New EOS flavours, Inspired by ALICE

Testimonials:

“This trio of releases comes in the most delicious flavours - our favorite one is definitely the Watermelon Wonderland! Thank you EOS!”

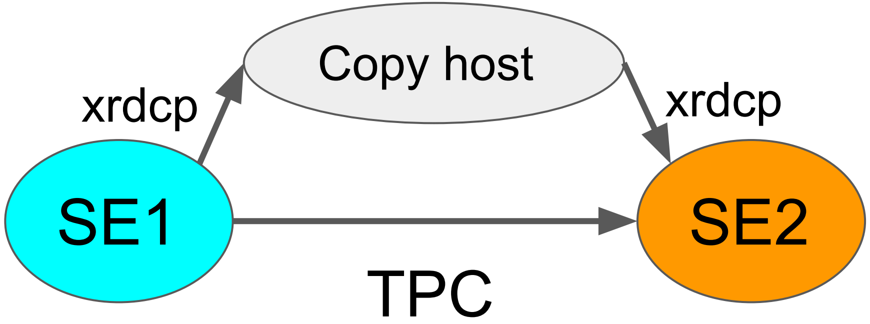
“At a mere zero euros, we feel that this is the perfect gift for any computing site, or stash it away and give it to your favourite sysadmin for Christmas!”

EOS in ALICE - present and future

L. Betev

ALICE use of storage in general

- On the Grid, ALICE uses exclusively ***xrootd*** protocol for all data write/read from ***local*** and ***remote*** storage
- No FTS - ***xrdcp*** and ***xrd3cp*** (now popularly known as TPC) to transfer data since beginning of times
- Initially, ALICE advisory was to install storage with vanilla xrootd management
    - Simplifies operation
    - No DB to worry about

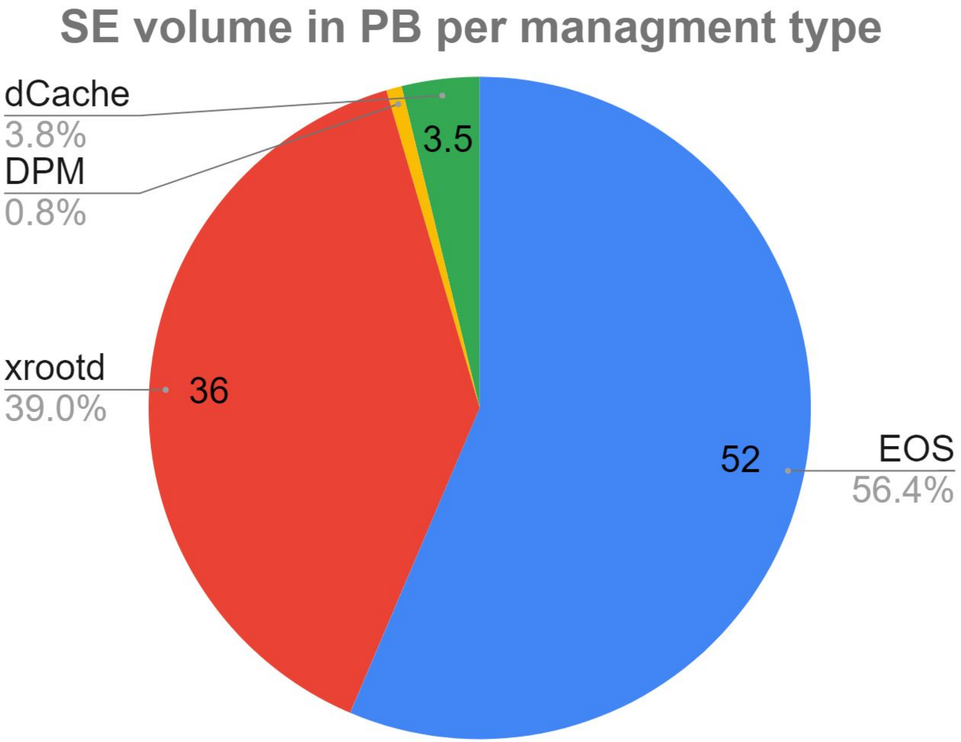


ALICE use of storage in general + EOS

- Since several years, we encourage sites to migrate to EOS
  - Especially for large chunks of new storage servers
- Clear advantages
  - Integrated admin tools for operation and debugging
  - Full support by developers and active user forum
  - Long-term strategic support and collaborative options
  - Cheapest hardware (JBODs with no HW RAID)
  - High-level data security by using erasure coding
  - No need for complicated and expensive cluster filesystems

