Improving Performance of Flash Based Key-Value Stores Using Storage Class Memory as a Volatile Memory Extension

Hiwot Tadese Kassa *(Presenter), Jason Akers, Mrinmoy Ghosh, Zhichao Cao, Vaibhav Gogte, Ronald Dreslinski*

USENIX
ATC '21

JULY 14–16, 2021

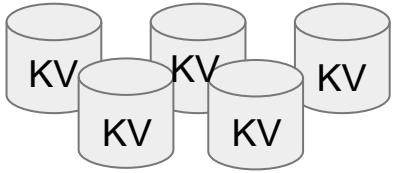
Data Center Memory Requirements

Facebook data is growing

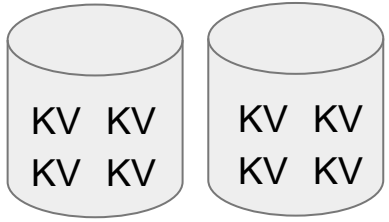
Videos, images … uploaded every second

More memory and power capacity

Multiple small servers

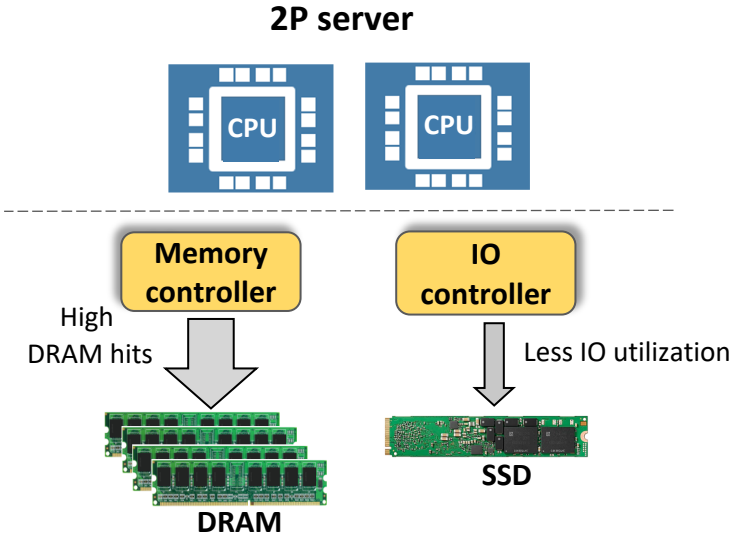


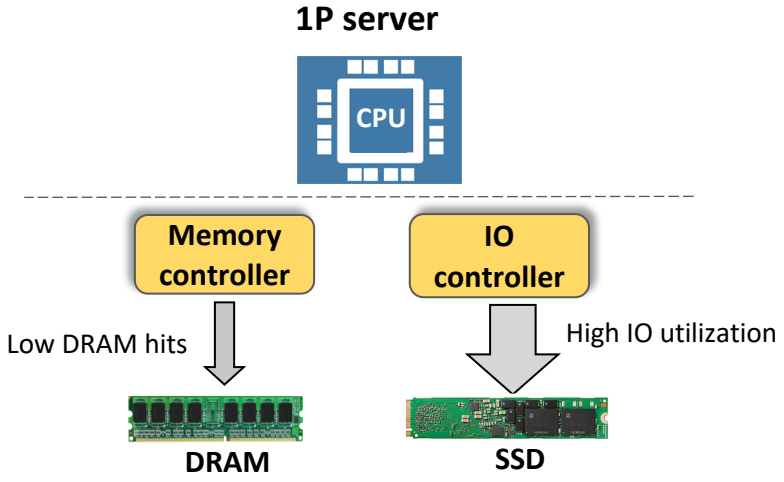
Fewer large servers



Demands fast, low latency access.

