The June Software Review

David Lawrence, JLab

Feb. 16, 2012

(some history)

May 10, 2011 IT Review

An internal JLab review of IT readiness was done on May 10th, 2011. This was intended as a "warm up" for the review coming this summer.

Charge to the review panel:

We request that the review panel address the following points for IT in the 12 GeV era:

- *An assessment of the state of current software and systems developments An assessment of planning for bringing all software to a suitable level of maturity, including software testing for correctness and performance*
- *An assessment of planning for an evolution of computing, networking and storage capacity and performance to address the needs of detector simulation and data analysis*
- *An assessment of the IT infrastructure to meet requirements including support for other areas, e.g. accelerator, light source, theory, operations*
- *An assessment of the quality and effectiveness of the management of the major efforts to prepare*
- *As assessment of the resources, budget and staffing, to meet the needs of the program*

- one day review
- afternoon session focused on non-ENP* software
  - Management Information Systems
  - Networking and Infrastructure
  - Accelerator Controls
- Hall-D had one 25-minute talk given by Mark Ito.

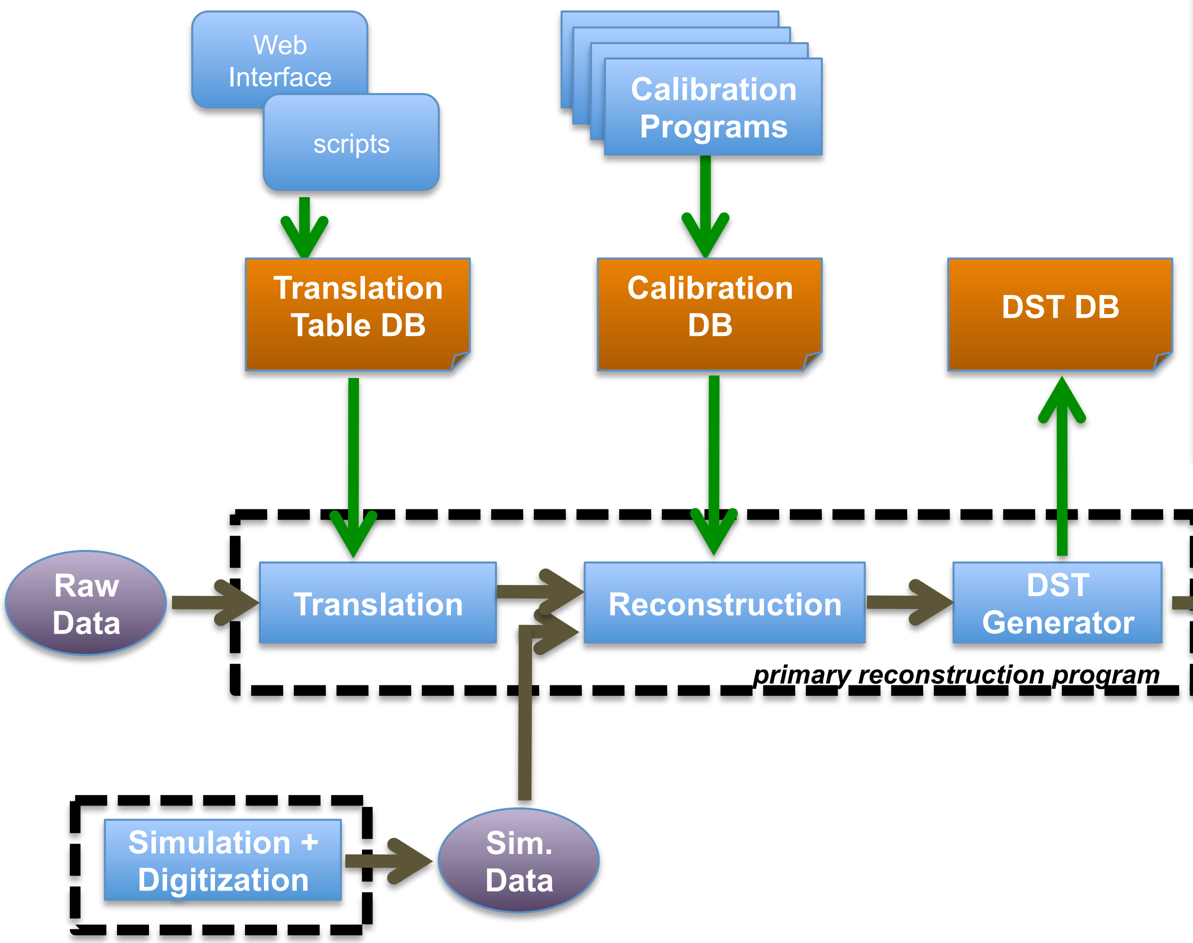
**ENP=Experimental Nuclear Physics*

From May IT Review closeout

… these items were specific to the experimental halls…

- Software: No common process for defining requirements, no common management structure
- 4 halls not sharing much software
- Hall D:
  - D's requirements not as well defined as other halls
  - Software head seems to have insufficient authority to direct software development priorities (i.e. software architect)
  - 2 FTE seems too small for 40% of effort planned for Jefferson Lab
  - Hall D Offsite computing & networking requirements nebulous
- Halls do not yet have robust plans for testing and reviewing readiness to operate.
- Identification of risks, and addressing risks, still needs to be done

Rough Diagram of GlueX Software



The software effort centers around *reconstruction* where a large fraction of the effort is spent.

Reconstruction requires inputs from databases to 1.) translate the geographic DAQ info into detector ids and 2.) convert digitized values into physical units.

Software is used to fill these databases with the bulk of that effort focused on the calibration programs.

Simulated data is required for the Amplitude Analysis. It does not require translation, but will be processed by the same reconstruction software as is used for the real data.

Configuration DB
Conditions DB

The configuration DB will hold information used to configure the online systems prior to data taking. The conditions DB will have values read from the online systems streamed into it during data taking.

For simplicity, not all connections are shown. (e.g. arrow from "Raw Data" to "Calibration Programs")

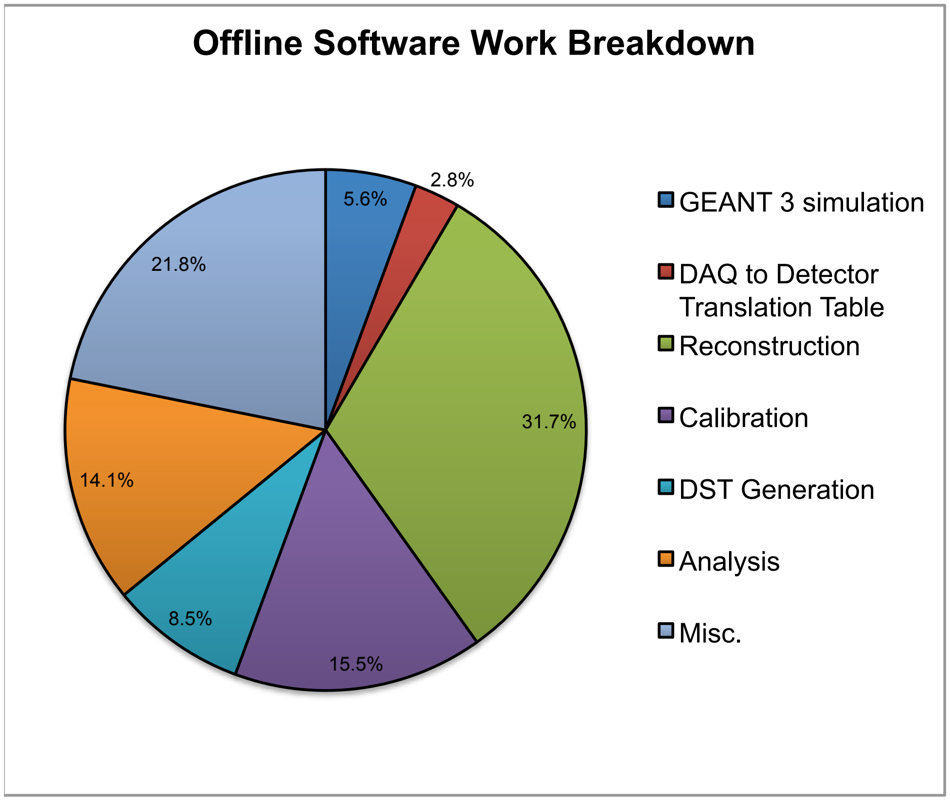
Hall-D Software Activity Schedule

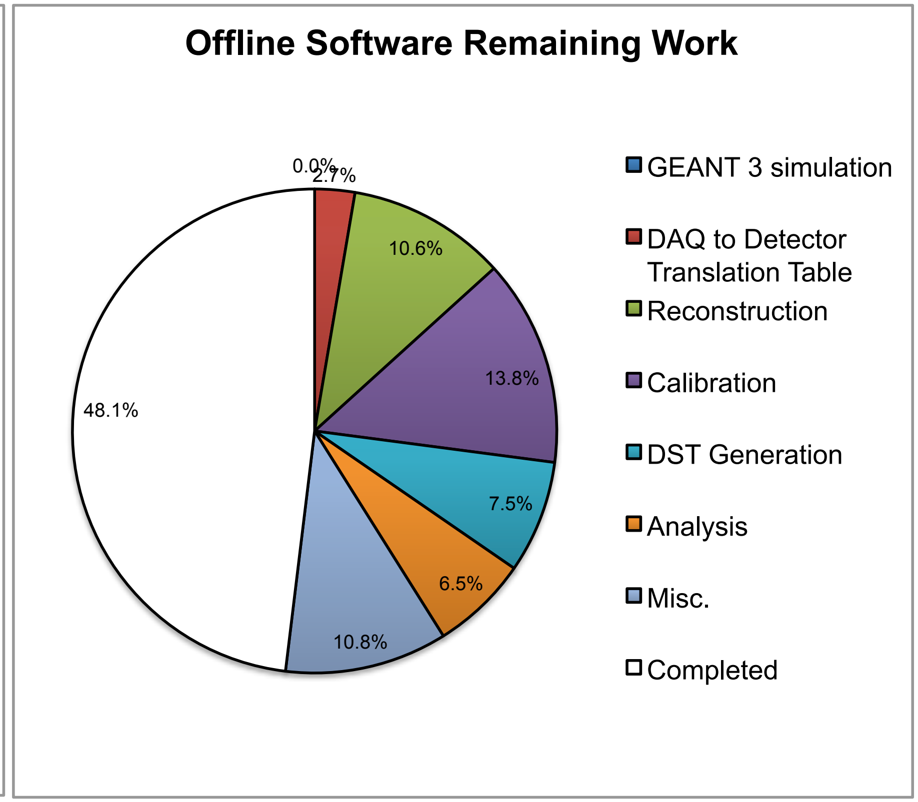
	Budgeted Labor Units (MW)	FTE-years	% complete	Responsible Persons	fraction of project
GEANT 3 simulation	88	2.0	100%	Richard Jones	5.6%
GEANT 4 simulation	88	2.0	0%		
DAQ to Detector Translation Table	44	1.0	5%	*JLab*	2.8%
Reconstruction	495	11.3	67%		31.7%
Reconstruction Framework	44	1.0	95%	David Lawrence	
CDC Reconstruction	33	0.8	85%	David Lawrence	
FDC Reconstruction	33	0.8	85%	Simon Taylor	
Track Finding	66	1.5	75%	Simon Taylor/David Lawrence	
Track Fitting	66	1.5	50%	Simon Taylor	
BCal Reconstruction	44	1.0	50%	Matt Shepherd/Zisis Papandreou	
FCal Reconstruction	33	0.8	75%	Matt Shepherd/Richard Jones	
TOF Reconstruction	33	0.8	50%	Paul Eugenio	
Tagger Reconstruction	33	0.8	0%		
Start Counter Reconstruction	22	0.5	50%	Simon Taylor/Werner Boeglin	
Particle ID	44	1.0	75%	Paul Mattione	
Kinematic Fitter	44	1.0	95%	Matt Shepherd	
Calibration	242	5.5	11%		15.5%
Calibration Database	33	0.8	80%	Dmitry Romanov	
CDC Calibration	33	0.8	0%	*CMU*	
FDC Calibration	33	0.8	0%	*Jlab*	
BCal Calibration	33	0.8	0%	*Univ. of Regina*	
FCal Calibration	33	0.8	0%	*IU*	
Tagger Calibration	33	0.8	0%	UConn/??	
Starter Counter Calibration	22	0.5	0%	*FIU*	
TOF Calibration	22	0.5	0%	*FSU*	
DST Generation	132	3.0	11%		8.5%
Data format	44	1.0	33%		
Micro DST Writer	22	0.5	0%		
Job Control Reconstruction	33	0.8	0%		
Job Control/Database for Simulation	33	0.8	0%		
Analysis	220	5.0	54%		14.1%
PWA Development	132	3.0	90%	Matt Shepherd/Ryan Mitchell	
PWA Challenge	44	1.0	0%		
Grid Implementation	44	1.0	0%	*UConn/??*	
Misc.	341	7.8	50%		21.8%
Event Viewer (adapted from online)	22	0.5	50%	David Lawrence	
Documentation	88	2.0	40%		
MC Studies for Detector Optimization	132	3.0	95%		
Integration of Slow Controls	33	0.8	0%	Elliott Wolin/Hovanes Egiyan	
Integration/QC	44	1.0	0%		
Coordination	22	0.5	0%	Mark Ito	
	Man-weeks	FTE-years			
Total	1562.0	35.5			100.0%

