Conceptualization of an S2I2 Institute for High Energy Physics

Peter Elmer (Princeton University)
Mike Sokoloff (University of Cincinnati)
Mark Neubauer (University of Illinois at Urbana-Champaign)

August 10, 2015

1 Overview and Objectives

Advanced software plays a fundamental role for large scientific projects - from designing experimental instruments to acquiring, reducing, and analyzing the data. In such projects, success requires large-scale collaboration; software is the glue which enables teams of researchers to work together to exploit accelerators, telescopes and other large scientific instruments. Building the requisite software is technically challenging because of rapid evolution of computing technologies and increases in data volume, requiring ever more sophisticated data processing and analysis. Individual projects, experiments and researchers are typically organized to succeed at specific goals. Significant organizational, funding, and career advancement challenges exist, especially for university researchers, to creating and sustaining software for an entire research community. The primary goal of the proposed project is to produce a well-defined strategy for developing software and computing models for use in high energy physics (HEP), particularly for the experiments collecting the very large data sets anticipated in the "High-Luminosity Large Hadron Collider" (HL-LHC) era. The community process envisioned will identify potential areas where U.S. university personnel can lead in key areas of software development to help realize the full potential of the HL-LHC program.

The quest to understand the fundamental building blocks of nature, and their interactions, is one of the most ambitious and enduring of human endeavors. Facilities such as the LHC, where our research is conducted, represent a leap forward in our ability to answer these questions. The discovery of the Higgs boson, the observation of exceedingly rare decays of B mesons, and strong constraints on possible theories of physics beyond the Standard Model (SM) of particle physics demonstrate the strong scientific reach of these experiments. Despite these achievements, fundamental questions remain, including: Why does nature express the symmetries embodied in the SM, and not other equally elegant symmetries? Why are there (only) three generations of basic building blocks of matter? Why are the masses of these building blocks so different from each other, both within a generation and between generations? What is the dark matter which pervades the Universe? Why is matter so dominant over antimatter in the Universe? Does space-time have additional symmetries or extend beyond the 3+1 dimensions of which we know? What mechanism stabilizes the Higgs mass from large quantum corrections at high energy? Are neutrinos their own anti-particles? Can gravity and quantum mechanics be described in a consistent theoretical framework? During Run 1 at the LHC, the ATLAS and CMS detectors stored data from integrated luminosities of 30 fb^{-1} using almost 150 PB of disk and tape storage each. During the HL-LHC era, these experiments plan to record data from 100 times as many collisions. The LHCb experiment operates at lower luminosity, but records data from a much larger fraction of its interactions, to study rare decays and matter-antimatter asymmetries. Already in Run 3 (starting in 2020), it will process tens of exabytes of raw data each year using a pure software trigger. Processing and analyzing multi-exabyte scale data sets requires new approaches to software and computing.

To help define a computing and software strategy for the HL-LHC era, we will organize working groups, workshops, and other discussions with participants from the HEP, scientific software engineering and computer science communities. The first major product of the project will be a *Community White Paper* (CWP) that identifies the key issues of computing infrastructure and software R&D required to enable full realization of the scientific potential of the large investment in upgraded detectors for the HL-LHC. The LHC experiments utilize state-of-the-art instrumentation built and operated by large international collaborations. The U.S. is the largest national group working in ATLAS and CMS, and it is positioned to make key contributions to the next generation of software infrastructure as well as to upgraded detectors, operations, and data analysis. In addition to the CWP, we will prepare a *Strategic Plan* which identifies areas where U.S. university-based researchers can lead important software infrastructure efforts that will complement those led by U.S. national laboratory-based researchers and international collaborators. This will provide the basis for a subsequent "Scientific Software Innovation Institute" (S^2I^2) implementation proposal.

2 Physics Context

The Standard Model (SM) of particle physics is a quantum field theory of quarks, leptons, and their interactions. It provides a coherent framework for understanding all experimental data, to date, at the shortest distance scales. The Higgs boson discovery in 2012 [1,2] answered the last open question of the Standard Model. It experimentally established the mechanism for electroweak symmetry breaking and confirmed that fundamental particles acquire their masses via their interactions with the Higgs field. However, the most interesting fundamental physics questions remain wide open.

The SM is incomplete: the "ordinary" matter it describes accounts for only 4.9% of the mass-energy budget of the universe while dark matter, which interacts with ordinary matter gravitationally, accounts for 26.8%. While we know something about dark matter at macroscopic scales, we know nothing about its microscopic, quantum nature, *except* that its particles are not found in the SM and they lack electromagnetic and SM nuclear interactions. This has motivated a large number of SM extensions using the formalism of quantum field theory to describe physics beyond the Standard Model (BSM physics). Constraints on BSM physics come from "conventional" HEP experiments plus others searching for dark matter particles either directly or indirectly.

BSM physics also addresses a key feature of the observed universe: the apparent dominance of matter over anti-matter. The fundamental processes of leptogenesis and baryongenesis are not explained by the SM, nor is the required level of CP violation (the asymmetry between matter and anti-matter under charge and parity conjugation). This motivates a diverse set of experiments with both quarks and neutrinos.

The LHC was designed to search for the Higgs boson and for BSM physics – goals in the realm of discovery science. The ATLAS and CMS detectors are optimized to observe and measure the direct production and decay of massive particles. They have now begun to measure the properties of the Higgs boson to test how well they accord with SM predictions.

