Scientific Software Innovation Institute for High-Energy Physics (S2I2-HEP)
Conceptualization Phase

January 19, 2017

Summary of Workshop 1:
Fostering Collaboration between HEP and Computer Science Communities

December 7-9, 2016
University of Illinois
National Center for Supercomputing Applications (NCSA)

Prepared by the Workshop Organizers:
P. Elmer, D. S. Katz, M. Neubauer (Chair), M. Sokoloff, D. Thain

Overview

The worldwide particle physics community is currently planning upgrades to the Large Hadron Collider (LHC) at CERN in Geneva. The LHC currently uses a worldwide distributed computing model to meet the needs of thousands of scientists to process and analyze some of the world's largest scientific datasets. The upgrades being planned will increase data volumes by more than two orders of magnitude and require significantly more complex data and analysis techniques.

The first workshop dedicated to conceptualizing an NSF Scientific Software Innovation Institute for High-Energy Physics (S2I2-HEP) was held on the campus of the University of Illinois Urbana-Champaign from December 7-9, 2016. The purpose of this workshop was to bring together a diverse set of attendees from the particle physics and computer science (CS) communities to develop an understanding of how the two communities could work together in the context of a future NSF Software Institute aimed at supporting particle physics research over the long term.

The workshop was jointly hosted by the University of Illinois HEP Group and the National Center for Supercomputing Applications (NCSA). The workshop agenda was organized around a series of topical presentations and breakout sessions involving all attendees but within smaller groups to facilitate discussion and active participation. The workshop main page, including links to the detailed agenda, presentations, logistics for participants, attendee profiles, and photos from the workshop can be found at:

http://hep.physics.illinois.edu/hepg/S2I2-HEP-CS-WKSHP/home.html

The workshop was very productive and successful in achieving the goals we set forth.

Workshop Goals

The primary goals of the S2I2-HEP workshop were to bring together members of the HEP and CS communities to identify

1. possible areas of collaboration between HEP and CS that can help facilitate future particle physics research, particularly in the context of the Community White Paper (CWP) process now underway in collaboration with the HEP Software Foundation (HSF).
2. possible areas of collaboration between HEP and CS in the context of a US-based Scientific Software Innovation Institute for HEP (S2I2-HEP), including particular areas where interests and expertise at US universities would enable key contributions to tackle software challenges for the HL-LHC.
3. specific opportunities and challenges in accomplishing the first two goals.

Attendees

Attendance at the workshop was by invitation by the organizers, working from a list of HEP and CS researchers generated in advance. Many of those invited from the HEP community expressed interest in S2I2-HEP by providing a letter of collaboration for the conceptualization proposal. Those invited from the CS community were identified to have leadership experience, interests, and activities aligned with relevant scientific software areas. The workshop had 50 attendees, with an approximately equal proportion of HEP physicists and those outside of HEP, primarily computer scientists and computer professionals. Of the 50 attendees, 24 held a faculty position, 21 held a staff position, three were postdoctoral fellows, and two were graduate students. Of the 40 attendees with a Ph.D. degree, more than half (28) received their degree in the last 20 years and 15 received their degree within the last 10 years. The full distribution is shown below:

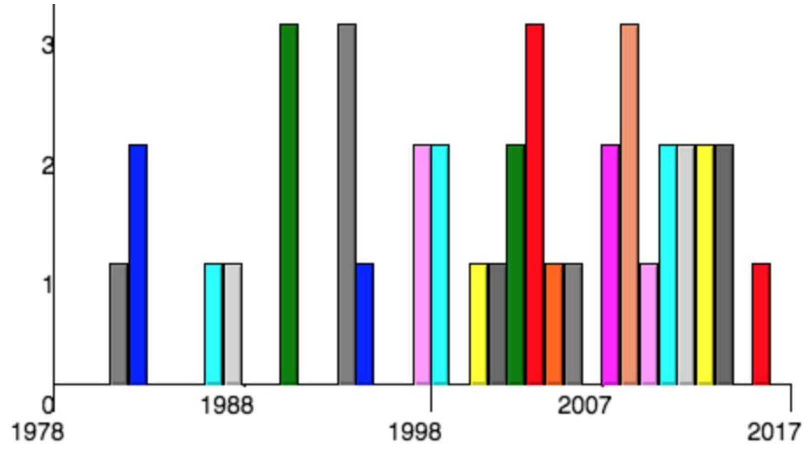

Figure: Year of Ph.D. for attendees with a Ph.D. degree in physics or computer science

