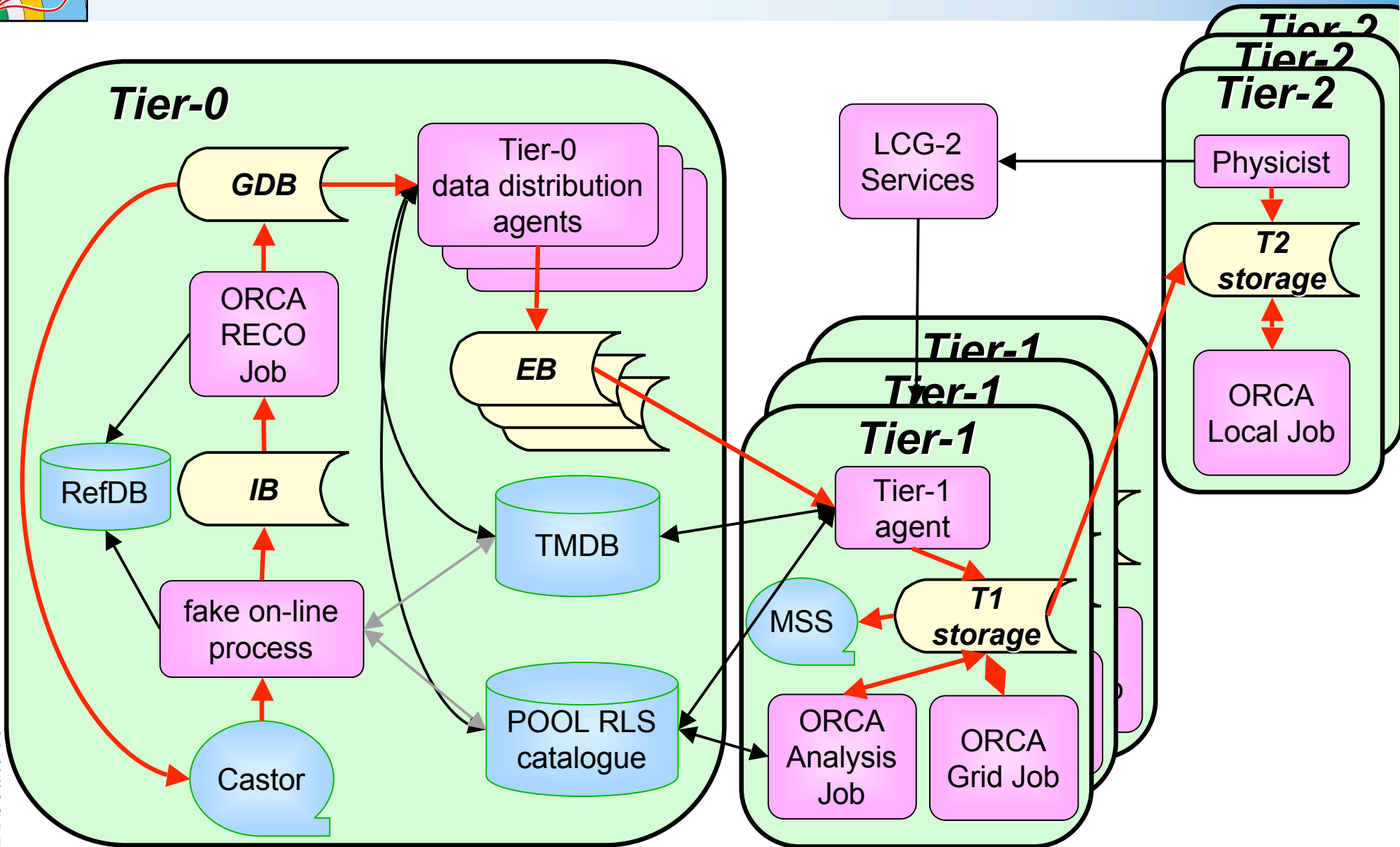


DC04 layout

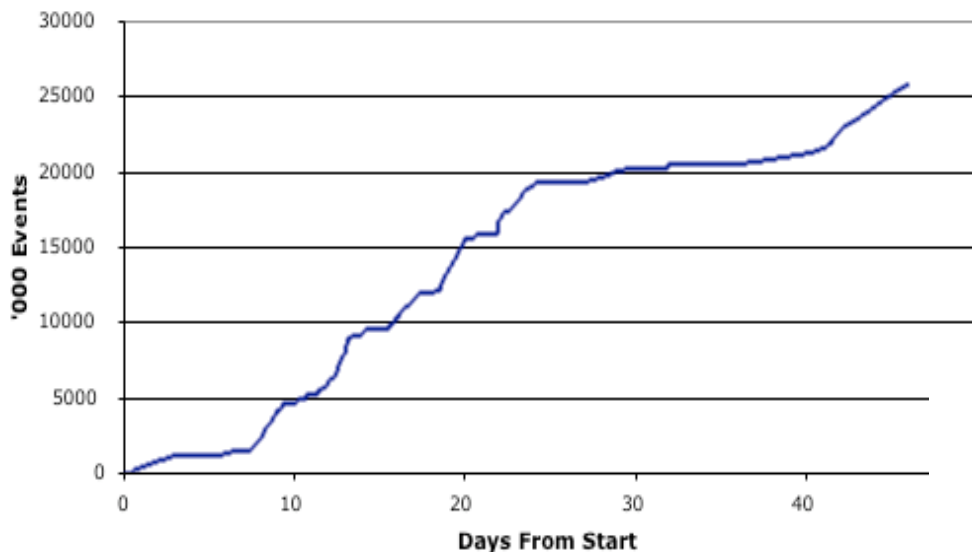


DPS June 04



DC04 Processing Rate

T0 Events Per Time



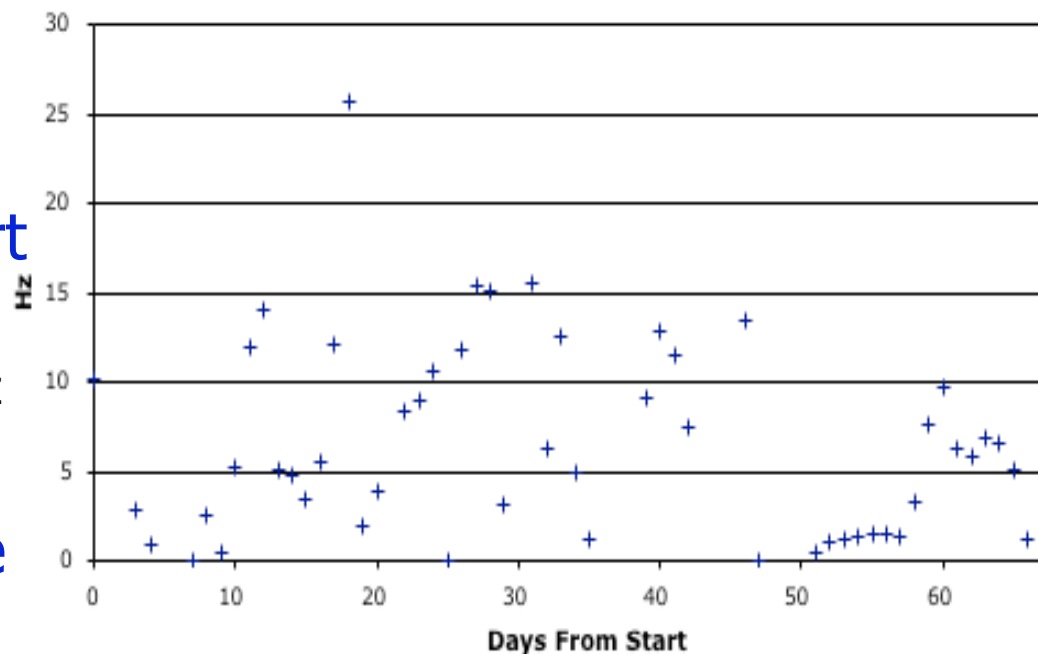
- ❖ Processed about 30M events
 - ◆ But DST “errors” make this pass not useful for analysis
- ❖ Generally kept up at T1’s in CNAF, FNAL, PIC

❖ Got above 25Hz on many short occasions

- ◆ But only one full day above 25Hz with full system

❖ Working now to document the many different problems

