Gordon Watts

Grew up 45 minutes from here
Got into programming in HS, and then DAQ systems for the AMY experiment
Hardware

Fermilab Experiments
Coding up SVX cooling interlock experiment (array programming language!)
But 90% physics (learned coding was not good for your cv (top discovery)

Low-level coding for a high-level trigger for DZERO. Distributed Computing and Hardware
Physics (top, single top, higgs)

Professor since 1999
DZERO, ATLAS, etc.
Got interested in DSL's, especially after tenure

IRIS-HEP
Large software Institute spanning 20 institutes, partially supporting ~60 people.
with PI's, Research Software Engineers, Post-docs, Graduate students, post-bac's, and undergraduates that have joined to help the field address the HL-LHC's computing challenges.
Got into this because I was not in a corner.
Help run the institute, but also get to work on something I'm pretty passionate about which is how difficult it has become to access our data and perform the technical parts of analysis

(see later slides for more details)

IRIS-HEP

G. Watts (UW/Seattle)

Computational HEP
Traineeship Summer School

2023-07-26

First, and foremost, we are…

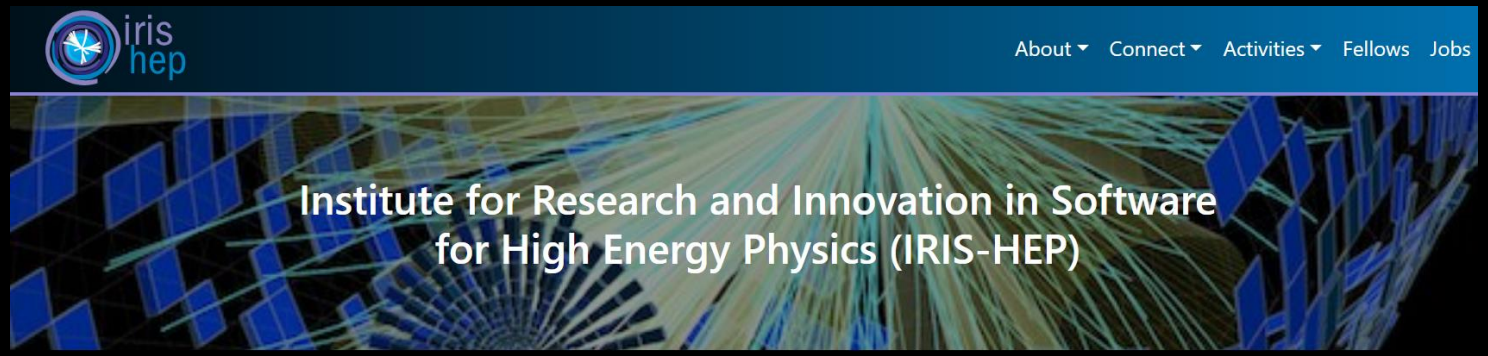
A group of universities with **PI's, Research Software Engineers, Post-docs, Graduate students, post-bac's**, and undergraduates that have joined to help the field address the HL-LHC's computing challenges.

Enable new approaches to computing and software that maximize, and potentially radically extend, the **physics reach** of the detectors.

Make improvements in software efficiency, scalability and performance and make use of the advances in CPU, storage and network technologies, that allow the experiments to **maximize their physics reach within their computing budgets**.

Significantly improve the long-term **sustainability** of the software through the lifetime of the HL-LHC.

Where Can You Find Us?



https://iris-hep.org

And what we do…

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 similated 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 mantain a regional cache that benefits all southern California CMS researchers.

