SNIA DEVELOPER CONFERENCE

SDC 25

By Developers FOR Developers

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

HDD Innovations for Hyperscale

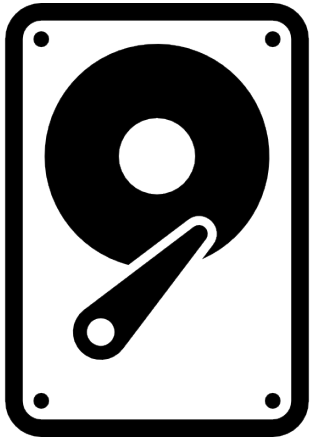
Rick Kutcipal, Product Planner, Broadcom
Damien Le Moal, Distinguished Engineer, Western Digital

STA™

SCSI Trade Association
A SNIA Community

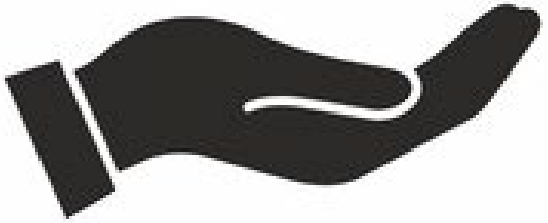
www.sniadeveloper.org

Agenda



The need for HDDs in modern hyperscale architectures

Innovations enabling HDDs in modern hyperscale architectures



Open-source contributions supporting these Innovations

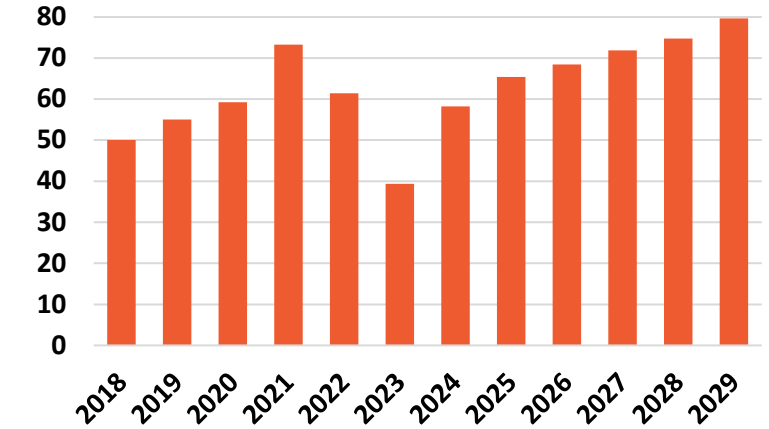
The need for HDDs in modern hyperscale architectures

The Correction

- ’24 – ’29 Exabyte CAGR 24%
- ’22 / ‘23 Post-Covid inventory correction

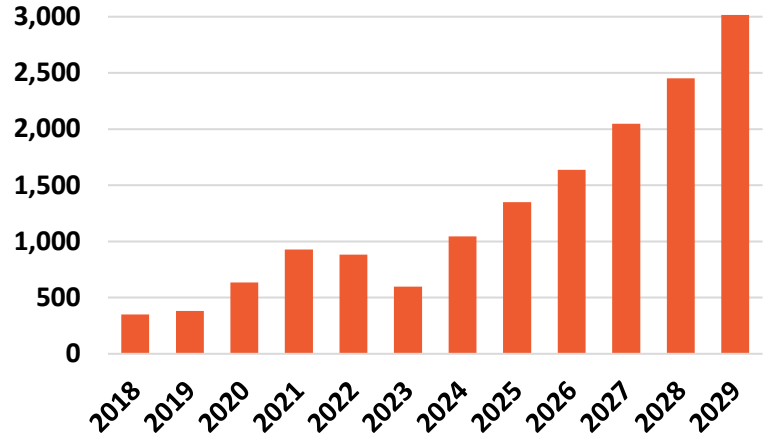
– TRENDFOCUS

NL HDD Units (M)



2024-2029 Unit | CAGR **6%**

NL HDD Capacity (Exabytes)

