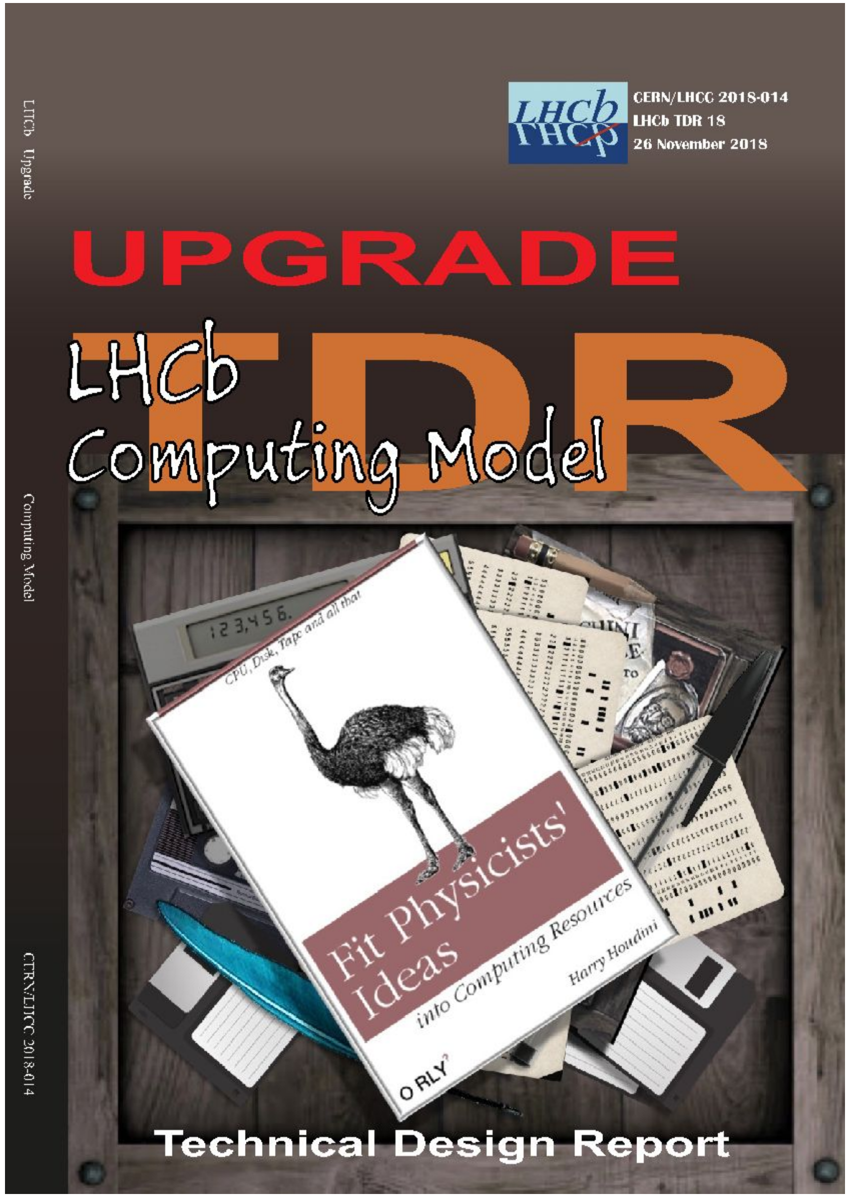
# Modèle de calcul vs réalité: LHCb

Federico Stagni
Concezio Bozzi
Journées LCG-France
Paris, June 7th, 2023

# Overview

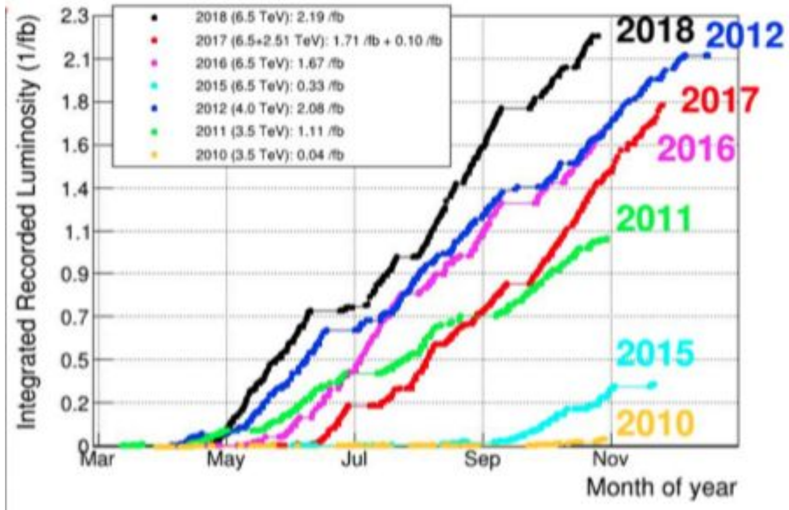
- LHCb Upgrade
- Run3 + Run4 computing model

