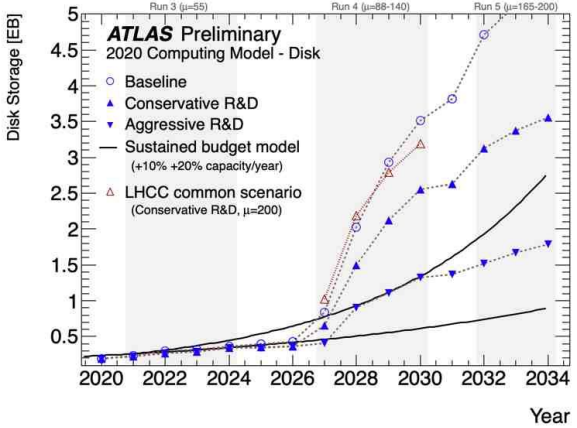
Seamless integration of commercial Clouds with ATLAS Distributed Computing

Johannes Elmsheuser (BNL)
on behalf of the ATLAS/Google/AWS R&D teams

May 2021
vCHEP21

Why Cloud services ?

Run2	$20*10^9$ data + $40*10^9$ MC events 50 kB/evt = 3 PB single DAOD >100 PB Run2 DAODs on disk
Run3	expect more data+MC events
HL-LHC	expect even more data+MC events



Goals:

- Process fraction of data in Google/Amazon by individual analysers through PanDA jobs or interactively using data formats with small event sizes
- Establish collaboration with Google team
- **Use generic solutions** and fit Google and/or Amazon compute and storage into Rucio and PanDA ecosystem as a Cloud site and **explore and learn** new technologies like Kubernetes, columnar storage etc.
- Explore interactive analysis possibilities with Cloud add-ons like ML or GPUs

N.B. Vera Rubin Observatory/LSST started to use Google Cloud as interim data facility for next 3 years (link)

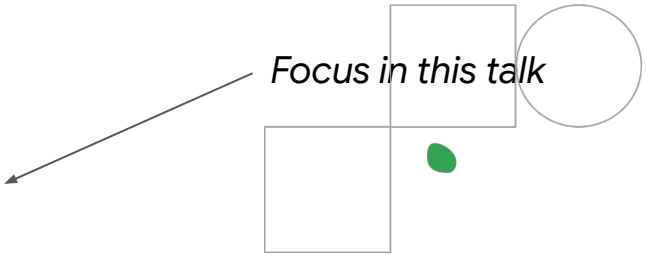
ATLAS+Google R&D and Work on Amazon in past years

Timeline

- 2013/15 Initial setups with prototypes in GoogleCloudPlatform
- 2017/18 Established USATLAS, Rucio + Google R&D
- 2020 Intensive R&D with different tracks (see below)
- 2020 Setups from Google explored on AmazonWebServices (funded by California State University and Amazon)
- Spring 2021 Continue Google R&D (funded by USATLAS and Google) with focus on user analysis

Google R&D Tracks (as initially defined in 2019)

- Track 1: Data Management Across **Hot/Cold Storage**
- Track 2: **Machine Learning**, TPU vs. GPU for GNN training
- Track 3: **Optimized I/O** and data formats for object storage
- Track 4: **End user analysis** conducted worldwide at PB scale
- Additional Track: LSST/Vera C. Rubin Observatory



