Experiences with Large Data Sets

Curtis A. Meyer

Data Handling

Baryon PWA analysis using the CLAS g11a data set.

$\gamma p \rightarrow p \pi^+ \pi^-$
$\gamma p \rightarrow p \eta \rightarrow p \pi^+ \pi^- (\pi^o)_{missing}$
$\gamma p \rightarrow p \omega \rightarrow p \pi^+ \pi^- (\pi^o)_{missing}$
$\gamma p \rightarrow p \eta' \rightarrow p \pi^+ \pi^- (\eta)_{missing}$
$\gamma p \rightarrow K^+ \Lambda \rightarrow p \pi^- K^+$

Consistent Analysis using same tools and procedures.

$N^* \rightarrow p\eta$, $p\omega$, $p\eta'$, $K\Lambda$, $p\rho$, $\Delta\pi$
$D^* \rightarrow p\rho$, $\Delta\pi$

∞	$\pi^+\pi^-$ events
15,000,000	ω events
1,400,000	η events
1,000,000	$K\Lambda$
300,000	η' events

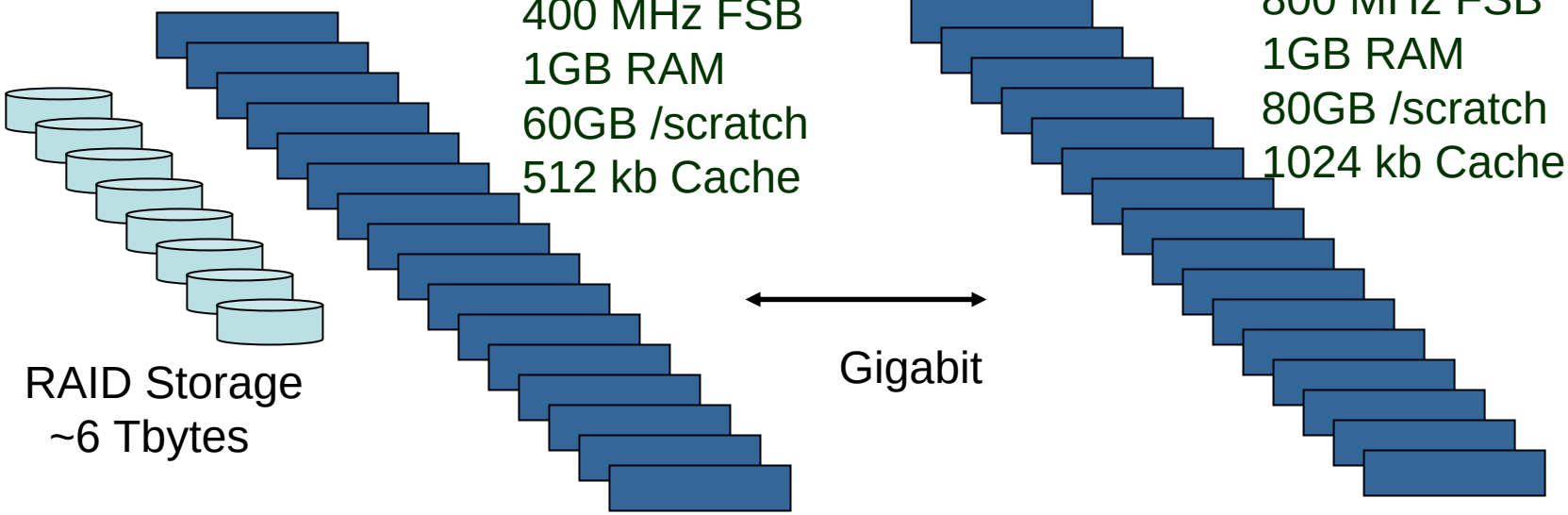
CMU Computers

fermion

Dual 2.4GHz Xeon
400 MHz FSB
1GB RAM
60GB /scratch
512 kb Cache

fermion gold

Dual 3.0GHz Xeon
800 MHz FSB
1GB RAM
80GB /scratch
1024 kb Cache



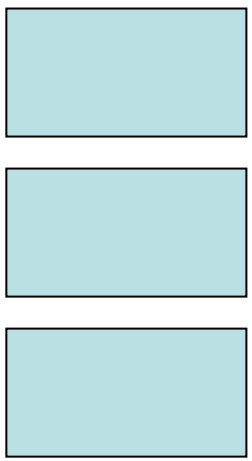
RAID Storage
~6 Tbytes

Gigabit

15 nodes

15 nodes

File Servers



Batch System: pbs
Scheduler: Underlord

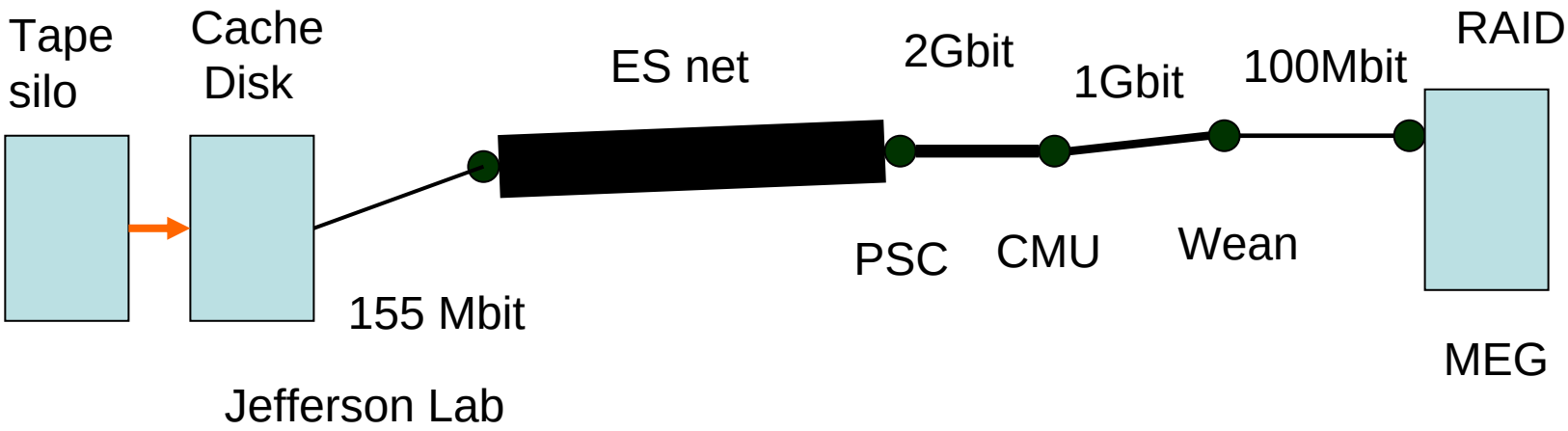
Dual Xeon 2.4, 2.8, 3.1 GHz

System shared with Lattice QCD

~$85,000 Investment for a 66 CPU system

800MHz FSB has brought no improvement (64bit kernel?)

Networks



Transfer Tools:

srmget JLab developed
stages data at JLab and then moves it to remote location.

bbftp LHC tool
similar to ftp, copies files

Can sustain 70-80 Mbits/second of data transfer

We peaked at about 600GB in one 24 hour period.

Tape to Cache is bottle neck

Data Sets

The DST data sets existed on the tape silos at JLab

2positive 1negative track filter:	10,000	~1.2 Gbyte files
flux files (normalization)	10,000	~ 70 Kbyte files
trip files (bad run periods)	10,000	~ 20 Kbyte files

30,000 files with about 12 Terabytes of data

Number of files is big (have trouble doing ls).

Relevant Data spread over multiple files creates a book keeping nightmare.

1.2Gbyte file size is not great 12-20 Gbytes would be more optimal

Some Solutions

The CLAS data model is effectively non-existent.

- Data is repeated multiple times in the stream (up to 6 times)
- Not all the allocated memory is used.
- It is extremely difficult to extract relevant event quantities.
- Little or no compression is implemented.

Assumption for our analysis:

- We do not need access to the raw data for physics.

