
Open. Together.
OCP
SUMMIT

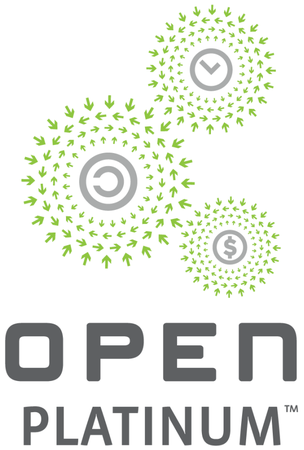
Storage

Multi-Actuator HDD Datacenter Deployment

Best Practices

James Borden – Cloud Architecture, Seagate Technology

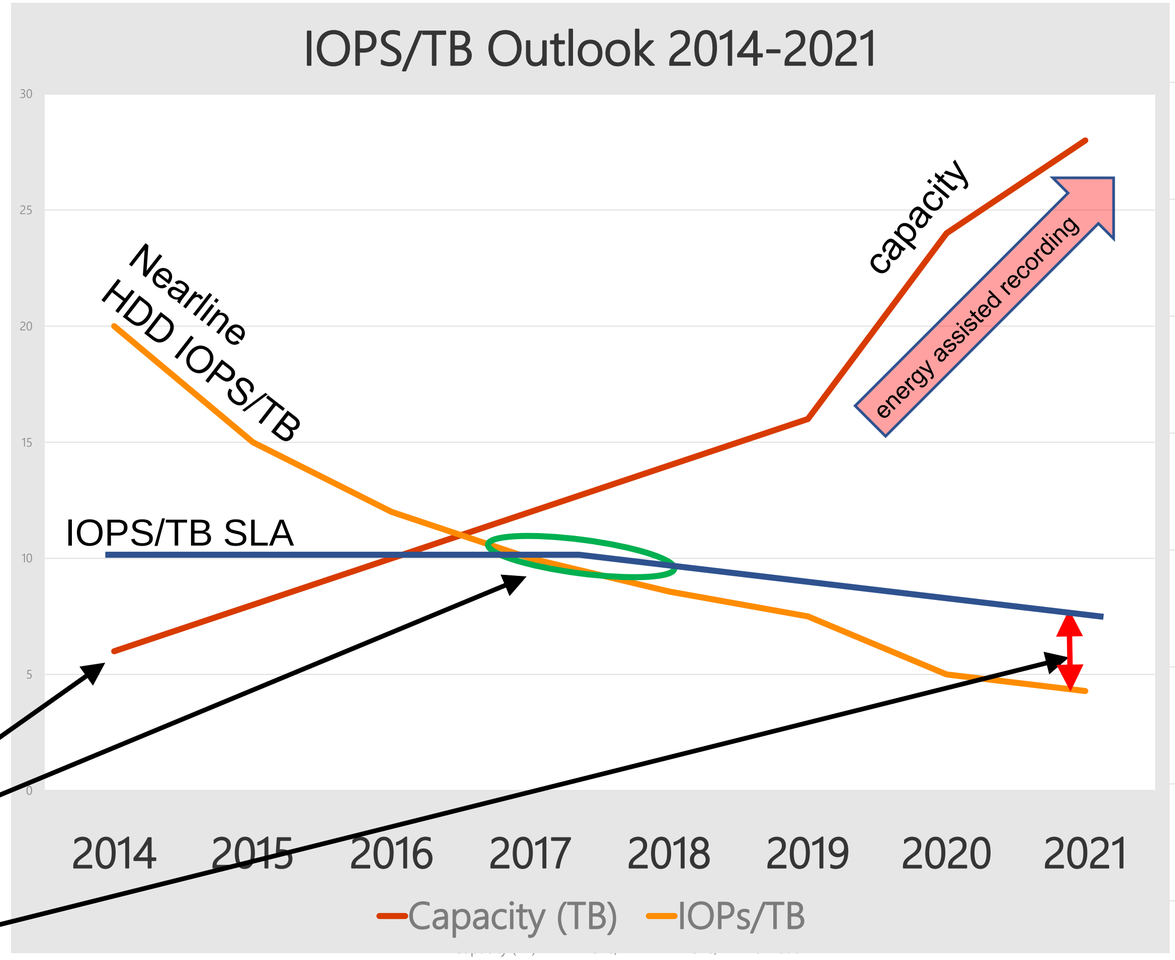
Tim Walker – Principle Engineer, Seagate Technology



Why? 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, customers may need:
  - short-stroke the HDD, leaving unused capacity stranded
  - Deploy lower capacity drives

