

# DOMA – Area Lead Chat



**MORGRIDGE**  
INSTITUTE FOR RESEARCH  
CORE COMPUTATION

**FEARLESS SCIENCE**

## Approach

- Trying to address the questions / requests directly.
  - I.e., one (or more) slides partitioned by the question text.
- The slides are formatted more like discussion reminders / generators. They assume the reader has passing familiarity with DOMA's work.
  - It's not fully cross-referenced and with pretty pictures as one would expect for a public talk.



# 1. Summarize the set of projects/activities and associated effort for your area

Summarize the set of projects/activities and associated effort for your area.

- **IDDS** Intelligent Data Delivery Service: works to deliver transformed events - or notifications of event / file arrival – to a processing framework.
  - **Effort:** Wen Guan @ UW-Madison. Sum: 15%
- **ServiceX:** Column delivery service. Deliver transformed columns from.
  - **Effort:** Marc Weinberg @ UChicago, Ben Galewsky @ UIUC. 27%
- **HTTP-TPC:** .
  - **Effort:** Diego Davila @ UCSD, Brian Bockelman @ Morgridge. 10%
- **XCache:**
  - **Effort:** Diego Davila, Edgar Fajardo, Igor Sfiligoi @ UCSD. 12%
- **SkyHook:**
  - **Effort:** Jeff LeFevre, Carlos Maltzahn, Ivo Jimenez, Aaron Chu @ UCSC. 30%
- **Hardware Acceleration:** Grad student at UChicago. ???
- **Modeling:**
  - **Effort:** Diego Davila, Frank Würthwein @ UCSD. 5%
- **Broader Impact / Training / Education / Outreach:** None.

**Note:** effort levels are estimates - percentage of DOMA's total effort; meant to reflect relative sizes. Does not add up to 100%; effort & throughput not fungible (e.g., 1 hour of grad student time is not the same as 1 hour of staff time).



## 2. Are there internal or external collaborations associated with each project or activity?

Are there internal or external collaborations associated with each project or activity? For external collaborations, is IRIS-HEP leading, contributing or simply “connecting/liaising”?

- **WLCG DOMA Working Groups:** (Co-)Leading the ACCESS and TPC areas.
  - Also serves to liaise with the Ops programs.
- **SkyHook** has external connections through CROSS, led by Maltzahn.
  - Working to build connections to Coffea.
- **ServiceX** has effort in AS, connects with other AS projects, and heavily utilizes SSL for dev & testing. Have connection with the U.S. ATLAS Ops program via a HL-LHC R&D postdoc.
  - Starting to work with the TreeMaker (Pedro et al) effort @ FMAL.
- **HTTP-TPC** and **XCache** touch a number of technical communities: XRootD, WLCG DOMA, storage system developers, wider OSG efforts (e.g., LIGO).



### 3. Which project/activities/goals are making progress and which are not?

Which project/activities/goals are making progress and which are not? (Area lead's opinion) For those that are not, what is impeding progress?

Given this is internal-looking, **I focus on the negative and not the positive.**

- HTTP-TPC is finally starting to see (halting) use in production – but is behind schedule. Goal was for a single site to hit 30% by 1 March. Currently at 2-10%, depending on the week and mix of transfers. Not particularly concerned.
- XCache is in use by ATLAS, CMS, and OSG. Next challenge is going to be usage outside current leading-edge sites. Example: SoCal cache is used heavily at Caltech & UCSD but model has not spread to other sites.
  - Effort to validate monitoring has largely failed to take off.
  - Biggest need: plan for the next few deployment steps. Technical pieces are largely in place.

### 3. Which project/activities/goals are making progress and which are not?

(Given this is internal-looking, I focus on the negative and not the positive.)

- **IDDs:** Starting to deliver value / processing events for ATLAS.
  - Not made a lot of traction outside ATLAS.
  - Starting on a hyperparameter optimization workflow (work mostly done by non-IRIS-HEP collaborator) – but need to figure a strategy to align this with DOMA.
- **ServiceX:** Nascent community is starting to form – critical to start pulling in external people over the next 3-6 months.
- **SkyHook:** Still very early days. Over the last month, been working with more a concrete use case and analyst (Pedro).
- **Modeling:** As contributions have been made to various planning documents (both internal and WLCG), a lot of these activities are wrapping up.

