CMS from STEP’09 to Data Taking:

CMS Computing experiences from the WLCG STEP’09 challenge to the first Data Taking of the LHC era

Oliver Gutsche
[ CMS Data Ops / STEP’09 coordination – Fermilab, US ]

Daniele Bonacorsi
[ deputy CMS Computing Coordinator / STEP’09 coordination – University of Bologna, Italy ]
CMS Computing and “steps”
# Coarse schedule

<table>
<thead>
<tr>
<th>Event</th>
<th>Dates and Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start of 7 TeV Running</td>
<td>March 26±2, 2010 (<em>proposed</em>)</td>
</tr>
<tr>
<td>ICHEP ’10 Conf.</td>
<td>July, 2010 (hopefully several pb(^{-1}) to analyze)</td>
</tr>
<tr>
<td>Shutdown for 2010 HI Run</td>
<td>mid October, 2010 (hopefully several hundred pb(^{-1}))</td>
</tr>
<tr>
<td>HI Run 2010</td>
<td>mid November 2010 ➔ mid December 2010</td>
</tr>
<tr>
<td>Technical Stop</td>
<td>December 2010 ➔ February 2011</td>
</tr>
<tr>
<td>7 TeV pp running</td>
<td>February/March 2011 ➔ October 2011 (aim to finish with at least 1 fb(^{-1}))</td>
</tr>
<tr>
<td>Heavy Ion Run 2011</td>
<td>mid November 2011 ➔ mid December 2011</td>
</tr>
</tbody>
</table>
**CMS involvement in STEP’09**

**STEP’09**: a WLCG multi-VO exercise involving LHC exps + many Tiers

CMS operated it as a “series of tests” more than as a challenge

- CCRC’08 for CMS was a successful and fully integrated challenge
- In STEP’09, CMS tested specific aspects of the computing system while overlapping with other VOs, with emphasis on:

  **T0**: data recording to tape
  - Plan to run high scale test between global cosmic data taking runs

  **T1**: pre-staging & processing
  - Simultaneous test of pre-staging and rolling processing in complete 2-week period

**Transfer tests**

- T0→T1: stress T1 tapes by importing real cosmic data from T0
- T1→T1: replicate 50 TB (AOD synchronization) between all T1s
- T1→T2: stress T1 tapes and measure latency in transfers T1 MSS → T2

**Analysis tests at T2’s**: Demonstrates capability to use 50% pledged resources with analysis jobs
CMS stores 1 ‘cold’ (archival) copy of recorded RAW+RECO data at T0 on tape

- Can CMS archive the needed tape-writing rates? What when other VO's run at the same time?

In STEP’09, CMS generated a tape-writing load at CERN, overlapping with other exps

- To maximize tape rates, CMS ran the repacking/merging T0 workflow (streamer to RAW conversion, I/O-intensive), in two test periods within Cosmic runs (CRUZET, MWGR’s)

Successful in both testing periods (one w/ ATLAS, one w/o ATLAS)

- Structure in first period, due to problems in Castor disk pool mgmt
- no evidence of destructive overlap with ATLAS
CMS Tier-1 sites in STEP’09

T1’s have significant disk caches to buffer access to data on tape and allow high CPU efficiencies

✦ Start with static disk cache usage…
  - At the start of data taking period 2009-2010, CMS can keep all RAW and 1-2 RECO passes on disk

✦ … fade into dynamic disk cache management
  - Later (and already now for MC), to achieve high CPU efficiencies data has to be pre-staged from tape in chunks and processed

In STEP’09, CMS performed:

✦ Tests of pre-staging rates and check of stability of tape systems at T1’s
  - ‘Site-operated’ pre-staging (FNAL, FZK, IN2P3), central ‘SRM/gfal script’ (CNAF), ‘PhEDEx pre-staging agent’ (ASGC, PIC, RAL)

✦ Rolling re-reconstruction at T1’s
  - Divide dataset to be processed into 1 days-worth-of-processing chunks, according to the custodial fractions of the T1’s, and trigger pre-staging (see above) prior to submitting re-reco jobs
Pre-staging

