

*ATLAS Software & Computing Workshop
CERN, November 29 – December 3, 2010*

Central dAOD and D3PD Production for the Top Group



*Ulrich Husemann
Deutsches Elektronen-Synchrotron DESY*





Data Access for Top Analyses



- Top group following ATLAS recommendation
 - Starting point: **common dAOD** (more precisely: D2AODM) format
 - **dAOD** = selected AOD events (“skims”), preserving full AOD content, no additional objects stored for now
 - **D3PD** = for all dAOD events: subset of content converted into “flat” TTree format (“ntuple”)
 - **Several D3PD formats** available in the top group
→ running on dAODs saves all of them a lot of time
- Top group data skimming requirements
 - Top lepton+jets and dilepton analyses start from **single high p_T lepton triggers**
→ single lepton (e, μ) skims with lower p_T cuts (currently 15 GeV)
 - **Data-driven** QCD background models → specialized QCD skims
 - All-hadronic top and tauonic final states → multijet skim
 - Today: total of **six dAOD skim flavors**



History of Top dAOD Production



- dAOD effort started in **Spring 2010**, including first attempts of group production with **limited success**
 - Top group's requirements didn't match ProdSys capabilities
 - Little progress → next problem → little progress → next problem ...
- dAODs were needed urgently: **dAOD production "by hand"** on the grid and data replication to all PHYS-TOP group disks
 - Large effort by very few people
 - Job **babysitting** and **bookkeeping nightmare**
 - A few **problems** with missing datasets, failed replication, ...
- Serious work on group production only after summer conference craze (after HCP, ~August)
 - Strong push for **automated top D3PD production** around the same time
 - **Improved contact** to group production team: Minoru Hirose & Junji Tojo
 - ProdSys got **more flexible** and top group **adjusted requirements** in dialog with group production experts



Getting Ready for Group Production



- Moved to group production from first post-HCP data period (G)
- Changes for group production
 - Adaptation of **standard job transforms** (Reco_trf.py, merge_trf.py) to read/write dAODs/D3PDs
 - Updated **skim definitions**: tighter criteria to accommodate more data on limited storage (was necessary anyway)
 - Minimalistic replication scheme: keep only two copies to PHYS-TOP group space
- **Testing** cycle (about 1.5 months in total)
 1. Test software changes in Athena (standalone mode)
 2. Build and install TopPhys cache with required changes
 3. Run full dAOD/D3PD production chain on single run using new TopPhys cache
 4. Compare output with AODs and previous versions of dAODs/D3PDs (event selection for top lepton+jets cross section analysis)
 5. Repeat 1.–4. until error free, then start “open-ended” production request



