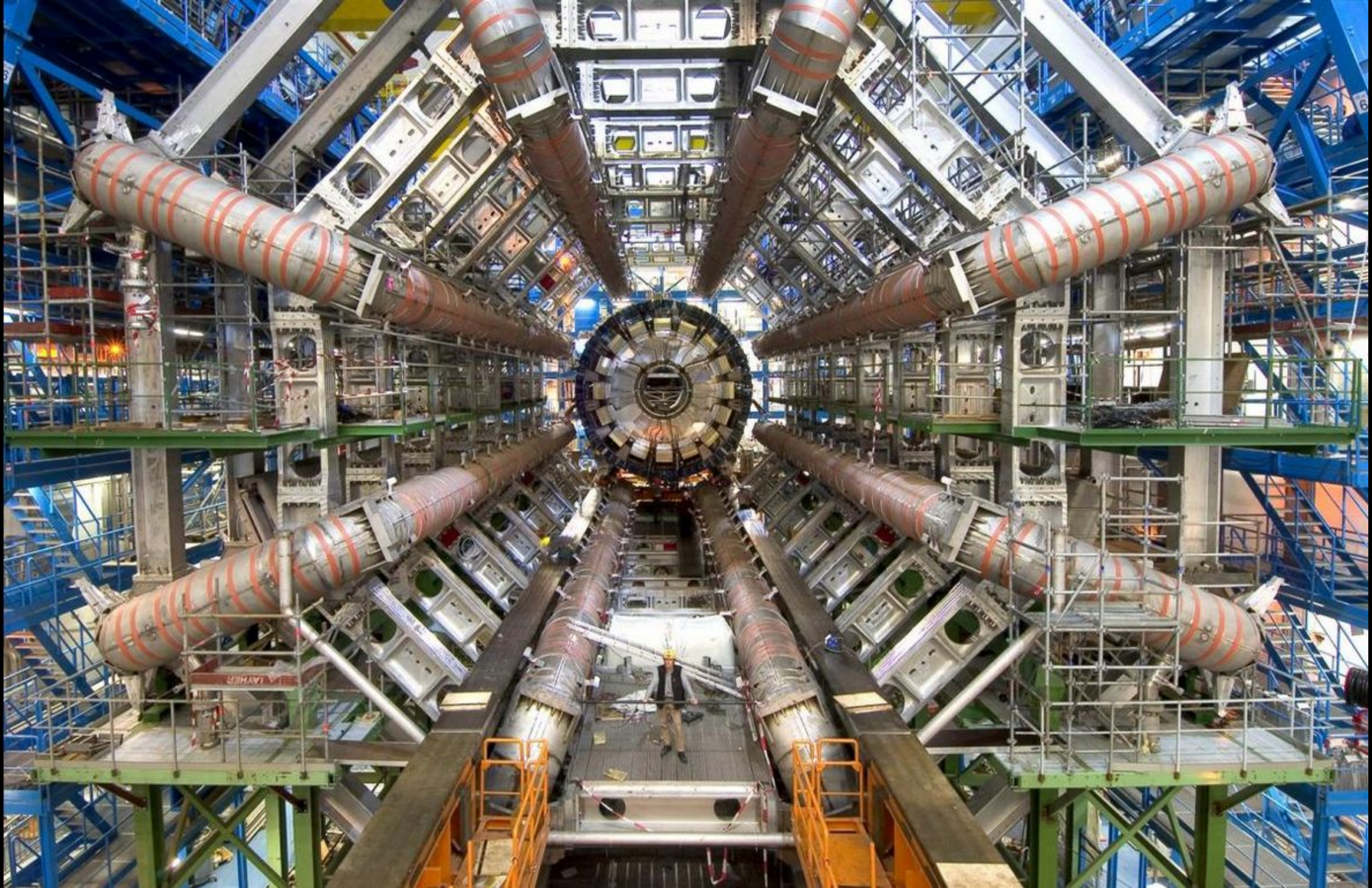
ATLAS Data Flows & FTS

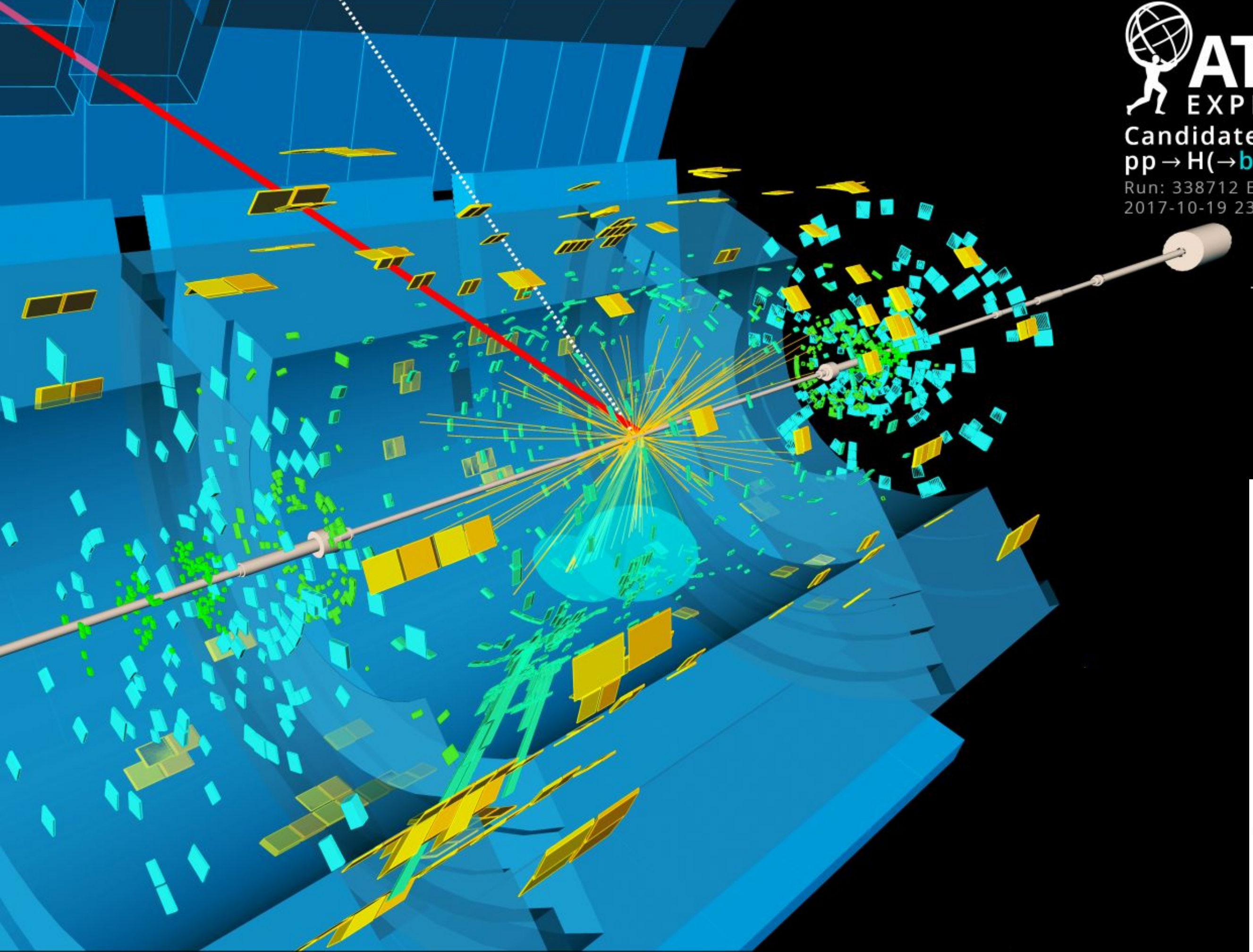
Mario.Lassnig@cern.ch

XRootD & FTS Workshop

27 March 2022 - 31 March 2023

https://indico.cern.ch/event/875381





ATLAS EXPERIMENT

Candidate Event: $pp \to H(\to bb) + W(\to \mu\nu)$

Run: 338712 Event: 335908183

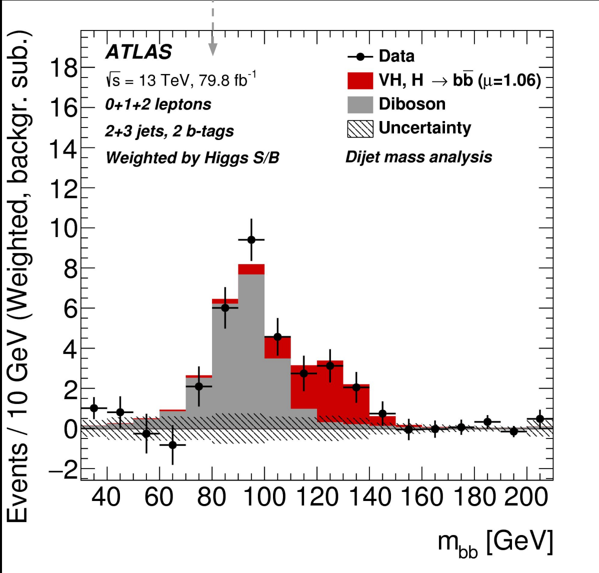
2017-10-19 23:31:18 CEST

13 TeV detector data

