Community White Paper and S^2I^2 Conceptualization

OSG Council Meeting
July 28, 2016

Peter Elmer
Princeton University

Mark Neubauer
University of Illinois at Urbana-Champaign

Mike Sokoloff
University of Cincinnati

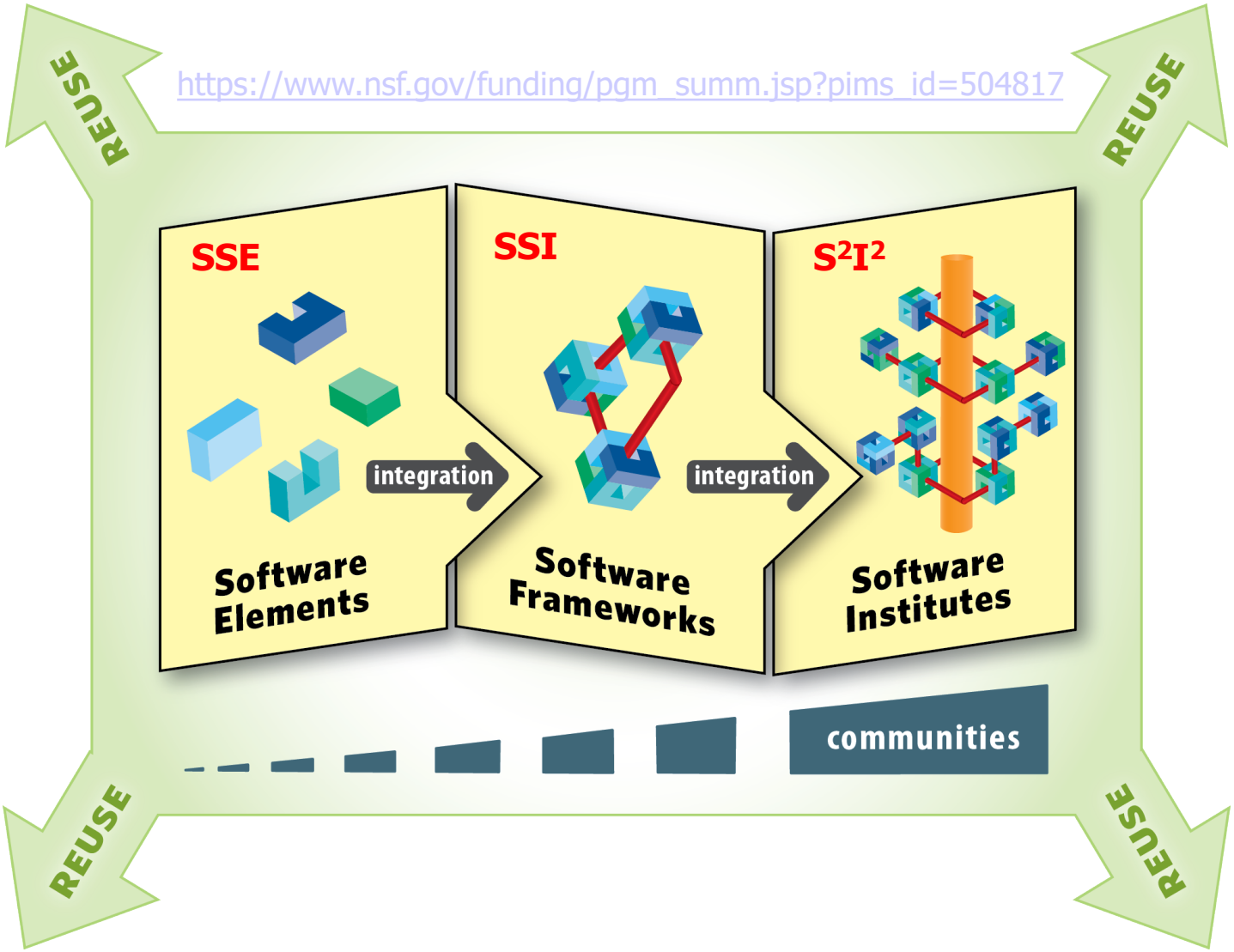
What this is all about

1. A process by which a roadmap document in the form of a ***Community White Paper*** (CWP) is produced which aims to broadly identify the elements of computing infrastructure and software R&D required to realize the full scientific potential of the HL-LHC