Event Processing Rate





Processing Cluster Utilization

Cluster info: CMS CPU nodes

01 May 2004 Sat 19:18

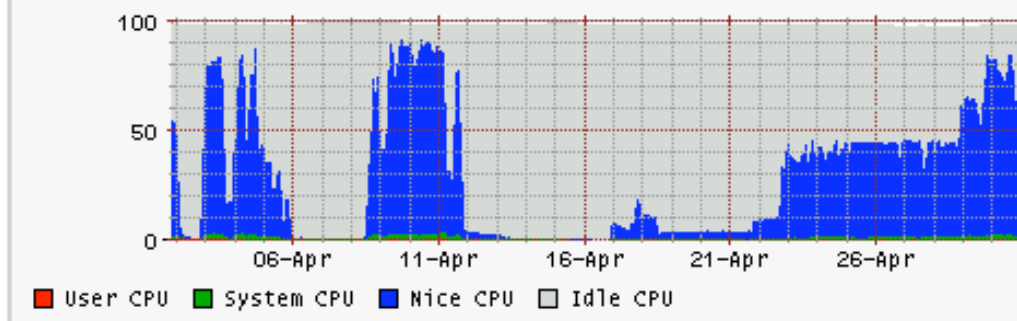
Cluster Information

of hosts (down): 220 (5)
operating systems: 2.4.20-30.7.cernsmp, 2.4.20-28.7.cernsmp
of CPUs (down): 440 (10)
average up time: 59 days, 16h:38m (boots per host)
hosts down: lxb0579, lxb1221, lxb1233, lxb1256, lxb1279

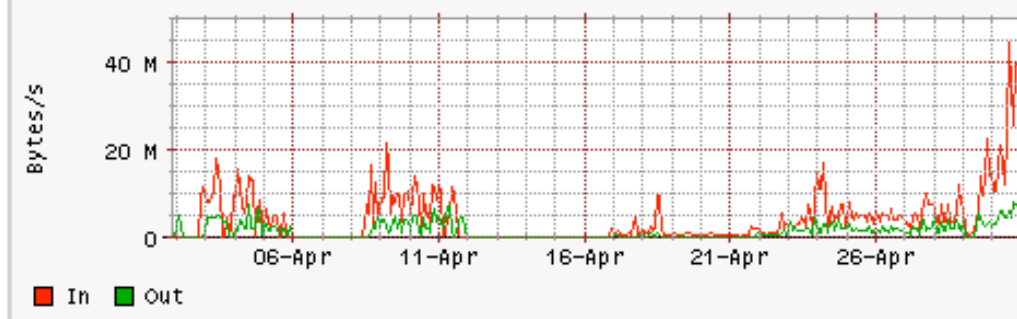
Select from hosts:

