LCIO Data Improvements

Norman Graf (SLAC)

HPS Software Meeting

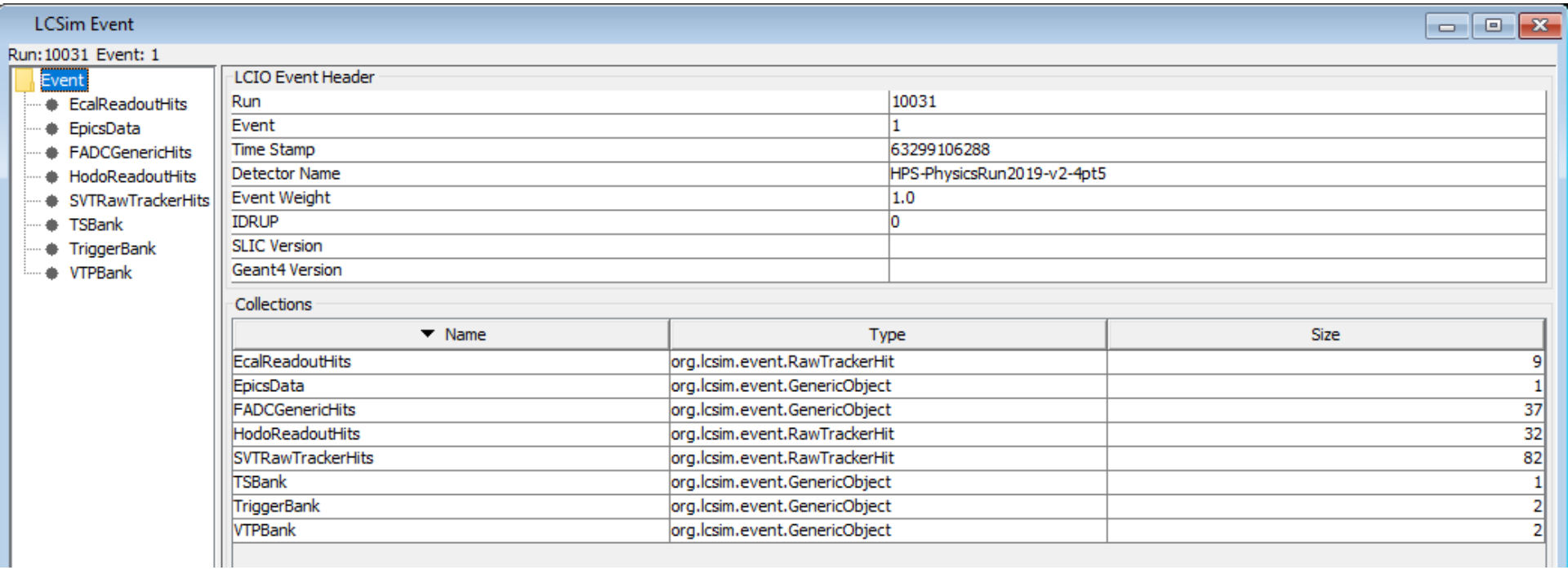
US Tax Day, 2020

Issue

- How does a 2GB Raw data file turn into a 5, 6 or 7GB Recon data file?
- What can we do about it?
- What do we need to do about it?
- What is the role of the recon file?
- What is the role of a DST file?

evio vs LCIO

- Is it just because LCIO is a lousy file format?
- Run a no-op steering file which only converts the trigger, raw SVT, Ecal and Hodoscope data from evio to lcio. (No evio mapping for rest of EDM)



evio vs LCIO

- Compare files sizes for files with same content.
- ls -lsh

  11M 10M Jul 15 08:22 hps_010031.1k_0.evio

  6.6M 6.4M Jul 15 08:23 hps_010031.1k_0.slcio
- sLCIO file is actually 1/3 SMALLER than the evio!
- Can we compress LCIO any further?