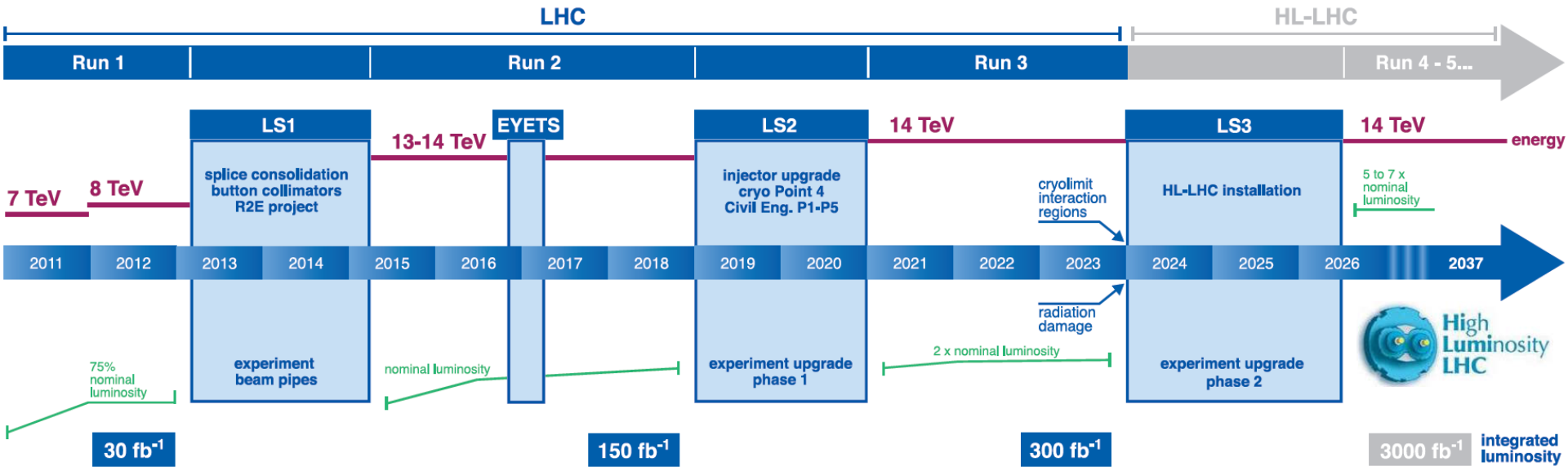
2. Conceptualization of a ***Scientific Software Innovation Institute*** (S^2I^2) where U.S. university-based researchers can play an important role in key software infrastructure efforts that will complement those led by U.S. national laboratory-based researchers and international collaborators

NSF Software Institutes

From NSF's Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) Vision, classes of software infrastructure typically fall into ***three*** classes:



A Software "Upgrade" for the HL-LHC



- Large investment in upgraded detectors for the HL-LHC are well underway
- Investment in R&D and upgrades to our software & computing infrastructure will be required to realize the full scientific potential of the HL-LHC

A Software "Upgrade" for the HL-LHC

- The HL-LHC will provide ***O*(100)** times the current data volume, with significantly increased data (pileup) and detector complexity.
- Most of the current software, which defines our capabilities, was designed **15-20 years ago**: there are many ***software sustainability challenges***
- Estimates of computing needs run faster than Moore's Law by factors of 3-30(?), but ***technology change*** will also make it challenging to exploit Moore's Law without ***software evolution***

HEP Software Foundation (HSF)

- The ***HEP Software Foundation*** (*HSF*) was created ~2 years ago as a means for organizing our community to address the software challenges of future projects like the HL-HLC
- An initial set of collaborative activities have begun (see recent HSF workshop at LAL-Orsay)
- The objectives of HSF as a community-wide organization are :
  - Share expertise and raise awareness of existing software and solutions
  - Catalyze new common projects
  - Promote commonality and collaboration in new developments to make the most of limited resources
  - Provide a framework for attracting effort and support to S&C common projects (new resources!)
  - Supporting career development for software and computing specialists
  - Provide a structure to set priorities and goals for the work
- The HSF is a ***HEP community effort*** and open enough to form the basis for collaboration with other sciences.

Where do we go from here?

- The HSF has demonstrated some initial collaborative activities between people working on different experiments. However, what is needed to address the future HEP software/computing challenges (HL-LHC and others) is ***additional dedicated resources for projects***.
- There are a couple of "common" software-focused projects today which have acquired "new" resources. Examples include:
  - DIANA-HEP software framework (link)
  - AIDA2020 software work package (link)

  Neither of these was really proposed or funded "as part of" HSF, but they are the kinds of projects that fit naturally under the HSF umbrella. How concretely do we go about doing that?
- Even more concretely: can we build something resembling a "software upgrade" project for HL-LHC?

