

CSA07 Physics Planning

J. Incandela

Sep. 20, 2007

Computing and Offline Meeting

Thanks to physics conveners & CSA07 contacts, and B. Hegener



CSA07 Primary Datasets (PD) and Physics Skims

- 46 Physics skims

1. Muons 1
2. E-Gamma 4
3. PF/tau 1+7
4. B tag 3
5. Jet-MET 4
6. Diffraction 2
7. QCD 5+3
8. EWK 6
9. Top 5
10. Higgs 6
11. Bottom 2
12. SUSY 7+4

- Green= other groups' skims.

- Primary Datasets (TB)

1. PDTau 3.4
2. PDPhoton 14.6
3. PDMuon 10.0
4. PDElectron 14.5
5. PDBJet 27.5
6. PDJetMET 31.2
7. PDAllEvents 83.4

- **Characteristics:**

- PDs Based on trigger paths
- Events do not occur more than once in any given PD
- ~21% of events appear in more than one PD (not including PD AllEvents)



Skims Info

- 59 PD x skim combinations (secondary datasets)
- Output size estimated by physics groups
 - Varies between 3 GB - 13 TB
 - Average size: 1.9 TB
 - 94 TB of skim output

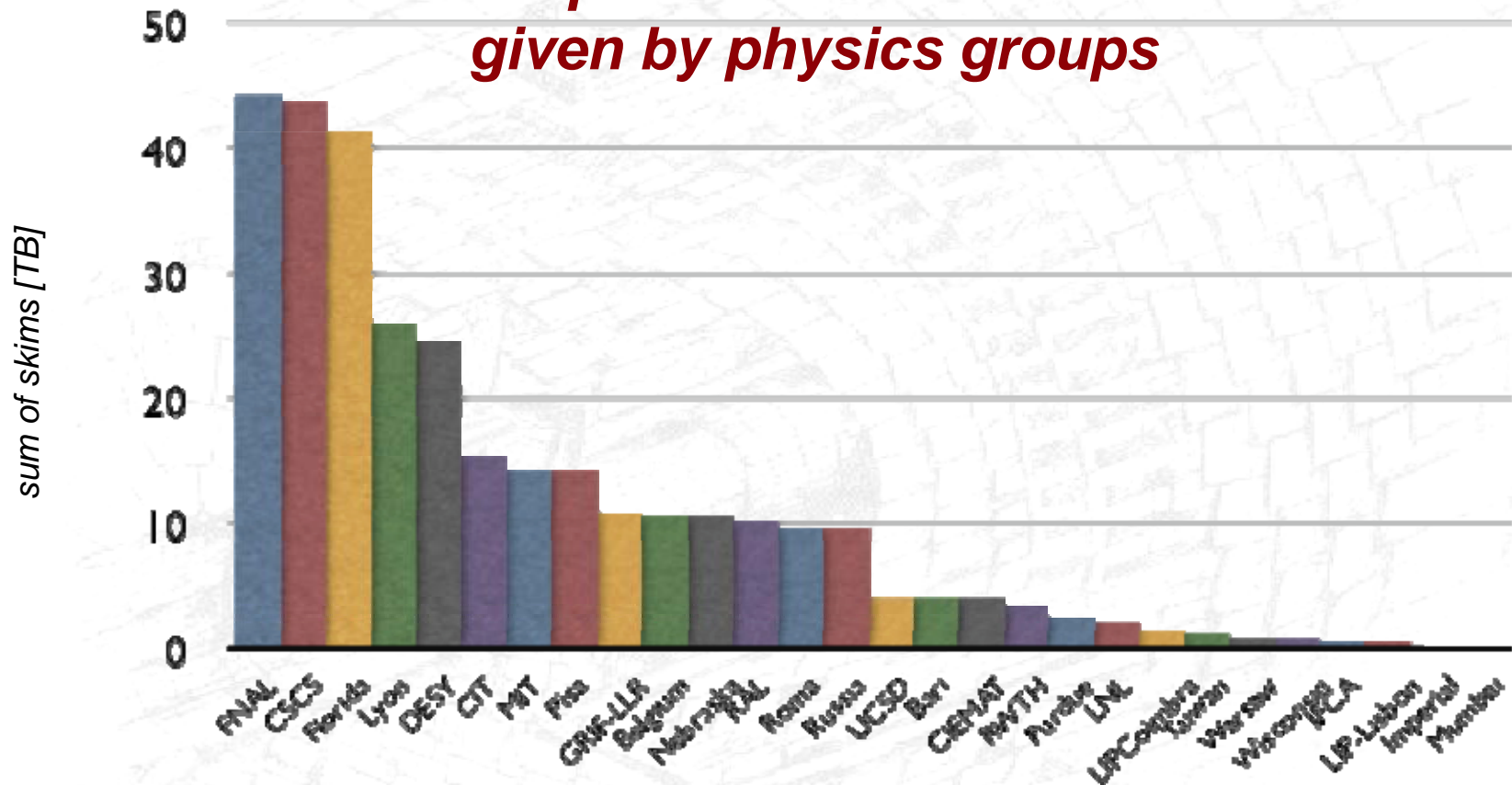
Full table distributed via hypernews

<https://hypernews.cern.ch/HyperNews/CMS/get/skims/8.html>

***Compiled from information
given by physics groups***

- *3.9 copies per skim*
- *Data not well distributed*
- *Current plan means between 1 - 44 TB per centre*
- *~ 300 TB in total*

Compiled from information given by physics groups





Physics Analyses

- Several Activities Planned for CSA07
 1. Monitoring skims (CSA07 “Data”)
 - Run (existing + new) validation code on skims ~daily to watch for problems
 - Groups have latitude here to try what they like
 2. Analyzing skims (CSA07 “Data”)
 - Some (few) groups will analyze skims as they appear.
 3. Analyzing Spring and Summer 07 samples
 - Most groups will run on 1_3 and 1_5 samples to prepare for the 2007 analyses pre-approvals scheduled for the 2-3 weeks following the October Physics week.
 4. Monte Carlo Production and analysis (CSA07 “MC”)
 5. Analyses on the CAF
 - 3 Analyses (Z' and $H \rightarrow WW$ searches, W/Z EWK)



Signal in Data & Express Stream

- Add signal to CSA “data”
 - Zprime, and SM Higgs
- Events corresponding to $\sim 1 \text{ fb}^{-1}$ to be distributed randomly throughout the CSA “data” sample.
 - Would propagate to PDs (all significant decays of H and Z' would be included)
 - Would like to skim a fraction of the leptons PDs to CAF where “real-time” monitoring would run. This would be a lepton or dilepton skim tuned to match CAF resources available for this.