High-Performance Storage Servers at Facebook



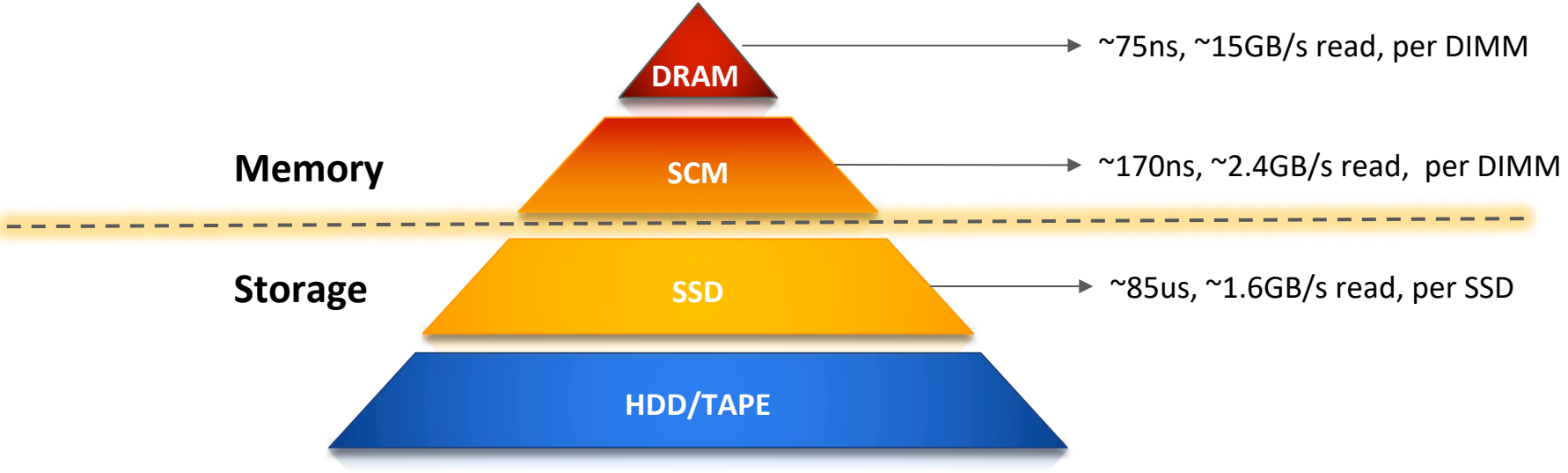


- 2 CPUs and 256 GB DRAM
- High performance but expensive

- 1 CPU and 64 GB DRAM
- Cost effective but lower performance per node

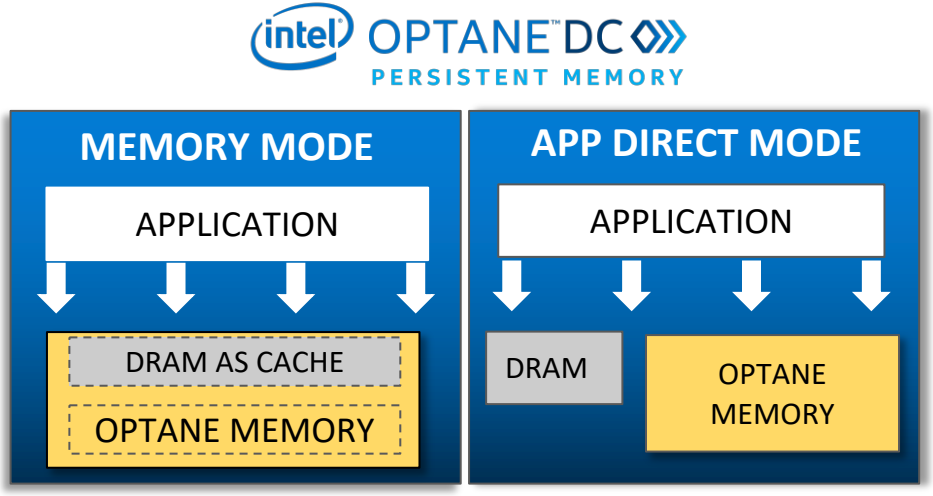
Source: https://engineering.fb.com/core- data/introducing-yosemite-the-first-open- source-modular-chassis-for-high-powered- microservers/.

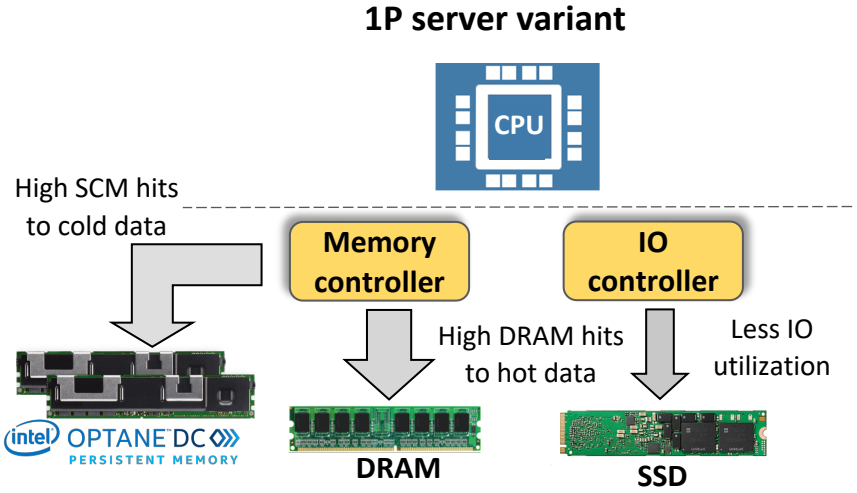
Storage Class Memory (SCM) Characteristics



- Slower than DRAM but are faster than SSD and give us very large memory density

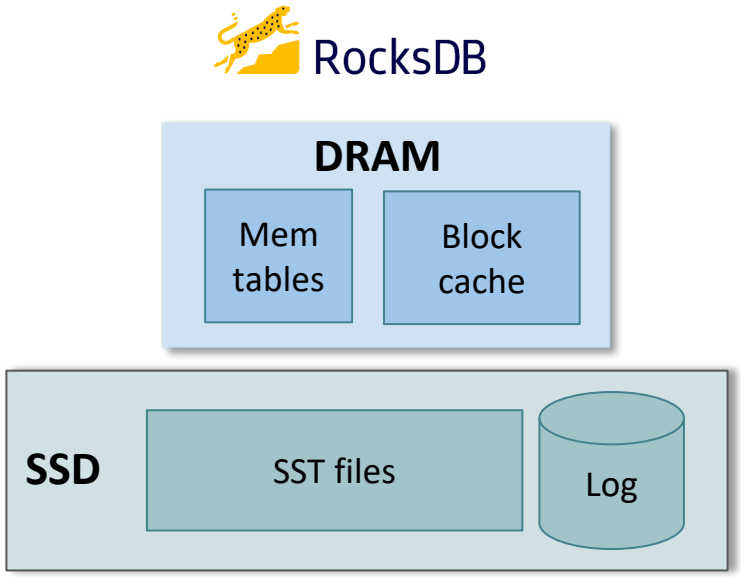
Intel Optane DC Persistent Memory (DCPMM)





- Byte addressable over **DDR-T** bus (pin compatible with DDR4)
- Are cheaper per GB than DRAM and comes in higher capacities than DRAM (upto 512 GB per module)
- **New high performance and cost effective 1P server variant with SCM addition**

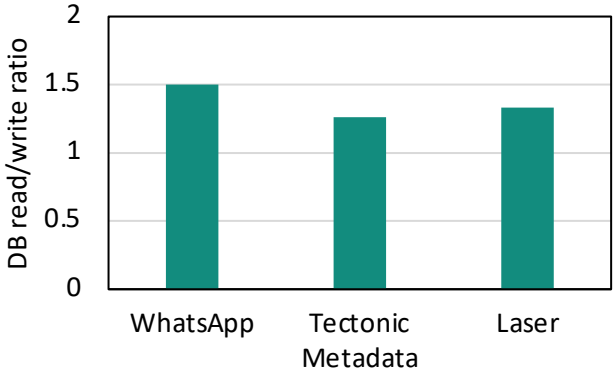
RocksDB Architecture

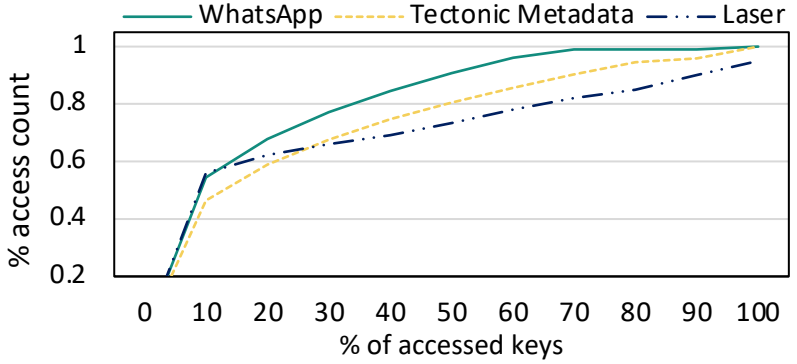


- RocksDB is a KV store used by many companies such as Facebook
- We studied representative production RocksDB workloads
- Largest memory consumption comes from block cache used for reads
- Read paths are critical and optimizing reads provides overall performance benefit

Facebook RocksDB Workload Characteristics

- We studied three large RocksDB workloads at Facebook, focused on read dominated workloads **(WhatsApp, Tectonic Metadata, Laser)**

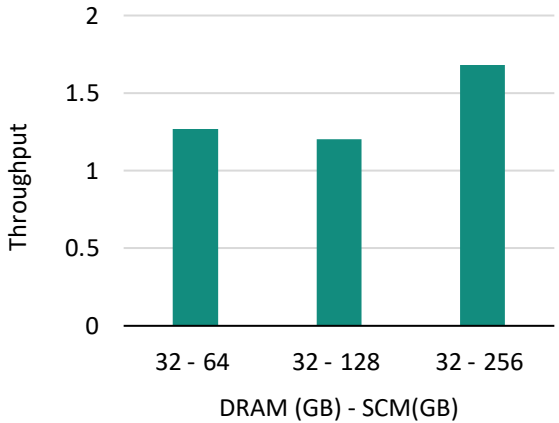




- Temporal locality in our workloads characterized by power law

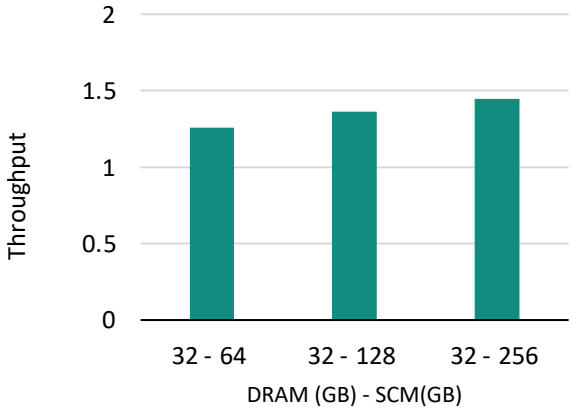
Intel Optane DC Persistent Memory Evaluations (WhatsApp)

Optimized App-direct throughput relative to Memory mode



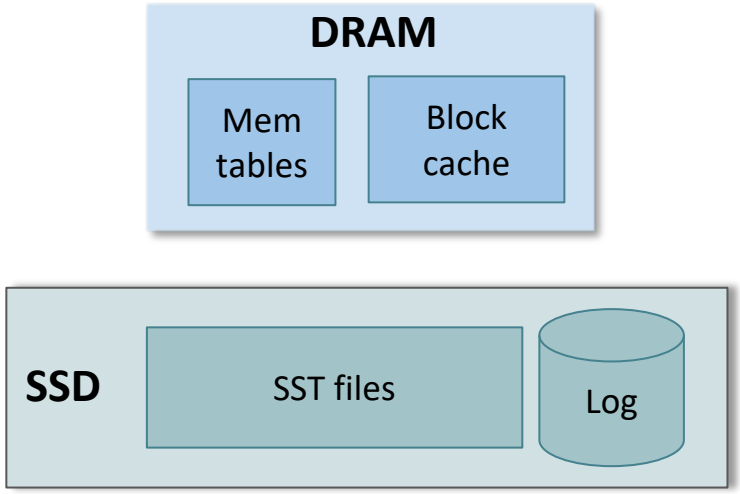
Optimized **App-direct** mode gives better performance than **Memory-mode**

