



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

Computational and data science research to enable discoveries in fundamental physics

IRIS-HEP is a software institute funded by the National Science Foundation. It aims to develop the state-of-the-art software cyberinfrastructure required for the challenges of data intensive scientific research at the High Luminosity Large Hadron Collider (HL-LHC) at CERN, and other planned HEP experiments of the 2020's. These facilities are discovery machines which aim to understand the fundamental building blocks of nature and their interactions. [Full Overview](#)

The IRIS-HEP project was funded on 1 September, 2018.

G. Watts, IRIS-HEP Steering Board Meeting #18



IRIS-HEP Steering Board Meeting #18

G. Watts

For the IRIS-HEP Executive Board

2023-10-17

“The IRIS-HEP Steering Board represents the Institute’s stakeholders to provide, to the Executive Board, the stakeholder’s input on the priorities, execution, and strategy of the Institute.”



Thank You

Danilo Piparo (CERN)
CMS

Paolo Calafiura (LBNL)
US ATLAS Ops Program

Simone Campana (CERN)
WLCG

David South (DESY)
ATLAS

Oliver Gutsche (FNAL)
US CMS Ops Program

Patrick Koppenburg (NIKHEF)
LHCb

Graeme Stewart (CERN)
HSF

Ken Herner (FNAL)
The OSG Council



Welcome

steering-board@iris-hep.org

(you)

exec-board@iris-hep.org

(us)



Next Meeting Dates

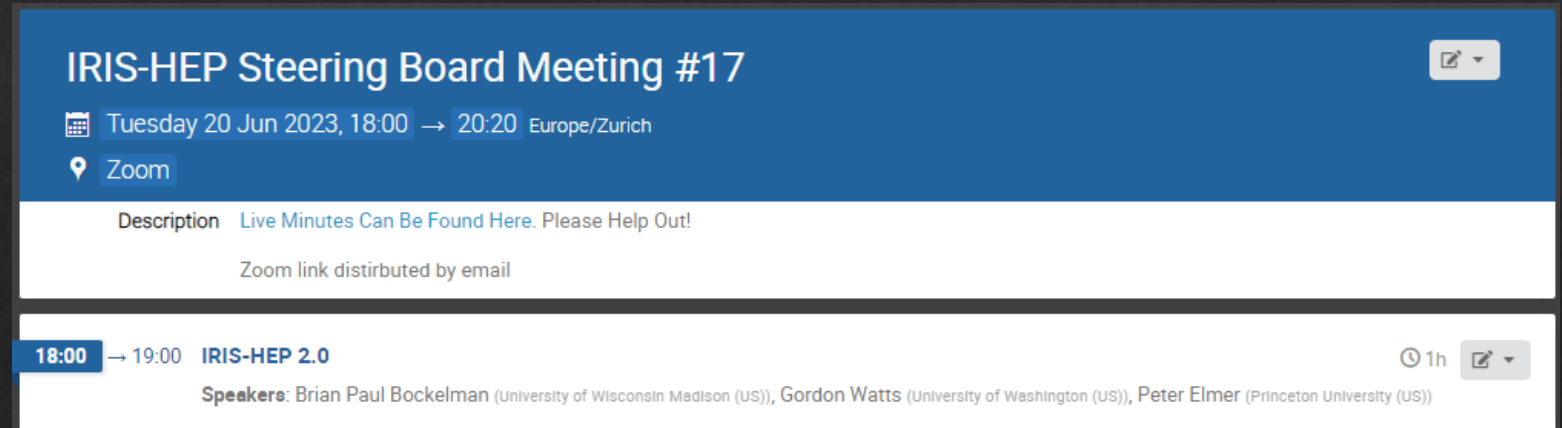
January 16, 2024



Today

IRIS-HEP 2.0

Please use the google doc circulated in the email to add comments or make any notes you want us to track!



The screenshot shows a Zoom meeting invitation for "IRIS-HEP Steering Board Meeting #17". The meeting is scheduled for Tuesday, June 20, 2023, from 18:00 to 20:20 in the Europe/Zurich time zone. The location is Zoom. The description states: "Live Minutes Can Be Found Here. Please Help Out!" and "Zoom link distributed by email". A specific agenda item is listed for 18:00 to 19:00, titled "IRIS-HEP 2.0", with speakers: Brian Paul Bockelman (University of Wisconsin Madison (US)), Gordon Watts (University of Washington (US)), and Peter Elmer (Princeton University (US)).



