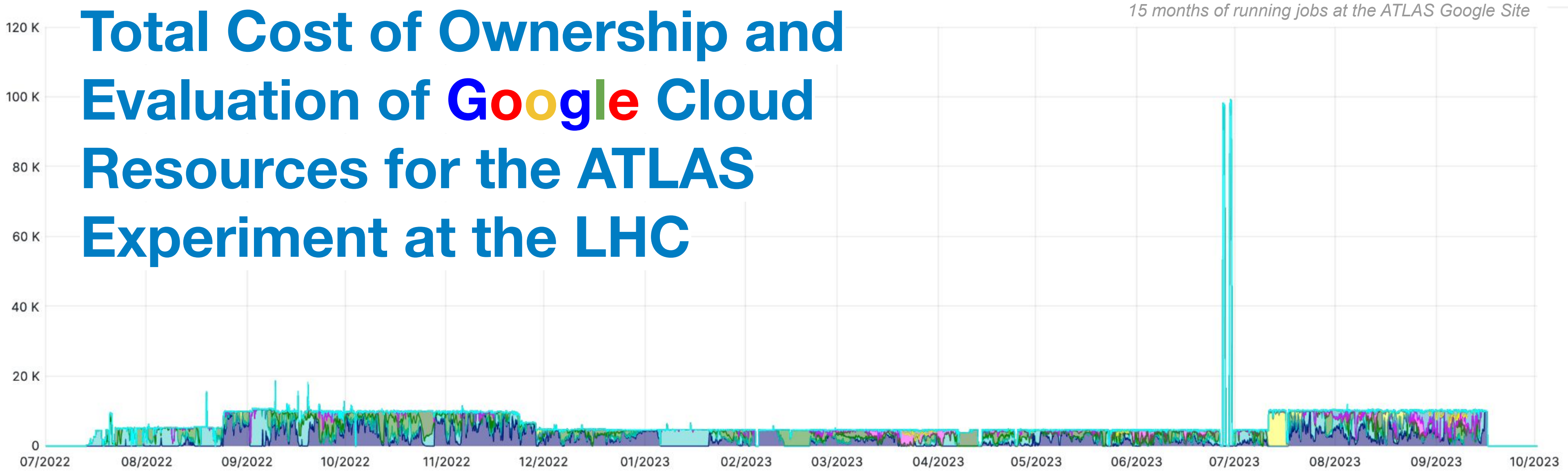
Total Cost of Ownership and Evaluation of Google Cloud Resources for the ATLAS Experiment at the LHC



David South (DESY) on behalf of the ATLAS Collaboration and Gonzalo Merino (PIC)

27TH INTERNATIONAL CONFERENCE ON COMPUTING IN HIGH ENERGY & NUCLEAR PHYSICS

CHEP 2024, Krakow, Poland, October 19-25th 2024

Accepted by CSBS https://arxiv.org/abs/2405.13695

DESY.

PIC
port d'informació científica

ATLAS EXPERIMENT

The LHC and ATLAS

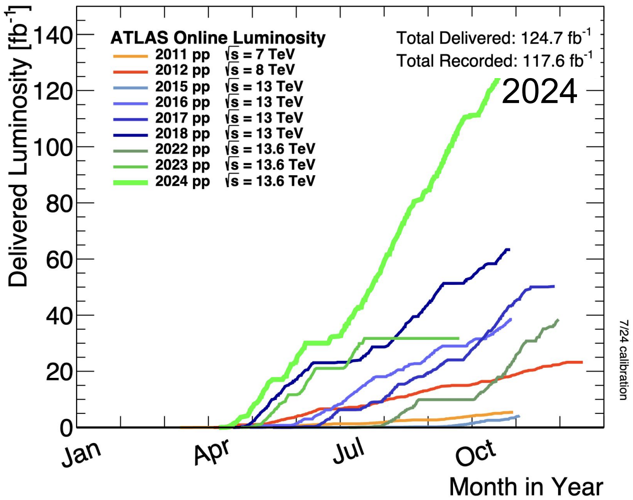
The LHC: Exploration of the energy frontier in pp and HI collisions

- 27 km circumference, up to 175 metres below the surface
- More than 10,000 superconducting magnets at 1.9K
- Bunches of 10^{11} protons, **2.6B collisions per second in Run 3**

The ATLAS trigger system and event selection reduces this to an **event rate of around 3.8 kHz**

- An ATLAS RAW data event in Run 3 is ~ 1.8 MB
- So an event almost* fits on an (HD) 3.5” disk
- The main physics stream from this year fits on a stack of floppy disks about 60,000km high

Museum of Obsolete Media
https://obsoletemedia.org



The LHC represents a new frontier in physics data volume

- Combined, the experiments generate ~**150 PB of collision data / year**
- But the RAW collision data are only the start..

** It's actually 1.25 floppy disks/event:*
~15B physics_Main pp events in 2024 x (1.8/1.44) x 3.3mm

The LHC experiments do a lot of processing

RAW collision data

- Sensor hits, energy deposits, timing information

Analysis Object Data (AOD)

- 4-vector momentum of tracks
- Energy in jet clusters
- Particle identification
- First calibrations

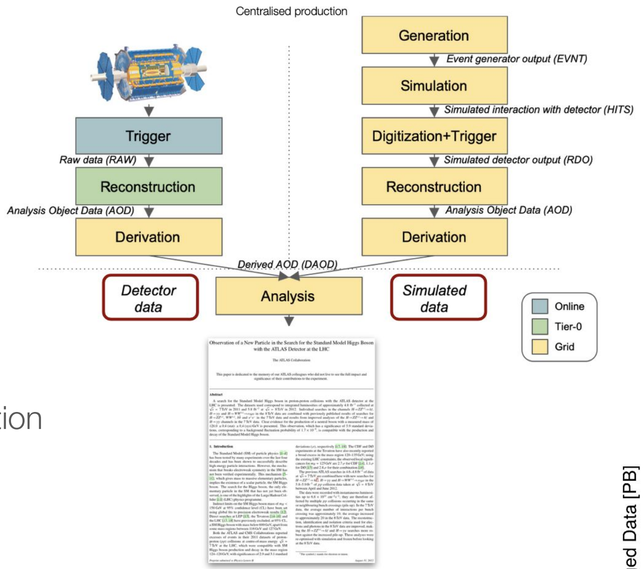
Derived AODs

- Selected analysis level information with full calibrations
- Starting point for analysis

Monte Carlo Simulation

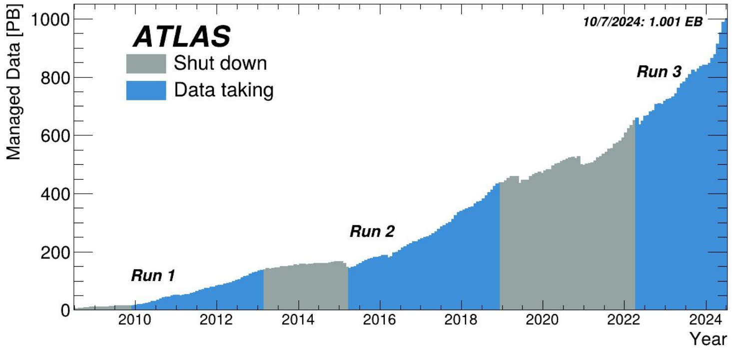
Event generation	EVNT	Calculated particle interactions
Simulation	HITS	Interactions with detector material
Digitization	RDO	Simulated energy → detector response
Reconstruction	AOD	Performed the same way as for data
Derivations	DAOD	Performed the same way as for data

ATLAS Data Processing Chain



ATLAS regularly produces new versions of data and simulated data in large scale production campaigns

- New features and methods, latest calibrations and improvements
- This adds to the total ATLAS data volume, now over an Exabyte, and all managed by RUCIO



The LHC experiments do a lot of processing

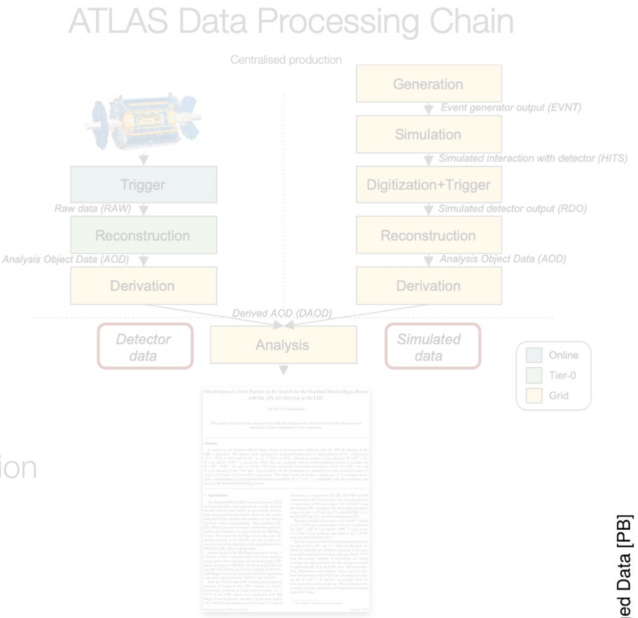
RAW collision data
- Sensor hits, energy deposits, timing information

Analysis Object Data (AOD)
- 4-vector momentum of tracks
- Energy in jet clusters
- Particle identification
- First calibrations