Where ATLAS and CMS were designed to study high mass particles directly, LHCb was designed to study heavy flavor physics where quantum influences of very high mass particles are manifest in lower energy phenomena. Its primary goal is to look for BSM physics in CP violation (asymmetries in the decays of particles and their corresponding antiparticles) and rare decays of beauty and charm hadrons. As an example of how one can relate flavor physics to extensions of the SM, Isidori, Nir, and Perez [3] have considered model-independent BSM constraints from measurements of mixing and CP violation. They assume the new fields are heavier than SM fields and construct an effective theory. Then, they "analyze all realistic extensions of the SM in terms of a limited number of parameters (the coefficients of higher dimensional operators)." They determine bounds on an effective coupling strength of the interaction in the limit that the dimensionless couplings unity. The critical point of their results is that kaon, B_d, B_s, and D^0 mixing and CPV measurements provide powerful constraints that are complementary to each other and often constrain BSM physics more powerfully than direct searches for high mass particles.

The Particle Physics Project Prioritization Panel (P5) issued their *Strategic Plan for U.S. Particle Physics* [4] in late May, 2014. It was very quickly endorsed by the High Energy Physics Advisory Panel and submitted to the DOE and the NSF. The report says, *we have identified five compelling line of inquiry that show great promise for discovery over the next 10 to 20 years. These are the Science Drivers:*

- *Use the Higgs boson as a new tool for discovery*
- *Pursue the physics associated with neutrino mass*
- *Identify the new physics of dark matter*
- *Understand cosmic acceleration: dark matter and inflation*
- *Explore the unknown: new particles, interactions, and physical principles.*

The HL-LHC will address the first, third, and fifth of these using data acquired at twice the

energy of Run1 and with 100 times the luminosity. As the P5 report says,

The recently discovered Higgs boson is a form of matter never before observed, and it is mysterious. What principles determine its effects on other particles? How does it interact with neutrinos or with dark matter? Is there one Higgs particle or many? Is the new particle really fundamental, or is it composed of others? The Higgs boson offers a unique portal into the laws of nature, and it connects several areas of particle physics. Any small deviation in its expected properties would be a major breakthrough.

The full discovery potential of the Higgs will be unleashed by percent-level precision studies of the Higgs properties. The measurement of these properties is a top priority in the physics program of high-energy colliders. The Large Hadron Collider (LHC) will be the first laboratory to use the Higgs boson as a tool for discovery, initially with substantial higher energy running at 14 TeV, and then with ten times more data at the High- Luminosity LHC (HL-LHC). The HL-LHC has a compelling and comprehensive program that includes essential measurements of the Higgs properties.

In addition to HEP experiments, the LHC hosts the one of world's foremost nuclear physics experiments. "The ALICE Collaboration has built a dedicated heavy-ion detector to exploit the unique physics potential of nucleus-nucleus interactions at LHC energies. [Their] aim is to study the physics of strongly interacting matter at extreme energy densities, where the formation of a new phase of matter, the quark-gluon plasma, is expected. The existence of such a phase and its properties are key issues in QCD for the understanding of confinement and of chiral-symmetry restoration." [5] In particular, these collisions reproduce the temperatures and pressures of hadronic matter in the very early universe, so provide a unique window into the physics of that era.

Summary of Physics Motivation: The ATLAS and CMS collaborations published letters of intent to do experiments at the LHC in October 1992, a bit more than 20 years ago. At the time, the top quark had not yet be discovered; no one knew if the experiments would discover the Higgs boson, supersymmetry, technicolor, or something completely different. Looking forward, no one can say what will be discovered in the HL-LHC era. However, with data from 100 times the number of collisions recorded in Run 1, the next 20 years are likely to bring even greater discoveries.

3 Software and Computing Context

The primary goal of the S^2I^2 conceptualization process is to understand the software requirements and challenges related to the high-luminosity LHC and to prepare a strategic plan for addressing them. The DOE and NSF jointly invest $\approx$ \$35M/year in ATLAS and CMS software and computing, about half in hardware plus operations, about half in software professionals. The LHC funding agencies, worldwide, invest about \$150M/year in these enterprises. In 2014, the LHC experiments used almost 175 PB of tape storage and slightly more disk storage. The event rate anticipated for the HL-LHC era is 100 times greater, and even assuming the experiments significantly reduce the amount of data stored per event, the total size of the datasets will be well into the exabyte scale; they will be constrained primarily by costs and funding levels, not by scientific interest. One long-term goal of a HEP S^2I^2 will be maximizing the return-on-investment to enable break-through scientific discoveries using the HL-LHC detectors.

Software is ubiquitous in the process by which data is acquired, managed, processed and analyzed to produce physics results from experiments such as those at the LHC. For example, over the past 15 years, more than 2000 CMS and ATLAS collaborators have written software well in excess of 10M source lines of code (SLOC) to support their activities. This includes application-level (algorithmic) software to support for number of stages in the typical data flow:

- data acquisition (DAQ)

- high level trigger (HLT) - a software process, runs in (quasi-) real time to select a small subset of data to be processed offline
- calibration and alignment of the detectors
- reconstruction - pattern recognition and related data reduction algorithms, typically in common for an entire experiment
- data analysis - activities initiated by individuals or groups within the experiment to select data subsets and extract physics results

Also, and perhaps less obviously, simulations of detectors and physics are equally critical in preparing experiments and analyzing data. The number of nominal stages is similar. The CPU resources used to generate simulated data often match, or even exceed, those used to process real data offline. In addition, the scale of the (geographically distributed) computing infrastructure requires significant software development to support data, workflow and resource management.

