SMART STORAGE ENGINE FOR INTEL® 3D XPOINT™ TECHNOLOGY AND QLC 3D NAND SSDs

Jack Zhang yuan.zhang@intel.com
Cloud & Enterprise Architect
Non-Volatile Solution Group, Intel Corp.

Legal Disclaimer

Intel® names on 3D XPoint™ Technology

- Media

Intel® Optane™ Memory Media

- SSD

Intel® Optane™ SSD

- Persist Memory

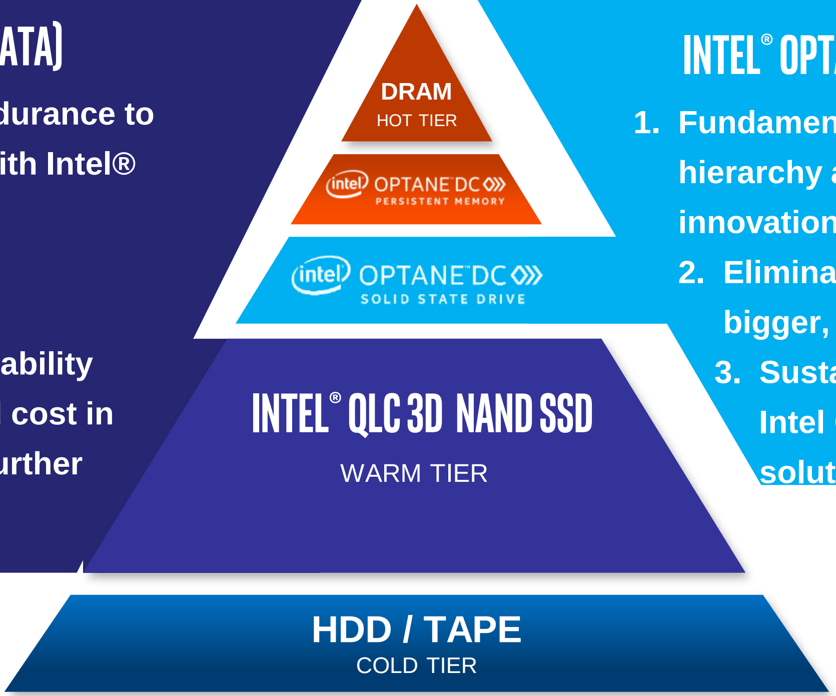
Intel® Optane™ DC Persistent Memory

Intel® Optane™ Technology and QLC Technology

Flash Memory Summit

QLC DC SSDs (CAPACITY DATA)

1. Improving Perf, QoS & Endurance to replace TLC (combining with Intel® Optane™ SSDs)
2. Accelerate warm storage innovations with better performance/capacity scalability
3. Replace HDDs for reduced cost in warm tier and push HDD further down to Cold tier

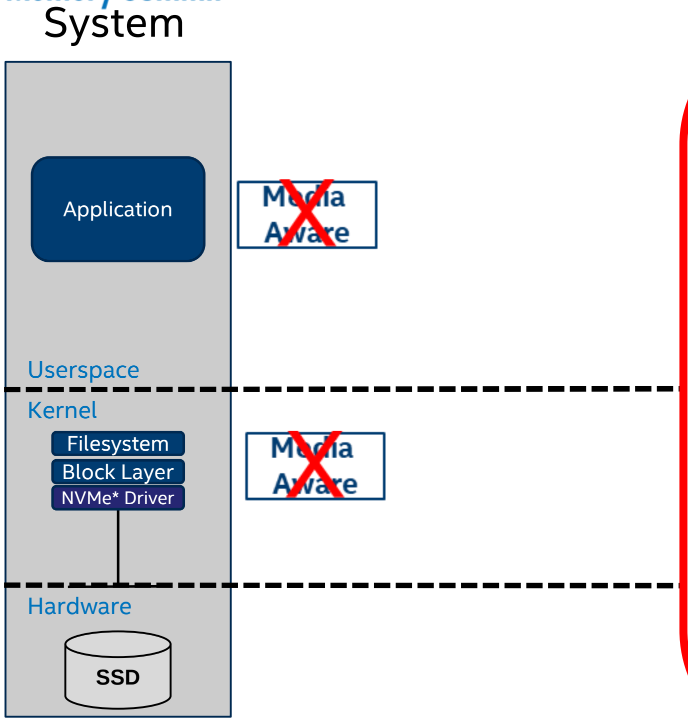


INTEL® OPTANE™ DC SSDs (WORKING DATA)

1. Fundamentally transform the storage hierarchy and accelerate arch innovations
2. Eliminate DC storage bottlenecks for bigger, more affordable data sets
3. Sustained cost competitiveness for Intel Optane™ SSDs + QLC vs. TLC solutions

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Problem Statements: Media Awareness

