Experimental Nuclear Physics

Jérôme Lauret (BNL), Jeff Porter (LBNL)

The RHIC and ALICE Nuclear Physics science programs have been generating extremely complex datasets at petabyte scales from nucleus-nucleus, proton-nucleus and proton-proton collisions at collision energies that span more than two orders of magnitude, from the energies accessible at RHIC to the highest proton-proton collision energies at the LHC. The size and complexity of these samples are expected to grow dramatically over the next few years, reaching the exa-scale regime in this decade, as the experiments and facilities pursue aggressive upgrade programs. Those upgrades will demand much faster data acquisition rates and add new data streams, some requiring continuous readout to properly sample unbiased (minimum-bias) collections, which can greatly inflate the dataset sizes to unmanageable scales if not tackled with care. It is, however, only with the full range of the rich datasets at RHIC, eRHIC and the LHC that physicists will be able to fully explore, with precision, properties of the strongly interacting Quark Gluon Plasma (QGP) and probe the partonic structures of both nucleons and nuclei to discover insights on the dynamic evolution that exists from cold nuclear matter to hot QGP.

The envisioned dataset growth has created a dire need for innovative data flows where computational power added closer to the data taking has to be considered to achieve sufficient in-situ data reduction for the experiments to properly store and manage the primary data. The experiments have begun to consider including in such data flows efficient real-time decision making filters (High Level Trigger or HLT, pattern recognition) as well as data reduction and repacking methods (fast online tracking, pile-up and hit rejection at the source for data reduction). By moving detector calibration processes closer to the data taking, real-time first pass track reconstruction in HLT and collision vertex reconstruction could lead to both highly filtered information per detected collision and better decision making in regards to event selection for collecting events that hold the highest potentials for key physics measurements.

The foreseen data compression and repacking benefits have been thought to hold the potential for data reduction by factors of x2 while the high precision selectivity could bring an order of magnitude in data size, a relief to the later data size challenge. However, the algorithmic approaches allowing us to consider such a processing transition are still in their early development. And while those data flow challenges unfold, modern computers have evolved toward highly parallel architectures often leveraging powerful co-processors (GPU, Phi, etc.) of different memory and bandwidth granularity, further requiring the architecting of highly flexible workflows and data paths at many levels. Real-time calibration,

quality assurance, event reconstruction, event aggregation, sampling, sorting or selection could all happen at different levels as seen in the nascent HLT farms. The morphing of the current computing landscapes toward online farm based pre-processing carries high risk for data loss and will require trigger and detector data flow simulations to be deeply integrated into the framework.

The realization of such transition and the full exploitation of the new computing architectures at an algorithmic and workflow level will require highly scalable and flexible asynchronous IO models allowing the seamless representation of data suitable for online and offline real or simulated data representation. Those data models must as well envision the fusion of signals: simulated and real data merging is essential to the evaluation of detector efficiencies as well as the development of a more complete understanding of the physics program goals.

On the other end of the data flow, the user analysis, the RHIC data streams are good examples of deep challenges: spanning over more than a decade of data taking, have also raised two new major difficulties exa-scale collaborations will also have to face: (a) the diversity of data and widely distributed datasets require advanced methods for data discovery and sharing to allow comparative analyses that span the experimental campaign and (b) the low-hanging fruits or “easy” science topics are far behind, bringing forth computationally challenging analysis that could be alleviated by enhanced data transport and data store via agile frameworks and modern adaptive data representation.

The diversity of tasks and the sharing of analysis and simulation loads as well as their variability would dictate a largely elastic computing resource where computational and data workflows can be efficiently planned and coordinated. Further, the NP need would be greatly served by the creation of collaborative environments that allow DOE science collaborators to easily save, share, and distribute the large-scale distributed data generated by various science experiments with the next generation high performance networks.