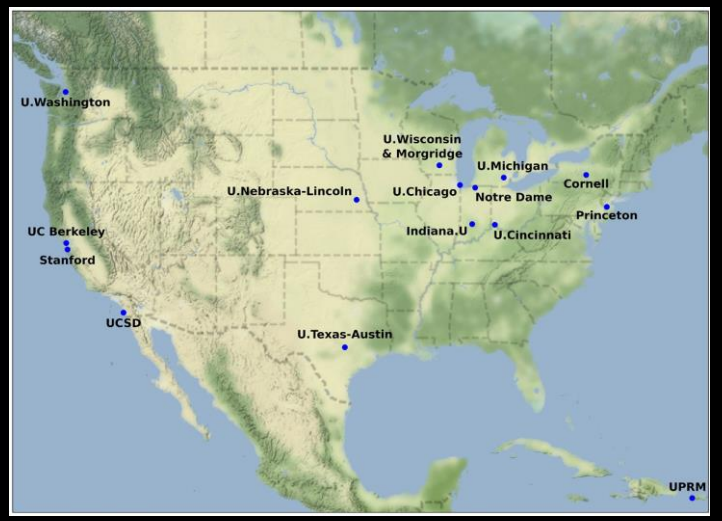
These in-production caches have shown to save up to a factor of three of WAN bandwidth compared with traditional data management techniques.

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 xcaches.", 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

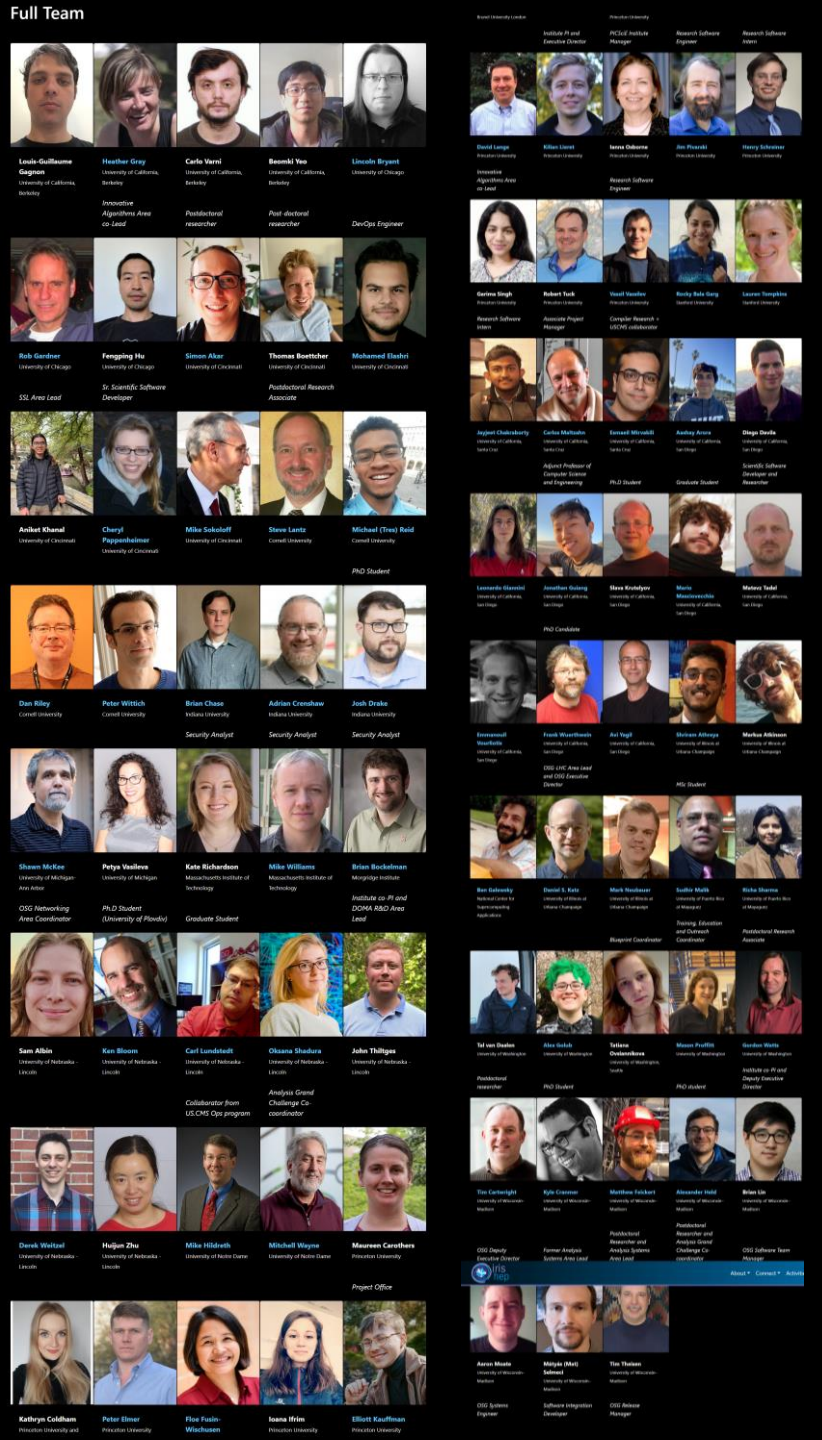
(often, but not always)

