

UC Berkeley

UC Berkeley Previously Published Works

Title

Developing a Modern Data Facility for the Genesis Mission

Permalink

<https://escholarship.org/uc/item/1mg4c5bg>

Authors

Berket, Karlo

Paine, Drew

Publication Date

2026-08-26



Developing a Modern Data Facility for the Genesis Mission

Karlo Berket, Software Architect, kberket@lbl.gov

Drew Paine, Product & User Experience Architect, pained@lbl.gov

August 26, 2026

Three key takeaways of this talk

1

Meeting core needs

Team is designing & building the reliable, resilient data services & storage urgently needed by Office of Science (SC) communities.

2

Accelerating delivery

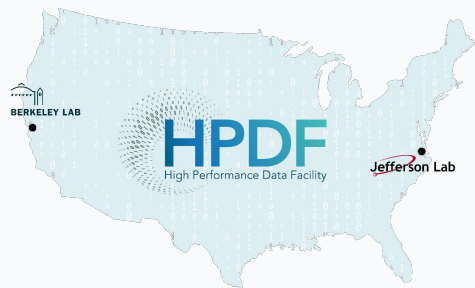
American Science Cloud & Genesis Mission are enabling rapid acceleration of our work on key resources for the SC community.

3

Enabling science

Focusing on delivering resources to help SC teams solve their biggest data problems while they focus on the science that matters most to them.

HPDF Overview – From Mission to North Star



A distributed, hybrid data facility with core staff and resources spanning both Jefferson Lab and Berkeley Lab.



North Star

Delivering reliable, resilient data services to meet DOE scientist's evolving needs across the data lifecycle.



Vision

Enabling scientists and AI systems to make new scientific discoveries by seamlessly working with distributed and federated datasets.



Mission

To enable and accelerate scientific discovery by delivering state-of-the-art data management infrastructure, capabilities, and tools.

HPDF and the Genesis Mission

Connecting foundational facilities, platform capabilities, and evolving scientific questions.

RFA Projects

Rapidly evolving scientific problems driven by expanding knowledge



Platform Projects

Transforming modes of interaction and data access layers

American Science
Cloud (AmSC)

Transformational AI
Models Consortium
(ModCon)

User Facilities

Foundational scientific resources

Office of Science

ASCR

HPDF

ESnet

NERSC

ALCF

OLCF

BES

BER

FES

HEP-NP

A user story

Imagine you are a light source experimentalist today...



Your team is using multiple DOE user facilities to collect data.

This data is rapidly processed on an HPC system. Now I have plenty to analyze today.



The LLM interface at this facility lets me rapidly ask questions of my data.



Each facility provides a different approach to collecting and managing data.

... but how do you **easily & reliably** work with your data after experimenting at different user facilities?

There is so much to keep track of and I don't know who has access to which data.

We can't store all of this locally...

How are we going to work with our data when it's all over the place?



