Storage considerations for Data Challenges and beyond

Lincoln Bryant
Judith Stephen

Motivation

- Share tunings with each other since we all have approximately the same workloads
- Try to quantify what the storage capabilities are at each site in terms of IOPS, throughput
- Consider storage trends, new technologies etc for evolving site designs in the future

Understanding site configurations

- We should be able to establish an approximate upper bound on site storage performance based on specs and configuration
- Let's take the opportunity to compare notes on storage and network tunings!
- Storage:
  - Strip(e) sizes
  - Controller caching parameters (Read Ahead, Write Back, etc)
  - Sysctls like I/O scheduler, queue depths, NR requests, etc
- Network:
  - Jumbo frames everywhere
  - IRQ balancing
  - Driver versions, Kernel versions
  - FasterData tunings

Tunings at MWT2

- These are not the ”best” tunings! Just the current tunings

Storage Configuration	@ MWT2
Stripe Size	512KB
Controller Caching	Read Ahead, Write Back
I/O Scheduler	noop
NR Requests	16384

Network Configuration	@ MWT2
MTU	9000 (Jumbo Frames)
IRQ Balancing	Enabled
Driver, Kernel Versions	Stock
FasterData Tunings	Yes

Performance grows slower than capacity

- IOPS/TB is trending downwards year-over-year
  - As we know, spinning disk capacity (TB) is going up every year
  - However, spinning disk performance (IOPS) is grows very slowly by comparison
  - Mechanical devices are fundamentally limited by rotation speeds, number of platters, read/write heads, etc
- We should be aware of this and other disk trends when planning our site upgrades for the HL-LHC era

Two disks, ten years apart

4TB Seagate Enterprise Disk

Constellation® ES.3

Specifications	4TB[1,2]
Standard Model Number	ST4000NM0023
SED Model Number	ST4000N
SED-FIPS Model Number	ST4000N
Features	
Protection Information (T10 DIF)	Y
Humdity Sensor	Y
Super Parity	Y
Low Halogen	Y
PowerChoice™ Technology	Y
Cache, Multisegmented (MB)	1
Reliability/Data Integrity	
Mean Time Between Failures (MTBF, hours)	1.4 million
Reliability Rating @ Full 24x7 Operation (AFR)	0.63%
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

5x capacity but only
1.5x throughput in 10 years

20TB Seagate Enterprise Disk

Specifications	SATA 6Gb/s
C	
S	007D
S	000D
S	
F	
H	
C	
P	
S	
L	
P	
PowerBalance™ Power/Performance Technology	Yes
Hot-Plug Support[1]	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)	200RPM
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

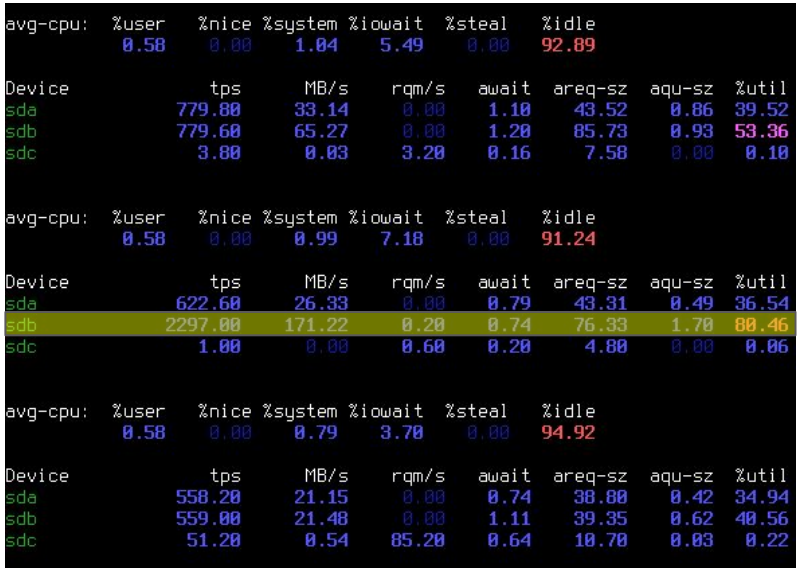
168 IOPS read

How does this apply to sites?

- 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 last slide are best possible performance
- Worst case performance **is** impactful:
  - 4K block size * 150 IOPS ~= 600KB/s per disk.
  - 600KB/s * 2,040 disks = **1.22GB/s (aggregate!!)**
- Real world will have a mix of random, sequential I/O, mix of read/write (70/30 maybe?)
- The bottom line: **Certain kinds of workloads can stress the storage before the network**



```
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)

Reading the HDD tea leaves..

- Capacities will continue to trend upward, and we don't see any indication that HDD IOPS will materially improve in the short term
- If you believe manufacturers' marketing, they'll be shipping 50 to 100TB size HDDs by 2030
  - Probably will have some non-negligible amount of solid state disk onboard
  - This helps with buffering and caching but not heavy sustained workloads
- This also has effects on raid rebuild times.
  - Today replacing a 24TB disk with 285MB/s sequential write → about **24 hours** for a rebuild, best case!
  - Imagine rebuilding a 100TB disk...

Let's talk NVMe

- At UChicago, we have demonstrated that a single server with 7x 15.36TB NVMe (~$300/TB) and 2x100Gbps NIC can use a substantial chunk (~160Gbps) of our WAN link
  - Just **one** NVMe would match all MWT2 spinning disks in terms of pure IOPS
    - a single Intel P5316 spec'd at 800,000 IOPS read, 7GB/s read/write
- Flash capacities are starting to surpass HDD (you can buy 30TB SSDs) and might catch up in terms of $/TB by the end of the decade
  - How will this affect our storage *and* network investments?
- At what price point do they make sense as a significant portion of T1/T2 storage? And how will we use them?
  - If caches, this will be driven by working-set-size
  - Do we just mix them into the HDD storage pools? Can we be smarter and place specific data types on NVMe?
- Other uses?
  - From the IRIS-HEP challenges we already see a need for a significant chunk of all-flash storage in the Analysis Facility

Comments/Questions?