LFC LHCb

Barbara Martelli
WLCG Workshop

CERN 22- 26 January 2007

Outline
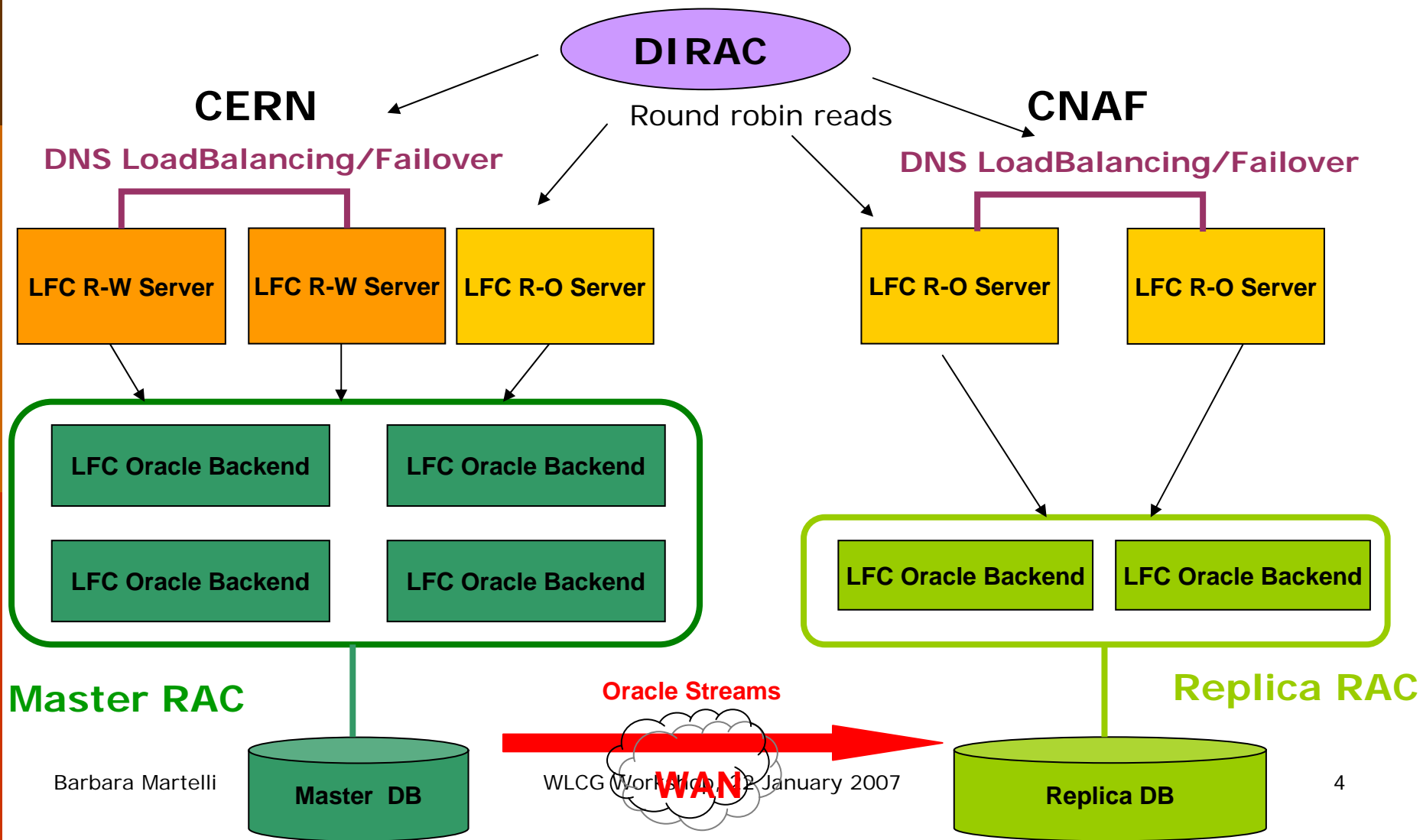
- LFC LHCb architecture and deployment
- Present usage
- Stress test on replication
- Conclusions

LHCb LFC read – only replica

- Why: LFC LHCb needs more scalability and fault tolerance.

- How: adding a read only LFC replica at CNAF. The CNAF slave is maintained consistent with the CERN master at the database level through the *Oracle Streams* replication.

- A *Streams* sustained rate of 170 KB/s is shown by preliminary throughput tests.

- LFC preliminary tests showed a sustained rate of 40 entries per second (entries inserted in the master and replicated at CNAF).

