Available on CMS information server **CMS CR 2004/054**

CMS Conference Report

October 6, 2004

Role of Tier-0, Tier-1 and Tier-2 Regional Centres in CMS DC04

T. Barrass, S. Metson, *Bristol University, United Kingdom*
J. Andreeva, W. Jank, N. Sinanis, *CERN, Switzerland*
N. Colino, P. Garcia-Abia, J.M. Hernandez, F.J. Rodriguez-Calonge, *CIEMAT, Madrid, Spain*
M. Ernst, *DESY, Germany*
A. Anzar, L. Bauerdick, I. Fisk, R. Harris, Y. Wu, *FNAL, Batavia, USA*
G. Quast, K. Rabbertz, J. Rehn, *Karlsruhe University, Germany*
N. De Filippis, G. Donvito, G. Maggi, *INFN-Bari, Italy*
P. Capiluppi, A. Fanfani, C. Grandi, *INFN-Bologna, Italy*
D. Bonacorsi, A. Chierici, L. Dell'Agnello, G. Lo Re, B.Martelli, P. Ricci,
F. Rosso, F. Ruggieri, *INFN-CNAF, Italy*
M. Biasotto, S. Fantinel, *INFN-Legnaro, Italy*
M. Corvo, F. Fanzago, M. Mazzucato, *INFN-Padova, Italy*
C. Charlot, P. Minè, I. Semeniouk, *LLR-Ecole Polytechnique, CNRS&IN2P3, France*
L. Tuura, *Northeastern University, Boston, USA*
M. Delfino, F. Martinez, G. Merino, A. Pacheco, M. Rodriguez, *PIC, Barcelona, Spain*
D. Stickland, T. Wildish, *Princeton University, USA*
D. Newbold, C. Shepherd-Themistocleous, *RAL, United Kingdom*
A. Nowack, *RWTH Aachen, Germany*

Abstract

The CMS 2004 Data Challenge (DC04) was devised to test several key aspects of the CMS Computing Model in three ways: by trying to sustain a 25 Hz reconstruction rate at the Tier-0; by distributing the reconstructed data to six Tier-1 Regional Centres (CNAF in Italy, FNAL in US, GridKA in Germany, IN2P3 in France, PIC in Spain, RAL in UK) and handling catalogue issues; by granting data accessibility at remote centres for analysis. Simulated events, up to the digitization step, were produced prior to the DC as input for the reconstruction in the Pre-Challenge Production (PCP04).

In this paper, the model of the Tier-0 implementation used in DC04 is described, as well as the experience gained in using the newly developed data distribution management layer, which allowed CMS to successfully direct the distribution of data from Tier-0 to Tier-1 sites by loosely integrating a number of available Grid components. While developing and testing this system, CMS explored the overall functionality and limits of each component, in any of the different implementations that were deployed within DC04.

The role of Tier-1's is presented and discussed, from the import of reconstructed data from Tier-0, to the archiving on to the local Mass Storage System (MSS) and the data distribution management to Tier-2's for analysis. Participating Tier-1's differed in available resources, set-up and configuration. A critical evaluation of the results and performances achieved adopting different strategies in the organization and management of each Tier-1 centre to support CMS DC04 is presented.

Presented at *Computing in High Energy and Nuclear Physics, CHEP 2004,*
Interlaken, Switzerland, 27 Sep - 1 Oct, 2004

1 Introduction

The Compact Muon Solenoid (CMS) experiment is one of the four high-energy physics experiments that will be collecting p-p collisions data at the CERN Large Hadron Collider (LHC).

The large amount of data, the scale of the required resources, the software complexity and the geographically distributed nature of the CMS Collaboration naturally imply a distributed computing model and a solution for data distribution and access. The preparation of the Computing System to be able to deal with the data being collected includes a series of planned computing challenges of increasing complexity.

