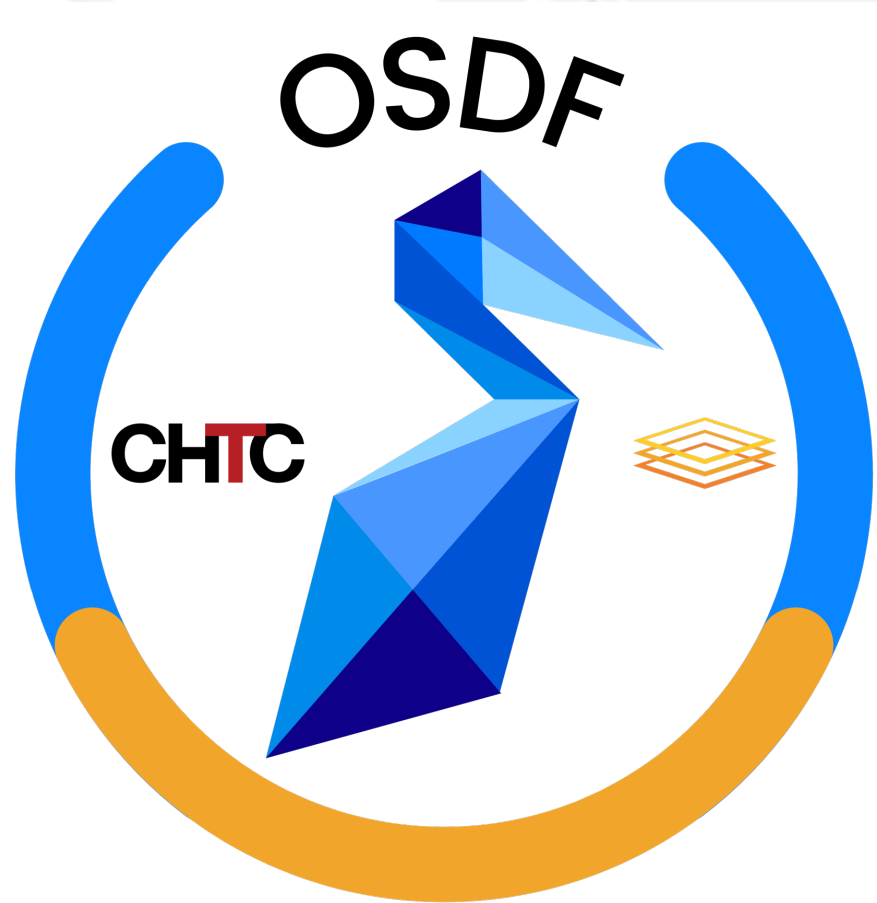




Dear CICI PI:  
Let's Be Partners



Brian Bockelman

 **MORGRIDGE**  
INSTITUTE FOR RESEARCH

# CHTC

Center For High  
**Throughput**  
Computing



**MORGRIDGE**  
INSTITUTE FOR RESEARCH



**School of Computer, Data  
& Information Sciences**  
UNIVERSITY OF WISCONSIN-MADISON





# Have you ever been frustrated by (scientific) data?

It is real work to curate a scientific dataset, including metadata and grow community interest in it!

But, as a PI, your work is not yet done:

- What tools do you provide to use the data?
- How do you manage access control?
- How do you connect your community to computation?
- How do you provide scalable access to the data?
  - S3 scales: but S3 egress fees can kill!
  - ... and S3 “user pays” is not loved!



# Option 1: Dear User, it's Your Problem

You *can* shift the problem to the user community:

- Put the data on your university web server / S3, create a website, tell your friends about it.
- Not my problem to provide you with compute!
- You figure out how to download the data.
  - You write the scripts to have `curl` scrape the website.
  - And figure out where to copy the data. And get the updates.
  - And figure out where to compute.

**Why not?** Simple: higher friction to use => smaller impact.





## Option 2: Dear User, We'll Do It!

- Scale out the data repository.
- Build tools to distribute the data.
- Build a platform (gateway?) for computing on the dataset.
- Build your authentication and authorization solution.
- Build build build ...

**Problem:** At some point you have to do the science!

Curating data is hard enough:  
Isn't there a better way?



## Option 3: Leverage the NSF ecosystem

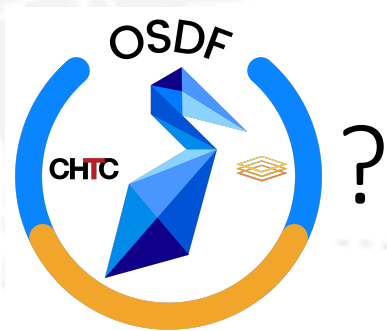
You understand your dataset better than anyone. However,

- Some of these problems are common across domains!
- NSF funds services that help stream your data (OSDF).
- NSF funds services to catalog & publish datasets (NDP).
- NSF funds computing platforms to explore & analyze data (OSPool, NRP).
- NSF funds service platforms for providing unique capabilities specific to your dataset (NDP, NRP).

**You can build the community & provide value – without the DIY!**



What is the



... By Analogy...



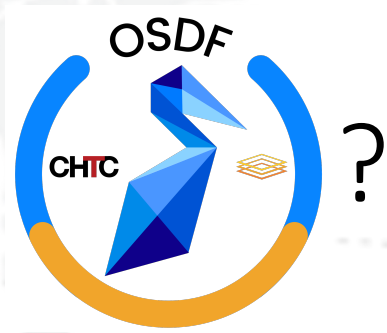
“Netflix for Science”

Allowing computational workloads to stream data





What is the



... By Analogy...

