ATLAS Data Flows & Rucio

Mario.Lassnig@cern.ch

DIRAC & Rucio Workshop
2023-10-16
https://indico.cern.ch/event/1252369

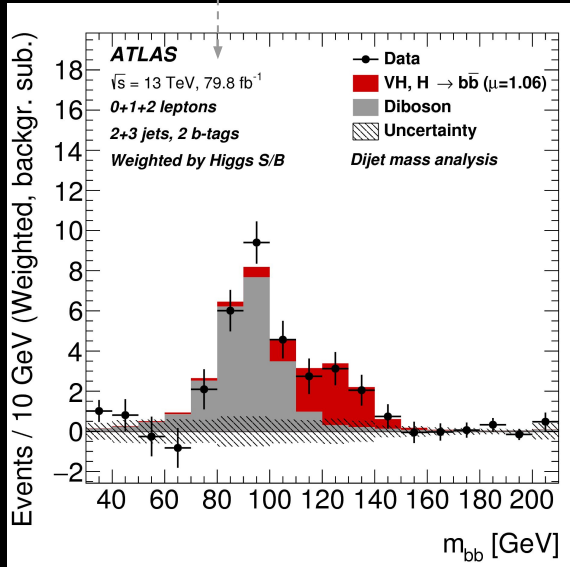
ATLAS EXPERIMENT

ATLAS EXPERIMENT

Candidate Event:
$pp \rightarrow H(\rightarrow b\bar{b}) + W(\rightarrow \mu\nu)$
Run: 338712 Event: 335908183
2017-10-19 23:31:18 CEST

13 TeV detector data
8 quadrillion collision candidates
92 petabytes
130 million files

13 TeV simulation data
166 petabytes
544 million files



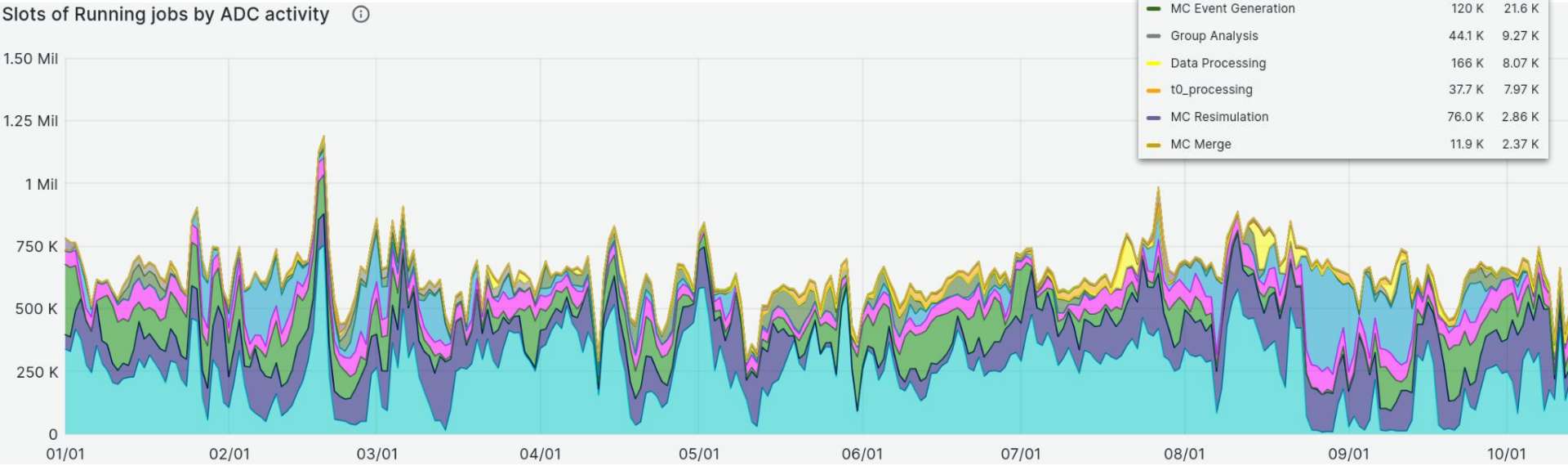
A candidate event display for the production of a Higgs boson decaying to two b-quarks (blue cones), in association with a W boson decaying to a muon (red) and a neutrino. The neutrino leaves the detector unseen, and is reconstructed through the missing transverse energy (dashed line). (Image: ATLAS Collaboration/CERN)

ATLAS computing usage

Global high-throughput computing system

Steady 500'000+ running jobs, with **full spread** of **experiment activities**

Spread across ~250 clusters **worldwide**



ATLAS computing usage

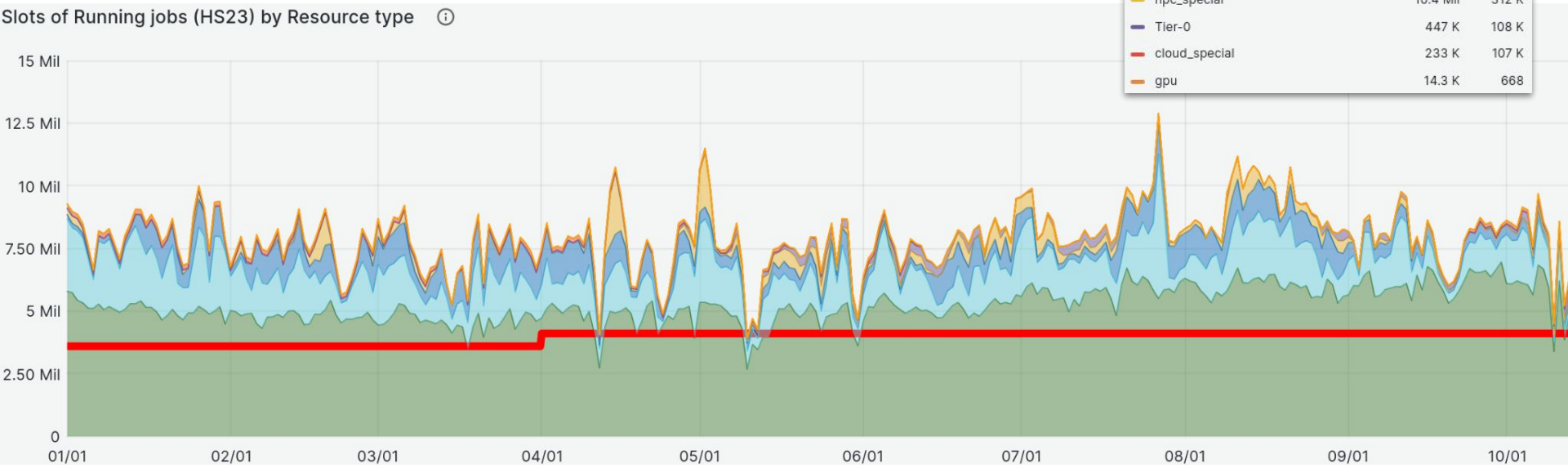
Computing power expressed in terms of HEPSPEC benchmark

