State of OSG

Frank Würthwein
OSG Executive Director
UCSD/SDSC

March 10th 2023

OSG “Statement of Purpose”

OSG is a consortium dedicated to the advancement of all of open science via the practice of distributed High Throughput Computing (dHTC), and the advancement of its state of the art.

Four categories of participants in the OSG Consortium

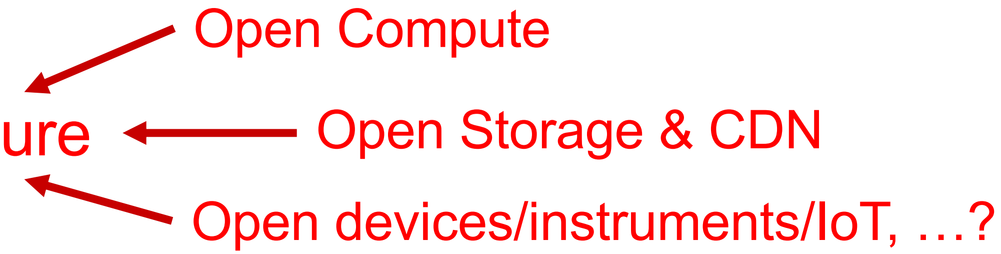
- The individual researchers and small groups through the Open Science Pool.
  - 91M jobs consuming 191M core hours within the last year
- The campus Research Support Organizations
  - Teach IT/CI organizations & support services so they can integrate with OSG
  - Train the Trainers (to support their researchers)
- Multi-institutional Science Teams
  - XENON, GlueX, SPT, Simons, and many many more
  - Collaborations between multiple campuses
- The 4 “big science” projects:
  - US-ATLAS, US-CMS, LIGO, IceCube

OSG Vision & Aspiration

Long Term Vision

- Create an Open National Cyberinfrastructure that allows the federation of CI at all ~4,000 accredited, degree granting higher education institutions, non-profit research institutions, and national laboratories.