  6.3M 6.2M Jul 15 08:23 hps_010031.1k_0.slcio.gz
- It appears that sLCIO (LCIO EDM with SIO file format) is already a pretty good solution.
  - Many years ago significant effort was expended in developing rLCIO, LCIO EDM + root persistency.
    - It was larger, slower, and suffered from problems with every new root release. It was abandoned.

Recon Status

- Status of the reconstruction is still in flux.
- Little (no?) effort has been devoted to limiting content or file size.
- Effort concentrated on understanding efficiency, resolution, etc. i.e.”physics” performance.
- Latest “pass0” steering file in git iss687_dev
  - Includes both SeedTracker/GBL & Kalman Filter to enable comparison of tracking.
- Production Reconstruction will differ substantially.
- Nevertheless…

Recon Output I (iss687_dev)

LCSim Event

Run: 10031 Event: 8573619

- Event
  - BeamspotConstrainedV0Candidates
  - BeamspotConstrainedV0Candidates_KF
  - BeamspotConstrainedV0Vertices
  - BeamspotConstrainedV0Vertices_KF
  - EcalCalHits
  - EcalClusters
  - EcalClustersCorr
  - EcalReadoutHits
  - EcalUncalHits
  - FADCGenericHits
  - FinalStateParticles
  - FinalStateParticles_KF
  - GBLKinkData
  - GBLKinkDataRelations
  - GBLStripClusterData
  - GBLStripClusterDataRelations
  - GBLTracks
  - HelicalTrackHitRelations
  - HelicalTrackHits
  - HodoCalHits
  - HodoGenericClusters
  - HodoReadoutHits
  - KFTrackData
  - KFTrackDataRelations
  - KalmanFullTracks
  - MatchedToGBLTrackRelations
  - MatchedTracks
  - OtherElectrons
  - RFHits
  - RotatedHelicalTrackHitRelations
  - RotatedHelicalTrackHits
  - SVTFittedRawTrackerHits
  - SVTRawTrackerHits
  - SVTShapeFitParameters
  - StripClusterer_SiTrackerHitStrip1D
  - TSBank
  - TargetConstrainedV0Candidates
  - TargetConstrainedV0Candidates_KF
  - TargetConstrainedV0Vertices

LCIO Event Header

Run	10031
Event	8573619
Time Stamp	902324918596
Detector Name	HPS_V0Align_OPAngleBOT_m0_8mrad_iter1
Event Weight	1.0
IDRUP	0
SLIC Version	
Geant4 Version	

Collections

Name	Type	Size
BeamspotConstrainedV0Candidates	org.lcsim.event.ReconstructedParticle	0
BeamspotConstrainedV0Candidates_KF	org.lcsim.event.ReconstructedParticle	0
BeamspotConstrainedV0Vertices	org.lcsim.event.Vertex	0
BeamspotConstrainedV0Vertices_KF	org.lcsim.event.Vertex	0
EcalCalHits	org.lcsim.event.CalorimeterHit	11
EcalClusters	org.lcsim.event.Cluster	2
EcalClustersCorr	org.lcsim.event.Cluster	2
EcalReadoutHits	org.lcsim.event.RawTrackerHit	10
EcalUncalHits	org.lcsim.event.CalorimeterHit	11
FADCGenericHits	org.lcsim.event.GenericObject	37
FinalStateParticles	org.lcsim.event.ReconstructedParticle	2
FinalStateParticles_KF	org.lcsim.event.ReconstructedParticle	3
GBLKinkData	org.lcsim.event.GenericObject	1
GBLKinkDataRelations	org.lcsim.event.LCRelation	1
GBLStripClusterData	org.lcsim.event.GenericObject	14
GBLStripClusterDataRelations	org.lcsim.event.LCRelation	14
GBLTracks	org.lcsim.event.Track	1
HelicalTrackHitRelations	org.lcsim.event.LCRelation	26
HelicalTrackHits	org.lcsim.event.TrackerHit	13
HodoCalHits	org.lcsim.event.CalorimeterHit	15
HodoGenericClusters	org.lcsim.event.GenericObject	6
HodoReadoutHits	org.lcsim.event.RawTrackerHit	32
KFTrackData	org.lcsim.event.GenericObject	1
KFTrackDataRelations	org.lcsim.event.LCRelation	1
KalmanFullTracks	org.lcsim.event.Track	1
MatchedToGBLTrackRelations	org.lcsim.event.LCRelation	1
MatchedTracks	org.lcsim.event.Track	1
OtherElectrons	org.lcsim.event.ReconstructedParticle	0

