# Database Services for Physics at CERN with Oracle 10g RAC

HEPiX - April 4th 2006, Rome

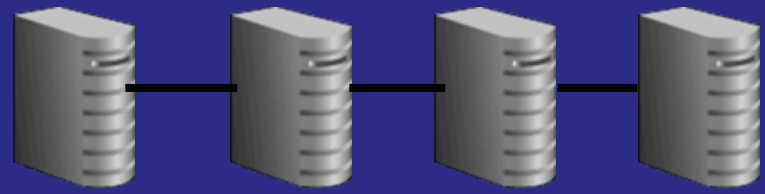
Luca Canali, CERN

# Outline

- Oracle 10g technology
  - Focus on RAC and ASM
  - Scale out vs. scale up
  - Scalable storage using low-cost components
- Oracle for HEP at CERN
  - Deployed hardware
  - Services provided
  - Latest improvements

# Architectural Goals

- Enterprise class performance + HA at low cost

Conventional -> Scale UP

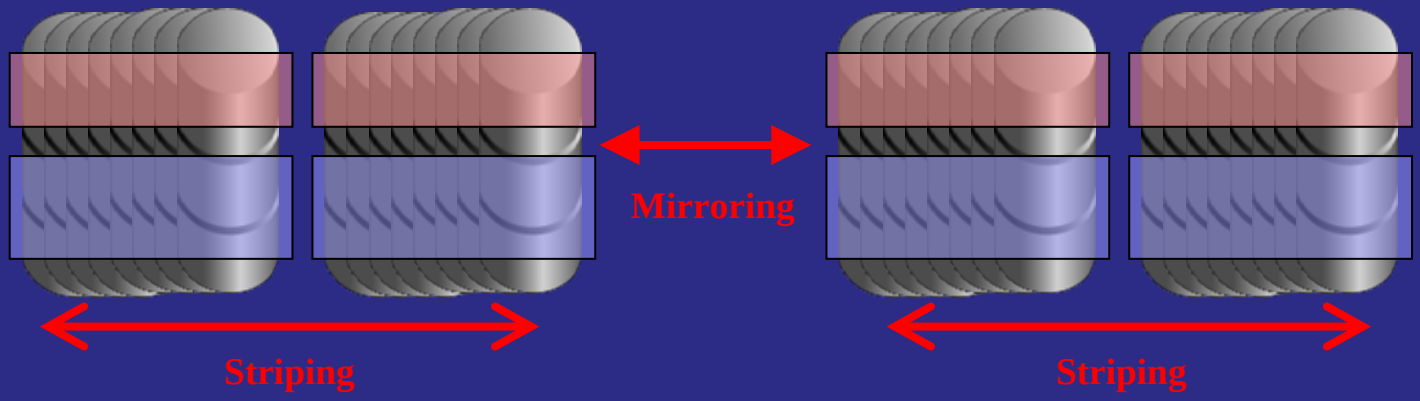
Grid-like -> Scale OUT


# Real Application Cluster

- RAC is a feature of the Oracle RDBMS engine
  - Very High Availability: failed cluster nodes don't stop the service + 'Rolling' software upgrades are possible.
  - Scalability: load balancing across cluster nodes
  - Low Cost: commodity hardware and Linux can be used
  - Deployment: no changes needed for most applications
- Database clustering technologies
  - Oracle RAC : shared-everything + distributed caches (cache fusion)
  - Other RDBMS typically provide 'shared-nothing' cluster architectures (DB2, MySQL, SQLServer)

# Automatic Storage Manager

- ASM is a volume manager and cluster filesystem for Oracle DB files
- Implements S.A.M.E. (stripe and mirror everything)
  - Similar to RAID 1 + 0: good for performance and HA
- Online storage reconfiguration (ex: in case of disk failure)
- Ex: ASM 'filesystems' -> disk groups: DiskGrp1 DiskGrp2



