Science & Technology Facilities Council

CASTOR Databases at RAL

Castor Face to Face,
RAL
18-19 February 2009

Carmine Cioffi

Database Administrator and Developer

Database Services Team

- Gordon Brown (not the prime minister!!!)
- Keir Hawker
- Richard Sinclair
- Carmine Cioffi

Science & Technology Facilities Council

Outline

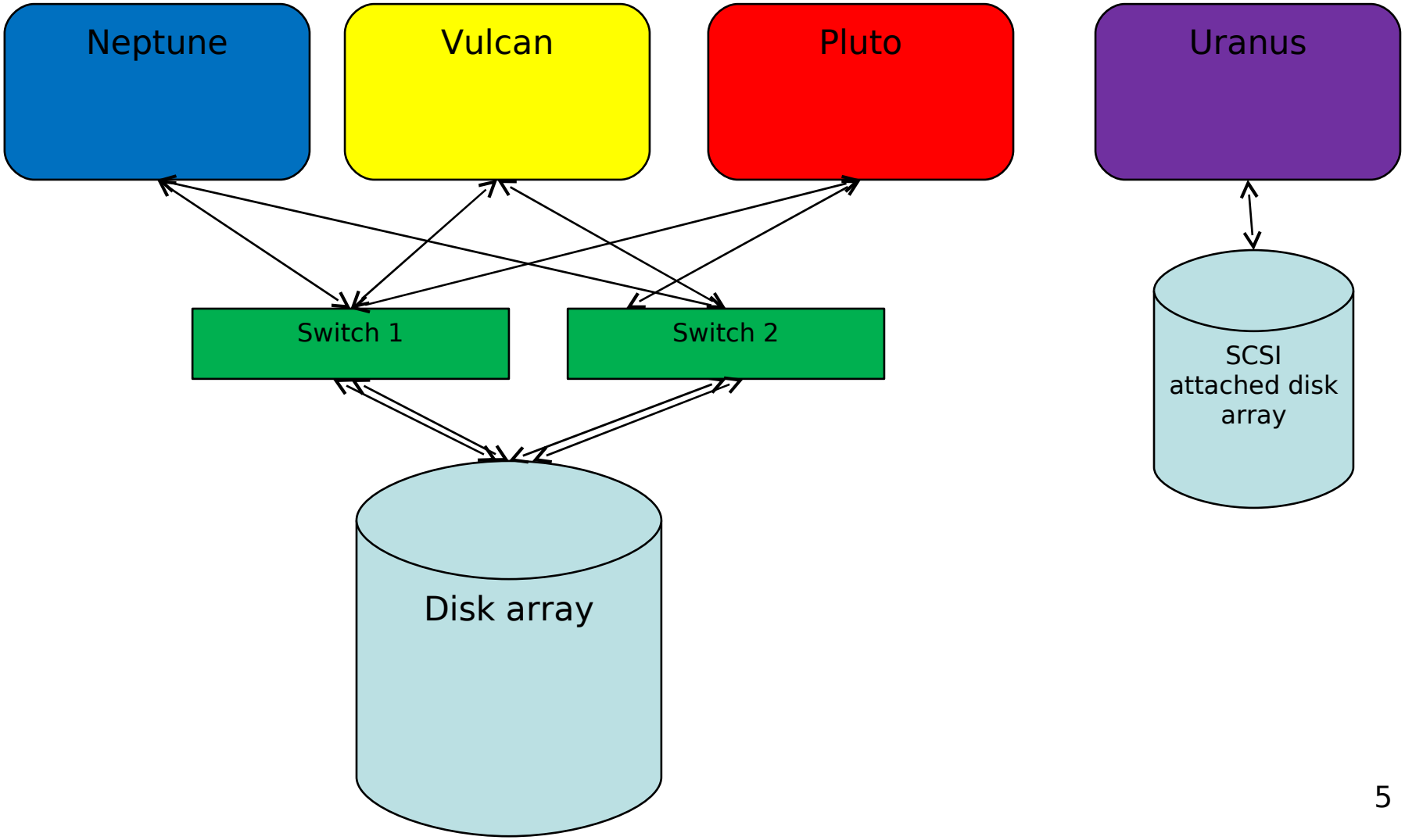
- Database Overview
- Schemas Size and Versions
- Oracle Installation
- Hardware Overview
- Hardware Specification
- Hardware Limitation and Future Plan
- New Hardware Architecture
- How we will move over the new hardware
- Problems hit during production
- Test Database
- Tuning and Monitoring
- Few Metrics
- Knowledge Exchange

