Oracle Storage Performance Studies

Luca Canali, IT-DM

WLCG Collaboration Workshop

Database track, April 25th, 2008

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

Outline

- Goal: deploy I/O subsystems that meet applications' requirements
  - Input 1: What is needed. Ultimately comes from application owners
  - Input 2: What is available. Technology- and cost-driven
- How to measure HW performance?
- Lessons learned and experience collected at CERN and T1s

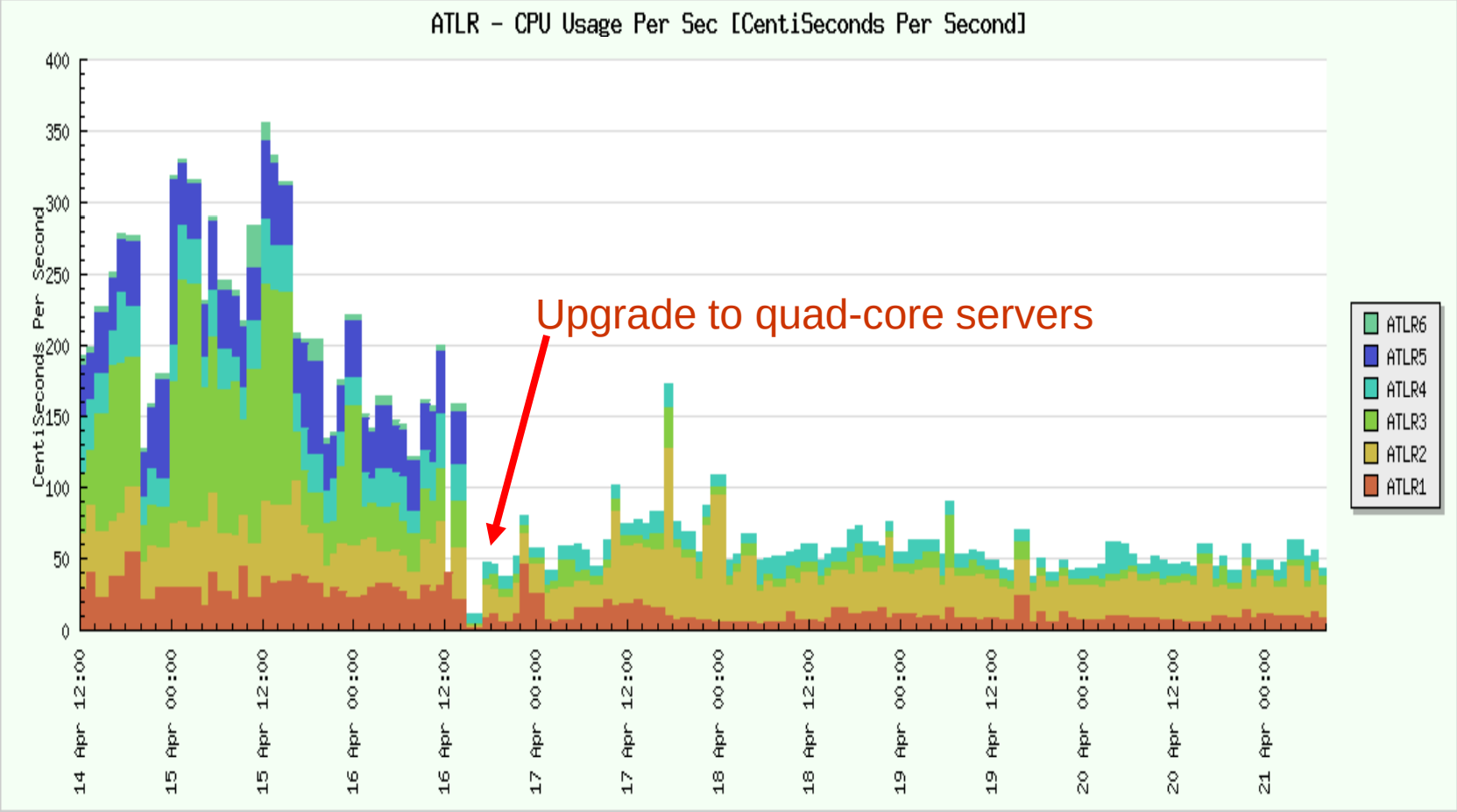
New Hardware for LHC startup

CERN IT Department

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

CPU and Memory Performance

- Measuring CPU and memory performance, is a relatively easy task.