Who are we?



A number of the people you've seen this week are associated with IRIS-HEP in one way or the other:

- Peter Elmer
- Jim Pivarski
- Oksana Shadura
- Alex Held
- David Lange
- Matthew Feickert



Where did IRIS-HEP come from? A community process.

January 2017
UCSD

June 2017
Annecy

[Computing and Software for Big Science](#) **volume 3**, Article 7 (2019)

"*The result: a Programme of Work for the field as a whole, a multifaceted approach to addressing growing computing needs on the basis of existing or emerging hardware.*"

Eckhard Elsen (CERN Director of Research and Computing), editorial published with CWP/Roadmap

Many workshops, involving a diverse group of 100's of researchers to build consensus:

- Multiple domains: physics, computer science, data science
- International participants
- Full engagement from the Experiments and Labs (DOE, CERN)

Individual whitepapers on the arXiv:

Careers & Training, Conditions Data, DOMA, Data Analysis & Interpretation, Data and Software Preservation, Detector Simulation, Event/Data Processing Frameworks, Facilities and Distributed Computing, Machine Learning, Physics Generators, Security, Software Development, Deployment, Validation, Software Trigger and Event Reconstruction, Visualization

Community White Paper & the Strategic Plan

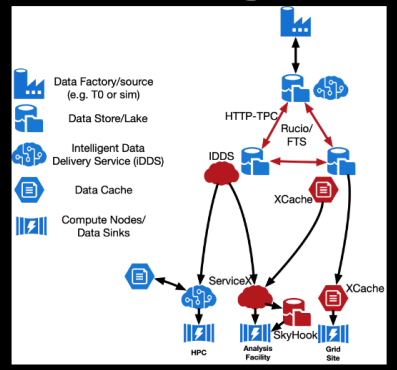
arXiv 1712.06982 arXiv 1712.06592

IRIS-HEP

ACI-1558216, ACI-1558219, and ACI-1558233

Areas of Research

Data Organization, Management and Access (DOMA)



An overview of the conceptual components of the data management ecosystem for HEP. The technologies that the DOMA team contributes to are highlighted in red.

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 and deliver data to processing endpoints; the DOMA area in IRIS-HEP is working on the R&D necessary to affect such change.

The bulk data transfer technologies were designed almost 15 years ago; the DOMA team is taking a fresh look at the transfer protocols and the authentication and authorization infrastructure used by the LHC community. This is resulting in a worldwide transition to new protocols and authorization approaches. The first phase of this work is done: the LHC community has successfully transitioned to HTTP as a foundation for bulk data transfers.

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 similated 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. Not only is the team researching new approaches for data delivery and implementing services but also working to integrate them together with into a coherent analysis facility, Coffea-Casa, for users.

