A Week in the Life of University X ATLAS Group, 2010

(worrying about analysis paralysis)

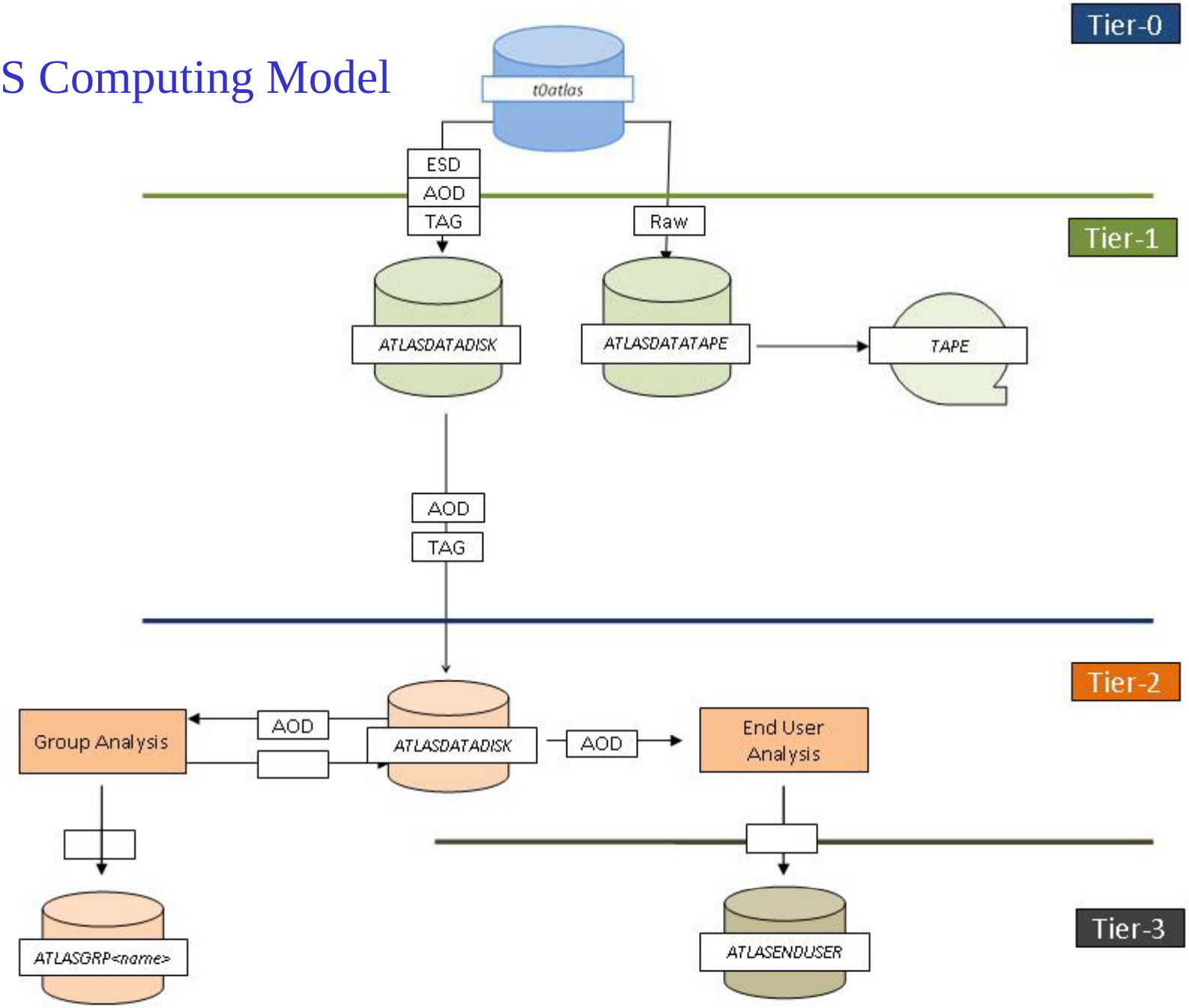
Jim Cochran
Iowa State Univ.

