THE OCTOBER PHYSICS EXERCISE: PREPARING FOR STARTUP

September 17, 2009

Joe Incandela
UC Santa Barbara
October exercise: Main Goals

- Deploy the Tier-2 analysis model
  - Begin to train people to do all of the tasks that are needed to enable individuals, groups, and CMS as a whole to efficiently access and analyze data.

- Begin working with 31X samples on a large scale and also start looking at SDs

- A scale-test (October 5th - 19th)
  - Approximate some of what we’ll face with early data - many key analyses simultaneously
    - T1-T2 transfers of (pseudo) Secondary Datasets
    - Widespread use of CRAB server (very important to address issues)
    - Group Skims of SDs by high priority users
    - T2-T2 transfers by some groups
    - T1-T2 subscriptions of MC samples by all groups
    - Analysis jobs exercising early analyses by as many people as possible
  - See if we can correct problems, overcome obstacles, eliminate bottlenecks during an intense two week period...
    - There will be an “OCT X” E-Log where people should regularly post comments/plots etc. regarding problems and successes. It will be monitored by Computing to be able to provide assistance where they can and it will provide a record of the problems/issues that arise.

- Post-mortem
  - Review what has been learned and iterate ahead of data-taking
  - Continue to work on improving the system continuously
Elements of plan

Tier-1’s focus on reconstruction – not used for data analysis

Tier-2’s

- Store SDs in RECO initially but need to switch to AODs at some point in the run.

Associated PAG, POG+DPG

- Groups skims (or subscribe SDs or even PD’s) to group space

NB: 2009-10 run much longer than 2008 was to be. Cannot put full PD’s in RECO format at Tier-2’s. SDs are a better choice. Need to switch to AODs mid-way through upcoming run..

*see: Ian Fisk “Introduction to Analysis on Tier-2s” **confId=41375 Sep. 23, 2008

**see: 2nd ECOM report https://twiki.cern.ch/twiki/bin/view/CMS/ECoMReports
Primary Datasets and their evolution

- The 8E29 PDs are ready, 1E31 PDs very nearly ...
  - They will be revised after we see the data.
    - Just like the triggers, what we have designed with MC will not be exactly right for real data.
  - OpenHLT is designed to allow us to use real data to quickly analyze the impact of changes in triggers on rates, correlations, and primary datasets.
  - Even so, they must change in a controlled way because they have to satisfy both physics and computing needs
    - For physics we define them by physics object
    - For Computing, we balance them in size and make sure that none are so big that they cannot fit in at least 3 of our biggest Tier-1’s
Secondary datasets (see R. Rossin’s talk)

- Each SD is derived from 1 PD using trigger info
  - NB Top & SUSY group pioneered use of OpenHLT for design of SD.
  - This is a great idea! We would like to follow through on this by incorporating all SD definitions in OpenHLT as done for PDs

- SD aren’t as constrained as PD in number and design
  - But there are real benefits of having fewer SD
    - If they are used by multiple groups they become better understood and thus better foundations for subsequent analysis steps.
  - Further selection is done with Group Skims.

- Iterate
  - We’ll start with a lean set of SDs (RECO+AOD) that will allow us to get at data quickly without jamming the system.
    - These will be used in the October exercise