Tape performance very good at ASGC, CNAF, PIC, RAL
✦ IN2P3 in scheduled downtime during part of STEP’09
✦ FZK tape system unavailable, could only join later
✦ FNAL failed goals in some days, then problems got resolved promptly

<table>
<thead>
<tr>
<th>Target</th>
<th>2-Jun</th>
<th>3-Jun</th>
<th>4-Jun</th>
<th>5-Jun</th>
<th>6-Jun</th>
<th>7-Jun</th>
<th>8-Jun</th>
<th>9-Jun</th>
<th>10-Jun</th>
<th>11-Jun</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASGC</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>140</td>
<td>170</td>
<td>190</td>
<td>160</td>
<td>145</td>
</tr>
<tr>
<td>CNAF</td>
<td>56</td>
<td>380</td>
<td>300</td>
<td>160</td>
<td>240</td>
<td>240</td>
<td>270</td>
<td>105</td>
<td>80</td>
<td>125</td>
</tr>
<tr>
<td>FNAL</td>
<td>242</td>
<td>280</td>
<td>200</td>
<td>200</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FZK</td>
<td>85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN2P3</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96</td>
<td>99</td>
<td>120</td>
</tr>
<tr>
<td>PIC</td>
<td>50</td>
<td>60</td>
<td>61</td>
<td>106</td>
<td>83</td>
<td></td>
<td></td>
<td>99</td>
<td>142</td>
<td>123</td>
</tr>
<tr>
<td>RAL</td>
<td>40</td>
<td>250</td>
<td>230</td>
<td>160</td>
<td>140</td>
<td>135</td>
<td>190</td>
<td>170</td>
<td>100</td>
<td>220</td>
</tr>
</tbody>
</table>

Pre-staging and CPU efficiency at CMS T1’s

CPU efficiency

(= CPT/WCT)

Measured every day, at each T1 site. Mixed results:
✦ Very good CPU efficiency for FNAL, IN2P3, (PIC), RAL
✦ ~good CPU efficiency for ASGC, CNAF
✦ Test not significant for FZK
Transfer tests in STEP’09

Area widely investigated by CMS in CCRC’08
✧ All routes: T0→T1, T1→T1, T1←T2
✧ CMS runs ad-hoc transfer links commissioning programs in daily Ops

STEP’09 objectives:
✧ Stress tapes at T1 sites (write + read + measure latencies)
✧ Investigate AOD synchronization pattern in T1→T1
   - Populate 7 T1’s (dataset sizes scaled as custodial AOD fraction), subscribe to other T1’s, unsuspend, let data flow and measure

STEP’09 (2 weeks)

1 GB/s
Reached 989 MB/s on a 3-day average
✧ complete redistribution of ~50 TB to all T1s in 3 days would require 1215 MB/s sustained Regular and smooth data traffic pattern
✧ (see hourly plot)
Transfer latency in STEP’09

General feature:

✧ Smooth import rates in T{0,1}→T1 and T1→T2
✧ Most files reach destination within few hrs
  - but long tails by few blocks/files (working on this)

Load sharing in AOD replication pattern

✧ evidence of WAN transfers pattern optimization via files being routed from several already existing replicas instead of all from the original source

Example of T0 → T1
  [ T0 → PIC ]

Example of T1 → T1
  [ all T1’s → FZK ]

Example of T1 → T2
  [ CNAF → LNL ]

In replicating one ASGC dataset to other CMS T1’s, eventually ~52% of ASGC files were not taken from ASGC as source
Analysis tests in STEP’09

Goal: assess the readiness of the global Tier-2 infrastructure

✦ Push analysis towards scale using most pledged resources at T2
  - Close to 16k pledged slots, about 50% for analysis
✦ Explore data placement for analysis
  - Measure how (much) the space granted to physics groups is used
  - Replicate “hot” datasets around, monitor its effect on job success rates

Before STEP’09:

Increase in the # running jobs: more than 2x in STEP’09

More running jobs than analysis pledge (~8k slots)

Few T2 sites host more data than 50% of the space they pledge, though
Analysis tests in STEP’09

Try to increase the submission load, and observe

Caveats:
✧ Several sites had at least one day downtime during STEP09
✧ CMS submitters in STEP did not queue jobs at all sites all the time
✧ Standard analysis jobs were run, reading data, ~realistic duration, but with no stage-out

Another analysis exercise (“Oct-X”, in Fall 2009):
✧ Addressed such tests with a wide involvement of physics groups
✧ Ran ‘real’ analysis tasks (unpredictable pattern, full stage-out, …)
STEP’09 lessons learned

STEP'09 for CMS focussed on specific key areas
✦ It was an efficient approach to test and measure:
  - tape system performance at T0 and T1 sites
  - several aspects of the transfer system
  - analysis at T2’s at a higher scale
✦ Sites profited of exercises to further mature and tune their infrastructure

STEP’09 summary in a nutshell:
✦ T0 OK, tapes OK
  - Only need a better Castor@CERN monitoring for tape writing speed
✦ T1 downtimes are a concern, tapes OK for most of the sites
  - re-confirmed that CPU efficiency is significantly better with good mechanisms to pre-stage data
  - although very sensitive to tape family setup which has to be optimized
✦ Transfers in good shape in all routes
  - Just impacted by tape access to files at T1
  - pre-staging activated for all T1 transfer endpoints now
✦ Multi-VO aspect also tested (and no special worries arose)

More info on the STEP’09 twiki portal
✦ [https://twiki.cern.ch/twiki/bin/view/CMS/Step09](https://twiki.cern.ch/twiki/bin/view/CMS/Step09)
Some tests re-runs were performed as an appendix of STEP’09

**T0**: scale tests with special MC

- Produced special MC samples emulating a realistic population of PD’s - worth several days of T0 Ops with input at 300 Hz - and ran a \{bulk,express\} processing test, including the 48-hrs conditions hold
  - T0 farm has 2300 slots[*]
- Results of the “bulk processing test”
  - Used on average 1900 slots, demonstrated to sustain repacking and prompt-reco for ~250 HZ at 13% overlap
- Results of the “express processing test”
  - 25 Hz express stream processing needed on average 120 slots

**T1′s**: re-processing tests to check the CPU efficiency improvements

- Performed in Oct’09 at IN2P3+KIT (still due) and at ASGC, CNAF (requested by sites)
  - highlights: CNAF ran on the new (GEMSS) storage system; FZK successful, peaks at 300 MB/s in reading (100-150 on average) and at 400 MB/s in writing
  - ASGC and IN2P3 profited of these STEP re-runs to review the tape families set-up

The October Analysis exercise (“Oct-X”) ran at **T2’s**

- Not really a STEP’09 appendix (more focused on involving the physics groups)
- But drew interesting peaks in the analysis usage of T2 resources

[*] if no RelVal are running
2009: Planning vs Beams

Previous planning expectations for late 2009 - early 2010:

- A first data-taking period from Oct-Nov 2009 (then another one in Apr 2010)
- 100 days at 20% live-time (20 days); Total # evts: ~726 M (NOTE: includes ~40% overlap)
- RAW: 1.5 MB/evt, RECO: 0.5 MB/evt
- Total Volume of Data: ~1 PB RAW, 359 TB RECO
- Integrated lumi: a few tens of pb⁻¹
- Data rate from P5: 450 MB/s

What LHC accelerator and CMS detector gave us so far:

- 2009 to present for the Minimum Bias sample
- nearly 16k lumi sections on the RAW Minimum Bias PD’s
- 17 days; 90 M evts
- Total # files: 2400 files
- Total size MinimumBias: 7.8 TB
- Collected lumi: ~10 μb⁻¹
- Selecting only the ‘good’ runs: ~870 ‘good’ lumi sections
  - 22 hrs; 6.8 M evts; ~1 TB
T0 workflows

Rolling workflows (fully automated)
- Express processing (at Tier-0 level)
- Prompt reconstruction (at Tier-0 level)
- Prompt skimming (at Tier-1 level - but scheduled by Tier-0 system)

The CMS online system records events and stores them in binary files (streamer files)

T0 ‘Bulk’ processing path (latency of few days) and T0 ‘Express’ processing path (latency of 1-2 hrs)

Repacking of streamers into ROOT files, splitting of evts into Primary Dataset (PD) according to trigger selections (RAW data tier)

