



New directions in storage, hard-disks with build-in networking

Presenter: Patrick Fuhrmann

dCache.org

Patrick Fuhrmann, Paul Millar, Tigran Mkrtchyan

DESY

Yves Kemp

HGST

Christopher Squires



Objectives



- Current storage systems are composed of sophisticated building blocks: large file servers, often equipped with special RAID controllers
- Are there options to build storage systems based on small, independent, low-care (and low-cost) components?
- Are there options for a massive scale-out using different technologies?
- Are there options for moving away from specialised hardware components to software components running on standard hardware?
- What changes would be needed to software and operational aspects, and how would the composition of TCO change?

Building storage systems with small, independent data nodes



- The actual storage system must be able to handle independent data nodes.
 - Like CEPH and dCache
- It must support protocols, allowing massive scale-out (no bottlenecks)
 - CEPH Ok (Client side proprietary driver)
 - dCache through NFS 4.1/pNFS, GridFTP, WebDAV
- It must support protection against failures and support data integrity features.
 - CEPH and dCache continue operation if data nodes fail.
 - CEPH and dCache support data integrity checks.

Selection



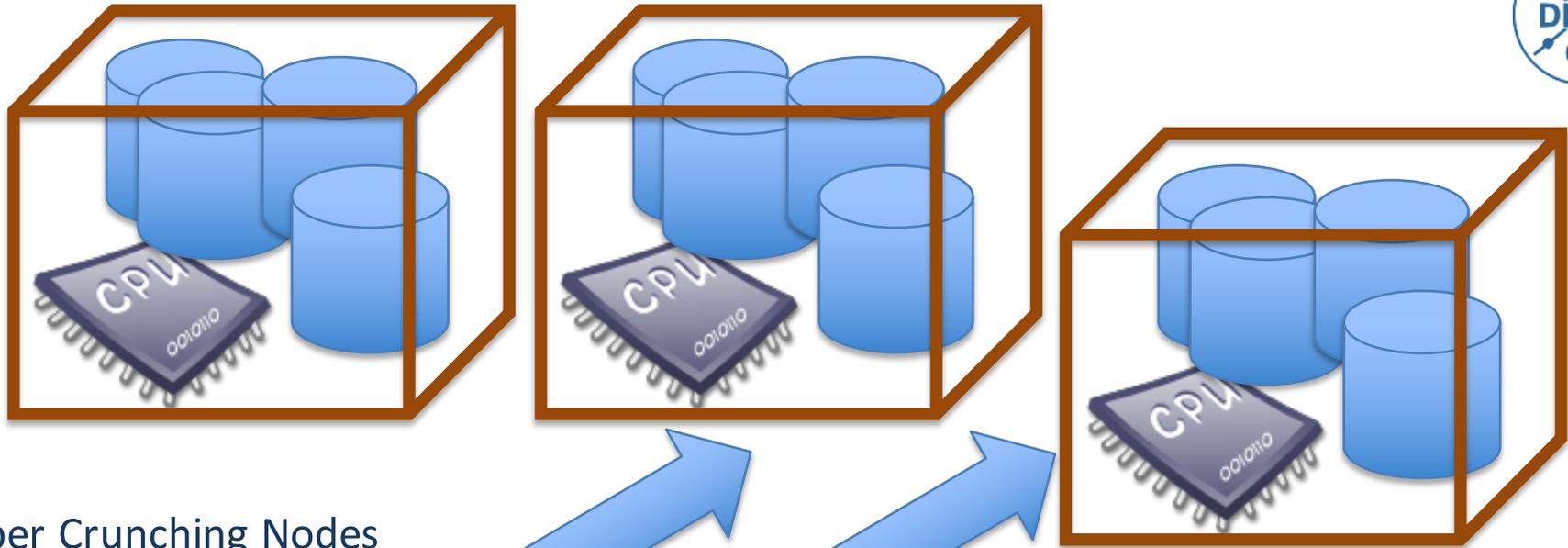
As we are running dCache at DESY, we focused our investigation on dCache, for now, however, we will build a CEPH cluster soon.

CEPH has already been successful ported to a HGST setup.

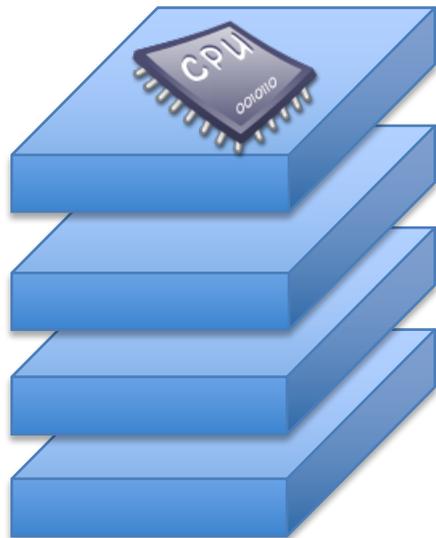
We believe that a merge of CEPH as storage backend and dCache as protocol engine would be exactly the right solution for DESY.