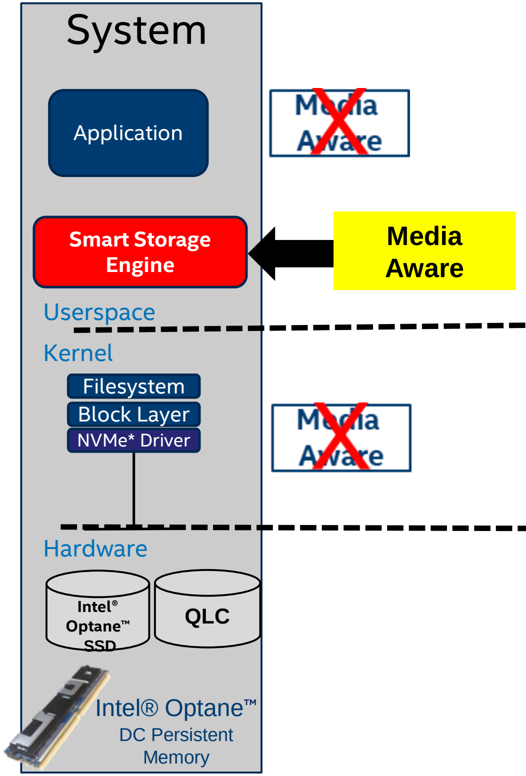


- "Awareness" needed for Intel® Optane™ SSD/Intel® Optane™ Persistent Memory
  - Write-in-place, symmetric write/read
  - Random = Sequential, efficiency on low QD
- "Awareness" needed for QLC SSDs
  - Data need to be sequentially write to QLC SSDs
  - Larger than 4K IU, 16K/64K etc
- "Awareness" needed for different storage infrastructures
  - Block (e.g., vSAN*, CEPH)
  - Object (e.g., S3, CEPH)
  - File (e.g., HDFS, CEPH)
  - KV (e.g., RocksDB, …)
- Today's typical solution like Intel Optane™ SSD for metadata/journal, QLC for data is not "media aware", could not survive heavy random write workloads

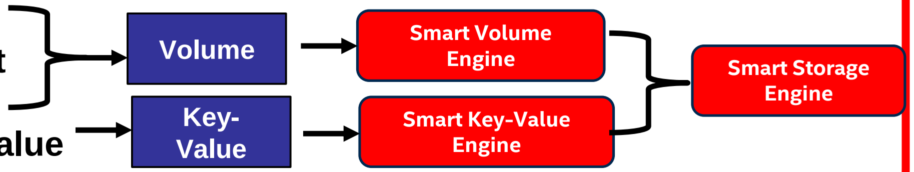
*Other names and brands may be claimed as the property of others.