Recon Output II (iss687_dev)

LCSim Event

Run: 10031 Event: 8573619

- Event
  - BeamspotConstrainedV0Candidates
  - BeamspotConstrainedV0Candidates_KF
  - BeamspotConstrainedV0Vertices
  - BeamspotConstrainedV0Vertices_KF
  - EcalCalHits
  - EcalClusters
  - EcalClustersCorr
  - EcalReadoutHits
  - EcalUncalHits
  - FADCGenericHits
  - FinalStateParticles
  - FinalStateParticles_KF
  - GBLKinkData
  - GBLKinkDataRelations
  - GBLStripClusterData
  - GBLStripClusterDataRelations
  - GBLTracks
  - HelicalTrackHitRelations
  - HelicalTrackHits
  - HodoCalHits
  - HodoGenericClusters
  - HodoReadoutHits
  - KFTrackData
  - KFTrackDataRelations
  - KalmanFullTracks
  - MatchedToGBLTrackRelations
  - MatchedTracks
  - OtherElectrons
  - RFHits
  - RotatedHelicalTrackHitRelations
  - RotatedHelicalTrackHits
  - SVTFittedRawTrackerHits
  - SVTRawTrackerHits
  - SVTShapeFitParameters
  - StripClusterer_SiTrackerHitStrip1D
  - TSBank
  - TargetConstrainedV0Candidates
  - TargetConstrainedV0Candidates_KF
  - TargetConstrainedV0Vertices

LCIO Event Header

Run	10031
Event	8573619
Time Stamp	902324918596
Detector Name	HPS_V0Align_OPAngleBOT_m0_8mrad_iter1
Event Weight	1.0
IDRUP	0
SLIC Version	
Geant4 Version	