dCache design

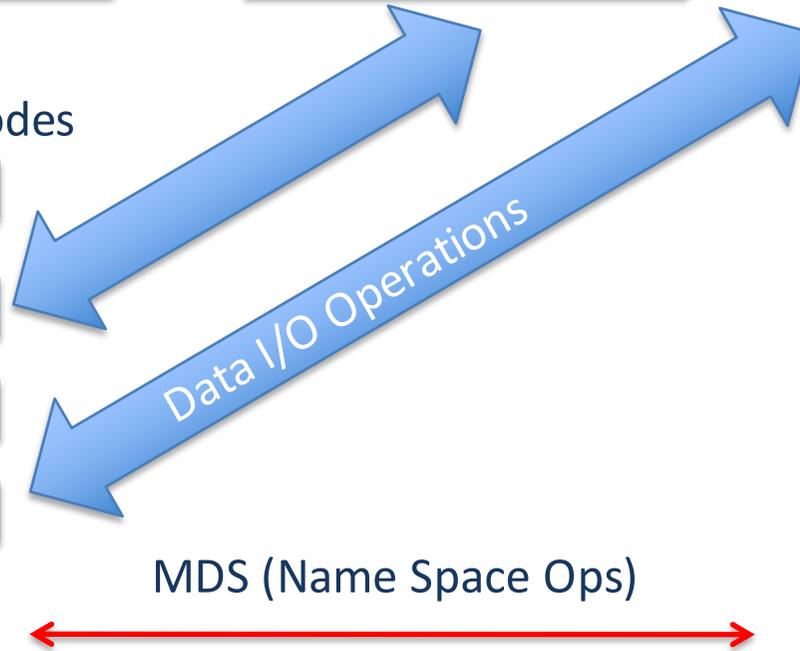
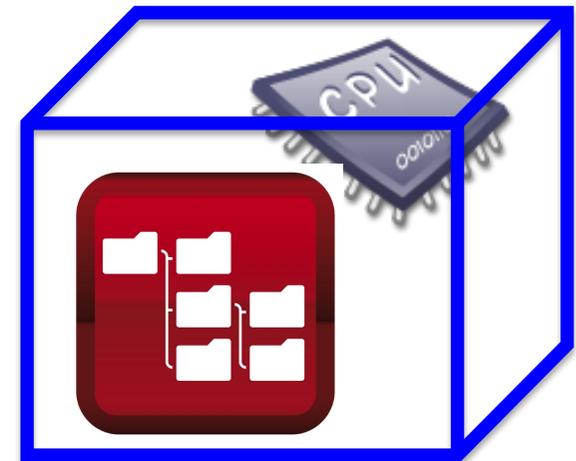
dCache pool (data) nodes



Number Crunching Nodes



dCache Head Node



Available 'small' systems



- DELL C5000 blades, HP moonshot, Panasas, ...
 - Those still share quite some infrastructure
 - Often more than one disk per CPU. Too large for a simple system, but too small for a serious RAID system.
- The extreme setup would be:
 - One disk is one system
 - Small CPU with limited RAM
 - Network connectivity included.
- HGST and Seagate announced such a device
 - Equipped with a standard disk drive, an ARM CPU and a network interface.
 - We have chosen the HGST Open Ethernet drive for our investigations.





- Small ARM CPU with Ethernet piggybacked on regular Disk.

Spec:

- Any Linux (Debian on demo)
- CPU 32-bit ARM, 512 KB Level 2
- 2 GB DRAM DDR-3 Memory
 - 1792 MB available
- Block storage driver as SCSI sda
- Ethernet network driver as eth0
- dCache pool code working!

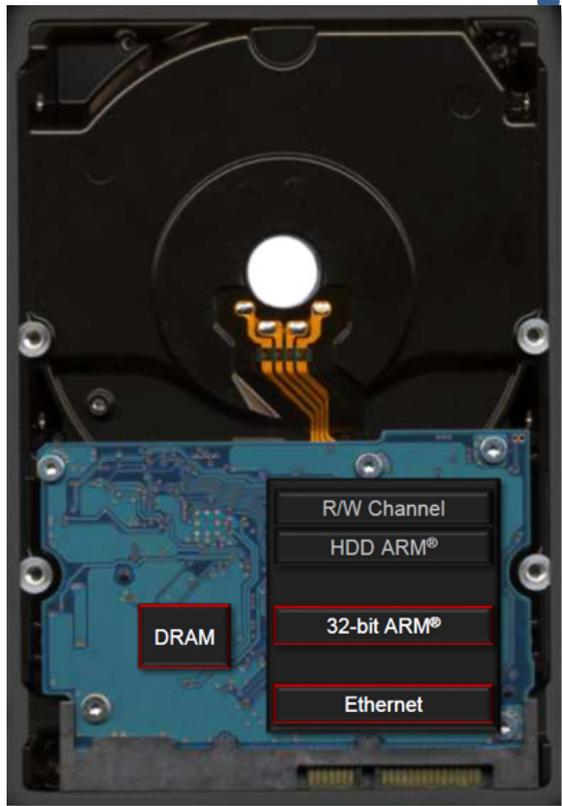
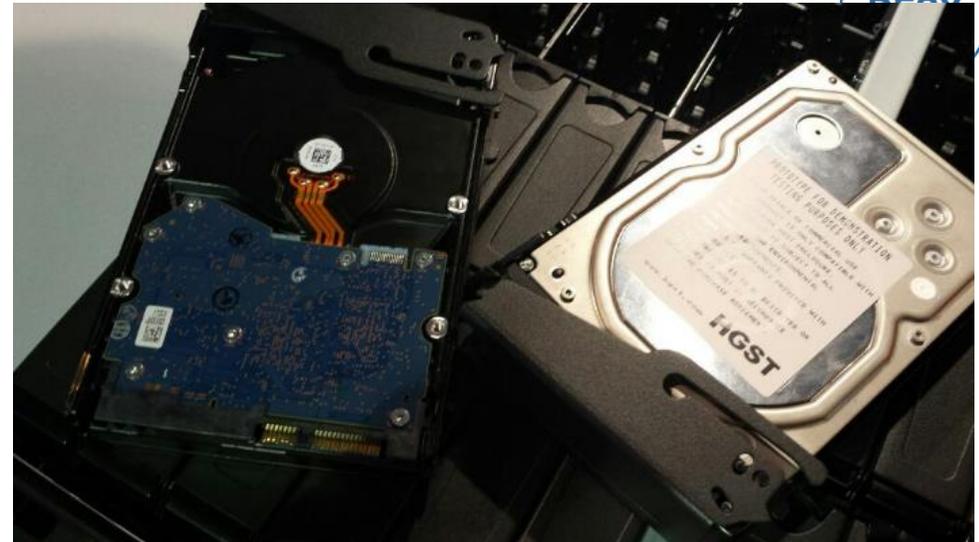


