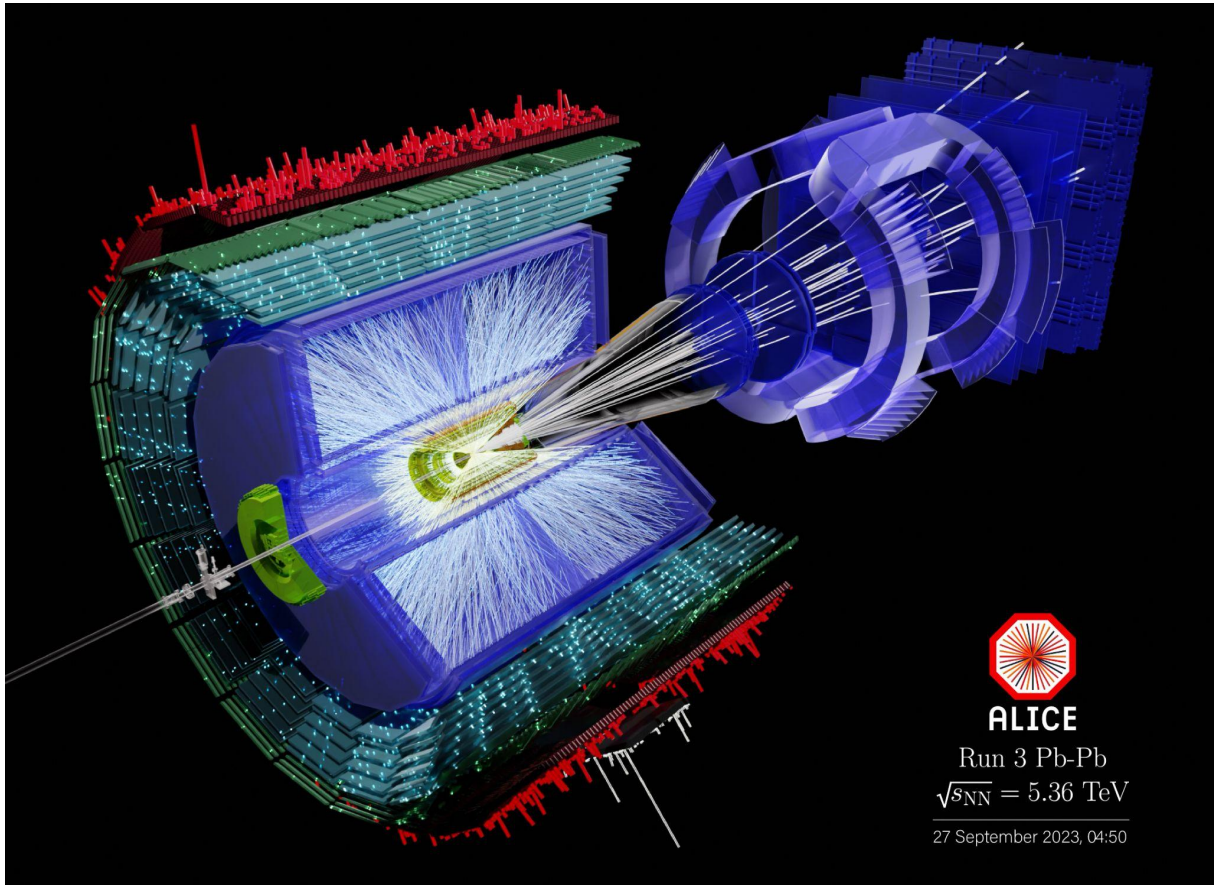
ALICE Computing in Run3 and processing plans for 2024

Latchezar Betev

7th Asia Tier Center Forum, Jeju-do, November 1-3, 2023

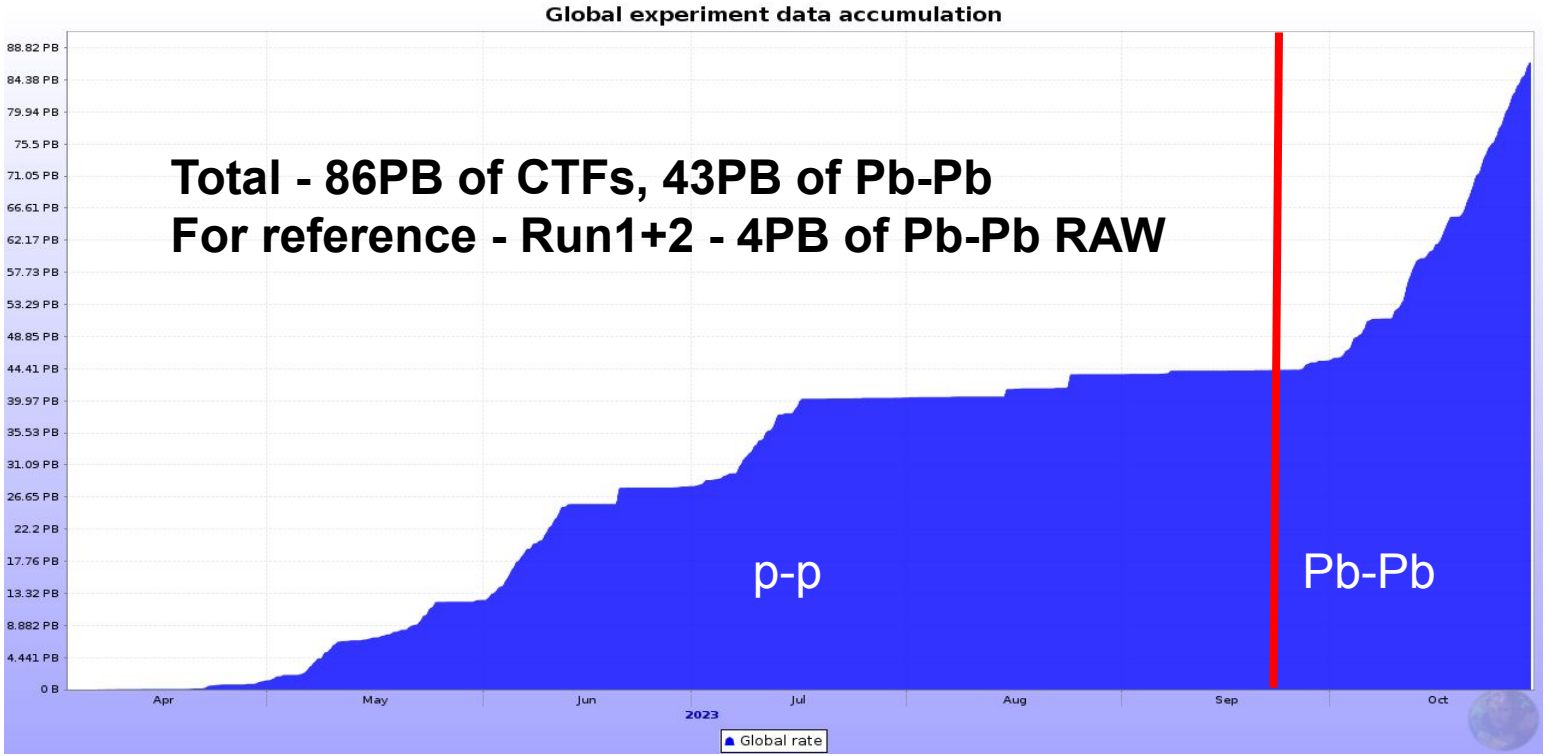
2023 - second year of Run3
First year of Pb-Pb beam