The CMS Data Challenge during March-April 2004 was planned to reach a complexity scale that corresponds to 5% of the LHC rate at full luminosity, i.e. 25% of that foreseen at LHC start-up. Its purpose was to run CMS data reconstruction at CERN (Tier-0) for a sustained period at 25 Hz input rate, distribute the data to the CMS Tier-1 Regional Centres (RC) and analyse them at remote sites (both Tier-1's and Tier-2's).

Prior to the DC04, a Pre-Challenge Production (PCP) phase allowed the simulation and the digitization of about 70 millions of events corresponding to different physics channels, needed for DC04.

2 The role of Regional Centres

The main objectives of the CMS DC04 were:

- data reconstruction sustained at 25 Hz at the Tier-0;
- data distribution to Tier-1/2 sites;
- data analysis at remote Tier-1/2 sites as data arrive;
- monitor and archive both resources and process information;

with a general aim to demonstrate the feasibility of the whole chain. The global DC04 layout is depicted in Fig.1. The specific role of each regional centre is outlined in the following.

2.1 Tier-0 reconstruction and data distribution

Digitized data from the PCP phase were stored on Castor [1] MSS at the Tier-0. A fake on-line process made the data available on an Input Buffer (IB) as input for the reconstruction jobs, that ran at the Tier-0 on a CERN computer farm.

During DC04, to sustain the 25 Hz target reconstruction rate, ~ 2200 jobs/day ran on ~500 CPU's, 40 MB/s of data were staged from the Castor MSS, 4 MB/s of data were produced, 0.4 files/s were registered to the CERN Replica Location Service (RLS) [2] with POOL [3] metadata. The output files (Data Summary Tapes, DST) were stored on a General Distribution Buffer (GDB), acting also as a Castor buffer area, for data distribution to the Tier-1's involved in the challenge. A DC04 dedicated Castor stager was set-up and maintained at the Tier-0, with two pools, one for the IB (10 TB) and another one for the GDB (4 TB). Some limitations concerning the use of Castor at CERN (too many files in the stager database, hardware problems with tapes) were found during DC04 operations.

Central database services, crucial for the challenge purposes, were set-up and maintained at the Tier-0. Redundant monitoring services were deployed on DC04 resources: MonALISA [4] for global monitoring of network and all CPU resources, LEMON [5] as a dedicated fabric monitoring tool on DC04 Tier-0 resources, and GridICE [6] to monitor LCG-2 resources.

CMS developed a file transfer management structure that allowed scheduling data transfers by implementing a multi-agent system design. A limited number of software agents, each dealing with a well-defined task, were deployed at several geographically distributed sites (Tier-0 and the Tier-1's). The inter-communication among agents and the propagation of information through the overall data distribution system was dealt with by allowing the agents to retrieve from (and post information to) a central Transfer Management Database (TMDB) [7].

A set of data distribution agents were deployed at the Tier-0 to steer the overall data and workflow management at the Tier-0 level, their tasks ranging from the discovery of new files available on the GDB to the registration of DST files and metadata to the POOL RLS catalogue, to the streaming of specific data for transfer to one (or more) Tier-1's via machines dedicated each to a particular transfer mechanism.

The data distribution to Tier-1 centres supported three data transfer strategies, corresponding to distinct flavours of the middleware used to address data transfer issues, namely the LCG-2 Replica Manager tools of the LHC Computing Grid (LCG) [8] middleware, the native Storage Resource Manager (SRM) [9] with dCache [10], and the Storage Resource Broker (SRB) system [11] According to local choices, different regional centres adopted one of the aforementioned approaches (CNAF and PIC Tier-1's used the LCG Replica Manager (RM) interface, FNAL exploited the SRM dCache chain and RAL, GridKA and IN2P3 used the SRB system; see Fig. 2). For each data distribution chain, the Tier-0 agents filled a distinct Export Buffer (EB) system with data to be transferred to the Tier-1's participating to that distribution chain. The disk-servers used as EBs were provided by and maintained at the Tier-0, namely the EB-SE (3 servers, 3.1 TB) for the LCG-2 chain, the EB-SRM (4 servers, 4.2 TB) for the SRM chain and the EB-SRB (4 servers, 4.2 TB) for the SRB chain.

