

# Sharing Over the WAN in the age of BigData

Dave Cohen, Intel  
ASCR Intelligent Optical Network  
Infrastructure Workshop 2014  
Plenary Talk

# Disclaimer

The contents of this presentation are derived from publicly available sources.

The slides presented here are a technical survey of the architecture common to Cloud and HPC platforms. It is intentionally Vendor Neutral. The slides and my presentation are not statements from my employer and are not intended to present my employer's position or strategy, nor is this presentation an assessment of products from other vendors, their position or strategy.

All copyrights used in this presentation are owned by their respective companies or individuals.

# Agenda

- I. Introduction
- II. Moore's Law vis-à-vis the Memory Hierarchy
- III. What Happens if Capacities Offered by Persistent Nonvolatile Memory (NVM) Increase and the Access Times they Offer Decrease?
- IV. A Deeper Dive into this Performance/Capacity, 2-tier Storage Architecture
- V. As this Shift to a Performance/Capacity Storage Architecture Unfolds What is the Impact on the Network?
- VI. What are the Opportunities for E-Sciences?
- VII. Conclusion

# Abstract

Advances in computational power has given rise to an unfathomable diversity of methods and tools for science across all disciplines. One challenge this activity presents is the unprecedented quantities of data it produces. This so-called “BigData” phenomena is a dominant consideration in planning for future computational infrastructure. In this talk we’ll take a look at the current status and future expectations/vision of how networks and storage systems will evolve to meet this BigData challenge. An emphasis will be on the influences of the emergence of SDN and storage system technologies on the solution space. Finally, we’ll take a look a BigData driven application.

# About Me

Dave is a Senior Principle Engineer in Intel's Communications and Storage Infrastructure Group (CSIG), part the larger Data Center Group (DCG). In his role as an architect he focuses on the intersection of Networking and Storage. Dave joined Intel last year from EMC where he provided technical leadership in the areas of Network Virtualization, Next-Generation Data Center architecture, and Cloud Storage. Prior to EMC Dave worked on a variety of distributed systems problems, most recently on Wall Street.

Section #1

# **MOORE'S LAW VIS-À-VIS THE MEMORY HIERARCHY**

# Moore's Law

- 1 Original prediction in 1965, revised in 1975
- 2 “The number of transistors will double every 2 years”
- 3 From 1971 to 2011 there was a million-fold increase: 40 (yrs) x 2x (density every 2 years)
- 4 This is expected to continue with projection for another factor of 100x through 2024

What's happened to the memory hierarchy over that same period?

# Over the same time period, the Memory Hierarchy has not changed

- L1 cache, 64KB, ~4 cycles (2ns)
- L2 cache, 256KB, ~10 cycles (5ns)
- L3 cache (shared), 8MB, 35-40+ cycles (20ns)
- DRAM/Main Memory, GBs up to TBs, 100 to 400 cycles
- Solid State/Nonvolatile Memory, GBs up to TBs, 5k cycles
- Disk, up to PBs, 1m cycles

Disk is not keeping up!

NVM is replacing disk to cover the gap

Section #2

**WHAT HAPPENS IF CAPACITIES OFFERED BY PERSISTENT NONVOLATILE MEMORY (NVM) INCREASE AND THE ACCESS TIMES THEY OFFER DECREASE?**

# Historical precedence

1. The Atlas Computer (1962) introduced Virtual Memory as a means of shielding the developer from dealing 3-tiers: Main Memory, Drum Memory, and Tape
  - Introduction of the concept of Shared Storage split between a Performance Tier and a Capacity Tier
  - Performance tier via Drum memory: low-capacity, fast access/nonvolatile and Capacity tier via tape: relatively high capacity/slow access
2. The Multics Computer (1964) extended this concept by introducing segmentation to better support sharing

