ETH zürich

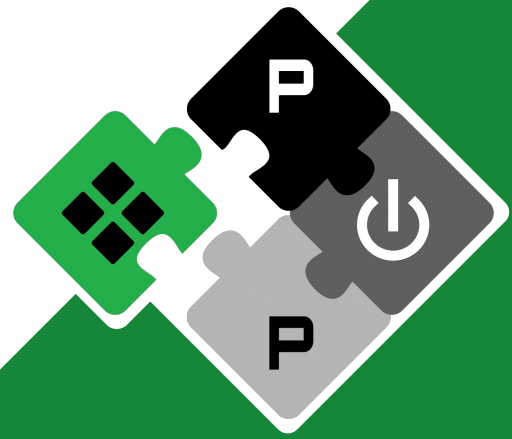
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

CachePool: Many-core cluster of Customizable, Lightweight Scalar-Vector PEs for for irregular L2 data-plane workloads

Integrated Systems Laboratory (ETH Zürich)

Yichao Zhang, Chi Zhang, Zexin Fu, Marco Bertuletti
yiczhang, chizhang, zexifu, mbertuletti@iis.ee.ethz.ch

Alessandro Vanelli-Coralli avanelli@iis.ee.ethz.ch
Luca Benini lbenini@iis.ee.ethz.ch

