Storage for HyperScalers

Mark Carlson
And the SNIA Technical Council

SNIA Technical Council White Paper

- Hyperscaler Storage
  - http://www.snia.org/sites/default/files/technical_work/Whitepapers/HyperscalerStorage.pdf
- *Hyperscaler storage customers typically build their own storage systems from commodity components. They have requirements for drives (SSDs and Hard drives) that are in some cases being put into standard interfaces. This paper explores that trend and highlights features of existing and new standards that will meet those requirements.*

Current Situation

- Hyperscalers are very large customers
  - One estimate notes that ½ of all bytes shipped now are to Hyperscalers
  - Total (Server + Storage) Market to grow to $71.2 Billion by 2022 with 20.7% CAGR*
  - They can and do request specific features from storage devices via the RFP acquisition process
- Drive vendors will add these features in order to sell to these customers
  - Each vendor differs in how these features are implemented and in how they extend standard interfaces to accommodate them
- Software Defined Storage (SDS) products will also benefit from these features as they are added
  - Many Enterprises are taking advantage of the Hyperscalers techniques by using SDS

* Allied Market Research

One Hyper-scaler's thoughts

- For FAST '16 (USENIX File And Storage Technologies) conference, Eric Brewer (VP of Infrastructure for Google and also credited with the CAP Theorem) talked about a white paper they had just published.
- The Metrics for disk are: IOPS, capacity, lower tail latency, security and lower TCO.
- They would like "APIs" to:
  - Control timing over background tasks (when tail latency is an issue)
  - Leverage disk's knowledge of details (which blocks are responding slow)
  - Prioritize requests, but still allow disk firmware to do the scheduling
  - Must be an abstraction layer with multiple implementations (a Standard does this)
  - Have a Per I/O retry policy (Try really hard, or fail fast)
- Mostly due to the tail latency problem (further slides).

The “long tail” of Latency

- Krste Asanović's keynote at the 2014 FAST conference cited The Tail At Scale by Dean and Barroso in drawing attention to the importance of tail latency at data center scale.
- At FAST ‘16, Mingzhe Hao *et al*'s *The Tail at Store: A Revelation from Millions of Hours of Disk and SSD Deployments* paid homage to Dean and Barroso's work as they analyzed:
  - storage performance in over 450,000 disks and 4,000 SSDs over 87 days.
  - an overall total of 857 million (disk) and 7 million (SSD) drive hours.
  - 0.2% of the time, a response to a particular I/O is more that 2x slower than the other I/Os in the same RAID group (e.g. one of the drives was slow in responding)
  - And 0.6% for SSD.
  - As a consequence, disk and SSD-based RAID stripes experience at least one slow drive (i.e., storage latency tail) 1.5% and 2.2% of the time respectively.

Hyperscale Infrastructure for drives

- Higher layer software handles data availability and is resilient to component failure.
  - Thus no need for expensive (No Single Point of Failure) storage systems
- Primary model has been Direct Attached Storage (DAS) with CPU (memory, I/O) sized to the servicing needs of however many drives of what type can fit in a rack's tray (or two).
  - See the OCP Honey Badger
- With the advent of higher speed interfaces (PCI NVMe) SSDs are moving off of the motherboard onto an extended PCIe bus shared with multiple hosts and JBOF enclosure trays.
  - See the OCP Lightning proposal
- Custom Data Center monitoring (telemetry), and management (configuration) software monitors the hardware and software health of the storage infrastructure.

Software Defined Storage (SDS)

- Presents Object (and file) access to a scale out storage infrastructure.
- Operates processes on multiple “hosts” to coordinate access to underlying drives.
- Stripes across drives (may be virtual) and adds redundant copies (or protection information) for both data availability and data protection.
- Also can speed response time by responding to the request when enough drives have responded with sufficient data to reconstitute the response.
  - But this is where the long tail comes in
- Scales out by adding identical nodes and then hosting some of the processes.
- Can be virtualized if needed.
- Can include Geographic replication.

Tail Latency Remediation

- A per I/O tag to indicate that the data is part of a stripe would allow the drive to fail fast and return an error.
  - NCQs are partial solution, but are not “per I/O”
- The Data Center monitoring software may detect these slow drives and mark them as failed (ref Microsoft Azure) to eliminate the latency issue.
  - Even when it’s only a small part of the media that is slow
- The Software Defined Storage then automatically finds new places to replicate the data and protection information that was on that drive.
- But this detracts from TCO, due to having to replace the drive in the field.
  - Not immediately, but typically on the next replacement pass through the data center
- If the drive kept track of these slow media locations and could not remap them to faster areas, a means to shrink the size of the media down to only fast locations (plus new spare) is needed.

Current Drives

- Don't allow per I/O hinting (yet?).
  - See RAID rebuild mode & IO hints
- Remap LBAs only on true media failure.
  - After vendor dependent number of retries (re-writing?)
- Obviously keep track of failed media areas.
  - But do they keep track of media areas with slow response times (to predict failure)?
  - Is slow response time repeatable (not always)? Does it discount background tasks?
- Start really failing when they run out of spare media to remap.
  - Spare media sized to give the drive a certain life (MTTF) on average, reduce factory returns, increase yields
  - But if it reduced spare media for slowdowns would this sizing need to take that into account? (Mean Time To Size Reduction)
- Try and keep the data and serve it up at all costs.
  - Again to minimize factory returns and aid TCO, but…

DePop

- Google, Microsoft (Azure) and other Hyperscale storage customers have specific requirements for Hard Drives and SSDs
  - Long tail of performance biggest problem today
- Repurposing and Data Preserving DePop proposals submitted to T13, T10
  - Addresses the long tail by removing slow physical elements from the LBA address space
- As LBA ranges starting responding slow (and causing long tail results), the physical element (at as small a grain as possible) status is indicated in a Physical Element Status Descriptor
- Assuming it is consistently slow due to media problems (not garbage collection, etc.)
- Host then performs an operation that results in a new, lower capacity with those physical elements removed
  - Drive is returned to service and filled with new data by Software Defined Storage

Drive would need to

- Keep track of Physical Element response times and log any that were slow to respond via corresponding LBAs.
  - May also remap any LBAs as a result
  - Might serve LBAs out of cache temporarily
  - Competitive Benefit to keep these as small as possible, but large element list as a result
- Update entries in the Physical Element Status log page.
  - Indicate Health status: Within, At, or Beyond design limit (vendor specific)
- Support the “REMOVE ELEMENT AND TRUNCATE” command.
  - Currently: Issued for each sub-element at host’s discretion, would re-issue if more than one

Streams

- Streams is a concept that associates multiple blocks with an upper level construct such as a file or object
  - It is likely that all these blocks will be deleted as a group and therefore knowledge of this grouping can reduce garbage collection of unused write locations
  - SSDs can consolidate the LBAs associated with the stream into one or more write blocks
  - When used with the TRIM command or UNMAP command, entire write blocks can then be erased
- This improves performance and minimizes write amplification due to garbage collection
  - This also improves the life of the device
- For the NVMe interface this is part of the Directives proposal targeted at version 1.3
- For SAS, this is supported by the WRITE STREAM command among others
- In SBC-4 this is the Stream Control sub-clause 4.24. For SATA, proposals are being processed

Advanced Background Operations

- Drives do perform operations asynchronously from host requests. These operations may include:
  - Garbage Collection
  - Scrubbing
  - Remapping
  - Cache flushes
  - Continuous Self Test
- These operations may delay an I/O operation from a host and lead to tail latency
- By giving the host some ability to affect the scheduling of these operations, these background operations may be scheduled at a time that reduces the impact on I/O operations, reducing the impact of this effect on tail latency

Advanced Background Standards

- Advanced Background Operation proposals are targeted for the NVMe interfaces
  - For NVMe this work is a proposal targeted at version 1.3
  - For SAS, this work has added a BACKGROUND CONTROL command as defined in SBC-4 Background Operation Control
  - For SATA, this work is defined in ACS-4 Advanced Background Operations feature set

Open-Channel SSDs

- An Open Channel SSD is a solid state drive which does not have a full firmware Flash Translation Layer (FTL), but instead presents the full physical address space of the storage media to a host
  - The FTL is then run on the host and may communicate through an interface such as NVMe
- An open source implementation of the required FTL is typically used to achieve wear leveling, garbage collection and sparing
  - This gives the Hyperscale customer control over the access to the physical media allowing it to control the FTL processes
  - The SDS is then able to manage tail latency
- The LightNVM project has software to enable these types of devices
  - Using this software, the bad block management, metadata persistence and extensions to support atomic I/O are still required to be handled by firmware on the device
- Other approaches may divide the operations between host and device differently

Fast-Fail / Rebuild Assist

- For SAS and SATA, when a drive is put into rebuild assist mode (it can be enabled/ disabled as needed), the drive does a couple of things differently:
  - It doesn't try so hard to get the data
  - Because in this mode, it is assumed that the host has another copy of the data available somewhere else
- This mode can be disabled on a per I/O basis (so that complex error recovery may be done for some I/Os)
- When an error is detected, the drive tells the host not only about the error on the requested block, but it tells the host about the other errors in a related contiguous chunk
  - The LBA of the next "good" block
  - This enables the host to not bother trying to access any of those "already known to be bad" blocks, and go ahead and get the data from the other source for those blocks

Rebuild Assist Standards

- This is described in T10/SBC-4
  - Sub-clause 4.20 Rebuild assist mode
- Also in T13 (ACS-4).
- NVMe no proposal so far

Per I/O Hints

- SAS and SATA support some hints using Logical Block Markup (LBM). Examples of hints that can be used include:
  - Subsequent I/O Hint – prioritizes this I/O with regard to other I/Os
  - Read Sequentiality – probability of subsequent reads to this LBA range
    - Enables intelligent cache pre-fetch
    - This allows the data to be removed from the cache after it is read
  - Write Sequentiality – probability of subsequent writes to this LBA range
- In SAS, there can be up to 64 LBMs and each LBM contains a combination of hints. Each I/O may reference an LBM.
- In SATA the Data Set Management command may assign an LBM to a range of LBAs.
- These hints can be used by the SDS to enable the drive to better manage workloads, possibly achieving higher throughput or IOPs.