The workshop attendees were:

D. Adrovic (DePaul), A. Aurisano (Cincinnati), L. Bauerdick (FNAL), P. Calafiura (LBNL), J. Carver (Alabama), K. Chard (Chicago), K. De (Texas, Arlington), A. Elliott (Aegis Research Labs), P. Elmer (Princeton), N. Ernst (Software Engineering Institute), A. Farbin (Texas, Arlington), M. Feickert (Southern Methodist), R. Gardner (Chicago), S. Gesing (Notre Dame), S. Gleyzer (Florida), D. Greenwood (Louisiana Tech), T. Hacker (Purdue), B. Hooberman (Illinois), K. Huff (Illinois), S. Jha (Rutgers), R. Kalescky (Southern Methodist), D. Katz (Illinois/NCSA), J. Kowalkowski (FNAL), A. Kumar (Southern Methodist), D. Lange (Princeton), D. Lesny (Illinois), M. Livny (Wisconsin), C. Maltzahn (Santa Cruz), S. Mannam (DePaul), A. Majumder (Wayne State), S. McKee (Michigan), M. Neubauer (Illinois), N. Niu (Cincinnati), P. Onyisi (Texas), D.Petravick (Illinois/NCSA), M. Paterno (FNAL), J. Pivarski (Princeton), J. Porter (LBNL), B. Riedel (Chicago), H. Schreiner III (Cincinnati), S. Seidel (New Mexico), E. Sexton-Kennedy (FNAL), M. Sokoloff (Cincinnati), D. Thain (Notre Dame), M.Turk (Illinois), J. Towns (Illinois/NCSA), I. Vukotic (Chicago), J. Wozniak (ANL), F. Wuerthwein (San Diego) and M. Zhang (Illinois).

Photo: Attendees of the S2I2-HEP/CS Workshop at the University of Illinois in December 2016

In advance of the workshop, each participant was requested to fill out a brief profile of themselves and their interests. The requested information included a basic profile, contact information, expertise, and research problems of interest both now and in the future. Requesting profiles from participants was motivated by the stated workshop goals and was useful to help “break the ice” and nucleate collaboration between HEP and CS members, many of whom had not met before this workshop. The profiles, which more than 90% of the participants completed, can be found at

http://s2i2-hep.org/downloads/s2i2-hep-cs-participants.pdf

Workshop Themes

While CS experience and expertise has been brought into the HEP community over the years, the aim of this workshop was to take a fresh look at current and planned HEP and CS research and brainstorm about engaging specific areas of effort, perspectives, synergies, and expertise of mutual benefit to HEP and CS communities, especially as related to a future NSF Software Institute for HEP.

In this context and with the previously stated goals in mind, the workshop was structured to have topical presentations and breakout sessions that considered a couple of themes with associated questions:

Physics-driven Software and Computing challenges in HEP for the HL-LHC era

What are the current HEP software and computing activities and why does our science demand these? What are the key software challenges to realize the full physics potential of the HL-LHC running? In what ways is the HL-LHC not just a simple extension (i.e., factor of 100 more data) of Run 2, and how do these inform the required software development?

Guidance from NSF Strategic Vision and Other Scientific Software Institutes

What is the strategic vision for the NSF Software Institutes within the ACI software program and how does this inform our S2I2-HEP conceptualization process? What lessons can be learned from other S2I2 conceptualizations, including the successfully funded Software Institutes?