Current Operating Mode



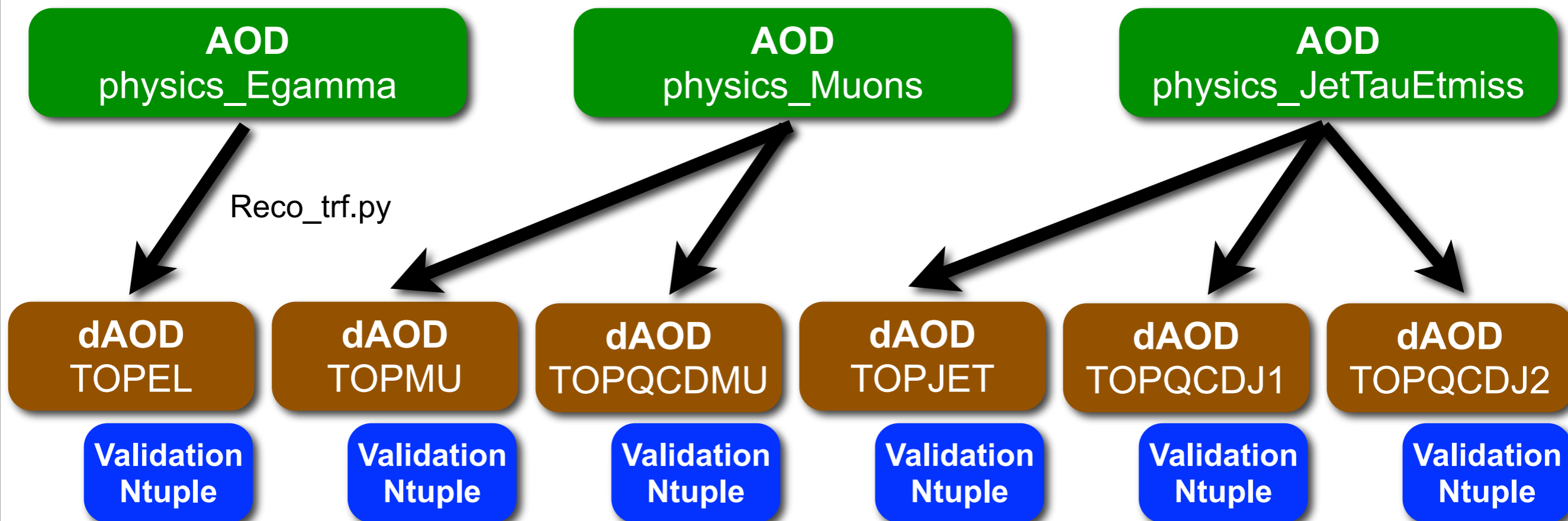
AOD
physics_Egamma

AOD
physics_Muons

AOD
physics_JetTauEtmis

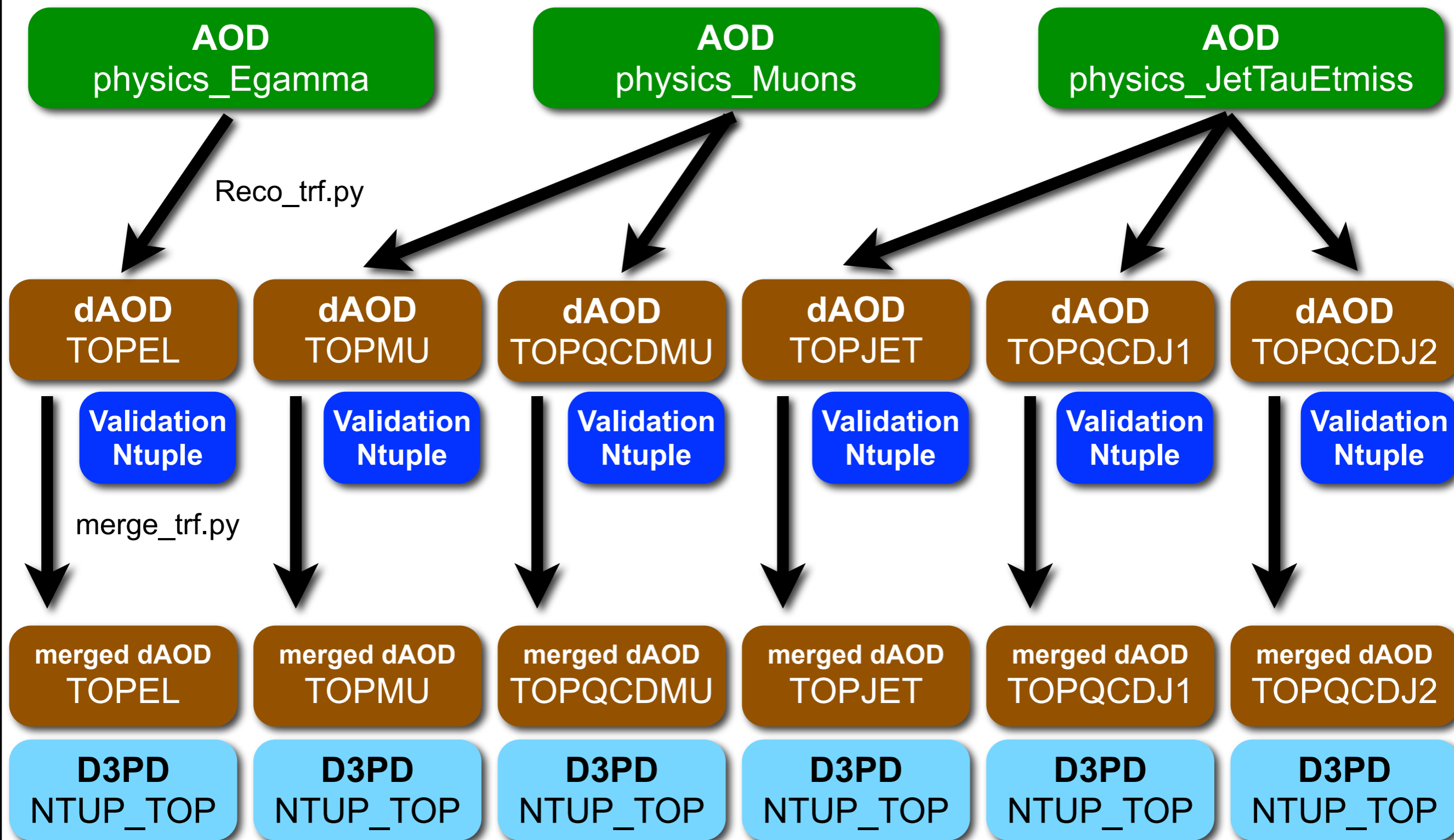


Current Operating Mode





Current Operating Mode





Current Dataset Sizes



- Periods G–H:
(18.4 pb⁻¹)

