SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

# Technology Breakthroughs in Mass Capacity Storage


## Seagate Technology

Dave Craton, Joel Schipper

Sr. Staff Product Manager(s) - Technical Marketing

# Agenda

1. Hard Drive Technology Breakthroughs
   - The Scale of HDDs for Datacenter and AI
   - HAMR Technology

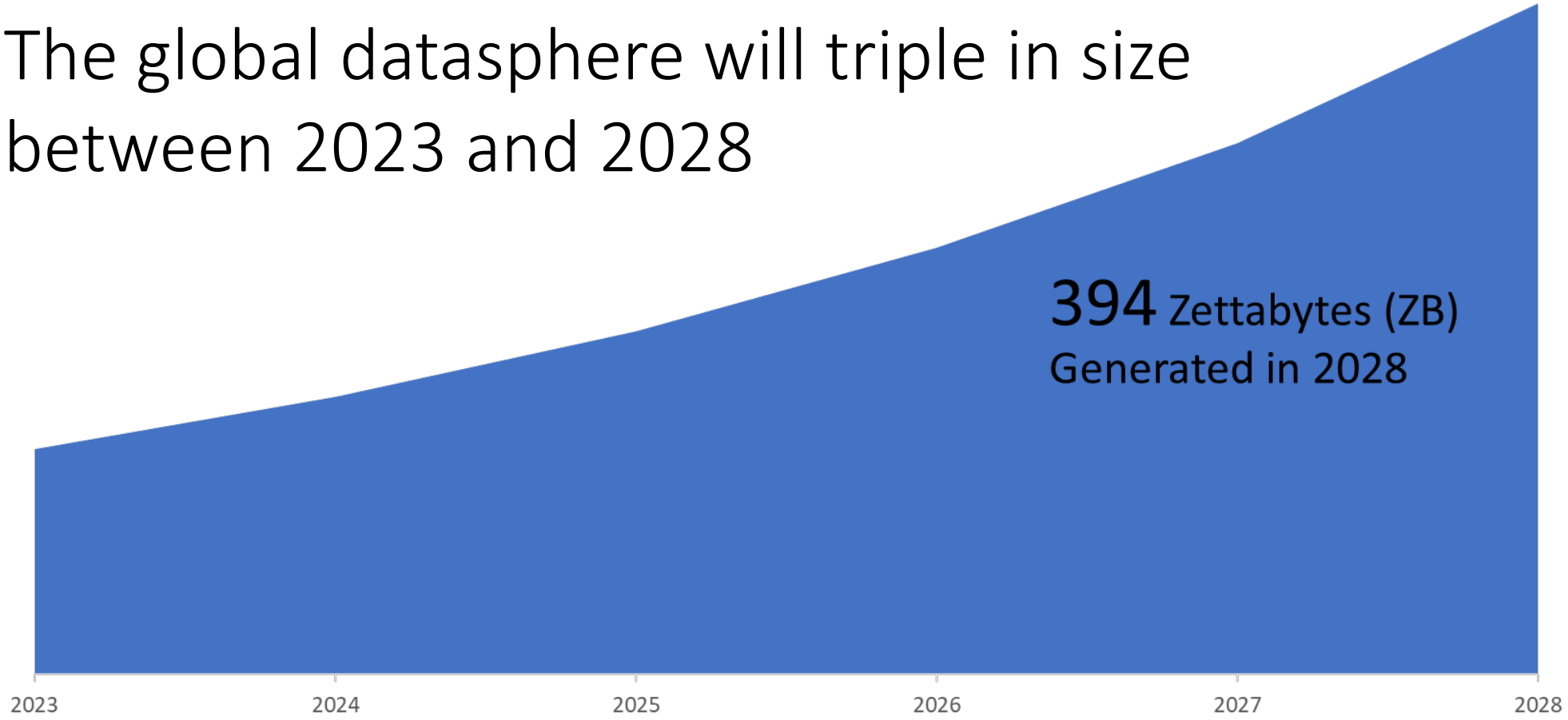
2. Storage Reliability & Sustainability Innovation
   - Drive Regeneration Features (Storage Element Depopulation)

