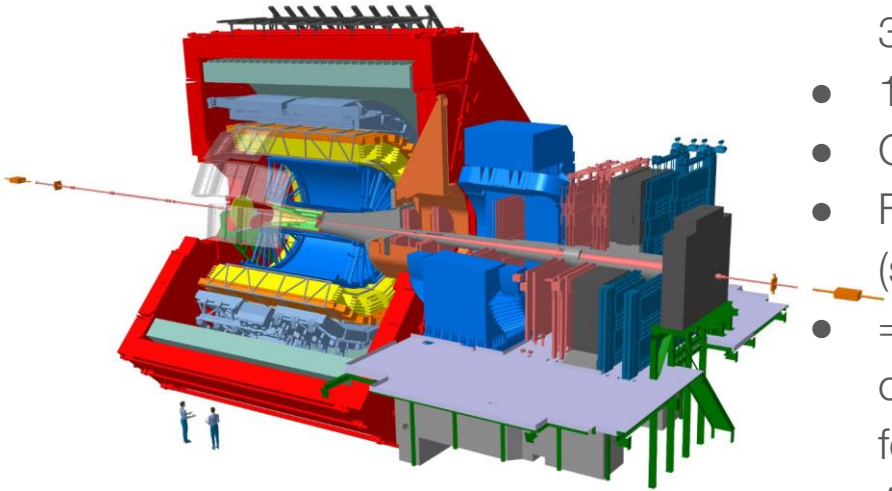
ALICE

WLCG Workshop, Lancaster, Nov 7, 2022

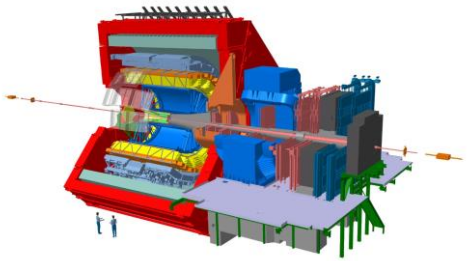
L. Betev, M. Litmaath

ALICE upgrade general



- p-p and HI physics
- 10x integrated luminosity L~10nb^{-1} (B=0.5T) + 3nb^{-1} (B =0.2T)
- 100x event rate of Run 1/2
- Continuous readout
- Focus on data compression and real time (synchronous) data reconstruction
- => Reasonable rates and data volumes after compression to storage and secondary data formats
- Adherence to ‘flat budget’ resources funding for data processing and analysis

Data flow and data rates



RAW data 3.5TB/s

First Level Processors (FLP)

STF - data from single FLP

600GB/s

Event Processing Nodes (EPN)

Compressed Time Frame (CTF) 10ms

100GB/s

Data buffer (80PB)