- Activity schedule adopted for BIA (Baseline Improvement Activity) schedule.
- Tracking of BIA stopped in 2009
- Minor tweaks including addition of a couple of lines (e.g. *Data Format* under *DST Generation*)
- % complete column added and based purely on my "engineering judgment"
- Responsible Persons column added

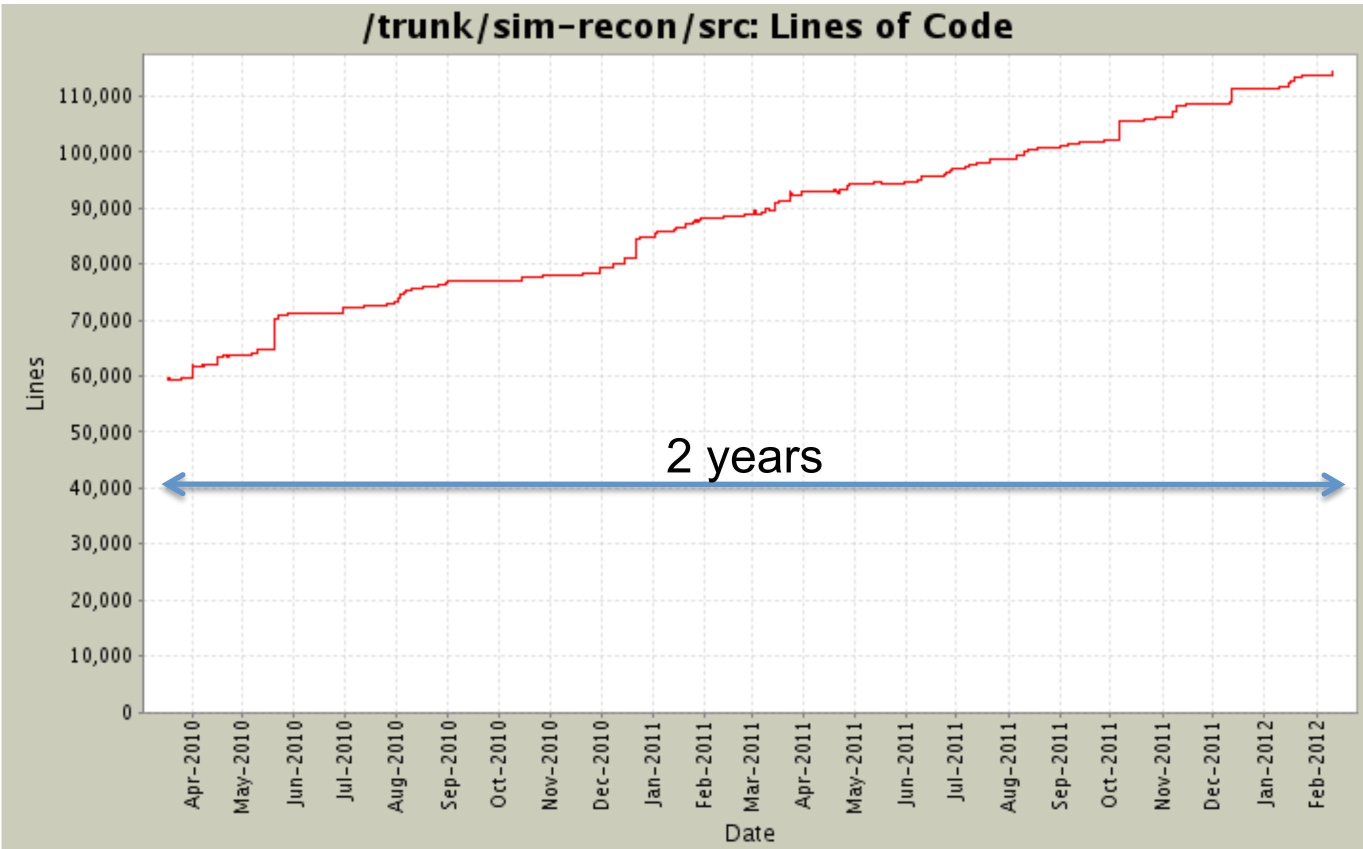
Overall status of Hall-D Software Activities