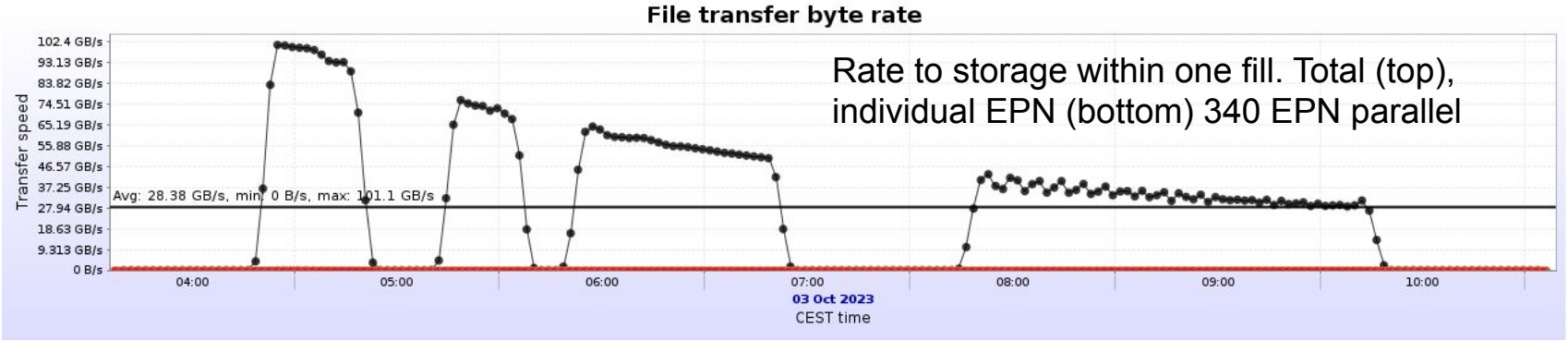
A Large Ion Collider Experiment



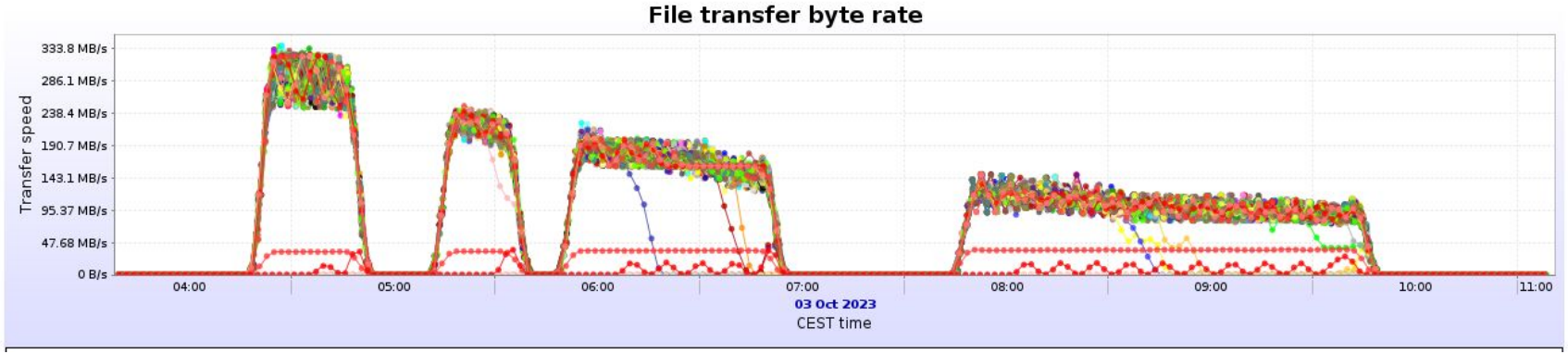
First Pb-Pb in 2023
(low IR - 6kHz)

A Large Ion Collider Experiment

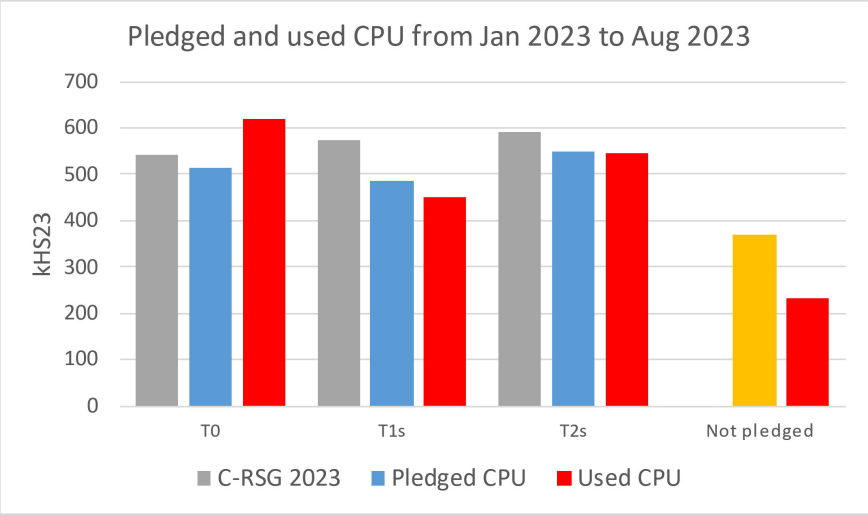




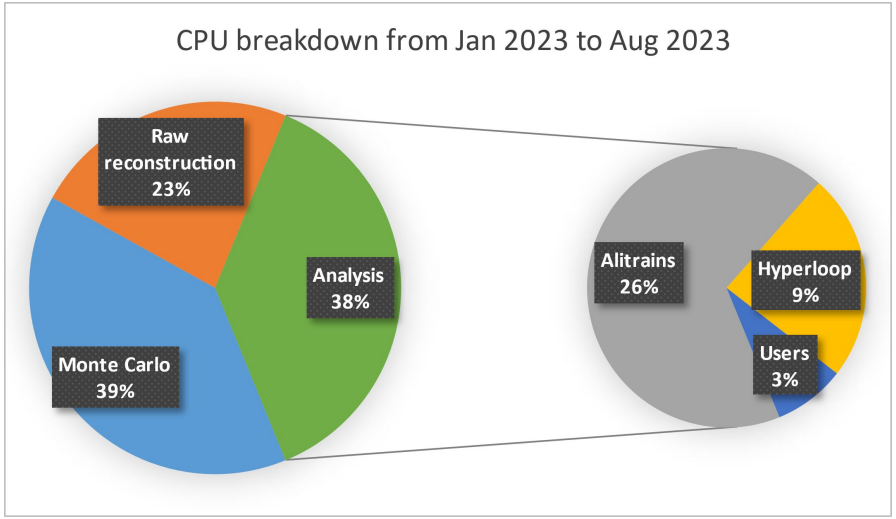
Rate to storage within one fill. Total (top), individual EPN (bottom) 340 EPN parallel



CPU utilization and breakdown by job types

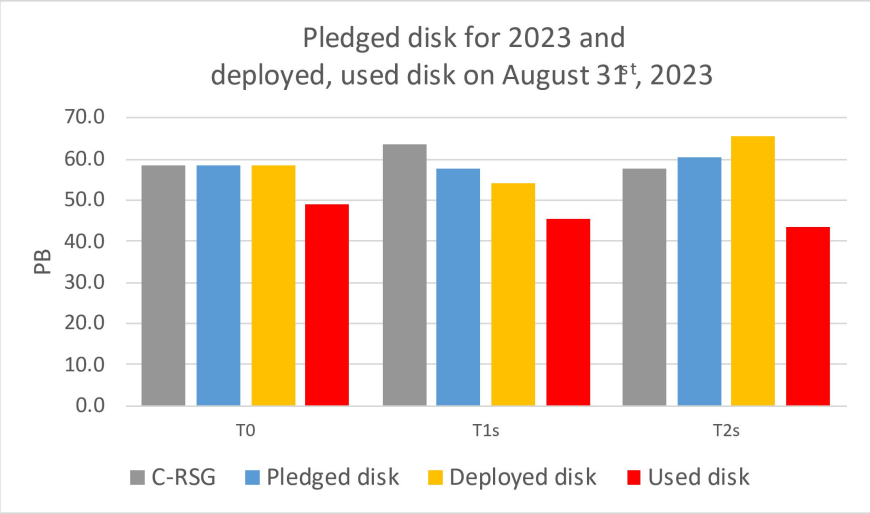


- Good utilization of pledged resources
- Opportunistic CPU usage at the T0 and LBNL, Japan, Wigner and EPN (230 kHS23 only CPU, with 2.5 GPU speedup factor from April => 370 kHS23)

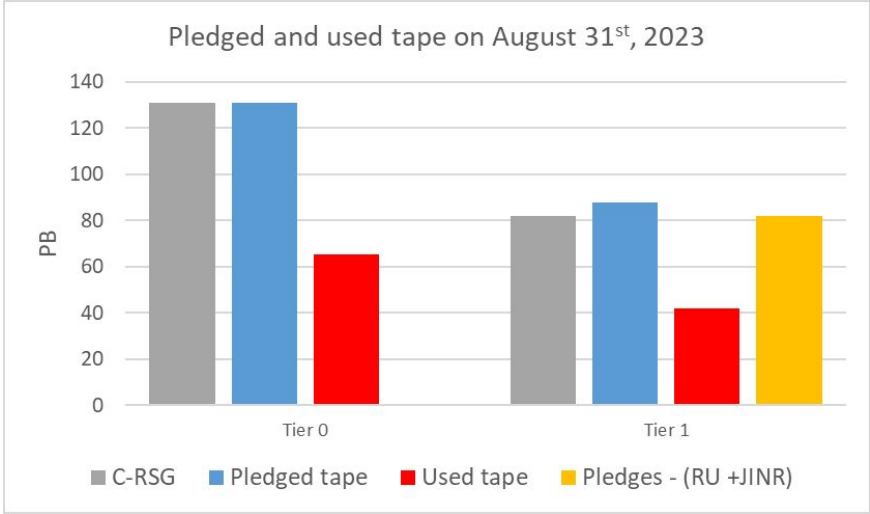


- High activity for raw calibration and reconstruction of Run 3 pp runs
- Growing analysis activity for conferences and publications both on Run 2 and Run 3 data
- Lower MC share affecting T1 T2 CPU usage (!)