Project Information

The screenshot shows the IRIS-HEP website with a navigation menu. The menu items are: Analysis Systems, Blueprint Activity, Data Organization, Management and Access (DOMA), Innovative Algorithms, Open Science Grid (OSG-LHC), Scalable Systems Laboratory, Training, Education and Outreach, Impact Beyond HEP, Presentations, Publications, and Projects. The 'Data Organization, Management and Access (DOMA)' item is highlighted in blue. Below the menu, there is a section titled 'Computational and research to enable fundamental physics' and another titled 'News and Featured Stories:' with a photo of three people.

Data Organization, Management and Access (DOMA)

The HL-LHC era will provide enormous challenges in the area of Data Organization, Management and Access (DOMA). The LHC will provide a significantly increased number of events and increased event complexity, both of which will drive much larger data sizes - with no changes in how the LHC community functions, the total increase in data volume may be a factor of 30.

Given the LHC experiments are, combined, managing nearly an exabyte of data, such a significant increase in volume is unmanageable. New mechanisms and techniques are necessary to more efficiently manage storage resources; the DOMA area in IRIS-HEP is working on the R&D necessary to affect such change.

It is not only data volumes that are potentially disruptive to the HL-LHC physics program; the extraordinarily large number of events (potentially 150 billion simulated and recorded events per year per experiment) presents a challenge in data management for users. Along with the analysis systems team within IRIS, DOMA is working on improved techniques for delivering events to users.

Contact us: doma-team@iris-hep.org

DOMA Projects



Caching Analysis Data

Cached-based placement of analysis datasets.
[More information](#)

Intelligent Data Delivery Service

Delivering Data. Better.
[More information](#)

Per-project information is available on all IRIS-HEP projects.

Caching Analysis Data

Significant portions of LHC analysis use the same datasets, running over each dataset several times. Hence, we can utilize cache-based approaches as an opportunity to efficiency of CPU use (via reduced latency) and network (reduce WAN traffic). We are investigating the use of regional caches to store, on-demand, certain datasets. For example, the UCSD CMS Tier-2 and Caltech CMS Tier-2 joined forces to create and maintain a regional cache that benefits all southern California CMS researchers.

These in-production caches have shown to save up to a factor of three of WAN bandwidth compared with traditional data management techniques.

Presentations

- 23 Apr 2020 - "How CMS user jobs use the caches", Edgar Fajardo, XCache DevOps SPECIAL
- 22 Apr 2020 - "XRootD Transfer Accounting Validation Plan", Diego Davila, S&C Blueprint Meeting
- 27 Feb 2020 - "XCache", Edgar Fajardo, IRIS-HEP Poster Session
- 5 Nov 2019 - "Creating a content delivery network for general science on the backbone of the Internet using xcache", Edgar Fajardo, CHEP 2019
- 5 Nov 2019 - "Moving the California distributed CMS xcache from bare metal into containers using Kubernetes", Edgar Fajardo, CHEP 2019
- 12 Sep 2019 - "OSG XCache Discussion", Frank Wuerthwein, IRIS-HEP retreat
- 31 Jul 2019 - "CMS XCache Monitoring Dashboard", Diego Davila, OSG Area Coordination
- 8 Jul 2019 - "XCache Initiatives and Experiences", Frank Wuerthwein, pre-GDB meeting on XCache

(often, but not always)