Community White Paper (CWP) Overview

- ***Broad overview*** of grand challenge science (HL-LHC, HEP)
  - How can new approaches to computing and software enable and radically extend the physics reach of the detectors?
  - What computing and software research will be required so that (for example) 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***

Community Roadmap for HEP S&C

- As a next step for HSF and as a step towards software upgrades for HL-LHC, we are proposing a ***Community Roadmap for HEP Software and Computing***.
- In the form of a ***Community White Paper*** (CWP), it should describe a global vision for software and computing for the HL-LHC era and HEP in the 2020s
- The ***CWP*** should
  - include discussions of elements that are ***common to the HEP community as a whole*** (LHC community, etc.) as well as those that are specific to the individual experiments
  - discuss the relationship of common S&C elements to the ***broader community*** that utilizes scientific computing
  - be an element of the ***HL-LHC planning process*** and play a role in discussions surrounding scenarios for funding a "software upgrade" to realize the full physics potential of the HL-LHC

Detector Simulation, Triggering, Event Reconstruction and Visualization

- ***Simulation***. 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 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. 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.
- ***Triggering***. 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.
- ***Event Reconstruction***. 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***. 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.

Data Access and Management, Workflow and Resource Management

- ***Data Access and 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. High-granularity (e.g. 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 and Resource Management***. 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

Physics generators, Data analysis and Interpretation, Data and Software Preservation

- ***Physics generators***. 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***. 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.

- ***Data Interpretation and Software Preservation***. 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.

CWP Status

- The CWP roadmap plan was presented and discussed at the HSF meeting at LAL-Orsay. It was generally agreed that this is a necessary next step. It has also been discussed with some key members the CS and Theory and also communities (DOE-CCE, CERN OpenLab, WLCG, OSG! …)
- The proposal for a community roadmap, to be ***carried out by HSF***, was presented to the LHCC.
- In an effort to kickstart the process, we/HSF have worked with Ian Bird/WLCG to generate a charge for the CWP as part of the HL-LHC planning process. The current draft of the charge can be found here.
- Key elements include establishing working groups with the experiments and others, specific charges and timeline for workshops and the CWP draft itself (end of August 2017)

CWP Process (evolving)

- We propose a series of workshops over the next year to build the community roadmap:
  - Organization of working groups with experiments ~September. Pre-CWP meeting within HSF is also being scheduled.
  - A half-day (Sun ***9 Oct***) CWP-planning meeting just before CHEP
  - A more general ***"kick-off" workshop late-Nov/Dec*** (likely in Europe)
  - Several dedicated ***"topical" workshops in Jan-Jun 2017*** covering software required in the various areas:
    - Detector Simulation, Triggering, Event Reconstruction and Visualization
    - Data Access and Management, Workflow and Resource Management
    - Physics generators, Data Analysis and Interpretation, Data and Software Preservation
  - A ***final workshop in summer 2017*** (near CERN?)
- We should build on existing community activities when possible (e.g. DPHEP, Reco Algorithms Forum, IML, …).