	Budgeted Labor Units (MW)	% of effort	% complete	% of total project remaining
GEANT 3 simulation	88	5.6%	100%	0.0%
DAQ to Detector Translation Table	44	2.8%	5%	2.7%
Reconstruction	495	31.7%	67%	10.6%
Calibration	242	15.5%	11%	13.8%
DST Generation	132	8.5%	11%	7.5%
Analysis	220	14.1%	54%	6.5%
Misc.	341	21.8%	50%	10.8%
Completed				48.1%





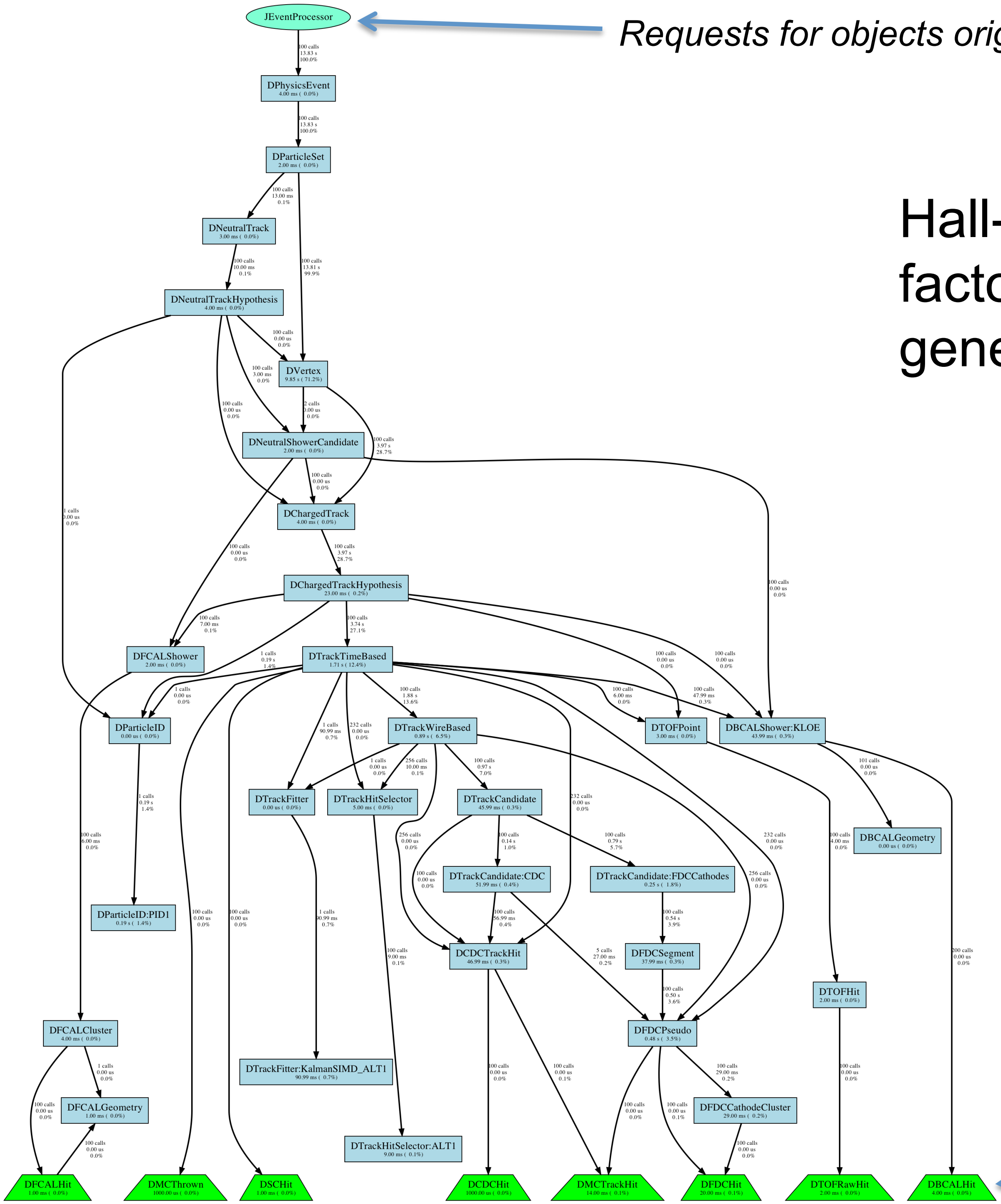
Hall-D Software Development



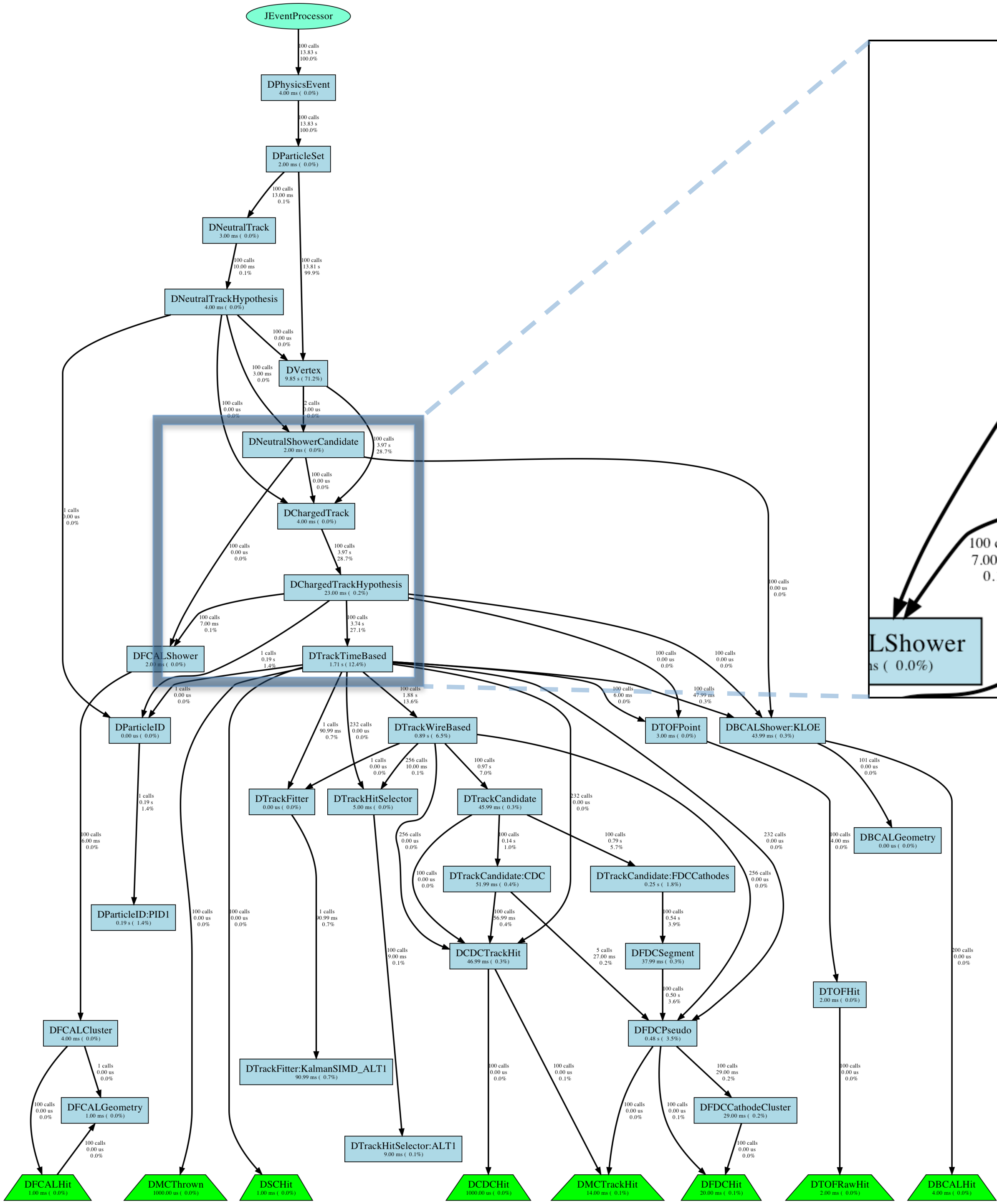