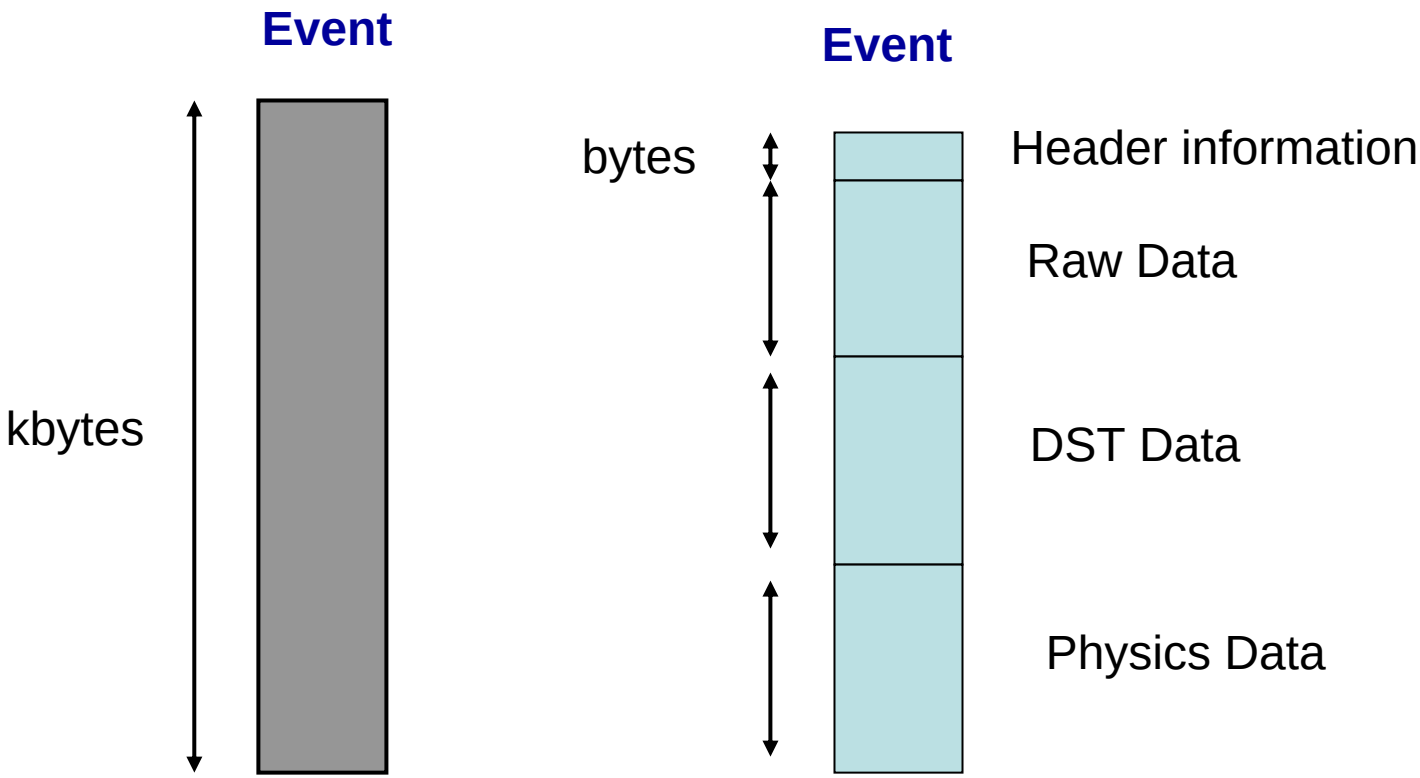
Goal: Reduce 10TB of data to fit on 1 800GB disk.
Make access to data transparent to the user.

Data I/O

Data i/o is a bottle neck, you want to be able To only load the data that you need into Memory.

Make decision on event type, Conf. Level, ….

Example: Tree structure with linked lists and branches



Data Compression

Data should be compressed as much as possible.

Integer words should be byte packed.

Some Floats can be truncated and byte-packed

Do not duplicate data.

Data should be accessed through function calls, rather than direct access.

