



Data management in Run3

Predrag Buncic

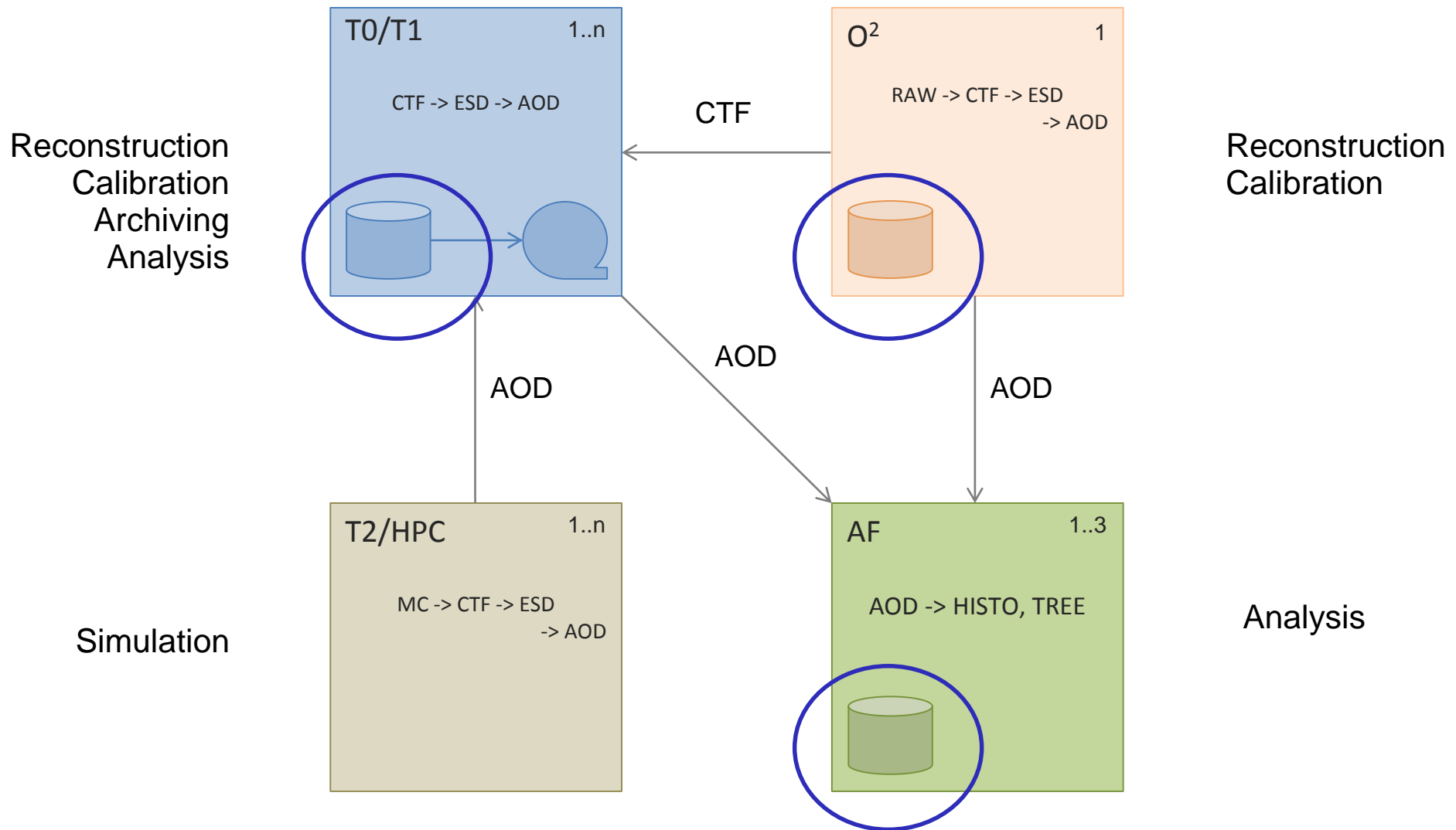
CERN



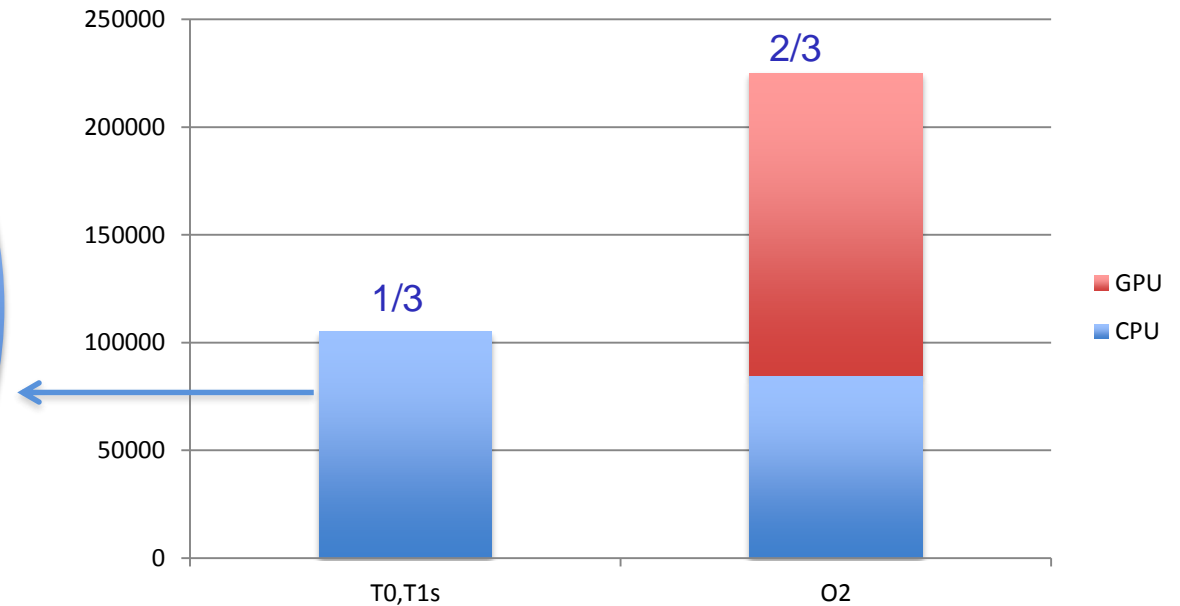
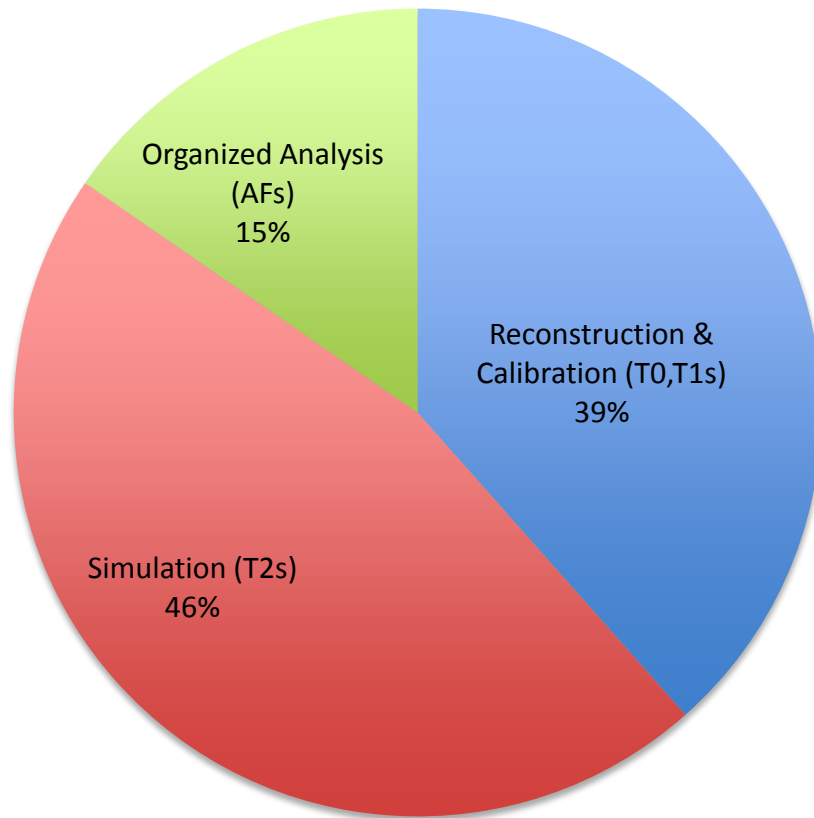
ALICE