 - H and Z' signals could be seen there and also at T2s
 - At T2s, can try to see other decay modes & would have more more complete samples for more detailed studies



Example: B Tag in CSA

- **B tag**
 - Analyses foreseen are described in <https://twiki.cern.ch/twiki/bin/view/CMS/BTagCSA07> under the section " Btag CSA07 Analysis Exercises ".
 - Main topic is the development and studies of the different methods to measure the performance of the b-tagging algorithms, the study of the effects of a displaced beam-spot and misalignment on vertex reconstruction and b-tagging, including recalibration of the algorithms under these conditions.
- **Tagger Recalibration and Performance Determination**
 - Recalibrate probability tag, combined tag, b->e tag and b->mu tag; measure performance using b tag and vertex validation code where possible; Produce nice performance nice plots for note; Do with and without recalibration of b tags.
 - "Data" samples: QCD, ttbar, e/mu from b
 - "MC" samples: bbbar, ccbar (?) + Bs->JPsi+Phi (?)



Example: EWK in CSA

- Highest priority 2007 Analyses and notes
- Want to run jobs at CERN on 1_5_2 SM samples for the 2007 analyses notes:
 - W, Z production into electron, muon, tau
 - ZZ and WZ production
 - Would not help to copy them to their regional T2s
- Skim monitoring
 - One person assigned to each of the EWK skims
- CAF
 - forward-backward asymmetry (because the people involved are not the same as for the notes)
- Real-time analysis of CSA data at T2
 - Investigating what can be done



Example: $t\bar{t}$ in CSA

- Estimated T2 activity
 - A quick-and-dirty calculation yields several 100k of jobs of 1000 events each for the analysis effort.
- Highest priority 2007 Analyses & notes
- FastSim for some systematic studies in the framework of the 2007 analyses.
 - This FastSim results (=RECO/AOD files) are also planned to be put at some T2's and published in DBS to be used by the full (for example) top quark PAG. This will not eat CPU, but will eat disk space. This is needed if we want to have some consistent sets of samples with different settings for systematics. Indeed both activities clash and the CERN T2 could maybe be (part) of the solution.



