



Flash Memory Summit

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

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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.

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.

Intel, Intel Optane, Xeon, and others are trademarks of Intel Corporation in the U.S. and/or other countries.

© Intel Corporation.

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



Flash Memory Summit

Intel® names on 3D XPoint™ Technology

- Media

Intel® Optane™ Memory Media

- SSD

Intel® Optane™ SSD

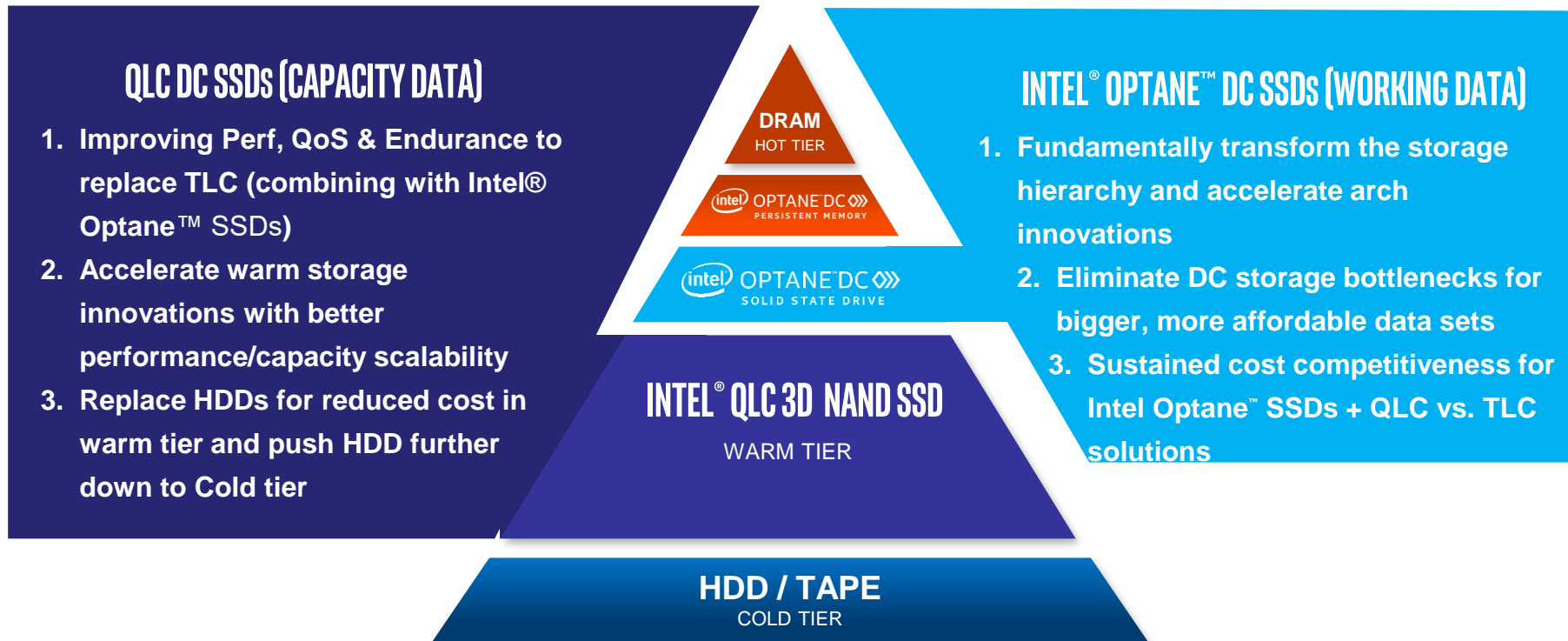
- Persist Memory

Intel® Optane™ DC Persistent Memory

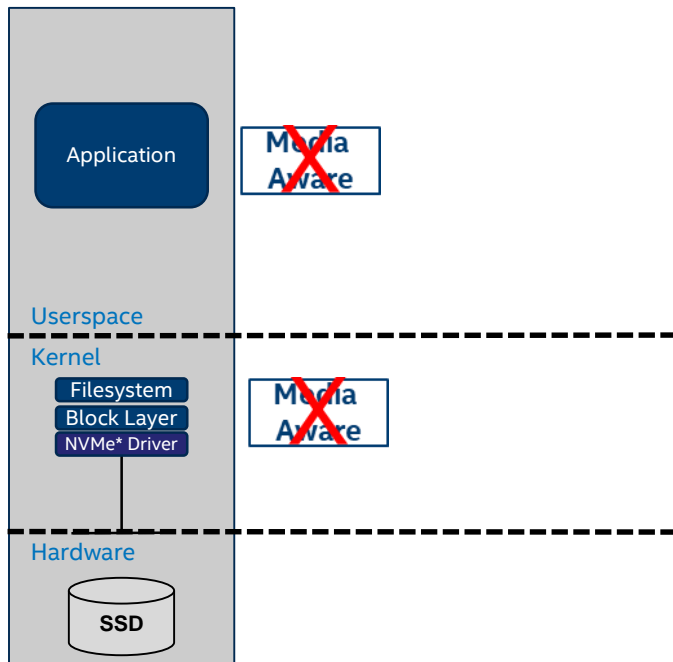


Intel® Optane™ Technology and QLC Technology

Flash Memory Summit



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.



