



EOS Roadmap

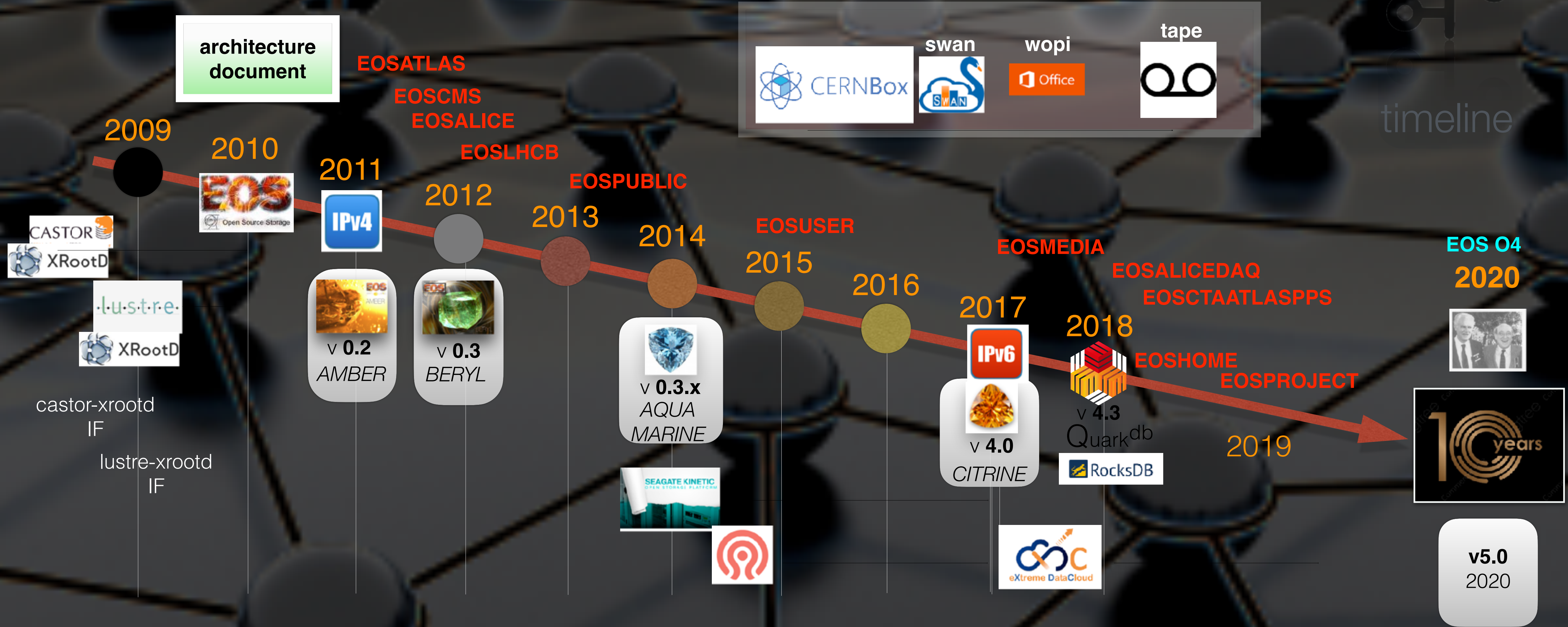
Progress, Direction, Outlook

Dr. **Andreas-Joachim Peters**
CERN IT-ST



Project History

timeline





EOS 2020

10 years





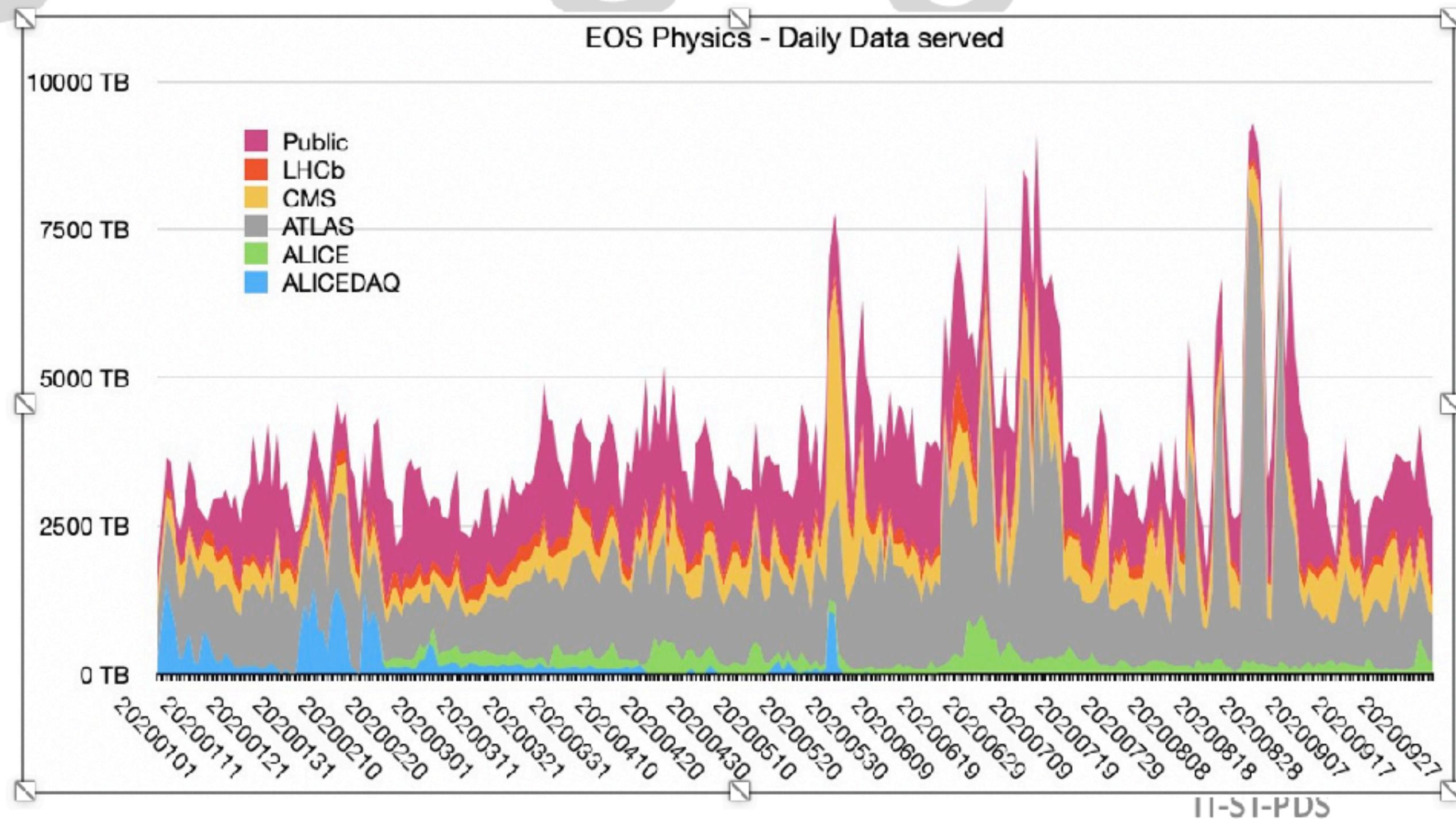
CERN-EOS for WLCG



>10 Collaborations

1081 PB
Data Served
 since 1/2020

6 Instances
284.05 PB
Total Space
+ 50 PB for RUN3



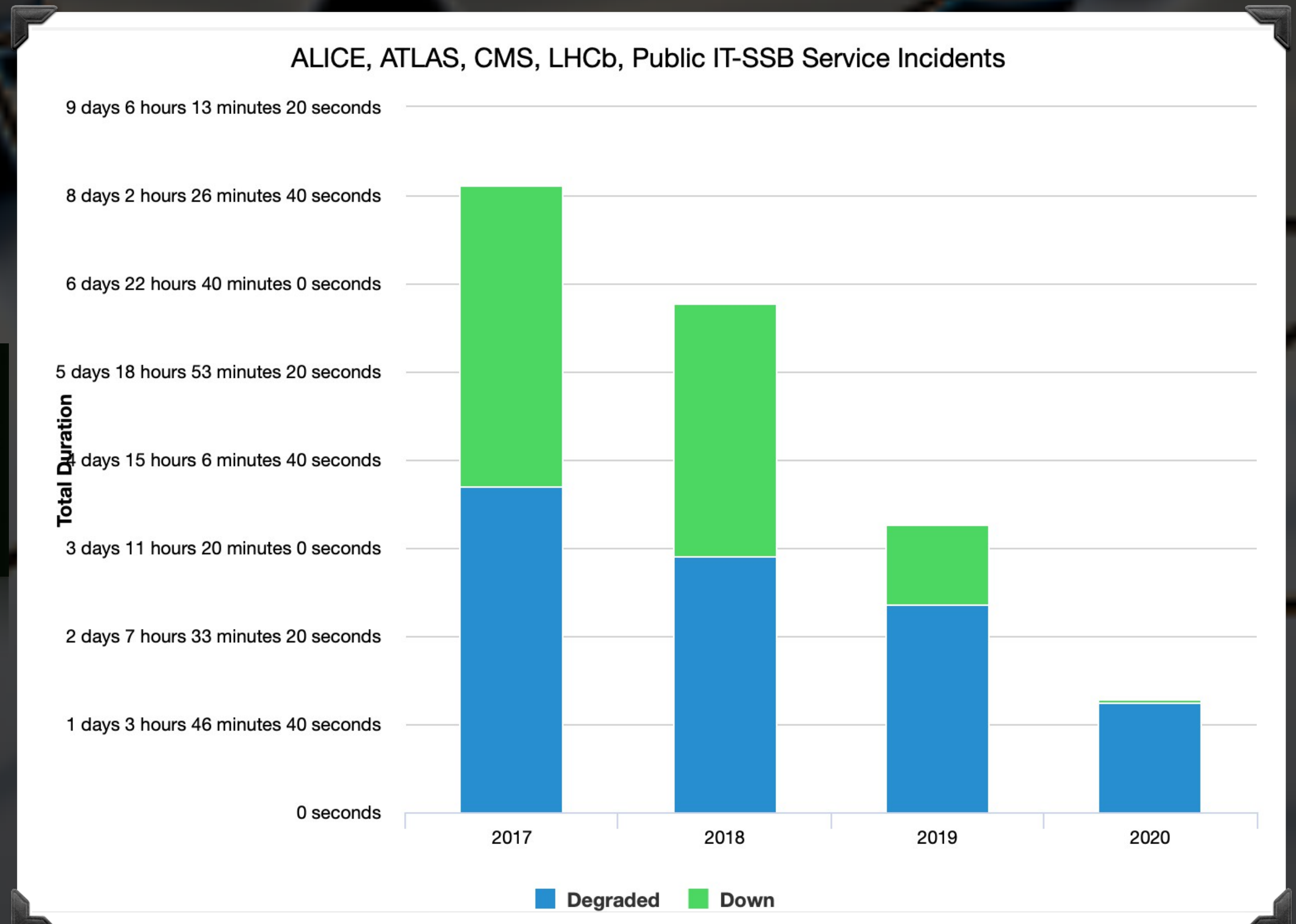
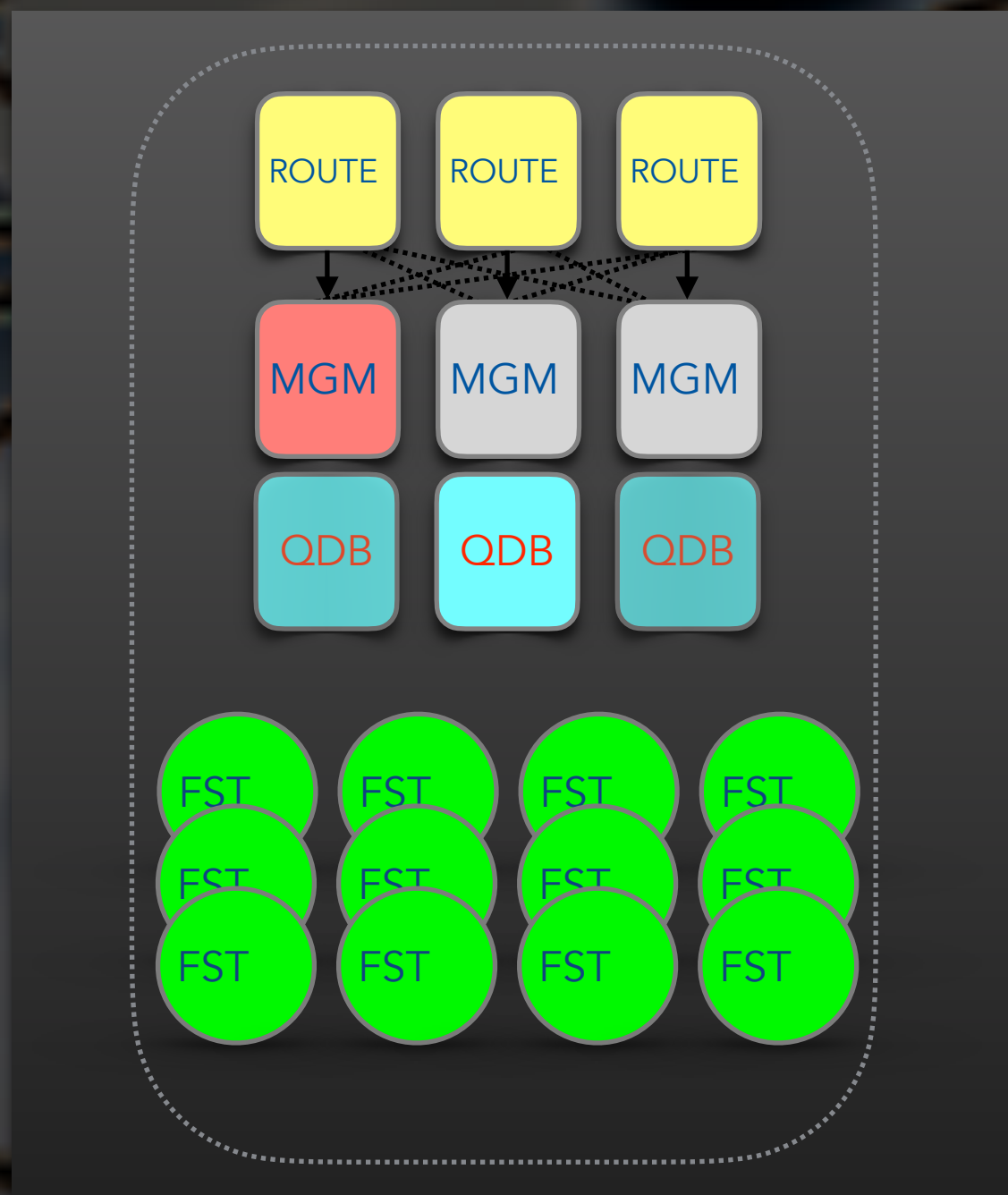
	Total space	Used space	Number of files
ATLAS	66.91 PB	51.62 PB	230 Mil
CMS	56.33 PB	42.76 PB	161 Mil
ALICE	65.33 PB	58.04 PB	867 Mil
LHCb	28.04 PB	18.56 PB	708 Mil
PUBLIC	54.93 PB	39.81 PB	334 Mil
ALICEO2	12.51 PB	2.98 PB	3.75 Mil
TOTAL	284.05 PB	213.77 PB	2303.75 Mil



Reliability - CERN-EOS in WLCG

New Architecture with QuarkDB persistency store

Drastic reduction of downtimes in physics instances

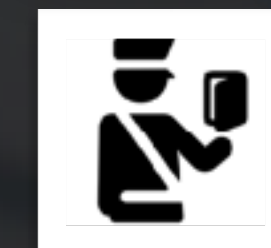
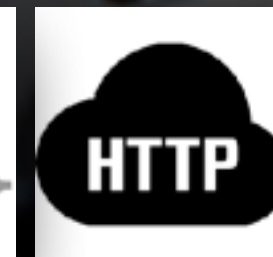
