

# Data sets for trigger & TPG studies

Klaudia Wawrowska  
klaudia.nicola.wawrowska@cern.ch

10/12/2024

# Introduction

## Recent Developments:

- We have a new Trigger Simulation module in LArSoft which facilitates development and physics performance tests for trigger algorithms.

## Current Challenges:

- Reliance on self-generated data due to absence of raw digit waveforms/TPs in production datasets.
- High-statistics trigger tests with raw digits are bulky & limited by individual storage quotas.

## Could be an ideal time to think about a more consistent and streamlined approach to trigger studies :

- A good first step might be adopting official datasets in studies:
  - Ensures reproducible results & consistent simulation configurations in studies.
  - Eliminates data duplication among users working on adjacent things.
  - Having readily accessible data would accelerate developments as each incoming person can dive right into analysis.
  - Resolves some storage issues.

## Overview

- DUNE's current approach to data management
  - And why it's a bit of a problem for trigger studies.
- What official data sets are available.
  - Can we use any of it for our purposes, or not?
- What aspects of trigger algorithm development pipeline should we prioritise?

# Storage spaces at FNAL machines

Direct usage for a lot of these spaces is slowly phased out as DUNE moves towards the use of official data sets.

In the past, users had write access to *persistent dcache* for studies involving high statistics.

- Decent storage space, grid accessible, permanent retention of files.

Currently the only grid-accessible storage for users is the *scratch dCache*:

- Unlimited space but wipes older files when new files are added.

The current permanent storage space is *NAS data*:

- ~1 TB quota per user (?), not grid accessible.

	Quota/Space	Retention Policy	Tape Backed?	Retention Lifetime on disk	Use for	Path	Grid Accessible
Persistent dCache	No/~100 TB/exp	Managed by Experiment	No	Until manually deleted	immutable files w/ long lifetime	/pnfs/dune/persistent	Yes
Persistent PhysGrp	Yes/~500 TB/exp	Managed by PhysGrp	No	Until manually deleted	immutable files w/ long lifetime	/pnfs/dune/persistent/physicsgroups	Yes
Scratch dCache	No/no limit	LRU eviction - least recently used file deleted	No	Varies, ~30 days (NOT guaranteed)	immutable files w/ short lifetime	/pnfs/<exp>/scratch	Yes
Tape backed	dCache No/O(10) PB	LRU eviction (from disk)	Yes	Approx 30 days	Long-term archive	/pnfs/dune/...	Yes
NAS Data	Yes (~1 TB)/ 32+30 TB total	Managed by Experiment	No	Until manually deleted	Storing final analysis samples	/exp/dune/data	No
NAS App	Yes (~100 GB)/ ~15 TB total	Managed by Experiment	No	Until manually deleted	Storing and compiling software	/exp/dune/app	No
Home Area (NFS mount)	Yes (~10 GB)	Centrally Managed by CCD	No	Until manually deleted	Storing global environment scripts (All FNAL Exp)	/nashome/<letter>/<uid>	No
Rucio	10 PB	Centrally Managed by DUNE	Yes	Each file has retention policy	Official DUNE Data samples	use rucio/justin to access	Yes

# Trigger study requirements

- Trigger algorithm design is driven by signal and background properties.
- Need a lot of radiological data to get realistic estimate of false trigger rates for a given trigger configuration.
- Getting decent statistics for raw digit files containing radiological signals will take a lot of storage.
  - E.g.: Raw digit detsim file consisting of 100 events containing lateral APA radiological signals has a size of 13 GB.
- To reuse data for different studies this storage should be semi-permanent & grid-accessible.

# Persistent, grid-accessible storage


Direct write access to the persistent repository is now granted to **working groups instead of individual users**:

***“Persistent dCache: the data in the file is actively available for reads at any time and will not be removed until manually deleted by user. There is now a second persistent dCache volume that is dedicated for DUNE Physics groups and managed by the respective physics conveners of those physics group. In general if you need to store more than 5TB in persistent dCache you should be working with the Physics Groups areas.”***

**Ticket system for production requests:** [https://wiki.dunescience.org/wiki/Production\\_and\\_Processing](https://wiki.dunescience.org/wiki/Production_and_Processing)

- The main takeaway is that there’s a fair bit of planning and authorisation needed to get this going.

In addition, we summarize the main steps in order to submit a request.