Outline

- reminder: ATLAS Computing Model
- reminder: ATLAS Analysis Model
- assumptions
- description of T3 at University X
- cast of characters in University X HEP group
- “experiences”
- outlook

reminder:

ATLAS Computing Model

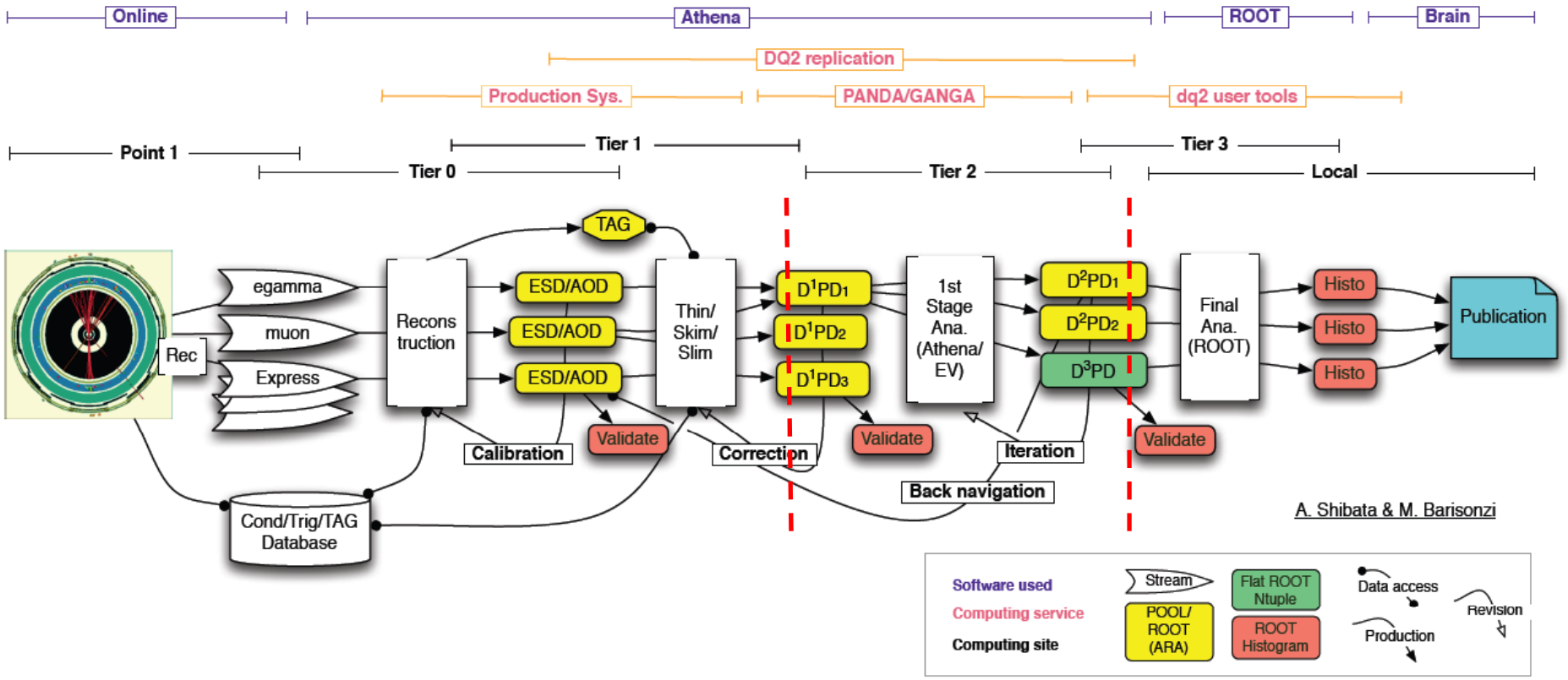