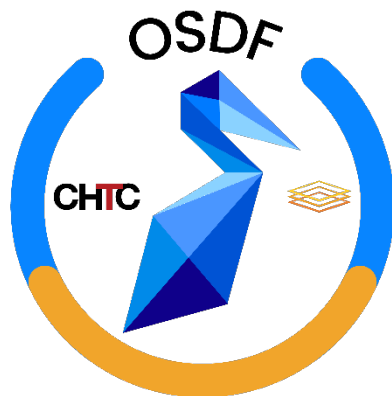


“Cloudflare for Science”

Infrastructure for scaling access to your storage



# Serving an (Inter)national Community





# OSDF In Detail

- The OSDF is a global service for data delivery



- Operated by the PATH project
- Using software from the Pelican Project
- Leveraging hardware from the NRP, Internet2, ESNet, and more!







# Connecting to the OSDF

- The place where your data lives is the “object store”
  - This can be a POSIX filesystem, S3-compatible, HTTP webserver, or Globus collection.
  - Exports immutable objects in your dataset.
- We – or you! - operate the **origin service**, a container that connects the object store to the OSDF.
  - The origin implements your access control policies and proxies .
- The rest of the OSDF scales via reuse and manages access.

This is  
where your  
data lives



**Object Store**



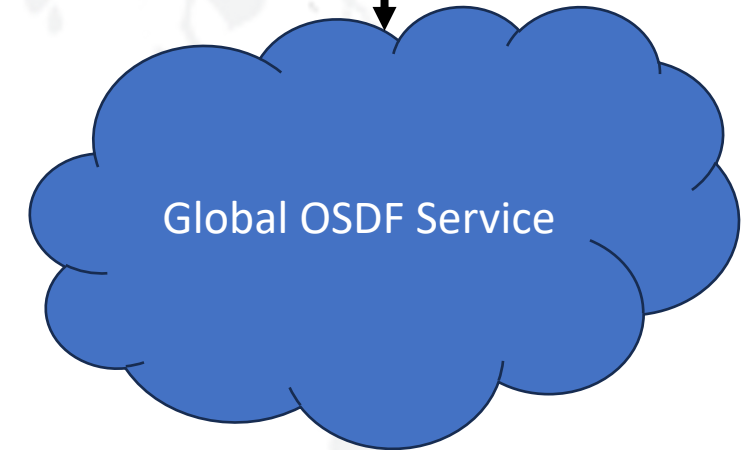
Implements  
your access  
policies!