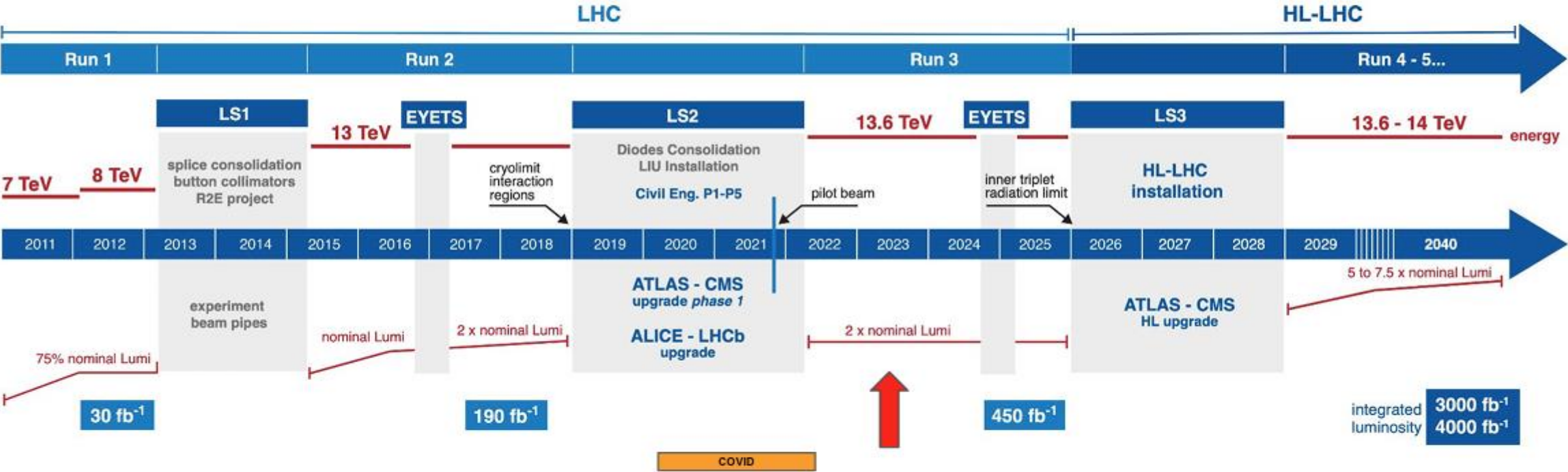
iris hep 2.0

**Institute for Research & Innovation
in Software for High Energy Physics**





LHC / HL-LHC Plan



Institute Conceptualization and Community White Paper Process

S2I2-HEP

IRIS-HEP Institute (Proposed Renewal)

Design Execution

Snowmass

U.S. HEP Community Planning Process

HL-LHC Software and Computing Gaps

The four software and computing gaps are:

- G1. **Raw resource gaps:** The HL-LHC dataset will be enormous. Event complexity and count will each go up by about an order of magnitude. If no improvements to algorithms or resource management techniques are made, the HL-LHC experiments will simply be unable to process and store the data necessary for the science program.
- G2. **Scalability of the distributed computing cyberinfrastructure:** It is insufficient to buy cores and disk alone – the cyberinfrastructure used by the experiments must also scale to support the volume of hardware. This challenge is especially acute when it comes to data transfers: both the software must be ready and the shared networking resources (e.g., ESNet in the US) must be appropriately managed.
- G3. **Analysis at scale:** Analysis at the HL-LHC will be markedly different for two reasons: (a) the scale of the datasets involved and (b) the use of next-generation techniques (such as the latest machine learning techniques) to increase the scientific reach of each result. The former will require users to heavily utilize dedicated ‘analysis facilities’, optimized for high data rate I/O and the latter will require new services and data management techniques to be developed.
- G4. **Sustainability:** HEP is a facilities-driven science - the cyberinfrastructure assembled for an experiment must last or evolve on the decadal scale. This limits some strategies to cyberinfrastructure - for example, it is impossible for LHC to “do it yourself” and own the entire software stack. Specific sustainability strategies must be implemented even at the R&D phase to ensure that the cyberinfrastructure put in place at the beginning of the experiment is one the community can afford.