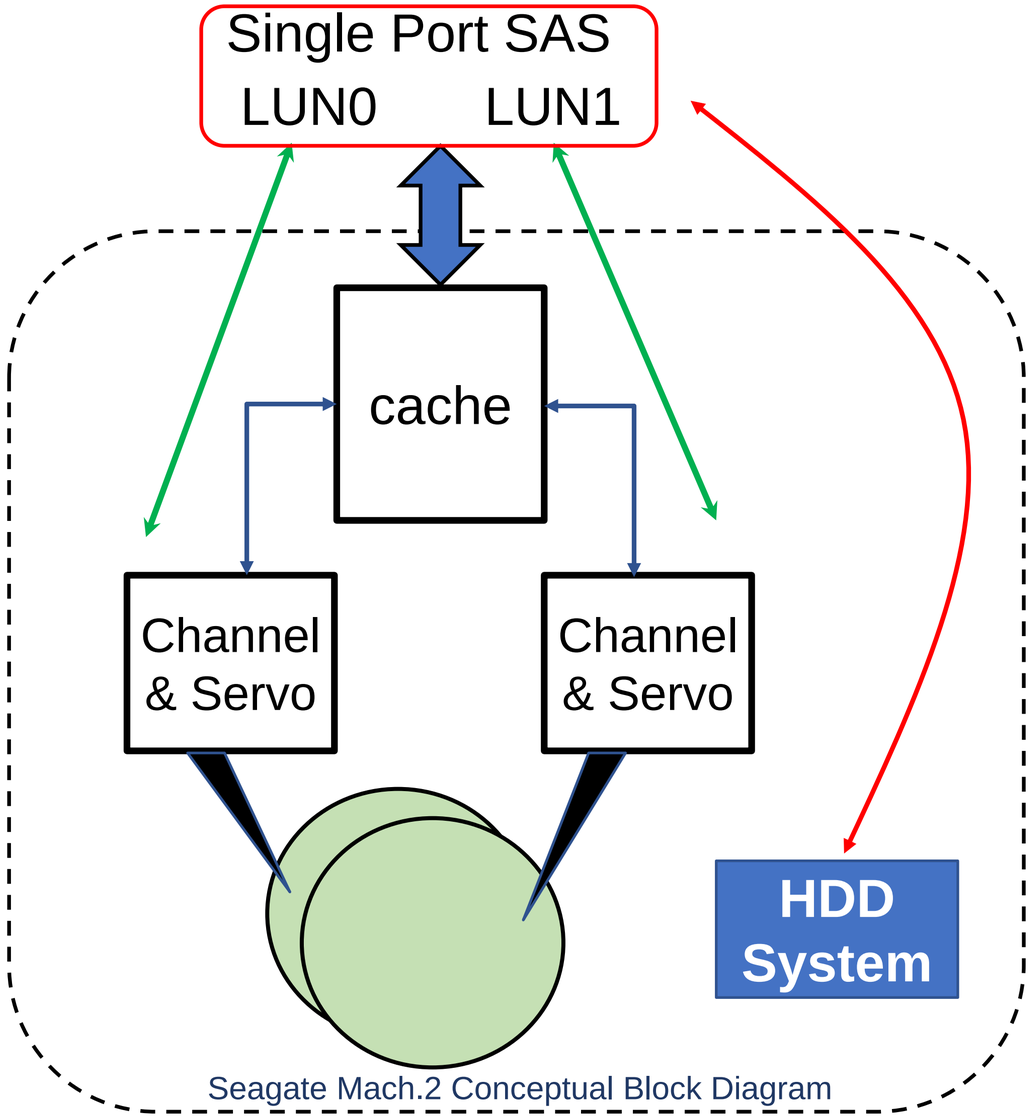
Both increase overall storage TCO



Solution: HDD Parallelism

Multiple parallel data channels

- Dual actuator:
  - 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. Can deliver up to 2X the throughput of a comparable single actuator drive
    - This design's data path is essentially independent; management and reporting path is common.



