SCALABLE ORACLE 10G

FOR THE PHYSICS DATABASE SERVICES AT CERN

Luca Canali, CERN, Feb 2006

OVERVIEW

The requirements from the physics experimental community for database services are stringent in terms of: high availability, performance, scalability and security. To fulfill such demands, the Database Services for Physics group at CERN (http://cern.ch/phydb) is currently deploying their production databases using Oracle 10g RAC and Oracle ASM on Linux as the main platform. Database storage, of the order of 100 TB, is provisioned using low-cost storage solutions: SATA disks on FC storage arrays (dual-ported). A 2Gb Fiber Channel SAN network is used to connect the database servers to the storage arrays via redundant SAN switches. Oracle ASM is used to build the disk groups with striping and mirroring across the available disks, where the Oracle databases are allocated. CPU scalability is provided by RAC with its cache fusion technology. Backup to disk (flash backup) is implemented together with tape backups to minimize downtime for database recovery. The implementation of flash backups leverages Oracle 10g flash recovery areas and RMAN 10g features. In the following a short description of the architectural components will be given.

ORACLE 10G RAC AND ASM FOR SCALABLE DATABASE SERVICES

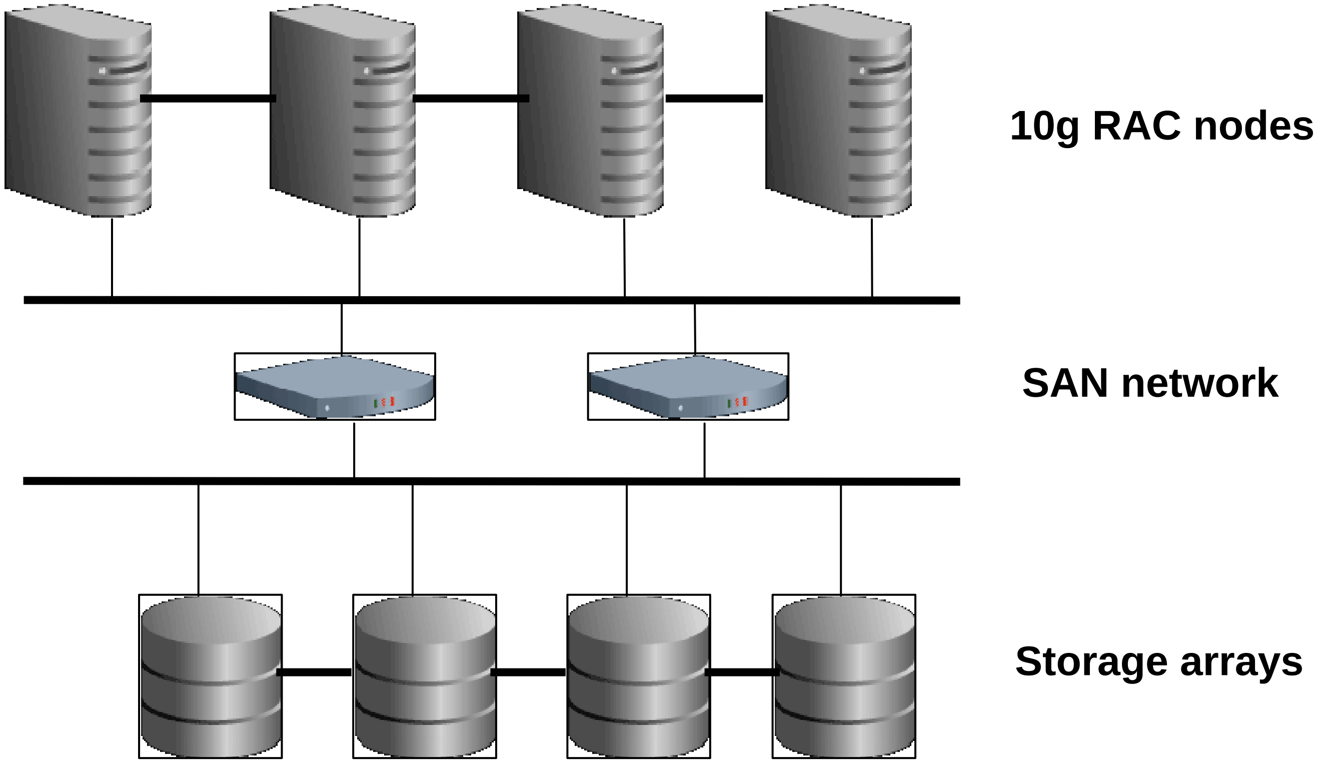

Fig 1: Pictorial representation of the scalable 10g Oracle architecture

A scalable Oracle 10g architecture is currently deployed in production at CERN using the following key elements:

- Oracle 10g RAC on Linux to scale out Oracle workload (e.g. CPU power)
- Fiber Channel SAN network
  - Storage can be reconfigured online using SAN zoning
  - Resilience to fabric failures and load balancing across SAN paths is implemented with multipathing
- Storage built on low-cost SATA disks managed by disk arrays with Fiber Channel controllers
- Oracle ASM used as the volume manager and cluster filesystem for Oracle
  - storage striping and mirroring for performance and HA