LFC LHCb Deployment

DIRAC

CERN

Round robin reads

CNAF

DNS LoadBalancing/Failover

DNS LoadBalancing/Failover

LFC R-W Server   LFC R-W Server   LFC R-O Server

LFC R-O Server   LFC R-O Server

LFC Oracle Backend   LFC Oracle Backend

LFC Oracle Backend   LFC Oracle Backend

LFC Oracle Backend   LFC Oracle Backend

Master RAC

Replica RAC

Oracle Streams

WAN

Master DB

Replica DB

LFC Read Only Configuration

At LFC level, very simple configuration, in:

/etc/sysconfig/lfcdaemon

set

RUN_READONLY="yes"

Then the LFC will prevent write operations:

$ lfc-mkdir /grid/dteam/hello cannot create /grid/dteam/hello: Read-only file system

This is supported from **LCG-2_7_0**

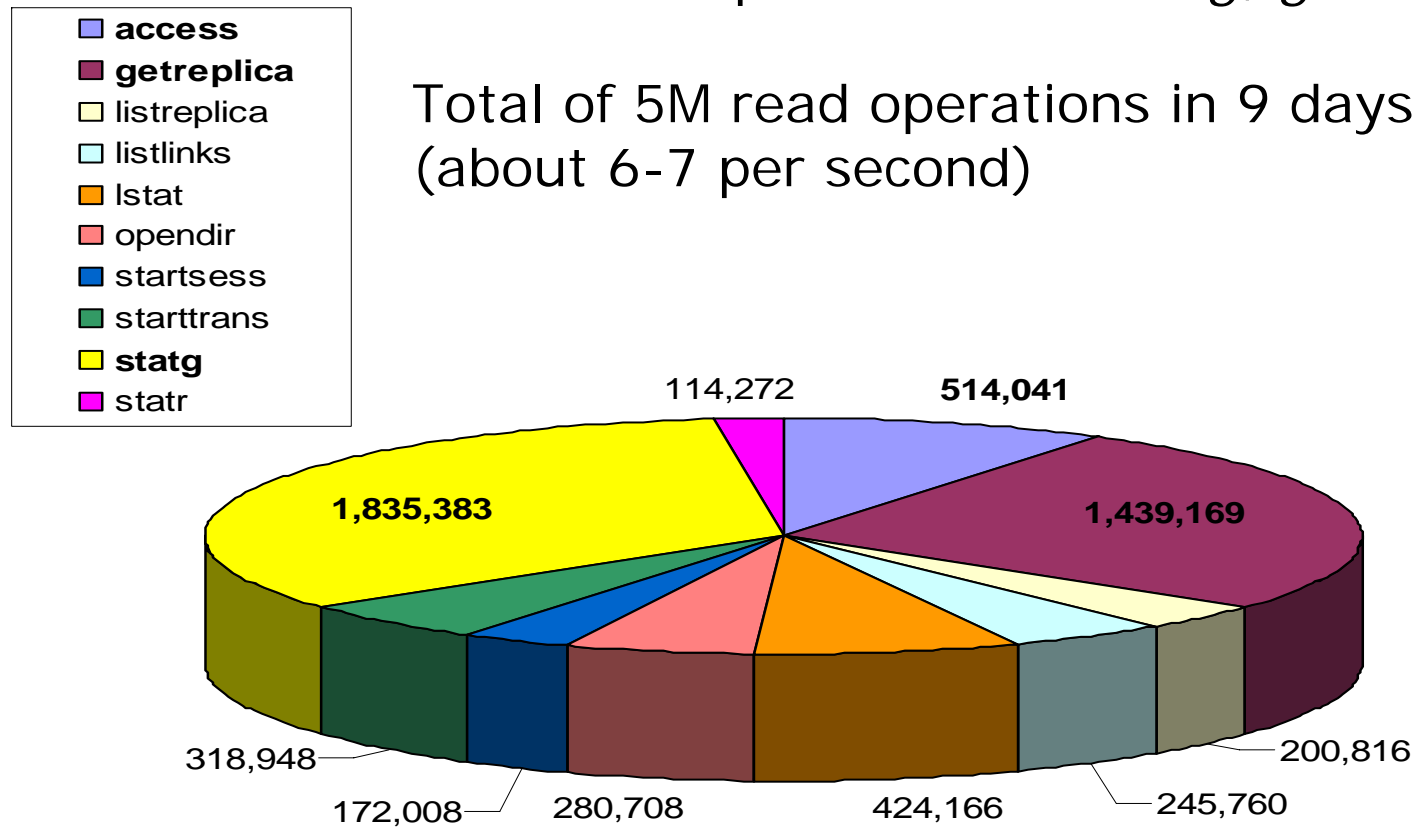At DB level we use an user with read only grants on LFC tables.

LFC Present Usage

- Jan 8 to Jan 16 (9 days) LFC logs analized.
- A rough estimate of only 10 LFC API calls per second (in total over the 5 LFCs).
- Write operations are the 0.5% of the total traffic.
- Write operations are well distributed among the 2 LFC R/W servers
- Read operations seem to privilege CERN LFCs.

