Technical Interchange Meeting on the ePIC Streaming Computing Model

Reference document: https://zenodo.org/records/14675920

Questions from the Advisory Committee:

[General] High-Level questions

1) It would be useful to have a brief overview of the EIC program schedule, for reference. Start date, planned extended shutdown periods, etc …

2) The document describes the online and computing aspects of ePIC. How much of the online part is in scope of the Advisory Committee we are part of? Could be “all”, “none” or “from this point”. It would be useful to know, as the part not in scope we should treat it as “for information” while the one in scope we should treat it as “for discussion”.

3) We still have difficulties understanding what does what in the DAQ. In particular, we do not understand in Figure 3 which elements 1) buffer the information from the different subsystems in timestamped datafiles, 2) recombine those timestamped datafiles into a timeframe, 3) compress the timeframe (or the previous datafiles) 4) push the timeframes into the offline.

4) We also do not understand (related to the above) if the timeframes are processed in any way online (except for being compressed). What is the role of the online data filters? Do they filter out timeframes and based on which criteria? We are trying to understand how much intelligence is in place to make sure that timeframes contain useful information, as we understand they are kept forever.

5) A key aspect of the computing model is the need to run first pass reconstruction at Echelon-2. This is because JLAB and BNL will not have enough resources to do prompt reconstruction. While Echelon-2s will provide more resources, the computing model will also be more complicated and that will probably imply further costs for JLAB and BNL. Has the cost been looked at in a holistic way?

6) Could you provide an explanation of the services and capabilities you expect to find at Echelon-0, 1, 2, and 3? In several parts of the document one finds “this workflow fits the capabilities of Echelon-X but for the other one we need X-1. It is not however clear what you expect to find in X-1 that you do not find in X. An example is section [4.8] on analysis: what do you need for quark-gluon structure analyses that you find in Echelon-2 and not in Echelon-3? E.g. high memory nodes, high IOPS storage, … ?

7) It would be useful to have a description of the data model, with formats of each data tier, a short description of what they contain and the expected event sizes (apart from the timeframes which contain many events). This will help understanding some of the numbers. For example in section [5.3.2] it is very hard to follow the logic of why a complete simulated data sample is 100PB.

8) Similarly to the data model, it would be useful to have a description of the workflows transforming from one data tier to another, and a rough understanding of the CPU time/event for each transformation. This will help understand the assessment of the CPU needs.

[1] Introduction

[2] The ePIC experiment

9) The text mentions 1.5 fb-1 per week with 60% efficiency. How many weeks of run are foreseen in a year (the question is related with the high-level one about the schedule)

[3] The streaming data acquisition system

10) Section [3.1] Different channels will have different readout times. So in order to build the frames you will need some buffering. Where do you buffer?

11) Section [3.1]. What happens if in a frame you do not find at least one event and how likely is that to happen? Do you throw away the frame? We guess no, as you assume improved SW triggers could find an event

12) Time-based frames. For our own understanding: do the time-based “frames” overlap with the preceding and following ones? We are asking because, if they are strictly time-sliced, there’s a (very small) chance—on the order of 1 in 60,000—that an event could be split between two frames. Has this scenario been considered?

13) Section [3.1]. ATLAS does not use data streaming. We guess the text intends ALICE. We challenge that LHCb thanks to streaming now can publish in weeks, while before it was years

14) Figure 8 is very hard to read, so we can not compare it with Fig 6. But from the text in [3.7] it seems the data into DAM is ~10 Tbps while from Fig 6 it reads 1.7 Tbps.May be we misunderstand Figure 6. I assume the RDO data rate is for the input, otherwise it is not coherent with the tape data volume column (input to tape)

15) The text in [3.7] says most bunch crossing will not result in interesting physics. But they could result in noise that gets read out and timeframed. You will not find interesting

events in there, but will you keep those timeframes? (same question as above, but now after some more information)

16) When referring to the DAQ computing farm in [3.8] do you refer to all the elements on the right hand side of the network switch in Fig 3. (apart from the Echelon-1 of course) ?

17) In [3.8] there is a broken link to Figure ?? or there is a missing figure. We do not find a figure that matches the text.

18) In [3.8] we understand that in both scenarios (DAQ in IP6 and DAQ in B725 enclave) the architecture of the online-offline interface will be the same (Echelon-1 in BNL will be a distinct object from Echelon-0 even if they are co-hosted). Did you consider a hybrid where you leverage the benefits of the co-location?

19) Do you have enough floor space/power in IP6 for the full online system?

20) We guess the IP6 option closer to the detector is presumably preferred—assuming space constraints are met. Is there an intermediate solution being considered, such as a surface-level room (like some LHC experiments use) to house online computing equipment?

21) In [3.9] you mention the condition and calibration databases. Where do you intend to deploy those databases and will you have online databases and offline databases ? If the condition and calibration databases have to be accessed for reconstruction and calibration (which is an offline task) they will need to be accessible by the online resources (at least Echelon-1s).

22) In [3.11] you foresee having a buffer depth of 1 week. This is about 7.5 PB of storage. It seems large as at least 1 echelon-1 is located in the same lab as the echelon-0. What are the arguments?

23) Data reduction for the RICH (PID) subsystem seems to be crucial to stay within the bandwidth budget. We couldn't find in the document where this reduction will be performed—presumably in the Readout Computer. Will the available processing resources be sufficient to keep up with the acquisition rate? Has the reduction algorithm been developed and its computational cost estimated?

