Database storage incident

Description

After a network cable replacement on the database storage, part of the Database on Demand instances lost access to the specific NFS filer serving instance volumes where the cable was replaced. Manual re-mounting of the affected volumes was required to bring back the DBoD instances. At 17h one of database storage High-availability pairs experienced a kernel panic. Over 900 NFS storage volumes were unavailable for about 2:45h. This affected databases in the Oracle Database Service and the Database on Demand service. Most services could be restarted in a few hours but some have required more complex interventions. The Oracle database for CMS offline (CMSR) was back online the day after.

Impact

- It has affected both DBoD and Oracle databases (among the most visible, it has affected accelerator databases, experiment databases, access control, EDMS, CASTOR/CTA, Indico, Drupal, 2 Factor Authentication, the new SSO, OpenStack management databases, etc. - many OTGs were created by the relevant services to highlight that their service was affected...
- The list of affected services is in the relevant OTGs:
  - https://cern.service-now.com/service-portal/?id=outage&n=OTG0056746
  - https://cern.service-now.com/service-portal/?id=outage&n=OTG0056791
  - https://cern.service-now.com/service-portal/?id=outage&n=OTG0056733

Time line of the incident

- 26-May-20 18:25 - After observing an inactive port in aggregated (LAG/LACP) interface on storage filer a check and replace request for cable was raised with IT-CS INC:2427315
- 27-May-20 09:30 - Port on switch was tested (with network tester) and came back OK. Therefore network cable was changed. After the intervention, the port remains down.
- 27-May-20 11:00 - Several DBoD instances connected to the NFS filer involved in the cable replacement, started to have issues to access their NFS storage volumes.
  - Attempts to access volumes resulted in `nfs timeouts` which looked at the time as network related or linked with accesing an specific filer IP address. The team tried to work around the situation by trying to stop the instance, re-mount the volumes and re-start the databases.
  - In several cases instances wouldn't shut down cleanly and left behind processes in unkillable 'D' state which were holding required unique resources like listening ports or socket files. In order to clean up those affected servers we attempted to restart the hosts, with the support of IT-CM (for both VMs and Ironic nodes).
  - This procedure worked partially, as it was succesful for some instances but not completely in all the cases.
- 27-May-20 17:11 - In order to solve the connectivity issues towards NFS filer/node, data interface in the problematic filer is migrated to its partner in the cluster (NetApp cluster consist in pairs of nodes connected together).
  - Immediately after interface migration the target node crashed with kernel panic
  - During automated failover of panicked storage node to partner, partner node also crashed with the same kernel panic.
  - From this moment, access to the data hosted in the disks attached to this pair of nodes, is interrupted. This lead to a complete stop/hang of some of the DBoD and Oracle databases. The HA pair does not recover in a clean status after panic reboot and node root metadata volume recovery is required.
- 27-May-20 17:35 - Priority 1 case opened with NetApp Support. NetApp case https://mysupport.netapp.com/site/cases/mine/2008329056

- 27-May-20 18:25 - on the phone with NetApp priority 1 deep technical support (NetApp senior engineer in the US involved).
- 27-May-20 18:42 - Indico database moved to its standby on the 2nd network hub.
- 27-May-20 19:45 - NetApp HA pair recovered and restarted. Storage back online and stable.
- 27-May-20 19:50 - Manual re-start of DBoD instances and Oracle databases started.
- 27-May-20 22:30 - All Oracle databases were available except CMSR database. Loss of transactions on CASTORNS
- 28-May-20 07:30 - All DBoD instances were back online.
- 28-May-20 16:30 - CMSR database back online, some transactions could be lost.
- 02-June-20 12:30 - Replication from the ATLAS offline database (ATLR) to the ATLAS T1s was fixed and restarted. All changes were applied in the next hours.

