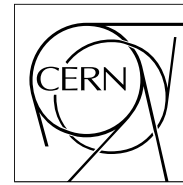**The Compact Muon Solenoid Experiment**

CMS Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

7 March 2007

CMS Computing, Software and Analysis Challenge in 2006 (CSA06) Summary

The CMS Collaboration

Abstract

This is the summary document on the CMS Computing, Software and Analysis Challenge of 2006. An overview of the software tools and the computing architecture deployed for the challenge is given, followed by summaries of the operational experience at the Tier-0, Tier-1, and Tier-2 centres. A summary of the analyses based on the produced data samples is given followed by detailed reports on specific analyses. A list of various lessons learned during the challenge completes the document.

Contents

2

1 Definition

The combined Computing, Software, and Analysis challenge of 2006 is an O(50) million event exercise to test the workflow and dataflow associated with the data handling model of CMS [1, 2]. It is designed to be a 25% capacity test of what is needed for operations in 2008. The main components include:

- Preparation of large simulated datasets (some with High Level Trigger tags)

- Prompt reconstruction at Tier-0, including:

    - Reconstruction at 40 Hz using the new CMSSW software framework
    - Application of calibration constants from the offline database (DB)
    - Generation of FEVT[1], AOD[2], and Alignment/Calibration skim datasets known as "AlCaReco" skims
    - Splitting of an HLT-tagged sample into O(10) streams

- Distribution of all AOD and some FEVT to all participating Tier-1s (as well as to some Tier-2s)

- Calibration jobs on Alignment/Calibration datasets at Tier-1, Tier-2 and CAF

- Re-reconstruction performed at a Tier-1

- Skim jobs at some Tier-1s with data propagated to Tier-2s

- Physics jobs at Tier-2s and Tier-1s on AOD and Reco

While this is an exercise to test the data handling workflow under as realistic conditions as possible, it is not explicitly required that the software components are fully validated for physics performance at the time of the challenge. However, where possible we tried to maintain the maximum utility of the simulated, reconstructed, and selected samples for the analysis component of the exercise. The CMS computing model and is described in [1] and [2].

2 Success Metric

Success of the CSA06 challenge was pre-defined (June 2006) as meeting a series of binary metrics (succeed/fail) as well as a list of quantitative numbers based on the performances anticipated a half-year before the CSA06 challenge (and Service Challenge 4) began. The quantitative metrics are placed at two levels: a minimum threshold below which we consider a definite failure, and a goal which is considered achievable if everything runs well. No specific goals were placed on the number or results from the calibration, alignment, and analysis exercises other than to meet the overall daily job submission goal and to demonstrate the workflow associated with prompt calibration.

The metrics were chosen to exercise a variety of important elements in the CMS computing model, though to enable the available effort to concentrate on particular functionality not all areas were tested. The metrics were chosen to ensure a broad participation of CMS computing facilities, to enable the experiment to demonstrate functionality critical to early experiment operations, and to encourage physics analysis.

2.1 Binary metrics

- Automatic FEVT+AOD transfer Tier-0 to Tier-1 via PhEDEx, the data placement tool

- Automatic transfer of part of FEVT+AOD Tier-1 to Tier-2 via PhEDEx

- Offline DB accessible via FroNtier/Squid (a caching layer between the reconstruction jobs and the Oracle DB) at participating sites

- Insertion and use new constants in Offline DB

- User submission of analysis/calibration/skim jobs via the grid job submission tool CRAB and using the developed Dataset Bookkeeping Service (DBS) and Data Location Service (DLS)

---

[1] "Full Event" data, i.e. RAW+RECO which for this challenge was combined into one file stream

[2] "Analysis Object Data" (AOD) is a reduced collection of reconstruction products (i.e. DST)

- Skim job output automatically moved to Tier-2 via PhEDEx

- Running re-reconstruction-like jobs at Tier-1 that access updated information from the offline DB and perform a new reconstruction on data distributed fromt he Tier-0 centre

2.2 Quantitative metrics

- Number of participating Tier-1 – Goal: 7 – Threshold: 5

  - Passing requires 90% uptime, or < 3 days downtime during challenge

- Number of participating Tier-2 – Goal: 20 – Threshold 15

- Weeks of running at sustained rate – Goal: 4 – Threshold: 2

  - This will be the period over which we measure the other metrics

- Tier-0 Efficiency – Goal: 80% – Threshold: 30%

  - Measured as unattended uptime fraction over 2 best weeks of the running period

- Running grid jobs (Tier-1 + Tier-2) per day (2h jobs typ.) – Goal: 50K – Threshold: 30K

- Grid job efficiency – Goal: 90% – Threshold: 70%

- Data serving capability at each participating site : Goal 1MB/sec/execution slot – Threshold : 400 MB/sec (Tier-1) or 100 MB/sec (Tier-2)

- Data transfer Tier-0 to Tier-1 to tape – Individual goals (threshold at 50% of goal):

  - ASGC: 10MB/s
  - CNAF: 25 MB/s
  - FNAL: 50 MB/s
  - GridKa: 20MB/s
  - IN2P3: 25MB/s
  - PIC: 10 MB/s
  - RAL: 10MB/s

- Data transfer Tier-1 to Tier-2 – Goal: 20MB/s into each Tier-2 – Threshold: 5MB/s

  - Overall "success" is to have 50% of the participants at or above goal and 90% above the threshold
  - Several Tier-2s have better connectivity and we will have higher targets for those
  - Goal for each Tier-2 is to demonstrate 50% utilization of the WAN to the best connected Tier-1
    * list was defined after SC4

3 Executive Summary

The CSA06 challenge was officially launched on 2 October 2006, with major operations lasting approximately 6 weeks. The challenge began with the reconstruction of minimum bias events at the Tier-0 center at CERN, but was preceded by the development of approximately 0.5M lines of software in the new CMSSW framework to conduct the challenge (i.e. simulation, reconstruction, data formats, and filtering) and by the pre-challenge production of over 60M simulated events during July-September. During the pre-challenge production, computing operations were able to produce approximately 5M minbias events per day. When signal samples were enabled, approximately 30M events per month were delivered. The produced datasets did not include pile-up for this challenge.

The first week of prompt reconstruction at Tier-0 processed primarily minbias events and, after a few days, events from an "electroweak soup" at a total rate of about 50 Hz. Given the small size of minbias events, it was difficult to match the goal of exporting FEVT data to all Tier-1s at a rate of 150 MB/s. By the start of the second week, however, a program was instituted of reconstructing and subscribing signal datasets at a rate that maintained the desired overall bandwidth and the individual Tier-1 rate goals. The reconstructed samples were subscribed to

particular Tier-1 centres based loosely on the needs of the analysis exercises, with minimum bias and electroweak events going to every centre. The average Tier-0 export rate met the target rate of 150 MB/s during the first 3 weeks of the challenge.

Due to time constraints, CSA06 was the first opportunity to commission on a large scale the access to calibration constants via a Frontier caching layer to the offline database using the CMSSW reconstruction code. Initial tests at Tier-0 did not succeed for several reasons: segmentation faults occurred, and the detailed logging of a large number of database queries per job eventually crashed the servers. This was diagnosed and patched in a subsequent software release, and access logging was limited, so that by the last week of the challenge access to calibration constants was a normal part of reconstruction.

In the fourth week of CSA06, the Tier-0 prompt reconstruction was significantly ramped up to a rate of over 100 Hz (peaking at 300 Hz for 10 hours), processing mainly signal samples and exercising simultaneously the full complexity of the CSA06 challenge: FEVT, AOD, and AlCaReco production, with calibration and alignment constants accessed via Frontier. The Tier-0 achieved 96% CPU utilization of 1400 CPUs. The export bandwidth to Tier-1 increased substantially to about 260 MB/s average in the fourth week. Occasional glitches, site outages and downtime were encountered during the challenge, but the prompt attention by site managers and the network performance led to rather quick recovery, with bursts sometimes exceeding 650 MB/s achieved. Six of 7 Tier-1 centres achieved better than 90% availability during the challenge. Of 27 validated Tier-2 sites, 21 conducted transfers for more than 24 days. Twelve sites met the transfer goal of 20 MB/s for at least 24 hours, while 20 met the threshold for success of 5 MB/s.

Along with the reception of data from Tier-0, two important workflows at Tier-1 centres were the organized production of secondary skim datasets and a demonstration of re-reconstruction. Approximately 25 skim filters were run, generating nearly 60 secondary datasets for the challenge. This was organized by 4 skim teams for the 7 Tier-1 centres, and secondary datasets were published and accessible just like primary datasets. Re-reconstruction of distributed datasets in FEVT format was demonstrated at 6 Tier-1 centres, where two reconstruction modules (out of nearly 100) had to be disabled due to a technical problem

The fifth and sixth week of the challenge focused on increasing the job submission rate to Tier-1 and Tier-2 centres, and evaluating the efficiency and latency of data transfers to Tier-2. For 6 days the grid job submissions exceeded the goal of 50K per day (30K from robot job submissions which have the same mechanics of user jobs). The job completion efficiency was 90%.

When the CSA06 exercise completed, the Tier-0 centre had maintained 100% uptime over 4 weeks (attended, but without the need for intervention to keep it running), and the reconstruction software ran stably on over 200M events. Overall, more than 1 petabyte of data was shipped across the network between Tier-0, Tier-1, and Tier-2 in 6 weeks.

A wide variety of analysis activities were performed during the challenge, which totaled approximately 30 and involved nearly 70 physicists. In addition to testing various analysis workflows, the demonstrations proved to be useful training exercises for collaborators in the new software and computing tools. Several calibration and alignment demonstrations were performed for ECAL (intercalibration from isolated electrons and phi symmetry), HCAL (intercalibration from isolated tracks and phi symmetry), tracker (alignment from $Z \rightarrow \mu\mu$ events), and muon (alignment from $Z \rightarrow \mu\mu$ events). Specialized reduced calibration data formats were developed for the challenge (the so-called "AlCaReco" formats), and two exercises demonstrated a complete calibration workflow by processing these datasets to obtain new calibration constants from initially miscalibrated data, inserting new constants into the offline database, and validating the correction after reprocessing at a Tier-1 centre. Many general physics studies also were conducted on the reconstructed samples using the CRAB job submission tool, and are reported herein.

The rest of this document covers the details of the CSA06 exercise and concludes with a list of specific lessons or recommendations to further improve upon the data-handling model of CMS.

4   Organization

Strategic planning for the CSA06 exercise was the responsibility of the CPT project manager Paris Sphicas, and overall decisions regarding the scope and implementation of the challenge were discussed in the weekly CPT management meeting between the CPT project manager and the coordinators of the Software, Computing, and PRS projects.

There was an overall coordinator of the CSA06 challenge (Darin Acosta) for general planning, facilitation of interactions between the Computing, Software, and Analysis teams, and the indentification of problems (and decisions on the appropriate course of action) as they arose in the planning and operations. General planning included the specification of datasets, order of processing, and the destination of distributed datasets. A general CSA weekly meeting was held Wednesdays at 15:00 (see http://indico.cern.ch/categoryDisplay.py?categId=4l55), and a Twiki site posting relevant information and news about the challenge was maintained (see https://uimon.cern.ch/twiki/bin/view/CMS/CSA06).

Two coordinators for the integration program (Ian Fisk and Michael Ernst) defined the scope and approaches of the computing tasks, led the preparations for activities, and handled the daily coordination. Daily coordination meetings were held at 15:30, where operational issues were discussed and operational decisions taken. Daily reports were posted to Twiki pages. For example, the last report is here: https://uimon.cern.ch/twiki/bin/view/CMS/ReportCSA06Day39.

4.1   Interactions with WLCG and CERN-IT

According to the established model and the associated division of responsibilities the majority of the generic, non-CMS specific processing, storage and network infrastructure is provided by or through the Worldwide LHC Computing Grid Project (WLCG) and for the CMS Tier-0 center by CERN-IT.

As some of the components, though released to the field, have never been utilized at the scale that was foreseen for CSA06, a plan was worked out in early summer 2006 that foresaw thorough evaluation of specific components and services to ensure their readiness in terms of functionality, performance and robustness for CSA06.

The CMS-WLCG Integration Task Force (chaired by M. Ernst) was chosen for the interaction between the middleware service providers and the CMS computing organization.

Guided by the functionality requirements and associated performance metrics as they were defined for CSA06 the Task Force was focusing on three areas

- Data transfer infrastructure
  The Task Force worked specifically on the architecture and implementation of the FTS (File Transfer Service) channel infrastructure following the requirements as they are stated in the CMS Computing Model [1]. Several suggestions made by the Task Force were accepted by the WLCG/EGEE developers and are already implemented or became part of the medium term development and integration plan.

- Grid job submission system
  A bottleneck that became obvious during Service Challenge 4 (SC4) with CMS aiming at submission and execution of 25 000 jobs/day to/at participating Tier-1 and Tier-2 centers was the performance limitation at the level of the LCG Resource Brokers (RB). Doubling the rate as was foreseen in CSA06 would have required an unmanageable number of physical RB instances. According to thorough analysis studies performed by Andrea Sciaba for CMS and Simone Campana for ATLAS the primary reason for the performance problems was due to the time required to submit individual jobs. Since the gLite Workload Management System, the successor of the LCG RB, that supports bulk submission and therefore is addressing this point was already part of the deployed LCG middleware stack, the Task Force decided to further evaluate this component. WLCG made one instance available to CMS in July exclusively and added two more close to the start of CSA06. Multiple performance and stability issues were observed when testing the gLite RB under heavy load. Given good interaction due to all relevant groups, including the EGEE JRA1 lead person, being represented in the Task Force all known problems were fixed by the EGEE developers on a timely basis and well ahead of the start of CSA06. CMS and in particular the Task Force has helped significantly to improve the maturity of this vital component, such that it can be used in production.

- Data storage and data access at CERN
  Data storage and data access at CERN, in particular in the context of data transfers from the CERN storage facilities to the Tier-1 centers, encountered lots of problems since CMS began to move large amounts of data in 2004 as part of the DC04 program. An issue that came up during the CSA06 preparation phase was that activities originated by different application areas could massively interfere with each other. The particular problem observed was that the serving of data from the Magnet Test/Cosmic Challenge (MTCC) to centers

outside CERN, required to be performed at low latency, was largely impacted by the PhEDEx-driven high-performance transfer tests between CERN and the CMS Tier-1 centers. After thorough analysis, the problem was traced down to a protocol mismatch between the Castor client and the PhEDEx stager agent, which led the stager agent to generate a massive amount of staging requests for files that were already available on disk—all of them being merged in the same Castor queue with transfer requests from all application areas. As stability and performance issues with data access services at CERN were considered a severe problem for CMS analysis and production activities, a dedicated working group, called the "Castor/CMS Collaboration" with participation from the Castor Team and CMS Computing was formed. It was agreed that data storage and data access services provided by Castor at CERN are to be considered a component that is deeply integrated with the CMS workflow and that a traditional client/service provider relationship is not appropriate. The working group is chaired by Frederic Hemmer, the CERN-IT Deputy Division Leader and CMS/IT liaison, and it reports to the CMS/WLCG Integration Task Force.

5 Offline Software

5.1 Sequence of Releases

The following releases of CMSSW software were employed for the CSA06 challenge and pre-challenge activities:

- CMSSW_0_8_x: available July 2006, validated for large-scale simulation
  - CMSSW_0_8_1: minimum bias events
  - CMSSW_0_8_2 and 0_8_3: for signal samples (as generator filters became available)
- CMSSW_1_0_x: available September 2006, validated for large-scale reconstruction, consisting of over 0.5M lines of code
  - CMSSW_1_0_2: minimum-bias reconstruction
  - CMSSW_1_0_3: fixes to improve robustness of signal samples
  - CMSSW_1_0_4: Alignment/Calibration skim dataset production
  - CMSSW_1_0_6: Frontier access ready, analysis skims ready

5.2 Generation

For the CSA06 exercise, 50 million events were requested to be generated, simulated and reconstructed; they consist of the 9 samples outlined below. For all samples, the PYTHIA generator interface was used (version 6.227), with the CTEQ5.1 Parton Density Functions. For the description of the underlying Event, Tune DWT (Rick Field) was used. Some of the samples were preselected using generator-level information. For these samples, EDFilters were invoked right after the generation step. The samples were:

1. *Minimum Bias*: 25 million events (produced with CMSSW_0_8_1). All non-elastic processes (including diffractive and double-diffractive) switched on.

2. $t\bar{t}$: 5 million events (produced with CMSSW_0_8_2). All decay channels open.

3. $Z \rightarrow \mu\mu$: 2 million events (produced with CMSSW_0_8_2)

4. $W \rightarrow e\nu$: 4 million events (produced with CMSSW_0_8_3) selected in an η, ϕ range as to illuminate 2 supermodules.

5. *Soft Muon Soup*: 2 million events (produced with CMSSW_0_8_3), of which 1 million of inclusive muon events filtered from Min Bias and 1 million J/ψ events with $P_T(\mu) > 4$ GeV.

6. *Electroweak soup*: 5 million events (produced with CMSSW_0_8_3) consisting of 2.6 million $W \rightarrow l\nu$, 2.2 million Drell-Yan (mass > 15 GeV), 0.1 million $H \rightarrow WW$ events and 0.1 million WW events. For the last two subsamples, the cross sections were artificially reweighted in order to produce the desired event mix. All 3 charged lepton generations are included.

7. *Jet Calibration Soup*: 1.2 million events (produced with CMSSW_0_8_3) consisting of dijet and Z+jet events, in various $\hat{p}_t$ bins reweighted to give the event numbers desired by the Jet-MET group.

8. *Exotics Soup*: 1 million events (produced with CMSSW_0_8_3) consisting of 0.22 million excited quarks (400 GeV) events (all decays), 0.39 million Z' (700 GeV) events (all decays) and 0.39 million SUSY LM1 events (all decays).

9. *HLT Soup*: 5 million events (produced with CMSSW_0_8_4) consisting of W (forced to charged leptons), Drell-Yan (mass > 20 GeV, forced to charged leptons), $t\bar{t}$ events (all decays) and QCD dijets ($\hat{p}_t > 350$ GeV)

5.3 Simulation and Geometry

The first step of the CSA06 challenge consisted of the preparation of large simulated datasets, some of which included High Level Trigger (HLT) tags.

The produced samples of over 60M events were obtained using the CMSSW_0_8_x series of releases. The physics generator input, based on Pythia, is described in the previous section along with a description of the generator datasets.

Simulation in the CMSSW_0_8_x series is based on the 7.1 version of the Geant4 simulation toolkit. The full simulation chain consists of the detailed description of the CMS detector setup in the 4 Tesla magnetic field, particle propagation and physics process modeling, hit collection in the sensitive detector elements and signal digitization taking into account all relevant effects. Pile-up, however, was not available at this stage of integration. In addition to the simulation chain, the simulation applications available with the CMSSW_0_8_x release are the GeometryProducer for visualization debugging, and a set of hit and digi level packages which are part of the Software Validation Suite (SVSuite).

The simulation chain, i.e. geometry, simulation and digitization, was extensively validated in terms of description correctness, detector response, physics quality and software robustness and performance, using the SVSuite.

5.3.1 Infrastructure

Simulation infrastructure in CMSSW_0_8_x included the following elements:

- User hooks or observers which allowed access detector simulation information at the beginning/end of a track, run, or event, and at every step in Geant4 tracking.

- Interface to a realistic magnetic field.

- QGSP and LHEP physics lists in Geant4, as well as interface to set cuts per sub-detector on the production of secondary particles by Geant4.

- Use of random number service provided by the Framework project.

- An exception catcher tool to skip events in cases of a divide by zero, invalid, overflow, underflow.

Some of the infrastructure features which were missing for CSA06 and will be added in the future are:

- Validation of a G4.8.1 based version of CMSSSW and subsequent migration.

- Commissioning of the geometry overlap detection tool.

- Local magnetic field management

- Optimization of simulation parameters for speed and accuracy

- Development of capability to overlap real data to Monte Carlo signal events.

- GFlash parameterization of hadronic showers.

5.3.2 Performance and Robustness

The simulation chain, which includes generation, Geant4 based detector simulation, and digitization is very robust from the point of view of crash rate. Crashes are very rare, on the order of 10^{-4}–10^{-6} per event. This robust performance, however, is partially due to the action of the exception catcher, based on the FloatingPointException (FPE) service. Otherwise, there would be a large crash rate due to Geant4 unsafe features, as well as overlap of CMS detector volumes. The fraction of events skipped by the exception catcher tool is approximately 0.5% for minimum bias, and 2% for QCD events. Preliminary studies using G4.8.1 based CMS simulations show a significantly improved behavior of Geant4. At the same time, the geometry is currently being debugged to remove volume overlaps. Time performance performed on Min-Bias, H(300 GeV)$\rightarrow ee\mu\mu$, and heavy ion events gave average values of 48, 247, and 5976 seconds per event, respectively.

5.3.3 Geometry, Hits, Digi Implementation and Verification

Before the CSA06 exercise, the geometry of all the sub-detector systems, except the electromagnetic calorimeter (ECAL), has been re-written using the Detector Description Database (DDD) xml based tools. It has also been verified against the old implementation based on a machine translation to xml of the CMSIM Geant3 geometry. The new material budget description of all tracker sub-systems except the Forward Pixels (FPix) and the Tracker Outer Barrel (TOB) is in excellent agreement with the old implementation. The FPix description has been improved and is more accurate than the previous one. The TOB differences are not understood and under investigation. The tracker group is performing a validation of the geometry description based on weighing the actual unite modules. The ECAL material budget in the new and old implementations was verified to be identical, as expected. A new post-CSA06 DDD/xml based geometry is now completed for the ECAL barrel and pre-shower detectors and is being validated. The agreement between the old and new Hcal geometry implementations is nearly perfect. The muon geometry was also re-developed and awaits verification. A Geometry SVSuite package was developed to test material budget. Hits and Digis were implemented in CMSSW for all sub-detector systems and verified versus OSCAR/ORCA results using their respective SVSuite packages. Corrections and updates were, however, incorporated as the code is tested from CSA06 and physics validation samples. The forward detectors (Zero Degree Calorimeter (ZDC), Totem, Castor) were also available for CSA06 as standalone CMSSW applications, outside the CMS detector simulation chain.

5.4 Filtering and Streaming

CMSSW included filtering technology to select events at the generator-level, at HLT, and for skimming datasets secondary datasets. The ability to write out multiple output files simultaneously was included. The ability to merge input files into a single output file was also provided.

In order to simplify the definition of filtering modules, we used a generic approach implemented with C++ templates. This could be achieved because all RECO/AOD objects have the same naming convention for member functions returning the same type of information (like p_T, E_T, and so on). Generic filter modules have been written for the most commonly used selection criteria. Events are selected by a filter module if they contain at least a specified number of reconstructed objects passing a specified selection criterion. "Primitive" selection criteria are defined based on a single variable cut and can be combined with Boolean operations to create a richer variety of selections. The generic modules have been instantiated for the objects types of interest (electrons, muons, jets, tracks, etc.) and for the selection criteria of interest for that object type, creating in a release a large "suite" of selector and filter modules ready to be plugged in cascade in any skimming sequence. All cuts for all modules are fully configurable via the parameter set mechanism provided by the Framework.

5.5 Higher Level Trigger

A total of 12 HLT triggers from the Physics TDR Vol.2 [3] trigger menu were implemented in CMSSW for the CSA06. The CSA06 HLT triggers are based on Monte Carlo information using generator-level photons, electrons, muons, taus and jets clustered from generator-level particles. The trigger thresholds, all denoting transverse momentum inside the relevant subdetector acceptance, are listed in the following table.

The above HLT triggers are identified by their mnemonics, or by a non-negative consecutive integer number (serving as an index into a C++ vector). The number is assigned based on the relative order in which the HLT trigger paths appear in the configuration file. Because the workflow management converts configuration files to python files and back, this order was mangled, and the expected bit positions did not correspond to the actual ones. To

Table 1: List of implemented HLT triggers, based on generator-level particles.

Name	Mnemonic	Threshold (GeV)
Single Gamma	p1g	80
Double Gamma	p2g	30,20
Single electron	p1e	26
Double electron	p2e	12,12
Single Muon	p1m	19
Double Muon	p2m	7,7
Single Tau	p1t	100
Double Tau	p2t	60,60
Single Jet	p1j	400
DiJet	p2j	350
TriJet	p3j	195
Quad Jet	p4j	80

alleviate this problem, the keyword *schedule* was introduced, during the CSA exercise, into the configuration language. This statement allows to specify the order of paths. A default schedule statement is implicitly generated from the order encountered in the configuration file in case the user does not specify one explicitly.

In general, access to specific trigger bits and trigger results should rely on the mnemonics rather than integer enumerations. Acess through mnemonics is not affected by re-ordering problems. However, using integer positions has the advantage of using less memory space and is thus used internally; recall that the HLT trigger table is the same for a large number of events.

When the prepared sample of generated events was split into separate datasets, similar triggers were grouped together such that 4 output datasets were created: photons, electrons, muons, and jets (essentially no events passed the generator-level thresholds for the tau triggers).

5.6   Reconstruction

The CSA06 related goals were (in order of importance)

- A release containing reconstruction modules for all high level objects

- Reconstruction code stable, able to run on tens of million of events without a significant crash rate

- Reconstruction code able to process a few thousand events per job, without suffering from memory leaks. On complex event samples, this meant more than 12 hours of running without glitches

- Reconstruction code usable for detector studies as well as first look at physics analysis.

While the first CSA06 oriented version was CMSSW_1_0_0 (end of Sept. 2006), stability/performance tests were run during the summer using integration releases, and led to fixing the majority of technical problems.

CMSSW_1_0_0 contained reconstruction components for

- Local detector reconstruction (clustering, segment building in muon chambers)

- Super Clustering in the ECAL with Brem recovery

- Full tracking using Kalman Filter and Combinatorial Track Finding

- Fast tracking (trigger like) using only the pixel subsystem

- Jet reconstruction (MidPoint and Iterative Cone algorithms)

- Missing E_T reconstruction

- Electron reconstruction seeded with ECAL Super Clusters

- Photon reconstruction

- Standalone and Global Muon reconstruction (Muon + Tracker links)

- Primary vertices with full tracking and pixel only tracking

- B Tagging using track counting

- Tau Tagging using calorimetric and tracker (isolation) variables

The modules were organized in framework sequences, local, global and high level, and were consistently run for all data samples. The code, after the summer tuning and some other late moment fixes, was able to run with a negligible crash rate and without showing memory problems, on all of the CSA06 samples. The two extreme conditions (minimum bias low p_T events and TTbar events) are shown in Table 2.

Table 2: Time and memory footprint during CSA06 in two extreme cases. The memory is measured when the steady state is reached.

	Minbias Events	TTbar Events
Time (sec/ev)	3.5	25
Memory Footprint (MB)	400	700

The main challenge we faced during summer and fall 2006 was to manage the rapid deployment of the reconstruction packages as well as maintain stability and backward-compatibility of data structures and code to enable the CSA challenge to be run. As this was also the first appearance of some of these packages in the new software environment, validation was and is an extremely important and urgent task.

While much of the reconstruction code was quite recent at the start of CSA06 T0 processing, previous and a-posteriori checks have shown that the reconstruction output is useful from physics point of view, and allows speeding up the validation process the developers are now focusing on. That apart, the huge and coordinated CSA06 effort, with its variety of use cases for reconstruction software (plain, with calibrations, reprocessing) has been valuable to uncover some weak points in the overall structure, e.g. configuration file structure being too naive and not flexible enough (which has been changed since). In fact many of the changes which went into the CMSSW_1_2_0 development release do come from the lessons learnt during CSA06. The reconstruction code as in CMSSW_1_0_x proved to be adequate for the scope and the goals of CSA06, and it serves as a solid base to continue the development toward data-taking.

5.7 Calibration & Alignment

In the context of CSA06 several exercises concerning the prompt alignment and calibration work-flow have been carried out and results of these exercises are described in Sections 13.1 and 13.2. The full chain work-flow for ECAL and HCAL calibration as well as for tracker alignment has been tested using several techniques that are foreseen to be used during CMS operation. The calibration and alignment part of the CSA06 challenge consisted of the following steps:

1. Prompt reconstruction at the Tier-0, reading alignment and calibration constants from the offline database;

2. Production of a special skimmed data format, the **AlCaReco**, which contains only the information relevant for a specific alignment or calibration task;

3. Running alignment/calibration algorithms at Tier-1/2 centres or the Tier-0 using misaligned/miscalibrated AlCaReco data;

4. Inserting the derived alignment/calibration objects into the database and deploying them to the Tier-1/2 centres;

5. Re-reconstruction at Tier-1 centres reading these updated constants;

6. Analysis jobs at Tier-2 centres comparing ideal, miscalibrated (misaligned) and calibrated (aligned) distributions.

The CMS software is set up such that misalignment and miscalibration do not necessarily have to be applied already during simulation or reconstruction, but can also be applied on the fly even at the level of a user analysis job. Therefore, in order to keep maximum flexibility, it was decided not to misalign and miscalibrate the data reconstructed at the Tier-0 (and hence the AlCaReco), but rather to use ideal constants for the prompt reconstruction and AlCaReco production, and to apply misalignment/miscalibration on the fly when running the alignment/calibration algorithms.

To achieve this, all calibration exercises made use of a common *miscalibration tool*, available under `CalibCalorimetry/CaloMiscalibTools` which allows to miscalibrate RecHits also at the reading stage based on predefined scenarios. Similarly, the alignment exercises used a dedicated tracker/muon *misalignment tool* (in `Alignment/TrackerAlignment` and `Alignment/MuonAlignment`) which is able to move/rotate all parts of the tracker and the muon detector.

In order to perform the exercise, several software ingredients were put in place:

- Implementation of the actual calibration and alignment algorithms;

- Implementation of the various AlCaReco stream producers;

- Preparation of a matrix defining which streams to create from which dataset, as well as a combined central configuration file;

- Insertion of the alignment and calibration objects into the offline database;

- Configuration fragments for reading database objects containing alignment and calibration constants from the offline database, and using them in the reconstruction.

The alignment and calibration algorithms were implemented for the CMSSW_1_0_0 release.

The following AlCaReco streams were defined and implemented:

1. `CSA06ZMuMu`
   Tracker alignment using $Z^0 \to \mu^+\mu^-$ events. Only the tracks to be used by the alignment algorithm, namely those corresponding to the muons from the Z^0 decay and satisfying $p_T > 10$ GeV are stored in the event, using the `AlignmentTrackSelector`.

2. `CSA06MinBias`
   Pixel tracker alignment using minimum bias events. Only tracks with $p_T > 1.5$ GeV and at least 6 reconstructed hits are considered, and a minimum of two such tracks was requested for the event to be kept. Again, the `AlignmentTrackSelector` is used.

3. `CSA06ZMuMu_muon`
   Muon alignment using $Z^0 \to \mu^+\mu^-$ events.

4. `AlcastreamElectron`
   Cell-wise ECAL calibration using E/p from isolated electrons. In the AlCaReco, only the information relevant to the calibration is retained: the RecHits associated to the selected electron and the track associated to it. The `ElectronSelector` package has been used to select tracks with transverse P_t above a threshold of 20 GeV. An AlCaReco event from $W^+ \to e^+\nu$ has a size of about 3 kBytes.

5. `AlcastreamEcalPhiSym`
   Inter-calibration of ECAL rings using the phi symmetry method. The only information which needs to be stored for each event is the subset of ECAL RecHits with energy above a certain threshold, which is introduced in order to prevent noise from contributing to the energy sums. The value of the threshold is 150 MeV for barrel crystals and 750 MeV for end-cap crystals. Only a few tens of RecHits per event have energy exceeding these thresholds. The AlCaReco data format for the ϕ symmetry calibration exercise is defined to consist solely of the filtered RecHit collections for the ECAL barrel and end-caps. The average size of ϕ-symmetry AlCaReco events is 120 bytes.

6. `AlcastreamHcalDijets`
   HCAL Calibration using dijet balancing.

Table 3: Matrix of produced AlCaReco streams.

| Output Stream Name | Input Dataset | | | | Purpose |
	$Z^0 \rightarrow \mu^+\mu^-$	min. bias	QCD Jets	$W^\pm \rightarrow e^\pm\nu$	
CSA06ZMuMu	X				Tracker Alignment
CSA06MinBias		X			Tracker Alignment
CSA06ZMuMu_muon	X				Muon Alignment
AlcastreamElectron				X	ECAL Calibration
AlcastreamEcalPhiSym		X			ECAL Calibration
AlcastreamHcalDijets			X		HCAL Calibration
AlcastreamHcalIsotrk		X	X		HCAL Calibration
AlcastreamHcalMinbias		X			HCAL Calibration

7. `AlcastreamHcalIsotrk`
   HCAL calibration using E/p from isolated pion tracks. The procedure is done independently for pions not interacting with ECAL and interacting with ECAL and is supposed to provide calibration constants as a function of energy on a tower by tower basis (HB and HE up to $|\eta| < 2.1$). The AlCaReco producer makes a selection of reconstructed tracks by requiring spatial isolation from other tracks in the event. An isolated track was defined by: (a) no other charged particles within a cone of 0.5; (b) a conservative cut on $p > 1$ GeV to reject tracks that will not reach HCAL. In addition the original CaloTower collection is kept in the AlCaReco.

8. `AlcastreamHcalMinbias`
   HCAL inter-calibration using the phi symmetry method. Mean value and variance of the energy distribution in the readout are used. If the energy in the readout is collected without zero-suppression, i.e. negative values after pedestal subtraction are kept, the variance is used. If the energy in the readout is collected in zero-suppression mode the mean value is used. The AlCaReco content is limited to the the HB/HE, HF and HO collections of RecHits with no additional selection applied.

The matrix that defines the streams to be produced for a given data set is shown in Table 3. In general, more than one stream can be created per dataset, and more than one dataset can be associated to a particular stream. Therefore, the capability of the framework to write more than one output stream in parallel was crucial for this part of the challenge.

A combined configuration file (`Configuration/Examples/data/AlCaReco.cfg`) was prepared, which reads a RECO file and writes one or more AlCaReco streams according to the matrix shown in Table 3. Technically, the reconstruction itself and the AlCaReco streaming were performed as two consecutive steps.

Furthermore, configuration fragments were added to the central reconstruction configuration in order to enable the reading of ECAL calibration constants as well as tracker and muon alignment constants from the offline database, either directly from ORACLE, or using the FRONTIER caching.

For more details on the alignment and calibration related offline software, see [4].

5.8   Event Content Definitions

Raw data for this exercise is defined to be the "digi" representation of the detector information. The RECO data consists of a set of products produced during prompt reconstruction from the raw data that allow an event to be re-reconstructed, whereas AOD is a selected subset of essential information for analysis. The Full event format (FEVT) includes the raw digi data and the RECO data, with only a few intermediate tracking objects dropped. As this was a challenge based on samples produced from Monte Carlo generators, the Monte Carlo generator and detector simulation data was also kept with the distributed samples, leading to "RECOSim" and "AODSim" formats.

5.8.1   RECO content

Data samples in the RECO format contain mainly the following information:

- Track collections, including associated rec-hits, allowing track refits

- Tracker rec-hits

- Primary vertices

- Muon collections, both reconstructed locally in the muon detector, and combined with tracks reconstructed in the tracker detector

- Muon detectors rec-hits (including DT, CSC, RPC)

- Electrons and photon collections

- Clusters and Super-clusters reconstructed in the electromagnetic calorimeter

- ECAL rec-hits

- Jet collections reconstructed with different algorithms and missing Et

- Calorimetric towers

- HCAL rec-hits

Together with the reconstruction output, the RECOSim format stores Geant-4 tracks and vertices plus the full generator output. Jet collections reconstructed at the Generator level are also stored.

5.8.2 AOD content

The AOD is a proper subset of the RECO. This allows event skimming producing AOD out of data samples in the RECO or FEVT format without need to run any software module for data conversion. The Content of the AOD is defined essentially by dropping rec-hits collections from the RECO format:

- Track collections without associated rec-hits (no track refits is possible with AOD)

- Primary vertices

- Muon collections, both reconstructed locally in the muon detector, and combined with tracks reconstructed in the tracker detector. Track fits have no associated rec-hits

- Muon detectors rec-hits (including DT, CSC, RPC)

- Electrons and photon collections

- Clusters and Super-clusters reconstructed in the electromagnetic calorimeter

- Jet collections reconstructed with different algorithms and missing E_T

- Calorimetric towers

Together with the physics objects and reconstruction information, the AODSim format stores the full generator output. Jet collections reconstructed at the Generator level are also stored.

The output sizes for these formats for a few CSA samples are given in Table 4.

6 Production and Grid Tools

6.1 CRAB: Cms Remote Analysis Builder

CRAB is a tool which allows the users to access and analyze the remote data using the grid infrastucture. CRAB queries the CMS Data Management services for data discovery, prepare the job with the specific user code and take care to interact with WLCG/gLite/OSG middleware from job submission to output retrieval.

CRAB version used for CSA06 introduced the capability of using gLite with "bulk submission," increasing significantly the submission speed and so the submission rate.

Table 4: Event size for various data formats and samples.

Format	Event size
Minimum bias events	
FEVT	843 kB/evt
RECOSim	202 kB/evt
AODSim	83 kB/evt
T-Tbar events	
FEVT	3408 kB/evt
RECOSim	781 kB/evt
AODSim	309 kB/evt
EWK Soup events	
FEVT	1730 kB/evt
RECOSim	417 kB/evt
AODSim	197 kB/evt

7 Data Management

7.1 Dataset Bookkeeping System

The Dataset Bookkeeping System (DBS) for CSA06 included the functionality needed for cataloging Monte Carlo data and tracking some of the processing history. Included were data-related concepts of Dataset, File, File Block and Data Tier. The processing related concepts of Application and Application Configuration were provided to track the actual operations that were performed to produce the data. In addition, data parentage relationships were provided. A client level API enabled the creation of each of the entities described above. File information including size, number of events, status and Logical File Name (LFN) are included as attributes of each file. A discovery service was developed that enabled users to find data of interest for further processing and analysis.

7.1.1 Deployment and Operation

The architecture of the DBS service included a middle tier server running a CG script under Apache. All client access to the server was through an HTTP API. The CG script was written in PERL and access the database via the PERL DBI module. All activity for CSA06 was established on the CMSR production Oracle database server, with one Global DBS account and a half dozen so-called Local DBS accounts. The procedure was to produce Monte Carlo data under the control of four "Prod Agents", each with access to its own Local DBS instance. When the data was appropriately merged and validated its catalog entries were migrated to the global catalog for use by CMS at large. This migration task used allowed block-by-block transfer of Datasets to be done through a simple API.

There were two servers provided for CSA06, a "test" and "production" machine. The production servers were both dual Pentium 2.8 Hz processor with 2GB memory. The test server was heavily used by remote sites, including the initial data production, CMOS Robot submissions, and final skimming operations. The production server was used by the Tier-0 reconstruction farm,and some skimming operations near the end of CSA06. There were also ongoing CMOS activities included in the loads for the test server that are not related to CSA06. Access statistics for the service were obtained by mining the Apache access log files for each of the servers. The activity for the production server is shown during the month of November in Fig. 1 and Fig. 2. The important features to observe in this data are the number of pages served (Pages), and the total amount of data transferred (Bandwidth). As an example of a particularly busy day, November 27 showed 220k pages (query requests) and over 10 GB of data. This is a request rate of over 2.5 Hz and the server CPU was around 50% loaded. Demand on the Test server was heavy in October and the first part of November with peak rates of around 3 Hz in Mid October and again in early November.

7.1.2 Experience from CSA06

The overall operation and performance of the system was very good throughout the course of the CSA06 exercise. The limited functionality provided by the schema and API was sufficient for the test although many additional features are needed for the ultimate system. The dataset propagation from Local-scopes to Global-scope worked seamlessly. Maintenance of the server code was simple and straightforward. The clients were easily integrated with DLS, CRAB and Prod agent. The support needed for the DBS production server was at its minimal.

Clients occasionally reported slow response during peak periods, but the servers held up well. The service can be
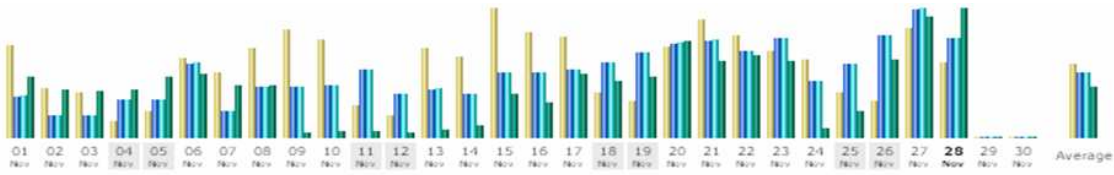
Figure 1: DBS production server November daily statistics. The bars on the chart for each day represent the number/amount of "Visits", "Pages", "Hits", and "Bandwidth" served. The scale and legend for the bar chart can be determined from the data in the table in Fig. 2.

Day	Number of visits	Pages	Hits	Bandwidth
01 Nov 2006	79	70174	71183	5.46 GB
02 Nov 2006	42	37701	38225	4.33 GB
03 Nov 2006	39	38208	38518	4.23 GB
04 Nov 2006	14	64200	64348	4.35 GB
05 Nov 2006	23	65038	65436	5.54 GB
06 Nov 2006	68	127344	127928	5.77 GB
07 Nov 2006	56	45813	46260	4.73 GB
08 Nov 2006	76	86642	87696	4.77 GB
09 Nov 2006	92	86267	87078	447.17 MB
10 Nov 2006	84	88627	89613	600.15 MB
11 Nov 2006	27	116329	116475	633.59 MB
12 Nov 2006	19	74603	74687	413.93 MB
13 Nov 2006	77	83033	83965	745.94 MB
14 Nov 2006	69	74494	75085	1.07 GB
15 Nov 2006	110	110457	111652	3.91 GB
16 Nov 2006	90	111812	112537	3.10 GB
17 Nov 2006	86	115972	116588	5.73 GB
18 Nov 2006	38	129160	129335	5.13 GB
19 Nov 2006	31	145523	145570	5.55 GB
20 Nov 2006	78	161702	162205	8.78 GB
21 Nov 2006	101	166394	167369	6.97 GB
22 Nov 2006	87	147144	147526	7.45 GB
23 Nov 2006	74	169733	169999	6.97 GB
24 Nov 2006	67	97496	97916	861.54 MB
25 Nov 2006	39	126041	126195	2.41 GB
26 Nov 2006	31	176068	176344	7.00 GB
27 Nov 2006	93	220007	220566	10.93 GB
28 Nov 2006	64	169911	170310	11.64 GB
29 Nov 2006	0	0	0	0
30 Nov 2006	0	0	0	0
Average	62.64	110924.75	111450.32	4.62 GB
Total	1754	3105893	3120609	129.42 GB

Figure 2: DBS production server November daily statistics.

15

easily scaled by adding additional machines and a load balancing mechanism, such as round robin DNS, and this will be examined for the final system. The loads of the CSA06 operation were artificially inflated because many Local DBS instances were being managed centrally at CERN, in addition to the Global instance. In the final system the local instances will not be operated at CERN. There were two incidents which resulted in service interruption, both caused by problems with the central CMSR database system. Ultimately, the DBS service rates needed will be reduced by the fact that only Global-instance traffic will go through the central CERN service.

There were several specific problems and concerns that can be noted based on the CSA06 experience:

1. Lack of proper communication between FNAL and Tier 0 for DBS needs resulted in some concepts missing in the schema and API functionality.

2. Parameter Set information was not properly stored and there were not enough APIs to relate datasets with these parameters sets.

3. Provenance information for a dataset and a file were not properly stored in DBS.

4. Block Management was not automated and was done externally on an irregular basis. This led to transfer of both opened and closed blocks. Also, initially blocks that were transferred could not be uniquely and universally identified but this was fixed.

5. The merge remapping API were not used and thus not tested because of unavailability of needed functionality with the Framework Job Report and Prod Agent.

6. Dataset migration from Local-scope to Global-scope was found to be slower than desired. It performed at rate of 1000 files per minute which caused problems for the "test" server behind the cmsdoc proxy server that timed out after 15 minutes.

These will be addressed in the next generation DBS being implemented post CSA06.

7.2 Data Location Services

The data location services (DLS) operated in CMS was based on the Local File Catalog (LFC) infrastructure provided by the EGEE grid project. Data is CMS is divided into blocks, which are logical groupings of data files. The block to file mapping is maintained in the DBS. The advantage of file blocks is that they reduce the number of entries that need to be tracked in the DLS catalog. Instead of an entry for each file there is an entry for every block, which in CSA06 typically contained a few hundred files.

7.2.1 Deployment and Operation

The LFC was deployed as a service provided by the WLCG and the DLS tools developed by CMS were deployed at locations that needed to query or update DLS entries. The tools were deployed and compiled from CVS and were deployed on a standard WLCG user interface machine (UI).

7.2.2 CSA06 Experience

The DLS performed stably over the challenge. New data blocks were created once per day for each dataset. This created a maximum of about twenty new entries per day, so the load for creating production datasets was small. The user analysis jobs and load generating job robot jobs queried the DLS to determine the location of data blocks, but only when creating new work flows. The query rate was larger than the new entry rate, but the DLS performed well.

The largest load in the system came from PhEDEx agents updating the DLS entries with data block locations at sites. The PhEDEx agents update the DLS with the status of all the complete blocks for a site on a ten minute time interval. This kept the latency for publishing complete blocks low, and was manageable with the small number of blocks used in CSA06. As the number of blocks grows, CMS may need to investigate local site caching of the DLS information and only update the DLS with changes from the previous block publication.

7.3 Data File Catalogs

CMS utilized a technique called the trivial file catalog (TFC) to provide the data catalogs for the site. The TFC utilizes a consistent namespace on each site to provide the catalog functionality that maps a logical file name to a physical file name in the storage system. There is a local site configuration file that points the applications to the common namespace.

7.3.1 Deployment and Operation

The TFC was deployed during the site validation phase of service challenge 4. The local site configurations were entered into the common CMS CVS repository, which provided tracking and aided with debugging from remote experts. The TFC was successfully deployed at all sites participating in the challenge activities and the feedback on deployment and operations was generally positive. All underlying storage systems could be accommodated and the site instructions were detailed. The local configuration file also provides the location of the local database cache and the local storage element to the application and could be used for other site specific elements.

7.3.2 CSA06 Experience

The TFC scaled well during the challenge. Even on sites with a high load of applications the logical to physical file name mappings were reliably resolved. The TFC did not represent too high a load on the name spaces of the underlying storage systems. An additional factor of four in the TFC rate should be possible in all the currently used storage systems.

7.4 Data Transfer Mechanism

Data in CMS was transferred between sites using the Physics Data Exporter (PhEDEx) system. The PhEDEx system relies on underlying grid file transfer protocols to physically move the files. While PhEDEx is capable of using bare gridFTP to replicate files, only File Transfer Service (FTS) driven transfers and Storage Resource Manager (SRM) transfers were operated during CSA06.