1 modern x64 core ≈ 10 HEPSPEC

Infrastructure is **consistently over pledge**

Opportunistic resources

Allows **scale out** to 1+ million jobs



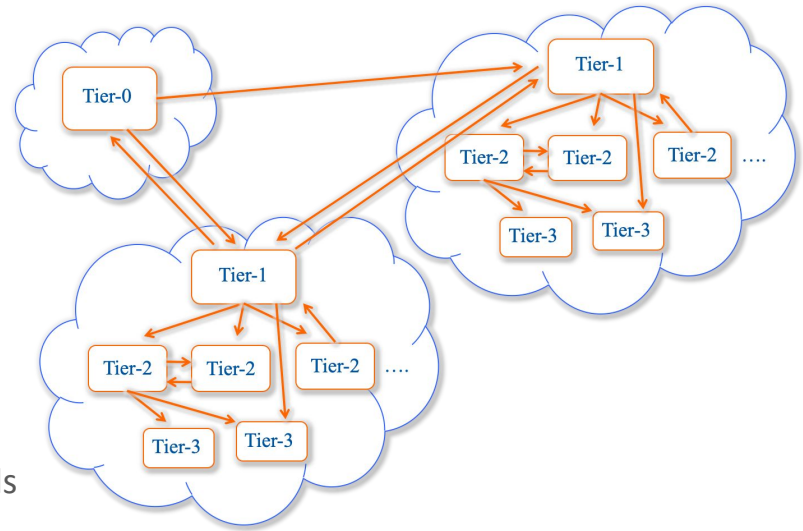
Basic experiment data flows 1/2

Original ATLAS computing model designed as static **clouds**

- ATLAS Clouds ≠ “Cloud computing”
- Mostly national or geographical **groupings of sites**
- **Common funding** agencies
- Support often using the **same language**

Model had a series of shortcomings

- Individual tasks **inflexibly executed** within a static cloud
- All tasks **output aggregated** at the 10 Tier-1s
- The **Tier-2 storage** was not optimally exploited
- **High priority tasks** were **occasionally stuck** at small clouds



Basic experiment data flows 2/2

WLCG networks have evolved significantly in the last decades

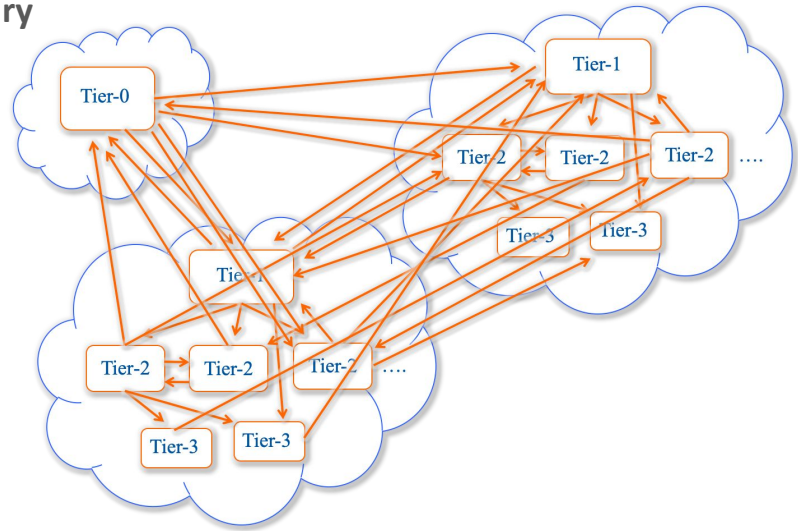
- **Limiting transfers** within a single cloud **no longer necessary**
- Now single **WORLD cloud** site concept

Nucleus

- **Any stable site** can aggregate the output of a task
- Site **can be manually assigned** as a nucleus

Satellites

- **Process the jobs** and **send the output** to the nucleus
- **Defined dynamically** for each task
- **No longer confined** inside the original cloud



Currently around **130 active sites** used by ATLAS

Experiment job types

Global shares are employed to allocate the available resources among the activities

