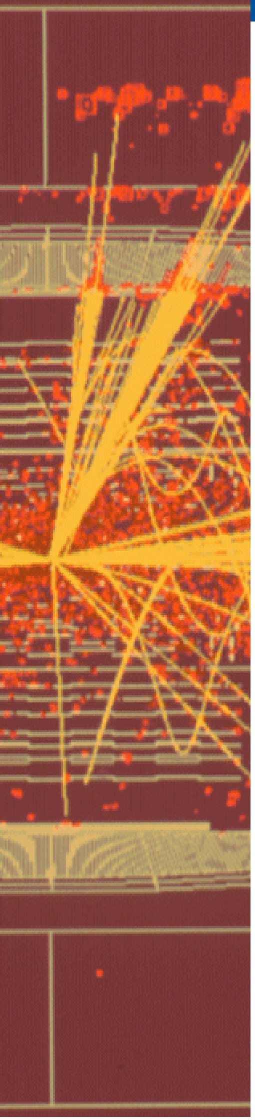
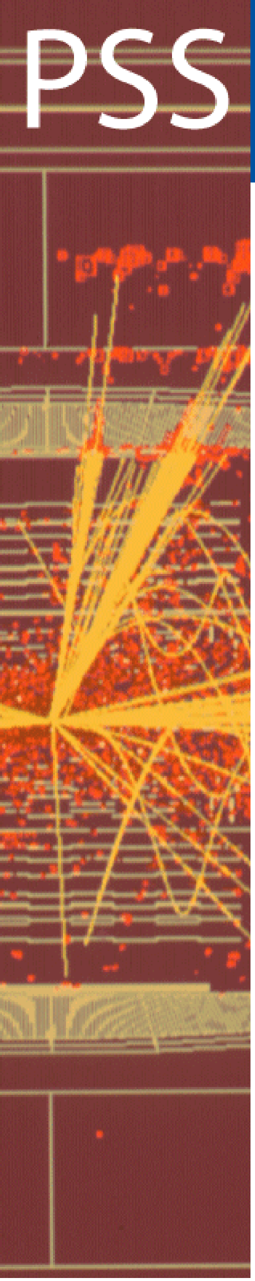


Architecture and Implementation of the Oracle Services for Physics at CERN

Database mini-Workshop

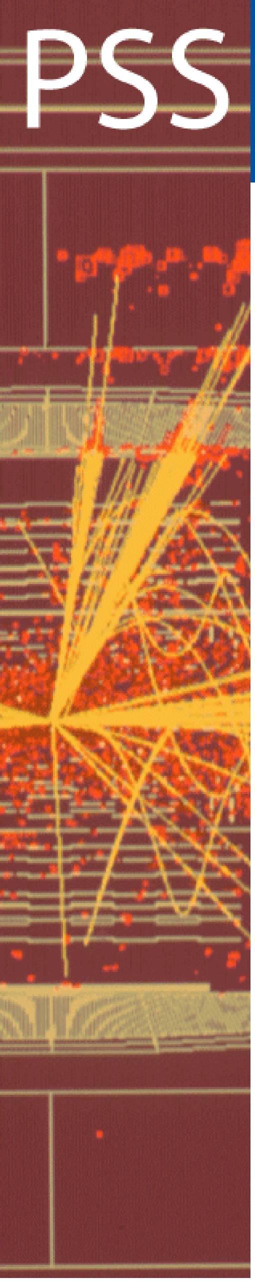
CERN, January 26th, 2007

Luca Canali and Jacek Wojcieszuk, CERN IT



Outline

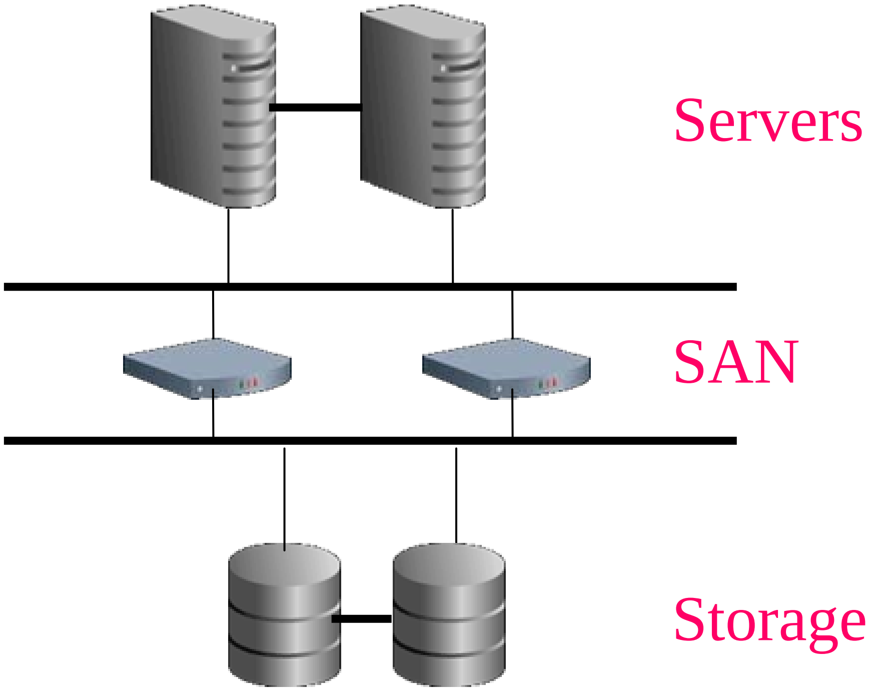
- Architecture
  - DB service for physics
  - Goals and key features
- Selected implementation details
  - Main infrastructural components
  - Lessons learned from CERN’s production DB services



