Report on HEP/ASCR Data Summit

Salman Habib, Rob Roser, and Craig Tull (HEP)
Bruce Hendrickson, Rob Ross, and Arie Shoshani (ASCR)

April 20, 2013

Introduction

Representatives from the HEP and ASCR communities met at Germantown on April 2-3, 2013, to discuss issues in carrying out science with large datasets and associated data-intensive computing tasks. The authors of this report acknowledge the important contributions made by all of the ASCR and HEP participants at the data summit and thank them for their efforts.

During the meeting, HEP scientists made a series of presentations highlighting how they as a community interact with data, what tools they have developed in doing so, and the challenges to be faced moving forward. From the presentations and discussions, it became clear that the emerging data-intensive challenges span all HEP frontiers. Furthermore, since HEP is building fewer machines and experiments, there is increased emphasis in further optimizing the extraction of science results from the data by staying informed about the latest technologies and taking advantage of those that are particularly relevant.

The joint summit proved to be very useful, as it not only familiarized the ASCR team with HEP science, but it also made apparent to the HEP group that there are significant common issues across the three HEP research frontiers that, in the past, have traditionally been addressed separately within each domain. A number of potential areas of common interest across the ASCR and HEP communities were identified. They cover the full range from pure R&D to focused assistance in the development and optimization of specific applications. It was generally agreed that HEP data management tools and techniques are already state of the art, and meet current needs. The community has extensive experience with management and analysis of large data sets. Properly planned joint HEP-ASCR research could incorporate transformational advances in developing tools for future applications. Both communities expressed the need to have viable models for maintenance of such tools and software, once developed.

Common Themes Expressed in Data Space

The common themes that emerged, along with some broader issues, are discussed below. Of particular interest were topics that cut across all the areas of HEP scientific engagement, i.e., the Energy, Intensity, and Cosmic frontiers.

(1) Data-Intensive Science: In order to extract scientific knowledge, a central aspect is the manipulation, exploration, and analysis of data. This requires the development and application of a host of tools and methods tailored to data-intensive applications –

scalable and approximate algorithms, robust machine learning methods, experiment and simulation design, and advanced statistical techniques (regression, solution of inverse problems). Applications include anomaly detection, coverage of gaps in observed data sets, design of simulation campaigns, and experiment optimization. Because of the complex nature of the data and the techniques involved, experts from both the ASCR and HEP communities must work closely together. As datasets grow in size and complexity, uncertainty quantification and verification and validation are topics of ever increasing concern to HEP scientists; these are also natural areas for collaborations to develop between ASCR and HEP. Finally, the size and complexity of the data analysis chain have reached the point where additional and/or improved workflow tools are essential for a large number of HEP science cases. Joint development of such tools with ASCR researchers would be extremely valuable.

(2) Data: Many aspects of future HEP/ASCR interactions relate to the nature of the data itself. These issues cover topics such as data representations (e.g., event models) and data structures (e.g., layout of events in storage) designed to optimize data manipulation and analysis, and data organization to enhance efficient selective data access (database design and indexing). Scalable metadata (e.g., parallel databases) was another area of interest for all HEP frontiers. Data archiving and curation (or knowledge preservation) also emerged as a significant cross-cutting theme. Real-time monitoring of experimental and observational data was identified as a challenging requirement as data volumes continue to grow.

(3) Throughput Maximization: A clear common requirement for the three frontiers and each of the represented HEP experiments is the need to maximize throughput both for simulation and for analysis. This requires modeling and optimizing, to the extent possible, the computational hierarchy, communication networks, and data organization. For the LHC experiments, as for others, improvements of throughput translate directly into either cost savings and/or expanded capability to do science. A global strategy to improve throughput typically involves making advances on several fronts that compound, resulting in a greater improvement than from any one alone. Such fronts include:

- Evaluation and potential adoption of emerging technologies and machine architectures.
- Monitoring and profiling of codes through low-impact methods to determine bottlenecks and the impact of code changes.
- Predictive modeling and simulation, including use of an overall cost model (cost in terms of money, time budget, FTE investment, etc.) to help make cost/benefit analyses of alternative approaches.
- Use of networks as active components of the overall dataflow and workflow, including intelligent applications interacting with instrumented, active networks and predictive, adaptive caching.
- Services to optimize data access; this requires services that adapt to data popularity, e.g., data replication based on dynamically changing workloads.

