Expression of Interest (EOI) for Software

Please indicate the name of the contact person for this submission:

Conveners of the Software Working Group:
- A. Bressan, M. Diefenthaler, and T. Wenaus
- eicug-software-conveners@eicug.org

Please indicate all institutions collectively involved in this submission of interest:

ANL	Argonne National Laboratory
BNL	Brookhaven National Laboratory
CEA/Irfu	IRFU at CEA /Saclay institute
EIC-India	Akal University, Central University of Karnataka, DAV College Chandigarh, Goa University, Indian Institute of Technology Bombay, Indian Institute of Technology Delhi, Indian Institute of Technology Indore, Indian Institute of Technology Patna, Indian Institute of Technology Madras, Malaviya National Institute of Technology Jaipur, Panjab University, Ramkrishna Mission Residential College Kolkata
IMP-CAS	Institute of Modern Physics - Chinese Academy of Sciences
INFN	Istituto Nazionale di Fisica Nucleare
JLab	Thomas Jefferson National Accelerator Facility
LANL	Los Alamos National Laboratory
LBNL and UC Berkeley	Lawrence Berkeley National Laboratory and University of California, Berkeley
NCBJ	National Centre for Nuclear Research
OhioU	Ohio University
ORNL	Oak Ridge National Laboratory
SBU	Stony Brook University
SLAC	SLAC National Accelerator Laboratory
SU	Shandong University

UManitoba	University of Manitoba
USTC	University of Science and Technology of China

Please indicate the items of interest for potential equipment cooperation and the level of potential contributions are for each item of interest:

Scope of the Software EoI for the EIC

The Software Working Group (SWG) of the EIC User Group (EICUG) has coordinated an Expression of Interest (EoI) for Software. The scope of the EOI includes all aspects of software from physics and detector simulations to online and offline analysis. It does not include computing infrastructure.

The EoI is based on the input from the EICUG at large and consists of contributions from various member institutions. It addresses software needs of the EIC but it is not, at the point of EoI submission, a comprehensive expression of the software needs. The SWG intends to carry the EoI forward as a living document that will evolve towards a work plan for the SWG, setting priorities for the next years and goals for the next decade.

The common projects listed in the EoI reflect the interests of the participating institutions. The goals are expressed with respect to two major milestones of the EIC project as follows.

Technical Design Report (TDR): As part of the Yellow Report Initiative (YRI), the Software Working Group has worked on physics and detector simulations that enable a quantitative assessment of the measurement capabilities of the EIC detector(s) and their physics impact. This has been an integral step in the preparation of the TDR(s) for the EIC detector(s). The common projects related to the TDR continue the work on validation of the simulation software as well as the workflow tools required for detector full simulations. These projects will be critical to show that the SWG will be able to sustain a common software effort for the EIC.

Completion of the EIC project: The EIC collaborations will determine for themselves what they do for software, but that will likely include common software. The SWG will continue the software development for the YRI and TDR and will focus on areas with the strongest prospects for common work, including areas with profound consequences for the software design:

- **Streaming readout** Analyzing and monitoring streamed data, at least partially in near real time, is a paradigm shift for the Nuclear Physics data models and data processing.
- **User-centered design** The software consortium will work with the EICUG to enable all scientists, regardless of their level, to participate in EIC simulations and analyses.

The software development will build upon the infrastructure of the EICUG, including the GitHub organization and the GitHub website.

Common Projects

These are common project areas that institutes participating in the EoI have identified thus far. They cover **Software Tools for Simulations and Reconstruction**, **Middleware and Preservation**, and **Interaction with the Software Tools**. We don't consider this list exhaustive and we expect that further involvements may extend it.

Software Tools for Simulations and Reconstruction

Monte Carlo Event Generators

The SWG has reviewed Monte Carlo event generators (MCEGs) for ep and eA processes and discussed their requirements and developments for the science program at the EIC.

For the preparation of the TDR, we will maintain a collection of MCEGs that are used in the EICUG and validate them against existing DIS data. We will provide an online catalogue of MCEGs where the community can select MCEGs and find documentation, validation plots, and examples on how to use the MCEG with existing EIC Software. The MCEGs will be deployed in the same manner as described in **Discoverable Software** and thus fully integrated in the **Workflows** as outlined in that section. Comparison to H1, ZEUS, and COMPASS measurements will be used to validate the MCEGs. For the MC-data comparison, we will use the RIVET tools as recommended by the worldwide MC community.