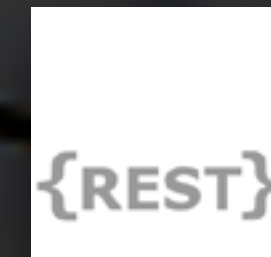
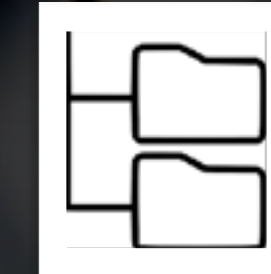
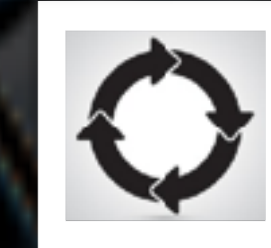




Development

Development field of work 19/20

- **namespace architecture** (MGM)
- **storage consistency** (FST)
- **filesystem access** (eosxd/ACLs)
- **tape integration** (CTA)
- **protocols/API**
(ProtBuf, XrdHttp, GRPC)
- **tokens & authorisation**
- **tokens & authorisation**



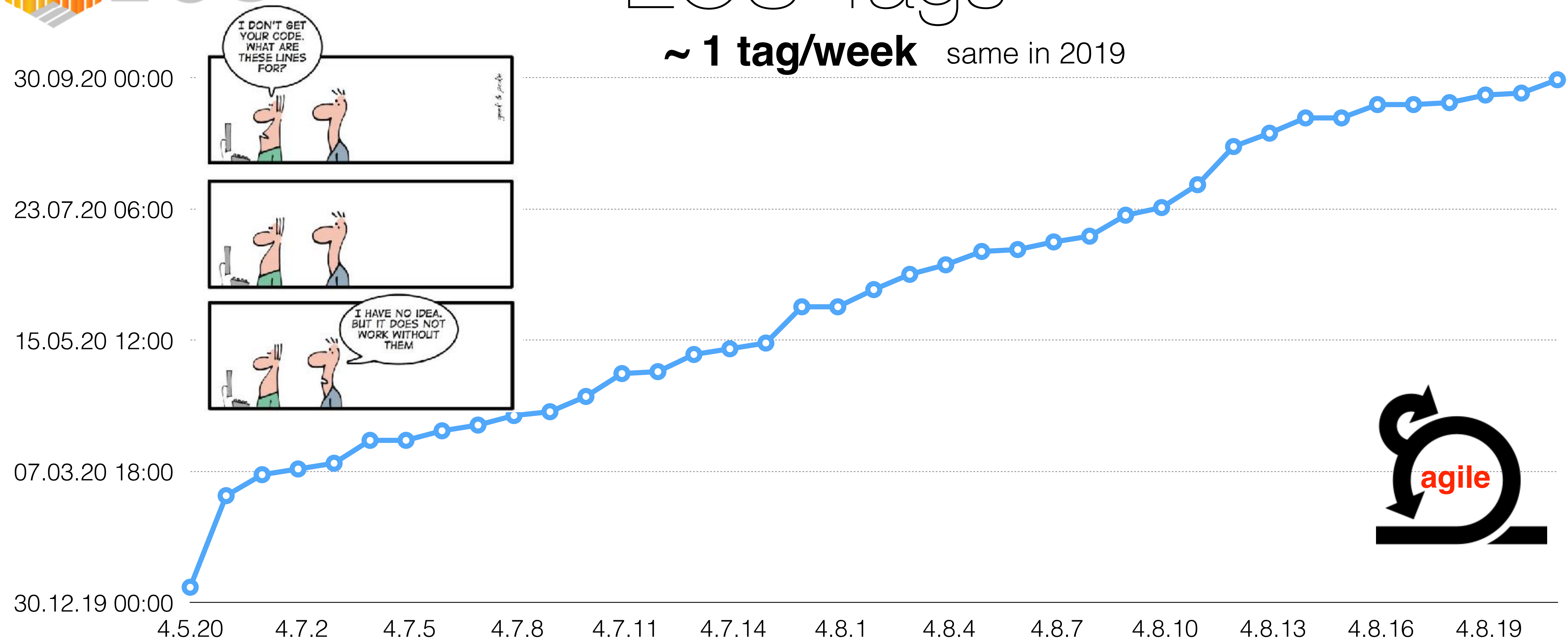
Few Highlights 2020

- **QuarkDB** persistency & **eosxd** fuse access on interactive & batch nodes
 - removed limitation in number of files per instance, more transparent software updates
- **GRPC** protocol usable by CTA & REVA
- **HTTPS** eco-system with TPC and WLCG tokens available
- **AliceO2** initial prototype testing achieved **20 GB/s WR** on 10 node cluster
- fully automated **CI infrastructure** with complete VM & kubernetes test facility



