Data Taking for the Heavy-Ion Run 2018

C. Contescu¹, G. Lo Presti¹, and H. Rousseau¹

¹CERN IT-ST

December 20, 2018

Abstract

The Heavy-Ion Run 2018 was the last part of the LHC Run 2 programme. The experiments depend on the statistics from this relatively short (24 days) data taking phase, before the LS2 maintenance period starts. This report provides a summary of this data taking exercise from the point of view of the data recording and distribution in the CERN Data Centre.

Contents

1 Data flow and data rates 1
   1.1 Common readiness tests .............................................. 2
   1.2 EOS rates (initial phase) ............................................ 2
   1.3 CASTOR rates (initial phase) ....................................... 2

2 Mid-run status 5
   2.1 November 18-19 activity ............................................ 5
   2.2 CASTOR activity ....................................................... 7

3 End of run summary 11

1 Data flow and data rates

Throughout the LHC Run 2, the data flow of the ATLAS and CMS experiments has been as follows: data is sent from the experiments’ Data Acquisition systems to the corresponding EOS instances. Data are then distributed to batch for the reconstruction, via FTS for the Tier1 replication and to CASTOR for tape archival. These accesses are essentially controlled by the experiments themselves. In 2018, ALICE modified their data flows along the lines of the ones of ATLAS and CMS.

We expect LHC to operate with a duty cycle of 50%-60% (in the stable-beam phase this number will be probably higher). The goal for ALICE is to sustain about 8 GB/s during collisions. ATLAS and CMS goal is about 4 GB/s. LHCb will run at 1-2 GB/s but their utilisation of the network link will be typical in anti-coincidence with the other experiments (data are preferentially transferred off-collision). They are still using the old workflow (CASTOR receiving the data directly form the pit).
1.1 Common readiness tests

In September 2018, a series of tests were scheduled and coordinated among LHC experiments and IT groups. A common test was organised on September 13-15 plus several follow up investigations.

- Preparation of the common tests can be found under https://indico.cern.ch/event/751715/
- Data challenge wrap-up https://indico.cern.ch/event/758256/

1.2 EOS rates (initial phase)

This is the experiments’ activity (ALICE rates are in red, ATLAS in blue and CMS in yellow). ALICE is routinely exceeding 8 GB/s as required and CMS their 4 GB/s level. The data from ATLAS are comparable to CMS as expected. The on-off behaviour is expected and corresponding to the way their transfer system works (alternating buffering and transfer phases every 30 minutes. The LHC "page 1" shows the LHC fill in a 24-h window and it is compared to the LHC rates (zoom on the last 24 hours).

On Friday November 17th, the read activity on EOSALICEDAQ is also very relevant (peaks at 25 GB/s; average on 6 hours about 12 GB/s) with no signs of contentions between readers and writes. We observed the usual 9 GB/s peaks (input from the pit) in coincidence with a sustained 15 GB/s read activities.

1.3 CASTOR rates (initial phase)

The CASTOR rates are shown in here in a 24-hour window. One should note that in this phase ALICE hasn’t started the data transfer from EOS (L. Betev private communication: they are acquiring data and doing the Tier1 transfers only for the time being). Thus, the bandwidth of the tape system is not currently being used at full capacity.
Figure 2: LHC Page One

Figure 3: EOS Data rates
Figure 4: EOSALICEDAQ Data rate

Figure 5: Castor overall throughput
It is interesting to note that no data stays on the CASTOR disks for longer than 8 hours (SLA target) as apparent from fig. 7.

One should note that during the same period of time the experiments are doing also data recalls (from tape) to a sizeable level without visible interference. This is expected to change as soon as ALICE starts their data transfers to CASTOR, and the residual capacity to support tape recalls will be reduced to a minimum.

On the other hand the present of sizeable recall activity (notably for ALICE) shows that the system is not yet saturated and CASTOR has still some headroom to start the data archival as discussed in the run preparation (i.e. longer delays to tape are acceptable due to the increase of rate requirements).

2 Mid-run status

2.1 November 18-19 activity

In the following plots we show the activity of ALICE, ATLAS and CMS (experiments to IT Computer Centre only). The time interval is between Sunday 18 at 9:00 and Monday 19 at 9:00.