3. Optimizing Performance
   - Priority Management with Command Duration Limits (CDL)

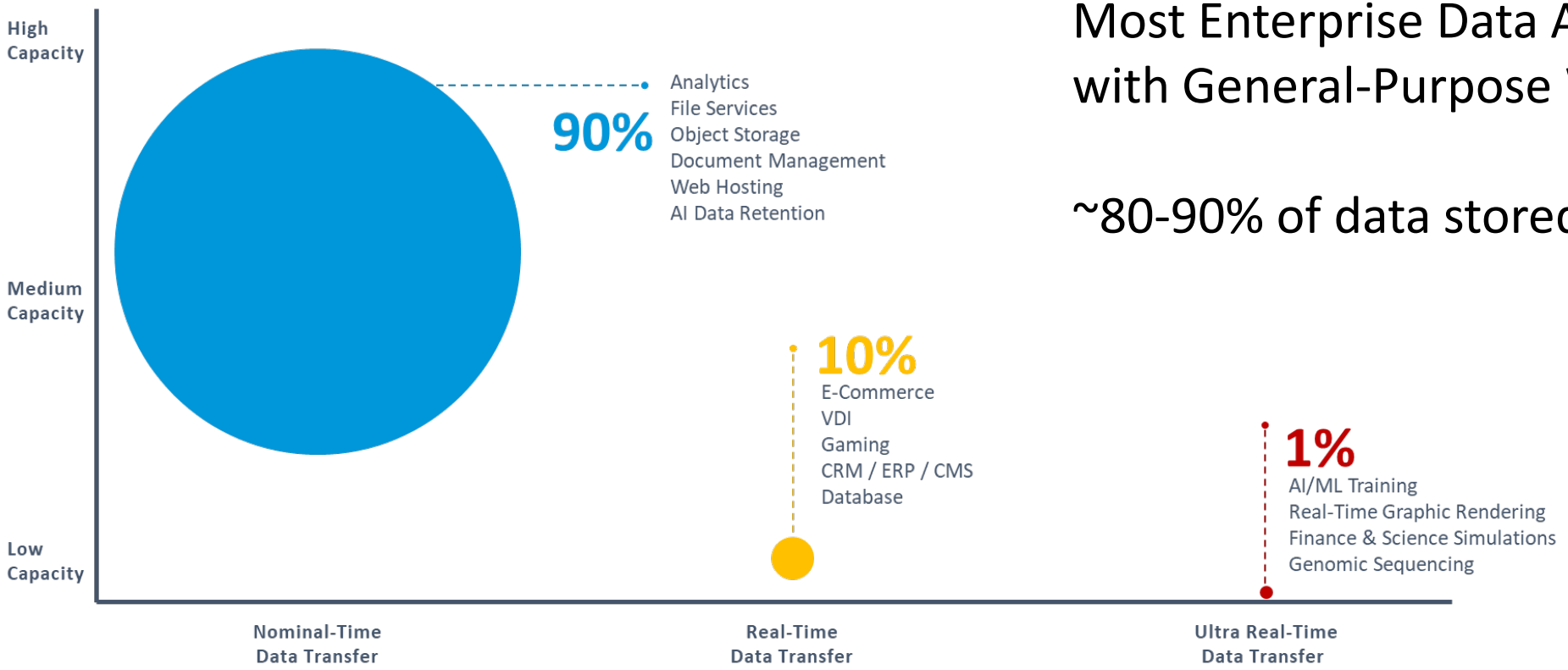
# Global Datasphere Growth

The global datasphere will triple in size between 2023 and 2028



# Mapping Enterprise Data




Most Enterprise Data Associated with General-Purpose Workloads

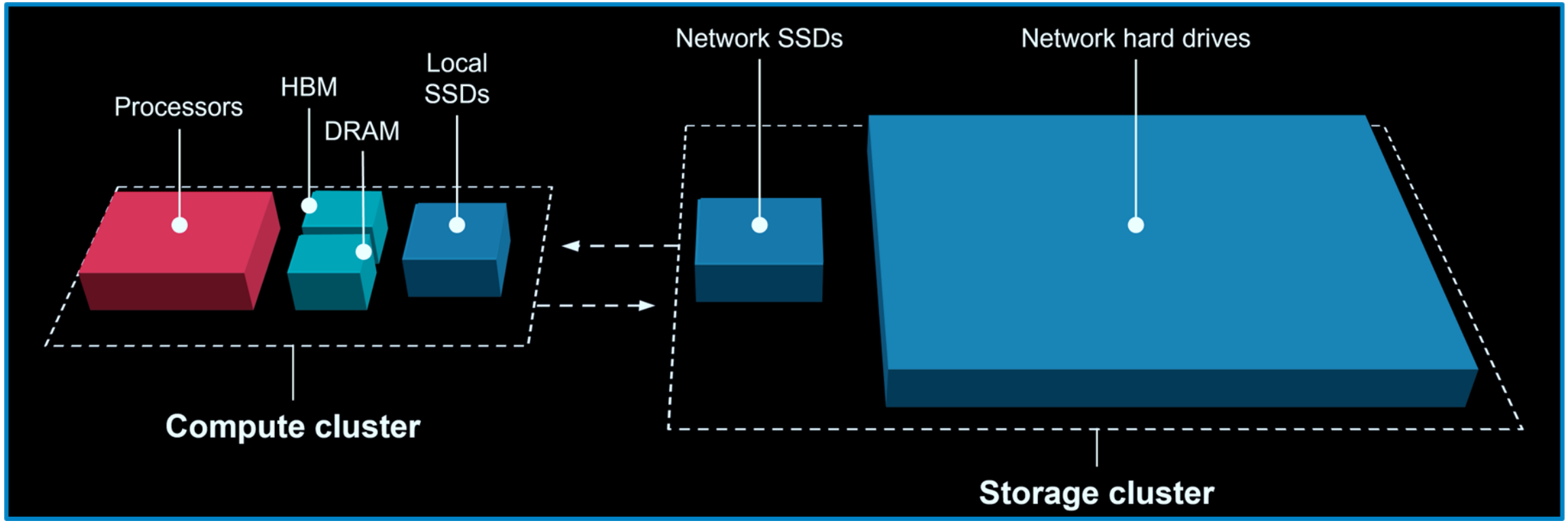
~80-90% of data stored on HDDs

Image: Mapping enterprise workloads to EBs stored.

Source: Seagate analysis of IDC, (May 2023). *Streaming and Real-Time Data Redefined and Superimposed on IDC's Global DataSphere*, 2022 (May 2023).

# AI Data Starts and Ends on Mass-Capacity Storage




| | | |
|---|---|---|
| Performance | PB/s ← | MB/s |
| Scale of Deployment | TB → | Multi EB |

* Graphic: Seagate

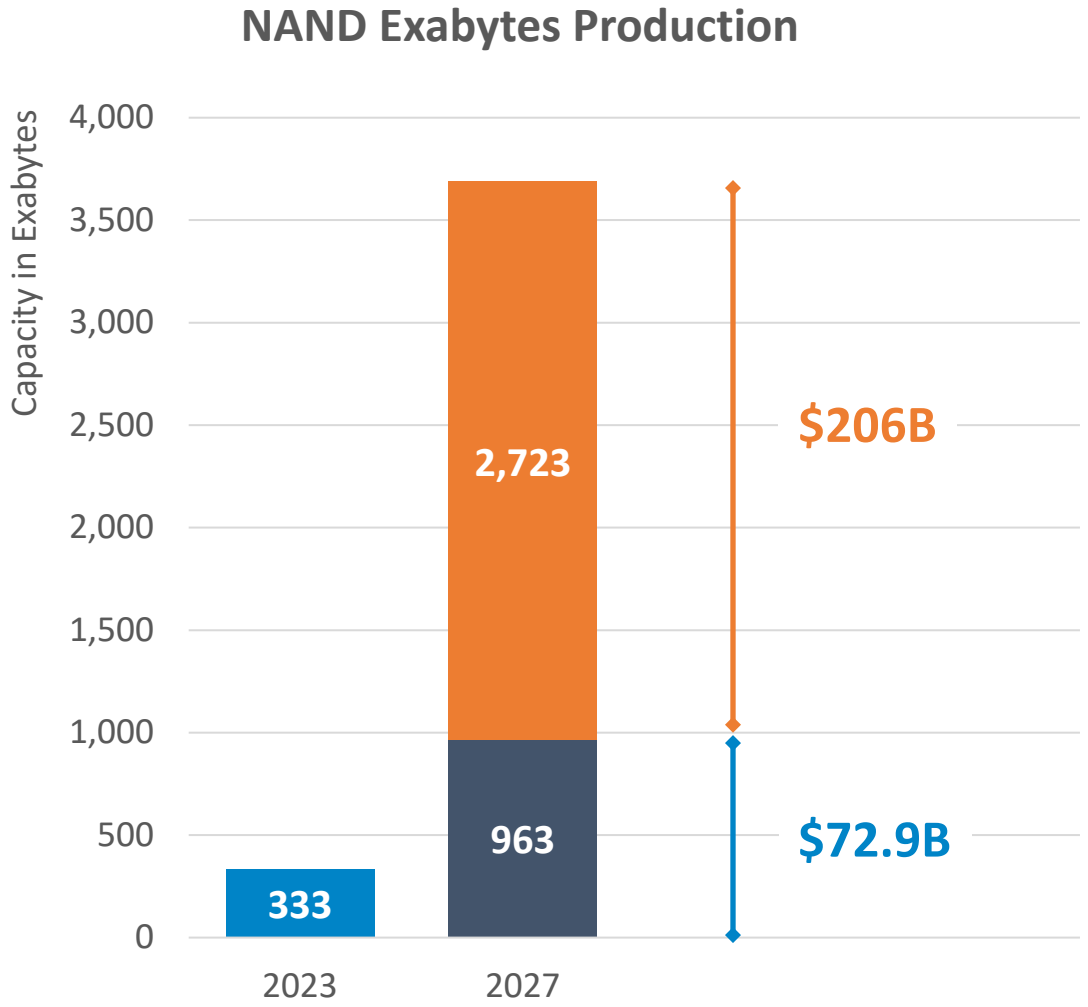
 

# Replacing all HDD-Exabytes w/ NAND is Cost-Prohibitive

**$10 investment required for a $1 return**

- **Exabytes the NAND industry produced**
- **Exabytes the NAND industry is projected**
- **Exabytes the NAND industry would need to produce, if they were to replace all HDDs**

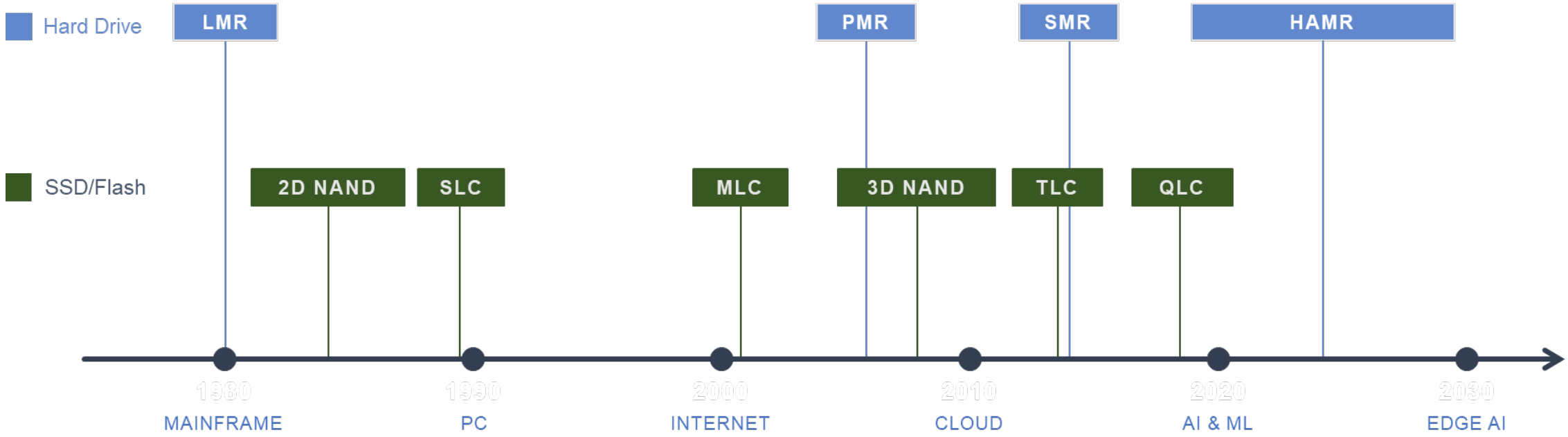
Image: Producing enough NAND Ebs to fulfill hard drive demand would be cost-prohibitive.
Sources: Seagate analysis based on NAND Flash Platinum Datasheet by TrendForce and IDC Global StoageSphere Forecast, 2023-2027 Doc. #US50851423, June 2023.



