Scalable Database Services for Physics: Oracle 10g RAC on Linux

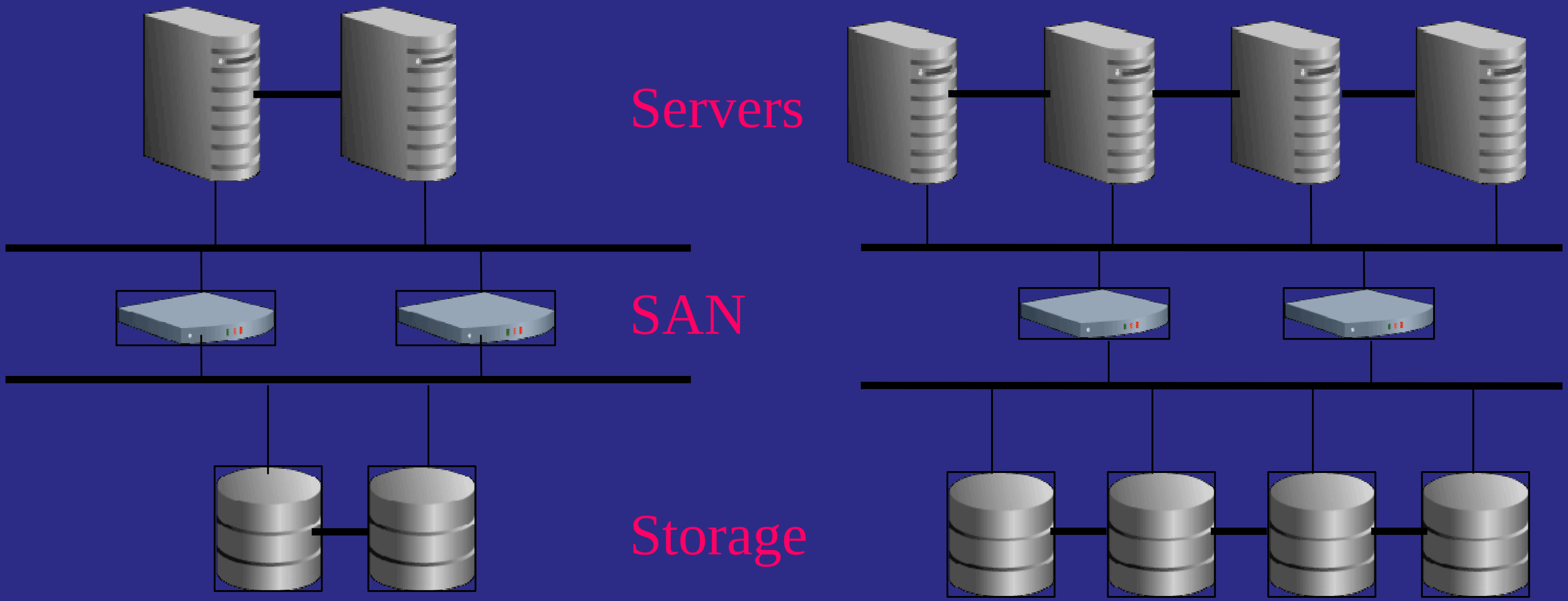
Luca Canali, CERN-IT

Goals

- Review of the architectural components and configurations for RAC 10g at CERN
  - Servers
  - Network
  - SAN
  - Storage
  - ASM
- Focus on installation (DBA)
- Discussion and feedback from Tier 1 installations

Performance and Scalability

- Cluster Nodes and Storage Arrays are added to match applications' growth



Servers

- Mid range PCs
  - Dual CPUs (Xeon 3 GHz)
  - 4 GB RAM, 3 NICs, 1 HBA
- Linux RHEL ES 3 U6
- Oracle 10g R2 (10.2.0.2)
  - Oracle Home installed on local filesystems (no OCFS2)
- Open points
  - 64 bit Linux
  - Larger memory (ex: 8 GB)
  - RHEL 4 (2.6 kernel)

Public Network

- TCP/IP over Gigabit Ethernet
- Redundant switches
  - Different cluster nodes are attached to different switches
- Open points for improvement:
  - More NICs may improve HA and performance
  - Management and backup network

Interconnect

- UDP and TCP /IP over Gigabit Ethernet
  - Oracle may certify RDS over Infiniband
- Two NICs are configured
  - RAC can failover and load balance over the NICs
- Gigabit switches are used

- Open points:
  - CRS can not failover over NICs. Possible solution: NIC bonding and the deployment of switches with L2 trunking