- Run database services to meet the requirements of the Physics community
- Key features:
  - High Availability
  - Performance and Scalability
  - Cost reduction with commodity HW
  - Consolidation
  - Solid backup and recovery
  - Security
  - Distributed databases
  - Operations and Monitoring 24x7

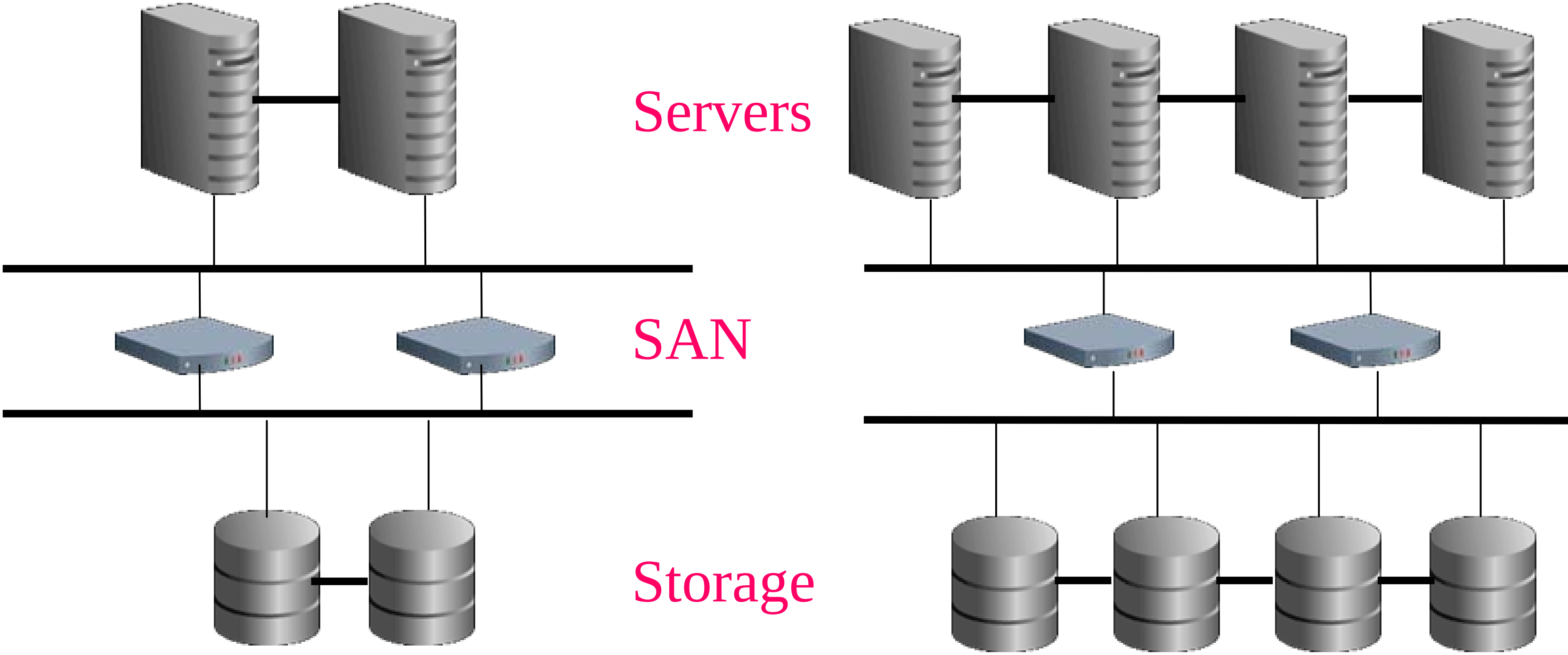
High Availability

- Clustering of redundant HW
- Eliminate single points of failure



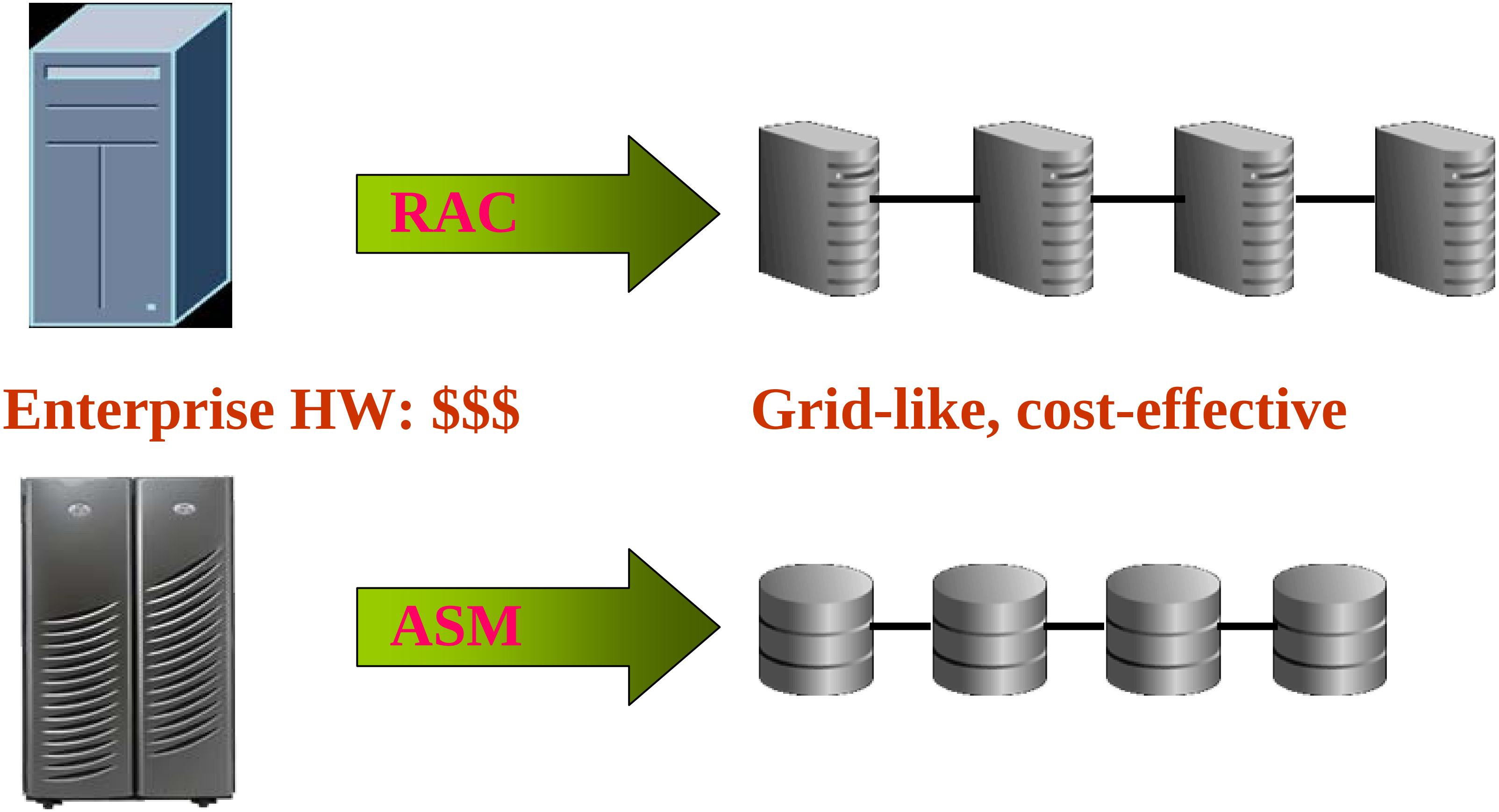
Scalability

- Clusters are expanded to meet growth.



