Skim for electroweak-electron group

Kalanand Mishra

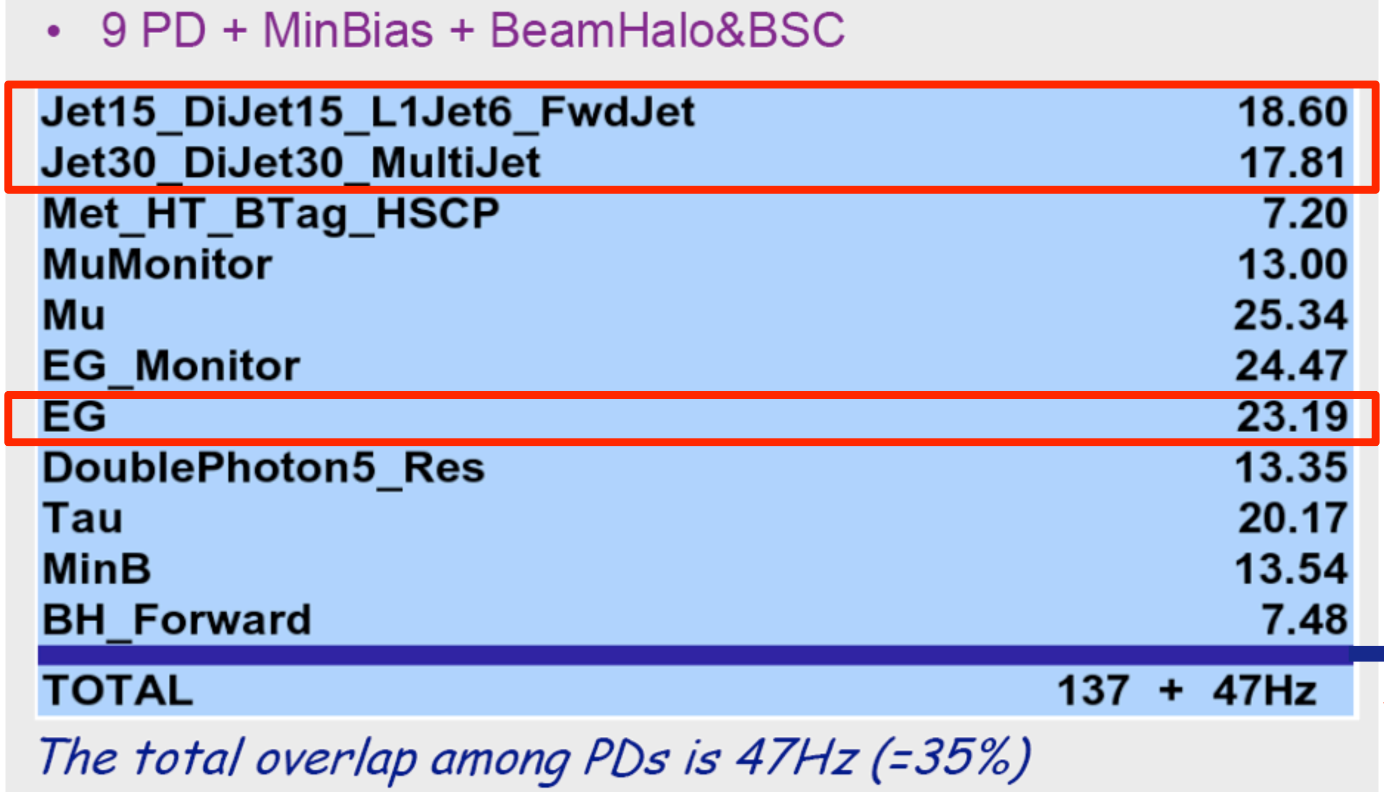
Fermilab

with inputs from Jeffrey Berryhill (Fermilab) & Markus Klute (MIT)

Electroweak electron meeting
(August 14, 2009)

