Possible Service Upgrade

Jacek Wojcieszuk, CERN/IT-DM

Distributed Database Operations Workshop

April 20th, 2009

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

Outline

- Hardware
  - CPU/Memory
  - Storage
- OS
  - RHEL 5
- Oracle Software
  - Data Guard
  - Oracle 11g
- Procedures
  - Backup&Recovery
  - Data Lifecycle Management

Hardware

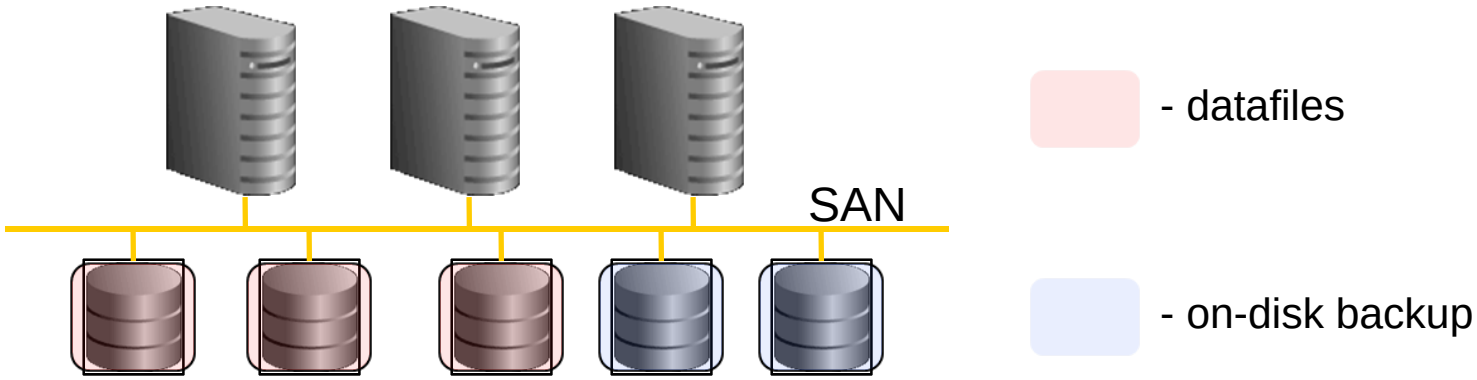
- Although the CPU clocks do not speed up anymore, new processors continue to offer better and better performance
  - More cores
  - Improved microarchitectures
  - Faster and bigger cache
- Intel Nehalem
  - Expected to be even twice faster than the currently used CPUs: http://structureddata.org/2009/04/10/kevin-clossons-silly-little-benchmark-is-silly-fast-on-nehalem/
  - To be released this year
  - Very promising for Oracle customers paying per-core licenses although Oracle licensing not clear, yet

More on hardware

- Memory
  - Is a critical factor for performance of Oracle RDBMS
  - Growing databases require bigger caches
    - To keep IOPS reasonable
- Storage
  - 10k RPM disks get cheaper
    - WD Raptor: ~140 IOPS, ~$300
    - Capacity grows as well
  - Bigger SATA disks
    - Up to 2 TB
    - Making on-disk backups cheaper and cheaper
  - New promising interconnect technologies:
    - iSCSI (1Gb or 10Gb)
  - See Luca's talk for more details

Latest hardware acquisition

- RAC7
  - Ordered last year, will pe put in production in coming weeks
  - 20 blade servers split into 2 chassis
    - Dual Quad Core, 16MB of RAM
  - 32 x 12-bay disk arrays equipped with WD Raptor disks (10k RPM, 300GB)
  - 12 x 12-bay disk arrays equipped with 1TB SATA disks (7.2k RPM)
  - Will host CMSR, LCGR, LHCBR and Downstream Capture databases



RAC7

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

Next hardware acquisition

- RAC8 and RAC9
  - Replacement for RAC3 and RAC4
  - Should be put in production in late autumn 2009
  - 40 blade servers grouped into 4 chassis
    - Nehalem CPUs
    - At least 24 GB of RAM
  - 60 12-bay disk arrays
    - Big SATA disks (2TB most likely)
    - iSCSI solutions considered

