The Hyperscale Challenge: Flash Deployed in a Disaggregated Model

Roark Hilomen, Engineering Fellow

Systems and Software Solutions, SanDisk Corp.

Forward Looking Statement

During our meeting today we will make forward-looking statements.
Any statement that refers to expectations, projections or other characterizations of future events or circumstances is a forward-looking statement, including
those relating to market growth, industry trends, future products, product performance and product capabilities. This presentation also contains forward-looking statements attributed to third parties, which reflect their projections as
of the date of issuance. Actual results may differ materially from those expressed in these forward-looking statements due to a number of risks and uncertainties, including the factors detailed under the caption “Risk Factors”
and elsewhere in the documents we file from time to time with the SEC, including our annual and quarterly reports.
We undertake no obligation to update these forward-looking statements,
which speak only as of the date hereof or as of the date of issuance by a third party, as the case may be.

Characteristic of Hyperscale @Scale Companies

- One or more Large common infrastructure
  - Limited HW SKU serving primary usecase, designed to be reused in ‘general’ use cases.
    - i.e; Search, Reused /shared SKU with Cloud
    - i.e; OLTP, Reused / shared SKU with online analytics
    - i.e; Offline Analytics, Reused/ shared SKU with Archival or Media serving
  - Most have DevOps capability
  - Mixture of Traditional IT (usually small) and Scale out/Web scale
  - Tech Friendly, Risk Adverse

Challenges - Infrastructure

- Networking
  - Tech Transitions (1,10,25-100,400)
- Real Estate (100’s of Racks 10’s/100’s of thousands servers)
- ‘Golden’ SKU’s (usually less then 20)
  - Compute, Memory, Storage
    - SSD(PCIE, NVME) Poor man’s memory
    - SSD (SAS, SATA) Enables use case overlay
  - Large Infrastructure drivers not necessarily T0 or T1 application
  - Everyone always believe they contribute to bottom line

Challenges - Storage

- Overlaying 'Other Use Cases' to Cloud or even Bare Metal with common SKU potential of high inefficiencies
  - Common SKU built to hold minimum requirements
    - Optimized for primary use case (cloud, search, big data, etc.)
  - Requires disaggregated storage through some orchestration means to 'normalize' other uses cases.
    - i.e. OpenStack Cinder
  - Cost of inefficiencies very high
    - 30% storage utilization (2TB per server across a rack of 40 servers @ $1/GB enterprise flash) for 10 racks == $ ½ Mil unused

Storage Disaggregation

Technology	Throughput	Latency (micro)	Notes
1Gbe	80MB/s	400+	Each side based on load/TOE
10Gbe	800MB/s	400+ (40+ RDMA)	
6G SAS/SATA	500+MB/s	NA	Based on device
25Gbe	2GB/s	400+ (40+ RDMA)	
40Gbe	3+GB/s	400+ (40+ RDMA)	
HDD	30-100MB/s	6ms	
6G SSD	500+MB/s	300-800	Based on vendor
12G SSD	700MB-1GB/s	250-600	Based on vendor (per port)
PCIE/NVME SSD	800-2GB/s	50-200	Based on vendor

** Data is Generalized, not specific to any vendor

Storage Disaggregation Protection

- Replicated Data
  - Fastest (no computation or remote data fetches)
  - Most expensive (whole number multiplier, 2x, 3x, etc)
  - Replication count based on resilience of data and origin
- Erasure Coded Data
  - Slowest (relative)
  - Least expensive (1.n multiplier)
  - Most resilient (done correctly, data can survive rack level or even data center failure)
  - Assumed Archival due to speed (on HDD)
- Local Raid / Rack Replicated
  - Fast
  - Moderate expense (2.n multiplier)
  - Expensive rebuild (limited n to y due to local raid)

Models of Storage Disaggregation

- Top of Rack Storage
  - Ignores Network OP (usually 1:3, 1:6 or more)
- Rack Adjacent Storage
  - Requires Line Rate Networking (1:1 to storage rack)
- Distributed Storage/Compute (local storage shared remotely)
  - Used local, protected by neighbor.
  - Used remote, protected by remote
- Centralized Storage Bubble
  - Network Bubble with Storage only.
  - Routed to Compute

DISAGGREGATED
STORAGE & COMPUTE

Use Case: NoSQL

- Deployment Model
  - Dedicated HW (Bare Metal), Local Flash, App replicated
    - May require traditional ‘storage array’ perform data protection (usually snap and enough capacity for days of recovery)
  - Cloud Enabled or Containers
    - Most NoSQL are low thread count limited io depth
    - Requires flash, but barely utilizes it
    - Cloud ‘stamp’ of S,M,L inefficient, requires shared storage through some type of storage disaggregation
    - Storage Disaggregation can right size VM and increase efficiencies

VIRTUALIZATION

Use Case: Virtualization

VM

- Optimized for primary use case (search, web, ecommerce, etc.)
    - Usually guarantees some IOPS (IOPS per size, Fixed IOPS per VM, Advanced models have QoS – ceilings and/or floors)
    - Single or dual networking (Important consideration when deploying flash)
        - Separation due to
            - Customer vs data traffic
            - Compliance
            - Management
    - High performance requirement potentially waste entire server to serve single VM..
        - Disaggregated Flash can minimize these waste

Use Case: Analytics

- Hadoop
  - Primary Challenge
    - Co$T
    - Perceived Endurance issue due to Shuffle/Map Reduce
    - Networking
  - Value
    - Flash in Shuffle and Map Reduce for IO bound 3-6x faster
    - Resilience (Operational Fatigue with at scale HDD)
    - Allows Orthogonal scaling of compute and storage
  - Solution
    - Erasure Coding on Flash
    - Tier to lower cost media (consumer grade hdd/ smr)
    - Remove Networking Overprovisioning (at least on Storage side)

Comparison

	HDD	SSD	PCIE/NVME
Replication vs EC (6,2) Media Latency lined up	EC: 48ms Rep: 6-20ms	EC: < 8ms Rep: < 1ms	EC: < 2ms Rep: < 1ms
Throughput (2GB/s) Large Blk Seq	12-16+	~4	1 (2 NVME)
Saturate 1Gbe port	1	0.16	0.04
Saturate 10Gbe port	7	~2	0.5
Saturate 40Gbe port	25	6	1.5

Questions?

SanDisk Booth #207

@BigDataFlash
#bigdataflash
ITblog.sandisk.com
http://bigdataflash.sandisk.com