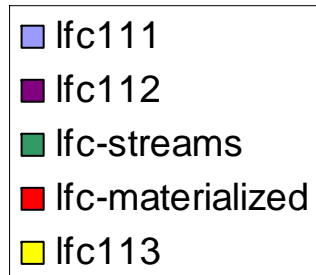# Present Usage – Read Operations
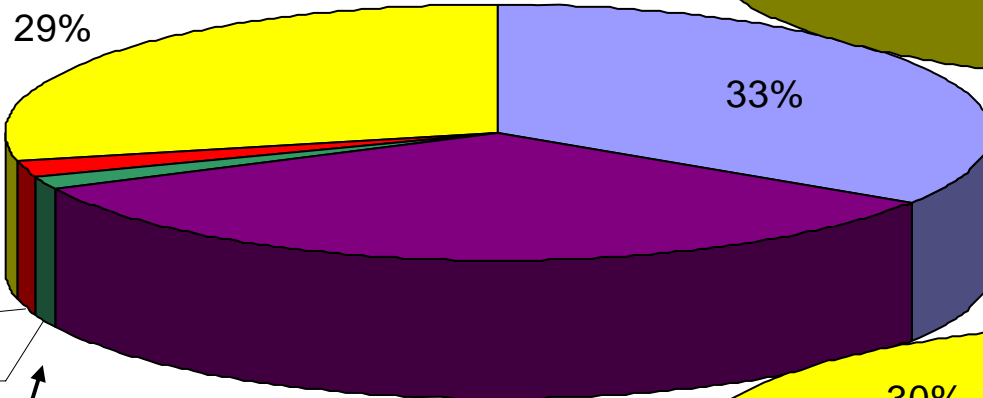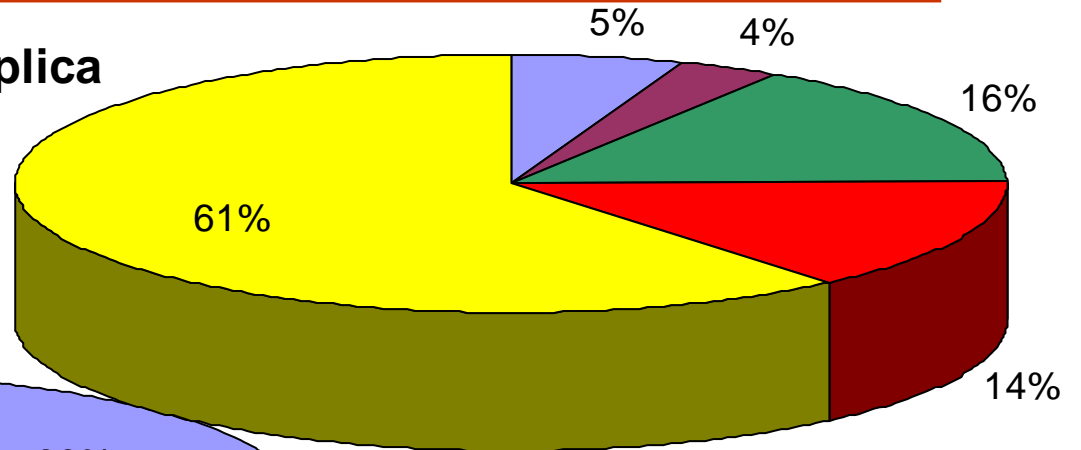
Most used operations are *statg, getreplica, access*

Total of 5M read operations in 9 days
(about 6-7 per second)

Legend:
- ☐ **access**
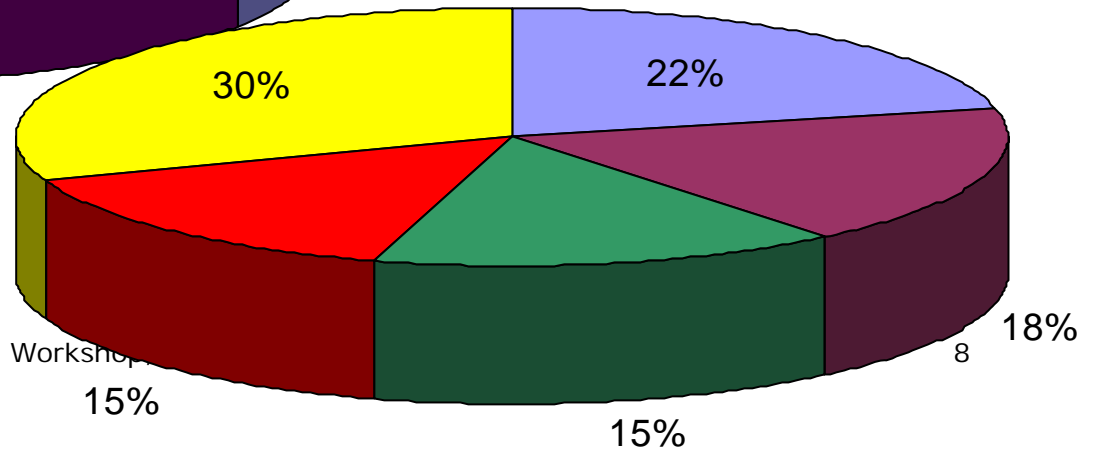- ☐ **getreplica**
- ☐ listreplica
- ☐ listlinks
- ☐ lstat
- ☐ opendir
- ☐ startsess
- ☐ starttrans
- ☐ **statg**
- ☐ statr

Pie chart values:
- 114,272
- 514,041
- 1,439,169
- 1,835,383
- 318,948
- 172,008
- 280,708
- 424,166
- 245,760
- 200,816