reminder: ATLAS Analysis Model - Overview

As Amir comments, this is still rather fluid

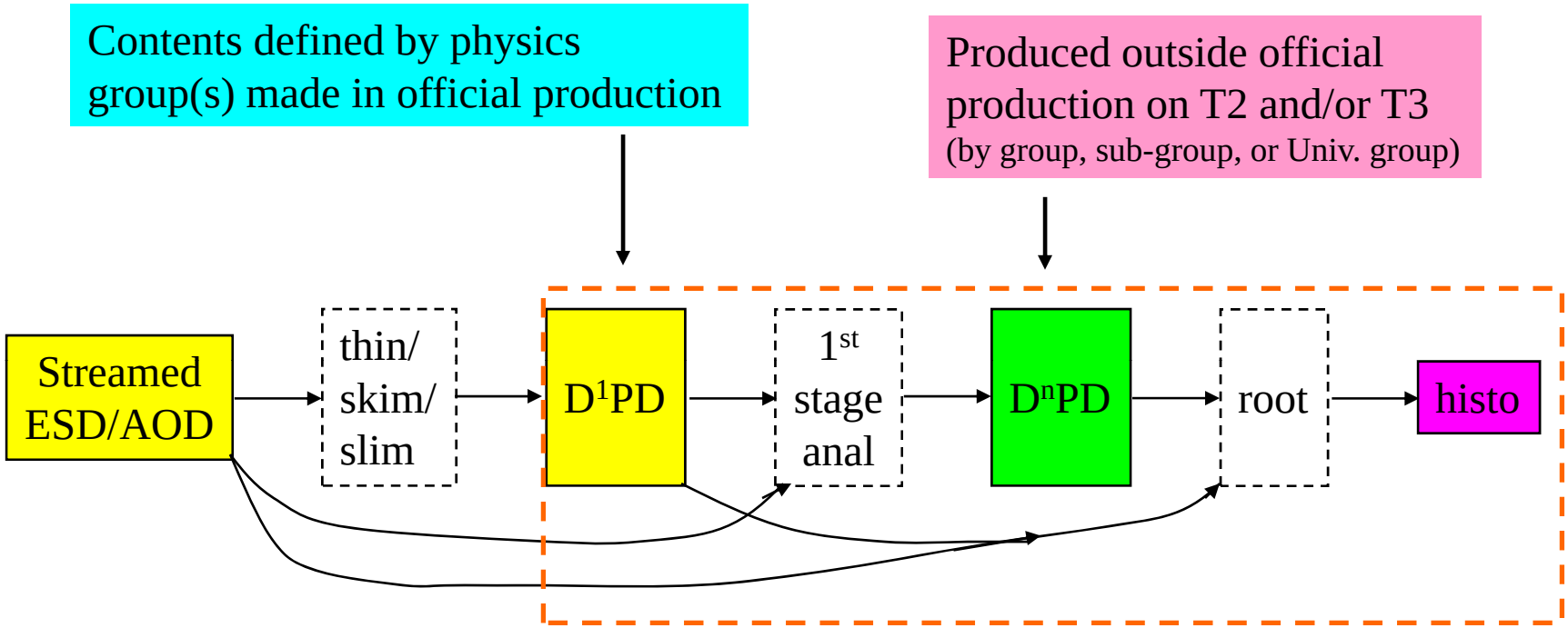
From analysis point of view:

T2 performs AOD/D^1PD → D^nPD [n=1,2,3]



A baseline model encompassing D^1PD, D^2PD and D^3PD/ntuple.

