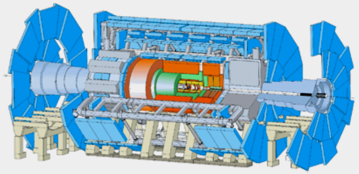
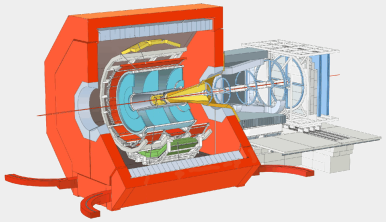
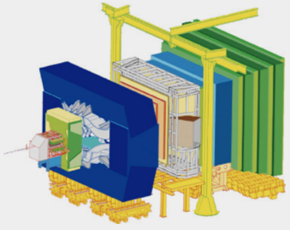
AGGIORNAMENTO SUI REQUIREMENT DI STORAGE DEGLI ESPERIMENTI

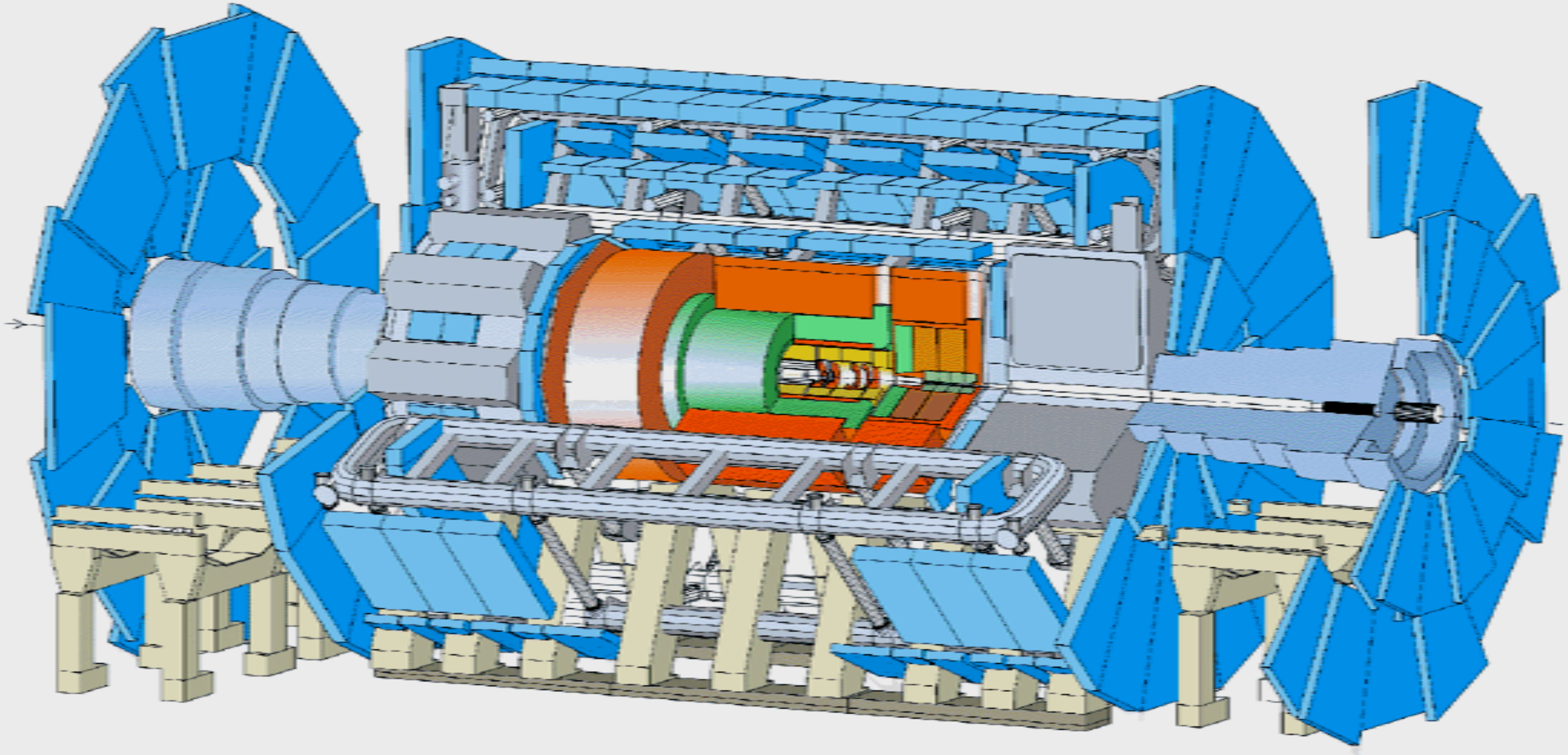
Stefano Bagnasco

INFN Torino

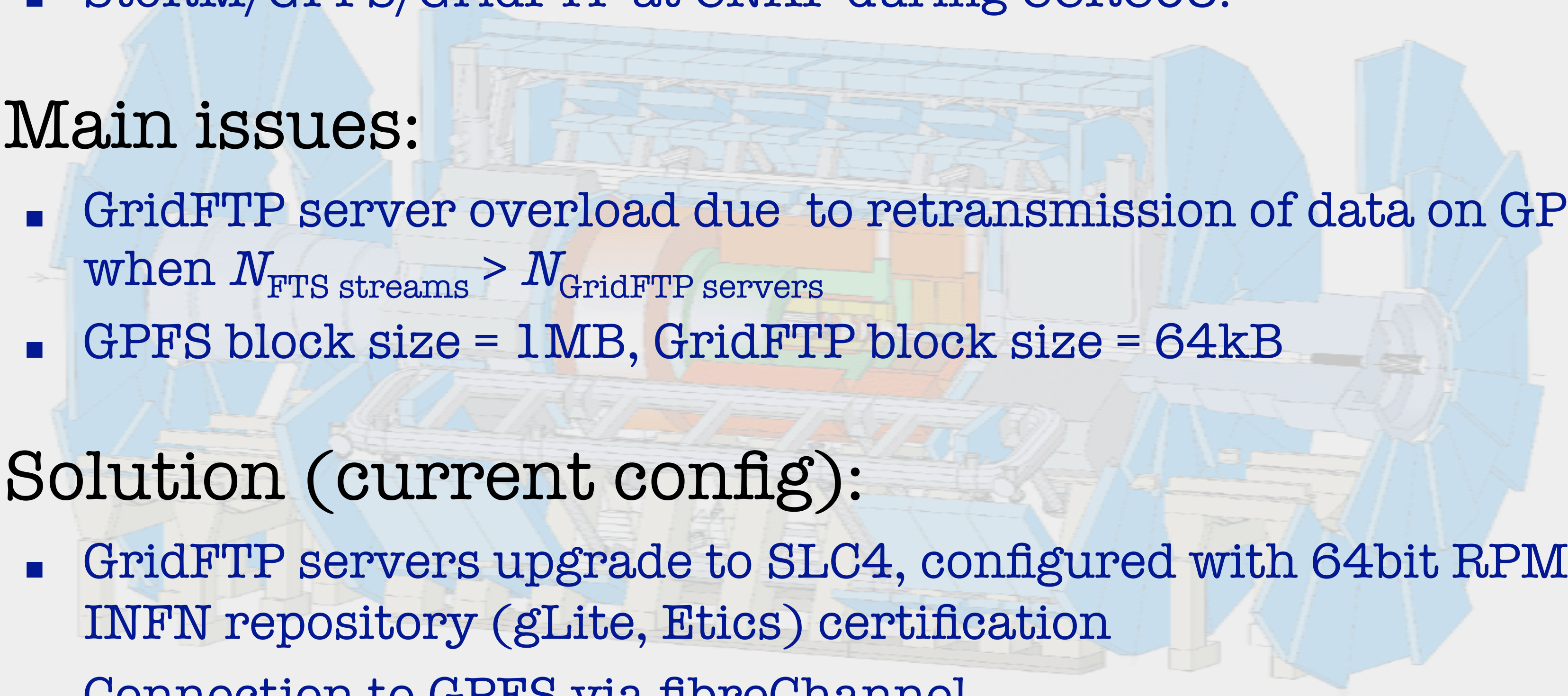
OUTLINE OF THE TALK

- 
-
- 
- 
- Conclusions

ATLAS

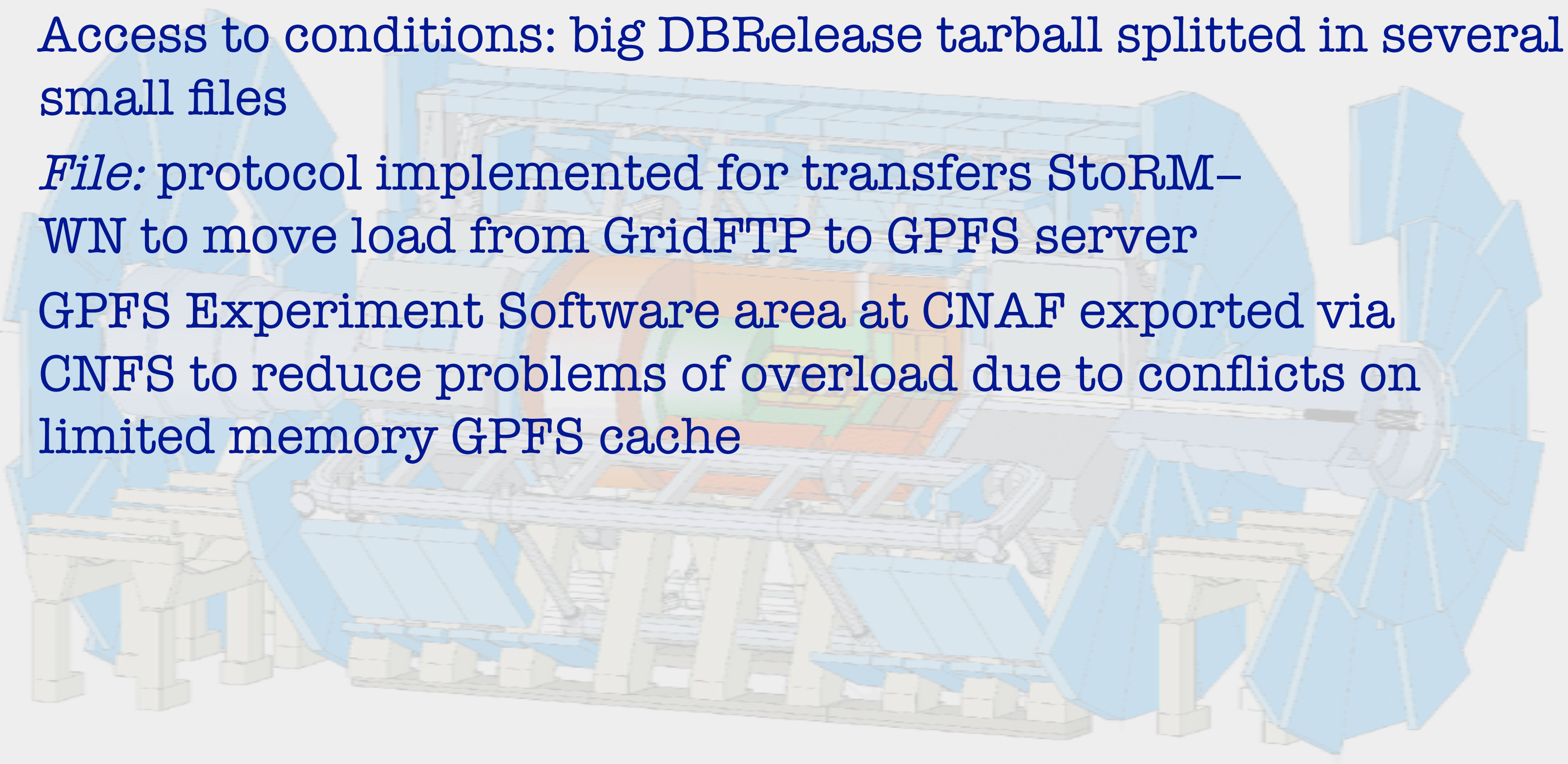


ATLAS EXPERIENCE 1



- Large experience, knowledge and improvement of D1T0 storage class
  - StoRM/GPFS/GridFTP at CNAF during CCRC08.
- Main issues:
  - GridFTP server overload due to retransmission of data on GPFS when $N_{\text{FTS streams}} > N_{\text{GridFTP servers}}$
  - GPFS block size = 1MB, GridFTP block size = 64kB
- Solution (current config):
  - GridFTP servers upgrade to SLC4, configured with 64bit RPM, INFN repository (gLite, Etics) certification
  - Connection to GPFS via fibreChannel
  - StoRM BE and MySQL DB on separate machines

ATLAS EXPERIENCE 2



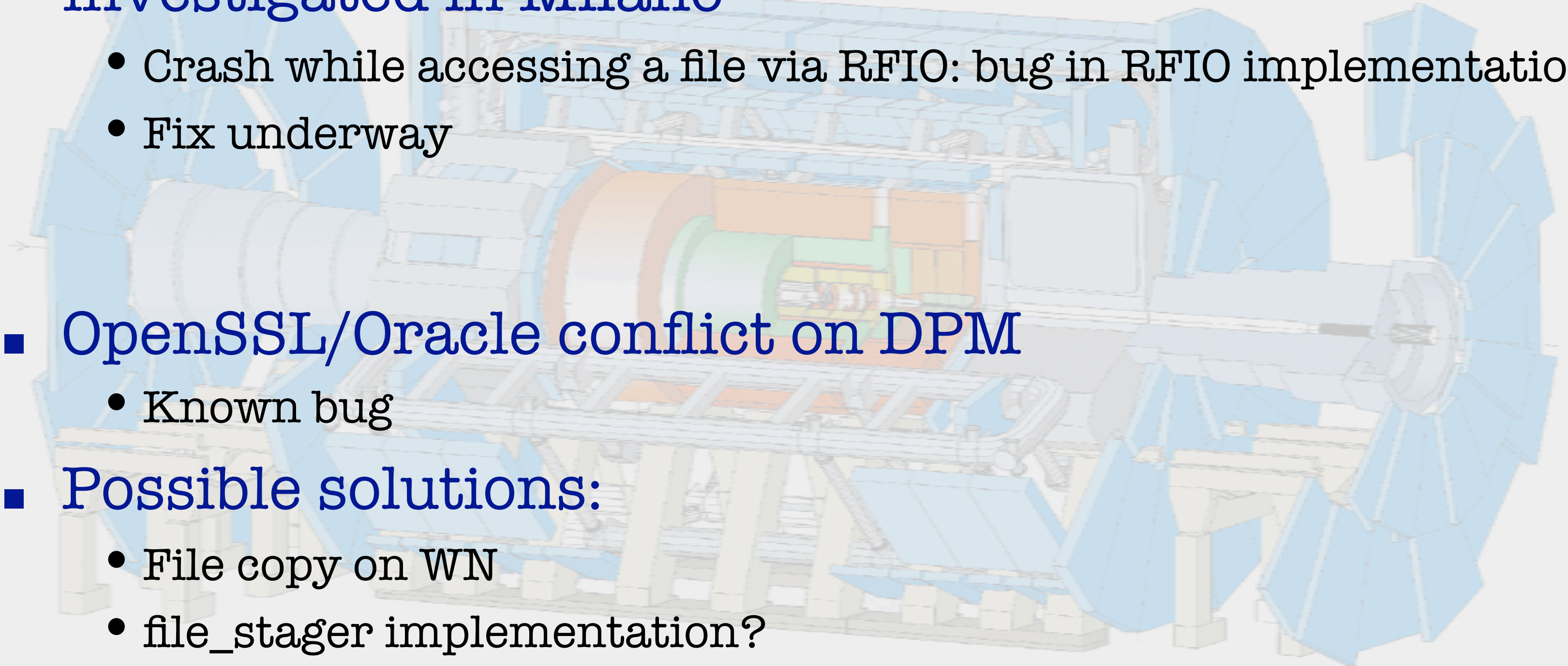
- Experience from reprocessing and analysis challenges:
  - Access to conditions: big DBRelease tarball splitted in several small files
  - *File:* protocol implemented for transfers StoRM–WN to move load from GridFTP to GPFS server
  - GPFS Experiment Software area at CNAF exported via CNFS to reduce problems of overload due to conflicts on limited memory GPFS cache

