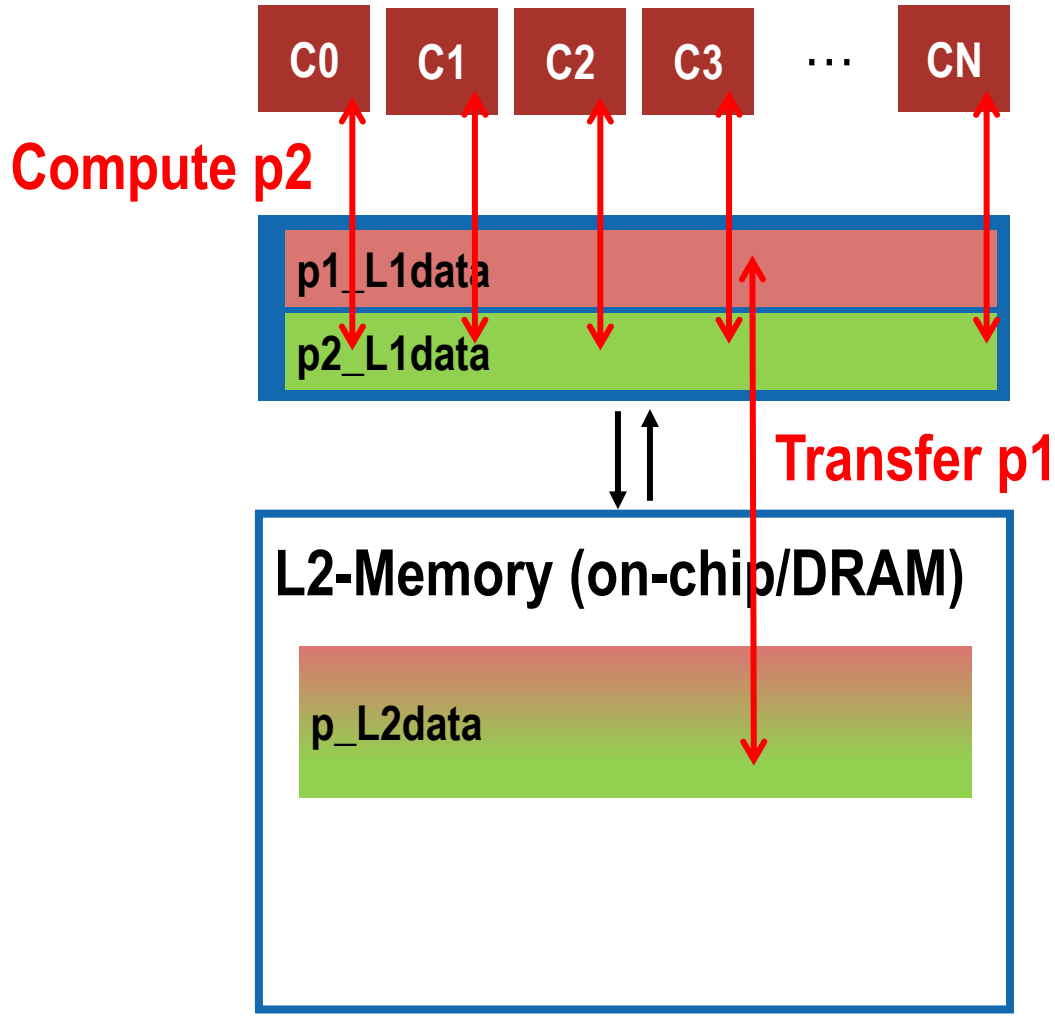
BIG workload → **MANY** cores & **BIG** memory

Example:

FFT 64/128 antennas, frequency range 275-RB → **1.6/0.8 MiB**

Beamforming, 64/128 antennas, 32 beams → **2.1/1.2 MiB**

We can **double buffer** the computation to hide **transfer latency**, but 2X **L1** space is required!



BIG workload → **MANY** cores & **BIG** memory

Example:

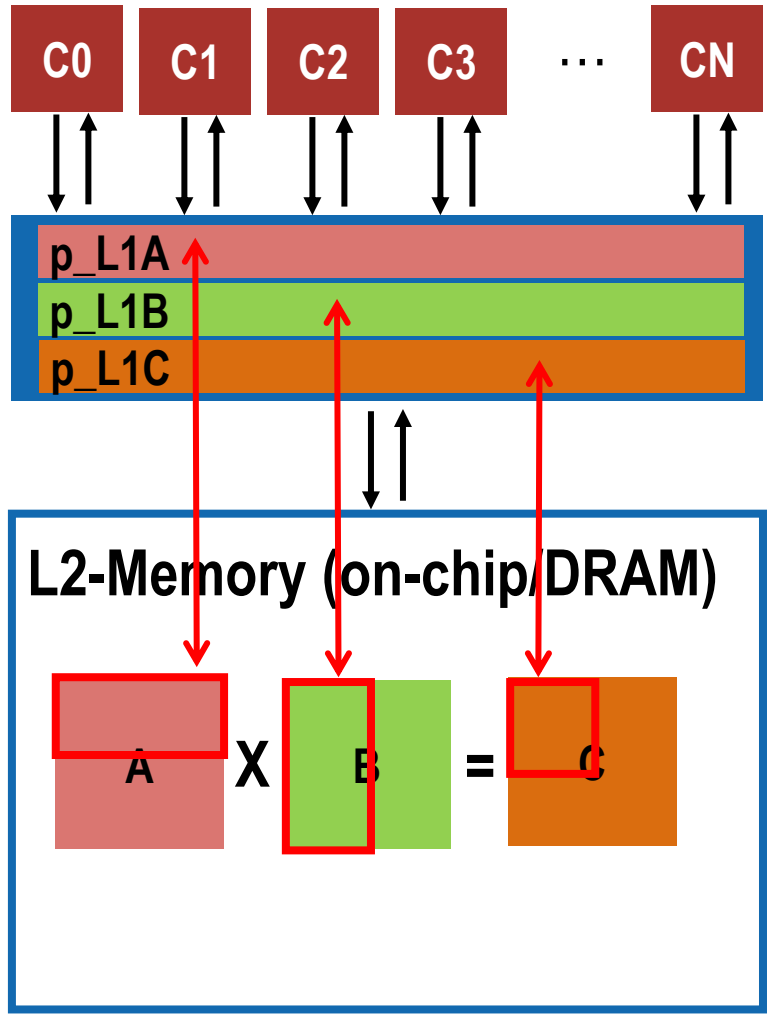
FFT 64/128 antennas, frequency range 275-RB → **1.6/0.8 MiB**

Beamforming, 64/128 antennas, 32 beams → **2.1/1.2 MiB**

Big workloads are not solved with one transfer: data in L2 is **tiled** and we do multiple rounds

→ Synchronization + transfer overhead

→ Difficult to optimize (best tiling for each kernel?)



