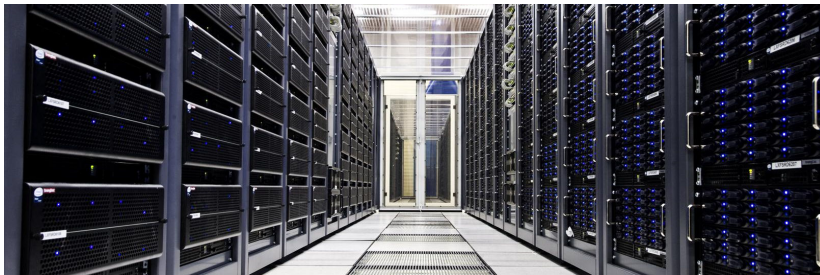


# ASM Configuration Review



Luca Canali, CERN-IT

Distributed Database Operations Workshop  
CERN, November 26<sup>th</sup>, 2009



## Outline

- ASM motivations and CERN's architecture
- ASM technology review
- ASM configuration parameters, instance management
- Configuration of storage arrays for ASM
  - RAID and ASM
  - Oracle clusterware and ASM
  - Normal redundancy or external redundancy?



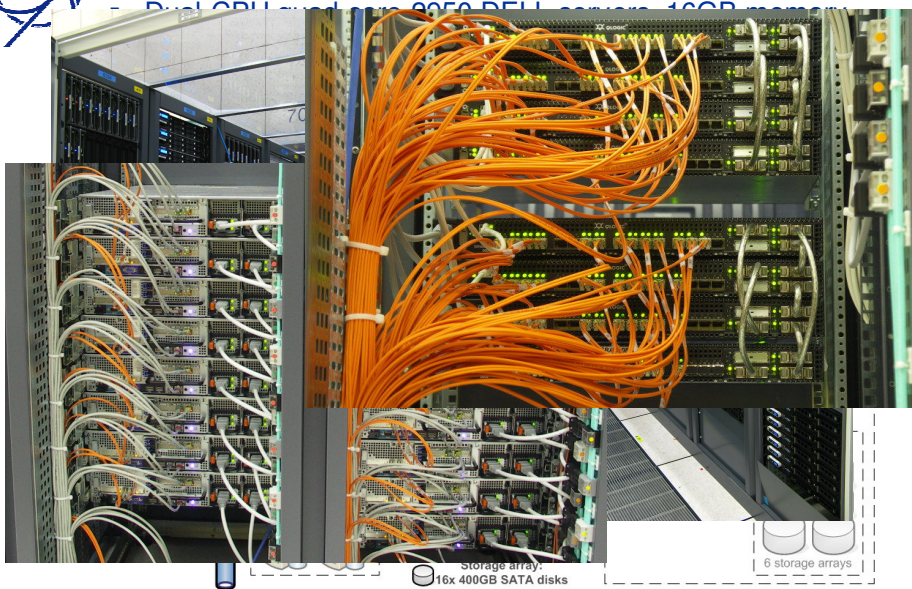
## Architecture and main concepts



- Why ASM ?
  - Provides functionality of volume manager and a cluster file system
  - Raw access to storage for performance
- Why ASM-provided mirroring?
  - Allows to use lower-cost storage arrays
  - Allows to mirror across storage arrays
    - arrays are not single points of failure



## CERN Set-up

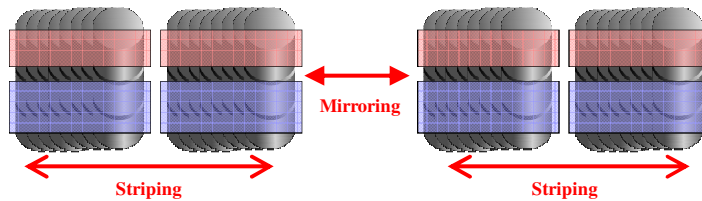




## Oracle Cluster Storage, ASM

- ASM for **mirroring and striping** across arrays and striping
- Allows the use of **commodity HW** (mirror across arrays)
- Disks can be added and removed online for scheduled and unscheduled changes
- Example:

**DATADG** **RECODG** disk groups: data and flash recovery



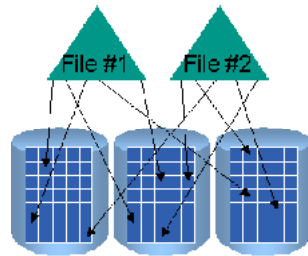
## ASM Storage Internals

- ASM **Disks** are divided in Allocation Units (**AU**)
  - Default size **1 MB** (`_asm_ausize`)
  - Tunable diskgroup attribute in 11g
- ASM **files** are built as a series of **extents**
  - Extents are mapped to AUs using a file extent map
  - When using 'normal redundancy', 2 mirrored extents are allocated, each on a different **failgroup**
  - RDBMS read operations access only the primary extent of a mirrored couple (unless there is an IO error)
  - In 10g the ASM extent size = AU size

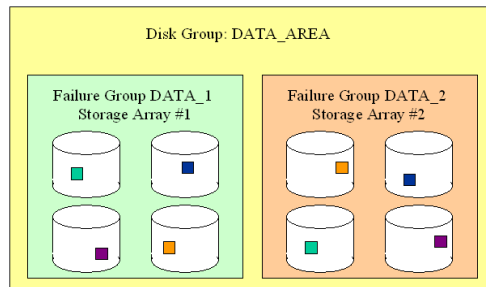


## ASM Files, Extents, and Failure Groups

Files and  
extent  
pointers



Failgroups  
and ASM  
mirroring



## ASM Is Not a Black Box

- ASM is implemented as an Oracle instance
  - Familiar operations for the DBA
  - Configured with SQL commands
  - Info in V\$ views
  - Logs in udump and bdump
  - Some 'secret' details hidden in X\$TABLES and 'underscore' parameters



## Selected V\$ Views and X\$ Tables

View Name	X\$ Table	Description
V\$ASM_DISKGROUP	X\$KFGRP	performs disk discovery and lists diskgroups
V\$ASM_DISK	X\$KFDSK, X\$KFKID	performs disk discovery, lists disks and their usage metrics
V\$ASM_FILE	X\$KFFIL	lists ASM files, including metadata
V\$ASM_ALIAS	X\$KFALS	lists ASM aliases, files and directories
V\$ASM_TEMPLATE	X\$KFTMTA	ASM templates and their properties
V\$ASM_CLIENT	X\$KFNCL	lists DB instances connected to ASM
V\$ASM_OPERATION	X\$KFGMG	lists current rebalancing operations
N.A.	X\$KFKLIB	available libraries, includes asmlib
N.A.	X\$KFDPARTNER	lists disk-to-partner relationships
N.A.	X\$KFFXP	extent map table for all ASM files
N.A.	X\$KFDAT	allocation table for all ASM disks

*ASM Configuration Review, Luca Canali*



## ASM PDB-Utilities

- Isdg
  - Shows diskgroup details
- Isop
  - Shows rebalancing operations
- listdisks.py
  - Shows available disks in ASM, space, status, etc
- New: utility to scan disks to find secondary extent corruption
- Developed in-house, available to share

*ASM Configuration Review, Luca Canali*



## ASM Parameters

- Notable ASM instance parameters:

```
*.asm_diskgroups='TEST1_DATADG1','TEST1_RECODG1'
*.asm_diskstring='/dev/mpath/itstor*p*'
*.asm_power_limit=5
*.shared_pool_size=90M
*.db_cache_size=80M
*.large_pool_size=20M
*.sga_max_size=200M
*.instance_type='asm'
*.processes=100
```



## ASM takes care of mirroring

- ASM with normal redundancy diskgroups
  - Provide for striping and **mirroring**
  - It's similar to RAID10, but there is no disk-to-disk mirroring
  - Oracle can take care directly of the disk access
    - No stripe on stripe, only 1 stripe size (1MB wide on 10gR2 by default)

