

# DOMA Steering Board #12



**FEARLESS SCIENCE**



No changes since last update –  
June 2021 (6 months ago)

## The DOMA Team

Wisconsin



UNL



UIUC



U Chicago



Morgridge



UCSD



Jeff Lefevre currently on leave.

UCSC



The DOMA team is a distributed team working across UIUC, Morgridge, U Chicago, UCSD, UCSC, UNL, UW.



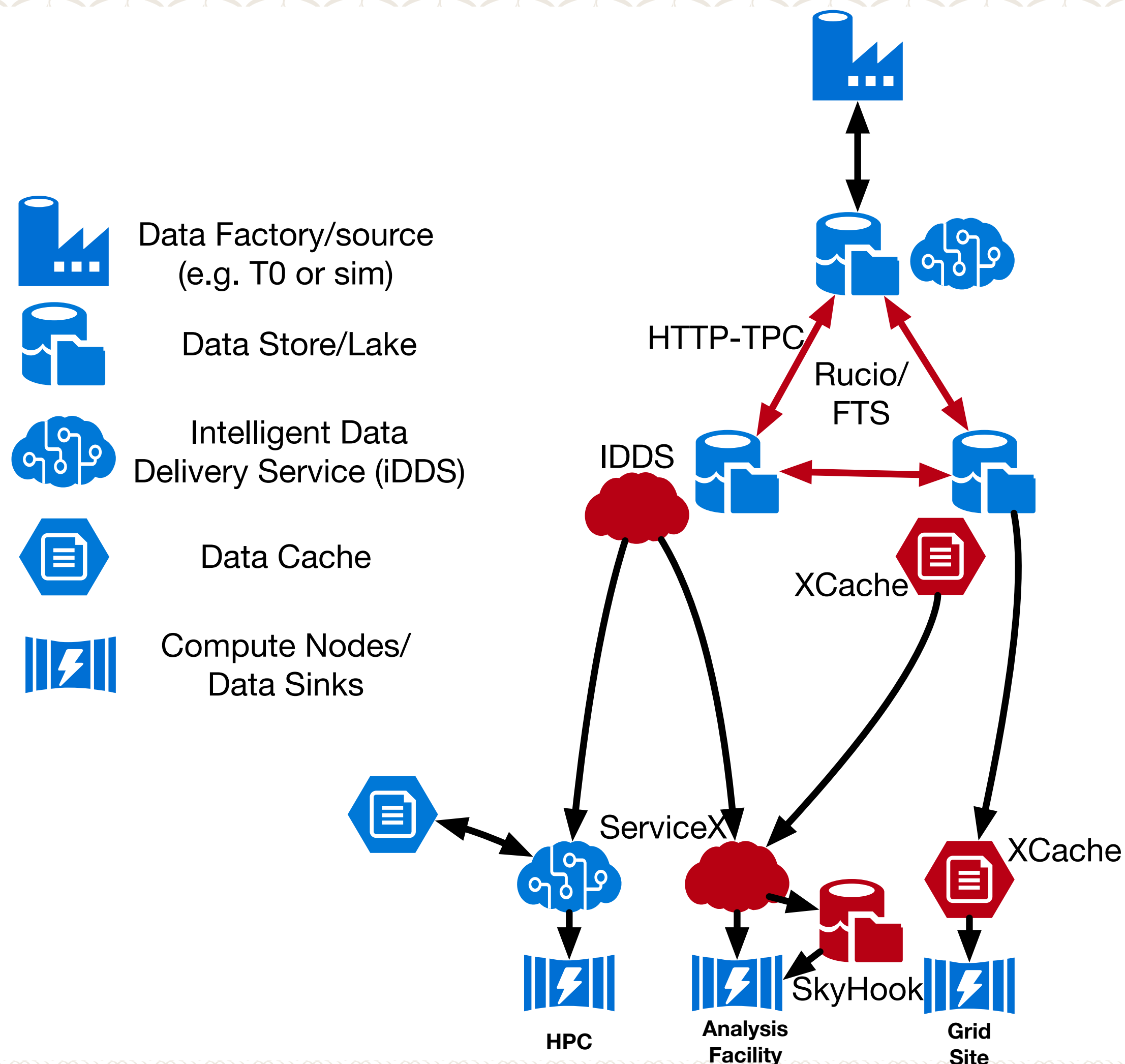
## The DOMA Challenges

What are the HL-LHC challenges the DOMA area is trying to tackle?