None

CPU utilization - last month



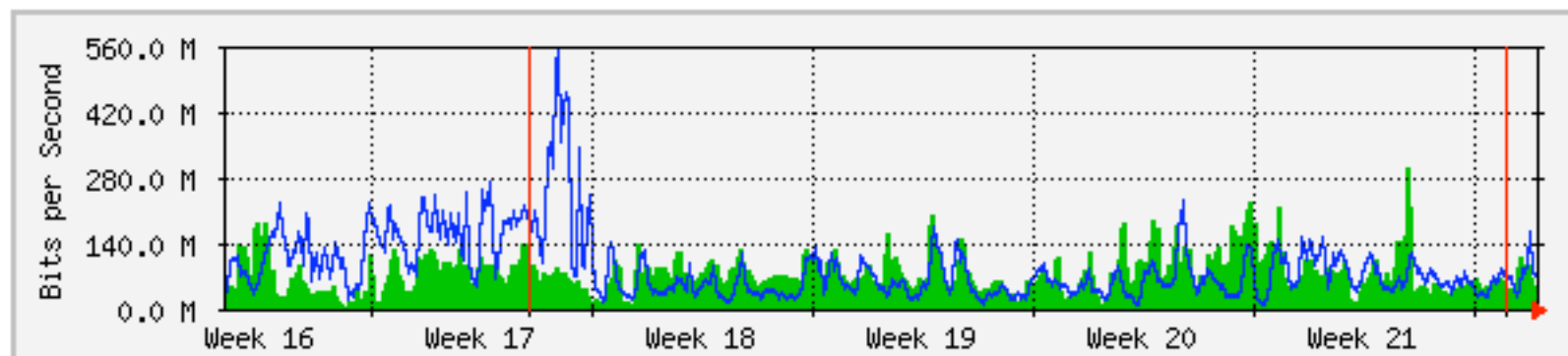
Network utilization - last month





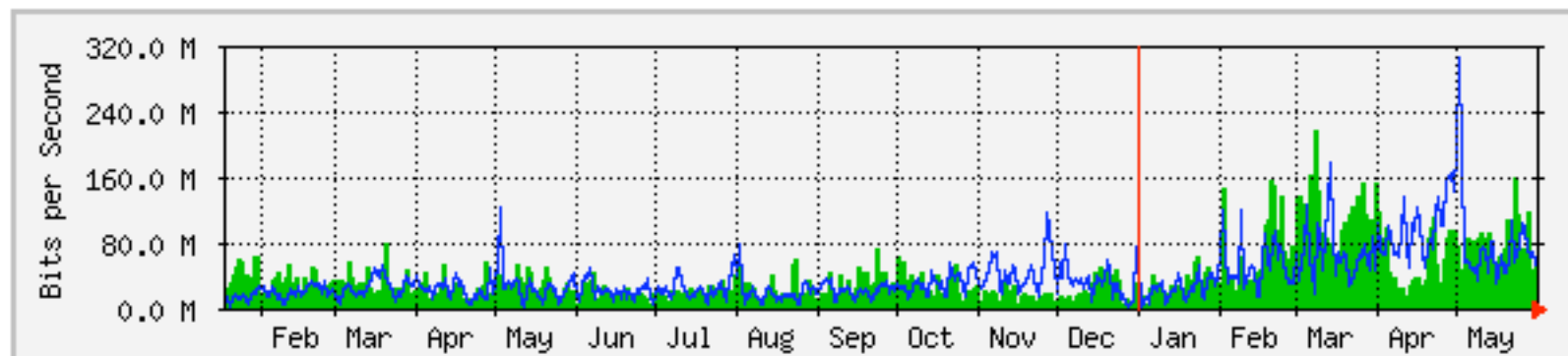
Pushing the Networks

'Monthly' Graph (2 Hour Average)



Max **In**: 305.2 Mb/s (30.5%) Average **In**: 78.0 Mb/s (7.8%) Current **In**: 53.4 Mb/s (5.3%)
Max **Out**: 552.1 Mb/s (55.2%) Average **Out**: 90.7 Mb/s (9.1%) Current **Out**: 67.8 Mb/s (6.8%)

'Yearly' Graph (1 Day Average)



Max **In**: 217.7 Mb/s (21.8%) Average **In**: 39.2 Mb/s (3.9%) Current **In**: 54.0 Mb/s (5.4%)
Max **Out**: 306.1 Mb/s (30.6%) Average **Out**: 39.2 Mb/s (3.9%) Current **Out**: 56.2 Mb/s (5.6%)



LCG-2 in DC04

Aspects of DC04 involving LCG-2 components

- ◆ register all data and metadata to a world-readable catalogue
 - RLS
 - ◆ transfer the reconstructed data from Tier-0 to Tier-1 centers
 - Data transfer between LCG-2 Storage Elements
 - ◆ analyze the reconstructed data at the Tier-1's as data arrive
 - Real-Time Analysis with Resource Broker on LCG-2 sites
 - ◆ publicize to the community the data produced at Tier-1's
 - Not done, but straightforward using the usual Replica Manager tools
 - ◆ end-user analysis at the Tier-2's (not really a DC04 milestone)
 - first attempts
 - ◆ monitor and archive resource and process information
 - GridICE
-
- ❖ Full chain (except Tier-0 reconstruction) could be performed in LCG-2



