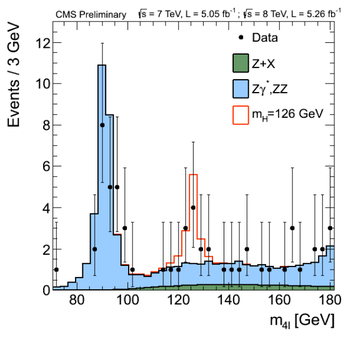
S2I2: Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)

http://iris-hep.org

PI: Peter Elmer Co-PIs: Brian Bockelman, Gordon Watts
Institutions: Princeton, Morgridge Institute, U. Washington, UC-Berkeley, University of Chicago, University of Cincinnati, Cornell University, Indiana University, MIT, U.Michigan-Ann Arbor, U.Nebraska-Lincoln, New York University, Stanford University, UC-Santa Cruz, UC-San Diego, U.Illinois at Urbana-Champaign, U.Puerto Rico-Mayaguez and U.Wisconsin-Madison

Award #: 1836650

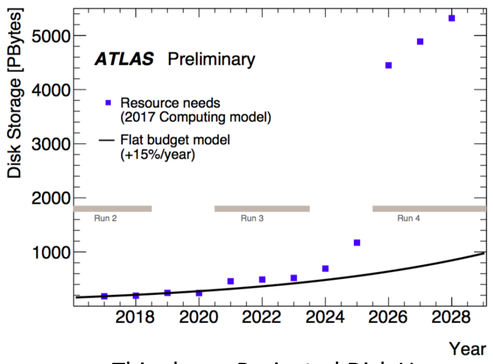
Science Drivers:
Discoveries beyond the Standard Model of Particle Physics



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

From "Building for Discovery - Strategic Plan for U.S. Particle Physics in the Global Context" - Report of the Particle Physics Project Prioritization Panel (P5):

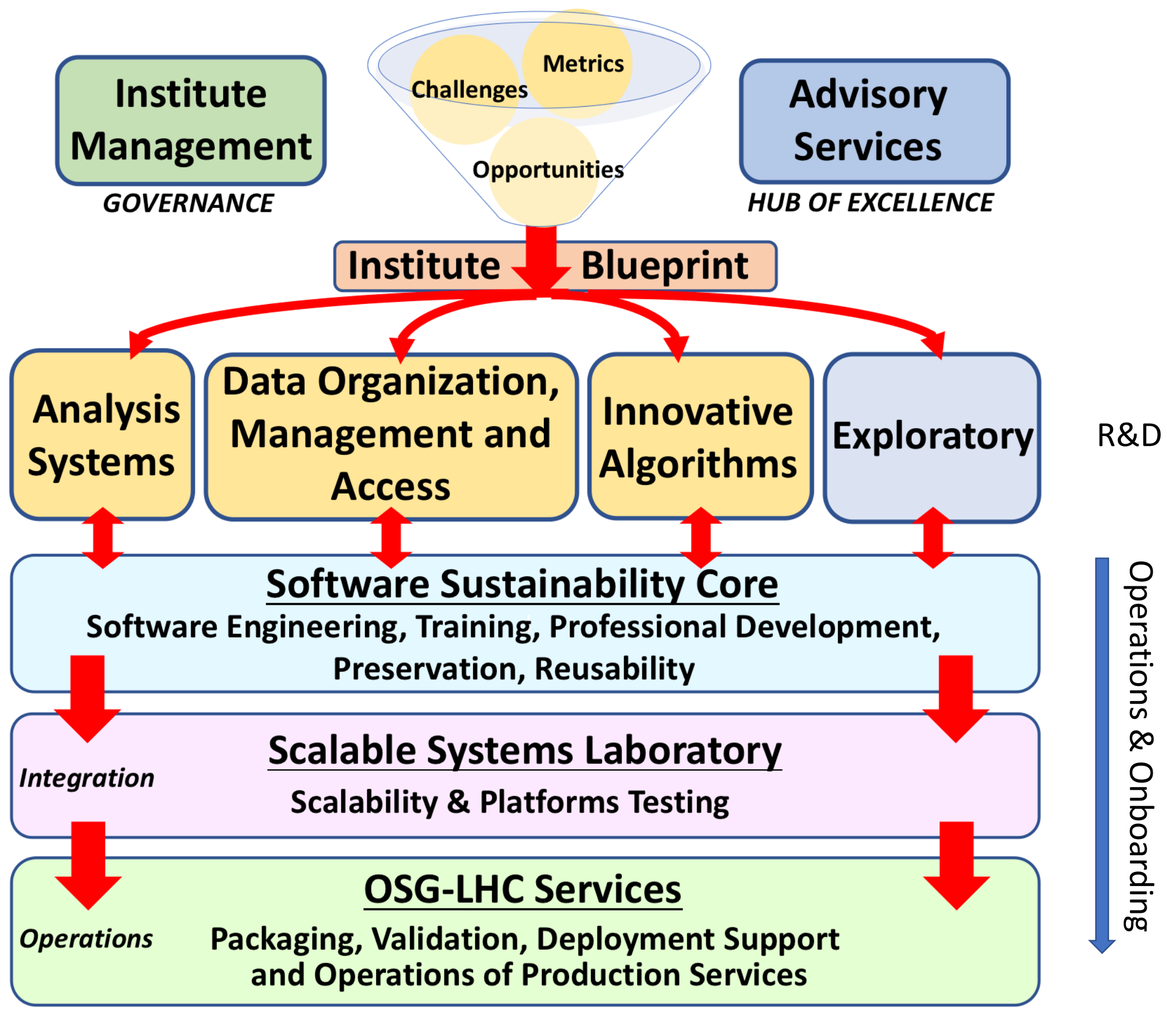
Computational and Data Science Challenges of the High Luminosity Large Hadron Collider (HL-LHC) and other HEP experiments in the 2020s



This shows Projected Disk Usage. Projected CPU usage looks very similar.

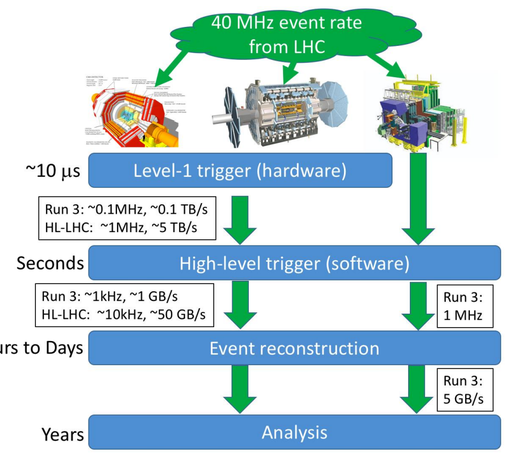
The HL-LHC will produce exabytes of science data per year, with increased complexity: an average of 200 overlapping proton-proton collisions per event.

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).



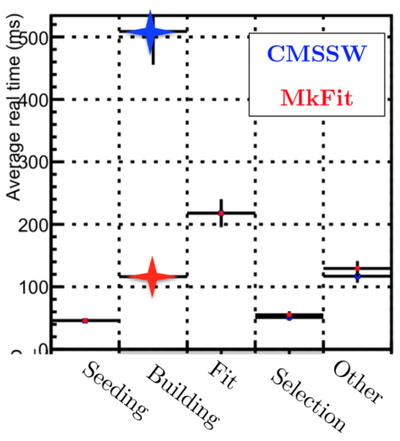
Innovative Algorithms - Trigger/Reconstruction

Algorithms for real-time processing of detector data in the software trigger and offline reconstruction are critical components of HEP's computing challenge. Focus on redesign tracking algorithms for HL-LHC and how take advantage of of major advances in ML.



MkFit – Fast Tracking for CMS

Develop track finding/fitting implementations that work efficiently on many-core architectures (vectorized and parallelized algorithms). See 4x faster track building with similar physics performance in realistic benchmark comparisons. Currently being integrated into CMS's production reconstruction program infrastructure.



Community Building

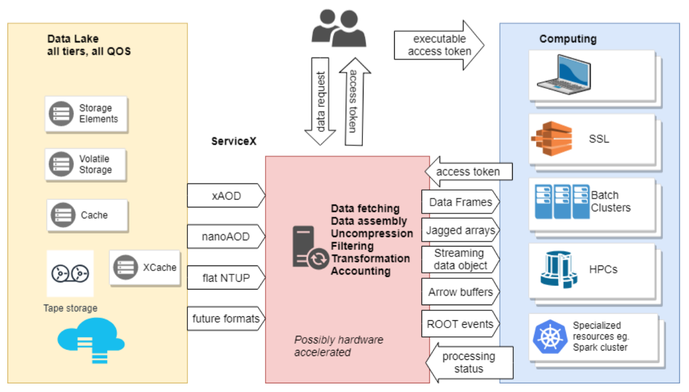
IRIS-HEP has community building as part of its original DNA: it was born from an NSF funded S2I2 Conceptualization Project: 17 workshops from 2016-2017 that took place all over the world. The participants in the workshop then created a community white paper which has been guiding the field since.

IRIS-HEP continues to build community with virtual weekly *topical meetings*, with small in person Blueprint workshops to build field consensus, and sponsorship of more general HEP conferences, like PyHEP to take place in Austin July 11-13, 2020, just adjacent to the SciPy2020 workshop.

