CMS Data Operations

Oliver Gutsche, Markus Klute

Fermilab, Massachusetts Institute of Technology

IN2P3 Visit
October 23rd, 2009

Disclaimer

- Most of the following has been discussed in the past.
- We want to repeat what CMS expects from a Tier-1.
- Workflows have been exercised already and the infrastructure has been used.
- (Hopefully) last chance to smooth out last edges before collision data arrives.

Tier-1 Responsibilities

- Host custodial copy of data and Monte Carlo on tape.
- Provide resources for processing workflows:
  - Skimming of incoming prompt-reco datasets.
  - Re-reconstruction of raw datasets.
  - Skimming of reco datasets.
  - In general: access files and write out new files.
- Serve data to other centers
- Network infrastructure responsibilities (FTS).

Expected Data Volume

- Parameters below are used of operational planning.
- Uncertainties are large.
- Assume 8E29 trigger table and rates.
- Split 2009/10 run in 3 periods of 1, 2 and 4E6s.

total run time ($\sim$ 70 days):	7E6s
total rate:	300Hz
mean overlap:	35% (later 20%)
RAW event size:	200kB
RECO event size:	400kB
AOD event size:	150kB
secondary data (SD) set fraction:	30%
number of full reco cycles:	3

Table: Parameters used for operational planning.

Expected Data Volume

- Assume 50Hz per PD (25Hz with safety factor of 2).
- We treat all PD equally.
- Custodial distribution of PDs to first order according to resource availability.
- More detailed and optimize planning will follow with first data experience.
- We will have copies of PDs at multiple sites, Tier-1s and Tier-2s.
- Currently, we plan to assign 1/11 PDs custodially to IN2P3.
- We will change custodial distribution following operational needs:
  - If PD rate and size not within margin.
  - Site performance is not sufficient.
  - To balance the distribution based on contributed resources.

Primary Dataset Unit

Period	1	2
RAW	50Hz · 1E6s · 200kB = 10TB	50Hz · 2E6s · 200kB = 20TB
RECO	50Hz · 1E6s · 400kB · 3Passes · 1.3SD-fraction = 78TB	50Hz · 2E6s · 400kB · 3Passes · 1.3SD-fraction = 156TB
AOD	50Hz · 1E6s · 150kB · 3Passes · 1.3SD-fraction = 29.25TB	50Hz · 2E6s · 150kB · 3Passes · 1.3SD-fraction = 58.5TB
Total	**117.25TB**	**234.5TB**

Table: Primary dataset unit.

- For each custodial primary datasets we allocate 120TB in period 1 and 240TB in period 2.
- Estimated 360TB custodial collision data till June 2010.
- We allocate 100 TB extra non-custodial storage at each sites to hold replicas of AOD.
- Additional resources are required for Monte Carlo.

Concerns

- We noticed an unusual high number of tape loses at IN2P3 (3 since 2007).
- Tier-1 - Tier-2 separation was not transparent to us.
  - We acknowledge that this is a one-time operation.
  - We also acknowledge that this is a Tier-2 problem affecting the Tier-1.
  - In general we ask Tier-1's to be proactive about site issues.
  - Currently, all links for T2_IN2P3 (except the link to T1_IN2P3) are deactivated to avoid being flooded with transfer errors due to namespace inconsistencies.
  - As we understood from communication with the T2 contact, it will take significantly more time to complete the copy of datasets.
  - After the copy is complete, we need a full consistency check between local MSS, PhEDEx and DBS for both T1 and T2 at IN2P3.
  - Then we can reactivate the links

Expectations: Support

- Facility and all central services: 24/7.
  - In case of alarm ticket response within 1 hour.
  - All central services (we expect: CEs, SEs, MSS, batch system, worker nodes, Frontier and Squids, access to installed CMS software).
- CMS contact: nominally business hours
  - Contact has to be familiar with CMS workflows.
  - Organize tape family setup.
  - Approve transfer requests.
  - Regularly run PhEDEx consistency tools to check for orphaned files and inconsistencies in bookkeeping systems.
  - Follow up on savannah tickets and triage to facility if needed.
  - Make sure that local hardware situation (available disk, tape) is close to pledges and update SiteDB regularly.
  - Proactive information about site issues.

Expectations: Infrastructure

- CMS software server.
- Squids and Frontier.
- Local workernode disk space (needed for caching input files (LazyDownload) and writing output), needs to be sufficiently dimensioned, CMS is working hard to restrict file sizes to 10 GB and optimize workflows not to overfill WN at the sites.
- Correctly working TFC for CMS application and PhEDEx.
- Role and cleanup of /store/unmerged (temporary small files, don't have to go to tape, should be cleaned up automatically by production systems, not perfect, need automatic cleanup by sites for files older than 30 days)
- FTS server for regional site support.
- Data incoming from T0 should stay on disk for a period of time, in the first year all data coming from CERN should stay on disk.
- New: checksum verification of incoming files via PhEDEx.

Expectations: Operational

- Pre-staging via mail or srm (tested in STEP'09, complaints?)
- Tape family creation (manual communication with the sites).
  - In case of transfers into the sites: requests are not auto approved, additional safety net if tape families are not setup.
  - In case of output produced at sites, tape family creation is checked before workflows are started.
- Transfer request approval within 24 hours during business hours (does not work reliably at all times, if necessary this is overruled by central approval).

Summary and Conclusion

- IN2P3 was not used much this summer.
- Expect ramp up usage of IN2P3 in the following weeks and months:
  - "Backfill" workflows to study workflow performance.
  - Transfer tests to Tier-2's.
  - Storage consistency checks.
  - Expect 24/7 support of the facility.
  - Support by CMS contacts during business hours.
- You will find more detailed information in backup.