8 quadrillion collision candidates
92 petabytes
130 million files

13 TeV simulation data

166 petabytes
544 million files

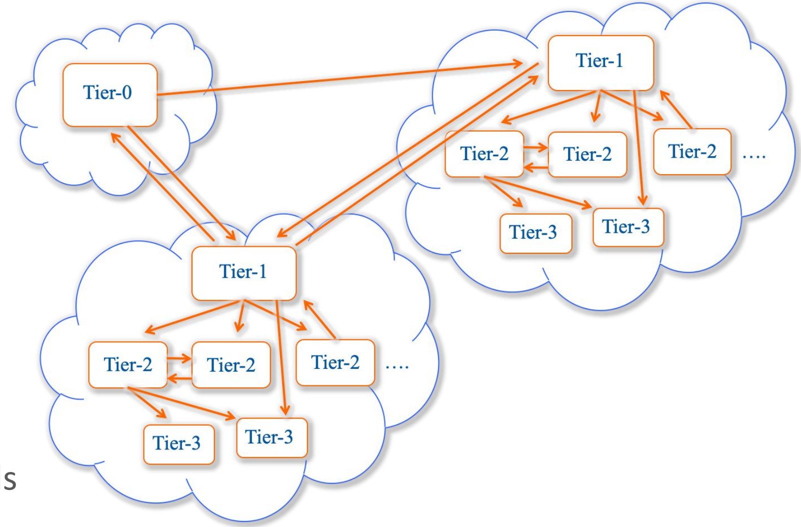


A candidate event display for the production of a Higgs boson decaying to two b-quarks (blue cones), in association with a W boson decaying to a muon (red) and a neutrino. The neutrino leaves the detector unseen, and is reconstructed through the missing transverse energy (dashed