2.2 The LCG-2 distribution chain

CNAF and PIC Tier-1's were installed as LCG-2 sites and exploited the LCG-2 Replica Manager interface for data movement. The LCG components used in DC04 are described in [12]. The LRC (Local Replica Catalogue) component of RLS provided the replica catalogue functionality. Castor Storage Elements (SE) providing an MSS interface were deployed at the Tier-1 sites, and "classic" (disk-based) SEs were deployed both at CERN (the machines acting as EBs) and at the Tier-1/2 sites (serving data for analysis). A set of software agents was deployed (in C/C++, Perl, bash) at both Tier-1's to perform basic data transfer, replication and management operations. A *transfer agent* was used for T0(SE-EB) → T1(Castor) data movement, exploiting native Castor tape migration. A separate *replication agent* dealt with T1(Castor) → T1(disk-SE)/T2(disk-SE) data movement to grant data availability for DC04 real-time analysis (see later). An *MSS agent* was also deployed to independently and more efficiently monitor the data migration to tape and the posting of the "safe" state onto the TMDB blackboard.

During DC04 operations, due to a significant overhead introduced by the use of the Java API in the LCG-2 Replica Manager command line tools, different solutions were sought. CNAF limited the use of the RM command-line interface to transfer and register operations, and moved to the use of LRC C++ API to query the RLS for filenames to start transfers. PIC instead used `globus-url-copy` to transfer files and the LRC C++ API to asynchronously register replicas to the RLS. The PIC approach resulted in faster transfers since any overhead by the java processes is removed, but on the other hand the CNAF approach allowed to transfer and inherently register files into the RLS in a unique operation with a file-size check included, thus offering more warranty against failed replications. Nevertheless, throughout DC04 both CNAF and PIC Tier-1 agents were able to sustain the rate of file generation and distribution from the Tier-0 to the Tier-1's (see Fig. 3a). Transfer rates to CNAF and PIC reached a sustained >30 MB/s during a large file-size transfer test undertaken at the end of DC04; CNAF sustained ~42 MB/s for ~5 hours (see Fig. 3b).

The large number of files with sizes in the unexpected small range of 500B-50kB raised severe Castor stager scalability issues and problems with the operation of the underlying MSS system at CNAF (too many entries in the stager database, too high number of segments on tape, bad tape read/write performances and repositioning failures, LTO-2 SCSI errors, inefficient tape space utilization). A new stager was rapidly made available during DC04 operations, and software agents had to be modified accordingly, and this allowed the challenge to continue.

PIC and CNAF were also able to distribute data on to Tier-2's real-time analysis (see below).

2.3 The SRM distribution chain

The FNAL Tier-1 centre deployed an SRM distribution chain for the DC04. It comprised an SRM Export Buffer at the Tier-0 providing access to a local dCache disk pool, and an SRM Import Buffer at the Tier-1 providing access to Enstore [13], again via dCache. Files to be exported were staged out of the Castor system to the dCache disk pool, and pinned until transferred. A transfer agent was used to copy files from the Tier-0 EB to the Tier-1 Import Buffer by initiating a third party SRM transaction to receive a TURL (Transfer URL) from the EB, then using GridFTP [14] to make the actual transfer.

At the start of DC04, some problems arose concerning the authentications, both the high number of authentication requests and the significant overhead introduced by the authentication process itself. The development of agents capable of dealing with multiple streams, thus allowing the transfer of multiple files in each stream, addressed the first item by reducing the number of authentications required. Additional performance improvements were achieved by the optimization of the behaviour of the Globus security layer for specific DC purposes. In common with other distribution chains, the FNAL Tier-1 operations encountered difficulties in dealing with the large number of small files. The resulting inefficient use of tapes forced the Tier-1 operators to increase the number of tapes available and to deploy a larger namespace service. A problem encountered only at FNAL Tier-1 was the difficulty to install monitoring technology, resulting in the fact that any hardware failures had to be identified by a human operator, so actions and restarts were handled manually.

