CERN IT
Disk Storage Services

Luca Mascetti
CERN IT-Storage

CERN IT-Storage

provide and operate storage solution and tools for data management and data analysis to experiments and users.

CERN IT-Storage Group

CERN IT-Storage
2020 Group (Re)Structure

ST Group Leader

Alberto Pace

CERN IT-Storage
2020 Group (Re)Structure

ST-TAB

Section Leader

Oliver Keeble

TAB: Tapes, Archives and Backups

CERN IT-Storage
2020 Group (Re)Structure

ST-PDS
Section Leader

Luca Mascetti

PDS: Physics Data Services

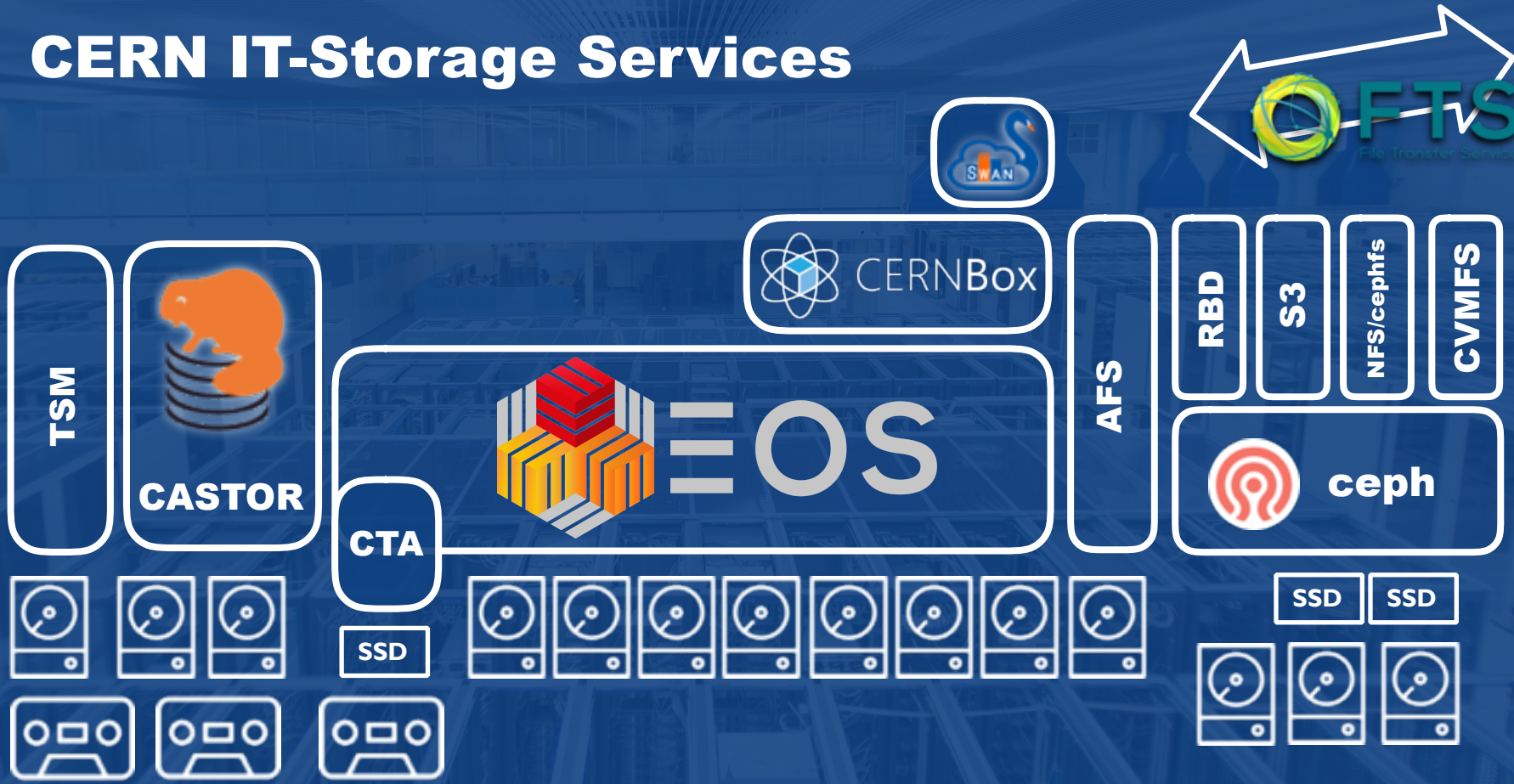
CERN IT-Storage
2020 Group (Re)Structure