Opportunities and Challenges for Future HEP and CS Collaboration

What are examples of successful HEP/CS collaborations and what properties have driven their success? How can a broader slice of the CS community be engaged with scientific computing when a commonly held belief in the CS community is that is a “niche” research area? What are the incentives for future HEP and CS collaboration in both incentive directions? How can CS research mechanisms (three-year grants, student developers, conference publications) be aligned with longer-term needs of big science (30-year project timelines, production software requirements, journal publications) like the LHC? How might a HEP Software Institute facilitate fruitful interactions between the HEP and CS communities?

Scope of S2I2-HEP Activities and Strategic Plan Development

What CS technologies, techniques, and trends could the HEP community adopt rather than (re-)invent, keeping in mind the long time scales and production needs of HEP software? What are the areas that S2I2-HEP should play a leading role in, informed by activities and interests within the US HEP and CS academic institutions?

Summary of Workshop Activities

The workshop was organized around a series of topical presentations and discussion sessions involving smaller groups of attendees. The first and third days were held at NCSA and the second day was held in Loomis Laboratory of Physics.

An organizational goal for the workshop was to balance time spent on presentations to inform and inspire with the time allotted for the discussions designed to generate the primary intellectual products of the workshop. This organizational goal was achieved by starting the workshop with an afternoon session of presentations on the first day followed by a second day comprised of breakout sessions interleaved with topical presentations.

A brief summary of the workshop activities is as follows:

Day 1 (December 7):

● **Afternoon**:	Topical presentations
● **Evening**:	Welcome Reception

Day 2 (December 8):

● **Morning**:	Topical presentations Breakout sessions
● **Afternoon**:	Topical presentations Breakout sessions
● **Early Evening**:	Tour of Blue Waters Supercomputer (30 people attended)
● **Evening**:	Workshop Banquet

Day 3 (December 9):

● **Morning**:	Summary and Closeout NCSA Colloquium on the Engineering of Scientific Software

The detailed agenda can be found at https://indico.cern.ch/event/575443,

Topical Presentations

The following presentations were given at the workshop:

- ***Welcome and Introduction*** (M. Neubauer)
- ***Overview of Computing in HEP*** (F. Wurthwein)
- ***NSF/ACI Software Programs and its Perspectives on Software Institutes*** (R. Ramnath)
- ***S2I2-HEP Conceptualization and the Community White Paper*** (P. Elmer)
- ***S2I2: Molecular Sciences Institute*** (S. Jha)
- ***S2I2: Science Gateways Institute*** (S. Gesing)
- ***Future Software Requirements*** (N. Ernst)
- ***Machine Learning*** (S. Gleyzer and A. Aurisano)
- ***CS & HEP Partnership: Past, Present and Future*** (M. Livny)
- ***What we have Learned about using Software Engineering Practices in Scientific Software*** (J. Carver) - NCSA Colloquium

The topical presentations were very informative and generated a great deal of discussion that was helpful for the conceptualization phase of the S2I2-HEP Project. All of the slides were posted to the workshop agenda page for to archive the content.

Breakout Sessions

On Day 2 of the workshop, held in the Physics Department (Loomis Laboratory), participants assembled into a large room configured to have four rectangular seating arrangements at each corner of the room. The purpose of this configuration was to break the workshop up into four equally-sized groups of about 12 people for discussions during breakout sessions. Each group was tasked with identifying technical and organizational challenges of a Scientific Software Institute, and proposing ways in which an Institute could be structured to address them. All participants contributed actively in these sessions to develop a set of questions and proposed responses to the questions. The room allowed content to be projected simultaneously on four screens throughout the room, so that the topical presentations could be seamlessly interleaved with the breakout sessions without need for the participants to move to a different room.

The morning breakout sessions were dedicated to developing key questions that were to be addressed in the afternoon breakout sessions. Each discussion group was referred to as a “Pod” and labeled as “Pod A”, “Pod B”, “Pod C” or “Pod D”. Before the morning breakout sessions, the membership of each Pod was randomized to avoid the natural “clumping” attendees by HEP or CS as they chose their seats (unrestricted) upon arrival in the room. Each Pod was given a link to its own Google Document which contained some seed questions from the Organizers but were asked to create a set of questions on their own in addition to listing the names of people in their Pod. Each Pod was asked to identify a "Pod Leader" that emerged naturally from their discussions who would be responsible for clearly articulating the questions generated by their Pod.