CTFs 8GB/s

Custodial storage T0 ⅔ T1s ⅓

CTF 20GB/s

T0 Async. Processing, Analysis

Local storage

AOSs, MCAODs ~10%

Analysis Facilities AODs, MCAODs

T2s MC production, Analysis

Local storage

AOD - Analysis Object Data
MCAOD - AOD from MC

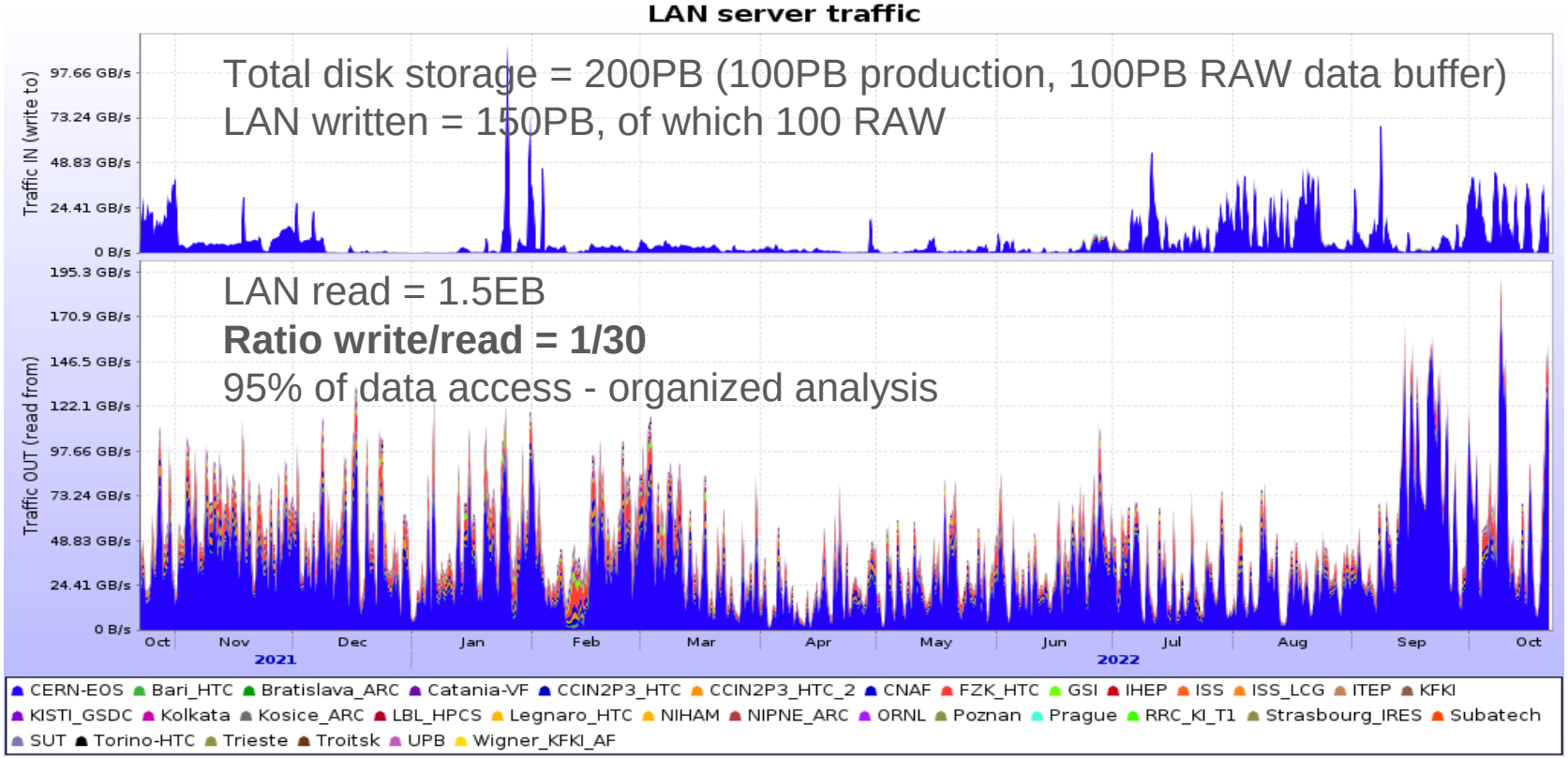
Data management model

- ALICE Run3/Run4 upgrade completed, first year of operation with new detector/readout/software
  - Some uncertainties to be expected in the numbers presented
- In general, the data management model and tools remain ~same as in Run2
  - Minimize the use of WAN
  - Most of the WAN use is RAW data transfers
  - Some details to follow on
    - data processing,
    - upgrade and data transfers
    - new elements of the computing system