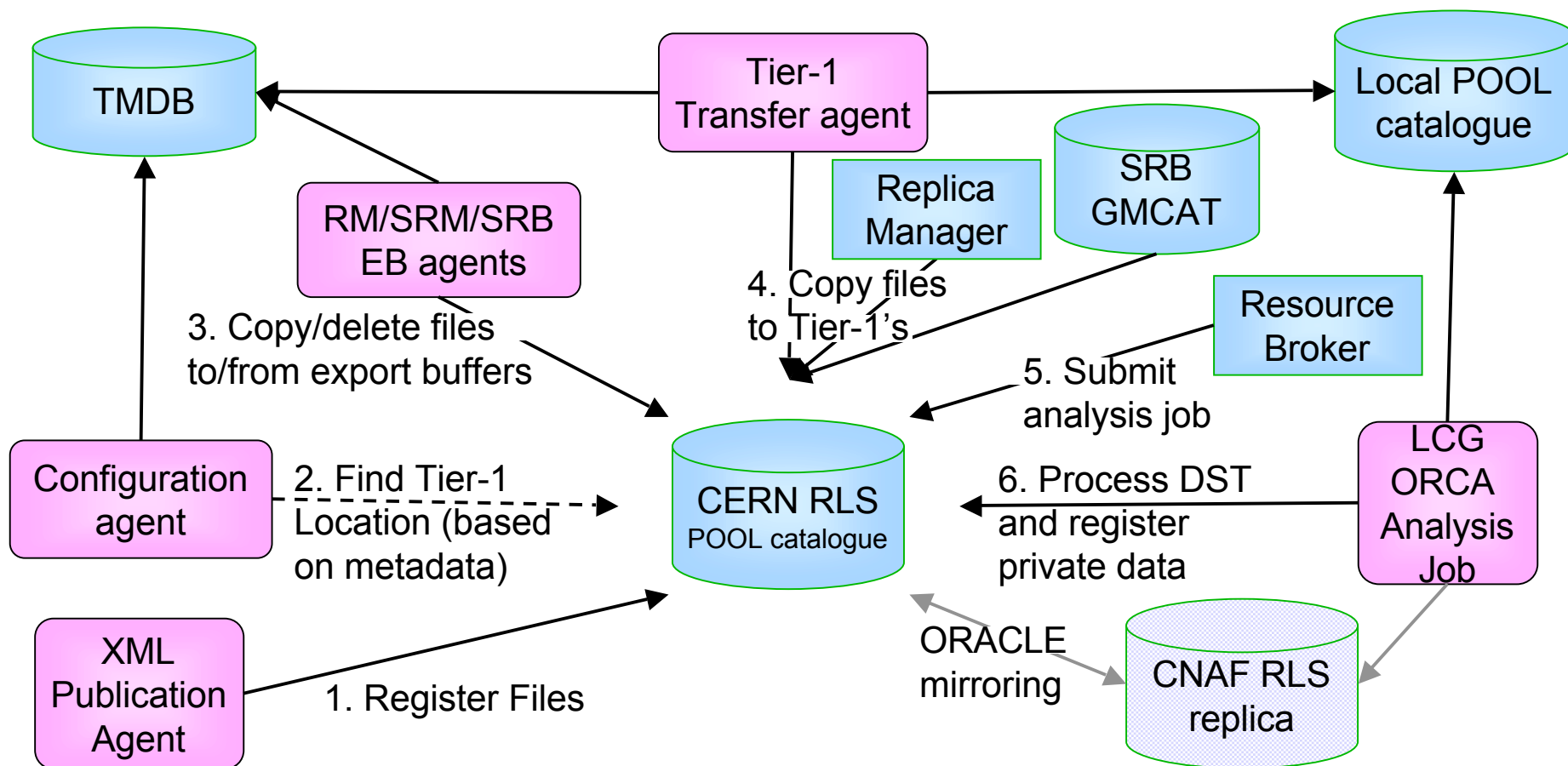
Interactions with RLS in DC04

RLS use as the global POOL catalogue:

- ❖ **Publishing Agent: Register files produced at Tier-0 into RLS**
 - ◆ Transfer POOL XML catalogue fragment into RLS; POOL CLI (`FCpublish`)
- ❖ **Configuration Agent: Query the RLS metadata to assign files to Tier-1**
 - ◆ POOL CLI (`FClistMeta`) ... all data sent everywhere, not truly used
- ❖ **Export Buffer Agents: Insert/delete the PFNs for the files in the buffer**
 - ◆ POOL CLI (`FCaddReplica`), C++ LRC API-based programs, LRC java API by GMCAT
- ❖ **Tier-1 Agents: Insert PFN for the destination location upon transfer**
 - ◆ In some cases copy the RLS into local MySQL POOL catalogue
 - ◆ POOL CLI (`FCaddReplica`, `FCpublish`), C++ LRC-API based programs, LRC java API by GMCAT
- ❖ **Analysis jobs on LCG**
 - ◆ Use the RLS through the Resource Broker to submit jobs close to the data
 - ◆ Register the private output data into RLS
- ❖ **Humans trying to figure what was going on (`FCpublish` commands)**



Description of RLS usage in DC04



Specific client tools: POOL CLI, Replica Manager CLI, C++ LRC API based programs, LRC java API tools (SRB/GMCAT), Resource Broker



RLS as Global POOL Catalogue

RLS used as a global POOL catalogue, with full file meta data

❖ File information by GUID

- ◆ LFN
- ◆ PFNs for every replica
- ◆ Meta data attributes

❖ Meta data schema handled and pushed into catalogues by POOL

- ◆ Some attributes are highly CMS-specific

❖ CMS does not use a separate file catalogue for meta data

Some sites had local catalogues that copied RLS partially

❖ POOL MySQL catalogue with all the files, but only local PFNs

- ◆ On file copy, copy also the catalogue entries to the local catalogue

❖ Dump the catalogue needed by an analysis job into a XML file

- ◆ Job looked up (some?) files in the XML catalogue



RLS issues

❖ Total Number of files registered in the RLS during DC04:

u ~ 570K LFNs each with ~ 5-10 PFN's and 9 metadata attributes

❖ Inserting information into RLS

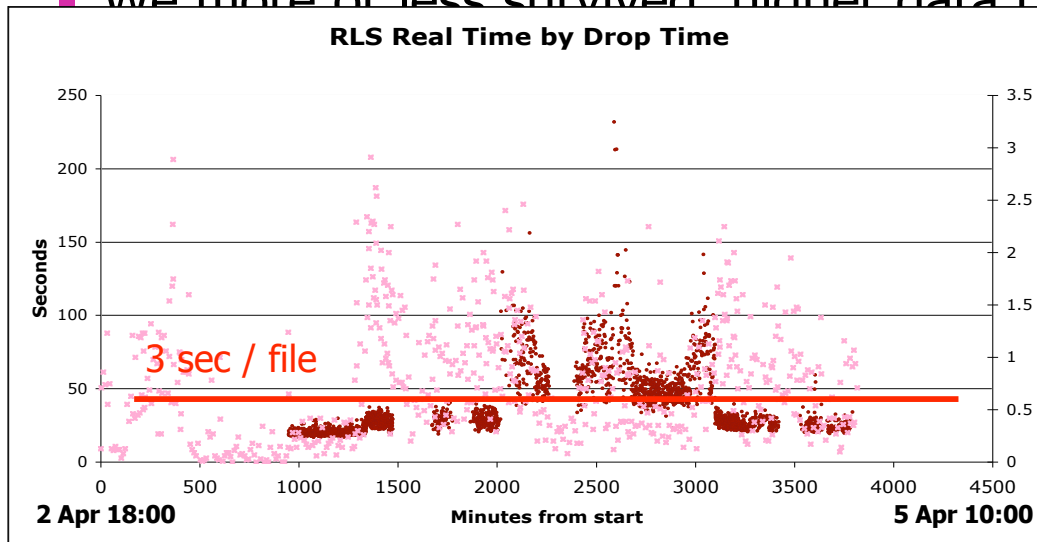
◆ Insert PFN (file catalogue) was fast enough if using the appropriate tools, produced in-course

▪ LRC C++ API programs (~0.1-0.2sec/file), POOL CLI with GUID (secs/file)

◆ Insert files with their attributes (file and metadata catalogue) was slow

▪ ~~We more or less survived, higher data rates would be troublesome~~

Time to register the output of a Tier-0 job (16 files)



Sometimes the load on RLS increases and requires intervention on the server (i.g. log partition full, switch of server node, un-optimized queries)

⇒ able to keep up in optimal condition, so and so otherwise



RLS issues (II)

❖ Querying information from RLS

- ◆ Looking up file information by GUID seems sufficiently fast
- ◆ Bulk queries by GUID take a long time (seconds per file)
- ◆ Queries on metadata are too slow (hours for a dataset-owner)
- ◆ Direct access to the RLS Oracle backend was provided during DC04 to allow fast checking and dump of the RLS content (two minutes to suck the entire catalogue with room for optimisation)
- ◆ Dump from a POOL MySQL or XML catalogue is *minimum* factor 10 faster than dump from POOL RLS
 - A global MySQL catalogue could sustain the scale of DC04?

