SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

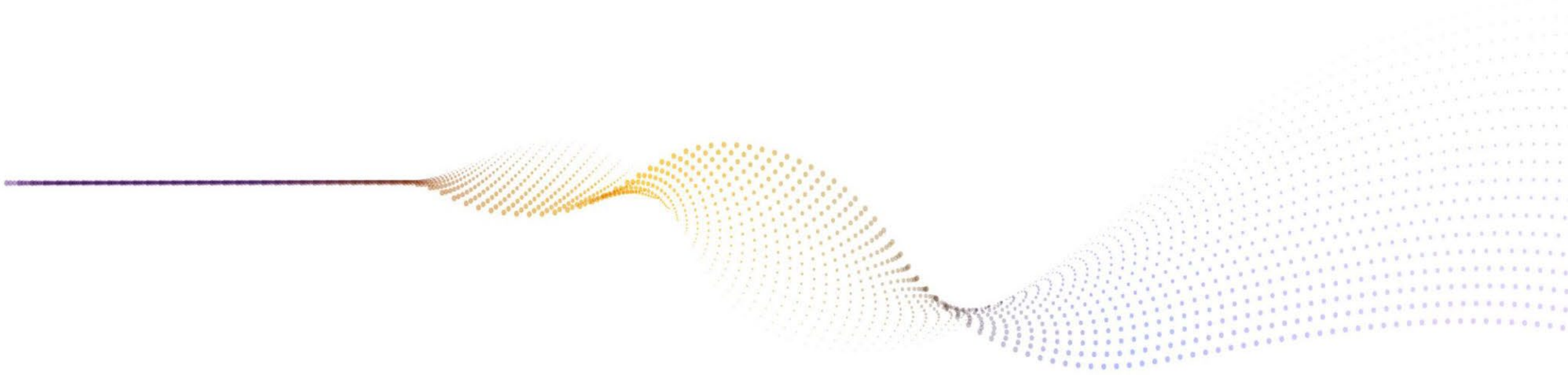
Hyatt Regency Santa Clara, CA
September 15-17, 2025

Drive Regeneration in Action: Enhancing Fault Tolerance in Datacenters

Seagate Technology
Curtis Stevens - Strategist
Dave Craton – Principal Product Manager

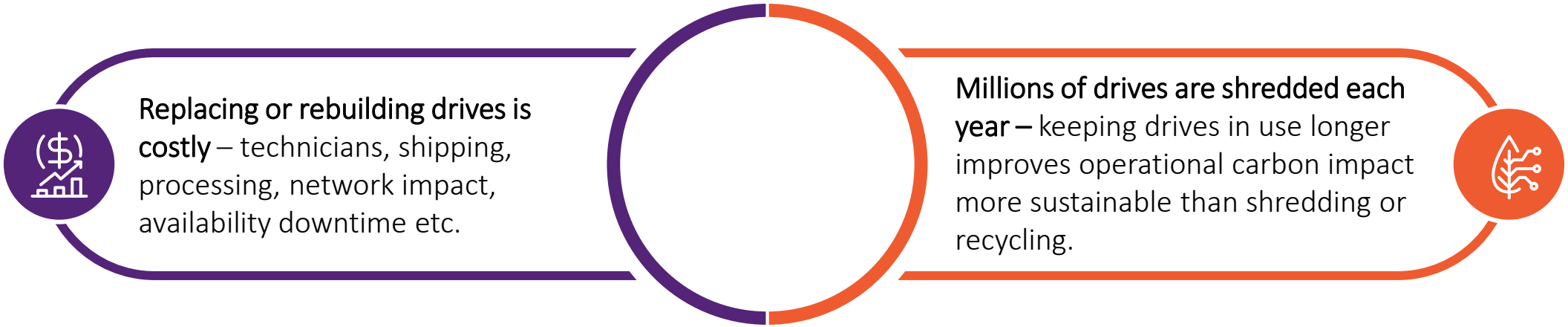
Agenda

1. Datacenter challenges with fault tolerance at scale
2. Expanding HDD fault tolerance from sector to head/surface level
3. Data-Safe Drive Regeneration / Depop Options