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.

Intelligent Data Delivery

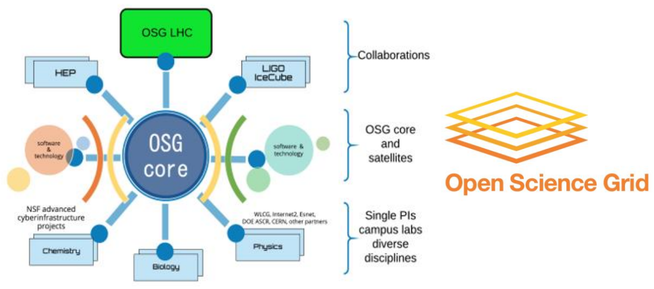


In the HL-LHC era, we must deliver more events - and at lower latencies - if the analysts want to make progress!

- Low-latency delivery of events requires transformation from long-term archival formats that we want to decrease data size.
- Users should be enabled to work on a multitude of data formats (esp. non-ROOT) without having to write them to disk.

We are currently prototyping an Intelligent Data Delivery service to extract events from a data lake for fine-grained processing and deliver events to analysis facilities at a high data rate.

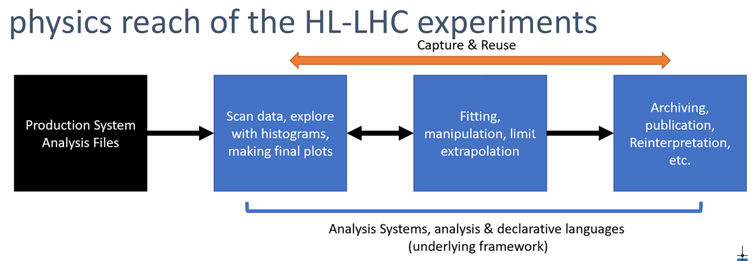
The Open Science Grid



The OSG is dedicated to the advancement of all open science via the practice of Distributed High Throughput Computing, and the advancement of its state of the art. The LHC makes major use of OSG for all its US computing responsibilities. This is funded through IRIS-HEP. This supports OSG in Organizing a transition away from Globus to third party tokens, the Central Services Paradigm using containers, etc., to better separate hardware and software. December 2019 Stats: 129 million hours of computing for US LHC

Analysis Systems

Develop sustainable analysis tools to extend the physics reach of the HL-LHC experiments

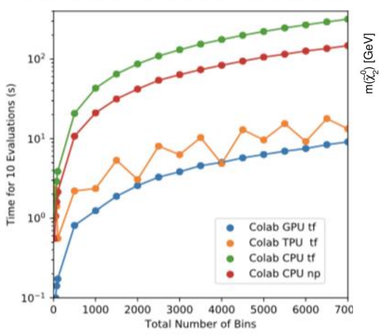


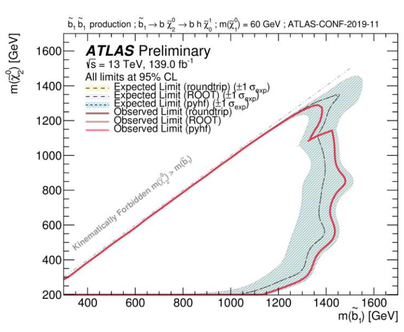
Statistical Models

Implementation of widely used statistical tool in modern frameworks. Implementation of HistFactory likelihood as a **computational graph of multi-dimensional array operations**. Use of array operations through a common API layer around high performance array libraries:

NumPy TensorFlow PyTorch

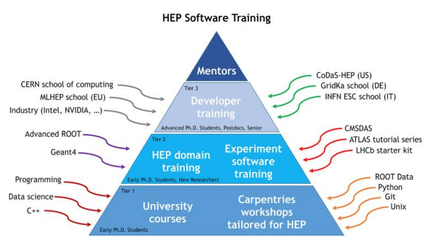
By using these tools we get benefits: like Auto-Differentiation, and a large library of minimization algorithms like **MINUIT** and **scipy.minimize**, and **hardware acceleration**.





The limit plot was made in less than 30 minutes with the **pyhf** tool and used to take more than 10 hours with our standard tools. Available on **pypi**.

Software Training



Using the Software Training Framework to guide our efforts, provide both base and expert training and support to the field by running standalone workshops (e.g. CoDaS-HEP 2019, pictured) as well as add-on workshops that are co-located with other large HEP physics meetings. In the past year we have been involved in more than 10 of these events.

We partner with other organizations, leveraging their expertise.

FIRST-HEP
OAC-1829707, 1829729