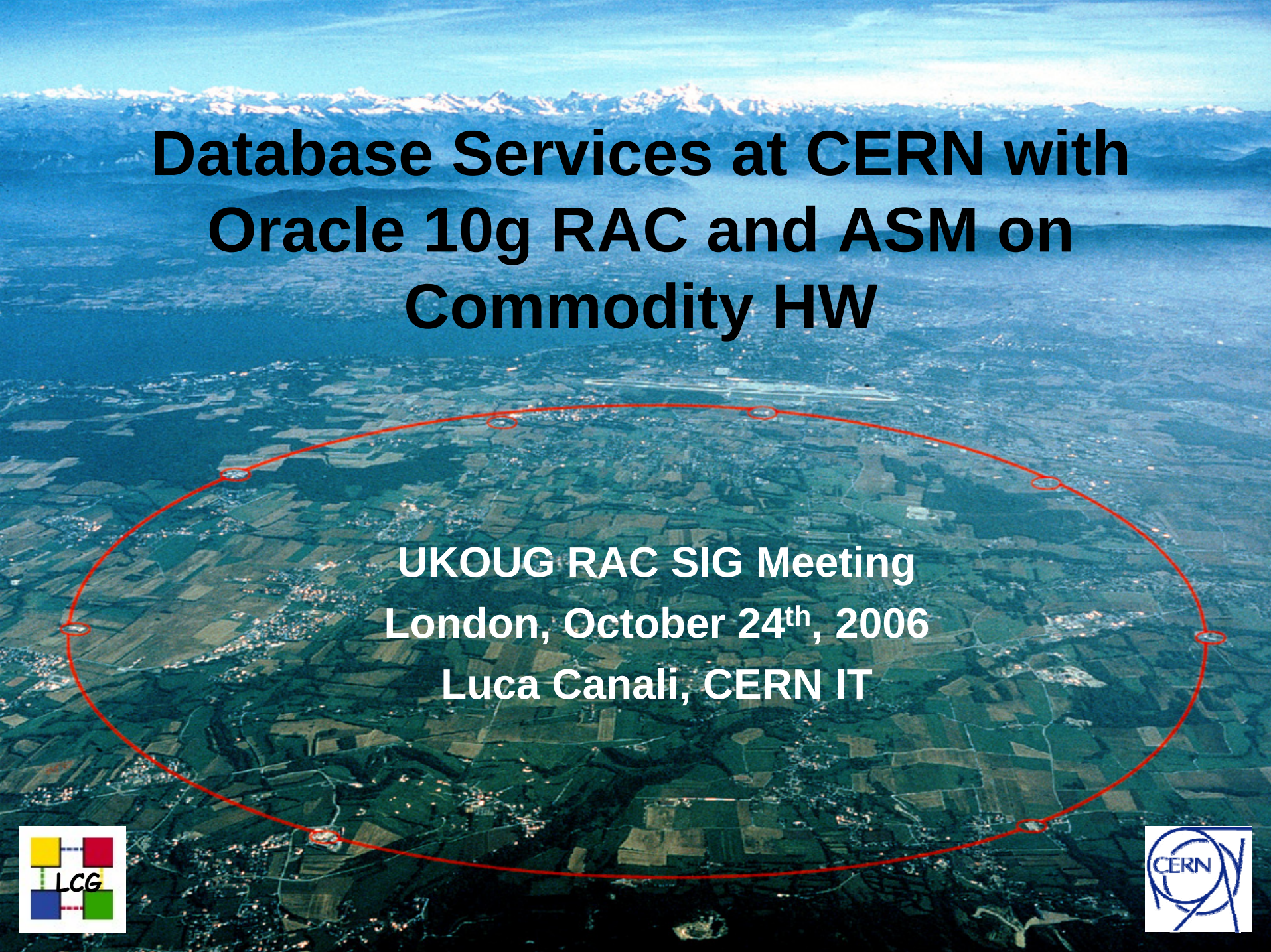
Database Services at CERN with Oracle 10g RAC and ASM on Commodity HW