DISK and TAPE utilization (to be updated)



- 2023 disk deployment: 100% at T0 and T2s, 95% at T1s
- Used 80% of capacity at T0, T1s and 75% at T2s
- Expected to fill up most of the disk by spring 2024 (Pb-Pb reco + MC)

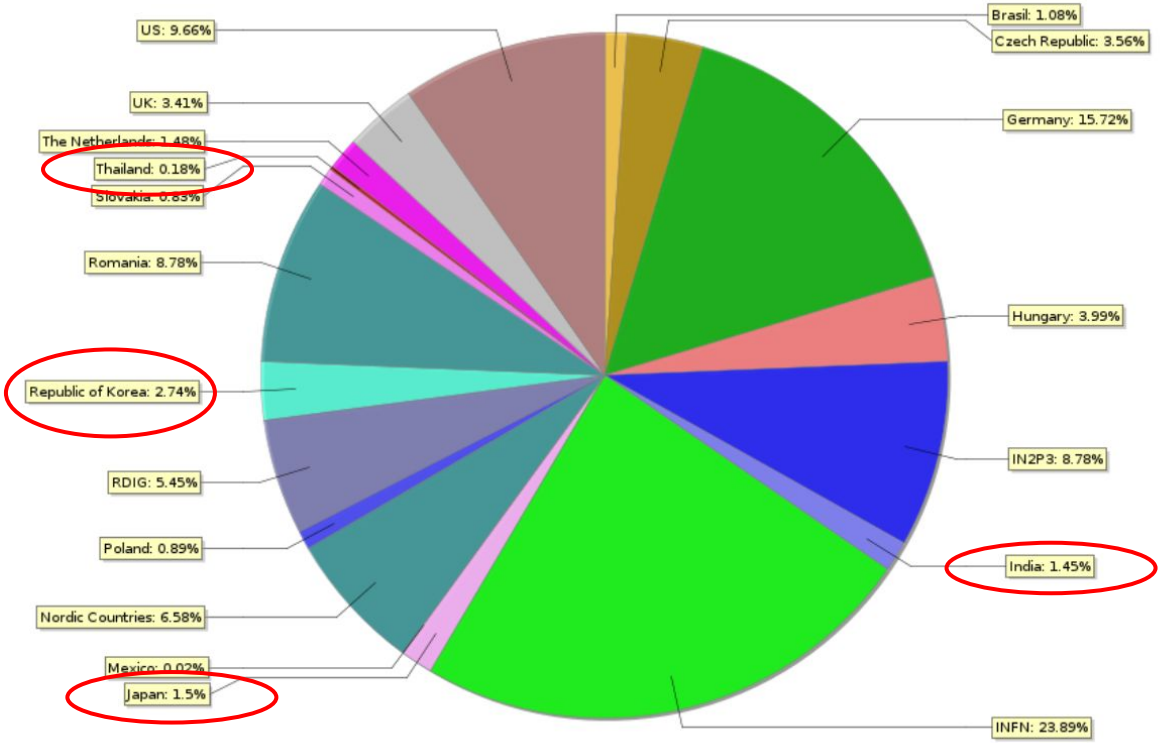


- Pledged tape 100%@T0 and surplus at T1s (+5.7 PB) - compensates the tape pledged by RU
- Enough for 5w of Pb-Pb (extended programme)

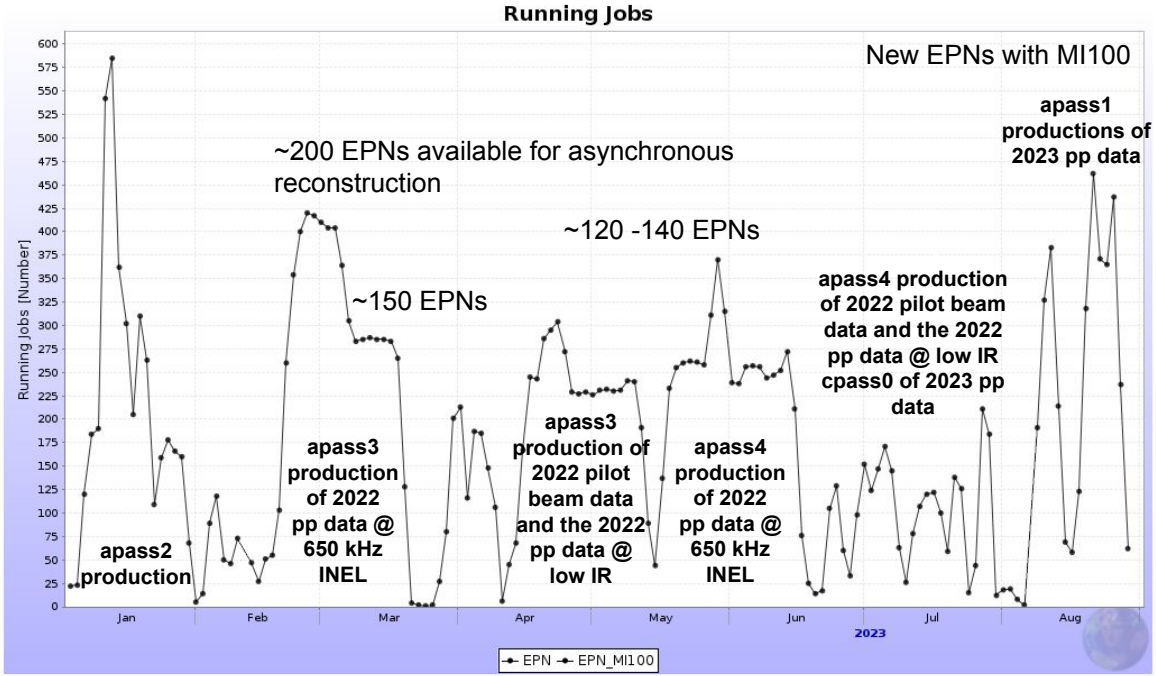
Regional contribution

Total contribution of Asian centres - 6%

Proportional growth with respect to the rest of the Grid



Asynchronous reconstruction on EPN (CPU+GPU)



2022 data calibrations and processing

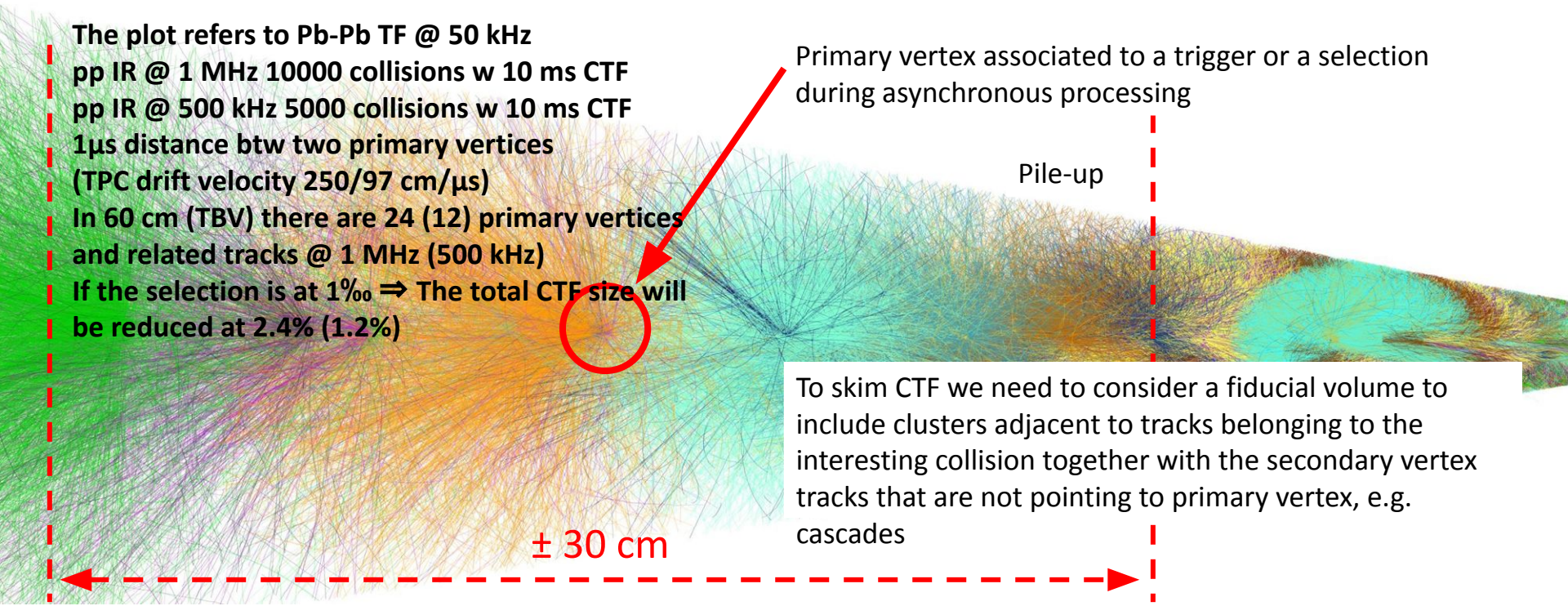
- Collected 15.6/pb of pp @ 650 kHz INEL IR:
  - Four processing calibration campaigns on the full statistics
  - Last pass (apass4) with TPC analytical map correction suitable for skimming
  - Skimming and validation - completed for 2022 pp data
- Offline selection factored in 4 steps (only for 2022):
  - Asynchronous reco ⇒ Event tagging ⇒ CTF skimming ⇒ Asynchronus reco of skimmed CTF for validation
  - Event tagging: selections by analysis tasks, tags about 0.1% of the collisions
  - CTF skimming: CTFs are cut keeping only info for the selected collisions
    - Not possible to apply a tight window cut (±30 cm of the PV of the selected event)
    - Needed to consider all the clusters of [-0.25, 1.25] TPC drift time
    - Compression factor increased from 1.5% to 6% for 2022 pp data
  - Tighter physics selections (~0.05%) applied to 2023 pp data to compress the CTF files at 3%