In the corresponding interval, the ATLAS activity from the pit is compared with the Tier0 activity (y axes on the right). The Tier0 uses the same EOS nodes for input and output and there is no sign of interference between the different activities. The Tier0 activity
Figure 8: CASTORALICE activity

Figure 9: Overall tapes recall queues

Figure 10: EOS data rates between 18th 9:00 and 19th 9:00
corresponds to jobs performing a copy of a RAW data file at the beginning of their activity and a copy back of the processed information at the end (both copy actions are done with xrdcp copy command).

During the same interval of time the global LHC EOS activity (all 4 experiments’ instances) is dominated by reading and only in the write part one can see some remnant of the HI activity. The mean value on the 24-hour interval is 37 GB/s (read) and 18 GB/s (write) with peaks exceeding respectively 60 GB/s and 30 GB/s.

The LHCb activity is summarised in this plot. It is not technically possible to isolate the LHC data taking from the other activities. Due to their mode of operating, the correlation with the beam is much less clear. Write activities (we assume dominated by data taking) amount to about 1.5 GB/s sustained over the 24-hour interval.

If we compare on the total EOS activities (LHC + PUBLIC + USER/HOME) the write activity is still modulated by the LHC data taking while the read activity exceed 50 GB/s for long periods of time over the 24-hour interval.

2.2 CASTOR activity

After ALICE started to copy data to tape, the system has reached a steady state whereby the system sustains influx rates of 2.5 to 3.5 GB/s for ATLAS, 4 to 6 GB/s for CMS, and
8 to 9 GB/s for ALICE. The LHC duty cycle is lower than expected at about 40%, which allowed to migrate to tape all data without any significant backlog. Had the duty cycle been higher, the system would have accumulated data on disk for longer periods: this scenario was anticipated to the experiments and it is expected that the SLA of 8 hours to tape is not always met. In particular, a typical ALICE session is shown in fig. 16, where the input rate is dominated by the influx traffic from EOSALICEDAQ and the output rate is the superposition of tape migrations at 6.6 GB/s (where 20 tape drives write data at their full speed) and occasional user accesses that create peaks of over 20 GB/s. To be noted that the deployed disk capacity consists of 20 disk-servers with 10 GB/s NICs and an expected network blocking factor of 2, whereas the observed blocking factor seems to be near 1, with the system delivering its full bandwidth capacity without significant contention on the tape streams.

The migration queue developed as follows during the same period of time, with a latency before going to tape that approached the 8 hours mark (fig. 19). During the same period of time, the lifetime of files on disk decreased as space was made to accommodate all the incoming data (fig. 20).

At the end of this run the accumulated backlog of 110 TB was migrated to tape in less than 5 hours, as apparent in fig. 21.

It is worth mentioning that ATLAS has a significantly lower average throughput to tape despite their data influx from P1 being only slightly lower than CMS. This is due to congestion of the EOS GridFTP gateways, as all traffic out of EOSATLAS goes by design through those gateways. To partially mitigate this, the number of parallel transfers from EOSATLAS to CASTORATLAS has been increased from 350 to 400, and other transfer ac-
Figure 16: CASTORALICE Network throughput

Figure 17: CASTORALICE Running transfers

Figure 18: CASTORALICE Migration queues
Figure 19: CASTORALICE Queued files

Figure 20: CASTORALICE Files lifetime before being garbage collected

Figure 21: CASTORALICE Migration status towards end of run
activities have been reduced. As a result, the tape capacity dedicated to ATLAS has been more than sufficient to timely write all data to tape, with a time-to-tape below 5 hours. The state of the tape migration queue during a 7-day time is shown in fig. 22.

3 End of run summary

As the LHC duty-cycle improved in the last days of the run, more data was received by the CC until the beam was dumped, as apparent from the following plots, which show the network interfaces of the network links receiving the data from the four LHC experiments, as well as the incoming data throughput to the four LHC EOS instances.

After the last beam was dumped, all experiments kept pushing their data out for several hours. During this period, CASTOR accumulated the largest backlog of data to be migrated to tape, and several recalls from tape for all experiments were delayed to leave more bandwidth to data recording, as visible in the following plots. ALICE kept recording data to tape for two more weeks, as most of the data was only stored on EOSALICEDAQ.
Figure 23: Combined EOS+CDR overview

Figure 24: CASTOR Tape overview