Expandable Computing Through the OSPool

Christina Koch, on behalf of the OSG Research Facilitation Team
SGX3 Webinar
March 21, 2024

What is your goal in attending this presentation?

Comment in the zoom chat:

Name
Institution
and…

a) Do you have a specific idea gateway idea or plan that you're ready to implement?
b) Are you curious about the OSPool and other OSG services?
c) Something else? (please comment :))

Who are we?

The OSG Consortium provides computational capacity and services for researchers, faculty, staff, and students at academic, government, and non-profit institutions.

What does this mean for *you*?

We provide a place (OSPool) for computations and help you with doing so!

Highlights to Consider

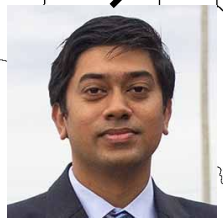
- Why use the OSPool?
  - The OSPool can be a **computational backend** for science gateways
  - The OSPool is a good fit for communities who can express their big computational problem as a **throughput computing** workload.
- Who can use it?
  - A gateway **operator** must be affiliated with a **US-academic, government, or non-profit institute**.
  - Gateway **users can be anywhere**
- What does it cost?
  - All capacity and services are open - **no proposal, no allocation, no fee**.

The OSG Research Facilitation Team

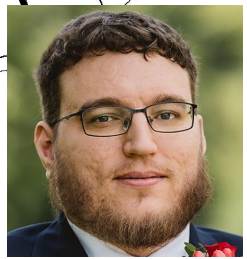
Christina Koch

Rachel Lombardi

Mats Rynge



Showmic Islam



Andrew Owen

See the rest of the OSG Team

Community Opportunities

Throughput Computing 2024 (July 8-12)

Annual high-throughput computing community conference

https://osg-htc.org/events/throughput-computing-2024/

OSG School 2024 (August 5 - 9)

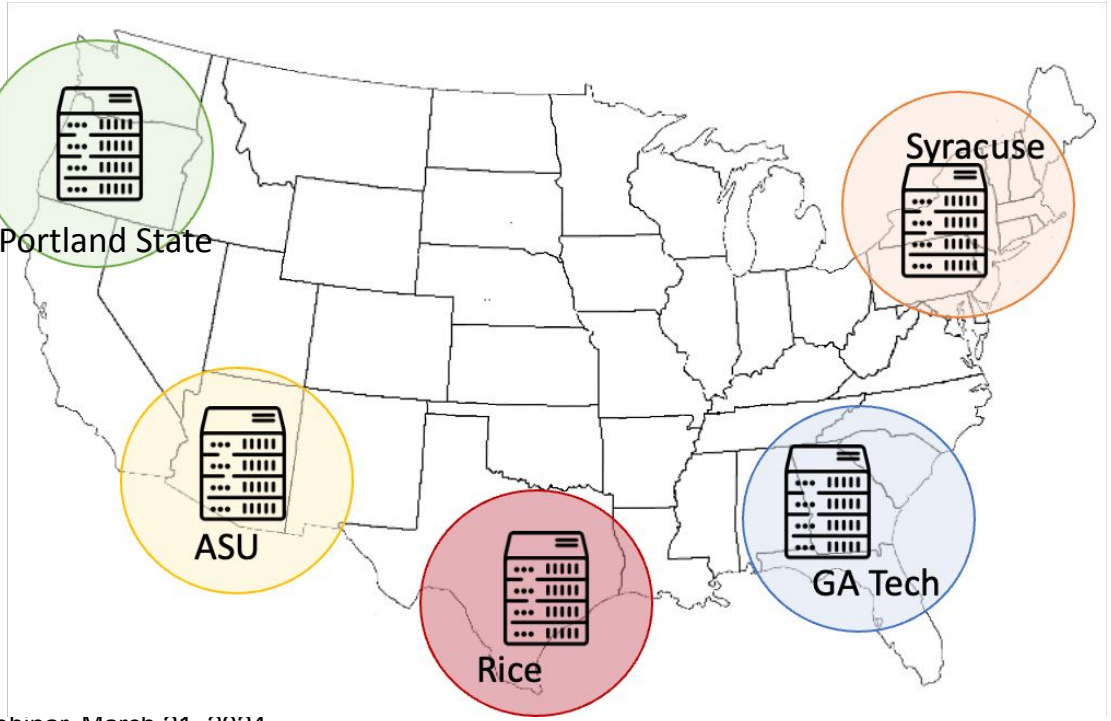
Weeklong summer school
Applications open now!

https://osg-htc.org/school-2024/

So…what IS the OSPool?

Local Computing

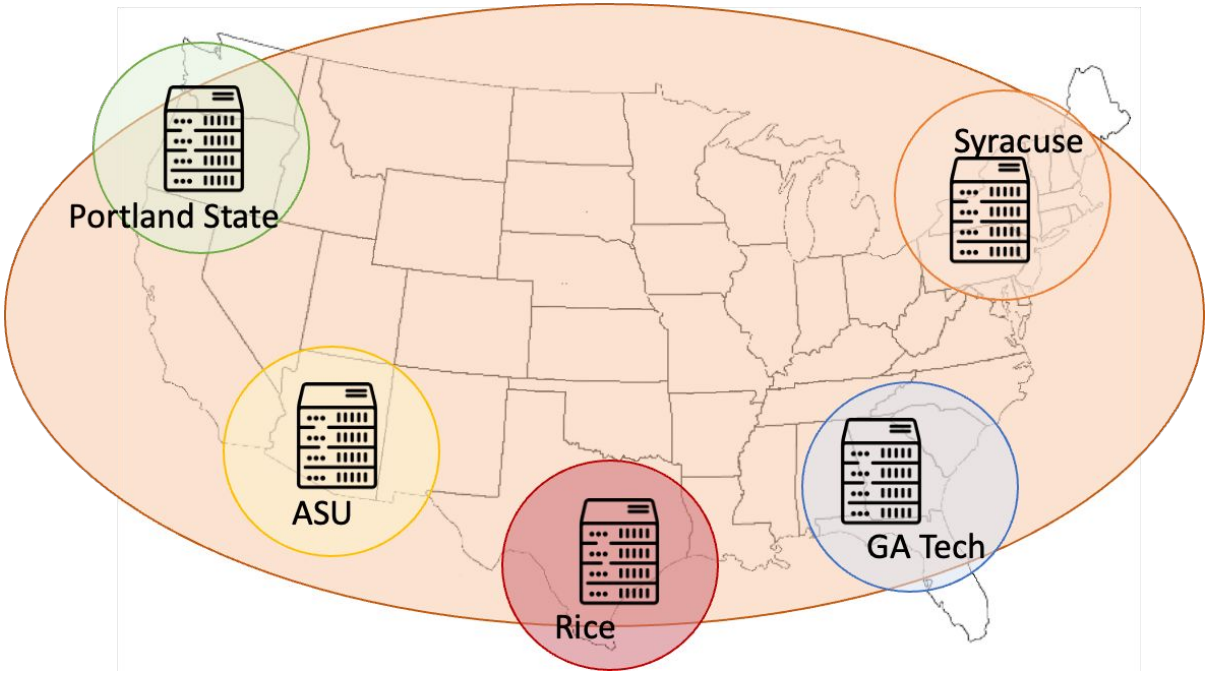
Often, if a campus wants to provide more computing to its researchers, it will invest in a local cluster.



