SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

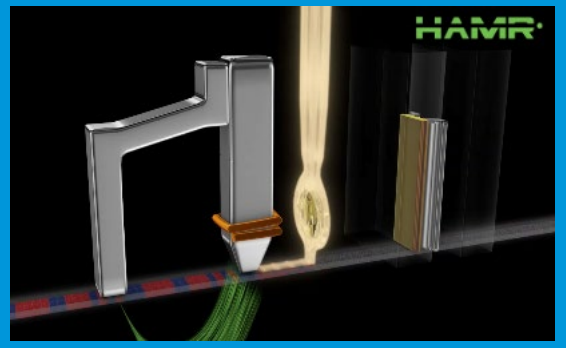
Cloud Storage Efficiency at Scale

Seagate Technologies

Mohamad El-Batal

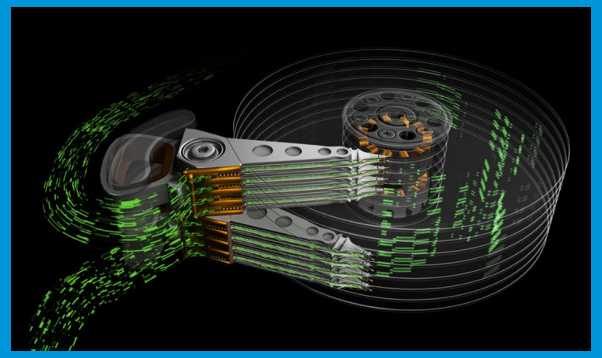
Seagate OCTO Technologist & CTO Storage Systems

Highly Efficient HDD Storage Innovation



Nanoscale Complexity
Deployment Simplicity

Efficiency through
Foundational Storage Technology
Superior Media Density Innovation



TCO Optimized
Data Accessibility

Efficiency through
Electro-Mechanical Innovation
OS & Application Optimization

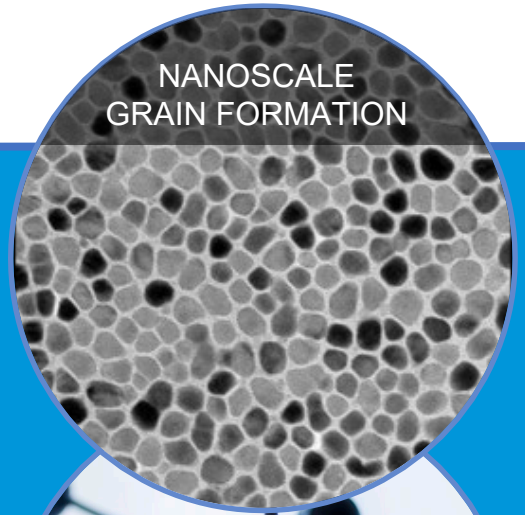
Mass Capacity
Systems Innovation

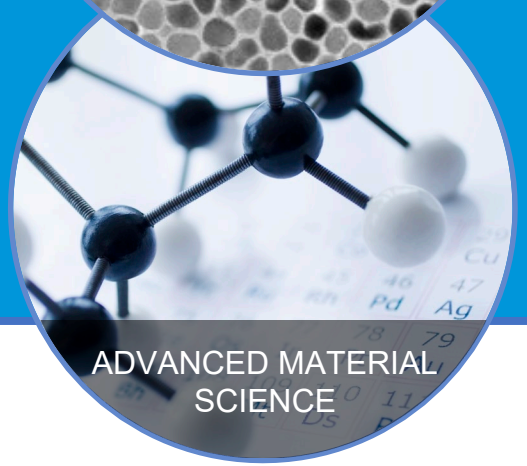
Efficiency through
Intelligent Resource Virtualization
Effortless HDD Deployment

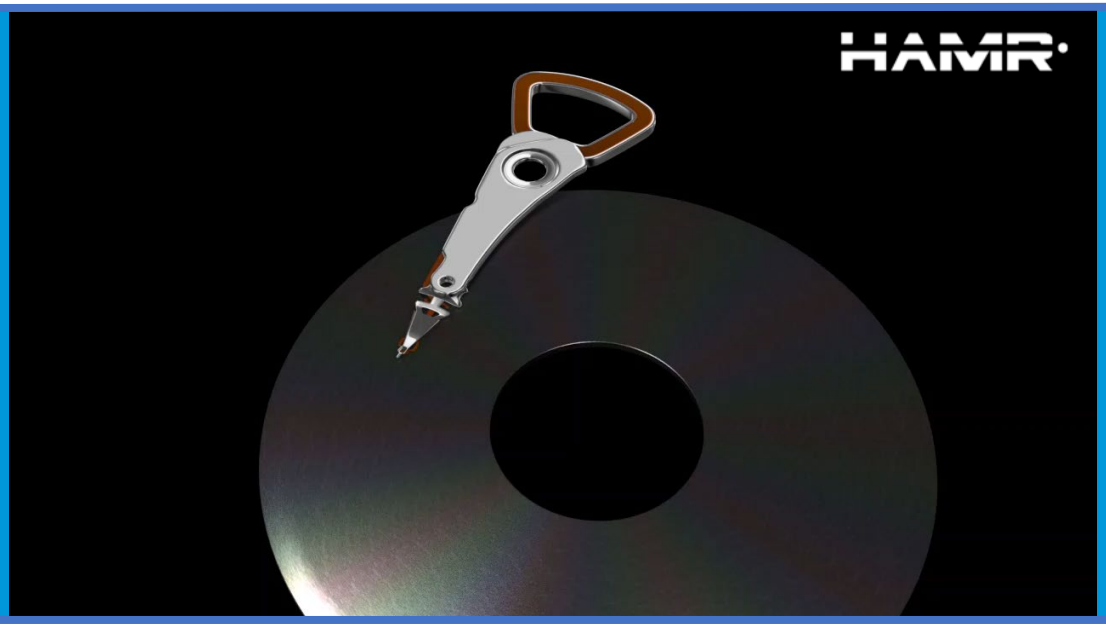
SDC 24

Fueling Storage Efficiency with Recording Density

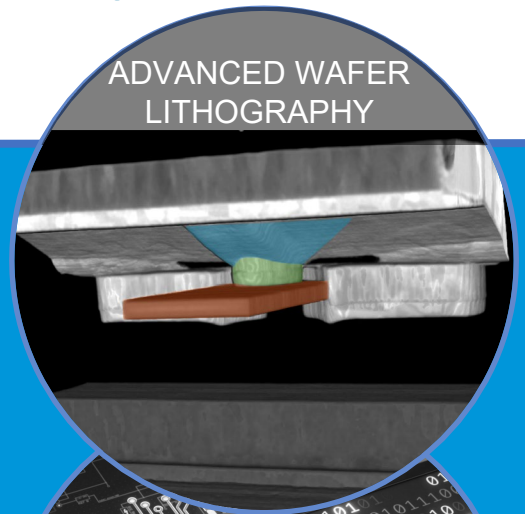
Nanoscale complexity, deployment simplicity