# A Look at the Nonvolatile Memory Roadmap

- Today - Solid State Drive (SSD) retained disk drive mechanicals while replacing spinning media with NAND
- Today - PCIe-based NAND moves the NAND on to PCIe
- Emerging - DRAM-based NAND moves the NAND on the memory bus
- Future - NAND gets replaced by Storage Class Memory (SCM)

# Anecdotal Evidence supporting the NVM Roadmap

1. The Emergence of I/O Forwarding in Leadership Class systems: Argonne's Intrepid (BlueGene/P), ORNL's Titan (Cray XK7), etc.
2. Samsung's V-NAND - a shift from 2D to 3D; stacking NAND cells vertically to increase capacity while decreasing space
3. HP's "The Machine" - a system purpose built to exploit disaggregated, "shared" memory based on HP's investments in Memristor technology

# A Look at the Spinning Disk Roadmap

- SATA-based drives
- Shingled Magnetic Recording (SMR) drives
- Emerging - Optical Storage (ala Frank Frankovsky's startup)

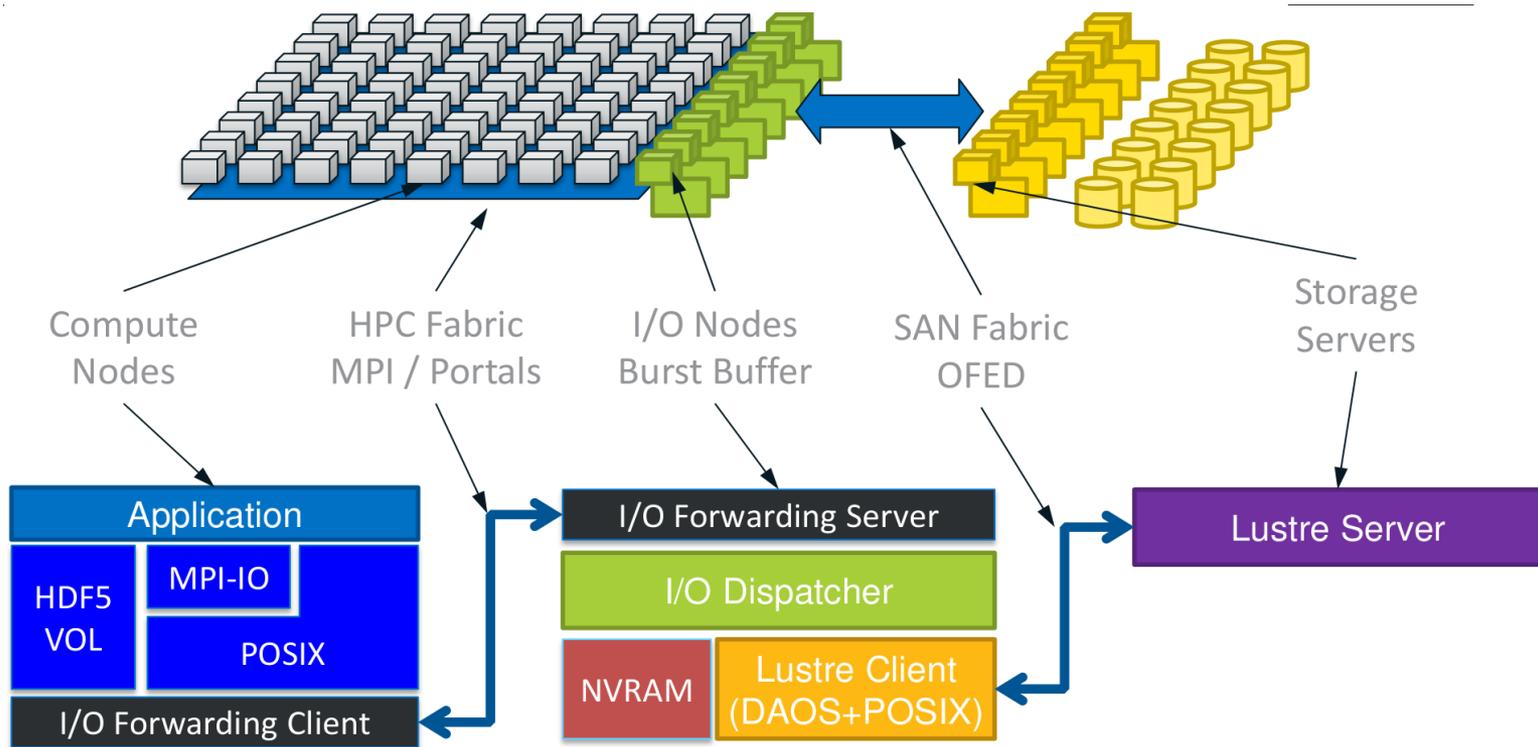