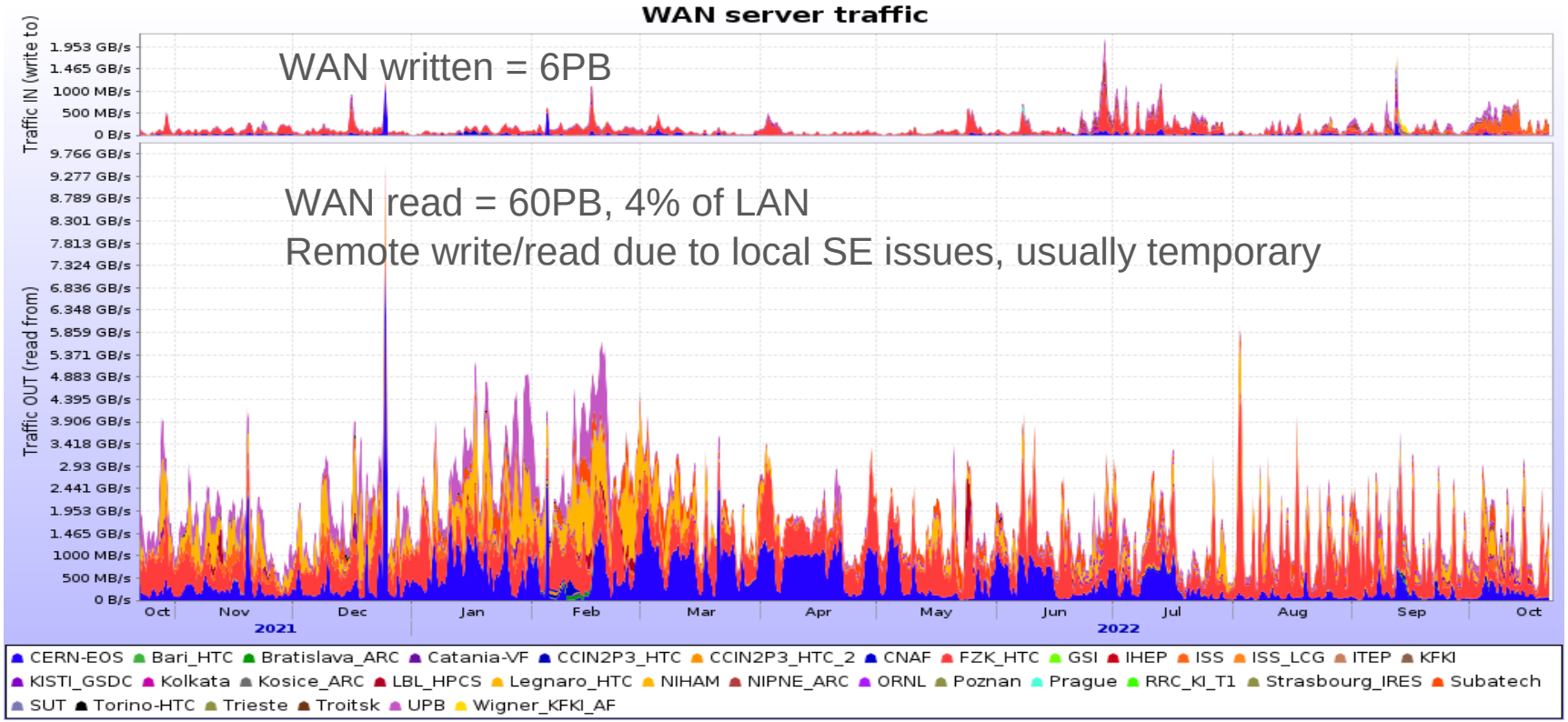
Data processing

- 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 LHCOPN / LHCONE

Data access - LAN



Data access - WAN (LHCONE/LHCOPN)