2023 data taking and readiness for HI

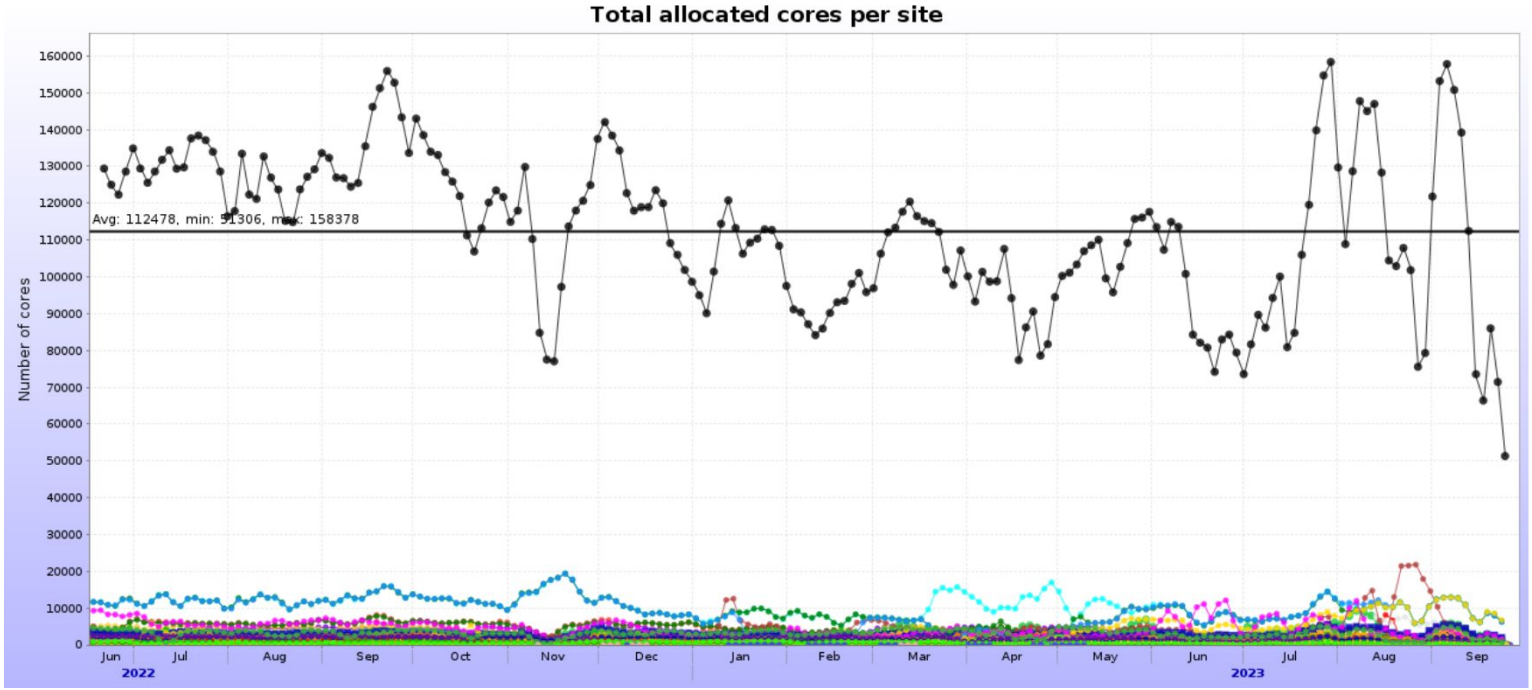
- Collected 9.4 pb^{-1} for pp physics programme
- Focus on commissioning for HI:
  - 0 B field data for alignment and low B field (0.2 T) for calibrations and physics
  - Interaction rate scan campaigns
    - 10 kHz - 1.5 MHz with different and fixed machine filling scheme conditions
    - 500 kHz - 4 MHz exceeds the equivalent charged track load of Pb-Pb at 50 kHz
    - Among other studies, test and validate TPC firmware with dense data format
- HI data taking:
  - 70 new EPN nodes with MI100 ordered, delivered and installed at ALICE Point 2
  - Planned to collect about 90PB of data
  - Collected ~½ of that

Skimming and rejection power

The plot refers to Pb-Pb TF @ 50 kHz
pp IR @ 1 MHz 10000 collisions w 10 ms CTF
pp IR @ 500 kHz 5000 collisions w 10 ms CTF
1μs distance btw two primary vertices
(TPC drift velocity 250/97 cm/μs)
In 60 cm (TBV) there are 24 (12) primary vertices and related tracks @ 1 MHz (500 kHz)
If the selection is at 1‰ ⇒ The total CTF size will be reduced at 2.4% (1.2%)

