CDF Offline Report

Donatella Lucchesi  
*INFN-Padova, U. of Padova*

Rick Snider  
*Fermilab*

Raw data production model  
Production operations  
Monte Carlo data production  
Analysis Computing

All Experimenter's Meeting  
October 27, 2008
Raw data production model

• Goals of offline production operations
  ◆ Deliver data required for analysis as close to data taking as possible
    ▶ Final compressed datasets from reconstructed raw data
  ◆ Ensure production is not the limitation in the rate of physics output

• The processing problem
  ◆ Log data at rate of 5 – 7 M events/day
  ◆ Calorimeters require re-calibration every ~3 months
    ▶ Need to accumulate ~150+ M events to calibrate (though not all used for calib)

• Strategy
  ◆ Divide data into “run periods” of 4 – 10 weeks
    ▶ Typically 200 – 400 M events
  ◆ Process data by run period
    ▶ Calibration, raw data reconstruction, ntuple creation
  ◆ Analyses use multiple run periods as needed for new results
Raw data production model

Raw Data
- Inclusive
- H-Pt Leptons
- Jets
- Hadronic B

Data Quality Monitoring
- Calibration pass
- Calib dataset
- Calibration DB

Electrons
Muons
N-tupling
Reco / split
40 GeV Jets
Had Bs

Prod Data
- Jets
- H-Pt Leptons
- Inclusive
- Hadronic B

N-tupling
- Stn
- TNt
- Stn

Root N-tuples

Analysis

Concurrent with data taking
Run periodically over "run period"
Prod Data
Root N-tuples
Raw data production model

- Raw Data
  - Inclusive
  - H-Pt Leptons
  - Jets
  - Hadronic B

- Data Quality Monitoring
  - Calibration pass
  - Calib dataset

- Calibration DB
  - Calibration procedure

- Reco / split
  - Electrons
  - Muons
  - 40 GeV Jets
  - Had Bs

- 40 GeV Jets

- Analysis
  - Concurrent with data taking

- Prod Data

- Root N-tuples

Oct 27, 2008
Raw data production model

- Raw Data
  - Inclusive
  - H-Pt Leptons
  - Jets
  - Hadronic B

- Data Quality Monitoring
- Calibration pass
- Calib dataset

- Calibration procedure
- Calibration DB

- Reco / split
  - Electrons
  - Muons
  - 40 GeV Jets
  - Had Bs

- N-tupling
- Prod Data

- Analysis
  - Stn
  - TNt
  - Stn

Run periodically over "run period"
The production cycle

• Detector calibrations
  ◆ Process about 30% of the raw data within a few days of data taking
  ◆ Calculate calibrations and perform validation for each run period
  Typically completed 3 – 6 weeks after end of run period

• Raw data production
  ◆ Reconstruction of data
  ◆ Split data into datasets into physics datasets based upon triggers
    ► 42 full + 9 compressed datasets
  Typically completed 3 – 6 weeks after calibrations ready

• Ntupling
  ◆ Performed on production output (after splitting)
    ► Prioritize processing to do most important first
  ◆ Three partially overlapping flavors: standard, top, Bs
  Typically 2 – 3 days behind raw data production
Raw data operations

- Event reconstruction
  - Average processing time
    - ~2 sec/event across all streams and luminosities (varies greatly event type)

### Data delivery for recent run periods

<table>
<thead>
<tr>
<th>Period</th>
<th>Start</th>
<th>End</th>
<th>Lum (pb⁻¹)</th>
<th>Events (M)</th>
<th>N-tuples ready</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>May 13, 07</td>
<td>Aug 4, 07</td>
<td>317</td>
<td>545</td>
<td>Nov 29, 07</td>
</tr>
<tr>
<td>14</td>
<td>Oct 28, 07</td>
<td>Dec 3, 07</td>
<td>45</td>
<td>59</td>
<td>Feb 21, 08</td>
</tr>
<tr>
<td>15</td>
<td>Dec 5, 07</td>
<td>Jan 27, 08</td>
<td>159</td>
<td>210</td>
<td>Apr 7, 08</td>
</tr>
<tr>
<td>16</td>
<td>Jan 27, 08</td>
<td>Feb 27, 08</td>
<td>142</td>
<td>168</td>
<td>May 21, 08</td>
</tr>
<tr>
<td>17</td>
<td>Feb 28, 08</td>
<td>Apr 16, 08</td>
<td>188</td>
<td>235</td>
<td>Jun 6, 08</td>
</tr>
<tr>
<td>18</td>
<td>Apr 18, 08</td>
<td>Jul 1, 08</td>
<td>407</td>
<td>436</td>
<td>Oct 25, 08</td>
</tr>
</tbody>
</table>

10 – 12 weeks for most

Continue to work on improvements to address rate limitations
Raw data operations

• Data processed on-site
  ◆ Past run periods processed on 600 node farm dedicated to CDF
    ► Also used for calibrations, N-tupling and analysis
  ◆ Currently migrating processing to Fermigrid-based farms
    ► Final stage of migrating all CDF computing into Fermigrid
      ▷ Better optimizes CPU utilization
      ► All processing for the next run period will be performed on Fermigrid

• Data re-processing
  ◆ About 30% of data is processed twice as part of production cycle
    ► Once for calibrations, once for physics datasets
  ◆ The experiment has no plans for large scale re-processing
Data volumes

- Data on tape
  - Total of 3.6 PB
  - Raw data
    - 7.9 billion events
  - Monte Carlo data
    - 4.6 billion events
    - Includes a combination of centrally produced MC and analysis-specific MC
Monte Carlo data production

• The “old” MC data production model
  ◆ Run-based MC that takes into account detector configuration and luminosity
  ◆ Required continuous MC production operations coordinated with data taking

• Changing the production model for new MC

• The new MC production model
  ◆ Luminosity profile scaling
    ► Generate MC asychronously with data taking
    ► Allows better scheduling of CPU usage
    ► Significantly reduces amount of MC needed relative to run-based approach
  ◆ Possible because the detector configuration is very stable
Monte Carlo data production

- Centralized MC produced off-site
  - Open Science Grid
    - US institutions
  - Same technology for Pacific Rim
  - LHC Computing Grid
  - INFN-CNAF
    - Priority access to CNAF T1
  - Barcelona

OSG usage by site

Pacific Rim usage by site

LCG usage by site
MC data production operations

- MC data generated
  - 1.1 G events produced last year
  - Some periods of concentrated production during “MC attacks”
Analysis computing

- Computing requirements scale with:
  - Full data set size
  - Complexity of analyses
  - Number of people / analyses

- Facilities
  - 5k CPUs on-site for data intensive analysis
    - Shared with production activities
    - Some large datasets also located at INFN-CNAF
  - Off-site computing also available for CPU intensive analysis
    - Matrix element analysis, pseudo-experiments, etc.

![CDF On-site CPU usage](chart)

- Production
- N-tupling
- Central MC
- Analysis
- Specific MC
- Core analysis
- Other analysis
Analysis computing

• Is it all effective?
• The bottom line is the physics that CDF produces
  ◦ 50+ new results at 2008 Winter conferences
  ◦ Another 50+ new results at 2008 Summer conferences
  ◦ Expect ~40 publications in 2008
Summary

- The CDF offline is successfully meeting the physics needs of the experiment
  - Due to the hard work of many collaborators at Fermilab and around the world
  - A close and productive collaboration with the Computing Division has been critical to this success
    “Thank you” to the CD!
- Will ensure continued success by working to improve the systems, increase efficiency and reduce the effort required to conduct computing operations.