ATLAS operations

Pre-GDB on experiment computing operations
24 February 2022

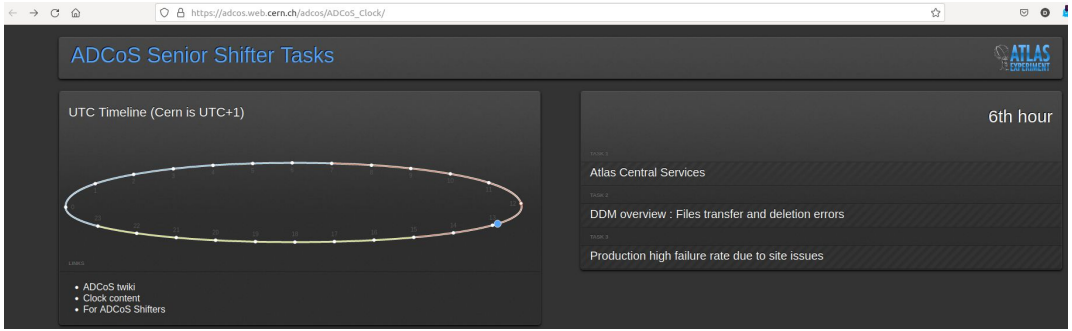
David Cameron (University of Oslo), David South (DESY)
With input from shift coordinators:
Peter Love, Helmut Wolters, Armen Vartapetian, Michal Svatos

Operations in ATLAS Distributed Computing

- Site admins (multi-VO)
  - People running the hardware and services at a site
  - Typically not a member of any experiment and little knowledge of experiment workflows
- Cloud squads
  - Clouds are regional groups of sites, historically a T1 and associated T2s
    - Today the connection is more for support
  - “Cloud squads” are a liaison between a region of sites and central ops
  - Typically ATLAS members who follow what goes on in ATLAS computing and help and advise site admins
- Shifters: ADCoS + CRC
  - ATLAS Distributed Computing Operations Shifts (ADCoS)
  - Computing Run Coordinator (CRC)
- Central ops team
  - Dedicated experts overseeing the management of data and jobs on the grid
- Central services operations
  - Sysadmins for the ATLAS computing services at CERN (Rucio, PanDA, software build machines, etc)

ADCoS shifts

- Institute-based, 8 hour shifts, 24/7 (~80% coverage)
- Checklist of tasks looking for problems, following shift clock (set list of tasks per hour)
- If any problem, escalate following well-established procedure
  - e.g GGUS ticket, JIRA ticket, mail to expert list, CRC
  - Detailed logging of all actions in ATLAS e-logbook
- Most common problems are site storage and transfers
  - Storage servers down, hardware failures, lost/corrupted files, site connectivity
- Less common, but harder to diagnose are job-related problems



CRC shifts

- Pre-COVID:
  - One week shift, at CERN, 8/7 (working hours incl weekend)
- From the start of 2021:
  - One month shift, from institute, 50% of working time
  - With this model, went from ~50% coverage to ~100%
- Shift model when Run 3 starts still TBD, probably a hybrid of the two

- CRC coordinates and reports current ADC operational issues
  - Reports summary of issues in daily ops meeting and summary in weekly ADC-wide meeting
  - Attends weekly WLCG ops
  - Follows up unanswered mails/tickets
  - During data taking follows LHC/ATLAS status, data export etc
- “Expert-level” shift for those who already actively follow or are involved in ADC

Central ops team

- Team of experts roughly split into two areas

- DPA: Distributed production and analysis
  - Ensure smooth running of central production and user analysis, with appropriate shares and priorities to each activity
  - Support for difficult failure modes, communication with sites, production managers and other ADC experts
  - Feedback of features/bugs in PanDA and related systems
- DDM: Distributed data management
  - Ensure smooth data transfers, with appropriate shares and priorities to each activity
  - Managing disk space on grid storage, running deletion campaigns, decommissioning storage
  - Investigating unexpected behaviour from Rucio or storage
  - Feedback of features/bugs in Rucio

- Discussion of day to day issues at daily ops meeting, dedicated weekly meeting for longer-term plans
  - Interaction with all relevant systems: monitoring, analytics, CRIC, Frontier, etc.

What works well and what doesn't

- Overall system is stable and mature
- Very well documented procedures lower the entry barrier for ADCoS
- CRC even at 50% relieves load from central ops team
- Storage issues usually well-understood (same problems happen often)

- Central ops team often in "fire-fighting" mode with no time to work on longer-term improvements or automation
- Diagnosing job-related issues is difficult
- System is stable and mature - so any changes require a lot of re-education
- Many individuals who are a single point of failure

Thoughts on sharing effort across experiments

- Overall difficult without knowing what other experiments do
  - So useful to have this meeting!
- Most important prerequisite is unified monitoring
  - Data challenge dashboard is an important step in that direction
- Areas where shared services and infrastructure are used are best candidates for shared operation
  - GGUS, FTS, storage, network
  - However seemingly experiment-independent problems can often require some experiment knowledge
- Higher-level services are too experiment-specific even when technology is common
  - Even Rucio, used by ATLAS and CMS, is run in very different way, especially interfaces to other systems
  - At this level the access barriers (technological and cultural) are much higher
- There is a culture of secrecy in each experiment, for good scientific reasons
  - Need to find a good balance between scientific independence and cost-savings
- Final thought: Maybe investing in automation is a better means of cost saving than sharing (labour-intensive) operations?