Seagate Mach.2 Conceptual Block Diagram

**** Seagate Mach.2™ Exos 2X14 (14TB CMR) is currently deployed for final evaluation in multiple customer environments, and is being qualified in Microsoft Project Olympus and Facebook Bryce Canyon***

System Integration – Device level

Major vendors' HBAs should preserve the dual-LUN presentation through the driver*: OS storage stack sees a traditional SCSI target for each of the two LUNs

Device management tools should recognize the common management plane across each HDD's LUN-pair

- Some commands affect both LUNs (e.g., SCSI Power Management)

Data management schemes should recognize the common failure domain for the LUNs in a given HDD

- Either OS or Application
- Few single-LUN failure modes (e.g., VCM driver)

Data layout should balance the workload across LUNs

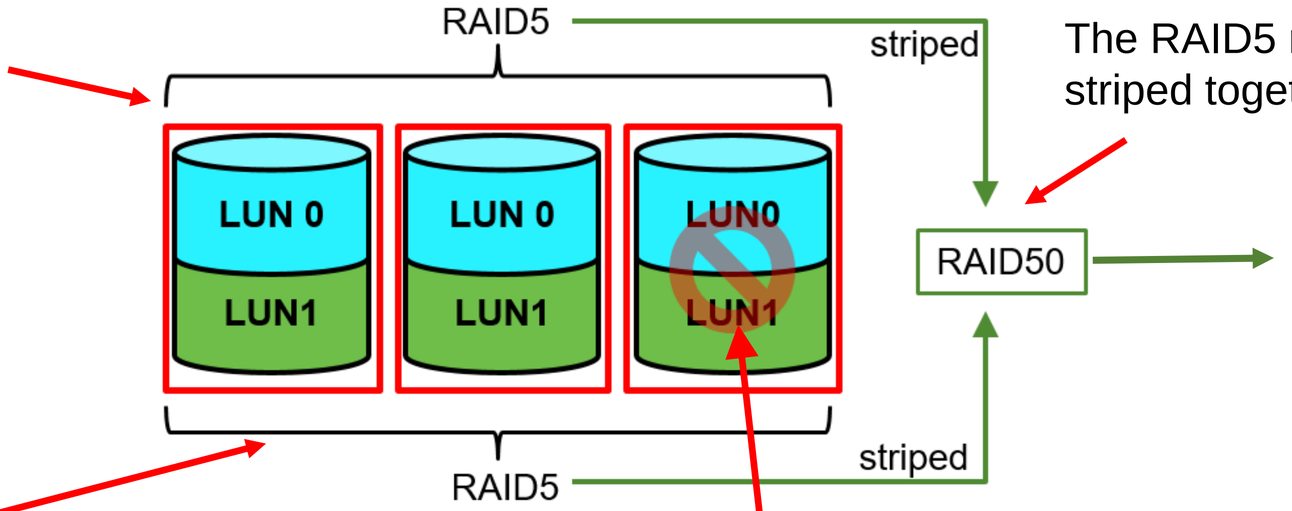
- Design criteria similar to traditional multi-LUN data layout design
- Installing a file-system-per-LUN or directly accessing device storage is appropriate

** **Broadcom**, in collaboration with Seagate, has updated and demonstrated full HBA functionality with Mach.2™ Dual-actuator HDDs in enterprise/cloud environments.*

System Integration – Hardware RAID

- The LUNs do not constitute different failure domains.
- One LUN per device assumption is prevalent throughout hardware RAID firmware and testing

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.



The RAID5 redundancy-groups are striped together into a RAID50

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

Reliability is the same as RAID5. A loss of an HDD degrades both redundancy-groups, but no data loss.

System Integration – LVM2 & GEOM

LVM2 and GEOM: volume managers on Linux & FreeBSD

Conceptually, OS-based volume manager dual-actuator approaches are similar to the hardware RAID case, but with added flexibility

Like hardware RAID, volume manager configuration design should:

- Observe failure domains
- Balance workload across actuators
- Optimize disk data layout (e.g., striping) for intended workload

System Integration – Windows Storage Spaces

Data path is clean – multi-LUN HDDs can be combined or used as needed to meet the storage workload

Storage Spaces recommends max 84 “disks” per pool

- Multi-actuator HDDs are a “disk” but present multiple LUNs per disk.
  - e.g., 4U96 chassis of Mach.2™ now reports 192 LUNs (3 vs 2 pools)

Windows Server pushes fault domains down to the physical disk, but does not yet recognize the LUNs of an HDD are in the same fault domain (get-storagefailuredomain)

Sysadmin should monitor LUNs from different *Slot:Adapter:Port:Target* tuples while assigning elements to a storage pool. Should not assign LUNs from the same HDD into a storage pool

Calls to Action

Device vendors:

- Continue to optimize the data and command paths to minimize cross-LUN interactions on multi-LUN devices

Storage vendors / stack developers:

- Continue to harden multi-LUN SCSI direct attached storage implementation, particularly for device management
- Enable more granular failure domain support for direct attached storage
  - Physical HDD LUNs are not redundant
- Watch for device=LUN stack assumptions in code and tools


Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT