Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)

PI: Peter Elmer (Princeton University)
co-PI: Gordon Watts (University of Washington)
co-PI: Brian Bockelman (University of Nebraska-Lincoln)

May 2, 2018

1 Introduction

The quest to understand the fundamental building blocks of nature and their interactions is one of the oldest and most ambitious of human scientific endeavors. Facilities such as CERN's Large Hadron Collider (LHC) represent a huge step forward in this quest. The discovery of the Higgs boson, the observation of exceedingly rare decays of B mesons, and stringent constraints on many viable theories of physics beyond the Standard Model (SM) demonstrate the great scientific value of the LHC physics program. The next phase of this global scientific project will be the High-Luminosity LHC (HL-LHC) which will collect data starting circa 2026 and continue into the 2030's. The primary science goal is to search for physics beyond the SM and, should it be discovered, to study its details and implications. During the HL-LHC era, the ATLAS and CMS experiments will record ∼10 times as much data from ∼100 times as many collisions as were used to discover the Higgs boson (and at twice the energy). The NSF and the DOE are planning large investments in detector upgrades so the HL-LHC can operate in this high-rate environment. A commensurate investment in R&D for the software for acquiring, managing, processing and analyzing HL-LHC data is critical to maximize the return-on-investment in the upgraded accelerator and detectors.

We propose to establish the Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP) to meet the software and computing challenges of the HL-LHC. The Institute will address key elements of the "Roadmap for HEP Software and Computing R&D for the 2020s" [1] and will implement the "Strategic Plan for a Scientific Software Innovation Institute (S2I2) for High Energy Physics" [2] submitted to the NSF in December 2017. These two documents represent, respectively, the outcome of international and U.S. HEP community planning processes; these were driven in part by the NSF-funded S2I2-HEP Institute Conceptualization Project [3]. Over the course of a dozen workshops during 2016 and 2017, more than 260 scientists and engineers from around the world were involved in building this community vision.

IRIS-HEP will serve as an active center for software R&D, function as an intellectual hub for the larger community-wide software R&D efforts, and transform the operational services required to ensure the success of the HL-LHC scientific program. Three high-impact R&D areas will leverage the talents of the U.S. university community: (1) development of innovative algorithms for data reconstruction and triggering; (2) development of highly performant analysis systems that reduce 'time-to-insight' and maximize the HL-LHC physics potential; and (3) development of data organization, management and access systems for the community's upcoming Exabyte era. IRIS-HEP will sustain investments in today's distributed high-throughput computing (DHTC) and build an integration path to deliver its R&D activities into the distributed production infrastructure.

As an intellectual hub, IRIS-HEP will lead efforts to (1) grow convergence research between HEP and the Cyberinfrastructure, Data Science and Computer Science communities for novel approaches to the compelling HL-LHC challenges, (2) bring in new effort from U.S. Universities emphasizing professional development and training, and (3) sustain HEP software and underlying knowledge related to the algorithms and their implementations over the two decades required. HEP is a global, complex, scientific endeavor. These activities will help ensure that the software developed and deployed by a globally distributed community will extend the science reach of the HL-LHC and will be sustained over its lifetime. It will also advance other HL-LHC era HEP experiments.

The plan for IRIS-HEP reflects a community vision. Developing, deploying, and maintaining sustainable software for the HL-LHC experiments has tremendous technical and social challenges. A university-based Institute to lead a "software upgrade" will complement the hardware investments being made. In addition to enabling the best possible HL-LHC science, IRIS-HEP will bring together the larger cyberinfrastructure and HEP communities to address the complex issues at the intersection of Exascale high-throughput computing and Exabyte-scale datasets in big science. This will advance the objectives of the 2016 National Strategic Computing Initiative [4].

2 Science Drivers

In the past 200 years, physicists have discovered the basic constituents of ordinary matter and they have developed a very successful theory to describe the interactions (forces) among them. All atoms, and the molecules from which they are built, can be described in terms of these constituents. The nuclei of atoms are bound together by strong nuclear interactions. Their decays result from strong and weak nuclear interactions. Electromagnetic forces bind atoms together, and bind atoms into molecules. The electromagnetic, weak nuclear, and strong nuclear forces are described in terms of quantum field theories. The electromagnetic and weak nuclear interactions are intimately related to each other, but with a fundamental difference: the particle responsible for the exchange of energy and momentum in electromagnetic interactions (the photon) is massless while the corresponding particles responsible for the exchange of energy and momentum in weak interactions (the W and Z bosons) are about 100 times more massive than the proton. A critical element of the Standard Model (SM) is the prediction (made more than 50 years ago) that a qualitatively new type of particle, called the Higgs boson, would give mass to the W and Z bosons. Its discovery at the LHC by the ATLAS and CMS Collaborations in 2012 [5, 6] confirmed experimentally the last critical element of the SM.

The SM describes essentially all known physics very well, but its mathematical structure and some important empirical evidence tell us that it is incomplete. These observations motivate a large number of SM extensions, generally using the formalism of quantum field theory, to describe new physics beyond the SM (BSM) physics. For example, "ordinary" matter as described by the SM accounts for only 5% of the mass-energy budget of the universe, while dark matter, which interacts with ordinary matter gravitationally, accounts for 27%. 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 nuclear interactions. 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 (how electrons and protons, and their heavier cousins, were created in the early universe) 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) present. Constraints on BSM physics come from "conventional" HEP experiments plus others searching for dark matter particles either directly or indirectly.

The Particle Physics Project Prioritization Panel issued a *Strategic Plan for U.S. Particle Physics* [7] in May 2014. It was quickly endorsed by the High Energy Physics Advisory Panel and submitted to the DOE and the NSF. The report identifies, *five compelling lines 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 an order of magnitude more data acquired at twice the energy and with 100 times the luminosity compared to the Higgs discovery. These are the Science Drivers for IRIS-HEP.

Summary of Physics Motivation: The ATLAS and CMS collaborations published letters of intent to do experiments at the LHC in October 1992, about 25 years ago. At the time, the top quark had not yet been discovered; no one knew if the experiments would discover the Higgs boson, supersymmetry, technicolor, or something completely different. Looking forward, no one can say

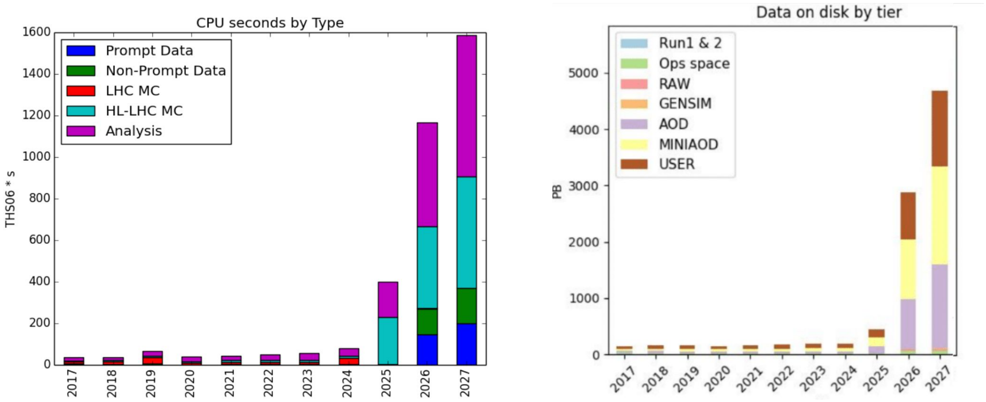

Figure 1: CPU and disk storage requirement evolution for the CMS experiment into the first two years of HL-LHC [8]. MC stands for Monte Carlo simulated data. RAW, GENSIM, AOD, MINIAOD and USER are various categories ("tiers") of data. HS06 is a HEP-specific SPECINT06-derived benchmark [9]. An average x86_64 core in recent years is approximately 10 HS06 units.

what will be discovered in the HL-LHC era. However, with major increases in luminosity and in data rates, the next 20 years are likely to bring even more exciting discoveries.

3 Computing Challenges

During the HL-LHC era (Run 4, starting circa 2026/2027), the ATLAS and CMS experiments intend to record 10x more events per year with a factor of 5x more complexity and 20x more data compared to Run 2. The planned integrated luminosity is $\mathcal{L}_{\text{int}} \sim 3000\,\text{fb}^{-1}$ at 14 TeV by 2035.

In terms of data volume, the LHC experiments expect a dramatic increase at the beginning of the HL-LHC program, as can be seen in Figure 1 and 2. Each of the experiments will generate several exabytes of science data per year. In contrast, the computing resource scrutiny group (C-RSG) [10] says that "growth equivalent to 20%/year [...] towards HL-LHC [...] should be assumed" when projecting available computing resources.

While no one expects such projections to be accurate over 10 years, it is clear that the projected needs of the experiments far exceed the resources that will be available assuming the typical growth rate of computational and storage capabilities and today's methods of doing processing and analysis. The magnitude of the discrepancy is illustrated in Figures 1 and 2 for CMS and ATLAS, respectively. Very crudely, the experiments need five times greater resources than will be available to achieve their full science reach in the first 1000 fb^{-1} of data. An aggressive and coordinated software R&D program is essential to mitigate this problem.

Aside from the purely science driven challenge of dealing with dramatically higher data rates and more complex events, HEP faces both hardware infrastructure and expert staffing challenges that must be addressed to ensure its continued effective use of computing facilities. Specifically:

- Processor technology evolution: The challenges for processor technologies are well known [12]. While the number of transistors on integrated circuits doubles every two years (Moore's Law), power density limitations and aggregate power limitations lead to a situation where "conventional" sequential processors are being replaced by vectorized and even more highly parallel architectures. Major changes to the algorithms implemented in our software are required to take advantage of these new architectures.

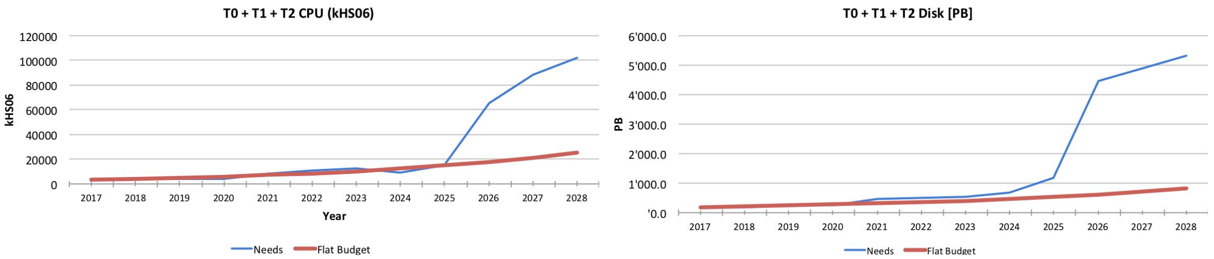

Figure 2: ATLAS CPU and disk requirement evolution into the first three years of HL-LHC, compared to growth rate assuming flat funding [11]. Units are as in Figure 1.

- New computing architectures: Understanding how emerging architectures, from low power processors to parallel architectures like GPUs to more specialized technologies like FPGAs, are one way for HEP to realize dramatic growth in computing power.
- Storage evolution: Limitations in affordable storage pose a major challenge, as does the I/O capacity of ever larger hard disks. While wide area network capacity will probably continue to increase at the required rate, the ability to use it efficiently will need a closer integration with applications.
- Training and expert development: The human and social challenges run in parallel with the technical challenges. All algorithms and software implementations are developed and maintained by people, many with unique expertise. HEP must understand how to effectively train large numbers of novice developers, and to cope with smaller numbers of more expert developers and architects, while fostering effective collaboration within software development teams and across experiments.

The overarching goal of IRIS-HEP is to maximize the return-on-investment in the upgraded accelerator and detectors to enable break-through scientific discoveries. Each of these software and computing challenges must be met with innovative ideas and approaches for HL-LHC to be a success. This requires a coherent effort engaged with the larger scientific software community. IRIS-HEP will play a central role in guaranteeing this success.

4 Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)

As described in Section 1, the mission of IRIS-HEP will be to serve as both an active software research and development center and as an intellectual hub for the larger R&D effort required to ensure the success of the HL-LHC scientific program. The proposed Institute will operate for a five year period from September 2018 to 2023 (inclusive) and therefore coincide with two important steps in the ramp up to the HL-LHC: (1) the delivery of the Computing Technical Design Reports (CTDRs) of ATLAS and CMS currently targeting ∼2020, and (2) Run 3 of LHC in 2021 through 2023. The CTDRs will describe the experiments' technical blueprints for building software and computing infrastructures to realize the maximum physics reach for HL-LHC, given financial constraints defined by the funding agencies.

For ATLAS and CMS, the increased size of the Run 3 data sets relative to Run 2 will not be a major challenge, and changes to the detectors will be modest compared to the upgrades anticipated for Run 4. As a result, ATLAS and CMS will have an opportunity to deploy prototype elements of the HL-LHC computing model during Run 3 as real operational tests, even if not at full scale. In contrast, LHCb is making a major transition in its scale of data processing at the onset of Run 3.

Some Institute deliverables will be deployed at full scale to directly maximize LHCb physics and provide valuable experience the larger experiments can use to prepare for the HL-LHC.

4.1 Institute Structure

The Institute will have a number of internal functional elements, as shown in Figure 3. The Management and Coordination Plan details effort levels for each of the elements.

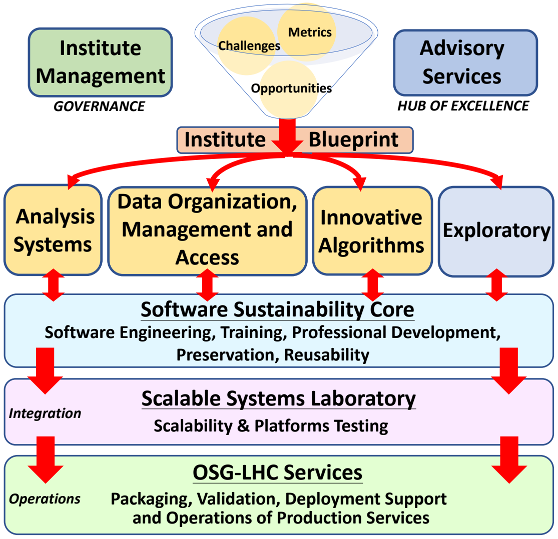
Institute Management
GOVERNANCE
Challenges
Metrics
Opportunities
Advisory Services
HUB OF EXCELLENCE
Institute
Blueprint
Analysis Systems
Data Organization, Management and Access
Innovative Algorithms
Exploratory
Software Sustainability Core
Software Engineering, Training, Professional Development, Preservation, Reusability
Integration
Scalable Systems Laboratory
Scalability & Platforms Testing
Operations
OSG-LHC Services
Packaging, Validation, Deployment Support and Operations of Production Services

Figure 3: Internal elements of the Institute.

Institute Management: The Institute will have a well-defined internal management structure, as well as external governance and advisory structures. Further information on these is provided in the supplemental "Management and Coordination Plan."

Institute Blueprint: The Blueprint activity will maintain the software vision for the Institute. Blueprint activities will be essential elements to building a common vision with other HEP and HL-LHC R&D efforts. One aspect of the Blueprint activity will be to bring together subject-matter experts to answer specific key questions within the scope of the Institute's activities or within the wider scope of HEP software/computing. Three or four workshops are anticipated each year. The blueprints will inform the evolution of both the Institute activities and the overall community HL-LHC R&D objectives in the medium and long term.

R&D Areas: The proposed Institute will initially have three focus areas that encompass the major R&D activities of the Institute. These areas are: Analysis Systems; Data Organization, Management and Access; and Innovative Algorithms, which will be described in Section 4.2. Each of these R&D areas will have its own specific plan of work and metrics for evaluation.

Exploratory: The Institute occasionally may deploy modest resources to carry out short-term exploratory R&D projects. These projects will be designed to inform the planning and overall mission of the Institute.

Sustainable Software Core: In addition to the specific technical advances that will be enabled by the dedicated R&D areas, a "core" activity will focus on the key elements of producing and utilizing *sustainable software*. This will impact all aspects of HL-LHC era physics from R&D to integration to operations. Backbone elements will include improving ways to communicate to students and researchers; identifying best practices in software engineering as well as possible incentives to adopt them, developing and providing training and professional development; and making data and tools that are modular, reusable, and available to the public.

Scalable Systems Laboratory: The SSL will provide the Institute and the HL-LHC experiments with a means to transition their R&D from toys to testbeds to production-scale prototypes. The process is capable of supporting innovation of novel analysis and data architectures, development of software elements and tool chains, functional and scalability testing of service components, and foundational systems R&D for accelerated services developed by the Institute.

OSG-LHC Services: Unlike the hardware composing the HL-LHC detectors, the computing services have to be in continuous operation throughout the LHC era into HL-LHC. The OSG project provides a production-quality set of services for a wide community. The Institute's operational components will include LHC-specific services within the OSG. This includes activities like packaging, validation, deployment, and operation of services as they pertain to the DHTC community.

Advisory Services: The Institute will play a role in the larger research software community (in HEP and beyond) by being available to provide technical and planning advice to other projects and by participating in reviews. The Institute will execute this functionality both with individuals directly employed by the Institute and by involving others through its network of partnerships.

4.2 Institute R&D Focus Areas

Analysis Systems: The goal of analysis systems is to realize the maximum scientific potential of the data in the least time. The final stages of data analysis are topically diverse and highly collaborative within small subsets of the experiment collaborations. Future analysis models must be nimble enough to adapt to new opportunities for discovery and large increases in data volume and analysis complexity, and still minimize "time-to-insight". Future analysis systems must provide high throughput capability and data interactivity. Furthermore, they must include reproducibility and reuse as fundamental capabilities of the entire system. IRIS-HEP will focus on four of the areas outlined in Refs. [2,13]:
Diversification: Over the last 20 years, HEP has developed and primarily utilized an analysis ecosystem centered on ROOT [14]. While this approach has certain advantages for HEP, maintenance and sustainability are significant challenges. Open source tools for data analysis have become widely available from industry and data science. This modern ecosystem includes analytics platforms, machine learning tools, and efficient data storage protocols. In many cases, these tools are evolving rapidly and surpass HEP in terms of in total software development effort and the size of communities that use and maintain them. IRIS-HEP will look to leverage these tools for HEP data analysis and achieve better HEP engagement with data science communities for software R&D.
Functional declarative programming: The dramatic increase in data and the availability of heterogeneous computing resources like HPC clusters, GPU's, and more traditional compute platforms poses both a challenge and an opportunity for HEP. Instead of physicists defining *how* data processing for analysis should proceed, they instead declare *what* they want to occur in their analysis. This functional, or declarative programming model allows (and delegates the responsibility to) the underlying infrastructure to optimize all aspects of the workflow, including data access patterns and execution concurrency. IRIS-HEP will pursue R&D projects needed to develop functional programming models tailored to physics use cases, along with the sophisticated algorithms to match declarative specifications to underlying compute resources within a global optimization framework.
High-throughput, low-latency systems: Current late-stage analysis involves data reduction via a series of time-intensive processing steps. A more modern approach is to employ high-throughput, low-latency analysis systems. Two types have been explored in HEP: a system based on a technology like Spark [15] and query-based systems. IRIS-HEP will extend existing work [16–18] to develop them beyond prototypes to understand how they can improve the analysis ecosystem infrastructure.
Analysis preservation and reuse: Reproducibility is the cornerstone of scientific results, yet currently it is difficult to repeat or reuse most HEP analyses. Furthermore, reuse of analysis-specific software is currently very inefficient even though the investment in a complete analysis approved by HEP experiments is many person-years. Building the capability to easily reuse or repurpose an analysis can reduce the time needed to produce new physics results from of years to months. Analysis preservation and reuse strategies need to be considered in all new approaches under investigation, so that they become a fundamental component of the system as a whole. In addition, the analysis workflows should be tightly integrated into the high-level cyberinfrastructure used to

dissiminate HEP results via the INSPIRE [19] literature system and HEPData [20] data archive.

Data Organization, Management and Access (DOMA): The DOMA focus area performs fundamental R&D related to the central challenges of organizing, managing, and providing access to exabytes of data from processing systems of various kinds.
Data organization: R&D focuses on distributed storage infrastructure that enables dynamic transformation of data from being optimized for archival storage to agile event reconstruction, and/or fast & efficient partial event access for analysis. This includes R&D in on-disk event structure, metadata and labeling for object selection, and organization optimized for function and quality of service. The latter may include compression; placement to minimize access latency; and high-volume delivery for reconstruction. A key component of the organization is more actively incorporating archival storage into the storage infrastructure, reducing the volume of data the LHC experiments are required to keep on disk and therefore reduce costs.
Data management and delivery: R&D focuses on how data is moved from the archival format and large data facilities. This includes interaction with experiment's larger data management systems for movement of bulk data between continents, interaction with transient storage or caches, and how data is delivered to users or streamed to centers doing bulk processing. The last case is where data is transformed into a desired format and will involve edge services that deliver data for large-scale processing at HPC centers.
Data access: R&D focuses on the interfaces, APIs, and infrastructure aspects necessary to provide fast access to data for analysis. The latter includes intelligent storage, efficient data organization, and evaluation of industry standard big data tools for use at the HL-LHC.

The R&D mentioned above will be fully integrated with the needs of other focus areas. This includes dynamically transforming data as it is imported from archive into Analysis System prototypes, and developing efficient data access mechanisms for such prototypes. DOMA works closely with SSL and OSG in order to support the software lifecycle from R&D to testing, and production scale prototyping.

Innovative Algorithms: Algorithms to perform the real-time processing in the trigger and the reconstruction of both real and simulated detector data are critical components of HEP's computing challenge. University personnel, including graduate students and post-docs working on physics research grants, frequently develop and maintain innovative algorithms and implementations. These algorithms face a number of new challenges in the next decade due to new and upgraded accelerator facilities, detector upgrades and new detector technologies, increases in anticipated event rates, and emerging computing architectures. Tracking for the HL-LHC is an area in particular need of novel approaches, though the Institute will pursue other high-impact applications. The Institute will employ a wide range of strategies for the development of Innovative Algorithms, for example:

Innovative Tracking Algorithms will explore solutions for track reconstruction at the HL-LHC. As tracking dominates the overall CPU budget of reconstruction, significant algorithmic improvements will decrease the overall cost of computing for a given performance, thus enabling selection based on tracking to be used on a bigger part of the incoming LHC data - both online trigger and offline. This effectively increases the size and physics quality of the LHC data that we can afford to write out of the detector.

A key target would be switching tracking software from being fundamentally sequential, to becoming massively vectorized and thus parallel. For CMS, the work will build on accomplishments achieved as part of the NSF funded project "Particle Tracking at High Luminosity on Heterogeneous, Parallel Processor Architectures" and extend it to novel compute architectures like FPGAs. It will research the fundamental limitations of GPU architectures for Kalman filter tracking, exploring new algorithms that are inherently more amenable to parallelization than the traditional Kalman Filter algorithm, and generalize the previous work done to optimize memory access to other data structures and algorithms beyond tracking. For ATLAS, the work will be performed in collaboration with A Common Tracking Software (ACTS) project which will provide parallelised algorithms in an open-source and detector-independent code base. We aim to deploy our tracking algorithms parasitically in the HLT during Run 3 data taking as a significant step towards full tracking in the High Level Trigger for ATLAS and CMS.

We will also develop tracking algorithms for new detector architectures and technologies. This will include developing algorithms with high efficiency and purity in the new forward regions of the tracking detectors and exploiting detector architectures with tilted sensors and many hits per track. We will also explore the use of techniques from machine learning in key areas. For example, for ATLAS, replacing the classical ambiguity solving function used to discriminate between duplicate tracks with a machine learning discriminant. This would be expected to improve the track reconstruction performance in the cores of jets.

Machine Learning (ML) is a rapidly evolving area of computer science providing novel algorithmic approaches for solving a wide variety of tasks based on data. Exploiting these developments will require a significant and coordinated effort to understand the pros and cons of ML based approaches compared to traditional techniques.

The HEP community is investigating the use of ML techniques across the gamut of HEP computing tasks. Modern ML frameworks have been designed to leverage GPU and FPGA technologies [21]. In addition to the use of ML for event selection and particle identification, there has been an explosion of effort to use ML for jet physics, tracking, and triggering applications. Recently, ML and data science algorithms have been developed to for measurements [22], data quality monitoring, the tuning of existing Monte Carlo simulations [23, 24], and neural network based simulations trained on real data [25–27]. Algorithmic enhancements and new approaches will enable extended physics reach even in more challenging detection environments. Moreover, algorithmic development is needed to take full advantage of enhanced detector capabilities such as timing detectors and high-granularity calorimeters.

Evaluation of traditional techniques, modernized approaches to traditional algorithms, and machine learning based approaches will require well-considered metrics and a coordinated benchmarking effort. The Institute will take a holistic view that incorporates not only algorithmic performance, but also considerations of computing hardware requirements, software maintenance and documentation, and deployment in our production and online systems. The Institute will be able to leverage the SSL and expertise in the OSG-LHC Services.

Institute R&D organization: These three focus areas will seed the intellectual contributions of the Institute. They are expected to facilitate the growth of transdisciplinary convergent research in the HEP community. The R&D focus of the institute will evolve over time, directed by stakeholder input and involving both funded and unfunded collaborators. This process will be facilitated by regular community workshops, blueprint activities, the results from exploratory R&D activities and input from the Institute's advisory groups and management team.

4.3 Cross-cutting Elements

In addition to the Sustainability Core, the Scalable Systems Laboratory (SSL) and the OSG Services for the LHC (OSG-LHC) represent important cross-cutting elements of the Institute. The SSL and OSG-LHC provide a mechanism for scalability & platforms testing and service operations necessary for the U.S. LHC community to continually evolve its infrastructure from now until the HL-LHC.

Scalable Systems Laboratory (SSL): The SSL is constructed to have a core team with expertise in scale testing and deploying services across a wide range of cyberinfrastructure. This core team will embed and partner with other areas in the Institute to define investigations, design concrete tests, execute and evaluate the results. The team will embed with the relevant area for each test, dynamically growing and allowing it to draw upon a wider pool of effort to accomplish its goals.

The scalable platforms provided by the SSL does not imply it is a large-scale infrastructure project. A key enabler for the success of the SSL will be the partnerships established outside the Institute for access to services and resources representative of the HL-LHC scale. We will rely on a number of resource types: (1) existing U.S. LHC cyberinfrastructure resources (including personnel and hardware platforms), (2) NSF-funded R&D programs such as SLATE [28, 29], (3) the OSG, (4) NSF supercomputers, and (5) institutional resources affiliated with the Institute. When a significant test is needed the SSL will work with these sources to access and aggregate allocations, perform validation and scale testing, and release the resources back to providers. In this context we will incorporate innovations in automation, service orchestration, and configuration management in building suitable DevOps environments for the software innovation teams. We have demonstrated this approach previously within the community [30–32].

As an example from the Conceptualization Plan [2], we consider innovating a new delivery network coupling a large capacity storage “lake” with distributed HPC centers. The delivery network would consist of a set of caches and other organizational data services embedded at various points in the distributed LHC cyberinfrastructure, deployed and operated by a central team of service developers. Systems of orchestrated, containerized services would be functionally tested and assessed for scalability and performance in realistic configurations using leveraged resources from the participating institutions and the U.S. LHC computing facilities.

The Open Science Grid Services for LHC (OSG-LHC): Since 2006, the Open Science Grid (OSG) has provided a fabric of services necessary for the production-grade distributed high-throughput computing infrastructure required by the LHC experiments. The OSG is organized as a consortium run by an independent council; members of the consortium - including the NSF-funded OSG project - contribute the effort required to execute its vision.

For a complex software ecosystem involving functioning components for today’s LHC needs and interoperating with the U.S.’s international partners, integrating new R&D from the Institute cannot be done as a “big bang”. Instead, there must be a continuous pipeline from the internal R&D into a running production infrastructure.

Hence, to provide this on-ramp while having continuity of the existing services for LHC, the Institute will include a team, contributing to the OSG consortium, that performs LHC-specific activities. This team will start with a focus on the following activities, identified in conjunction with the U.S. LHC operations programs: (1) software packaging and integration; (2) operations of software distribution, job accounting, registration, and monitoring services; (3) cybersecurity infrastructure and operational cybersecurity; and (4) coordination with international distributed computing partners such as the Worldwide LHC Computing Grid (WLCG), as well as the wider DHTC community beyond the LHC. As IRIS-HEP matures, these startup activities will adapt so the team can provide an “on-ramp” of R&D products into the U.S. LHC’s cyberinfrastructure.

4.4 The Institute as an Intellectual Hub

IRIS-HEP will serve as an *intellectual hub*, not only for the HL-LHC R&D effort, but also for the larger HEP software and computing (S&C) community *and* for external communities with overlapping interests. It will build on activities started in the S2I2-HEP conceptualization project and the associated Community White Paper roadmap process. In this role, IRIS-HEP will focus on seeding, building, and enabling collaborations across experiments, countries, and disciplines. It will serve as a focal point for convergent research by individuals and groups from the HEP, computer science, data science, software engineering, and cyberinfrastructure communities. IRIS-HEP will provide a forum for domain experts outside HEP –and from the private sector– to advise the HEP community on sustainable software development. The Institute will partner with the HEP S&C community (writ large) to provide the computer science community access to HEP data sets, environments and systems for use in CS-specific research. Similarly, the Institute will serve as a center for the HEP community to disseminate knowledge related to the current S&C landscape, emerging technologies, and tools. IRIS-HEP will provide critical evaluation of new algorithms, implementations, sustainability, and provide recommendations to collaborations on training, workforce, and software development.

Building on the experience of the S2I2 conceptualization project and preparing the Community White Paper, the Institute will facilitate Blueprint activities for the HEP S&C community as well as for itself. The primary mechanism will be bringing together key personnel for small workshops on specific aspects of the full R&D effort. These will serve to develop and document a common vision for the broad R&D landscape.

The OSG-LHC activities will provide a bi-directional connection to the larger open science community. This will provide opportunities for collaboration and for disseminating IRIS-HEP R&D results, particularly for the DOMA and SSL work related to distributed computing.

Education and outreach, discussed in Section 5, is another area where the Institute will provide intellectual leadership. The core team will undertake some of the outreach activities itself (especially those related to packaging software or datasets, as an example), but most will be undertaken in cooperation with partners of the Focus Area research groups. Many examples of possible projects are discussed in the S2I2 Strategic Plan [2]. Specific choices will be made annually by the leaders of these teams, following discussions with the Steering Board and Executive Board and will be reviewed by the Advisory Panel. These activities will provide special opportunities for advancing diversity: the Institute will proactively reach out to ensure broad representation at both the instructor and student level. Amongst other benefits, these activities will provide individuals with opportunities to directly engage with peers and gain visibility in the community.

5 Training, Education, and Outreach

People are the key to successful software. Working together across disciplines, experiments, and generations, they are the real cyberinfrastructure underlying sustainable software. Developing, maintaining, and evolving the algorithms and software implementations for HEP experiments will continue for many decades. The HEP community is currently planning hardware upgrades for the HL-LHC era which will *start* collecting data 8 or 9 years from now, and then acquire data for at least another decade. Building the necessary software requires a workforce with a mix of HEP domain knowledge, advanced software skills, and strong connections to other related disciplines. The investments to grow this workforce must begin today and IRIS-HEP will play a leading role.

Current Practices: Education and training for software-related activities in HEP is uneven and consists of a patchwork of activities with significant holes. Although most universities do provide some relevant computer science and software engineering courses, and many are starting to provide introductory “data science” courses, many HEP graduate students and postdocs are not required to take these classes as part of the curriculum. No “standard” recommendations exist for incoming research students, from the individual HEP experiments or the HEP field as a whole. Some universities are developing curricula for STEM training in general and/or “certificate” programs for basic data science and/or software training, but these are by no means universal.

HEP collaborations do typically provide opportunities for members to learn the software tools developed by and/or used within the experiments. For example, the week-long CMS Data Analysis School (CMSDAS) [33] pairs software experts with new collaborators to build and run end-to-end examples of real analysis applications. Similarly, LHCb has a training program and workshops called the “Starter Kit” [34] and the ATLAS collaboration has maintained an “ATLAS Analysis Workbook” to provide information and examples for new (and experienced) ATLAS scientists doing physics analysis. The goals of these programs are primarily to make new collaborators effective *users* of the complex experiment software ecosystems, rather than effective developers of that ecosystem, even if the latter will be often an important part of their eventual research contribution. In addition, these programs need to train collaborators with very uneven backgrounds in basic ideas of computer science and software engineering, as described above.

A number of summer schools focused on more advanced software and computing topics also exist in the global HEP community. These include, among others, the CERN School of Computing [35], the GridKa school [36] in Germany organized by the Karlsruhe Institute of Technology, the "Developing Efficient Large Scale Scientific Applications (ESC)" [37], school organized by the Istituto Nazionale di Fisica Nucleare (INFN) in Italy and (more recently) the "Computational and Data Science for High Energy Physics (CoDaS-HEP)" school [38] in the U.S. Similarly, the laboratories also organize some "short-course" training activities. For example, the LHC Physics Center (LPC) at Fermilab also offers half-day targeted training on specific topics.

A Vision for Training: The HEP community planning process during 2017 triggered numerous discussions regarding training. Training is central to building the community skills needed to address the computing challenges of the HL-LHC era. One key insight is the need to think of training not as a set of individual, disconnected activities, but as part of a larger framework, as shown in Figure 4.

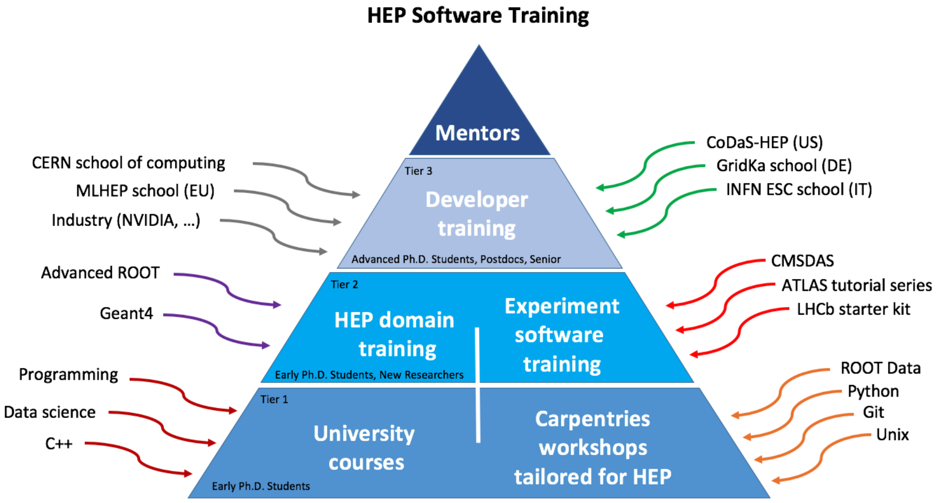

Figure 4: A vision for training in HEP: researchers progress (vertically) from basic skills training, through "user" training in existing software to training in skills needed to development new research software.

The highest impact role for the Institute regarding training will be to *coordinate* training related activities and to assemble and communicate this *coherent vision* of a training program for HEP graduate students, postdocs and more senior researchers in software and computing. In addition to its own direct training activities, it will develop a process with the community for implementing and updating that vision over time. It can build a "federated" view over the possible training opportunities in the experiments, at the labs, in dedicated summer schools and from other sources (HEP and non-HEP). It will bring together the people organizing those training activities not only to articulate the vision, but also to develop plans to enhance the sustainability, reusability and impact of the training activities. Finally, it will work with the community to build an assessment framework for the ensemble of activities that allows us to measure the impact of our activities.

In addition, IRIS-HEP will devote some resources to specific training activities, with the goal of making a complete training program accessible for all U.S. graduate students and postdocs.

We will support the existing CoDaS-HEP school and will provide travel support for students and postdocs to participate. Personnel funded by IRIS-HEP will serve as lecturers for the CoDaS-HEP school, for other schools and will give short courses at labs and universities. In order to provide opportunities for further professional development, including mentoring, we will establish the IRIS Fellows program. IRIS Graduate Fellows will each spend 3 months intensively developing software tools in conjunction with either IRIS-HEP personnel or with collaborating institutions. Similarly, IRIS Undergraduate Fellows will work 10 - 12 weeks during the summer, either developing or using data-intensive tools. IRIS Fellows will also be a key ingredient to building an ever larger community around the Institute as an intellectual hub. IRIS-HEP personnel will also engage with and serve as mentors for other student programs at their universities, at labs (e.g. CERN, DOE labs), with Google Summer of Code and other similar programs.

Education and Outreach: Outreach will include both educational activities aimed at targeted communities and activities aiming to include the general public. The institutions participating in the IRIS-HEP project already have good track records in these activities, and the Institute will work with them to expand efforts that have strong software components.

Diversity: In addition, the Institute leaders have proactively sought diversity in building its own research teams and advisory bodies. The institute will use its Outreach programs to establish mentoring relationships with a diverse community through sponsoring undergraduate and graduate student Fellowships. Women and minorities are poorly represented in HEP, and even more so in software and computing areas. While the Institute can not tackle this problem alone, it will work to minimize bias in the pipeline for students by partnering with its institution's efforts. The Institute will strive to introduce exciting HEP science at all educational levels, from high-school outreach, to Undergraduate REU programs to programs at predominantly minority universities. Partnerships are already established between senior IRIS-HEP personnel and the ENLACE [39] and STARS [40] programs at UCSD and the INCLUSION [41] program at U.Illinois. INCLUSION is an REU program run out of NCSA that provides opportunity for undergraduates from underrepresented communities and Minority Serving Institutions to work with each other and mentors (e.g. IRIS-HEP personnel) on interdisciplinary, socially-impactful projects that develop and use open source software. Similarly, we will partner with the Graduate and Undergraduate Societies for Women in Physics (SWiP) at the University of Chicago, the University of Nebraska and UC San Diego.

6 Timeline and Metrics of Success

The primary goal of the Institute is fostering, coordinating, and performing research and development that leads to deployment of high impact software. The aim is to significantly advance the physics reach of the HL-LHC experiments. Because the Institute will exist within a larger context of international and national projects, a second major goal will be to build a more cooperative, community process for developing, prototyping, and deploying sustainable software. The timeline and metrics for success must address both of these goals.

Timeline The first two years of the Institute are designated as the Ramp Up Phase, to be followed by a Full Execution Phase. Figure 5 provides an overview of some the activities during these phases. The first year efforts will be focused on fully staffing the Institute and developing its infrastructure. The Focus Areas will establish work plans and begin to execute them. This will include defining initial goals, identifying specific use cases, and starting the execution of work plans. Example initial goals would be to include a parallelized Kalman filter implementation for track reconstruction (Innovative Algorithms area) for Run 3 or initial deployment of a distributed storage facility for delivering event data to HPC centers (DOMA and SSL areas) for Run 3. Tools for internal communication, sharing of code, unit and integration testing, external communication, etc., will be determined. The HEP software community already has tools for these activities; the efforts in these areas will be choosing appropriately to create a coherent environment.

A key effort during the Ramp Up will be holding small, focused, Blueprint workshops to guide the internal Institute plans and also to align the Institute's plans with those of other entities doing related software R&D for HL-LHC era experiments. Topics to be discussed will be defined by the Focus Area teams. These will also help to form partnerships with the larger community and incorporate their input in the Focus Area work plans. The first IRIS-HEP annual workshop (an "all-hands" meeting) will be held in the first year. This will be both a forum for planning activities, and also an opportunity to build a sense of community.