Done on **agreement** between the various production and physics groups

Hierarchical implementation

Related activities have the opportunity to **inherit unused resources**

Essentially two categories of jobs

Production Data reprocessing

Event generation / Simulation / Reconstruction

Group production

Analysis User analysis

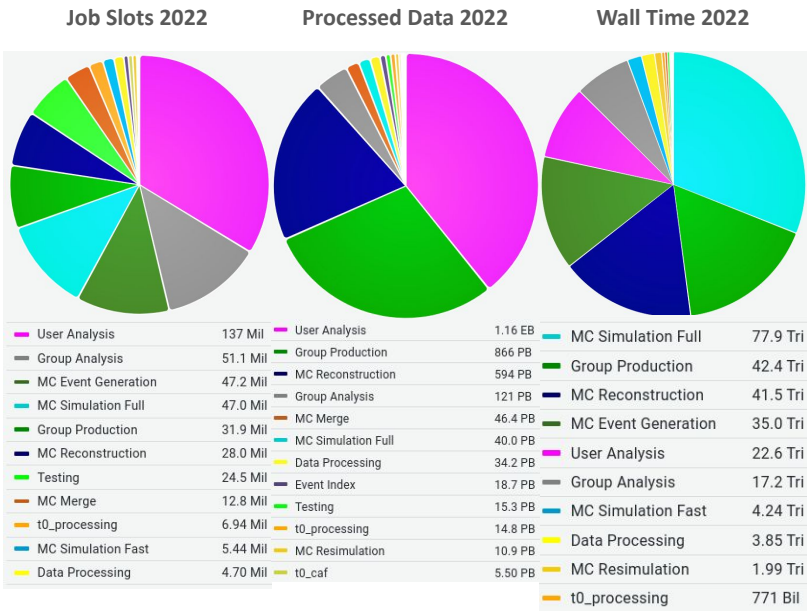
Group analysis

The main activity at a given time can depend on many things

Data **reprocessing** or Monte Carlo **production** campaigns

Conference deadlines, need for an increase for user analysis

Global **pandemics**

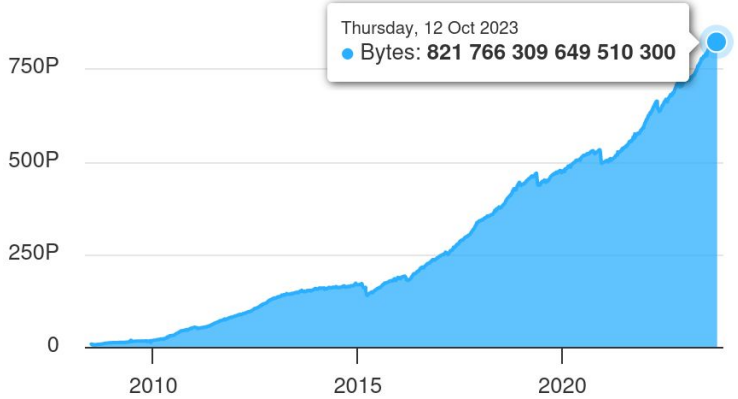


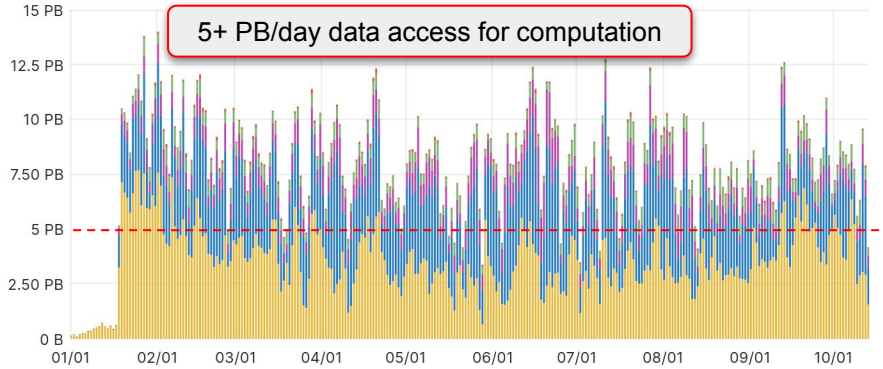
Data transfer rates

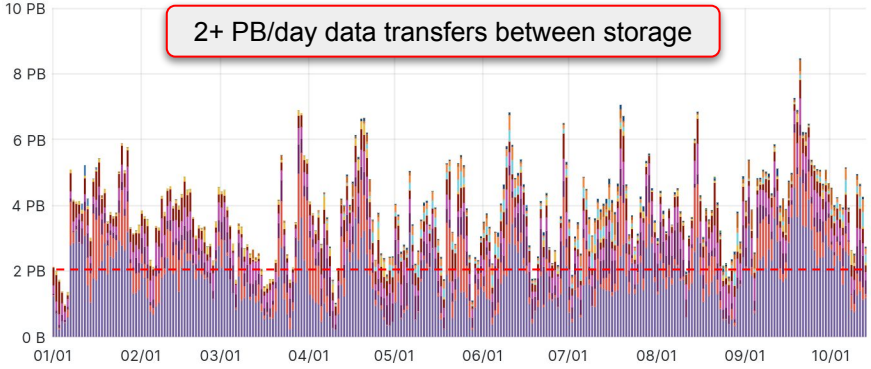
A few numbers showing the ATLAS scale

- 1B+ files, 800+ PB of data, 400+ Hz interaction
- 120 data centres, 5 HPCs, 3 clouds, 1000+ users
- 1.5+ Exabytes/year transferred
- 3+ Exabytes/year uploaded & downloaded

Increase 1+ order of magnitude for HL-LHC







Data management

Rucio handles all data management for ATLAS

Creation, location, transfer, deletion, annotation, and access
Orchestration of dataflows with both low-level and high-level policies
Coherent interface required to allow smooth data handling for production and users
We also have data management **internal flows** (recovery, rebalancing, …)

ATLAS sites are not homogeneous

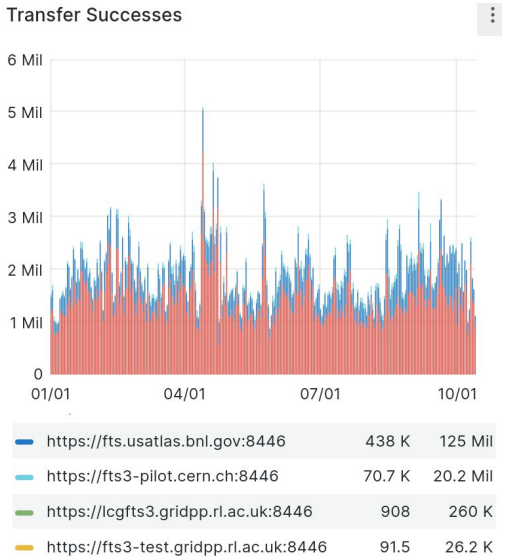
Different storage, **different** protocols
Abstracted by **FTS, GFAL** and **Davix**

