G2M
RESEARCH

G2M Research Multi-Vendor Webinar: Can Your Servers Handle the Size of Your SSDs?

Tuesday January 26, 2021

Webinar Agenda

9:00-9:05	Ground Rules and Webinar Topic Introduction (G2M Research)
9:06-9:29	Sponsoring Vendor presentations on topic (8 minute each)
9:30-9:36	Key Question 1 (1-minute question; 2 minutes response per vendor)
9:37-9:37	Audience Survey 1 (1 minute)
9:38-9:44	Key Question 2 (1-minute question; 2 minutes response per vendor)
9:45-9:45	Audience Survey 2 (1 minutes)
9:46-9:52	Key Question 3 (1-minute question; 2 minutes response per vendor)
9:53-9:59	Audience Q&A (7 minutes)
9:59-10:00	Wrap-Up

G2M Research Introduction and Ground Rules

Mike Heumann

Managing Partner, G2M Research

Panelists

Matt Hallberg
Sr. Product Marketing Manager
www.kioxia.com

Jonmichael Hands
Sr Strategic Planner, Prod Mgr
www.intel.com

Josh Goldenhar
VP, Product Marketing
www.lightbitslabs.com

Mike Heumann
Managing Partner
www.g2minc.com

SSDs and Servers – An Uneasy Relationship?

- SSDs have continued to increase in capacity with Moore’s Law
  - Max capacities of 16TB are here, 32TB are close
- However, the growth in processing power of servers has plateaued
  - Clock speeds, feature size improvement rates have slowed
- A single server may have needed 4-6 SSDs a five years ago
  - One SSD can accommodate most workloads today
- This causes issues for redundancy, data sharing, data management, etc.

G2M RESEARCH

Rebalancing the Relationship

- How Can We Balance SSD Size, Server Processing Capacity, Application Needs, and Data Management?
  - Networked SSDs – Sharable over Ethernet
  - Scale-Out Flash Storage (SOFS) software
  - Smart Storage/Computational Storage – adding processing capabilities to storage
- All these approaches have tradeoffs
  - Application software impacts
  - Data management
  - Network congestion management
- These are the subjects that we will explore today

Kioxia

- Matt Hallberg

  Senior Product Marketing Manager

  www.kioxia.com

SSD Failure Mitigation

NVMe™/ PCIe® Specification Features

Asynchronous Events
SMART Health Logs

KIOXIA Features

- Internal RAID6 / Die Failure Recovery
- Wear Leveling
- Endurance Throttling
- End of Life Behavior
- High Availability / Redundancy (Dual Port)

PCIe is a registered trademark of PCI-SIG. NVMe is a trademark of NVM Express, Inc.

SSD Failure Mitigation: NVMe™ 1.4 Spec / PCIe® Gen 4.0, Asynchronous Events*

- Asynchronous Events
  - Asynchronous Events are used to notify the host of status, errors, and health information.
  - When an asynchronous event occurs, the device (controller) will notify what kind of event occurred
    - Error Events (not necessarily related to discussion)
      - Invalid Doorbell Write (01h): Related the host doing things “out of boundary”, i.e. adding to full submission queue, out of range values
      - Diagnostic Failure (02h): Diagnostic Failure detected. This may include device self-test.
      - Persistent Internal Error (03h): Failure that is persistent and unable to be isolated. Host should perform reset.
      - Transient Internal Error (04h): Transient error occurred specific to set of commands, no reset needed
      - Others (Firmware Load, Reserved)
    - **SMART / Health Status Events**
      - NVM Subsystem Reliability (00h): Reliability has been compromised. Could be due to significant media errors, an internal error, read-only mode, volatile memory back-up device failing
      - Temperature Threshold (01h): Device has reached a temperature lesser/greater than defined minimum/maximum threshold
      - Spare Below Threshold (02h): Available spare capacity has fallen below threshold (i.e. spare area is used up, drive can no longer be written to)
      - Others
    - Others: Notice Event, Command Set Event, Vendor Specific Event
  - Device keeps bugging the host that an Asynchronous Event has occurred until host takes necessary steps to “clear” the event

*** - NVM-Express-1_4-2019.06.10_Ratified used for this presentation**

SSD Failure Mitigation: NVMe™ 1.4 Spec / PCIe® Gen 4.0, SMART Health Logs*

- SMART Health Logs
  - This log is used to provide SMART and general health information to the host and is provided over the life of the SSD
  - Get Log Page – SMART / Health Information (02h), relevant bytes listed (others not specified)
    - Critical Error (Byte 00h)
      - Bit 0 - Available spare capacity has fallen below threshold, i.e. spare area is used up, drive can no longer be written to
      - Bit 1 - Device has reached a temperature lesser/greater than defined minimum/maximum threshold
      - Bit 2 - Reliability has been compromised due to significant media errors, an internal error that degrades NVMe subsystem
      - Bit 3 – Device has been placed in read-only mode
      - Bit 4 – Volatile memory back-up device failing (Power Loss Protection capacitors have bad health / failing)
    - Available Spare (Byte 03h) – contains a normalized percentage (0 to 100%) of the remaining spare capacity
    - Available Spare Threshold (Byte 04h) – when the available spare capacity falls below the value specified in this field (can be set by user), an asynchronous event may occur
    - Percentage Used (Byte 05h) – contains a vendor specific estimate of the SSD life based on actual usage and manufacturer's prediction
      - Can go above 100% without incurring an asynchronous event
    - Media and Data Integrity Errors (Bytes 160-175) – Contains the number of occurrences where the SSD detected an uncorrectable data integrity error. Includes UECC, CRC checksum, LBA Tag mismatch

*** - NVM-Express-1_4-2019.06.10_Ratified used for this presentation**

KIOXIA CM6 Series Enterprise NVMe SSDs

- Enterprise PCIe® 4.0, NVMe™ 1.4 SSDs
- Form factors: 2.5-inch, 15mm Z-height
- Proprietary KIOXIA architecture: controller, firmware and BiCS FLASH™ 96-layer 3D TLC memory
- SFF-TA-1001 conformant (U.3) works with Tri-mode controllers and backplanes
- Dual-port design for high availability applications
- 6th generation die failure recovery and double parity protection
- High performance with lower power consumption
- Power loss protection (PLP) and end-to-end data protection
- Suited for 24x7 enterprise workloads
- Data security options: SIE, SED, FIPS 140-2
- Six power mode settings
- 5-year warranty; 2.5 million hour MTBF; AFR 0.35%

			CM6 (Mixed-Use)					CM6 (Read-Intensive)					
Endurance		DWPD	3					1					
User Capacity*		GB	800	1600	3200	6400	12800	960	1920	3840	7680	15360	30720
Sequential Read	128KB(QD32)	MB/s	6900	6900	6900	6900	6900	6900	6900	6900	6900	6900	6850
Sequential Write	128KB(QD32)	MB/s	1400	2800	4200	4000	4000	1400	2800	4200	4000	4000	4000
Random Read	4KB(QD256)	KIOPS	800	1300	1400	1400	1400	800	1200	1400	1300	1400	900
Random Write	4KB(QD32)	KIOPS	100	215	350	325	330	50	100	170	170	170	70

* KIOXIA Corporation definition of capacity: 1 GB = 1,000,000,000 (10^9) bytes (see end of presentation for full capacity disclaimer).

