IU Mini Data Challenge

Kei Moriya
Indiana University
GlueX Offline Meeting
October 30, 2013

Generation Summary

- Processed 50M bggen events - 10 hrs worth of data[1]
- 10k events x 5000 files
- All events had primary vertex smeared (smear_thrown_vertex)
- E_γ range was 8.4 - 9.0 GeV
- No EM background added

- Using 192 nodes on IU cluster, can process 2000 files/day = 20M events/day (hdgeant, mcsmear, REST)
- 4739 files had usable REST output (94.8%)
- 161 files got stuck at REST, remaining 100 finished REST but had unusable output (more detailed logs available)

1. according to GlueX Data Challenge Report, $10^7\gamma$/s

Details of Failures, File Size

File size	Number of files	Usable
20MB	22	22
19MB	4719	4717
10-18MB	39	0
1-10MB	52	0
<1MB	7	0
REST fail	161	0
TOTAL	**5000**	**4739**

- Most unusable files are identifiable from file size
- Use hddm_merge_files (modified to work on hddm_r files)
- Sometimes REST will finish, but file will be unusable
- If this is the case, hddm_merge_files will crash
- In most failures, REST will get stuck at a particular event and stay in the queue until the job is killed

Question on Run Number

- I generated all events without specifying the run number, and this gives me RunNumber 2 for all files
- This is rather inconvenient when I merge files together for ease of analysis, since I can't tell which file the event came from
- How do I change the run number, and will the CCDB complain if the number is not within a given range?

Analysis Summary

- Analysis speed depends strongly on processor (channel to analyze) - 1~20 Hz
- For 1Hz, to analyze 50M events takes 2.89 days using 200 cores → Will need to use grid to do more
- REST output files are 19MB each, so that 50M events ~100GB on disk
- Analysis Trees created by GlueX software will be rather large
  → 250GB - 3TB for 50M events/analysis channel (10 hrs of data!!)
- Currently only have minimal cuts on events - DReaction::Set_MinCombinedTrackingFOM(1.e-5)

	channel	file size/1M events
no photons	$K^+\Lambda$	5.5GB
	$K^+\Sigma^0$	21GB
	$K^+\Sigma^+\pi^-$ ($\Sigma^+ \rightarrow p\pi^0$)	62GB
no photons	$K^+\Sigma^-\pi^+$	10GB
	$K^+\Lambda\eta$	65GB

$K^+\Lambda$ Analysis on bggen

- Run $K^+\Lambda$ processor on 50M bggen events, check potential backgrounds
- Processor will run at ~20Hz, can run relatively quickly
- After processing 50M events, 165 events were left with kfit CL > 0.01 (includes vertex constraints) [5.4M combos with converging kfit]
- 120 $p\pi^+\pi^-$, 22 $K^+\Lambda$, 2 $p\pi^+\pi^-\pi^0$, 2 $p\pi^+\pi^-\pi^0\omega$
- Other single backgrounds: $p\pi^+\pi^-\rho^+$,$K^+\Lambda\pi^0$,$pp\bar{p}\pi^0$, $p\rho^+\pi^+2\pi^-\pi^0$, $n\pi^+\rho^0$, $p\pi^+\pi^-\omega$, $p\eta\pi^0$, ...
- Most pressing background issue is π,K separation
- Previous studies show that $p\pi^+\pi^-$ (mostly through ρ^0) is ~10% of total cross section, main background, rejection of 120/5M ~ 10^{-4}

$K^+\Sigma^0$ Analysis on bggen

- Run $K^+\Lambda$ processor on 50M bggen events, check potential backgrounds
- Processor will run at ~4Hz
- After processing 50M events, 31 events were left with kfit CL > 0.01 (includes vertex constraints) [6.1M combos with converging kfit]
- 15 $p\pi^+\pi^-$, 5 $p\ \pi^+\ \rho^-$, 3 $p\pi^+\pi^-\pi^-\rho^+$, 3 $K^+\Sigma^+(1385)\pi^-$, 2 $K^+\Lambda$

$K^+\Sigma^+\pi^-$ Analysis on bggen

- Run $K^+\Sigma^+\pi^-$ processor on 4M bggen events, check potential backgrounds (final state: $pK^+\pi^-\pi^0$)
- Processor will run at ~1Hz, rather slow
- After processing 4M events, 661 events were left with kfit CL > 0.01 (includes vertex constraints) [642k combos with converging kfit]
- 432 $p\pi^+\pi^-\pi^0$, 53 $K^+\Sigma^+\pi^-$, 37 $pK^{*+}K^-$, 33 $p\rho^+\pi^-\pi^0$, 28 $p\pi^+\pi^-$
- Seems like reasonable backgrounds ($\pi \leftrightarrow K$ confusion)

Also processing other channels

such as $K^+\Sigma^0$, $K^+\Lambda\eta$, $K^+\Sigma^-\pi^+$

Summary

- Processed 50M bggen events at IU using 192 CPUs
- Generation took 3 days, 95% success rate
- Analysis is slow without cuts, file sizes are large
- Currently using more than 2TB of disk from this data challenge (mostly analysis)
- Truth information on background events can be reconstructed for most cases; complicated final states are difficult but rare
- Further generation/analysis would require running at either Big Red II (IU-wide cluster), or scientific grid
- Have started looking into setting these options