SCALABLE ORACLE 10G

FOR THE PHYSICS DATABASE SERVICES AT CERN

Luca Canali, CERN, June 2006

OVERVIEW

The Physics Database Services (PDB) group at CERN (http://cern.ch/phydb) runs database services for the Physics community at CERN. With the upcoming startup of the LHC experiments, several dozens of new database-oriented applications are being deployed. The main challenges that the service is addressing are: provide high availability for the mission critical applications, performance and scalability of several multiterabyte applications, contain HW and DB administration costs. Moreover CERN database services are part of a distributed database network connecting several large scientific institutions (Tier 1 sites).

Oracle 10g RAC and ASM on Linux has been chosen as the main platform to deploy the database services. Oracle RAC provides scalability where a **cluster** of many small nodes can be used to load balance the DB workload against shared database storage (shared everything clustering technology). At the same time Oracle **RAC provides high availability,** because the failure of one single cluster node does not bring down the service. Oracle ASM is a volume manager and specialized cluster filesystem which allows the use of low-cost storage array to build **scalable and highly available storage** solution for Oracle 10g. Oracle clusters are composed of a relatively large number of cluster nodes and storage elements, allocated on commodity (low-cost) hardware with homogeneous characteristics. This has the additional advantage of **simplifying administration**, hardware provisioning and service growth. The database architecture deployed at the PDB group embodies the key ideas of grid computing, as implemented by Oracle10g.

There are currently more than 50 dual process nodes and more than 150TB of raw storage deployed by PDB at CERN using this architecture.

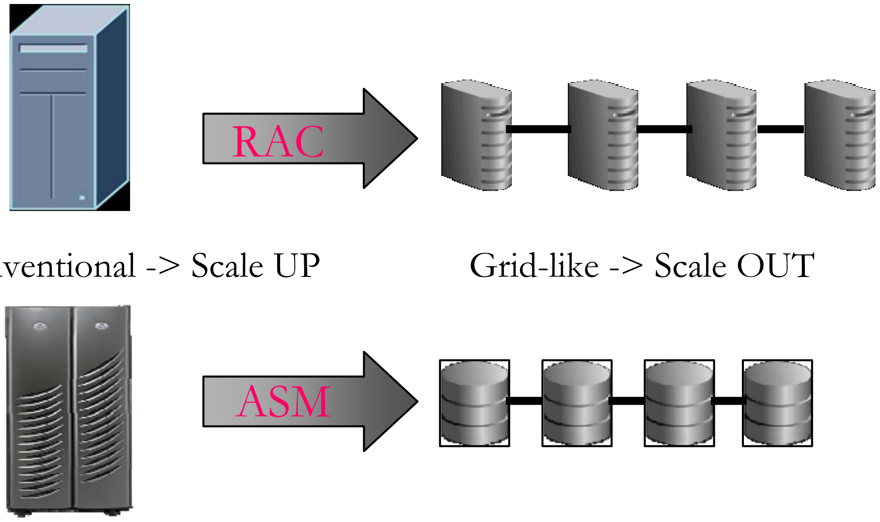

FIGURE 1: *The 'conventional approach' for mission critical database services is to deploy them on large SMP servers and storage high-end arrays (here referred as scale up). Oracle RAC and ASM allow to build clusters with smaller and lower cost components (scale-out the server and storage workload) to achieve the same goals.*

