LHCb data managment

Angelo Carbone
INFN-Bologna

Workshop CCR INFN GRID 2011
19 Maggio 2011

Outline

- Introduction
  - Computing model at a glance
  - LHCb data flow
- LHCb Data Management with Dirac
- Experience of data management with real data
  - Change of computing model
- Data analysis management

Introduction

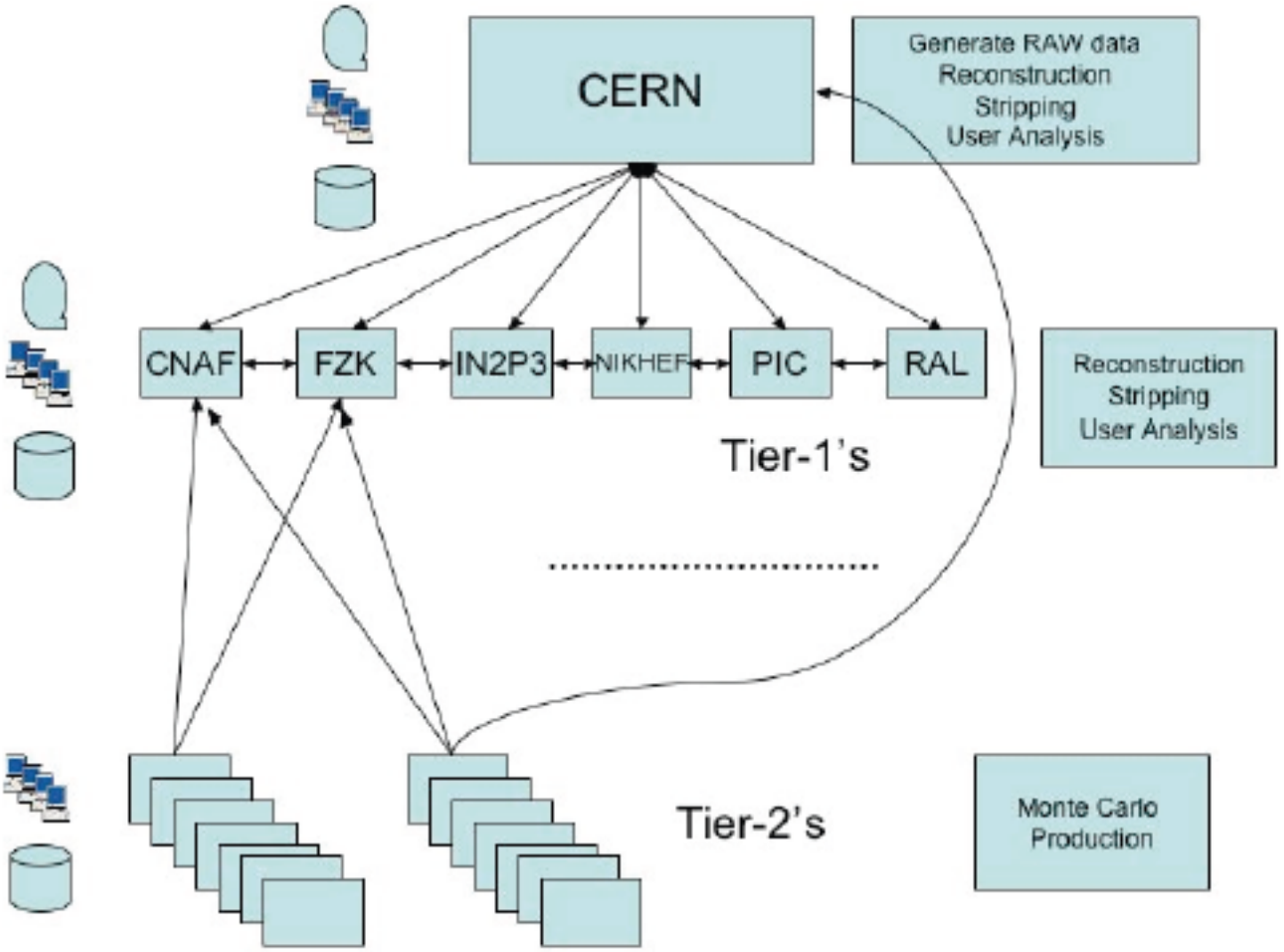
LHCb Data Characteristics

- Data size and rates are modest compared to other LHC experiments
  - 35 kB RAW event size
  - Trigger rate: 2000-3000 events/s
  - 25 kB RDST (a.k.a. ESD), 85 kB DST (a.k.a. AOD)
  - Typical reconstruction time: 12 HS06.s/event
- Physics research channels are rare
  - b-quark CP violation decay modes (BR ~ 10^{-9} to 10^{-6})
  - Typically a few 10'000s to a million events per year (2 fb^{-1})
    - A needle in a haystack
  - Easier to extract b decay events if only one primary vertex
  - Metrics = average number of visible interactions per beam crossing (μ)
    - For LHC design characteristics μ=0.4 at LHCb

Guidelines for the Computing Model

- Small processing time, but high trigger rate
  - 24 kHS06 required for reconstruction
    - Typically 2000 CPU slots
  - Tier-0 could not provide the necessary CPU power
  - Use Tier1s as well for reconstruction (first pass)
- Most problems for analysis jobs are related to Data Management
  - SE accessibility, scalability, reliability…
  - Restrict the number of sites with data access
  - Use Tier1s for analysis
- High requirements on simulated data
  - Background identification, efficiency estimation for signal
  - Typically 360 HS06.s per event
  - Use all possible non-Tier1 resources for simulation

LHCb Computing Model



Computing activities

- Production activities
  - Simulation, reconstruction, stripping, WG analysis (µDST)
  - Use DIRAC and the LHCb Production System
- User analysis
  - Data and MC analysis
  - For testing, use local resources (including local batch system)
  - For large datasets, use Grid Computing
- Toy MC and fits
  - Use Grid Computing for large samples
- Non-Grid user analysis
  - Mostly interactive analysis on local clusters (Tier-3), desk/lap-top (Tier-4)