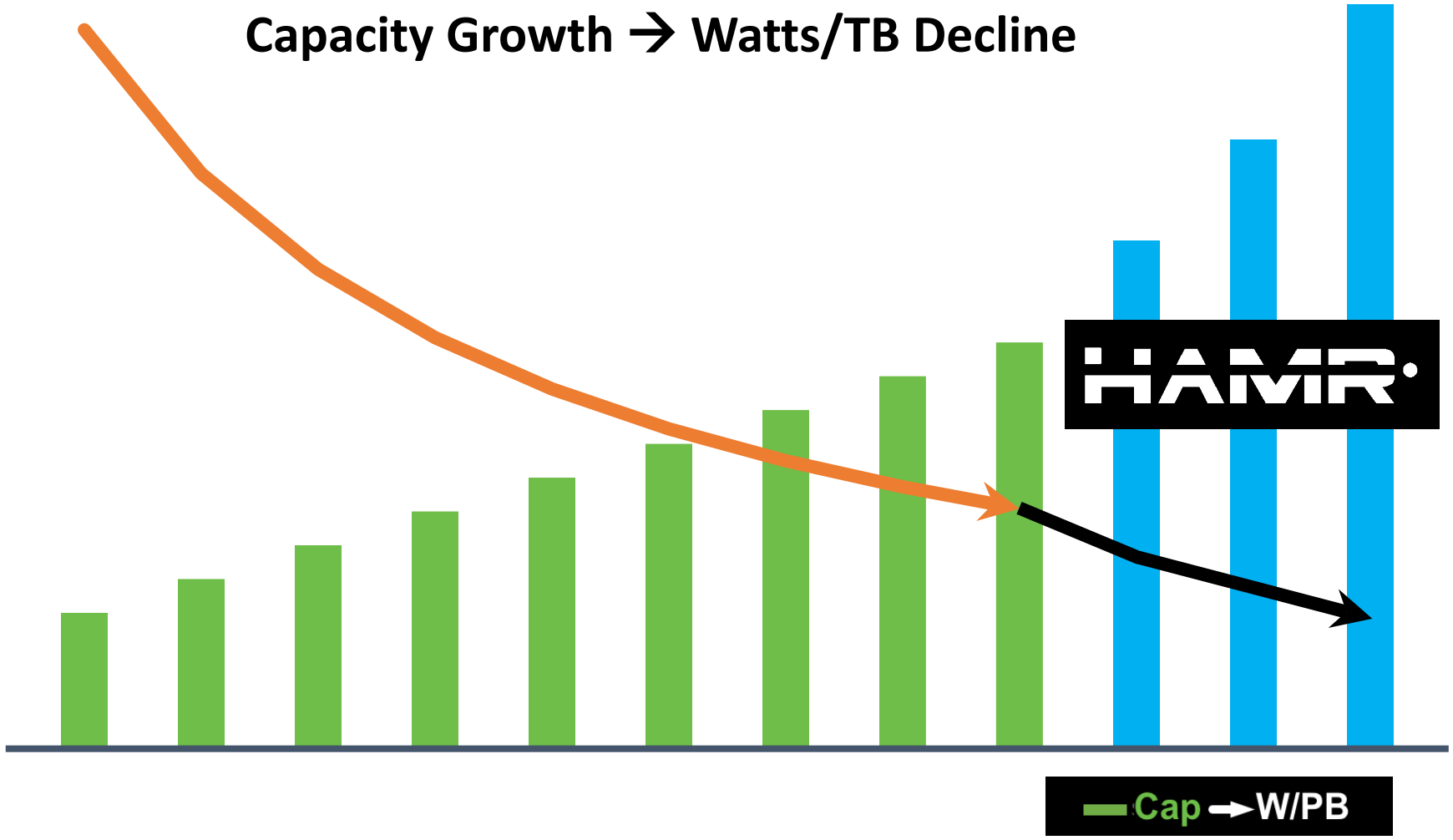




HEAT-ASSISTED MAGNETIC RECORDING



Higher-Capacity Hard Drives Significantly Reduce Power of Storage Infrastructure



SDC 24

Efficient Data Accessibility

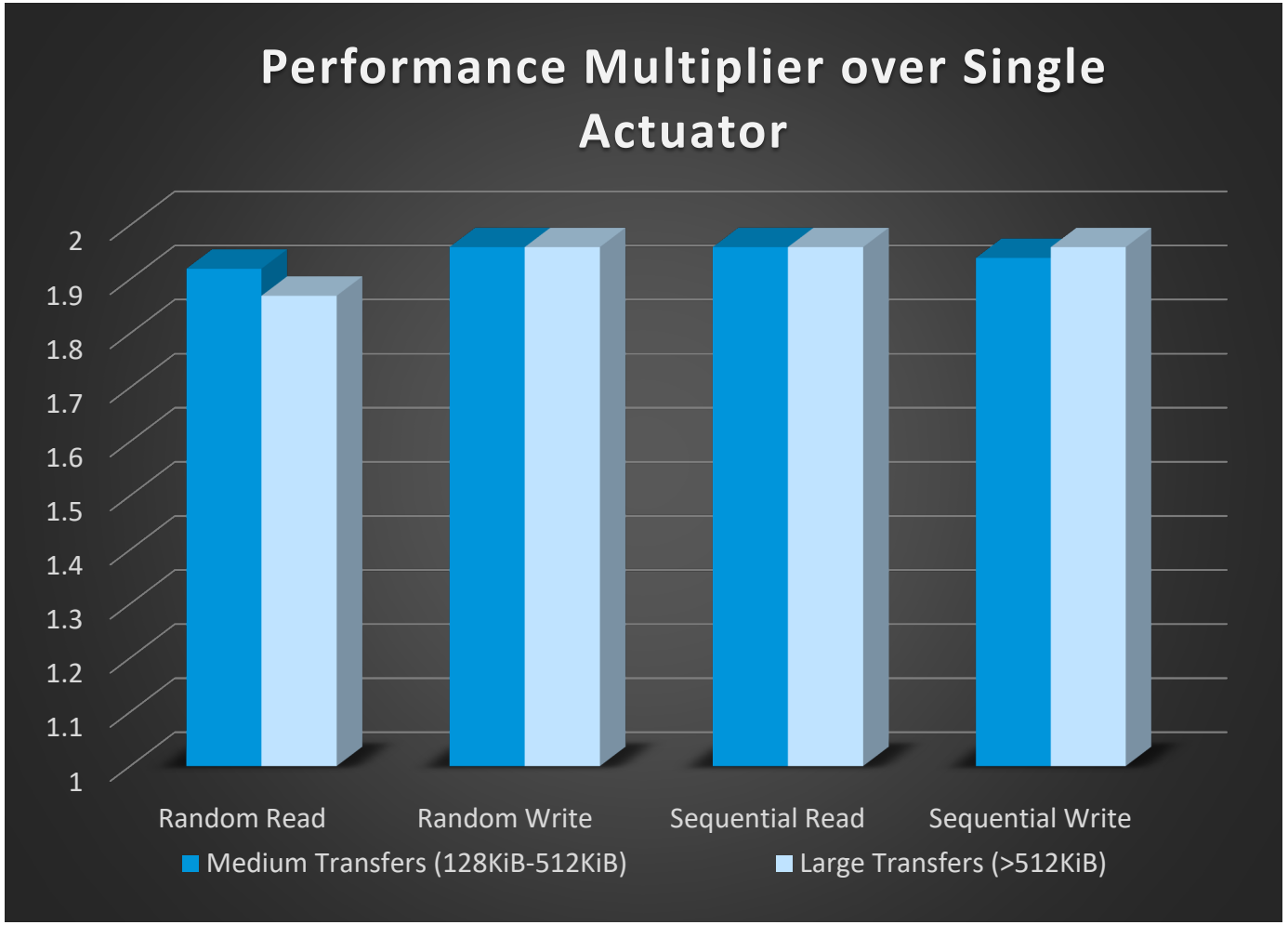
Parallel data streams enable performant deployment of dense storage

Up to 2X BANDWIDTH & IOPS

Optimized Cost, Power & Density

MULTI-ACTUATOR TECHNOLOGY

Dual-Actuator Performance Benefits Over Single Actuator



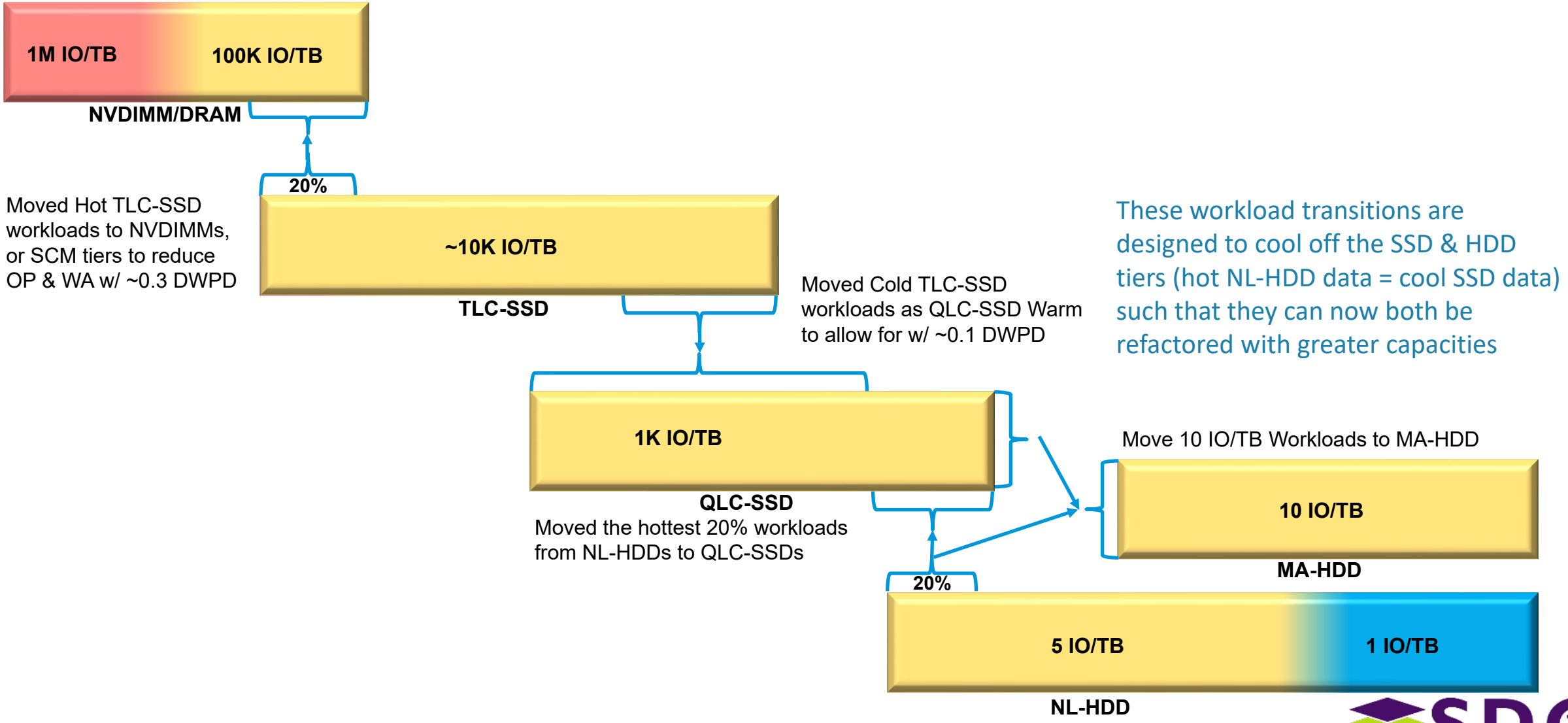
Ideally Maintain ~2X Bandwidth at same Power of Single Actuator!