Read Operations Share on LFCs

Legend:
- lfc111
- lfc112
- lfc-streams
- lfc-materialized
- lfc113

getreplica

- 5%
- 4%
- 16%
- 14%
- 61%

statg

- 29%
- 33%
- 2%
- 2%
- 34%

access

- 22%
- 18%
- 15%
- 15%
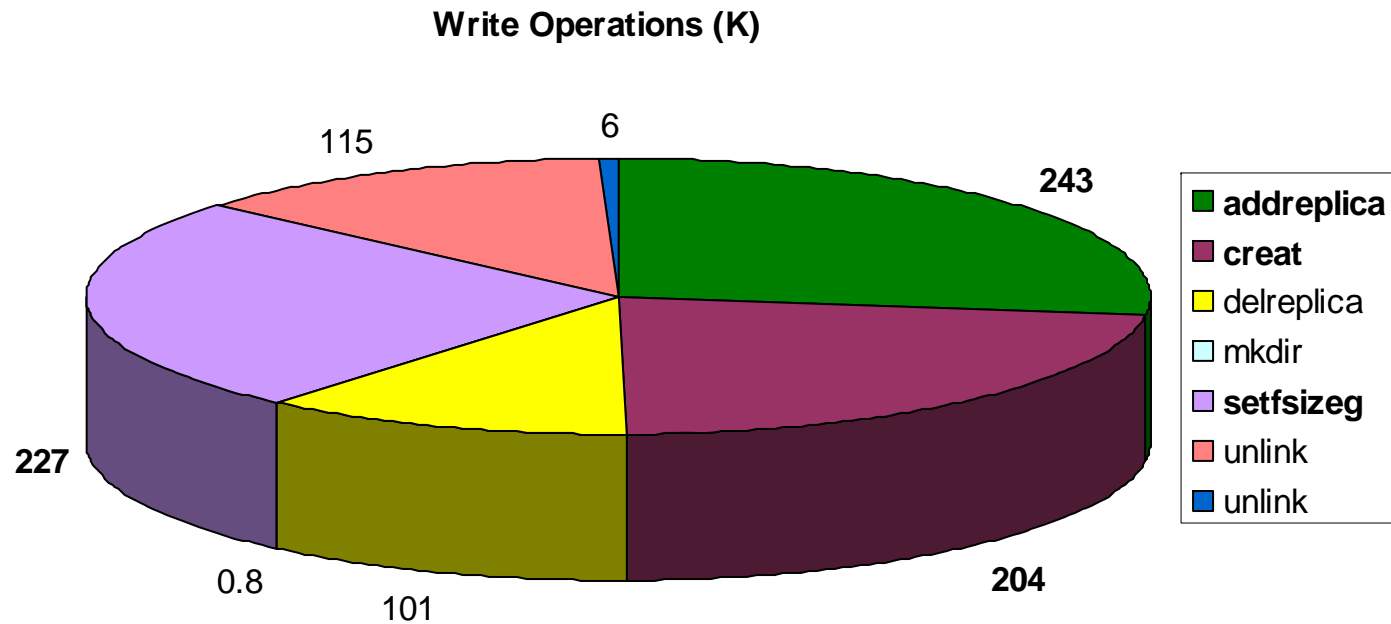- 30%

CNAF's LFC less used

Barbara Martelli

WLCG Workshop

8

Present Usage -Write Operations

Most used operations are *addreplica*, *setfsizeg*, *creat*

Total of 121K write operations in 9 days
(about 9 per *minute*)

Write Operations (K)

Legend:
- **addreplica** (dark green)
- **creat** (dark red/maroon)
- delreplica (yellow)
- mkdir (light cyan)
- **setfsizeg** (light purple)
- unlink (salmon/pink)
- unlink (blue)

Values shown: 6, 243, 115, 204, 227, 0.8, 101

Stress Test Setup I

- Python script which adds files in
  /grid/lhcb/lfc_test/cnaf

- For each file 5 replicas are added.

- For each file the related replicas are added all in the same session (to avoid authentication overload in LFC frontends).

- Test duration: 25K seconds (~7 hours), 800K files, 2M replicas.