Several major software sustainability challenges exist given the physics goals and 15-20 year time scale of HL-LHC: (1) the evolution of computing hardware (processors, storage, networks) and the need to find cost-effective solutions, (2) increased complexity of the data and/or detectors, and (3) increased sophistication of analyses required from larger datasets.

The challenges for processor technologies are well known [6]. While the number of transistors on integrated circuits doubles every two years (Moore's Law), power density limitations and aggregate power limitations have led to a situation where "conventional" sequential processors will be replaced by parallel architectures, with consequent implications for changes to the algorithms implemented in our software. Understanding how emerging architectures (from low power processors to parallel architectures like GPUs to specialized technologies like FPGAs) will allow HEP computing to realize the exponential growth in computing power required to achieve our science goals will be a critical element of any R&D effort. Similar challenges exist when storage and network at the scale of HL-LHC [7], with implications for the persistency and computing models and the software supporting them.

The software landscape itself is quite varied. The preparations for LHC have yielded important community software tools for data analysis like ROOT [8] and detector simulation GEANT4 [9,10], both of which have been critical not only for LHC but in most other areas of HEP and beyond. Other tools have been shared between some, but not all, experiments. Examples include the GAUDI [11] event processing framework, RooFit [12] for data modeling and fitting and the TMVA [13] toolkit for multivariate data analysis. Tools developed in collaboration with computer scientists include FRONTIER [14] for cached access to databases, XROOTD [15] and dCache [16] for distributed access to bulk file data, CERNVM-FS [17] for distributed and cached access to software, IgProf [18] for profiling very large C++ applications like those used in HEP, and the Parrot [19] virtual file system. In addition many software tools written by theorists are used heavily by the experiments, including event generators such as SHERPA [20] and ALPGEN [21] and jet finding tools like FastJet [22,23]. This is not meant to be an exhaustive list. The conceptualization process will need to produce a full software map, along with specific ideas of the sustainability challenges for each software package and relevance to the long-run HL-LHC and HEP programs.

More interesting is the 10M lines of code mentioned earlier that has been developed within individual experiments, often as a by-product of the physics research program and typically without with the explicit aim of producing sustainable software. While much of the code is experiment specific due to actual differences in the detectors and appropriate techniques, some of this is simply redundant development of different implementations of the same functionalities. Long term sustainability issues exist in many places in such code. One obvious example is the need to develop parallel algorithms and implementations for the increasingly computationally intensive charged particle track reconstruction. A challenge for the conceptualization process will be to forge effective collaborative efforts between the experiments to solve the software issues.

In developing a broad vision for HL-LHC computing, we need to understand how experts from the worlds of software engineering, data science, computer science, and project management can contribute to our efforts. The areas they can help us address include:

- developing "enterprise computing models" to balance investments in software engineering compared to hardware and to balance computing power versus storage requirements, also accounting for operating costs (personnel, wall power, cooling, real estate, etc.);
- determining the life-cycle cost of purchasing and maintaining hardware compared to buying resources from commercial cloud services;
- understanding the trade-offs between software optimized for individual experiments and software which is commonly developed and used, but optimized for no one.
- understanding how to develop and maintain sustainable software, including design, documentation, and training.
- understanding what tools already exist in other scientific disciplines or are being planned.

Several scales define the computing needs of the experiments. One of the simplest metrics is the size of the datasets. At CERN's Resource Review Board meeting in April, 2015 the Computing Resources Scrutiny Group (CRSG) reported their findings. The 2014 tape and disk storage requirements for each of the experiments is summarized in Table 1. In the HL-LHC era, the number of events produced in the detectors will increase by a factor of 100. The experiments are all financially constrained – while all want to record data from (at least) 100 times as many events, it will be necessary to record less data per event to fit into the anticipated budget envelopes. In addition to the data stored on tape and disk, the experiments will need to process proportionately more data in their high level triggers, and they will need to generate and analyze proportionately more simulated data. The CSRG made several comments and recommendations. The first two are:

Table 1: Mass storage used by the LHC experiments in 2014, corresponding to the full Run 1 statistics. Numbers extracted from the CRSG report, CERN-RBB-2015-014 [24]. During the HL-LHC era, the experiments plan to record data from 100 times as many collisions, although it will be necessary to store less data per event.

Experiment	Disk Usage (PB)	Tape Usage (PB)	Total (PB)
ALICE	20	15	35
ATLAS	86	73	159
CMS	59	69	128
LHCb	15	16	31
Total	180	173	**353**

1. As reported previously, the experiments' requests for Run 2 are made keeping in mind a flat funding profile (not adjusted for inflation). However, we see a tendency as Run 2 progresses for the experiments' requests to outstrip the growth that can be accommodated by this profile.
2. The CRSG strongly supports software engineering development and recommends that sufficient effort be funded to support this. Improving the efficiency of software, including making optimal use of new hardware designs is essential to mitigate the growth in resource use. There have been substantial improvements made for Run 2. *In the longer term, with orders-of-magnitude increases in the expected computing needs for Run 3, this work is even more essential* (emphasis added).

