ATLAS Computing Operations

A first, partial view

Jaroslava Schovancová (CERN IT)
with ideas and contributions from Alessandro Di Girolamo, Mario Lassnig, Michal Svatos, Armen Vartapetian, Mayuko Maeno, Farida Fassi, Sabine Crepe, Helmut Wolters

Operational Intelligence meeting, 2019-04-01

Operational Intelligence context

- Quasi-real time analytics of streams of variate data, based on algorithms[to be developed], take operational actions.
- Really very wide activity
  - requires simple tools for some things, but might require other very complex tools for other things
- We would like to understand what are the problems to be addressed

⇒ Let's start simple & small, and build on top of that.

A proposal:

Let's start with DDM-related stuff:

a window of 4 hrs of failures at src/dest,
create a digest email to shifters
with suggestions what to do.

ATLAS Computing Operations at a glance

- Operating the **ATLAS Distributed Computing (ADC) infrastructure** in order to support the ATLAS Physics programme mission
  - **Resources**: heterogeneous (grid, cloud, HPC, opportunistic; compute, storage, GPU, network, ...) at 120+ sites (T0/T1s/T2s/T3s); **100+ people, ~ 50 FTEs**
  - **Systems**: DDM Rucio, WMS PanDA (components: JEDI, Harvester, Pilot, ...), AGIS (ATLAS Grid Information System)
  - **Central operations** teams: DDM, DPA (Distributed Production and Analysis)
  - **Shifters**: CRC (Computing Run Coordinator), ADCoS (ADC Operations Shifts), DAST (Distributed Analysis Support Team)
  - **Automation**: HammerCloud (full-chain experiment test jobs, automatic exclusion & recovery of compute resources on job failures), Switcher (automatic exclusion & recovery of compute resources on downtime), DDM blacklisting (space area full, storage on downtime), ASAP (ATLAS Site Availability and Performance); AGIS controller, ...

Communication

- **Summary**: emails, GGUS, JIRA, elog, chat, regular meetings
  - **Sites**: GGUS, emails, email notifications
  - **Central teams**: emails, JIRA, GGUS
  - **Shifters**: emails, elog, chat, GGUS, JIRA
  - **Services**: emails, JIRA

Tasks & responsibilities 1 - DDM Ops

- **DDM Operations.**
- **Activity**: data transfers, assuring & balancing data replication policies, deletion, dark data, communication with clients, ...
- **Communication**: emails; GGUS, JIRA tickets; daily&weekly meetings
- **Wish list**: "how can we save ops people time?"
  - bring back analytics results to operations — analytics themselves don't help much
  - automate triage & report of issues with storage and transfers (site, network, FTS)? monitor RSE usage (dark data cleanup, hot spot rebalancing), why is my transfer taking so long? why is my data not available? auto-recovery/replication of files?
  - extra challenge for natural language processing: automated replies to emails asking about standard problems on DAST ;-)
- **Existing work**: anomaly detection (also in DeepLearning variant), simulation (also in hybrid variant), auto labelling of failures, data popularity, transfer time estimation

Tasks & responsibilities 2 - DPA

- **“Operations of distributed data processing”**
- **Activity**: follow up issues with compute, coordinate fixes with teams across the board: PanDA, Rucio, AMI, production managers, shifters, ...
- **Communication**: emails, elog; GGUS, JIRA tickets; daily&weekly meetings
- **Wish list**: automation
  - long tails of tasks, follow up the “last jobs” to finish? automate issue triage & report? daily report of issues (for the morning meeting, where to focus)? estimate when my task will finish? planning and prioritizing campaigns?

Tasks & responsibilities 3 - ADCoS

- **Production shifters**. ADCoS TWiki (copy https://cernbox.cern.ch/index.php/s/4obdn3G1mpfpiU1)
- **Activity**: spot, triage, chase, escalate, follow up issues
- **Communication**: emails, chat, elog; GGUS, JIRA tickets
  - **GGUS**: bug report site issues
  - **JIRA**: bug report SW issues (new type of failure of production tasks), stalled tasks, data transfer/deletion issues (which are not due to site issues)
  - **elog**: manually record every ticket, important updates for next shifters, shift summary report
  - **emails**: intra- and cross-team communication. automatic notifications.
- **Wish list**: automation
  - identify the source of a problem. correlation between various sources.
  - automate the triaging activity? ticket creation? issue resolution? predict issue from service behavior?
    - *how???* extract error information from job/transfers logs? observe & record steps to create a ticket? analyze service logs?

Tasks & responsibilities 4 - DAST

- **Analysis support shifts.**
- **Activity**: 1st line user support: support physicists in their distributed analysis on grid, follow up users (accidentally) abusing the system to offer a more sensitive approach, collaboration with teams/services
- **Communication**: emails; GGUS, JIRA tickets; weekly meetings
- **Wish list**: automation
  - frequently asked questions ⇒ opportunity for a classifier system? “What does this error in the monitor mean ?”, “What does this job/task state mean ?” “Why did a job fail even it works locally ?”
  - checklists? suggestions from error to solution? why my job does not run/fails? what does this error mean? my code was working last week but not now? what do you mean I filled all the SCRATCHDISKS with my 0.5PB data request?
  - organize tutorials & trainings in collaboration from other shifters teams, and WMS/DDM experts

Tasks & responsibilities 5 - CRC

- **Computing Run Coordination**
- **CRC Shifts:** at CERN, 1 week long
- **Activity**:
  - get a good **overview of the whole system** gathering information from monitoring, expert and shifter teams; find not obvious issues and not optimal resource use; coordinate and spread information between all the teams and outside
- **Communication**: emails; GGUS, JIRA tickets; daily and weekly meetings, chatroom
- **Wish list**: automation
  - All requested automation from the previous team will help
  - Statistics of failures by category on the medium term : for instance what are the most frequent failures for transfer, so that we can find and solve the problem at the source => tickets analytics ?
  - Find non used / not enough used resources : our monitoring/organisation is made to find issues/problems at site, a site not running has no failed transfers/jobs => alarm on less than usual used resources (not in downtime) ?

Tasks & responsibilities 6 - Automation

- **Automatic exclusion and recovery of resources**
- **Activity**: testing (HC) & auto-exclusion/recovery of compute (HC on job failures, Switcher on downtimes), auto-exclusion/recovery of storage (full; downtime), “ranking” usefulness of resources (ASAP)
- **Communication**: emails, email notifications
- **Wish list**: more automation
  - HC: get rid of flip flops in blacklisting, ML anomaly detection?
  - ???

ATLAS Computing Operations

A first, partial view

A proposal:

Let's start with DDM-related stuff:

a window of 4 hrs of failures at src/dest,
create a digest email to shifters
with suggestions what to do.

What do you think?

Jaroslava Schovancová, Alessandro Di Girolamo, Mario Lassnig, Michal Svatos, Armen Vartapetian, Mayuko Maeno, Farida Fassi, Sabine Crepe, Helmut Wolters


CERN