Reconstruction Assessment MicroBooNE

LAr TPC Reconstruction Assessment Workshop
October 19-20, 2015

Tracy Usher
Representing the MicroBooNE Collaboration

Overview

- Reconstruction vs Analysis
- Overview of MicroBooNE reconstruction paths
- Some use cases
- List of possibly interesting lessons and/or discussion items
- Summary

Reconstruction vs Analysis

Reconstruction vs Analysis

- The task of reconstruction is to provide the basic building blocks common to all physics analyses
  - Start with the “raw” data from the detector
  - Apply calibrations, convert from detector coordinates to space coordinates, etc.
  - Output high level objects for analysis
    - e.g Tracks, Showers, etc.
- The task of Analysis is to perform the final event reassembly, make final event event selections and produce the output for the specific analysis at hand
  - There may be be some massaging of the input objects at this level
    - e.g. final track fits based on constraints

MicroBooNE Reconstruction Output

- Hits - Giving time and deposited charge on wires
- Clusters - Groups of logically associated hits in 2D
- PFParticles - Association of 2D objects into 3D objects, also can give pattern recognition hierarchy (event topology) for associated objects
- OpFlash/OpHits - PMT Flash information
- Tracks - “fit” trajectories of particles
- SpacePoints - 3D position from associated 2D hits
- Showers - position/direction/energy of particle showers
- CosmicTag - cosmic tagging score for tracks
- Calorimetry/Particle ID
- etc.
- The above is all produced “automatically”

MicroBooNE Analysis Paths

- "AnalysisTree"
  - root tuple output which summarizes the information in recon output
  - Allows quick analysis over large data sets
- "LArLite"
  - A "lightweight" framework which operates on data objects
  - Output of LArSoft reconstruction converted to LArLite analogues
  - Can do more extensive event reassembly in the context of LArLite
    - MicroBooNE is developing and making great use of the ERTools package which is aimed at getting the final event topology using the output of the reconstruction
  - This framework can also be (and is) used to do code development
    - Advantage of being able to rapidly process large data sets

Overview of MicroBooNE Reconstruction

Liquid Argon TPC Reconstruction

- Liquid Argon TPC's are very interesting detectors, simultaneously a
  - a high resolution 3D Tracking device,
  - an electromagnetic calorimeter
- Reconstruction is a challenging problem
  - Readout is wire/drift time and 3D imaging is achieved with multiple stereo wires
    - Primary problem: a priori don't know how to match hits between the different stereo planes
      - Forming 3D space points can lead to many ambiguous (ghost) combinations
    - Two basic approaches have been developed:
      - Associate hits in each plane individually to form 2D objects (e.g. tracks), then match object features across planes to form 3D objects
      - Form 3D hits from "allowed" matches of 2D hits between planes and use 3D algorithms to resolve any remaining 3D hit ambiguities and to form the output 3D objects
    - It should be noted that the above are certainly not the limit of paths that might be followed to attack this problem
  - There is no IP constraint to help guide the reconstruction
    - At MicroBooNE energies you can't even assume tracks are in beam direction
- There have been several pattern recognition algorithms developed over the course of the past few years which have effectively explored most (all?) of the phase space for an approach that starts with hits reconstructed on wires.

MicroBooNE Reconstruction Paths

- "Fuzzy Cluster"
  - Strictly 2D approach relying on a hough transform to associate hits into common objects
  - Relies on Track fits and/or shower reconstruction to form 3D objects
- "ClusterCrawler"
  - Strictly 2D approach utilizing trajectory following techniques to identify track candidates in each plane
  - Companion "ccTrackMaker" will then do feature matching across planes to produce 3D objects including the PFParticle hierarchy, plus candidate tracks, vertices, etc. Until then relies on downstream algorithms to form 3D objects
- "Pandora"
  - Utilizes Pandora SDK and reconstruction philosophy to identify objects in each of the 2D planes and then do feature matching across planes to form the 3D objects. Runs in two passes - first optimized to reconstruct CR tracks, the second runs on those hits not associated to tagged CR's and is optimized for neutrino topologies.
  - Outputs a PFParticle hierarchy representing pattern recognition decay topology, plus candidate tracks, vertices, etc.
- "Cluster3D"
  - Forms all "allowed" combinations of 2D hits to make 3D hits. Forms 3D objects directly from the 3D hits (resolving any ambiguous combinations as part of the algorithm).
  - Outputs a PFParticle hierarchy

- "Wire-Cell"
  - 3D approach but employing a tomographic approach using the waveforms (no hit finding). This is a new approach which is progressing rapidly but is not yet fully integrated into LArSoft - not yet part of the MicroBooNE mainline reconstruction.

