HPS Software Review: Offline Data Analysis

Matt Graham,SLAC

HPS Software Review

January 27, 2014

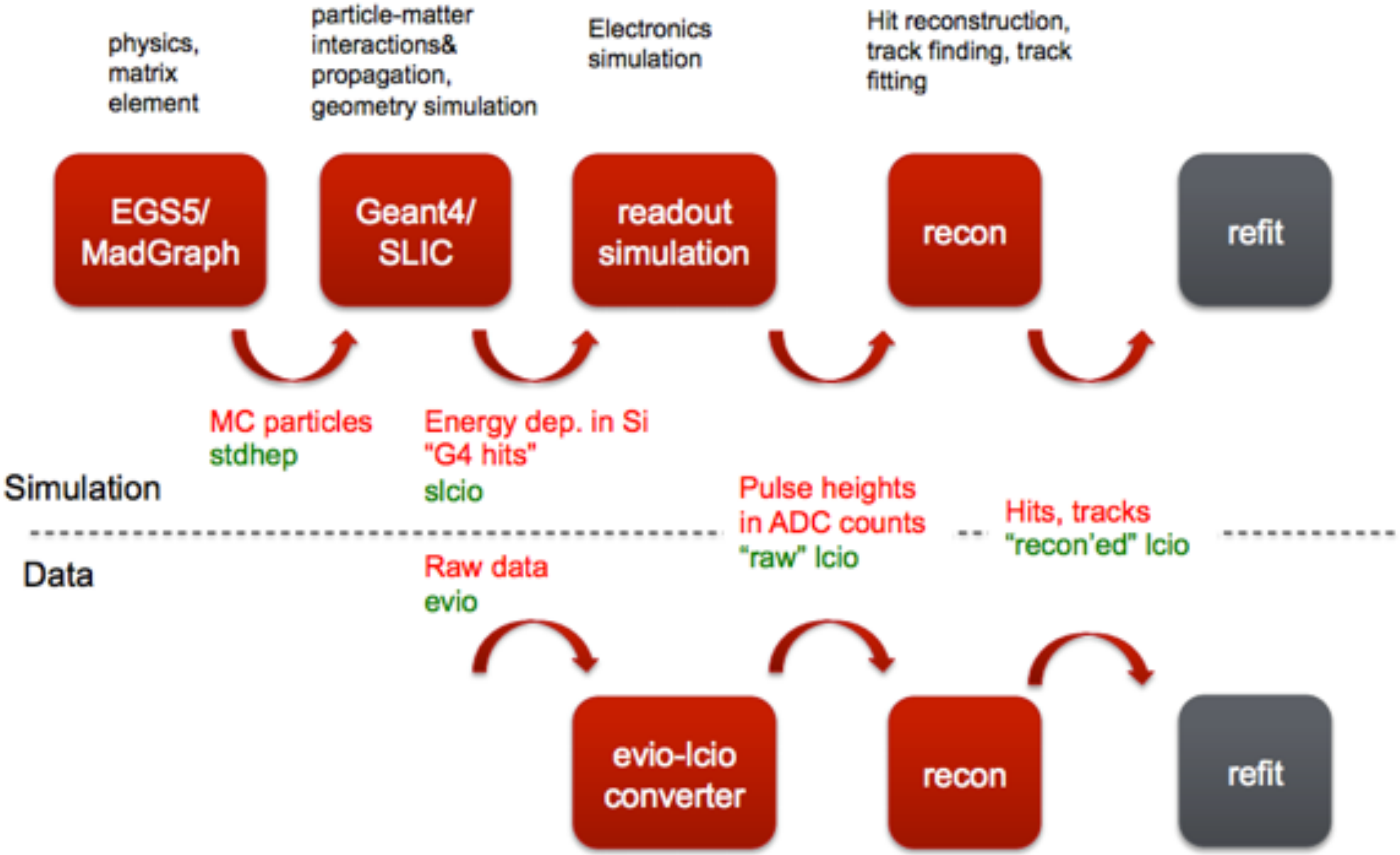
Data Analysis Working Group

- Sub-group of the software group; created to organize and coordinate the work required to go from having data on disk→publishable physics result
- Charges to the group:
  - *Data production & distribution*: analysts get the data they need in a timely fashion
  - *Data quality*: quickly verify that data is physics quality
  - *Physics analysis*: guide physics analyses through to publication; serve as "first line of review"; develop and maintain common physics analysis tools

