IT-DB Physics Services

Planning for LHC start-up

Maria Girone, IT-DB

6 July 2004

http://cern.ch/db/

Introduction

- Summary of Current Situation

- Requirements for LHC and proposed solution

Physics Services – Today

- Replica Location Service (RLS)
  - One WN as AS per VO (6 VO) for production service
  - Two DS, hosting the DB, coupled via hot standby
  - 2-3 spare nodes for failover
  - Other parallel set-up for test
    - One WN (AS) and one DS (DB) for POOL
    - Two WN (AS) and one DS (DB) for IT/GD certification team
- >10 dedicated disk servers (COMPASS, HARP, Cristal II, CMS Tracker, …)
- Sun Cluster (public 2-node for applications related to detector construction, calibration and physics production processing)
- Run Oracle 9iR2 (DB) 9.0.4 (iAS) on RHEL 2.1

Dedicated Nodes (Disk Servers)

- COMPASS: 8 Disk Servers, +2/3 per year
  - Typically only 1-2 DBs actively used – all others idle
  - Good candidate for consolidation into a RAC
- HARP: 2 Disk Servers
- Cristal II: 1 Disk Server
- X5Tracker: 1 Disk Server
- CASTOR nameserver: 1 Disk Server
- PVSS: 1 Disk Server (foreseen)
- Basically everything that doesn't 'fit' in Sun cluster (next…)

Sun Cluster

- 2 x Sun 280Rs (ex-Sundev)
- 360GB of SAN storage
  - Roughly 100GB used for DB files…
- Hosts all physics applications with storage requirements << 100GB
  - Production DBs (e.g. LHCb scanbook, CMS TMDB, …)
  - Detector construction DBs
- Neither CPU capacity nor storage sufficient for LHC exploitation phase
  - Current requests suggest that storage will be insufficient before end 2005!
- **Needs to be replaced 2005 - 2006**

Future Service Needs

- Increasing Service needs from the LCG side
  - Grid File Catalog (high availability)
  - CASTOR name server (high availability)
  - Event-level Metadata
    - Upcoming Conditions DB and POOL RDBMS production services
- Uncertain requirements on
  - storage volume
  - resource split among different services
  - access I/O patterns

Investigating Oracle RAC / Linux

- RAC - Real Application Clusters provides
  - High availability – proven technology used in production at CERN to avoid single points of failure with transparent application failover
  - Consolidation – allows a smallish number (8? 16?) of database services to be consolidated into a single management entity
  - Scalability – SAN-based infrastructure means CPU nodes and storage can be added to meet demand
- RAC (OPS) used on Sun since 1996
  - 2 – 3 node clusters

SAN-based DB infrastructure

Mid-range Linux PCs: dual power supply, mirrored system disk, with dual HBAs multiple GbitE (3)

F/C connected switches: up to 64 ports (Qlogic)

(Infortrend) Storage: 16 x 256GB disks

Short Term (July 2004-Aug 2004)

- Build prototype Linux RAC using existing h/w
  - Farm nodes, Infortrend storage, Brocade/Qlogic switch etc
  - Plan to use Oracle 10g with 9iR2 as fall-back
- Obtain offers for Qlogic Switches and HBAs
- Obtain offer(s) for 'mid-range' PCs (FIO)
- Order pre-production RAC h/w
- Define test cases for all target apps
  - Castor n/s, RLS, COMPASS, …

Medium Term (Sep 2004 – Dec 2004)

- Build 2 RACs using shared SAN infrastructure to handle RLS and CASTOR production in 2005
- Using same h/w, evaluate needs for COMPASS/HARP, Sun cluster replacement, POOL RDBMS b/e
- Understand deployment and manageability impact
  - How many RACs do we need?
- Order additional h/w and plan migration of these services
  - Extra PCs, HBAs, storage as required…

Long(-er) Term (mid-2006?)

- Scalable SAN-based infrastructure for all DB services for Physics
- Can add storage and / or CPU capacity as needed
- Can re-configure resources to meet demand
- Runs Oracle 10g R2?
  - Announced December 2004; available H2 2005?
- Re-use same architecture for other services?
  - AIS; IDB ?

Conclusions

- Today's database services need to evolve to be ready for the LHC era
  - Consolidation, manageability, high availability, scalability
- Investigating RAC as possible solution
  - Linux based F/C Storage seems to be an affordable way to meet requirements
- IT-DB proposes a prototype phase using RAC for RLS and CASTOR for 2005 Data Challenges
  - Plan to participate in the LCG Service Challenges beforehand
  - All other services to follow later
- These services will be prototyped on Oracle 10g

CPU / Disk Needs (Estimates)

# of PCs	Disk Space (TB)	Application	RAC Needs
12	1	Grid File Catalog	H/A
4	1	CASTOR n/s	H/A
10	5	COMPASS/Harp	Scalability
10(?)	10(?)	Event MetaData	Scalability

Above table assumes that existing Sun cluster and future metadata needs can be consolidated into a single RAC

Resources for Grid File Catalog assume DB node / VO, including CERT, TEST, DTEAM and non-LHC VOs such as SIXT

Storage Requirements – Estimates

- Based on COCOTIME and other requests, expect similar data volumes from all LHC experiments for Detector Construction DBs
  - Guestimate 1-2TB total (one storage array?)
- + conditions, production control DBs, POOL / RDBMS b/e etc.
  - Also TB scale? (another storage array?)
- COMPASS / HARP – like usage ('event HDRs') → ~10TB / experiment / year
- Need to understand scalability up to 100TB?