line). (Image: ATLAS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

Experiment data flow 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**



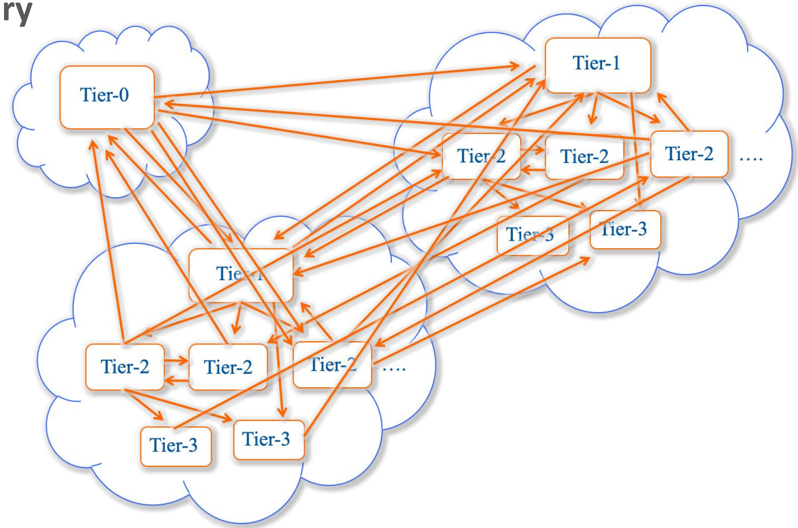
Model had a series of shortcomings

- 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

Experiment data flow 2/2

WLCG networks have evolved significantly in the last two 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