- NB: Eventually (in the upcoming run) we can’t afford to distribute RECO so we must worry about what we want in the AOD.
<table>
<thead>
<tr>
<th>Month</th>
<th>OP scenario</th>
<th>Max number bunch</th>
<th>Protons per bunch</th>
<th>Min beta*</th>
<th>Peak Lumi</th>
<th>Integrated</th>
<th>% nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beam commissioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Pilot physics combined with commissioning</td>
<td>43</td>
<td>$3 \times 10^{10}$</td>
<td>4</td>
<td>$8.6 \times 10^{29}$</td>
<td>~200 nb$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>43</td>
<td>$5 \times 10^{10}$</td>
<td>4</td>
<td>$2.4 \times 10^{30}$</td>
<td>~1 pb$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>156</td>
<td>$5 \times 10^{10}$</td>
<td>2</td>
<td>$1.7 \times 10^{31}$</td>
<td>~9 pb$^{-1}$</td>
<td>2.5</td>
</tr>
<tr>
<td>5a</td>
<td>No crossing angle</td>
<td>156</td>
<td>$7 \times 10^{10}$</td>
<td>2</td>
<td>$3.4 \times 10^{31}$</td>
<td>~18 pb$^{-1}$</td>
<td>3.4</td>
</tr>
<tr>
<td>5b</td>
<td>No crossing angle – pushing bunch intensity</td>
<td>156</td>
<td>$1 \times 10^{11}$</td>
<td>2</td>
<td>$6.9 \times 10^{31}$</td>
<td>~36 pb$^{-1}$</td>
<td>4.8</td>
</tr>
<tr>
<td>6</td>
<td>Shift to higher energy: approx 4 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4 – 5 TeV (5 TeV luminosity numbers quoted)</td>
<td>156</td>
<td>$7 \times 10^{10}$</td>
<td>2</td>
<td>$4.9 \times 10^{31}$</td>
<td>~26 pb$^{-1}$</td>
<td>3.4</td>
</tr>
<tr>
<td>8</td>
<td>50 ns – nominal Xing angle</td>
<td>144</td>
<td>$7 \times 10^{10}$</td>
<td>2</td>
<td>$4.4 \times 10^{31}$</td>
<td>~23 pb$^{-1}$</td>
<td>3.1</td>
</tr>
<tr>
<td>9</td>
<td>50 ns</td>
<td>288</td>
<td>$7 \times 10^{10}$</td>
<td>2</td>
<td>$8.8 \times 10^{31}$</td>
<td>~46 pb$^{-1}$</td>
<td>6.2</td>
</tr>
<tr>
<td>10</td>
<td>50 ns</td>
<td>432</td>
<td>$7 \times 10^{10}$</td>
<td>2</td>
<td>$1.3 \times 10^{32}$</td>
<td>~69 pb$^{-1}$</td>
<td>9.4</td>
</tr>
<tr>
<td>11</td>
<td>50 ns</td>
<td>432</td>
<td>$9 \times 10^{10}$</td>
<td>2</td>
<td>$2.1 \times 10^{32}$</td>
<td>~110 pb$^{-1}$</td>
<td>12</td>
</tr>
</tbody>
</table>

Mike Lamont, Sep. 7 in Bologna
Adding new SDs and Central Skims

- The 1st iteration of the SDs is for commissioning:
  - Learn how to manage the reliable splitting of PD at Tier-1’s into SD and establish smooth quasi-continuous T1-T2 transfers ...
  - Use them to learn more about the detector, understand problem elements, severity flags, good run selection, clean up cuts...
  - Adjust triggers, modify PDs....
  - Commission Group Skims, analyze data, study what we really need, and refine what we use...

- In other words there is plenty to do without trying to manage specialized Tier-1 skims at startup.
  - Add SDs naturally, as they are understood & demonstrated to be 1) broadly useful and 2) extensively tested, before adding to T1 load!
  - Add Central Skims naturally when there is a clear motivation.

- We do not gain anything by overwhelming DataOps!

- We will have the support of “AnOps”
  - Frank Wuerthwein’s talk.
We have 3 weeks before October exercise

1st meeting today (17 Sep.)
- Hash out general plans. – SDs presented by Roberto R. – Computing plans by F. Wuerthwein – Organizing POG-PAG plans by Roberto T.

Next meetings
- Wednesday plenary meetings 23rd Sep.
  - Presentations from PAG and POG on their organization/plans for Oct.X
    - We will want to see the most ambitious plans the groups can mount, but we will then need to discuss with Computing what can actually be done. We will do that and circulate a draft plan that then gets presented the following week in the plenary.
  - Possible tutorials on use of key tools, skimming, data transfers etc.
- Wednesday plenary meetings 30th Sep.
  - A presentation of the overall plan for Oct.X (as discussed above)
  - Possible tutorials on use of key tools, skimming, data transfers etc.

In parallel
- R. Rossin is finalizing configuration files for 1st iteration SDs to be run by dataOps to define “pseudo” SDs and DataOps will begin placing datasets in central spaces.

Another goal will be to develop guidelines for a variety of tasks
- dataset and skims design
- effective CRAB-based analyses
  - More generally, how individuals should do analysis to increase their success rate and use resources as efficiently as possible.
- We have many pieces of the puzzle in place
  - T2 associations and data transfer tools
  - Lean trigger tables
  - PDs and 1<sup>st</sup> iteration of SDs
  - AnOps: New branch of computing that will actively support analysis

- Get ready to take the system for a test drive
  - We will certainly find problems.
  - Do not panic! October exercise is a dry run which will help a lot to commission and improve the mode.
  - Take it seriously and help get ready for real data.
Additional Info
Overlap among triggers is *expected* to decrease when increasing lumi

