LHC Pb-Pb Ion Run preparation: storage workflows and requirements

26th September 2024: https://indico.cern.ch/event/1459784/

***Attendance**: Latchezar Betev, Liz Sexton-Kennedy, Michael Davis, Luca Mascetti, Ben Couturier, Jakub (Kuba) Moscicki, Julien Leduc, Andreas Peters, Costin Grigoras, Xavier Espinal (notes), David South (R), Stefano Piano (R), Jan van Eldik (R), Phat Srimanobhas (R)*

Topic: Review the LHC experiment plans, workflows and requirements from the disk and tape storage perspective for the upcoming heavy-ion run 6th-24th of November (18 days of run with one MD day in the schedule)

Summary of the discussion:

Tape infrastructure:

- Julien from the IT Storage Group gave an update about the status of the tape infrastructure, also reminding the baseline bandwidth availability for tape recording and support.
- There is no substantial change on the tape infrastructure with respect to the tests performed during the Data Challenge 2024: **40GB/s is the nominal available throughput** for the 4 LHC experiments and any extra throughput beyond that figure cannot be guaranteed. Julien also reminded that in case of throughput saturation the tape team has the means to arbitrate across throughputs in almost real time, should be noted this support is on best effort basis outside working hours.

The LHC experiments plans for Pb-Pb data taking.

- ATLAS and LHCb: **no substantial changes with respect to p-p run**. Throughputs to tape staying around the 10GB/s metric and enough local buffer contingency to sustain data recording load for a long weekend.
- ALICE: will make use of EOSALICEO2 to record on disk the full Pb-Pb run and **agreed to start tape migration after the 26th of November**, hence relaxing the pressure on the tape infrastructure and leaving some extra margin for peaks and for other experiments/baseline tape activities.
- CMS reported some limitations on the P5 Lustre infrastructure and is setting up a **new extra workflow** from P5 to EOSCMS during the Pb-Pb to record minimum bias data. The average rate to EOSCMS is expected to be a combined average of 16GB/s (18-20 GB/s peaks) constant during the run and coming from both systems alternatively (fills vs interfills):
    - The new workflow is set up to record **opportunistic** minimum bias data during **fills**. The data is streamed to EOSCMS from P5 SSD-based nodes **directly**, hence skipping the recording/consolidating step in the P5 Lustre system. This is alleviating the pressure on the local system and gives margin for the standard physics data recording. *NB: The nature of this setup makes it vulnerable to glitches (data only on flash memory before transferring) but it is considered a "good enough" setup to cater for such opportunistic data type.*

- It is understood that if for any reason (e.g. service failure in P5, IT Data Center or network) and for any duration, the opportunistic data cannot be transferred, it won’t be recorded.
-
- The “normal/standard” workflow will continue to run during **interfills**, with no changes. Data is streamed to EOSCMS in CERN-IT from the SSD-based nodes **reading from** the Local P5 Lustre system.
- It is pointed out that there is an **inherent risk** in case the two workflows (normal vs. minimum-bias) overlap. CMS is aware of it, mentioning they have monitoring in place and the means to give priority to either stream. CMS will be favouring the “normal/standard” physics stream over the opportunistic “minimum bias” stream.
- The target tape rate is expected to be around **15 GB/s**, above the fair share but absorbable by the tape system thanks to the relief provided by ALICE by delaying the tape recording until the end of the Pb-Pb run.
- It was pointed out that the Lustre buffer at P5 has a capacity of 1.3 PB, which caters for **21 hours** of buffering without any transfer to EOSCMS. This is based on the following assumptions: a) constant data-taking rate of 17 GB/s during fills and b) 50% LHC duty cycle allowing enough time for data to be transferred to EOS, preventing backlog accumulation.
- It was noted that current sizing of the CMS Lustre buffer may not, under unfavourable circumstances, provide enough local buffer contingency to sustain data recording load during the weekend.