Roles of Tiers in Run 3



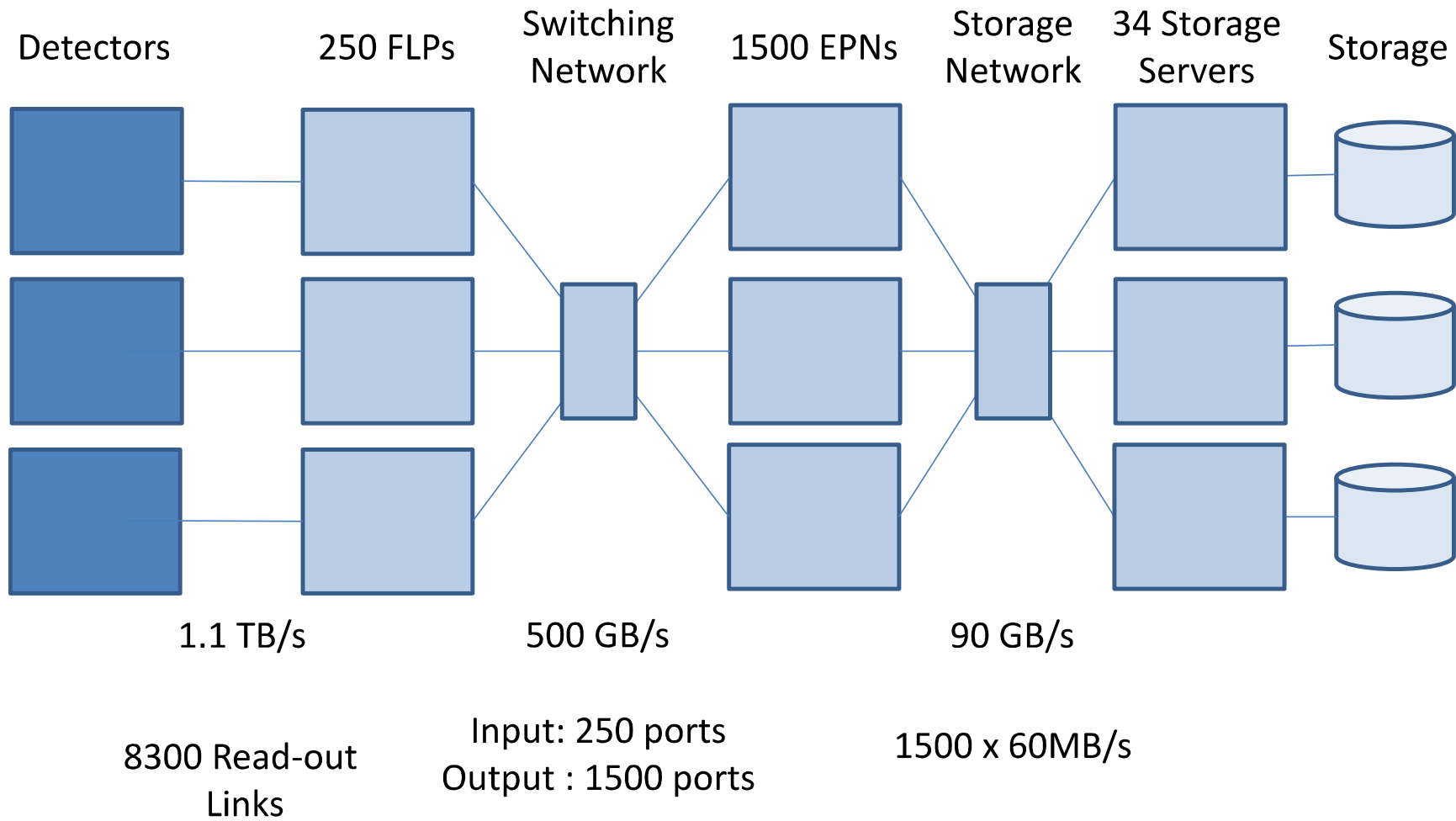
Relative shares of various workflows in Run 3



- Relative importance of O2 vs Grid may change with time if Grid resources continue to grow as expected