Stress Test Setup II
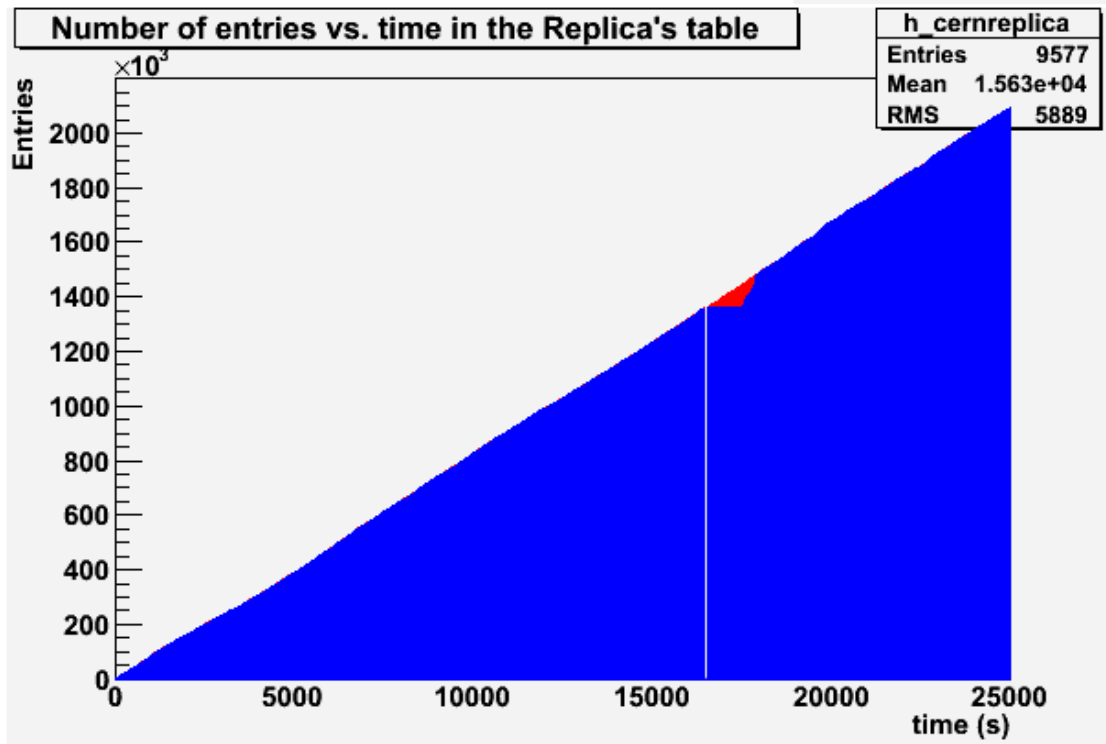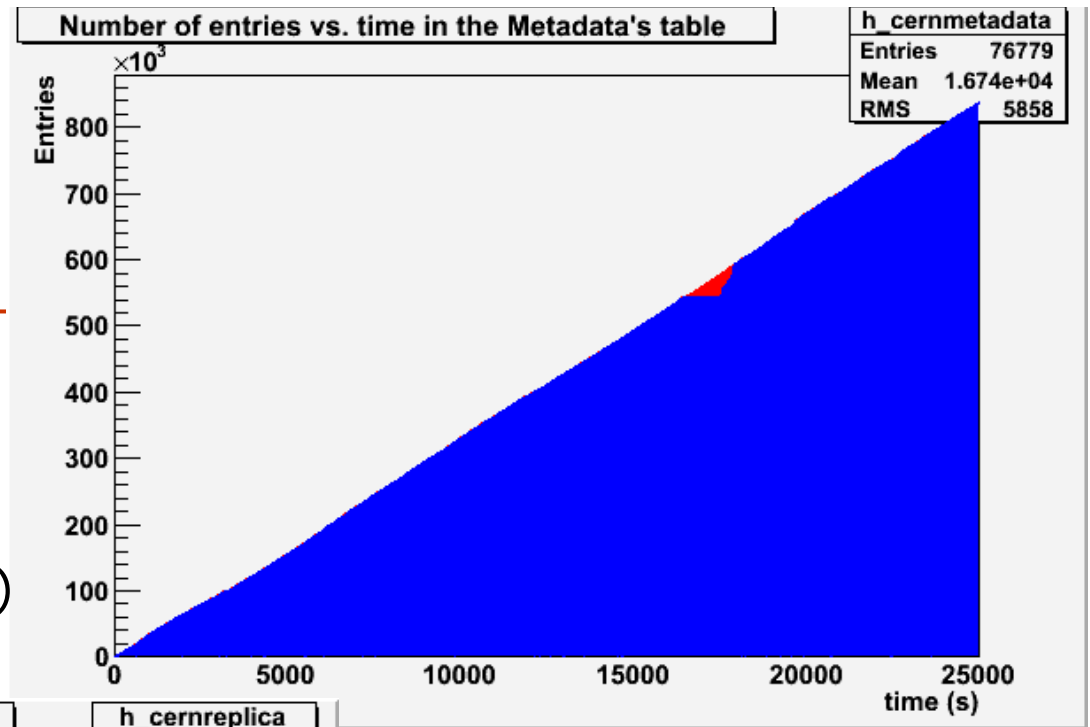
- At database level for each new file a row is inserted in the CNS_FILE_METADATA table. For each replica a row is added into CNS_FILE_REPLICA table.

- The script monitors the status of the master and slave database backends once a second. It counts the number of files and replicas present in each of them.