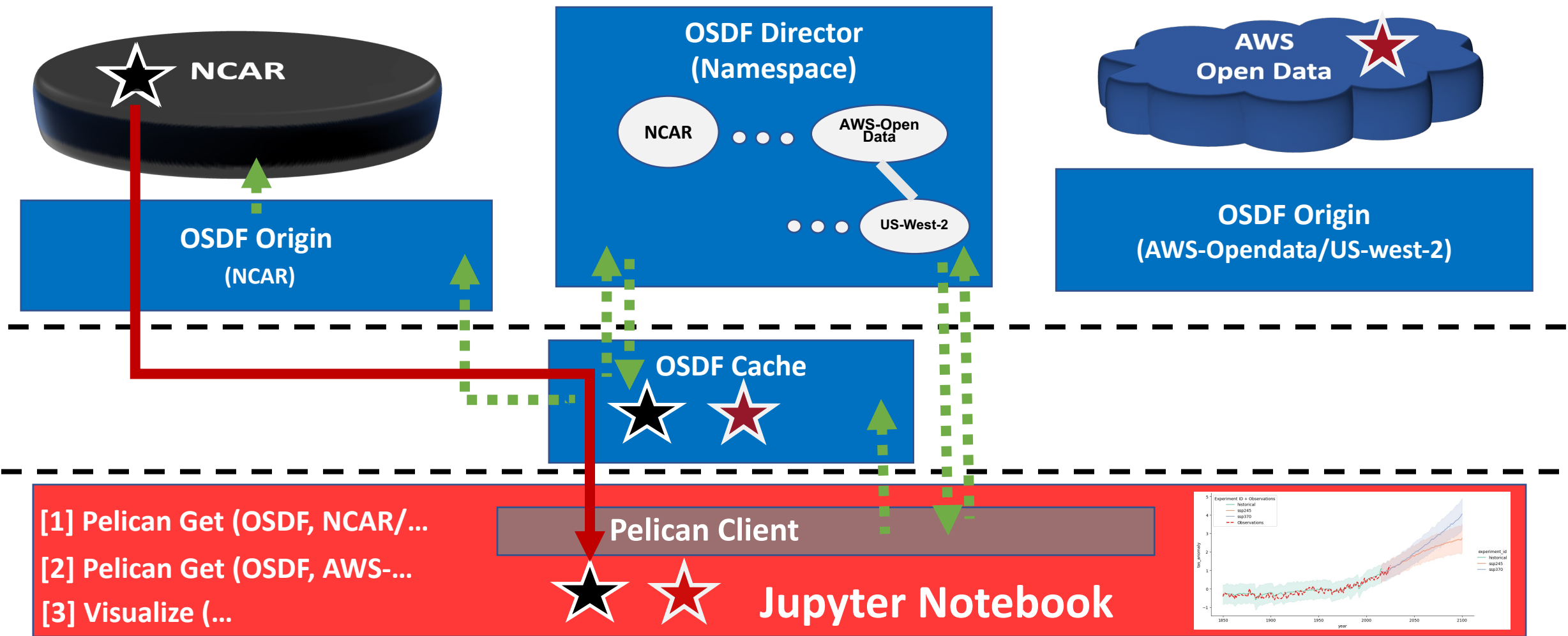


**Origin Service**



**Global OSDF Service**





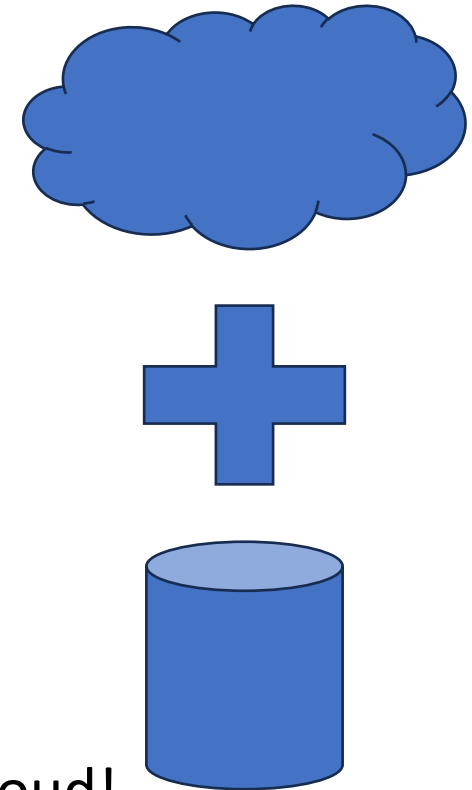
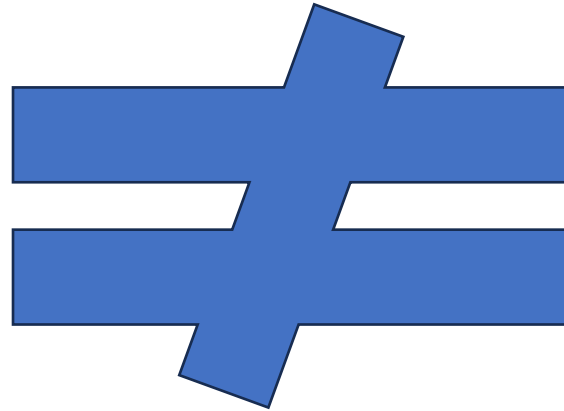
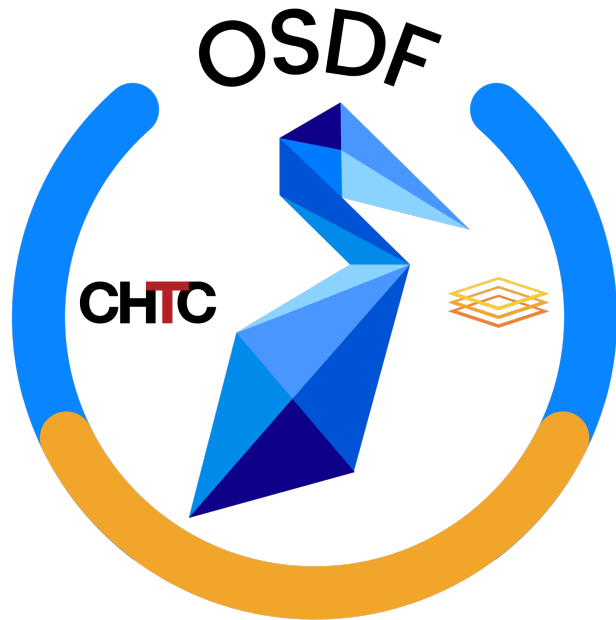
Researcher uses a Jupyter Notebook to create a visualization that requires two objects:

★ `NCAR/rda/harshah/osdf_data/HadCRUT.5.0.2.0.analysis.summary_series.global.monthly.zarr`

★ `AWS-OpenData/US-West-2/cmip6-pds/CMIP6/CFMIP/NCAR/CESM2/aqua-4xCO2/r1i1p1f1/Amon/co2mass/gn/v20190816`



# How to leverage the OSDF?



The OSDF is not some dusty disk-based FTP server in the cloud!  
The OSDF is not a “free S3”!

How can your community use the service effectively?





## OSDF's value for a CICI PI:

Stream your data to your user community => Saves your project \$\$\$

Implement your authorization policy => Saves your project time

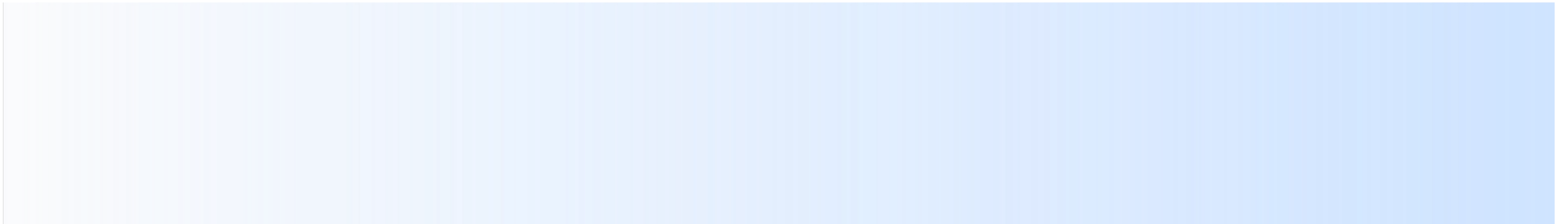
Tracks usage data => Saves you PI time

Allow users to use NSF compute to scale => Grows your community

**Important note:** If you do not have a location to store a disk copy of the dataset, PATH can also host a time-limited copy via the sister OSStore service.



