Performance Requirements for WLCG Storage

US ATLAS Perspective

Lincoln Bryant
University of Chicago

WLCG Open Technical Forum
10 December 2024

Storage is Real Estate

- Storage is the most expensive service that sites provide
  - Requires significantly more care and attention than compute
  - Inaccessible or failed storage is painful for sites, DDM, users
- Sites primarily pursue storage capacity per unit money, because capacity is pledged
  - Availability is also monitored by WLCG, and therefore important
  - Performance is not pledged or accounted, and is largely considered an internal site matter

HDD performance-per-TB is eroding

- Large pools of spinning disk have been the best combination of capacity, price, and performance for a long time
- However, throughput and IOPS per TB are trending downwards year-over-year!
- This will have impacts on performance, reliability in the next 5-10 years unless the technology improves significantly

Two disks, ten years apart

4TB Seagate Enterprise Disk (2011)

Constellation® ES.3

Specifications	4TB[1,2]
Standard Model Number	ST4000
SED Model Number	ST4000
SED-FIPS Model Number	ST4000
Features	
Protection Information (T10 DIF)	
Humdity Sensor	
Super Parity	
Low Halogen	
PowerChoice™ Technology	
Cache, Multisegmented (MB)	
Reliability/Data Integrity	
Mean Time Between Failures (MTBF, hours)	1.4
Reliability Rating @ Full 24x7 Operation (AFR)	0.
Nonrecoverable Read Errors per Bits Read	1 sector per 10E15
Power-On Hours per Year	8760
Bytes per Sector	512, 520, 528
Limited Warranty (years)[5]	5
Performance	
Spindle Speed (RPM)	7200
Max. Sustained Transfer Rate OD (MB/s)	175
Average Latency (ms)	4.16
Interface Ports	Dual
Rotation Vibration @ 1500 Hz (rad/s²)	12.5

175MB/s sustained throughput

(IOPS unspecified, but benchmarks[1] around 80 IOPS)

20TB Seagate Enterprise Disk (2021)

Specifications	SATA 6Gb/s
PowerBalance™ Power/Performance Technology	Yes
Hot-Plug Support[3]	Yes
Cache, Multisegmented (MB)	256
Organic Solderability Preservative	Yes
RSA 3072 Firmware Verification (SD&D)	Yes
Reliability/Data Integrity	
Mean Time Between Failures (MTBF, hours)	2,500,000
Reliability Rating @ Full 24×7 Operation (AFR)	0.35%
Nonrecoverable Read Errors per Bits Read	1 sector per 10E15
Power-On Hours per Year (24×7)	8760
512e Sector Size (Bytes per Sector)	512
4Kn Sector Size (Bytes per Sector)	4096
Limited Warranty (years)	5
Performance	
Spindle Speed (RPM)	7200RPM
Interface Access Speed (Gb/s)	6.0, 3.0
Max. Sustained Transfer Rate OD (MB/s, MiB/s)	285/272
Random Read/Write 4K QD16 WCD (IOPS)	168/550
Average Latency (ms)	4.16
Interface Ports	Single
Rotation Vibration @ 20-1500 Hz (rad/sec²)	12.5

285MB/s sustained throughput

168 IOPS read

5x capacity but only 1.5x throughput in 10 years

[1] https://techgage.com/article/seagate-constellation-es-3-4tb-enterprise-hard-drive-review/4/

What does this mean for sites?

- Expect sites' throughput per TB (i.e., 1.5x throughput / 5x capacity) to go **down** in the future with HDDs
- With ideal (highly sequential) workloads, HDD performance will continue to be good enough and push the bottleneck toward network
- The read/write mix and sequential/random mix of real-world WLCG workloads should probably be studied
  - See *backup slide for an example of how I/O mixture affects HDD performance