Steady development over the last few years. (Repository restructuring 2 years ago limits reach of this plot).

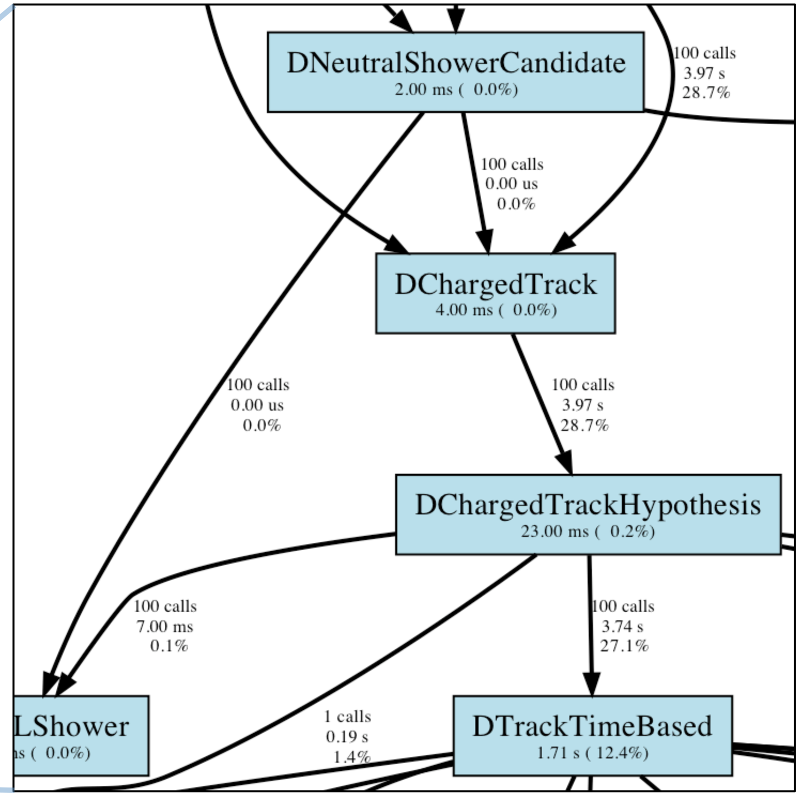
Hall-D reconstruction factory call graph auto-generated by JANA

Requests for objects originate here



Objects read from file enter here





Detailed profiling of entire reconstruction chain provided by JANA framework for diagnosing where time is spent.

Profiling at subroutine level can be done using standard C/C++ debugging utilities.