  - Open Science
  - Open Data
  - Open Source
  - Open Infrastructure



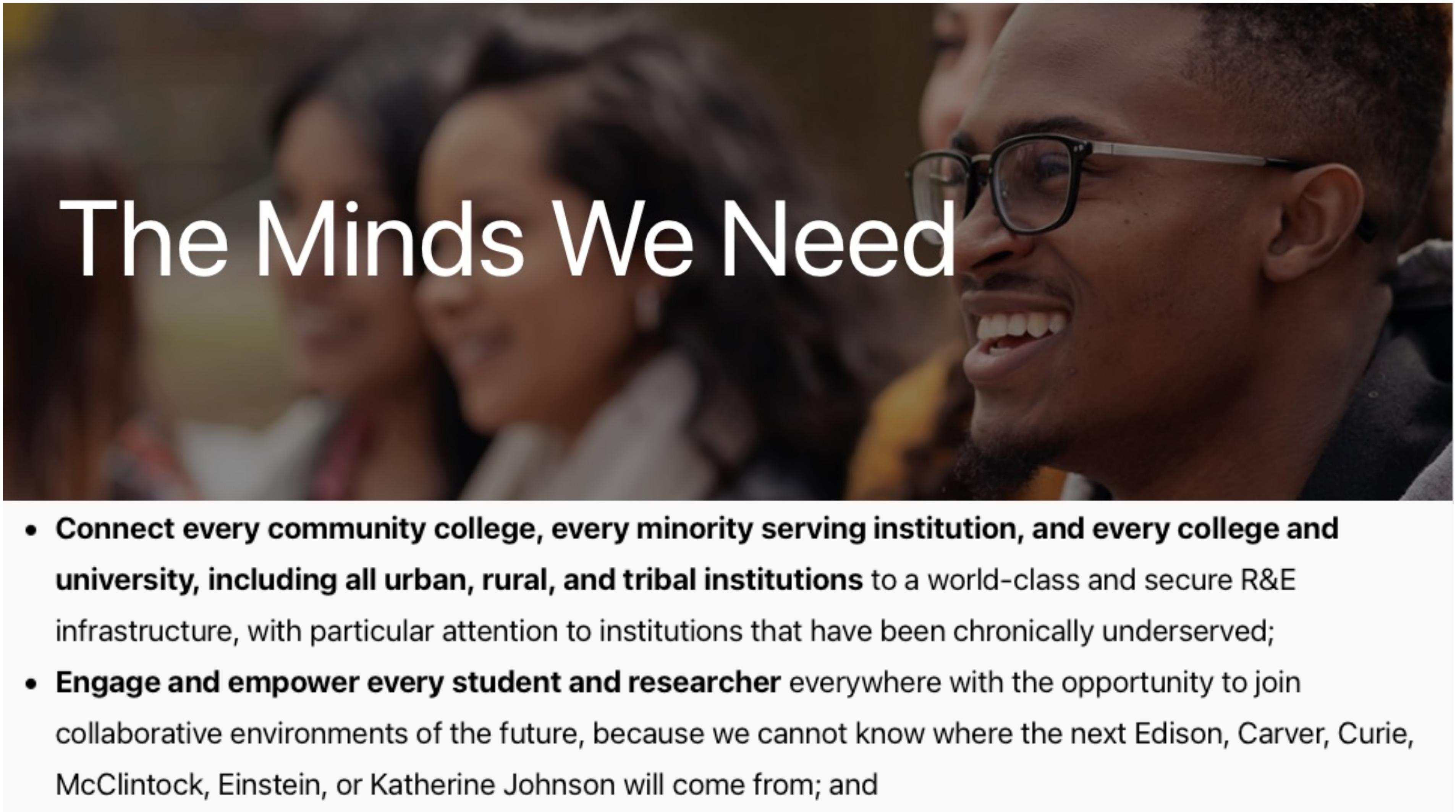
Open Compute

Open Storage & CDN

Open devices/instruments/IoT, …?

Openness for an Open Society

Democratizing Access



The Minds We Need

- **Connect every community college, every minority serving institution, and every college and university, including all urban, rural, and tribal institutions** to a world-class and secure R&E infrastructure, with particular attention to institutions that have been chronically underserved;
- **Engage and empower every student and researcher** everywhere with the opportunity to join collaborative environments of the future, because we cannot know where the next Edison, Carver, Curie, McClintock, Einstein, or Katherine Johnson will come from; and

