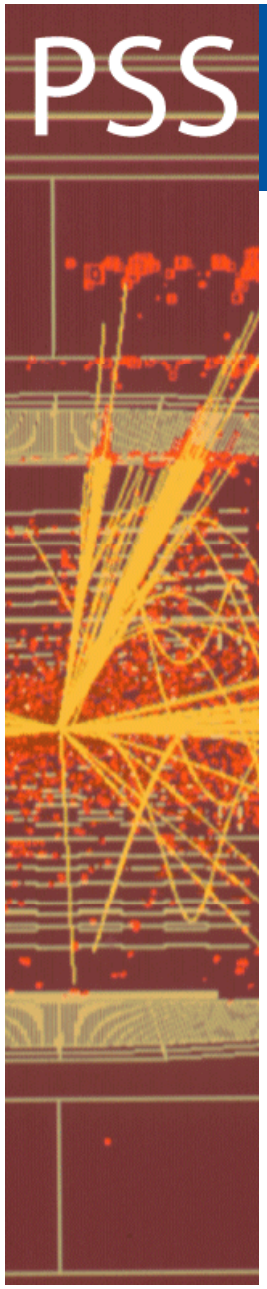


PSS Physics Services Support

CERN IT Department

CERN Database Services for the LHC Computing Grid

Maria Girone, CERN

CHEP'07 VICTORIA, BC

International Conference on Computing in High Energy and Nuclear Physics

2-7 Sept 2007 Victoria BC Canada

PSS

Outline

CERN IT Department

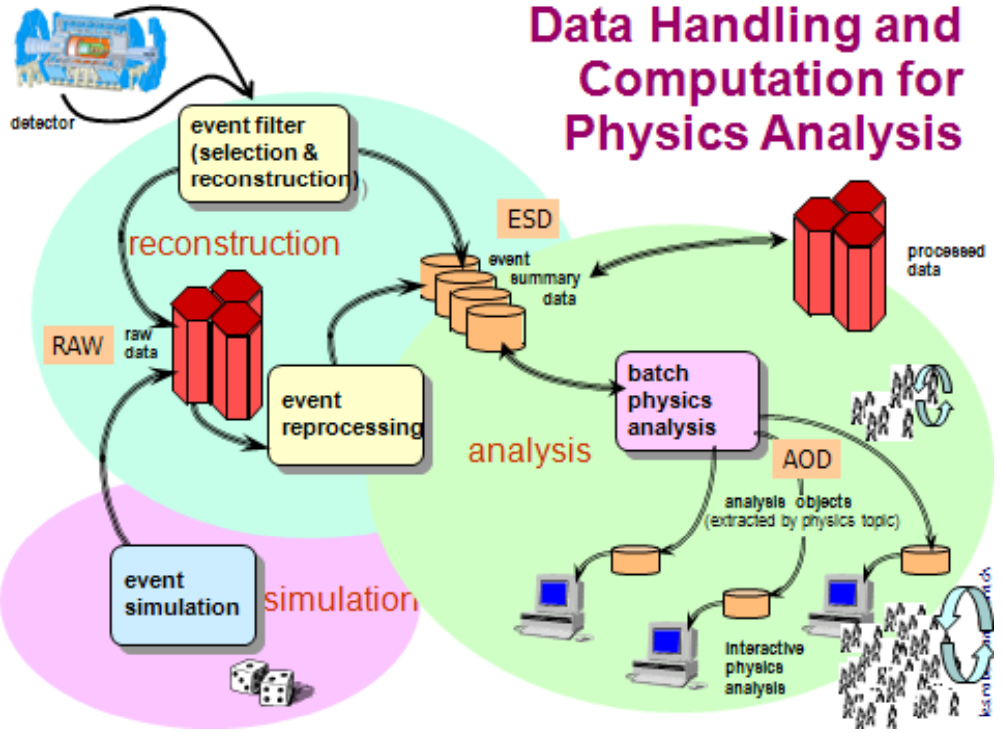
- CERN database services for physics goals
  - **How we address scalability, performance & reliability needs**
- Service set-up and operations
  - Highly available clusters
  - Development, test and production levels
  - Backup and update policies
- Replication set-up to the Tier1
- Service evolution for the LHC start-up
- Conclusions

See also workshop sessions on Robust&Reliable Services

CHEP'07 VICTORIA, BC

Introduction

- Physics meta-data stored in relational databases play a crucial role in the LHC experiments and in the operation of the Worldwide LHC Computing Grid (WLCG) services
  - Detector conditions, calibration, geometry, production bookkeeping
  - Core grid services for cataloguing and distributing LHC data
- Key features:
  - High Availability
  - Performance and Scalc
  - Cost reduction with co
  - Consolidation
  - Solid backup and recov
  - Security
  - Distributed databases
  - Operations and Monito