Estimated Resource Requirements

(from Mark's spreadsheet prepared for May 2011 IT review)

1	**GlueX Computing Model**			
2				
3	**parameter**	**value**	**units**	**comments**
4	event rate	20000	events/s	raw data rate out of the counting room when beam is on
5	annual running	35	weeks	amount of running in a year
6	running efficiency	0.5		fraction of wall time when beam is on, either due to beam unavailable or detector not ready
7	effective event rate	6707.734428	events/s	Event rate averaged over time
8	"	2.12E+11	events/year	Event rate averaged over time
9	CPU time per event	0.133333333	CPU-s/event	Time to reconstruct a single event on a 2.8 GHz Nehalem machine, per thread, from Simon's email of 1/31/2011
10	single Pass 1 CPU needed	894.3645905	CPU's	number of threads to keep up with the raw event rate
11	raw event size	15000	bytes	size of a single raw event
12	raw instantaneous storage rate	300	MB/s	data rate when beam is on
13	raw effective storage rate	3.2E+15	bytes/year	average data volume rate
14	raw effective storage rate	3173.026694	TB/year	average data volume rate
15	pass 0 event fraction	0.05		fraction of events from raw data stream to perform calibrations
16	pass 1 repetition factor	2		number of times event reconstruction will be repeated
17	pass 0 repetition factor	2		number of times calibration will be repeated
18	pass 0 CPU need	89.43645905	CPU's	number of threads of calibration to keep up
19	pass 1 CPU need	1788.729181	CPU's	number of threads of reconstruction to keep up
20	stream/pass-1 CPU ratio	0.1		ratio of CPU time required for a skim stream to that needed for reconstruction
21	stream output to input size ratio	0.1		ratio of data volume output for a stream to that of input
22	stream multiplicity factor	5		number of streams to be produced
23	single stream CPU need	89.43645905	CPU's	number of threads for one stream to keep up
24	stream repetition factor	2		number of times streaming will be repeated
25	stream CPU need	894.3645905	CPU's	number of threads for streaming to keep up
26	single stream output data rate	2.012320329	MB/s	
27	total stream output data rate	0.63504	PB/year	
28	MC CPU ratio per event, generation	0.5	CPU-s/event	ratio of CPU time required for generating a Monte Carlo event to that needed for reconstruction
29	MC CPU ratio per event, reconstruction	1		ratio of CPU time to reconstruct Monte Carlo events to that to reconstruct real data
30	MC/raw data event rate ratio	2		ratio of number of Monte Carlo events needed to number of raw data events
31	MC event size	15000	bytes	size of a single generated Monte Carlo event
32	MC multiplicity factor	2		number of times MC data will need to be generated
33	MC effective event rate	26830.93771	events/s	event rate averaged over time of MC event generation
34	MC CPU need	5366.187543	CPU's	numbers of threads needed for generating Monte Carlo
35	MC pass 1 output event size	3000	bytes	size of a single reconstructed Monte Carlo event
36	MC effective data rate	80.49281314	MB/s	
37	MC effective data rate	2.54016	PB/year	
38	analysis/pass-1 CPU ratio	0.1		ratio of CPU time required for performing a physics analysis to that needed for reconstruction
39	analysis multiplicity factor	10		number of analyses to be conducted
40	analysis CPU need	894.3645905	CPU's	number of threads needed for analysis
41	total CPU need	9033.082364	CPU's	total number of threads needed for all activities
42	total CPU need exclusive of MC	3666.894821	CPU's	total number of threads needed for all activities
43	data rate, tape to cache disk	100	MB/s	average rate from tape library to cache disk
44	data rate, cache disk to local disk	3	MB/s	average rate from cache disk to local farm node disk
45	raw data recording tape need	1.006160164		
46	Pass 1 output to input size ratio	0.2		ratio of output event size to input event size
47	pass1 processed event size	3000	bytes	reconstructed event size
48	Single pass 1 output data rate	20.12320329	MB/s	data rate for a single pass 1 output stream
49	total pass 1 output data rate	1.27008	PB/year	data rate for all pass 1 output streams
50	Single pass 0 output data rate	1.006160164	MB/s	data rate for a single pass 1 output stream
51	total pass 0 output data rate	0.063504	PB/year	data rate for all pass 0 output streams
52	single pass 1 tape need	1.207392197	drives	number of tape drives needed to support pass 1, one iteration
53	Pass 1 tape need	2.414784394	drives	number of tape drives needed to support pass 1, all iterations
54	single pass 0 tape need	0.06036961	drives	number of tape drives needed to support pass 0, one iteration
55	Pass 0 tape need	0.12073922	drives	number of tape drives needed to support pass 0, all iterations
56	single stream input tape need	2.012320329	drives	number of tape drives needed to support input for streaming, one iteration
57	single set of stream output tape need (all str	1.006160164	drives	number of tape drives needed to support output for streaming, one iteration
58	total stream tape need	6.036960986	drives	number of tape drives needed to support streaming, all iterations
59	MC tape drive need	0.804928131	drives	number of tape drives needed to archive reconstructed MC data
60	total tape drive need	10.3835729		total number of tape drives needed for all activities
61	disk usage per analysis	20	TB	permanent disk space used by an analysis
62	disk usage total	200	TB	permanent disk space used by all analyses
63	total output rate	7.681810694		

Estimated Resource Requirements

(from Mark's spreadsheet prepared for May 2011 IT review)