EOS Tags

~ 1 tag/week same in 2019



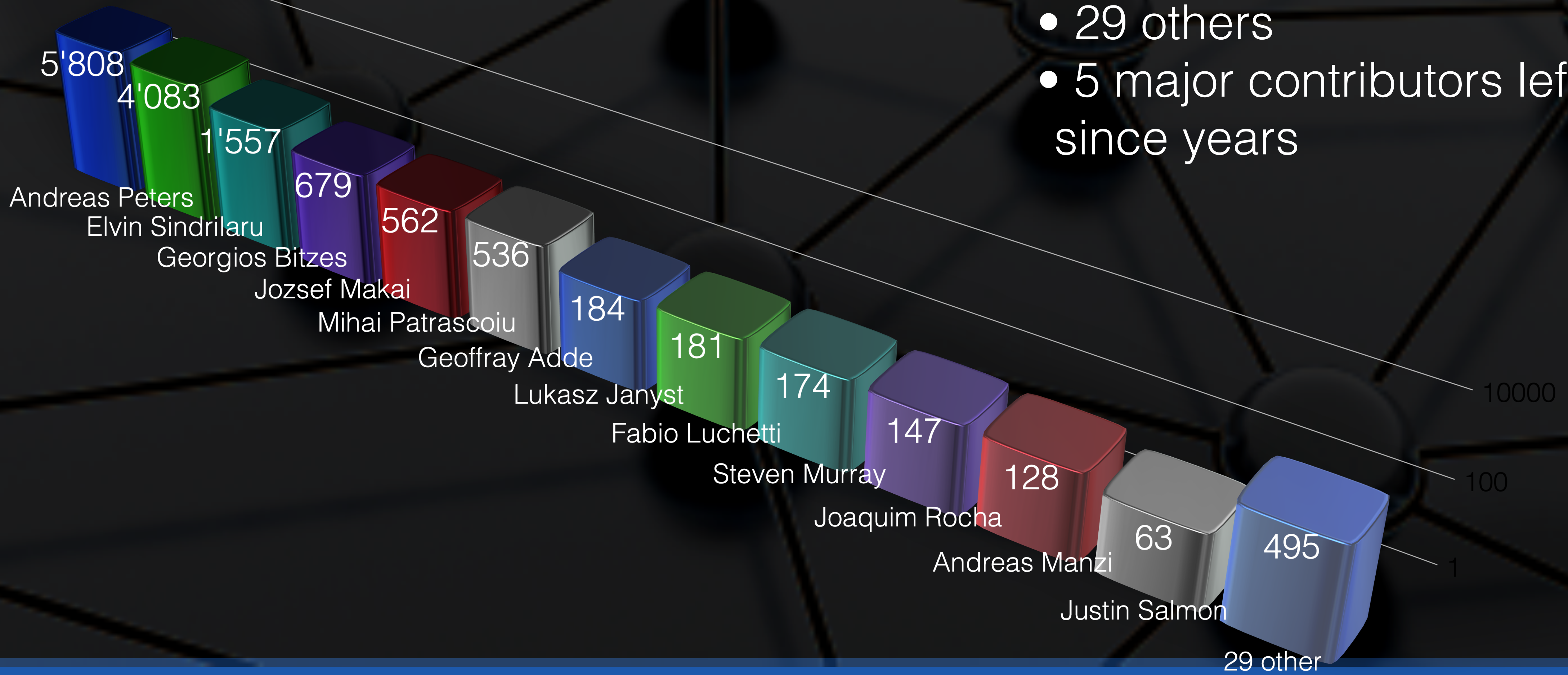
Production Release CITRINE **4.8.28**

Development & release process is very quick in reacting to urgent issues!

Contributors to EOS

#commits since begin

- 12 major contributors
- 29 others
- 5 major contributors left since years



vivid pool of developers



Performance - did we kill performance with QDB?

EOS QDB namespace

	EOS MGM* VM	Ceph MDS** VM
creation	1kHz	1..4kHz
1 x find	25kHz	14kHz
1 x find (ss cached)	100kHz	25kHz
10 x find (uncached)	100kHz	30kHz
10 x find (ss cached)	850kHz	30kHz
10x find —count	1.1MHz	n.e.
	* scale-out multi instance	** scale-out multi MDS

Is it not too informative to compare a pure POSIX namespace to the EOS namespace, which executes multiple transactions behind simple commands like 'create' 'rm' e.g. creating a version involves normally at least three operations create,mkdir,rename (+purging rm's)

HTTP Support

#dasselbeingrün



- HTTP support in EOS based on vanilla XRootD plugins (+extHandler)
 - **XrdHttp** - HTTP protocol
 - **XrdTpc** - third party transfers over HTTP
 - **XrdMacaroons** - Macaroon support
 - **XrdSciTokens** - SciTokens/WLCG JWT support
- deployed for **ALICE** (testing), **ATLAS** (testing), **CMS** (testing), **LHCb** (testing) at CERN

HTTP Support

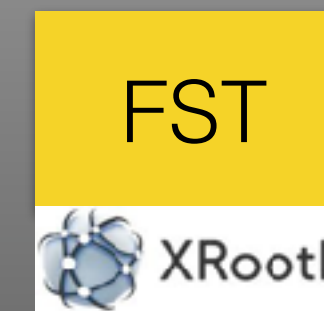
#dasselbeingrün

gateway-free deployment model

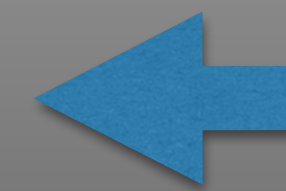
WLCG
Instance



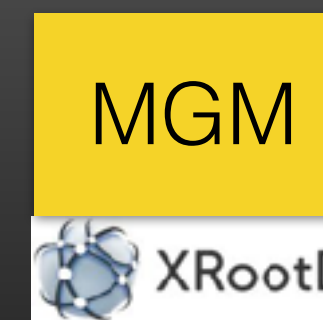
storage server
data



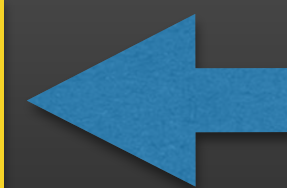
HTTP/TPC
PLUGINS



meta-data
namespace



HTTP/Token
PLUGINS

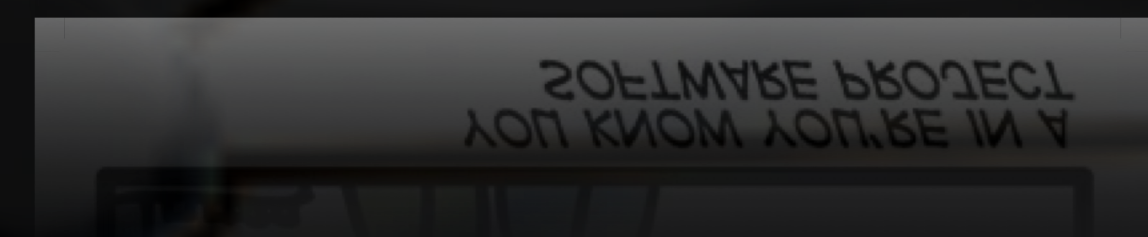


HTTP Support

#dasselbeingrün

Story & Outlook

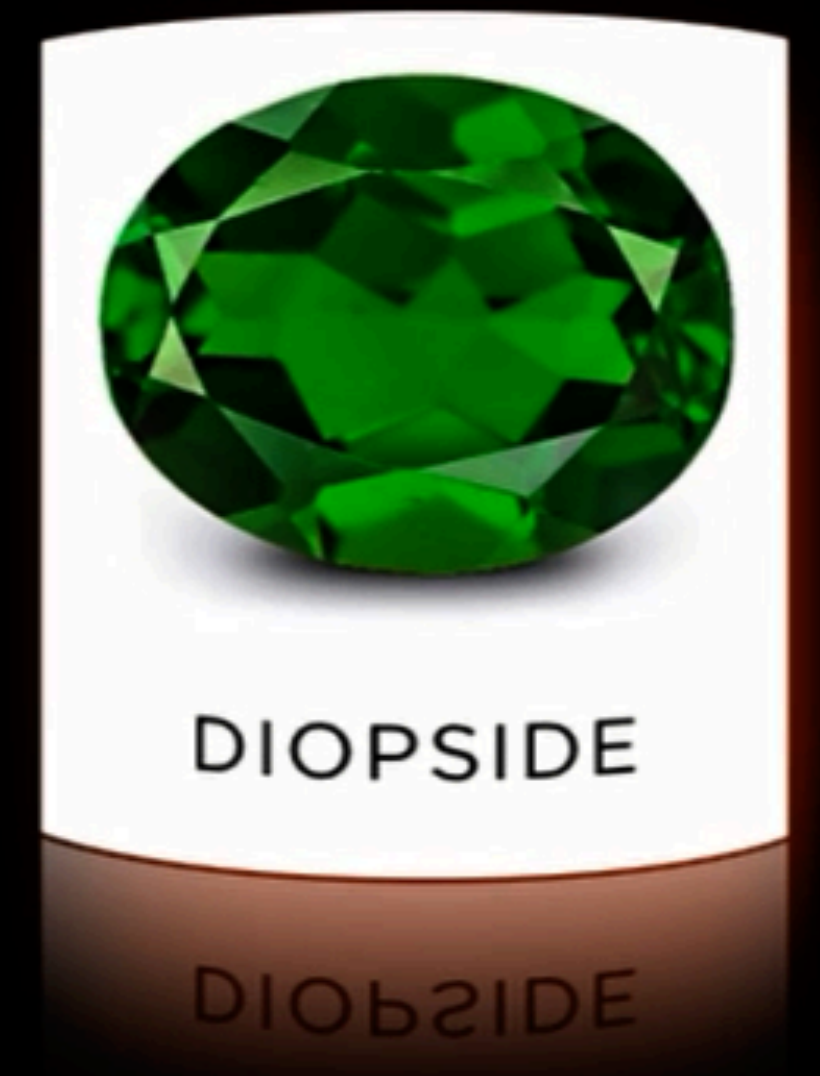
- **Several key issues fixed** in plug-ins since September
 - handling of proxy, multi-proxy certificates with gridmap-files
 - looping requests handling large/imcomplete headers
 - TPC pull race condition for checksum query
 - TPC missing URL encodings
- **group-based authorisation** to be implemented inside EOS
- use of **XRootD5** to get symmetric token/TPC support for HTTP+XRootD protocol





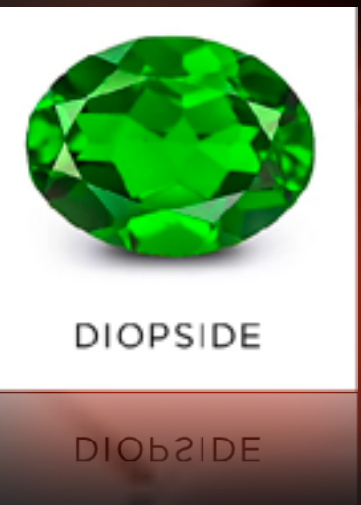
What's next: EOS 5

- **XRootD5**
 - brings **encryption** to make sense of token authorisation
 - **HA** EOS setup with so called redirect collapse
 - high **performance** IO with splicing, compound requests & new client API
- **Deprecation** of in-memory namespace, libmicrohttpd, async, transfer queues, rip-out unused code
- **Reduction** of MGM lock contention, target: latency per request under 1ms





What's next: EOS 5



- **Simplifications**

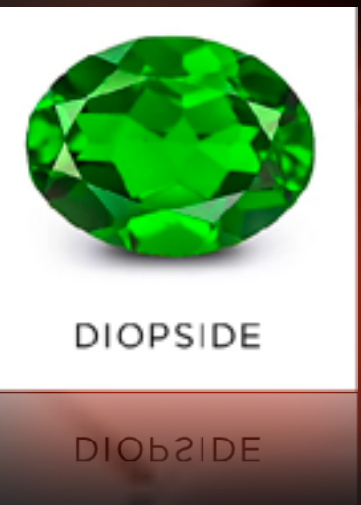
- only QuarkDB pub-sub for messaging - deprecate MQ service
- simplify EOS configuration and integration of QuarkDB as an EOS service