Photo Evidence

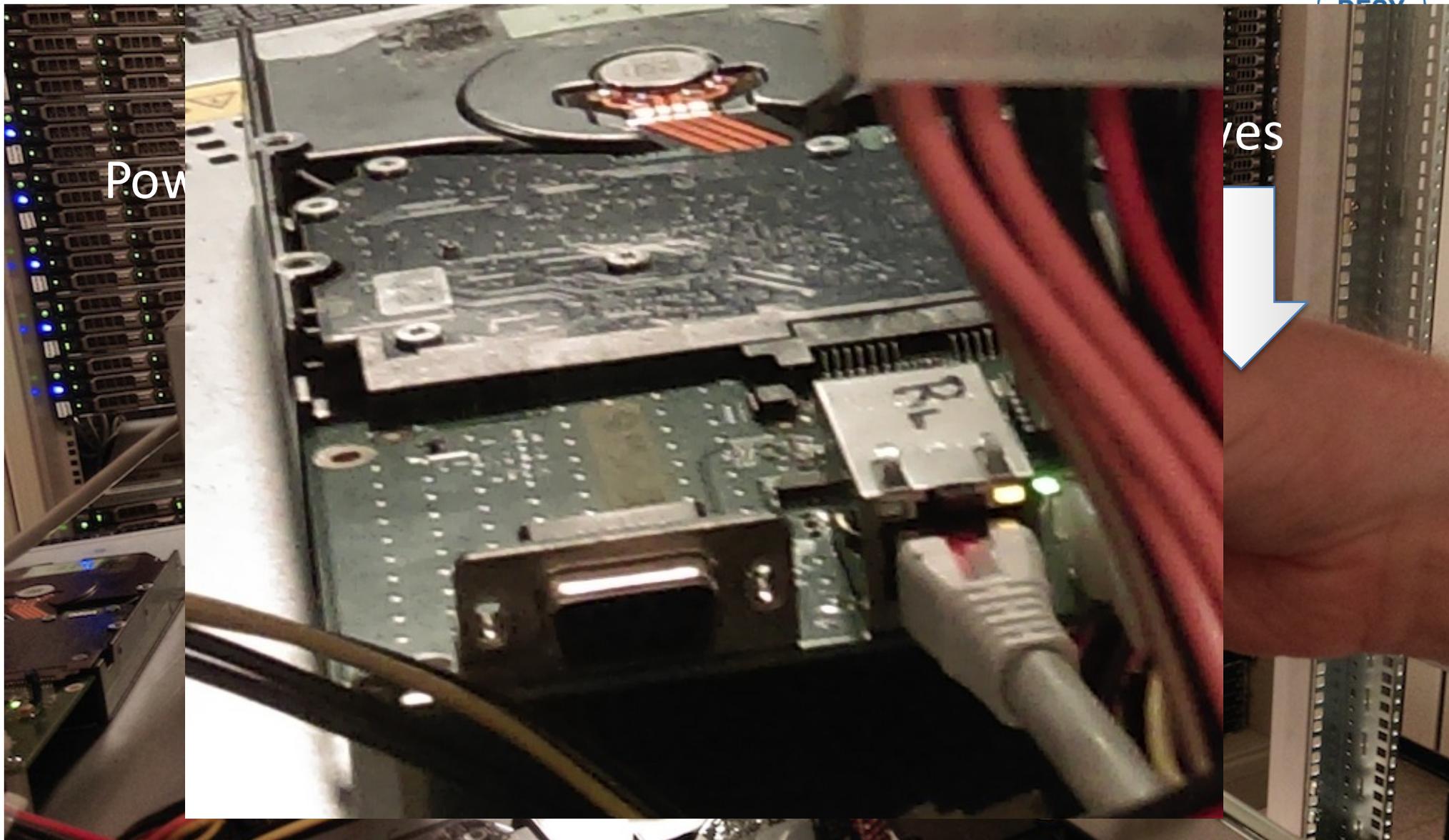
Single Disk



Enclosure



Professional DESY Setup



dCache Instance for CMS at DESY



- dCache instance for the CMS collaboration at DESY
- 199 File Servers in total
 - 172 with locally attached disks
 - 12 disks: RAID-6
 - Varying file size
 - 27 with remote attached (fibre channel or SAS)
 - All file servers are equipped with 24 GB RAM and 1 or 10 GBit Ethernet
- dCache with 5 PBytes net capacity
 - Some dCache partitions are running in resilient mode, holding 2 copies of each file and reducing the total net capacity.
- Four head nodes for controlling and accounting
 - 2 CPU + 32 GByte RAM each
- 10 Protocol initiators (doors)
 - Simple CPU 16 GByte RAM

