FDR in Top Physics WG

-- A full dress rehearsal of Analysis Model --

Akira Shibata, New York University
Marcello Barisonzi, DESY

16.January.2008

Motivation

- Need to address the question: "how do we analyse the real data?"
- Identify how all the software tools fit into the big picture of physics analysis. Stream AOD, database, DPD, TAG, EventView, ARA etc.
- Develop strategy for physics analysis with early data in the top working group.
- We will present our ideas today. On going discussion with D. Charleton (FDR), M. Bosman (top), K. Assamagan (TAG), A. Holloway (Stream), T. LeCompte & L. D. Ciaccio (SM) et al. We would love feedback today.
- **Inviting contributors. Take part in physics analysis aimed at early data in the way we will analyse real data. See page 6 and 7.**

Aim of the Project

- ✓ Production of common $D^{1/2/3}PD$.
- ✓ Event selection using TAG/Trigger/Condition DB.
- ✓ Study physics trigger menus for all luminosities.
- ✓ Replicate DPD and analyse remotely (down to T2).
- ✓ Development of common tools for analysis.
- ✓ Study usability of ARA (how fast?)
- ✓ Exercise physics analysis with early data using all the goodies above...

FDR Parameters

- FDR-1: February '08
  - Integrated lumi ~1 pb-1.
  - Mostly at 10^{31} cm^{-2} s^{-1} and short period at 10^{32} cm^{-2} s^{-1}. No pile-up
  - Lowest unprescaled thresholds: e*10 (21 Hz)*, **μ***10(18 Hz)*, **τ***60 (10 Hz)*, **γ***20 (7 Hz)*, *j120 (9Hz)*.
  - egamma, muon, jetTauEtmiss, minbias, express and calibration streams. Might be more for 10^{32}.

- FDR-2: May '08
  - Integrated lumi ~50-100 pb-1.
  - Include 10^{33} cm^{-2} s^{-1}.
  - With pile-up

cm^{-2} s^{-1}	10^{31}	10^{33}	10^{34}
pb^{-1} s^{-1}	10^{5}	10^{3}	10^{2}
s to $1pb^{-1}$	10^{5}	10^{3}	10^{2}
h to $50pb^{-1}$	1500	15	1.5

(Just a reminder)

Physics Contents in FDR

	σ (pb)	σ *BR	Eff.*Acc.	Num at 1 pb^{-1}	Num at 50 pb^{-1}
pp->J/ψ (μ6μ4)				30,000	1,500,000
W		20510	15%	3,077	153,825
Wbb		111	15%	17	833
Z		2015	15%	302	15,113
tT	833	461	10%	46	2,305
t-chan single t	246	80	1%	1	40
	(Eff.*Acc. is a rough estimate. BR includes e/mu/tau)				

- FDR includes Min-bias, QCD, DY, W/Z, B, top etc.
- Signal study in FDR-I not feasible but thousands of Ws to start studying background and reco and trigger perf.
- FDR-II gives enough data for many tT analyses.

Physics Analysis Cases: FDR1

- Little signal for top signal studies but relevant studies possible and the following discussed:
  - Measure efficiency of lepton triggers with tag'n'probe.
  - Study fake rate at 10^{31} with unprescaled trigger.
  - Study strategy for combined trigger, trigger overlap and monitoring trigger.
  - Enough high-pt jet to study jet energy scale. (~ μb for pt > 100 GeV)
  - Multi jet rate and study p_{T3}/p_{T1} and p_{T4}/p_{T1} as a function of eta/E.
  - Ratio of W+1/W+2/W+3 and more. Low statistics for 4 jets.
  - Estimate QCD events in W/top samples.
- **May overlap with SM activity? Contributors are welcome!**