Databases Overview

- Two 5 nodes RAC + one single instance in production:
  - Pluto
  - Neptune
  - Uranus
- One two nodes RAC for development and testing:
  - Vulcan

Science & Technology Facilities Council

Database Overview



Databases Overview

- Our databases are named after Roman Gods:
  - Pluto: god of the underworld and the judge of the dead.
  - Neptune: God of sea
  - Uranus: god of the sky/heaven
  - Vulcan: god of fire and manufacturer of arms

Databases Overview

- Each database host the following schemas:
  - Pluto: Name Server, VMGR (Volume Manager), CUPV (Castor User Privilege Validator), CMS Stager, Gen Stager, Repack, SRM Alice, SRM CMS
  - Neptune: Atlas Stager, LHCb stager, SRM Atlas, SRM LHCb.
  - Uranus: DLF for all VOs

Schemas Size and Versions

Science & Technology Facilities Council

Pluto

Schemas	Version	Size
Name Server	n/a	2GB
VMGR	n/a	1.5MB
CUPV	n/a	0.2MB
CMS Stager	2_1_7_19_2	1GB
Gen Stager	2_1_7_19_2	2.5GB
Repack	2_1_7_19	24MB
SRM Alice	1_1_0	556MB
SRM CMS	2_7_12	155MB

Neptune

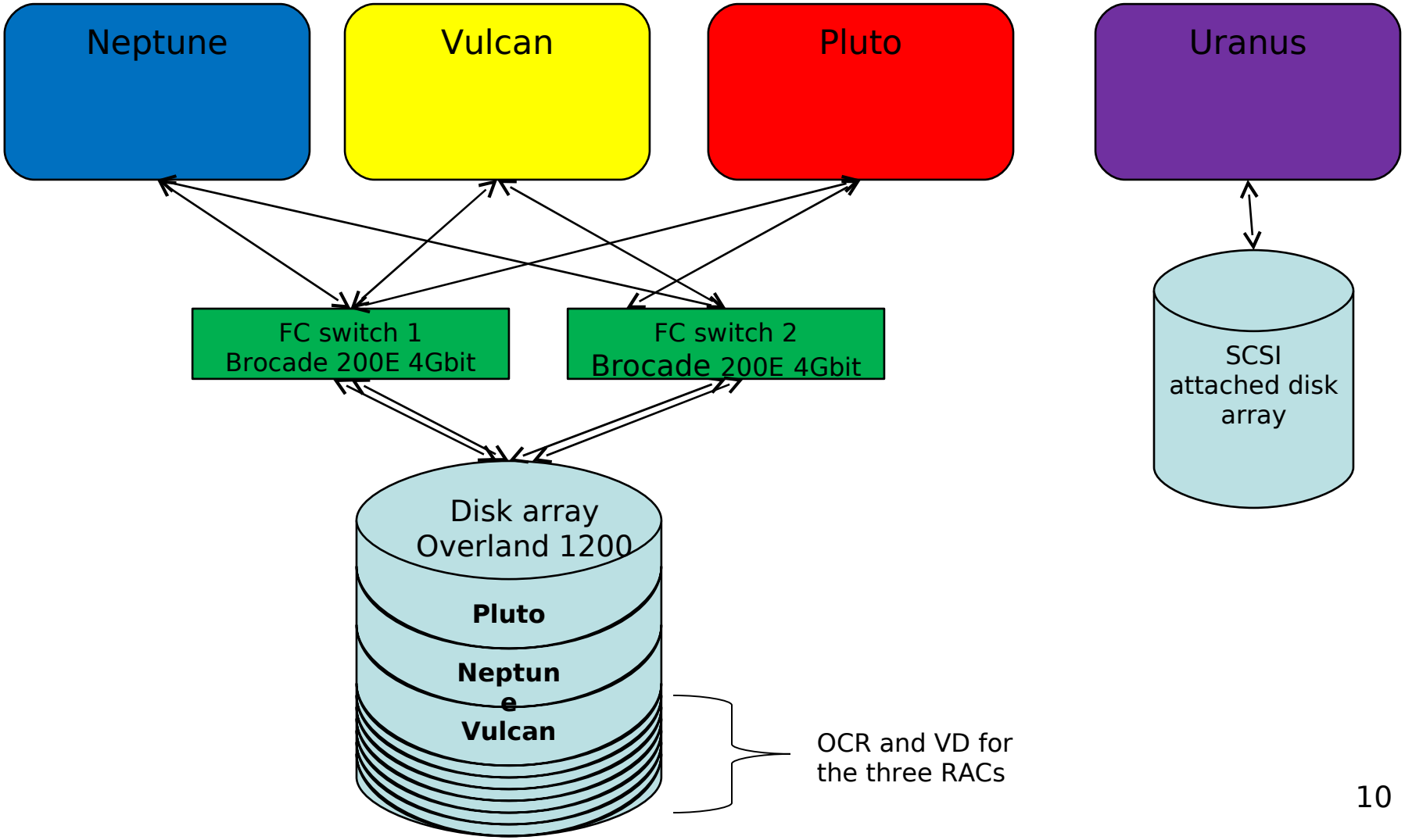
Schemas	Version	Size
Atlas Stager	2_1_7_19_2	134GB
LHCb stager	2_1_7_19_2	600MB
SRM Atlas	2_7_12	3GB
SRM LHCb	2_7_12	76MB