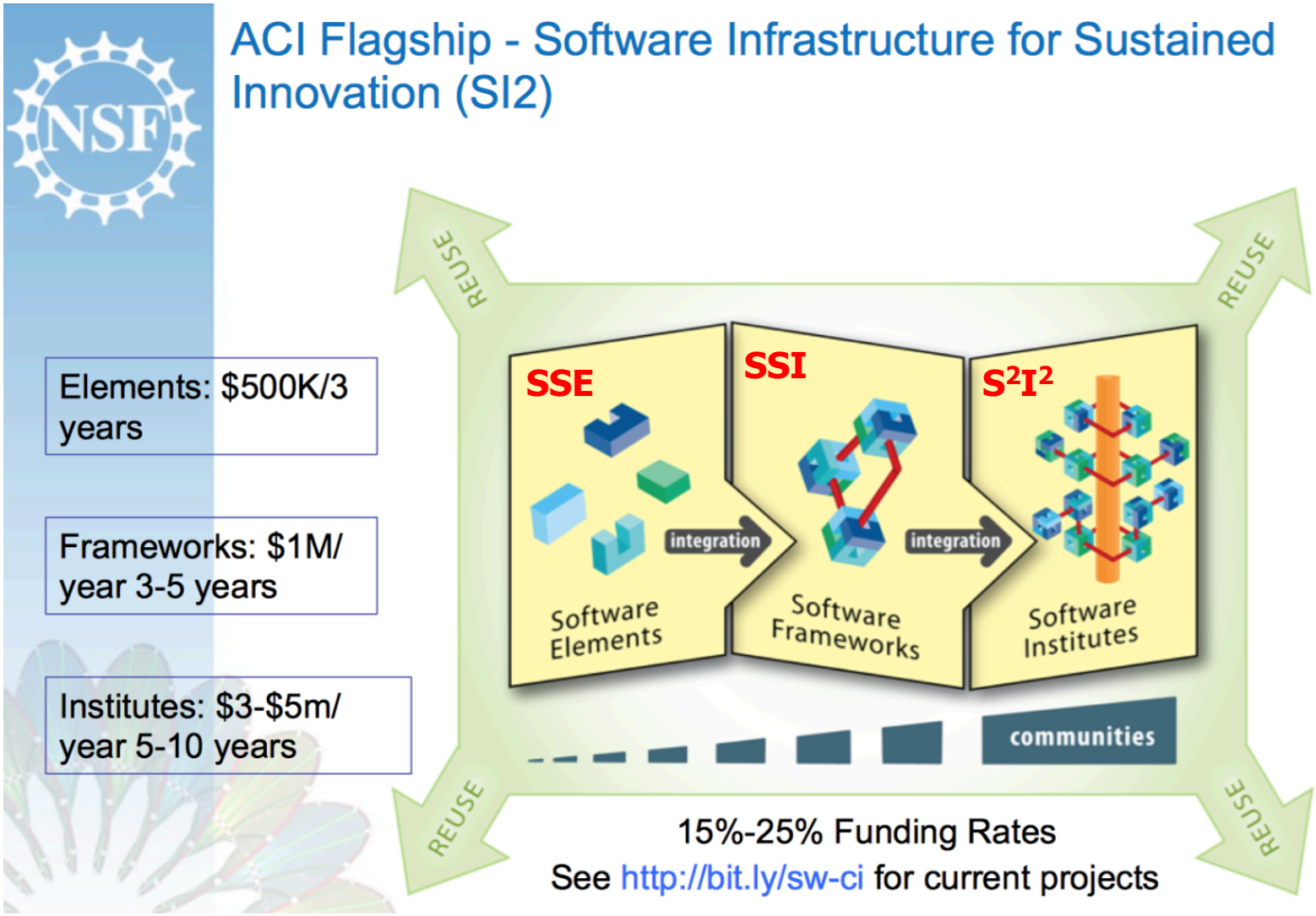
What could the CWP process accomplish?

- Going back to a subset of HSF goals listed earlier:
  - Catalyze new common projects
  - Promote commonality and collaboration in new developments to make the most of limited resources
  - Provide a framework for attracting effort and support to S&C common projects (new resources!)
  - Provide a structure to set priorities and goals for the work
- The workshop process, an eventual community roadmap white paper and (simultaneously) the pursuit of specific plans and proposals, will support precisely these goals

US-Specific Notes

- If we are aiming at a larger “software upgrade” project towards the HL-LHC, an additional ingredient is to find (or liberate) the resources to realize this roadmap.
- We need both initial exploratory R&D and eventual development projects!
- Both the NSF and the DOE have at least the notion of eventual resources and/or organization for new common projects in HEP (NSF: SI2, DOE: HEP CCE)
- We talk about the NSF here as (a) we involved in that part of the process and (b) the possible path indicated to us is a bit clearer
- A 2-day workshop involving the US HEP & CS communities is planned for late Fall somewhere in the US (likely UIUC/NCSA)

Software Infrastructure for Sustained Innovation (SI^2)



- Here I talk about the "Software Institute" (S^2I^2) class of awards.