BIG workload → **MANY** cores & **BIG** memory

Example:

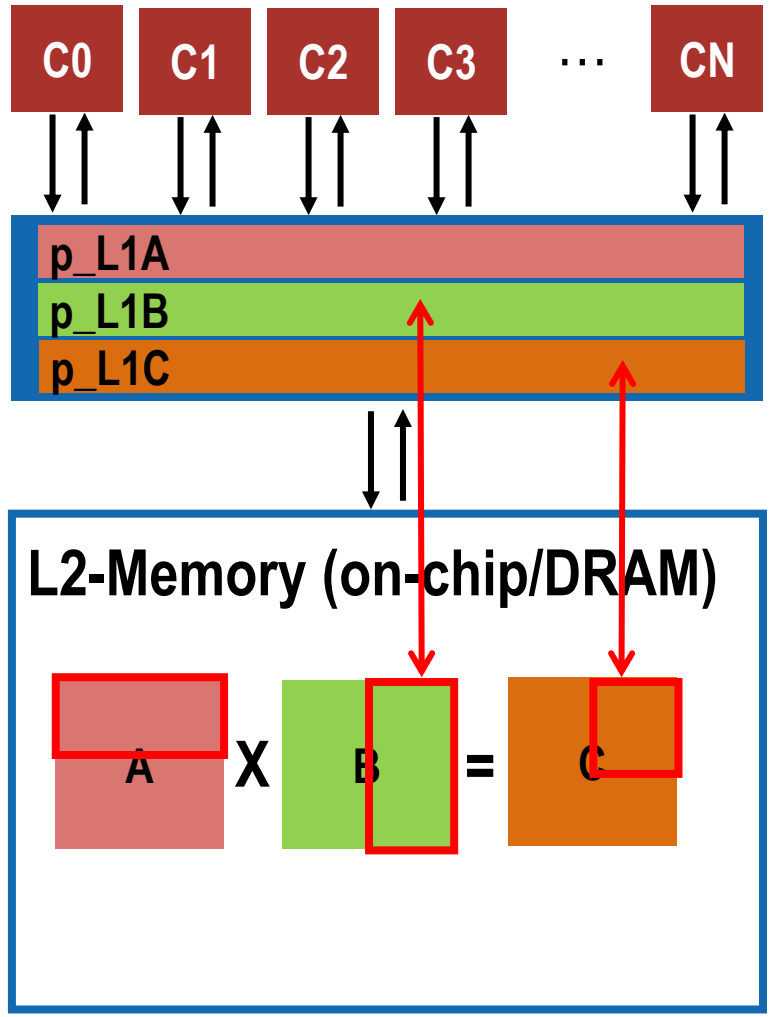
FFT 64/128 antennas, frequency range 275-RB → **1.6/0.8 MiB**

Beamforming, 64/128 antennas, 32 beams → **2.1/1.2 MiB**

Big workloads are not solved with one transfer: data in L2 is **tiled** and we do multiple rounds

→ Synchronization + transfer overhead

→ Difficult to optimize (best tiling for each kernel?)

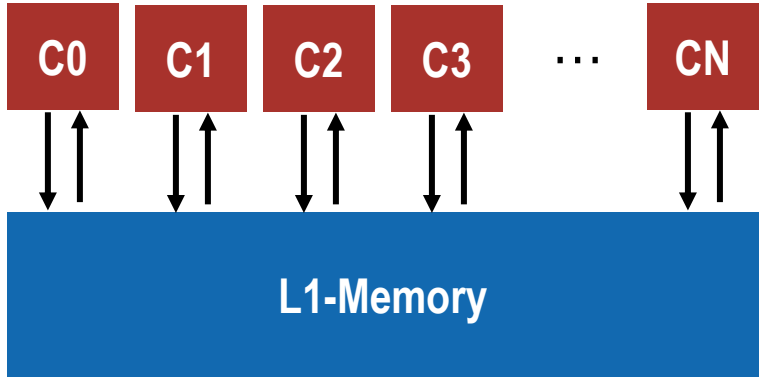


BIG workload → **MANY** cores & **BIG** memory

Example:

FFT 64/128 antennas, frequency range 275-RB → **1.6/0.8 MiB**

Beamforming, 64/128 antennas, 32 beams → **2.1/1.2 MiB**



MemPool[1] has «only»	1MiB L1/256 cores
ET-SoC1[2] has «only»	1KiB L1/core, 4MiB L2/Cluster
Kalray MPPA[3] has «only»	8KiB L1/core, 2MiB L2/Cluster

It is not enough to **double-buffer** the computation, **requires tiling**

[1] Riedel et al., 2023 IEEE ToC, https://ieeexplore.ieee.org/document/10227739
[2] Ditzel et al., 2022 IEEE Micro, https://ieeexplore.ieee.org/abstract/document/9670721
[3] Dinechin et al., 2013 HPEC, https://ieeexplore.ieee.org/abstract/document/6670342

TeraPool-SDR: 4MiB L1 for 1024 cores!

- **We propose TeraPool-SDR 4MiB multi-banked L1 for 1024 Snitch cores**
  - Reduces data movement and splitting (no tiling)
  - Allows to double buffer computation, hiding copy latency
  - **Is it still physically feasible?**
- **ManyCore will achieve high performance on 5G kernels**
  - **Can we keep high energy-efficiency?**

1. **Physical aware** architecture
2. Physical implementation
3. Performance on SDR