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

Operating System

- RHEL 5 seems to be mature enough to be used for production
  - 3rd update already released
- Few minor improvements:
  - Bigger file systems allowed
  - Improved TCP/IP stuck (important for iSCSI)
  - Improved (?) DevMapper management
- RedHat stops support for RHEL 4 in 2012
- Some features of Oracle 11gR2 not supported on RHEL 4
- At least 2 possible migration scenarios:
  - Rolling upgrade: node-by-node
  - Data Guard and switchover

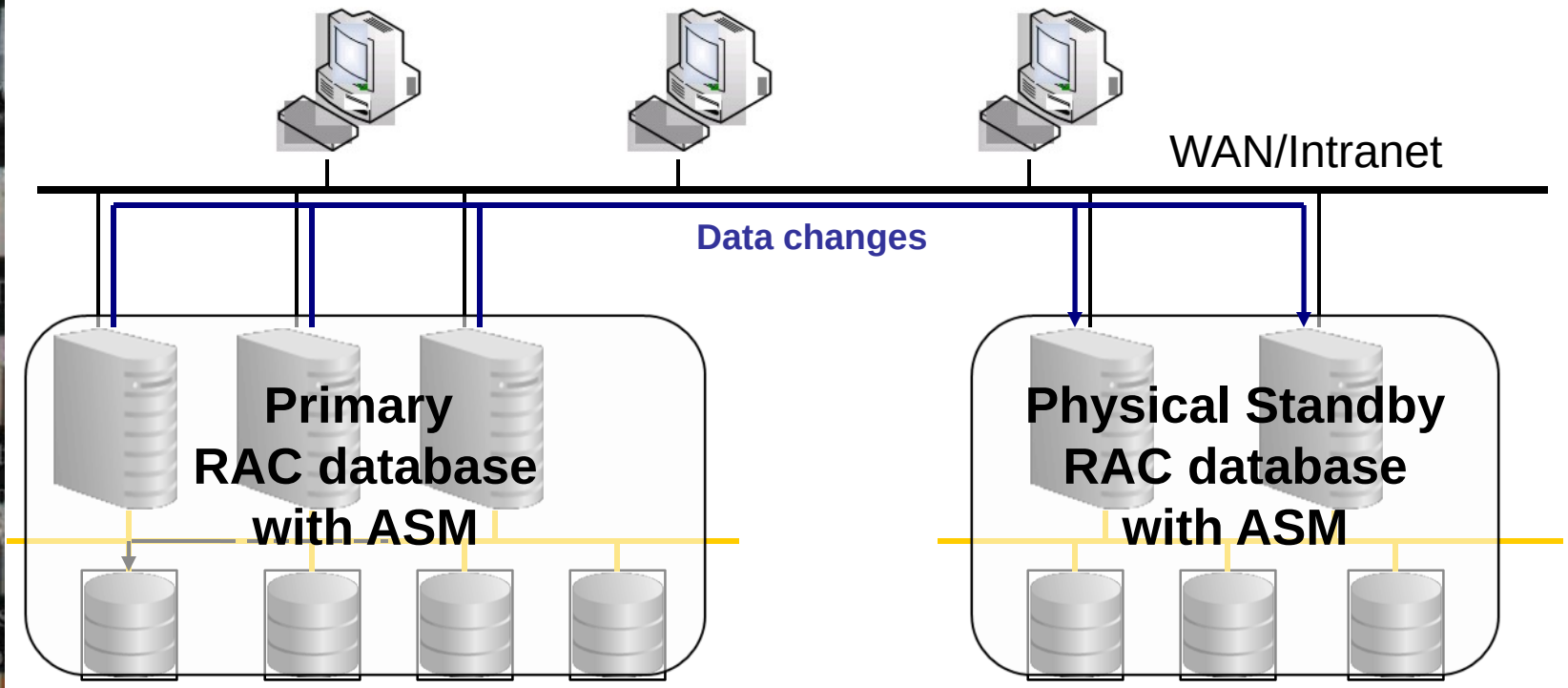
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Data Guard Physical Standby Database

- Mature and stable technology
- The most reliable Oracle recovery solution
  - Simplifies and speeds up both: full database recovery and handling human errors
  - Less error-prone than RMAN-based recovery
  - Very important for on-line databases of CMS, LHCB and Alice which are not backed up to tapes
- In conjunction with flashback logging it allows to run various tests of applications
- A lot of experience gathered in past years:
  - Data Guard for migrations
  - Pilot setup during the 2008 run
- Plans to deploy standby database for majority of production systems