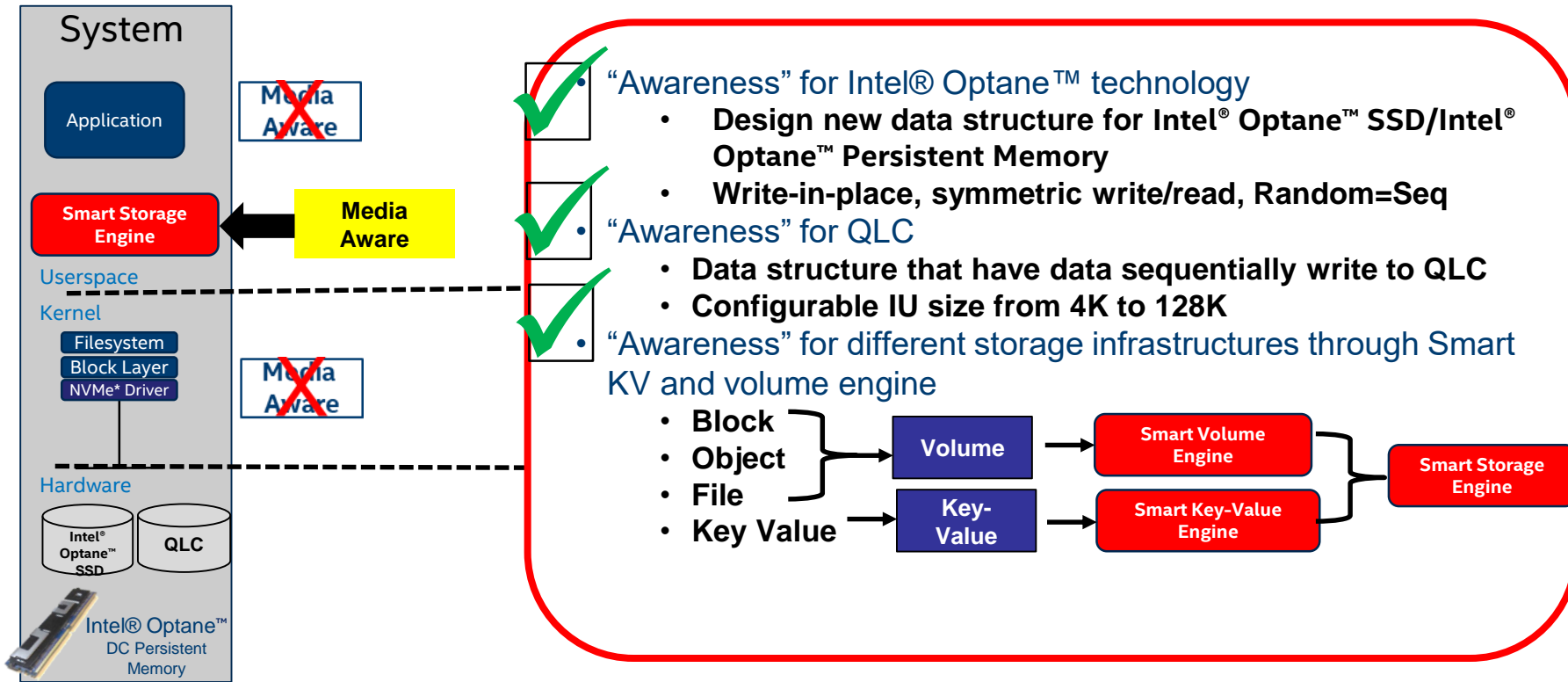
Problem Statements: Media Awareness

- “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



Solution: "Media Aware" Smart Storage Engine

Flash Memory Summit

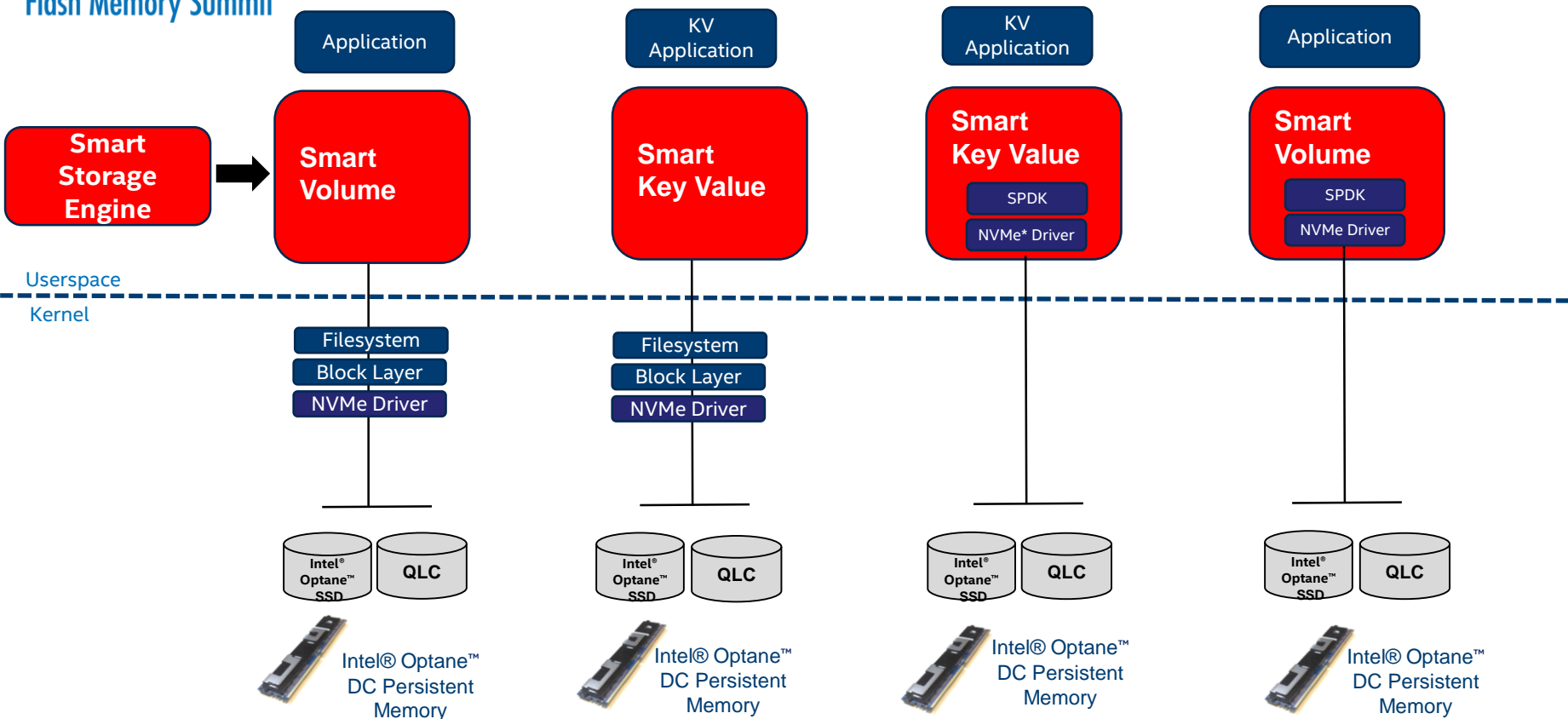


*Other names and brands may be claimed as the property of others.



Smart Storage Engine

Flash Memory Summit

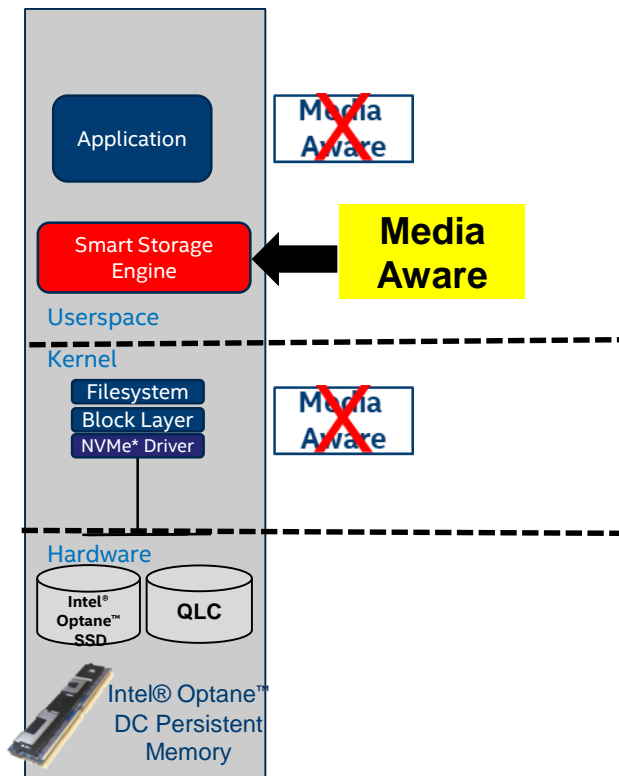


*Other names and brands may be claimed as the property of others.



Configurable Good/Better/Best

Flash Memory Summit System



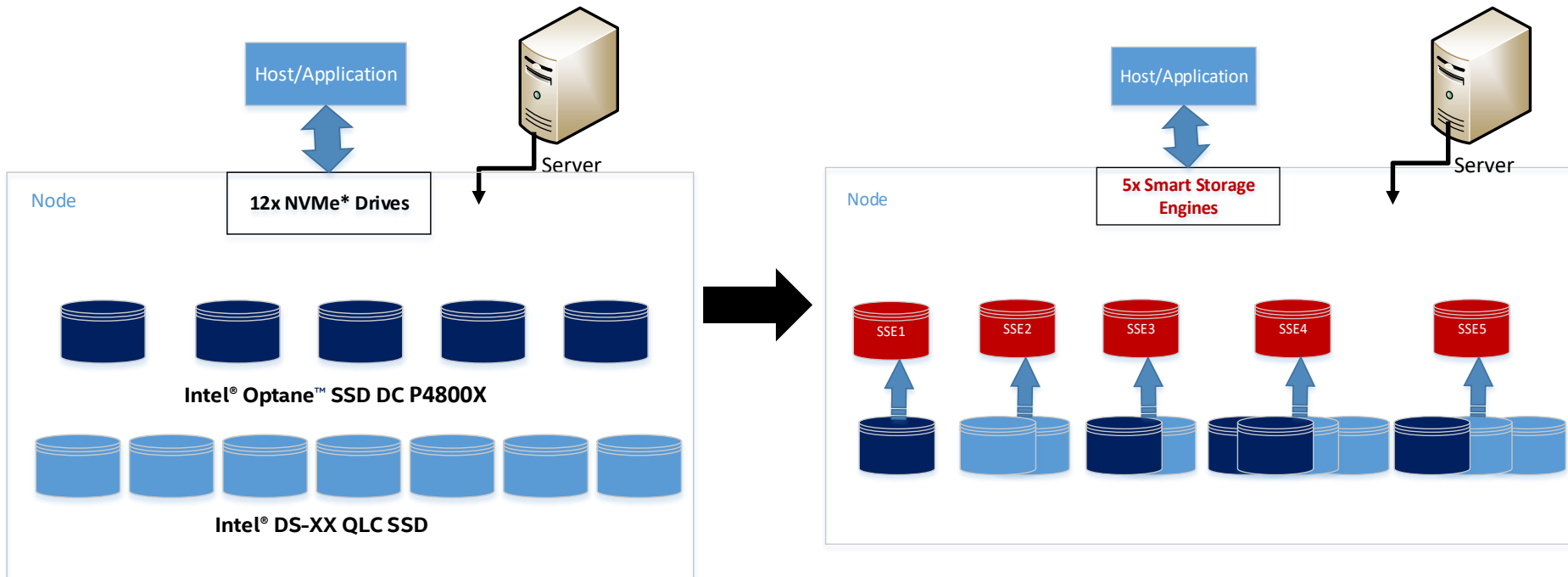
- **Good**
QLC(s) Only – good performance/best cost reduction
- **Better**
Intel® Optane™ SSD + QLC SSDs -
> replacing TLC SSDs
– better performance, better cost
- **Best**
Intel® Optane™ SSDs Only – best performance