Data & Simulation Production

- data production: evio(raw hits)→lcio (clusters,tracks,vertices..etc)
  - this is hps-java…code exists and is constantly being refined
  - automated scripts for submitting jobs to batch & bookkeeping exist; exercised for test run
- simulation production: multi-step process
  - event generation (MadGraph), beam overlay (EGS)
  - detector simulation (slic/GEANT4)
  - readout simulation (hps-java)
  - reconstruction (hps-java…just like data)
  - all of the above steps are in good shape (see previous talks by Takashi, Sho); currently writing scripts to automate & link each of these steps.
    - this is "phase-0" of the mock data challenge
- All of the data & sim production will take place at JLAB (see Homer's talk for resources)
- Data Production Manager: overseer of data and simulation production
  - currently: **Sho Uemura**

Data & Simulation Production: Processing Chain



physics, matrix element

particle-matter interactions& propagation, geometry simulation

Electronics simulation

Hit reconstruction, track finding, track fitting

EGS5/ MadGraph

Geant4/ SLIC

readout simulation

recon

refit

MC particles
stdhep

Energy dep. in Si
"G4 hits"
slcio

Pulse heights
in ADC counts
"raw" lcio

Hits, tracks
"recon'ed" lcio

Simulation

Data

Raw data
evio

evio-lcio converter

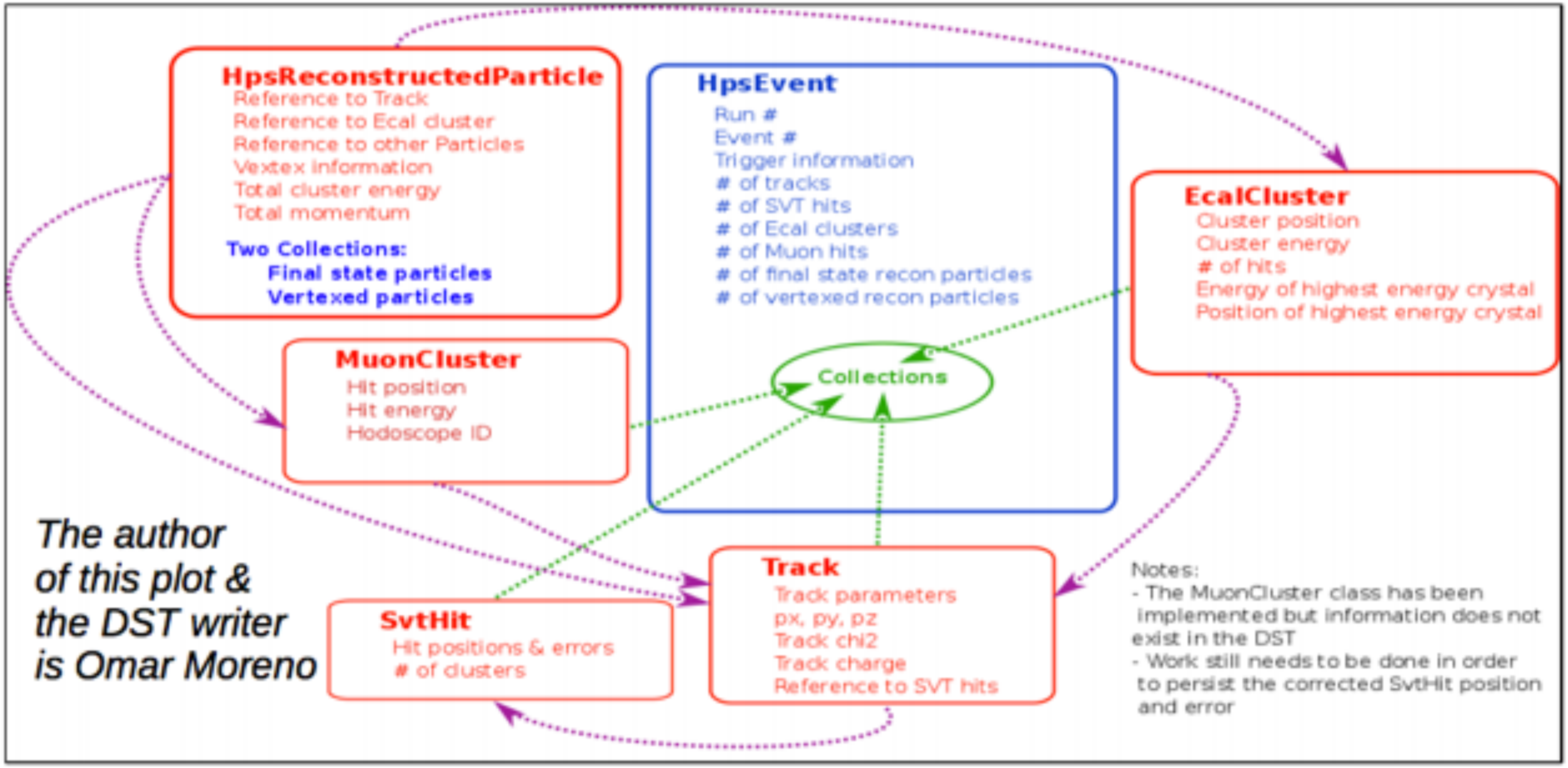
recon

refit

Data & Simulation Production: Reco'd lcio & DSTs

- At the end of the primary processing chain → reconstructed lcio file which contains:
  - all low-level objects (SVT & ECAL ADC counts)…everything the evio file has
  - higher-level objects based on default algorithms (SVT clusters, tracks, vertices, reconstructed particles..etc).
  - this file may be accessed using lcsim/hps-java, c++ (using lcio libraries), directly in ROOT…but it has a fairly complicated structure; just just a list of hits/tracks/etc
- Collaboration requested higher-level output format →DST
  - converts the recon'ed lcio to ROOT TTree, using lcio c++ libraries
  - Omar has written a "dst writer" with a default format (see next page)
  - DST is not intended to be loss-less or to used for low-level tasks like reconstruction…intended to be a light-weight dataset for high-level analysis