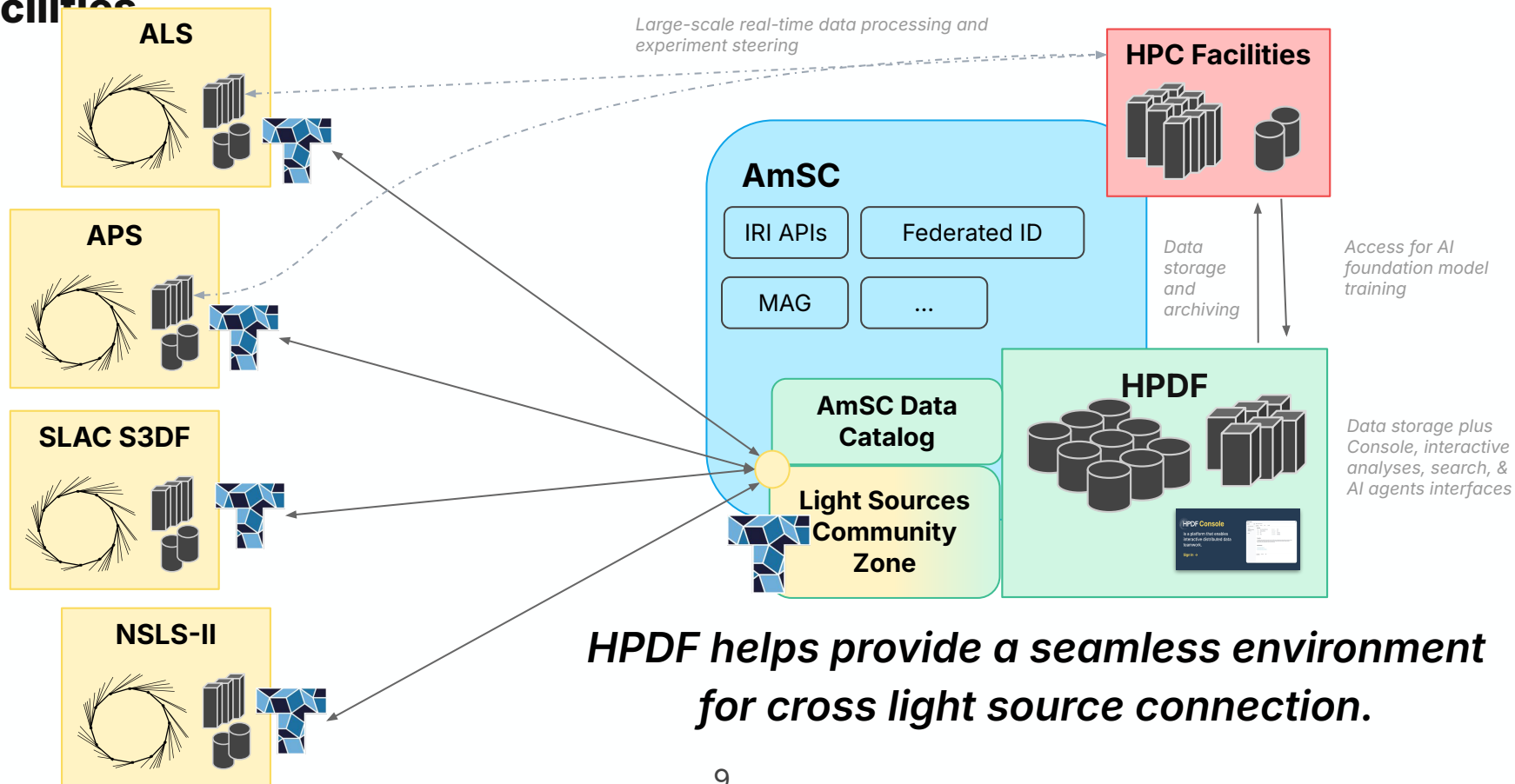
Varied experiences at user facilities make post-experiment work more challenging.

... but with HPDF as a partner, end user communities can access a **streamlined , federated environment** for data work filled with community managed resources



Partnership between light sources & HPDF can deliver a more seamless user experience.

... and key to a federated environment are interconnected user facilities



Value Proposition & Mission Delivery

Addressing the broader needs of scientists

Value Propositions for DOE Science

- Provide a **stable, long-term home** for (meta)data across the Office of Science
- Enable data discovery and use through improved federation, breaking down silos
[Even more pressing in Genesis Mission era]
- Provide **reliable, resilient, and highly available** infrastructure and services
- Reduce effort and complexity for teams delivering services and end users building workflows
- Provide resources for enabling increased data stewardship within and among DOE communities

These value propositions and the challenges they capture are both technical but just as importantly social and cultural!

Highlighted User Experience Lessons



SC communities require flexibility + reliable and streamlined resources

- Data services must offer nuanced access controls
 - “open” data has variable meanings across the data lifecycle
- Range of community software stacks & data expectations result in need for hybrid environments to leverage:
 - Facility services (staff, systems, etc.)
 - Community developed & managed solutions



Ongoing agentic AI user study is showing

- HPDF must store more than datasets. Users want key project context from teamwork (git issues & PRs, publications, etc.) to be accessible to their agents to enable effective work
- Users don't want simple, context free answers from agents—they want outputs they can easily gut check, refine, and reuse in later work.