Oracle Installation

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

Science & Technology Facilities Council

Hardware Overview



Science & Technology Facilities Council

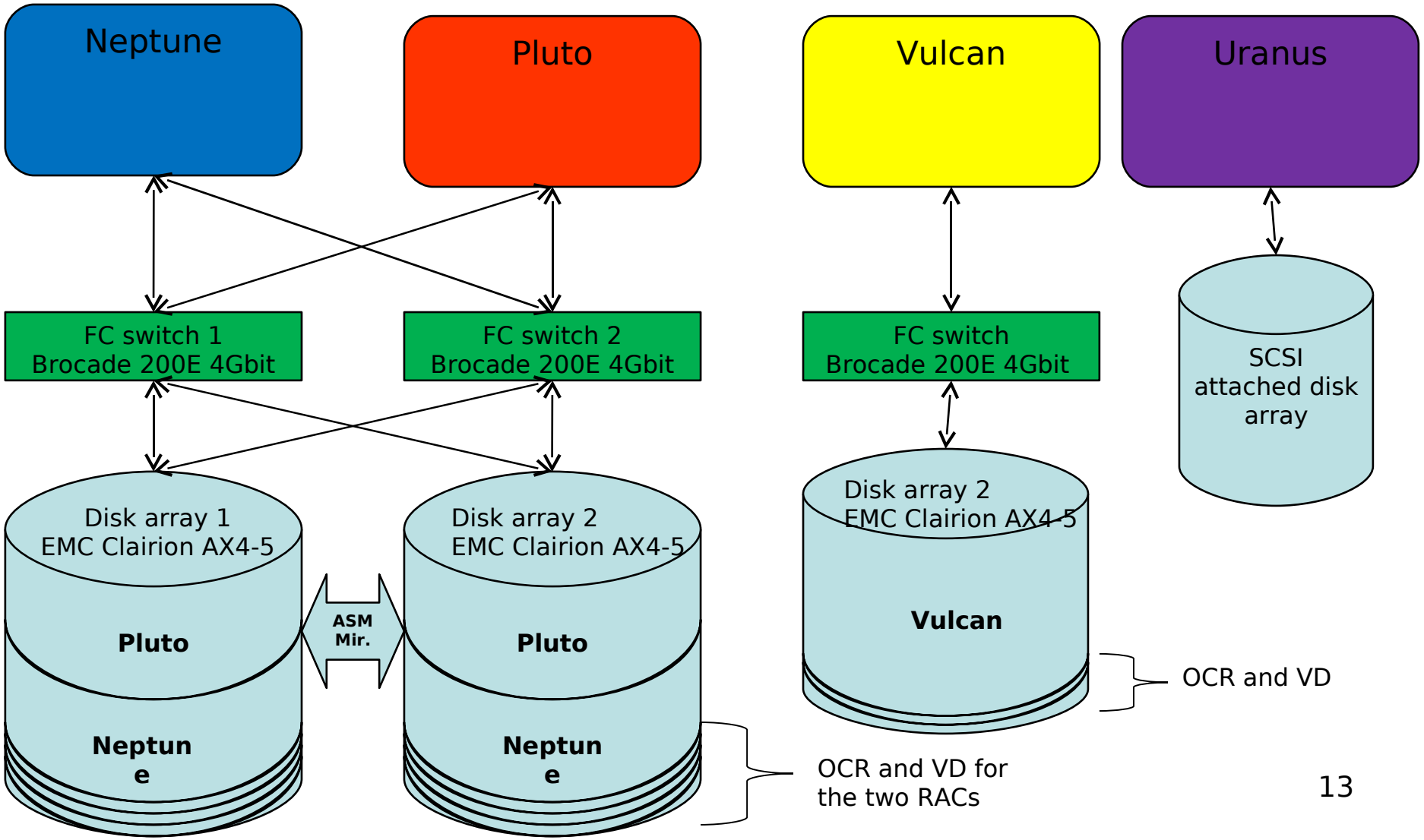
Hardware Specification

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

Hardware Limitation and Future Plan

- A concern is raised by the current hardware :
  - There is not enough storage
  - There is no disk array redundancy: if the disk array breaks down the Castor database goes down
- Because of the above concern we are planning to move over to new hardware:
  - Two new disk arrays with more space
  - One disk array will be used to mirror the

New Hardware Architecture



New Hardware Specifications

- Storage:
  - Pluto: 2TB
  - Neptune: 3TB
- Two Overland 1200 disk array each one with
  - twin controller
  - twin Fibre Channel ports to each controller
  - 20 SAS disks (300GB each 6TB total gross space)
- Two Brocade 200E 4Gbit switched
- Raid 5 (5TB total net space)

How we will move over the new hardware

- The current ASM disk group has redundancy external which cannot be modified therefore we will proceed in this way:
  - Create a new disk group with redundancy normal
  - Shutdown the database and migrate data from the current disk group to the new disk group.
  - Adjust the control file content so that the tablespaces are associate with the data files in the new location
  - Start-up the database

Problems hit during production

1. Big Id when updating the Id2Type table:
   - It does happen randomly and we are trying to recreate it on demand so that Oracle support can investigate it
