Experimental Computing Overview

Graham Heyes

Data Acquisition Support Group lead

ENP computing coordinator

CLAS12 software group co-lead

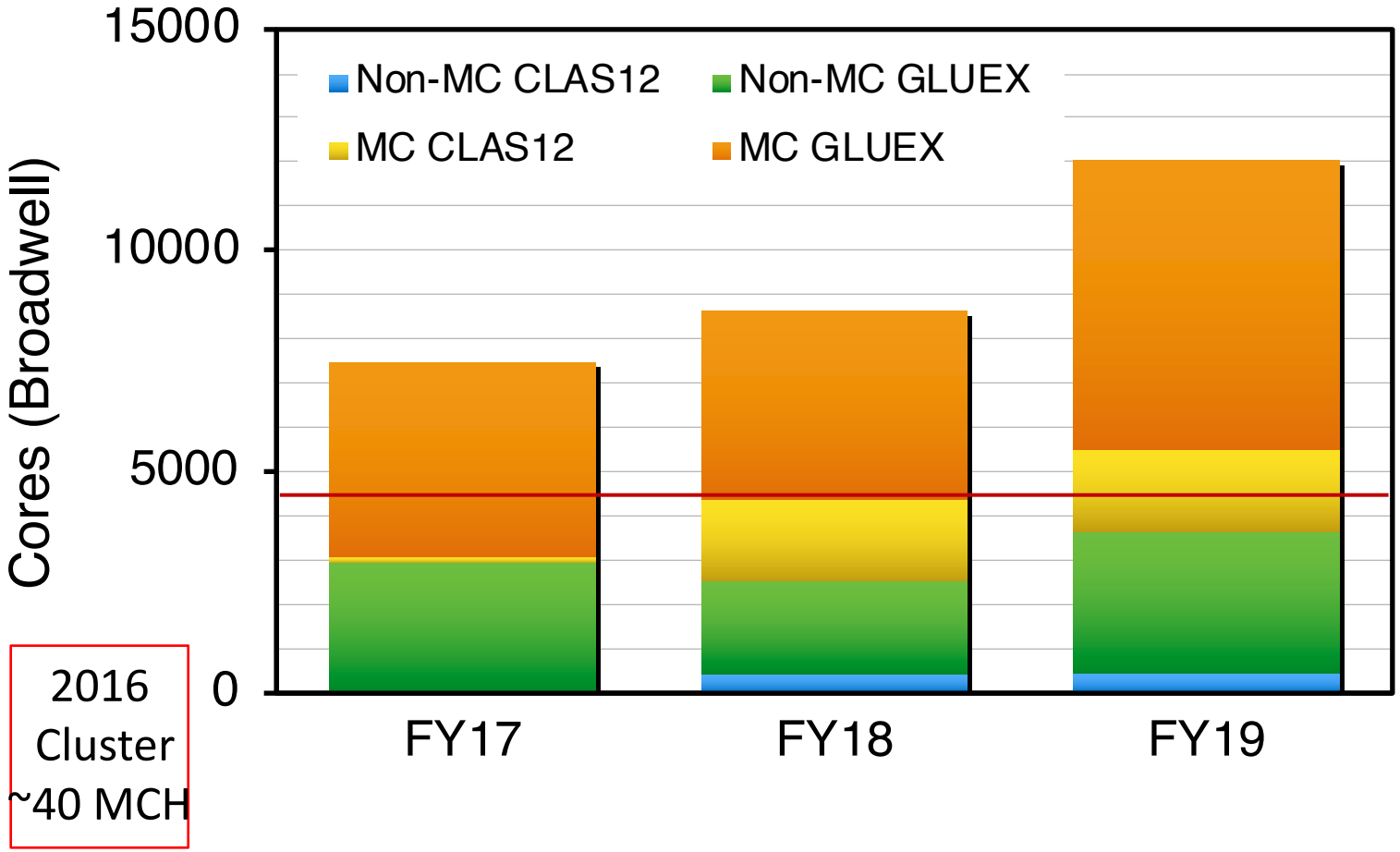
Jefferson Lab

Experimental Computing Overview

- Recap of requirements as of last review and what was presented to S&T
- S&T recommendations
- DOE/NP report – response to S&T
- Roll-up of current requirements
- Roll-up of current off-site planning assumptions
- Drivers for facility planning – lead in to Chip + Sandy.