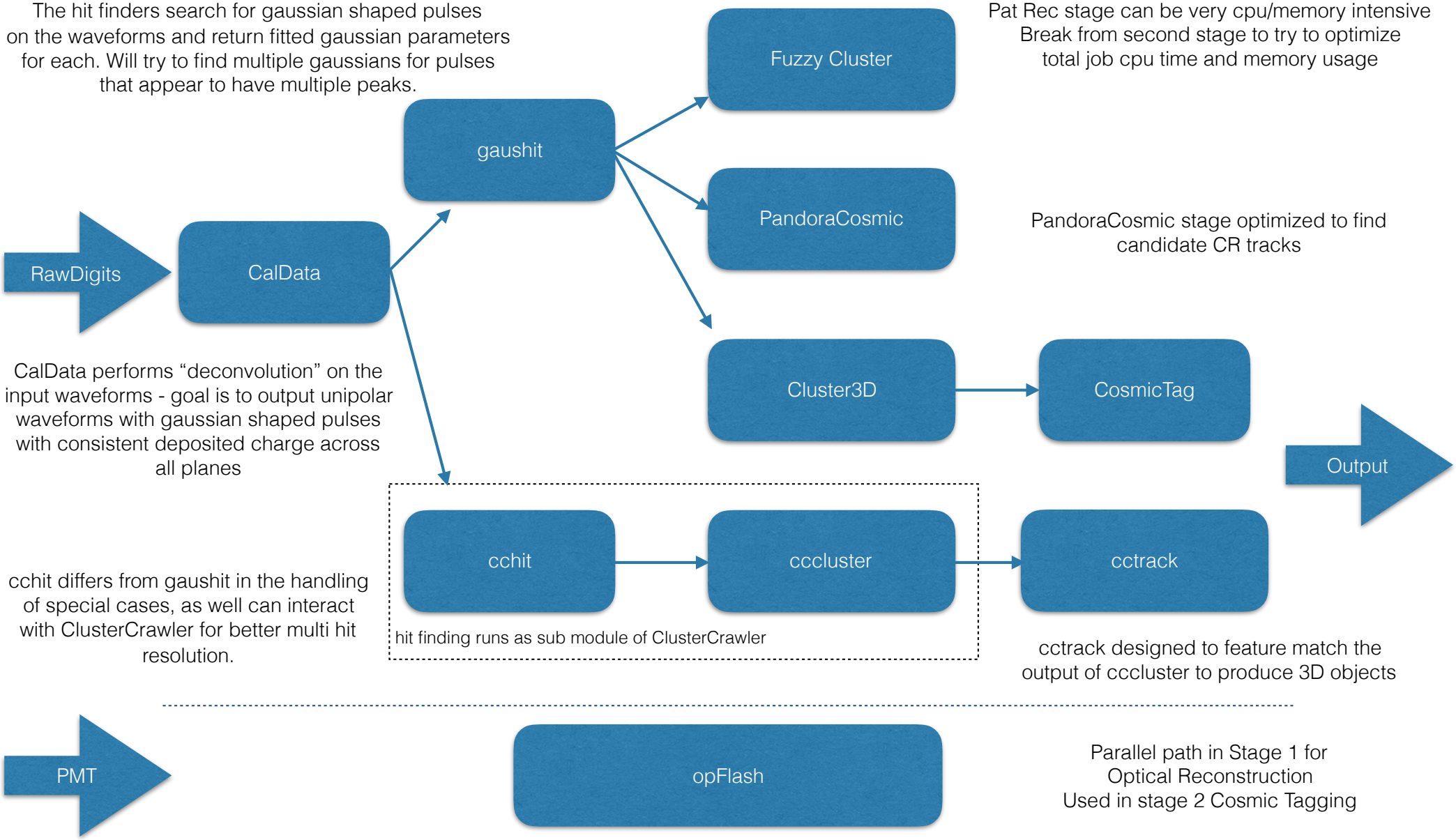
Monte Carlo Challenge 6

- The sixth MC data challenge for MicroBooNE: MCC6
  - Most realistic detector simulation to date
  - Significantly improved reconstruction efficiencies for downstream analysis
- Primary goals
  - Provide large scale MC datasets to the Analysis Groups so each could develop end-to-end analysis to provide a feedback loop with the reconstruction group
  - Evaluate the reconstruction paths and determine if we can eliminate approaches from the mainline
    - Ideally could reduce to a single reconstruction path
- Desirable secondary goals
  - First MC data set with PFParticle hierarchy - develop path for using this to seed analysis event topologies, e.g. in ERTools
  - Develop metrics which new/alternate pattern recognitions would need to meet before being considered to enter the “mainline”
- Launched ~5-6 months ago
  - MC Data sets started to become available around the time MicroBooNE commissioning began

MCC 6 Stage 1 Recon

The hit finders search for gaussian shaped pulses on the waveforms and return fitted gaussian parameters for each. Will try to find multiple gaussians for pulses that appear to have multiple peaks.

Pat Rec stage can be very cpu/memory intensive
Break from second stage to try to optimize total job cpu time and memory usage



PandoraCosmic stage optimized to find candidate CR tracks

CalData performs “deconvolution” on the input waveforms - goal is to output unipolar waveforms with gaussian shaped pulses with consistent deposited charge across all planes

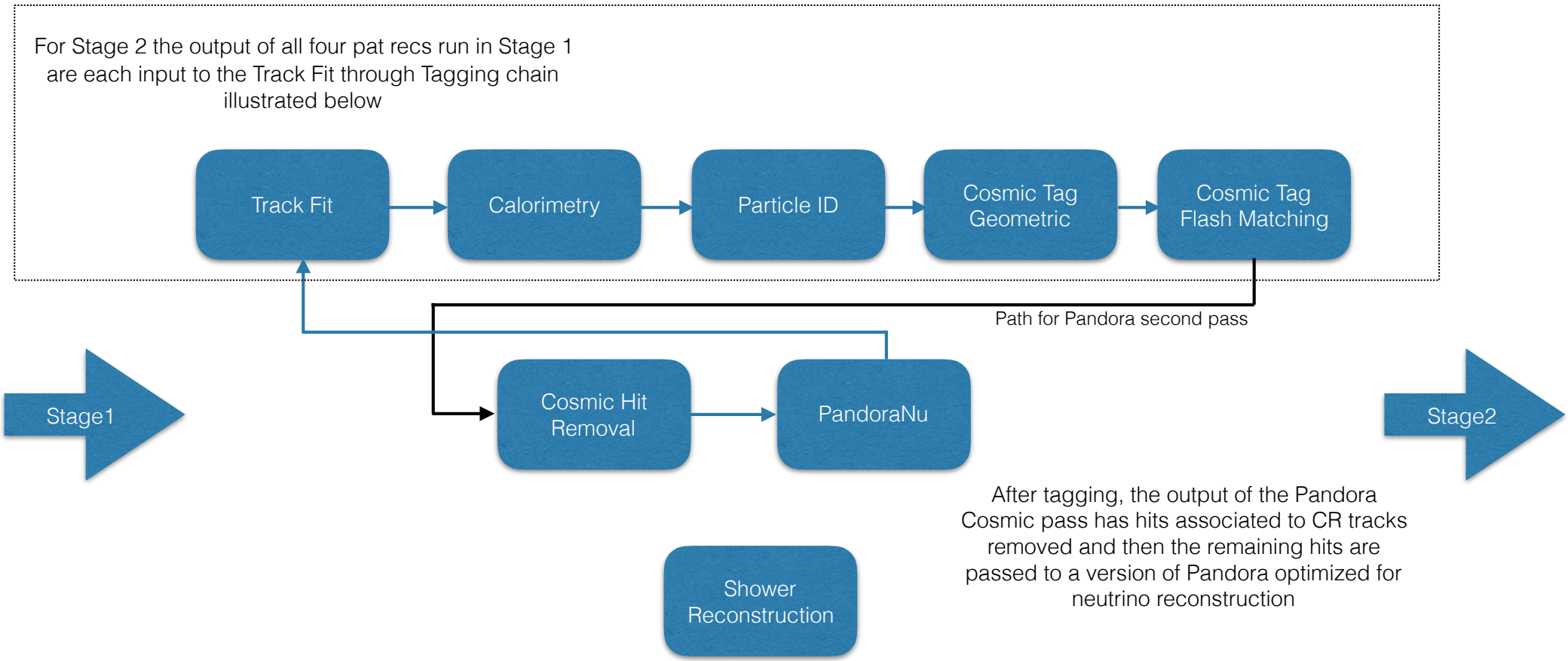
cchit differs from gaushit in the handling of special cases, as well can interact with ClusterCrawler for better multi hit resolution.