Sharing Capacity

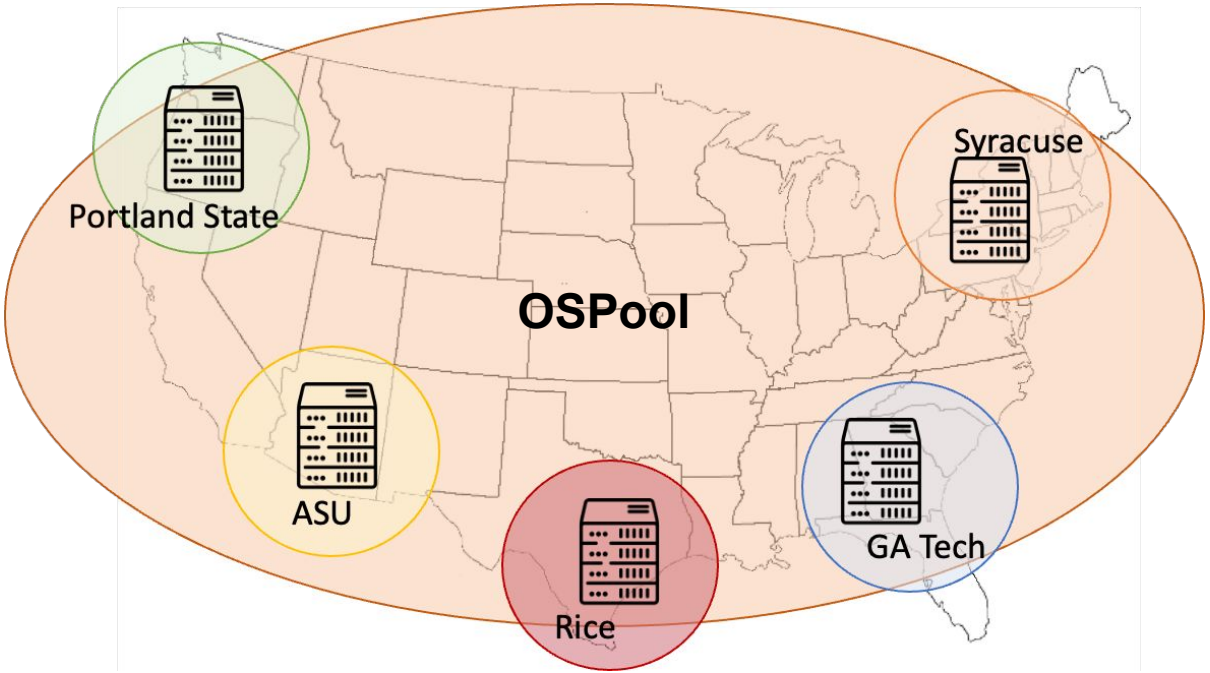
But what if institutions pooled their “extra” resources together in a virtual cluster?

(and provided access to researchers without a local cluster?)

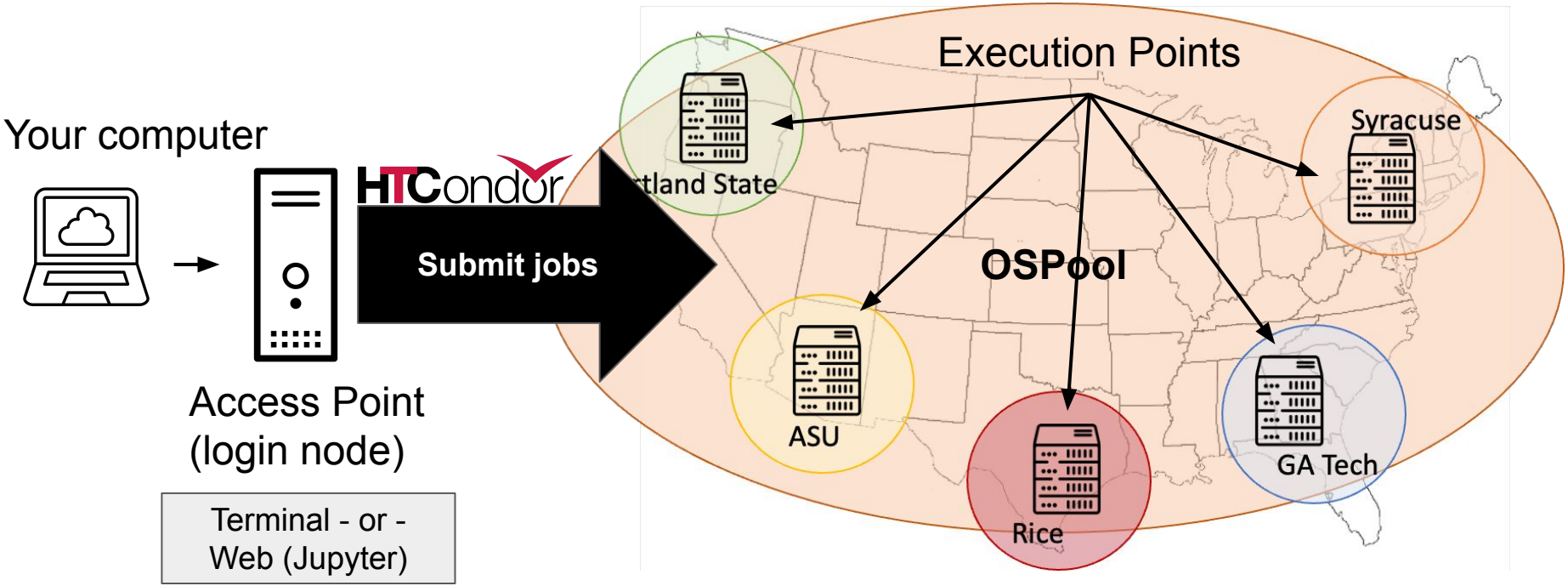


The OSPool

This is exactly the idea behind the OSPool.