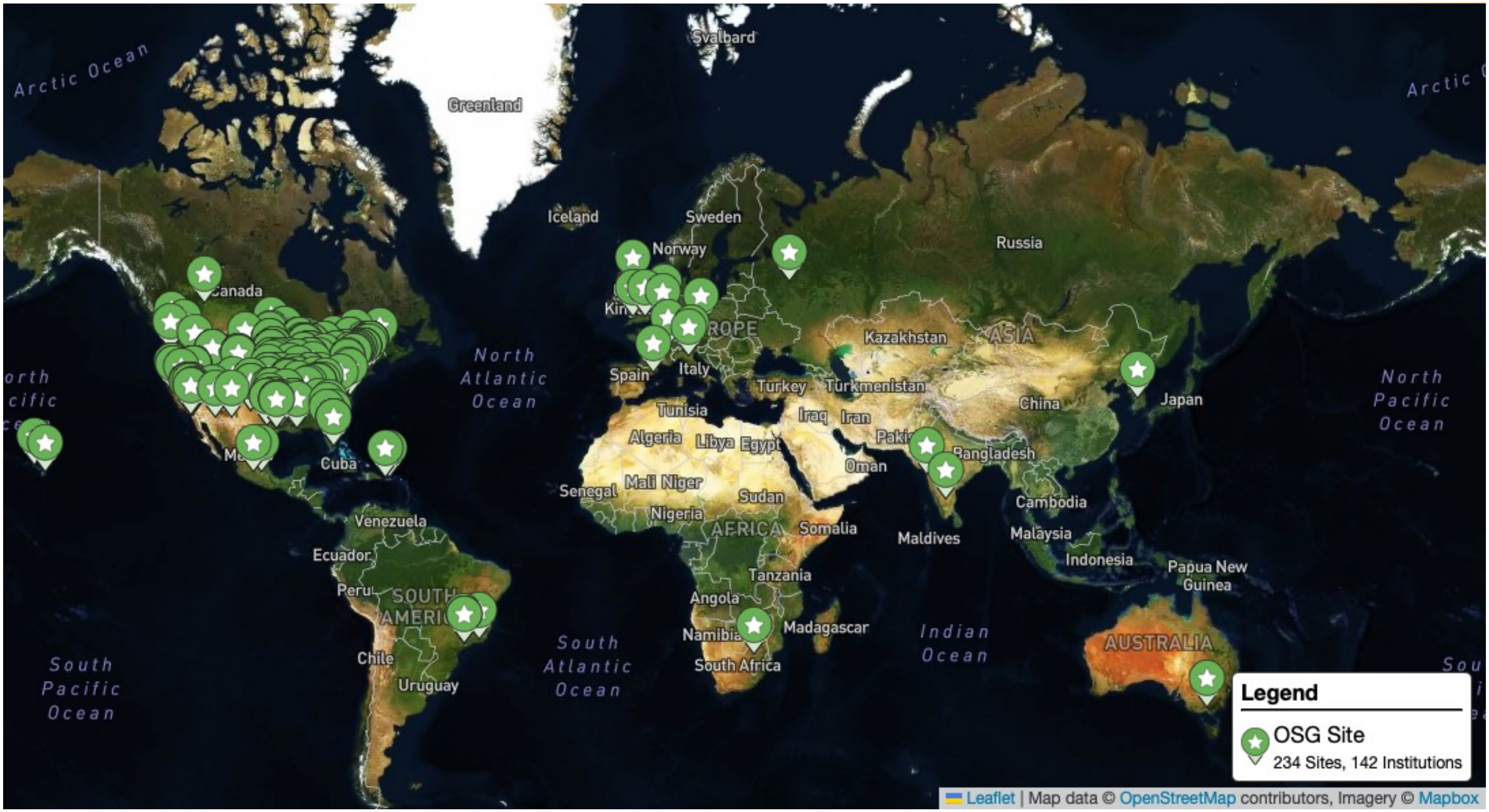
https://mindsweneed.org

OSG

UCSD

OSG Distributed Computing viewed as Maps

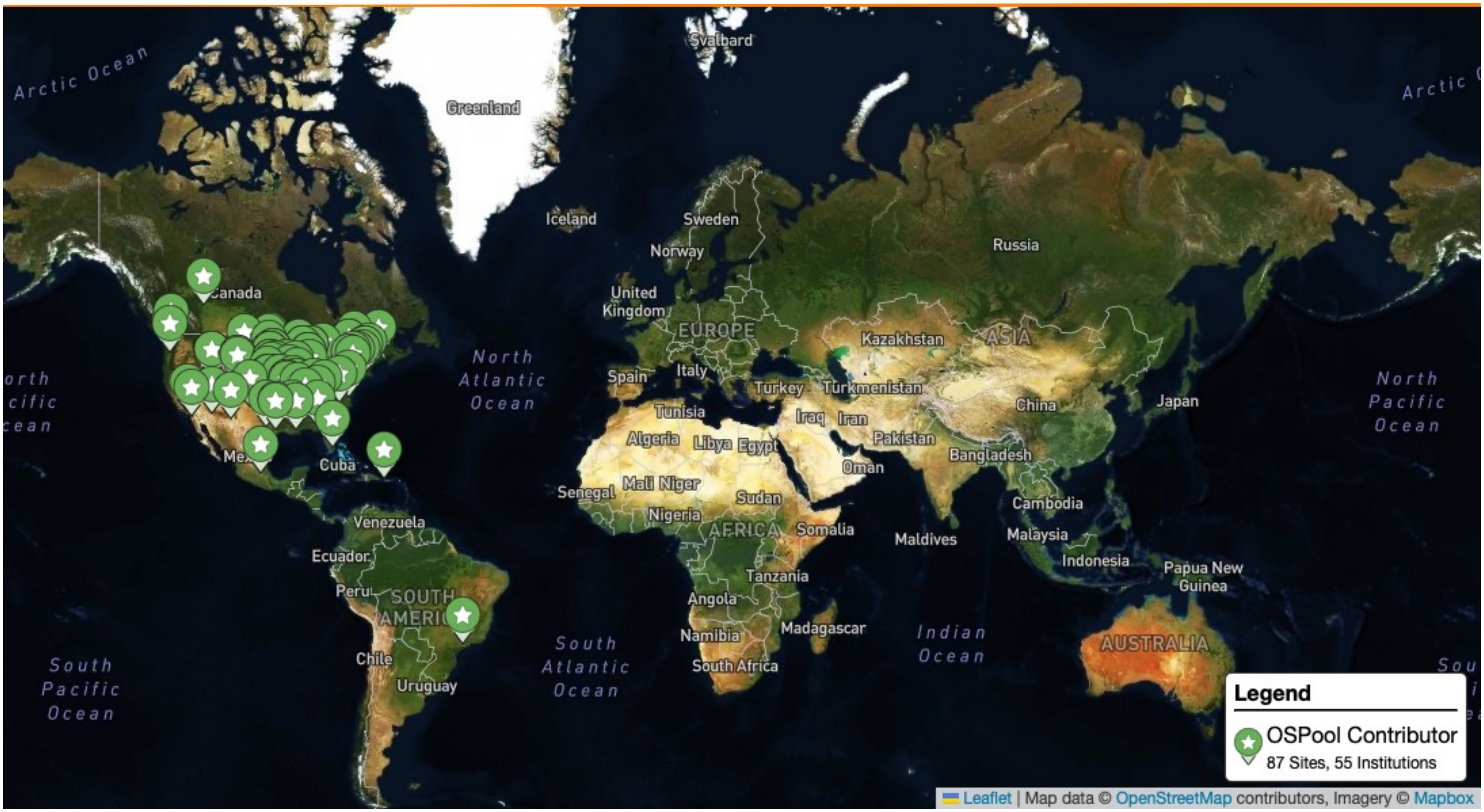
OSG Sites & Institutions



Leaflet | Map data © OpenStreetMap contributors, Imagery © Mapbox