Note: Specifications are subject to change

PCIe is a registered trademark of PCI-SIG. NVMe is a trademark of NVM Express, Inc.

KIOXIA CD6 Series Data Center NVMe SSDs

- Data Center PCIe® 4.0, NVMe™ 1.4 SSDs
- Form factors: 2.5-inch, 15mm Z-height
- Proprietary KIOXIA architecture: controller, firmware and BiCS FLASH™ 96-layer 3D TLC memory
- SFF-TA-1001 conformant (U.3) works with Tri-mode controllers and backplanes
- Single-port design, optimized for data center class workloads
- 6th generation die failure recovery and double parity protection
- Consistent performance and reliability in demanding 24x7 environments
- Power loss protection (PLP) and end-to-end data correction
- Data security options: SIE, SED, FIPS 140-2
- Five power mode settings
- 5-year warranty; 2.5 million hour MTBF; AFR 0.35%

			CD6 (Mixed-Use)					CD6 (Read-Intensive)				
Endurance		DWPD	3					1				
User Capacity*		GB	800	1600	3200	6400	12800	960	1920	3840	7680	15360
Sequential Read	128KB(QD32)	MB/s	5800	5800	6200	6200	5500	5800	5800	6200	6200	5500
Sequential Write	128KB(QD32)	MB/s	1300	1150	2350	4000	4000	1300	1150	2350	4000	4000
Random Read	4KB(QD256)	KIOPS	700	700	1000	1000	750	700	700	1000	1000	750
Random Write	4KB(QD32)	KIOPS	90	85	160	250	110	30	30	60	85	30

* KIOXIA Corporation definition of capacity: 1 GB = 1,000,000,000 (10^9) bytes (see end of presentation for full capacity disclaimer).

Note: Specifications are subject to change

SSD Failure Mitigation: KIOXIA Features

- KIOXIA's Enterprise and Datacenter SSDs support features to protect the drive from failing prematurely and/or protecting the data on the drive from further corruption
    - Wear Leveling – We use an in-house developed algorithm that is always ON ensuring the drive's media is worn evenly to prevent premature drive failure due to focused overuse (i.e. writing constantly to the same LBA address range)
    - Endurance Throttling - When enabled, the drive manages how frequently it is written to ensure it meets the 5 year warranty period by throttling the writes to media
    - End of Life Behavior - The drive continues to support read and write operations before it meets critical endurance state, informing host via SMART flags that the drive needs to be replaced
    - RAID6 / Die Failure recovery - We use an internal RAID6 architecture to ensure if a NAND Die fails, we are able to recover without loss of drive functionality
- CM6's dual port functionality allows for redundancy or high availability usage
    - CM6 can operate in single port or dual port mode depending on the DualPortEn pin on the host
    - Dual port allows two paths to the same drive, so if one path goes down, the data is still accessible through the other path

Intel

- Jonmichael Hands
  Senior Strategic Planner, Product Manager
  www.intel.com

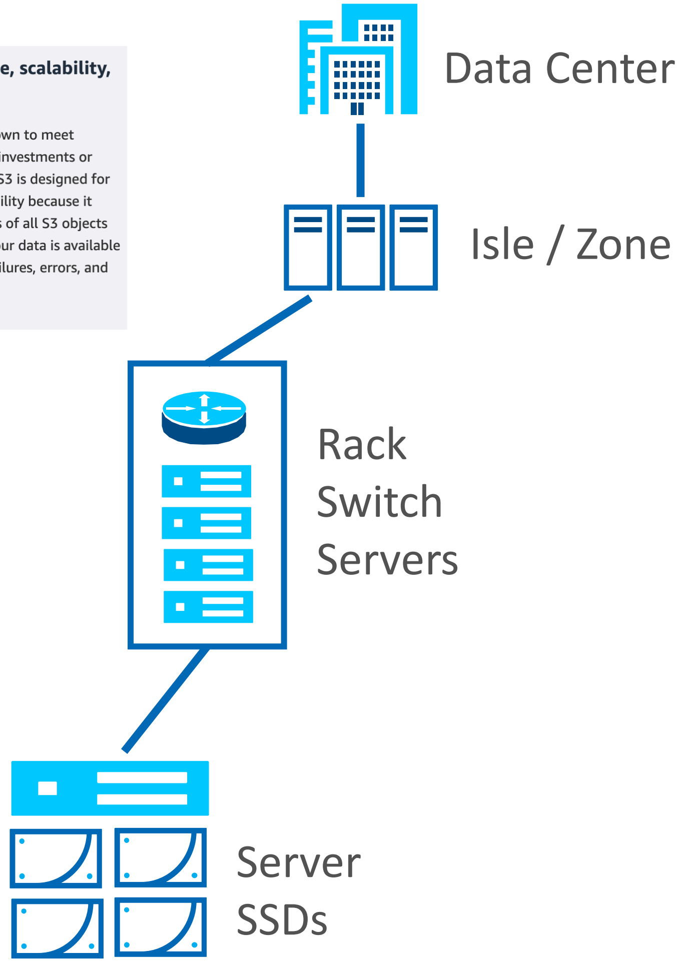
Blast Radius of SSD – a distraction (from March 2020 G2M Webcast)

- Top cloud vendors don't talk about blast radius. They talk about durability, availability, and TCO
- Durability and availability requirements vary drastically by deployment scale
- SDS / HCI distribute data (e.g. Ceph CRUSH, VMware vSAN)
- Blast radius a function of SSD bandwidth, network bandwidth, and replication schema (RAID & EC)
- Small deployments (server, AFA, storage array) rely on rebuild time

Industry-leading performance, scalability, availability, and durability

Scale your storage resources up and down to meet fluctuating demands, without upfront investments or resource procurement cycles. Amazon S3 is designed for 99.999999999% (11 9's) of data durability because it automatically creates and stores copies of all S3 objects across multiple systems. This means your data is available when needed and protected against failures, errors, and threats.