Random reads are fast across all transfer sizes!

Random write and sequential Rd/Wr Larger transfers = More benefit

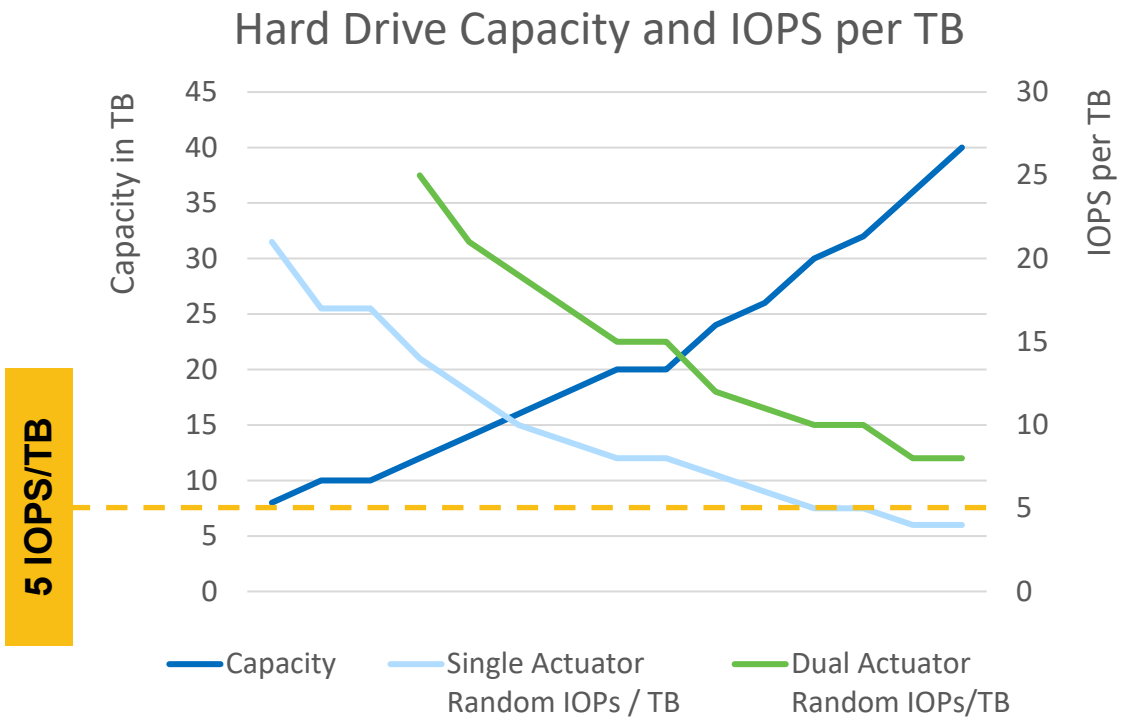
SAS/NVMe the preferred solution for Sequential performance

Storage & Memory Tiers Workload TCO Optimization



These workload transitions are designed to cool off the SSD & HDD tiers (hot NL-HDD data = cool SSD data) such that they can now both be refactored with greater capacities

Dual-Actuator Hard Drive Workload Target



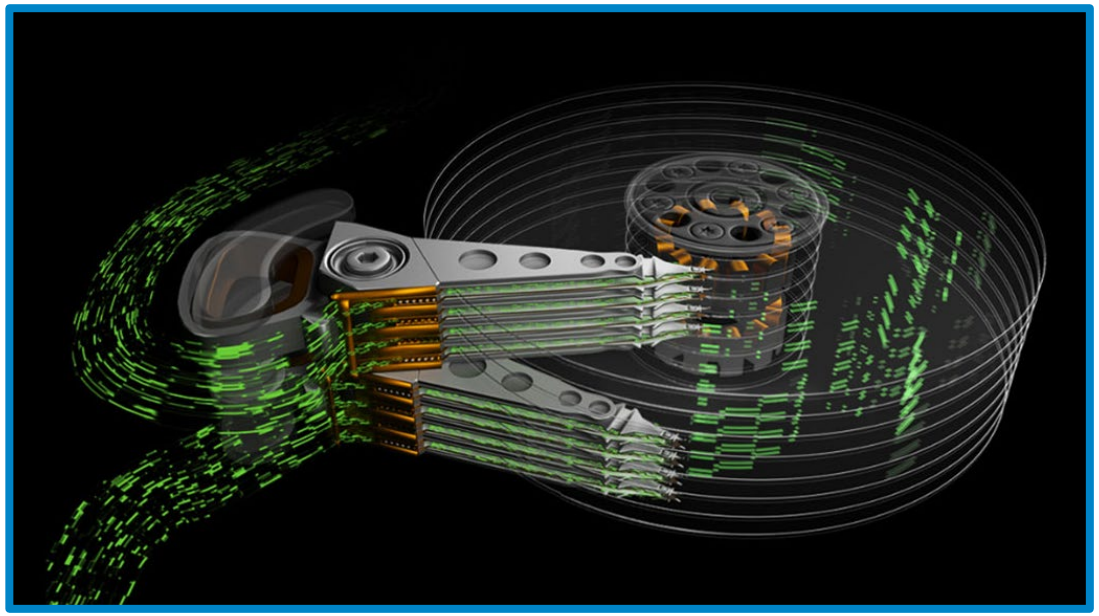
10 IOPS/TB

8TB	14TB	18TB	30+TB
Key Partner Initial Dual Actuator limited Deployment	First-Generation Dual Actuator Production Drive	Second-Generation Dual Actuator Production Drive	Next-Generation Multi Actuator Production Drive

Dual-Actuator
Maintain ~5-10 IOPS/TB as capacity scales

Design Target IOPS/TB

Dual-Actuator: Reduces Power and Embodied Carbon per TB – Easing the pressure on Flash



Sustainability at Scale

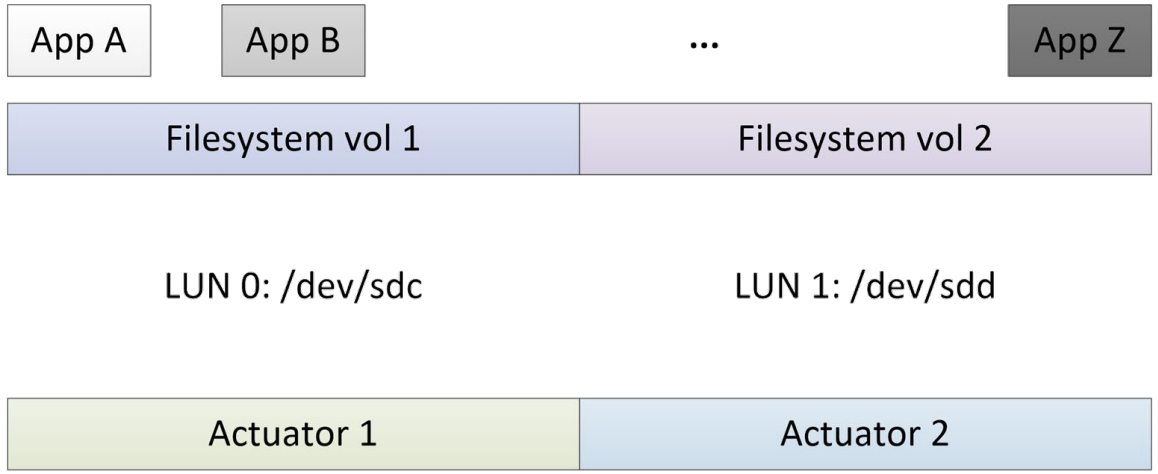
4x less operating power
10x lower EB/TB
Embodied Carbon (EB) per TB.[1]

1. Hotcarbon.org MS Azure white paper, SSD Storage Rack vs. Hard Drive Storage Rack: A call for research on Storage Emissions.

SAS – One Actuator Per Linux “Disk”

14TB SAS Exos 2x14

- Dual LUN
- 1 Filesystem per actuator
- One “disk” per actuator
- Need to manage failure domains