ISSUES SEEN BY ATLAS 1

- Storage system too fragile in 2008
  - Storage system downs cause inefficiencies
  - Sometimes caused by interference from other VOs
- SRM endpoint often shared
- Some files cannot be recalled from tape
  - Manual intervention needed
- Checksum mismatches (FTS)
  - Require cleaning and re-transfers
- Prestaging
  - Very low performances at some sites, good at CNAF (100MB/s)
  - Still untested under stress and VO concurrency

ISSUES SEEN BY ATLAS 2

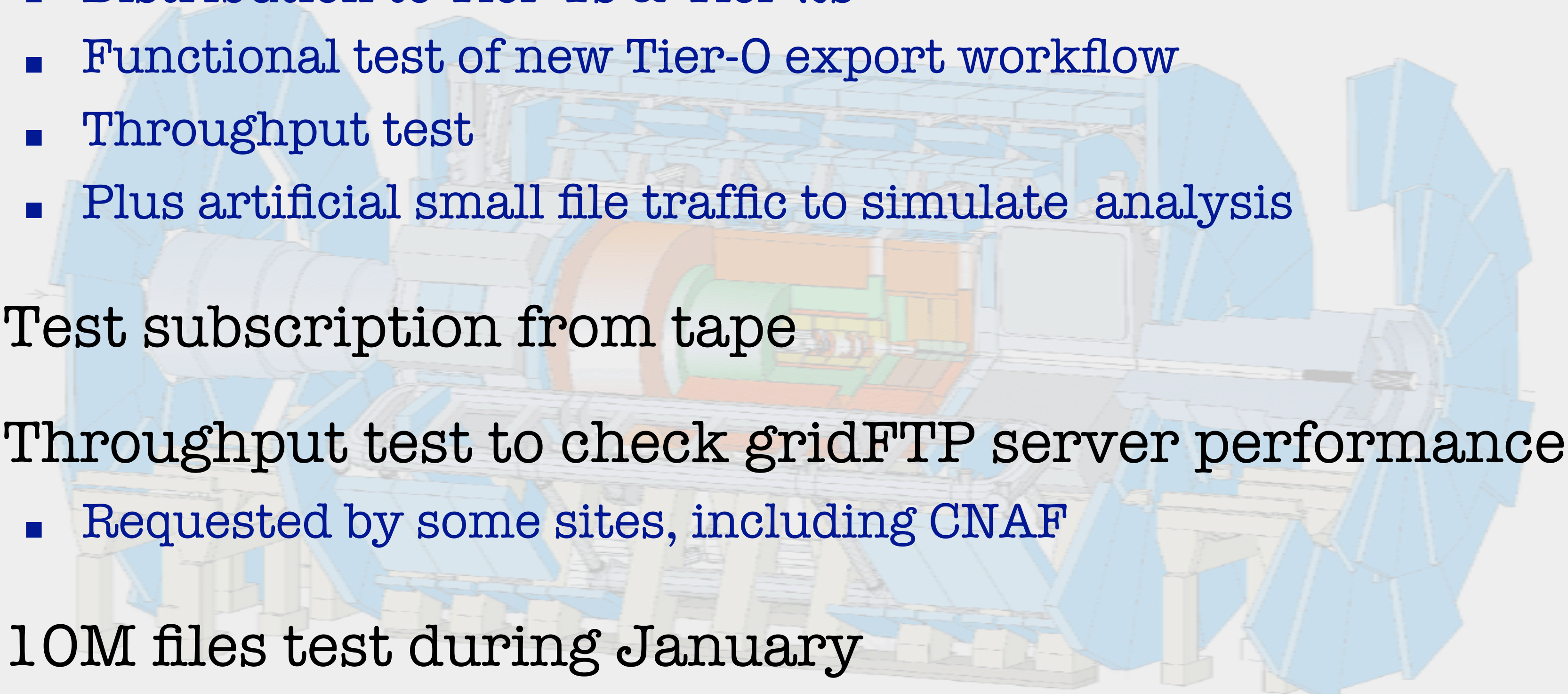
- Analysis



  - Problems with RFIO protocol on DPM at Tier-2s investigated in Milano
    - Crash while accessing a file via RFIO: bug in RFIO implementation
    - Fix underway
  - OpenSSL/Oracle conflict on DPM
    - Known bug
  - Possible solutions:
    - File copy on WN
    - file_stager implementation?
    - Migration to StoRM?

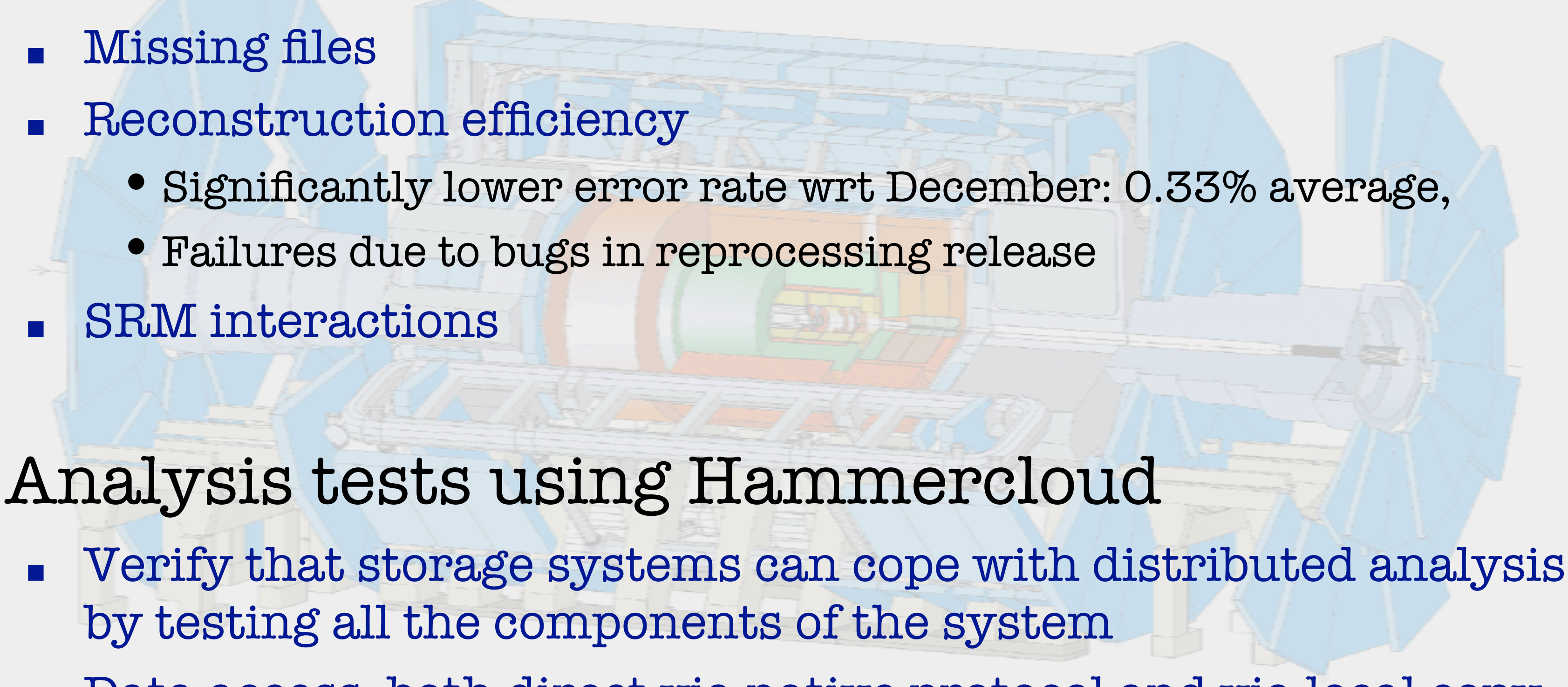
PLANS FOR 2009

- STEP09 Full scale test
  - Tier-0 reconstruction
  - Distribution to Tier-1s & Tier-2s
  - Functional test of new Tier-0 export workflow
  - Throughput test
  - Plus artificial small file traffic to simulate analysis
- Test subscription from tape
- Throughput test to check gridFTP server performance
  - Requested by some sites, including CNAF
- 10M files test during January
  - 10M small files distributed over 10 days in Tier-1 sites
  - Checked how many files can be collected in a day at Tier-1 (in case of stop export for few days)



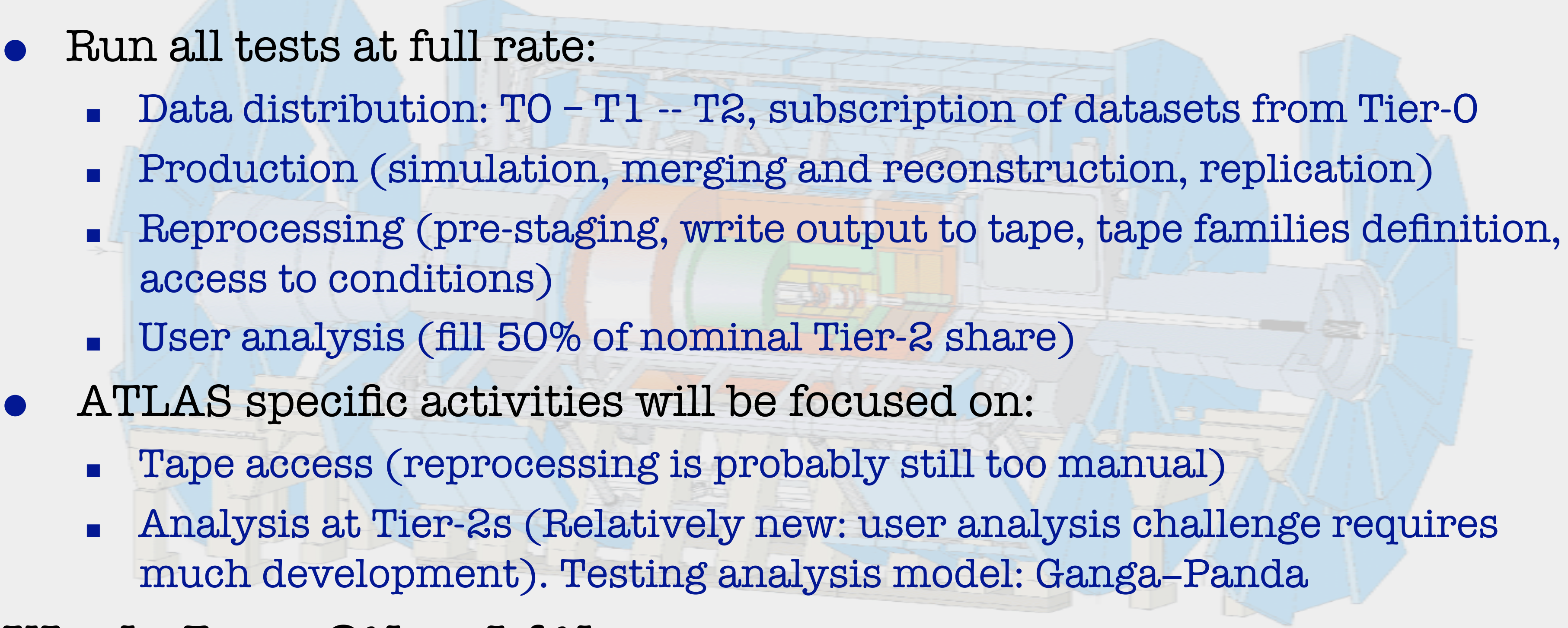
PLANS FOR 2009

- Reprocessing tests during April. Checked:
  - Bulk pre-staging
  - Achievable throughput
  - Missing files
  - Reconstruction efficiency
    - Significantly lower error rate wrt December: 0.33% average,
    - Failures due to bugs in reprocessing release
  - SRM interactions



- Analysis tests using Hammercloud
  - Verify that storage systems can cope with distributed analysis by testing all the components of the system
  - Data access: both direct via native protocol and via local copy to WN are tested

STEP 2009



Week May 25th – 31th:

- Setting up all tests at low rate

Week June 1st – 7th:

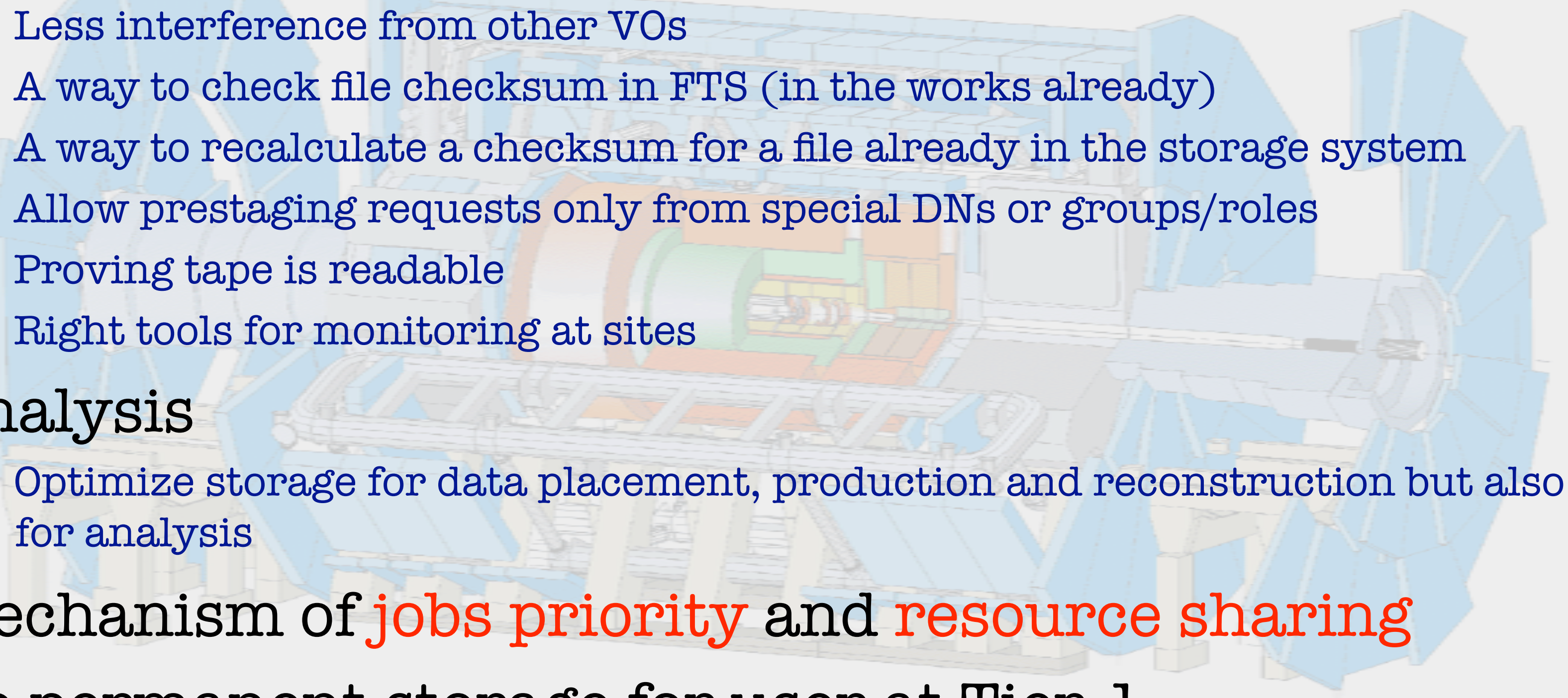
- Run all tests at full rate:
  - Data distribution: T0 – T1 -- T2, subscription of datasets from Tier-0
  - Production (simulation, merging and reconstruction, replication)
  - Reprocessing (pre-staging, write output to tape, tape families definition, access to conditions)
  - User analysis (fill 50% of nominal Tier-2 share)
- ATLAS specific activities will be focused on:
  - Tape access (reprocessing is probably still too manual)
  - Analysis at Tier-2s (Relatively new: user analysis challenge requires much development). Testing analysis model: Ganga–Panda

Week June 8th – 14th :

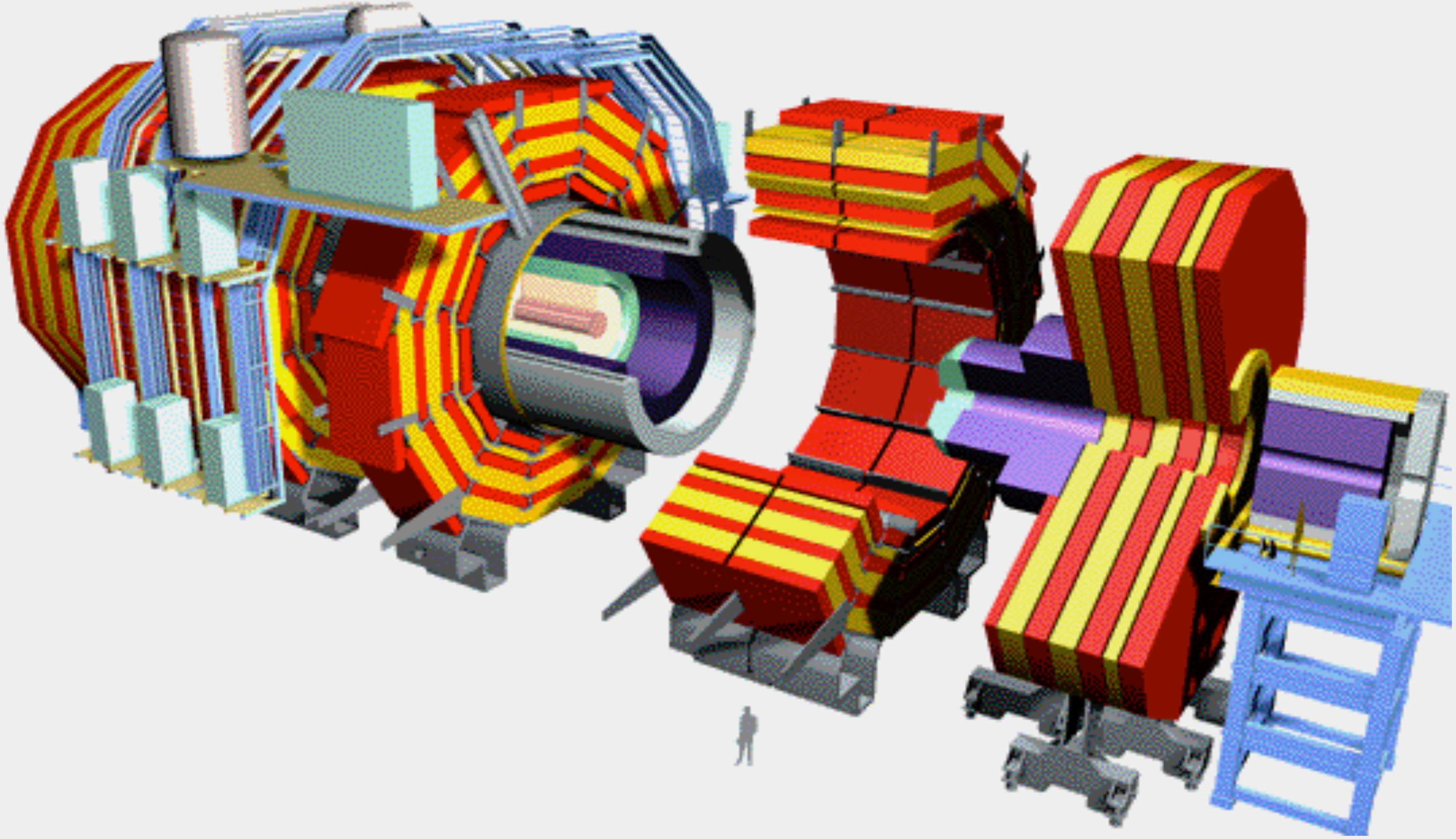
- Run all tests at full rate – combined

REQUIREMENTS FOR 2009

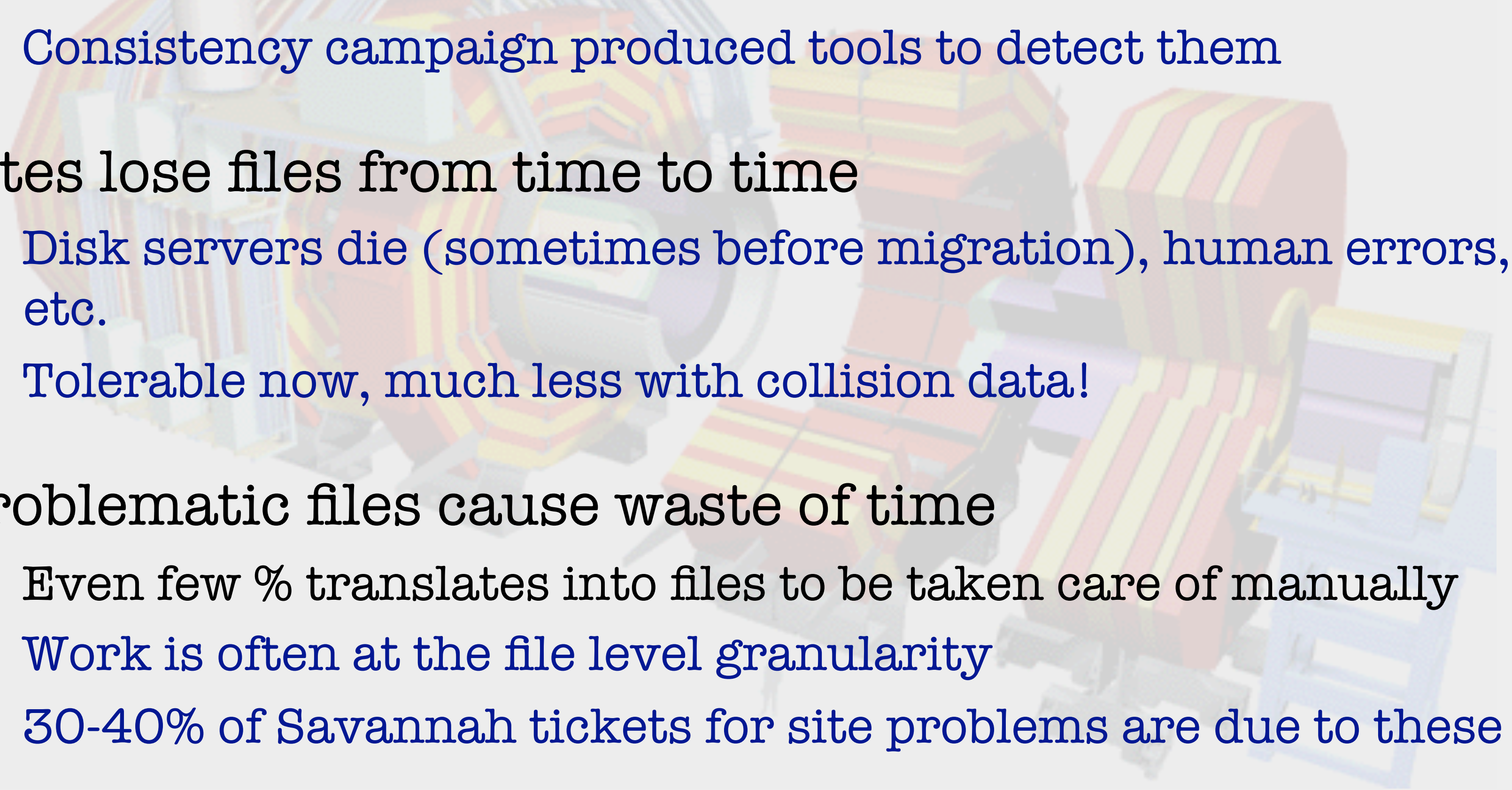
- The focus should be on strengthening the services
- Questions and requests:
  - Is it a good practice to use srmLs, e.g. to get file size, checksum, etc? Can any user do that or only a central service? srmLs should not block SRM server!
  - Less interference from other VOs
  - A way to check file checksum in FTS (in the works already)
  - A way to recalculate a checksum for a file already in the storage system
  - Allow prestaging requests only from special DNs or groups/roles
  - Proving tape is readable
  - Right tools for monitoring at sites
- Analysis
  - Optimize storage for data placement, production and reconstruction but also for analysis
- Mechanism of jobs priority and resource sharing
- No permanent storage for user at Tier-1
- Optimize access to conditions data