Learn about S3 data durability »



switch gear
UPS
1,000 Kilowatts
Peter's sleep quality
INSOMNIA
RESTLESS NIGHTS
DEEP SLEEP
Blast radius
Complexity
6:51

Expanded 3D NAND Portfolio

Compelling Opportunities

Intel® SSD D7-P5500 & D7-P5600, 96-Layer TLC

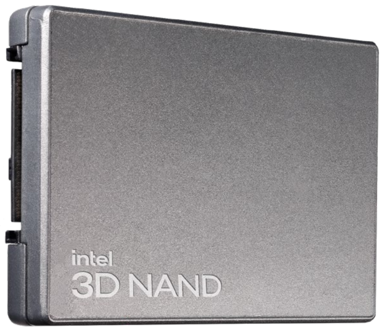
First Intel PCIe 4.0 NVMe for enterprise servers

Shipping since 1H'20

Intel® SSD D7-P5510 144-Layer TLC

NEW

Cloud storage acceleration



Available starting in Q4' 20

Intel® SSD D5-P5316 144-Layer QLC

NEW

Warm storage optimized

Available in 1H'21

Intel® SSD D5-P5316 Key Specifications

Comparison	Performance[1] Spec	Gen to gen
4K Rand. Read	Up to 800K IOPS	38% up to higher[4]
128K Seq. Read	Up to 6800 MB/s	2x+ up to higher[5]
Endurance (Total PB Written)	Up to 18PB (3K P/E Cycles)	4x up to higher[2]

Form Factor & Capacity	
Form Factor	U.2 15mm/E1.L
Storage capacity	Industry-leading QLC storage capacity[3] up to 30.72TB

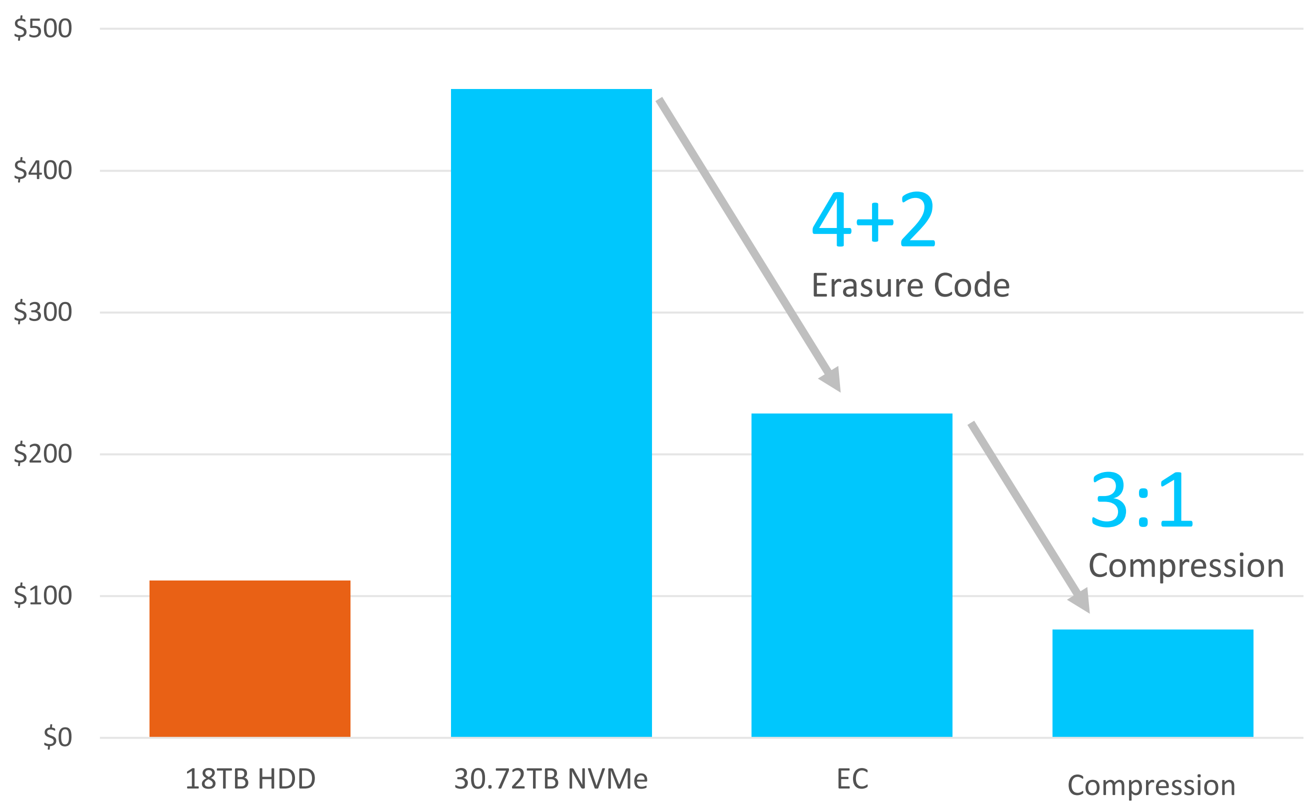
See Appendix for workloads and configurations. Results may vary.

Delivering Disruptive TCO for Capacity Storage[6]