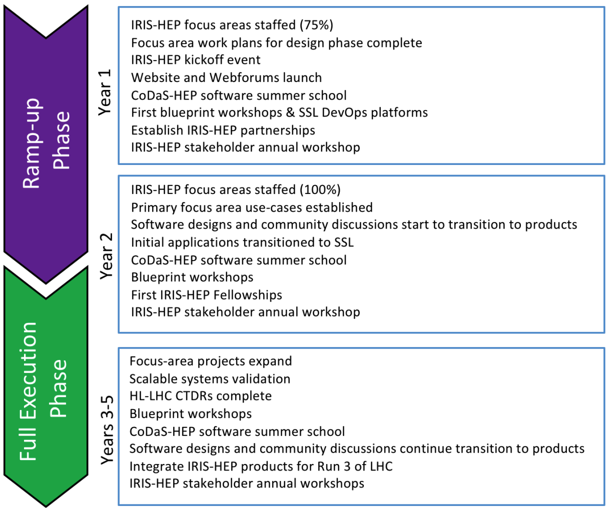
Ramp-up Phase
Year 1
IRIS-HEP focus areas staffed (75%)
Focus area work plans for design phase complete
IRIS-HEP kickoff event
Website and Webforums launch
CoDaS-HEP software summer school
First blueprint workshops & SSL DevOps platforms
Establish IRIS-HEP partnerships
IRIS-HEP stakeholder annual workshop
Year 2
IRIS-HEP focus areas staffed (100%)
Primary focus area use-cases established
Software designs and community discussions start to transition to products
Initial applications transitioned to SSL
CoDaS-HEP software summer school
Blueprint workshops
First IRIS-HEP Fellowships
IRIS-HEP stakeholder annual workshop
Full Execution Phase
Years 3-5
Focus-area projects expand
Scalable systems validation
HL-LHC CTDRs complete
Blueprint workshops
CoDaS-HEP software summer school
Software designs and community discussions continue transition to products
Integrate IRIS-HEP products for Run 3 of LHC
IRIS-HEP stakeholder annual workshops

Figure 5: Timeline framework for the Institute.

The Institute's software summer school (continuing the CoDaS-HEP [38] series) will be an annual event, with the first offering in 2019. Similarly, the first IRIS Fellows will be selected towards the end of the first year. The SSL will be working with the focus areas and the OSG-LHC services to develop its roadmap, with the first DevOps platforms available in 2019. During the Full Execution Phase, Focus Area projects that are maturing will move to integrate their products with the SSL and the experiments. As projects reach completion, the teams will transition to new projects. Some activities, like CoDaS-HEP and the Fellows program are anticipated to operate over the course of the Institute with only minor tweaks. Meetings with the Steering Board and the Advisory board during the Full Execution phase will be particularly important as the Institute focuses on getting its work ready for testing during Run 3 of the LHC. Blueprint meetings will continue as all aspects of the HL-LHC software efforts become more focused on Run 4.

Metrics No single set of metrics can evaluate the performance of all aspects of the Institute over the course of its full lifetime. Different questions must be asked to evaluate its performance during the Ramp-up Phase and during the Full Execution Phase. Specific goals and assessment metrics are required for all Institute activities, specifically including, (i) Focus Area R&D, (ii) other Institute R&D, (iii) Institute services, including SSL and OSG-LHC, (iv) Institute role as an intellectual hub, and (v) training, education, and outreach. The goals and the assessment metrics in each area will evolve over time. A detailed discussion of these ideas is presented in the S2I2 Strategic Plan [2].

The Institute's progress will be formally evaluated in several places. The Steering Board will work closely with the Executive Board to define goals and metrics for each forthcoming year, and will review progress on a quarterly basis. In addition, the Advisory Panel will be asked to review the Institute's plans and provide feedback annually. It will also be asked to judge whether proposed goals and metrics for the forthcoming year are appropriate.

7 Intellectual Merit of the Proposed Work

Developing cyberinfrastructure for use by domain scientists has two related types of *intellectual merit*: (i) enabling the domain science, and (ii) advancing understanding of algorithms, implementations, and software engineering. Fully exploiting the anticipated hardware investments in the HL-LHC, to do the promised transformative science, demands increased investments in software and computing. The primary goal of the Institute will be fostering the requisite software research and development, and helping the experiments deploy the innovative tools they need. In addition, the Institute will help maintain and advance community elements of the operational infrastructure.

The R&D to be done in the *Focus Areas* of Data Analysis Systems, DOMA, and Innovative Algorithms, will transform the HL-LHC's ability to do science by enabling the experiments to ingest, digest, and analyze the flood of data that will be produced. Data Analysis Systems will introduce techniques that will deliver decreased time-to-science by optimizing frameworks for high-volume I/O, developing interfaces that decrease the complexity exposed to the end user, and better leverage existing industry tools to complement HEP-specific ones. DOMA will focus on how data is organized within the existing distributed storage systems utilized by the LHC community, more actively using tape and reducing replication levels. DOMA will also study how data is delivered (either in terms of bulk movement or from an edge service on a computational resource) and how to optimize the format used for delivery. Innovative Algorithms will work to produce transformational advances in the physics algorithms used by the experiments; by applying modern programming techniques, leveraging new approaches in detector designs, and machine learning, this area will extend the physics reach of the HL-LHC.

The SSL will provide the Institute, and its partners, opportunities to test software and ideas at full production scale, by pulling together resources within IRIS-HEP and from the larger LHC community. It will provide the Institute with a diverse set of environments for integrating new tools; a capability that will be unique within the community. Specifically, it will help assure that R&D leads to successful deployment of useful software. A critical element in the success of the LHC, to date, has been its use of distributed high throughput computing (DHTC). The OSG-LHC activities to be incorporated into the Institute will provide a natural bridge for HEP computing from the current set of services to those required for the success of the HL-LHC.

8 Broader Impacts of the Proposed Work

Developing cyberinfrastructure to enable the science program of the HL-LHC will have broad impacts in fields as diverse as HEP and DHTC, and in designing sustainable software. Much of the software will have applications beyond the HL-LHC experiments, and even beyond HEP. Most students who earn Ph.D.'s in experimental particle physics, and even most post-docs, move into the private sector taking their skills with them. Hence, the education and training provided by the Institute will contribute to a highly qualified STEM workforce.

Many of the problems to be addressed by IRIS-HEP teams of physicists, computer scientists, data scientists, and software engineers will serve as the specific and compelling use cases of more general questions. As the non-physicists bring their knowledge, theories, and methods to bear on these problems, this transdisciplinary effort will drive additional research in their fields. This will serve as an example of the convergence research advocated by the NSF [42].

The ATLAS and CMS experiments each have ∼3000 collaborating physicists, and LHCb ∼800. The software to be developed and fostered will enable the full physics program of all these scientists. About 30% of the ATLAS and CMS collaborators come from U.S. institutions, as do ∼5% of the LHCb collaborators. Beyond the HL-LHC experiments, many of the tools will be used by other HEP experiments at Fermilab, BNL, and elsewhere. Past experience with the OSG demonstrates that DOMA and DHTC software developments in HEP can be deployed to benefit other sciences. Examples include Rucio [43], gWMS [44], CVMFS [45] and xrootd [46, 47]. As a vehicle for growing convergence research with the CS and Cyberinfrastructure communities, the Institute's research areas will impact and drive innovation in other domain sciences and the private sector. For example, machine learning algorithms developed as part of DIANA-HEP to mitigate systematic uncertainties [48] can be adapted by the private sector to ensure fairness (mitigate against implicit bias) [49–51]. Similarly, likelihood-free inference algorithms developed by DIANA-HEP [22, 24] are being used for genome-wide association studies [52, 53] and cardiac simulators [54].

9 Results from Prior NSF Support

Princeton: PI Elmer is supported by OAC-1450377 ("Collaborative Research: SI2-SSI: Data-Intensive Analysis for High Energy Physics (DIANA/HEP)", 05/01/2015-04/30/2019, $1,145,564), PHY-1521042 ("Collaborative Research: Particle Tracking at High Luminosity on Heterogeneous, Parallel Processor Architectures", 08/01/2015-07/31/2018, $494,529) and OAC-1558216 ("Collaborative Research: S2I2: Cncp: Conceptualization of an S2I2 Institute for High Energy Physics", 07/01/2016-06/30/2018, $95,036). **Intellectual Merit**: Interoperability of HEP tools with the Big Data software ecosystem [16–18, 55, 56]. Development of key algorithms for charged particle tracking [57–63]. **Broader Impacts**: Training of the national community of HEP physicists to use advanced software/computing tools, through schools like CoDaS-HEP [38]. Organization of dedicated workshops to explore common research interests between the HEP and Computer Science communities [64–66] and a multi-disciplinary (HEP, Astronomy, Genomics) workshop on data management [67]. **Publications**: As listed above. **Research Products**: The S2I2-HEP conceptualization project produced the Strategic Plan for U.S. university contributions to an HL-LHC "software upgrade" [2] and was a key driver of the development of the international community roadmap [1]. DIANA/HEP has produced numerous software products including uproot, Histogrammar and OAMap. For a full list see [68].