Storage is a different game

- There is not yet a big jump in performance for storage technologies
  - This can quickly change in a few years (SSD)
- Measuring and sizing storage subsystems is not as easy as servers (CPU)
- For Oracle there is traditionally a lot of FUD focused towards high-end solutions
  - Traditionally storage configuration is not the DBA's job

Criteria for PDB Storage Sizing

- We have selected the following main criteria for sizing PDB storage:
  - IOPS: small random I/O performance (e.g. index-based read, nested loops join)
  - TB: total capacity after mirroring
  - MBPS: sequential IO performance (e.g. backups, full scans, hash joins)
- High Availability of the solution is a must
  - Storage is shared in a RAC cluster

Cluster Storage for Oracle

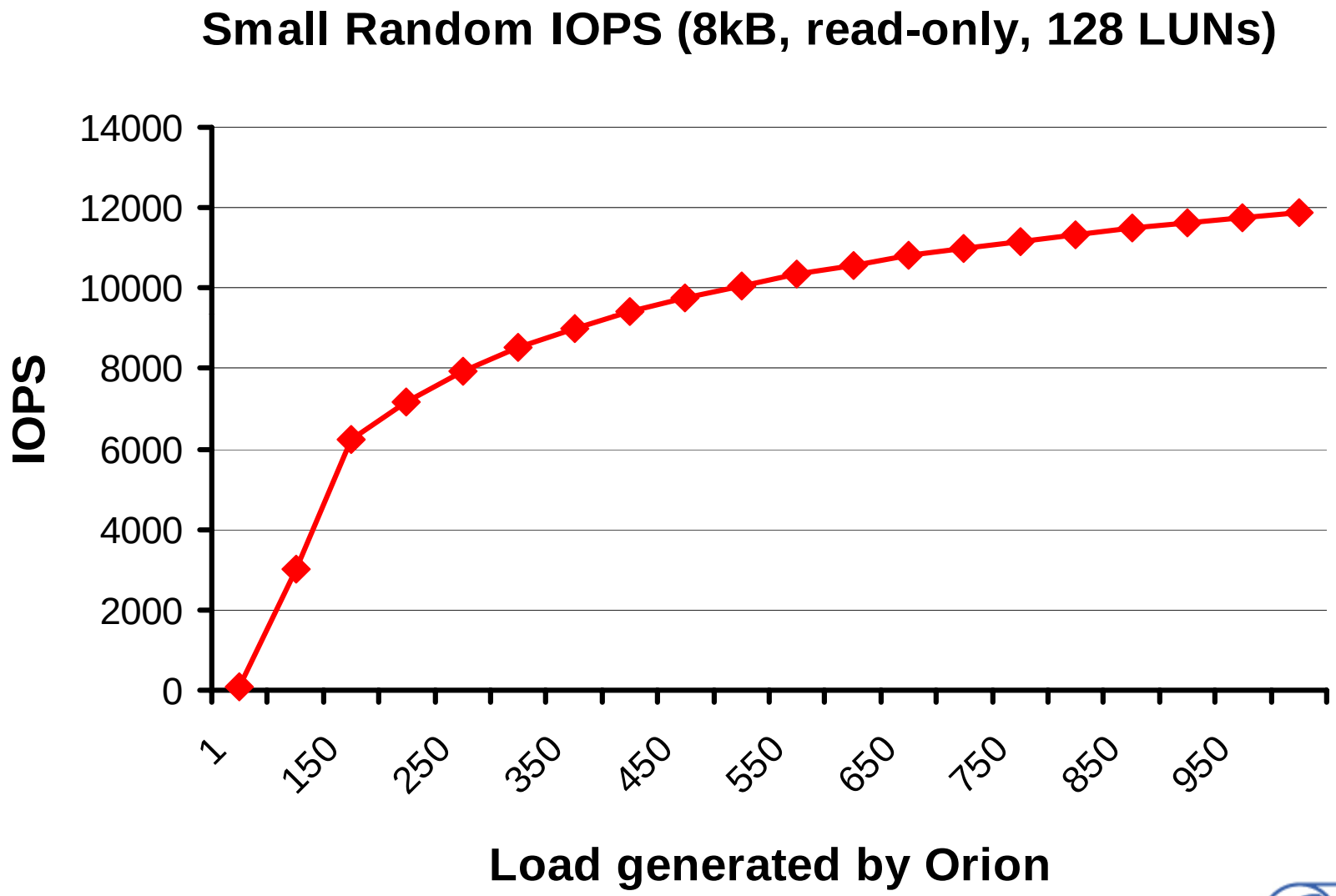
- Several layers between Oracle tablespaces and physical storage
  - Arrays, controllers, storage network, HBAs, volume managers, filesystems
  - Each of them can be a bottleneck
- How can we find this out?
  - Let's measure it!