For the completion of the EIC project, we will define with the community what part of the MCEG collection can be substituted and what part needs to be preserved for the EIC. We will make the necessary changes to the legacy MCEGs to ensure their future compatibility, e.g., adding HEPMC3 as output format and integrating them in the RIVET workflow. At the same time, we will work with PARTONS and other initiatives to integrate new and upcoming MCEGs in the MCEG collection. We will collaborate with the worldwide MCEG community on a global DIS tune based on our validation work.

Participating institutions (alphabetical order) ANL, BNL, CEA/Irfu, EIC-India, IMP-CAS, INFN, JLAB, LANL, LBNL and UC Berkely, NCBJ, OhioU, ORNL, SBU, SU, USTC.

Detector Simulations

Towards the TDR, while fast simulation will continue to be important, accurate full simulation will assume a larger role in selecting detector technologies, optimizing the configurations and confirming the discovery capability of the detector system.

During the detector design phase, ease of switching detector options with comparable levels of detail is also essential. To meet these requirements, a comprehensive and centrally maintained simulation framework with a library of potential detector options is desirable.

The engine of such a simulation framework is Geant4. Physics models in Geant4 should be carefully evaluated and chosen to meet EIC physics requirements and well validated. Also, given Geant4 is a living software tool steadily being updated, the physics quality of chosen models and potential alternatives should be continuously monitored. We will maintain a Geant4 physics list.

Accurate detector simulation is also essential for calibration and alignment studies of the detector towards the project completion. Calibration data both from early test beam data of the individual detector components as well as the actual experiment data of the entire detector system should be carefully compared with detailed simulation, and any discrepancy between real data and simulation results must be understood. In regular meetings, we will follow up with detector projects on their simulation work and integrate their updates in the centrally managed software stack.

An automatic validation tool to check new revisions before publishing them will be put in place. This will be built on the ongoing validation of the Geant3- and Geant4-based simulations tools that are used for the YRI. Running the full chain over a given list of Monte Carlo data will allow us to monitor performances and results and eventually identify unexpected behaviours.

In the validation, we will ensure that fast simulations are in agreement with full simulations. We will work on accelerating full simulations simply by adjusting the level of details in the full simulations or by utilizing ML approaches. This will allow us to use the same tool for both fast and full simulations.

Participating institutions (alphabetical order) ANL, BNL, CEA/Irfu, EIC-India, IMP-CAS, INFN, JLab, LBNL and UC Berkeley, LANL, SBU, SLAC, SU, UManitoba, USTC.

Reconstruction

The EIC software will use event reconstruction algorithms to turn detector signals into kinematic quantities. This will include track finding and track reconstruction, vertex reconstruction, calorimeter reconstruction, particle identification, and jet-finding. Some algorithms have been around for decades (tree search, Kalman filters), but they are being reevaluated using machine

learning techniques (e.g. trackML). We intend to approach reconstruction with a focus on modularity. This will allow us to compare performance, and exploit the unique strengths of some algorithms in schemes that combine fast methods in early stages with more accurate methods in later stages of reconstruction. Modularity will also enable independent development of the various reconstruction algorithms.

In preparation for the TDR, we intend to survey the available algorithms for track finding and track reconstruction, vertex reconstruction, calorimeter reconstruction, particle identification, and jet-finding. We will identify their strengths and weaknesses, and test existing implementations on Monte Carlo simulations akin to offline reconstruction conditions. In particular, we will gain expertise with ACTS (A common tracking software) which encapsulates the ATLAS track reconstruction software into a generic, framework- and experiment-independent software package.

On the path towards project completion, we will develop modular reconstruction software that includes tracking, vertexing, calorimetry for central and far-forward regions, and particle identification, and that very likely takes advantage of machine learning for performance. With a modular approach, we will be able to identify those algorithms for online reconstruction that will allow us to achieve near real-time reconstruction.

Participating institutions (alphabetical order) ANL, BNL, CEA/Irfu, EIC-India, IMP-CAS, INFN, JLab, LANL, OhioU, ORNL, SU, USTC.

Middleware and Preservation

Workflows

For the preparation of the TDR, we will identify and deploy software and tools available for EIC-related computations on federated computing resources. Tools will include those for managing job submission and prioritization, opportunistic use of resources when and where available, and management of stored data. We will apply tools able to integrate data and analysis preservation from the beginning, by e.g. making a permanent record of full workload specifications including software versions, configuration and data.