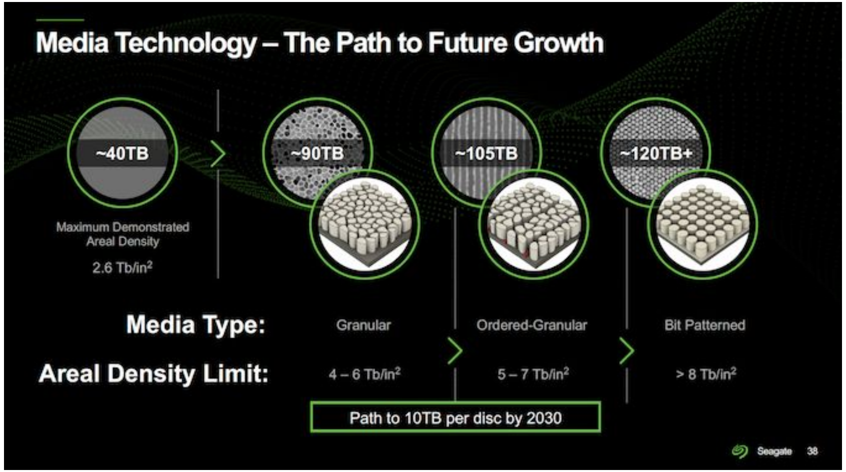
HDD reliability risks in Run 4

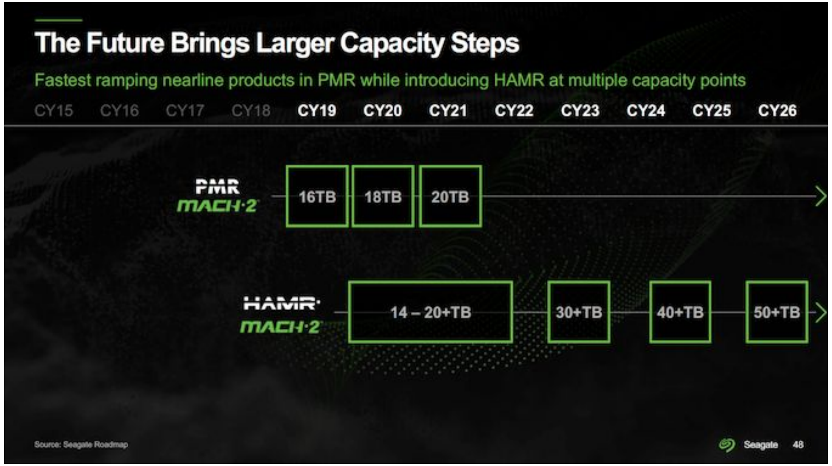
- Huge HDDs bought during Run 4 will probably have longer, riskier RAID rebuild times
  - Speculation: 3 days or more to replace one 100TB disk
- Sites may feel pressure to switch to object stores?
  - Self-healing; amortize the rebuild time across the cluster

YEAR	2011	2021	2031 (crude extrapolation)
Capacity	4TB	20TB	100TB
IOPS	80 IOPS	168 IOPS	250 IOPS
Throughput	175MB/s	285MB/s	425MB/s
RAID Rebuild Time (Perf./Capacity)	6h	19h	**65h**

Compared to marketing material

- Capacity not too far off from the manufacturer's predictions
  - https://www.tomshardware.com/news/seagate-technology-roadmap-2021





Let’s talk about solid state storage

- Despite the hype, solid state storage has not eliminated HDDs and probably will not for some time to come
  - Street price 30.72TB NVMe: $130/TB
  - Street price 24TB HDD: $20/TB
- NVMe is starting to hit attractive price points for certain types of workloads (e.g. Analysis Facilities) where the performance is worth the cost
- Sites in Run 4 may see a mix of HDD and NVMe
  - Caching layers perhaps, but even better if DDM software is aware of it

What about caching?