Optimized App-direct throughput relative to naive SCM block cache

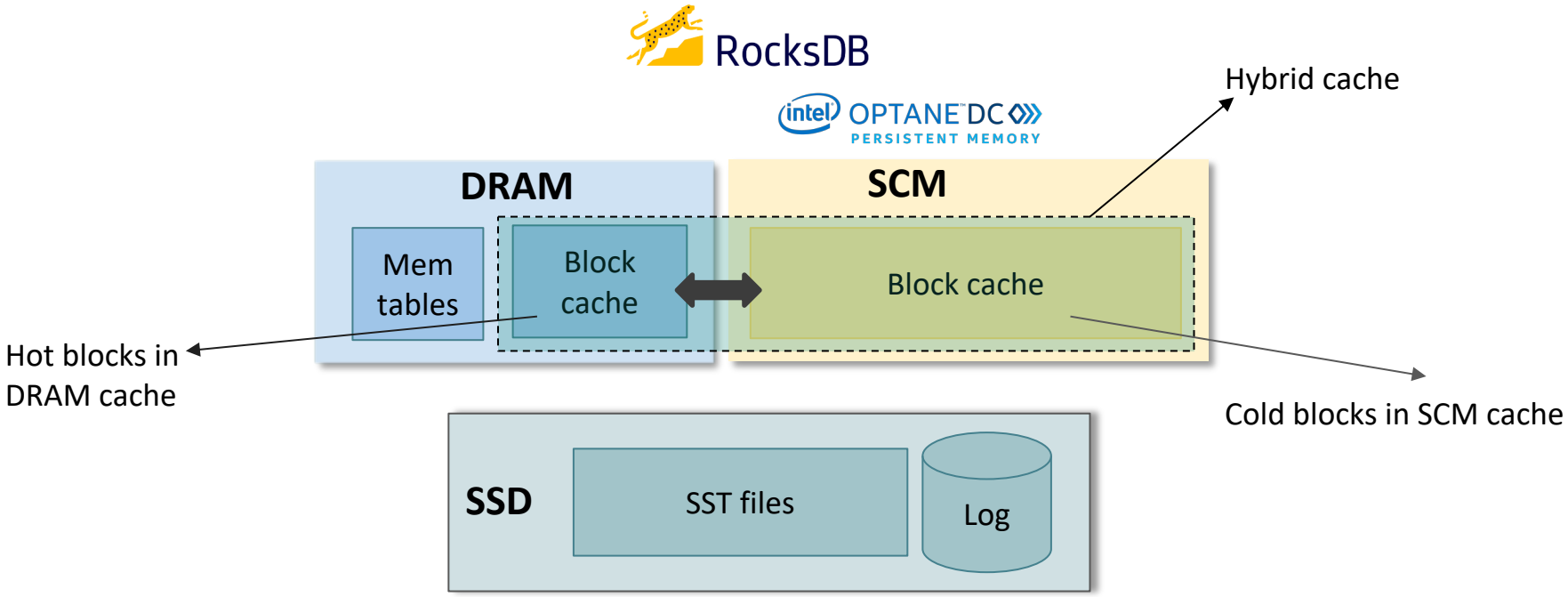


Optimized **App-direct** mode gives better performance than **naïve SCM block cache**