405TB/U

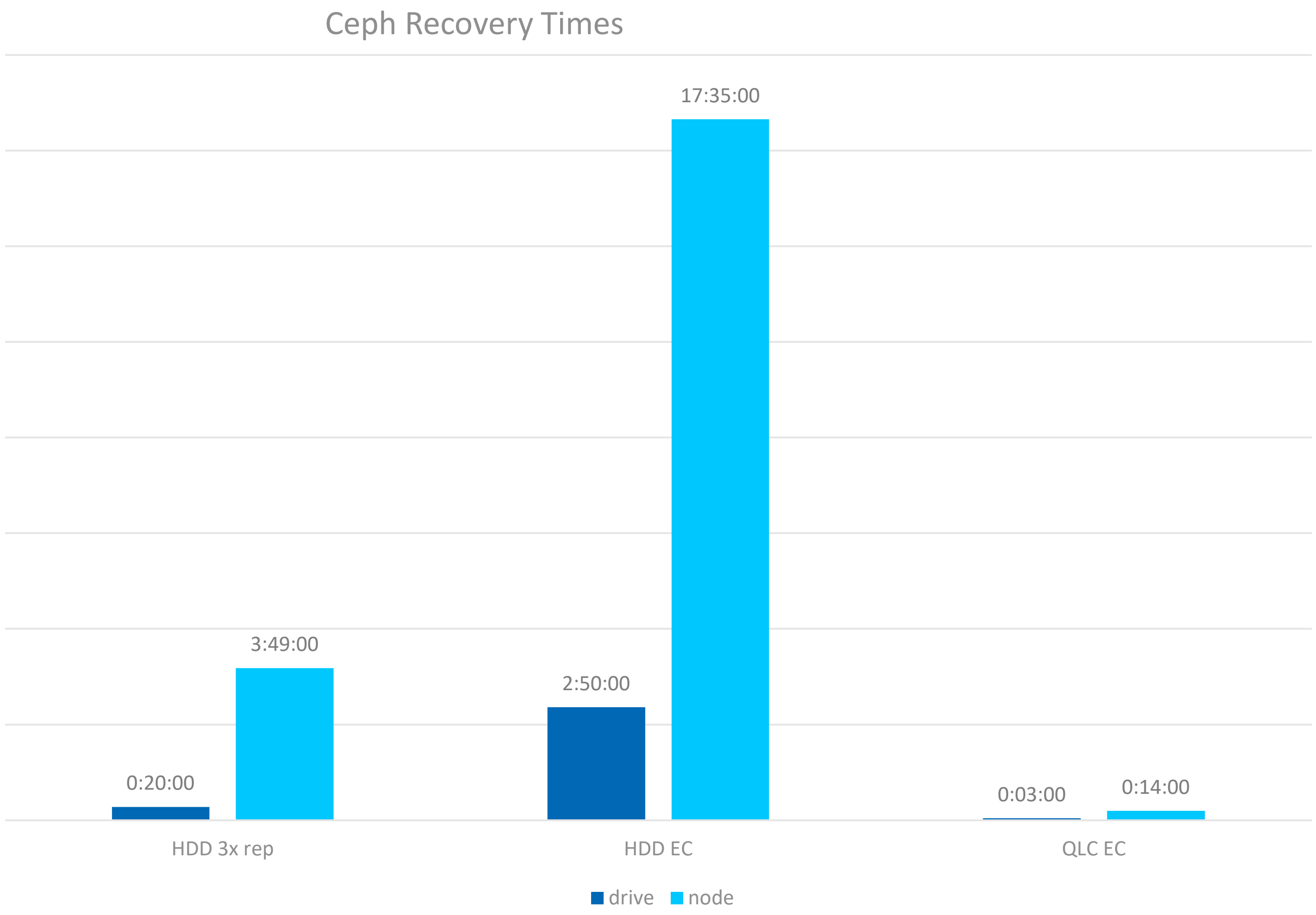
1PB/U

TCO $/TB Effective Per Rack



See Appendix for workloads and configurations. Results may vary.
Source: [Seagate](), Intel TCO model
https://www.supermicro.com/en/products/system/1U/1029/SSG-1029P-NEL32R.cfm
https://www.supermicro.com/en/products/system/4U/6049/SSG-6049SP-E1CR90.cfm

Ceph Rebuild Times – HDD vs QLC[7]



- 5% fill…initial testing takes a LONG time to execute
- Similar platform configs, both using Intel® Optane™ SSD cache & Intel® Xeon® SP, 5 node cluster
- 8TB HDD, 15.36TB QLC…2x the amount of data
- 4-2 EC on QLC is still 16x faster rebuild than 3x replication on HDD, and 75x vs HDD EC[7]

See Appendix for workloads and configurations. Results may vary.

Legal Disclaimers

All product plans and roadmaps are subject to change without notice.

Intel optimizations, for Intel compilers or other products, may not optimize to the same degree for non-Intel products.

Intel technologies may require enabled hardware, software, or service activation.

Performance varies by use, configuration, and other factors. Learn more at www.intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in the configurations and may not reflect all publicly available security updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Results have been estimated or simulated.

Appendix - Intel

1. Test and System Configuration: Mainboard: Intel® Server Board S2600WFT, Version: R2208WFTZS, BIOS: SE5C620.86B.00.01.0014.070920180847, Platform architecture: x86_64, CPU: Intel® Xeon® Gold 6140 CPU @ 2.30GHz, CPU Sockets: 2, RAM Capacity: 32G, RAM Model: DDR4, OS version: centos-release-7-5, Build id: 1804, kernel: 4.14.74, NVMe Driver: Inbox, Fio version: 3.5, G4SAC Gen4 switch PCIe card with Microsemi switch. P5510, P5316 were tested on JCV10020 and ACV10005 firmware respectively .
2. 4x higher endurance gen over gen – Comparing endurance (64K random write) between Intel® SSD D5-P5316 30.72TB (18,940 TBW) and Intel® SSD D5-P4326 15.36TB (4,400 TBW).
3. Industry-leading QLC storage capacity – P5316 capacity up to 30.72TB.
4. Up to 38% higher random read – Comparing 4K random read between P5316 15.36TB (800K IOPS) and P4326 15.36TB (580K IOPS).
5. 2x+ higher sequential read – Comparing 128K sequential read between P5316 15.36TB (6.8GB/s) and P4326 15.36TB (3.2GB/s).
6. Used Intel TCO model, IDC average HDD Prices (Dec 2019). MSRP pricing reflected in Dec 2020, not commitment for official Intel pricing.
7. Ceph rebuild times measured using following configurations for HDD (left) and QLC NVMe (right)