- **Scheduling**

- allow file updates on draining filesystems



What's next: EOS 5



- **Balancing**

- higher transaction rates

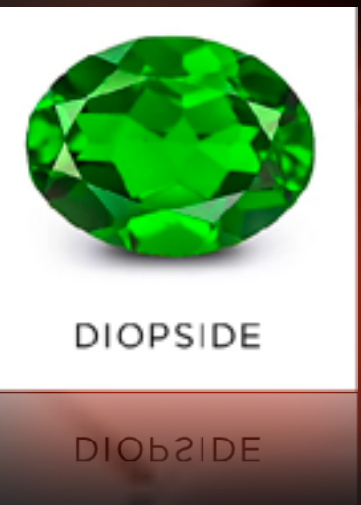
- **Conversion [QOS]**

- space placement by name and size (SSD for small files etc.)
- conversion jobs
 - dynamic erasure encoding (increase/decrease redundancy according to available space)
 - dynamic policy based conversion defined by subtree, name, size and layout





What's next: Operations



- **Latency improvements**

- use SSD storage and conversion policies and XFS SSD journal

- **Space usage reduction**

- EC configuration/operation for LHC instances, VDO compression





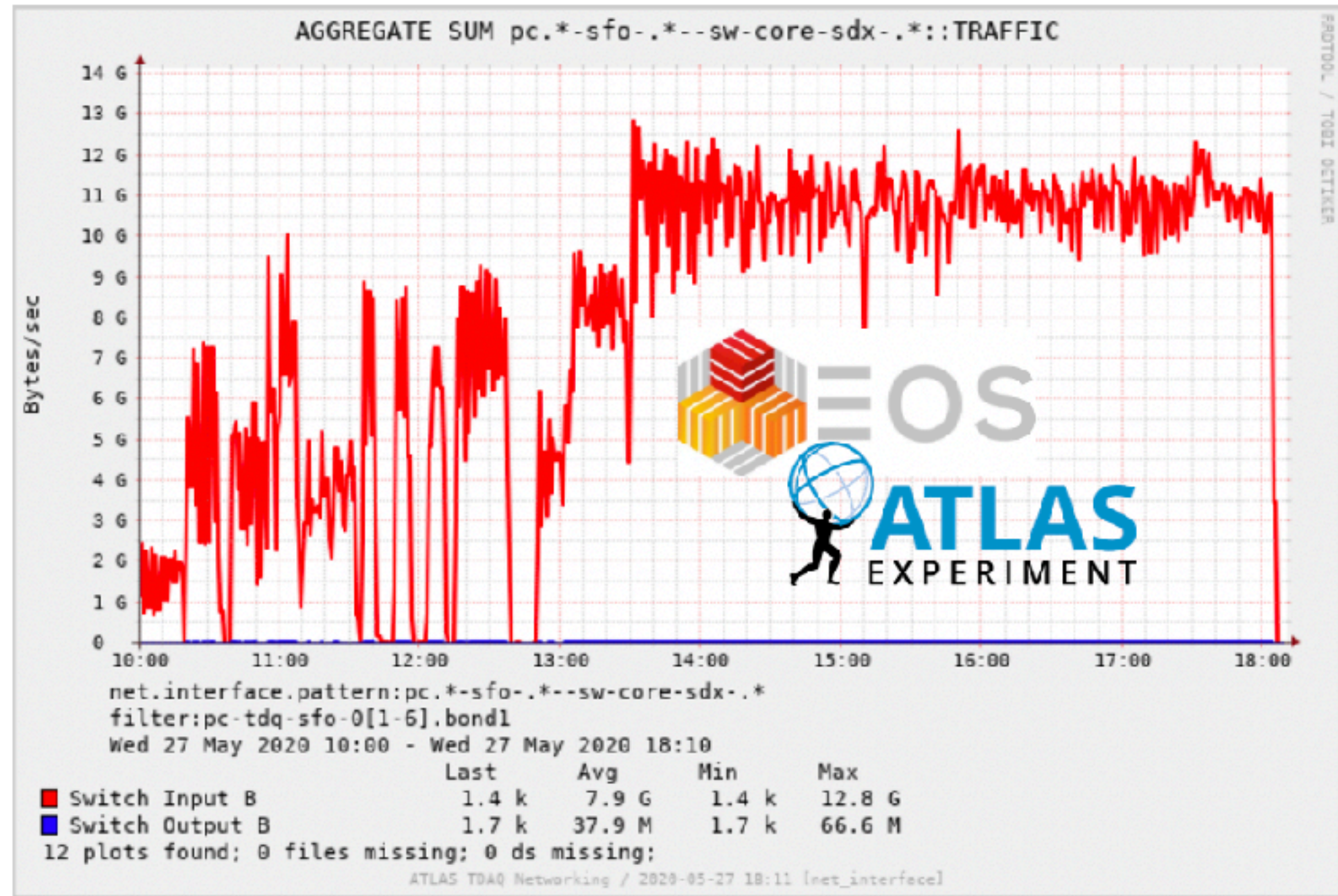
Outlook

2020+

RUN 3 Preparation: ATLAS

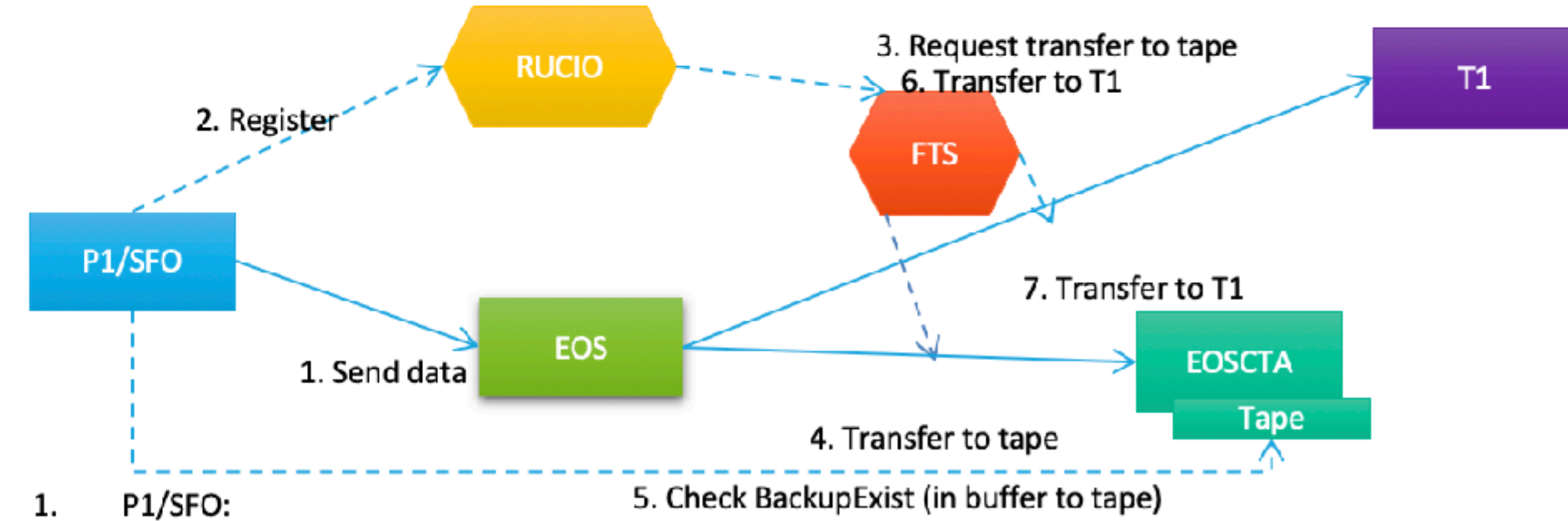
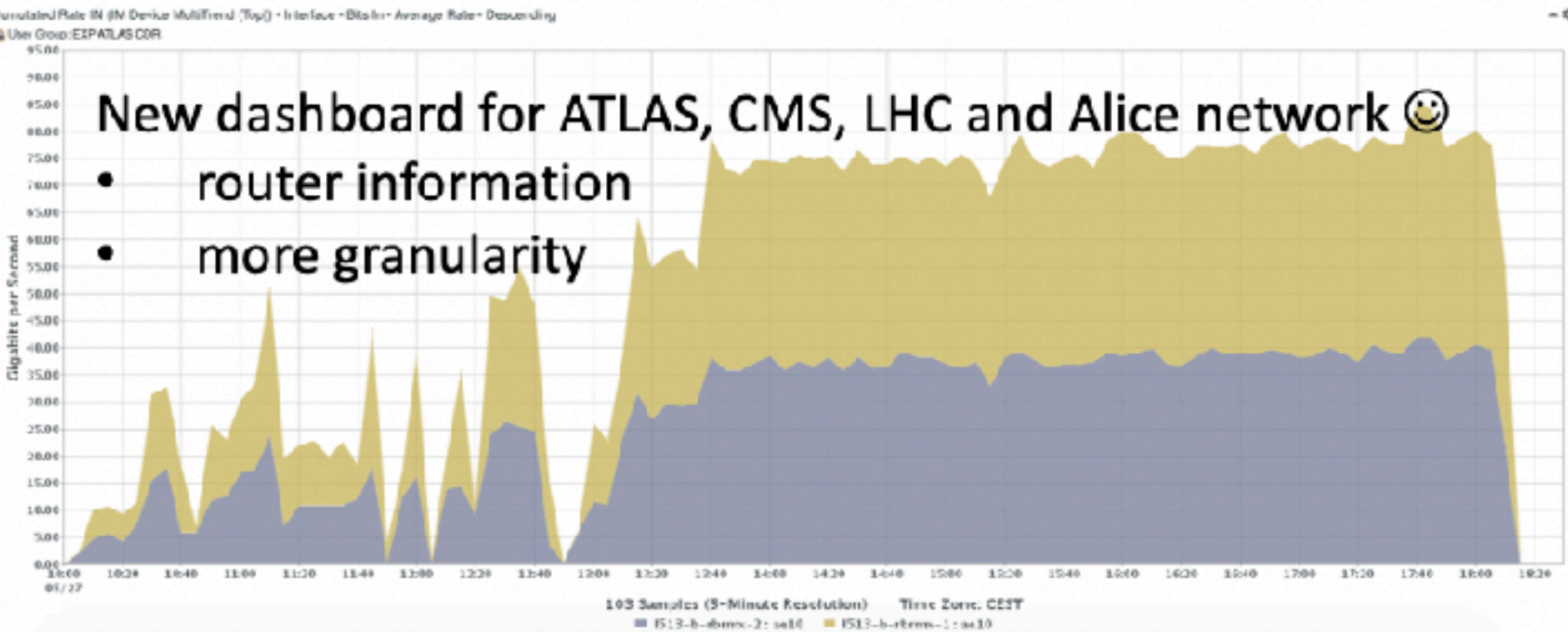
Atlas already performed 2 stress tests.

