OVERVIEW OF THE ATLAS DISTRIBUTED COMPUTING SYSTEM

Johannes Elmsheuser, Alessandro Di Girolamo on behalf of the ATLAS collaboration
11 July 2018
CHEP 2018, Sofia
ATLAS computing manages the resources to:

- Store the ATLAS data on disk and tape
- Process the data at the Tier0
- Reprocess it once year at Tier1/2 sites
- Run continuous MC simulation and reconstruction at all sites
- Run continuous user analysis

Again, LHC is delivering lots of data:

The ATLAS distributed computing system is performing very well
The ATLAS distributed computing system is centered around:

- **Workflow management system**: Panda (Talk 143)
- **Data management system**: Rucio (Talk 137)
- **Many additional components**: AGIS, ProdSys, Analytics, ...
- **Resources**: WLCG grid sites, Tier0, HPCs, Boinc, Cloud
- **Shifters**: Grid, Expert and Analysis (ADCoS, CRC, DAST)

More details: Poster 144, Poster 141
• Smooth steady running on the grid with full utilisation
• Well over pledge
• Additional CPUs from HPCs, Boinc and Sim@P1 (HLT farm) mainly for simulation (Talk 127, Talk 150)
• Note that HPC CPUs are usually "weaker" than grid CPUs
**Data Transfers and Disk Usage**

- Moving >1 PB, >20 GB/s, 1.5-2M files per day
- Limited by the simultaneously transferred number of files the file transfer service (FTS) can push through
- Data Volume exceeds 350 PB on disk and tape (below only datadisk) - Primary data (yellow) is partially replicated
• Large fraction of 2017 CPU usage in MC simulation, generation, reconstruction
• HPC and Cloud (a large fraction of it is the HLT farm when LHC is in a technical stop) add to MC simulation
Tier0 processing

Running job slots:

- Data taking pushed the infrastructure to the limit
- Powerful worker nodes: SDD, 4GB memory/core, Hyper threading switched off

Queued jobs:

- Tier0 data processing jobs and Grid jobs running on Tier0 when idle
How is the data analysed? Distributed Analysis beginning of 2018

- Large fraction of events are processed from DxAODs, but some walltime is devoted to process/generate other formats
- xAOD: primary physics output of reconstruction
- DxAOD: physics data products derived from xAOD
Continuous improvements and new features for BigPanda monitor (Talk 148)
**Evolution of the Systems - Workflow management**

**Resources**

- Efficient job scheduling, diverse resource usage capability with developments of Harvester, Event Service, Pilot2 (*Poster 137, Poster 130*)

**Global Fair-shares and unified queues (Poster 129)**

- Limit the cpu slots per activity and boost activity when requested
- Hierarchical implementation - sub-activities within activity have partial shares
- Unify the currently partitioned production and analysis queues into a single queue per site

**Harvester (Talk 135, Poster 161)**

- Developed for resource provisioning for timely optimisation of CPU allocation among various resource types (grid/HPCs/Clouds)
- First version for HPCs and Grid (see )

**Event Service and Event Streaming Service (Poster 155, Talk 154)**

- Fine grained event level processing instead of file level processing
Evolution of the Systems - Data management & Site infrastructure

Rucio: (Talk 137)

- Interest in Rucio by other HEP experiments and communities - Very successful Rucio Community Workshop in March
- Examples of new features: Adding rucio input file mover to panda pilot, site file cache awareness (Poster 138), object stores (Poster 162), dynamic data placement (Talk 140) zip archive file creation, Tape carousel

Data lake/ocean:

- A very successful R&D project with Google: integration with Rucio and distributed analysis using Harvester (Talk 133)
- Ramping up in the WLCG DOMA (data organisation, management, access) project to explore possibilities to overcome disk shortage in HL-LHC

Singularity and Containers: (Poster 163)

- Enabled Singularity with Docker images distributed via CVMFS
- Enable also container usage in user analysis containers and analysis preservation
Conclusions

- Production and data management system is robust, performant and can easily cope with higher Run-2 LHC performance
- Processing power, storage capacities and network used to its limits
- There are presently no scaling issues. Each subsystem able to sustain the large loads
- New workflows and technologies are being integrated but storage shortage has to be addressed in the future