# Integrating Datasets with the OSDF





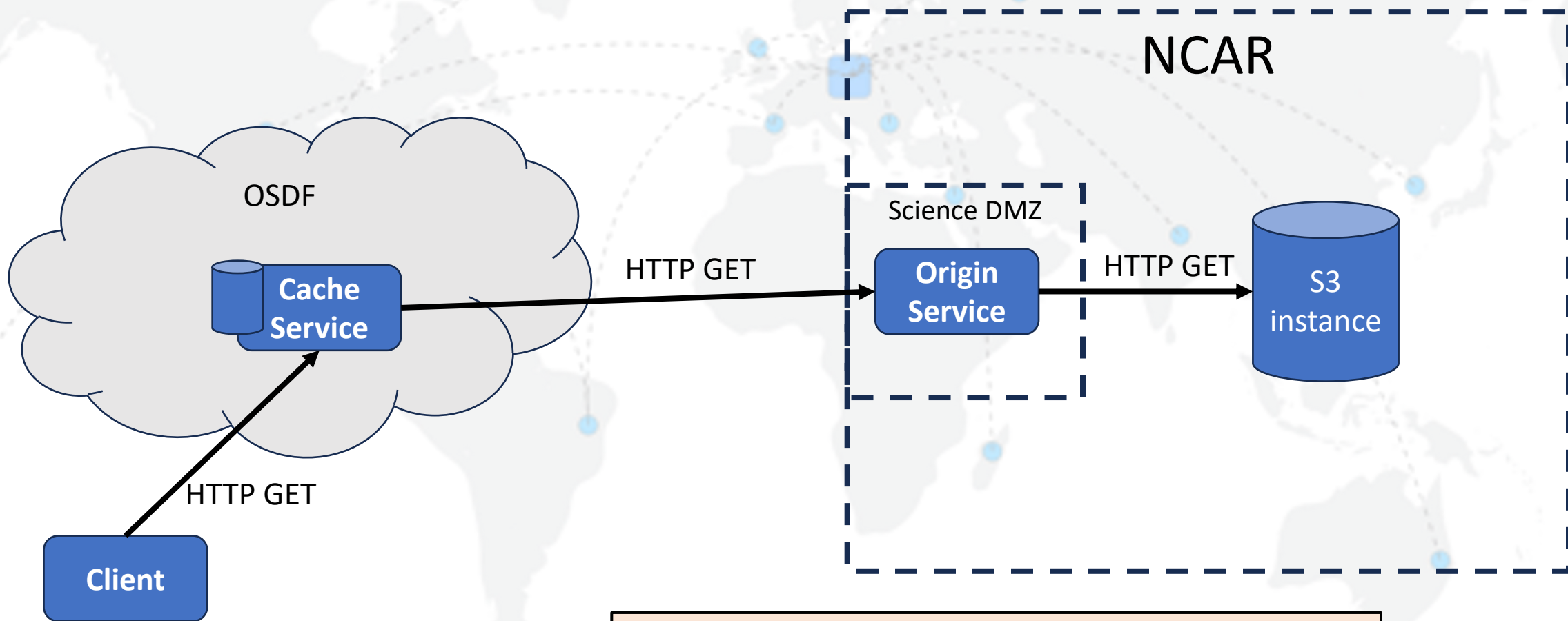
# Recipe #1: NCAR

- **Scenario:** NCAR has a big S3 instance with petabytes of public datasets it wants to effectively distribute to large community.
  - Transition community from “download then compute” to “stream to compute”.
- **Solution:**
  - PATH deployed an **origin service**, pointing at NCAR’s on-prem S3 instance.
  - All data (objects) in the bucket are exported to OSDF via the origin.
  - NCAR updated their online portal to provide HTTP download links to OSDF.
  - NCAR updated sample Jupyter notebooks to use Pelican Python client instead of assuming locally-downloaded data.
- **Example:** Collaborators in Korea now leverage local cache instead of streaming from across the planet.





# NCAR Setup



**Important note:** the origin service can also be hosted externally in Internet2; need not be on campus.



# New Feature #1: AWS Caching

Coming  
October 2025

What if NCAR's data lived in AWS S3, not on-prem?