1. P1 to EOS achieving 9.5GB/s + peak of 12GB/s



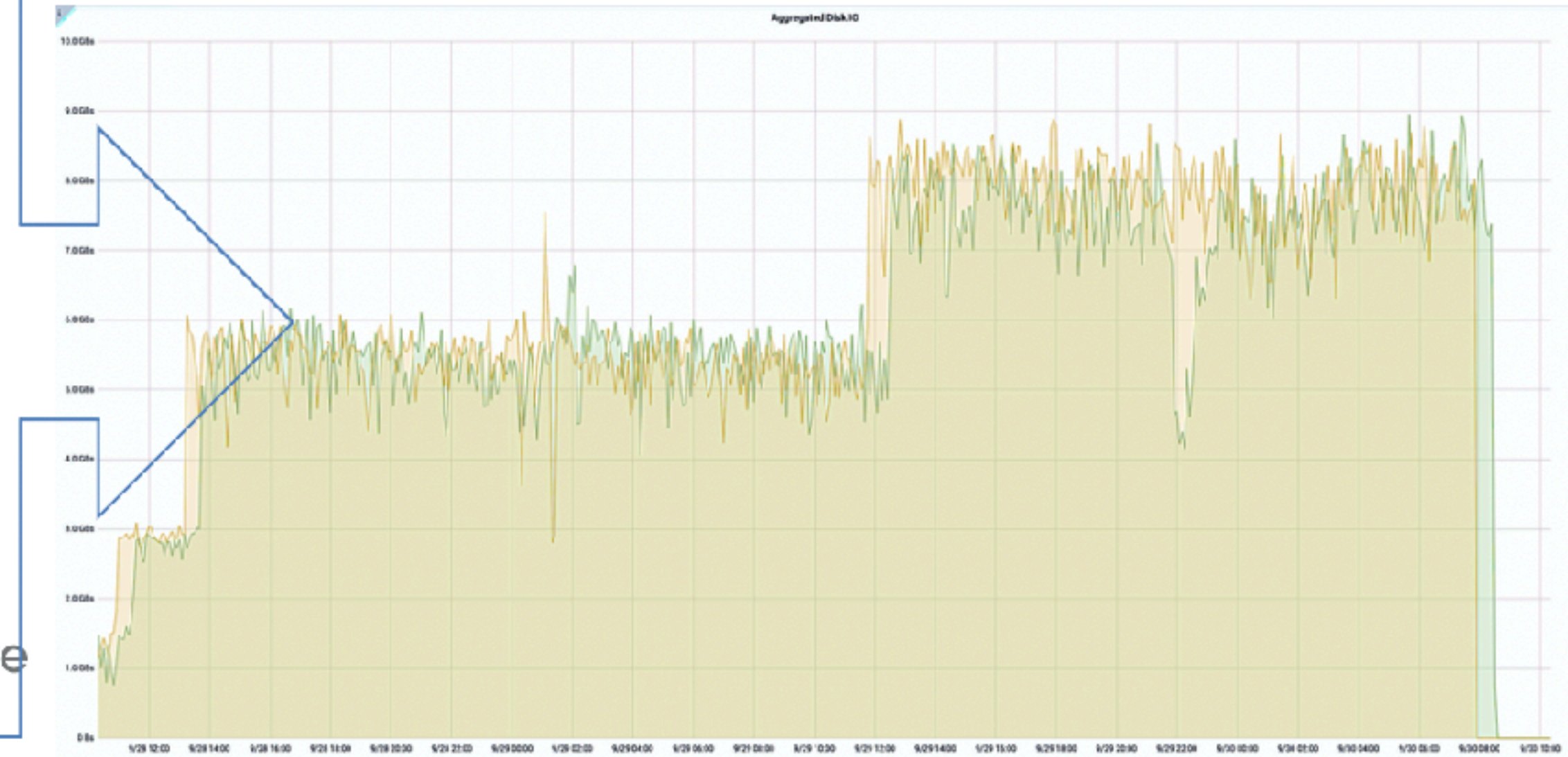
New dashboard for ATLAS, CMS, LHC and Alice network 😊

- router information
- more granularity



1. P1/SFO:
- Daq -> Write
 - Reconstruction -> Read/Write
 - Merging -> Read/Write

2. 1PB from EOS to CTA achieving more than 5GB/s + peak of 8GB/s with the available tape hardware

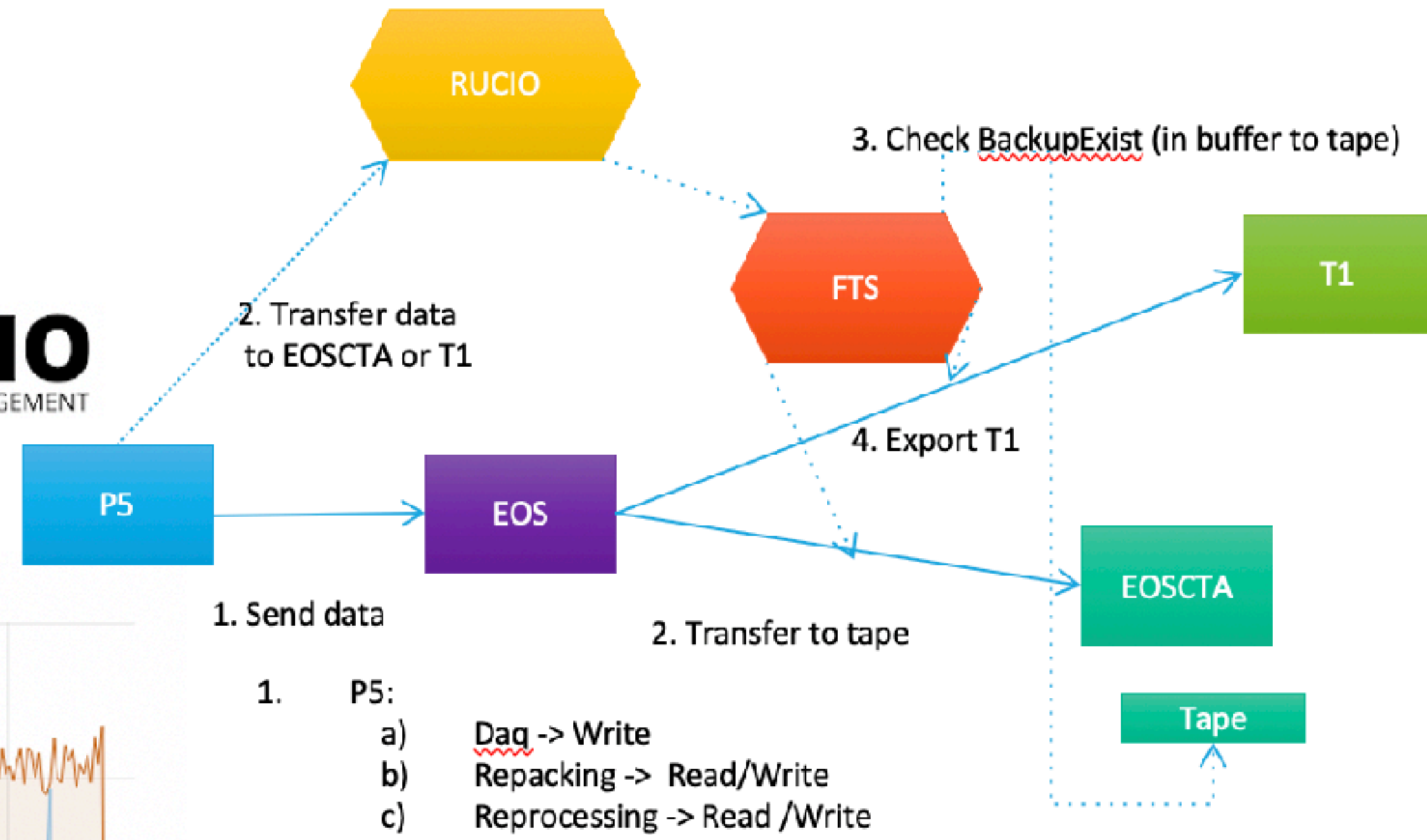


RUN 3 Preparation: CMS

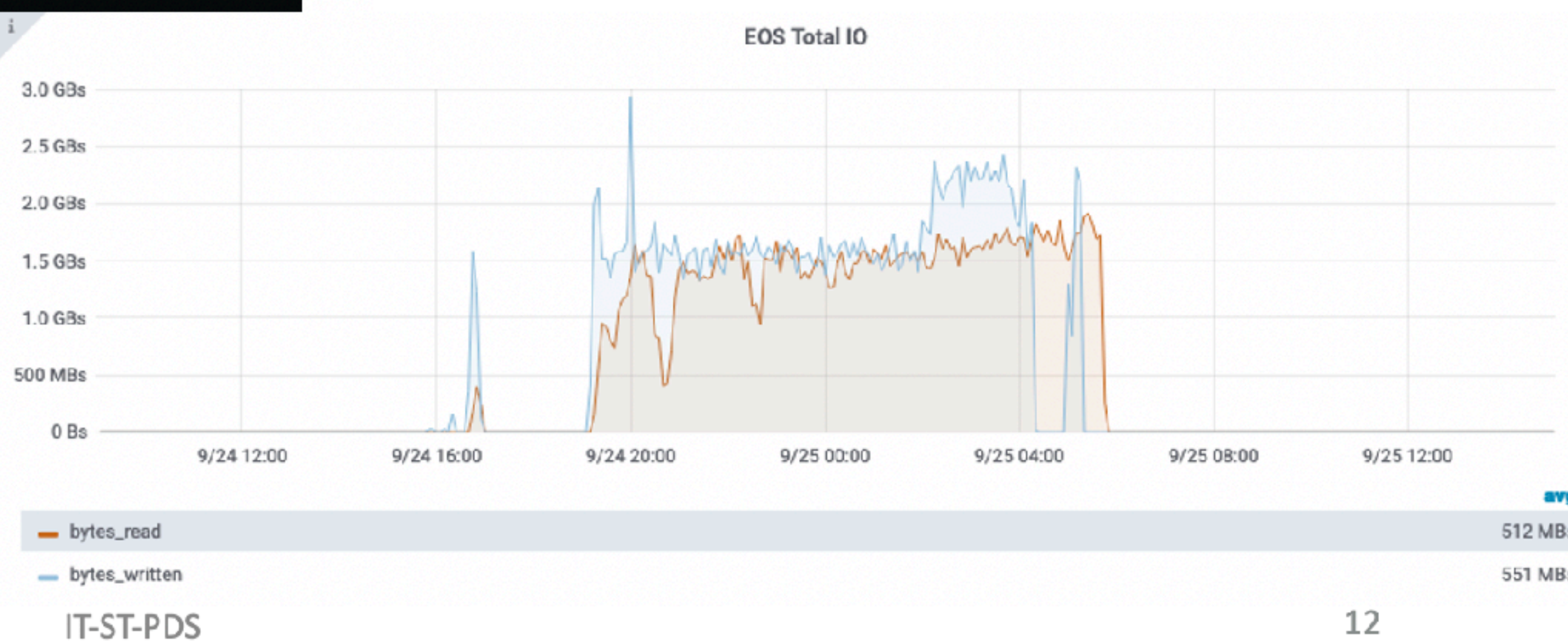
Challenge of CMS : Migration Phedex to RUCIO



