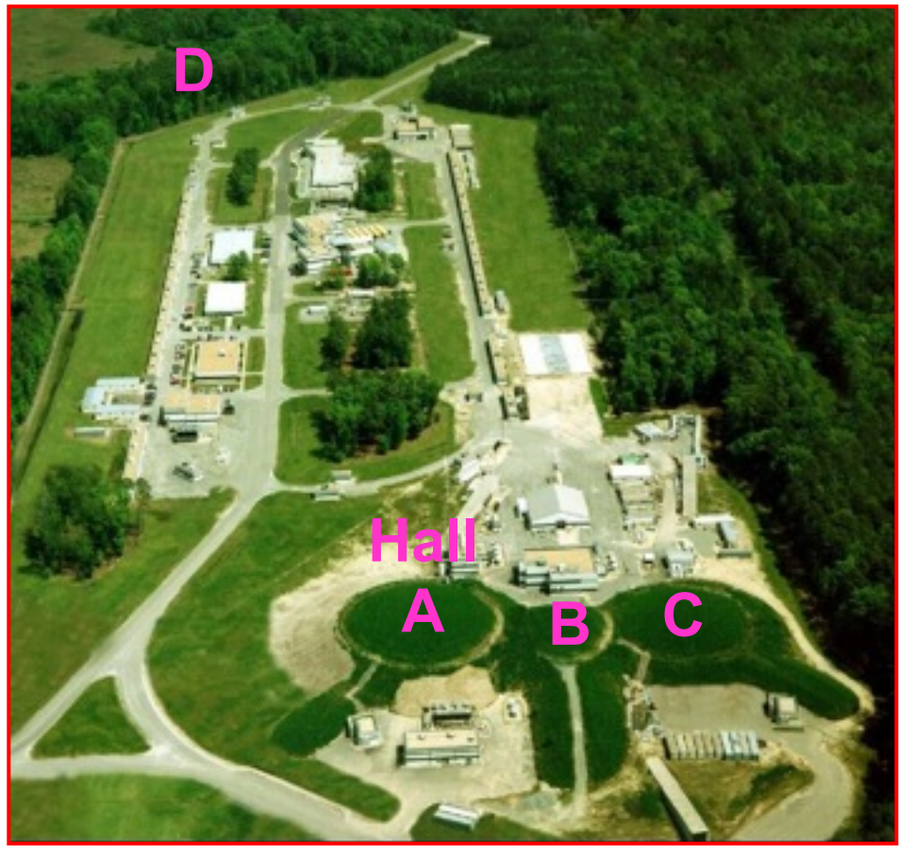
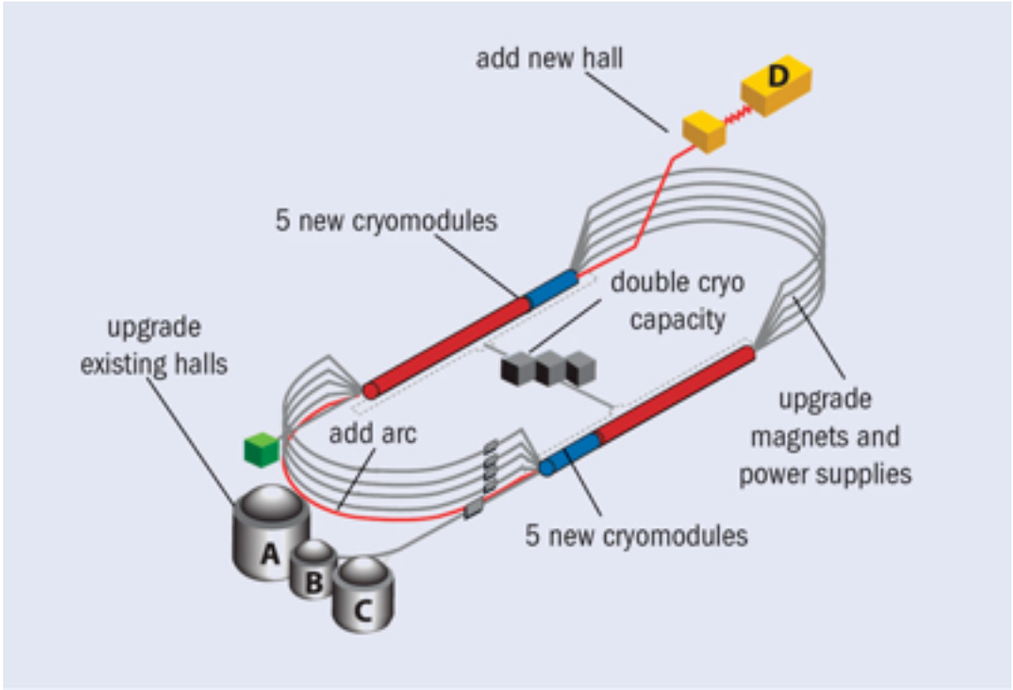
Experimental Nuclear Physics computing at Jefferson Lab