Some key numbers

- Service based on Oracle 10g Real Application Clusters on Linux
- Service Size
  - 110 mid-range servers and 110 disk arrays (~1100 disks)
  - In other words: 220 CPUs, 440GB of RAM, 300 TB of raw disk space!!
- Several production clusters
  - One production cluster per LHC experiment for offline applications, up to 8-node clusters
  - Online test Atlas cluster
  - COMPASS cluster
- Several validation and test clusters
  - 1 or 2 per LHC experiment of 2-nodes
  - Some hardware allocated for internal use/tests
- Service responsibilities
  - 5 DBAs in the team
  - 24x7 service on "best effort" for the production service

PSS

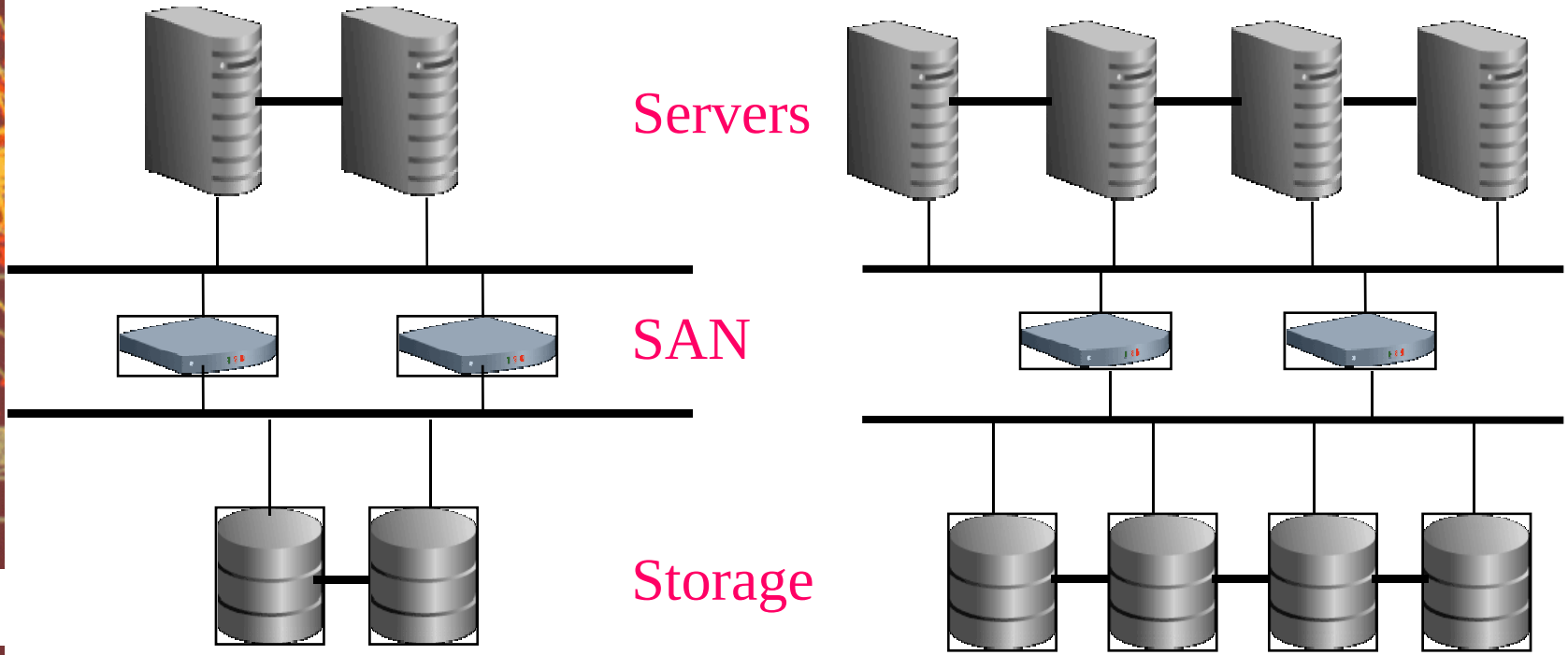
Current set-up

CERN IT Department

CHEP'07 VICTORIA, BC

High Availability & Scalability

- Clustering of redundant HW
- Eliminate single points of failure
- Clusters are expanded to meet growth.