UKOUG RAC SIG Meeting

London, October 24th, 2006

Luca Canali, CERN IT

Outline

- Oracle at CERN
- Architecture of CERN Physics DB Services
- Sharing experience from CERN's Oracle production. Focus on:
  - Oracle on commodity HW
  - 10g RAC and ASM on Linux

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Why DBs at CERN

- Latest-generation accelerator (LHC) will start next year
- Ultimate goal: solving open questions in particle physics and cosmology
- A staggering amount of data will be produced for analysis
- Long-running collaboration with Oracle

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Physics Services Support

- Run database services to meet the requirements of the Physics experiments
  - Central repository for many LHC applications
  - Mission-critical (respect SLA)
- Requirements
  - High Availability
  - High Performance and Scalability
  - Consolidation
  - Provide a cost-effective solution

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

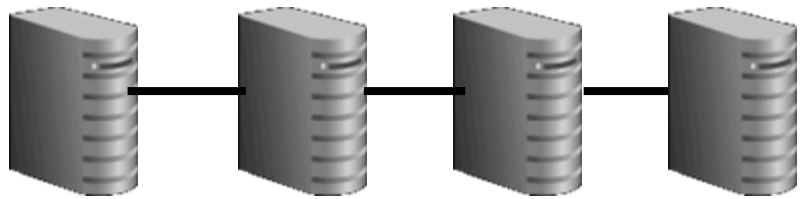
Architecture

- Database Clusters
  - An implementation of grid computing for the database tier for HA and load balancing
- HW
  - Many interchangeable ‘pieces’
  - Cost-effective HW
- Software
  - Cluster database (Oracle RAC)
  - Cluster volume manager and filesystem (ASM)