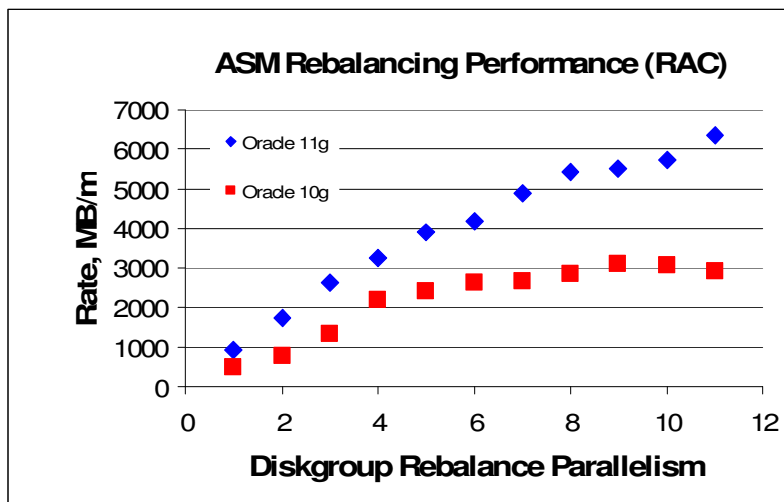


## ASM Rebalancing

- Rebalancing is performed (and mandatory) after space management operations
  - Goal: **balanced** space allocation across disks
  - Not based on performance or utilization
  - ASM spreads every file across all disks in a diskgroup
- ASM instances are in charge of rebalancing
  - Extent pointers changes are communicated to the RDBMS
    - RDBMS' **ASMB** process keeps an open connection to ASM
    - This can be observed by running strace against ASMB
  - In RAC, extra messages are passed between the cluster ASM instances
    - **LMD0** of the ASM instances are very active during rebalance



## Rebalancing, an Example



Data: D.Wojcik,  
CERN IT



## Rebalancing Workload

- When ASM mirroring is used (e.g. with **normal redundancy**)
  - Rebalancing operations can move more data than expected
- Example:
  - 5 TB (allocated): ~100 disks, 200 GB each
  - A disk is replaced (diskgroup rebalance)
    - The total IO workload is 1.6 TB (8x the disk size!)
    - How to see this: query v\$asm\_operation, the column **EST\_WORK** keeps growing during rebalance
- The issue: excessive repartnering



## ASM Disk Partners

- ASM diskgroup with normal redundancy
  - Two copies of each extents are written to different **'failgroups'**
- Two ASM **disks** are **partners**:
  - When they have at least one extent set in common (they are the 2 sides of a mirror for some data)
- Each ASM disk has a limited number of partners
  - Typically 10 disk partners: **X\$KFDPARTNER**
  - Helps to reduce the risk associated with 2 simultaneous disk failures





## Fast Mirror Resync

- ASM 10g with normal redundancy does not allow to offline part of the storage
  - A transient error in a storage array can cause several hours of rebalancing to drop and add disks
  - It is a limiting factor for scheduled maintenances
  - 11g has new feature 'fast mirror resync'
    - Redundant storage can be put offline for maintenance
    - Changes are accumulated in the staleness registry (file#12)
    - Changes are applied when the storage is back online



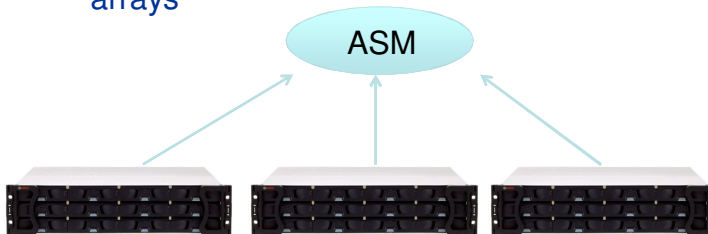
## Setup of the Array

- Should I use normal redundancy or RAID (or both)?
- Let's review some use cases:
- CERN: high performance and high availability clusters.
  - RAC and ASM.
  - Many physical disks for IOPS performance
  - Many TB of storage
  - Low cost
  - ~10 disk arrays per DB cluster
  - ASM diskgroups with normal redundancy
    - Disks configured as JBOD, ASM does mirroring and striping



## Small Config with ASM and CRS redundancy

- Data diskgroup created over 3 storage arrays
  - Each array is an ASM failgroup
  - Disks configured as JBOD
  - Three copies of cluster voting disk distributed over 3 arrays



## Notes on a config with 3 arrays

- Data in a ASM diskgroup with normal redundancy and 3 arrays
  - is simply mirrored (i.e. 2 copies of extents are maintained)
  - Note small difference of ASM mirroring vs. RAID1:
    - Ex: stripe of 1 MB on diskarray 1, mirror copy somewhere on disk array 2 or 3



## ..with 2 arrays

- With 2 arrays it is also possible to profit from ASM mirroring across arrays.
  - Although in case of failure redundancy cannot be restored
  - Rebalance will only be possible when HW is restored



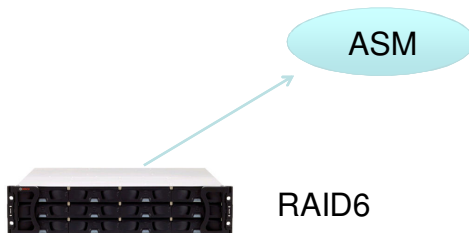
## Notes on a config with 2 arrays

- Clusterware voting disk need be deployed in a odd number of copies
  - Typically 3
  - A limitation of the algorithm used by Oracle
  - Therefore for a 2-array config one array failure can bring down the cluster
    - There is the possibility of a work-around, but not too elegant..