  - this default DST will be produced as a part of the processing chain…DST-maker is very fast and the output is a small fraction of the reconned lcio

Example DST Content



HpsReconstructedParticle
Reference to Track
Reference to Ecal cluster
Reference to other Particles
Vextex information
Total cluster energy
Total momentum

Two Collections:
Final state particles
Vertexed particles

HpsEvent
Run #
Event #
Trigger information
# of tracks
# of SVT hits
# of Ecal clusters
# of Muon hits
# of final state recon particles
# of vertexed recon particles

Collections

EcalCluster
Cluster position
Cluster energy
# of hits
Energy of highest energy crystal
Position of highest energy crystal

MuonCluster
Hit position
Hit energy
Hodoscope ID

Track
Track parameters
px, py, pz
Track chi2
Track charge
Reference to SVT hits

SvtHit
Hit positions & errors
# of clusters

The author of this plot & the DST writer is Omar Moreno

Notes:
- The MuonCluster class has been implemented but information does not exist in the DST
- Work still needs to be done in order to persist the corrected SvtHit position and error

- we envision that this DST will satisfy most analysts but, if not, the “dst writer” is easily…Omar is currently the DST tzar.

Offline Data Quality Monitoring

- during the run:
  - if possible and appropriate, contribute to online monitoring so that it includes physics-level measurements of data quality
- after the run:
  - assess the data quality of the run...for example:
    - % of good SVT/ECAL/Muon channels
    - tracking/trigger efficiency
    - tracks/event; vertices/event etc....
    - resolutions...
- maintain run conditions & quality list
- name a Run Quality Manager to lead this effort

Physics Analysis Tools

- Help development of analysis tools:
  - physics objects in lcsim (e.g. “particles” as combination of track,cluster, muon detector object; types of vertices; etc)
  - “good” hit/cluster/track/vertex/event definitions...multiple layers (good/better/best)
  - PID selectors
  - event generation/conversion tools (...going from MadGraph to stdhep for example...)
  - ...other stuff...
  - ...some of these things probably be best done by/in collaboration with sub-system guru