ST-GSS
Section Leader
& Deputy GL

Jakub Moscicki

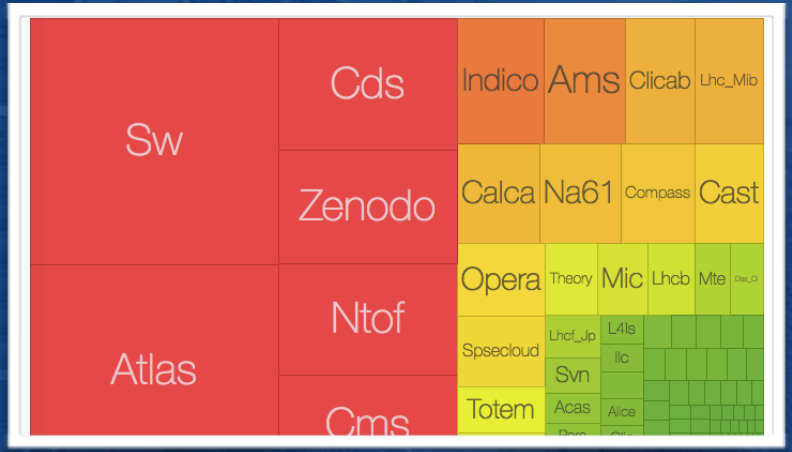
GSS: General Storage Services

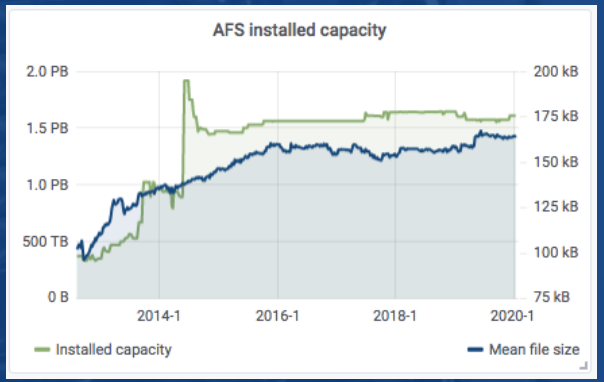
CERN IT-Storage Services

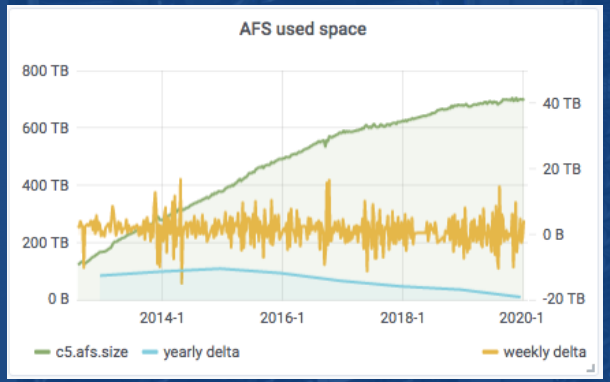


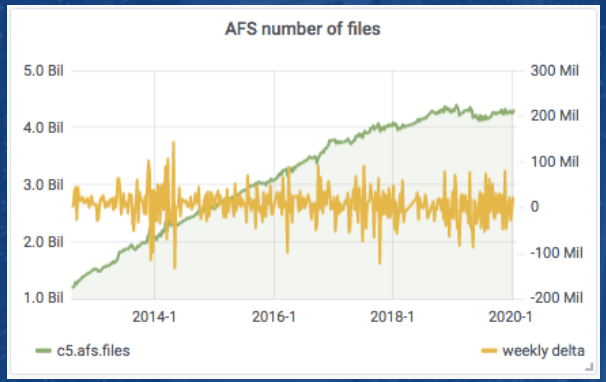
AFS

- 1.480 PiB total installed capacity
  - 642 TiB used
  - 4.2 B files (avg files size:160kB)
- ~100k Volumes
- ~25k clients
  - 400-500 Mio. Reads/Writes









