Science & Technology Facilities Council

CASTOR Issues at RAL

3D Workshop,
Barcelona (ES),
20-21 April 2009

Carmine Cioffi

Database Administrator and Developer

Outline

- Database Overview
- Schemas Size and Versions
- Oracle Installation
- Hardware Specifications
- Problems hit during production
- Test Database
- Key Metrics
- Future plan

Databases Overview

- Two 5 nodes RAC + one single instance in production

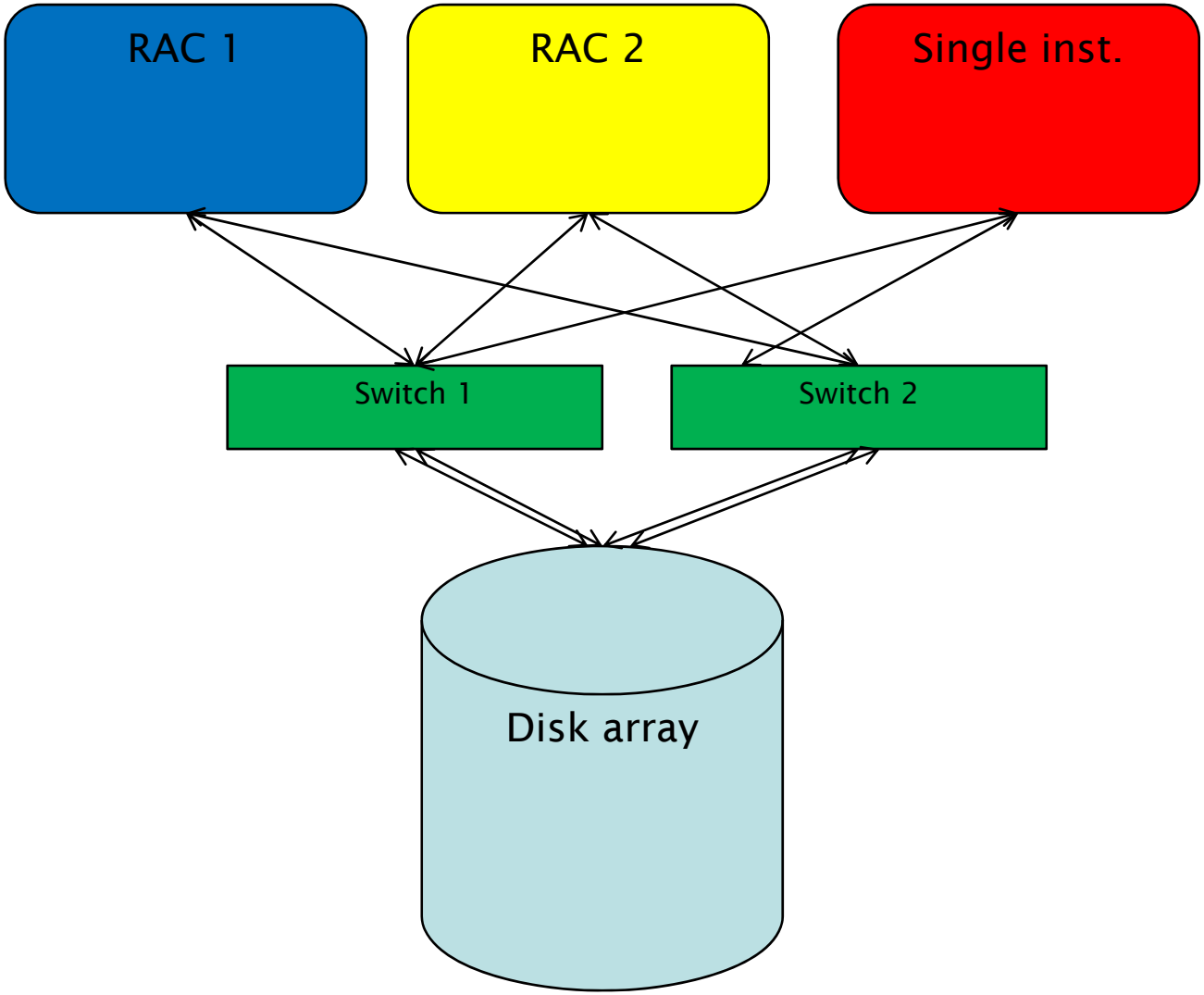
- One two nodes RAC for development and testing