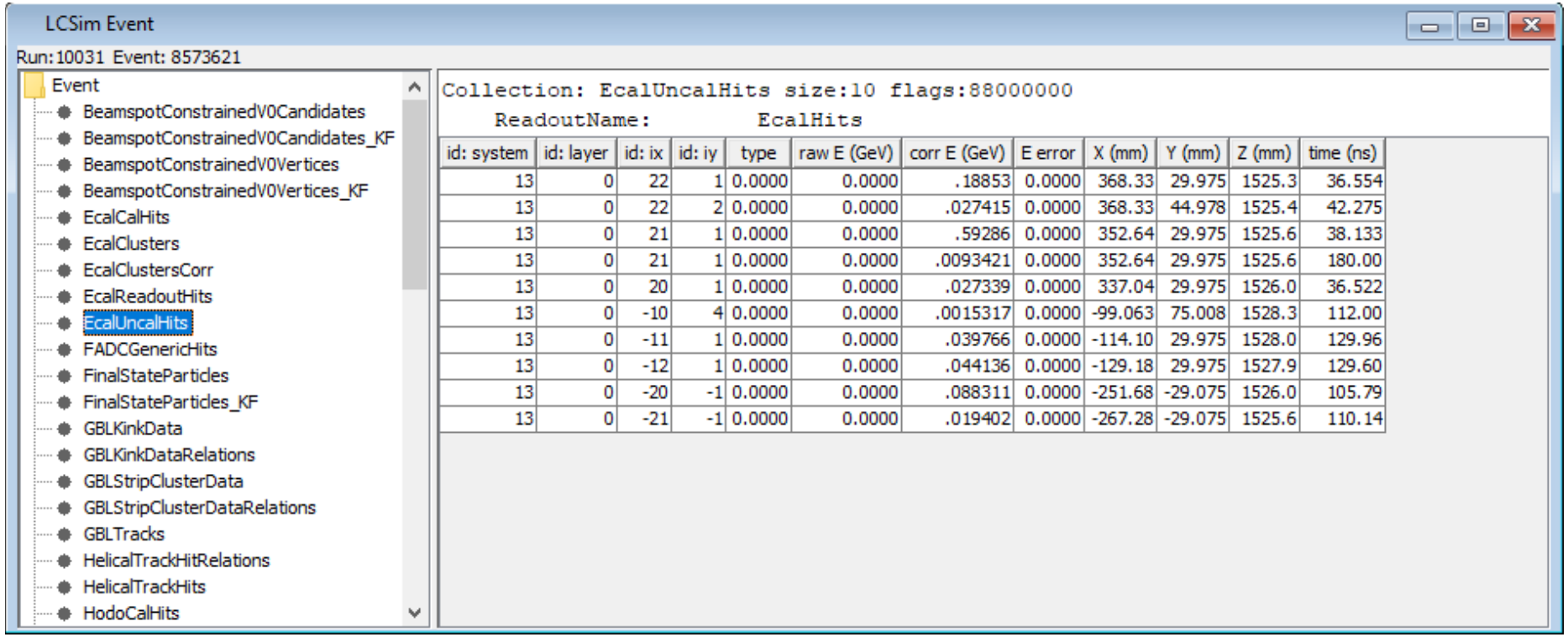
Collections

Name	Type	Size
OtherElectrons	org.lcsim.event.ReconstructedParticle	0
OtherElectrons	org.lcsim.event.ReconstructedParticle	0
RFHits	org.lcsim.event.GenericObject	1
RotatedHelicalTrackHitRelations	org.lcsim.event.LCRelation	13
RotatedHelicalTrackHits	org.lcsim.event.TrackerHit	13
SVTFittedRawTrackerHits	org.lcsim.event.LCRelation	109
SVTRawTrackerHits	org.lcsim.event.RawTrackerHit	94
SVTShapeFitParameters	org.lcsim.event.GenericObject	109
StripClusterer_SiTrackerHitStrip1D	org.lcsim.event.TrackerHit	49
TSBank	org.lcsim.event.GenericObject	1
TargetConstrainedV0Candidates	org.lcsim.event.ReconstructedParticle	0
TargetConstrainedV0Candidates_KF	org.lcsim.event.ReconstructedParticle	0
TargetConstrainedV0Vertices	org.lcsim.event.Vertex	0
TargetConstrainedV0Vertices_KF	org.lcsim.event.Vertex	0
TrackData	org.lcsim.event.GenericObject	1
TrackDataRelations	org.lcsim.event.LCRelation	1
TrackResidualsGBL	org.lcsim.event.GenericObject	1
TrackResidualsGBLRelations	org.lcsim.event.LCRelation	1
TriggerBank	org.lcsim.event.GenericObject	2
UnconstrainedV0Candidates	org.lcsim.event.ReconstructedParticle	0
UnconstrainedV0Candidates_KF	org.lcsim.event.ReconstructedParticle	0
UnconstrainedV0Vertices	org.lcsim.event.Vertex	0
UnconstrainedV0Vertices_KF	org.lcsim.event.Vertex	0
UnconstrainedVcCandidates	org.lcsim.event.ReconstructedParticle	0
UnconstrainedVcCandidates_KF	org.lcsim.event.ReconstructedParticle	0
UnconstrainedVcVertices	org.lcsim.event.Vertex	0
UnconstrainedVcVertices_KF	org.lcsim.event.Vertex	0
VTPBank	org.lcsim.event.GenericObject	7