- Current status
- Future evolution

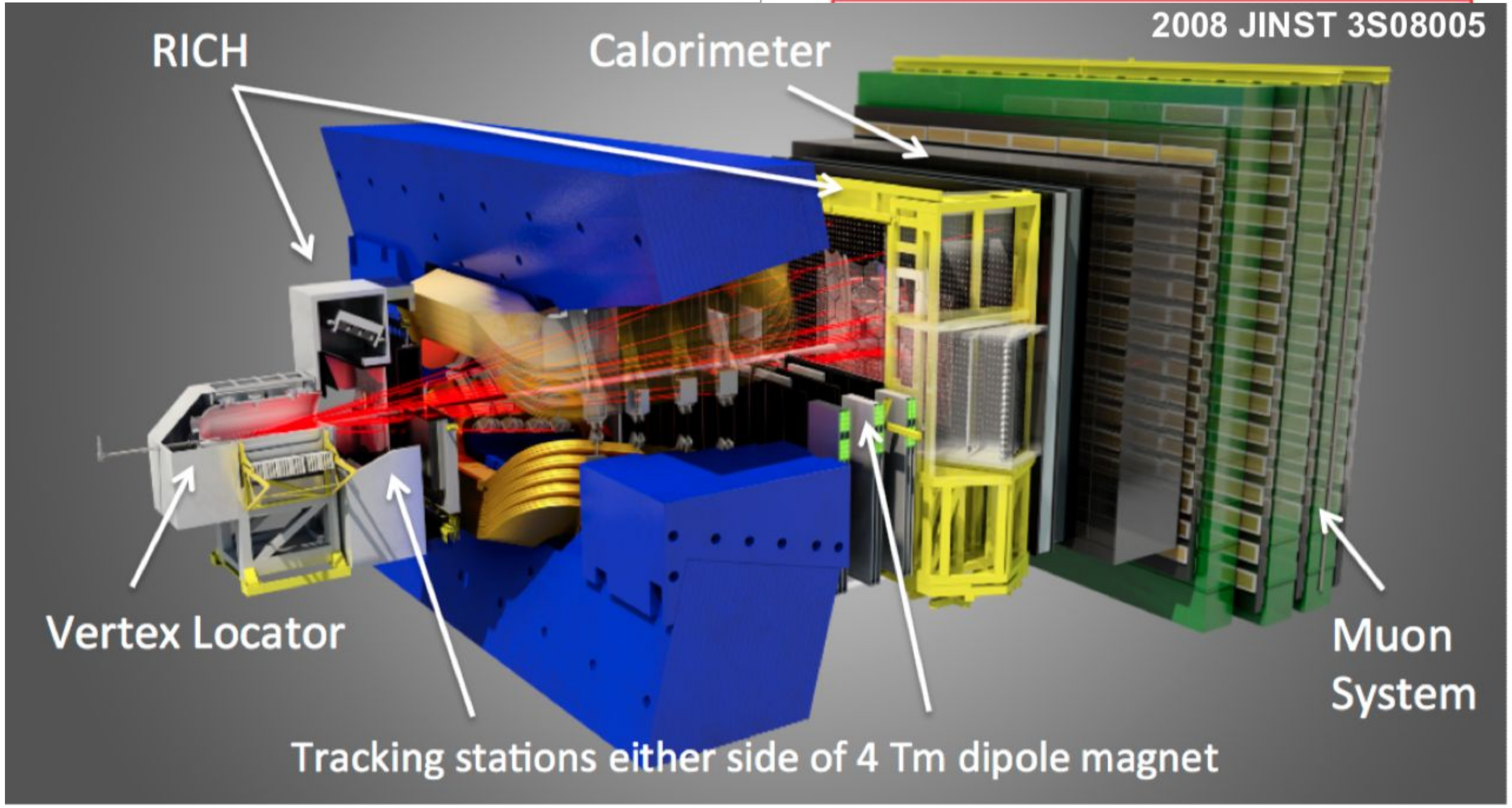



LHCb-TDR-018

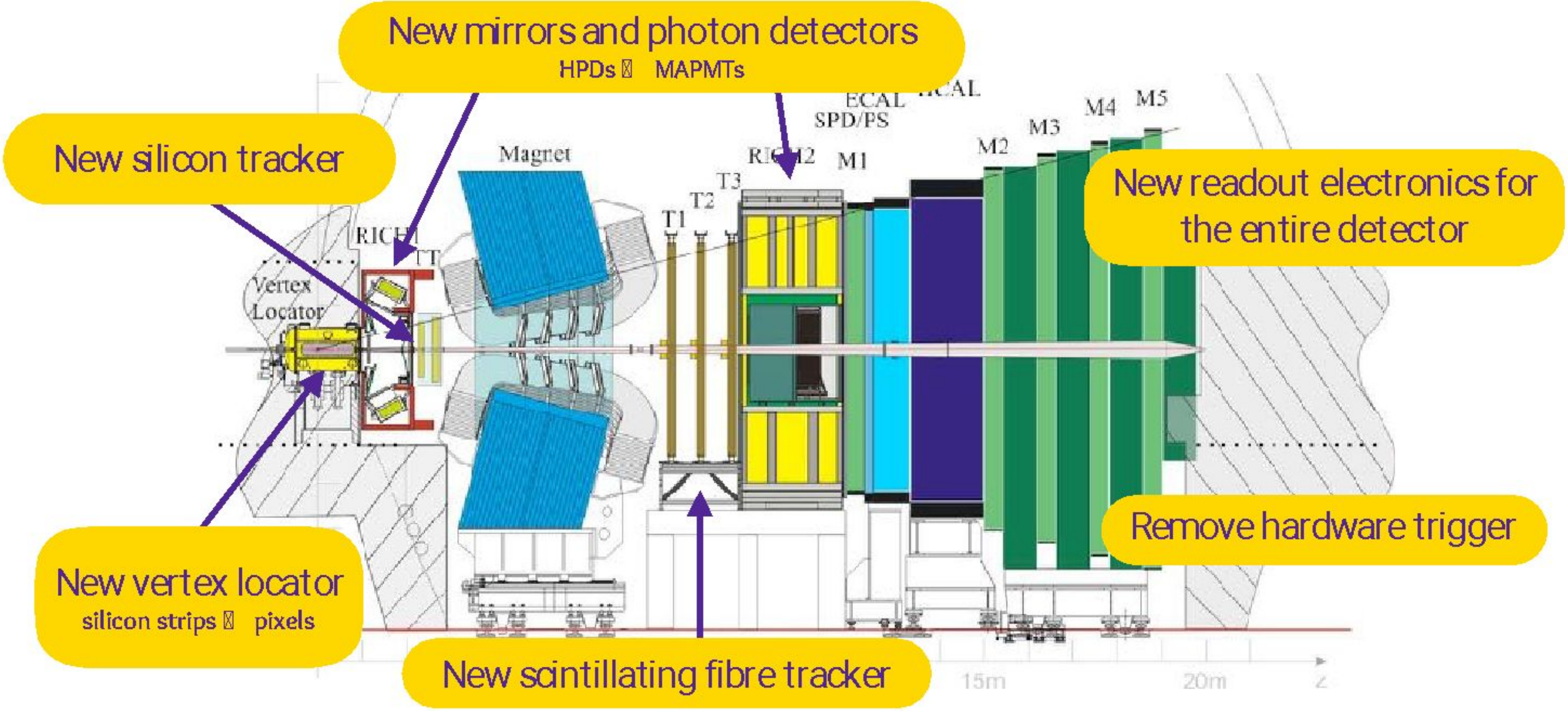
# The LHCb experiment



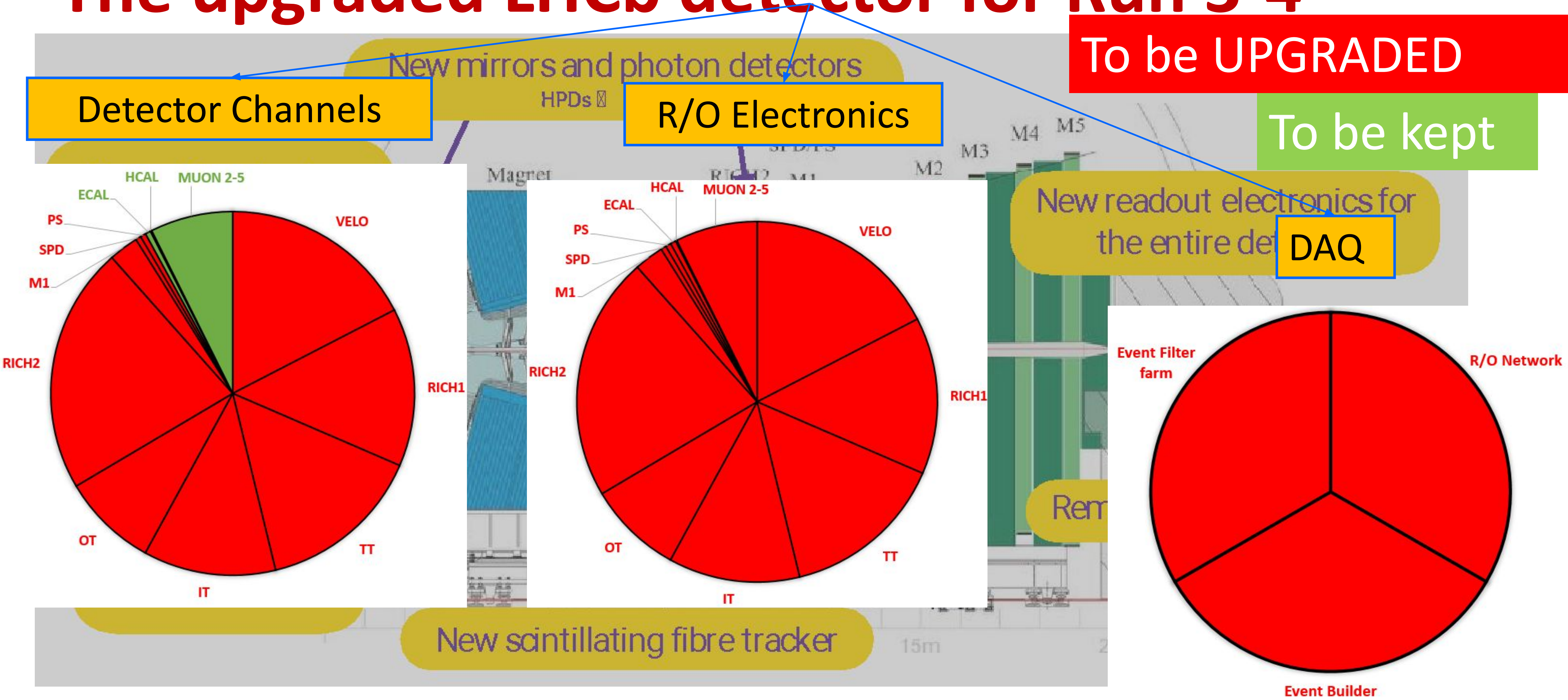

- Many of LHCb results obtained in Run1 and Run2 are dominated by statistical uncertainties
- An upgrade of LHCb has therefore been achieved to take data in Run3 and beyond



# The upgraded LHCb detector for Run 3-4



# The upgraded LHCb detector for Run 3-4



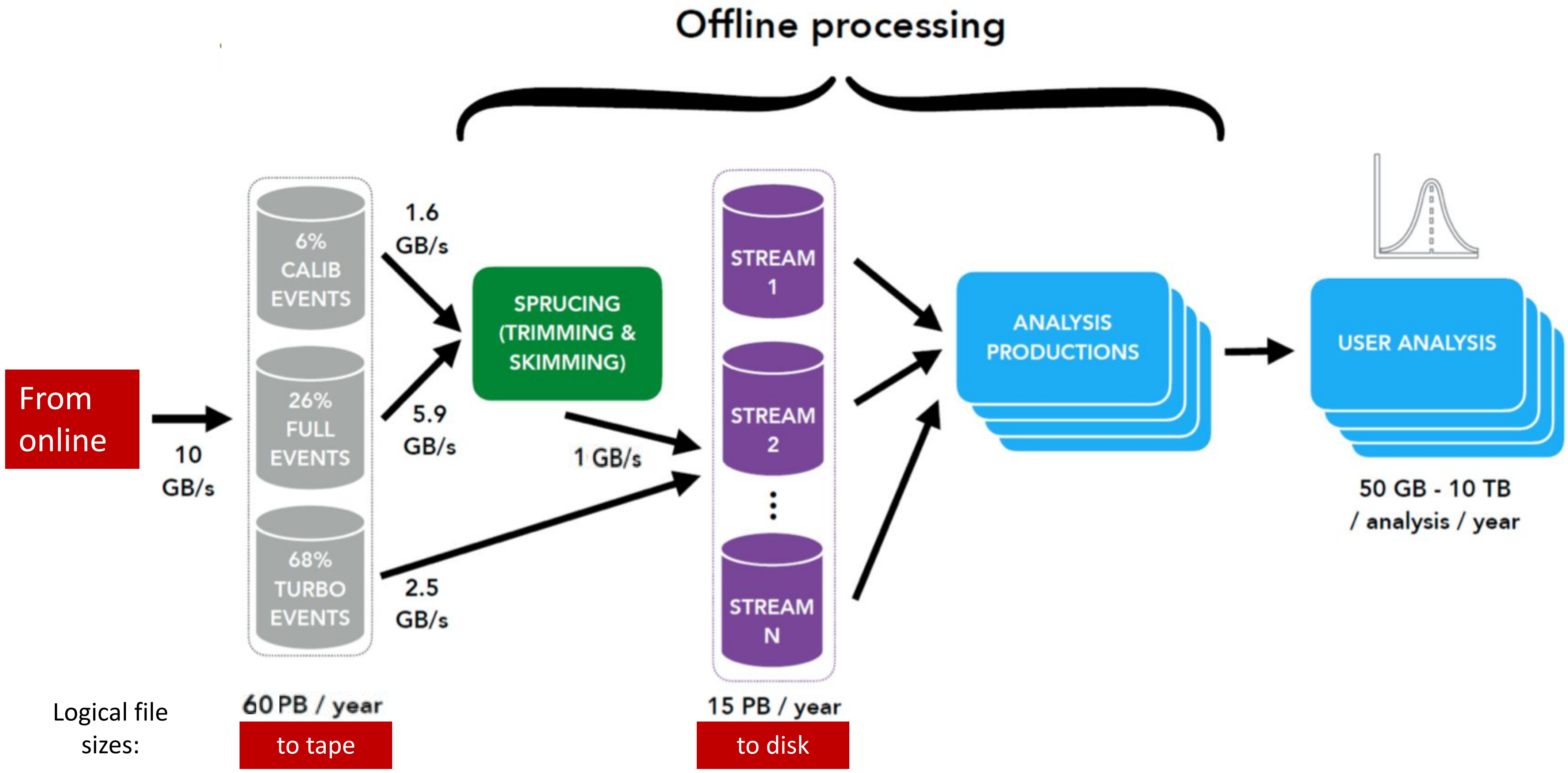
# A big challenge in data handling

- Major expansion of LHCb physics programme through:
  - 5-fold increase in instantaneous luminosity
    - $4 \times 10^{32}$ to $2 \times 10^{33}$ $cm^{-2}s^{-1}$
  - Full software trigger at 30MHz inelastic collision rate
    - Factor 2 increase in trigger selection efficiency
- Order of magnitude increase in physics event rate to storage
- Pile-up increase
    - Factor 3 increase in average event size
- 30x increase in throughput from the upgraded detector
  - Without corresponding jump in offline computing resources
- Full software trigger and selective persistency to mitigate throughput from online to offline
  - Nevertheless, from ~0.65GB/s (Run2) to 10GB/s (Run3-4)

# Data streams and dataflow

- Data from the LHCb detector organised in 3 streams; in all cases; events are reconstructed online at the HLT farm
  - **FULL:** «classic» stream, where information from the entire event is persisted in DST format and input to offline «sprucing» i.e. «slimming and skimming» for subsequent physics analysis
  - **TURCAL:** calibration stream, with both reconstruction output and (some) RAW banks. To be «spruced» offline and used for performance studies.
  - **TURBO:** introduced in Run2, implements selective persistency thus saving selected info that can range from a couple of tracks to the entire event contents. Data ready to be analysed, no further processing needed
- Sprucing is performed at T0 and T1s, concurrently with data taking and during winter shudown («re-sprucing»)
  - T0 for LHCb is equivalent to any other T1 from processing PoV
- Further processing (e.g. tupling) done in centralised Analysis Productions
- Additional analysis steps done on user / local resources

# Data streams and dataflow



# Run3 Computing model in a nutshell

- LHCb Upgrade computing model accommodates a trigger output BW of 10 GB/s
  - Massive usage of novel event selection (Turbo) and event size reduction (selective persistence) techniques
  - Save the full bandwidth on cheap storage (tape)
  - Reduce by a factor 3 disk requirements using the above techniques
- CPU needs dominated by MC production
  - Massive use of faster simulation techniques
- In summary:
  - Substantial reduction of expensive resources
  - Maintain the full breadth of the physics programme
  - Flexible: incorporate future technology advancements

| LHCb Run3 Computing Model assumptions | |
|---|---|
| L ($cm^{-2}\ s^{-1}$) | $2\times10^{33}$ |
| Pileup | 6 |
| Running time ($s$) | $5\times10^{6}$ ($2.5\times10^{6}$ in 2021) |
| Integrated luminosity | 10 $fb^{-1}$ (5 $fb^{-1}$ in 2021) |
| Trigger rate fraction (%) | 26 / 68 / 6 Full/Turbo/TurCal |
| Logical bandwidth to tape (GB/s) | 10 (5.9 / 2.5 / 1.6 Full/Turbo/TurCal) |
| Logical bandwidth to disk (GB/s) | 3.5 (0.8 / 2.5 / 0.2 Full/Turbo/TurCal) |
| Ratio Turbo/FULL event size | 16.7% |
| Ratio full/fast/param. MC | 40:40:20 |
| HS06.s per event for full/fast/param. MC [a] | 1200 / 400 / 20 |
| Number or MC events[b] | $2.3\times10^{9}$ / $fb^{-1}$ / year |
| Data replicas on tape | 2 (1 for derived data) |
| Data replicas on disk | 2 (Turbo); 3 (Full, TurCal) |
| MC replicas on tape | 1 (MDST) |
| MC replicas on disk | 0.3 (MDST, 30% of the total dataset) |

| Resource requirements | | | | | | |
|---|---|---|---|---|---|---|
| WLCG Year | Disk (PB) | | Tape (PB) | | CPU (kHS06) | |
| 2021 | 66 | 1.1 | 142 | 1.5 | 863 | 1.4 |
| 2022 | 111 | 1.7 | 243 | 1.7 | 1579 | 1.8 |
| 2023 | 159 | 1.4 | 345 | 1.4 | 2753 | 1.7 |
| 2024 | 165 | 1.0 | 348 | 1.0 | 3467 | 1.3 |
| 2025 | 171 | 1.0 | 351 | 1.0 | 3267 | 0.9 |

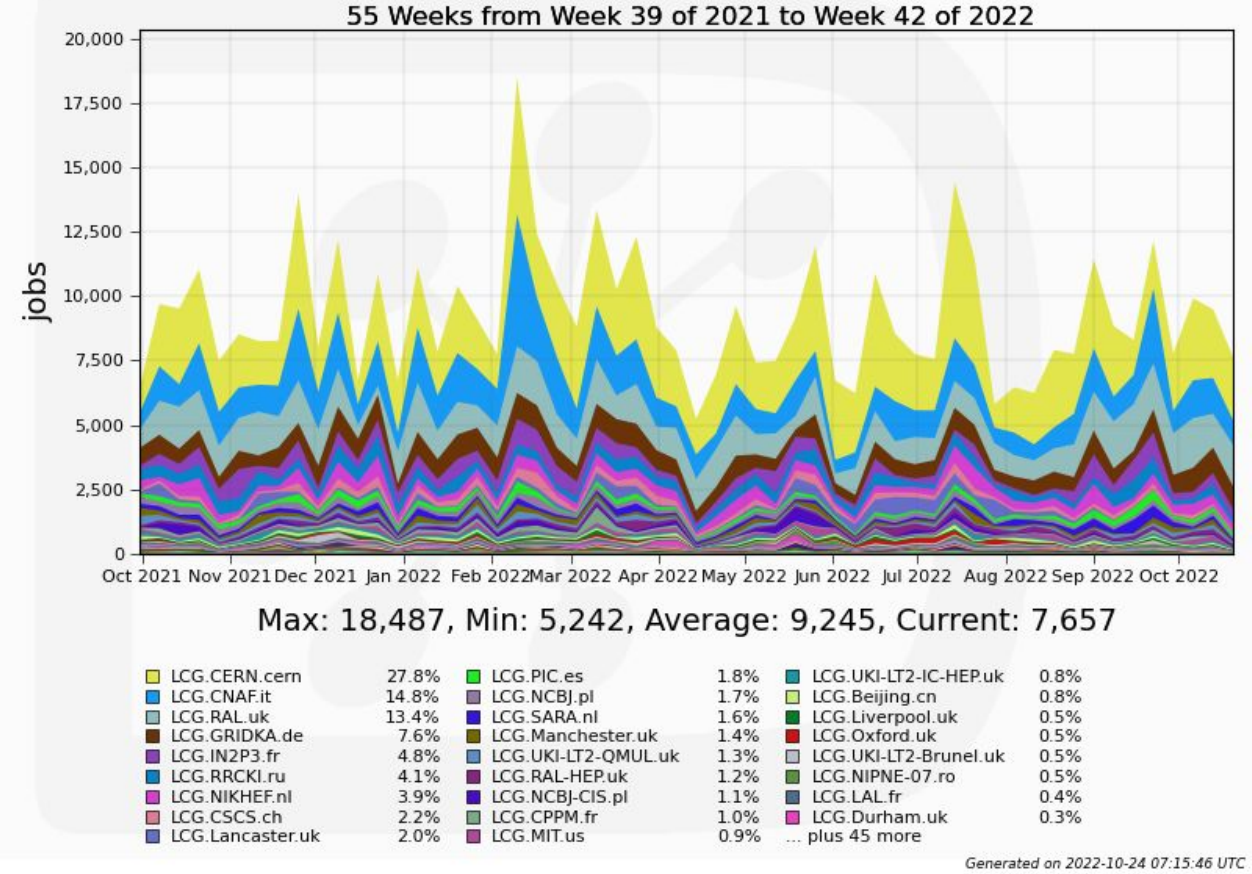
[a] corresponding to 120, 40, 2s on a 10HS06 computing core