Computing Tile in PULP

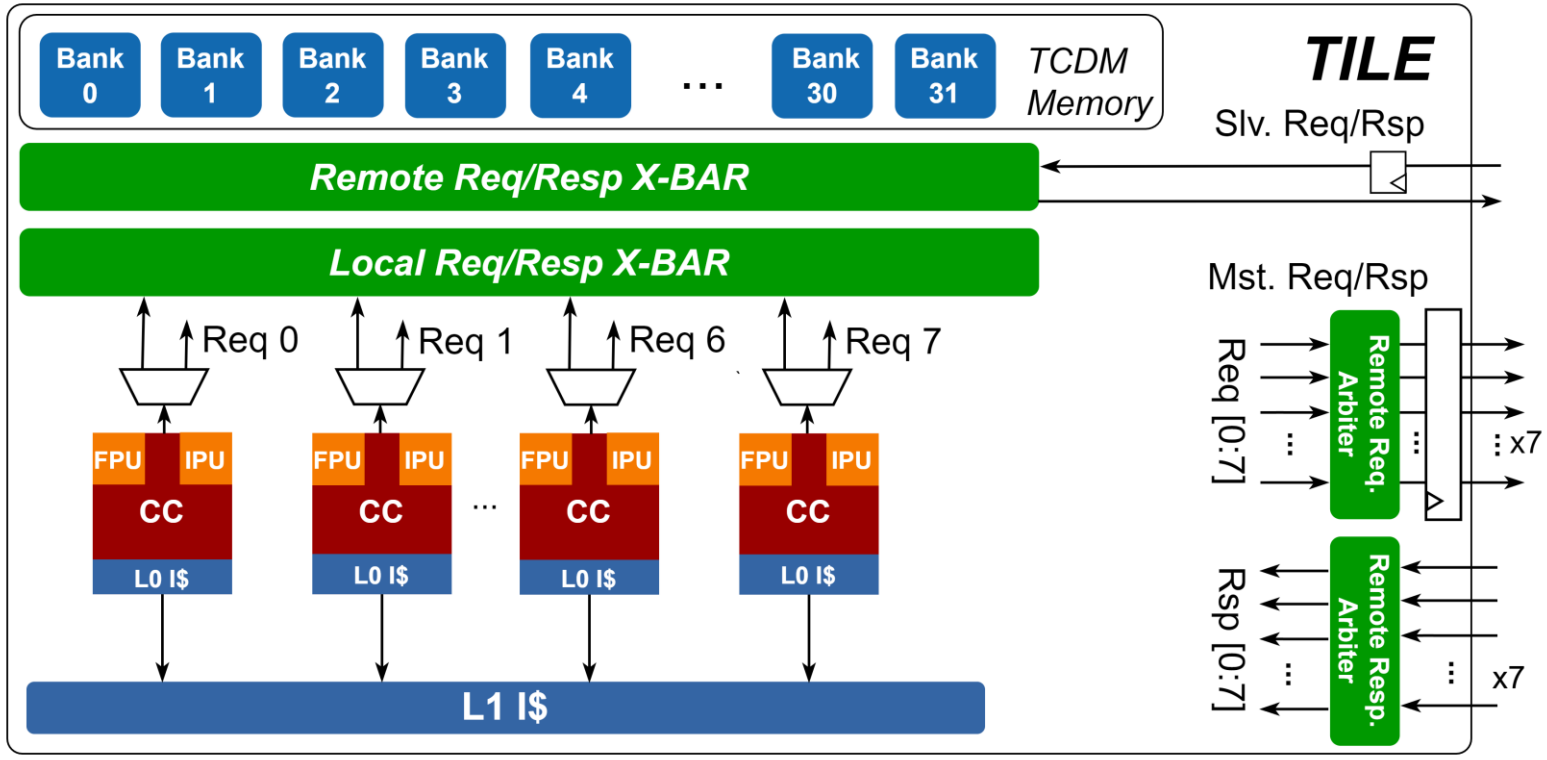
Tile = **building block**

- **Scalable**
  - 8x RV32IMAXpulpimg cores
  - Tightly Coupled Data Memory
  - Shared I$

TCDM in other Tiles is **accessible** via a Request/Response

Fully combinational interconnect:

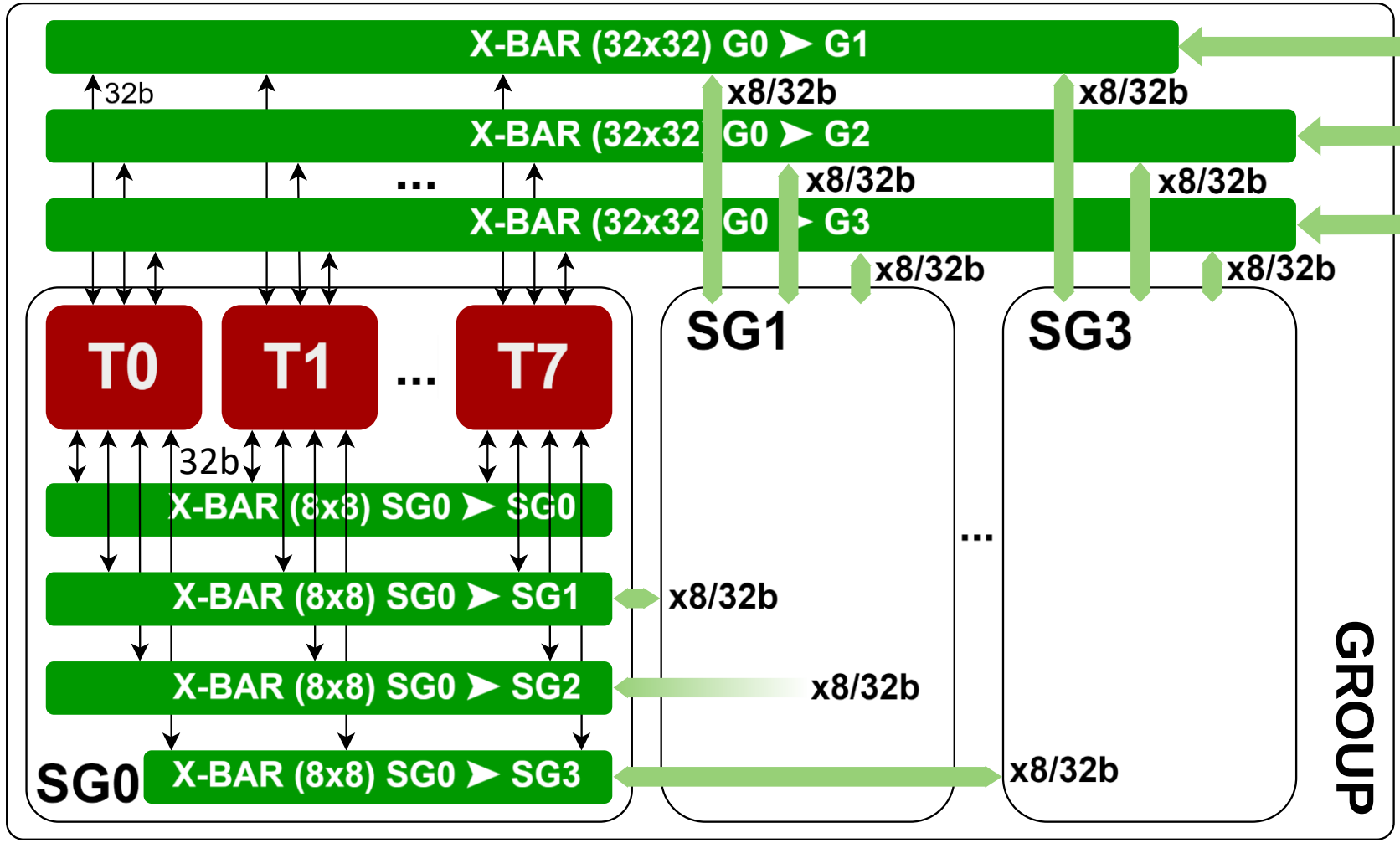
1 cycle access latency in the Tile



TeraPool-SDR hierarchical interconnect

Hierarchical Design:

- **8x cores/**Tile
  1 cycle
- **8xTiles/**SubGroup
  3 cycles
- **4xSubGroups/**Group
  5 cycles
- **4xGroup/**Cluster
  7 cycles



Interconnect Placement

In the top hierarchy, we assemble 4 **Groups**:

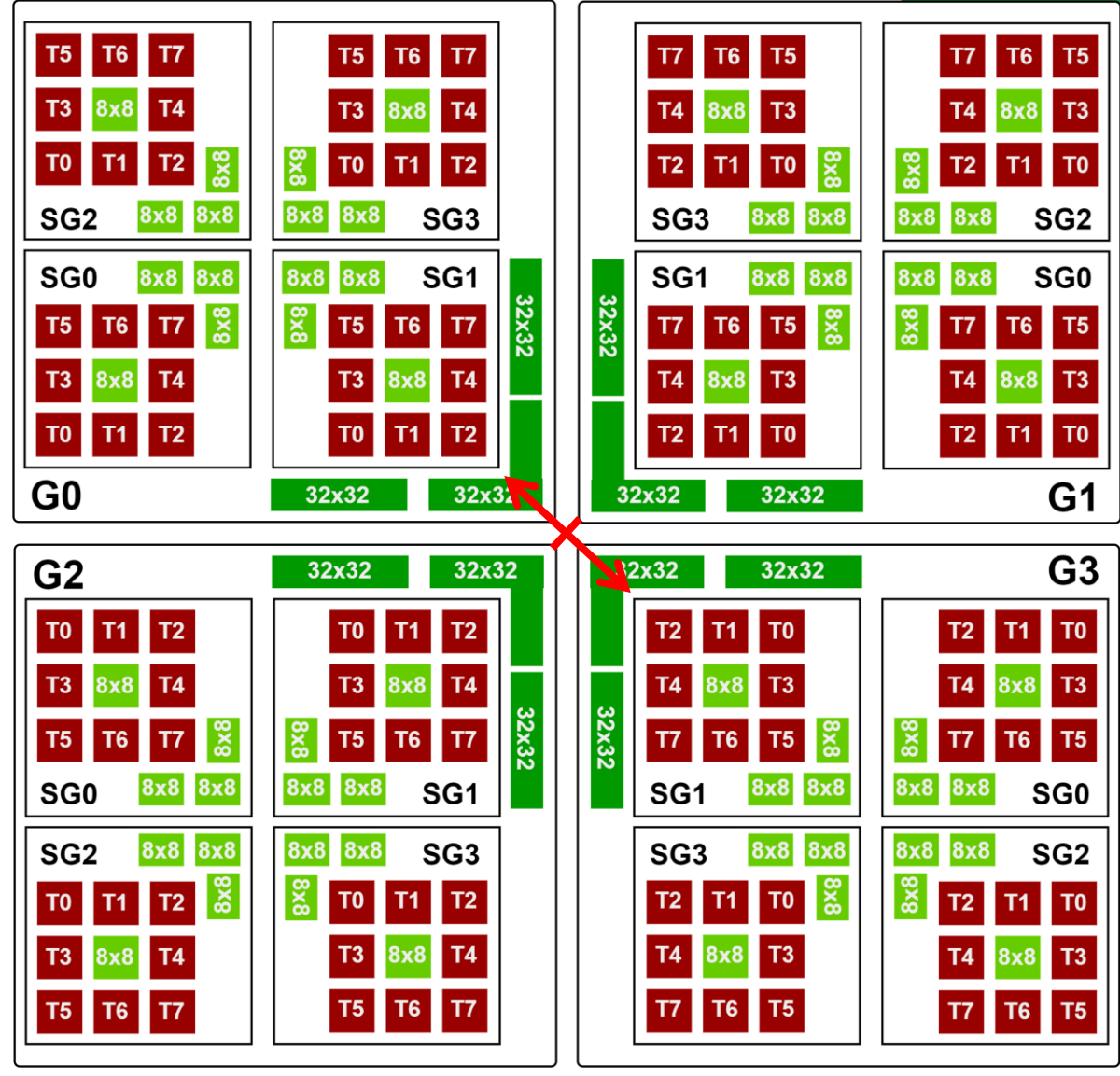
- We keep central symmetry
- We route point-to-point connections between the 32x32 interconnects of each Group

Flexibility on where to put spill registers (do we target low access latency or high frequency) 7, 9, 11 cycles

Largest design so far in PULP:

1024 cores RV32IMA-Xpulpimg

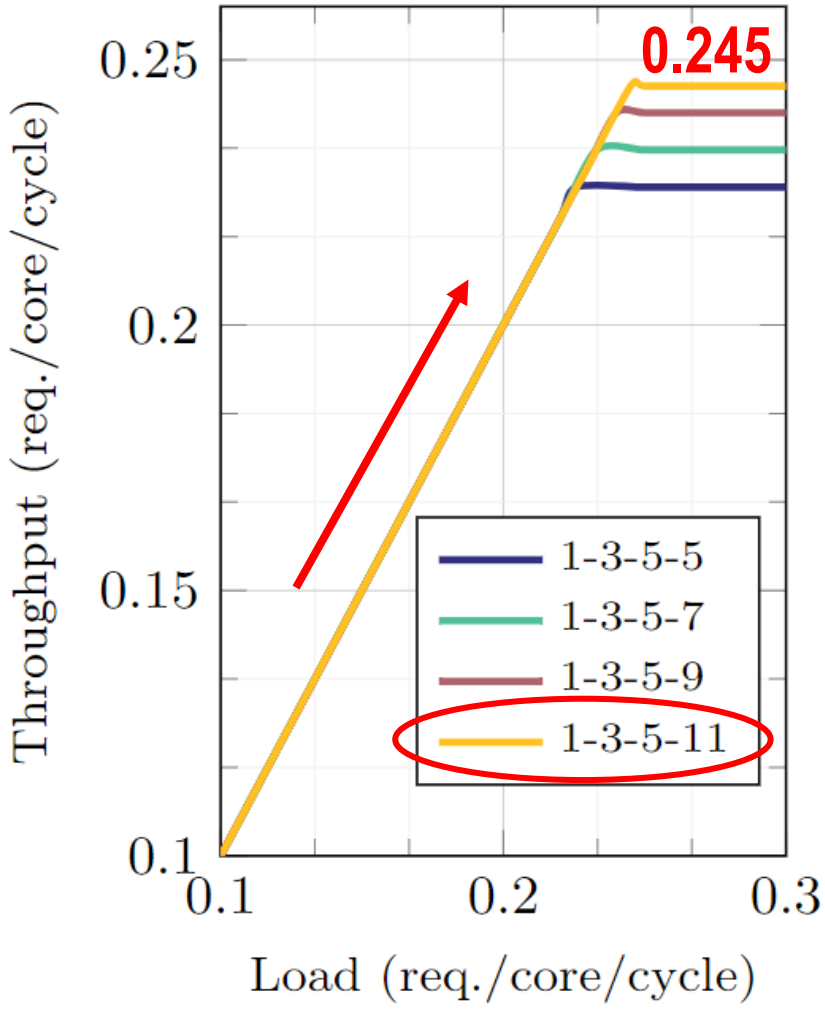
4096 8b-MAC/cycle



Throughput analysis

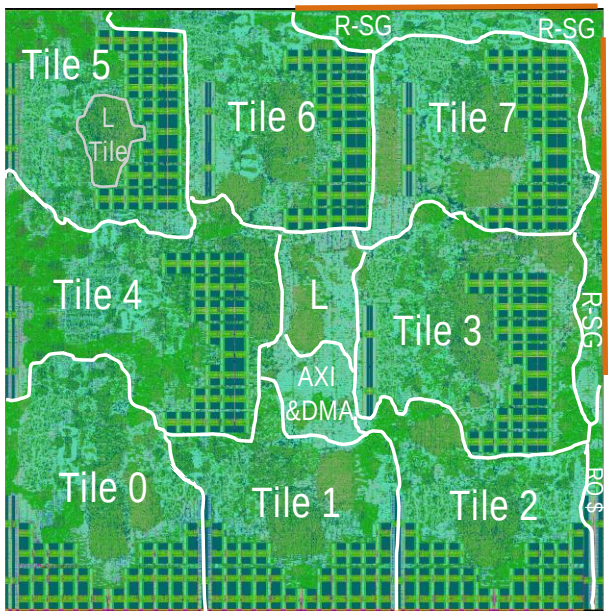
- **Each core injects random requests**
  - Poisson process, rate = req./cycle
- **Throughput increases linearly, until congestion in the interconnect**
- **Saturation point depends on number of pipeline stages**

