Computational resources for EIC data acquisition and analysis.

Graham Heyes
Scientific Computing Department
Jefferson Lab
heyes@jlab.org

Abstract— This document is a preliminary study of the computational resources required by an EIC detector read out in streaming mode. Since this document is being written when detector and data acquisition electronics are at an early stage of design it is expected that the requirements will be subject to future refinement. At this stage the document will be confined to presenting estimates of upper and lower bounds. Possible strategies for meeting the computing requirements are also discussed.

1. INTRODUCTION

This document is relevant to a detector read out in streaming mode. In a traditional system a trigger provides a signal when an interesting event has occurred that initiates acquisition. The data is associated with a trigger, and so to an event, as it is acquired. The data transported through the data acquisition system consists of sequences of parts of events that are at some point combined by an Event Builder into a sequence of events. In streaming mode system, the detector is read out continuously and, at the time of acquisition, there is no concept of an event. The data transported through the data acquisition system is a time ordered sequence representing signals that occurred in various detectors during well-defined time periods. It is only after the data is acquired that the time ordered data is examined to identify events. With a streaming system the acquired data may contain random noise since there is no trigger to filter it out, though thresholds, zero suppression and other local filtering can still occur. Similarly, background events, such as beam-gas interactions and cosmic, are not filtered by a trigger. If the acquired data rate is low enough that the cost is not prohibitive then all data can be archived and processed later. If not, then the data must be processed near-line to remove the background and reduce the data rate. The implication of this is that there is no simple relationship between final event rate, data rate to storage, and acquired data rate. The rates must be treated separately.

The data rate off the detector determines the scale of data transport systems used to buffer data, as well as the I/O bandwidth requirement for data transport, storage hardware, and any real-time filter. The unfiltered event rate, which includes the background rate, and the cost in processing time per unfiltered event to run the filter, gives the scale of compute system required to run the filter. Whether a filer is run or not there is an archived event rate, and archived data rate, out of the data acquisition system that goes to storage and is the input for further data processing and analysis. This data rate determines the scale of near-line data cache and the system and media requirements for long term data archival. The archived event rate is a factor in calculating the total compute cost of further data processing. Typically, the total nuclear physics data processing workload for a dataset has the components: calibration, reconstruction, analysis, and simulation. It is possible to convert the dataset size, into compute load by multiplying total event count by compute time per event and a workflow related scaling factor. The workflow factor can be simple, for example ten trial runs on 1% of the data followed by complete reconstruction would lead to a reconstruction workflow factor of 1.1. When workflows, event type ratios, and per event processing times are well known the calculation can be complex since processing times may vary considerably for different event types. Since the EIC detector is not well defined at this time we shall restrict this document to relatively simple arguments that should still adequately set the scale.

2. ASSUMPTIONS

This document assumes an EIC detector based on the sPHENIX detector at RHIC .[1] The detector is expected to come online for EIC data taking some time after 2030 and is projected to run for 28 weeks per year with a 70% facility efficiency.[2] The fraction of the year that the detector would be taking production data is:

$$F_{pd} = 28/52*0.7 = 0.376$$

According to the presentation by Jin Huang [1] (slide 33), the EIC minimal bias collision rate is:

$$R_e = 500 \text{ kHz}.$$

In a collider, all collision can be considered useful for one analysis or another. Therefore, it has been suggested that storing all minimal bias EIC collisions, without pre-scaling and trigger selection, should be encouraged. Based on this collision rate the non-background data rate for all EIC collision signals, is estimated to be:

$$R_{nb} = 60 \text{ Gb/s}.$$

The acquired rate out of the streaming readout system will be higher since it includes background. Assuming 10^{-9} mbar beam vacuum, the beam gas interaction rate from the proton beam will be 12 kHz within the experimental region of $|z|<4.5$ m. There is nothing useful from these collisions but at this rate it is a small contribution that makes a negligible change to the overall rates. The rest of the background are random and continuous, such as detector dark noise, and synchrotron background. Currently we do not have a full estimation of rates for these. Detector dark noise rate depends on the detector design. As EIC is a low rate collider, it is advised that the detector group NOT design an intrinsically noisy detector. A good example of a low noise detector is a modern MAPS silicon pixel tracker, which can control the noise to very low level (10^{-9} noise hit/pixel/time frame). Therefore, for the purpose this document, it is safe to assume a noise rate that is less than the collision signal rate. Jin quotes an upper limit on acquisition rate that includes intrinsic noise of:

$$R_a \sim 100 \text{ Gb/s}.$$

Jin also notes that it requires a very careful detector and readout design to achieve this. Also, for example, ASIC algorithms in detector electronics and other components of the streaming readout hardware could effectively remove much of this noise. So the true rate will be somewhere between R_a and R_{nb}.

Synchrotron background is currently the major unknown. Jin will try to provide a range that can be included in a later iteration of this document.

An acquisition rate of <100 Gb/s is not a challenging rate for current (2020) technology. This rate is a factor of 2x lower than the rate for the sPHENIX detector at BNL and only a factor of four larger than the achieved bandwidth of the GLUEX DAQ at JLab. Given that sPHENIX will run before the EIC, and that the EIC startup is at least 2030, and that technology improvement is expected, transport and storage at this rate is not a concern.