- **Challenge 1:** Data delivery in the exabyte era.
- **Challenge 2:** Modernization of the LHC's bulk data movement infrastructure.
- **Challenge 3:** Data delivery for new analysis systems and techniques during HL-LHC.

In addition to the technical work, we lead community working groups (WLCG DOMA TPC) and workshops (Analysis Systems for the HL-LHC; whitepaper in draft form) relevant to these challenges.

## Our view of the world



DOMA has fundamental impact on how we process and move data for the HL-LHC.

- Red indicates areas where IRIS-HEP is actively working
- You can cleave the contributions in two themes:
  - Production work:
    - **C1:** IDDS, XCache.
    - **C2:** HTTP-TPC
  - Analysis facility work (**C3**): XCache, ServiceX, SkyHook.

## Activities in DOMA

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.
- **HTTP-TPC:** Moving data between storage services over HTTPS.
  - **Includes participation in the WLCG data challenges.**
  - **Effort:** Diego Davila @ UCSD, Brian Bockelman @ Morgridge.
- **XCache:** Delivering file data through the use of large caches.
  - **Effort:** Diego Davila, Igor Sfiligoi @ UCSD.
- **ServiceX:** Column delivery service. Transform experiment data and deliver it as columns.
  - **Effort:** Suchandra Thapa @ UChicago, Ben Galewsky @ UIUC.
- **SkyHook:**
  - **Effort:** Esmail Mirvakili, Jeff LeFevre, Carlos Maltzahn, Ivo Jimenez, Aaron Chu @ UCSC.
- **Facilities:** Combing new services along with tools from AS and deployment techniques from SSL
  - **Effort:** Oksana Shadura @ Nebraska.

**Deliver data to  
production systems**

**Deliver data to  
analysis**

**New integration activity; aligned closely  
with AS and subject of other talk.**

# **DOMA & Analysis**





## Analysis Impact

The DOMA area works to deliver data in new ways to analysis services. A few projects I'll highlight:

- **ServiceX:** Deriving analysis-specific datasets from experimental data and delivering it them to analysts.
- **SkyHook DM:** Enabling computational storage for HEP data; using the Arrow Dataset API to push down queries to Ceph object storage.
- **Coffea-Casa:** A facility for integrating new services into a coherent vision for analysis. Currently provides the ability to process experiment data through a Dask interface, using either a browser (via JupyterHub) or command-line.

Two Coffea-Casa instances were used for the recent AGC workshop.

# ServiceX

ServiceX is a Kubernetes-based data delivery service.

Highlights of Year 3:

- Additional file types: CMS NanoAOD, CMS OpenData AOD. Complements existing support for ATLAS xAOD.
- Improved frontends for physicists (func\_adl).
- Additional security models (for Nebraska integration).

**In progress / TODO:**

- Continuous integration tests; ensure we don't regress functionality.
- Improve integration with analysis facilities to make it more approachable to users.
- Improve monitoring / logging (visibility of activities to user).

**Year 4 goal: 5 analysis groups utilizing the service.**

- **Progress:** one new user group after the AGC workshop.

