LHCb data at T1 centers

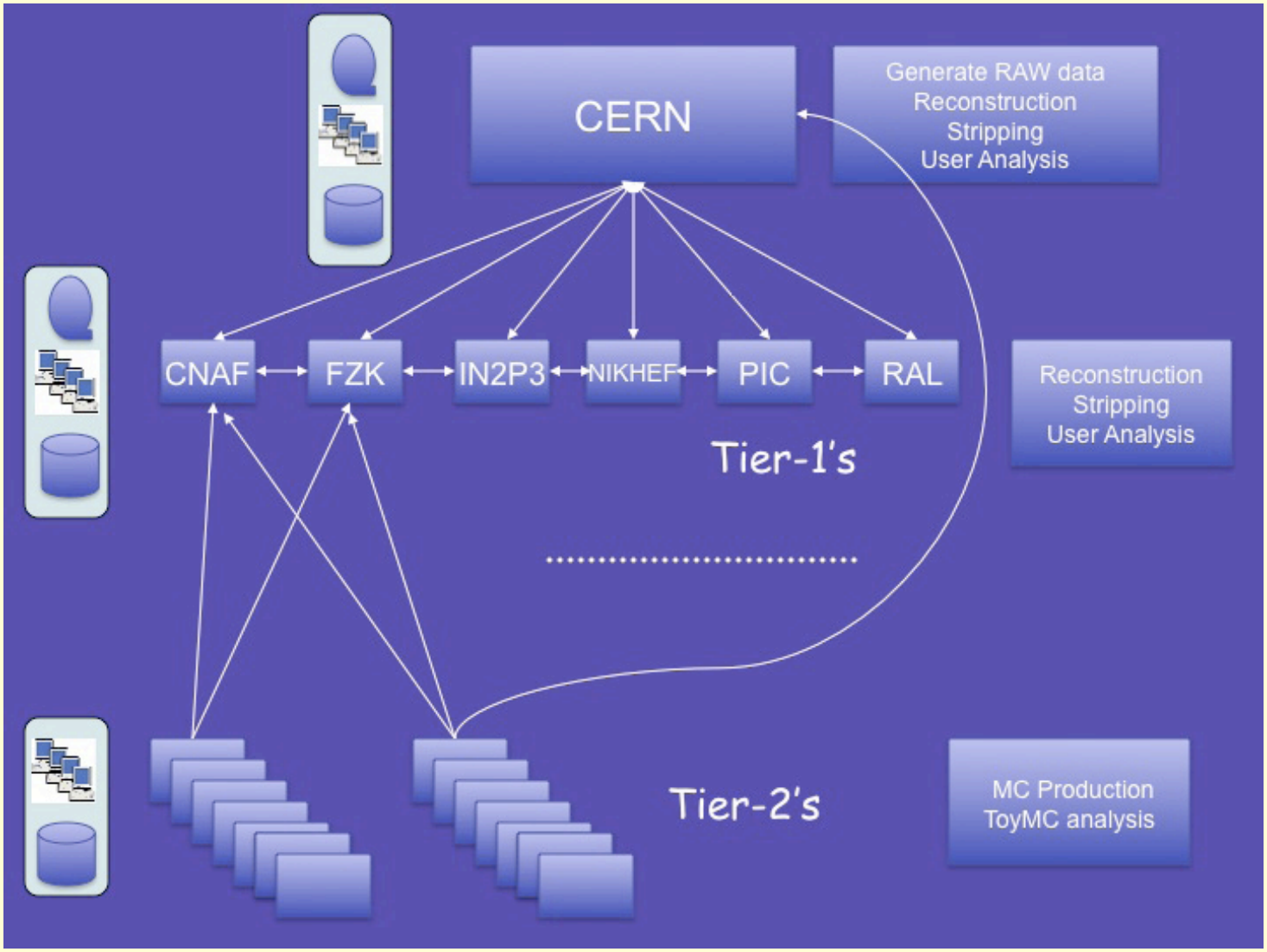
A.Tsaregorodtsev,
CPPM-IN2P3-CNRS, Marseille