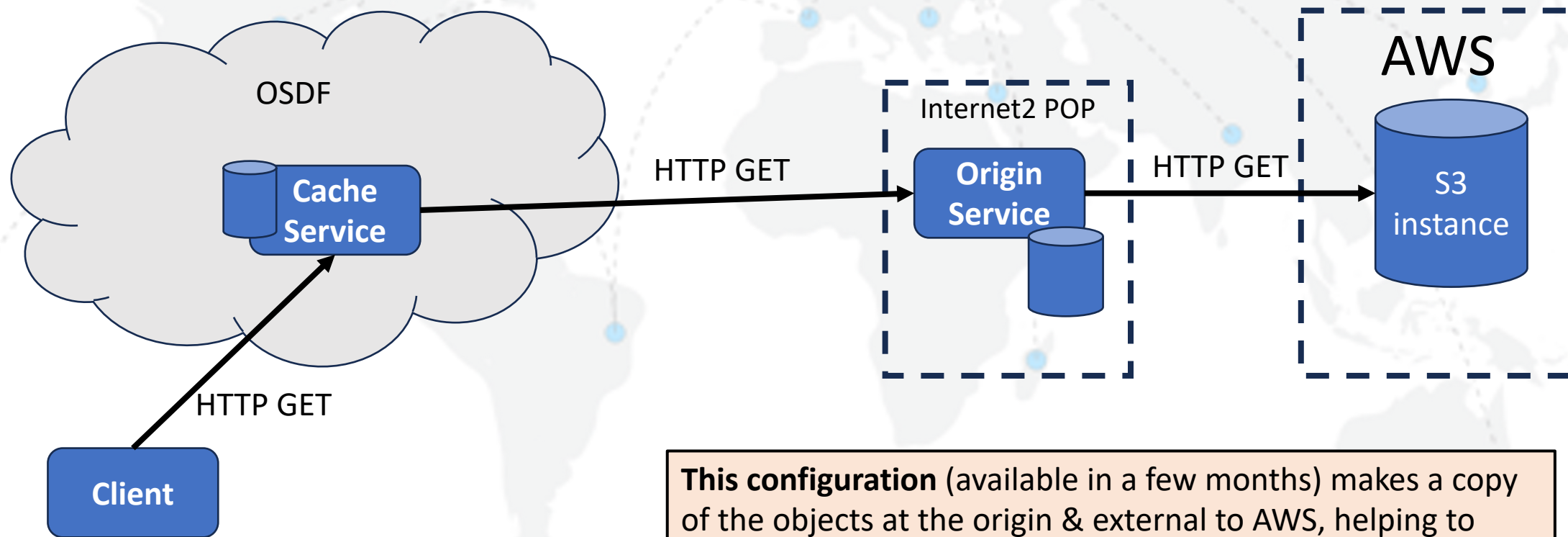
- In that case, streaming publicly can be **very expensive**.
- But NSF funds significant compute outside AWS! How to manage **egress costs**?

OSDF can operate the origin in the NRP and attach large (slow) storage

- Data egresses once!



# Alternate Setup



**This configuration** (available in a few months) makes a copy of the objects at the origin & external to AWS, helping to manage egress costs.



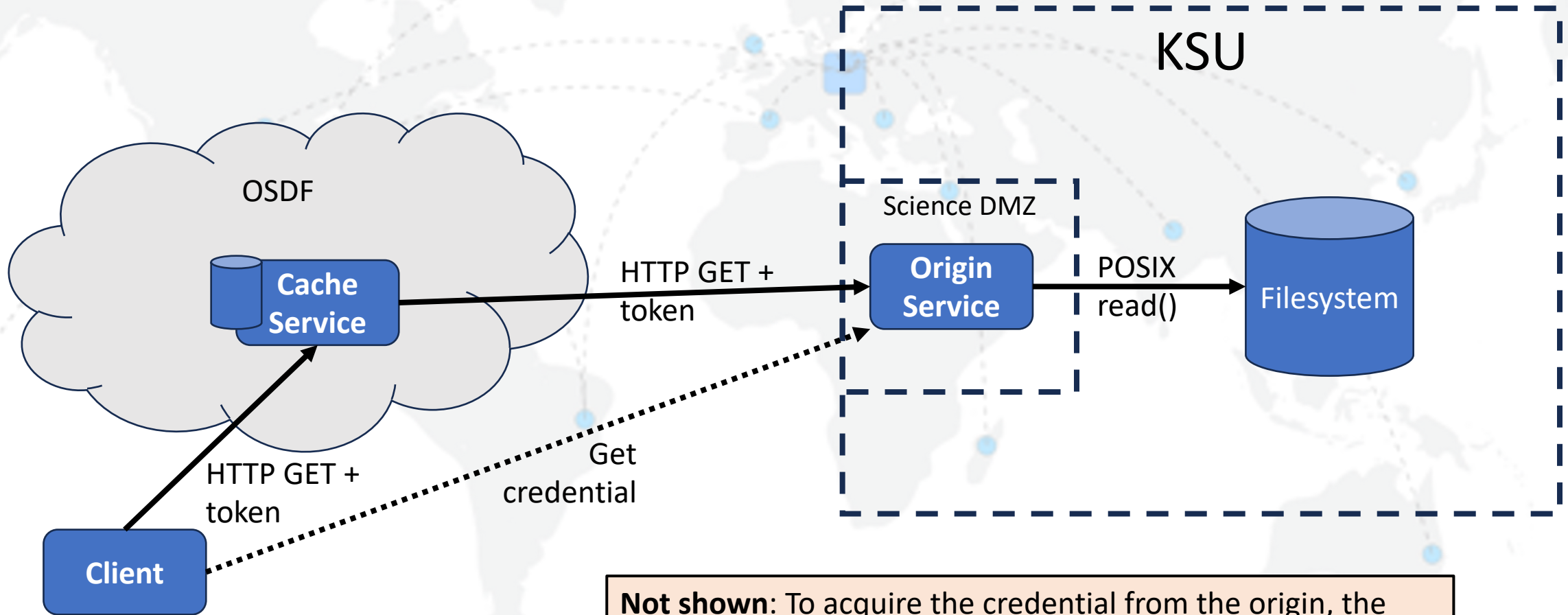
## Recipe #2: FlamingoSim @ KSU

- **Scenario:** PI at Kennesaw State University (Ramazan Aygun) wants to share a single dataset “FlamingoSim”, located on a shared filesystem, with his collaborators.
- **Solution:**
  - PATH deployed an **origin service** on a host with the filesystem mounted.
  - PATH configures the origin to only authorize specific collaborators to download data.
  - Authentication is done via their institution’s Single Sign On (no new user accounts!) using CILogon.
  - Ramazan hands collaborators the OSDF object names and teaches them to use the `pelican` CLI to download files.