Job types drive the data volume

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

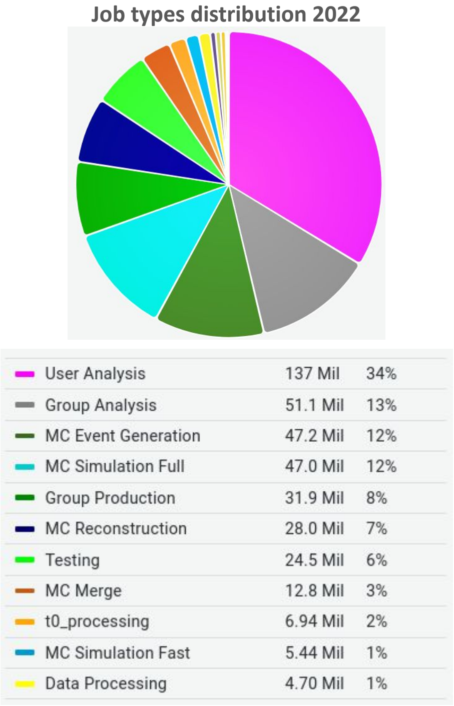
Production	Data reprocessing Event generation / Simulation / Reconstruction Group production
Analysis	User analysis Group analysis

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**



Job types distribution 2022

▬	User Analysis	137 Mil	34%
▬	Group Analysis	51.1 Mil	13%
▬	MC Event Generation	47.2 Mil	12%
▬	MC Simulation Full	47.0 Mil	12%
▬	Group Production	31.9 Mil	8%
▬	MC Reconstruction	28.0 Mil	7%
▬	Testing	24.5 Mil	6%
▬	MC Merge	12.8 Mil	3%
▬	t0_processing	6.94 Mil	2%
▬	MC Simulation Fast	5.44 Mil	1%
▬	Data Processing	4.70 Mil	1%

Data transfer rates

A few numbers showing the ATLAS scale

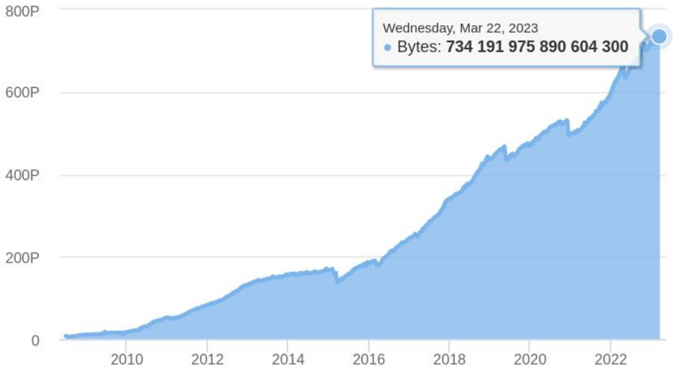
1B+ files, 700+ PB of data, 400+ Hz interaction

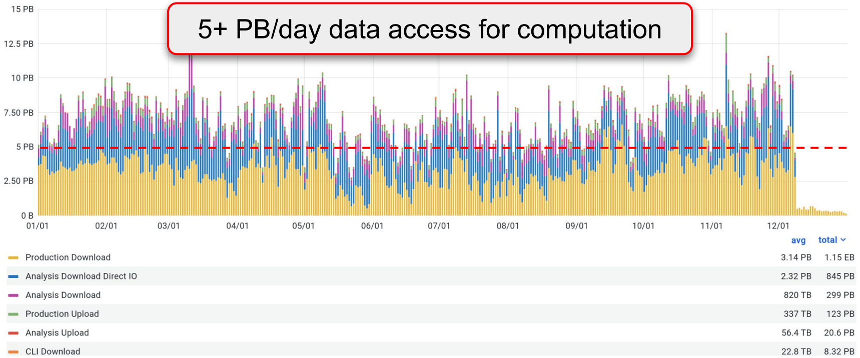
120 data centres, 5 HPCs, 3 clouds, 1000+ users