cctrack designed to feature match the output of cccluster to produce 3D objects

Parallel path in Stage 1 for Optical Reconstruction
Used in stage 2 Cosmic Tagging

MCC 6 Stage 2 Recon

For Stage 2 the output of all four pat recs run in Stage 1 are each input to the Track Fit through Tagging chain illustrated below



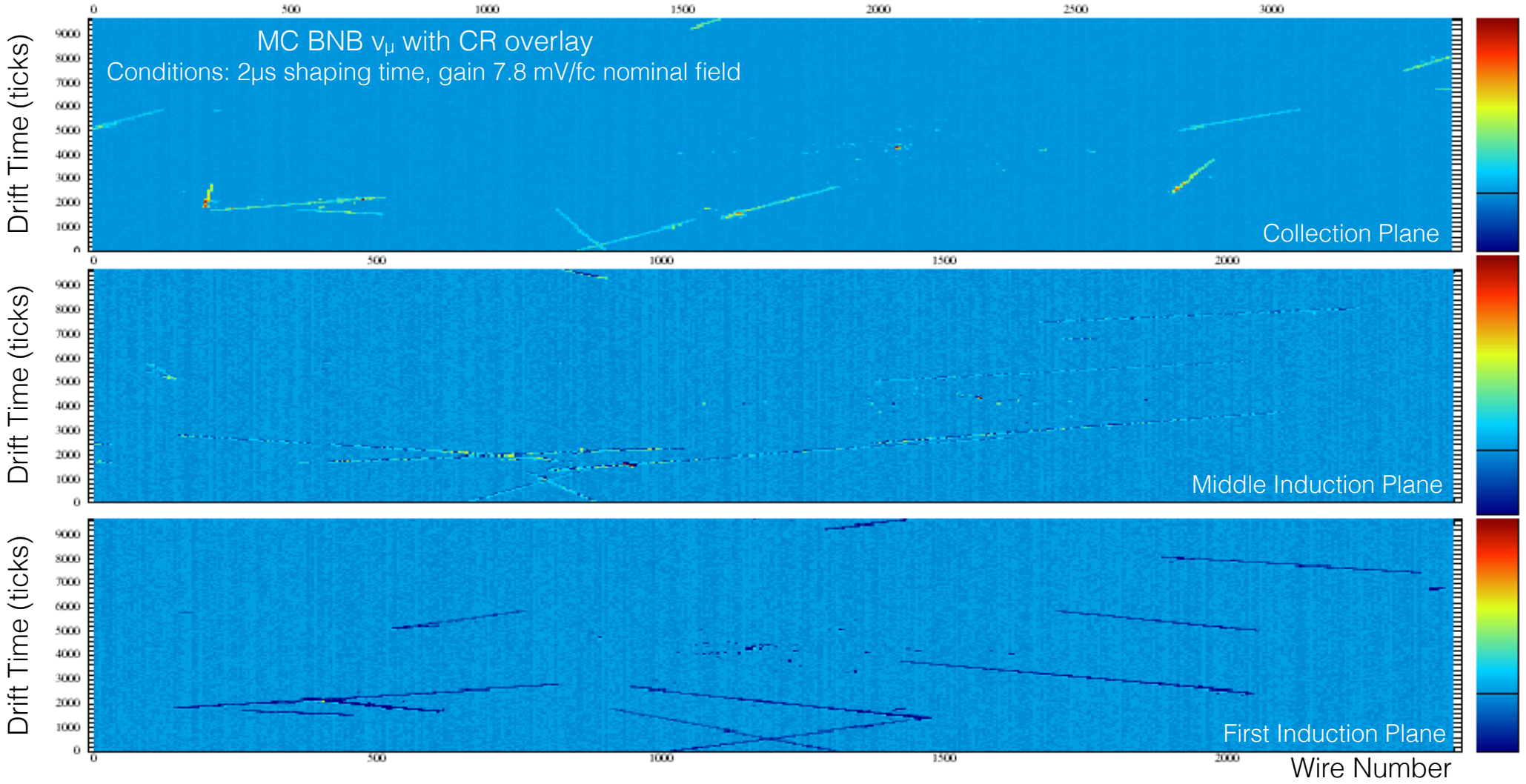
After tagging, the output of the Pandora Cosmic pass has hits associated to CR tracks removed and then the remaining hits are passed to a version of Pandora optimized for neutrino reconstruction

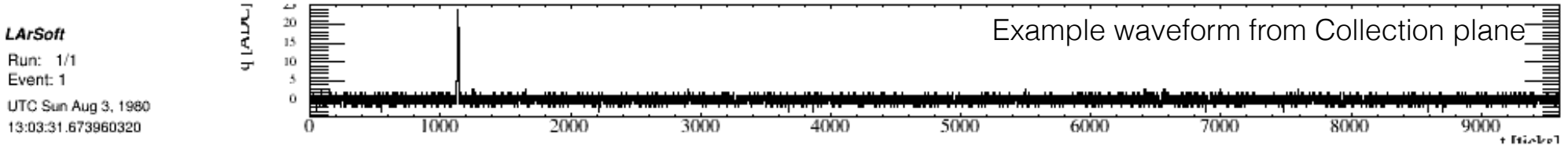
The final step in the Stage 2 reconstruction is to run the Shower Reconstruction separately on the output of the Fuzzy Cluster path and on the output of the neutrino optimized Pandora path