CMS physics primary dataset definition & rates

Outlined in Joe Incandela's talk (CRB meeting: June 18, 2009)
http://indico.cern.ch/contributionDisplay.py?contribId=8&confId=55408



- 9 PD + MinBias + BeamHalo&BSC

Jet15_DiJet15_L1Jet6_FwdJet	18.60
Jet30_DiJet30_MultiJet	17.81
Met_HT_BTag_HSCP	7.20
MuMonitor	13.00
Mu	25.34
EG_Monitor	24.47
EG	23.19
DoublePhoton5_Res	13.35
Tau	20.17
MinB	13.54
BH_Forward	7.48
TOTAL	137 + 47Hz

The total overlap among PDs is 47Hz (=35%)

Jet triggers (Ewk-electron background): **36.4 Hz**

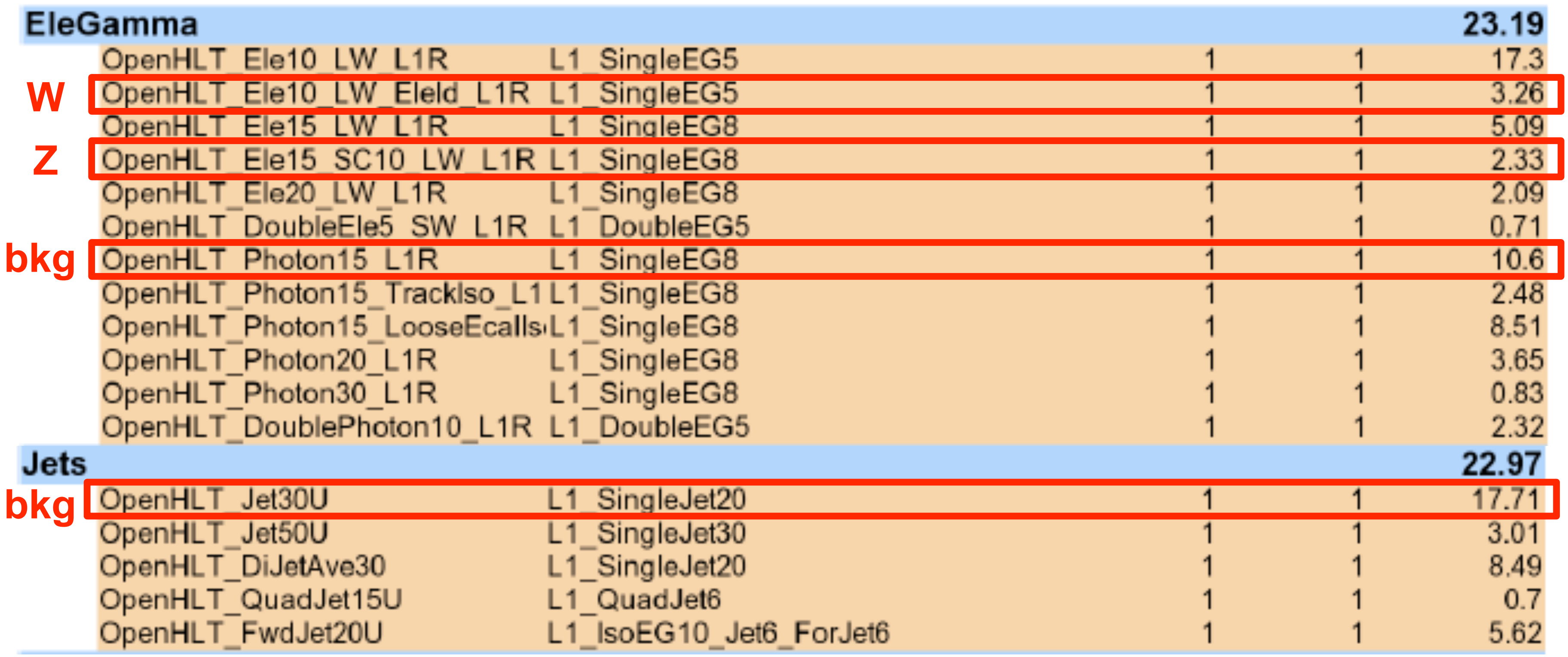
e/γ triggers (Ewk-e signal): **23.2 Hz**

59.6 Hz

These numbers are for rough estimations only. They keep changing !