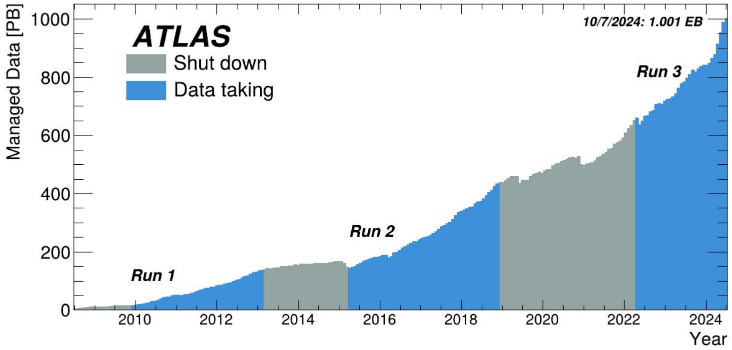
Derived AODs
- Selected analysis level information with full calibrations
- Starting point for analysis

Monte Carlo Simulation

Event generation	EVNT	Calculated particle interactions
Simulation	HITS	Interactions with detector material
Digitization	RDO	Simulated energy → detector response
Reconstruction	AOD	Performed the same way as for data
Derivations	DAOD	Performed the same way as for data



ATLAS regularly produces new versions of data and simulated data in large scale production campaigns
- New features and methods, latest calibrations and improvements
- This adds to the total ATLAS data volume, now over an Exabyte, and all managed by RUCIO

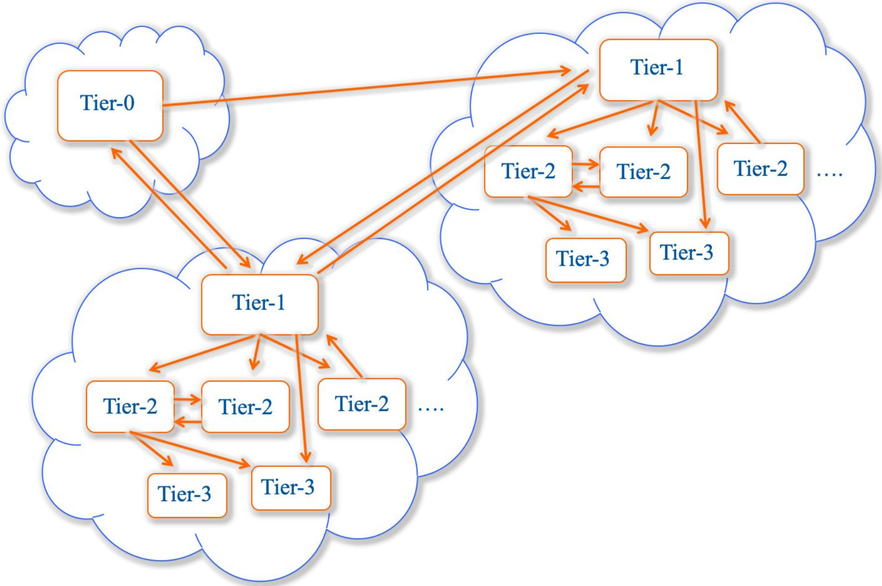


Distributed Computing Infrastructure

To produce and analyse such large quantities of data requires a significant computing infrastructure

The Worldwide LHC Computing Grid (WLCG) needs no introduction to this audience

- Originally organised as static “clouds”, mostly national or geographical groupings of sites, with common funding agencies and languages

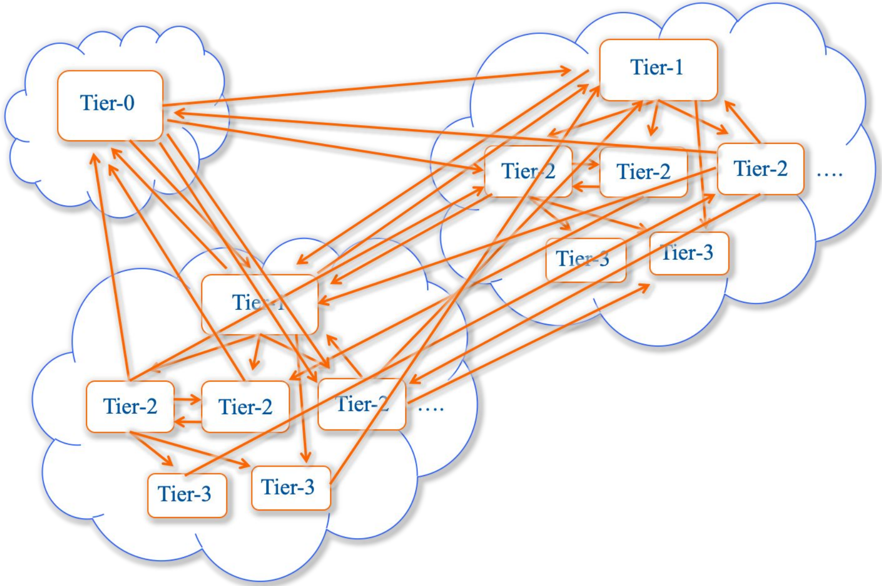


Distributed Computing Infrastructure

To produce and analyse such large quantities of data requires a significant computing infrastructure

The Worldwide LHC Computing Grid (WLCG) needs no introduction to this audience

- Originally organised as static “clouds”, mostly national or geographical groupings of sites, with common funding agencies and languages
- Networks have evolved significantly in the last two decades and limiting transfers to within a single cloud no longer necessary



Distributed Computing Infrastructure

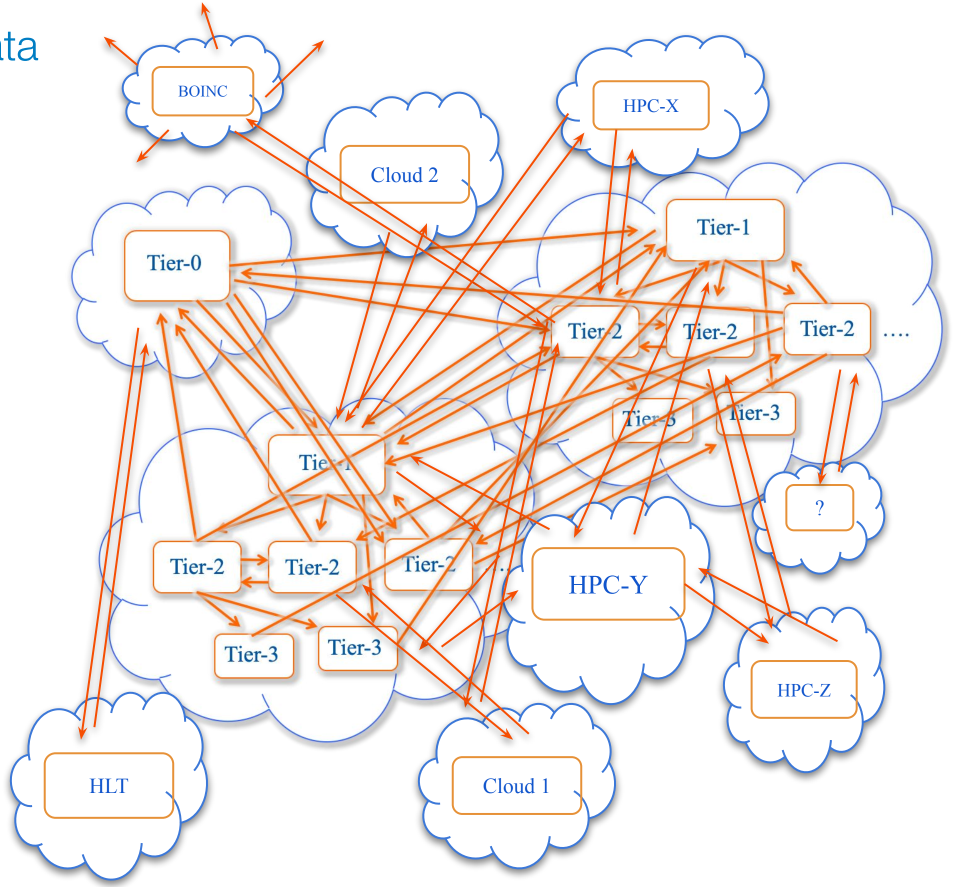
To produce and analyse such large quantities of data requires a significant computing infrastructure

