



CHTC



FabAID:

Fabric for AI-Driven Science

IDSS Kickoff; NSF #2609485

PI: Brian Bockelman, Morgridge Institute for Research

co-PI: Miron Livny, Jason Simms, Anthony Gitter, AnHai Doan,



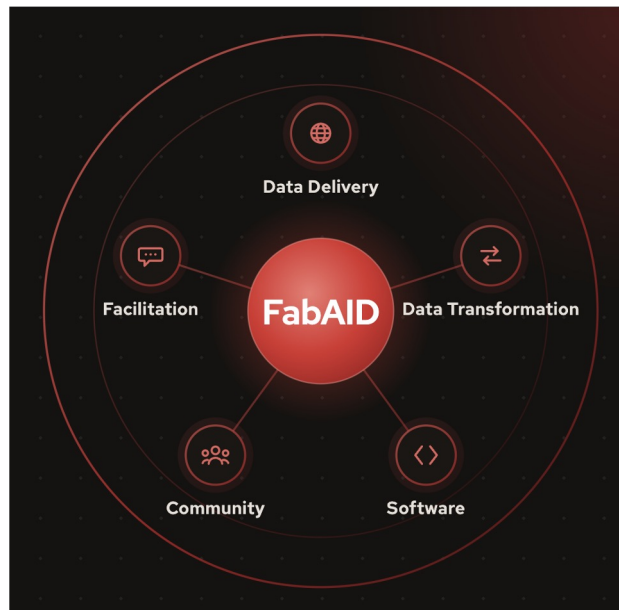
UC San Diego

FabAID - Building a Fabric for Accelerating Science

FabAID builds on **services** (*OSDF, OSPool*), **software** (*Pelican, HTCondor*), and proven **methodologies** (Facilitation) to allow the science and engineering community to accelerate their research.

Key ideas:

- Integrate existing investments,
- Involve All of America, and
- Pull together computing and data into a single fabric.



<https://fabaid.io>



FabAID

CHTC

FabAID - Ingredients

Services



Software



FabAID

CHTC



FabAID Targets

Starting with these ingredients, what do we want to tackle?

1. **Delivering datasets** to and from compute capacity **across (inter) national CI**.
2. **Generating & managing** datasets for use with AI training.
3. **Interfacing** the national CI with the evolving AI-driven ecosystem.
4. **Facilitating** the use of advanced CI services for science.
5. Active **metadata management and extraction**.
6. **Integrating archives** into the national CI.

We use Translational Computer Science to impact these 6 targets!



FabAID

CHTC

Example - Processing LIGO Data

LIGO Lab hosts the LIGO datasets at **Caltech**



How to get the LIGO data to the scientist's available capacity?

LIGO Lab stewards the LIGO data but does not have the compute capacity to meet the needs of the scientific collaboration.

The collaboration has compute capacity (cluster) or allocations (cloud, ACCESS): how do they transparently deliver data?