Role of the IRIS-HEP Institute

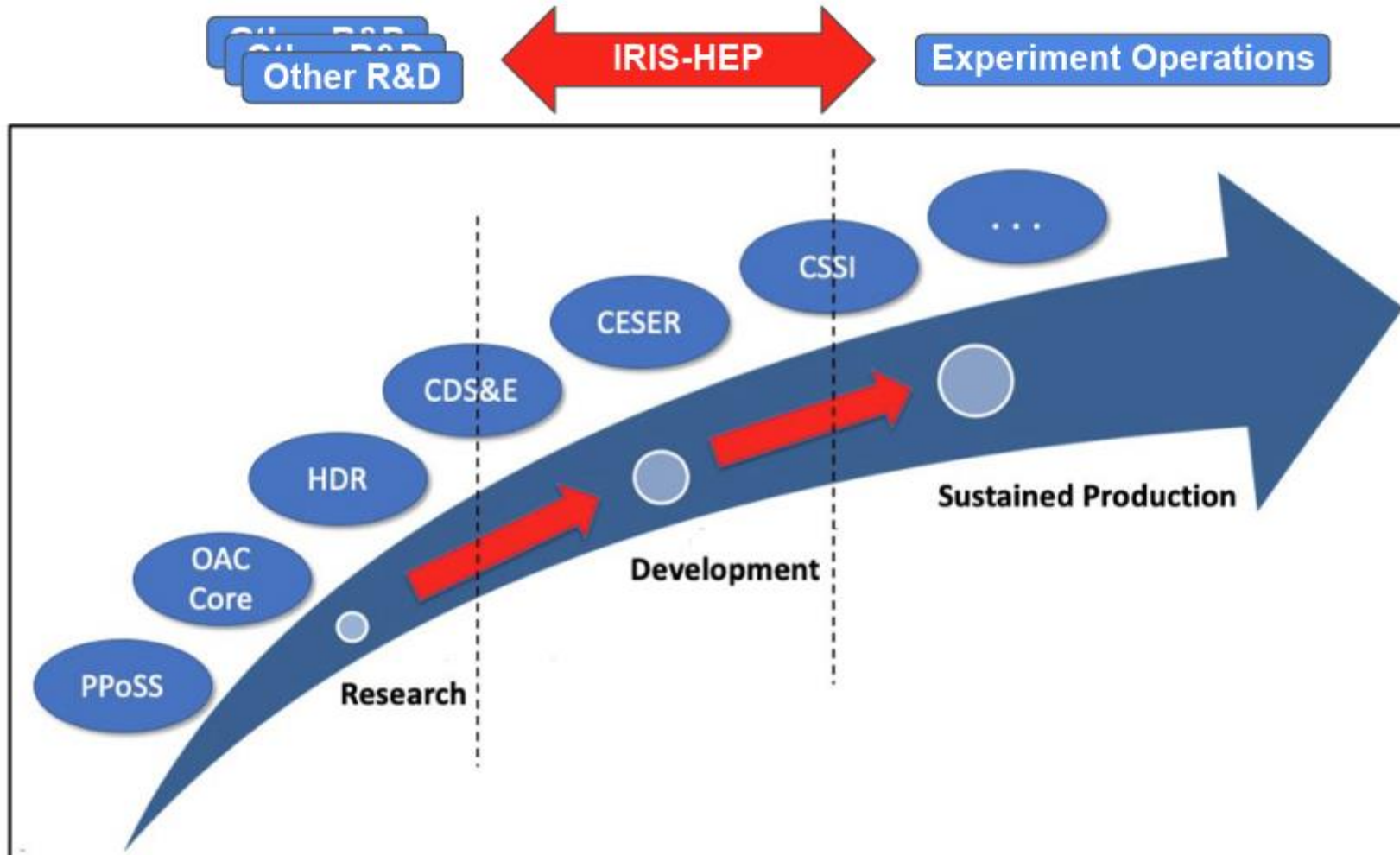


Figure 2: Data and Software CI pathways to production.

From ["OAC Vision: Blueprint for a National Data and Software Cyberinfrastructure"](#)



Key Evolutions for IRIS-HEP in the next 5 years

1. Evolution from early stage “research” to “development” activities (sometimes referred to as “R&d → r&D”)
2. Increasing focus on integration activities as well as the paths to deliver output to the experiments/collaborations
3. Sustainability and leaving a significant community legacy, including career paths for the next generation of HEP software/computing leadership



Systems Integration and Delivery

The following presentations are structured to align to integration/delivery paths:

Analysis Grand Challenge (AGC): The AGC aims to execute realistic analyses at the scale and complexity envisioned by the HL-LHC using a set of tools, facilities, and services developed within IRIS-HEP.

Tracking Challenge: Efficient reconstruction of charged particle tracks is a key computational challenge in the high pileup environment at the HL-LHC.

Data Grand Challenge (DGC): The DGC is realized as a set of global data challenges coordinated with the Worldwide LHC Computing Grid (WLCG).

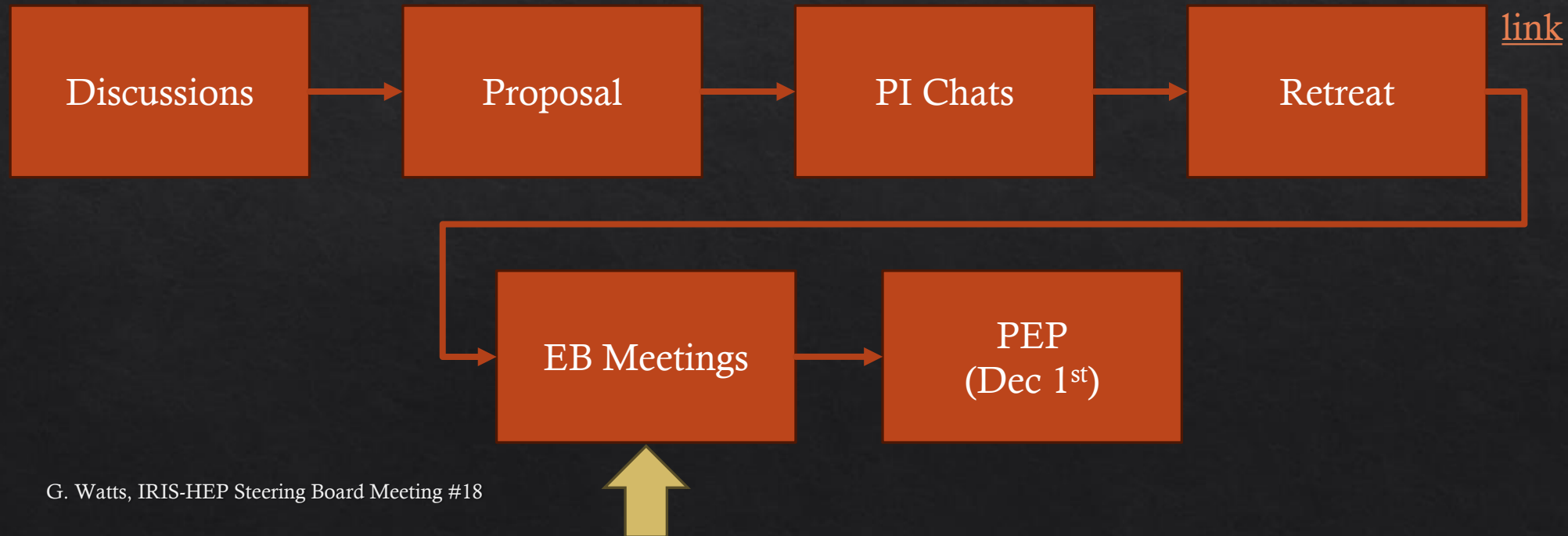
OSG Services for LHC (OSG-LHC): Evolution of the fabric of services necessary for the distributed high-throughput computing infrastructure required by the LHC experiments.