CMS

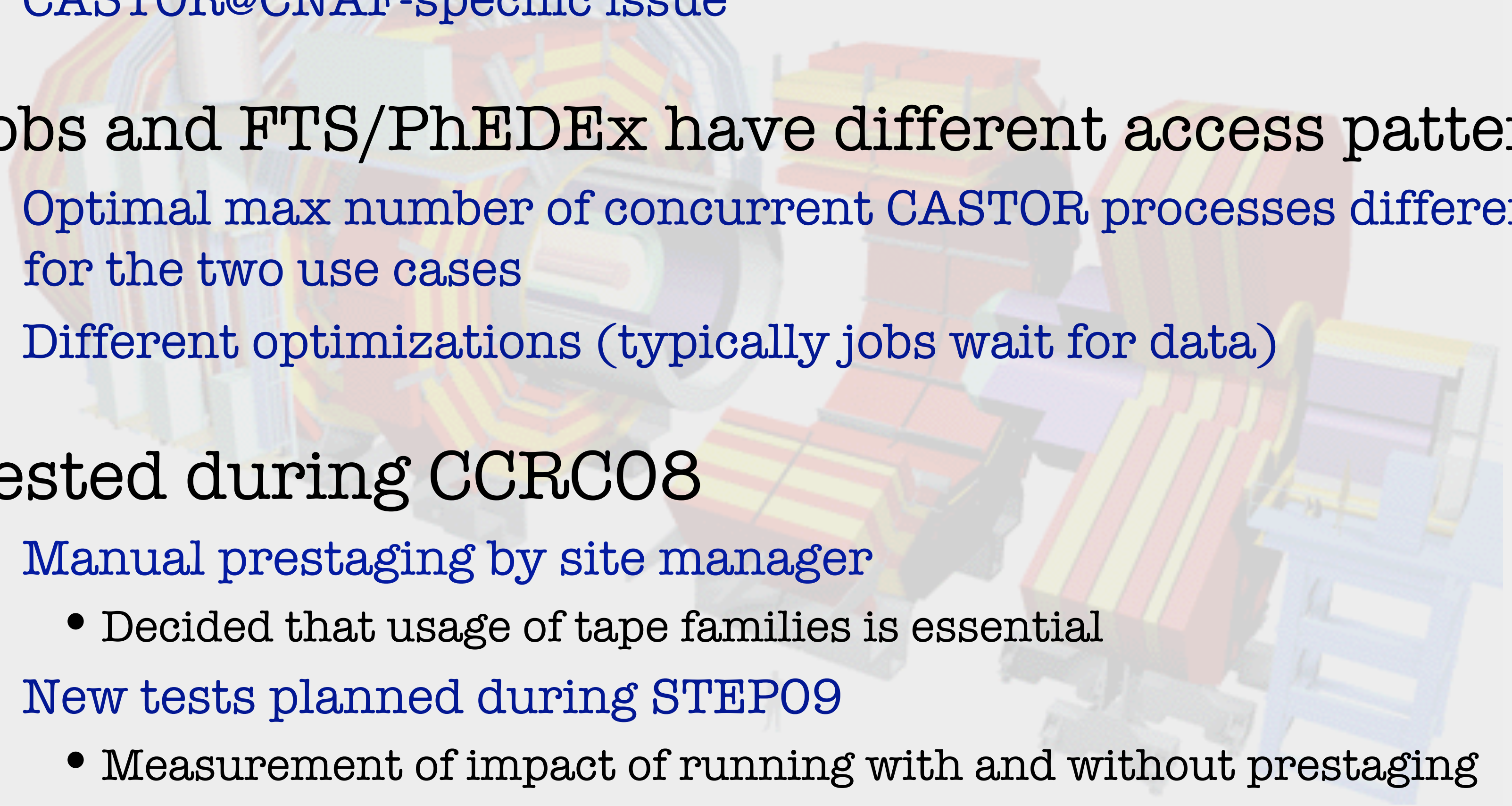


ISSUES SEEN BY CMS



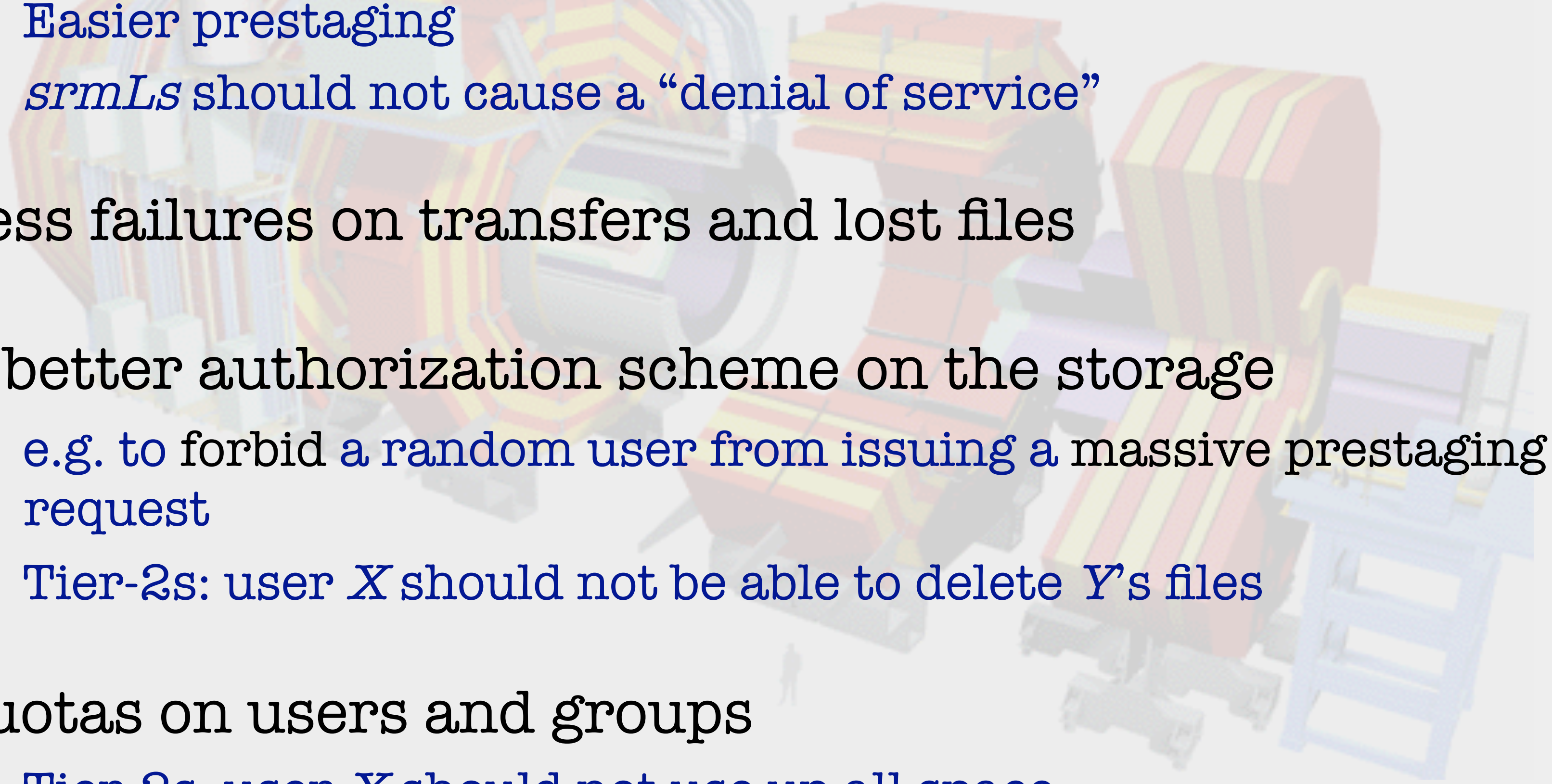
- Internal inconsistencies can arise among DBS, PhEDEx database and storage
  - Consistency campaign produced tools to detect them
- Sites lose files from time to time
  - Disk servers die (sometimes before migration), human errors, etc.
  - Tolerable now, much less with collision data!
- Problematic files cause waste of time
  - Even few % translates into files to be taken care of manually
  - Work is often at the file level granularity
  - 30-40% of Savannah tickets for site problems are due to these
- Storage system instabilities
  - E.g. dCache SRM port

PRESTAGING



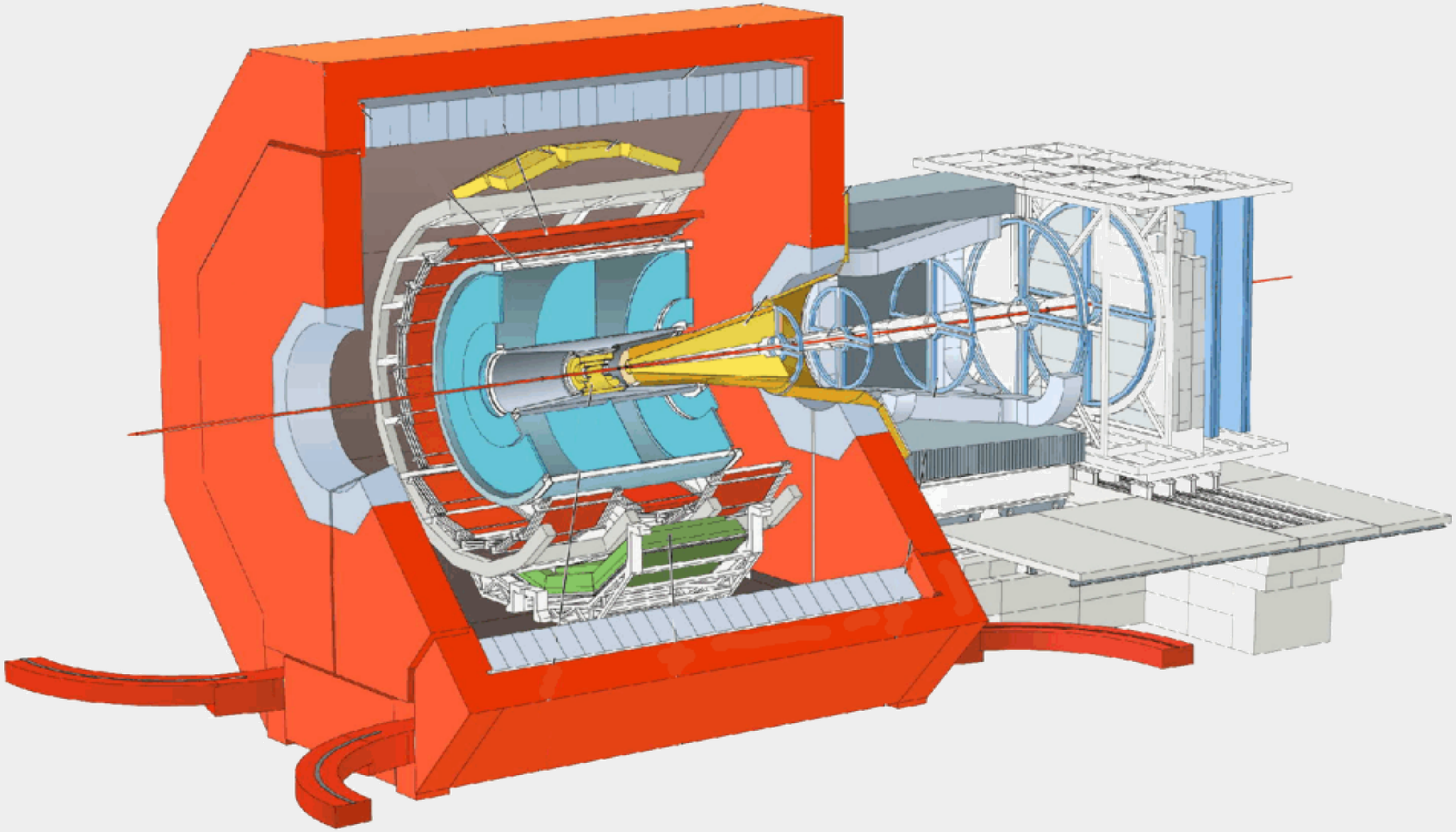
- Improve efficiency of jobs accessing data on tape
  - CASTOR@CNAF-specific issue
- Jobs and FTS/PhEDEx have different access pattern
  - Optimal max number of concurrent CASTOR processes different for the two use cases
  - Different optimizations (typically jobs wait for data)
- Tested during CCRC08
  - Manual prestaging by site manager
    - Decided that usage of tape families is essential
  - New tests planned during STEP09
    - Measurement of impact of running with and without prestaging
    - Planned to be integrated in the workflow at a later time

CMS WISH LIST

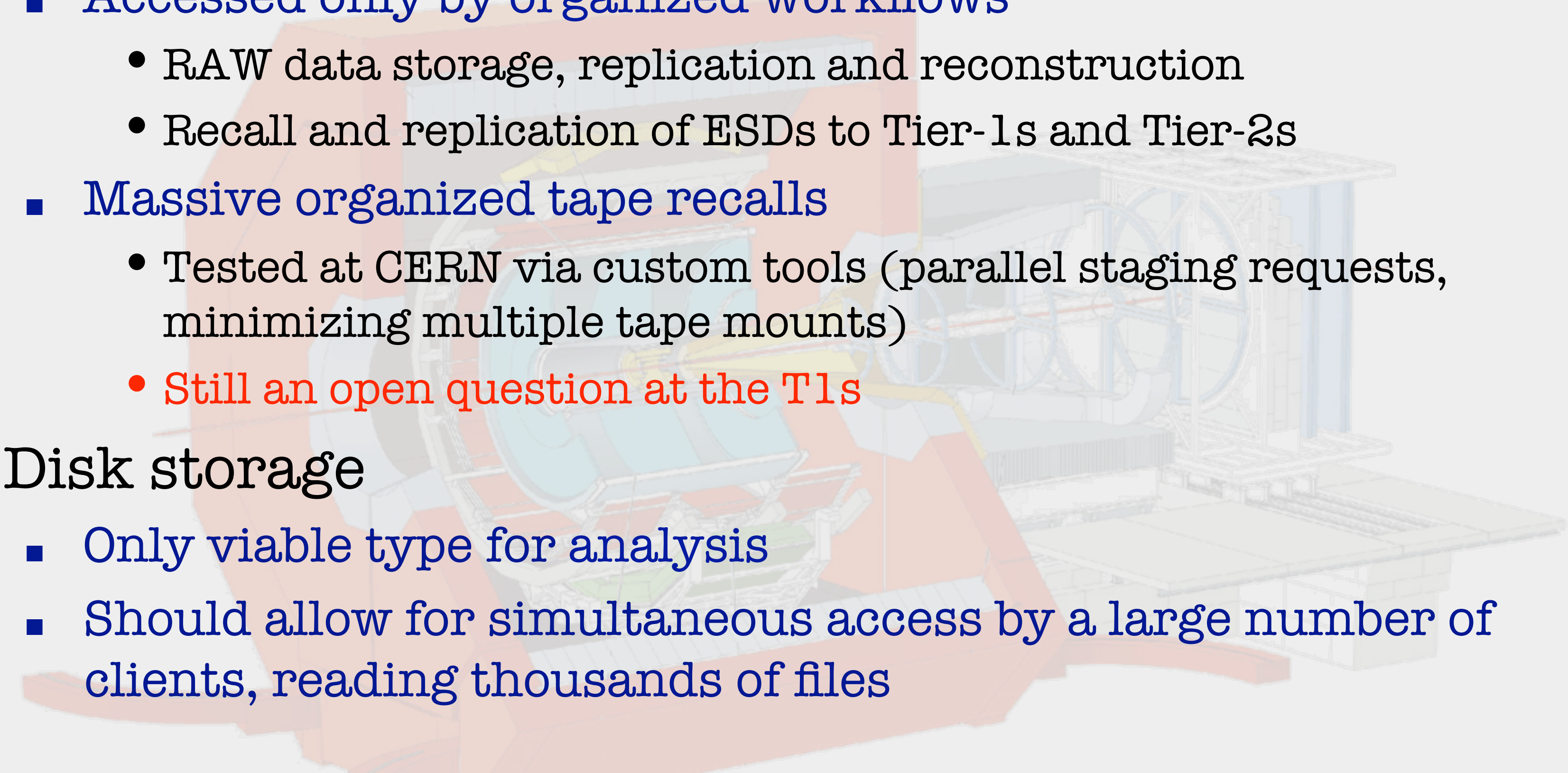


- A much better dCache SRM scaling (by a factor of 5-10)
  - Would pull data management farther from the sites
  - Easier prestaging
  - *srmLs* should not cause a “denial of service”
- Less failures on transfers and lost files
- A better authorization scheme on the storage
  - e.g. to forbid a random user from issuing a massive prestaging request
  - Tier-2s: user *X* should not be able to delete *Y*’s files
- Quotas on users and groups
  - Tier-2s: user *X* should not use up all space