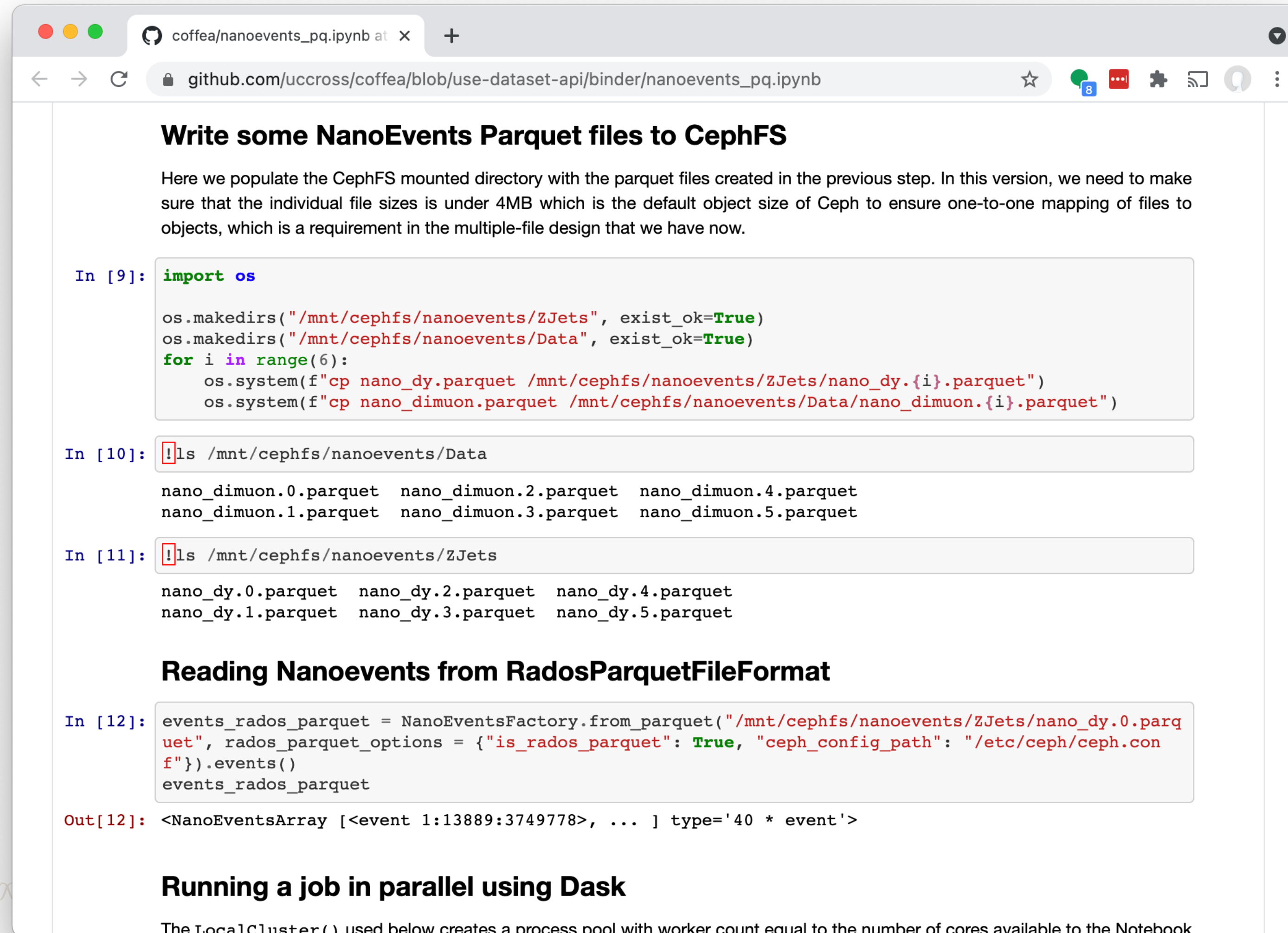
**'Spin-off' CSSI awarded for**  
**collaborating with a Dark Matter**  
**experiment on ServiceX**



# SkyHook DM

SkyHook provides a mechanism to access data kept in CephFS through the popular Arrow libraries.

- Allows filters & projections to push down to the storage hosts.
- Significant reduction in client CPU and network costs.
- Last year, hit milestone to be able to read/write in Coffea.
- Coffea itself knows nothing about SkyHook: it's a generic Arrow backend to Coffea. The Arrow libraries are configured to work with SkyHook.



The screenshot shows a Jupyter Notebook titled "coffea/nanoevents\_pq.ipynb" at a Binder URL. The notebook content includes:

### Write some NanoEvents Parquet files to CephFS

Here we populate the CephFS mounted directory with the parquet files created in the previous step. In this version, we need to make sure that the individual file sizes is under 4MB which is the default object size of Ceph to ensure one-to-one mapping of files to objects, which is a requirement in the multiple-file design that we have now.

```
In [9]: import os

os.makedirs("/mnt/cephfs/nanoevents/ZJets", exist_ok=True)
os.makedirs("/mnt/cephfs/nanoevents/Data", exist_ok=True)
for i in range(6):
    os.system(f"cp nano_dy.parquet /mnt/cephfs/nanoevents/ZJets/nano_dy.{i}.parquet")
    os.system(f"cp nano_dimuon.parquet /mnt/cephfs/nanoevents/Data/nano_dimuon.{i}.parquet")
```

```
In [10]: !ls /mnt/cephfs/nanoevents/Data

nano_dimuon.0.parquet  nano_dimuon.2.parquet  nano_dimuon.4.parquet
nano_dimuon.1.parquet  nano_dimuon.3.parquet  nano_dimuon.5.parquet
```

```
In [11]: !ls /mnt/cephfs/nanoevents/ZJets

nano_dy.0.parquet  nano_dy.2.parquet  nano_dy.4.parquet
nano_dy.1.parquet  nano_dy.3.parquet  nano_dy.5.parquet
```