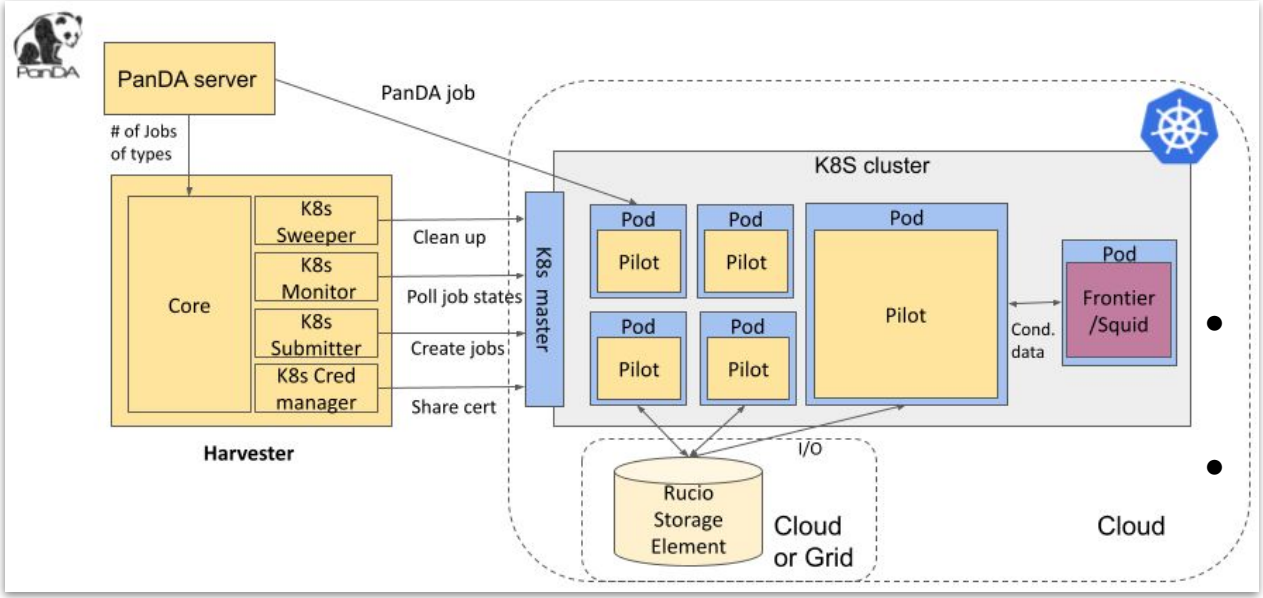
Step 1: Cloud Storage support in Rucio

- Cloud storage setup as Rucio storage with **3rd-party-copy File Transfer Service (FTS) transfers**
- **Fully validated** at 10 grid sites with transfers up to 15 GB/s over hours
  - For GoogleCloudStorage: need to have Google CA cert in IGTF in the long term
- Balance transfer to/from Cloud storage through General Public Network
- Future large scale tests put on hold due to **large egress (exit cloud network) costs**
- Upload/download implemented in Rucio in generic way for all types of clouds, such as GCS, AWS, S3-compatible (MinIO), SWIFT-compatible (OpenStack)
- Open point: User quota/limit handling - planned to be addressed through the ACL

Step 2: Kubernetes batch integration in PanDA

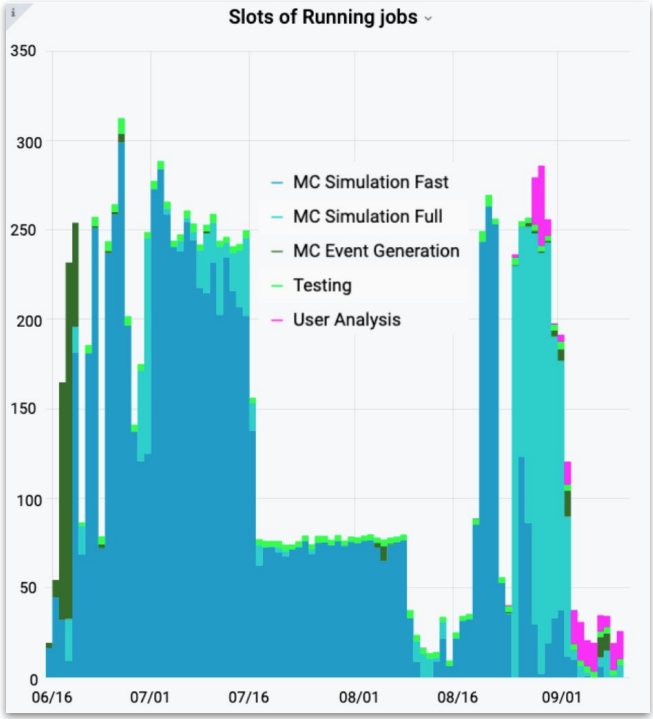
Started with Geant4/Fast simulation with storage at CERN