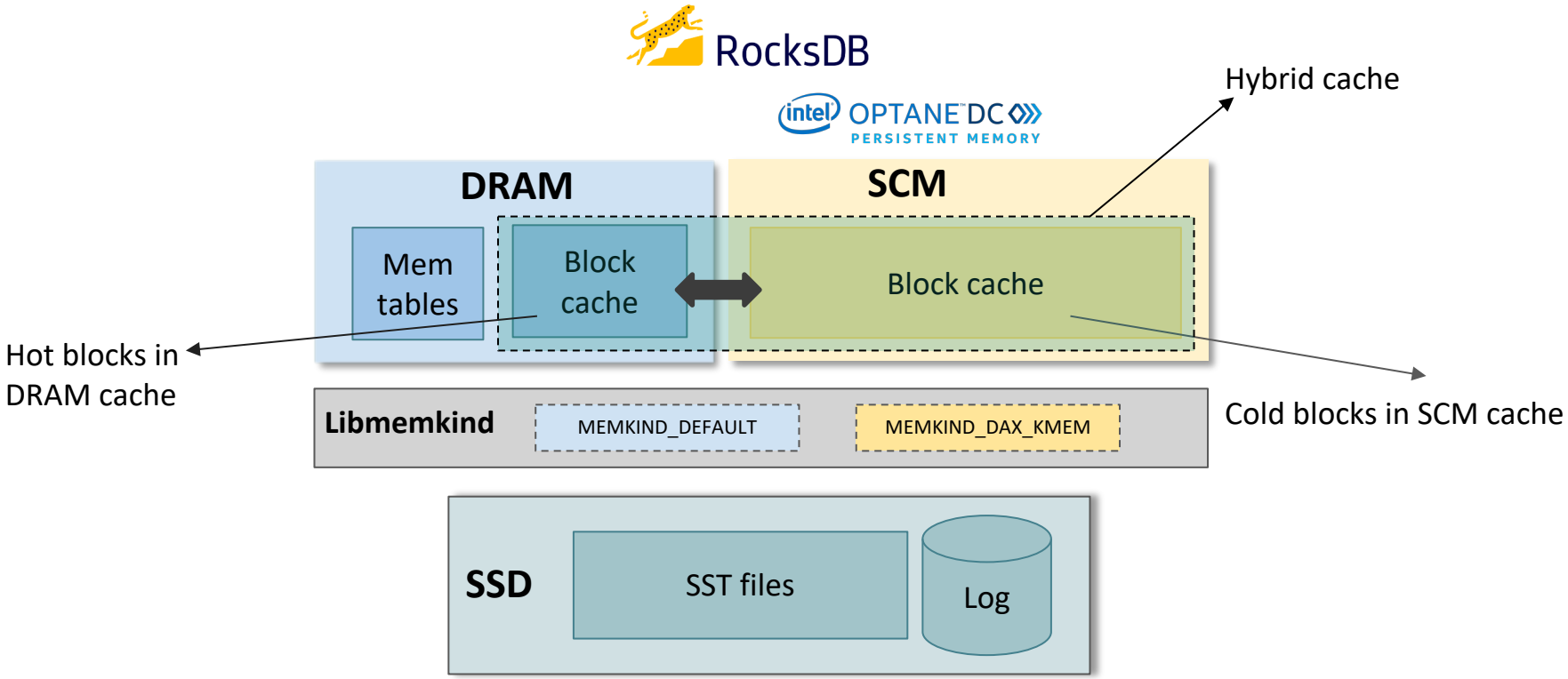
Facebook RocksDB extended with SCM



Facebook RocksDB extended with SCM

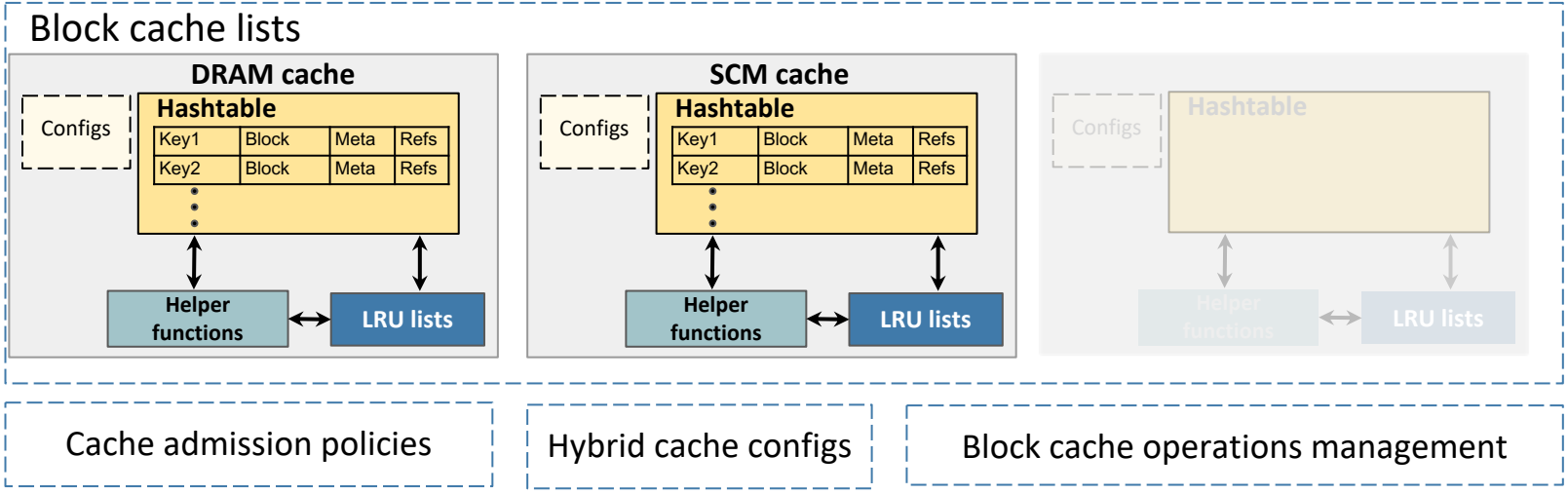


Facebook RocksDB extended with SCM

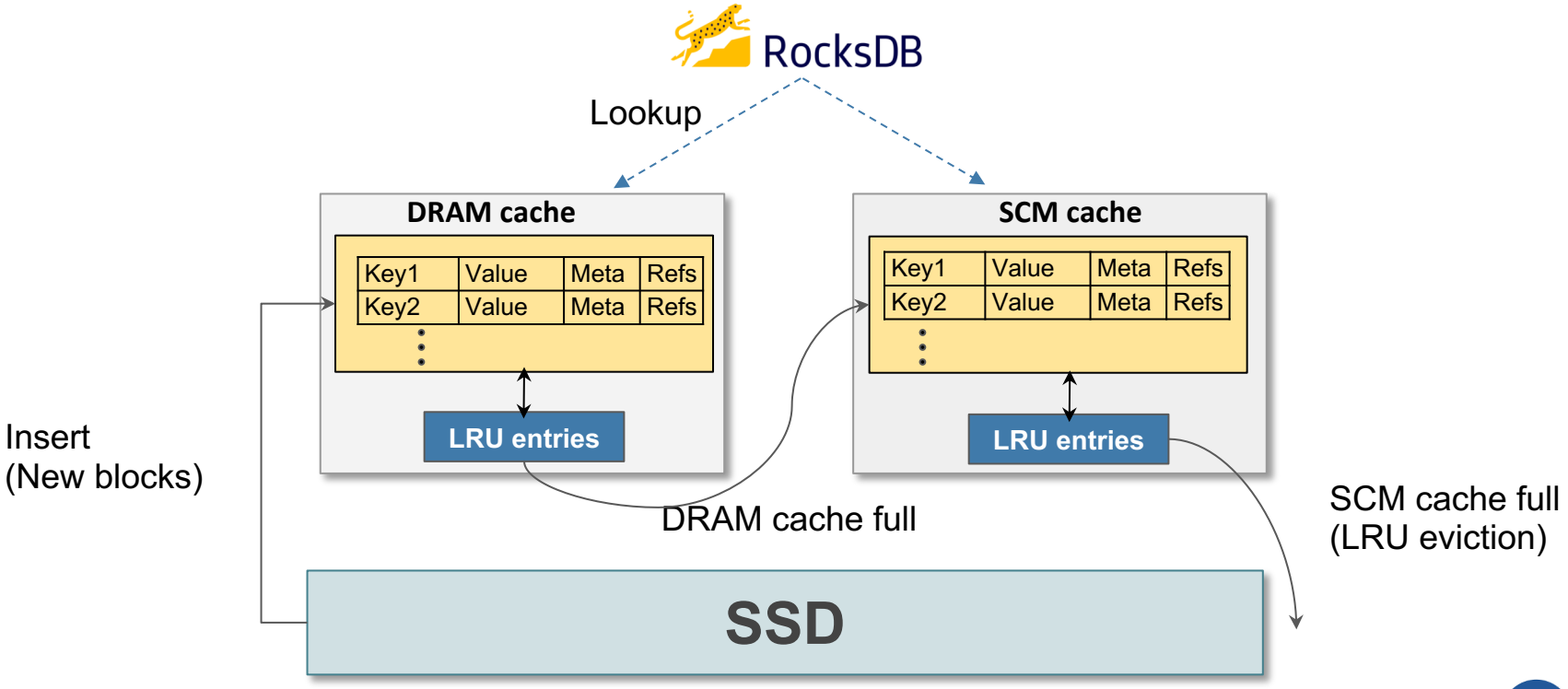


Hybrid Cache Design and Components

- New top level module in RocksDB to manage hierarchical block cache
- Supports multiple block caches and multiple memory types
- Admission policies to handle block transfer between multiple block caches