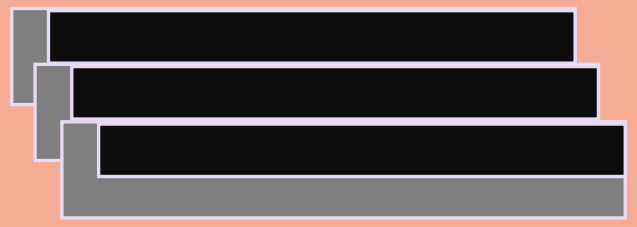


2024-2029 Capacity | CAGR **24%**

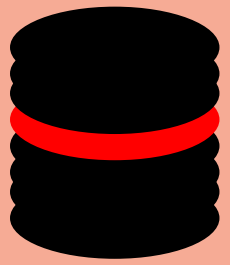
Source: TRENDFOCUS CQ2 ’25 Revised Long-Term Forecast through 2029

Innovations enabling HDDs in modern hyperscale architectures

Hyperscale HDD Innovations



Shingled Magnetic Recording (SMR)

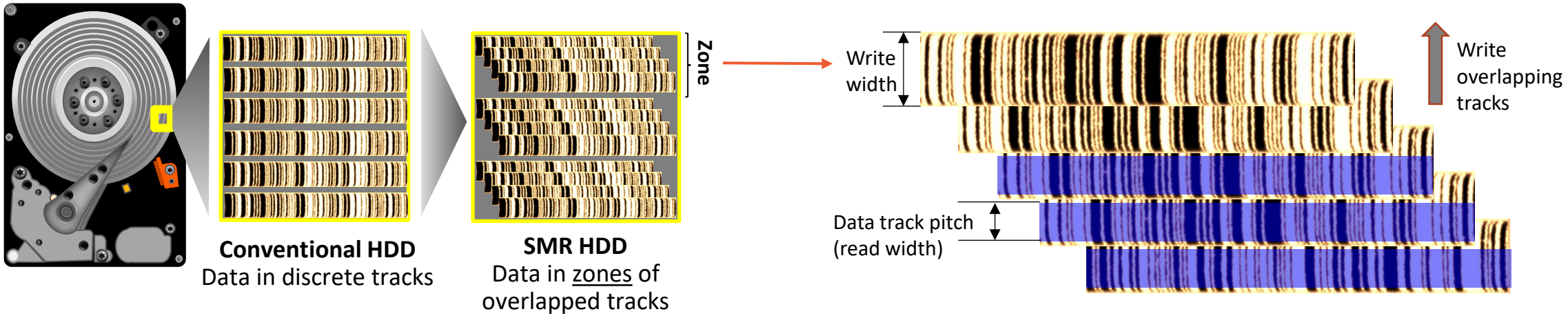


Repurposing Depopulation (Depop)

Command Duration Limits (CDL)

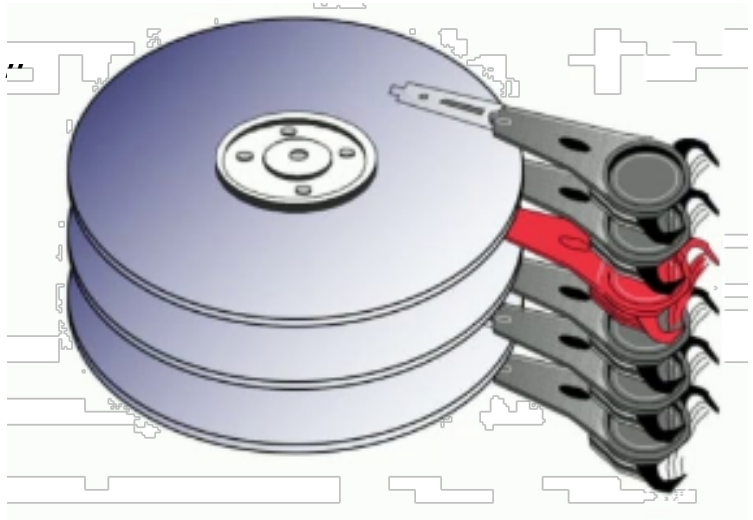
Shingled Magnetic Recording

- SMR enables higher drive capacity by overlapping written tracks
  - Tracks are organized into zones
  - Requires sequential writes in each zone, no read performance impact
- Standards
  - ZBC – SCSI Zone Block Commands
  - ZAC – Zone ATA Commands
  - SAT-4 – Defines SCSI to ATA translation between ZBC and ZAC
- Host managed SMR started shipping around 2014
  - Broad ecosystem support in 2025



Repurposing Depopulation

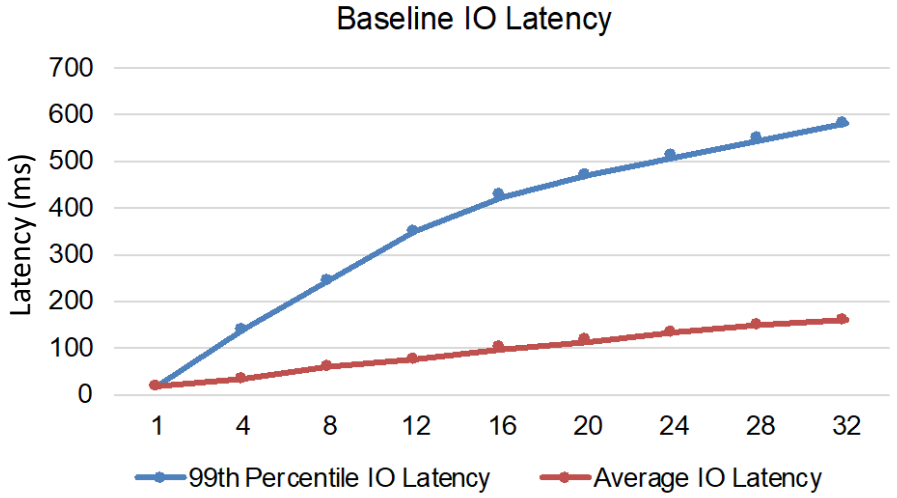
- Growing capacity of HDD presents challenges for large-scale Data Center deployments
  - Increased frequency of correctable errors increases tail latencies and decreases performance
  - Failing a drive results in vast amounts of capacity remaining offline until a failed unit is replaced
  - Significant number of returned drives found fault is with a single failed head
- T10 and T13 standards bodies defined "Offline Logical Depop"
  - Capacity backed by the failed head is removed from namespace
  - Drive is reformatted at the lower capacity
  - Drive brought back online with the lower capacity
  - T10: SBC-4/5, ZBC-2 and SAT-5
- The Offline Logical Depop ecosystem is maturing, Hyper-Scale deployments are immanent



Command Duration Limits

- HDD tail latency is important in large data centers
  - For implementing different service level agreements
  - For overall system performance – the aggregate system performance is throttled by the drive with the longest access time
- OCP published "Cloud HDD – Fast Fail Read" in 2018
- In 2019, T10 introduced Command Duration Limits and proposed it to SPC-6

One Read to One Location



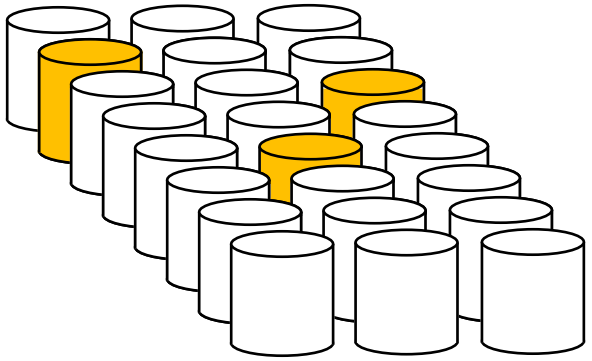
Latency data source:

Damien Le Moal, SDC21 - *Be On Time: Command Duration Limits Feature Support in Linux*[tm]

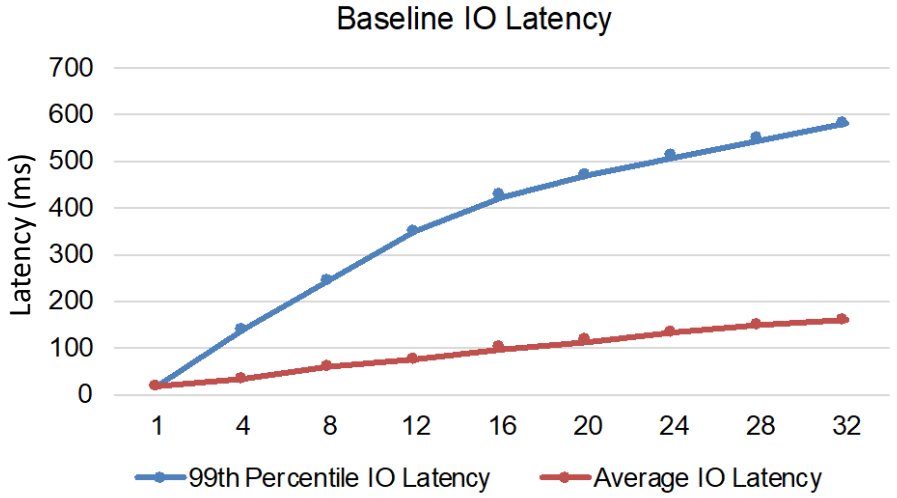
Tail Latency Grows w/ Queue Depth

Command Duration Limits

- HDD tail latency is important in large data centers
  - For implementing different service level agreements
  - For overall system performance – the aggregate system performance is throttled by the drive with the longest access time
- OCP published "Cloud HDD – Fast Fail Read" in 2018
- In 2019, T10 introduced Command Duration Limits and proposed it to SPC-6



One Read to Three Locations



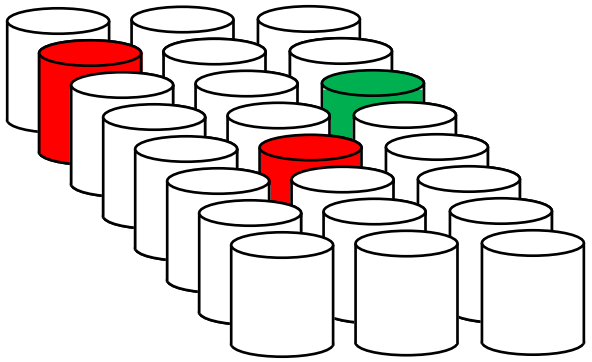
Latency data source:

Damien Le Moal, SDC21 - *Be On Time: Command Duration Limits Feature Support in Linux*[tm]

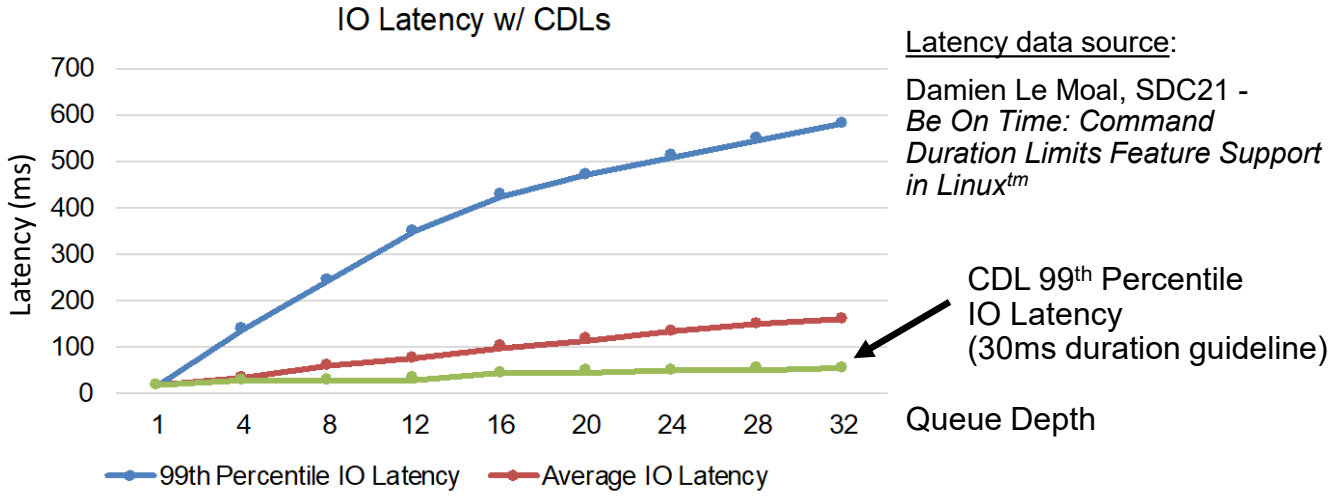
Tail Latency Grows w/ Queue Depth

Command Duration Limits

- HDD tail latency is important in large data centers for 2 reasons
  - For implementing different service level agreements
  - For overall system performance – the aggregate system performance is throttled by the drive with the longest access time
- OCP published "Cloud HDD – Fast Fail Read" in 2018
- In 2019, T10 introduced Command Duration Limits and proposed it to SPC-6



One Read to Three Locations



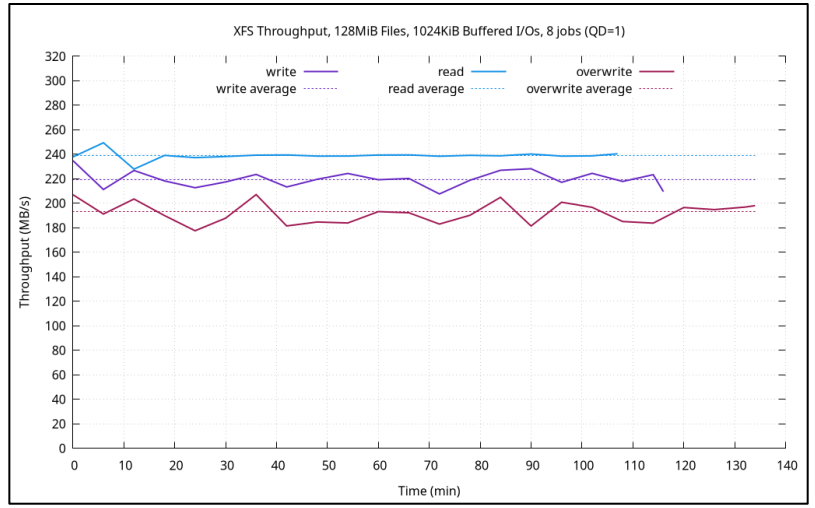
CDLs Greatly Improve IO Tail Latency

Open-source contributions supporting these Innovations

Linux Kernel and user tools

Linux Kernel SMR Support

- Support introduced with kernel v4.10 in 2017
  - (Zoned) block device file interface only
- Support has grown and improved significantly
  - Better (and complete) zone management user API
  - Better handling of sequential write command ordering per zone
    - Re-implemented in kernel 6.10 with zone write plugging
  - Support in device mapper and file systems
    - dm-crypt target (among other), BTRFS and XFS
      - F2FS support is still valid but unusable with SMR disk capacities now exceeding 16TiB
- Support for SMR is now enabled by default in most Linux distributions



Zoned XFS: Transparent and Efficient Zoned Storage Support For Scalable Storage Systems
File Systems & Protocol session this afternoon

SMR Tools and Libraries

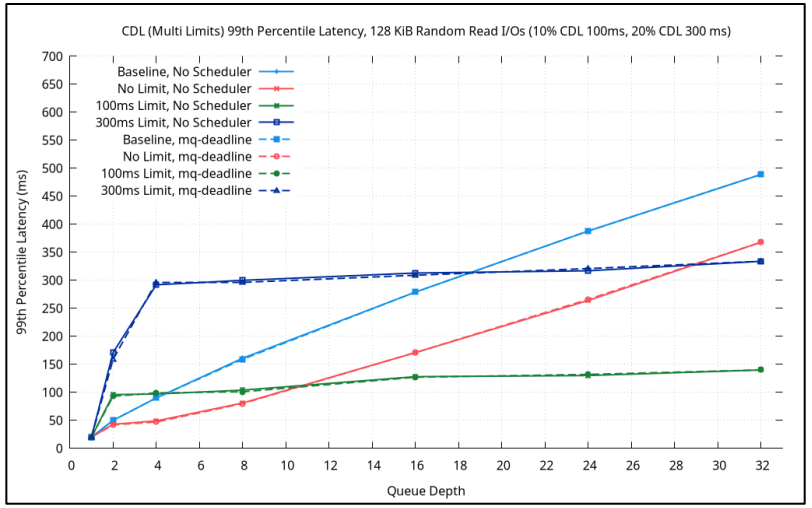
- Libraries
  - **libzbc** (https://github.com/westerndigitalcorporation/libzbc) and **sg3utils** (http://sg.danny.cz/sg/sg3_utils.html)
    - SCSI/ATA passthrough library and utilities for zone management commands
    - Beware !
  - **libzbd** (https://github.com/westerndigitalcorporation/libzbd)
    - Zone management API using the kernel *ioctl* interface
- Utilities
  - **util-linux** (https://github.com/karelzak/util-linux)
    - *blkzone* command line utility
  - **fio** support introduced with version 3.9 in 2018
    - *zonemode=zbd* option