Storage Reliability Challenges

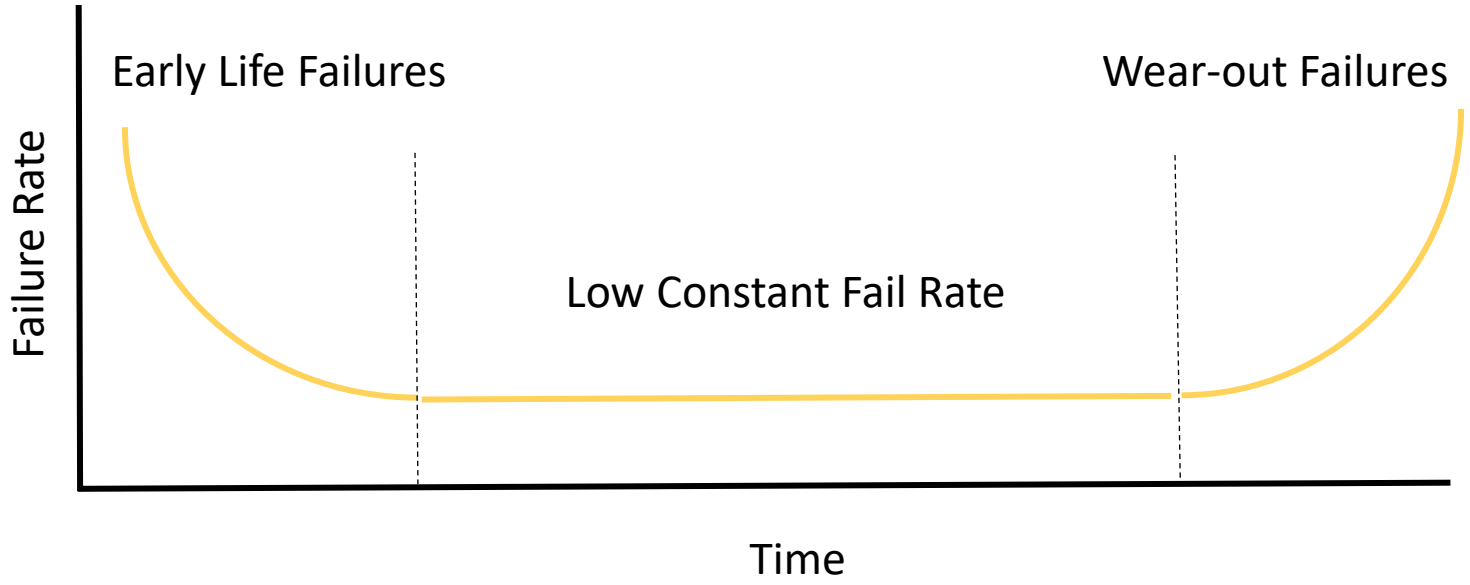
Key Storage Reliability Challenges



Enterprise HDD Reliability - Baseline

Standard Enterprise HDD Reliability:

- 2.5M MTBF = 0.35% Annualized Failure Rate
- 1.75% CFR (5-year warranty)
- 8760 hours/year



Drives in a server or storage system in the datacenter may be rejected at a higher rate than baseline.

Examples:

- System-wide offline events, performance variations, excursions, etc.
- Some heavy datacenter workloads and high temperatures may accelerate some HDD fail modes.

Typical rejection target < 1% / yr

Categorizing Production HDD Rejections

GROUP 1: NTF

No trouble found with a return to service recommendation. Some portion of these may fail again and land in a category below.

GROUP 2: SINGLE-HEAD FAILURES

May be resolved by **Drive Regeneration / Depop.** Can represent 40%+ of failures.

GROUP 3:

Typically the smallest group of failures recommended for decommissioning.

On 20 head/10 disk drives, 95%+ of the good capacity may be preserved rather than discarded.