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

Data Guard Physical Standby Database



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

Data Guard - configuration details

- Standby databases configured as RAC
  - Servers sized to handle moderate applications' load
  - Enough disk space to fit all the data and few days of archived redo logs
- Identical Oracle RDBMS version on primary and standby
  - The same patch level
- Asynchronous data transmission with the LGWR process
  - Standby redo logs necessary
  - Only few last transactions can be lost
- Shipped redo data applied with 24 hours delay
- Flashback logging enabled with 48 hours retention
- Data Guard Broker
  - to simplify administration and monitoring

Oracle 11g

- Many interesting features:
  - ASM fast mirror resync
    - Short unavailability of one side of the ASM mirror will not result in disk eviction
  - Active Data Guard
    - Standby database can be continuously used for query/reporting activity
  - Real Application Testing (Database Reply)
    - Allows replaying captured workload on demand
    - Potentially very useful for testing and optimization

Oracle 11g

- Other interesting 11g features:
  - SQL Plan Management
    - Prevents Oracle from switching to a worse execution plan
  - Oracle Secure files (LOBs)
    - Improved LOB performance
    - Compression and encryption
  - Interval based partitioning
- Even more features in 11gR2
  - Hopefully will be released in autumn
- We plan to go straightaway to 11gR2
  - On production after first patchset
  - On integration probably several months earlier

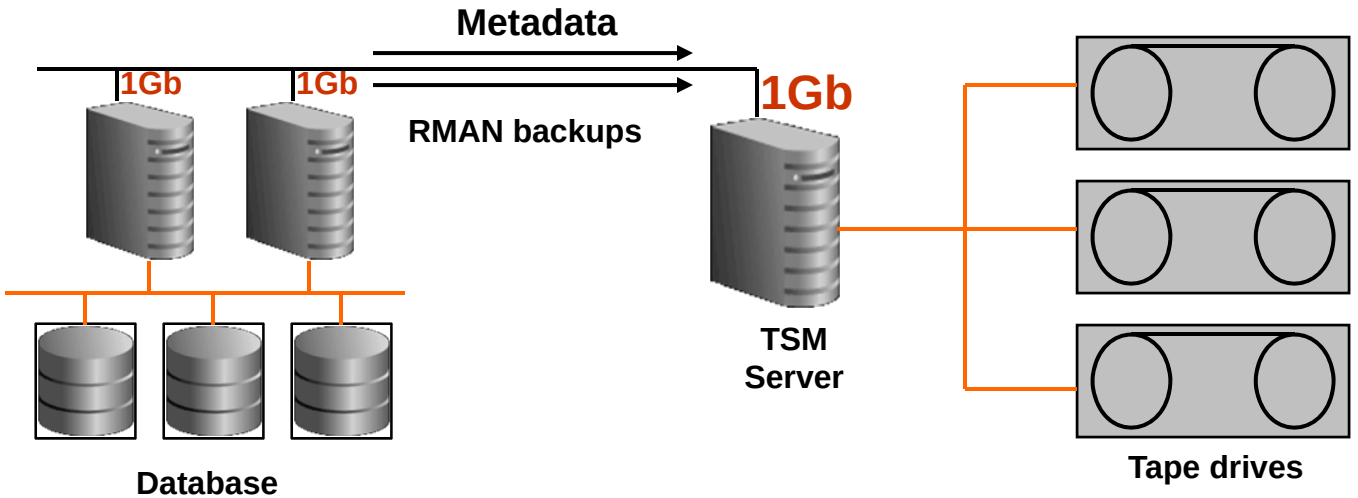
Backup&Recovery

- Even with standby databases in place tape backups will stay as the key element of the backup&recovery strategy
  - Certain type of failures require restore from tapes:
    - Disasters
    - Human errors discovered long time after they were made
- Backing up to tapes and restoring those backups is a challenging task as the databases get bigger and bigger