ALICE



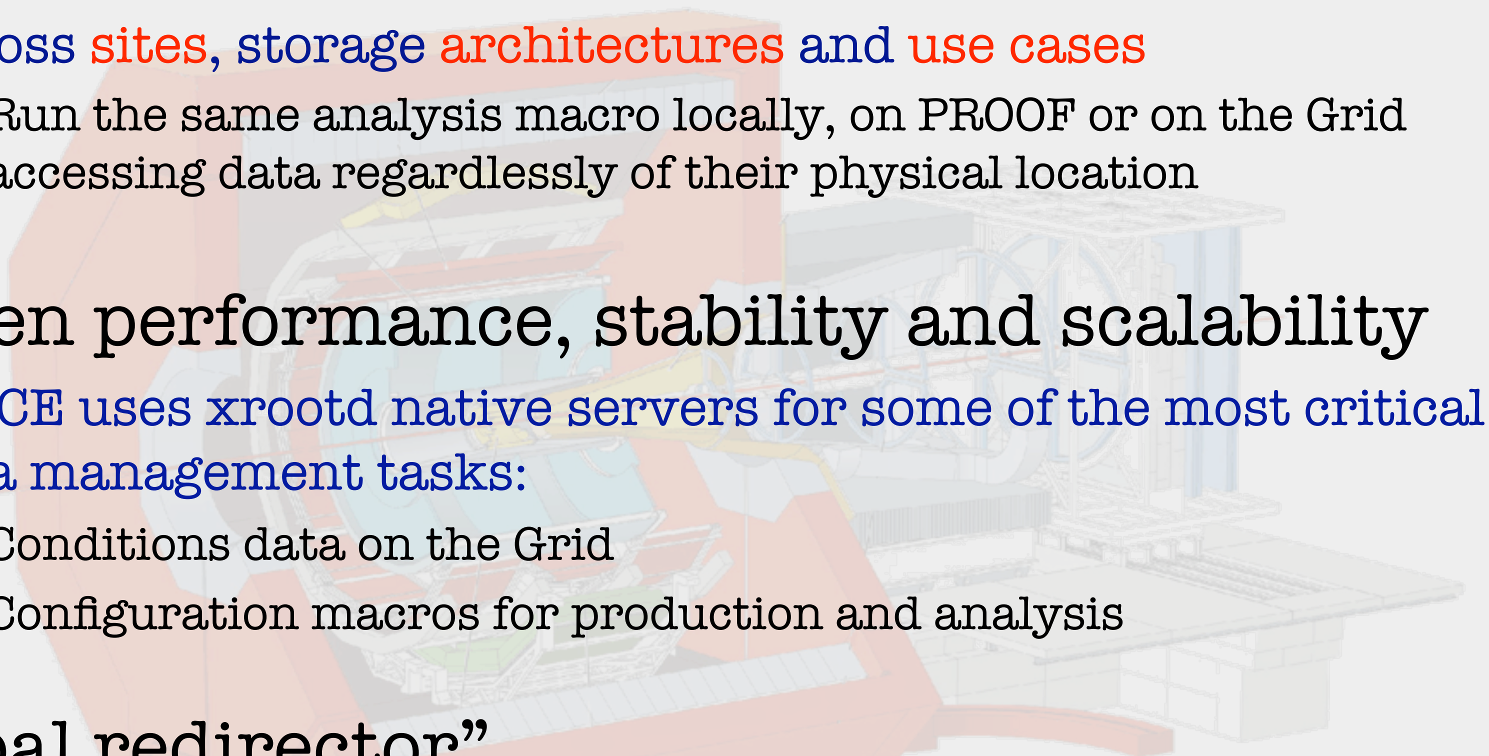
ALICE DATA MANAGEMENT



- Mass storage
  - Accessed only by organized workflows
    - RAW data storage, replication and reconstruction
    - Recall and replication of ESDs to Tier-1s and Tier-2s
  - Massive organized tape recalls
    - Tested at CERN via custom tools (parallel staging requests, minimizing multiple tape mounts)
    - Still an open question at the T1s
- Disk storage
  - Only viable type for analysis
  - Should allow for simultaneous access by a large number of clients, reading thousands of files

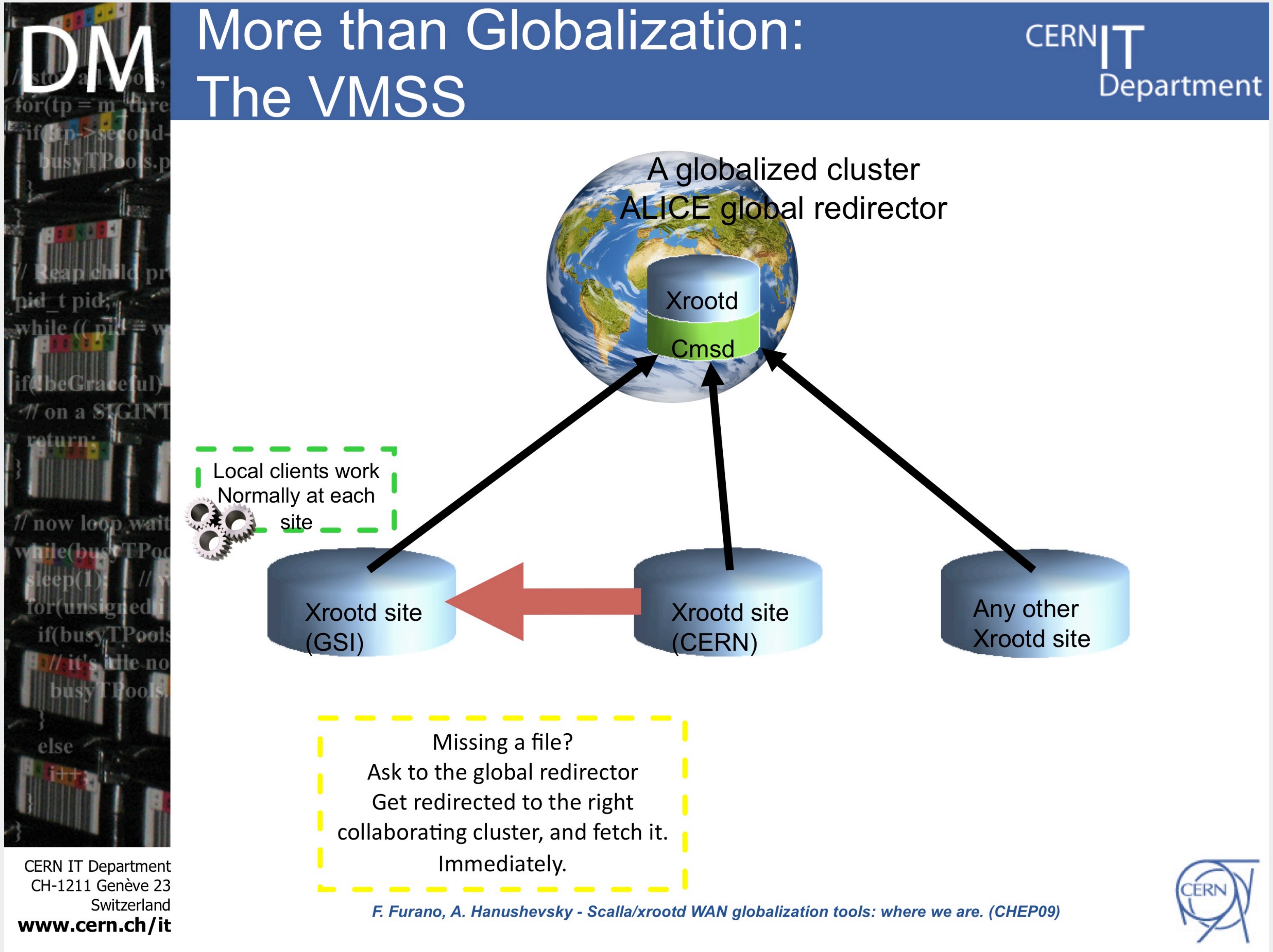
All ALICE Grid sites are required to provide an xrootd enabled storage

XROOTD RATIONALE



- Uniform access protocol
  - Across sites, storage architectures and use cases
    - Run the same analysis macro locally, on PROOF or on the Grid accessing data regardlessly of their physical location
- Proven performance, stability and scalability
  - ALICE uses xrootd native servers for some of the most critical data management tasks:
    - Conditions data on the Grid
    - Configuration macros for production and analysis
- “Global redirector”
  - Xrootd has a highly optimized “WAN mode”
  - Torrent-like “extreme copy”
  - See next slide…

GLOBAL REDIRECTOR



DM

More than Globalization: The VMSS

CERN IT Department

A globalized cluster
ALICE global redirector

Xrootd

Cmsd

Local clients work
Normally at each
site

Xrootd site
(GSI)

Xrootd site
(CERN)

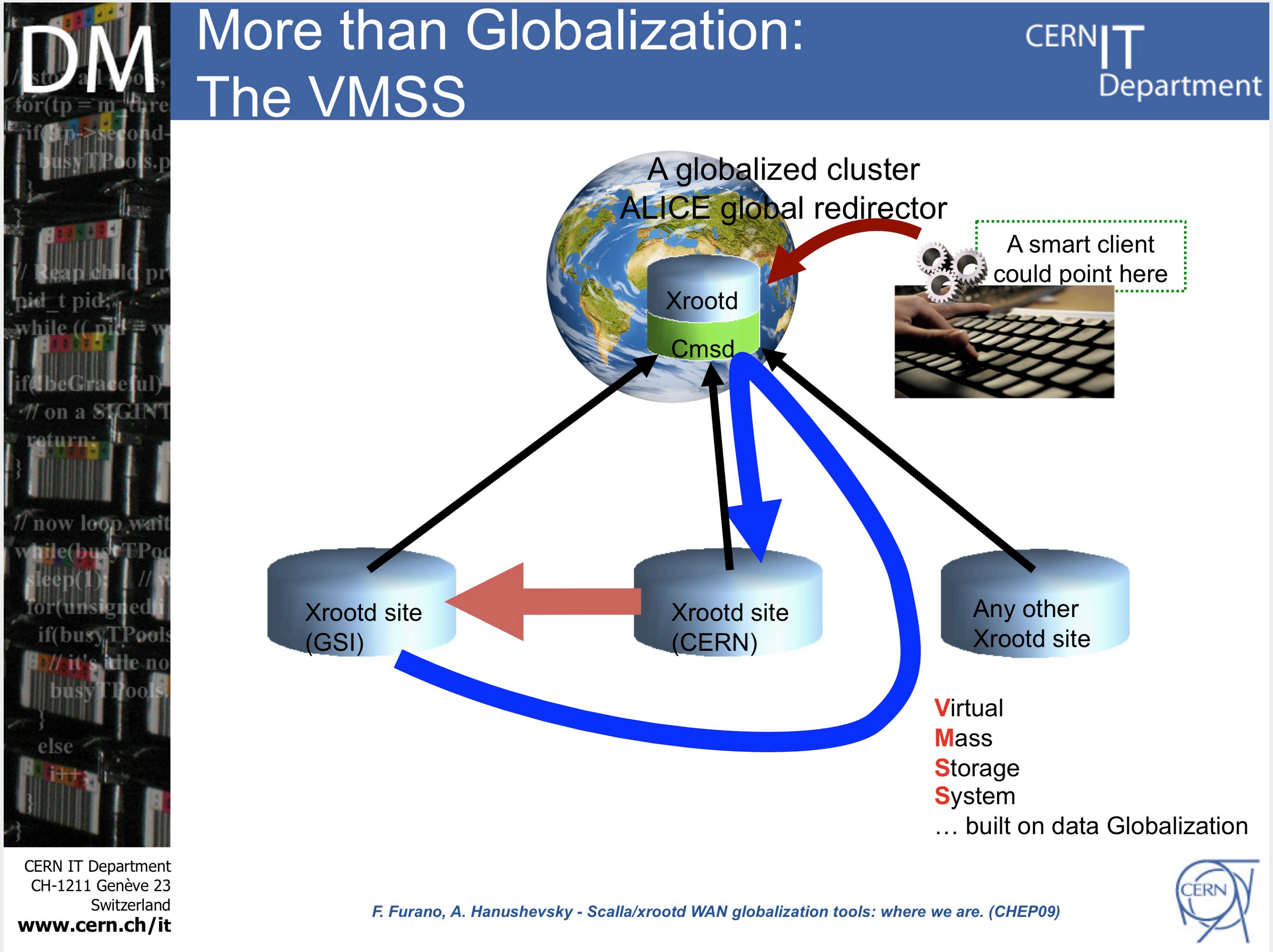
Any other
Xrootd site

Missing a file?
Ask to the global redirector
Get redirected to the right
collaborating cluster, and fetch it.
Immediately.

