BROOKHAVEN
NATIONAL LABORATORY

Auf dem Weg zu schlankeren Sites

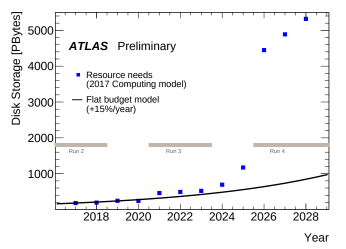
Johannes Elmsheuser

28 September 2018

Perspektiven für HEP Software und Computing in Deutschland, Wuppertal

Introduction

- In the past years LHC experiment built their customised workflow and data management systems on heterogenous resources
- Growing compute needs and data volumes of HL-LHC require to utilise many diverse resources, but also simplifications and more commonality among experiments

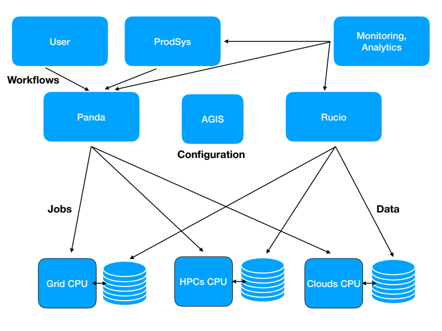


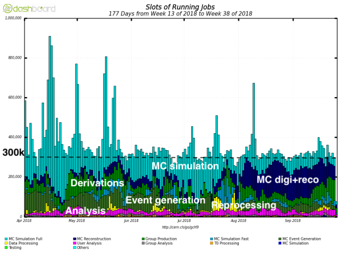
- General ideas in the following apply to all VOs - examples from ATLAS distributed computing - apologies for this slight bias

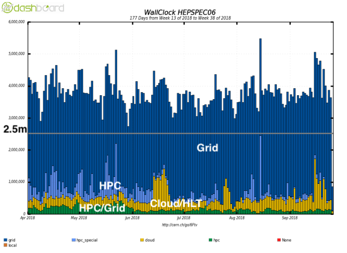
ATLAS DISTRIBUTED COMPUTING OVERVIEW

The ATLAS distributed computing system is centered around:

- Workflow management system PanDA and data management system Rucio
- Diverse resources: WLCG grid sites, Tier0, HPCs, Boinc, Cloud
- Many workflows, users, running jobs, >350 PB data on disk and tape

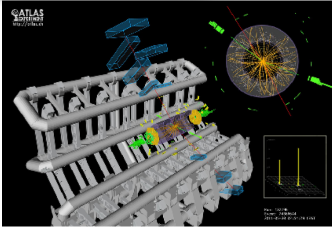






Simplified ATLAS data workflow

1 pp-collision event:



1 event:

Array of objects with sub-detector infos

Calorimeter ...

Inner detector ...

Muon detector ...

...

Array of objects with kinematic infos of physics objects

Electrons ...

Muons ...

Jets ...

...

Collision events are independent

Simulation

EVNT — Generation

HITS — Simulation

RDO — Reconstruction

ROOT file formats:

Data

RAW

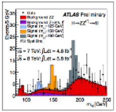
AOD — Derivation/Filtering

DAOD — Analysis

used in statistical analysis of many events

1 ROOT file:

Array of events: ...



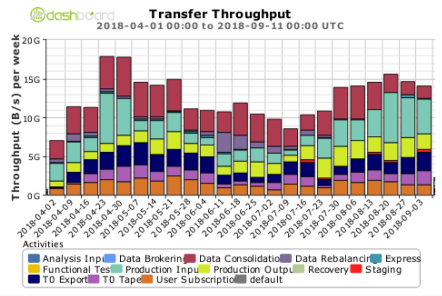
Workflow examples, metrics and typical costs - today

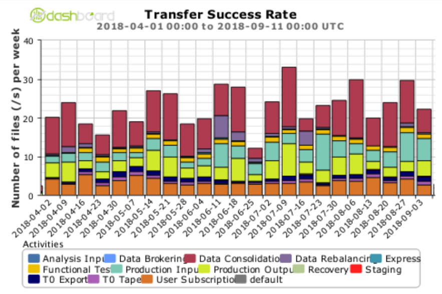
- Ideally: fast turn-around with high events/s - but every workflow has some **cost**
- Costs of workflows: **CPU, memory, disk** and **network I/O**
- Examples loosely categorised by:

Category	Workflow	Time/evt. [s]	Evt. size [MB]	CPU/Walltime [%]
CPU heavy	MC simulation	30-600	1	80-95
CPU + I/O	MC digitisation/reco	10-40	0.1-0.5	50-80
I/O heavy	data reco derivations analysis	0.1-10	0.1-0.5	30-80

- **Memory:** Fit into $\approx$ 2 GB/CPU core grid slot (can vary)
- **Network:** usually not directly specified, since input files replicated to sites in advance and then locally read (in ATLAS), only remote conditions DB access, in other VOs also remote input file access
- → Workflows fit differently on different resources

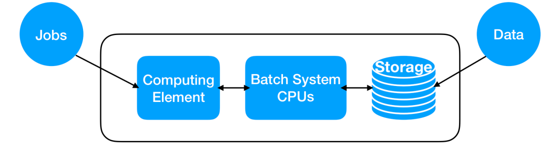
ATLAS data transfers - since April 2018





- Moving >1 PB, >20 GB/s, 1.5-2mio files per day
- Limited by the simultaneously transferred number of files the file transfer service (FTS) can push through
- Using FTS at CERN and BNL, FTS at RAL used at small scale only
- Data is asynchronously shipped to available CPUs

Light site setup - Compute



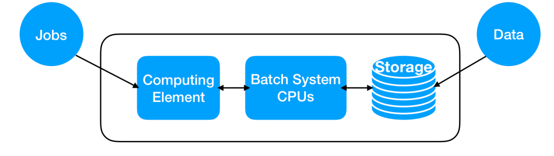
Computing element

- All LHC experiments use Pilot job based workflow management systems
- Use ARC CE or HTCondor-CE
- Future: Use something like Kubernetes, OpenShift for payload scheduling ?

Batch and CPUs

- Use HTCondor or Slurm
- Ideally use native OS (SL7) or Singularity/Container for worker nodes
- Backfill spare CPUs with Boinc ? ATLAS uses ≈11k CPU cores constantly for Simulation

Light site setup - Data



Storage

- Most of the daily operations errors are due to storage → more error resilient or simplification
- Popular SEs: dCache, DPM, EOS, Storm - if the site pledge is large enough and there is personpower available use one of these
- For smaller sites: Cache storage like ARC cache in production on NDGF or XCache (available/development)
- SRM-less setup with xrootd/webdav (example Uni Bonn)

Management

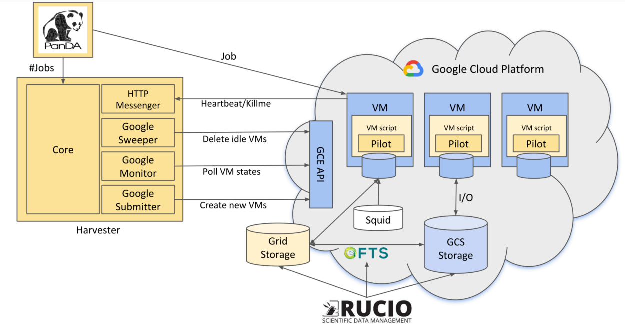
- Rucio centrally manages storage and transfers using FTS (asynchronous data delivery)

DOMA - data lake

- WLCG DOMA (data organisation, management, access) activities ramping-up - LHC experiments. WLCG and more are active (IMeetings in indico)
- Topics and working groups are forming:
  - 3rd-party-copy replacement of gridFTP by xroot/http
  - Storage access, caching, latency hiding, distributed storage
  - QoS (Quality of Service) of storages, Networks, Storage resource reporting
- Site caches well suited for light-weight sites
- Combine cache content info with workflow management system
- In production lower cache resuse - should be better for analysis

Google and/or other Clouds and site extensions

- R&D project with Google: full integration with Rucio and distributed analysis using PanDA/Harvester
- Data ocean idea: Store analysis inputs (DAOD/nanoAODs) on Google storage - fast access from Cloud or Grid
- Could be easily used as simple or bursty site extension
- Compare costs of Cloud vs. Grid



Personpower

- Shifters, developers and **local** experts are absolutely essential to support the heterogenous systems

Summary

- HL-LHC challenges: manage and store the large amounts of data
- Consolidate and improve present system with common projects and systems