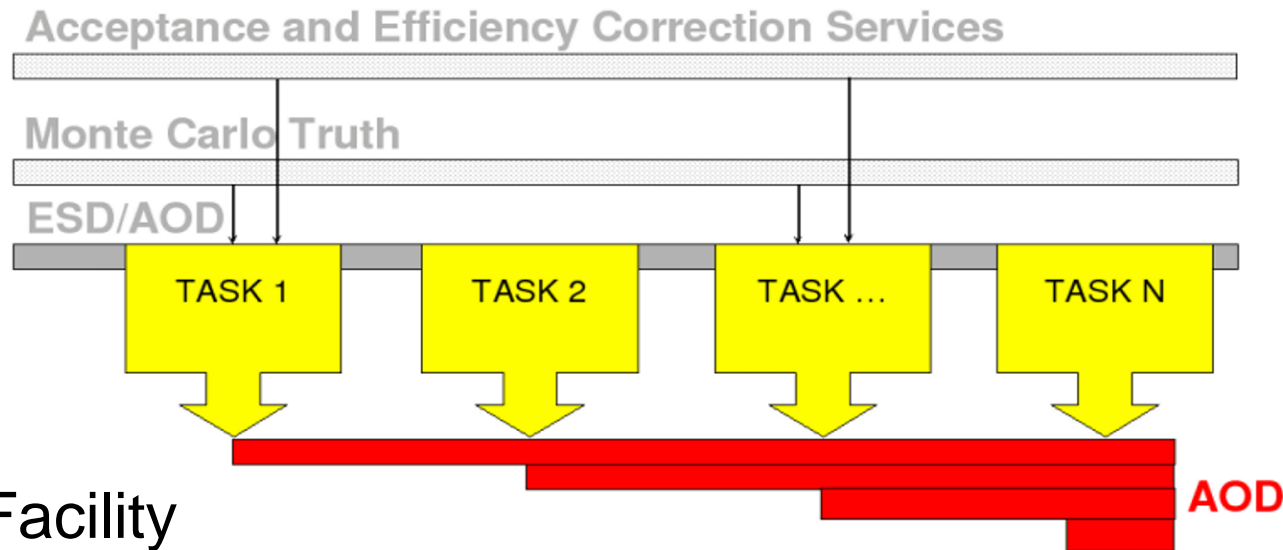


O2 Facility





Analysis Facilities



- Analysis Facility

- Collect AODs on a few dedicated sites (AFs) that are capable of locally processing quickly large data volume
- The AF needs to be able to digest more than 4PB of AODs in a 12 hours period
 - Typically (a fraction of) HPC facility (20-30'000 cores) and 5-10 PB of disk on very performant file system
- Analysis trains need on average 5 MB/s per job slot to be reasonably efficient.
- We require the cluster file system able to serve 20,000 job slots at an aggregate throughput of 100 GB/s.



GPFS is commercial, more mature than the others (been around since late 90's), and has some nice features in terms of HA/failover, distributed metadata, HSM/tiering storage, moving LUNs around, migrating data, filesets (with directory-level quotas), etc.



The GlusterFS architecture aggregates compute, storage, and I/O resources into a global namespace. Each server plus attached commodity storage (configured as direct-attached storage, JBOD, or using a storage area network) is considered to be a node. Capacity is scaled by adding additional nodes or adding additional storage to each node. Performance is increased by deploying storage among more nodes.



Lustre is probably the next mature option, it's free, it's been picked up by Intel and they're serious about improving it. Together with GPFS, it is used in many flagship HPC installations.

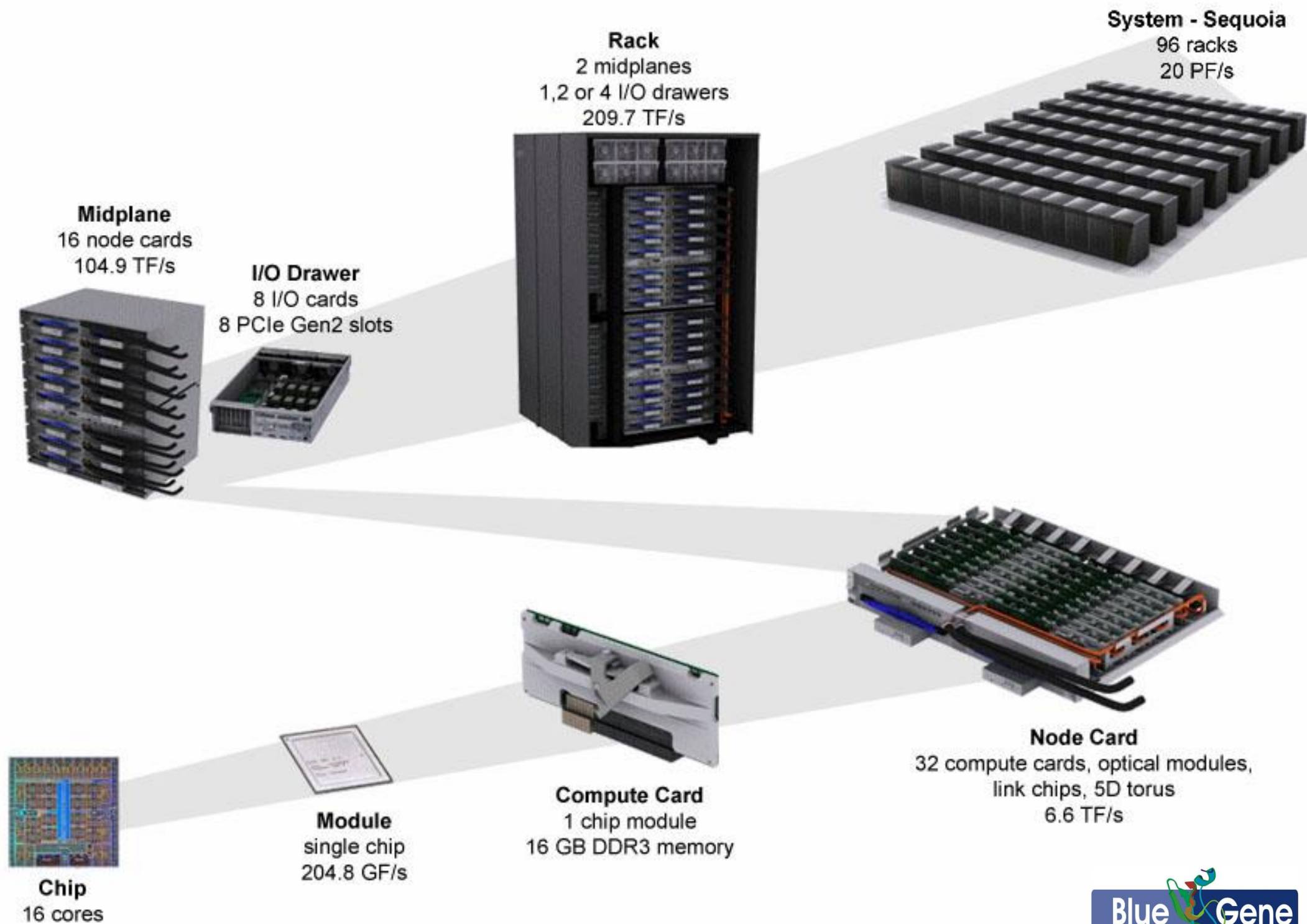


Ceph is an object storage based free software storage platform that stores data on a single distributed computer cluster, and provides interfaces for object-, block- and file-level storage

Think BIG



- Sequoia is a 20 Petaflop IBM BG/Q system sited at the Lawrence Livermore National Laboratory in Livermore, CA
 - 98,304 nodes with 16 cores/node; 1,572,864 total cores, 64-bit, IBM PowerPC A2 processor, 1.6 petabytes of memory; 16 GB/node, 96 refrigerator sized racks, 7.9 MWatts total power