Cost-effective solution

- Enterprise-class HW vs. commodity HW

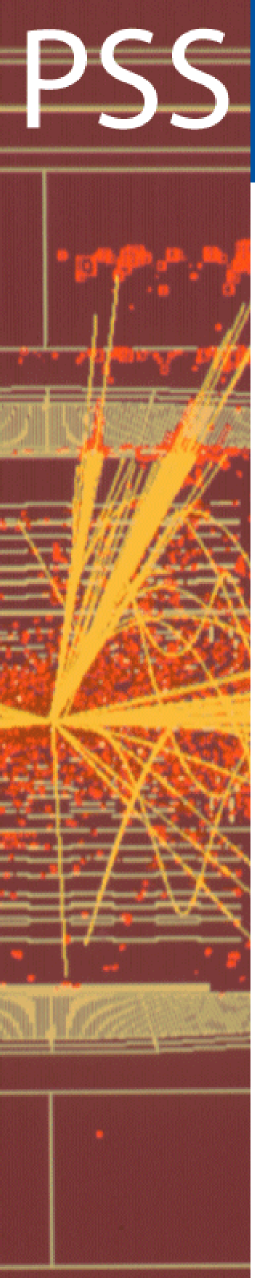


Enterprise HW: $$$

Grid-like, cost-effective

Economies of Scale

- Homogeneous HW configuration
  - A pool of servers, storage arrays and network devices are used as ‘standard’ building blocks
  - Hardware provisioning and setup is simplified
- Homogeneous software configuration
  - Same OS and database software on all nodes
    - Red Hat Linux and Oracle 10g R2
  - Simplifies installation, administration and troubleshooting

