Webscale, All Flash, Distributed File Systems

Avraham Meir
Elastifile

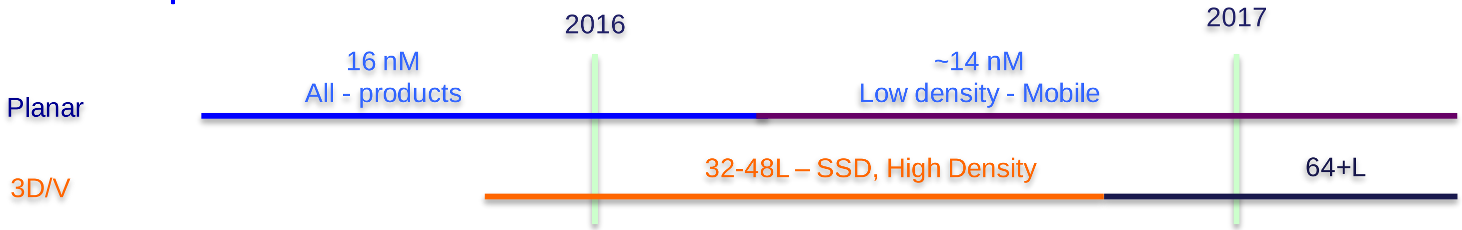
Outline

- The way to all FLASH
- The way to distributed storage
- Scale-out storage management
- Conclusion

Storage Technology Trend

NAND Technology Trends

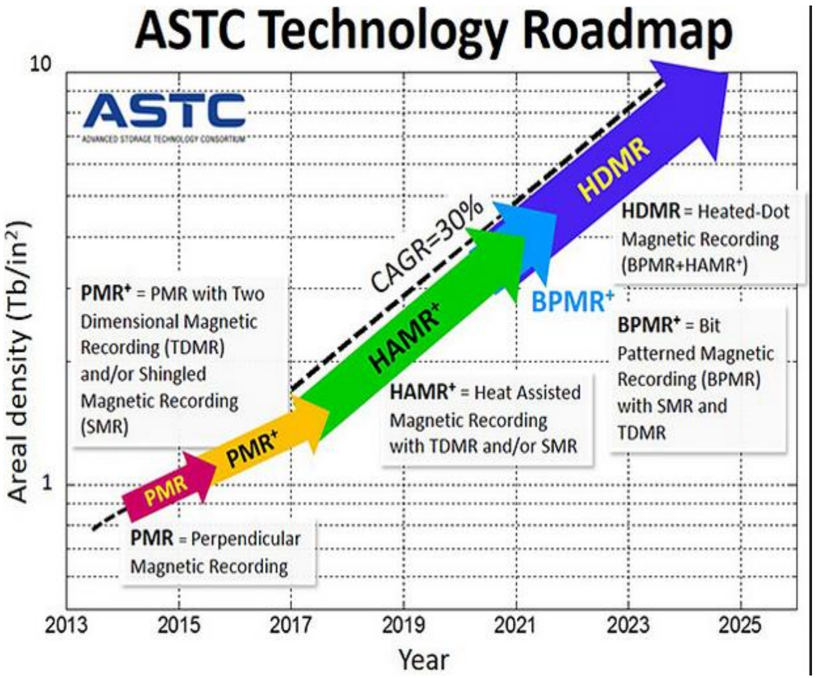
- Roadmap



- 2D/3D – 3D is not efficient for small density, 2D will continue for low density applications
- 3D Impact
  - Better endurance/performance – 3B/C main stream, potential for 4B/C
  - Clear way for cost reduction at least till 2019 (128 Layers)
- Density/Device
  - Higher capacity per package (16 dies per package becomes standard)