Graham Heyes - ENP division JLab

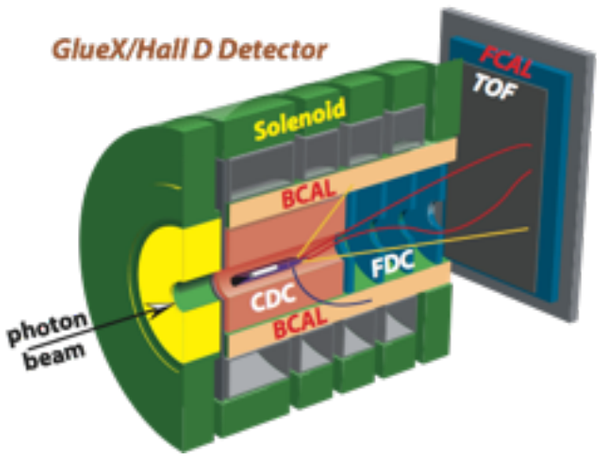
Jefferson Lab

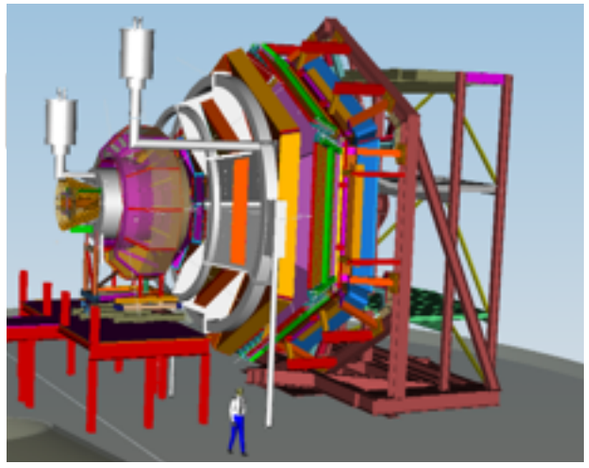
- Nuclear physics research using the CEBAF electron accelerator.
  - Two superconducting LINACs with recirculating arcs.
  - Simultaneous beam to multiple halls.
  - High intensity.
  - Polarized beam.
- Accelerator operated for about 20 years with up to 6 GeV beam energy.
- Shutdown in fall 2012 for upgrade:
  - Increase beam energy to 12 GeV.
  - Upgrade existing detectors.
  - Add a fourth end station.
- First beam after upgrade was a few weeks ago.

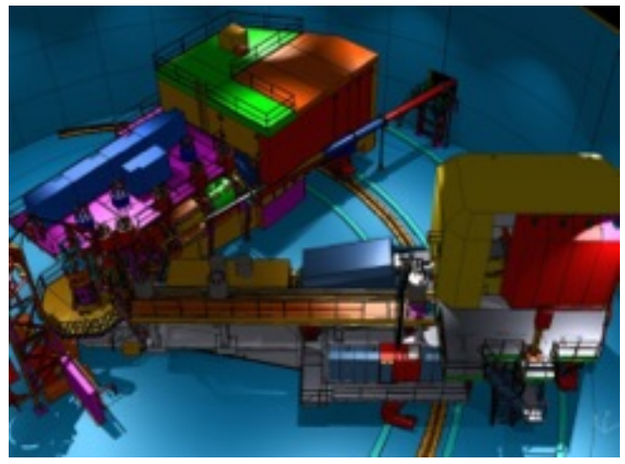


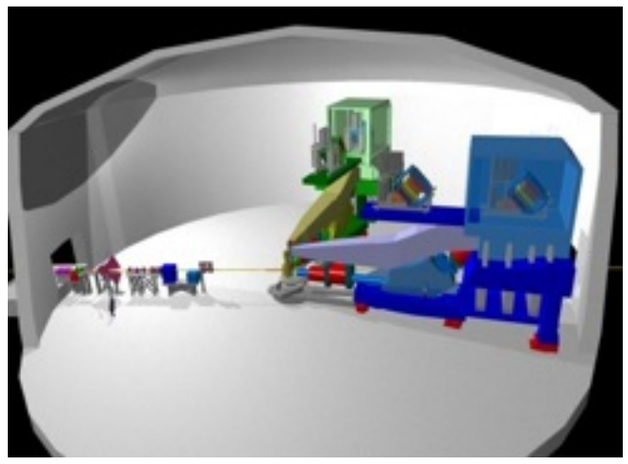


Upgraded Detectors









Hall D	Hall B	Hall C	Hall A
excellent hermeticity	luminosity 10	energy reach	custom installations
polarized photons	hermeticity	precision	
E_γ	11 GeV beamline		
10	target flexibility		
good momentum/angle resolution		excellent momentum resolution	
high multiplicity reconstruction		luminosity up to 10	
particle ID			

12 GeV Approved Experiments by Physics Topics

Topic	Hall A	Hall B	Hall C	Hall D	Other	Total
The Hadron spectra as probes of QCD (GluEx and heavy baryon and meson spectroscopy)		1		2		3
The transverse structure of the hadrons (Elastic and transition Form Factors)	4	3	2	1		10
The longitudinal structure of the hadrons (Unpolarized and polarized parton distribution functions)	2	2	6			10
The 3D structure of the hadrons (Generalized Parton Distributions and Transverse Momentum Distributions)	5	10	4			19
Hadrons and cold nuclear matter (Medium modification of the nucleons, quark hadronization, N-N correlations, hypernuclear spectroscopy, few-body experiments)	4	2	6		1	13
Low-energy tests of the Standard Model and Fundamental Symmetries	2			1	1	4
Total	**17**	**18**	**18**	**4**	**2**	**59**

12 GeV Approved Exp. by PAC Days

Topic	Hall A	Hall B	Hall C	Hall D	Other	Total
The Hadron spectra as probes of QCD (GluEx and heavy baryon and meson spectroscopy)		119		320		439
The transverse structure of the hadrons (Elastic and transition Form Factors)	144	85	102	25		356
The longitudinal structure of the hadrons (Unpolarized and polarized parton distribution functions)	65	120	165			350
The 3D structure of the hadrons (Generalized Parton Distributions and Transverse Momentum Distributions)	409	982	161			1552
Hadrons and cold nuclear matter (Medium modification of the nucleons, quark hadronization, N-N correlations, hypernuclear spectroscopy, few-body experiments)	159	120	179		14	472
Low-energy tests of the Standard Model and Fundamental Symmetries	513			79	60	652
Total	**1290**	**1426**	**607**	**424**	**74**	**3821**

More than 7 years of approved experiments

Computing

- Computing infrastructure for ENP is funded by the ENP division as part of the OPS budget but installed and managed by the scientific computing (SCI) group in IT. Networking and other support is provided by the computing and network infrastructure (CNI) group in IT.
- The use of the ENP computing infrastructure is coordinated on a per hall basis by computing coordinators.
- For historical reasons I am the overall computing coordinator for ENP:
  - An intermediary between ENP and IT to arbitrate, advocate and coordinate on behalf of the computing infrastructure users.
  - Gathering computing requirements and working with IT to fulfill them.
    - Technical review of plans from the SCI group - annual work plan process.
    - Managing the ENP computing budget.
  - Point of contact for computing related topics:
    - Data management plans, cyber security assessments, software quality assurance, IT steering committee, ENP rep on internal reviews.

Data processing

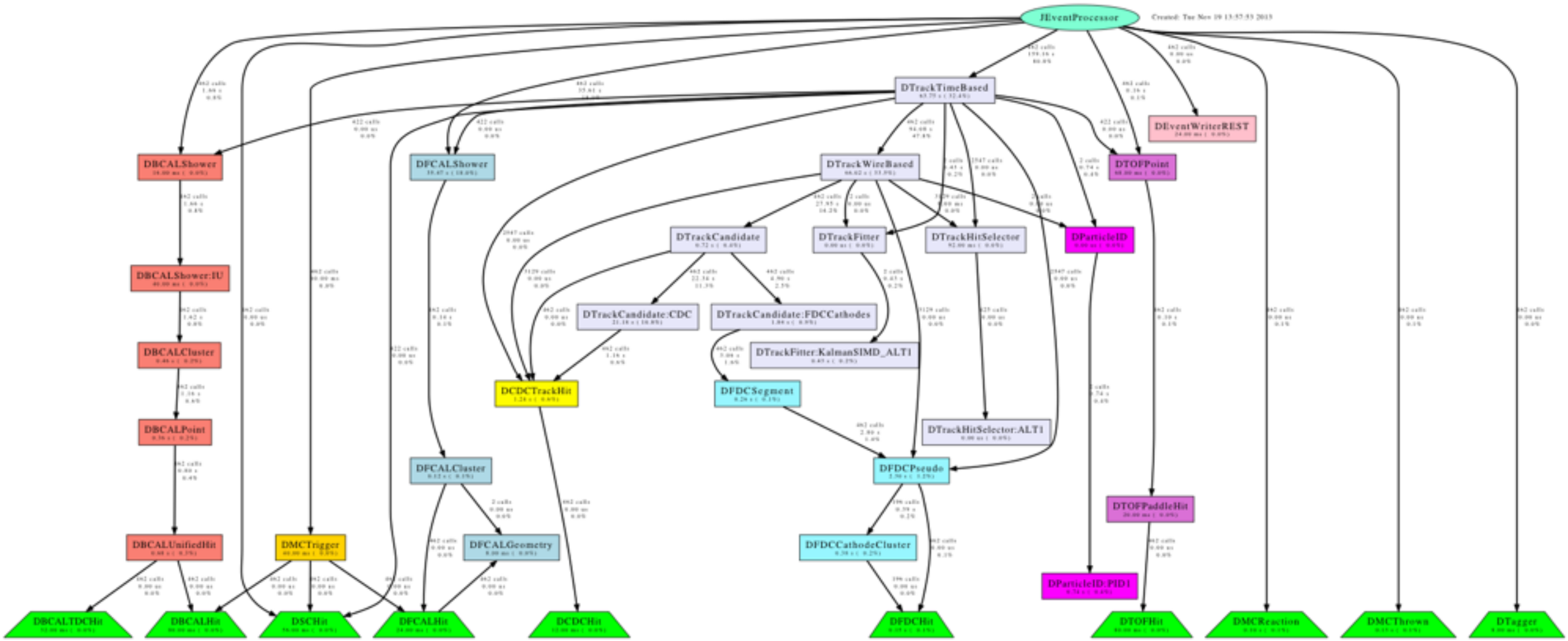
- There are four types of computation associated with ENP.
  - Simulation - Modeling of physics using GEANT simulation package.
  - Reconstruction - Processing of the raw data from the detector to convert patterns of hits etc into particle tracks, identities, momentum vectors etc.
  - Calibration - Processing of a sample of raw or reconstructed data to calibrate detector parameters.
  - Analysis - Processing of output of reconstruction using ROOT and similar packages.
- By far the largest computational workloads are reconstruction and simulation.
  - How much simulation is required depends on the physics.
  - The output from the simulation must also be fed through reconstruction