7.4.1 Deployment and Operation

CMS deployed an architecture where the FTS servers were located at each Tier-1 and supported channels for groups of "associated" Tier-2 centers. The association between Tier-1 and Tier-2 centers was intended for channel hosting, as the data can be sourced from any Tier-1 center in the CMS computing model [1, 2]. The FTS channels relied on SRM transfers and the FNAL Tier-1 center also supported SRM transfers driven directly from srmcp. The stability of the SRM service at the sites varied, but the percentage of time the transfer succeeded on the first attempt was improved over similar tests during service challenge 4, indicating the services are maturing.

7.4.2 CSA06 Experience

The architecture deployed in CMS for FTS transfers with channels hosted at the associated Tier-1 centers for the supported Tier-2 centers leads to a large number of FTS channels. The number of FTS channels supported at Tier-1 centers was larger than the number of FTS channels supported at CERN. The deployed FTS architecture will be re-examined for scalability and supportability.

7.5 Data Access

The CMS application was able to successfully read from the local storage element using RFIO, RFIO2, and dCache during CSA06. During the challenge the local file access was largely sequential and all protocols were able to meet the application input and output needs. The initial goal of the challenge was to reach 1MB/s per batch slot for Tier-1 and Tier-2 centers. On average CMS was able to reach approximately half the anticipated rate, which was improved after the end of the challenge with protocol specific tuning for the application.

8 Resources

8.1 Tier-0

The Tier-0 hardware consisted of two standard Castor pools, the Input and Export pools, a set of batch nodes running under LSF, and a management node.

The Castor pools consisted of disk-servers with 5 TB of space in 3 or 4 filesystems per server, with 13 and 32 servers in the input and export pools for a total of 64 TB and 155 TB, respectively. Garbage-collection was disabled, tape migration from the export pool was allowed. The input pool was populated with input data before the challenge began.

The batch nodes were 700 dual-CPU worker nodes with 2.8 GHz CPUs, with 2 GB of RAM, running SLC4. They were configured with only 2 LSF jobs per node, instead of the usual 3 per node that IT department usually require to achieve reasonable efficiency.

The management node, lxgate39.cern.ch, is also a dual 2.8 GHz node with 2 GB RAM, and a 120 GB local disk used for log files and drop-box files.

8.2 Tier-1

All 7 of the CMS Tier-1 centers participated in CSA06 and in the pre-challenge commissioning activities. The anticipated processing, storage and network capacity compiled in September, before the challenge began, can be seen in Table 5. The resources are divided by relative fraction in the Memorandums of Understanding (MOU). The estimates for required resources necessary to complete the challenge were 300 CPUs for processing and 70TB of storage capacity for a nominal Tier-1 center. Taking the resource requests into account, the resources provided roughly match with MOU pledges with the exception of FNAL, which provided more than was requested.

Table 5: Tier-1 Computing Resources for CSA06. Items within parenthesis were installed during the challenge or were available opportunistically

Site	CPU[kSI2k]	CPU[#]	Disk[TB]	Network[Gb/s]
ASGC	(228)	(140)	48+(36)	2
CNAF	225	300	40	10+(10)
FNAL	2200	1800	700	11
GridKa	220+(220)		40+(30)	10
IN2P3		250	70	10
PIC	150	120	50	1
RAL	200		60+(20)	11

8.3 Tier-2

In total 26 Tier-2 facilities committed to participate in CSA06 and participated in pre-challenge commissioning activities. The anticipated processing, storage and network capacity compiled in September, before the challenge began, can be seen in Table 6. The requested resources at a nominal Tier-2 center were 100CPU and 10TB of disk space. The average number of CPU across the centers is 120 and the average storage capacity is 15TB, though there is a large variation in the resources available at the CMS Tier-2 centers

8.4 Site Validation

In order to prepare the sites for CSA06 each site was asked to participate in Service Challenge 4 (SC4). The worldwide LHC Computing Grid (WLCG) service challenges were exercises to demonstrate specific capabilities of the services and sites. CMS asked the sites to perform 6 steps to demonstrate readiness.

1. Demonstrate that the site grid interfaces are functional.

Table 6: Tier-2 Computing Resources for CSA06. Items within parenthesis were installed during the challenge or were available opportunistically

Site	CPU[kSI2k]	CPU[#]	Disk[TB]	Network[Gb/s]
Aachen		94	5	1
Bari		30	5	1
Belgium_IIHE	150	84	25	0.1
Belgium_UCL	110	38	25	1
Budapest		85	2	0.5
Caltech	330	256	55	10
CIEMAT	260	200	10	1
CSCS	30	20	4	1
DESY		100	35	10
Estonia		100	15	1
Florida	320	240	25	10
IFCA		100	15	0.6
ITEP		20	1	1
JINR		10	1	1
KNU		56	15	10
Legnaro		150	16	1
London_IC		264	4	1
MIT	182		5	1
Nebraska	350	252	21	0.6
Pisa	76	50	6	1
Purdue		228+(256)	30 +(170)	10
Rome		48	5+(10)	1
SINP		24	1	1
SPRACE	311	242	28	1
Taiwan		20	6	2
UCSD		150	10	10
Wisconsin	547	428	90	10

- The site should pass the site functional tests in the EGEE framework
- The site should appear green in the grid catalog for the OSG framework

2. Install the PhEDEx service on the site and enable the agents for data import

3. Install the CMS software on the site and publish the existence in the grid information system. The first version installed was CMSSW_0_6_0, but software releases were upgrades over the summer.

4. Transfer a small test sample to the site and publish the existence

   - Transfer the sample using PhEDEx. The original test sample was fewer than 100 files
   - Verify the files were correctly entered into the Trivial File Catalog (TFC) on the site
   - Publish the data as accessible in the Dataset Location Service (DLS)

5. Successfully verify that a CRAB submitted job can execute on the site and access the data

6. Measure the IO from the local storage element with applications running on the site

9   Pre-challenge Monte Carlo Production

The Monte Carlo Production for CSA06 started in mid July. The original aim was to produce 50 million events in total to be used as input for prompt reconstruction.

Four teams from CIEMAT, DESY/RWTH, INFN/Bari, and University of Wisconsin, Madison, volunteered to run production using the Production Agent for the first time at large scale. All production related job submissions were Grid-based, we refrained from using local submissions entirely. After a short ramp-up when sites prepared for

production (e.g. most of the sites received the CMSSW software via a centrally managed installation mechanism, while a few managed the installation manually) a total of 28 sites offered resources for the pre-production step.

Table 7 shows the datasets by event category and the associated number of events that were requested and actually produced. All four teams started production with the simulation of minimum bias events.

Table 7: CSA06 Pre-challenge Production by event category.

CMSSW		Nb Events produced	Nb Events requested
0_8_1	minbias	39.8	25.0
0_8_2	TTbar	5.8	5.0
0_8_2	Zmumu	2.2	2.0
0_8_3	Wenu	4.6	4.0
0_8_3	SoftMuon	2.0	2.0
0_8_3	EWK Soup	5.6	5.0
0_8_3	Jets	1.2	1.2
0_8_3	Exo Soup	1.0	1.0
0_8_4	HLT Soup	5.0	5.0
	Total	67.2	50.2

The average event processing time observed on a 3.6GHz Xeon processor and the event size is shown in Figure 3.
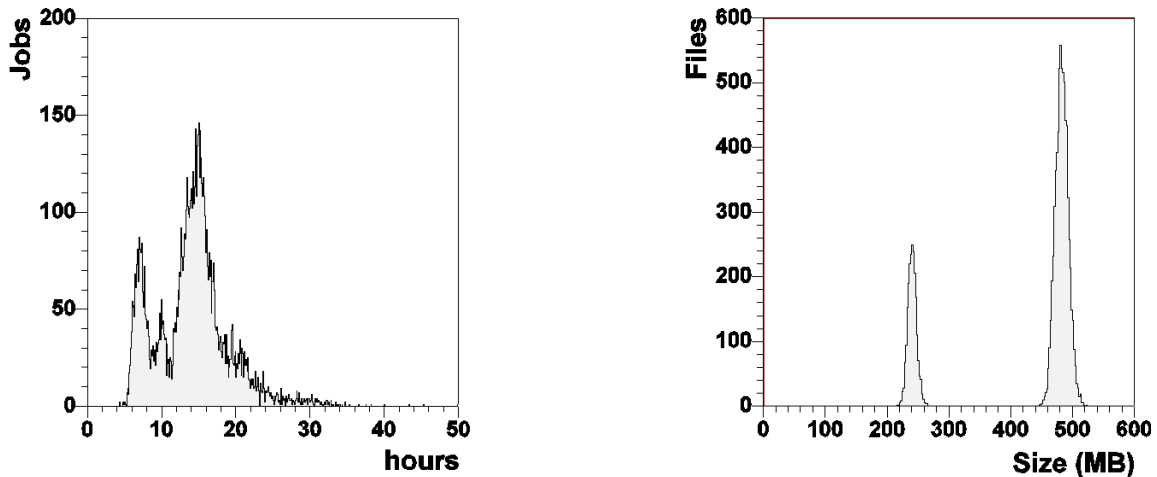
Figure 3: Minimum Bias Event processing time and Event size, where some jobs have 1000 events/job and some less than 500 events/job.

More complex signal events like those associated with the TTbar data sample take significant more time to simulate. According to the experience, it required about 4 minutes per event to complete.

Regarding the job submission strategy, the teams found that the information about resource usage at sites as it is published by the Grid Information System is not useful to build the ranking since it is lacking the job information associated with the particular VO role. Therefore a static ranking was used that was built according to the available resources as they were discovered by the ProdAgent's Job Tracking component.

With an average of up to 100 jobs/hour, per agent the performance of job submission by the ProdAgent was rather low. With the level of resources available to the teams it took a day or more until all CPUs could be utilized. Given the anticipated scale of production for CSA06 and the fact that there were four teams running two instances of ProdAgent each this was not a problem for the CSA pre-production, however needs to be taken into account for future production activities and is an area that needs to be improved. Moving to the new gLite resource Broker with its bulk submission feature may help to some extent, but this is certainly not the only area that needs to be looked at.

Rather than using the output sandbox for the produced data, files are staged out to the local Storage Element (SE). The performance of the process copying files from the the Worker Node disk to the SE, illustrated by Figure 5, is very good for all the prominent SE's CMS is using at sites (Castor, dCache and DPM).

As was reported by the production teams, early production was affected by instabilities in the JobTracking and
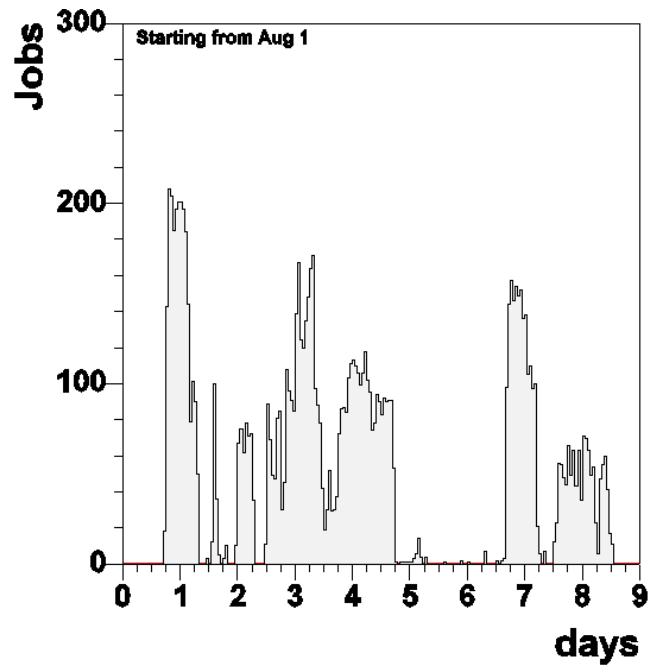
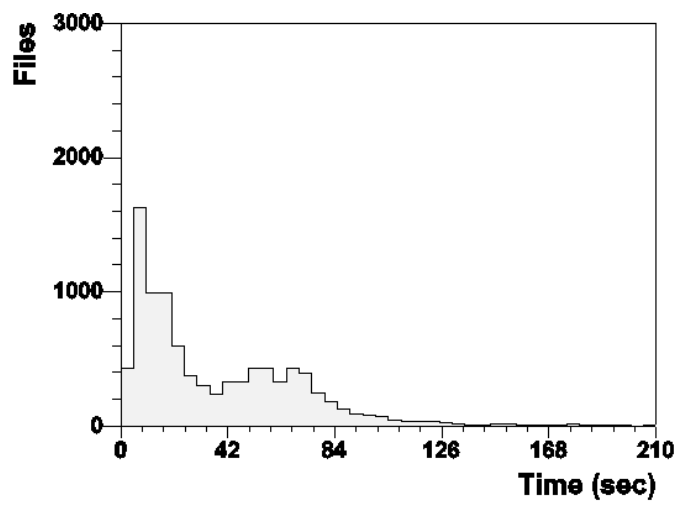Figure 4: ProdAgent job submission rate

Figure 5: Local stage-out performance

MergeSensor components of ProdAgent and required continuous attention by the operators. Fortunately the problems were solved in late August/early September.

Regarding operational problems one that turned out to be common across almost all participating sites was with data access between the farm of Worker Nodes and the local SE, i.e. for stage-out and in particular the merge process. Given the many processes running in parallel the latter has shown to stress some of the deployed SE's up to their limit. It is therefore important to maintain a suitable CPU to storage access bandwidth ratio.

To help operate the ProdAgent more efficiently people from INFN Bari developed a monitoring tool that allows a comprehensive overview of the current state and access to log files from a single web page. A screenshot is shown in Figure 6 and 7.
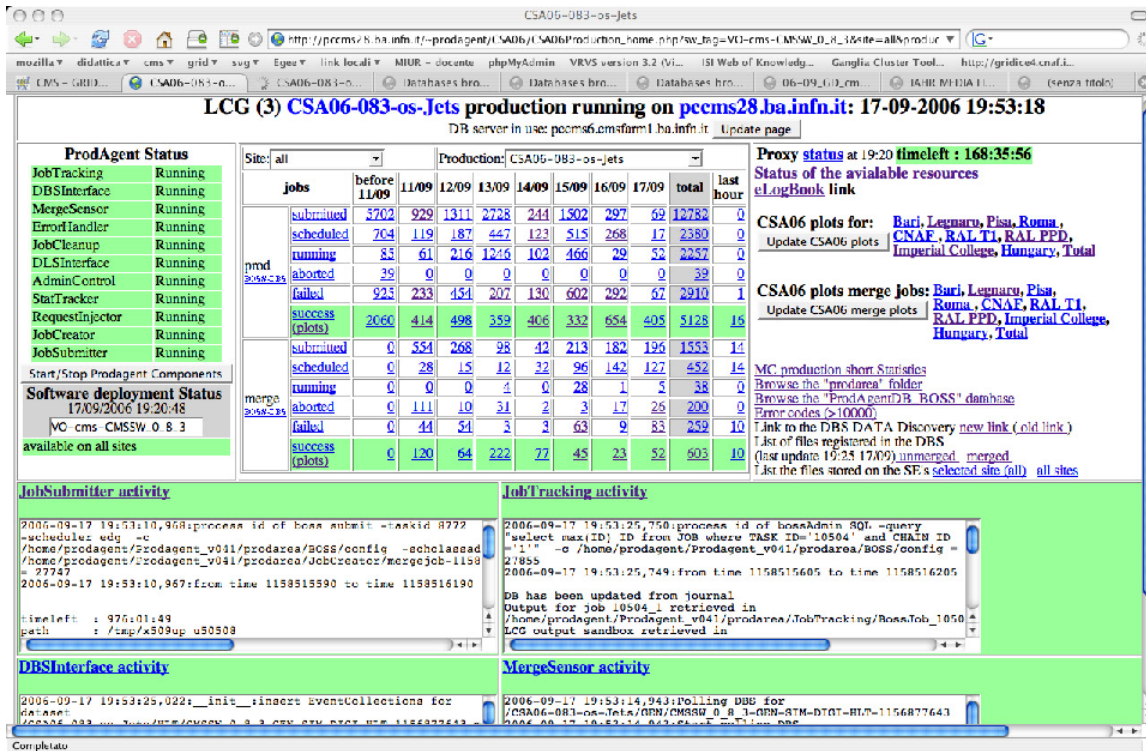
Figure 6: ProdAgent monitoring tool developed by INFN/Bari

10 Offline Database and Frontier

10.1 Frontier

The Frontier infrastructure was installed and tested prior to CSA06 and was used for the T0 operation and at T1 and T2 Centers. The goal was to observe the behavior of the frontier central servers at CERN, referred to as the "launchpad", and monitor the squids deployed at each participating site. The setup at CERN is shown in Fig. 8. There were three production servers, each running in tandem a tomcat server and squid server in accelerator mode. Load balancing and failover among the three servers is done via DBS round robin, and this worked flawlessly. The squids were configured in cache peer sharing mode which reduces traffic to the database for non cached objects.

Monitoring was in place to observe the activity for each squid through its SNMP interface and plots for 1) request rate, 2) data throughput and 3) number of cached objects was available for each installed squid. Lemon was used to monitor CPU, Network I/O, and other important machine operating parameters on the servers at CERN.

Initial tests with 200 T0 clients running CMSSW_0_8_1 were successful for the calibrations available at the time, ECAL and HCAL. However, when CMSSW_1_0_3 was tried a significant fraction ($\sim 5\%$) of jobs ending with segmentation faults was observed. Several additional problems emerged associated with the Si alignment when the software was run for the first time on the T0 system for the CSA exercise. The Si Alignment C++ object comprises a vector of vectors, which are translated in POOL-ORA into a very large number of tiny queries to the database. This makes loading the object quite slow, and frontier somewhat slower than direct oracle access. Due to the large
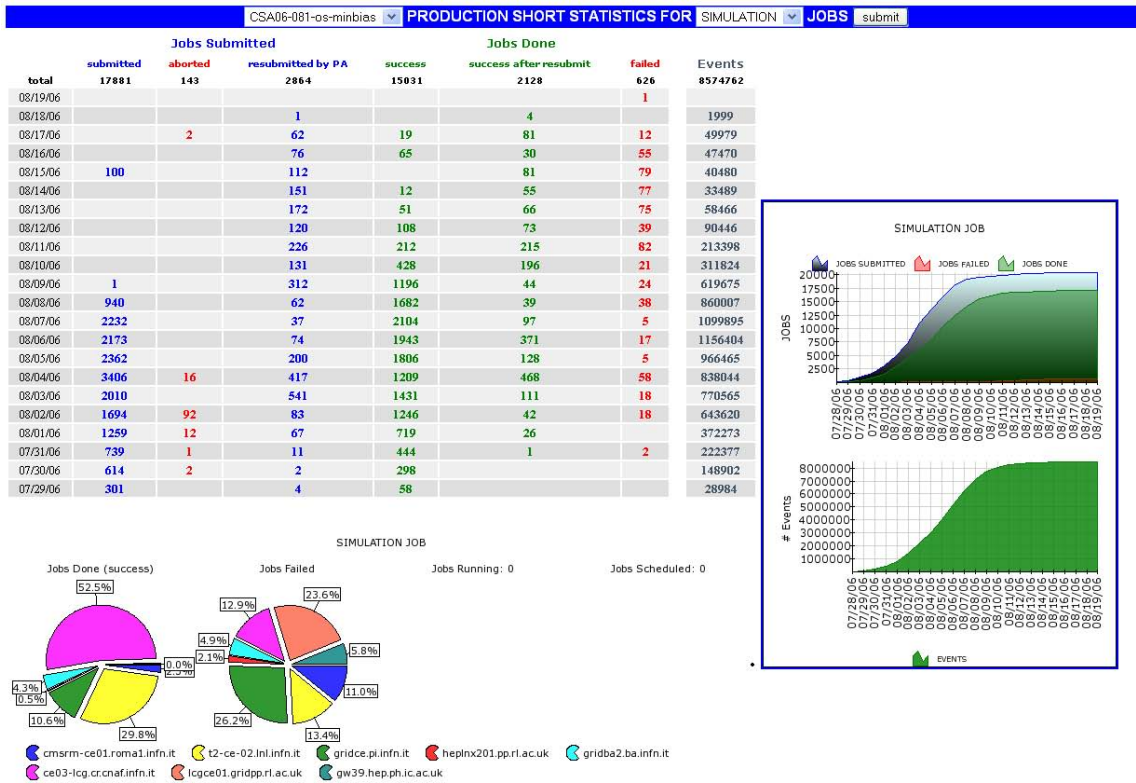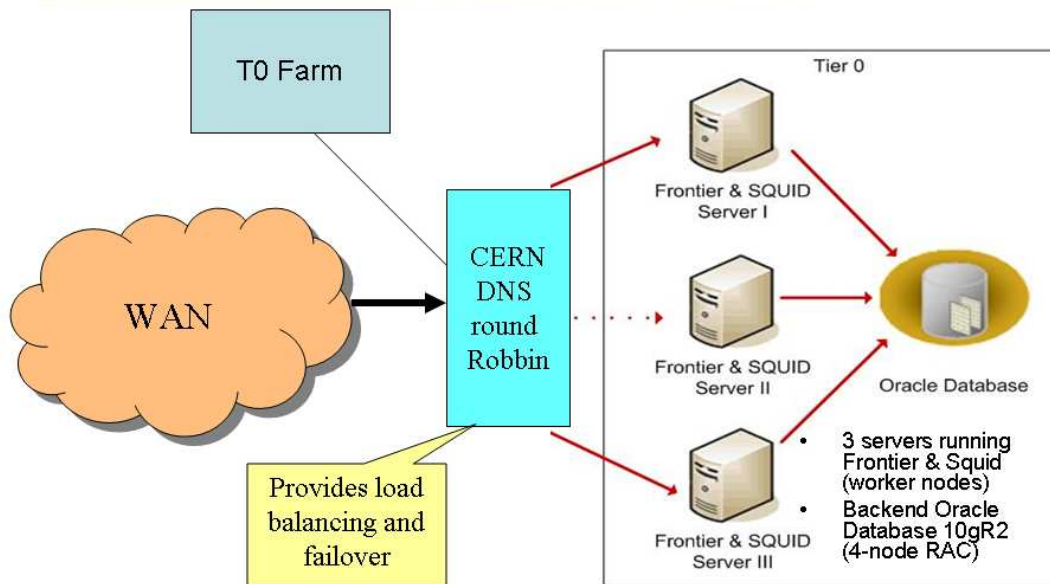
Figure 7: Bari ProdAgent monitoring summarizing important production job information

| | Jobs Submitted | | | Jobs Done | | | |
	submitted	aborted	resubmitted by PA	success	success after resubmit	failed	Events
total	17881	143	2864	15031	2128	626	8574762
08/19/06						1	
08/18/06			1		4		1999
08/17/06		2	62	19	81	12	49979
08/16/06			76	65	30	55	47470
08/15/06	100		112		81	79	40480
08/14/06			151	12	55	77	33489
08/13/06			172	51	66	75	58466
08/12/06			120	108	73	39	90446
08/11/06			226	212	215	82	213398
08/10/06			131	428	196	21	311824
08/09/06	1		312	1196	44	24	619675
08/08/06	940		62	1682	39	38	860007
08/07/06	2232		37	2104	97	5	1099895
08/06/06	2173		74	1943	371	17	1156404
08/05/06	2362		200	1806	128	5	966465
08/04/06	3406	16	417	1209	468	58	838044
08/03/06	2010		541	1431	111	18	770565
08/02/06	1694	92	83	1246	42	18	643620
08/01/06	1259	12	67	719	26		372273
07/31/06	739	1	11	444	1	2	222377
07/30/06	614	2	2	298			148902
07/29/06	301		4	58			28984

Figure 8: Frontier overview of launchpad and connection to WAN and T0 Farm.

23

number of calls to frontier, the squid access logs filled more quickly than we had observed in our previous testing and we were forced to temporarily turn them off.

A patch was found for the segmentation fault problem and this was implement and released in CMSSW_1_0_6. The root cause of the seg faults was non-threadsafe code in the SEAL library. By commenting out logging in the CORAL Frontier access libraries it was found the failure rate could be reduced to a few per mil. Additional work is underway to solve this problem.

10.1.1 Performance under T0 processing load

After the problems were resolved, there was a week of extensive operation and the T0 farm was ramped up to 1000 nodes. The number of requests and data throughput is shown in Fig. 9 and Fig. 10. This shows how the system behaved under loads ranging from 200 (Sunday) to 1000 (Wednesday) concurrent clients. These figures are for one of the three frontier server machine, although the other two servers looked very similar indicating that the load balancing was working as expected. The blue line in these plots indicate the requests that were not in the squid cache and had to be retrieved from the central database. The observed throughputs for each of three servers was at a maximum of around 660kB/s, which indicates that the 100Mbps network was not a bottleneck. The total throughput for the three servers was 1.8 MB/s.

Fig. 11 shows the server CPU for one of the servers during this same time period. Spikes are observed when new objects are brought into the cache, but there are no severe loads observed. Fig. 12 shows the CPU load for the same server under steady load during the 1000 client T0 test and it remains below 10% for the duration. The I/O during this same time is shown in Fig. 13. The fact that the input to the server is almost two-thirds that of the output was somewhat surprising, but is the result of the HTTP and TCP overhead, and it is significant because its size is the same order as the payload itself for the small objects.
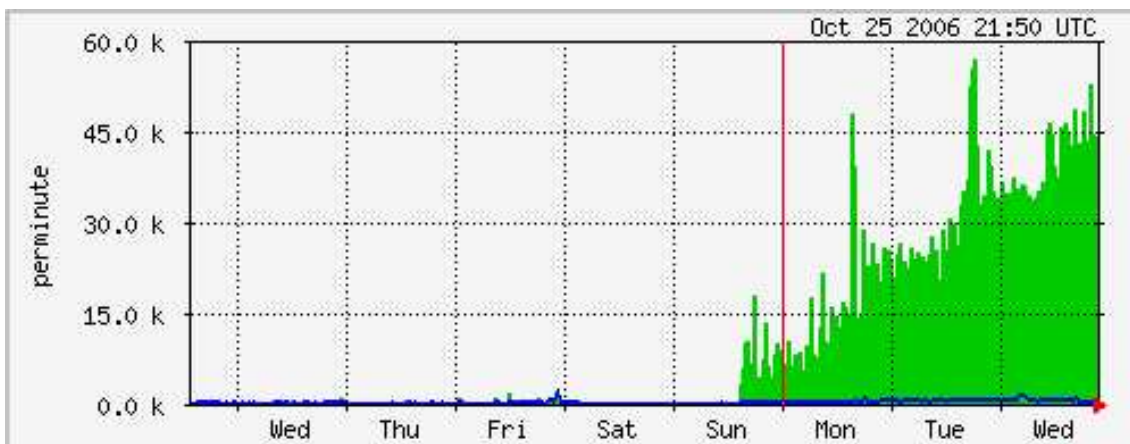
Figure 9: Requests to one of the three frontier servers from the T0 processing farm. The number of T0 nodes is ramped up form 200 to 1000 nodes during the time shown on the chart.

10.1.2 Tier N operation

In addition to the launchpad at CERN, there were 28 Tier-1 and Tier-2 sites where squid was installed and properly configured. Each of these squids is monitored through the SNMP interface and activity and history is available at the web site http://cdfdbfrontier4.fnal.gov:8888/indexcms.html. No remarkable issues were observed during this testing, however the large number of tiny objects problem makes a typical client startup take 15 minutes or more. For data not in the local squid caches, the startup was observed to be as long as 40 minutes.

10.1.3 Si Alignment Object Characteristics

To understand the characteristics of the Si object better we looked at the size and number of objects that were being requested. A single run of the RECO081_onlyCkf.cfg with the patched FrontierAccess had the following frontier statistics:
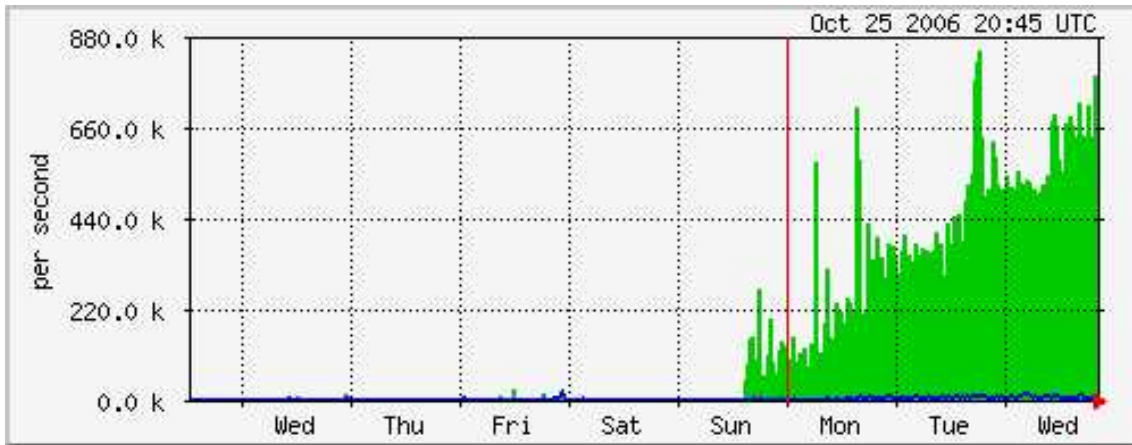
```
28116 queries
```

24

Figure 10: Data throughput for one of the three frontier servers from the T0 processing farm. The number of T0 nodes is ramped up form 200 to 1000 nodes during the time shown on the chart.
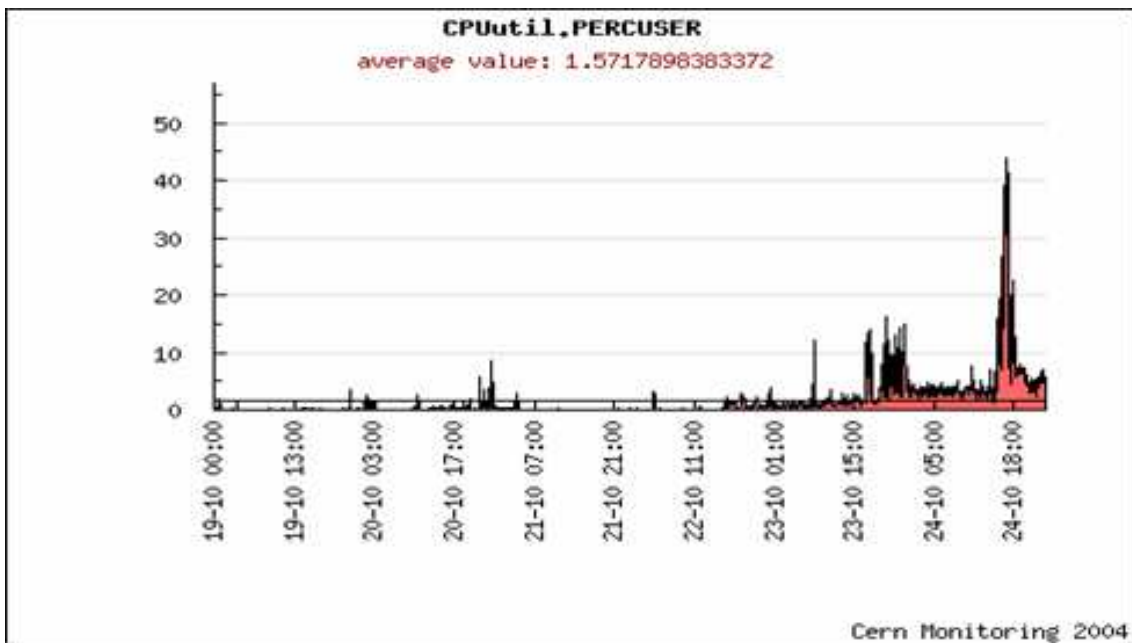

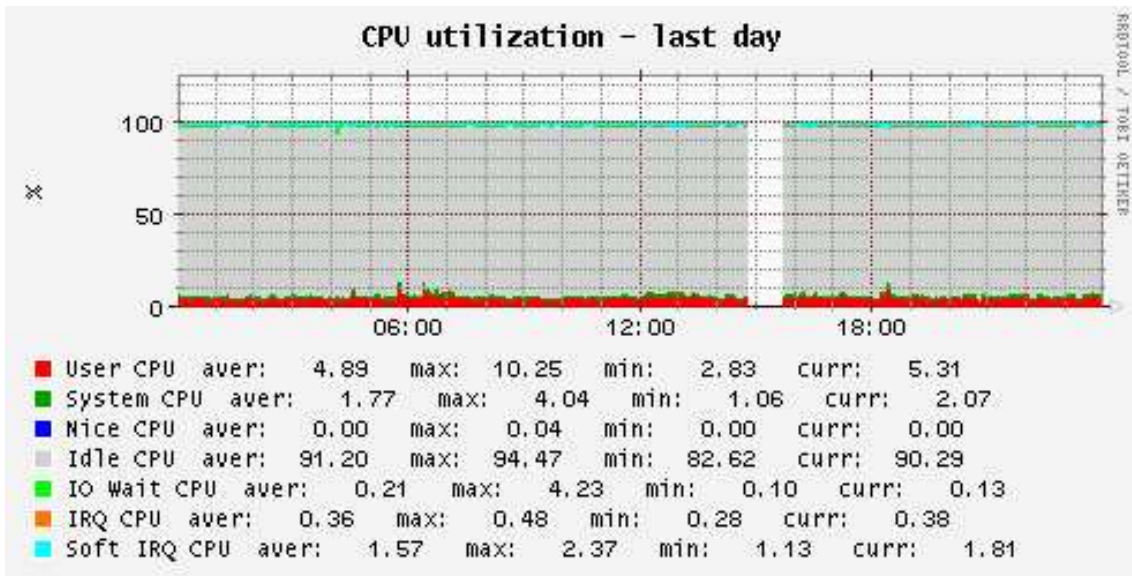Figure 11: Frontier server CPU usage during the ramp up of T0 activity.

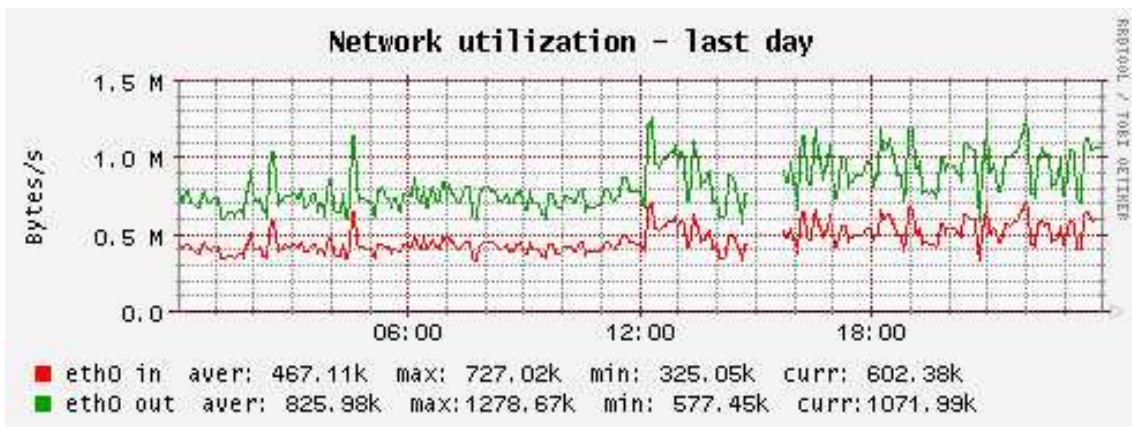Figure 12: Steady state CPU usage on Frontier server node during 1000 node T0 operation.

Figure 13: Steady state IO on Frontier server node during 1000 node T0 operation.

```
138 no-cache queries
342 queries of the database version
27502 unique queries

These are the largest payloads (full size = uncompressed):

  1369 byte (full size 12630),            25033 byte (full size 152389)
 20221 byte (full size 157849),           54251 byte (full size 575482)
 57316 byte (full size 597911),          109821 byte (full size 843757)
392046 byte (full size 2948642),         419859 byte (full size 3250885)
411531 byte (full size 3555809),         431981 byte (full size 6728489)
```

Everything else is under 4000 bytes full size, and 99% of the total queries are 317 bytes full size or smaller. The data was compressed by Frontier for the network transfer, the "full size" numbers refer to the uncompressed size.

The performance effect of this very large number (27k+) of small requests per job is being investigated. Job startup time for frontier is about 25% longer than direct oracle access at CERN (when running one job at a time and when the squid cache has been preloaded). We have prototyped reusing a single persistent TCP connection for all the frontier queries, but it only appears to account for about half of that difference in job startup time. Even with the persistent TCP connection, the small packets keeps the maximum network throughput with many parallel jobs down to around 1 Megabyte/second. By contrast, we have seen as high as 35 Megabytes/second throughput with larger queries over Gigabit ethernet (at Fermilab). The large number of requests are also responsible for producing squid access log entries at the rate of about 2GB/hour when 400 jobs run in parallel. The bottom line is that many tiny objects is not good for overall performance and must be avoided.

10.1.4 Site Configuration

Admins at each site are responsible for configuring their squid(s) to coincide with the hardware being used. One important question was whether the instrumentation we have is sufficient to diagnose specific site problems and help the site administrators fix them. One example we encountered during the CSA06 tests was an improperly configured cache at one of the sites. We noticed cached objects (# in cache) had "hair" as seen in Fig. 14. The requests per minute chart, Fig. 15, showed that there was a correlation to the unusual features. The precise problem was that their squid configuration had a very small disk cache, causing the objects in the cache to be "thrashed" quickly out.

The other important part of the site configuration is the so-called site-local-config file. This file is a bootstrap for jobs running at the site, and contains the frontier server URL and local squid proxy URLs. Many local-site-configs have been debugged and fixed over the course of CSA06. The CMS job robot started submitting jobs that include frontier access, to an ever increasing number of Tier-1 and Tier-2 sites.
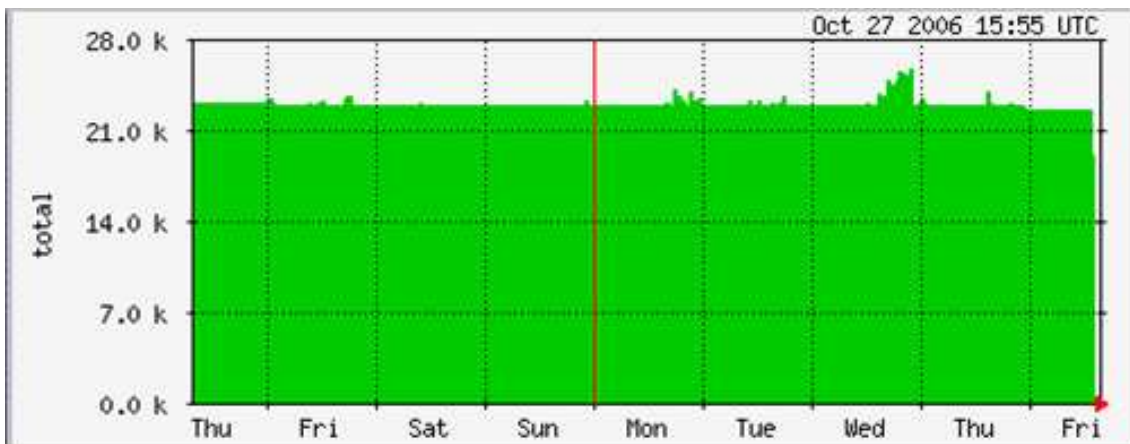

Figure 14: Squid configuration problem caused cache thrashing as indicated by the "hair" on the number of cached objects chart.
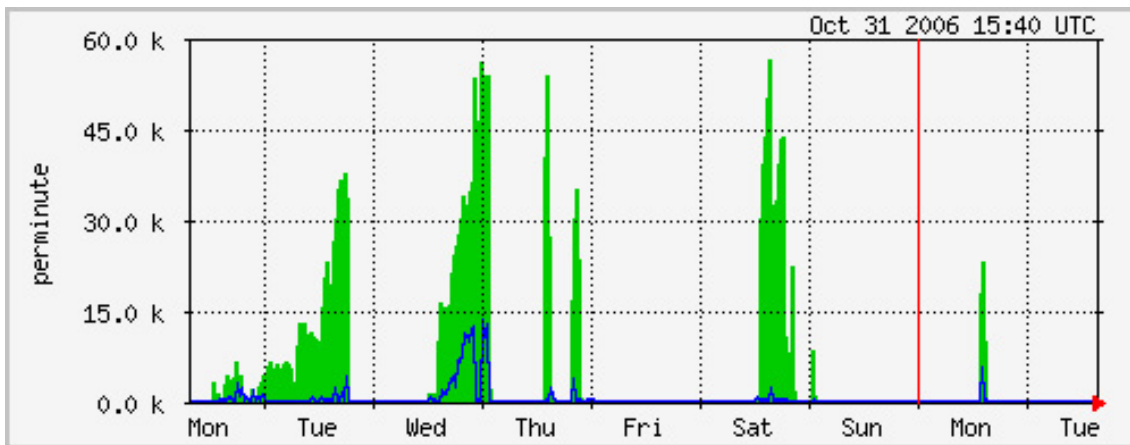
Figure 15: Requests rate during cache configuration problem.

10.1.5 Cache Coherency

One issue of concern for the objects in the Squid cache is cache coherency with the object stored in the central database. CMS has agreed to a policy of never changing objects that are stored into the central database, and ultimately this and other cache refresh options will be implemented. During the startup period, however, it was desired to have a mechanism that would provide periodic cache refresh in case the object was changed. This mechanism is implemented as an expiration time included in the HTTP header of each object which causes it to expire at 5 AM CERN time (3:00 AM UTC) the next day. The effect of this can be seen in Fig. 16 and Fig. 17. At 22:00 UTC the cache was dumped by hand, and the servelet installed that writes the expiration time in the header. Subsequently, it is observed that the objects expire and are refreshed between 3:00 and 4:00 UTC time.
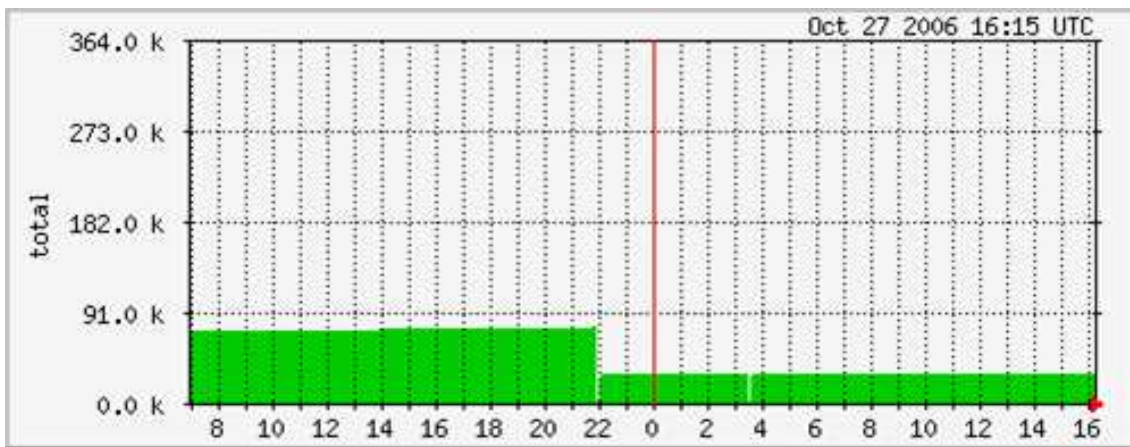
Figure 16: Object count on launchpad server when refresh is done by hand, and through the expiration at 3 am UTC.

This is an adequate solution for the short term and solves the cache coherency problem to within a few hours. However, reloading every cached object at every site all over the world will have significant performance implications and we must have a better solution for the final system. The impact on the launchpad servers is apparent in Fig. 18 and Fig. 19, which show spikes in the the CPU and I/O for one of the Frontier server machines as the caches are reloaded.

10.1.6 Conclusion

CSA06 Calibration and Alignment DB access via The Frontier infrastructure has been successfully exercised for up to 1000 clients on the T0 farm, and at T1 and T2 sites. The monitoring we have in place is extremely useful to observe the activity and understand performance at several levels of the system. The activity helped to uncover several issues that need additional work including:
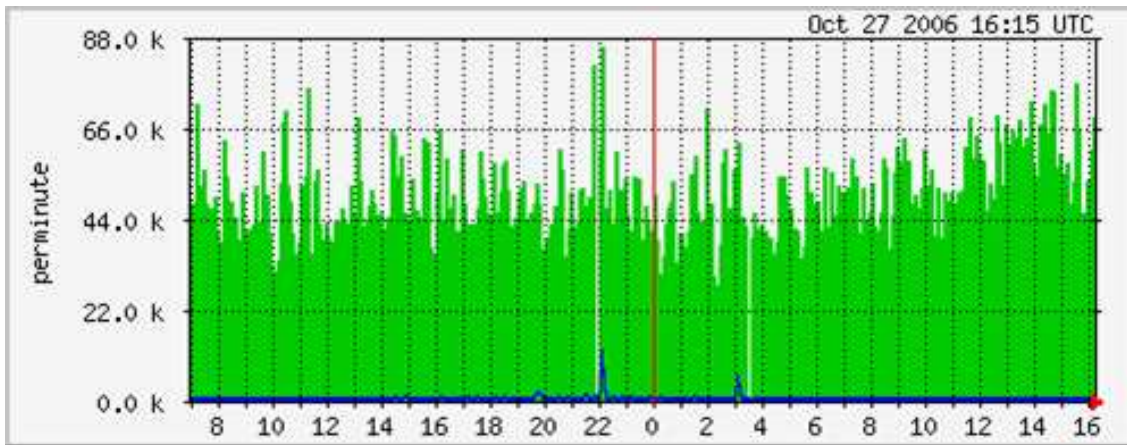
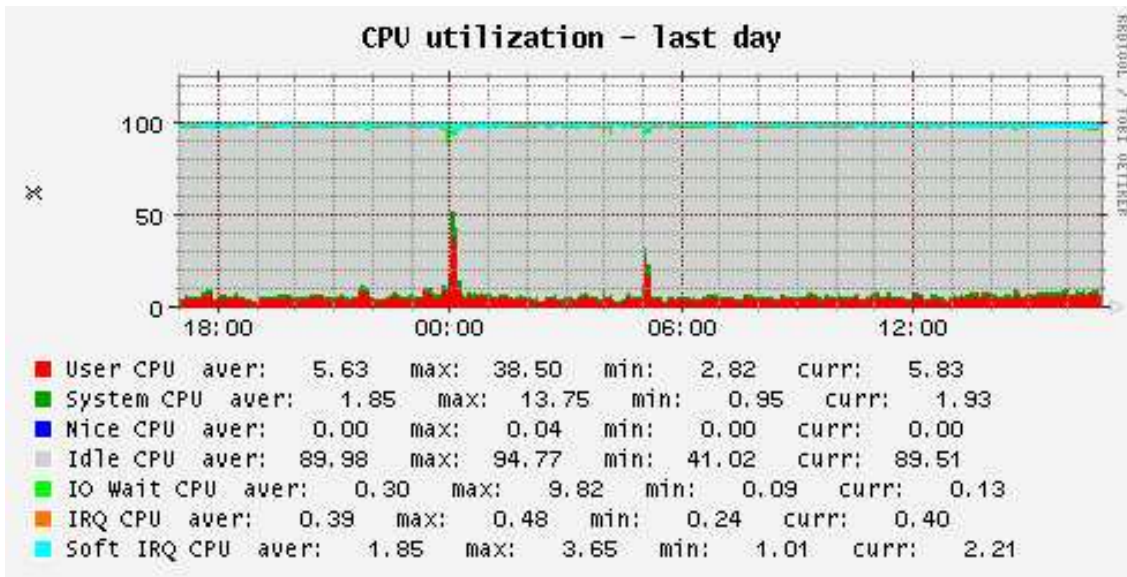Figure 17: Requests on launchpad server during cache refresh and expiration.