16 October 2009, Journées "Grilles France”, IPNL, Lyon

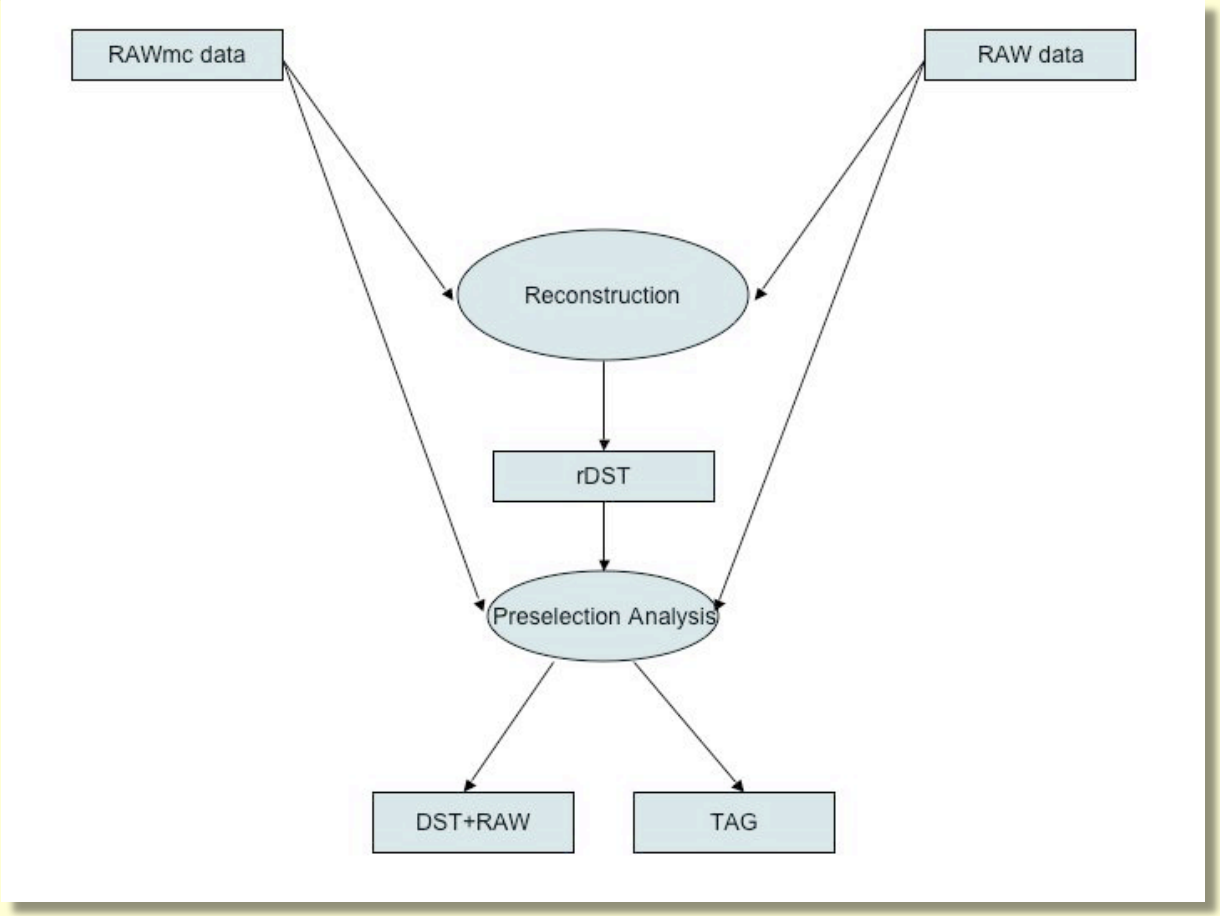
Outline

- LHCb, Computing and Data Model
- Data Distribution
- Storage monitoring
- Data Integrity
- Data Access

Computing Model

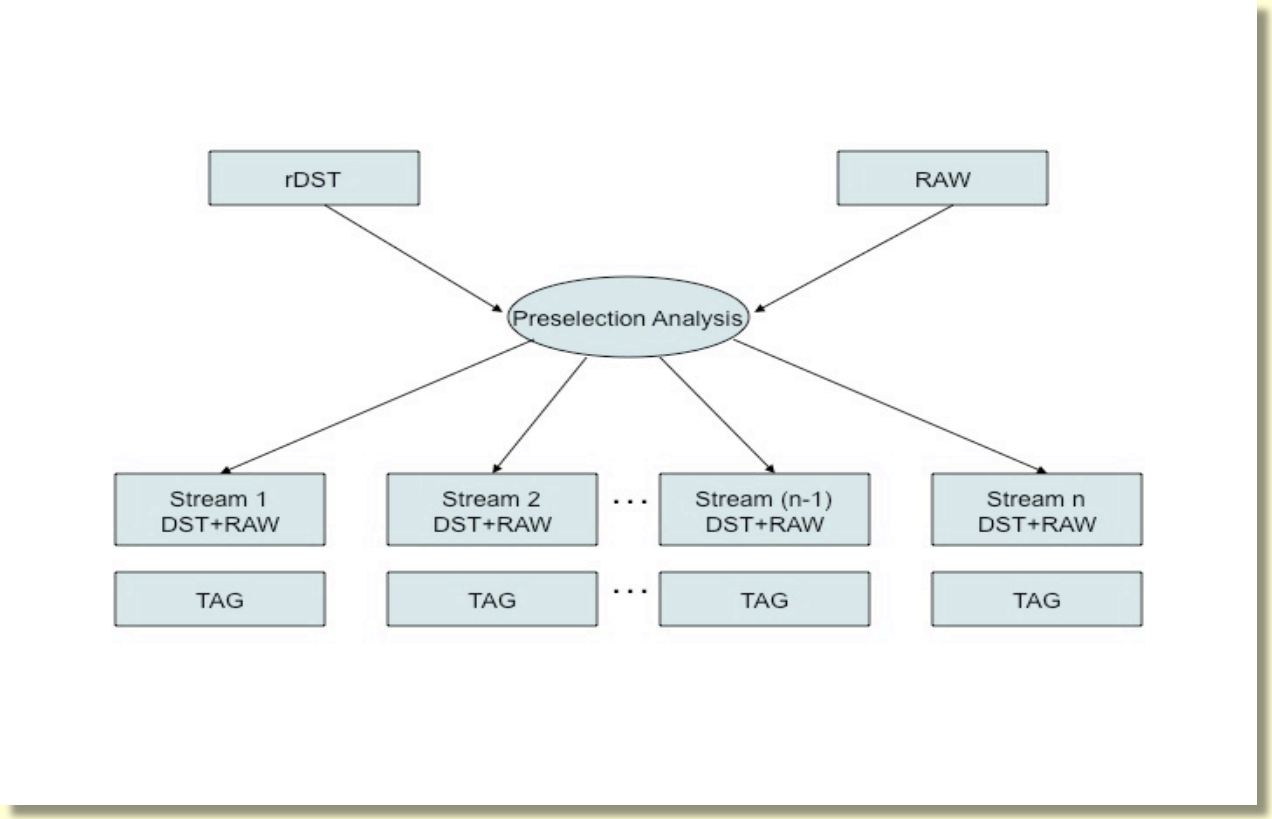


LHCb Data Model (1)