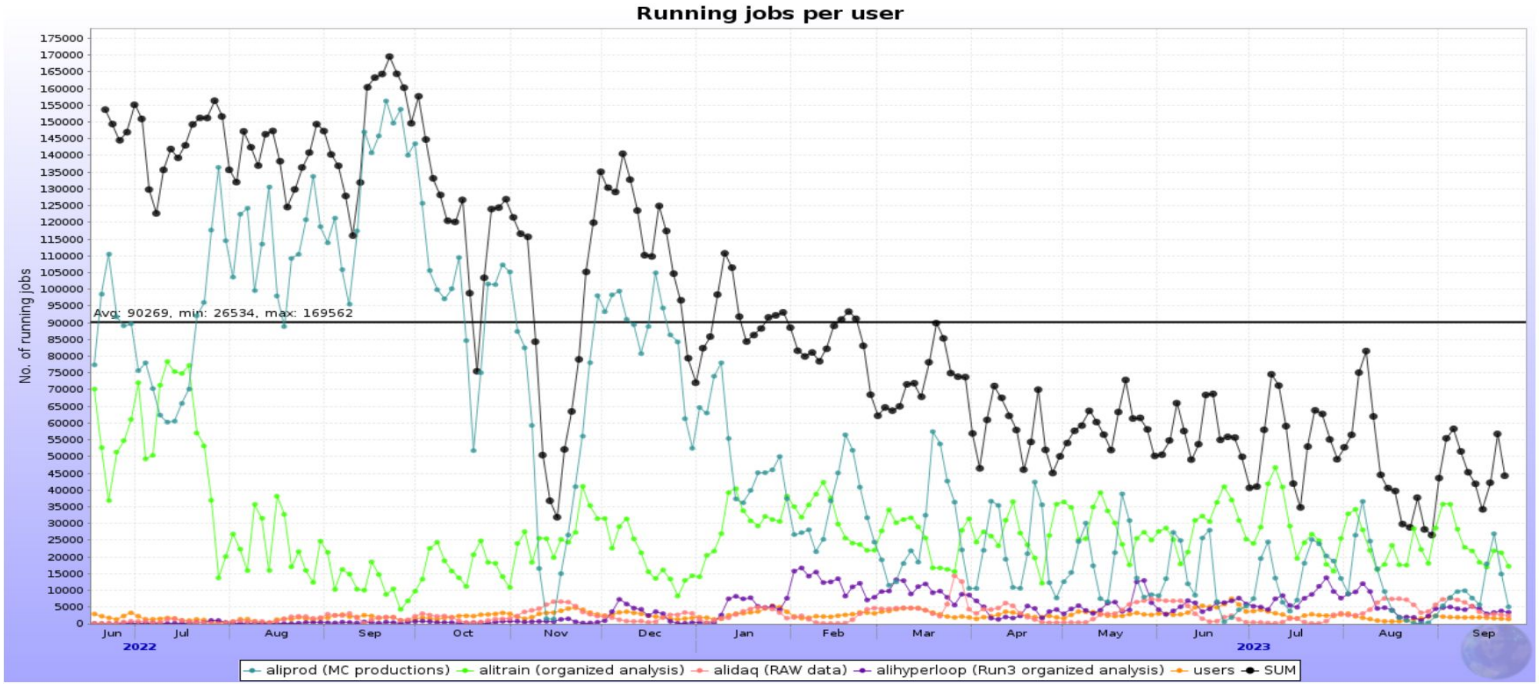


Core allocation profile



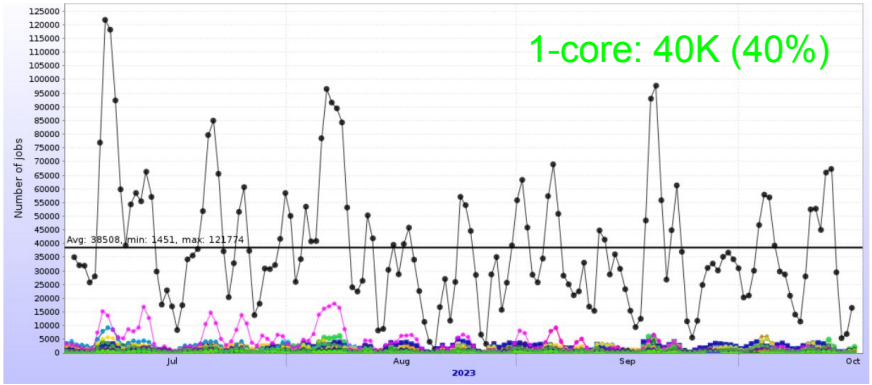
- Mix of single-core (alitrain), 1-2-4 core (hyperloop), 8-core (O2 MC and O2 RAW)

Job profile per user

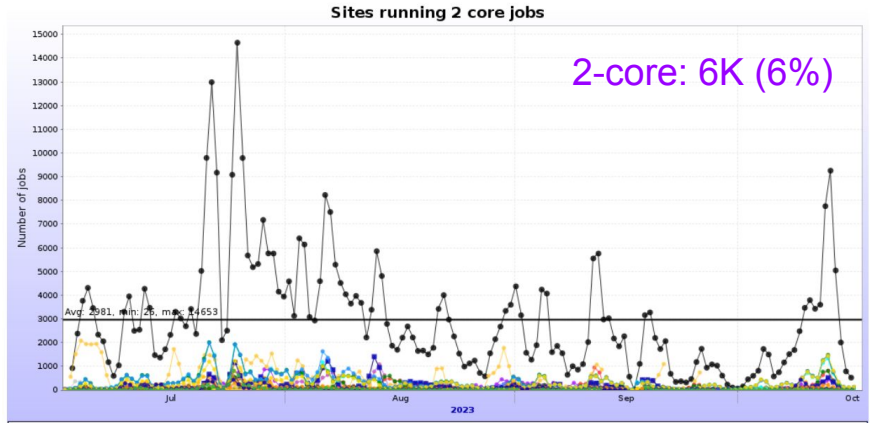


- Reduction of number of jobs by ~x3 - move to multicore processing

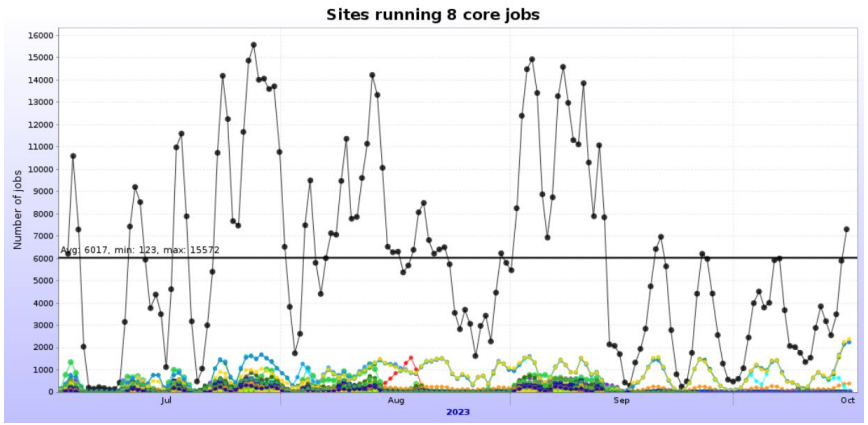
Job usage by core count



- Single core jobs (legacy MC, alitrain, users) diminishing
- Multi-core jobs are split into several categories, MC and CTF reco: 8+-core
- hyperloop analysis: 2-core

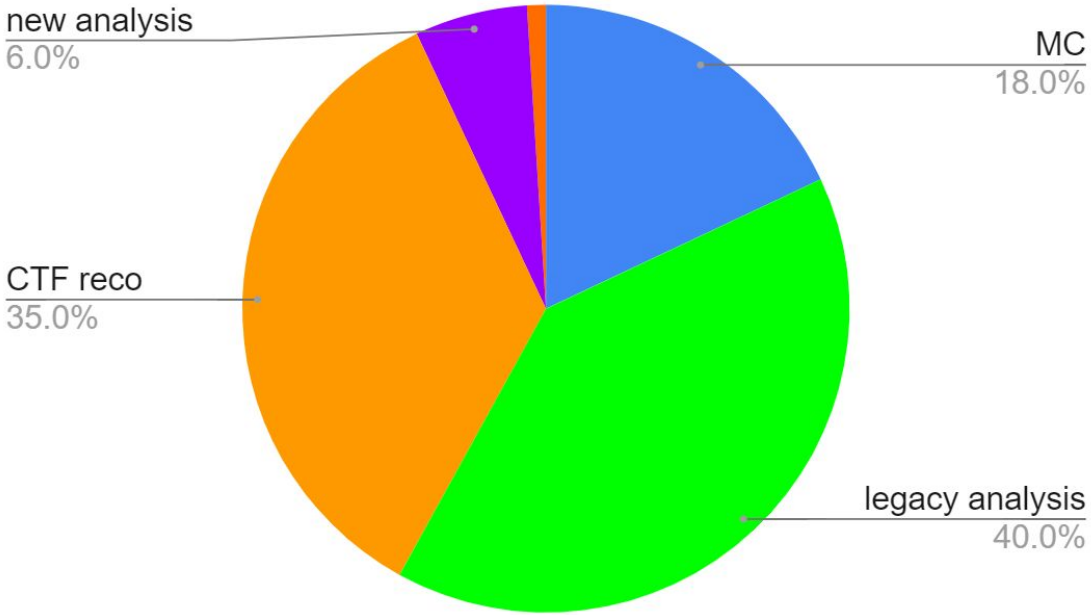


8+ cores:54K (54%)



CPU occupancy by activity (July-October)

- Analysis is dominating, CTF reco not far behind
- Substantial modification with respect to the previous period



Capacity distribution

- Most of the sites are providing 8-core queues (compatible with the requirements of other VOs)
- Specific sites are providing multi-core queues (1-NUMA isolation)
- Preferred option is a whole-node queue (job agent decides the split per job)
  - Allows for combination of different types of jobs to compensate for high memory request, I/O balancing, TTL optimization
  - Given a good network, perform tasks of another site (pull data remotely)