Backup: For Further Discussion.

Jobs Submission

- Use ProdAgent with gLite WMS or glideIn Pilot GRID submission infrastructure to submit jobs to sites.
- Site needs enough CEs to sustain submission load and keep the latency low (we are seeing latency problems with T2 sites who don't have enough CE's and connections to the WMS time out).

Jobs Software

- Jobs use pre-installed CMS software installations, only job configurations are sent to the local workernode (WN).
- Software is installed centrally via special GRID roles and is accessed on all WN (usually, software is installed on nfs mounted shares, installation jobs have write access while all WN have only read access, several nfs servers might be needed to serve large installations).
- Jobs read conditions data from the central CMS databases via local Squid caches and the Frontier system.
- Large sites might need several Squid servers to distribute the load.

File access

- Files should be accessed directly from MSS disk to keep the CPU efficiency high.
- CMS currently does not use a central pre-staging system, we either ask the sites per mail to pre-stage a sample or use a SRM based pre-staging script which essentially sends srm-bring-online commands and then forgets about them.
- It is enough to get to get of the order of 90% of the files pre-staged to keep the CPU efficiency high, the remainders are handled by the MSS internal staging mechanisms.
- File access should be fast, sufficient network bandwidth is needed. The access to files is very erratic, meaning that the application jumps within the file a lot. CMS can use special caching mechanisms on the local workernode like LazyDownload which reads files in 128 MB chunks. All chunks of a file are kept on local disk, so the maximum local worker node disk consumption is the size of an individual input file (10 GB max.).

Job Output

- All output the application is producing including the data output files are written to the local workernode disk.
- An application can write out files for different output datasets with different LFN structures.
- This includes also all log files.

Job Stage-Out

- After the application execution, CMS stages out the data output files into the CMS namespace.
- If the output file size is large enough to be suited for tape storage (more than about 2 GB per file), files are directly staged into the CMS namespace /store using the appropriate logical file name (LFN) (policies are listed here: https://twiki.cern.ch/twiki/bin/view/CMS/DMWMPG_Namespace).
- Sites will be asked beforehand to create tape families according to the LFNs to be written in production. This will have to be done and checked before any processing starts.
- The LFN to physical file name (PFN) resolution is using the site's trivial site catalog (TFC) which is accessible via the software installation.

Job Stage-Out

- If an output file is not sufficiently large in size, it is staged out using the same mechanism into a portion of the CMS namespace called /store/unmerged .
- This namespace portion is temporary and does not have to be migrated to tape.
- Although the production system is trying to clean up after itself, the remote clean up is not perfect. Sites are asked to delete files in /store/unmerged older than 30 days.
- In addition, archives of the produced log files (called logArchive tarball) are staged into /store/unmerged for later debugging and archiving purposes. To enable remote debugging, /store/unmerged should be reachable via srm to be able to copy logArchive tarballs. logArchive tarballs are later archived at CERN and deleted.

File Merging

- Files which have been written to /store/unmerged are merged in a dedicated step and this output is written into the CMS namespace using the appropriate LFNs.
- Stage out of merged files follows the above described principles.

Data Transfer

- CMS uses PhEDEx to initiate transfers and FTS to perform transfers and report back to PhEDEx about the status.
- CMS follows a pull model, that means the FTS of the T1 which is pulling the data is used. With exceptions:
  - Transfers from T0 to the T1 sites are handled by the CERN FTS server
  - Transfers from a T2 to a T1, the FTS of the respective T1 is used
  - Transfers from a T1 to a T2, also the FTS of the respective T1 is used
  - Transfers form a T2 to a T2, the FTS server of the regionally associated T1 of the receiving end is used
- PhEDEx is an agent based system which runs at each site. It uses a TFC to resolve LFN to PFN which ideally is the same as for processing

Data Transfer

- Tier-0 → Tier-1:
  - All data from T0 is archived at the sites and written to tape.
  - There might be processing steps needed immediately after the data arrives at a site (skimming) -¿ all arriving data from T0 should also be kept on disk for a time period. In the first year, all data coming from T0 should be kept on disk.
- Tier-1 → Tier-2:
  - Data and MC samples are served to the T2 sites via PhEDEx in burst mode.

Data Transfer

- Tier-2 → Tier-1:
  - MC production is only run on T2 sites and the outputs have to be archived on tape at the T1 sites. There is no need to keep MC production on disk after they have been migrated to tape. The MSS will handle necessary staging if a sample is requested to be transferred to a T2.
  - There might be dedicated processing on MC samples. Like for data not on disk, pre-staging requests would be made to the sites either via email or srm pre-staging scripts.
- Checksum verification
  - PhEDEx is able to provide adler32 and cksum checksums of all files transferred into a site.
  - Sites will be asked to check these with the checksums of the local files to verify the correctness of the transfer.
  - Currently only FNAL is checking checksums of incoming files and repeats transfers if the checksums do not agree.

Primary Datasets

PD Name	Rate [Hz]
JetMonitor	14
Jets	23
Met_HT_BTAG_HSCP	7
MuMonitor	13
Mu	25
EleGammaMonitor	25
EleGamma	23
DoublePhoton5_Res	13
Tau	20
MinB	14
BH_Forward	7

Table: 8E29 Primary Dataset Table.

- Table sums up to 185Hz.
- Large uncertainties on trigger rates estimated from MC.
- There are plans to scale trigger rates to 300Hz.
- This naively translates to a rate of 405Hz in PDs (300 Hz with 35% overlap).
- Total of 11 primary datasets.