- E.g. jets, at higher thresholds they decouple more readily from other objects

Did not happen for us because the $10^{31}$ menu contains:

- Different paths for inherently similar objects with very loose definitions
  - For example: electrons and photons are not really distinguished
  - Work ongoing (more on this later)

- Multiple object triggers.
  - By construction they build correlation among PDs
    - This has been recently solved by changing the definitions of several Ele+X and Mu+X triggers, cutting the overlap in this sector from $O(10)$ Hz to negligible values.
• The total overlap is 21% (acceptable) when clustering all the single Electron/Photon triggers in a single, oversized (unacceptable), Primary Dataset
• This remaining issue is currently under investigation (see also Chris Seez’s talk)
Solving the e/γ overlap problem:
- E.g. might be to prescale Photon25 by a factor 10-20 and use Photon25_TrackIso for physics, which has little overlap with electron triggers, while keeping Photon25_TrackIso_EcalIso as backup.
  - Roberto Covarelli is studying the efficiencies on gamma+jet events (results expected next week).
  - We will check the change in total overlaps with OpenHLT (results expected next week)

Impact of e/γ overlaps on the total PD overlaps:
- Single Elegamma PD has a total rate of 38Hz
- The full menu has a total rate of 140Hz and 38/140=0.27 so
  - Any overlap X between new Electron and Gamma PDs will impact the total overlap as 0.27*X
  - Total overlap will be 0.21+0.27*X
  - We need X to be less than ~0.1 to get the overlap down under 25% which is within striking distance of the 20% we need in real data... (more work ahead)
Main Goals: (Best use of our resources; especially people’s time.)
  - Datasets structured and placed to make life as easy as possible
  - A common analyzer (PAT)
  - Maintain an “open” analysis climate - key code stored in CVS*

Special issues for CMS:
  - Computing geographically distributed at unprecedented level
  - Very broad physics program with a diverse needs
    - But ... a limited number of basic physics objects

The plan
  - Primary Datasets (PD) – immutable (trigger based) and split at Tier-0
    - Limited distribution to Tier-2’s central storage
  - Secondary Datasets (SD) - RECO or AOD format, also trigger-based
    - Produced from PDs at Tier-1’s and distributed to central storage at Tier-2’s.
  - Central Skims produced at Tier-1’s by dataOps.
    - Very few initially – for key applications that cannot be served otherwise
  - Group skims
    - Run on datasets stored at Tier-2’s. Flexibility in choice of event content but provenance must be maintained. Approved by conveners and expect to have a tool allowing them to be registered to global DBS. Subscribable to Group space.
  - User Analysis skims
    - A dataset that is no more than one skim away from provenance.

*Draft of analysis code guidelines from Analysis Tools Group currently under review
Secondary Datasets

- **Design**
  - RECO initially, AOD at some point
    - For RECO format, there will be no more than ~3 SD per PD
    - Each SD a central produced subset of one PD
      - Select no more than ~30% of events

- **Purpose and functionality**
  - Distributed to central data storage at Tier-2s
  - Provides more targeted and manageable datasets
  - When small, will be stored in FEVT format at Tier-1s
    - To allow re-reconstruction without accessing the PD
  - When large, will not be stored
    - Re-reconstruction of full PD required, with reconstitution of SD from immutable information
Group & User Skims

- Group Skims
  - Designed by groups, run on Tier-2’s
    - Run on any data stored at Tier-2’s and stored in group space
    - Secondary Datasets likely the main source for Group skims
    - Output expected to be PAT-tuples in most cases

- User Skims
  - Whatever you need
    - Final skims by users specific to their unique needs for analysis
    - Only restriction is that any analysis up for public approval must be done on datasets that are at most one step from provenance tracking
For each trigger path define an ES trigger bit (possibly 2?) to select “interesting” subset(s) of events to go to low latency CAF queue for:
- DQM
- Basic physics analysis
- Calibration

The _ES bit(s) set by the main trigger path
- Just make additional requests (they do not reject events)

[ Simple Example ] Modify the threshold:
- HLT_Ele15_LW_L1R
  - threshold = ET > 15
  - ~15% bandwidth \( \Rightarrow \) ET > 25 \( Z \rightarrow ee \sim 90\% ; W \rightarrow ev \sim 80\% ; t\bar{t} \sim 70\% \) (efficiency relative to original)

NB: Low latency ES at CAF will not have the most up-to-date AlCa results – so will not be used for publication of physics!

NB: Secondary datasets can be defined with ES bits
- Secondary datasets can be used for physics publications.