

# LCIO Data Improvements

---

Norman Graf (SLAC)  
HPS Software Meeting  
US Tax Day, 2020

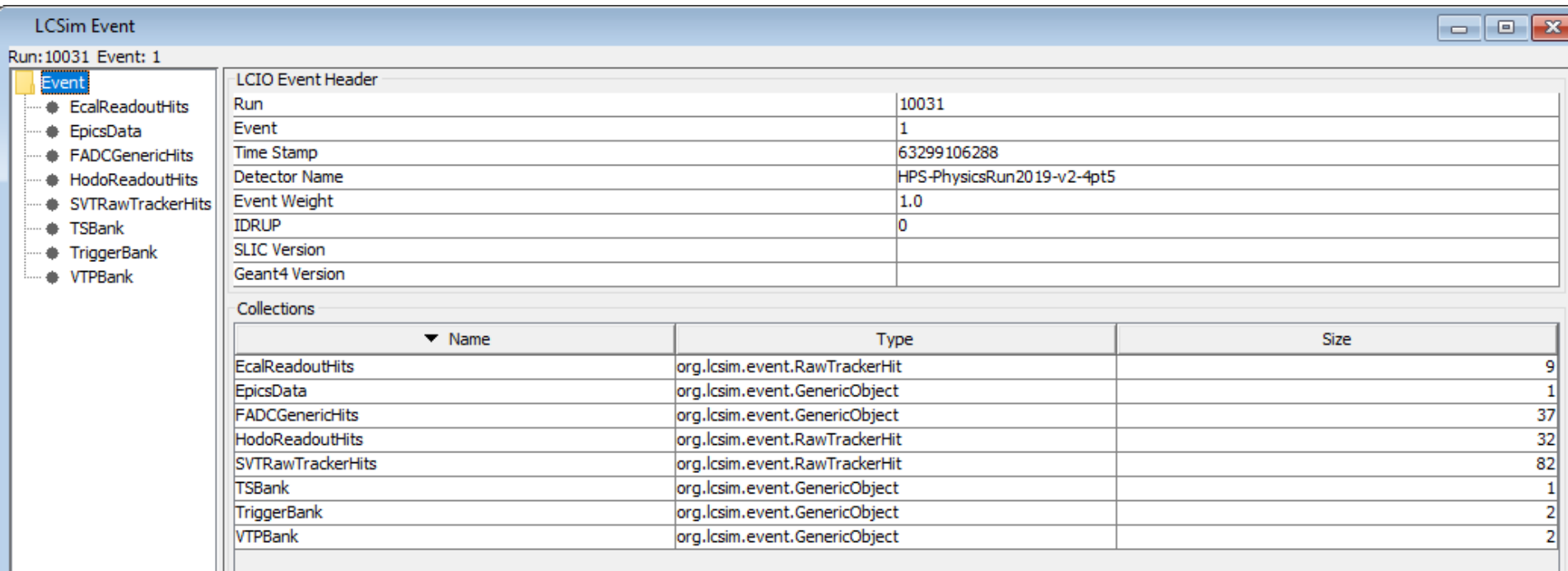
---

# 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)



The screenshot shows the 'LcSim Event' window for Run: 10031, Event: 1. The left sidebar lists event components: Event, EcalReadoutHits, EpicsData, FADCGenericHits, HodoReadoutHits, SVTRawTrackerHits, TSBank, TriggerBank, and VTPBank. The main area displays the LCIO Event Header with the following fields:

Field	Value
Run	10031
Event	1
Time Stamp	63299106288
Detector Name	HPS-PhysicsRun2019-v2-4pt5
Event Weight	1.0
IDRUP	0
SLIC Version	
Geant4 Version	

Below the header is a table of Collections:

Name	Type	Size
EcalReadoutHits	org.lcsim.event.RawTrackerHit	9
EpicsData	org.lcsim.event.GenericObject	1
FADCGenericHits	org.lcsim.event.GenericObject	37
HodoReadoutHits	org.lcsim.event.RawTrackerHit	32
SVTRawTrackerHits	org.lcsim.event.RawTrackerHit	82
TSBank	org.lcsim.event.GenericObject	1
TriggerBank	org.lcsim.event.GenericObject	2
VTPBank	org.lcsim.event.GenericObject	2

# 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
- FinalStatePartides\_KF
- GBLKinkData
- GBLKinkDataRelations
- GBLStripClusterData
- GBLStripClusterDataRelations
- GBLTracks
- HelicalTrackHitRelations
- HelicalTrackHits
- HodoCalHits
- HodoGenericClusters
- HodoReadoutHits
- KFTrackData
- KFTrackDataRelations
- KalmanFullTracks
- MatchedToGBLTrackRelations
- MatchedTracks
- OtherElectrons
- RFHits
- RotatedHelicalTrackHitRelations
- RotatedHelicalTrackHits
- SVTFittedRawTrackerHits
- SVTRawTrackerHits
- SVTShapeFitParameters
- StripClusterer\_SITrackerHitStrip 1D
- TSBank
- TargetConstrainedV0Candidates
- TargetConstrainedV0Candidates\_KF
- TargetConstrainedV0Vertices
- TargetConstrainedV0Vertices\_KF
- TrackData
- TrackDataRelations
- TrackResidualsGBL
- TrackResidualsGBLRelations
- TriggerBank
- UnconstrainedV0Candidates
- UnconstrainedV0Candidates\_KF
- UnconstrainedV0Vertices
- UnconstrainedV0Vertices\_KF
- UnconstrainedVcCandidates
- UnconstrainedVcCandidates\_KF
- UnconstrainedVcVertices
- UnconstrainedVcVertices\_KF
- VTPBank