2.4 The SRB distribution chain

The RAL, GridKa and IN2P3 T1 deployed an SRB distribution chain for the DC04. It comprised a shared EB at the Tier-0 and Import Buffers (and underlying different MSS solutions) at each Tier-1. Files to be transferred were copied from Castor onto the EB, where they were inserted (via `Sput` command) into the SRB space, the files' GUIDs being added later as additional SRB metadata (via `SmodD` command). The data replication to the Tier-1's was then performed using either `Sreplicate` or `Sget/Sput`.

Prior to the DC04, the SRB system was successfully used by CMS Production for data-management throughout the pre-challenge production (PCP) phase, and was thought to be a valid component of the DC04 system. This was true especially for those Tier-1's where the LCG-2 middleware was not yet deployed on resources. In addition, for RAL and IN2P3 the SRB represented the only mechanism available to automatically and transparently place files in local MSS with consistent catalogue information, using the GMCat [15] application developed at RAL, which linked the name spaces of SRB and the RLS by publishing SRB replica information into the LRC at CERN periodically. During DC04, the SRB showed an unexpected poor performance, and the overall operations on the SRB distribution chain in the DC were severely hampered by technical issues. A first issue was related to the unavailability of the MCat metadata catalogue (hosted at RAL) in SRB version 2, causing serious operating problems on 22 of 56 challenge days. Loss of performances at different levels were observed (lengthy directory query times; long transaction times causing the transfer agents to time out; core dumps; etc.); in addition, a number of bugs in both SRB client/server, and in Oracle Linux implementations, as well as the use of SRB command-line interface (return codes were not reliable enough; killed transfers continued to run in the background, etc.) additionally hampered the overall operation of the SRB chain. Before the end of DC04, the metadata catalogue service was stopped as the system no longer responded in a useful timescale.

In common with other distribution chains, the DC04 operations with SRB encountered difficulties in dealing with the large number of small files. At SRB sites this showed up as a particularly troublesome injection process of the initial entries onto the EB at the Tier-0. The strategy adopted to enter data into the EB assumed files of the order of 1 GB in size, and the `Sput` command was chosen, since the PCP experience showed it to be particularly efficient with large files. But dealing with small files in DC04 caused considerable and unexpected inefficiencies.

On the other hand, the transfer speeds from the Tier-0 to the Tier-1 were respectable. As an example, transfers to IN2P3 reached 80 Mbps sustained over a period of hours, although with a typical average of only 30 Mbps over a whole day. These transfer rates are due to the small file size, and pre-DC04 tests indicated that transfers could be sustained at a rate greater than this. Due to major SRB problems, the Tier-1's of the SRB chain could not take part in the large files size transfer tests at the end of DC.

2.5 "Real-time" data analysis at Regional Centres

Reconstructed data delivered out of the Tier-0 to the Tier-1 sites were also replicated and made available for data analysis in quasi "real-time" at Tier-1 sites and also at several selected Tier-2 sites. The examples of the LCG-2 and the SRM distribution chains are quoted below.

In the LCG-2 distribution chain, CNAF and PIC Tier-1's replicated data to Legnaro and CIEMAT Tier-2's respectively, and automatic procedure were developed to advertise the arrival of new data on dedicated disk-SEs and hence automatically trigger the job submission on LCG-2 resources via the Resource Broker. Real-time analysis at PIC measured a median delay of ~20 minutes between files being ready for distribution at the Tier-0 and analysis jobs being submitted at the Tier-1/2 sites. More than 17k analysis jobs were submitted in the last

two weeks of the challenge [16].