User research and community engagements repeatedly affirm need for flexible but reliable, resilient, distributed data services.

Complementing Existing Facilities

Missions today are stretcheddddd...

Taking on growing expectations around data (services, stewardship, AI-ready transformation, etc.) can be burdensome alongside HPC+AI priorities.

How is HPDF partnering to complement existing Office of Science facilities and their resources?



Mission driver is specifically the wide variety of SC data and co-designing solutions to pressing problems

By design we are building for distributed, high-availability, & federated data use



This helps relieve mission burden of other facilities

NOT aiming to be the central location for model training, large-scale simulation or compute, or quantum

Activities providing support across the Data Lifecycle

Value propositions, mission, vision, & north star illustrate how HPDF will support scientists with solving their biggest data problems across five core activities.

1

Co-Design

Partnerships to address community needs by integrating core resources in community zones, co-developing new software, improving usability, and growing data management expertise.

2

Register

Ensure discoverability, interoperability, & reproducibility with xCard registration, a critical resource for the Genesis Mission.

3

Explore

Interactive data exploration tools to curate & carefully share data with colleagues, agents, and broader communities.

4

Analyze

Reliable, AI-enabled, collaborative analysis tools (e.g., Jupyter, shared notes, workflows) to abstract repeated tasks and automate their execution and monitoring.

5

Store

Resilient, distributed storage facilities spanning Jefferson Lab, Berkeley Lab, and beyond.

Five components to enable support the data lifecycle



Community Science Zones

Our hybrid architecture is built from the ground up to enable DOE communities to integrate their essential software tools and solutions with HPDF's resources.



CO-DESIGN

Partner with us to advance your community's needs. We are ready to support your science. Integrate your core science resources in a community zone. Co-develop new software to address challenges. Improve usability with user experience (UX) consulting. Leverage data management expertise to simplify work.



Reliable Registries

Store, register and ensure discovery of data, models, agents, and other types of scientific resources for cross-cutting DOE research.



REGISTER

Ensure discoverability, interoperability, and reproducibility by registering and storing cards describing your scientific products. HPDF's registry furnishes critical resources for the Genesis Mission.



Queryable Data

Seamlessly and reliably automate data ingestion to transform raw outputs into valuable, discoverable, and AI-ready resources.



EXPLORE

Probe new avenues of inquiry with interactive data exploration tools. Curate and carefully share your data with colleagues, agents, and broader communities. Create opportunities for groundbreaking discoveries.



Collaborative Environment

Easily collaborate across your community, generate novel artifacts, and derive and share new insights.



ANALYZE

Accelerate your team science with reliable, AI-enabled, collaborative analysis tools (e.g., Jupyter, shared notes, workflows). Simplify your work by abstracting repeated tasks and automate their execution and monitoring.



