HDD FEATURES FOR THE FUTURE

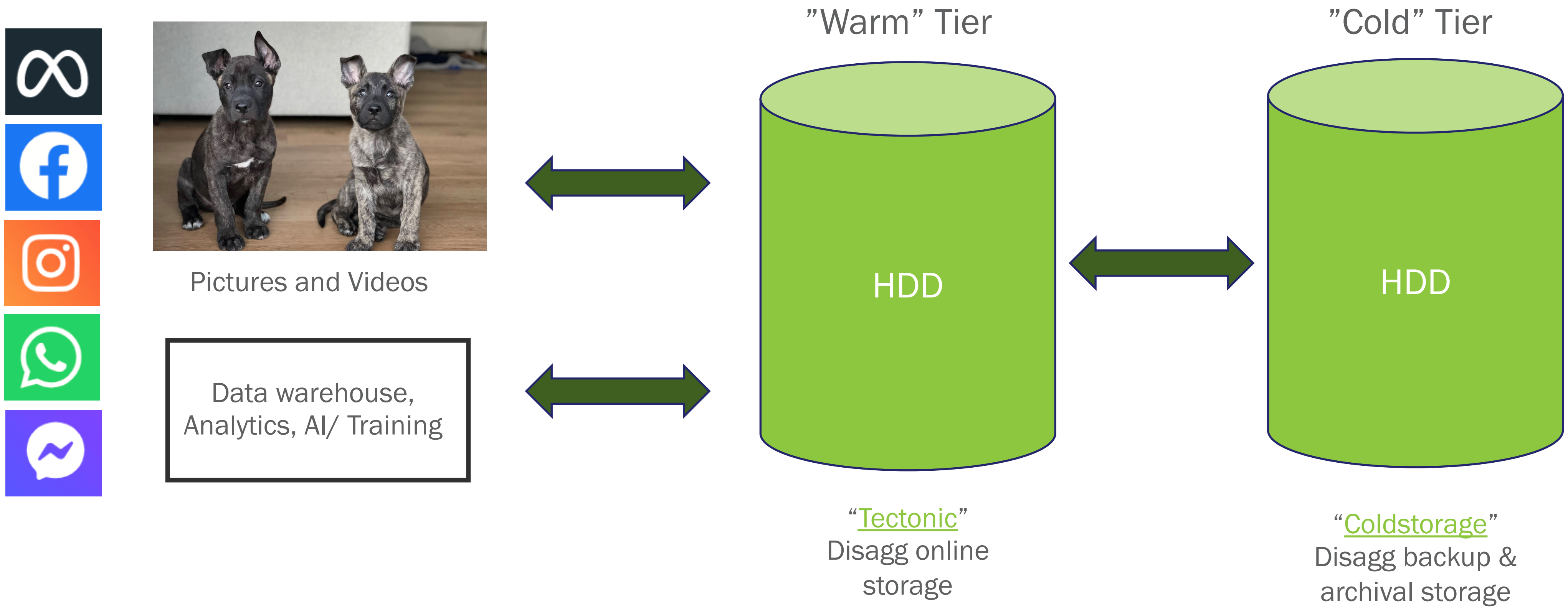
Meta’s perspective

Madhavan Ravi, Meta

Connect. Collaborate. Accelerate.