Details of the Storage setup

Storage for Oracle databases is built following the ideas underlying the low-cost storage resilient storage initiative (Ref 2). A 2 Gb Fiber Channel SAN network is used to connect mid-range storage arrays. The disk arrays contain 'consumer-quality' SATA disks but have dual-ported Fiber Channel controllers. Disks are mapped from the storage array to the Linux servers directly as LUNs spanning whole physical disks (that is no RAID configuration is used at the array level). Mirroring and striping are done instead at the software level by ASM. This has the advantage to allow mirroring across two different storage arrays (and therefore storage controllers). Storage arrays are dual-ported, where each port is connected to a different SAN switch. Analogously, Linux servers mount HBAs that are dual-ported and connected to both SAN switches.

Oracle RAC databases with 2 and 4 nodes are currently deployed in production. The typical characteristics of the storage arrays used in production are: 16 SATA disks with a raw capacity of 6 TB per array. Storage is configured such that each SATA disk in visible as a LUN to the Linux servers (for example /dev/sdb), there disks are partitioned in 2 halves (i.e. partitions /dev/sdb1 and /dev/sdb2 are created): the outer half of the disk (/dev/sdb1) is used to build data disk group, while the inner half is used to build the flash recovery area disk group. Disk groups are created with ASM and used to allocate Oracle files: data disk groups are used, among others, for datafiles, controlfiles and redologs. Flash recovery area disk groups are used to allocate the Oracle 10g flash recovery area, where, among others, archivelogs and backups to disks are stored.

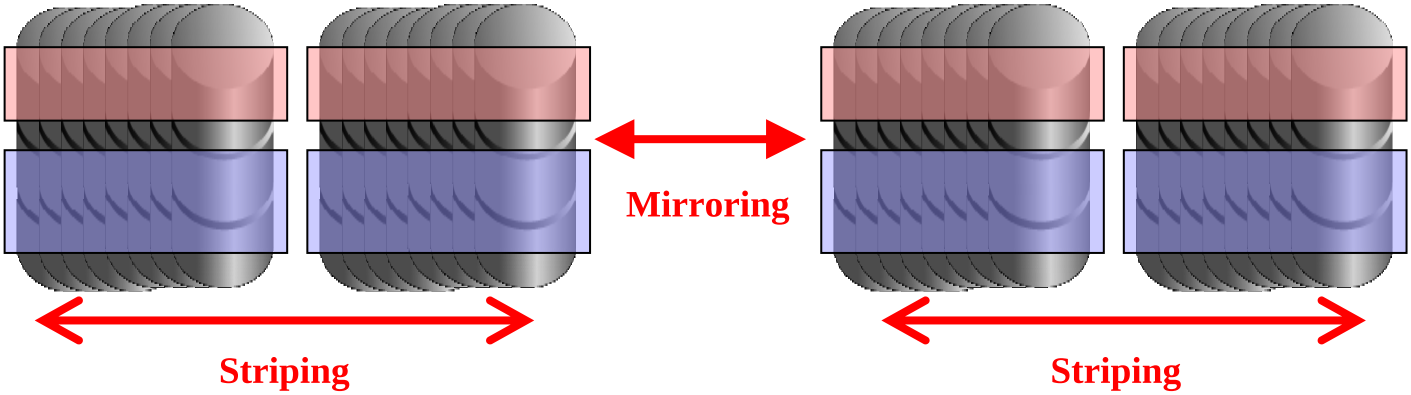

Fig 2: ASM organizes storage into diskgroups, an implementation of the 'SAME' (stripe and mirror everything) guidelines. Two diskgroups are show in the figure (marked with pink and blu colors). Each diskgroup is built by striping and mirroring HD partitions. External disk partitions are used for the data diskgroup, internal partitions are used for the recovery disk group.

A SPECIAL CONFIGURATION

An optimized configuration has been deployed for some production RACs to improve resilience, recovery time and performance. There two Oracle RAC databases are partially coupled together (see figure below): data and flash recovery areas for a production database are interleaved with the storage for a non-production database (such as test, validation, development). The non-production (test) databases that are coupled with production are chosen because they have a relatively low workload.

The benefit of this configuration is that the production database can take advantage of the whole set of disks allocated in the 4 storage arrays. A negligible amount of I/O contention between the production and test database is expected because the latter has a low workload. Another advantage is that, in case of failure or logical corruption of the disk arrays containing the production database, database recovery can be done by switching to the flash backup held in the flash recovery area (see also the paragraph on flash backup).

A study of the measured performances of the disk configuration as described here can be found in Ref 3 and 4.