Depop Support

- No kernel support
  - Head depop is not initiated by any in-kernel component, this is a user/sysadmin decision
  - Offline depop reformats the disk
    - Regular disk revalidation captures the disk capacity change
  - Data preserving depop (SMR) switch zones under the depop head to the “offline” state
    - File systems will not like this (corruptions): unmount first !
    - All SMR-aware software should already be coded to handle (and ignore) offline zones
- User tools
  - **sg3utils** (http://sg.danny.cz/sg/sg3_utils.html) can be used to issue remove/restore element command
    - *sg_rem_rest_elem* utility (send SCSI remove or restore element command)

Linux CDL Support

- Support introduced with kernel v6.5 in 2023
  - Defines an interface to pass CDL limits per IO from user process down to the device
- User tooling released at the same time
  - **cdl-tools**
    - https://github.com/westerndigitalcorporation/cdl-tools
    - Utility to consult and update a disk duration limits descriptors
  - **fio** also supports CDL
    - *ioprio_hint* and *cmdprio_hint* options allow specifying a duration limit for I/Os, per job and per I/O



Command Duration Limits - Improving IOPS per TB in HDDs
Data Architecture / Storage Architecture session later this morning

Follow STA

SCSI Trade Association
A SNIA Community

https://www.snia.org/sta-forum

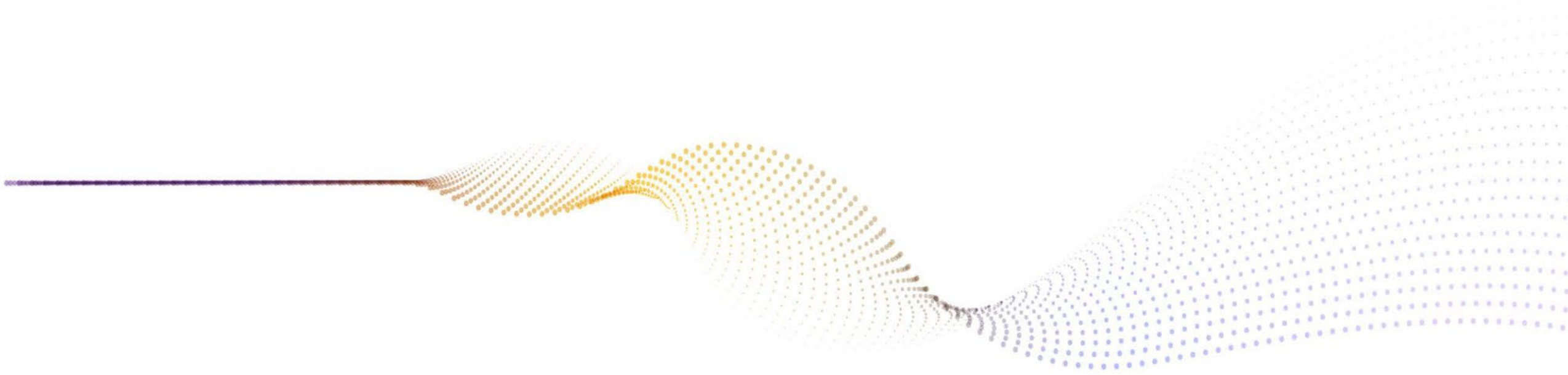
https://x.com/SNIA

YouTube

Serial Attached SCSI Playlist on SNIAVideo
https://www.youtube.com/@SNIAVideo

LinkedIn

https://www.linkedin.com/company/snia/



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

