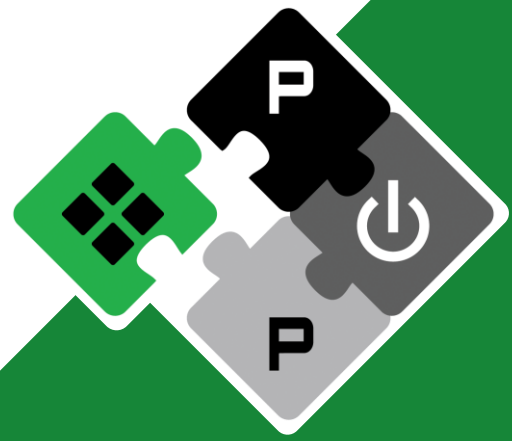
ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

CachePool: Many-core cluster of customizable, lightweight scalar-vector PEs for irregular L2 data-plane workloads

Integrated Systems Laboratory (ETH Zürich)

Zexin Fu, Diyou Shen zexifu, dishen@iis.ee.ethz.ch

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



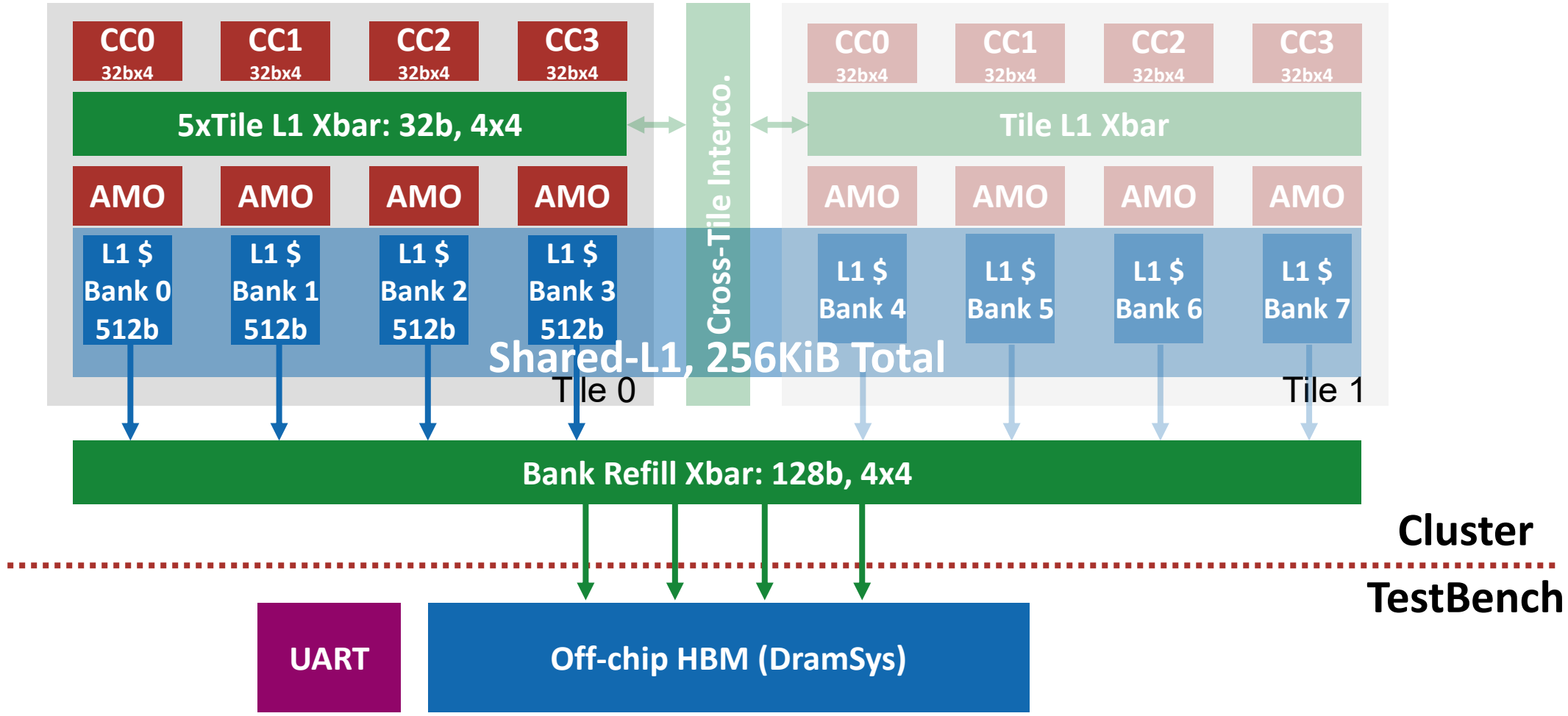
@pulp_platform

pulp-platform.org

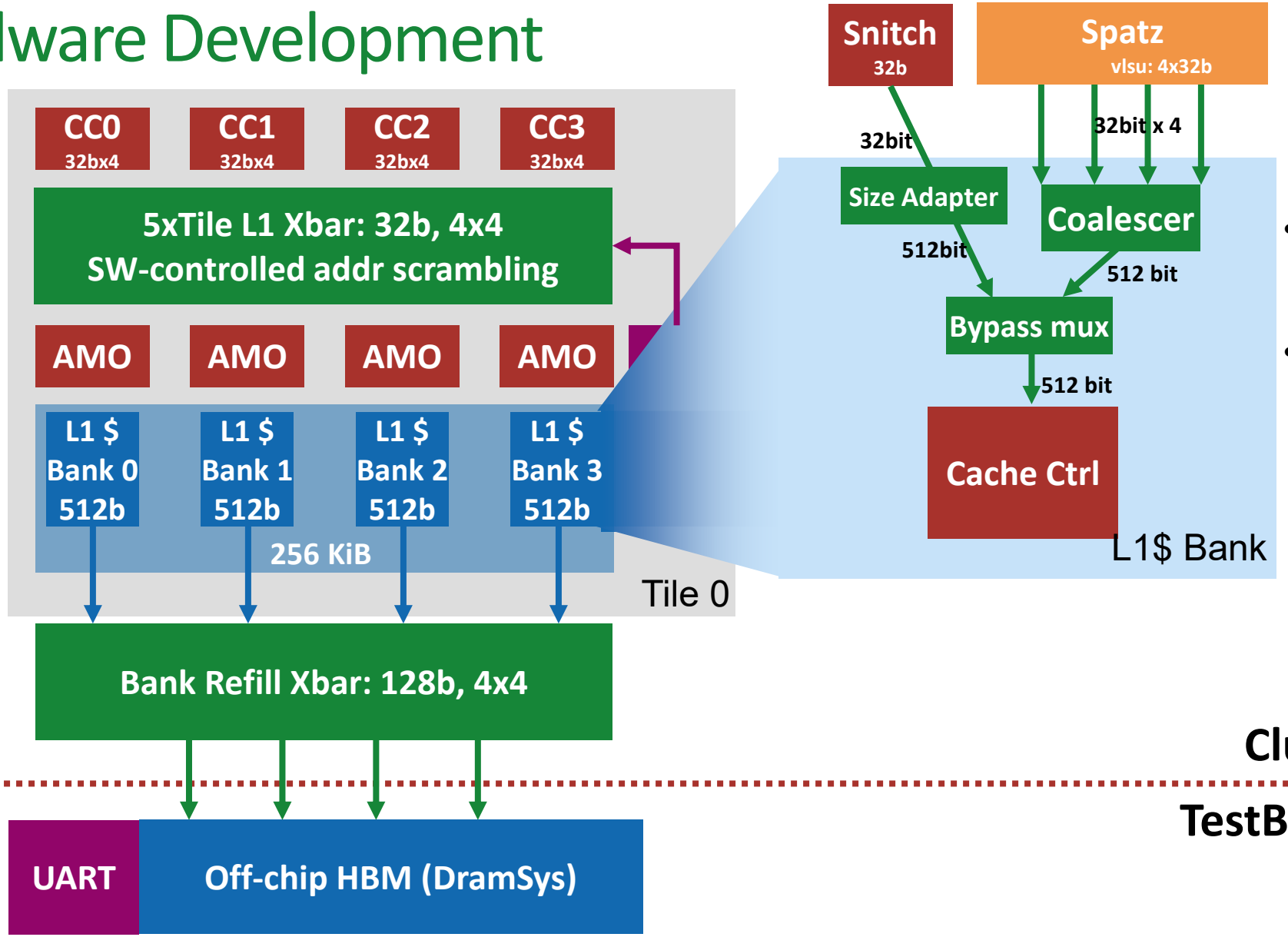
youtube.com/pulp_platform

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

Hardware Development



Hardware Development



- Let snitch req bypass coalescer
- Cut 4 cycle latency for snitch memory access