The Worldwide LHC Computing Grid (WLCG) needs no introduction to this audience

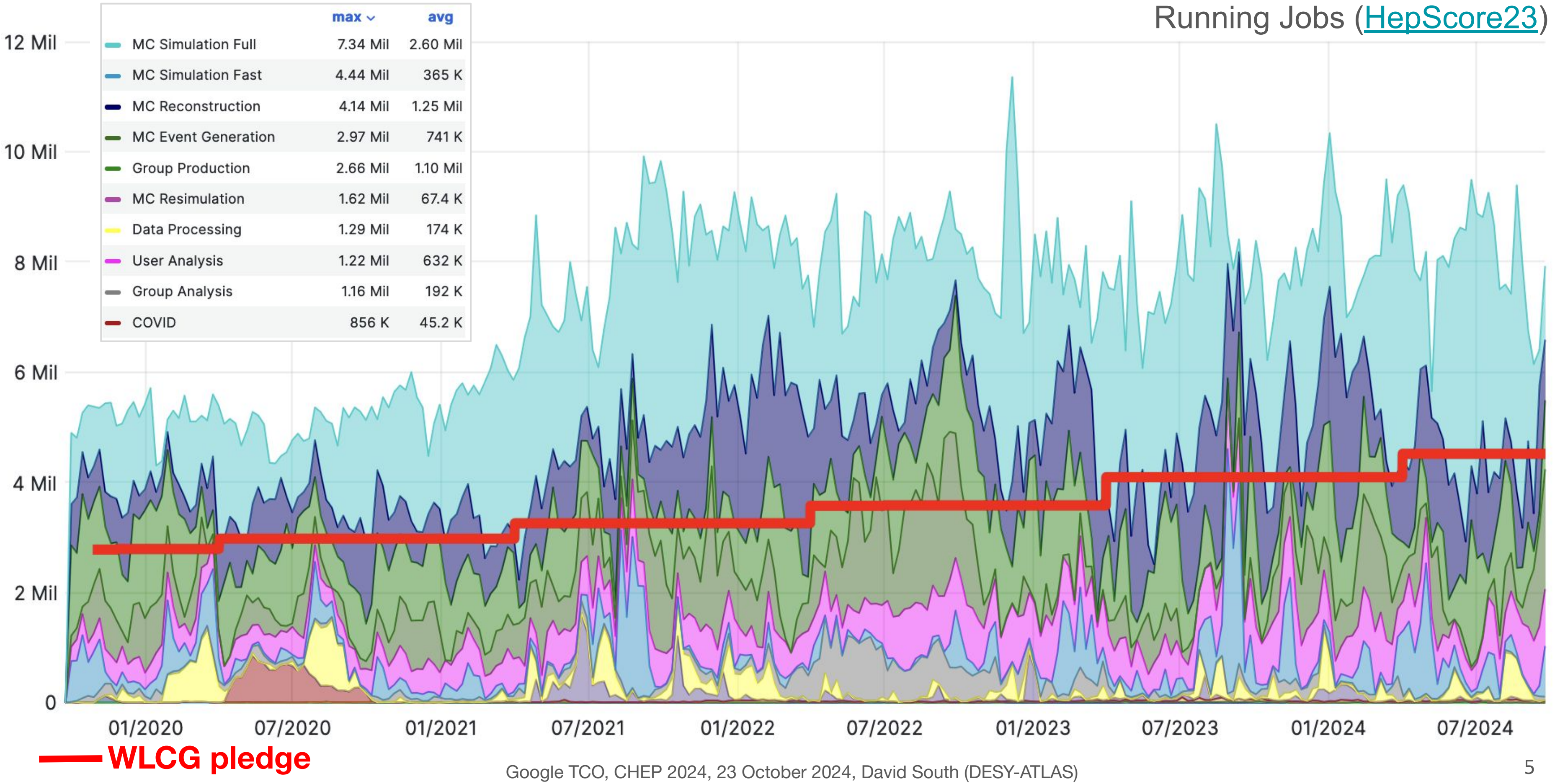
- Originally organised as static "clouds", mostly national or geographical groupings of sites, with common funding agencies and languages
- Networks have evolved significantly in the last two decades and limiting transfers to within a single cloud no longer necessary

Grid now complemented by additional, beyond pledge resources such as HPCs and Cloud