[b] simulation of year N starts in year N+1

# Data distribution for physics analysis

- Data distribution model quite simple
- User jobs run where data is
  - Mostly at Tier0 and Tier1s
- Number of sites with data relatively small
  - 1 T0, 7 T1s, 14 T2-Ds
- Well-balanced CPU and disk resources
  - Grid user jobs are given the highest priority anyway
- No need for caches, pre-placement, etc
- Little impact on WAN other than dataset replication (2 copies)

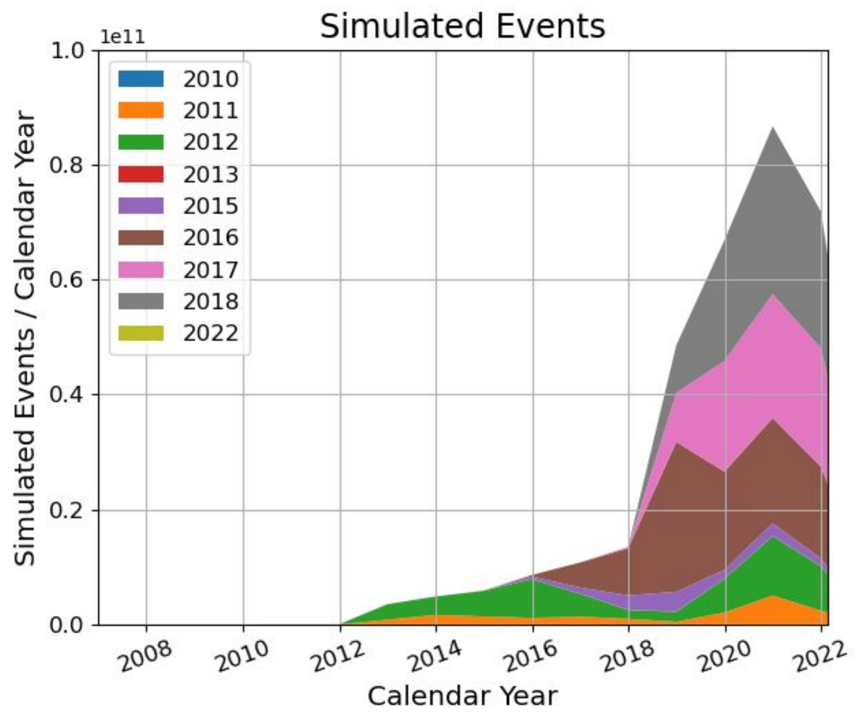
User jobs by Site



# Monte Carlo simulation

- No input data required. Starting from random seed!
  - Pile-up significantly smaller than GPDs
- Simulation dominates (95%) CPU work, runs everywhere
  - Improvements in simulation and introduction of fast simulation significantly decrease CPU work per event
    - big jump in number of simulated events per year!
- Simulation reconstruction is heavily filtered
  - E.g. 70-80B events simulated in 2021-2022 but much less stored
- Simulation is continuously running, with a given data-taking year being simulated for the following N years

| Data taking year | Simulation year | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | X | X+1 | X+2 | X+3 | X+4 | X+5 | X+6 | X+7 | X+8 | X+9 | X+10 |
| X | | | | | | | | | | | |
| X+1 | | | | | | | | | | | |
| X+2 | | | | | | | | | | | |
| X+3 | | | | | | | | | | | |

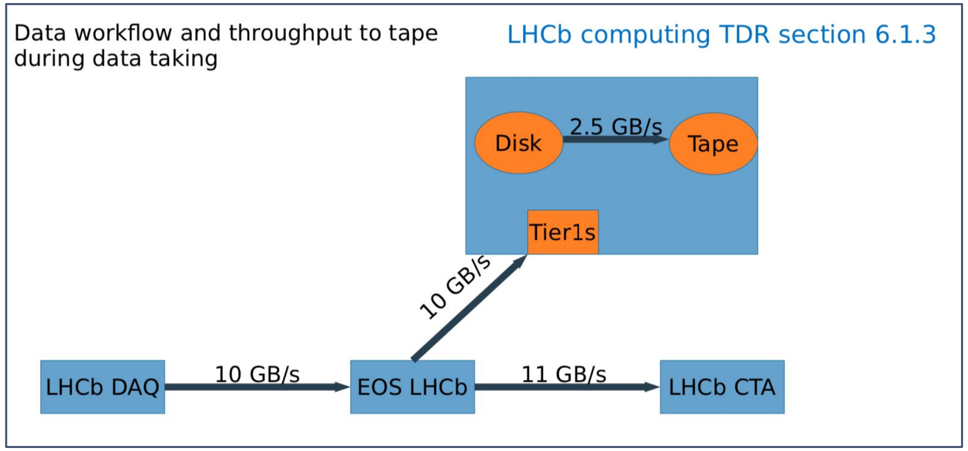


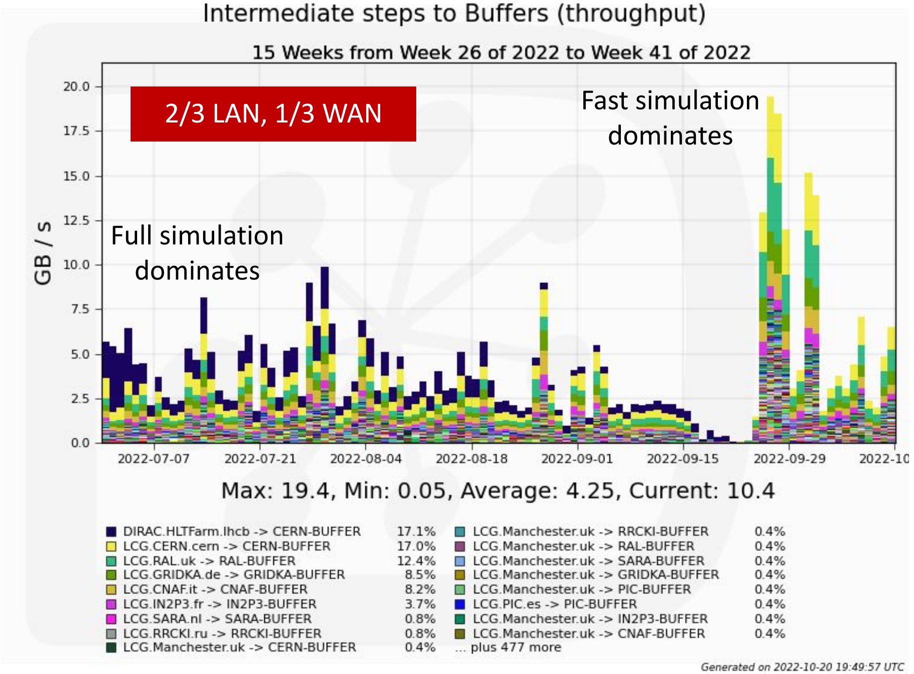

| Year | Simulated events ($10^9$) | Stored events ($10^9$) | Ratio | CPU work kHS06.y | CPU per event kHS06.s | LFS TB |
|---|---|---|---|---|---|---|
| 2017 | 10.3 | 4.2 | 40.3% | 817 | 2.50 | 640 |
| 2018 | 12.0 | 3.0 | 25.3% | 1009 | 2.65 | 550 |
| 2019 | 45.0 | 6.9 | 15.2% | 1290 | 0.90 | 1110 |
| 2020 | 67.0 | 16.8 | 31.7% | 1357 | 0.81 | 2010 |
| 2021 | 80.0 | 11.1 | 13.9% | 1815 | 0.72 | 2030 |
| 2022 | 78.4 | 3.2 | 4.1% | 2243 | 0.98 | 2490 |

More details given at LHCONE/LHCOPN workshop, Oct 25 2022

# Network

- LHCb increases network usage in Run3 and beyond
  - Dominated (one order of magnitude!) by real data coming from the detector
  - A factor two expected for simulation
    - Fast simulation requires more BW
- Run3 requirements have been successfully tested with data challenges in 2022
- Fast and reliable network is at the basis of our successful computing operations and ultimately of the physics productivity of LHCb
- In general:
  - we favour LAN over WAN
  - when running on a Tier2, we favour the national network before going abroad.



# Analysis productions

- The Analysis Productions infrastructure allows a user-friendly, declarative approach to ntupling
- user processing of data and simulation are supported using the DIRAC transformation system
- Historically analysts were responsible for running O(10,000) grid jobs to produce ROOT files
- Centralised production ensures e.g. better validation hence more efficient use of resources
- Job details / configuration / logs automatically preserved in LHCb bookkeeping / EOS
- Automated error interpretation / advice
- Intuitive web interface for requesting / testing / browsing outputs
- Integration of testing and monitoring using gitlab CI/CD - web based monitoring

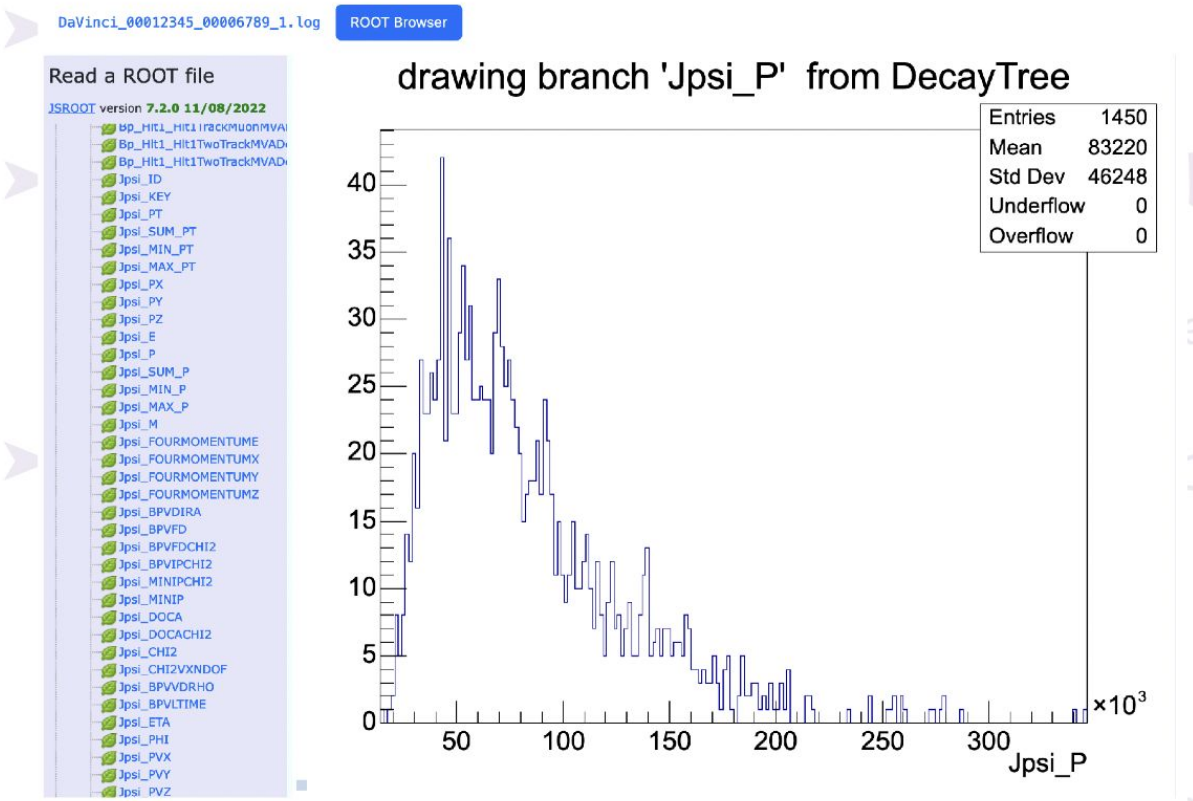
C. Burr, CHEP 2023 talk

```
MySample:
  wg: WG
  inform:
    - someone@cern.ch

  application: DaVinci/v45r4
  options:
    - make_ntuple.py
  output: DVNtuple.root

  input:
    bk_query: /some/bookkeeping/path/
```