Flash Memory Summit

Smart Storage Engine @ system

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



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

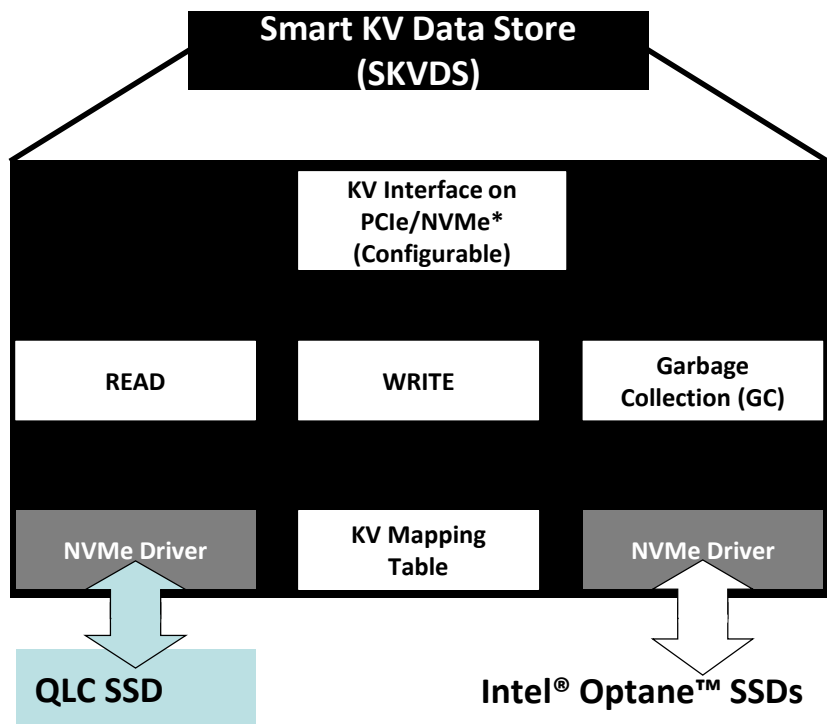


Flash Memory Summit

Smart K-V Data Store (SKVDS) Architecture



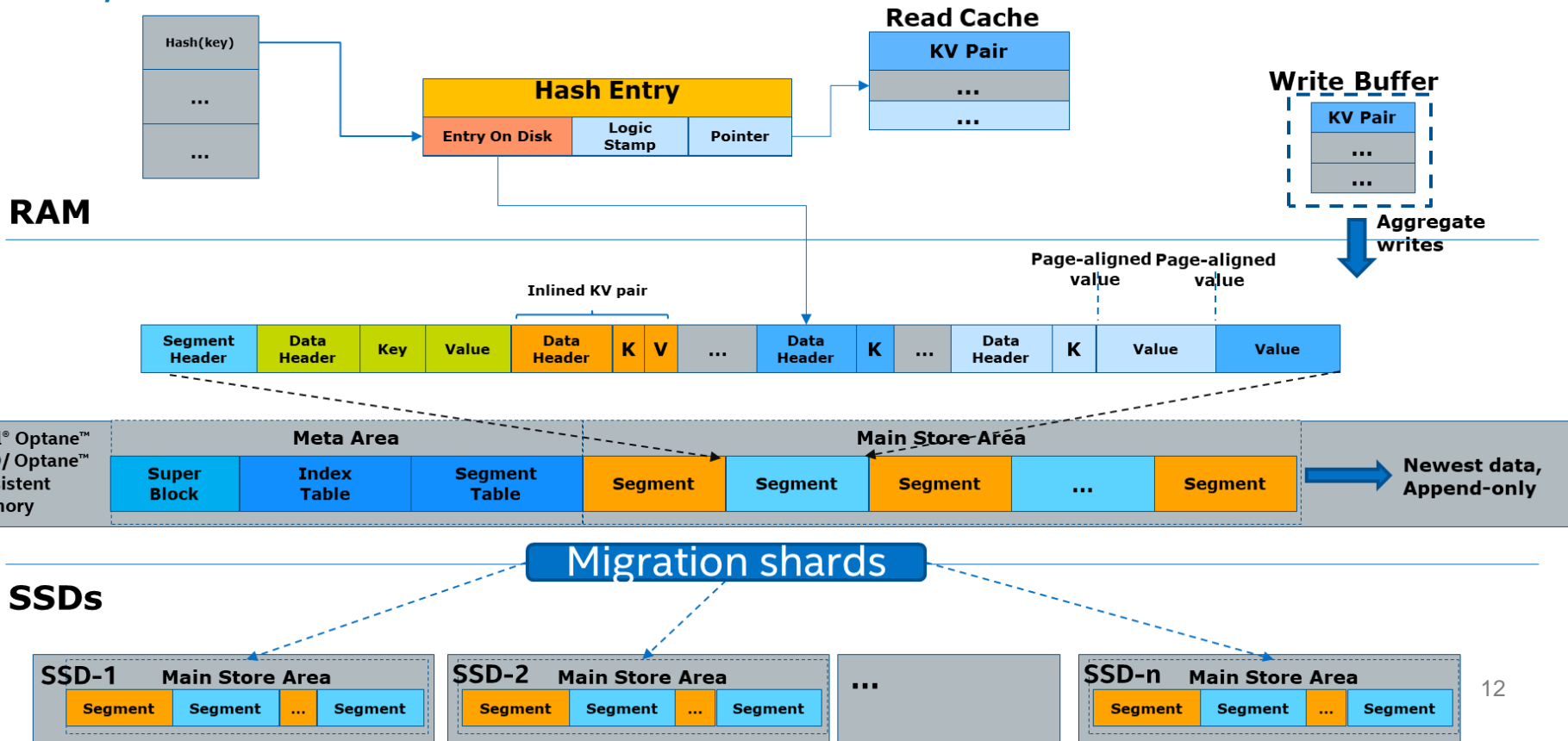
Smart Key Value Data Store Architecture



- **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)