# Enabling the World's Digital Infrastructure

Hard drives and SDDs have upheld the world's digital infrastructure for decades—and will continue to coexist in the market.



# HAMR Technology

# What makes up a Hard Drive?

- A hard drive is made up of many precision parts with the head and the disk being two of the most important.
- Semiconductor-type wafer processing allow for economical manufacturing of heads that read/write data onto the disk.
- Wafer --> Bars --> Sliders --> Head Gimbal Assembly (HGA) --> Head Stack Assembly (HSA)
- HSA + Media + Electronics/Other Components + Case = HDD!

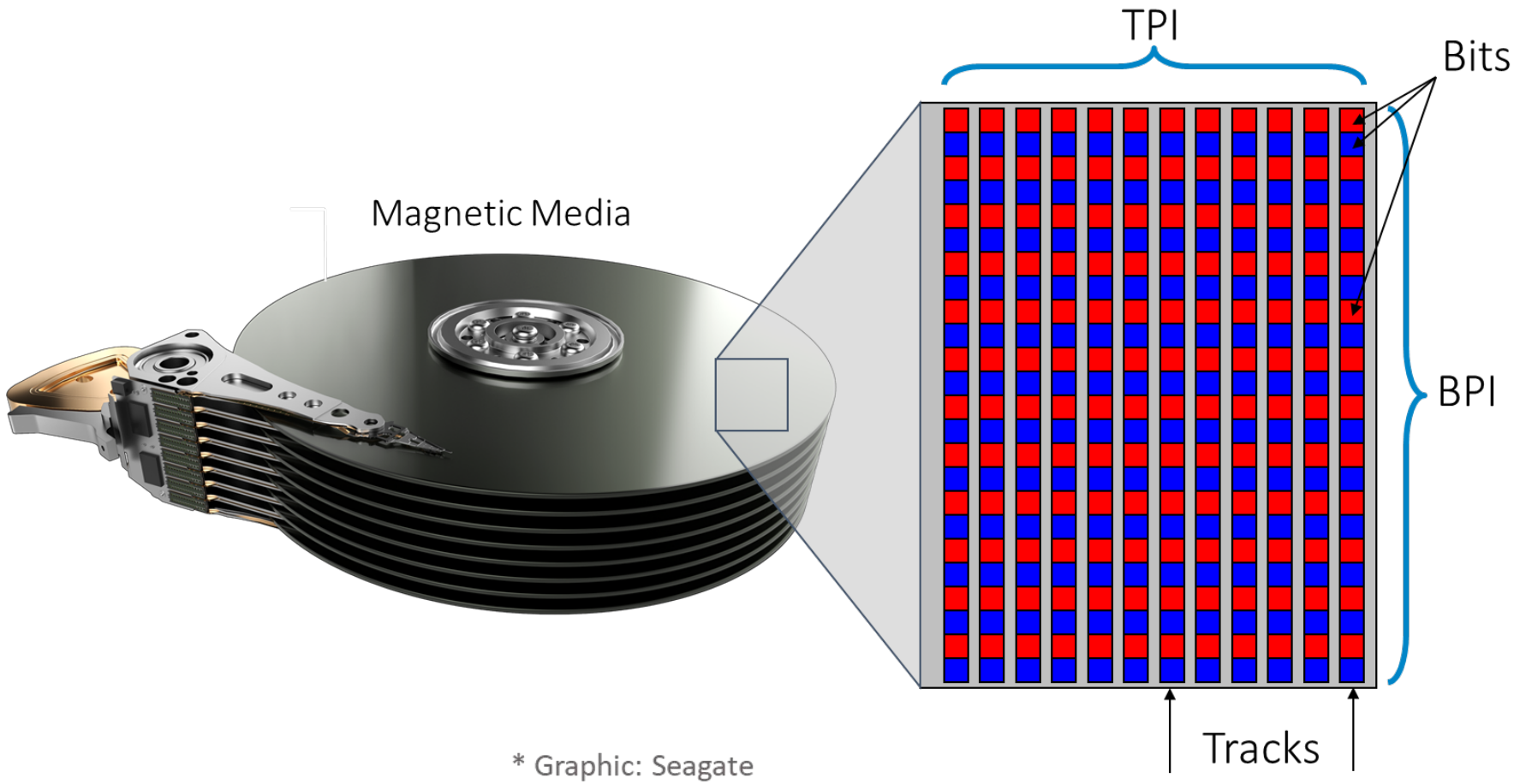

NEW! ← Key Areas of HAMR Innovation → NEW!

# HDD Areal Density

**Areal Density (AD) = Linear Density x Track Density**

(bits/in$^2$) (bits/in) (tracks/in)