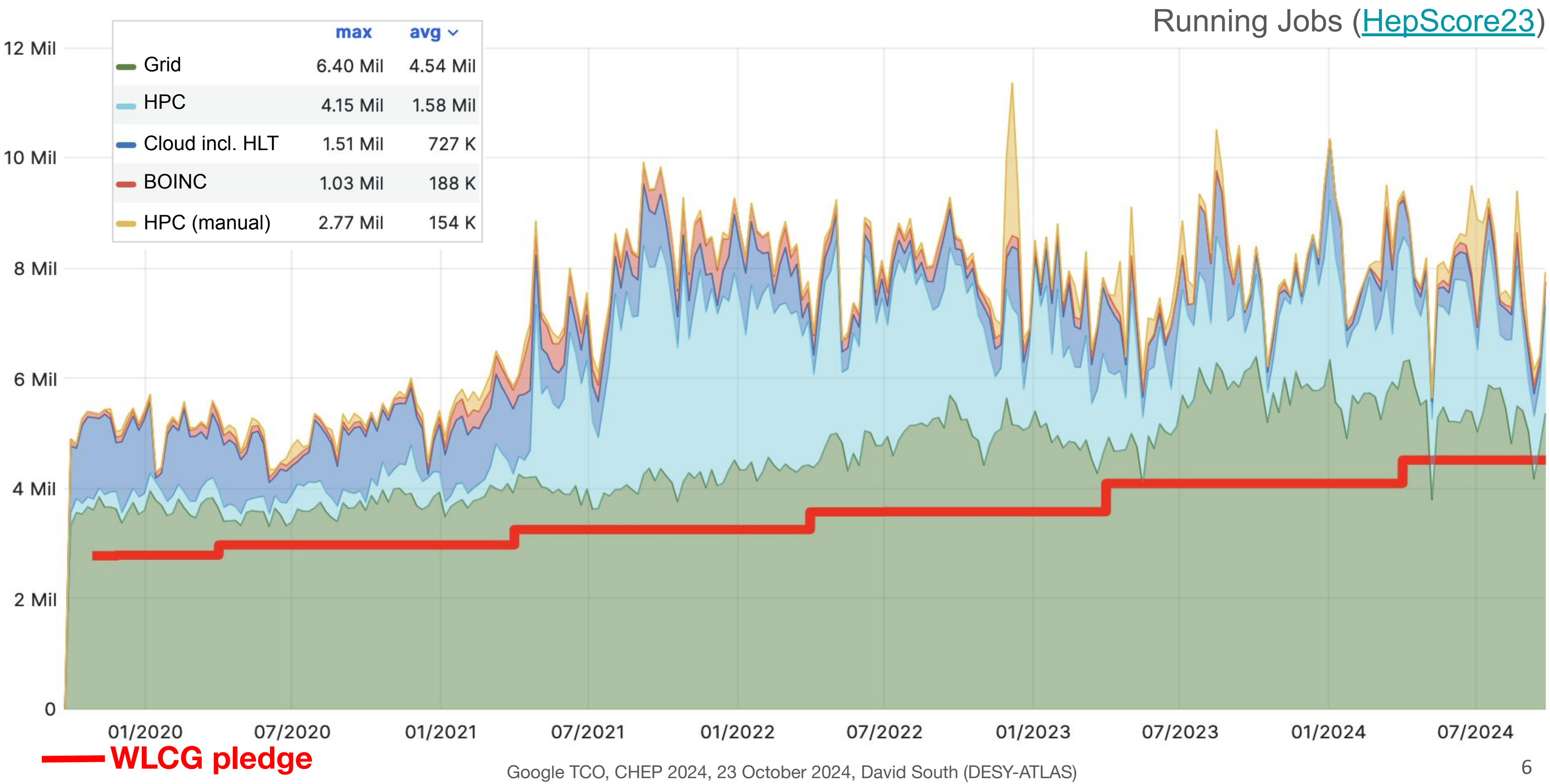
Coherent workload management using



ATLAS offline processing, last five years: Workflows

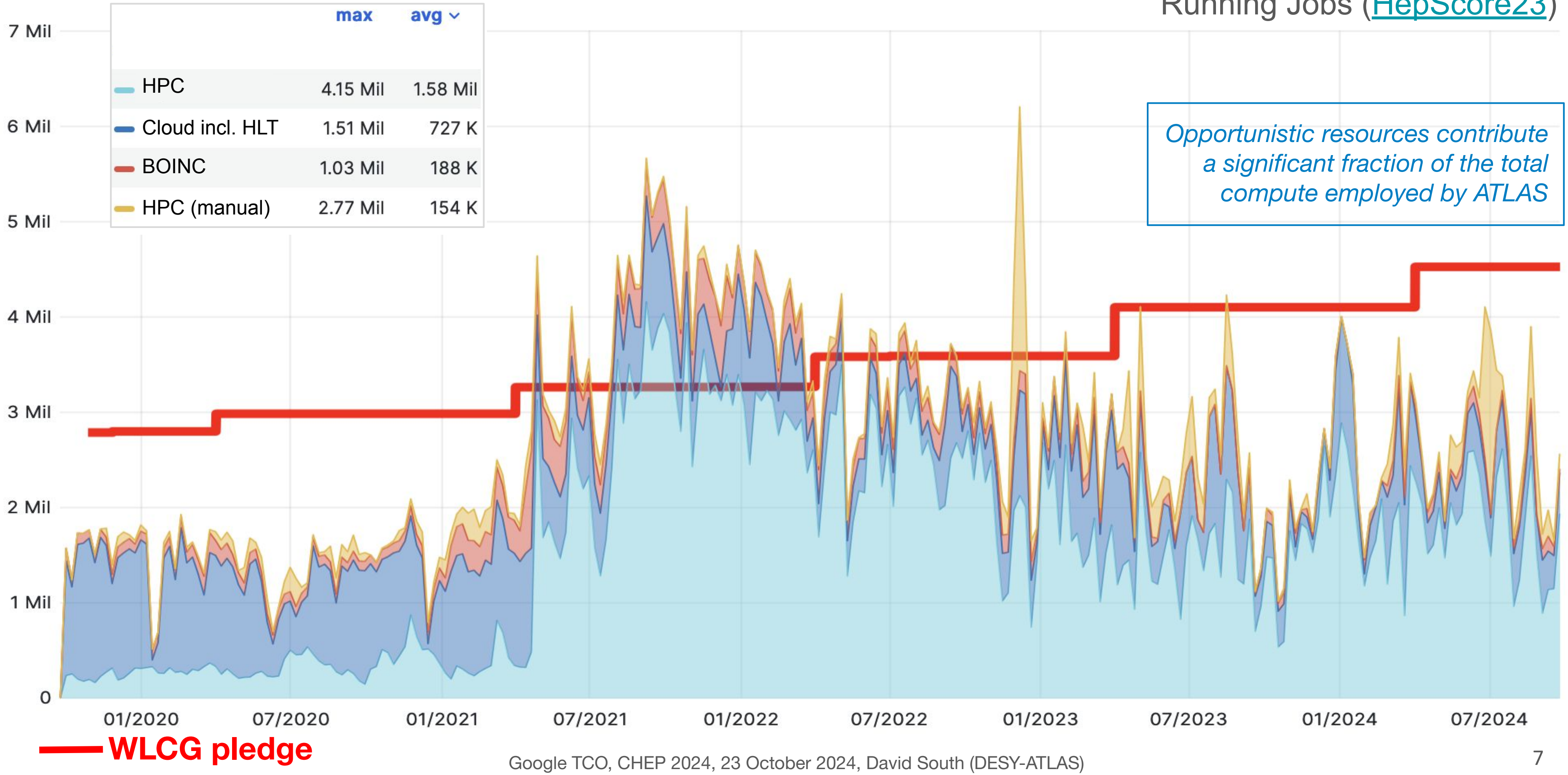


ATLAS offline processing, last five years: All resources



ATLAS offline processing, last five years: Opportunistic

Running Jobs (HepScore23)

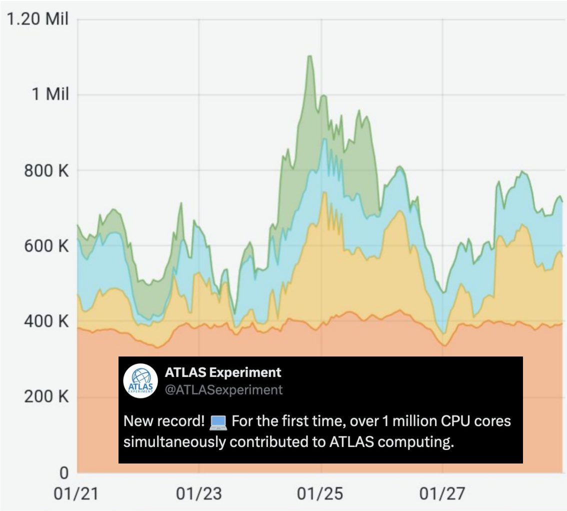


Opportunistic resources contribute a significant fraction of the total compute employed by ATLAS

Opportunistic resources: HPCs and Clouds

ATLAS has harnessed HPCs for over a decade now

- First generation e.g. **Cori** and **Titan** (US), **SuperMUC** (DE), the majority requiring dedicated task submission
- More recently we have had significant success running all ATLAS workflows using the EuroHPCs Vega in Slovenia and Karolina in Czechia
- Continue with Next Gen in US: Perlmutter and TACC
- Additionally, several sites deploy HPCs as part of their pledge including MareNostrum (Spain) and CSCS (CH)

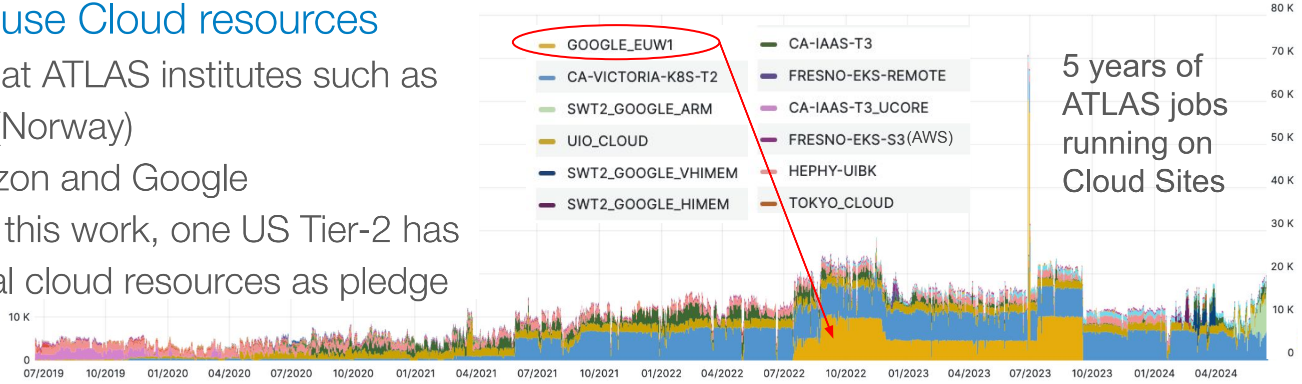


ATLAS running job slots

Feb 2023

ATLAS has also been able to use Cloud resources

- On-premise Cloud projects at ATLAS institutes such as **Victoria** (Canada) and **UiO** (Norway)
- Commercial clouds at Amazon and Google
- Very recently, and following this work, one US Tier-2 has begun to deploy commercial cloud resources as pledge



5 years of ATLAS jobs running on Cloud Sites

Introduction to ATLAS and Google

The ATLAS Google Project is long established as an R&D effort to demonstrate that the experiment can effectively use commercial cloud resources, within the distributed computing infrastructure

- The need to **secure additional resources** continues to grow towards HL-LHC
- WLCG sites are increasingly exploring methods to utilise new resources: **Major motivation to explore cloud**

Focus of previous projects:

- **Nov 2019 - Oct 2022**: Demonstrate integration of Google using ADC tools (PanDA, Harvester, Rucio,..)
- **Apr 2021 - Mar 2022**: Google Cloud Platform as analysis facility (Columnar Analysis, REANA, Active Learning)

This project: Running a cloud site in production for an extended period, **July 2022 - Sept 2023**

- Gain operational experience of running a cloud site at large scale, including:
  - **Evaluation all ATLAS workflows, including data reprocessing, and user analysis**
  - **Demonstrate rapid and efficient bursting to additional, large scale resources**
- Appraisal of a new GCP fixed-price subscription agreement model
  - **Evaluate the Total Cost of Ownership of employing a commercial cloud site at scale**
- (Further R&D building upon experience already gained, exploit unique opportunities available at Google)

Google data centre in St. Ghislain, Belgium

ATLAS EXPERIMENT

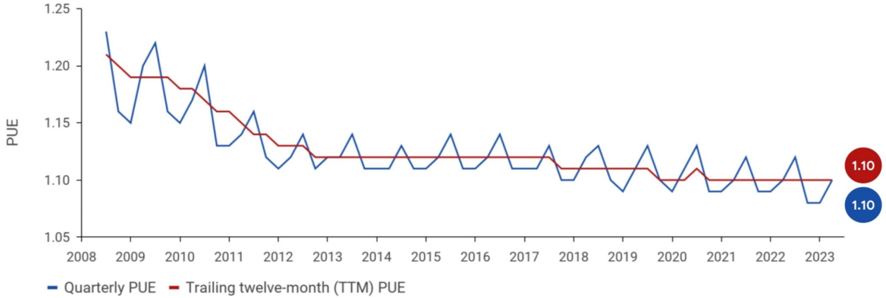
The main Google site used in this project is *europe-west1*, in Belgium

Google operates data centres at very large scale to achieve very high energy efficiency

Low carbon intensity, operating with net-zero greenhouse gas emissions

Google

Continuous PUE Improvement
Average PUE for all data centers



PUE = [Power Usage Effectiveness]

https://www.google.com/about/datacenters/efficiency

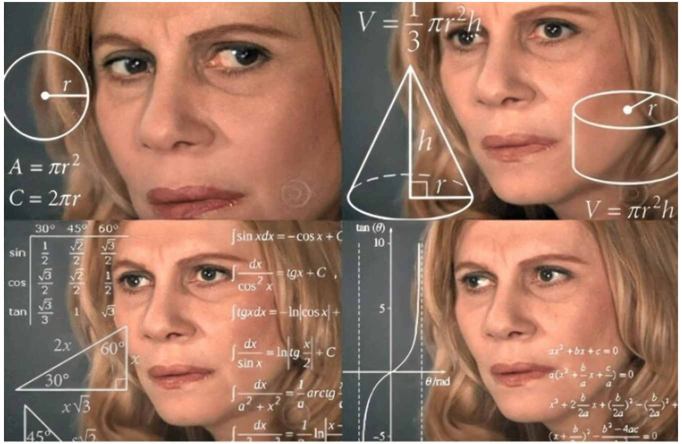
https://cloud.google.com/sustainability/region-carbon

Cost modelling

Commercial Cloud providers tend to use highly detailed, granular and itemised cost breakdowns for all services used during a contract, based on the so called *list-price*

Conversely, evaluating the effective costs of the usual resources employed by ATLAS is often difficult

- Large variation in funding methodologies with respect to the resource type, country... in some cases even from different Funding Agencies within the same country
- Local, unknown arrangements for example where some sites, typically at universities, do not directly pay electricity or WAN access costs
- They also might simply not want to tell us..