MicroBooNE Commissioning

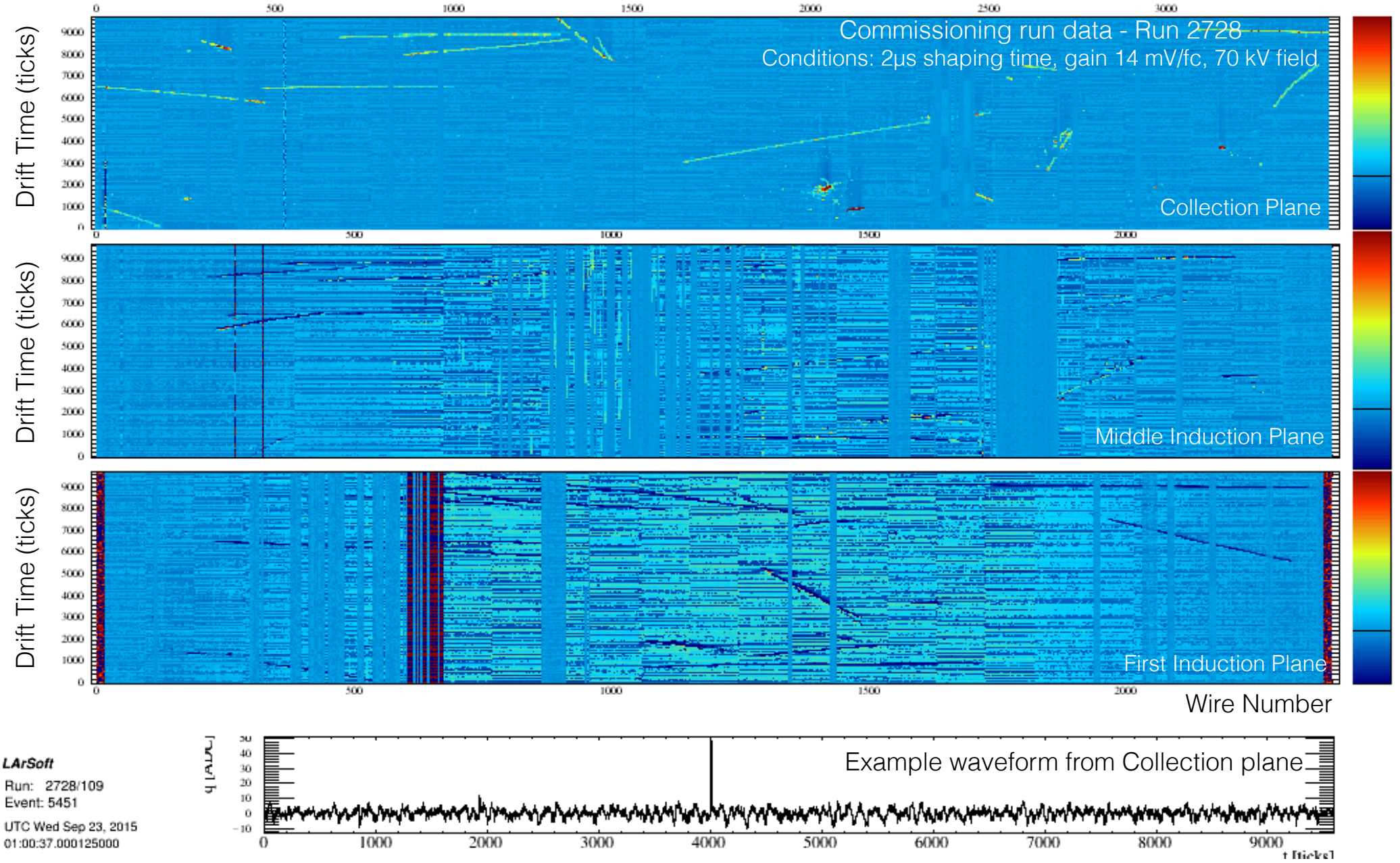
- Bringing complex detectors online for the first time is rarely a smooth process
  - In particular, there are almost always surprises
- Two issues directly impacting reconstruction
  - Dead channels
    - Tend to be in groups as opposed to the assumed isolated dead channels one might have studied in developing algorithms
  - Noisy channels with several different signatures
    - “zig-zag” - high frequency tick-to-tick oscillations in randomly distributed short bursts
    - “correlated” - low frequency (~20 kHz) correlated across wires
    - “chirping” - transient issue, switching between “dead” and “live” with large baseline excursions
    - “high noise” - steady state very high rms noise - effectively dead channels for recon
- Redirection of reconstruction resources to address these issues
  - Attacking noise issues by developing algorithms aimed at filtering out as much noise as possible
  - Developing more sophisticated channel status information mechanism
  - Pattern recognition algorithms will need to be able to handle gaps with no information