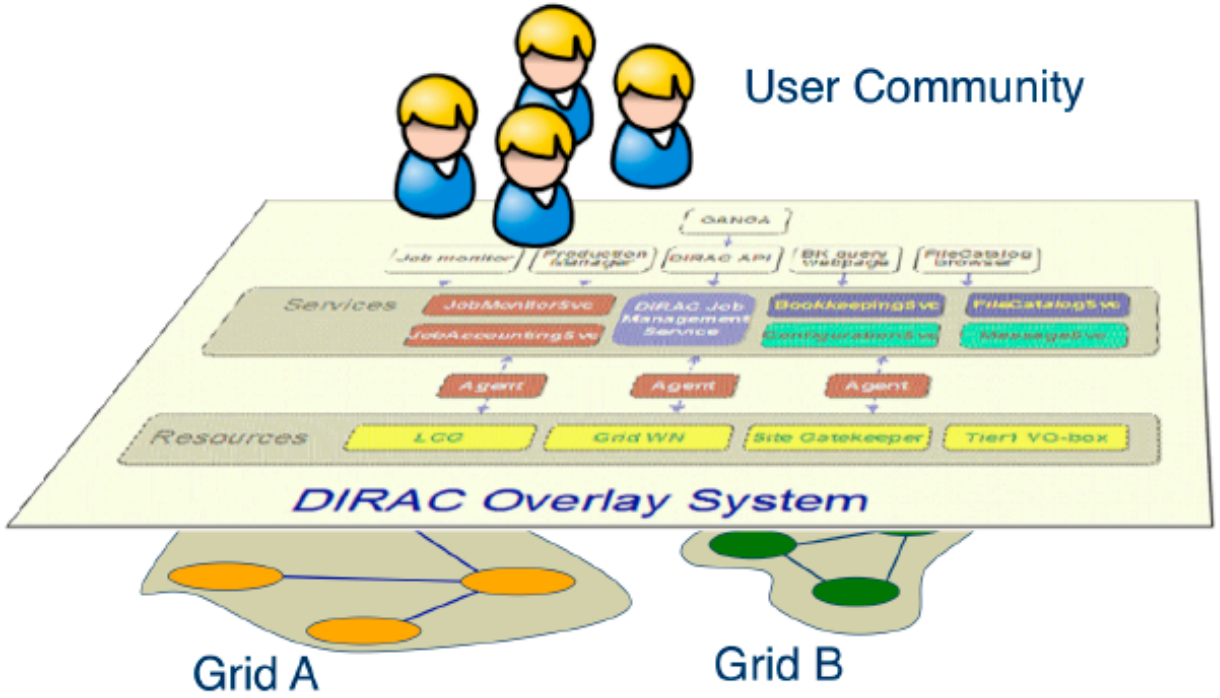
Data flow

- As soon as the raw data are recorded by the detector they are sent to Tier-1's+CERN-CAF to be reconstructed
- After the reconstruction, the Stripping process is performed
  - pre-selections provided by analysis working groups are applied in order to
    - write on disk different streams where each contains similar selected events, such as B-hadron, V0 decays, charm decays, etc…
    - reduce the data sample to a handy size in order to perform a finely tuned analysis
- The stripping can be performed several times on the same datasets (according to the availability of resources) if pre-selection algorithms change.

Data management : DIRAC

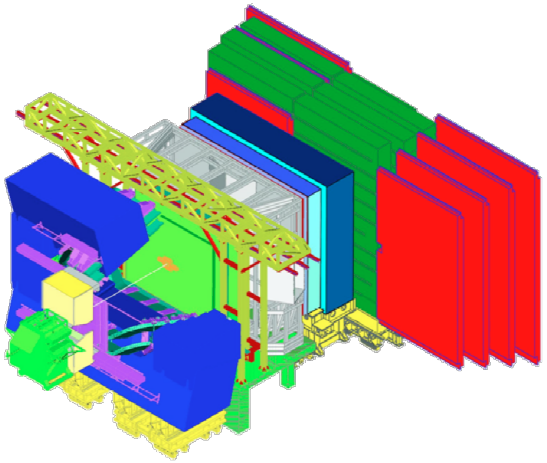
DIRAC: Community Grid Solution

The DIRAC project is a complete Grid solution.
Designed to be used by a community of users.
Services and agents of DIRAC overlay resources.
Transparent use of different grids e.g. gLite, NorduGrid etc.
Integration of non-grid resources e.g. local, clouds, batch systems etc.
Grid compliant security framework (OpenSSL with X509 certificates).