U.Nebraska: Co-PI Bockelman is senior personnel on PHY-1148698 ("THE OPEN SCIENCE GRID The Next Five Years: Distributed High Throughput Computing for the Nation's Scientists, Researchers, Educators, and Students", 6/01/2012-11/30/2018, $22,399,980) **Intellectual Merit**: The OSG is a national, distributed computing partnership for data-intensive research. It provides a fabric of services and a framework for sharing and utilizing a heterogeneous set of computational resources. **Broader Impacts**: The OSG underlies the US portion of the computing strategy of several major NSF investments, including the ATLAS and CMS experiments at the Large Hadron Collider (LHC). Approximately two-thirds of the OSG usage is US LHC experiments, the remainder usage is split between other High Energy Physics (HEP) experiments and other fields of science. **Publications**: OSG is utilized for a large range of publications by our stakeholders. For a broad overview of the OSG itself, see [69], [70]. **Research Products**: See [71] for links to further publications, software, and software packaging produced by this proposal. OSG maintains approximately monthly releases and has enabled hundreds of publications across many disciplines [72].

U.Washington: Co-PI Watts is a co-PI on PHY-1509257 ("Beyond the Standard Model Physics", 7/15/2015-6/30/2018, $2,490,000). Senior personnel on "Data and Software Preservation for Open Science (DASPOS) (sub-award from Notre Dame for DASPOS, 9/14/2012-8/31/15, $166,758). **Intellectual Merit**: Leading several searches for Long Lived Particles in the ATLAS detector, using boosted decision trees to differentiate jets coming from long lived particles and standard model jets [73, 74]. Work on developing a new surface detector at the LHC to search for ultra-long-lived particles and construction of a test-stand to demonstrate techniques [75]. Work on data preservation for the DZERO experiment at the Tevatron. In particular, the design of the cvmfs back-end for the software environment preservation [76]. Ontologies for preservation of physics analysis and software for reproducability [77]. **Broader Impacts**: Publications of trigger efficiencies for signature-driving triggers are used by phenomonologists to predict performance of their models. The workshop on Advanced Computing and Analysis Techniques (ACAT) 2017 was hosted at the University of Washington. 200 people attended [78]. Preserving data for the DZERO experiment for future use by the HEP community [76]. **Publications**: See above.

References

[1] Antonio Augusto Alves et al. A Roadmap for HEP Software and Computing R&D for the 2020s. 2017. [arXiv 1712.06982] https://arxiv.org/abs/1712.06982.

[2] Peter Elmer, Mark Neubauer, and Michael D. Sokoloff. Strategic Plan for a Scientific Software Innovation Institute (S2I2) for High Energy Physics. 2017. [arXiv 1712.06592] https://arxiv.org/abs/1712.06592.

[3] S2I2-HEP project webpage: http://s2i2-hep.org.

[4] National Strategic Computing Initiative. https://www.nsf.gov/cise/nsci/.

[5] G. Aad *et al.* [ATLAS Collaboration]. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1, 2012.

[6] S. Chatrchyan *et al.* [CMS Collaboration]. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716:30–61, 2012.

[7] 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.

[8] L.Sexton-Kennedy, presentation to the LHCC, 12 Sep 2017. https://indico.cern.ch/event/570975/contributions/2309740/subcontributions/208383/attachments/1521736/2377733/cms_lhcc_wlcg_Sept_2017.pdf.

[9] Michele Michelotto, Manfred Alef, Alejandro Iribarren, Helge Meinhard, Peter Wegner, Martin Bly, Gabriele Benelli, Franco Brasolin, Hubert Degaudenzi, Alessandro De Salvo, Ian Gable, Andreas Hirstius, and Peter Hristov. A comparison of hep code with spec 1 benchmarks on multi-core worker nodes. *Journal of Physics: Conference Series*, 219(5):052009, 2010.

[10] S. Foffano. Computing Resources Scrutiny Group Report. Technical Report CERN-RRB-2017-067, CERN, Geneva, April 2017.

[11] S.Campana, presentation to the LHCC, 12 Sep 2017. https://indico.cern.ch/event/570975/contributions/2309740/subcontributions/208382/attachments/1521583/2377282/ATLAS-LHCC-09-2017-V2.pdf.

[12] 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.

[13] Lothar Bauerdick et al. HEP Software Foundation Community White Paper Working Group - Data Analysis and Interpretation. 2018.

[14] Fons Rademakers and Rene Brun. ROOT: an object-oriented data analysis framework. *Linux J.*, page 6.

[15] Apache Spark. http://spark.apache.org.

[16] Oliver Gutsche, Luca Canali, Illia Cremer, Matteo Cremonesi, Peter Elmer, Ian Fisk, Maria Girone, Bo Jayatilaka, Jim Kowalkowski, Viktor Khristenko, Evangelos Motesnitsalis, Jim Pivarski, Saba Sehrish, Kacper Surdy, and Alexey Svyatkovskiy. CMS Analysis and Data Reduction with Apache Spark. 2017. [arXiv 1711.00375] http://arxiv.org/abs/1711.00375.

[17] Jim Pivarski, David Lange, and Thanat Jatuphattharachat. Toward real-time data query systems in HEP. 2017. [arXiv 1711.01229] http://arxiv.org/abs/1711.01229.

[18] Jim Pivarski, Peter Elmer, Brian Bockelman, and Zhe Zhang. Fast access to columnar, hierarchical data via code transformation. *CoRR*, abs/1708.08319, 2017.

[19] INSPIRE HEP website:. http://inspirehep.net/.

[20] HEPData website:. https://hepdata.net.

[21] Javier Duarte, Song Han, Philip Harris, Sergo Jindariani, Edward Kreinar, Benjamin Kreis, Jennifer Ngadiuba, Maurizio Pierini, Nhan Tran, and Zhenbin Wu. Fast inference of deep neural networks in FPGAs for particle physics. 2018.

[22] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

[23] Philip Ilten, Mike Williams, and Yunjie Yang. Event generator tuning using Bayesian optimization. 2016.

[24] Gilles Louppe and Kyle Cranmer. Adversarial variational optimization of non-differentiable simulators. *arXiv preprint arXiv:1707.07113*, 2017.

[25] Anders Andreassen, Ilya Feige, Christopher Frye, and Matthew D. Schwartz. JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics. 2018.

[26] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks. 2017.

[27] Charlie Guthrie, Israel Malkin, Alex Pine, Kyle Cranmer. (2017) GitHub: `https://github.com/pinesol/hep-calo-generative-modeling`.

[28] CIF21 DIBBs: EI: SLATE and the Mobility of Capability. `https://nsf.gov/awardsearch/showAward?AWD_ID=1724821&HistoricalAwards=false`.

[29] R Gardner, J Breen, L Bryant, and S. McKee. Slate and the mobility of capability. *Gateways Conference, October 23-25, 2017, Ann Arbor, Michigan*, 2017.

[30] J Balcas, B Bockelman, D Hufnagel, K Hurtado Anampa, F Aftab Khan, K Larson, J Letts, J Marra da Silva, M Mascheroni, D Mason, A Perez-Calero Yzquierdo, and A Tiradani. Stability and scalability of the cms global pool: Pushing htcondor and glideinwms to new limits. *Journal of Physics: Conference Series*, 898(5):052031, 2017.

[31] J Balcas, S Belforte, B Bockelman, O Gutsche, F Khan, K Larson, J Letts, M Mascheroni, D Mason, A McCrea, M Saiz-Santos, and I Sfiligoi. Pushing htcondor and glideinwms to 200k+ jobs in a global pool for cms before run 2. *Journal of Physics: Conference Series*, 664(6):062030, 2015.

[32] E M Fajardo, J M Dost, B Holzman, T Tannenbaum, J Letts, A Tiradani, B Bockelman, J Frey, and D Mason. How much higher can htcondor fly? *Journal of Physics: Conference Series*, 664(6):062014, 2015.

[33] S. Malik, F. Hoehle, K. Lassila-Perini, A. Hinzmann, R. Wolf, et al. Maintaining and improving of the training program on the analysis software in CMS. *J.Phys.Conf.Ser.*, 396:062013, 2012.