reminder: ATLAS Analysis Model – analyzer view



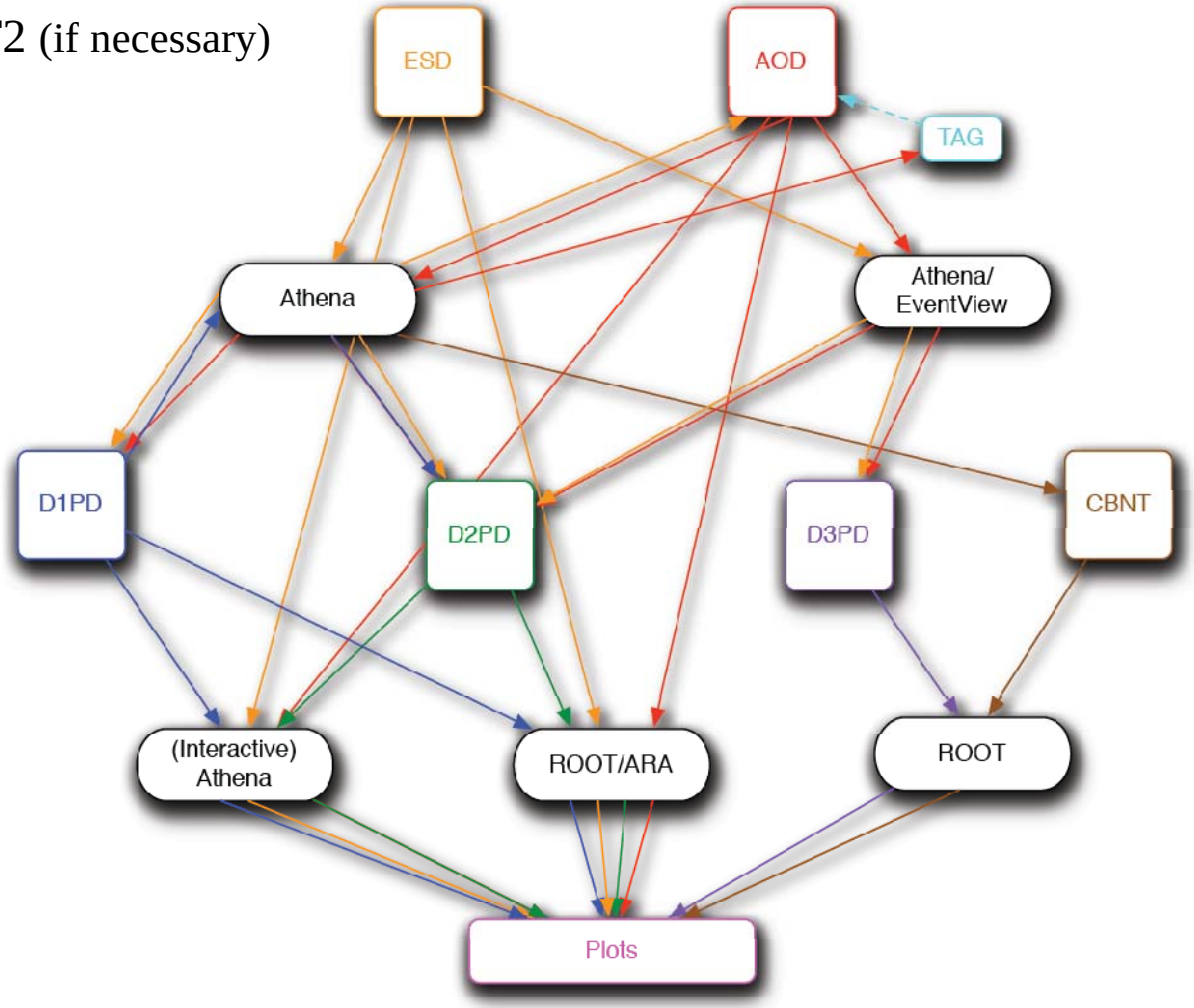
ESD/AOD, D^1PD, D^2PD - POOL based

D^3PD - flat ntuple

Expect there will also be the BAF (BNL Analysis Facility)
Primarily for high-volume parallel (PROOF) DPD analysis
perhaps also for AOD/RAW analysis ?

