# Data Management issues

R. Graciani

LHCb-Tier1 Jamboree

Lyon 7-8 March 2010

# Overview

- Storage Element Status
- Data Management
- Data Access
- Summary

# Disk Usage

| Disk Summary | Pledge (TB) | Seen by SLS | | Seen by LHCb | |
|---|---|---|---|---|---|
| 20/12/2010 | | TB | | TB | |
| | | Total | Used | Used | Pledge-Used |
| **FZK** | 495 | 500 | 331 | 339.9 | 155.1 |
| **IN2P3** | 610 | 641 | 334 | 320.7 | 289.3 |
| **CNAF** | 450 | 463 | 392 | 391.6 | 58.4 |
| **NL-T1** | 560 | 563 | 339 | 254.5 | 305.5 |
| **PIC** | 240 | 255 | 138 | 138.3 | 101.7 |
| **RAL** | 505 | 791 | 562 | 453.3 | 51.7 |
| **Tier1s** | **2860** | **3213** | **2096** | **1897.5** | **962.5** |
| **CERN** | **1135** | **1175** | **922** | **763.6** | **371.4** |

✔ There is space available.

✘ It is highly fragmented, many full SEs.

# CERN as of Feb 2011

|  |  | Pledge (TB) | Seen by **SRM** |  |  | Seen by **LFC** |  |
|---|---|---|---|---|---|---|---|
|  |  |  | TB |  |  | TB |  |
| SRM Space Token |  |  | Total | Used | Avail. | Used | **Pledge-Used** |
| LHCb_RAW | T1D0 | 380 | 35 | 31.4 | 3.6 | 179.4 | **200.6** |
| LHCb_RDST | T1D0 | 325 | 71.5 | 64.3 | 7.2 | 130.4 | **194.6** |
|  |  |  |  |  |  |  |  |
| LHCb_M-DST | T1D1 | 350 | 352.8 | 265.2 | 87.6 | 239 | **111** |
| LHCb_DST | T0D1 | 0 | 87.3 | 12.6 | 74.7 | 8.1 | **-8.1** |
| LHCb_MC_M-DST | T1D1 | 580 | 510.9 | 378.6 | 132.3 | 326.5 | **253.5** |
| LHCb_MC-DST | T0D1 | 0 |  |  | 0 |  | **0** |
| LHCb_USER | T0D1 | 205 | 200.1 | 192.1 | 8 | 160.7 | **44.3** |
| LHCb_HIST | T0D1 | 0 | 20 | 11.7 | 8.3 | 2.4 | **-2.4** |
| LHCB_FAILOVER | T0D1 | 0 | 4.5 | 1 | 3.5 | 0.1 | **-0.1** |
| CERN-disk | T0D1 | 0 |  |  | 0 |  | **0** |
| CERN-tape | T1D0 | 0 |  |  |  |  | **0** |

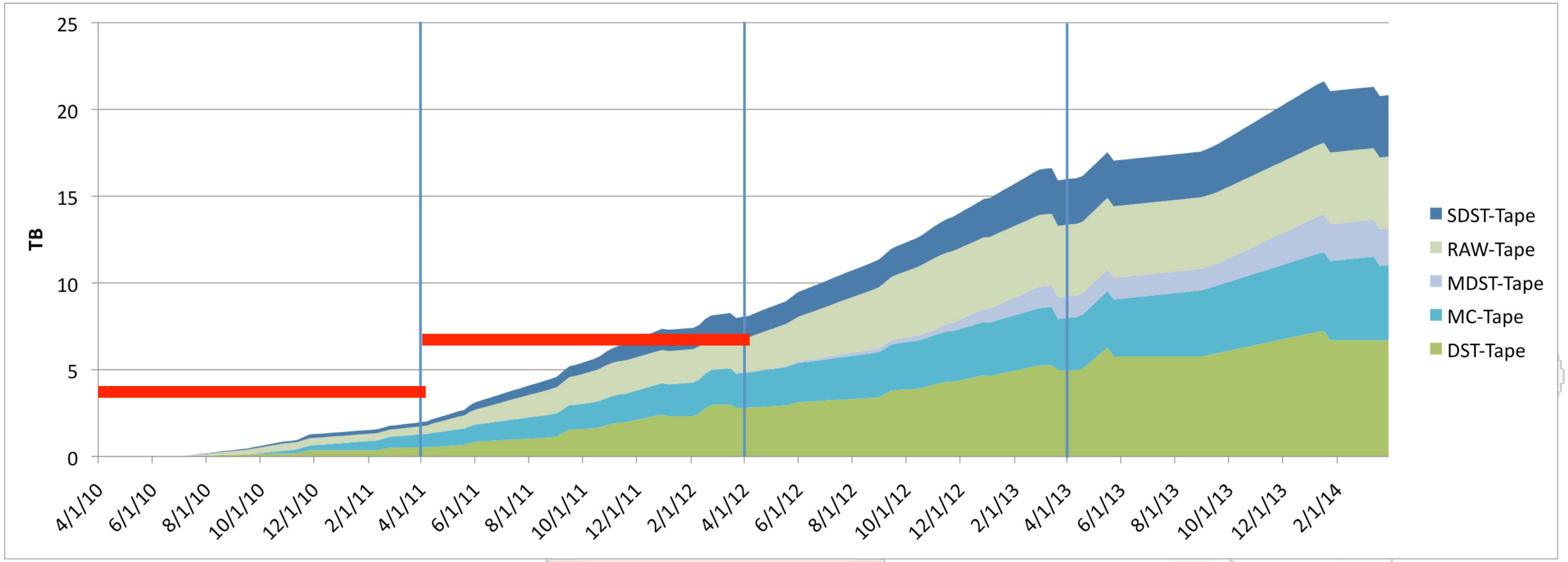
✔ There is space available.