AOD	:	19774.79	GB	=	1074.72	GB/pb ⁻¹
NTUP_TOP	:	1926.15	GB	=	104.68	GB/pb ⁻¹
TOPEL	:	1566.56	GB	=	85.14	GB/pb ⁻¹
TOPMU	:	1000.31	GB	=	54.36	GB/pb ⁻¹
TOPJET	:	1398.39	GB	=	76.00	GB/pb ⁻¹
TOPQCDMU	:	326.60	GB	=	17.75	GB/pb ⁻¹
TOPQCDJ1	:	21.76	GB	=	1.18	GB/pb ⁻¹
TOPQCDJ2	:	1.51	GB	=	0.08	GB/pb ⁻¹
Total dAOD	:	4315.13	GB	=	234.52	GB/pb ⁻¹

- Period I
(23 pb⁻¹)

AOD	:	12175.43	GB	=	529.37	GB/pb ⁻¹
NTUP_TOP*	:	1396.23	GB	=	60.71	GB/pb ⁻¹
TOPEL	:	1652.65	GB	=	71.85	GB/pb ⁻¹
TOPMU	:	963.20	GB	=	41.88	GB/pb ⁻¹
TOPJET*	:	690.94	GB	=	30.04	GB/pb ⁻¹
TOPQCDMU*	:	372.98	GB	=	16.22	GB/pb ⁻¹
TOPQCDJ1	:	11.92	GB	=	0.52	GB/pb ⁻¹
TOPQCDJ2*	:	0.65	GB	=	0.03	GB/pb ⁻¹
Total dAOD	:	3692.35	GB	=	160.54	GB/pb ⁻¹

*not yet fully processed (status: November 10, 2010)



Software Dependencies



- TopPhys software project
 - Collection of software for top data processing and analysis
 - Key ingredient: **TopInputs** → common object selection code
 - Code for **dAOD production**: skimming algorithms, ...
- TopInputs may require **non-standard** software
 - Example: had to rerun jet and MET reconstruction adjusted to top requirements
 - Solution: build **analysis cache** on top of existing release (5-digit release IDs, latest 15.6.13.3.2) → works well, fairly short turnaround times, but still creates overhead
 - Plan for rel16: trying to keep **in sync** with regular releases
- Further dependencies:
 - Standard job transforms used in group production must recognize output formats
 - Trigger names keep changing
- Bottom line: keeping everything up to data involves a large and diverse group of experts



Summary, Plans, and Wishes



- Experience with group production generally **very good**: experts very responsive and helpful
- Some difficulties to get group production going: standardized ProdSys vs. non-standard top group requests
- Next steps:
 - dAODs/D3PDs from reprocessed data in rel16 (tests ongoing)
 - Efficient production of more than one D3PD format (currently two common formats in top group)
 - Preparation for 2011 data taking: careful look at expected data volume (skimming criteria, D3PD definition, replication scheme, ...)
- Improved replication scheme would be very useful
 - Load balancing: request initial replication to m out of n PHYS-TOP sites so that data is distributed equally
 - Trigger additional replicas if datasets are used often

Backup Slides



- Top group's definitions of dAODs and D3PDs
 - **dAOD** = selected AOD events (“skims”), preserving full AOD content
 - **D3PD** = for all dAOD events: subset of content converted into “flat” TTree format (“ntuple”)
- dAOD strategy (starting from period G)
 - dAODs produced on **central production system** (ProdSys),
 - At least two replicas stored on grid **group space** (PHYS-TOP)
- Currently six different skim **flavors**:
 - Single e, single μ ,
 - Multijet, QCD μ , QCD jets (2x)
 - Details in backup and under <https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TopPhysD2PDEventFilters>