Special Alignment / Calibration (AlCa) datasets are produced and copied directly to the CAF

Reconstruction of RAW data for the first time (PromptReco) (RECO data tier), including AOD extraction

All steps of ‘bulk’ path combined into a single process run on ~10% of all events selected online from all the recorded data, output is copied to CAF for express AlCa workflows and prompt feedback by physics analysis

All RAW, RECO, AOD data is stored on tape at CERN and transferred to T1’s for storage on tape
CMS streams from the Online

**Stream A**: is the source of the Primary Datasets (PD’s). In the planning: it was expected at 300 Hz for 16 hrs with 8 hrs to catch-up, sustained is ~200 Hz, and corresponds to 10 PD’s. With 2009 collisions: it was 200 Hz (with spikes to more than 1 kHz), and in the first run there were only 2 PD’s populated

[ # evts: ~730M evts, size: ~100 TB ]

**Express**: expected to be ~40 Hz. Generally stayed within 40-60 Hz, with occasional spikes to 3 kHz

[ # evts: ~80M evts, size: ~12 TB ]

**Stream B**: was proposed before the run as insurance. It’s a very high rate stream of ZeroBias Data. Averages 1 kHz after the intervention.

[ # evts: ~278M evts, size: ~20 TB ]
Data volume: Streams and PD’s

Planning called for 726 M evts in data taking 2009

- 770 M simulated \( \rightarrow \) good agreement between real and simulated # evts

Event size and complexity of processing much lower than planned, though

- The fraction of “interesting” to “taken” events is much lower...

Some figures:

- Total streamer size: \( \sim 190 \text{ TB} \), total RAW size: \( \sim 150 \text{ TB} \)
  - Stream A: \( \sim 730 \text{ M evts} \), PD’s out of Stream A \([\ast]\) add up to \( \sim 723 \text{ M evts} \), MinimumBias RAW only \( \sim 90 \text{ M evts} \) (~8 TB)

**Table:**

<table>
<thead>
<tr>
<th>Stream</th>
<th>#Events</th>
<th>Size [GB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLTMON</td>
<td>19,454,692</td>
<td>3,935.81</td>
</tr>
<tr>
<td>Express</td>
<td>80,478,349</td>
<td>12,335.44</td>
</tr>
<tr>
<td>B_Buffered</td>
<td>130,167,201</td>
<td>25,478.95</td>
</tr>
<tr>
<td>A</td>
<td>731,269,373</td>
<td>98,467.30</td>
</tr>
<tr>
<td>B</td>
<td>279,019,843</td>
<td>20,111.04</td>
</tr>
<tr>
<td>RPCMON</td>
<td>145,150,042</td>
<td>540.61</td>
</tr>
<tr>
<td>FEDErrors</td>
<td>457</td>
<td>0.17</td>
</tr>
<tr>
<td>Calibration</td>
<td>209,228,981</td>
<td>24,186.10</td>
</tr>
<tr>
<td>ALCAP0</td>
<td>40,154,649</td>
<td>401.21</td>
</tr>
<tr>
<td>ALCAPHSYM</td>
<td>253,569,603</td>
<td>2,488.50</td>
</tr>
<tr>
<td>OnlineErrors</td>
<td>89,297</td>
<td>24.87</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,887,582,487</strong></td>
<td><strong>187,970.01</strong></td>
</tr>
</tbody>
</table>

**Table (continued):**

<table>
<thead>
<tr>
<th>PD</th>
<th>#Events</th>
<th>RAW Size [GB]</th>
<th>#Events</th>
<th>Prompt Reco</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinimumBias</td>
<td>90,052,126</td>
<td>7,822.32</td>
<td>89,791,258</td>
<td>6,964.36</td>
</tr>
<tr>
<td>RandomTriggersOP</td>
<td>46,969</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RandomTriggers</td>
<td>47,444,572</td>
<td>5,441.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZeroBias</td>
<td>78,086,537</td>
<td>6,445.00</td>
<td>78,038,521</td>
<td>4,671.24</td>
</tr>
<tr>
<td>ZeroBiasB</td>
<td>404,057,754</td>
<td>37,999.02</td>
<td>20,122,018</td>
<td>1,369.00</td>
</tr>
<tr>
<td>LogMonitor</td>
<td>86,462</td>
<td>19.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TestEnables</td>
<td>209,228,219</td>
<td>19,352.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhysicsMuonBkg</td>
<td>91,890,670</td>
<td>9,411.69</td>
<td>67,386,127</td>
<td>4,536.36</td>
</tr>
<tr>
<td>BeamHalo</td>
<td>123,852</td>
<td>17.40</td>
<td>80,971</td>
<td>31.21</td>
</tr>
<tr>
<td>Test</td>
<td>763,109</td>
<td>147.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALCaPhiSymCal</td>
<td>253,569,603</td>
<td>1,896.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MinimumBiasNoCal</td>
<td>0</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALCa0</td>
<td>40,154,649</td>
<td>326.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HcalHPrimDNoise</td>
<td>1,674,393</td>
<td>257.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPCMonitor</td>
<td>145,150,042</td>
<td>219.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEDMonitor</td>
<td>3,293</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calo</td>
<td>417,676,986</td>
<td>14,634.55</td>
<td>81,452,298</td>
<td>4,428.22</td>
</tr>
<tr>
<td>Cosmics</td>
<td>407,437,569</td>
<td>42,588.37</td>
<td>383,880,277</td>
<td>31,803.55</td>
</tr>
<tr>
<td>HcalIn2S</td>
<td>6,916,109</td>
<td>1,203.36</td>
<td>61,619</td>
<td>6.77</td>
</tr>
<tr>
<td>ZeroBiasBnotT0</td>
<td>4,129,290</td>
<td>473.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,888,761,905</strong></td>
<td><strong>148,229.52</strong></td>
<td><strong>700,628,076.00</strong></td>
<td><strong>59,332.73</strong></td>
</tr>
</tbody>
</table>

NOTE: Sums do not reflect overlaps in PDs
PD’s: event rates and RAW data rates

Average PD event rates per lumi section

Average event rate [Hz]

Average PD RAW data rates per lumi section

Average Datarate (MB/sec)

Individual PD rates lower than planning number but overlap very high
In general T0 job success/failure rates were irrelevant in terms of data usability for physics

- Reco and express failure rates dominated by:
  - Trigger rate explosion runs in pre-collision Cosmic runs data creating files too large to process
  - Issues with the Cosmics sequence with redundant beamsplash/collision cosmos triggers

Collisions data taking period (below) is a higher efficiency subset of BeamCommissioning09 (left):

**BeamCommissioning09 era**
(include also data taking)

<table>
<thead>
<tr>
<th>Job type</th>
<th>Succeeded</th>
<th>Failed</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repack</td>
<td>86982</td>
<td>69</td>
<td>99.92%</td>
</tr>
<tr>
<td>Repack Merge</td>
<td>23838</td>
<td>5</td>
<td>99.98%</td>
</tr>
<tr>
<td>Prompt Reco</td>
<td>209773</td>
<td>2875</td>
<td>98.63%</td>
</tr>
<tr>
<td>Reco Merge</td>
<td>16316</td>
<td>9</td>
<td>99.94%</td>
</tr>
<tr>
<td>Express</td>
<td>404546</td>
<td>9442</td>
<td>97.67%</td>
</tr>
<tr>
<td>MergePack</td>
<td>56117</td>
<td>30</td>
<td>99.95%</td>
</tr>
<tr>
<td>AlcaSkim</td>
<td>17631</td>
<td>431</td>
<td>97.56%</td>
</tr>
<tr>
<td>AlcaMerge</td>
<td>12357</td>
<td>78</td>
<td>99.37%</td>
</tr>
<tr>
<td>DOMHarvest</td>
<td>3172</td>
<td>0</td>
<td>100.00%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>830732</td>
<td>12939</td>
<td><strong>98.44%</strong></td>
</tr>
</tbody>
</table>

Success = job completed processing OK and histos staged to Castor

**Comparison 1**: BeamCommissioning09 Data taking period [run 12[234]XXX]

<table>
<thead>
<tr>
<th>Job type</th>
<th>BeamCommissioning09</th>
<th>Data taking period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repack</td>
<td>99.92%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Prompt Reco</td>
<td>98.63%</td>
<td>99.96%</td>
</tr>
<tr>
<td>Express</td>
<td>97.67%</td>
<td>99.99%</td>
</tr>
<tr>
<td>AlcaSkim</td>
<td>97.56%</td>
<td>99.98%</td>
</tr>
</tbody>
</table>
“Express at CAF” and “RAW at T1” latency

1 Latency from receiving first streamers of run at T0 to first express files on the CAF

- Very empty events...