1. What to do

2. What input data

*Also support chaining productions, e.g.*
*Run HLT1 then*
*use as input for HLT2 then*
*as input for ntupling*



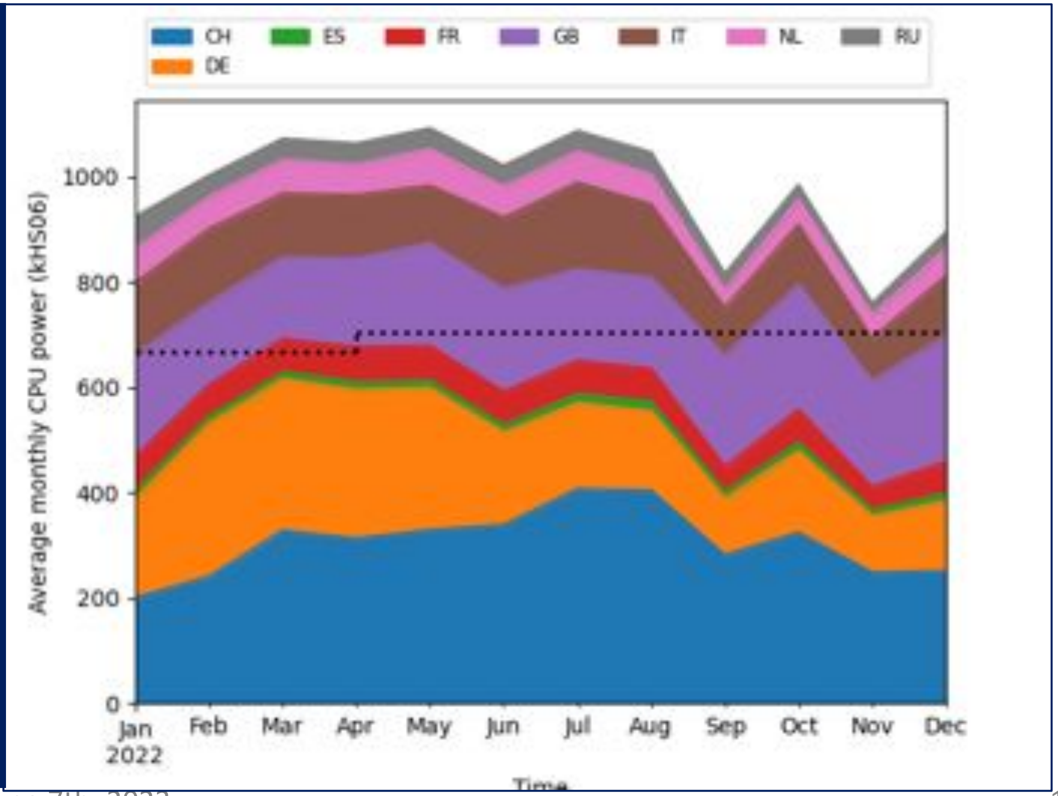
# Reality bites

- LHC schedule and installation of LHCb upgrade sub-detectors was slowed down by Covid-19 pandemics
- 2022 was a commissioning year for LHCb
- Upstream Tracker (UT) installed during Year-End Technical Stop (YETS 2022-2023)
  - Currently under commissioning
- A failure of the LHC vacuum system of the VELO resulted in an incident in the VELO vacuum volume on January 10th 2023
  - VELO «RF foil» deformed, precluding the possibility to fully close the VELO around the LHC beam in 2023
  - RF foil to be replaced in YETS 2023-2024
- 2023 is mostly commissioning for LHCb
  - Expecting to take heavy-ion collision data with all subdetectors included
  - Perhaps some proton collision data shortly before

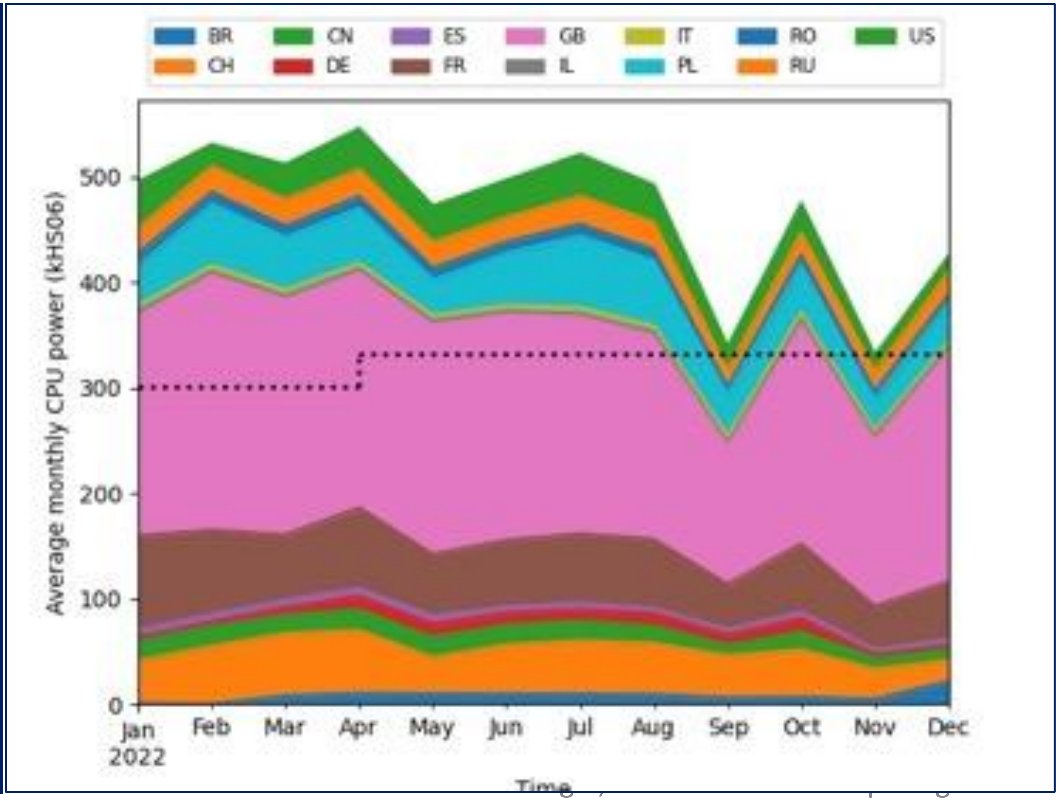
# CPU usage in 2022

- Nevertheless, CPU usage on WLCG resources has been above the pledges
  - Decreasing from Q3/Q4: no Run3 events to simulate, only Run1+Run2 "tails"

