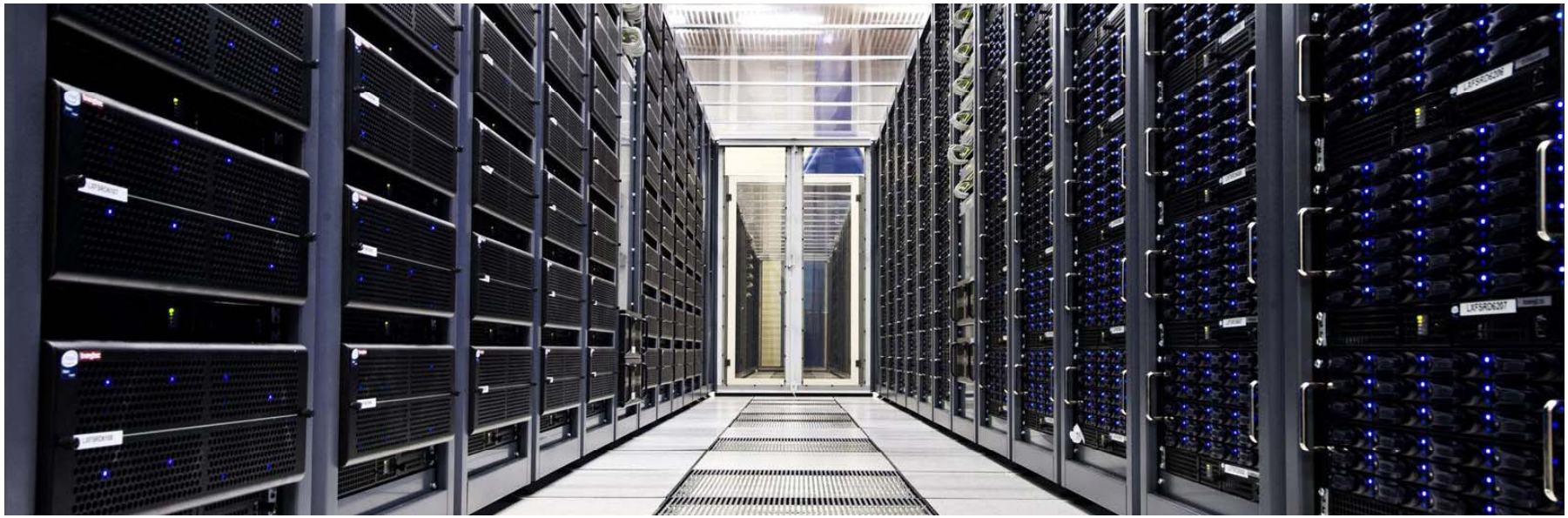
Databases Services at CERN for the Physics Community

Luca Canali, CERN

Orcan Conference, Stockholm, May 2010



Outline

- Overview of CERN and computing for LHC
- Database services at CERN
- DB service architecture
- DB service operations and monitoring
- Service evolution

What is CERN?

- CERN is the world's largest particle physics centre

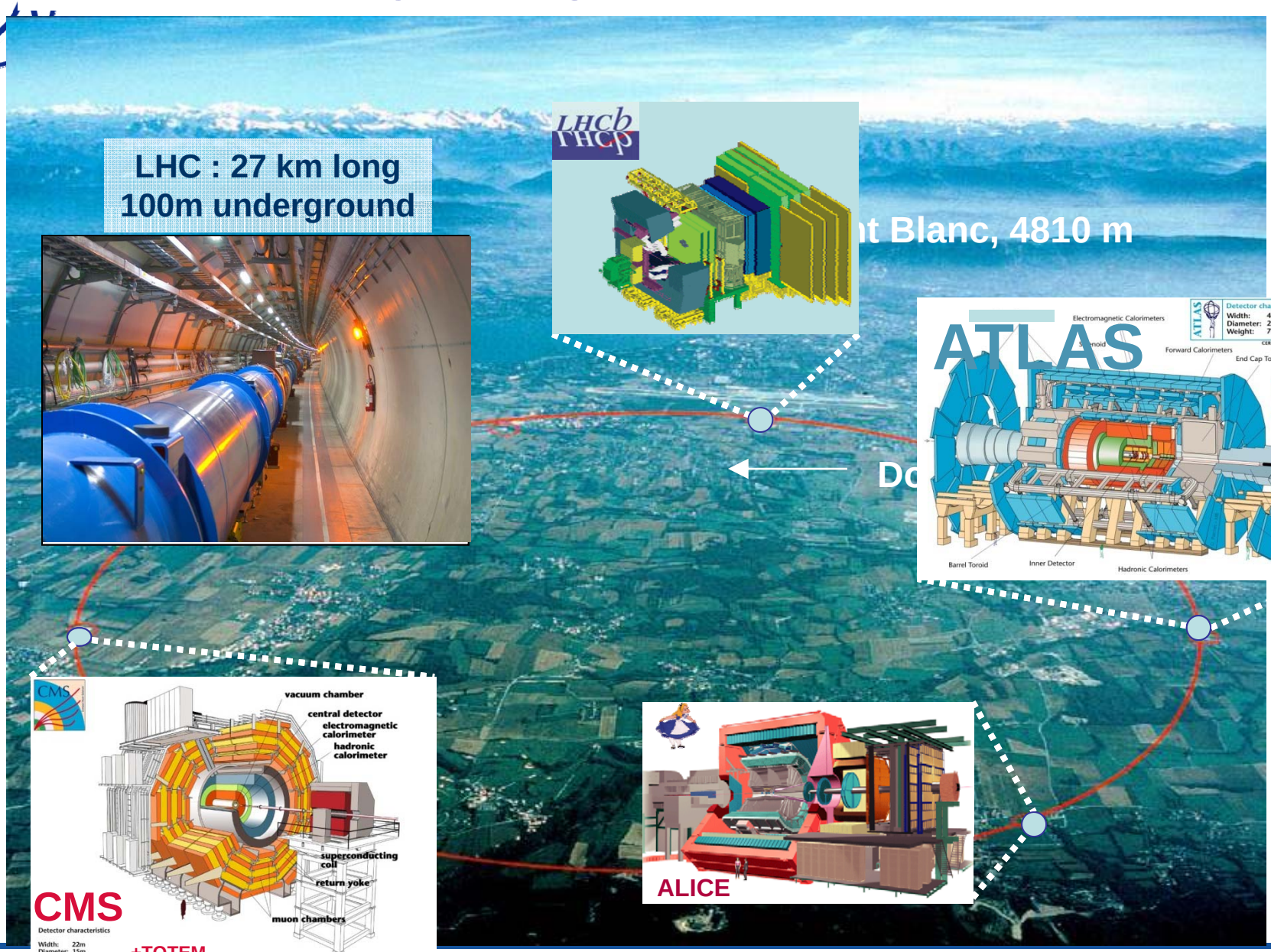
- Particle physics is about:
  - elementary particles and fundamental forces
- Particles physics requires special tools to create and study new particles

  - ACCELERATORS, huge machines able to speed up particles to very high energies before colliding them into other particles
  - DETECTORS, massive instruments which register the particles produced when the accelerated particles collide

CERN is:
-~ 2500 staff scientists (physicists, engineers, …)
- Some 6500 visiting scientists (half of the world's particle physicists)

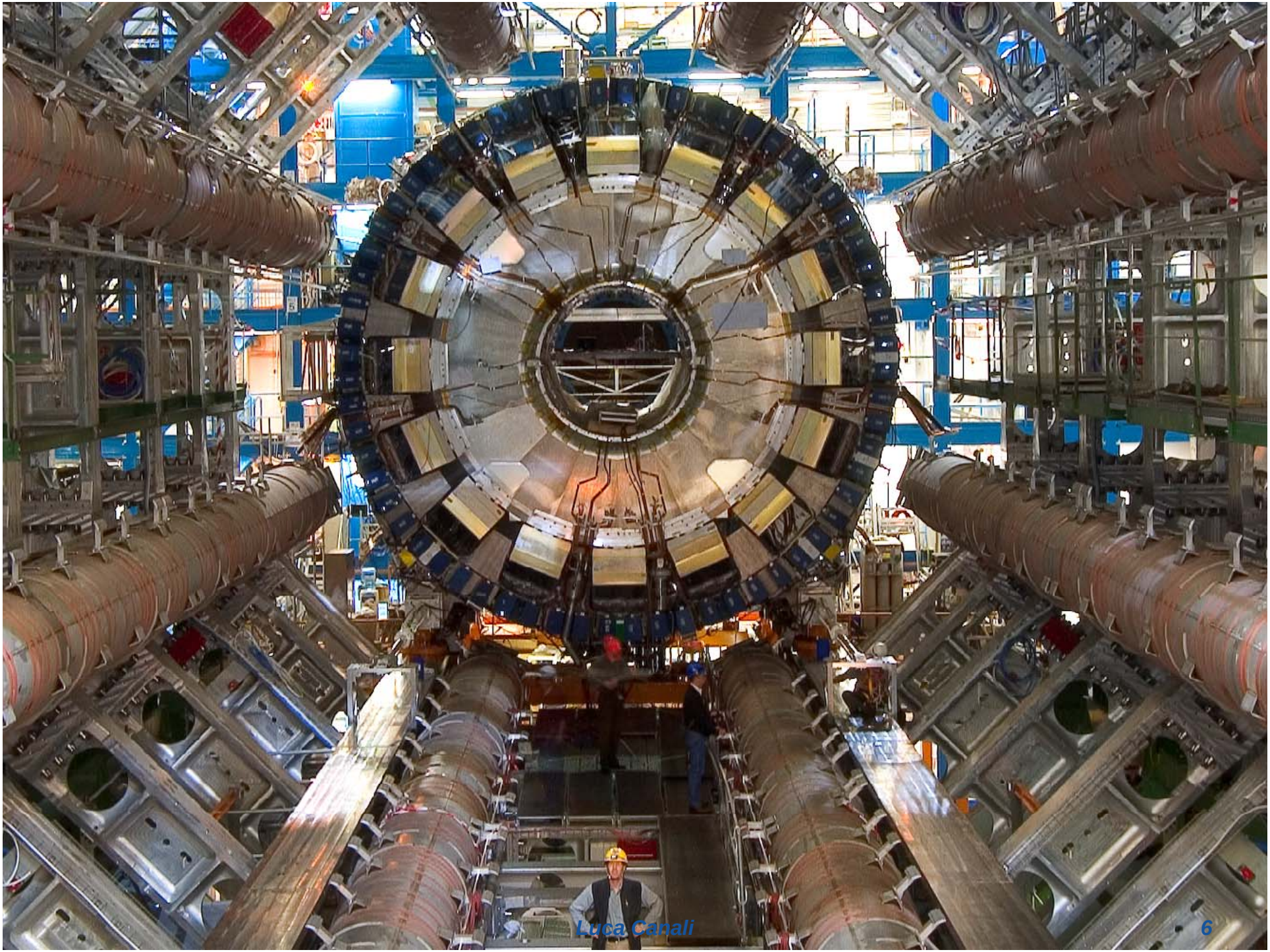
They come from 500 universities representing 80 nationalities.