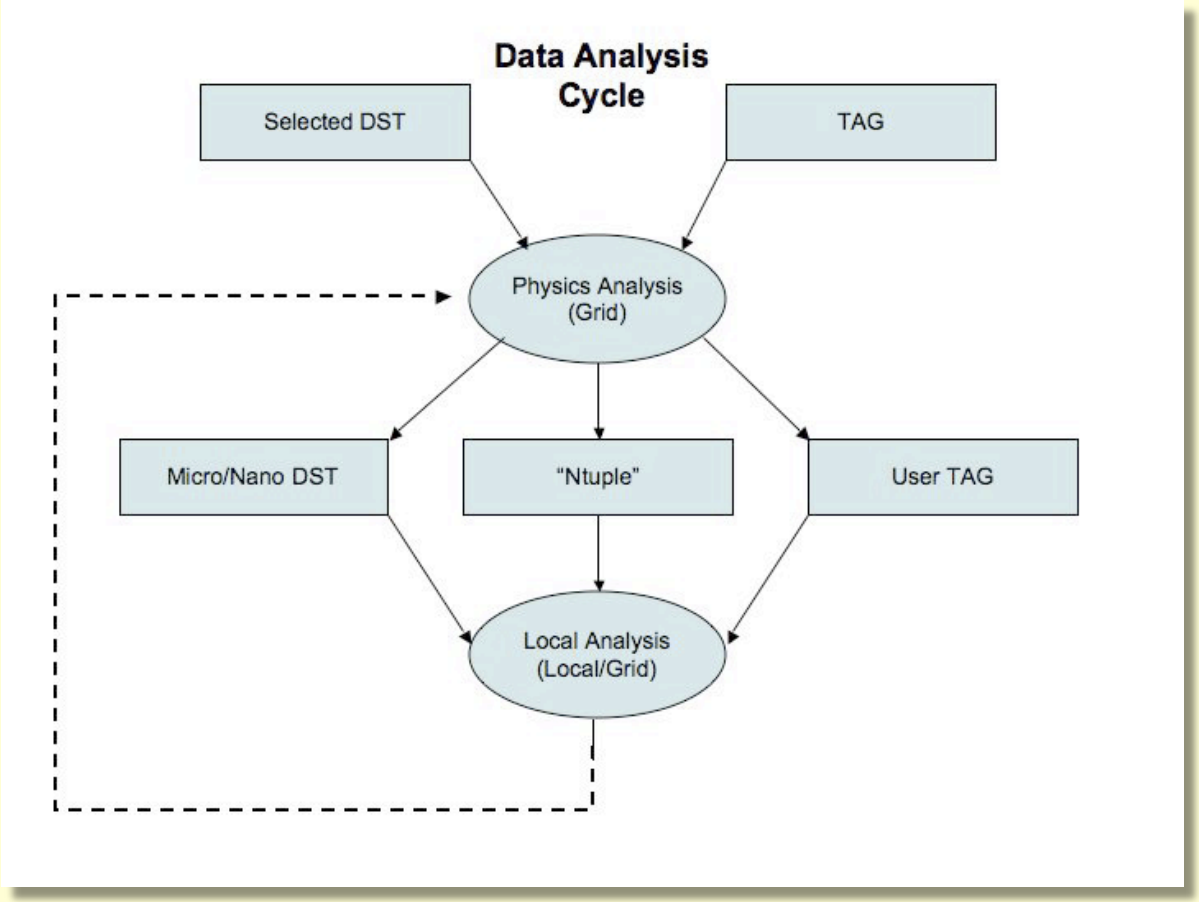
- rDST (reduced DST) – only objects allowing to preselect events for analysis

LHCb Data Model (2)



- DST – preselected events with RAW information included – main analysis input
- TAG – tuples for fast event selection

LHCb Data Model (3)



- Micro/Nano DST – reduced number of objects suitable for particular user analysis

Event statistics

Event type	Real Data, kB	MC Data (b), kB	MC data (non-b), kB
RAW	35	250	10
rDST	35	35	15
DST	50	50	20

Events	2009	2010
Real data	$5x10^5$	$6x10^6$
MC data (b)	$5x10^8$	10^9
MC data (non-b)	$3x10^9$	$2x10^9$

Data Processing

- 3 reconstruction passes per year
  - 1 copy for each pass of rDST data is kept on tape
- 4 (or more in 2010) stripping passes per year
  - Copies of DST data for the last 2 passes on disk at all the 7 T1 centers
  - A copy for each pass is archived on tape at CERN and one T1 center