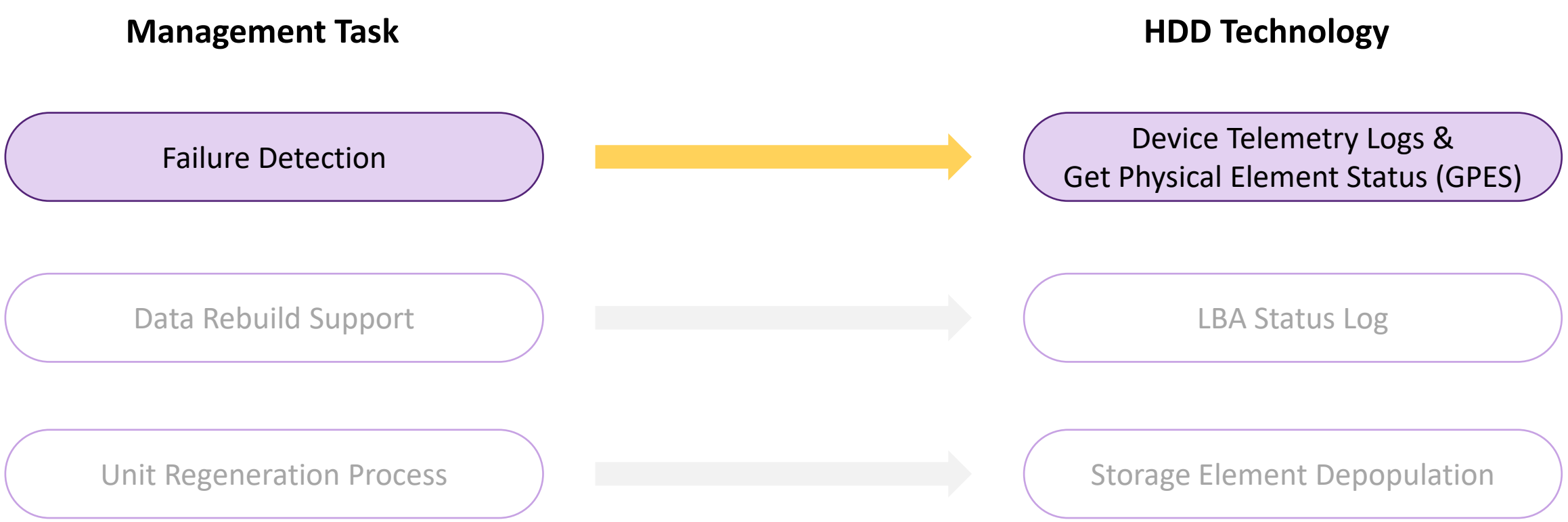
Hundreds or thousands of rebuilds may be significantly reduced or prevented entirely.