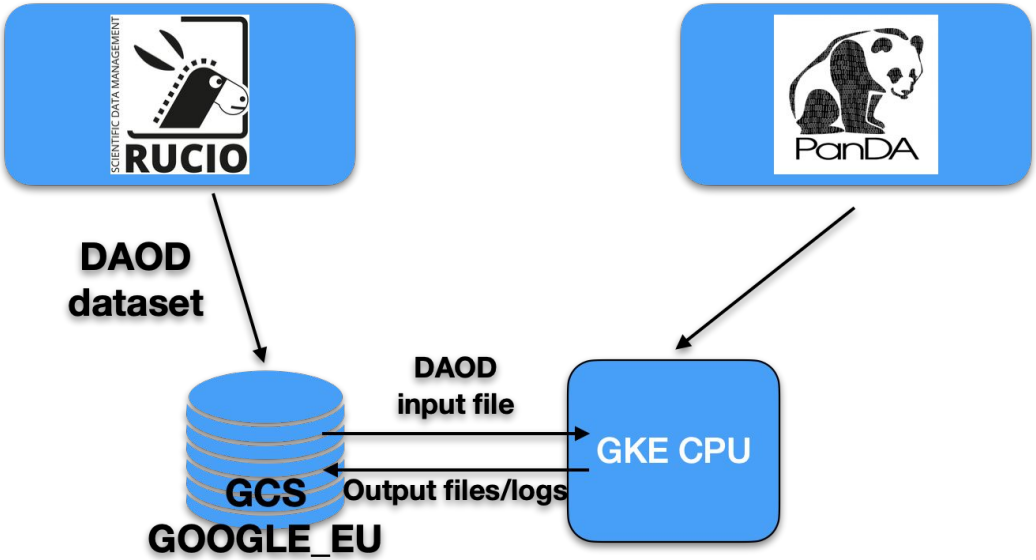
- GKE (Google) / EKS (Amazon) setup for compute
- Very light I/O jobs



- **CVMFS:** Installed through daemonset + Kubernetes (K8s) volumes
- **Frontier/Squid:** Installed on dedicated Pod
- **Preemptible nodes (Google) or Spot (Amazon):**
  - Causing most of the failures
  - Limiting job duration to <5 hours
  - Attractive deal: 80% cost reduction, slightly higher failure rate
- **Autoscaled cluster**
  - Cluster ramps down and lowers the cost when no jobs queued
- **Costs** (Google, remote storage, 120-160 cores)
  - July: 2.3k USD/mo. (76.6 USD/day)
  - Aug: 1.7k USD/mo. (54.4 USD/day)

Step 3: User Analysis with GKE and GCS





How ?

- Replicate dataset to GCS Rucio storage GOOGLE_EU
- Run regular ATLAS analysis submitted to PanDA with/without systematics (30 min/20h)
- Store outputs back to GOOGLE_EU

Difficult learning/setup process due to heavy I/0 and throttling on Google and quirks with direct I/O

R&D in 2021

- Plan to continue the R&D on Google and Amazon **using generic solutions**
- Tracks with focus on **user analysis**:
  - Columnar Analysis:
    - Process 100 TB in compact ROOT or parquet format on S3 storage
    - Explore uproot and parallel processing with dask and jupyter notebooks
  - Machine Learning (GPU/TPU)
  - Robust Kubernetes GKE and Rucio setup on the Cloud

- Lessons learnt also important for Analysis Facilities under discussion for HL-LHC
- Many easy to use Google services: Bigquery, AutoML, ...
  -> added value for an Analysis Facility in the Cloud

Summary and Conclusions

Summary

- Successfully integrated Google and Amazon Cloud into PanDA/Rucio
- Important lessons learnt toward the HL-LHC data and CPU challenges
- Will focus on user analysis related topics in further R&D

Not discussed in detail in the talk here, but in the paper:

- A simulation of Cloud data management
- More technical aspects the Rucio and Kubernetes setup in the Cloud
- Details about distributed analysis and production on the worker node:
  - Direct I/O and copy-to-scratch of inputs with signed TURLs
  - Additional workflows using e.g. GPUs could be added
- Cloud site list prices are higher than Grid sites, but cloud resources can be added very easily on-demand to existing facilities in situations where additional capacity is required at short notice.

