STORAGE DEVELOPER CONFERENCE
SDC21 EMEA
BY Developers FOR Developers

Virtual Conference
June 8, 2021

MASSé:
Media Aware Smart Storage Engine

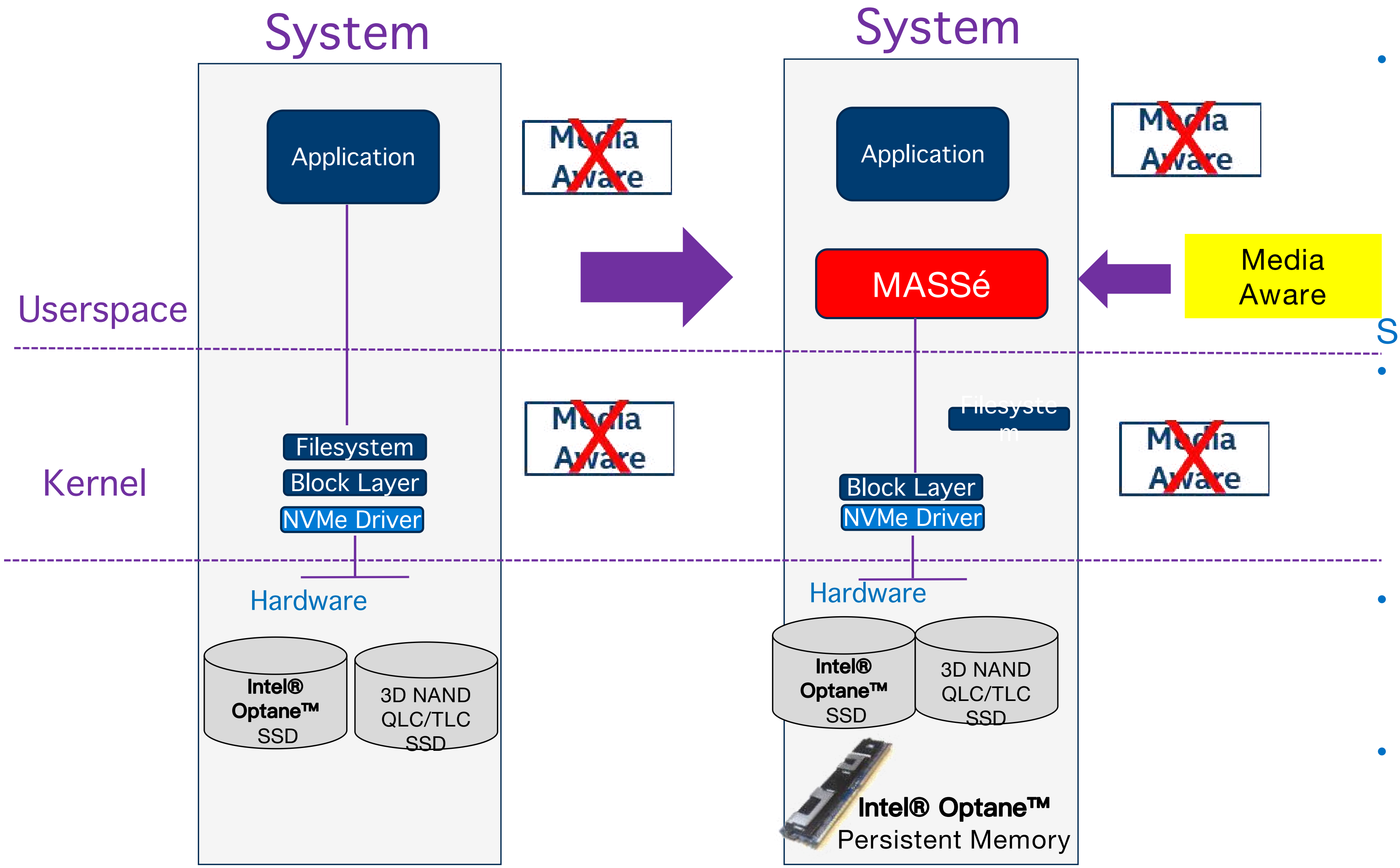
Jack Zhang
Cloud & Enterprise Architect
yuan.zhang@intel.com

Agenda

- MASSé introductions, Tiered storage for Optane+QLC
- MASSé Evaluation and Proof
- What Comes Next

MASSé = Media Aware Smart Storage Engine

MASSé : Overviews



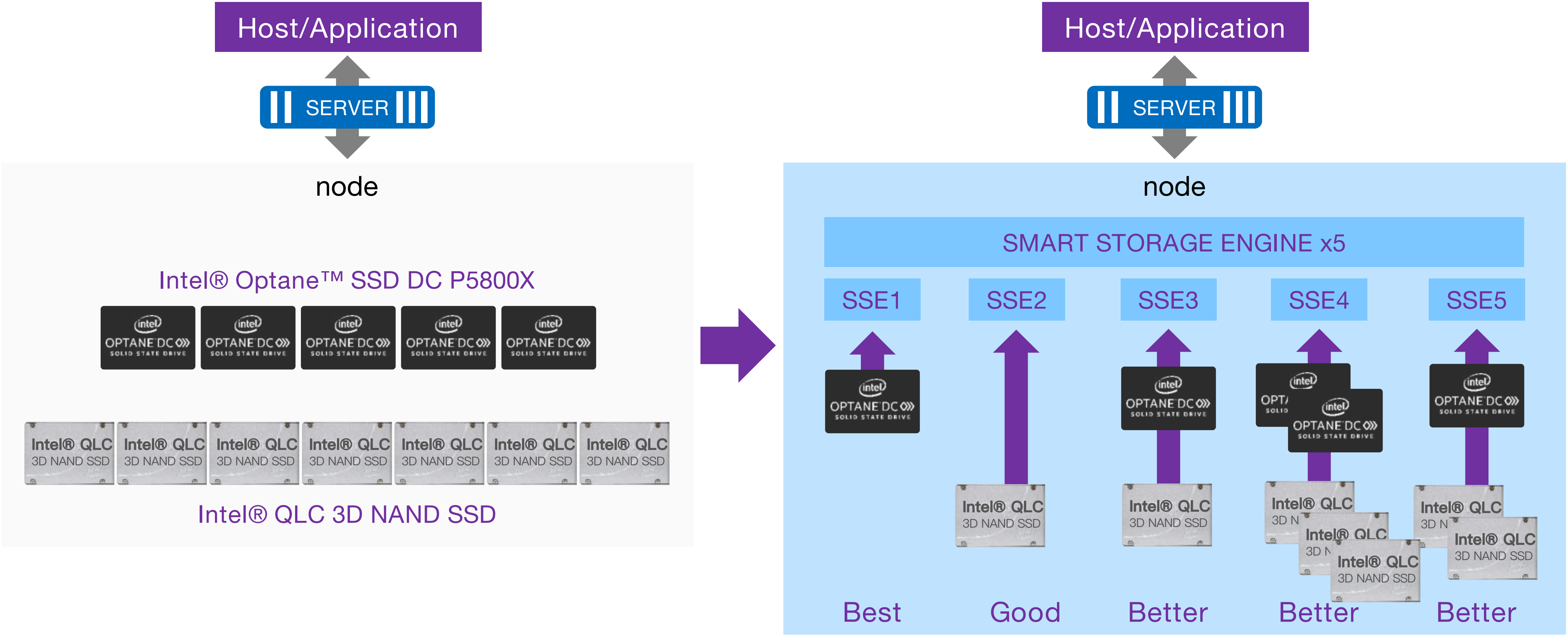
Feedbacks:

- “Why do I not see x number of times improvement over flash SSDs when dropped in an Intel® Optane™ SSD?”
- “Re-shaping writes into larger datasets and sequentially sending to a QLC SSD requires additional software investments, and implementations differ from application to application…is there a generic solution that supports this?”

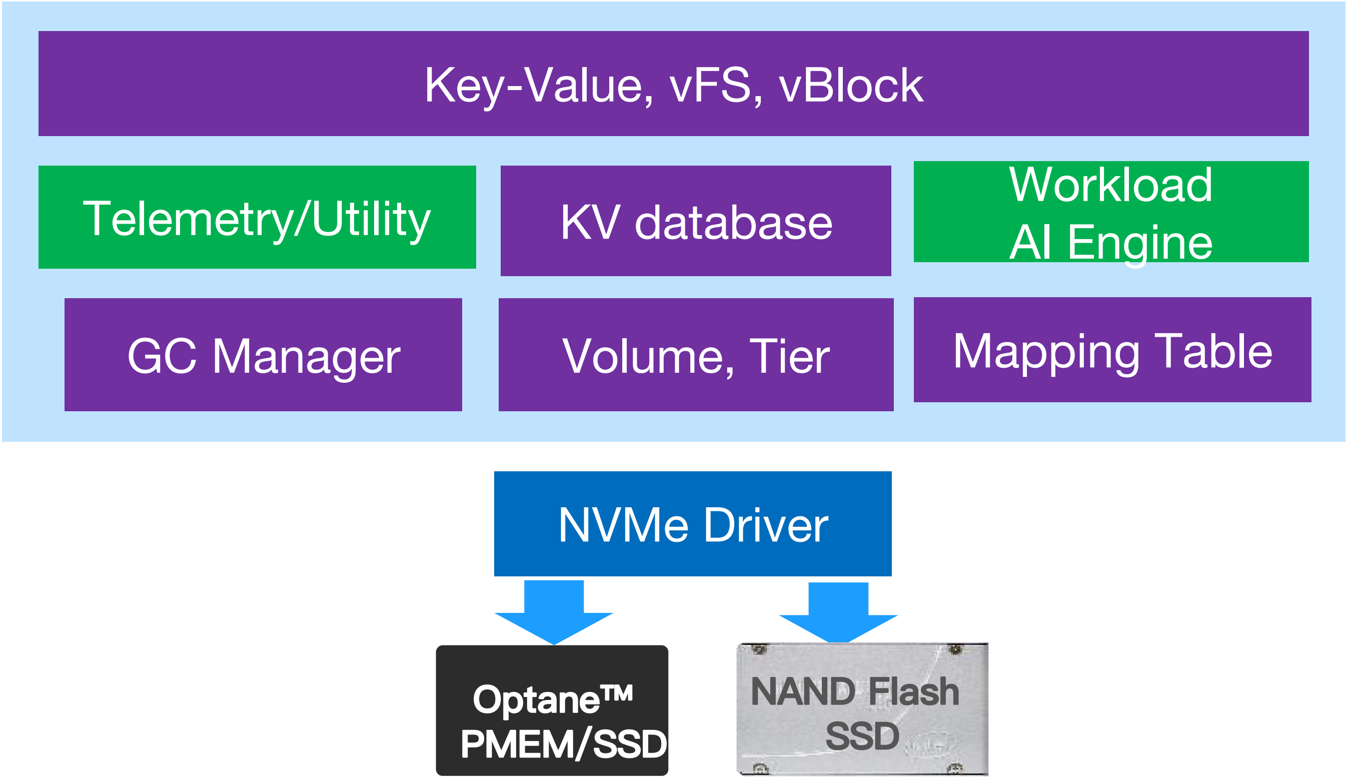
Solution:

- Media Aware --uniquely identifies and classifies heterogeneous SSDs by their media type, and builds inclusive data structures and algorithms, accordingly, helping to release maximum SSD capabilities to applications
- Smart -- intelligent module features such as data placements, IO re-shaping, key-value/virtual filesystem/virtual block APIs, workload pattern AI engine etc,
- Storage Engine --replacement of filesystem and managing raw SSD blocks without modifying SSD firmware and kernel modules