Figure 18: CPU usage on launchpad server during cache refresh and expiration.
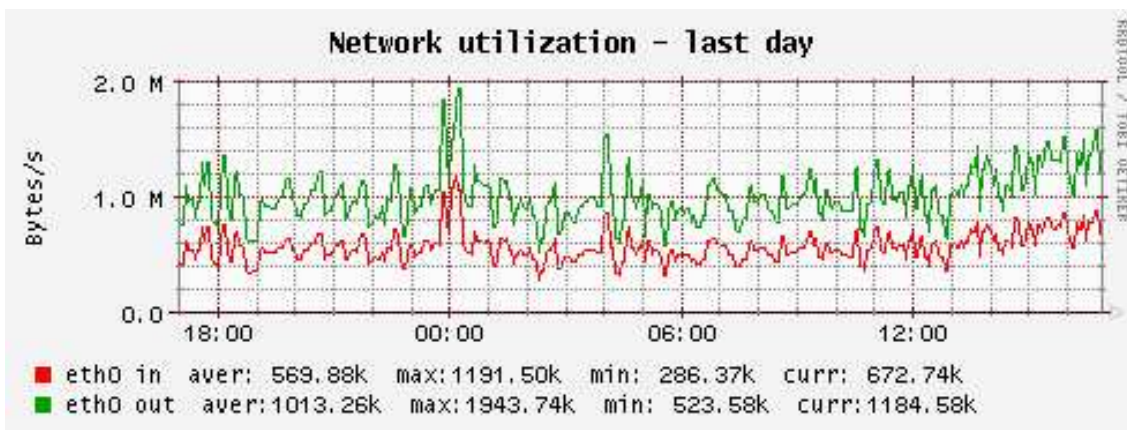

Figure 19: I/O on launchpad server during cache refresh and expiration.

- Threading problem found that causes seg faults.

- Lots of tiny objects in the SI alignment need to be consolidated.

- Logging of Squid access information can be copious.

- TCP connection overhead should be improved.

- Squid config and local-site-config.

- Cache coherency concern has temporary solution, but needs more work.

These areas and others will be addressed in the future. The configuration of the service at CERN is not final and work is needed to provide a dedicated Squid for the T0 farm. More work at Tier-1 centers will be done to provide failover solutions that will make the service more reliable at that layer, although there were no problems encountered over the course of the CSA06 tests.

11 The Tier-0

The Tier-0 had a pivotal role in CSA06. If it failed, nothing else could succeed, for lack of data. Robust error-handling and recovery file-by-file were therefore less important than continuous smooth operation of the total system. Failure of any given processing step resulted in the output being discarded, the rest of the system supplying enough data to ensure a steady flow. Unreliable reconstruction algorithms were therefore excluded, after initial testing, to ensure that the reconstruction software used in the Tier-0 was performant and robust.

The fundamental properties of the dataflow and workflow had already been verified at scales beyond the physics-startup using the 'July prototype', an emulation which allows exploration of the behavioural phase-space without relying on real events or CMSSW software. The basic design was thus known to be sound and workable, and the Tier-0 for CSA06 therefore concentrated on exploring the operational aspects.

The Tier-0 workflow for CSA06 encompassed most of the complexity required for first-physics: Prompt Reconstruction, creation of AOD and AlcaReco streams, merging of small output files to larger files for archive and export, insertion of files in DBS, injection into PhEDEx, and retrieval of calibration and alignment constants from Frontier. Only the communication with the Storage Manager and the repacking step were missing, since the input Monte Carlo data was not delivered in a format that was appropriate for such an exercise.

The Tier-0 ran for 4 weeks with 100% uptime, with no intrinsic scaling or behavioural problems. Operator interventions were required to introduce new CMSSW versions, modify the workflow as more features became available in CMSSW, adjust the rate of the different physics channels, and deal with minor problems caused by trivial bugs in the Tier-0 framework. The Tier-0 achieved all of its target metrics for CSA06.

11.1 System Architecture

The Tier-0 uses a Perl-based message-passing framework. Components in the workflow send messages to a central dispatcher/logger, these messages are then forwarded to other components. The forwarding is based on subscriptions, components declare what they are interested in receiving. This gives a modular, pluggable system; the workflow is determined by the interaction of the components only through their message contents, components are not directly coupled. The workflow can be changed on the fly to accommodate new requirements that may appear, and components can be stopped and restarted with no impact on the overall system behaviour.

The component structure used for CSA06 was to have one Manager component and zero or more Worker components for each step of the workflow (Prompt Reconstruction, AOD, AlcaReco, Fast Merge, DBS Update, and Exporter). The managers subscribe to messages that indicate input files are ready for them, then build queues of payloads for the workers. Workers ask the managers for work when they become idle. All components can report information to MonaLisa, and through it to the dashboard. Workers report per-task statistics (e.g. reporting every 50 events), the managers report per-step aggregates (e.g. reporting the total number of active Prompt Reconstruction processes).

The entire system is lightweight, flexible, robust, and has very low latency. This makes it very well suited to the operational environment expected for the Tier-0 during real data-taking.

The processing rate of the Tier-0 was determined by the rate at which files were injected into the Prompt Reconstruction. For a given input dataset, files were injected at intervals corresponding to a given rate in MB/sec. This rate was adjusted to correspond to the desired event rate in Hz, using the average event size per dataset.

11.2 Prompt Reconstruction

Prompt Reconstruction began at noon on October 2nd with CMSSW 1_0_2. Only minbias data was used in the first few days, following to the planning set out in the CSA06 wiki page. The EWKSoup sample (used as a pseudo express-line stream) was added on October 5th, in order to increase data-transfer rates. Job success-rates were over 99.7% in all channels.

Successive versions of CMSSW were used over the following weeks, as the code matured for reconstruction of the signal channels and for the other activities needed at the Tier-0 (AOD and AlcaReco production). At first, new versions were tested standalone, in parallel with the main running of the Tier-0. Later versions were deployed live in the Tier-0, without separate testing, but at a low enough level that any failures would not harm the smooth running of the total system. Once they were seen to be stable the event-rate with the new version was increased, and the older version was gracefully retired.

As each new version of CMSSW was deployed, reconstruction was restarted from the beginning of the input data. If a version of CMSSW was retired before the input data was completely processed, that input channel was left incomplete for that version of the software. With CMSSW_1_0_6, all input channels were run to completion, to provide a complete and coherent dataset for all subsequent activities. Essentially, the Tier-0 part of CSA06 was repeated from scratch in the last week, with CMSSW_1_0_6.

CMSSW_1_0_3 was deployed for the second and third weeks of running, being stable enough for the reconstruction of the signal channels. CMSSW_1_0_5 was used from the 19th to the 24th of October, and included the first AlcaReco streams, from minbias data. CMSSW_1_0_6 was used from the 22nd to the 30th, when the Tier-0 participation in CSA06 ended. This final version had all the AlcaReco streams, the AODs, and Frontier-access to conditions data.

The output of the Prompt Reconstruction contained the original input event as well as the reconstructed data, because of limitations in the CMSSW framework. This made the RECO output larger than the original input, so merging of RECO files was not useful. Prompt Reconstruction was therefore a one-file-in/one-file-out process. Event sizes and reconstruction times for CMSSW_1_0_6 are shown in Table 8.

Table 8: Prompt Reconstruction with CMSSW_1_0_6

Channel	Reconstruction Time (CPU sec)	Input Event Size (MB)	Output Event Size (MB)
EWKSoup	6.7	1.1	1.7
ExoticSoup	18.5	1.8	2.8
HLTElectron	8.6	?	1.8
HLTGamma	37.4	?	3.5
HLTJet	42.0	?	3.7
HLTMuon	8.4	?	1.8
Jets	22.8	1.6	2.6
minbias	2.9	0.5	0.8
SoftMuon	8.0	1.2	1.9
TTbar	19.3	2.0	3.4
Wenu	8.0	1.2	1.8
ZMuMu	8.4	1.2	2.0

11.3 AlcaReco Production

AlcaReco streams are produced according to the map shown in Table 3 in Section 5.7, relating input datasets with output AlcaReco streams. AlcaReco streams were first produced using CMSSW_1_0_3 running on minbias RECO data produced with CMSSW_1_0_2, for just the AlcastreamElectron stream, but not until CMSSW_1_0_6 were all input/output stream combinations available and useful. As with Prompt Reconstruction, all channels were run to completion with CMSSW_1_0_6.

The AlcaReco files are mostly small, by definition, so merging was required before they could be written to tape.

Simply writing and reading the files from tape would have been effectively impossible without this step, let alone the problems of analysing so many files once they were on disk.

AlcaReco production was essentially error-free, with no problems from the CMSSW application for any channels.

11.4   AOD Production

AOD production was run only with CMSSW_1_0_6, there being no suitable config file beforehand. As with AlcaReco, all channels were run to completion, and the output merged for efficient tape and analysis access.

AOD production was also trouble-free, with negligible failure rate.

11.5   Fast Merge

Merging is an integral part of the Tier-0. To minimize load on the storage system and our own data management, we have certain requirements for the minimum size of files we keep. But for various reasons (mostly due to optimizing workflows) we do have jobs that write output files which are significantly smaller then our requirements. In CSA06 AOD and especially AlcaReco output files were smaller then what is optimal. Due to the RAW+RECO output of the Prompt Reconstruction in CSA06, the Prompt Reconstruction output files were actually bigger then the RAW input files and didn't need to be merged.

The Merge component consisted of two parts, a manager and a number of workers. The manager subscribed to notifications of AlcaReco and AOD job completions and queued them internally. There were multiple queues separated by DataType (AlcaReco or AOD), dataset, CMSSW version and the configuration used (Pset hash) in the AOD or AlcaReco job. Once a new entry was added to a queue, the content of that queue was checked against three thresholds.

- Number of input files (FileThreshold)

- Number of events in input files (EventThreshold)

- Combined size of input files (SizeThreshold)

The FileThreshold was set to 32. This limit was introduced because of the next step in the workflow, the registration of the merged file into DBS. The registration was done with a shell command and all the parent files (ie. the input files of all the AlcaReco or AOD jobs whose output was being merged) were passed as command line arguments. Because we were concerned about shell command length limits, we restricted the number of input files to 32. In standalone longterm testing (without DBS registration) we already successfully explored scenarios with up to 150 input files.

The EventThreshold was set to 100000 since this is a useful number of events for AlcaReco studies.

The SizeThreshold was set to 3.9GB to prevent overly large files.

The actual merge operation on the workers was performed by the EdmFastMerge application. The input files were not directly accessed through Castor, instead we staged in all the files to the local disk with rfcp, ran the merge locally with local input and output and then staged out the output file back to Castor. For our operational requirements (merging of many small files) this was shown in earlier tests to be much faster then merging directly from Castor.

Since the CMSSW 1_0_6 processing cycle went through all input data within the last week of CSA06, only performance numbers for that period are quoted here. The performance numbers for merges run with CMSSW 1_0_5 are similar. Only for the CMSSW 1_0_3 cycle of AlcaReco merges did we see a significant amount of errors. About a third of these merge jobs failed with Castor stage in errors.

In the CMSSW 1_0_6 processing cycle there were a total of 5263 merge jobs submitted, 2436 for AOD and 2827 for AlcaReco. 33 jobs failed due to Castor stage in errors. Out of these 33 failed jobs, 31 were successfully rerun. The stage in error that was observed (which is the same that caused much more havoc in the CMSSW 1_0_3 cycle) is a bug in Castor that can be triggered under certain conditions. The Castor team is aware of the problem.

The remaining 2 jobs were rerun twice but failed both times. Further analysis showed the merge input files to be corrupted, ie. even an interactive rfcp of the files would always fail. This kind of problem is surprising since the job that created these corrupted files checked the rfcp exit code when it staged them out to Castor.

No failures during the merge itself or for the stage out were observed.

Figs. 20 and 21 show the distribution of number of input files for AlcaReco and AOD merges. The AlcaReco plot shows nicely that many merge jobs are triggered at the threshold of 32 input files.
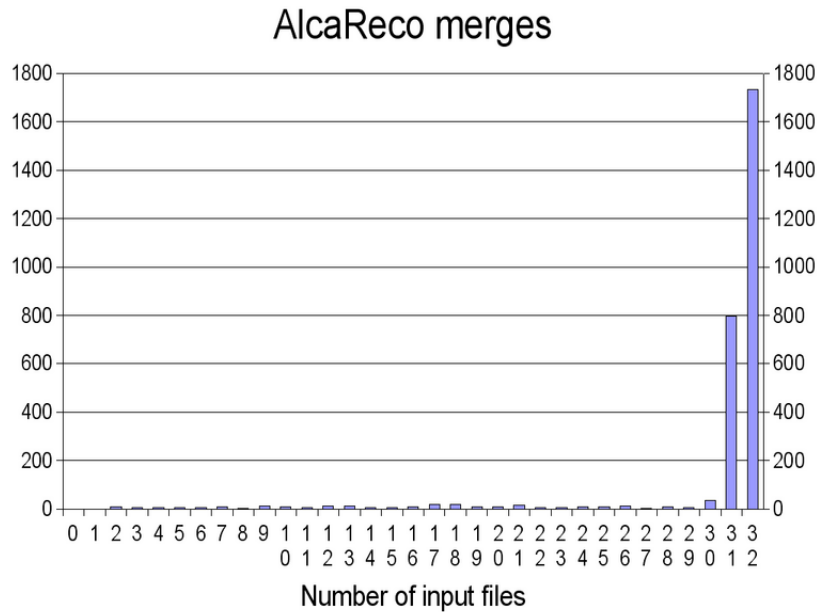

Figure 20: Number of input files for all CMSSW 1_0_6 AlcaReco merges at Tier-0.

Figs. 22 and 23 show the distribution of output file size for AlcaReco and AOD merges. One can see that most AOD merges are triggered at the 3.9GB file size threshold. AlcaReco merges are more diverse, some are quite large (almost reach the 3.9GB file size threshold), while many have quite small output file sizes.

And lastly, Figs. 24 and 25 show the distribution of number of events for AlcaReco and AOD merges.

11.6 Data Registration

Figs. 26 and 27 show the latency between the RecoReady notification (i.e. Prompt Reconstruction job finished) and the completion of the PhEDEx drop for the RECO file. Only RECO jobs run with CMSSW 1_0_6 software were considered. In Figure 26 there is a long tail up to about 2000 seconds, but it is a flat tail without structure or spikes. Figure 27 shows the average latency for the day in October including the statistical errors.

Figs. 28 and 29 show the latency between the RegisterMerged notification (i.e. completion of a merge job) and the completion of the PhEDEx drop for the merged file. Only merge jobs run with CMSSW 1_0_6 software and with input files produced with CMSSW 1_0_6 software were considered. In Figure 28 there is a long tail up to almost 3000 seconds, but it is a flat tail without structure or spikes. Figure 29 shows the average latency for the day in October including the statistical errors. There is no value for day 24 shown on this plot since the number of merges is offscale with a large error: 206 with an error of 1200. These plots contains a mix of AOD and AlcaReco merges with various different merge scenarios (number of files, size etc.). None of these parameters should effect the DBS registration except maybe the number of input files because of the registration of parentage in DBS. If there were large differences depending on number of input files, one would expect that to show up as a big statistical error on the average. For the days that show low latencies, the effect is not seen.

11.7 PhEDEx Data Injection

Each PhEDEx drop prepared by the Tier-0 workflow needs to be parsed and its meta data informations made available to PhEDEx. This task is performed by a dedicated PhEDEx process, which analyzes the drop, extracts all the informations and feeds them to the Transfer Management Data Base (TMDB). This whole process is called "data injection".

Since this process takes a non-negligible amount of time, the latency caused by this workflow step was analyzed.
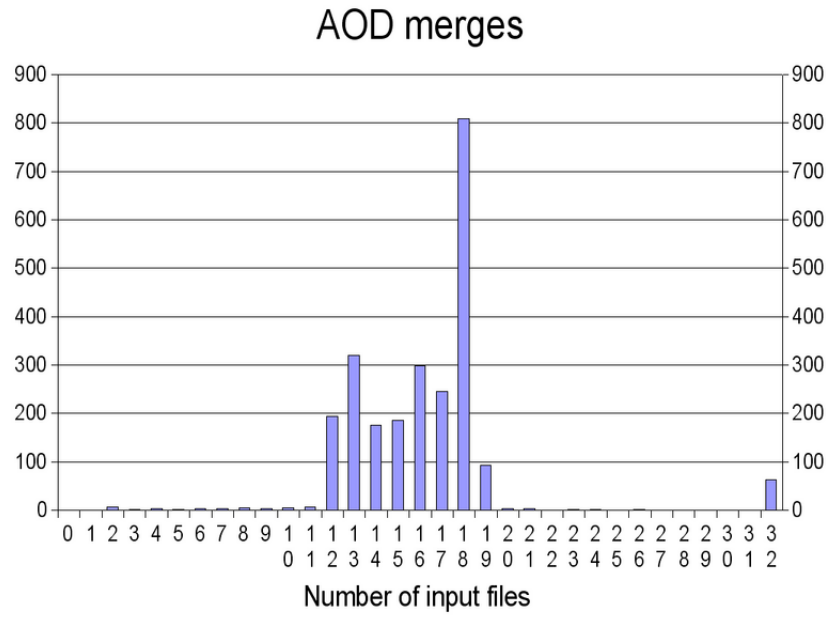
Figure 21: Number of input files for all CMSSW 1_0_6 AOD merges at Tier-0.
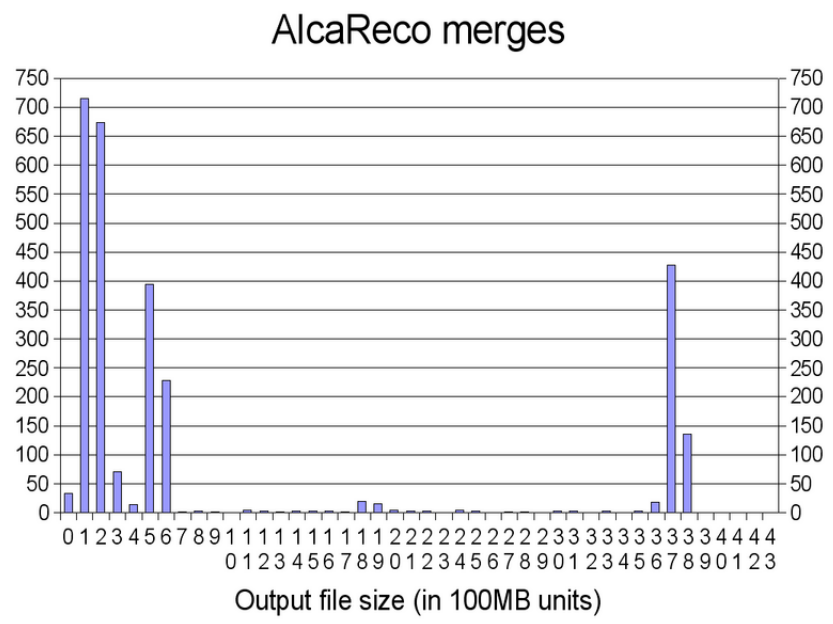

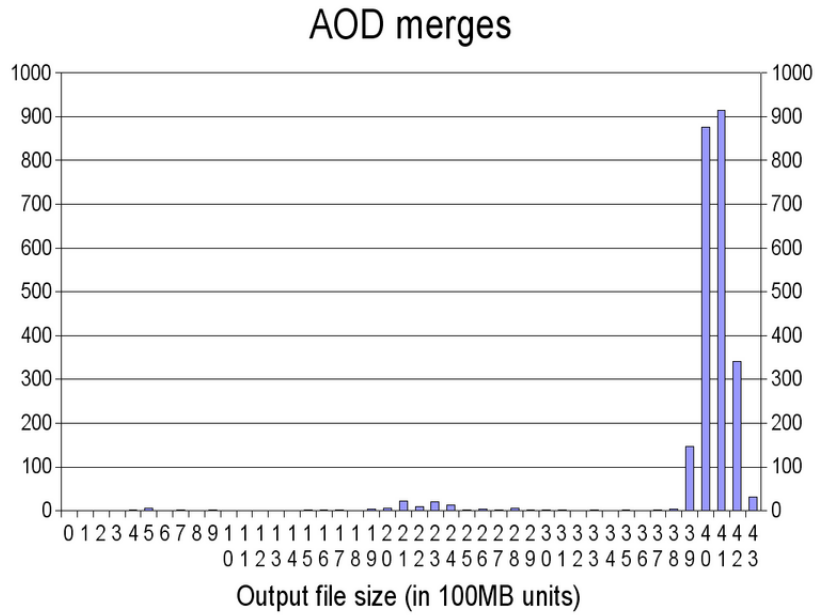Figure 22: Output file size for all CMSSW 1_0_6 AlcaReco merges at Tier-0.

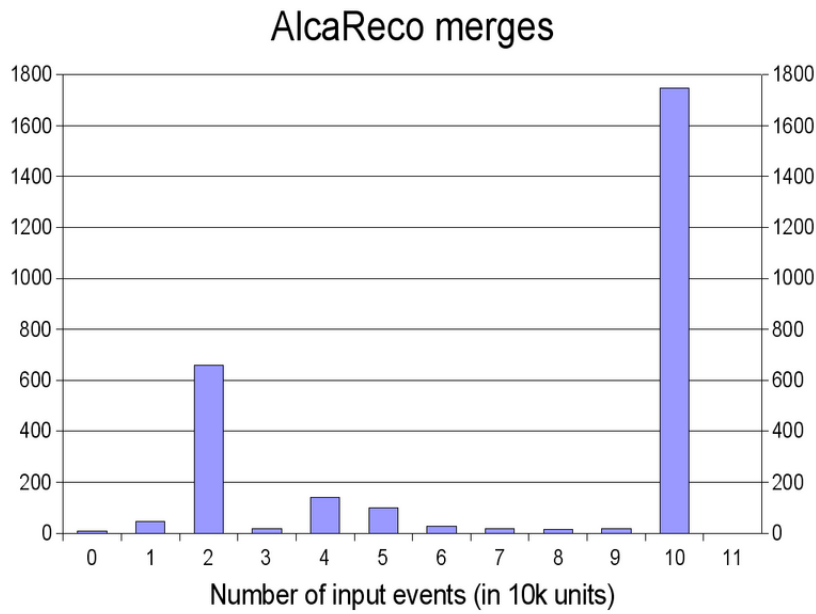Figure 23: Output file size for all CMSSW 1_0_6 AOD merges at Tier-0.

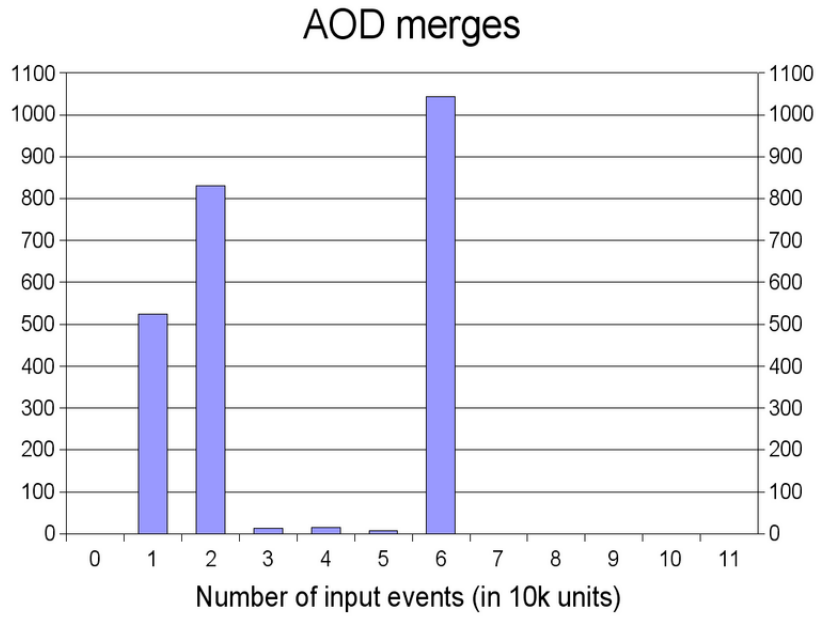Figure 24: Number of events for all CMSSW 1_0_6 AlcaReco merges at Tier-0.

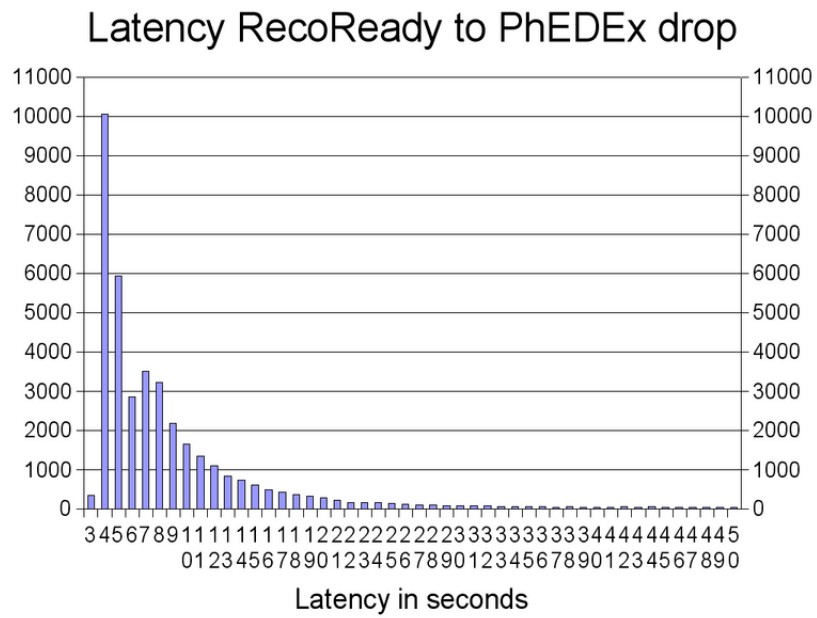Figure 25: Number of events for all CMSSW 1_0_6 AOD merges at Tier-0.

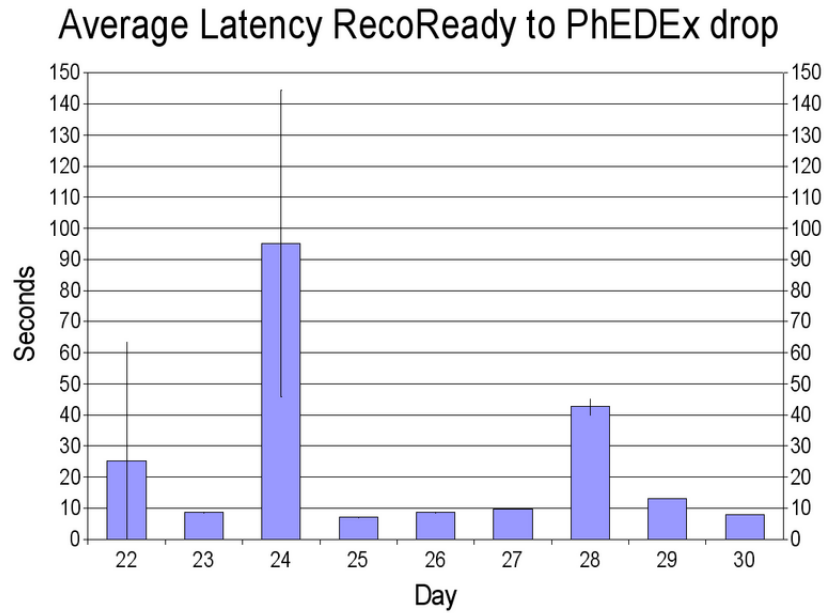Figure 26: Latency between RecoReady and PhEDEx drop for all CMSSW 1_0_6 jobs at Tier-0.

Figure 27: Average latency by day between RecoReady and PhEDEx drop for all CMSSW 1_0_6 jobs at Tier-0.
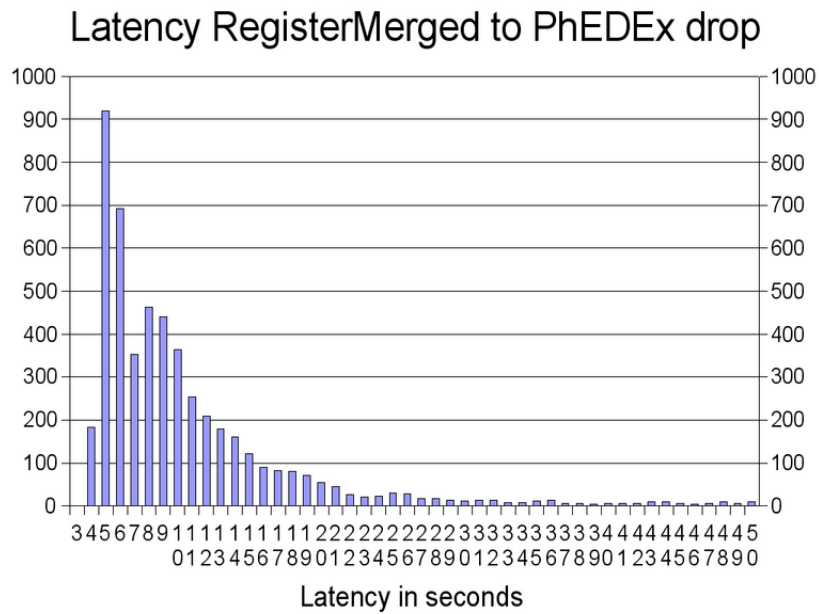
Figure 28: Latency between RegisterMerged and PhEDEx drop for all CMSSW 1_0_6 jobs at Tier-0.

Average Latency RegisterMerged to PhEDEx drop
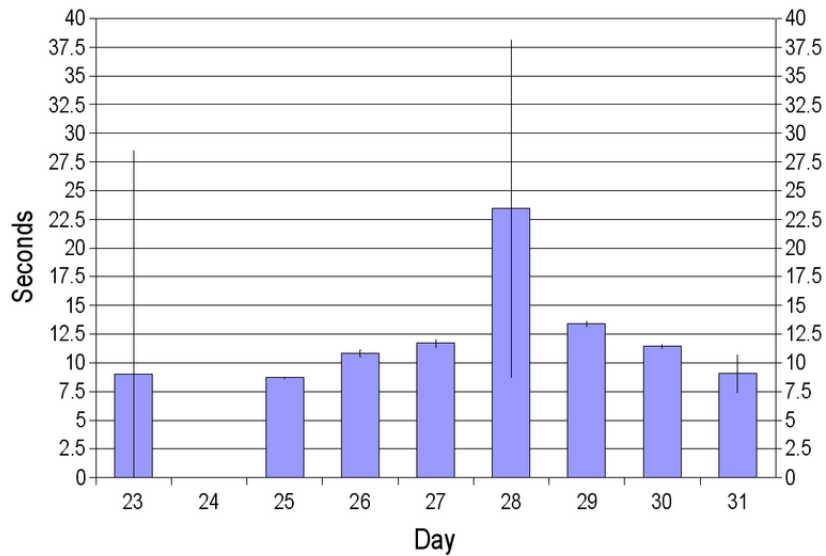
Figure 29: Average latency by day between RegisterMerged and PhEDEx drop for all CMSSW 1_0_6 jobs at Tier-0.

Fig. 30 shows the average injection time per day and the corresponding statistical error for this measurement. Up to day 25 only one injection agent was used, while from day 26 onwards, 5 parallel injectors were started in order to improve performance. During the first few days of CSA06 the average time per injected file was at a constant

Injection performance
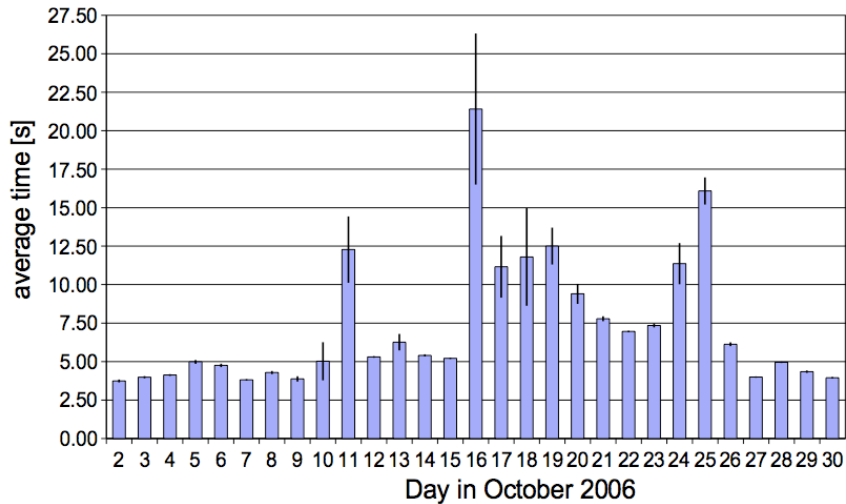
Figure 30: Latency between reception of PhEDEx drop and finalization of data registration in PhEDEx. Up to day 25 only one injection process was used, while from day 26 onward 5 parallel injectors were running.

level of about 3 seconds with negligible statistical errors. However, up to day 25 a slow increase of the average injection time is visible. This effect is most likely correlated to the growth of data registered in TMDB, which slows down the injection performance. Starting with day 26, the latency was reduced to what was observed during the first days, since from that date on five injection processes were started in parallel until the end of CSA06.

During the whole period of CSA06 a PhEDEx drop contained only informations for one file. The overhead of reading an XML drop per file could be reduced by grouping multiple files of the same block in one drop. A brief test was conducted using about 50 files of the same block per drop and a speed up of the injection operation to

about 0.5s per file was observed, which corresponds to a performance boost of a factor six. It is recommended to implement such a feature in the drop creation process of the final system in order to further optimize the workflow performance.

11.8   Performance

The event processing rate during CSA06 is shown in Figs. 31 and 32, where the latter plot shows the peak of the processing during the last day. The cumulative volume of produced events is shown in Fig. 33, and totals 207M events at the end of the challenge.

The processing rate in events/sec is not in fact a particularly meaningful metric for the Tier-0 in CSA06. The event-mix was varied considerably to accommodate external requirements, which gives a wide variety of mean reconstruction times throughout the challenge. The set of reconstruction algorithms was explicitly pruned to achieve the stability needed from CMSSW, and the algorithms thereby excluded tended to be the most complex, time-consuming algorithms. Finally, pile-up was not included in the simulated events. This makes the tracking, in particular, rather fast compared to realistic events.
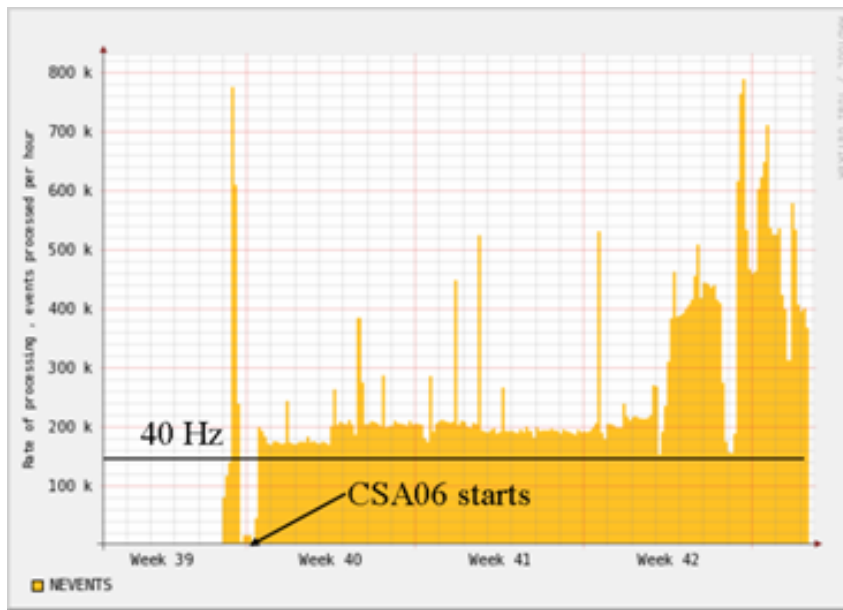
Figure 31: The processing rate at Tier-0. The target rate of 40 Hz is illustrated.

12   Tier-1 and Tier-2 Operations

12.1   Data Transfers

The Tier-1 centers were expected to receive data from CERN at a rate proportional to the 25% of the 2008 pledge rate and serve the data to Tier-2 centers. The expected rate into the Tier-1 centers is shown in Table 9. Note that while the listed rates are significantly less than the bandwidth to the WAN (see Table 5), they fit within the storage capability available for a 30 day challenge.

Table 9: Expect transfer rates from CERN to Tier-1 centers based on the MOU pledges.

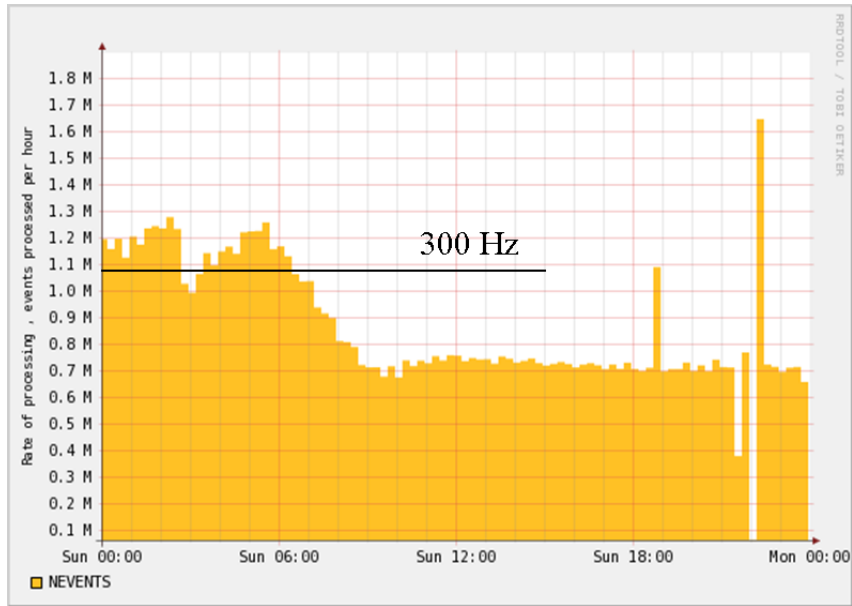| Site	Goal Rates (MB/s)	Threshold Rates (MB/s)
ASGC	15	7.5
CNAF	25	12.5
FNAL	50	25
GridKa	25	12.5
IN2P3	25	12.5
PIC	10	5
RAL	10	5

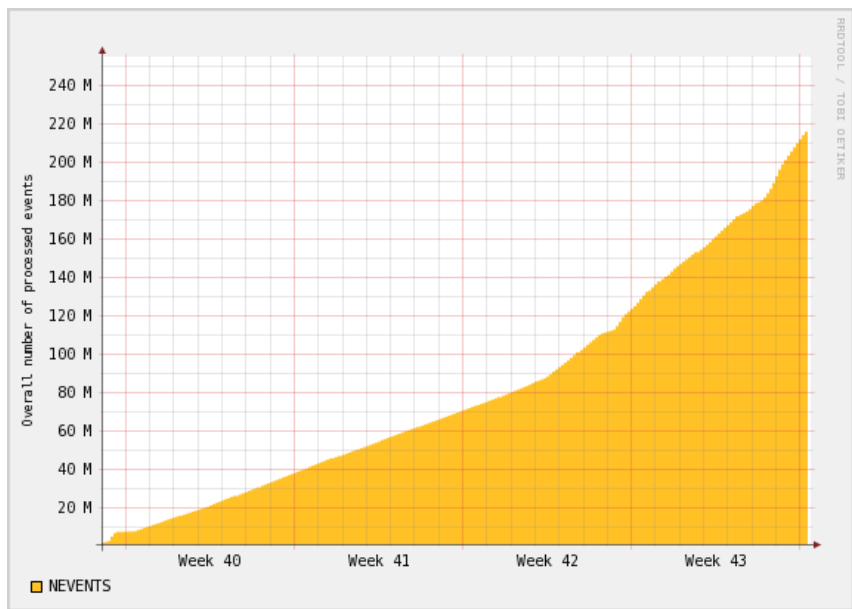Figure 32: The peak processing rate at Tier-0 in the last day of Tier-0 operations.

Figure 33: The total number of produced events at Tier-0.

The Tier-2 centers are expected in the computing model [1, 2] to transfer from the Tier-1 centers in bursts. The goal rate in CSA06 was 20MB/s, with a threshold for success of 5MB/s. Achieving these metrics in the computing model was defined as hitting the transfer rate for a 24 hour period. At the beginning of CSA06 CMS concentrated primarily on moving data from the "associated" Tier-1 centers to the Tier-2s. By the end of the challenge most of the Tier-1 to Tier-2 permutations had been attempted.

The total data transferred between sites in CSA06 is shown in Figure 34. This plot only includes wide area data transfers, additionally data was moved onto tape at the majority of Tier-1 centers. Over the 45 days of the challenge CMS was able to move more than 1 petabyte of data over the wide area.
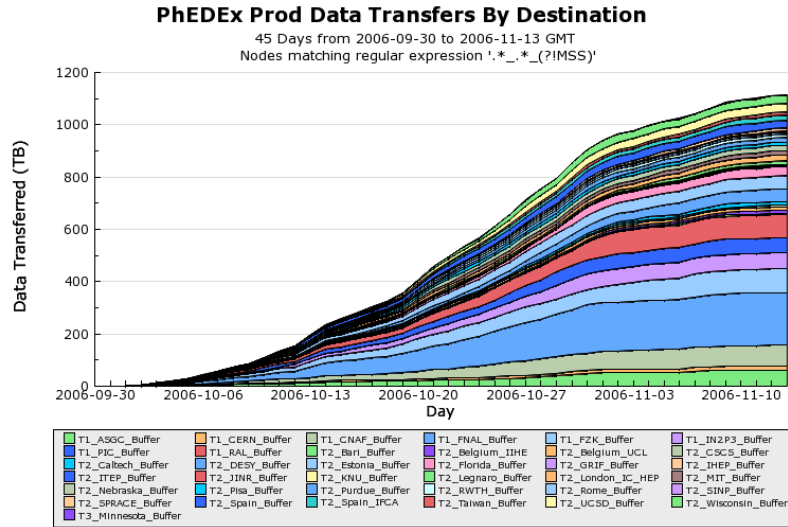
Figure 34: The cumulative data volume transferred during CSA06 in TB.

Timeline:

- October 2, 2006: The Tier-0 to Tier-1 transfers began on the first day of the challenge. In the first few hours 6 of 7 Tier-1 centers successfully received data. During the first week only minimum bias was reconstructed and at 40Hz the total rate out of the CERN site does not meet the 150MB/s target rate.

- October 3, 2006: All 7 Tier-1 sites were able to successfully receive data and 8 Tier-2 centers were subscribed to data samples: Belgium IIHE, UC San Diego, Wisconsin, Nebraska, DESY, Aachen, and Estonia. There were successful transfers to 6 Tier-2 sites.

- October 4, 2006: An additional 11 Tier-2 sites were subscribed to data samples: Pisa, Purdue, CIEMAT, Caltech, Florida, Rome, Bari, CSCS, IHEP, Belgium UCL, and Imperial College. Of the 19 registered Tier-2 sites, 12 were able to receive data. Of those, 5 exceeded the goal transfer rates for over an hour, and an additional 3 were over the threshold rate.

- October 5, 2006: Three additional Tier-2s were added increasing the number of participating sites above the goal rate of 20 Tier-2 centers. New hardware installed at IN2P3 for CSA06 began to exhibit stability problems leading to poor transfer efficiency.

- October 9, 2006: RAL transitioned from a dCache SE to a Castor2 SE. The signal samples began being reconstructed at the Tier-0.

- October 10-12, 2006: The Tier-1 sites had stable operations through the week at an aggregate rate of approximately 100MB/s from CERN. IFCA has joined the Tier-1 to Tier-2 transfers and their average transfer rate over the day was observed at 14MB/s at a low error rate.

- October 13, 2006: Multiple subscriptions of the minimum bias samples were made to some of the Tier-1 centers to increase the total rate of data transfer from CERN. The number of participating Tier-2 sites increased to 23.

41

- October 18, 2006: The PhEDEx transfer system held a lock in the Oracle database which blocked other agents from continuing with transfers. This problem appeared more frequently in the latter half of the challenge when the load was higher.

- October 20, 2006: The reconstruction rate was increased at the Tier-0 to improve the output from CERN and to better exercise the prompt reconstruction farm. The data rate from CERN approximately doubles. An average rate over an hour of 600MB/s from CERN was achieved.

- October 25, 2006: The transfer rate from CERN was large with daily average rates of 250MB/s-300MB/s. The first observation of transfer backlogs begin to appear.

- October 30, 2006: Data reconstruction at the Tier-0 stopped.

- October 31, 2006: PIC and ASGC finished transferring the assigned prompt reconstruction data from CERN.

- November 2, 2006: FNAL and IN2P3 also completed the transfers.

- November 3, 2006: RAL completed the transfers. The first of the Tier-1 to any Tier-2 transfer validation began. The test involved sending a small sample from a Tier-1 site to a validated Tier-2, in the test case DESY, and then sending a small sample to all Tier-2 sites.

- November 5, 2006: CNAF completed the Tier-0 transfers

- November 6, 2006: The Tier-1 to Tier-2 transfer testing continued.

- November 9, 2006: GridKa completed the Tier-0 transfers

12.1.1 Transfers to Tier-1 Centers

During CSA06 the Tier-1 centers met the transfer rate goals. In the first week of the challenge using minimum bias events the total volume of data out of CERN did not amount to 150MB/s unless the datasets were subscribed to multiple sites. After the reconstruction rate was increased at the Tier-0 the transfer rate easily exceeded the 150MB/s target. The 30 day and 15 day averages are shown in Table 10. For the thirty day average all sites except two exceeded the goal rate and for the final 15 days all sites easily exceed the goal. Several sites doubled and tripled the goal rate during the final two weeks of high volume transfers.