1. Large test from EOS to CTA (~600TB) with the Rucio production instance



2. Multihop test JINR -> EOS -> CTA with Rucio in production

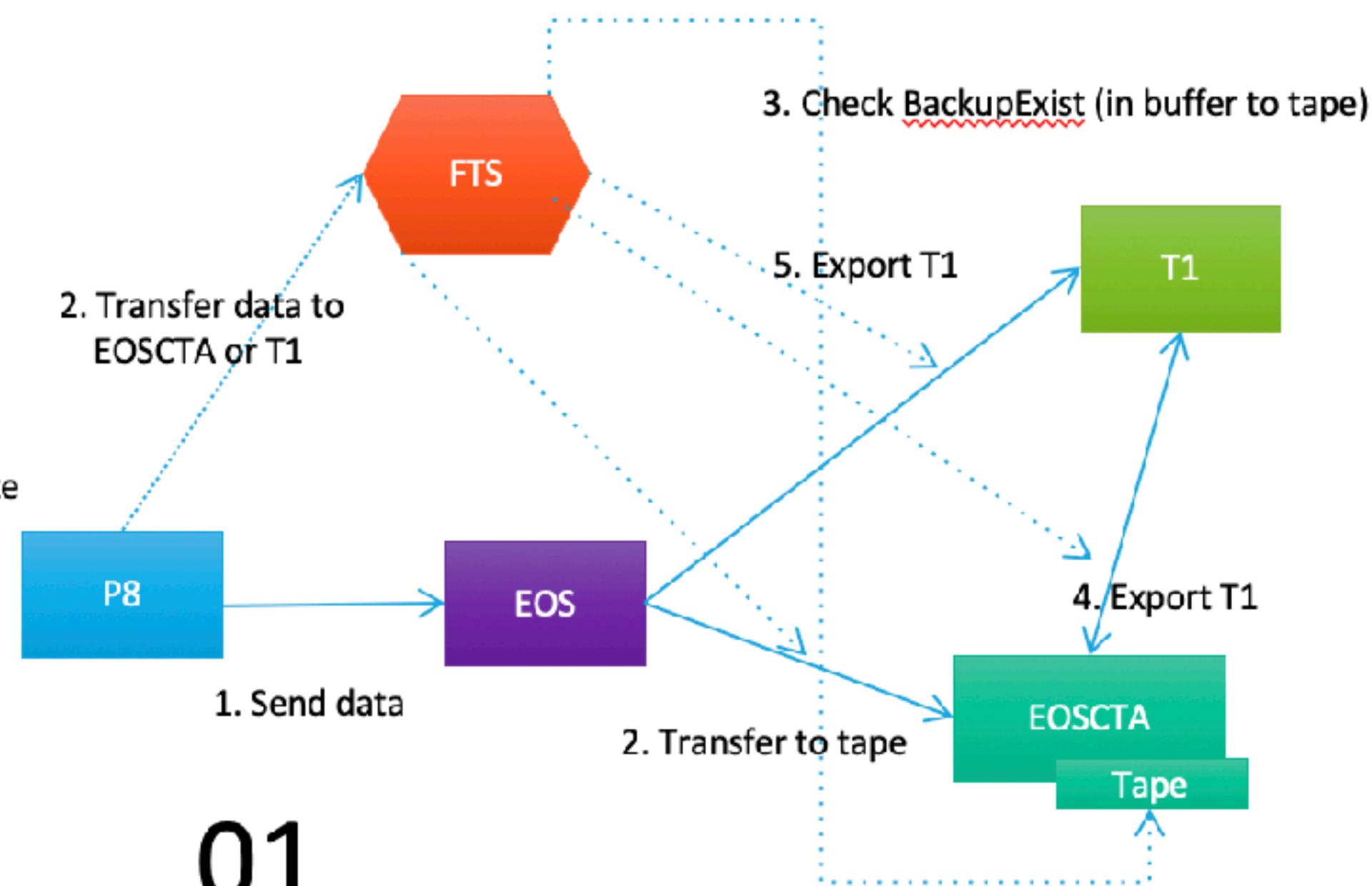


RUN 3 Preparation: LHCb

Challenge of LHCb : SRM-less + TPC constraints



1. P8:
 - a) Daq -> Write
 - b) Repacking -> Read/Write
 - c) Reprocessing -> Read /Write



01

Helping in following up SRM-less process in dCache namespace migration to enable LHCb exports

Gridka was migrated (27/07)

The migration of IN2P3, PIC and SARA will start in October

02

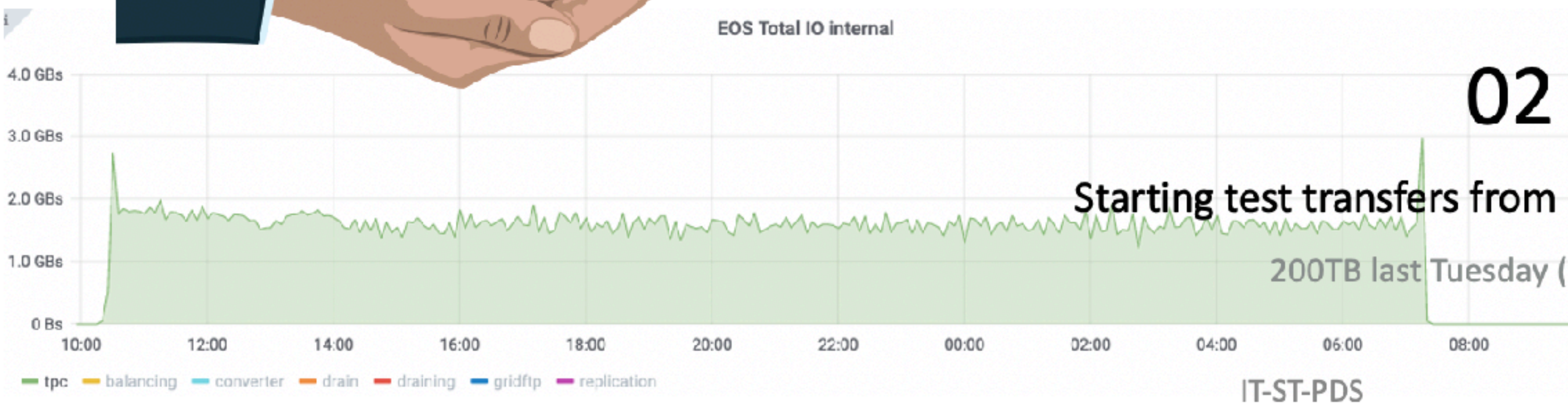
Starting test transfers from EOS to CTA

200TB last Tuesday (13/10)

03

Validate export with HTTP TPC

Already in place in EOSLHCb



RUN 3 Preparation: ALICE

Challenge of Alice: Increasing performance for ALICE O2: O2

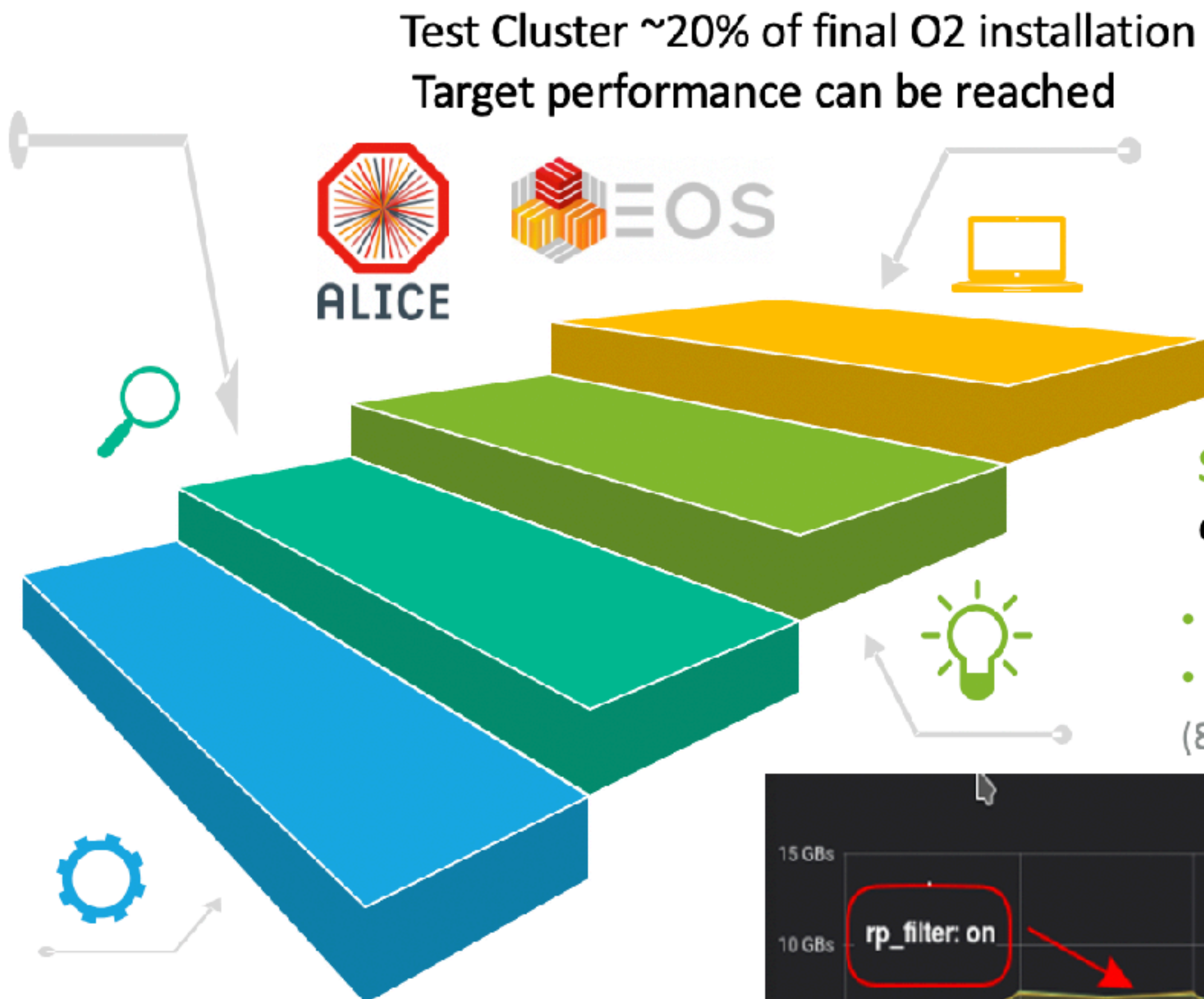
TEST cluster 10 nodes with 100GE and 96 disks each