Flash Memory Summit

Solution: “Media Aware” Smart Storage Engine

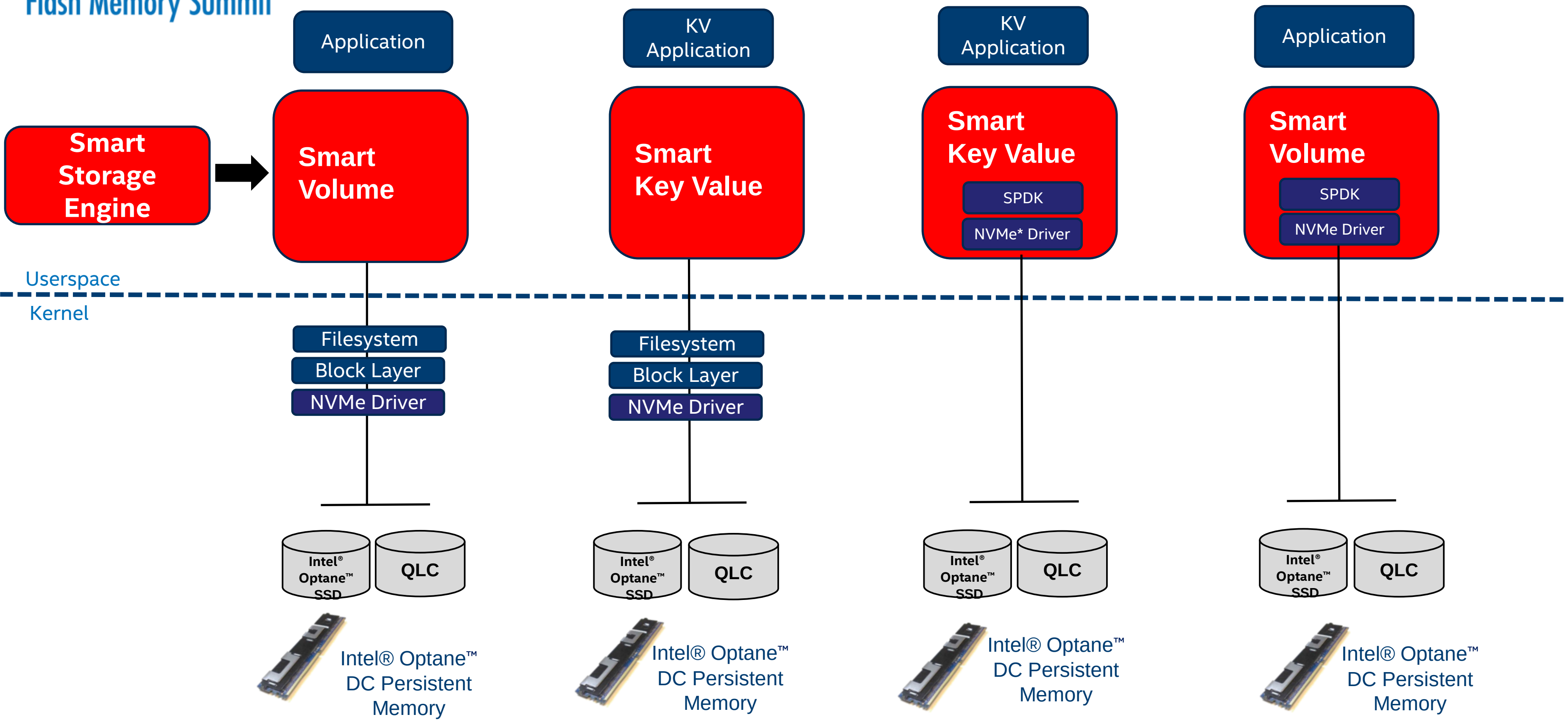


- “Awareness” for Intel® Optane™ technology
  - Design new data structure for Intel® Optane™ SSD/Intel® Optane™ Persistent Memory
  - Write-in-place, symmetric write/read, Random=Seq
- “Awareness” for QLC
  - Data structure that have data sequentially write to QLC
  - Configurable IU size from 4K to 128K
- “Awareness” for different storage infrastructures through Smart KV and volume engine
  - Block
  - Object
  - File
  - Key Value



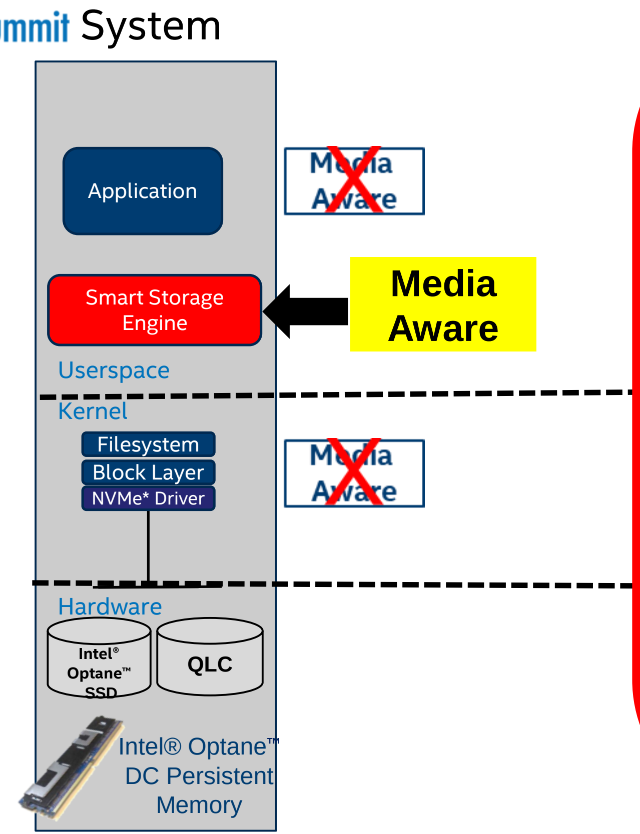
*Other names and brands may be claimed as the property of others.

Smart Storage Engine



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

Configurable Good/Better/Best



- **Good**

QLC(s) Only – good performance/best cost reduction

- **Better**

Intel® Optane™ SSD + QLC SSDs -> replacing TLC SSDs – better performance, better cost