The WLCG metric for availability this year is 90% for the Tier-1 sites. If we apply this to the Tier-1 participating in CSA06 transfers we have 6 of 7 Tier-1s reaching the availability goal.

Table 10: Transfer rates during CSA06 between CERN and Tier-1 centers and the number of outage days during the active challenge activities. In the MSS column the parentheses indicates the site either had scaling issues keeping up with the total rate to tape, or transferred only a portion of the data to tape.

Site	Anticipated Rate (MB/s)	last 30 day average	last 15 day average	Outage (Days)	MSS used
ASGC	15MB/s	17MB/s	23MB/s	0	(Yes)
CNAF	25MB/s	26MB/s	37MB/s	0	(Yes)
FNAL	50MB/s	68MB/s	98MB/s	0	Yes
GridKa	25MB/s	23MB/s	28MB/s	3	No
IN2P3	25MB/s	23MB/s	34MB/s	1	Yes
PIC	10MB/s	22MB/s	33MB/s	0	No
RAL	10MB/s	23MB/s	33MB/s	2	Yes

The rate of data transferred averaged over 24 hours and the volume of data transferred in 24 hours are shown in Figures 35 and 36. The start of the transfers during the first week is visible on the left side of the plot as well as the transfers not reaching the target rate shown as a horizontal red bar. The twin peaks in excess of 300MB/s and 25TB of data moved correspond to the over-subscription of data. The bottom of the graph has indicators of the approximate Tier-0 reconstruction rate. Both the rate and the volume figures show clearly the point when the Tier-0 trigger rate was doubled to 100Hz. The daily average exceeded 350MB/s with more than 30TB moved. The hourly averages from CERN peaked at more than 650MB/s.

The transferrable volume plot shown in Figure 37 is an indicator of how well the sites are keeping up with the volume of data from the Tier-0 reconstruction farm. During the first three weeks of the challenge almost no backlog of files is accumulated by the Tier-1 centers. A hardware failure at IN2P3 resulted is a small accumulation. The
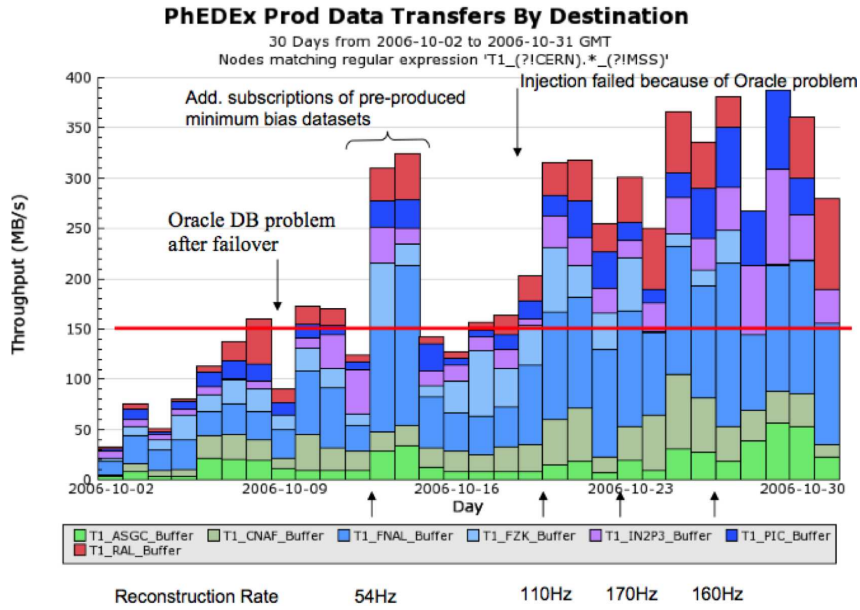
Figure 35: The rate of data transferred between the Tier-0 to the Tier-1 centers in MB per second.
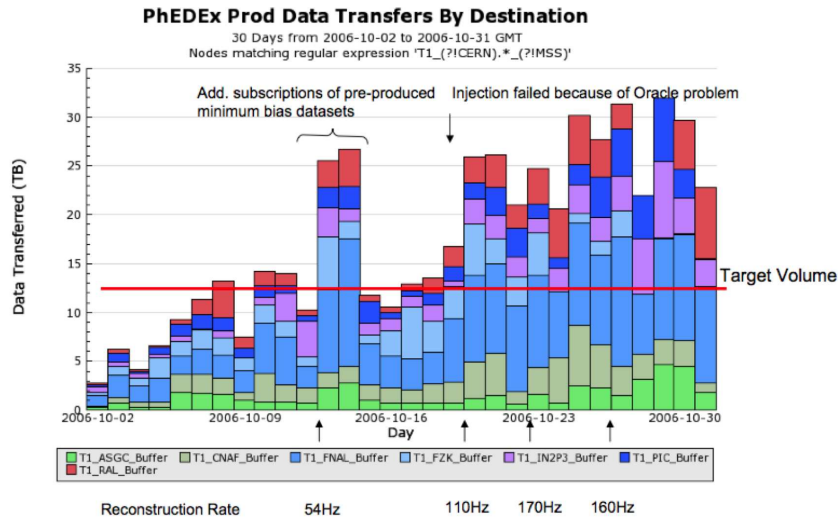

Figure 36: The total volume of data transferred between the Tier-0 to the Tier-1 centers in TB per day.

additional data subscriptions leads to a spike in data to transfer, but is quickly cleared by the Tier-1 sites. The most significant volumes of data waiting for transfer come at the end of the challenge. During this time GridKa has performed a dCache storage upgrade that resulted in a large accumulation of data to transfer. CNAF suffered a file server problem that reduced the amount of available hardware. Additionally RAL turned off the import system for two days over a weekend to demonstrate the ability to recover from a service interruption. The Tier-1 issues combined with PhEDEx database connection interruptions under the heavy load of the final week of transfers to accumulate a backlog of approximately 50TB over the final days of the heavy challenge transfers. During this time CERN continued to serve data at 350MB/s on average.
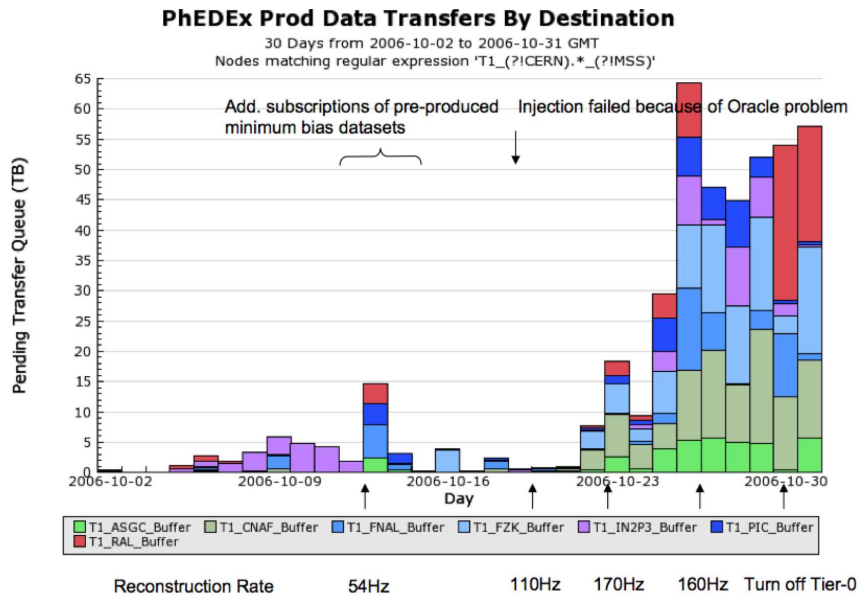
Figure 37: The total volume of data waiting for transfer between the Tier-0 to the Tier-1 centers in TB per day.

The CERN to Tier-1 transfer quality is shown in Figure 38. In CMS the transfer quality is defined as the number of times a transfer has to be attempted before it successfully completes. The link between two sites with 100% transfer quality would have had to attempt each transfer once, while a 10% transfer quality would indicate each transfer had to be attempted ten times to successfully complete. Most transfers eventually complete, having low transfer quality uses the transfer resources inefficiency and usually results in a low utilization of the network.
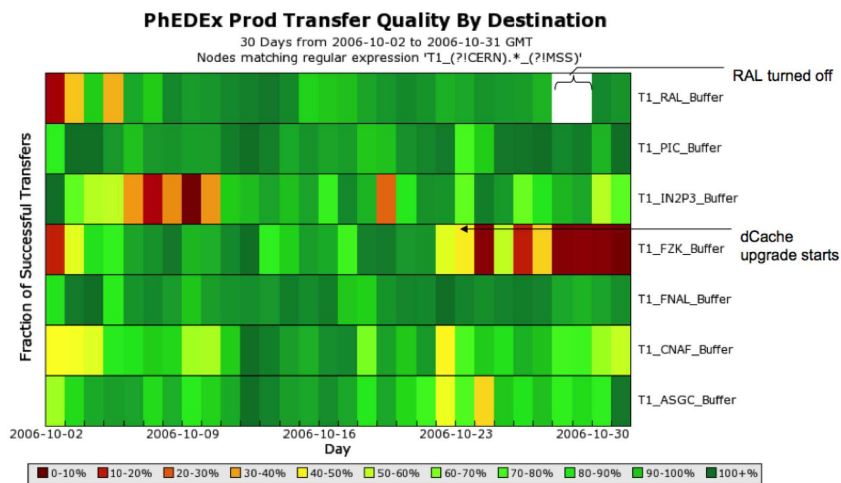

Figure 38: Transfer quality between CERN and Tier-1 centers over 30 days

The transfer quality plot compares very favorably to equivalent plots made during the spring. The CERN Castor2 storage element performed very stably throughout the challenge. There were two small configuration issues that were very promptly addressed by the experts. The Tier-1s also performed well throughout the challenge with

44

several 24 hour periods to specific Tier-1s with no transfer errors. The stability of the RAL SE before the transition to CASTOR2 can be seen at the left side of the plot, as well as the intentional downtime to demonstrate recovery on the right side of the plot. The IN2P3 hardware problems are visible during the first week and the GridKa dCache upgrade is clearly visible during the last week. Most of the other periods are solidly green. Both FNAL and PIC are above 70% efficient for every day of the challenge activities.

Tier-1 to Tier-1 transfers were considered to be beyond the scope of CSA06, though the dataflow exists in the CMS computing model. During CSA06 we had an opportunity to test Tier-1 to Tier-1 transfers while recovering from backlogs of data when the samples were subscribed to mulitple sites. PhEDEx is designed to take the data from source site where it can be efficiently transferred from. Figure 39 shows the total Tier-1 to Tier-1 transfers during CSA06. With 7 Tier-1s there are 84 permutations of Tier-1 to Tier-1 transfers, counting each direction separately. During CSA06 we successfully exercised about half of them.
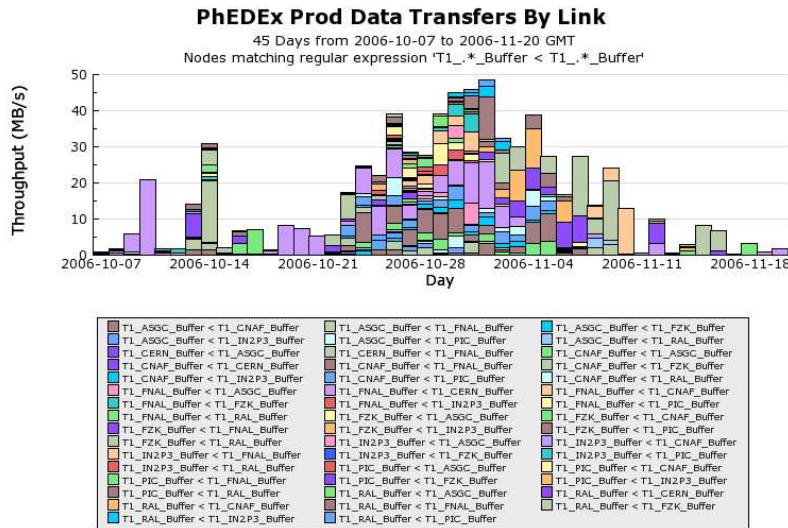

Figure 39: Transfer rate between Tier-1 centers during CSA06

12.1.2 Transfers to Tier-2 Centers

In the CMS computing model the Tier-2s are expected to be able to receive data from any Tier-1 site. In order to simplify CSA06 operations we began by concentrating on transfers from the "Associated" Tier-1 sites, and in the final two weeks of the challenge began a concerted effort on transfers from any Tier-1. The associated Tier-1 center is the center operating the File Transfer Service (FTS) server and hosting the channels for Tier-2 transfers.

The Tier-2 transfer metrics involved both participation and performance. For CSA06 CMS had 27 sites that signed up to participate in the challenge. Participation was defined as having successful transfers 80% of the days during the challenge. By this metric there were 21 sites that succeeded in participating in the challenge, which is above the goal of 20.

The Tier-2 transfer performance goals were 20MB/s and the threshold was 5MB/s. In the CMS computing model the Tier-2 transfers are expected to occur in bursts. Data will be transferred to refresh a Tier-2 cache, and then will be analyzed locally. The Tier-2 sites were not expected to hit the goal transfer rates continuously throughout the challenge. There were 12 sites that successfully averaged above the goal rate for at least one 24 hour period, and an additional 8 sites that rated averaged the threshold rate for at least one 24 hour period.

The transfer rate over the 30 most active transfer days is shown in Figure 40. The aggregate rate from Tier-1 to Tier-2 centers was not as high as the total rate from CERN, which is not an accurate reflection of the transfers expected from the CMS computing model. In the CMS computing model there is more data exported from the Tier-1s to the Tier-2s than total raw data coming from CERN because data is sent to multiple Tier-2s and the Tier-2s may flush data from the cache and reload at a later time. In CSA06 the Tier-2 centers were subscribed to specific samples at the beginning and then specific skims when available.

The ability of the Tier-1 centers to export data was successfully demonstrated during the challenge, but several sites indicated interference between receiving and exporting data. The quality of the Tier-1 to Tier-2 data transfers

PhEDEx Prod Data Transfers By Destination
30 Days from 2006-10-02 to 2006-10-31 GMT
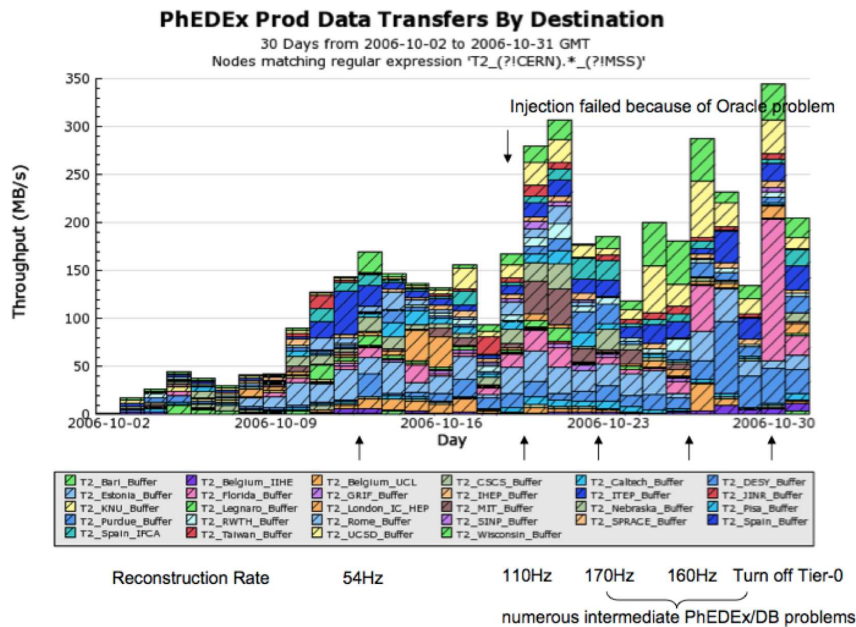Nodes matching regular expression 'T2_(?!CERN).*_(?!MSS)'

Figure 40: Transfer rate between Tier-1 and Tier-2 centers during the first 30 days of CSA06

is shown in Figure 41. The quality is not nearly as consistently green as the CERN to Tier-1 plots, but the variation has a number of causes. Not all of the Tier-1 centers are currently exporting data as efficiently as CERN, especially in the presence of a high load of data ingests, in addition most of the Tier-2 sites do not have as much operational experience receiving data as the Tier-1 sites do.

The Tier-1 to Tier-2 transfer quality looks very similar to the CERN to Tier-1 transfer quality of 9-12 months ago. With a concerted effort the Tier-1 to Tier-2 transfers should be able to reach the quality of the current CERN to Tier-1 transfers before they are needed to move large qualities of experiment data to users.

PhEDEx Prod Transfer Quality By Destination
30 Days from 2006-10-02 to 2006-10-31 GMT
Nodes matching regular expression 'T2_(?!CERN).*_(?!MSS)'
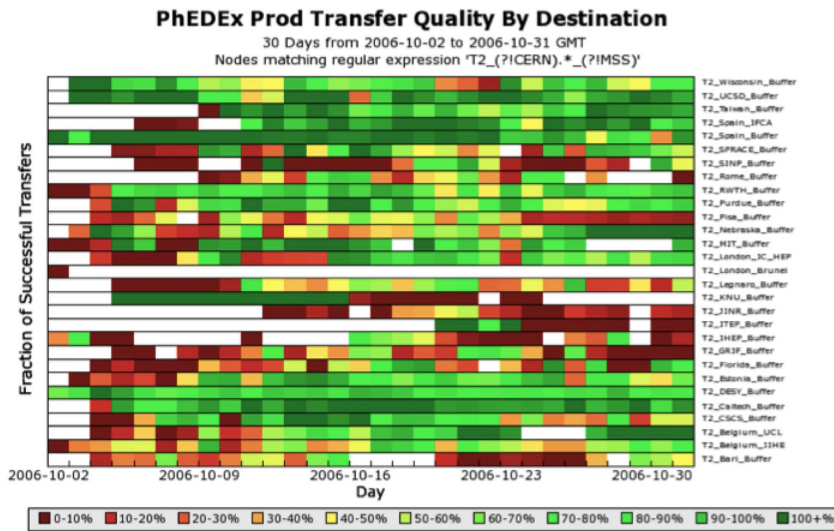
Figure 41: Transfer quality between Tier-1 and Tier-2 centers during the first 30 days of CSA06

There are a number of very positive examples of Tier-1 to Tier-2 transfers. Figure 42 shows the results of the Tier-1 to all Tier-2 tests when PIC was the source of the dataset. A small skim sample was chosen and within 24 hours 20 sites had successfully received the dataset. The transfer quality over the 24 hour period remained high with success transfers to all four continents participating in CMS.

Figure 43 is an example of the very high export rates the tier-1 centers were able to achieve transferring data to Tier-2 centers. The peak rate on the plot is over 5Gb/s, which was independently verified by the site network
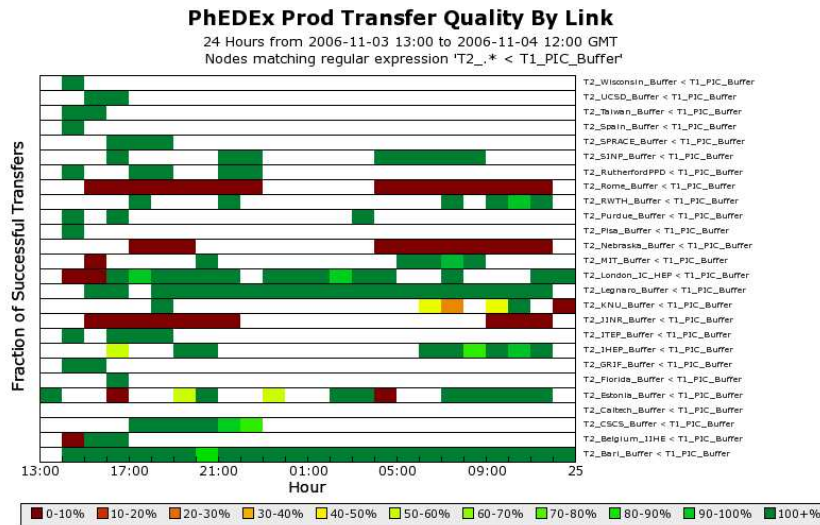
Figure 42: Transfer quality between PIC and Tier-2 sites participating in the dedicated Tier-1 to Tier-2 transfer tests

monitoring. This rate is over 50% of the anticipated Tier-1 data export rate expected in the full sized system.
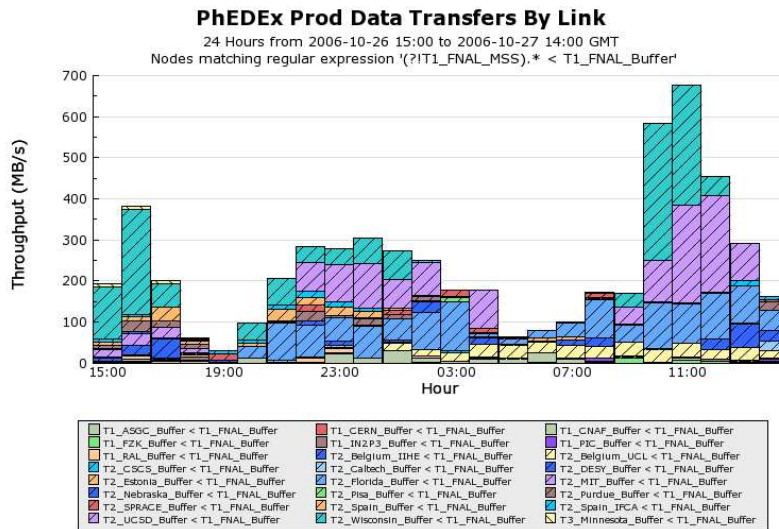

Figure 43: Transfer Performance between FNAL and Tier-2 sites participating in the dedicated Tier-1 to Tier-2 transfer tests

Figure 44 is an example of the very high rates achieved at both Tier-1 export and Tier-2 import observed in CSA06. The plot shows both the hourly average and the instantaneous rate. DESY achieved an import rate to disk of higher than 400MB/s.

12.2   Tier-1 Skim Job Production

CSA tested the workflow to reduce primary datasets to manageable sizes for analyses. Four production teams provided centralized skim job workflow at the Tier-1 centers. The produced secondary datasets are registered into Dataset Bookkeeping Service and accessed like any other data. Common skim job tools were prepared based on Monte Carlo generator information and reconstruction output, and both types were tested (see Section 5.4). There was overwhelming response from the analysis demonstrations, and about 25 filters producing nearly 60 datasets were run as compiled in Table 11. A variety of output formats for the secondary datasets were used (FEVT, RECO, AOD, AlCaReco), and the selected number of events range from $< 1\%$ to 100%. Secondary dataset sizes ranged from < 1 GB to 2.5 TB.

Table 11: List of requested skim filters to run during CSA06 by group, filter name, primary input dataset, efficiency, and input/output data formats.

| Group	Filter	Samples	Efficiencies	Input format	Output format
hg	CSA06_Tau_Zand1lFilter.cfg	EWK	14%	FEVT	RECOSim
hg	CSA06_HiggsTau_1lFilter.cfg	EWK	36%	FEVT	RECOSim
hg	CSA06_HiggsTau_1lFilter.cfg	T-Tbar	47%	FEVT	RECOSim
hg	CSA06_HiggsWW_WWFilter.cfg (bkgnd)	EWK	1%	FEVT	FEVT
hg	CSA06_HiggsWW_WWFilter.cfg (signal)	EWK	1%	FEVT	FEVT
hg	CSA06_HiggsWW_TTb_Filter.cfg	T-Tbar	4%	FEVT	FEVT
hg	CSA06_Higgs_mc2l_Filter.cfg	EWK	10%	FEVT	RECOSim
hg	CSA06_Higgs_mc2l_Filter.cfg	Jets	2%	FEVT	RECOSim
hg	CSA06_Higgs_mc2l_Filter.cfg	HLT(e,mu)	1%	FEVT	RECOSim
hg	CSA06_Higgs_mc2gamma_Filter.cfg	EWK	0	FEVT	RECOSim
hg	CSA06_Higgs_mc2gamma_Filter.cfg	Jets	34%	FEVT	RECOSim
hg	CSA06_Higgs_mc2gamma_Filter.cfg	HLT(gam)	0.4%	FEVT	RECOSim
hg	CSA06_Higgs_mc2l_Filter.cfg	TTbar	14%	FEVT	RECOSim
hg	CSA06_Higgs_mc2gamma_Filter.cfg	TTbar	8%	FEVT	RECOSim
sm	CSA06_TTbar_1lFilters.cfg (skim1efilter)	T-Tbar	20%	FEVT	RECOSim
sm	CSA06_TTbar_1lFilters.cfg (skim1mufilter)	T-Tbar	20%	FEVT	RECOSim
sm	CSA06_TTbar_1lFilters.cfg (skim1taufilter)	T-Tbar	20%	FEVT	RECOSim
sm	CSA06_TTbar_dilepton.cfg	T-Tbar	~10%	FEVT	RECOSim
sm	CSA06_MinimumBiasSkim.cfg	minbias	100%	FEVT	RECOSim
sm	CSA06_UnderlyingEventJetsSkim.cfg (reco)	Jets	~100%	FEVT	RECOSim
sm	CSA06_UnderlyingEventDYSkim.cfg	EWK	~10%	FEVT	RECOSim
eg	CSA06_ZeeFilter.cfg (zeeFilter)	EWK	3%	FEVT	RECOSim
eg	CSA06_ZeeFilter.cfg (AlCaReco)	EWK	3%	FEVT	AlcaReco
eg	CSA06_AntiZmmFilter.cfg	Jets	85%	FEVT	FEVT
mu	CSA06_JPsi_mumuFilter.cfg	SoftMuon	50%	FEVT	FEVT
mu	CSA06_JPsi_mumuFilter.cfg	Zmumu	50%	FEVT	FEVT
mu	CSA06_JPsi_mumuFilter.cfg	EWK	10%	FEVT	FEVT
mu	CSA06_WmunuFilter.cfg (reco)	EWK	20%	FEVT	AODSim
mu	CSA06_WmunuFilter.cfg (reco)	SoftMuon	60%	FEVT	AODSim
mu	CSA06_ZmmFilter.cfg	Zmumu	50%	FEVT	RECOSim
mu	CSA06_ZmmFilter.cfg	Jets	?	FEVT	FEVT
mu	recoDiMuonExample.cfg (reco)	EWK	20%	FEVT	RECOSim
mu	recoDiMuonExample.cfg (reco)	Zmumu	67%	FEVT	RECOSim
su	CSA06_Exotics_LM1Filter.cfg	Exotics	39%	FEVT	FEVT
su	CSA06_BSM_mc2e_Filter.cfg	Exotics	2%	FEVT	FEVT
su	CSA06_BSM_mc2e_Filter.cfg	EWK	$\sim 40\%$	FEVT	FEVT
su	CSA06_BSM_mc2e_Filter.cfg	HLT(e)	?	FEVT	FEVT
su	CSA06_Exotics_ZprimeDijetFilter.cfg	Exotics	~30%	FEVT	FEVT
su	CSA06_Exotics_QstarDijetFilter.cfg	Exotics	~20%	FEVT	FEVT
su	CSA06_Exotics_XQFilter.cfg	Exotics	22%	FEVT	FEVT
su	CSA06_Exotics_ZprimeFilter.cfg	Exotics	39%	FEVT	FEVT
su	CSA06_Exotics_LM1_3IC5Jet30Filter.cfg (reco)	Exotics	25%	FEVT	FEVT
su	CSA06_TTbar_2IC5Jet100ExoFilter.cfg (reco)	T-Tbar	5%	FEVT	FEVT
jm	CSA06_QCD_Skim.cfg (21 samples)	Jets	100%	FEVT	FEVT
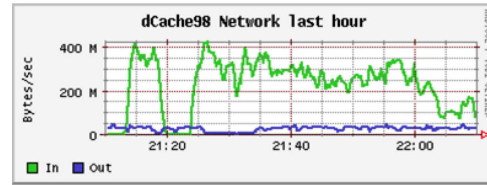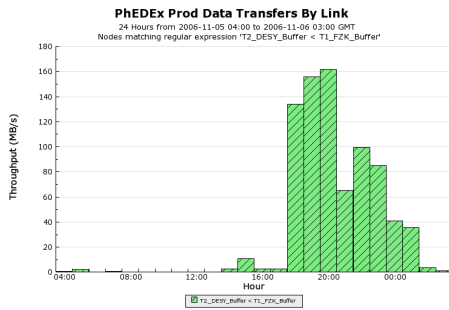
Figure 44: The plot on the left is the hourly average transfer rate between GridKa and DESY. The plot on the right is the instantaneous rate between the two sites measured with Ganglia.

12.3 Tier-1 Re-Reconstruction

The goal was to demonstrate re-reconstruction at a Tier-1 centre on files first reconstructed and distributed by the Tier-0 centre, including access and application of new constants from the offline DB. Four teams were set up to demonstrate re-reconstruction on at least 100K events at each of the Tier-1 centres.

12.3.1 Baseline Approach

Since re-reconstruction had not been tested before the start of CSA06, a technical problem was encountered with a couple of reconstruction modules when re-reconstructed was first attempted November 4. The issue has to do with multiple reconstruction products stored in the Event, and the proper mechanism of accessing them. Once diagnosed the Tier-1 re-reconstruction workflow dropped pixel tracking and vertexing out of about 100 reconstruction modules, and the processing worked correctly. Re-reconstruction was demonstrated on >100K events at 6 Tier-1 centres. For the Tracker and ECAL calibration exercises (see Section 13.1), new constants inserted into the offline DB were used for the re-reconstruction, and the resulting datasets were published and accessible to CRAB jobs. Thus, CSA06 also demonstrated the full reprocessing workflow.

12.3.2 Two-Step Approach

While the reconstruction issue described above was being diagnosed, a brute-force two-step procedure was conducted in parallel to ensure re-reconstruction at a Tier-1 centre. The approach consisted of first skimming off the original Tier-0 reconstruction products in analogy with the physics skim job workflow described in Section 12.2, and then run reconstruction on the skimmed events (i.e. two ProdAgent workflows). This approach was also successfully demonstrated at the FNAL Tier-1 centre.

12.4 Job Execution at Tier-1 and Tier-2

12.4.1 Job Robot

The processing metrics in CSA06 as they were defined foresaw that sites offering computing capacity to CMS and participating in CSA06 were expected to complete an aggregate of 50k jobs per day. The goal was to exercise the job submission infrastructure and to monitor the input/output rate.

- About 10k per day were intended as skimming and reconstruction jobs at the Tier-1 centers

- About 40k per day were expected to be a combination of user submitted analysis jobs and robot submitted analysis-like jobs

The job robots are automated expert systems to simulate user analysis tasks using the CMS Remote Analysis Builder (CRAB). Therefore they provide a reasonable method to generate load on the system by running analysis on all datasamples at all sites individually. They consist of a component/agent based structure which enables parallel execution. Job distribution to CMS compute resources is accomplished by using Condor-G direct submission on the OSG sites and gLite bulk submission on the EGEE sites.

The job preparation phase comprises four distinct steps

- Job creation

    - Data discovery using DBS/DLS
    - Job splitting according to user requirements
    - Preparation of job dependent files (incl. the jdl)

- Job submission

    - Check if there any compatible resources in the Grid Information System known to the submission system
    - Submit job to the Grid submission component (Resource Broker or Condor-G) through the CMS book-keeping component (BOSS)

- Job status check

- Job output retrieval

    - Retrieve job output from the sandbox located on the Resource Broker (EGEE sites) or the common filesystem (OSG sites)

The job robot executes all four steps of the above described workflow on a large scale.

Apart from job submission the monitoring of the job execution over the entire chain of all steps involved plays an important role. CMS has chosen to use a product called Dashboard, a development that is part of the CMS Integration Program. It is a joint effort of LCG's ARDA project and the MonAlisa team in close collaboration with the CMS developers working on job submission tools for production and analysis. The objective of the Dashboard is to provide a complete view of the CMS activity independently of the Grid flavour (i.e. OSG vs. EGEE). The Dashboard maintains and displays the quantitative characteristics of the usage pattern by including CMS-specific information and it reports problems of various nature.

The monitoring information used in CSA06 is available via a web interface and includes the following categories

- Quantities - how many jobs are running, pending, successfully completed, failed, per user, per site, per input data collection, and the distribution of these quantities over time

- Usage of the resources (CPU, memory consumption, I/O rates), and distribution over time with aggregation on different levels

- Distribution of resources between different application areas (i.e. analysis vs. production), different analysis groups and individual users

- Grid behaviour - success rate, failure reasons as a function of time, site and data collection

- CMS application behaviour

- Distribution of data samples over sites and analysis groups

Timeline:

- October 15, 2006: The job robots have started analysis submission. 10k jobs were submitted by two robot instances, with 90% of them going to OSG sites using Codor-G direct submission and 10% going through the traditional LCG Resource Broker (RB) to EGEE sites. In preparation of moving to the gLite RB, thereby improving the submission rate to EGEE sites, bulk submission was integrated into CRAB and is currently being tested.

- October 17, 2006: Job robot submissions continue at a larger scale. There was an issue found with the bulk submission feature used at EGEE sites leaving jobs hanging indefinitely. The explanation was parsing of file names in the RB input sandbox failed for file name lengths of exactly 110 characters. The problem, located in the gLite User Interface (UI), was solved by rebuilding the UI code to include a new version of the libtar

library. A new version of the UI was made available to the job robot operations team within a day.

A total of 20k jobs were submitted in the past 24 hours. A large number of jobs seemed not to report all the site information to the Dashboard, which results into a major fraction marked as "unknown" in the report. The effect needs to be understood.

Apart from the jobs being affected by the problem mentioned above the efficiency regarding successfully completed jobs is very high.

- October 19, 2006: Robotic job submission via both the Condor-G direct submission and the gLite RB bulk submission is activated. The job completion efficiency remains very high for some sites. Over the course of the past day nearly 2000 jobs were completed at Caltech with only 5 failures.

- October 20, 2006: The number of "unknown" jobs is decreasing following further investigations by the robot operations team. The job completion efficiency remains high though the total number of submissions is lower than in the previous days. A large number of sites running the PBS batch system have taken their resources off the Grid because of a critical security vulnerability. Sites applied a respective patch at short notice and were back to normal operation within a day or two.

- October 23, 2006: Over the weekend significant scaling issues were encountered in the robot. Those were mainly associated with the mySQL server holding the BOSS DB. On the gLite submission side a problem was found with projects comprising more than 2000 jobs. A limit was introduced with the consequence that the same data files are more often accessed.

- October 24, 2006: There were again scaling problems observed in the job robots. Switching to a central mySQL data base for both the robots has lead to the databases developing a lock state. Though the locks automatically clear within 10 to 30 minutes the effect has an impact on the overall job submissions rate. To resolve the issue two data bases were created, one for each robot. While the Condor-G side performs well the gLite robot continues to develop locking. A memory leak leading to robot crashes was observed in CRAB/BOSS submission through gLite. The robot operations team is working with the BOSS developers on a solution.

- October 25, 2006: The BOSS developers have analyzed the problem yesterday reported as a "scaling issue" and found that an SQL statement issued by CRAB was incomplete, leading to long table rows being accessed resulting in a heavy load on the data base server. The CRAB developers have made a new release available the same day and the robot operations team found that the robots are running fine since.

- October 26, 2006: Following the decision to move from analysis of data that has been produced with CMSSW_1_0_3 to more recent data that was produced with CMSSW_1_0_5 a lot of sites were not selected and therefore not participating since they are still lacking respective datasets.

- November 1, 2006: The submission rate reached by the job robots is currently at about 25k jobs per day. To improve scaling up to the desired rate 11 robots were set up and are currently submitting to OSG and EGEE sites.

- November 2, 2006: The total number of jobs was in the order of 21k. Due to more sites having datasets published in DBS/DLS that were created with CMSSW_1_0_5 the number of participating sites has increased. The total application and Grid efficiency is both over 99

- November 6, 2006: The number of submitted and completed jobs is still increasing. 30k jobs have successfully passed all steps in the past 24 hours. 24 Tier-2 sites are now publishing data and are accepting jobs from the robot. The efficiency remains high.

- November 7, 2006: The combined job robot, production and analysis submissions exceeded the goal of 55k per day. The activity breakdown is shown in Figure 45.

  The job robot submissions by site are shown in Figure 46. Six out of seven Tier-1 centers are included in the job robot. As expected the Tier-2 centers are still dominating the submissions. The addition of the Tier-1 centers has driven the job robot submission rates past the load that can be sustained by a single mySQL job monitor.
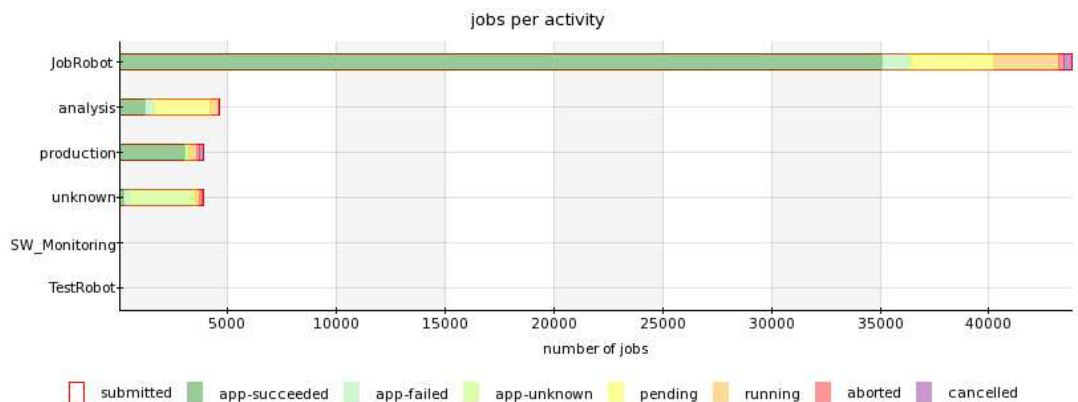
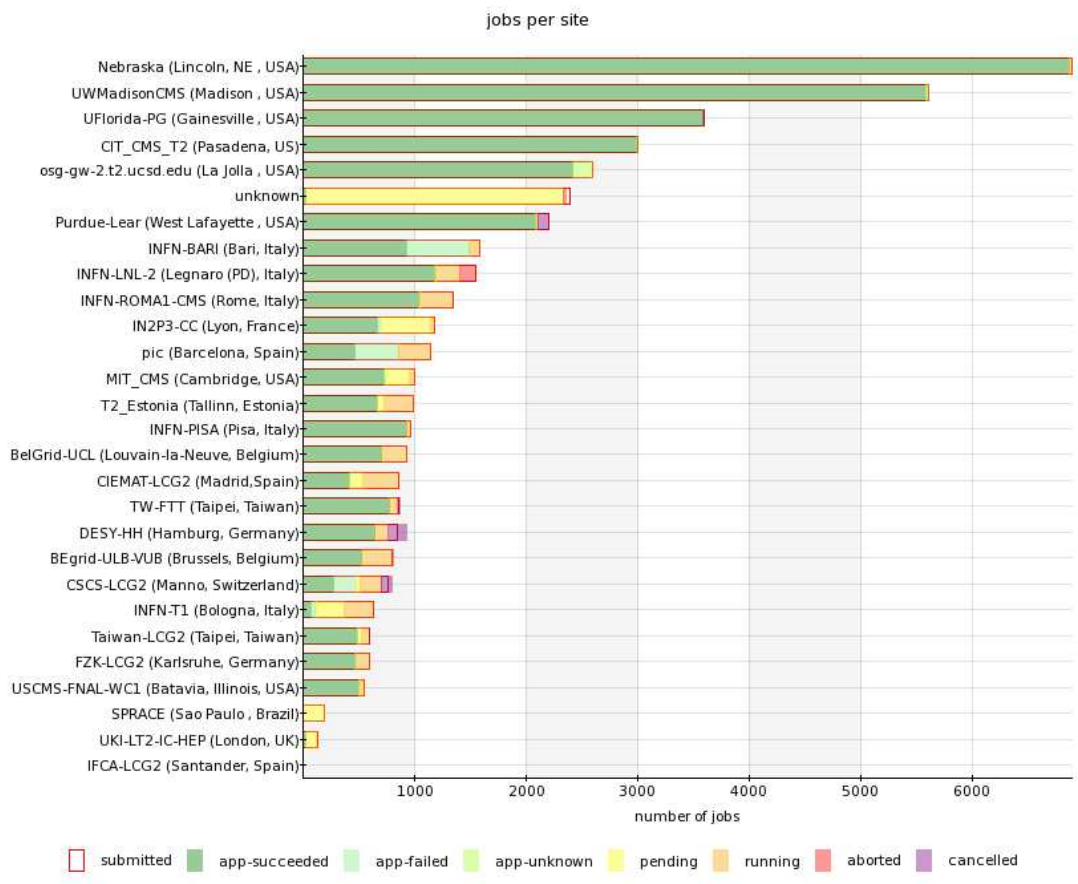Figure 45: Dashboard view of job breakdown by activity (7 Nov - 8 Nov)

Figure 46: Dashboard view of job breakdown by site (7 Nov - 8 Nov)

13 Analysis Demonstrations

A wide variety of physics analysis demonstrations were prepared by the physics groups for CSA06 in order to test the analysis workflow for CSA06. The analyses conducted on the CSA-produced data samples numbered approximately 30, with nearly 70 active participants. These demonstrations proved to be useful training exercises for collaborators in the new software and computing tools as well. The list of specialized calibration and alignment streams produced at the Tier-0 and analyzed is discussed in Section 5.7. The list of general physics analysis skims produced at the Tier-1 centres is covered in Table 11. Table 12 lists the physics analyses that were conducted as part of CSA06. Details on these analyses are contained in the following subsections.

13.1 Calibration

13.1.1 ECAL - single electron calibration analysis

This calibration exercise is performed using the CMSSW_1_0_6 release. It is based on the comparison between the energy reconstructed in a 5 by 5 matrix centered around the maximum containment crystals in the ECAL and the momentum reconstructed in the Tracker. The so-called "L3 iterative algorithm" is applied to derive the constants.

The workflow for the single electron calibration can be summarized in the following steps:

1. AlCaReco Production as detailed in Section 5.7

2. Calibration: using as input the AlCaReco objects previously produced, the calibration is performed using an iterative algorithm. To get reliable calibration constants, a basic selection which aims to reduce events that radiated a lot of energy through bremsstrahlung is applied. To verify the complete functionality of the algorithm, two types of exercises are performed. In the first exercise, the initial RecHits are perfectly calibrated. In the second exercise the initial RecHits are miscalibrated using a 4% gaussian miscalibration around one. The calibration constants obtained are written out in an XML file.

3. Propagation of the constants to the DataBase: the XML file containing the calibration constants is converted to an SQLite file and the constants are propagated to the DataBase.

4. Re-reconstruction: after the propagation of the calibration constants to the DataBase, the event reconstruction is performed once again and a comparison to the originally reconstructed data is performed.

To carry out this exercise, a sample of 4 million events of $W^+ \to e^+\nu$ was generated with the electron pointing into a region limited to two supermodules of the ECAL barrel. This increases the number of events per crystal vs. complete azimuthal symmetry for the same number of generated events, making the calibration more precise. The transverse momentum distribution of the AlCa Electrons is shown in Fig. 47 where the cut at 20 GeV introduced in the AlCaReco production is visible.

Since the precision reachable by this calibration technique is strongly dependent on the amount of energy lost by bremsstrahlung, a basic selection has been implemented. The following cuts were applied: $0.7 <$ s25/P< 1.4, s9/s25> 0.93. The distribution of these variables is shown in Fig. 48. An overall selection efficiency of about 50% is obtained with respect to the generated sample. The calibration precision is also dependent on the number of electrons per crystals, the distributions of the calibration constants derived for different statistics are shown in Fig. 49.

Fig. 50 shows the map of the calibration coefficients for the two supermodules. It is visible the effects of module and supermodule boundaries where the calibration constants are increased to account for the energy lost.

The effect due to the staggering of the crystals along the eta direction (as well as the effect due to module boundaries) is visible from Fig. 51 where each point is computed averaging the calibration coefficients over the phi direction.

The last step of this CSA06 exercise aims to compare re-reconstructed data to reconstructed data. The re-reconstruction is computed accessing the calibration constants obtained from the perfectly calibrated data sample via Frontier. An event-by-event comparison between re-reconstructed and reconstructed data is shown in Fig. 52, which shows the relative difference of the rechit energy between re-reconstruction and reconstruction. This distribution is in fact equivalent to the distribution of the calibration constants obtained from perfectly calibrated data.