In the SRM distribution chain, the FNAL Tier-1 deployed a MySQL [17] POOL catalogue to enable access to the DC04 transferred data in the US, the performance of which was adequate. The publishing of entries from the RLS to the FNAL POOL catalogue performed poorly at the beginning of DC04 due to slow RLS queries, but close collaboration between the agent developers at CERN and FNAL addressed this issue and achieved a factor ~100 increase in time performance. Data access at FNAL T1 was attempted through dCache via a ROOT [18] plug-in, allowing for COBRA [19] based applications to access the data. The software environment was based on access to applications over AFS at CERN, which was proved to be quite stable. Data were transferred to University of Florida (UFL) and Caltech Tier-2 sites and UFL was able to use the same software environment as the Tier-1 to analyze the data.

Good read performances were achieved, but the large number of small files exacerbated the bottleneck of the file opening operations. In addition, the high number of files also made it difficult to find the needed data among the transferred files. This was due to the fact the files were organized by date ranges rather than their physics content and hence the files needed for data analysis could be stored on many different tapes; the result was a high number of tape stages to make complete file sets available to analysts.

3 Conclusions

The full reconstruction-transfer-analysis chain was demonstrated to be feasible, and could run at 25 Hz but for limited amount of time. The main areas for future improvements have been identified. Most of the issues are connected to the non-optimal size of the files transferred. Increasing the ratio <#events>/<#files> would *i)* allow a more efficient use of the bandwidth *ii)* address scalability of MSS systems, and *iii)* avoid that "start-up" dominates command execution times. The different Tier-1 performances are strictly related to the data transfer strategies adopted by each Regional Centre.

Throughout DC04, the LCG-2 distribution chain showed a good overall performance. The main difficulty derived from some specific implementation of underlying MSS solution (e.g. at CNAF but not at PIC) in a topology with a Castor-SE directly receiving files from the data distribution system. These difficulties were due to specific DC04 operational conditions that triggered automatic migration policies on an unexpected high number of small files Alternative solutions with a disk-based Import Buffer instead of a Castor buffer as a front-end to the data distribution system have already been designed at CNAF and deployed successfully in the post-challenge CMS activities.

The use of storage media that present a uniform SRM interface to the outside world emerged as an interesting model for the future. The experience with SRM in DC04, as tested by FNAL Tier-1, may allow the creation of generic simpler transfer agents, with some of the basic operations being handled by the underlying system.

The performance of the SRB chain (at least version 2) was severely hampered by technical issues, and more time and work is needed in post-challenge activities to reach the production quality requirements. The MCat in SRB version 2 was identified as a single point of failure since all user authentications, replica and metadata lookups are undertaken using this single service, and disruption to the MCat service crippled the whole of the SRB chain during the challenge. Most of the problems clearly identified in the challenge are already addressed in SRB version 3.

Real-time analysis at Tier-1/2 sites was demonstrated to be possible, and the time window between the availability of reconstructed data at the Tier-0 and the start of analysis jobs was in general reasonably low.

Acknowledgements

We would like to acknowledge the DC04 Task Force, all Tier-0/1/2 site managers, the CERN IT/DB and IT/FIO divisions for the invaluable help and support, and the LCG Deployment team.

References

[1] Castor (Cern Advanced Storage Manager), http://cern.ch/it-div-ds/HSM/CASTOR
[2] D. Cameron et al., "*Replica Management in the European DataGrid Project*", Journal of Grid Computing 2004, in press
[3] POOL Project, http://lcgapp.cern.ch/project/persist
[4] I.C. Legrand et al., "*MonALISA: a distributed monitoring service architecture*", CHEP'03, La Jolla
[5] LEMON Fabric Monitoring, http://cern.ch/lemon
[6] S. Andreozzi et al., "*GridICE: a monitoring service for the Grid*", 3rd Cracow Grid Workshop, Oct. 2003
[7] T. Barrass et al., "*Software agents in data and workflow management*", CHEP'04, Interlaken
[8] LCG Project, http://lcg.web.cern.ch/lcg
[9] SRM Project, http://sdm.lbl.gov/srm-wg
[10] dCache Project, http://www.dcache.org
[11] A. Rajasekar et al., "*SRB, managing distributed data in a Grid*", Computer Society of India Journal, special issue on SAN, 33(4):42-54, 2003
[12] A. Fanfani et al., "*Distributed Computing Grid Experiences in CMS DC04*", CHEP'04, Interlaken
[13] The Enstore Mass Storage System at Fermilab, http://computing.fnal.gov/docs/products/enstore
[14] The Globus GridFTP protocol and software, http://www.globus.org/datagrid/gridftp.html
[15] T. Barrass et al., "*Integrating the Storage Resource Broker with the GIGGLE Framework*", NIM A, in press: http://authors.elsevier.com/sd/article/S0168900204014809
[16] N. De Filippis et al., "*Tier-1 and Tier-2 real-time analysis experience in CMS DC04*", CHEP'04, Interlaken
[17] MySQL, http://www.mysql.com
[18] The ROOT Analysis Framework, http://root.cern.ch
[19] V. Innocente, "*COBRA, Coherent Object-oriented Base for Reconstruction, Analysis and simulation*", http://cobra.web.cern.ch/cobra

