Initial Processing Stages for DUNE FD TPC Data

Brett Viren

July 26, 2019

Contents

Abstract

This note briefly describes a proposed, initial data processing of the largest category of data produced by the DUNE far detector, that of the TPC waveforms. It provides an estimate of the input and output data volumes, describes the processing in terms of data tiers, stages, chains, branches and joins. An singular initial stage is proposed followed by a branch allowing each of two existing models for the ionization electron distributions to be executed as parallel stages. Subsequent stages are left to other notes for consideration.

1 Input Data from DAQ

The DUNE FD TDR DAQ chapter[1] gives details on rate of data the DAQ will produce for archive storage (tape) and which will be input to the offline processing. The annual data volume is nominally limited to be less than 30 PB. The initial running with one single-phase (SP) detector module is expected to produce at about half this rate if the data read out from all channels of the detector for each trigger is output in uncompressed form. Lossless compression may be applied and it will result in a factor of at least 2 and likely 4 data volume reduction. During commissioning and initial operation advanced trigger and readout schemes will be developed and validated against this early data. These are expected to reduce the amount of data sent to offline by a factor of at least 10 as they will not emit data that is consistent with being from either electronics noise or low energy radiological decay. Along with the optimistic lossless compression factor these advanced schemes will allow capturing the same physics information with while producing only about 400 TB/year from one SP detector module.

Subsequently, the other three detector modules will be deployed. The final makeup of the FD is not yet finalized and may be comprised of another SP module, possibly with a fourth wire plane, a dual-phase (DP) module and/or a module with pixel readout. Until these details are understood, their contribution to the data rate input to the offline processing is uncertain. A rough estimate may be taken by simply multiplying the 3-plane SP estimate by four.

The production of simulated data, itself, is not considered deeply here. Its result is also input to the same offline processing chains described below. At least as much simulated data as detector data will be required and so represents at least a factor of two further inflation.

[1] `https://dune.bnl.gov/docs/dune-tdr/vol-sp/vol-sp-ch-sp-daq.pdf`

2 Processing Data Tiers, Stages and Chains

Here, a *tier* refers to an object schema and file format to which some data adheres. A processing *stage* refers to an input data tier, a transformation procedure ("job") and an output data tier. Stages may be connected via an intermediate data tier serving as output of one transformation and the input of one or more other transformations. A series of connected stages make up a *chain*. Where multiple stages consume data from the same tier, a chain is *branched* and where data from multiple tiers are consumed by the same stage two chains are *joined*.

3 Signal processing stage

Here it is proposed that a singular first stage be applied to the majority of the raw data which can be said to come from a *raw TPC data tier*. The remaining detector data can be said to come from the *raw PDS data tier* and is not a subject of this proposal although it is of course expected to be processed by other stages. The full transformation performed by this stage will consist of the following *steps*:

1. **Decoding** raw TPC ADC waveform data consists of reformatting data from the packing required by the DAQ into a form in memory which is suitable for efficient consumption by the next step in this stage. This step provides data in a "just in time" manner as prompted by the next step so that the amount of memory required may be minimized.

2. **Noise filtering** removes unwanted features in the ADC waveforms that are due sources of noise beyond those the inherent, thermal noise of the electronics. This includes coherent/common noise across multiple channels and "harmonic" noise that has most of its power near discrete frequencies. It may also include other mitigation such as unwanted transient features.

3. **Signal processing** applies two major and related substeps. First, a deconvolution of the ADC waveforms is performed across both time and channel within a plane. The deconvoluion kernel is formed from the (frequency-space) product of the field response (induced current due to drifting charge in the electrostatic field) and the electronics response (anti-alias and other shaping circuits). Second, signal region-of-interest (signal ROI) selection is performed to localize waveform segments consistent with drifting ionization and perform baseline corrections.

4. **Data reduction** that is enabled by the previous steps is applied and is described below.