Broader Impacts and Intellectual Hub (including Training Challenge): The Institute's larger role in the LHC & physics community and its broader impacts (training, education and outreach).

Getting To the Project Execution Plan

This document is our official list of Milestones, Metrics, and Risks that we agree upon with the NSF.

- See the current dashboards for [milestones](#), [metrics](#), and [risks](#) (the risks is out of date).



IRIS-HEP All-Hands Retreat



Second in-person get-together since COVID! Had a large range of people from our team, to US OPS members



EB Meetings

- Oct 9 – DOMA
- Oct 16 – Analysis Grand Challenge
- Oct 23 – Scalable Systems
Laboratory, Innovative Algorithms
- Oct 30 – Analysis Systems,
SSC/Training

Draft of WBS and milestones, metrics, risks

OSG and Translational AI are to be scheduled



DOMA (DRAFT)

DOMA WBS – IRIS-HEP v2

WBS Areas:

- **W4.1 Scaling to HL-LHC Data Rates:** Work on technologies and tests toward scaling the reference platform (XRootD) to HL-LHC data rates.
- **W4.2 Authorization technology overhaul:** Help the community transition authorization technologies to use capability tokens.
- **W4.3 Delivering columnar data:** Have ServiceX become a platform in the LHC community for data transformation & delivery.
- **W4.4 Coffea-Casa:** Deploy an integrated ‘showcase’ of IRIS-HEP and community analysis ecosystem technologies.
- **W4.5 XRootD core:** Sustain and help grow the use of XRootD within the community.

DOMA Milestones and Deliverables - Y1

ID	Description	Date	WBS
D4.1	Coffea-Casa used as part of one production analysis facilities in both ATLAS and CMS.	Dec 2023/ Y1Q1	W4.4
D4.2	All the U.S. LHC T2s in the US support bearer tokens on their storage endpoints for Third Party Copy transfers.	Dec 2023/ Y1Q1	W4.2
D5.2	(With SSL) ServiceX deployed inside FABRIC at CERN.	Jan 2024/ Y1Q2	W4.3
D4.3	Rucio/SENSE integration is included as part of the DC24.	Mar 2024/ Y1Q2	W4.1
D4.4	Successful execution of DC24, meeting its data transfer and technology goals.	Mar 2024/ Y1Q2	W4.1
D4.5	ServiceX used for >=2 physics analyses as part of Coffea-Casa.	June 2024/ Y1Q3	W4.3

DOMA Milestones and Deliverables

ID	Description	Date	WBS
D4.6	Demonstrate analyses running at 200Gbps as part of the Analysis Grand Challenge.	Dec 2024/ Y2Q1	W4.3, W4.4, W4.5
D4.7	50% of production transfer volume is performed with tokens	Mar 2025/ Y2Q2	W4.2, W4.5
D4.8	Demonstrate the capability of XRootD to scale beyond 400Gbps.	June 2025/ Y2Q3	W4.1, W4.5
D4.9	All production transfers executed using tokens	Jan 2026/ Y3Q2	W4.2
D4.10	Successful execution of DC26, meeting its data transfer and tech. goals	Mar 2026/ Y3Q2	W4.1, W4.2
D4.11	Demonstrate analysis running at full HL-LHC scale as part of the AGC.	Mar 2027/ Y4Q2	W4.3, W4.4, W4.5
D4.12	Successful execution of DC28, showing readiness for HL-LHC transfers	Mar 2028/ Y5Q2	W4.1, W4.2