Simulated events - MC 6

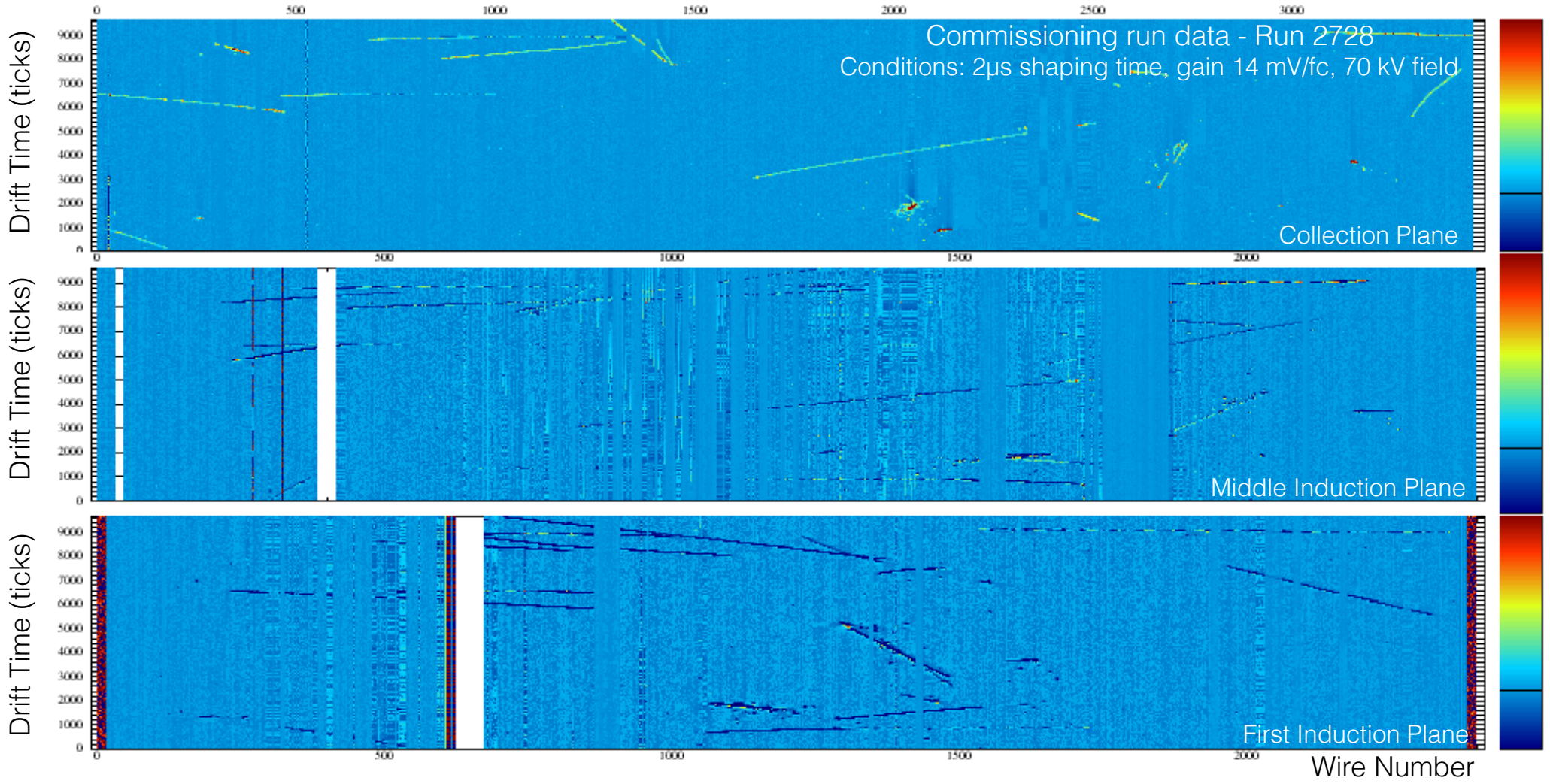


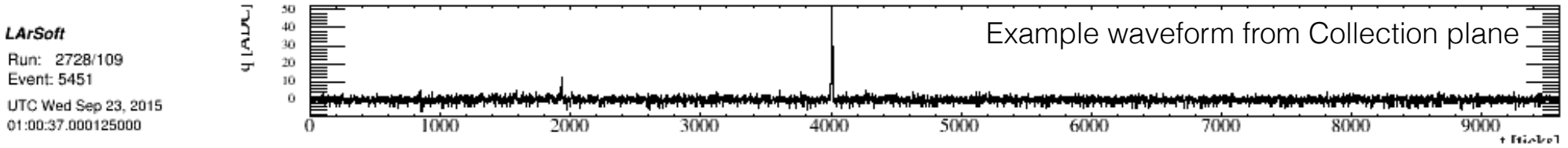


Noise Run Event - Unfiltered



Noise Run Event - Partial Filter





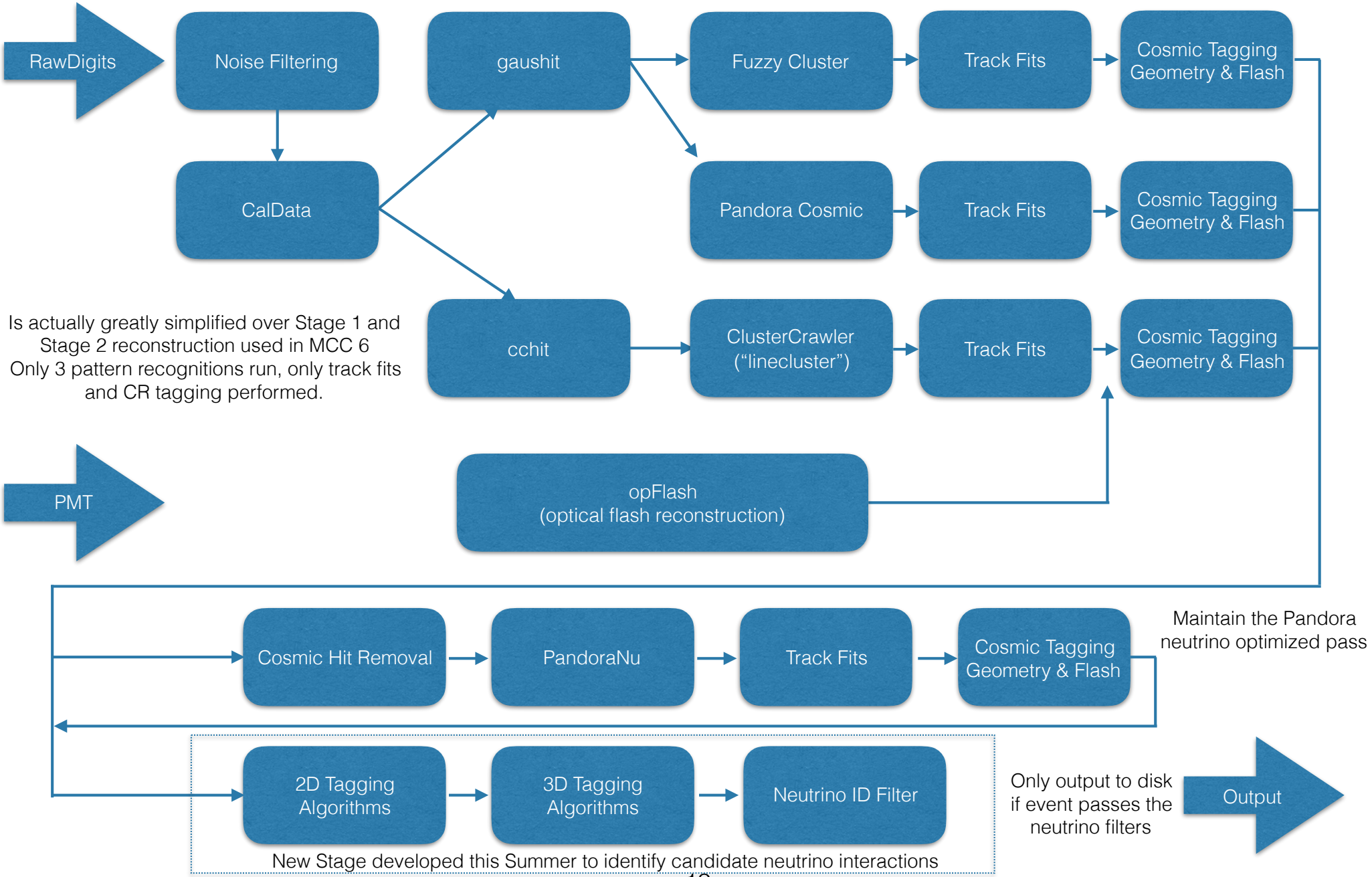
Impacts on Reconstruction

- Less compression of RawDigits than expected
  - Uncompressed RawDigits ~152 MB/event, root compression in MC was ~38 MB/event, root compression for unfiltered RawDigits is ~43 MB/event and for filter ~38 MB/event
- Significant increase in the number of reconstructed hits, in particular from the chirping wires
  - Leads to a significant increase in cpu time for the pattern recognition algorithms
- “Grouped” dead channels, transient dead channels impact reconstruction efficiency
- Need to address these issues before willing to run the MCC6 two stage reconstruction on beam data
  - Would take a lot of cpu time per job, generate a lot of output data and might be of limited use for analysis