Running work



OSG Services for US Researchers

Open Science Pool (OSPool):

Open computing capacity for high *throughput* workloads

OSPool Access Points:

OSG-operated service to submit jobs to the OSPool

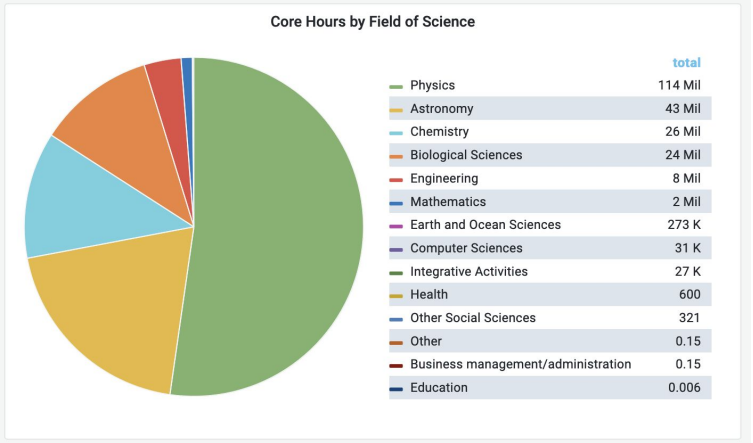
Open Science Data Federation (OSDF):

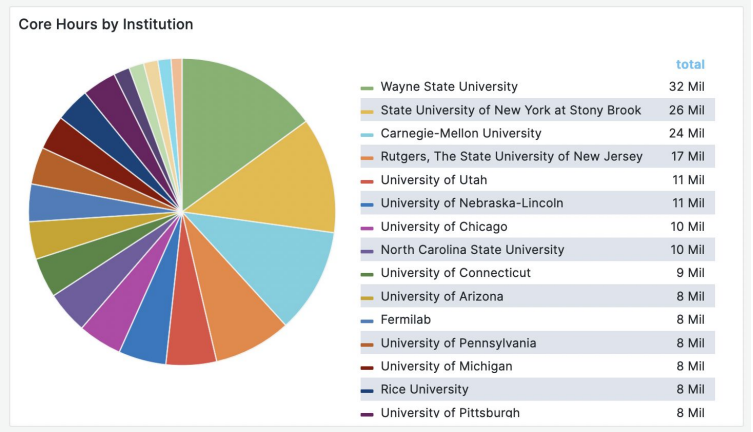
Network of data origins (file servers) and caches for data accessibility

https://display.osg-htc.org/

OSPool: The Last 365 Days

Total Jobs	~150 million
Active Users	220
Active Groups	140
File transfers	~3.3 billion

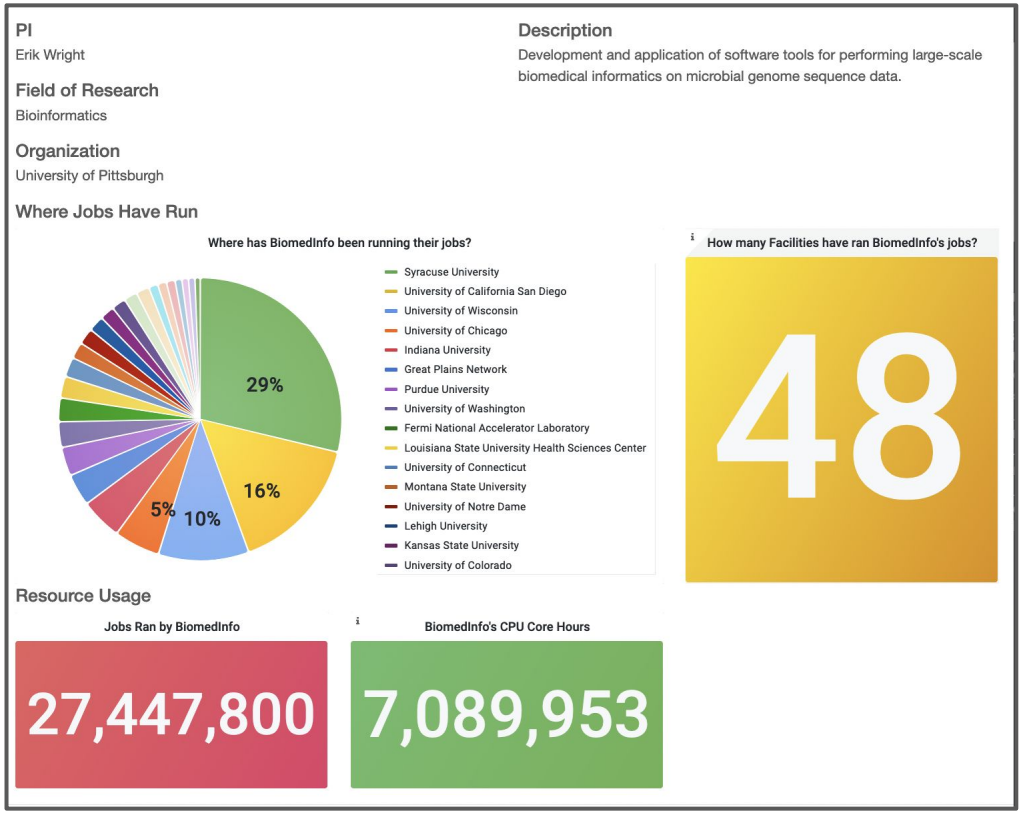




Single Project Usage

This is a project based at the University of Pittsburgh, usage from the past year.

Screenshot from: https://osg-htc.org/projects?project=BiomedInfo



By combining the power of the OSPool (and relevant OSG services), and science gateways, this scale of computing can become more accessible to all.

What can I do on the OSPool?

High Throughput Computing (HTC)

One of our favorite HTC analogies: baking the world's largest/longest cake

In computational terms: solving a big problem (the world's longest cake) by executing many small, self-contained tasks (individual cakes) and joining them.

Photos: Arun Sankar via https://www.theguardian.com/world/2020/jan/16/indian-bakers-rise-to-task-of-making-worlds-longest-cake

What workloads are good for the OSPool*?