142 “brick & mortar” Institutions provide resources to the OSG Consortium

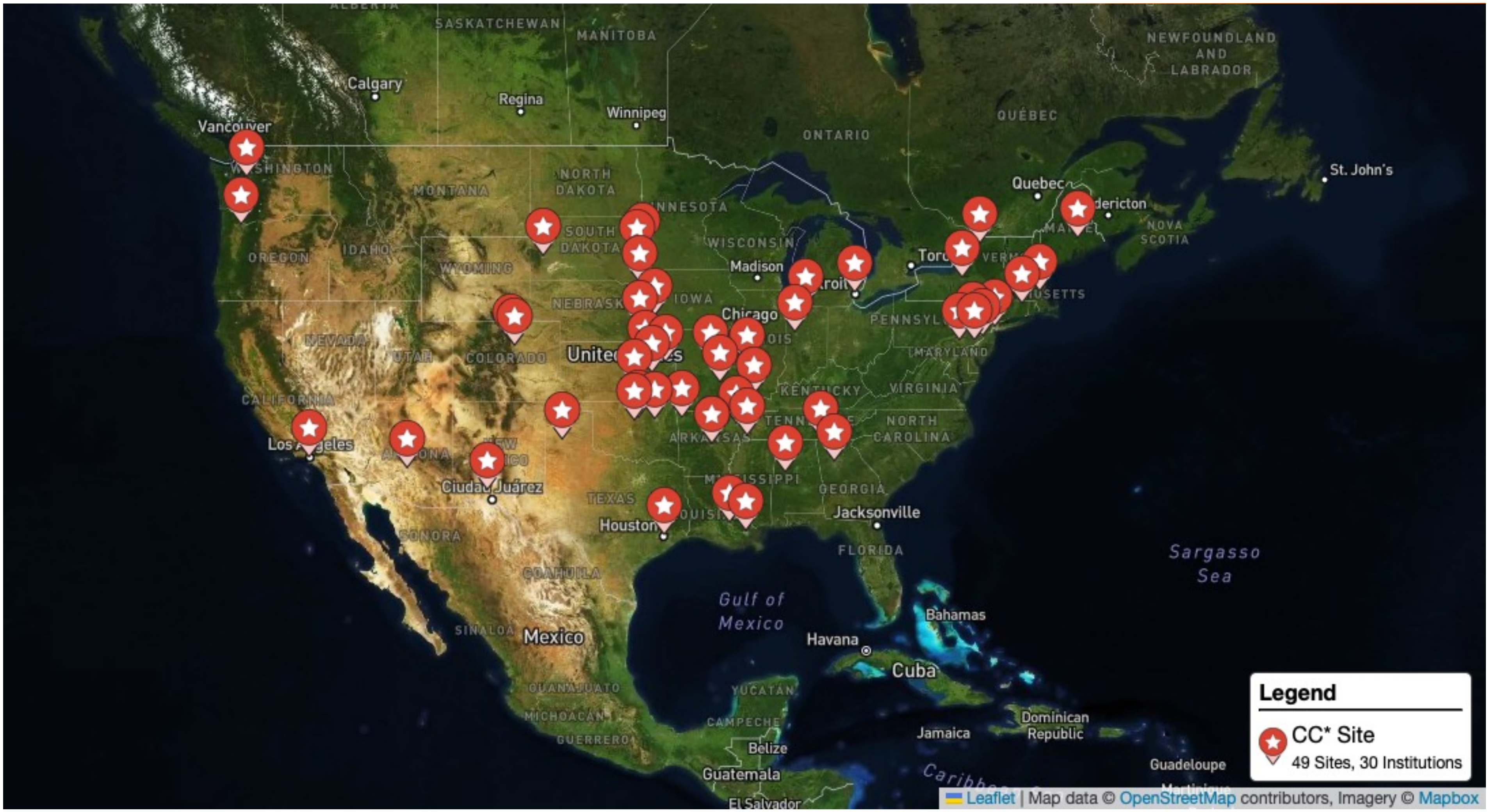
The Open Science Pool



Leaflet | Map data © OpenStreetMap contributors, Imagery © Mapbox