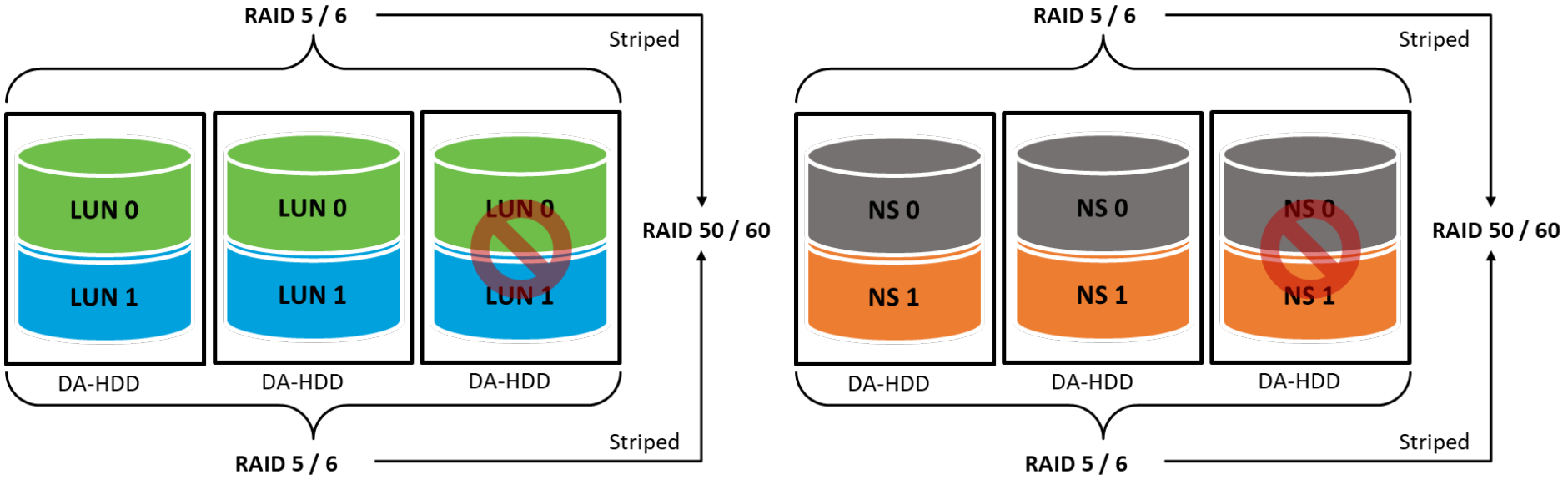


Linux device listing

```
[0:0:0:0]    disk    SEAGATE  ST14000NM0001     K003  /dev/sda   /dev/sg0
[0:0:0:1]    disk    SEAGATE  ST14000NM0001     K003  /dev/sdb   /dev/sg1
[0:0:1:0]    disk    SEAGATE  ST14000NM0001     K003  /dev/sdc   /dev/sg2
[0:0:1:1]    disk    SEAGATE  ST14000NM0001     K003  /dev/sdd   /dev/sg3
[1:0:0:0]    disk    ATA      CT240BX500SSD1    R013  /dev/sde   /dev/sg4
[2:0:0:0]    disk    ATA      SAMSUNG MZ7WD480  NS00  /dev/sdf   /dev/sg5
```

Dual-Actuator – SAS/NVMe RAID Integration

- The LUNs/NSs do not constitute different failure domains
- Multi LUN/NS per device assumption is now prevalent throughout software RAID solutions



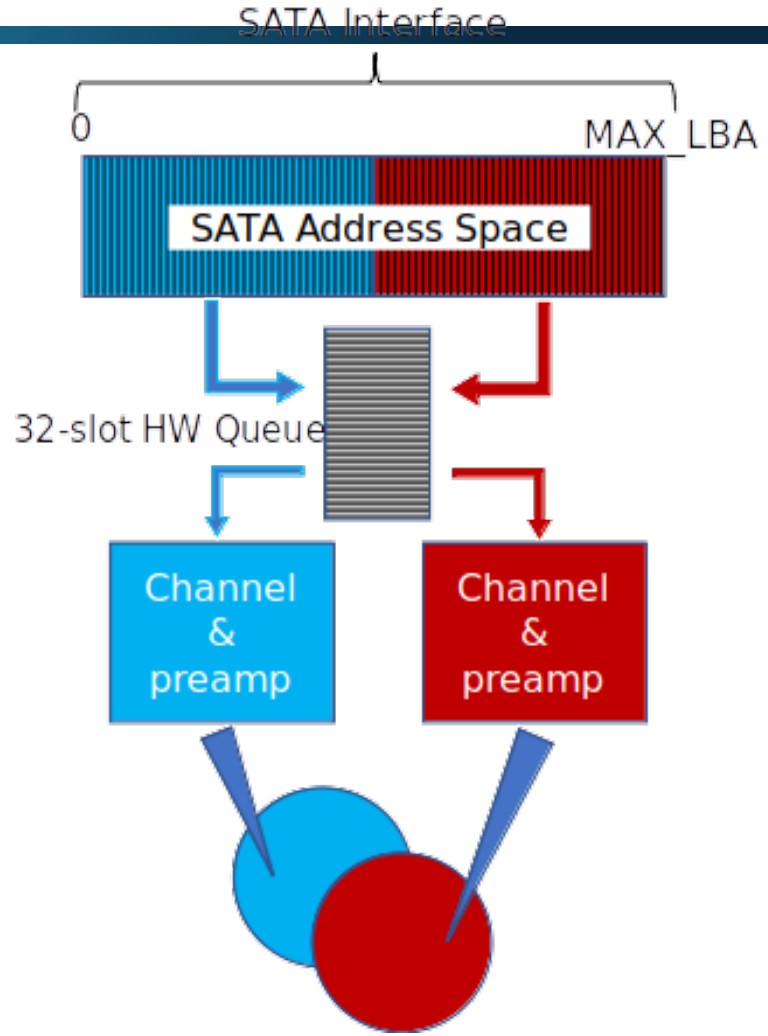
Striping the Data across the two LUN/NS based RAID5 or RAID6 groups producing a single RAID50 / RAID60 group with the same resiliency and availability

Dual-Actuator HDD with independent Media Capacity per LUN/NS

Group one LUN/NS from each HDD into a separate RAID5 or RAID6 redundancy-group

Single-namespace Split Actuator HDD

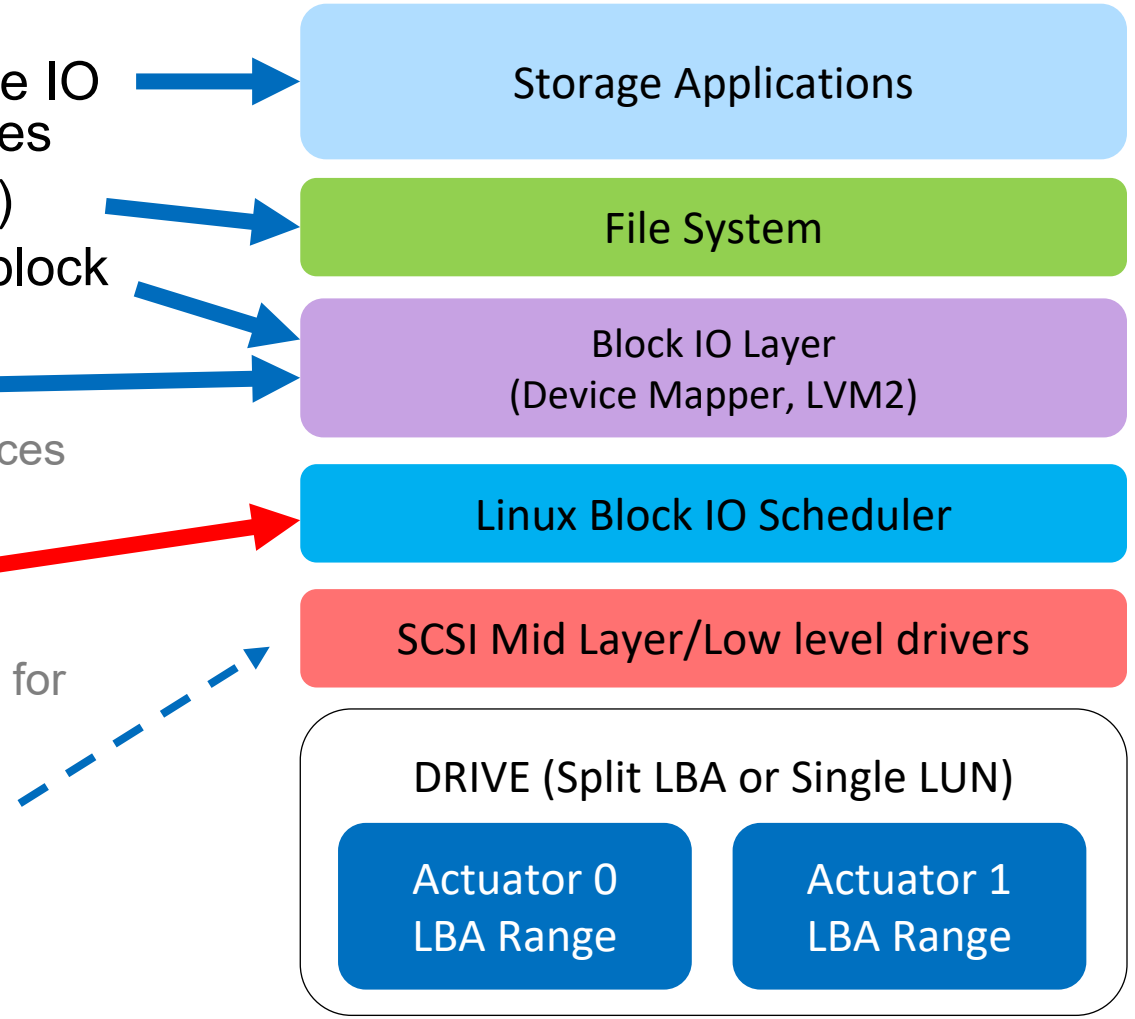
- The split-actuator approach divides the disks into groups, each addressed by an independent actuator. A given sector is reachable by only one actuator - none of the address space is shared.
- For example, Seagate's SATA version contains two actuators and maps the lower half of the SATA LBA space to the lower actuator and the upper half to the upper. There are no changes to the IO protocol, except for a log page to report the LBA:Actuator mapping.
- SATA allows 32 commands to be queued in the NCQ hardware queue. The 32-slot HW queue is a shared resource that services both actuators.
- Like any shared resource, we can either manage it, or it will manage us...



Note: The single-namespace, split-actuator design is also compatible with single LUN SAS.