Tier0 + Tier1

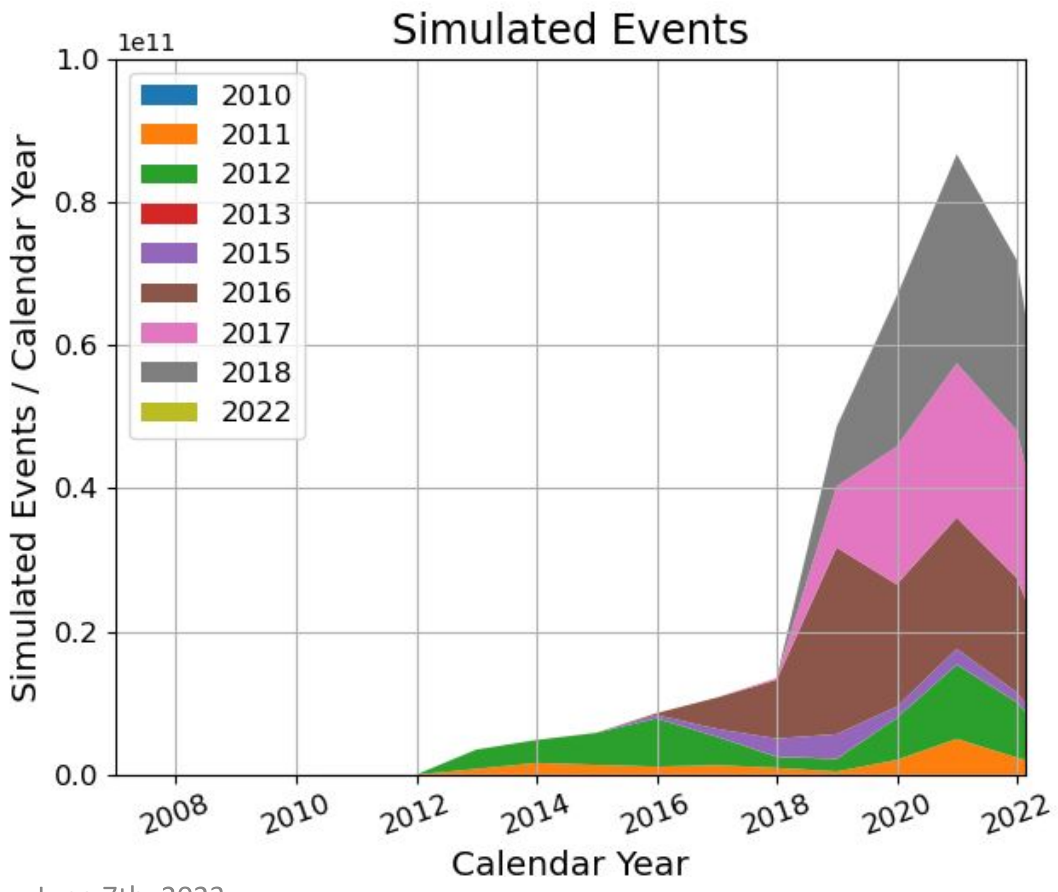


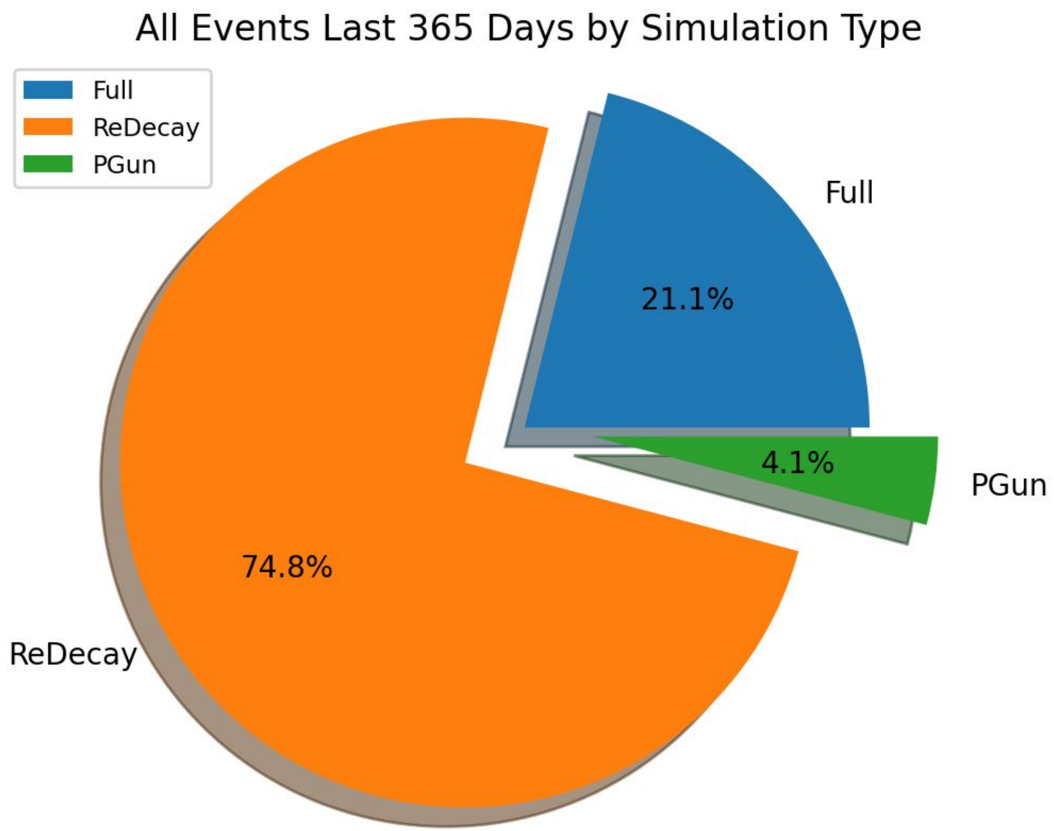

Tier2



# CPU usage in 2022

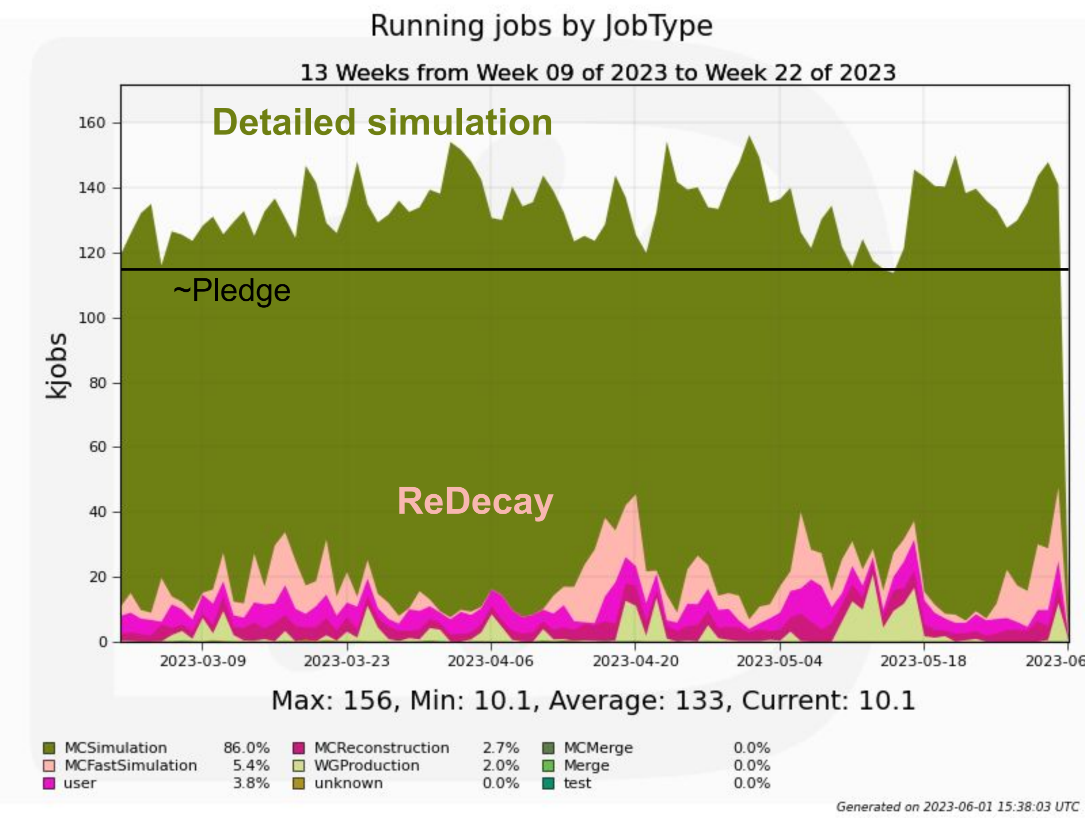
~75 billion events simulated in 2022, 80% with fast simulations

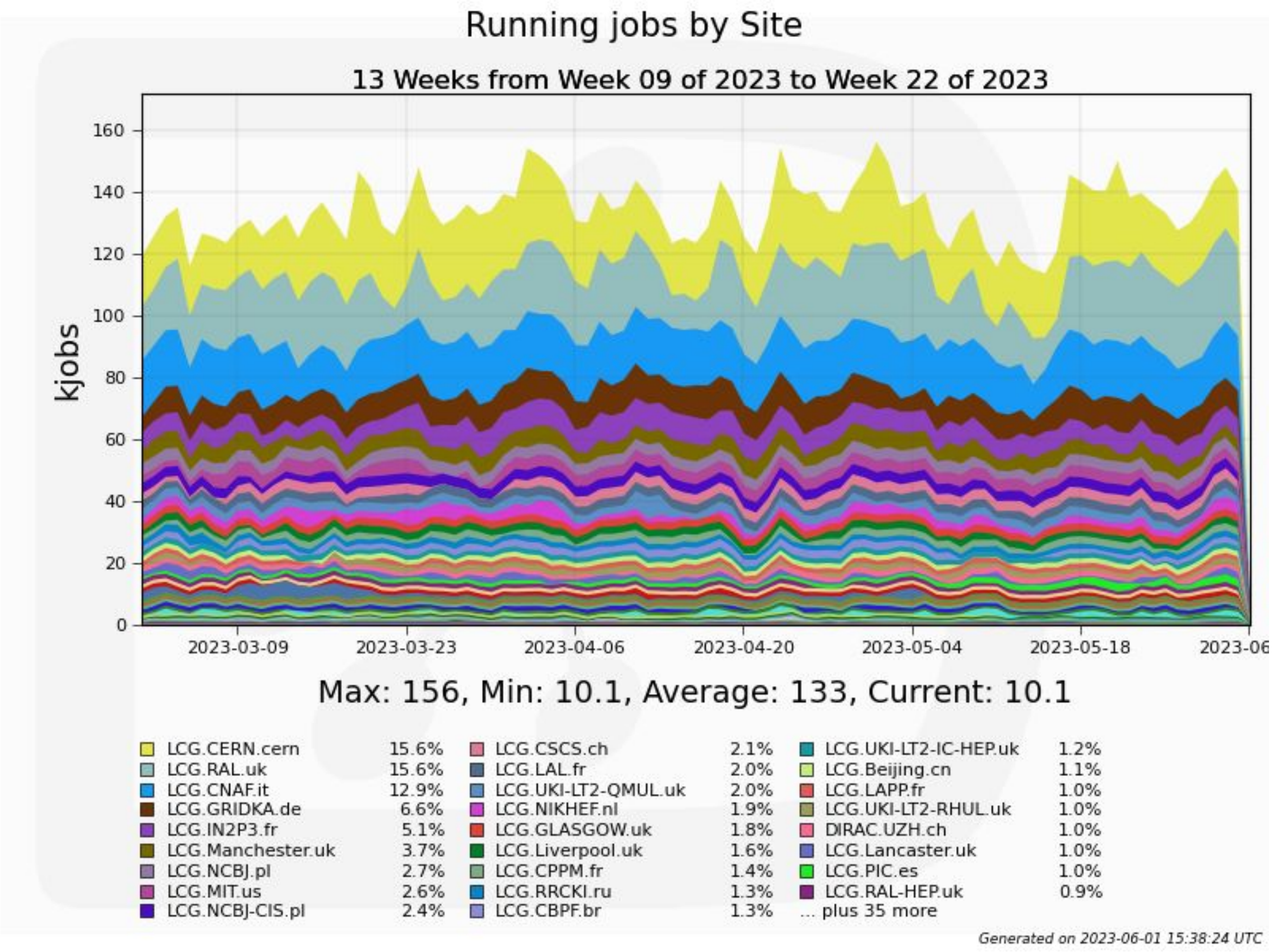


# Distributed computing operations

- Computing work: MC production (94%), physics analysis (6%)
  - 1/3 of physics jobs run through **Analysis Productions**

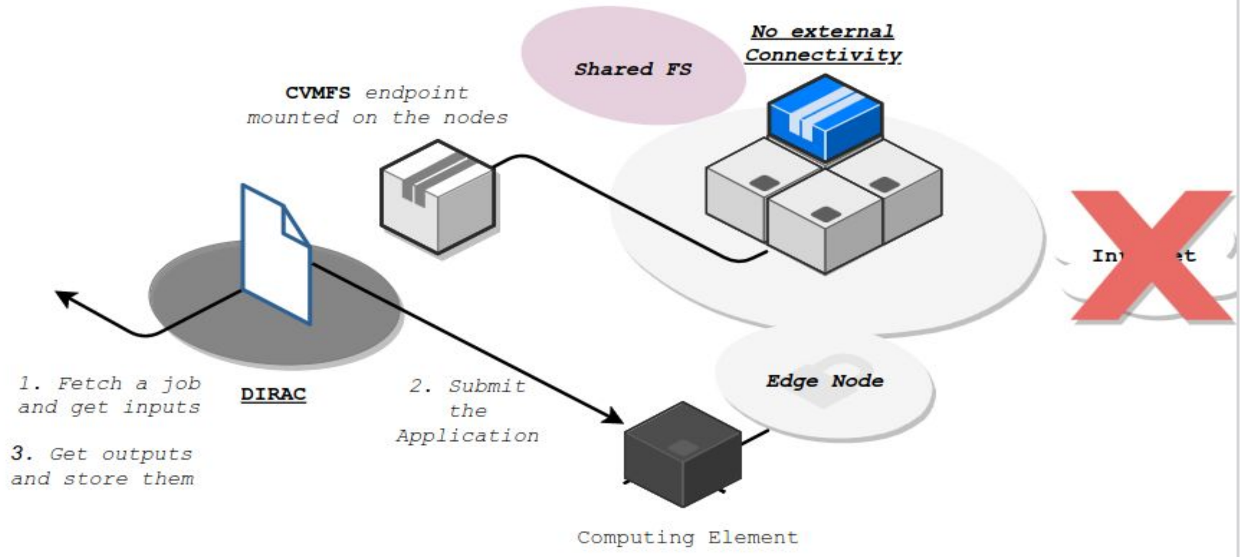


# DIRAC equipped to run on HPCs

***PushJobAgent***

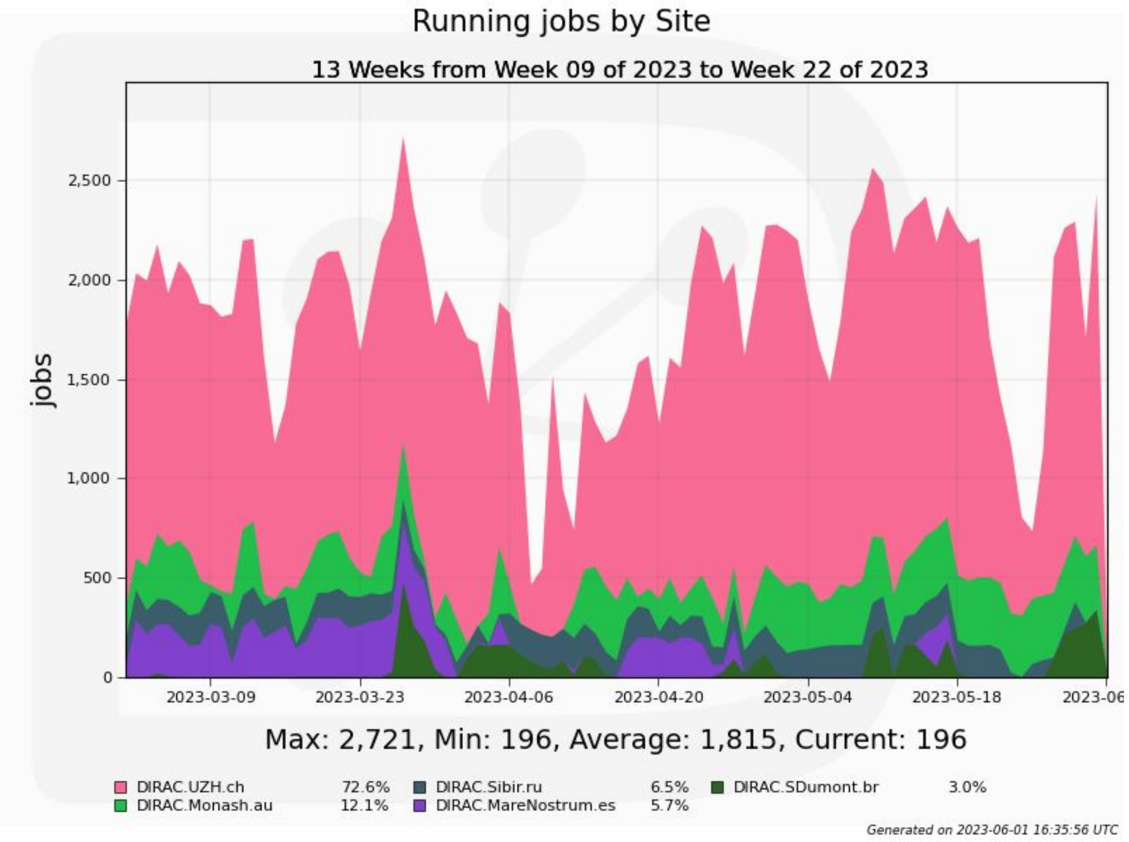
- Used in absence of network connectivity on the worker nodes
- Works as a Pilot-Job that would be executed outside of the HPC
- Fetches jobs, manages their input and output data, and solely submit the application to the HPC.
- Requires a direct access to the LRMS.