Last software review (October 2016)

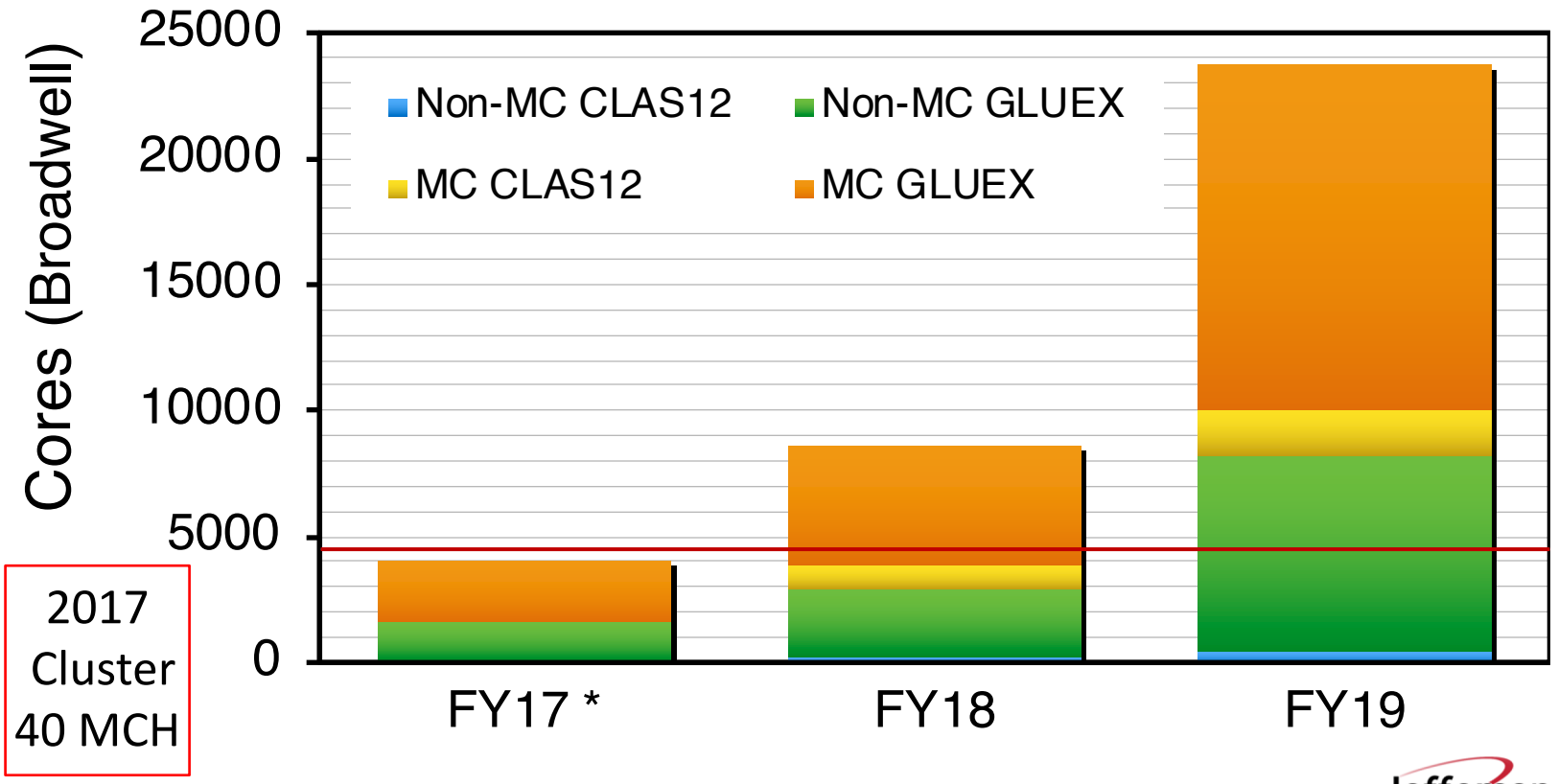
- The total non-MC workload was always less than 4000 Broadwell cores - 2016 JLab farm was 4500 cores (red line).