2 Latency from run end (MinBias PD) to RAW at custodial T1

- Long tails correspond to:
  - Few day period when the MinBias PD first appeared at T0 and subscription to the custodial site was pending
  - Operational first experiences with multi-custodial sites in PhEDEx

Design spec: 1 hr
Observed (mean): ~25 min

Very tiny tails

Long tails (again: mostly transfer request approval latency)

Observed (mean): ~6 hrs

Mean: 5.895
RMS: 4.597
PromptReco latencies

1. Latency from run start (when T0 first saw streamers) to when first reco job started
   - Most runs started PromptReco within ~2 hrs of data taking

2. Latency from first Reco job starting to first Reco data becoming available at T0 (post merge)
   - First evts for most runs were promptly reco’ed and available on the CAF within 2 hrs from reco start

3. Latency from run end to Reco block complete at T1
   - Most blocks complete at T1 ~10 hrs after run ended
   - Longer tails though

Observed (mean): ~1.4 hrs
Tails here correspond to runs with high rates (repacking takes longer)

Observed (mean): ~1.7 hrs

Observed (mean): ~15 hrs
Long tails (again: mostly transfer request approval latency)
Tier-1 sites: ready?

T1 sites readiness and stability has improved

✦ In 2009 collisions data taking, CMS distributed custodial data to 6 T1’s out of 7, though Goal is to distribute multiple ‘hot’ copies at T1’s (+1 ‘cold’ archival copy at CERN)

✦ As long as the resources permit in 2010

![Site Readiness Status for CMS Tier-1 sites](image)
Transfers: T0 → T1

> 0.9 PB transferred out of CERN to T1’s during last 4 months

A good balance in data distribution to T1’s was kept in e.g. Dec 2009 (“hot” month)

✦ Too ‘few’ data to play with, though: better tuning will be hopefully possible in 2010

6 T1 sites received data

✦ the ‘hot’ MinBias dataset was sent at 4 T1 sites (and then to many T2’s, and also T3’s)

~ Dec 09

Interesting and “hot” month.

~ Dec 09

[ NOTE: IN2P3 was repopulated of a fraction of the data ]
**T1 re-reconstruction**

T1's involved in all scheduled workflows
- Re-reconstruction (at Tier-1 level)
- Skimming (at Tier-1 level)
- MC production (mostly at T2 level - but low-latency ones at T1 level as needed)

8[^*] re-reco passes of good runs list for the 2 PD's we had in 2009
- MinBias re-reco pass: ~ 22 M evts, total RECO size 2.3 TB plus skims
- ZeroBias re-reco pass: ~ 23 M evts, total RECO size 2.2 TB plus skims

Latency: 1-2 days
- Planning expectations: 1-2 weeks

CPU efficiency for reprocessing jobs: ~80-90%
- No accurate measures for all re-reco rounds, though
- Main time consumption:
  - Long running jobs (many evts in input file while splitting by file to keep lumi sections intact)
  - Debugging and bookkeeping
- Failures: still a few, due to monitoring and memory applications

[^*]: 9 passes as we speak
Transfers: T1→T1

Data transfer between T1’s driven by needs
✦ E.g. dominated by some repopulation of IN2P3
✦ Plot below also includes:
  - ~3 TB from ‘old’ FZK to ‘new’ KIT T1 PhEDEx node in Germany
  - ~8 TB to repair samples at ASGC
  - ~23 TB going to T1_CH_CERN