Ewk-electron data populates 2 PDs that amounts to about 44% of all data.

Need for centralized skims



	EleGamma				**23.19**
	OpenHLT_Ele10_LW_L1R	L1_SingleEG5	1	1	17.3
W	OpenHLT_Ele10_LW_EleId_L1R	L1_SingleEG5	1	1	3.26
	OpenHLT_Ele15_LW_L1R	L1_SingleEG8	1	1	5.09
Z	OpenHLT_Ele15_SC10_LW_L1R	L1_SingleEG8	1	1	2.33
	OpenHLT_Ele20_LW_L1R	L1_SingleEG8	1	1	2.09
	OpenHLT_DoubleEle5_SW_L1R	L1_DoubleEG5	1	1	0.71
bkg	OpenHLT_Photon15_L1R	L1_SingleEG8	1	1	10.6
	OpenHLT_Photon15_TrackIso_L1	L1_SingleEG8	1	1	2.48
	OpenHLT_Photon15_LooseEcalIs	L1_SingleEG8	1	1	8.51
	OpenHLT_Photon20_L1R	L1_SingleEG8	1	1	3.65
	OpenHLT_Photon30_L1R	L1_SingleEG8	1	1	0.83
	OpenHLT_DoublePhoton10_L1R	L1_DoubleEG5	1	1	2.32
	Jets				**22.97**
bkg	OpenHLT_Jet30U	L1_SingleJet20	1	1	17.71
	OpenHLT_Jet50U	L1_SingleJet30	1	1	3.01
	OpenHLT_DiJetAve30	L1_SingleJet20	1	1	8.49
	OpenHLT_QuadJet15U	L1_QuadJet6	1	1	0.7
	OpenHLT_FwdJet20U	L1_IsoEG10_Jet6_ForJet6	1	1	5.62

- The primary datasets are large !
- A PAG/POG may need small part of PDs ➔ need for centralized skims.
- Standard OpenHLT codes and samples to be used for central skims.
- One strategy is to filter events by trigger bit and slim the event content.

Proposed specifications for EWK-e skim(s)

✦The EWK skims for W and Z will be of use to the Egamma, JetMET, TOP, SUSY, and Higgs groups.
✦This has some impact on max. rates and thresholds.
✦Can keep relevant RECO+AOD as event content (need to finalize !).
✦We should try to design central skims now, group skims to be implemented later (for example by using higher thresholds)
✦A proposal for central skims could look like the following (initial /low lumi case).

•**W→ev signal:** PD=EleGamma; Trigger=Ele10_SW_eID; skim definition= trigger+loose id e P_T >10 GeV (can be higher, but 10 GeV make it useful for other groups).
•**Z→ ee signal:** PD=EleGamma; Trigger=Ele15_SW; skim definition= trigger+ e P_T >15 GeV

