Multi-Actuator: Performance & Data Protection Best Practices

James Borden & Tim Walker
Seagate Technology
Agenda

- Stranded Capacity and HDD Parallelism
- Data Durability and Performance in a Multi-LUN Environment
- Best Practices
- Summary
Multi Actuator and Stranded Capacity

New recording technology is driving HDD capacity to 60TB+ per spindle

Servo-mechanical capability has not scaled with areal density, so IOPS/TB are falling

Latency driven workloads cannot utilize the capacity gains as IOPs/TB drops below minimum workload QOS

To meet read latency QOS:
- short-stroke the HDD, leaving unused capacity stranded
- Deploy lower capacity drives
- Both increase overall storage TCO

IOPS/TB Outlook 2014-2021

HDD capacity increasing

Data management advances trending minimum IOPS/TB requirements down

HDD IOPS/TB Performance Gap
HDD Parallelism and Mach.2™

Multiple parallel data channels via multiple actuators:

Dual-actuator solution leverages highly-developed HDD hardware technology and established SAS transport and protocol.

Seagate Mach.2™ separates media into two groups of platters, each with its own independent actuator, read channel, and disk manager for ~2X throughput.

This design’s data path is essentially independent; management and reporting path is common.
Performance and Redundancy

1. Multi-actuator HDDs are effectively as reliable as a corresponding capacity single-actuator drive
   - Additional components are small contributors to overall AFR (annual failure rate)
   - LUNs are considered to be in the same failure domain
     - Single head failures are the primary single-LUN failure mode

2. Dual-actuator HDDs data-redundancy implementations
   - Provide 1.75-2X reduction in array rebuild time for drive failure
   - Increase MTTDL (mean time to data loss) by 2X for RAID5 and 4X for RAID6
   - Provide better (1.75-2x) performance vs. single-actuator drives, improving TCO
   - Enable more granular array data placement to reduce the noisy-neighbor effect and reduce the impact of data hot spots

3. Current array technology can support multi-actuator devices, but a reliable configuration requires:
   1. Clear understanding of the underlying failure domains
   2. Solid familiarity with updated administration tools
Multi-Actuator Failure Domains in Arrays

Spoiler: Do not put strips/shards from the same stripe onto the LUNs of the same device.

Multi-actuator drives change the commonly accepted definition of failure domains of a disk array where single drive failure can affect multiple LUNs.

Data redundancy schemes must address common failure modes between LUNs that are not present in single LUN devices.
Data Availability

Array data availability is driven by drive AFR and rebuild time.

Reducing AFR
- Multi-actuator drives have about the same AFR as single actuator nearline.
- Multi-actuator AFR is better than two separate drives (\(\Delta \text{perf} > \Delta \text{reliability}\))

Reducing rebuild time
- Each actuator-media set deployed independently as smaller, full-throughput LUNs.
- Higher throughput and IOPS/TiB reduce rebuild time, decrease time in critical mode, reducing probability of data loss.

\[
\text{Markov chain: Describes a sequence of possible events, such as drive failure and rebuilds, where the probability of each event is based on the previous state of the system, not the entire history.}
\]

\[
\text{n = array size, mttf = mean time to failure, mttr = mean time to rebuild}
\]
Array Availability & Rebuild Time

\( n \) = number of drives in array
\( \lambda \) = drive failure rate = AFR / 8766 (hrs/yr)
\( \mu \) = drive rebuild rate = rebuild transfer rate / capacity
MTTDL\(_x\) = mean time to data loss with redundancy of \( x \) drives

\[
\text{MTTDL}_0 = \frac{1}{n\lambda} \quad \text{no redundancy} \rightarrow \text{the basic failure rate equation}
\]

\[
\text{MTTDL}_1 = \frac{\mu}{n(n-1)\lambda^2} \quad \text{single drive redundancy} \rightarrow \text{MTTDL is proportional to rebuild rate.}
\]

\[
\text{MTTDL}_2 = \frac{\mu^2}{n(n-1)(n-2)\lambda^3} \quad \text{two drive redundancy} \rightarrow \text{MTTDL is proportional to rebuild rate squared. A 2x improvement in rebuild rate produces a 4x increase in MTTDL.}
\]

Analysis by Serkay Olmez (Seagate)
Best Practices

- Basic hardware RAID (MegaRAID, etc)
- Sophisticated hardware RAID (NetAPP, IBM)
- Software RAID and SDS
  (LVM, MDRaid, StorageSpaces, object store, Ceph)
- Multi-volume file systems (btrfs, ZFS)
Drive groups (sets) map stripes to a subset of LUNs. By placing LUN0s and LUN1s into separate drive groups, no slice will have multiple chunks on the same device. Aggregate performance is 1.75-2X a comparably-sized single actuator RAID. Downside: drive groups alone produce at least two volumes, one for each group.