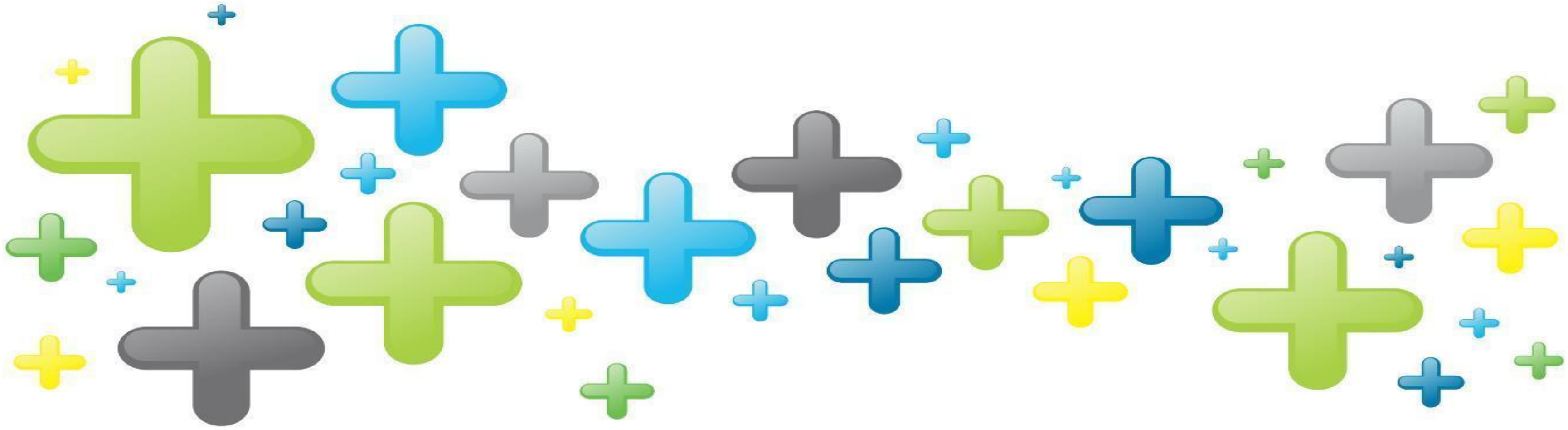
HDD Technology Trends

- Data Density
  - Increasing data density becomes harder and harder (See graph)
  - Data Dependency – Data in one track impacted by other track→ Need to read many tracks in order to recover the intermediate track data
- Outcomes
  - Cost per GB/s reduction slows down
  - Latency deteriorating



Future Data Center

- SSD – HDD Technology Parameters
  - $\GB – SSD will be closer to HDD
  - Latency\IOPS – The gap is huge and it will widen
- SSD Role
  - First tier - High end drives (3D MLC\TLC)
  - Second Tier – Client grade drives (3D TLC\QLC)
- HDD Role
  - Content storage
  - Backup systems



Distributed or Centralized

Legacy Approach Using SSD

- EMC HDD Box Example – VNX
  - 75-1,000 HDD (300 IOPS each)
  - 22K-300K IOPS → 90-1,200 MB/s
- Replacing HDD with modern Enterprise SSD
  - 75-1,000 SSD (1M IOPS each)
  - 75,000-1M KIOPS → **300-4,000 GB/s**

This design is not practical

Existing AFA Storage Box

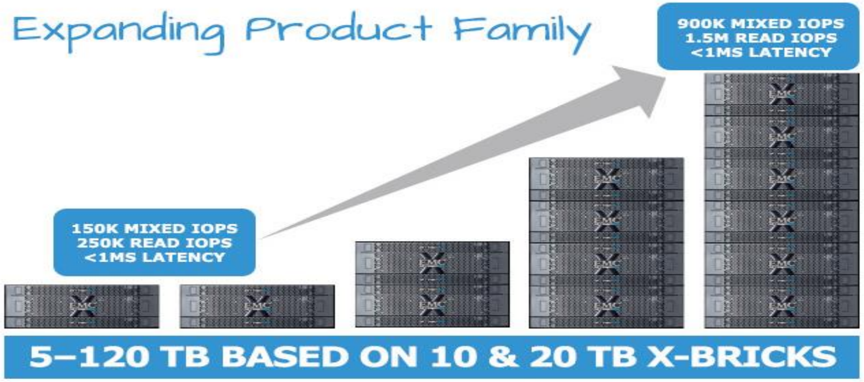
- Making the Storage Box Smaller
  - Legacy ~25 SSDs per Brick
  - 150K IOPS per brick