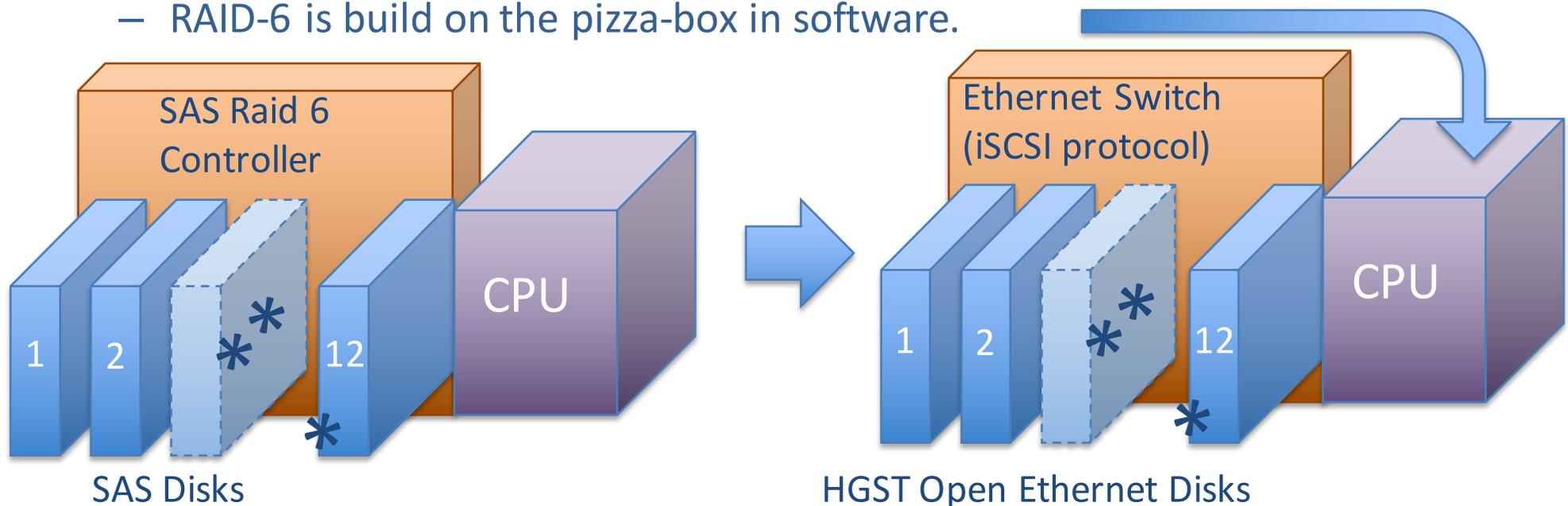
Three main use cases



Potential setup with HGST Open Ethernet drive / iSCSI

The standard 12 disks in RAID-6, attached via SAS to RAID controller card is replaced by :

- One pizza-box with Intel CPU and performant network, connects to 12 HGST drives.
- Each HGST drives acts as one iSCSI target
- RAID-6 is build on the pizza-box in software.



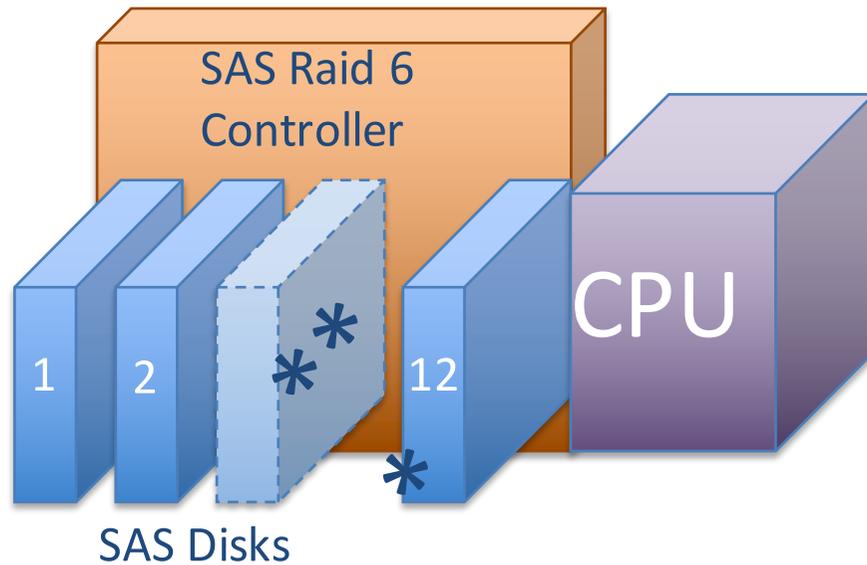
Potential setup with HGST Open Ethernet drive / CEPH



The standard 12 disks in RAID-6, attached via SAS to RAID controller card is replaced with:

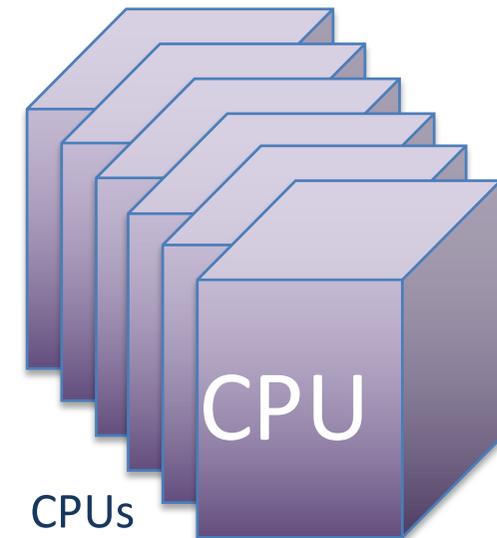
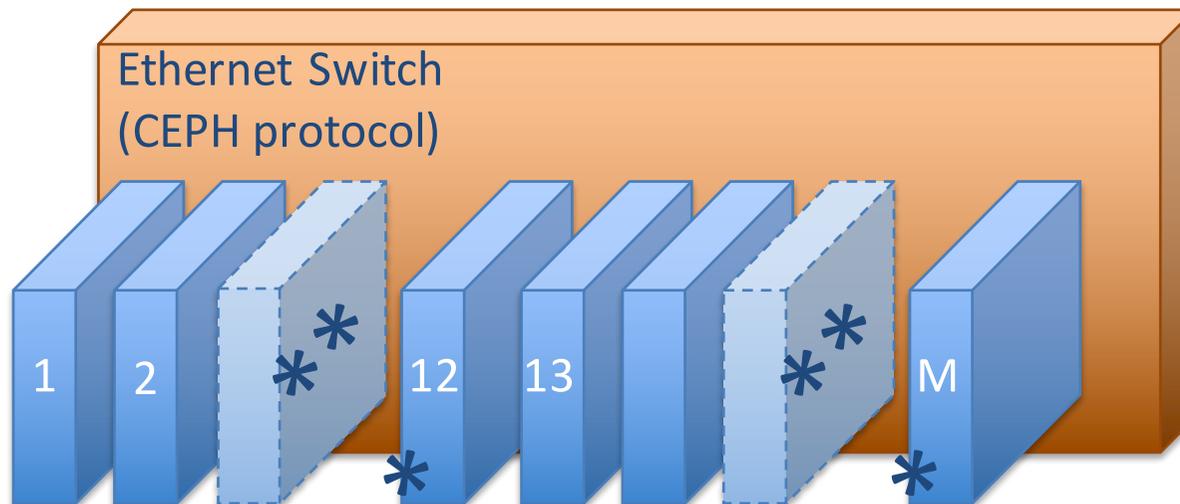
- All HGST drives acting as one large object store
- Pizza Boxes act as CEPH clients can run higher level services, e.g. dCache Storage System.

Potential CEPH deployment



CPU
Running
CEPH plus
high level services
e.g. dCache pool

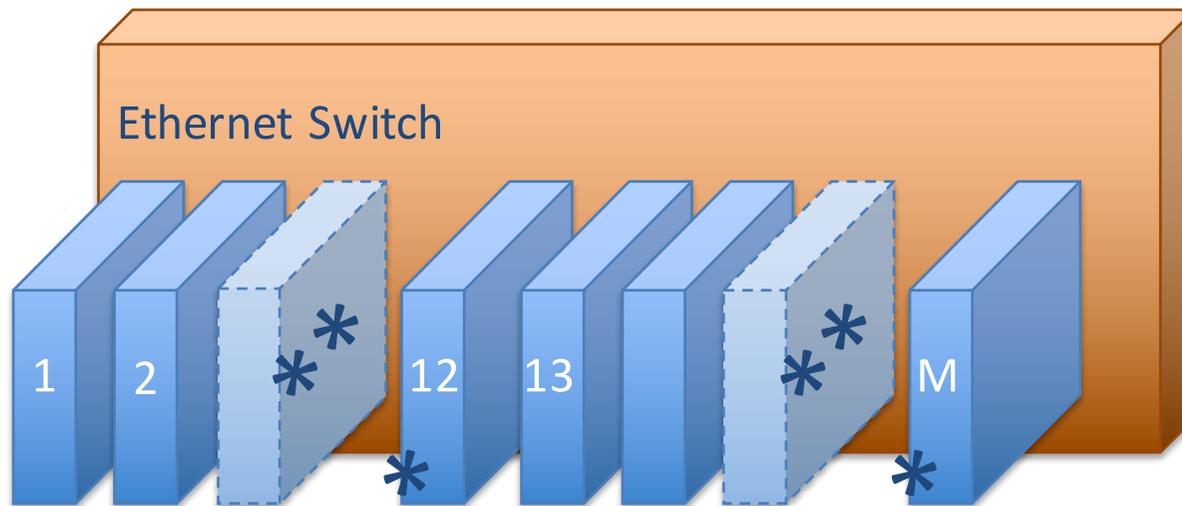
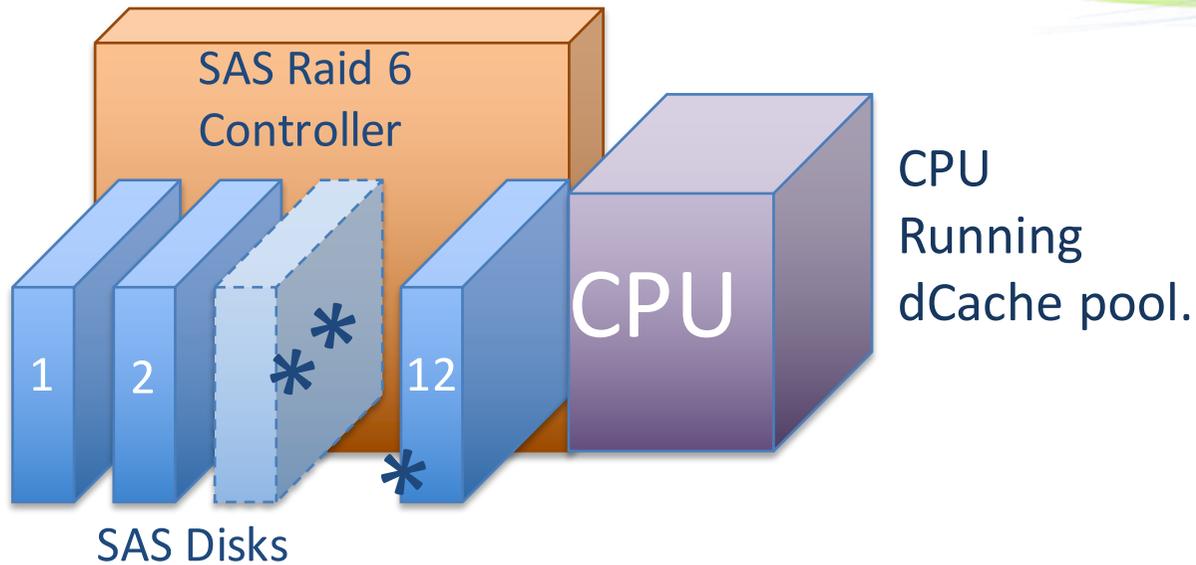
* N