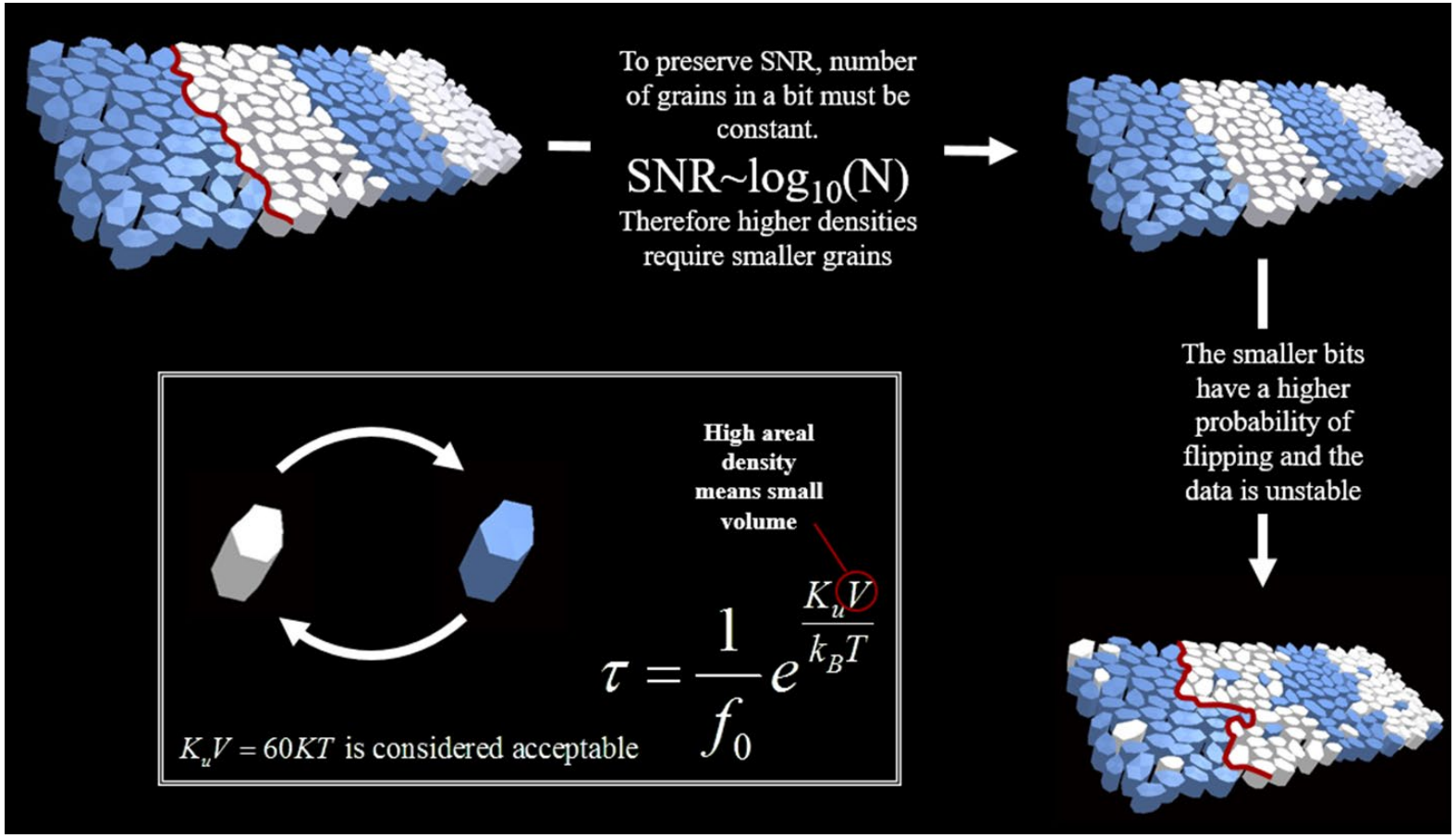



* Graphic: Seagate

# The Challenge with Capacity Growth

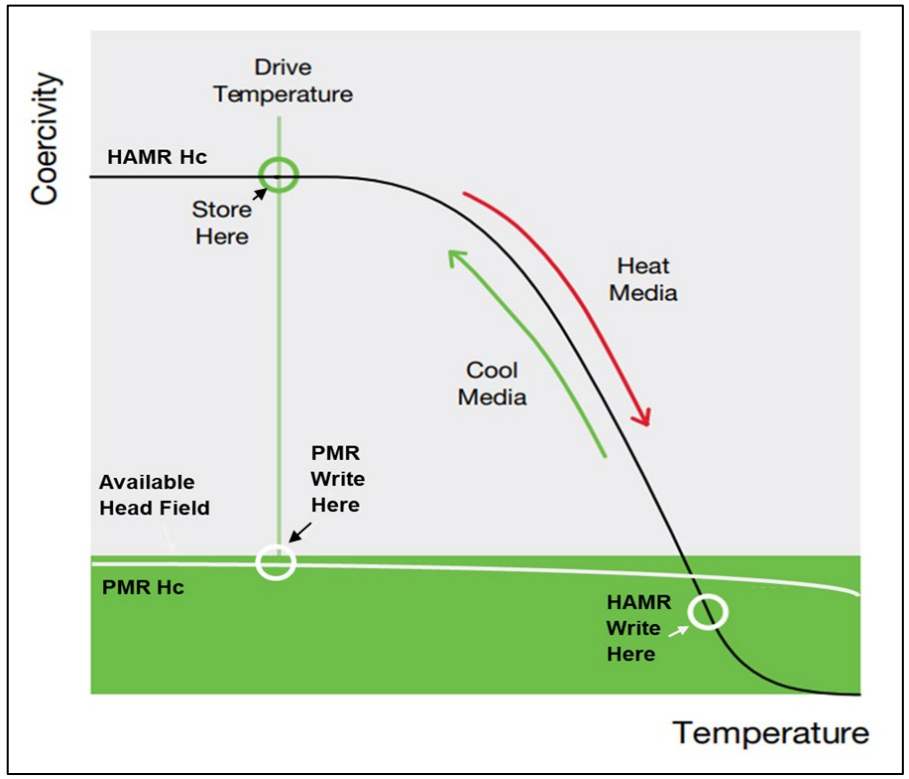
- Increased Density = Smaller Bits
- Smaller Bits = Less Thermal Stability

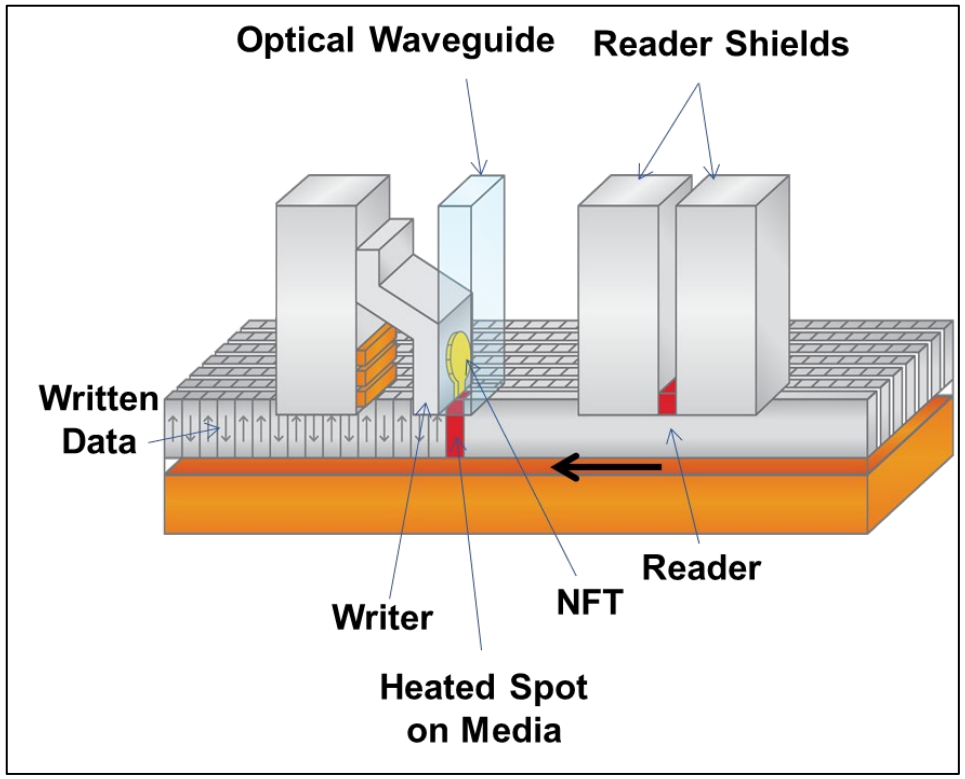


# The Solution to Capacity Growth: HAMR!

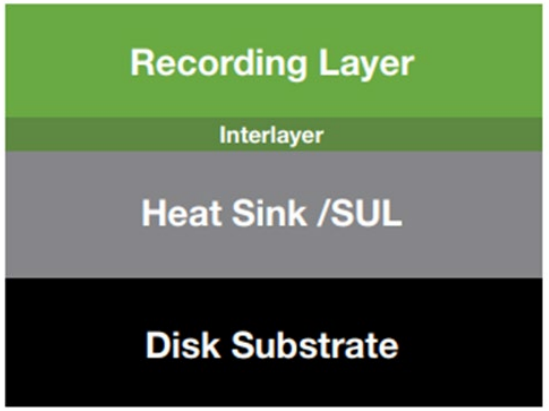
- **New Media**: High coercivity media with FePt lattice gives thermal stability and can be softened during the write process
- **New Heads**: Laser & Near-field transducer heats write spot on media to heat, write, and cool in a matter of nanoseconds

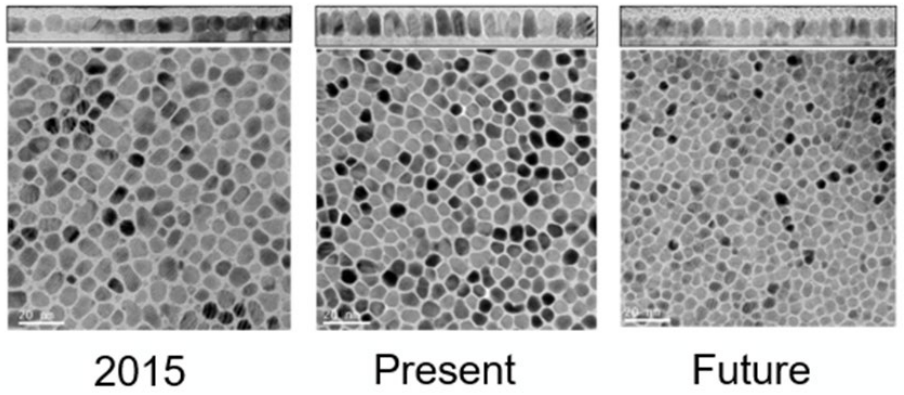


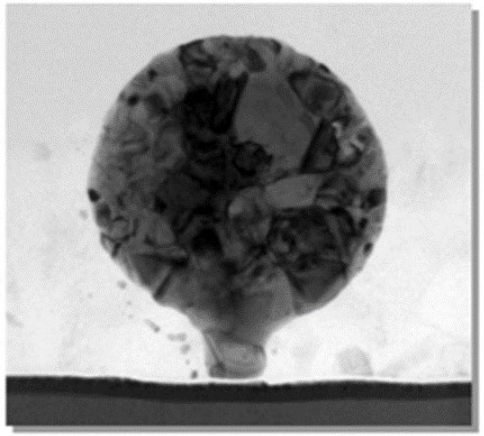
# HAMR Media and Head Structures

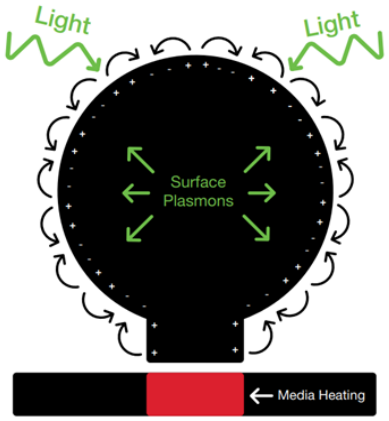



Media consists of high anisotropy FePt with optimized thermal/optical properties

Plasmonic NFT device delivers EM energy in sub-diffraction spot for high density recording