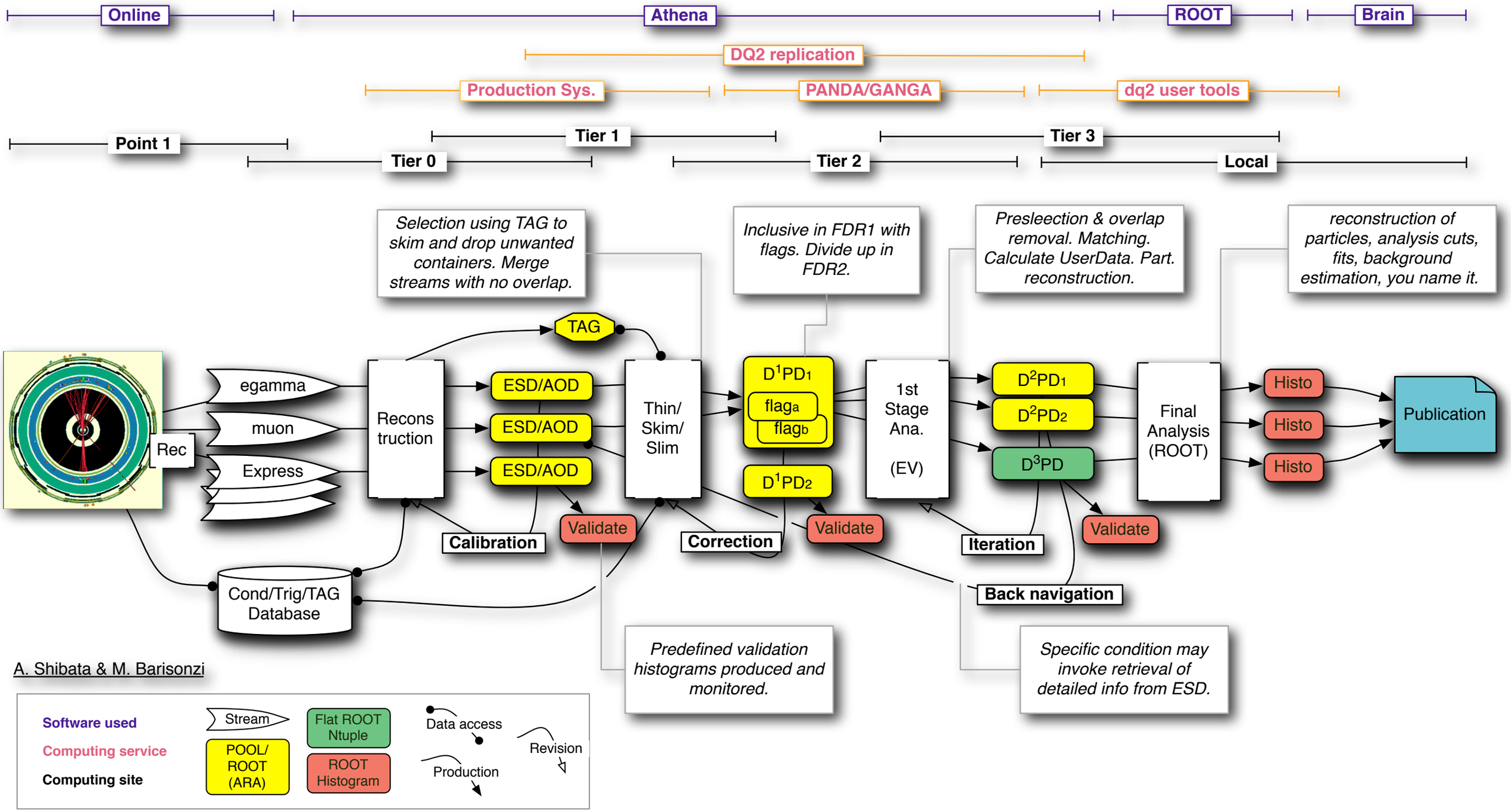
From BNL Jamboree in Dec. 20

Physics Analysis Cases: FDR2

- Good prospect for tT reco already seen with (pre) commissioning analysis.
- Feasible studies that goes beyond the current studies:
  - tT cross section without missing E_T.
  - Differential cross section as a function of top p_T
  - Delta phi/eta between tops (or just measure bbbar directions to approximate esp. with high pt top.)
  - Study soft muon tagging for cleaner top sample and study heavy flavour fraction (Wbb, Wcc, Wcj, Wbj) in W+jets.
  - Lower trigger threshold by combining triggers and increase efficiency with ORing triggers.
  - Jet multiplicity and JES versus luminosity (pile-up). More discussion needed.
- **Let us know if you are interested in working on these.**