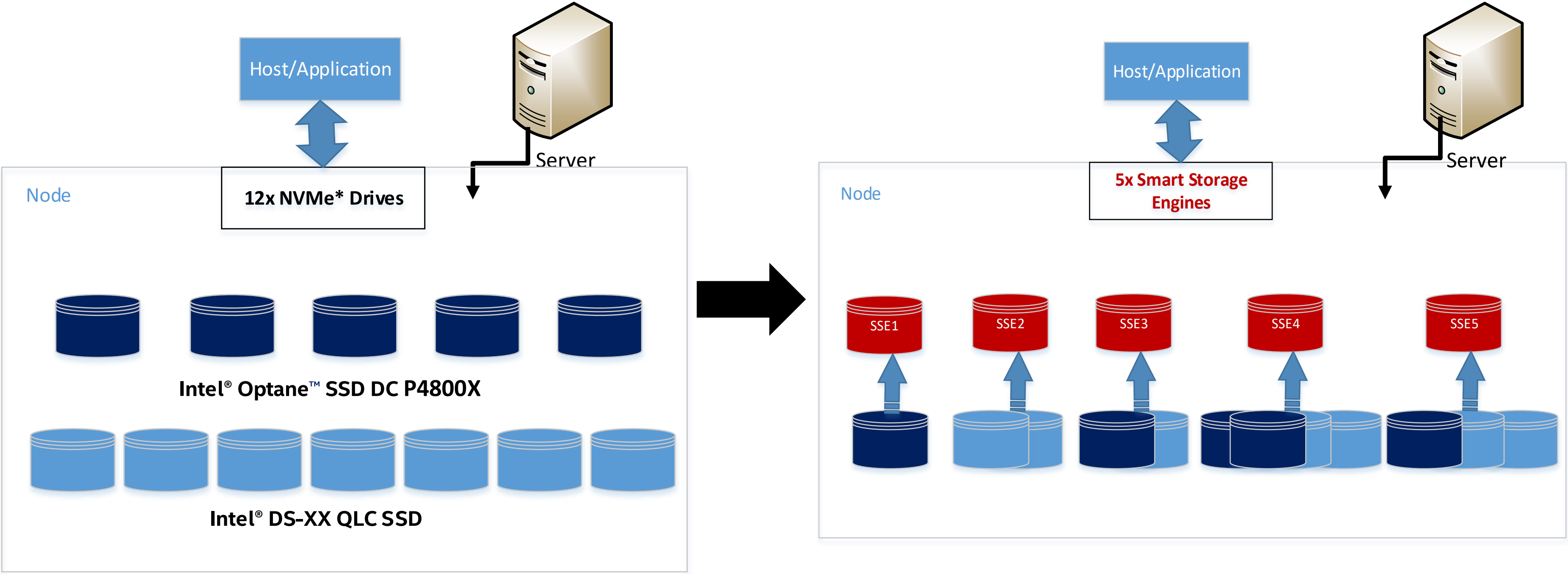
- **Best**

Intel® Optane™ SSDs Only – best performance

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

Smart Storage Engine @ system

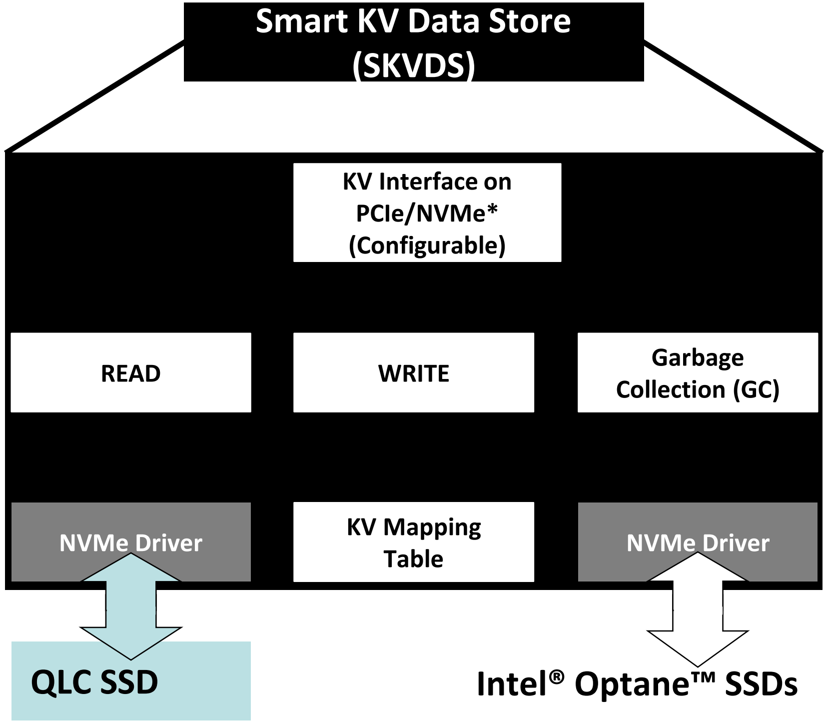
---Configurable ratio Intel® Optane™ SSD : QLC SSDs