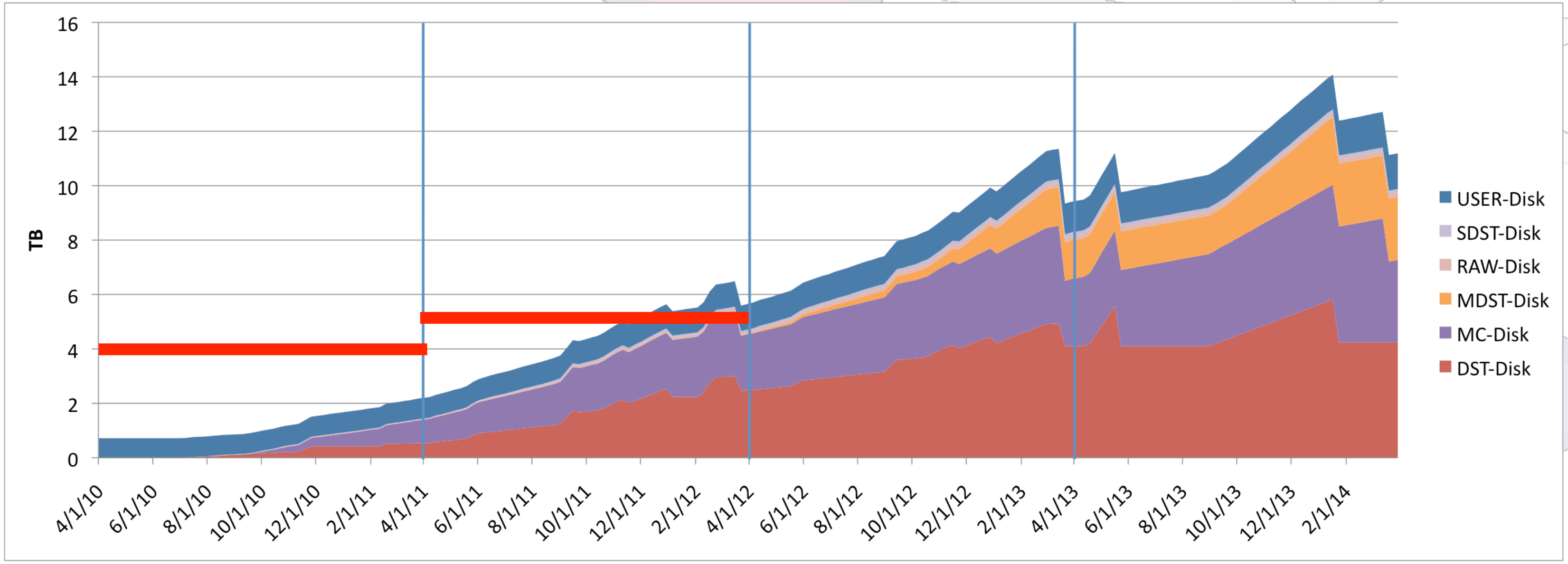
✘ It is highly fragmented, many full SEs.

✘ 100 TB Disk as Tape Cache

✘ Small Space Tokens are very inefficient

# 2011 Re-assessment Storage





# 2011 Re-assessment Storage

| Disk | 2011 | |
|---|---|---|
| | PB | % |
| Tie0 | 1.9 | 26 |
| Tier1 | 5.3 | 74 |

| Tape | 2011 | |
|---|---|---|
| | PB | % |
| Tie0 | 5.6 | 57 |
| Tier1 | 4.3 | 43 |