	Ideal Jobs! (up to 10,000 cores across Jobs, per user!)	Still Very Advantageous!	Less-so, but maybe
Cores (GPUs)	**1** (1; non-specific type)	**<8** (1; specific GPU type)	**>8 (or MPI)** (multiple)
Walltime	**<10 hrs*** *or checkpointable	**<20 hrs*** *or checkpointable	**>20 hrs**
RAM	**<few GB**	**<10s GB**	**>10s GB**
Input	**<500 MB**	**<10 GB**	**>10 GB**
Output	**<1 GB**	**<10 GB**	**>10 GB**

* the “less-so, but maybe” column could still be an HTC workload, but one that would run more effectively on a local, dedicated HTC system instead of the OSPool

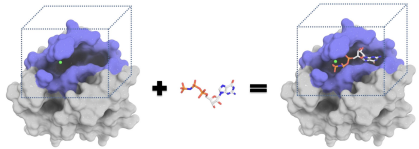
Examples: Processing Many Data Files

Potential domains:

- Bioinformatics
- Medical images
- Audio files
- Mapping data
- Hyperspectral images
- Protein docking
- Natural language processing

(and more)

Rousselene Larson
Utah State University
Identification of plant-based natural compounds that interact with multiple myeloma; drug development



Number of OSPool jobs:
1.7 million
40,453,470 total docking runs in one month!

Presentation from Throughput Computing 2023

Examples: Many Simulations

Potential domains:

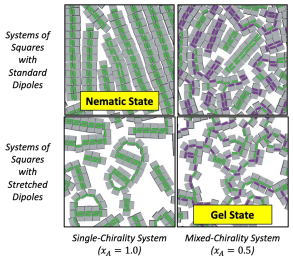
- Statistics
- Economics
- Physics
- Mechanical Engineering
- Chemistry
- Materials Science
- Earth/Ocean/Atmospheric Sciences

(and more)

Matthew Dorsey

North Carolina State University

Design new types of magnetic materials from colloidal building-blocks.

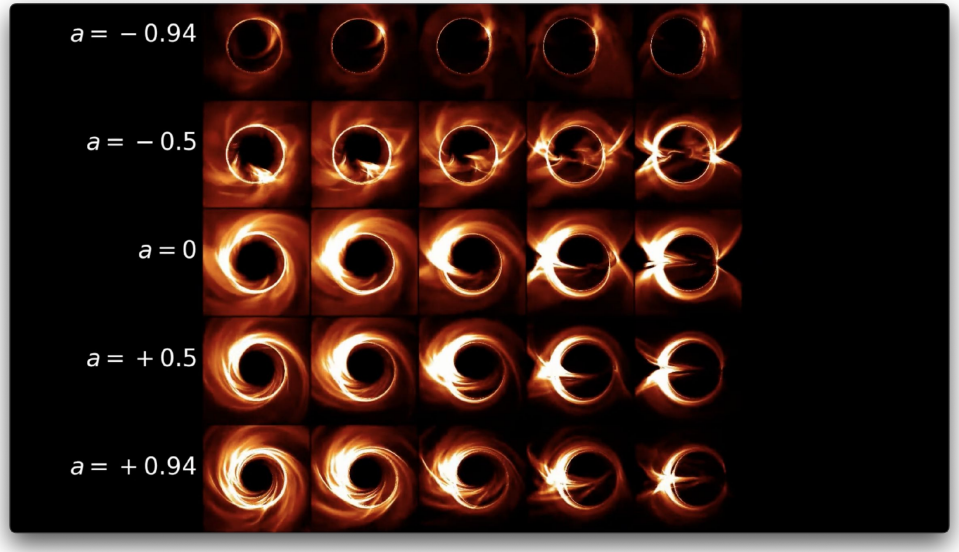


Number of OSPool jobs in last year:

6.5 million

Sample Gateways Using the OSPool

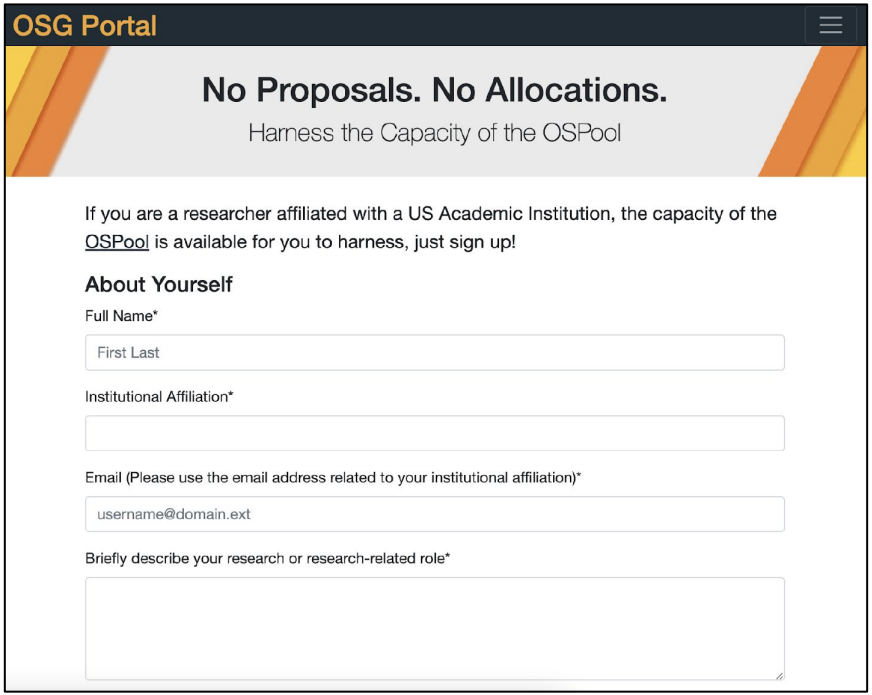
- Event Horizon Telescope
  - https://eht.scigap.org/
  - Can run many independent simulations at once
  - Still under development
- CLAS12
  - Internal gateway for submitting thousands – millions of event simulations.



EHT simulations, image courtesy of CK Chan
Full Video

Access OSPool Capacity

Request an Account



https://portal.osg-htc.org/application

Orientation Meeting

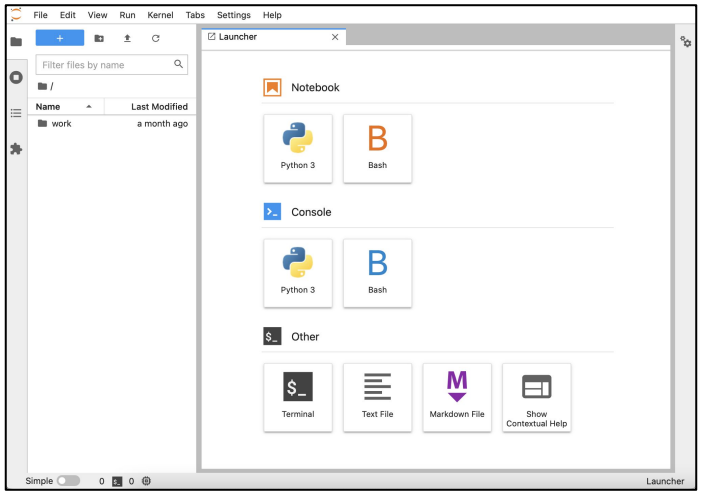
- Discuss science and computation details, identify goals and motivation.
- Communicate about HTC approaches / OSPool capacity (New User Slides)
- Provide specifics for getting started (Roadmap)

End with ongoing support options and account creation details.