Jefferson Lab

The 2017 S&T Review (July 2017)

- GLUEX experience with production data, reconstruction grew 1.6x.
- Geant 3 -> 4 transition MC was half as fast.
- Note CLAS12 requirement were relatively unchanged except for a correction based on updated schedule.
  - Little experience at the time with real data from the detector…



Jefferson Lab

S&T recommendations

- 2017 S&T Recommendation:

Generate a cost-effective plan to ensure sufficient computing resources for data analysis and simulation in FY18 and a longer-term approach to address the needs in FY2019 and beyond.

The plan for FY2019+ resources should include a time line and detailed plan to evaluate the feasibility of the proposed approach, including off-site computing resources, such that the plan can be in place and tested before the FY2019 running.

Synergy with the theory computing needs should be considered.

DOE/NP report – response to S&T

- In response to the S&T recommendation a plan was developed and a report written. The two main parts of the plan are :
  - Investment in local resources to expand them to the point where they support the basic needs for calibration, reconstruction and analysis.
  - MC simulation to take place using offsite resources. Investigate OSG, Cloud and other.
- In this it was assumed that the halls would continue to gain experience and refine code and workflows to reduce the compute requirements.
- As noted in earlier presentations CLAS12 had a similar experience as GLUEX: their “real world” reconstruction rate was much slower than anticipated, but they have worked hard to improve it.

Compute requirements process

- Each hall is encouraged to “own” their computing requirements.
  - I check for consistency, advise where needed, merge the requirements and work with Chip on ways to meet them.
- Up to last year the computing requirements were, for the most part, calculated based on expected rates and assumptions about workflow.
- By last year GLUEX had significant experience of real workflow patterns and how their offline software and DAQ behave.
  - They now have a script that calculates computing requirements based on a set of observables.
- CLAS12 are not yet this advanced but the requirements presented today are now based on real experience, however limited.

Hall A

		2019		2020		2021		2022		2023	
Run		Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall
Exp		APEX	PREX/CREX	CREX	-install-	SBS-GMn	SBS-GEn	-install-	SBS-GEp	-install-	SBS-SIDIS
Weeks		13	13	9	0	9	14	0	13	0	13
MC	MCH*	0.0	0.0	0.0	0.0	0.2	0.3	0.0	1.2	0.0	2.3
Non-MC	MCH*	0.0	0.0	0.0	0.0	0.8	1.3	0.0	2.9	0.0	1.1

- The table shows the hall-A experiments during the next five years.
  - Note that the Fall run periods of 2020 and 2023 are installation so no data.
  - Multiplying the core weeks of compute load per week of running by the weeks on the schedule gives the million core hours required for hall-A reconstruction – Peaks at about 4 M core hr/yr in 2022.
  - Simulation peaks at 2.3 M core hr/yr in 2023
- Hall-A's current quota is 5% of the local 80M core hr/year cluster, which is 4 M core hr/yr.

* Note : hall A measured with Skylake, the load here is converted to Broadwell core hours to be consistent across halls. Any requirement below 0.1 MCH is reported as 0.0.

Hall B

		2018		2019		2020	
Run		Spring	Fall	Spring+Fall	Summer	Spring	Fall
Exp		RG-A	RG-A/K	RG-A/B	RG-I (HPS)	RG-F	?
events	Billions	22.5	43	49	36	8	?
Load	M core hr	2.4	4.5	5.2	3.8	0.85	?

- Reconstruction rate has been significantly improved.
- RG-A and RG-B generate the largest compute loads – high rate, long duration.
- Backlog from spring is ~22 B events or 2.4 M core hours.
- Backlog + added load through end of 2019 is ~12 M core hours
- Hall-B's current quota is 45% of the local 80M core hr/yr. cluster, which is 36 M core hr/yr.
- Hall-B simulation ~63 to 80 M core hr/yr.