Expanding Fault Tolerance From Sector to Disk

Methods for Managing Hard Drive Failures

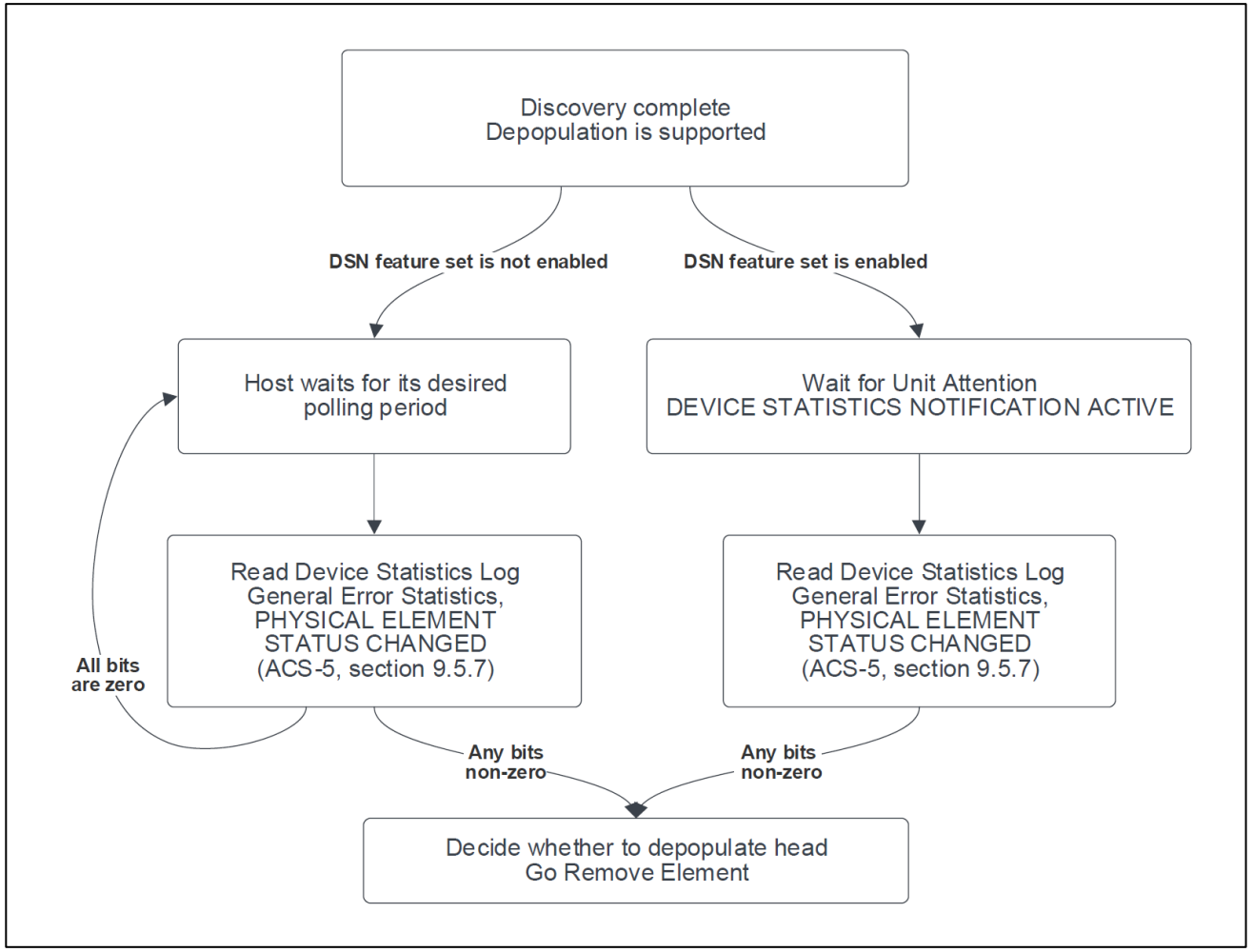
Management Task	HDD Technology
Failure Detection	Device Telemetry Logs & Get Physical Element Status (GPES)
Data Rebuild Support	LBA Status Log
Unit Regeneration Process	Storage Element Depopulation

