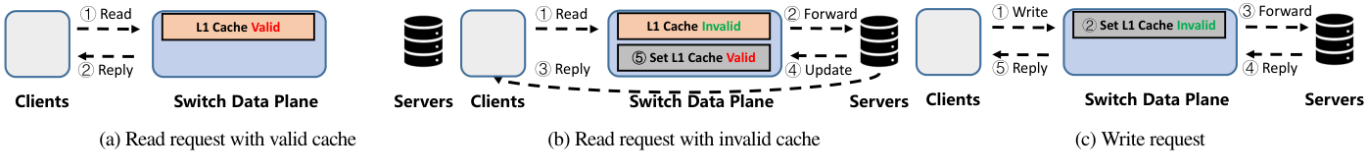
UniCache: Unified and Scalable Distributed Key-Value Storage with Multi-level Caching

Jiaqi Zheng, Chang Liu, Li Wang, Jialin Li, Guihai Chen

Frontiers of Computer Science, DOI: 10.1007/s11704-026-50475-8

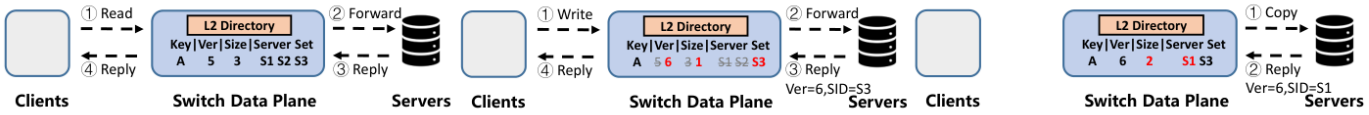
Problems & Ideas

- Problems: Skewed workloads create hotspots; pure sharding and server-only caches struggle to balance load and maintain consistency at scale.
- Ideas: UniCache unifies in-network (switch SRAM) and server-side caching with lightweight coherence directories to route reads/writes to the right replicas.



(a) Read request with valid cache (b) Read request with invalid cache (c) Write request

Fig. 3 Hot items stored in switch on-chip SRAM



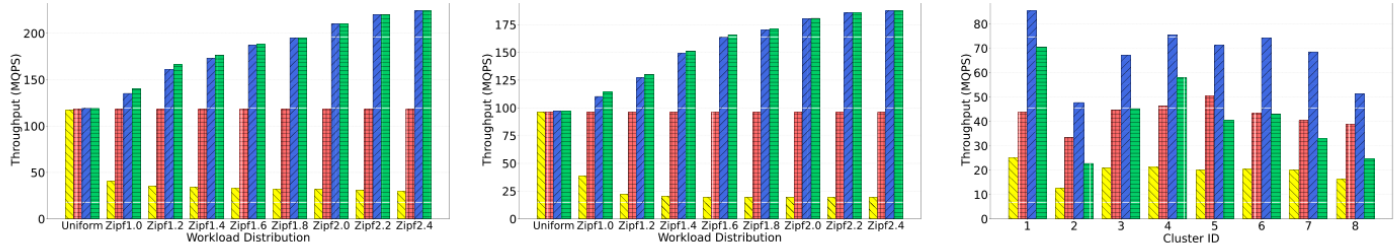
(a) Read request (b) Write reply with larger version number (c) Write reply with same version number

Fig. 4 Hot items stored in server memory with directories in switches

Hot items in switch SRAM (L1) and switch-managed directories for server-memory replicas (L2).

Main Experimental Results

- UniCache achieves higher throughput under skewed workloads, outperforming server-side caching baselines and traditional sharding.
- Performance remains stable on real-world traces, indicating effective hotspot mitigation and scalable coordination.



(a) 64B values (b) 128B values (c) Real-world

Fig. 5 Throughput under skewed workloads

Throughput under skewed workloads and real-world traces (Fig. 5).