LHC: a Very Large Scientific Instrument

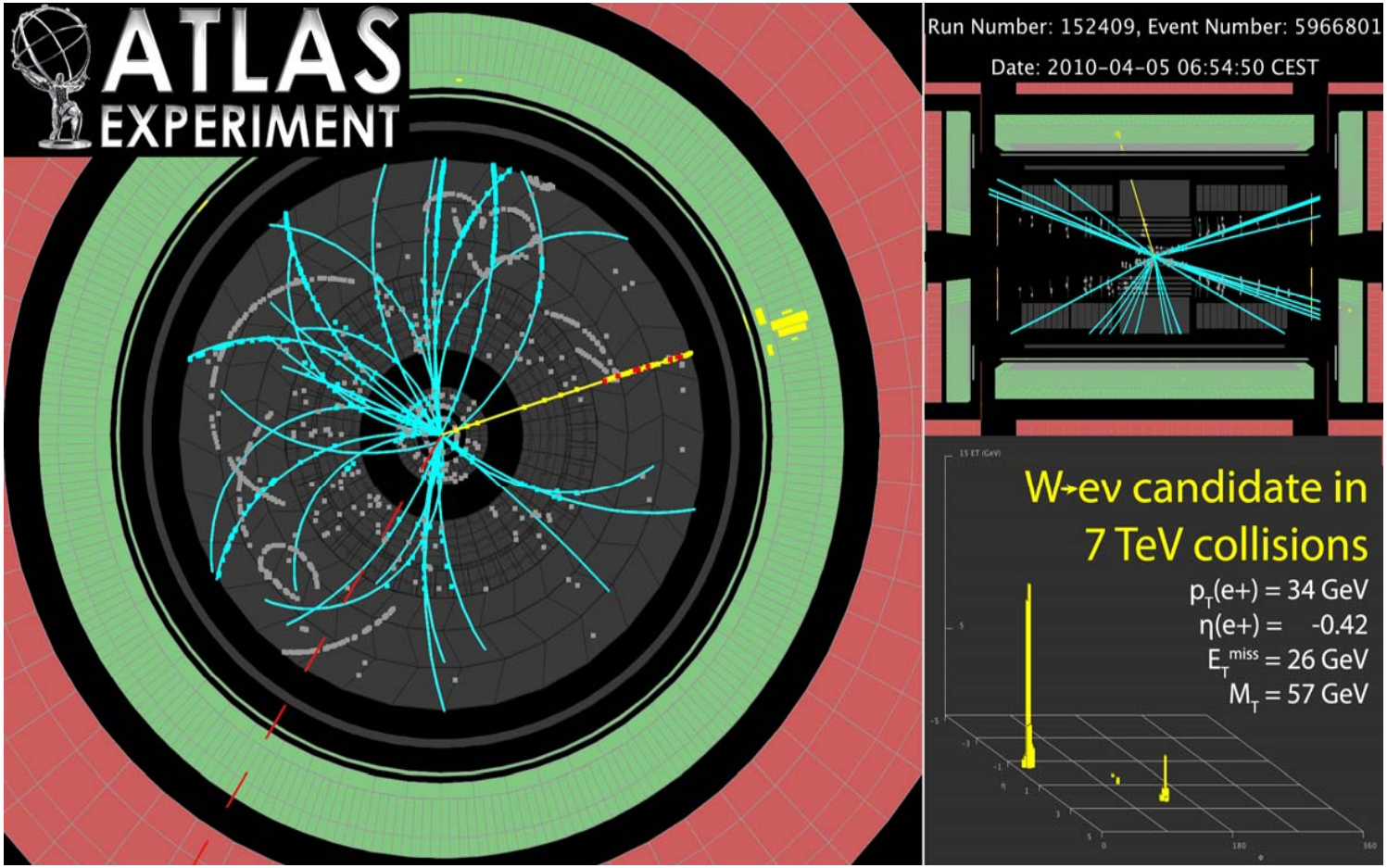


… Based on Advanced Technology

27 km of superconducting magnets
cooled in superfluid helium at 1.9 K

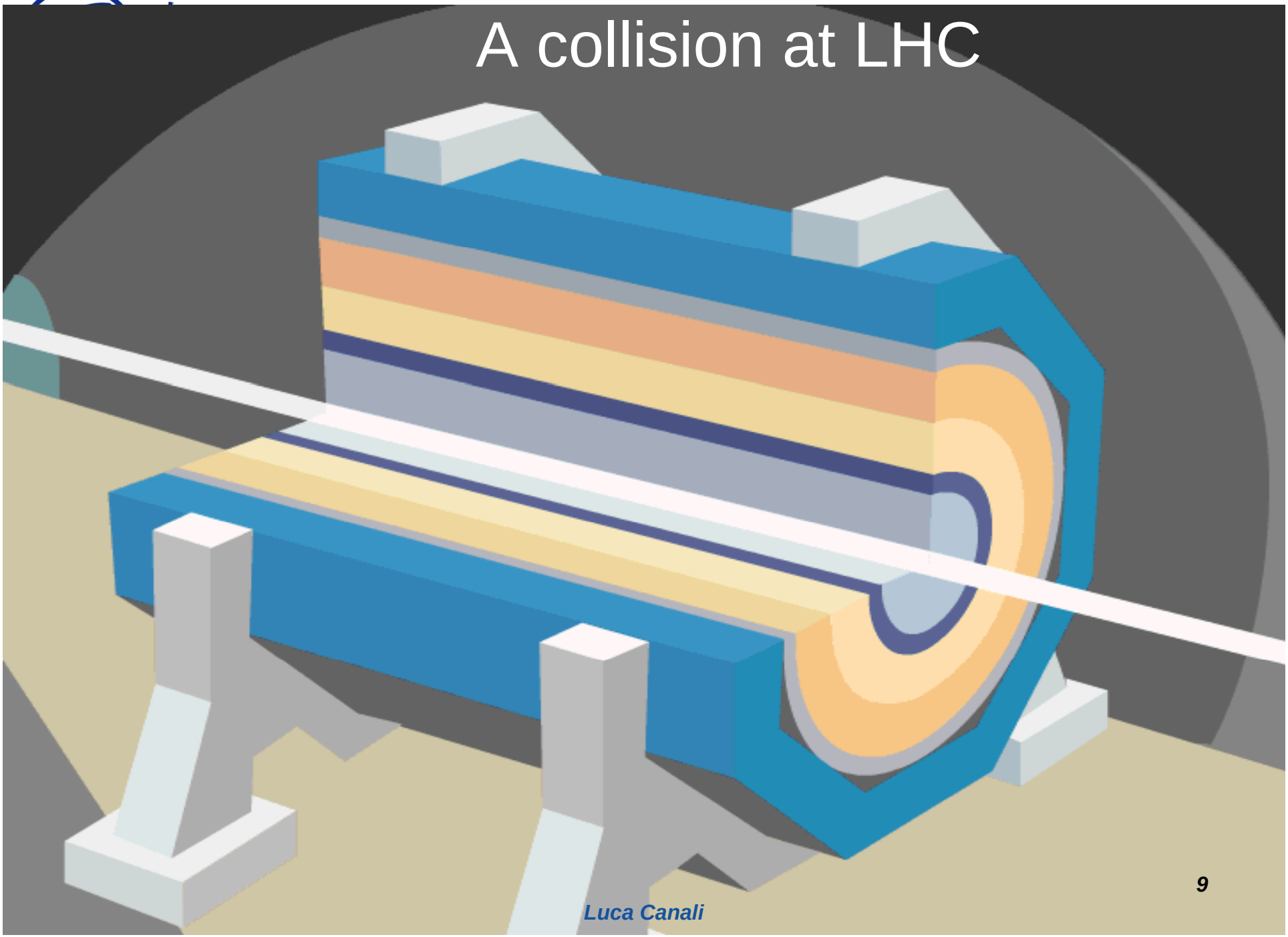


7 TeV Physics with LHC in 2010

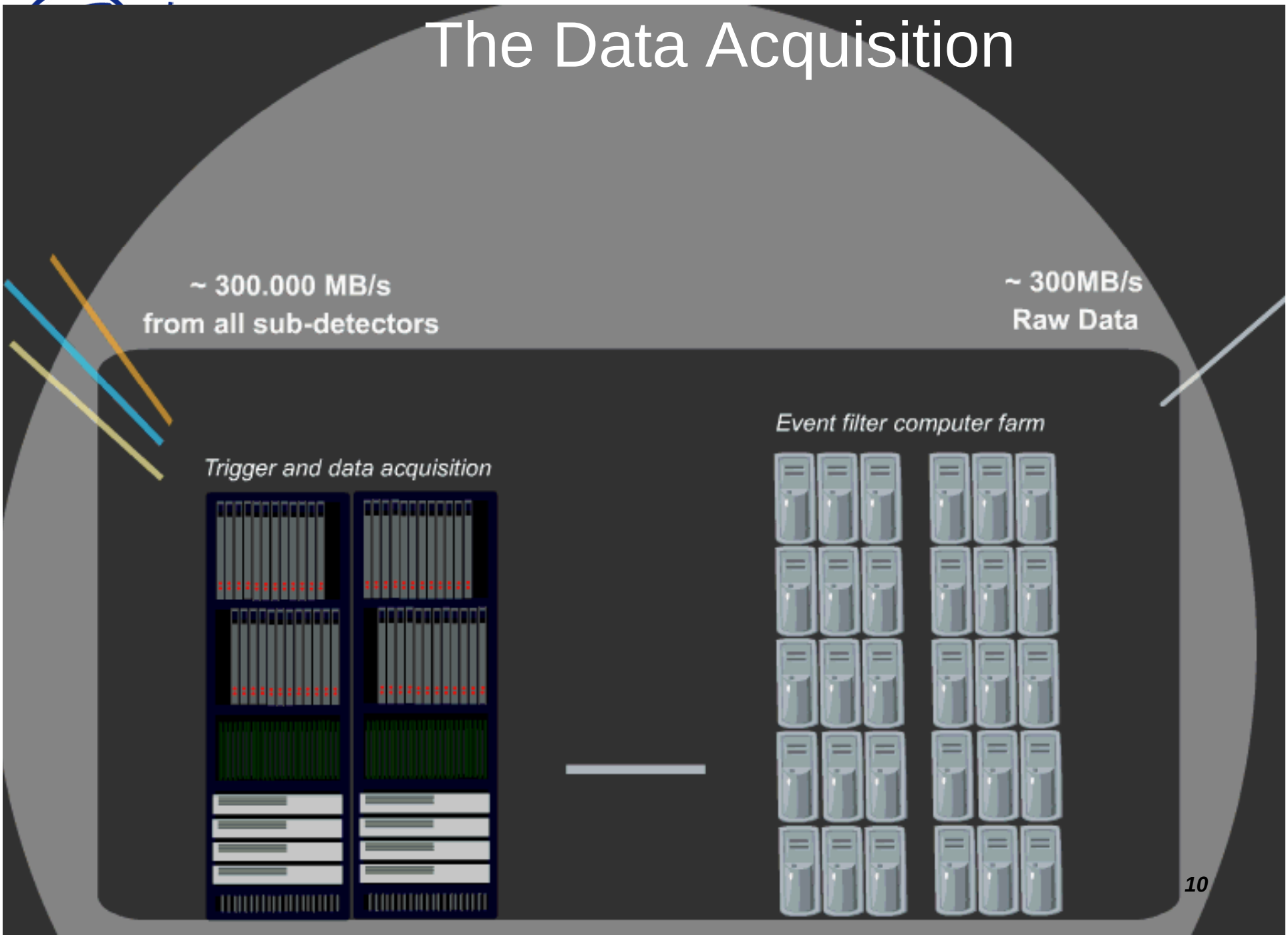


The LHC Computing Grid

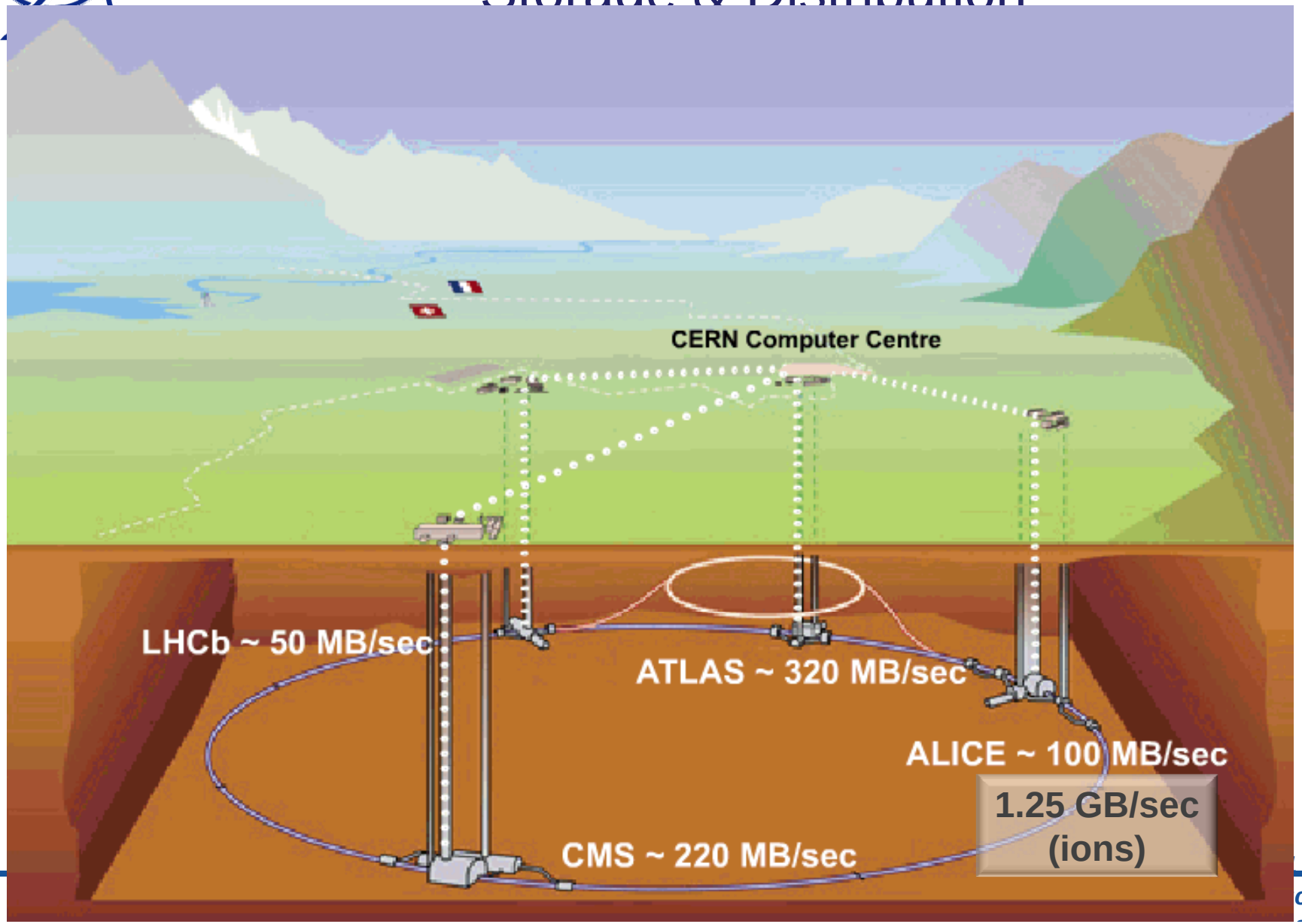
A collision at LHC



The Data Acquisition



Tier 0 at CERN: Acquisition, First pass processing
Storage & Distribution



The LHC Computing Challenge

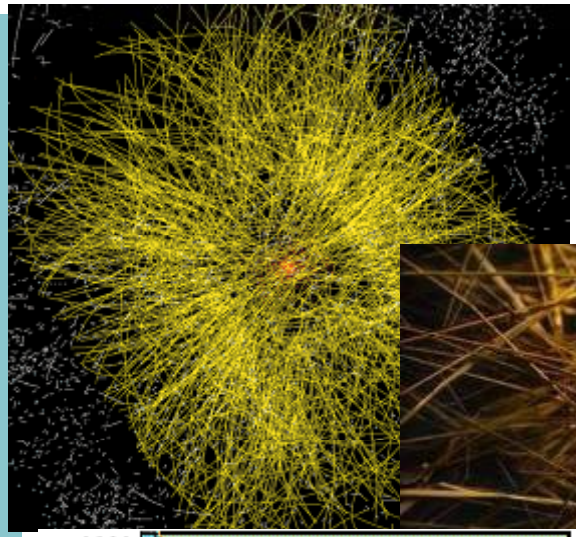
- Signal/Noise: 10^{-9}
- Data volume