It's important to point out that this TCO evaluation is *not* attempting to directly compare the cost of running ATLAS jobs in the Google Cloud with the cost of running an ATLAS grid site

TCO methodology

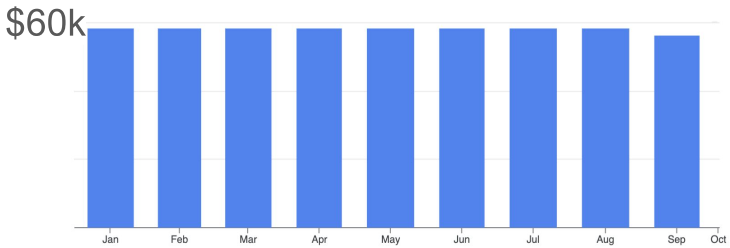
A *Subscription Agreement* contract was negotiated with Google, based on initial resources usage estimates, with significant discount with respect to the *list-price*

- Flat rate cost of around $57k/month or around $1900 per day
- CPU cores are provided as "spot instances", meaning preemption is possible
- Use available resources at any time at any scale, e.g. one month could be 10 PB / 3k cores, next month 1 PB / 20k cores

www.anyrgb.com

So where does that leave the TCO?

- Taking this literally, the TCO is the total price over the duration of the 15 month contract, so ~ $850k
- *Not really informative or very useful..*



https://makeameme.org

Therefore, what we have done in this evaluation is to **examine the relative contributions** of compute, storage and network to the **total *list-price* cost** under **different operating models** and for **different workflows** to understand how to best use such a resource

Setting up the site: Integration

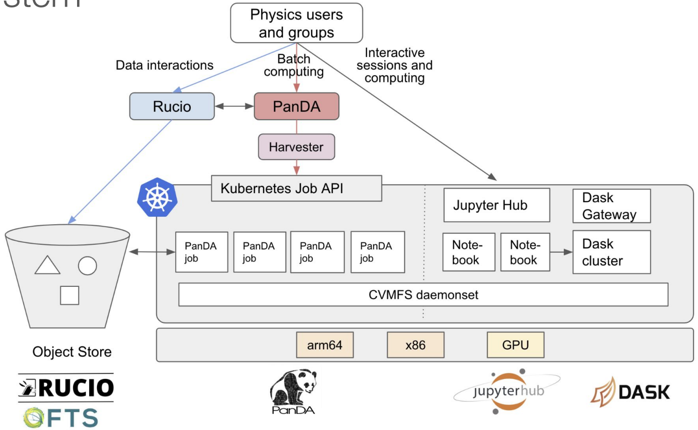
Google Cloud Compute within ATLAS Distributed Computing

- Transparent technical integration into ATLAS Workflow/Workload System
  - ProdSys (Tasks), PanDA (Jobs), Harvester (Cluster interaction)
  - Google Cloud Compute set up as Kubernetes cluster
  - Harvester interacts with standard Kubernetes API in Google
  - Applies various limits and selections based on workload

Software is delivered via CVMFS, same as for grid sites

Google Cloud Storage within ATLAS Distributed Computing

- Lots of developments in Rucio and FTS, including gfal/davix libraries
- Significant technical obstacles running it like a grid storage
  - Custom developments of cloud signature algorithms (S3v4, gclouds)
  - Custom load-balancers for cloud storage host certificates



F. Barreiro et al., *Operational experience and R&D results using the Google Cloud for high energy physics in the ATLAS experiment*, arXiv:2403.15873

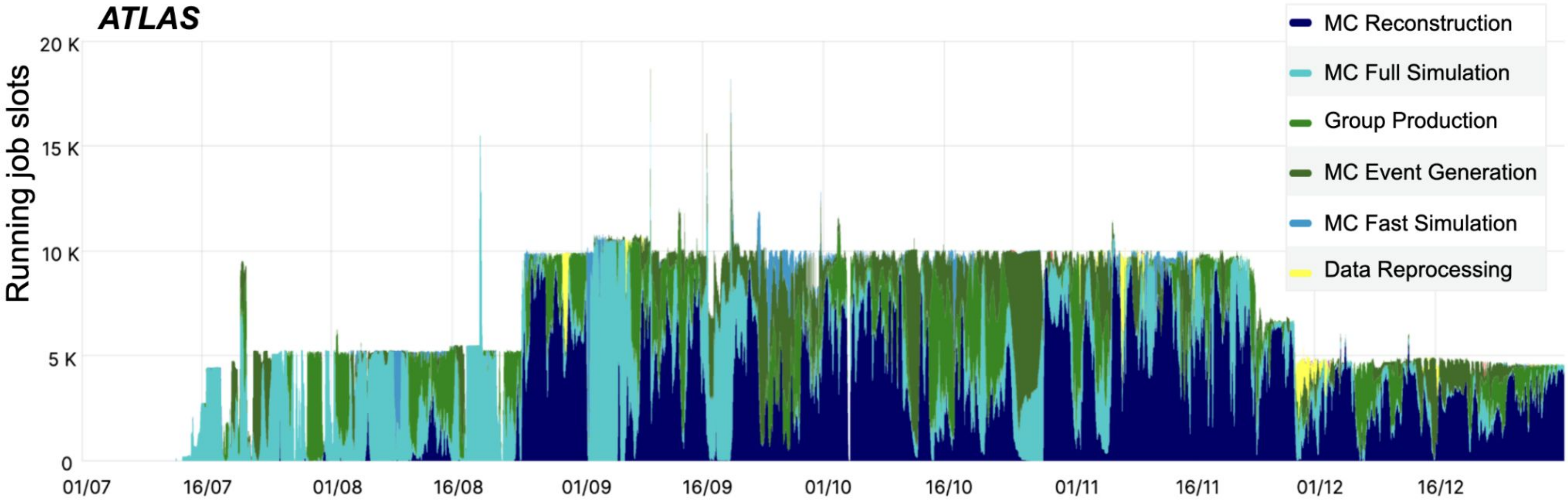
F. Barreiro et al., *Accelerating science: The usage of commercial clouds in ATLAS Distributed Computing*, CHEP 2023

M. Lassnig et al., *Extending Rucio with modern cloud storage support*, CHEP 2023

Running the site: A learning process

First few months: Gaining experience of running jobs on the site

- Alternating between 5k and 10k slots, understanding any set up issues
- Incorporating different production workflows via brokerage and fair shares



Running the site: Taming the egress

Meanwhile.. running all of those jobs, we started to accumulate significant data at the site

- Mostly AODs but also HITS, RDOs.. up to 6PB

These output data are attractive to other sites as production input to tasks running elsewhere

- Resulting egress up to 300 TB / day

Lower plot shows the impact on the *list-price*

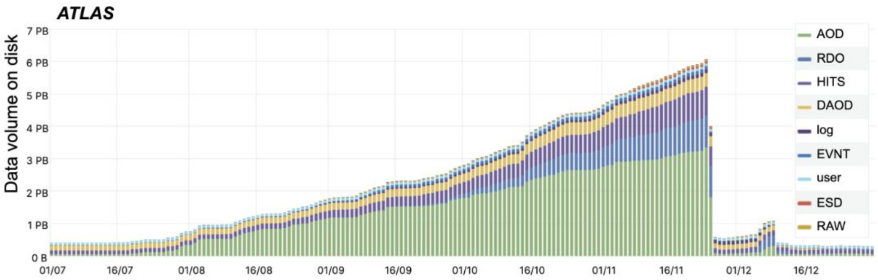
- Clear increase in cost from egress and storage

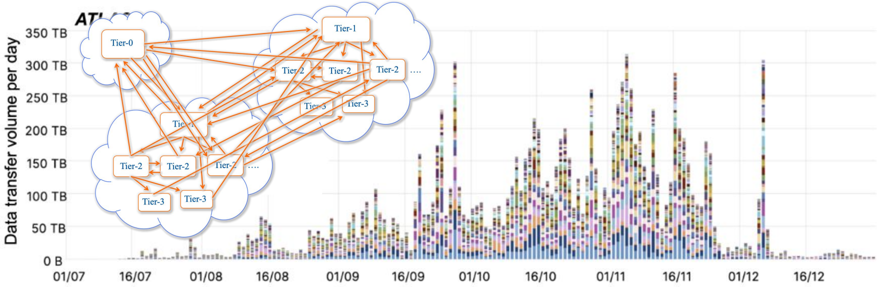
Re-establish control in two ways:

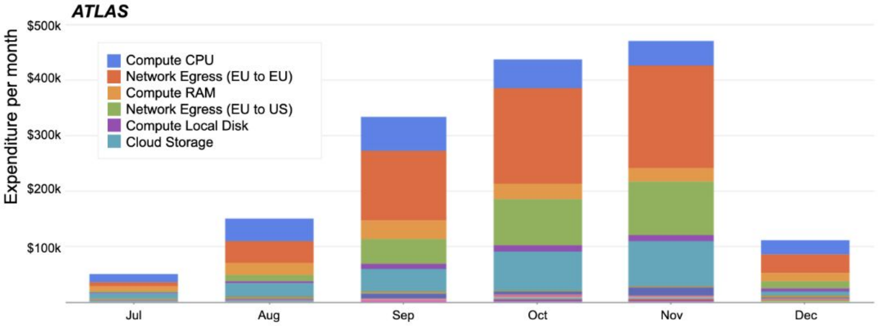
- Data deleted as soon as replication rules expire
- Distance to Google RSE from all other ATLAS sites set to very large value