Early Data Taking Strategy

- First Beam Data
  - Time in the PMT trigger system to match flashes to beam gate
  - No reconstruction run
- Initial Data Taking period
  - “Swizzle” events with an optical flash PE cut in either the BNB or NUMI window
    - Initially ~11% of all data events
    - See Mike Kirby presentation for details, docdb 4934
  - Run a limited reconstruction with subset of paths and only through track reconstruction
  - Have developed and are applying algorithms to identify candidate neutrino interactions and output event to disk only if a candidate found
    - When output, keep both the unfiltered and filtered RawDigits to allow more development of upstream filtering algorithms
- Primary Goals:
  - Gain more experience with detector issues and make necessary improvements in recon, in particular in developing better noise mitigation algorithms and dealing with gaps
  - Provide a data set that enables us to start understanding cosmic ray tagging, neutrino reconstruction, etc.
  - Couple all of this with improvements from experience gained with MCC6 to improve overall recon
- Note: all raw data from detector is written to tape so plan will have ability to reconstruct all data at the appropriate time

Neutrino ID Recon Flow



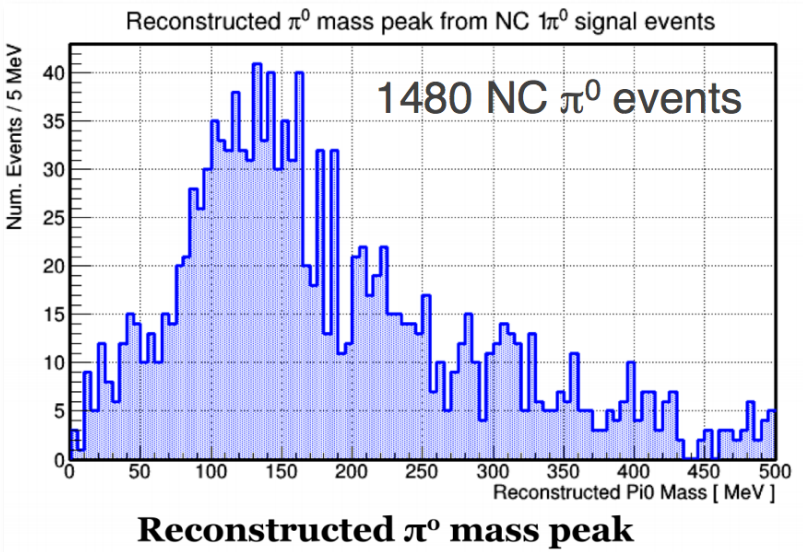
MicroBooNE Reconstruction Summary

- Message: MCC 6 put us in good position to do a systematic evaluation of the reconstruction performance
  - In spite of the demands of commissioning the detector work has progressed and the first end-to-end analyses are in place and have identified two primary areas for reconstruction improvement
    - Shower reconstruction
    - Final track fits
- The commissioning period has indicated important areas of further development
  - Noise mitigation
  - Channel status handling
  - Event filtering algorithms
  - cpu time and data volume
- Touch on a few key points in what follows

A Few Use Cases

Need for End-To-End Analyses

- MicroBooNE displayed some of its first end-to-end analyses using MCC 6 at the Fermilab PAC meeting this past June
  - See slide 35 of docdb 4570
- The π^0 analysis drew scrutiny
  - Efficiency for reconstruction is low
  - Mass resolution is poor
- But… this is actually a significant achievement:
  - One of the first examples of an end-to-end analysis resulting from "automatic" reconstruction
    - Find and reconstruct showers
    - Identify them as electron showers
    - Vertex the showers to find common start point
    - Reconstruct the mass
- Important: served primary purpose to help focus shower reconstruction group on needed areas of improvement…



Reconstructed π^0 mass peak

Shower Reconstruction Development

- Shower reconstruction is a very active area right now
  - Significant improvements have been made already in efficiency, axis pointing resolution, shower start point, dQ/dx, etc.
  - Driven by a dedicated group already very busy with commissioning tasks but still finding time to push this forward
  - Very promising progress, will soon see new π^0 reconstruction plots
- Infrastructure issue:
  - Their work is facilitated by working in the LArLite framework
    - They can quickly run through entire MCC 6 data sets to evaluate their performance
  - Unfortunately, code is not yet plug compatible with the LArSoft framework
    - Requires the developer to make small interface changes before porting to LArSoft - a hurdle
  - LArSoft is still running MCC 6 era shower reconstruction
    - What you saw on the previous page
  - There is strong need to find a mechanism to facilitate the plug compatibility of code developed in the LArLite framework with LArSoft
    - In particular it would be nice to get their tagged updates into the weekly releases

Improvements to Tracking

- The MicroBooNE cross sections group has recently distributed an internal note detailing the status of their first end-to-end analysis - the ν_μ CC inclusive cross section.
  - MicroBooNE internal document - docdb 4925
- They highlight two primary areas where for improvement:
  - Cosmic Ray tagging
    - To get good rejection their efficiency for neutrino survival is lower than desired
    - Big issue here are the broken/stopping Cosmic Rays which can mimic neutrinos
  - When trying to extend to extend to single differential cross sections they run into issues with track resolution
    - Angular resolution
      - Subset of this is track start point - sometimes track is “backwards”
    - Momentum resolution
    - Improved tracking efficiency for short tracks (aiming to get proton multiplicity)

Probable Cause of Angular Issues

- MicroBooNE uses Track3DKalmanHit
  - Does Kalman Filter fit of 2D hits associated to Track
  - Does internal pattern recognition
    - It is given a collection of 2D hits and it runs its own pattern recognition to find and return as many tracks as it can with those hits
      - For Fuzzy Cluster and ClusterCrawler it gets ALL the 2D hits in ALL clusters
      - For Pandora it operates on hits associated to individual PFParticles - more likely to return a single track in this case
- Current track fit uses an assumed (and fhicl controllable) track momentum currently set to 200 MeV/c
  - Low energy of fit is good for efficiency in pattern recognition stage
    - But multiple scattering errors are large and allow fit to “move around” resulting in poor angular resolution
- Track fit also finds its own “seed” for starting the track
  - It often will start the track at the end
  - (note that a recent addition is to allow use of a pat rec seed)