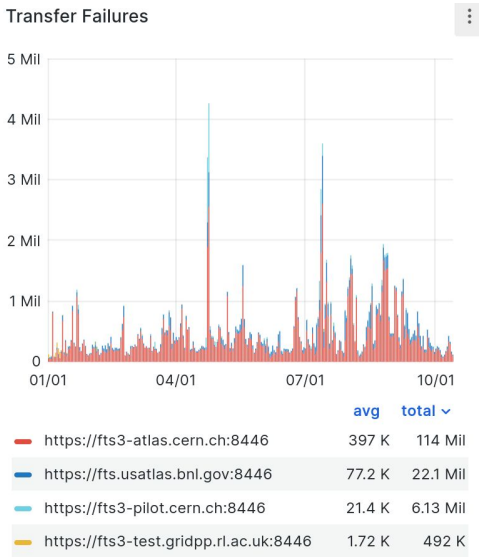
ATLAS deployment

Two FTS servers in production
Plus regularly the pilot & test services

Average file flow rate

1+ million successful transfers per day
200k failed transfers per day
Constant background failures
Biased because of **quick retries**
Peaks mostly site configuration problems





Data policies

Vast majority via subscriptions

RAW Export
DAODs to T2 disks
T0 spillover
…

Special use cases

Replication **P**olicy on the **G**rid
e.g. migrate small files to T2

RPG functionality

Being merged into subscriptions
Going away hopefully soon

Subscriptions

Show 100 entries

Name	Ok		Stuck
*Functional Test	109792		2722
*MC, EVNT	119927		11
Backup from INFN-T1	2609		0
data DAOD to non-nucleus disk	133278		26
DATA TAGS to CERN DATADISK	265		0
data15_13TeV TAGS p2685 to CERN DATADISK	92		0
Enhanced bias for HLT	110		0
EVNT to 2 T1s	31629		0
EVNT to 2 T1s 1 year	24529		0
EVNT to T0 with 1 year lifetime	12848		0
group.phys-gener to CERN-PROD_PHYS-GENER	5344		0
MC TAGS to CERN DATADISK	714		0
MC15 DAOD to T2 disk	30434		4
mc15_13TeV TAGS p2671 to CERN DATADISK	6		0
mc15_13TeV TAGS to CERN DATADISK	101		0
mc15_13TeV TAGs to CERN DATADISK	1		0
MC16 NTUP_PILEUP datasets to CERN DATADISK	179090		0
non-T0 HIST to CERN disk	761		0
PHYS-GENER_inputs	65968	0	0
sit*PAC archival to TAPE	1	0	0
T0 AOD to non-nucleu			
T0 AOD to non-nucleu			
T0 AOD to nucleus			
T0 AOD to nucleus (D			
T0 AOD to nucleus (T			
T0 DAOD to T1 disk			
T0 DESD to T1 tape			
T0 DRAW to T1 disk			
T0 DRAW to T1 tape			
T0 HIST to CERN disk	16544	1	0
T0 RAW to T1 tape	73835	2	1
T0 spillover to CERN-PROD_DERIVED	25	0	0
Validation HITS and NTUP	6717	0	0
Validation RDO, ESD, AOD and HIST	2199	0	0

```
{
  "scope": [
    "mc.*_.*TeV"
  ],
  "datatype": [
    "DAOD_.*"
  ],
  "transient": [
    "None",
    "0"
  ],
  "prod_step": [
    "merge",
    "deriv"
  ],
  "did_type": [
    "DATASET"
  ],
  "split_rule": true
}
```

```
[
  {
    "lifetime": 2592000,
    "copies": 1,
    "rse_expression": "(tier<3&type=DATADISK&datapolicyanalysis=true&datapolicynucleus=false)
    "activity": "Data Consolidation",
    "weight": "freespace"
  }
]
```

- MigrateArchiveData.conf
- MigrateArchiveMC.conf
- MigrateDataDAOD.conf
- MigrateMCDAOD.conf
- MigrateSmallFilesToT2.conf
- MigrateToContainerMC.conf
- MigrateToContainerMC15Evnt.conf
- MigrateToTapeData.conf
- MigrateToTapeDataAOD.conf
- MigrateToTapeMC.conf
- MigrateToTapeMCReconAOD.conf
- MigrateToTapeMCT01.conf
- MigrateToTapeMCT2.conf
- MigrateToTapeValid.conf
- README.md
- RPG.conf.template
- RPG.crontab
- RPG.py