Future Proposals

- Hyperscalers (Google, Microsoft Azure, Facebook, Amazon) are trying to get SSD vendors to expose more information about internal organization of the drives
- Would like 200 µs Read response and 99.9999% guarantee for NAND devices
- I/O Determinism means the host can control order of writes and reads to the device to get predictable response times
  - Understand when Reads will be behind reads
  - Understand when Reads will be behind Writes
- Proposed a mathematical model to convey this to the host
  - May require changes in how sparing and allocation of LBAs to devices is done

Hyperscaler Techniques in the Enterprise

- Enterprises have traditionally purchased highly available storage systems with built-in redundancy and dedicated storage controllers with proprietary firmware. However growth of these types of systems has slowed and revenues are declining.
  - *Total worldwide enterprise storage systems factory revenue declined 2.2% year over year to $10.4 billion during the fourth quarter of 2015*

Changing Supply Chain

- There is a new category of storage vendors called Original Design Manufacturer (ODM) direct who package up best in class commodity storage devices into racks according to the customer specifications and who operate at much lower margins.
- They may leverage hardware/software designs from the Open Compute Project (OCP) or a similar effort in China called Scorpio, now under an organization called the Open Data Center Committee (ODCC), as well from as other available hardware/software designs.

Large Bank Example

- The bank has over 20 datacenters around the world
- They create an internal private cloud for the entire bank's IT project usage
- They cannot use the public cloud for this data as they currently need to comply with over 200 different country government regulations
- Their storage budget dwarfs the revenue of most medium sized storage vendors
- Enterprise companies that embrace a blended value model, offering software defined storage, long term retention, data life cycle management and traditional work horse storage are better suited to benefit from this shift in the industry

Large Bank Example

- Their deployed storage has 10s of thousands of nodes with around 200 petabytes of active data and 1⁄2 exabytes of inactive data
- Their overall data footprint is growing at 45% annually
  - They process 10s of trillions of transactions daily, and downtime is very expensive
- They are big enough that vendors will custom build for them, but they also have a policy of no single source for any of their hardware
  - They buy storage in 6 PB pods that are 1⁄2 CPU and 1⁄2 storage drives, pre-assembled by the ODM direct vendor
  - They installed their first such pod in 2015
  - Their next pod will be all flash
- Most importantly their cost savings is projected to be 50% over traditional storage

Software

- They use SDS with the best in class commodity hardware to create a private cloud for internal IT projects and customer facing services
  - They are currently deploying about 11% of their storage as SDS today, but have a goal of 20% by the end of 2016
  - Their goal is to grow this to 50% by 2020
- They license their SDS from a major vendor (site license) and are training up their staff in the new approach
  - They virtualize the hardware to abstract away any differences between the multiple vendors
  - From the storage service they serve up an S3 compatible interface for new projects and mainly provide block services for existing applications
- They plan to look at Ceph for SDS and NVMe for SSD interfaces in the future

New SNIA Effort

- Datacenter Storage Task Force
  - SNIA.org/datacenter
- Investigating organization approach
  - Technical Work Group (TWG) to work out how to address requirements
    - APIs, Software, Proposals to T10, T13, NVMe
  - Initiative to promote adoption of solutions and standards

Summary

- Increasing attention to the fast growing Hyperscaler storage market by drive and SDS vendors
- Current fractured approach of new features via RFP procurement does not scale
- Coordination of Hyperscaler requirements needed, starting to happen in organizations such as NVM Express
- The SNIA has the right stakeholders to help accelerate an industry response
  - Get involved in the Datacenter Storage Task Force

Hyperscale Storage at SDC

- BoF: Hyperscaler and Datacenter Storage Update
  - 7-8pm Tuesday – Cypress
- Standards for Improving SSD Performance and Endurance
  - 3:05-3:55pm Wednesday – Lafayette **Bill Martin**
- Open Storage Platform Delivers Hyper-scale Benefits for the Enterprise
  - 4:05-4:55pm Wednesday – Lafayette **Eric Slack**
- Software-Defined Flash: Tradeoffs of FTL, Open-Channel, and Cooperative Flash Management
  - 5:05-5:55pm Wednesday – Lafayette **Craig Robertson**

Questions?