Logically, multi-actuator HDD with media dedicated to each actuator, and a LUN assigned to each, are very similar to multiple, independent HDDs in common failure domains.

Device failure tolerance is the same as RAID5/6. A loss of an HDD degrades both redundancy-groups, but no data loss.

Two drive groups, each comprised of no more than one LUN per device.

Because of the dual-LUN parallelism, RAID rebuild time is ½ of an array with the same number of similar sized single actuator drives.
Scale-out RAID 50/60

- Combine multiple RAID 5/6 into a single RAID 50/60 for better scale out with fewer volumes.
- Larger arrays combined into fewer volumes

Logically, multi-actuator HDD with media dedicated to each actuator, and a LUN assigned to each, are very similar to multiple, independent HDDs in common failure domains.

We group one LUN from each HDD into a RAID5 or 6 redundancy-group.

Device failure tolerance is the same as RAID5/6. A loss of an HDD degrades both redundancy-groups, but no data loss.

The RAID5/6 redundancy-groups are striped together into a RAID50/60.

Because of the dual-LUN parallelism, RAID rebuild time is ½ of an array with the same number of similar sized single actuator drives. MTTDL is 4X (2X for RAID5).
Scale-out RAID 10

Stripe multiple RAID 1 Mirrors into a single RAID 10 for better scale-out with fewer volumes.

The RAID1 redundancy-groups are striped together into a RAID10.

Mirrored LUNs are not on the same device

Because of the dual-LUN parallelism, RAID rebuild time is ½ of an array with the same number of similar sized single actuator drives. MTTDL is 2X due to one drive redundancy.
Multi-LUN Distributed Array

Dynamic Disk Pools (DDP) & Software Defined Storage (SDS) Placement Policies

<table>
<thead>
<tr>
<th>Drive</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• SDS & DDP RAID supports multiple simultaneous stripe widths and placement policies. For durable multi-LUN data placement:
  • the stripe width must not be greater than the number of physical drives (pool width)
  • the slice-placing algorithm must know the failure domain boundaries

• Example shown:
  • Simple periodic 4+2 RAID 6 with 1/7 standby overprovisioning on an eight drive dual-LUN HDD pool
  • Since the stripe width is \( \leq \) the pool width we can place data chunks such that no drive has >1 chunk
  • The RAID6 data layout gives 2X faster rebuild, and MTTDL goes up by 4X.
  • The *inherent DDP rebuild advantage* from total LUNs > stripe-width is doubled with dual actuator due to 2X LUN targets to simultaneously receive rebuild traffic.

2019 Storage Developer Conference. © Seagate Technology. All Rights Reserved.
Summary: Performance and Redundancy

1. Multi-actuator HDDs are effectively as reliable as a corresponding capacity single-actuator drive
   - Additional components are small contributors to overall AFR
   - LUNs are considered to be in the same failure domain
     - Single head failures are the primary single LUN failure mode

2. Dual-actuator HDDs data redundancy implementations
   - Provide 2X reduction in array rebuild time for drive failure
   - Mean time to data loss (MTTDL) increases by 2X for RAID5 and 4X for RAID6
   - Provide better (1.75-2x) performance vs. single-actuator drives, improving TCO
   - More LUNs enable more granular array data placement to reduce the noisy-neighbor effects and reduce the impact of data hot spots

3. Current array technology can support multi-actuator devices, but a reliable configuration requires:
   1. Clear understanding of the underlying failure domains
   2. Solid familiarity with administration tools
Call to Action

- Improve Multi-LUN Storage Device Support in RAID Add-in Card Silicon
- Update Administration Tools to Incorporate Different Failure-domain Models
- Improve and Standardize Multi-LUN Storage Device Topology Discovery and Reporting
- Update SDS and Distributed RAID Placement Policies to take Advantage of Multi-LUN Storage and New Failure Domains
RAID 06

- Multi-level RAID implementation option.
- Implementation alternative to RAID60
- Viable when LUN AFR ≈ Drive AFR

We group both LUNs from a single HDD into a RAID0 stripe. No additional redundancy is assumed.

Each individual RAID0 set gives ~2X performance, so rebuild time is ¼ ~½ of an array with the same number of similar sized single actuator drives.

Combine the set of RAID0’s into an upper parity RAID. Data protection applied at this level.

Device failure tolerance is the same as RAID5/6. Since HDD LUNs are striped together into a single volume, on HDD failure the upper parity RAID only sees a loss of one volume.