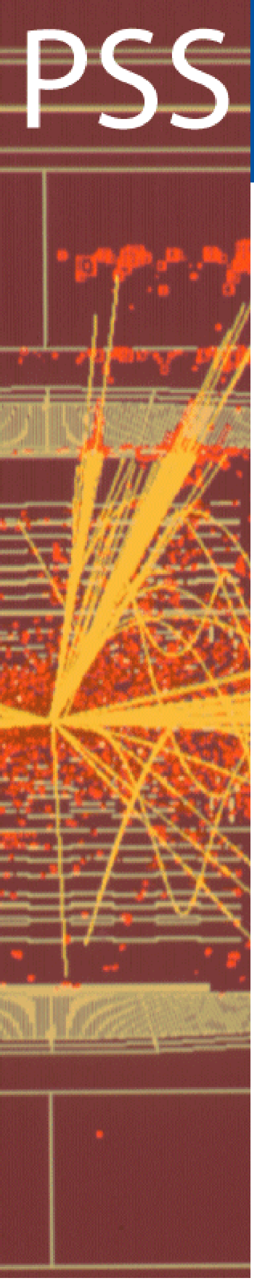


RAC Nodes Configuration

- Current commodity HW
  - Most nodes are dual CPUs
    - Xeon @ 3GHz with 2MB L2 + 4GB of RAM
    - ‘mid-range PC’ with dual power supply and HBA
    - Running Red Hat Linux
  - RAC clusters up to 8 nodes
- Most likely evolution:
  - Scale-up and scale-out, combined:
  - Leverage multi-core CPUs + 64bit Linux
    - Good for services that don’t scale over multiple nodes
    - ‘Dual cores’ currently best power efficiency

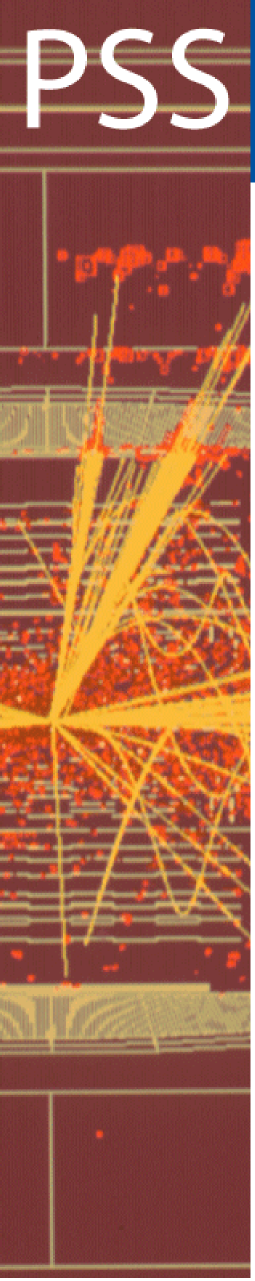
Server Sizing

- How many CPUS?
  - Size CPU power to match the number of concurrent active sessions
  - DB sessions are mostly idle in our workloads, 200-500 DB sessions measured per server
- How many nodes?
  - Architect to grow (don’t need to start with large clusters)
  - Isolate workloads of different applications
  - Leave ‘extra node’ for contingency
- How much RAM?
  - A rule of thumb: 2-to-4 GB per CPU
  - From production: Oracle SGA =2.3 GB, PGA aggregate 1.4GB



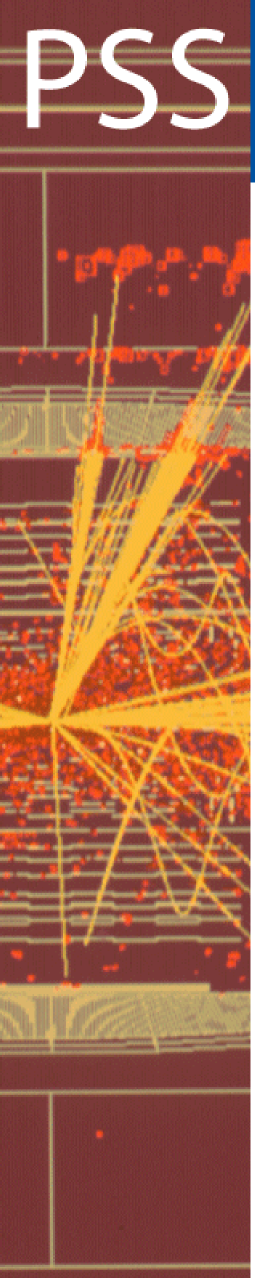
Storage Sizing

- How Much Storage do I need?
  - – Metrics: TB needed, IOPS and MB/sec
    - Requirements ideally from application owners and from stress testing/experience
- Current guidelines from CERN production
  - – 1 TB of ‘usable tablespace data’ -> ~ 2 (mirrored) storage arrays with 8 disks each
  - – IOPS is the most critical metric
    - Consider random I/O (index range scan)
    - 64 disks -> ~7000 IOPS (measured)
    - 8KB Oracle block x 7000 -> ‘only’ 56 MB/sec



Storage Configuration - HW

- SAN at low cost (not an oxymoron)
- FC Storage Arrays
  - Infortrend (A08F-G2221)
  - SATA disks
  - FC controller (dual ported, 2GB cache, 8 disks)
- FC switches
  - QLogic SANBox 5600 (4Gbps)
- Qlogic HBAs
  - Dual ported QL2462
- Redundant fiber connections (multipathing)