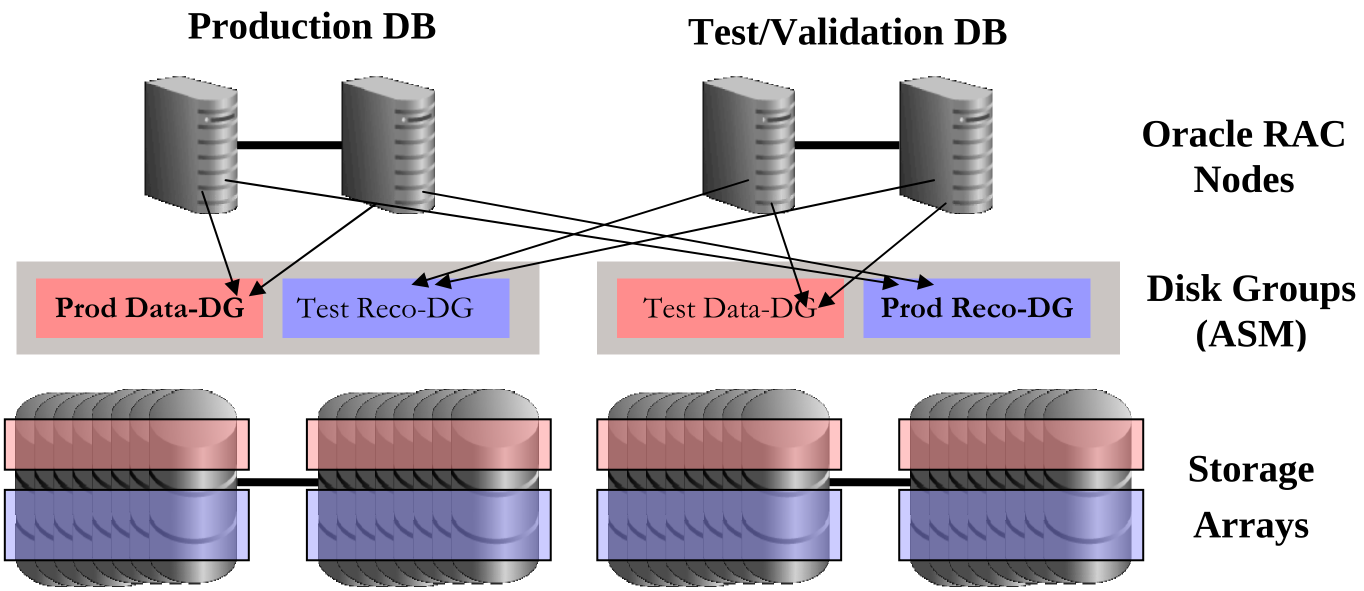

Fig 2: Oracle RAC deployment where production and test databases are partially coupled together. Each couple of ASM storage arrays is mirrored with ASM. Four disk groups have been created (data and flash recovery areas for 2 databases). Production data-DG contains the production database and allocates storage from the first 2 storage arrays from the left, while the flash recovery area contains the backup to disk and is allocated on the remaining 2 arrays. In case of failure or logical corruption of the production data the production database can be switched to the recovery area. I/O contention between production and test is small because test has low workload.

FLASH BACKUPS

Backus to disk (flash backups) have been implemented to complement 'traditional' tape backups and to reduce recovery times for most failure scenarios. A copy of the database is kept in the flash recovery area and is refreshed daily. A copy of the archivelogs needed to recover the flash backup is also kept on disk. For example, in case multiple I/O failures or logical corruption of the production database, the DBA can perform a 'switch operation' to the flash copy on disk, then performs a recovery with the redologs and finally put the production back online. This operation can be performed in a relatively small time window even for very large databases, while the full recovery of hardware and restore from tape backup can take several hours and/or days.

Flash backups come at basically no additional cost using the storage configuration described above. This is because of the large flash recovery areas that are allocated in the storage configuration described above. SATA disks used in the storage array are 400 GB in size each, for performance reasons (see also measurements in Ref 3) only the outer 200 GB are used for data, while the rest is used for the flash recovery areas.

Flash backups are performed using 10g RMAN. In particular Oracle 10g new features are leveraged to incrementally maintain flash backups for large databases (incremental refresh of database copies and block change tracking are two new features that make this possible). In this way the I/O resources needed are proportion to the amount of transactions and not to the overall size of the database.

SUMMARY

Oracle 10g database architecture deployed for the physics database services at CERN has been discussed. Some of the key infrastructural components that allows for a scalable database service are: Oracle RAC for CPU scalability and Oracle ASM with SAN-based low-cost storage for I/O scalability. Backup to disk has also been implemented for higher resiliency and reduced recovery time.

REFERENCES

1. A. Shakian, OOW 2005, take the guesswork out of db tuning, http://www.oracle.com/pls/wocprod/docs/page/ocom/technology/products/database/asm/pdf/take%20the%20guesswork%20out%20of%20db%20tuning%2001-06.pdf

2. J Loaiza and S Lee, OOW 2005 session 1262, http://www.oracle.com/technology/deploy/availability/pdf/1262_Loaiza_WP.pdf

3. L. Canali, https://twiki.cern.ch/twiki/pub/PSSGroup/HAandPerf/Orion_tests_Dec05.pdf

4. L. Canali, https://twiki.cern.ch/twiki/pub/PSSGroup/HAandPerf/RAC_storage_performance.pdf