Expectation is that the “average user” will

- have as a starting point officially produced streamed AOD/D^{1}PD
  (possibly D^{n}PD [n=2,3] produced by group, sub-group, University group)
- produce D^{n}PD [n=2,3] on T2 (if necessary)
- **perform analysis on T3**



Ultimately many paths
to final plots

assumptions

(1 week in 2010)

500 pb^{-1} of data reconstructed
(reprocessing [if necessary] would require 3 months)

~10^9 events

Raw data: (10^9 events) × (1.6 MB/event) = 1600 TB
(including calibration data ~ 160 TB)

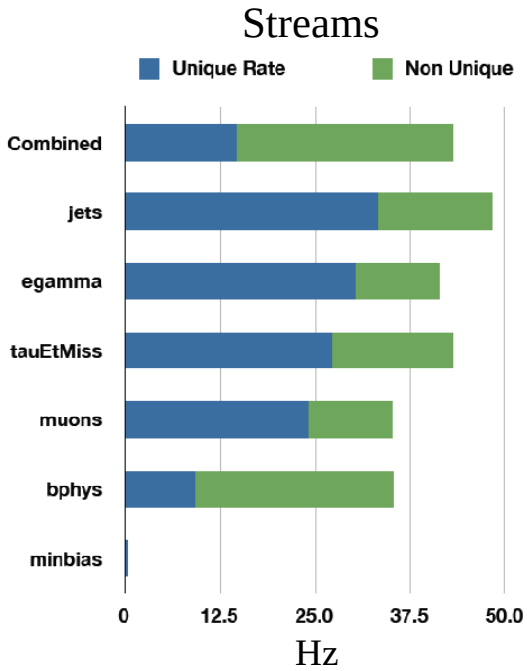
ESD: (10^9 events) × (0.5 MB/event) = 500 TB

AOD: (10^9 events) × (0.1 MB/event) = 100 TB

D^1PD: ~0.15-0.3 × AOD ≈ 15-30 TB

D^2PD: 0.1-1.1 × D^1PD ≈ 1.5-33 TB

D^3PD: < 0.3 D^1PD ?



physics streams are of course smaller

jets	24%
egamma	19%
tauEtMiss	21%
muons	18%
bphys	18%

T3 at University X

OSG site hanging from nearest T2

40 cores

30 TB (not enough for all D[1]PD streams, not to mention needed MC)

PROOF installed

cast of characters in University X HEP group

Postdoc #1  Cell intercalibration of EM calorimeter using $Z \rightarrow ee$

Postdoc #2  D^1PD & D^2PD production for top group (& sub-groups)

Grad. Student #1  Search for SUSY in tri-lepton final state

Grad. Student #2  Measurement of W-boson mass

Grad. Student #3  Measurement of $tt \rightarrow e$+jets cross section

Prof.  Overseer and participates in t-channel single-top search

“Experiences”

Comments

- it’s very hard to predict the future

- some failures/concerns based on input from a user poll I took in Dec. 2007

- others are from my own experiences on ATLAS, Dzero, and BABAR

- still others are just speculation …

- there may be additional resources that I’m not aware of

Postdoc #1

Cell level calibration of EM calorimeter using $Z \rightarrow ee$

Need ESD/RAW

Preliminary studies done using a few ESDs copied to T3

Further studies require access to more data!

Attempts to do this at CAF at CERN (T0) failed
- staging ESDs from tape takes forever, T0 system is simply too overloaded

Next attempt is to use T1 (non-grid)
- staging is faster
- unfortunately T1 CONDOR queues optimized for “shorter” AOD jobs
  also worker nodes are shared with (higher priority) production jobs
  so ESD jobs go into hibernation and rarely revive (personal experience)
  and when they do revive staged ESD has been deleted and job crashes

a possible solution … ?

3rd attempt is made using the newly created BAF (BNL Analysis Facility)
[recall that the BAF is designed primarily for parallel DPD analysis]

Jobs finish before ESDs are removed from the staging area but completion rate is still unacceptably slow (analyzing only a few ESD files/day)
- repeat passes (which are of course necessary) are not possible

Postdoc #1 estimates that skimmed ESDs of the needed events would only occupy a few TB

- she requests and receives 10 TB of dedicated space at BAF (through RAC ?)