### Reading Nanoevents from RadosParquetFileFormat

```
In [12]: events_rados_parquet = NanoEventsFactory.from_parquet("/mnt/cephfs/nanoevents/ZJets/nano_dy.0.parquet", rados_parquet_options = {"is_rados_parquet": True, "ceph_config_path": "/etc/ceph/ceph.conf"})
events_rados_parquet
```

```
Out[12]: <NanoEventsArray [<event 1:13889:3749778>, ... ] type='40 * event'>
```

### Running a job in parallel using Dask

The `LocalCluster()` used below creates a process pool with worker count equal to the number of cores available to the Notebook

## SkyHook – Recent Progress

SkyHook code was recently merged into Apache Arrow – will be in the 7.0.0 release.

- With this, a much larger community can benefit from push-down from the Arrow Dataset API into the Ceph RADOS layer. Benefits beyond HEP; much of the original work done by prior IRIS-HEP fellow (now PhD student at UCSC).

Year 4 goals:

- Virtual join of multiple derived datasets. Progress:
  - Investigated writing through Databrick's DeltaLake. Results: meets many of the requirements but all scheduling has to go through Spark.
  - Current approach: working on the “join” primitive at the Arrow level (a-la ROOT “Friend” TTrees). Still looking
- **Production instance(s)**. Progress: requirements baked into hardware refresh of the Coffea-Casa instance at Nebraska. Should be on-target for this winter.



# Putting it all together: “Coffea Casa”

Effort led by Oksana  
Shadura



The Coffea-Casa effort is an attempt to integrate the DOMA-developed technologies with tools and techniques from Analysis Systems – all deployed with techniques promoted by SSL area.

- Meant to be a proving ground where R&D meets real users.
- IRIS-HEP has invested \$150k of hardware at both Chicago and Nebraska to ensure there’s enough capacity to attract real use.
  - Due to hardware procurement delays, these are still being installed.
- Three instances: Two at Nebraska (one can access CMS data, the other CMS OpenData), one at Chicago (ATLAS data).
- **Usage:**
  - 100 users in the last 6 months.
  - 24 users in the last month
  - 14 in the last 24 hours.
- Recently completed a strategy meeting to set out 12 months of detailed goals for improved functionality.

## Year 4 goals

- **Running in production with SkyHook and ServiceX integrated.**
- **Five involved analysis groups.**
- **Two additional instances outside Nebraska.**