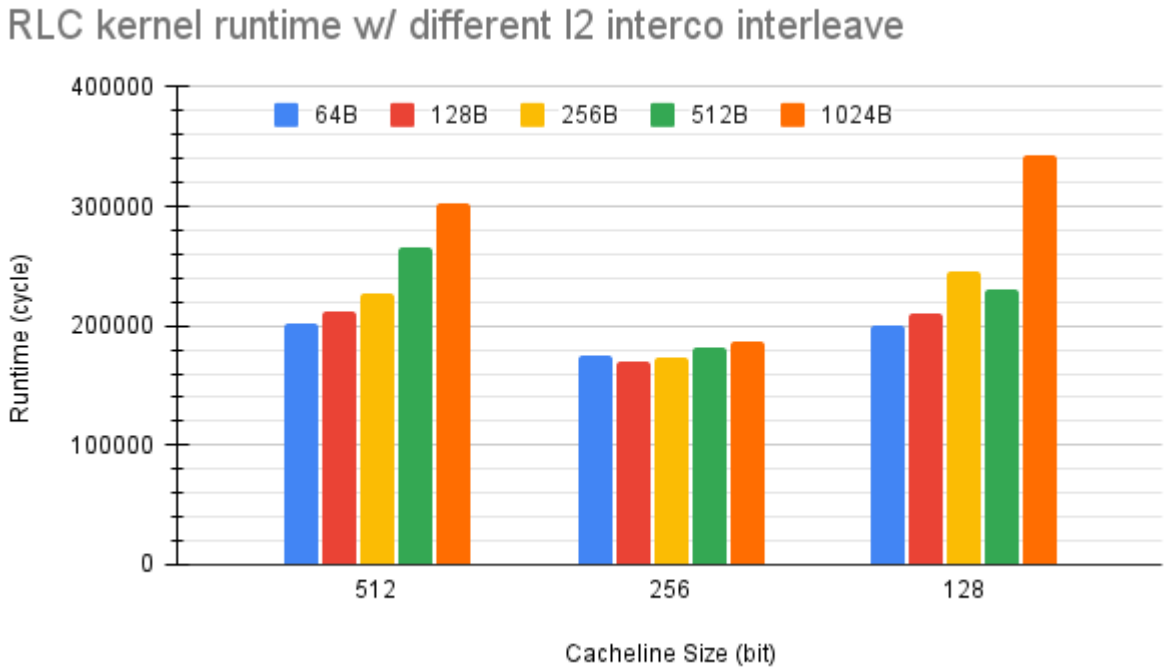
Software Analysis: Performance under Different Cacheline Width

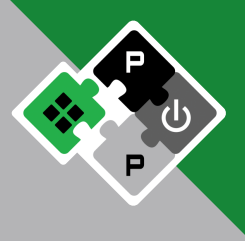
- Workload: 300 RLC package workload
- L2 interleave 1024 byte
- Strange: 256b has the best perf
- 512b & 256b has similar miss rate, but 512b has a much higher avg (per banks) miss stall cycle
  - 512-bit config has fewer cache lines with the same total cache capacity → more conflict misses and trigger cache eviction
  - The evictions can saturate the memory channel and block the following vector loads
- Plan
  - Stream data pollute the cache
  - Add non-cachable support for stream data (RLC pkg data)

Cache line size	Total$ access	$ Hit number	$ Miss Rate	Avg Miss Stall cycle	Total Run Cycle
512bit	61733	34641	**43.9%**	**47,848**	306,870
256bit	62137	34576	**44.4%**	**2,448**	187,255
128bit	81816	32866	59.8%	132,667	342,871

Software Analysis: L2 interconnect interleave impact

- **Performance impact of different L2 interconnect interleave factors**
  - The interleave granularity of 4 L1 $ banks access to 4 dram channels
  - Workload: 300 RLC package workload
  - 512b cacheline config is more sensitive to the L2 interconnect interleave factor
    - Not very clear the reason for now, may because more conflict miss and thus higher memory BW requirement for 512b





Thank you!