→ The deeper the interconnect, the more requests it can handle

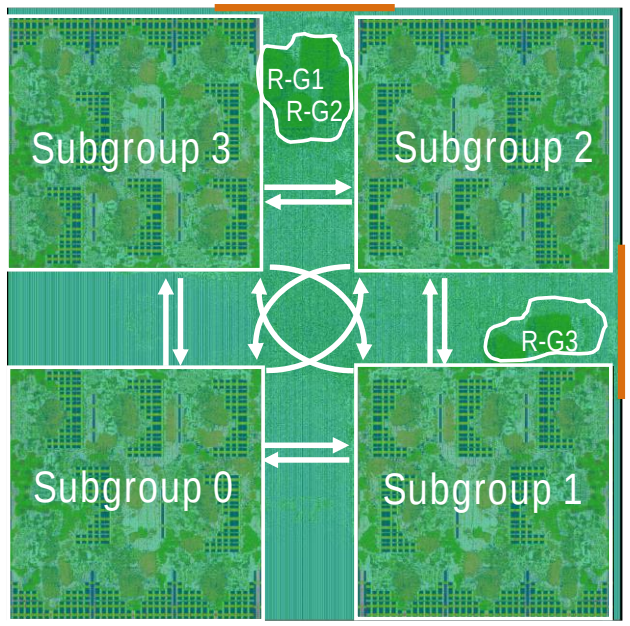


Physically Feasible Design

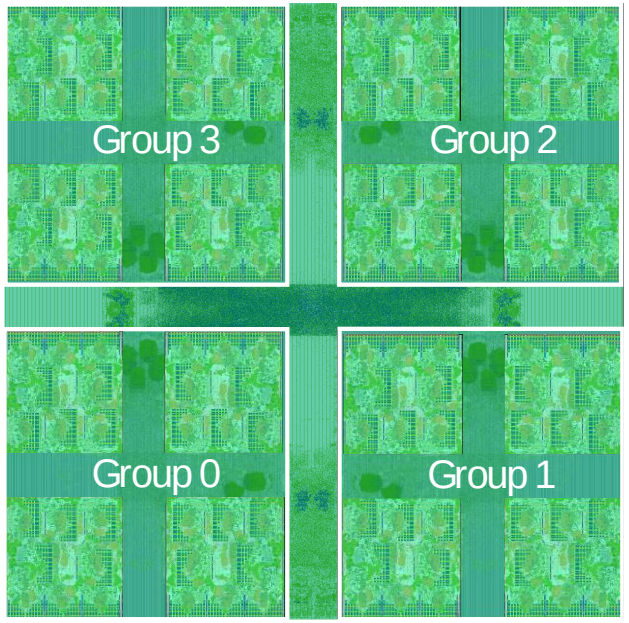
Subgroup



Group



Cluster



Methodology:

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

Area:

- Subgroup: 1.52 x 1.52 mm^2 (58% utilization)
- Group: 3.8 x 3.8 mm^2
- Cluster: 8.3 x 8.3mm^2

Performance:

- 740MHz in WC, 920MHz in TT

Benchmarks from 5G SDR

- **FFT**

  → **Radix-4 Cooley-Turkey** decimation in frequency

- **Beaforming**

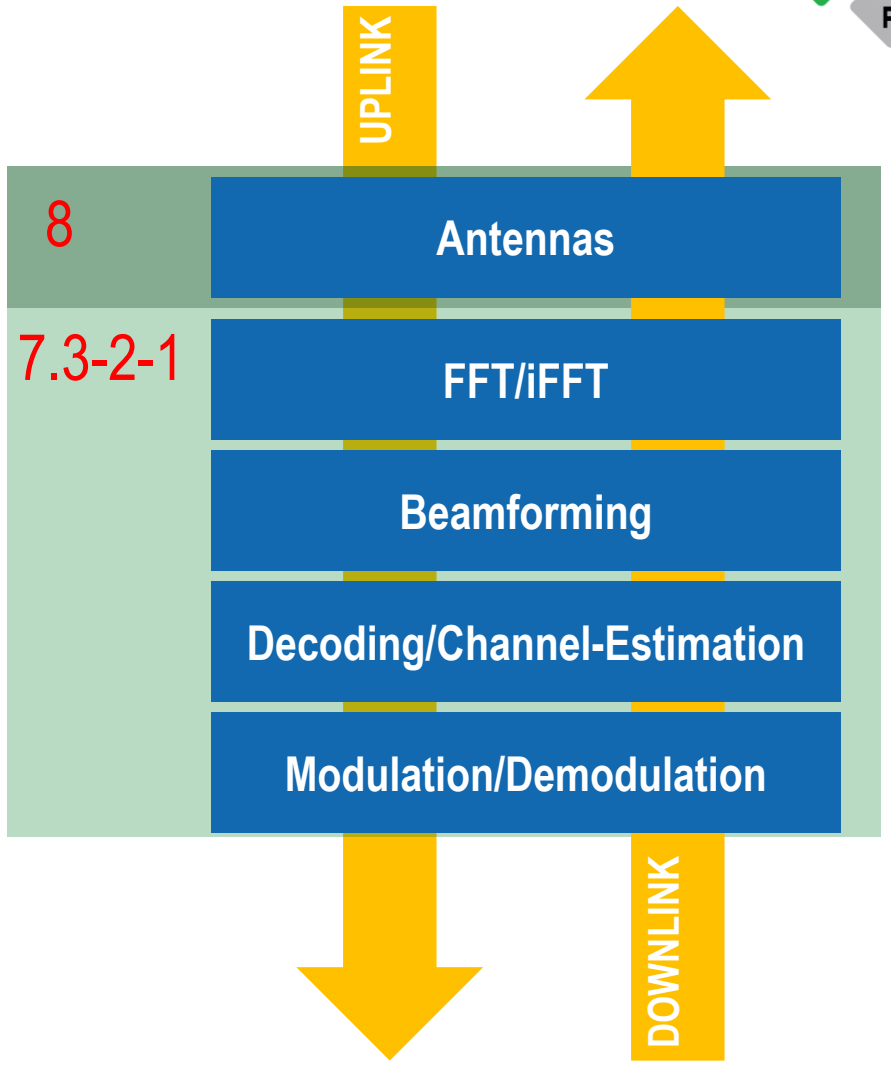
  → **Matrix-matrix multiplication** with known coefficients

- **Channel estimation**

  → Reconstucts the transmission channel via signals known at the transmitter and at the receiver (**element by element division**).