As a reminder, Run 2 is starting now (summer 2015), and Run 3 will start circa 2020 or 2021. Run 4 will mark the beginning of the HL-LHC era. Investing in sustainable and innovative software infrastructure to reap the benefits of the upgraded accelerator and detectors is necessary to achieve the promised scientific reach of these instruments.

4 Conceptualization Activities and the HEP Community

A HEP S^2I^2 with a focus on the physics of the HL-LHC will need to be part of a coherent international effort including U.S. university-based scientists, DOE-funded laboratory staff, and

scientists from outside the U.S. The questions to be addressed during the conceptualization process include both possible projects for an S^2I^2 as well as the wider community context for these efforts. The success of this conceptualization process requires engaging a diverse group of people from the LHC software and computing community, as well the more general HEP, scientific computing, and computer science worlds. The specific workshops and deliverables we will describe later (Sec. 5) take into account the need to understand both the global context and eventual specific S^2I^2 activities of the U.S. university community.

To prepare this proposal we have had preliminary discussions with many key individuals from the community. In particular we have made contact with all 4 LHC experiments, the CERN and FNAL software groups, the Open Science Grid (OSG) and the Worldwide LHC Computing Grid (WLCG), all of whom are significant producers or consumers of software within HEP. To build on the expertise of others, we have contacted the Software Sustainability Institute (SSI) in the UK and the NSF-sponsored "Data and Software Preservation for Open Science" (DASPOS) project. Several specific individuals from the Computer Science and HEP Theory communities have also agreed to take the lead in helping us identify and bring on board others from those communities. Last, but not least, we have engaged a large number of U.S. university groups, as well as a handful of international groups, all with interests and experience in software and computing.

An aggregate description of those we have contacted and who have already committed to participation in various ways in the conceptualization process is included as "Other resources" in the Princeton Facilities Plan. Specific details are in the individual letters of commitment. Elmer, Neubauer and Sokoloff will act as the initial Steering Committee to guide the process. Starting from these key players, during the first three months we will organize the workshops and working groups with specific charges for the kick-off workshop. The set of key players we chose was meant to guarantee that we had a basic set of collaborators representing all of the relevant areas. If this proposal is funded, we will begin an even broader process to involve our community. We include here some additional notes on specific entities within the community.

HEP Software Foundation (HSF): A recent effort within our community is the HEP Software Foundation (HSF) [25]. The HSF aims to "facilitate coordination and common efforts in high energy physics software and computing internationally". Several workshops have been held over the past year at CERN, SLAC and at the CHEP 2015 conference in Okinawa [26–28]. More than ∼200 individuals have participated, representing most HEP experiments and labs. The significant interest in these initial HSF workshops demonstrates recognition by the community of the challenges ahead and a widespread interest and motivation for finding common software solutions.

The HSF is an "umbrella" for community activities, not a project itself. Eventual substance is to be provided as an aggregate of individually funded community-oriented projects and initiatives, in a fashion similar to the Apache Software Foundation [29]. Two recent examples of such projects are the NSF SI2-funded DIANA project [30] and the software work package in the EU-funded AIDA 2020 project [31], both of which feature cross-experiment activities aimed at producing community software. In addition several topical working groups have been formed on software licenses, software packaging, and educational and training activities. Although the HSF aims more broadly in HEP than the HL-LHC, its goals are completely consistent with the proposed S^2I^2 conceptualization process. The S^2I^2 conceptualization workshops we propose will themselves be co-branded as HSF workshops and build on and accelerate the work done thus far by HSF.

DOE HEP Forum for Computing Excellence (HEP-FCE): Following the Snowmass computing activities [32] the DOE organized a workshop "Topical Panel Meeting on Computing and Simulations in High Energy Physics". The resulting report [33], along with the P5 panel report [4], led to the creation of a "Forum for Computing Excellence" (HEP-FCE) to coordinate efforts between the DOE labs. The HEP-FCE aims at at supporting not only HL-LHC (Energy Frontier) activities, but also the cosmic and intensity frontiers. The HEP-FCE is very much in the formative

stage and its co-leads have committed to participating in the activities described here. The collaboration of HEP-FCE with the NSF S^2I^2 Conceptualization process should lead to a coherent and complementary plan for the principal U.S. entities.

5 Deliverables and Timeline

There are two major deliverables for the S^2I^2 conceptualization process:

(1) A **Community White Paper (CWP)** describing a global vision for software and computing for the HL-LHC era; this will include discussions of elements that are common to the LHC community as a whole and those that are specific to the individual experiments. It will also discuss the relationship of the common elements to the broader HEP and scientific computing communities. Many of the topics discussed here will address issues required for a HEP S^2I^2 implementation proposal, but which are also generic to the larger community. These will include

- a broad overview of the grand challenge science of the HL-LHC;
- how new approaches to computing and software can enable and radically extend the physics reach of the detectors;
- what computing and software research will be required so that computing and software Technical Design Reports can be prepared several years before Run 4 of the LHC begins; this will include studies of hardware and software architectures and life-cycle processes and costs.
- identify specific software elements and frameworks that will be required for the HL-LHC era which can be built and tested during Run 3.
- organizational issues for the common software and for coordinating research of common interest, even when the final products will be specific to individual experiments.
- software development and documentation tools for writing sustainable software;
- identifying and mitigating potential risks

This document will be written with our DOE Laboratory and foreign colleagues. In addition to providing material for possible S^2I^2 proposals, it will provide a roadmap for members of the international HL-LHC community to approach their funding agencies. A first draft of this document should be ready by the end of first year of the grant.