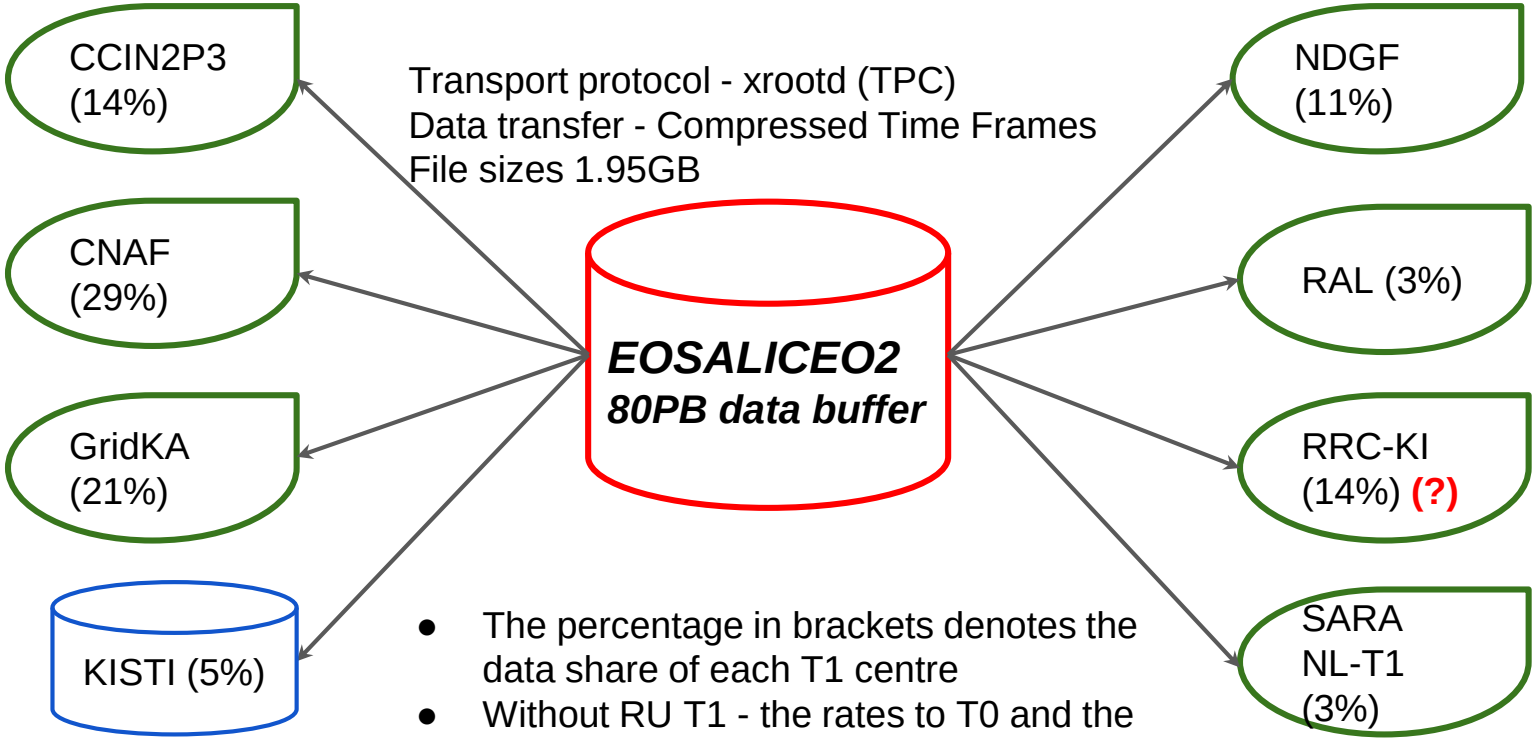


Summary of data access for past year

Description	Data volume PB
LAN write	150 (50 processing, 100 RAW data)
LAN read	1500 (95% organized analysis)
WAN write (LHCONE/LHCOPN)	6 (inaccessible local storage)
WAN read (LHCONE/LHCOPN)	60 PB (inaccessible local storage)
Other WAN transfers	20PB (file recovery/storage decommissioning, data replication)

- We expect these numbers to increase ~15% per year
- Strongly dependent on LHC programme, maximums reached after Pb-Pb data taking