Ceph, CVMFS and S3

CERN Ceph Clusters		Size	Version
OpenStack Cinder/ Glance	Prod.	6.4 PiB	mimic
	~~Wigner~~	~~1.6 0 PiB~~	~~nautilus~~
	Hyperconv.	245 TiB	mimic
CephFS (HPC+Manila)	Prod.	1.09 PiB	luminous
	Pre-prod.	164 TiB	mimic
	Hyperconv.	356 TiB	mimic
CASTOR	Disk Buffer	5.5 PiB	nautilus
S3+SWIFT	Prod. (4+2EC)	1.92 PiB	luminous

Ceph

- Openstack RBD remain the biggest use-case
  - ~1PB/year
- Openstack Manila Share in production during 2018
  - NFS-like share
- Storage for Kopano

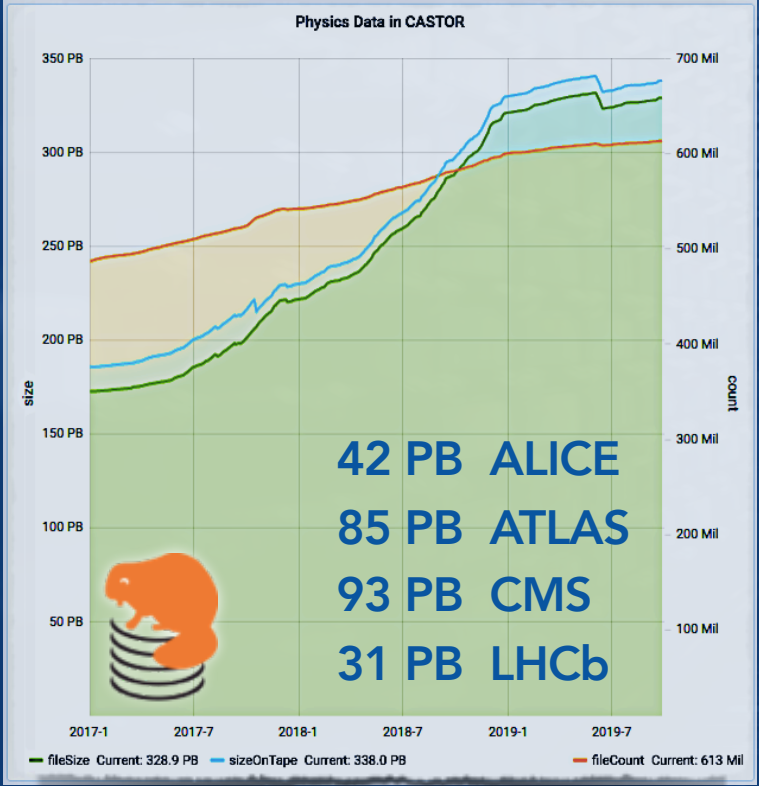
CVMFS

- 49 repositories with 868 M files and 52 TB
  - 43 using block storage 6 using S3 objects

S3

- IT applications, ATLAS+LHC@home
- CVMFS repositories
- CERNBox new backups using Restic

CASTOR and CTA



330PB of data (+10 PB dual copy)

~0.6 EB capacity

Minimal space increase in 2019

"Only" 16 PB of disk cache on CC7

Preparing the migration to CTA

- ATLAS will be the first
  - Ongoing data reprocessing on CTA
  - Scheduled 1PB write stress-test

CERNBox and SWAN

CERNBox

- 5 years of production
  - still growing service with 18k users
  - 4PB+ data 1B+ files 110k+ shares
  - new EOS backend deployment
- New core daemon (Reva) in production
- Consolidating HOMEs into CERNBox
  - Migration out of DFS ongoing
  - HA SAMBA gateway in production
- Central Hub for CERN data and apps
  - draw.io, onlyoffice, collabora, kopano, …

SWAN

- Turn-key data analysis platform
- Accessible from everywhere via a web browser
- Support for ROOT/C++, Python, R, Octave
- Fully integrated in CERN ecosystem
  - Storage on EOS, Sharing with CERNBox
  - Software provided by CVMFS
  - Massive computations on Spark