Reconstruction Frameworks

- To prevent duplication of effort the groups develop a reconstruction framework that is configured to suit the detector configuration and physics being studied.
  - Since halls A and C have similar spectrometers, low rates and, particularly in A, a rapid turnover of experiments they have a common framework with a strong focus on usability.
  - GLUEX in hall-D have developed OO multi-threaded framework, JANA.
    - Traditional monolithic batch job.
    - OO factory model: reconstructed objects on demand multi-threaded, event-level parallelization
    - Linear scaling observed up to 32 cores, memory cost per thread 30% that of single-threaded job.
    - integrated access to calibration constants in relational database.
  - CLAS12 in hall-B have a service oriented data driven architecture, CLARA.
    - Distributed array of loosely coupled services.
    - Services are chained together at run time.
    - Services communicate locally via memory or remotely via network.
    - Doesn't need a batch system.
    - Adapts well to cloud computing model.

JANA

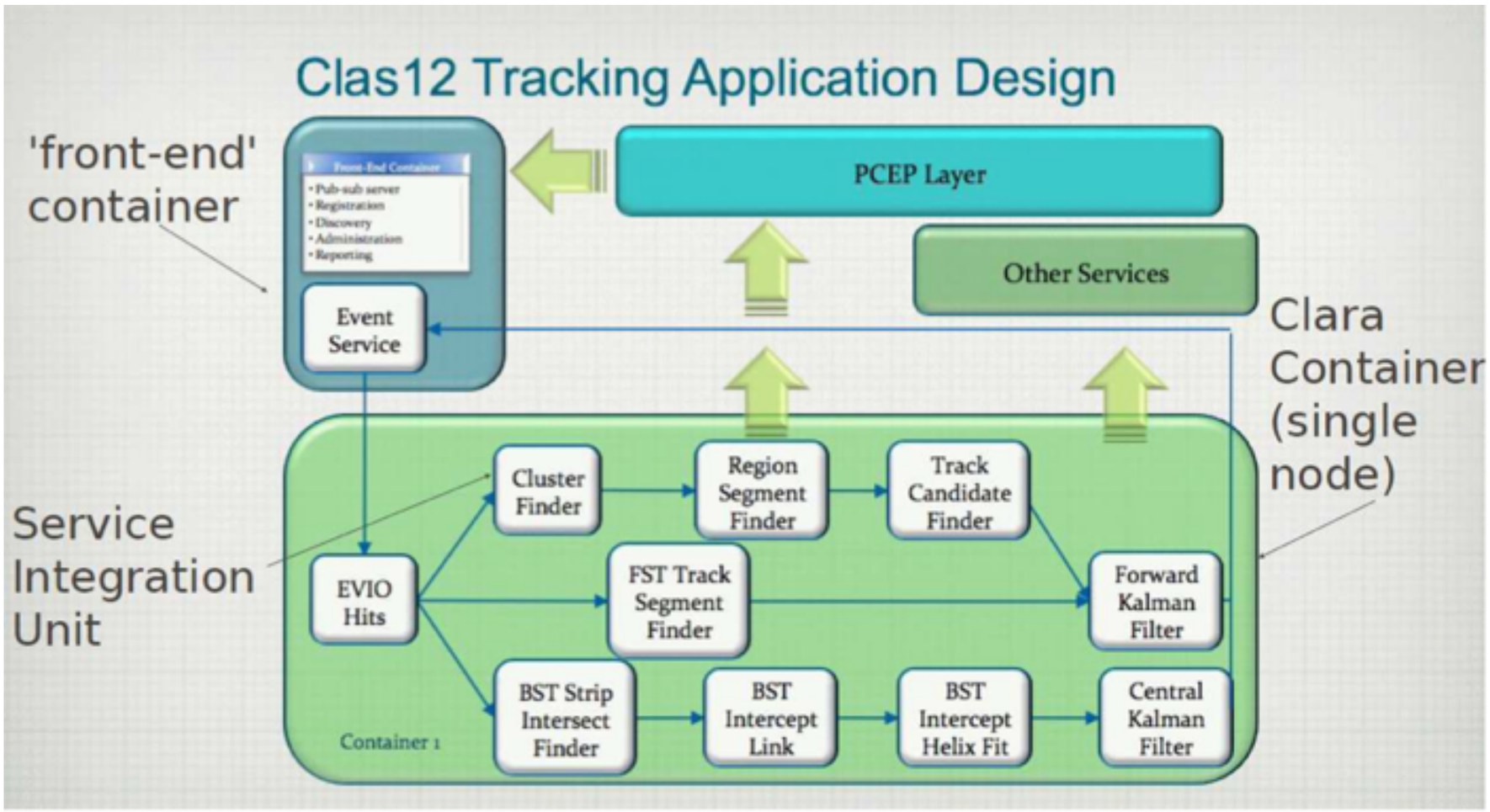
- JANA builds a top down tree of objects, for example:
  - Showers are identified from hit clusters.
  - Hit clusters are groups of points.
  - Points are groups of unified hits.
  - Unified hits are hits from different detectors.