Methods for Managing Hard Drive Failures

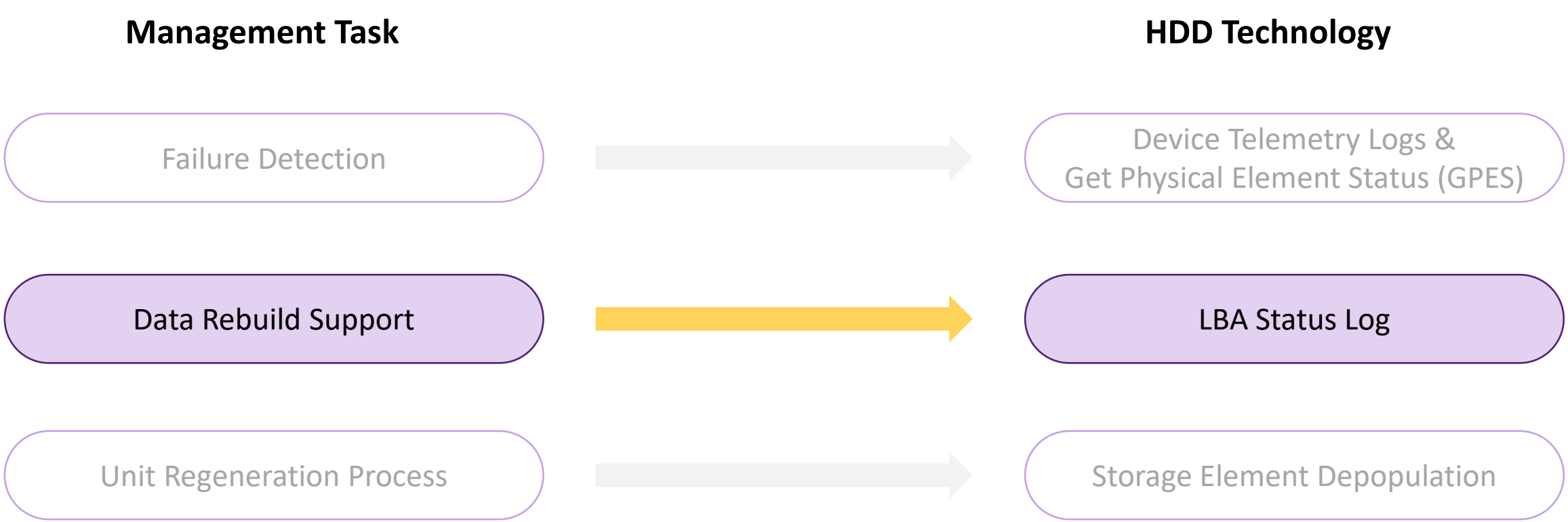


Hard Drive Telemetry & GPES

- Failures are detected through a combination of:
  - System monitoring
  - Drive telemetry data
  - **Get Physical Element Status (GPES)** field, defined in ACS
- GPES may be polled by the host, or set up for asynchronous alert using the Device Statistics Notification (DSN) feature on supported devices
- GPES is critical to managing head-level faults, using drive algorithms to detect and report which recording head has degraded and may be depopulated



