# User Support and Workflows

Nils Krumnack (Iowa State University)

# Analysis Facilities Uses

- various uses for analysis facilities:
  - n-tuple analysis
  - machine learning
  - code development
  - grid submission
  - n-tuple production
  - …

- main uses:
  - n-tuple storage and compute for "interactive" n-tuple analysis
    - including ML training
  - access to full ATLAS environment for interactive work
  - easy/fast running of "heavy" compute loads in batch
  - access to GPUs
  - access to JupyterLab web interface

# N-Tuple Size

- n-tuple size is very important for analysis:
  - determines how much disk is needed
  - larger n-tuples (normally) take longer to process
- typically need space for at least two sets of n-tuples
  - the current set and the next/last set
  - can be shared by all users of analysis

- size for full Run 2 dataset (based on ttbar cross section analysis):
  - 10s of TBs with Run 2 frameworks
  - <1TB with CP algorithms
  - CP algorithm n-tuple still has room for improvements
- very dependent on framework and analysis
  - these numbers are probably at the upper end of use cases

# Interactive Analysis

- generally want "interactive" analysis
  - user submits request, gets result after a "short" wait

- assumptions vary wildly about what is a "short" wait
  - minutes to hours, definitely by the next day
  - hours still doable with regular batch jobs
  - shorter turnaround times may be harder to support

- hard to know what's realistic in the future
  - depends on how well we do with analysis/n-tuple optimization
  - may in part depend on choices at the site itself
  - getting shorter turnaround can have big payoffs in productivity

# Shell Login vs JupyterLab

- not clear whether people prefer JupyterLab or ssh logins
  - probably will always need both
  - likely depends a lot on personal preferences
  - even for python analysis not everybody uses/prefers Jupyter
  - some things will always need users to log into a shell

- trying to advertise Jupyter more to people
  - seems easier to teach and support
  - (hopefully) also simpler to use
  - native environment for machine learning tools
- JupyterLab widely used in industry
  - can make use of tools from outside of HEP
  - generally a robust development/analysis environment
  - potential advantage for graduates leaving the field
  - caveat: most industry users have smaller datasets

# PHYSLITE

- recommendation from AMSG: use PHYSLITE to replace n-tuples
  - i.e. produce histograms directly from PHYSLITE
  - run CP tools on-the-fly during histogramming step
    - not optional: MET/OR/SFs/systematics not in PHYSLITE
  - have full set of PHYSLITE datasets at analysis facilities
  - main goal: save disk space (at analysis facilities)

- my guess: at least 1-3 years in the future
  - CP tools across the board too slow for on-the-fly use
  - CP tools do more work than ideal for use on PHYSLITE
  - PHYSLITE is currently 22kB/evt, I'd prefer 2-4kB/evt
  - CP tools not compatible with columnar analysis

- may need to do further skimming/thinning/etc. on PHYSLITE
  - i.e. smaller/more focused files specific to analysis
  - would have to evaluate usefulness based on performance

# On-The-Fly Corrections

- goal: run CP tools on PHYSLITE/n-tuple when filling histograms
  - avoid storing SFs/systematics in n-tuple/PHYSLITE
    - reduce n-tuple size by factor 3-10 (very rough estimate)
  - calculate MET/OR, not directly present in PHYSLITE

- assume this is faster than reading data from input file
  - most calculations are technically very simple
  - analysis (normally) limited by i/o speed
  - even if it's slower: trade disk for CPU

- this is a fair amount of work to implement:
  - CP tool infrastructure/implementations need to be faster
  - need to change how some corrections are done
  - need more flexibility in how we call CP tools
- still in the early stages of this effort

# PHYS

- don't expect any on-the-fly processing for PHYS
  - larger and slower to process than PHYSLITE
  - needs more corrections to be applied
  - PHYS corrections can't be optimized in the same way
- expect will only use PHYS for n-tuple/PHYSLITE production

- normally n-tuples from PHYSLITE more practical
  - smaller and easier to process
  - can skip corrections already in PHYSLITE (not yet possible)
  - PHYSLITE available at more sites

- some niche advantages of PHYS:
  - PHYSLITE needs to be remade after recommendation change
  - can run corrections with non-standard settings
  - some extra data available in PHYS

# Training Events

- main event: Induction Day + Analysis Software Tutorial
  - one week introduction (mostly) for new students
  - online during pandemic, back to in-person this year
  - still discussing how best to combine online and in-person
- new format this fall: analysis walkthrough
  - follow leptoquark analysis from beginning to end
    - analysis basics, mc generation, CP corrections, n-tuple analysis, analysis optimization, statistical analysis…
  - more complete/coherent view of analysis workflow/tools
  - based on (limited) experience so far: engaging format for students

- analysis facilities onboarding events by Amber and Cecilia
- other more specialized events:
  - e.g. offline software developer tutorial every 1-2 years
  - plans to track these other tutorials in a single place
- also events/courses from HEP-SF and Software Carpentry

# Tutorial in a Box

- analysis tutorial is meant to be portable
  - meant to be doable at any analysis facility
  - need both shell login and JupyterLab
  - would like a fast turnaround batch queue
- testing of portability in progress
  - managed to run at CERN, SLAC, U-Chicago so far
  - plans to add support for running at BNL as well
- requires some local customizations/support:
  - pre-place all the data files needed on local disk/storage
  - provide instructions for login, batch access, etc.
  - (fast) technical support for problems that arise

- can be used to organize/host local training event
  - new tutorial originally developed for US tutorial at SLAC
  - still need local tutors+organizers, but material essentially ready
- allow tutorial participants to work on their local analysis facility

# Documentation & Support

- several sources of documentation
  - twiki pages and atlassoftwaredocs
  - tutorial pages written to double as documentation
  - doxygen pages generated for athena

- usual avenues for software support: mailing lists and mattermost

- support for analysis facilities:
  - discourse server at BNL: https://atlas-talk.sdcc.bnl.gov/
  - static documentation on the analysis facilities: https://usatlas.readthedocs.io/projects/af-docs/en/latest/

- helpful to have support people for each analysis facility: Shuwei (BNL), Cecilia (UChicago), Wei (SLAC)