NSF SI²-S²I² Software Institute

- NSF S²I² includes two subclasses of awards (two steps!):
  - ***Conceptualization Awards*** - which are planning awards aimed at organizing an interdisciplinary community and understanding their software requirements and challenges ($500k, 1-2 years)
  - ***Implementation Awards*** - which will be made to implement community activities that support software infrastructure, for example, such as those developed by the conceptualization awards ($3-5M/year, 5 years)
- The first solicitation for implementation proposals was last June; announcement of these awards expected to be soon
- We submitted a conceptualization proposal to NSF last August (2015): "***Conceptualization of an S2I2 Institute for High Energy Physics"*** (link). http://s2i2-hep.org/
- Status: awarded ***11 July***, with 1 July official start date

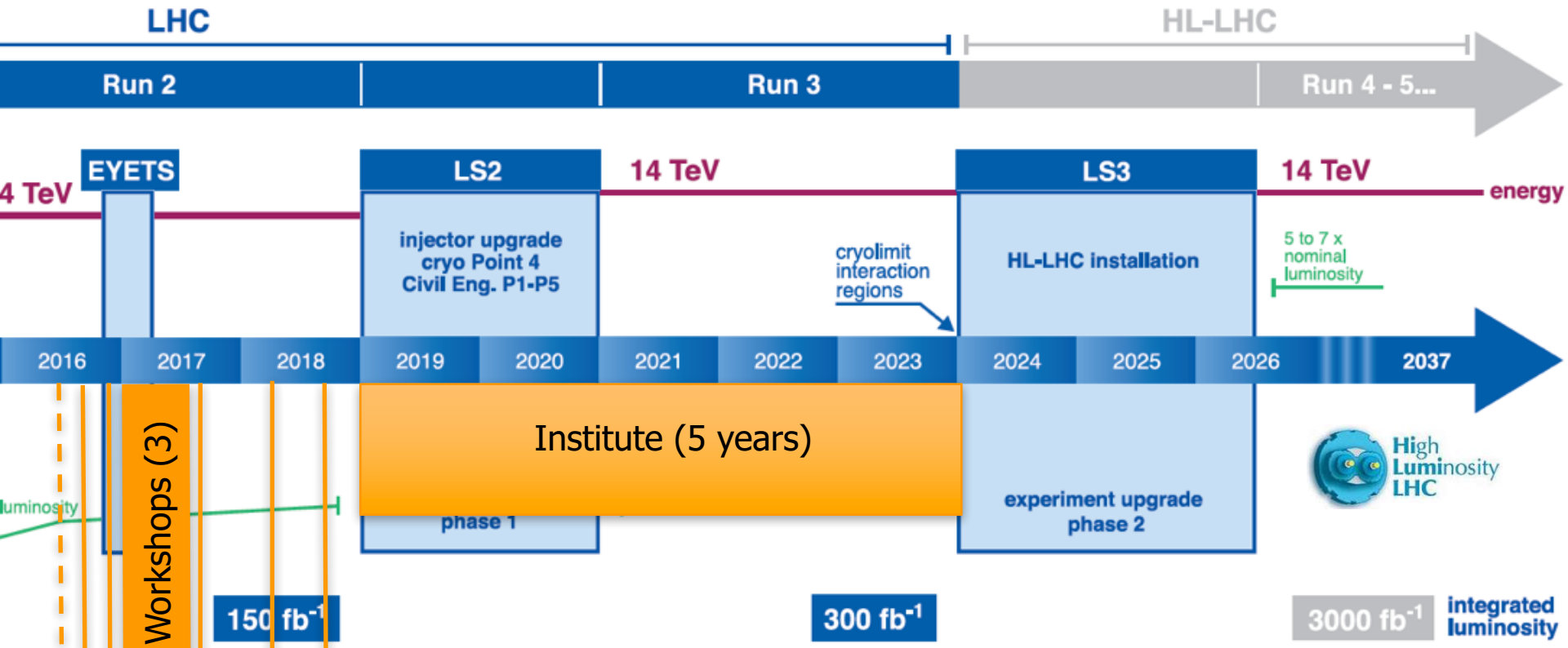
S^2I^2-HEP Deliverables

Major deliverables of the S^2I^2-HEP project are:

1. A roadmap document in the form of a ***Community White Paper*** (CWP) which aims to broadly identify the key issues of computing infrastructure and software R&D required to enable realization of the scientific potential of the large investment in upgraded detectors for the HL-LHC

2. A ***Strategic Plan*** (SP) for conceptualization of a *Scientific Software Innovation Institute* (S^2I^2) where U.S. university-based researchers can play an important role in key software infrastructure efforts that will complement those led by U.S. national laboratory-based researchers and international collaborators. The SP would presumably form the basis for a proposal for implementation of the Institute

CWP and S²I²-HEP (Success-driven) Timeline



Topical Workshops (3)

Institute award decision (?)

Institute proposal (?)

Final Workshop (July'16) – CWP completed few months thereafter

Kick-off (Dec'16)

Workshop Organization (Sep'16 and Oct'16)

Conceptualization recommended for funding (Jun'16)