EOS report

WLCG meeting 20180913

Massimo Lamanna

Instabilities

- Unusual level of service interruption earlier this year
  - Particularly visible on EOSATLAS and EOSPUBLIC
- Amplified by additional problems of rebooting the instance
  - Reboot of the MGM (Catalogue)
  - Reads back ~ 1kB * #entries (typical "boot" time ~ 1000 s)
  - Scheduled reboot have a similar impact
    - Related activity (in background): MGM memory compaction (to avoid the boot time to be artificially longer)
- Detailed analysis (EOSATLAS) available
  - Corrective actions in production on all instances
  - **https://hackmd.web.cern.ch/xOCz_B3WRYq5yqBW38ehtQ#**

Why? Why now?

- Why instabilities pop up suddenly?
  - No clear correlations with new bugs
    - New undetected issues in a new release
    - Very rare these days
  - Activation of sleeping bugs
    - Workflow changes (different activities)
    - Pressure changes ("rogue" users)
  - Follow-up procedure
    - Get info (stack traces, logs, ...)
    - Development team involved
    - If fix not available, workarounds/mitigation
      - As a matter of fact a new version is generated overnight, goes in the testing pipeline and is deployed the day after
    - Within several days the other instances follow to protect all other users from known issues

Visibility of the problem

- Impact
  - Instance size
    - e.g. rogue user impacts all (but only) colleagues (LHC instances)
    - large instance (n. of files) --> long restart
- Present situation
  - Isolation: one instance for LHC experiment
    - One "catch-all" for NA62, AMS, COMPASS, ...
    - One instance for EOSUSER (CERNBox)
- Transition phase (new deployment)
  - EOSUSER --> 600M entries distributed across 5 instances (impact/5)
  - New MGM with HA backend
    - reboot time from ~1000s to ~10s (duration/100)
    - RocksDB + RAFT consensus algorithm
- By default: no change applied to running instances before LS2

More changes

- Present system:
  - Resilience improved by 2 MGMs (one master, one slave)
  - The slave helps only for read
- New MGM:
  - In production on EOSHOME (EOSUSER successor)
  - In EOSBACKUP up to 1.3B entries
  - Tested up to 4.4B entries (max size today is 600 M entries)
    - The $\sum \quad LHC$ is ~800 M entries
- Next steps
  - Multiple MGMs
    - The status is given by the disk-resident catalogue
    - Fully benefit from the disk-resident catalogue high-availability

Conclusions

- We encouter instabilities areas
  - Presently, moving out of an instability area is amplified
    - Large instances
    - Long reboot time
    - Known features of the system
- We are aware of this
  - Continue to tackle issues avoiding them to hit us multiple times, if possible
  - Improve quality assurance
  - Keep instances up-to-date to avoid resurgence of issues fixed in other instances
  - New deployment
    - Reduce cross-section for problems
    - Reduce their duration
    - Reduce their visibility