# FlamingoSim @ KSU Setup



**Not shown:** To acquire the credential from the origin, the user must authenticate to their home institution's SSO (or ORCID) in a browser.



# New Feature #2: Self-Service Group Management

Available now!

- PATH currently configures the authorization policies for FlamingoSim to match the PI's requests.
  - Pro: Uses institutional SSO, simple for user.
  - Con: Requires a different human to change settings!
- Datasets can now be onboarded using group-based access.
- Dataset owners can add or remove collaborators via the web portal.
  - Portal uses the Comanage software operated by CILogon.
- **This fall:** self-service APIs/web portal for more fine-grained access, simplifying process of publishing many datasets.



# Portal Integration & Advanced Topics

Each community tends to have unique authorization setups:

- If there's an existing web portal, it can generate the authorization token directly and embed it into a URL (a-la S3's pre-signed URLs).
- Options exist for direct creation of long-lived credentials for automation and web browser-based flows: **talk to us!**

Under the hood, OSDF is “just” HTTPS:

- Instead of using our tools, you can create a HTTPS URL for your users.
  - If it speaks HTTPS, it can leverage OSDF! Opens a world of possible integrations.
- The simplest OSDF client is a browser!



# Unlocking Value with the OSDF

What have existing communities done with the OSDF?



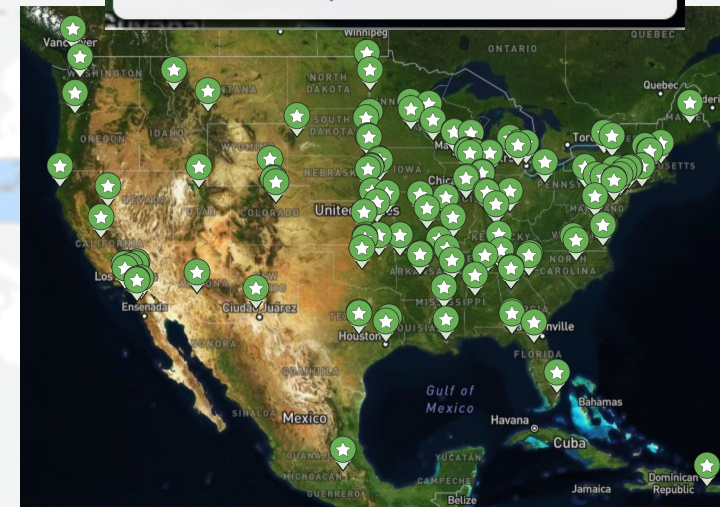
Experience the OSPool environment at <https://notebook.ospool.osg-htc.org/>

# OSPoo - Compute

- The OSPool is a PATH-run service that aggregates compute capacity from across the nation into a single service.
- Users can place their workloads at an access point and process inputs via the OSDF at scale.
- On a typical day, >600K jobs take inputs from the OSDF.
- ~150 OSPool projects used the OSDF in the last year.
- See more: <https://osg-htc.org/services/osdf/projects>

Give your users the “gift” of computing at scale on your dataset!

OSPoo Contributor  
Updated 8/20/2025, 12:12:53 PM  
145 Sites, 99 Institutions



**On August 19**

**2M jobs completed**

**Placed by 78 researchers**

**Triggering 29M file transfers**  
**Consuming 840K core hours**





# Notebooks - Visualize

OSDF can be accessed directly from a Jupyter notebook, using the Pelican Python library.

- Libraries that support FSSpec (Pandas, XArray, Zarr, ) can stream from OSDF.
- Don't have a JupyterHub? Can always leverage the [National Research Platform](https://nationalresearchplatform.org/).

The screenshot shows a web browser window displaying a Jupyter notebook titled "Analyzing Salinity Patterns in". The URL is [projectpythia.org/osdf-cookbook/notebooks/envistor-technical/](https://projectpythia.org/osdf-cookbook/notebooks/envistor-technical/). The notebook content includes a text block and a code block.

We use `BytesIO` to read the content as a stream before passing it to `pandas.read_excel()`. Each resulting DataFrame includes a "Station" column to identify its source location.

```
pelfs = OSDFFileSystem()
file_buoy1 = pelfs.cat('/envistor/CREST_Buoy_2_NW_Biscayne_Bay_-_S_of_Biscayne_Canal_04')
file_buoy2 = pelfs.cat('/envistor/CREST_Buoy_3_Haulover_Inlet_100518_-_073020_updated.')
file_buoy3 = pelfs.cat('/envistor/CREST_Buoy_3-2_Little_River_042121-050624.xlsx')

excel_file1 = BytesIO(file_buoy1)
df_file_buoy1 = pd.read_excel(excel_file1)
df_file_buoy1['Station'] = 'Buoy - Biscayne Bay'

excel_file2 = BytesIO(file_buoy2)
df_file_buoy2 = pd.read_excel(excel_file2)
df_file_buoy2['Station'] = 'Buoy - Haulover Inlet'

excel_file3 = BytesIO(file_buoy3)
df_file_buoy3 = pd.read_excel(excel_file3)
df_file_buoy3['Station'] = 'Little River'
```

**Clean and Combine the Data**

The right sidebar shows a "CONTENTS" menu with the following items: Imports, Load the Curated Salinity Datasets, **Clean and Combine the Data** (highlighted), Resample and Aggregate, Visualize the Salinity Patterns, Interpret the Results, and Next Steps.