Teams

Alexei Klimentov, Brookhaven National Laboratory

Kaushik De, University of Texas Arlington

Fernando Barreiro Megino, University of Texas at Arlington

Johannes Elmsheuser, Brookhaven National Laboratory

Mario Lassnig, CERN

Cedric Serfon, BNL

Misha Borodin, University Iowa

Tobias Wegner, University of Wuppertal

Xianyang Ju, Lawrence Berkeley National Laboratory

Paolo Calafiura,Lawrence Berkeley National Laboratory

Andy Hanushevsky, SLAC National Accelerator Laboratory

Ricardo Rocha, CERN

Siarhei Padolski, Brookhaven National Laboratory

Doug Benjamin, Argonne National Laboratory,

Karan Bhatia, Google Cloud

Ema Kaminskaya, Google Cloud

Miles Euell, Google Cloud

Usman Qureshi, Google Cloud

Ross Thomson, Google Cloud

Kevin Jameson, Google Cloud

Dom Zippilli, Google Cloud

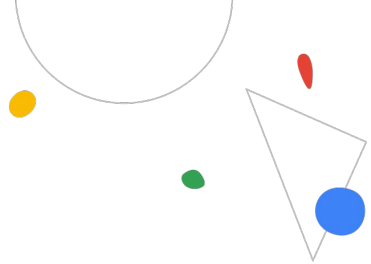
Kevin Kissel, Google Cloud

Harinder Singh Bawa (California State University)

Nikolai Hartmann (LMU Munich)

Lukas Heinrich (CERN)

Fang-Ying Tsai (Stony Brook)



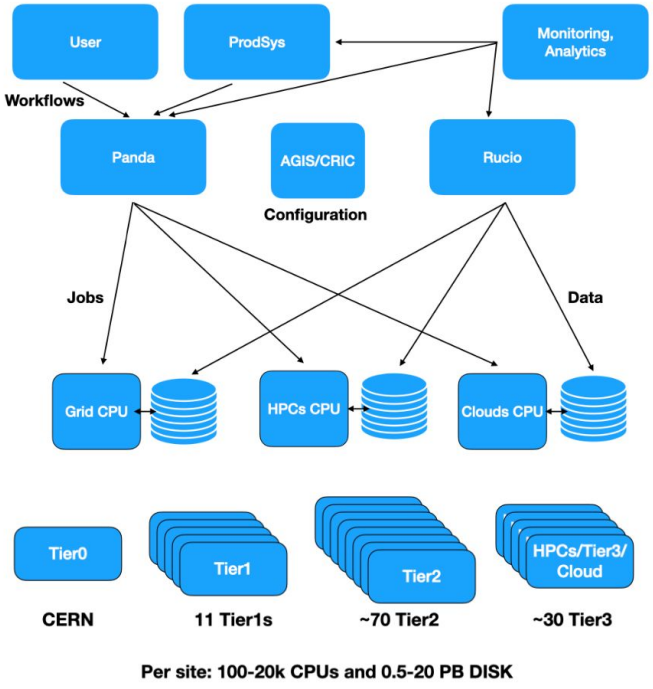
Backup

Reminder: ATLAS distributed computing setup

ATLAS DISTRIBUTED COMPUTING OVERVIEW

The ATLAS distributed computing system is centered around:

- **Workload management system**: PanDA
- **Data management system**: Rucio
- Many **additional components**: AGIS/CRIC, ProdSys, Analytics, ...
- **Resources**: WLCG grid sites, Tier0, HPCs, Boinc, Cloud
- **Shifters**: Grid, Expert and Analysis (ADCoS, CRC, DAST)
- Runs 24/7 all 365 days per year



Step 3: GKE/GCS user analysis

What works:

- **Successful GKE/GCS integration for the first time with full Rucio/PanDA workflow**
- PoC for analysis works stable after extensive iterations of GKE node setup with copy-to-scratch input
- Essential to use powerful well connected GKE nodes
- Usage of preemptible nodes seems only useful for workloads under a few hours

What does not work (so far):

- “weak” GKE nodes
- ROOT direct I/O via DAVIX access of inputs to GCS broken - could avoid parts of local storage troubles

ToDo:

- Detailed cost estimation per wall clock hour or processed TB
- Scale to really large datasets, many users and long payloads