Section #3

# **A DEEPER DIVE INTO THIS PERFORMANCE/CAPACITY, 2-TIER STORAGE ARCHITECTURE**

# Warm Storage and Log-structured Memory as the basis for the Performance Tier

- All data path accesses are via DRAM
- DRAM is backed by Nonvolatile Memory (NVM), initially NAND/Flash and over time a more DRAM-like media w/faster access times and much more granular segment size (as opposed to a 4k page, as an example)
- DRAM-resident Data Sets are “Checkpointed,” continuously or periodically to a distributed key-value store optimized for NVM-based persistence
- Aggregate capacities of this tier are sufficient to hold between 30 and 90 days of an organization’s “warm” data
- Once data reaches an inactivity threshold it is moved to “cold” storage, ie the capacity tier.

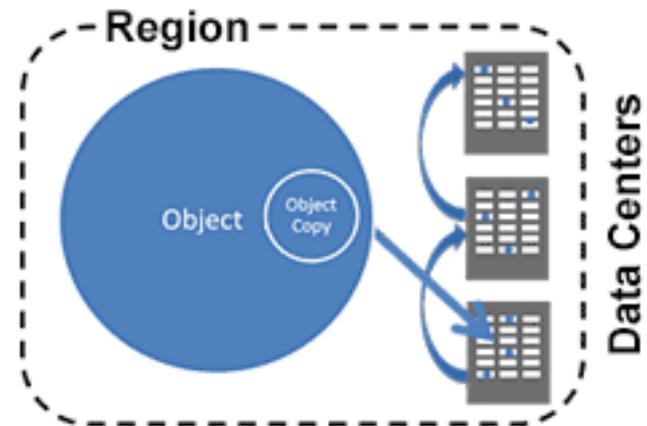
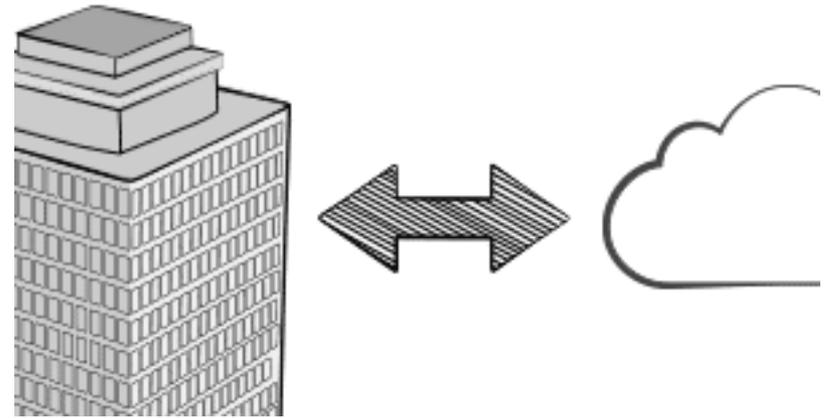
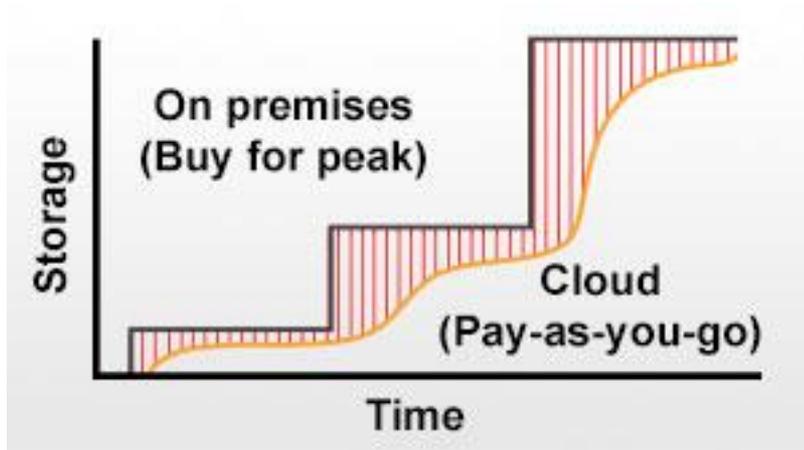
# I/O Forwarding as an illustration of how this might look