- Converting sites to diskless / cache-only sites could reduce a lot of operational expense
  - Disk becomes easy to operate, easy to scale, and fungible like compute
    - especially at sites where personpower is lacking
  - NVMe provides excellent performance, if a bit expensive still
  - Perhaps concentrate storage at the most reliable sites
- But poorly designed caches cause all kinds of problems:
  - If the working set size (~ratio of compute to cache) is too large for the cache, the cache will be nearly useless
  - Performance/TB problem of HDDs is much amplified (see backup slide)

Improving site networking

- Adding NVMes to our sites will also necessitate improving our network infrastructure considerably
    - One server full of NVMes can easily saturate a 100Gbps+ link
- Happily, 100Gbps networking is becoming affordable within the datacenter, even if sites are a ways off from Tbps WAN links
- These sort of site networking overhauls are largely invisible to the WLCG, but are essential to site operations

Summary

- Sites are incentivized to prioritize capacity over performance
- Performance/TB is trending downward for HDD, making them less suitable devices in the future
- Sites adding NVMe may become common in Run 4
  - Software support, pledging/accounting these high-performance resources would help
- Well-designed, diskless (cache-heavy) sites can have excellent performance characteristics with a simpler operations model

Comments/Questions?

MWT2 2014 vs MWT2 2024

- MWT2-UC *2014*
  - About 4PB total
  - 1,620 disks, ranging from 1TB – 3TB in size
  - Assuming 175MB/s throughput and 100 IOPS per disk (100% sequential read):
    - 1,620 * 100 IOPS = 162,000 IOPS
    - 1,620 * 175MBps = 283GB/s
- MWT2-UC *2024*
  - About 21PB total
  - 2,040 disks, ranging from 6TB – 20TB in size
  - Assuming 250MB/s throughput and 150 IOPS per disk (100% sequential read):
    - 2,040 * 150 IOPS = 306,000 IOPS
    - 2,040 * 250MBps = 510GB/s
- Today MWT2-UC has, compared to 2014:
  - **500% capacity with 25% more disk, but only ~40-50% more IOPS and throughput per disk**

*Worst case and the real world

- As always, spec sheet numbers and benchmarks are purely synthetic
- The numbers showed in the spec sheet slide are best possible performance
- Worst case performance **is** impactful:
  - 4K block size * 150 IOPS ~= 600KB/s per disk.
  - Even with 2000 disks (e.g. MWT2), this is only a bit over 1GB/s *for the whole storage pool* in the worst case
- Real world will have a mix of random, sequential I/O, mix of read/write (70/30 maybe?)
- The bottom line: **Bad workloads can seriously impact HDD storage pool performance**

```
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.58    0.00    1.04    5.49    0.00   92.89

Device             tps    MB/s    rqm/s   await  areq-sz  aqu-sz  %util
sda             779.80   33.14     0.00    1.10    43.52    0.86  39.52
sdb             779.60   65.27     0.00    1.20    85.73    0.93  53.36
sdc               3.80    0.03     3.20    0.16     7.58    0.00   0.10

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.58    0.00    0.99    7.18    0.00   91.24

Device             tps    MB/s    rqm/s   await  areq-sz  aqu-sz  %util
sda             622.60   26.33     0.00    0.79    43.31    0.49  36.54
sdb            2297.00  171.22     0.20    0.74    76.33    1.70  80.46
sdc               1.00    0.00     0.60    0.20     4.80    0.00   0.06

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.58    0.00    0.79    3.70    0.00   94.92

Device             tps    MB/s    rqm/s   await  areq-sz  aqu-sz  %util
sda             558.20   21.15     0.00    0.74    38.80    0.42  34.94
sdb             559.00   21.48     0.00    1.11    39.35    0.62  40.56
sdc              51.20    0.54    85.20    0.64    10.70    0.03   0.22
```

sda and sdb here are two 12-disk RAID-6 arrays on a random, newer storage node at UChicago
Note the `%util`, `MB/s` and `TPS`
(sampled at 5 second intervals)

**HDD cache back-of-the-envelope perf.:

- Suppose a typical 2U cache server with HDDs:
  - 24x 24TB disks in two RAID 6s = 480TB usable
  - 280MB/s (100% sequential read, from Seagate spec sheet)
    - Assume 50-100MB/s read per disk with ongoing mixed read+write operations
  - 50 to 100MB/s * 24 disks = 1.2 to 2.4GB/s
  - Assume each job reading 4MB/s continuously on avg
  - 2.4GB/s / 4MB/s = 600 job slots before the cache is completely stressed!