CPUs
Running
High level services e.g.
dCache

HGST Open Ethernet Disks running CEPH

Potential dCache deployment



dCache composed of HGST disks (A quantitative Investigation)

dCache.org



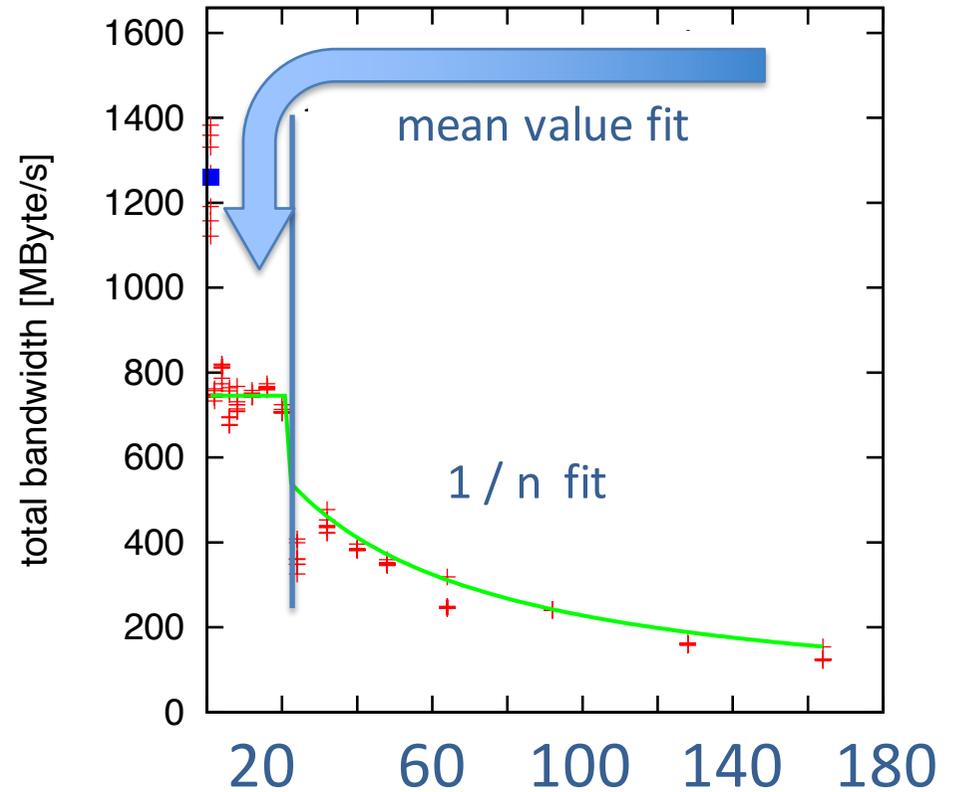
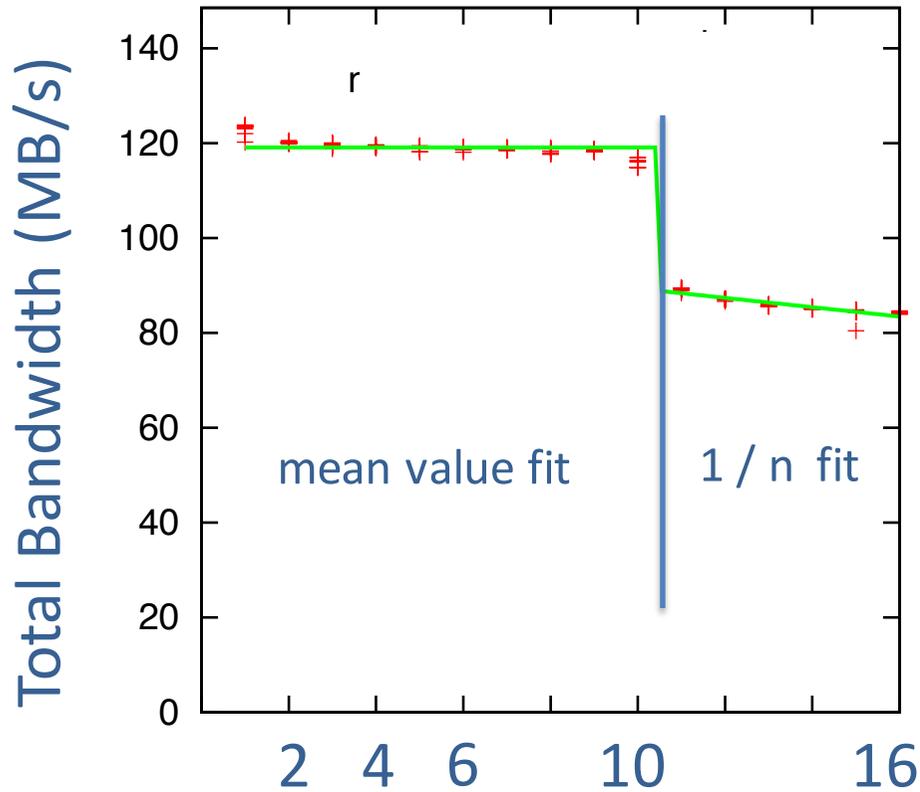
- Qualitative comparison between a setup based on HGST and RAID dCache pool nodes.
- Assumptions:
 - All files identical in size (4 GBytes)
 - Streaming read only
 - Files are equally distributed (flat, random)
- Initial measurement: Total bandwidth versus number of concurrent streams.
 - Using ioZone (stream, read)
 - Applying a fit to the measured data.

