SDC 19

September 23-26, 2019
Santa Clara, CA

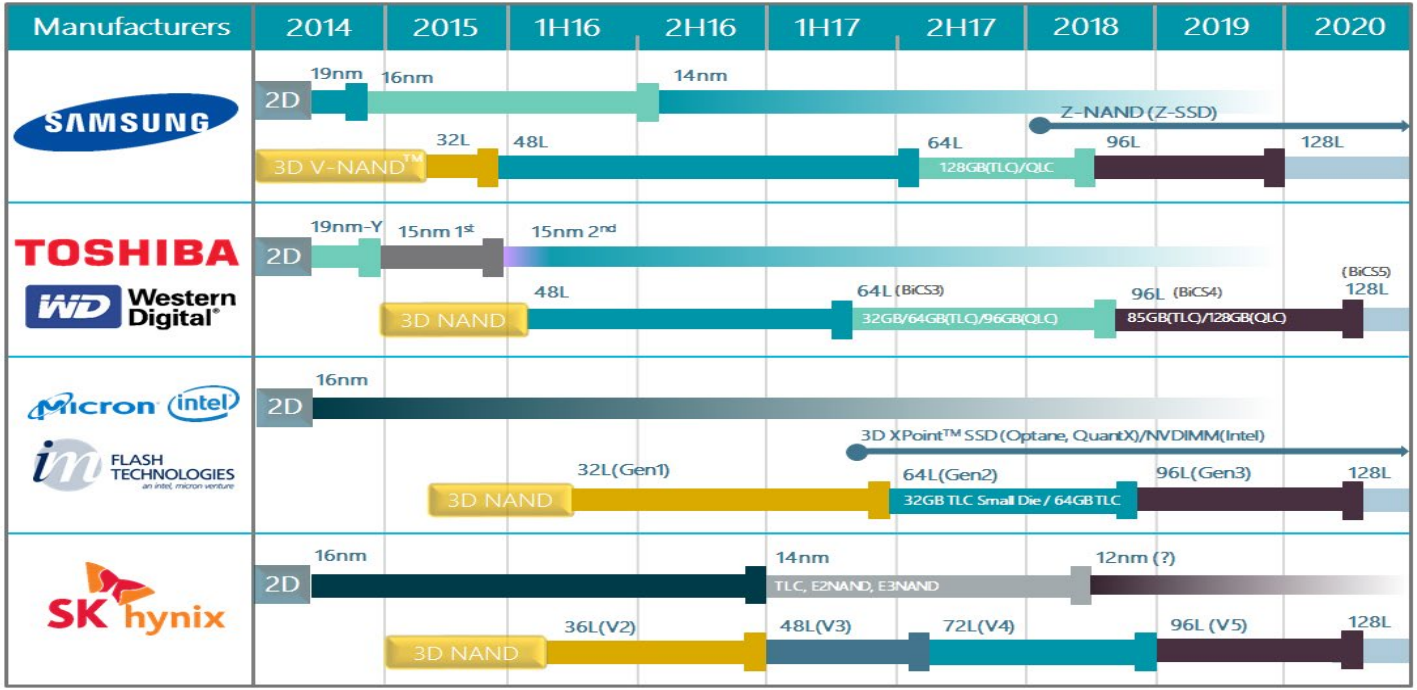
Fast Resilvering in High Capacity All Flash Systems

Shailendra Tripathi
Fellow, Filesystem Development
Western Digital

Agenda

- Flash Capacity Trend
- Performance Trend
- Data Redundancy Models
- Resilvering Mechanism
- Evaluation

Flash Capacity Trends



Tech Insights

https://www.forbes.com/sites/tomcoughlin/2018/01/02/digital-storage-projections-for-2018-part-2/#5d0e29303b07