Methods for Managing Hard Drive Failures



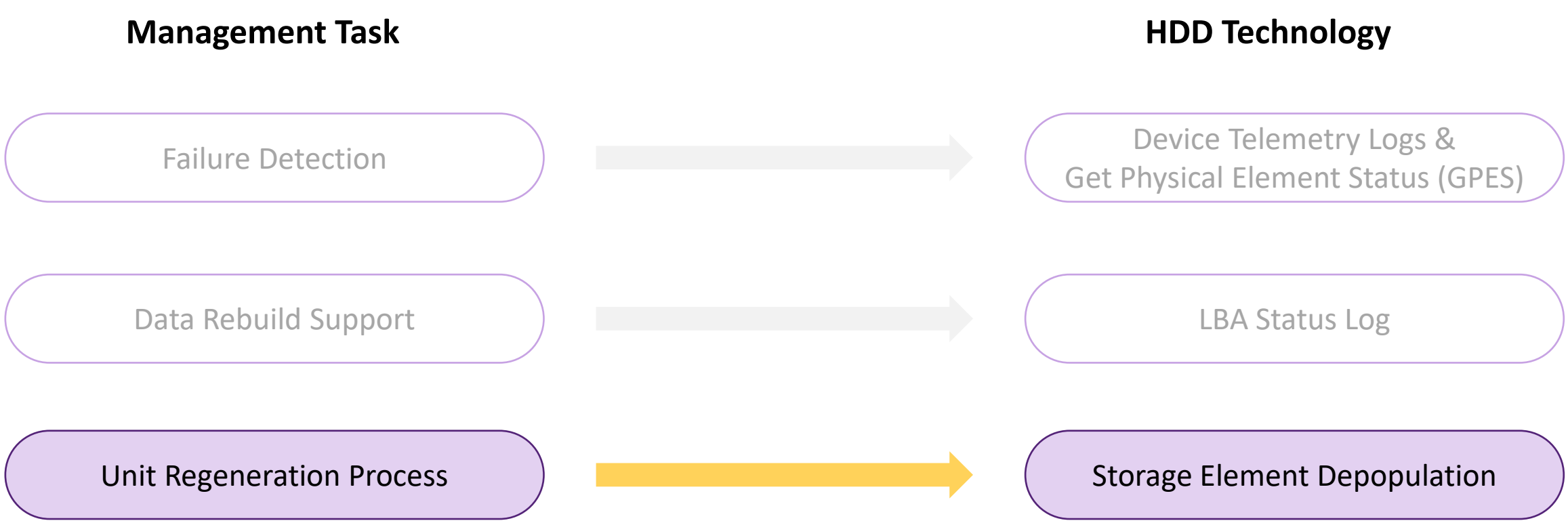
LBA Status Log

Bit Byte	7	6	5	4	3	2	1	0
0	OPERATION CODE (9Eh)							
1	Reserved			SERVICE ACTION (12h)				
2	(MSB)							
...	STARTING LOGICAL BLOCK ADDRESS							
9								(LSB)
10	(MSB)							
...	ALLOCATION LENGTH							
13								(LSB)
14	REPORT TYPE							
15	CONTROL							

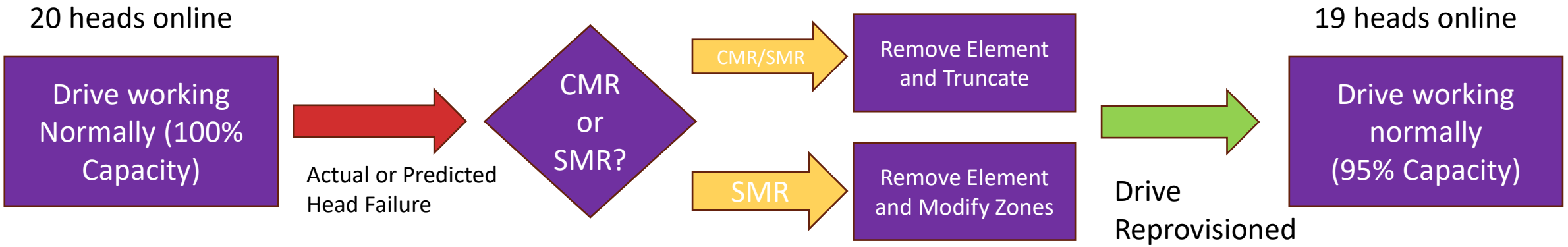
Get LBA Status (16) Command

- The LBA Status Log is a drive-reported list of LBAs and their status details (including provisioning and error status)
- Hosts may use the log to understand which LBAs are healthy or unhealthy, providing additional options for managing issues:
  - Continue using the good LBAs in place without depop
  - Speed up rebuild process by moving good LBAs prior to depop