Analysis

- In order to understand the NetApp issues, the "core files have been already uploaded to the NetApp support site.
- Observed kernel panic is related to NFSv4.1 race condition for open/close file operation in high concurrency environment (our DBoD infrastructure)
- There is a fix in a higher version of ONTAP, but there is a catch. The bug is likely to be triggered *during* interface failover from one storage node to another while serving NFSv4.1 data. Therefore, there is a high chance we would hit the same problem during upgrade where a controller interface (and storage) failover is needed.
    - NetApp is recommending a downtime maintenance window for the upgrade OR remount all NFSv4.1 volumes to NFSv3 or NFSv4.0 and disabling NFSv4.1 protocol before proceeding to assure non-disruptive upgrade.
- At this moment, we are in an "at risk situation" because in case of a storage failover in RAC52 for a any reason, we might hit this bug again.

- Some Oracle databases did not come back in a clean state. A number of checks and manual operations have been required.
    - ORA-600 internal error code, arguiments: [KCRATR_SCAN_LASTBWR] observed on SCADAR and ATLARC, required database recover before being able to open them.
    - Control files corruption was observed on ATLR. One of the copies was in good state and could be used to restart the database.
    - Lost write (essentially inconsistency at the database level between the redo log stream and the data files) on the redo log file prevented CASTOR Name server database re-start. It also required database recover in order to open the database.
    - Block corruption observed on CMSR database. Several attempts to restore and recover datafiles from backup which took many hours (quite large database ~23TB) . Once the database was open, the standby database detected a lost write on primary database and it was decided to failover to the standby (with few minutes of lost transactions) as we were not able to assess the integrity of the primary database.
    - One of the database archived logs on the ATLAS ATLR database, used by the database LogMiner process to "mine" past transactions was corrupt. The log was restored from the standby database and replication resumed.

Follow up

- We should try to understand what made DBoD affected in the morning/early afternoon and not Oracle databases (impact of NFS version, etc.).
    - Comments on DBOD vs Oracle setup differences [Ignacio]:
        - DBOD mounts using NFS v4.1
        - DBOD hosts reach the NetApp clusters via different network paths (not direct) as

generally host are VMs or physical machines which can cross-mount (i.e. RAC54 host, using RAC52 volumes).
    - ◊ There have been DBOD instances affected by NFSv4 related problems in the past couple of months experiencing temporary interruptions of service https://its.cern.ch/jira/browse/DBOD-2195.
    - ◊ Oracle uses it's own nfs implementation compatible with NFS v3, v4.0 and v4.1 (direcnfs) which can react to either bugs or network problems in a different way from the CC7 NFS stack.
  - ♦ Comments from Miro:
    - ◊ We are experiencing 2 separate issues
      1. NFS lock reclaim problem for NFSv4.1 mounts appearing on RAC52 *and* RAC54 that hangs the mounts when occurring.
      2. Race condition with kernel panic when there are many OPEN/CLOSE requests (like for example when force remounting many moutns and/or rebooting many cliens with many mounts - as is the case when we experience #1). More details in bug description attached here: https://cern.service-now.com/service-portal?id=outage&n=OTG0056746
    - ◊ RAC54 runs version that has this kernel panic patched but it also has much lower number of NFSv4.1 mounts
    - ◊ We have to treat #1 and #2 separately with priority on #2 since it's disrupting to Oracle DBs as well.
- Intervention to upgrade ONTAP software needs to be carefully planned in order to be able to do it safely
- NetApp senior engineer should also analyse the lost write issues we have faced to better understand the risk for our databases. Update 4th June: Involved Escalation Engineers and WAFL engineers
- Oracle standby databases of primaries version 12.2 affected by the NFS issue and higher will be checked for lost writes
- Problem #1 was also amplified by the fact that Oracle dNFS file access doesn't work well with NetApp DNS load-balancing because of the Oracle cluster aggressiveness of I/O retry where the load-balanced alias can change filer that is responding to I/O leaving multiple parallel I/O requests "in-flight" that might cause corruption in a certain cases.