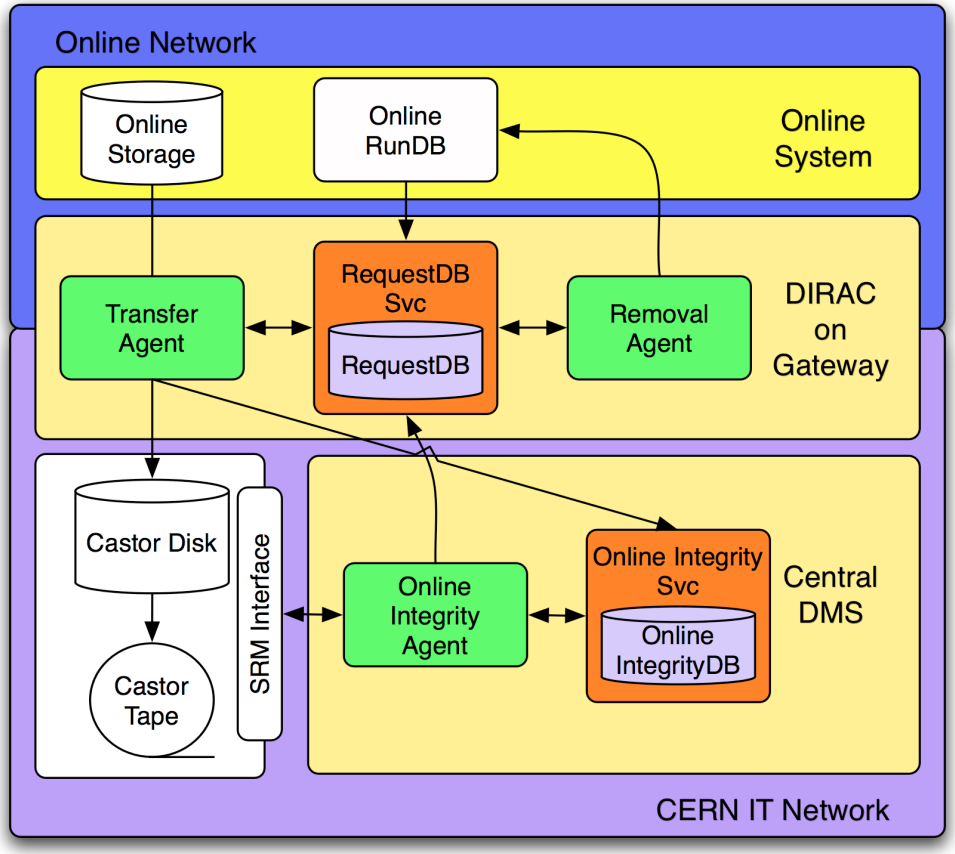
LHCb to Grid

- RAW physics data from LHCb written to Online storage
  - 100 MB/s online data acquisition rate (3KHz)
- Replicate RAW data to Castor at CERN
  - Dedicated 1Gb link per host
- Remove data Online when **'safe'**
  - Correctly migrated to tape
    - Checksum compared against Online file



LHCb to Grid

- DIRAC install at Gateway
  - Replicate RAW data to Castor pools
  - Register data to Online Integrity DB



Persist files Online till **'safe'**

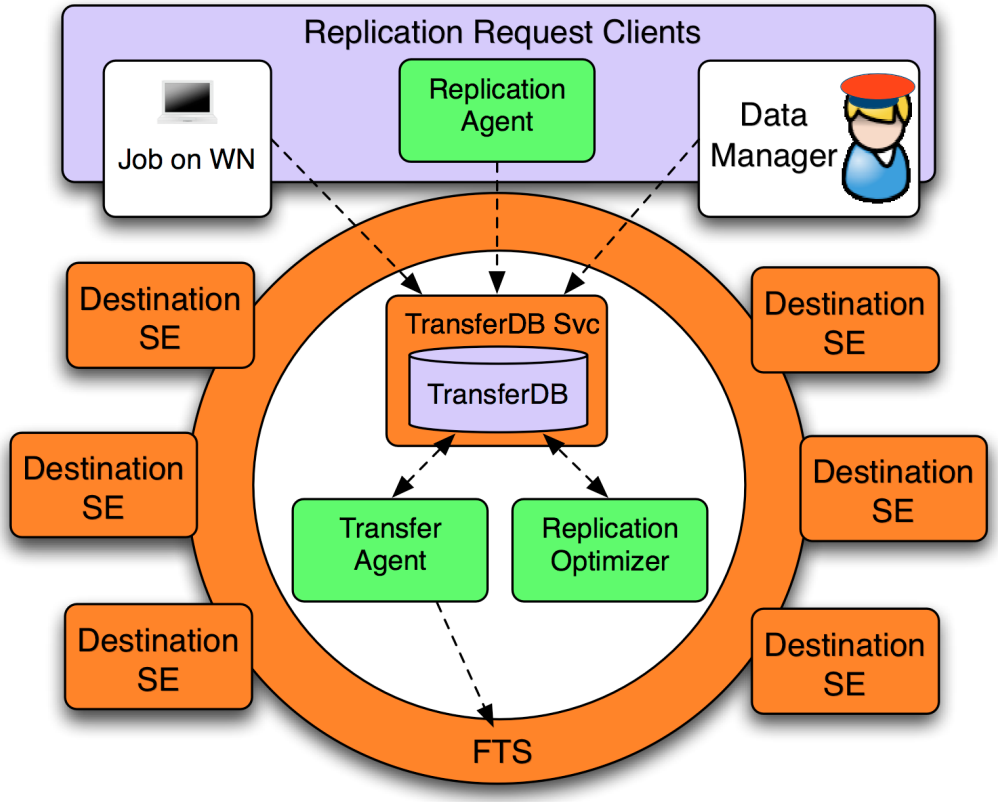
- File checksum calculated online at write time
- File checksum calculated by Castor on migration to tape
- Online Integrity Agent interrogates Castor for file checksum
- If **'safe': place removal** request in DIRAC RequestDB at Gateway
- Pass removal request to Online system

Bulk Data Replication

- Transfer requests centrally managed
  - Maintained in TransferDB
  - Create bulk transfers by aggregating requests

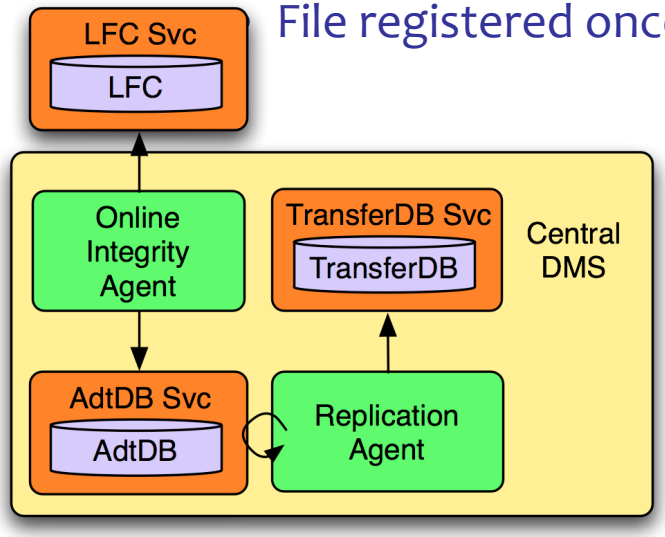
Transfer Agent polls TransferDB

- Obtains bulk transfer requests
- Submits and monitors transfers through FTS CLI
- Requests retried till success