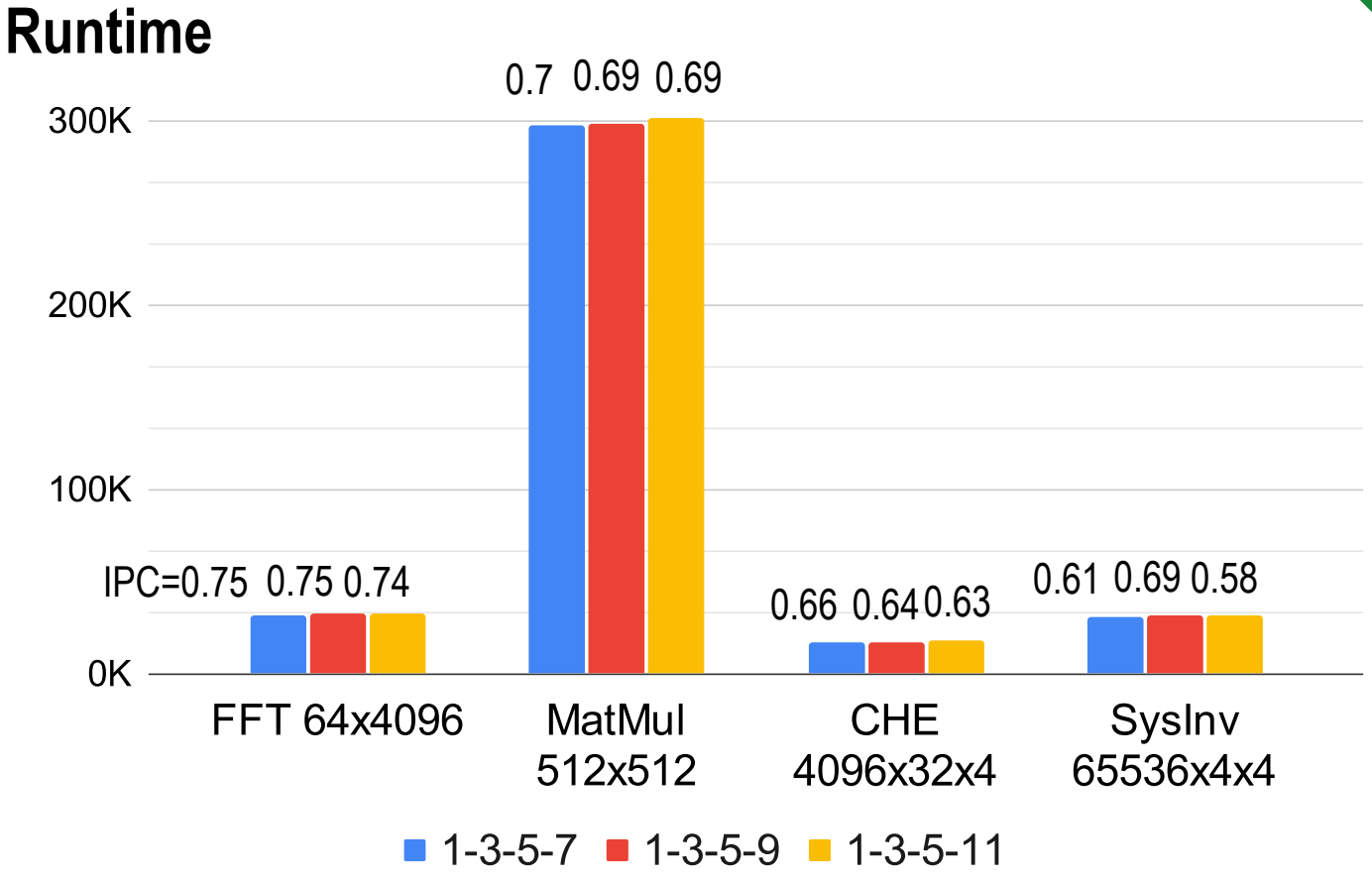
- **Decoding**

  → To reconstruct the input signal MMSE estimator, based on **matrix inversion**.



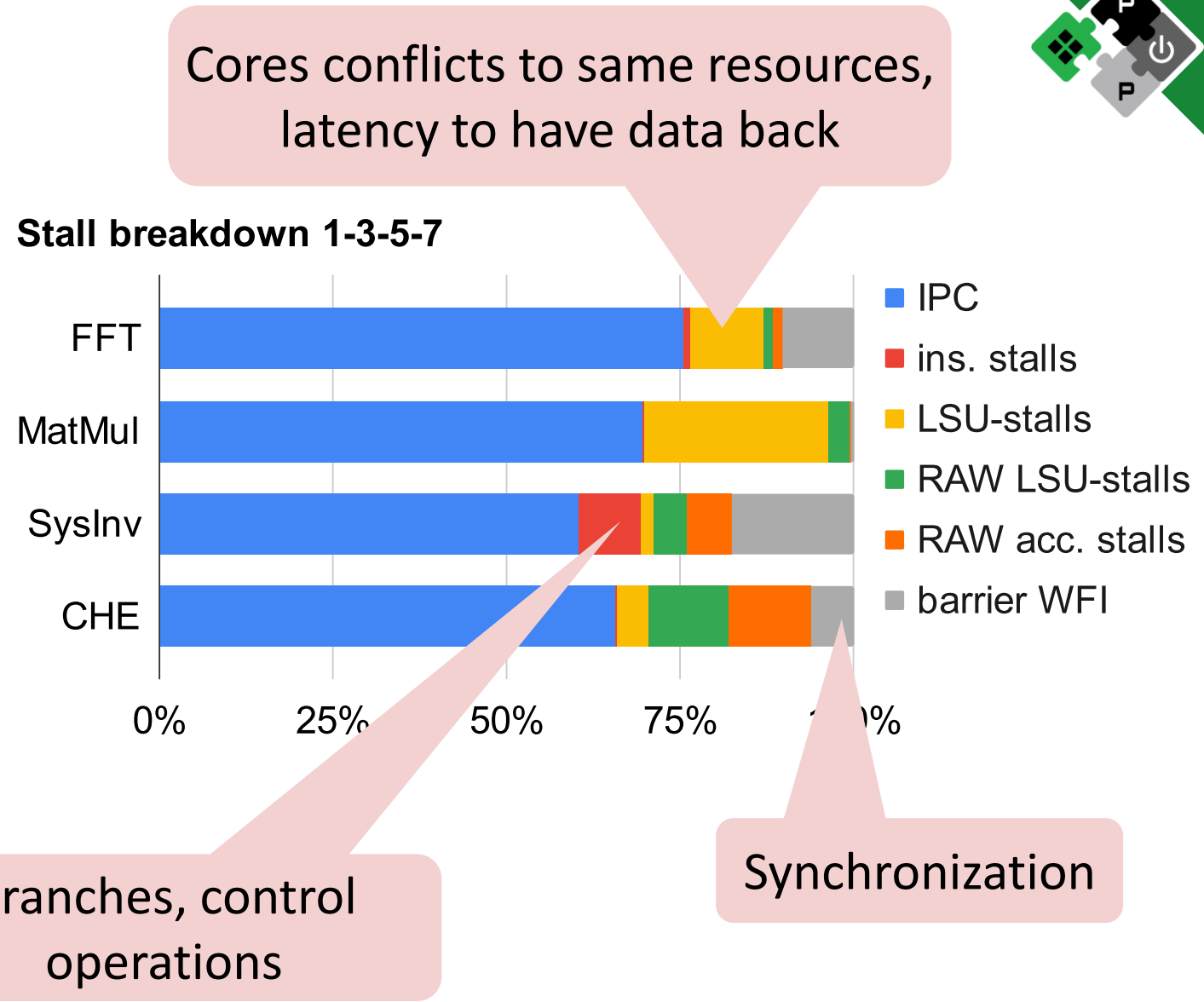
High IPC on SDR-kernels

- Up to **0.75 IPC**, no significant variations depending on configuration
- Minimize LSU stalls for FFT, CHE, and SysInv by computing in **local memory**
- MatMul (many out-of-Tile accesses) achieves 0.7 IPC