DRAM First Admission Policy



Evaluation Setup and Implementations

Hardware spec

- Two Socket Cascade Lake with Intel Apache Pass DIMMs (48 cores)
- 192GB DRAM, 1.5TB SCM
- Evaluated 1P variants based on DRAM and SCM component granularities and minimum DRAM to run linux OS/software.

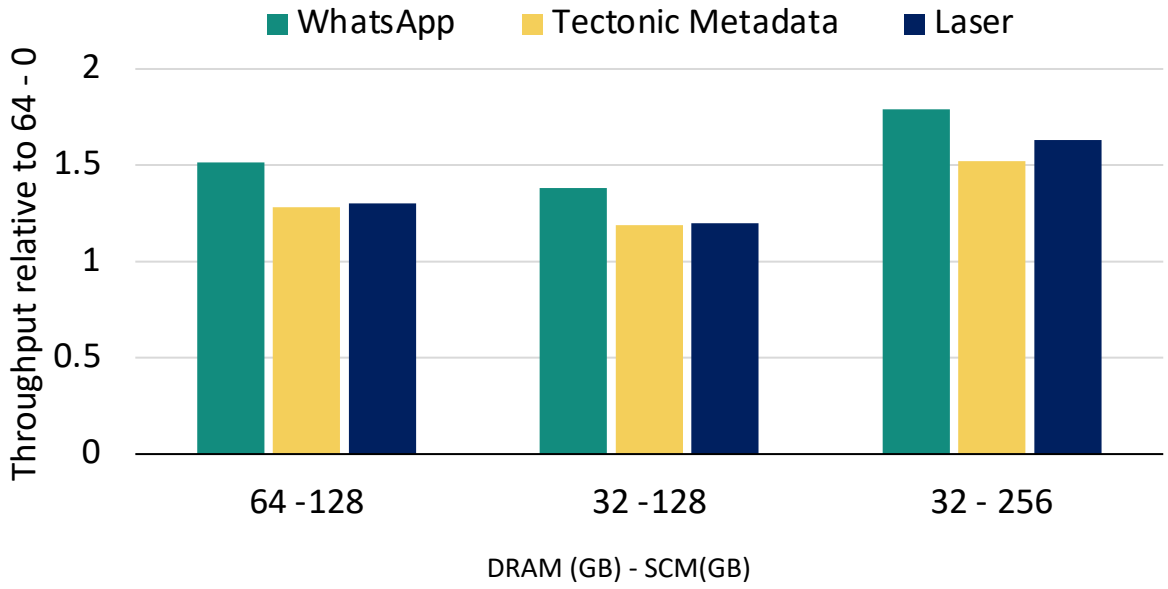
Server Type	1P (64 - 0)	1P variant (64 - 128)	1P variant (32 – 128)	1P variant (32 – 256)
DRAM size (GB)	64	64	32	32
SCM size (GB)	0	128	128	256