Hall C

- Hall C compute load is fairly steady at around 2.5 M core hr/yr.
- A third spectrometer arm is planned at a future date. This development must be monitored but is not expected to significantly change hall C's contribution to the compute workload of the lab.
- Hall-C's current quota is 5% of the local 80M core hr/yr cluster, which is 4 M core hr/yr.

Hall-D

		2017	2018	2019	2019	Out
Exp/		Low intensity GLUEX	Low intensity GLUEX	PrimEx	High intensity GLUEX	High intensity
Load MC	M core hr	3	11	1	8	37
Load non-MC	M core hr	21	62	6	40	122

- GLUEX has two phases, low and high intensity.
- The PrimEx experiment is scheduled in spring 2019.
- The switch to high intensity is scheduled for the fall of 2019
- Out years are “full years” of high intensity GLUEX operation.

- It is clear that the requirements presented by GLUEX differ significantly from earlier projections.
  - Simulation requirement has fallen – lower statistics, faster code.
  - Non-MC has gone up.
- Hall-D’s current quota is 45% of the local 80M core hr/yr. cluster, which is 36 M core hr/yr.
  - Excluding MC GLUEX this allocation is only 30% of the out year requirement.

Summary of requirements

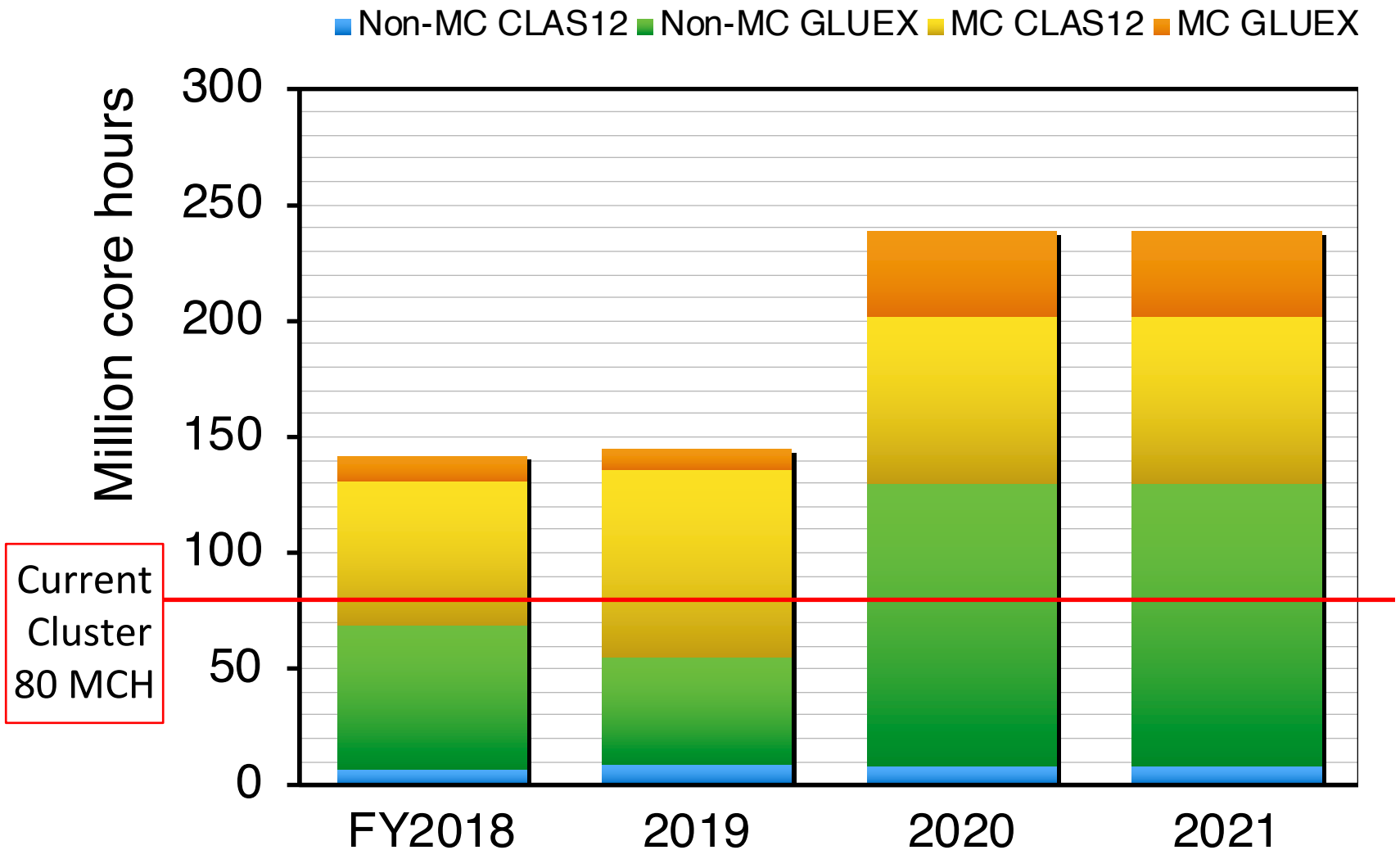
- In 2016 simulation was projected to be the big workload – it has decreased by a significant factor for GLUEX and CLAS12.
  - Since this is the workload with the largest uncertainties and more suitable to offsite we stick to our plan to run it offsite.