Data Driven RAW Replication

- Dataflow for RAW files
  - Master copy at CERN
  - Replicated at Tier1 site
    - 40 MB/s aggregated out of CERN
  - Reconstruction based on pledged resources
    - Tier1s and CERN
- Data driven replication using AutoDataTransferDB as hook



File registered once **'safely' migrated**

Replication Agent splits files according to site share defined in configuration
- Places transfer requests into TransferDB
- Handled together with other types of transfer request

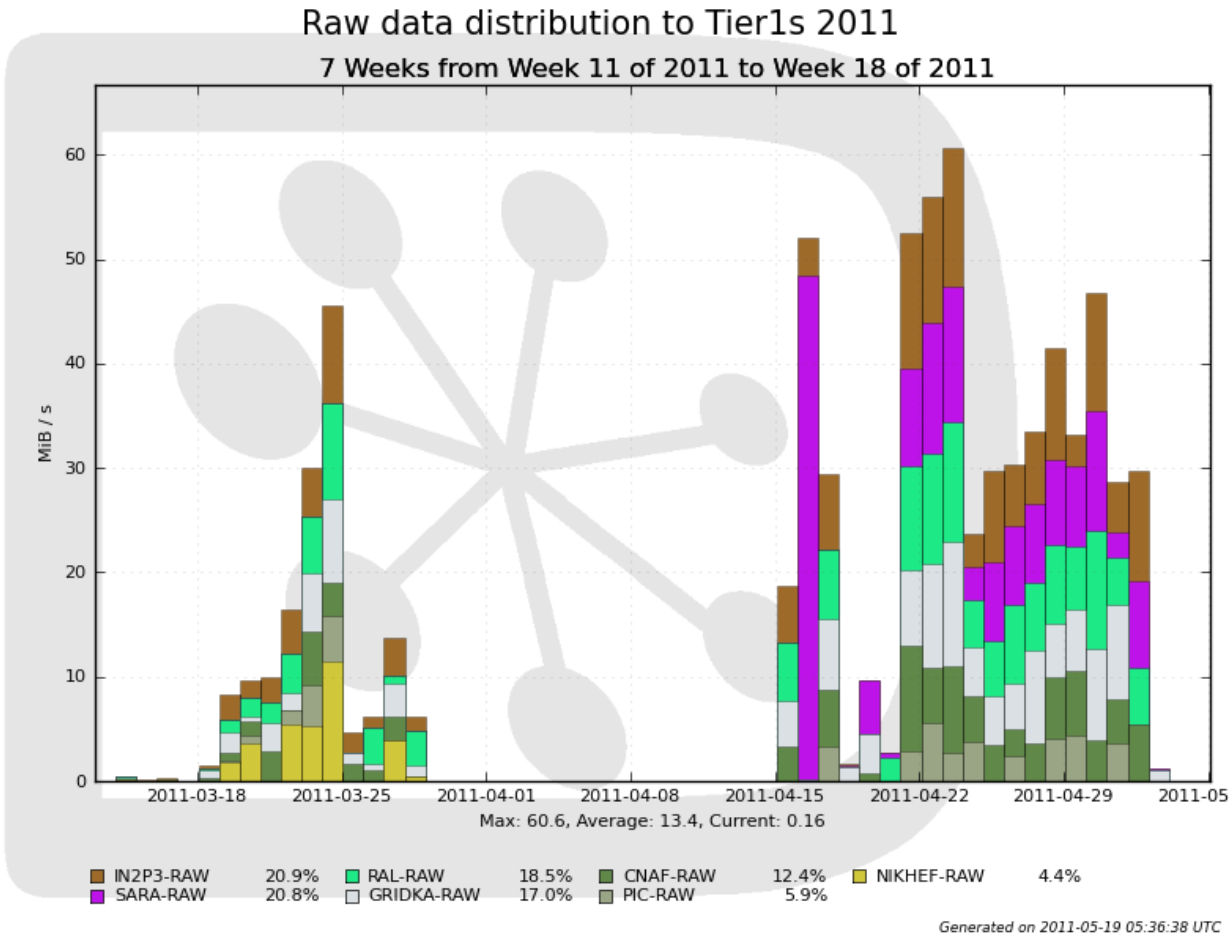
Data Driven RAW Replication

- AutoDataTransferDB (AdtDB) contains simple file catalogue
  - Exposes standard DIRAC file catalogue API
  - Each DM operation specifications contained in DB
    - Source and target SEs
    - File selection (based on LFN namespace)
    - Threshold number of files
- Replication Agent uses AdtDB API
  - Checks for files to be processed
  - Applies selection for desired files
  - Checks file location
  - Files selected if matching operation specification
  - Request created when threshold number of files found
- Replication Agent logic generalized to support multiple operation types
  - Single source to many destinations
  - Splitting files based on configurable share

Data Driven Reconstruction/Stripping

- Similar components exist in the Workload Management System for job creation and submission
  - ProcessingDB
    - Analogous to AdtDB
  - Transformation Agent
    - Analogous to Replication Agent
- Files registered in ProcessingDB may be selected for processing
  - Transformations define specific processing activities
    - Reconstruction/stripping/re-processing
    - Based on file properties
- Transformation Agent groups files for processing
  - Subsequently submits jobs
- Registration of RAW data in AdtDB initializes the chain of replication and data processing

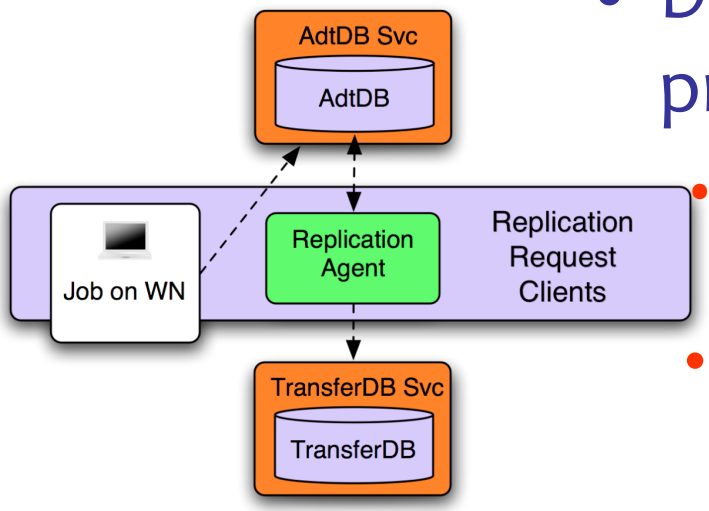
Data Driven RAW Replication



- Surpass computing Model requirements (40MB/s)