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

Smart K-V Data Store (SKVDS) Architecture

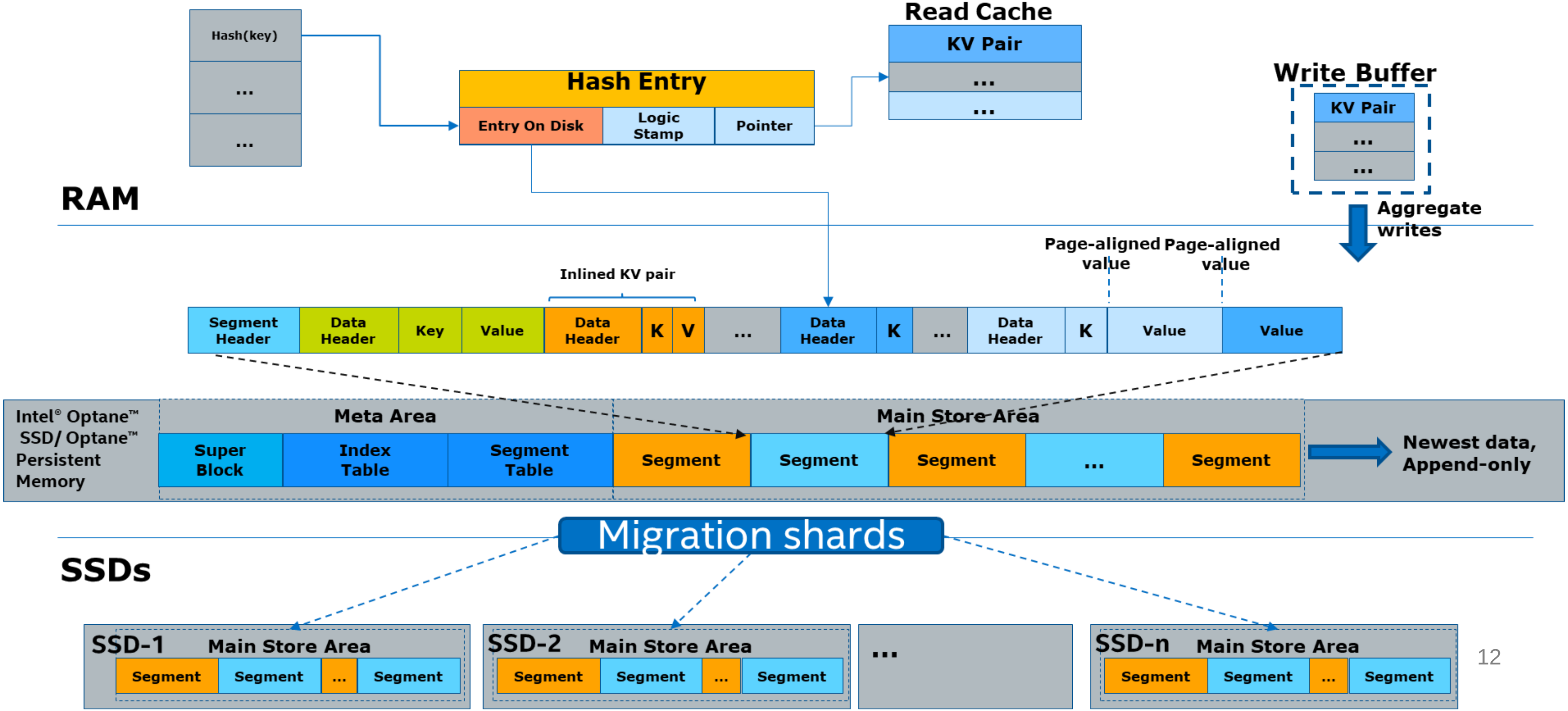
Smart Key Value Data Store Architecture



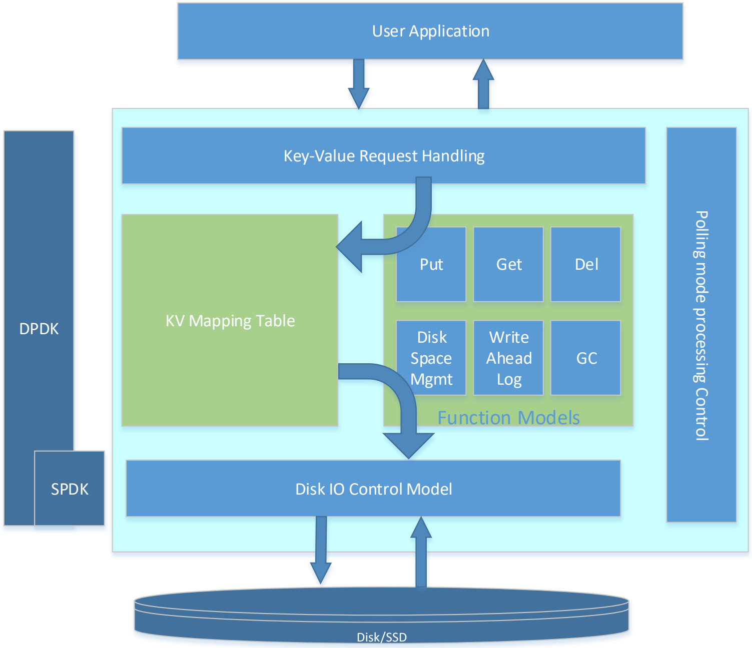
- **Key-Value APIs over PCIe*/NVMe***
  - Put, Get, Del
- Bypass kernel and filesystem
- Efficient KV mapping table
- Disk space management, WAL, GC
- Full disk log write
  - Randwrite -> seqwrite (pipelined)
- Three task threads:
  1. Read from QLC SSDs
  2. Write direct to Intel Optane SSDs or QLC SSDs (optimize for Intel Optane SSDs as write buffer)
  3. Garbage Collection on QLC SSDs (minimize QLC garbage collection + special functions, e.g., TRIM)