ORACLE 10G RAC AND ASM FOR SCALABLE DATABASE SERVICES

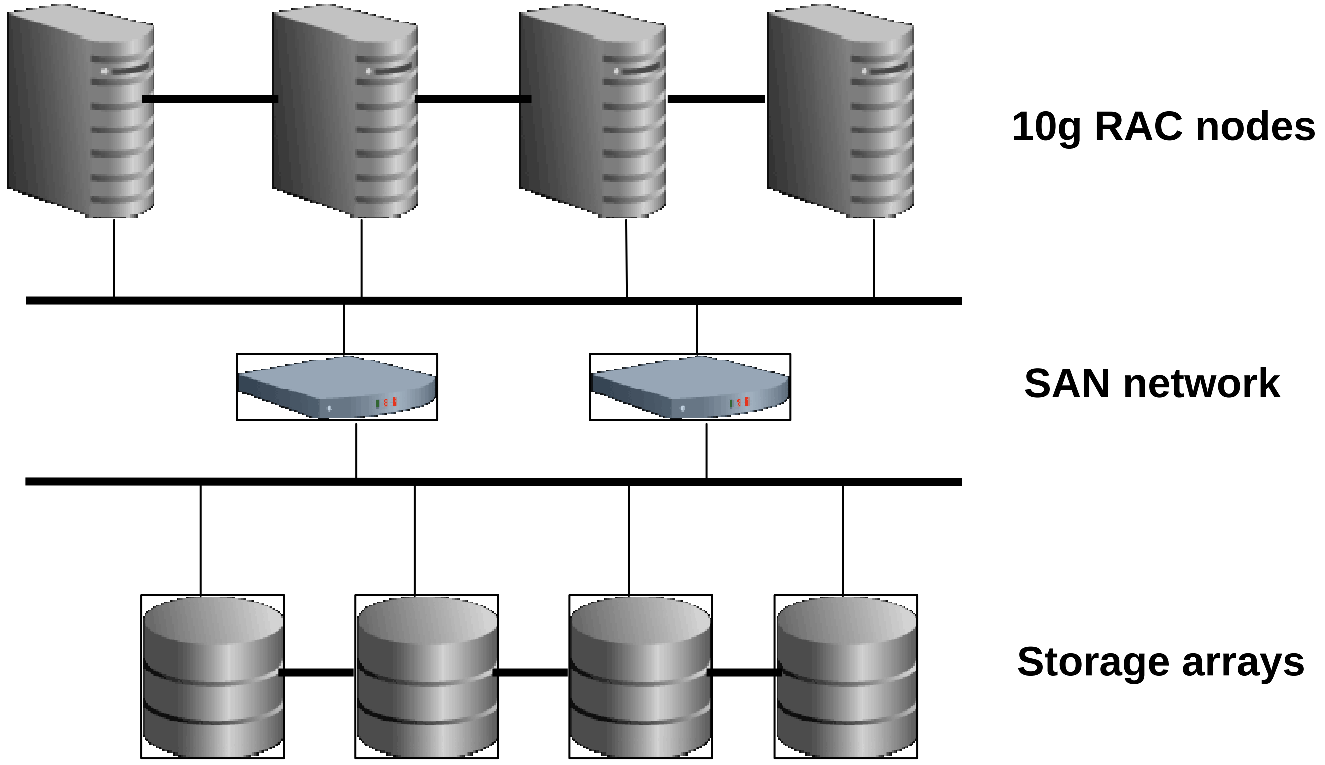

FIGURE 2: *Pictorial representation of the scalable 10g Oracle architecture at PDB*

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

FURTHER DETAILS

Oracle RAC clusters are typically deployed over 4 nodes. Each node is has 2-CPU x86 compatible 32bits @ 3GHz with 4GB RAM. The cluster interconnects are 2 Gbps Ethernet networks. The public network is also on Gbps Ethernet (in some cases 2 NICs).The servers mount HBAs that are dual-ported and connected to redundant SAN switches.

Storage for Oracle databases is built following the ideas underlying the low-cost storage resilient storage initiative: a 2-Gbps 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.

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, then the disks are partitioned in 2 halves under Linux: the outer (faster) half of the disk 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.

A study of the measured performances of the disk configuration as described here can be found in the PDB wiki pages (see links and references section further in this document).

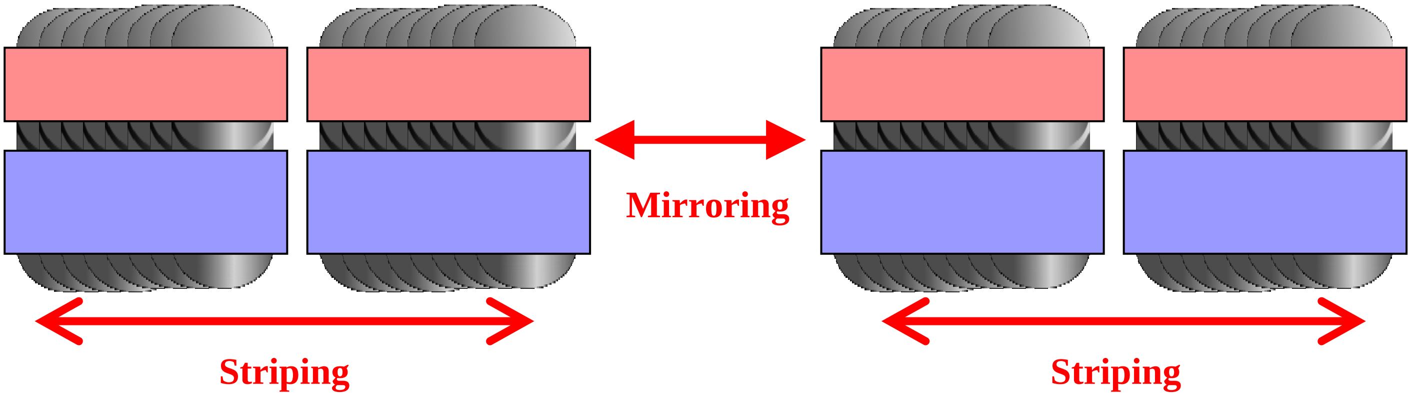

FIGURE 3: *ASM organizes storage into diskgroups, an implementation of 'SAME' (stripe and mirror everything) guidelines. Two diskgroups are shown in the figure (marked with pink and blue). Each diskgroup is built by striping and mirroring HD partitions. External disk partitions are used for the data diskgroup, internal partitions are used for the recovery disk group.*

BACKUPS

Tape backups are currently performed using RMAN with Tivoli storage manager. An incremental strategy is used because of the very large size of the databases: Level 0 (full) backups, level 1 cumulative and level 1 differential backups are scheduled to optimize recovery time and at the same time not overload the database and the backup system.

Backus to disk (flash backups) have also 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 links paragraph) 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

The Physics Database Services at CERN are deployed using Oracle 10g RAC and ASM on Linux. Clusters of relatively low-cost hardware are used to achieve scalability and high availability for the service without paying the price of high-end Unix servers and enterprise storage solutions. The hardware deployment model, where a large number of homogenous servers are bound together via Ethernet and FC networks, allows also for additional savings in provisioning, simplified management and a flexible architecture for growth.

LINKS AND REFERENCES

http://cern.ch/phydb - PDB home page

https://twiki.cern.ch/twiki/pub/PSSGroup/HAandPerf - twiki page with this document and other links on performance and HA