F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

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

GLOBAL REDIRECTOR



More than Globalization: The VMSS

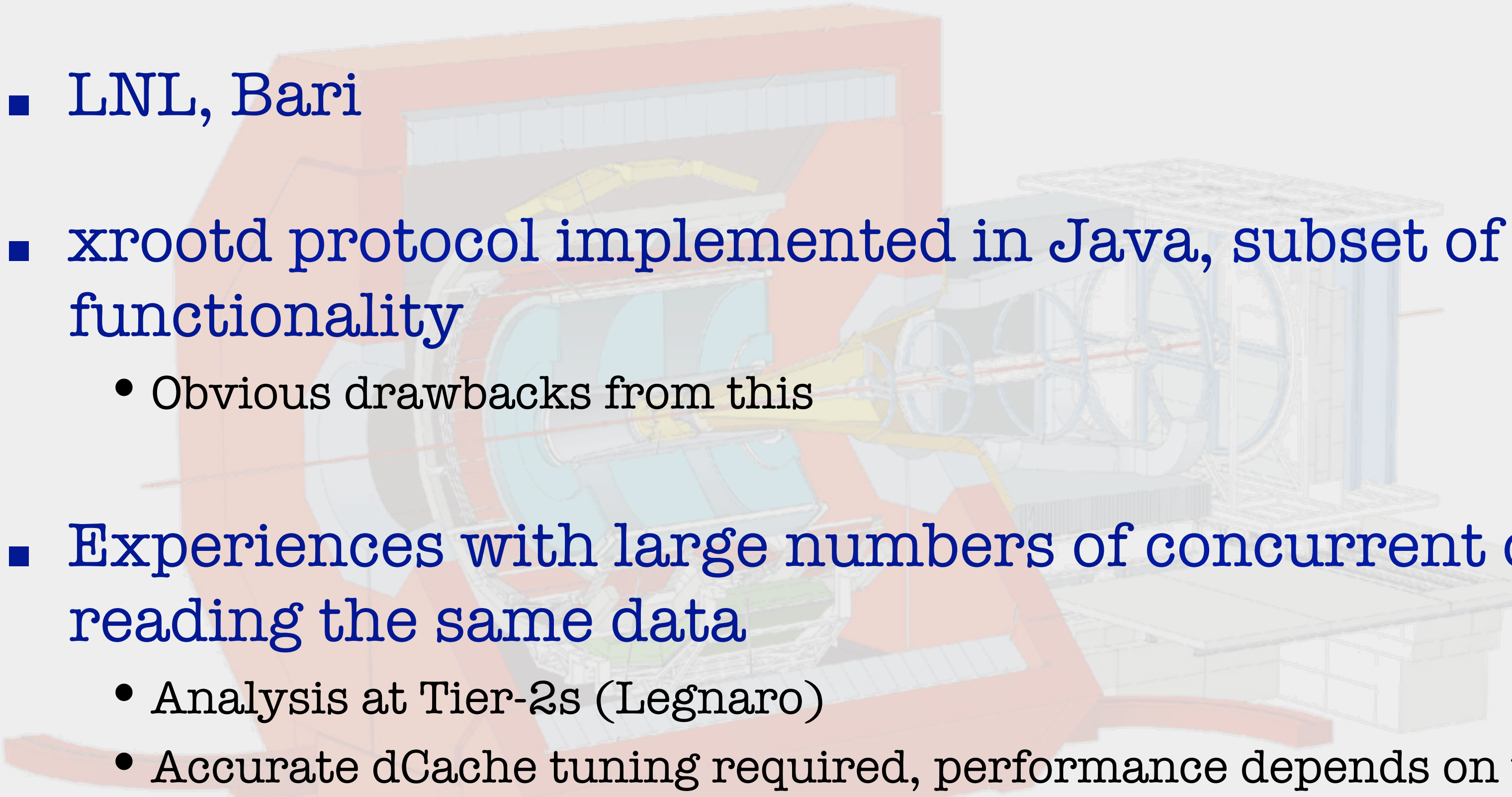
CERN IT Department

Virtual
Mass
Storage
System
… built on data Globalization

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

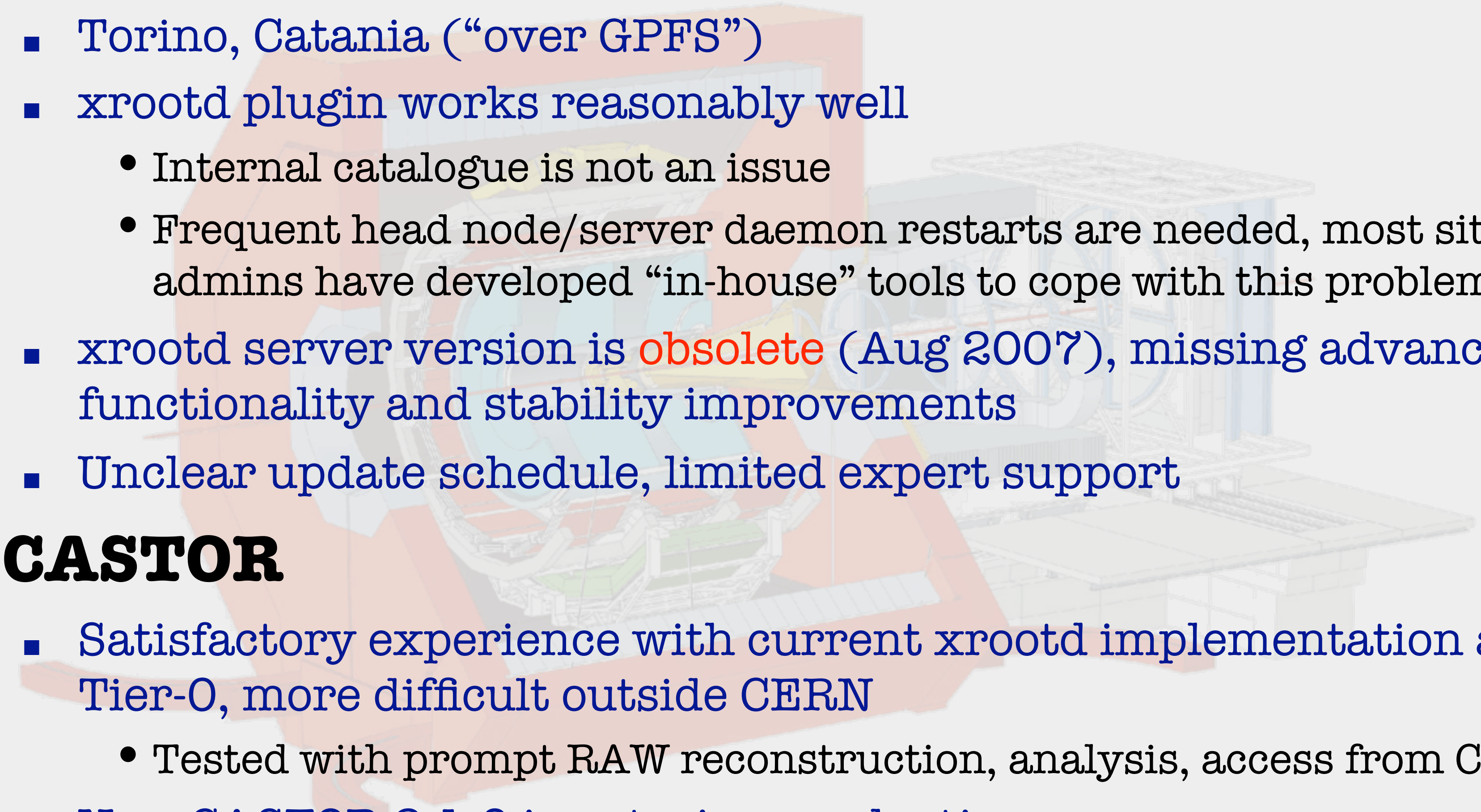
F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

DISK ACCESS VIA XROOTD 1



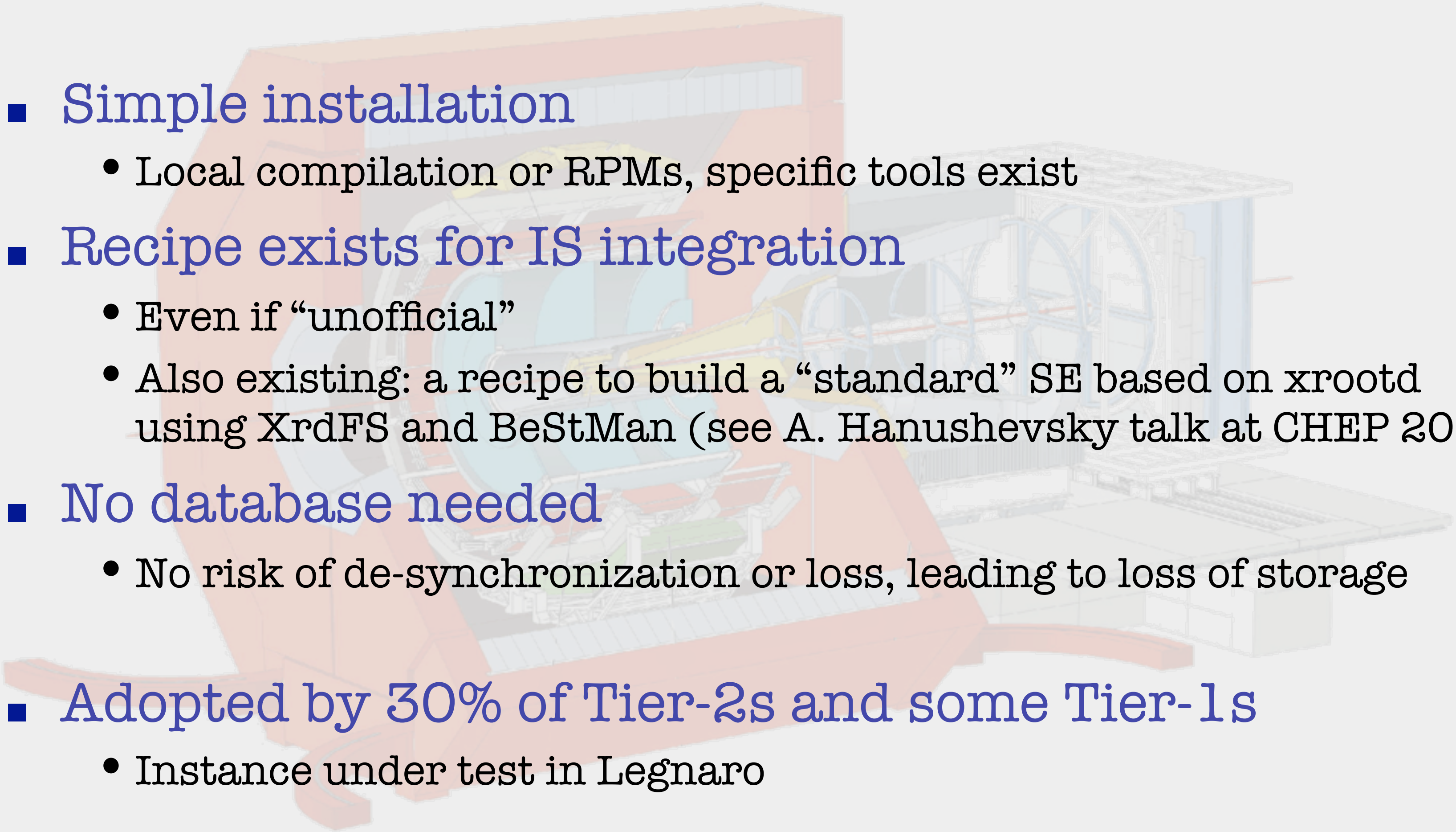
- **dCache**
  - LNL, Bari
  - xrootd protocol implemented in Java, subset of full functionality
    - Obvious drawbacks from this
  - Experiences with large numbers of concurrent clients reading the same data
    - Analysis at Tier-2s (Legnaro)
    - Accurate dCache tuning required, performance depends on the level of expertise at the centre
    - Internal dCache server protection causes clients to wait, resulting in under-utilization of CPU

DISK ACCESS VIA XROOTD 2



- **DPM**
  - Torino, Catania ("over GPFS")
  - xrootd plugin works reasonably well
    - Internal catalogue is not an issue
    - Frequent head node/server daemon restarts are needed, most site admins have developed "in-house" tools to cope with this problem
  - xrootd server version is obsolete (Aug 2007), missing advanced functionality and stability improvements
  - Unclear update schedule, limited expert support
- **CASTOR**
  - Satisfactory experience with current xrootd implementation at Tier-0, more difficult outside CERN
    - Tested with prompt RAW reconstruction, analysis, access from CAF
  - New CASTOR 2.1.8 is entering production
    - Further improvement in xrootd-CASTOR interoperability

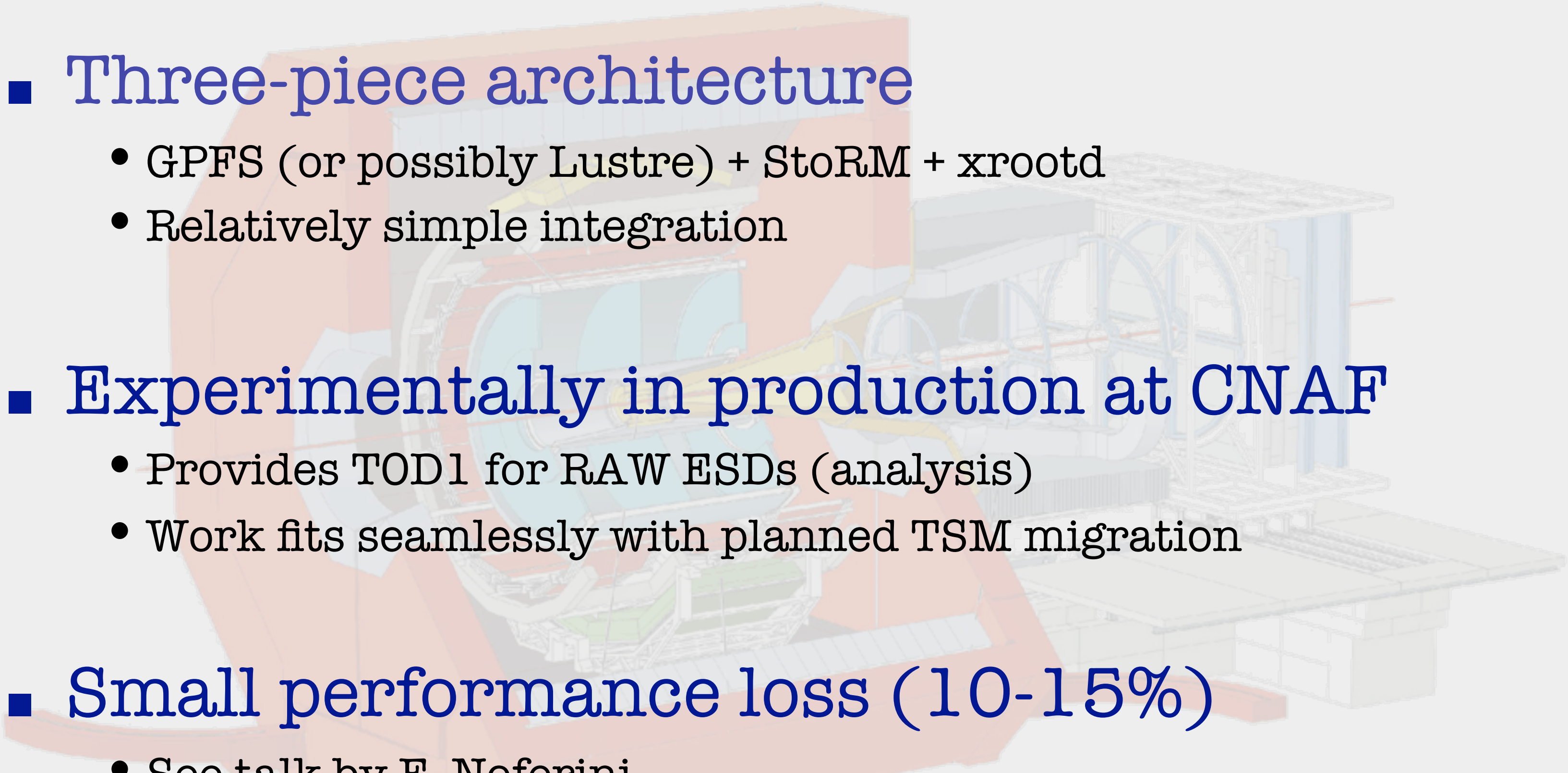
DISK ACCESS VIA XROOTD 3



● “Native” xrootd

- Simple installation
  - Local compilation or RPMs, specific tools exist
- Recipe exists for IS integration
  - Even if “unofficial”
  - Also existing: a recipe to build a “standard” SE based on xrootd using XrdFS and BeStMan (see A. Hanushevsky talk at CHEP 2009)
- No database needed
  - No risk of de-synchronization or loss, leading to loss of storage
- Adopted by 30% of Tier-2s and some Tier-1s
  - Instance under test in Legnaro
- Strongly endorsed by the collaboration