55 Institutions provide resources to the Open Science Pool

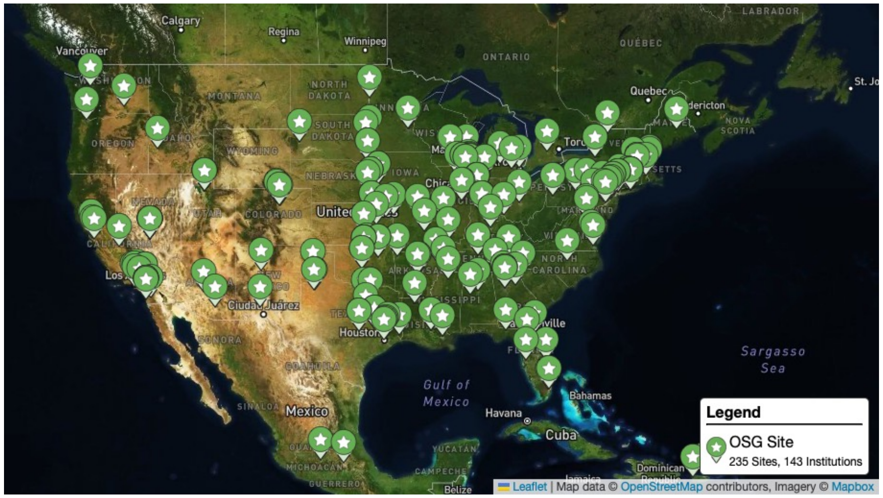
The CC* Program

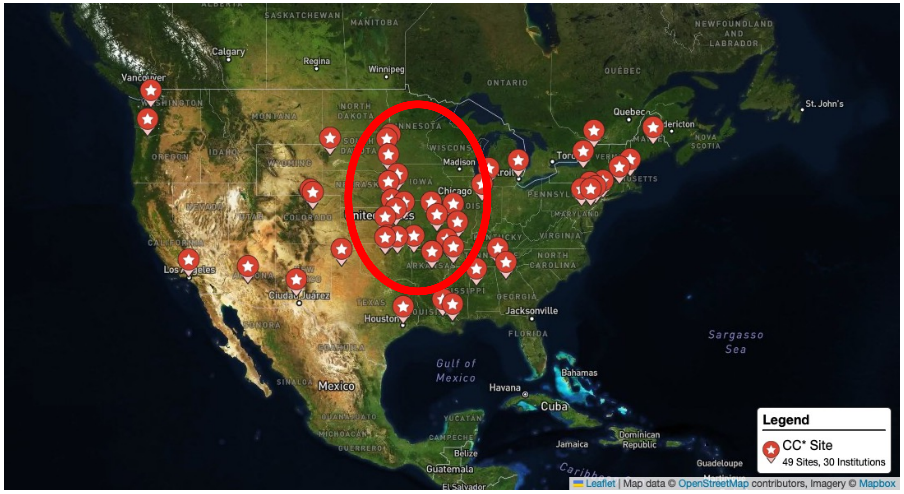


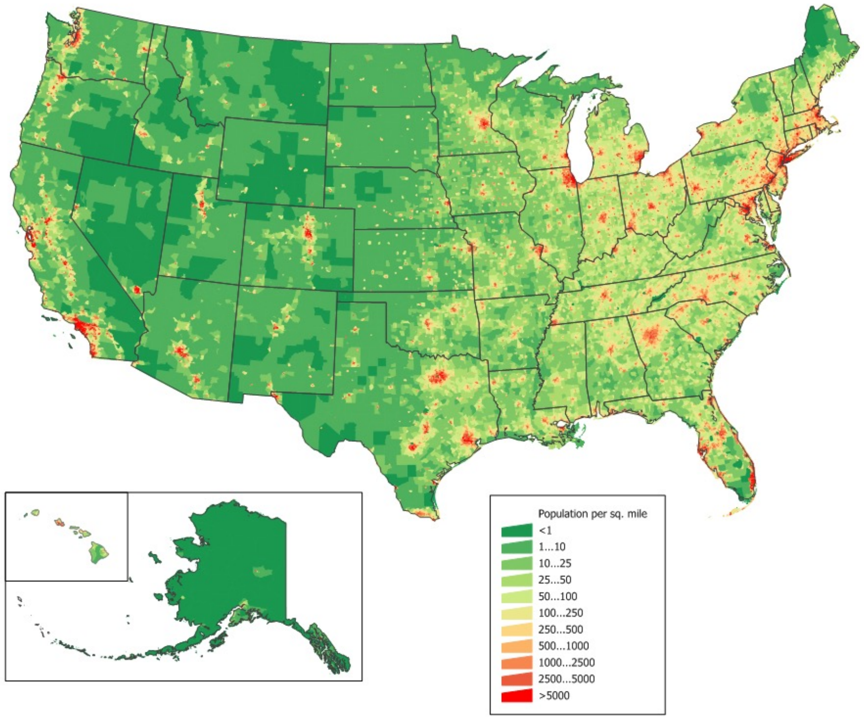
Leaflet | Map data © OpenStreetMap contributors, Imagery © Mapbox

Of the 55 contributing to OSPool, 30 are/were CC* funded

US Population Statistics vs OSG sites vs CC* sites


Legend
OSG Site
235 Sites, 143 Institutions
Leaflet | Map data © OpenStreetMap contributors, Imagery © Mapbox