**Progress: one at Chicago; in discussion with BNL.**

# **DOMA & Production**

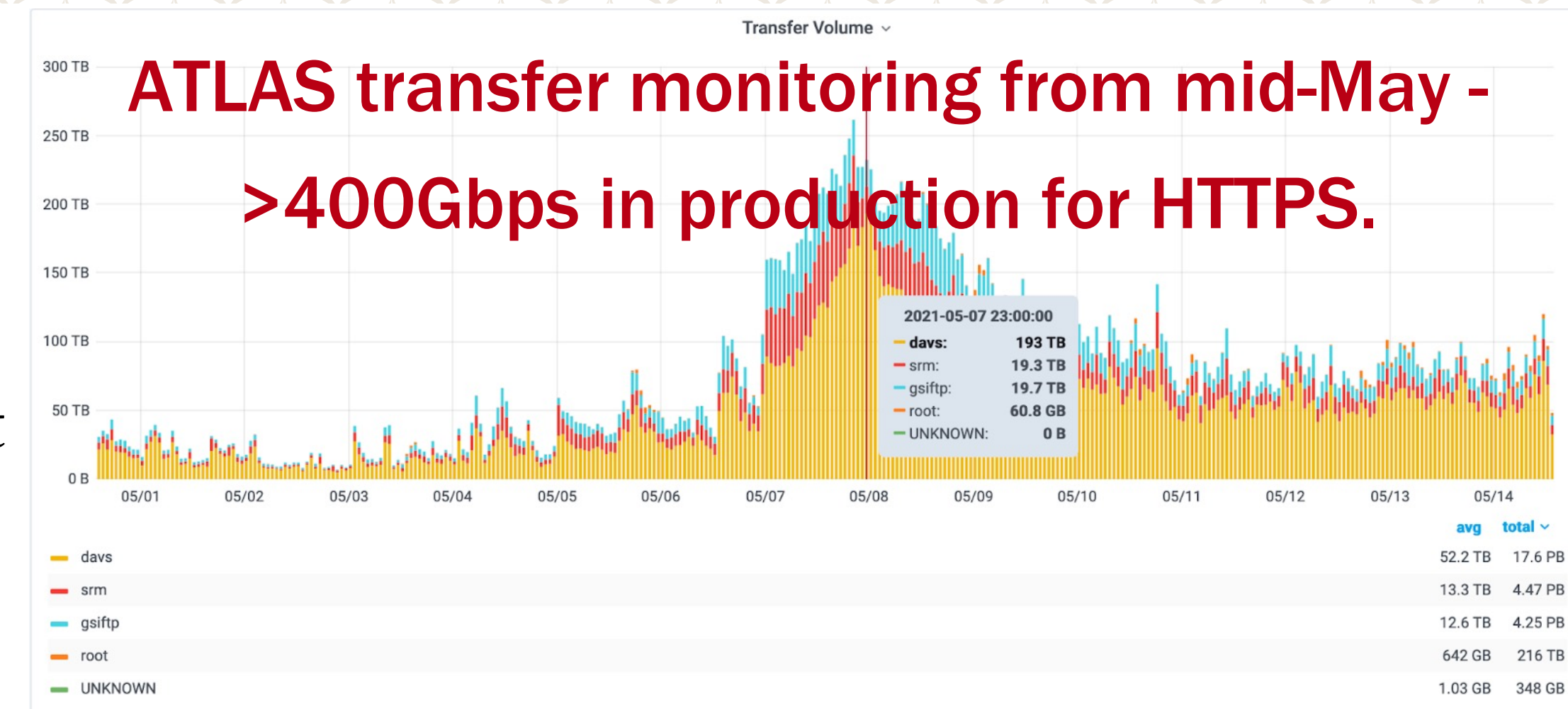




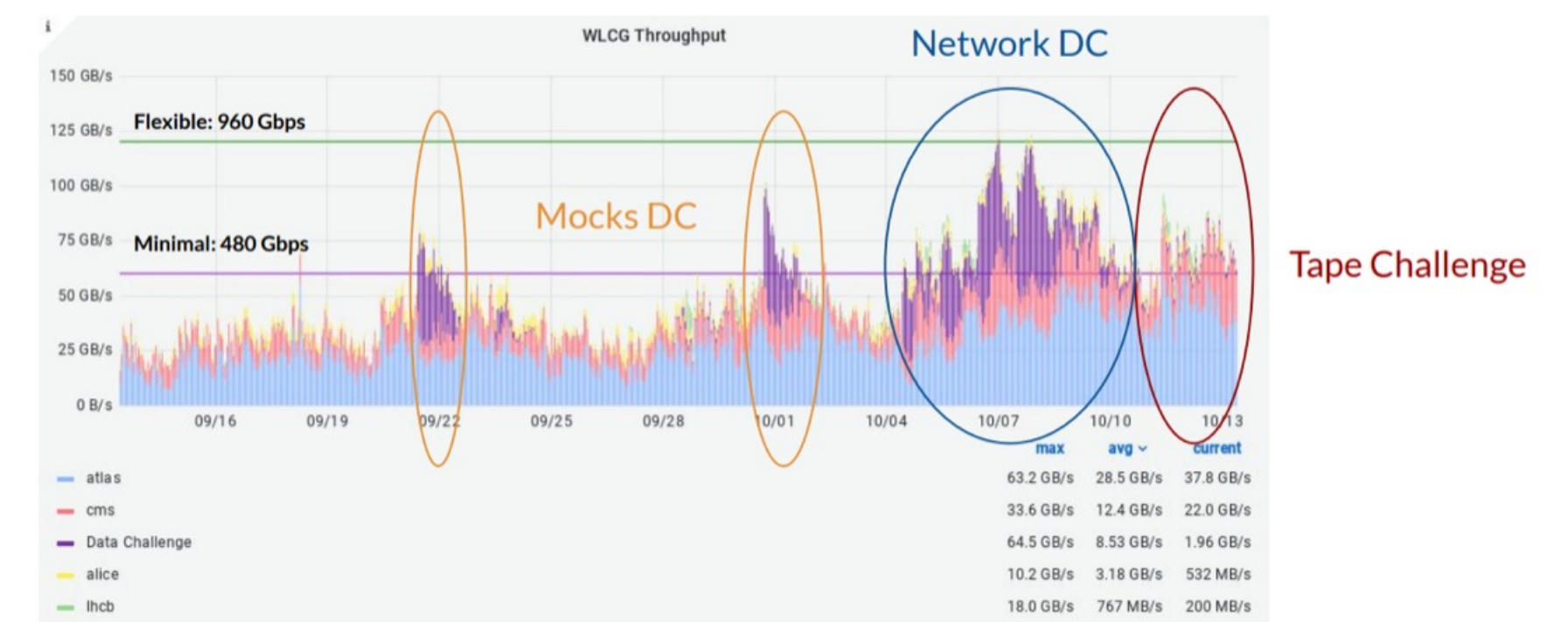
# Modernizing Bulk Data Transfer: Deploying HTTP-TPC for WLCG