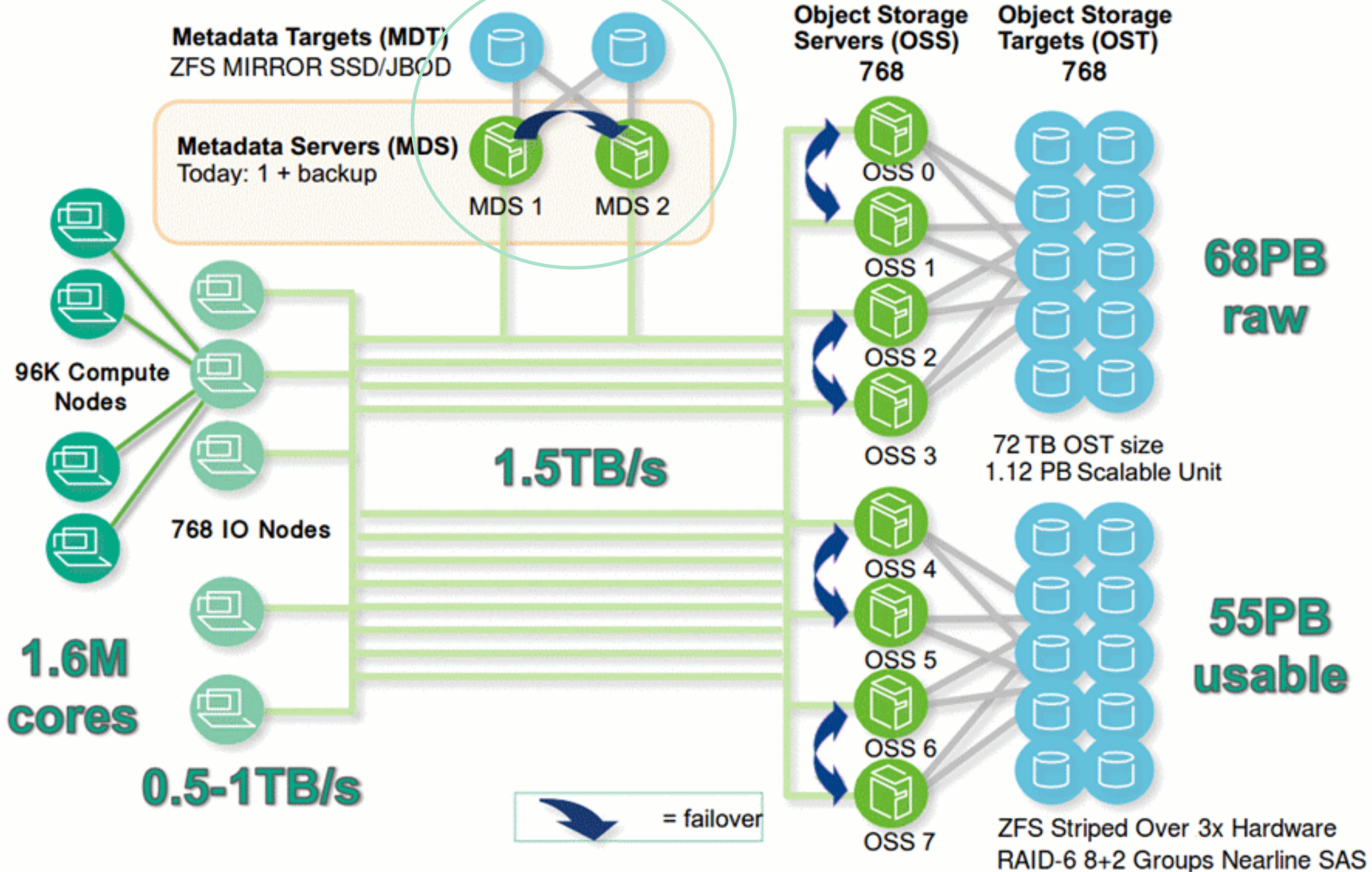


GPFS vs Lustre (apples-to-apples, on the same hardware)

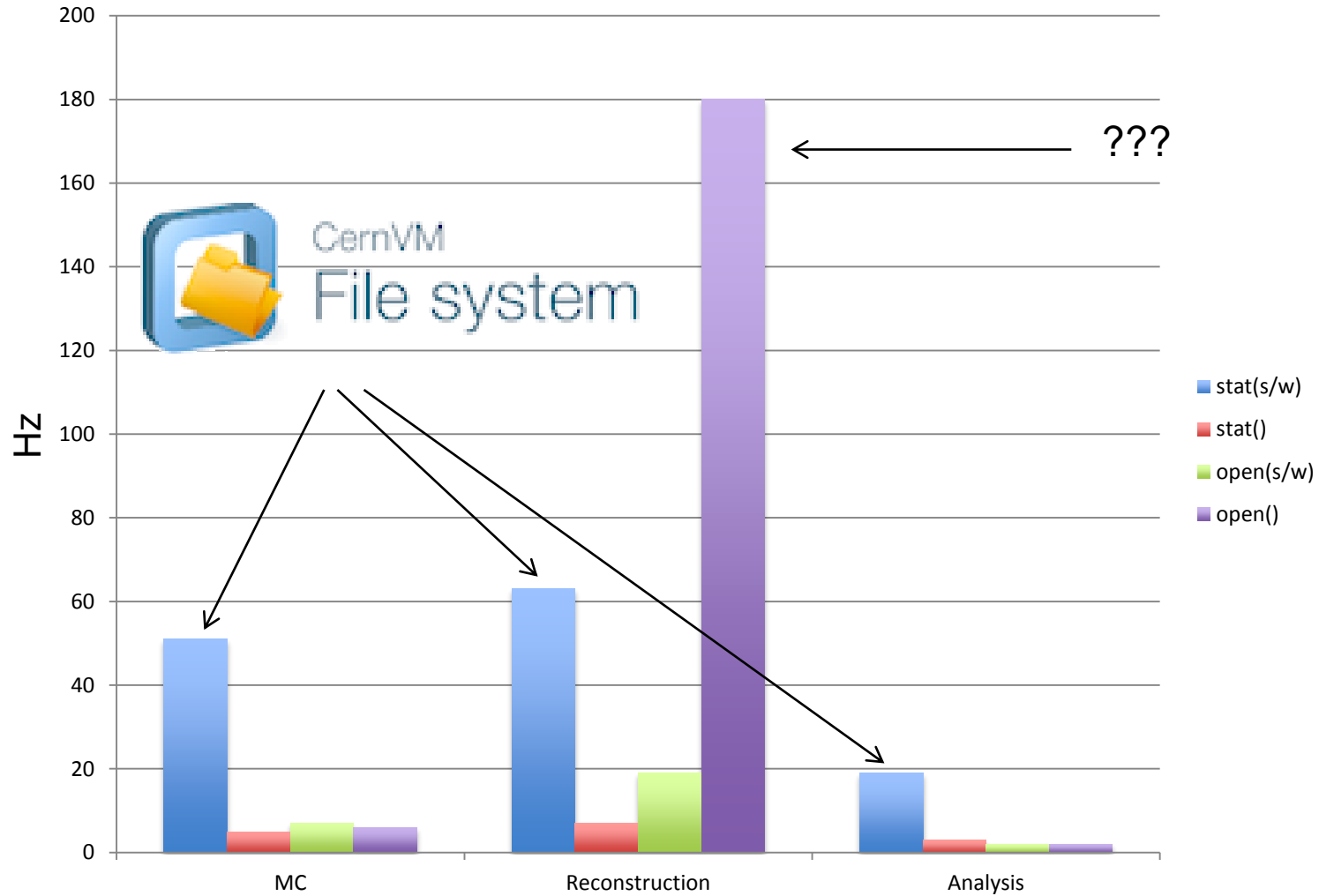
- Planning the environment for Sequoia, a 20 PetaFlop/s IBM system with an I/O target of **512 GB/s**, and a stretch goal of 1TB/s.
 - Both file systems are able to drive hardware at high rates.
 - GPFS has the advantage for throughput tests due to larger blocksize, and since Lustre has the additional overhead of internal checksumming.
 - Lustre is generally significantly faster for the metadata operations that are most important in our workload: operations where data cannot be locally cached on the client.
 - GPFS is better able to spread data over multiple servers and to use multiple cores in the client.
 - Where metadata can be cached on the client, GPFS will have an advantage.
 - For stating a large number of files in a shared directory GPFS can beat Lustre, with the advantage increasing as the number of files increase.
 - File creations in a shared directory have been special optimized algorithm for GPFS with “fine grained directory locks”