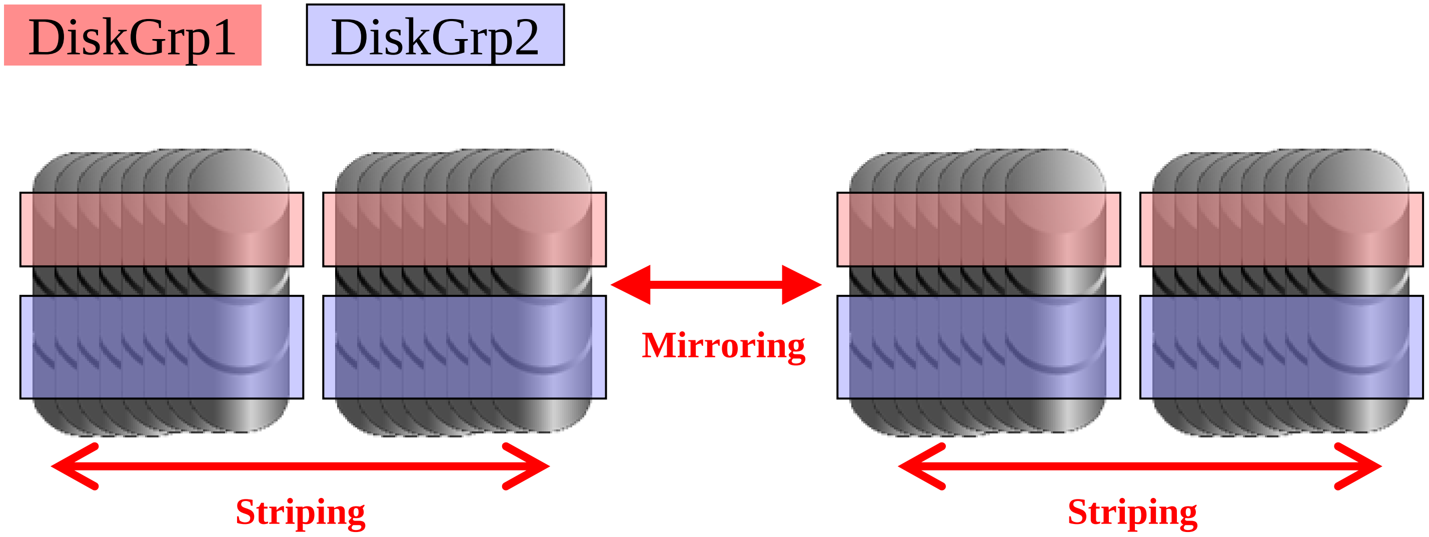


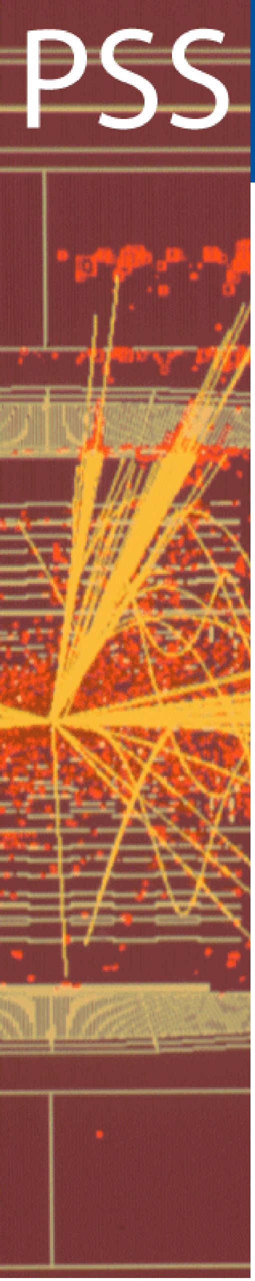
Storage Configuration - OS

- Device name persistency and multipathing
  - RHEL3: devlabel, asmlib and Qlogic multipathing
  - RHEL4: Devmapper
- DevMapper:
  - 2 rpms shipped with RHEL4
  - Only 1 config file (/etc/multipath)
    - Aliases assigned to devices (names persistency)
    - DM ‘block devices’ can be used directly by ASM 10g
  - IO performance, DM vs QLogic multipath
    - Measurements with ORION show no difference

Oracle Cluster Storage, ASM

- Follow the ideas of S.A.M.E. as much as possible
- ASM for mirroring across arrays and striping
- Two diskgroups per DB (data,flash recovery area)
- Destroking: use (mostly) the external part of the disk
- Example:





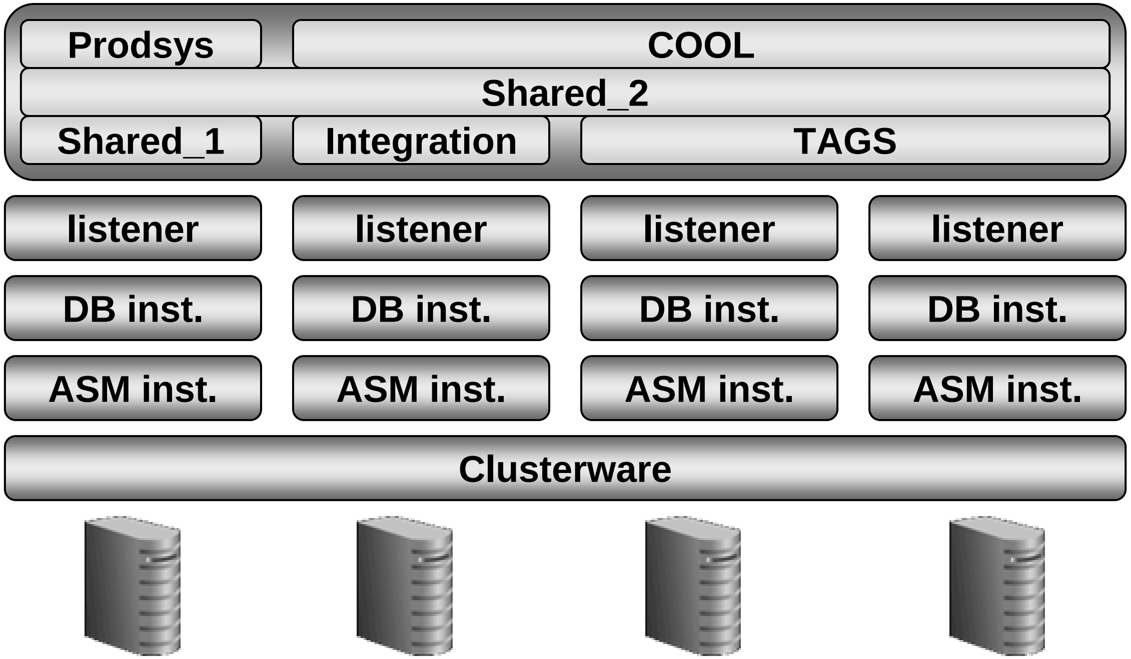
2nd part

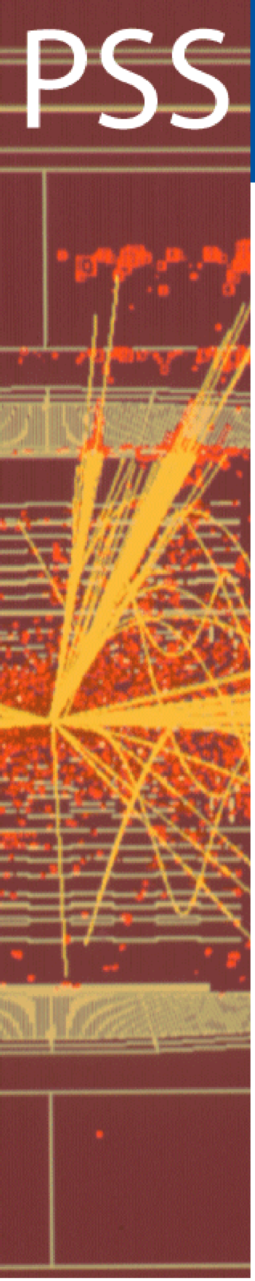
Questions so far?

More Q&A at the end of part 2

Consolidation

- Applications are consolidated on large clusters, per experiment
- Cluster resources distributed among applications using Oracle 10g services
  - Each big application is assigned to a dedicated service
  - Smaller applications share services



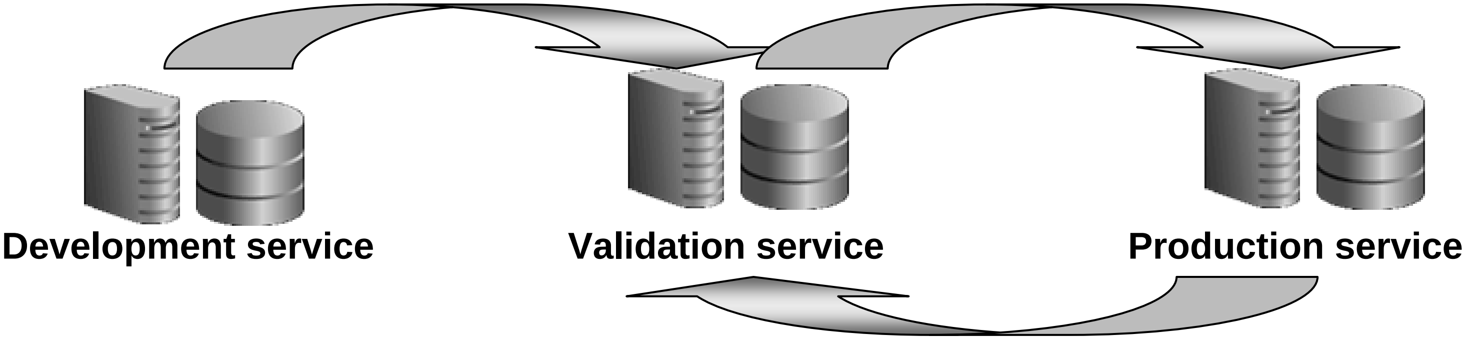


Deployment

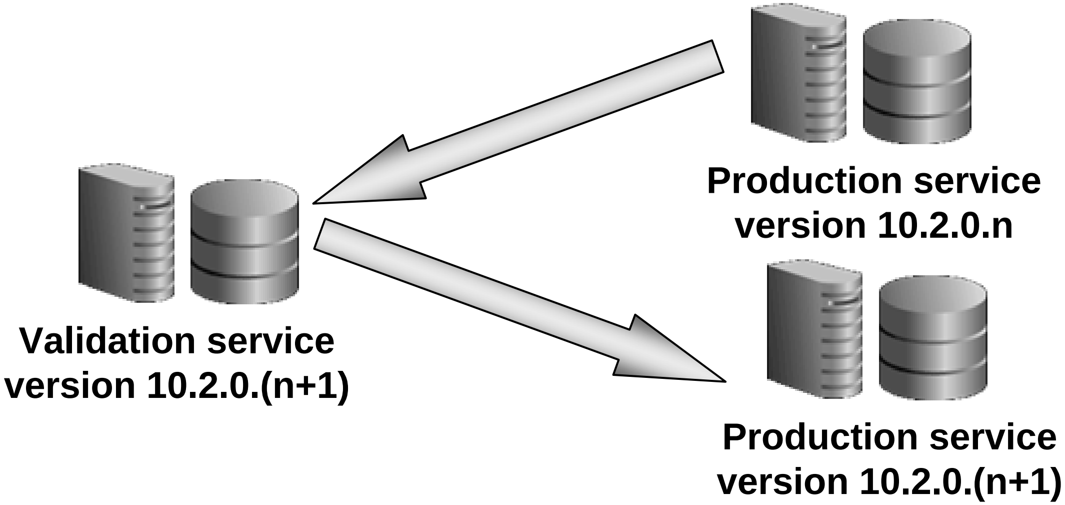
- Deployment model depends on the application:
  - Model 1: service is run on all cluster nodes (default)
  - Model 2: service is assigned to a preferred node (for applications that don’t scale well across multiple RAC nodes)
  - Validation of applications before deployment in production is vital
    - Validation for new application
    - Validation for each new application release