- she then uses most of this space as an additional staging area and sets up an ESD skimming utility to put on disk ESD files of only the $Z \rightarrow ee$ events

- this takes several months to complete but once finished studies can be run and re-run with minimal overhead (still considerable congestion at BAF)
Other analyzers also express interest in using this skimmed ESD data set

what next … ?

Assuming this study indicates a global need to modify the e/γ energy, what should we do ?

- should we reprocess or apply an AOD/DPD/Ntuple level correction ?
- how to decide ?

PD#1 message:
Do we need to provide dedicated resources for limited ESD/RAW analyses (especially during the early running period) ?

Postdoc #2

D1PD & D2PD production for top group

Prepares transforms and handles validation for production of D^1PDs

Uses T2 system to produce D^2PDs as specified by the top sub-groups (one of team)

As D^1PD production happens automatically, once defined this is not much effort
- validation and user support (variable definitions, etc.) require some work

However, top group has had a proliferation of D^2PDs - by stream, by subgroup, by approach, … is becoming difficult to manage (espec with frequent software updates)

Regeneration of D^1PDs leads to a cascade of D^2PD production which taxes both the available queues and available top group T2 project space (does this exist ?)

One month before a conference, problem is found with reco - no time to reprocess

how to proceed ?

Due to management bureaucracy, it soon becomes clear that producing “fixed” D^1PDs (the easiest solution) will not happen in time

PD#2 thus undertakes to instead make “fixed” D^2PDs (for both data and all MC)

Producing D^2PDs from AODs would be best in terms of implementing the fix but AODs are missing some derived data found in D^1PD - code change needed

Producing D^2PDs from D^1PDs is less desirable in that some of the fixed quantities are not in the D^1PD so aspects of the “fixed” D^2PD would be inconsistent

It is decided to go with the 1st option

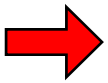
PD#2 message:
Those physics groups with their own D^2PD production should try to achive a balance between minimum D^2PD size and flavors of D^2PD

Grad. Student #1

Search for SUSY in tri-lepton final state

Needs to use soft-e id not available in ElectroWeak D^1PDs

(I've assumed SUSY group elected for the initial running **not** to produce its own D^1PD but will instead use the ElectroWeak (and Jet ?) D^1PDs)



GS#1 tries to run his own D^2PD maker on the T2 analysis queues

Initial jobs pend for a long time and once started run very slowly (many many users [most unfamiliar with the limitations] - system is overloaded)

GS#1 receives notice that his jobs have finished but cannot find the output D^2PDs

With help from panda expert it is determined that shortly after the jobs finished the disk containing the output D2PDs died (I don't know if the panda monitor already has a way of checking on this)

the saga continues

Next initial test goes much better and D^2PDs are copied to T3
- a few more passes and he perfects his D^2PD maker

Now ready to produce the full set of D^2PDs (running over both the eγ & μ streams)
GS#1 launches 4000 jobs to the T2 analysis queues

- due to high usage & GS#1 now having a rather low priority (due to his previous attempts), it takes 4 days for his jobs to start …

Apparently the T2 production and analysis queues share the same disks

– 8 days into his D^2PD production, many production jobs fail and the disks fill up, causing all the analysis queues to fail
– about 75% of GS#1's jobs had finished but he had no easy mechanism to determine which ones failed and resubmit those which failed
(& he was too lazy to write a script to do this!)

So, under conference pressure he re-submits all 4000 jobs directly to the ANALY_BNL_ATLAS_* queues (since all AODs are also at T1)

The ANALY_BNL_ATLAS_* queues are even more overloaded and job completion is extremely slow

some progress

Speaking with colleagues GS#1 learns of the production disk incident (which was in hypernews if he'd looked in the right place …)

- he resubmits to the general T2 ANALY queue & eventually all his jobs finish

He misses the conference deadline but it turns out he wasn't ready anyway
- sub-group convener failed to get MC request in on time

During Univ. X group meeting it becomes apparent that GS#1 doesn't know what several of his D^2P2 variables actually are(!) [he copied them from official D^1PD maker]

Further investigation reveals corresponding AOD variables also not understood (this is a common complaint in 2008)

Contact people are soon identified & the variables in question are understood

Analysis reveals soft-e id needs to be optimized for this analysis

GS#1 message:
We should expect users to submit thousands of jobs at a time - should they be provided with job management tools (cleanup, resubmission, etc) ala production ?