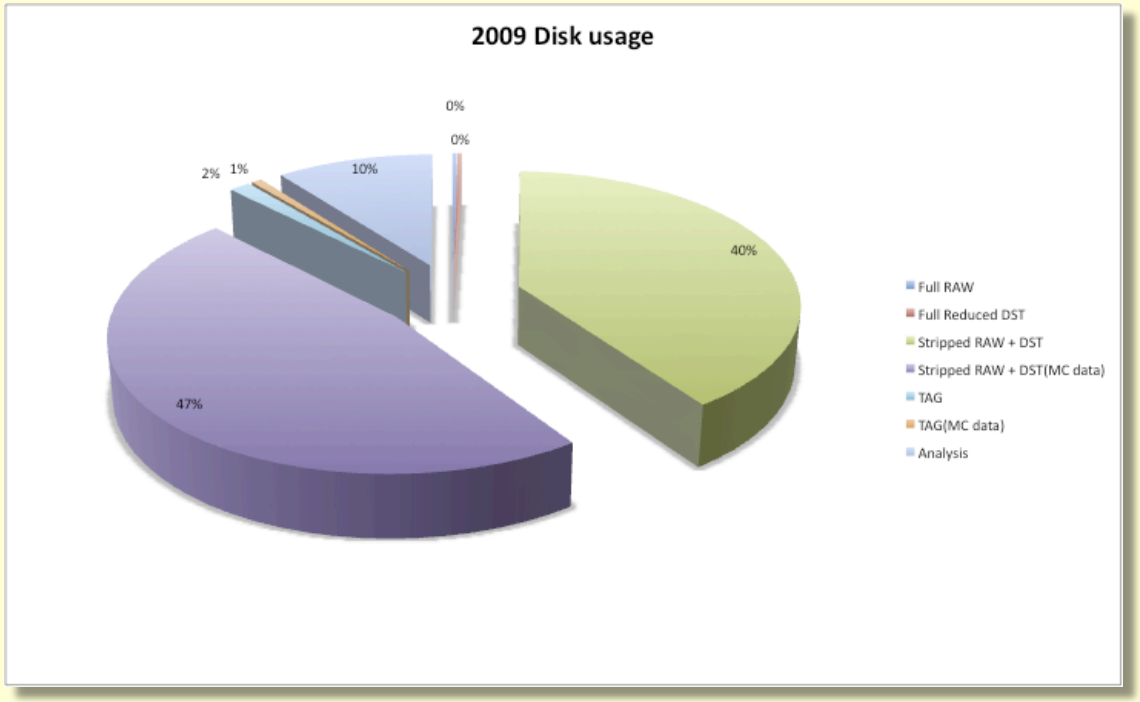
Storage types and requirements

- 2009 illustration

Space tokens	Type	Total T0-T1, TB 2009	T1-IN2P3 TB 2009
LHCb-RAW	T1D0	175	15
LHCb-RDST	T1D0	110	20
LHCb_M-DST	T1D1	140	30
LHCb-DST	T0D1	585	85
LHCb_MC_M-DST	T1D1	655	65
LHCb_MC_DST	T0D1	470	125
LHCb-FAILOVER	T0D1	20	5
LHCb-USER	T0D1	230	30

Storage requirements

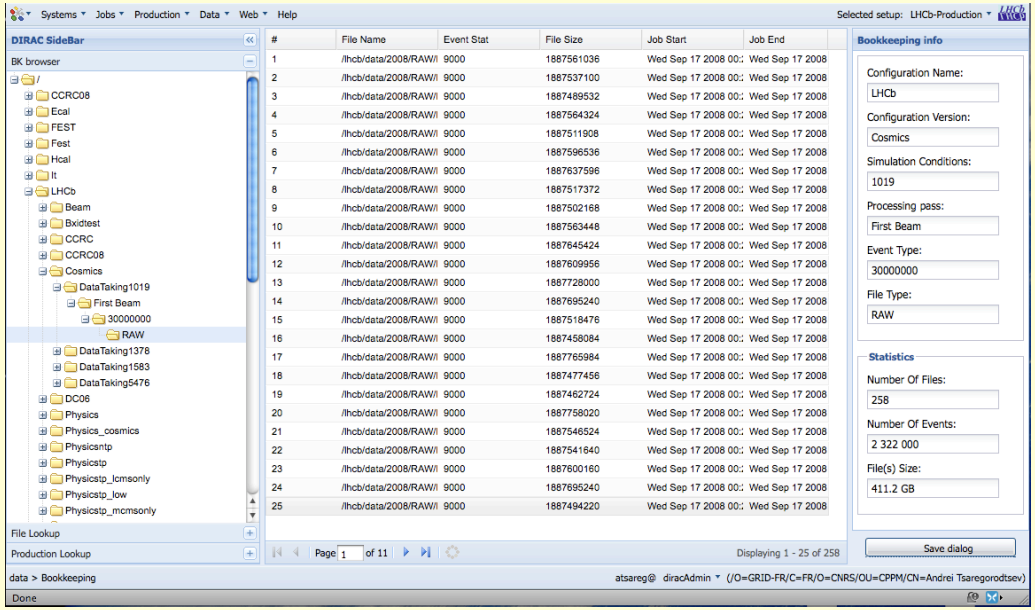
	Disk, PB 2009	Tape, PB 2009	Disk, PB 2010	Tape, PB 2010
T0(CERN)	0.63	0.49	1.22	1.64
All T1s	1.58	0.45	2.87	2.06



- Storage requirements are dominated by the MC data needs

Data bookkeeping

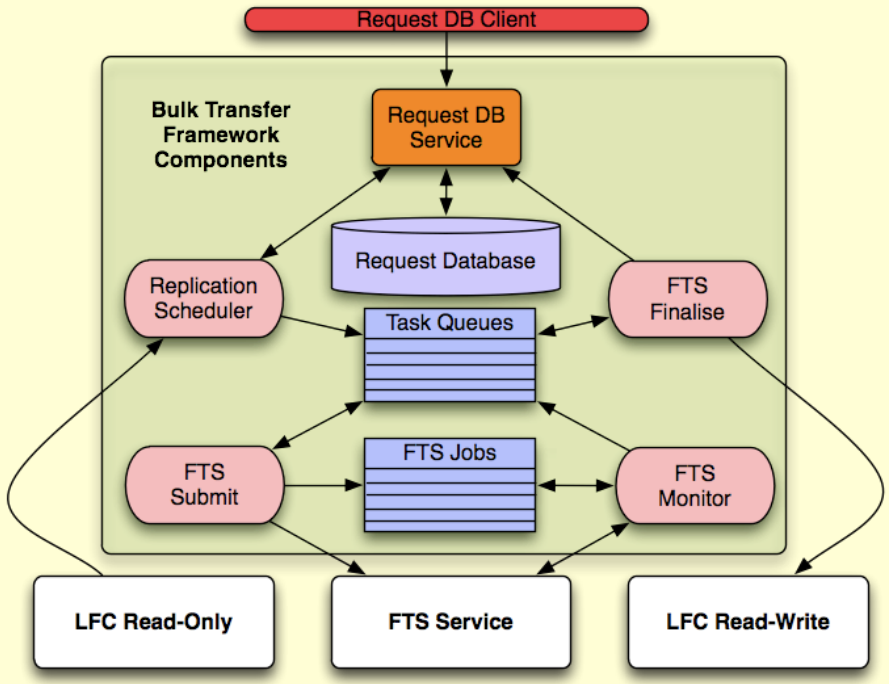
- Central Bookkeeping database
  - Metadata catalog
  - Allows users to select the desired data sets
  - DIRAC service with an ORACLE backend