The data output by this stage is (here) said to be from the *signal data tier* as it retains all available information which is inconsistent with being due to just inherent (thermal) electronics noise and identifiable excess noise. As such, these waveforms have unipolar shape, are sparse in time and channel and are measured in units of ionization electrons.

The filtering preformed in channel, channel periodicity, time and frequency allow for a substantial data reduction to be applied just prior to output of the results. Although technically lossy, from the point of view of desired signal information it is effectively lossless. The reduction applies in the following way.

Only waveform samples inside a signal ROI are meaningful and the rest may be discarded. A sparse representation of the result is expected to be reduced by at least a factor 100. This is based on an estimate[2] that scales MicroBooNE data to DUNE as well as the more direct experience with ProtoDUNE-SP. Being on the surface, these detectors have far more activity from cosmic muons than will be observed in the DUNE FD underground environment and so this factor represents an underestimate.

Furthermore, the low-pass filter which is applied during the signal processing leaves the signal-ROI waveforms oversampled by a factor of four. A rebinning in time can thus provide an information-lossless reduction by up to this same factor.

[2] `http://docs.dunescience.org/cgi-bin//ShowDocument?docid=2089`

An estimate for the size of this *signal data tier* for the SP module is based on the dominant cosmic muon sample. It has been estimated that when a cosmic muon traverses the SP module on average about 10% the APAs see activity. If sparseness factor of 100 based on the MicroBooNE scaling and the ProtoDUNE-SP experience, a total factor of 4 for lossless compression and rebinning then the data for cosmic muons is reduced by a factor of at least 4000. This is about 800 MB for one "event" or 2.5 TB/year.

An advanced trigger and readout scheme was described above. After it is enacted the 10% factor will be applied in the DAQ. Prior to that, this 10% factor does not take into account the fact that the 90% of the data that has no cosmic muon activity does have 39Ar activity. Each 39Ar decay will lead to a signal-ROI "patch" approximately 3 channels by 20 us seen by each wire plane facing a drift volume. Subject to the same factor of 4 reduction from the combination of lossless compression and rebinning (again, technically lossless), the 39Ar decays will contribute about 10MB which is negligible.

The initial signal processing stage thus represents a crucial reduction in data volume. In principle it may be employed once on any trigger and its output saved and reused. Only when improvements in the steps of this stage are developed need it be rerun.

4 Ionization imaging stages

After this first stage of processing, a branch is expected to occur as there are two major and divergent approaches to reconstruction. They diverge in how they model or "image" the the sparse ionization signal waveforms from the *signal data tier* just described. The two modeling approaches are defined here as:

1D the signal-ROI waveforms associated with a given channel are modeled as (fit to) a collection of Gaussian distributions over sample time ("hits").

3D portions of signal-ROI waveforms in a common time slice and over all channels of all planes (in each APA) along with knowledge of the geometry of the wires are used to model the likely location of ionization charge in both transverse directions using computed tomographic techniques ("blobs"). This intermediate result is checked for consistency across neighboring slices of time producing a collections of 3D locations ("clusters") in the drift volume that localize ionization electrons.

The data representing the model parameters for both approaches is of a further reduced volume compared to their common input of sparse waveform segments. The 3D approach further reduces the data volume to some extent as any excess noise that survived the previous stage tends to be removed as it makes patterns that are found to be inconsistent with a 3D ionization electron source.

The output of these two stages are here called *hit data tier* and *blob data tier* respectively. They each start a branch on a parallel chain. The subsequent stages of these chains are not considered here.

5 Summary

The singular first stage and two subsequent stages which each represent a branch of the DUNE FD processing are described. The data input to the first stage is from the *raw TDC data tier* and its output populating the *signal data tier* and is dramatically reduced compared to the input. The output represents a branch point being consumed by two different approaches to modeling the signal. One uses 1D model and the other a 3D model. These stages output data from the *hit data tier* and *blob data tier*, respectively.