Analysis Grand Challenge Effort

Changing how the AGC works internally

2 new WBS areas

WX.1 New sub-activity which focus delivering **integration test** for IRIS-HEP (SSL, DOMA, AS) = **internal**

- Run technical demos (AS, DOMA, SSL)
- High-level design of training workshop content (targeting end-users)
- Benchmarking & performance goals (yearly milestones)
- Stability / reliability of pipeline

WX.2 New sub-activity which will focus on building **community** = **external**

- Community events (e.g. contributions from AFs)
- Engaging with experiments and operations programs
 - AGC on (US) facilities
 - Providing expertise to experiment-specific demonstrators
- Blueprint workshops
- Running training workshops

Splitting the internal technical and external interface into two overlapping teams

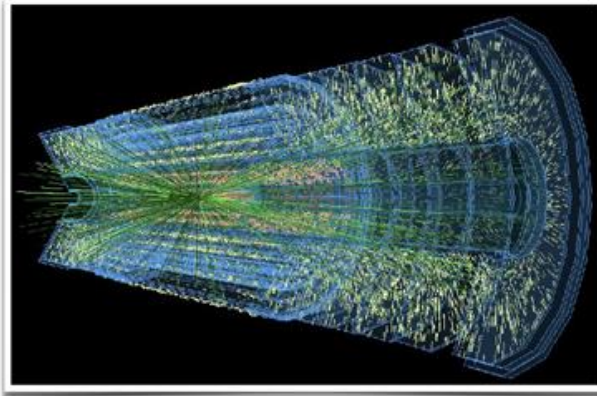
- Spread some of the effort
- Bring new people into the process
- Internally focus more on scale and functional benchmarks

Also change the name to reduce confusion!



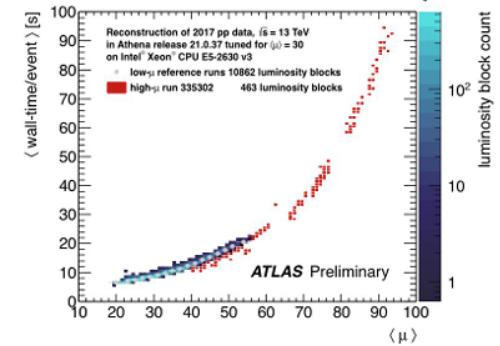
Tracking challenge for HL-LHC

- Online trigger will need to process **1 trillion events/yr** (10^{15} tracks)
 - 10x Run 3
- Up to **500 billion events/yr** will need to be processed offline
 - Up to 10x Run 3



Tracking Challenge for HL-LHC

- Upgraded accelerator
 - **non-linear increase** in collisions per bunch crossing (pileup)
- Detector upgrades
 - new detector technologies
 - additional **channels**
- Increased **event rates** due to trigger upgrades
- Evolving **heterogeneous** computing architectures



IRIS-HEP Tracking Strategy

- To address this challenge, our strategy is as follows
 - Re-engineer existing algorithms for new technologies and new detectors
 - MkFit, ACTS, LST
 - Explore novel algorithms, e.g. using machine learning
 - OCT
- Important to pursue multiple R&D strategies in parallel

ACTS = A Common Tracking Software
LST = Line Segment Tracking
OCT = Object Condensation Tracking

The OSG Consortium

- **OSG** is dedicated to the advancement of all of open science via the practice of Distributed High Throughput Computing, and the advancement of its state of the art.
- It is a collaboration between IT, software, and science organizations.
 - The consortium aggregates human and computing resources from many funded projects.
- It is governed by the OSG Council, maintaining its by-laws, and electing an executive director for 2 year renewable terms to coordinate a program of work.

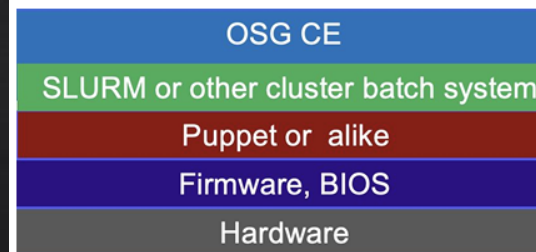
OSG-LHC is IRIS-HEP's contribution to this consortium