<table>
<thead>
<tr>
<th>Destination Site</th>
<th>Total Transfer Volume [TB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1_DE_KIT</td>
<td>0.39</td>
</tr>
<tr>
<td>T1_ES_PIC</td>
<td>1.51</td>
</tr>
<tr>
<td>T1_FR_CCIN2P3</td>
<td>105.15</td>
</tr>
<tr>
<td>T1_IT_CNAF</td>
<td>4.55</td>
</tr>
<tr>
<td>T1_UK_RAL</td>
<td>19.27</td>
</tr>
<tr>
<td>T1_US_FNAL</td>
<td>12.29</td>
</tr>
<tr>
<td></td>
<td><strong>143.16</strong></td>
</tr>
</tbody>
</table>

![CMS PhEDEx - Transfer Rate graph]

120 Days from Week 44 of 2009 to Week 09 of 2010

Maximum: 336.47 MB/s
Minimum: 0.00 MB/s
Average: 70.68 MB/s
Current: 20.65 MB/s

Daniele Bonacorsi
Tier-2 sites: ready?

T2 sites readiness has plateaued in late 2009 to ~40 usable T2’s

- Many structures visible though
  - e.g. SL5 migrations for bunches of sites at a time
MC production in 2009/2010

MC production continued in parallel to data taking

- Baseline is at T2’s. Special high-priority MC request go to T1’s also
  - mostly MinBias MC samples for comparison with data

Produced at the T2 sites (during Xmas break):

- 3 MinBias requests (2 for 900 GeV, 1 for 2.36 TeV), 10 M evts each

Produced at the T1 sites (late 2009 - early 2010):

- 63 production workflows ➔ 189 output datasets
- 385 M evts produced in total (RAW, RECO, AOD ~ 1/3 each)
- total output size: 58 TB

Produced at FNAL-T1 / CERN:

- “RelVal”: over 235M evts, 32 TB of tape space in 2567 datasets for 17 CMSSW releases

Latency:

- T1 level: ~ 2 days between request and samples available at T1
- T2 level: ~ 4-5 days between request and samples available at T1
  - Latency dominated by transfers to T1 sites and the fact that it was the last weekend before Xmas
- RelVal latency: ~24 hrs
  - Fixed # slots at CERN (500), could be eventually faster in FNAL
MC production in 2009/2010

Each color is a T2

Planning period started in Oct’09

- In late Jan’10: $1.2 \times 10^9$ evts = ~400M individual simulation events
- It roughly scales where we expected to be
  - 3-4 months through 6 month period, and we have more than half of ~750M

Over ~230 TB of MC produced only in last 3 months

Over ~200 M of MC evts produced only in last 3 months
A new AnalysisOps team in CMS Computing was launched in 2009
- Provide technical support for analysis infrastructure
- Manage centrally controlled space at T2’s by subscribing samples
  - AnalysisOps has access to 50 TB of space at each of ~50 existing T2’s

The team and the sites ran an Analysis Exercise in October 09 (“Oct-X”)
Consistent data traffic corresponding to datasets needed for analysis
- ~1.5 PB transferred to CMS T2’s in the last ~90 days (not necessarily with T1’s as sources)
- ~300 individuals submitting distributed analysis jobs in a given week
Analysis: slots usage and job success rate

~11k jobs slots are available for Analysis at T2 level

- Reaching ~75% utilization around the beginning of 2010
- In any given week 47±2 T2’s run analysis jobs

Success rate remains a persistent issue

- Improvement over last year though, when we had ~65%
  - Half of errors are related to remote stage-out of produced files
Summary

The 2009 data taking gave us few collisions events but plenty of interesting operational observations

- All digested so far, including the CMS-internal communication channels in Ops, now established and tested to work

The CMS T0 system was very stable during operations

- A predominant part of the effort spent on monitoring incoming data rates and on occasional mods of thresholds to adapt to changing data taking conditions

The CMS Tier-1/2 sites have reached a remarkable operational maturity

- Quite clear what could be more fragile and where
  - E.g. work-in-progress on risk-assessment analysis for different crisis scenarios at T1 sites

New limitation might appear in 2010 collisions data taking though

- Have to keep an eye on increasing data volumes, mostly
- More thorough planning and monitoring of data placement and WAN transfers

We are ready for the next round of data taking.