Updated Skim Definitions from Period G



- D2AODM_TOPEL:
 - “ElectronEvent” (“LooseElectronEvent” or “IsolatedElectronEvent”)
 - At least one electron with: $P_t > 15 \text{ GeV}$, Author "Electron" (1 or 3), PID "RobustLoose"
 - Or at least one electron with: $P_t > 15 \text{ GeV}$, Isolation $etcone20 < 4 \text{ GeV} + 0.023 * P_t$
- D2AODM_TOPMU:
 - “MuonEvent” (“StacoMuonEvent” or “MuidMuonEvent”)
 - At least one Staco or Muid muon with $P_t > 15 \text{ GeV}$
- D2AODM_TOPQCDMU:
 - “QCDMuonEvent” (“QCDStacoMuonEvent” or “QCDMuidMuonEvent”):
 - At least one Staco or Muid muon with $P_t > 15 \text{ GeV}$, Quality = "Tight"
- D2AODM_TOPJET:
 - “JetEvent” (“FourJetsEvent” or “FiveJetsEvent”)
 - At least four jets with $P_t > 20 \text{ GeV}$ and at least two jets with $P_t > 40 \text{ GeV}$
 - Or at least five jets with $P_t > 20 \text{ GeV}$



Updated Skim Definitions from Period G



- D2AODM_TOPQCDJ1
 - “QCDJetEvent”
 - At least two jets with $P_t > 15$ GeV and at least one jet with $P_t > 15$ GeV, $EMFraction > 0.8$
- D2AODM_TOPQCDJ2
 - “QCDLeptonJetEvent” (“QCDElectronJetEvent” or “QCDStacoMuonJetEvent” or “QCDMuidMuonJetEvent” or “QCDTrackJetEvent”) :
 - At least one electron with $Author = 1 \ || \ Author = 3$, $P_t > 15$ GeV, $|ClusterEta| < 2.47$, $etcone20 < 6 + 0.04 * P_t$
 - Or at least one Staco or Muid muon with $P_t > 15$ GeV, $|Eta| < 2.5$, number of pixel hits ≥ 1 , number of SCT hits ≥ 2 , $ptcone30 < 6$ GeV and $etcone20/P_t < 0.15$
 - From Period E: Or at least one TrackParticle with $P_t > 15$ GeV, $|Eta| < 2.5$, number of pixel hits ≥ 1 , number of SCT hits ≥ 2 , $ptcone30 < 10$ GeV (sum of all tracks with $pt > 0.5$ GeV) and $etcone20/P_t < 0.15$
- Both require **jet trigger** (names keep changing): EF_j20 (Periods A–D) or EF_j20_jetNoCut (Periods E–F) or EF_j20_jetNoEF (from Period G)



Current dAOD/D3PD Availability



- Periods A–F:
 - All dAODs and PAT D3PDs ready (produced “by hand” on the grid)
 - dAODs replicated to all PHYS-TOP sites (except CERN)
 - PAT D3PDs on various LOCALGROUPDISKs
- Periods G–H (18 pb^{-1})
 - All dAODs and PAT D3PDs ready (produced centrally by Production System)
 - dAODs and D3PDs replicated to two PHYS-TOP sites (in different clouds)
- Period I (23 pb^{-1})
 - Main dAODs and D3PDs ready (e and μ skims), others still queued
- Further details:
 - AMI: search for “data10_7TeV.%D2AODM_TOP%p297%”, “%NTUP_TOP%p297%”
 - dAODs: <https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TopD2PD2010>
 - PAT D3PDs: <https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TopD3PDProduction>