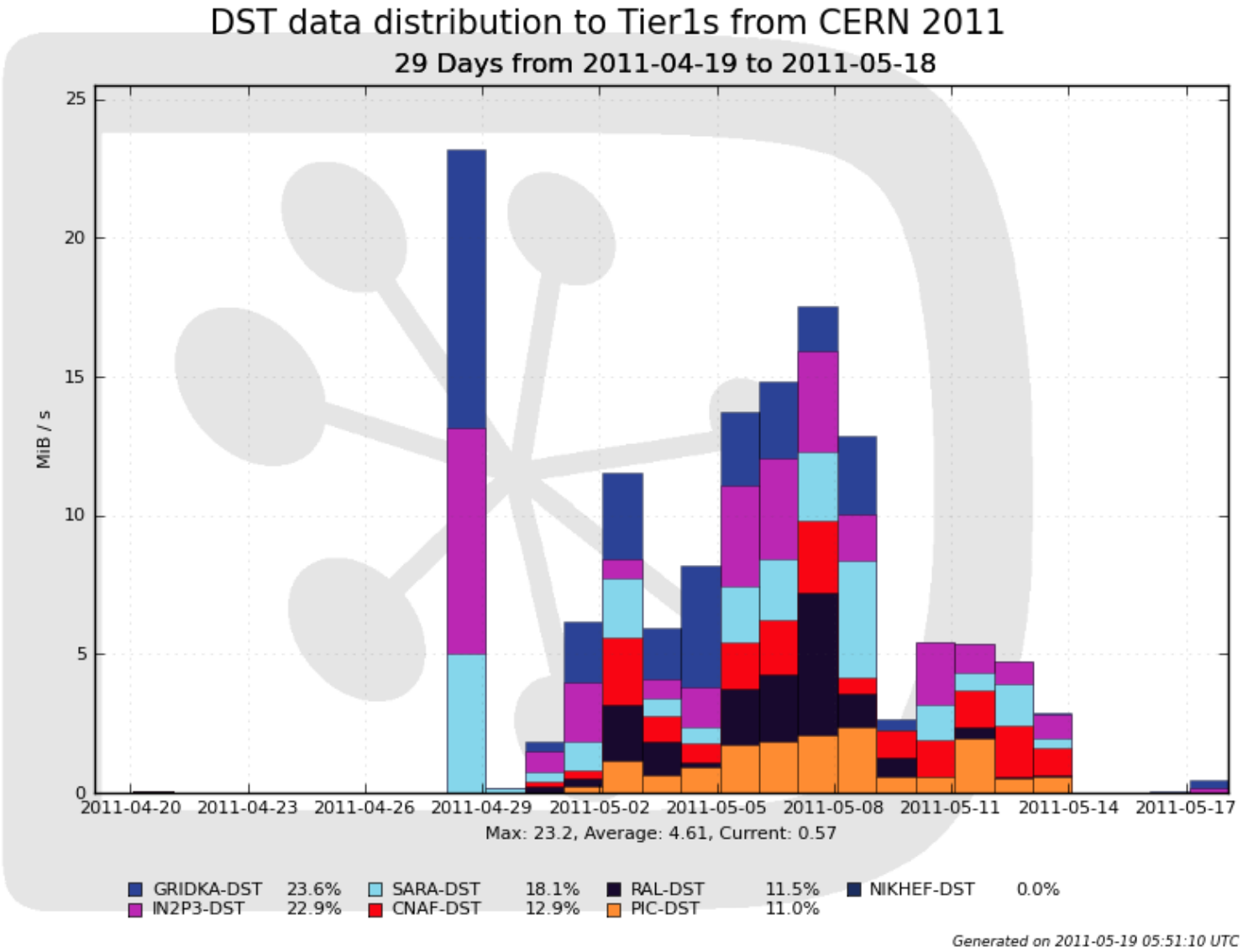
Optimized DST Replication

- Processing activities produce user analysis files (DST)
  - DSTs from real data must (should) be replicated to all (4) Tier1s
  - Average ~11MB/s in and out
  - Shared 10Gb network available using FTS
- DSTs from MC data present on a selection of Tier1s



- Define output SEs in AdtDB per production
  - Job uploads output to SE associated Tier1
    - File registered to AdtDB
  - Replication Agent determines output SEs
    - Defined in configuration per production
    - Places replication request to TransferDB
  - This allows load balancing for MC DSTs

DST 2011 data distribution: CERN → Tier1s



Data Management: Real data

LHCb real data 2010-2011

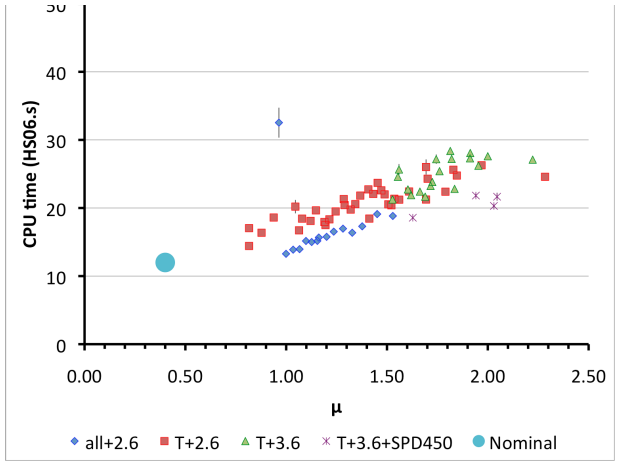
- LHC started with very low luminosity
  - Very few colliding bunches
  - Not worth for rare b-physics decays
    - Minimum bias trigger for 2 months (early 2010)
    - Introduce tighter triggers when luminosity increases
- LHC change of strategy for higher luminosity
  - Large number of protons per bunch
  - Small squeezing
  - Consequence: larger number of collisions per crossing
  - μ=1 to 2.3 !!!
  - Much higher pile-up (1.6 to 2.3 collisions per trigger)
- Effects on Computing
  - Larger events
  - More complex events to reconstruct
  - Larger pre-selection retention