Upgrades and New Releases

- Applications' release cycle

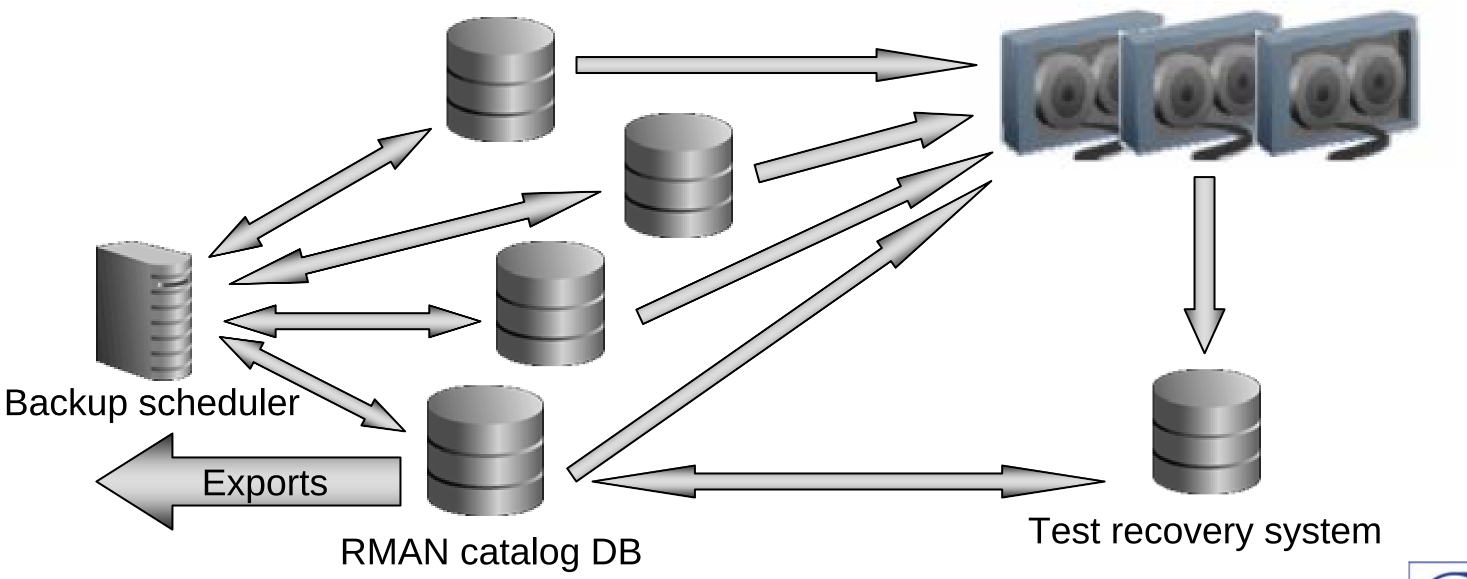


- Database software release cycle


Production service version 10.2.0.n
Validation service version 10.2.0.(n+1)
Production service version 10.2.0.(n+1)