Recon Output

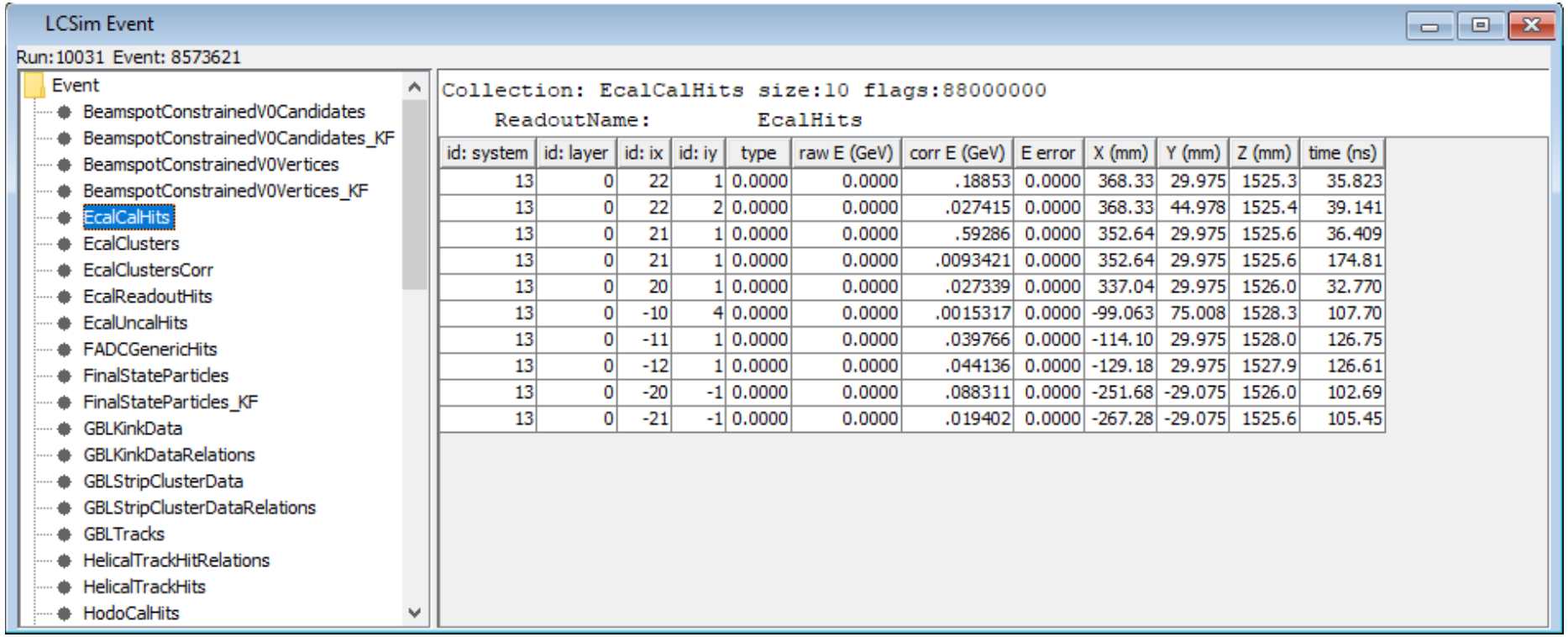
- So, it's clear that there is a LOT of extra data included in this file.
  - For instance, we won't have both SeedTracker/GBL and Kalman Filter tracks and ReconstructedParticles.
- Won't try to analyze every collection here, but it's clear that we need to survey what's going into the output and justify what's there.

Recon Output EcalUncalHits



id: system	id: layer	id: ix	id: iy	type	raw E (GeV)	corr E (GeV)	E error	X (mm)	Y (mm)	Z (mm)	time (ns)
13	0	22	1	0.0000	0.0000	.18853	0.0000	368.33	29.975	1525.3	36.554
13	0	22	2	0.0000	0.0000	.027415	0.0000	368.33	44.978	1525.4	42.275
13	0	21	1	0.0000	0.0000	.59286	0.0000	352.64	29.975	1525.6	38.133
13	0	21	1	0.0000	0.0000	.0093421	0.0000	352.64	29.975	1525.6	180.00
13	0	20	1	0.0000	0.0000	.027339	0.0000	337.04	29.975	1526.0	36.522
13	0	-10	4	0.0000	0.0000	.0015317	0.0000	-99.063	75.008	1528.3	112.00
13	0	-11	1	0.0000	0.0000	.039766	0.0000	-114.10	29.975	1528.0	129.96
13	0	-12	1	0.0000	0.0000	.044136	0.0000	-129.18	29.975	1527.9	129.60
13	0	-20	-1	0.0000	0.0000	.088311	0.0000	-251.68	-29.075	1526.0	105.79
13	0	-21	-1	0.0000	0.0000	.019402	0.0000	-267.28	-29.075	1525.6	110.14