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

Current tape backup architecture at CERN

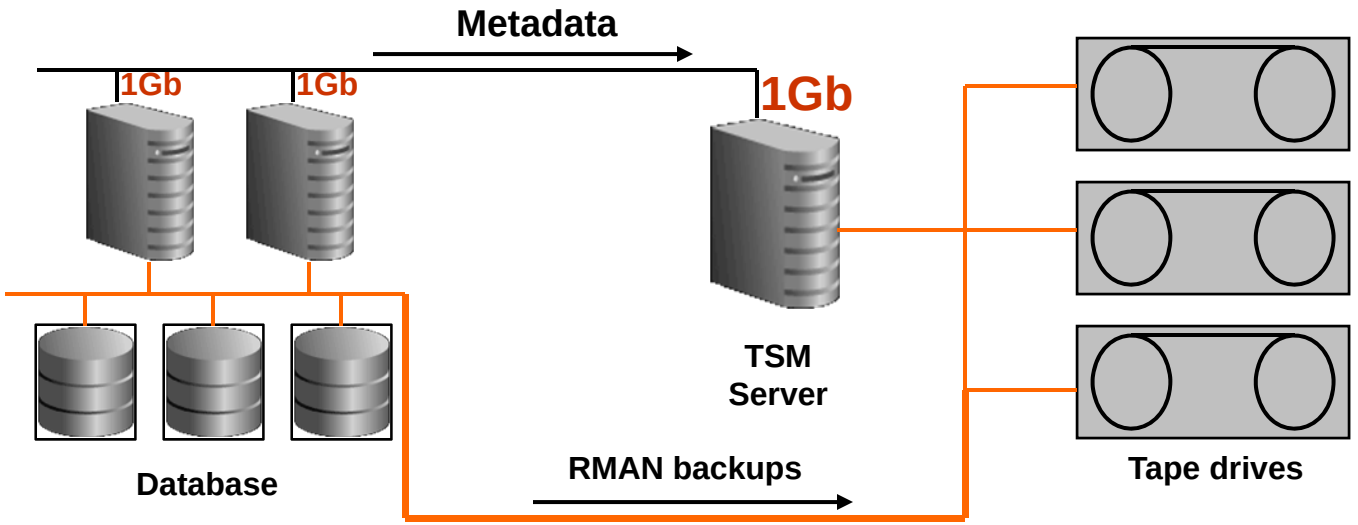
- Traditionally at CERN tape backups are sent over a general purpose network to a media management server:
  - This limits backup/recovery speed to ~100 MB/s
  - Backup/restore of a 10TB database takes almost 30 hours!
  - At the same time tape drives can archive data with the speed of 160 MB/s compressed



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

LAN-free tape backups

- Tivoli Storage Manager supports so-called LAN-free backup
- When using LAN-free configuration:
  - Backup data flows to tape drives directly over SAN
  - Media Management Server used only to register backups
  - Very good performance observed during tests (see next slide)



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

LAN-free tape backups - tests

- 1 TB test database with contents very similar to one of the production DB
- Different TSM configurations:
  - TCP and Shared Memory mode
- Backups taken using 1 or 2 streams

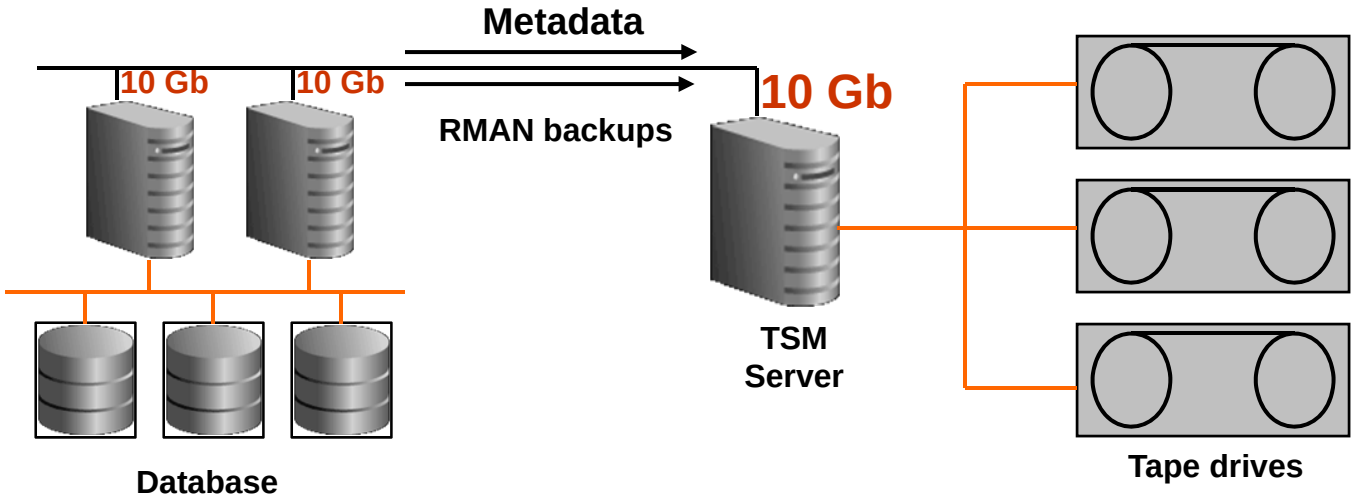
	TCP	Shared mem
1 stream	198 MB/s	231 MB/s
2 streams	361 MB/s	**402 MB/s**