# Cold Storage and Multi-Site Erasure Coding as the basis for the Capacity Tier

- Erasure Codes provides a means of reducing the data redundancy factor while satisfying availability, protection, recovery, and retention objectives
- Distributing data and parity blocks across more than two sites allows for data redundancy factors of less than 2x

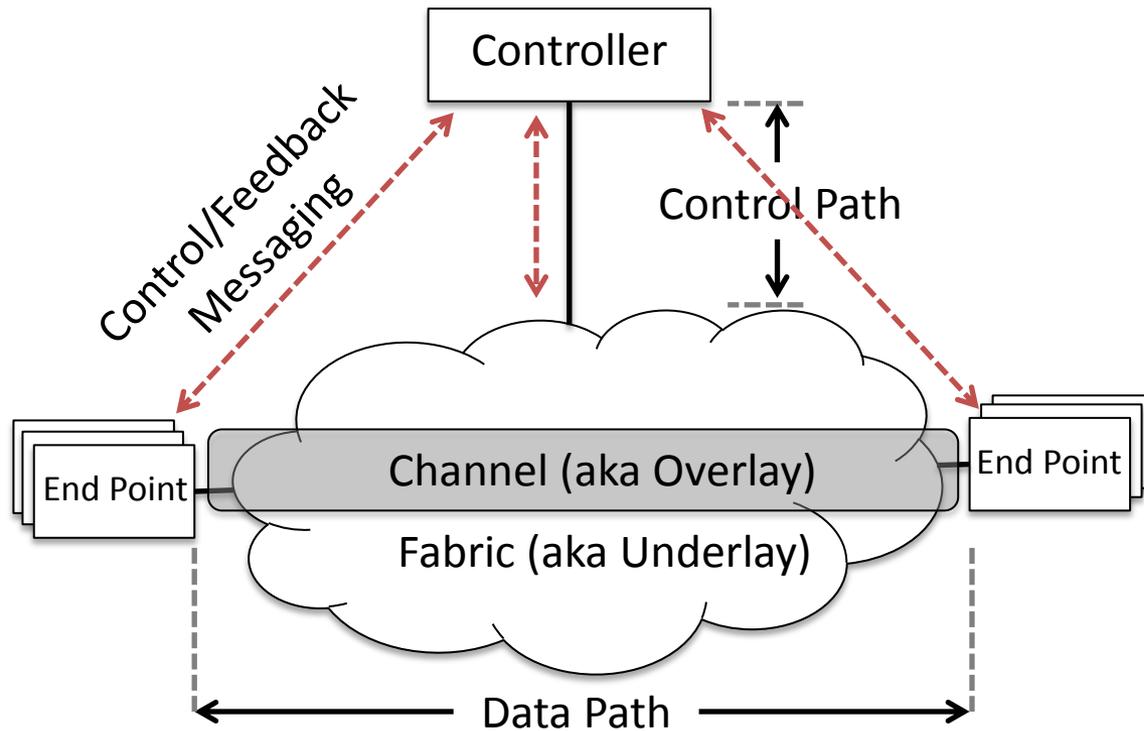
# Amazon's Glacier Cold Storage Service as the Illustration