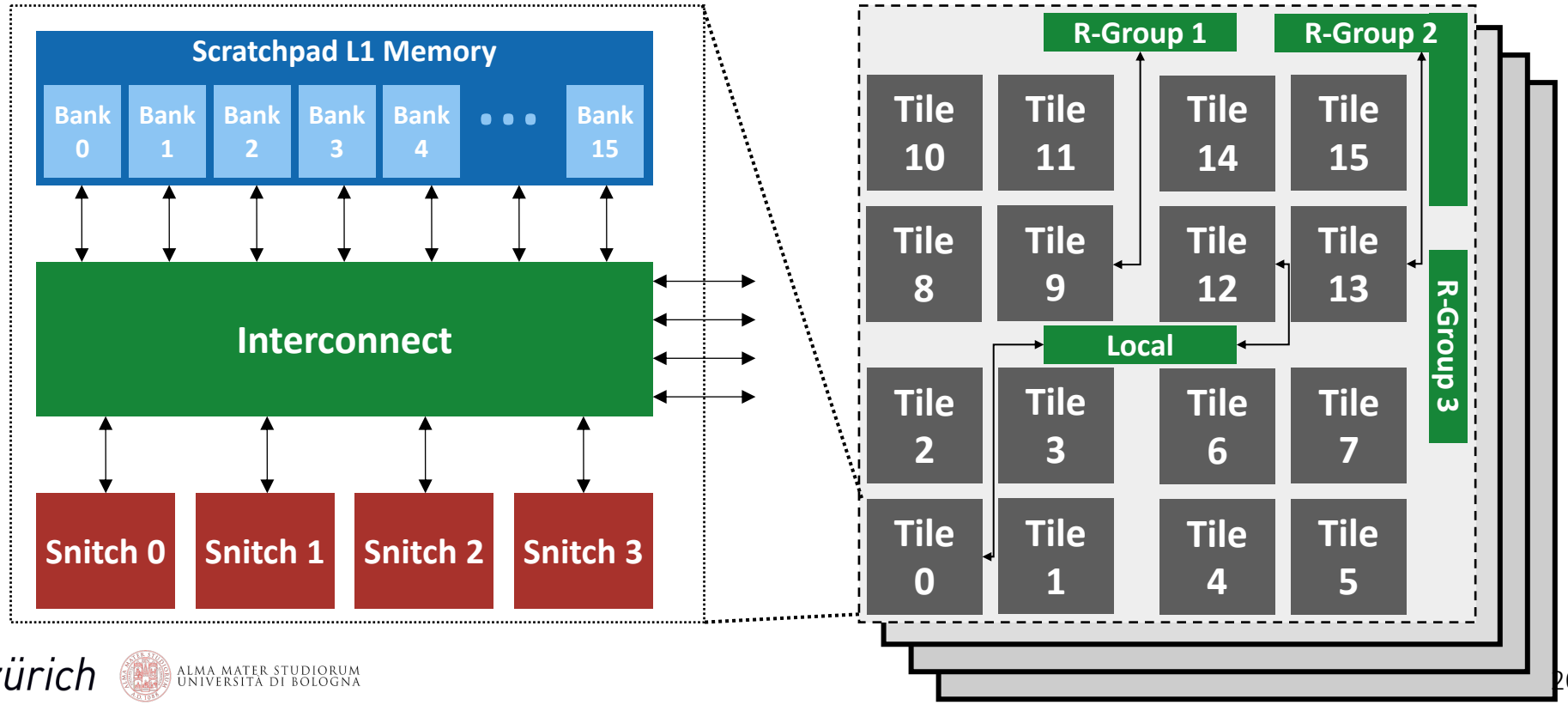


PULP Platform
Open Source Hardware, the way it should be!