Comparison: Ethernet Disk - RAID

Measured

HGST (single drive)

RAID (12 drives)



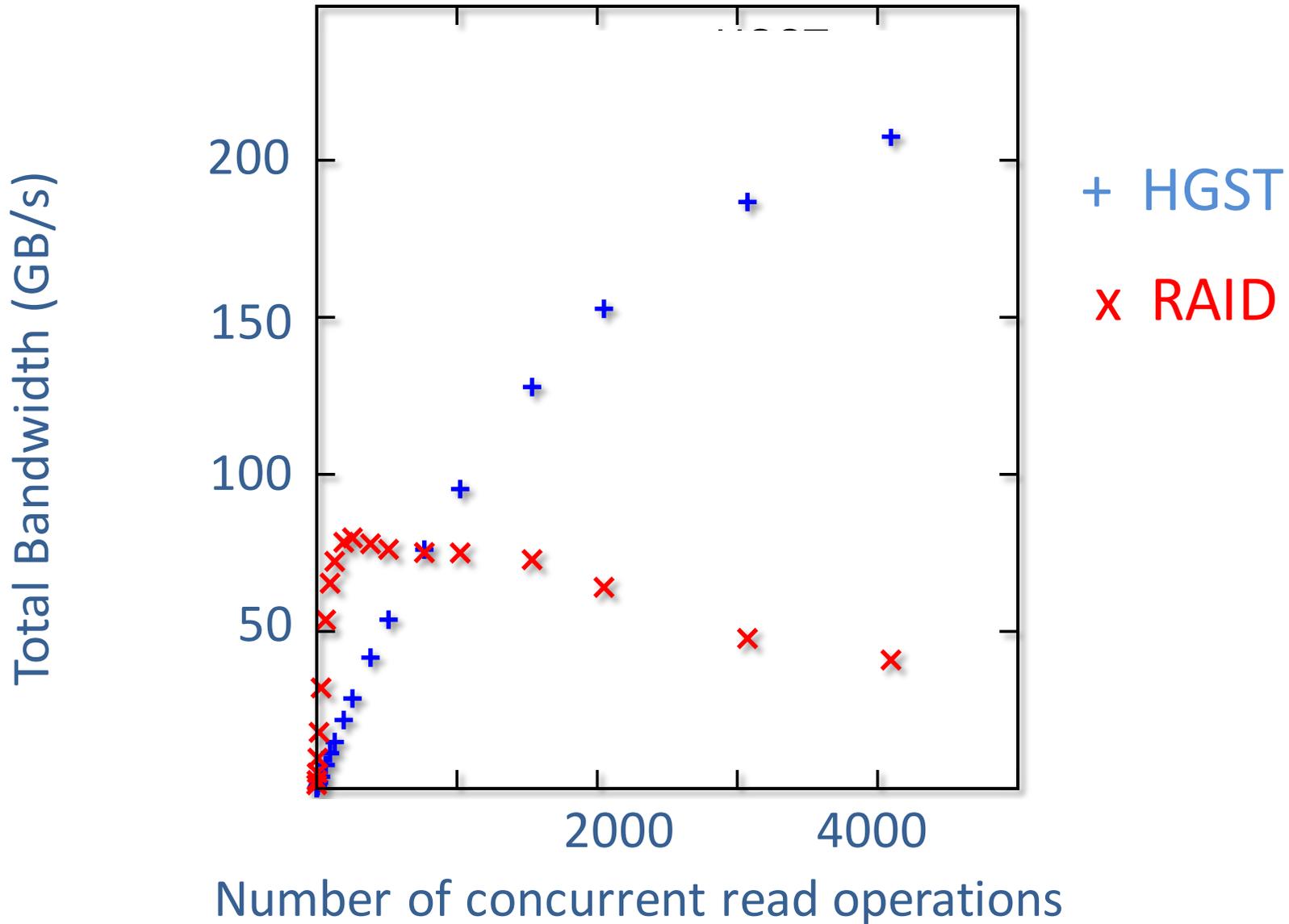
Number of concurrent sequential read operations

Simulation



- Simulating and comparing two 4 PByte dCache instances
 - A) 100 RAID servers (12 disks a 4 TB), **single file copy**.
 - B) 2000 HGST disks a 4 TB, **two file copies**
 - dCache only supports replication of entire files (no erasure code).
- Computing total bandwidth for a varying number of streams.
 - IOzone measurements from previous slides are used.
 - Counting streams/servers and looking up bandwidth to be added up.