Hide the data from the end user

Classes provides a very nice mechanism to do this.

The compression scheme can change as long as the data are tagged with a version.

Eliminates user mistakes in extracting data.

DON'T ADD OVERHEAD TO THE DATA STREAM.

MAKE IT AS LEAN AS POSSIBLE.

Put intelligence in access tools

Results

We have been able to reduced the data set from 12 TBytes to 600 GBytes With no loss of physics information. Each of the 10,000 files is now about 60MB in size.

Better compression occurred with the Monte Carlo. 1.7GB compressed to 8 Mbytes.

We have not looked at the CLAS raw data sets. While I expect some compression Possible, it is probably less than what has been achieved here.

2 Gbytes	2 Gbytes	100's Mb	5 Gbytes per "tape"
RDT	DST	Mini-Dst	
2 Gbytes	60 Mbytes	5 Mbs	2.065 Gbytes per "tape"

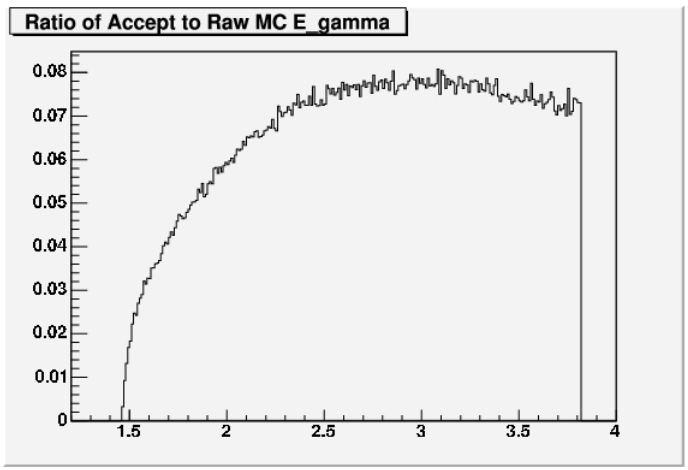
Exported to outside users

Monte Carlo Production

3.2 GHz Xeon Processor 32-bit kernel RHE3

100,000 events (4-vector input)		4MB	keep
i)	Convert to Geant input	30MB	flush
ii)	Run Geant	170MB	flush
iii)	Smear output data	170MB	flush
iv)	Reconstruct data	1700MB	flush
v)	Compress output file	8MB	keep