❖ Tests of RLS replica

- ◆ During DC04 the ORACLE single-master configuration was tested between CERN and CNAF



RLS issue (III)

- ❖ Several workarounds or solutions were provided to speed up the access to RLS during DC04
 - ◆ Replace (java) replica manager CLI with C++ API programs
 - ◆ POOL improvements and workarounds
 - ◆ Index some meta data attributes in RLS (ORACLE indices)

- ❖ Requirements not supported during DC04
 - ◆ Transactions
 - ◆ Small overhead compared to direct RDBMS catalogues
 - ◆ Fast queries



RLS Summary

- ❖ We survived
- ❖ Important performance issues found
 - ◆ Some are being addressed
 - ◆ Promising reports
- ❖ Some usability issues remain
 - ◆ Transactions anyone?
- ❖ Long-term future is seems blurry...
 - ◆ We like the file catalog concept: directory of files viewed different ways
 - ◆ RLS model may not be what CMS would ideally like to have
 - See discussion on file catalogue's role in resource broker
 - Our use of catalogues may be different in the next data challenge
 - ◆ RLS developers are changing implementation
 - ◆ Couldn't yet determine what EGEE is thinking of

Let's try to make sure we converge on something usable soon!



RLS Current Status

- ❖ Bulk functionalities are now available in RLS
 - ◆ Promising reports from the POOL developers
- ❖ Transactions still not supported
 - ◆ Bulk transfer poor man's replacement?
- ❖ Tests of RLS Replication current carried out by IT-DB
 - ◆ ORACLE streams-based replication mechanism
 - ◆ If RDMBS handles it, then RLS design needs revision?
- ❖ We are providing IT-DB statistical "dummy" agents
 - ◆ Act like our transfer agents (see previous slides)
 - ◆ Using timing statistics from TMDB and other agent logs

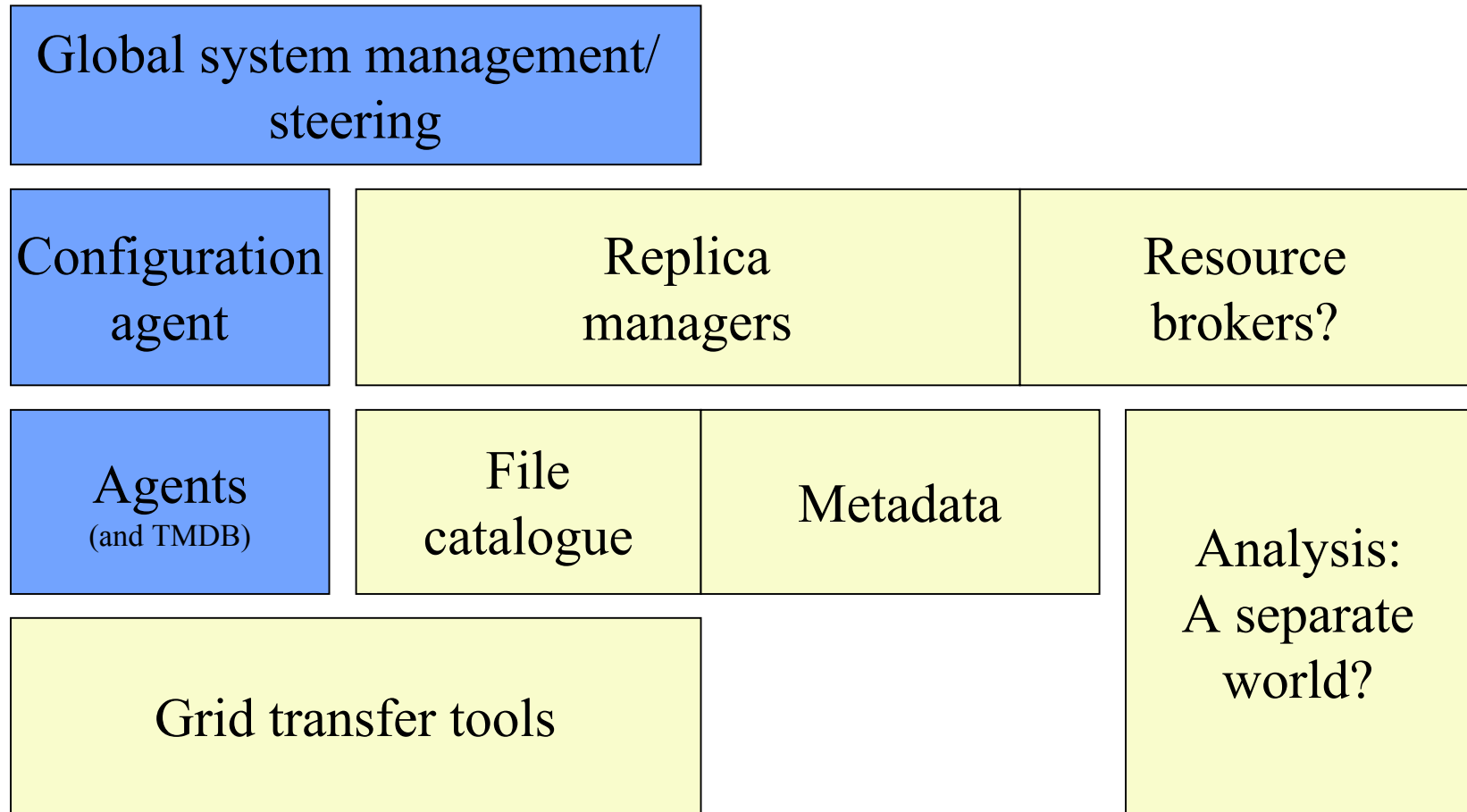


Data Transfer: Where Are We?