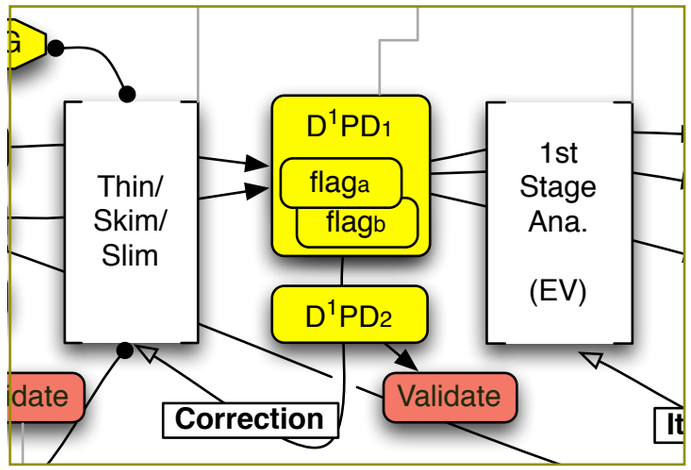
From BNL Jamboree in Dec. 20

Full Analysis Flow Model I



A model encompassing D^1PD, D^2PD and D^3PD/ntuple. The main analysis model to be exercised in FDR.

D¹PD with FDR-1



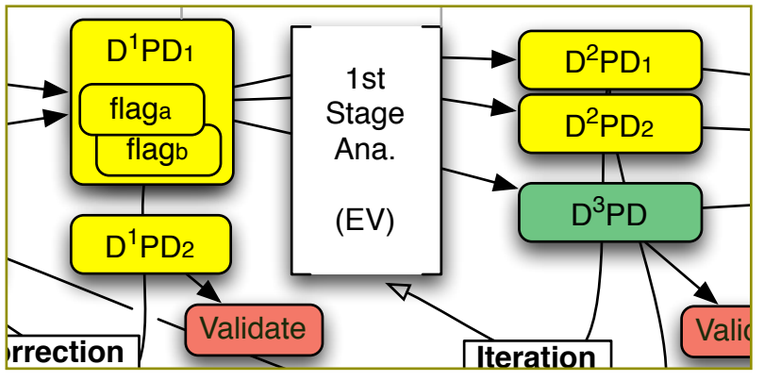
e/mu inclusive D¹PD:

ElectronCollection / PhotonCollection
MuidMuonCollection / StacoMuonCollection
Cone4H1TopoParticleJets / Kt4H1TopoParticleJets
MET_* / ObjMET_*
VxPrimaryCandidate
TrackParticleCandidate
CaloCalTopocluster
Trigger Decisions / Objects

Need input to finalise the thinning and slimming. More discussion in PAT later.