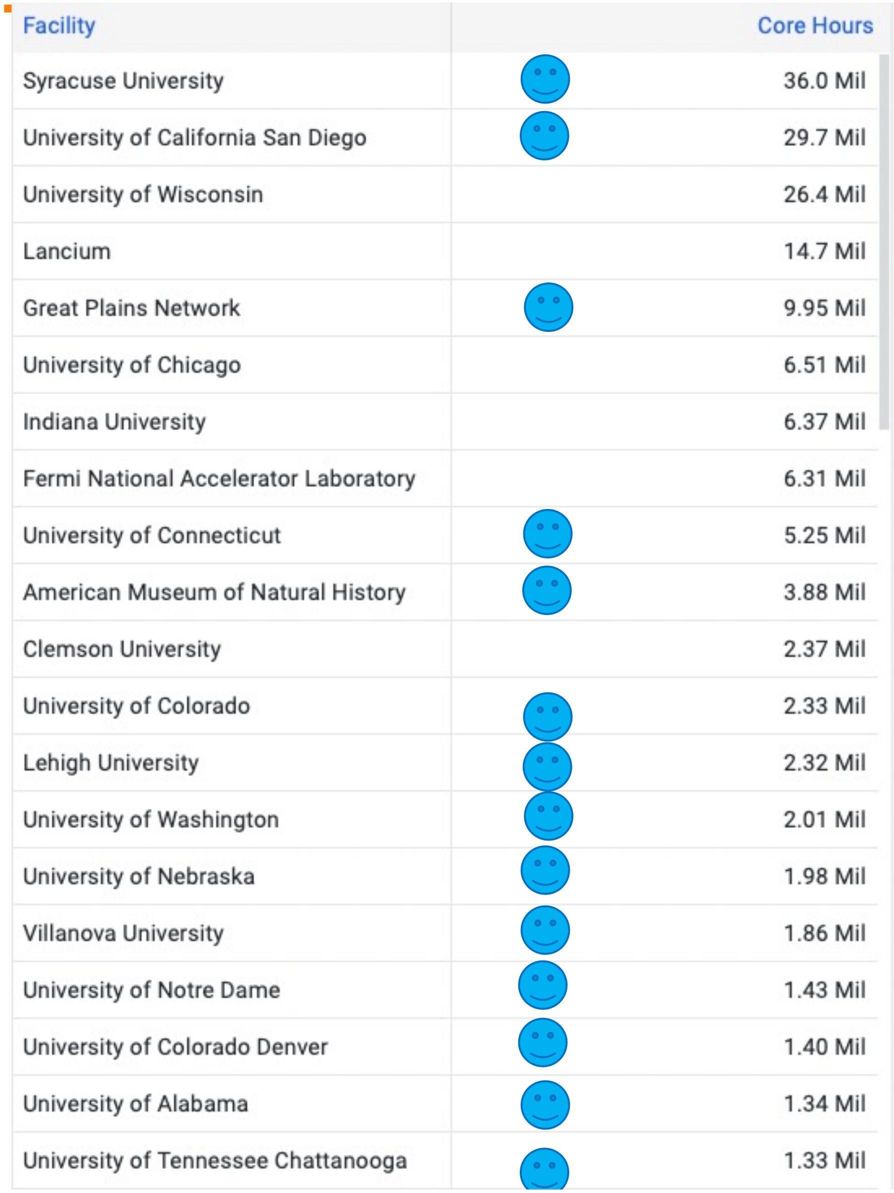

Legend
CC* Site
49 Sites, 30 Institutions
Leaflet | Map data © OpenStreetMap contributors, Imagery © Mapbox



There seems to be a region in the Midwest that knows especially well how to make use of CC* program.

More on that in Tuesday morning session before lunch.

Top OSPool Contributors



Facility		Core Hours
Syracuse University	☺	36.0 Mil
University of California San Diego	☺	29.7 Mil
University of Wisconsin		26.4 Mil
Lancium		14.7 Mil
Great Plains Network	☺	9.95 Mil
University of Chicago		6.51 Mil
Indiana University		6.37 Mil
Fermi National Accelerator Laboratory		6.31 Mil
University of Connecticut	☺	5.25 Mil
American Museum of Natural History	☺	3.88 Mil
Clemson University		2.37 Mil
University of Colorado	☺	2.33 Mil
Lehigh University	☺	2.32 Mil
University of Washington	☺	2.01 Mil
University of Nebraska	☺	1.98 Mil
Villanova University	☺	1.86 Mil
University of Notre Dame	☺	1.43 Mil
University of Colorado Denver	☺	1.40 Mil
University of Alabama	☺	1.34 Mil
University of Tennessee Chattanooga	☺	1.33 Mil