Central data management operations

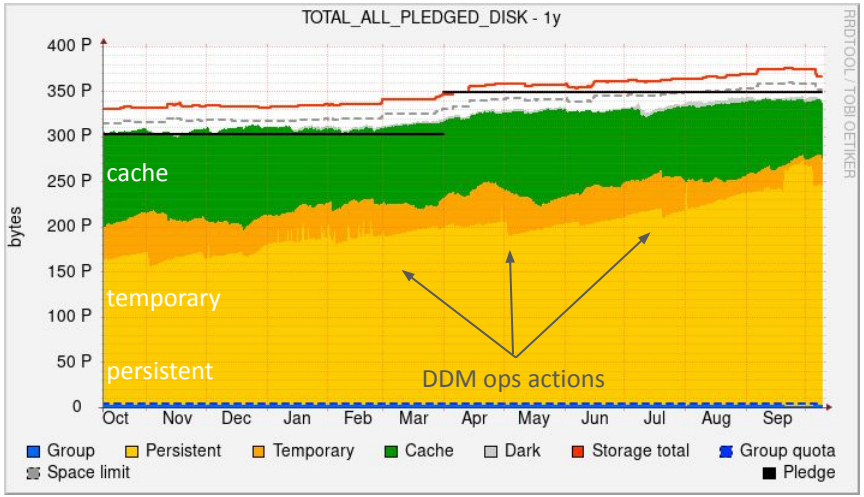
Following up transfer issues

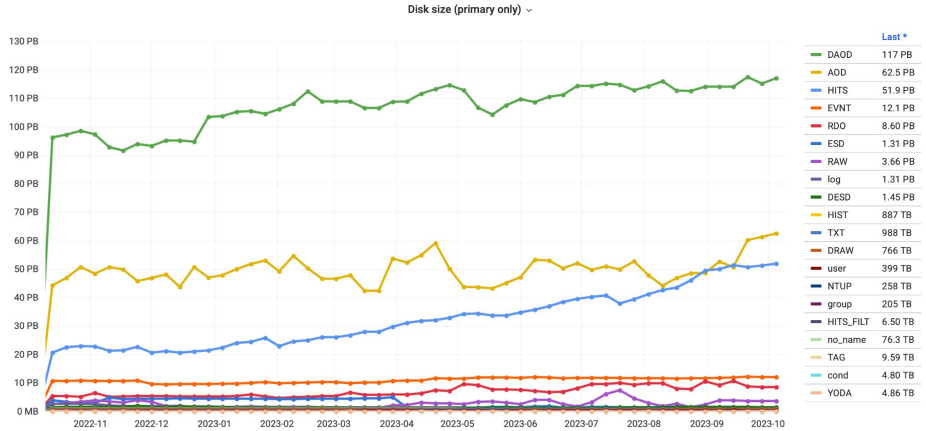
Rucio deployment operations

User/client support

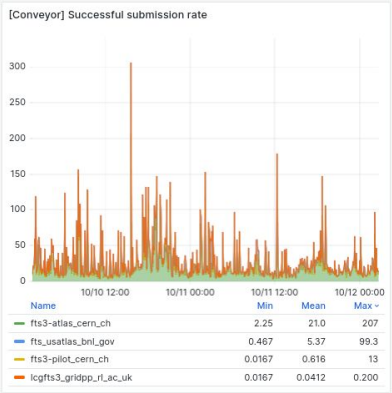
Getting disk space under control

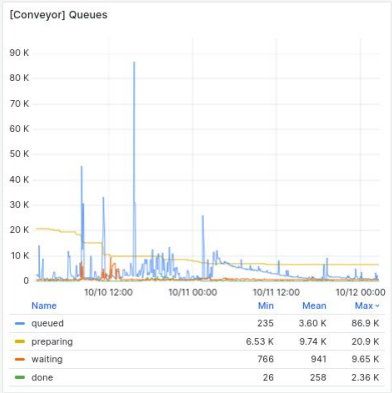
- Obsoletion campaigns
- Lifetime models (and exceptions)