The LHC is switching its bulk data movement to modern protocols  and authorization infrastructure; this is co-led by IRIS-HEP (but involves a worldwide set of people).

- In 2020, we moved from ‘development’ to ‘early production use’. In 2021, moved to ‘full production’.
- In October 2021, WLCG led a network data challenge (see plot) where IRIS participated.
  - Activities of the HTTP-TPC transition wrapping up.
- Current testing is around using SRM + HTTP-TPC.



## WLCG (FTS+XRootD) Throughput - 30-Day Plot



## Data Challenge

In 2021, the scale goal was modest: 10% of HL-LHC scale.

- A reusable infrastructure for monitoring transfers and generating additional load was created.

Open question for the community:

- What is needed for 2022?



## Modernizing Bulk Data Transfer Next Challenge: Authorization Changes

In Year 4, we are focusing on the authorization changes (tokens) for the bulk data transfer infrastructure. For WLCG, key challenges are:

- Ensure all storage implementations can be configured to accept token-based authorization.
  - We are contributing to the scitokens-cpp & xrootd-scitokens libraries.
- Ensure FTS has a consistent way of using (and renewing, as needed) access tokens.  
Ensure Rucio interacts with FTS in the same way
  - WLCG authz working group is forming a sub-group focused on defining the token flows.

Activities in 2022 in this area will look a lot like the early days of HTTP-TPC in Year 1 of IRIS-HEP.

**Goal: By end of Year 4, we move one byte in  
production using these workflows**

# IDDS

The Intelligent Data Delivery Service (IDDS) manages input data and provides data locality for job management systems such as PanDA. Initial use case was the ATLAS “Data Carousel”, managing the processing of data as it is staged from archival system.