Table 12: List of CSA06 analyses.

| Group	Analysis	People	Dataset
eg	ECAL isol. electron calib.	L.Agostino (CERN), P. Govoni (Milan), L.Malgeri (CERN), R.Ofierzynski (CERN)	AlCaReco
eg	ECAL Phi symmetry calib.	D.Futyan (IC)	minbias AlCaReco
eg	Z$\to ee$ reco.	P.Meridiani (Rome)	Zee AlCaReco
jm	HCAL Phi symmetry calib.	O.Kodolova (Moscow)	minbias AlCaReco
jm	HCAL isol. trk. calib.	M.Szleper (Northwestern), S.Petrushanko (Moscow)	minbias AlCaReco
jm	Jet calib.	R.Harris, M.Cardaci (FNAL)	Jets
tk	Partial Tracker alignment	L.Edera, F.Ronga, and O.Buchmuller (CERN), F.-P. Schilling (Karlsruhe),	Zmumu AlcaReco
tk	Misalignments effects on track reco	M.Abbrescia, G.Cuscela, N.De Filippis, G.Donvito, G.Maggi, S.My (Bari)	Zmumu skim
mu	Muon alignment	J.Fernandez, P.Martinez, F.Matorras (IFCA, Santander)	Muon AlCaReco
mu	W$\to \mu\nu$	M.Biasotto, U.Gasparini, M.Margoni, E.Torassa (Padova)	EWK, Soft Muon skims
mu	J/Psi and Z$\to \mu\mu$	J.Alcaraz, J.Hernandez, J.Caballero P.Garcia Abia (CIEMAT)	EWK, Soft Muon skims
mu	Z$\to \mu\mu$ and muon effic.	C.Liu and N.Neumeister (Purdue) M.Schmitt (Northwestern)	EWK skim
mu	Dimuon reconstruction	B.Kim (Florida)	SoftMuon, minbias
hg	Selection of Z$\to 2\tau$	K.Petridis (IC), A.Kalinowski (Warsaw)	EWK skims
hg	jet$\to \tau$ mis id	F.Blekman (IC), C.Siamitros (Brunel)	Z+jet
hg	tau tagging efficiency from Z$\to 2\tau$	S.Gennai and G.Bagliesi (Pisa), A.Goussiou and R.Vasques Sierra (FNAL)	EWK skims
hg	tau HLT with pixel trigger using Z$\to 2\tau$	D.Kotlinski and P.Trueb (PSI)	EWK skims
hg	Background to H$\to WW \to \ell\ell$	F. Stoeckli (ETH)	TTbar
hg	Z+2Jet background to qqH, H$\to$inv	J.Brooke and S.Metson (Bristol), K. Mazumdar (TIFR), S.Bansal (Panjab)	EWK skims
sm	underlying event/minbias	L.Fano and F.Ambroglini (Pisa) P.Bartalini and R.Field (Florida) F.Bechtel (Hamburg)	m.b., Jets, D-Y skims
sm	T-Tbar dilepton selection	I.Gonzalez-Caballero (IFCA, Santander), J.Cuevas Maestro (Oviedo)	TTbar skim
sm	T-Tbar inclusive	I.Gonzalez-Caballero (IFCA, Santander), J.Vizan (Oviedo) J.Heynick (Brussels) J.Vizan (Oviedo)	TTbar skims
sm	W mass	M.Malberti (Milan)	EWK skim
su	LM1 Jets + MET	M.Tytgat, M.Spiropulu (CERN)	Exotic, TTbar skims
su	Di-tau + MET	L.Houchu, D.J.Mangeol (Strasbourg)	Exotic
su	LM1 electron cleanup	F.Moortgat, L.Pape	Exotic
su	LM1 b-tagging	R.Stringer (Riverside)	Exotic
su	Dijet mass	R.Harris, M.Cardaci (FNAL)	QCD, Exotic skims
su	High energy e^+e^-	P.Vanlaer (Brussels), D.Evans (Bristol), C.Shepherd-Themistocleous (RAL)	Exotic, EWK skims
su	Z$' \to \mu^+\mu^-$	M.Tytgat, M.Spiropulu (CERN)	Exotic
su	Z$' \to \mu^+\mu^-$	A.Lanyov, S.Shmatov (Dubna)	Exotic
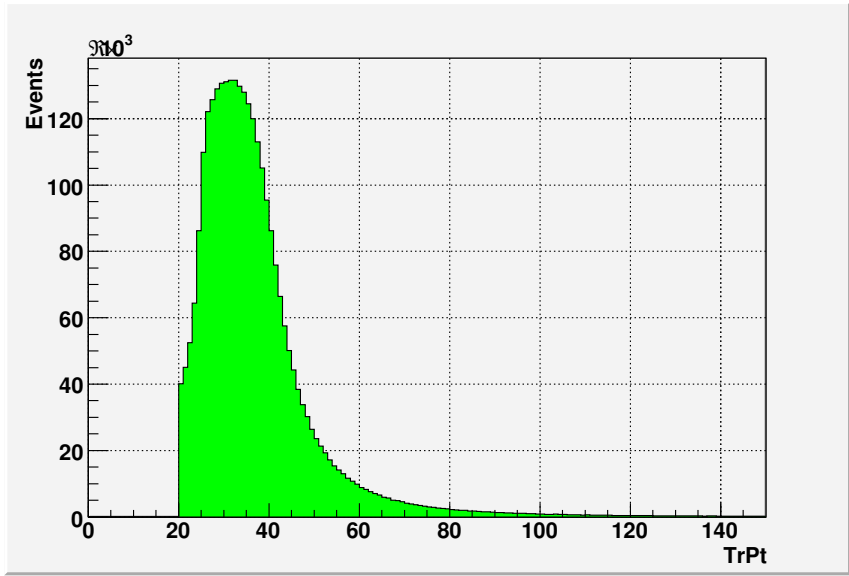
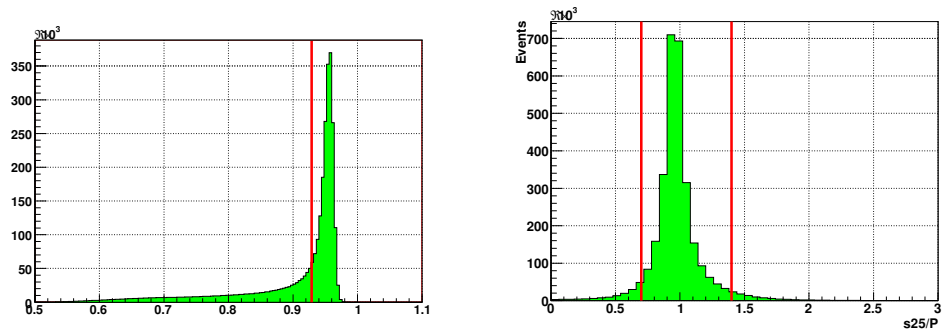Figure 47: Transverse momentum distribution of the AlCa Electrons.

Figure 48: Selection variables for the single electron calibration.
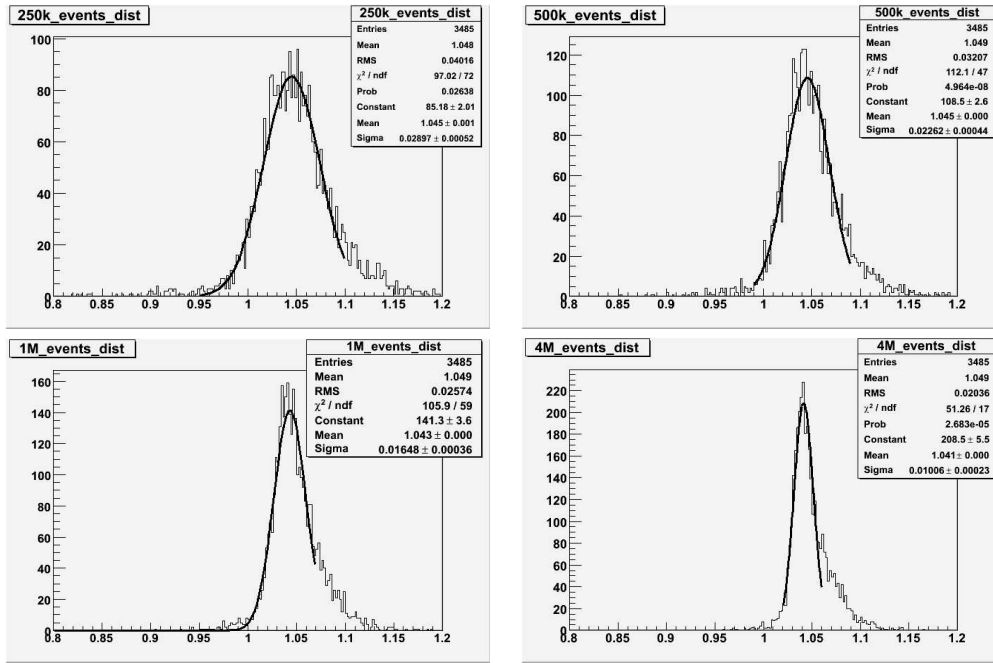
Figure 49: Distributions of the calibration constants derived from the single electron calibration exercise for different statistics. 1) 250k events; 2) 500k events; 3) 1M events; 4) 4M events. The overall shift of about 4% is expected from cluster containment.
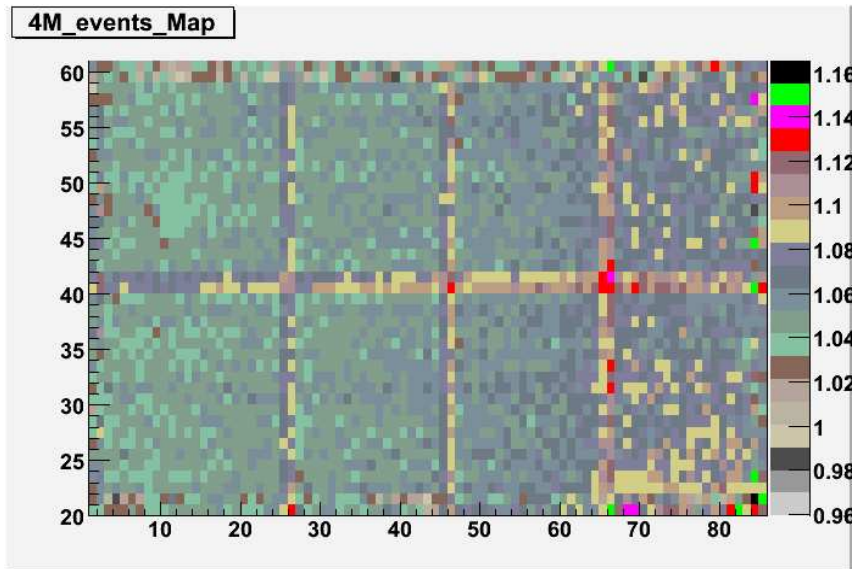

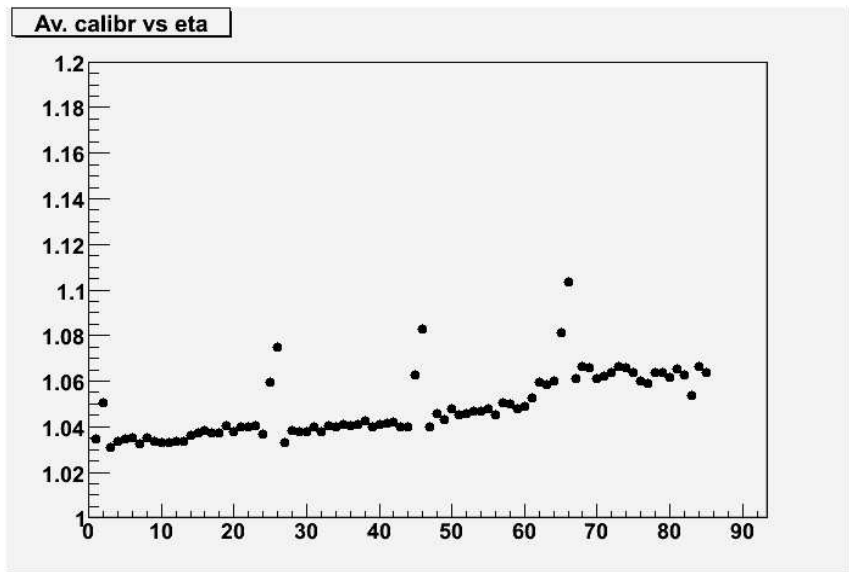Figure 50: Calibration coefficients map from the single electron calibration exercise.

Figure 51: Derived single electron calibration coefficients versus η.
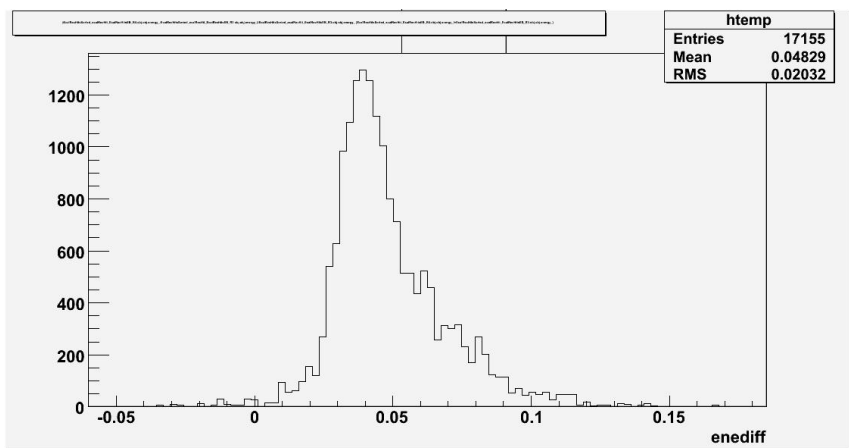
Figure 52: Comparison between reconstructed and re-reconstructed rechit energies from the single electron exercise.

13.1.2 ECAL - ϕ-symmetry calibration analysis

The symmetry of energy deposition in ϕ resulting from the detector symmetry can be exploited to perform relative intercalibration within rings of crystals at constant $|\eta|$. The technique forms part of a strategy at start-up to rapidly perform calibration of the ECAL to a precision of around 2%. Intercalibration is performed by comparing the total energy deposited in each crystal with the mean of the distribution of total energies for all crystals at the same pseudorapidity. In this way, the number of intercalibration constants is reduced from the total number of crystals ($\sim 76\,000$) to the number of pairs of fixed η rings (124).

The aim of the analysis is to use the 38 million minimum bias events generated during CSA06 to carry out ϕ symmetry calibration of the ECAL by applying the techniques detailed in [5].

Initial calibration coefficients with a 4% random spread were read in from the conditions DataBase via the Frontier caching layer. The AlCaReco data was then read from Castor at CERN and processed with the calibration algorithm. The accumulation of energy sums was performed by running 38 jobs in parallel, each processing one million events. The data processing rate was around 100 events/second using the lxbatch system at CERN.

The outputs of the parallel jobs were combined to give the total energy sums, from which the new calibration coefficients were calculated and output in xml format, then converted to SQLite format and written back into the conditions DataBase.

Figure 53 shows, as a function of η, the Gaussian width of the distribution of residual miscalibrations after correction given by:

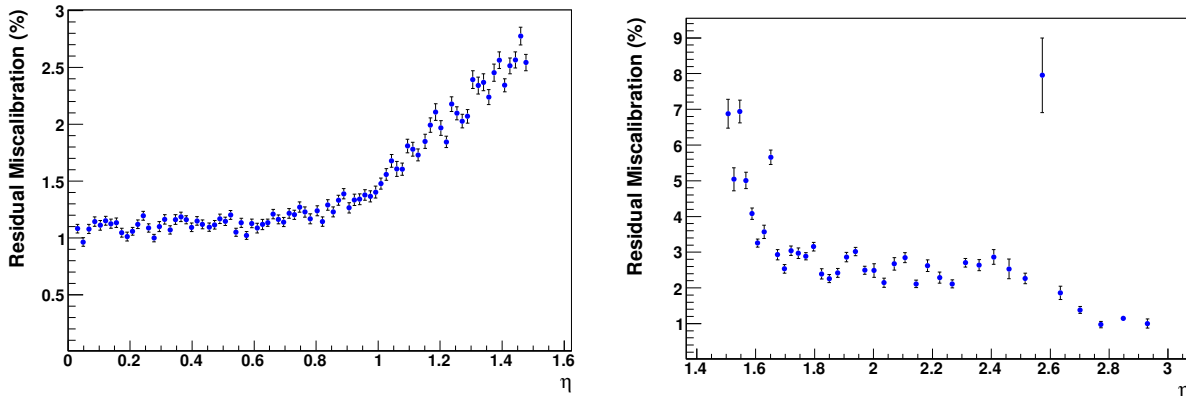residual miscalibration = initial miscalibration / measured miscalibration.

Figure 53: Residual miscalibration from the ϕ symmetry exercise after correction using 38 million minimum bias events, as a function of η, for the ECAL barrel (left) and endcaps (right). Each point correspond to an individual ring of crystals.

The residual miscalibration is limited by the variation in the amount of material in front of the ECAL as a function of ϕ. Large variations due to services are responsible for the poor result in the endcap for $\eta < 1.6$, while the anomalous points at $\eta = 1.66$ and $\eta = 2.6$ correspond to the outer and inner edges of the preshower disk, which, due to the $x - y$ crystal arrangement, overlaps crystals in the corresponding rings by differing fractions.

The results were compared to the results of the original study and found to be consistent.

13.1.3 HCAL - ϕ-symmetry calibration analysis

The sample used to provide the calibration coefficients is divided in subsets of $10\,000$ events each. The mean value ($\langle E_{ij} \rangle$ and variance ($D(i,j) = \langle E_{ij}^2 \rangle - \langle E_{ij} \rangle^2$) of the energy distribution per subset in readout (i) are calculated over the subset (j). The mean value and variance of the energy distribution in the readout are calculated as $\langle E_i \rangle = 1/N_{subsets} \times \sum_j(\langle E_{ij} \rangle)$ and $\langle D_i \rangle = 1/N_{subsets} \times \sum_j(\langle D_{ij} \rangle)$. The mean value and variance in η ring are calculated as $\langle E \rangle = 1/N_{readouts} \times \sum_i(\langle E_i \rangle)$ and $\langle D \rangle = 1/N_{readouts} \times \sum_i(\langle D_i \rangle)$. The calibration coefficients for the readout i in the ring is:

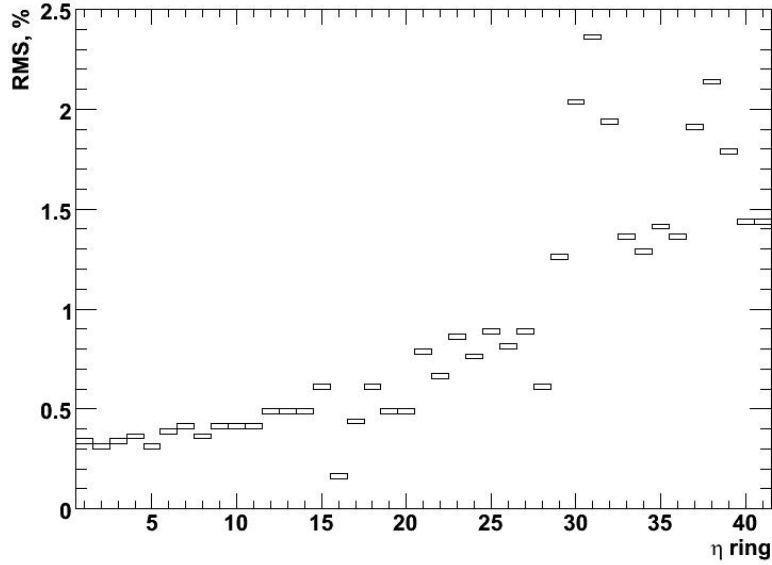$$\alpha_i = \langle E_i \rangle / \langle E \rangle \tag{1}$$

Figure 54: Accuracy of the HCAL azimuthal symmetry calibration in % for rings with Z>0 and depth is equal one or

$$\alpha_i = \sqrt{\langle D_i \rangle / \langle D \rangle} \tag{2}$$

The calibration coefficients are applied to the prompt sample and energy deposition per readout is rescaled:

$$E_i^{new} = E_i / \alpha_i \tag{3}$$

A calibration sample of $1\,000\,000$ events is chosen and calibration coefficients are calculated. The execution is organized in 100 jobs running in parallel with $10\,000$ events each. Each job takes about 600 KSI2K. Root macros are used to provide the calibration constants. The prompt sample consists also from $1\,000\,000$ events. The different values of miscalibration is chosen: 5%, 10%, 20%. Zero suppression is applied only to barrel and endcap, i.e. readouts with energy deposition with $E < 0.5$ GeV are kept in events. In the case of HF, no zero suppression is applied.

The mean values method is used to get the calibration coefficients in the barrel/endcap parts, and the variances method is used for HF.

The recalibration with $1\,000\,000$ min. bias events restores the azimuthal symmetry with a precision less then 1% for barrel and endcap and 1–2.5% for HF for the depth equal to one. It should be noted that $1\,000\,000$ events correspond to 10 seconds at a trigger rate of 100 kHz.

13.1.4 HCAL - isolated tracks calibration analysis

The CSA06 minimum bias sample was used to carry out this analysis. The average number of tracks per event before selection was 5.3 and their average momentum was 3.4 GeV as shown in Figure 56.

Figure 57 shows the number of selected isolated tracks per event after selection. The sample reduction was thus 40%.

Analysis of the AlCaReco Isotrk samples has helped tune the proper track isolation criteria for future producer versions. Issues of multiple scattering in the low momentum track extrapolation to ECAL and HCAL were studied and a method to associate those tracks to the right HCAL towers was proposed. Track extrapolation to the surface of ECAL will be implemented in the future versions of the Isotrk producer. The Track Extras will thus be redundant.

CaloTowers were included in the AlCaReco Isotrk samples for simplicity and as a proof of principle. Proper calibration will require ECAL and HCAL RecHits. Thus, the future AlCaReco Isotrk stream will include:

- reduced track collection,
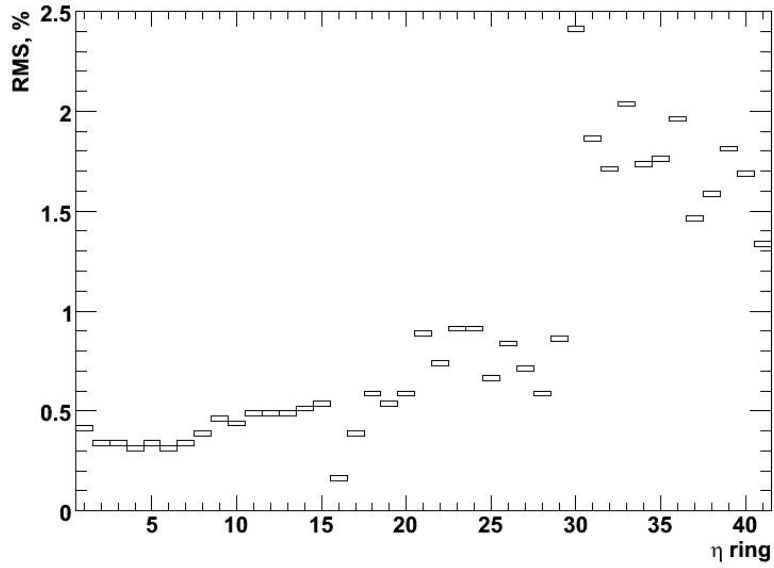- reduced ECAL and HCAL RecHits.

Figure 55: Accuracy of the azimuthal symmetry calibration in % for rings with Z<0 and depth is equal one
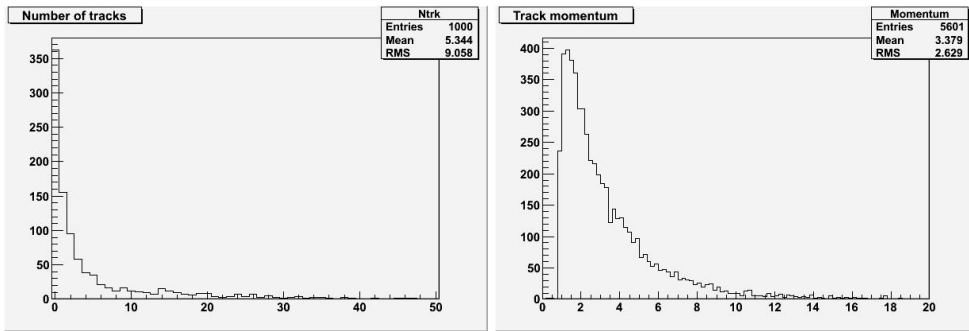

Figure 56: Average number of tracks and average momentum before selection in the HCAL isolated track calibration exercise.
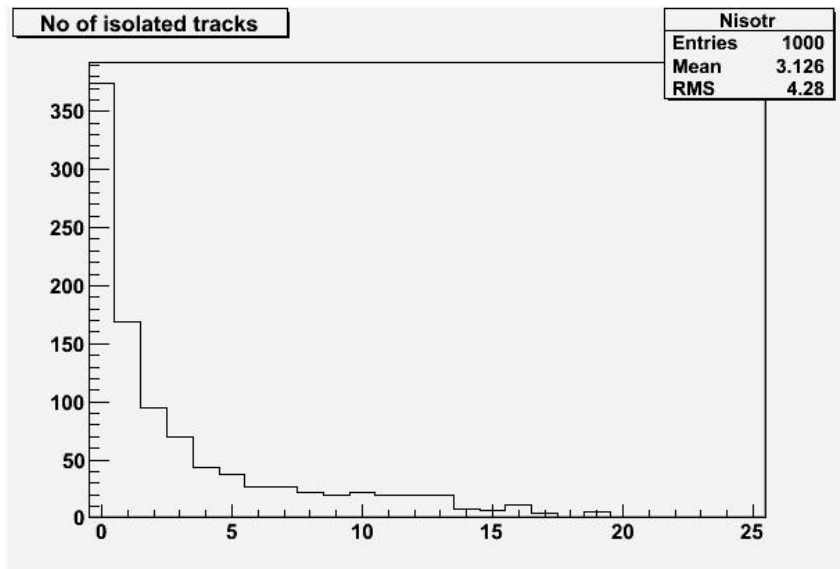

Figure 57: Distribution of number of tracks after selection in the HCAL isolated track calibration exercise.

Purity and statistics after the cut

MinBias CSA06 sample

		w/o Cut	5 GeV	3 GeV	1 GeV
Cone 0.15	Purity	88.0%	88.9%	91.6%	93.3%
	Statistics	100%	95%	83%	42%
Cone 0.2	Purity	88.0%	88.7%	91.5%	92.8%
	Statistics	100%	96%	87%	53%

Figure 58: Purity and statistics after the isolation cut against neutral hadrons

However, CaloTowers without thresholds could be in principle an option, depending on the technical aspects of the calibration procedure.

For a future producer the criteria for isolation from neutral hadrons also are needed. It can be done using the difference between ECAL energies in a big (0.5) and small (0.1–0.2) η-ϕ cone around the point of interaction of an isolated track with the ECAL calorimeter. To remove isolated tracks with neutral hadrons in a cone of 0.5, a cut (1-2 GeV) on this parameter has to be used. The purity and survived event yield after the cut for min-bias events can be found in Table 58.

13.2 Alignment

13.2.1 Tracker Alignment

A tracker alignment CSA06 exercise was carried out with the goal to demonstrate the full work- and dataflow of the alignment process. The exercise followed closely the ideas and concepts developed during the T0-RTAG [6]. The exercise comprised the following steps:

- Reading of alignment constants from the offline database during prompt reconstruction;

- Writing dedicated AlCaReco streams for alignment;

- Defining a misalignment scenario and insertion of the corresponding object into the offline database;

- Running an alignment algorithm at the Tier-0;

- Inserting the resulting alignment corrections into the database;

- Running re-reconstruction at a Tier-1 centre reading this alignment object;

- Running analysis jobs in which ideal, misaligned and aligned distributions are compared.

The steps regarding production of the dedicated AlCaReco streams as well as enabling the prompt reconstruction to read alignment constants from the offline database were already described in section 5.7. In the following, a short summary of the exercise is given. For more details, see [4].

The following misalignment scenario was defined for this exercise: The sensors of the Tracker Inner Barrel TIB as well as the Rods of the Tracker Outer Barrel TOB were misaligned. Random shifts drawing from a flat distribution in the range $\pm 100 \, \mu$m were applied in u, v, w (local coordinate system) for layers built from double sided modules,
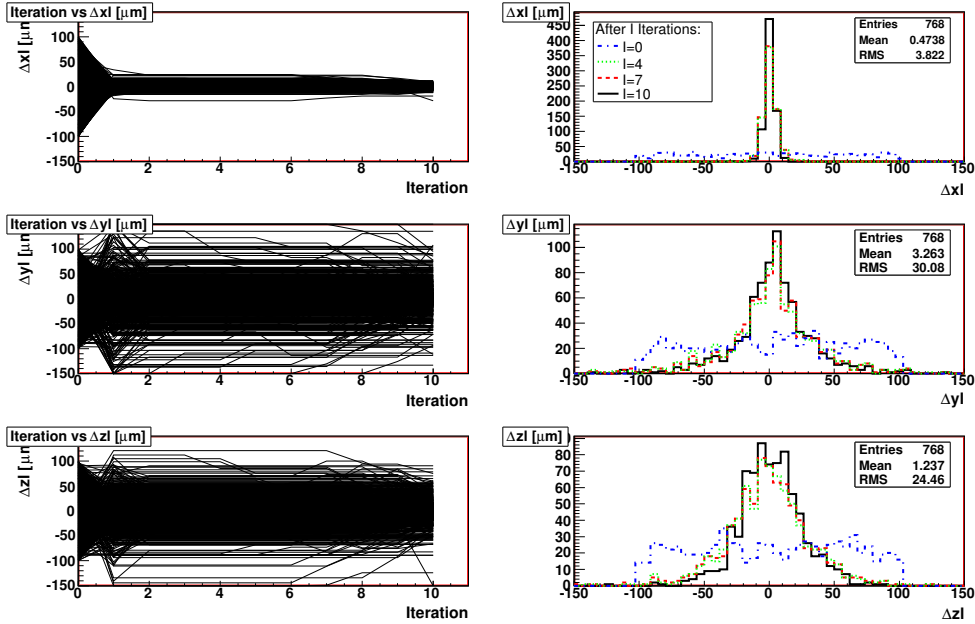
Figure 59: Alignment exercise: The convergence of the alignment algorithm in $\Delta u, \Delta v, \Delta w$ is shown for double sided TIB modules as a function of the iteration number (left), as well as projected for initial misalignment (labeled "after 0 Iterations") and after 4, 7 and 10 iterations (right).

and in u (precise coordinate) only for layers built from single sided modules. In addition, random rotations around all three local coordinate axes of size ± 10 mrad were applied to the modules/rods of both single and double sided TIB and TOB layers. The pixel detector was kept fixed in order to define a reference system. In addition, the outermost TOB layer was also kept fixed in order to improve the convergence. This misalignment scenario was inserted as an alignment object into the offline database. In addition, another object corresponding to the ideal tracker geometry was inserted, to be used during prompt reconstruction.

The alignment was performed running the HIP alignment algorithm [7] as implemented in CMSSW_1_0_6 over approximately 1 M AlCaReco $Z^0 \rightarrow \mu^+\mu^-$ events produced during prompt reconstruction at the Tier-0, reading the above mentioned misalignment object from the offline database. The algorithm was run on 20 dedicated CPUs in parallel at CERN, iterating 10 times over the data sample. The result of the alignment was obtained in less than 5 h, and the corresponding tracker alignment object was inserted into the database. Figure 59 illustrates the convergence of the alignment for the double sided TIB sensors.

Once the alignment and calibration constants were inserted in the database, they were deployed to the Tier-1/2 centres via Frontier. Subsequently, re-reconstruction of some of the CSA06 datasets was launched at various Tier-1 centres as described in Section 12.3. For instance, the $Z^0 \rightarrow \mu^+\mu^-$ data set was re-reconstructed at PIC (Barcelona) using the new alignment object. In order to demonstrate the final missing piece of the workflow, grid analysis jobs were submitted to PIC to process these re-reconstructed $Z^0 \rightarrow \mu^+\mu^-$ events. The reconstructed invariant dimuon mass is presented in Figure 60 for three cases:

- Using the ideal geometry, reading the RECO produced during prompt reconstruction;

- Using the misaligned geometry, reading the AlCaReco and the misalignment scenario database object;

- Using the realigned geometry, reading the events re-reconstructed with the alignment database object.

As can be seen, the invariant mass resolution is degraded in the case of misalignment. After the alignment algorithm has corrected the tracker geometry, the resolution is recovered close to the original value. This demonstrates that the work- and dataflow of the full alignment was successfully carried out.

13.2.2 Muon alignment

A similar example was run for the Muon system alignment. The goal of this analysis was to demonstrate the possibility of performing the Muon System standalone alignment with tracks at a remote Tier-2 (Spanish Tier-2, in
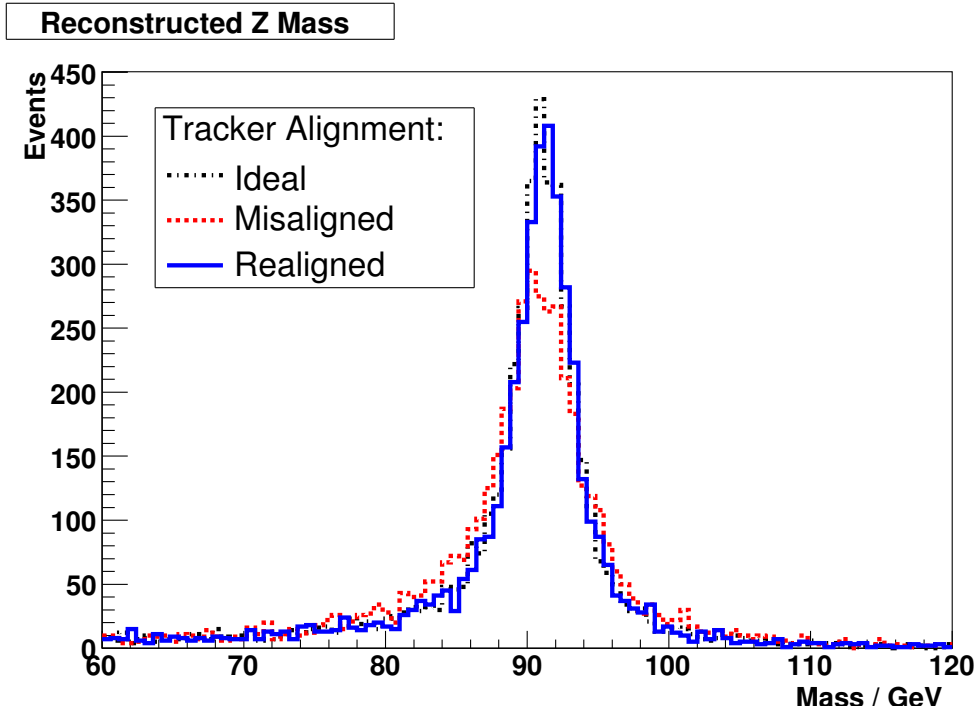
Figure 60: Invariant mass distribution from $Z^0 \to \mu^+\mu^-$ events, obtained from events produced by the prompt reconstruction at the Tier-0 ("ideal"), from events processed with misalignment as used as input for the alignment algorithm ("misaligned") and from events re-reconstructed at a Tier-1 centre (PIC) using the alignment constants derived from the alignment algorithm ("realigned").

this case). The exercise was foreseen to prove the data and workflow, but also the performance of the algorithms was tested. The exercise was split in several subtasks:

- Data import to Tier-2 and prompt analysis

- Re-reconstruction: remote access to geometry DB

- Extraction of alignment constants and production of re-aligned geometry file

- Re-Reconstruction with re-aligned geometry

The basic sample for this exercise was a 2 Million event Z, decaying to muon pairs. The full RECO sample was imported for the different versions as they were made available (4.4 TB for the last version) although many of the tasks could be performed onto the AlCaReco produced at CERN-Tier-0 (as described in Section 5.7) with a much smaller size of about 20 GB for the same 2 Million events. Other samples and formats were also used for checks, like 0.5 Million W decays to muons (AOD), ttbar (semileptonic and di-leptonic decays) or minimum bias. As soon as the sample was fully transferred, prompt analysis was performed using 90 CPUs in which these jobs were prioritized. The jobs, submitted with the standard CRAB tools, consisted in running over the whole sample checks on the data quality from basic distributions derived from standalone muon and global track quantities (transverse momentum and invariant mass, for the dimuon case). The latency was dominated by the different steps in the data transfer. For the AlCaReco sample, results were available within the next day of the sample availability at Tier-0.

Re-reconstruction of the samples was performed locally at the Tier-2, starting from the DIGI, in the RECO sample. Two different tests were done. In the first case, a distorted geometry in the Database is read from by the Tier-0 reconstruction jobs. The one defined in the so-called ShortTerm Scenario, mimicking the misalignment conditions at the very beginning of data taking, was used. The access was made with Frontier through the local SQUID cache. The obtained invariant mass in each case is shown for the Global (using tracker and muon hits) and Standalone Muon (muon hits only) reconstruction in Figure 61. A clear widening of the distribution is observed due to the wrong position of the muon chambers. The effect is less evident for the Global reconstruction, because the momentum resolution for these energies largely depends on the Tracker, which is assumed perfectly aligned in this case. Alternatively, the re-reconstruction was tested using a local DB, previously created offline, which was

supplied during the job submission, together with other input files. No significant delay was found due to the remote access to DB.
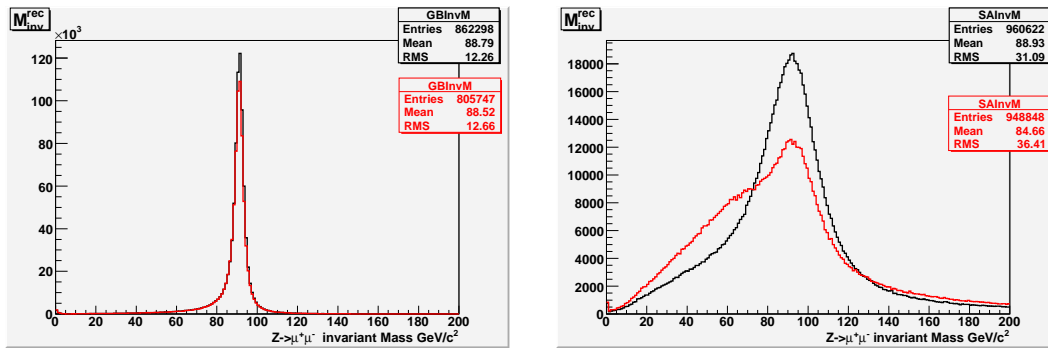

Figure 61: Reconstructed mass for Z→ $\mu\mu$ sample with Global or Standalone Muons for ideal or ShortTerm geometry.

A simplified version of the Muon alignment with tracks algorithm was applied to the sample (AlCaReco). It is based in the so-called Blobel method [8, 9], which provides a relatively simple procedure to simultaneously fit track parameters and alignment constants. The method is in principle iterative, but for this application the problem is nearly linear and only one step is needed. This has the advantage that it is enough to provide track fit residuals and local coordinates of the hit, when the fit is performed ignoring the alignment constants. No refit or iterations are needed, nor access to the geometry and, hence, it is very fast and can be run in a standalone mode if necessary. For this exercise only the barrel muon chambers are aligned in the more relevant d.o.f. displacement along the rφ direction. The resultant linear system has 240 unknowns, which is small enough to handle the problem with standard methods. However, the alignment with tracks problems without additional constraints leads to a singular matrix, which cannot be inverted. The method proposed here, looks for the degrees of freedom corresponding to the singularity characterized as eigenvectors of eigenvalue 0 and stores them. They correspond to invariants of the problem, that cannot be resolved without the introduction of additional constraints. For this exercise they were ad-hoc fixed to zero, although in a real case external constraint will be included (mechanical references, hardware alignment or other physic constraints).

This procedure was applied to the Z→ $\mu\mu$ sample reconstructed with a distorted geometry in which Barrel chambers are displaced from their nominal position alternatively by 1 mm. The alignment constants obtained in this way are stored with MuonAlignment tools into a geometry DB, which constitutes an approximation of the nominal geometry. Muon reconstruction was again performed taking as input this geometry. Figure 62 shows the invariant mass for Standalone Reconstruction in the barrel, for each of the three geometries. It can be appreciated that the alignment recovers almost totally the precision given by the ideal geometry.

13.3 Physics Analysis Exercises

13.3.1 Effect of tracker misalignment on track reconstruction performances

The alignment uncertainties of the CMS Tracker detector, made of a huge amount of independent silicon sensors with an excellent position resolution, affect the performances of the track reconstruction and track parameters measurement. The purpose of the analysis exercise performed by the team at Bari during the CSA06 was to study the effect of the CMS tracker misalignment on the performance of the track reconstruction [10]. Realistic
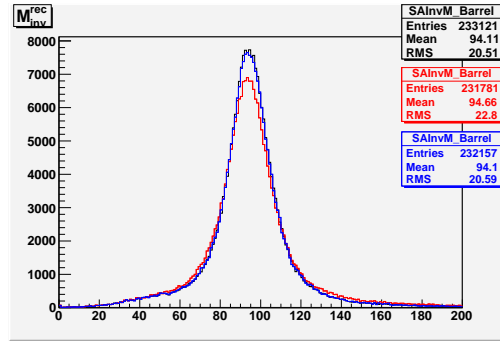
Figure 62: Reconstructed mass for Z→ $\mu\mu$ sample with Global or Standalone Muons for ideal (black), misaligned (red) and aligned (blue) geometry.

estimates for the expected displacements of the tracking systems were supplied in different scenarios as specified in the following:

- the ideal scenario with a perfect tracker geometry;

- the short term misalignment scenario representing the typical mis-alignment conditions during the first data-taking when the uncertainties on the position of the sub-structures of the CMS tracker will be between $10\,\mu$m for pixel detectors and $400\,\mu$m for microstrip silicon detectors in the endcaps. Detector position and errors are read from the offline database at CERN by caching the needed information locally via frontier/squid software.

- the long term scenario when the alignment uncertainties are expected to be a factor 10 smaller because of the improvement obtained by using alignment algorithms with a high statistics of tracks.

- the CSA06 aligned scenario by using the tracker module position and errors as obtained by the output of the alignment procedure (Section 13.2) that was run at the CERN Tier-0 to verify the efficiency of the alignment procedure on the track reconstruction. The refit of tracks is performed also in this case.

Track reconstruction is based on the Kalman Filter formalism [11] for trajectory building, cleaning and smoothing steps and uses hits from pixel detector as seeds to provide initial trajectory candidates. Because of the misalignment, the analysis requires to refit tracks with a misaligned tracker geometry. Global efficiency of track reconstruction and track parameter resolutions for muons were compared in all the cases. The association between simulated track and reconstructed tracks is performed by comparing the corresponding track parameters at the closest approach point and choosing the pair which gives the minimum χ^2 from the best fit procedure.

Events from CSA06 $Z \to \mu\mu$ sample were firstly skimmed by selecting events with Hep MC muons from Z decay with pseudorapidity, η, in the tracker acceptance, $|\eta| < 2.55$, with a transverse momentum larger than $5\,\mathrm{GeV}/c^2$ and a dimuon invariant mass in the following range around the Z peak: $50 < m_{\mu\mu}(\mathrm{GeV}/c^2) < 130$. The efficiency of this selection is between 50 and 60% mainly due to the cut on the acceptance, for a final sample of 1 million events. The output files in RECOSIM format were needed for the subsequent analysis.

Jobs executing the misalignment analysis were submitted at Bari with CRAB_1_4_0 in the LCG infrastructure. A total of about 2.5 thousands jobs (45 at most in parallel) ran with a grid efficiency of 90% and an application efficiency of 80%, by accessing detector position and errors from the offline database via frontier.

Some results of the misalignment analysis are summarized below. The global efficiency of track reconstruction of muons coming from Z decay is shown in Fig. 63 as a function of the pseudorapidity, η, in the tracker acceptance. In the case of a perfect geometry the global track reconstruction is not fully efficient over all the η range because of the track associator algorithm itself, which discards tracks with χ^2 of the fit larger than 25. The effect of misalignment is relevant observed ib the short term scenario and causes a partial inefficiency of the track reconstruction that can be recovered if the intrinsic position resolution of the tracker detector is combined with the alignment uncertainties to make larger the error on the position of the reconstructed hit (called the alignment position error, or APE) so improving the track fit at the expense of a larger rate of fake tracks.

The transverse momentum resolution as a function of the transverse momentum is reported in Fig. 64; the degradation of the transverse momentum resolution at large p_T because of the misalignment is in a factor between 2 and 3 with respect to the perfect geometry case. At low transverse momentum (less than few GeV/c the multiple scattering is the most important contribution to the resolution so the effect of misalignment is negligible.

The residual of the Z mass obtained as the invariant mass of muons coming from Z decays for the case of perfect tracker geometry and for the short-term and long term misalignment scenarios is shown in Fig. 65; the σ of the Gaussian fit of the residual distribution can be quoted as the Z mass resolution which is degraded by a factor 2 because of the tracker misalignment in the short term scenario.
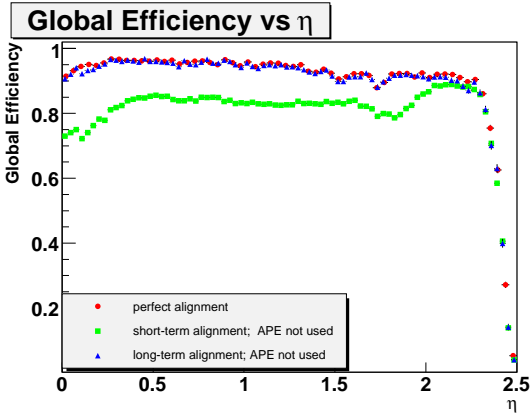

Figure 63: Global track reconstruction efficiency vs pseudorapidity for muons coming from Z decay in the case of perfect tracker geometry and in short-term and long term misalignment scenarios when the APE is not used.
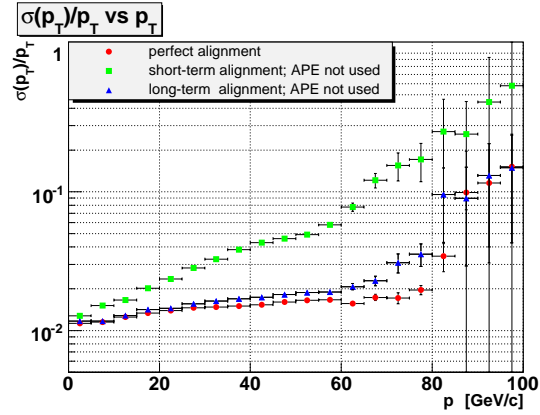
Figure 64: p_T resolution vs p_T in the case of perfect tracker geometry and in short-term and long term misalignment scenarios.

13.3.2 Selection of W$\rightarrow \mu\nu$ events

Different Monte Carlo data samples with different signal compositions were analyzed to study the $W \rightarrow \mu\nu$ preselection yield. The goal of the skimming is to reduce the disk space for the data storage and physics analysis at Tier-2 whilst preserving the signal and suppressing the backgrounds. As an example of a simple analysis performed at Tier-2 on skimmed data, the efficiency and resolution of the muon reconstruction obtained with the CMSSW software were studied as a function of p_T and η, the $W \rightarrow \mu\nu$ signal was selected with isolation cuts and the transverse mass was compared with the same distribution obtained from the background. The following data samples were considered:

- the electroweak soup (3.4M events, 50% $W \rightarrow \mu\nu$ and 50% DY)

- the soft muons (1.8M events, 50% minimum bias and 50% J/Ψ, $p_T > 4$ GeV)

The skim jobs were running at the PIC and CNAF Tier-1 over FEVT (full events). The output was sent at the LNL Tier-2 in AODSIM format (Analysis Object format). For the $W \rightarrow \mu\nu$ channel the skim consists of the selection of events having at least one muon (StandAlone muon reconstructed object) with $p_T > 10$ GeV. Table 13 summarizes
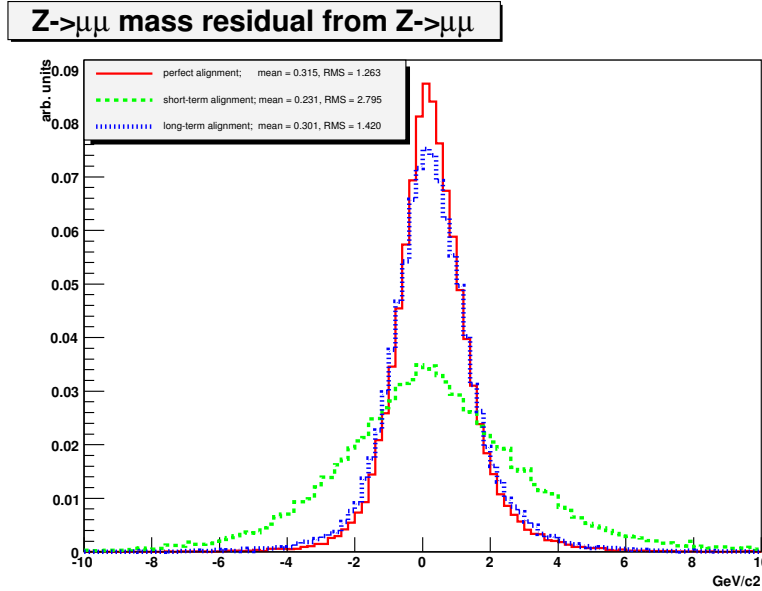
Figure 65: Residual of Z mass obtained as the invariant mass of muons coming from Z decayd for the case of perfect tracker geometry and for the short-term and long term misalignment scenarios.