Tools – Oracle's Orion

- Oracle ORION simulates Oracle workload:
  - IOPS for random I/O (8KB)
  - MBPS for sequential I/O (in chunks of 1 MB)
  - Latency associated with the IO operations

- Simple to use
  - Get started: `./orion_linux_em64t -run simple -testname mytest -num_disks 2`
  - More info: `https://twiki.cern.ch/twiki/bin/view/PSSGroup/OrionTests`

ORION output - example



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

ORION results – HW tests

- Small random read operations, IOPS as measured with ORION

System Type	RAID Level	IOPS	IOPS / DISK(*)	Usable Capacity
Infortrend 128x SATA	ASM 'RAID10'	12000	100	24 TB
Infortrend 96x Raptor	ASM 'RAID10'	16000	160	6.5 TB
Netapps SAN 80x SAS	RAID-DP 'RAID6'	17000	210	7.5 TB

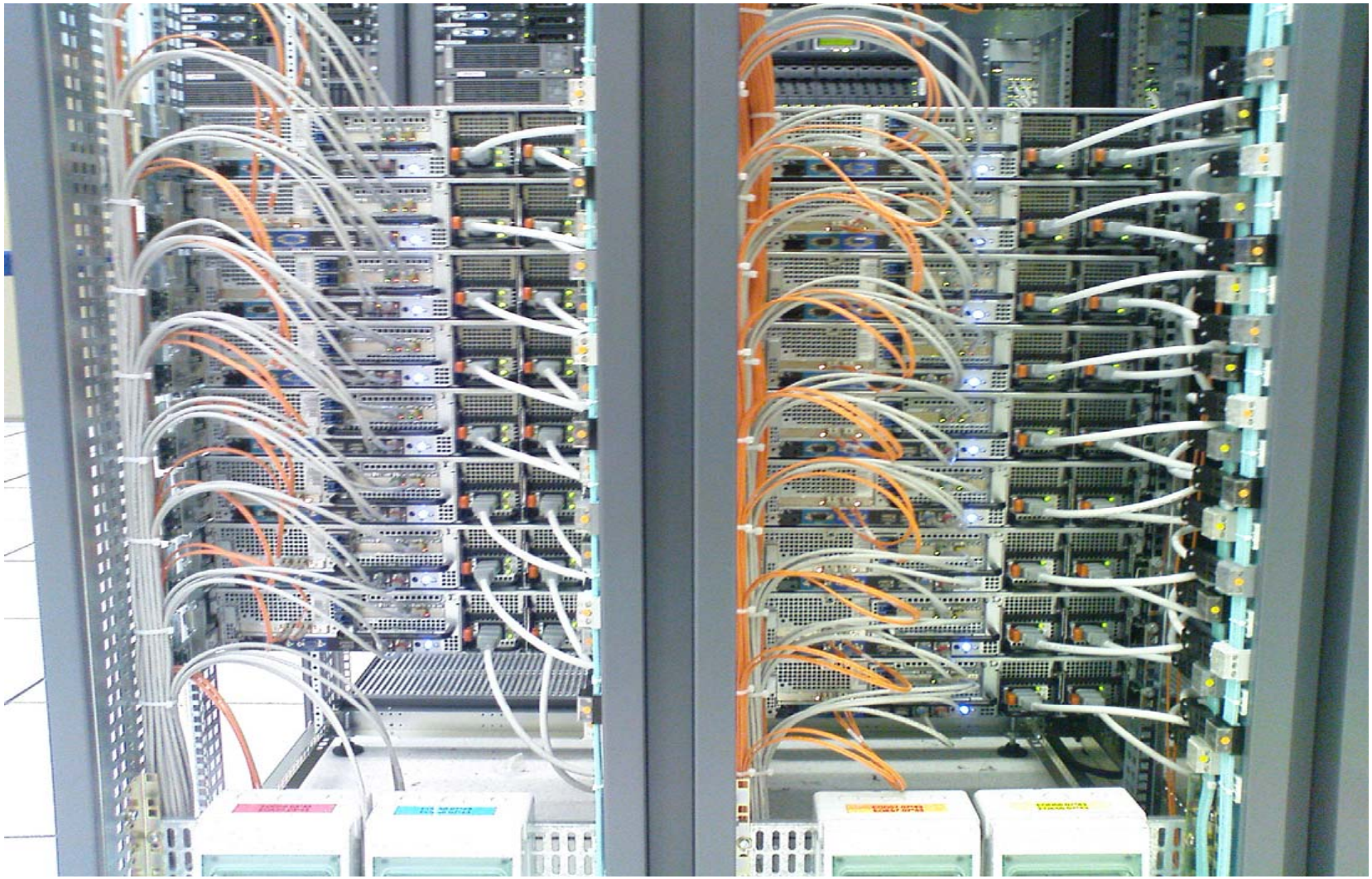
ORION - Lessons learned

- (*) IOPS measured by ORION are saturation levels (~10% higher than expected)
- Write tests are important
  - RAID5 and 6 volumes are OK for reads but have degraded performance with write operations
- Take cache in consideration when testing:
  - (ex: `-cache_size 8000`)
- Orion is very useful also when tuning arrays (parameters, stripe sizes, etc)