❍ Disk pledge: 1.5/3.8 TB

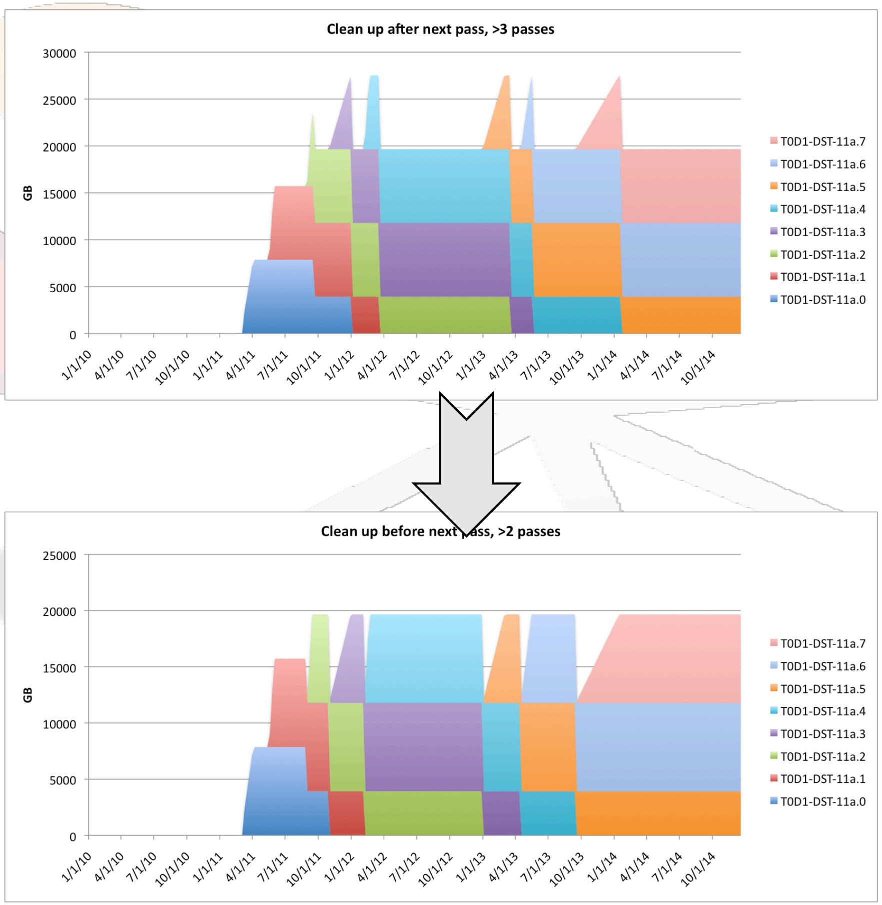
❍ Tape pledge: 2.5/3.9 TB

✔ It is a kind of worst case scenario

✘ With 60% usage we are at the limit

# Change in the model

- Old model:
  - 2 x T1D1
    - eventually becomes T1D0
  - 5 x T0D1
    - first reduced,
    - then removed
- New model
  - T2D0 (CERN) + T1D0
    - never removed
  - 2 x T1D1
    - first reduced,
    - then removed
  - 2 x T0D1
    - First reduced
    - Then removed



# Further improvements

- Not less than 2 archive copies
- Can we reduce "master" replicas (T1D1)?
  - "Active" data requires 2 replicas
  - Since have now 2 archive copies
    - Is it really transparent recovery?
    - How often are we able to recover from T1?
  - Can we recover from other replicas?
    - We need the procedure to recover T0D1 replicas
  - This might save on Tape
- "extra" replicas on T0D1
  - They are static at the moment
  - We are working into a dynamic model
    - Depending on the fraction of "hot" data
  - Might or might not save on Disk
- Target:
  - 2 "master" + 0-5 "extra"

# Dynamic data replication

- Base on usage
  - All usage goes through DIRAC
  - Need to implement metric
- Requires
  - Replication policies
  - Cleanup policies
  - Proactive consistency of SE vs LFC check
- Hard to predict Data vs MC ratios
  - Dynamic allocation of shares
  - Single configuration point
    - Reduce number of Space Tokens
    - Make DIRAC handle the shares

# Scenarios

- **Aggressive** 3 Tokens
  - 1 T1D0:
    - RAW, SDST (write, + n read)
    - Archival (write + 0 read)
  - 1 T0D1:
    - "master" replicas
    - "dynamic" replicas
    - "disk caches" merging, failover, freezer…
  - 1 T0D1:
    - Users
- **Conservative** 5 Tokens
  - 2 T1D0
    - Separate RAW/SDST from "archival"
  - 1 T1D1
    - "master" replicas
  - 1 T0D1
    - "dynamic" + caches
  - 1 T0D1
    - Users

# Data Access issues

- Reprocessing
  - ✔ Stage + local copy on WN
- Stripping
  - ✘ Stage + protocol access
    - ✘ Depends on configuration of the tape cache
  - ? Stage + local copy on WN
- Merging
  - ✔ Local copy on WNs
  - ? Protocol access
  - ? Dedicated WNs
- Analysis
  - ❑ Protocol access
    - ☆ Must scale up factor > 5
  - ❑ Need error recovery at protocol level

# Other work

- Historical usage of Storage (LFC/SE)
  - Ready for Users
  - Almost ready for low level data categories
    - Data, MC, users, test
  - Working on high level data categories
    - Run #, MC channel, Reco Pass,…
- Consistency
  - Detection
  - Correction
  - Feed back
- FTS
  - Improve traceability, error handling,…
- Stager
  - Improve traceability, error handling,…
- Removal
  - Implement as asynchronous operations
- Replications
  - Further development needed for dynamic

# Summary

- We are in reasonable shape but…
  - Will be working much closer to the limit
  - Will require extra flexibility
- We are aware and need to
  - Simplify the ground
  - Improve/develop tools
  - Evaluate performance
  - Iterate with your help
- For 2011 DM will be the real challenge
- But, we should not forget data access