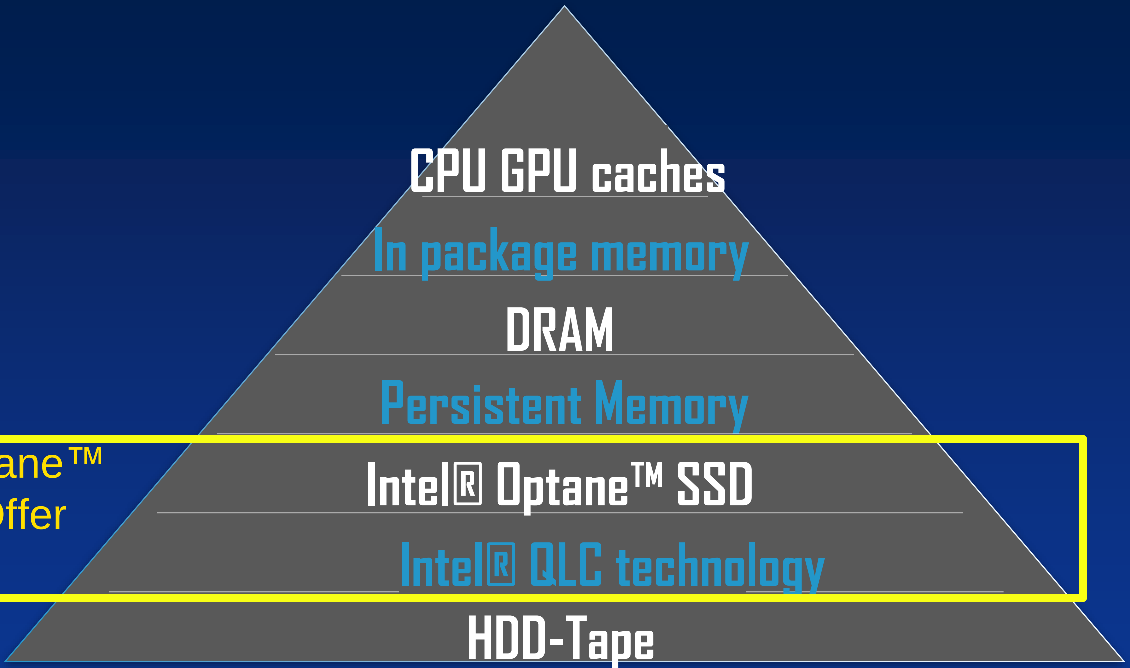
Flash Memory Summit

Approaches that Combine Intel® Optane™ SSDs with Intel® QLC technologyOffer Solutions Competitive with TLC



Kapil Karkra, Intel

Michal Wysoczanski, Intel

Piotr Wysocki, Intel

Santa Clara, CA
August 2019

Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Performance results are based on testing as of July 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

Intel® Optane™ SSD + Intel® QLC 3D NAND SSDs

Intel® QLC 3D NAND SSDs

- Compelling **cost** and **density**
- Good **read performance**

But...

- Write bandwidth dependent on workload (WAF)
- Low endurance

Intel® Optane™ SSD:

- **Write bandwidth** independent of workload (no WAF, in-place overwrites)
- Very low **latency** and great **QoS**
- Maximum performance even at **low queue depth**
- High **endurance**

Why not to combine both to create an optimal solution?

A Note on Workloads Suitable For Intel® QLC Technology

Flash Memory Summit

- Workloads with Write Amplification Factor equal to unity (WAF=1) are not always suitable for QLC due to its low endurance
  - e.g., a streaming workloads with a Time to Live (TTL=18 days) policy might be suitable, but not a workload with TTL = 1 minute

*QLC Media Writes = WAF * Host Writes*

With NVMe features like Streams and ZNS, the best software can achieve is WAF =1, at which point:*