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

Innovative Algorithms

:o perform the real-time processing in the trigger and the reconstruction of both real and simulated detector data are critical components of , uting challenge. 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. Projects in the IA area are focused in three areas:

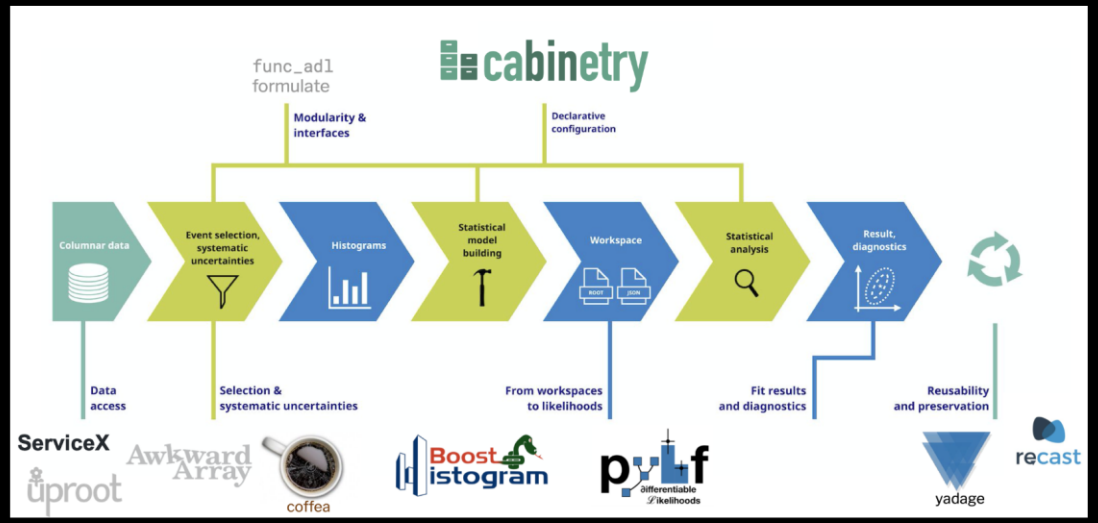
- Developing tracking algorithms for the HL-LHC
- Re-engineering algorithms for hardware accelerators
- Exploiting major advances in machine learning

Tracking for the HL-LHC is an area in particular need of novel approaches as it is expected to require a large fraction of the CPU budget. IA is working to develop more efficient and performant tracking algorithms. Hardware accelerators are expected to be the way forward to speed up reduce infrastructure cost. IA is exploring the use of hardware accelerators for tracking and the use of ML on accelerators in realistic HEP applications. In the area of machine learning, IA aims to capitalize on industry and data science techniques and tools. In particular, we are investigating new HEP applications of ML and to apply new ML techniques to HEP.

Areas of Research

Analysis Systems

The goal of the Analysis Systems focus area is to develop sustainable analysis tools to extend the physics reach of the HL-LHC experiments by creating greater functionality, reducing time-to-insight, lowering the barriers for smaller teams, and streamlining analysis preservation, reproducibility, and reuse.



Focus Area Strategies:

- Establish declarative specifications for analysis tasks and workflows that will enable the technical development of analysis systems to be decoupled from the user- facing semantics of physics analysis.
- Leverage and align with developments from industry and the broader scientific software community to enhance sustainability of the analysis systems.
- Develop high-throughput, low-latency systems for analysis for HEP.

Training, Education and Outreach

Quick links

- Upcoming and recent events
- Training Modules and videos
- HSF Training Events

Introduction

The long-term sustainability of the research software ecosystem is particularly important for HEP, given that the HL-LHC and other facilities of the 2020s will be relevant through at least the 2030s. The IRIS-HEP Software Sustainability Core (SSC) focuses on activities which promote the long-term sustainability of the software. The SSC has primarily focused on **Training, Education and Outreach**.

To implement this vision, IRIS-HEP has developed a program to provide software training paths from a researcher's first steps through more regular, active contribution along with its partners: HSF, FIRST-HEP and the Carpentries. In particular, it has developed an introductory HEP software curriculum and several software modules on techniques and methods for computing and data science. The training program has enabled HEP users to jump start their research and contribution to the field. The common efforts on training across HEP has helped build a strong sense of community.

Our efforts have established a platform from where we can scale and sustain our training efforts. Our training vision is reflected in the pyramid below.