- ❖ Transfer Management Data Base and system of autonomous agents put together in a remarkably few weeks and performed very well
- ❖ We have a system able to sustain high-rate data transfer (About 1Gb/s), but with intervention by experts.
 - ◆ Can stream data from T0 to T1s for storage and realtime analysis (only 20 minute lag).
- ❖ A good idea of components:
 - ◆ Central database (TMDB)
 - Provides a context for communication between agents.
 - ◆ Agents to link data sources with distribution.
 - Publish to file catalogue, enter into distribution, merge files.
 - ◆ Agents to handle distribution.
 - Transfer files, manage intermediate buffers.
 - ◆ Agents to handle real time analysis.

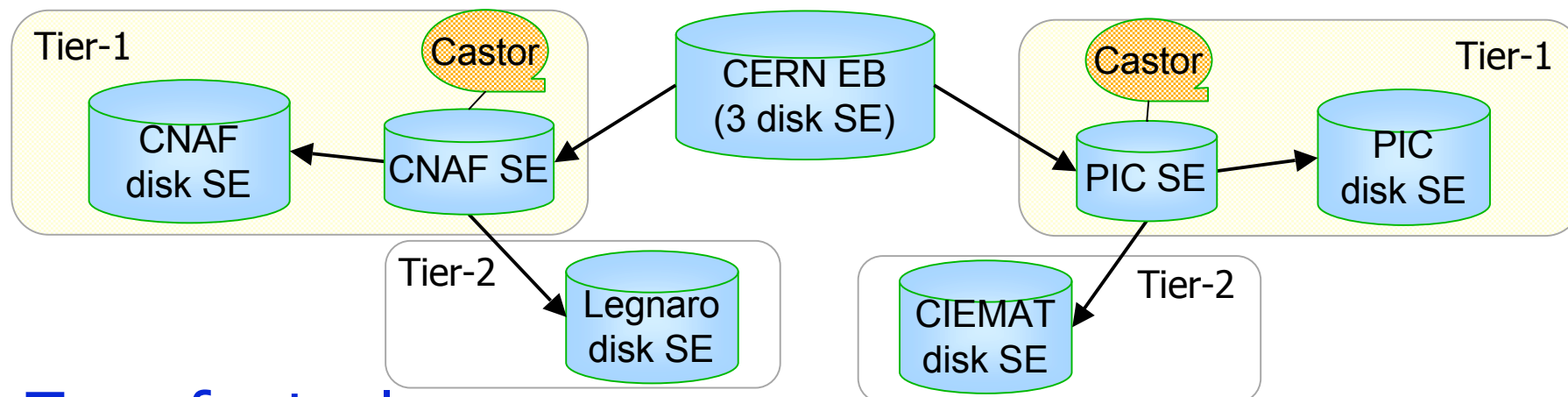


Context for the agent system





(LCG) Data Transfer



❖ Transfer tools:

- ◆ Replica Manager CLI used for EB → CNAF and CNAF → Legnaro
 - Java-based CLI introduces non negligible overhead at start-up
- ◆ globus-url-copy + LRC C++ API used for EB → PIC and PIC → Ciemat
 - Faster

❖ Performance has been good with both tools

- ◆ Total network throughput limited by small file size
- ◆ Some transfer problem caused by performance of underlying MSS
 - Always use a disk SE in front of an MSS in the future?



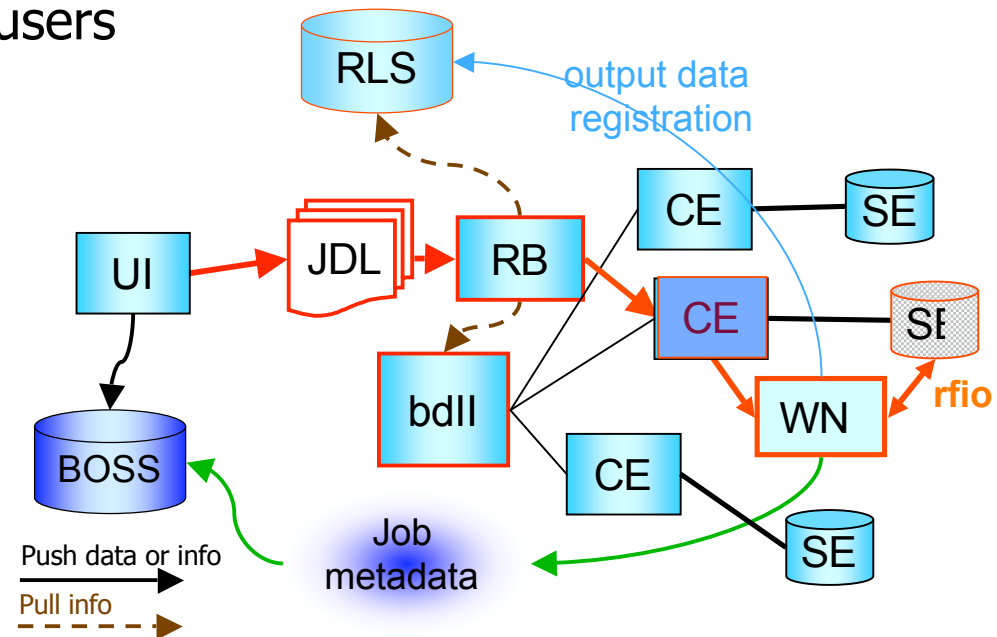
Analysis

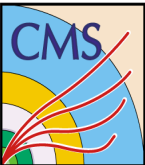
❖ CMS software installation

- ◆ CMS Software Manager installs software via a grid job provided by LCG
 - RPM distribution based on CMSI or DAR distribution
 - Used at CNAF, PIC, Legnaro, Ciemat and Taiwan with RPMs
- ◆ Site manager installs RPM's via LCFGng
 - Used at Imperial College
- ◆ Still inadequate for general CMS users