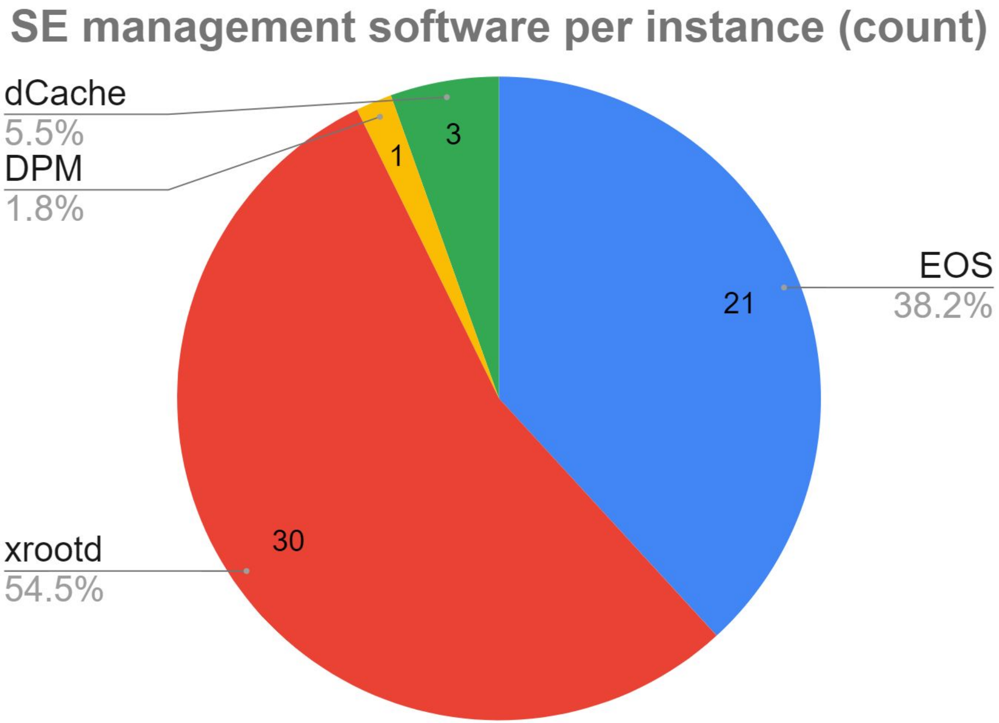
Storage today - volume management

- ~100PB of disk SEs
- Picture is different for tape instances, but we do not discuss these here



Storage today - instance management

- Largest count are still xrootd-managed instances
  - Tend to be smaller capacity SEs
  - Still easiest to install
- Individual storage behaviour does not depend on management software



ALICE data management policy

- All files on Grid storages anywhere in the world are annotated in the central catalogue
  - No exceptions, no private/group direct access to storage
  - No roles defined on the storage element, all accesses mapped to the only “ALICE” account
  - Token authentication, signed by central services (similar: Macaroons, Sci/WLCG Tokens)
- All of the above simplifies SE operation
  - Quotas and ACLs are managed centrally
  - Data transfers are managed centrally
  - Goal-minimize load on site admins, SEs are like block devices for the VO

Important storage metrics and consequences

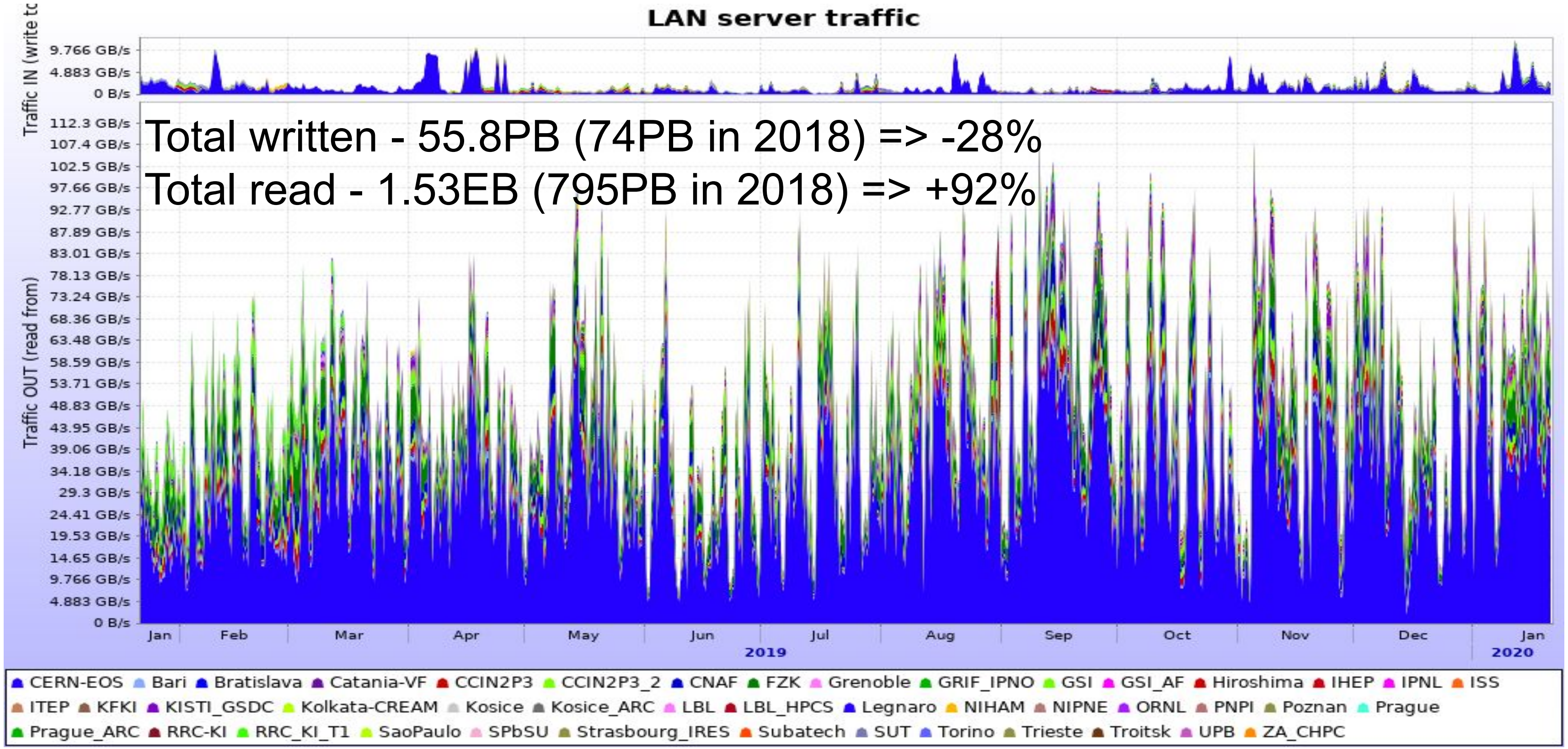
- Unrestricted and fast local read access to data
  - Read/write ratio = ***15/1*** *(!)* (was 11/1 a year ago)
- Storage should not be a bottleneck
  - In terms of client access rate and throughput
  - Jobs go to data - remote WAN reading <5%

=> Most important is to have the site network fabric/WNs and SE correctly paired in terms of performance

- 1. deploy cheap and reliable storage, 2. invest in network fabric

=> EOS provides the answer to the first requirement

Storage access - always increasing!



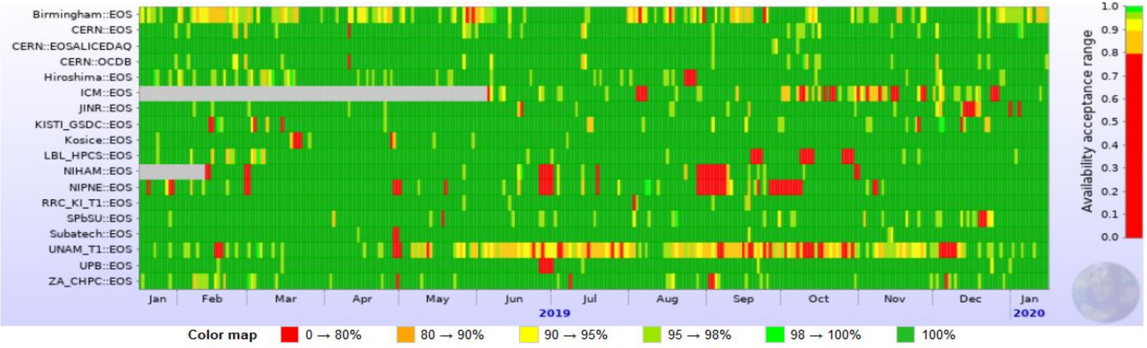
Total written - 55.8PB (74PB in 2018) => -28%

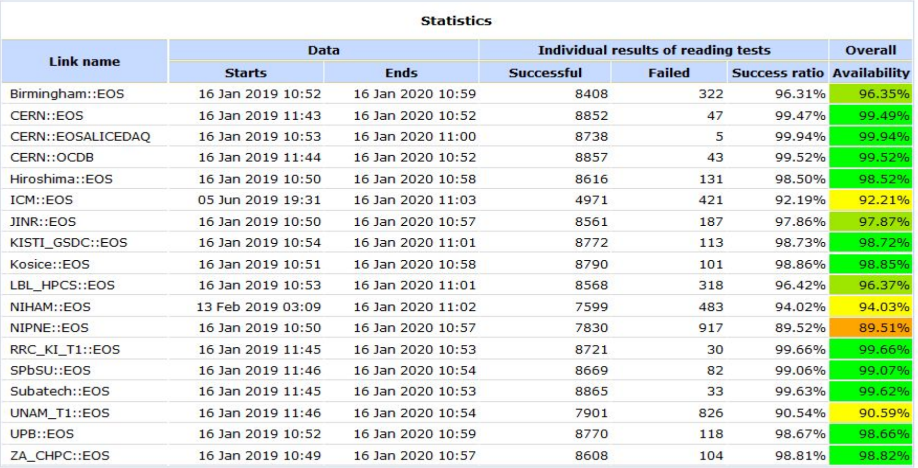
Total read - 1.53EB (795PB in 2018) => +92%

Availability of the storage

- Minimizing remote reading and absence of replicas => individual storage availability is critical for operation

- Target availability for SE >95%





Statistics

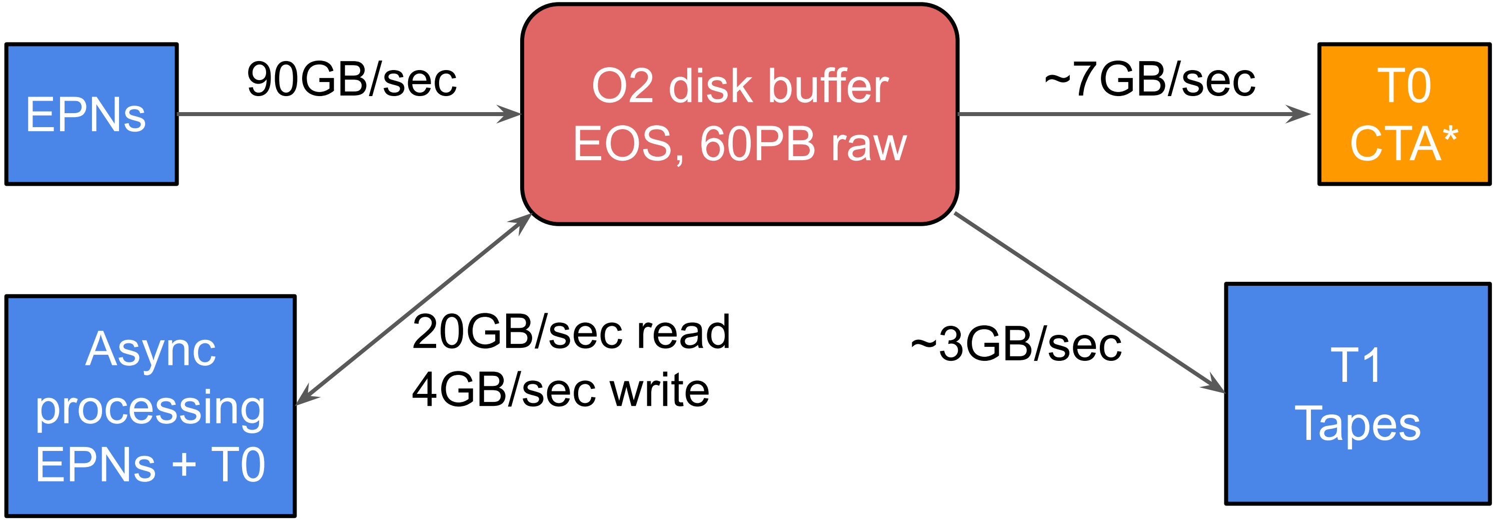
Link name	Data		Individual results of reading tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
Birmingham::EOS	16 Jan 2019 10:52	16 Jan 2020 10:59	8408	322	96.31%	96.35%
CERN::EOS	16 Jan 2019 11:43	16 Jan 2020 10:52	8852	47	99.47%	99.49%
CERN::EOSALICEDAQ	16 Jan 2019 10:53	16 Jan 2020 11:00	8738	5	99.94%	99.94%
CERN::OCDB	16 Jan 2019 11:44	16 Jan 2020 10:52	8857	43	99.52%	99.52%
Hiroshima::EOS	16 Jan 2019 10:50	16 Jan 2020 10:58	8616	131	98.50%	98.52%
ICM::EOS	05 Jun 2019 19:31	16 Jan 2020 11:03	4971	421	92.19%	92.21%
JINR::EOS	16 Jan 2019 10:50	16 Jan 2020 10:57	8561	187	97.86%	97.87%
KISTI_GSDC::EOS	16 Jan 2019 10:54	16 Jan 2020 11:01	8772	113	98.73%	98.72%
Kosice::EOS	16 Jan 2019 10:51	16 Jan 2020 10:58	8790	101	98.86%	98.85%
LBL_HPCS::EOS	16 Jan 2019 10:53	16 Jan 2020 11:01	8568	318	96.42%	96.37%
NIHAM::EOS	13 Feb 2019 03:09	16 Jan 2020 11:02	7599	483	94.02%	94.03%
NIPNE::EOS	16 Jan 2019 10:50	16 Jan 2020 10:57	7830	917	89.52%	89.51%
RRC_KI_T1::EOS	16 Jan 2019 11:45	16 Jan 2020 10:53	8721	30	99.66%	99.66%
SPbSU::EOS	16 Jan 2019 11:46	16 Jan 2020 10:54	8669	82	99.06%	99.07%
Subatech::EOS	16 Jan 2019 11:45	16 Jan 2020 10:53	8865	33	99.63%	99.62%
UNAM_T1::EOS	16 Jan 2019 11:46	16 Jan 2020 10:54	7901	826	90.54%	90.59%
UPB::EOS	16 Jan 2019 10:52	16 Jan 2020 10:59	8770	118	98.67%	98.66%
ZA_CHPC::EOS	16 Jan 2019 10:49	16 Jan 2020 10:57	8608	104	98.81%	98.82%

Other critical use cases - Conditions data

- **Run1+Run2** - set of ROOT files distributed over several Grid SEs
  - Used for offline tasks (reco/MC/analysis)
  - Primary source was CERN::OCDB EOS instance with multiple internal replicas
  - Backup in CVMFS
- **Run3** - combination of online stream for synchronous (realtime) processing + ROOT/other objects for asynchronous (offline) processing
  - New REST API to access conditions data, HTTP access to storage is explored
  - All objects in CERN::OCDB EOS instance
  - Will see order of magnitude increase of data volume (not critical) and access frequency
  - Tested and confident that the schema will work