Simulation



Real Monitor

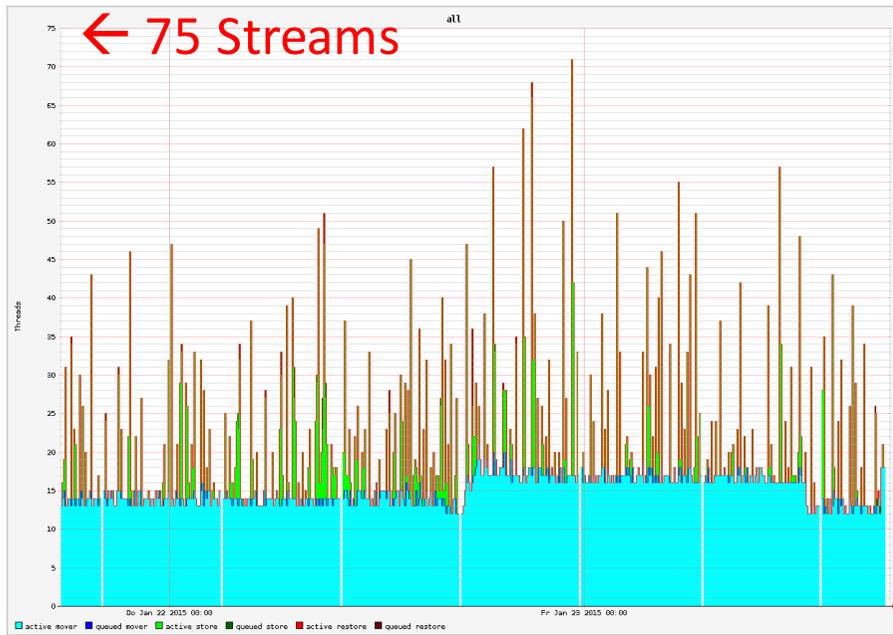


Photon dCache

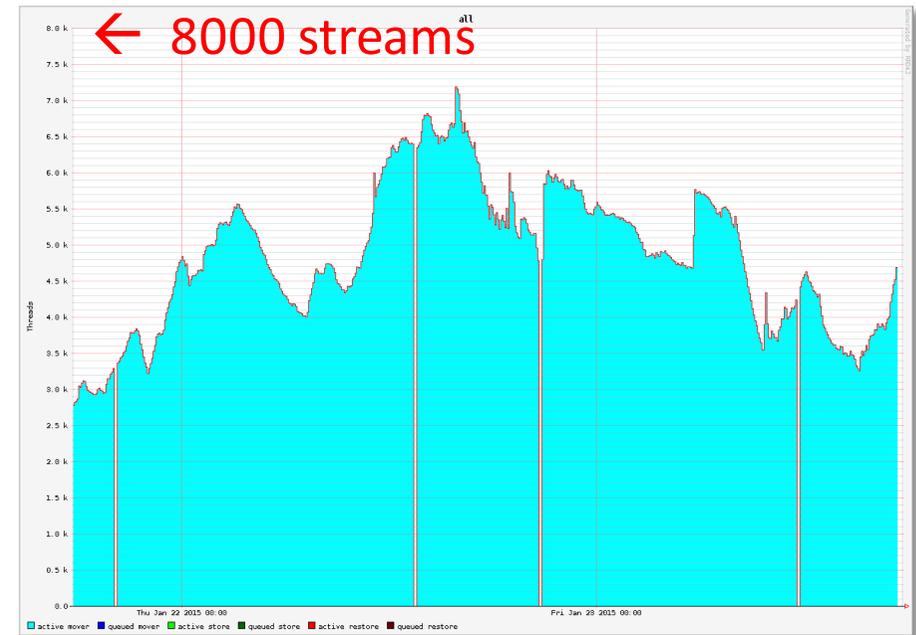
Typically \ll 100 simultaneous reads

CMS dCache

Typically \gg 1000 simultaneous reads



RAID setup would have best performance



HGST setup would have best performance

Total Cost of Ownership

(some considerations)



- At the time being, neither the price nor the delivery units are known.
- Assumption for a HGST product:
 - 4U box with 60 disks and internal switch
 - 4 * 10 GE network
- Reminder: dCache only supports full replica (Erasure codes would be more efficient)

What?	100 RAID	2000 HGST	Comment
Network: Ports	100 Network ports	2000/60=133 ports	Small overhead
Network: IP	100 IP addresses	2000 IP addresses	Public IPv4 NO Private IPv4 & IPv6 OK
Power	1200 disks + 200 Intel CPU + 100 RAID controller	2000 disks + 2000 ARM CPU + 133 switches	Roughly similar power consumption to be expected
Space	100x2U = 200 U (usual DESY setup)	2000/60x4 U = 130 U Potentially two computing rooms	Caveat: Other form factor exists for RAID which are denser
Management	100 Systems, with RAID controller	2000 systems, without RAID controller	Anyway, need a scaling life-cycle management system
Operations	100x12 disks in RAID6, no resilience	2000 disks with resilience	HGST setup: No RAID rebuild times, no need for timely change of defective hardware, just re-establishing resilience

Summary



- Performance:
 - To preserve a high overall throughput with a high number of streams, many single storage units are preferred.
 - If the focus is on the single stream performance of a medium or small number of streams, RAID system would be preferable.
- As an example for small storage units we selected the HGST Open Ethernet drive
- Operations considerations:
 - A dCache setup with two copies as data integrity method is roughly similar in TCO compared to traditional RAID system setup
 - A setup based e.g. on CEPH with more advanced and efficient data integrity methods would clearly shift the TCO towards an advantage for the HGST setup compared to traditional RAID system setup

Thanks

Further reading:

- HGST: www.hgst.com
- dCache: www.dCache.org
- DESY: www.DESY.de