- Replacing SSD with fast NVMe
  - 25 modern SSDs @ 1M IOPS each
  - 25M IOPS → 100 GB/s (per brick)

- Scaling Out Bricks
  - Scaling is Limited with existing SSDs (6-8 Bricks)
  - Scaling becomes major issue with modern SSDs

- Bottleneck

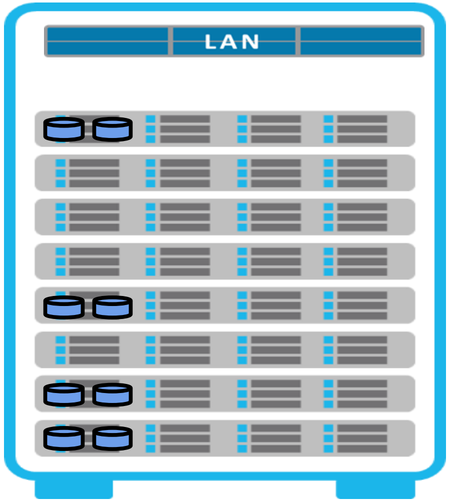
The legacy bottleneck moved from backend (drives), to the controllers and network



Reducing the Storage Box Further

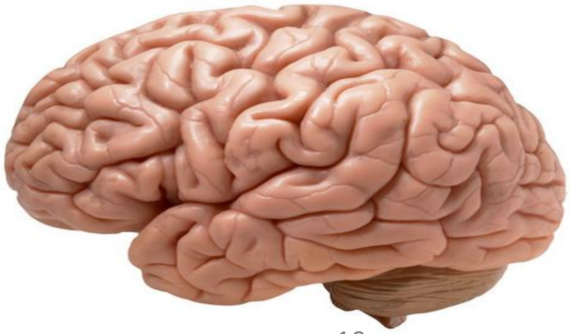
- Distribute SSDs between servers
  - Each server has 1-8 SSDs
  - Balanced Compute-Storage-Bandwidth
  - Storage SW, manages all SSDs as one name-space

    (Although it looks like DAS, it is actually NAS)
- Scale Out Easily
  - Increasing performance → just add more servers and/or SSDs