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_SITrackerHitStrip 1D	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

LCSim Event  
Run: 10031 Event: 8573621

Event

- BeamspotConstrainedV0Candidates
- BeamspotConstrainedV0Candidates\_KF
- BeamspotConstrainedV0Vertices
- BeamspotConstrainedV0Vertices\_KF
- EcalCalHits
- EcalClusters
- EcalClustersCorr
- EcalReadoutHits
- **EcalUncalHits**
- FADCGenericHits
- FinalStateParticles
- FinalStateParticles\_KF
- GBLKinkData
- GBLKinkDataRelations
- GBLStripClusterData
- GBLStripClusterDataRelations
- GBLTracks
- HelicalTrackHitRelations
- HelicalTrackHits
- HodoCalHits

Collection: EcalUncalHits size:10 flags:88000000  
ReadoutName: EcalHits

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

LCSim Event  
Run: 10031 Event: 8573621

Event

- BeamspotConstrainedV0Candidates
- BeamspotConstrainedV0Candidates\_KF
- BeamspotConstrainedV0Vertices
- BeamspotConstrainedV0Vertices\_KF
- **EcalCalHits**
- EcalClusters
- EcalClustersCorr
- EcalReadoutHits
- EcalUncalHits
- FADCGenerichits
- FinalStatePartides
- FinalStatePartides\_KF
- GBLKinkData
- GBLKinkDataRelations
- GBLStripClusterData
- GBLStripClusterDataRelations
- GBLTracks
- HelicalTrackHitRelations
- HelicalTrackHits
- HodoCalHits

Collection: EcalCalHits size:10 flags:88000000  
ReadoutName: EcalHits

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 Icio files.
- At this point, we should be able to drop the raw waveforms and only save the fitted  $t_0$  and pulse area.
  - Saves  $\sim 2/3$  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  $t_0$  and pulse area.
  - Saves  $\sim 2/3$  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...

LCsim Event  
Run:10031 Event: 8573623

Collection: SVTShapeFitParameters size:93 flags:80000000

index	nInt	intValues	nFloat	floatValues	nDouble	doubleValues
0	0	0	0		5	[14.731,3.9685,1801.4,472.24,.63315]
1	0	0	0		5	[30.009,7.5991,1133.0,319.97,.90737]
2	0	0	0		5	[9.1874,27.757,129.58,302.63,.73648]
3	0	0	0		5	[-50.202,11.417,261.94,304.45,.92477]
4	0	0	0		5	[-27.389,NaN,200.04,297.99,.95356]
5	0	0	0		5	[6.5441,NaN,210.48,305.43,.91209]
6	0	0	0		5	[31.000,NaN,157.24,290.87,.74943]
7	0	0	0		5	[-67.213,28.811,603.15,330.87,.97816]
8	0	0	0		5	[-60.165,NaN,180.98,302.10,.99128]
9	0	0	0		5	[28.066,2.0690,1773.9,265.07,.65742]
10	0	0	0		5	[31.412,8.1867,1072.2,309.23,.14044]
11	0	0	0		5	[33.157,1.3886,1758.1,272.22,.95845]
12	0	0	0		5	[-64.836,14.432,773.61,304.97,.75370]
13	0	0	0		5	[-58.195,2.0665,2841.7,285.01,.69962]
14	0	0	0		5	[8.1432,9.8795,240.38,255.10,.69962]
15	0	0	0		5	[-22.435,13.378,215.70,286.92,.87739]
16	0	0	0		5	[14.558,NaN,202.25,306.66,.91204]
17	0	0	0		5	[-62.715,NaN,1063.5,337.84,.81699]
18	0	0	0		5	[-65.032,NaN,1168.2,324.02,.91270]
19	0	0	0		5	[-27.402,NaN,1206.3,351.49,.17072]
20	0	0	0		5	[-30.742,NaN,420.05,358.81,.57599]
21	0	0	0		5	[-52.102,1.0998,1234.7,273.26,.55838]
22	0	0	0		5	[-3.8423,31.782,159.14,272.55,.55838]
23	0	0	0		5	[-23.786,1.1386,2282.1,312.62,.71003]
24	0	0	0		5	[48.528,5.5947,253.60,260.45,.71003]
25	0	0	0		5	[28.283,24.962,145.77,316.27,.85803]
26	0	0	0		5	[-76.677,79.464,399.61,340.77,.85378]
27	0	0	0		5	[-28.587,5.2947,1910.9,355.97,.36912]
28	0	0	0		5	[-62.609,25.862,587.68,304.91,.56710]
29	0	0	0		5	[9.5494,NaN,1065.2,318.57,.46879]
30	0	0	0		5	[17.434,NaN,953.14,342.40,.71427]
31	0	0	0		5	[-71.898,104.75,214.00,358.26,.24235]

---

# 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  $\text{cov}(x,y,z)$ [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 our 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.