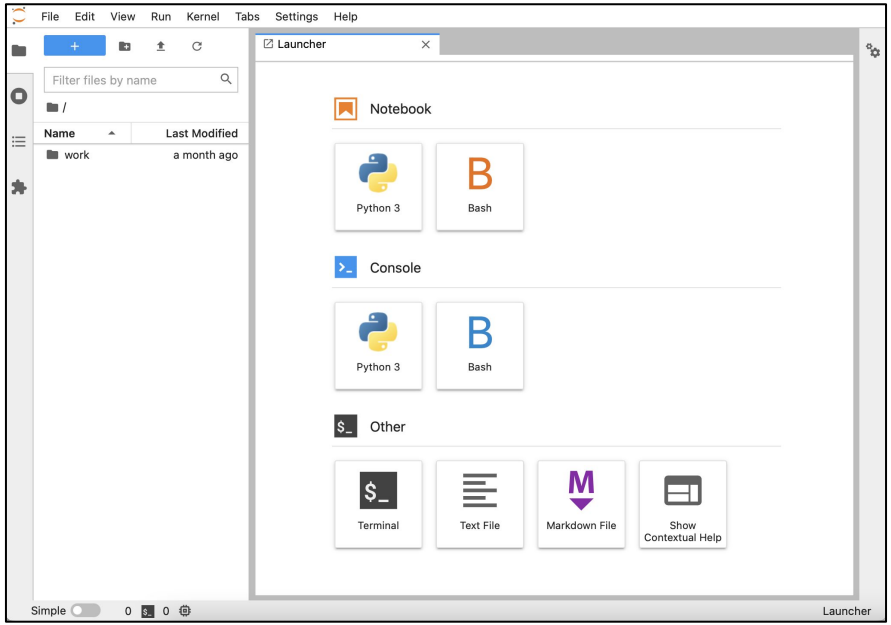
Get Started Right Away

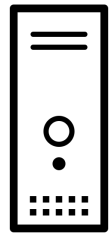
OSPool Notebooks: Jupyter-based interface to an HTCondor Access Point

https://notebook.ospool.osg-htc.org



OSPool Notebooks (Guest Version)



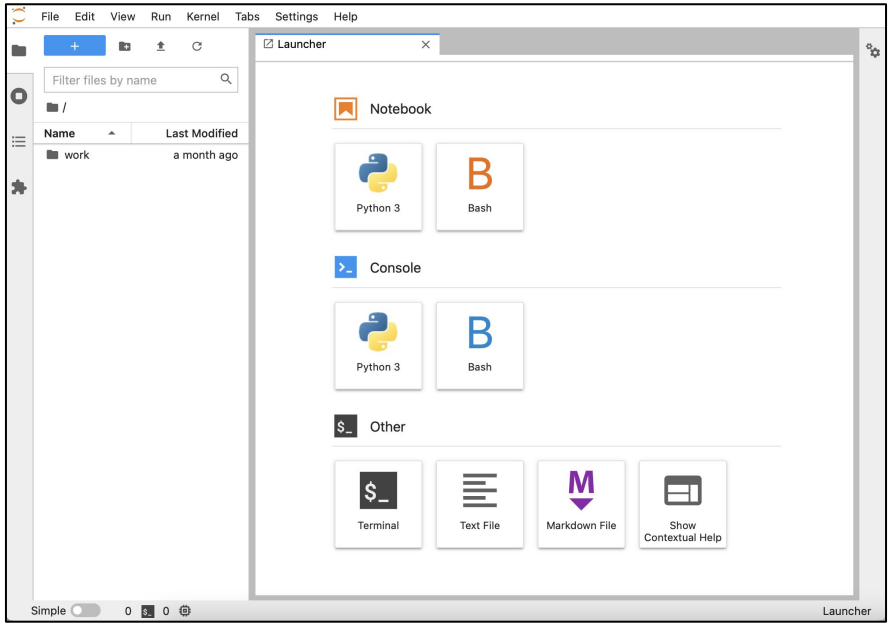


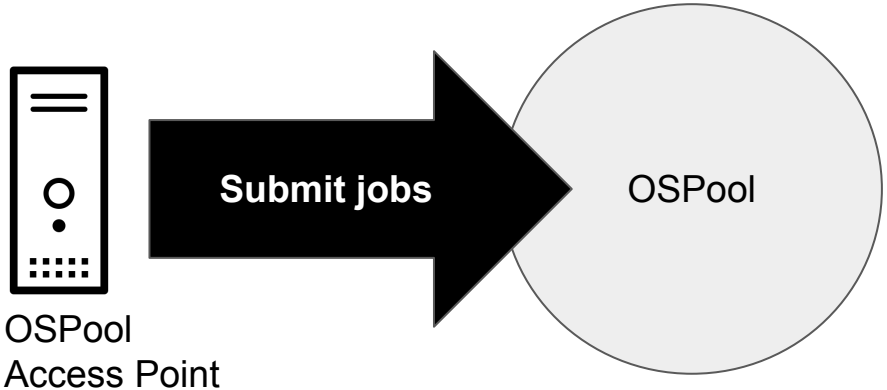
Temporary Access Point

Submit jobs

Temporary Test Pool

OSPool Notebooks (Full Account Version)