Distributed storage is not new in the market

This machine proves it

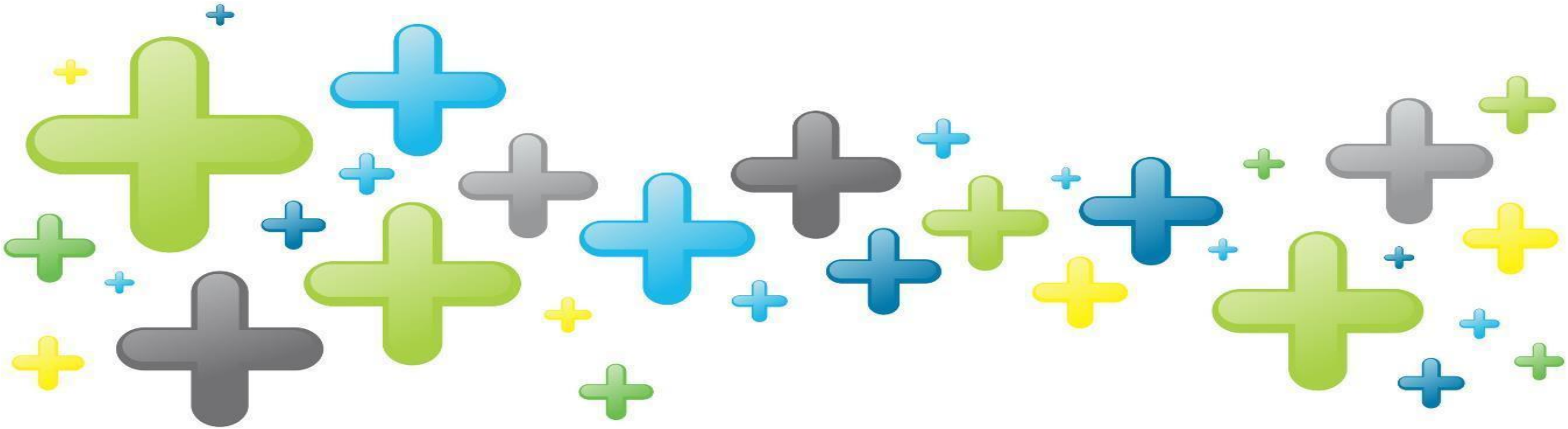


Scale-Out Architecture Benefits

- Capabilities Expansion

  System can be expanded gradually

- Heterogeneous HW
  - Servers can be purchased from multiple vendors
  - SSDs can be purchased from multiple vendors
  - HW can be upgraded to newer generation easily
- Low cost HW
  - Single port SSD vs. dual port SSD as an example



Scale-out Storage Management

Storage SW Requirements

- Cost Sensitivity
  - Should target low operating cost
  - Make use of HDD based systems like SAN/NAS/cloud-storage
- Single Name Space Management
  - Storage should be virtualized, hide all drives as one name space
- Classical Enterprise Storage Features
  - High availability, redundancy
  - Snapshots
  - Backup
- Virtualization
  - Support multiple hypervisors

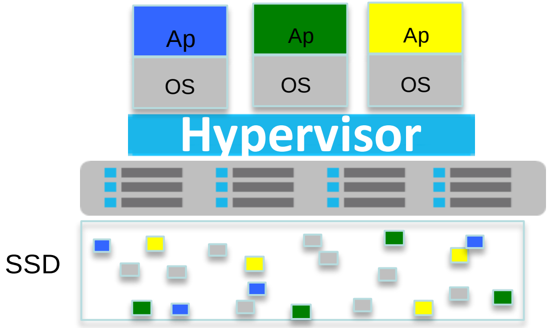
Cost Reduction Mechanisms (1)

- Data reduction
  - Compression (Online)
  - De-Duplication (online/offline)
- FLASH Tiering
  - Use high-end SSD (eMLC) for hot written data
  - Use Client SSD (cMLC, TLC, QMLC?) for cold updated data
  - Performance impact –non, cSSD has same latency and very high read IOPS
- Life cycle Extension

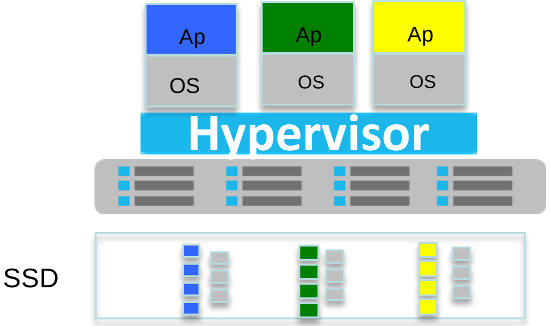
  …SSD life cycle is limited by endurance.
  - Classifying data (see next slide)

Life Cycle Extension Examples

- **Locality**
  - Virtualization cause all data to look random
  - Centralized file system can arrange the data[1]



- **Hot/Cold separation**
  - Mixing hot/cold (metadata & media for example) causes unnecessary garbage collection and reduces drive life
  - File system is cable of separating data by temperature[1]



1. Using “streams” interface is an example file-system can manage data on SSD

Cost Reduction Mechanisms (2)

- Information Life Cycle (ILM) based on External HDD
  - Move Rarely accessed data
  - Latency insensitive data objects (Media)
- Heterogeneous SSDs
  - Use different SSDs from different vendors
  - Use different NAND generations SSD[1]
  - Balancing endurance – As drives may have different age, different endurance should be considered

1. AFA uses certain generation NAND/SSD. Upgrading/Maintenance of AFA is expensive due to the need for old generation NAND/SSD.