Big Picture Objective

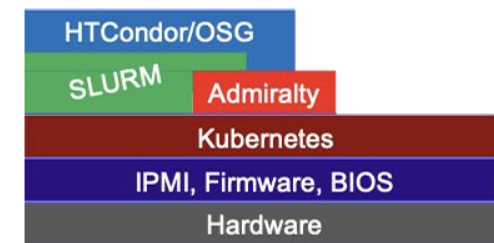
- Maintain software APIs, Containers, & services to support continued operational stability and global integration => don't rock the boat !!!
- Integrate evolutionary & revolutionary new services and paradigms => rock the boat without sinking the boat, or letting passengers fall overboard.
- **Evolution:**
 - Scaling out of data capabilities (See DGC)
 - Complete Token Transition for data access (See DGC)
 - Computing beyond x86: GPUs, FPGAs, and other "Domain Specific Architectures"
 - HPC integration
- **Revolution:**
 - Introduction of "engineered networks" (See DGC)
 - Integration of Analysis Facility into the standard Tier-2 hardware infrastructure (see AF)
 - Co-scheduling Dedicated AI hardware for Inference
 - Change in operations responsibility towards a global container model

Opportunity for Radical Change

Traditional stack deployed by staff at each Tier-2



Modern stack deployable from remote



Opportunity to radically rethink who should be responsible for what services & at what granularity, because everything can be operated from remote => See NRP



Challenges and Opportunities

Not all HEP students can attend university-offered software courses No standard curricula for HEP students exists

Democratize science by making software prerequisites accessible to everyone

Experiments need **Cyberinfrastructure professionals** and lifelong learners

We need a **unified, scalable, and sustainable** software training framework powered by the entire community

IRIS-HEP is leading training efforts and powered the **HSF Training WG**



3

Challenges and Opportunities

We need a **unified, scalable, and sustainable** software training framework

Unified

- Material and events should be **centrally listed** and **discoverable**
- Concentrate efforts by prioritizing **cross-experiment** content
- A community must **guide, support, and coordinate**

Scalable

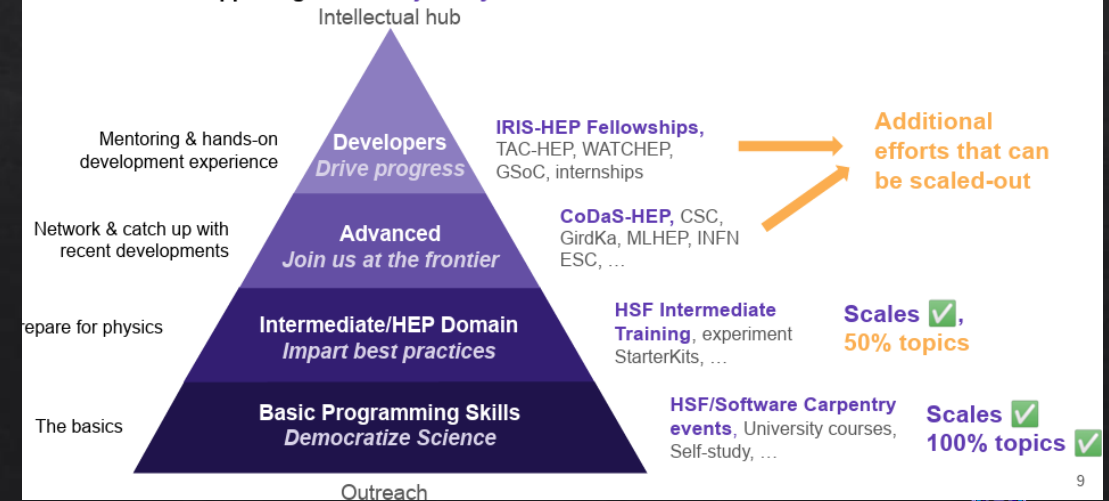
- The training initiatives need to reach **O(1000) students/year** ⇒ material must be teachable by **multiple instructors**
- **Self-study** must not be an afterthought

Sustainable

- Material must be **open source** and **maintained collaboratively**
- **Incentives & recognition** important motivators

Training Cyberinfrastructure professionals

IRIS-HEP is supporting the whole journey



9

Blueprint Meetings

We are re-doing our blueprint landing page (not quite public yet)

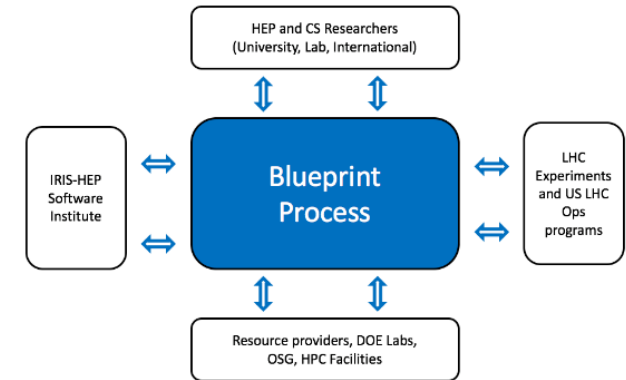
- Moving off google docs and using the website infrastructure
- [Comments welcome](#) (but please don't share yet).