CSA07 & Physics

- Physics groups highest priorities:
 - The 2007 analyses which will be presented in Oct. Physics Week and pre-approved shortly thereafter
- CSA07 thus like a “normal” data-taking situation.
 - Analyses running on older, well-understood data
 - New data has to be validated before it is used in analyses. Rarely used immediately for final physics.
- Testing the Computing model
 - Issue at hand is whether or not we can do all the simultaneous effort according to the Computing model
 - Would require 1_3 and 1_5 samples at T2's



T2 CSA07 Subscriptions & Capacities (TB)

This table is not an official tally and not expected to be totally accurate or up to date...

A rough indication of the current situation...

T2	CSA	Capacity
Bari	2	15
Belgium	11	70
CC-IN2P3AF	23	30+10
CIEMAT	2	35
CIT	15	50
CSCS	44	22
DESY	25	50
FNAL	44	700
FLORIDA	40	126
GRIF-LRR	9	15
IFCA	1	45
LIP-Coimbra	1	3
LIP-Lisbon	1	2
LNL	2	35
Lyon	1	170
MIT	14	30
Nebraska	11	111
Pisa	14	15
Purdue	3	180
RAL	10	?
RWTH	3	10+25
Rome	10	34+6
RDMS	10	?
Taiwan	2	21
UCSD	2	48
Warsaw	1	9+7
Wisconsin	1	100

- Tier 2s
 - Much space
 - Even some of those that were heavily or over subscribed (as indicated in table) have already adjusted subscriptions downward
 - Can we fit it all?
 - both the CSA07 data and the Spring and Summer 07 samples
 - Alternatively
 - Should we reduce numbers of copies and push for more remote access?



Thoughts on Analysis & Data Placement (for discussion)

- Tier 1 analyses
 - Concerns about space and reliability of access to T2s- somewhat correlated
 - Concerns about access \Rightarrow physics groups want local copies of skims
 - Several physics groups have requested to operate their analyses for 2007 papers using 1_3 and 1_5 at T1s (particularly CERN)
 - This may be possible to some extent but needs to be limited and coordinated
 - Not a true test of the Computing model
- Alternative that we appear to be adopting
 - Place important pieces of 1_5 and 1_3 samples at T2s of related physics groups +/- or T2s which are ~reliably linked at present to those locations
 - Migrate “almost” all analyses to T2s, including analyses of CSA07 samples which will occur later
 - For the small number of cases where huge samples are needed, (validation/calibration of algorithms, triggers etc.) consider leaving these at T1s if there are few such cases, and few overall passes.
 - Consider reducing number of skim copies if necessary/desirable



Pros and Cons

- Pros

- Test of something resembling expected conditions
- All in all, we learn a lot early
 - If Computing has problem with links, data placement, etc.
 - The much is learned + strong motivation to improve things
 - Many people would start to use T2s for analysis
 - Again, much learned + strong motivation to improve things

- Cons

- Physics analyses may be disrupted
- Potentially higher level of frustration all around
 - Better now than when we have real data?



Recent Developments

- Signal Mix and CAF
 - Incorporate some signal into the samples
 - Agreed this will go into the PYTHIA soup and is done
 - CAF Skim
 - Leptons only => 3 skims of 3 PDs *at most*
 - Agreed to use/modify existing lepton skim cfgs and:
 - AOD output
 - Conveners of Higgs, SUSY-BSM, EWK will send info to Christoph
 - Need to understand available CPU & storage and tune skim accordingly
 - Oliver has gotten some indication resources will be available
 - It has tentatively agreed to do the CAF skim at T0
- The data and analysis placement model during CSA07
 - Agreed to start moving heavily used 1.3 and 1.5 samples to some T2s