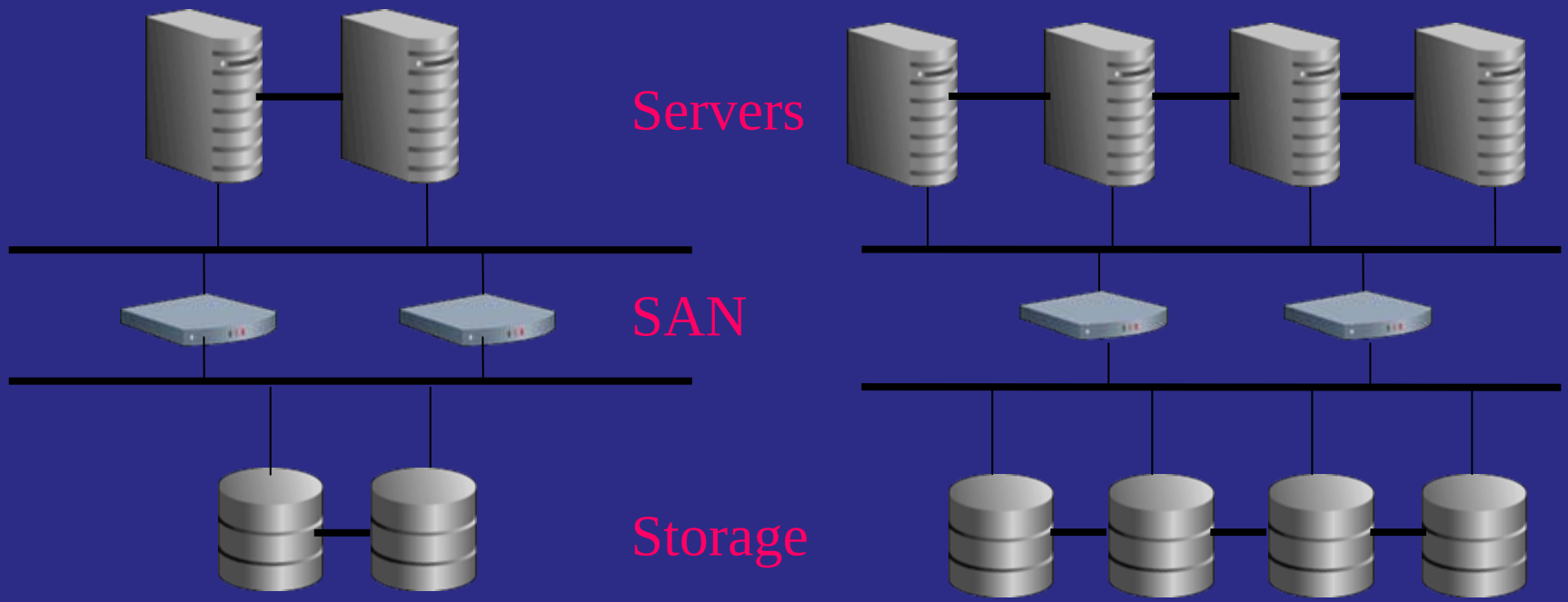
# Performance, Capacity and Cost

- ASM can be used to 'scale out' low-cost storage:
  - I/Os per second:
    - Tests at CERN showed nearly linear scalability up to 64 HDs
    - ~ 100 IOPS per disk (SATA disks, small random IO)
  - Sequential throughput:
    - Limited by fabric to 2Gbps (per HBA)
    - Tests on a 4 node RAC at CERN -> ~800MB/s for seq. read
  - High capacity: leverages SATA disks (typical DB size 5-10 TB)

- Comparison with the top performers: Solid State Disks (SSD)
  - SSD has highest performance: ~100K IOPS, latency < 1 ms
  - BUT cost/capacity (SSD vs. SATA) > 1000, while cost/IOPS ~ 1

# Scalable DB Services for Physics

- Cluster nodes and storage arrays are added to match experiments demand.



# Oracle 10g Deployment at CERN

- Oracle RAC 10g R2 on Linux
- Clusters with 4 nodes and 64 HDs for production DBs
  - 2 nodes for validation and other services
- HW deployed in Q2 2006:
  - ~ 40 RAC nodes
  - ~ 400 HDs
- Plans for Q3:
  - Double server and storage capacity

# Users Community

- LHC experiments
  - Offline processing
  - Validation/preproduction environments
  - Some Online setups
- LCG
- Distributed environment with Tier 1 sites (3D)
- Other Physics users, notably
  - COMPASS
  - HARP

# Conclusions

- Physics Database Services at CERN migrated in 2005 to
  - Scalable databases clusters setup based on Oracle 10g RAC
  - Linux mid range servers connected via redundant IP/FC networks
  - Low-cost storage used, but more reliable than IDE 'diskservers'
- Increased availability
  - Fewer interventions, more interventions done transparently
  - More flexible setup. Can more easily grow to meet the demands of the experiments during LHC startup

More info:

- http://www.cern.ch/phydb/
- https://twiki.cern.ch/twiki/bin/view/PSSGroup/HAandPerf