- Halls A and C have a small footprint that fits comfortably within their allocation of the existing ENP cluster
- Reconstruction for CLAS12 can be performed onsite within the footprint of their existing quota on the ENP cluster.
- Reconstruction for GLUEX will use a mix of local, NERSC and OSG resources.
  - The local cluster provides only 50 to 30% of the required core hours (low number is in the out years).
  - NERSC is a cost effective fit to the bursty nature of the workflow.
  - GLUEX are on track to use offsite resources but we will monitor.

Roll-up of current off-site planning assumptions

- When we investigated offsite resources commercial cloud was found to be affordable but not cost effective at the scale required.
- Initially there were reservations regarding OSG but these have been resolved.
- GLUEX and CLAS12 are large collaborations involving institutes with considerable computing resources of their own.
  - Work with OSG and member institutes to give GLUEX and CLAS12 access to these resources.
  - GLUEX have taken the lead with this and CLAS12 is following.
- Assume that most MC work will take place offsite.
- GLUEX acquired an allocation at NERSC who prefer that their resources are NOT used for MC. As a result GLUEX are planning to run up to 90 MCH per year of reconstruction at NERSC which covers their shortfall

Summary in chart form



Jefferson Lab

Consideration for facility planning

- The main drivers are:
  - Local farm nodes.
  - Mass storage
    - Bandwidth
      - Raw to tape
      - Copying of raw for backup
      - Playback of raw for reconstruction etc.
      - Writing of intermediate results
    - Disk and temporary storage of “live” data.
    - These are the same irrespective of where the compute happens
  - LAN and gateways
    - OSG gateway nodes – handle submit requests
    - LAN bandwidth to transmit raw data offsite and receive results.
  - Workflow management, batch and monitoring tools.

Input for Facility planning II

- The existing 80M core hour/yr cluster includes the old LQCD 12s cluster which, as it's name implies, was procured in 2012.
  - A priority should be given to ensuring that this resource is at least replaced before it is turned off.
  - Since the local cluster will be heavily loaded any additional boost would be good
- Disk for staging data for local reconstruction and analysis as well as offsite tasks. The halls have submitted requests.
  - Total request is 1.4 PB increase to the spinning disk pool.
- We are investigating using SSD as a passthrough for raw data to tape and to boost IO intensive workflows. Request a modest boost to SSD to allow R&D.

Summary

- The total requirements for all four halls are greater than we have provisioned in the local farm.
- The workload is bursty in nature and requirements are still evolving so we are cautious about overinvesting.
- Offsite resources have been identified and are being exploited.
  - CLAS12 MC exceeds local allocated resources. Contribution from the collaboration, NERSC and OSG will fill the gap.
  - GLUEX has requirements that exceed local resources, between OSG and NERSC have access to resources that meet this need.
- Our current plan has been submitted and accepted by NP.
- As defined the plan allows us to be both proactive and reactive as needs evolve.
- We will pay close attention to all of the halls and adjust as necessary.