Low performance with stock CC7 kernel and firewall issue in CC8

- **60 GBps per node** => 600 GBps for 10 nodes
- **Erasure encoding: 20GB/s**
(80 streams and RS(12,10) (clients on server))
- **Cause of the problem for CC8:** Reverse path filtering for IPv6
 - **rp_filter** ensures a packet would go back on the same interface as the input interface, otherwise it would be dropped

Low disk performance testing due to HBA card

Solution: HW HBA card upgrade from 6 GB/s to 12 GB/s per node



IT-ST-PDS

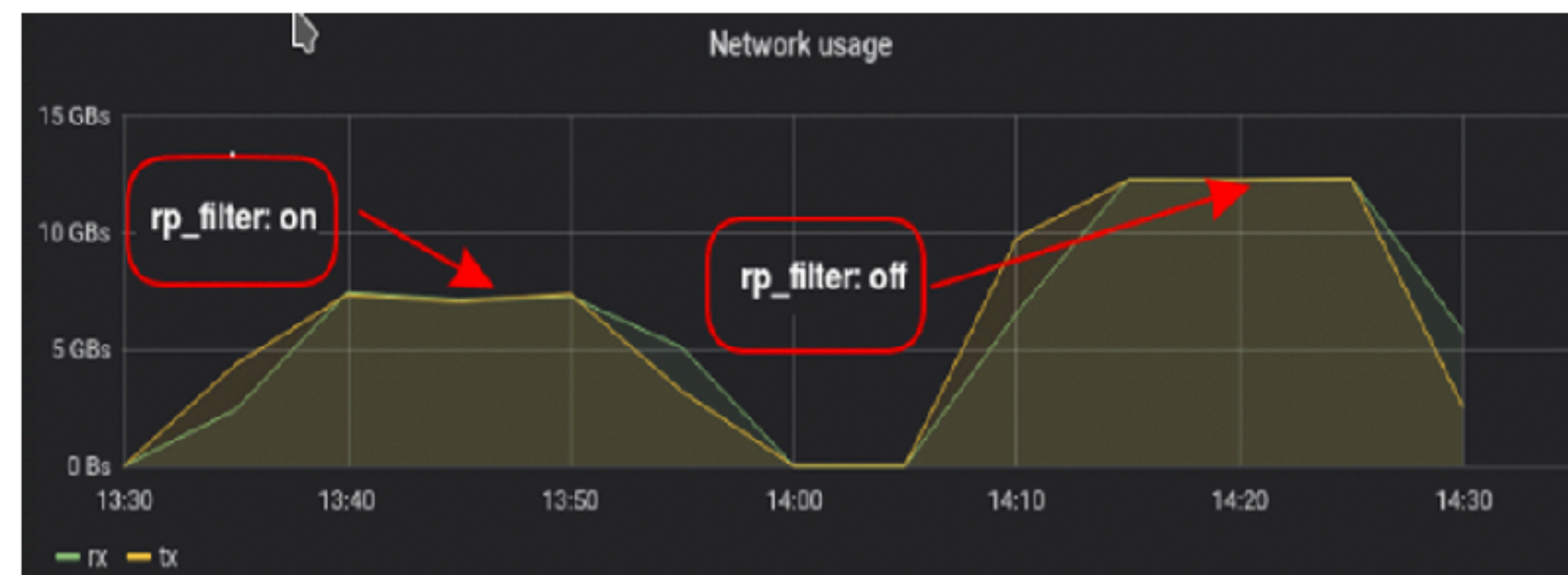
- P2:
 - Daq -> Write
 - Repacking -> Read/Write
 - Reprocessing -> Read /Write



- **main remaining problem:** variation (jitter) of transfer durations when many clients push data as fast as possible
- **possible solution:** bandwidth limitation per client

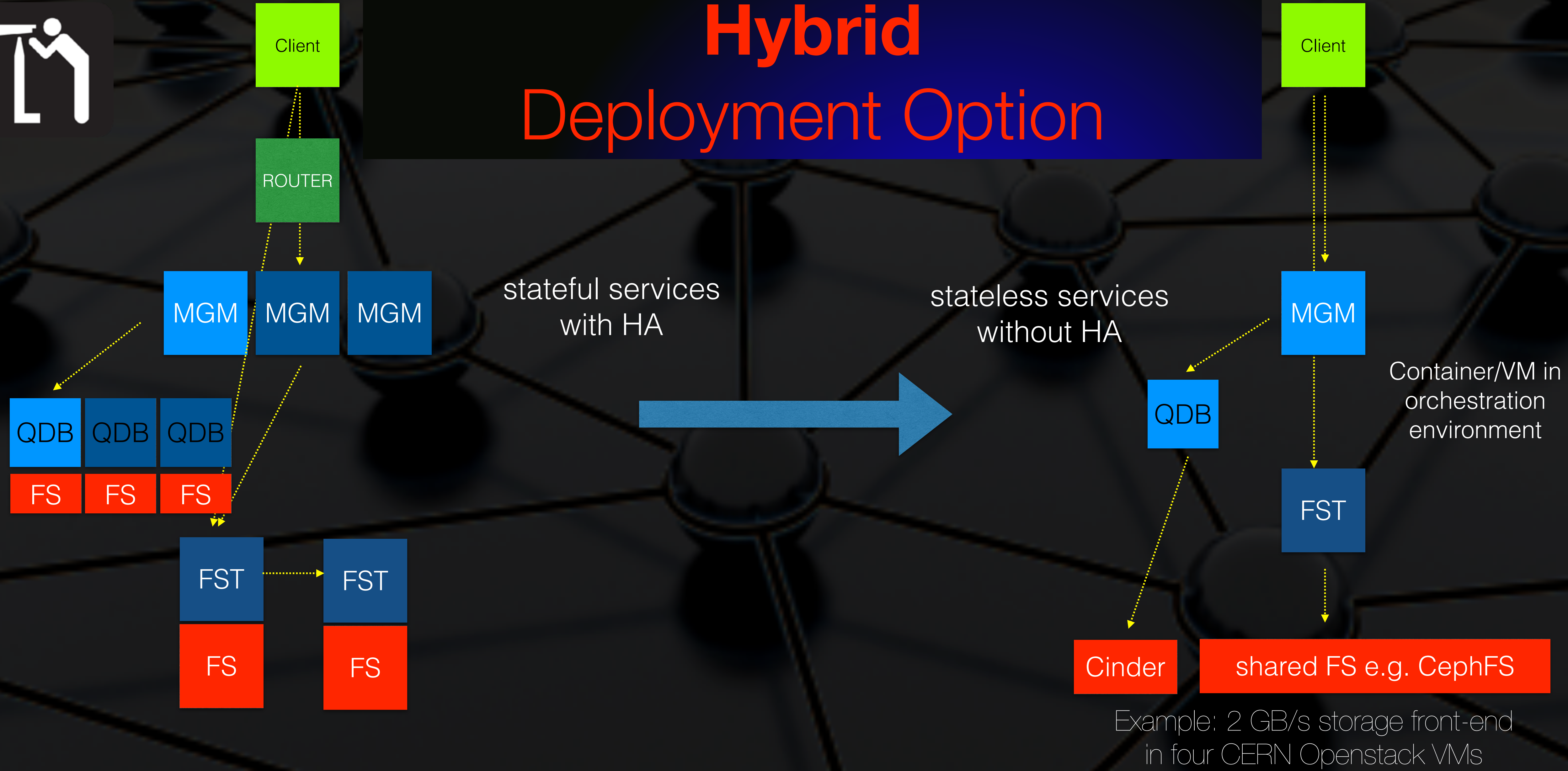
Solution: disabling the **rp_filter** since our storage nodes only have one IP

- **~92GBps per node**
- **Erasure encoding: 30 GB/s**
(80 streams with RS(12,10) (clients on server))





Hybrid Deployment Option



native deployment
physical nodes - native HA

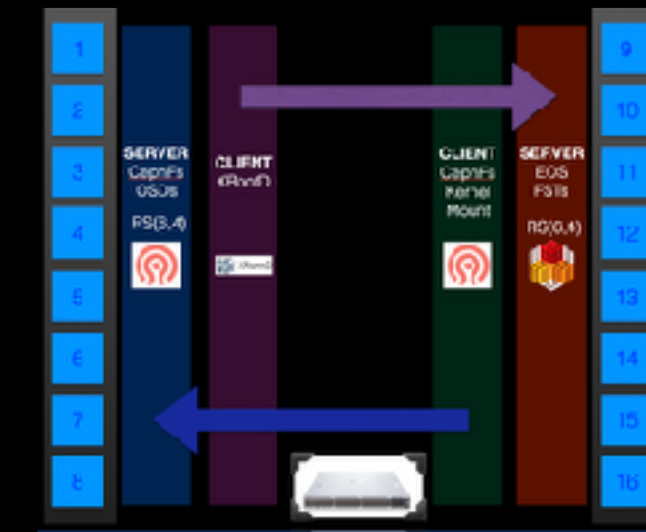
fabric deployment
virtual nodes + backend - fabric HA