3-4 hours of CPU time 0.1 to 0.15 second/event

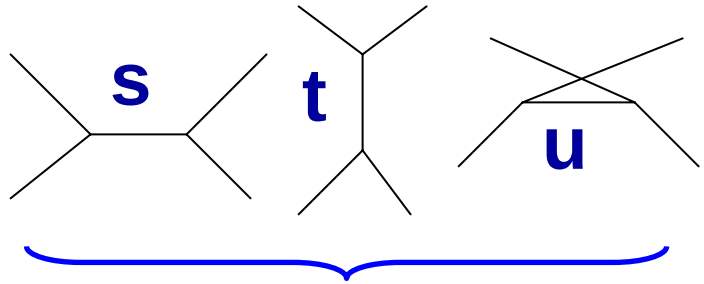


Typical CLAS Acceptance is about 10%

60,000,000 Raw η ‘
130,000,000 Raw η
250,000,000 Raw ω

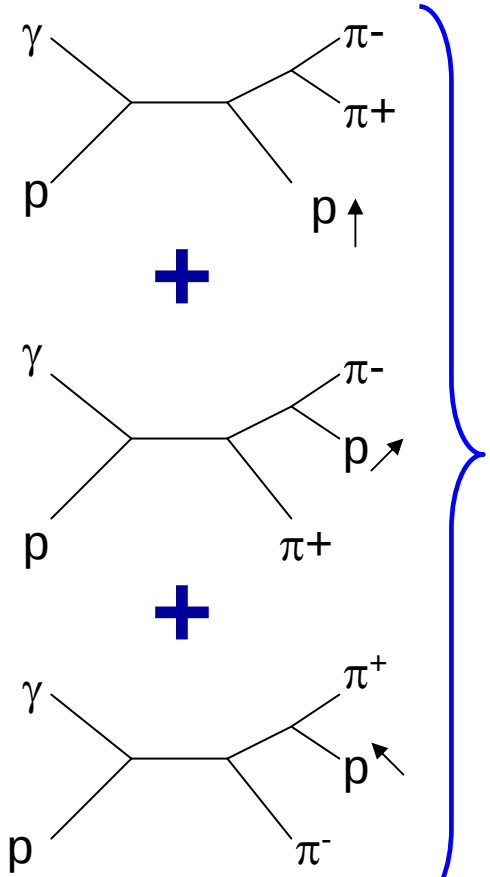
More intelligence in generation needed

Baryon Issues



All known to be important in Baryon photoproduction

$\gamma\, p \rightarrow p\, \pi^+\pi^-$



To coherently add the amplitudes, the protons need to be in the same reference frame for each diagram

Write all amplitudes as Lorentz scalars

"covariant tensor formalism"

Chung, Bugg, Klempt, …

Amplitude Tool

Created a tool that very efficiently evaluates amplitudes given four vectors as input.

Up to spin 9/2 in the s-channel

Up to L=5

Correctly adds all s,t and u diagrams

Input based on Feynman Diagrams, has been tested against know results and identities. To evaluate 100,000 events with spin 7/2 takes a couple of hours. This is VERY FAST.

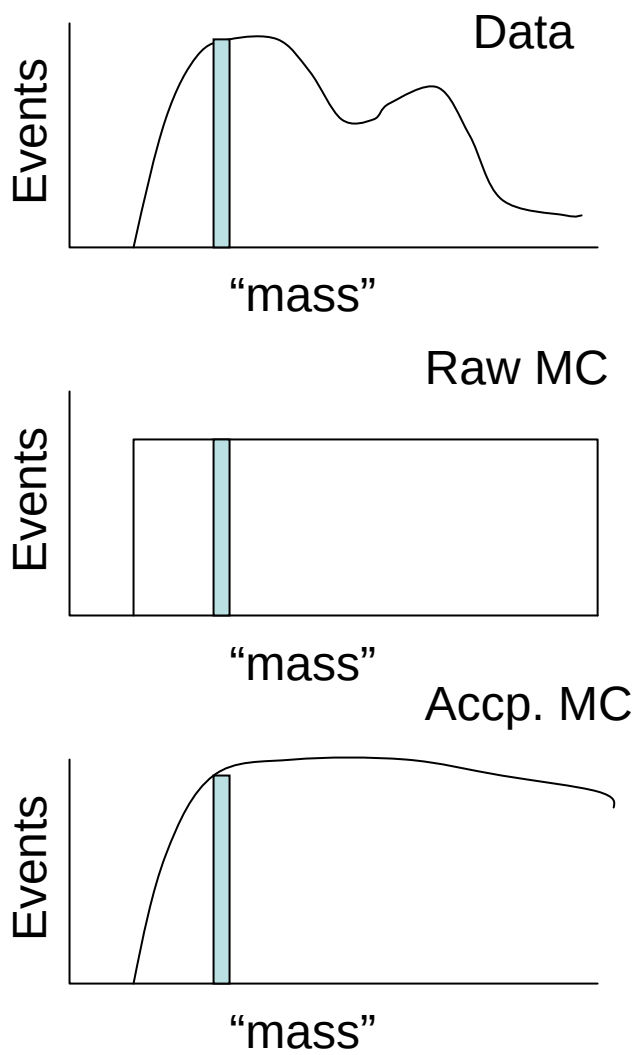
Production Amplitudes are written as electric and magnetic mulitopoles.

PWA Issues

20,000 Data
50,000 Accp. M.C.
100,000 Raw M.C.

The actual minimization is driven by Amplitudes time Data

1GB of memory/Dual Processor may be our limit.



Summary

Put effort into making the data size as small as possible.

Design Data to facilitate easy skimming of data

Hide the data from the user.

We think that the multi-channel PWA is doable now with few 100,000 to 10 million event PWA's

100,000 events per bin may cap what we can do (memory)