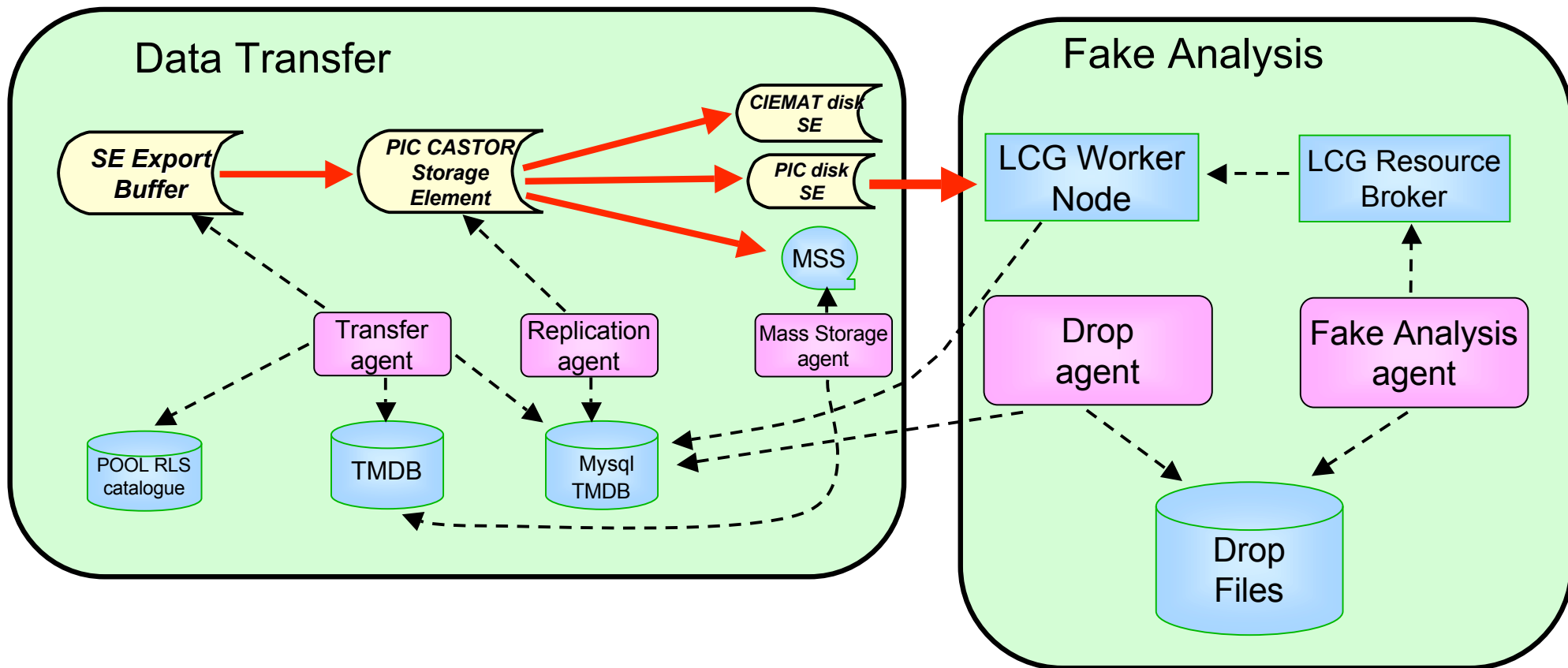
❖ Real-time analysis at Tier-1

- ◆ Main difficulty is to identify complete file sets (i.e. runs)
 - Information today in TMDB or via findColls
- ◆ Job processes single runs at the site close to the data files
 - File access via rfio
- ◆ Output data registered in RLS

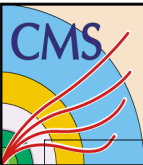




Fake Analysis Architecture

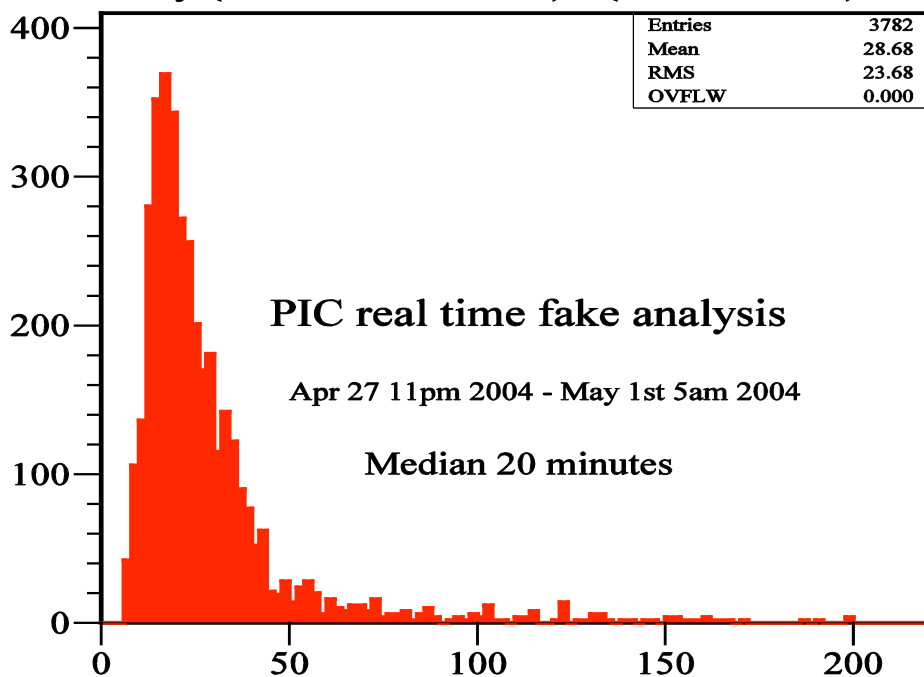


- ❖ Drop agent triggers job preparation/submission when all files are available
- ❖ Fake Analysis agent prepares xml catalog, orcarc, jdl script and submits job
- ❖ Jobs record start/end timestamps in mysql DB