Site operating as essentially as CPU only

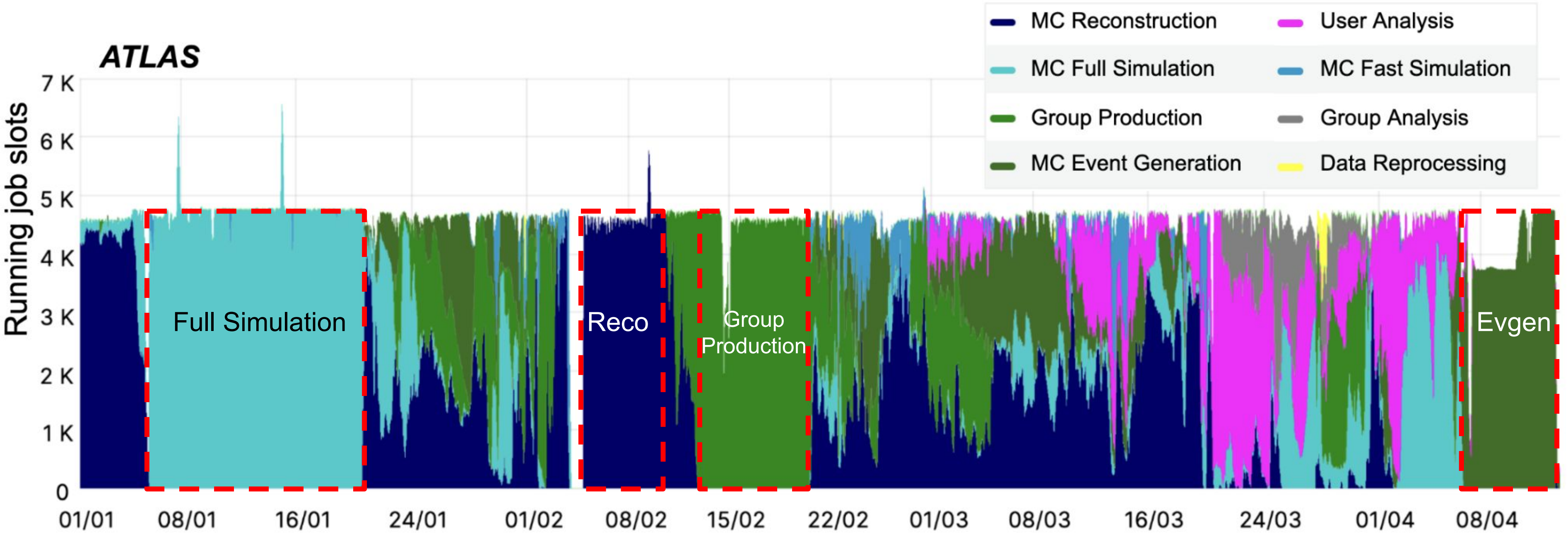
- Storage ~ 300 TB, egress ~5-10 TB / day
- Which is fine, but not really “grid-site-like”...







Running the site: Different workflows (1)



Next: Run with single workflows to analyse different relative contributions to total cost

- Aligned, mostly successfully, to ATLAS activity schedule (not always possible to plan too far in advance)

Also visible here: User analysis jobs started to run on the ATLAS Google Site from March

- Was not possible to completely fill the 5k slots due to unpredictable nature of analysis workflows

Running the site: Different workflows (2)

July brought an opportunity to run some of a major data reprocessing campaign at Google

- Usual Data Carousel mechanism for staging

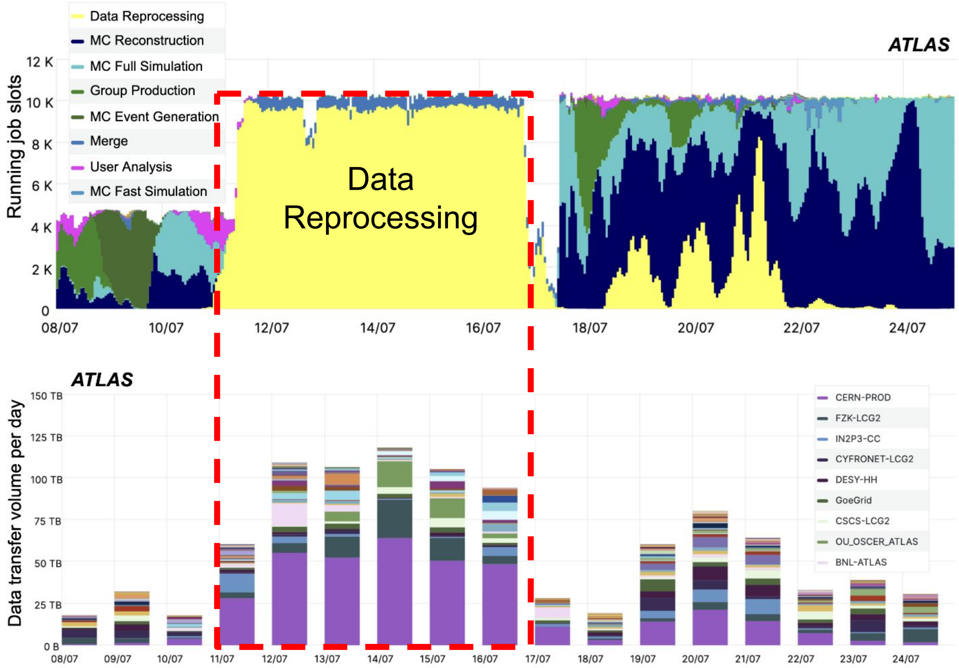
Rough calculation: 150 HS23s per event, 10 HS23/core = **10% of the 2022 data** on 10k cores (1250 x 8-core jobs) in 1 week

- And it was about correct! 😎
- As expected, Google CPU cores are better than 10 HS23/core, added more runs, and did 600M events in the end, about **15%** of the 2022 data

Egress during campaign up to around 100 TB per day, half coming from production output

- Remainder from moving other data out

Google also used as part of the holistic ATLAS resources for remainder of the data reprocessing



Running the site: Different workflows (3)

	MC Full Simulation 06/01–19/01	MC Reconstruction 05/02–09/02	Group Production 11/02–09/02	MC Event Generation 07/04–12/04	Data Reprocessing 12/07–16/07	Full Project 07/22–09/23
Compute	73%	65%	52%	47%	21%	28%
Storage	10%	10%	9%	32%	11%	20%
Network egress	7%	11%	26%	3%	63%	46%
Other	10%	14%	14%	19%	4%	6%

Looking at the relative cost of components different workflows, a few trends appear

- Running Group production only has markedly more egress than other MC workflows
- Evgen looks like it needs more storage, but this was due to an unrelated, concurrent replication of 1PB of DAODs to the ATLAS Google site: Multi-dimensional distributed system
- Most striking difference is for data reprocessing, where egress contributes 63% of the total cost

Running the site: Bursting (1)

Cloud computing is intrinsically highly elastic

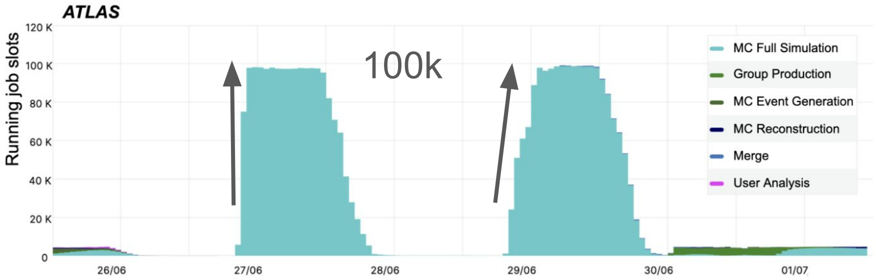
- Opportunity to acquire a significant number of additional resources, potentially at short notice
- Previous tests used to tens of thousands of cores to process a small number of events

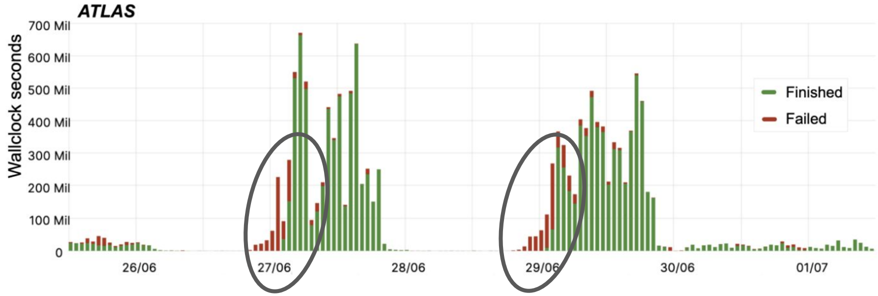
Large scale test: Produce a MC Full Simulation of 50M ttbar sample on 100k cores