Methods for Managing Hard Drive Failures



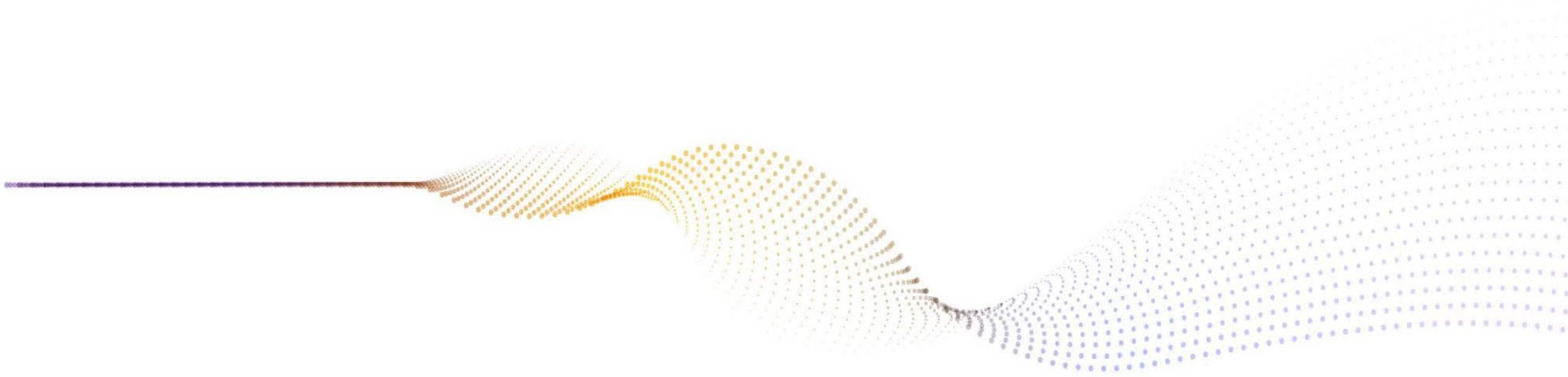
Storage Element Depopulation



- **Storage Element Depopulation** allows for removal of a head/surface from the addressable LBA space
- After the depop process, the addressable LBA space is reduced, and the drive comes back online at 95% of its original capacity (20 head drive)
- The change in capacity makes the process a better fit for software defined and erasure coded redundancy than HW RAID.

Feature Support

Feature	Function	Ecosystem Support
Standard Offline Regen (Remove Element and Truncate)	Drive completes full format, re-provisioned to lower capacity	Supported in Linux on HDDs today
Data-Safe Regen – SMR/ZBD Only (Remove Element and Modify Zones)	Leaves good data intact. Zones with failed heads are marked offline, re-provisioned to lower capacity.	Supported in Linux on SMR HDDs today
LBA Status Log	LBA log indicating which LBAs are no longer available due to head failure. Allows continued use of good LBAs for use or rebuild before issuing remove element commands.	In development



Data-Safe Drive Regeneration Options

The Importance of Data-Safe Drive Regeneration

Objective: Keep the good data valid, eliminating reformat and drive replacement steps.

Improve storage availability

- Typical datacenter has 95% storage availability – 5% unavailable
- Slots may go for weeks or months before replacement storage is added
- Each drive has a replacement cost beyond the cost of the storage
- Sustainability – shredding or recycling failed drives carry significant cost

Prevent Rebuilds

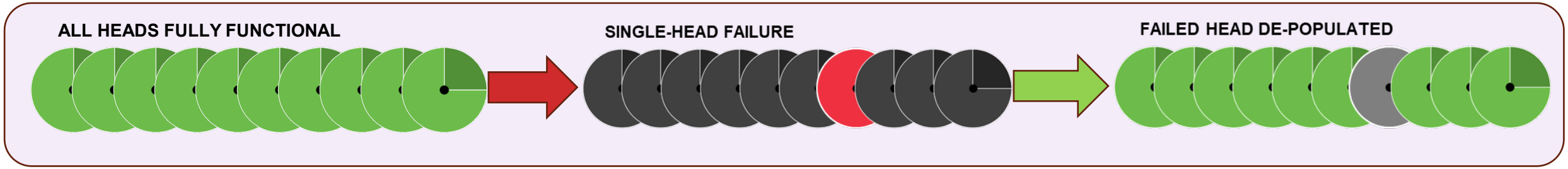
- Offline methods are already available, but they destroy all the existing data triggering a rebuild operation
- SMR drives have online options, need to expand to cover CMR

Advantages for Sealed Datacenters

- Keeping data online and available extends the useful life of sealed datacenters without costly maintenance

Existing Zoned-based Regen/Depop

For systems with *Host Managed Shingled Magnetic Recording (SMR) only*



20 heads online

SMR Drive working Normally (100% Capacity)

Actual or Predicted Head Failure

Host backs up content of target zones

Remove Element and Modify Zones

Drive Reprovisioned

19 heads online

Drive working Normally (95% Capacity)

- Once the head removal is complete, REPORT ZONES EXT is used to determine which LBAs were removed from the zones marked OFFLINE
- Existing data in remaining ONLINE zones are unaffected – no format required

Expanding Zoned-based Depopulation to CMR

Solution

Reporting zone structure would need to be enabled for standard CMR HDD's

- All zones need to be the same size; need to agree on a standard zone size, for SMR it's 256MB
- Drive would still report head health as it does today
- When a head becomes unhealthy, the zone report indicates the media that would go offline line when the head fails
- Based on this information, the host moves the affected data elsewhere
- The host then stops using those sectors.