Storage Stack Solutions for Per Actuator Control

- Applications can be modified to schedule IO across both concurrent positioning ranges
- File System: (kernel dependent, difficult)
- Device mapper target splits underlying block device at the actuator split
- Linux Block Device Partitioning
  - Use GPT to create two independent devices
  - Persistent / Kernel Dependent
- IO Scheduler Optimization
  - Manages commands (using a variety of algorithms) to provide command streams for the ”best” overall IO performance.
- SCSI Subsystem
  - Normally shouldn't redistribute workload
  - Kernel and Legacy complications limit flexibility



HDD Foundational Technologies Innovation Key Takeaways

- The traditional small bumps in legacy technology enhancements are a thing of the past, as new and more aggressive product strategies must be embraced
- Many challenges are ahead of us, yet full stack software architects will continue to find solutions if the commodity vendors are willing to listen and able to adapt
- Memory and Data-Storage vendors are accelerating their foundational technology innovation investments to keep up with the predicted Generative AI demands

Systems Innovation Enables Dense Deployments

Intelligent, Efficient and Reliable

- Fabric Friendly Virtualization
- Ultra-High Density
- High Availability
- Data-Reduction
- Data Protection
- Self-Healing
- Faster Recovery
- Acoustic Management

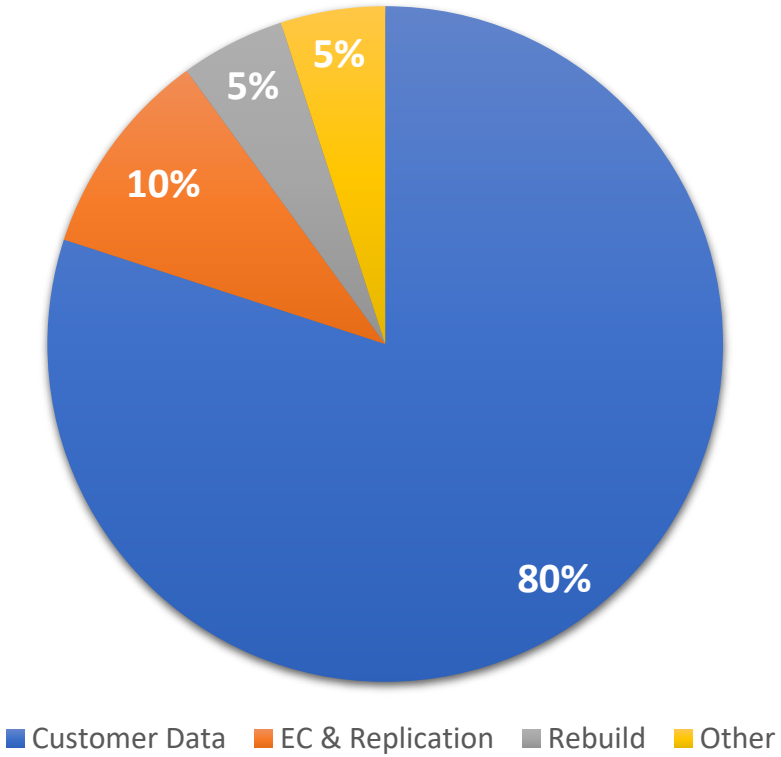
Mass Capacity Systems Innovation

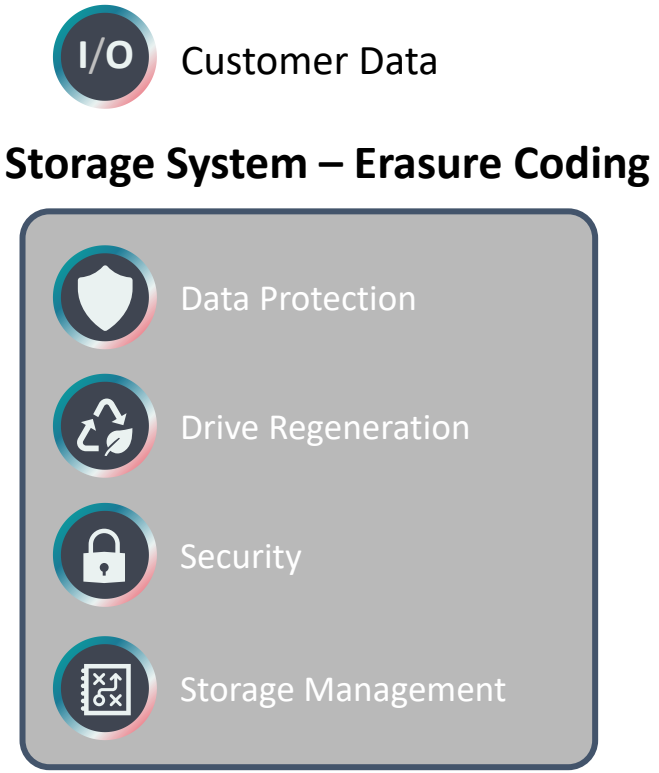
Efficiency through
Intelligent Resource Virtualization
& Fabric HDD Deployment

SDC 24

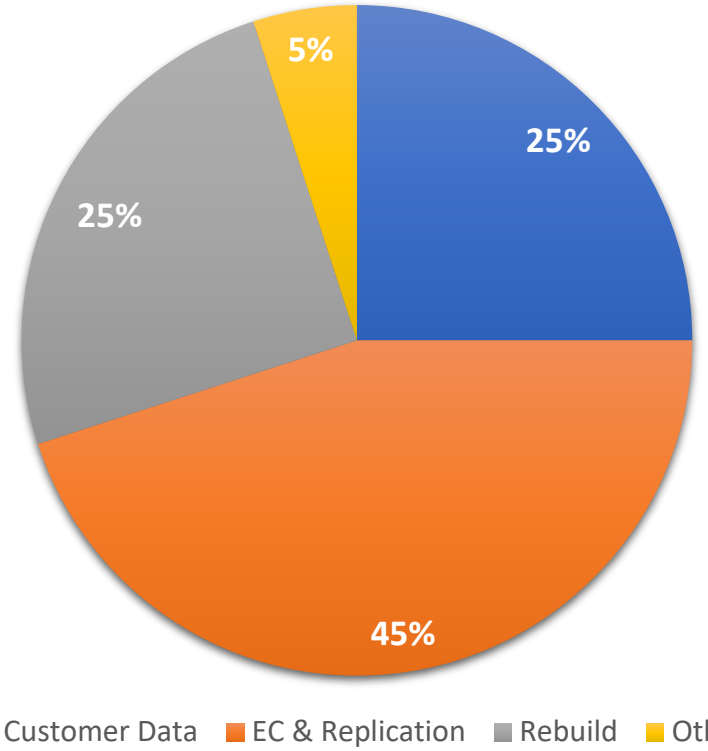
Hidden Cost of Data Management

Breakout of Network Traffic at Cloud Datacenters
Storage System – Erasure Coding





Breakout of Host Resources at Cloud Datacenters
Traditional JBOD Storage



Storage providers are forced to sacrifice host resources to protect their customers data

Superior Scale-Out TCO

Best In-Class Data Center Efficiency –Storage Systems with built-in Erasure Coding for Exabyte-Era

20PB Solution

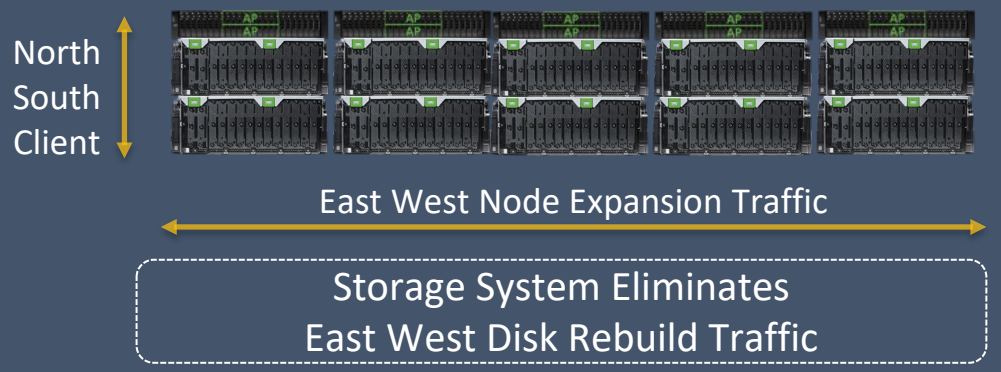
256 CPU Cores

↓ 60% Less CPU

4,096 GB RAM

↓ 40% Less RAM

Scale-Out CEPH Storage Cluster



Multi-Layer EC

- Disk rebuild traffic is localized to each Storage System
- Rebuild Traffic is eliminated
- 14 nine's data durability
- CEPH [16+2] + ADAPT

* Saving is Storage System dependet and calculated with Seagate Exos AP Storage Server and CORVAULT 4U106

CAPEX Savings

- ❑ Racks
- ❑ Switches
- ❑ Cabling

OPEX Savings

- ❑ Power & Cooling
- ❑ Real Estate

SDC 24

Enhanced Durability w/ ~Same Efficiency = <TCO

Design a 20PB solution

Storage Server / JBODs

- 53 HDDs per Server = ~ 27 Cores Per Server Required
- Efficiency (16+4) = 80%
- Durability = 10 nines (99.99999999%)