1.2 Exabytes/year transferred

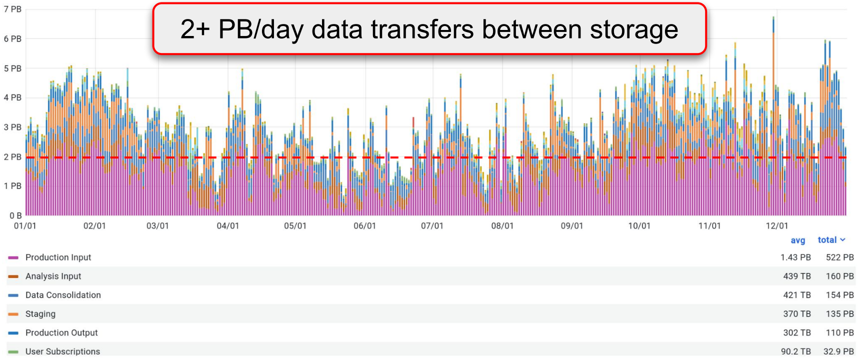
2.7 Exabytes/year uploaded & downloaded

Increase 1+ order of magnitude for HL-LHC




5+ PB/day data access for computation

	avg	total ⌄
▬ Production Download	3.14 PB	1.15 EB
▬ Analysis Download Direct IO	2.32 PB	845 PB
▬ Analysis Download	820 TB	299 PB
▬ Production Upload	337 TB	123 PB
▬ Analysis Upload	56.4 TB	20.6 PB
▬ CLI Download	22.8 TB	8.32 PB


2+ PB/day data transfers between storage

	avg	total ⌄
▬ Production Input	1.43 PB	522 PB
▬ Analysis Input	439 TB	160 PB
▬ Data Consolidation	421 TB	154 PB
▬ Staging	370 TB	135 PB
▬ Production Output	302 TB	110 PB
▬ User Subscriptions	90.2 TB	32.9 PB

Data management

Rucio handles the data management

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

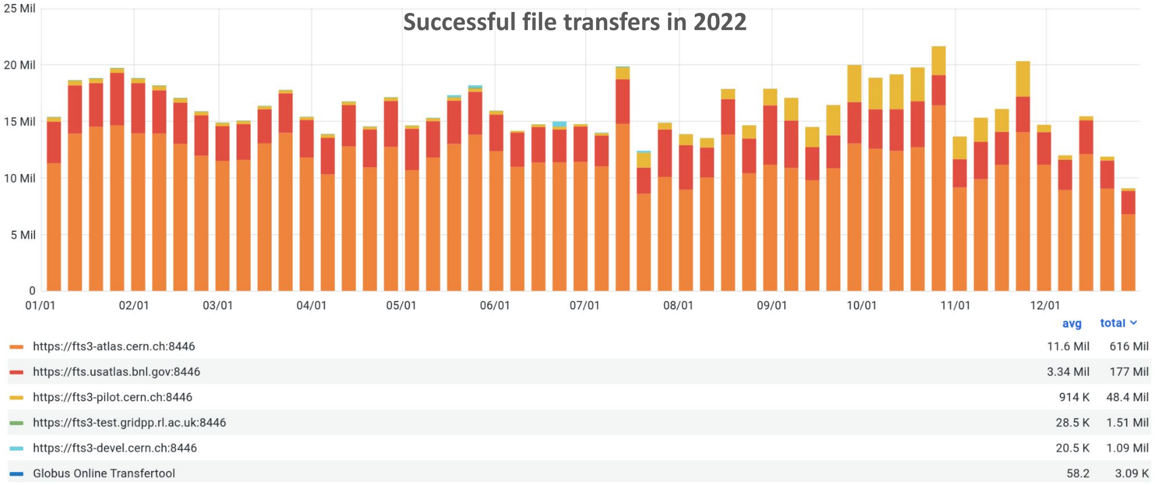
Different storage, **different** protocols
Hello **FTS, GFAL and Davix** :-)

ATLAS deployment

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

Average file flow rate

15 million successful transfers per day
2 million failed transfers per day
Mostly site configuration problems


Successful file transfers in 2022

	avg	total ⌄
🟧 https://fts3-atlas.cern.ch:8446	11.6 Mil	616 Mil
🟥 https://fts.usatlas.bnl.gov:8446	3.34 Mil	177 Mil
🟨 https://fts3-pilot.cern.ch:8446	914 K	48.4 Mil
🟩 https://fts3-test.gridpp.rl.ac.uk:8446	28.5 K	1.51 Mil
🟦 https://fts3-devel.cern.ch:8446	20.5 K	1.09 Mil
🟦 Globus Online Transfertool	58.2	3.09 K