A. Boyer, CHEP talk

# Opportunistic resources

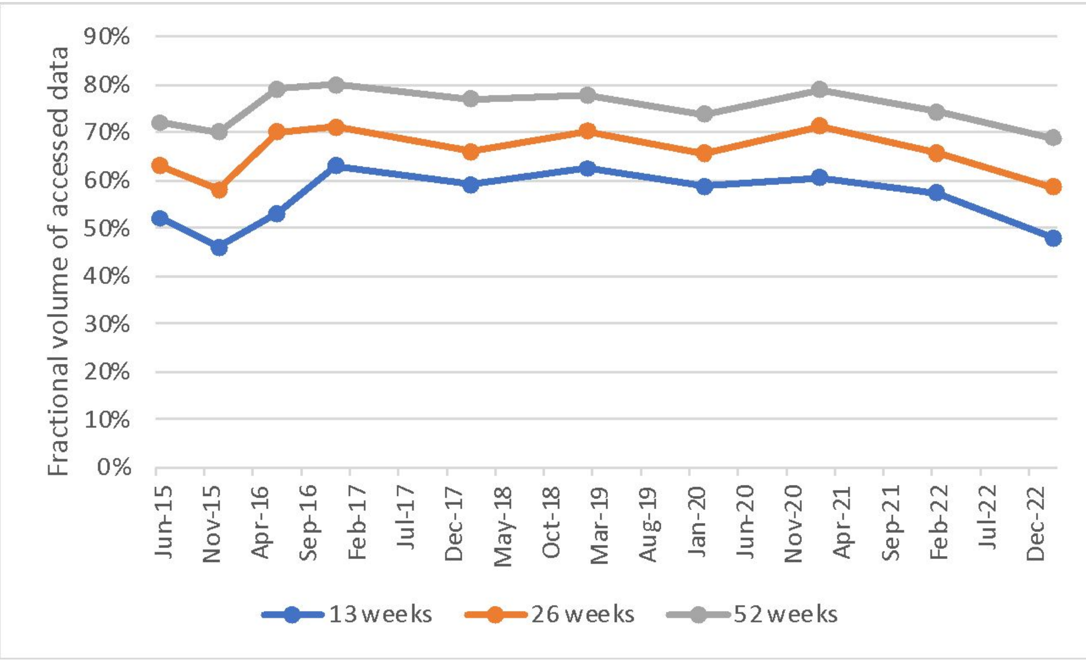
- Fraction of jobs executed on totally opportunistic resources stable at a few percent level
- Zurich (CH), Monash (AU), Sibir (RU), Barcelona (ES)
- Barcelona Supercomputing Center (MareNostrum) and SDumont.br are production HPCs
- WLCG sites not pledging to LHCb are also utilised opportunistically (at a few percent level)



# Storage (under)usage in 2022

- Disk and tape usage way below requirements/pledges/deployed capacities
- 2022 has been a commissioning year…
- Data popularity is also somewhat decreasing; Run1+Run2 analysis tails…

| LHCb | | 2022 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Request | Pledge | Pledge/CRSG | Used | Used/CRSG | Deployed capacity | Deployed cap./CRSG |
| WLCG CPU | Tier-0 | 189 | 189 | 100% | 305 | 161% | 189 | 100% |
| | Tier-1 | 622 | 515 | 83% | 676 | 109% | 515 | 83% |
| | Tier-2 | 345 | 333 | 96% | 470 | 136% | 333 | 96% |
| | HLT | 50 | 50 | 100% | 271 | 541% | n/a | n/a |
| | Sum | 1206 | 1086 | 90% | 1721 | 143% | 1036 | 86% |
| Others | | 50 | 50 | 100% | 53 | 105% | n/a | n/a |
| Total | | **1,256** | **1,136** | **90%** | **1,773** | **141%** | **1036** | **86%** |
| Disk | Tier-0 | 26.5 | 26.5 | 100% | 10.5 | 39% | 12.1 | 46% |
| | Tier-1 | 52.9 | 47.8 | 90% | 30.6 | 58% | 45.1 | 85% |
| | Tier-2 | 10.2 | 6.9 | 68% | 4.0 | 40% | 7.5 | 74% |
| | Total | **89.6** | **81.2** | **91%** | **45.1** | **50%** | **64.7** | **72%** |
| Tape | Tier-0 | 81 | 81 | 101% | 29.8 | 37% | | |
| | Tier-1 | 139 | 116 | 83% | 47.1 | 34% | | |
| | Total | **219.9** | **197.3** | **90%** | **76.9** | **35%** | | |



# LHCb software on ARM

- Gauss simulation application successfully built on ARM
  - Simulation takes 90% of CPU work on the grid
- Physics validation needed
  - LHCb performance & regression test suite (LHCbPR)
- Continuous integration / nightly builds a prerequisite
  - server available from CERN/IT
- LbPlatformUtils needs to be properly extended
  - An LHCb-developed library (but really generic) that
    - identifies the platform, of the node where it ran
    - used for matching of LHCb jobs
      - actually, very little of LHCb in there
    - handles compatibilities
    - finds out the possibility to use containers
    - recognizes the instructions set available

# Proto-Tier1 sites

NATIONAL CENTRE FOR NUCLEAR RESEARCH ŚWIERK

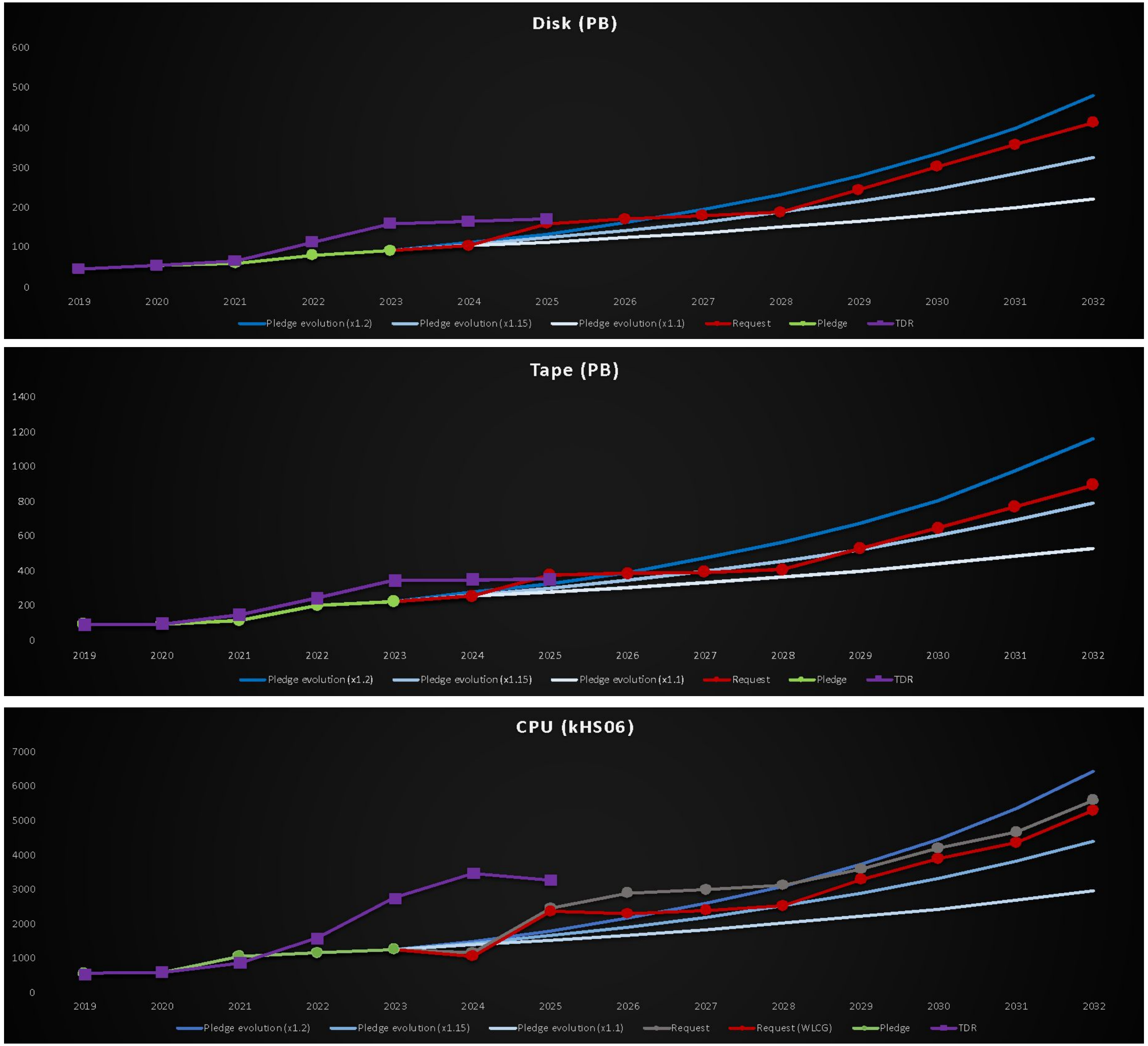
- The WLCG Overview Board (OB) approved on Dec 8th 2022 the plan presented by the NCBJ (Warsaw) and IHEP (Beijing) Tier2 sites to become Tier1 sites for LHCb
  - Contributing ~5% computing resources each
- Both sites must comply with the needed requirements in terms of network, storage (most notably: tape), services, service level agreement
- LHCb distributed computing team engaged to define tasks/deliverables/milestones/etc.
  - Minimal requirements shown at MB on Feb 14th 2023
- Current status: CPU and storage OK, NCBJ connected to LHCOPN, IHEP network to be finalised (in a couple of weeks)

Institute of High Energy Physics
Chinese Academy of Sciences

# Resource evolution in Run3 / LS3 / Run4

- Taking VELO incident into account
- Pledge evolution from 2023
- 2024 requests ~ 2023 pledges
- After step in 2025, requests are within 1.15-1.2x pledge evolution through Run4
- Two-years shift wrt TDR

Legend:
**Requirements**
**Pledges**
**TDR**
Evolution: 1.2x, 1.15x, 1.1x

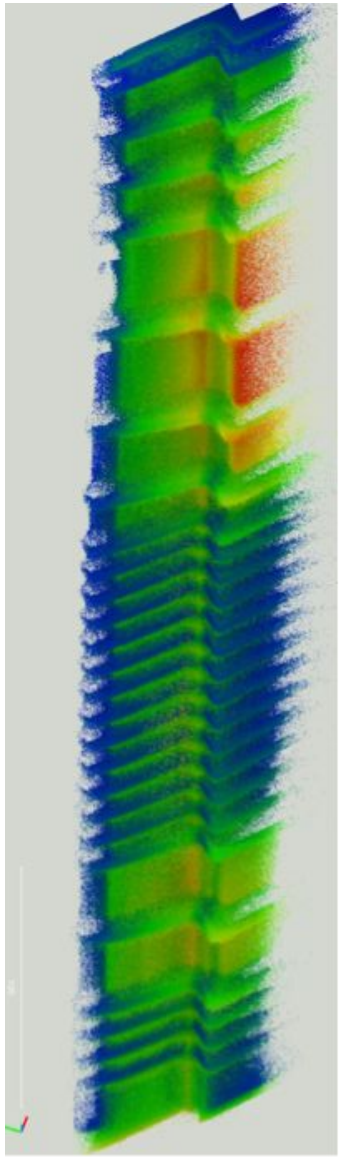


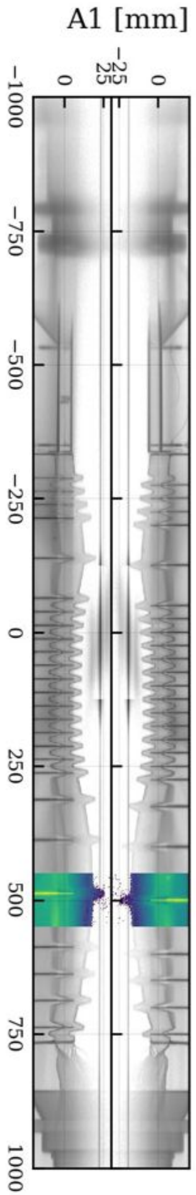
# Summary

- Run3 + Run4 computing model
  - 30x larger data volume from detector mitigated by aggressive triggering strategy, filtering, selective persistency
  - Network utilisation one order of magnitude larger than Run2
    - Requirements validated by data challenges in 2022
    - Still small wrt other LHC VOs
- Resource usage
  - CPU dominated by simulation production
  - Fast simulation significantly mitigates requirements
- Status so far: commissioning LHCb sub-detectors, waiting for data
- Offline resources:
  - Two-years delay with respect to computing model TDR
  - 2024 ~ 2023; big jump expected in 2025, then within “flat budget” in LS3 & Run4
  - Two new Tier1 sites helpful to alleviate pressure on storage (tape)

# backup

# VELO Vacuum Volume Incident : Status & Recovery






- On 10th January 2023 incident occurred due to a failure of the LHC vacuum system of the VELO.