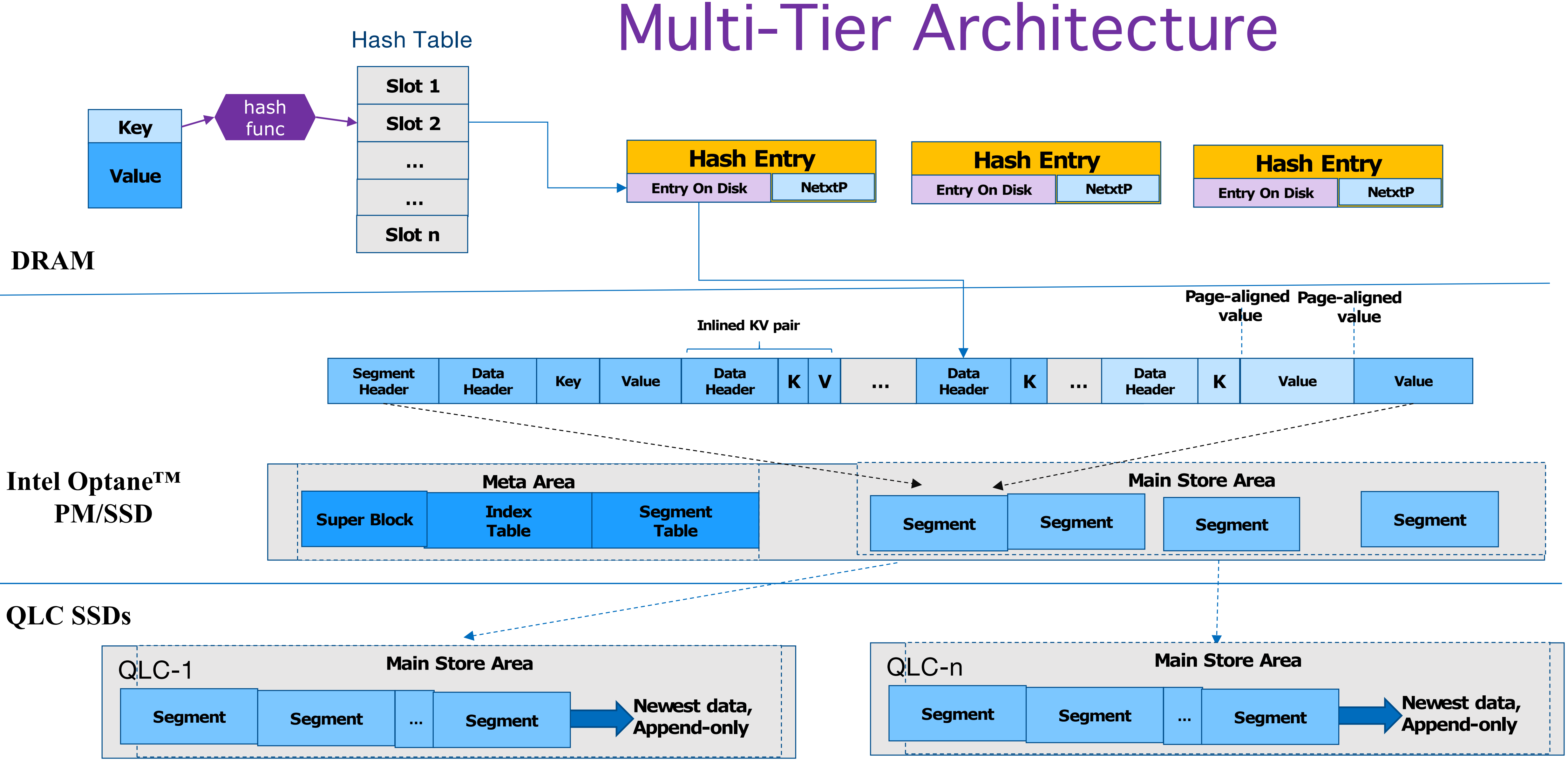
Configurable Engine



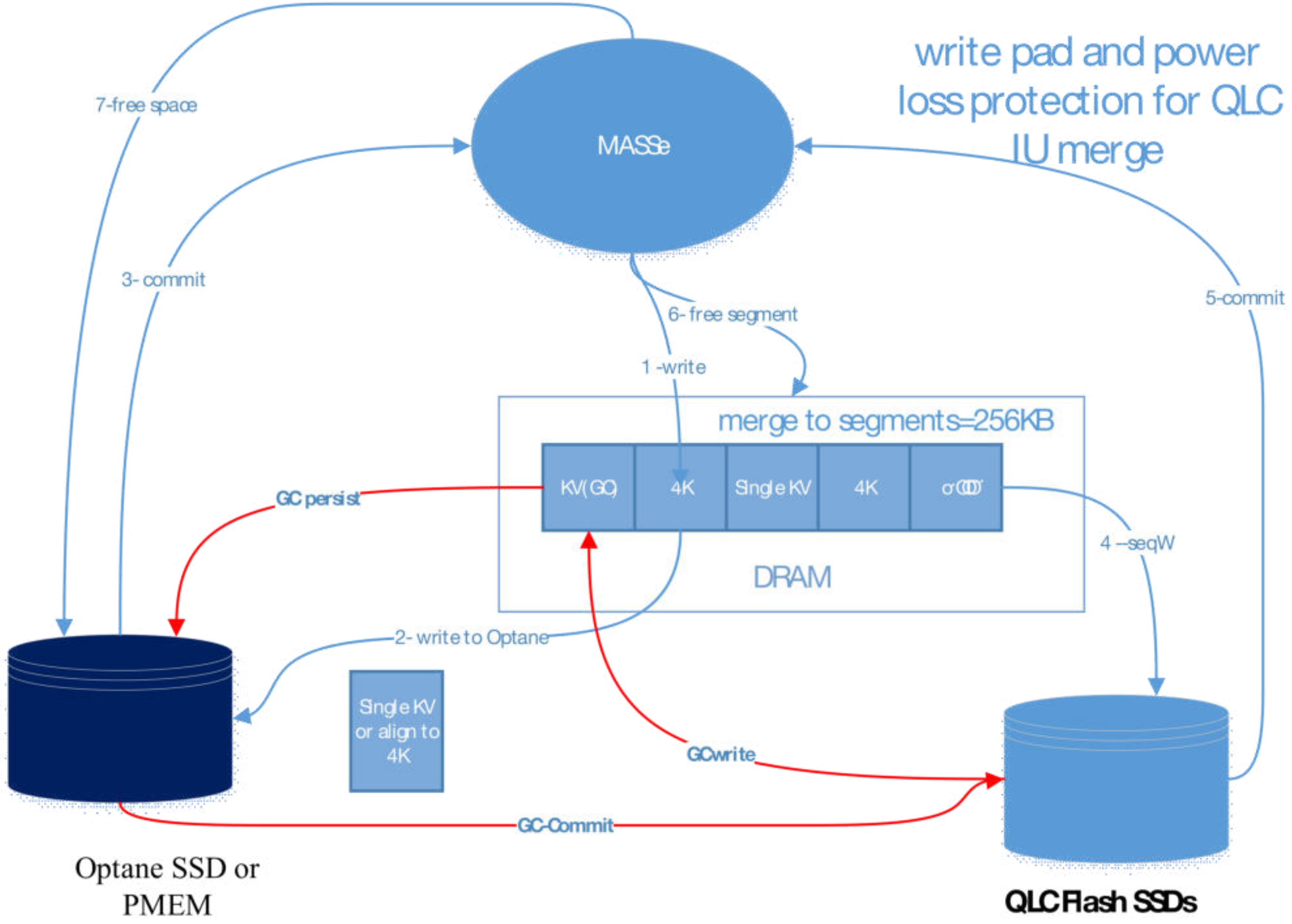
Software Architecture



Multi-Tier Architecture

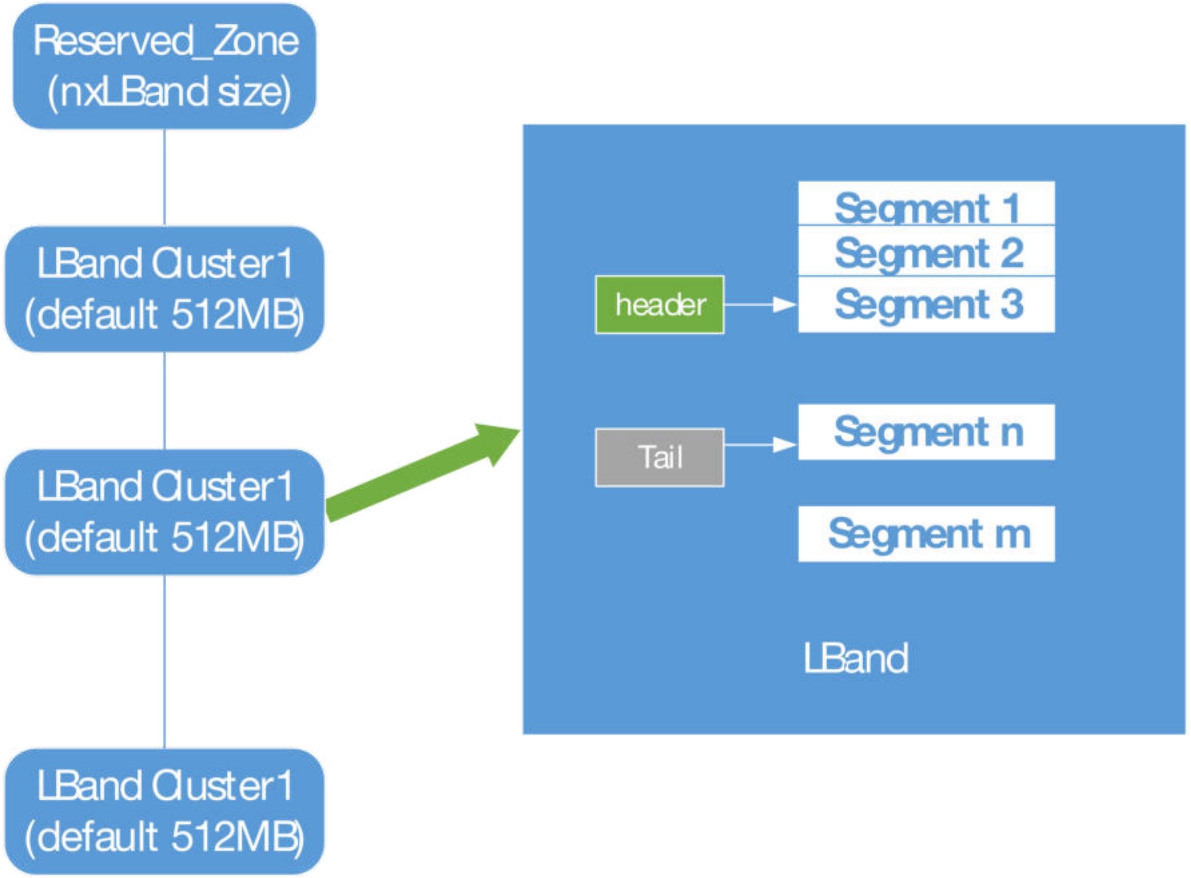