Cloud Storage

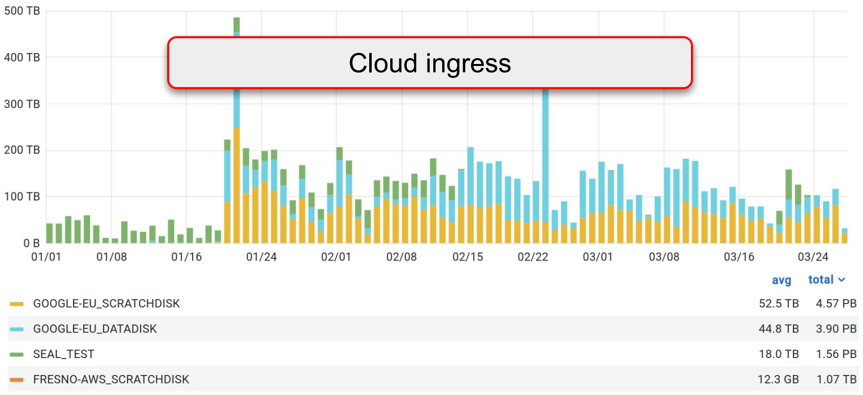
ATLAS has cloud R&D projects ongoing with Amazon, Google, and SEAL Storage

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

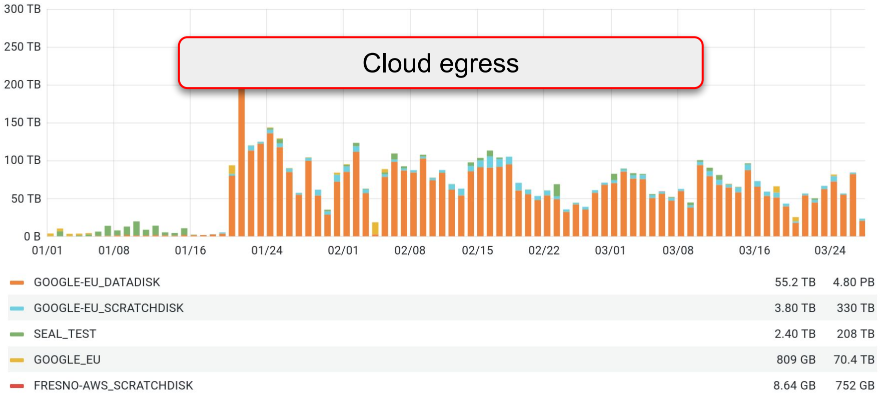
Very **close development collaboration** across the full stack

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

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



	avg	total ⌄
🟨 GOOGLE-EU_SCRATCHDISK	52.5 TB	4.57 PB
🟦 GOOGLE-EU_DATADISK	44.8 TB	3.90 PB
🟩 SEAL_TEST	18.0 TB	1.56 PB
🟧 FRESNO-AWS_SCRATCHDISK	12.3 GB	1.07 TB



🟧 GOOGLE-EU_DATADISK	55.2 TB	4.80 PB
🟦 GOOGLE-EU_SCRATCHDISK	3.80 TB	330 TB
🟩 SEAL_TEST	2.40 TB	208 TB
🟨 GOOGLE_EU	809 GB	70.4 TB
🟥 FRESNO-AWS_SCRATCHDISK	8.64 GB	752 GB

HTTP TAPE REST API

ADC wants to move to the new HTTP TAPE REST API earlier than later

- By the way, we need a better name for this… HTRA ? Doesn't work… :-D
- Four volunteer sites: CERN, FZK, DESY, BNL
- All CTA endpoints use the `archive_timeout=86400` functionality

State of the manual functional tests

- **Successful archive & recall** at CERN CTA
- **Successful archive** at FZK :: dCache version upgraded on Monday :: New tests coming asap
- BNL **ready for testing**
- DESY **waiting for configuration**

Plan for putting it in production

- Once manual functional tests are successful, change the **LOCALGROUPTAPE** at the site
- Once we're confident it works well, switch the remaining tape endpoints at the site

HL-LHC 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%
- Potentially need to reconsider due to **new HL-LHC schedule** and hardware purchasing

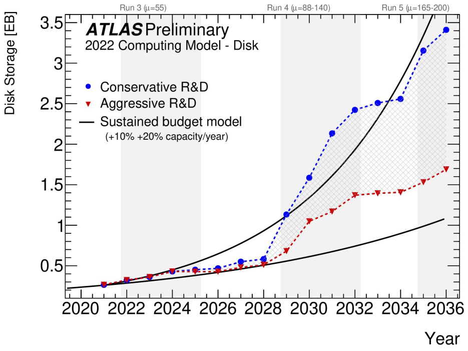
Token-based authentication will be deployed and tested at scale during DC24