## ..with 1 array

- When using only 1 array, probably better use RAID6
  - ASM diskgroup with external redundancy
  - May be good for **test**/non critical DBs
  - or simply **low budget/lower** maintenance **effort**



## Notes on a config with 1 array

- Often total disk capacity is quite high
- Sizing should take in consideration number of disks for IOPS
- Create 1 or 2 volumes with RAID6
  - then create LUNs of 2TB
  - ASM 10g does **not want LUNs > 2TB**
  - No need to use all the available space
    - **Leave spare LUNs** to be added to ASM in the future is needed



## ASM over RAID

- ASM 'external redundancy' over RAID volumes
  - Makes sense for the 1-array config described
  - Makes sense for HA configs when 'enterprise storage' is used
  - Does not protect against controller failure
  
- ASM 'normal redundancy' over RAID volumes
  - Provides protection for more than 1 simultaneous failure
  - That is **more failures** can happen while redundancy is being restored
  - Reduced performance: in particular writes suffer
  
- ASM with high redundancy
  - Protects against 2 simultaneous HW failures, no experience there



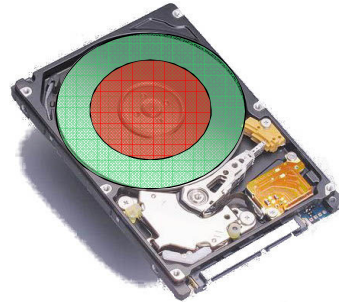
## Data Protection, Redundancy and More

- From architecture point of view
  - How much protection to build with redundancy
  - Where do one start with backup protection
  - Where to put the line for disaster recovery protection?
  
- Example from CERN (see Jacek's presentation):
  - ASM normal redundancy
  - Flash copy of the DB
  - Backup to tape
  - Dataguard protection



## ASM and Flash Copy

- **Flash copy** of DB stored in the recovery diskgroup
- Recovery diskgroup created on:
  - The internal part of the disk (slower part)
  - While data is in the data diskgroup on faster part of the disk
- Recovery diskgroup with 2 failgroups
  - One of the **2 failgroups** on dedicated storage **for critical DBs**
- This is also a way to make use of extra storage when using large disks

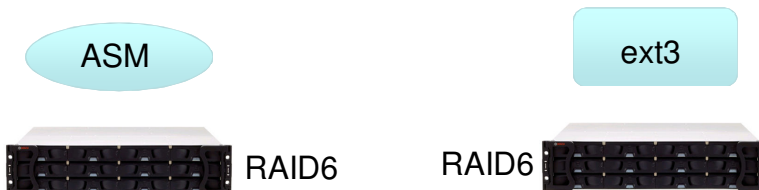


Note: storage sizing is often on IOPS, that is number of spindles



## When tape is not available

- Create 2 (or more) filesystems on RAID6 volume
  - On dedicated storage array
  - Mount ext3 filesystem locally on one node
    - ASM diskgroup considered too risky as this is a backup solution
  - RMAN backups to filesystem
    - Parallelized on multiple channels, one per filesystem





## Notes on installation

- **Prerequisites:**

- Create volumes and export LUNs
- Setup FC zoning
- Partition disks OS level
- Set permissions for oracle user at os level
- Device mapper setup for multipathing
- HBA driver parameters, for example ql2xmaxqdepth
- Raw devices for CRS and ASM spfile

- **Sharing info:**

<https://twiki.cern.ch/twiki/bin/view/PDBService/InstallationProcedureRHEL5>

[https://twiki.cern.ch/twiki/bin/view/PDBService/Installation\\_verbose](https://twiki.cern.ch/twiki/bin/view/PDBService/Installation_verbose)