Optane as write pad, QLC as capacity store



write pad and power loss protection for QLC IU merge

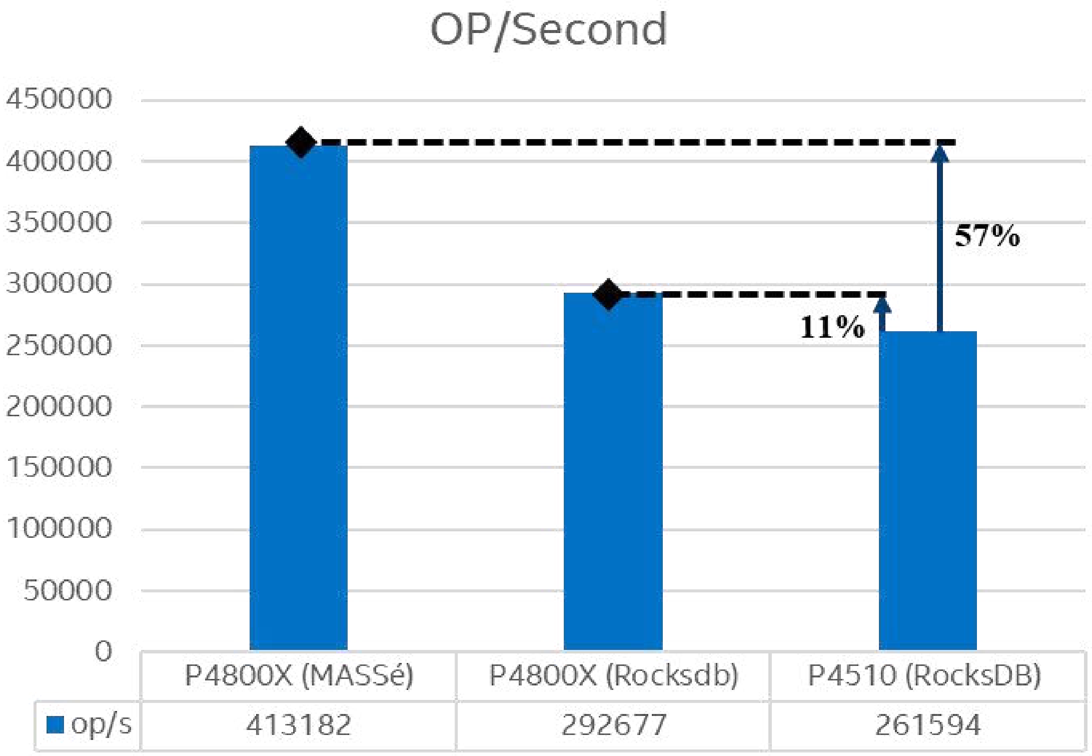
Data layout in QLC Flash



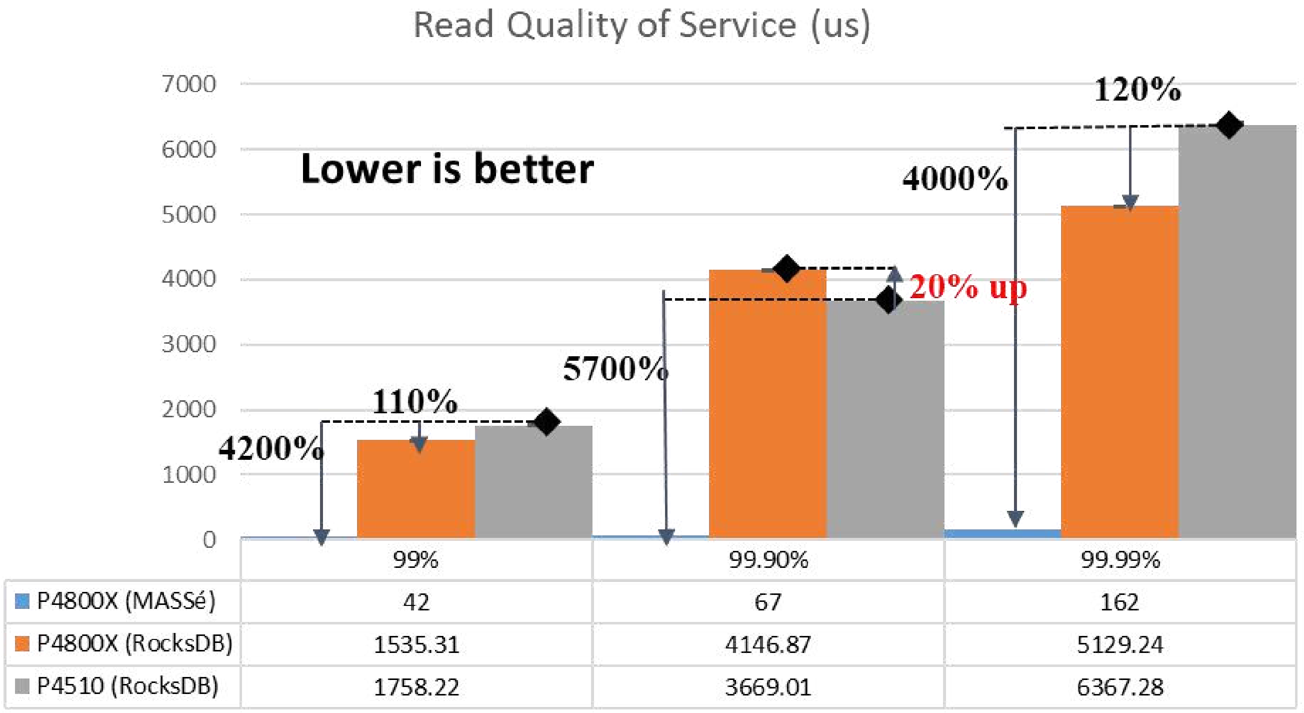
MASSé Evaluation and Proof

1. MASSé vs RocksDB (media un-aware engine) performance comparison
2. MASSé performance with different SSD media
3. MASSé case study in real customer application, Bytedance TerarkdB