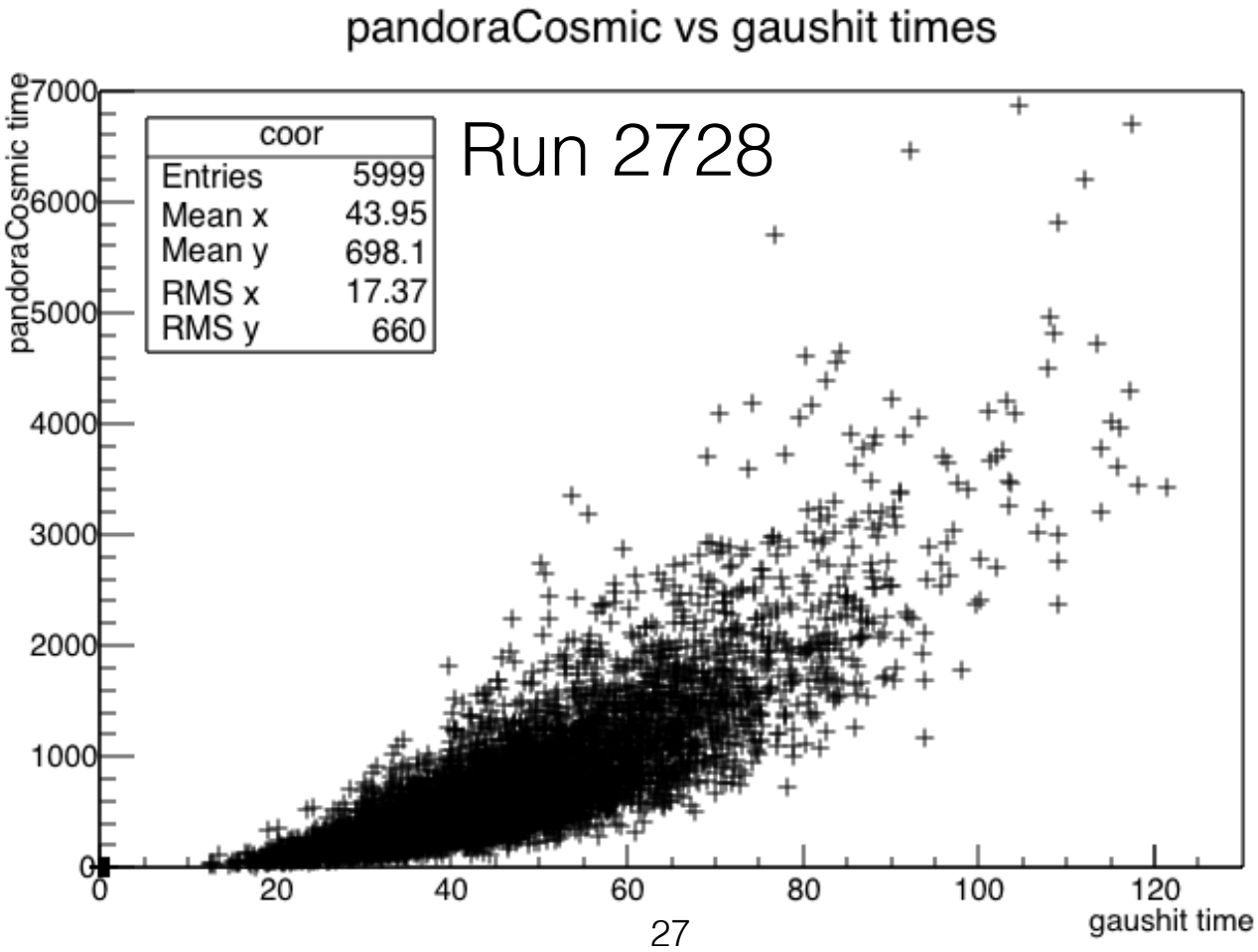
How To Improve?

- Age old Cart/Horse Problem:
  - Need track to do calorimetry and particle id
  - Need calorimetry and particle id to get optimal track fit
- Need a final track fit stage once energy estimated
  - Probably will also want ability to do a final final track fit at the analysis stage too
- Subject for discussion sections?
  - Current track fit module incorporates track finding
  - There is also a hit ordering problem which can be sensitive to changes in track parameters
  - Is there a way to separate the actual Kalman Filter fit from the hit finding?

Noise Impacts on CPU Time

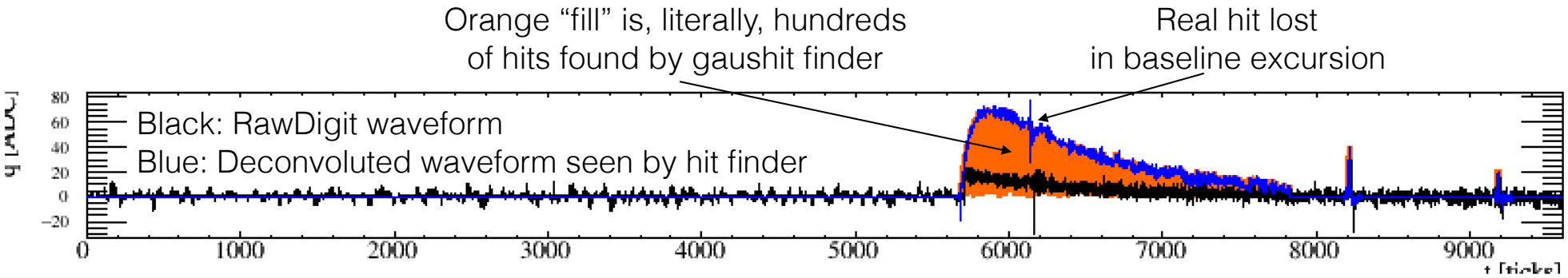
- Leon Rochester did a quick study of module timing in the single pass neutrino ID reconstruction by parsing the output of the log files.