EOS Project History

Andreas J. Peters
EOS Project Lead

timeline

2009 architecture document

2010 EOS Open Source Storage

2011 IPv4 — EOSATLAS, EOSCMS, EOSALICE, EOSLHCB

2012 — EOSPUBLIC

2013

2014

2015 — EOSUSER

2016

2017 IPv6 — EOSMEDIA

2018 — EOSALICEDAQ, EOSCTAATLASPPS, EOSHOME

2019 — EOSPROJECT

2020

CASTOR
XRootD
castor-xrootd
IF

lustre
XRootD
lustre-xrootd
IF

v 0.2 *AMBER*

v 0.3 *BERYL*

v 0.3.x *AQUA MARINE*

v 4.0 *CITRINE*

v 4.3 Quarkdb RocksDB

v 4.6

sync&share — CERNBox

swan — SWAN

wopi — Office

tape

SEAGATE KINETIC — kinetic R&D

librados R&D

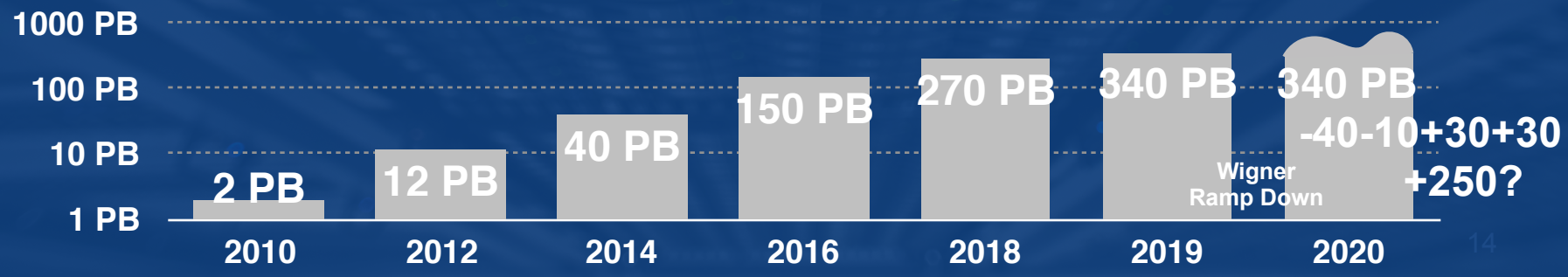
eXtreme DataCloud — QOS, caches & data lakes

EOS

Production Instances:

- 5 for the LHC experiments
- 9 CERNBox
  - EOSUSER + 5 EOSHOME + 3 EOSPROJECT
- EOSMEDIA (photo/audio/video)
- EOSPUBLIC (Open Data and non-LHC experiments)
- EOSBACKUP (backup for CERNBox)

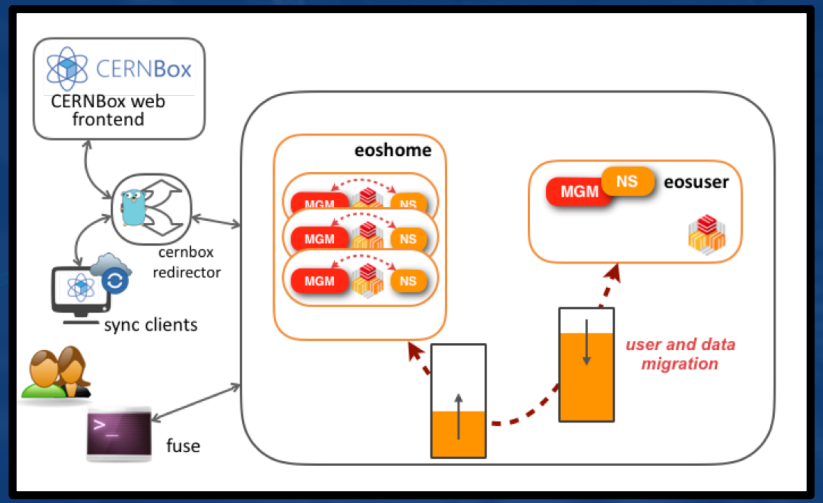
Total Space	Files Stored	# Storage Nodes	# Disks
340 PB	5.67 Bil	~1600	~60000

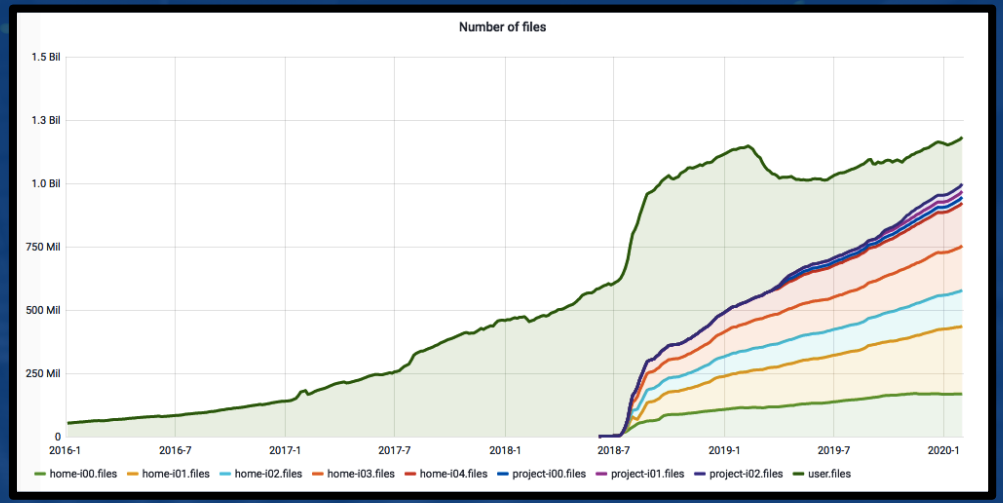


EOS

CERNBox instances

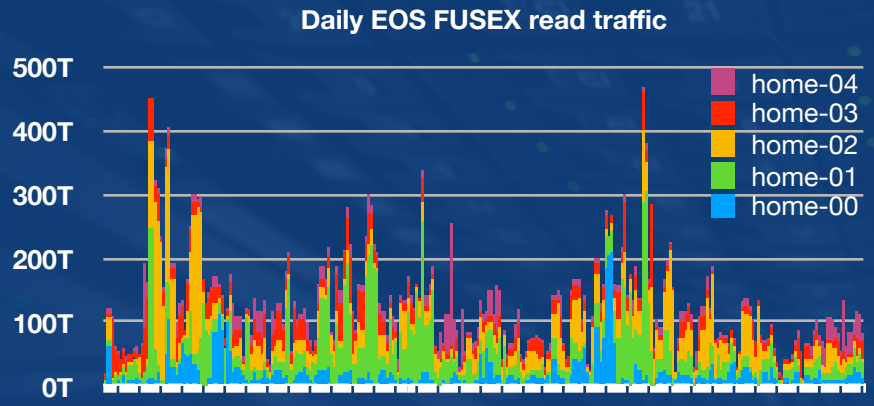
- Ongoing Projects migration
- Users migration completed

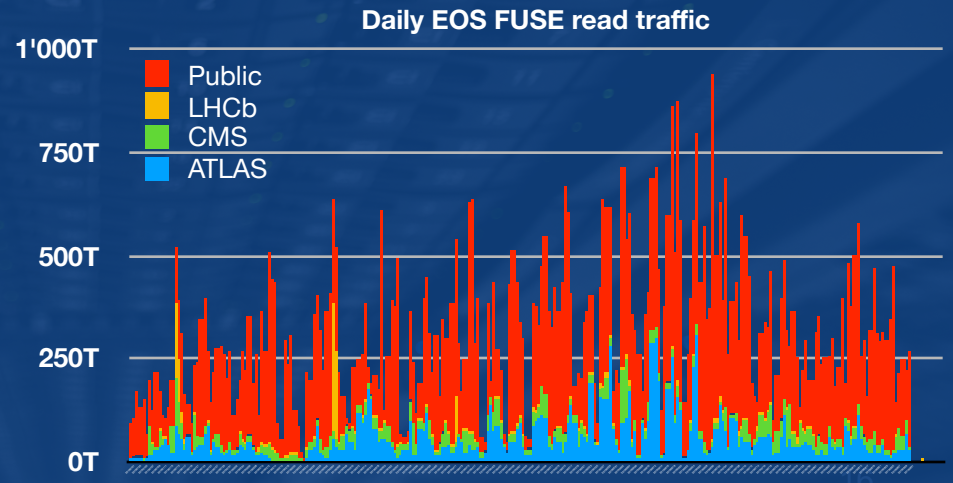




EOS

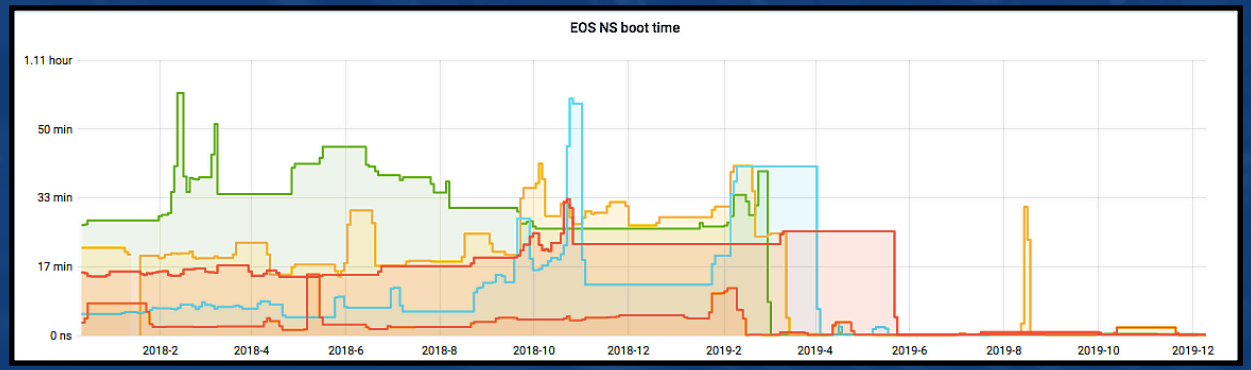
- Deployment of a new EOS architecture for CERNBox instances
  - latest generation EOS mounted filesystem (fusex) deployed

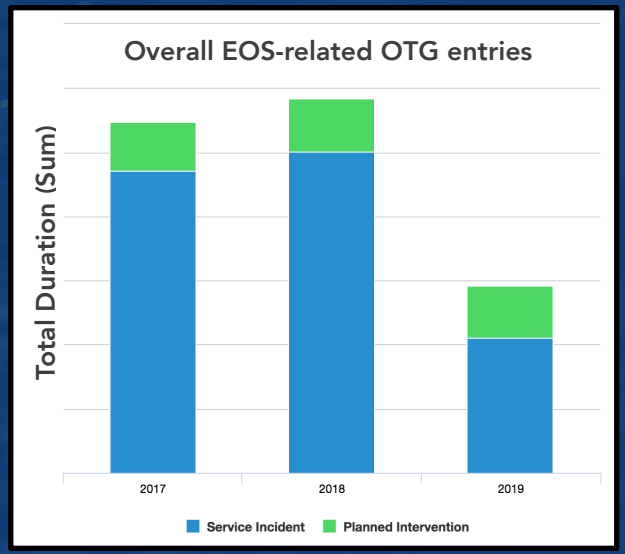




EOS

- Full deployment of the latest generation namespace (quarkdb)
  - reduced restart time
  - reduced memory requirements
  - improved availability during 2019