  - High rate * large number of channels * 4 experiments

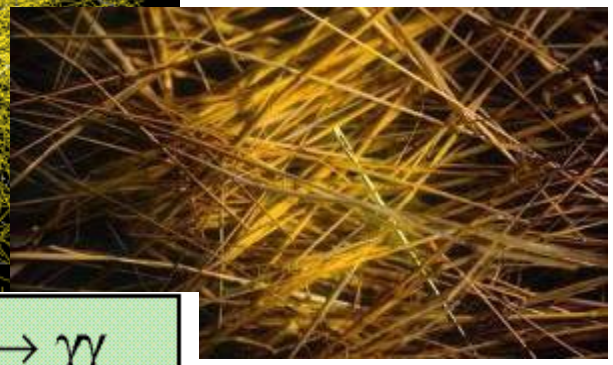
  ➔ **15 PetaBytes of new data each year**
- Compute power
  - Event complexity * Nb. events * thousands users

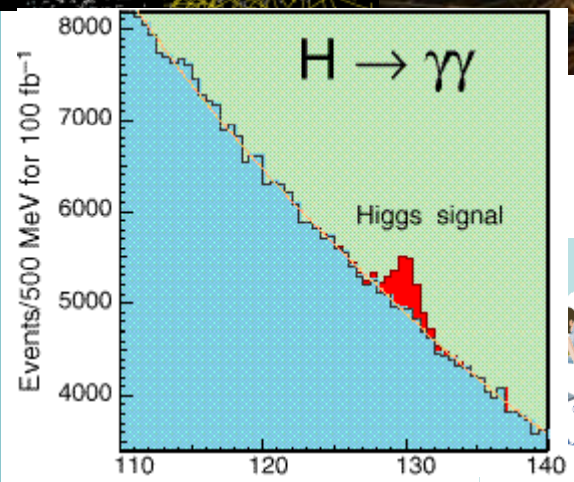
  ➔ **100 k of (today's) fastest CPUs**

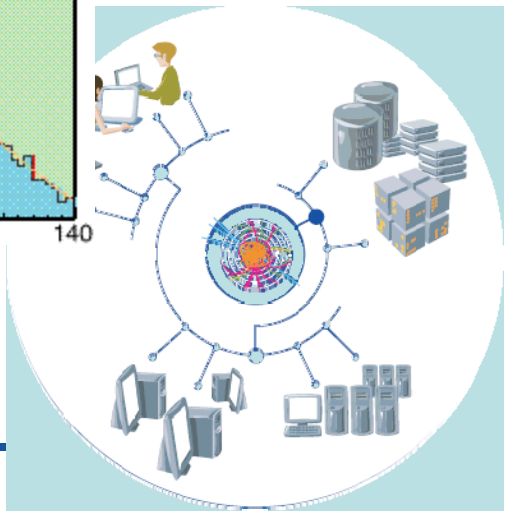
  ➔ **45 PB of disk storage**
- Worldwide analysis & funding
  - Computing funding locally in major regions & countries
  - Efficient analysis everywhere

  ➔ **GRID technology**
- Bulk of data stored in files, a fraction of it in databases (~30TB/year)