Interoperable Resources

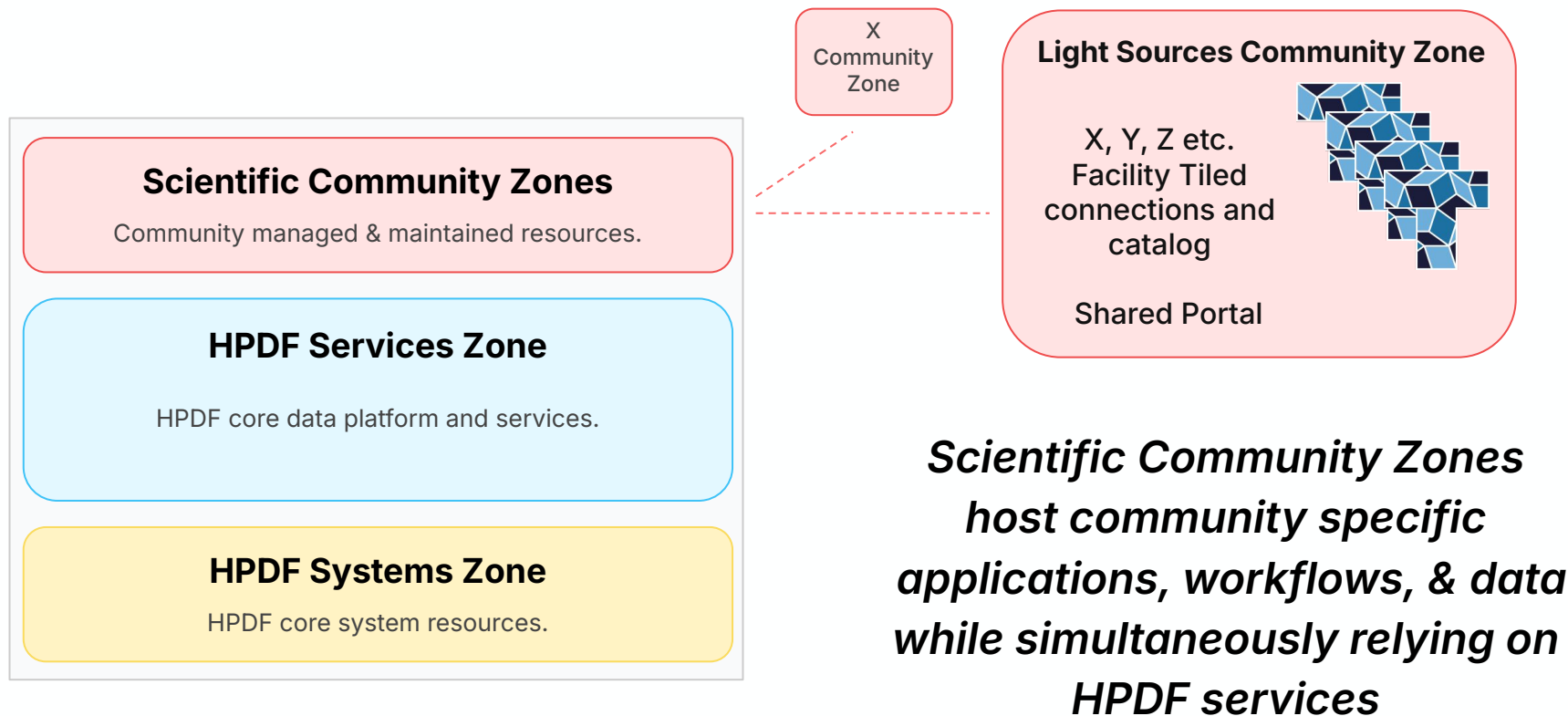
Leverage resources from the Genesis Mission platform to amplify your AI use and readily address cutting edge problems using a variety of data.



STORE

Expand your data capabilities with HPDF's resilient, distributed storage facilities capable of handling DOE's anticipated exabytes of data. Our storage and compute will span Jefferson Lab, Berkeley Lab, and beyond.

Example Component: HPDF Scientific Community Zones



HPDF as American Science Cloud Infrastructure Partner

Delivering for the Genesis Mission

HPDF in AmSC and GM



Community Science Zones

Our hybrid architecture is built from the ground up to enable DOE communities to integrate their essential software tools and solutions with HPDF's resources.



CO-DESIGN

Partner with us to advance your community's needs. We are ready to support your science. Integrate your core science resources in a community zone. Co-develop new software to address challenges. Improve usability with user experience (UX) consulting. Leverage data management expertise to simplify work.



Reliable Registries

Store, register and ensure discovery of data, models, agents, and other types of scientific resources for cross-cutting DOE research.



REGISTER

Ensure discoverability, interoperability, and reproducibility by registering and storing cards describing your scientific products. HPDF's registry furnishes critical resources for the Genesis Mission.



Queryable Data

Seamlessly and reliably automate data ingestion to transform raw outputs into valuable, discoverable, and AI-ready resources.



EXPLORE

Probe new avenues of inquiry with interactive data exploration tools. Curate and carefully share your data with colleagues, agents, and broader communities. Create opportunities for groundbreaking discoveries.



Collaborative Environment

Easily collaborate across your community, generate novel artifacts, and derive and share new insights.