Source: <http://www.pdsw.org/pdsw10/resources/posters/parallelNASFSs.pdf>

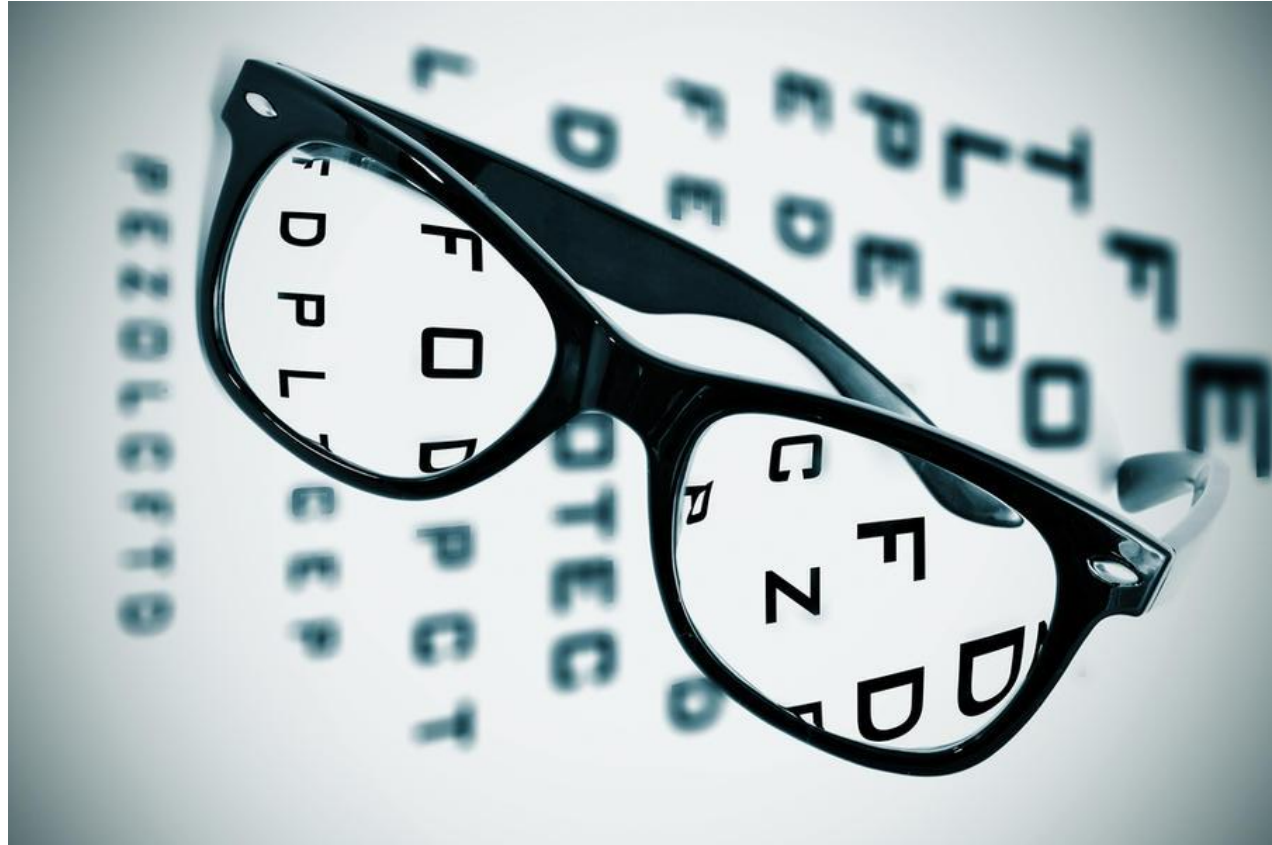
LLNL Sequoia Lustre Architecture



Frequency of stat() and open() calls per job type



C. Grigoras



Look closer



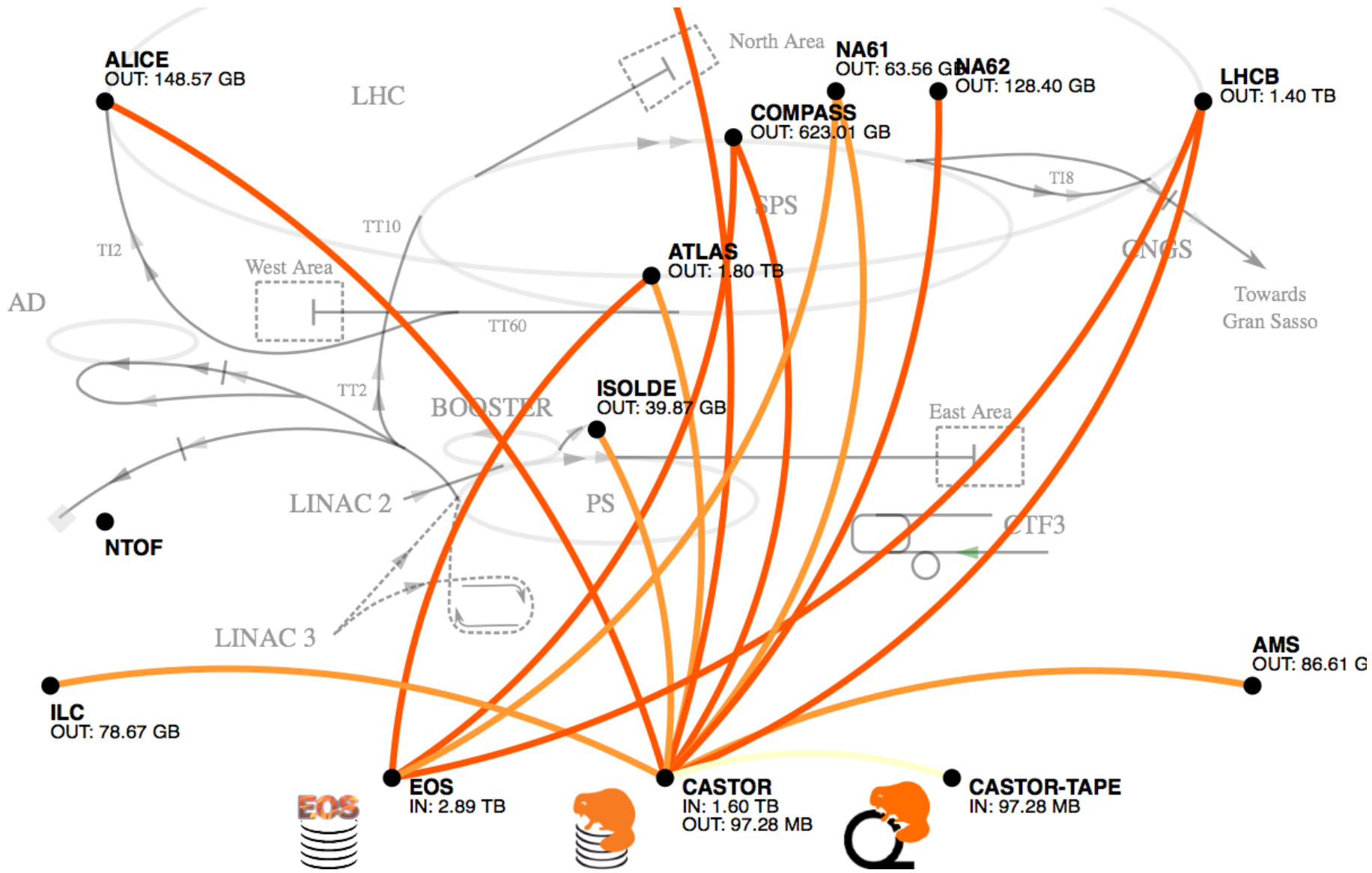
Data TV

Last Min

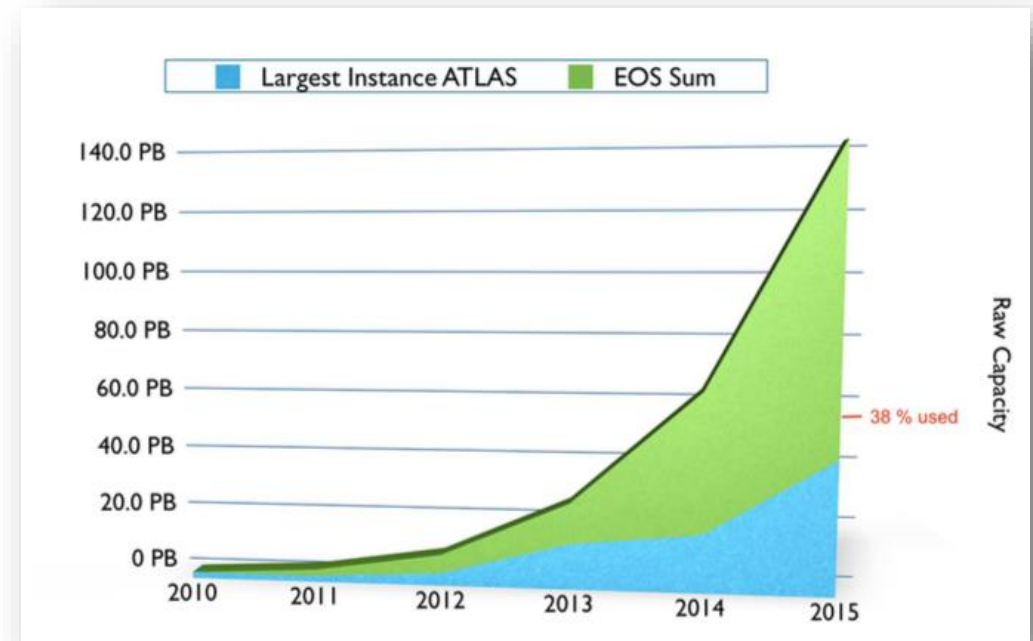
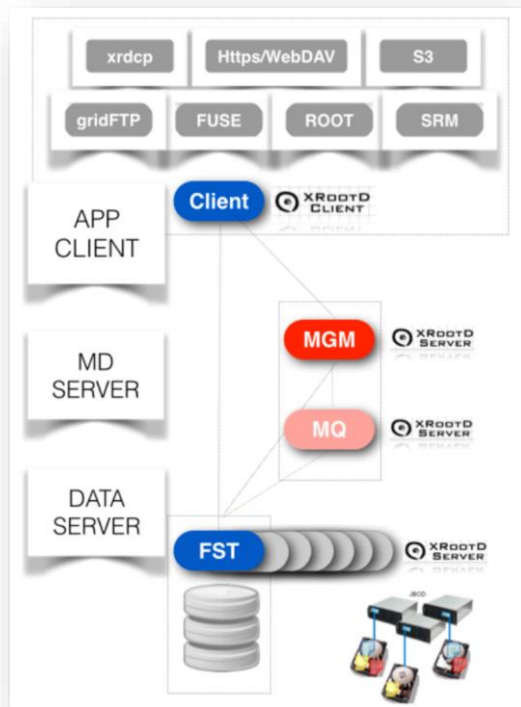
Last 24h

Last Week

Run2



EOS



A. Peters, Journal of Physics: Conference Series **664** (2015) 042042

EOS provides a highly-scalable hierarchical namespace implementation. Data access is provided by the XROOT protocol. The main target is the physics data analysis use case often cases characterized by many concurrent users, a significant fraction random data access and a large file open rate. For user authentication EOS supports Kerberos (for local access) and X.509 certificates for grid access.