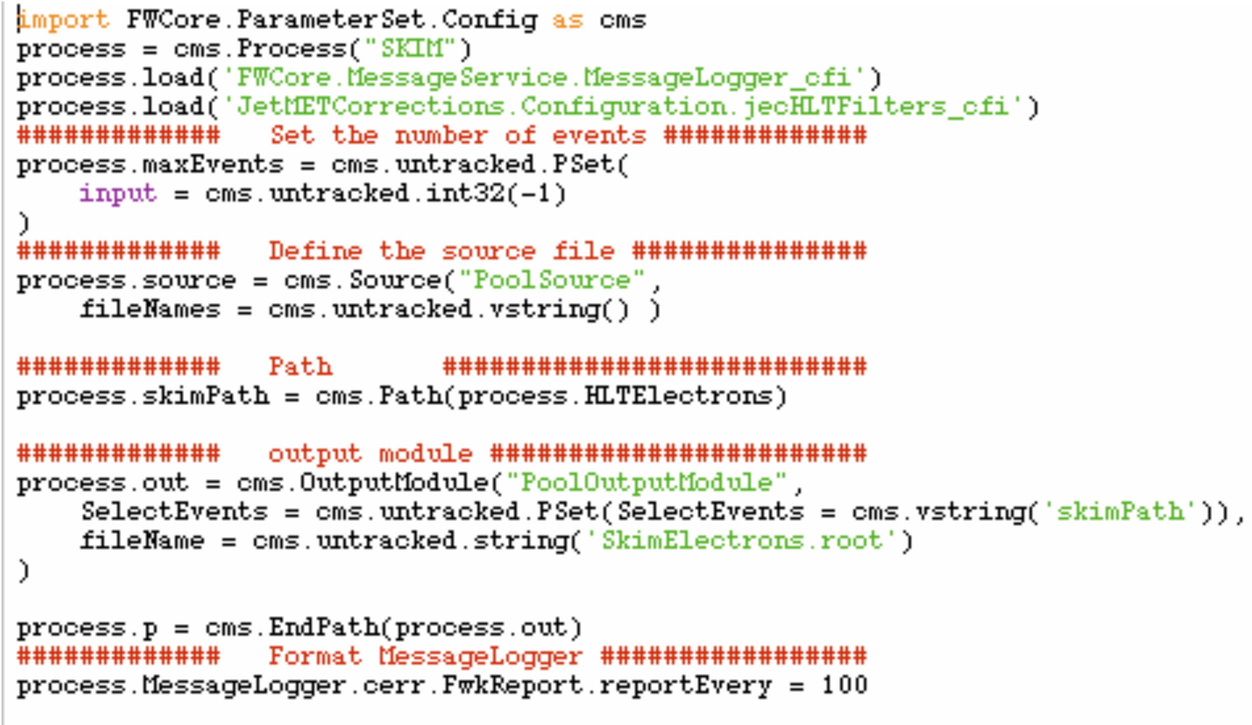
•**W→ev background:** PD=Jets; Trigger=Jet30; skim definition= trigger+loose e P_T > 10 GeV (should not be lower than the one for the signal skim).
•**W→ev background:** PD=EleGamma; Trigger=Photon15; skim definition= trigger+loose e P_T > 10 GeV (should not be lower than the one for the signal skim).
•**Z→ee background:** can use the W signal and background skims.

Note: For background samples, can run over secondary datasets.

Logistics and code

- ✦Have used official HLT machinery for this purpose
- ✦Tested extensively an ED filter to just filter events using trigger bit
- ✦Tested extensively an ED filter to filter + slim event content in CRAFT data and Summer09 samples

Simple code:
Developed for filtering electron-triggered events



```
import FWCore.ParameterSet.Config as cms
process = cms.Process("SKIM")
process.load('FWCore.MessageService.MessageLogger_cfi')
process.load('JetMETCorrections.Configuration.jecHLTFilters_cfi')
#############   Set the number of events #############
process.maxEvents = cms.untracked.PSet(
    input = cms.untracked.int32(-1)
)
#############   Define the source file ###############
process.source = cms.Source("PoolSource",
    fileNames = cms.untracked.vstring() )

#############   Path       ###########################
process.skimPath = cms.Path(process.HLTElectrons)

#############   output module ########################
process.out = cms.OutputModule("PoolOutputModule",
    SelectEvents = cms.untracked.PSet(SelectEvents = cms.vstring('skimPath')),
    fileName = cms.untracked.string('SkimElectrons.root')
)

process.p = cms.EndPath(process.out)
#############   Format MessageLogger #################
process.MessageLogger.cerr.FwkReport.reportEvery = 100
```