RocksDB 6.10 with libmemkind library used for memory allocations

Workload Creation

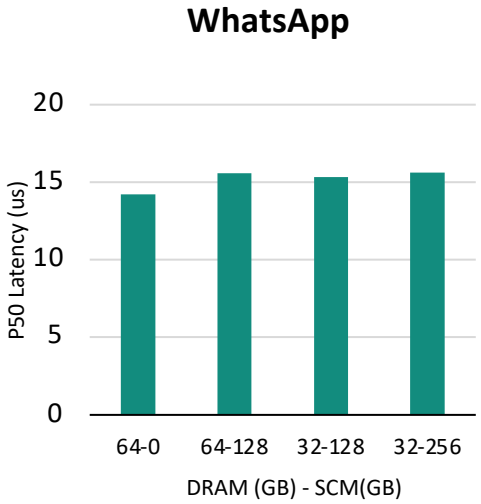
- We run db_bench with multiple dbs sharing the block cache
- Number of threads to simulate number of readers
- Realistic workload patterns sampled from production hosts
  - KV access distribution
  - Value distribution per put query
  - Average key and value sizes when creating db
  - Query composition for get, put, and seek
- Three sample workloads; WhatsApp, Tectonic Metadata and Laser

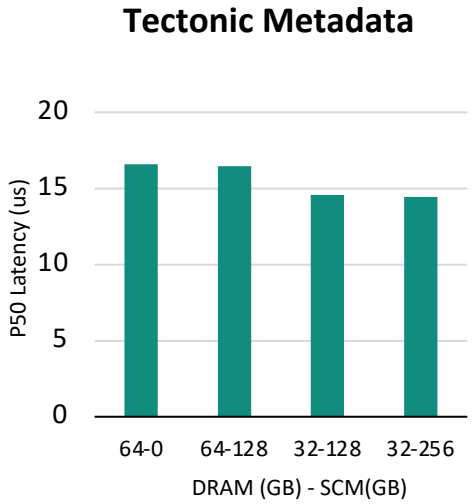
DRAM First Policy Throughput all workloads

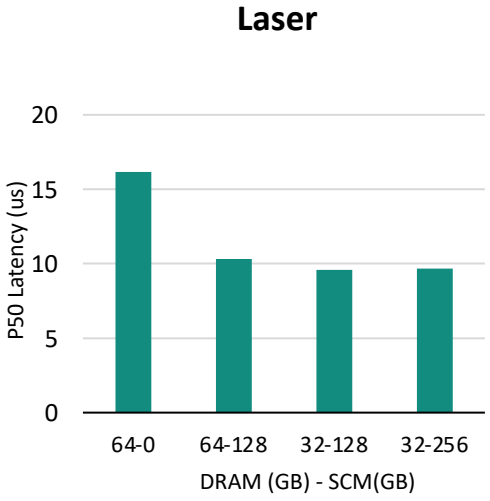


- Adding SCM to 1P servers with 64GB DRAM provide 25% - 80% throughput improvement in all configuration for all workloads

DRAM First Policy Latency improvement for all workloads (P50 latency)

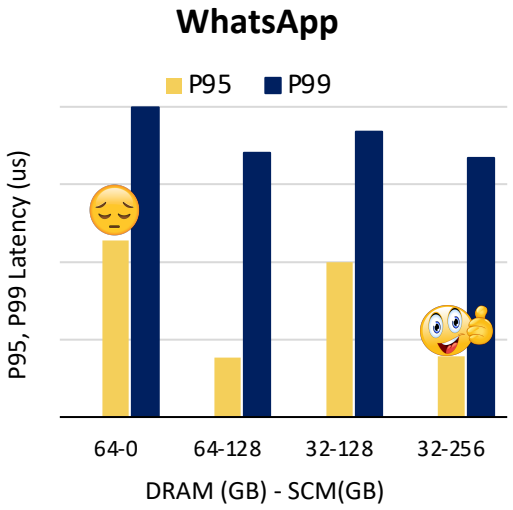


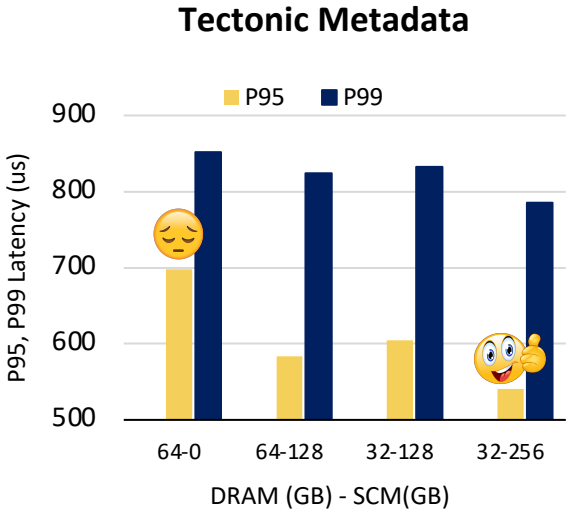


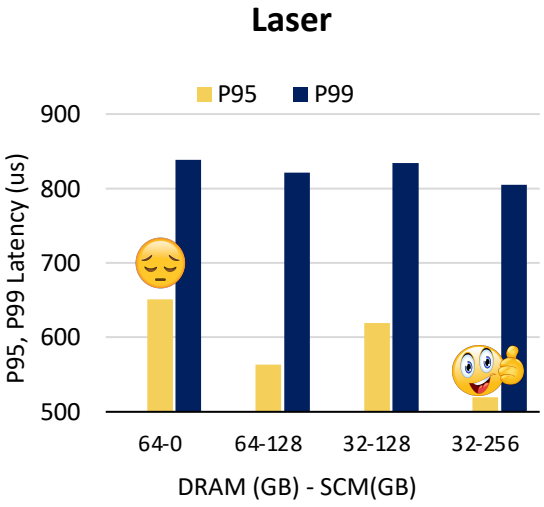


- P50 latency stays the same with small variations

DRAM First Policy Latency improvement for all workloads (P95 and P99)