1	**GlueX Computing Model**			
2				
3	**parameter**	**value**	**units**	**comments**
4	event rate	20000	events/s	raw data rate out of the counting room when beam is on
5	annual running	35	weeks	amount of running in a year
6	running efficiency	0.5		fraction of wall time when beam is on, either due to beam unavailable or detector not ready
7	effective event rate	6707.734428	events/s	Event rate averaged over time
8	"	2.12E+11	events/year	Event rate averaged over time
9	CPU time per event	0.133333333	CPU-s/event	Time to reconstruct a single event on a 2.8 GHz Nehalem machine, per thread, from Simon's email of 1/31/2011
10	single Pass 1 CPU needed	894.3645905	CPU's	number of threads to keep up with the raw event rate
11	raw event size	15000	bytes	size of a single raw event
12	raw instantaneous storage rate	300	MB/s	data rate when beam is on
13	raw effective storage rate	3.2E+15	bytes/year	average data volume rate
14	raw effective storage rate	3173.026694	TB/year	average data volume rate
15	pass 0 event fraction	0.05		fraction of events from raw data stream to perform calibrations
16	pass 1 repetition factor	2		number of times event reconstruction will be repeated
17	pass 0 repetition factor	2		number of times calibration will be repeated
18	pass 0 CPU need	89.43645905	CPU's	number of threads of calibration to keep up
19	pass 1 CPU need	1788.729181	CPU's	number of threads of reconstruction to keep up
20	stream/pass-1 CPU ratio	0.1		ratio of CPU time required for a skim stream to that needed for reconstruction
21	stream output to input size ratio	0.1		ratio of data volume output for a stream to that of input
22	stream multiplicity factor	5		number of streams to be produced
23	single stream CPU need	89.43645905	CPU's	number of threads for one stream to keep up
24	stream repetition factor	2		number of times streaming will be repeated
25	stream CPU need	894.3645905	CPU's	number of threads for streaming to keep up
26	single stream output data rate	2.012320329	MB/s	
27	total stream output data rate	0.63504	PB/year	
28	MC CPU ratio per event, generation	0.5	CPU-s/event	ratio of CPU time required for generating a Monte Carlo event to that needed for reconstruction
29	MC CPU ratio per event, reconstruction	1		ratio of CPU time to reconstruct Monte Carlo events to that to reconstruct real data
30	MC/raw data event rate ratio	2		ratio of number of Monte Carlo events needed to number of raw data events
31	MC event size	15000	bytes	size of a single generated Monte Carlo event
32	MC multiplicity factor	2		number of times MC data will need to be generated
33	MC effective event rate	26830.93771	events/s	event rate averaged over time of MC event generation
34	MC CPU need	5366.187543	CPU's	numbers of threads needed for generating Monte Carlo
35	MC pass 1 output event size	3000	bytes	size of a single reconstructed Monte Carlo event
36	MC effective data rate	80.49281314	MB/s	
37	MC effective data rate	2.54016	PB/year	
38	analysis/pass-1 CPU ratio	0.1		ratio of CPU time required for performing a physics analysis to that needed for reconstruction
39	analysis multiplicity factor	10		number of analyses to be conducted
40	analysis CPU need	894.3645905	CPU's	number of threads needed for analysis
41	total CPU need	9033.082364	CPU's	total number of threads needed for all activities
42	total CPU need exclusive of MC	3666.894821	CPU's	total number of threads needed for all activities
43	data rate, tape to cache disk	100	MB/s	average rate from tape library to cache disk
44	data rate, cache disk to local disk	3	MB/s	average rate from cache disk to local farm node disk
45	raw data recording tape need	1.006160164		
46	Pass 1 output to input size ratio	0.2		ratio of output event size to input event size
47	pass1 processed event size	3000	bytes	reconstructed event size
48	Single pass 1 output data rate	20.12320329	MB/s	data rate for a single pass 1 output stream
49	total pass 1 output data rate	1.27008	PB/year	data rate for all pass 1 output streams
50	Single pass 0 output data rate	1.006160164	MB/s	data rate for a single pass 1 output stream
51	total pass 0 output data rate	0.063504	PB/year	data rate for all pass 0 output streams
52	single pass 1 tape need	1.207392197	drives	number of tape drives needed to support pass 1, one iteration
53	Pass 1 tape need	2.414784394	drives	number of tape drives needed to support pass 1, all iterations
54	single pass 0 tape need	0.06036961	drives	number of tape drives needed to support pass 0, one iteration
55	Pass 0 tape need	0.12073922	drives	number of tape drives needed to support pass 0, all iterations
56	single stream input tape need	2.012320329	drives	number of tape drives needed to support input for streaming, one iteration
57	single set of stream output tape need (all str	1.006160164	drives	number of tape drives needed to support output for streaming, one iteration
58	total stream tape need	6.036960986	drives	number of tape drives needed to support streaming, all iterations
59	MC tape drive need	0.804928131	drives	number of tape drives needed to archive reconstructed MC data
60	total tape drive need	10.3835729		total number of tape drives needed for all activities
61	disk usage per analysis	20	TB	permanent disk space used by an analysis
62	disk usage total	200	TB	permanent disk space used by all analyses
63	total output rate	7.681810694		

About 115 computers with 32 cores each will be needed just to keep up with time-averaged data acquisition + calibration

Another 170 will be needed for simulation (+recon.)

Software Sharing Among Halls

- Meeting was held on Jan. 26th to discuss areas where halls could share software, minimizing duplication of effort.

- All halls were represented

- Rolf Ent asked halls to get together to discuss specific topics and explore sharing opportunities

- Two items were given to Halls B and D to discuss (a few others for all halls):
  - Tracking Algorithms
    - Multiple, organized discussions have taken place between primaries
    - Hall-B has read-access to our repository and is using it as a reference as they develop their own tracking package
  - CLARA and JANA
    - (see next two slides)

Primary Differences between JANA and CLARA

CLARA	JANA
• "Loosely Coupled":	• "Tightly Coupled":
• Allows multiple languages to be combined since each module is a separate process	• Single language, all modules contained within a single process
• Data passed between modules by value	• Data passed between modules by reference
• Built-in ability to distribute reconstruction job over multiple computers (cloud)	• Utilizes external distributed computing mechanisms like the GRID and Auger

CLARA is designed to provide interactive access to a system of services hosted either on a single node or distributed over a cloud

JANA is designed to make maximal use of a local, multi-core resource

Functionality common to both JANA and CLARA

- Framework for event reconstruction
  - Modular:
    - allow easy replacement of one or more algorithms
    - allow independent development of modules by separate groups
  - Provides mechanism to parallelize reconstruction using multiple cores on the same computer
  - Plugin mechanism to allow extension of existing functionality at run time

How JANA and CLARA might used in conjunction

JANA could be used to implement CLARA services that need to be highly efficient.

