DOMA/QoS Workshop
The ALICE view

Nikola Hardi, for the ALICE collaboration
nikola.hardi@cern.ch

07-02-2020

Outline

1. ALICE Grid storage QoS policy
2. ALICE QoS orchestration agents
3. DOMA/QoS whitepaper feedback

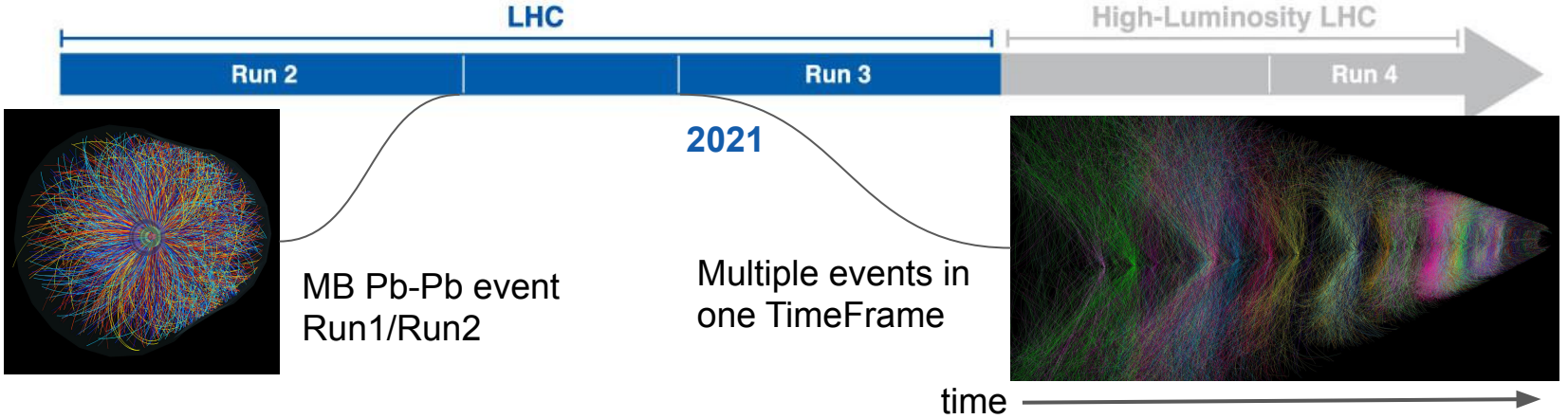
ALICE Grid storage QoS policy

ALICE data storage in Runs 1 & 2

- Current data storage
  - ~70 PB of data on tape (raw data, 2 replicas)
  - ~70 PB of data on disk (many files have 2 replicas)
- Replication policy
  - Files usually have two replicas
  - Some files have more replicas (conditions data, 5 replicas)
  - Replicas are written directly by jobs, one replica in local SE and other replicas in close SE
- Jobs are almost always dispatched to sites that already have the data
- 95% of read access is on LAN - 1.5 XB
- The read/write ratio is nowadays 15:1

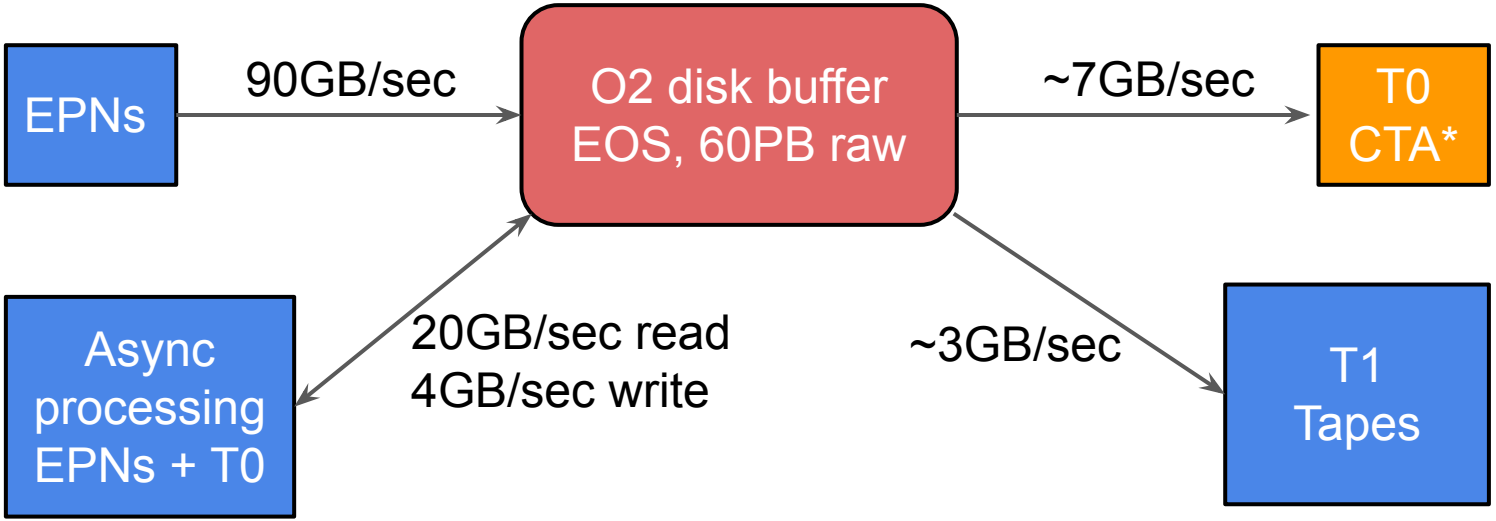
ALICE upgrades for Run 3

- Major detector, readout and software upgrade during LS2
- In Run 3 ALICE experiment will
  - Observe and record more collisions: 100x more events, higher luminosity
  - Continuous readout in time frames: no trigger possible due to type of physics studies
  - Highly efficient (new) compression algorithms allow ALICE to fit into the standard ‘flat funding’ computing growth scenario



O2 Disk buffer

- 60PB raw capacity (some degree of safety to be included)
- Based on cheap JBODs, SATA drives, managed through EOS



*CTA = CERN Tape Archive

Basic data rates in Run3

Description	Data volume, rates and provenance
Compressed Time Frames (CTFs) / year	~50 PB (max 100GB/sec) from O2 facility, to CERN IT on dedicated fiber
CTF transfer from T0 to T1s	⅓ of CTF (~3 GB/s) from O2 disk buffer
AODs from CTF processing to disk and archive	2x10PB/year, one replica (LAN) on T0/T1s
AODs from CTF and MC processing at T0/T1s	partial replica, ~1GB/s aggregate from T0/T1s to T2s
AODs from MC generation at various tiers	At today's level + 10-15%, from T2s to T2s/T1s
AODs to the AFs (~5PB per AF)	10 PB @ 100 Gb/s in 12 months
WAN traffic from analysis activities	At today's level + 10-15%, from all centres to all centres

ALICE QoS orchestration agents

Data Storage in the AliEn Grid

- Files represented by logical filenames (LFNs)
- Replicas represented by physical filenames (PFNs)
- Users interact with LFNs only, replicas and PFNs are managed automatically
- Physics data stored as ROOT files, accessed directly (streamed)
- Tape storage exposed as a separate storage element
- AliEn Grid middleware is to be superseded by JAliEn

ALICE Grid operations - storage

- Remote data access: read 5% / write 20% (2019, no data taking)
- This includes grid operations: recovery, manual data movement
- Manual transfers and recovery, using TPC when possible
- Storage element failures - data corruption
  - Hardware failure and human error
  - Replicas in different data centers help
  - Recovery from spare replica
- Replicas in separate SEs are more valuable than replicas stored together
- Single replica CTFs in Run 3 increase importance of QoS per site
  - Expected to get some results while data is still both on disk and tape

DOMA/QoS whitepaper feedback

QoS classes and attributes

- The ALICE QoS model is compatible with QoS classes (disk and tape)
- Looking forward to tapeless custodial storage (KISTI / EOS)
- Directly exposed storage elements are preferred
  - Geographic location
  - Data center
  - Connectivity to other sites
- Very fast storage would saturate site network fabric,

  however SSDs are appreciated on worker nodes
-

QoS and pledges

- There are many hidden costs
  - The “infant mortality” curve
  - Testing hardware before commissioning
  - Different ways to aggregate storage (RAID variants, EOS, dCache)
  - Training and operations
- If pledges are defined in raw disk space, can experiments choose how that storage will be combined into a storage element?

Other comments

- Make an example QoS class and policy schema for purposes of discussion
- What QoS values can be tested and how?
- Design guidelines for relaxing over-constrained QoS policies
- Replica geolocation dispersity as QoS class attribute for data lakes

Thank you!

Questions?

DOMA/QoS Workshop
The ALICE view

Nikola Hardi, for the ALICE collaboration
nikola.hardi@cern.ch

07-02-2020