OPEN Compute Project®

HDD Usecases – 30,000ft view



Pictures and Videos

Data warehouse, Analytics, AI/ Training

”Warm” Tier

HDD

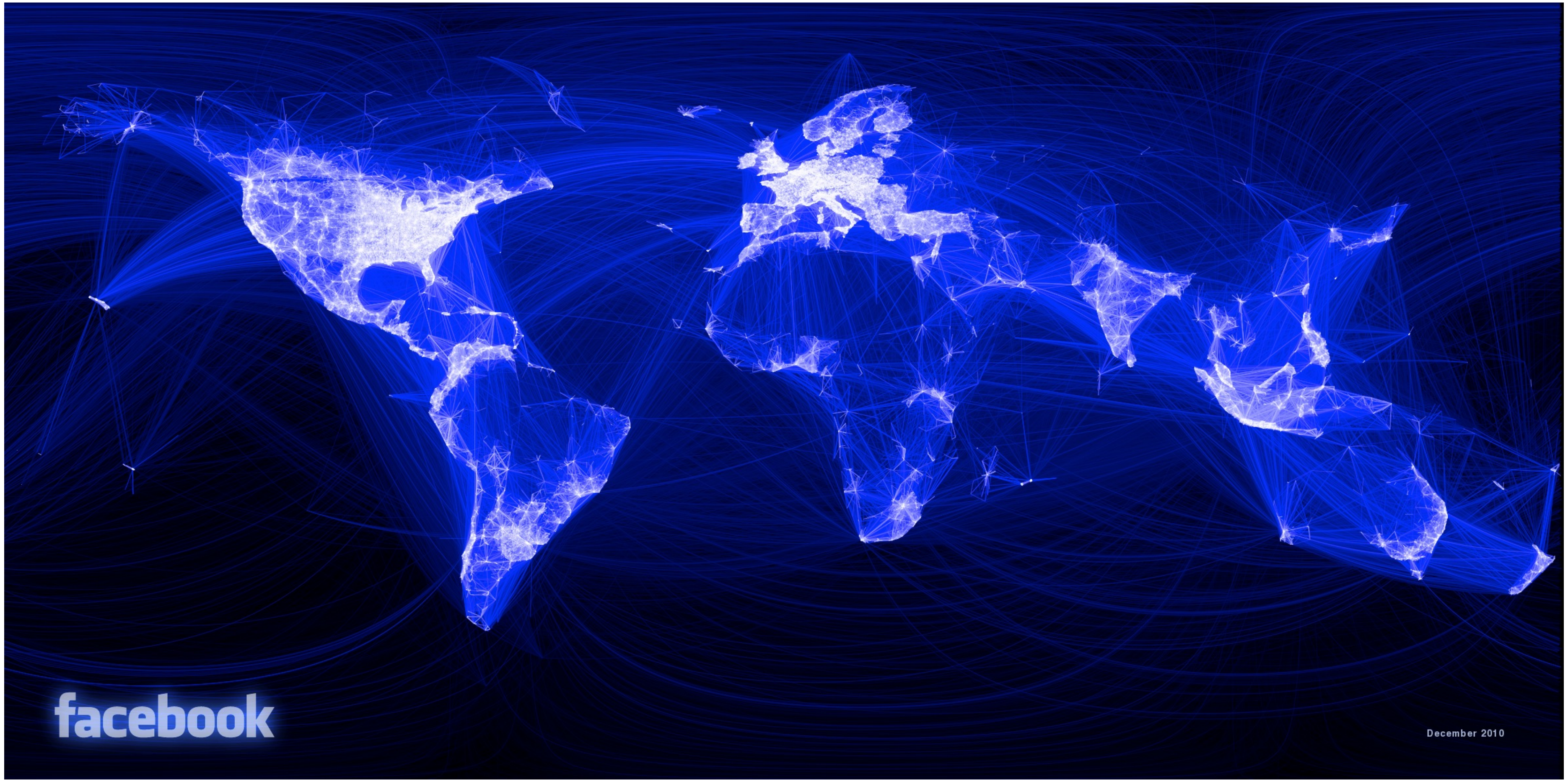
“Tectonic”
Disagg online storage

”Cold” Tier

HDD

“Coldstorage”
Disagg backup & archival storage

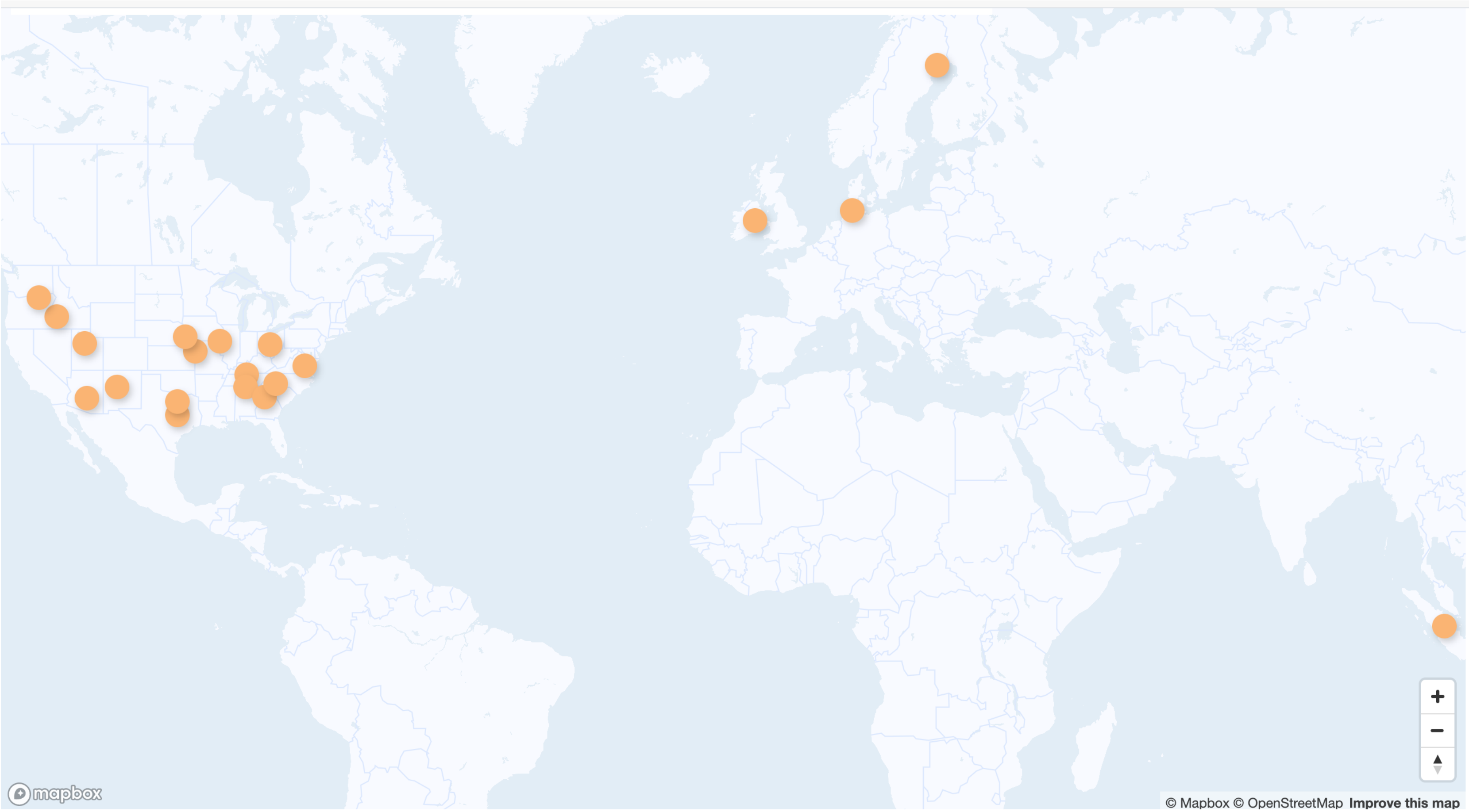
Global Connectivity snapshot



- 2010 snapshot above
- We've grown ~10x in our DAU metric from 2010 to now just on the Facebook product. Imagine how this map morphs. Now add Instagram, Whatsapp, Messenger, ...

Connect. Collaborate. Accelerate.

Meta’s Datacenters



Source: https://datacenters.fb.com/

Connect. Collaborate. Accelerate.