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

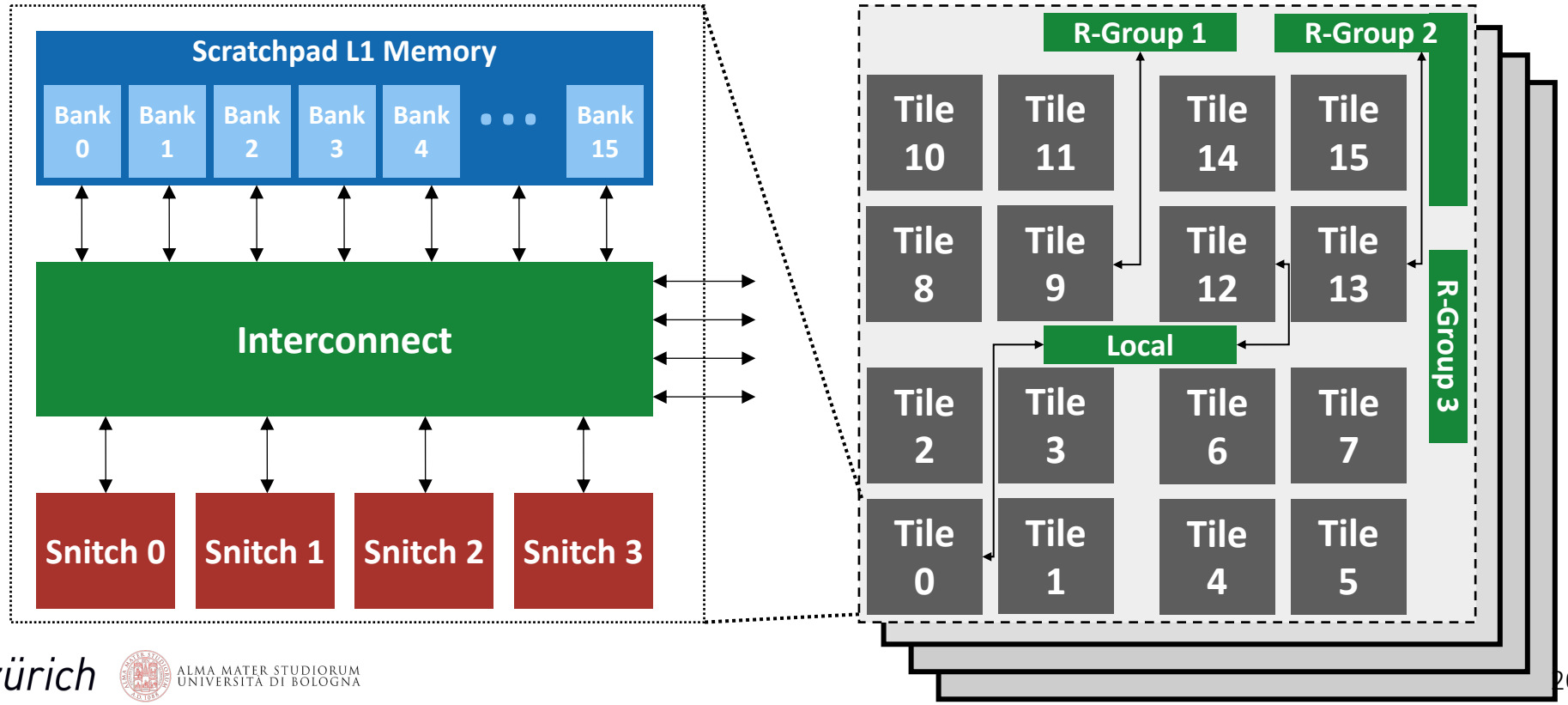
Our Baseline: MemPool

- **MemPool Family: scalable many-core shared L1-TCDM cluster**
  - Physically Feasible, scale-up to 1024 **extendable tiny RISC-V** cores (TeraPool-SDR)
  - Energy-efficient for B5G workload (**average 6W for 5G-PUSCH**)
  - Explicit DMA-based data transfer from L2+: tiling & double buffering for latency hiding



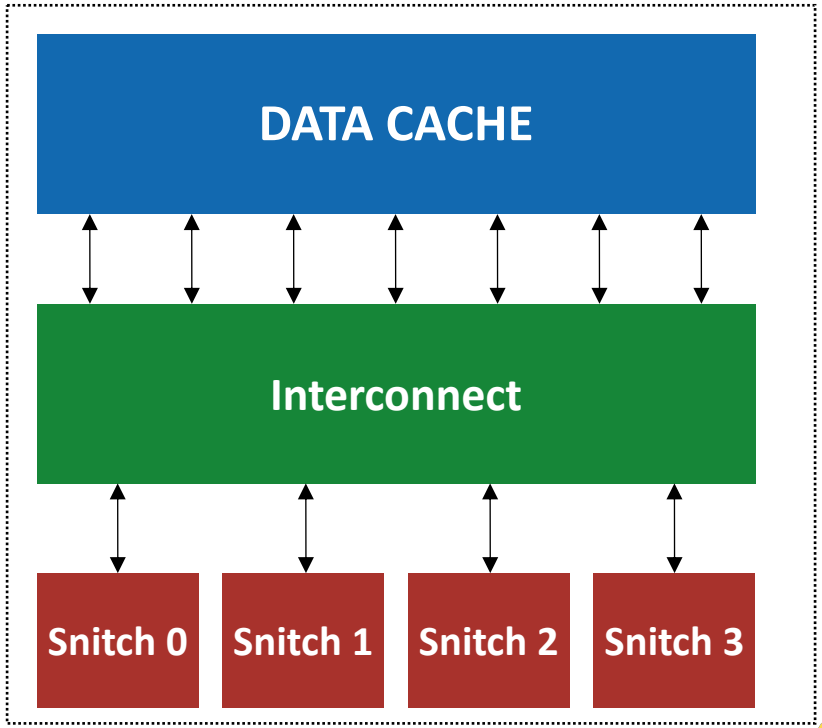
Our Baseline: MemPool

- **However, for large, irregular (sparse) workload in large-space main memory:**
  - DMA is a not efficient design choice for small, data dependent transfers: hard to tile and double-buffer, requires massive software rework