More importantly – what meetings should we be helping to organize in the future?

- Your input is always helpful – this group has a global view of S&C in HEP that few other groups do.
- One of the few groups where (almost?) everyone is aware of what the LHCC is going.

IRIS-HEP Blueprints

The Blueprint Activity is designed to inform the development and evolution of the IRIS-HEP strategic vision. At its core, it is a series of workshops that bring together IRIS-HEP team members, key stakeholders and domain experts from disciplines of importance to the Institute's mission. The Blueprint Activity facilitates the Institute's role as an intellectual hub for software and computing R&D in high-energy particle physics and beyond.



Blueprint Dashboard

Completed / Scheduled

Topic / Title	Focus Area(s)	Dates	Location	Status	Summary Report / Notes
Analysis Systems R&D on Scalable Platforms	AS, SSL	2019-06-21	NYU	Complete	report
Fast Machine Learning & Inference	IA, SSL	2019-09-10	FNAL	Complete	report
A Coordinated Ecosystem for HL-LHC Computing R&D		2019-10-23	CUA	Complete	
Software Training	SSC	2020-02-20	Virtual	Complete	report
Sustainable Software in HEP	SSC	2020-07-22	Virtual	Complete	report
Future Analysis Systems and Facilities	AS, DOMA, SSL, OSG-LHC	2020-10-26	Virtual	Complete	
Fast Machine Learning for Science	IA	2020-11-30	Virtual	Complete	report
Portable Inference	IA	2020-12-04	Virtual	Complete	
Virtual Meeting on Virtual Meetings		2021-05-05	Virtual	Complete	report
Differentiable Programming for the AS Grand Challenge	AS	2021-12-01	Virtual	Complete	report
HSF/WLCG Analysis Facilities Forum Kick-off	AS	2022-03-25	Virtual	Complete	AF Forum
A Coordinated Ecosystem for HL-LHC Computing R&D		2022-11-07	DC	Complete	
Software Citation and Recognition in HEP	SSC	2022-11-22	Virtual	Complete	

Proposed

Topic / Title	Focus Area(s)	Dates	Location	Status	Summary Report / Notes
Analysis Preservation & Open Access Data	SSC, AS			Proposed	
Analysis Software Ecosystem	AS			Proposed	
Integration of Data Management & Analysis Services (Coffee-Casa, ServiceX, Skyhook, etc)	DOMA			Proposed	
SSL and the IRIS-HEP Grand Challenges	AS, DOMA, SSL, SSC		Virtual	Proposed	
Strengthening Theory & Experiment Connections	AS, IA			Proposed	

Field-Wide Blueprint Proposals

- ◇ Analysis Preservation
 - ◇ Experiments seem to be pulling back a little from HEPData, and perhaps more from REANA
 - ◇ Time to gather experimentalists and theorists to see what the appropriate level is, tooling?
 - ◇ Should coordinate with [LHC Reinterpretation Forum steering group](#) and [FAIROS-HEP](#)
 - ◇ How does this work with new Services that will be used (e.g. dask, ServiceX, ML, etc.)
 - ◇ Summer 2024?
 - ◇ Also, [Nelson memo](#) (OpenData)
- ◇ What is a HL-LHC Analysis (Benchmarks for Analysis)
 - ◇ Data sizes? Methods? Precision vs Discovery
 - ◇ Representative analyses? (“a analysis”)
 - ◇ Ignore solutions explicitly - just talk about framing the problem
 - ◇ Can we build a new ADL benchmarks
 - ◇ LHCC request coming up shortly?
- ◇ XROOT Sustainability (very tentative)



USA-Wide Blueprint Proposals

- ◇ Coordinated Ecosystem for HL-LHC Computing
 - ◇ Should occur ~ every two years. Last one in Nov 2022 -> Fall 2024.
 - ◇ Follow up earlier with the **international benchmarking** and **P5 report** released? Spring 2024?
 - ◇ The Nelson memo (OpenData)?
 - ◇ DPF's Coordinating Panel for Software and Computing



Next Steps

- ◇ Planning will continue in the Executive Board Meetings
 - ◇ By end of October should finish draft presentations
 - ◇ Careful drafting and integration will occur in November
 - ◇ EB has explicit ex-officio members from the US OPS program, as well as many of us who are part of the experiments. We welcome others who are interesting, please get in touch.
- ◇ Looking forward to International Benchmarking Study and P5 planning document release in November and December
- ◇ Next Steering Board Meeting is January 16th
 - ◇ Paolo will give his talk “Experiments and Large External Collaborators – Improving integration and Planning.”



Questions & Comments?