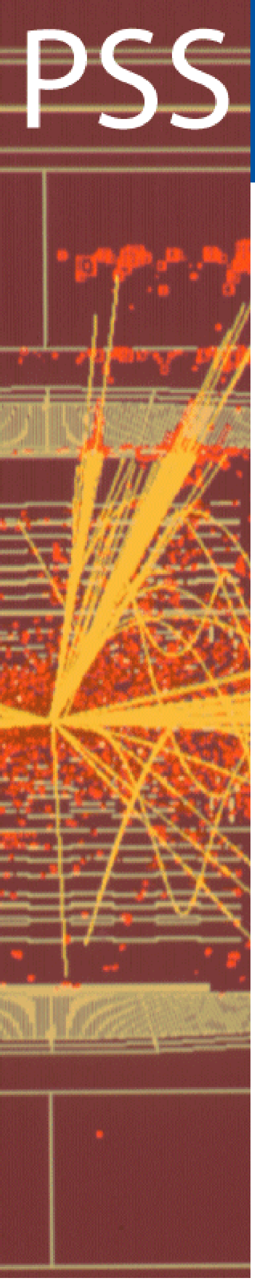
Backup

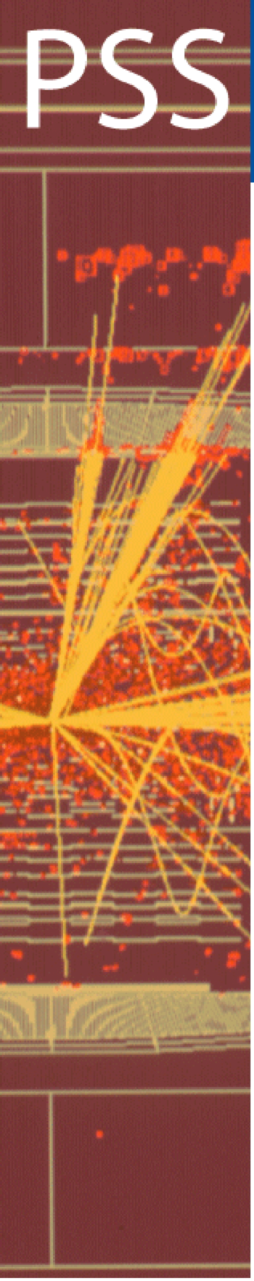
- Solid backup and recovery infrastructure:
  - RMAN is proven technology
  - Interfaced with tape backup system (Tivoli)
  - Dedicated machine for scheduling and running backup jobs
  - Well protected RMAN catalog
  - Dedicated hardware for test recoveries



…and Recovery

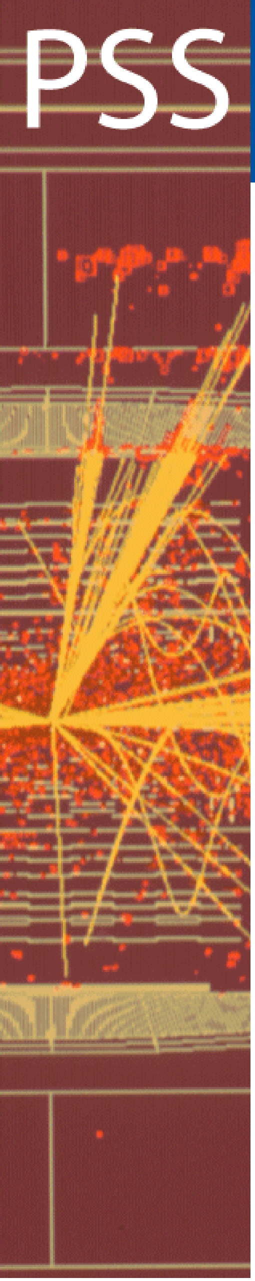
- Test recoveries performed on a regular basis
  - Different scenarios:
    - Full recovery
    - Database point in time
    - Tablespace point in time recovery
    - Recovery from controlfile loss
    - RMAN catalog loss
    - Database duplication
  - Different backups
- Comprehensive documentation
- Evaluation of other recovery methods
  - Data Guard
  - Flashback functionality





Security

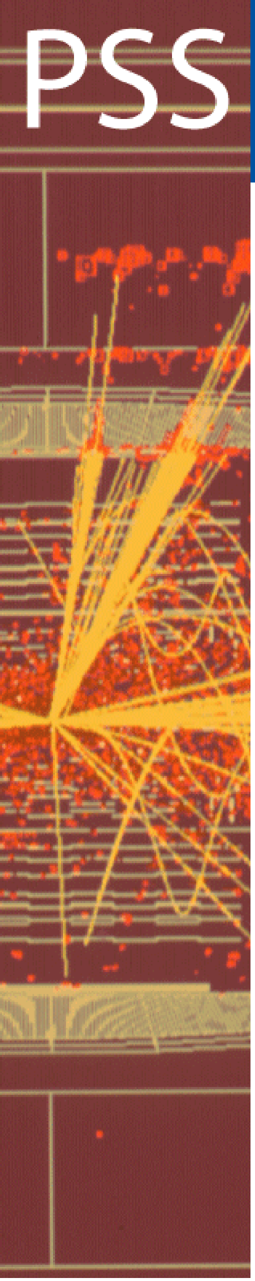
- Extremely important in a distributed environment
- Tools:
  - Several features across the RDBMS engine
  - Firewalls
- Actions taken:
  - Hardware and software firewalls
  - Non-default listener port
  - Password protection
  - Password scans
  - Granting minimum set of necessary privileges
  - Profiles
- Quarterly Oracle security patches (CPU)
- OS security patches



Monitoring & Datacenter Operations

CERN IT Department

- HW is deployed in a datacenter
  - Production is on critical power (UPS and diesels)
  - Leverage the expertise of several groups for installation and maintenance
  - One interface with all the vendors
- 24x7 reactive monitoring
  - Sys-admins, Operators
    - Hardware failures, OS problems
  - Net-admins
    - Network infrastructure
  - DBAs
    - DB instance and host availability (home-grown)
    - ASM diskgroups and DB services (see next talk)
    - Backups (home-grown)
- Pro-active monitoring with OEM and Lemon



Conclusions

- Physics Database Services at CERN run production and integration Oracle 10g services:
  - Positive experience after 1.5 years of production
  - 10gR2 RAC and ASM on commodity HW
  - Currently ~220 CPUs, ~1100 HDs
  - 6 DBAs
- Links:
  - **http://www.cern.ch/phydb**
  - **http://twiki.cern.ch/twiki/bin/view/PSSGroup**