From MemPool to CachePool

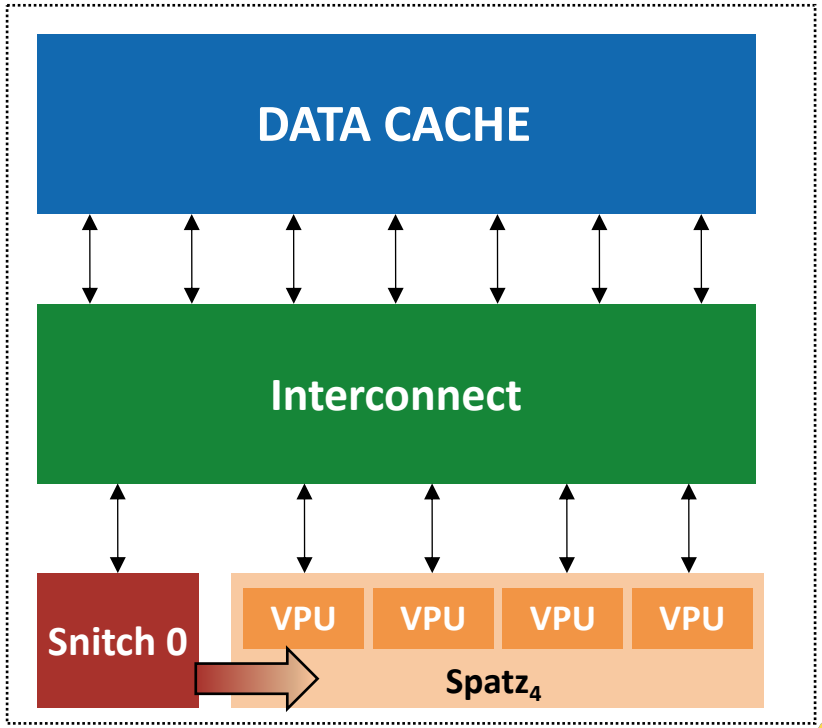
- **However, for large, irregular (sparse) workload in large-space main memory:**
  - DMA is a not efficient design choice for small, data dependent transfers: hard to tile and double-buffer, requires massive software rework
  - **Cache based** → hardware gathering data near to PEs.



