Table of Contents

STREAMS post mortem

Network hardware problem affecting availability of several production databases and Oracle replication (Atlas: online->offline and Tier0->Tier1, LHCb: Tier0->Tier1) on Monday 1st June

Description of the problem

The XFP in the router port which connects the public switch of RAC5 to the General Purpose Network failed on Monday 1st June around 8:20 am CEST causing unavailability of the following production services:

- ATLR (ATLAS offline database),
- COMPR (Compass DB),
- ATLDSC and LHCBDSC (downstream capture databases of Atlas and LHBb).

The hardware problem was fixed by CS around 10:00 a.m. and all the databases became accessible again around 10:15. The problem did not lead to data loss.

CS actions (provided by Andreas Hirstius):

- The logs on the router show that the port suddenly became faulty:
    - changed state to "down" at 8:19am and 5 seconds later to "up" again
    - "down" again at 8:22am and "up" again at 8:27am
    - finally "down" at 8:27am
- Operator investigated until ~8:45am, incl. reboot of switch (to no effect), then called Firstline
- Firstline started processing on site at 9:30am
    - the port received "light" from the switch at nominal power levels
    - Firstline cleaned the fibers on the router and the switch side to no effect
    - XFP of the router port was exchanged ~10:03am
- switch and the machines behind were reachable again at 10:04am
- Firstline closed ticket at 10:10am

Original firstline report: "pas de link sur l'uplink fibre. nettoyage des connecteurs optiques : inactif; sh log sur routeur B513-C-RFTE6-1 int 0/7 : en receive, tx perte signal sur la fibre OK (<-3) --> remplacement du module XFP cote' routeur : link et tests OK Appel le 01/06 a 8h45, fin d'intervention a 10h10"

Oracle Streams propagation jobs from the affected databases (except from ATONR) were automatically disabled after 16 attempts to connect to the destination database (around 9:46 a.m.) and needed to be restarted manually. Replication to Tier1 sites was completely reestablished around 12:00.

Some connectivity anomalies might also have affected the Atlas on-line database (ATONR) which is primarily accessed over the Atlas experiment network and is connected to GPN for monitoring purposes mainly. Due to the aforementioned XFP failure one of Oracle listener processes went down and could not be re-started with the standard procedure. While resolving this issue 2 out of 3 nodes of the cluster rebooted unexpectedly (around 12:25) causing all database activity to fail including online to offline replication. The database was restarted successfully few minutes later (at ~12:30).

The failure of the XFP had so serious impact on services because at the moment it is not possible to implement redundancy at the public switch level:

- using 2 or more IP services would not help because Oracle 10g is not able to use interchangeably 2 or more IP addresses belonging to different subnets
- according to CS experts configuring 2 (or more) public switches for redundancy in a way which would be transparent to Oracle is extremely complicated.

It should be said that a hardware failure of the network infrastructure is extremely rare (this is the first occurance of a public switch failure in 5 years DB operations).

Impact

- Databases:
  - ATLAS online (ATONR) - possible connectivity anomalies between 8:24 and 12:25, effectively down for few minutes between 12:25 and 12:30
  - ATLAS offline (ATLR) effectively down between 8:24 and 10:15 aprox.
  - Compass (COMPR) effectively down between 8:24 and 10:15 aprox.
  - ATLAS and LHCb downstream capture databases (ATLDSC and LHCBDSC) effectively down between 8:24 and 10:15 aprox.
- Replication:
  - ATLAS pvss and conditions data ONLINE -> OFFLINE replication showed several delays from 8:24 to 12:30 aprox.
  - ATLAS conditions OFLLINE -> Tier1s replication stopped from 8:24 to 12:00 aprox.
  - LHCB conditions OFFLINE -> Tier1s replication stopped from 8:24 to 12:00 aprox.
  - LFC LHCb OFFLINE -> Tier1s replication stopped from 8:24 to 12:00 aprox.

Problem with ATLAS replication from ATONR to ATLR (12.12.2008)

Description of the problem

On Friday morning 12.12 both capture processes (conditions and pvss data replication) running on the ATONR database aborted with the following error ORA-00600: internal error code, arguments: [krvrdCBmddlSQL1]. This error was caused by an index recreation using the parallel option executed by the ATLAS dbas. This problem is caused by an Oracle bug which introduces inconsitencies in the logminer dictionary.

Impact

- ATLAS pvss data ONLINE -> OFFLINE replication stopped
- ATLAS conditions ONLINE -> OFFLINE replication (online schemas) stopped
- ATLAS conditions OFLLINE -> Tier1s replication: online schemas out of the replication (but offline schemas still replicated)

Actions

Open Service Request 7239707.994

In the meantime, we tried to re-start the capture process, re-create the logminer components, ignore the transaction causing the problem, ... but without any success. Capture processes aborted at the same point.

Workaround proposed by Oracle support: Rebuild the logminer dictionary and re-create the capture processes to start using this new dictionary. This solution is not suitable for our production environments because some transactions are not captured (capture process starts in a higher scn than the last scn captured before the error)

and this means data loss in the destination database/s. Oracle support was not able to propose another valid workaround.

Final solution

It was decided to re-create both Streams setups. On Monday 15.12 conditions replication was reestablished between ATONR and ATLR and on Tuesday 16.2 PVSS replication was recovered also between ATONR and ATLR. It was necessary to import all the data in order to synchronize both databases (8 GB for conditions, 111 GB (transportable tablespaces) + 2GB for PVSS). Conditions replication (online schemas) between ATLR and Tier1 sites was reestablished on Wednesday 14.01.2009 (after the Xmas break).