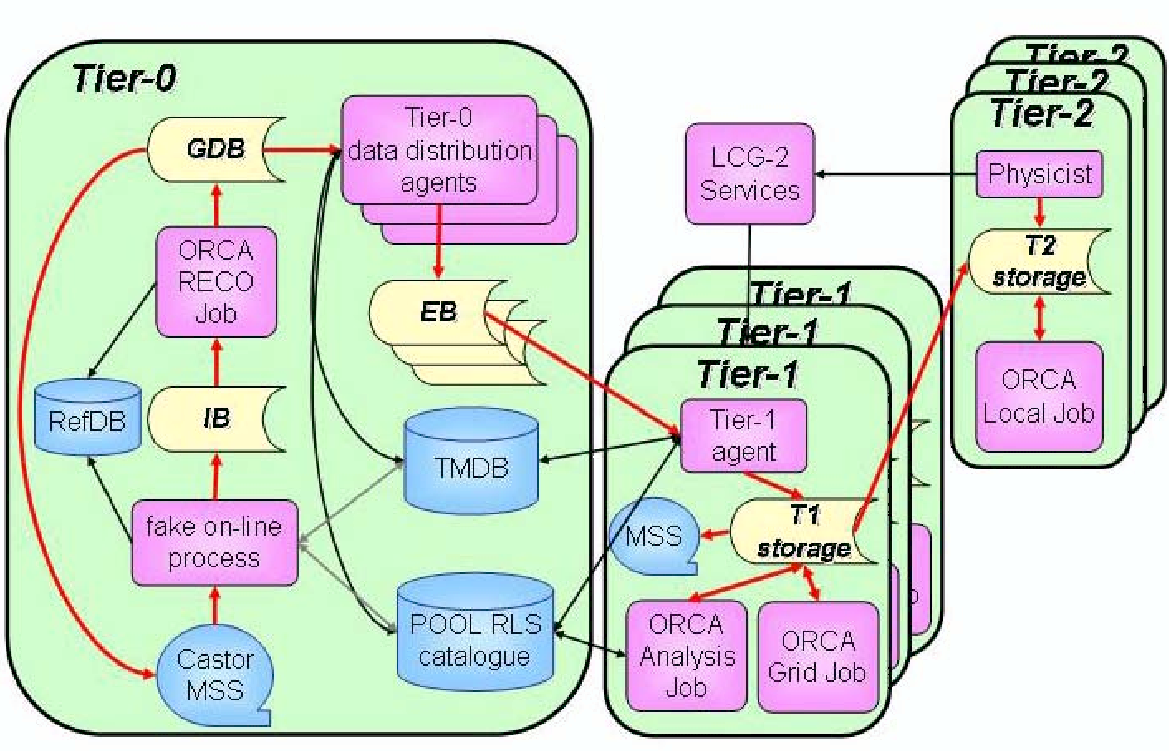

Fig.1: Global DC04 layout. Red lines depict data-flow, black lines show control-flow.