NVMe TLC Performance Stats

Ultrastar® DC SN200

DATA SHEET | **NVMe™ DATA CENTER SSDs**

Specifications

Configuration	HH-HL Add-in Card (AIC)		U.2 Drive	
Interface	PCIe 3.0 x8 NVMe 1.2		PCIe 3.0 x4 or 2×2 NVMe 1.2	
Form Factor	HH-HL add-in card		U.2 2.5-inch drive	
Capacity[1]	6.4TB / 3.2TB / 1.6TB	7.68TB / 3.84TB / 1.92TB	6.4TB / 3.2TB / 1.6TB / 800GB	7.68TB / 3.84TB / 1.92TB / 960GB
Endurance (Drive Writes per Day)[2]	3	1	3	1
Flash Memory Technology	15nm MLC NAND			
Performance[3]				
Sequential Read (max MiB/s, 128KiB)	6,170		3,350	
Sequential Write (max MiB/s, 128KiB)	2,200		2,100	
Random Read (max IOPS, 4KiB)	1,200,000		835,000	
Random Write (max IOPS, 4KiB)	200,000	75,000	200,000	75,000
Mixed Random Read/Write (max IOPS 70%R/30%W, 4KiB)	580,000	240,000	550,000	240,000
Write Latency[4] (µs)	20		20	
Reliability				
Uncorrectable Bit Error Rate (UBER)	< 1 in 10^{17}		< 1 in 10^{17}	
MTBF[5]	2M hours		2M hours	

https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/product/data-center-drives/ultrastar-dc-ha200-series/data-sheet-ultrastar-dc-sn200.pdf

QLC Data Perf Spec

Features At-a-Glance	
Model	Intel® SSD D5-P4326
Capacity and Form Factor	U.2 15mm: 15.36TB E1.L 9.5mm: 15.36TB, 30.72TB (Available Q3 2019) E1.L 18mm: 15.36TB, 30.72TB (Available Q3 2019)
Interface	PCIe* Gen3.1 x 4
Media	64-layer, QLC 3D NAND
Performance	Sequential R/W up to 3,200/1,600 MB/s
	Random R/W up to 580K/15K IOPS[6]
Endurance	0.18 DWPD (Random) / 0.9 DWPD (Sequential) DWPD = drive writes per day
Reliability	AFR: ≤ 0.44% UBER: 10^{17} bits read MTBF: 2 million hours
Power	Active: 20W Idle: 5W Enhanced power-loss data protection
Operating Temperature	0° C to 70° C
Warranty	5-year limited warranty

https://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/d5-p4326-series-brief.pdf

Failure and Rebuild

- Increasingly Dense Capacity per U
- QLC stepping up TLC soon
- QLC stepping up HDDs likely
- AFR not trending down
- Increasing Blast Radius

Fast Resilvering Key Differentiator

Resilver Theoretical Times

- 300 MB/s Random at 4K
- 960 MB/s Random at 128KB
- 2GB/s Seq at 128KB – by spec
- About 8 min / TB to 55 min / TB
- 16 TB Drive – 2h15m to 15h

Sequential at 128KB best model

RAID Mirror / 1- Replica

- RAID Mirror or 1 Replica
- Surviving Drive – Unpredictable impact
- Full Resilvering – Crowds out client I/Os
- Throttling needed
- Increased Resilver Times
- Client Read Latency much higher impact

Triple Mirror / 2-Replica

- Surviving Drives – Up to 50% impact
- Full Resilvering – May crowd out client I/Os
- Throttling still needed
- Potentially higher resilver time
- Client Read Latency impacted

Erasure Coding

- More Drives, Amortized I/Os
- Network performance
- Write Latency Impact
- Manageable if not across network
- Computational Overhead – Higher
- Throttling potentially needed

Reasonable Choice vs Space Tradeoff

RAID- 5/6/7

- More Drives, Amortized I/Os
- Write Latency Amortized
- Network Latency Impact
- Manageable if not across network
- Computationally manageable
- Throttling potentially needed