ANALYZE

Accelerate your team science with reliable, AI-enabled, collaborative analysis tools (e.g., Jupyter, shared notes, workflows). Simplify your work by abstracting repeated tasks and automate their execution and monitoring.



Interoperable Resources

Leverage resources from the Genesis Mission platform to amplify your AI use and readily address cutting edge problems using a variety of data.



STORE

Expand your data capabilities with HPDF's resilient, distributed storage facilities capable of handling DOE's anticipated exabytes of data. Our storage and compute will span Jefferson Lab, Berkeley Lab, and beyond.

Use Case: Data Lakehouse Proof of Concept

1 Catalog

Registered metadata
from existing catalogs
in OpenMetadata

2 Transform & Register

Transformed those datasets
into Iceberg tables

Stored those table
representations

Registered them in AWS Glue
as a data lakehouse catalog

3 Query & Analyze

Connected with MCP and
skills

Put a conversational interface
on top of them

I need data



*I can do
science*

Use Case: ModCon Hackathon Co-Design

CHALLENGE

There is a class of scientific data that remains on individual laptops or in ad hoc storage, but for which the path into today's larger data infrastructure can be too difficult.

ModCon Hackathon



Data on laptop

Difficult to collaborate, query across, and make available to tools and agents



Data in lakehouse

Readily explorable and queryable

HPDF is helping close this gap

Register: The AmSC Data Catalog

Build on data and metadata that already exists

Co-design Make it easier for AmSC infrastructure providers to register metadata from their own catalogs

Reuse Leverage existing metadata to avoid manual recreation

Connect Build interoperable paths that allow metadata to be discoverable across the broader ecosystem

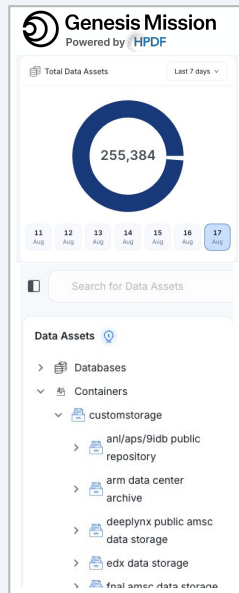
This gives scientists a broader discovery surface across resources that previously had to be searched independently