SAN Network

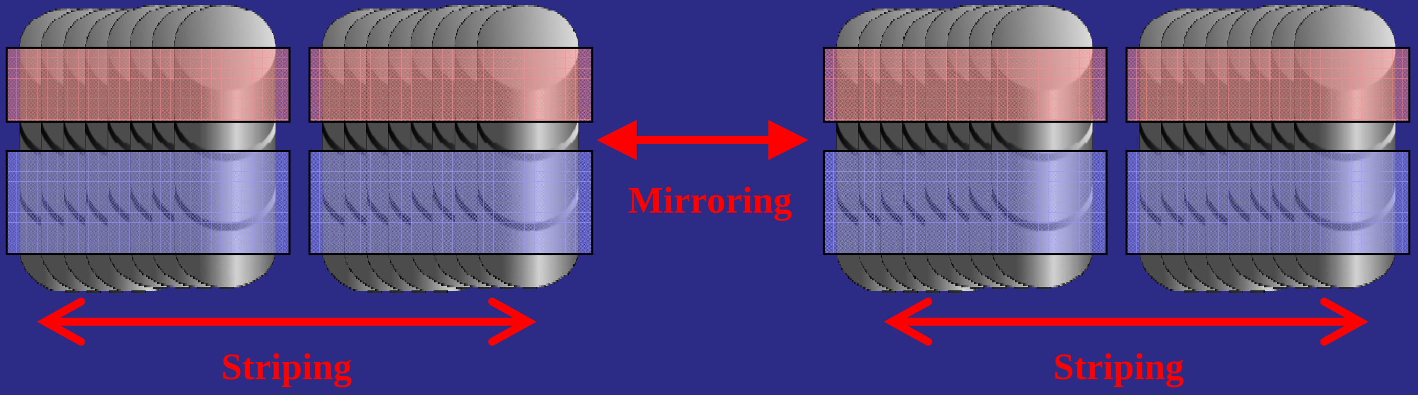
- Fiber Channel SAN (2Gb FC)
- Redundant connections
  - Dual ported HBAs
  - Two SAN switches
  - For failover and load balancing
- Multipathing
  - Leverage the QLogic HBA driver
  - Requires additional configuration

Storage

- Infortrend storage arrays
  - 2 Gb dual ported FC controller
  - SATA HDs (from 8 to 16 disks)
  - Battery backed cache
- We don't use the array's RAID
  - We map the HDs directly as LUNs visible by Linux
  - An extra 1 GB LUN is allocated for CRS (raw devices)
  - ASM is used to stripe and mirror

ASM Storage Configuration

- ASM disk groups created with ‘horizontal’ slicing
  - External part of the disk used for data disk groups -> 🟥
  - Internal part for recovery areas and backup to disk -> 🟦
  - ASM implements SAME (stripe and mirror everything)



Linux LUN Configuration

- Disk partitioning and labeling
  - Each physical disk is mapped as a LUN and visible under Linux as /dev/sd..
  - Two partitions are created (external and internal part of the disk)
  - ASMlib is used to label the partitions and provide persistency across reboot and storage reorganizations
- Special case for CRS files
  - They are allocated as raw devices from the extra 1 GB LUN
  - devlabel (udev on 2.6 kernel) is used to provide persistency for these raw devices

Other Configurations

- Oracle managed files
  - db_create_file_dest='+DATA_DG1'
- Oracle flash recovery area
  - db_recovery_file_dest='+RECOVERY_DG1'
- Connection Management
  - Dedicated Server is used
- Character Set
  - WE8ISO8859P1

Selected init.ora Parameters

- db_block_size = 8192
- parallel_max_servers = 0
- Not set: db_file_multiblock_read_count (autotuned to 128 with 10gR2)
- processes=500
- pga_aggregate_target = 1600m
- sga_target = 1700m
- undo_retention = 3600
- audit_trail = db (audit session is used)
- recyclebin = off
- db_domain='cern.ch'
- global_names=TRUE
- job_queue_processes=10

Oracle Listener Security

- Choose listener port (1521 or non default)
  - Configure firewall (HW and/or netfilter)
- Security has many layers:
  - Oracle's security checklist
  - Scan for weak or default passwords
  - Check for published info on the web also by other sites
    - 'Social engineering' is a threat for complex environments
  - Timely installation of the latest CPU patch
    - A 'must' but not necessarily enough: unpublished vulnerabilities exist
- Other configurations to consider:
  - Encryption
  - Listener password
  - Remove EXTPROC services from the listener
  - XDB can be used to open ftp and http

Backups

- RMAN - backups to tape. Current incremental strategy:
  - Level 0, every 2 weeks
  - level 1 cumulative, twice per week
  - level 1 differential, every day (except when the cumulative backup is done)
  - archivelogs backups, every 30 minutes
  - Retention: recovery window of 31 days (may change)
- RMAN - backups to disk:
  - Daily refreshed with incremental recovery
  - Image copy delayed from production (2 days)
  - Allows for very fast recovery, for many failure scenarios
- Regular tests of recovery procedure recovery
- Open point: disaster recovery / dataguard

Conclusions

- Review of the 10g RAC architecture and configuration deployed at CERN
- More details on WIKI: https://twiki.cern.ch/twiki/bin/view/PSSGroup/LCG3DWiki
  - Installation documentation
  - Init.ora parameters