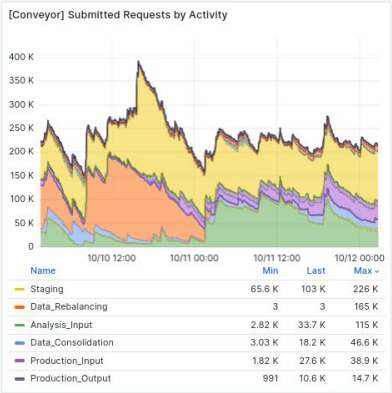


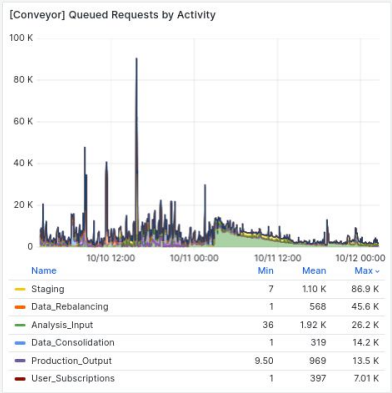


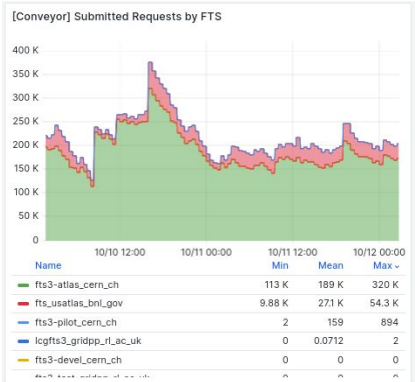
Ops dashboard 1/2

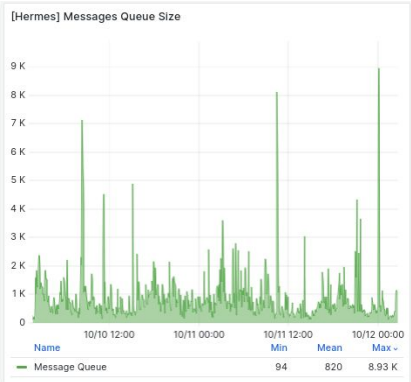


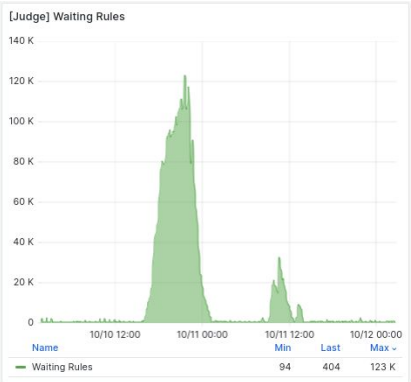


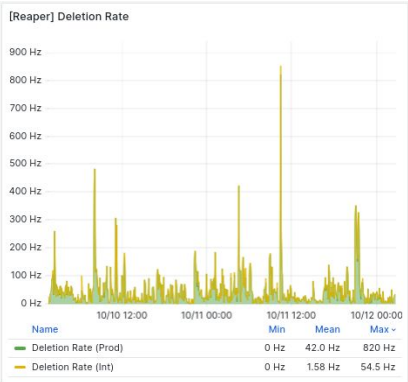




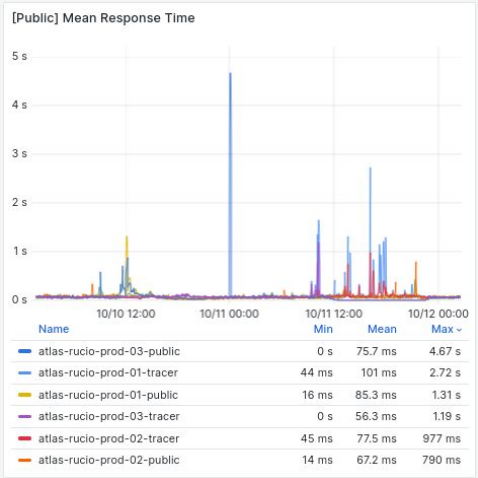


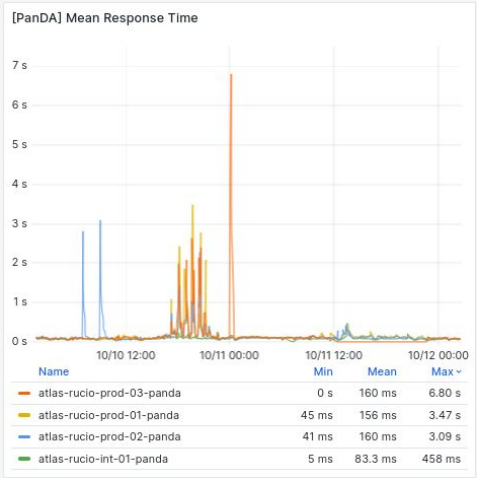


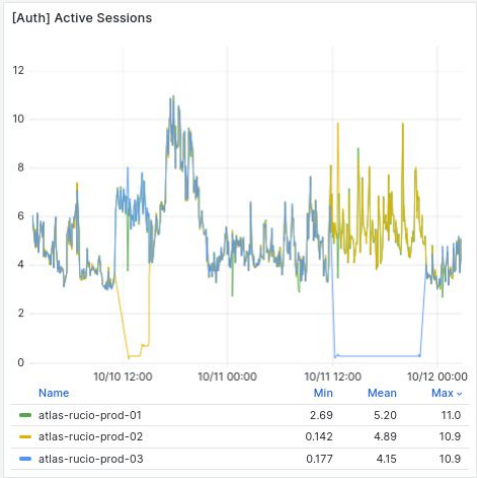


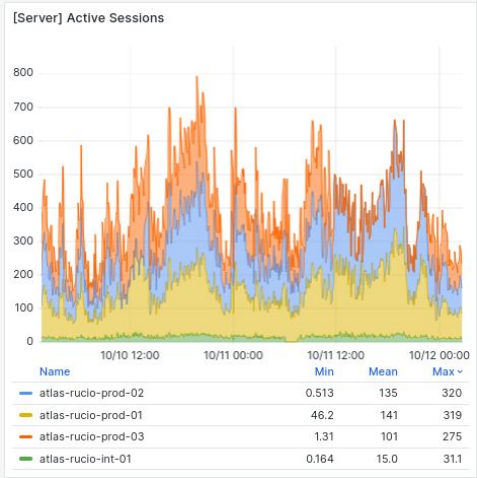


Ops dashboard 2/2

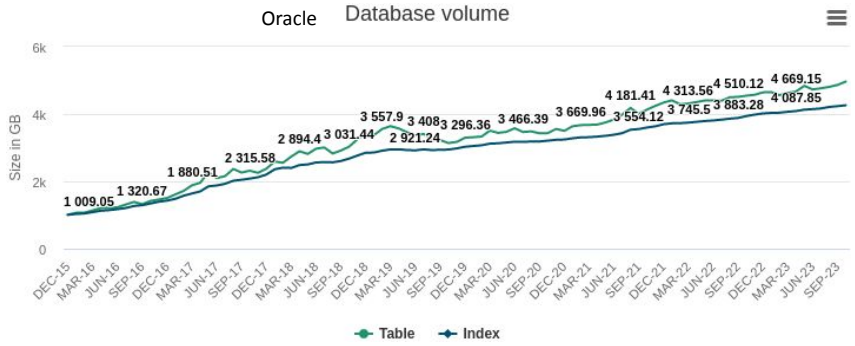


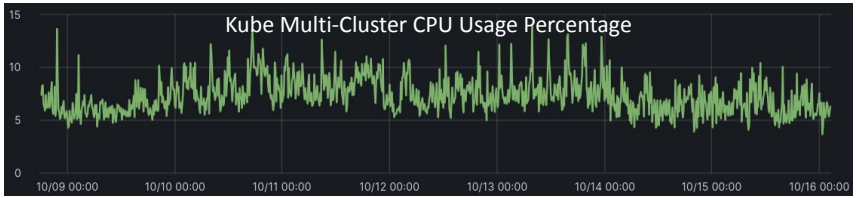


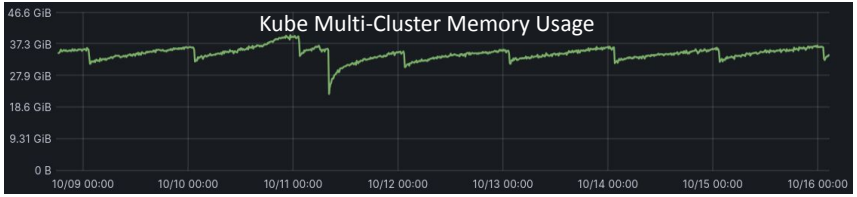


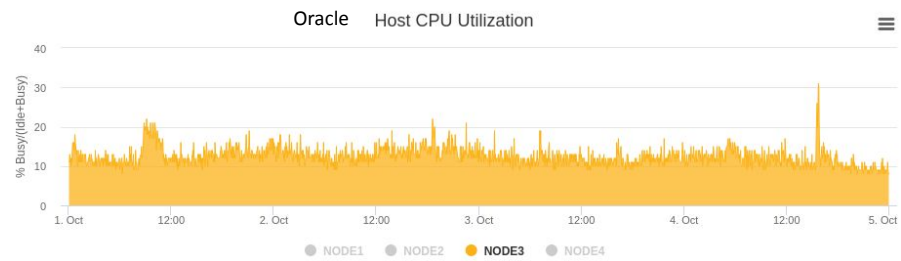


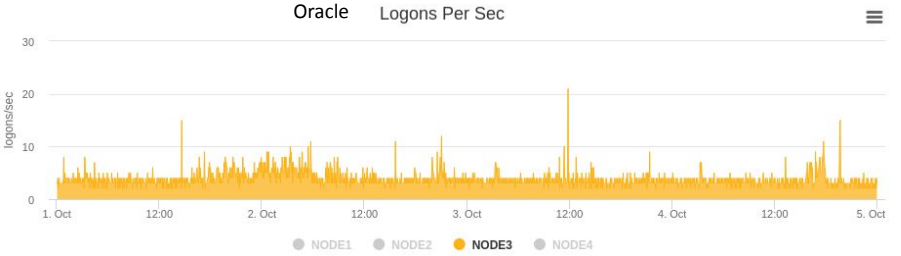
Deployment / usage

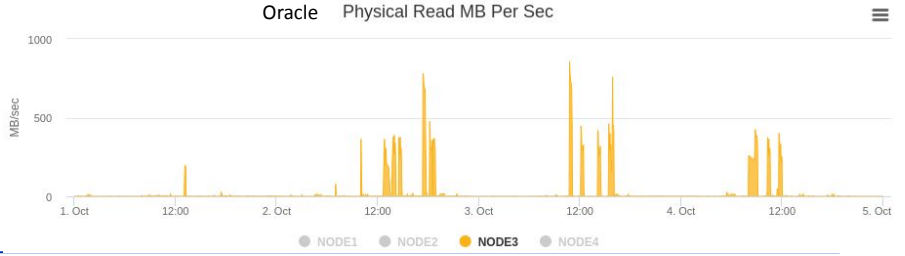










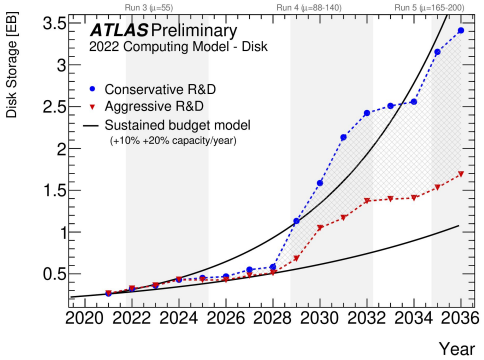