Warm HDD Tier Storage Evolution

Past

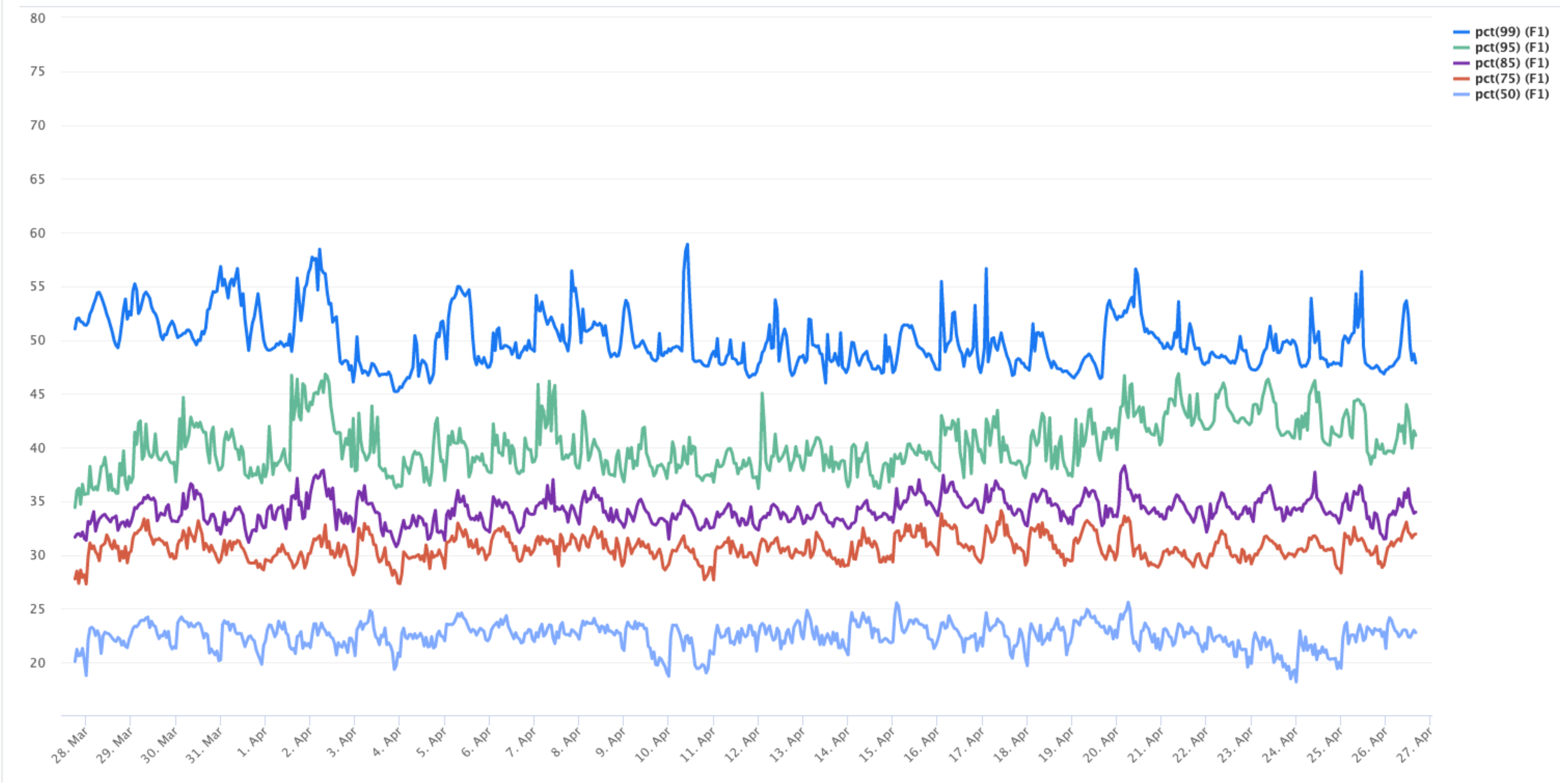
- 1 Customer per cluster
- Homogeneous workload per cluster
- Localized workload extremes across clusters
- 100s of clusters
- No rebalancing to optimize for IO demand
- No caching layer in Datacenter
- Trends
  - BLOBstore → Storage bound on HDD
  - Data Warehouse → IO bound on HDD