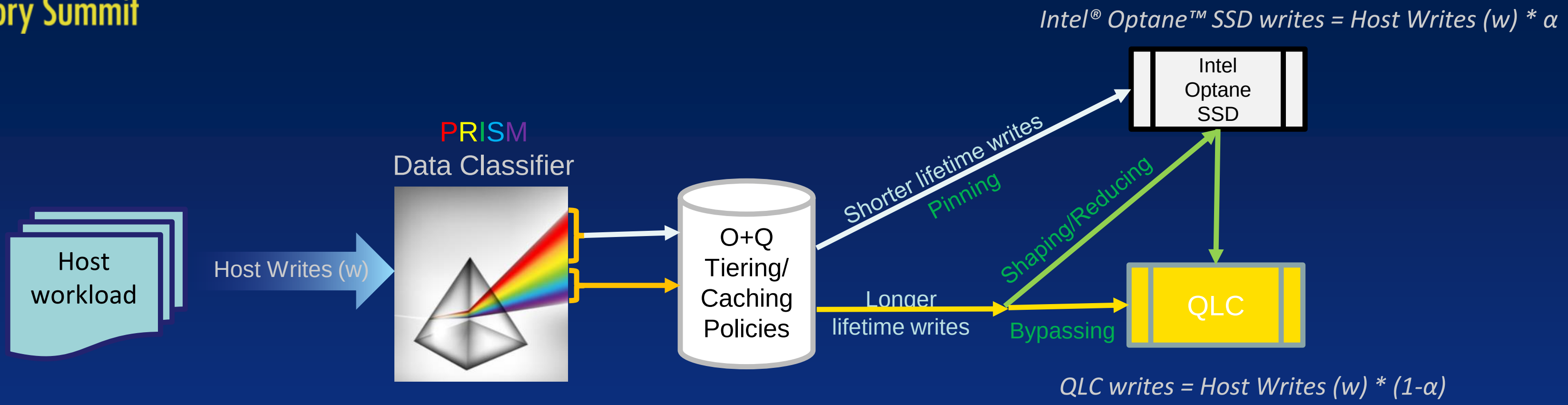
QLC Media Writes = Host Writes

- How do we make QLC Media Writes << Host Writes?
- Hint: Trap some host writes upstream in the memory/storage hierarchy in the Intel® Optane™ SSD

The key question that we answer next is: What kind of data is suitable for Intel® Optane™ SSDs?

*Other names and brands may be claimed as the property of others.

How to combine Intel® Optane™ SSDs and Intel® QLC Technology into O+Q?



- **Goal:** Reduce amount of writes that goes to Intel® QLC technology (α – Write Reduction Factor).
- **How:** Place on Intel® Optane™ SSD data that is small enough and generates a lot of writes (β – Write Invalidation Factor).
- **Key:** PRISM – Data classification and separation according to WIF. Need to separate different data classes that meet WIF and WRF requirements.
- **Extra:** Shape/reduce the workload part that goes to QLC to improve WAF (e.g., classify data based on data lifetime into classes in Intel Optane SSDs and place them on separate zones/streams in QLC; or stage data on Intel Optane SSDs, compress it to reduce it, and place compressed data on QLC.

There is a need for software that would provide PRISM.
But what kind of data should we look for?

What kind of data should go to Intel® Optane™ SSDs?

- Through several more macro and micro benchmarks, a thumb rule emerged: Place TML+H on Intel® Optane™ SSDs

- **T**emp data – The intermediate data that's discarded upon arriving at final result (e.g., data swapped out from memory)
- **M**etadata – Metadata (e.g., indexes that are updated and read a lot)
- **L**ogs – Journals, write ahead logs, redo logs, undo logs, binary logs, transaction logs, Parallel raft or Paxos logs
- **H**ot data – The data that is read frequently (e.g., popular songs)

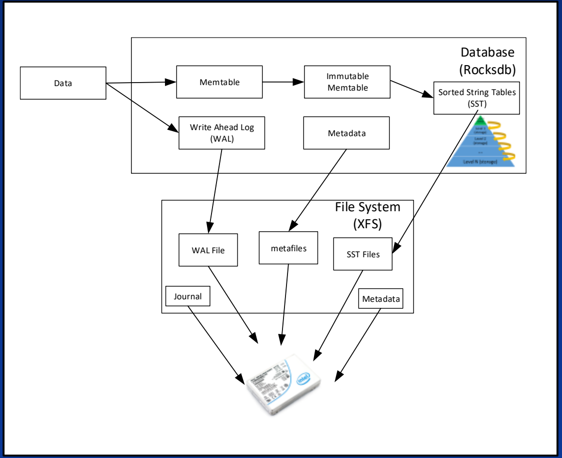
Use case 1: RocksDB

PRISM for RocksDB:

- Write lifetime hints – Open CAS in block layer consumes WLTH provided by RocksDB
  - WAL and levels 0-3 are placed on Intel® Optane™ SSD
  - Lower levels go directly to Intel® QLC technology
- Additionaly, file system metadata is also placed on Intel Optane SSD

79% of α can fit within Intel Optane SSD

Rocksdb workload through the Prism	Occupancy (GB)	Writes (GB)	WRF (α)	WIF (β)	Type of data (HTML)
XFS Metadata + Journal	2.0	5.7	0%	3	M, L
WAL	4.0	688.9	5%	172	L
L0	1.0	689.6	5%	690	T
L1	1.0	1012.3	8%	1012	T
L2	11.8	3157.0	25%	268	T
L3	100.0	4511.9	36%	45	T
L4/L5	1000.0	2601.9	21%	3	X
Total (GB)	1119.8	12667.3			



Rocksdb (CLS=64, md, XFS) read QoS numbers for level 0	Endurance/Performance Metrics			Relative	
Benchmark: fillseq+readwhilewriting	TLC	QLC	O+Q	O+Q vs. TLC	O+Q vs. QLC
EPBW	24.92	8.86	37.22	149%	420%
EDPWD	1.71	0.63	2.65	156%	420%
write bandwidth (MB/s)	29.83	25.96	40.32	135%	155%
p50	163.93	522.03	74.07	45%	14%
p75	423.68	1263.73	240.07	57%	19%
p99	5041.73	7961.48	1281.36	25%	16%
p99.9	17866.50	18953.32	4213.98	24%	22%
p99.99	30199.82	73848.21	13682.56	45%	19%

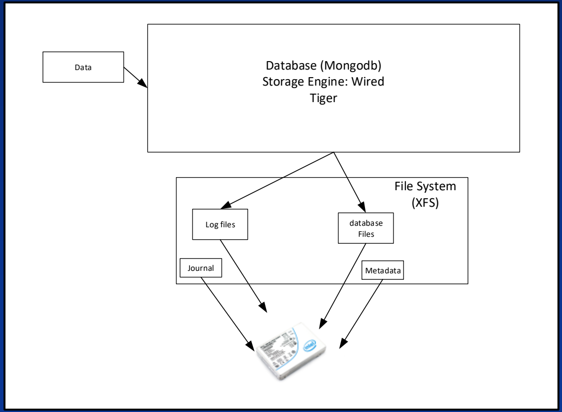
See „Notes on benchmarks" for benchmark specification

Use case 2: MongoDB

PRISM for MongoDB:

- Use Open CAS file path classification to put directory with MongoDB journal on Intel® Optane™ SSD

Mongodb workload through the prism	Occupancy (GB)	Writes (GB)	WRF (α)	WIF (β)	Type of data (HTML)
Journal	3.0	27478.3	97%	9159	L
Data	750.0	776.7	3%	1	X
Total	753.0	28255.0			