Custodial data transfers over LHCOPN



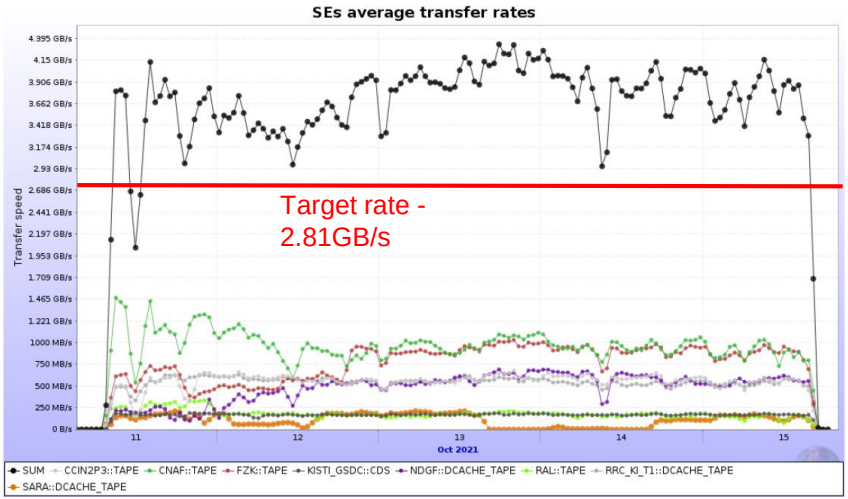
- The percentage in brackets denotes the data share of each T1 centre
- Without RU T1 - the rates to T0 and the other T1s will be proportionally increased

Data rates - from October 2021 data challenge

T1 Centre	Target rate GB/s	Achieved rate GB/s
CNAF	0.8	0.94 (116%)
IN2P3	0.4	0.54 (130%)
KISTI	0.15	0.16 (106%)
GridKA	0.6	0.76 (123%)
NDGF	0.3	0.47 (144%)
NL-T1	0.08	0.1 (122%)
RRC-KI	0.4	0.53 (128%)
RAL	0.08	0.17 (172%)

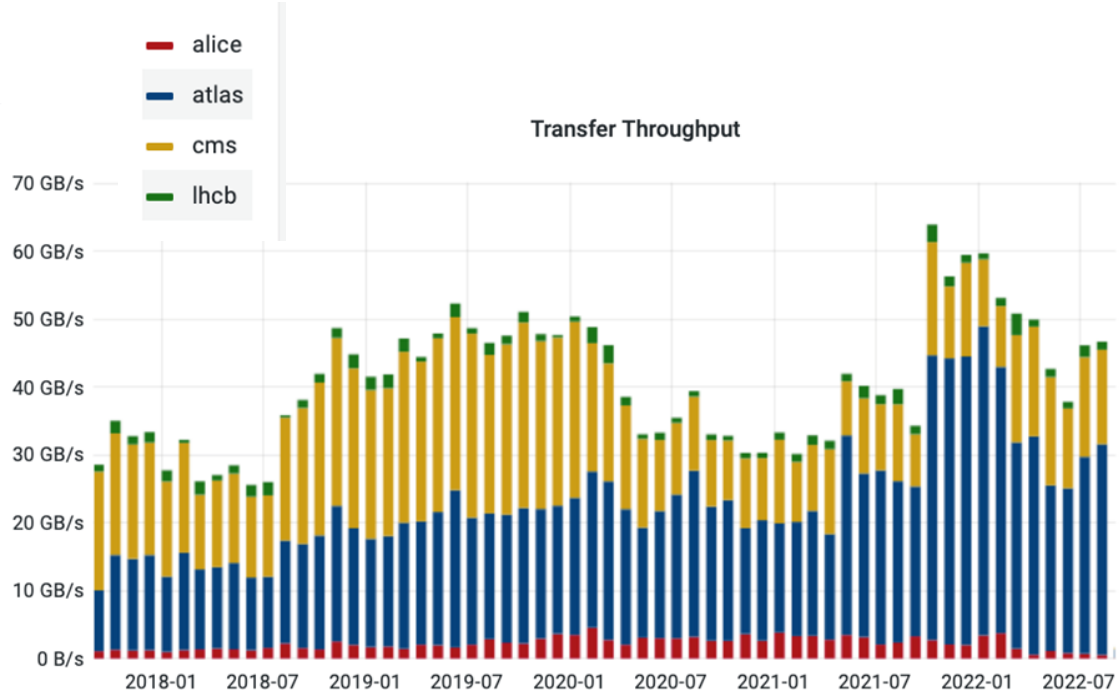
Sum 2.81GB/s

- Successful completion
- Channels tuned to slightly above the target rate, within reasonable limit
- The bulk of the bandwidth will be used after the Pb-Pb data taking period, for ~3 months
  - Since there is no Pb-Pb this year, we remain at the level of data challenges