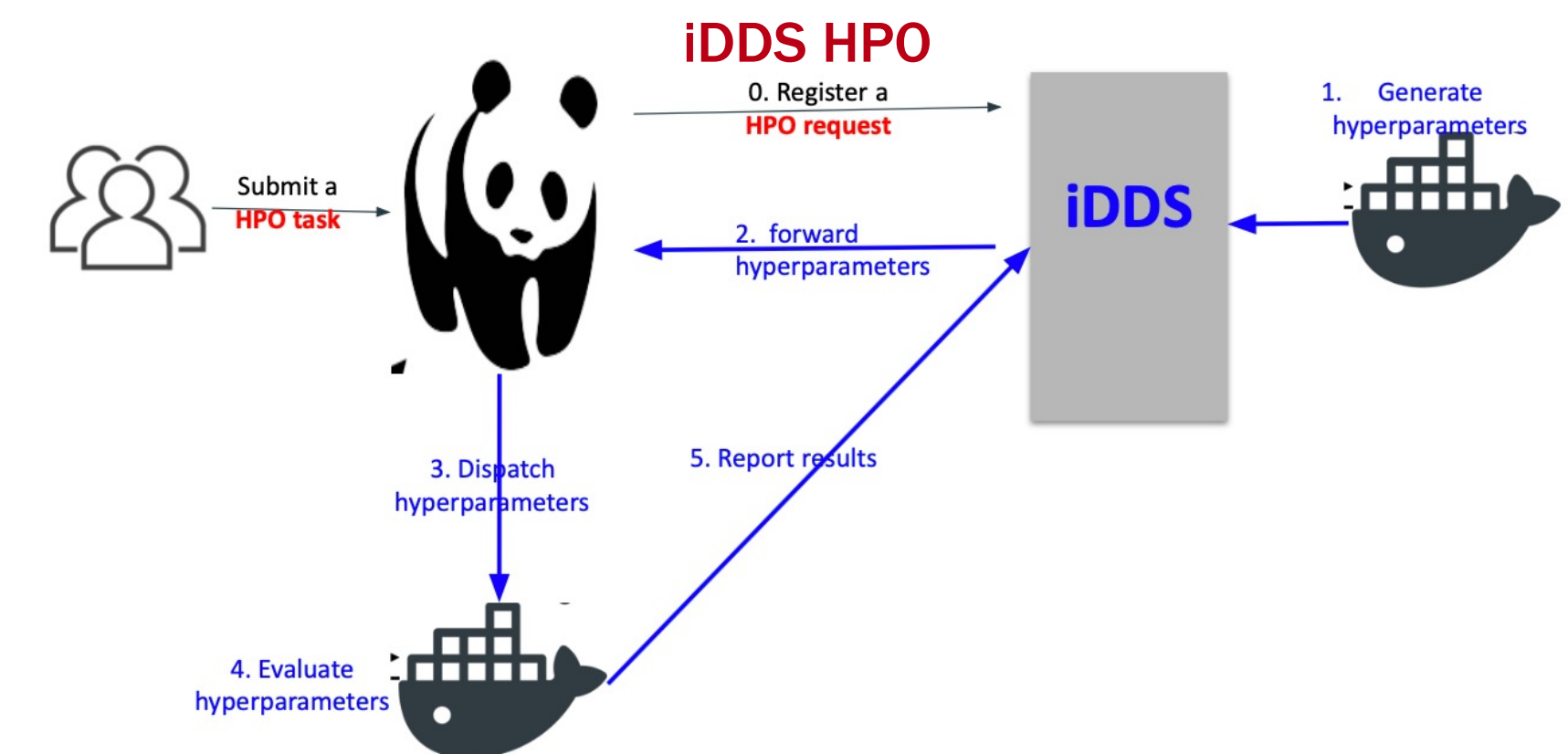
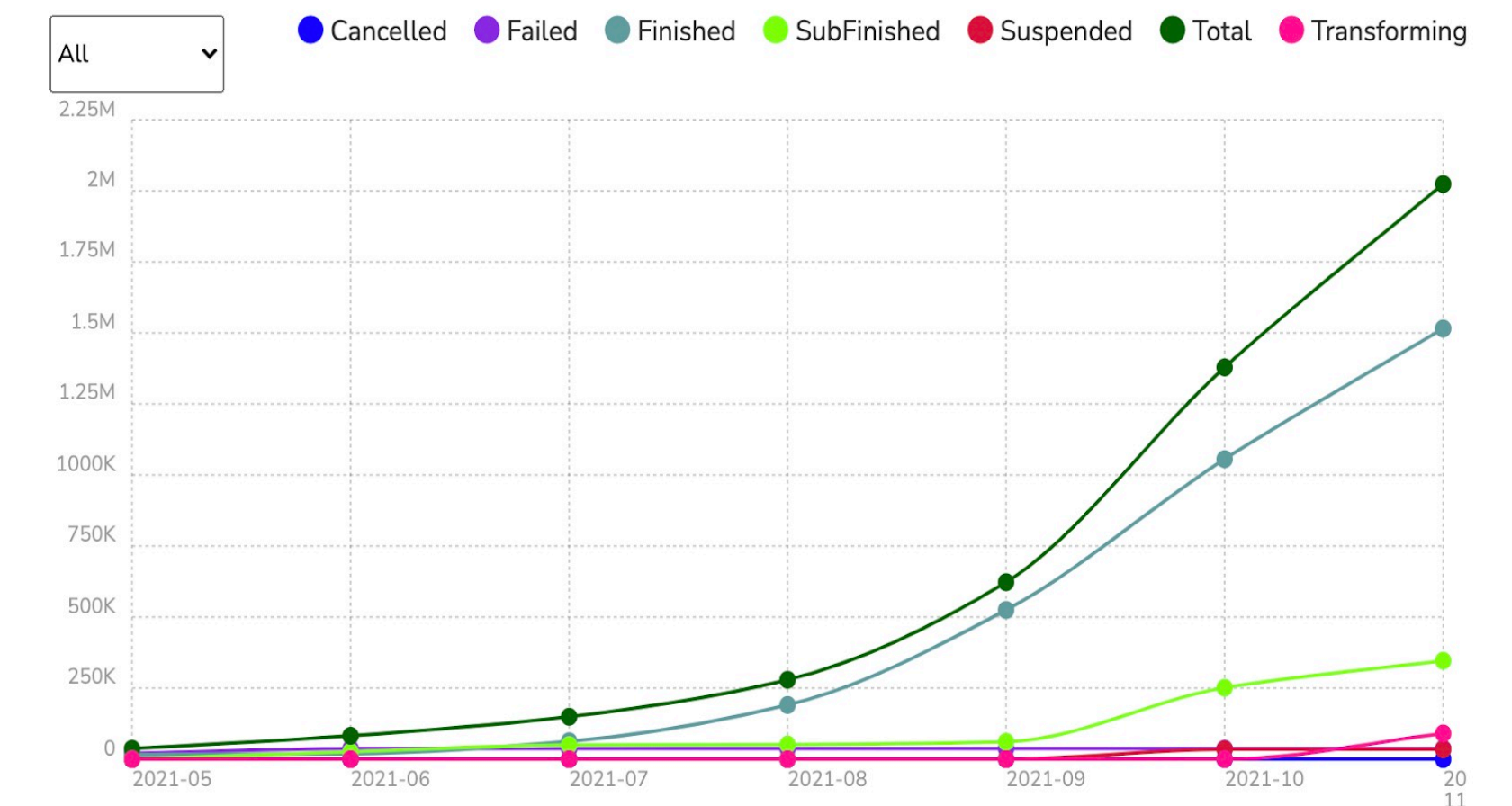
Next use case was to manage machine learning training parameters for Hyper Parameter Optimization.