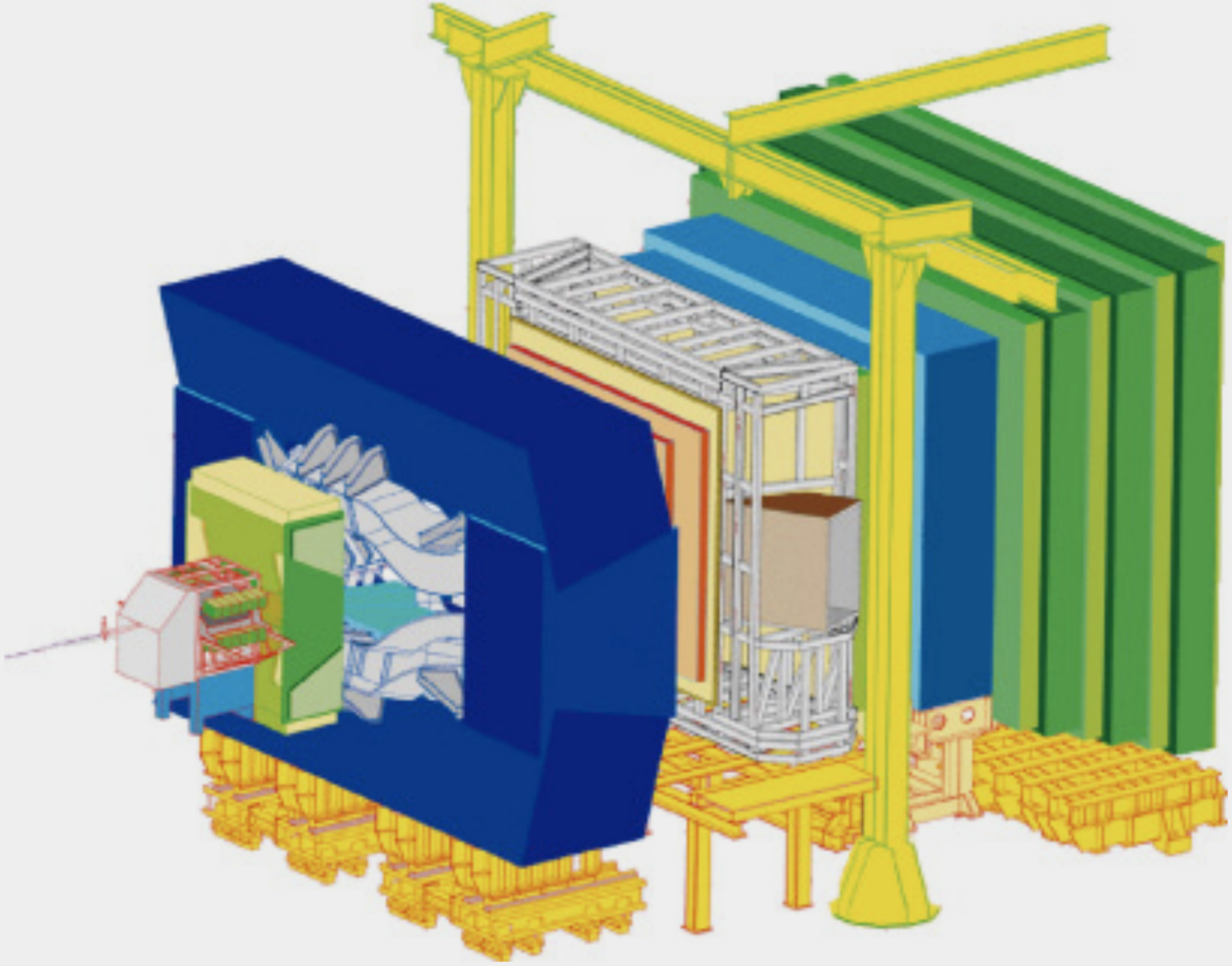
DISK ACCESS VIA XROOTD 4



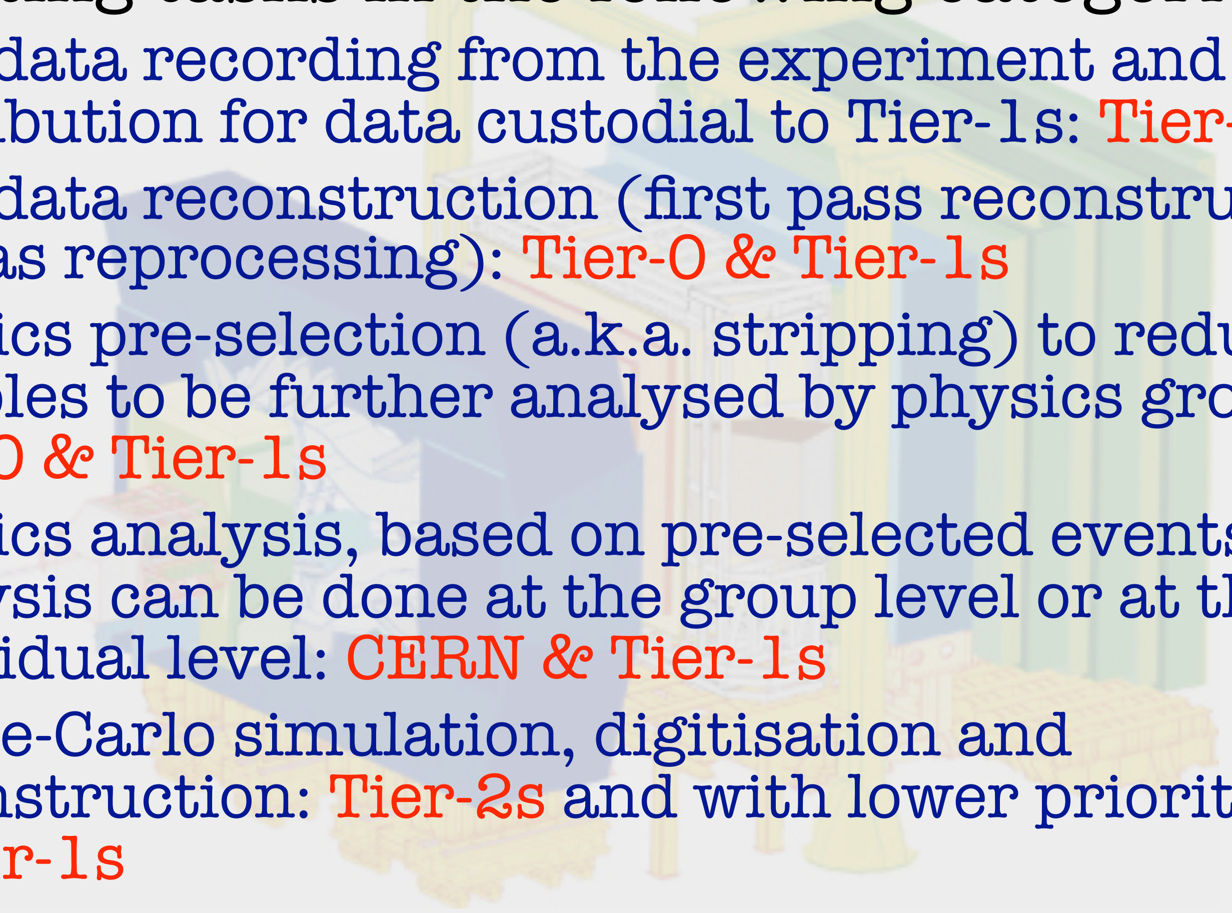
StoRM?

- Three-piece architecture
  - GPFS (or possibly Lustre) + StoRM + xrootd
  - Relatively simple integration
- Experimentally in production at CNAF
  - Provides TOD1 for RAW ESDs (analysis)
  - Work fits seamlessly with planned TSM migration
- Small performance loss (10-15%)
  - See talk by F. Noferini
  - No co-optimization, possible conflicting settings in GPFS/xrootd
  - Large network traffic generated by parallel file system under control

LHCB

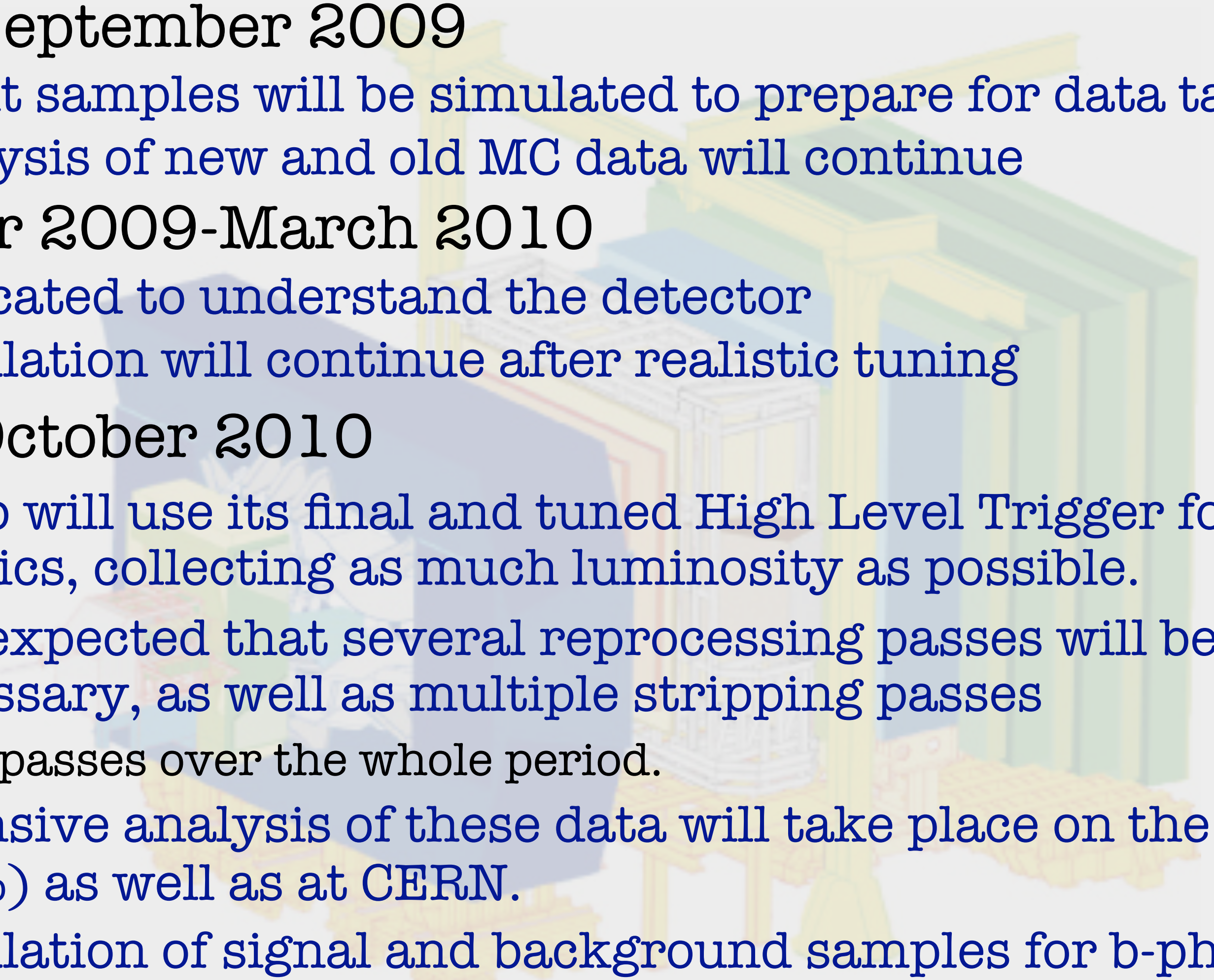


LHCB COMPUTING TASKS



- The LHCb Computing Model splits the computing tasks in the following categories:
  - Real data recording from the experiment and distribution for data custodial to Tier-1s: Tier-0
  - Real data reconstruction (first pass reconstruction as well as reprocessing): Tier-0 & Tier-1s
  - Physics pre-selection (a.k.a. stripping) to reduce data samples to be further analysed by physics groups: Tier-0 & Tier-1s
  - Physics analysis, based on pre-selected events. This analysis can be done at the group level or at the individual level: CERN & Tier-1s
  - Monte-Carlo simulation, digitisation and reconstruction: Tier-2s and with lower priority Tier-0 & Tier-1s