- We use the *strmmon* monitor to mesure the replication latency at DB level.
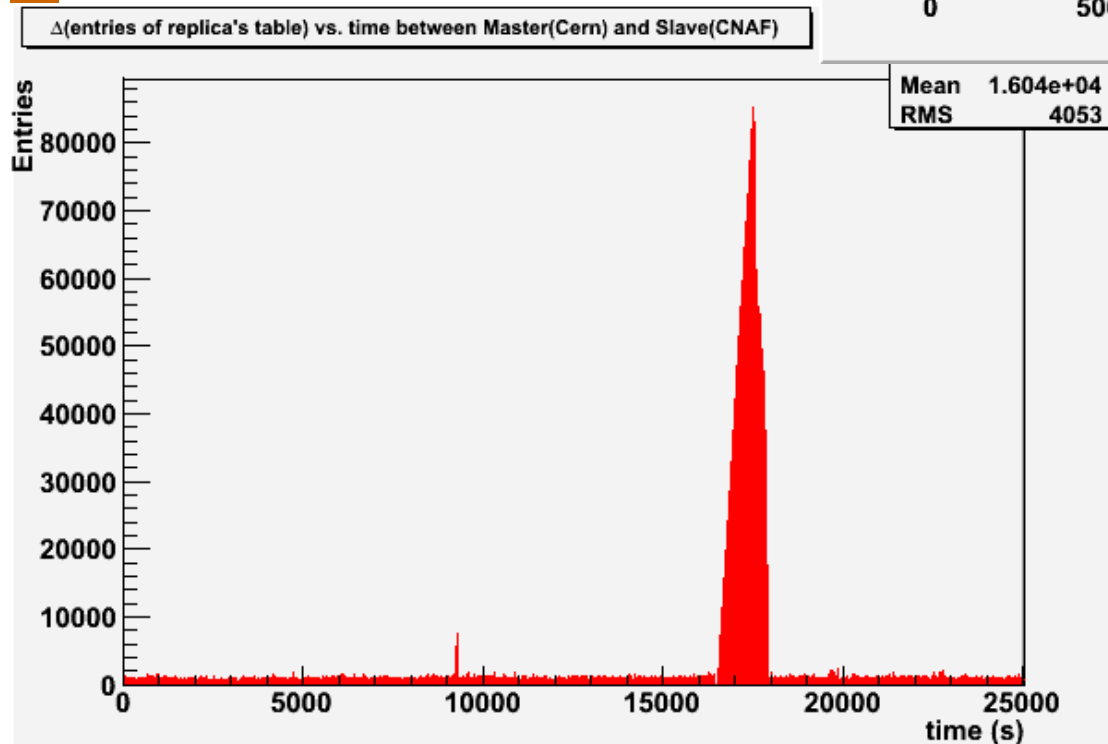
Population Rate

■ CERN

■ CNAF

25000 seconds (7 hours test)

Number of entries vs. time in the Metadata's table

h_cernmetadata
Entries   76779
Mean   1.674e+04
RMS   5858

Number of entries vs. time in the Replica's table

h_cernreplica
Entries   9577
Mean   1.563e+04
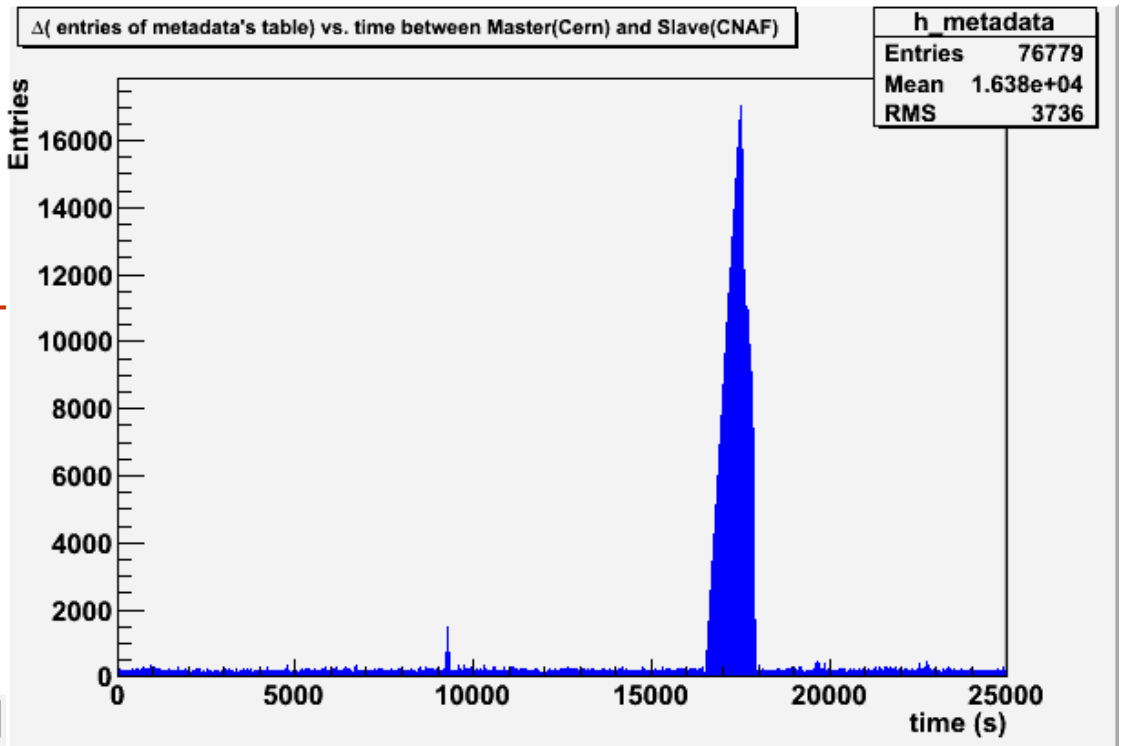RMS   5889