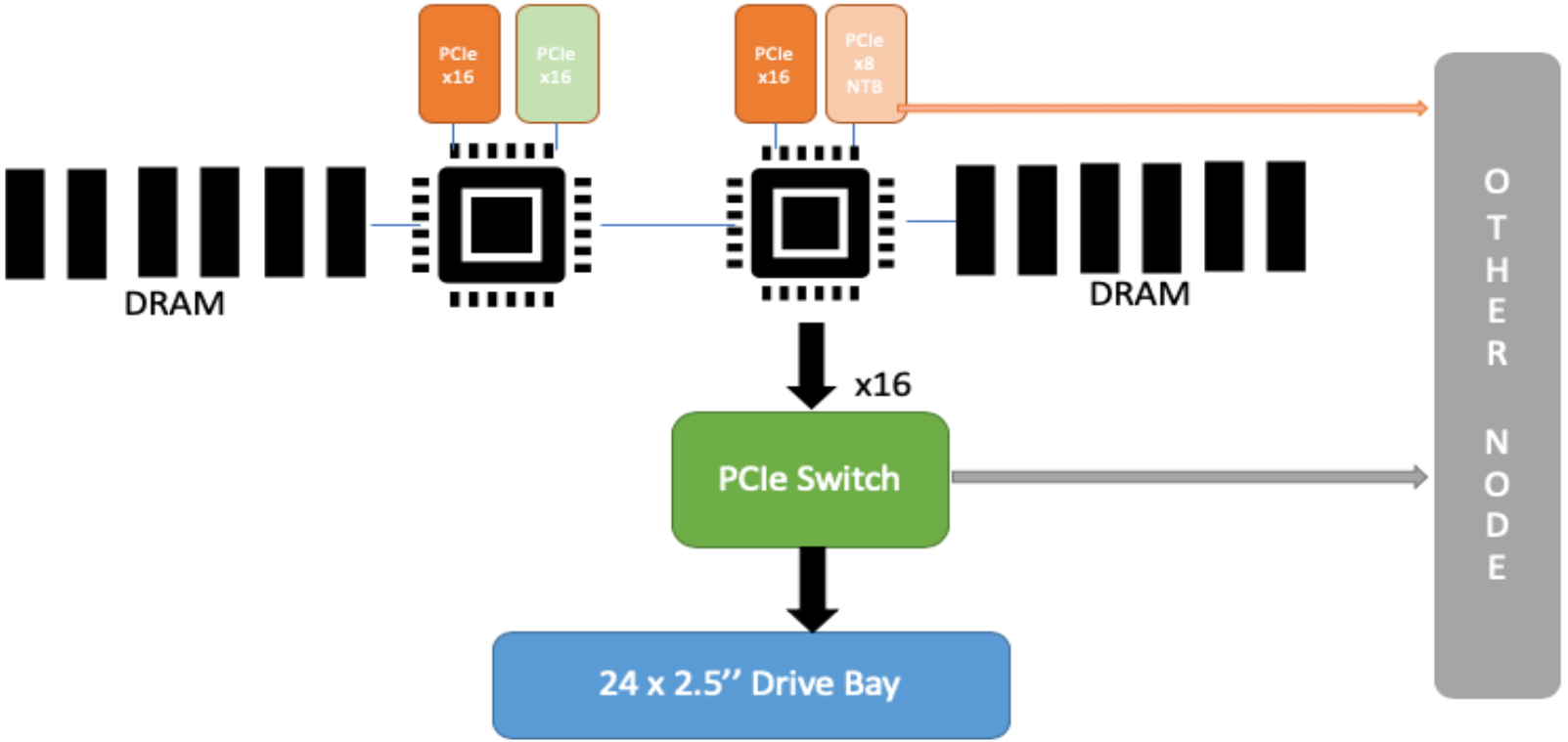
Space and Complexity Tradeoff

Observations

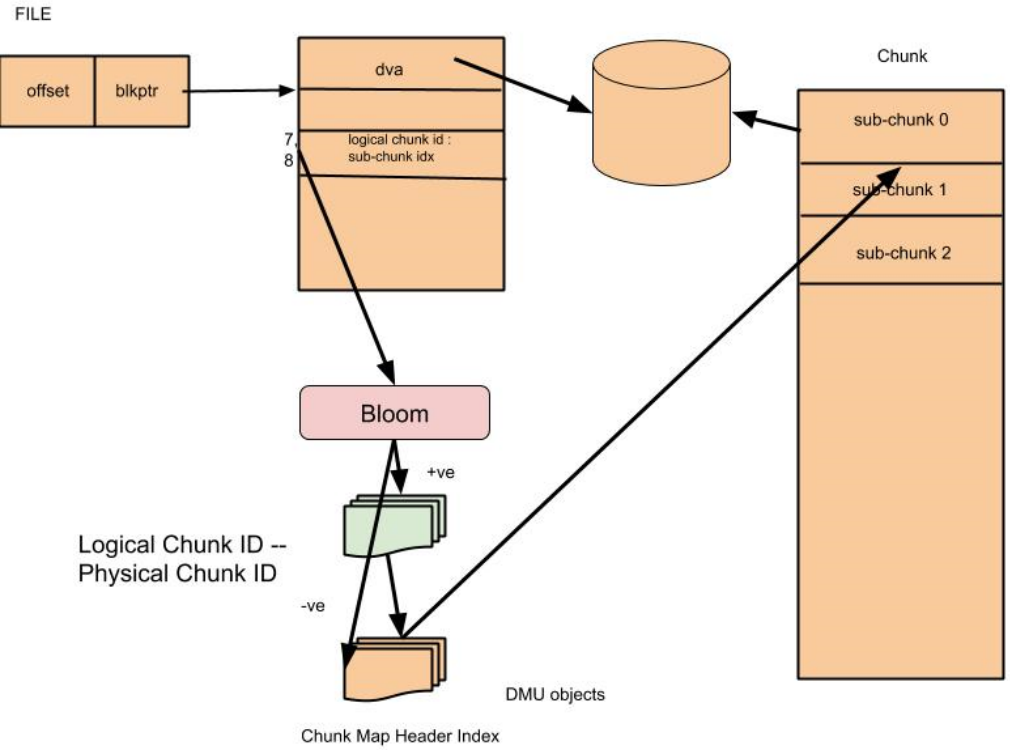
Observations

- N x Drive – Better
- Exploit Delta of Read vs Write
- Computational Overhead – Lower the better
- Resilver Only Used Data
- Seq Writes – Higher I/O Size
- Random Sequential Reads – Higher I/O Size

Intelliflash Appliance Block Diagram



iRAID



- Client I/O and Drive I/O Independent, Separated
- Stream of 1 MB Chunks
- Small I/Os mapped in Chunks
- Concurrent Allocator
- Concurrent I/O queues
- Meta / Data Separation
- Headers Sequential

Resilvering Process

- Scan Chunks Sequentially First
- Headers Contain – In-Use data map
- Data / Meta in Parallel
- Read – Higher I/O size, Sorted, Rand-Seq
- Write – Higher I/O size, Sorted, Rand-Seq
- Example - RAID – Double Parity
- For 2 GB/s, read per drive 200MB/s
- Read Delta utilized

Limited By Single Drive Write Bandwidth

Resilvering Beyond One Drive BW

- Client Write is not in synchronous path
- Throttling needed as client write shared
- Sustained Write Case – Latency spills
- Impacted on read
- Internal Fragmentation on the chunks

Resilvering Beyond One Drive BW

- Combination of Re-write Chunks + Resilver
- Re-write chunks – spreads data uniformly
- Chunk with low-space – recompacted
- Others resilvered
- Dynamically adaptable to exploit BW
- Even sharing – Least Impact

Capacity Increase and PCIe

- PCIe speeds doubling (Gen4, Gen 5 each)
- Drive Read and Write BW as well
- Similar Level of Resilver Times
- Full Drive Failure vs Part Failure
- Endurance Groups
- Zoned Namespaces

Erasure Coding and RAID
Space and Performance Tradeoff

Thank You