- LFC File Catalog
  - Replica Catalog
  - 1 write/read master at CERN + 6 read-only mirrors at 1 centers
    - Synchronized via ORACLE streams
  - Seen as a single redundant service for clients

Data Management System

- All the Data Distribution operations
  - Pit to CERN transfers
  - T0-T1 transfers
  - T1-T1 transfers
- Based on the Request and Production Management Systems
  - Automatic transfer scheduling
  - Full monitoring of ongoing operations
- Using FTS for bulk data transfers
  - Full failure recovery
- Comprehensive checks of data integrity in SEs and File Catalogs



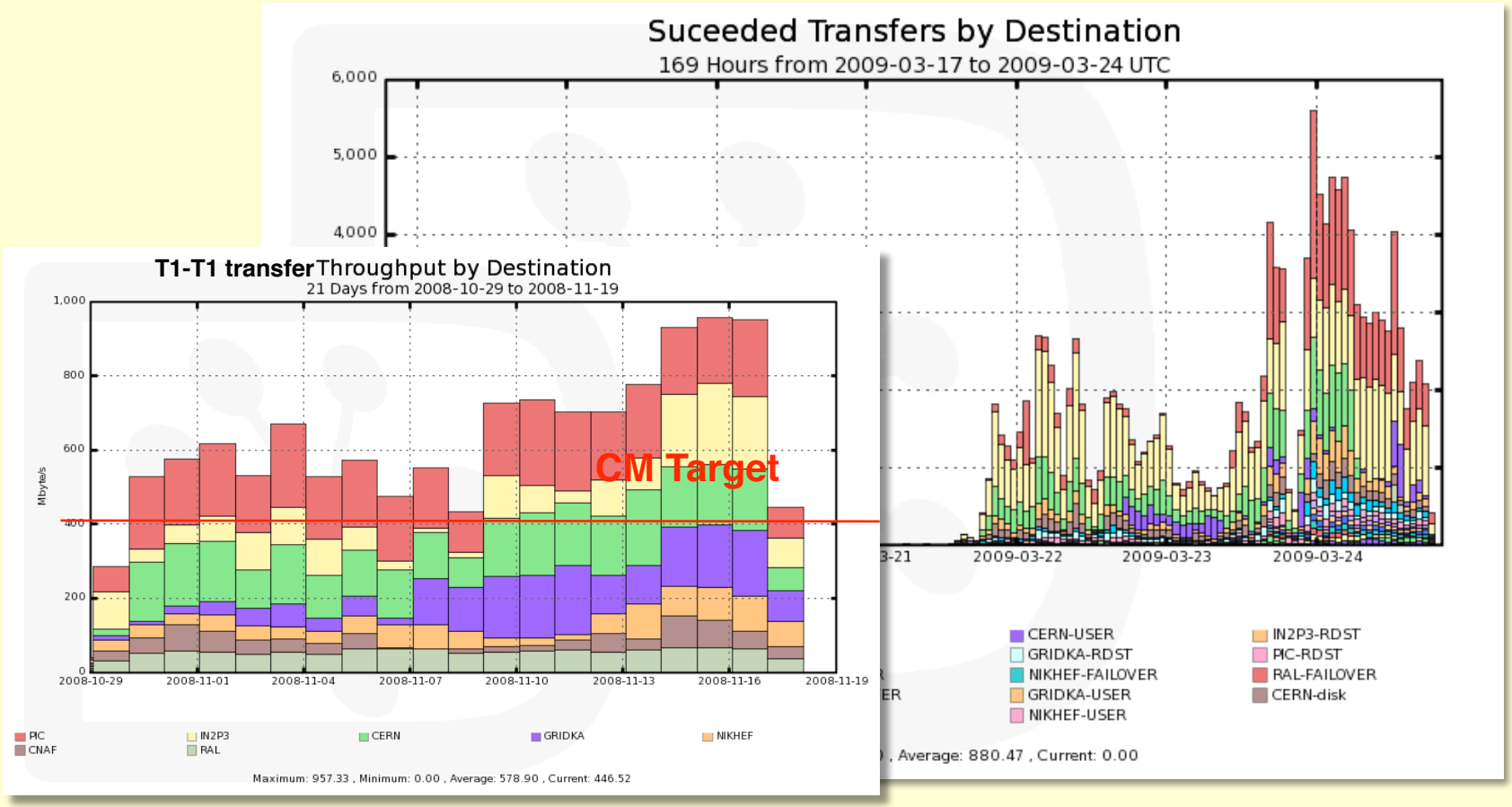
Data Distribution (1)

- LHCb Online Farm output exported in real time to CERN/Castor storage
  - RAW data - Express stream
    - Fast treatment at CERN by the Data Quality (DQ) group
  - RAW data - Full stream
    - Thorough integrity checking
      - ➡checksums after migration to tape
- Full stream exported to T1 centers as soon as approved by the DQ procedures
  - Respecting T1 site shares

Data Distribution (2)

- rDST (reconstructed) data archived on tape at CERN as well as at the production T1 center
  - To allow extra stripping passes as necessary
- DST (analysis) data replicated to all the 7 T1 centers
  - To maximize resources for the end user analysis
  - One copy is archived on tape at CERN also