[4] The computing use cases

24) In [4.1] (and after) we think it would be good to have a diagram that explains for different use cases tha data flow with data rates, etc.. Can we get something along these lines?

25) Section [4.1] and below. When does the Echelon-0 have the green light to flush data on the buffer? E.g. if there is a disk copy at Echelon 1, if there is a tape copy at Echelon-1, if

there are two tape copies at both Echelon-1 … This has an impact on the buffer sizes and the level of sophistication of the agents responsible for transfers and archive.

26) Section [4.4] If we understand correctly, the prompt reconstruction produces events from timeframes. Are the events still “RAW” after prompt reconstruction, in the sense that they are not in the form of physics objects? Maybe another way to phrase this: are “prompt reconstruction” and “first pass reconstruction” the same things in terms of input and output data formats? And, following from that, is the output format of the two ready for analysis?

27) Section [4.6]. Reprocessing seems to be a more relaxed activity wrt first pass reconstruction and in fact can run everywhere, including opportunistic resources even for full reprocessing. Why is that the case? Analyses for publications will run very likely on reprocessed data, particularly for precision measurements.

28) Section [4.7]. we do not fully understand the last part of the section (technical and sociological considerations). If you apply new sw algorithms in reconstruction of real data, I guess you want to apply the same algorithms to reconstruct the simulated data, no?

29) Frame-based Simulation vs Event-based [4.7]. You propose simulating “frames” instead of discrete events to match the real-data reconstruction workflow. While this offers consistency, could the complexity outweigh the benefits? Generating frames requires overlaying many events (long processing time in Geant4) and including various machine backgrounds, which may be hard to model realistically. If the splitting of frames into events happens early in reconstruction anyway, might it be more practical to simulate individual events directly and avoid this overhead?

30) Use Echelon2 for Analysis [4.8]. What is the motivation for using Echelon 2 resources for analysis as well? Wouldn’t this complicate the infrastructure and software environment, since analysis has different characteristics: it is more interactive, less predictable, may require more varied and complex software stacks, and often benefits from fast cache storage. Would it make more sense to build dedicated facilities optimized for analysis instead of trying to generalize Echelon 2 for all use cases?

31) Digital twin development [4.9]. Has there been an evaluation of the potential benefits of building a digital twin, relative to the significant effort required to develop and maintain it?

[5] Computing Resources

32) In section [5.3.1] it mentions that Echelon-2 should also do prompt processing. This is a bit in contrast with section [4]. Also, this will mean Echelon-1 will need an output bandwidth of 1/6th of the RAW stream. Where does the number 6 come from?

33) RAW data archival [5.3.2]. The handling of RAW data is also a bit unclear. The document mentions that RAW data will only be deleted once the reconstruction artifacts are complete and stored, but it's not explicitly stated whether a full set of RAW data is archived permanently. It may help to clearly lay out the storage strategy: what data products are stored, for how long, and where (e.g., temporary, disk, tape, etc.).

34) Echelon-2 processing. It's not clear whether both Echelon 2 sites will process all frames, or if they will share the load. The document suggests the goal is to maintain two complete data copies, but if both sites are fully reconstructing the same data, the computing cost is effectively doubled for the first reconstruction pass. Could this be clarified?

35) Section [5.3.1]. Why do you need monitoring, and slow control data at Echelon-2? Even in the case of running reconstruction, you need conditions data, not the full slow control time series. Also, why in Table-3 monitoring, calibration and slow controls are incoming from the other Echelon-1 and Echelon-2's? The slow control information is collected at the detector ..

36) Fig.10. Are the numbers for disk or tape? Streaming data (timeframes) are ~70PB/year and they go to each Echelon-1, but they remain on disk for ~3 weeks (the time to reconstruct with up-to-date calibrations). I guess you do not need to keep the full 70 PB of disk. On the same figure, why does 35PB of streaming data also need to be at Echelon-2? I guess it is to support first pass processing at Echelon-2s, but for that you would need only a small buffer of the data you want to process at a given point in time. 35PB is roughly the volume that each of the two Echelon-1s will first pass process in one year (70PB divided by two Echelon-1s)

37) Fig.11 (and related text in [5.3.3]. we do not understand why alignment and calibration need to happen also at Echelon-2s. They are complex workflows, they likely require dedicated input streams, they need to upload back constant at Echelon-1s condition databases, etc .. They require 2k cores out of 150k, which is on the order of 1%. We guess BNL and JLAB can provide these 2k cores. What are we missing?

[6] Distributed Computing

[7] Software

38) Section 7.1. There is large repetition in the section (bad cut and paste we guess). The Round Tables and the Workshop are repeated twice. BTW, a pity that the S&C Round Tables are not continued.

39) Programming Languages. There's no mention of programming languages in the document. We assume C++ is the primary language for simulation, reconstruction, and

analysis, but it would be good to explicitly state this. Will you continue with the C++/Python model, or are you considering evaluating other languages such as Julia or some level of language interoperability?

[8] Serving users

40) JANA2 Agnostic algorithms [8.1]. We were surprised to read that developers can write reconstruction algorithms completely agnostic of the JANA2 framework. While we understand that the main purpose of these algorithms is to transform data objects in EDM4hep format into other objects, thus independent of JANA2, it seems unlikely that they could be fully decoupled from the framework. In practice, algorithms often need access to configuration parameters, detector conditions, geometry, and other shared services. All of this typically requires interaction with the framework. Could you clarify how this separation is achieved in JANA2, and whether there are limitations or trade-offs in this approach?

[9] Project organization and collaboration

[10] Long term software and computing plan