From MemPool to CachePool

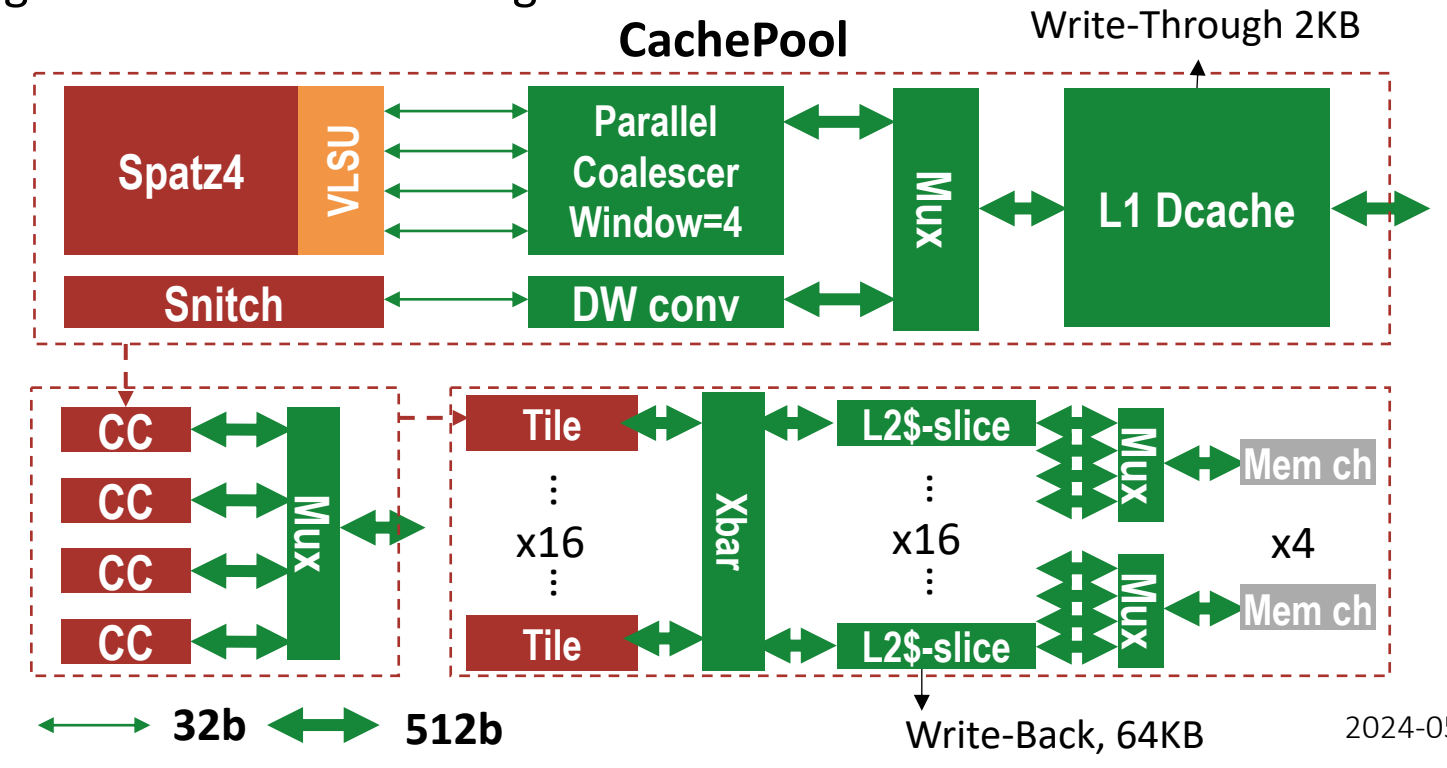
- **However, for large, irregular (sparse) workload in large-space main memory:**
  - DMA is a not efficient design choice for small, data dependent transfers: hard to tile and double-buffer, requires massive software rework
  - **Cache based** → hardware gathering data near to PEs.
  - A compact **Vector** processing unit (Spatz):
    - Enabling **SIMD processing**, energy-efficient;
    - **Latency Hiding**:

    From large L1 (MemPool) → **large VRF** (MemPool-Spatz)



CachePool: Cache-based, Manycore, Vector-Scalar PEs

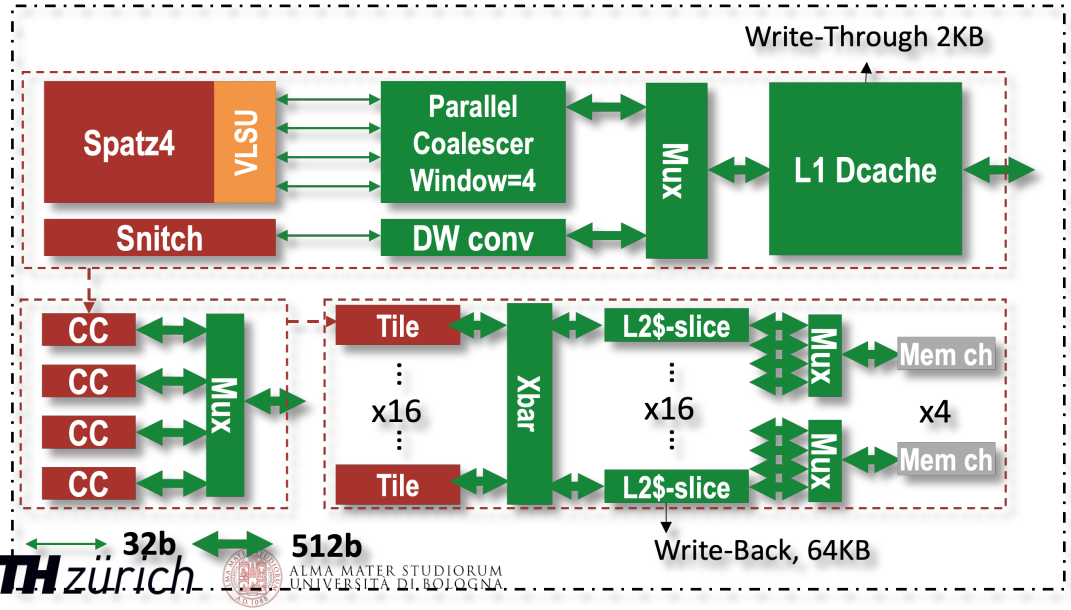
- **Cluster Architecture from PE upwards:**
  - Snitch (flow control) → Spatz (SIMD processing)
  - Flexible (configured Nr. Scalar cores and Vector units) and Scalable.
  - 64 Core-Complex (Snitch-Spatz4), totally 256 VPUs.
  - Parallel Coalescer design for efficient accessing cache line.