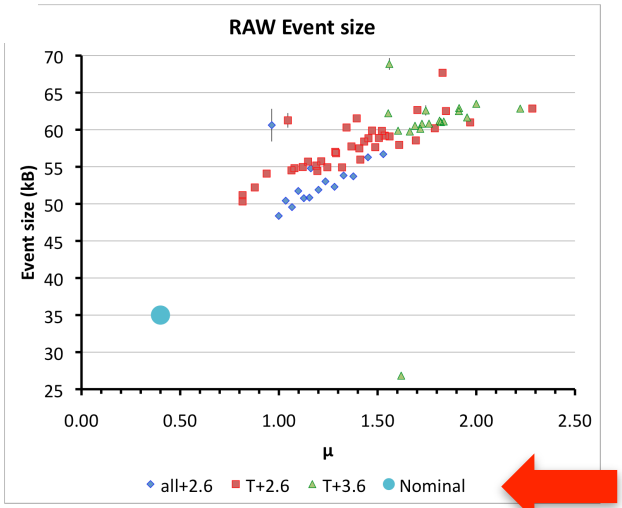
Recent change on Computing Model (I)

- Increase in μ from 0.4 to implies changes to basic parameters of the model:

Process	CPU (HS06·s/evt)		Data Type	Storage (kB/evt)	
	New	Old		New	Old
Data Taking			RAW	**50**	30
Reconstruction	**25**	12	SDST	**40**	25
Stripping	**1.75**	0.8	DST	**130**	80
			MDST	**13**	
Simulation	**1700**	376	DST	**400**	300

Reconstruction CPU time





LHCb decide to fix the working condition to μ=1.5

Different triggers

Recent change on Computing Model (II)

- To accommodate larger events, reduce number of replicas:
  - 4 disk only (T0D1) (secondary) replicas (one at CERN)
  - (was 2 T1D1 (master) plus 5 T0D1 (secondary))
- Archival copy (tape only, T1D0):
  - 2 copies (one at CERN)
    - Plus second CERN copy for RAW (three copies in total for RAW)
  - master copy written directly to archival storage
    - In previous model T1D1 masters migrated to T1D0 at archival time
    - Protection in data management tools to prevent deletion of archive copies
- Previous model prone to errors:
  - Last disk copy could be deleted before archival
  - Confusion between archive copy (never deleted) and T1D1 master copies (can be deleted after archival)

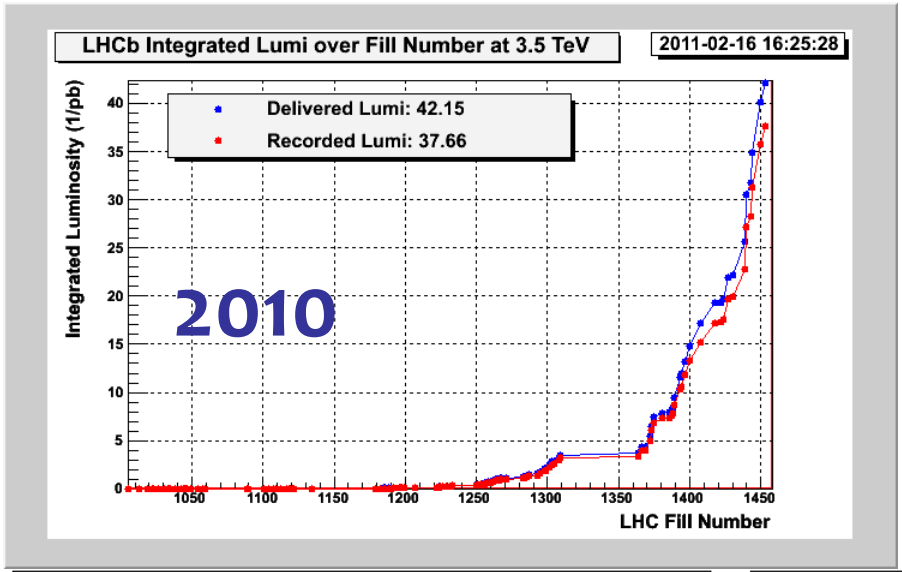
Recent change on Computing Model (III)

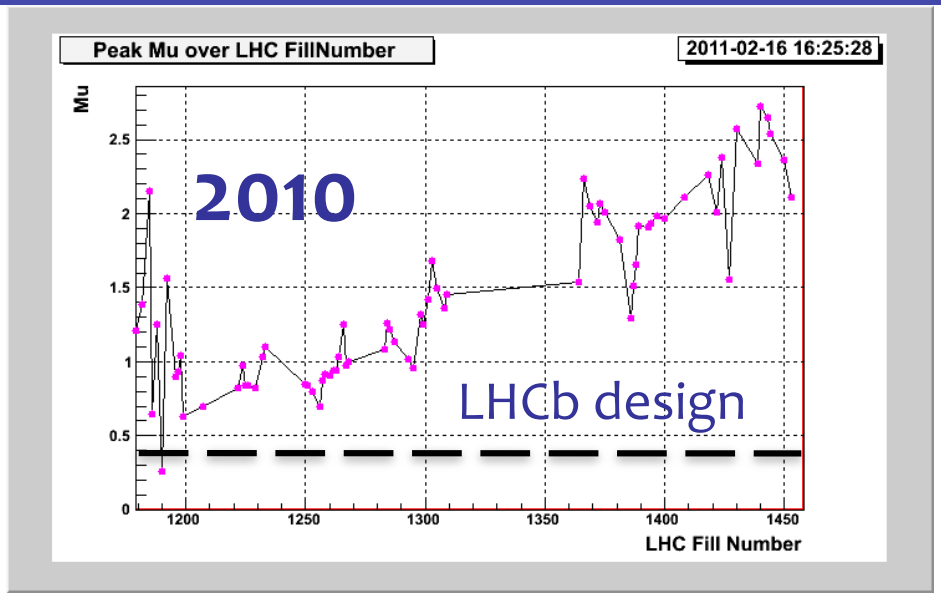
- More aggressive reduction of old data from disk:
  - Keep latest (N) processing and previous (N-1) processing
  - Halve number of disk copies of (N-2) processing
    - This is done BEFORE starting processing "N"
  - Completely remove (N-3) processing
    - Before starting processing N
    - Exceptions only at expense of delaying reprocessing