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

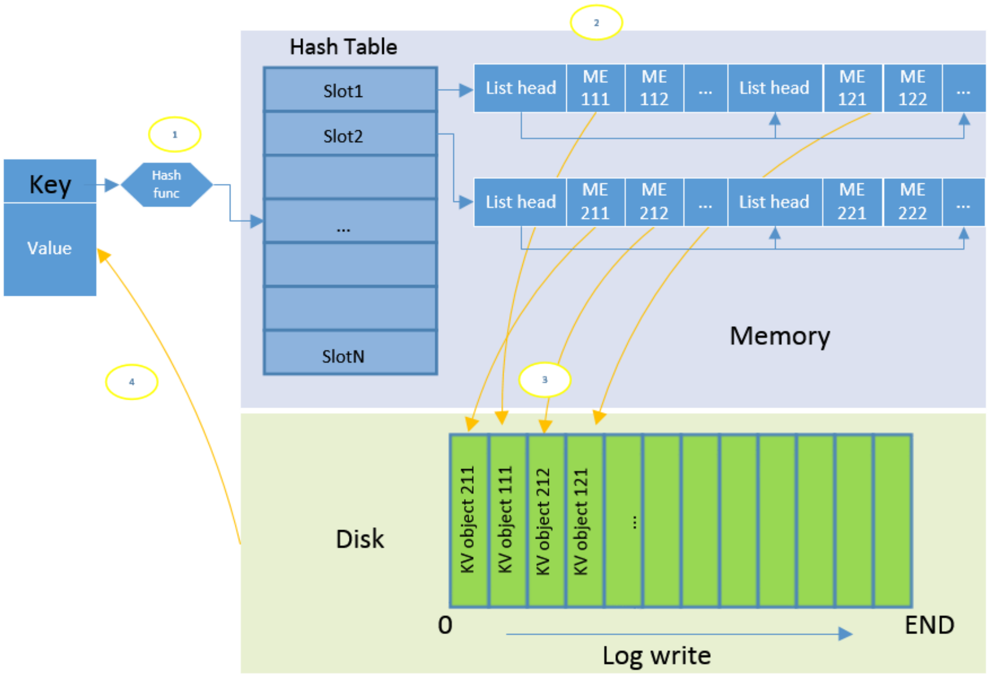
Multi-Tier Architecture



Software Architecture



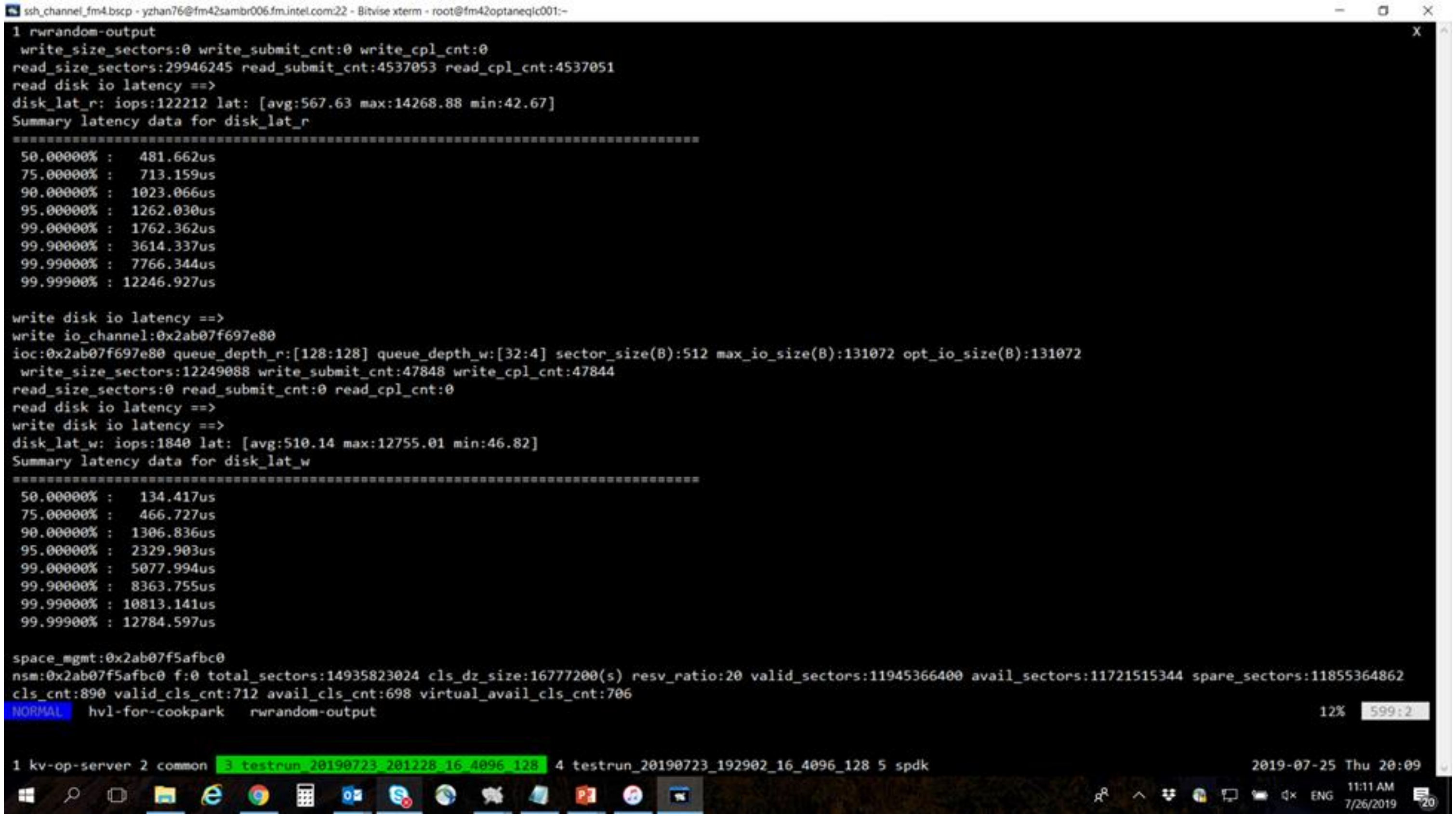
High-Level Architecture



a. Fast in-memory indexing structure with optimized per-key memory usage, along with power-loss recovery algorithms.