(2) A separate **Strategic Plan** identifying areas where the U.S. university community can provide leadership and discussing those issues required for an S^2I^2 which are not (necessarily) relevant to the large community; topics will include:

- where does the U.S. university community already have expertise and important leadership roles;
- which software elements and frameworks would provide the best educational and training opportunities for students and postdoctoral fellows;
- what types of programs (short courses, short-term fellowships, long-term fellowships, etc.) might enhance the educational reach of an S^2I^2;
- possible organizational, personnel and management structures and operational processes;
- how the investment in an S^2I^2 can be judged and how the investment can be sustained to assure the scientific goals of the HL-LHC.

This document will be prepared by members of the U.S. university community. Broad participation in working groups will be encouraged. Although it is not a project deliverable, a goal of the conceptualization exercise will be building a team for submitting an S^2I^2 implementation proposal, should there be an appropriate solicitation. This plan will be completed 15 months after the beginning of the conceptualization project.

5.1 Timeline

The timeline for executing the project will have *three phases* over approximately 15 months:

Project Establishment (October 2015 - January 2016): During this phase, we will produce a detailed framework for the project. It will culminate in a "kick-off" workshop where the participants will develop detailed plans for subsequent workshops, prepare an outline of the CWP, and establish a clear set of expectations for the community. Leading up to this, the PIs will identify conveners for working groups, develop draft charges for the working groups (to be refined during the first workshop), and make initial plans for the later workshops. Most of this will be done "virtually", (i.e., via email/phone/video-conferencing). We may also invite a small number of participants to a short, in-person meeting of area conveners to prepare a detailed agenda for the first workshop and templates for working groups to use in writing their chapters of the CWP. We tentatively plan to hold the first workshop in January 2016 at the San Diego Supercomputing Center. Its primary purpose will be to focus on "big picture" software and computing questions for the HL-LHC era and define detailed charges to the working groups for the topical workshops.

Community White Paper (CWP) Preparation (January 2016 - August 2016): The primary goal of this phase of the project will be producing the CWP. Its content will be developed by working groups whose activities will generally correlate with particular chapters. The working groups will generate ideas and questions in advance of three *topical workshops* (discussed in Section 5.2), to be held in the first half of 2016. At the end of each of these, the participating working groups will identify specific software elements needed for the HL-LHC era and prepare preliminary reports identifying how development might proceed. Between the workshops, the working groups will meet "virtually" to flesh-out and polish their reports. The final workshop will be held in summer 2016 at or near CERN. The working groups will present their findings to the full group for discussion and integration into coherent, draft CWP. The PIs will then serve as primary editors to produce a publicly distributed document.

Strategic Planning (September 2016 - December 2016): During the final phase of the project, we will work (primarily) with U.S. University personnel to prepare a *Strategic Plan* for a possible NSF-funded scientific software institute to enable both basic science and education. The scope of possible projects will be coordinated with other stake-holders, including DOE-funded laboratories in the U.S. and international collaborators, to assure a coherent community-wide effort. In addition to identifying areas where the U.S. university-based community has special interests and strengths, the plan will identify possible management structures, mechanisms for continuing assessment, and appropriate education and outreach activities. The plan will specifically address all issues identified in the program solicitation, including those not discussed in detail already:

- development, testing and deployment methodologies, validation and verification processes, end usability and interface consideration, and required infrastructure and technologies;
- the requirements and necessary mechanisms for human resource development, including integration of education and training, mentoring of students, postdoctoral fellows as well as software professionals, and proactively addressing diversity and broadening participation;
- potential risks including risks associated with establishment and execution, necessary infrastructure and associated technologies, community engagement, and long-term sustainability.

As in the Project Establishment phase, we expect to meet with participants "virtually". If necessary, we will have face-to-face meetings with specific U.S. university groups.

At each stage, the process will be as open as possible. The initial workshop workshop where the specific issues to be addressed are defined, and the three intervening workshops, will be held

in the U.S. to make it as easy as possible for the broad U.S. community (physicists, computer scientists, and software engineers) to participate in the detailed planning which will be done at those meetings. The choice of location for the final workshop at, or near CERN, is to encourage the widest possible participation by our international partners. Even before the first workshop, we will establish a portal on the web, similar in spirit to that of the P5 panel [34], to communicate with the broad HEP and related computing communities.

Although the focus of the work will be to tackle issues related to ensuring the scientific success of the HL-LHC program, many of the issues relate to other scientific projects which will collect very large data sets. Between the time the CWP is prepared and the Strategic Plan is written, we will again reach out to the U.S. community for reactions to the CWP and contributions to the Strategic Plan. A preliminary draft of each document will be distributed at least a month before the final draft is formally submitted.