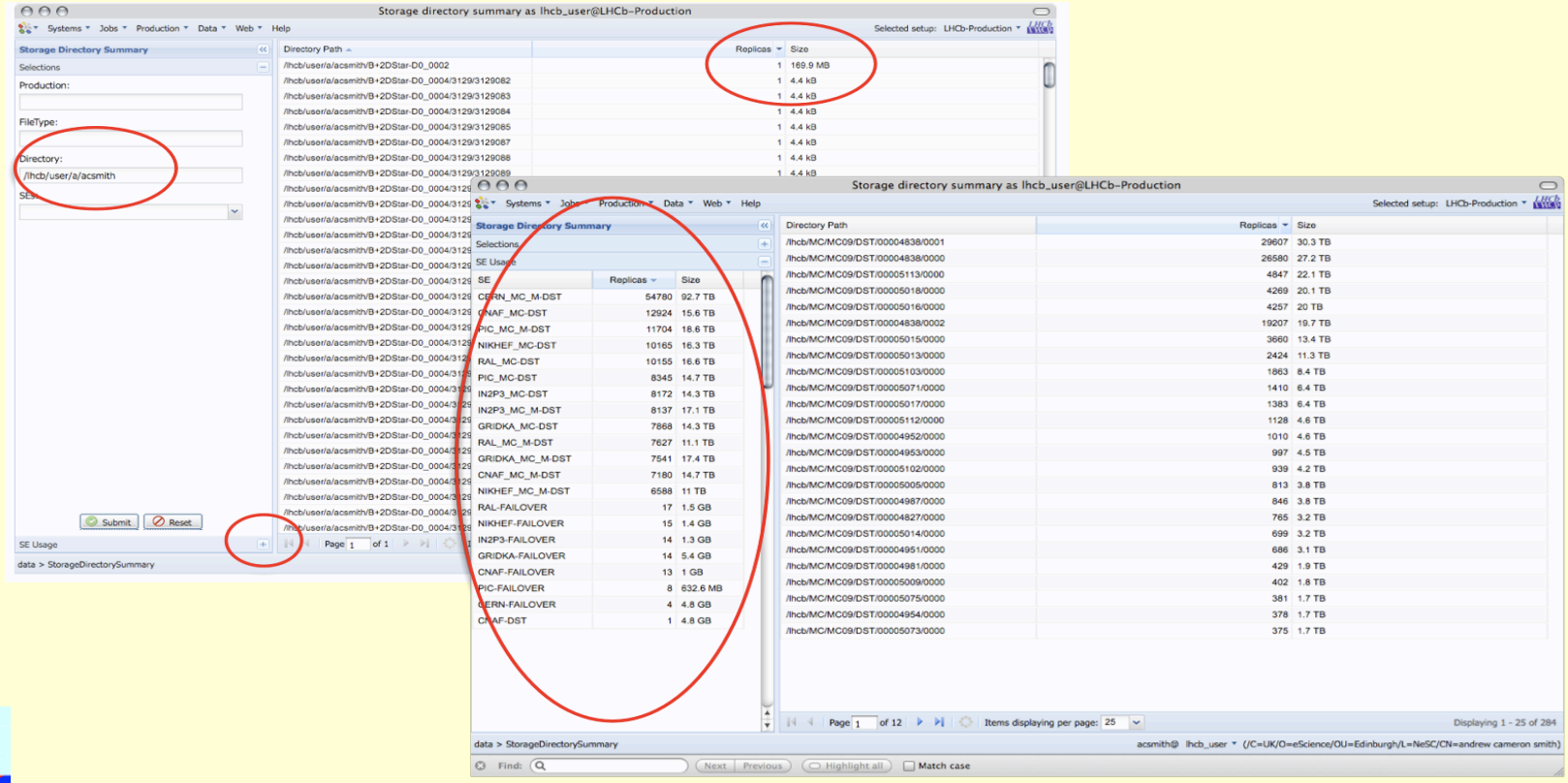
Data moving performance



- Extensively tested in a serious of tests (CCRC, FEST'09, …)
  - Proven to support the LHCb Computing Model targets