Still to resolve

- Choice of two options for the Re-reco step in CSA07
 1. Re-reco the CSA07 “data” with 100 pb-1 align/calib scenario
 2. Reco the original MC samples with 100 pb-1 align/calib scenario to basically reconstruct the current 1.5 samples but now with HLT and the 100 pb-1 align/calib scenario
 - Physics conveners were undecided, have one week to consider

More Information



Estimated Sizes of Physics Skims

- There are 46 -49 skims
 - Muons (1)
 - 23 TB ? (AOD+RECO)
 - E-Gamma (4)
 - 0.16 TO 3.2 TB (RECO)
 - 0.025 to 0.5 TB (AOD)
 - PF/tau (1+7)
 - ?
 - B tag (3)
 - 4 - 10 TB (AOD)
 - Jet-MET (4)
 - 4 – 10 TB (AOD)
 - 25 TB (FEVT)
 - Diffraction (2)
 - 3 - 470 GB (AOD)
 - QCD (5+3)
 - 0.074 – 13 TB (AOD)
 - EWK (6)
 - 0.15-0.8 TB (AOD)
 - 1.0 TB (RECO)
 - Top (5)
 - 0.15-0.470 TB (AOD)
 - Higgs (6)
 - ?
 - Bottom (2)
 - 0.2 TB (AOD)
 - 1.0 TB (RECO)
 - SUSY (7+4)
 - 0.005 – 2.2 TB (AOD)



Primary Dataset Usage 1

- PDBJet (1 skim)
 - Top
 - topFullyHadronic
- PDJetMET (7 skims)
 - Higgs
 - vbf_jets + MET
 - JetMET
 - METHIGH_SKIM
 - QCD
 - High PT Jets Event Filter
 - Ultra High PT Jets Event Filter
 - Very High PT Jets Event Filter
 - SUSYBSM
 - JetMet
 - JetMet_HLT
- PDElectron (14)
 - Diffraction
 - gammagammaEE
 - EWK
 - Di-electron
 - Single electron
 - Higgs
 - 2 tau
 - multilepton
 - single lepton skim
 - SUSYBSM
 - ElectronPhoton
 - ElectronPhoton_HLT
 - egamma (AOD&RECO)
 - W+jet-like events
 - Z+jet-like events
 - electron validation
 - high-pT EM validation
 - topDiLepton2Electron
 - topSemiLepElectron



Primary Dataset Usage 2

- PDMuon (19)

- Bphysics

- onia
 - tauTo3Mu

- Diffraction

- gammagammaMuMu

- Bphysics

- onia
 - tauTo3Mu

- Diffraction

- gammagammaMuMu

- EWK

- Di-muon
 - Single mu

- Higgs

- 2 tau
 - multilepton
 - single lepton skim

- QCD

- 2mu

- SUSYBSM

- Muon
 - Muon_HLT
 - MuonsHits

- egamma (AOD&RECO)

- W+jet-like events
 - Z+jet-like events

- Top

- topDiLeptonMuonX
 - topSemiLepMuon



Primary Dataset Usage 2

- PDPPhoton (3)
 - Higgs
 - diphoton
 - SUSYBSM
 - ElectronPhoton
 - ElectronPhoton_HLT
- PDTau (4)
 - EWK
 - Di-tau
 - Single tau
 - Higgs
 - tau_jet + MET
 - SUSYBSM
 - JetMet
- PDAIIEvents (7)
 - JetMET
 - 1JET_SKIM
 - METLOW_SKIM
 - PHOTON_JET_SKIM
 - QCD
 - Jet+X
 - B tag
 - btagDijet
 - btagElecInJet
 - btagMuonInJet



Multi-PD Usage

- PDElectron+PDMuon (5)
 - Higgs
 - 2 tau
 - multilepton
 - single lepton
 - egamma (AOD&RECO)
 - W+jet-like events
 - Z+jet-like events
- PDElectron + PDPhoton(3)
 - SUSYBSM
 - ElectronPhoton
 - ElectronPhoton_HLT
- PDJetMET + PDTau (1)
 - SUSYBSM
 - JetMet

• Summary

– 46 skims total

- 37 use one PD
- 9 use two PD

– Overlaps

• Ave.=1.24 PD/skim