Storage System with integrated Erasure Coding

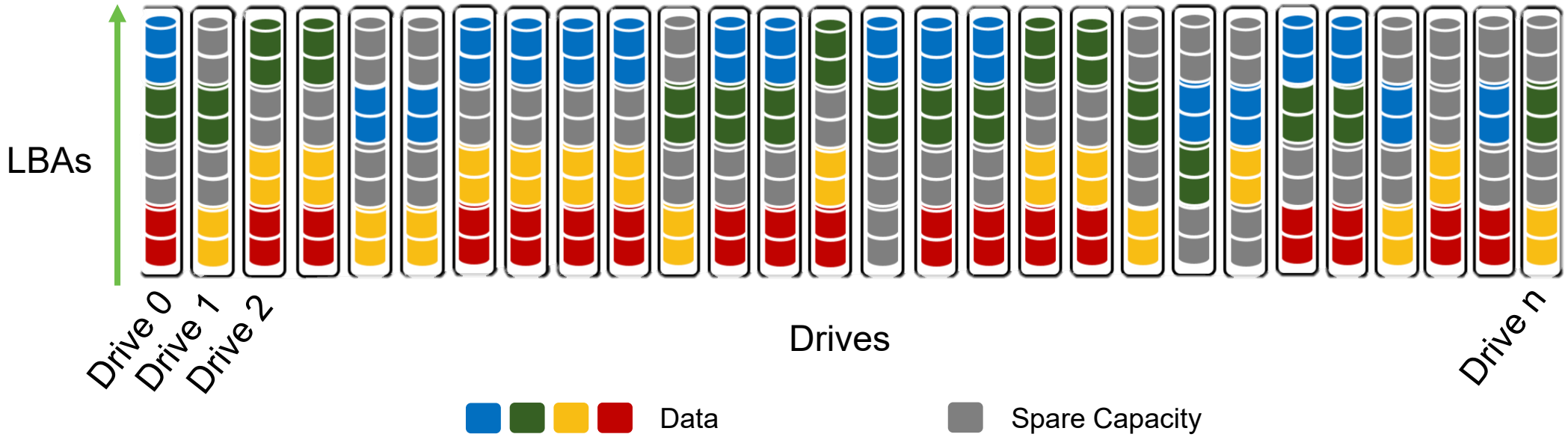
- 20 LUNs per Server = ~ 10 Cores pers Server Required
- Deliver about the same efficiency = 79% (16+2) x 2
- Deliver MUCH BETTER durability = 14 nines
  - (99.99999999999%)
  - **~60% Reduction** in CPU Cores
  - ~ $1000/PB Saving on BOM Cost

20 X $$$ storage servers

20 X $ servers + 10 Erasure Coding protected Systems

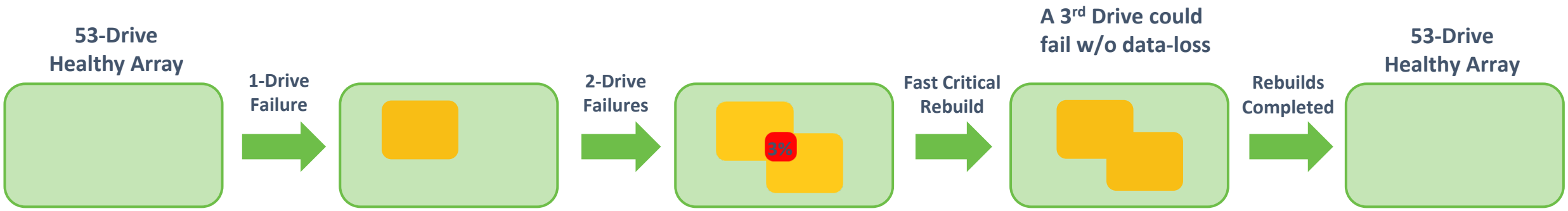
Dispersed Erasure Coding Provides Optimal Data Protection

Erasure Coding dispersed across Stripe Zones w/ contiguous N+2+2 Stripes



Data configuration: 8 or 16 data stripes, 2 parity & 2 spare

Dispersed Erasure Coding Provides Fast Rebuild of Critical Data



- Only **3%** of data is **"Critical"** after a second drive failure
- Protection restored **"5x Faster"** than typical RAID6

"Fault-Tolerant" Stripes
"Degraded" stripes
"Critical" stripes

HDD Head Failure Reman/Logical-Depop Self-Healing

Minimizes e-waste, downtime, and human intervention by allowing HDDs with head failures to be healed and re-introduced to the drive pool

1 Drive posts hard error, controller alerted

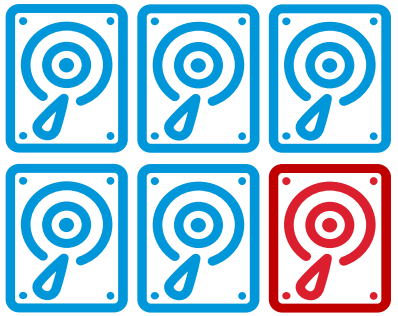
2 Controller offloads data to other drives and removes drive from logical volume

3 Drive and controller diagnose, bypass error, and regenerate drive

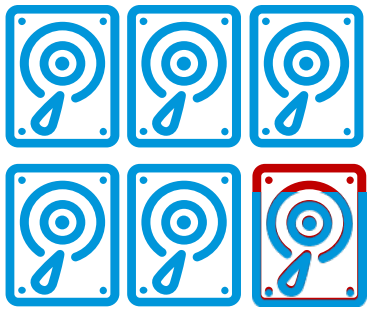
4 Volume is rebalanced with renewed drive Logical Depop

Data Durability & Sustainability

Erasure Coding (EC) in HDD Storage Systems
+ Reman/Logical-Depop technology
reduces human intervention & e-waste thus
enabling HDD Durability & Sustainability



EC Spare Pool:
Drives & Capacity



Reman EC Spare Pool:
Drives & Reduced Capacity

- Extending HDD lifetime saves 275x more CO^2 than recycling and avoids e-waste[1]
- Drive replacements cost data centers over $1,000 per device replacement
- All HDDs feature Instant Secure Erase for easy reuse or retirement
- Additional benefits: Compute & Networking Software licenses savings, faster critical data recovery, higher data resiliency & endurance with minimal performance impact

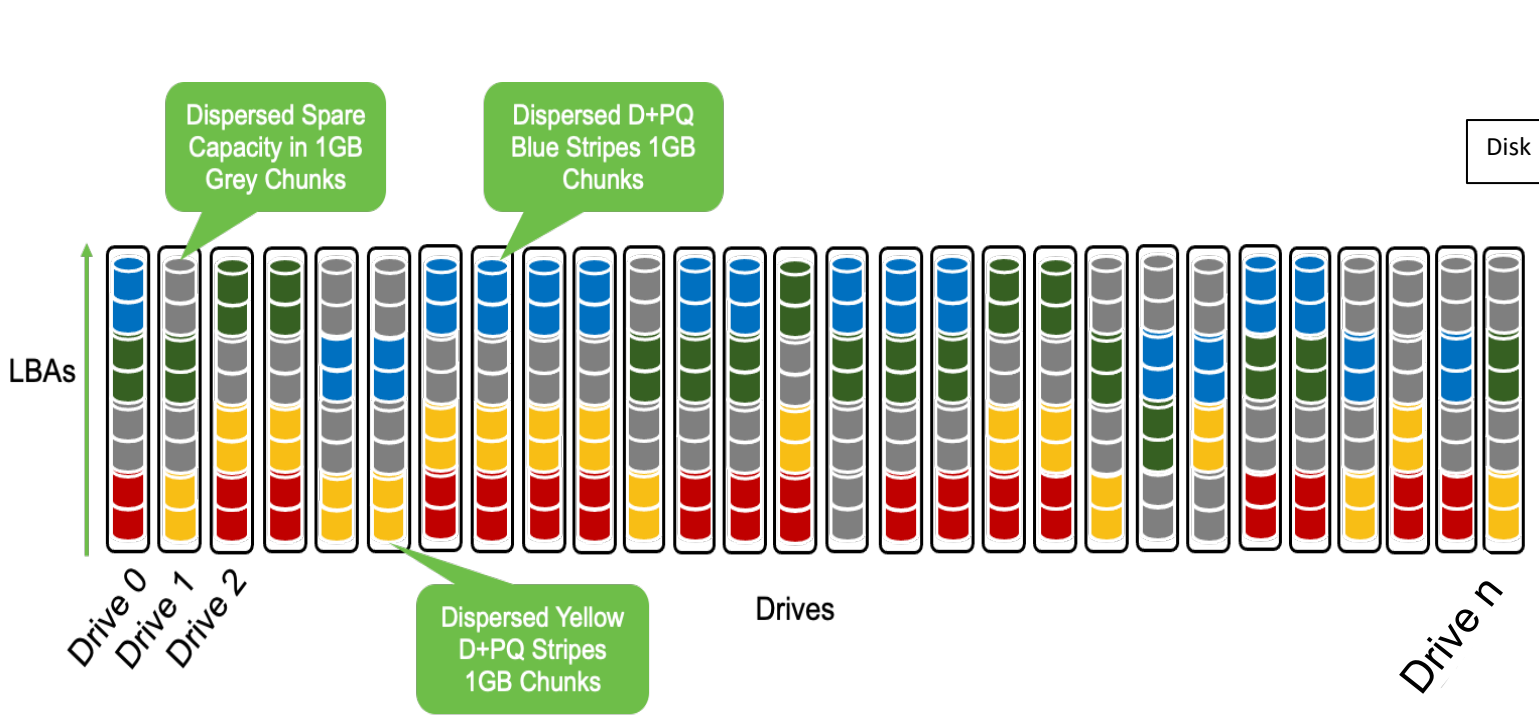
1: Jin, H., Frost, K., Sousa, I., Ghaderi, H., Bevan, A., Zakotnik, M. and Handwerker, C., 2020. Life cycle assessment of emerging technologies on value recovery from Hard Drives. *Resources, Conservation and Recycling*, 157, p.104781.