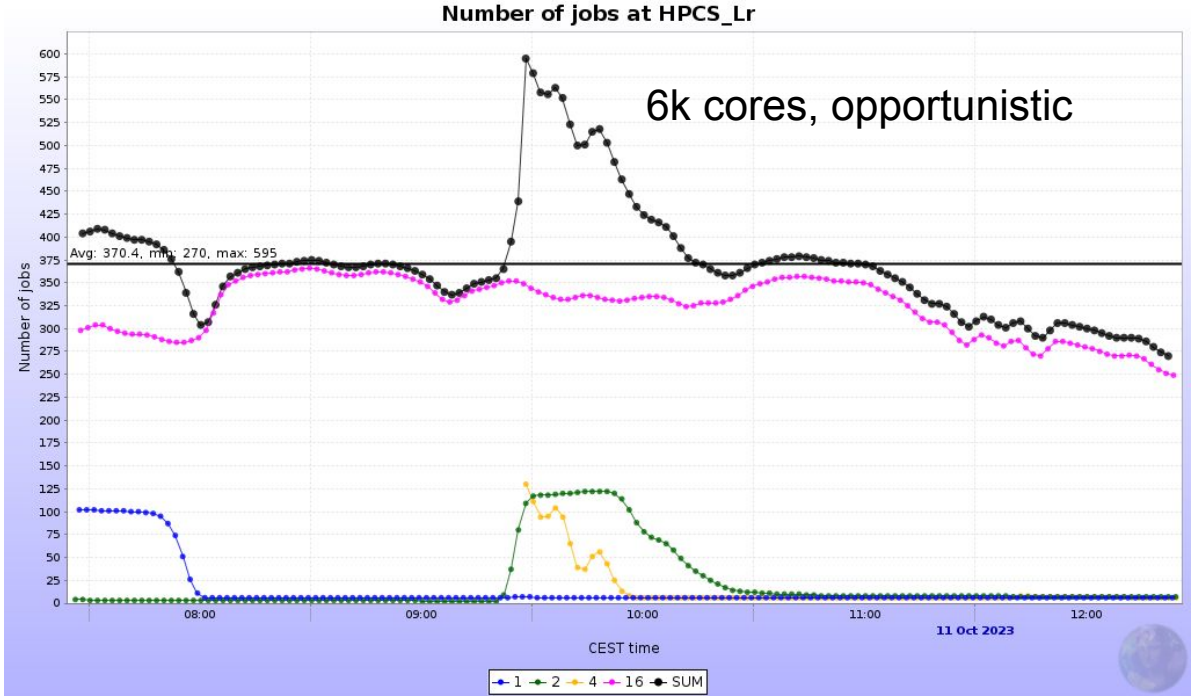
	LDAP	
Service	Stat	Cores
51. Perlmutter		256
21. EPN_MI100		96
20. EPN		64

Service	Stat	Cores
24. FZK_KIT		0
31. HPCS_Lr		0
37. KISTI_GSDC		0
40. LBL_AFP		0
41. LBL_HPCS		0
45. NIHAM		0
48. ORNL		0

Remote job execution

- LBNL/NERSC initiative - use Lawrencium and Perlmutter for remote Pb-Pb reconstruction

- Fully opportunistic resources
- Adequate network - data pulled from CERN, no saturation
- 6-10k cores in addition to T0 (up to 10% more resources)
- More slots for p-p reco (2023 apass2) @CERN



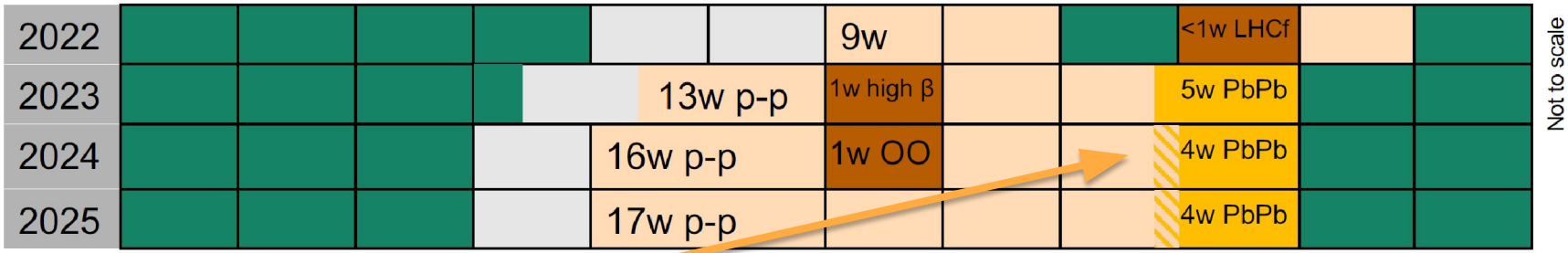
ALICE

Analysis of computing resources use

- Two major trends
  - Multicore processing (expected), but with nuances
  - Data driven processing: 60% of computing resources (partially expected)
- Consequences
  - Less MC jobs, which usually act as a filler and smoother of resources use
  - More ‘spikes and valleys’ in CPU utilization at the T2 farms, T0/T1s are less affected
- Mitigation - move more data-intensive tasks to T2s
  - Cost - more storage and increased network use
  - Is this feasible in medium-term?
  - What is the best computing centre envelope in which to achieve it?

Computing resource and processing plans 2023 - 2024

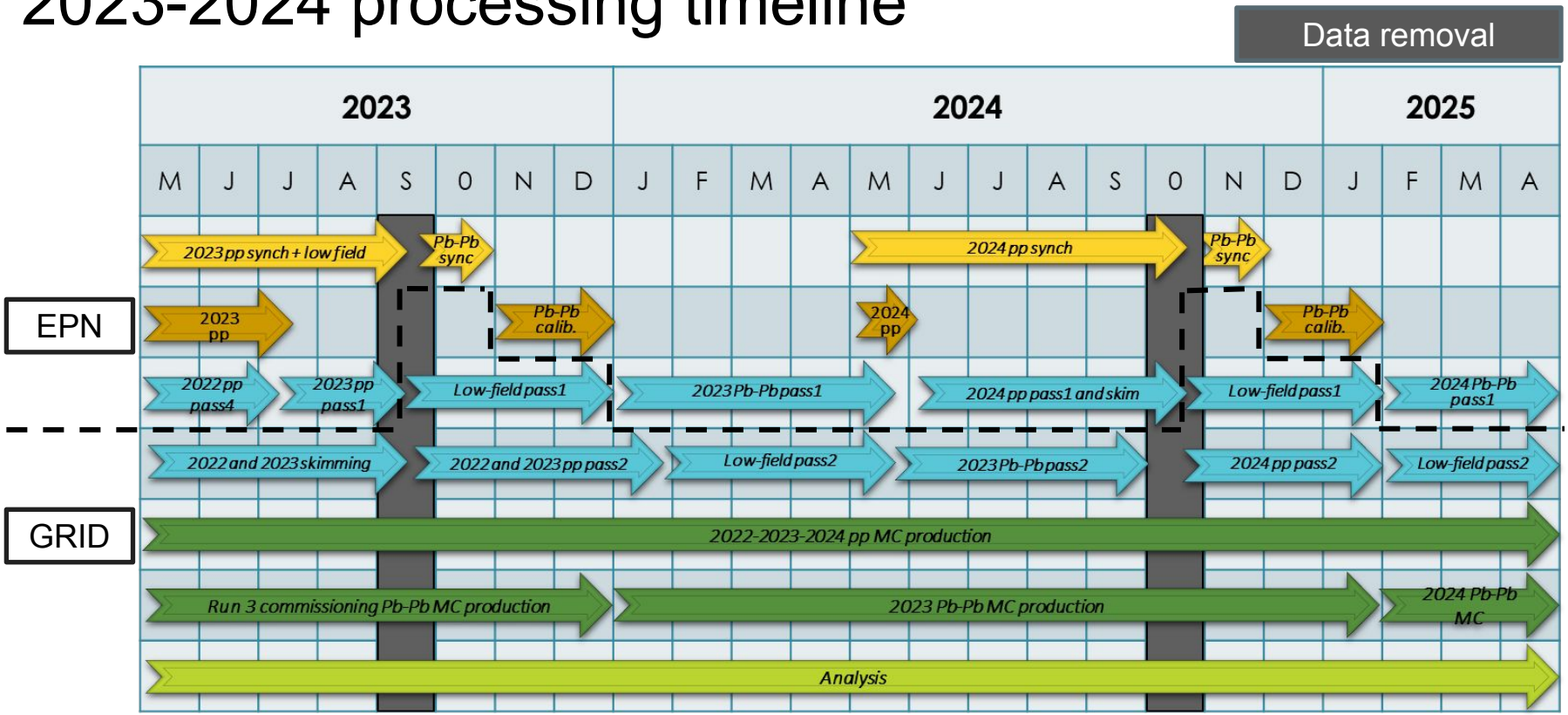
Baseline scenario for 2024



- **Assumed that the HI run in 2024 could be extended to 5 weeks**
- **Same luminosity goals of 2023 for Pb-Pb and pp ref runs:**
  - **3.25 nb^{-1} of Pb-Pb collisions (strategy B aggressive)**
  - **3 pb^{-1} of pp ref run**
- Such an assumption accommodates with some margin, all the different possible scenarios for the HI period in 2024.