Networking research

- We observe network errors towards the IP service of RAC52. See attached *network_errors.png* file. Those errors start around 9.30 in the morning of the day of the incident. There is not exact timing match with the cable replacement [1].
- CS team has reported an incident causing traffic lost in the Barn, potentially due to a bug in the Juniper routers [2]
  - ♦ Despite it was reported to start on Saturday the 2nd of June, IT-CS confirms [1] that the routers affected are the same that give service to RAC52. Waiting for confirmation that this could be the same issue that caused the connectivity issues on DBOD on Wednesday morning
  - ♦ In that case, still not understood why the connectivity issues were observed only with connections to the specific filer where the network cable was replaced
- Still not understood why the network port where the cable was replaced still down
  - ♦ NetApp support doesn t observe hardware issues and IT-CS confirms [1] that before changing the cable, the port on the switch was validated
  - ♦ Action that could be performed is to reallocate that cable in another port. That would require an update of the configuration of the interface aggregation (LAG/LACP) in the side of the stack of switches
  - ♦ **UPDATE**: Asked NetApp support for a deeper analysis on that port and it is identified a hardware issue in the network card. Replacement is shipped. Most probable, the DBOD clients having problems on Wednesday morning were clients whose ips got mapped, router level, to the mac address of the defective network card. Asked NetApp clarification about that

point (NetApp case 2008352219). Quote from reply:

The fact that the card observed ECC errors is indication that the card is no longer functioning as it should and we cannot predict how the card will react to actions such as cable reseat / replace in that state. That also implies we cannot really figure out the root cause of the behavior, if the card is already - by definition - not working as intended.

- Some ports flapping are observed in the RAC52 filers (prior the incident)
  - After further debugging with IT-CS, it is discovered that the ports flapping are all connected to UNIT 4 (4th switch out of 4 giving uplink connectivity to RAC52 installation)
  - IT-CS recommends to replace the unit
  - **UPDATE**: NetApp software update also requires the upgrade of the stack of switches for link aggregation compatibility
- RedHat support has been contacted (prior the incident - not directly related) [3] regarding the recurrent issues observed dring the last months in DBOD
  - Clients mounting data volumes via NFSv4.1 get blocked and report nfs_lock_reclaim related errors
  - RH analysis observes in the logs provided TCP re-transmissions, and point to network instabilities or firewall as root cause
  - DBOD servers have added iptables rules to ensure that both outgoing and incoming new connections to the storage filers, are accepted (was not the case for incoming connections)
  - (NFSv4 is a stateful, session oriented protocol. In case of network interruption or network interface failover, the storage server --entity exporting the NFS volume-- might request new connections to the client --server mounting the NFS export--
  - **UPDATE**: further follow up with NetApp support shows that the minimum Linux Kernel supported for NFSv4.1 is kernel-3.10.0-1127.el7.x86_64 and previous versions are not considered bug-free for NFSv4.1. DBOD instances that showed the issue were running Linux Kernel version 3.10.0-1062.4.3.el7.x86_64

Final Actions

- **Several urgent actions were identified with the help of IT-CS (network configuration and faulty parts/hw repair) and NetApp (storage vendor, configuration and software upgrades). Therefore, a major intervention was organised on 27th June 2020 to fix those with an impact on the Oracle database and DBoD services running on RAC52.**
  - **More details on the intervention: https://cern.service-now.com/service-portal?id=outage&n=OTG0057263**
  - **Impact on the Oracle database services details: https://cern.service-now.com/service-portal?id=outage&n=OTG0057201**
  - **Impact on the DBoD service details: https://cern.service-now.com/service-portal?id=outage&n=OTG0057252**
- This complex intervention has fixed all issues described in this document.

- [1] https://cern.service-now.com/service-portal?id=ticket&table=incident&n=INC2427315
- [2] https://cern.service-now.com/service-portal?id=outage&n=OTG0056846
- [3] https://access.redhat.com/support/cases/#/case/02658941

-- EvaDafonte - 2020-05-28

This topic: DB > PostMortem27May2020
Topic revision: r31 - 2020-09-25 - EvaDafonte