MASSé vs RocksDB



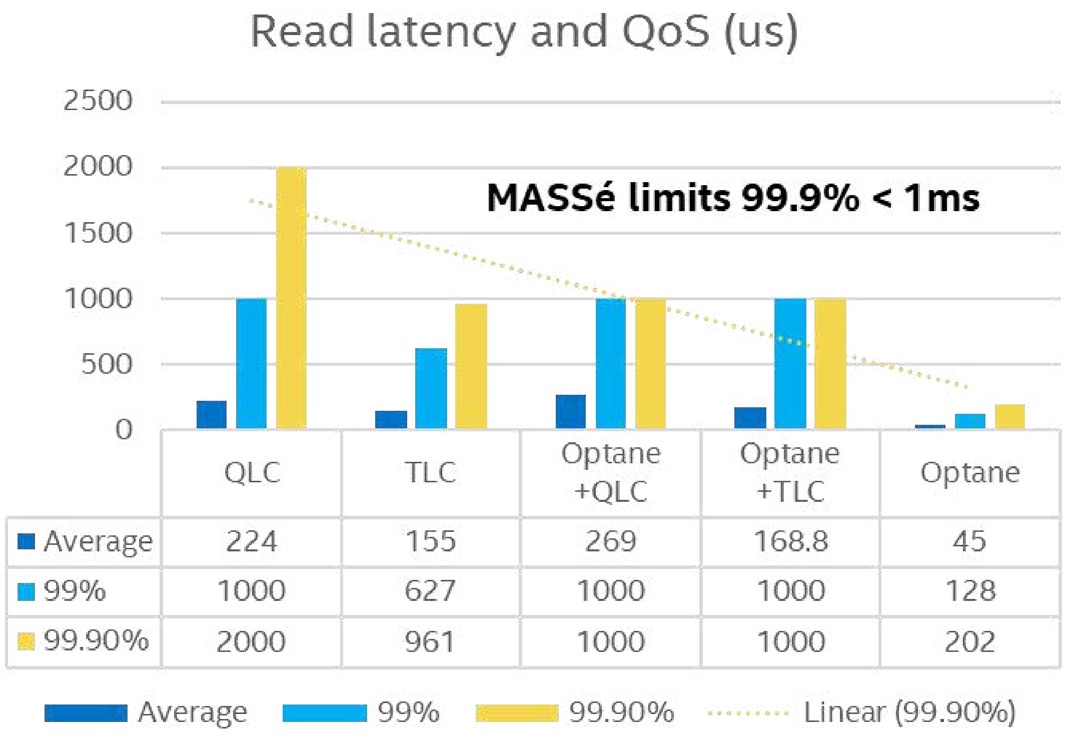
	P4800X (MASSé)	P4800X (Rocksdb)	P4510 (RocksDB)
■ op/s	413182	292677	261594



	99%	99.90%	99.99%
■ P4800X (MASSé)	42	67	162
■ P4800X (RocksDB)	1535.31	4146.87	5129.24
■ P4510 (RocksDB)	1758.22	3669.01	6367.28

Test configurations:
CPU: Intel(R) Xeon(R) Gold 6142M CPU @ 2.60GHz, Memory: 384GB, Storage: Intel® Optane™ SSD P4800X 375GB, Intel® SSD DC P4510 8TB
Workloads: Index search.
db_bench, 64threads KV(23B, 100B), 1Billion kv pairs, readwhilewriting 50/50 r/w
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

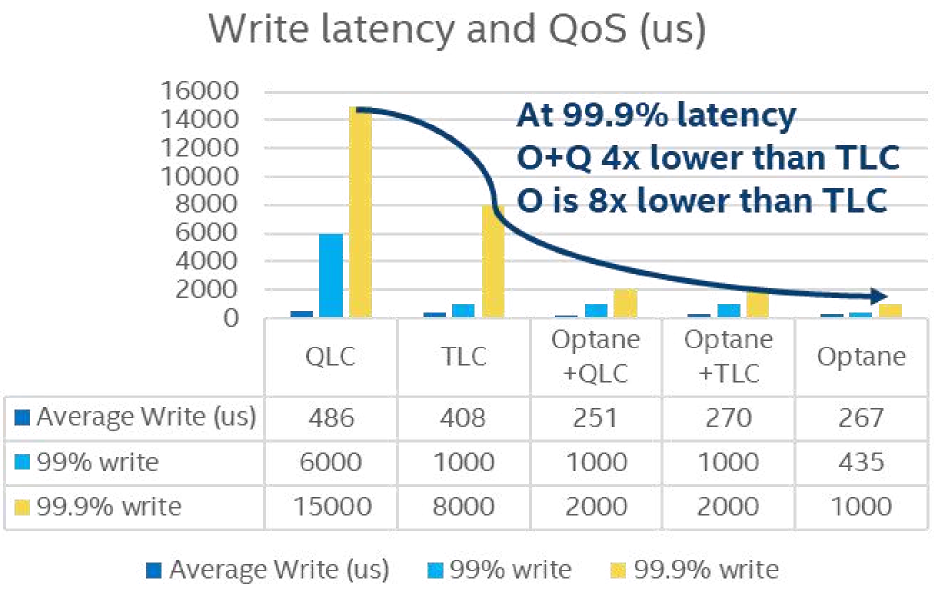
MASSé w/ different SSD media



Read latency and QoS (us)

MASSé limits 99.9% < 1ms

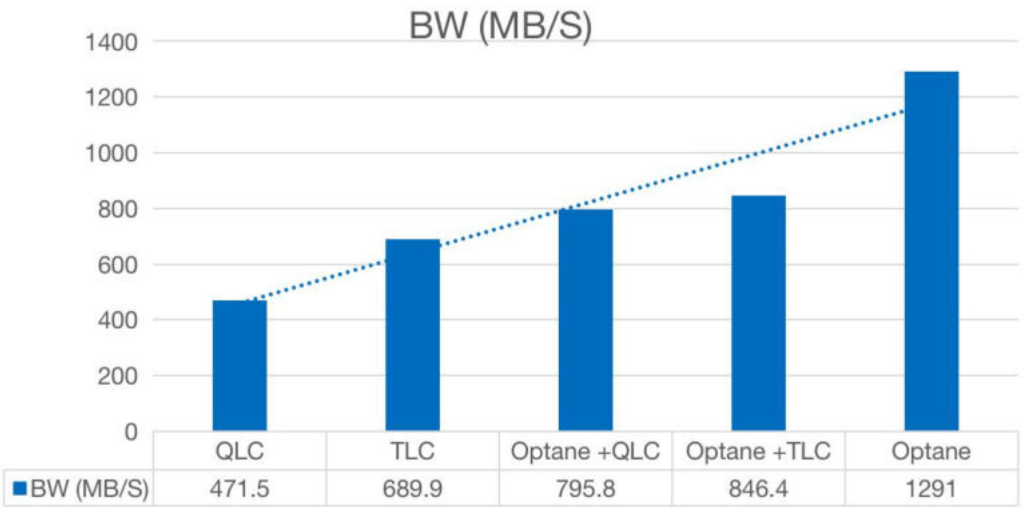
	QLC	TLC	Optane +QLC	Optane +TLC	Optane
Average	224	155	269	168.8	45
99%	1000	627	1000	1000	128
99.90%	2000	961	1000	1000	202