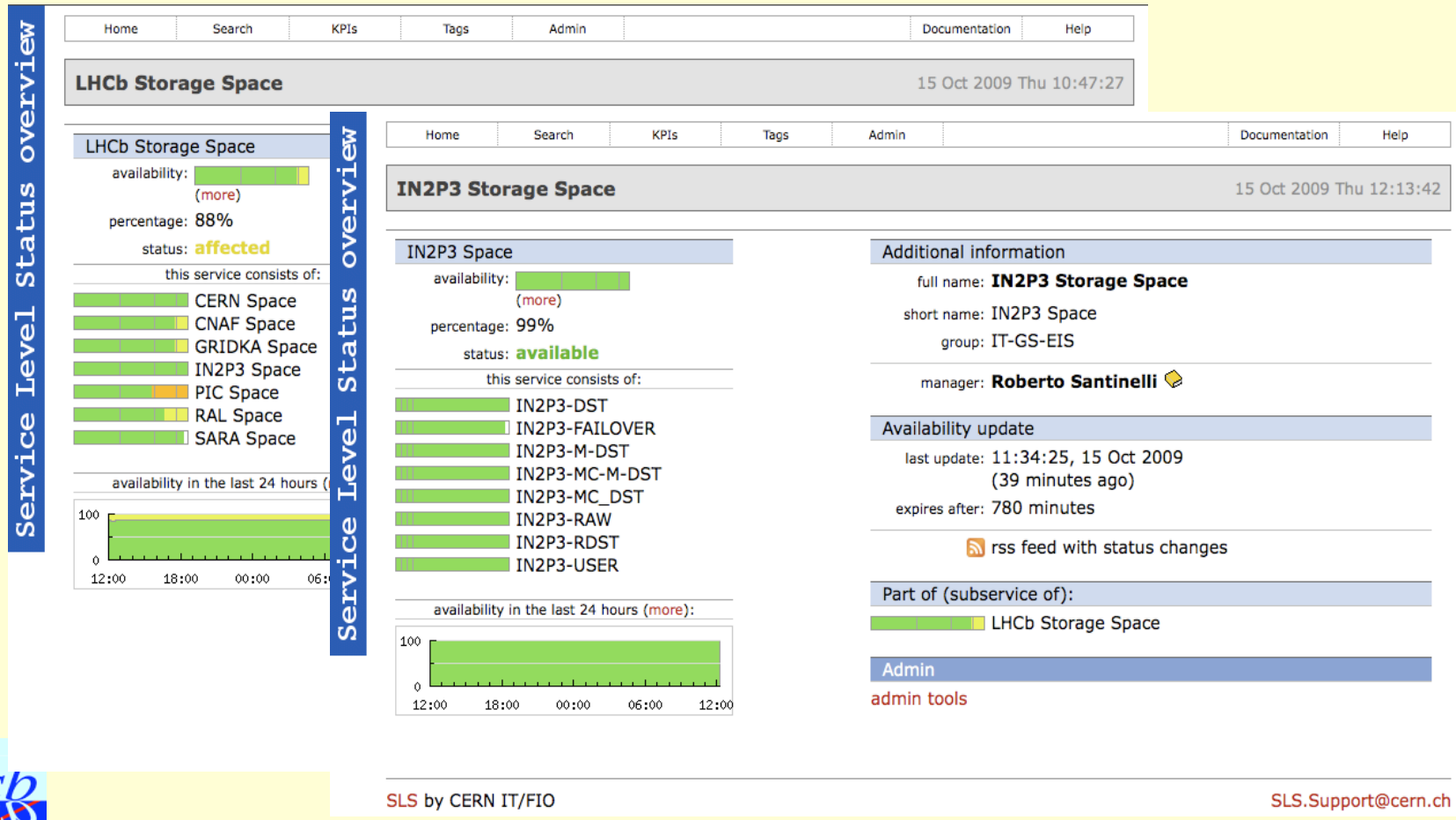
Data Mgmt: Storage monitoring (1)

- Permanent Storage Usage monitoring based on the LFC information
  - Per Space Token
  - Per logical name space directory



Data Mgmt: Storage monitoring (2)

- SLS sensors
  - Storage allocations, Physical Storage usage
  - Alarms are sent in case of misbehaving or approaching the limits



Data Integrity (1)

- Data Integrity can be broken in a variety of ways
  - Catalog corruptions or registration failures
  - Physical storage failures
  - Human errors
  - Everything above and even more happens
- Data Integrity needs permanent monitoring
  - Spotting problems before they are hitting users

Data Integrity (2)

- Production data validation
  - LFC <-> Bookkeeping
    - Relatively simple, needs efficient bulk catalog queries
  - LFC <-> SE
    - Needs efficient inspection of the SE namespace
    - srmLs is not very useful currently
      - ➡ Asking site managers for dumps of the storage name space
- Incidents with the data access are reported to a specialized Integrity DB
  - Automatic agents or human intervention for the incident resolutions

Data access: prestaging

- Part of the Workload Management System
- Marks jobs for execution only once the data is brought on-line from tape
  - Issues SRM “bring online” requests, waits for their execution
  - Fails jobs unable to get data online within a predefined time interval
    - Possibly reassigning jobs to other sites having the required data
- Current evolution
  - Site disk cache management using SRM file “pinning” facility
  - Throttling jobs with high I/O requirements to avoid site I/O systems collapse

Data access at the WN

- LHCb model: data access from the WN through SRM:
  - Remote access by obtained TURL:
    - rfio, (gsi)dcap, xrootd
- During the STEP'09 had to download file local as remote access was very unstable
  - Not a long term solution
- Considering Xrootd
  - Xrootd sharing disk cache with dCache does not seem to have advantages
  - Dedicated Xrootd server
    - Can not obtain TURL from SRM
    - Security

Data access problems (1)

- File locality at dCache sites
  - “Nearline” reported even after BringOnline (IN2P3/SARA)
- SRM overloads (all)
- gsidcap access problem (incompatibility with ROOT plugin)
  - Fixed by quick release of dcache_client (and our deployment)
- SRM spaces configuration problems
  - Fixed at site, need for a migration of files (CNAF)
- Massive files loss at CERN and RAL
  - 7,000 files definitely lost (no replicas anywhere else)
  - Others could be located and replicated back to CERN
- Slowness observed deleting data at CERN (race condition with multiple stagers)
- Hardware reliability: sites need to be able to quickly give VOs the list of files that are affected by hardware / disk-server problems.
- On CASTOR sites globus_xio error rising when gridftp servers exhaust connections and new ones cannot be honored (in case a client is abruptly killed)
  - (script in place to monitor and keep tidy gridftp servers)

Data access problems (2)

- Firewall issue in the file server causing jobs to not receive back data connection remaining stuck (IN2p3).
- Sites should follow dCache.org and WLCG prescriptions regarding versions than gLite releases
- dCache pool which got stuck and could not process any request. (PIC)
- dcap movers to be cleaned up (GridKA/SARA/IN2P3)
- Mis-configuration on the number of slots per server (SARA)
- Not adequately dimensioned servers with too few slots/connections defined per server
  - sites should consider 2 requests: the amount of disk requests AND the necessary number of disk servers for serving all jobs _and_ for allowing redundancy, i.e. always more than one server on T1Dx spaces to allow recalling from tape missing file if a server is down.
- In general when the client is killed (whatever the reason), dcap does not close the connection with the server, which remains pending orphan. This reduces the number of available slots, which makes the lack of available slots issue to become even worse (and the vicious circle is started).