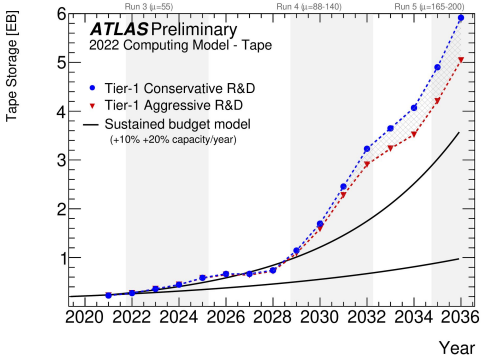
HL-HLC Data Roadmap

Next data challenge jumps from 10% (960 Gbps) to 25% (2400 Gbps) of HL-LHC needs
- **Large single step increase** of volume in the decade-long plan - had to reduce from 30%
- Need to reconsider due to **new HL-LHC schedule** and hardware purchasing

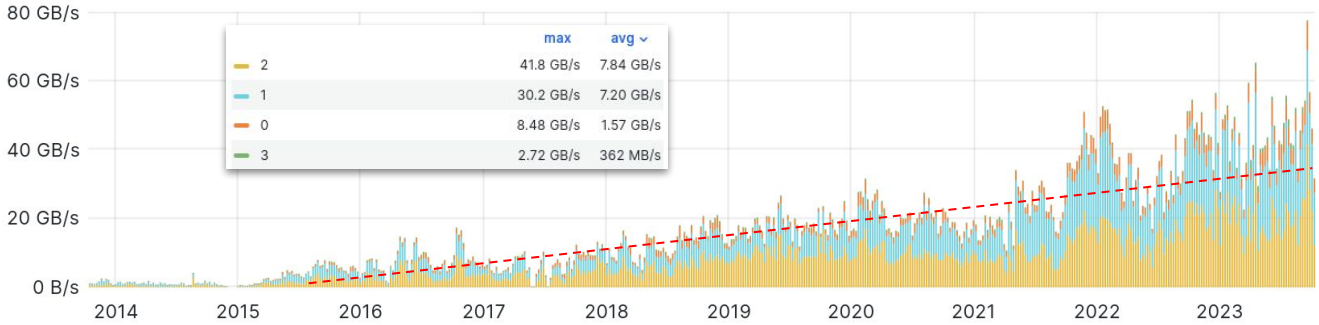
With communities beyond WLCG, such as DUNE, SKA, Belle II, JUNO, …and the NRENs
- We spend a considerable effort to **share our data management stack**
- Allows us to **work together** on these shared challenges

One interesting point: For the middleware stack, the volume is rather irrelevant
- **Number of files total**, and **number of files processed** is the key metrics
- ATLAS stance on ***big files vs. lots of files*** not yet decided