Write latency and QoS (us)

At 99.9% latency
O+Q 4x lower than TLC
O is 8x lower than TLC

	QLC	TLC	Optane +QLC	Optane +TLC	Optane
Average Write (us)	486	408	251	270	267
99% write	6000	1000	1000	1000	435
99.9% write	15000	8000	2000	2000	1000

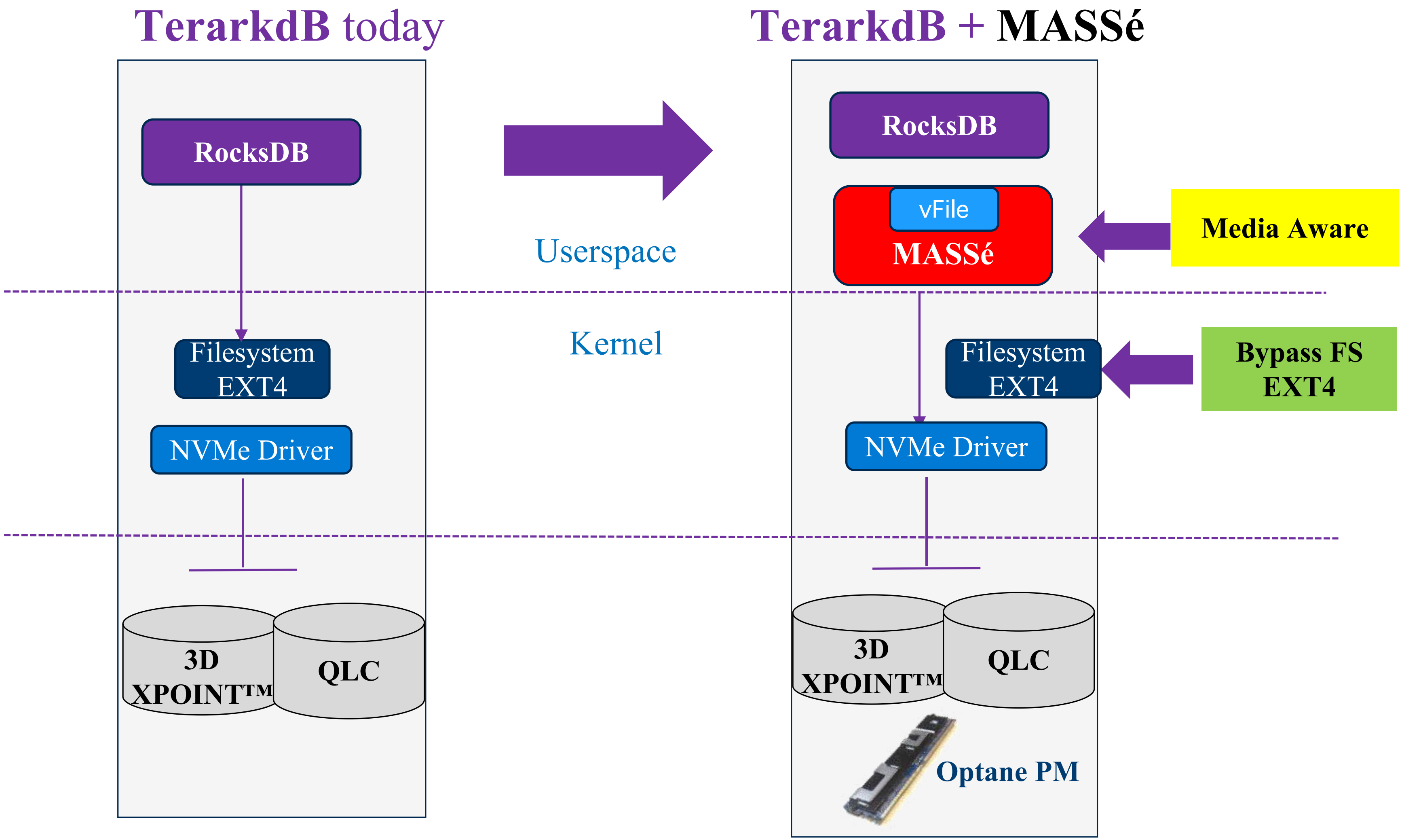


BW (MB/S)

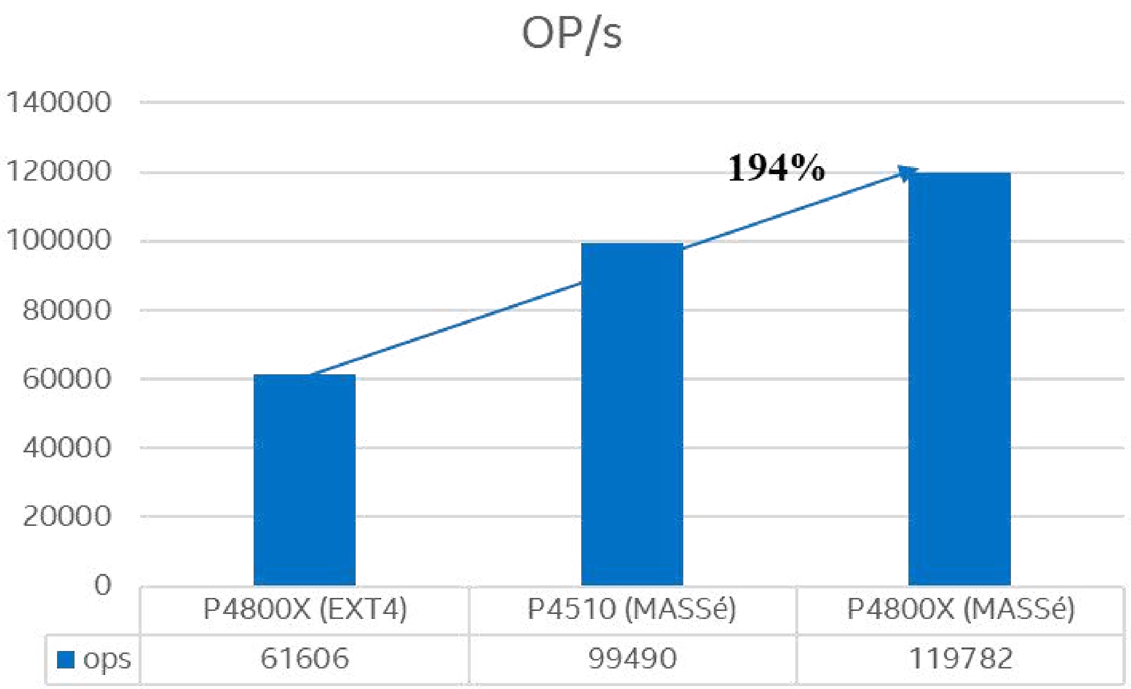
	QLC	TLC	Optane +QLC	Optane +TLC	Optane
BW (MB/S)	471.5	689.9	795.8	846.4	1291