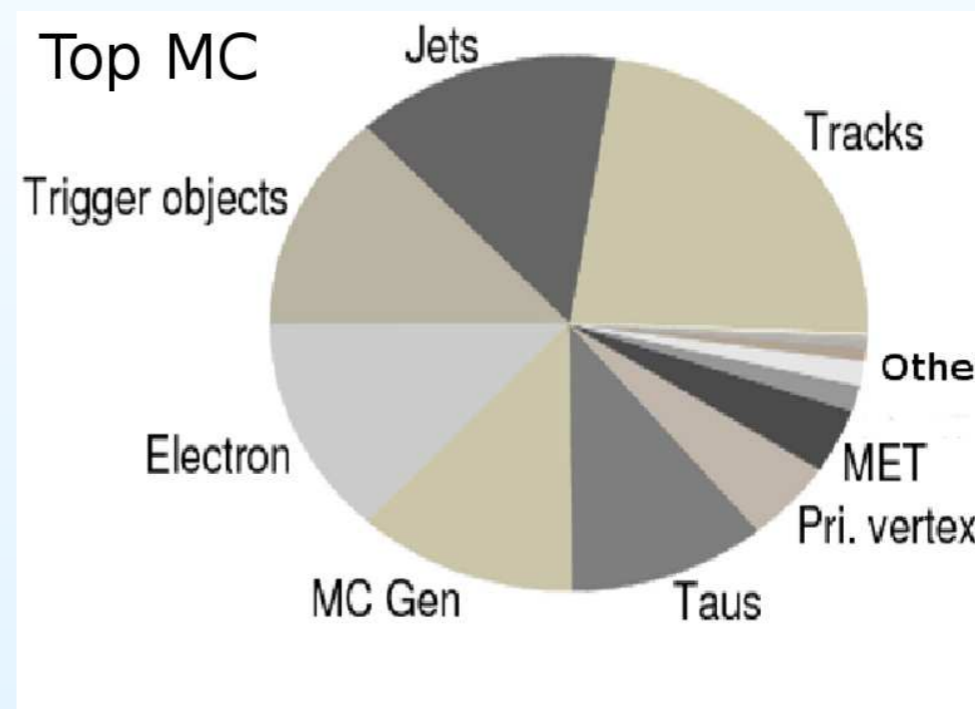
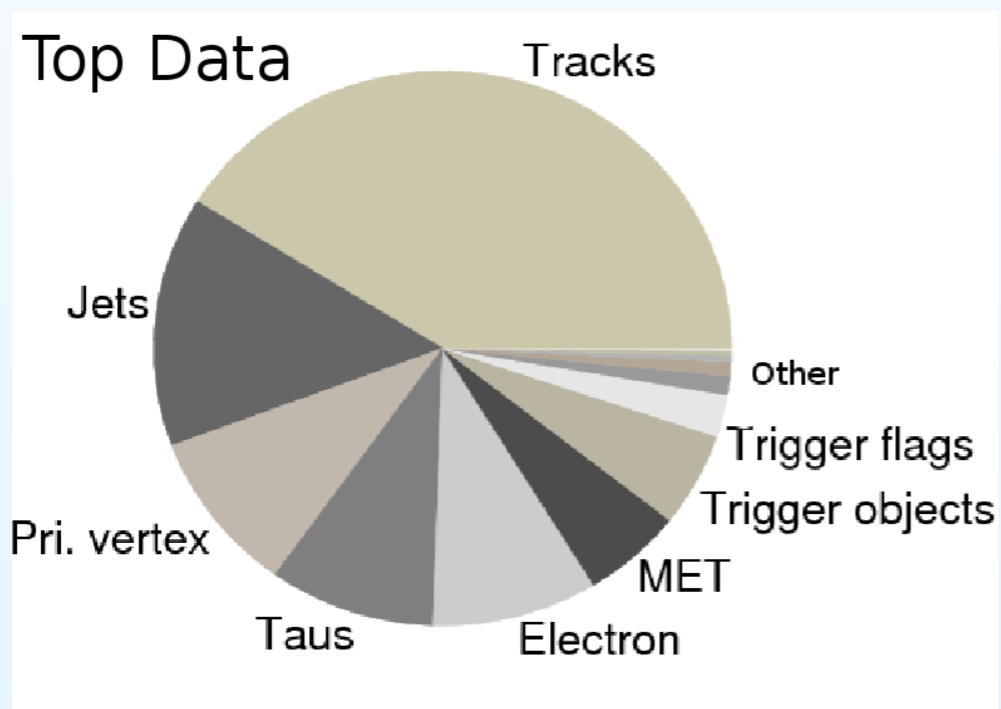
ProdSys Technicalities



- ProdSys step 1: dAOD production
 - Flavor identified by name and 3rd tag
 - p291: D2AODM_TOPEL from Egamma stream
 - p292: D2AODM_TOPMU from Muons stream
 - p293: D2AODM_TOPQCDMU from Muons stream
 - p294: D2AODM_TOPJET from JetTauEtmisss stream
 - p295: D2AODM_TOPQCDJ1 from JetTauEtmisss stream
 - p296: D2AODM_TOPQCDJ2 from JetTauEtmisss stream
 - Validation ntuples for all of the above
- ProdSys step 2: dAOD merging & D3PD production
 - Identified by 4th tag: p297
 - D3PD data type: NTUP_TOP
 - Merged validation ntuples, data type: NTUP_TOPVAL (4th tag: p288)

Size

- From Will the event size for tag 00-00-18
- Data (period F)- 36 per kb event
- TTbarMC- 60 per kb event



[J. Searcy, Top Reconstruction Meeting, 11/04/10]