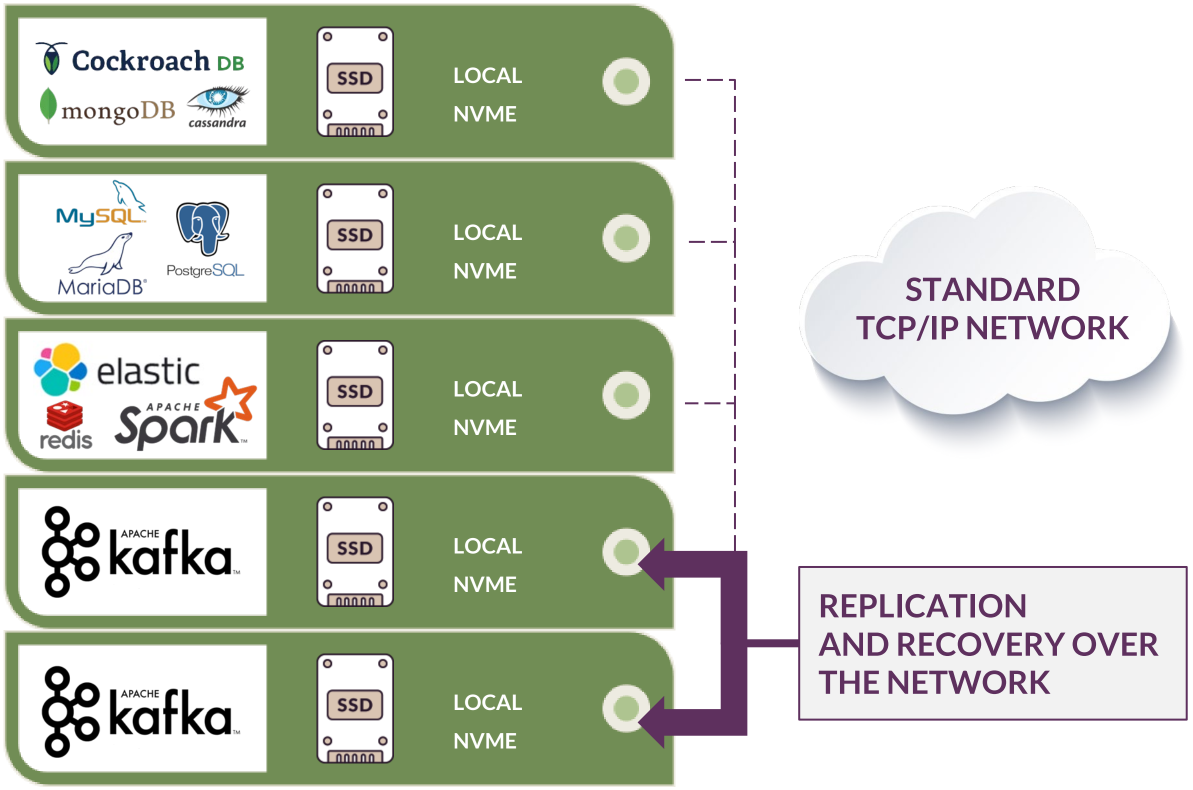
Node Config	5 x OSD Node Configuration
CPU	Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz, 2Sockets
DRAM	187 GB
Metadata Storage	2x Intel® Optane™ SSD 375 GB SSD DC P4800X (Wall devices)
Data Storage	12x7.6TB HDD
Cache Device	N/A
Boot	SATA
NIC	Intel® Ethernet Converged Network Adapter X550T,Intel® Ethernet Network Adapter XXV710
Chassis	Rack Mount Chassis
Board/PS	YZMB-00882-104
Software Configuration	
Operating system	redhat:enterprise_linux:7.9:GA:server Kernel: 3.10.0-1160.6.1.el7.x86_64
Ceph	RHCS 4.0
Ceph OSD backend	Bluestore
Ceph deployment type	ceph-ansible

		5x Ceph OSD Nodes
Codename		qct1-5
Chassis		QCT
Rack Units		1
Motherboard	Model	QuantaGrid D52B-1U/S5B-MB (LBG-1G)
	BIOS	3B13
	# of Sockets	2
CPU	SKU	Intel Xeon Platinum 8280
	# of Cores	28
	Core Frequency	2.7 GHz
	Cache Size	1 MB L2, 3.84 MB L3
Memory	# of Memory Channels	12
	DIMM Type	2666 DDR4
	# of DIMMs/Channel	1
	DIMM Size	16 GB
	Total System Memory	192 GB
Network	Model	Mellanox ConectX-5
	Bandwidth	100GbE
Storage	OS Disk	Intel SSD DC S4610 960 GB
	Data	6x Intel SSD D5-P4326 15.36 TB
	Metadata	2x Intel SSD DC P4800X 375 GB
Software	OS	RHEL 8.1
	Kernel	4.18.0-147.8.1.el8_1.x86_64
	Ceph	14.2.8 Nautilus
	FIO	N/A

Lightbits Labs

Josh Goldenhar
Vice President, Product Marketing
www.lightbitslabs.com

Cloud-Native Applications: The New Normal



NoSQL, In-memory, Distributed

They All Need:

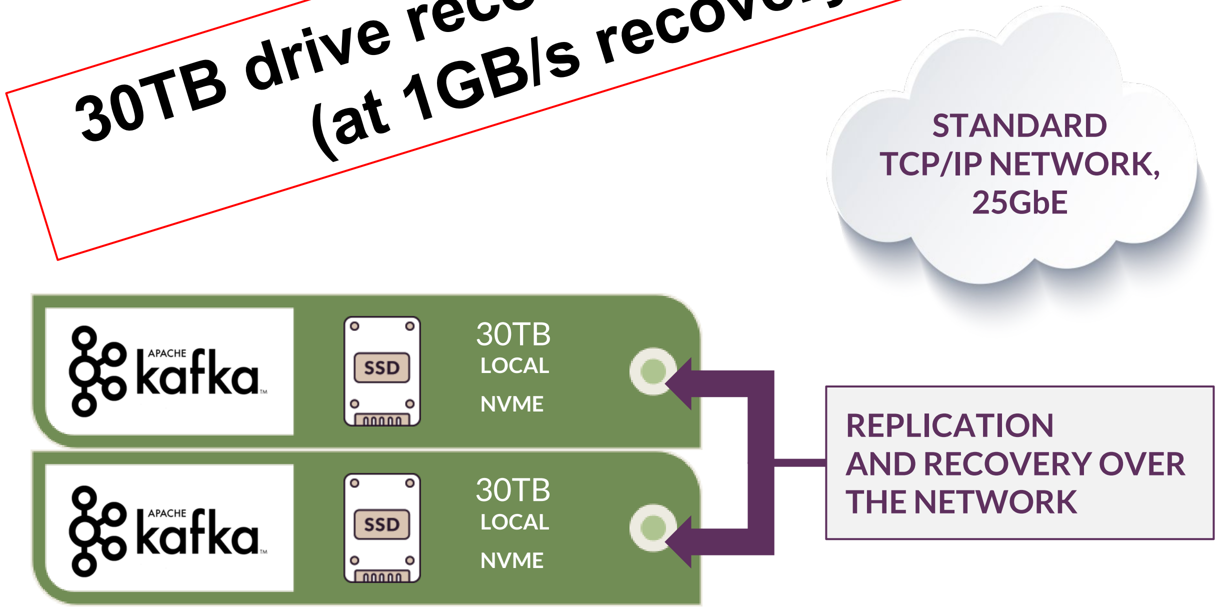
- Low latency and high bandwidth
- Consistent response time
- Applications perform data protection
- Local flash (NVMe)