Recon Output EcalCalHits



id: system	id: layer	id: ix	id: iy	type	raw E (GeV)	corr E (GeV)	E error	X (mm)	Y (mm)	Z (mm)	time (ns)
13	0	22	1	0.0000	0.0000	.18853	0.0000	368.33	29.975	1525.3	35.823
13	0	22	2	0.0000	0.0000	.027415	0.0000	368.33	44.978	1525.4	39.141
13	0	21	1	0.0000	0.0000	.59286	0.0000	352.64	29.975	1525.6	36.409
13	0	21	1	0.0000	0.0000	.0093421	0.0000	352.64	29.975	1525.6	174.81
13	0	20	1	0.0000	0.0000	.027339	0.0000	337.04	29.975	1526.0	32.770
13	0	-10	4	0.0000	0.0000	.0015317	0.0000	-99.063	75.008	1528.3	107.70
13	0	-11	1	0.0000	0.0000	.039766	0.0000	-114.10	29.975	1528.0	126.75
13	0	-12	1	0.0000	0.0000	.044136	0.0000	-129.18	29.975	1527.9	126.61
13	0	-20	-1	0.0000	0.0000	.088311	0.0000	-251.68	-29.075	1526.0	102.69
13	0	-21	-1	0.0000	0.0000	.019402	0.0000	-267.28	-29.075	1525.6	105.45

What is the role of the recon file?

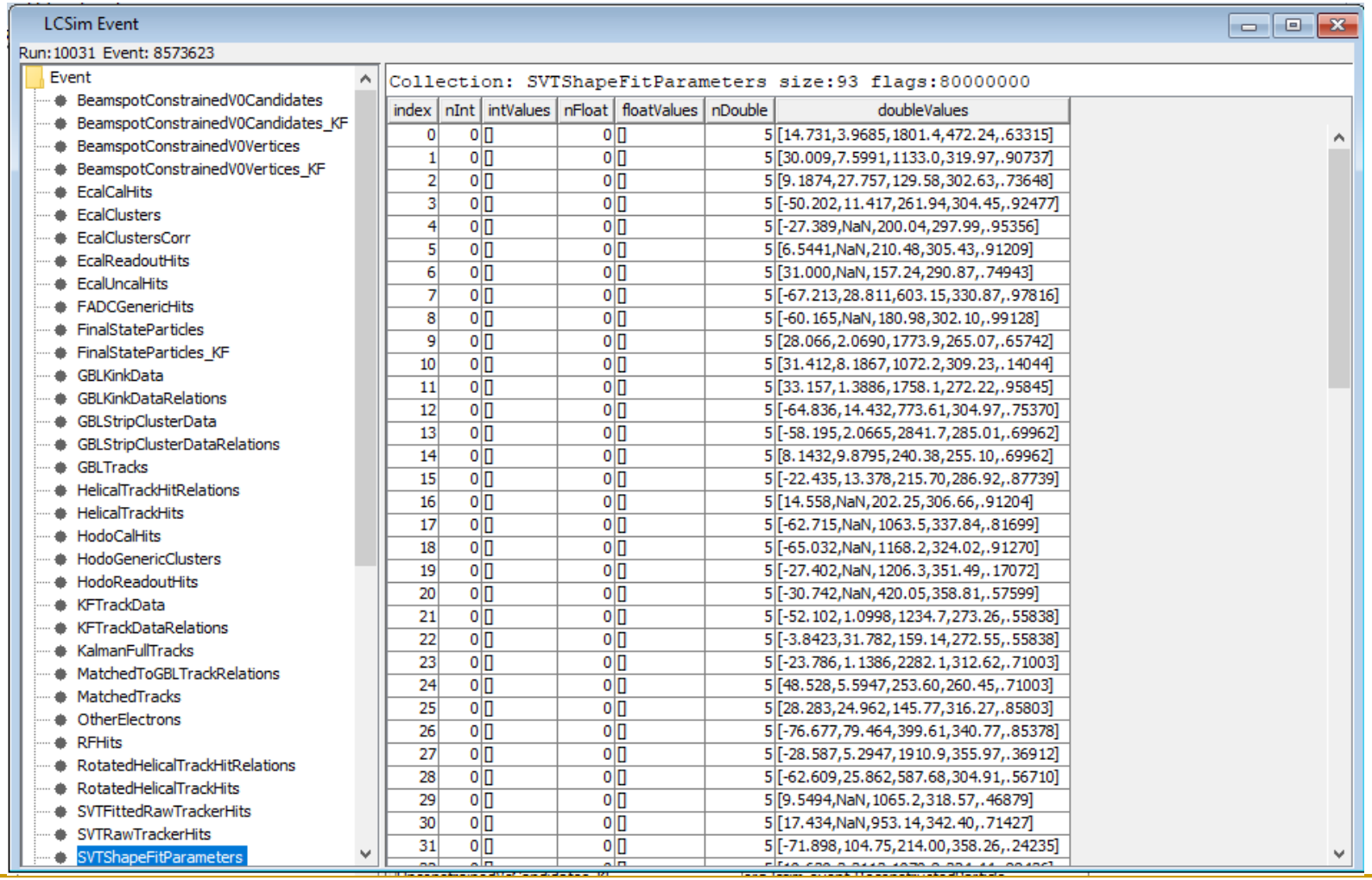
- Historically we have kept all of the data, including the raw data, to enable re-reconstruction from the lcio files.
- At this point, we should be able to drop the raw waveforms and only save the fitted t0 and pulse area.
  - Saves ~2/3 of the original evio file size.
- Do we need to save individual Ecal crystals or SVT readout channels, or can we live with just ECal clusters or StripClusterer_SiTrackerHitStrip1D?