High IPC on SDR-kernels

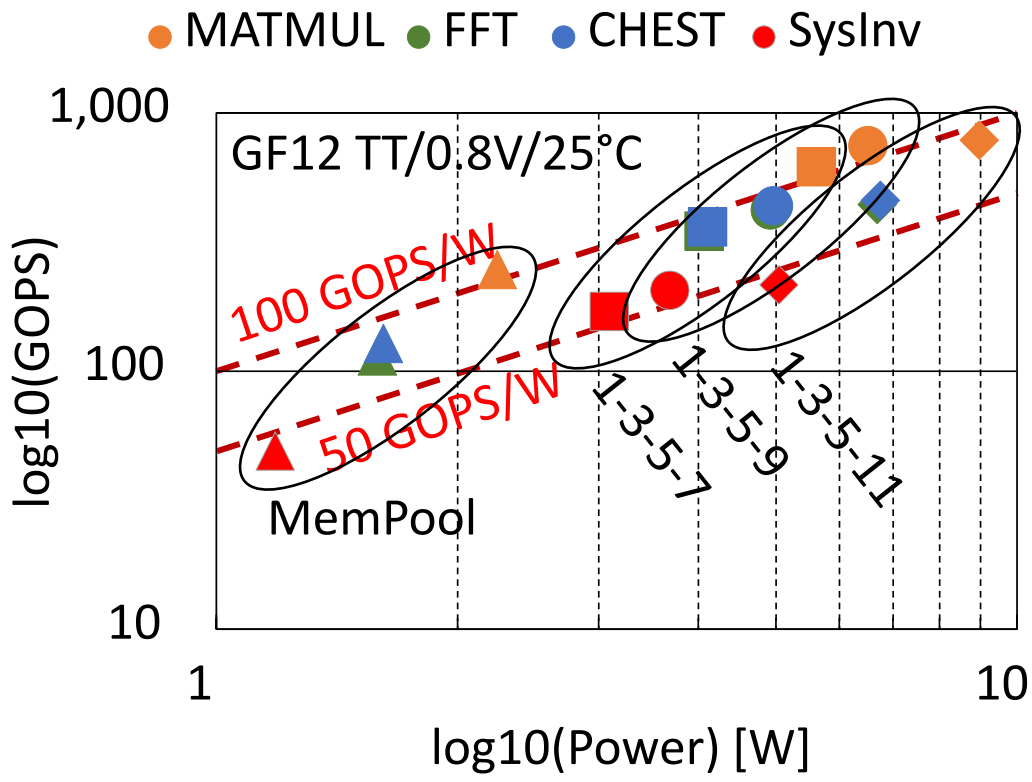
- Up to **0.75 IPC**, no significant variations depending on configuration
- Minimize LSU stalls for FFT, CHE, and SysInv by computing in **local memory**
- MatMul (many out-of-Tile accesses) achieves 0.7 IPC



Scaling does not reduce energy efficiency!

MemPool is a ManyCore cluster with **256 cores and 1MiB** shared scratchpad

- similar interconnection hierarchy (Tile, Group, Cluster)
- Performance analysis with L1 saturated

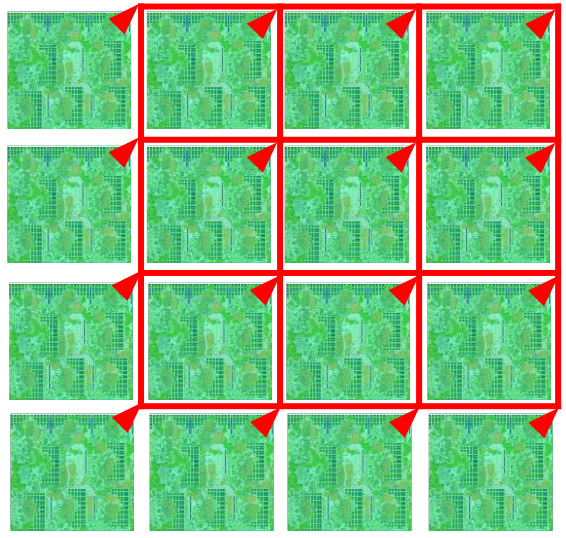


- **Performance are up to 3.2X**
  - Top performance configuration 1-3-5-11
- **E.Efficiency (GOPS/W) are equal or higher than MemPool**
  - Top energy-efficiency configuration 1-3-5-9

In the roadmap…

- Implementation of a full **6G application** on the platform
- Pack **more functions** in the SubGroup hierarchy (64cores) e.g. floating point support
- Explore other interconnects e.g. **NoC routers** at the SubGroup Boundary

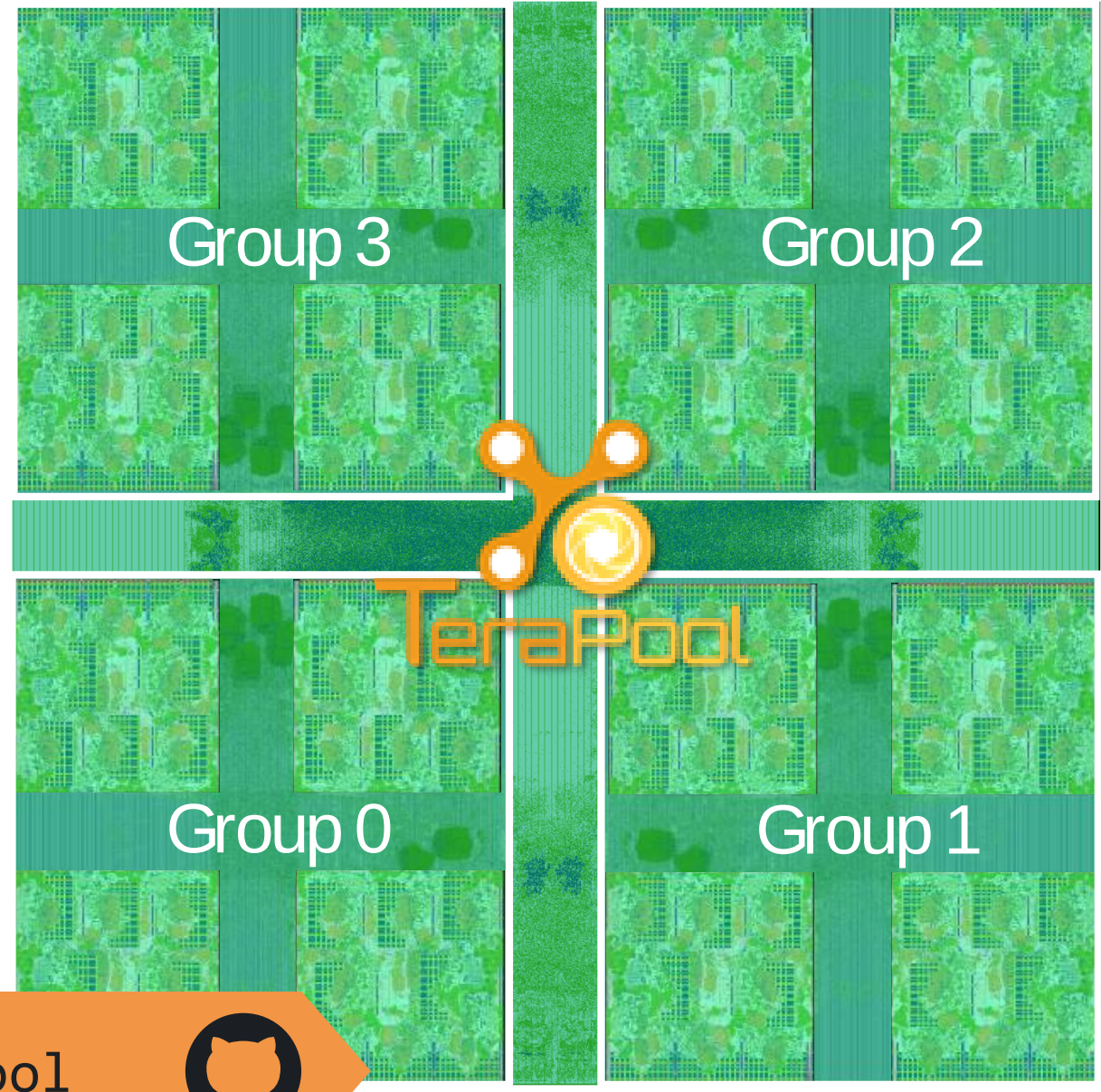
→ Might increase area utilization, **still tolerable latency?**