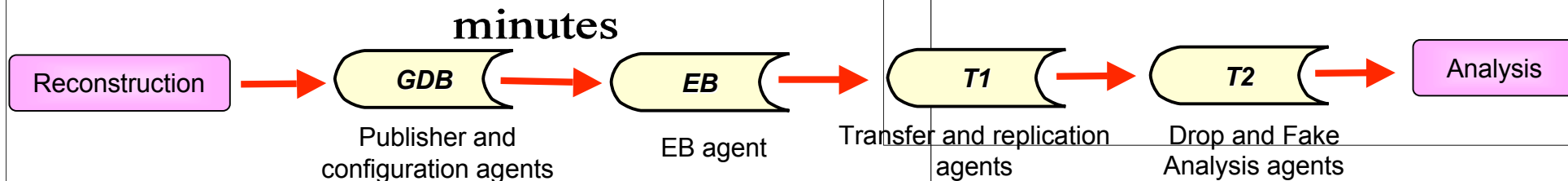
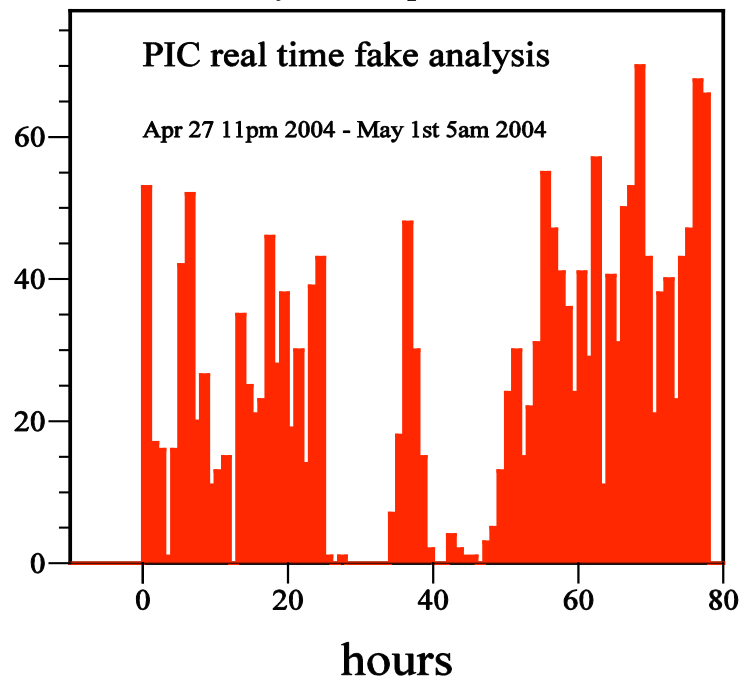


From GDB to analysis at T1

Time delay (File analyzed at T1) - (File on GDB)



Fake Analysis Jobs per hour



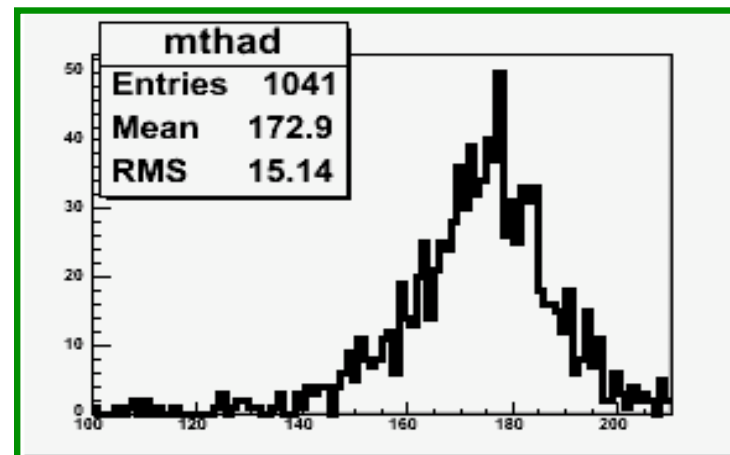


tTH analysis results

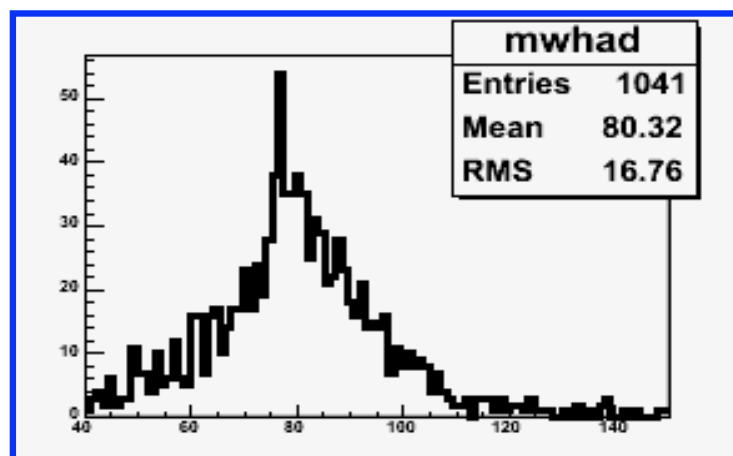


Reconstructed Masses:

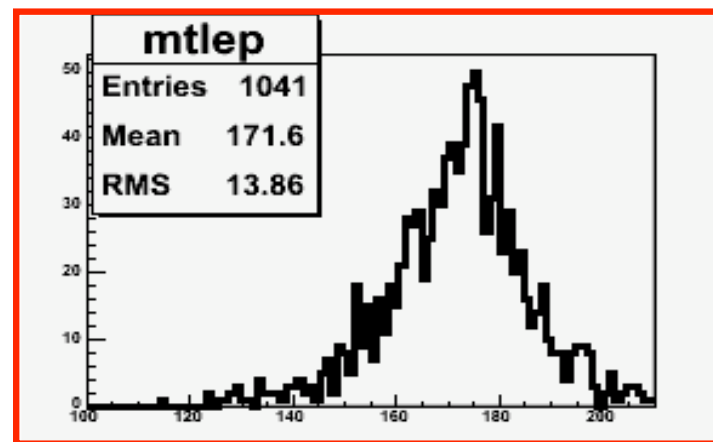
Hadronic Top



Hadronic W



Leptonic Top





Next plans

- ❖ Establish a Data Management task
 - ◆ Use LCG model of RTAG to report this summer
 - Physicists/ Computing to define CMS Blueprint, relationships with suppliers (LCG/EGGG.. CMS Dm task in Computing group
 - Expect to draw heavily on newly available effort and experience from Run II
- ❖ Establish a Workload Management task
 - ◆ Work closely with CMS-ARDA "DAPROM" and CMS-DM task
 - ◆ Make the Grid useable to CMS users
- ❖ Establish high-level Physics/Computing panel between t1 countries to ensure Collaboration Ownership of Computing Model for MoU and RRB discussions
- ❖ In broad strokes:
 - ◆ Our requirements/expectations from LCG are at the base component level.
 - ◆ We anticipate that CMS will need to provide the higher-level tools and services to map these to our computing model