As advances fan out across these fronts, and throughput increases across software and experiments, science output will increase, as will the ability to process more data.

(4) HEP/ASCR Partnership: Because of the diverse and distinctive nature of the ASCR and HEP portfolios, it is important to consider the different modes of interaction between researchers from these two areas. Continuity in the HEP/ASCR dialog was considered highly desirable, enabled for example by future workshops and by the establishment of joint teams to tackle priority problems. The OSG effort is an example of previous engagement with ASCR in important HEP endeavors. A key aspect of the interaction is the importance of expert knowledge that, in many applications, needs to flow in both directions, from domain scientists to ASCR researchers and vice versa. One suggestion was to form a HEP/ASCR future technologies working group to assist HEP with roadmapping activities in data-intensive science.

Larger Issues

Several important topics that were touched upon go beyond the specificity of HEP applications and are important to mention. It is hoped that the proposed HEP/ASCR partnerships can help further the attainment of a broader objective of a common data vision across the sciences.

For a HEP/ASCR data initiative to be successful and result in powerful products that are truly useful and adopted by the HEP community, close and continuous collaboration between the two communities is required. We stress that it is not sufficient to just define a set of requirements and wait for a "result". We strongly advocate a collaborative model where joint HEP/ASCR teams are supported without the imposition of hard funding boundaries on the researchers, scientists, and software developers. Furthermore, a final "result" that is useful for the HEP (or any science) community needs to be of production quality and not just be a research prototype.

Virtual data facility concepts that are now under discussion at ASCR would be a useful and important future strategy for engagement with HEP and the other SC offices. We welcome the construction of such a facility.

In the context of the cross-cutting workshop that followed the HEP/ASCR Summit, several other issues were discussed that could involve HEP activities and lead to collaborations between HEP and other offices mediated via ASCR partnerships. Tools developed through HEP/ASCR partnerships, or existing HEP tools, could be adapted to other disciplines via cross-cutting initiatives engaging expertise from both (HEP and ASCR) communities as well as the intended new users. Future advances and tools made possible through cross-cutting initiatives could also benefit HEP science. Such tools should have an associated long-term management plan as they are designed to benefit a broad community.

Summary Remarks and Priorities

Members of the HEP and ASCR communities met in Germantown to discuss the topic of future data needs in HEP. The conversations were very productive and the challenges facing HEP were well articulated. The importance of research directions that impact the broad scope of HEP activities was emphasized. The hope is that this 2-day meeting is the beginning of a series of interactions that will support fruitful partnerships, helping to solve some of the most critical data challenges facing HEP research areas.

We believe that the formation of a joint ASCR/HEP future technologies group (hardware/software/networks) would be very useful and can be initiated right away. The other topics mentioned above are listed below. The priorities take into account the breadth of impact across the HEP scientific frontiers. We recommend that these should constitute the initial focus of the partnership.

(1) Machine learning/statistics methods for classification, regression, and solution of high-dimensional inverse problems; associated uncertainty quantification (UQ) and verification and validation (V&V). These methods can be applied in a number of applications including the design of triggers, event classification, data pipelines, and cosmological surveys.

(2) Scalable and approximate algorithms for data analysis (e.g., anomaly detection, clustering); associated UQ/V&V. This broad area covers a number of applications in energy, intensity, and cosmic frontier data management and analysis pipelines.

(3) Complex workflows; integrated analysis platforms (e.g., science gateways, PANDA, PDACS). This area was seen as a good platform for collaboration with ASCR across all three HEP frontiers, which have their own individual efforts in this direction.

(4) Joint projects in data organization and data structures to optimize data selection, manipulation, and analysis. This is an important emerging area of particular interest to the energy frontier data "software stack" and the analogous situation for cosmological surveys and includes parallel databases and indexing as an important sub-topic.

(5) Sampling/Experimental Design. This topic covers the optimal choice of experimental parameters given constraints in resources and maximization of science output. It is of interest to all three frontiers.

(6) Throughput maximization via multiple strategies. Given resource constraints, it is very important for HEP to maximize science throughput across all of its efforts. This requires optimization of available computational and network capabilities, and their modeling, to decide on the most effective future investment directions.

(7) Data archiving and curation strategies. This is a topic of interest across all three frontiers, especially in improving facilities for scientific analysis of archival data.