Problem with ATLAS replication from ATLR to Tier1 databases (18.10.2008-20.10.2008)

Description of the problem

Following some ATLAS stress tests launched on Friday 17.10.2008 at 18:00, on Saturday morning 18.10.2008, around 09:30 a high load at NDGF caused the database to be totally unresponsive. The propagation job connection to NDGF got blocked due to this situation and could not deliver the LCRs. When the LCRs are not consumed by all the destination databases, they are maintained in the capture queue (spilled LCRs). The capture process and all the propagation jobs to the other 9 Tier1 sites were still working fine. The replication was completely stopped on Sunday 19.10.2008 morning, around 07:30 when the large number of messages build up in the capture process's buffered queue kicked up the flow control state causing the capture process to temporarily "pause" capturing any new messages until some messages are removed from the queue.

On Monday morning 20.10.2008, NDGF had to reboot the host in order to fix the problem with the database. When NDGF was back, LCRs started to be consumed (consuming spilled LCRs is slower than consuming LCRs in memory, this explains why the apply at NDGF was running slower than during normal activity). Once the LCRs were consumed and removed from the queue, capture process started capturing changes again and all the Tier1 sites recover from the accumulated backlog during the afternoon.

As all components of streams system were in a healthy state during that time, the streams monitoring did not reported any problem.

Actions

We've introduced additional latency monitoring into STRMMON to spot situations when one Tier1 that is having difficulties and affects all the streaming.

We are currently investigating the possibility of increasing the memory dedicated to the capture queue. This would allow the capture process to capture LCRs during a longer period of time (and possibly cover the weekends) when one destination site is not consuming the LCRs and they are being spilled.

Problem with LFC for LHCB replication to Gridka (31-7 till 6-8)

Description of the problem

Replication of LFC for LHCB from CERN to Gridka stopped shortly after the upgrade to 10.2.0.4. Propagation from CERN to Gridka would end with an error (TNS error of connection dropped). At the same time Gridka DBAs reported several crashes of CRS on cluster node 1.

Actions

Several attempt to re instantiate the replication to gridka cluster node 1 failed. Gridka DBAs applied a maintenance to node N.1 which stabilized the cluster. Still transfer of files to node N.1 was not possible. Streams reinstantiation was performed using cluster node N.2.

Conclusions and further actions

Streams replication has been re activated since 6-8 using Gridka cluster node N.2 as destination. Further investigations are being performed by Gridka DBAs to see weather the network/communication issue with node N.1 is still present.

As the problem has reappeared again, we have opened a SR with Oracle. They have provided a diagnostic patch which was installed at Gridka. We have to wait until the problem reproduces again.

Problem spotted on 29th of July 2008 (evening)

Description of the problem

Initial problem

Due to a known bug in Oracle streams when dropping tables with referential constraints, one needs to drop the constraints before dropping the object. Otherwise it will cause instantiation of other objects to be dropped, thus will cause all apply processes to abort with an error. This error is encounter every now and then when conditions DB user does not notify experiment DBA before dropping tables. The problem can be easily fixed within an hour.

Main problem discovered

After fixing the initial problem it turned out that propagation of ATLAS conditions was not working for some time. From the monitoring perspective it was reporting as waiting for new changes to be captured and then propagated to Tier1s, with a reported latency of just around 1h (more or less normal state). More investigation showed that archivelogs of 26th of July where missing on downstream capture machine. There were no information either in logs on source (ATLR) or downstream capture box.

Details

- ATLAS conditions capture stuck at archivelog from 26th July 2-3 a.m. (3,5 days accumulated lag to most of the sites until problem realized 29th of July evening)
- Problem fixed around 15:15 on 30th of July (missing logfiles identified, copied and registered, capture & propagation successfully restarted)
- new logfiles found missing on the 31st of July (due to ATLDSC restarts, when fixing the streams setup) - missing logfiles identified, copied and registered, capture & propagation successfully restarted

Impact

- ATLAS conditions ONLINE -> OFFLINE replication interrupted for a short time (1h break, the usual thing when hitting the Oracle bug when drop tables)
- ATLAS conditions OFLLINE -> Tier1s replication was not working properly (for most of the sites except SARA) since 26th of July, till 30th of July afternoon
    - additional problems on the 31st of July, short outage of few hours due to missing logfile

Root cause

Initial problem

The constraints were not dropped beforehand on Tuesday evening by ATLAS DBAs (documented procedure that each person dropping objects from ATLAS conditions DB must report it to ATLAS DBAs first), which caused streams from ATLAS online DB to ATLAS offline to abort with an error. Around 19:30, after some investigation, the propagation was back and working ok.

Main problem discovered

Main problem root cause identified to missing logfiles in ATLAS downstream capture. The capture was stuck due to unknown problem, what caused old log files to be removed. After capture has been restarted it started mining logs from far behind and required them to be manually shipped from ATLR.

Actions

- problem identified and fixed
- notifications to WLCG meetings sent on 30th of July and 1st of August
- improvements in the monitoring will be needed to spot similar problems (assigned to development)

This topic: PSSGroup > StreamsPostMortem
Topic revision: r11 - 2009-06-02 - JacekWojcieszuk