5.2 Topical Workshops

We plan to hold three topical workshops. Each will bring together members of specific working groups to address questions developed in the kick-off workshop and via virtual meetings. At the end of the workshops, each group will produce a written report to be used as material for the CWP. The precise definitions of working groups and topics for specific workshops will be determined in the Project Establishment phase. Our role will be to coordinate and guide the process. We have already had extensive discussions with people who are interested in participating in the process, and have identified a number of broad areas which serve as as a starting point pending further discussion. With these considerations, we propose a possible set of topical workshops (in no particular order) along with a description of some of the issues that might be discussed:

Workshop 1: *Detector Simulation, Triggering, Event Reconstruction and Visualization*
Challenges surrounding high pile-up simulation, including the CPU resources needed for large statistics samples needed to compare with data from high trigger rates, high memory utilization, generation and handling of the large (min-bias) samples needed to achieve accurate description of high pile-up collision events, and a flexible simulation strategy capable of a broad spectrum of precision in the detector response, from "fast" (e.g. parametric) simulation optimized for speed to full simulation in support of precision measurements and new physics searches (e.g. in subtle effects on event kinematics due to the presence of virtual particles at high scale). Software required to emulate upgraded detectors (including the trigger system) and support determination of their optimal configuration and calibration. Software in support of triggering during the HL-LHC, including algorithms for the High-level Trigger, online tracking using GPUs and/or FPGAs, trigger steering, event building, data "parking" (for offline trigger decision), and data flow control systems. New approaches to event reconstruction, in which the processing time depends sensitively on instantaneous luminosity, including advanced algorithms, vectorization, and execution concurrency and frameworks that exploit many-core architectures. In particular, charged particle tracking is expected to dominate the event processing time under high pile-up conditions. Visualization tools, not only in support of upgrade detector configurations and event displays, but also as a research tool for data analysis, education, and outreach using modern tools and technologies for 3D rendering, data and geometry description and cloud environments.

Workshop 2: *Data Access and Management, Workflow and Resource Management*
Data handling systems that scale to the Exabyte level during the HL-LHC era and satisfy the needs of physicists in terms of metadata and data access, distribution, and replication. Increasing availability of very high speed networks removes the need for CPU and data co-location and allows for more extensive use of data access over the wide-area network (WAN), providing failover capabilities, global data namespaces, and caching. Event-based data streaming as complementary to the more

traditional dataset-based or file-based data access, which is particularly important for utilizing opportunistic cycles on HPCs, cloud resources, and campus clusters where job eviction is frequent and stochastic. Workflow management systems capable of handling millions of jobs running on a large number of heterogeneous, distributed computing resources, with capabilities including whole-node scheduling, checkpointing, job rebrokering, and volunteer computing. Systems for measurement and monitoring of the networking bandwidth and latency between resource targets and the use of this information in job brokering. Software-defined networking technologies which enable networks to be configurable and schedulable resources for use in the movement of data.

Workshop 3: *Physics generators, Data Analysis and Interpretation, Data and Software Preservation*
There are many theory challenges in the HL-LHC era, among them are improving the precision of SM calculations, better estimation of systematic uncertainties, and elucidation of promising new physics signals for the experiments. Software needed to make connection between observations and theory include matrix element generators, calculation of higher-order QCD corrections, electroweak corrections, parton shower modeling, parton matching schemes, and soft gluon resummation methods. Physics generators that employ concurrency and exploit many-core architectures will play an important role in HL-LHC, as well better sharing of code and processing between LHC experimenters and phenomenologists. Data analysis frameworks that include parallelization, optimized event I/O, data caching, and WAN-based data access. Analysis software that employs advanced algorithms and efficiently utilizes many-core architectures. Tools and technologies for preservation and reuse of data and software, preservation and re-interpretation of physics results, analysis providence and workflow ontologies, analysis capture, and application packaging for platform abstraction. Future software repositories and build platforms that leverage advances in these areas and improved software modularity and quality control that will allow a broader community of people to effectively contribute to software in the HL-LHC era.

In addition to addressing topical issues, each group will be expected to address questions which cut across boundaries, including:

- What are the specific challenges for the HL-LHC?
- What opportunities exist to exploit new or advanced algorithms (e.g. deep learning)?
- How can emerging architectures improve the bang-per-buck and what software evolution is needed to exploit them?
- Which problems are specific to individual experiments and which are common to the HL-LHC experiments or to HEP and nuclear physics experiments more generally?
- What is required to make common software packages sustainable?

6 Broader Impacts of the Proposed Work

The focus of the CWP will be an overall software and computing plan for the HEP experiments and ALICE in the HL-LHC era. It will inform thinking for the preceding "phase-l" upgrade era at the LHC similarly, and for other scientific experiments collecting very large data sets starting circa 5 years from now. Many of the issues related to effectively using emerging architectures, providing access to highly distributed data, and developing sustainable software will be of general interest. Both the CWP and the Strategic Plan will be published in open repositories (such as the arXiv) and posted on the project's portal. In addition to addressing overarching issues, each of these will identify well-defined, self-contained projects that members of the broader HEP and computing communities can pursue. The scope of projects discussed in the CWP will be much greater than what an NSF-funded S^2I^2 can fund. The process of preparing it should encourage additional efforts

to address big-picture needs. In part, the Strategic Plan will also serve as a template for those seeking resources from other U.S. or foreign agencies, foundations, or private partners.