Test configurations:
CPU: Intel(R) Xeon(R) Gold 6142M CPU @ 2.60GHz
Memory: 384GB
Storage: QLC=Intel® SSD D5-P4326, TLC= Intel® SSD DC P4510 8TB “Optane” =Intel® Optane™ SSD DC P4800X 375GB
db_bench: readwhilewriting, random 50% / 50%
64threads KV(16B, 4096B), 1Billion KV datasets
For more complete information about performance and benchmark results, visit
www.intel.com/benchmarks

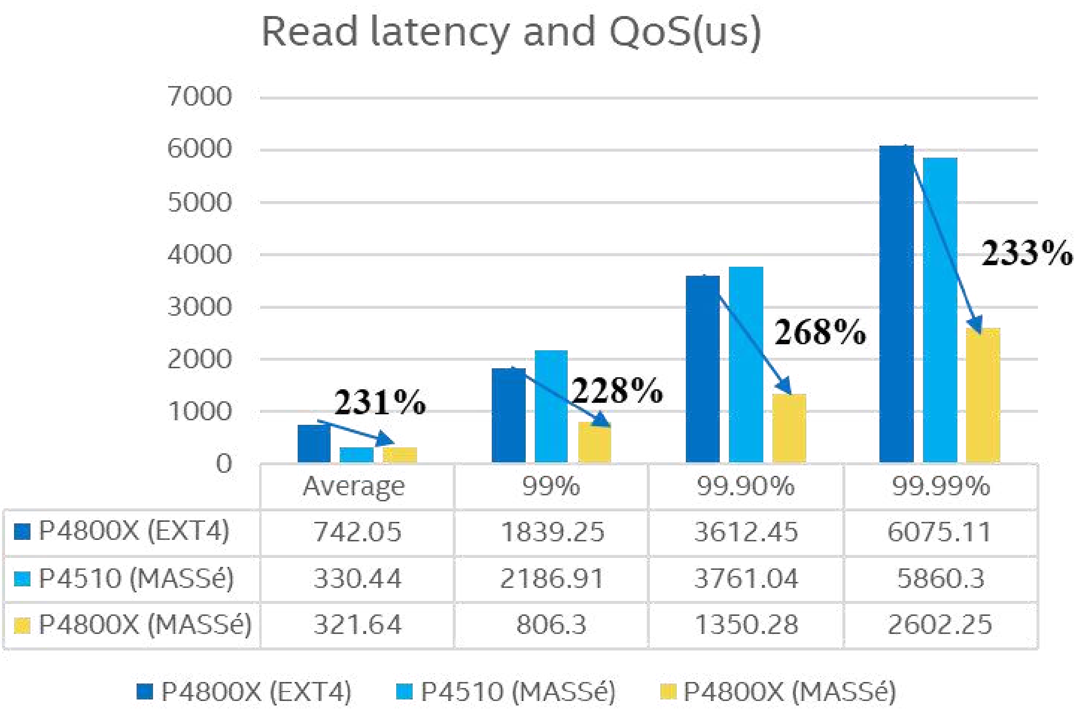
Replaces EXT4 FS



Case Study TerarkDB: **MASSé** replacement of EXT4



	P4800X (EXT4)	P4510 (MASSé)	P4800X (MASSé)
■ ops	61606	99490	119782



	Average	99%	99.90%	99.99%
■ P4800X (EXT4)	742.05	1839.25	3612.45	6075.11
■ P4510 (MASSé)	330.44	2186.91	3761.04	5860.3
■ P4800X (MASSé)	321.64	806.3	1350.28	2602.25

workloads

Key=20B, Value=400B
readrandomwriterandom 70/30
100M entries, no read cache
3.2Billion Operations

```
./db_bench --skvds=false (or true) --db=/mnt/Xdb ( or /test)--
benchmarks=readrandomwriterandom --threads=32 --readwritepercent=70 --
num=100000000 --key_size=20 --value_size=400--options_file=../skvds_options --
statistics=1 --histogram=1
```

What Comes Next

- Conclusions

  1) MASSé is a high-performance and effective storage solution that releases the maximum power of heterogeneous SSD media. It is an inclusive design that reduces application burdens and encourages investments in new storage technologies.

  2) By making the combination of Optane and QLC SSDs more effective, MASSé meets the growing demands of cloud and datacenter to improve performance while reducing cost

- Next steps

  1) Design standard MASSé lib and userspace module, standardize vFile and vBlock interfaces

  2) Design media aware RocksFS to replace RocksDB filesystems-- improve RocksDB performance especially with Optane, in general, RocksFS = abstract POSIX FS + MASSé

  3) Opensource, MASSé revision 1.0 released at private https://github.com/TeamSKVDS/skvdsmaster

  4) white paper, https://software.intel.com/content/www/us/en/develop/download/masse-a-high-performance-storage-solution.html?wapkw=masse

Please take a moment to rate this session.

Your feedback is important to us.