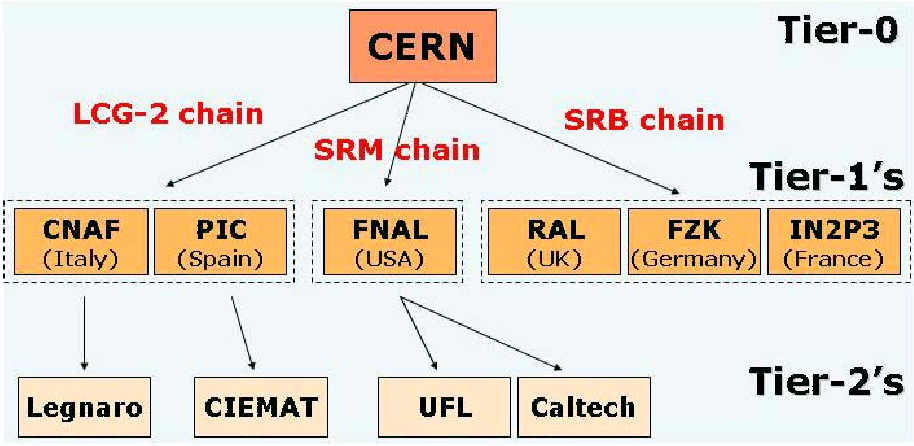

Fig.2: Data transfer topology and distribution chains in CMS DC04. All involved Regional Centres are depicted.

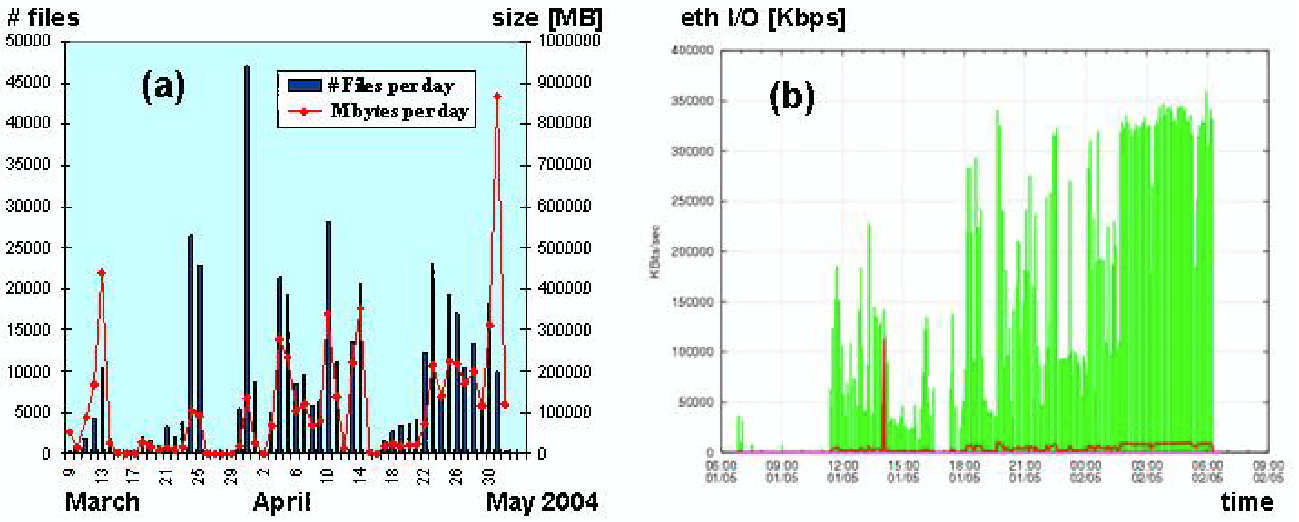

Fig. 3: (a) Number and size of files transferred per day at PIC Tier-1 during DC04. (b) Transfer rates measured at CNAF Tier-1 during a short network stress test with big files transferred at the end of DC04. The plateau at ~330 Mpbs sustained for several hours is visible on the right.