They were many excellent questions produced by the Pods and insightful comments embedded the each of the documents. During lunch, the Organizers took the Pods' content from the morning breakout sessions and used this to produce a set of five questions that could be answered in the time allotted for the afternoon breakout session. At the start of the afternoon breakout session, the membership of each Pod was again randomized so that the participants were required to interact with a new set of participants, this time in answering the same five questions that were previously referenced.

The full, unedited content from the Google docs generated in the Thursday breakout sessions by the four Pods can be found at:

http://s2i2-hep.org/downloads/s2i2-hep-cs-pods-questions-answers.pdf

In this report, we list the five questions given to each of the Pods to answer in the Thursday afternoon breakout session along with a summary of the answers and some discussion.

1) *How could we proceed put together a document in the next 6 months summarizing HEP computing challenges in a language that CS people understand and map it to established discipline areas in CS? (useful for developing future synergistic and collaborative projects/relationships with CS faculty?)*

It was broadly expressed that the CS community should be brought into the fold as early as possible as a true partner in the process of producing such a document and that this process should be an iterative one. In this way, they can understand the challenges early on and help frame the problems in a manner that maps on to CS research and solutions. It was suggested the HEP community recruit some CS researchers who would be willing to partner with them during the CWP process and invite the CS community to give early feedback on the CWP process.

One suggestion was that it might be beneficial to dedicate a future S2I2-HEP workshop to this topic with sufficient time to understand the challenges and develop strategies for potential solutions. It was also asked if there was the inverse case where the CS

This work was supported by NSF grants ACI-1558216, ACI-1558219 and ACI-1558233

community sought out domain scientists to engineer scientific software. An example of a workshop along these lines was given of a "Software Engineering Workshop for Sciences" (http://se4science.org/workshops/se4science17).

Another suggestion was to start from a list of problems together with priorities, which would help create a mapping between problems and the CS domain that is most relevant to addressing the given challenges.

2) *What are the incentives for such collaboration for HEP people? For CS people? For non-CS people? E.g. recognition, funding, publications, students, new problems to solve, new places to apply technologies, new solutions to current problems, pride in working on a global-scale problem. How could an S2I2-HEP institute create the relevant incentives and promote HEP/CS research collaborations?*

The groups identified several HEP and CS incentives to answer this question. HEP incentives included CS expertise and new ideas, access to people with skillsets different from their own, guidance (e.g., should we really build what we think we want to build?), and CS perspectives on industry trends and tools that can advance HEP research. CS incentives included handling of large data sets and challenging practical problems, Ph.D. and Masters theses, internship, research visits, future funding opportunities for CS proposal (domain applications), interdisciplinary research, and contributing to "cool" science.

Regarding the second question about S2I2-HEP creating the relevant incentives, responses included funding for people and projects at the HEP-CS computing interface, a mechanism to facilitate credit for software work, and fellowships.

It was commented that the main concepts behind the MoISSI (Molecular Sciences Institute) fellowships make sense for an S2I2-HEP as long as high standards of quality and relevance to the scope of the Institute are maintained, which requires strong supervision and mentoring. In the MoISSI, this oversight is planned by attaching 10% of a staff software engineer to each and every Fellow. Additionally, the ultimate success and eventual impact of a fellowship program will depend on the Fellows visibly advancing in their careers (both inside and outside of HEP) through the work and research they did as a Fellow.

3) *What can an S2I2-HEP institute do to create an environment of increased communication and awareness by individual HEP and CS researchers of each other's problems, expertise and research interests?*

In answering this question, groups came up with a pretty comprehensive list of suggested activities for the Institute. These included providing sabbatical support for visiting faculty, curating an overlay journal, editing/continuing the Big Data Science Springer journal (HSF), organizing new workshops (e.g., Hackathons) and strengthening existing ones (e.g., CHEP), instituting management structures (e.g., a Steering Committee) with HEP and CS researchers, developing CS curriculum applicable to HEP graduate students, supporting people to attend external conferences, hosting datasets of broad interest, and creating challenge problems at the intersection of HEP and CS. It was also suggested that the Institute could help facilitate the sociological elements of trust, respect, and affirmative acknowledgement of mutual and disparate disciplinary needs.