Table 13: Data reduction due to the skimming in term of number of events and disk space

	Input	Output
EWK soup	FEVT , 3.4 Mevts , 5.1 TB	AODSIM , 0.55 Mevts , 0.82 GB
Soft muons	FEVT , 1.8 Mevts , 2.7 TB	AODSIM , 0.94 Mevts , 1.4 GB

the reduction of the disk space for the data storage due to the skimming. The efficiency of the electroweak soup data skimming results to be 16% (0.55M events / 3.4M events).

The selection efficiency for the $W \rightarrow \mu\nu$ skimming of the soft muon sample was 50% (0.94M events / 1.8M events). The "Soft Muon soup" consisted of the mixing of approximately 1M events of J/Ψ inclusively produced in pp collision with 1M events of $pp \rightarrow \mu + X$ events from a minimum bias sample, preselected with a requirement of $p_T(m) > 4$GeV/c at the generator level. The analysis was performed with a sample of 0.12M events. Figure 66 shows the transverse momentum of muons for the different electroweak and soft muon soup components with cross section normalization. Figure 67 shows the transverse mass for the W boson candidate (reconstructed with the missing E_T estimation). Figures 68 and 69 show the muon p_T and W boson E_T distributions after muon isolation cuts.

13.3.3 Dimuon Reconstruction efficiency

A CMSSW skim job was prepared to select events with at least two reconstructed muons with $p_T > 5$ GeV/c or at least one reconstructed muon with $p_T > 20$ GeV/c. This event selection was applied to the $Z \rightarrow \mu\mu$ and the EWK-Soup datasets. The skim jobs ran at two Tier-1 centers (FNAL and Spain) and the selected events were transfered to the Tier-2 center at Purdue. The selection efficiency is about 23%.

Using these skimmed datasets a study of calculating muon reconstruction efficiency from data was performed. The idea is to compare and cross-check muons reconstructed in the tracker system, the standalone muon system, and the full tracker+muon system, thus allowing to obtain reconstruction efficiencies without relying on Monte Carlo information.

We calculate the dimuon invariant mass spectrum using opposite-sign muon pairs from

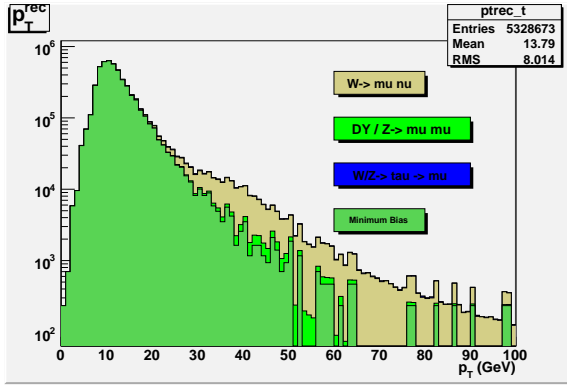- Monte Carlo simulation,

- global muon reconstruction,

Figure 66: Transverse momentum of muons after the skimming selection for the different electroweak soup and soft muons components with the cross section normalization.
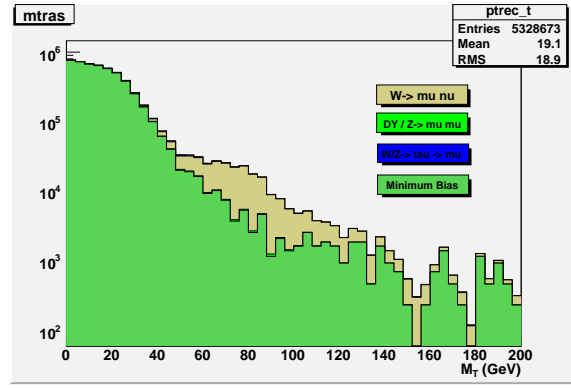

Figure 67: Transverse mass of W bosons candidate after the skimming selection.
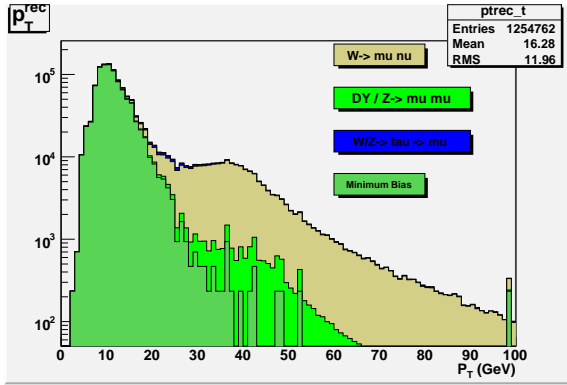

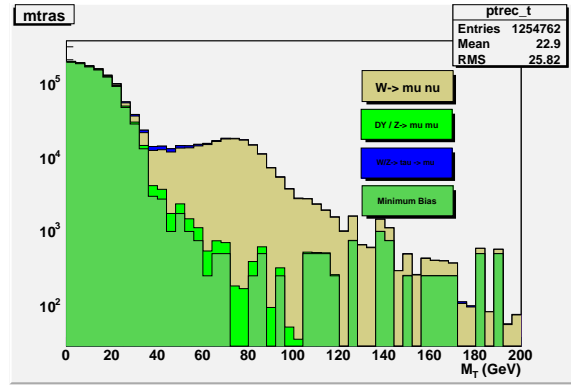Figure 68: Transverse momentum of muons after the isolation cuts.

Figure 69: Transverse mass of W bosons candidate after the isolation cuts.

- combinations of global and standalone muons,
- combinations of one global muon and one tracker track.

To avoid double counting a minimal separation of $\delta R = 0.3$ between the two muons is required.

Figure 70 shows the muon p_T distribution for selected events with only a single muon. Figure 71 shows the opposite-sign dimuon mass spectrum.

13.3.4 Reconstruction of Z$\rightarrow e^+e^-$ events

This exercise has multiple goals: one is the selection of Z$\rightarrow e^+e^-$ out of the "Electroweak Soup" (EWK) and their use to evaluate the performance of the electron reconstruction; and another one is the demonstration of the workflow for running offline ECAL calibration at a remote Tier-2 centre, starting from an AlcaReco stream produced at the Tier-1 centre.

For this reason a special skim filter job has been designed, producing as output two streams for the same events: one is the standard electron AlcaReco stream, as described in section 5.7, the other one is the standard RECOSIM output. The events are selected on the basis of Monte Carlo generator information, requiring only Drell-Yan e^+e^- events in a mass range from 50 to 130 GeV/c^2. The electrons are also required to have p_T greater than 5 GeV/c and $|\eta| < 2.7$. This selection has an efficiency over the Electroweak Soup of 3.7%.
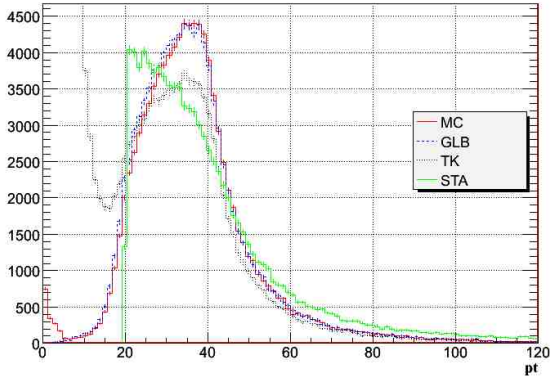
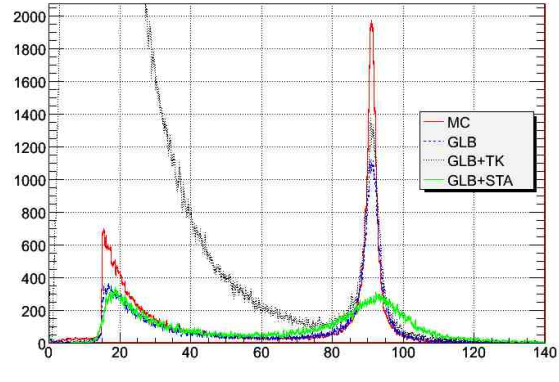Figure 70: The muon p_T distribution for selected events with only a single muon.

Figure 71: The opposite-sign dimuon mass spectrum.

Skim jobs were processed at the CNAF Tier-1, starting from less than 3 million EWK soup events (not the full EWK dataset have been processed using the CMSSW_1_0_5 release). Out of these events, a sample of 76494 events have been selected and the two created streams were transferred to the Rome Tier-2. The total amount of data amounts to about 33.5GB, of which around 500MB are occupied by the AlcaReco electron stream. Analysis jobs are sent to the Rome Tier-2 using CRAB_1_3_0 and CMSSW_1_0_5 on both the RECOSIM and the AlcaReco stream.

The main purpose is to validate the electron reconstruction, evaluating the electron reconstruction performance. The electron reconstruction efficiency both versus p_T and η is shown in Fig. 72.
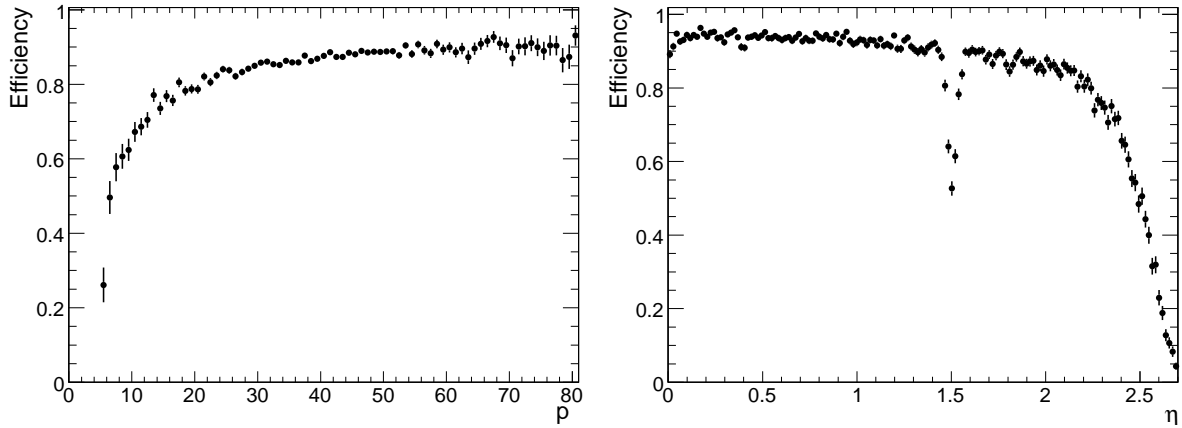

Figure 72: (left) Electron reconstruction efficiency as a function of p_T and (right) η.

The two main ingredients for the electron reconstruction, the supercluster and the electron track, are then evaluated separately. The ratio of reconstructed supercluster energy over the true energy as a function of η is visible in Fig. 73, showing a well-known problem in the CMSSW_1_0_5 release for what concerns the endcap supercluster reconstruction. The projected distribution for the barrel is presented in the right plot of Fig. 73.

The quality of the track reconstruction is evaluated looking at the E/p distribution, using the track parameters both at the vertex and at the outermost state, as it can be seen in Fig. 74.

At the time of the CMSSW_1_0_5, the reconstructed electron track was not the track after the smoothing step, and for this reason the quality of the E/p distribution with the track parameters at the vertex is rather poor. A refitting of the electron track has been tried in the offline analysis, improving the quality of the E/p matching, as it visible in Fig. 75.

The invariant mass computed from the electron pair nearest two the nominal Z mass is presented in Fig. 76.

The distribution of the difference between the reconstructed mass and the generated mass is shown in Fig. 77, where the right plot displays the difference as a fuction of η. The distribution shows the expected behaviour due to
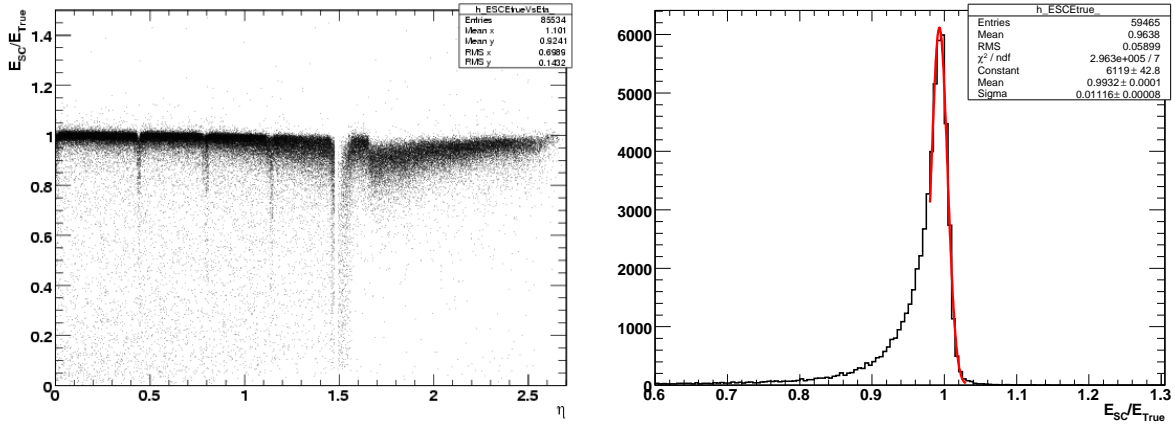
Figure 73: (left) Distribution of the electron reconstructed supercluster energy over the true energy as a function of η and (right) the corresponding distribution in the ECAL barrel acceptance.
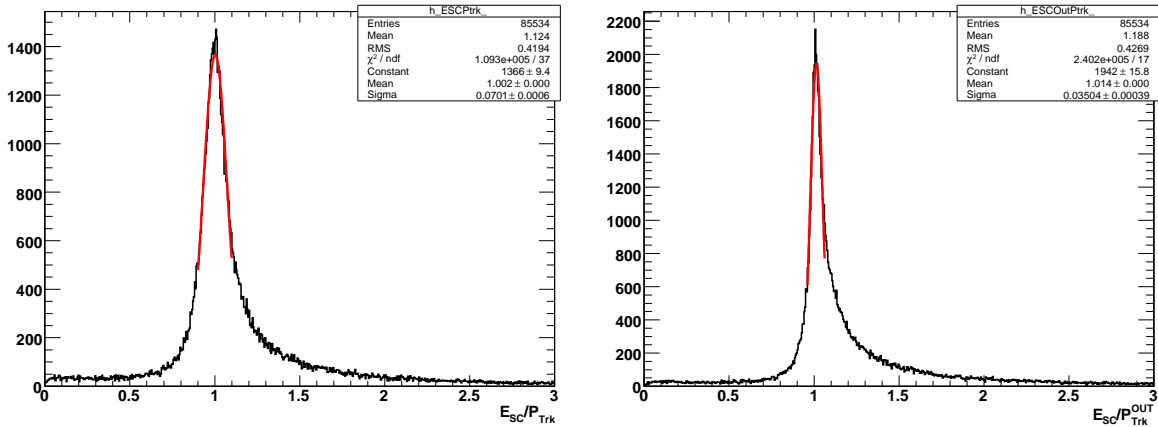

Figure 74: Distribution of the E/p variable using track parameters at the (left) vertex (right) outermost state.
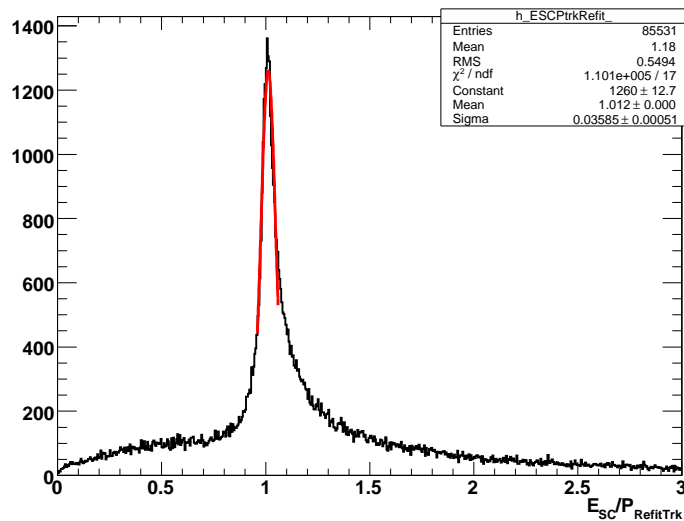

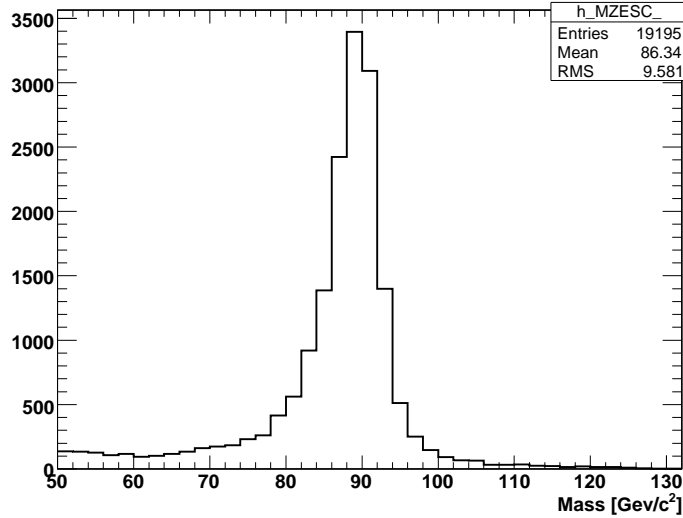Figure 75: Distribution of the E/p variable using refitted track parameters at the vertex.

Figure 76: Electron pair invariant mass in Z→ e^+e^- events.

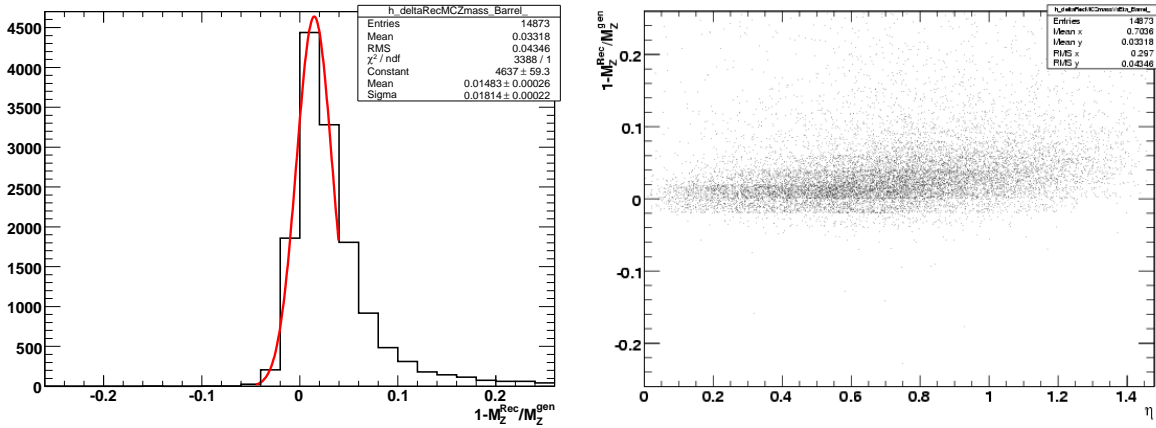the effect of the increasing tracker material towards η over the electron reconstruction.

Figure 77: (left) Relative difference between the reconstructed and the generated mass in Z→ e^+e^- events and (right) its bahaviour as a function of η.

Using the corrections present in CMSSW_1_0_5, the invariant mass peak is about 1.4% off, while the Z invariant mass resolution is about 1.8%,

13.3.5 M_W measurement exercise

The analyzed dataset for this exercise is /CSA06-103-os-EWKSoup0-0/RECOSIM/CMSSW_1_0_5-eg_ZeeFilter-1161162341 which contains 76 494 events. The jobs have been submitted using CMSSW 1.0.5 and CRAB 1.4.0, and have run on the Rome Tier-2 CPUs.

$Z \rightarrow e^+e^-$ events are treated by ignoring one randomly chosen lepton to mimic $W \rightarrow e\nu$ events. Z events with an invariant mass of the electron-positron pair $M_{e^+e^-} > 50$ GeV are then selected using the criteria applied in the W mass analysis [12]: one electron inside the pseudorapidity region $|\eta| <2.4$ and with transverse energy greater than 29 GeV; missing transverse energy $E_{missing}^T > 25$ GeV; the momentum of the system recoiling against the boson below 20 GeV; and, finally, no jets in the event with tranverse momentum greater 30 GeV. Electrons are chosen among electron candidates, requiring a geometrical matching between the supercluster and the associated track ($|\phi_{SC} - \phi_{track}| <0.1$ and $|\eta_{SC} - \eta_{track}| <0.006$) and a matching between the energy of the supercluster and the momentum measured from the track ($0.8< E_{SC}/P_{track} <2.5$). The missing transverse energy is measured from the energy deposited in the calorimeters towers and the jets are obtained with the iterative cone algorithm.

Variables relevant to this analysis are the transverse energy of the lepton, the missing transverse energy and the invariant transverse mass of the lepton pair. Their spectra and the corresponding resolutions are shown in Fig. 78, 79 and 80.
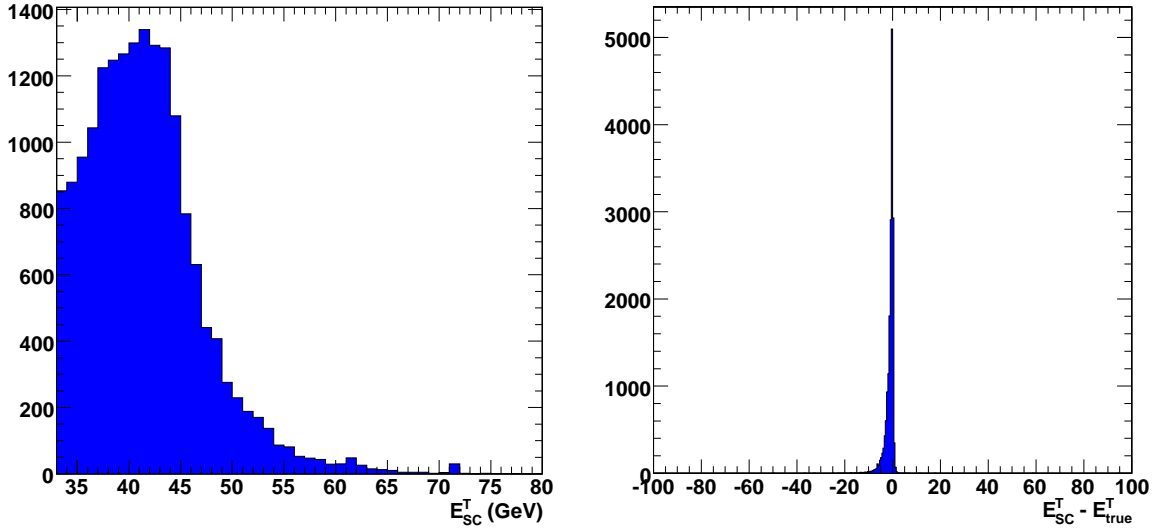
Figure 78: The electron transverse energy distribution in $Z \rightarrow e^+e^-$ decays (left) and the corresponding resolution (right).
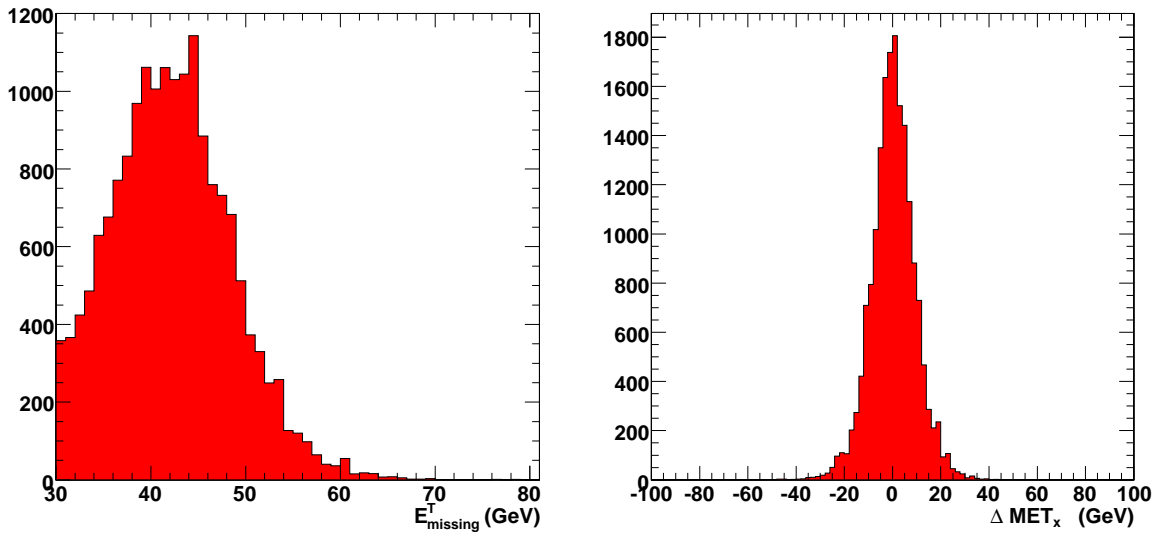

Figure 79: The missing transverse energy distribution in $Z \rightarrow e^+e^-$ events with one lepton removed (left) and the resolution for one MET component (right).

13.3.6 Selection of Z$\rightarrow \tau\tau \rightarrow \mu + \text{jet}$ events

The goal of this exercise was the selection of Z$\rightarrow \tau\tau \rightarrow \mu + \text{jet}$ events out of "Electroweak Soup" and reconstruction of the effective mass of the Z boson. The skim filter selects events of Drell-Yan $\ell\ell$ ($\ell = e, \mu, \tau$) production with electron or muon transverse momentum $p_T^{e,\mu} > 15$ GeV/c based on Monte Carlo "truth." The RECOSIM output of the filter was moved to Tier-2 centers in UK, Italy, Switzerland and US and accessed through Grid. The high level objects were reconstructed as follows: calorimeter jets with an iterative cone algorithm using cone size 0.5; missing E_T (MET) was calculated from calorimeter towers and muon transverse momentum; muons were reconstructed using global muon reconstruction algorithm. No jet energy corrections were applied. The τ-jet tagging uses tracker isolation criteria and cut on the transverse momentum of the leading track in the signal cone. An example of one analysis is described below.

About 380K skimmed events were processed and further preselected with loose cuts on muons and τ-jet tagging. The preselection efficiency is 0.023. Selected events (8.7K events of the size 650 MB) were stored to local PC, where the final selection cuts were optimized. The final selection criteria are:
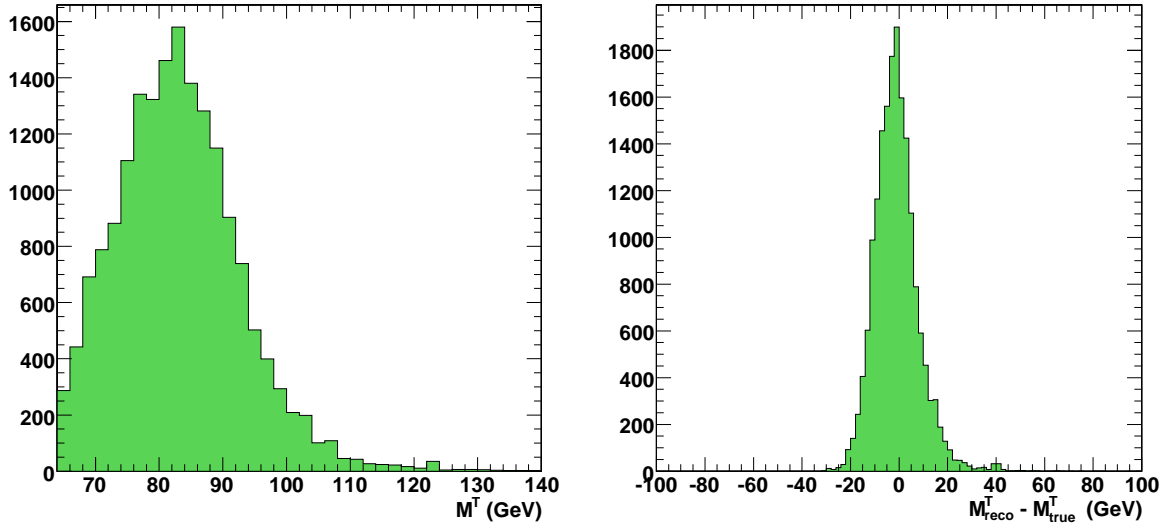
Figure 80: The transverse mass spectrum in $Z \rightarrow e^+e^-$ events with one lepton removed (left) and its resolution (right).

- only one muon with $p_T^\mu > 15$ GeV/c, $|\eta| < 2.1$

- at least one jet with $E_T > 25$ GeV, $|\eta| < 2.1$ passed τ-jet identification

- the τ-jet algorithm uses the following parameters: a matching cone size of 0.1; a signal cone size of 0.07; an isolation cone size of 0.35; a cut on the momentum of the leading track in the signal cone is 10 GeV/c; only tracks with a z impact parameter within 2 mm window around the leading track z impact parameter counted for the isolation and a requirement of no tracks with $p_T > 1$ GeV in the isolation cone; and only one or three tracks required in the signal cone.

- muon and τ jet have opposite charge

- energy of neutrinos from τ decays reconstructed using missing E_T and collinear approximation is positive.

- angle between the muon and τ-jet in the transverse plane $\Delta\varphi_{\mu,\tau} < 175^o$

- electron rejection criteria: $0.2 < R < 1.1$, where $R = E_{\tau-jet}^{HCAL}/p_T^{ltr}$

The total selection efficiency with the final cuts is 2.75×10^{-4}. Fig. 81 shows reconstructed di-tau effective mass after all selections but without electron rejection. The black histogram is the sum of $Z \rightarrow \tau\tau \rightarrow \mu +$ jet and other (background) events passed the selection criteria. The gray histogram shows background events only. The dominant contribution to the background consists of $Z \rightarrow \tau\tau \rightarrow \mu e$ events when electron is misidentified as τ-jet. The contamination of electrons is removed applying electron rejection criteria, $0.2 < R < 1.1$. The di-tau effected mass after electron rejection is shown in Fig. 82.

13.3.7 Measurement of jet mis-τ-tagging efficiency from $Z \rightarrow \mu^+\mu^-$ plus jet events.

The goal of this exercise was the selection of $Z \rightarrow \mu^+\mu^-$ plus jet events out of "Electroweak Soup" and use it for the measurement of jet mis-τ-tagging efficiency as a function of p_T^Z (and η^Z) exploiting the p_T balance of the Z boson and recoiling jet. The skim filter selects events of Drell-Yan $\ell\ell$ ($\ell =$ e, μ, τ) production with electron or muon transverse momentum $p_T^{e,\mu} > 15$ GeV/c based on Monte Carlo "truth." The RECOSIM output of the filter was moved to Tier-2 centers and accessed through Grid. The high level objects were reconstructed as follows: calorimeter jets with an iterative cone algorithm using a cone size of 0.5; muons were reconstructed using global muon reconstruction algorithm. No jet energy corrections were applied. An example of one analysis is described below.

About 600K skimmed events were processed, a sample equivalent to an integrated luminosity of approximately 0.4 fb^{-1}. The useful events were selected with the following criteria:
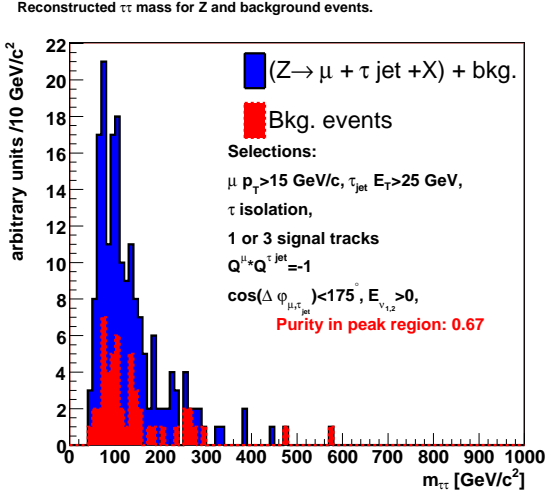
Figure 81: The reconstructed di-tau effected mass after all selections but without electron rejection. The black histogram - Z→ $\tau\tau$ → μ + jet plus background events. The gray histogram - background events only.
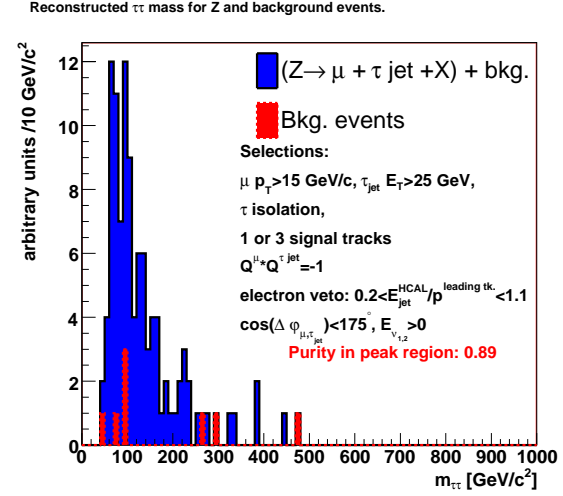
Figure 82: The same as Fig. 81, but after electron rejection criteria applied.

- two muons with $p_T^\mu > 10$ GeV/c, $|\eta| < 2.5$,

- mass of dimuon pair is within Z boson mass window, $70 < M_{\mu\mu} < 110$ GeV/c^2,

- dimuon transverse momentum $p_T^{\mu\mu} > 15$ GeV/c,

- only one 0.5 cone jet in the event with raw $E_T >15$ GeV, $|\eta| < 2.5$,

- back-to-back $\mu\mu-$jet configuration in the transverse plane, $\Delta\varphi_{\mu\mu-\text{jet}} >135^o$.

About 7K events were selected. For the events which have a jet with $|\eta| < 2.1$ the jet was passed through the τ-jet identification algorithm. It uses tracker isolation criteria and cut on the transverse momentum of the leading track in the signal cone. The following parameters were set in the algorithm: the matching cone size is 0.1; the signal cone size is 0.07; the isolation cone size is 0.4; the cut on the momentum of the leading track in the signal cone is 6 or 10 GeV/c; only tracks with z impact parameter within 2 mm window around the leading track z impact parameter were counted for the isolation and a requirement of no tracks with $p_T >1$ GeV in the isolation cone.

Fig. 83 shows reconstructed dimuon effective mass after selection of events with two muons with $p_T^\mu > 10$ GeV/c, $|\eta| < 2.5$. The grey part of histogram shows the events selected for the further analysis. Fig. 84 shows the angle in the transverse plane, $\Delta\varphi_{\mu\mu-\text{jet}}$, between dimuon momentum and the jet for events with $70 < M_{\mu\mu} < 110$ GeV/c^2 and requirement to have only one jet in the event with $E_T >15$ GeV, $|\eta| < 5.0$. The grey part of histogram shows the back-to-back events selected for the further analysis. The dimuon transverse momentum for selected events is shown in Fig. 85. The efficiency of the τ-jet tagging was evaluated as a function of the dimuon transverse momentum, $p_T^{\mu\mu}$. The $p_T^{\mu\mu}$ should be approximately equal to the recoiling parton (gluon or quark) momentum after all selections. Therefore it is possible to evaluate the tagging efficiency as a function of parton p_T and η as well as a function of the reconstructed jet E_T and η. Fig. 86 presents the result of the exercise. It shows the jet mis-τ-tagging efficiency vs the dimuon transverse momentum, for two different working points. The average tagging efficiency in simulated muonic Z boson events was observed to be $(3.4 \pm 0.2)\%$ and $(2.1 \pm 0.1)\%$ for the two different working points.

13.3.8 τ validation package

In addition to the analysis exercises described in the previous sections the τ reconstruction validation package has been run on skimmed $t\bar{t}$ events. The skim filter, based on Monte Carlo "truth," selects events with at least one electron or muon with a transverse momentum of $p_T^{e,\mu} >15$ GeV/c with a selection efficiency of $\approx 47\%$
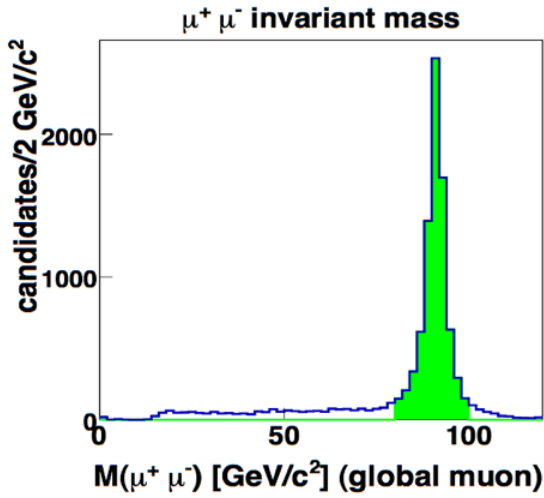
Figure 83: The reconstructed dimuon effective mass after selection of events with two muons with $p_T^\mu > 10$ GeV/c, $|\eta| < 2.5$. The grey part of histogram shows the events selected for the further analysis
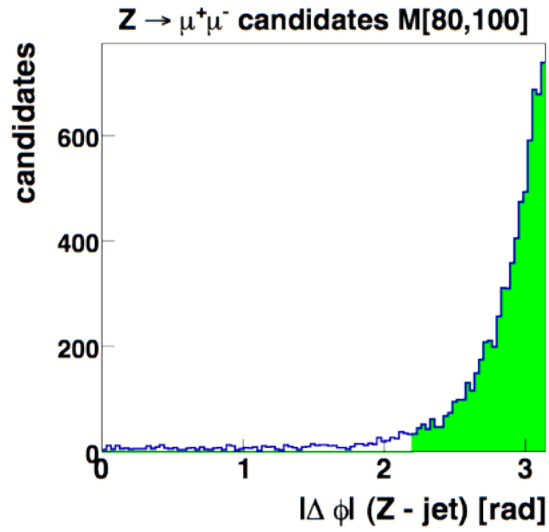
Figure 84: The angle in the transverse plane, $\Delta\varphi_{\mu\mu-\text{jet}}$, between dimuon momentum and the jet for events with $70 < M_{\mu\mu} < 110$ GeV/c^2 and requirement to have only one jet in the event with $E_T > 15$ GeV, $|\eta| < 2.5$. The grey part of histogram shows the back-to-back events selected for the further analysis.
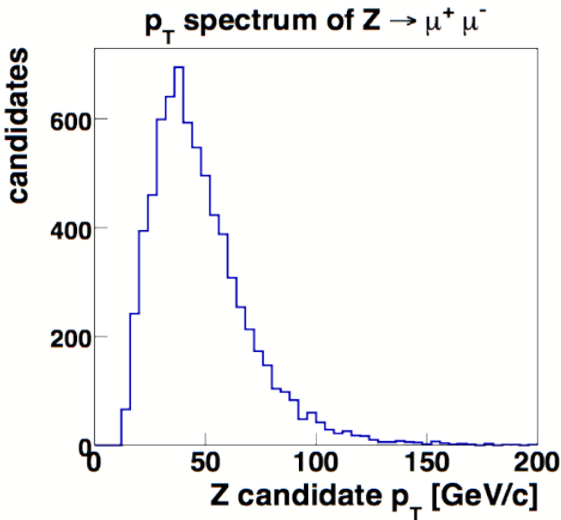

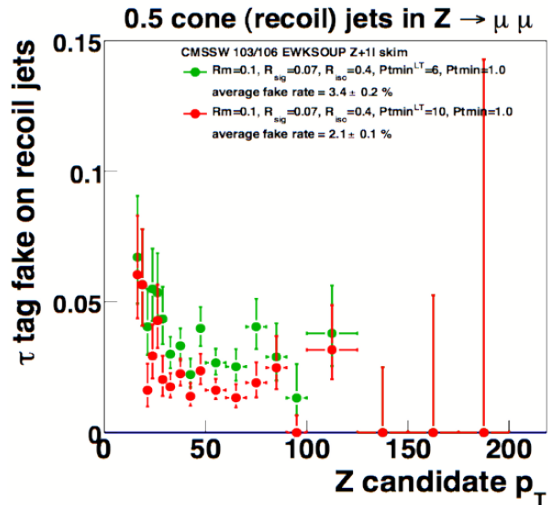Figure 85: The dimuon transverse momentum after all selections.

Figure 86: The jet mis-τ-tagging efficiency vs the dimuon transverse momentum.

A total of 2.15M $t\bar{t}$ events were produced with a total data size of $\approx$1.5TB (FEVT). The RECOSIM output of the filter ($\approx$180 GB) was moved to various Tier-2 centers in UK, Italy, Switzerland and Germany. The jobs, submitted through CRAB from CERN, have been run over 250k events in Pisa Tier-2.

The τ validation package, whose scope is to monitor the performance of the various τ tagging methods[13], includes a number of standard histograms which be produced periodically for each release of CMSSW and during data taking. Some of the most relevant histograms are the Efficiency vs Isolation Cone, Efficiency vs Jet transverse momentum, Efficiency to find the Leading Track, Number of Signal Tracks, ΔR between the leading track and the jet. As an example Fig. 13.3.8 shows the τ tagging efficiency versus the isolation cone (left) and versus the jet transverse energy (right). The various curves correspond to different selection of the primary vertex: for $t\bar{t}$ events without pileup, the various definitions are almost equivalent, as expected.
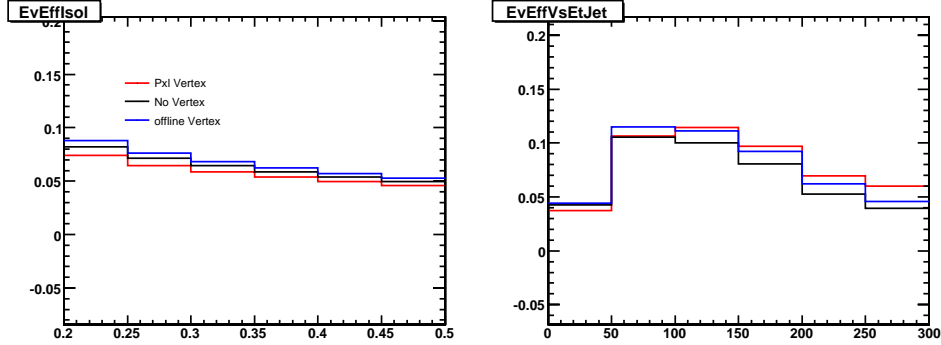

Figure 87: $t\bar{t}$ events: τ tagging efficiency vs isolation cone (left); τ tagging efficiency vs jet transverse energy (rigth). The various curves correspond to different selection of the primary vertex.

13.3.9 Selection of H→WW→ $\ell\ell$ events

The goal of this exercise was the selection of an early Higgs boson discovery channel, inclusive H→WW→ $\ell\ell$ with M_H=165 GeV/c^2, out of "Electroweak Soup" and $t\bar{t}$ background events. The non skimmed datasets of RECOSIM format located in CERN were accessed with CRAB. The single W and Drell-Yan $\ell\ell$ production processes in the "Electro Weak Soup" were not taken in to account in the analysis. For inclusive $t\bar{t}$ dataset only events with leptonic decays of W bosons were analyzed. The high level objects were reconstructed as follows: calorimeter jets with an iterative cone algorithm using a cone size of 0.5; missing E_T (E_T^{miss}) was calculated from two high p_T leptons in the final state; muons were reconstructed using global muon reconstruction algorithm and electron candidates matched with the hits in the pixel detector were used. The number of signal and background events corresponding to luminosity 5 fb^{-1} were analyzed.

Event selections were used the same as in the CMS Physics TDR Vol.2 analysis[3]. First, events with two central and high p_T isolated leptons, $p_T^\ell > 20$ GeV/c and $|\eta^\ell| < 2.0$ were selected. Electron isolation requires that the sum p_T of the tracks (with the p_T >0.9 GeV/c) within a cone 0.2 around the η, ϕ direction of the electromagnetic cluster at the vertex should be less than 5% of the cluster energy. It was also required that ratio of cluster energies in the HCAL and ECAL should be less that 5%. The E/p ratio should be bigger than 0.8 and $|1/E - 1/p|$ <0.05. Muon isolation requires that the energy in the calorimeter around the muon candidate within a cone of 0.3 must be smaller than 5 GeV and the sum of the p_T of the tracks within a cone of 0.25 around the muon candidates must be smaller than 2 GeV. The following kinematic selections were applied:

- $E_T^{miss} > 50$ GeV

- $\phi_{\ell\ell} < 45^o$ (angle between the leptons in the transverse plane)

- 12 GeV/$c^2 < m_{\ell\ell} < 40$ GeV/c^2 (invariant mass of two leptons)

- jet veto. No jets with E_T^{raw} >15 GeV and $|\eta| < 2.5$

- 30 GeV/$c < p_T^{\ell\,max} < 55$ GeV/c (lepton with maximal p_T)

- $p_T^{\ell\,min} > 25$ GeV/c (lepton with minimal p_T)

In order to suppress fake jets with E_T^{raw} between 15 and 20 GeV the tracker information was used. In particular, the sum p_T of tracks (with $p_T > 2$ GeV/c from the vertex and within cone of 0.5 around jet direction should be bigger than 20% of the raw jet E_T. The total selection efficiency was 2% for the signal, 0.148% for the WW and 0.012% for $t\bar{t}$ backgrounds. Fig. 13.3.9 (left plot) shows the angle between the leptons in the transverse plane for the signal and backgrounds after all selections for 5 fb^{-1}.
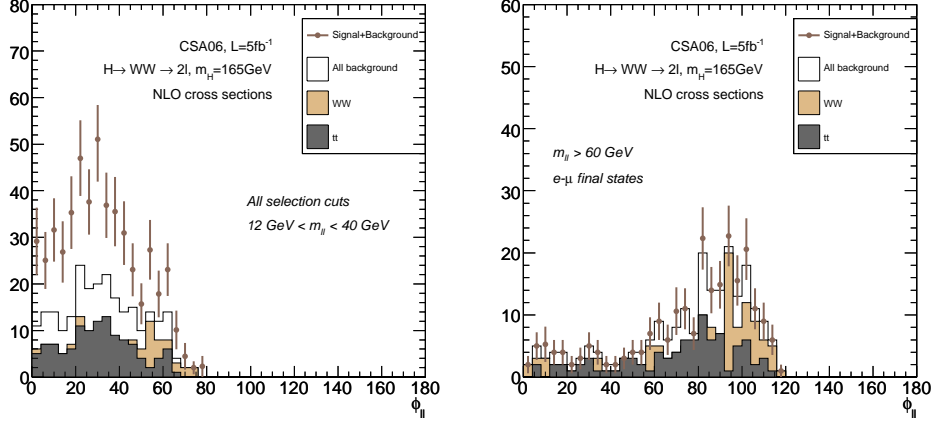

Figure 88: Left plot: the angle $\phi_{\ell\ell}$ between the leptons in the transverse plane for the signal and backgrounds after all selections for 5 fb^{-1}. Right plot: the angle $\phi_{\ell\ell}$ for the WW background normalization region.