Databases Overview

- Each database has the following schemas:
  - RAC 1: Name Server, VMGR (Volume Manager), CUPV (Castor User Privilege Validator), CMS Stager, Gen Stager, Repack, SRM Alice, SRM CMS, VDQM2
  - RAC 2 : Atlas Stager, LHCb stager, SRM Atlas, SRM LHCb.
  - Single Instance: DLF for all VOs

Science & Technology Facilities Council

Database Overview



Schemas Size and Versions

Science & Technology Facilities Council

RAC 1

Schemas	Version	Size
Name Server	n/a	2.2GB
VMGR	n/a	1.5MB
CUPV	n/a	0.2MB
CMS Stager	2_1_7_24	1.2GB
Gen Stager	2_1_7_24	3GB
Repack	2_1_7_19	5MB
Gen SRM	2_7_15	419MB
SRM CMS	2_7_15	400MB
VDQM2	Coming soon	

RAC 2

Schemas	Version	Size
Atlas Stager	2_1_7_24	16GB
LHCb stager	2_1_7_24	688MB
SRM Atlas	2_7_15	4GB
SRM LHCb	2_7_15	280MB

Oracle Installation

- Version 10.2.0.4
- Last patch CPUJan09
- Non default initialisation parameter:
  - _kks_use_mutex_pin = FALSE
  - sga_target = 2GB
  - pga_aggregate_target = 600MB
  - processes = 800
  - cursor_sharing = EXACT
  - open_cursors = 300

Hardware Specification

- OS: Red Hat Enterprise Linux AS release 4 (Nahant Update 7)
- RAM: 4GB
- CPU: Dual quad Intel(R) Xeon(TM) 3.00GHz
- Storage:
  - RAC 1: 560GB
  - RAC 2: 560GB
  - Development RAC:93GB
  - Single instance: 1.8TB
- Overland 1200 disk array
  - twin controller
  - twin Fibre Channel ports to each controller
  - 10 SAS disk (300GB each 3TB total gross space)
  - Raid 1(1.5 TB net space)
- Two Brocade 200E 4Gbit switches

Problems hit during production

1. Big Id (or better Random ID) when updating the Id2Type table:
   - It does happen randomly and CERN proved Oracle bug no 8439390
2. Cross talk experienced during the synchronization process:
   - was resolved by removing permanently the synchronisation process from CASTOR. The new CASTOR version will allow to disable dynamically the synchronization.
3. Atlas stager tables stats becomes stale because of high throughput:
   - We had this problem for a while but then it did disappear probably due to the Atlas Stager cleanup.

Problems hit during production

- 4 Atlas Stager was too big:
  - The cleanup jobs were disabled and the schema size reached 135GB. It took two weeks to clean it up
- 5 CMS Stager BULKCHECKFSBACKINPRODJOB failing:
  - Due to inconsistency in the schema probably related to the Big Id problem. Shaun has got a solution from CERN and is planning to test it before applying it on the production system
- 6 Backup been slow over NFS:
  - We generate around 300GB per day and the backup process was too slow to remove them so ASM did fill up. The solution was to remove NFS and use locally attached disks

Test Database

- Oracle version 10.2.0.3
- Oracle Software not updated/patched since crosstalk did happen
- It was used to try to recreate the:
  - Big Id problem
  - Cross talk problem
- Schema version and size:

Schemas	Version	Size
SRM ATLAS	1_1_0	96MB
Atlas Stager	2_1_7_15	16MB
LHCb stager	2_1_7_15	16MB

Key Metrics

- 350 tran/sec on Neptune
- 290 tran/sec for Atlas Stager
- 780 physical write/sec (6MB/sec) on Neptune
- 700 physical write/sec (5.5MB/sec) for Atlas Stager
- 930 physical read/sec (7.3MB/sec) on Neptune
- 910 physical read/sec (7.15MB/sec) for Atlas Stager
- Orion results:
  - MBPS=182.88
  - IOPS=1652
  - Latency msec=13.45

Future Plan

- Provide more hardware resilience
- Use a new storage configuration:
  - Two disk arrays
  - Config ASM with normal redundancy
  - Mirror disks over different disk array
  - Mirror/multiplex OCD and VD over the two disk arrays
- We are testing the migration procedure and failover:
  - Data transfer to a new disk group with external redundancy
  - Failure of one of the disk array

Science & Technology Facilities Council

New Hardware Architecture