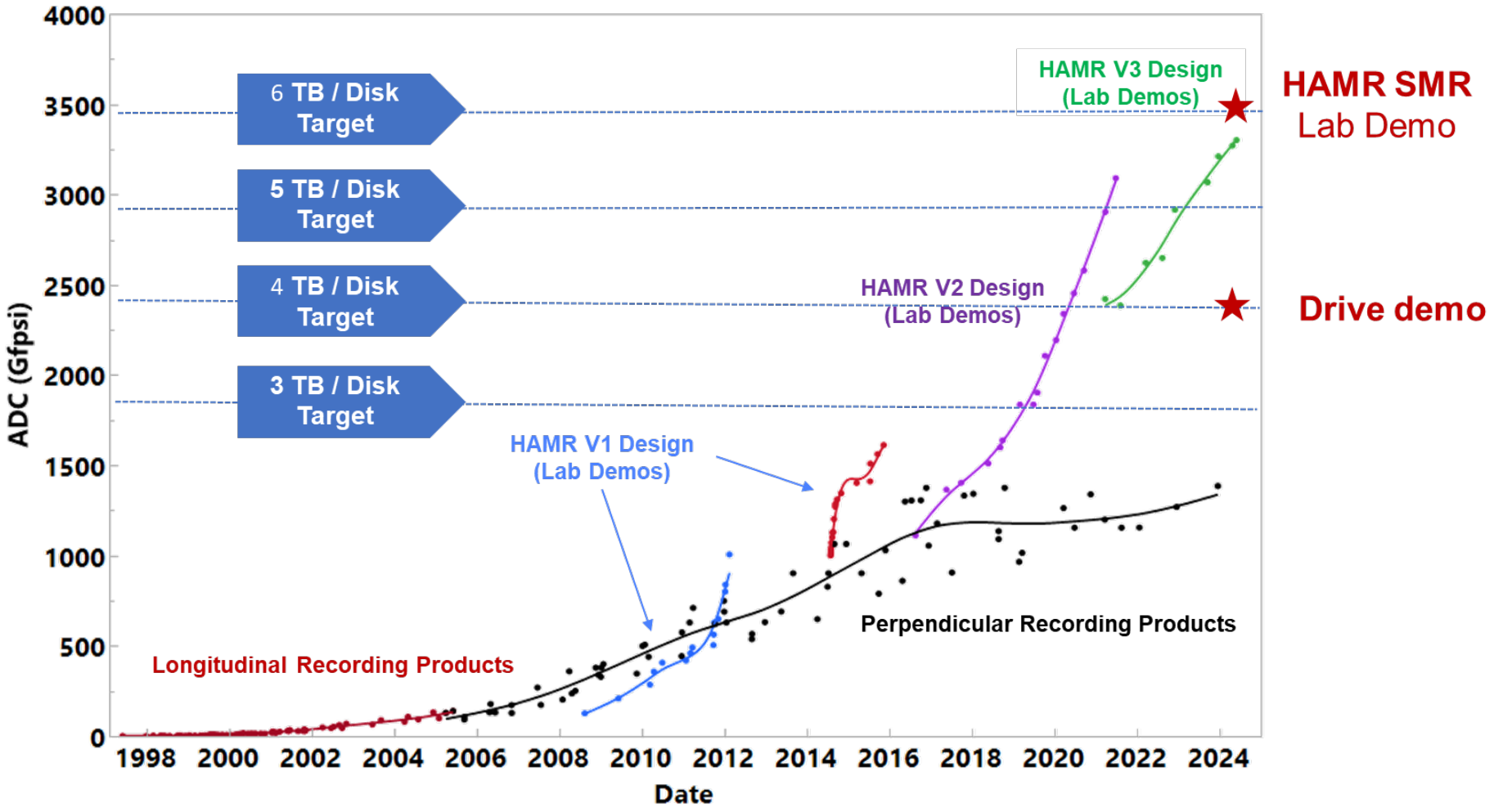





Near-Field Transducer (NFT)

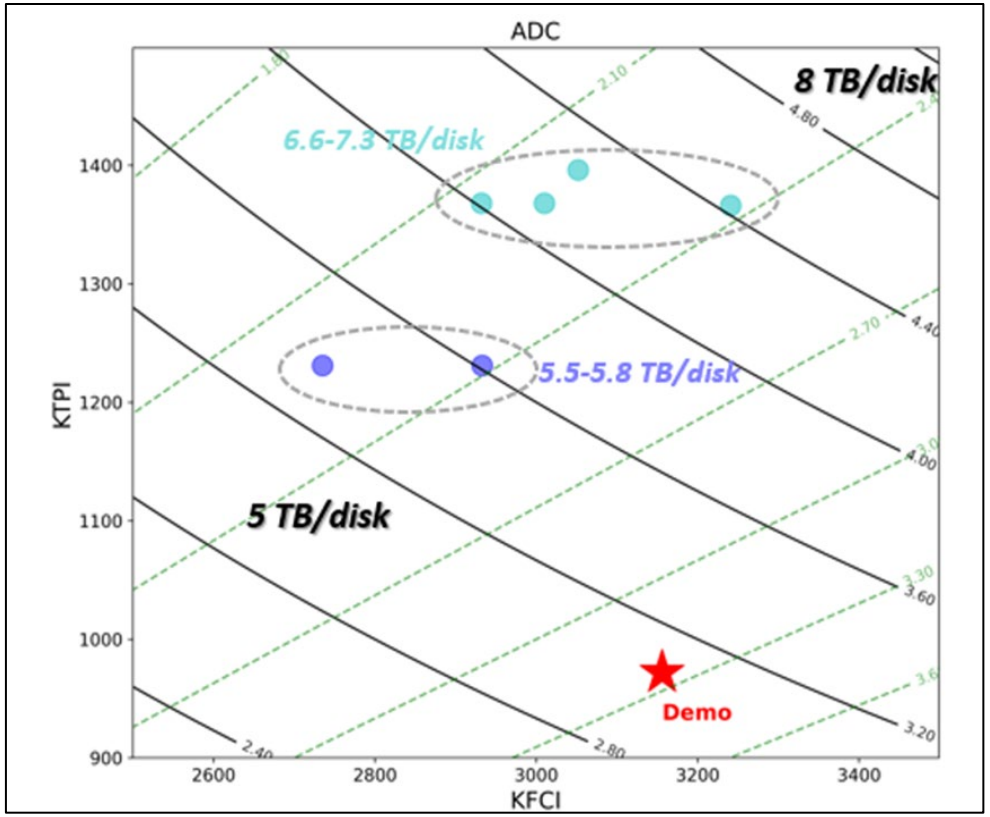
 

# HAMR Areal Density Progression



# Underpinning The Future HDD Roadmap

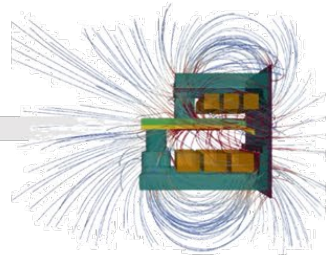



- RSS modeling is used to evaluate advanced writer, reader, and media designs.
- Designs for > 5 TB/disk have been identified and are being matured internally and through external research.