Present

- Datacenter building
  - Storage fabric hosting many tenants
- Operate 10s of clusters
  - Each cluster is up to 10x larger than the past
    - More HDDs to share the IO load
- Rebalance data to move more cold data to bigger HDDs
  - All spindles share the IO load equally
- Read caching on SSD in both
  - HDD Storage Node
  - Network-connected SSD Storage Nodes

Goal: Remain Storage bound on HDDs

Warm Tier : HDD IO utilization



Higher IO utilization on HDD –> longer tail latencies
Prevents HDDs from being used to their full IO capacity

Connect. Collaborate. Accelerate.

Cold HDD Tier Evolution

- HDD cold tier has not evolved over time
  - Highly power- and cost-optimized solution
  - Only power ON 7-8% of HDDs per Server at a time
  - Workloads are random, large blocksize requests (MBs to GB)

Trend: Remain Storage Bound

What Meta cares about in HDDs

- Keep up the HDD Capacity CAGR
  - Translates to TCO savings and W/TB (power footprint) savings
    - Is 10 platters per HDD the limit?
    - Accelerate HAMR/MAMR product delivery

- Power optimizations/opportunistic power savings
  - W/TB savings while meeting latency requirements

- IO priority mechanisms
  - All IO requests are not equal
  - Ability to use 100% of available IO capacity per HDD, without tail latency impact

- Bring modest throughput increments per HDD
  - W/TB is the metric for improvement
  - Excited about bit per inch (BPI) improvements in HAMR and MAMR to help with this

- Data and metadata persistence mechanism improvements on power loss
  - Write cache data safety (<10 seconds time-to-persist requirement)
  - Write pointer hardening on open SMR zones

HDDs features to use and/or explore

<table>
<tr><th>Feature</th><th>“Warm” Tier</th><th>“Cold” Tier</th></tr>
<tr><td>HDD Write Cache</td><td>Required</td><td>Disabled</td></tr>
<tr><td>Secure Boot, Signed FW update, chain of trust, secured/disabled debug ports</td><td colspan="2">Required</td></tr>
<tr><td>PRIO (high and low IO priority)</td><td>Evaluating</td><td>N/A</td></tr>
<tr><td>Command Duration Limits (CDL)</td><td>High Interest</td><td>N/A</td></tr>
<tr><td>Power Balance/Adv. Power Mgmt</td><td colspan="2">Evaluating</td></tr>
<tr><td>Attestation</td><td colspan="2">Highly Interest</td></tr>
<tr><td>Encryption at Rest/Full disk Encryption</td><td colspan="2">Interest</td></tr>
<tr><td>SMR</td><td colspan="2">High Interest</td></tr>
<tr><td>Reman/Head depop</td><td colspan="2">Interest</td></tr>
</table>

NVMe HDD Thoughts

- Benefits
  - User software stack unification for SSD and HDD storage nodes
  - No proprietary drivers needed
  - Able to leverage existing SSD tools

- Critical features for consideration
  - Command Duration Limits
  - SMBus interface for out-of-band temperature polling & management
  - Support for Attestation (SPDM over MCTP) via SMBus and in-band via VDM
  - T10 DIF support

- Feature Concerns
  - Resource benefits from IOC+Expander → PCIe Switch
    - Benefits Very small – Not an NVMe HDD driver
  - HDD connector to optimize out SMBus
    - Very small resource saving
    - Eliminates one of the main value drivers for going to NVMe HDD

Dual Actuator HDD Thoughts

- High power footprint
- Lagging capacity vs single actuator HDDs makes it unattractive today
- Past: Was evaluated due to small, localized cluster challenges
  - Issue is resolved via shared clusters and SSD based caching
- For most high IO load use cases, read caching on SSD will be the power optimized answer

Conclusion : Could be of interest in the future but no application need today

Thank You

Any Questions?

Connect. Collaborate. Accelerate.

OPEN Compute Project®