- Working group leaders must formally request the initiation of a production campaign through the DUNE Computing Service Request/Production.
- Short description of the physics motivations
- Estimated deadline for the delivery of the output datasets
- Code version to be used (it has to be a tagged version and from a DUNE official repo )
- Description of the workflow: list of processing steps (generation, g4, detector simulation, reconstruction, analysis), and fcl file to be used for each step
- Validation sample, a test sample to be used as validation before full production, [docdb 29278](#)  “Production Policy”
- For each submitted job, the list of output files to be copied on tape
- Estimation of needed resources : CPU, memory and storage. If possible, it would be helpful to specify the statistic on which these values are based
- List of samples/runs to be processed, ordered by priority.
- Number of events/sample(run)
- Valid metadata
- If you have any questions, please do not hesitate to contact the Production team

# Existing data sets

- **Full list for the most up-to-date data:**

- [https://wiki.dunescience.org/wiki/Data\\_Collections\\_Manager/data\\_sets](https://wiki.dunescience.org/wiki/Data_Collections_Manager/data_sets)

- **What's there:**

- Beam data for HD & VD (fully reconstructed)
- SN data for HD & VD (hit reconstructed)
  - Separate CC and ES signal
  - Data with and without backgrounds available
    - Different radiological background models: decay0, lateral APA, central APA
- Cosmics and pure Ar simulation for protoDUNE

# Action plan?

## Do we request official data and if so, what data do we need?

- Can we “branch off” at the detsim stage using existing files?
- Are additional physics events beyond the official list needed (e.g. purely radiological data, single particle)?

## What’s the “stage” at which we want the data? Detsim (i.e. raw digit), TPs..?

- Brings up the issue of the entire trigger simulation workflow & what we need to focus on.
  - We’re at the awkward stage where things are simultaneously developed and tested.
    - There’s no official trigger simulation pipeline & it’s hard to say what exactly do we need.
    - Requesting raw digit data offers best flexibility, but is problematic from storage point of view.
- Storing raw digits is a necessity for TPG design finalisation and development.
  - We need better understanding of thresholds, noise filtering, & signal transformations across different algorithms.
- For trigger studies (TAs/TCs), we can only store TPs and discard raw digits.
  - This would require having some validated TPG chains (algorithm, threshold etc.).

**All of this depends on which milestones for the trigger are prioritised.**



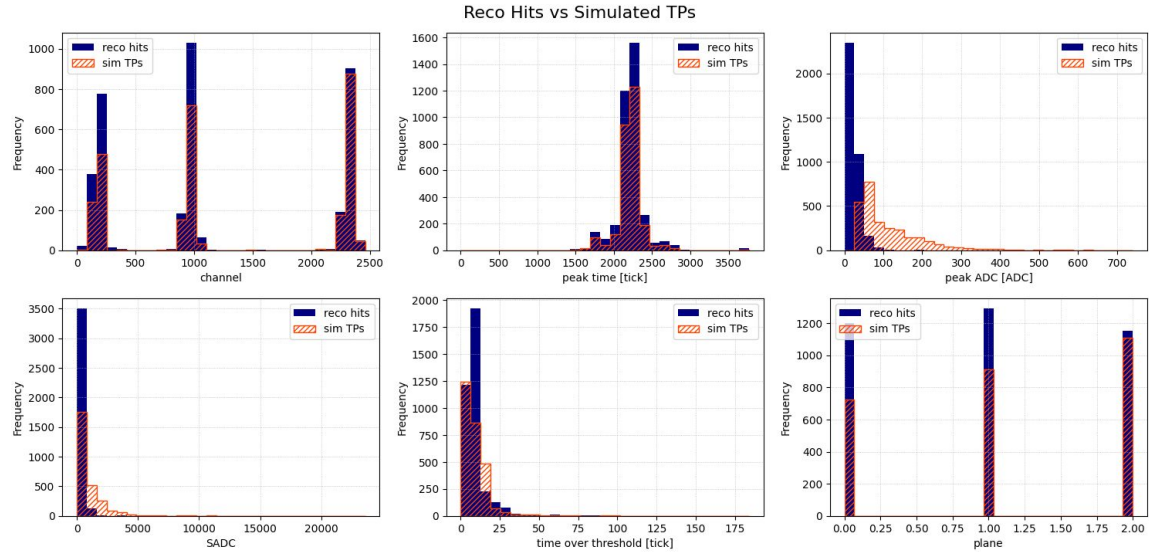
# Back-up

# Hits vs TPs

Reconstructed hits go through various signal processing steps we can't do online.

- Signal energy estimation (e.g., peak ADC) differs from offline algorithms.
- Collection signals are reliably located, but many induction TPs may be missed.

Without studying how signal transforms across TPG algorithms, hits can't be reliably used to develop more sophisticated trigger algorithms.



Simulated *SimpleThreshold* TPs vs reconstructed offline hits for muon tracks .