2. Cross talk experienced during the synchronization process:
   - was resolved by removing the synchronisation process from Castor. We are trying to recreate it over the testing system
3. Atlas stager tables stats becomes stale because of high throughput:
   - When the problem appeared we recollected the statistics. We are changing the script to collect statistics similar to CERN (i.e. separate statistics gathering scripts for internal Oracle tables and Castor tables.

Problems hit during production

1. Running out of space because of the huge amount of redo logs generated (~300GB day)
2. RMAN is very slow to work over NFS mounting points:
   - The backups were done over an NFS mounting point and this caused RMAN to be very slow in removing redo logs. Because Oracle could create redo logs faster than RMAN could remove them many times we got close to fill up the ASM. We solved the problem by doing backup over locally attached disks which sped up RMAN

Science & Technology Facilities Council

Test Database

- Vulcan is used for testing
- Oracle version 10.2.0.3
- Oracle Software not updated/patched
- It is 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

Tuning and Monitoring

- Grid Control
- AWR
- CPU load (top)
- Oracle advisors
- Explain execution plan

Few 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

Knowledge Exchange

- Do we want to have regular Castor database workshops?
- Exchange on regular basis Metrics?
  - I/O sec
  - Tran/sec
  - AWR reports
- I would like to see the system architectures for Tier(0/1)s sites
- I found useful the Castor time slot into the 3D fortnight meeting (even though I'm the only one to report problems ;o)
- Any suggestion?