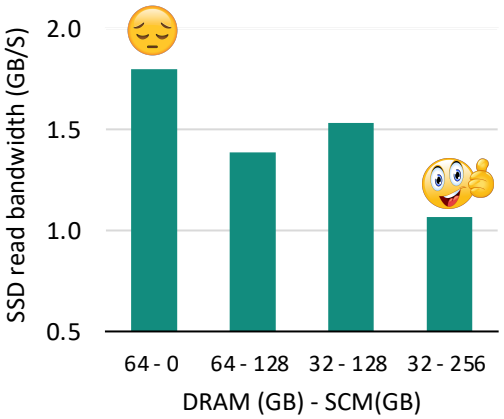


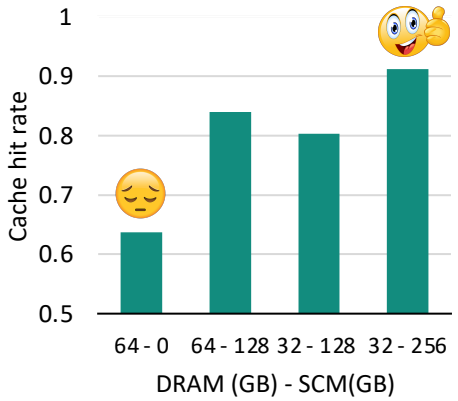


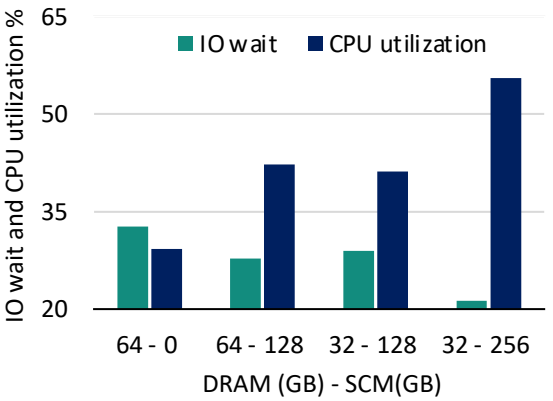


- P95 and P99 are affected much more by the reduction of DRAM size, an average of ~20% and ~10% improvement in P95 and P99

Deep Dive Analysis

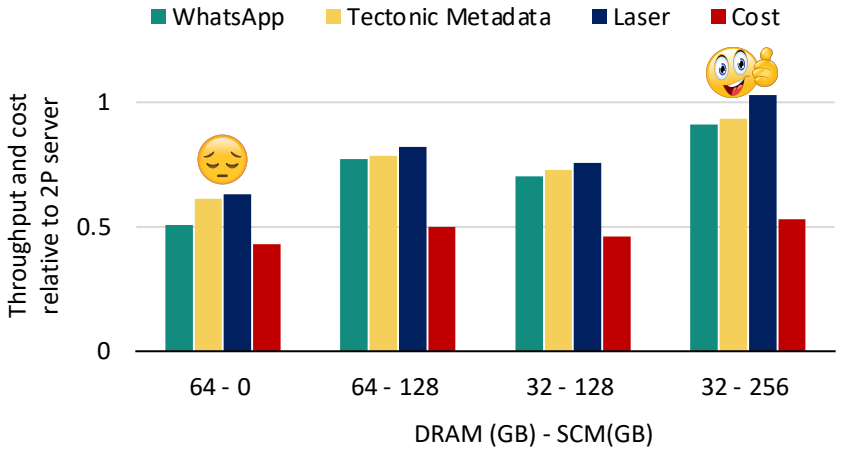






- **WhatsApp** IO bandwidth, cache hit rate, and IO wait and CPU utilization for dram 1st policy shown above
- SCM increases cache hit rate upto 30% and IO bandwidth reduced by 0.8 GB/s.
- CPU utilization increases but still within the one CPU limit
- Other workloads also follow the same trend as above

Cost and Performance



Sever types	Relative TCO
2P server	1.0
1P server	0.72 – 0.86
1P (32 – 256)	0.52 – 0.57

- 1P variants (32 – 256) achieve ~95% performance of 2P servers
- SCM cost relative to DRAM per GB is 0.38
- Reduces TCO by 43% - 48% compared to 2P servers, leading to power reduction per service

Limitations of SCM Deployments

- **Workloads**: A number of write-heavy workloads at Facebook that would not benefit from SCMs.

- **Reliability**: SCMs are not widely available in the market, the reliability of SCMs is a concern until they have been proven in mass deployments.

- **SKU diversification:** Adding a new hardware configuration into the fleet incurs additional qualification, maintenance, and scale deployment costs not included in this TCO analysis.

Conclusion

- We studied the benefit of SCM using Intel Optane memory for 3 read dominated RocksDB workloads
- We extended RocksDB to add new hybrid cache with multiple block cache that uses different kinds of memories
- Demonstrate up to ~50% throughput improvement and ~20% latency decrease by creating new 1P server variants

Acknowledgement

- Yanqin Jin
- Jesse Koh
- Siying Dong
- Pallab Bhattacharya
- Shumin Wu
- Darryl Gardner
- Vijay Balakrishnan
- Anonymous reviewers for their feedback

Thank You

Contact Info: Hiwot Tadese Kassa (hiwot@umich.edu)