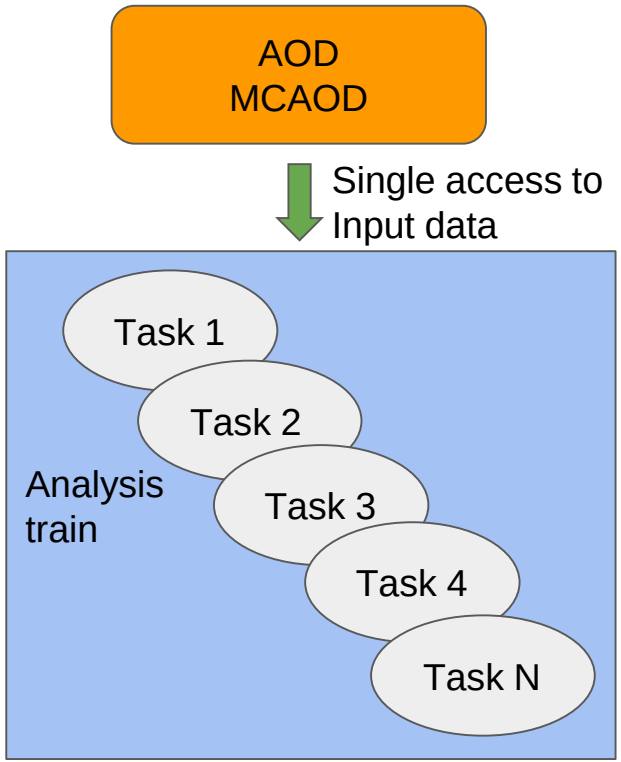


ALICE in the big picture - WLCG data transfers

- Includes RAW data distribution and other LHCONE/LHCOPN transfers



Analysis facilities (AFs)



- New element of the computing model
- Goals
  - Provide a location with comprehensive data samples from asynchronous and MC data processing at ~10% statistics
  - Fast tuning of analysis algorithms - once ready, run on full sample on the Grid
  - First data and low statistics analysis (if compatible)
- Incorporated in the Grid framework
- Sites tuned for fast I/O between storage and CPU
  - Approximate total size 6-8k cores, 10PB storage
  - ~15MB/s/core throughput
- As of today - 2 AFs (EU), possibly 1 more in US

AF data transfers

- Data is transferred to the AFs from T0/T1s/T2s
- Current AFs are co-located with T2s
- Data is transferred in blocks from the originating site
  - Can be anywhere in the world
  - Our tools take into account the network distance and copy the closest replica
- External network is not (yet) a limiting factor
  - May become one if the data turnover is greater than anticipated
- Bandwidth allocation / L3P2P service could be of interest to cover the AF use case
  - To speed up the transfers and responsiveness of the AF to analysis priorities

Multicore jobs and whole-node submission

- ALICE would like as many sites as possible to support multicore job slots
  - New ALICE framework exploits variable number of cores, from 1 to 8 in any combination
  - ~40% of the ALICE grid capacity is used through 8-core job slots
  - A campaign will soon be launched to ramp up that fraction
    - Starting with the biggest sites still on single-core job slots for ALICE