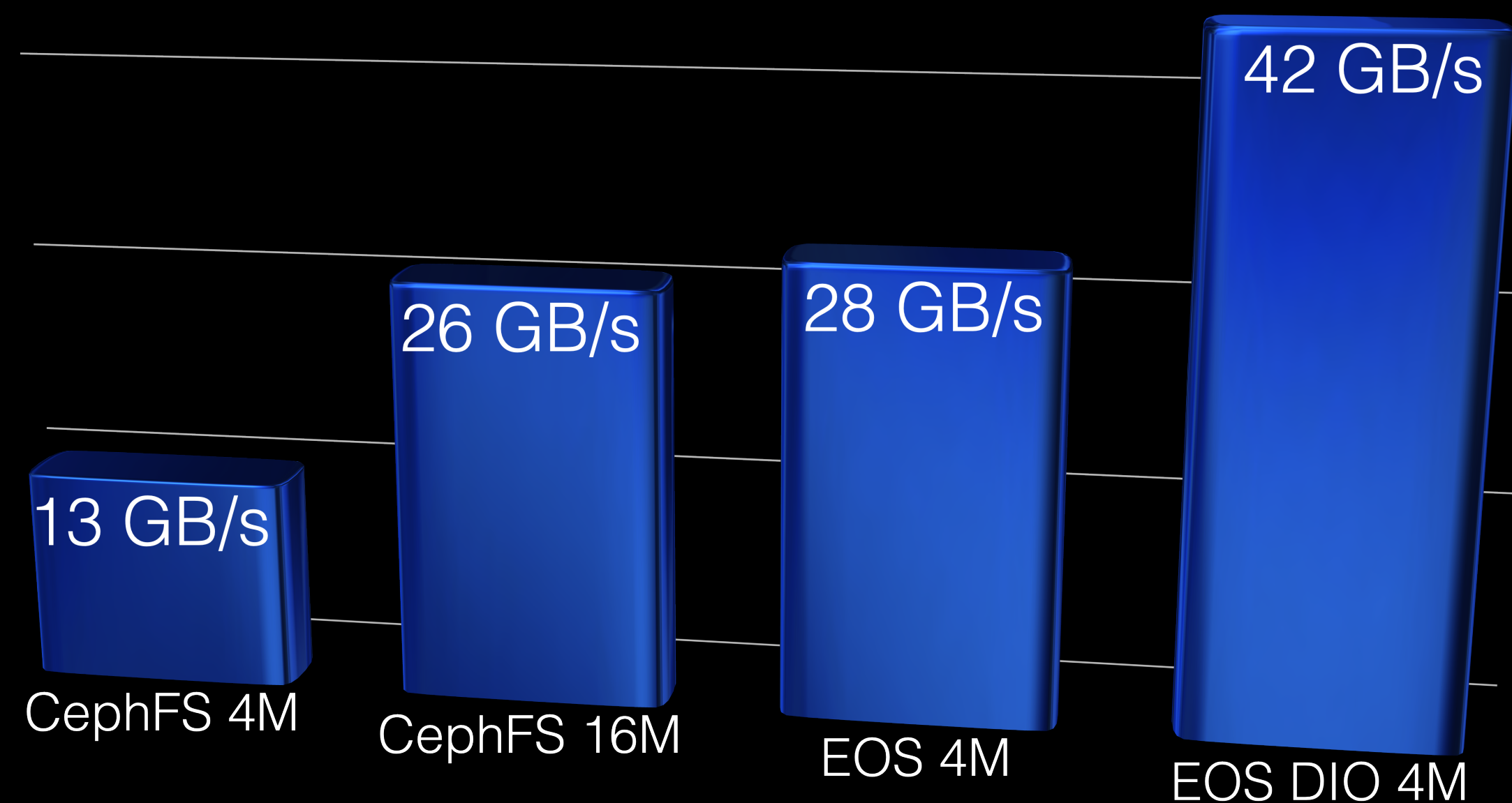
Erasure Coding for Streaming

1:1 Comparisons on same HW

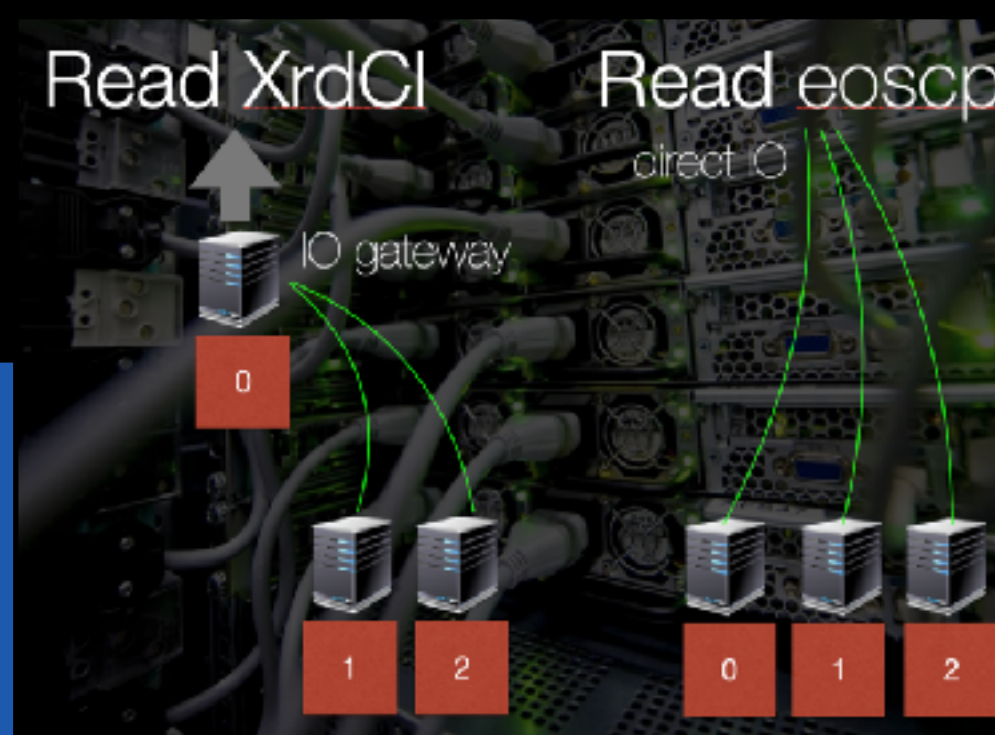


Streaming Read Performance

Large Files (GBs) - 240 [x4] streams on 480 disks



- 8 x 100 GE server with 60 disks - 192 GB mem
- 8 x 100 GE clients - all on same switch
- **EC IO network path** limits performance
- **CephFS** works well with adjusted parameters
 - a lot of advantages not discussed here
 - one observation: write performance decreases by 30% with filling of disks, another steep drop at 85% - default write stop at 95% @ 60% of empty system bandwidth
 - if you need maximum stream performance the write-/read-through OSD path is a bottleneck
- EOS **DIO** (direct client connection provides highest throughput by design)
- EOS development team start adding **DIO** also for writing now





EOS for Tier-1/2 Storage

- **EOS** includes a very powerful quota, access and user mapping system, accidental deletion protections, versioning, OAuth2, WLCG token/TPC support, realtime configuration and monitoring, **GRPC/WebDAV/XRootD/CIFS** access, active user community and developer support from CERN
- **EOS** brings a high-performance **tape**-storage solution with **CTA**
- **EOS** provides a drop-box-like **sync&share** platform with **CERNBOX**
- **EOS *natively*** can be deployed and configured in few minutes for a Tier-2 like setup
 - cost effective storage using HDDs with erasure encoding - simple management interface
- **EOS** is suitable with a **fabric** approach, where persistency is provided by a common infrastructure e.g. **CephFS**[Rados] and/or **Kubernetes** providing service **HA**

EOS in WLCG

Some of the sites running EOS ...

- Birmingham
- CERN
- Hiroshima
- IHEP
- IPNL
- JINR
- KFKI
- KISTI
- KOLKATTA
- KOSICE
- LBL
- NIHAM
- ORNL
- RRC-KI
- SARIFTI
- SPBSU
- SUBATECH
- TRIESTE
- VIENNA
- ZA_CHPC
- ...

1. Birmingham - EOS	ALICE::Birmingham::EOS	2	1.084 PB	453.7 TB
2. CERN - EOS	ALICE::CERN::EOS	0	29.72 PB	26.09 PB
3. CERN - EOSALICEDAQ	ALICE::CERN::EOSALICEDAQ	0	10.66 PB	265.8 GB
4. CERN - EOSALICEO2	ALICE::CERN::EOSALICEO2	0	10.22 PB	0.226 KB
5. Hiroshima - EOS	ALICE::Hiroshima::EOS	2	640.3 TB	381.5 TB
6. IPNL - EOS	ALICE::IPNL::EOS	2	109.1 TB	26.11 TB
7. ISS - EOS	ALICE::ISS::EOS	2	434.8 TB	138.7 TB
8. JINR - EOS	ALICE::JINR::EOS	2	1.456 PB	711.2 TB
9. KFKI - EOS	ALICE::KFKI::EOS	2	64.47 TB	0.226 KB
10. KISTI_GSDC - EOS	ALICE::KISTI_GSDC::EOS	1	1.953 PB	1.464 PB
11. Kolkata - EOS2	ALICE::Kolkata::EOS2	2	997.7 TB	153.4 TB
12. Kosice - EOS	ALICE::Kosice::EOS	2	571 TB	230.9 TB
13. LBL_HPCS - EOS	ALICE::LBL_HPCS::EOS	2	2.749 PB	1.667 PB
14. NIHAM - EOS	ALICE::NIHAM::EOS	2	3.4 PB	3.078 PB
15. NIPNE - EOS	ALICE::NIPNE::EOS	2	672.5 TB	259.5 TB
16. ORNL - EOS	ALICE::ORNL::EOS	2	2.725 PB	213.3 TB
17. RRC_KI_T1 - EOS	ALICE::RRC_KI_T1::EOS	1	4.152 PB	3.85 PB
18. SARFTI - EOS	ALICE::SARFTI::EOS	2	201.5 TB	93.64 TB
19. SPbSU - EOS	ALICE::SPbSU::EOS	2	136.4 TB	30.89 TB
20. Subatech - EOS	ALICE::Subatech::EOS	2	1.336 PB	1.069 PB
21. UNAM_T1 - EOS	ALICE::UNAM_T1::EOS	2	401.6 TB	268.1 TB
22. UPB - EOS	ALICE::UPB::EOS	2	2.643 PB	1.894 PB
23. Vienna - EOS	ALICE::Vienna::EOS	2	227.4 TB	5.415 TB
24. ZA_CHPC - EOS	ALICE::ZA_CHPC::EOS	2	348.8 TB	194.2 TB
Total			76.79 PB	42.2 PB

EOS team is interested in collaborations with WLCG sites to

- simplify deployments
- share monitoring tools
- explore hybrid and virtual deployments
- become experts in the community
- get code contributions to the project
- provide feedback on
 - EOS/CTA/CERNBOX

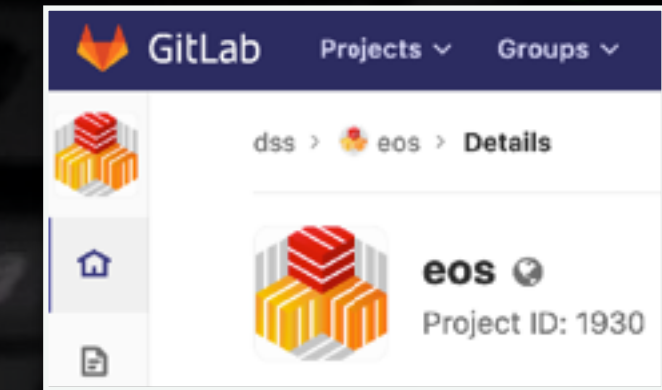
...



Web Page <https://eos.cern.ch>



Git Repository <https://gitlab.cern.ch/dss/eos>



Community Forum <https://eos-community.web.cern.ch/>

email: eos-community@cern.ch



Documentation <http://eos-docs.web.cern.ch/eos-docs/>



Support email: eos-support@cern.ch



THANK YOU!