4) *Will HEP have anything interesting to offer in 5–10 years for CS researchers?*

All of the groups gave a strong “yes” answer to this question. To date, CS-HEP collaborations have focused primarily on the aggregation of computational power and the movement of data. With this raw computational ability now in production, a number of new challenges emerge on the 5-10 year time scale. Responses to this question included robustness and performance of large distributed systems within a scientific research ecosystem; practical challenges of storage, discovery, preservation, and management of Exabyte-scale data; an arena to apply new CS ideas (e.g., in Machine Learning), use of FPGAs for “in-flight” data analysis, data on the scientific software development process for “anthropological” studies, and a case study in industry-scale academic openness.

5) *The S2I2-HEP will not be trying to solve all problems for HL-LHC or HEP for that matter. Rather, it will be laying out a set of software activities for US Institutions for which the US can play a leading role. What are the areas in which the S2I2-HEP should play a leading role, informed by activities and interests within the US HEP and US CS communities?*

The groups came up with a number of areas in which the US HEP and CS communities are particularly strong and could play a leading role in addressing important problems for the HL-LHC era. This included Machine Learning, Distributed Computing, Workload Management, Data Management and Delivery, Remote Data Access, Data-Intensive Analysis Tools, Languages, Advanced Computing Architectures, Immersive Technologies (e.g. Virtual Reality), Outreach and Education, and Professional Training.