Data volume estimate

Based upon the fraction of the year that the detector would acquire production data and <100 Gb/s acquisition rate the acquired data volume would be 136 PB/y. This is a conservatively high rate since it should be reduced by online filtering. We will use this rate since the reduction factor is not known at this time. Since reconstruction generally leads to simplification of the data, for example replacing sequences of hits with parameterized tracks, reconstruction output is typically compact and a size 10% of the raw data volume is commonly quoted.

Archiving all the raw and reconstructed data would require a storage capacity of:

$$\text{Archive rate } 1.1 * 136 \approx 150 \text{ PB/yr}.$$

Since a similar scale mass storage system is required for sPHENIX it should already exist at BNL before the EIC startup. Media cots should be calculated but the information to do that is not available.

We typically would aim to keep 20% of the annual dataset on disk so near-line storage need:

$$\text{Cache Disk} \sim 30 \text{ PB}.$$

At current 2020 prices of $100k/PB this would be a $3M resource, $300k is reasonable if bought close to 2030.

In 2020 a 100 Gb/s links are common, and the rate R_a can comfortably transit ESNET if needed. This implies that we are not constrained to a computing model that is tied to local resources due to difficulty in transporting a large dataset. Furthermore, R_a is the acquisition rate, the average rate during data taking. Since the facility efficiency is 70% the effective daily rate from the experiment is:

$$0.7*12.5 = 8.75 \text{ GByte/s.}$$

The experiment runs 28 out of 52 weeks so the effective annual average data accumulation rate is:

$$28/52*8.75 = 4.7 \text{ Gbyte/s.}$$

Processing requirement

Reconstruction and calibration

Assuming the event rate R_e quoted above, of 500 kHz, and the time required to process a single event, T_e, we can calculate the number of parallel jobs required to process in real time as the data is taken, N_r.

$$Nr = 500{,}000 * T_e$$

Unfortunately, T_e is dependent on the event structure which is in turn dependent on the detector. The value is not known but some assumptions about the scale can be made. The EIC is an electron ion collider and so sPHENIX with and EIC is expected to have simpler events than, for example sPHENIX running with RHIC. The events should be closer in complexity to a Jlab detector such as GLUEX and CLAS12. Based on this it is conservative to assume a reconstruction assuming 2020 vintage processors of:

$$T_e = 600 \text{ ms/event}$$

$$N_r = 500{,}000 * 0.6 = 300{,}000$$

This is the number of parallel reconstructions to keep up with the acquisition rate. The latest 2020 vintage AMD compute nodes at Jlab run 128 jobs per node so to keep up with the acquisition rate would need:

$$N_r/128 \sim 2{,}300 \text{ nodes}$$

Historically, the cost per node has been relatively constant, while the performance per dollar increases with each generation of processor. The 2020 vintage nodes were ~$7,000 per node so estimated cost of 2,300 nodes would be $16M if bought today.

Based on historical trends and current industry roadmaps we can expect a factor of 10x improvement between 2020 and 2030, so we can conservatively plan for 230 nodes at an approximate cost of $1.6M, assuming the static price per node still holds. The infrastructure to support a 230-node cluster is likely already to exist in 2030 since BNL has sPHENIX running before EIC. The cost could be inflated by 10% for racks, cables, and network switches but this is lower than the accuracy of the estimate. Similarly, there will be a need for compute resources for offline calibration checks and for final stage analysis. This workload is typically of the order of 10% of that of the full reconstruction and not a large perturbation when the other uncertainties are considered.

Data acquisition

The acquisition rate R_a is low enough that, although it presents and I/O bandwidth challenge for current DAQ systems it should not be a limiting factor for an EIC detector in ten years. Computationally, for a streaming readout, the data flow is already parallelized, and the main consideration is how many parallel threads are needed to perform online processing tasks at a rate that can keep up with the event rate. If complete reconstruction is not needed for a final stage online filter, then a reasonable estimate is 10% of the number of parallel threads required for reconstruction, a cluster of 23 nodes should be adequate and would cost ~$160k in 2030

Simulation

Simulation is the dominant compute intensive task. It is likely that at least ten times more simulated events would be required compared with acquired data. Since simulated data must be generated and then reconstructed a conservative estimate is 20x the compute resource required for reconstruction. However, the earlier calculation was to keep up with the acquisition rate. Data is acquired for only 0.38, say ½,of a year while simulation can be run at any time. So the number of nodes needed for simulation is:

$$230 * 20 * ½ = 2{,}300.$$

Simulation can also be run anywhere an so is a good candidate for distributed computing. A resource of 2,300 nodes as distributed resources should not be hard to find in 2030.

Summary

System.	Scale.	2030 estimated cost.
Mass storage rate	≈ 150 PB/yr.	Part of BNL infrastructure – need media cost estimate.
Cache disk	≈ 30 PB	$300k
Cluster for reconstruction.	≈ 230 nodes	$1.6M
Cluster for DAQ	≈ 23 nodes	$160k

Cluster for simulation	2,300 nodes	Offsite and onsite resources
Datacenter and support infrastructure		Assumed to already exist.

References

[1] Based on presentation by Jin Huang (BNL) https://indico.bnl.gov/event/7449/contributions/35877/attachments/27208/41609/EIC_DAQ.pdf

[2] Private email from Elke-Caroline Aschenauer.