[34] LHCb Starter Kit webpage. `https://lhcb.github.io/starterkit/`.

[35] CERN School of Computing webpage. `https://csc.web.cern.ch/`.

[36] GridKa School (KIT) webpage. `http://gridka-school.scc.kit.edu/`.

[37] ESC17 school webpage: `https://web.infn.it/esc17/index.php`.

[38] CoDaS-HEP school webpage: `http://codas-hep.org`.

[39] UCSD ENLACE program homepage. `http://graeve.ucsd.edu/ENLACE/`.

[40] UCSD STARS program homepage. `https://grad.ucsd.edu/degrees/summer-researches/stars/index.html`.

[41] **I**ncubating a **N**ew **C**ommunity of **L**eaders **U**sing **S**oftware, **I**nclusion, Innovation, Interdisciplinary and **O**peN-Science (INCLUSION) Program. `https://reu.ncsa.illinois.edu`.

[42] Convergent Research at NSF. `theirknowledge,theories,methods`.

[43] Rucio website. `https://rucio.cern.ch`.

[44] I Sfiligoi. glideinWMS - a generic pilot-based workload management system. *Journal of Physics: Conference Series*, 119(6):062044, 2008.

[45] J. Blomer, P. Buncic, R. Meusel, G. Ganis, I. Sfiligoi, and D. Thain. The Evolution of Global Scale Filesystems for Scientific Software Distribution. *Computing in Science Engineering*, 17(6):61–71, Nov 2015.

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

[47] L Bauerdick, D Benjamin, K Bloom, B Bockelman, D Bradley, S Dasu, M Ernst, R Gardner, A Hanushevsky, H Ito, D Lesny, P McGuigan, S McKee, O Rind, H Severini, I Sfiligoi, M Tadel, I Vukotic, S Williams, F Wuerthwein, A Yagil, and W Yang. Using Xrootd to Federate Regional Storage. *Journal of Physics: Conference Series*, 396(4):042009, 2012.

[48] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In *Advances in Neural Information Processing Systems*, pages 982–991, 2017.

[49] Towards fairness in ML with adversarial networks (2018). `https://blog.godatadriven.com/fairness-in-ml`.

[50] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, pages 585–596, 2017.

[51] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5969–5978, 2017.

[52] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5529–5539, 2017.

[53] Dustin Tran and David M Blei. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*, 2017.

[54] Adam McCarthy, Blanca Rodriguez, and Ana Minchole. Variational inference over non-differentiable cardiac simulators using bayesian optimization. *arXiv preprint arXiv:1712.03353*, 2017.

[55] Oliver Gutsche, Matteo Cremonesi, Peter Elmer, Bo Jayatilaka, Jim Kowalkowski, Jim Pivarski, Saba Sehrish, Cristina Mantilla Surez, Alexey Svyatkovskiy, and Nhan Tran. Big Data in HEP: A comprehensive use case study. *Journal of Physics: Conference Series*, 898(7):072012, 2017.

[56] Brian Bockelman, Zhe Zhang, and Jim Pivarski. Optimizing ROOT IO for analysis. *CoRR*, abs/1711.02659, 2017.

[57] Giuseppe Cerati, Peter Elmer, Steven Lantz, Ian MacNeill, Kevin McDermott, Dan Riley, Matevz Tadel, Peter Wittich, Frank Wuerthwein, and Avi Yagil. Traditional Tracking with Kalman Filter on Parallel Architectures. *J. Phys. Conf. Ser.*, 608(1):012057, 2015. [arXiv:1409.8213].

[58] Giuseppe Cerati, Peter Elmer, Steven Lantz, Kevin McDermott, Dan Riley, Matevž Tadel, Peter Wittich, Frank Würthwein, and Avi Yagil. Kalman Filter Tracking on Parallel Architectures. 2015. Submitted to proceedings of the 21st International Conference on Computing in High Energy and Nuclear Physics (CHEP2015), Okinawa, Japan. [arXiv:1505.04540].

[59] G. Cerati, M. Tadel, F. Wurthwein, A. Yagil, S. Lantz, K. McDermott, D. Riley, P. Wittich, and P. Elmer. Kalman-filter-based particle tracking on parallel architectures at hadron colliders. In *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–4, Oct 2015.

[60] Cerati, Giuseppe, Elmer, Peter, Krutelyov, Slava, Lantz, Steven, Lefebvre, Matthieu, McDermott, Kevin, Riley, Daniel, Tadel, Matev, Wittich, Peter, Wurthwein, Frank, and Yagil, Avi. Kalman filter tracking on parallel architectures. *EPJ Web Conf.*, 127:00010, 2016.

[61] G Cerati, P Elmer, S Krutelyov, S Lantz, M Lefebvre, K McDermott, D Riley, M Tadel, P Wittich, F Wurthwein, and A Yagil. Kalman filter tracking on parallel architectures. *Journal of Physics: Conference Series*, 898(4):042051, 2017.

[62] Cerati, Giuseppe, Elmer, Peter, Krutelyov, Slava, Lantz, Steven, Lefebvre, Matthieu, Masciovecchio, Mario, McDermott, Kevin, Riley, Daniel, Tadel, Matev, Wittich, Peter, Wurthwein, Frank, and Yagil, Avi. Parallelized kalman-filter-based reconstruction of particle tracks on many-core processors and gpus. *EPJ Web Conf.*, 150:00006, 2017.

[63] Giuseppe Cerati et al. Parallelized Kalman-Filter-Based Reconstruction of Particle Tracks on Many-Core Architectures. In *18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2017) Seattle, WA, USA, August 21-25, 2017*, 2017.

[64] S2I2 HEP/CS Workshop 2016. `https://indico.cern.ch/event/575443/`.

[65] S2I2 HEP/CS Workshop 2016 Report. `http://s2i2-hep.org/downloads/s2i2-hep-cs-workshop-summary.pdf`.

[66] 2nd S2I2 HEP/CS Workshop. `https://indico.cern.ch/event/622920/`.

[67] Data Organisation, Management and Access (DOMA) in Astronomy, Genomics and High Energy Physics. `https://indico.cern.ch/event/669506/`.

[68] DIANA/HEP Activities/Products webpage. `http://diana-hep.org/pages/activities.html`.

[69] Ruth Pordes, Don Petravick, Bill Kramer, Doug Olson, Miron Livny, Alain Roy, Paul Avery, Kent Blackburn, Torre Wenaus, Frank Wuerthwein, Ian Foster, Rob Gardner, Mike Wilde, Alan Blatecky, John McGee, and Rob Quick. The open science grid. *Journal of Physics: Conference Series*, 78(1):012057, 2007.

[70] Mine Altunay, Paul Avery, Kent Blackburn, Brian Bockelman, Michael Ernst, Dan Fraser, Robert Quick, Robert Gardner, Sebastien Goasguen, Tanya Levshina, Miron Livny, John McGee, Doug Olson, Ruth Pordes, Maxim Potekhin, Abhishek Rana, Alain Roy, Chander Sehgal, Igor Sfiligoi, and Frank Wuerthwein. A science driven production cyberinfrastructure - the open science grid. *Journal of Grid Computing*, 9(2):201–218, 2011.

[71] Open Science Grid webpage: `https://www.opensciencegrid.org`.

[72] OSG Annual Report (2017). `http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=1238`.

[73] et. al. G. Aad. Search for long-lived neutral particles decaying into jets in proton-proton collisions at $\sqrt{s} = 8$ tev with the atlas detector. *Journal of High Energy Physics*, 2014(11):88, Nov 2014.

[74] et. al. G. Aad. Search for long-lived neutral particles decaying into jets in the hadronic calorimeter of atlas at $\sqrt{s} = 8$ tev in 3.2 fb^-1 of data. *ATLAS Conference Note*, 2016, Sept. 2016.

[75] H.J. Lubatti J-P. Chou, D. Curtain. New detectors to explore the lifetime frontier. Sept. 2016.

[76] S. Amerio et al. Data preservation at the Fermilab Tevatron. *Nucl. Instrum. Meth.*, A851:1–4, 2017.

[77] Git Repro of Computational OWL Specification. `https://github.com/Vocamp/ComputationalActivity`.

[78] ACAT 2017 conference webpage: `https://indico.cern.ch/event/567550/`.