Experience with CMS Offline and Computing from Commissioning to Collisions

ICHEP, July 24th 2010

Markus Klute [MIT / CMS DataOperation Coordinator]
for the CMS collaboration

CMS world wide distributed computing system

Tier-0 @ CERN

Tier-1 @ DE, ES, FR, IT, TW, UK & US

Tier-2 @ 50 sites on 4 continents

Tier-3 @ many, many locations
Computing Activities & Resources

- **Governing Principles**
  - 2 safe copies of RAW data on tape at CERN and Tier-1 sites
  - 2-3 large re-reconstruction passes per year in first years at Tier-1 sites
  - Monte Carlo production matches collision data
  - Production and user jobs go where the data is
  - Full “mesh” network connectivity. All sites are connected with each other
  - A lot a work went into preparation and testing
  - Provides flexibility to tackle unforeseen scenarios (very high turn-around before ICHEP)

- **Tier-0 activities**
  - Prompt data processing
  - Prompt calibration and alignment
  - Storage of RAW data backup

- **Tier-1 activities**
  - Custodial storage of RAW data
  - Prompt skimming
  - Reprocessing of data and MC
  - MC production

- **Tier-2 activities**
  - MC production
  - User analysis

- **Tier-3 activities**
  - User analysis
Computing Activities & Resources

- Excellent site readiness
- Key ingredient for successful operations
- Close relationship with sites through contact person and data manager

Installed resources as of 2010:

- **Tier-0**: 55 kHS06, 3 PB disk, 9 PB tape
- **Tier-1**: 100 kHS06, 11 PB disk, 20 PB tape
- **Tier-2**: 192 kHS06, 12PB disk
Offline Project

- Huge enterprise to provide stable software while incorporating latest developments
- Hundreds of code developer
- Offline sub-projects
  - Simulation
  - Fast Simulation
  - Data & Workflow Management
  - Reconstruction
  - Core Software Framework
  - L1 Software
  - Data Quality Monitoring
  - Databases
  - Validation
Offline Releases and Performance

- **Release cycles**
  - Patch release mechanism deployed
  - Deployed train model for release cycle
  - Detailed and frequent monitoring of software quality and performance

- **Software performance**
  - Extensive optimization program
  - Looking into multi-core usage
  - Reconstruction of collision data (MinBias)
    - 0.6 seconds per event
    - 400 kB RECO, 150 kB AOD
    - 900 MB memory
  - Simulation of Monte Carlo (ttbar)
    - 90 s/evt (50 s for low-pT QCD)
    - 1400 kB RAW SIM
    - 980 MB memory
Collision Data in 2009 and 2010

- **First 900 GeV Collisions**
  - Nov 23rd 2009
  - 350k pp collision (10 ub⁻¹)

- **First 2.36 TeV Collisions**
  - Nov 30th 2009
  - 20k pp collision (<1ub⁻¹)

- **First 7 TeV Collisions**
  - Mar 30th 2010
  - L = 303 nb⁻¹

- **Luminosity Goals:**
  - 100 pb⁻¹ in 2010
  - 1 fb⁻¹ in 2011

- Computing resource plan input: event counts and size, not luminosity
  - Trigger rate and overlap (300 Hz, 20-40%)
  - LHC duty cycle (20-50%)
  - Event sizes (300kB RAW, 500kB RECO, 200 KB AOD)
Central Processing @ CERN

CMS (P5)

- HLT
  - Storage Manager

CAF

- Alignment & calibration
- Commissioning, Physics, DQM

Tier-0

- Repacker
- Express reconstruction (within 1-2 hours)
- Disk buffer

Tier-1’s

- Prompt skims

Tier-1’s

- Prompt reconstruction

Feedback planned but not implemented

Conditions

- Offline Conditions Database

Tier-0

- Primary Datasets

Tier-1’s

- Express reconstruction

Tier-0

- Repacker

Tier-1’s

- Prompt skims
Central Processing @ CERN

- Rolling workflows are fully automated
- Express processing provides quick feedback for commissioning, data quality monitoring and physics
- Alignment and calibration loop to improve quality of prompt reconstruction
- Operational experience at Tier-0 is excellent. Success rate of 99.9%
- Categorize data according to trigger selection in primary dataset

Express FEVT Latency

- Mean latency 1h
- Tails dominated by untypical conditions

- Data volume from 7 TeV collisions
  - 987M raw events
  - 88 TB
  - 11 primary datasets
Data Transfer from CERN to Tier-1's

- Resources provisioned for steady data stream from Tier-0 to Tier-1's
- Current reality looks different
- Total volume of 1 PB since April
- Very good transfer quality
Central Processing @ Tier-1

- All Tier-1 sites used in production
- Upon arrival at Tier-1’s, data is being processed and stored on tape
- Prompt skimming
  - see poster by Si Xie
- Produce small datasets based on trigger selection or reconstructed objects
- Fully automatized system
- Reprocessing of data and MC
  - Improved software, calibration and alignment
  - 10 data reprocessing passes for 7 TeV
  - 3 MC reprocessing passes for 7 TeV
Data Distribution for Analysis

- Data distribution to Tier-1 organized centrally to balance resource utilization.
  - Jobs go where the data is
- Data storage serves as temporary buffer
  - Refresh with hot datasets
- Data distribution on Tier-2 organized
  - Centrally (Analysis Operations)
  - By physics groups
  - By local users

Tier-2 storage breakdown (typical example)
Data Distribution for Analysis

- Data transferred from Tier-1's
  - 49 Tier-2 sites received data
  - > 5 PB transferred in last 120 days
  - average rate 562 MB/s
  - max rate 1407 MB/s

- Data transferred between Tier-2's
  - 41 Tier-2 sites received data
  - > 2.5 PB transferred in last 120 days
  - average rate 254 MB/s
  - max rate 853 MB/s
  - full mesh approach
  - Data distribution re-balances itself
  - Datasets produced at Tier-2’s can be distributed to others
Analysis Activities @ Tier-2/3's

- 500 individual CMS users active using grid resources
- Maximum reached in preparation for ICHEP
- Tier-2 resource usage currently dominated by analysis activities

**Terminated jobs**
30 Days from 2010-06-15 to 2010-07-25

**Running jobs**
30 Days from 2010-06-15 to 2010-07-25

- Tier-2 Activity
- 150k terminate jobs
- 20k parallel jobs

Markus Klute, MIT
ICHEP - July 2010
Monte Carlo Production

- Successfully exercised for years
- 64 Tier-1/2/3 sites participating
- MC production preparation started Summer 2009
- Multiple production validation cycle

- Productions:
  - Switched from 10 to 7 TeV
  - 200M full MC events in 2 weeks
  - MC for 900 GeV and 2.36 TeV collision
  - Mostly “Data-like” MC production in 2010 (MinimumBias & low-Pt QCD)

Approx batch slot usage
58 Days from Week 26 of 2009 to Week 34 of 2009

MC Production in 2009

8k slots

MC in 2009/2010: Simulated Events per Month

Events, total ~2.3 Billion RAW events

- Events per month from January 2009 to May 2010
Summary and Conclusion

- **CMS Distributed Computing Model** has proven very successful
  - Able to cope with load in all sectors
  - (Rare) backlogs or service problems w/o impact on physics
- **CMS Computing** is truly a distributed system
  - Excellent performance of sites with start of LHC operations
- **CMS Offline project** reached steady state
  - Provides stable and highly efficient software
- Not operating in resource constrained environment
  - Total data volume still small
  - Allows very fast turn-around to incorporate new software, calibration and alignment
  - Will change this Fall and is a new challenge!