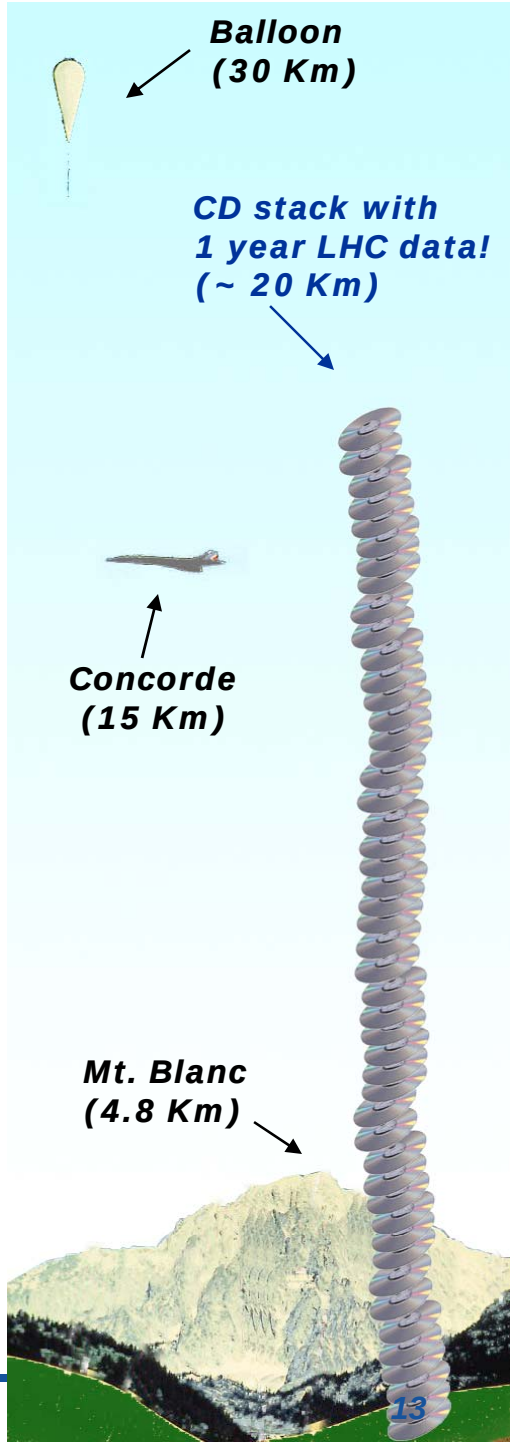




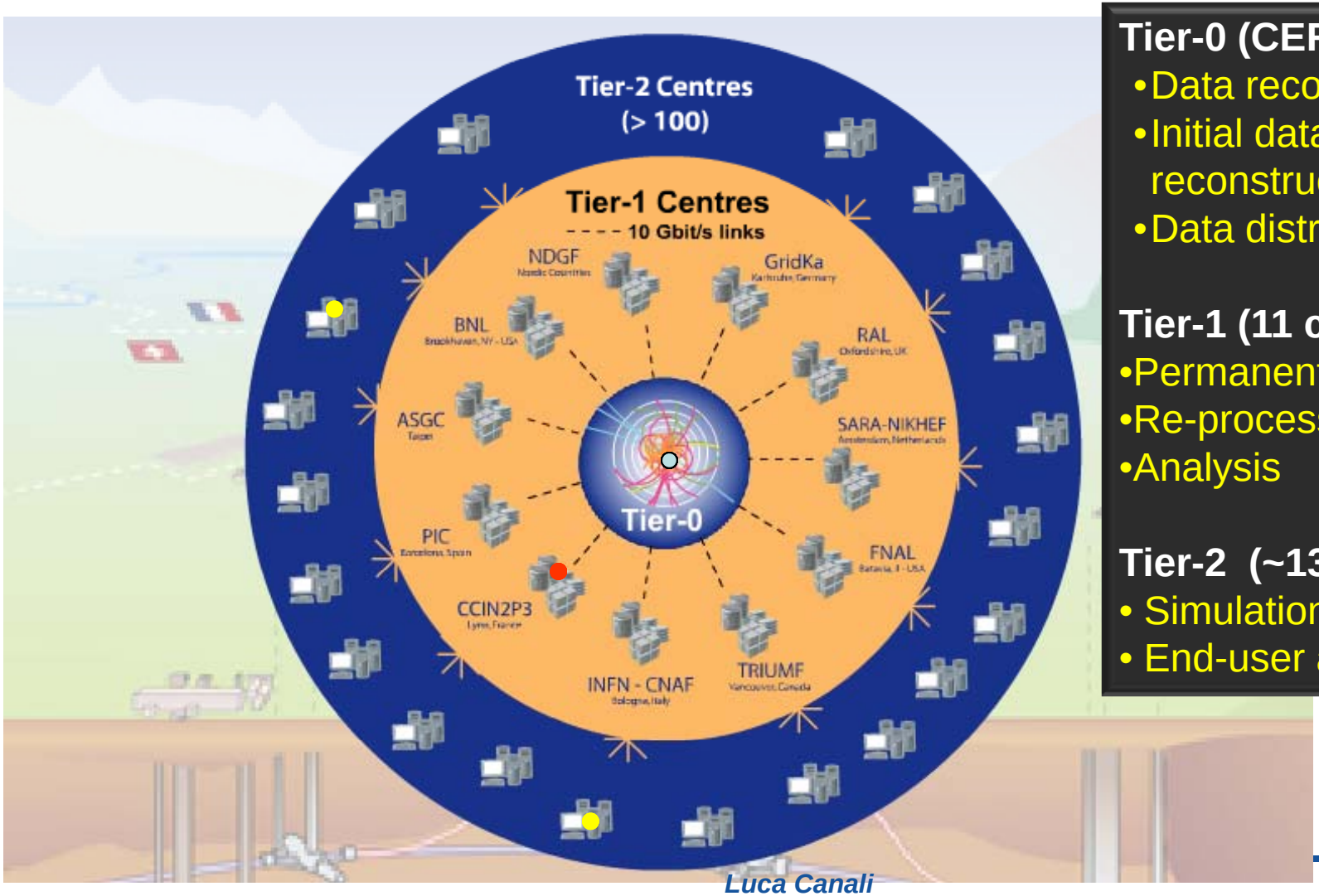
LHC data

LHC data correspond to about 20 million CDs each year!

Where will the experiments store all of these data?



Tier 0 – Tier 1 – Tier 2



Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):
- Simulation
- End-user analysis

Databases and LHC

- Relational DBs play today a key role in the LHC production chains
  - online acquisition, offline production, data (re)processing, data distribution, analysis
    - SCADA, conditions, geometry, alignment, calibration, file bookkeeping, file transfers, etc..
  - Grid Infrastructure and Operation services
    - Monitoring, Dashboards, User-role management, ..
  - Data Management Services
    - File catalogues, file transfers and storage management, …
  - Metadata and transaction processing for custom tape storage system of physics data
  - Accelerator logging and monitoring systems

DB Services and Architecture

CERN Databases in Numbers

- CERN databases services – global numbers
  - Global users community of several thousand users
  - ~ 100 Oracle RAC database clusters (2 – 6 nodes)
  - Currently over 3300 disk spindles providing more than 1PB raw disk space (NAS and SAN)
- Some notable DBs at CERN
  - Experiment databases – 13 production databases
    - Currently between 1 and 9 TB in size
    - Expected growth between 1 and 19 TB / year
  - LHC accelerator logging database (ACCLOG) – ~30 TB
    - Expected growth up to 30 TB / year
  - ... Several more DBs on the range 1-2 TB

Service Key Requirements

- Data Availability, Scalability, Performance and Manageability
  - Oracle RAC on Linux: building-block architecture for CERN and Tier1 sites
- Data Distribution
  - Oracle Streams: for sharing information between databases at CERN and 10 Tier1 sites
- Data Protection
  - Oracle RMAN on TSM for backups
  - Oracle Data Guard: for additional protection against failures (data corruption, disaster recoveries,...)

Hardware architecture

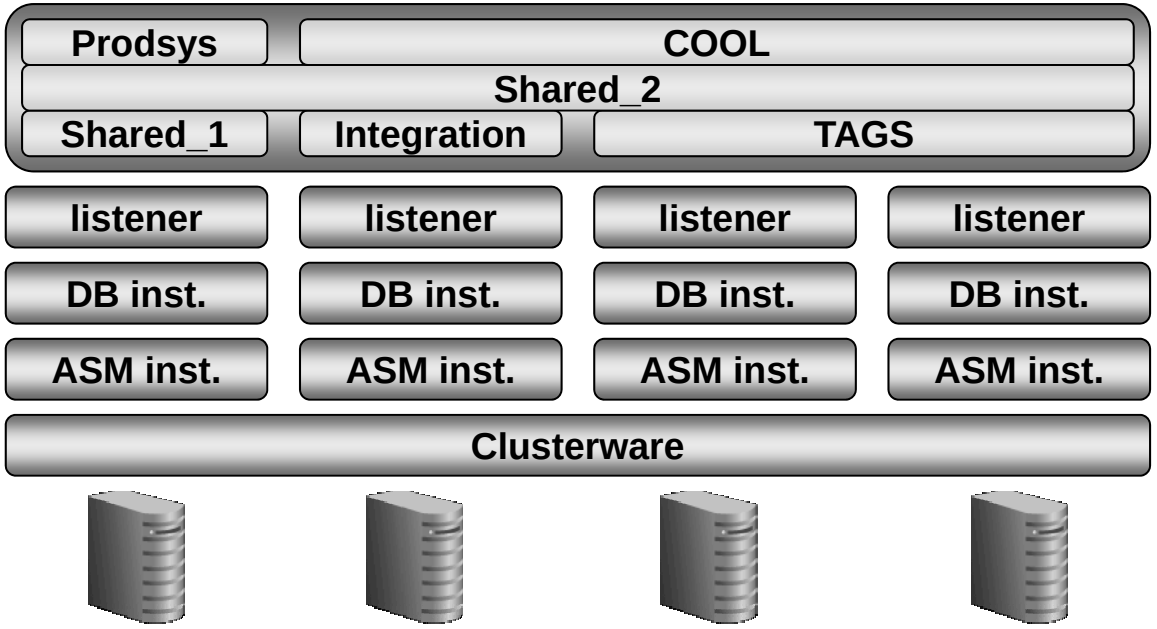
- Servers
  - "Commodity" hardware (Intel Harpertown and Nahalem based mid-range servers) running 64-bit Linux
  - Rack mounted boxes and blade servers
- Storage
  - Different storage types used:
    - NAS (Network-attached Storage) – 1Gb Ethernet
    - SAN (Storage Area Network) – 4Gb FC
  - Different disk drive types:
    - high capacity SATA (up to 2TB)
    - high performance SATA
    - high performance FC

High Availability

- Resiliency from HW failures
  - Using commodity HW
  - Redundancies with software
- Intra-node redundancy
  - Redundant IP network paths (Linux bonding)
  - Redundant Fiber Channel paths to storage
    - OS configuration with Linux's device mapper
- Cluster redundancy: Oracle RAC + ASM
- Monitoring: custom monitoring and alarms to on-call DBAs
- Service Continuity: Physical Standby (Dataguard)
- Recovery operations: on-disk backup and tape backup

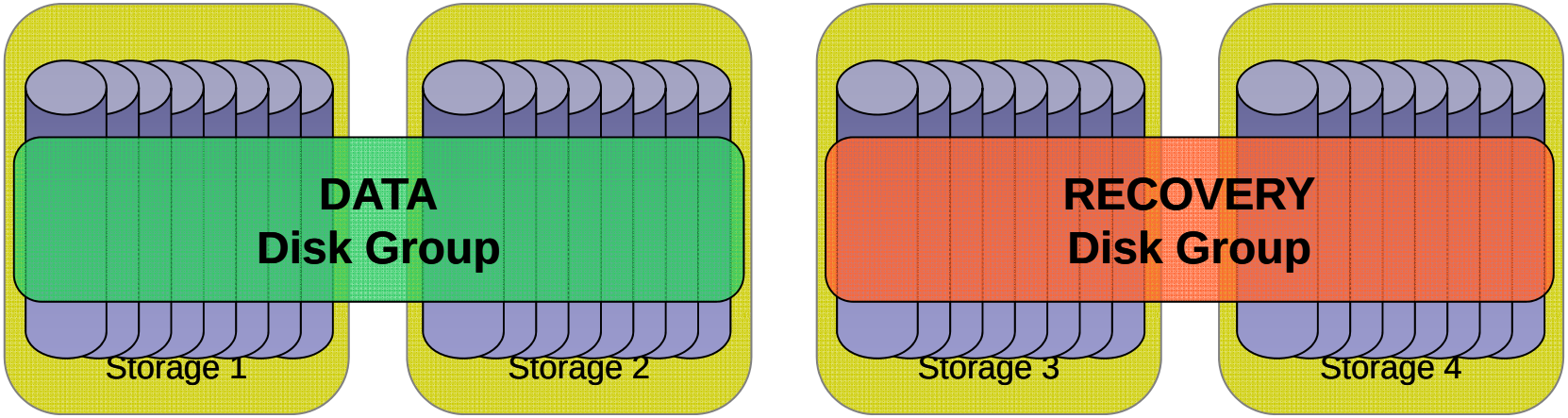
DB clusters with RAC

- Applications are consolidated on large clusters per customer (e.g. experiment)
- Load balancing and growth:leverages Oracle services
- HA: cluster survives node failures
- Maintenance: allows scheduled rolling interventions