The WW background normalization region was defined by keeping the same cuts that the signal but requiring $\phi < 140^o$ and $m_{\ell\ell} > 60$ GeV/c^2. Moreover only opposite flavor leptons are considered in order to reduce the Drell-Yan and WZ contribution. Fig. 13.3.9 (right plot) shows the ϕ distribution for the WW and $t\bar{t}$ precesses in this normalization region.

13.3.10 Measurement of Z plus two jets background for invisibly decaying Higgs boson produced in vector boson fusion.

The Z plus two jets process with Z$\to \nu\nu$ decay is a background for Higgs boson production in vector boson fusion (VBF) with the invisible decay mode of the Higgs boson. The goal of this exercise was the selection of Z$\to \mu^+\mu^-$ plus two jets events in the signal like VBF topology. In a real analysis it will give a data-driven prediction of the Z+2j background exploiting the known ratio of branchings Z$\to \nu\nu$ to Z$\to \ell\ell$. The skim filter on "Electroweak Soup" selects events of Drell-Yan $\ell\ell$ (ℓ = e, μ, τ) production with electron or muon transverse momentum $p_T^{e,\mu} > 15$ GeV/c based on Monte Carlo "truth." The RECOSIM datasets, output of the filter, located in Tier-1 centers at DESY (GridKa) and FNAL were accessed with CRAB. The statistics corresponding to approximately 300 pb^{-1} of data were analyzed. The high level objects used in the analysis were reconstructed as follows: calorimeter jets with an iterative cone algorithm using cone size 0.5; muons were reconstructed using global muon reconstruction algorithm; the missing E_T (E_T^{miss}) was reconstructed from the calorimeter towers. Muons were not added to E_T^{miss} in order to reproduce Z$\to \nu\nu$ plus two jets topology. No jet energy corrections were applied.

Events with two reconstructed muons of $p_T^\mu > 15$ GeV/c and 80 GeV/$c^2 < m_{\mu\mu} < 100$ GeV/c^2 and at least one jet with $E_T^{jet} > 20$ GeV were first selected. Fig. 13.3.10 shows the the E_T^{jet} distribution for events with at least one jet (upper histogram), at least two jets (middle histogram) and at least three jets (lower histogram). The further event selections on jets and E_T^{miss} were the same as for the signal selection:

- at least two jets with E_T of the two highest E_T jets (tagging jets) is bigger that 40 GeV
- $|\eta_{j1} - \eta_{j2}| > 4.0$
- $M_{j1,j2} > 1$ TeV
- $E_T^{miss} > 100$ GeV; muons were not added to mimic Z$\to \nu\nu$ decay mode
- veto events with third jet of $E_T > 20$ GeV in the η interval between two tagging jets.

The selected number of events provides the statistical accuracy of the Z+2j background determination from the data. For PYTHIA Drell-Yan generation the number of events is underestimated and hence is not provided.
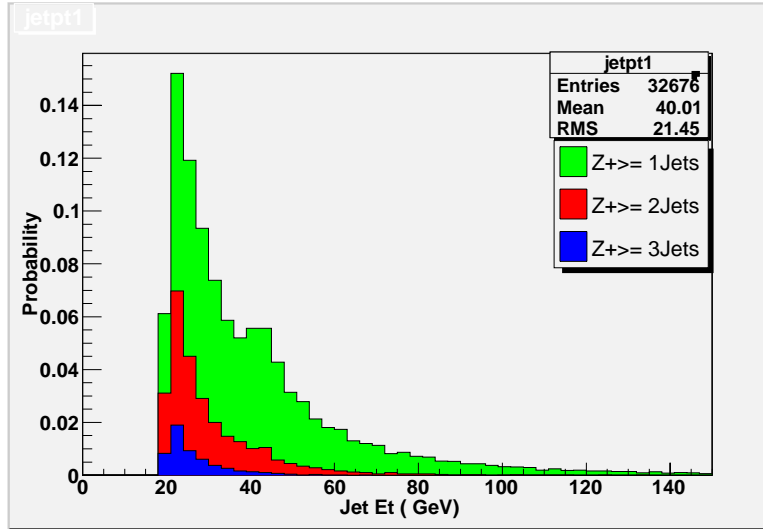
Figure 89: The E_T^{jet} distribution for events with at least one jet (upper histogram), at least two jets (middle histogram) and at least three jets (lower histogram).

13.3.11 Minimum Bias and Underlying Event

Minimum Bias (MB) collisions at the LHC are generic inelastic proton-proton interactions, including hard scattering and diffractive components. Studying the features of the MB at the LHC and cross-checking them with the predictions of the Monte Carlo generators is particularly important in order to describe the unavoidable background of pile-up at high luminosity. The Underlying Event (UE), instead, is the softer component of a single proton-proton collision accompanying the hard scattering, and accounting for a large fraction of the activity in terms of multiplicity and momentum of the observed particles.

The phenomenology of MB and UE is not the same, but the experimental methodology to study them are similar, mostly relying on the reconstruction of charged tracks.

For what concerns the UE, one can use the topological structure of hadron-hadron collisions to study this activity looking only at the outgoing charged particles. Jets are constructed from charged particles using a simple clustering algorithm and then the direction of the leading charged particle jet is used to isolate regions of eta-phi space that are sensitive to the UE. The transverse region to the charged particle jet direction, is almost perpendicular to the plane of the hard 2-to-2 scattering and is therefore very sensitive to the UE.

The Underlying Event (UE) analysis is based on the selection of the following events:

- MinimumBias

- Hadronic Jet

- Drell-Yan ($Z \rightarrow \mu\mu$)

For each kind of events a filter was created that defines the output stream based on the RECOSIM format. In detail the selection criteria that have been adopted are:

- Minimum Bias: a Minimum Bias trigger strategy is not yet defined, as it is not defined higher level selection criteria for this kind of events. A random selection is applied in order to select the events without introducing any additional bias.

- Hadronic Jets: the selection is performed requiring at least one calorimetric jet with a Transverse momentum greater than 20 GeV/c.

- Drell-Yan ($Z \rightarrow \mu\mu$): the stream is defined requiring at least two muons in the central region ($|\eta| < 2.5$) having a minimum transverse momentum of 3 GeV/c and invariant mass above 15 GeV/c.
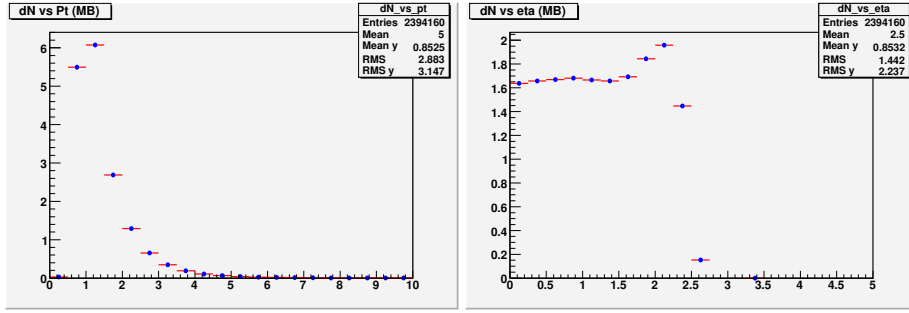
Figure 90: (right) Pt distribution of reconstructed track (left) multiplicity of reconstructed track as function of η
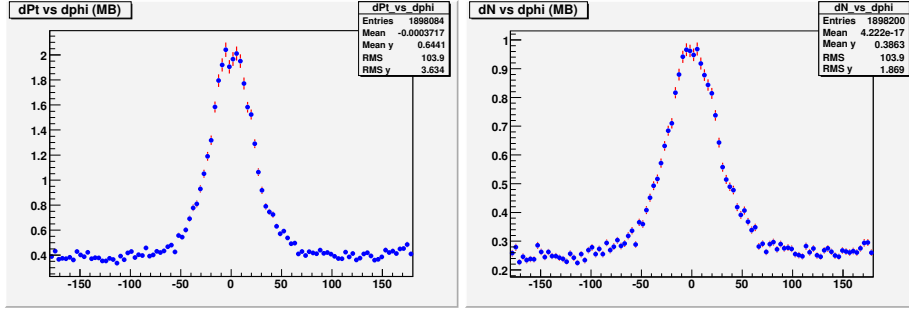

Figure 91: UE variables evaluated in Minimum Bias events, (right) energy density, (left) charge density.

Of course the first filter is also used in the Minimum Bias analysis.

These three filters have been run at Tier-1 (FNAL and CNAF) and the output was skimmed to the subscribing Tier-2 (Pisa and Florida). The MB skim output was transferred to the Pisa Tier-2 while the output of the Hadronic Jet sand the Drell-Yan skims were transferred to the Florida Tier-2.

The analysis step was performed using CRAB. Two different analyses for the MB and UE observables were run at Pisa and Florida. Starting from a total sample of 3×10^6 MB events, 1.2×10^6 Hadronic Jets and 0.5×10^6 DY, we finally performed the analysis on 2×10^5 MB, 10^5 QCD Jet and 10^5 DY events.

The analysis code is committed in CMSSW under the packages AnalysisExamples/MinimumBiasUnderlyingEvent. The modules using the reconstructed and generator level quantities provide directly the relevant plots of the analysis. In the MB analysis we are particularly interested on the event track multiplicity and transverse momentum distribution Figure 90.

In the UE analysis for the hadronic topologies, instead, we concentrate on the observables in the activity in different regions with respect to the charged jets, studying in particular the energy density ($\delta P_{Tsum}/\delta\eta\delta\phi$) and the charge density ($\delta N_{chg}/\delta\eta\delta\phi$). Figure 91 and Figure 92 report the energy and charge density against the distance with respect to the leading charged jets, for Minimum Bias and Hadronic Jet events respectively.
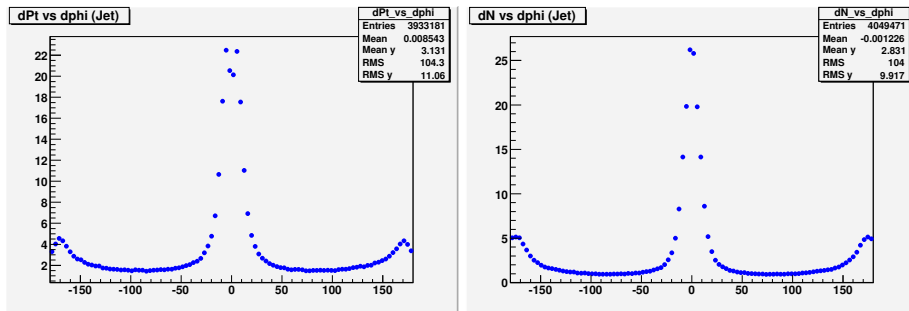

Figure 92: UE variables evaluated in Hadronic jet events, (right) energy density, (left) charge density.
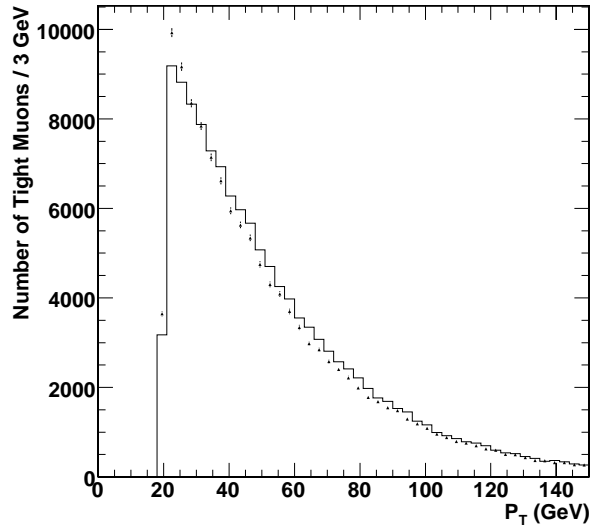
Figure 93: Distribution of reconstructed μ p_T (dots) compared with respect to the generated values (solid histogram) for μ selected after a p_T cut of 20 GeV/c in the $t\bar{t}$ dilepton exercise.

13.3.12 Selection of $t\bar{t}$ quark pairs in the dilepton channel

The aim of this exercise was to perform a selection of pairs of top quarks in the dilepton final state at a remote Tier-2, in this case T2_Spain_Ifca. The dilepton decay channel denotes the case where the two W bosons from the decaying $t\bar{t}$ pair both decay to final states containing an electron or a muon, accounting for about 5% of all $t\bar{t}$ SM decays.

The original sample used consisted of about 5 million $t\bar{t}$ inclusive events. Skim jobs run on that sample in order to select events having two high-p_T leptons in the final state. Based on the content of generator information in the event (HepMC), skim jobs required events having 2 leptons of opposite code, and then charge, without any restriction in η or ϕ, so basically selecting all dilepton events present in the sample in RECOSIM format. Skimmed data were transferred from FNAL Tier-1 to T2_Spain_Ifca then analyzed almost *on-line*, using CRAB and selection software adapted to CMSSW_1_0_6 version. The whole $t\bar{t}$ in AODSIM format was also transferred to the Tier-2 in order to have an estimation of the main background to the dilepton final state that comes from the semi-leptonic final state.

Elements used in the analysis were generation information (based on HepMC), muons, jets (from CaloJets), and without energy corrections, tracker information (ctfWithMaterialtracks), and missing E_T. The part of the selection that used electrons and b-tag info was not used.

Muons were selected if they had $p_T > 20$ GeV in the pseudo-rapitity ranges $|\eta| < 2.4$. A muon was considered isolated, if the sum of the p_T of all the tracks present in a cone of $\Delta R \leq 0.3$ minus p_T of the μ is less than 2 GeV/c. The efficiency obtained taking the two highest-p_T isolated muons in the event with opposite charge was of the order of 36%, that is similar to the one obtained in CMS-Note 2006/077. A selection of jets similar to the one described in CMS-note 2006/077 was also performed. Jets were selected when the electromagnetic fraction of their energy was below 0.8 Events were selected if they had two jets passing this selection and having uncorrected energy larger than 20 GeV in the pseudo-rapidity range $|\eta| < 2.5$. As b-tagging information was not present in the version used for this analysis, the fraction of jets correctly selected was checked based only in this information. Around 65% of the events had both jets within an angle of 20° with respect to the generated b-quark.

Figures 93 and 94 show the reconstructed muon spectra compared with the generated spectra, and the isolation variable for muons when they were matched and non-matched with respect to generated muons. The μ is considered matched if a generated μ is found within a cone of $\alpha = 0.15$ rad with respect to the reconstructed μ direction.
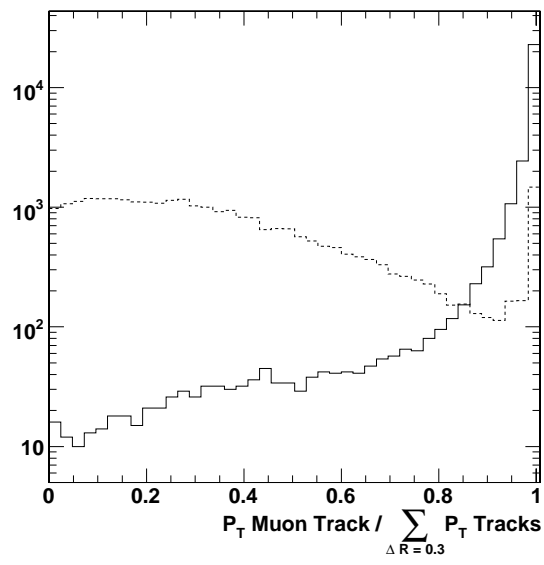
Figure 94: Distribution of variables used in μ isolation as described in the text for the $t\bar{t}$ dilepton exercise. Solid line corresponds to matched μ, and dashed line to non-matched μ.

13.3.13 SUSY Exercise using Jets plus Missing Transverse Energy at mSUGRA point LM1

Computing and Software

The analysis exercise reported here is entirely based on the CSA06 `ExoticsSoup` and `TTbar` data samples. The analysis code was written as a CMSSW 1.0.6 EDAnalyzer module.

For the LM1 analysis on the Exotics Soup, the `LM1_3IC5Jet30` Skim Filter was used. This is a combination of a LM1 Pythia process ID filter and a jet filter, in which 3 `iterativeCone5 reco::CaloJet` reconstructed jets with raw $E_t > 30$ GeV are required. A total filter efficiency of about 25 % with an FEVT event size of roughly 3.1 MB was obtained. For the $t\bar{t}$ analysis the `2IC5Jet100Exo` Skim Filter was applied. This is a similar jet filter as the one used for the Exotics Soup, but with the minimum requirement of 2 `iterativeCone5 reco::CaloJet` reconstructed jets with raw $E_t > 100$ GeV. A filter efficiency of about 7.5 % was obtained with an FEVT event size of about 3.8 MB. For the CSA06 samples that were accessed through the Grid, CRAB version 1.4.0 was used to submit jobs to the EDG scheduler with CERN as RB. The output ROOT files were stored on Castor at CERN. For part of the results presented in the analysis section the data was simply read from Castor at CERN. A summary of the Grid access to the samples with CRAB at the different Tier sites is given in Table 14. In the table a CRAB job is called "successful" if the ROOT output file was produced and stored on Castor at CERN correctly; this does not necessarily mean that no other problems occured. The initial number of jobs was the number of jobs submitted to run over the entire available data set; the total number of jobs includes jobs that failed for some reason and were resubmitted. In case of the Exotics Soup RECO sample two times 11 jobs were submitted, once with the jet skim filter and once without. For the first group of jobs, 1 job failed on GridKa[3] and was resubmitted successfully to IIHE. The second group of jobs had 2 jobs failing at CERN due to RFIO problems reading the ROOT input files[4] and were resubmitted to CERN successfully. For the TTbar sample, 1 job failed at CERN with a segmentation fault during running[5]. When trying to access the `ExoticsSoup Zprime` skimmed sample, all 6 initial jobs at FNAL failed due to a dCache error[6] because the maximum number of connections per dCap door, set at 1500, was reached. When this limit was increased to 3000 all jobs were resubmitted and terminated successfully, however 3 of the resubmitted jobs ended with status `aborted` even though the ROOT files were correctly produced. In general, when retreiving the job output reports after all jobs were finished a lot of corrupted output errors were encountered, most probably because the maximum size of the output Sandbox was exceeded.

Table 14: Summary of Grid access with CRAB to the CSA06 data samples.

Primary dataset	Data Tier	Initial # jobs	Total # jobs	Successful # jobs
/CSA06-106-os-ExoticSoup0-0/RECO /CMSSW_1_0_6-RECO_Hee5e8972ca2b9fe5859521c27772ba77	CERN, IIHE, RAL, GridKa	22	25	19
/CSA06-106-os-TTbar0-0/FEVT /CMSSW_1_0_6-su_TTbar_2IC5Jet100ExoFilter-1162574027	CERN	8	8	7
/CSA06-103-os-ExoticSoup0-0/FEVT /CMSSW_1_0_6-su_Exotics_ZprimeFilter-1161788458	FNAL	6	12	6

In the EDAnalyzer module the following objects from the ROOT input files were selected :

- `reco::GenJetCollection::iterativeCone5GenJets`

- `reco::CaloJetCollection::iterativeCone5CaloJets`

- `reco::MuonCollection::globalMuons`

- `reco::CaloMETCollection::met`

- `reco::GenMETCollection::genmet`

---

[3] max. retry count exceeded, job `exit code NULL` and no ROOT output

[4] job `exit code 134`, SEALException, RFIO error 0/1004

[5] job `exit code NULL`

[6] job exit 8000, dCache error 26 Server rejected "hello"

- `HepMC::HepMCProduct::VtxSmeared`

- `edm::TriggerResults::TriggerResults`

For the $t\bar{t}$ jet studies reported below, also the `midPointCone5`, `midPointCone7` and `kt GenJet` and `CaloJet` collections were considered. The scope of the analysis exercise is to validate part of the kinematic event selections reported in the Physics TDR Vol.2 [3] for SUSY searches at LM1 and to perform some simple jet and E_T^{miss} resolution studies with CMSSW.

Analysis

Part of the event selection criteria used in the E_T^{miss} + *multi-jet* SUSY search as reported in the Physics TDR Vol.2 were applied to the LM1 signal sample and to the $t\bar{t}$ background sample. The results are summarized in Table 15. Using the LO cross sections $\sigma_{LM1} = 49$ pb^{-1} and $\sigma_{t\bar{t}} = 430$ pb^{-1}, 7442 LM1 signal events and 241 $t\bar{t}$ background events are obtained for 1 fb^{-1} after the requirements listed in the table. In the Physics TDR analysis several additional requirements, e.g. event cleanup criteria, and topological requirements were applied and 6319 LM1 and 53.9 $t\bar{t}$ events survived the full selection path for 1 fb^{-1}. The obtained LM1 signal efficiency in the present analysis with CMSSW is in fair agreement with the ORCA results reported in the Physics TDR as expected [7] This is not the case for the application of the same selection on the top sample given the differences in jet and E_T^{miss} reconstruction performace between ORCA and the CMSSW version used for CSA06.

Table 15: E_T^{miss} + *multi-jet* SUSY search event selection.

selection	LM1 signal eff.	$t\bar{t}$ eff.
$NJet \geq 3$ ($E_T > 30$ GeV)	67 %	52 %
+ $E_T^1 > 180$ GeV, $E_T^2 > 110$ GeV, $E_T^3 > 30$ GeV	33 %	2,5 %
+ $E_T^{miss} > 200$ GeV	18 %	0.081 %
+ $H_T(\equiv E_T^{miss} + E_T^2 + E_T^3 + (E_T^4)) > 500$ GeV	15 %	0.056 %
Total PTDR LM1 signal eff. = 13 %		
Total PTDR $t\bar{t}$ eff. $\approx$ 0.013 %		

In the *inclusive muon + jets + E_T^{miss}* analysis as reported in the Physics TDR, events were selected with a leading muon with $P_T > 30$ GeV, $E_T^{miss} > 130$ GeV, $E_T^1 > 440$ GeV, $E_T^2 > 440$ GeV, $|\eta^1| < 1.9$, $|\eta^2| < 1.5$, $|\eta^3| < 3$, $\cos(\Delta\phi(j_1, j_2)) < 0.2$, $-0.95 < \cos(\Delta\phi(E_T^{miss}, j_1)) < 0.3$, $\cos(\Delta\phi(E_T^{miss}, j_2)) < 0.85$. In the present analysis, a total of 403 LM1 signal events for 10 fb^{-1} are obtained after these requirements are applied, whereas zero $t\bar{t}$ events pass the section requirements (in a $t\bar{t}$ sample corresponding to about 0.25 fb^{-1}). These numbers roughly agree with the Physics TDR results where for 10 fb^{-1} 311 LM1 signal events and 2.54 events for the total SM background were reported.

Using the event selection criteria from Table 15, some kinematical distributions for the LM1 signal events are obtained here with CMSSW and directly compared to the ORCA results directly taken from the Physics TDR. The comparisons are shown in Fig. 95 for 1 fb^{-1}. The level of agreement in the signal distributions between CMSSW and ORCA is promising. The background distributions are the ones from the Physics TDR.

HLT for LM1 and $t\bar{t}$

The HLT trigger paths and the bit order for the CSA06 samples are listed in Table 16. See Table 1 for the thresholds applied for each trigger. The actual implementation of the trigger simulation is based on generator level (HepMC) information. In Fig. 96 the HLT bits are shown for an LM1 event and $t\bar{t}$ sample before and after the application of the E_T^{miss} + *multi-jet* selection. Clearly, for the LM1 sample the selection path enhances the contribution of events passing a 1, 2, 3 or 4 jet HLT trigger. For none of the trigger paths the 100 % level is reached due to the stringent requirements on the HLT jet p_T ($|\eta| \leq 5$, single jet : $p_T^{min} = 400$ GeV; dijet : $p_T^{min} = 350$ GeV; triJet : $p_T^{min} = 195$ GeV; quad Jet : $p_T^{min} = 80$ GeV). In the $t\bar{t}$ case, pronounced enhancement is observed in the single jet and in the two-photon HLT path.

---

[7] Although i)the generator is different (SPYTHIA versus ISAPYTHIA), ii) jet and missing energy reconstruction performace is different compared to ORCA in the version of CMSSW used, the SUSY event characteristic signature of large missing energy and mutlijets provides a robust selection.
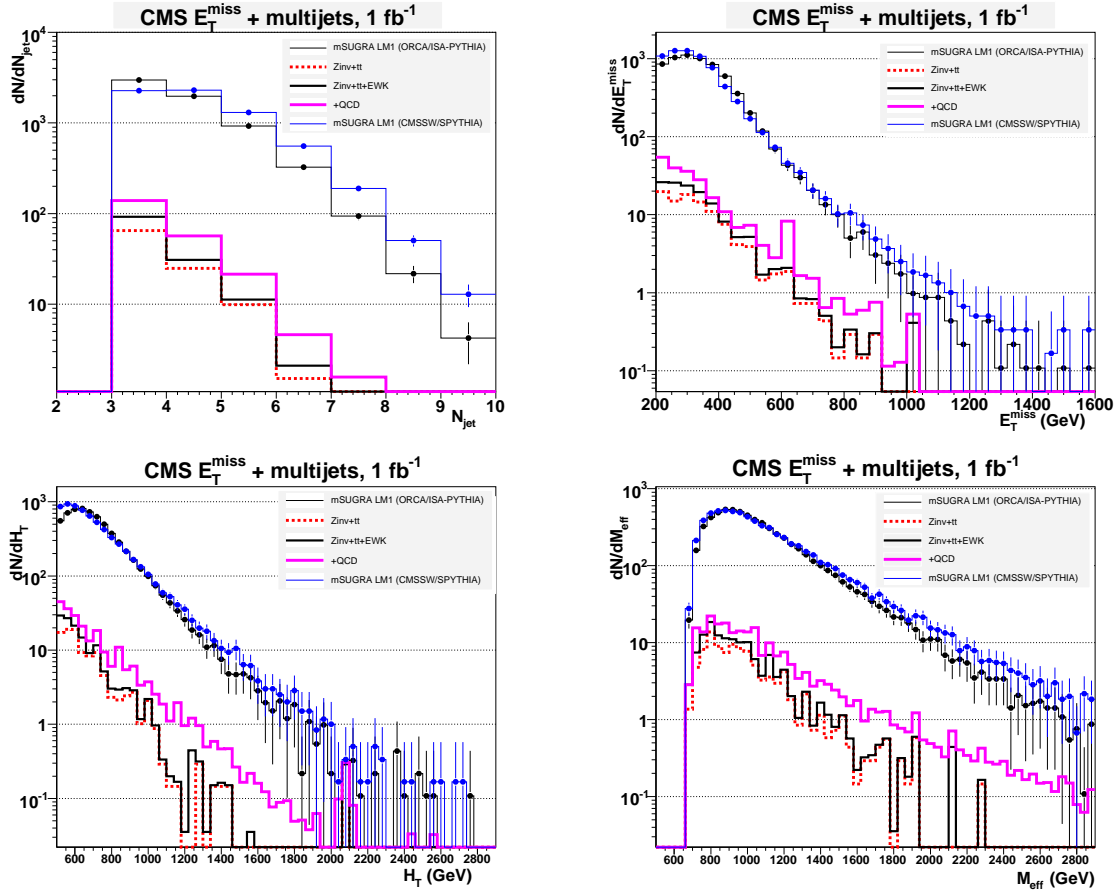
Figure 95: Distribution of $NJet$, E_T^{miss}, H_T and M_{eff} for LM1 events after application of the "E_T^{miss} + multi-jet SUSY search" selection, normalized to 1 fb^{-1} and compared to the ORCA results from the Physics TDR.

Table 16: The HLT trigger paths.

HLT bit	HLT path	HLT bit	HLT path
0	p2e	7	p1j
1	p3j	8	p1m
2	p2g	9	p2j
3	p1	10	p2t
4	p1e	11	p1t
5	p1g	12	p4j
6	p2m		

13.3.14 SUSY Exercise using di-τ final states at mSUGRA point LM1

Computing and Software

This analysis uses 322k events generated at the LM1 test point [3] corresponding to an integrated luminosity of 6.7 fb^{-1}. Those events were extracted from the so-called ExoticSoup sample (CSA06-106-os-ExoticSoup-0) using the LM1 filter (826k). In addition 4.7M inclusive $t\bar{t}$ events, 736k QCD events and 3.8M events from the Electroweak soup (CSA06-106-os-TTbar-0, CSA06-106-os-Jets-0 and CSA06-106-os-EWKSoup-0 respectively) are used to estimate the background contamination to the signal. The analysis is performed with CMSSW_1_0_6 at the CERN Tier 1 on the RECO samples. No sample access problem was encountered.

Analysis

The analysis focuses on di-τ production in the following cascades :

- $\tilde{g} \rightarrow q\tilde{q} \rightarrow qq\tilde{\chi}_2^0 \rightarrow qq\tau\tilde{\tau} \rightarrow qq\tau\tau\tilde{\chi}_1^0$,
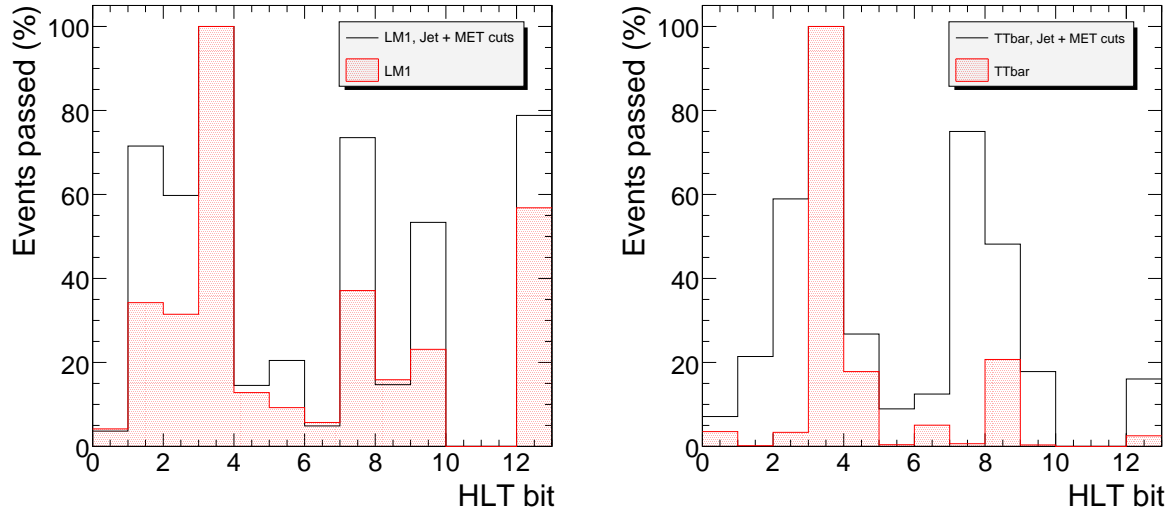
Figure 96: The HLT bits for the LM1 and $t\bar{t}$ event sample, before and after the application of the SUSY search E_T^{miss} + *multi-jet* selection path.

- $\tilde{q} \rightarrow q\tilde{\chi}_2^0 \rightarrow q\tau\tilde{\tau} \rightarrow q\tau\tau \, \tilde{\chi}_1^0$.

The mass spectra generated by SPYTHIA (used for the CSA06 data production) and ISASUGRA 7.69 (used in the Physics TDR) are compared and the main differences are found between the squark masses. They are 30 GeV lower when generated with SPYTHIA. The mass of the $\tilde{\chi}_2^0$ is also found to be lower by 5 GeV. This results in a lower E_T distribution of the τ-jet from the neutralino decay chain and shifts the di-τ invariant mass distribution compared to the study with ISASUGRA in the Physics TDR. The E_T^{miss} distribution for the generated particles (shown in Fig. 97 together with the reconstructed one) is very similar to the one obtained with ISASUGRA.

The τ reconstruction used here is similar to the one described in the Physics TDR corresponding analysis. The efficiency and purity obtained with the LM1 sample as a function of the E_T of the τ candidates are shown in Fig. 97. Discrimination between τ's decaying hadronically and leptonically is performed by comparing the energy deposited in the ECAL and the HCAL with the P_T of the track and removing the regions where the τ's are found to decay predominantly to muons and electrons as shown in Fig. 98. A large number of the τ candidates appear to have no energy deposit in the HCAL ($\sim$95% of the electronic τ's and a non-negligeable fraction of the hadronic τ's) which should be corrected in next versions of CMSSW.

The transverse energy of a τ candidate is obtained by reconstructing all the basic clusters within the candidate and identifying among them those which belong to a track ($\Delta R(\mathrm{track}, \mathrm{cluster}) < 0.015$ where the track position is taken at its entry point into the calorimeter). These clusters are then replaced by the corresponding tracks. This leads to an improvement of the overall energy resolution of the τ candidates as shown in Fig. 99.

Jets are reconstructed with the iterative cone jet-algorithm using a cone size of $\Delta R=0.5$. They are calibrated using the "γ+Jet" ORCA based calibrations. Furthermore, the HCAL towers are re-weighted in order to correct a mis-calibration of the recHits in the Endcap region of the HCAL in this version of CMSSW. The E_T^{miss} is obtained using all the reconstructed jets having an E_T higher than 15 GeV and calibrated with the so-called "Type-1" correction.

To retain the signature with two opposite sign τ's in the final state the following requirements are applied:

- E_T^{miss} larger than 150 GeV,

- at least two hadronic τ candidates,

- at least two jets with E_T larger than 150 GeV,

- $\Delta R(\tau, \tau) < 2$.

The fractions of the LM1 sample and of its main backgrounds remaining after each requirement are given in Table 17 in terms of cross section. Based on these numbers the ratio S/B is 0.96 and a 5σ discovery could be possible with 380 pb^{-1} which is at least four times higher than in the ORCA analysis.
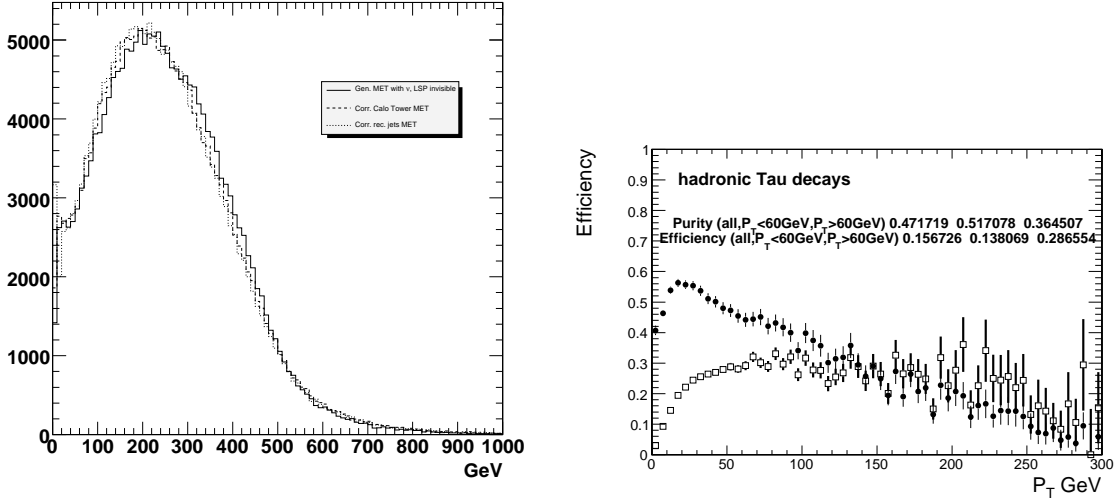
Figure 97: (left) E_T^{miss} distributions obtained with generated particles, CaloTowers and jets (right) Purity (full dot) and efficiency (open square) for hadronic τ's as function of the E_T of the τ candidates

Table 17: Remaining cross section for LM1 and Standard Model physics processes after each requirement.

Physics process	no sel. (pb)	$E_T^{\mathrm{miss}} > 150 GeV$ (pb)	2 τ cand. (pb)	2 Jets$^{\mathrm{E_T}>150\mathrm{GeV}}$ (pb)	$\Delta R(\tau,\tau) < 2$ (pb)
LM1 sample	47.67	35.2	0.427	0.166	0.084
QCD $\hat{p}_T < 170 GeV$	57×10^{10}	16.7	0	0	0
QCD $170 < \hat{p}_T < 470 GeV$	132515	197.4	0.94	0	0
QCD $470 < \hat{p}_T < 1000 GeV$	923.7	54	0.2	0.2	0.075
QCD $\hat{p}_T > 1000 GeV$	12.24	2.8	0.0066	0.003	0.0008
$t\bar{t}$ inclusive	489.7	18.26	0.174	0.0125	0.00446
EWK soup	22490	54	0.45	0.0116	0

The end-point analysis and sparticle mass reconstruction is performed as in the Physics TDR analysis. The invariant mass distributions are obtained by combining the two most energetic jets of the event to pairs of hadronic τ candidates. Two sets of invariant mass distributions are built, one considering only pairs of oppositely charged τ's (susceptible to contain signal cascades), the other made with pairs of τ's of same charge (containing only combinatorial background). Contamination from $t\bar{t}$ background is taken into account.

The end-points are obtained from the fits of the various invariant mass distributions taking into account the large combinatorial background due to multiple τ and jet associations as well as the remaining physics backgrounds (see the Physics TDR [3] and references therein for details). Examples of invariant mass distributions are shown in Figs. 100, 101, 102 and 103 including the combinatorial background and after it has been removed. The two physically allowed mass spectra obtained from these end-points are shown in Table 18. The most probable mass hypothesis is then chosen as the one for which the E_5 calculated from the mass spectra is the closest to the measured one. The one corresponding to the LM1 mass hierarchy is the closest to the measured end-point value. The measured mass spectra is also found to be in good agreement with its theoretical counterparts.
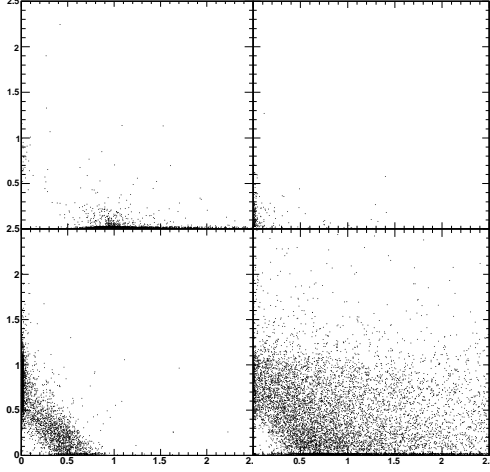
Figure 98: Fractions of the track energy deposited in the HCAL and the ECAL for various τ decay modes: (top left e, top right : μ, bottom left : π, bottom right : multi-hadrons)
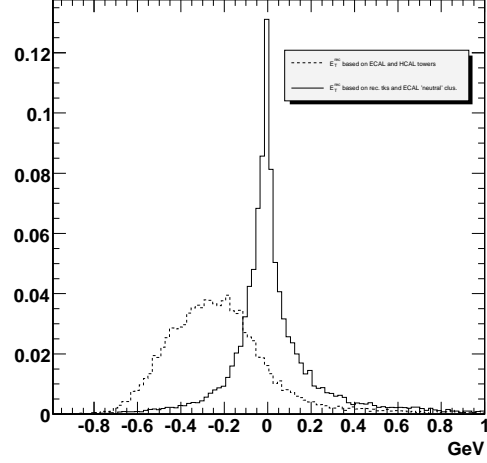
Figure 99: E_T resolution of hadronic τ candidates ($E_T^{\mathrm{gen}} < 80 GeV$)

Table 18: End-points obtained in the context of the CSA06 exercise with a log-normal fit and sparticle masses measured with the end-point technique for LM1 for integrated luminosities around 6.7 fb^{-1} . Only statistical uncertainties are shown.

End-points (GeV)	case 1 (GeV)	case 2 (GeV)
$m(\tau_1\tau_2)^{\mathrm{max}} = 70 \pm 3$	$M(\tilde{\chi}_1^0) = 176 \pm 21$	$M(\tilde{\chi}_1^0) = 108 \pm 24$
$m(\tau_1 q)^{\mathrm{max}} = 410 \pm 14$	$M(\tilde{\chi}_2^0) = 260 \pm 22$	$M(\tilde{\chi}_2^0) = 188 \pm 11$
$m(\tau_2 q)^{\mathrm{max}} = 244 \pm 10$	$M(\tilde{\tau}) = 238 \pm 22$	$M(\tilde{\tau}) = 124 \pm 15$
$m(\tau_1\tau_2 q)^{\mathrm{max}} = 449 \pm 14$	$M(\tilde{q}) = 663 \pm 29$	$M(\tilde{q}) = 579 \pm 51$
$E_5^{meas} = 571 \pm 30$	$E_5^{calc} = 620 \pm 29$	$E_5^{calc} = 591 \pm 7$

13.3.15 SUSY exercise for electron cleanup at mSUGRA point LM1

Computing and Software

The `ExoticsSoup` dataset and LM1 filter skim was used in this exercise. The destination of the skim was the Tier-2 CSCS,at Manno, Switzerland but the skimmed data never arrived to the intended Tier-2. However the skim was reached at FNAL and CERN. `CMSSW_1_1_1` was used for the analysis and the `SusyAnalyzer` framework. No crashes or other problems were observed. The following objects were selected: electrons (`SiStrip`), muons (`Global`), jets (`IC0.5`) and missing tranverse energy from jet recoil.

Analysis

The analysis scope was to access the electrons, muons and jets and perform the cleanup of electrons. No HLT trigger path is used. It was found that there were problems in accessing the electron track and H/E fraction which were solved in the version `CMSSW_1_2_0`. The access of the other objects was non-problematic. After basic requirements on the electron candidate E/P and H/E, removal of duplicate electrons and tracker isolation the
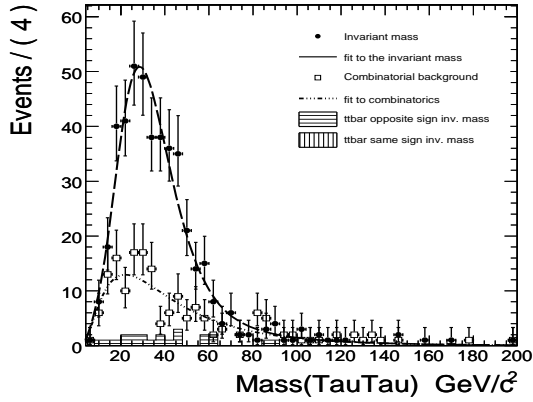
87

Figure 100: Di-τ Invariant mass distributions for signal and combinatorics together with the fitting functions for LM1.
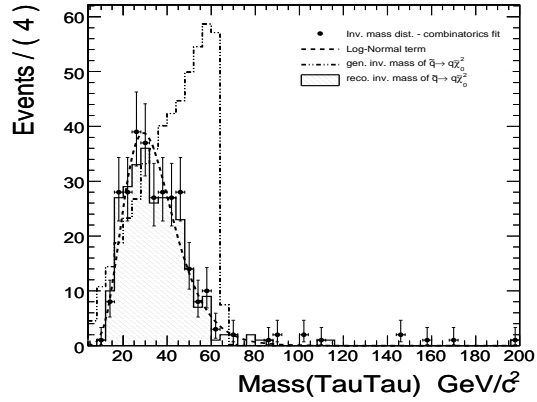

Figure 101: Di-τ invariant mass distribution of the signal obtained after subtracting the combinatorial fit together with the log-normal term used in the invariant mass fit.
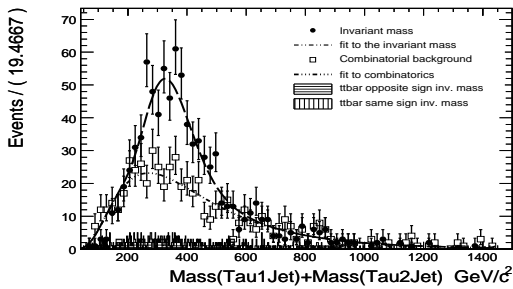

Figure 102: Distributions of $m(\tau_1 q) + m(\tau_2 q)$ for both signal and combinatorics together with the fitting functions for LM1.
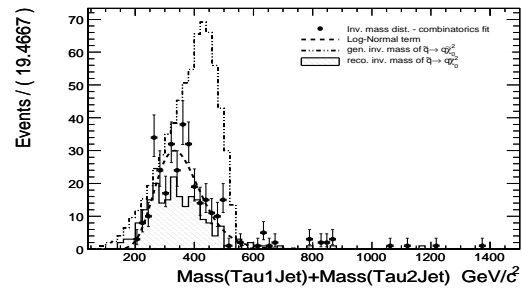

Figure 103: $m(\tau_1 q) + m(\tau_2 q)$ invariant mass distribution of the signal obtained after subtracting the combinatorial fit together with the log-normal term used in the invariant mass fit.

purity is found to be 88%. The electron candidate multiplicity after the "cleanup" and for $P_T > 10$ GeV is shown in figure 104.
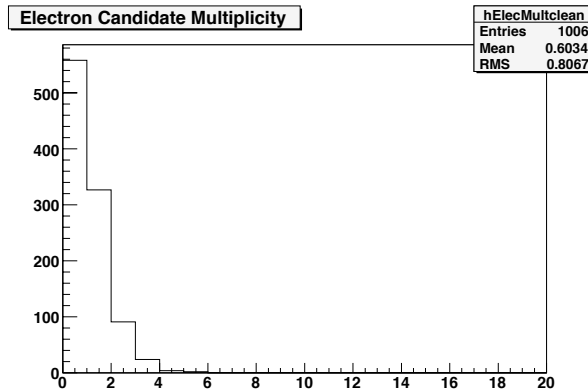

Figure 104: (The electron candidate multiplicity after the electron cleanup.

13.3.16   SUSY Exercise with b-tagging at mSUGRA point LM1

Computing and Software

The dataset to be used was CMSSW_1_0_6-su_Exotics_LM1Filter, a subset of the CSA06-106-os-ExoticSoup0-0 dataset, which was to be generated and placed on the CalTech Tier-2 site. However, this dataset was never sent to Caltech so my analysis was done using the data on the FNAL Tier-1. This exercise started late in the timeline of the CSA06 project and the work descibed was done two weeks before the end of the project.