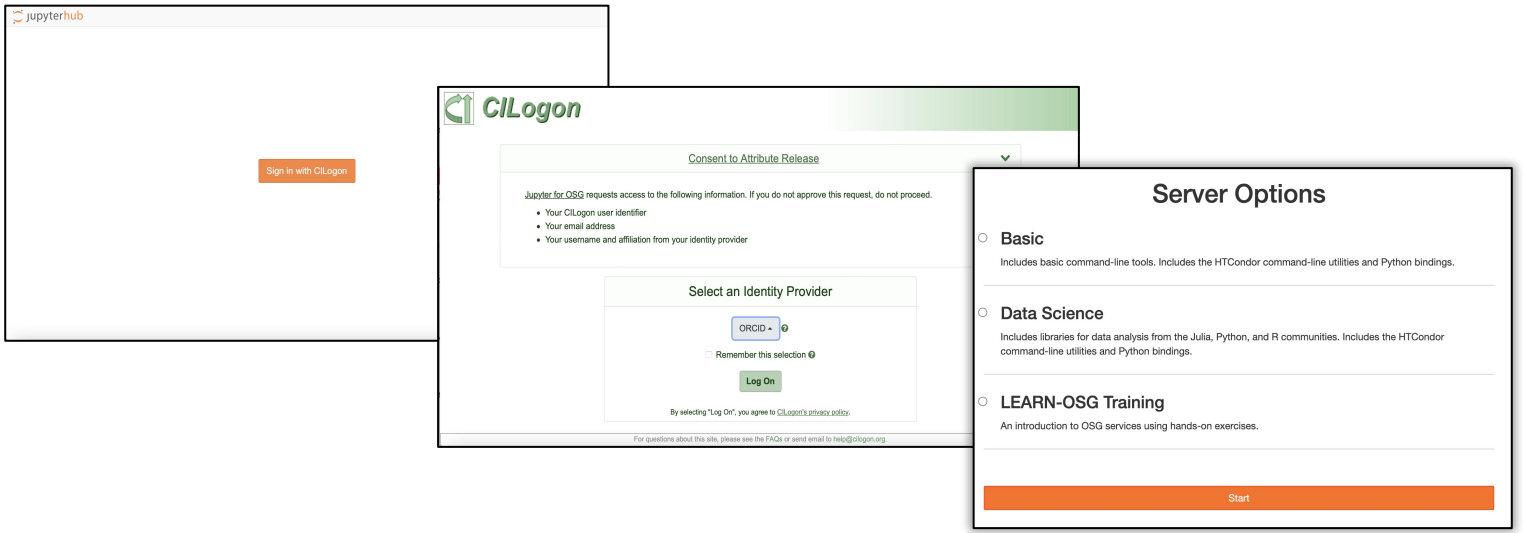




Sample Workflow

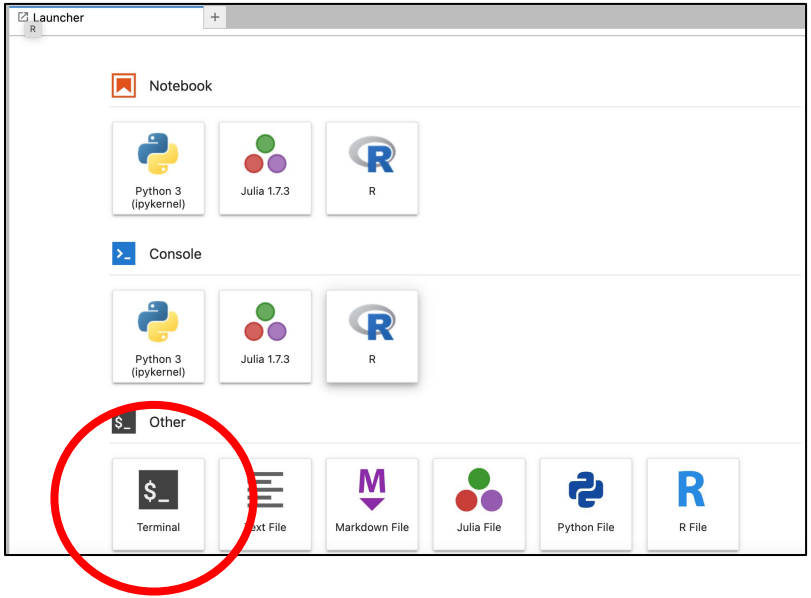
Get Connected

- Go to **https://notebook.ospool.osg-htc.org**
- “Log in” and choose the basic server option



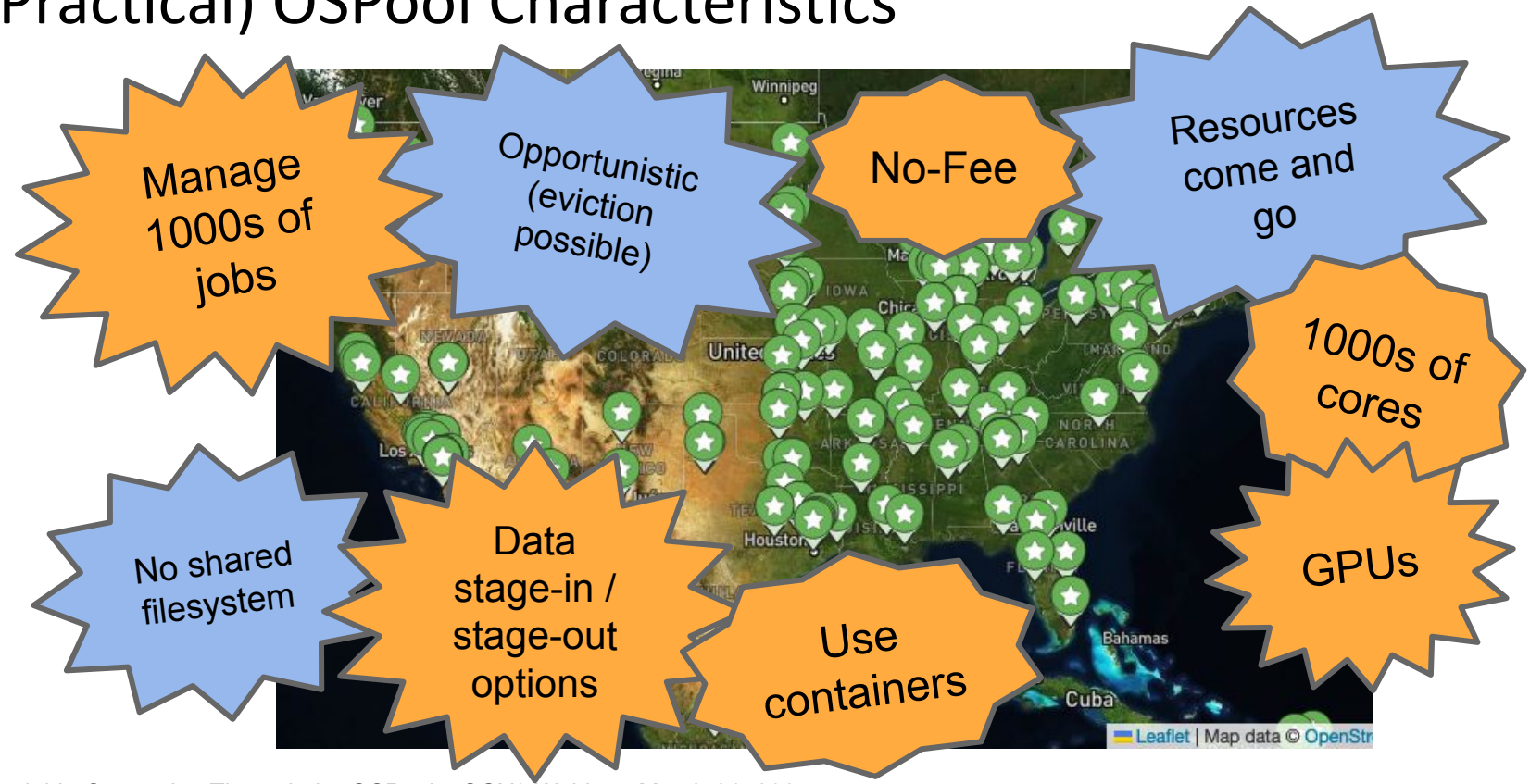
Get the Materials

- Open a terminal
- Run the command:
  `tutorial AutoDockVina`
- Open the README.ipynb file