# Future Outlook Summary

10 TB/disk



Modeling suggests 10TB/disk is achievable with HAMR and component innovations

5 TB/disk

Achieved well over 5 TB/disk with HAMR at spin-stand with aggressively scaled dimensions and clearances

4 TB/disk

Achieved 4 TB/disk in fully formatted factory processed HAMR drives

# Storage Reliability & Sustainability Innovation

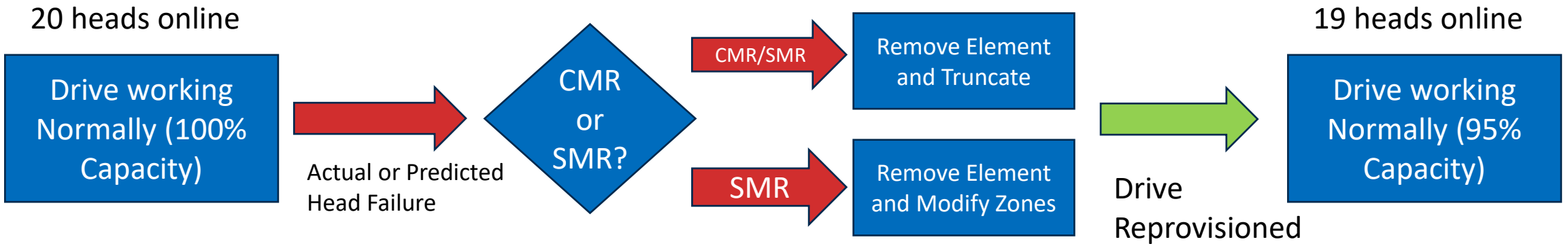
# Enterprise HDD Reliability & Sustainability

- Standard Enterprise HDD reliability:
  - 2.5M MTBF = 0.35% AFR = 1.75% CFR
  - 8760 hours/year and 5-year warranty
- Estimated 40%+ of failures are "single-head" failures eligible for regeneration
- With 20 heads per drive, a single-head failure means 95%+ of the capacity still good!
- Replacing a drive can cost thousands of $'s – technicians, shipping, processing, availability downtime etc.
- Millions of drives are shredded each year, keeping drives in use is significantly more sustainable than shredding or recycling
- There's a better way: **Drive Regeneration (Storage Element Depopulation)**

# Drive Regeneration Process

- Drive Regeneration known in standards as **Storage Element Depopulation** allows removal of a single head & surface from usable LBA space.
- The drive is re-provisioned at ~95% of original capacity (20 head drive) and may continue normal operation without leaving the system.



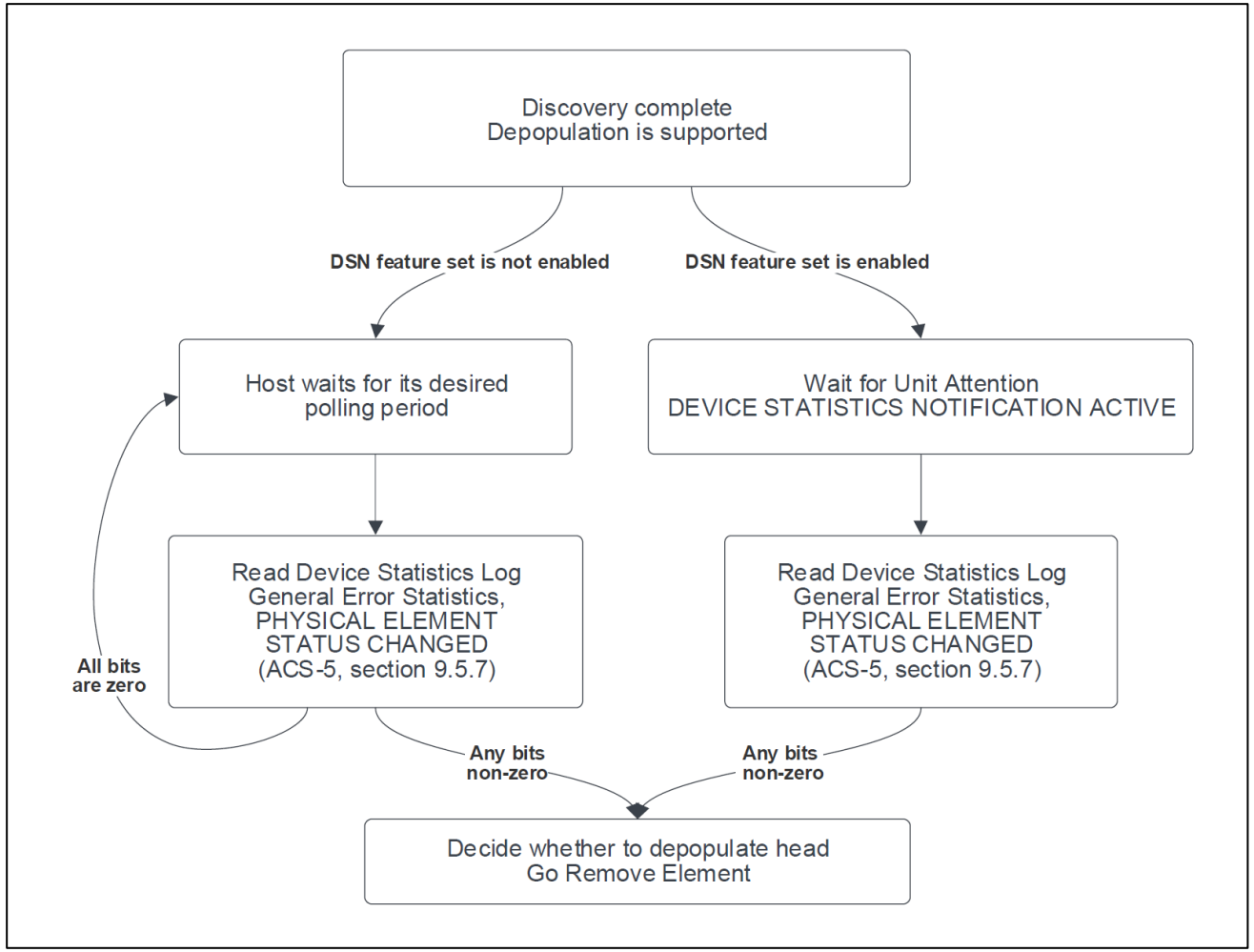
# Drive Regen Features and Support

| 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 |
| Online 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 |