This picture automatically generated by JANA framework

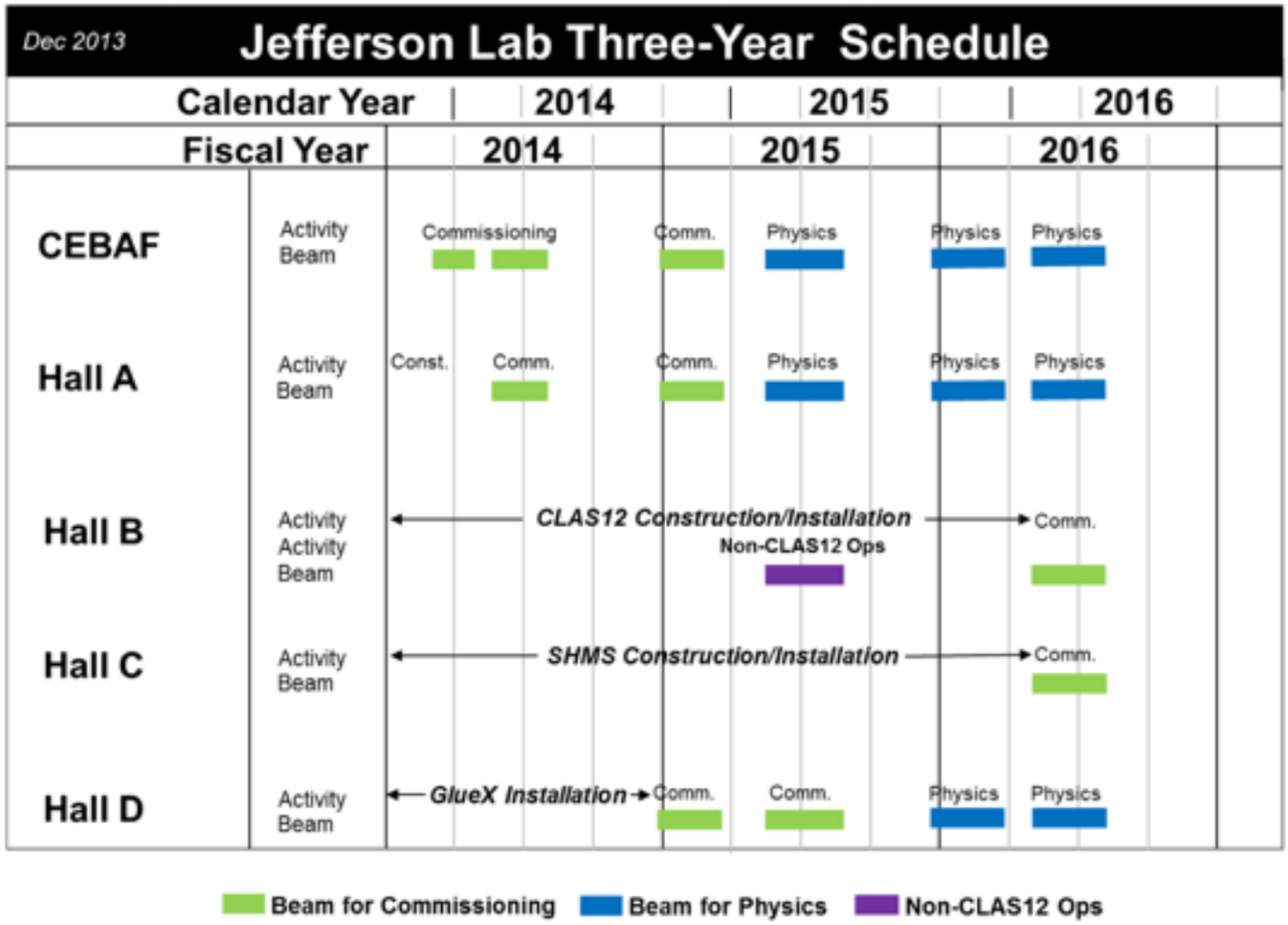
CLARA

- CLARA runs a Java "container" on each node. The container assembles a network of "service integration units" each containing a service. Data flows through the network and is processed at each stage. For example:
  - Cluster finder reads hits and outputs clusters.
  - Segment finder reads clusters and outputs segments of tracks.
  - Track candidate finder reads track segments and outputs track candidates.
  - Filter reads track candidates and outputs the most likely track.