Need inspiration? Check out <https://projectpythia.org/osdf-cookbook/>



# Browser - Download


You can use the fact OSDF is HTTP-based to embed URLs in your web portal – and combine your website with OSDF's distribution!

NCAR does this for their RDA service.

**Note:** by adding credentials generated by the portal into the URL, you can leverage existing authorization.

NCAR RDA Dataset d010092 x +

rda.ucar.edu/datasets/d010092/dataaccess/ ☆ New Chrome available ⋮

 **Community Earth System Model v2 Large Ensemble (CESM2 LENS)**  
d010092 ☆

ASK A QUESTION >

DESCRIPTION DATA ACCESS CITATION DOCUMENTATION

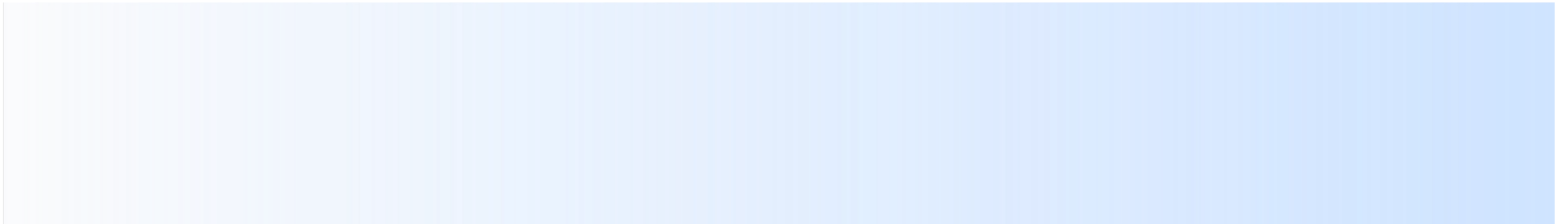
SOFTWARE METRICS

Mouse over the underlined table headings for detailed descriptions

DATA DESCRIPTION	DATA FILE DOWNLOADS		NCAR-ONLY ACCESS
UNION OF AVAILABLE PRODUCTS	<u>Web Server Holdings – Powered by OSDF</u>	<u>Globus Transfer Service (GridFTP)</u>	<u>Central File System (GLADE) Holdings</u>
	Web File Listing	Globus Transfer	GLADE File Listing



# What Next?





# Browse Some Examples

Interested in what others have done?

- We maintain a website highlighting various integrations and available datasets.
- Includes sample public objects and links to further info.
- Enjoy your RouteViews data!

<https://osg-htc.org/services/osdf/data>

Explore the OSDF | OSG

osg-htc.org/services/osdf/data?repository=routeviews

## RouteViews Close

The RouteViews dataset provides a map of the Internet, as seen by participating sites. The information, collected from the **BGP** tables of routers, includes both current and historic "snapshots". This allows operators of major Internet services to detect changes to the map in near-real time and for researchers to understand the historical evolution of the Internet.

The RouteViews dataset is funded by University of Oregon's **Advanced Network Technology Center**, and by grants from the **National Science Foundation**, **Cisco Systems**, the **Defense Advanced Research Projects Agency**, **Juniper Networks**, Sprint Advanced Technology Laboratories, **Catchpoint** and the providers who graciously provide their BGP views.

[View Datasets](#)

Organization	Field of Research
University of Oregon	Computer Systems Networking and Telecommunications

### Download a Public Object

With Pelican Client on the Command Line

```
pelican object get osdf:///routeviews/chicago/route-views.chicago/bgpdata/2025.03/RIBS/rib
```

From Your Browser

[Click to Download Public Object](#)

[Contact Us!](#)





# Leveraging more Data Services

What do you need beyond dataset delivery?

- Building out Jupyter notebook-based cookbooks?
- Building out a catalog of the published datasets?
- Generating metadata to attach to the data?
- Want a runtime environment for teaching classes?
- Have some bespoke computation service needed for your data?



The next talk about the National Data Platform (NDP) covers these higher-level services!





# CHTC uses Translational Computer Science

From the Laboratory to the Community,  
Advancing Throughput Computing Through  
**Translational Computing**





# Working with you is how we get our research done!

Translational CS requires a locale (the OSDF) to test out ideas and a community (CICI) to provide feedback

We are not doing this for altruism: Your feedback and partnerships are how we improve!



**Let's work together:**

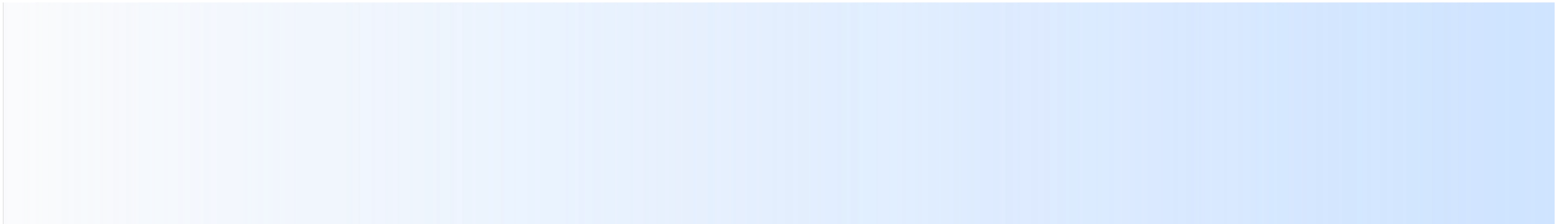
**[https://path-cc.io/cici-awardees/  
support@osg-htc.org](https://path-cc.io/cici-awardees/support@osg-htc.org)**

## Questions?

This project is supported by the National Science Foundation under Cooperative Agreements OAC-2331480. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.



# Backup Slides





Operated by PATH  
(NSF #2030508)



# A national object delivery network



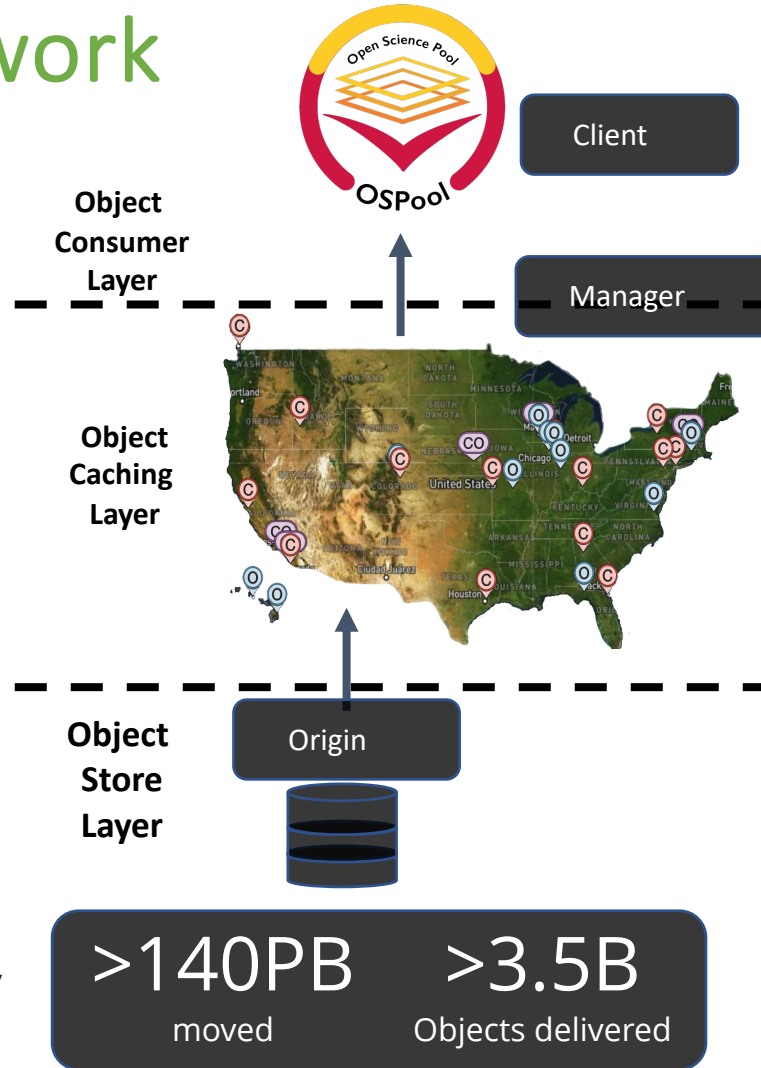
OSDF is powered by the  
Pelican software  
(NSF #2331480).

## Who can benefit?

- **Open** to any federally-funded science repository or object store.
- **Researchers** can access/connect their data to local or national compute by browser or notebook
- **Data providers** can share via FAIR principles.
- **Compute providers** can cache datasets on-site
- **Community platform providers** can build gateways and portals to large-scale datasets.

## Features

- No direct cost for integration.
- Supports automation of data-intensive workloads
- Unified name space preserves local autonomy
- Unified authorization infrastructure for restricting object access.



**34 caches** throughout the world, at points of presence within major compute centers and the global R&E networks (ESNet, Internet2).

## Example use cases

- Enable large-scale workloads that have significant **reuse and redelivery** of objects at remote autonomous object stores.
- Manage loads on object stores and **avoid egress costs** via caching & staging.
- **DOE**: FNAL (LHC, DUNE), JLab, Fusion
- **NOAA**: Via AWS OpenData
- **NSF**: NCAR, NRAO, LIGO, EHT

## Key concepts

- **Object**: both data & non-data objects (e.g., containers, models, libraries).
- **Object store**: a filesystem, S3-compatible endpoint, or HTTP webserver.
- **Origin service**: connects the backend object store with national infrastructure. Objects are read (GET) and written (PUT) through Origin.







*“The Partnership to Advance Throughput Computing (PATh) project will expand **Distributed** High Throughput Computing (dHTC) technologies and methodologies through innovation, **translational** effort, and large-scale adoption to advance the Science & Engineering goals of the broader community.”*

**PATH Proposal 04/21/2020**