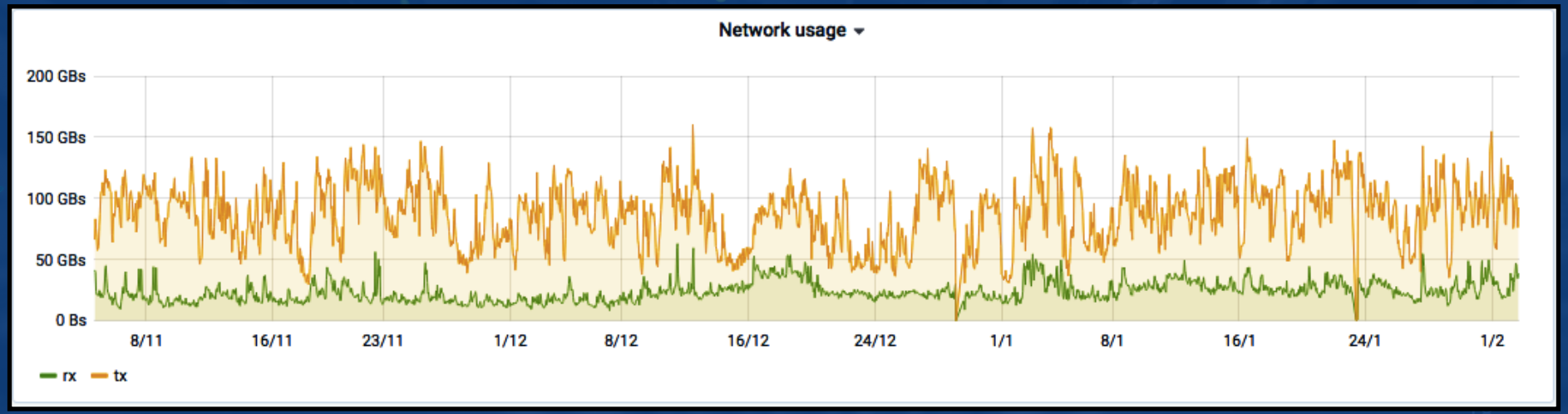




EOS

LHC and non-LHC instances

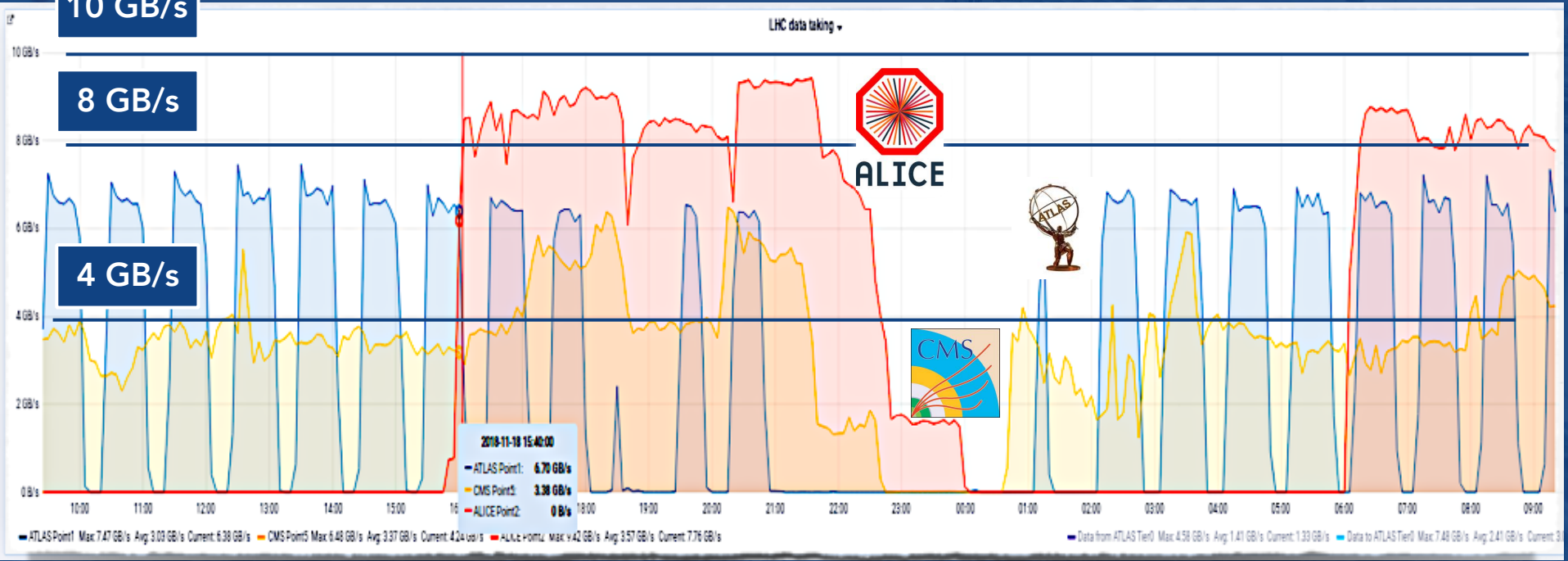
- “LS2”: Long Shutdown 2
  - higher activity than last data taking…



LHC Run2: Heavy Ions

EOS Rates during data taking (HI Run2 2018)

For Run3 expecting rates up to 10x higher…



EOS - Storage Hardware evolution

- Profiting from economy of scale
  - minimise price per TB

- Latest generation of storage servers
  - 8 trays (24x disks) per system unit
    - ~2300 TB (12TB drives)
  - 4 trays (24x disks) per system unit
    - ~1150 TB (12TB drives)
    - ~1340 TB (14TB drives)

- High Density JBOD
  - 2 trays (60x disks) per system unit
    - ~1680 TB (14TB drives)