- Restore tests done using 1 stream only
  - Performance of a test with 2 streams affected by Oracle software issues (bug 7630874)

	TCP	Shared mem
1 stream	150 MB/s	**158 MB/s**

LAN-free backup alternatives

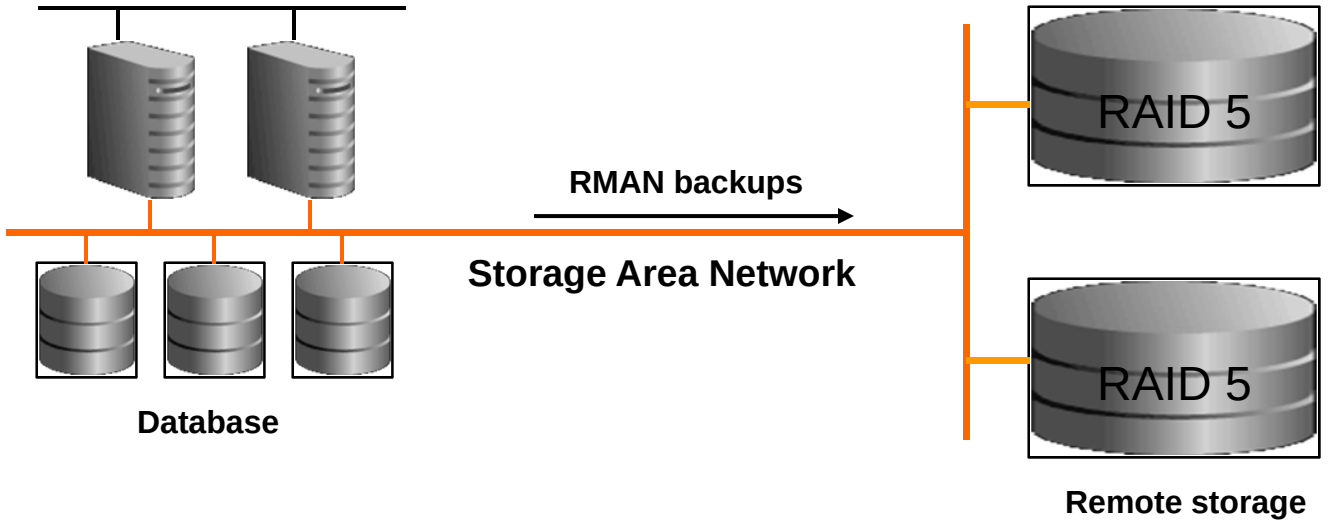
- 10 Gb network for backups
  - Will be tested in next few weeks



- Using a disk pool instead of tape drives

A disk pool instead of tapes - test

- Two tested configurations:
  - SAN-attached storage with a file system
  - SAN-attached storage with ASM
  - 1 disk array, 2 x 8 disk RAID 5
  - Test performed with 4 streams



	ext3	ASM
4 streams (backup)	235 MB/s	**369 MB/s**

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

Summary

- No revolutionary changes this year
  - Move to RHEL5 should be quite smooth
  - Migration to 11gR2 will happen next year the earliest
- We are ready for data growth
  - Enough hardware resources allocated
- Many procedure improvements
  - Backup&Recovery
  - Dealing with big data volumes
  - Security

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

Thank You

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