Data Mgmt: measures to take

- Main problems are mostly related to various site misconfiguration problems
  - All sites should increase the number of slots per server to a reasonable number (several 100's depending on the size of disk-servers)
  - All storage services must be adequately dimensioned for supporting peaks of activities
  - Improving the monitoring tools on Storage Service to minimize the occurrences of these annoying incidents

Data Mgmt: Banning faulty SEs

- Storage Elements can be unavailalbe
  - Failures, scheduled or unscheduled shutdowns
- This should be taken into account
  - While job scheduling
    - User and production jobs
- SEs now can be declared as unavailable
  - This is equivalent to banning sites for jobs needing input data on these sites
  - Banning specifically for Read or Write access
  - For jobs without input data, the sites are still available

LHCb data at T2-T3 centers

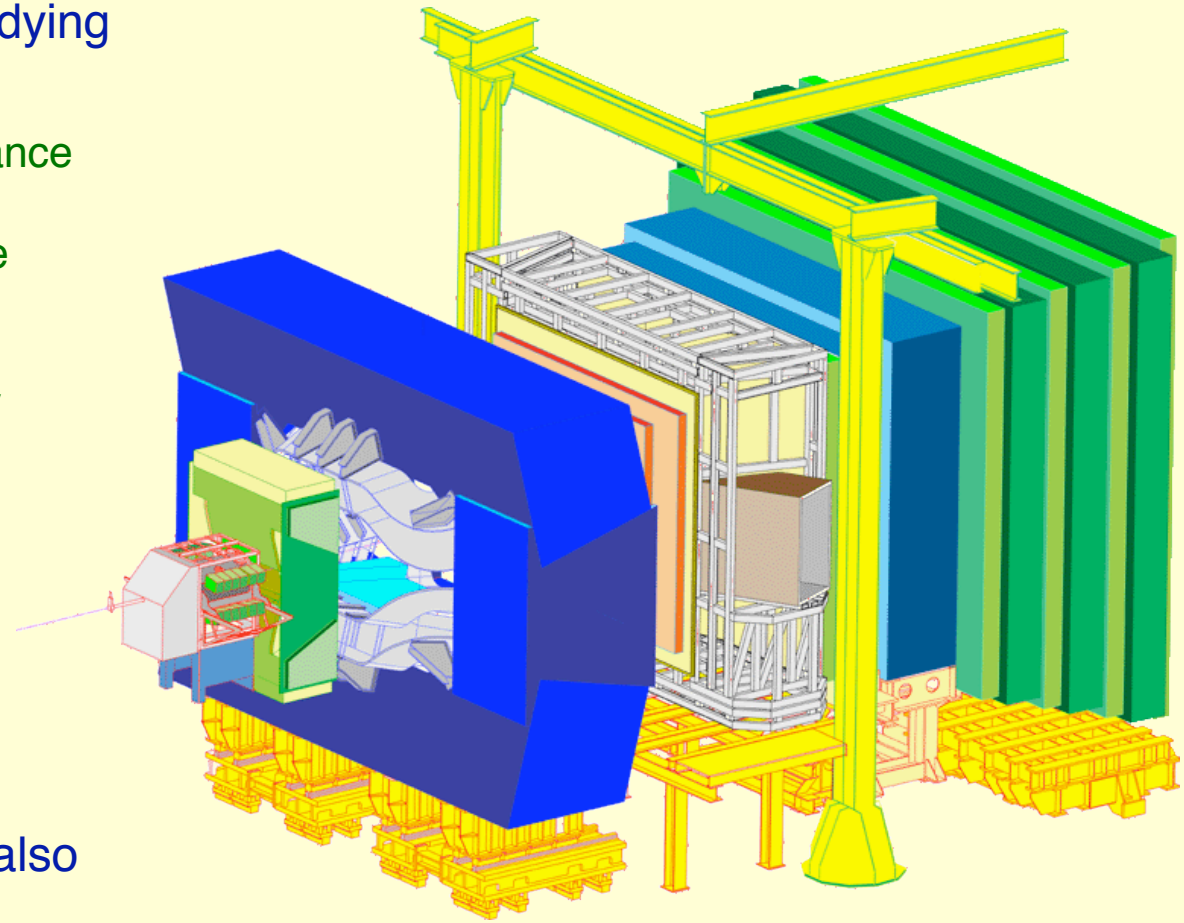
- Now LHCb analysis data at T2-T3 according to the mainstream Computing Model
- Still there are storage resources at some T2-T3 centers available to LHCb users
  - Data can be replicated using standard LHCb tools
  - The data usage is opportunistic – if the site is available, more resources available for the analysis jobs
- Grid storage at T2-T3 centers can be accessed directly by the local users from non-grid CPUs

Conclusions

- Preparation of 2009-2010 data taking is going on
  - Simulating running full RAW data steam
  - FEST regular activities
- Data moving (Pit-T0,T0-T1,T1-T1) is in good shape
- Data access issues and instabilities of services are still the main problem.
- Site storage systems misconfiguration problems are being cleaned up slowly

LHC*b* in brief

- Experiment dedicated to studying CP-violation
  - Responsible for the dominance of matter on antimatter
  - Matter-antimatter difference studied using the b-quark (beauty)
  - High precision physics (tiny difference…)
- Single arm spectrometer
  - Looks like a fixed-target experiment
  - Smallest of the 4 big LHC experiments
  - ~500 physicists
- Nevertheless, computing is also a challenge….



DMS: User Storage quotas

- Storage space on the grid is not unlimited
  - Users are supposed to clean their spaces but rarely do
    - Unless they are notified about exceeding quotas
- The user storage consumption is periodically checked by a dedicated agent
  - The results are available to users
  - Currently they can be just consulted
    - Command line and API tools available
  - Eventually the user space will be locked for writing if the quotas are exceeded
    - The quotas are defined in the CS per user