The CMSSW_1_0_6-su_Exotics_LM1Filter dataset was one of the last sets to be generated and thus was not initially available. As an exercise, I generated a "skim" from 1000 events using the LM1Filter.cfg from the CVS repository and compared the number of filtered events to the expected 39%. My skim contained 359 events (35.9%). Repeating the skim on 2000 events generated 752 events (37.6%).

10000 events were then analyzed from the LM1Filter dataset stored at FNAL using CRAB and CMSSW_1_2_0_pre5.
Analysis

Jet flavor tagging was done using the `TrackCounting` class with the `JetTracksAssociator` and `PrimaryVerticesProduce` The `JetFlavourIdentifier` class was used to distinguish jet flavours. The results are shown in figure 105.
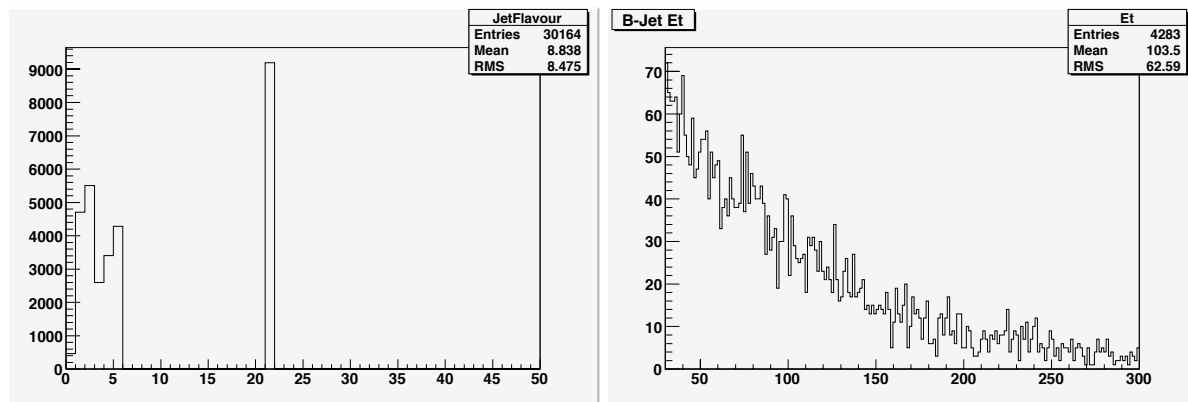

Figure 105: (left)The jet flavor in jets at mSUGRA point LM1 (right) the b-taggedjet E_T.

Jet types and values are as follows:

Further analysis beyond the CSA06 project is planned including determining b-taggings efficiency and work towards stop and sbottom content in the sample.

Table 19: Jet tags

Bin	Flavour	#
1	u	4709
2	d	5509
3	c	2600
4	s	3408
5	b	4283
21	gluon	9188

13.3.17 $Z' \to \mu^+\mu^-$ Exercise

The Z' analysis is based on the Exotics Soup with the `Zprime` Pythia particle ID Skim Filter. The computing and software details are described in section 13.3.13. In this sample the $Z' \to \mu^+\mu^-$ invariant mass is reconstructed using the two leading opposite sign muons with $p_T > 20$ GeV. The Z' branching fraction to the $\mu^+\mu^-$ channel as extracted from the total Z' sample was 3.1 %. Fig. 106 shows the generated $Z' \to \mu^+\mu^-$ invariant mass distribution. The corresponding invariant mass distribution from the reconstructed muon pairs is displayed in Fig. 107. A Z' reconstruction efficiency of 64 % is obtained. The efficiency decreases to 57 % with a 600 to 800 GeV mass window requirement imposed on the generated and reconstructed invariant mass. Additionaly the JINR group accessed the dataset `CMSSW_1_0_6-su_Exotics_ZprimeFilter-1161788458` at FNAL using CRAB and 274429 events were selected and reconstructed with `CMSSW_1_1_2`. Results for the events with $Z' \to \mu^+\mu^-$ decays are summarized in figures 108-109.

13.3.18   Dijet Mass Exercise

Computing and Software

The CSA06 QCD DataSets were produced using the QCD Skims of the CSA06 Jet Soup divided in range of pT-hat (hard collision transverse momemtum). The DataSets used to perform the dijet template analysis are summarized in Tab. 20.

Table 20: QCD dataSets and locations

Pt_hat (GeV)	DataSet	location
0_15	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_0-15-1161127916	cmssrm.fnal.gov
15_20	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_15_20-1161127916	cmssrm.fnal.gov
20_30	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_20_30-1161127916	cmssrm.fnal.gov
30_50	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_30_50-1161127916	cmssrm.fnal.gov
50_80	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_50_80-1161127916	cmssrm.fnal.gov
80_120	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_80-120-1161127916	cmssrm.fnal.gov
120_170	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_120-170-1161127916	cmssrm.fnal.gov
170_230	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_170-230-1161127916	cmssrm.fnal.gov
230_300	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_230-300-1161127916	cmssrm.fnal.gov
300_380	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_300-380-1161127916	cmssrm.fnal.gov
380_470	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_380-470-1161127916	cmssrm.fnal.gov
470_600	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_470-600-1161127916	cmssrm.fnal.gov
600_800	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_600-800-1161127916	cmssrm.fnal.gov
800_1000	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_800-1000-1161127916	cmssrm.fnal.gov
1000_1400	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_1000-1400-1161127916	cmssrm.fnal.gov
1400_1800	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_1400-1800-1161127916	cmssrm.fnal.gov
1800_2200	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_1800-2200-1161127916	cmssrm.fnal.gov
2200_2600	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_2200-2600-1161127916	cmssrm.fnal.gov
2600_3000	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_2600-3000-1161127916	cmssrm.fnal.gov
3000_3500	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_3000-3500-1161127916	cmssrm.fnal.gov
3500_40000	/CSA06-103-os-Jets0-0/FEVT/CMSSW_1_0_4-su_QCDSKIM_jet_3500-40000-1161127916	cmssrm.fnal.gov

These DataSets are available via CRAB, with DataSet names registered in the DBS/DLS lookup web pages. They were accessed by means of CRAB (version _1_4_0_pre1_IIHE) using CMSSW_1_1_1. The major problem encountered was the abortion of the jobs in the E_HOST, cmslcgce.fnal.gov. The same problem was encountered with CRAB_1_4_1_IIHE and CRAB_1_4_2_IIHE. To perform the analysis the jobs run in pnfs using the condor batch at Fermilab.

The dijet resonance datasets were produced usings the skims of the Exotic Soup dataset. They were produced by running filters on the inclusive Z' (700 GeV/c^2) and q^* (400 GeV/c^2)data samples and selecting out the dijet processes based on generator level information. For the $Z' \to q\bar{q}$ process 22 291 events were processed and analyzed while for the $q^* \to qg$ 22 457 events were processed and analysed. For both samples more data was available than were analyzed.

A simple EDAnalyzer and a ROOT macro was used to analyze the data and form the dijet invariant mass distributions in all samples.
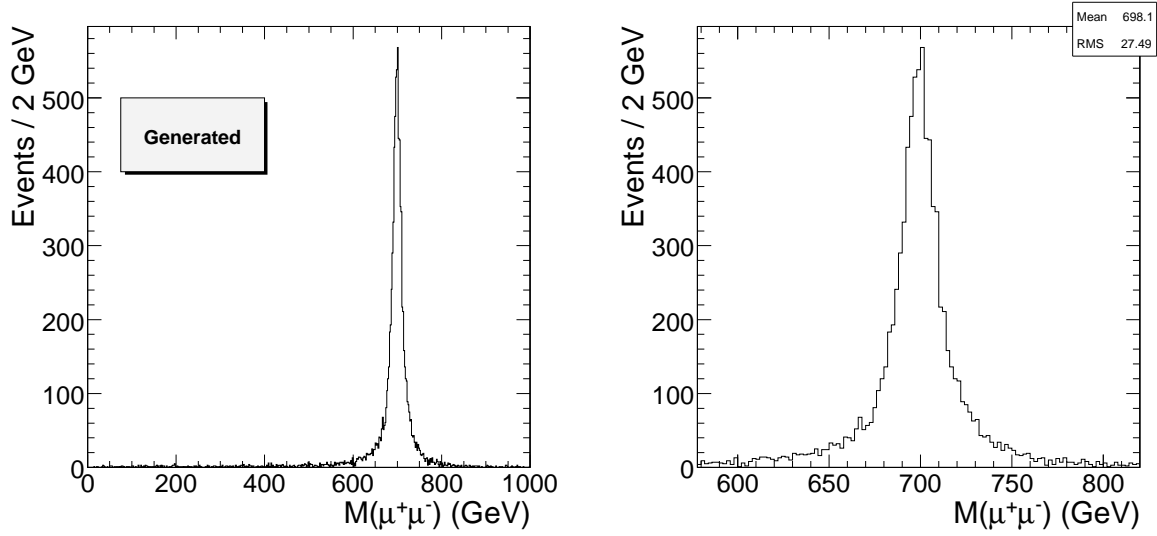
Analysis

Figure 106: The generated $Z' \rightarrow \mu^+\mu^-$ invariant mass distribution.
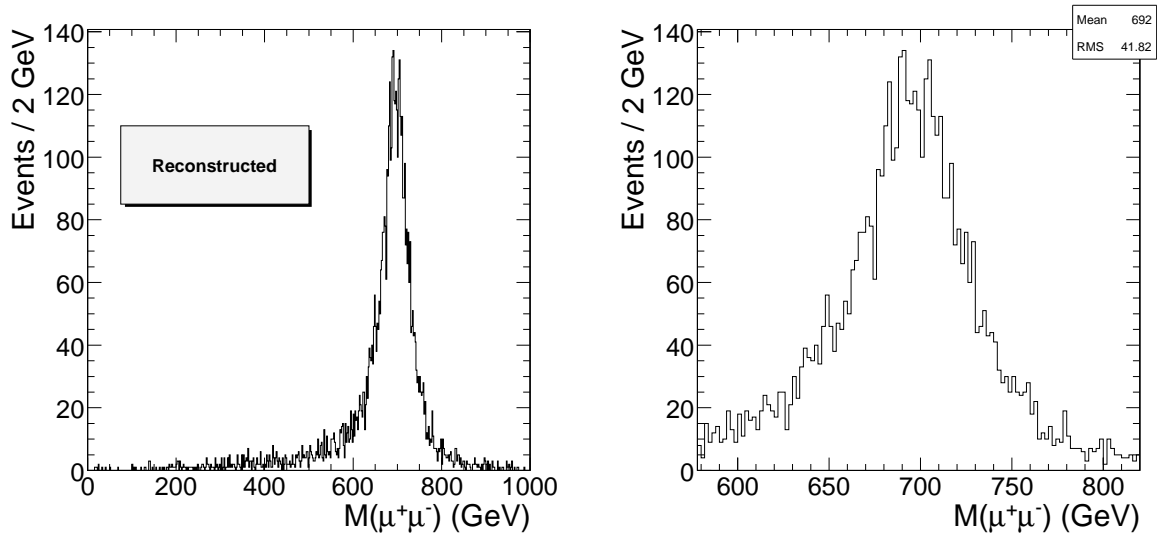

Figure 107: The reconstructed $Z' \rightarrow \mu^+\mu^-$ invariant mass distribution using two leading muons.

The invariant mass distribution of the the two leading jets in the QCD samples and for two jet collections, the GenJets (generator level) and CaloJets (reco level from CaloTowers) for a total of $922\,791$ events is shown in Fig. 110. The distribution is rebinned to take in account the variable resolution, and filled with reweighted entries. The weights on the entries are directly proportional to the cross sections and inversely proportional to the number of events processed. Finally a conversion factor is applied to have the distribution in pb.

The dijet invariant mass using the two leading jets with $|\eta| < 1.0$ for the Z' dijet sample is shown in Fig. 111 and for the q^* dijet sample in Fig. 112. Three jet algorithms were used, the midpoint with cone size $R = 0.5$, the iterative with cone size $R = 0.5$ and K_T with $D = 1.0$. The inputs were either CaloTowers (CaloJets), or HepMCparticles (GenJets). For the cone algorithms, jet corrections available from ORCA data were used to form corrected CaloJets (CorJets).

13.3.19   High Energy Electron Pair Exercise

Computing and Software

The Exotic soup samples at various processing levels were accessed:

- DIGI: /CSA06-083-os-ExoticSoup/GEN/CMSSW_0_8_3-GEN-SIM-DIGI-HLT-1156877641-merged;
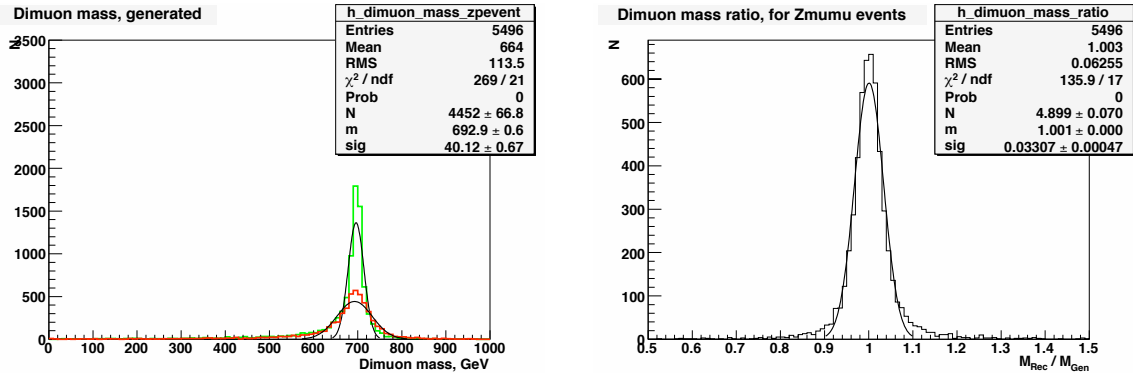
91

Figure 108: (left)Generated and reconstructed dimuon invariant mass and (right) their ratio.

local fragments produced at Belgian Tier-2

- RECO:

    - /CSA06-103-os-ExoticSoup0-0/RECO/...
      CMSSW_1_0_3-RECO_Hcc50df9a16717df4367a80c47fe190b8; at CERN

    - /CSA06-106-os-ExoticSoup0-0/RECO/...
      CMSSW_1_0_6-RECO_Hee5e8972ca2b9fe5859521c27772ba77; at Belgian Tier-2 (transferred from FNAL); at CERN

- skimmed RECO: /CSA06-103-os-ExoticSoup0-0/FEVT/...
  CMSSW_1_0_6-su_BSM_mc2e_Filter-1161788458; at Belgian Tier-2 (transferred from FNAL)

- HLT: /CSA06-083-os-ExoticSoup/HLT/...
  CMSSW_0_8_3-GEN-SIM-DIGI-HLT-1156877641-merged; at Belgian Tier-2

As an additional computing exercise, the files below were transferred.

- RECO:

    - /CSA06-106-os-EWKSoup0-0/RECO/...
      CMSSW_1_0_6-RECO_Hee5e8972ca2b9fe5859521c27772ba77; at RAL Tier-2

    - /CSA06-106-os-ExoticSoup0-0/RECO/...
      CMSSW_1_0_6-RECO_Hee5e8972ca2b9fe5859521c27772ba77; at RAL Tier-1

- skimmed RECO: /CSA06-103-os-ExoticSoup0-0/FEVT/...
  CMSSW_1_0_4-su_BSM_mc2e_Filter-1161045561; at RAL Tier-1 and Tier-2

- AOD: /CSA06-106-os-ExoticSoup0-0/AOD/...
  CMSSW_1_0_6-AODSIM_H15a59ba7b4c3d9e291172f60a399301f; at RAL Tier-1

Before CSA06-specific samples were produced or made available in DBS, the standard $Z \to e^+ e^-$ samples were used for validation in order to speed up the preparation of event selection, skimming and analysis.

Files containing the whole FEVT were processed and analyzed.

The mc2e skim filter was prepared and tested. This filter selects events with 2 electrons generated with $p_T > 15 \text{GeV/c}$ and $|\eta| < 3$. The mc2e filter was tested on the above-mentioned subset of the DIGI ExoticSoup sample, that was produced on the Belgian Tier-2. The $|\eta| < 3$ requirement leads to a selection inefficiency of 5%. Out of $10\,000$ ExoticSoup events, the filter retained 120 events, which is compatible with the proportion of Z' events simulated in the Exotic soup (39%) multiplied by the branching ratio $BR(Z' \to e^+ e^-) = 3.36\%$ times the selection efficiency of 95%.
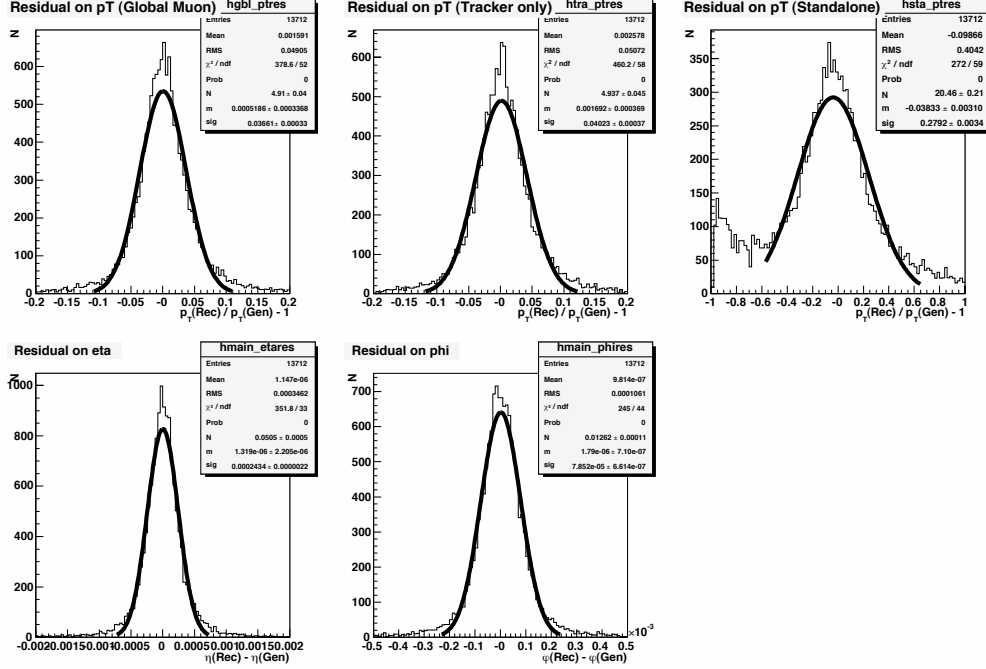
The samples were accessed in two ways:

Figure 109: The residuals of p_T, η and ϕ for single muons in the Z' sample. The p_T residuals are given for Global, Tracker-only and StandAlone muon reconstruction.

- Direct access to the Tier-2 storage: fast prototyping of event selection and analysis code is best done running CMSSW interactively.

- Remote access with CRAB. The most recent CRAB versions and prereleases were needed in order to access the samples as soon as they were made available in DBS. This appears due to incompatibilities between DBS and CRAB versions. These were solved by the CRAB team in a matter of days, but we believe that such a coupling should be avoided.

Early prototyping using the `CMSSW_0_8_3` release validation samples was done with `CRAB_1_3_0_pre1`. Later, `CRAB_1_4_0_pre1`, `CRAB_1_4_0` and `CRAB_1_4_1` were used.

The following problems were encountered:

- incompatibilities between DBS versions and CRAB versions; could have been severe but were solved by the CRAB team in a matter of days;

- publication of RECO and skimmed samples for the ExoticSoup 2-3 weeks late with respect to the initial CSA06 schedule;

- mismatches between a CMSSW version used to develop analysis code, and the version installed at some sites hosting the wanted data, were not detected by CRAB before job submission. This was traceable from the CRAB job error messages. The fix consists in white-listing a site with matching resources.

- in case of resource mismatch (either data or CMSSW version), debugging information provided was often unclear. In addition, as no simple compilation of the user requirements into a JDL file or similar was provided from the CRAB configuration file, it was difficult to independently investigate job submission errors.

- the feature to begin an analysis at an arbitrary event number within the chosen dataset was absent.
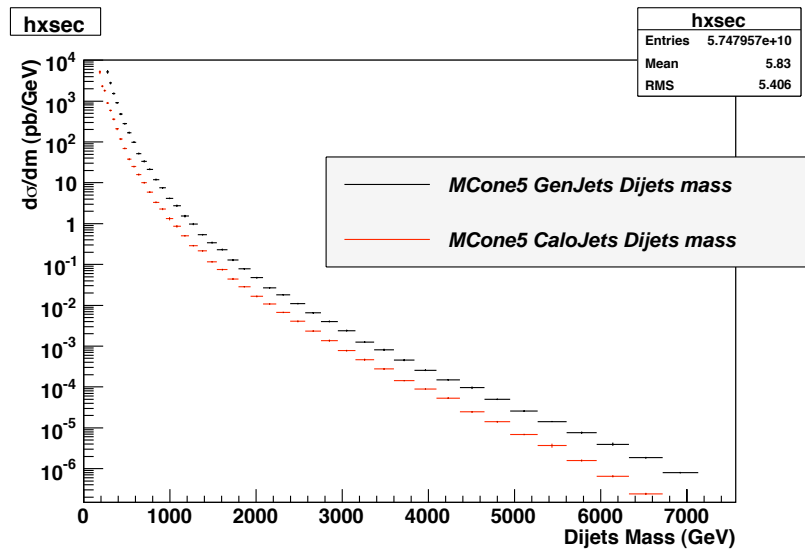
Figure 110: QCD spectrum rebinned and reweighted from GenJets and CaloJets collections.
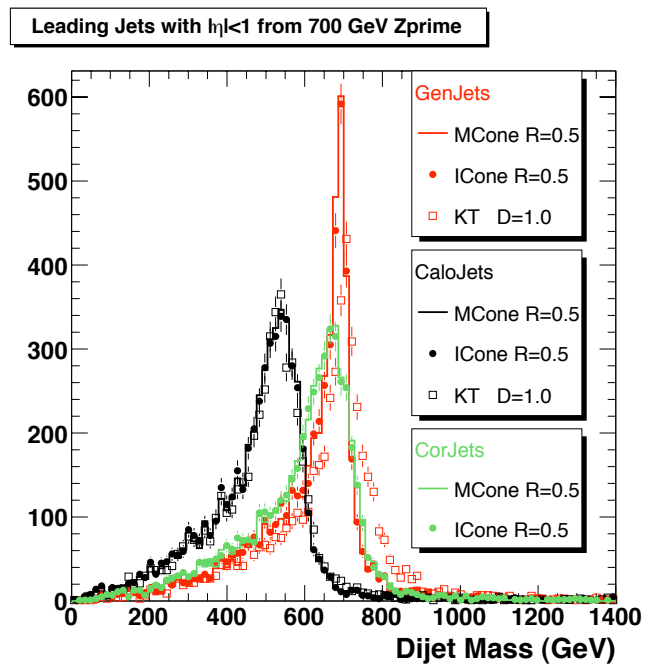

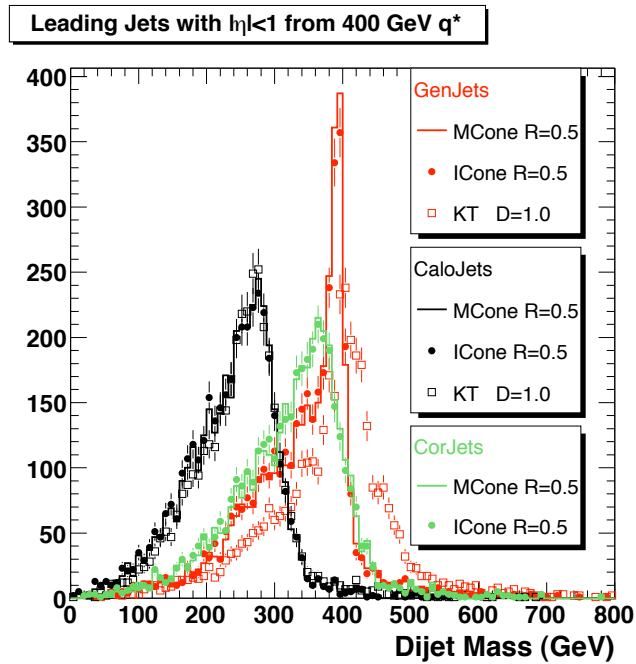Figure 111: Dijet inavriant mass for the Z' skim.

Figure 112: Dijet invariant mass in the excited quark skim.

- the management of records of submitted jobs by CRAB was found to be ineffective. When managing a large number of submitted jobs it was found to be easier to extract the jobId from the CRAB log and record this manually, querying job status manually with the edg tools than using `crab -status`.

- the **mc2e** skimmed sample transferred to the Belgian Tier-2 was not properly closed, and a request had to be made to have it closed and published in DBS.

The following CMSSW versions were used: `CMSSW_0_8_1`, `CMSSW_0_8_3`, `CMSSW_1_0_1`, `CMSSW_1_0_4`, `CMSSW_1_0_5`.

The following objects were used in this exercise:

- MC generator information: HepMC.

- ECAL Island SuperClusters. This collection was selected because it is provided for the whole ECAL acceptance, in contrast with the Hybrid SuperClusters which are only provided for the ECAL barrel. We note that a SuperCluster collection covering the whole acceptance is convenient. In addition a single collection would reduce confusion when performing analyses.

- generator-level jets and RECO jets, from k_T and Iterative Cone $\Delta R = 0.5$ algorithms. Jets from k_T algorithm were found problematic (some jets with 0 energy; jets not ordered in p_T).

- HLT bit information (computed from generator information in CSA06).

The purpose of the analysis was to exercise and set up the tools for an early analysis of real data. These tools comprise:

- event analyzers to access and study HepMC, ECAL SuperCluster, jet and HLT information;

- event filters to select the events to be transferred to the Tier-2's;

- possible criteria for an Express stream for high-mass resonances decaying into electrons;

- instructions for using CMSSW, the CMS grid computing environment and protocols related to the Tier-2's.
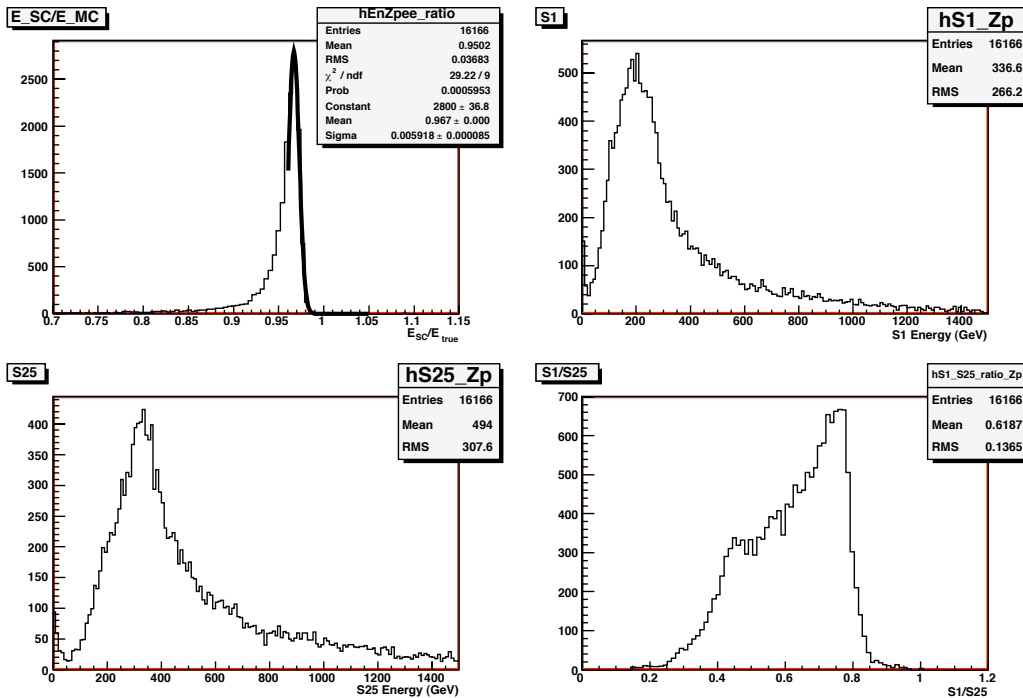
Figure 113: Distributions of energy and shape variables of the Island SuperClusters. The SC's are matched to generated electrons from Z'-decays within a cone of aperture $\Delta R = 0.1$.

Analysis

Distributions of the ECAL uncorrected Island SuperCluster (SC) energy and shape are shown in Fig. 113. The SC's are matched to the generated Z' decay electrons within a cone of $\Delta R = 0.1$. The ratio between the reconstructed SC energy and the generated electron energy is 97% in average and is compatible with previous studies of SC's from high-energy electrons. The distributions of the energy contained in the hottest cell $S1$, the 5×5 crystal matrix around the hottest cell $S25$ and their ratio $S1/S25$ are also in agreement with the distributions expected. The energy resolution, defined as the RMS of a Gaussian fit to the peak of the E_{SC}/E_{true} distribution, is 0.6% and matches the resolution expected. Corrected SC energy is not used since it is known to be inadequate for electrons of several hundreds GeV.

In CSA06 the HLT selection was faked by applying trigger threshold requirements at the generator level. The consistency of the HLT bit information with the mc2e event skimming criteria was studied and the results are shown in Fig. 114. The 2-electron HLT selection shows a slightly lower efficiency than the mc2e filter, due to the more restrictive $|\eta|$ requirement ($|\eta| < 2.4$). The efficiency of the 1-electron path (p1e) is slightly higher than the 2-electron path (p2e) efficiency where events with 1 electron beyond $|\eta| > 2.4$ are recuperated.

The high-mass di-electromagnetic object trigger is one of the Express line triggers considered for CMS [6]. In this exercise a few possible combinations of event selection criteria for a di-EM Express stream trigger (using the objects validated) are studied.

The criteria considered are:

- the presence of 2 EM SC's with $E_T > 15$ GeV (2SC);

- the presence of 2 EM SC's with different transverse energy cuts at $E_T > (100, 15)$ GeV (2AsymSC);

- the presence of 2 EM SC's with $E_T > (100, 15)$ GeV, the one with largest E_T being associated to a jet with $E_{jet}/E_{SC} < 1.1$ (2AsymSC-HoEVeto). This criterion implements a purely calorimetric (ECAL + HCAL) selection of EM objects.

In Fig. 115 the invariant mass spectra for events passing the different selection criteria, computed out of the 2
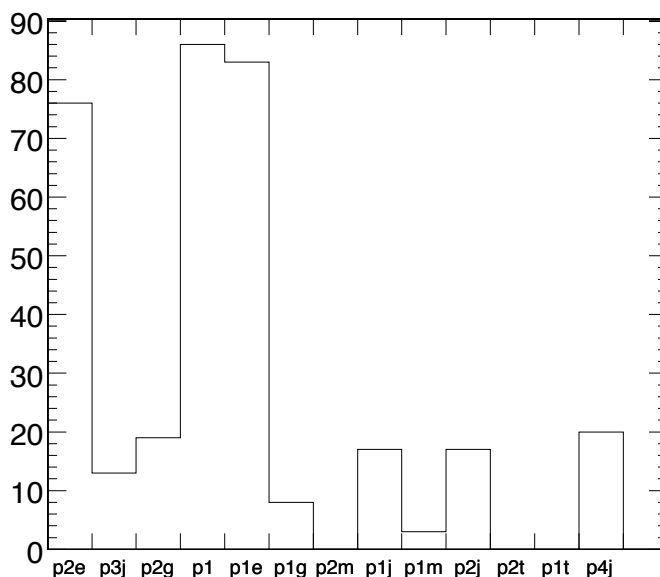
Figure 114: HLT conditions met by the events selected by the `mc2e` skim filter.

uncorrected SC's with largest E_T, are shown. The vertex is taken at (0, 0, 0). The efficiency of the criteria considered is higher than 80%. The most promising criterion is the calorimetric (ECAL + HCAL) selection of EM objects, yielding a purity around 75% for a cut at $E_{jet}/E_{SC} < 1.1$. Table 13.3.19 summarizes the efficiencies and purities for $Z' \rightarrow e^+ e^-$ decays.

Table 21: Number of events analyzed from the ExoticSoup sample, and number of events retained by the 2SC, 2AsymSC and 2AsymSC-HoEVeto criteria (see text for a detailed definition of the criteria).

	Nb. events	Nb. $Z' \rightarrow e^+ e^-$
ExoticSoup	30864	383
2SC	23126	358
2AsymSC	8798	346
2AsymSC-HoEVeto	416	311

14   Conclusions and Lessons Learned

As a complete exercise CSA06 was extremely successful. The technical metrics were all met and some were exceeded by large factors. While there is still considerable work do to, especially in the integration with data acquisition and on-line computing, the intended functionality was demonstrated in the challenge. With the success there are valuable lessons for CMS as we transition to operations and stable running.

14.1   General

There are a number of general lessons CMS can take away from the challenge. The first is in the area of the transition to operations. CMS needs development work to ease the operations load. CSA06 was very successful, but it required a higher level of effort and attention than could reasonably be expended for an experiment running for years. As CMS transitions from development into successful scale demonstrations to stable operations it is to be expected that development activities will be identified to reduce the operational load. During the challenge several specific areas were identified that will be listed in the sections below. More fine-grained operator control was identified for several elements, while more automation was identified for others.
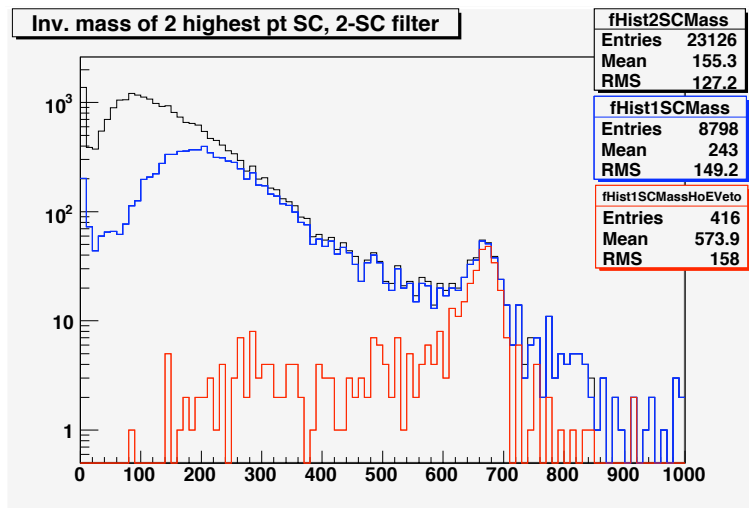
Figure 115: Distributions of the invariant mass spectra for events passing the selection criteria considered: 2 EM SC's with $E_T > 15$ GeV (black), 2 EM SC's with different transverse energy cuts at $E_T > (100, 15)$ GeV (blue) and 2 EM SC's with $E_T > (100, 15)$ GeV, the one with largest E_T being associated to a jet with $E_{jet}/E_{SC} < 1.1$ (red). The invariant mass is computed out of the 2 uncorrected SC's with largest E_T.

Another general lesson was that the strong engagement with Worldwide LHC Computing Grid (WLCG) and the computing sites themselves was extremely useful. There were a few problems encountered with grid services which were addressed very promptly. Sites generally responded to problems and solved them. An item related to the lesson about operations is the communication of problems that needed to be addressed. The sites and services were generally repaired promptly as soon as problems were identified, but frequently it was attentive operators and not automated systems that saw the problem first.

Scale testing continues to be an extremely important activity. Initial scaling issues were identified and solved in several CSA06 components. Most of the problems identified were related to components that were relatively new, or were used in at a new scale without being thoroughly tested. CMS was lucky that all of the scaling problems seen in CSA06 were straightforward to solve and were fixed promptly. CMS needs to achieve nearly a factor of four in scale for some of the components before high energy running in 2008, and the sooner scaling issues are identified the more time will be available to solve them.

14.2 Offline Software

There were three important lessons from the offline software experience of CSA06. The first is that the software in the configuration used was able to sustain greater than a 25% load for the prompt reconstruction activity. The software error rate, performance and memory footprint were well within expectations contributing to smooth running of the Tier-0 farm. The performance was somewhat faster than the time budget allows, but several of the slower and less stable reconstruction algorithms were intentionally left out of the prompt reconstruction workflow. As the reconstruction software evolves the performance and stability should be watched.

The second lesson was that the ability to promptly create and distribute a new software release was invaluable. During CSA06 operations CMS released four versions of the software to address issues that were encountered and to add functionality. The ability to make a release promptly and to equally promptly install it on the remote sites was extremely helpful in meeting challenge goals.

The last lesson is that CMS needs a more formal validation process and checklist for the application before a release is tagged. A problem in the re-reconstruction application was identified in the final two weeks of the challenge. With a more rigorous validation procedure the problem might have been seen in the opening days of the challenge giving more time to solve it. While it is impossible to test for all possible conditions, a validated list of validation checks would be useful.

14.3   Production and Grid Tools

14.3.1   Organized Processing

There are a number of lessons to take away from the experience with production and grid tools. The first is that the Prod_Agent worked well in the pre-challenge production. CMS was able to meet the very ambitious goals of 25M events per month of simulated event production. The one pass production chain contributed to the high efficiency of the production application during July and August. The production was performed by four teams, which is a decrease in operations effort over previous production exercises. CMS will need to maintain the efficiency and the flexibility as the simulation becomes more complicated for the physics validation.

The Prod_Agent infrastructure also worked well for accessing existing data and applying the user selections. The teams operating the agents were able to apply multiple selections simultaneously. The merging and data registration components worked well and could be reused from the simulated event production workflow.

Even though the exercises were successful there is clearly room for improvement. CMS needs to continue to improve automation of workflow for re-reconstruction, selection and skimming of events. The infrastructure of request, to validation, to scheduling, to large scale execution has components that involve people. The human interactions can be reduced and more automated workflows can be implemented. In the CSA06 workflows the work assignments and output destinations were conveyed by e-mail. During CSA06 the production teams were responsible for combining the skims, testing the configurations, and executing the skims.

In addition to improving the automation for scheduleable items like skimming, we also need to improve the transparency. In general users and groups need a consistent entry point to see the status of the requests and the location of the output. The request system, the data transfer system, and the dataset booking systems need to be tied together for consistent end-to-end user views.

The re-reconstruction activity was, by design, a demonstration of functionality and not a demonstration of the final production workflow. There is work left to do to make to ensuring every event is re-reconstructed and processing failures are tracked and addressed.

14.3.2   User Analysis Workflows

The largest success for the analysis workflows was the successful demonstration that the gLite and Condor-G job submission systems can achieve the goal of 50k jobs per day. Integration and scale testing continues to be very important. CMS integrated CRAB with the gLite bulk submission only shortly before the challenge began. There had been testing of the underlying infrastructure through the CMS WLCG Integration task force but no scale testing with the CMS submission system. The two problems in achieving scale were both related to the CMS implementation and not in the underlying infrastructure and both issues were promptly addressed by the CMS developers. As the number of people participating and the number of jobs increases the importance of scale testing will only increase.

In order to reach the target submission rate CMS needed to make heavy use of load generating "job robots". While the robots generate workflows that closely resemble user analysis jobs, the robots are not a substitute for an active user community for testing. For the next series of challenges CMS should ensure a larger number of individuals performing analysis.

Though only about 10% of the total job submissions, the user driven analysis in the challenge was successful with CRAB functioning well on both EGEE and OSG sites. This document highlights some of the types of analysis that were successfully completed. Nevertheless, there are a number of lessons. The first is that CMS needs to improve the user support model. Currently user support is provided by a mailing list in a community support model that works well for a size of the community currently being supported. It is not clear if this informal support will scale to the larger collaboration. It is possible for requests to fall through the cracks. CMS should look at hybrid support models that assign and track tickets while ensuring that a large enough community of people see the support requests to continue to provide a broad base of supporters.

14.4   Offline Database and Frontier

The offline database infrastructure was successful in the challenge. The calibration data could be distributed to remote locations from a single database instance at CERN using the Frontier infrastructure. The initial attempt in the Tier-0 workflow identified scaling limitations in the CMS web cache configuration for Frontier and stability

issues in the application code. Both of these were promptly addressed, but they underscore the need for validation and scale testing.

The other offline database lesson is related the way CMS stores calibration constants in the database and the frequency with which they are invalidated. Currently CMS stores the calibration information as a large number of small objects, which are treated as independent queries by the offline databases and they invalidated daily. The first application of the day can expect almost an hour updating the database information in the offline cache, which is not reasonable in the long term.

14.5   Data Management

The CMS data management solution relying on central components for data bookkeeping, data location, and data transfer management and site components for data resolution worked well and reduced the effort required by the site operators. The changes in the CMS event data modem significantly simplified the access of the data by analysis applications.

The general lesson from data management is that CMS needs to ensure that all the data management components have a consistent picture of the data. The synchronization of the various views needs to be better automated. CMS has data management information in the dataset bookkeeping system (DBS), the data transfer system (PhEDEx), and the dataset location service. CMS was able to fall out of sync in the various data management components. Maintaining consistency currently involves some manual operations.

A specific element that was identified in CSA06 was the need to examine the DBS performance in the presence of merging output. The initial performance needs estimates did not include this use-case, which introduces a heavy load on the DBS. For many output streams the performance of the bookkeeping system limited the performance to prepare data selections. The performance limitation is being addressed in the next generation of the DBS.

Data publication and the trivial file catalog resolution of the logical to physical file names both worked well. The trivial file catalog scaled well and applications were able to consistently discover data file locations with minimal additional services required at the sites.

14.6   Workflow Management

Workflow management components both at CERN and at remote centers were able to perform the achieve the required level of activity expected in the challenge. There is some overlap in the implementation of the Tier-0 workflow and the Prod_Agent workflow used at the Tier-1 and Tier-2 centers, which should be re-examined after the challenge with an eye for long term maintainability and support.

14.7   Central Services

Central services and facilities at CERN from IT and WLCG, including the batch resources and FTS, were carefully monitored and problems were solved. CASTOR support at CERN was excellent. As an export system, CASTOR2 performed at a higher rate and more stably than in past CMS exercises. CMS ran into an issue with the SRM release for files greater than 2GB in DPM file which was solved the next day.

14.8   Tier-0

The Tier-0 workflow and dataflow management tools performed better than required for CSA06, showing no significant problems throughout the challenge. The flexibility of the message-based architecture allowed adaptation of the running system to the changing operational conditions, as the challenge progressed, without any interruption of service.

No inherent scaling problems were found, and key Tier-0 components (hardware, software and people) were far from being stressed during the challenge. The system achieved the low latency response required for real data-taking.

Most of the full range of complexity of the final system was explored during the challenge. Other aspects were already explored with the "July prototype". The design of the Tier-0 can therefore be deemed validated.

Operationally, the Tier-0 can be installed, configured, and run by non-experts already. The Tier-0 internal goals of exploring the operations during CSA06 have therefore also been met.

14.9 Tier-1

While 6 of the 7 Tier-1 centers met the complete goals for full participating in the challenge with successful transfers on 90% of the days. The transfer quality, defined in CMS as the number of times a transfer was attempted before successful, was significantly improved for CERN to Tier-1 transfers during the challenge as compared to previous service challenge exercises. There are several elements to improve in the final year of experiment preparation. Several Tier-1 centers had problems importing and exporting data simultaneously. The Tier-1 centers either experienced unstable data export or limited performance. The majority of Tier-1 sites demonstrated successful migration of data to tape, but there is a substantial work left to demonstrate CMS can write the full data rate to tape at Tier-1 centers and serve the data to all Tier-2 centers when requested.

A specific technical item was identified in the FTS timeouts too tight for sites with low access bandwidth and high latency when CMS moved to files that were larger than 4GB. In the final experiment the raw data files should be between 5GB and 10GB, so CMS will need to revisit the transfer timeouts again.

14.10 Tier-2

The number of Tier-2 centers participating in the challenge was larger than the original goal and a broader variety of activities was successfully performed by the Tier-2 centers. An item to improve is the amount of effort required to make Tier-2 transfers work. Some sites accepted data only from particular sites. Early in the challenge PhEDEx dynamic Routing led to unpredictable Tier-1-to-Tier-2 paths through intermediate Tier-1s. Early in the challenge the PhEDEx operations team modified the path cost metrics in PhEDEx to avoid multi-hop transfers and make the route more static and prescribed, which makes transfers more look like baseline computing model.

The poor transfer quality on the PhEDEx monitoring plot is not necessarily a Tier-2 site issue. Some Tier-1s could better import data from Tier-0 than export to Tier-2s. One item that was identified is that lots of transfer requests could clog the queues and lead to component failures. The FTS system is designed to throttle transfer requests but developers initially focusing on protection of import rather than export. CMS is continuing the discussion on architecture and implementation of throttling in FTS with the developers.

One area where the general lesson about operations load was felt the strongest was the data management at the Tier-2 centers. The data stored at a Tier-2 center is defined by the supported community and a clear need for tools to allow the Tier-2 to control the resident data was identified during the challenge.

References

[1] "The CMS Computing Model," edited by C. Grandi, D. Stickland, L. Taylor, CMS NOTE 2004-031 (2004).

[2] G.L. Bayatian et al. (CMS Collaboration), "The CMS Computing Project Technical Design Report" CERN/LHCC 2005-023 (2005).

[3] G.L. Bayatian et al. (CMS Collaboration), "The CMS Physics Technical Design Report, Volume 2, Physics Performance," CERN/LHCC 2006-021 (2006).

[4] N. N. "Alignment and Calibration at CSA06", CMS Note in preparation.

[5] D.Futyan and C.Seez, "Intercalibration of ECAL Crystals in Phi Using Symmetry of Energy Deposition," CMS Note **2002/031**, (2002).

[6] D. Acosta et al., "A T0 Architecture for the CMS experiment", CMS Note **2006/095** (2006).

[7] V. Karimaki, T. Lampen and F.-P. Schilling, "The HIP Algorithm for track based alignment and its Application to the CMS pixel detector", CMS Note **2006/018** (2006).

[8] V. Blobel and C. Kleinwort, "A New Method for High-Precision Aligment of Track Detectors," Advanced Statistical Techniques in Particle Physics, Durham, Conference Proc., arXiv-hep-ex/0208021 (2002).

[9] A. Calderón et al., "Muon system alignment with tracks," CMS Note **CMS 2006/016**, (2006).

[10] N. De Filippis et al., "Impact of CMS silicon tracker misalignment on track and vertex reconstruction", CMS Note **2006/029**, (2006).

[11] R. Fruhwirth, Nucl Instr. and Meth. A 262 (1987) 444.

[12] CMS NOTE 2006/061 V. Büge, Ch. Jung, G. Quast, A. Ghezzi, M. Malberti, T. Tabarelli de Fatis, *Prospects for the precision measurement of the W mass with the CMS detector at the LHC.* (2006)

[13] S. Gennai et. al., "Tau jet reconstruction and tagging with CMS", Eur. Phys. J. C (2006) Digital Object Identifier (DOI) 10.1140/ep jcd/s2006-02-001-y