## Storage Sanity Check

- **Test performance and stress test for robustness**

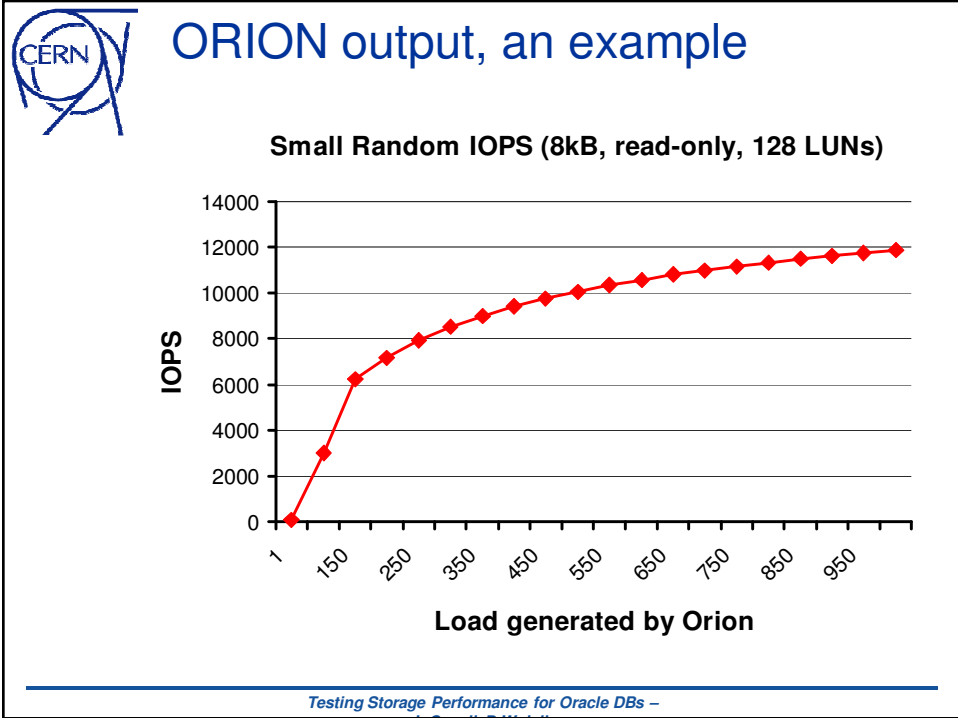

- **ORION** is the easiest
  - Basic IO metrics measured by ORION
  - IOPS for random I/O (8KB)
  - MBPS for sequential I/O (in chunks of 1 MB)
  - Latency associated with the IO operations

- **Simple to use**

- Get started:
 

```
./orion_linux_em64t -run simple -testname mytest -num_disks 2
```
- More info:
 

<https://twiki.cern.ch/twiki/bin/view/PDBService/OrionTests>

## ORION results, small random read IOPS

Disks Used	Array	IOPS	IOPS / DISK	Mirrored Capacity
128x SATA	Infotrend 16-bay	12000	100	24 TB
120x Raptor 2.5"	Infotrend 12-bay	17600	150	18 TB
144xWD 'Green disks'	Infotrend 12-bay	10300 <i>12600</i>	70 <i>90</i>	72 TB <i>22 TB</i>
96x Raptor 3.5"cmsonline	Infotrend 16-bay	16000	160	6.5 TB
80x SAS Pics	Netapps RAID-DP	17000	210	7.5 TB

*Testing Storage Performance for Oracle DBs –*





## Actual ASM Installation

- The easiest part when all of the above is done!
- Preferred method for 10g
  - Same ORACLE\_HOME for RDBMS and ASM
  - Oracle home installation with cloning
  - Example of diskgroup creation on partition 1 (external part) with 6 disk arrays:

```
create diskgroup test2_datadg1 normal redundancy
  failgroup itstor625 disk '/dev/mpath/itstor625_*p1'
  failgroup itstor626 disk '/dev/mpath/itstor626_*p1'
  failgroup itstor627 disk '/dev/mpath/itstor627_*p1'
  failgroup itstor628 disk '/dev/mpath/itstor628_*p1'
  failgroup itstor629 disk '/dev/mpath/itstor629_*p1'
  failgroup itstor630 disk '/dev/mpath/itstor630_*p1';
```

*ASM Configuration Review, Luca Canali*

33



## Conclusions

- Storage setup most important for HA and performance
- ASM and SAN storage for CERN and Tier1
  - Considerable experience cumulated over time
  - Sharing experience
- ASM can have different 'good' configs
  - depending on available HW and desired protection
  - Higher performance needs more DBA maintenance

*ASM Configuration Review, Luca Canali*

34



## Acknowledgments

- Physics DB team, in particular discussions with Maria, Jacek and Dawid
- Discussions with Carmine and Jason