Goals for 2017

- By 2017 halls A and D will have been taking production data for a year. Halls C and B are scheduled to end commissioning in 2016.
- Computational goal for 2017 is to be prepared to process data from all four halls.
- Scientific goals are:
  - To follow the scientific program with several “A” rated experiments in the first two years.
  - To understand the operation and limits of the upgraded detectors.



Computing requirements

- The computing requirements for each hall are owned by the offline working groups and are driven by the scheduled experiments, the capabilities of the detectors, the data analysis workflow adopted by each hall and their analysis frameworks.
- There have been two software and requirements reviews, June 2012 and Nov. 2013.
  - The Hall working groups have been working on benchmarks and data challenges that have allowed them to refine the input parameters to the computing requirements calculations.
  - Lab management is more aware of the budgetary requirements.
- While halls A and C are not completely negligible halls B and D dominate the computing requirements. Three main factors
  - Lower trigger rates and/or event sizes, sometimes by factor of 10x.
  - Factor of 10x less time per event to process.
  - Low volume of simulated data, often not using JLab ENP cluster.

Hall-D, GLUEX requirements

- GLUEX ramps up their data taking and analysis in several phases starting with commissioning and ending in 2016 with a steady state
- Base the calculations the final state and scale rates and days running to get other phases.
- Assumptions
  - Steady state raw event rate of 20 kHz.
  - 35 weeks of running per year with 50% efficiency = 2E+11 events / year.
  - Pass 0 (calibration) on 5% of the data repeated twice.
  - Pass 1 (reconstruction) on 100% of the data repeated twice.
  - Analysis 10x faster per event than reconstruction.
  - Analysis output 10x smaller than input.
  - 2 simulated events per raw event, must be generated + reconstructed. MC rate ~1/4 of reconstruction rate.

GLUEX continued

- Since we had a new cluster of dual 8-core 2.0 GHz Intel Xeon E5-2650 (Sandy Bridge) delivered in 2013 we used that as our testing baseline assume one thread per physical core.
- Measured reconstruction rate 7.5 events/s/thread. Ratios between reconstruction, simulation, calibration etc are known.
- Number of simultaneous threads for steady state are:
  - simulation + reconstruction of simulation ~ 5000
  - Everything else ~5000
- Total ~ 10000 simultaneous threads on the reference 2.0 GHz Xeon E5-2650 cores.

- In reality what we buy in future will have a different clock speed and probably more cores per CPU. We fold that in to the funding model by estimating the cost to provide the equivalent of a 2.0 GHz Xeon E5-2650 core.

CLAS12 model

- Assumptions, similar to GLUEX except :
  - Raw event rate 10 kHz.
  - Raw event size 10 kByte.
- Use same standard 2013 model reference CPU.
  - Several tasks that scale with data volume
    - Calibration 5% of data = 200 simultaneous threads.
    - Reconstruction = 1400
    - Validation = 1400
    - Analysis = 400
    - Simulation = 8000 - larger ratio of simulated to raw events ~10x
  - CLAS12 also assume a steady background load, equivalent to ~600 threads from users.
  - Total load ~11400 simultaneous threads.

Requirements

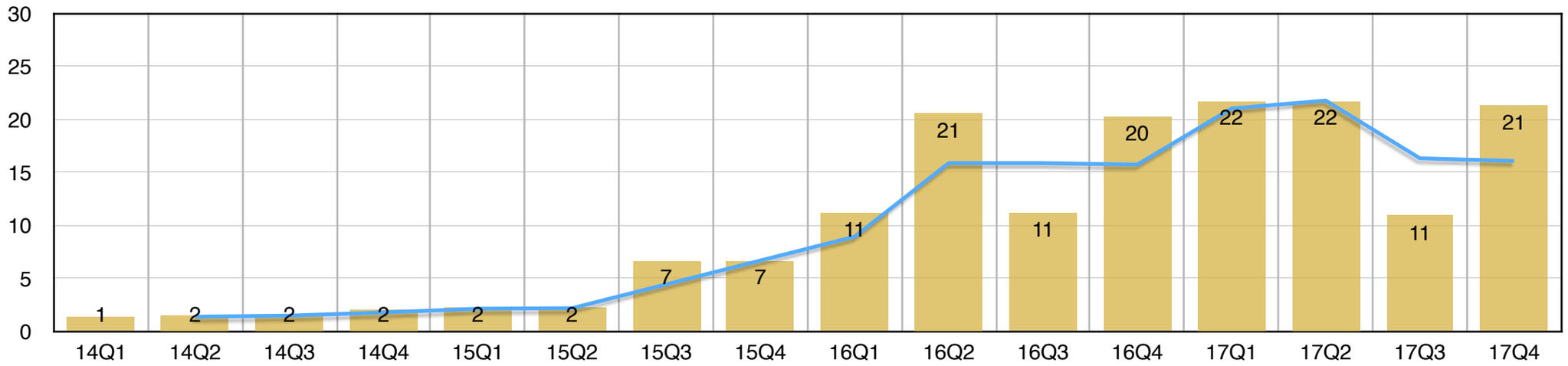
- The tables on the next three slides present the CPU, disk and tape requirements broken down by quarter over the next four calendar years.
- The assumption is steady state data taking so that in each quarter we must at least process one quarter's worth of data to not fall behind.
  - There are periods where the accelerator is down and no new data is being generated. We can take advantage of this to reduce the CPU requirement to meet the average load.
  - For reference the blue line is a moving average of the bars.