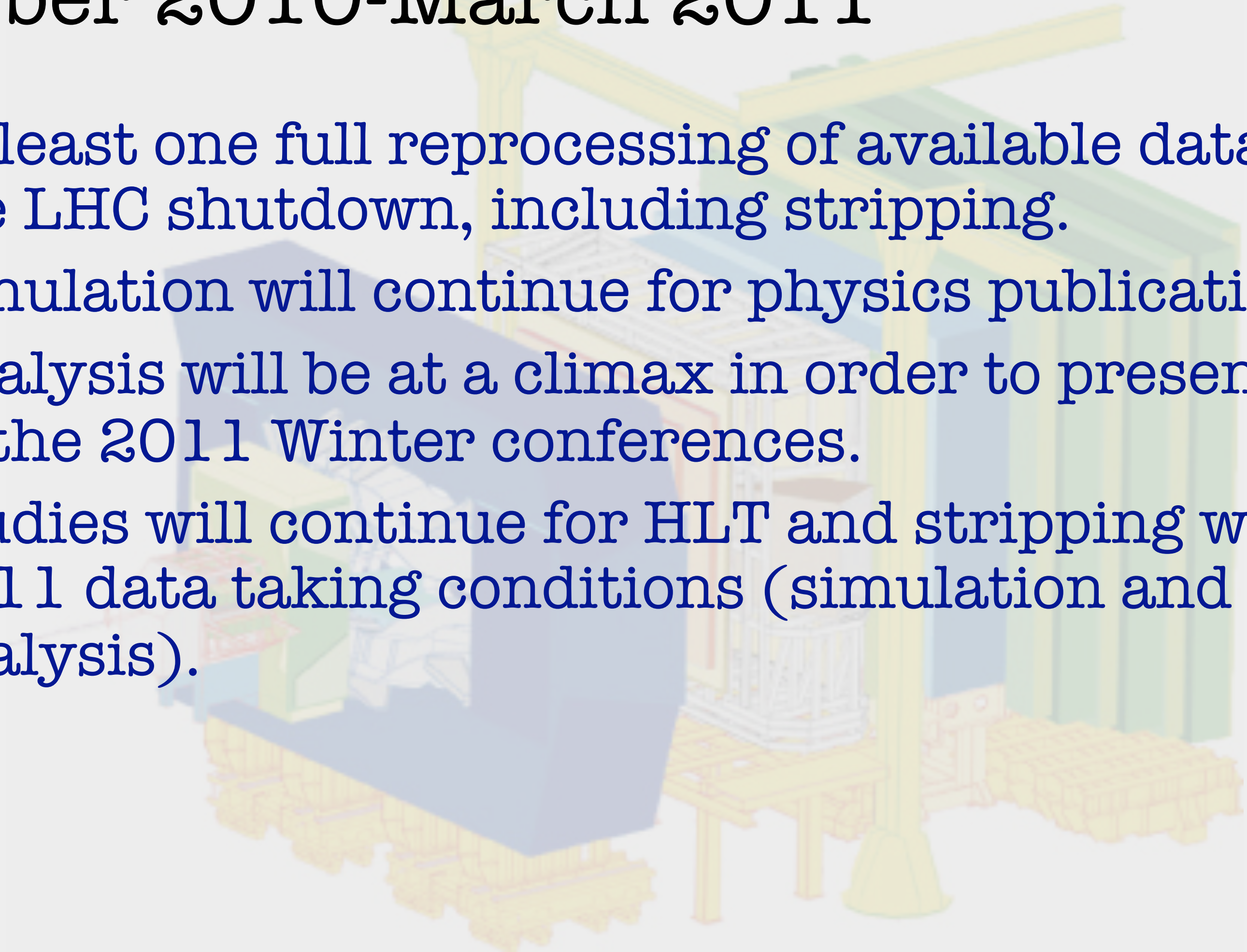
WORKING ASSUMPTIONS 1



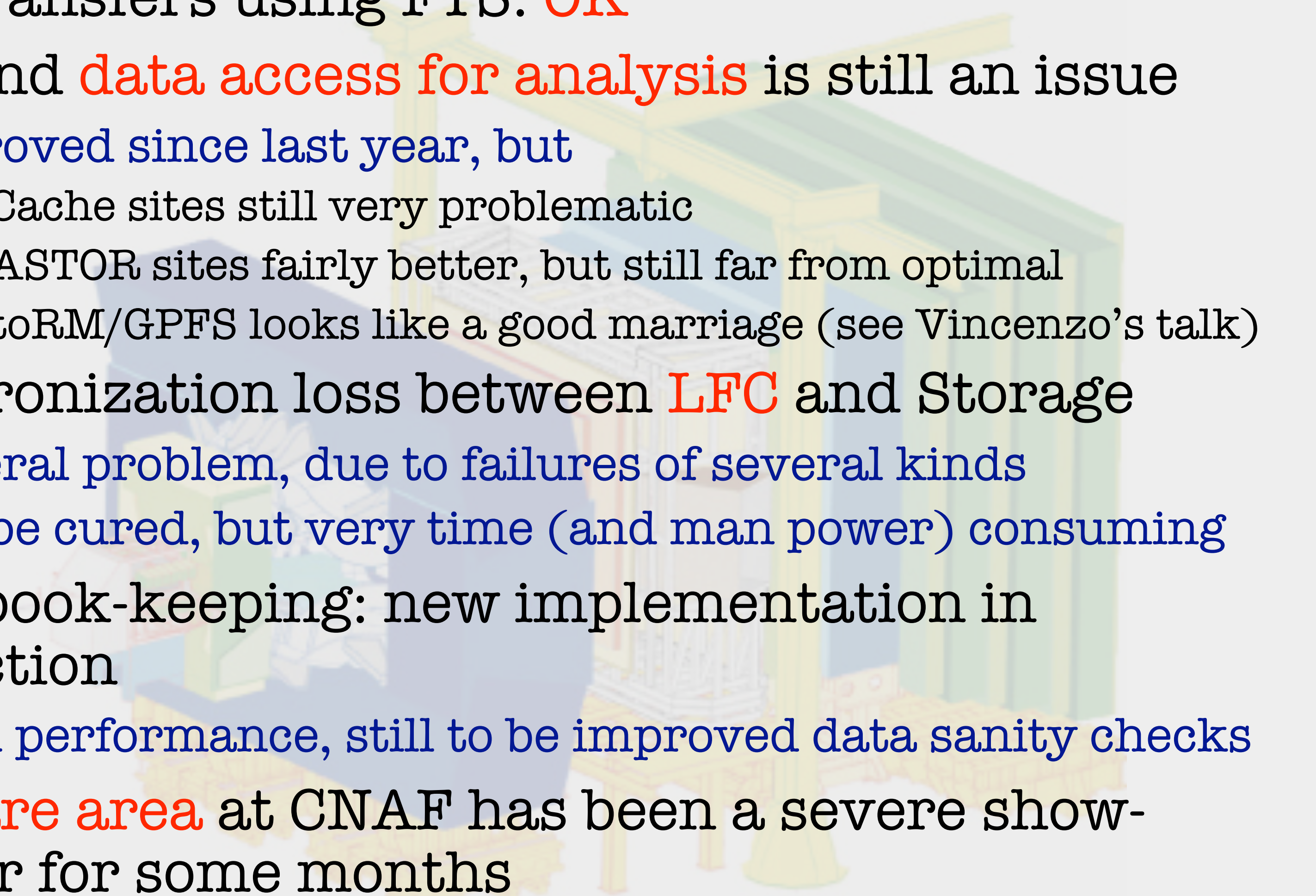
- April-September 2009
  - Event samples will be simulated to prepare for data taking.
  - Analysis of new and old MC data will continue
- October 2009-March 2010
  - Dedicated to understand the detector
  - Simulation will continue after realistic tuning
- April-October 2010
  - LHCb will use its final and tuned High Level Trigger for b-physics, collecting as much luminosity as possible.
  - It is expected that several reprocessing passes will be necessary, as well as multiple stripping passes
    - 3 passes over the whole period.
  - Intensive analysis of these data will take place on the Grid (60%) as well as at CERN.
  - Simulation of signal and background samples for b-physics, as well as preparatory simulation at the LHC nominal settings

WORKING ASSUMPTIONS 2

- October 2010-March 2011
  - At least one full reprocessing of available data during the LHC shutdown, including stripping.
  - Simulation will continue for physics publications
  - Analysis will be at a climax in order to present results at the 2011 Winter conferences.
  - Studies will continue for HLT and stripping with the 2011 data taking conditions (simulation and analysis).

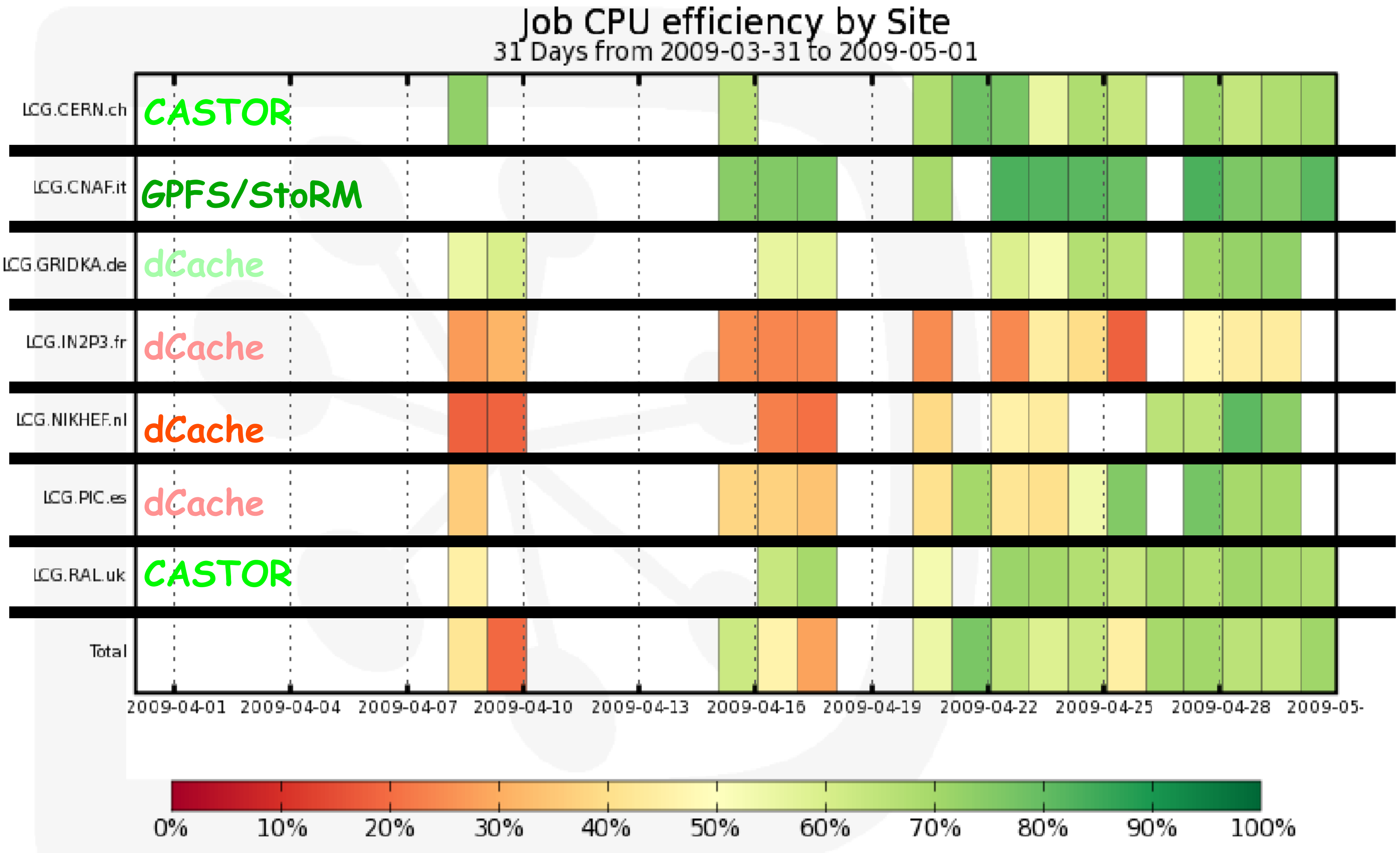


DATA MANAGEMENT ISSUES



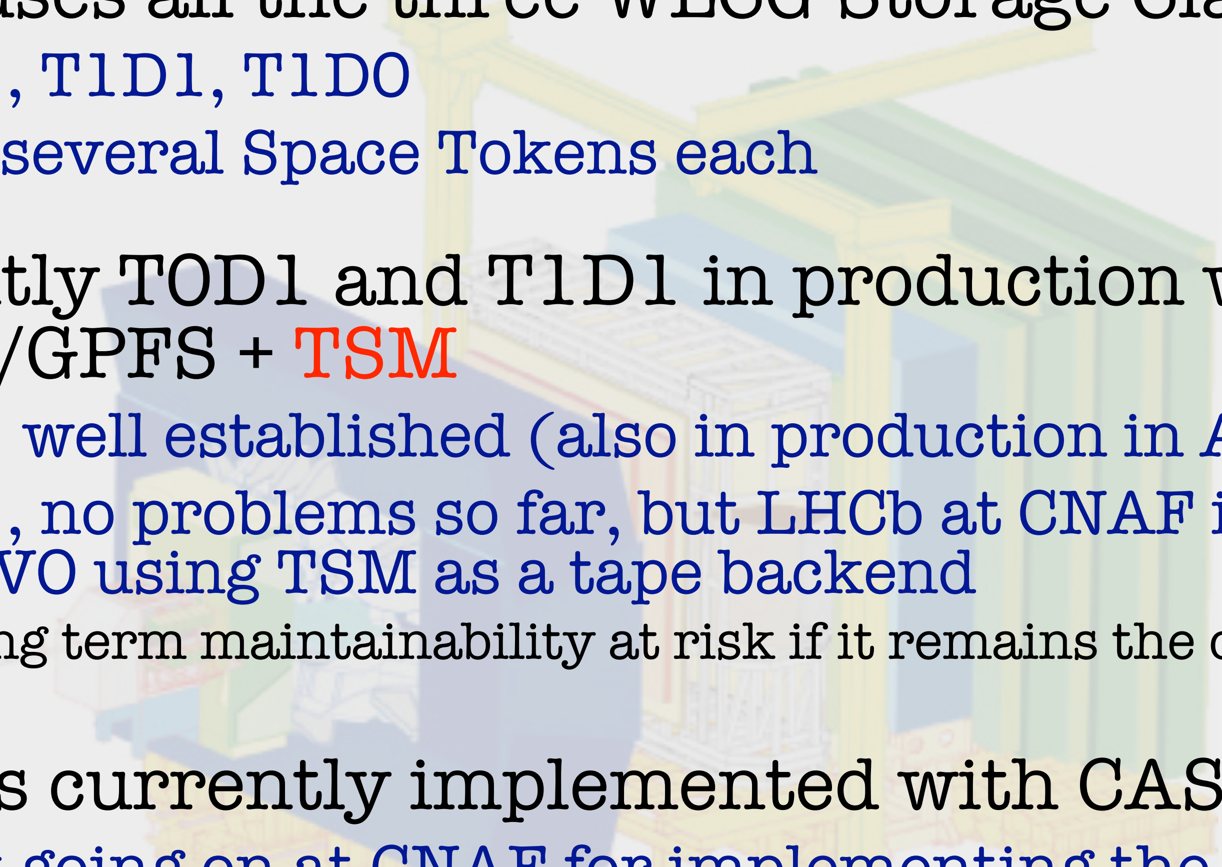
- Data transfers using FTS: OK
- SRM and data access for analysis is still an issue
  - Improved since last year, but
    - dCache sites still very problematic
    - CASTOR sites fairly better, but still far from optimal
    - StoRM/GPFS looks like a good marriage (see Vincenzo's talk)
- Synchronization loss between LFC and Storage
  - General problem, due to failures of several kinds
  - Can be cured, but very time (and man power) consuming
- LHCb book-keeping: new implementation in production
  - Good performance, still to be improved data sanity checks
- Software area at CNAF has been a severe show-stopper for some months
  - Problems seem solved after SW area moved from GPFS to CNFS over GPFS (see Vladimir's talk)

DATA ANALYSIS PERFORMANCE



CPU/WallClock for Successful Jobs

CNAF STORAGE SOLUTIONS



- LHCb uses all the three WLCG Storage Classes
  - T0D1, T1D1, T1D0
  - With several Space Tokens each
- Presently T0D1 and T1D1 in production with StoRM/GPFS + TSM
  - T0D1 well established (also in production in ATLAS)
  - T1D1, no problems so far, but LHCb at CNAF is the only VO using TSM as a tape backend
    - Long term maintainability at risk if it remains the only one
- T1D0 is currently implemented with CASTOR
  - Work going on at CNAF for implementing the T1D0 functionality with StoRM/GPFS/TSM
    - For LHCb it would be crucial in order to have a uniform access to all storage resources and namespaces

CONCLUSION: COMMON CONCERNS

- Prestaging strategies and tools
- SRM scaling (e.g. srmLs usage)
- GPFS optimisation
- TSM migration for T1D0
- Migration towards StoRM at Tier-2s

THANKS!