They All Suffer:

- Poor flash utilization
- Recoveries:
  - LONG, degraded service
  - Severe network impact
- Applications tied to servers

Cloud-Native Applications: Drive (or server) recovery = 100% rebuild over the network

30TB drive recovery will take 8+ hours (at 1GB/s recovery rate)



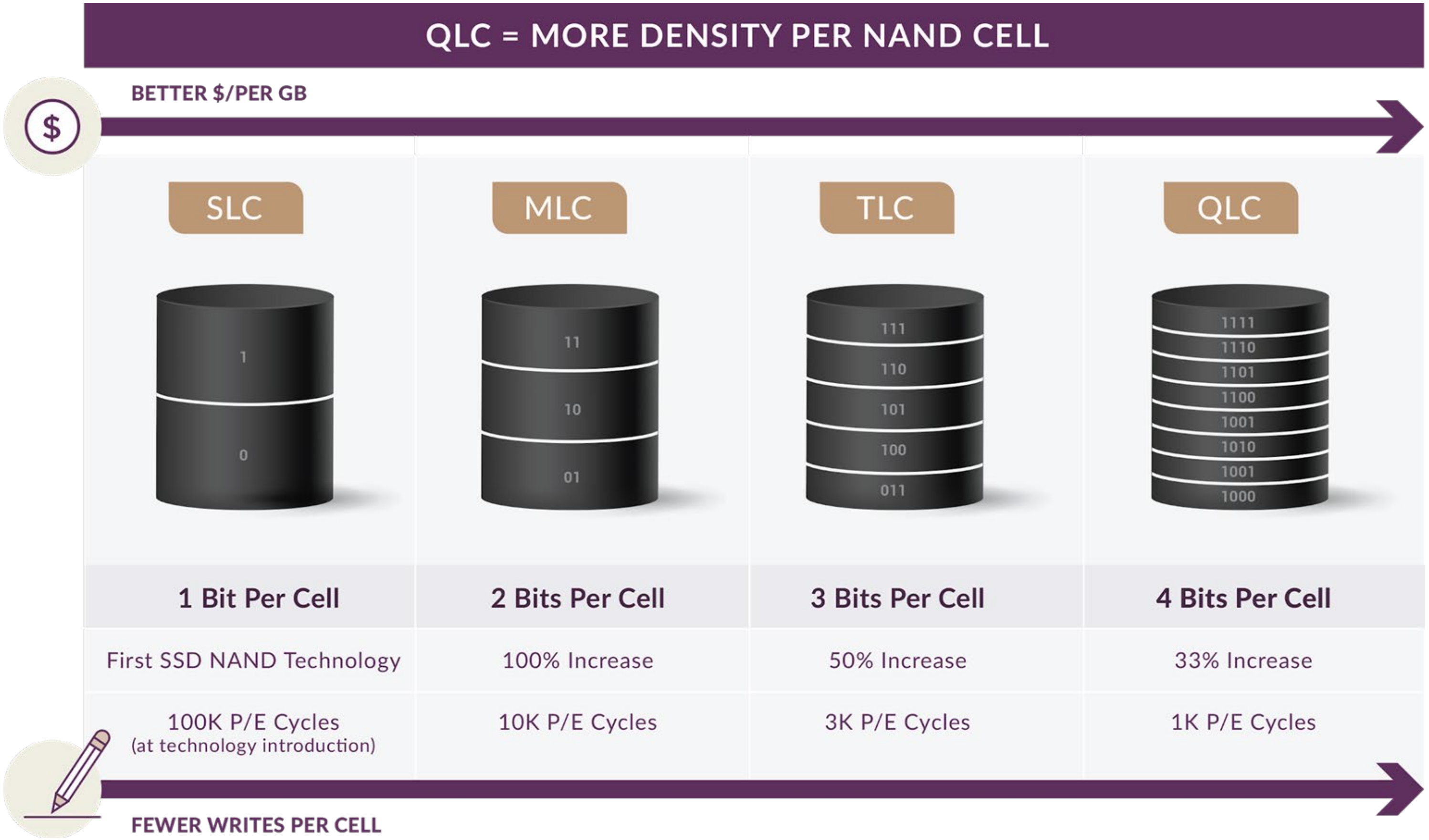
Recovery via Replication

- Likely not intelligent
- May or may not be throttled
- Hinders application performance during rebuild
- Can take a LONG time
- Limited by drive write speed, network or application prioritization

QLC NAND needs special write treatment for high performance and longer endurance

- Small random writes shorten drive life and perform poorly
- Local SW RAID is not optimized for QLC
- 500K IOPs 4K read – but only 11K IOPs 4K write = 44MB/s
- **128K sequential writes = 1600MB/s = 36 times higher!**