Physics Analysis Coordination

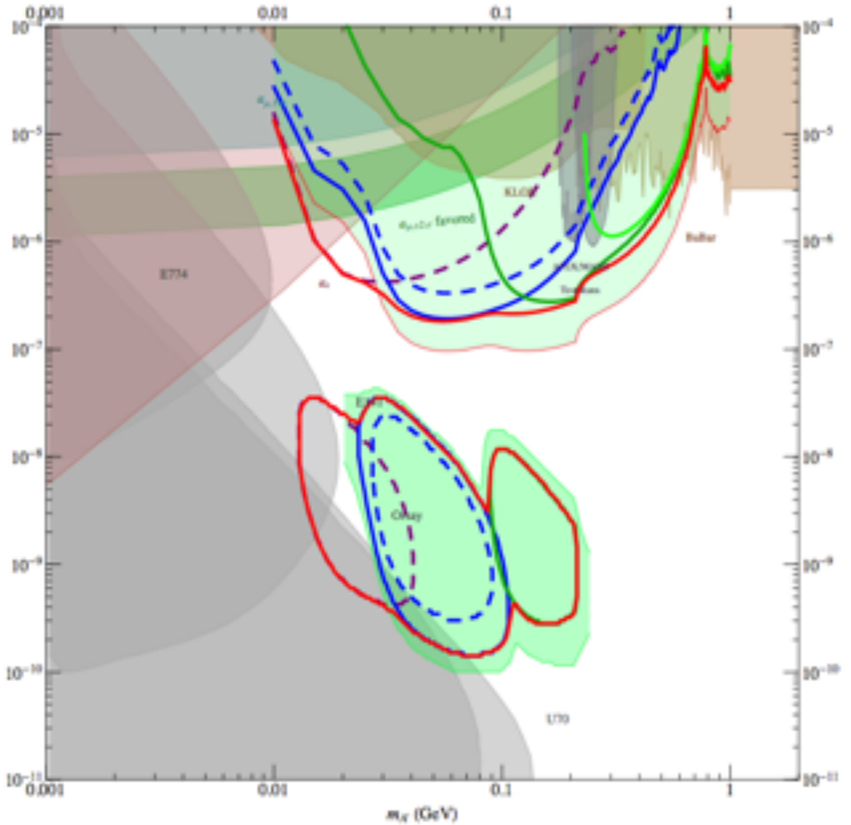
- Currently, we've got main topics important in the near-term/low energy running:
  - standard A' searches (bump-hunt & vertexing)
  - detailed study of trident events: cross-section, shapes, Rad/BH determination
- Longer term/higher energies:
  - true muonium
  - multi-leptons (complex hidden sector)
  - polarization studies
  - di-muon production asymmetries
- Physics analysis will be a team effort…"grad student wanders in the woods for years and returns with result" paradigm not a good one
  - these analyses are fairly complicated with many parts; crosschecks needed; timely result is vital → should have a clear picture of how ***full*** analysis of data will work ***before*** we have data on tape

Physics Analysis: From Proposal to Publication…

- The reach calculation in proposal was based on a primitive analysis/calculation…
- rates from MadGraph
- resolutions from simulations with detailed (but likely still sub-optimal) cuts
- signal extracted via simple cut-and-count

Good enough for a proposal, but there is a lot of work to be done to make a publishable analysis:

- track/event selection optimization
- cross-checks
- systematics
- cross-checks
- signal extraction/limit setting procedures



Physics Analysis Coordination: Mock Data Challenge

- Getting the A’ search analysis work going before first data is a priority for us:
  - help identify potential issues we can address “on-the-floor” (e.g. special runs for calibrations, etc)
  - quick turnaround from data taking to publication
- at DOE reviewer’s suggestion, we’re having a mock data challenge
  - beginning-to-end analysis on a data-sized chunk (1 week, 2.2 GeV) of MC, with MC samples available for tuning
    - first large scale production
  - include some realistic conditions (some sample of noisy, dead SVT channels) but assume detector is aligned/calibrated
  - expect to have datasets ~February (although MDC is starting now), end at the summer collaboration meeting
  - expect this will get many new collaborators involved with analysis

Physics Analysis Coordination: Mock Data Challenge

- What sort of analysis issues:
  - track & event selection optimization
  - adding extra stuff (e.g. recoil electron [...hopefully this will be in default reconstruction by then, but maybe not...even if it is how do you use it in analysis], ecal information, better track finding/fitting etc.)
  - signal extraction
  - limit setting
  - discovery criteria
  - blinding plan

Summary

- Getting HPS results published in a timely fashion is our #1 priority
  - We have (or will soon have) framework in place to process data and produce/process simulation
  - We have plans for ensuring the data is of high quality as we are recording it, both online and offline
  - Physics analysis will soon be ramping up
    - Mock data challenge
    - Software/analysis workshops (this week at SLAC)
    - Many new collaborators eager to get involved