CPU

CPU in 1000 of 2013 equivalent cores

	2014				2015				2016				2017			
	14Q1	14Q2	14Q3	14Q4	15Q1	15Q2	15Q3	15Q4	16Q1	16Q2	16Q3	16Q4	17Q1	17Q2	17Q3	17Q4
6 GeV	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.50	0.50	0.25	0.25	0.25	0.25	0.25
Hall A	0.01	0.01	0.01	0.01	0.04	0.04	0.04	0.04	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07
Hall C	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Hall B	0.2	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6	10.0	0.6	10.0	11.5	11.5	0.6	11.0
Hall D	0.1	0.1	0.1	0.5	0.5	0.5	5.0	5.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Total	**1.3**	**1.5**	**1.5**	**2.1**	**2.2**	**2.2**	**6.7**	**6.7**	**11.2**	**20.6**	**11.2**	**20.3**	**21.8**	**21.8**	**10.9**	**21.3**

CPU requirement (units of thousand 2013 equivalent cores)



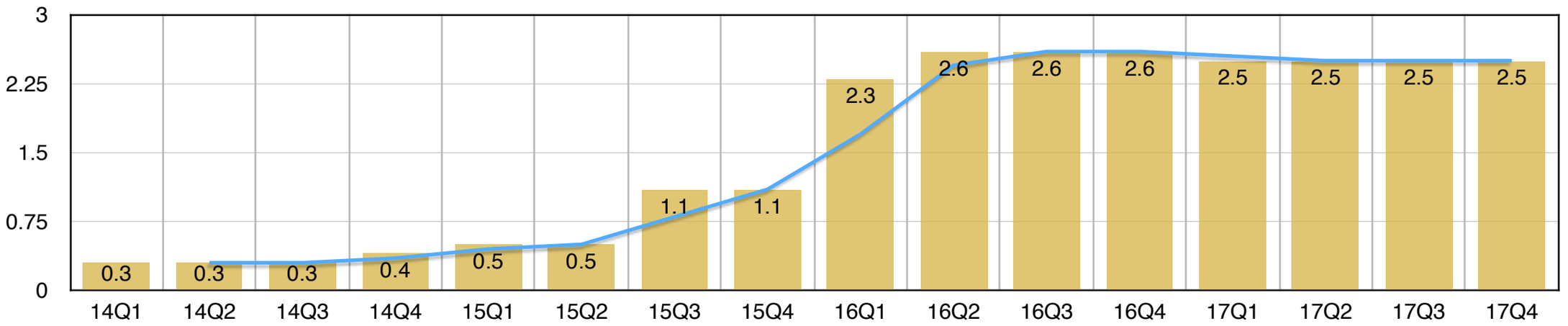
- There is still 6 GeV work so assume that starts with the current observed load and diminishes with time over the next four years.
- Halls A and C are a negligible load compared with B and D.
- Assume large data challenges use resources borrowed from the LQCD clusters.
- Yellow cells commissioning or low volume data taking, green high volume data taking.
- For reference, the size of the current cluster is ~1500 units.

Disk

Disk - volatile + work in TB

	2014				2015				2016				2017			
	14Q1	14Q2	14Q3	14Q4	15Q1	15Q2	15Q3	15Q4	16Q1	16Q2	16Q3	16Q4	17Q1	17Q2	17Q3	17Q4
6 GeV	180	180	180	180	180	180	180	180	90	90	90	90	50	50	50	50
Hall A	20	20	20	20	30	30	30	30	130	130	130	130	130	130	130	130
Hall C	0	0	0	0	30	30	30	30	40	40	40	40	40	40	40	40
Hall B	30	50	50	80	80	80	80	80	80	300	300	300	300	300	300	300
Hall D	30	30	30	150	150	150	750	750	2000	2000	2000	2000	2000	2000	2000	2000
Total (PB)	0.3	0.3	0.3	0.4	0.5	0.5	1.1	1.1	2.3	2.6	2.6	2.6	2.5	2.5	2.5	2.5