TeraPool SDR is a physically feasible shared-memory 1024-cores cluster

- In GlobalFoundries' 12P+ FinFET, **11 cycles** latency, **740MHz@WC** **920MHz@TC**, in 8.3 x 8.3mm^2
- **1.89TOPS** peak-performance, **3.2X** on SDR kernels vs MemPool (256-cores)

Interconnect-scaling does not impact energy efficiency **50-100GOPS/W**



`github.com/pulp-platform/mempool`

PULP Platform

Open Source Hardware, the way it should be!

Yichao Zhang yiczhang@iis.ee.ethz.ch
Marco Bertuletti mbertuletti@iis.ee.ethz.ch
Samuel Riedel sriedel@iis.ee.ethz.ch
Alessandro Vanelli-Coralli avanelli@iis.ee.ethz.ch
Luca Benini lbenini@iis.ee.ethz.ch

Institut für Integrierte Systeme – ETH Zürich
Gloriastrasse 35
Zürich, Switzerland

DEI – Universitá di Bologna
Viale del Risorgimento 2
Bologna, Italy

@pulp_platform

pulp-platform.org

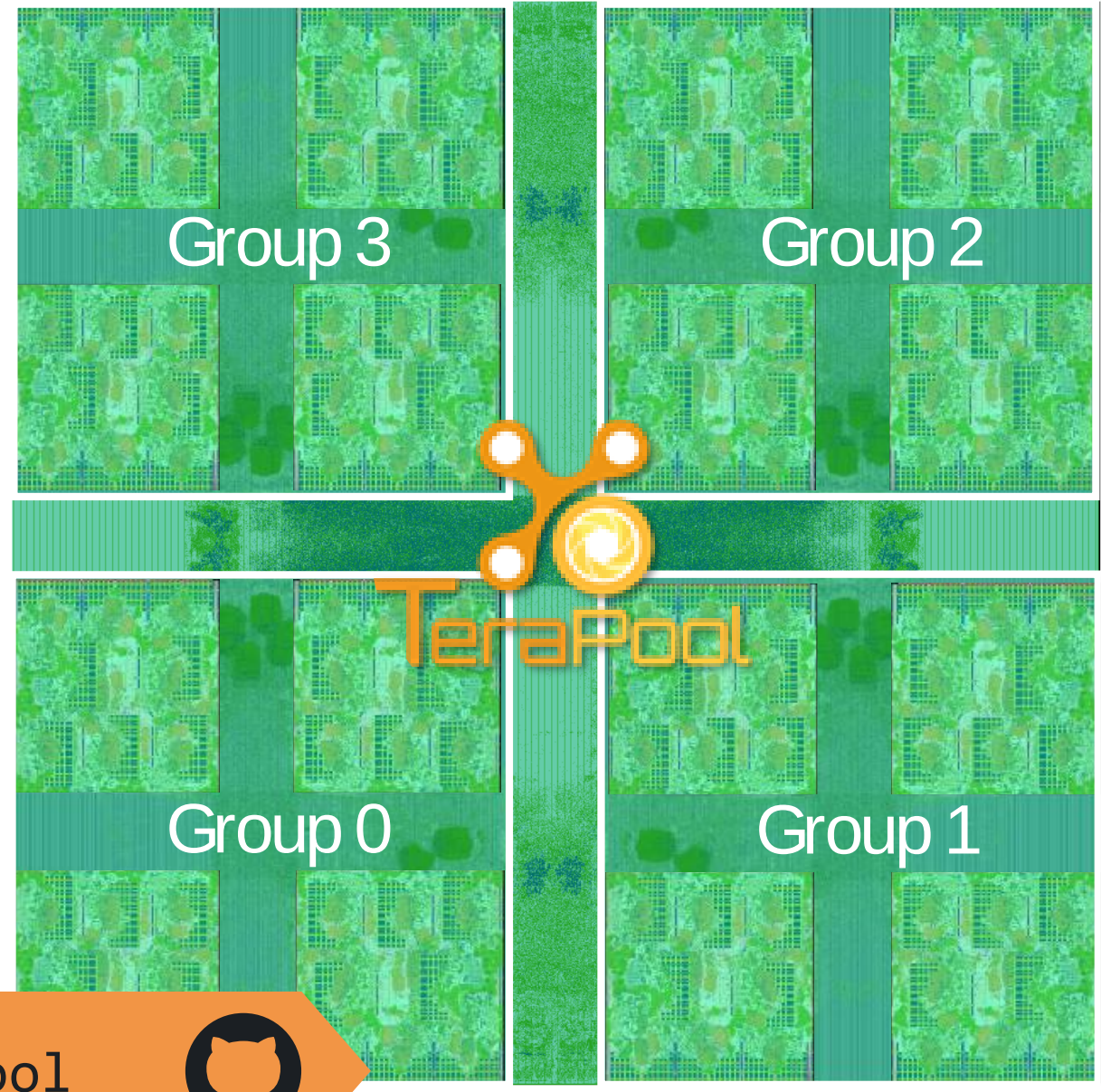
youtube.com/pulp_platform

ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

TeraPool SDR is a physically feasible shared-memory 1024-cores cluster

- In GlobalFoundries' 12P+ FinFET, **11 cycles** latency, **740MHz@WC** **920MHz@TC**, in 8.3 x 8.3mm^2
- **1.89TOPS** peak-performance, **3.2X** on SDR kernels vs MemPool (256-cores)

Interconnect-scaling does not impact energy efficiency **50-100GOPS/W**



`github.com/pulp-platform/mempool`