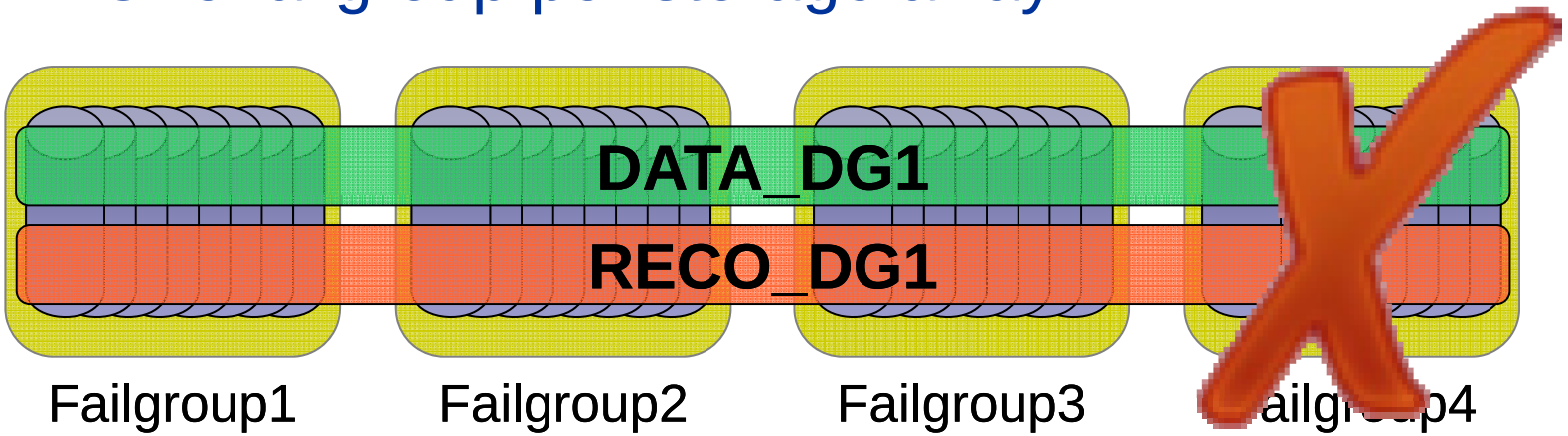
Oracle's ASM

- ASM (Automatic Storage Management)
  - Cost: Oracle's cluster file system and volume manager for Oracle databases
  - HA: online storage reorganization/addition
  - Performance: stripe and mirroring everything
  - Commodity HW: Physics DBs at CERN use ASM normal redundancy (similar to RAID 1+0 across multiple disks and storage arrays)



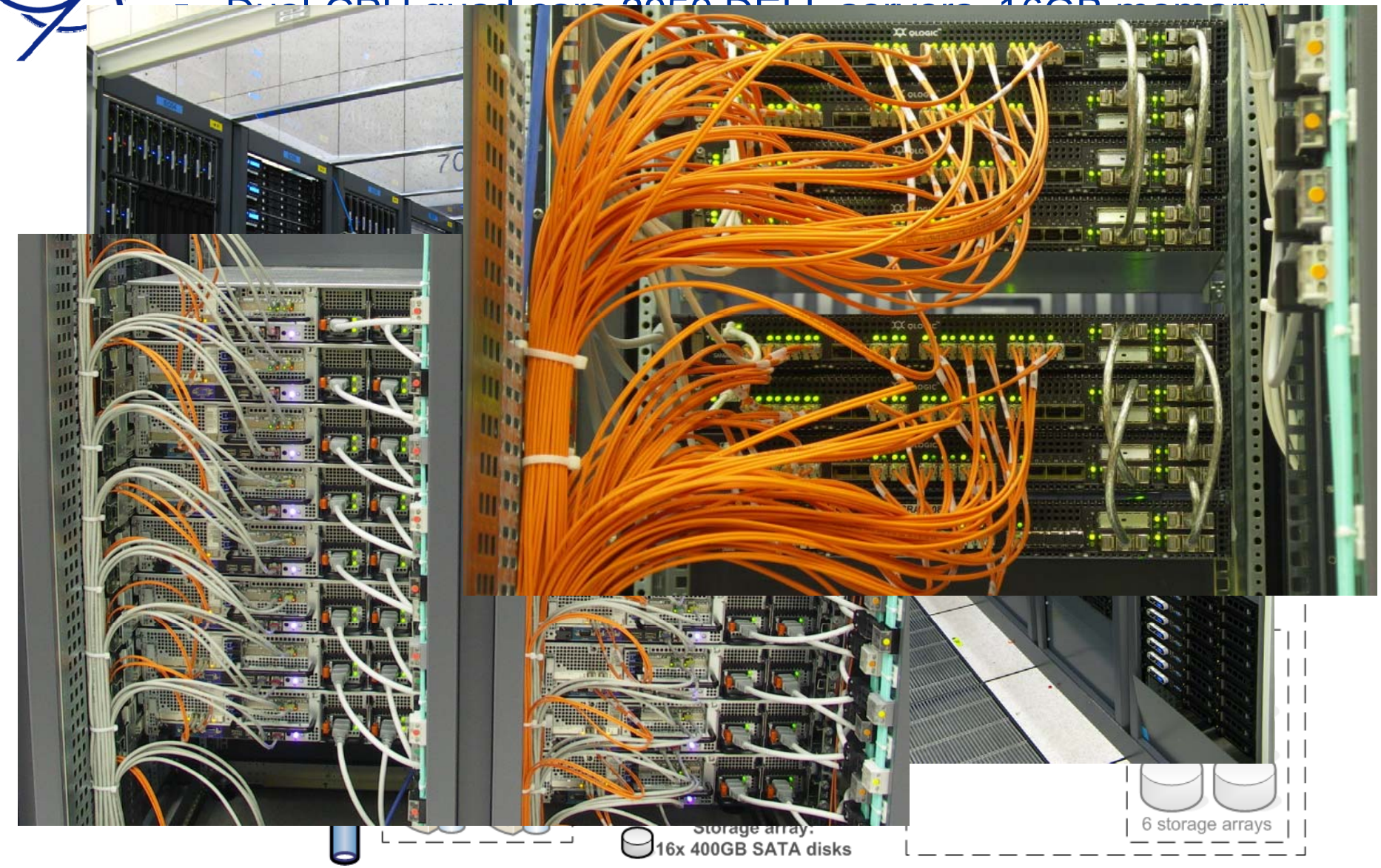
Storage deployment

- Two diskgroups created for each cluster
  - DATA – **data files** and **online redo logs** – outer part of the disks
  - RECO – **flash recovery area destination** – **archived redo logs** and **on disk backups** – inner part of the disks
- One failgroup per storage array



Physics DB HW, a typical setup

- Dual CPU quad core 2950 DELL servers, 16GB memory



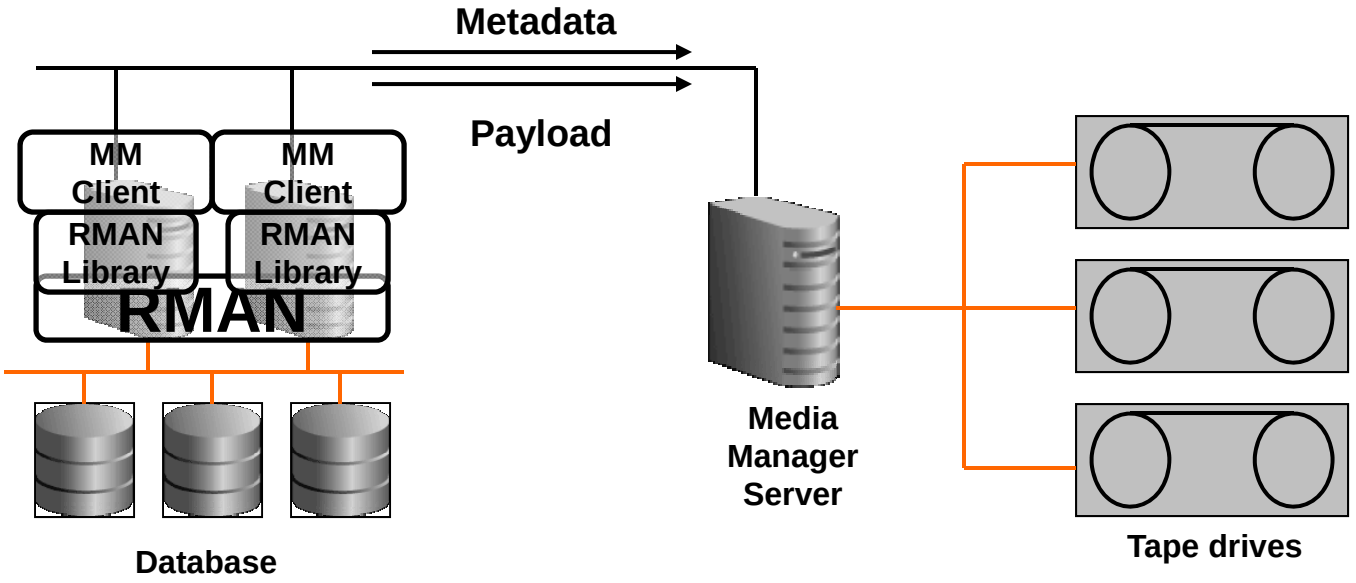
ASM scalability test results

- Big Oracle 10g RAC cluster built with mid-range 14 servers
- 26 storage arrays connected to all servers and big ASM diskgroup created (>150TB of raw storage)
- Data warehouse like workload (parallelized query on all test servers)
  - Measured sequential I/O
    - Read: 6 GB/s
    - Read-Write: 3+3 GB/s
  - Measured 8 KB random I/O
    - Read: 40 000 IOPS

- Results – “commodity” hardware can scale on Oracle RAC

Tape backups

- Main ‘safety net’ against failures
- Despite the associated cost they have many advantages:
  - Tapes can be easily taken offsite
  - Backups once properly stored on tapes are quite reliable
  - If configured properly can be very fast



Oracle backups

- Oracle RMAN (Recovery Manager)
  - Integrated backup and recovery solution
  - Backups to tape (over LAN)
    - The fundamental way of protecting databases against failures
    - Downside – takes days to backup/restore multi TB databases
  - Backups to disk (RMAN)
    - Daily updates of the copy using incremental backups
    - On disk copy kept at least one day behind - can be used to address logical corruptions
    - Very fast recovery when primary storage is corrupted
      - Switch to image copy or recover from copy
    - Note: this is a ‘cheap’ alternative/complement to a standby DB

Tape B&R strategy

- Incremental backup strategy example:
  - Full backups every two weeks

backup force tag ‘full_backup_tag’ incremental level 0 check logical database plus archivelog;

  - Incremental cumulative every 3 days

backup force tag ‘incr_backup_tag' incremental level 1 cumulative for recover of tag ‘last_full_backup_tag' database plus archivelog;