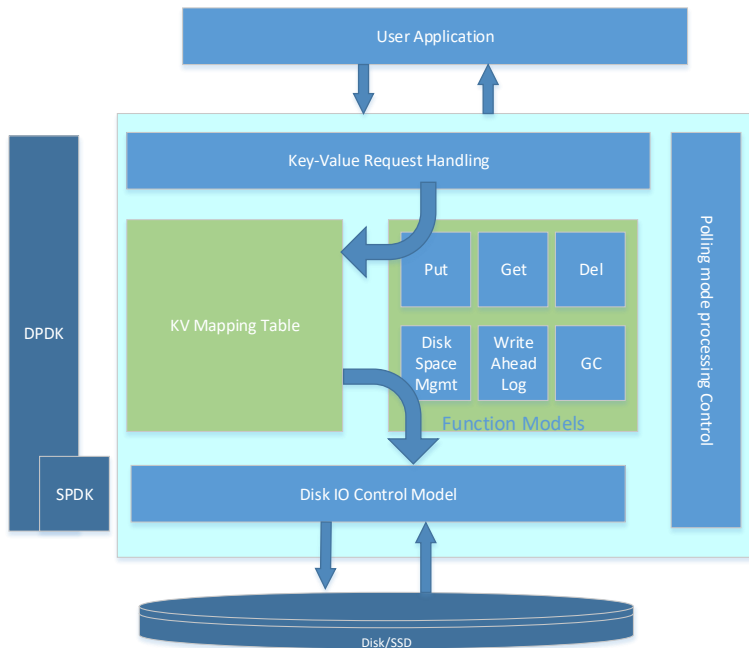


Multi-Tier Architecture





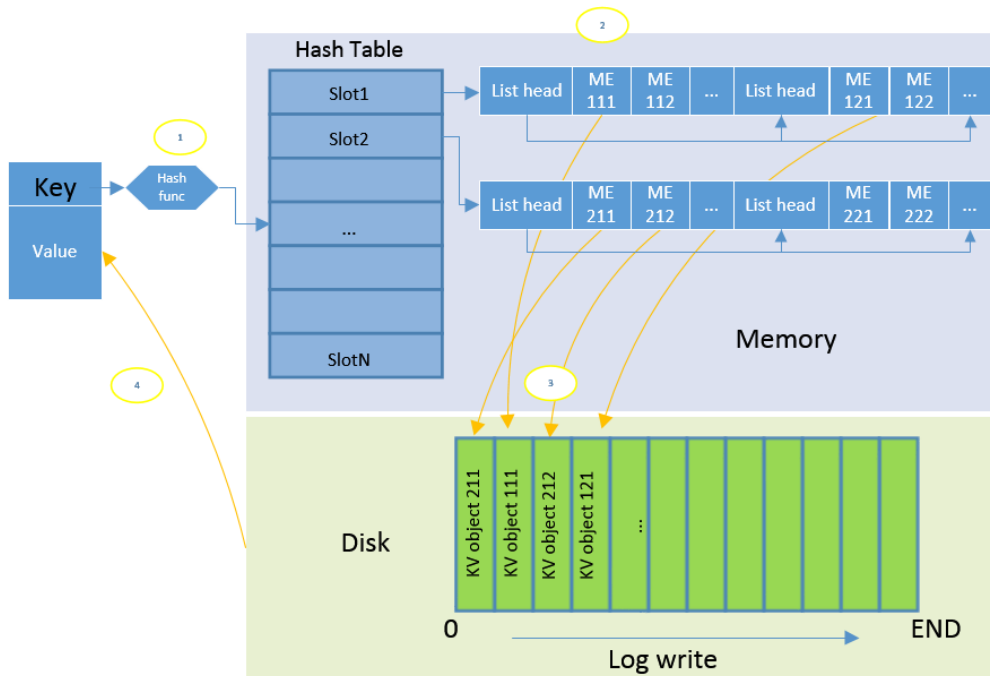
Software Architecture





High-Level Architecture

Flash Me!



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



Flash Memory Summit

Smart KV Data Store early results



Flash Memory Summit

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

```
ssh_channel_fm4.bscp - yzhan76@fm42sambr006.fm.intel.com:22 - Bitvise xterm - root@fm42optaneq001-
1 rwrandom-output
write_size_sectors:0 write_submit_cnt:0 write_cpl_cnt:0
read_size_sectors:29946245 read_submit_cnt:4537053 read_cpl_cnt:4537051
read disk io latency ==>
disk_lat_r: iops:122212 lat: [avg:567.63 max:14268.88 min:42.67]
Summary latency data for disk_lat_r
=====
50.00000% : 481.662us
75.00000% : 713.159us
90.00000% : 1023.066us
95.00000% : 1262.030us
99.00000% : 1762.362us
99.90000% : 3614.337us
99.99000% : 7766.344us
99.99900% : 12246.927us

write disk io latency ==>
write_io_channel:0x2ab07f697e80