Disk requirement - work + volatile in PB



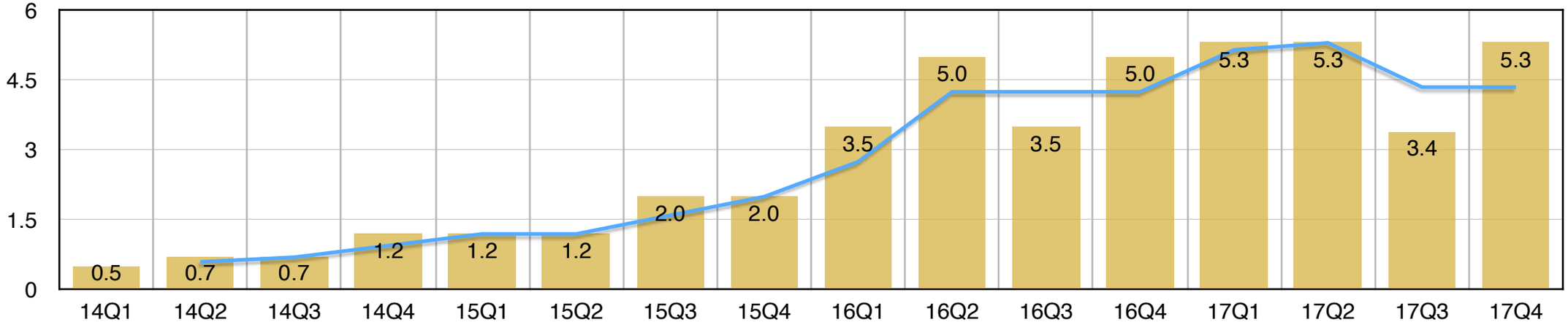
- Disk is in three flavors, work, volatile and write-through.
  - There is also cache disk which is part of the mass storage and isn't listed here.
  - The different types of disk use the same underlying hardware.
  - The ratio flavors of disk depends upon the volume of raw data and the analysis workflow and can be adjusted later.

Tape

Tape TB per quarter (total in PB/Q)

	2014				2015				2016				2017			
	14Q1	14Q2	14Q3	14Q4	15Q1	15Q2	15Q3	15Q4	16Q1	16Q2	16Q3	16Q4	17Q1	17Q2	17Q3	17Q4
6 GeV	200	200	200	200	200	200	200	200	100	100	100	100	50	50	50	50
Hall A	20	20	20	20	60	60	60	60	250	250	250	250	250	250	250	250
Hall C	0	0	0	0	0	0	0	0	350	350	350	350	350	350	350	350
Hall B	253	505	505	758	758	758	758	758	758	2313	758	2313	2624	2624	758	2624
Hall D	0	0	0	200	200	200	1000	1000	2000	2000	2000	2000	2000	2000	2000	2000
Total in PB/Q	0.5	0.7	0.7	1.2	1.2	1.2	2.0	2.0	3.5	5.0	3.5	5.0	5.3	5.3	3.4	5.3

Tape requirement in PB/quarter



- Tape storage costs are controlled by three factors:
  - Cost of library infrastructure and maintenance.
  - Cost of “shelf space” in the library.
  - Cost of media.
- We store duplicate copy of the raw data outside the library.
- Media cost is paid from operating budget of hall - an incentive not to waste tape
- Must eject processed and raw data quickly to keep library costs down.

Cost breakdown by quarter

Schedule of cpu and disk purchases.

	2014				2015				2016				2017			
	14Q1	14Q2	14Q3	14Q4	15Q1	15Q2	15Q3	15Q4	16Q1	16Q2	16Q3	16Q4	17Q1	17Q2	17Q3	17Q4
CPU req k cores	1.3	1.5	1.5	2.1	2.1	2.1	6.6	6.6	11.1	20.3	11.1	20.3	21.9	21.9	10.8	21.9
Disk req PB	0.2	0.3	0.3	0.4	0.5	0.5	1.0	1.0	2.3	2.5	2.3	2.5	2.5	2.5	2.3	2.5
Tape req PB/Q	0.5	0.7	0.7	1.2	1.2	1.2	2.0	2.0	3.5	5.0	3.5	5.0	5.3	5.3	3.4	5.3
			100			90			65			50				
CPU budget k$	0	0	360	0	0	360	0	0	360	0	0	270	0	0	0	0
Cluster size	1.5	1.5	5.1	5.1	5.1	9.1	9.1	9.1	14.6	14.6	14.6	20.0	20.0	20.0	20.0	20.0
Disk budget k$	0	0	0	32	0	0	60	0	180	0	0	0	0	0	0	0
Tape k$	8	12	12	20	19	19	32	32	41	60	41	60	42	42	27	42

- Just in time procurement to take advantage of technology improvements.
- In the table the bold outline is calendar year, the colored areas fiscal year. Red cells are major procurements.
- CPU and disk is procured from ENP Ops budget, tape media from hall Ops.
- In Q1 of calendar 2016 buy ~9k cores, to meet average load of ~15k.
- Avoid a large single year bump in spending by using the boundaries between fiscal years.

NERSC

- The raw datasets are large enough to be impractical to move offsite.
  - Reconstruction must be done at JLab.
- Simulation is 50% of the workload:
  - The code is relatively portable.
  - The output dataset is of a manageable size.
- Data management:
  - 20 PB/yr will fill the tape library quickly, need to eject tapes after reconstruction.
  - Managing data provenance and data preservation is an issue:
    - Raw, reconstructed, calibration, logbooks run conditions.
  - How do we ensure that future researchers have access to data taken today?
  - Perhaps NERSC or others have useful tools or experience?
- Archiving:
  - As part of the data management process we need to archive which data was used to generate the published results.
  - Can NERSC or some other similar body play a role in archiving this information?