The HL-LHC is expected to take data starting more than 10 years from now, and continue for many years. Consequently, it is critical to engage today's younger scientists in the S^2I^2 process, starting with the conceptualization phase. To this end, at least a quarter of the participants attending the workshop with project funding will be graduate students or post-docs; at least another quarter will be relatively early career scientists, those who earned Ph.D.'s fewer than 20 years ago.

An important element of any S^2I^2 is developing sustainable software. The HEP community has a long history of writing and supporting community software. The CERN Program Library, also known as CERNLIB [35], dates back to the 1960s [36]. It is no longer maintained, but many of the algorithms and packages have migrated from their FORTRAN implementations to C++ libraries like ROOT [8] (a data analysis framework) and GEANT [9,10] (for detector simulation). In moving forward, we will actively engage with requirements engineering experts from the computer science world. This will provide mutual benefits. They will help the HEP community understand generally accepted best practices for designing software intended for use over extended periods as hardware, operating systems, and languages evolve. At the same time, we will grant them access to our community: they will be able to study the practice of large-scale scientific software development from the inception of the S^2I^2 process through the development of the software to its long-term use. By working directly with the practitioners, they will be able to better understand both the technical and social challenges encountered, and help develop real-world solutions. Not only will this help us, it will also allow them to incorporate what they learn into recommendations for other communities developing large scientific infrastructures.

7 Summary

The confirmation that all observed matter-antimatter asymmetries (CP violation) in the quark sector are dominated by the single Kobashyi-Maskawa phase of the CKM matrix, plus the discovery of Higgs boson by ATLAS and CMS, have cemented the role of the Standard Model of particle physics as *the* correct description of the electro-weak and strong nuclear interactions of leptons and quarks. The focus of research in the field is now BSM physics, including: what is dark matter? why are there (only) three generations of quarks and leptons? why do the masses of the fundamental particles vary so strongly, both within generations and from one generation to the next? are neutrinos their own antiparticles? does nature express additional symmetries? does spacetime extend past the three-plus-one dimensions we know? The LHC is poised to address many of these questions uniquely well, and others from unique perspectives. The accelerator energy was just upgraded from 8 TeV to 13 TeV and the luminosity is planned to increase so that the experiments can collect data from 100 times as many *pp* interactions in the high-luminosity era as were studied in Run 1. The increased reach for transformative scientific discoveries is similar to that achieved between the inception of the LHC and its discovery of the Higgs boson.

The LHC detectors are all being upgraded to read out and record the higher data rates anticipated. To keep up with these higher rates, the software and computing infrastructures must evolve similarly. During Run 1 the experiments stored a third of an exabyte of data on tape and disk. In Run 3, LHCb will process exabytes of data per year flowing out of the detector, into a pure software trigger; ATLAS and CMS will each store exabytes of data offline. Detailed projections for the HL-LHC era (Run 4 and beyond) are not yet credible, but the number of electronics channels reading out the detectors will increase, and the occupancies will increase at least in proportion to the luminosity. To realize the promise of higher rates and improved instruments, software and computing must improve in parallel. The challenges are multi-dimensional: they span the fields of physics, computer science, and software engineering; they range from low-level algorithms to data structures and frameworks; they include understanding how to use emerging architectures and how

to design and write sustainable software for an enterprise that will extend for 15 to 20 years.

Developing sets of questions and identifying software projects that are required for the LHC experiments to take full advantage of the investments being made in the accelerator and the detectors are the primary goals of the conceptualization project. In addition, the process of holding workshops and preparing the Community White Paper (CWP) and Strategic Plan will be a community-building exercise. It will provide a forum for U.S. university-based researchers to collaborate with DOE-funded laboratory researchers and peers from outside the United States. It will bring together physicists with computer scientists and software engineers. The HEP Software Foundation was recently created as a light-weight umbrella organization for groups from across the field to bring resources and share responsibilities for creating common tools. Building on this spirit, the conceptualization project we propose will provide a mechanism for moving forward coherently. The Strategic Plan will serve as a conceptual design for U.S. university-based researchers to prepare an S^2I^2 implementation proposal. The CWP will provide a bigger picture of the software required to ensure the success of the HL-LHC. It will also serve as a starting point for discussions about where other groups will focus their efforts, and it will help them prepare proposals to their funding agencies. With a bit of luck, it will help create an efficient international effort. The large number of letters of collaboration from individual collaborators, the experiments' computing managements, and members of the computer science, software engineering, and larger scientific computing communities inspires confidence that the project will achieve its goals.

References

[1] G. Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett.*, B716:1–29, 2012.

[2] Serguei Chatrchyan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.*, B716:30–61, 2012.

[3] Gino Isidori, Yosef Nir, and Gilad Perez. Flavor Physics Constraints for Physics Beyond the Standard Model. *Ann.Rev.Nucl.Part.Sci.*, 60:355, 2010.

[4] Particle Physics Project Prioritization Panel. Building for Discovery: Strategic Plan for U.S. Particle Physics in the Global Context. `http://science.energy.gov/~/media/hep/hepap/pdf/May%202014/FINAL_DRAFT2_P5Report_WEB_052114.pdf`.

[5] ALICE Collaboration public website. `http://aliceinfo.cern.ch/`.