File vs. Block Comparison

	Block	File Container
Interface	Block read/write	File (NFS/SMB)
Complexity to Develop	Low	High
IO Efficiency	Low (Local FS overhead)	High
Tiering / ILM	Limited	Highly Efficient
Flash Optimized	Limited	Highly Optimized
Sharing Semantics	Complex/Limited	Simple
Backup/Restore	Complex (Image Based)	Simple (File Based)
Snapshot	Restore all volume	Restore single file
Networking cost	High (FC)	Low (Ethernet)

Conclusions

Enterprise Grade Storage
with
Public Cloud Agility

All-Flash vNAS

File, Object & Block

High Performance
File Services

Millions IOPS, < 2msec latency
@ ~ $0.5 per GB (Usable Capacity, Incl. Media & SW)

“SpecSfs2008 like” Benchmark Performance

@ AWS All-Flash, Linear

Flash Memory SUMMIT

Elastifile — Scale-out, Performance, Manageable

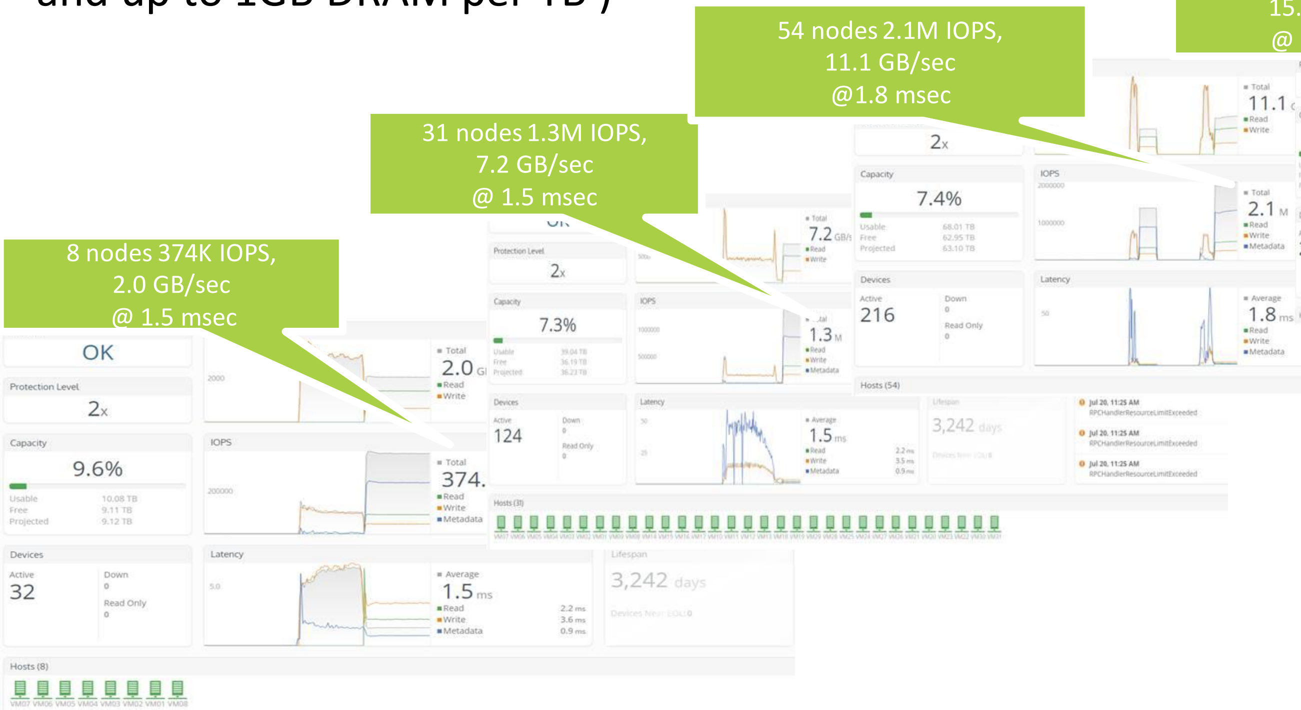
From 8 to 64 Hyper-converged nodes, 4 Local SSDs each, 10GbE, NFSv3 connectivity up to 64 clients “Specsfs like” workload, 2 way replication (utilizing approx. 20% of the Core count and up to 1GB DRAM per TB)