Benefits

Keeps data online, preventing rebuilds for 95% of capacity

Works better for RAID type arrangements, especially if your stripe size is a multiple of the zone size...

Smaller zone sizes allow the HDD to wear-level its data better, allowing for more consistent performance

Challenges

The granularity of this data (zones) is a bit large

- There will be collateral damage from zones that span heads

Average LBA loss is approx. 7% (due to collateral damage)

Use LBA Status Log to Extend Use

Solution

When the drive reports a head is unhealthy, the host looks to the LBA Status log to find a list of the affected LBAs. The host may then relocate the good data.

- Consider using new command WRITE GATHERED EXT to move data without using the datacenter infrastructure
  - This command takes a list of LBA ranges and copies the data to a new location
  - Normal read and write commands may also be used, but these will use the datacenter infrastructure to move the data
- The host may then avoid using the LBAs that are about to fail

Benefits

The average LBA loss using this method is 5% and collateral damage is minimized

Works better for filesystems with a cluster size under 1MB

Challenges

Can result in millions of entries, potentially challenging host-management capabilities.

Drive Managed Regeneration / Depopulation

Based on the thin provisioning concept – Currently coined *Optimized Provisioning*

Solution

When a head becomes unhealthy or fails

- The drive automatically redirects LBA accesses targeted at the failing head to a good head
- The host is informed that a head is unhealthy, but no action needs to be taken by the host
- Operation continues normally

Requires the drive to have available free space

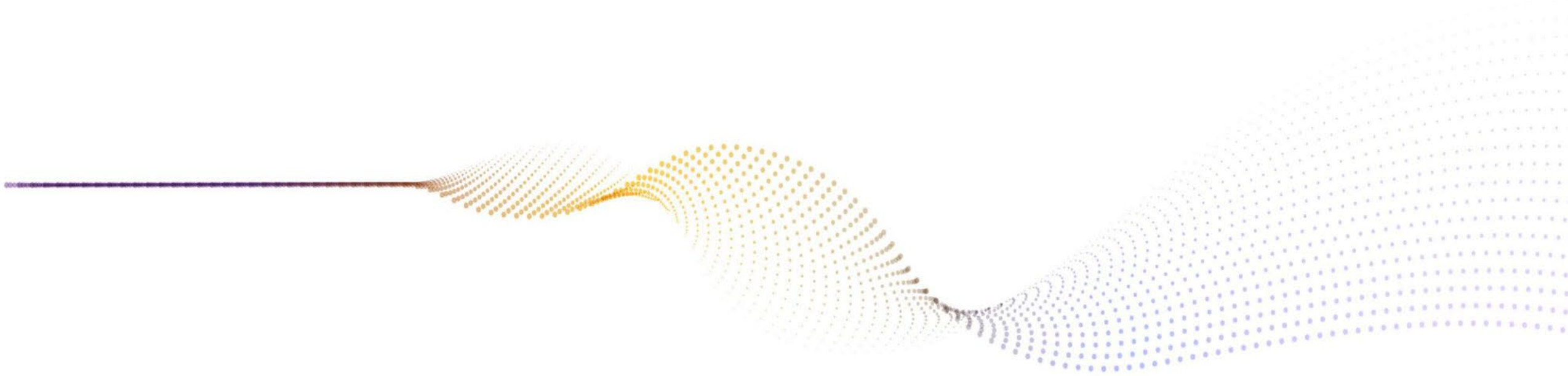
- The host issues trim commands to inform the drive which LBAs are unused
- Optimized provisioning reports when the free media (media the LBAs point to) has dropped to a critical level

Benefits

Depopulating a head becomes seamless
Host is notified, but no action is required
Works for both RAID and erasure coded systems

Challenges

The drive needs to receive trim commands
Free space is required to remap the LBAs when a head fails
Access times to remapped LBAs may change
    May get faster or slower, but remain within the average access times of the drive.



Thank You!