Other critical use cases - data buffer for O2 facility

- 60PB raw capacity, RS erasure coded (level of security to be defined)
- Based on cheap JBODs, SATA drives, EOS managed



For details - see talk of M. Lamanna

*CTA = CERN Tape Archive

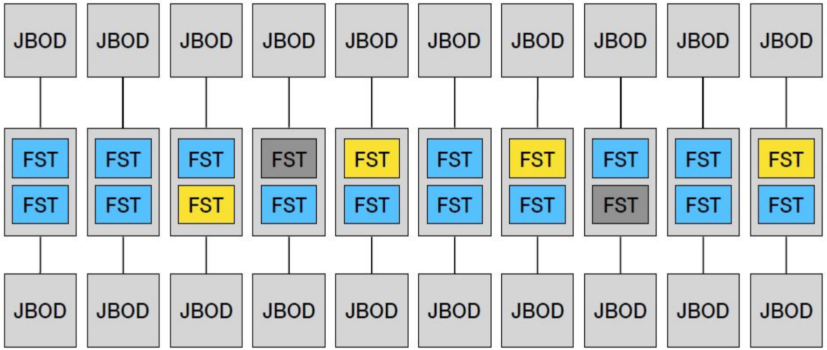
Grid evolution

- **ALICE Computing Model for Run3 -** continues to track the 'flat funding' resources growth scenario (+10-15%/year)
- Growing interest in SE consolidation
  - Mostly in terms of sharing of responsibilities/experience for operation
  - Country borders still a thing - common investment in SEs is not happening soon
  - … even between sites of the same country
  - Not exactly a 'Data lake' scenario, yet
- Having a common SE management system is a compulsory first step
- Even more sparse replica scenario - RAW data will not have a second copy
  - Smart storage solutions with high data protection
  - Temporary unavailability - better tolerated if data is secure

Yet another EOS application - diskless custodial SE

- Project of the KISTI T1 centre (S. Korea) - replace the tapes with inexpensive, but secure disk storage
  - Simplify the operation of the T1 centre, reduce exposure to a shrinking tape market
- Storage designed around EOS with EC, inexpensive JBODs
- Extensive fit-for-purpose studies of selected HW

EOS RAIN6 (14,4)


JBOD JBOD JBOD JBOD JBOD JBOD JBOD JBOD JBOD JBOD
FST FST FST FST FST FST FST FST FST FST
FST FST FST FST FST FST FST FST FST FST
JBOD JBOD JBOD JBOD JBOD JBOD JBOD JBOD JBOD JBOD

- RS(14,4) = 77.7% of RAW capacity
- $5x10^{-9}$ theoretical file loss probability
- Easy to upgrades nodes without degrading performance
- Further security and data integrity methods will be applied
- Power consumption 1.75W/TB (tape 0.5W/TB)

See talk of Sang-Un

General takeaway for ALICE

- Disk storage is and will continue to be one of the integral assets of distributed computing
- Data volumes increase in line with the expected yearly Grid growth
- In our experience - the storage load is not linear with increase of data volume
  - Storage management solutions must be future-protected in this respect
  - Computing models must also take this into account (local vs. remote access)
- Even less data replication
  - More pressure on storage to ‘never lose data’
  - Must learn how to live with temporary data unavailability (longer maintenance/interruptions of service), but know that the data is safe
  - Rely on storage solution (see erasure coding) to protect data
- ALICE upgrade will add a few more SE-dependent projects (CCDB, large disk buffer, tape replacement solution
- Storage consolidation requires uniformity of storage management solutions

Acknowledgements

- To all experts at the computing centres providing resources for ALICE - thank you for your support and dedication in the past 10 years of operation!
  - More will be asked of you in the next years
- To the CERN storage group and EOS experts - thank you for the storage and for being behind it 100%!
  - See above sub-bullet :-)