Section #4

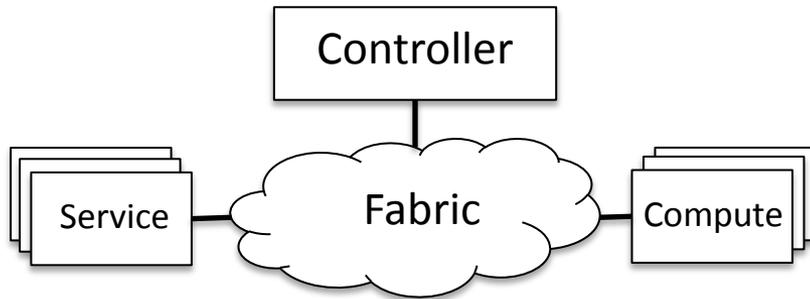
**AS THIS SHIFT TO A  
PERFORMANCE/CAPACITY STORAGE  
ARCHITECTURE UNFOLDS WHAT IS THE  
IMPACT ON THE NETWORK?**

# What is a Fabric?

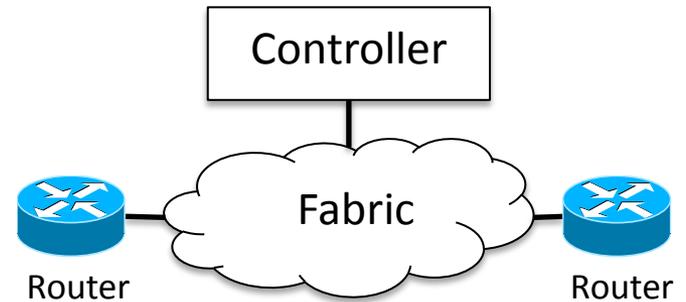


# Network Deployment Scenarios

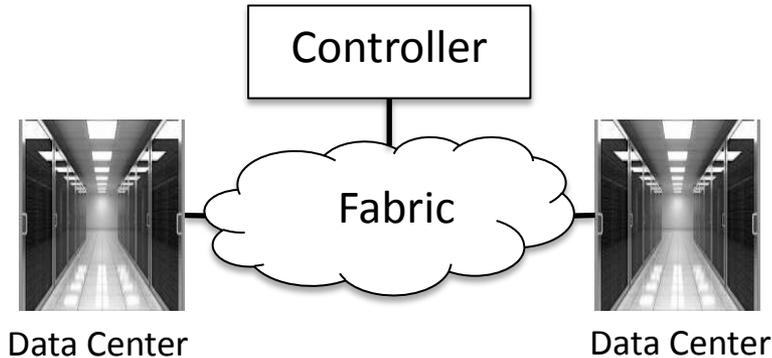
## 1. Data Center Fabric



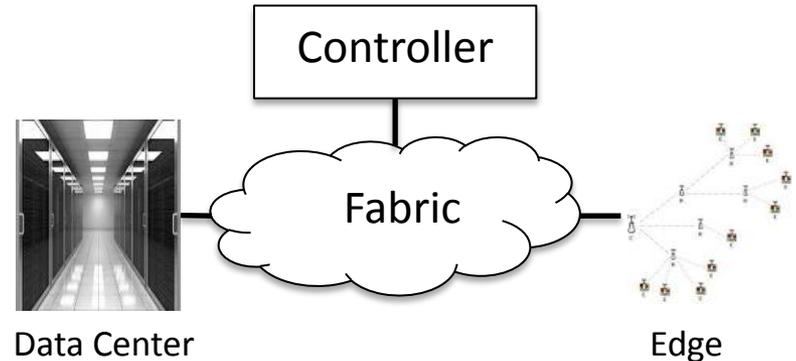
## 2. Internet Exchange Point (IXP)



## 3. Data-Center-to-Data-Center



## 4. Edge (e.g. Campus, End-User/Mobile, Internet-of-Things/IoT)

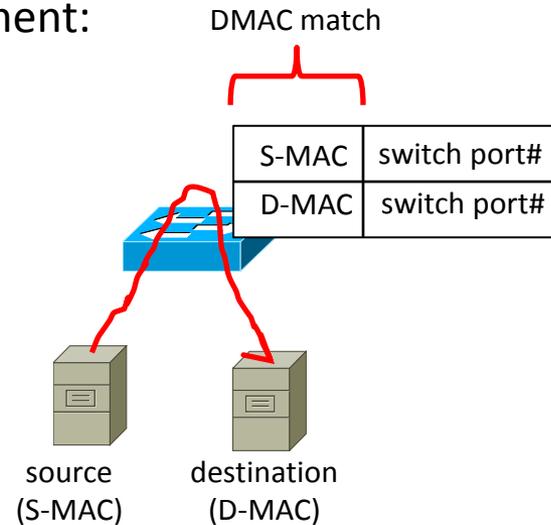


# How does forwarding work when there's more than one egress port?

## 1. Forwarding on a single Ethernet segment:

*Destination Address lookup produces one and only one egress switch port*

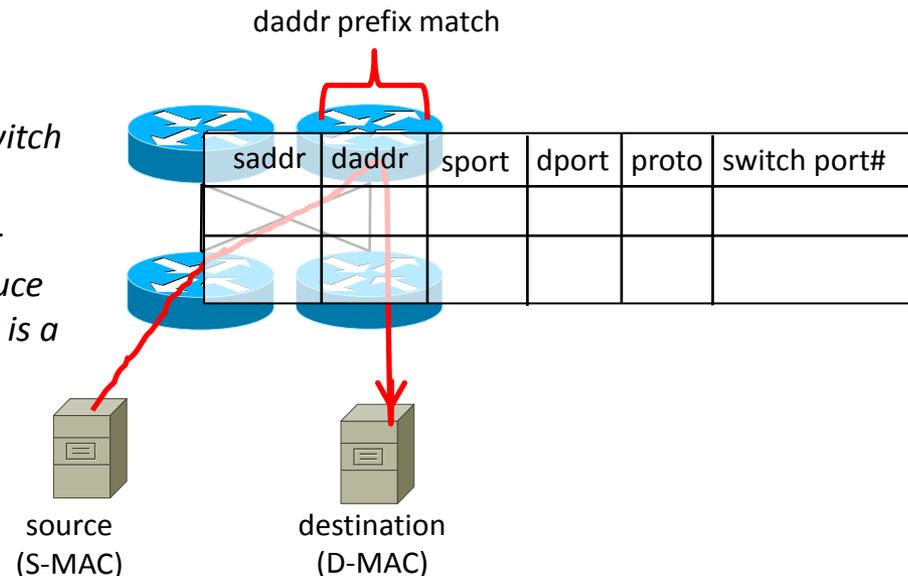
*Note: Match-Action rules that produce one and only one egress switch port is a form of L2 forwarding.*



## 2. L3 Forwarding

*Destination Address lookup may produce more than one egress switch port; requiring path selection*

*Note: Match-Action rules with (or without) wildcards that can produce more than one egress switch port is a form of L3 Forwarding.*



Section #6

# **WHAT ARE THE OPPORTUNITIES FOR E-SCIENCE?**

# Transferring Data Over the WAN

Sharing data amongst geographically distributed collaborators is very common.

Still lacks functionality at the edge:

- Edge/end-points within a site are quite capable
  - Many sites have multi-node, distributed/parallel file systems
  - Computational clusters with access to the file system
  - Nodes in these clusters have limited access to WAN
- Multi-tenancy
  - At a single site there are often multiple transfers occurring simultaneously.
  - There is little to no mediation amongst those sharing the resources
- Scale-Out Parallelism
  - Can we schedule transfers between sites using horizontal scaling techniques?
  - Can we apply cluster resource management techniques to this workload?