[6] Samuel H. Fuller and Editors; Committee on Sustaining Growth in Computing Performance; National Research Council Lynette I. Millett. *The Future of Computing Performance: Game Over or Next Level?* The National Academies Press, 2011.

[7] M. Butler, R. Mount, and M. Hildreth. Snowmass 2013 Computing Frontier Storage and Data Management. *ArXiv e-prints*, November 2013.

[8] ROOT home page. `http://root.cern.ch/drupal/`.

[9] V.N. Ivanchenko. Geant4 toolkit for simulation of HEP experiments. *Nucl.Instrum.Meth.*, A502:666–668, 2003.

[10] John Allison, K. Amako, J. Apostolakis, H. Araujo, P.A. Dubois, et al. Geant4 developments and applications. *IEEE Trans.Nucl.Sci.*, 53:270, 2006.

[11] G. Barrand et al. GAUDI - The software architecture and framework for building LHCb data processing applications. In *Proceedings, 11th International Conference on Computing in High-Energy and Nuclear Physics (CHEP 2000)*, pages 92–95, 2000.

[12] Wouter Verkerke and David P. Kirkby. The RooFit toolkit for data modeling. *eConf*, C0303241:MOLT007, 2003.

[13] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[14] Kosyakov S. et al. FRONTIER: HIGH PERFORMANCE DATABASE ACCESS USING STANDARD WEB COMPONENTS IN A SCALABLE MULTI-TIER ARCHITECTURE. In *Proceedings, 14th International Conference on Computing in High-Energy and Nuclear Physics (CHEP 2004)*, 2004.

[15] A Dorigo, P Elmer, F Furano, and A Hanushevsky. XROOTD - A highly scalable architecture for data access. *WSEAS Transactions on Computers*, 4.3, 2005.

[16] Patrick Fuhrmann. dCache: the commodity cache. In *In Twelfth NASA Goddard and Twenty First IEEE Conference on Mass Storage Systems and Technologies*, 2004.

[17] Jakob Blomer, Carlos Aguado-Sanchez, Predrag Buncic, and Artem Harutyunyan. Distributing LHC application software and conditions databases using the CernVM file system. *Journal of Physics: Conference Series*, 331(4):042003, 2011.

[18] Eulisse G. and Tuura L. IgProf profiling tool. In *Proceedings, 14th International Conference on Computing in High-Energy and Nuclear Physics (CHEP 2004)*, 2004.

[19] Douglas Thain and Miron Livny. Parrot: Transparent user-level middleware for data-intensive computing. *Scalable Computing: Practice and Experience*, 6(3), 2005.

[20] T. Gleisberg, Stefan. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter. Event generation with SHERPA 1.1. *JHEP*, 02:007, 2009.

[21] Michelangelo L. Mangano, Fulvio Piccinini, Antonio D. Polosa, Mauro Moretti, and Roberto Pittau. ALPGEN, a generator for hard multiparton processes in hadronic collisions. *Journal of High Energy Physics*, 2003(07):001, 2003.

[22] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J.*, C72:1896, 2012.

[23] Matteo Cacciari and Gavin P. Salam. Dispelling the N^3 myth for the k_t jet-finder. *Phys. Lett.*, B641:57–61, 2006.

[24] Computing Resources Scrutiny Group Report, April 2015. `http://cds.cern.ch/record/2002240/files/CERN-RRB-2015-014.pdf`.

[25] HEP Software Foundation (HSF) website. `http://hepsoftwarefoundation.org`.

[26] HSF Workshop at CERN (April 2014). `https://indico.cern.ch/event/297652/`.

[27] HSF Workshop at SLAC (January 2015). `http://indico.cern.ch/event/357737/`.

[28] HSF Workshop at CHEP 2015 (April 2015). `https://indico.cern.ch/event/385984/`.

[29] Apache Software Foundation. `http://www.apache.org`.

[30] DIANA/HEP website. `http://diana-hep.org`.

[31] AIDA 2020 website. `http://aida2020.web.cern.ch`.

[32] L. A. T. Bauerdick, S. Gottlieb, G. Bell, K. Bloom, T. Blum, D. Brown, M. Butler, A. Connolly, E. Cormier, P. Elmer, M. Ernst, I. Fisk, G. Fuller, R. Gerber, S. Habib, M. Hildreth, S. Hoeche, D. Holmgren, C. Joshi, A. Mezzacappa, R. Mount, R. Pordes, B. Rebel, L. Reina, M. C. Sanchez, J. Shank, P. Spentzouris, A. Szalay, R. Van de Water, M. Wobisch, and S. Wolbers. Planning the Future of U.S. Particle Physics (Snowmass 2013): Chapter 9: Computing. *ArXiv e-prints*, January 2014.

[33] Report from the topical panel meeting on computing and simulations in high energy physics. `http://hepfce.org/files/2014/08/Computing-Meeting-Report_final.pdf`.

[34] Avery, Paul and Habib, Salman (co-Chairs) and Others. Computing in High Energy Physics. `http://science.energy.gov/~/media/hep/pdf/files/Banner%20PDFs/Computing_Meeting_Report_final.pdf`, March 2014.

[35] CERN Program Library home page. `http://cernlib.web.cern.ch/cernlib/`.

[36] Ian Mclaren. A Brief History of Cernlib (1966-???). `http://ref.web.cern.ch/ref/CERN/CNL/2001/001/cernlib/`.