- **Considered as upperlimit:**
  - **112 days of pp in 2024:**
    - ~42 pb^{-1} of pp full-field
    - ~2.8 pb^{-1} of pp low-field
  - **Short O-O and p-O run:**
    - 1 nb^{-1} and 5 nb^{-1}, respectively

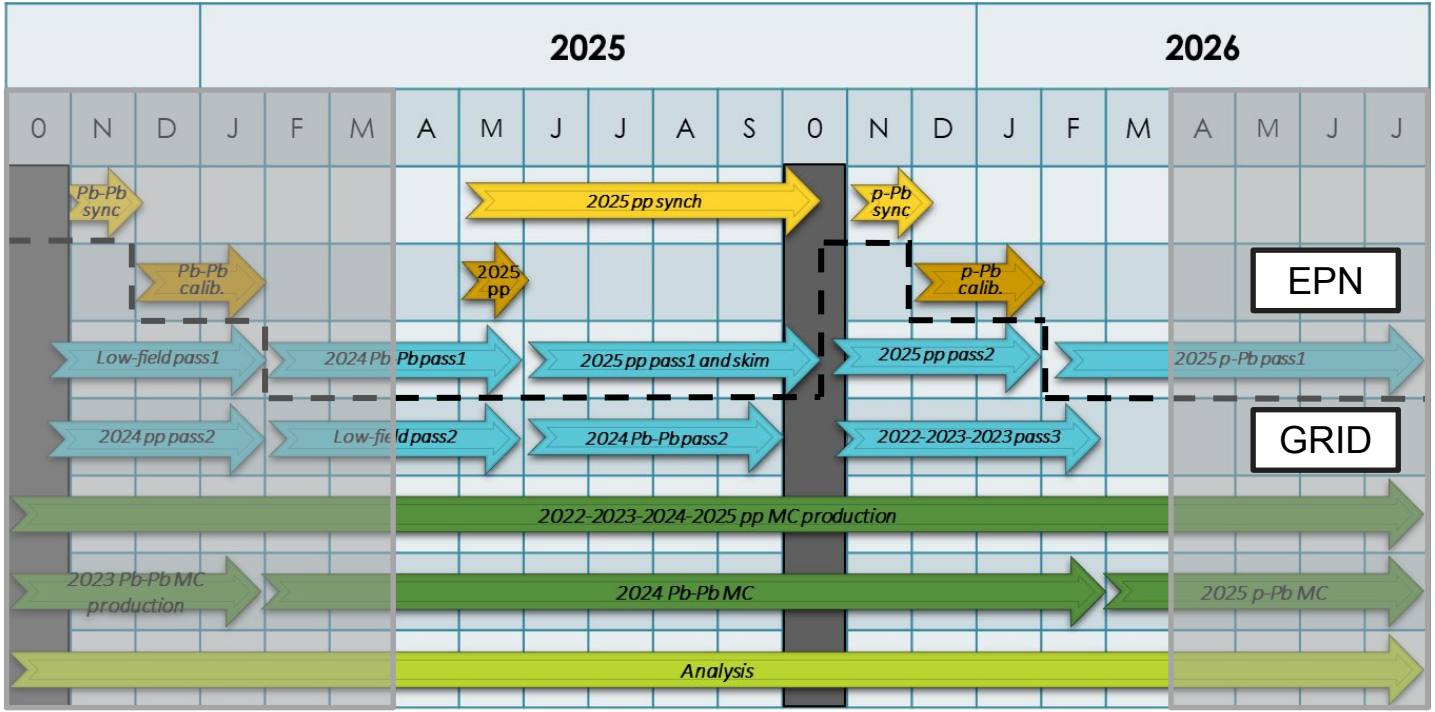
2023-2024 processing timeline



2025 processing timeline and resource needs

Data removal in 2024 and 2025

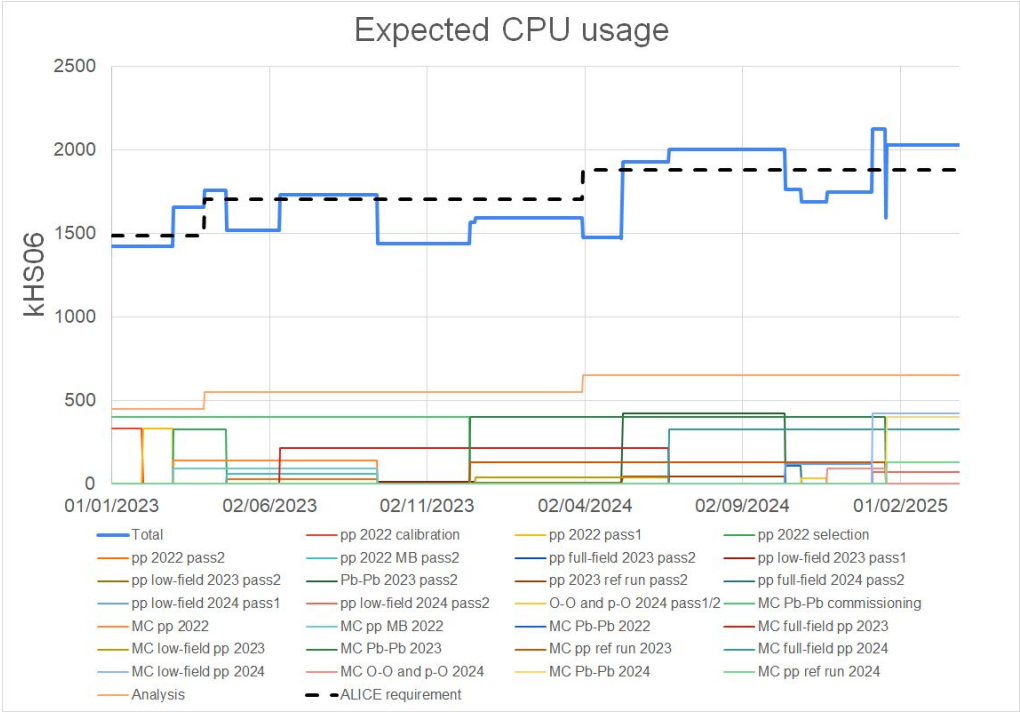
No data removal in 2026



Tape:
2025 p-Pb 49 PB,
2025 pp 3 PB + 14 PB
increased selections
of the pp full field data

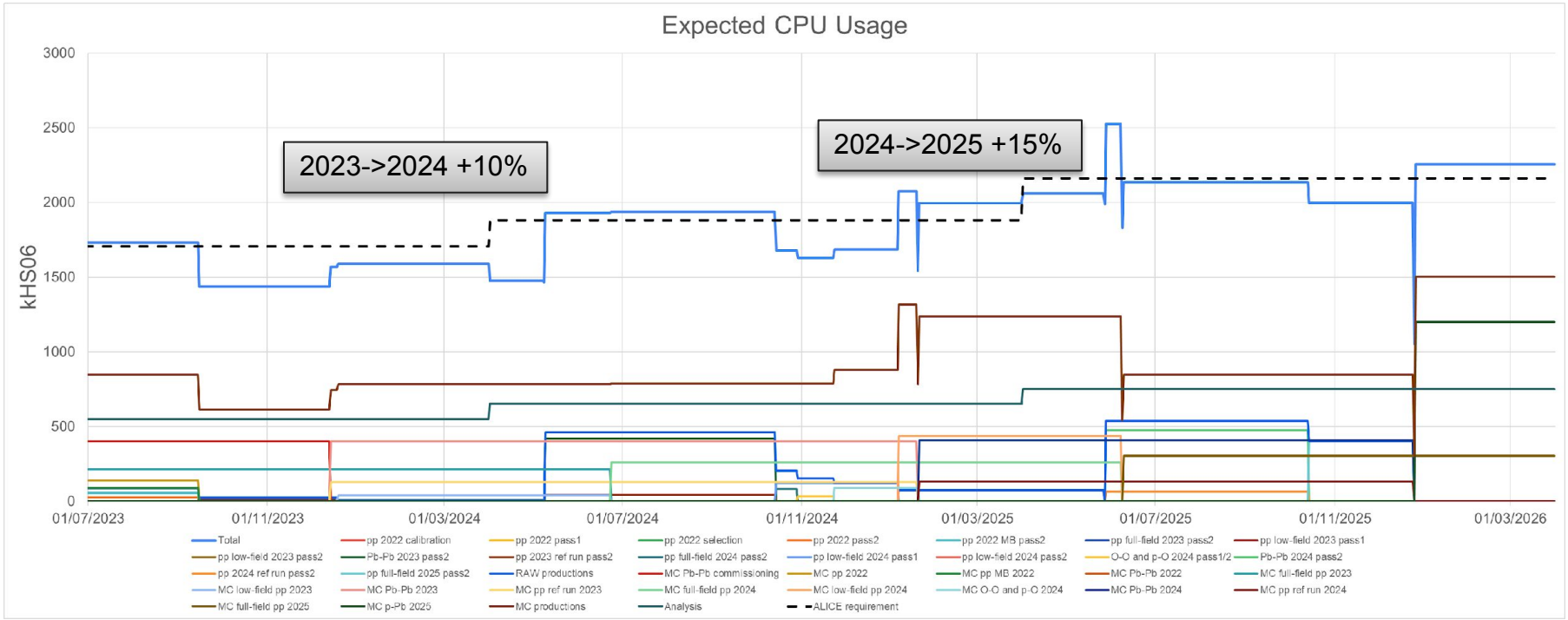
Most of the disk to accomplish 2024 data productions (27 PB), 6 PB for 2025 ones