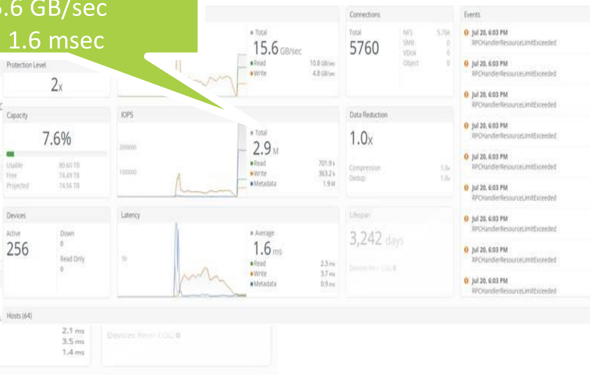
64 nodes 2.9M IOPS, 15.6 GB/sec @ 1.6 msec

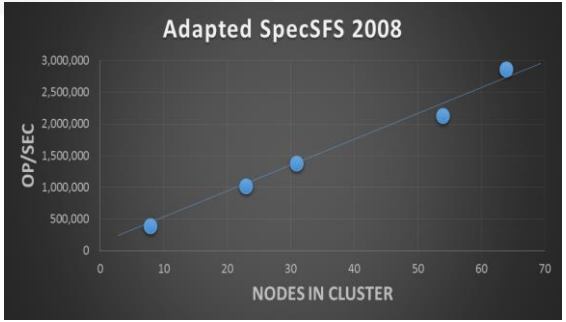
54 nodes 2.1M IOPS, 11.1 GB/sec @1.8 msec

31 nodes 1.3M IOPS, 7.2 GB/sec @ 1.5 msec

8 nodes 374K IOPS, 2.0 GB/sec @ 1.5 msec







Thank you for your attention

Conclusions

- Mainly FLASH: very attractive approach from roadmap perspectives
- Balanced Compute-Storage architecture: Enables easily scaling @ low cost
- File based storage management – looks as the preferred solution for scale-out storage systems

Technology Acronyms

- MLC vs. eMLC (HET in Intel)
  - eMLC is basically same as MLC
  - It has better screening, and different tuning
  - Schedule – Lags after MLC by 6-9 months
  - Usage – high endurance enterprise drives with special controller
- OVP Grades
  - Clint SSD - ~5% (1024 GB) → 0.3 DWPD
  - Read Intensive/Value – ~15% (960 GB) → 1-3 DWPD
  - High/Mid - ~30% (800GB) – 10 DWPD
  - High - >30% (800 GB) – SLC? 25 DWPD

Title of Presentation

Subtitle

And

Presenter(s)

Some Notes about this template

- The first action you should take is to save this presentation
  - You have opened a design template (.pot)
    - Need to save as .ppt
- A master exists for:
  - Slides
  - Handouts - default is 3 to a page
    - You can print a different number, but no guarantees about appearance
  - Notes

Free-format slide title