- Sustained rate of 84 entries/s on CNS_FILE_REPLICA
- Sustained rate of 36 entries/s on CNS_FILE_METADATA
- Total: 120 entries/s
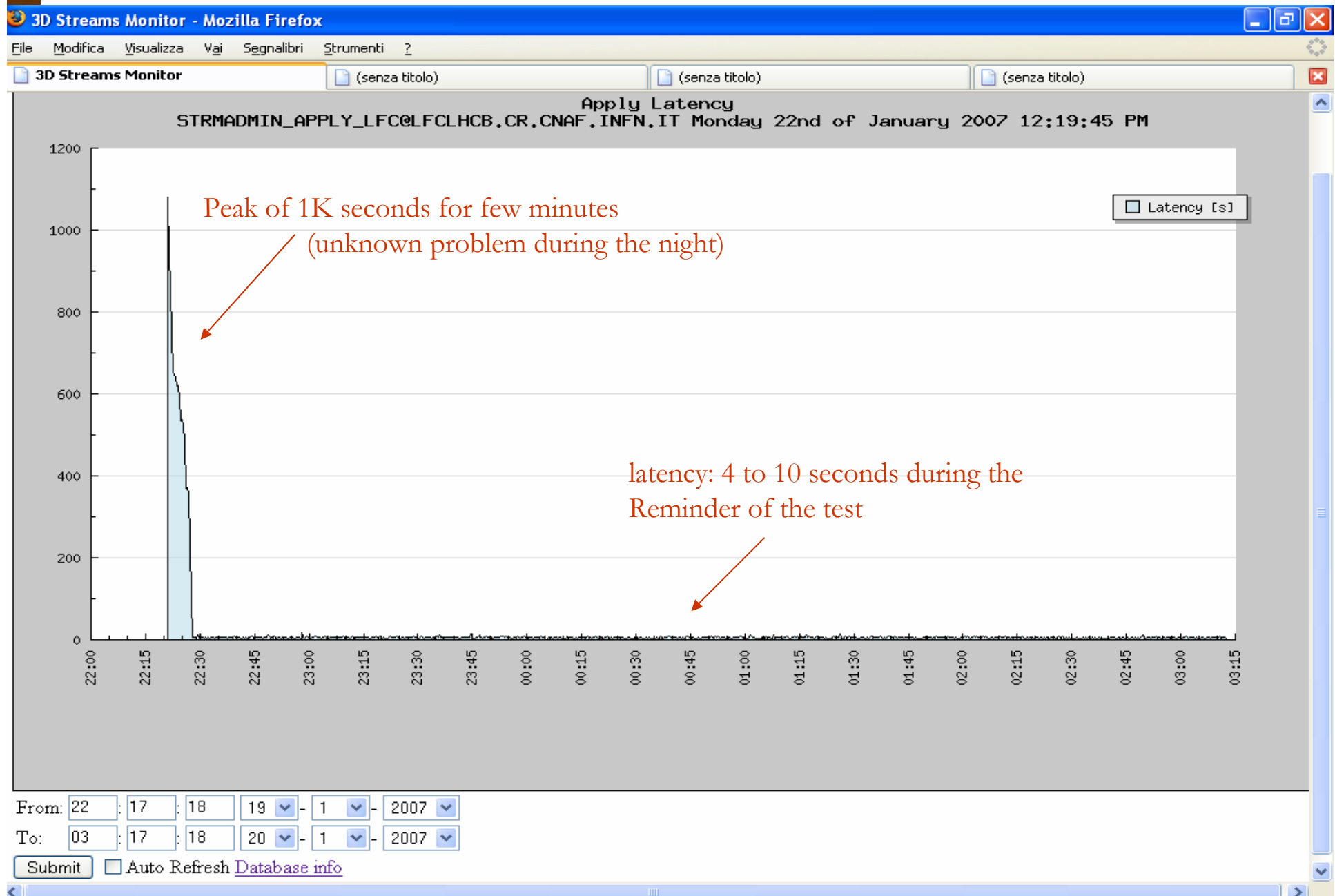- Only once CNAF's slave become misallignmed for about 1 hour

Difference between entries in Master LFC and CNAF's Slave

Δ(entries of metadata's table) vs. time between Master(Cern) and Slave(CNAF)

h_metadata	
Entries	76779
Mean	1.638e+04
RMS	3736

Δ(entries of replica's table) vs. time between Master(Cern) and Slave(CNAF)