# Implementing Drive Regen – Head Discovery

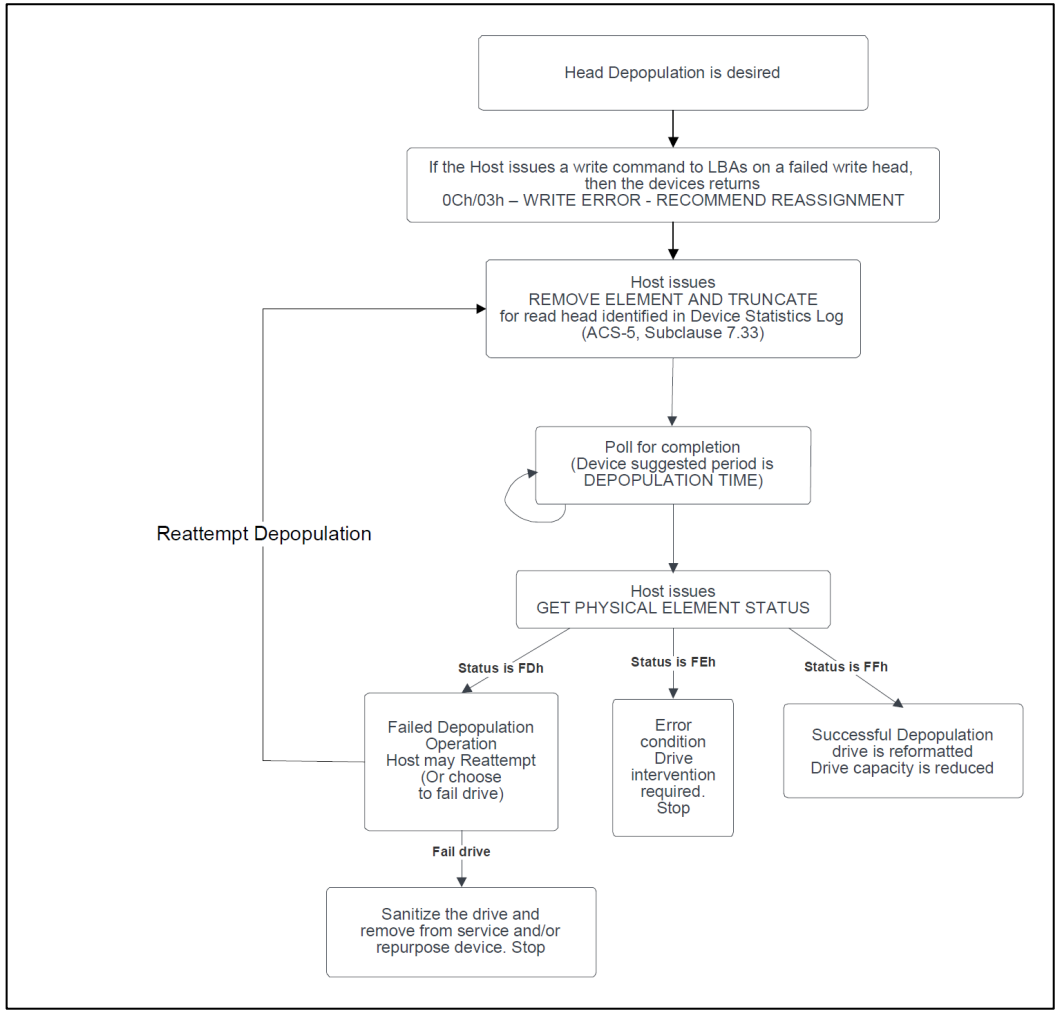
- Host confirms depopulation is supported
- Enterprise HDD internal algorithms monitor head health metrics and populate "**Get Physical Element Status**" field to notify host when a head is eligible for regen
- Host decides whether to regularly poll, or use Device Statistics Notification asynchronous alert via "**Physical Element Status Changed**"



# Implementing Drive Regen – Remove Element

- After identifying head for depopulation, host issues "**Remove Element and Truncate**" for offline regen
- Host may poll the device for estimated time to complete the operation (format and re-provisioning of LBAs)
- After complete, host issues another "**Get Physical Element Status**", receiving status back from the drive. Status FFh indicates successful depopulation and drive capacity is reduced.
- Drive re-provisioning changes capacity, posing challenges for Hardware RAID. Systems utilizing Erasure Coding already using Drive Regen today!



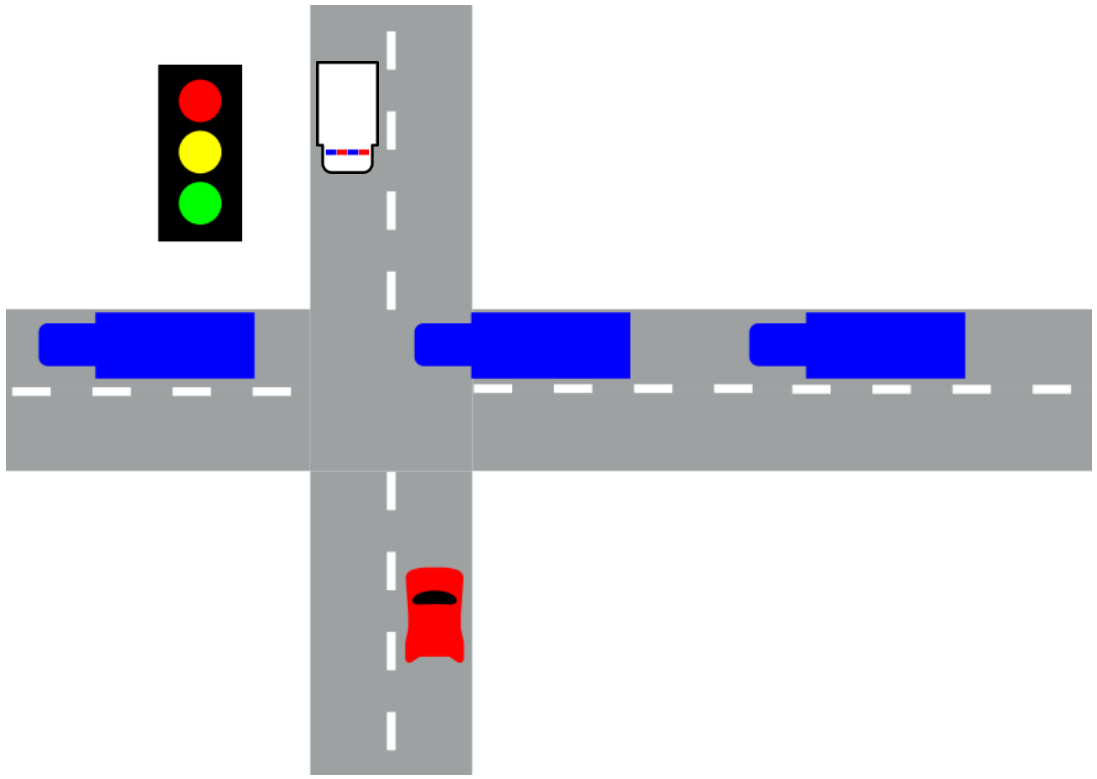

 

# Optimizing Performance

# Throughput vs Latency

- High bandwidth Write street (Blue)
- Latency sensitive Read street (Red)
- Priority Read (Ambulance)



# Command Duration Limits (CDL)

- As capacities increase, customers see more users per spindle, requiring performance to scale with capacity to satisfy service contracts
- CDL allows the host to supply time limits for read and write commands and an associated action to take if the limit can’t be met.
- The ability to provide command specific latency expectations, enabling opportunity to increase queue depth and gain additional throughput.

# How has this been addressed in the past?

- Limiting Queue Depth on the host.
  - Large opportunity cost
  - Commands age on host rather than drive
  - Gives up throughput even when latency would have been naturally constrained
  - Periodic queue drains to constrain worst case latencies further impact performance
- NCQ Prio
  - Specifies relative priority between commands, but interpretation of that priority is device dependent
    - Questions around how much a command can be pushed out are not addressed
    - Has no effect for commands at the same priority level
    - Fails to capture the weight of the relative priority (Excludes lower priority, heavily weighted to high priority, mixed fairness, etc.)

# Putting a reasonable bounds on time outstanding

- CDL provides seven separate Duration Limit Descriptors (DLDs) for reads and seven for writes.
- Each DLD can specify a different time threshold for Inactive, Active, and Total Time periods:

| Time Period | Description |
|---|---|
| Inactive | The  time a command waits in the drive's internal queue |
| Active | The time that a drive can spend operating on a command |
| Total | The time from command acceptance until status is returned by the drive |

- Read and write commands can be assigned to a DLD, providing a hint to the drive as to how sensitive the command is to latency
- Allows for different classes of latency needs to be expressed on a single host
- Low latency commands can be mixed with high latency commands