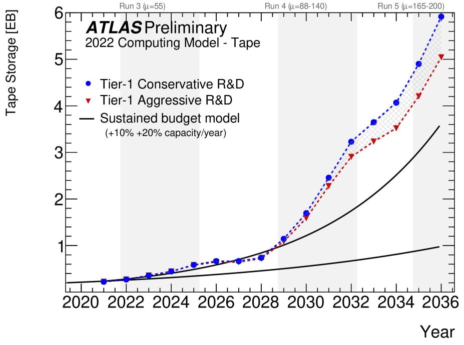
With communities beyond WLCG, such as DUNE, SKA, Belle II, JUNO, …and the NRENs

- 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





Our input to FTS development & operations

Major topics to address

- ~~Database performance and scalability~~
- Consistent configuration
- Timely upgrades across all FTS instances
- Global scheduling algorithm improvements
- Limit enforcement
- Fair-sharing per endpoint
- Re-prioritisation of transfers
- Resurrect steering meetings
- Battle-tested OIDC Token support
- ~~Commercial cloud support~~
- ~~Improve web interface~~
- Timeout handling for slow transfers
- Improved error reasoning & messages

Medium term topics

- ~~Bulk methods for tape interaction (HTTP REST API)~~
- Better automatic source selection
- Automatic session reuse revisited
- Easier debugging of failed FTS transfers
- SDN integration and support

Long term topics

- Backpressure mechanism from storage to FTS
- Labelling of transfers for networks
- Network awareness for transfer scheduling
- Load balancing across multiple storage endpoints at destination
- Community contributions for protocol support
- Cross-experiment scheduling

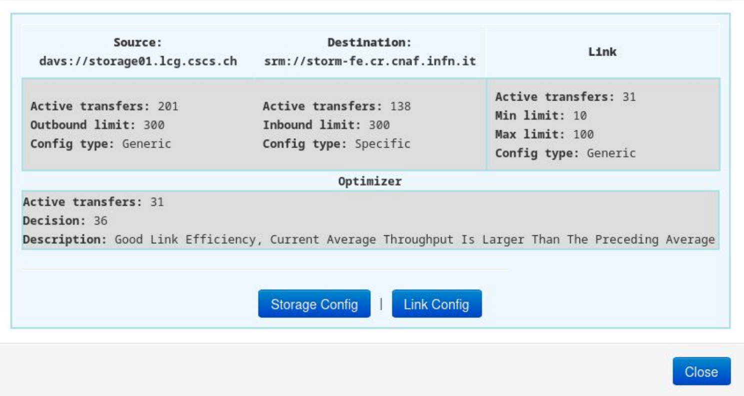
Our input to FTS development & operations

Major topics to address

- ~~Database performance and scalability~~
- Consistent configuration
- Timely upgrades across all FTS instances
- Global scheduling algorithm improvements
- Limit enforcement
- Fair-sharing per endpoint
- Re-prioritisation of transfers
- Resurrect steering meetings
- Battle-tested OIDC Token support
- ~~Commercial cloud support~~
- ~~Improve web interface~~
- Timeout handling for slow transfers
- Improved error reasoning & messages

Medium term topics

- ~~Bulk methods for tape interaction (HTTP REST API)~~
- Better automatic source selection
- TS
- points at destination
- Community contributions for protocol support
- Cross-experiment scheduling



Source: davs://storage01.lcg.cscs.ch	Destination: srm://storm-fe.cr.cnaf.infn.it	Link
Active transfers: 201 Outbound limit: 300 Config type: Generic	Active transfers: 138 Inbound limit: 300 Config type: Specific	Active transfers: 31 Min limit: 10 Max limit: 100 Config type: Generic

Optimizer

Active transfers: 31
Decision: 36
Description: Good Link Efficiency, Current Average Throughput Is Larger Than The Preceding Average

Summary

FTS is absolutely essential for ATLAS

Software is **stable and efficient**

Development and Operations teams are **friendly, diligent, and quick**

(even on very minor topics! thanks a lot!)

Strong long-term support of FTS team by CERN IT mgmt is **crucial**

Long list of topics for continuous collaboration

Data management as a whole is **progressing** at a nice pace

Many communities joining in due to our **shared software stack**

We look forward to continuing to **work together** in the future!