- He found examples of long reconstruction times in the “cosmic” pass of Pandora and found a correlation to the hit finding:



Noise Impacts on CPU Time

- Chirping wires are the primary source of the excessive hit finding:



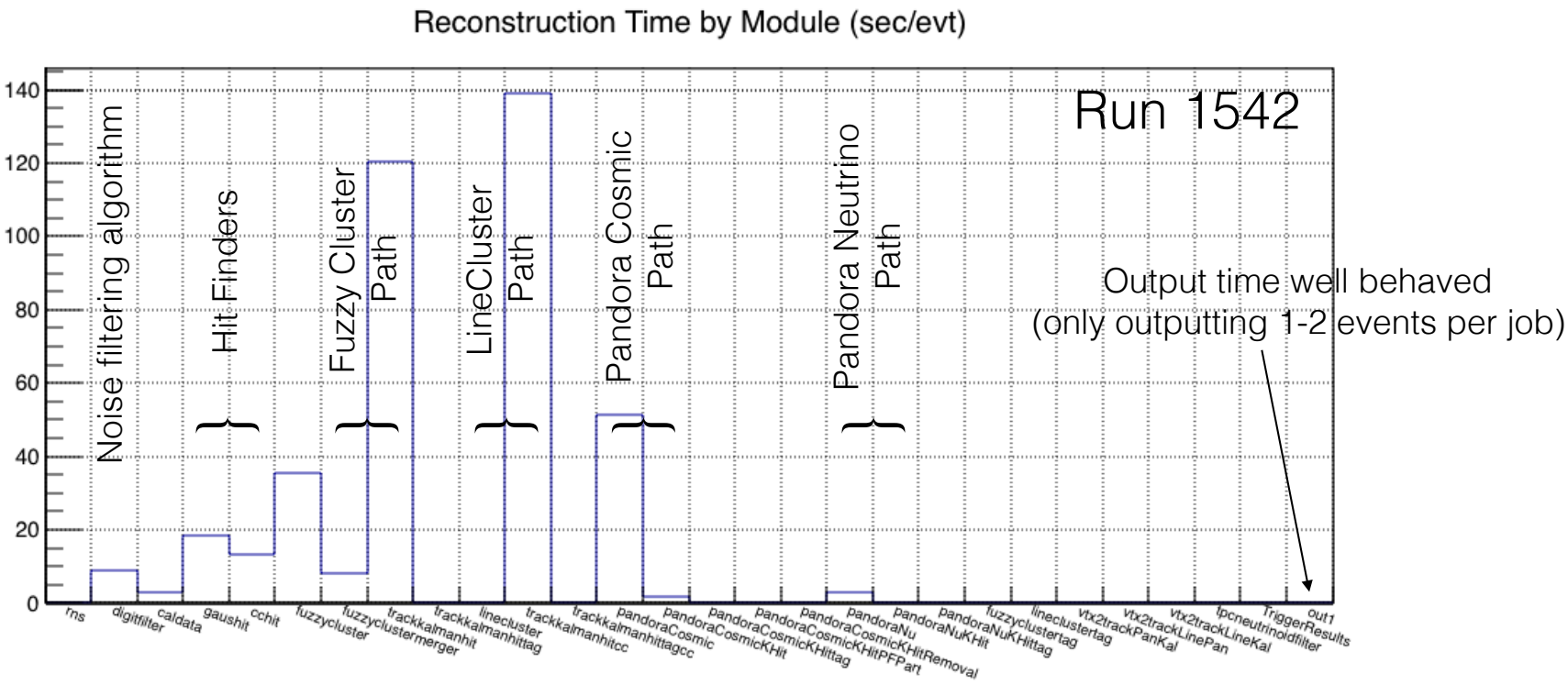
- When the channel transitions into the “live” state there is a baseline shift which then decays away back to pedestal
- The deconvolution algorithm reacts to the step from quiescent to live with an even larger baseline excursion
- The guashit finder is set to find pulses of an assumed maximum width, otherwise it assumes it has multiple hits and essentially fills the area under the chirping signal until it returns to baseline

Noise Impacts on CPU Time

- Obviously, more noise turns into more work for the hit finder, resulting in more hits output to the downstream reconstruction
  - Best fix is to do better at reducing the noise input to the hit finder
  - In addition, should make the gaushit finder “smarter”
    - Incorporate some of the same strategies for handling wide pulses as adopted by Bruce Baller in the cchit finder?
- Questions for improving the hit finder (for later discussion?)
  - The hit finders are “public” code in LArSoft
    - gaushit finder is in common use amongst all experiments?
  - Original author has other responsibilities now and has moved on, who assumes ownership of this code now?
  - SCD did a nice profiling study of reconstruction last Winter with specific suggestions on how to improve the hit finder… how can we get those suggestions incorporated into the module?
  - Is it possible to consolidate into a single hit finder for LArSoft
    - e.g. merge the best parts from gaushit and cchit?
  - This is not a new conversation

Noise Impacts on CPU Time

- Note that it is not only the Pandora path that is affected:



- Fuzzy Cluster and LineCluster rely on the track fit algorithm to do its own pattern recognition to find and fit tracks. The overall impact of the extra hits from noise significantly impacts those reconstruction paths as well.

Partial List of Various Lessons Learned

For Discussion Section?

- Channel Status information
  - Previously was contained in a single C++ class definition - need ability to communicate transient status info in recon
  - This is high priority and needs to be done in concert with the LArSoft folks
- Access to conditions information in reconstruction
  - Drift field, wire bias, live time, etc. - currently all set by fhicl parameters
- Database access
  - Requires network access - not always available for developers
- I was asked to mention pedestals
  - There is need for a good solution
- Release coordination
  - Excellent system in place yet we still have problems - e.g. the event display is currently broken
- Other noteworthy discussion items might include:
  - Realities of data volume impact general reconstruction philosophy - e.g. you can't simply assume you will have the luxury of running your favorite reconstruction in the data processing so will need mechanism for evaluating which to run
  - The issue tracker/gannt system is extremely useful for organizing the work effort
  - Cosmic Ray tagging and two pass reconstruction
- etc.

Summary

Summary

- MicroBooNE presently has several areas of focus
  - Data driven areas:
    - Developing signal processing algorithms to help mitigate the effects of the noise currently in the raw data stream
    - Improving the ability of the pattern recognition and tracking algorithms to handle dead channel gaps
  - Monte Carlo driven areas:
    - Improving the shower reconstruction with the current primary focus to improve both the efficiency and resolution of the pi-zero reconstruction
    - Improving the track fits to address issues with the track start point, angular resolution and energy reconstruction
    - Developing end-to-end analyses with the goal of providing feedback to further needed improvements in reconstruction
- One cannot emphasize enough the importance of developing end to end analysis chains to provide feedback to reconstruction