This work was supported by NSF grants ACI-1558216, ACI-1558219 and ACI-1558233

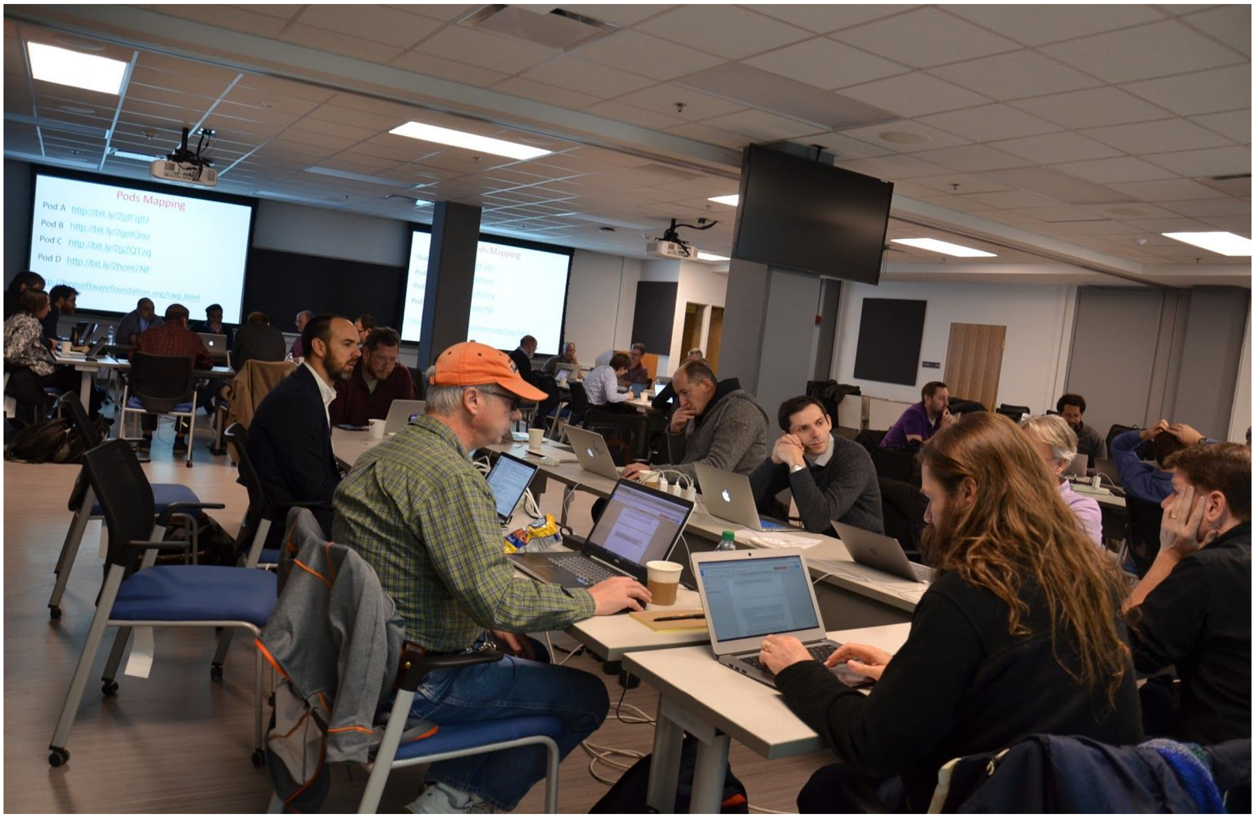

Photo: Pods in action at the S2I2-HEP/CS Workshop at the University of Illinois in December 2016. Each of four 12-member groups were asked to develop questions about conceptualization of a HEP Software Institute and answer a set of five questions in an afternoon session on Day 2

Summary and Closeout

On the last day of the workshop, each Pod Leader presented their answers to the five questions asked of them by the organizers. These five questions were a refined subset of the questions developed by the Pods in the morning session. A summary of the answers to these five questions can be found in the previous section of this report.

Conclusions and Next Steps

The workshop was very productive and successful in achieving the goals set forth by the organizers before the workshop. The “Pods” format in a single room for breakout sessions was highly successful as means to brainstorm ideas and achieve a productive dialog between HEP and CS researches, as demonstrated by the extensive body of work on questions and answers development in the Pods Google docs. We think this type of format could be very useful in future S2I2-HEP and CWP Workshops.

As a practical matter, this workshop succeeded in beginning the process of fostering collaboration between HEP and CS communities by bringing together a diverse set of attendees from these communities to develop an understanding of how the they could work together in the context of a future NSF Software Institute aimed at supporting the software the enables particle physics research over the long term. This workshop represented an important start to the S2I2-HEP conceptualization process, but it is still only a start.

Some reflection is warranted on a couple of aspects of the workshop. It is clear from the discussions that there is a divide to be bridged between the two communities before a true partnership in solving the most important challenges through an Institute can proceed efficiently. The CS community is large, with a diverse set of expertise, interests, and goals that drive their research activities and focus. In this regard, the HEP community comes off to them as a bit naive, such that more effort is needed to really understand what drives the CS community to collaborate with physicists on tackling their problems for mutual benefit. It is clear that the CS community is not interested in being used to solve HEP's problems for sole benefit of HEP (or any science domain), but rather as a respected partner that is brought in early in the process to help understand the challenges and devise strategies to tackle them that they find intellectually interesting and worthy of devoting research time of themselves and their students. In some cases, the best solution to HEP problems will require rethinking current practices and being open to change. The workshop discussion brought many these questions to the surface and in this regard it was very enlightening and productive.

Regarding the development of a Strategic Plan for a potential HEP Scientific Software Institute, more work is needed on developing the details of what activities such an Institute would propose and how the process will proceed and converge. This is expected, given that this was a kick-off workshop and it is still very early in the conceptualization phase of the S2I2-HEP project. Some care is also needed in being sure that the right set of HEP and CS partners broadly engaged from US Universities are represented in the conceptualization process.

The next step in the process is the CWP kick-off Workshop at UC San Diego in late January. Through the CWP process, the broad scope of required software and computing challenges will be identified and prioritized, which is a critical ingredient to the S2I2-HEP Conceptualization process.