  - Daily incremental differential backups

backup force tag ‘incr_backup_tag' incremental level 1 for recover of tag ‘last_full_backup_tag' database plus archivelog;

  - Hourly archivelog backups

backup tag ‘archivelog_backup_tag' archivelog all;

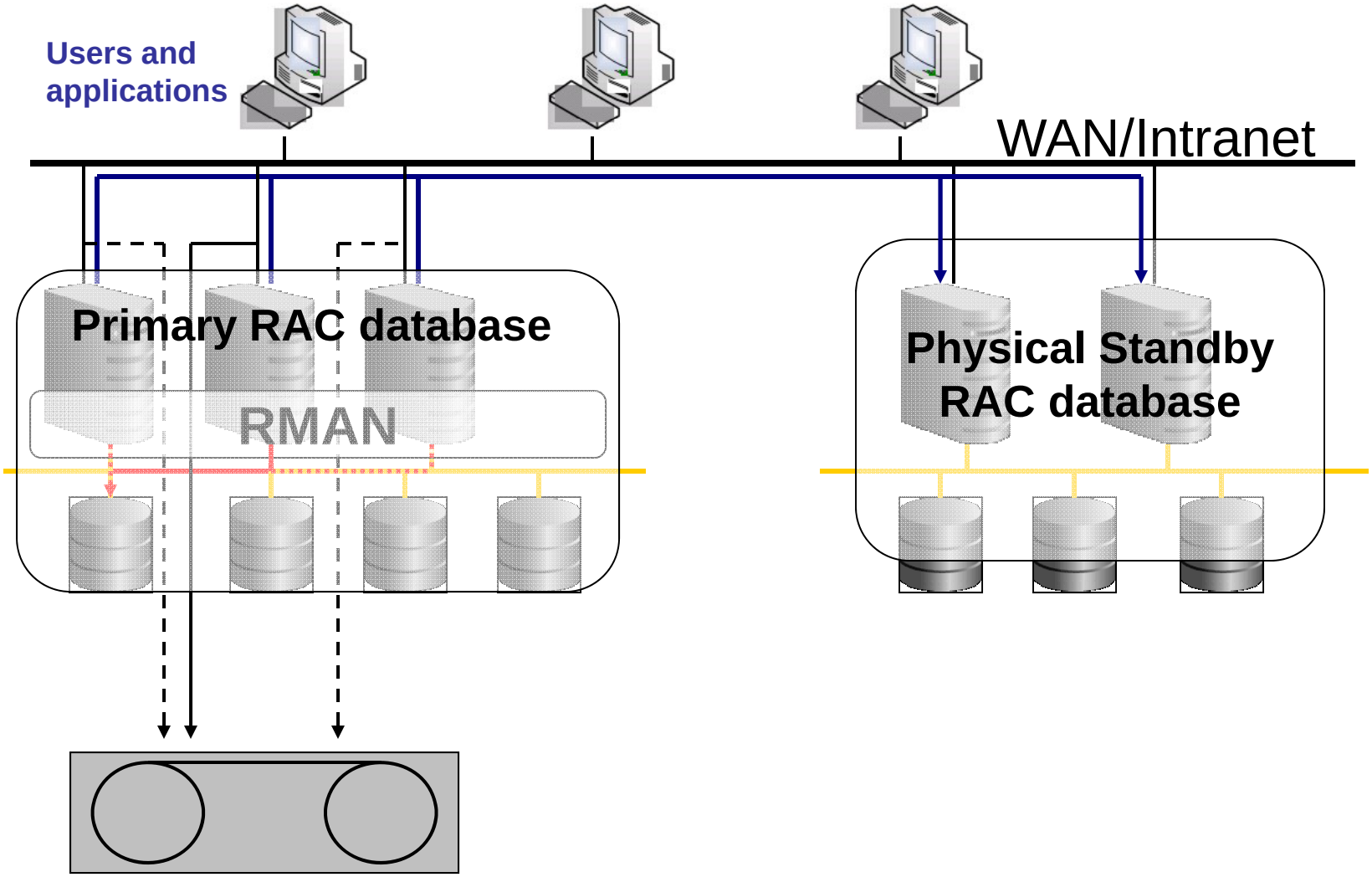
  - Monthly automatic test restore

Backup & Recovery

- On-tape backups: fundamental for protecting data, but recoveries run at ~100MB/s (~30 hours to restore datafiles of a DB of 10TB)
  - Very painful for an experiment in data-taking

- Put in place on-disk image copies of the DBs: able to recover to any point in time of the last 48 hours activities
  - Recovery time independent of DB size

CERN implementation of MAA

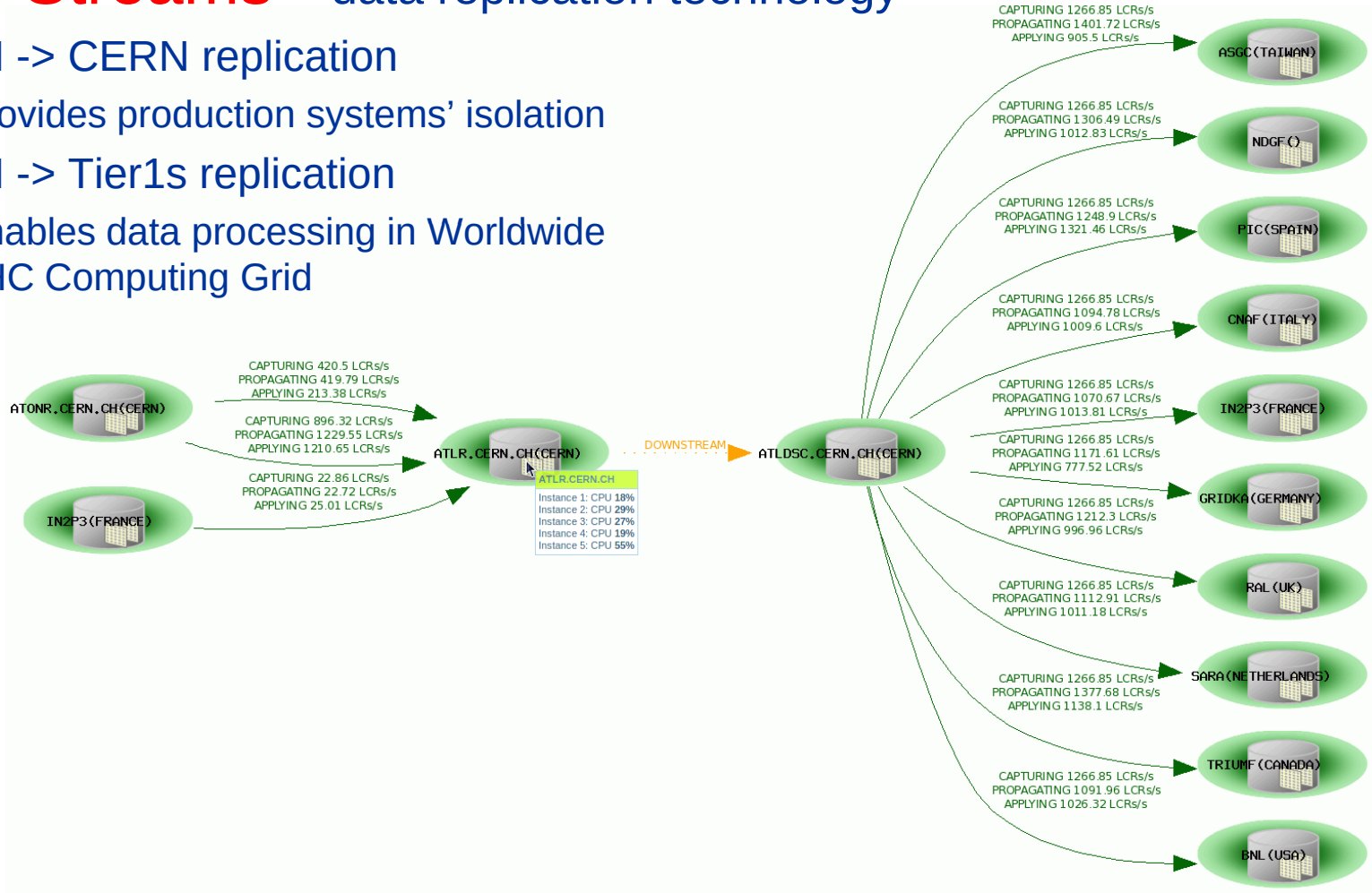


Service Continuity

- Dataguard
  - Based on proven physical standby technology
  - Protects from corruption of critical production DBs (disaster recovery)
  - Standby DB apply delayed 24h (protection from logical corruption)
- Other uses of standby DBs
  - Standby DBs can be temporarily activated for testing
    - Oracle flashback allows simple re-instantiation of standby after test
  - Standby DB copies used to minimize time for major changes
    - Standby allows to create and keep up-to-date a mirror copy of production
    - HW migrations
      - Physical standby provides a fall-back solution after migration
    - Release upgrade
      - Physical standby broken after intervention

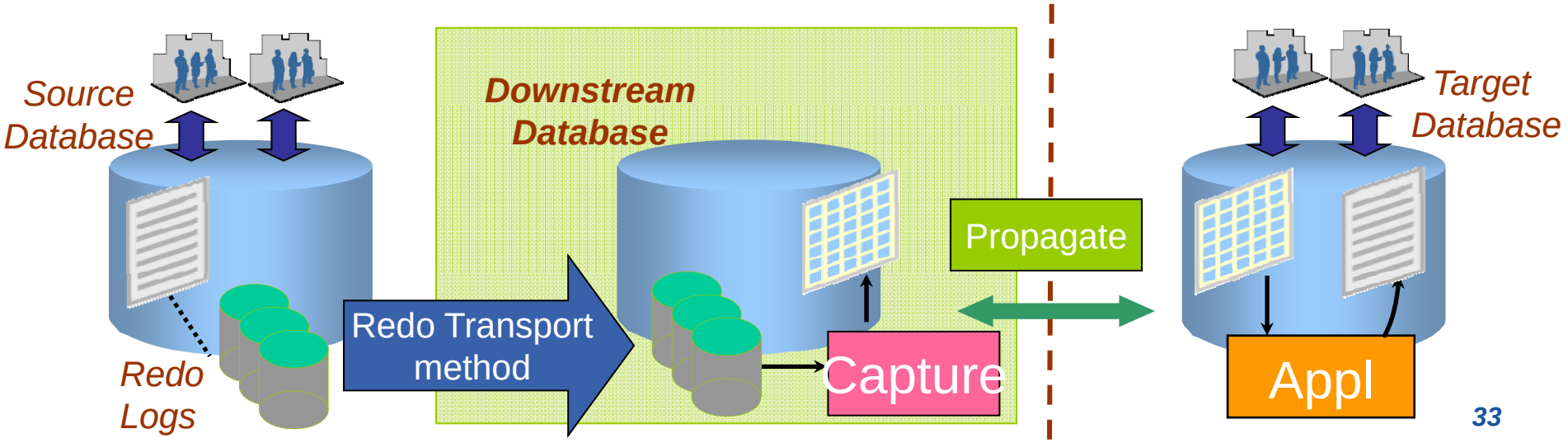
Software Technologies – replication

- Oracle Streams – data replication technology
  - CERN -> CERN replication
    - Provides production systems' isolation
  - CERN -> Tier1s replication
    - Enables data processing in Worldwide LHC Computing Grid