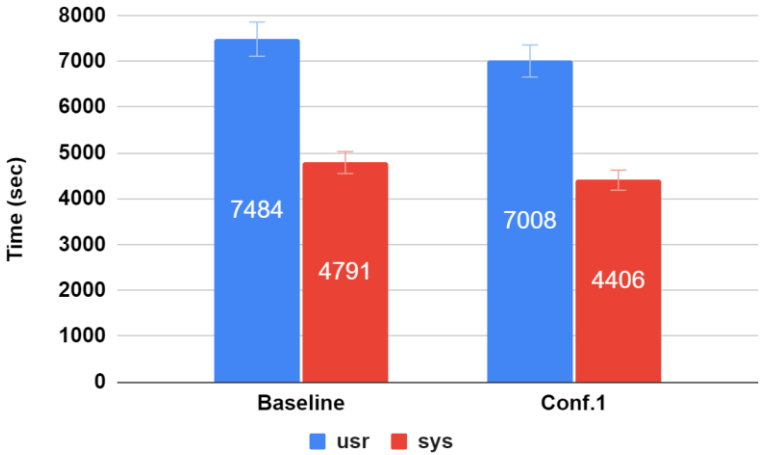
- More control of the cores would allow for more isolation of any payload type
  - Single and multicore of any flavour
  - Cgroups v2 functionality will be used on EL8 (where enabled) and EL9, taskset on EL7

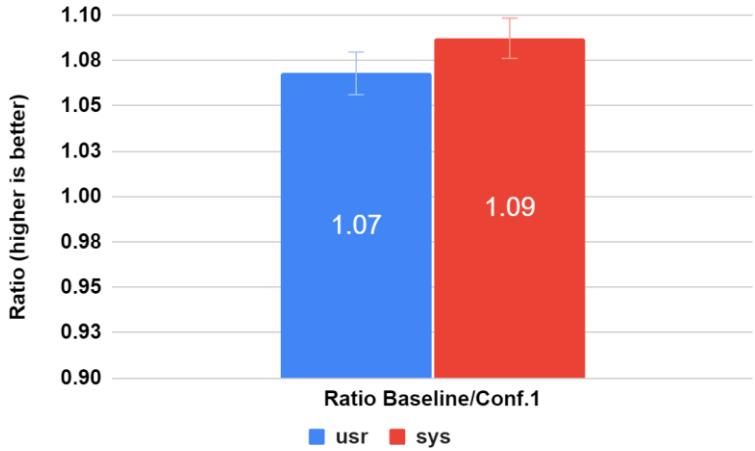
- Whole-node submission gives the most control plus further benefits
  - Appropriate NUMA-aware allocations
  - Node oversubscription - speculatively start more CPU-intensive tasks when IO-intensive ones leave cores idle and there still is enough memory and disk space

NUMA and cache configuration effects

- Standard MC job executed with no NUMA pinning and executed in most optimal configuration: same NUMA node, independent L1/L2 cache

~8% combined effect





HPC

- In production at LBNL: *Lawrencium*, *Cori* → *Perlmutter*
- MC and limited analysis ← thanks to reasonable bandwidth to local T2 storage
- Usage possible thanks to CVMFS and outgoing network access
- Whole-node job slots being used successfully
- GPUs could become interesting in the next years, ARM ditto

- In general the HPCs differ considerably from each other and the bar to integrate them in the Grid system of ALICE is medium to high

- Minimum set of requirements: CVMFS, outgoing network access, non-exotic local submission mechanism

Storage consolidation

- Storage would best be located at those sites which can operate it at scale
  - At least 1 PB today

- The storage MW should be from a project that has a long-term future ensured and helps site admins operate their storage at steadily increasing scales
  - Common advice, recipes etc. through community forums, HEPiX, etc.

- ALICE recommend EOS not only for those reasons, but also because:
  - RAIN works with cheap JBODs, obviating expensive RAID systems
  - EOS integrates well with MonALISA → convenient monitoring almost for free

Summary

- ALICE has started operations according to the new model designed for Run3 and Run4
- Our computing model favors local data access
- File transfers (data recovery and storage rebalancing) will continue at the current level
- T0 to T1s data transfer of Pb-Pb data - higher LHCOPN use for 2-3 months/year
- ALICE would welcome further storage consolidation
- Most jobs will be multicore and job isolation will be further improved through cgroups v2
- More HPCs could be integrated if minimum requirements are met