Participating institutions (alphabetical order) ANL, BNL, INFN, JLab, ORNL, UManitoba.

Data and Analysis Preservation

Already during the Yellow Report and TDR period, data and analysis preservation is an important issue. Decisions on detector design and conclusions on detector performance and physics requirements will be made on the basis of software and data constructs which must be

well defined, preserved and documented if these important studies are to be reproducible, and available for use as a well understood basis for progressing to more sophisticated studies. Accordingly, beginning now we will seek to establish a DAP activity that includes:

- ❏ Documentation and preservation of simulation and reconstruction tools, analysis code, data products and workflows. Data and analysis preservation will be an integral part of the common project on **Workflows**.
- ❏ Documentation and preservation of data and software required for detector development, e.g. test beam experiments. This will include a catalogue of MC data samples for the TDR, including the reference event generator data as well as full simulations data.

Based on our experience for data and analysis preservation for the TDR, we will inform the EICUG on possible strategies for data and analysis preservation at the EIC. In preparation for these discussions, we will develop expertise in state-of-the-art research data management systems such as Invenio RDM and community tools such as the HEPData portal. We will systematize the detector geometry description and documentation, with a focus on long-term preservation and expand our workflow tools for the TDR for management and preservation of software, data and detector characterization metrics.

Participating institutions (alphabetical order) ANL, BNL, JLab, SBU, UManitoba.

Interaction with the Software Tools

Explore User-Centered Design

All scientists of all levels, worldwide, should be enabled to actively participate in EIC simulations and analyses. To achieve this goal, we must develop simulation and analysis software using modern and advanced technologies while hiding that complexity .

In user-centered design the development of software is strongly informed by the requirements, environments, and characteristics of anticipated users. Early in the development process, users are included through surveys and focus groups. Throughout the development process, users participate in testing which informs the developers about where to focus attention or change technical approaches. This stands in contrast to some traditional software development practices where the users only play a role when development is completed.

While in smaller experimental collaborations it is easier to gain insight in user expectations without dedicated efforts, this is unlikely to be the case for the 1000+ international user community of the EIC. We anticipate that there will be a spectrum of user classes that will span from new undergraduate or graduate students to research software engineers. Our initial focus will be on actively including new students in the early software testing processes, since more

advanced users are likely to be familiar with methodologies. However, even advanced users will be included since they may not accept some approaches that they may feel stifled by.

Our main focus toward the TDR will be the creation of a multi-tiered, representative test user base who will be contacted for early feedback on our software products. Instead of allowing self-selection (e.g. by using users who seek out new software), we will reach out to users instead. This will increase the likelihood of our software being tested by a representative cross-section of future users.

Participating institutions (alphabetical order) ANL, BNL, JLab, LANL, UManitoba.

Discoverable Software

Growing the active EIC user community requires building entry points for new users at all levels of experience. Experimental physicists often prefer to try out new software before reading documentation that describes the intended workflows. Thus, we have to make it easy for new users to discover which EIC software exists and make its use as intuitive as possible, starting from a single point of entry.

All our EIC software delivery mechanisms will need to behave consistently, whether they be individual user laptops and desktops, shared systems at universities and national labs, large scale HPC sites and distributed systems (e.g. Open Science Grid), cloud computing environments, and web interfaces. The EIC SWG is using Spack as the center of our software delivery strategy to these various systems. For networked systems we distribute a fully operational EIC environment at /cvmfs/eic.opensciencegrid.org. The software stack will also be available as containers (heavy-weight and cvmfs-integrated) and for continuous integration in GitHub workflows. Since users expect a consistent experience over time and across systems, we anticipate that the level of abstraction offered by Spack will be positively received.

As we have been working out the details of our Spack software delivery, we have not advertised these efforts to the wider EIC community, in part to ensure we have a stable working product and in part to avoid changing up workflows late in the YR process. We will be working on documentation which, as described above, will not replace discoverability of interfaces.

Our main focus toward the TDR will be completing the list of relevant software packages we are supporting, and automating the software packaging development, testing, and deployment process. Our goal towards project completion is to build out the community of contributors to the EIC Spack repository, where development activity will be automatically tested and deployed within less than 24 hours to the cvmfs and container systems.

Participating institutions (alphabetical order) ANL, BNL, EIC-India, JLab, UManitoba.

Data Model

Traditional data models in NP experiments are typically event-oriented and with deep hierarchies reflecting the logical detector arrangement. Considerations of interfacing efficiently to ML and other scientific software tools (such as the SciPy stack), accelerators and modern HPCs, as well as the envisioned streaming readout design for the DAQ, make it important to switch to flat data structures with a column-oriented design to achieve best performance.