- **Detector modules & cooling are not damaged**
- Currently operating with VELO in retracted position
- Expectation to close to 16mm (where 0mm is nominal position)

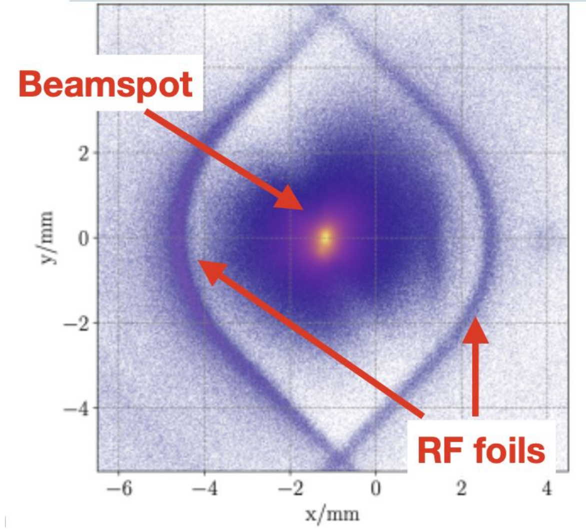
  after June Technical Stop, subject to checks

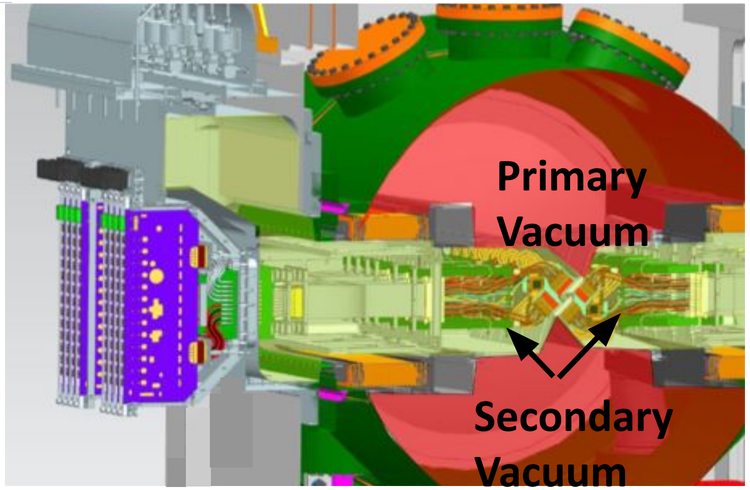
- **RF foil has undergone plastic deformation**
- Replacement in the shutdown at the end of 2023
- Planning advanced

- Commissioning of Upgrade I systems proceeds and physics opportunities in ‘23 remain

# VELO incident

RF foils imaged in 2022

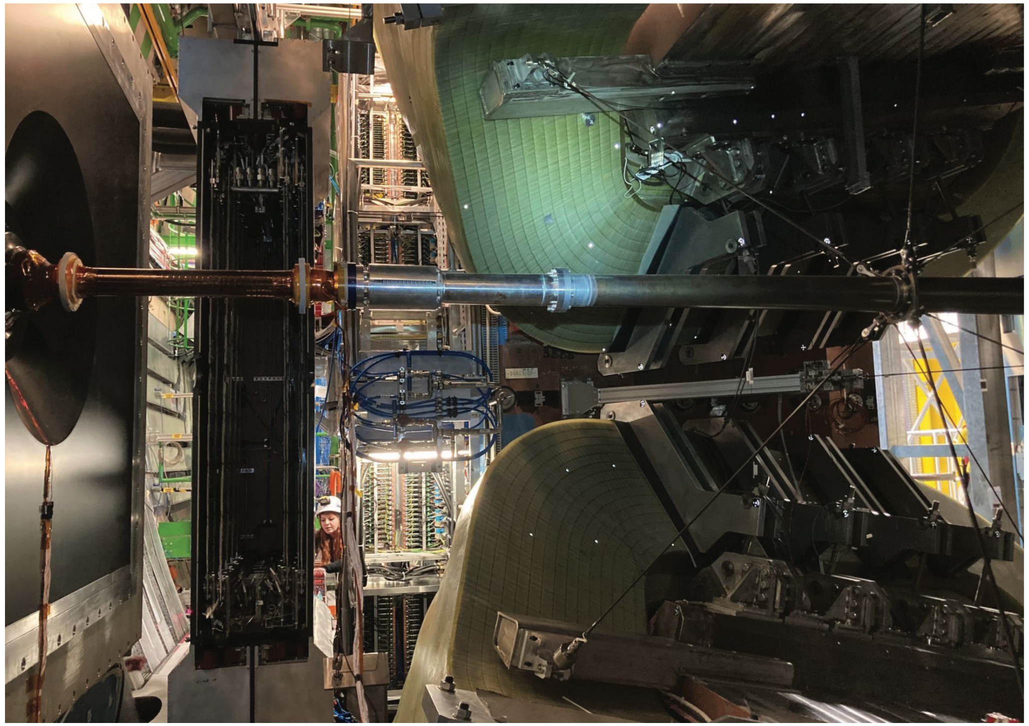



- Damage of the RF box between VELO and Primary Vacuum 10/1/23
  - multiple equipment failures resulted in a build up of pressure beyond specification between VELO and beam volumes
  - RF foils have been deformed. VELO modules do not show damage
  - Foils to be replaced in year end shutdown

- System expected to be operating with the VELO in an open or partially open position.
- Reduced physics programme though many opportunities are being explored to utilise the system.
- Significant impact on offline computing requests

# UT Installation Status

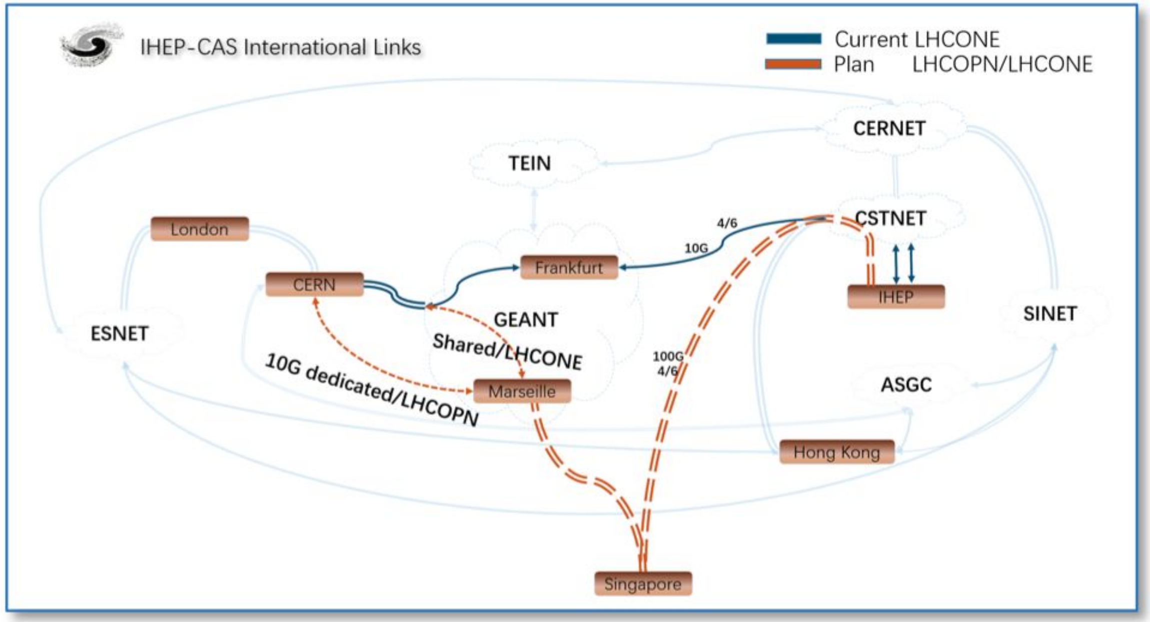


- UT was closed in time for start of LHC run, shortly after we last met
- 95.9% included readout channels
- Commissioning proceeding as anticipated
  - Will take time
  - Particular emphasis on firmware and software work in this period

# Current situation (NCBJ)

- Network: Connection with LHCOPN established on May 25th
  - testing/fixing things (e.g. some issues with IPv6 and routing to other T1s)
  - everything should be ready for data challenge very soon
    - targeting next week
- Computing hardware and configuration ready.
  - two additional servers added, disk capacity is now 3PB.
  - Purchase of tape drives is currently postponed
    - no immediate need for the additional throughput
    - existing hardware is sufficient to meet the requirements.
- LHCb team visited NCBJ on March 28-30 and was impressed by infrastructure and effort

# Current situation (IHEP)




## Computing:

- Hardware is ready
  - 3216 CPU cores with 63.5k HS06 (or 67.116 HS23)
  - 1280 Intel + 256 AMD CPU cores newly purchased
  - 1680 Intel CPU cores from current LHCb T2
- Software is ready
  - HTCondor CE deployed on newly purchased server

## Storage:

- Hardware is ready
- Disk storage capacity of 3.2PB
- Tape library is ready
  - 4 tape drivers and 3PB tape capacity.
  - Additional 7PB will be purchased this year
- Software: almost ready
  - EOS ready soon (end point root://eoslhcb.ihep.ac.cn)
  - CTA is ready (end point root://ctalhcb01.ihep.ac.cn)
    - Throughput of 4-6GB/s, ~3x requirement

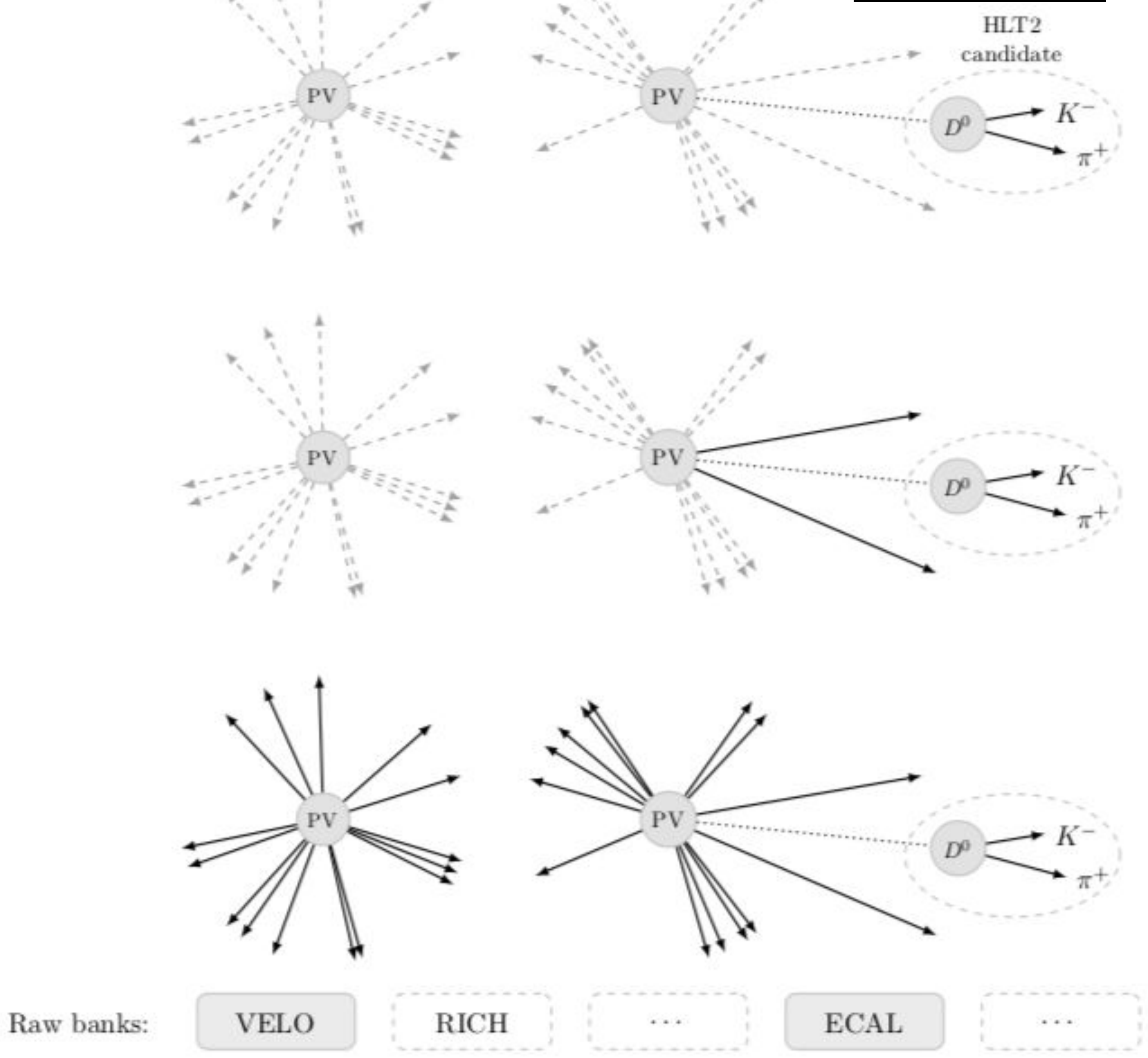
## Network:

- 100Gbps link (CSTNet - GEANT) will be ready on 15th June (instead of April)
- dedicated 10Gbps link (LHCOPN, Marseille-CERN) will be ready on 15th June