Downstream Capture

- Downstream capture to de-couple Tier 0 production databases from destination or network problems
  - source database availability is highest priority
- Optimizing redo log retention on downstream database to allow for sufficient re-synchronisation window
  - we use 5 days retention to avoid tape access
- Dump fresh copy of dictionary to redo periodically
- 10.2 Streams recommendations (metalink note 418755)



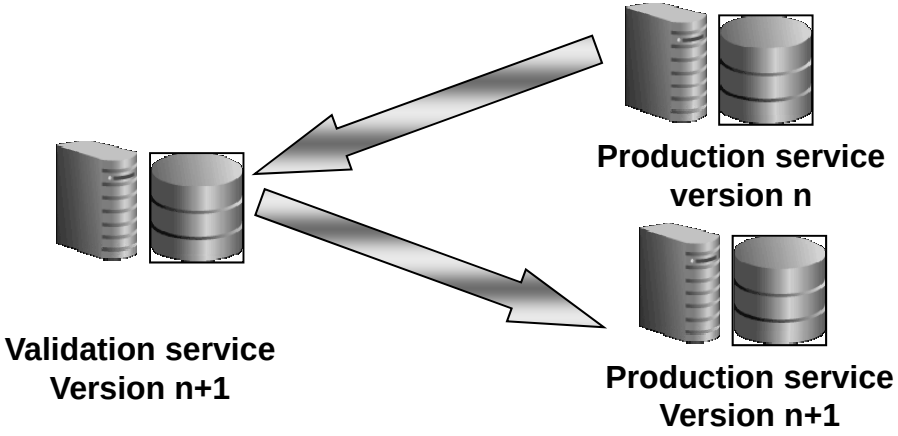
Monitoring and Operations

Application Deployment Policy

Policies for hardware, DB versions, applications testing

- Application release cycle

- Database software release cycle



Patching and Upgrades

- Databases are used by a world-wide community: arranging for scheduled interventions (s/w and h/w upgrades) requires quite some effort
  - Services need to be operational 24x7

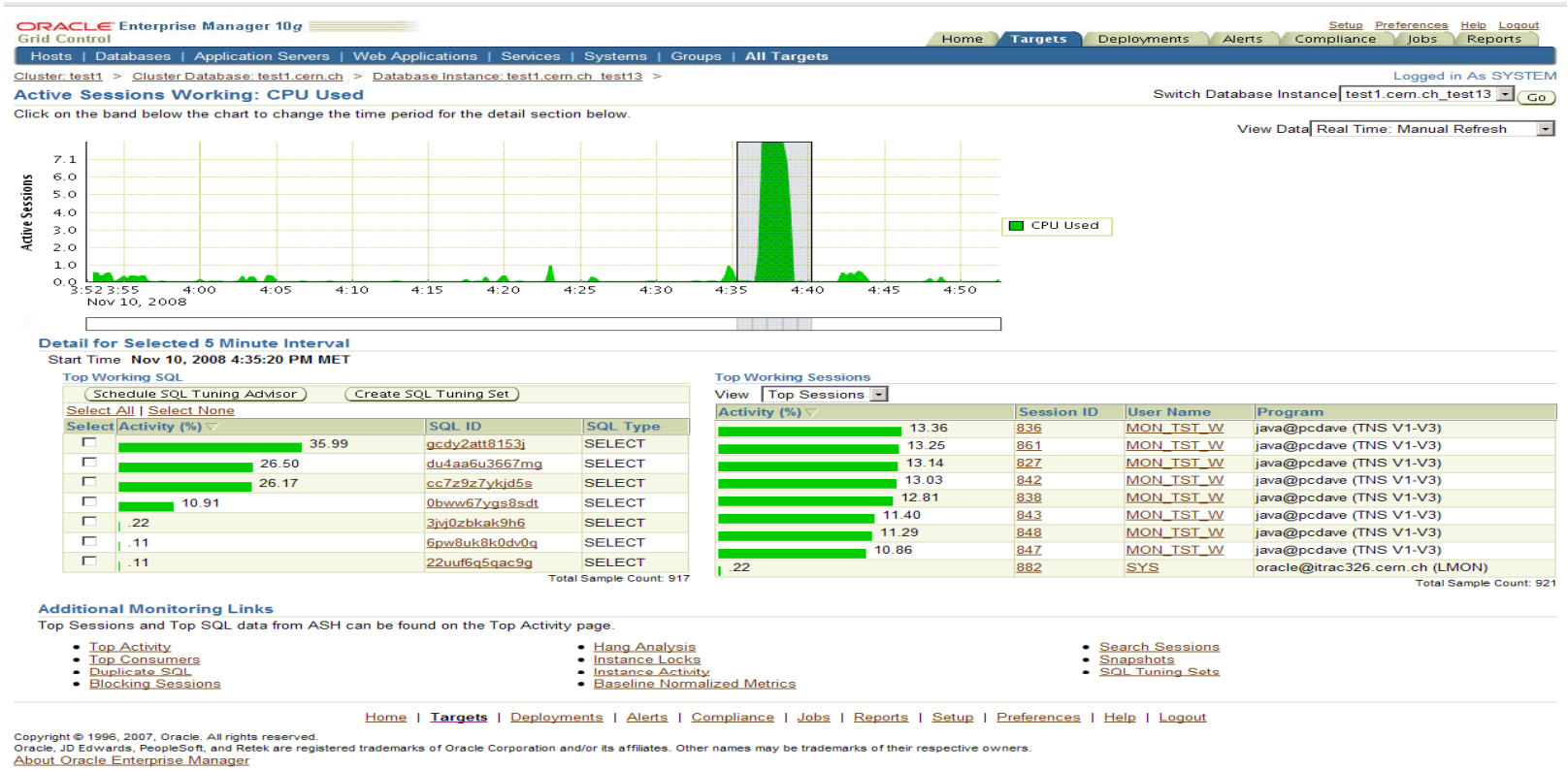
- Minimize service downtime with rolling upgrades and use of stand-by databases
  - **0.04% services unavailability = 3.5 hours/year**
  - **0.12% server unavailability = 9.5 hours/year (Patch deployment, hardware)**

DB Services Monitoring

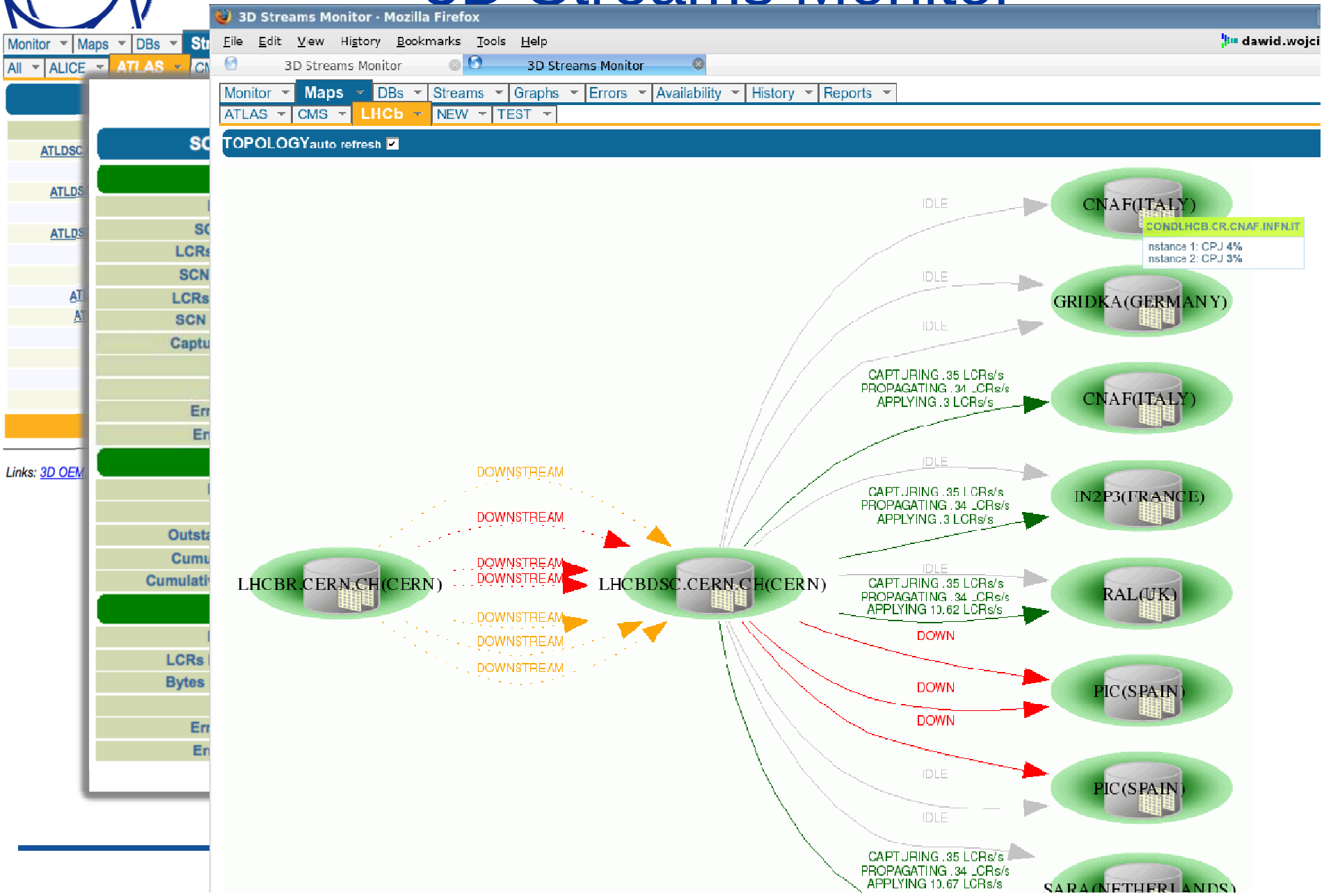
- Grid control extensively used for performance tuning
  - By DBAs and application ‘power users’
- Custom applications
  - Measure of service availability
    - Integrated to email and SMS to on-call
  - Streams monitoring
  - Backup job scheduling and monitoring
  - ASM and storage failures monitoring
  - Other ad-hoc alarms created and activated when needed
    - For example if a repeated bug hits production and need several parameters need to be checked as a work-around
  - Weekly report on the performance and capacity used in production DB is sent to ‘application owners’

Oracle EM and Performance Troubleshooting

- Our experience: simplify tasks and leads to correct methodology for most tuning tasks:

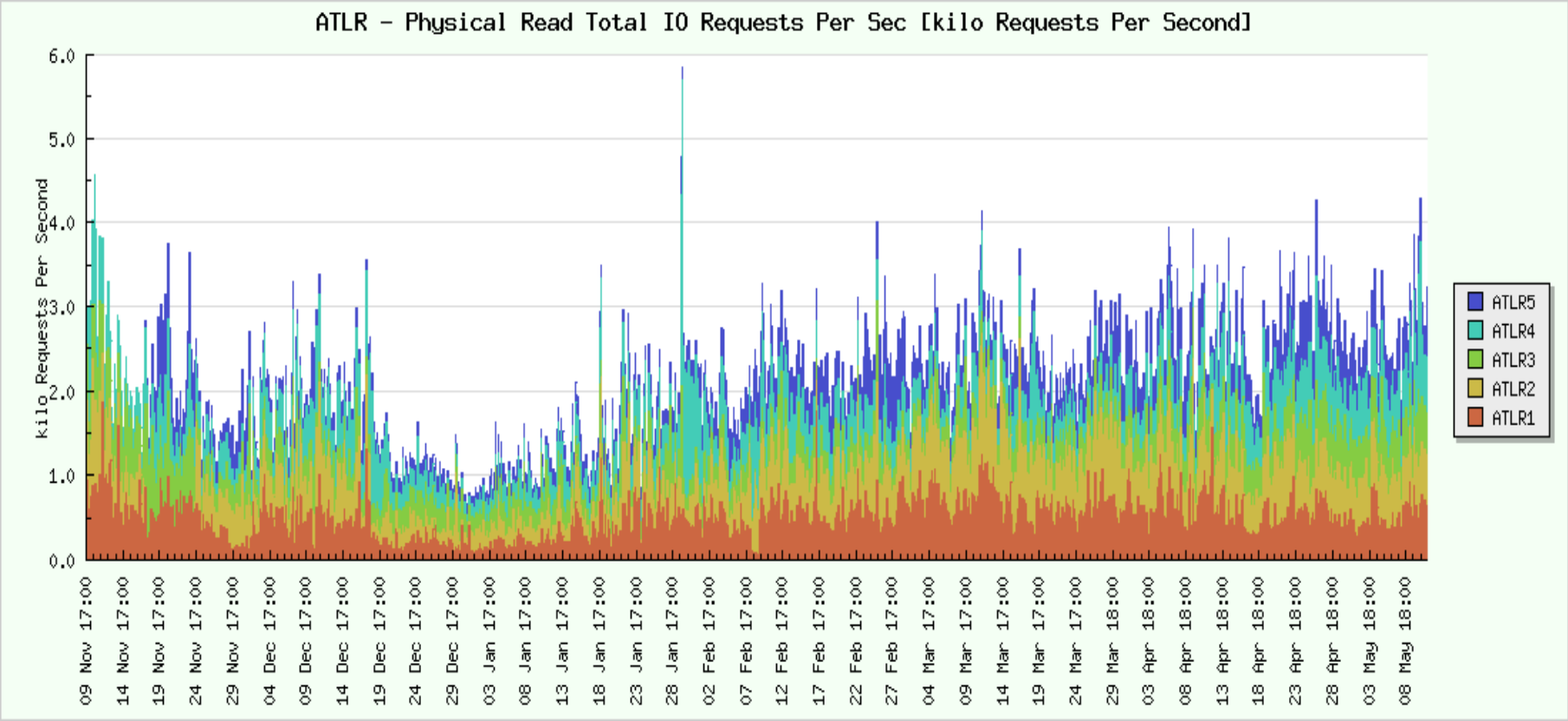


3D Streams Monitor



AWR repository for capacity planning

- We keep a repository from AWR of the metrics of interest (IOPS, CPU, etc)



Storage monitoring

- ASM instance level monitoring

Cluster: rac9 - RAC for GRID applications (monitoring enabled) last update: 2008-11-25 17:48

	Disk Groups											
	RAC9_DATADG1 RSTOR610	RAC9_DATADG1 RSTOR611	RAC9_DATADG1 RSTOR612	RAC9_DATADG1 RSTOR613	RAC9_DATADG1 RSTOR614	RAC9_RECODG1 RSTOR610	RAC9_RECODG1 RSTOR611	RAC9_RECODG1 RSTOR612	RAC9_RECODG1 RSTOR613	RAC9_RECODG1 RSTOR614	RAC9_RECODG1 RSTOR615	RAC9_RECODG1 RSTOR616
Reference: disks (failgroups)	16 (RSTOR610)	16 (RSTOR611)	16 (RSTOR612)	16 (RSTOR613)	16 (RSTOR614)	16 (FG1)	16 (FG1)	16 (FG1)	16 (FG1)	16 (FG1)	32 (FG2)	32 (FG2)
gridr1 (racn604) gridr2 (racn605) gridr3 (racn612) gridr4 (itrac613)	OK	OK	15 out of 16	OK	15 out of 16	OK	OK	15 out of 16	OK	OK	OK	OK

new failing disk on RSTOR614

- Storage level monitoring

rstor903
FW Version: 3.47B.06
Seq Optimize: Disable

Slot	LD	LUNs	Size	Speed	Status
1	0	0,1	238470	150MB	On-Line
2			238470	150MB	Frmt
3	2	3	238470	150MB	On-Line
4	3	4	238470	150MB	On-Line
5	4	5	238470	150MB	On-Line
6	5	6	238470	150MB	On-Line
7	6	7	238470	150MB	On-Line
8	7	8	238470	150MB	On-Line

new disk installed on RSTOR903 slot 2

Security

- Schemas setup with ‘least required privileges’
  - account owner only used for application upgrades
  - reader and writer accounts used by applications
  - password verification function to enforce strong passwords
- Firewall to filter DB connectivity
  - CERN firewall and local iptables firewall
- Oracle CPU patches, more recently PSUs
  - Production up-to-date after validation period
  - Policy agreed with users
- Custom development
  - Audit-based log analysis and alarms
  - Automatic pass cracker to check password weakness

DBAs and Data Service Management

- CERN DBAs
  - Activities and responsibilities cover a broad range of the technology stack
  - Comes natural with Oracle RAC and ASM on Linux
  - In particular leveraging on lower complexity of commodity HW
- Most important part of the job still interaction with the customers
  - Know your data and applications!
- Advantage: DBAs can have a full view of DB service from application to servers

Evolution of the Services and Lessons Learned

Upgrade to 11gR2

- Next ‘big change’ to our services
  - Currently waiting for first patchset to open development and validation cycle
  - Production upgrades to be scheduled with customers
- Many new features of high interest
  - Some already present in11gR1
  - Active Dataguard
  - Streams performance improvements
  - ASM manageability improvements for normal redundancy
  - Advanced compression

Active Dataguard

- Oracle standby databases can be used for read-only operations
- Opens many new architectural options
  - We plan to use active dataguard instead of streams for online to offline replication
  - Offload production DBs for read-only operations
  - Comment: active dataguard and RAC have a considerable overlap when planning a HA configuration
- We are looking forward to put this in production

ASM Improvements in 11gR2

- Rebalancing tests showed big performance improvements (a factor four gain)
  - Excessive re-partnering of 10g and 11gR2 fixed
- Integration of CRS and ASM
  - Simplifies administration
- Introduction of Exadata which uses ASM in normal redundancy
  - Development benefits ‘standard configs’ too ☺
- ACFS (cluster file system based on ASM)
  - performance: faster than ext3,

Streams 11gR2

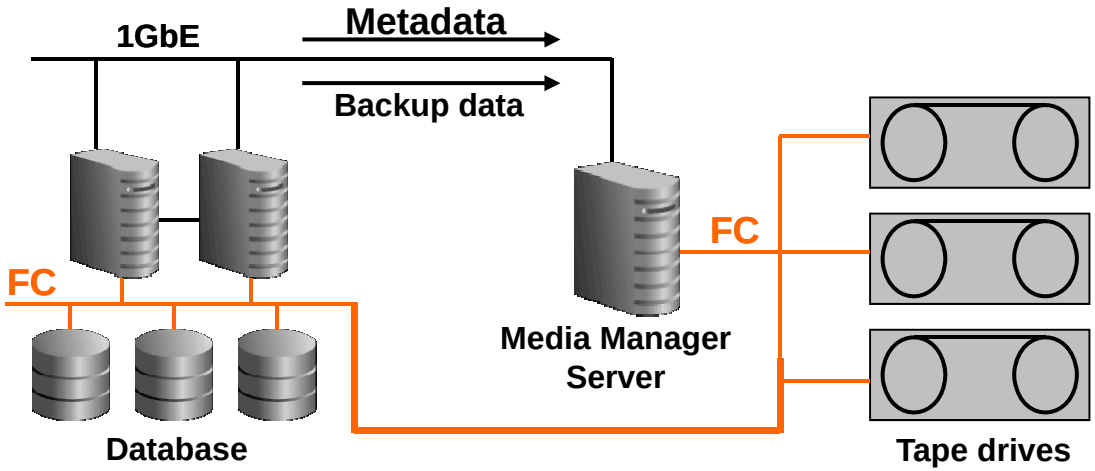
- Several key improvements:
- Throughput and replication performance has improved considerably
  - 10x improvements in our production-like tests
- Automatic split and merge procedure
- Compare and Converge procedures

Architecture and HW

- Servers cost/performance keeps improving
  - Multicore CPUs and large amounts of RAM
  - CPU-RAM throughput and scalability also improving
  - Ex: 64 cores and 64 GB of RAM are in the commodity HW price range
- Storage and interconnect technologies less straightforward in the ‘commodity HW’ world
- Topics of interest for us
  - SSDs
  - SAN vs. NAS
  - 10gbps Ethernet, 8gbps FC

Backup challenges

- Backup/recovery over LAN becoming problem with databases exceeding tens of TB
  - Days required to complete backup or recovery
  - Some storage managers support so-called LAN-free backup
    - Backup data flows to tape drives directly over SAN
    - Media management server used only to register backups
    - Very good performance observed during tests (FC saturation, e.g. 400MB/s)
  - Alternative – using 10Gb Ethernet



Data Life Cycle Management

- Several Physics applications generate very large data sets and have the need to archive data
  - Performance-based: online data more frequently accessed
  - Capacity based: Old data can be read-only, rarely accessed, in some cases can be put online ‘on demand’

- Technologies:
  - Oracle Partitioning: mainly range partitioning by time
  - Application-centric: tables split and metadata maintained by the application
  - Oracle compression
  - Archive DB initiative: offline old partitions/chunks of data in a separate ‘archive DB’

Conclusions

- We have set up a world-wide distributed database infrastructure for LHC Computing Grid
- The enormous challenges of providing robust, flexible and scalable DB services to the LHC experiments have been met using a combination of Oracle technology and operating procedures
  - Notable Oracle technologies: RAC, ASM, Streams, Data Guard
  - Developed in-house relevant monitoring and procedures
- Going forward
  - Challenge of fast growing DBs
  - Upgrade to 11.2
  - Leveraging new HW technologies

Acknowledgments

- CERN-IT DB group and in particular:
  - Jacek Wojcieszuk, Dawid Wojcik, Eva Dafonte Perez, Maria Girone

- More info:

  http://cern.ch/it-dep/db

  http://cern.ch/canali