Grad. Student #2

Measurement of W-boson mass

ElectroWeak D^1PD is a good starting point for $W \rightarrow e\nu$ and $Z \rightarrow ee$ (for calib)

However, EW D^1PD is rather big (~5TB) & contains much unnecesary info

GS#2 decides also to make D^2PD maker (but using ElectroWeak D^1PDs as input)

Preparations here also find confusion in meaning of some D^1PD (& AOD) variables (as with GS#1, confusion resolved only by identifying appropriate experts)

After several false starts $D^1PD \rightarrow D^2PD$ jobs run well (& quickly) on T2 analysis queues (jobs finish in only 3 days - thanks to smaller size and more localized EW D^1PD)

D^2PD is < 1 TB in size and easily transferred to T3

Likewise, requisite MC samples are run through D^2PD maker & copied to T3

on to calibration, fitting, & systematics

All calibration and fitting is performed on the T3

Fitting procedure proves to be very cpu intensive - must cooperate with other T3 users

While working on the systematics it is announced that there is a just discovered temperature dependence to the measured EM calorimeter energy

Too large to treat as single systematic - must be corrected on an event by event basis

Unfortunately, temperature (conditions info) is not available in AOD (or D^1PD)

GS#2 thus makes a new D^2PD maker which has to query the conditions db for each run and corrects the EM energy

She has lots of problems but eventually makes it work although it runs much more slowly than before

Initial comparisions between $W \rightarrow e\nu$ and $Z \rightarrow ee$ events points to a problem in the E_T^{miss} algorithm and also problem in modeling of detector response

GS#2 message: expect CPU intensive local tasks to grow - sufficient resources ?

Grad. Student #3

Measurement of t$\bar{t}$ → e+jets cross section

Can begin with the e+jets D^2PD produced by the top group from the eγ stream

Size is ~5 TB so is easily copied to T3

Fortunately this data sample can also be used to estimate multijet backgrounds

However, obtaining luminosity and conditions (bad runs) info at T3 is problematic - probably pilot error but pilot needs more explicit instructions

As the initial results didn't make sense, it was soon realized that GS#3 had used the wrong MC datasets for his efficiency studies (didn't understand file names)

Although there is a naming convention note, GS#3 was unaware of it - should make this better known (it doesn't appear to be in the workbook, but it should be)

GS#3 message: we need to define a basic analysis T3 with the functionality to perform a typical analysis (maybe even provide some part-time setup support ?)

Prof. 

participates in t-channel single-top search

Working with former student to understand the effect of various systematics on the multi-variate selection tool

- very cpu intensive

Local resources already fully saturated (primarily with W mass analysis)

Tries to use T2 but has naming convention problems with both his input and output files - and can't figure out how to run his jobs under pathena

Rather than solve this problem, he looks for other resources, eventually gaining access to underused farm in Padua

Prof. message:
As we saw with GS#2, it's not clear if all groups will have sufficient resources for CPU intensive tasks - if we want to encourage the use of T2 for more than DPD making, should we prepare to examples to follow ? (similar problem with one of our students)

Outlook

Analysis sometimes appears to be “Whack-a-mole” where the analyzer is the mole

During the first few years of running we should expect a significant amount of reprocessing at each stage (although reco from RAW/ESD will be difficult)

Physics & University groups should prepare resource needs estimate for planned tasks (and indicate potential failure concerns)

If the system gets in the way, users **will** go around it - usually to detriment of all

Thoughtful communication between users and computing management is essential