CASTOR Data Loss Incident

Gordon D. Brown

Rutherford Appleton Laboratory

3D Workshop
CERN, Geneva
26th/27th November 2009

Outline

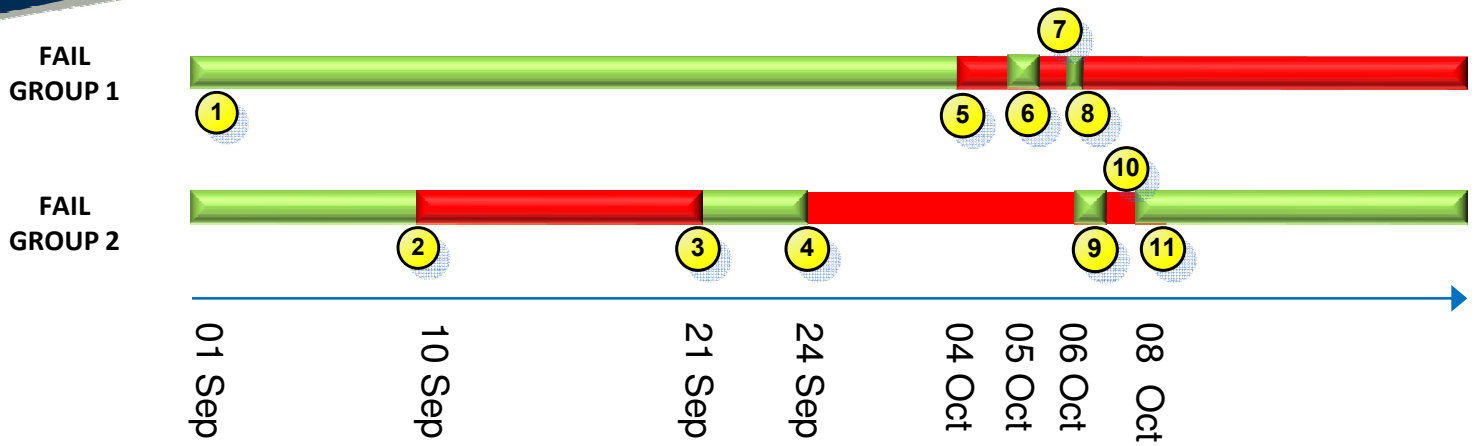
- Setup before data loss
- Data loss incident
- Fallout
- Architecture review
- Procedures
- Database resilience
- Plans

Setup Before Data Loss

- Before May 2009:
  - 2 x 5 node RAC
  - Single copy of data on ASM (external redundancy)

- May 2009:
  - 2 x 5 node RAC
  - Moved to high-performance EMC Storage
  - Two fail groups in ASM, normal redundancy
  - Move to new machine room, in to UPS room
  - Capacity to add new fail groups without downtime

Data Loss



1. Pluto running using Fail Group 1, all changes mirrored to Fail Group 2
2. EMC crash killing FG2, FG1 has poor performance for 2 hours. FG2 left down until Oracle come back with patch to fix hanging ASM.
3. Patch applied, FG2 bought back up
4. EMC crash killing FG2. FG2 left down due to ongoing EMC errors.
5. EMC crash killing FG1, database down
6. FG1 restarted, attempted database open but node space issues cause problems
7. Database remains down while work with Oracle to resolve space problem
8. Oracle resolve node space issues, database started on FG1. EMC crash killing FG1 after 20 minutes, database down.
9. Database opened again, but using undetected old copy of data on FG2. Database backed up.
10. Database taken down to do restore.
11. Database restored to 6th Oct, using backup from 6th Oct (only includes data to 23rd Sept)

Overview

- UPS Room power problems
- Database architecture unsound
- Backup & Recovery procedures did work
- No checks on database “age” on startup
- CASTOR team spent half day checking for consistency - concluded OK
- Filesync and garbage collector on CASTOR ran
- Disk server files not backed up (as disk0)
- Problem noticed when file IDs reused
- Database was restored – but too late as garbage collector had deleted files
- Now on old hardware, just as LHC starts up

Immediate Actions

- Move database off of EMC hardware
- Old hardware still around
- LFC/FTS moved to other (non-Tier-1) database
- Review and fix architecture
- Test database “start up” procedures
- Look at database resilience
- Quarterly Database Services team day due anyway
- Stop – and get it right!