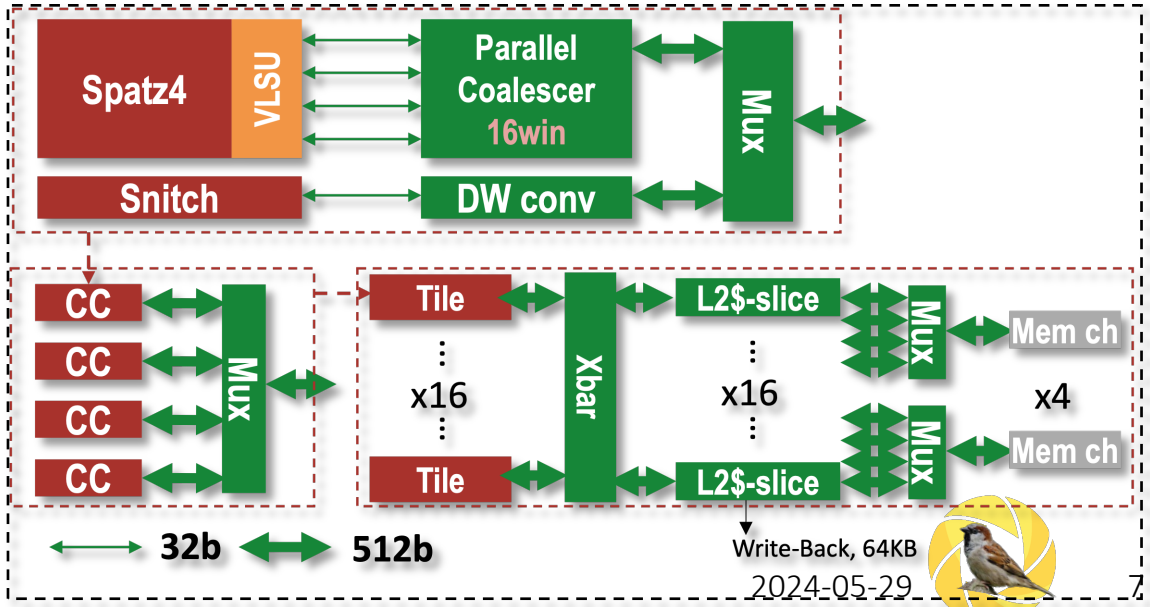
CachePool: Cache-based, Manycore, Vector-Scalar PEs

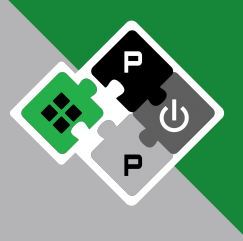
- **Architecture Research:**
  - L1 D-cache, or L1-less, coalescer exploration and design
  - Interconnection design for large-L2 cache slices.
  - How to reduce large, non-blocking cache MSHR overhead.
  - How to handle cache coherence.
  - Exploring Physical Implementation: feasibility, optimization

L1 D-Cache



L1 D-Cache Less





Thank you!

ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA