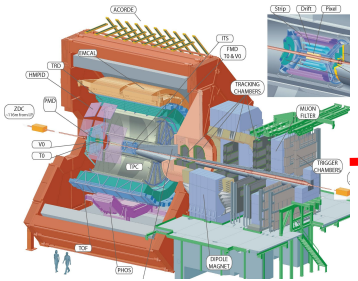
LHC OPN/One Workshop
ALICE feedback

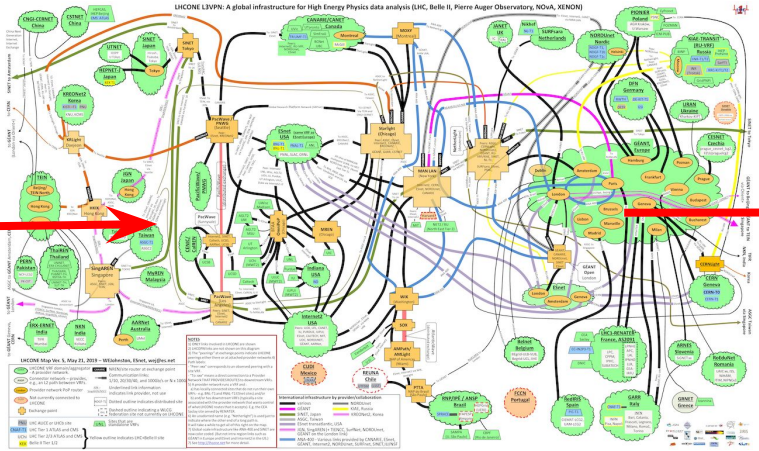
Nikola Hardi, for the ALICE collaboration
nikola.hardi@cern.ch

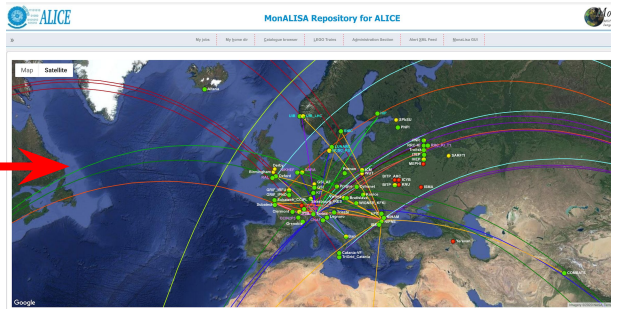
13-01-2019

Introduction / general feedback

The networking requirements for the ALICE experiment were ***fully satisfied*** during LHC Run 2.







Outline

1. The ALICE Computing Model in **Run 2**
2. The ALICE Computing Model in **Run 3**
3. LHC OPN / LHC One feedback

Computing Model in Run 2 from network perspective

Data processing

- Grid site local file access (95%), remote (5%)
  - Remote access due to local SE issues, usually temporary
- Multiple replicas sorted topologically: apps first access local replica, then the next closest
  - Sorting by network topology, availability, network quality, geo-location and other metrics
- Jobs are dispatched to the Grid sites that already have the data *
  - Minimizes WAN traffic and RTT efficiency penalty
- Storing multiple replicas
  - One replica is written to the local storage element
  - The other replicas are written to the remote (but close) storage elements
  - Remote writes might go through LHC OPN / LHC One

* the constraint is relaxed if needed after a predefined period to speed up the processing

Data rates in Run 2

- Raw, reconstructed and AOD data access
  - 95% LAN reads
  - 5% WAN reads (potentially LHC OPN)
- Monte Carlo (20% remote writes that would go through LHC OPN/One)
- File recovery / storage decommissioning / other data manipulation : ~2 PB/y *

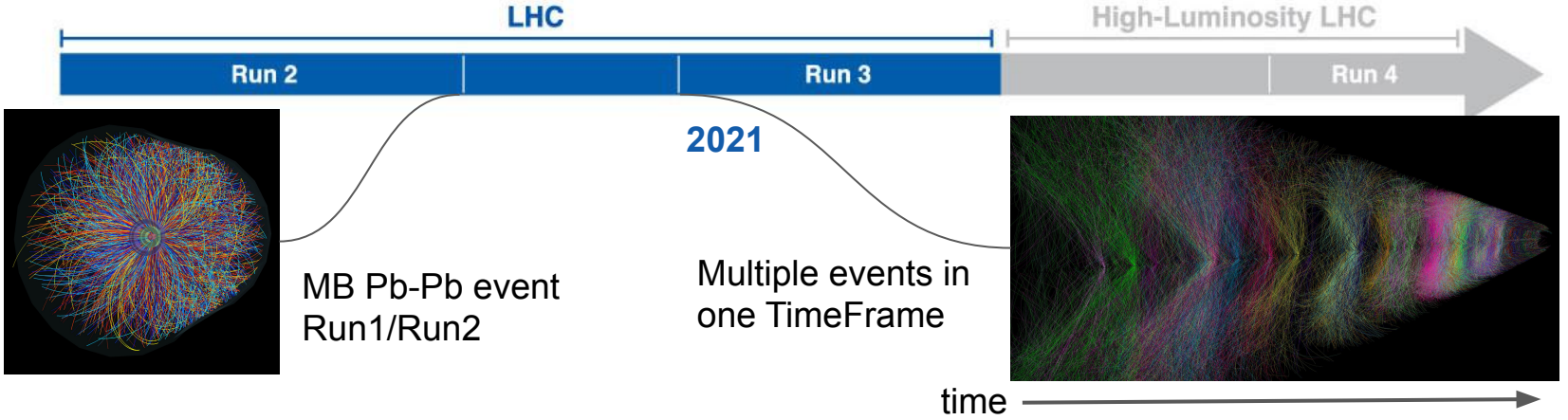
Description	**Transferred data**
Total data stored (disk / tape)	70 PB / 70 PB
Total traffic on storage (write/read) (1y)	65 PB / 1.5 XB
- site ***LAN*** traffic (write / read) (1y)	53 PB / 1.4 XB
- site ***WAN*** traffic (write / read) (1y)	11 PB / 74 PB (5% of total)

* one year of Grid operations during LS2 - no raw data transfers

Computing Model in Run 3 from network perspective

ALICE upgrades for Run 3

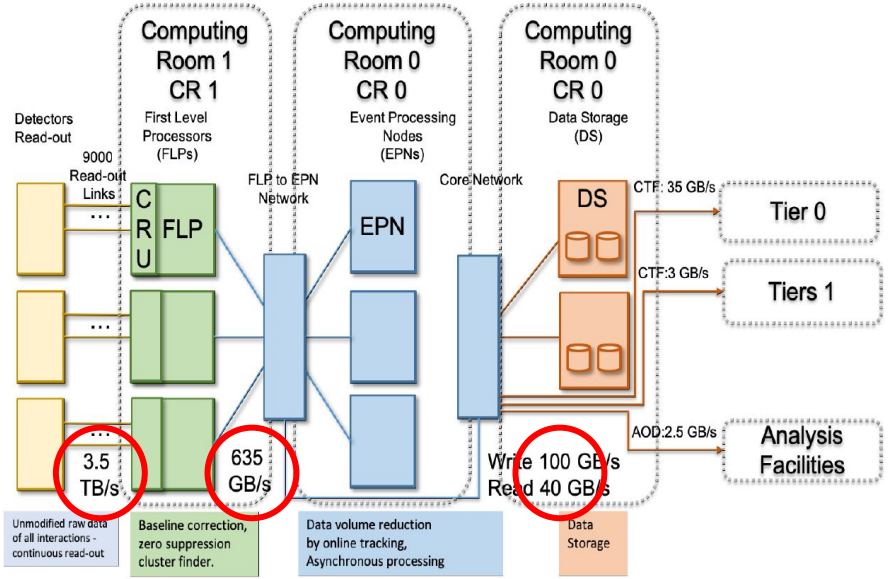
- Major detector, readout and software upgrade during LS2
- In Run 3 ALICE experiment will
  - Observe and record more collisions: 100x more events, higher luminosity
  - Continuous readout in time frames: no trigger possible due to type of physics studies
  - Highly efficient (new) compression algorithms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



Computing system upgrade

- A new framework will serve both online and offline functions
- The detector data will go through a pipeline of processing elements
  - **FLP** - First Level Processing nodes - collect data from individual detectors, produce sub-timeframes
  - **EPN** - Event Processing Nodes - Each EPN in turn receives sub-timeframes from all FLPs and produces the Compressed Timeframe (CTF), which is immutable.
  - **Disk Storage** - local disk buffer (60 PB) - stores the CTFs for subsequent dispatching to archive and for asynchronous data processing
- FLPs and EPNs are located in the O2 facility close to the detector@P2
- ALICE largest data taking period is 24 days of Pb-Pb / year, produces 90% of the CTFs and subsequent derived data
- CTFs are transferred to
  - ⅔ to T0 (CERN) archive storage
  - ⅓ to Tier1s archive storage: CNAF, NDGF, CCIN2P3, RAL, KISTI, RRC-KI, NL SARA, GridKA.

Hardware and software architecture



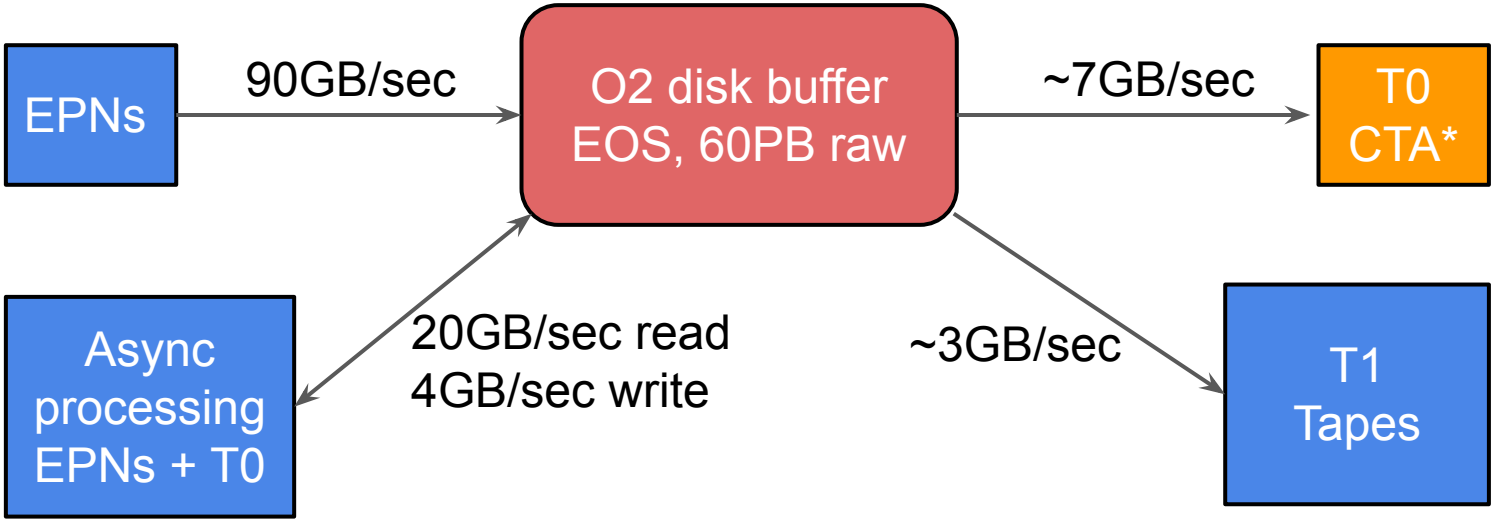
Task name	CPU Time [s]	GPU Time [s]
TPC sector track finding	706	11
TPC track merging	40	2
TPC track fit	300	6
TPC looping track following	150	6
TPC data track-based compression	100	2
Sum	1296	27
ITS clustering	10	
TPC-ITS track matching	1	
Global track matching to TRD	1	
Global track matching to TOF	1	
ITS tracking	10	
ITS tracklet vertexer (seeding)	1	
ITS (MFT) data compression	3	
TPC data entropy compression	35	
TPC gain calibration	10	
TPC distortions calibration with residuals	20	
Sum	92	
Total	1388	

Data compression factor ~35-40x

Emphasis on GPU algorithms for TPC reco: substantial reduction of overall processing time