Transfer throughput per destination Tier



Death by spreadsheet

DC24 is coming in February

Lots of lessons learnt from previous Data Challenge

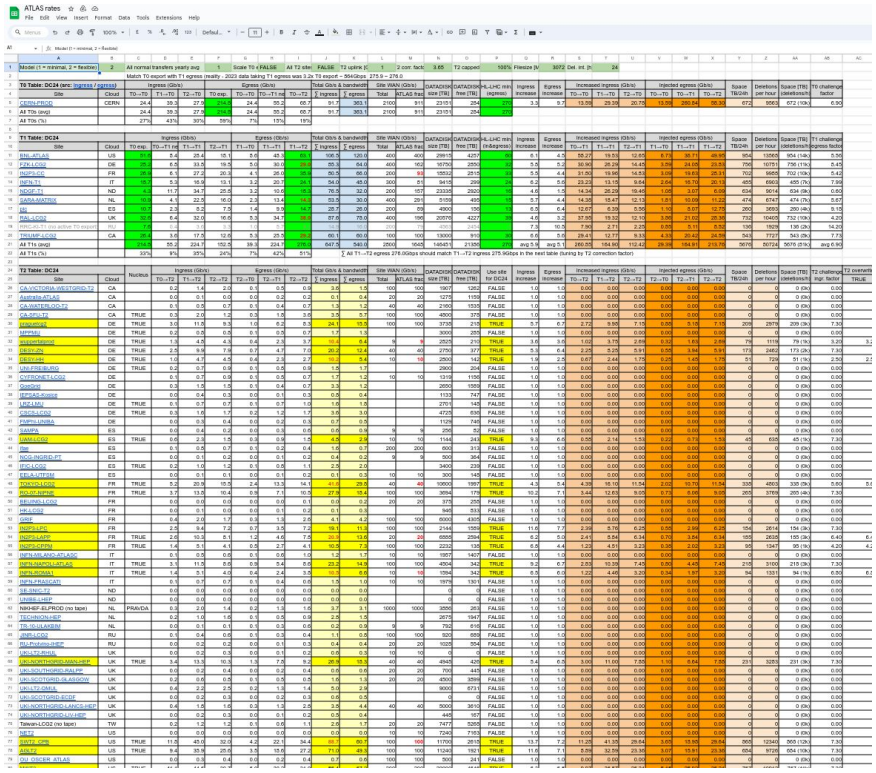
- **Rucio did very well** (and so did FTS!)
- However, injection had a distinct sawtooth-pattern
- Multi-hour cycle now revised to 15 minutes

Updated rates and new methodology

- Original distinction in minimal and flexible model
- Adapted to reality
  - Tier-0 export flows match LHC machine rates
  - Tier-1 and Tier-2 flows match processing

Ramp up challenges with new Rucio injection tools

- **Expect first plots at Data Challenge workshops**



Cloud

ATLAS has **cloud R&D projects** ongoing with **Amazon**, **Google**, and **SEAL** Storage

Integration into ADC systems PanDA & Rucio, and in turn FTS, GFAL, Davix

Very **close development collaboration** across the full stack

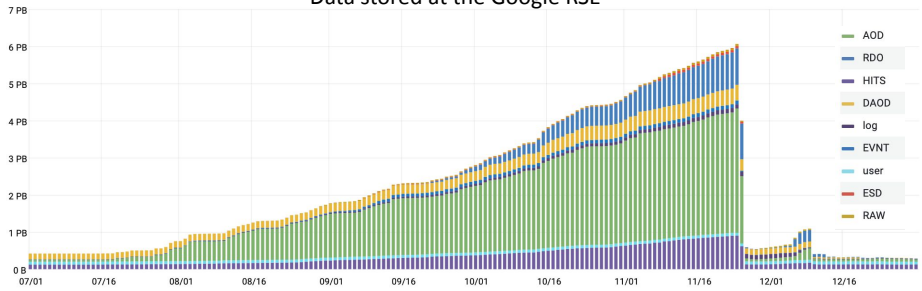
Two major angles to consider when discussing clouds

Technical	Access tools, transfer protocols, monitoring, authn/z, accounting, billing, storage, …	
Organisational	Deployed on-site or off-site	Centralised or distributed
	Public (institute, laboratory, …) or commercial	In-kind contribution or paid service

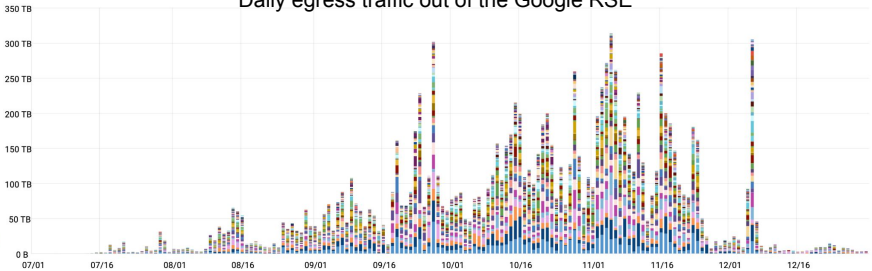
Large development programme in front of us to make cloud storage viable

Throughput **control**, access **control**, peering **control**, cloud transfer tool **control**, lifetime **control**, cost **control**, …

Data stored at the Google RSE



Daily egress traffic out of the Google RSE



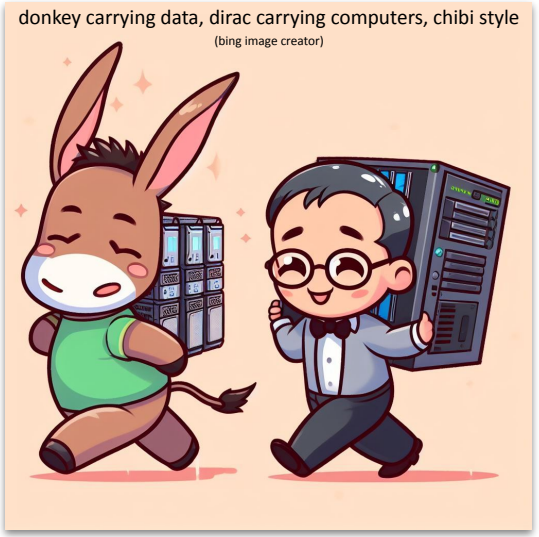
Summary

Rucio is working great for ATLAS!

- Thanks to the dedication of a great team
- We are happy and grateful for this big community

The ATLAS data needs are immense and continuously increasing

- Data flow complexity, incl. system topology and experiment policies
- Throughput and file rates are ever increasing
- Crazy R&D projects to keep things interesting ;-)



ATLAS will continue to contribute to the development and support of Rucio into the HL-LHC era!


CERN


ATLAS
EXPERIMENT