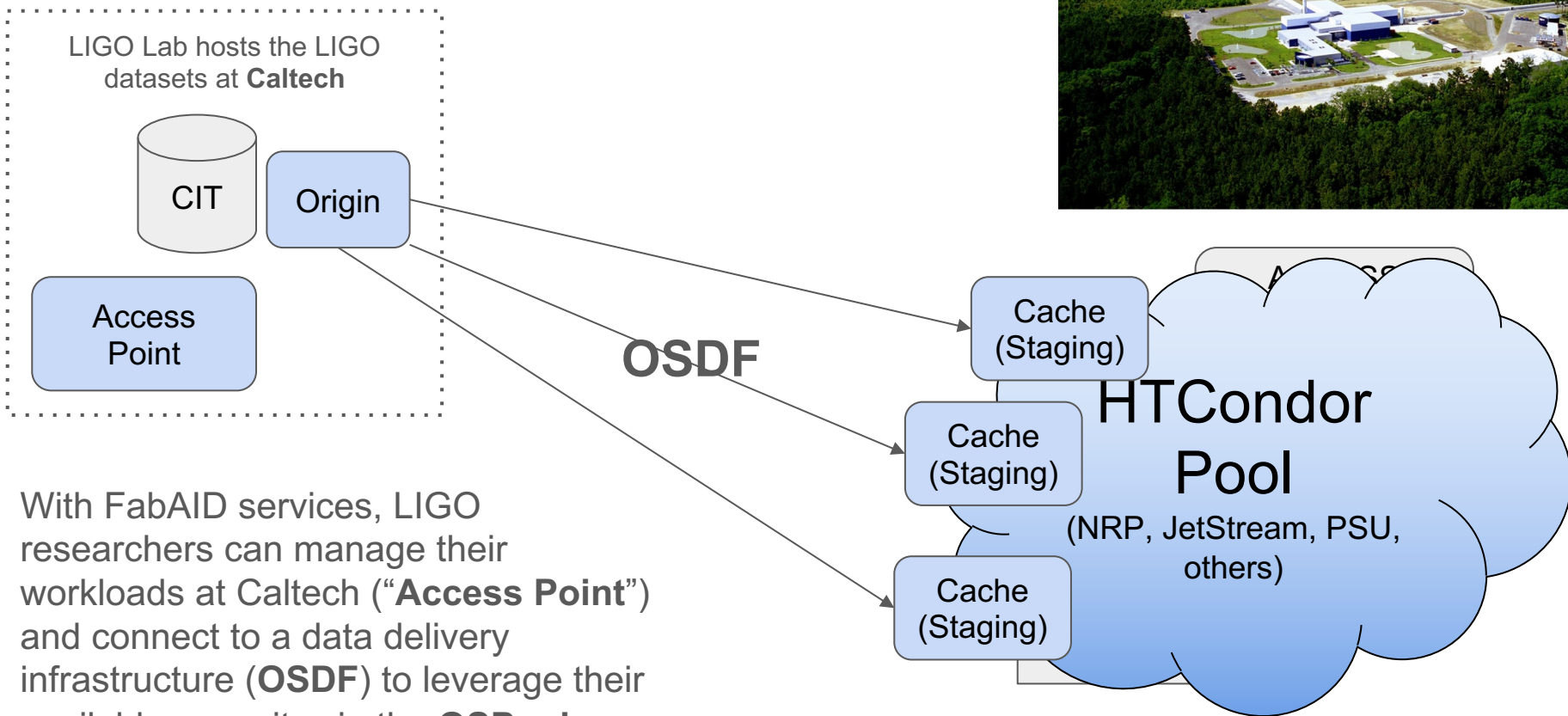


NRP

PSU

Australia
(OzSTAR)

Example - Processing LIGO Data



With FabAID services, LIGO researchers can manage their workloads at Caltech (“**Access Point**”) and connect to a data delivery infrastructure (**OSDF**) to leverage their available capacity via the **OSPool**.



Vision - Weaving the National Cyberinfrastructure Together



The nation's universities and labs hold a wealth of knowledge. *Leverage* the specialization by simplifying their ability to **bring and share across** a national fabric.

FabAID builds on the foundation of an existing fabric of **distributed services** that delivers compute and data capabilities across institutional boundaries.



FabAID

CHTC

Vision

Any Researcher ...

Can Connect from Any
Campus ...

To run inference against their
dataset ...

Through sharing, we turn
this to a virtuous loop!

FabAID

Regardless of where the data
resides in the US!

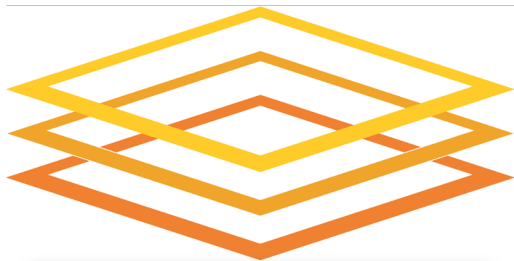


Building on a Partnership

FabAID builds on 20+ years of collaboration between the OSG Consortium and CHTC.

- The CHTC provides software, technologies, and innovations for Throughput Computing, including Pelican and HTCondor.
- The OSG Consortium's Fabric of Services provides a national community with curated services and solutions for Throughput Computing.

The CHTC + OSG Consortium are in a decades-long Partnership to advance education and research in the US. **FabAID takes the Partnership into the AI era.**



The human factor:
CHTC and OSG bring
decades of experience
in CI.



The **FabAID** Project

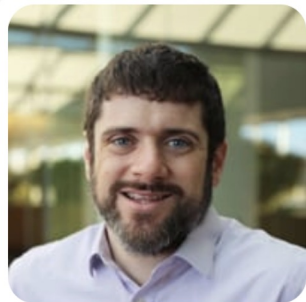
The FabAID Team

FabAID is led by the team at CHTC in Madison, split between two institutions (Morgridge & UW-Madison).

- Combines decades of experience across software, services, and methodologies.
- Core presence within the NSF CI ecosystem.

Brian Bockelman

Principal Investigator
Morgridge Institute for Research



Miron Livny

Co-Principal Investigator
University of Wisconsin–Madison

The FabAID Team

Jason Simms

Co-Principal Investigator
Swarthmore College



Not only CI practitioners, FabAID weaves in experts from across scales and domains!

Simms helps FabAID connect to the national community beyond the R1s.

Doan brings in a wealth of industry & academic experience in metadata extraction and management.



AnHai Doan

Co-Principal Investigator
University of Wisconsin–Madison

Tony Gitter

Co-Principal Investigator
Morgridge Institute for Research



Gitter brings experience in AI-driven science through domain science projects like AI protein factories.

A Project At The Starting Line

FabAID builds its fabric on top of existing services and software: what is the “starting point” for the project?

7 NSF Major Facilities: ATLAS, CMS, IceCube, LIGO, NCAR, NRAO, Rubin.

OSDF

BYTES READ

286,077 TB

↑ 9 GB/s Last 1 Year



OBJECTS READ

6,504,083,635

↑ 206/s Last 1 Year



>180 projects used the OSDF as part of computations on the OSPool in the last 12 months.

OSPool

On July 17

2M jobs completed

Placed by 55 research projects

Triggering 18M file transfers

Consuming 1M core hours

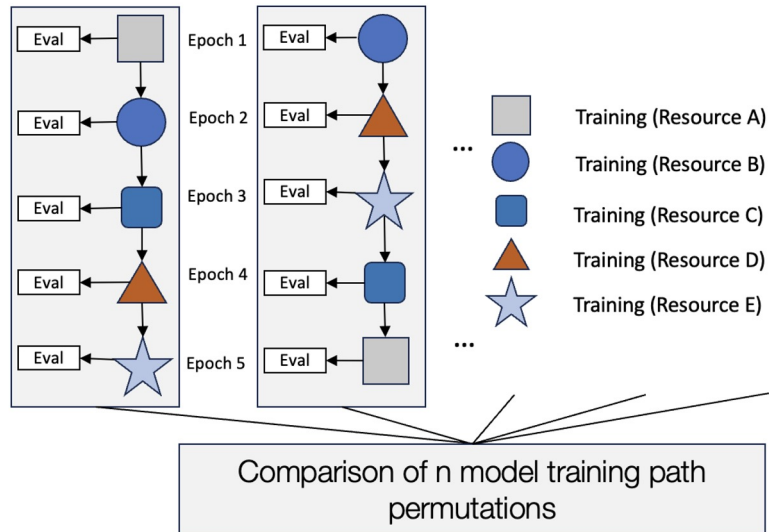
Harnessing capacity from 81 institutions.

IDSS(&S!)

We think of our project in terms of the IDSS(&S) program:

Integrated	Integrate data repositories and compute capacity at institutions into a coherent fabric. Create an environment that is functional, trustworthy, supported, and with E2E training.
Data	Extend prior projects to focus on the data lifecycle : new interfaces for data generation, transformation, and incorporate deposit into the (tape) archives.
Services	Operate dependable national-scale services that meet the needs of emerging AI-driven workloads (scale-out data reuse, copying, transformation, and archive).
Systems	Together with FabAID user-facing systems (APs, origins) connected to the fabric, institutions can bring their autonomous storage/compute systems.
(Bonus!) & Software	FabAID's services are powered by open-source software projects led by CHTC: HTCondor is used at hundreds of locations across science, government agencies, and industry. <u>Broadens overall project impact.</u>

Integrated Example: NAIRR Throughput AI Project



Can you run AI training workloads across all available resources?

Can we apply throughput computing principles to training? In our demo, we:

- Placed our training workload & dataset at an AP in Wisconsin.
- The AP targeted each epoch at a different allocation (or first-available).
- OSDF/Pelican moved training dataset to caches near site; copied back checkpoints to Wisconsin.



Work done by
PATH, NSF
#2030508.

Integrated Example: NAIRR Throughput AI Project

Result?

~40 models trained, NAIRR allocations utilized (plus several gracious extensions!), plus 13 other universities.

Services incorporated in FabAID ensures seamless delivery of data and storing of checkpoints.



Work done by
PATH, NSF
#2030508.

Total epochs	1,261
Total GPU hours	19,695
Total data moved via OSDF	~43.4 TB
Unique GPU types	9
Unique sites	6 (OSPool, Expanse, Delta, Bridges2, Anvil, AWS)
Unique sites, with OSPool sites breakout	13

FabAID Services

OSDF

The Open Science Data Federation connects data repositories via the origin, using strategically-placed staging services (caches) to **scale out data reuse**.



OSStore

The OSStore allocates (time, duration-limited) storage to projects from NSF-funded distributed storage systems.
Accessible via the OSDF.

FabAID

OSPool

Aggregates capacity from across the US, making it available to connected APs, executing data processing workloads for a range of research.



CHTC



OSDF: A National Object Delivery Service

Value Proposition

- **Shares** your data (Objects) with authorized consumers.
- **Manages** the load of Object reuse on your Object Store.
- **Fuses** Objects from diverse Object Stores through a common namespace and API.
- **Stages** Objects in network-embedded storage.

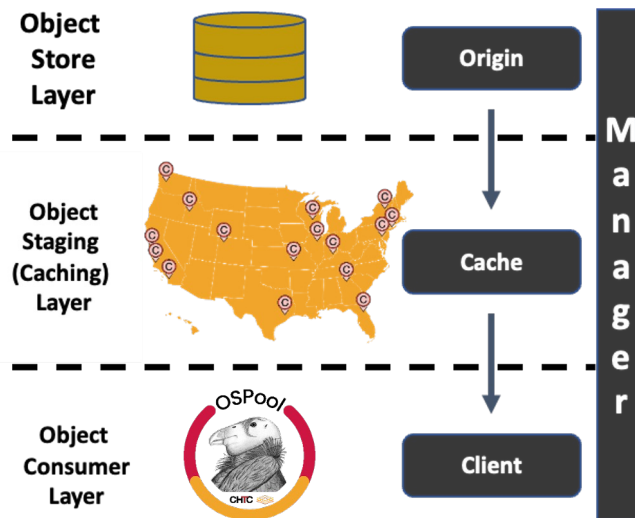
Who can benefit?

- **Open** to any US researcher, data (object) provider, or computing facility.

>280
PB moved

>6.5B
Objects delivered

In the last 12 months.



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

Who is using the OSDF?

- Workloads that have significant **reuse and delivery** of objects at remote autonomous Object Stores.
- Object Stores with limited capacity or egress costs.
- **NOAA**: Via AWS OpenData
- **NSF**: NCAR, NRAO, LIGO, EHT
- **DOE**: FNAL (LHC, DUNE), JLab, Fusion

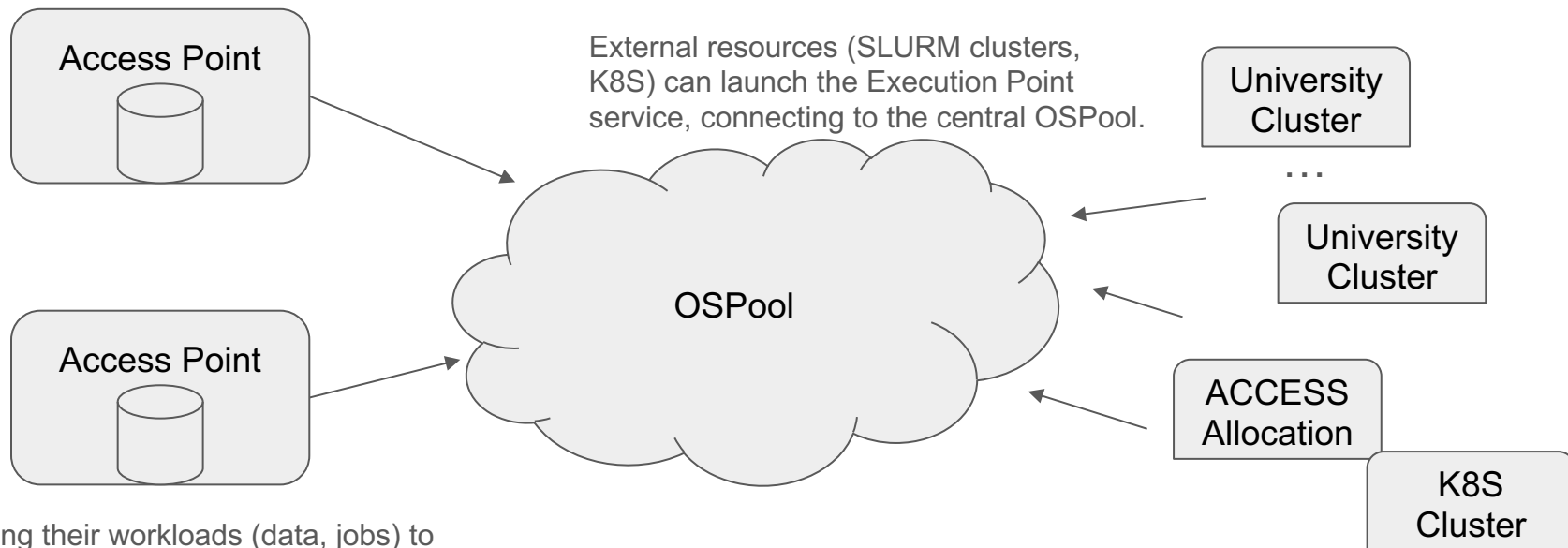
Key concepts

- **Object**: both data & non-data objects (containers, models, libraries).
- **Object Store**: a S3-compatible endpoint, filesystem, or HTTP webserver.
- **Origin**: connects an Object Store with national infrastructure. Objects are moved from (GET) and to (PUT) the store through Origin.



FabAID Services: OSPool

The OSPool is a distributed HTCondor system spanning 100 institutions.



Users bring their workloads (data, jobs) to an independent **Access Point** which can

Note: The PATH project (NSF #2030508) is responsible for connecting campuses to the pool. ²⁰

FabAID Systems

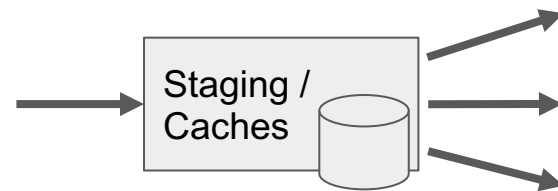


Origins

The *origin* service connects storage to the wider OSDF.

The origin can be run on-prem or hosted by FabAID if storage is remotely accessible or accessible via NRP.

Currently ~50 hosted origin services.



Staging (Caches)

The staging space provides management and scale-out close to the compute. “Hot” objects are staged once and delivered many times.

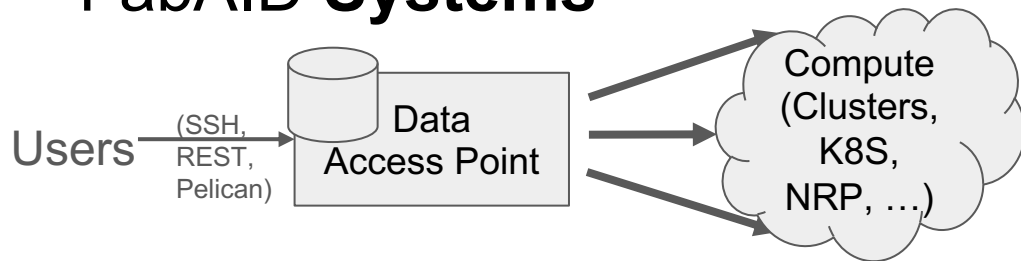
FabAID runs ~3 dozen in the network (I2, ESNet) via NRP and at major sites.



FabAID

CHTC

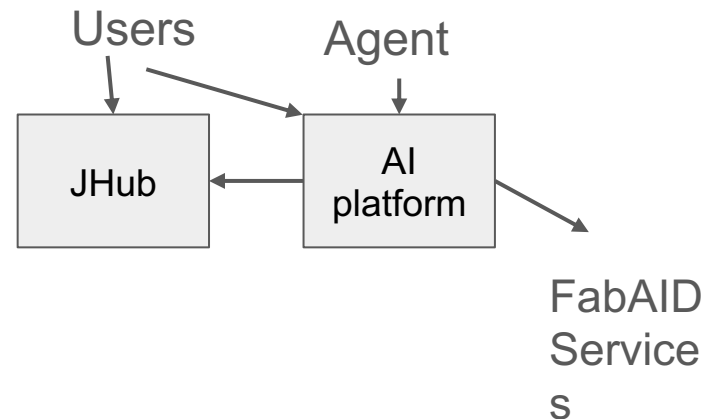
FabAID Systems



(Data) Access Points

The AP manages the users' workflow and connects to remote resources for execution.

In FabAID, we host general-purpose APs and support others' use; an increased focus on E2E data movement.



Interactive Computing (JupyterHub, AI Agent)

FabAID has an increased focus on new interfaces, broadening user communities.

This includes in-browser (JupyterHub) and plugins for agents to other services (driving workflows or data movement).



Systems Example - Inference with AlphaFold3

AlphaFold3 is a great hybrid data-intensive / AI workflow:

- ~Half of runtime is CPU-based, querying large, 100's GB databases (sequence alignment step).
- Half is traditional inference, feeding aligned data into a GPU model.

Graduate student at FIU Pienaar lab approached the facilitation team: **can you help with running AF3 on 1B pairwise sequences?**



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



FabAID

CHTC

Systems Example - Inference with AlphaFold3

Key observation: creating a repository of the alignment outputs (a cache) allows one to skip the CPU-heavy portion.

New: OSG's alignment service allows researchers to perform inference and save the alignment step.

Builds on AP & storage systems.

OSDF delivers the 750GB database, returns alignments, and delivers cached output of prior runs.



440k alignments cached

10k source organisms

371k unique sequences

62k community contributions



FabAID

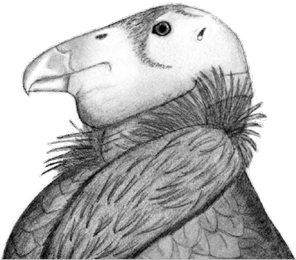
CHTC

FabAID Software

Each of FabAID's systems & services are based on software from the CHTC: anyone can install and setup their own services.

Industry and other government agencies are the biggest deploys by hardware (\$\$\$). Shows up in varied places like SpaceX ([link](#)) or Google's Quantum simulators ([link](#)).

FabAID services are a *flagship instance* with much broader impact.



Over 200 sites have self-reported running our software

FabAID Target Audience - The “Who” of the IDSS&S

Data Producer:

Research group generating datasets (via inference?)

Data Provider:

Established domain repository exporting datasets.

Data Mover:

Copying large datasets between origins.

HTC Consumer:

Processing datasets at scale on distributed resources

Interactive Consumer:

Downloading objects via the browser or Jupyter

Metadata Services:

Cataloging metadata across repositories

Federation Builder:

Setting up standalone services (data xform, delivery)

Message: FabAID integrates across the national ecosystem! We shrink the gaps for your data-intensive and AI-driven science!



FabAID

CHTC

Co-Federations: Linking together collaborators

FabAID sees its services as the *flagship* instance for our Translational Computer Science.

We envision independent instances (HTCondor pools for data processing; Pelican data federations for delivery) for autonomous projects

- **Example:** CMS and LIGO both run their own pools on their dedicated capacity.

Opportunity: share/leverage resources at the national scale.

- This is **flocking** for HTCondor, where one AP connects to multiple pools.
- Pelican envisions **co-federations**: sharing the nation's staging capacity across multiple independent data federations.



FabAID

CHTC

Race To Operations

We have a 12 month “**race to operations**” for the new Fabric (while keeping existing customers going). What are the themes?

- **Scale down:** How easy/lightweight can we make deploying our connectors (origins, APs)? Easy enough for any lab?
- **Extending Data Reach:** Tackle gaps **onboarding data** (e.g., bringing it to the AP) and **offboarding** (onto an archive), part of the natural lifecycle.
- **New Interfaces:** Meet users “where they are” by building out capabilities for web UI and agentic AI interfaces.
- **IDSS Community:** Strengthening connections across the new IDSS projects, particularly with the National Data Platform.



FabAID

CHTC

How does FabAID Impact the Nation?

Impact at the National Scale - Small Institutions

Jason Simms

Co-Principal Investigator
Swarthmore College



More than **75% of American students are outside R1 universities.**

- Impacting the nation's education and research means growing the presence outside R1 schools.
 - Larger research projects struggle understanding the “on the ground” environment outside R1s and how to advance their science.
- Co-PI Simms' focus is **providing feedback** on running FabAID services at a PUI (Swarthmore), directly **facilitating**/empowering other non-R1s, and advocating for FabAID at a national scale (RCSI, CaRRC Emerging Centers).
 - “Lessons learned” get rolled back into the FabAID services and development practices.
 - Goal: Have robust annual cohorts of small institutions working together, leveraging FabAID.

See Simms' HTC26 presentation: <https://agenda.hep.wisc.edu/event/2432/contributions/37631/>

Impact - Summer School

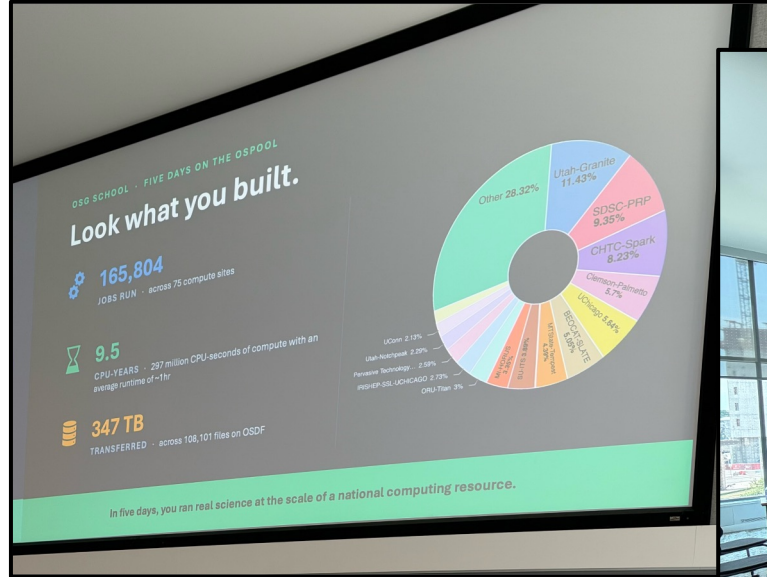


Each Year, FabAID will bring ~60 learners to the “Data-Intensive Summer School” in Madison.

- Week-long experience targeting learners (undergrad through staff) who have an existing research computing problem.
- Focus on solving real-world problems through hands-on experience on distributed systems.
- Overhaul existing successful curriculum to increase data intensity and leveraging agentic AI through new FabAID services.

Mark your calendars - **Summer 2027, Madison Wisconsin!**

CHTC's Hosted Summer Schools - Building on Excellence



Last week, we completed the 2026 version, the “OSG Summer School” - the last before the “data-intensive” overhaul for FabAID.
>1/4 PB moved by students!

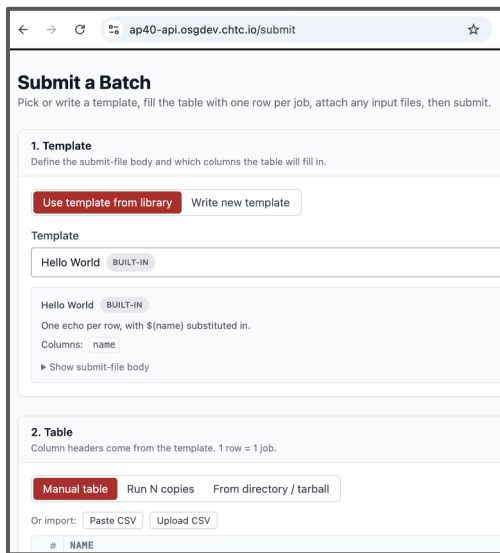


Impact at the National Scale - New Interfaces

A goal for the new fabric is to meet educators and researchers “where they are”.

We excel at the CLI - time to add web and agentic AI interfaces!

Web UI



The screenshot shows the 'Submit a Batch' page of the HTCondor API web interface. It includes a '1. Template' section with a 'Hello World' example and a '2. Table' section for defining job columns. The interface is clean and modern, with a sidebar for navigation.

Submit a Batch
Pick or write a template, fill the table with one row per job, attach any input files, then submit.

1. Template
Define the submit-file body and which columns the table will fill in.

[Use template from library](#) [Write new template](#)

Template

Hello World **BUILT-IN**

Hello World **BUILT-IN**

One echo per row, with \$(name) substituted in.
Columns: name

► Show submit-file body

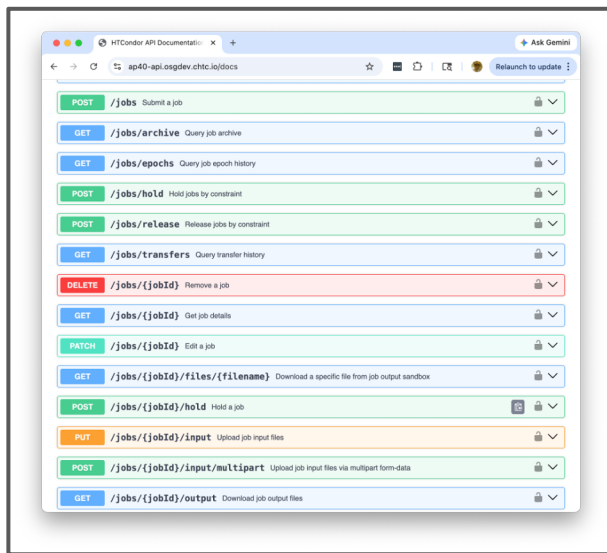
2. Table
Column headers come from the template. 1 row = 1 job.

[Manual table](#) [Run N copies](#) [From directory / tarball](#)

Or import: [Paste CSV](#) [Upload CSV](#)

NAME

REST API



The screenshot shows the REST API documentation for the HTCondor API. It lists various endpoints and their methods, such as POST /jobs, GET /jobs/archive, and DELETE /jobs/{jobId}.

HTCondor API Documentation

POST /jobs Submit a job

GET /jobs/archive Query job archive

GET /jobs/epochs Query job epoch history

POST /jobs/hold Hold jobs by constraint

POST /jobs/release Release jobs by constraint

GET /jobs/transfers Query transfer history

DELETE /jobs/{jobId} Remove a job

GET /jobs/{jobId} Get job details

PATCH /jobs/{jobId} Edit a job

GET /jobs/{jobId}/files/{filename} Download a specific file from job output sandbox

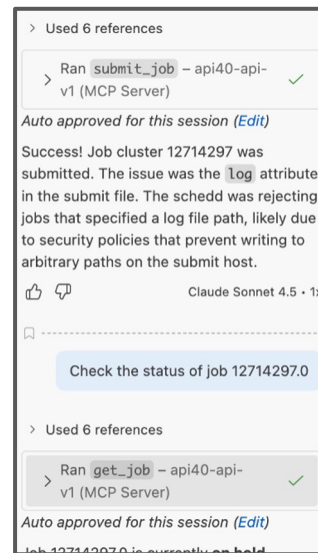
POST /jobs/{jobId}/hold Hold a job

PUT /jobs/{jobId}/input Upload job input files

POST /jobs/{jobId}/input/multipart Upload job input files via multipart form-data

GET /jobs/{jobId}/output Download job output files

Agentic AI



The screenshot shows the Agentic AI interface, which displays a list of references and a status check for a specific job. It includes a 'Used 6 references' section and a 'Check the status of job 12714297.0' button.

Used 6 references

Ran submit_job - api40-api-v1 (MCP Server) ✓

Auto approved for this session (Edit)

Success! Job cluster 12714297 was submitted. The issue was the log attribute in the submit file. The schedd was rejecting jobs that specified a log file path, likely due to security policies that prevent writing to arbitrary paths on the submit host.

👍 🗨 Claude Sonnet 4.5 • 1x

Check the status of job 12714297.0

Used 6 references

Ran get_job - api40-api-v1 (MCP Server) ✓

Auto approved for this session (Edit)

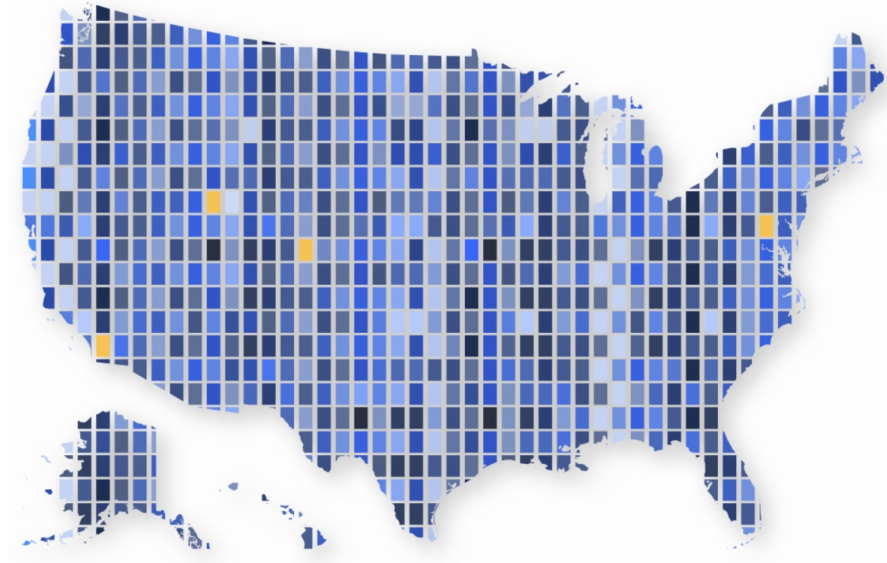
Job 12714297.0 is currently on hold

Agents show exciting potential as a universal “natural language” interface: enormous potential but still much to prove around impact on science!

Impact for IDSS - National Data Platform

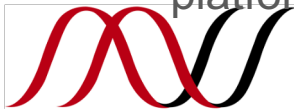
FabAID focuses on delivering computing, data, and metadata; we integrate with higher-level data services like the **National Data Platform** for E2E functionality. Year 1 projects:

- Reliable metadata delivery to NDP catalogs.
- Integrated dataset onboarding (upload to OSDF & NDP).
- Integrated online data exploration platform with unified authorization.



<https://nationaldatapatform.org/>

NSF #2609447



FabAID

CHTC

NDP & FabAID - Example joint project

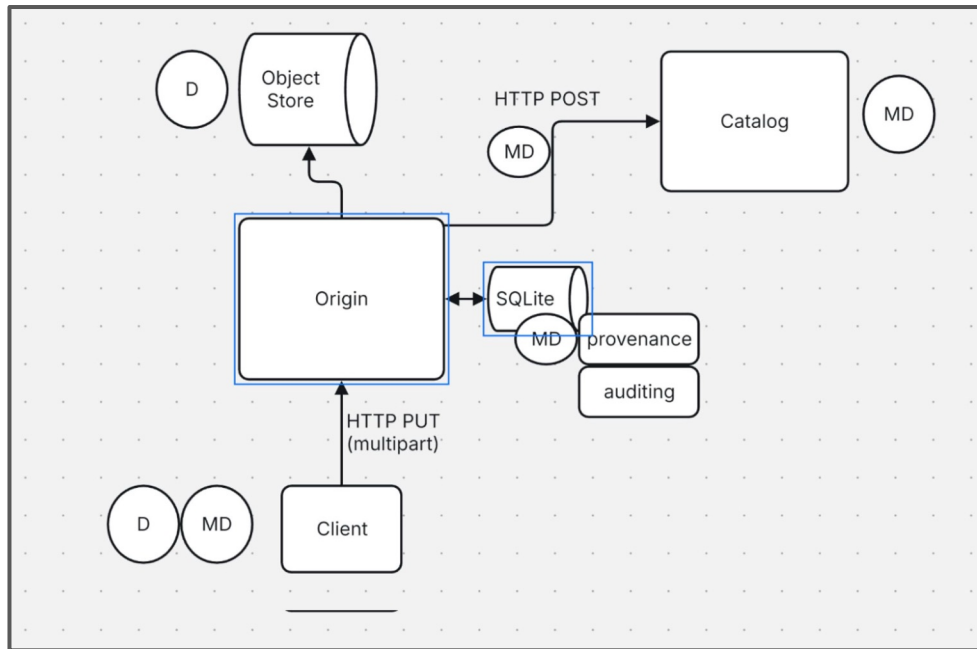


Diagram is from last design session between NDP & FabAID - current draft of the joint data & metadata upload path for clients.

Co-design allows “best of breed” between the different projects, letting FabAID handle data delivery and the NDP catalog handle metadata / higher-level services.

Hosting NDP developer in Madison in ~2 weeks!



Impact at the National Scale - You!

FabAID succeeds through integrating across the CI - **not lone projects!**

- Can we reliably extract & deliver metadata to your services?
- Can we deliver your datasets reliably to national computing infrastructures?
Power your integrated environment's connection to large-scale compute?
- Can our technologies & software be used to move computing to your data?
- Can our interfaces help empower your agentic AI work?
- **Can our facilitation team help users understand your services?**

Excited to **consult, plan, and collaborate**, especially with Category III projects, to help achieve national-scale vision!



FabAID

CHTC

FabAID ❤️ IDSS

<https://fabaid.io>

To quote the solicitation:

IDSS [...] supports operations-level **national-scale** cyberinfrastructure systems and services that **broadly advance and facilitate open**, data-intensive and artificial intelligence-driven science and engineering research, innovation, and education

FabAID is building a national platform with more than 100 institutions, weaving data repositories and computational services closer together.

We want to be a foundation that can serve your projects.



Questions? Reach out!

<https://fabaid.io>

support@osg-htc.org - Support for our Services

contact@fabaid.io - Questions on the FabAID project

This project is supported by the National Science Foundation under Cooperative Agreements OAC-2609485. 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.



FabAID

CHTC