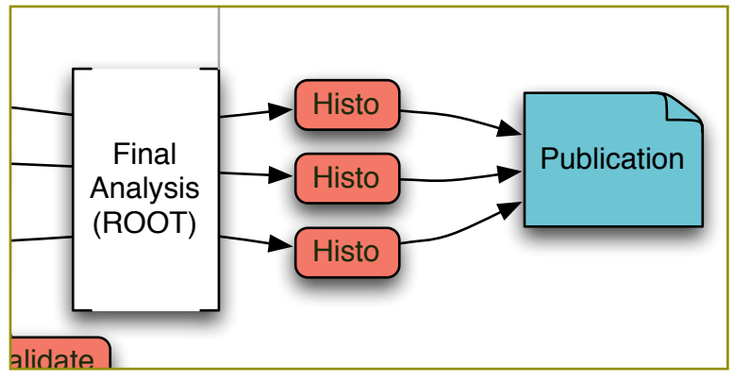
- D¹PD defines analysis datasets and their contents. Only relevant containers and details are selected to reduce data size without compromising physics.
- Generic contents for FDR-1 possibly shared by a number of groups. We'd start with "inclusive lepton (e/mu)" D¹PD similar to the one discussed in the AM report.
- Include egamma/muon streams removing overlapping events and bad quality events using TAG database.
- Common validation plots should be implemented
- Tools based on: AODtoDPD by Sven plus production transformation.
- For FDR-2 add specific Z/W/top-like D¹PDs using TAG objects. More later.

$D^{2/3}PD$ Contents



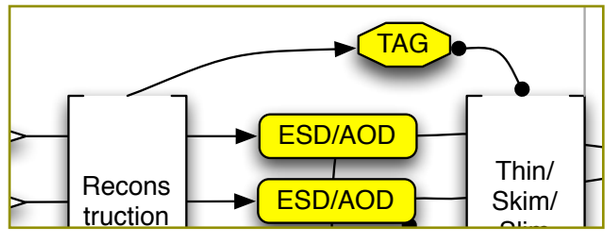
- D^2DP and D^3DP have more specific contents for specific analysis.
- D^2DP is in POOL format. D^3DP has equivalent contents in flat ntuple format.
- Contents follows CSC TopView ntuple. TopView analysis with common object preselection, overlap removal, trigger match, and additional UserData coming from existing tools, including "commissioning" tT analysis.
- Replicate the $D^{2/3}PDs$ to Tier2.
- Produced using PANDA/GANGA at group level.
- What's the time-scale for $D^{1/2/3}PD$ production?

ROOT analysis



- Recommended to use ARA in final ROOT analysis. Portability issues (only works on linux) keeps flat ntuples still relevant but one should try to move to ARA.
- Several use-cases. e.g.
  - Download $D^{2/3}DP$ to local disk using dq2_get and analyse on local cluster. Size and speed issue should be monitored.
  - Send GANGA job to the computing site that holds the DPD and process remotely. Marcello has an example that works with SFrame.
  - Use PROOF to analyse DPD in parallel using TSelector (e.g. BNL).
- Good public frameworks available such as SPyRoot and SFrame but needs to become compatible with ARA.
- Try and use Tier 2/3. Higher navigability on Grid.

TAG



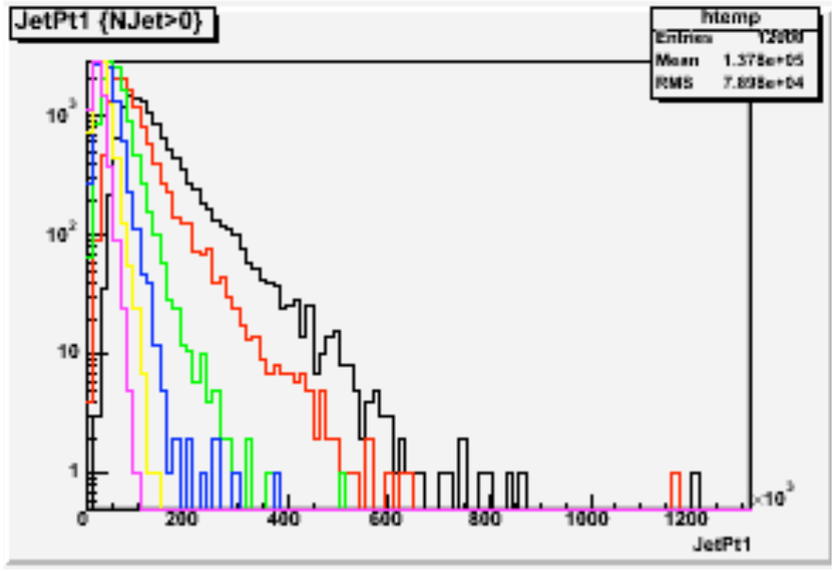
- So far, the weakest-link in analysis but more relevant with real data.
- We have come up with use-cases exercising both file and DB:
  - ✓ Merge streams with no overlap to produce DPD.
  - ✓ Reject events with bad data quality.
  - ✓ Selection on reco mass and other derived quantity as well as trigger decision. Define reproducible common selection for common samples such as inclusive W, Z and Top.
  - ✓ Fast/first method to validate the contents of AOD.
  - ✓ Re-generating iteration of TAG using cache.
- What's the status of the tools and DB? Do APIs exist for DB access? How about trigger DB?