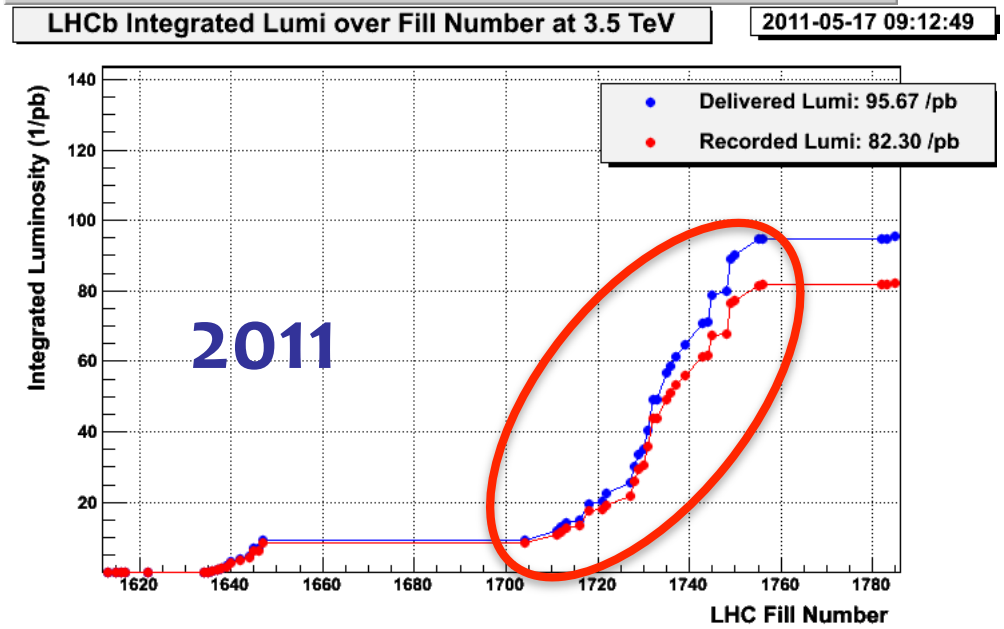
Recent change on Computing Model (IV)

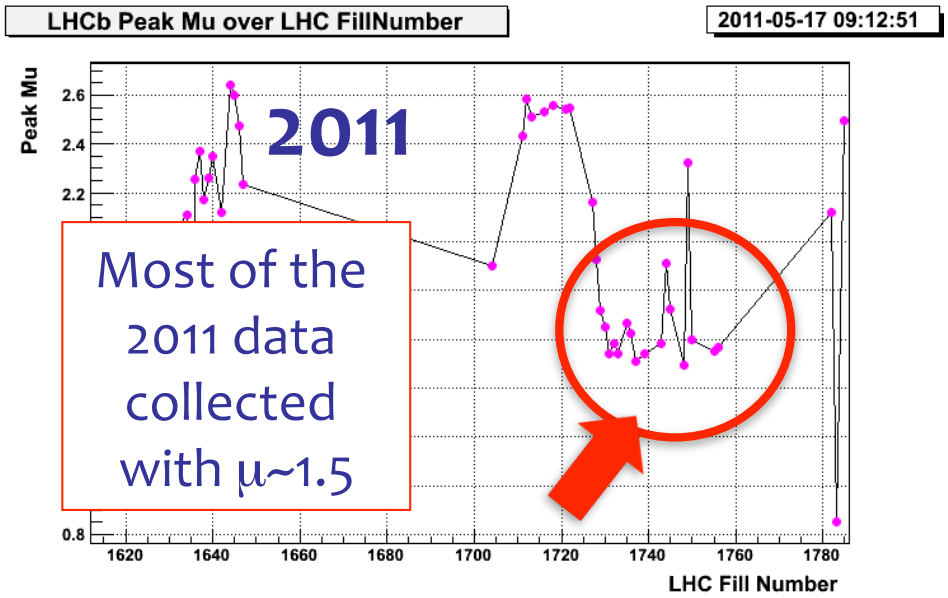
- Baseline: b-physics:
  - 2 kHz of Trigger
  - 10% retention in Stripping
    - RAW: 500 TB/year
    - SDTS: 400 TB/pass
    - DST: 130 TB/pass
- In addition: c-physics
  - 1 kHz of Trigger
  - 10% reduction in size after Stripping
    - RAW: 250 TB/year
    - MDST: 65 TB/pass
- Monte Carlo dedicated to b/c samples
  - DST: 300 TB/year

Data collection









Raw data distribution 2010

LHCb 2010 RAW data		
SE	**Size (TB)**	**# of Files**
CERN-RAW (T1D0)	97.4	87233
CERN-RDST (T1D1)	83.4	76711
CERN	**180.8**	**163944**
CNAF-RAW (T1D0)	19.5	17847
GRIDKA-RAW (T1D0)	27.5	25141
IN2P3-RAW (T1D0)	34.4	31666
NIKHEF-RAW (T1D0)	33.8	31024
PIC-RAW (T1D0)	9.1	8520
RAL-RAW (T1D0)	30.6	27978
Tier1s	**154.9**	**142176**

Only physics data (155 TB in total) is replicated to Tier1s.