What is the role of the recon file?

- Use for future re-reconstruction.
- Historically we have kept all of the data, including the raw data, to enable re-reconstruction from the lcio files.
- At this point, we should be able to drop the raw waveforms and only save the fitted t0 and pulse area.
  - Saves ~2/3 of the original evio file size.
- Do we need to save individual Ecal crystals or SVT readout channels, or can we live with just ECal clusters or StripClusterer_SiTrackerHitStrip1D?

Are we ready to discard raw waveforms?

- SVTShapeFitParameters has quite a few NaN entries…



Optimization

- It's clear we can gain quite a lot simply by not writing out unnecessary collections of objects.
- Much of our data has been shoe-horned into existing LCIO objects or into GenericObject collections which are not optimized for HPS.
- Time to consider our own custom HPS LCIO Objects?

Tracker Hits

- Unlike the SeedTracker, the Kalman Filter only uses 1-D hits. Furthermore current TrackerHits employ a byzantine set of LCRelations to link the 3D “cross” hits to 1D strip cluster hits to 1D channel hits.
- LCIO TrackerHit is a 3D hit.
- Can gain substantial amount of space reduction by moving to a 1D hit class.
- Instead of (x,y,z) [3] and cov(x,y,z)[6] we simply store u[1] and du[1].

Ouptu Data Size Reduction

- A number of strategies can gain us a substantial reduction in the size of our recon output files.
- Dropping the “raw” waveforms is easiest.
  - Are we satisfied with out current pulse fitting?
- Not running the SeedTracker is straightforward
  - Need to validate Kalman Filter.
    - More, better tracks faster.
- Pruning un-needed collections is next.
- Can consider DST set of collections which is optimized for “physics” analysis.
  - Just ReconstructedParticles?
- Implement custom HPS LCIO classes and restructure our EDM and code to accommodate would take more time and effort.