TAG and Data Quality

- Problematic events will be introduced in FDR. First option for physics analysis is to ignore these dodgy events.
- LAr cell problem planned and flagging of bad events. Is this going to be reflected in condition/TAG database? (will "Good for physics" word be used in FDR?
- In any case there'll be bad lumi blocks (4/3000 in streamtest), bookkeeping would be minimal if TAG knows about them.
- Condition database would reflect the most up to date status (and once "bad" may turn "good"), but physics analysis may prefer to say "we used TAG version xx-xx-xx" rather than "we used condition database on this year, this day, this hour, this minute and this second".
- Strong use-cases for TAG selection for early physics analysis. Selection based on TAG object will come later when we are more confident with it.

TAG Contents

- The TAG contains many objects of interest but mostly just the four vector + a little addition:
  - First 6 jets in event, ordered by descending Pt (Jet variables include Bjet Likelihood)
  - First 4 Electrons, 4 Muons, 2 Taus, 2 Photons, MissingEt (Tracking info also included)
  - Trigger words
  - Detector Status & Data Quality
  - Luminosity block
  - A 32-bit wide Physics WG-specific TagWord produced by running Athena algorithm.
- See https://twiki.cern.ch/twiki/bin/view/Atlas/PhysicsAnalysisWorkBookTAG

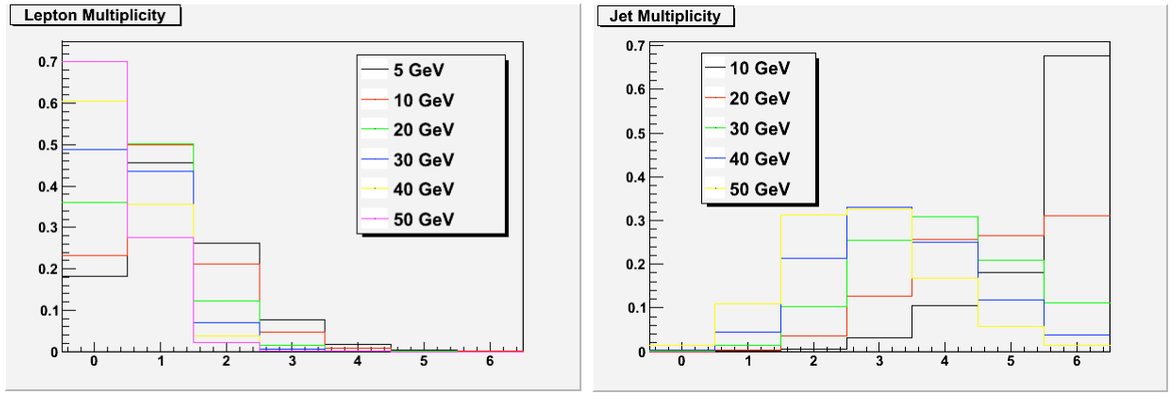


(Jet Pt from tT sample)

Group Word and DPD

- FDR-1:
  - Assuming very early run, objects in TAG may not be well understood.
  - Use TAG to merge streams and remove bad quality data and produce **inclusive lepton (e/mu) D^{1}PD**. To start testing group words, is it possible to flag (not remove) events based on group word (or any other TAG info)?