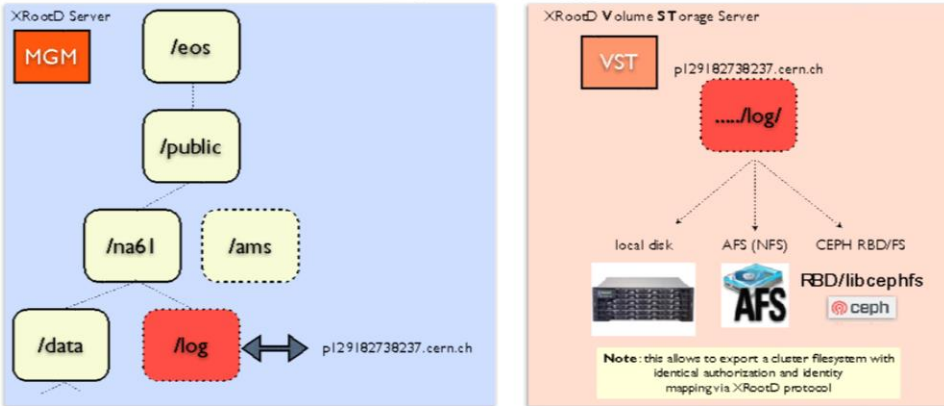
<http://eos.web.cern.ch/>

CITRINE VST



Infinity ∞ • EOS Infinity

- AFS-like attached volumes hosting data+meta data of a subtree
- small/many file use cases
- allows to attach any mountable FS tree into EOS namespace
- allows to have extended attributes on file and directory level for meta data tagging



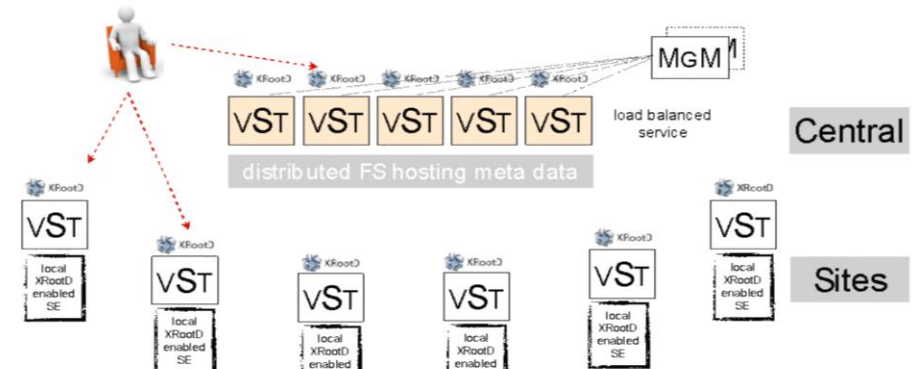
Wednesday, November 8, 12

CITRINE VST



Unity • EOS Unity

Today's Federations provide a redundant functionality via a read-only overlay network. A complete storage federation should have also placement capabilities, honor replication policies and a global reliable namespace. We can use a group of VSTs to host the global logical namespace redirecting read and write requests to VSTs hosting a logical or physical namespace (sites). A site VST is just a redirection and report gateway to any regular XRootD enabled SE or a local EOS setup. For placement and file access we can extend the already existing geo placement/scheduling capabilities of EOS used for the CERN/Wigner CC setup.

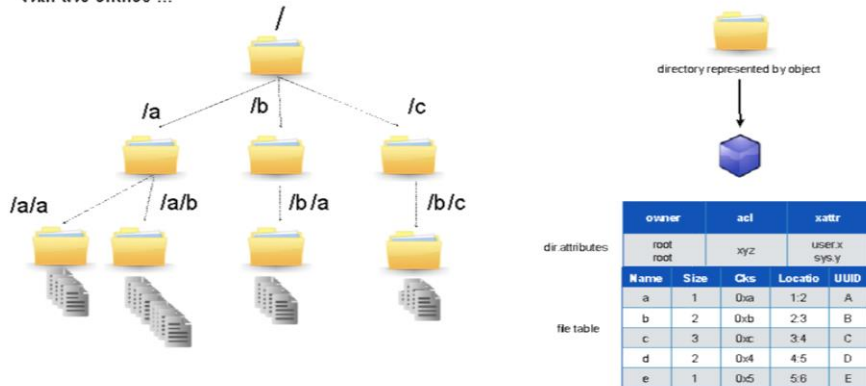


Wednesday, November 8, 12

Diamond R&D



- trivial idea: store a namespace in a scalable object store
 - we can represent data in a *hierarchical structure* using directories and files and we *don't need* to group an infinite amount of files into a single directory
 - each *file* is a *list entry* with meta data in a directory
 - each *directory* is represented as an *object* in an object store
 - to circumvent central locking we can allow a conflict if two files get created with the same name and different contents and make it visible in the namespace like a conflict in DropBox with two entries ...





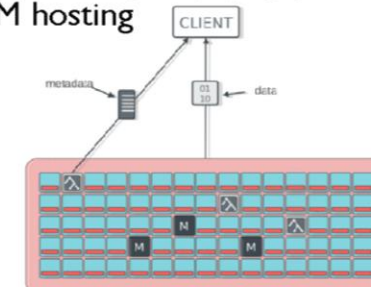
Wednesday, November 8, 12

Diamond R&D



Scalable Object Store/Namespace using CEPH

-  **ceph** is an open source implementation of an object store providing features like *dynamic resizing*, *self-healing*, *guaranteed consistency*, *low read latency*, *async object IO*, *extended attributes + key-value map per object*, *object notifications*
- IT-DSS provides now a  (rados) object store **service** with 1 PB capacity [x3] (~50 nodes) - initially for VM hosting