ioc:0x2ab07f697e80 queue_depth_r:[128:128] queue_depth_w:[32:4] sector_size(B):512 max_io_size(B):131072 opt_io_size(B):131072
write_size_sectors:12249088 write_submit_cnt:47848 write_cpl_cnt:47844
read_size_sectors:0 read_submit_cnt:0 read_cpl_cnt:0
read disk io latency ==>
write disk io latency ==>
disk_lat_w: iops:1840 lat: [avg:510.14 max:12755.01 min:46.82]
Summary latency data for disk_lat_w
=====
50.00000% : 134.417us
75.00000% : 466.727us
90.00000% : 1306.836us
95.00000% : 2329.903us
99.00000% : 5077.994us
99.90000% : 8363.755us
99.99000% : 10813.141us
99.99900% : 12784.597us

space_mgmt:0x2ab07f5afbc0
nsm:0x2ab07f5afbc0 f:0 total_sectors:14935823024 cls_dz_size:16777200(s) resv_ratio:20 valid_sectors:11945366400 avail_sectors:11721515344 spare_sectors:11855364862
cls_cnt:890 valid_cls_cnt:712 avail_cls_cnt:698 virtual_avail_cls_cnt:706
NORMAL hvl-for-cookpark rwrandom-output 12% 599:2

1 kv-op-server 2 common 3 testrun_20190725_201228_16_4096_128 4 testrun_20190723_192902_16_4096_128 5 spdk
2019-07-25 Thu 20:09
11:11 AM
7/26/2019
```

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.



