Joint Institute for Nuclear Research

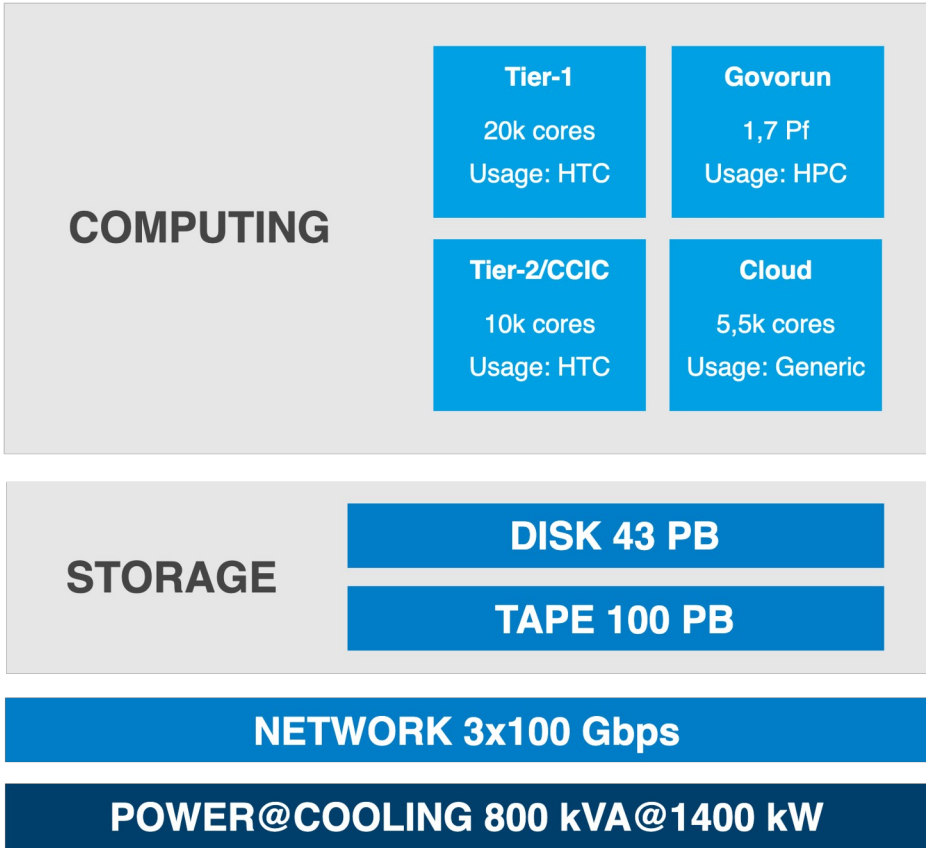
Shared EOS instance at JINR

Nikita Balashov
On behalf of the JINR storage team

8th EOS Workshop
15 March 2024

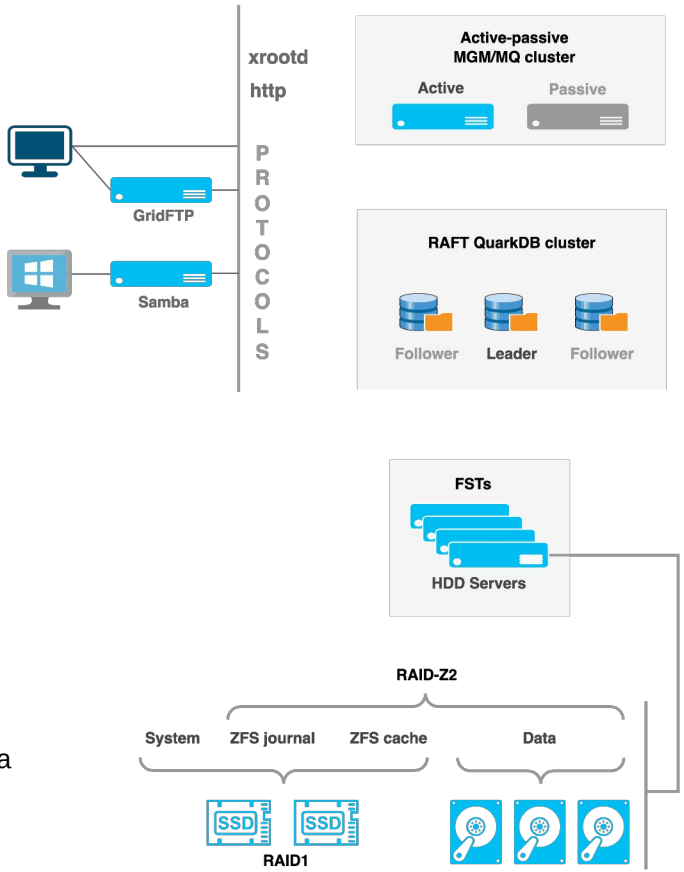
Multifunctional Information and Computing Complex

- Experimental data
  - dCache
  - EOS
  - CephFS
- HPC
  - Lustre
- Cloud computing (Virtualization)
  - Ceph RBD
- Miscellaneous (CI artifacts, backups, etc)
  - Ceph S3



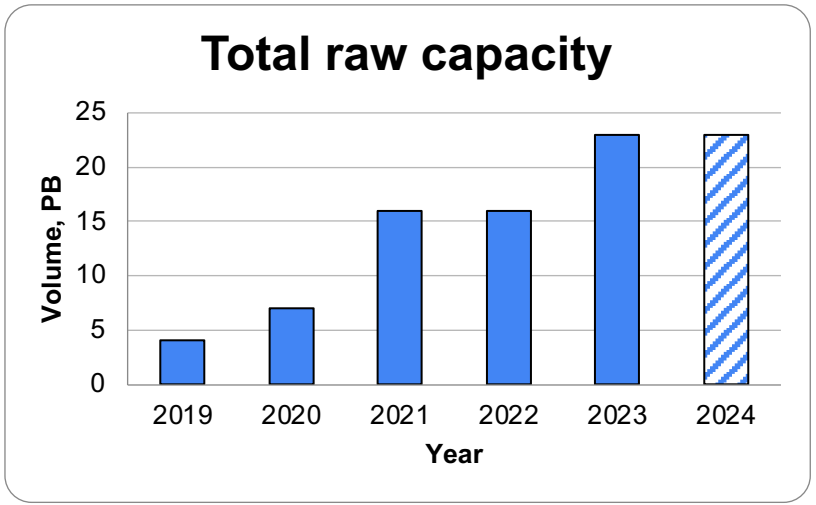
EOS Configuration Overview

- In operation since 2019
- Single instance shared among multiple projects
- Fault tolerance
  - Active-passive HA for MGM/MQ
  - RAFT HA for QuarkDB
  - Replica layout
  - **ZFS** on FSTs (RAID-Z2, 4 partitions per server)
- Software
  - Scientific Linux 7 everywhere
  - EOS/QDB 5.2.17
- Protocols:
  - **eos shell** and **xrdcp** are most widely used
  - **Fuse** mounts are available, but users are recommended not to use them to write data
  - **HTTPS/WebDAV** configured
  - **TPC** configured for both WebDAV and xrootd



Hardware and Usage

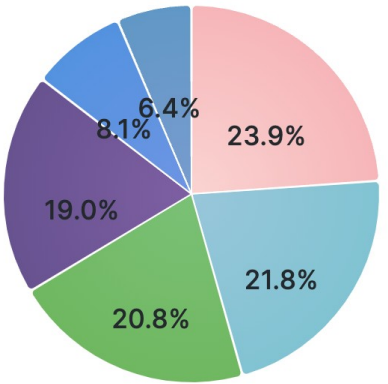
- ~23 PB raw capacity
- 92 FST servers (4 FSTs daemons per server, 1 per partition, 368 in total)
- 1832 drives
- 2x10 Gbps network connectivity
- Mostly Dell servers
- Evaluating other server vendors due to sanction restrictions



Usage Stats

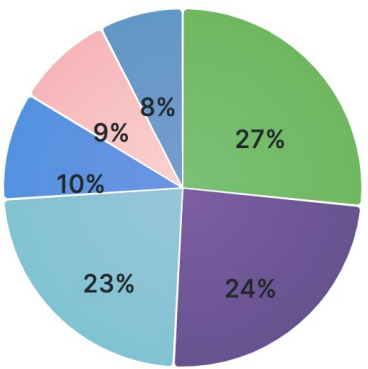
- ~8 PB of data stored in over 50 million files
- ~14 PB of physical space occupied
- Top consumers are neutrino and local experiments on NICA
- Experiments are pretty demanding, e.g. JUNO plans on ~1 PB per year increase

Space Used



	Value
/eos/nica/mpd/	1.8 PB
Other	1.7 PB
/eos/baikalgvd/	1.6 PB
/eos/nica/bmn/	1.5 PB
/eos/dayabay/	620.5 TB
/eos/juno/	487.8 TB

Files Stored



	Value
/eos/baikalgvd/	13.6 Mil
/eos/juno/	12.3 Mil
Other	11.8 Mil
/eos/dayabay/	4.93 Mil
/eos/star/	4.44 Mil
/eos/flnp-admin/	3.82 Mil

Data Consistency Issues

- We had numerous data consistency problems
  - Incorrect checksums
  - Data not visible to users, but present on FSTs
  - Corrupted files
- Hard to detect, in most cases discovered when users try to access the data
- In some cases the data was not recovered
- We never found out the exact reason of the problems
- Hard to define the scale, but one of experiments reported 0.2 % data corruption (out of ~400k files) of “passive” data in half a year
- Some projects consider developing external data consistency checkers
- Community of the forum is responsive to the problems and we usually get the help

QRAIN Migration

- Massive data loss during Replica→QRAIN data migration
  - JINR network to blame, but handling data consistency by EOS may need to be revised
  - Basically all the files larger than 500 MB were lost or abandoned due to hardness of detection of damaged data
  - Documentation was not quite clear at the moment or some important points were missing
- Consequences
  - We had to reload lots of data from other data centers
  - Some of the data had no replicas in other DCs, but luckily it was MC and the data was generated again
  - Some experiments moved to other storages
- RAIN documentation is not clear

Summary and Plans

- Data consistency issues are making us worry, some projects moved to other storages
- Some projects had to move to other storages
- Despite the issues we don’t plan abandoning EOS
- Move to new OS due to SL 7 EOL (presumably AlmaLinux 9)
- Consider moving large data consumers to dedicated EOS instances for reliability

P.S. Documentation needs improvement!

Thanks!

Nikita Balashov
balashov@jinr.ru