# Data persistency

LHCb-TDR-018

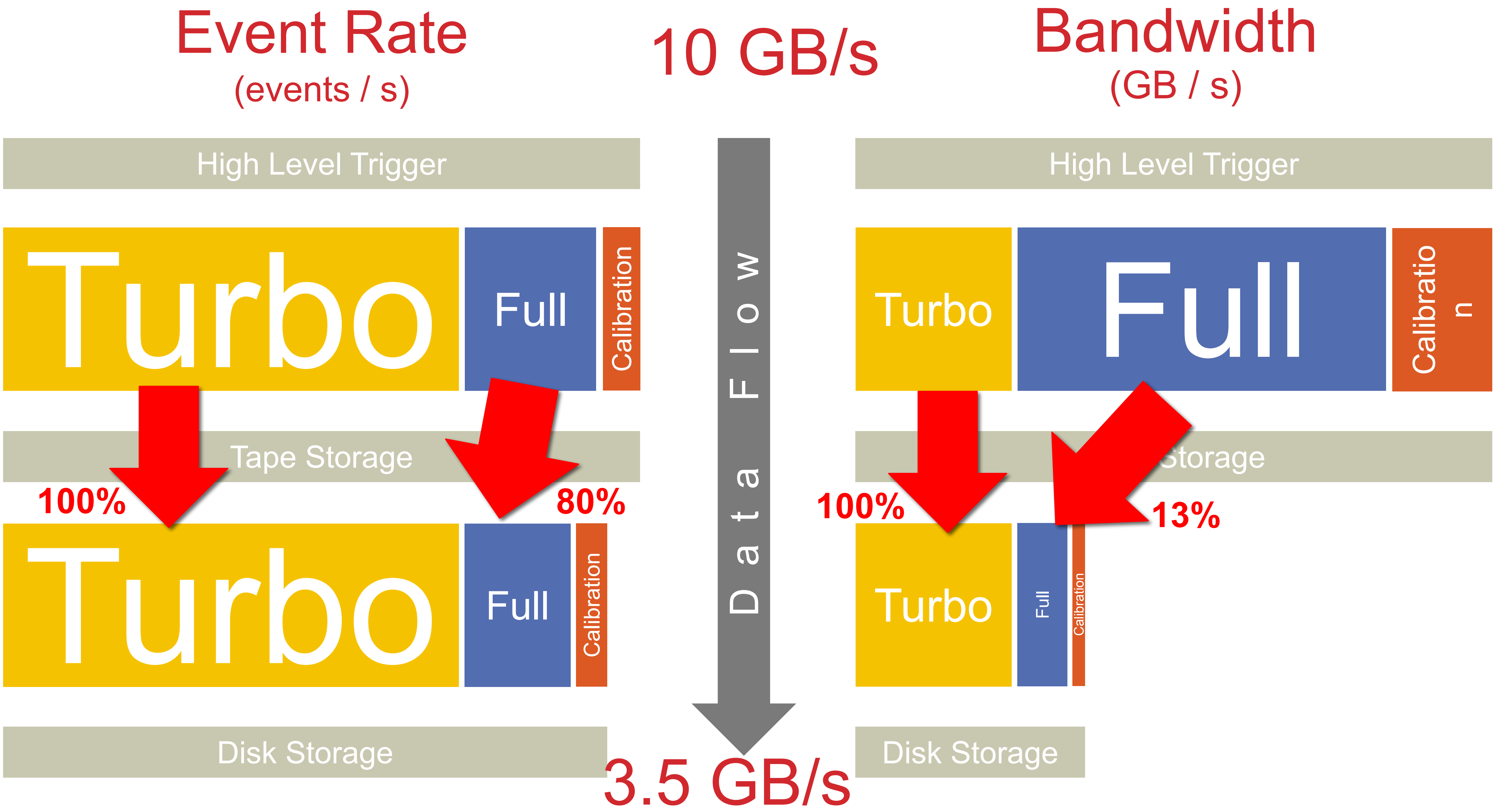
- Different levels of persistency:
  - FULL and TURCAL: the full event is persisted
  - TURBO: selective persistency, ranging from candidate firing the trigger to the entire event, optionally including some RAW subdetector data banks



# HLT output bandwidth

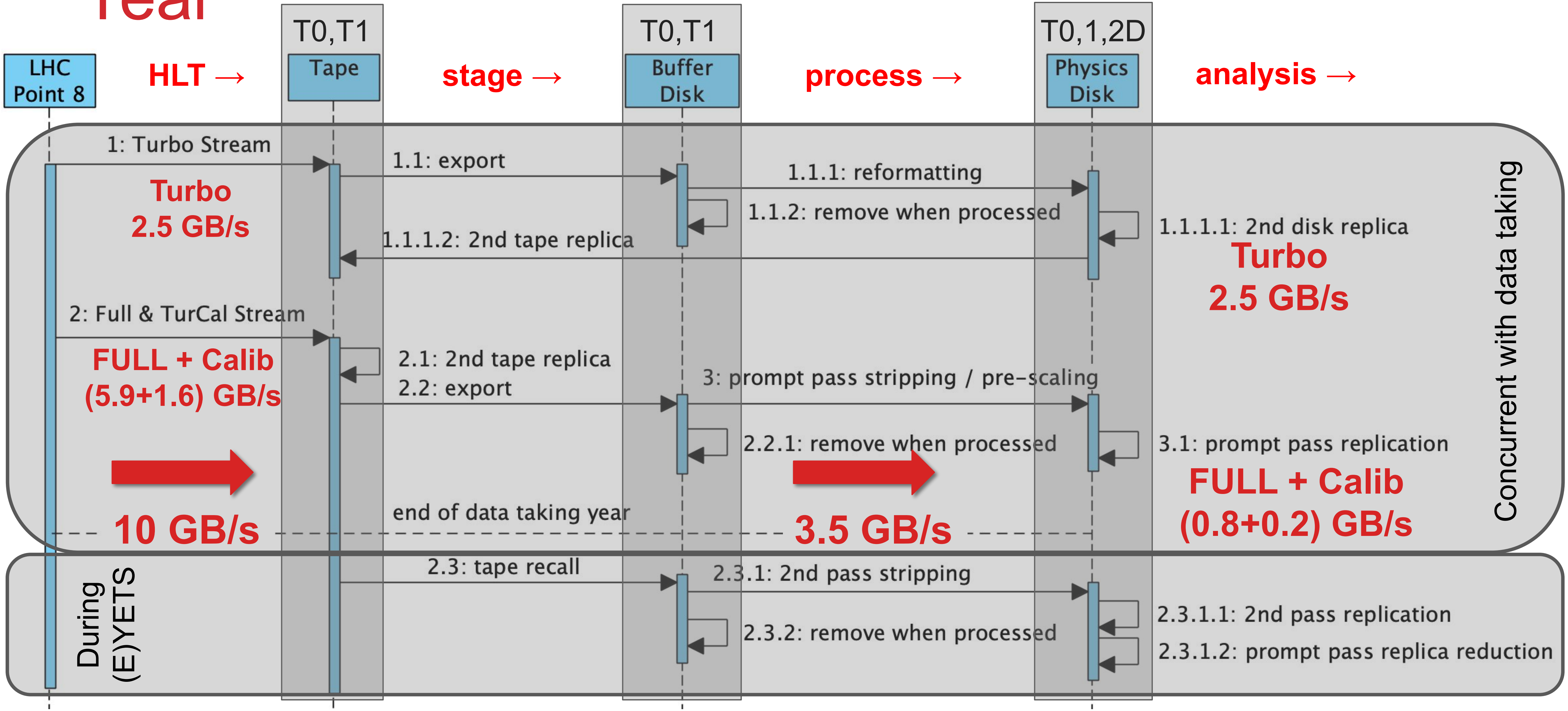
- Due to selective persistency, emphasis has shifted from trigger rate (Hz) to bandwidth (bytes/s)
  - save less information and give more rate for a given bandwidth!
- About 60% of the physics selections on FULL in Run2 are migrating to TURBO in Run3
  - Massive migration, not trivial!
- Logical bandwidth to tape: 10 GB/s
- Logical bandwidth to disk reduced to 3.5GB/s by sprucing FULL and TURCAL more aggressively (select substantial fraction but slim by factor 6)
- This gives requirements of O(100PB) tape and O(50PB) disk per data taking year

| | | Logical Throughput to tape | | Logical Throughput to disk | |
|---|---|---|---|---|---|
| stream | rate fraction | throughput (GB/s) | bandwidth fraction | throughput (GB/s) | bandwidth fraction |
| FULL | 26% | 5.9 | 59% | 0.8 | 22% |
| Turbo | 68% | 2.5 | 25% | 2.5 | 72% |
| TurCal | 6% | 1.6 | 16% | 0.2 | 6% |
| total | 100% | 10.0 | 100% | 3.5 | 100% |



Event Rate
(events / s)
10 GB/s
Bandwidth
(GB / s)
High Level Trigger
Turbo
Full
Calibration
Tape Storage
100%
80%
Turbo
Full
Calibration
Disk Storage
Data Flow
3.5 GB/s
High Level Trigger
Turbo
Full
Calibratio n
Storage
100%
13%
Turbo
Full
Calibration
Disk Storage

# Data Processing Workflow per Data Taking Year



# Data streams from the LHCb detector

- Due to selective persistency, emphasis has shifted from trigger rate (Hz) to bandwidth (bytes/s)
  - save less information and give more rate for a given bandwidth!
- Example of rate and bandwidth division for 2018 data taking

| stream | event size (kB) | event rate (kHz) | rate fraction | throughput (GB/s) | bandwidth fraction |
|---|---|---|---|---|---|
| FULL | 70 | 7.0 | 65% | 0.49 | 75% |
| Turbo | 35 (*) | 3.1 | 29% | 0.11 | 17% |
| TurCal | 85 | 0.6 | 6% | 0.05 | 8% |
| total | 61 | 10.8 | 100% | 0.65 | 100% |

(*) Turbo event size is an average. It ranges from a few kB (minimal persistence) to full event size

# Reconstruction / Stripping

- Reconstruction of FULL is performed at Tier1s (80% of events) and Tier0 (20%)
  - Output as RDST files
  - saved on tape ARCHIVE (1 copy only)
- TURBO does not need to be reconstructed, but only reformatted. Same T0/1 share
- No event re-reconstruction!
  - Alignment and calibration performed online on the trigger farm and applied on HLT
- RDST files are «stripped» according to selection criteria specific to each analysis. Stripping takes place at the same site as reconstruction. Output as
  - DST: full event information; stripping = event filtering
  - mDST: selective persistency; stripping = filtering + slimming
    - The offline equivalent of the TURBO stream
- (m)DST files are merged and grouped in O(10) streams and
  - Stored on tape ARCHIVE (1 copy) and DST disk
  - Replicated to DST disk on either another Tier1 or a Tier2 with disk (Tier2D)
    - 3 copies in total

# Run3 LHCb Upgrade

- With the upgrade conditions several factors need to be applied
  - Luminosity $4*10^{32}$ $cm^{-2}s^{-1}$ to $2x10^{33}$ $cm^{-2}s^{-1}$
  - HLT efficiency increase because of removal of L0 hardware trigger
  - Raw event size increase due to pileup, according to simulation
- Without any changes the HLT output rate would increase in Run 3 to 17.4 GB/s

| | Run 2 (GB/s) | Lumi | No L0 | Raw size | Run 3 (GB/s) |
|---|---|---|---|---|---|
| Full | 0.49 | x5 | x2 | x3 | 14.7 |
| Turbo | 0.11 | x5 | x2 | x1 | 1.1 |
| Calibration | 0.05 | x5 | x2 | x3 | 1.6 |
| Total | 0.66 | | | | 17.4 |

Event size:
Turbo/FULL ~0.167

# General considerations

- The current trend of allowing users to run docker/singularity images could impact network utilization, since this ultimately requires downloading the image on each worker node. That should be carefully thought about

- In terms of features, could a minimal QoS per user of the network be introduced?
  - Network is the only resource for which there is no pledge nor fairshare

- Our main concern for network in future is bandwidth availability
  - Non-LHC users are coming with large requirements

- Monitoring and performance: we regard ourselves as just users; of course we are willing to help with requirements/use cases/providing info/etc

# Monte-Carlo production in Run3 onwards

- Amount of events to be simulated scales with integrated luminosity
- Limit CPU by increasing usage of fast simulations
  - But this has a big impact on network traffic
- Limit storage and network usage by
  - Filtering in generation and stripping
  - Saving output in mDST format
- As a result, expect to generate a volume of O(10 PBs) of simulated data per year
  - 1/3 is kept on (MC_DST) disk, the rest is parked on tape
  - One disk replica is made, this gives an estimation of O(1 GB/s) network traffic
- If MC reconstruction is split and fast simulation dominates, then transfers of simulation output from Tier2 sites becomes dominant
  - O(5-10GB/s) as a ballpark estimate, to be further discussed

# Resource evolution in Run3 / LS3 / Run4

- Taking VELO incident into account
- Pledge evolution from 2019
- 2024 requests ~ 2023 pledges
- After step in 2025, within 1.2x pledge evolution through Run4

Legend:
**Requirements**
**Pledges**
Evolution: 1.2x, 1.15x, 1.1x