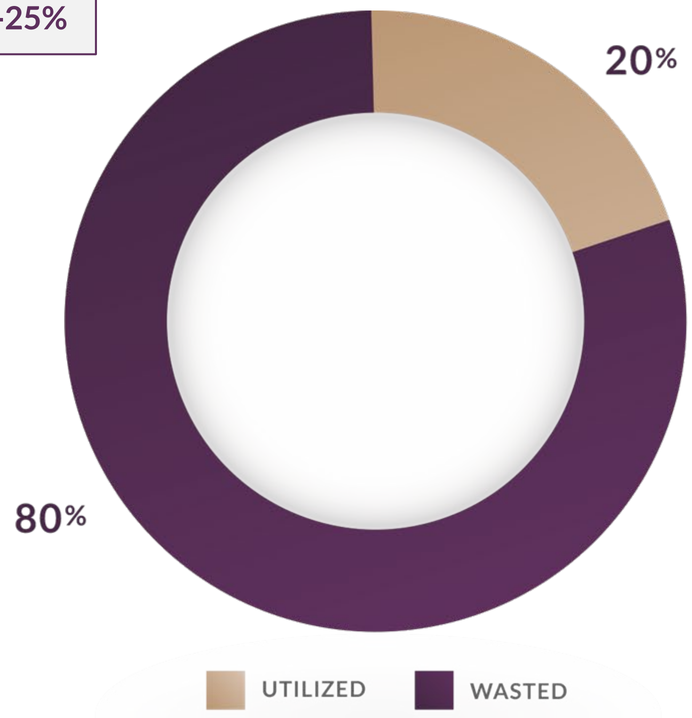
Big Drives = QLC NAND



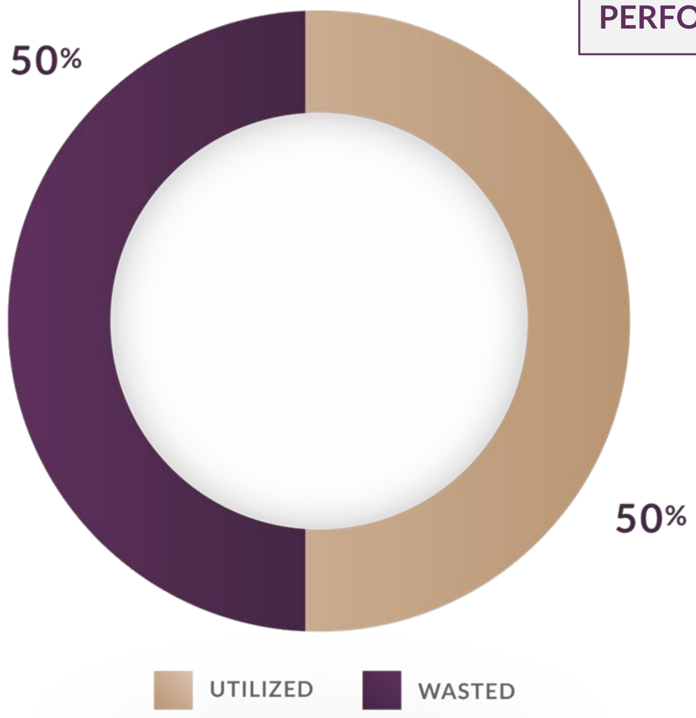
...But can you change your applications?

Severe Flash Under-Utilization

CAPACITY UTILIZATION 15-25%



PERFORMANCE UTILIZATION 50%



50-85% of flash is wasted!

LightOS High Level Features

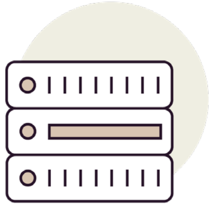
Local flash performance, feature-full data services and high availability

Scalable, Clustered Software Defined

- 3-16 nodes (64 coming soon!)
- Online automatic node replacement
- Online Cluster expansion
- Dynamic data rebalancing

Data Services

- Thin-provisioning
- Inline compression
- Space/Time efficient snapshots
- Thin clones



High Availability and Data Protection

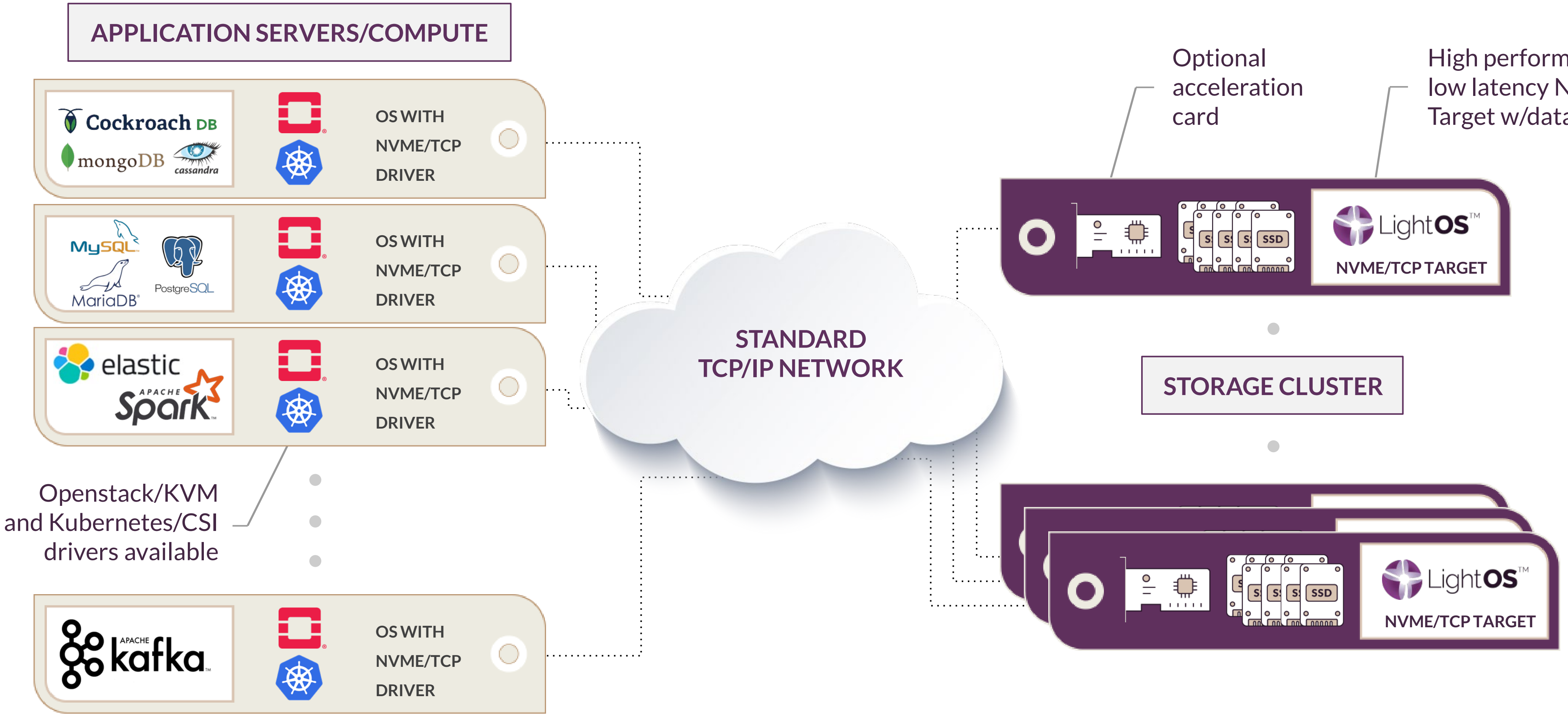
- NVMe multipathing (ANA)
- User-defined failure domains
- LARGE Drive Optimizations
  - DELTA log recovery (partial rebuild)
  - SSD failure protection with distributed, ElasticRAID

Management & Monitoring

- REST API and CLI interfaces
- Ecosystem integrations w/ multi-tenancy:
  - Kubernetes via CSI
  - OpenStack via Cinder
- Monitoring stack: Prometheus and Grafana