Disk buffer

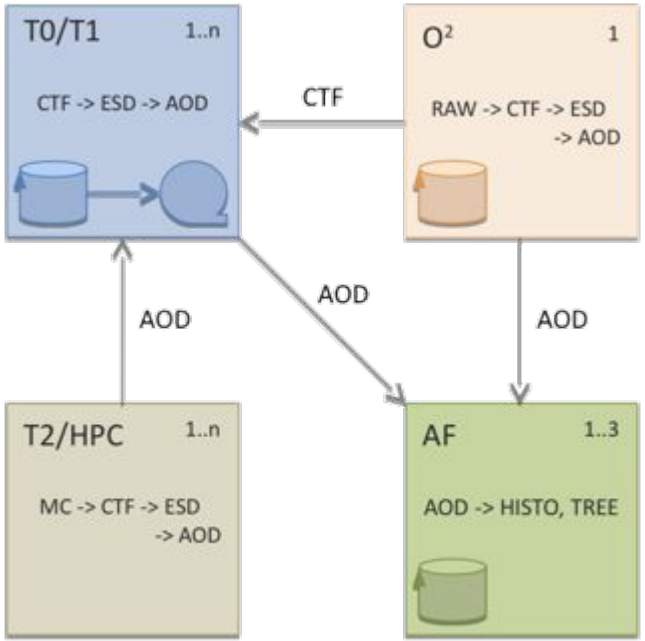
- 60PB raw capacity (some degree of safety to be included)
- Based on cheap JBODs, SATA drives, managed through EOS



*CTA = CERN Tape Archive

Computing Model in Run 3 - wider perspective

- CTFs are transformed into AODs in T0/T1s and the O2 facility
- Monte Carlo AOD productions on T2s and HPCs
- Specialized Analysis Facilities are used for fast analysis done on fraction of AODs
  - AFs will receive the AODs from T0/T1s/T2s
- Bulk of organized analysis will be done on the Grid
  - More planning, regularization of this process



LHC OPN / LHC One Feedback

ALICE bandwidth requirements in Run 3

- We expect network bandwidth to increase (the canonical 10-15% per year)
  - This will adequately cover the expected increase in CPU/storage and associated data processing
- LHCOPN transfers between T0 and T1s will be moderately higher for the period immediately following Pb-Pb data taking
  - Expected ~3GB/s for 2-3 months
- LHCOne community (additional replica writing / occasional remote reading from site WNs)
  - Similar to the current model for analysis at the standard computing centres
  - The Analysis Facility case: part of AODs produced anywhere on the Grid are written to the AF for rapid analysis and software tuning - still under discussion

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

Features we wished for in Run 2

- IPv6 is not critical, but we can make use of it
  - IPv4 is still required to be available alongside IPv6 until all/majority is on IPv6
  - ~60% SEs can already serve content on IPv6
- Bandwidth and network reliability fully compatible with our model and use
  - ALICE Storage Elements served more than 75 PB through WAN
- How much of this data was routed through LHCOPN and LHCOne?
  - Would be good to know, but we have no handle on this

Features we would like to have in Run 3

- Reserved bandwidth and temporary larger allocation of bandwidth (L3P2P)
- LHCOne is widely adopted - ***make it mandatory*** and ensure the rules are applied
  - Symmetric routing - the same IP route for incoming and outcoming traffic to/from a remote site
  - Bypass slow firewalls for storage elements and worker nodes (L3VPN)
- LHCOne network topology to be predefined and published
  - Currently we discover it ourselves
  - The LHCOne traffic optimisation affects the WAN portion of the ALICE total (WAN+LAN) network traffic

Networking main concerns, in order of importance

- From Edoardo’s list, our priority
  - Bandwidth
  - Availability and reliability
  - Monitoring

L3VPN

- Connected sites are listed in twiki
  - Where can we check if ALICE traffic is routed as expected?
- With IPv6 on all sites - will WN IP addresses be made public?
  - To avoid IPv4/NAT gateway box overloading

L3P2P

- This service is interesting to us
  - We don’t have practical experience with it yet
- The documentation isn’t very precise or is missing
  - I had difficulty understanding how to use this service
- Any future development planned?
- Who and how can request and/or open an L3P2P link?

Summary

- ALICE is happy with LHC OPN/One and in general with the network performance during Run 2 - ***always one step ahead of/above the needs***
- Our computing model favors local data access
  - WAN access for file replication and in case of issues with local storage
  - Run 3 model will continue using the same principles
- File transfers (data recovery and storage rebalancing) use will continue at the current level
- T0 to T1s data transfer of Pb-Pb data - higher LHCOPN use for 2-3 months/year
- More data from the experiment, but no general increase on the pressure for LHC networking
- Bandwidth allocation / L3P2P service is interesting to us
  - Especially to cover the AF case

Thank you!

Questions?

LHC OPN/One Workshop
ALICE feedback

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

13-01-2019