Case Study: The largest cluster I have ever installed, RAC5

CERN IT Department

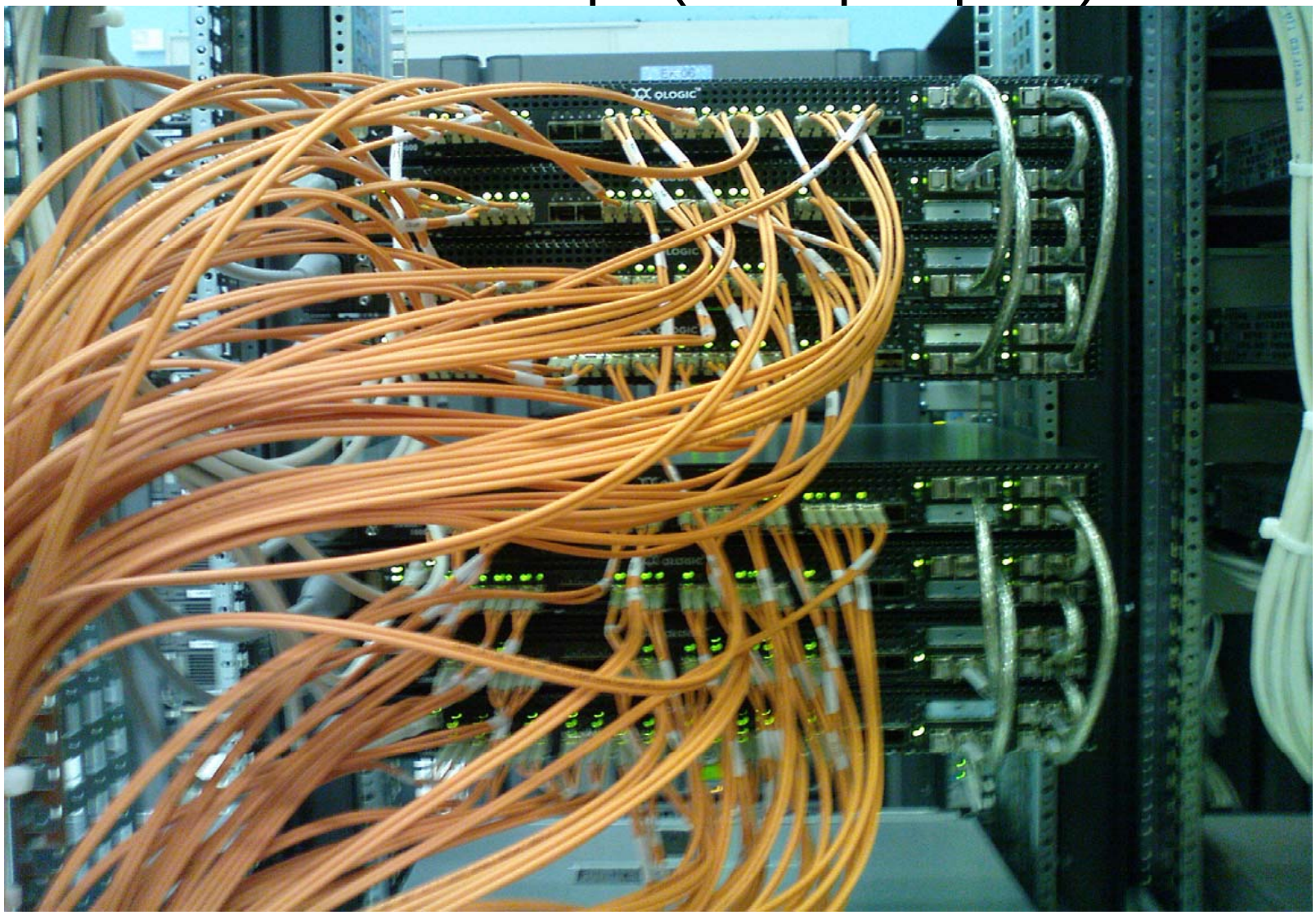
- The test used:14 servers



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

Multipath Fiber Channel

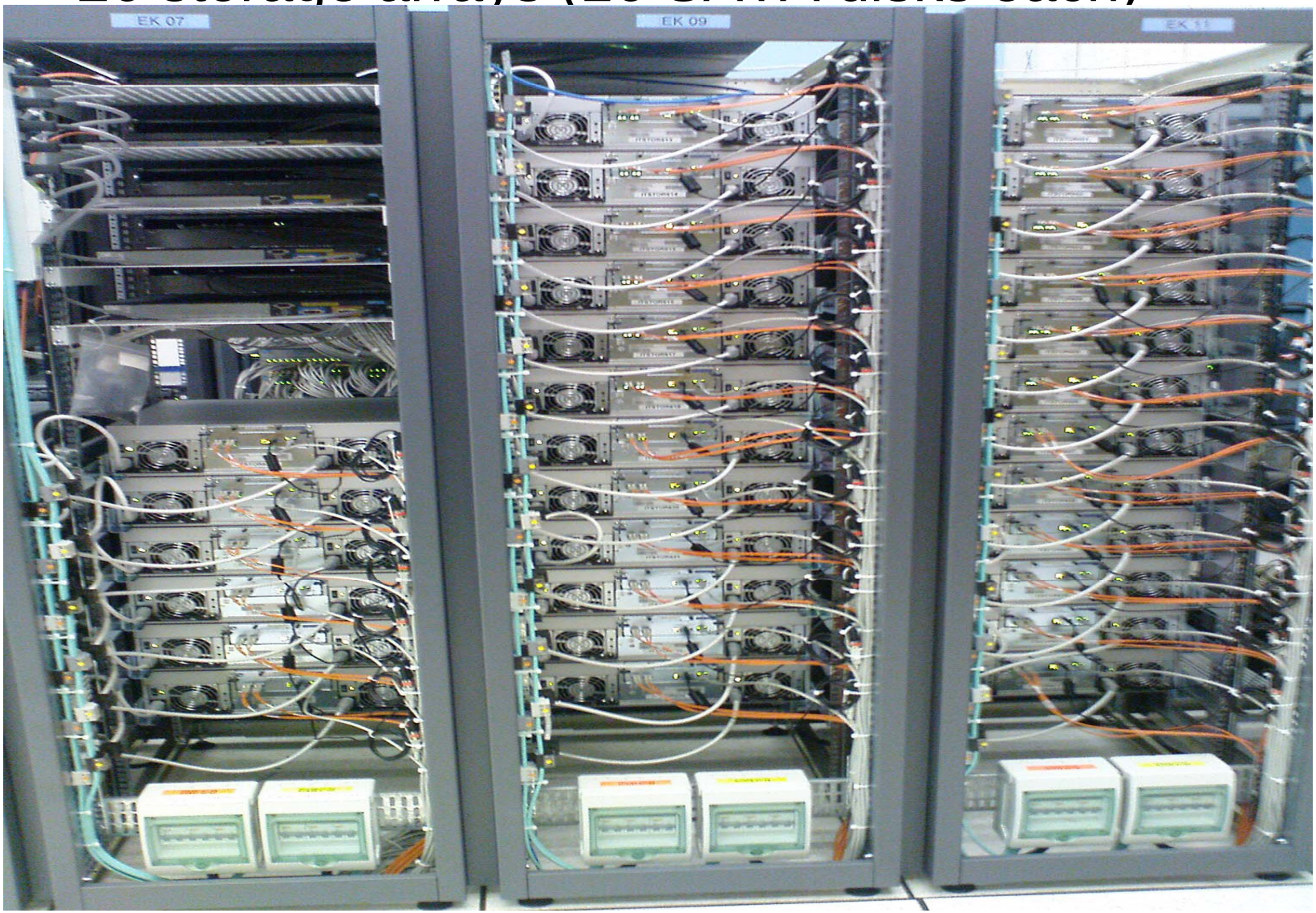
- 8 FC switches: 4Gbps (10Gbps uplink)



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

Many Spindles

- 26 storage arrays (16 SATA disks each)



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

Case Study: I/O metrics for the RAC5 cluster

CERN IT Department

- Measured, sequential I/O
  - Read: 6 GByte/sec
  - Read-Write: 3+3 GB/sec
- Measured, small random IO
  - Read: 40K IOPS small random
- Note:
  - 410 SATA disks, 26 HBAS on the storage arrays
  - Servers: 14 x 4+4Gbps HBAs, 112 cores, 224 GB of RAM

Tools – Custom SQL

- A custom SQL-based DB workload:
  - IOPS: Probe randomly a large table (several TBs) via several parallel queries slaves (each reads a single block at a time)
  - MBPS: Read a large (several TBs) table with parallel query
  - Scripts are available on request
- The test table used for the RAC5 cluster was 5 TB in size
  - created inside a diskgroup of 70TB

Case Study: Conclusions

- Commodity Storage on SAN and Linux servers can scale (tested up to 410 disks)
- RAC and ASM can scale up to 410 disks

- Lessons learned:
  - disk and controller ‘infant mortality’ can be a problem.

Conclusions

- Configuration of the storage subsystems for WLCG databases is critical
  - HA, IOPS and capacity are the most critical parts
- ASM and SAN on commodity HW have proven to perform and scale
- Being able to measure HW performance is an important part of the process
  - Arrays with SATA, Raptor and SAS have been tested at CERN and T1s
  - SATA are proven, SAS are promising
  - more tests are planned

Acknowledgments

- This presentation contains the work of the IT-DM DBA team: Dawid, Eva, Jacek, Maria, Miguel.
- Many thanks to Luis (PICS) for the Netapps tests.
- Thanks also to the T1 DBAs who showed interest in testing their storage with Orion: Andrew, Carmine, Carlos.