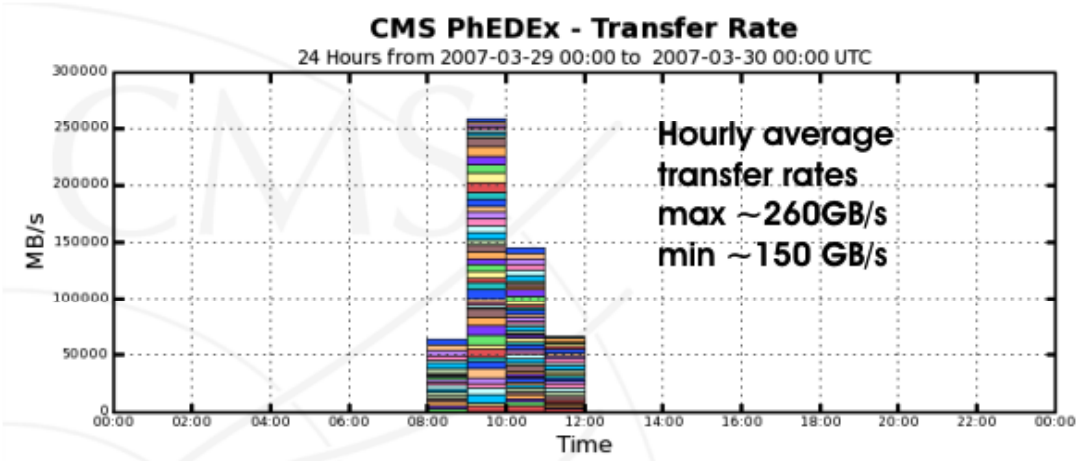
Economies of Scale

- 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 Enterprise Linux RHEL4 and Oracle 10g R2
  - Simplifies installation, administration and troubleshooting

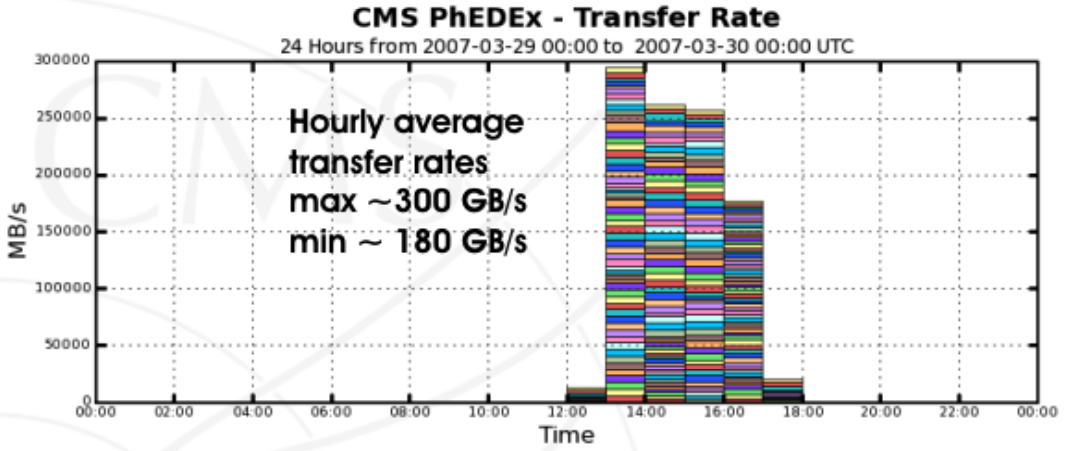
Current Set-Up

- RAC on commodity hardware - Full redundancy!!
  - Linux RHES4 32bit as OS platform
  - Oracle ASM as volume Manager
  - Dual-CPU P4 Xeon @ 3Hz servers with 4GB of DDR2 400 memory each
  - SAN at low cost
    - FC Infortrend disk arrays, SATA disks, FC controller
    - FC QLogic switches SANBox (4Gbps)
    - Qlogic HBAs dual ported (4Gbps)
- 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
    - Tests on 'Quad cores' look promising

Quad-core performance testing



6-node RAC



Quad-core server

A single quad core server is able to handle PhEDEx-like workload (a transaction oriented application) even more efficiently then a 6-node RAC

Server Sizing

- How many CPUS?
  - Size CPU power to match the number of concurrent active sessions
  - Look at the workload on the current production service
  - Leave 'extra node' for contingency
- How much RAM?
  - A rule of thumb: 2-to-4 GB per core
  - DB sessions are mostly idle in our workloads, 200-500 DB sessions measured per server

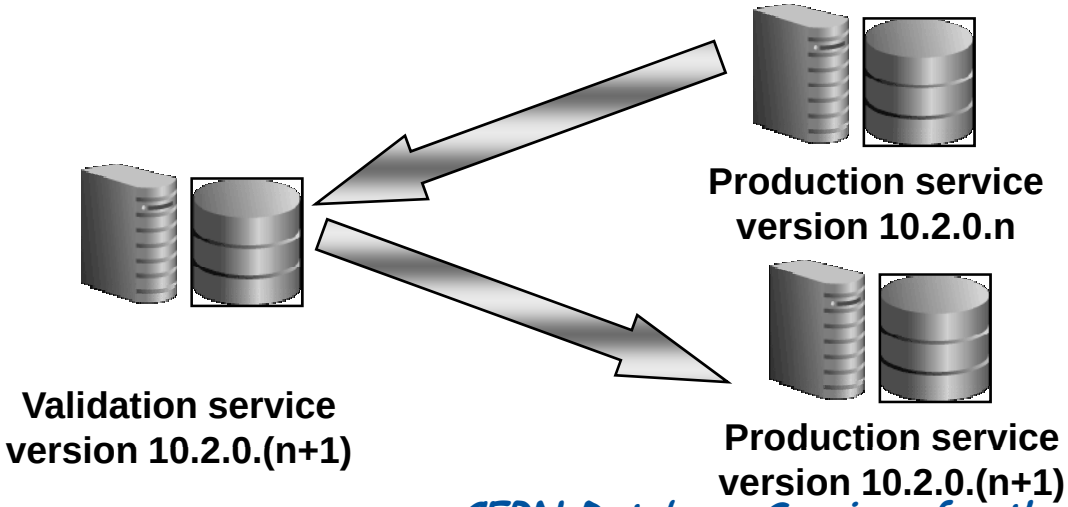
Storage Sizing

- How Much Storage?
  - Metrics: data volume needed and performance for IOPS and throughput
    - Requirements are gathered from experiments and from stress tests
- For our storage set-up
  - IOPS determines the number of disks
    - Consider random I/O (index range scan)
    - 64 disks -> ~7000 IOPS (measured)
  - For data 25% of the row storage capacity is used
    - we implement on-disk backups on the free disk space

Service Levels & Release Cycles

- Applications' release cycle

- Database software release cycle



Backup & Recovery

- Reliable backup and recovery infrastructure
  - Oracle Recovery Manager (RMAN) is proven technology
  - Backup policy agreed with the experiments and accepted also by Tier1 sites
- Backup on tape using IBM technology (TSM), with 31 days retention
- Backup on disk, with 2 days retention
- Automatic test recoveries in place

Security and s/w Updates

- Policy in place for Oracle and OS security patches
  - typically within two weeks, after validation
- Oracle software upgrades are typically performed once or twice per year
  - one month validation
- Oracle patches are only made for recent versions and therefore it is essential to update accordingly

Monitoring & Operations

- Procedures in line with the WLGC services
- Transparent operations
  - Disks replacement (ASM)
  - Nodes reboot (Oracle Clusterware)
  - Security and O/S upgrades
- HW is deployed at the IT computer center
  - Production is on critical power (UPS and diesels)
- 24x7 reactive monitoring
  - Sys-admins, Net-admins, Operators
  - DBAs
- Overall availability above 99.98% over the last 18 months
- Pro-active monitoring with OEM and Lemon

Replication Set-up for 3D

- ATLAS conditions replication setup
  - Tier0 online -> Tier0 offline -> Tier1's
  - All 10 Tier1s are in production
- LHCb replication setup
  - Tier0 online (pit) -> Tier0 offline -> Tier1's for conditions
  - LFC replication Tier0 offline -> Tier1's
  - All 7 Tier1s are in production
- Currently, 8x5 intervention coverage
  - Archive log retention on disk covers weekends
- See for more [171]: "Production experience with distributed deployment of databases for the LCG"

Hardware allocation in 2008

- Production databases for LHC:
  - 3 or 4-nodes clusters built with quadcore CPU machines (24-32 cores per cluster)
  - 48-64 GB of RAM per cluster
  - Planning for >10k IOPS
  - TBs of mirrored space
- Integration and test systems:
  - Single core CPU hardware
  - Usually 2 nodes per cluster
  - Usually 24-32 disks
- 64bit version of Linux and Oracle software
- Migration tools have been prepared and tested to minimize the downtime of the production RACs

Conclusions

- Database Services for physics at CERN run production and integration Oracle 10g services
  - **Designed to address the reliability, performance and scalability needs of WLCG user community**
    - One of the biggest Oracle database cluster installation
    - Approached by the LHC experiments for service provision for the online databases
- Recently connected to the 10 Tier1 sites for synchronized databases
  - Sharing policies and procedures
- Well sized to match the needs of the experiments in 2008
- Planning now the service growth for 2009-2010

https://twiki.cern.ch/twiki/bin/view/PSSGroup/PhysicsDatabasesSection

Sharing cluster resources

- Applications are consolidated on large clusters, per experiment
- We use the Oracle Service concept: partition of a larger cluster available to a application
- Can allocate resources (CPU, num of connects) per service
- Cluster resources distributed among applications using Oracle 10g services
  - Each big application is assigned to a dedicated service
  - Smaller applications share services