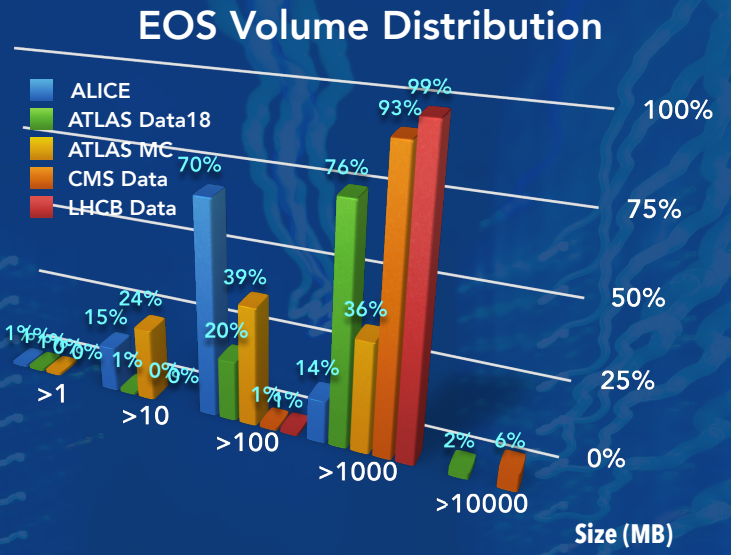
Future Challenges

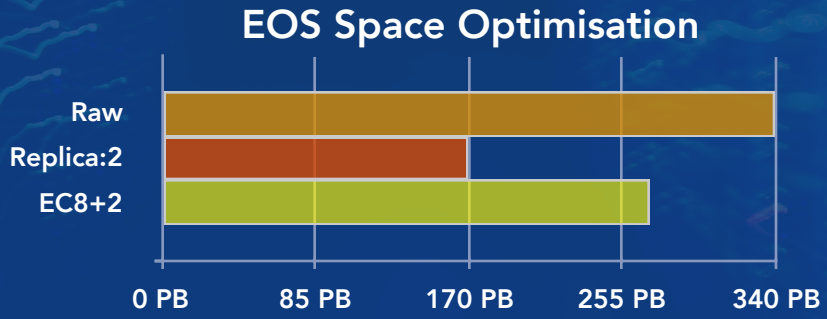
Disk Space Optimisation

- Large Scale Erasure Encoding
  - EC4+2 EC8+4 EC10+2

High Density Storage Solutions

- Run3 Storage purchasing
- ALICE O2 Storage Tests
- LHCb Storage Synergies





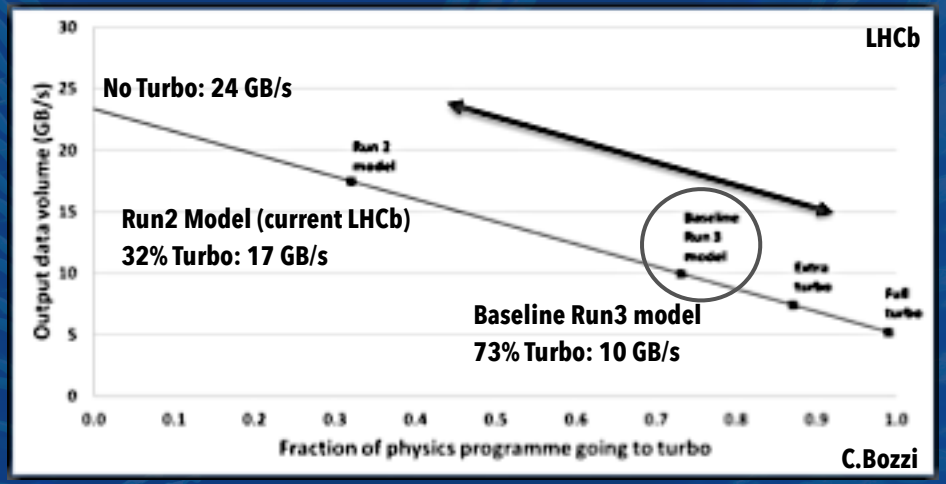
Future Challenges

New Workflows and New Tools

- CTA instead of CASTOR
- T1s data export
  - xrootd, https or gridftp?
- CMS: Phedex -> Rucio migration

Experiments New Data Rates

- LHCb 10x Bandwidth increase
  - 10GB/s
- ALICE: 90+GB/s ?
- ATLAS and CMS 1.5x increase?



Preparing Run3 Common Data Challenge

- Starting the Coordination from Q1 2020
- Testing Each Part of the Workflow
  - from Px to EOS to CTA to Export

Thanks for the Attention!

Questions?

Accélérateur de science