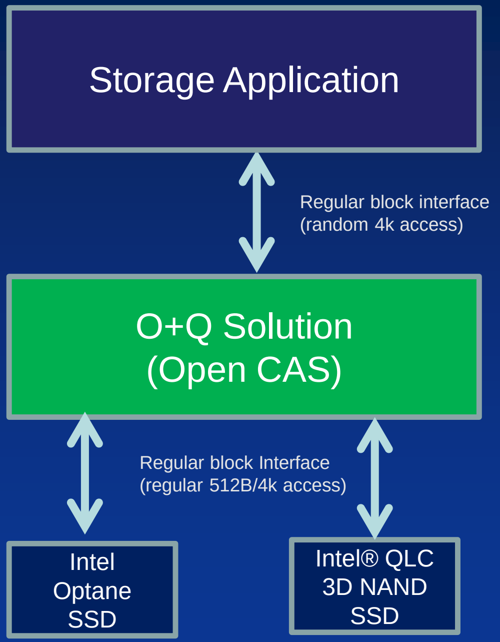


Mongodb (journal on Intel® Optane™ SSD)	Endurance/Performance Metrics		Relative
Benchmark: YCSB Workload A rd:wr 50:50 (zipf θ = 0.99)	TLC	O+Q	O+Q vs. TLC
EDPWD	0.27	3.01	1120%
Document Size 400B			
OVERALL Throughput (Ops/s)	28935	50703	175%
UPDATE 99thPercentilLatency (u	909	419	46%
OVERALL Throughput (Ops/s)	6155	1124	18%
Document Size 8kiB			
OVERALL Throughput (Ops/s)	34853	41621	119%
UPDATE 99thPercentilLatency (u	705	543	77%
OVERALL Throughput (Ops/s)	1752	1399	80%

See „Notes on benchmarks” for benchmark specification

Intel® Optane™ SSD + Intel® QLC technology solution vision

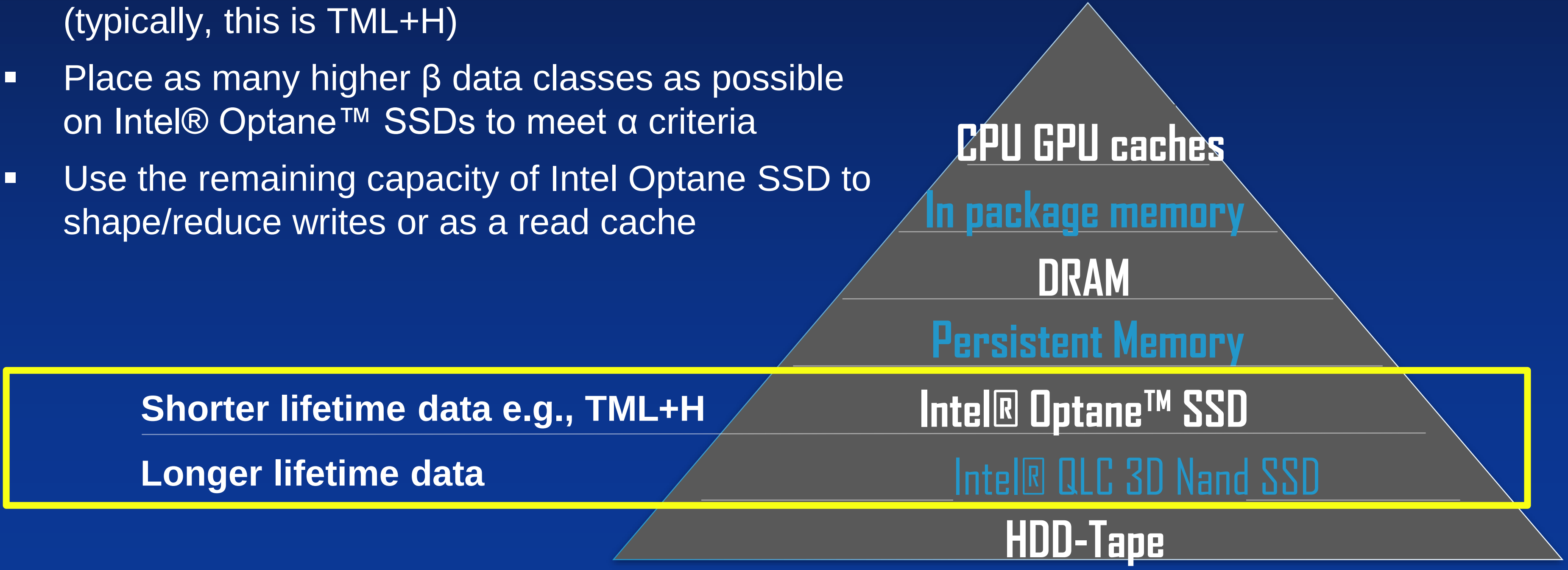
- Transparent to existing applications – middleware that uses regular block interface API
- Intel® Optane™ SSD and Intel® QLC technology accessed through regular block API
- Uses Open CAS as O+Q solution vehicle (PRISM, shaping, reducing, and read caching)
- Could be implemented as Linux kernel block device or SPDK bdev