(TAG from tT events)



- FDR-2:
  - Loose cuts but start defining more specific DPDs. Share with SM and others.
  - Can we use TAG for D^{1}PD->D^{2}PD? Or we have to make new TAG from D^{1}PD?
  - Current suggestion to go into group word (need feedback):
    - **Z-like D^{1}PD**: >=2 loose e/mu (one with Pt>20GeV). Lep. Inv. mass > 50 GeV.
    - **W-like D^{1}PD**: >=1 loose e/mu (Pt>5GeV) and MEt>20 GeV.
    - **top-like D^{1}PD**: W-like and >=2 jets with Pt>30 GeV.
    - Maybe dileptonic, fullhad and tau top DPDs?
- TAG need to be copied to D^{x}PD to study group word selection.

Backup Slides (DPD contents)

At the Tip of Analysis Model

As per Analysis Model Report (click to download the file):

D^1PD Primary DPD, POOL format

AOD Thinned, Skimmed and Slimmed to reduce size.

D^2PD Secondary DPD, POOL format

Output of framework analysis. Preselected/Overlap-removed objects and additional UserData

D^3PD Tertiary DPD, Flat Ntuple

The same content as D^2DP but in Flat Ntuple format

It also suggests:

"A comparison to an AOD analysis implementing the final DPD in a single job is straightforward and ensures that results obtained from the DPD can be validated relatively easily" **So we call:**

DPD All-in-one-go DPD

Same as $D^{2/3}$DP above but produced in one go without D^1DP. DPD' to indicate it's format is in Flat Ntuple.

Terms

As per our current convention:

Skimming: removal of events
e.g. Reject events based on TAG selection.

Slimming: removal of details of object information
e.g. Remove b-tag information from PJet.

Thinning: removal of container or object
e.g. Only keep tracks near reconstructed object.

1st Stage Analysis: common framework analysis
e.g. Object selection based on official selection. Common reference analysis. Calculate common UserData. Typically event level study.

Final Analysis: private analysis in ROOT
e.g. You name it! Event level study as well as sample level study.

D²PD Contents, detail

- We have produced functional DPD for CSC, "TopView Ntuple" and we will convert it into POOL based format. Let's quickly review the feedback on TopView and the DPD contents:

  - Pros:
    - Rich set of tools enables common object preselection and overlap
    - Calculation of non-trivial UserData and full analysis.
    - Configurable set of tools to customise the DPD contents.

  - Cons:
    - There is always VERY good reasons not to remove objects.
    - Overlap removal removes objects and things not recoverable.
    - Finalising one common ntuple that makes everyone happy is a mission impossible and can shorten the life of the responsible person.
    - Besides, no one should constrain our analysis more than necessary.
    - Digital divide: for some it's useful, for others it's just a black box.

- Removal of object should be minimised while keeping the pros and transparency.

So what do we do?

1. **No removing objects from primary DPD due to preselection/overlap.**
   Loose objects are frequently useful to estimate background normalisation, and signal efficiency.

2. **Remove tracks and clusters far from reco objects.**
   These are the heaviest containers, which that needs thinning.

3. **Keep the standard POOL DPD.**
   Non EV-aware analysis must run on it just fine.

4. **Use the same datasets for relevant analyses.**
   Keep the contents generic

5. **Analysis specific information should nonetheless be available.**
   e.g. PJet_TRFTag2incl (object info calculated for a particular analysis)

6. **Association should just be a link, no copying redundant information.**
   e.g. Not Electron_Truth_mother_pdgId but Electron.matchTo("Truth").mother(0).pdgId()

7. **UserData should be accessible through "factorised method".**
   e.g. PJet.UserData("TRFTag2incl"), PJet.UserData("Wdecay") etc.