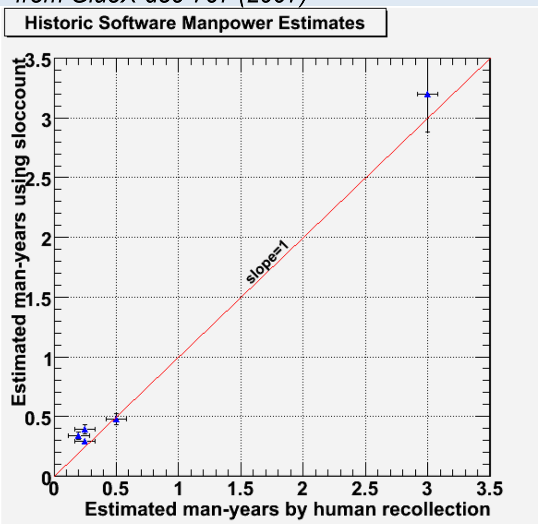
CLARA could be used to deploy JANA applications as shared services in a network distributed cloud computing environment.

The primary benefit to CLAS12 users of integrating JANA-based components into a CLARA-based system could be overall faster reconstruction for a fixed set of resources.

The primary benefit to Hall-D users of wrapping JANA-based programs as CLARA services would be gaining an interactive distributed computing environment that could provide a faster simulation/analysis cycle for specific studies.

Manpower

from GlueX-doc-767 (2007)



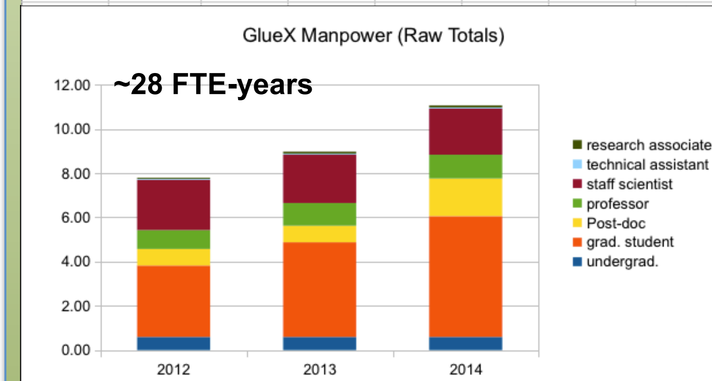
- Use standard COCOMO model to estimate man-years put into CLAS offline software
- Estimate was ~53 man-years for core CLAS offline
- GlueX was estimated to need **~40** man-years to be ready for start of operations

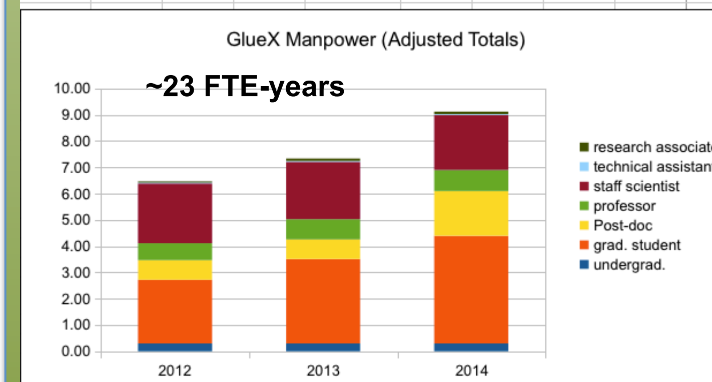
Raw Totals

	undergrad.	grad. student	Post-doc	professor	staff scientist	technical assistant	research associate
2012	0.59	3.23	0.75	0.86	2.27	0.05	0.05
2013	0.59	4.28	0.75	1.03	2.18	0.05	0.09
2014	0.59	5.46	1.70	1.07	2.09	0.05	0.09

Adjusted Totals

	undergrad.	grad. student	Post-doc	professor	staff scientist	technical assistant	research associate
2012	0.30	2.42	0.75	0.64	2.27	0.03	0.05
2013	0.30	3.21	0.75	0.77	2.18	0.03	0.09
2014	0.30	4.10	1.70	0.80	2.09	0.03	0.09
efficiency factor	0.5	0.75	1	0.75	1	0.75	1





- It is estimated that we will need approx. 40 FTE-years of offline software effort total for GlueX*
- Estimate is that we have done ~ 50% of work for offline software*
- Remaining 20 FTE-years of estimated work is well matched with manpower commitments from collaboration*

**every one of these estimates could be completely wrong*

Summary

- Software review is scheduled for early June 2012.
- Focus will be on having offline software development on track to be ready for analysis by the start of data taking
- Integrated GlueX manpower seems to be well-matched with what is needed to meet this goal

Backup Slides

outline

- Software Review details (charge, scope, …)
  - May review charge
  - May review recommendations
- Mark’s spreadsheet numbers for resources needed
- Existing software
  - LOC vs. time plot
  - janadot call graph
- BIA schedule
  - Rough diagram
  - Activity list
  - Pie charts
- Brainstorming session on collaborative efforts
  - Results of Tracking discussion
  - Results of JANA/CLARA discussion (3 slides)
- Manpower