To this end, for the preparation of the TDR we plan to define common input and output formats for the common software packages to facilitate easy exchange between software components, exchange of software modules, and data preservation.

For the completion of the EIC project, in close collaboration with the community involved in the readout design we will define data models suitable for streaming readout and the processing of streamed data in the analysis, both online and offline.

Generally, tools for validation and reference implementations will be made available.

Participating institutions (alphabetical order) ANL, BNL, EIC-India, INFN, JLab, OhioU, SBU.

Future Technologies

Software development is one of the most rapidly changing fields. Anticipating EIC running in the next decade, we should be open and do research on opportunities to productively apply new technologies and methods. In addition to working on **Common Projects**, we will explore emerging technologies and evaluate their relevance for the EIC. An examples of technologies is listed here:

- ❏ Artificial Intelligence (AI): AI/ML is one of today's game changing technologies. All aspects of developing and operating EIC accelerators and detectors, data processing, simulation and analysis can benefit from applications of various AI/ML methods. Research will be done to adopt and develop these technologies in our field.
- ❏ Heterogeneous computing: Extending software to effectively use heterogeneous environments is a challenging task. Along with AI/ML, GPU use is rapidly widening, not least as a platform for ML, and one also could anticipate the continuing growth of TPUs (Tensor processor units), FPGAs and even ASIC based algorithms in our field. Studying their utility and developing applications, in both online and offline environments, will be an important task.
- ❏ New languages and tools: Today, Python has become a mainstream scientific data processing tool replacing old analysis code in many areas. Furthermore a number of new

languages (Rust, Julia, Swift for ML, etc.) and technologies are emerging and will compete to become mainstream in future. Investigating the opportunities and capabilities that such new technologies offer, and how they could be integrated coherently in terms of APIs and modular design, will be an interesting area to investigate.

- ❏ Collaborative software: Collaborative tools continue to evolve rapidly, to the great benefit of developers. Cloud-like workflows and collaborative workspaces support easy effective access and support for users and teams. Today such functionality can be partially achieved with open source tools such as Jupyter, and we anticipate such technologies to be greatly improved in future and applied for EIC software.

We will also follow up with the theory community on advances in Lattice QCD and Quantum Computing and discuss how their work can be connected to the EIC Software efforts.

Please indicate what, if any, assumptions you made as coming from the EIC Project or the labs for your items of interest:

The scope of our EOI does not include computing infrastructure. We assume that the computing infrastructure for the EIC will be provided by Brookhaven National Laboratory and Thomas Jefferson National Accelerator Facility.

Please indicate the labor contribution for the EIC experimental equipment activities:

Six DOE National Laboratories and 11 other member institutions of the EICUG have joined this EoI with over 15.6 FTE's already identified who are willing and able to contribute to our common software effort. Please find a table summarizing the labor contribution of all participating labs and institutions according to their best knowledge. A detailed table is available in a separate Google sheet. We will update the Google sheet when more lab and/or institutions join our common software effort.

Institute	Estimated FTE
ANL	3+
BNL	
CEA/Irfu	
EIC-India	5
IMP-CAS	2
INFN	

JLab	
LANL	1.4
LBNL and UC Berkeley	
NCBJ	
OhioU	0.6
ORNL	1.7
SBU	3.7
SLAC	
SU	
UManitoba	1.2
USTC	

Please indicate if there are timing constraints to your submission:

Designing the detectors needed to realize the science program of the EIC requires a plethora of software tools. Many of these tools have yet to be developed or need to be expanded and tuned for the physics reach of the EIC. Still, various groups use disparate sets of software tools to achieve the same or similar analysis tasks such as Monte Carlo event generation, detector simulations, event reconstruction, and event visualization, to name a few examples. With a long-range goal of the successful execution of the EIC scientific program in mind, it is clear that early investment in the development of well-defined interfaces for communicating, sharing, and collaborating, will facilitate a timely completion of not just the design of an EIC but ultimate delivery the physics capable with an EIC.

It is important that we work now not only on the software for the TDR but also on the software for the EIC. This will allow both developers and the users to gain experience with modern and new approaches and don't get stuck in legacy approaches. In the list of the common software projects, we describe the goals with respect to two major milestones of the EIC project, the TDR and project completion.

Please indicate any other information you feel will be helpful:

None.