RAW data share according to CPU pledges of Tier1s

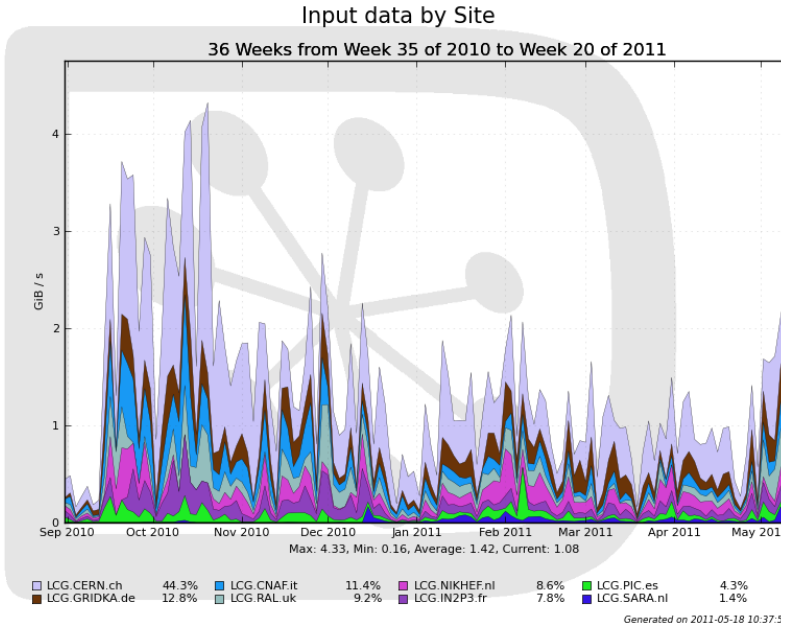
Disk usage (January 2011)

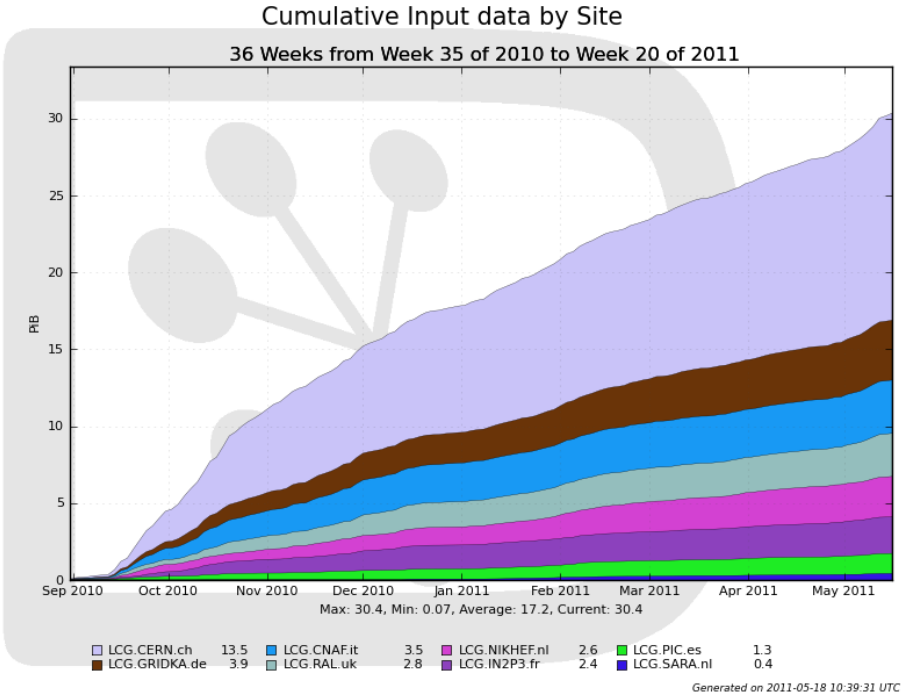
Disk Summary (08/02/2011)	Pledge (TB)	Seen by SLS			Seen by LHCb	
		TB			TB	
		Total	Used	Avail.	Used	Pledge-Used
FZK	495	500	330.2	169.8	336.3	158.7
IN2P3	610	638.5	298.7	338	284.4	325.6
CNAF	450	462.7	386	74.7	384.1	65.9
NL-T1	560	563	350.4	209.6	244.9	315.1
PIC	240	255	149.5	100.5	149.2	90.8
RAL	505	1174.9	453.5	183.9	443.9	505
Tier1s	**2860**	**3594.1**	**2359.7**	**1224.5**	**1840.2**	**1019.8**
CERN	**1135**	**1175.6**	**861.2**	**314.4**	**736.8**	**398.2**

NB: Good agreement between what is really on disk (386TB) and what LHCb seen (384.1) → big discrepancy observed on other sites due to lowest reliability of others storage system (CERN and RAL with castor but also NIKHEF with dcahce)

Data analysis managment

User job: input data





CERN

GRIDKA
CNAF
RAL
NIKHEF
IN2P3
PIC
SARA

16 PiB analyzed since September 2010.

Peak of 3-4 GiB/s achieved in September 2010 considering all the lhcb user jobs

User quota

- User grid storage is part of total Tier1 disk storage allocation
  - Competes with space for production data
  - Should be kept under control
- Up to now, no quotas
  - 41 users > 2TB
  - 17 users > 5TB
  - 6 users > 10TB
  - 1 user > 60TB (more than Reco DST!)
- "Enforce" quota of 2TB / user
    - (2 replicas of 1TB each)
  - Currently just mail warnings, but eventually block job submission if more than 50% overquota
    - Experience is that users are cooperative
  - Special arrangements if good reason for large quota need
    - e.g. datasets for group analysis, should be justified by convener with timescale for extended quota

Conclusions

- LHCb Data Management is well integrated within DIRAC project
  - all the tools provided by DIRAC worked well so far and satisfied the Collaboration needs
  - starting now the discussion about dynamic transfer of data based on popularity measurements
- The experience on real data shows the necessity to update the computing model
  - LHC reached the LHCb designed instantaneous luminosity with a number of visible interactions per beam crossing higher than what is expected
    - LHCb design μ=0.4, today μ~1.5
  - Main points:
    - reduced the DST replicas (from all Tier1s to 4 Tier1s)
    - More aggressive reduction of old data from disk
  - Increase of resources needs (in particular disk)