- Demonstrated a fully-automated platform for hyper-parameter optimization on top of geographically distributed GPU resources on the grid, HPC, and clouds.
- Active ATLAS users: FastCaloGAN and ToyMC workflows.
  - Current status: Working to improve overall latency, especially reducing latency on short-running tasks.

Impact beyond HEP:

- New: Vera Rubin Observatory (LSST) has decided to adopt IDDS for its data processing needs.
- IDDS still is driven by HL-LHC use cases – but happy to see additional impact.

Number Of Files By Transform Status





# XCACHE

The XCache software provides an authenticated caching proxy which can reduce the WAN usage, hide latency, and reduce number of on-disk copies for LHC sites.

Recent activities (**shared work with OSG-LHC**):

- Expand the usage and monitoring of the XCache service for cache-based data delivery.
  - Additional operational deployments “on the edge” using Kubernetes.
- **Overhauling the monitoring infrastructure.** Developing a new XRootD component, the “shoveler”, which moves the connection from XCache host to central monitoring from UDP to TCP. Also adds token-based authorization for the monitoring.
- **Deploy token-based authentication / authorization.** While actual LHC use of tokens on the worker node is likely not until Year 5, we are ensuring the services are ready.
  - Hopefully can roll out token authentication on Coffea-Casa during Year 4 (stretch goal! Also would be useful for EOS access..)

# Conclusions





## Conclusions

Largely unchanged from last update! We make targeted contributions across the vertical stack:

- We are building systems to deliver data to new approaches in analysis, resulting in efforts like Coffea-Casa to build an integrated analysis facility. (Challenge 3)
- We co-lead the modernization of the bulk transfer system (Challenge 2), which is the heart of the network utilization necessary for HL-LHC.
- DOMA delivers data to processing sinks with services like IDDS and XCache, preparing for the exabyte era.

While an R&D area, we leverage the integration capabilities of the SSL and work closely with the OSG-LHC to deploy & iterate on our software and services.

Highlights of the last 6 months: finishing up this phase of the **HTTP-TPC transition** (and **WLCG data challenge**) and continued ‘real world’ usage growth of Coffea-Casa.



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

This material is based upon work supported by the National Science Foundation under Grant No. 1836650. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

**FEARLESS SCIENCE**