b. Log write to maximize write performance and minimize write amplification.

Smart KV Data Store early results

Quality of Service -- QLC SSD ONLY



Test setup:
Intel® Xeon® CPU E5-2699 v4 @ 2.20GHz
DRAM = 128GB
Storage
1x Intel® SSD D5-P4320 7.68TB

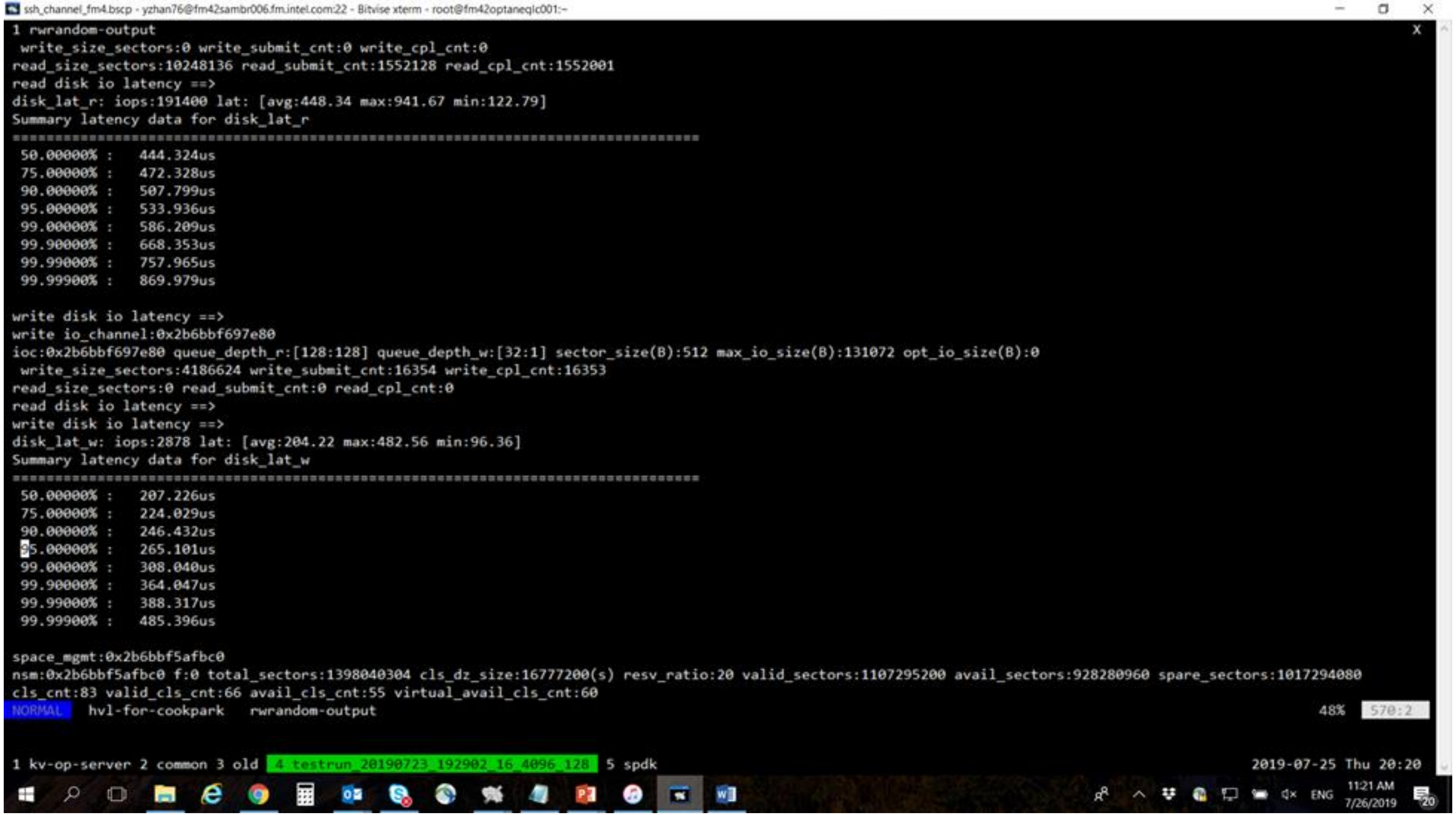
Test parameters
Key=16B,
Value=4096B
100M key pairs
Random mixed
70% read
30% write

