MDC production

Sho Uemura

Status

- Mock data production chain complete
- 1/25 of mock data complete
- Scripts, parameters, data locations: `https://confluence.slac.stanford.edu/display/hpsg/Finding+Monte+Carlo+data+at+JLab`

What's done?

- About 65% of total batch processing complete
- 1/5 (0.4 million triggers) of beam background
- 1/25 (13 million triggers) of trident triggers (with truth info)
- 1/25 (13 million triggers per data set) of mock data

Data sets

- Beam background: unbiased triggers on scattered beam + tridents + hadrons (250 seconds of beam)
- Three mock data sets: background only, bump-hunt signal, vertex signal
  - Each data set: trident triggers and A' from 1 week of beam
  - Parameters set by Natalia and Rouven, hidden from everyone else
- Mock data has no truth information (obviously); additional background-only data set with truth information
- A' samples for testing: work in progress
- Other data sets (A', BH/rad tridents) on request

Data management

- Most data on tape
- Keep:
  - Primary generator output (particles coming out of the target)
  - Readout simulation output (simulated raw data)
  - Recon output
  - DST
- Store in scratch space, and delete after MDC production is done:
  - SLIC output (detector simulation)
- Directory path shows event/data type, file name shows version numbers and run number
  - /mss/hallb/hps/production/dst/mock/mock2/mockv1-egsv3-triv2-g4v1_s2d6-readout20140522-recon20140614-dst20140616_17.root

Batch scripts

- XML Auger scripts, and shell scripts for automating job submission
  - Run mock data readout for runs 1-10000: ./runjob.sh readout/mock.xml 2pt2 1 10000
- Job scripts verify output files, so only successful runs are written to tape
  - No system for checking which jobs succeed and which fail

CPU and storage

- Normalized to 10k trident triggers (85k tridents, or 7.7 ms of beam background)
- Primary generators and SLIC dominate compute; SLIC dominates storage but can be deleted after readout

Sample	Time (CPU-h)	Size (MB)
MadGraph	19	4.5
EGS5/Geant4	0.8	52
SLIC	22	(11000)
Readout	11	180
Recon	2.2	690
DST	0.5	110

Data transfer

- Globus Online transfers between JLab and SLAC
  - About 50 MB/sec
- Using web GUI for now; will eventually be scripted
- Mock data DSTs will be transferred to SLAC soon

Lessons learned

- Find a data production manager
- Better communication with JLab scientific computing
  - We learned about various unenforced batch resource limits by overrunning them
  - Report farm errors instead of just rerunning the jobs
- Farm availability can be variable (average 600 cores some weeks, 200 cores other weeks)