Conclusion

Flash Memory Summit

- Build PRISMs (data classifiers) with Open CAS to identify data classes in the workload with high β (typically, this is TML+H)
- Place as many higher β data classes as possible on Intel® Optane™ SSDs to meet α criteria
- Use the remaining capacity of Intel Optane SSD to shape/reduce writes or as a read cache



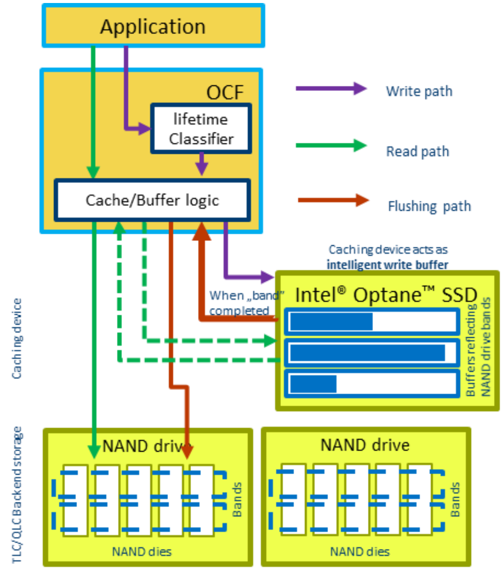
Extra: Shape/Reduce/Read Cache

Shaping

- Writes that are not pinned on Intel® Optane™ SSDs go to Intel Optane SSD write buffer portion
- Data in buffer partitioned, based on streams classifier
- Flushing of data performed in buckets of size equal to NAND drive band size
  - Only one bucket at a time
  - Band filled with data with same stream (e.g., same lifetime)
- Reads are handled directly from NAND drive (except for data that has not been flushed yet)

Reducing

- Compression
- Triple replication in Intel Optane SSD and Erasure Coding in QLC

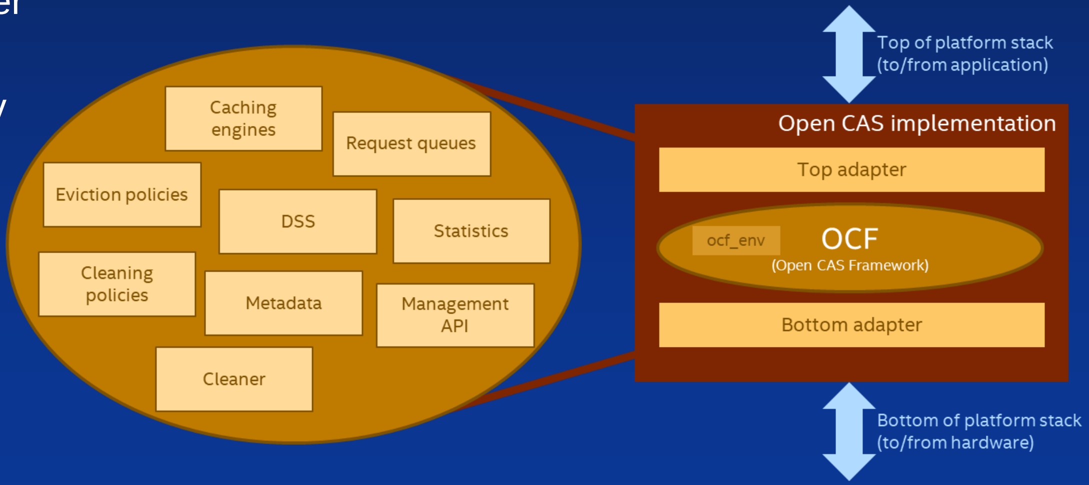


	Baseline	Intel® Optane™ SSD Buffer
Host writes	223,395	223,623
NAND writes	549,940	280,223
WAF	2.46	1.25

NAND drive: Cliffdale P4500 1TiB, Model: INTEL SSDPE2KX010T7
Optane drive: P4800X 280GiB, Model: INTEL SSDPED1D280GA
Benchmark: FIO, 3 sequential write streams (different velocity), 1 random write stream, 12h