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. See configurations in Legal Disclaimers for details. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Quality of Service – Intel® Optane™ SSD + QLC SSD

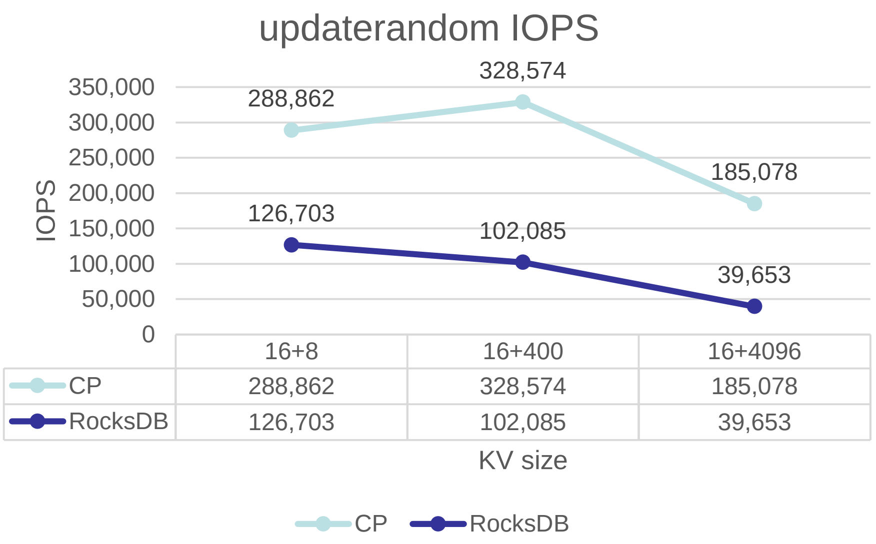
Test setup:
Intel® Xeon® CPU E5-2699 v4 @ 2.20GHz
DRAM = 128GB
Storage
1x Intel® SSD DC P4800X 375GB
1x Intel® SSD D5-P4320 7.68TB

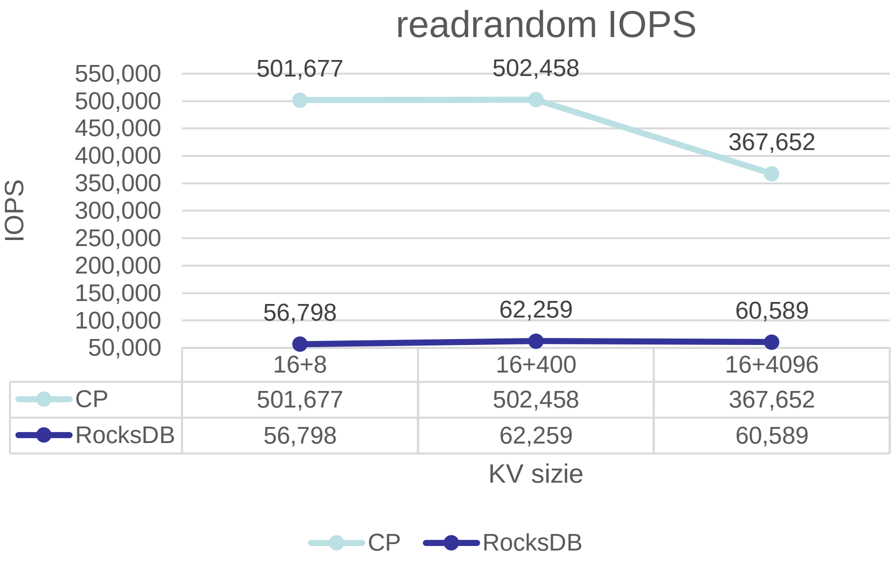
Test parameters
Key=16B, Value=4096B
100M key pairs
Random mixed
70% read
30% write



Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. See configurations in Legal Disclaimers for details. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

SKVDS vs RockDB





Test Configuration	
Server	DP
CPU	Xeon 2699v4 2.2GHz 22cores x2
Memory	64GB x2
SSD	P3700 2TB FW: 8DV101H0
OS	CentOS 7.4
Kernel	3.10.0-693.el7.x86_64
DPDK	17.08
SPDK	17.10
RocksDB	spdk-v5.6.1

Flash Memory Summit

Thank you