Services and Features

(Practical) OSPool Characteristics



HTCondor for Workload Management

- Submit multiple jobs (and organize their inputs/outputs)
- Automate retries, catch errors
- Checkpoint long-running tasks
- Run multi-step workflows with tools like DAGMan and Pegasus
- Submit jobs using a command line or Python API

Software Portability

Anything* that can be run from the command line and on Linux should be runnable on the OSPool.

Container support via native Singularity (apptainer) images or Docker images converted to Singularity.

* potential exceptions: restrictively licensed code, other edge cases

Data at Scale

What's available

HTCondor has a suite of **file transfer plugins** that can fetch input and place output for jobs in many ways: http, s3, Google Drive, Box…

(And you can write your own, if you want!)

The **Open Science Data Federation** is an OSG service that caches data across a national network to make it more accessible to distributed computing resources.

Implications for Gateways

Have many different options for fetching/moving gateway job data:

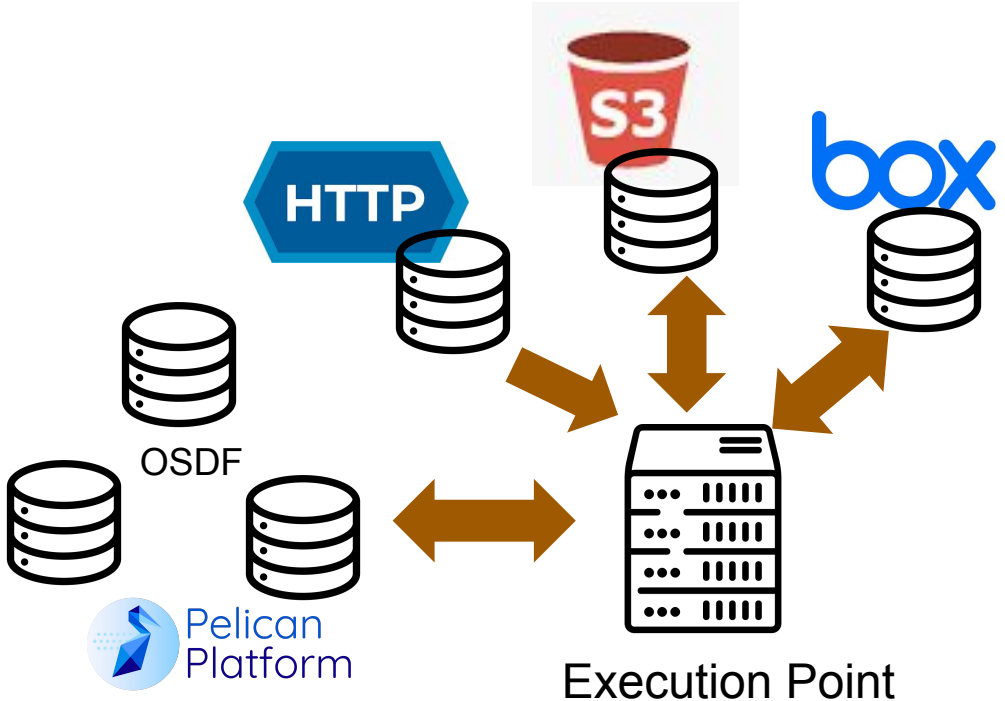
- Web server (hosted by the gateway, or elsewhere)
- Data with an S3 interface (Amazon, but also self-hosted data services)
- Use an OSDF Data Origin

Data at Scale

HTCondor has a suite of **file transfer plugins** that can fetch input and place output from different sources.

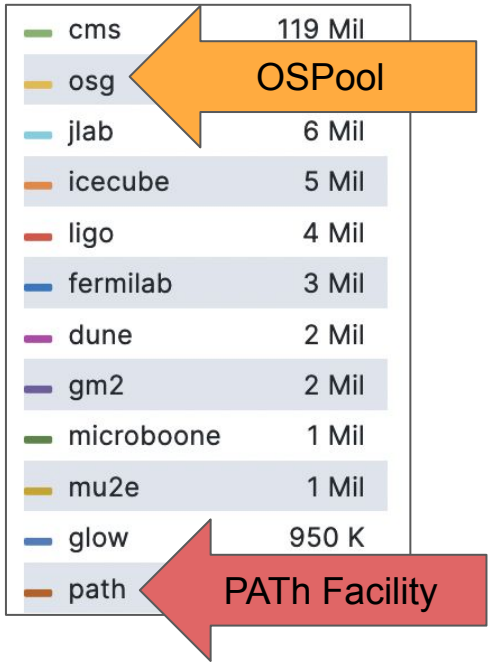
The **Open Science Data Federation (OSDF)** is an OSG service for data delivery.

The underlying technology, **Pelican** (https://pelicanplatform.org/), will make it possible to easily federate data into the OSDF.



Capacity Beyond the OSPool

HTC Pools



The PATh Facility: a dedicated high throughput pool. (https://path-cc.io/facility/).