Open CAS

- Open source caching engine
- Environment independent core and platform specific adapters
- Available adapters:
  - Linux kernel block device driver
  - Storage Performance Development Kit (SPDK) bdev
- Available on GitHub
  (https://github.com/Open-CAS)
- Tools for building solution:
  - Classifiers (PRISM)
  - Shaping/cleaning policies



Flash Memory Summit

Notes on benchmarks

RocksDB

CPU Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz, 2 sockets, 22 cores, memory 256GiB, BIOS Version: SE5C610.86B.01.01.0016.033120161139. Release Date: 03/31/2016
Fedora 25 (kernel 4.13.16), RocksDB v 5.17.2.
Drives used:
- Intel® SSD DC P4510 8TB
- Intel® SSD D5-P4320 7.68 TB
- Intel® Optane™ SSD DC P4800X 375GB

Preparation phase:
db_bench --db=/mnt/rocksdb --num_levels=6 --key_size=32 --value_size=1024 --block_size=4096 --cache_size=$((8 * GiB)) --cache_numshardbits=6 --compression_type=none --compression_ratio=0.5 --hard_rate_limit=2 --rate_limit_delay_max_milliseconds=1000000
--write_buffer_size=$((1024 * MiB)) --max_write_buffer_number=4 --target_file_size_base=$((128 * MiB)) --max_bytes_for_level_base=$((1024 * MiB)) --max_bytes_for_level_multiplier=10 --sync=0 --verify_checksum=1 --delete_obsolete_files_period_micros=$((60 * MiB))
--statistics=1 --stats_per_interval=1 --stats_interval=$((1 * M))
--histogram=1 --memtablerep=skip_list --bloom_bits=10 --num_multi_db=1 --open_files=$((20 * KiB)) --max_background_compactions=32 --max_background_flushes=32 --level0_file_num_compaction_trigger=7 --level0_slowdown_writes_trigger=16 --
level0_stop_writes_trigger=24 --benchmarks=fillseq --use_existing_db=0
--num=${key_no} --threads=1

Benchmark phase:
db_bench --db=/mnt/rocksdb --num_levels=6 --key_size=32 --value_size=1024 --block_size=4096 --cache_size=$((8 * GiB)) --cache_numshardbits=6 --compression_type=none --compression_ratio=1 --hard_rate_limit=2 --rate_limit_delay_max_milliseconds=1000000 --
write_buffer_size=$((1024 * MiB))
--max_write_buffer_number=4 --target_file_size_base=134217728 --max_bytes_for_level_base=1073741824 --sync=0 --verify_checksum=1 --pin_l0_filter_and_index_blocks_in_cache=false --cache_index_and_filter_blocks=false --mmap_read=0 --
max_background_compactions=32 --max_background_flushes=32
--disable_auto_compactions=0 --statistics=1 --stats_per_interval=2 --histogram=1 --memtablerep=skip_list --bloom_bits=10 --use_direct_reads=1 --open_files=-1 --level0_file_num_compaction_trigger=8 --level0_slowdown_writes_trigger=16 --level0_stop_writes_trigger=24
--benchmarks=readwhilewriting
--use_existing_db=1 --stats_interval=5000000 --num=$((600*M)) --threads=4

MongoDB

CPU Intel(R) Core(TM) i7-4960X CPU @ 3.60GHz, 1 socket, memory 16GiB
CentOS 7.6 (kernel 3.10.0-957), MongoDB v 4.0.6, YCSB 0.15.0
Drives used:
- Intel® SSD DC P4500
- Intel® Optane™ SSD DC P4800X 375GB

Preparation phase:
ycsb load mongodb -s -p recordcount=100000000 -threads 16 -P workloads/workloada -p fieldlength=1024 -p fieldcount=8 -p requestdistribution=zipfian -p mongodb.url=mongodb://localhost:27017/ycsb?j=true
Benchmark phase:
ycsb run mongodb -s -t -p operationcount=2000000000 -threads 16 -P workloads/workloada -p fieldlength=1024 -p fieldcount=8 -p requestdistribution=zipfian -p mongodb.url=mongodb://localhost:27017/ycsb?j=true