AmSC Data Catalog

255,000+
data assets

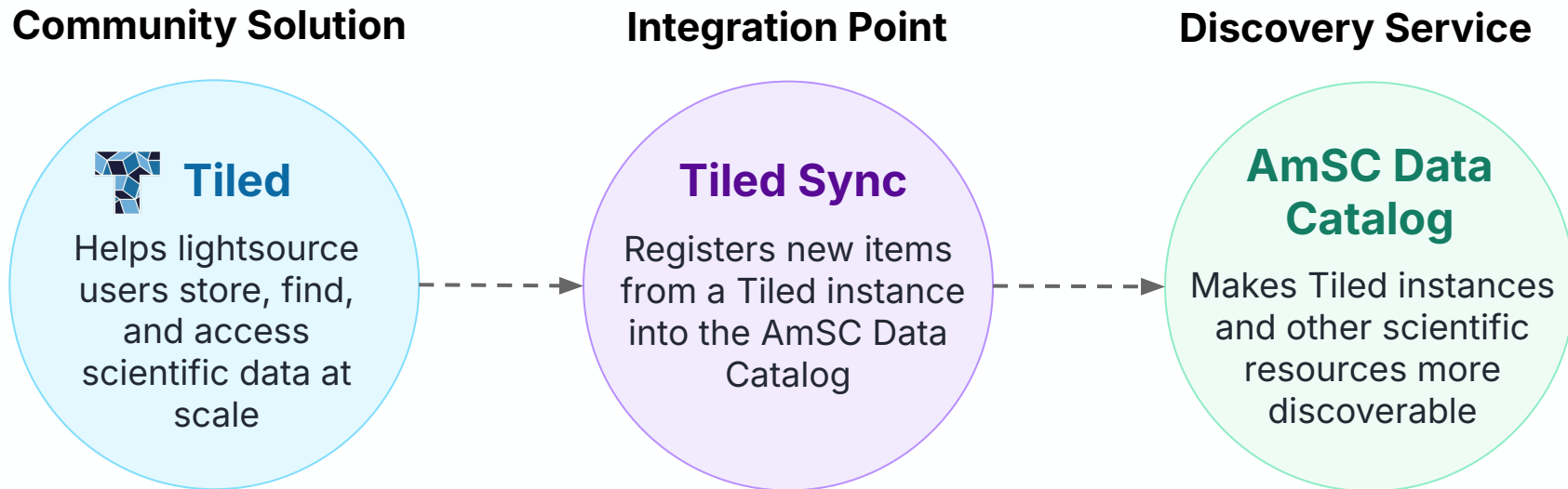
10
scientific data
catalogs

3
lakehouse
data catalogs



Co-design Example: Tiled Connector to AmSC

Integration with existing community workflows



HPDF will work with community users to help make scientific resources more discoverable without forcing every community into the same workflow

Reliable Registries: xCards

xCards

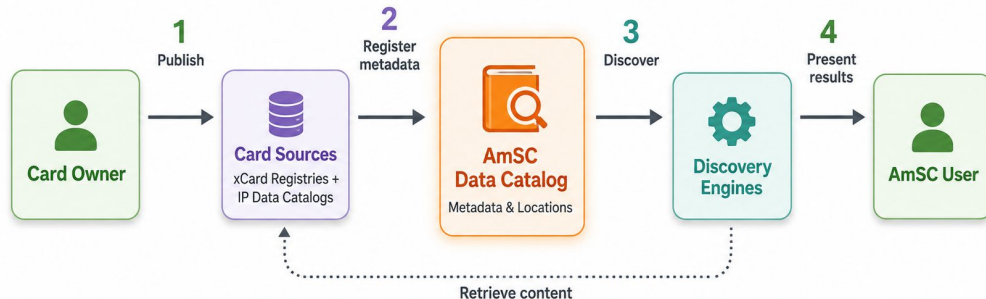
Standard representations for data, models, and agents

xCard Registry

A service for hosting and managing xCards.

HPDF is building this as a reference implementation that others can use.

xCard Discovery Flow



HPDF's role is to help create reliable infrastructure around which different parts of the Office of Science ecosystem can interoperate.

Lessons reinforced by our AmSC work

Scientists need reliable data infrastructure

Multiple scales:

- Research groups trying to get important datasets off laptops
- Large scientific facilities and infrastructure providers trying to make thousands of data assets discoverable

Co-design improves what we build

- Science teams tell us where the friction actually is
- Infrastructure providers tell us what interoperability requires in practice
- AmSC + Genesis let us test ideas against real systems

Deliver useful capabilities early

- We can deliver value before the full HPDF infrastructure is in place

HPDF's Roadmap for Continued Growth

Meeting Genesis Mission Demand

Facility Progression

A phased approach to deliver increasingly capable operational systems

EEAS

Early Early Access System



Cloud-based, production
quality capabilities

EAS

Early Access System



Hybrid environment connecting
the HPDF sites and cloud

Production HPDF



Resilient, integrated
data infrastructure

Closing

Closing

1

Meeting core needs

Team is designing & building the reliable, resilient data services & storage urgently needed by Office of Science (SC) communities.

2

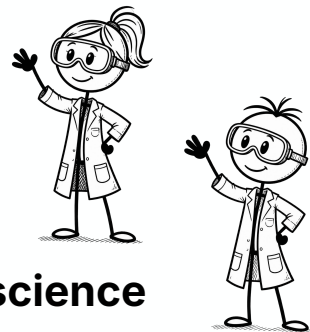
Accelerating delivery

American Science Cloud & Genesis Mission are enabling rapid acceleration of our work on key resources for the SC community.

3

Enabling science

Focusing on delivering resources to help SC teams solve their biggest data problems while they focus on the science that matters most to them.





Questions?

Thank you!

Karlo Berket, kberket@lbl.gov

Drew Paine, paided@lbl.gov



Want to talk more?
Reach out today.