Wednesday, November 8, 12

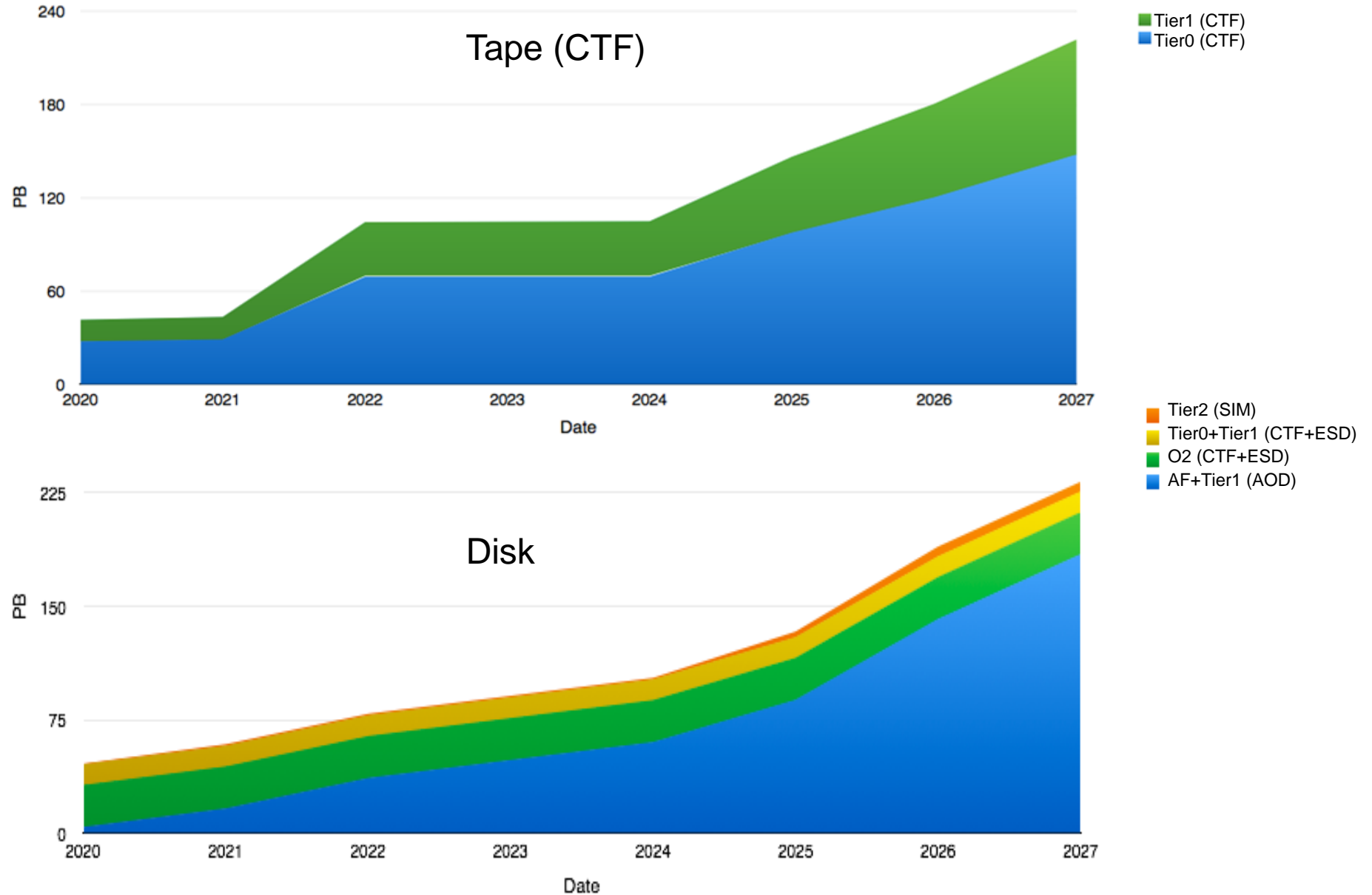
Reducing complexity



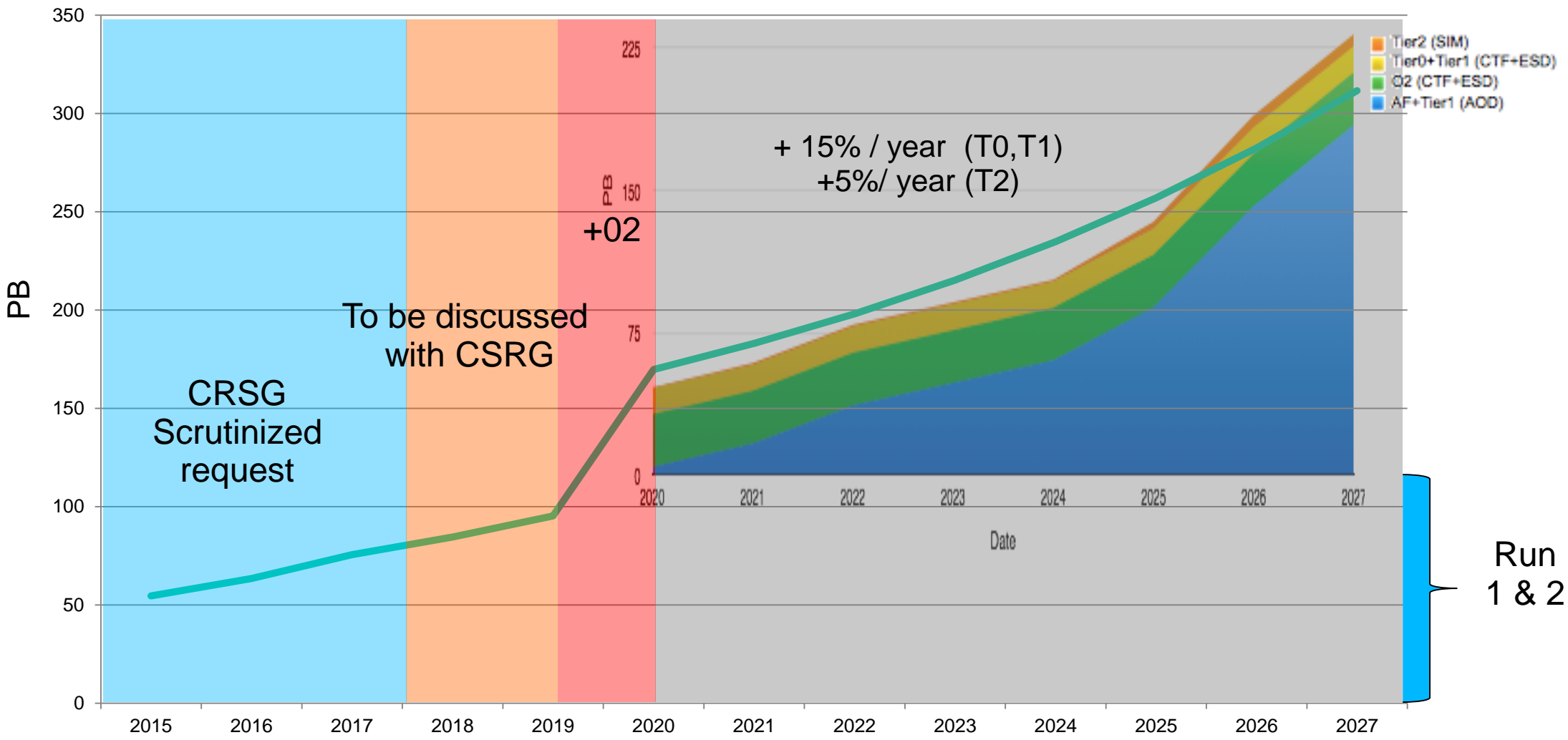
- Virtually joining together the sites based on proximity (latency) and network capacity into Regional Data Clouds
- Each cloud/region has to provide reliable data management and sufficient processing capability
 - Dealing with handful of clouds/regions instead of the individual sites



Expected tape and disk needs



Expected disk space evolution (T0+T1s+T2s+O2)



Conclusions

- Data storage/management problems in Run 3 will be significantly bigger than in Run 1&2 but not on the scale that is beyond of what is commonly deployed today
- HPC file systems may solve part of our problem
- EOS is already tested to the scale required to manage ALICE internal disk buffer@P2
 - Some extras might be needed
 - Media aware caching (SSD, fast disk, shingled disk...)
 - Sophisticated disk pool monitoring, visualization
 - Provides data access via
 - high performance private xroot protocol
 - standard HTTP protocol
 - On the Grid side, it has all we need to manage scalable global name space
- Prototyping and testing is needed to assess performance and TCO