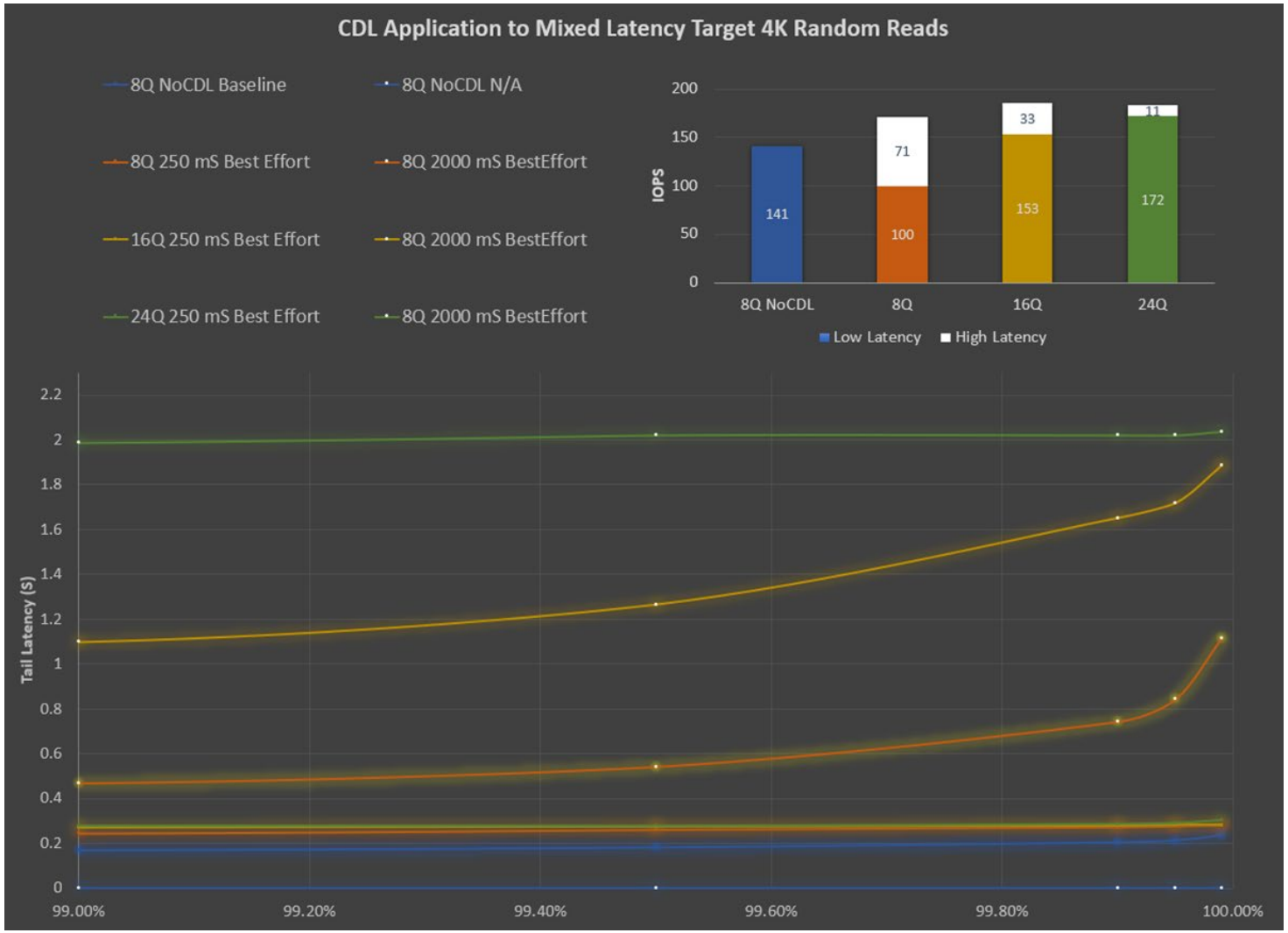
# What happens when the drive can’t meet the latency ask?

- When the drive fails to meet the host latency request, there are a limited number of actions that it can take. These actions are mapped to CDL Policy values, allowing the host to specify which action to choose for the given DLD by timer.

| Policy Value | Name | Description |
|---|---|---|
| 3h | Continue Limited | If unable to meet the timing requirement, push this command out to the values specified in the next DLD |
| 4h | Best Effort | If unable to meet the timing requirement, prioritize this command to the detriment of IOPs and other commands’ latencies |
| 5h | Continue Unlimited | If unable to meet the timing requirement, push this command out indefinitely to prioritize IOPs and other commands’ latencies |
| Dh | Early Return without Error | Complete the command without performing the data transfer but send good status. This will preserve the rest of the queue. |
| Fh | Early Return with Error | Complete the command without performing the data transfer, sending error status. The rest of the queue will not be preserved. |

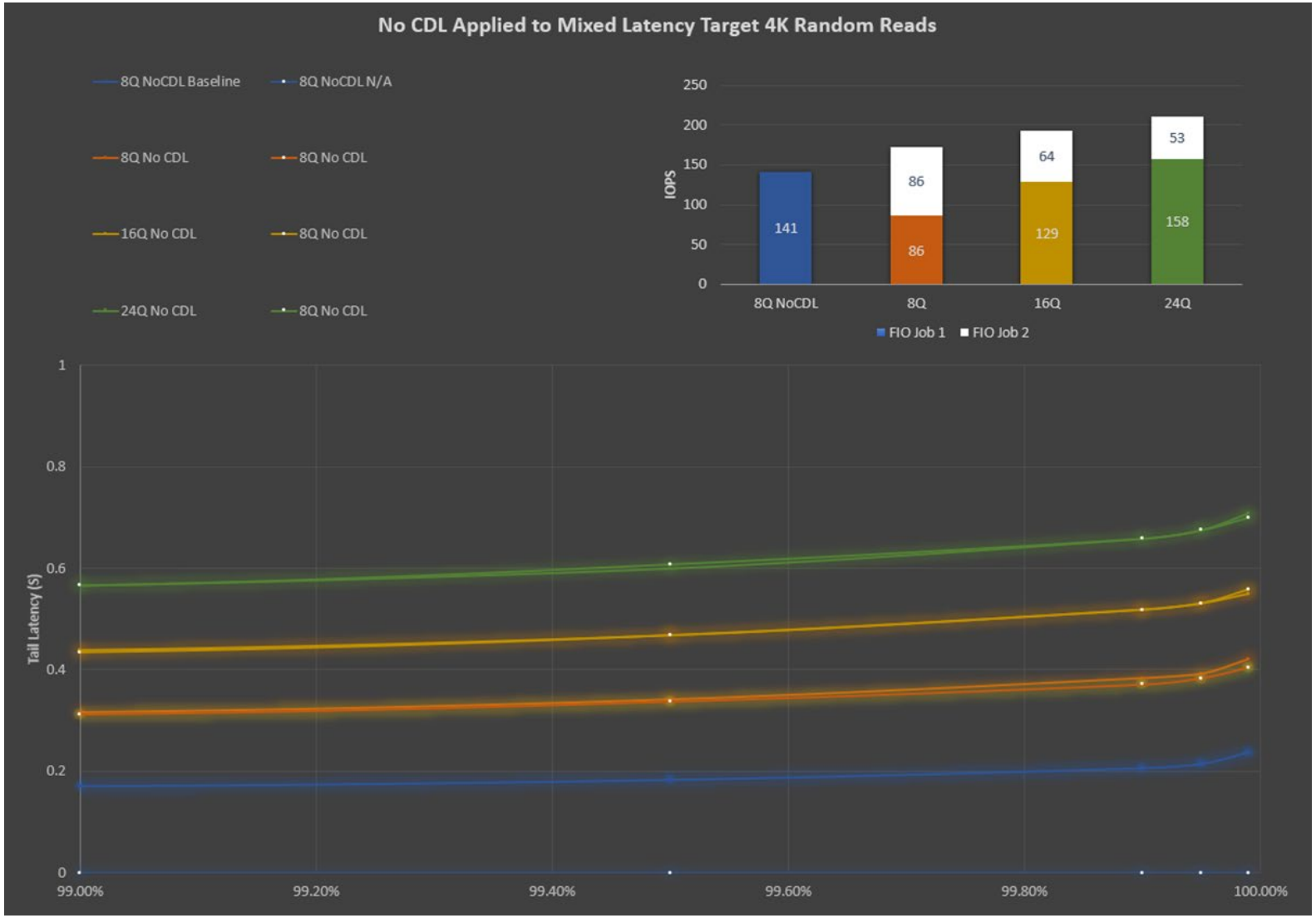
# CDL Throughput and Tail Latency Control




- CDL allows us to increase throughput and manage workload latency agreements for multiple IO classes.
- Utilizing two duration limit descriptors to shape latencies and increase throughput by altering the qdepth ratio between two simultaneously executed 4K random read workloads.

# Without CDL Throughput and Tail Latency




- Without applying CDL an overall increase in tail latency is spread across the two simultaneous workloads.
- IO/s increases by 5 to 15% but this comes at the expense of giving up latency control.

# Key Takeaways

- Massive datasphere exabyte growth continues, with HDDs storing 80-90% of bytes in datacenters now and in the future
- HAMR technology breakthroughs significantly accelerate and extend the areal density growth capabilities of HDD technology
- Regen/Depop features significantly improve storage availability and sustainability of HDDs in the datacenter
- Command Duration Limits enables higher queue depths with more flexible latency management for multiple classes of IO

# Please take a moment to rate this session.

Your feedback is important to us.