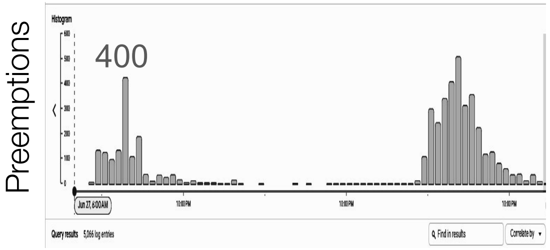
- 25k 8-core jobs with 2000 events each
- Each job 6-8h; each request took ~1 day
- Two iterations with different ramp up profiles

Between 11-13% lost wall time, mainly from rate of cvmfs deployment

Preemptions evict 4-8 jobs on full nodes, but still relatively low level, even at 100k running cores







Running the site: Bursting (2)

Cloud computing is intrinsically highly elastic

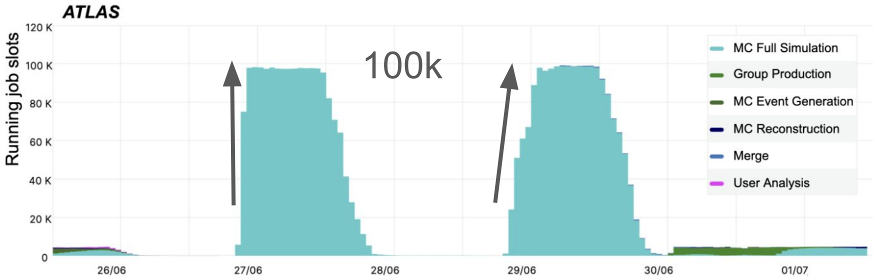
- Opportunity to acquire a significant number of additional resources, potentially at short notice
- Previous tests used to tens of thousands of cores to process a small number of events

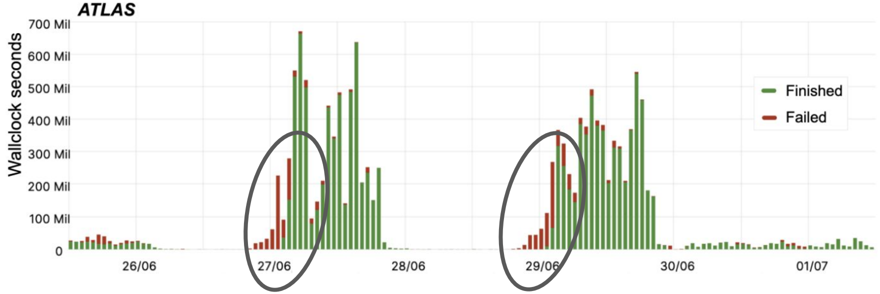
Successfully processed 50M events in 24h

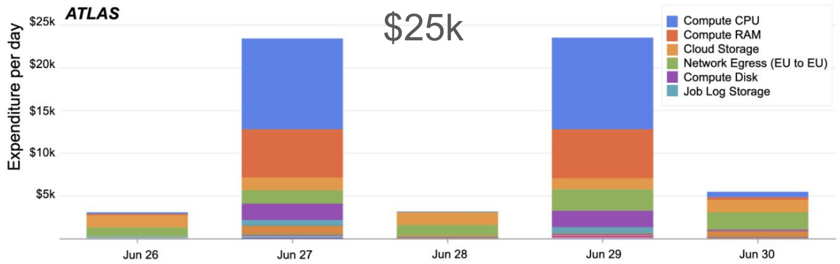
- At a *list-price* cost of around $25k each time
- Shows the clear benefit of bursting on cloud

Control tests on all resources took 8-10 days, even when most jobs on special (large) resources like P1 and Vega

- Advantage of bursting is to start from clean, empty resources, something not usually available to ATLAS on demand







The last few months: Scaling up the site again

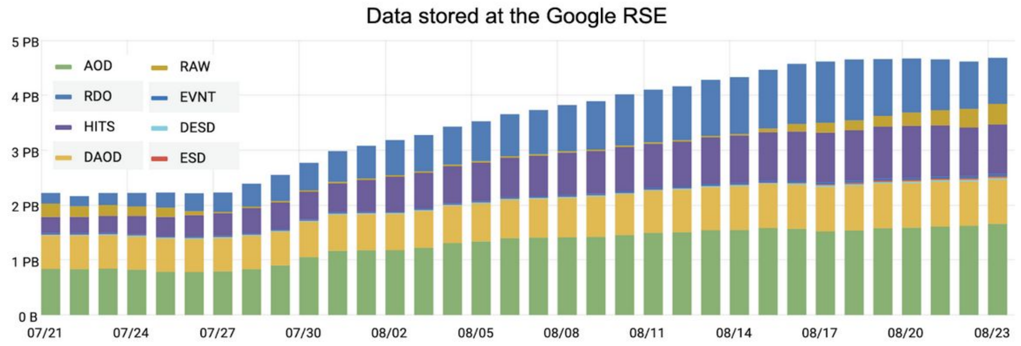
Following the data reprocessing campaign, scaled up site again for the last few months of the project

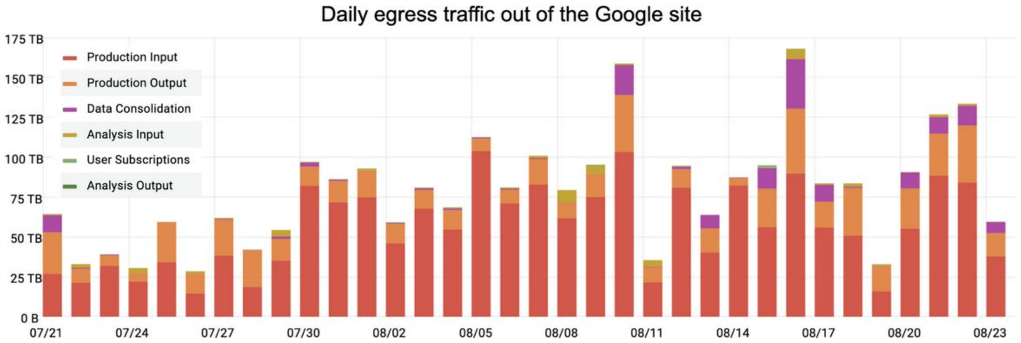
- Up to 5 PB of storage alongside the 10k running cores

Some control left in place, Rucio distances only decreased to German T1 and sites within the DE cloud

Slow increase of cached data as expected, along with a corresponding increase in egress of data

- Dominated by production input moving to other sites





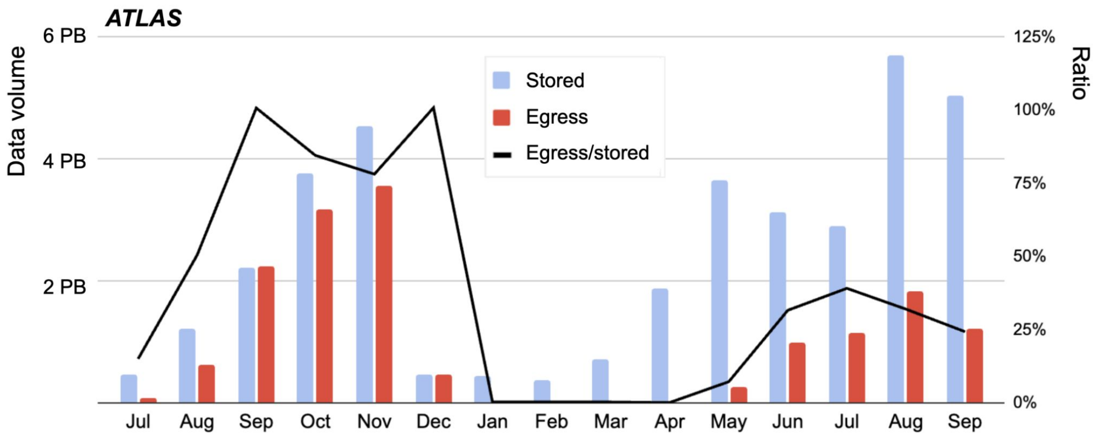
Network overview for the duration of the project

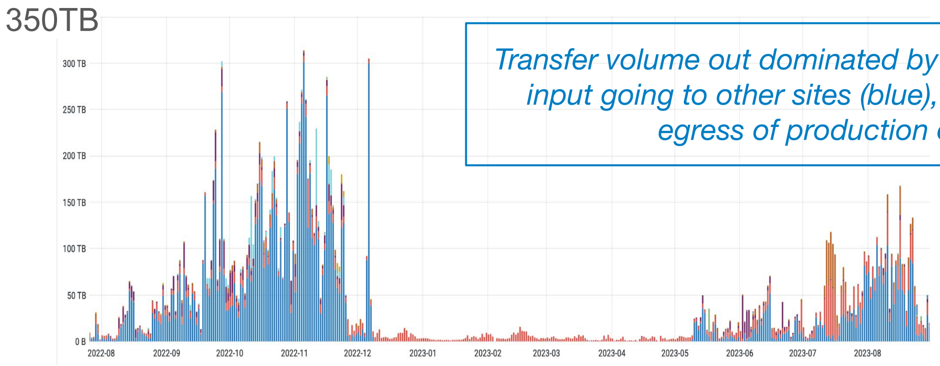
Clear distinction visible between when the site is operating with significant storage and more like a typical grid site and when it is essentially only CPU resources

In the first period, almost all data produced at the site is egressed out

In the second period, significantly less so, for two main reasons:

- The site was already established, with data accumulated over time rather than all being newly produced
- Only limited attractiveness of data stored at the site for use as input elsewhere





Transfer volume out dominated by production input going to other sites (blue), rather than egress of production output (red)

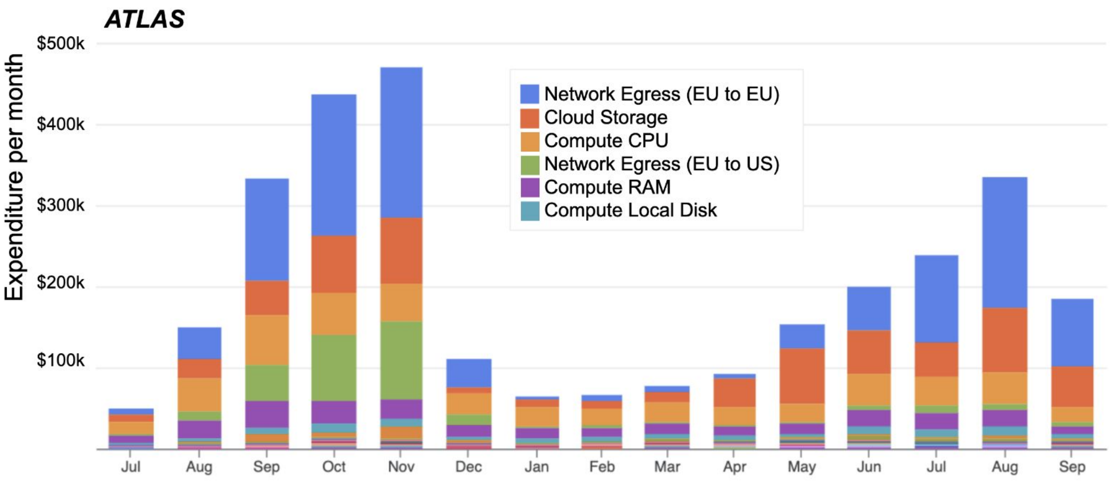
Cost overview for the duration of the project

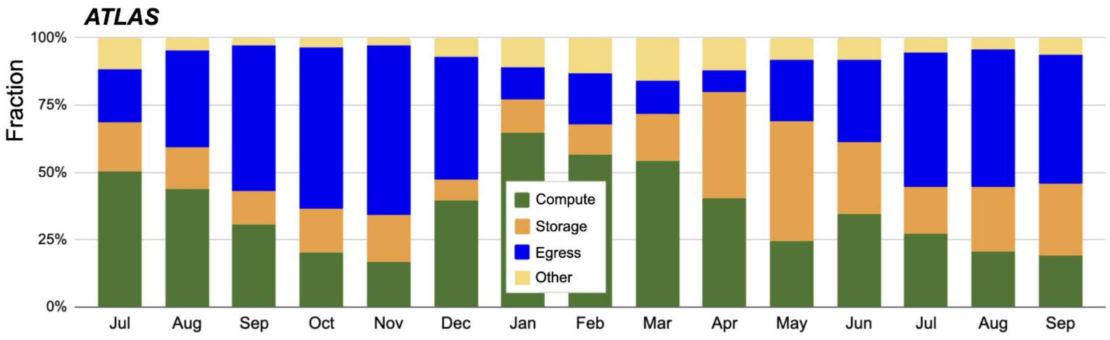
Monthly *list-price* cost profile

- The whole project in one plot
- Total for compute rather stable, but variations in storage and network

Looking at the relative contributions

- Different components dominate the cost, depending on the main activity
- Egress is most significant when the site is configured more openly and has significant cached data
- When there is only limited egress, compute is the main contribution





Summary: What is the cloud good for?

Technically the project was successful, establishing an effective ATLAS site in a commercial cloud

- Operation of the site at a very large scale and for an extended period of time is feasible and requires little effort
- ATLAS workflow and data management tools are adequate to adjust the cloud site configuration
- Integration solutions developed in this project essentially are **cloud-agnostic**, avoiding vendor lock-in

Commercial cloud computing is an ideal solution for providing additional, **stable** CPU resources

- We saw an overall rate of 5% lost wall-clock from failed jobs over the duration of the project (20% of which came from preemptions), resulting in an **observed eviction rate between 1-2%**
- Additionally, the Google site had **negligible downtime**

Seamless integration of cloud-based storage was also demonstrated, but *list-price* network costs incurred from egress by some workflows can be significant

- We cannot yet say commercial cloud is completely suitable for **all ATLAS workflows**

Resource bursting has been shown to be **very effective**, albeit at significant *list-price* cost

Summary: How about the costs?

The TCO evaluation has shown that without a subscription agreement, the cost of commercial cloud resources is **significantly greater** than running such resources on-premise

- A favourable agreement that makes sense to both the cloud provider and the client is a necessity

The Google Cloud resources used during this project cost a total of $3.2M at *list–price* compared to the $850k paid via the subscription agreement, representing a discount of 73%

ATLAS used **3.7 times more** resources than were purchased via the subscription agreement

- i.e. the resources used during this project would have been **270% more expensive** at *list-price*

This most obvious savings were in the costs associated with the 24h bursting tests at **$25k each, which is considerably more than the $1900 per day** rate of the subscription agreement

Summary: What's next?

Further investigations into running an ATLAS Google Site, with longer single workflow studies and a proper examination of analysis jobs, which are unique due to their highly variable nature

- The first ATLAS Tier-2 site has begun to provide resources to ATLAS via Google (also happen to be ARM)

Investigate widespread peering with commercial cloud for a cost-effective route into LHCONE

- If dedicated network links can be established, this can drastically reduce egress costs
- Controlling costs like egress may lead to more favourable future subscription agreements

ATLAS, together with the WLCG, also need to understand how such resources fit into the pledge structure especially if commercial clouds are employed by a significant number of sites

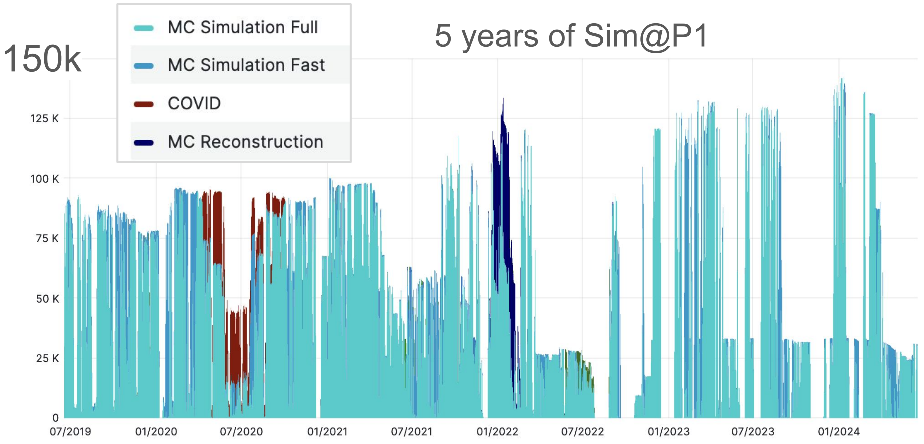
- Varying regional regulations and procurement processes add complexity to the management and planning

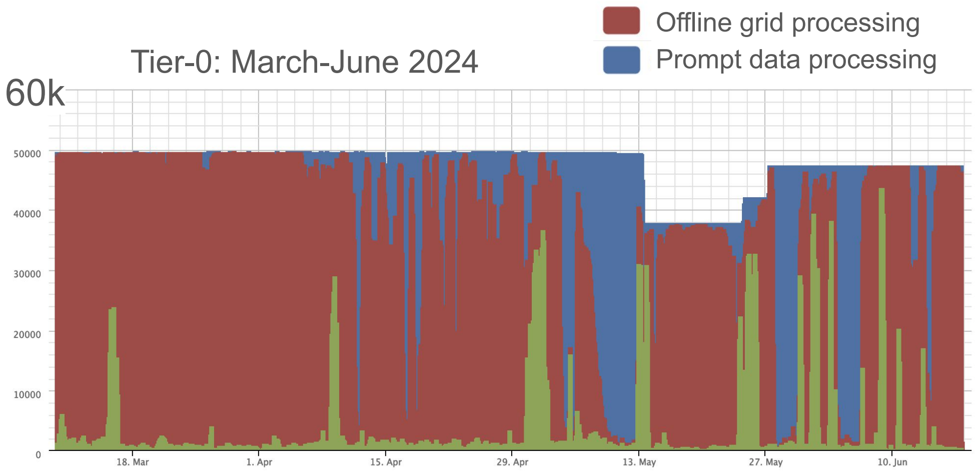
Continued access to non-standard resources like GPUs and ARM CPUs to assist validation

- Google Cloud resources have allowed rapid and elastic utilisation of a wide variety of hardware options
  - *For further details on this, see talk by J. Elmsheuser this afternoon*

Back Up

Opportunistic resources: CERN





ATLAS has significant computing resources at CERN in order to promptly process the data

- When there's no collisions, they are available for production!

The ATLAS High Level Trigger farm has almost 150k job slots

- Available for offline when there are no LHC collisions, mostly used for simulation jobs as a Cloud resource

The Tier-0 farm has around 50k job slots and is used to promptly process the collision data

- Only fully employed when data taking fully ramped up, otherwise used for offline processing as grid