CPU needs for 2023 - 2024



- Blue line - minimum CPU capacity needed to process all planned productions
- Dashed line - ALICE requests
- The achieved performances of the asynchronous reconstruction on EPN - allows to lower 2024 CPU request from 1960 kHS06 to 1880 kHS06

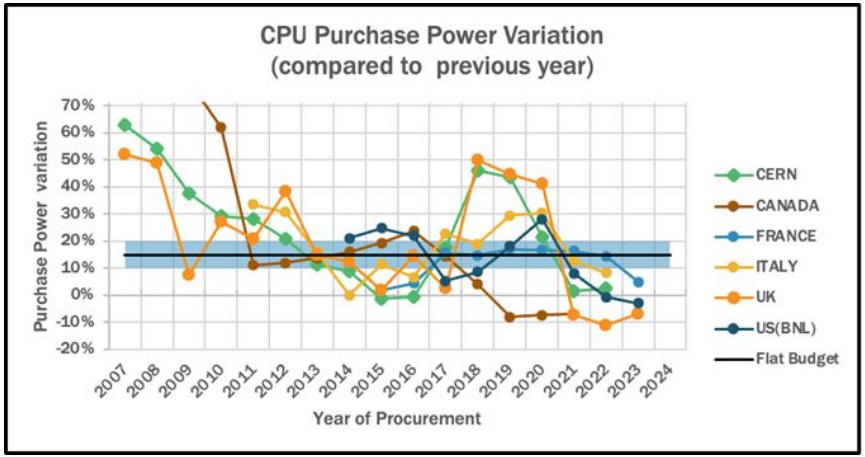
CPU needs for 2025



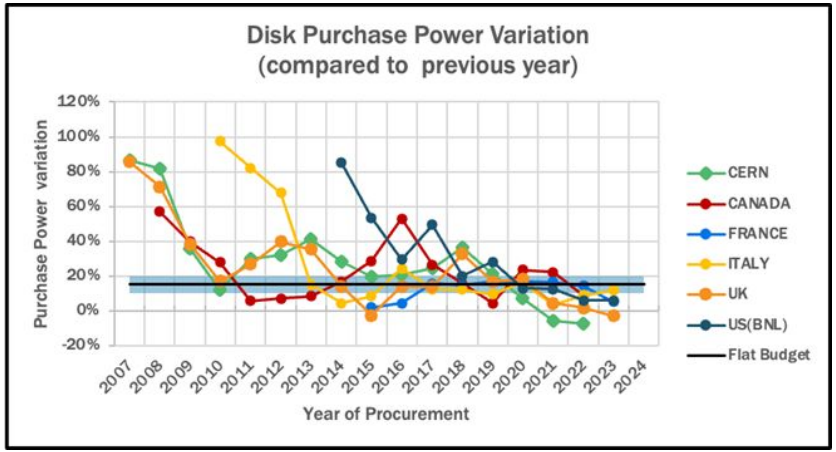
Hardware trends

- The WLCG ”flat budget model” assumption: +15% CPU, disk and tape every year **with the same level of funding**
- Monitor the HW trends in many countries. Results for UK and CERN were presented at the lats RRB. The study is now more complete (6 countries)

CPU average variation (5 years): +14%



DISK average variation (5 years): +15%



Disk and tape needs for 2024

- Disk: AOD average event sizes are unchanged with respect to 2022 and 2023 requests
- Tape: considered the adoption of compression strategy B (aggressive) in 2024:
  - **CTF average event size at +30%** as an **upper limit for strategy B as well**

ALICE		2024										
		pp 2023	pp ref 2023	Pb-Pb 2023	pp low field 2023	pp 2024	pp ref 2024	Pb-Pb 2024	O-O and p-O 2024	pp low field 2024	Total	Total - carry over from 2023
		To be processed in 2024										
Disk [PB]	Tier-0	0.0	1.4	4.9	0.2	1.6	0.7	2.3	0.2	1.5	12.8	9.3
	Tier-1	0.0	1.3	4.7	0.2	0.8	0.7	2.3	0.2	1.5	11.7	8.2
	Tier-2	0.0	1.4	5.1	0.2	0.5	0.7	2.4	0.2	1.6	12.1	8.2
	Total	**0.1**	**4.1**	**14.7**	**0.5**	**2.9**	**2.0**	**7.0**	**0.7**	**4.6**	**36.7**	**25.7**
Tape [PB]	Tier-0	0.0	0.0	0.0	0.0	1.6	3.7	41.3	0.4	5.4	52.4	55.0
	Tier-1	0.0	0.0	0.0	0.0	0.8	1.9	20.6	0.2	2.7	26.2	19.9
	Total	**0.0**	**0.0**	**0.0**	**0.0**	**2.4**	**5.6**	**61.9**	**0.6**	**8.1**	**78.7**	**74.9**

Disk and tape and CPU needs for 2025

ALICE		2023			2024			2025	
		C-RSG	Pledge	RU + JINR Pledge	C-RSG	Req. 2024 / C-RSG 2023	Req. 2024 / (Pledges - RU) 2023	Est.	Est. 2025 / C-RSG 2024
CPU [kHS23]	Tier-0	541	541		600	111%	111%	690	115%
	Tier-1	572	506	33	630	110%	133%	725	115%
	Tier-2	592	567	35	650	110%	122%	750	115%
	Total	**1705**	**1614**		**1880**	**110%**	**116%**	**2165**	**115%**
Disk [PB]	Tier-0	58.5	58.5		67.5	115%	115%	78.5	116%
	Tier-1	63.5	57.6	4.5	71.5	113%	135%	82.5	115%
	Tier-2	57.5	60.4	3.0	66.5	116%	116%	77.5	116%
	Total	**179.5**	**176.5**		**205.5**	**114%**	**116%**	**238.5**	**116%**
Tape [PB]	Tier-0	131	131		181	138%	138%	226	125%
	Tier-1	82	88	6	107	130%	131%	135	126%
	Total	**213**	**219**		**288**	**135%**	**132%**	**361**	**125%**

- **Resource estimates for 2025 submitted to C-RSG (October RRB)**
- Standard growth for CPU (+10%,+15%) and disk (+14%, +16%) in 2024 and 2025 compatible with flat budget
- Large step for tape, where for 2024 and 2025 compression strategy B has been considered with larger average event size (+30%) wrt estimates based on MC

Summary (1)

- **Computing resource utilization:**
  - ~Full utilization of CPU resources
  - EPN CPU and GPU resources successfully exploited for the processing of pp data
  - The postponed 2022 HI data taking lowers our GRID disk needs in 2023, but
    - 2022 pp skimmed CTF files and 2022 pp pass4 AO2Ds temporarily parked
    - Expected to fill up most of the disk with the processing of 2023 HI
  - Estimated a tape deficit of 14 PB for the archival of 2022, 2023 and 2024 pp skimmed CTFs
- **2022 and 2023 pp data processing:**
  - Tight schedule to balance reconstruction and skimming of 2023 pp data
  - Removal of 2023 pp CTFs before HI run - changed to ‘remove as you need the space’
- **Resource requests for 2024 and estimates for 2025:**
  - CPU and disk compatible with flat budget
  - Step for tape despite considering the adoption of aggressive compression in 2024
  - Uncertainty around Russian resources remains; requesting other FAs to cover if needed

Summary (2)

- **Computing resource utilization:**
  - Full utilization of CPU resources
  - EPN CPU and GPU resources successfully exploited for the processing of pp data
  - Disk and tape expected usage in line with the requested resources excluding Pb-Pb

- **Computing resources needs for 2023 with the updated Run 3 schedule:**
  - The postponed 2022 HI data taking lowers our CPU and disk needs in 2022-2023
  - Re-assessed tape needs with strategy A with larger average event size (+30%)
  - and with longer HI period in 2023

- **Resource requests for 2024:**
  - Considered the carryover from 2023, step for tape (+75 PB)
  - CPU and disk in 2024 compatible with flat budget considering our 2023 requests

- **Sizeable impact of the war in Ukraine: RU resources needed to be replaced by 2024**