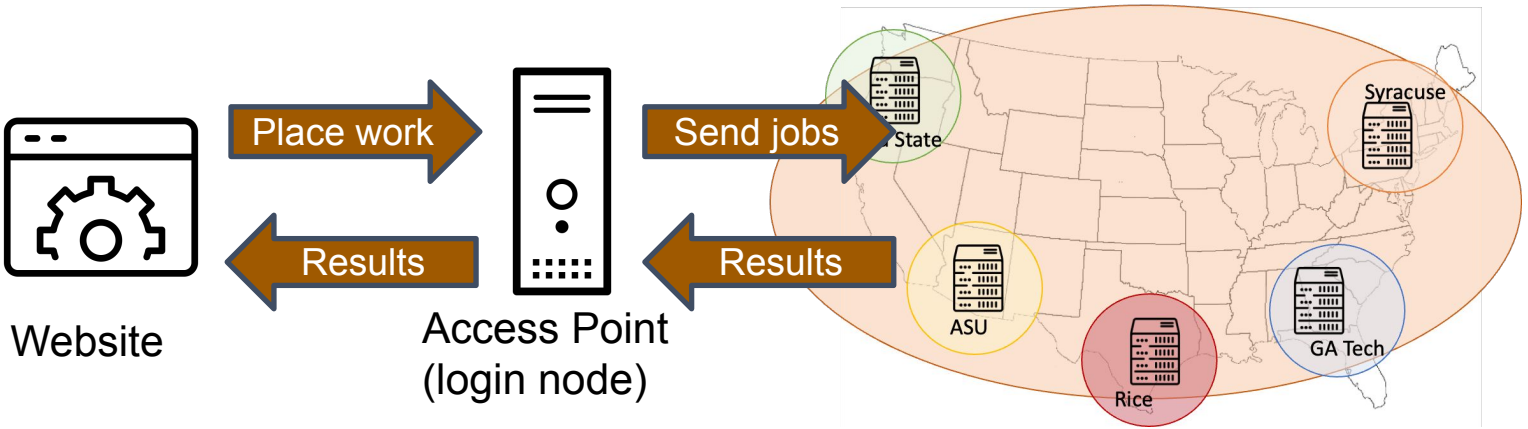
If you have a high throughput workload, **and…**

- You need more cores or memory per job than on the OSPool.
- Your data is larger than normally recommended for the OSPool/OSDF.
- You want a guaranteed amount of time for jobs to run.
- You have NSF funding.

Talk to us! support@path-cc.io

Gateway Setup Details

- Set up an Access Point or get a service account
- Run your own front end / web stack



Take Advantage of OSPool Services and Our Community

Throughput Computing 2024 (HTC24)

Annual HTC community conference, Madison, WI, 2024 July 8 - 12.

https://osg-htc.org/events/throughput-computing-2024/

Last year's schedule: https://path-cc.io/htc23

OSG School

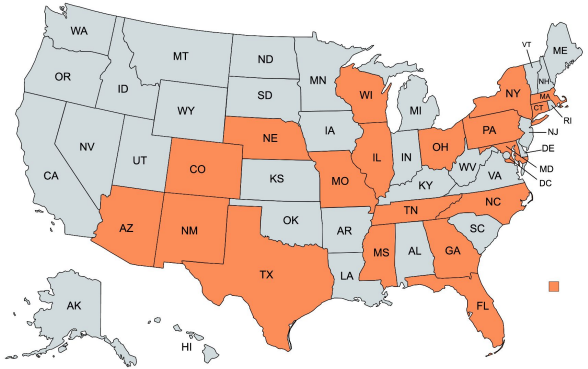
Annual, week-long summer school where researchers (and research staff!) learn how HTC systems work and how to apply tools like the OSPool to their *own research or environment.*

OSG School 2024 - Applications Open NOW!

https://osg-htc.org/school-2024/

In 2023, 58 participants originating from 19 states, 4 countries (Mali, Uganda, Australia, USA), and 30+ institutions attended.

Travel/hotel/attendance expenses covered



Above: ACE students with CHTC staff from the 2023 User School
Left: Institutional locations for 2023 School participants

Knowledge Sharing

OSG Campus Meet-Up

- Weekly session for campus staff who are contributing to or using the OSPool.
  - 1st/3rd weeks: planned topic
  - All other weeks: open discussion
- Would be an appropriate venue to discuss gateway questions!

OSPool User Office Hours

- Drop-in virtual help sessions twice-a-week, every week for OSPool users.

Additional Community Resources

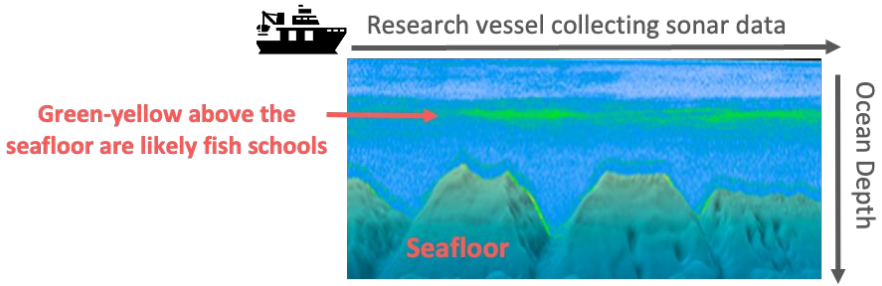
Synchronous

- Monthly trainings
- OSG School
- Throughput Computing

Asynchronous

- Documentation
  - https://portal.osg-htc.org/documentation/
- Tutorials
- OSPool Notebooks
  - https://notebook.ospool.osg-htc.org
- Recordings of events and trainings
  - 160 videos uploaded on OSG- affiliated YouTube channels
  - Partnership to Advance Throughput Computing

The Goal: Transforming Research



Slide content courtesy of Carrie Wall Bell and Brian Bockelman

The NOAA National Centers for Environmental Information stewards over **270 TB** of fisheries sonar data and makes them publicly accessible through a dedicated data portal and AWS.

Raw sonar files are complex and difficult to work with in the cloud. **To improve AI-readiness and efficiently apply scaled processing, we need to translate the sonar files to Zarr stores** - a crucial first step in building an optimized data lake and the foundation for applying deep learning (NSF Award # 2311843).

Using OSPool, a CU software developer adapted his local workflow and processed **55 research cruises conducted between 2007 and 2022 comprising over 100,000 files in 11 hours at no cost**!

- Processing this volume of data would have taken over 30 weeks on a local desktop and over 13 hours using parallelized lambas on AWS
- OSPool also gave us **the space to fail, adjust, and proceed**

Join Us!

Get an account:

https://portal.osg-htc.org/application

Chat with us:

- OSPool User Office Hours:
  - Tues 4-5:30pm ET / Thurs 11:30am - 1pm ET
  - Zoom Link: https://osg-htc.org/OfficeHoursZoom
- OSG Campus Meet-Up
  - Wednesdays 12-1pm ET
  - Zoom Link: Link: https://osg-htc.org/campus-meet-up-zoom
- (Or email support@osg-htc.org)

Join events and training:

- Throughput Computing 2024:
  - https://osg-htc.org/events/throughput-computing-2024/
- OSG School 2024:
  - https://osg-htc.org/school-2024/
- Monthly Training:
  - https://portal.osg-htc.org/documentation/support_and_training/training/osgusertraining

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2030508. 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.