NDGF All Hands 2019-2

Tape pool hardware

Tape pool storage config revisited

- Tape carousel means more usage
- All affordable storage devices today have some kind of wear based on usage
- Back-of-the-napkin calculations showed that we’ll need to cater for this

Spinning disks - Workload rating

- Both read and writes count (it’s the time the HDD head is lowered to the platter)
- Warranty valid even if you are outside the workload rating
  - Expect more HDD failures
  - Expect performance impact due to close-to-failure conditions
- Typically 500 TB/year for midline/nearline HDDs
- Be aware of Cloud/Archival/NAS HDDs (around 200 TB/year)
- Avoid desktop HDDs (not even rated to spin 24/7)

SSD/flash - Endurance

- Only writes count
- Out of warranty if endurance exhausted
  - Most devices today do throttling to ensure 3y/5y lifetime
- Two ways of defining the same thing
  - (Full) Drive Writes Per Day - DWPD
  - Terabytes Written - TBW
- Three main classes (definitions from Dell, but similar for all vendors)
  - Read Intensive - 1 DWPD
  - Mixed Use - 3 DWPD
  - Write Intensive - 10 DWPD
- Also boot-grade and similiar which can be whatever, 0.1 DWPD or worse

Murphy - didn’t get vendor replies in time

- This presentation was supposed to be based on actual configs/pricing from vendors, but I didn’t get any responses before leaving home
- Got configs from one vendor 06:50 this morning…
  - Not optimal configs due to budget constraints
  - No clear answer on throttling behavior
- And from another vendor after lunch (not included here)
- Boundary conditions:
  - ~30k EUR budget for four machines
  - 10GigE hosts: min 2500 MB/s storage bandwidth
  - 25GigE hosts: min 6250 MB/s storage bandwidth

HDD based tape pool

- Typical config: 24xLFF HDDs to meet 2500 MB/s bandwidth requirement
- RAID60 -> 20 HDDs holding data
  - Even with 2 TB HDDs we'll get 40 TB raw storage
- Workload: 20*500 => 10000 TB/year
- Tape pool workload: Assume data is written and read once respectively
  - 10000 / 2 => 5000 TB/year
- Side note: The offer we got for a HDD config within budget was 12 HDDs per server which won't cut it performance wise

SSD Read-Intensive based tape pool

- 16 SSDs @ 520 MB/s in RAID50 for a total 7280 MB/s storage bw
  - Guessing that RAID controller won't do full 7 GB/s though?
  - 960 GB SSDs
    - 0.96 * 14 => 13.44 TB raw storage
    - 1.0 DWPD => 13.44 TB/day
    - 9.6 * 365 => 4906 TB/year which is also the tape pool workload
- 8 SSDs @ 520 MB/s in RAID5 for a total 3640 MB/s storage bw
  - 1.92 TB SSDs
    - 1.92 * 7 => 13.44 TB raw storage
    - 1.0 DWPD => 13.44 TB/day
    - 9.6 * 365 => 4906 TB/year which is also the tape pool workload

SSD Mixed-Use based tape pool

- 6 SSDs @ 520 MB/s in RAID5 for a total 2600 MB/s storage bw
  - 1.92 TB SSDs
    - 1.92 * 5 => 9.60 TB raw storage
    - 5.0 DWPD => 48 TB/day
    - 48 * 365 => 17520 TB/year which is also the tape pool workload

SSD Mixed-Use based tape pool examples

- 4 SSDs @ 800 MB/s in RAID5 for a total 2400 MB/s storage bw
  - 1.92 TB SSDs
    - 1.92 * 3 => 5.76 TB raw storage
    - 3 DWPD => 17.28 TB/day
    - 17.28 * 365 => 5307 TB/year which is also the tape pool workload
- 8 SSDs @ 800 MB/s in RAID5 for a total 5600 MB/s storage bw
  - 960 GB SSDs
    - 0.96*7 => 6.72 TB raw storage
    - 3 DWPD => 20.16 TB/day
    - 23.04 * 365 => 7358 TB/year which is also the tape pool workload

Random observations

- NVME is cool bw-wise but no RAID controller means sw raid
  - Might not be a showstopper, but still untested
- HDD based pools get expensive when we only look at bandwidth
- For 25GigE bandwidth SSD is likely the only sane option
- If next-gen tape drives are 800+ MB/s and we are expected to deliver effective use of multiple of those, 25GigE will be required
  - Or LanFREE, while still quirky it’s doable mapping tape devices identically using Linux/udev
- To do config/procurement we need to know what workload to expect
  - Shooting too high wastes money
  - Too low and it’ll all go down in out-of-warranty flames