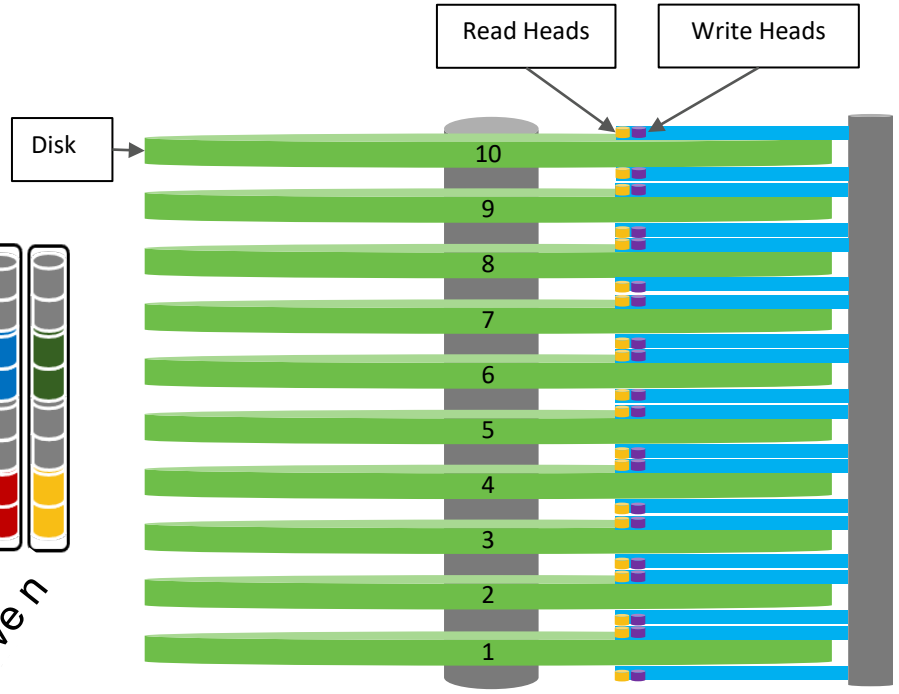
SDC 24

Exabyte Era: Faster Data Recovery with Reman Rebuild

Prerequisite: Erasure Coding (EC) at File System or Storage System and HDD supporting Reman Rebuild

1. Data is distributed in Storage Cluster or Drive Pool with Data Strips, ERC Strips, and Spare capacity to all available hard drives
2. 10x Disk platters + Two Heads per Disk Platter Surface = 20x recording surfaces per hard drive
3. Write-Head failures are much more likely to occur while the Read-head continues to function

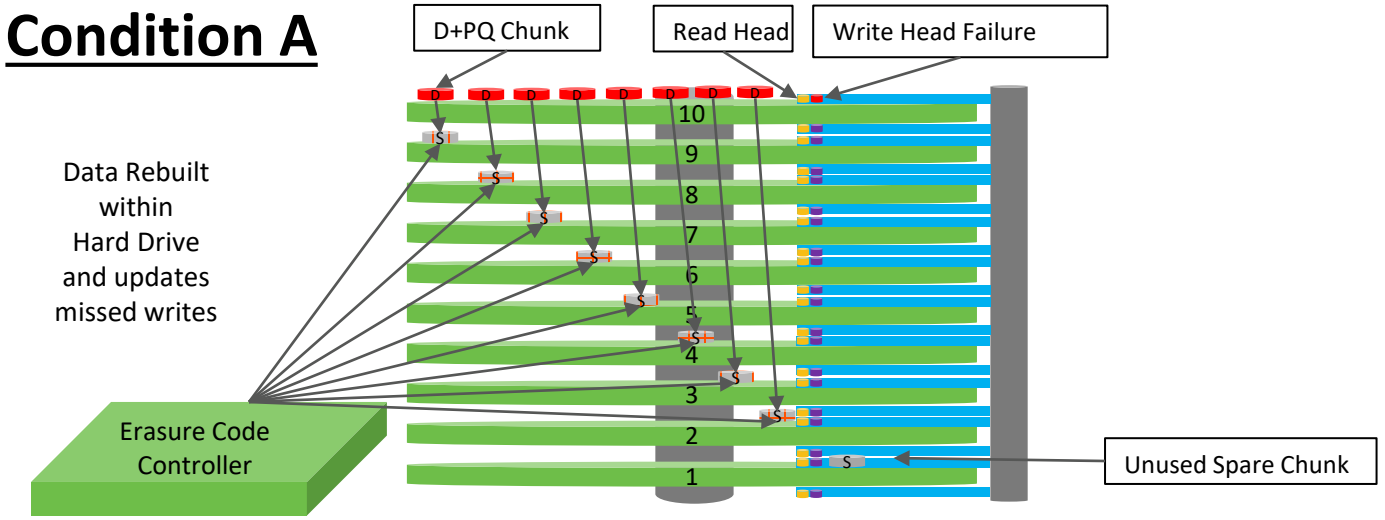




Exabyte Era: Faster Data Recovery with Reman Rebuild

Condition A: Read head functional and HDD has enough spare chunks to self heal

1. EC algorithm uses HDD internal copy function and commands it to copy impacted data chunks to a spare chunk list
2. EC algorithm keeps track of Missed-Writes and updates them within spare chunks
   - Benefit #1: No need for EC to rebuild data from P+Q chunks across entire hard drive pool
   - Benefit #2: Respond to all Read requests normally from impacted drive
   - Benefit #3: Reduce storage fabric congestion



Exabyte Era: Faster Data Recovery with Reman Rebuild

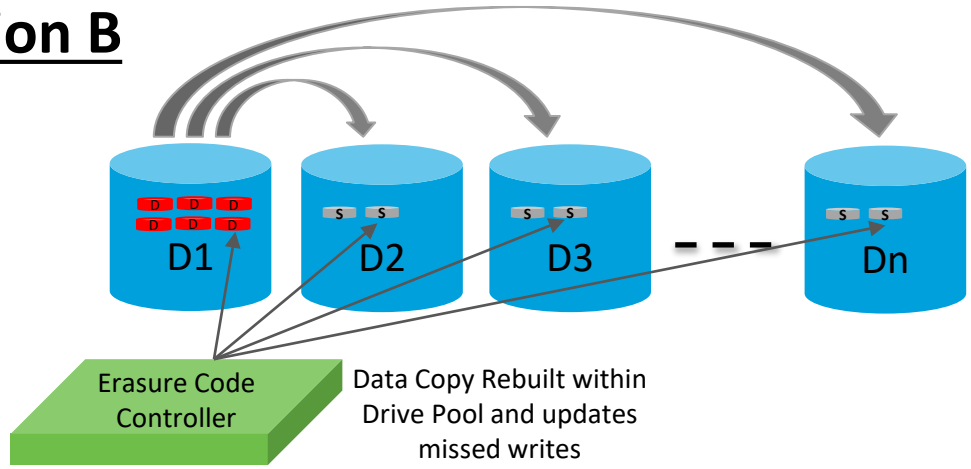
Condition B: Read Head functional and HDD doesn't have spare chunks

1. EC algorithm copies data chunks from impacted surface to available spare chunks within the drive pool
2. EC algorithm keeps track of Missed-Writes and updates them within spare chunks on all drives as it goes

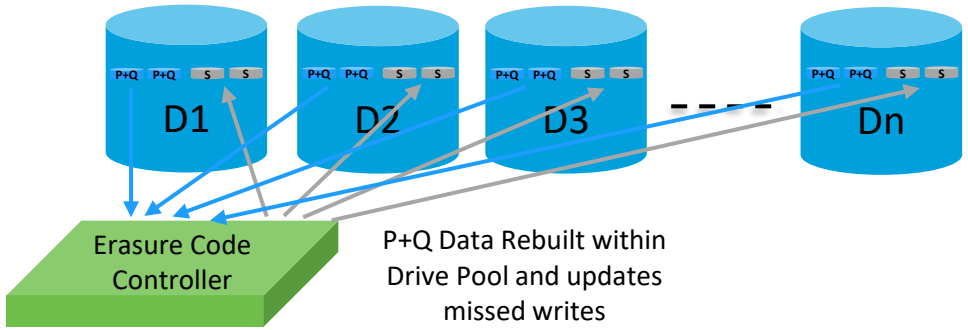
Condition C: Read Head not functional

1. EC algorithm regenerates impacted data chunks from P+Q chunks and writes them to available spare chunks
2. EC algorithm keeps track of Missed-Writes and updates them within spare chunks on all drives as it goes

Condition B



Condition C



Intelligent Resource Virtualization Key Takeaways

- As Data Center Storage matures, their Fabrics are getting too congested with EC, Replication and Rebuild traffic, one way to reduce traffic without compromising on data capacity, resiliency or availability is with Intelligent-Fast-EC-Rebuild & Head-Failure Self-Healing-Reman at the drive and enclosure level

The Myths

"HDDs are too slow to handle the data fragmentation effects of Data-reduction"

"Flash uses far less physical storage to meet the logical capacity demands due to Data-Reduction"

SDC 24

Myths Busted!