# The Network is in Good Shape?

“On-Demand Secure Circuits and Advance Reservation System (OSCARS)” as the example.

1. A node submits a request for a “container” to be copied or moved between two other nodes, with both source and destination distinct from the requesting node.
  - requests a virtual circuit between source node and destination node
  - provisions a sender node at the source and a receiver node at the destination, both sender and receiver are configured for I/O Forwarding
2. Sender node MMAPs the shared dataset
3. Receiver node MMAPs the shared dataset
4. Transfer is initiated

Break a Data Set up into chunks and transfer these in parallel, however, OSCARS only provisions a single circuit

Can we reserve/provision multiple circuits to increase the throughput of this parallel transfer?

# Employing Circuits to Build a Fabric between I/O Forwarders

“Parallel Resource-Optimized Provisioning of End-to-End Requests (PROPER)” as the example.

1. A node submits a request for a “container” to be copied or moved between two other nodes, with both source and destination distinct from the requesting node.
2. Requests multiple virtual circuits between source node and destination node
  - provisions several sender nodes at the source and several receiver nodes at the destination, both senders and receivers are configured for I/O Forwarding
  - sender nodes each MMAP a non-overlapping partition of a shared dataset
  - receiver nodes each MMAP a non-overlapping partition of a shared dataset
  - transfer is initiated

Wrapping it up

# **CONCLUSION**

# References

- (1) Cohen, “Cloud/IaaS Platforms,” 2013  
[https://www.usenix.org/sites/default/files/conference/protected-files/cohen\\_lisa13\\_slides.pdf](https://www.usenix.org/sites/default/files/conference/protected-files/cohen_lisa13_slides.pdf)
- (2) Bechtolsheim, “Moore’s Law and Networking,” 2012  
<http://www.nanog.org/meetings/nanog55/presentations/Monday/Bechtolsheim.pdf>
- (3) Samsung V-NAND - World’s First 3D Vertical NAND Flash Memory  
<http://www.samsung.com/global/business/semiconductor/html/product/flash-solution/vnand/overview.html>
- (4) Liu et al, “On the Role of Burst Buffers in Leadership-Class Storage Systems,” 2012  
<http://www.mcs.anl.gov/papers/P2070-0312.pdf>
- (5) Asanović et al, “FireBox: A Hardware Building Block for the 2020 WSC,” 2014 (see slide 18 - Next Step: New possibilities)  
[https://www.usenix.org/sites/default/files/conference/protected-files/fast14\\_asanovic.pdf](https://www.usenix.org/sites/default/files/conference/protected-files/fast14_asanovic.pdf)
- (6) Rumble et al, “Log-structured Memory for DRAM-based Storage,” 2014  
[https://www.usenix.org/system/files/conference/fast14/fast14-paper\\_rumble.pdf](https://www.usenix.org/system/files/conference/fast14/fast14-paper_rumble.pdf)
- (7) Cully et al, “Strata: High-Performance Scalable Storage on Virtualized Non-volatile Memory,” 2014  
[https://www.usenix.org/system/files/conference/fast14/fast14-paper\\_cully.pdf](https://www.usenix.org/system/files/conference/fast14/fast14-paper_cully.pdf)
- (8) Jimenez, “The Fast-Forward I/O and Storage Stack,” 2013  
<https://users.soe.ucsc.edu/~ivo//blog/2013/04/07/the-ff-stack/>
- (9) Allen et al, “Software as a Service for Data Scientists,” 2012  
<https://www.globus.org/sites/default/files/saas-for-data-scientists.pdf>
- (10) Kimpe et al, “Integrated In-System Storage Architecture for High Performance Computing,” 2012 (see section 4.2 Distributed Container Access; specifically the reference to support for 3rd party data transfers)  
<http://www.mcs.anl.gov/papers/P2092-0512.pdf>