Fallout

- Most serious RAL Tier-1 incident to date
- Two disaster procedures:
  - Hardware failures
  - Data loss
- Many meetings, reports
- Everybody becoming involved in databases
- Stress

Architecture Review

- Priority
  - Stop database opening using “old” fail group
- Considerations
  - Current hardware
  - Knowledge/experience
  - Not changing too much
  - Timescales – short
  - Advice from Oracle/CERN

Architecture Review

- Looked at:
  - Hardware RAID
  - ASM
    - Two fail groups with additional check
    - Three fail groups
      - Normal redundancy
      - High redundancy
    - How many fail groups?
    - What happens if problems during resync?
    - What is the optimal number of disk arrays/fail groups?
  - Data Guard
  - Logical Standby

Architecture Review

- Stick with existing architecture
  - Proven
  - Hardware available
  - Used by CERN
- Must have method of database not using old fail group
  - Lots and lots of testing
  - Check script with Oracle/CERN

Procedures

- What to do during an investigation
  - "Timeline" document
  - Communication
- Backup/Recovery Procedures
  - Ongoing
  - 18 tests
  - Wouldn't have helped solve this
  - Semi-automated recovery tests every two weeks

DB “Age” Tests

```
alter session set NLS_DATE_FORMAT = 'mm-dd-yyyy HH24:mi:ss'
/

spool restorecheck.lst
SET SERVEROUTPUT ON
column BACKUP_DATE format a30

prompt Check application last transaction

declare
  app_timestamp date;
begin
  for rec in (select OWNER from DBA_TAB_COLUMNS WHERE TABLE_NAME='SUBREQUEST'
   AND COLUMN_NAME='LASTMODIFICATIONTIME')
  loop
    EXECUTE IMMEDIATE 'SELECT
   TO_DATE(''19700101000000'',''YYYYMMDDHH24MISS'')+NUMTODSINTERVAL(MAX(CREATION
   TIME),''SECOND'') FROM '||rec.OWNER||'.SUBREQUEST WHERE
   LASTMODIFICATIONTIME=( SELECT MAX(t.LASTMODIFICATIONTIME ) FROM
   '||rec.OWNER||'.SUBREQUEST t)' into app_timestamp;
    DBMS_OUTPUT.PUT_LINE('Owners '||rec.OWNER||' timestamp
   '||TO_CHAR(app_timestamp,'dd_mm_yyyy hh24:mi:ss'));
  end loop;
end;
/
```

DB “Age” Tests

```
prompt Check of last 10 backup dates

SELECT TO_CHAR(BACKUP_DATES,'DD_MM_YYYY') BACKUP_DATE
  FROM (SELECT DISTINCT TRUNC(COMPLETION_TIME) BACKUP_DATES
          FROM V$BACKUP_SET
         ORDER BY 1 DESC)
 WHERE ROWNUM <11
/

prompt The history of uninterrupted database work

SELECT LAG(TIME_DP,1) OVER (ORDER BY TIME_DP ) START_DATE, TIME_DP STOP_DATE
FROM (
select LAG(TIME_DP,1,TO_DATE('01_01_2000','DD_MM_YYYY')) OVER (ORDER BY TIME_DP
   ) PREV_DP, TIME_DP
, MAX(TIME_DP) OVER(PARTITION BY 1) MAX_DP
from sys.smon_scn_time )
WHERE TIME_DP-PREV_DP>10/(24*60) OR TIME_DP=MAX_DP
ORDER BY 2
/

spool off

exit
```

DB Resileince

- Ongoing
- Offer options to Tier-1
- Considerations
  - Human error
  - Complexity
  - Risk
  - Cost (h/w, licenses)
  - Operation
- Scoring
- Oracle Maximum Availability Architecture

Current Status

- Machine room power problems nearly solved
- Plan to move back to EMC hardware
  - Two weeks of testing
  - Potentially 5th January 2010
  - New architecture almost complete
- Future Resilience Plans
  - Recommendations drafted
- GridPP Review on 14th December 2009

Acknowledgements

- Thanks to the following:
  - Luca Canali
  - Jacek Wojcieszuk
  - Eric Grancher
  - Maria Girone
  - Ruben Domingo Gaspar Aparicio