And how do we do?

1. **No removing objects from primary DPD due to preselection/overlap.**
   Selection will just flag: PJet.UserData("PassTopSel"), PJet.UserData("OverlapPJet")

2. **Remove tracks and clusters far from reco objects.**
   Remove if they are away from any reco/trigger/tru objects.

3. **Keep the standard POOL DPD.**
   No new format, just add info.

4. **Use the same datasets for relevant analyses.**
   Per analysis info is saved in persistified EventView. "SemilepView", "SingleTopView" etc.

5. **Analysis specific information should nonetheless be available.**
   As above. They are mostly collection of links and event level UserData.

6. **Association should just be a link, no copying redundant information.**
   Just keep element link of associated object. Also save

7. **UserData should be accessible through "factorised method".**
   Factorise UserData into EDM interface.

Backup Slides (FDR)

Representative sample of trigger items

- Some of the lowest threshold triggers that can run un-prescaled at 10^{31}

Signature	Physics Coverage	Rate @ 10^{31} (Hz)
e10, 2e5	b,c→e, DY, J/ψ, Υ, W,Z, tt	21, 6
g20, 2g15	Direct Photon, photon pairs, γ-jet balance	6, <1
μ10, 2μ4	W, Z, tt, B-physics, DY, J/ψ, Υ	19, 3
j120, 4j23	QCD, high PT final states, multi-jet final states	9, 5
τ20i+e10, τ20i+μ6	Z→ττ	1, 3
τ20i+xE30	W, tt	10
Minimum Bias	Pre-scaled trigger	4

S. Rajagopalan, FDR meeting for U.S.

Summary of EF rates from 13.0.30.4

Event Filter		
Slice	**Rate (Hz)**	
Jet	34.9	(± 0.01)
bjets	14.3	(± 0.06)
Electron	33.7	(± 0.08)
Photon	8.99	(± 0.02)
Tau	33.5	(± 0.07)
Muon	34.7	(± 0.7)
Missing E_T	3.73	(± 0.009)
Total E	0.925	(± 0.003)
Total Jet E	1.67	(± 0.05)
Topological + B-physics	13	(± 1)
Combined	46	(± 1)
Minimum Bias	0.0994	(± 0.0002)
Calibration	206	(± 5)
Total	382	(± 0.1)

→ ~ 4 Hz in 13.0.40

→ Fixed in 13.0.40

S. Rajagopalan, FDR meeting for U.S.

Stream rates

- Streams reflect configuration in 13.0.30.4
- Removed combined stream
  - Overlap with other streams considerably high.
- Muons & Bphysics merged
- Jets and tauEtmiss merged.
- Studies with different stream configuration ongoing.
  - But a baseline has been put in place for FDR-1.

Stream	Rate (Hz)	Unique Rate (Hz)
Combined	43	14.5
jets	48	33
egamma	41	30
tauEtMiss	43	27
muons	35	24
bphysics	13	9
minbias	0.1	0.1
TOTAL	223.1	137.6
express	20	0
calibration	207	199

S. Rajagopalan, FDR meeting for U.S.

FDR-1 Time Line

- January:
  - Sample preparation, mixing events
- Week of Feb. 4: FDR-1 run
  - Stream data through SFOs
  - Transfer to T0, processing of ES and CS.
  - Bulk processing completed by weekend.
    - Including ESD and AOD production
  - Regular shifts: DQ monitoring, Calibration and Tier-0 processing shifts
    - Expert coverage at Tier-1 as well to ensure smooth data transfer.
- Week of February 11:
  - AOD samples transferred to Tier-1s
  - DPD production at Tier-1.
- Week of February 18/25:
  - All data samples should be available for subsequent analysis.
- At some later point:
  - Reprocessing at Tier-1's and re-production of DPDs.
  - FDR-1 should complete before April and feedback into FDR-2

S. Rajagopalan, FDR meeting for U.S.