High Performance Software Defined Storage

Standard servers, NICs and SSDs, optional hardware accelerator

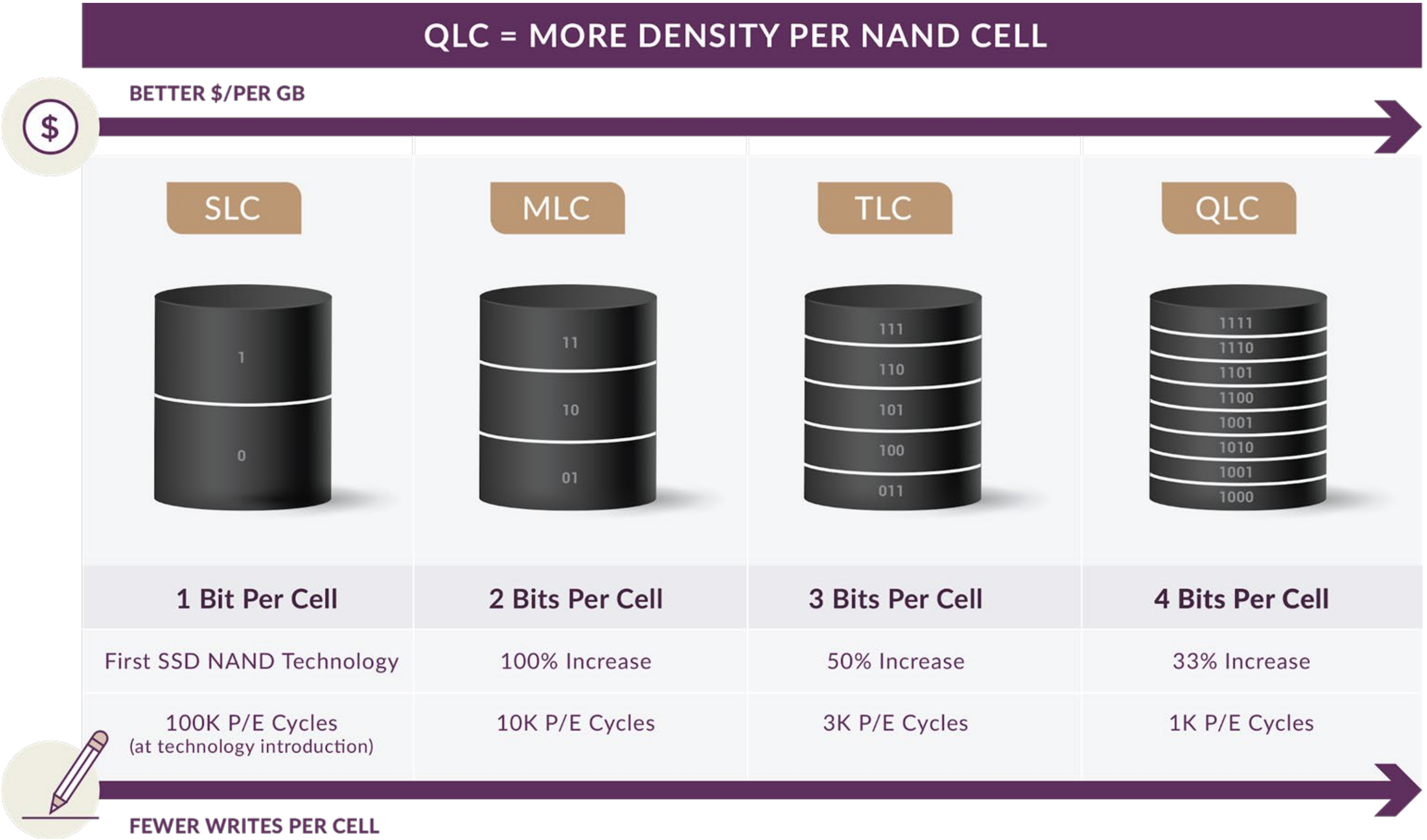


LightOS™

Enable LARGE QLC Drives, Reduce TCO

LightOS enables LARGE QLC Flash:

- Up to 5 times endurance
- Aggregates writes for higher performance
- Allows for high capacity and high density at low cost
- ElasticRAID perfectly suited for QLC NAND



LightOS: Optimized for Large SSDs

- **Share large NVMe SSDs**
  - Local performance, but shared for high utilization and hence maximum ROI
- **Avoid network saturation issues due to drive failures**
  - ElasticRAID rebuilds data due to failed drives within storage servers
    - Distributed rebuild optimized for large drives
- **Optimized for QLC drives**
  - Intel® Optane™ Persistent Memory:
    - Fast NV write buffer making writes to NAND sequential and large
    - Very large memory configurations for metadata needed for large SSDs

Panel Questions and Audience Surveys

Panel Question # 1

- “Blast Radius” is a growing concern. How do we manage blast radius so that servers and applications aren’t impacted?
  - Kioxia
  - Lightbits Labs
  - Intel

Audience Survey Question #1

- To what extent is the SSD “blast radius” problem an issue for your organization? (check one):
  - It is a significant problem across most of our mission-critical workloads today: 17%
  - It is a problem for a number of our mission-critical workloads today: 14%
  - It is a problem for a couple of our workloads today: 9%
  - It is not a workload-specific problem, but a general “IT efficiency” concern:20
  - It is not a problem today, but we expect it to be in the next 2-3 years: 29%
  - We do not see it as an issue for our workloads in the next 2-3 years: 11%

Panel Question #2

- The terms “Smart Storage” and “Computational Storage” continue to be one that is brought up as a solution to this issue. For what type of workloads is this a viable solution?
  - Lightbits Labs
  - Intel
  - Kioxia

G2M RESEARCH

Audience Survey Question #2

- When looking at solutions for the blast radius problem, which of these approaches has your organization explored? (check all that apply):
  - Scale-Out Flash Storage (SOFS) software solutions: 37%
  - Distributed File Systems: 33%
  - Networked SSDs (Ethernet, NVMe-oF, etc.): 30%
  - Composable Infrastructure: 10%
  - Centralized storage arrays: 37%
  - Other: 17%

Panel Question # 3

- What role can NVMe ™ and NVMe-oF™ play in helping to balance the relationship between SSDs, processing power, data management, and application demands?
  - Kioxia
  - Intel
  - Lightbits Labs

Audience Q&A

G2M RESEARCH

Thank You For Attending

G2M

RESEARCH