Proposed event content

```
CaloTowersSorted_towerMaker__RECO.
edmTriggerResults_TriggerResults__HLT.
edmTriggerResults_TriggerResults__RECO.
floatedmValueMap_eidLoose__RECO.
floatedmValueMap_eidRobustLoose__RECO.
floatedmValueMap_eidTight__RECO.
recoCaloClustersToOnerecoClusterShapesAssociation_hybridSuperClusters_hybridShapeAssoc_RECO.
recoCaloJets_antikt5CaloJets__RECO.
recoCaloJets_kt4CaloJets__RECO.
recoCaloJets_sisCone5CaloJets__RECO.
recoCaloMETs_metHO__RECO.
recoCaloMETs_metNoHF__RECO.
recoGsfElectrons_gsfElectrons__RECO.
recoPFCandidates_particleFlow_electrons_RECO.
recoPFJets_iterativeCone5PFJets__RECO.
recoPFJets_kt6PFJets__RECO.
recoPFJets_sisCone7PFJets__RECO.
recoPreshowerClusters_multi5x5SuperClustersWithPreshower_preshowerYClusters_RECO.
recoSuperClusters_correctedMulti5x5SuperClustersWithPreshower__RECO.
recoSuperClusters_multi5x5SuperClustersWithPreshower__RECO.
recoVertexs_offlinePrimaryVerticesWithBS__RECO.
triggerTriggerEvent_hltTriggerSummaryAOD__HLT.
EventAuxiliary
edmTriggerResults_TriggerResults__HLT8E29.
edmTriggerResults_TriggerResults__SKIM.
floatedmValueMap_eidRobustHighEnergy__RECO.
floatedmValueMap_eidRobustTight__RECO.
recoBeamSpot_offlineBeamSpot__RECO.
recoCaloClusters_hybridSuperClusters_hybridBarrelBasicClusters_RECO.
recoCaloJets_iterativeCone5CaloJets__RECO.
recoCaloJets_kt6CaloJets__RECO.
recoCaloJets_sisCone7CaloJets__RECO.
recoCaloMETs_metNoHFHO__RECO.
recoCaloMETs_met__RECO.
recoPFCandidates_particleFlow__RECO.
recoPFJets_antikt5PFJets__RECO.
recoPFJets_kt4PFJets__RECO.
recoPFJets_sisCone5PFJets__RECO.
recoPreshowerClusters_multi5x5SuperClustersWithPreshower_preshowerXClusters_RECO.
recoSuperClusters_correctedHybridSuperClusters__RECO.
recoSuperClusters_hybridSuperClusters__RECO.
recoSuperClusters_multi5x5SuperClusters_multi5x5EndcapSuperClusters_RECO.
recoVertexs_offlinePrimaryVertices__RECO.
triggerTriggerEvent_hltTriggerSummaryAOD__HLT8E29.
```

Summary: Keep the following

- ✦Trigger objects, beam spot, primary vertex information
- ✦Electron: gsfElectrons, id / isolation maps,
- ✦Super cluster: EB+EE super clusters, pre-shower SCs, corrected SCs
- ✦Jets: all jet algorithms, MET: CaloMET
- ✦Particle flow objects

Typical event size goes down from ~1MB to ~30 kB ➡ a factor of 30 improvement !!!

Disk space requirements

1 snow mass year = **10^7 second** [assumes 30% duty cycle]
Rate of our PD (signal+background) = 60 Hz
Typical event size in PD = 1 MB (say)

For 1 year of continuous CMS running, we can expect the disk storage requirement to be ≈ $10^7 \times 60 \times 1$ MB ≈ **600 TB**

☑ By reducing the data size by a factor of 20 (just by filtering on trigger bit & slimming the event content), we will need to store ≈ **30 TB** at T2/T3 sites.

We can further reduce this data size by applying kinematic thresholds for our analysis skims.

Bottom line: Even with minimal skim (filtering+slimming) we should be fine.