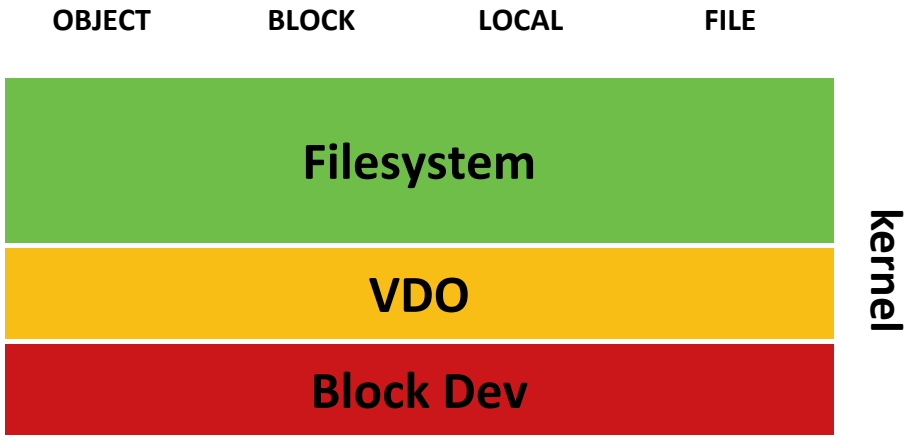
The **Perceived** HDD performance issue is a **Data-Reduction Application Defect**

Some Data-Reduction Applications' **Inability to coalesce & streamline outputs** are the **Culprit**

Typical Culprit Linux Data-Reduction

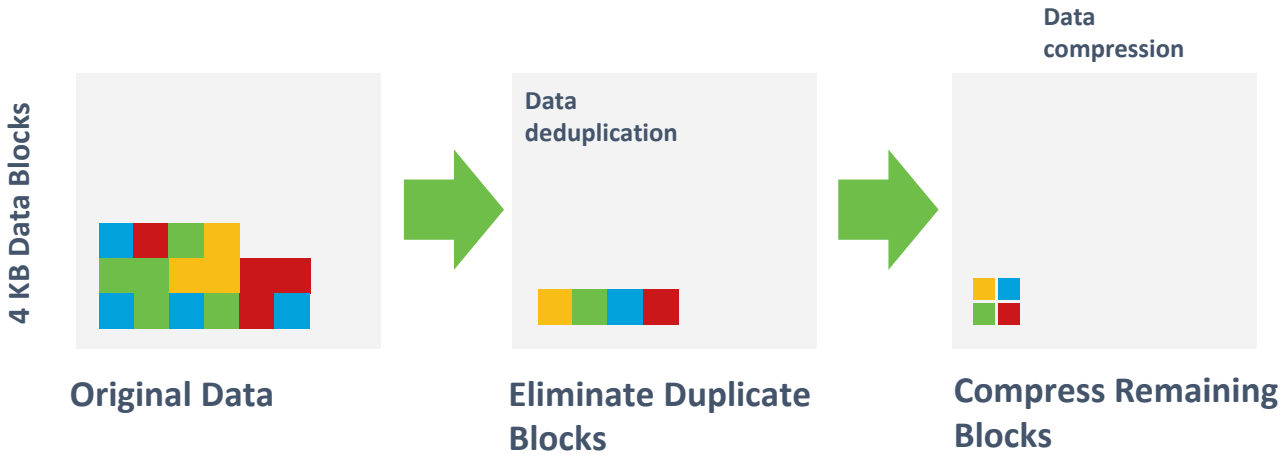
Degrades HDD performance | Increases SSD Write-Amplification

Virtual Data Optimizer (VDO)



Local or Networked Storage

VDO Data Reduction Processing
Dedup 2:1 + Compression 2:1 = 4:1 AVG Reduction



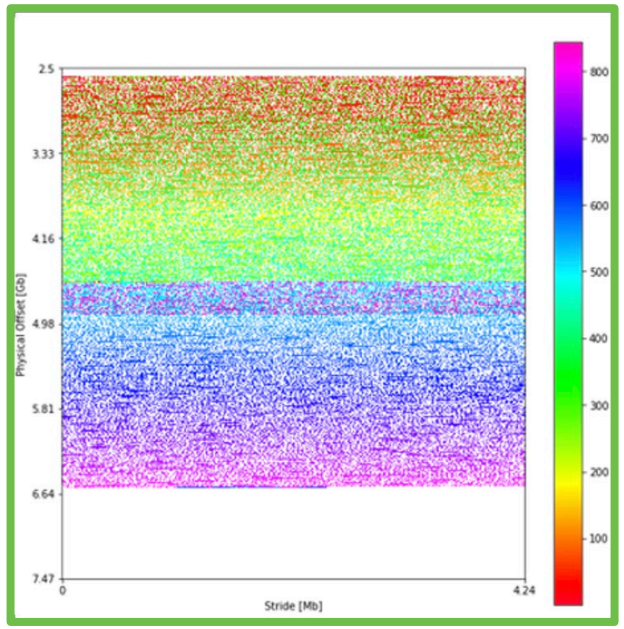
Source: https://www.redhat.com/en/blog/introducing-virtual-data-optimizer-reduce-cloud-and-premise-storage-costs

HDD Vendor Sponsored VDO Utility Optimizations

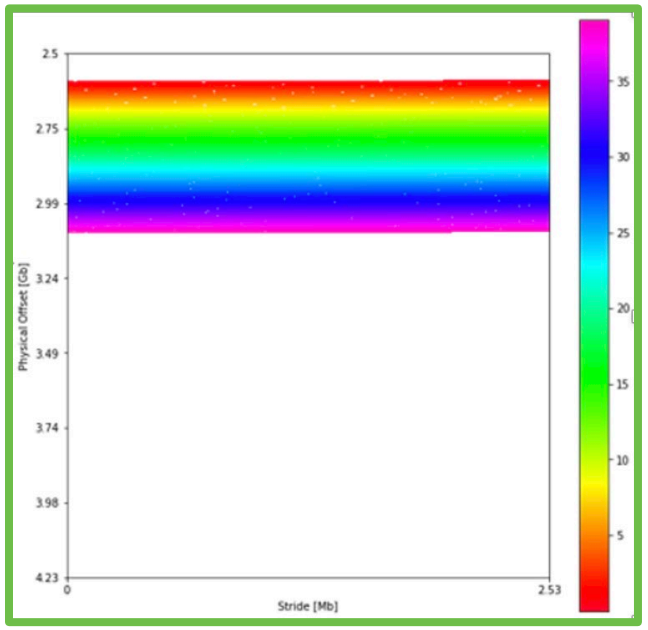
Architectural Enhancement Summary

- Delayed block-allocation
- Prevented need of duplicate checks
- Reduced regions between duplicates
- Cached block & journal maps
- Added read prefetch cache

Original



Optimized



JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

U of Mainz: Patrick Raaf
Advisor: Prof. Dr. André Brinkmann

SDC 24

VDO Optimization HDD Performance Impact

Results with 4:1 Data-Reduction

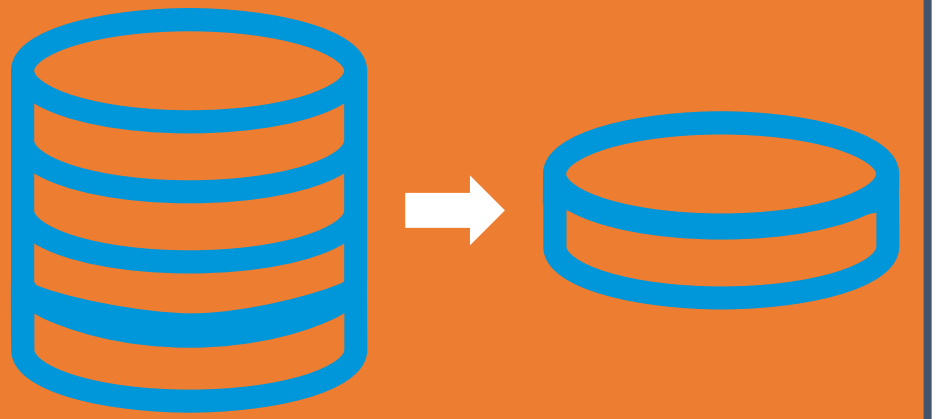
18x
Sequential Write Faster

86x
Random Write Faster

1.5x
Seq/Rand Read Faster

HDD Post Data-Reduction Performance

4:1
Reduction



1.5x Normal HDD
Speed

50%
faster

Results with 4:1 Data-Reduction

Data-Reduction Key Takeaways & Observations

- Regardless of what software or hardware data-reduction tools you use, always make sure to characterize the downstream workload hitting your storage devices
- Data-Reduction output workload streamlining/sequentialization benefit backend SSD & HDD storage devices alike
- Powerful HW fingerprinting and Hashing for Dedupe function as well as lossless HW Compression/Decompression improve TCO & Performance by eliminating CPU overhead and application checks for False-duplicates & Data-integrity
- Reduced Sequential Data should always lead to better Performance & TCO

Please take a moment to rate this session.

Your feedback is important to us.