## 4. How are each of these projects/activities connected to ... experiments?

How are each of these projects/activities connected to, being informed by or planning on delivering (eventually) to the experiments? Are there relevant blueprint meetings or workshops that should happen to make progress?

- ServiceX is developing connections to the experiments – the U.S. ATLAS side is more with university groups while U.S. CMS side is with the FNAL / Coffea group.
- SkyHook is not very visible at the experiment level. Need to find forums to showcase the technology (and work on a few demos).
- **Needed blueprint:** Analysis Facilities for HL-LHC. I believe each Ops program has had various smaller organizing discussions – now's likely the time to bring things together.
- **Needed blueprint:** Future plans for XCache usage (a long-form discussion is actually happening next week – but we probably need to open it up more).
- **Needed blueprint:** Long-term and transition plans for TPC protocols. Likely needs to be 'led' by WLCG but could have IRIS-HEP as a catalyst.



## 5. What would be potential Year 3 milestones for each of the projects?

What would be potential Year 3 milestones for each of the projects? (First ideas, to be iterated with PIs and the whole team as this process moves forward.)

- Haven't prepared exact milestones: the themes should be:
  - Contributing to 'challenges' (see next slide). Note this may require extra or redirected effort – an integrated challenge should take real effort.
  - Integrating across projects internally (particularly, dataflow of XCache -> ServiceX -> SkyHook). This has already started in Year 2.
  - Deploying functionality to production and building a user community.
    - For new services, this might mean O(dozen) users by Year 4.
    - For XCache or HTTP-TPC, this may mean replacing the prior generation of technology by Year 4.



## 6. What “grand challenges” would be useful to organize ...?

What “grand challenges” would be useful to organize involving your area during Year 3 of IRIS-HEP? How would these challenges depend on efforts from other areas of IRIS-HEP, the US LHC Ops programs or the experiments?

I'd suggest two challenges:

- [Data Processing Challenge](#) (from FKW): Perform a 10PB data reduction campaign on a resource ‘external’ to WLCG. Use k8s to setup necessary services on-demand, transfer data from WLCG sources, and do reduction processing in  $O(1)$  day).
- Analysis Facility Challenge: Demonstrate the ability to process 1PB of input data (note: not necessarily *move* 1PB) into a physicist-focused analysis using tools from AS.
  - Less ‘grand’: repeat the ‘Higgs discovery’ demo done for KubeCon using IRIS-HEP tools and services.

## 7. Are there new opportunities where effort from IRIS-HEP can make an impact?

Are there new opportunities where effort from IRIS-HEP can make an impact? Is the alignment of the focus areas in IRIS-HEP appropriate?

- I think there's a huge potential in the Coffea effort – and the biggest missing piece for them is data management (a-la SkyHook and/or ServiceX).
- We likely don't need new projects here but rather need help doing 'translational' work (non-developer one-on-one with prototype users, service operations).

## 8. How are projects currently managed in your area?

How are projects currently managed in your area? What tools are being used? How is progress measured? How are risks recorded, identified and mitigated?

- Projects are independently managed – DOMA acts like a coordination area, setting goals and trying to grow inter-project relationships.
  - Would welcome help in putting together quantitative measures.
  - Measures of progress have been difficult for R&D work. ServiceX, for example, is nearing the 1.0 release. I expect a phase shift for the execution phase.
- The PEP has served as the ultimate reference for metrics and deliverables. Wouldn't mind a tool for tracking both the NSF-level items and internal ones.



## 9. Are the metrics being used to measure success clearly defined?

Are the metrics being used to measure success clearly defined? How well do metrics in your area measure progress, success or impact? Where can the metrics be improved or refined to better measure progress, success or impact?

- Fairly happy with the metrics we have for the things they cover well (XCache, HTTP-TPC). The deployment metric for IDDS/ServiceX is less useful.
- Biggest change I'd like to make is increase coverage to services that didn't exist 2 years ago.



**MORGRIDGE**  
INSTITUTE FOR RESEARCH  
CORE COMPUTATION

[morgridge.org](http://morgridge.org)

**FEARLESS SCIENCE**