Flash Memory Summit

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

```

ssh_channel_fm4.bscp - yzhan76@fm42sambr006.fm.intel.com:22 - Bitwise xterm - root@fm42optaneqlc001:~
1 rwrandom-output
write_size_sectors:0 write_submit_cnt:0 write_cpl_cnt:0
read_size_sectors:10248136 read_submit_cnt:1552128 read_cpl_cnt:1552001
read disk io latency ==>
disk_lat_r: iops:191400 lat: [avg:448.34 max:941.67 min:122.79]
Summary latency data for disk_lat_r
=====
50.00000% : 444.324us
75.00000% : 472.328us
90.00000% : 507.799us
95.00000% : 533.936us
99.00000% : 586.209us
99.90000% : 668.353us
99.99000% : 757.965us
99.99900% : 869.979us

write disk io latency ==>
write io channel:0x2b6bbf697e80
ioc:0x2b6bbf697e80 queue_depth_r:[128:128] queue_depth_w:[32:1] sector_size(B):512 max_io_size(B):131072 opt_io_size(B):0
write_size_sectors:4186624 write_submit_cnt:16354 write_cpl_cnt:16353
read_size_sectors:0 read_submit_cnt:0 read_cpl_cnt:0
read disk io latency ==>
write disk io latency ==>
disk_lat_w: iops:2878 lat: [avg:204.22 max:482.56 min:96.36]
Summary latency data for disk_lat_w
=====
50.00000% : 207.226us
75.00000% : 224.029us
90.00000% : 246.432us
95.00000% : 265.101us
99.00000% : 308.040us
99.90000% : 364.047us
99.99000% : 388.317us
99.99900% : 485.396us

space_mgmt:0x2b6bbf5afbc0
nsm:0x2b6bbf5afbc0 f:0 total_sectors:1398040304 cls_dz_size:16777200(s) resv_ratio:20 valid_sectors:1107295200 avail_sectors:928280960 spare_sectors:1017294080
cls_cnt:83 valid_cls_cnt:66 avail_cls_cnt:55 virtual_avail_cls_cnt:60
NORMAL hvl-for-cookpark rwrandom-output 48% 570.2

1 kv-op-server 2 common 3 old 4 testrun 20190723 192902 16 4096 128 5 spd

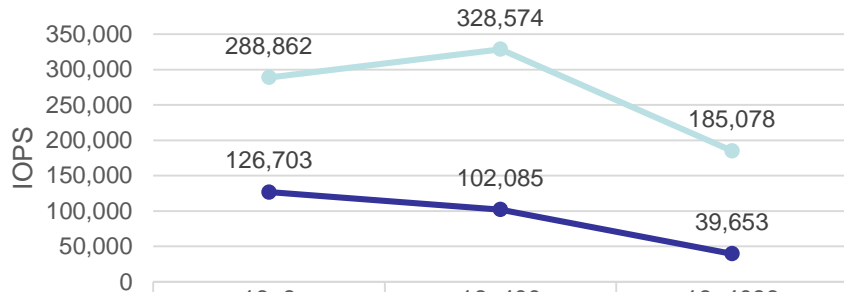
2019-07-25 Thu 20:20
11:21 AM
7/26/2019
  
```

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

updaterandom IOPS

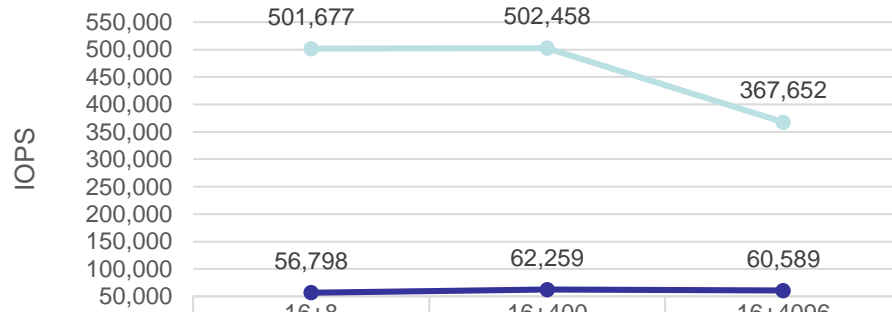


	16+8	16+400	16+4096
CP	288,862	328,574	185,078
RocksDB	126,703	102,085	39,653

KV size

CP RocksDB

readrandom IOPS



	16+8	16+400	16+4096
CP	501,677	502,458	367,652
RocksDB	56,798	62,259	60,589

KV size

CP RocksDB

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

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.



Flash Memory Summit

Thank you