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

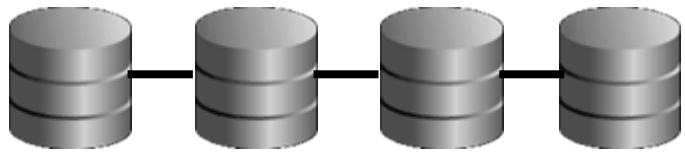
Oracle 10g RAC and ASM

- Enterprise-class HW vs. commodity HW



SMP, Scale UP

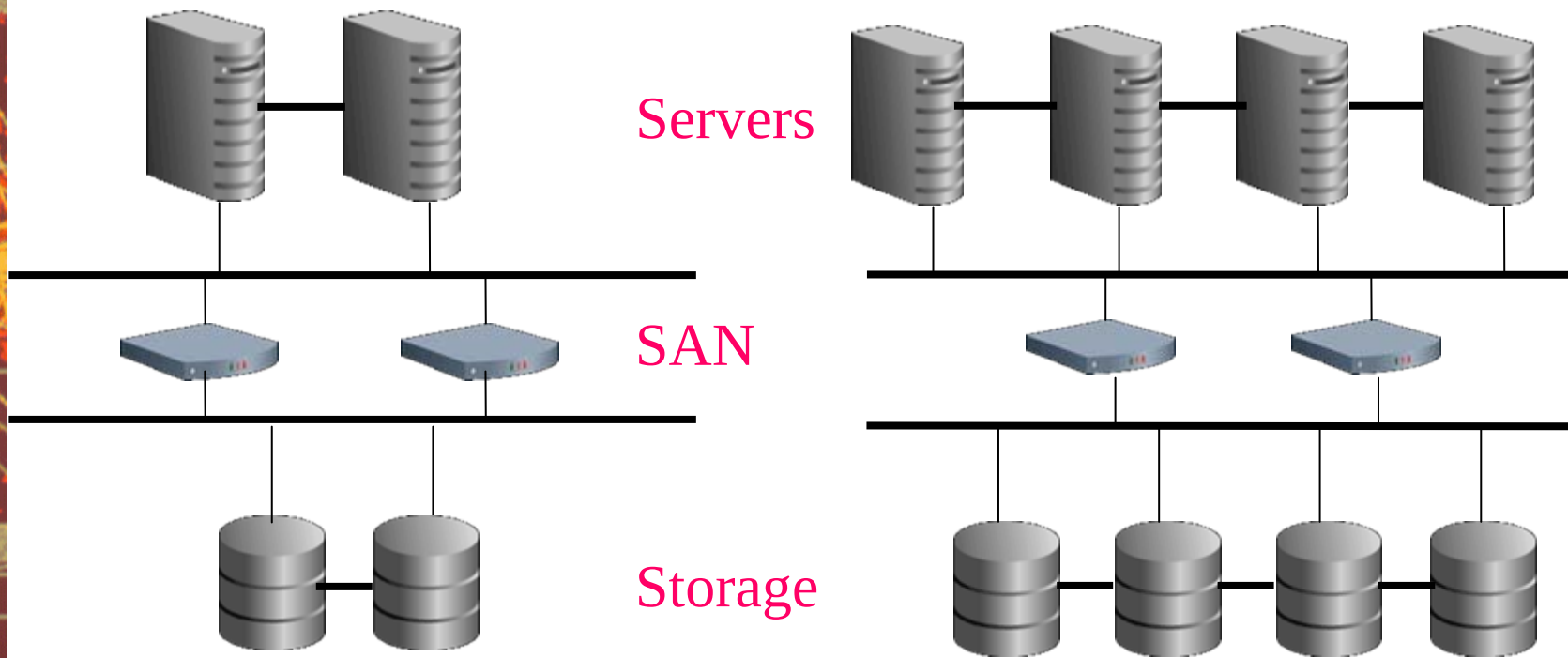
Grid-like, Scale OUT



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Growth on Demand

- Clusters are expanded to meet growth.



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Storage Configuration - HW

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

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

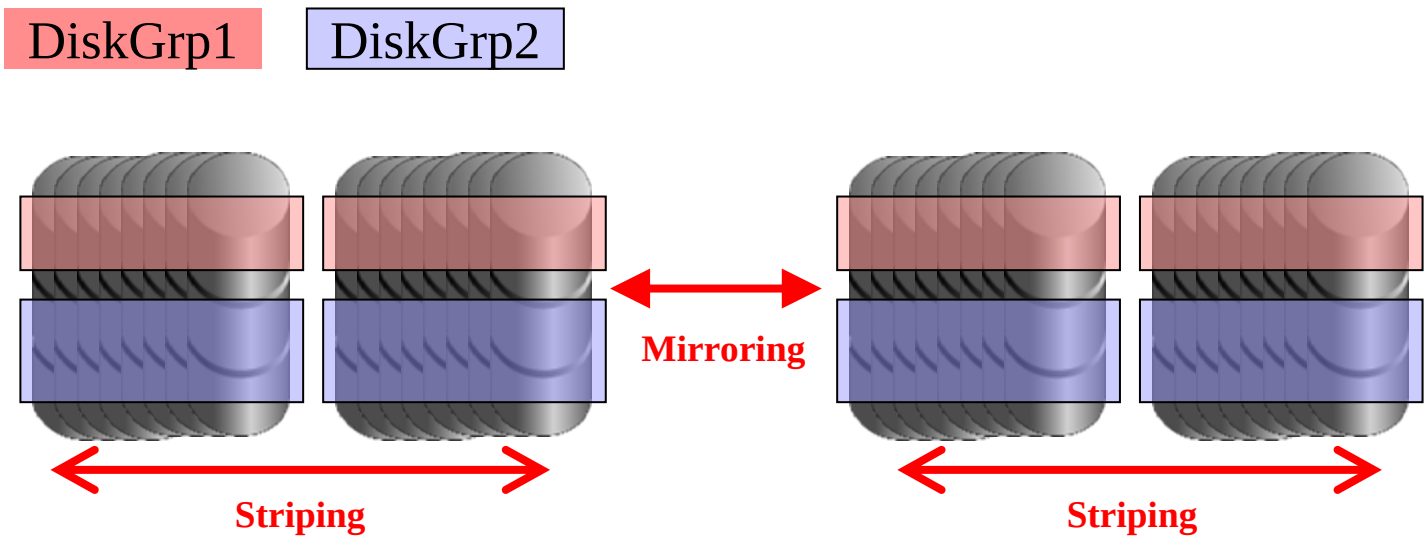
Storage Configuration - SW

- Oracle ASM 10g R2
  - Implements SAME (stripe size tuned for Oracle)
  - Mirroring across storage arrays
  - Comes from Oracle
    - Works with RAC
    - Takes storage configuration and tuning closer to DBAs
- Device name persistency
  - Asmlib
    - Simple way to ‘mark’ disks for persistency
    - Reduces N# of open file descriptors
  - Devlabel (RHEL 3)
- Multipathing with load balancing
  - Qlogic driver for Linux

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Disk Configuration with ASM

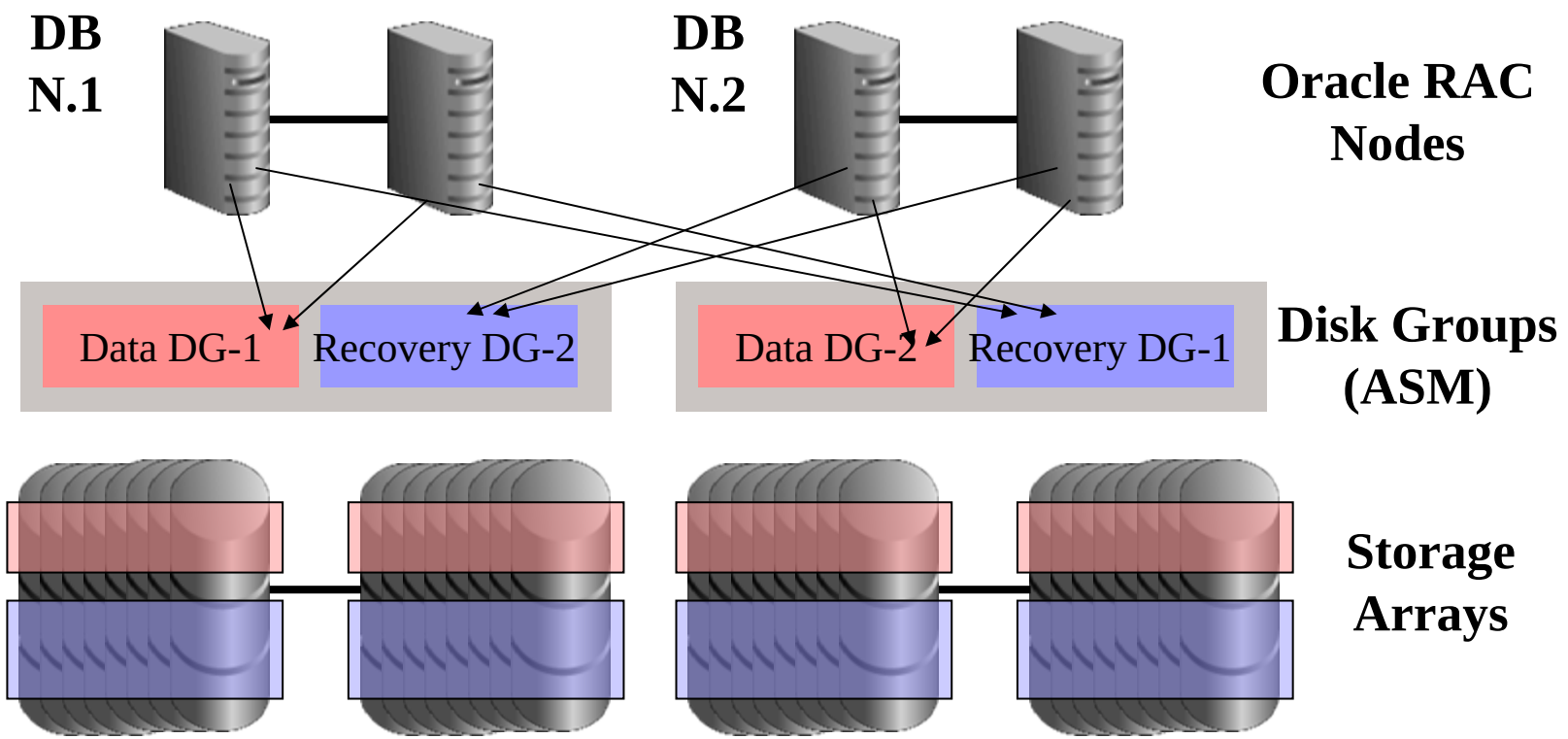
- Follow the ideas of S.A.M.E. as much as possible (J. Loaiza 1999)
- Two diskgroups per DB (data,flash recovery area)
- Destroking
  - Old 'trick of the trade' for performance
  - Even more important for SATA disks (high capacity, low IOPS)
- Example:



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

ASM Configuration

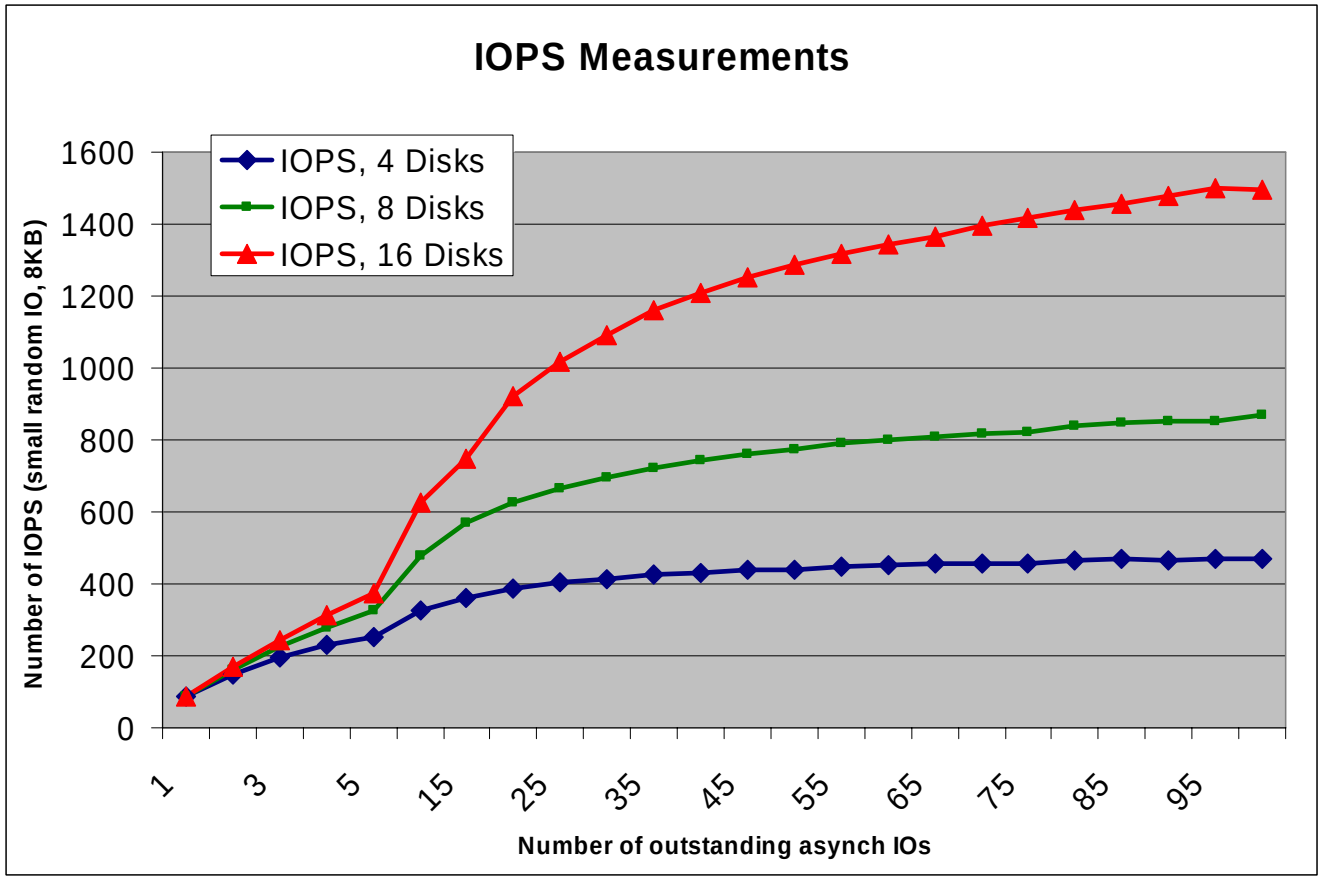
- Coupled configuration:
  - Production (DB1) with low-IO DB (DB2)



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Storage Performance

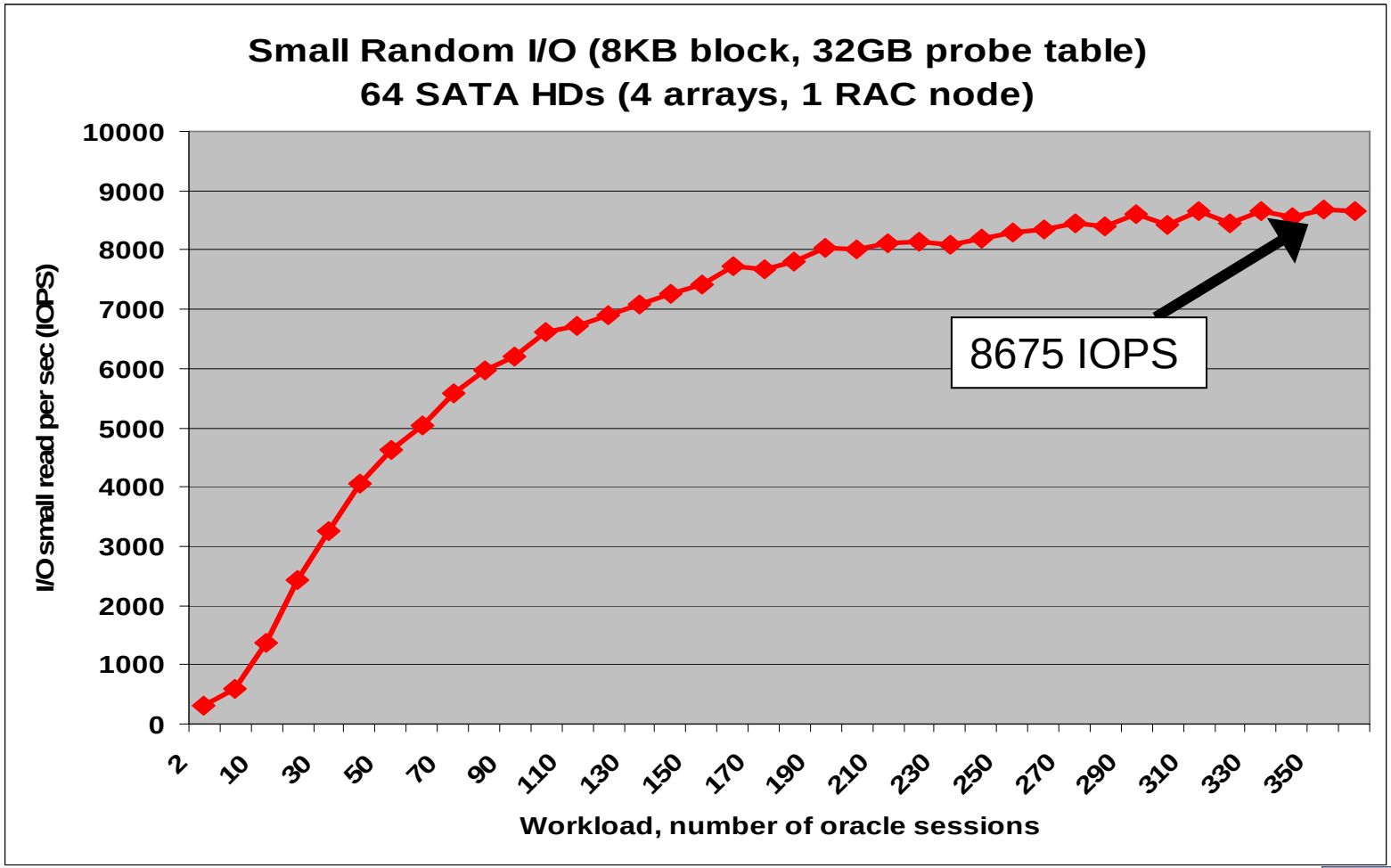
- Orion for performance measurements
  - http://otn.oracle.com/deploy/availability/htdocs/lowcoststorage.html



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

ASM performance

- IOPS and scalability from the DB (synthetic test)



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

ASM – Wrap-up

- Performance
  - ASM and ASMLib scale out, tested with 64 HDs.
    - Small random IO: ~100 IOPS per disk
    - Sequential IO (test 800 MB/s on 4 arrays, FC limit)
  - With ASM: uniform utilization of the HDs (for HA and perf)
  - High capacity/cost
- Administration:
  - We experienced a few stability issues with 10.1, much better in 10.2
  - We deploy a custom ASM config using JBOD (HA and performance in return for the added complexity).
  - 'ASM DBA' can streamline operations, but requires additional storage admin skills
  - Single disk failure rate requires DBA time, so far rate as expected: MTBF ~ 60 years

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

RAC Configuration

- Cost-effective HW
  - Most nodes are dual Xeon with 4GB of RAM
  - ‘mid-range PC’ with dual power supply and HBA
- Linux – RHEL 3, 32 bit
- Gigabit Ethernet
  - 2 private interconnects + public network(s)
- Cluster size
  - 4 nodes for most production
  - Larger clusters for scalability tests
  - Planned upgrade from 4 to 8 nodes

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Economies of Scale

- Homogeneous HW configuration
  - Clusters can be easily built and grown
  - A pool of servers, storage arrays and network devices are used as ‘standard’ building blocks
  - Hardware provisioning and installation is simplified
- Software configuration
  - Same OS and database version on all nodes
    - Ex: Red Hat Linux and Oracle 10g R2
  - Simplifies installation, administration and troubleshooting

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Operations

- Commodity HW simplifies troubleshooting
  - Node replacement and node reinstallation
  - Often offloaded to junior sysadmin
- DBAs have root password
  - Can install Oracle independently
  - Configure Storage at the OS level (asmlib etc)
  - Look at system logs
- FC storage
  - DBAs can logon and perform basic operation on storage arrays and fiber channel controllers

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

10g for Consolidation

- Our model is of an application service provider
- We assign dedicated DB clusters to each customer
- Each application is assigned a service
  - Services are run in load balanced mode if possible (easier to manage in case of failure)
  - Some application don't scale with RAC
    - They are assigned to one preferred node
    - Some RAC nodes are deployed only to provide redundancy in case of failover (available nodes)

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

10g Backup and Recovery

- Technology:
  - RMAN (Oracle's primary solution for HA)
  - Media manager (Tivoli)
- On-disk backups
  - Our disk layout allows for large recovery areas
  - Low overhead with 10g incremental recovery
- Incremental backups
  - With block change tracking we reduce performance impact and tape usage

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

10g Backup Performance

- Measurements from a production DB
  - Mixed workload, mainly read-only
- Full backup
  - Full backup to tape: 3.5 TB in 16 hours
- Incremental backups
  - We use block change tracking
  - Average daily activity for the month August 2006
    - Block_change_tracking file = 420 MB
    - Daily, incremental level 1 (differential): 9 minutes
    - Backup size to tape: 16 GB
    - Corresponding archived redo log size (daily): 58 GB of redo, 230 files

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Distributed Databases

- Streams replication
  - DB changes are captured at source, propagated at destination and then applied
  - Data transfer inside CERN (LAN)
  - Stream to other laboratories around the world (WLAN)
- Experience
  - 10gR2 improved stability and performance
  - Successfully tested functionality
  - Challenging to tune the 3-component architecture
    - Apply process is often the bottleneck on LAN
    - Propagation strongly affected by network latency on WAN
    - Capture is CPU intensive (downstream capture under investigation)

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Conclusions

- Physics Database Services at CERN run production Oracle 10g services
  - Positive experience after 1 year of production
  - 10gR2 RAC and ASM on commodity HW
  - Currently 100 CPUs, 200TB of raw data
  - Ramping up: ongoing deployment of new applications and expansion of the RAC clusters
- Links:
  - **http://www.cern.ch/phydb**
  - **http://www.cern.ch/canali**

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it