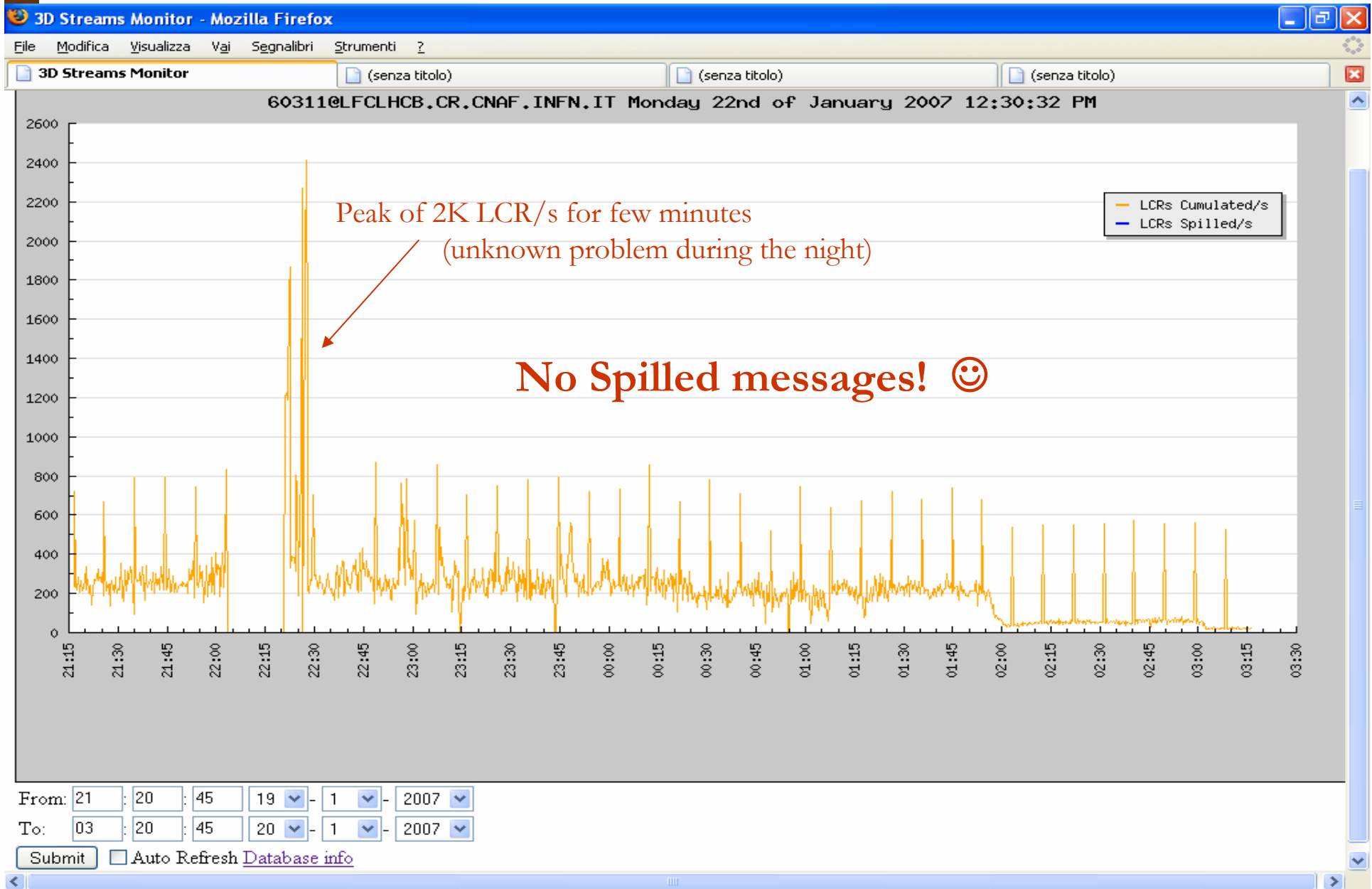|	
Mean	1.604e+04
RMS	4053

- Usually not higher than 200 entries in CNS_FILE_METADATA and 1000 in CNS_FILE_REPLICA
- Peak of some K entries for about 1hr
- CNAF's DB succeeded to apply all missing entries.

Strmmon View: Replication Latency

Strmmon View: Spilled Messages

3D Streams Monitor - Mozilla Firefox

File Modifica Visualizza Vai Segnalibri Strumenti ?

3D Streams Monitor | (senza titolo) | (senza titolo) | (senza titolo)

60311@LFCLHCB.CR.CNAF.INFN.IT Monday 22nd of January 2007 12:30:32 PM

Peak of 2K LCR/s for few minutes
(unknown problem during the night)

No Spilled messages! ☺

Legend:
— LCRs Cumulated/s
— LCRs Spilled/s

From: 21 : 20 : 45 19 - 1 - 2007
To: 03 : 20 : 45 20 - 1 - 2007

Submit   ☐ Auto Refresh Database info

Conclusions

- LFC LHCb deployment is stable.
- Sustained rate of 120 entries/s has been achieved
- NO SPILLING has been detected in case of database misalignment between master and slave. The messages are accumulated by the master only in the archived log.
- The Master-slave synchronization has been exercised.
- LFC at CNAF are less used for read operations than the CERN ones. We should analyze this anomalous behavior.
- Many thanks to Angelo Carbone and Gianluca Peco who made the stress test for LHCb.