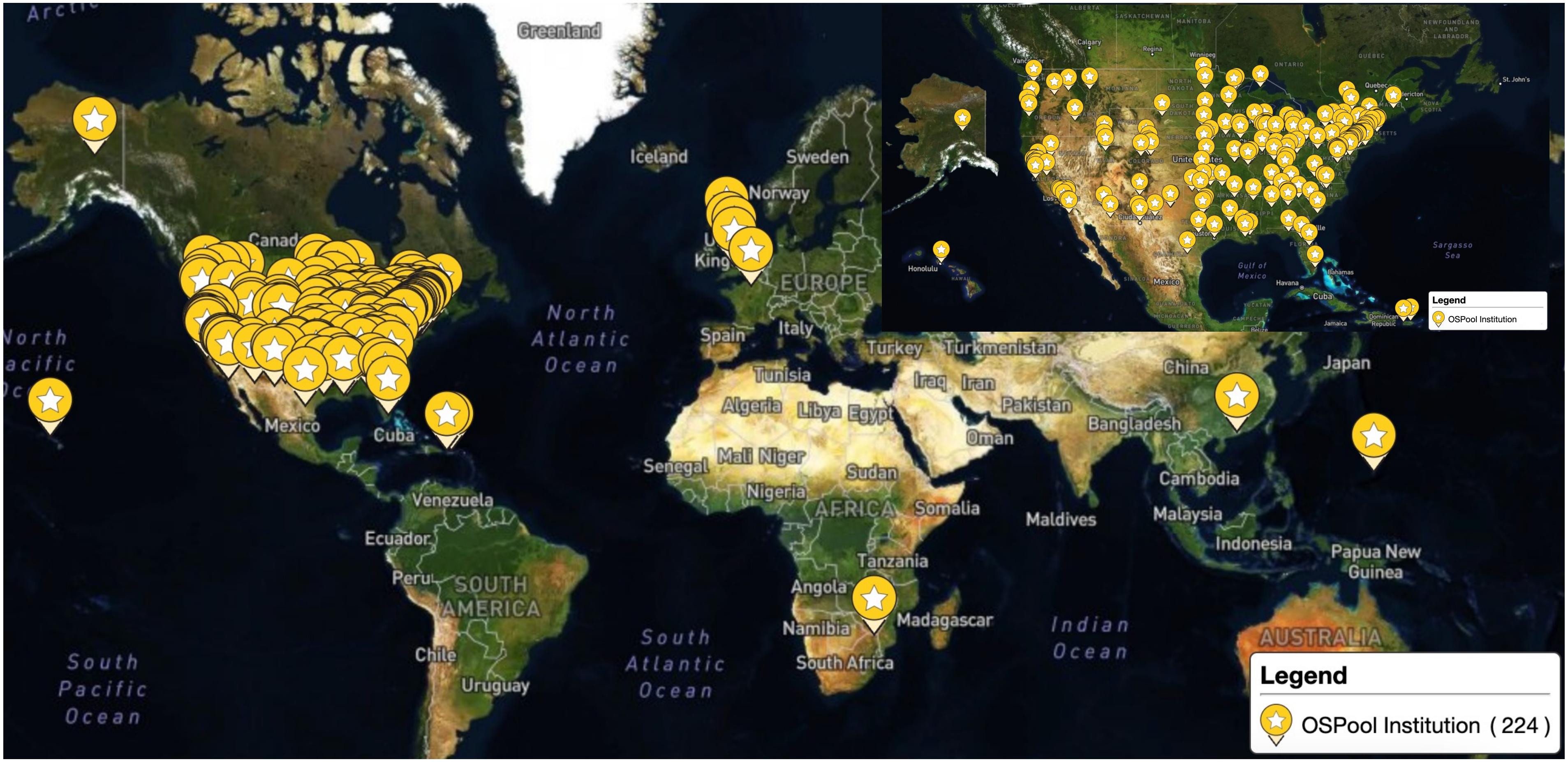
Institutions that were in the past,
or are today active awardees of
the CC* program (compute or storage)
Dominate the OSPool contributions.

Often, they contribute via resources independent of their CC* awards, and beyond the time of the award.

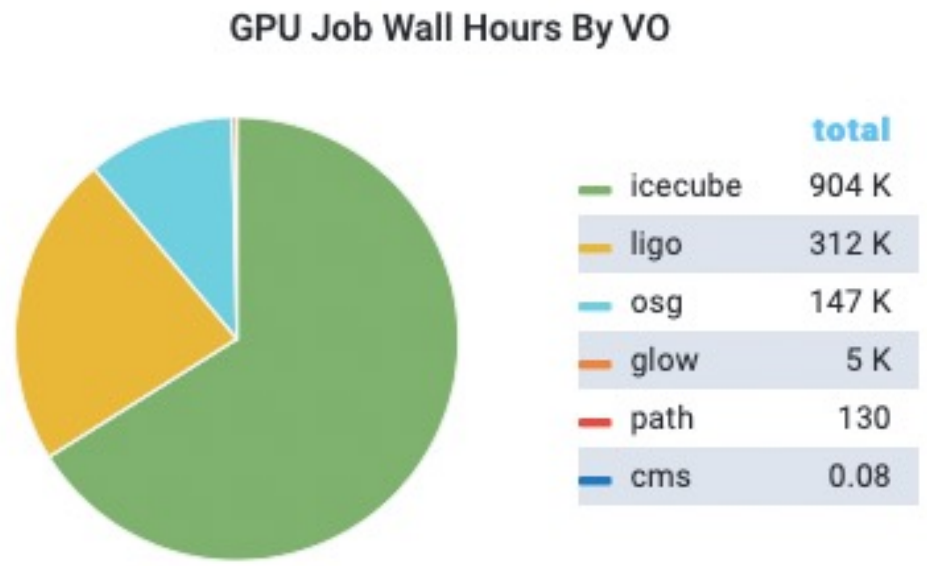
The CC* program is thus a catalyst for more general resource sharing.

91M jobs completed last year

PIs from 224 institutions use the OSPool



Aside on GPUs and AI/ML



GPU Job Wall Hours By VO

	total
icecube	904 K
ligo	312 K
osg	147 K
glow	5 K
path	130
cms	0.08

Within the last year, the OSPool provided more than $1M GPU hours to open science.

Almost all of it went to IceCube & LIGO.

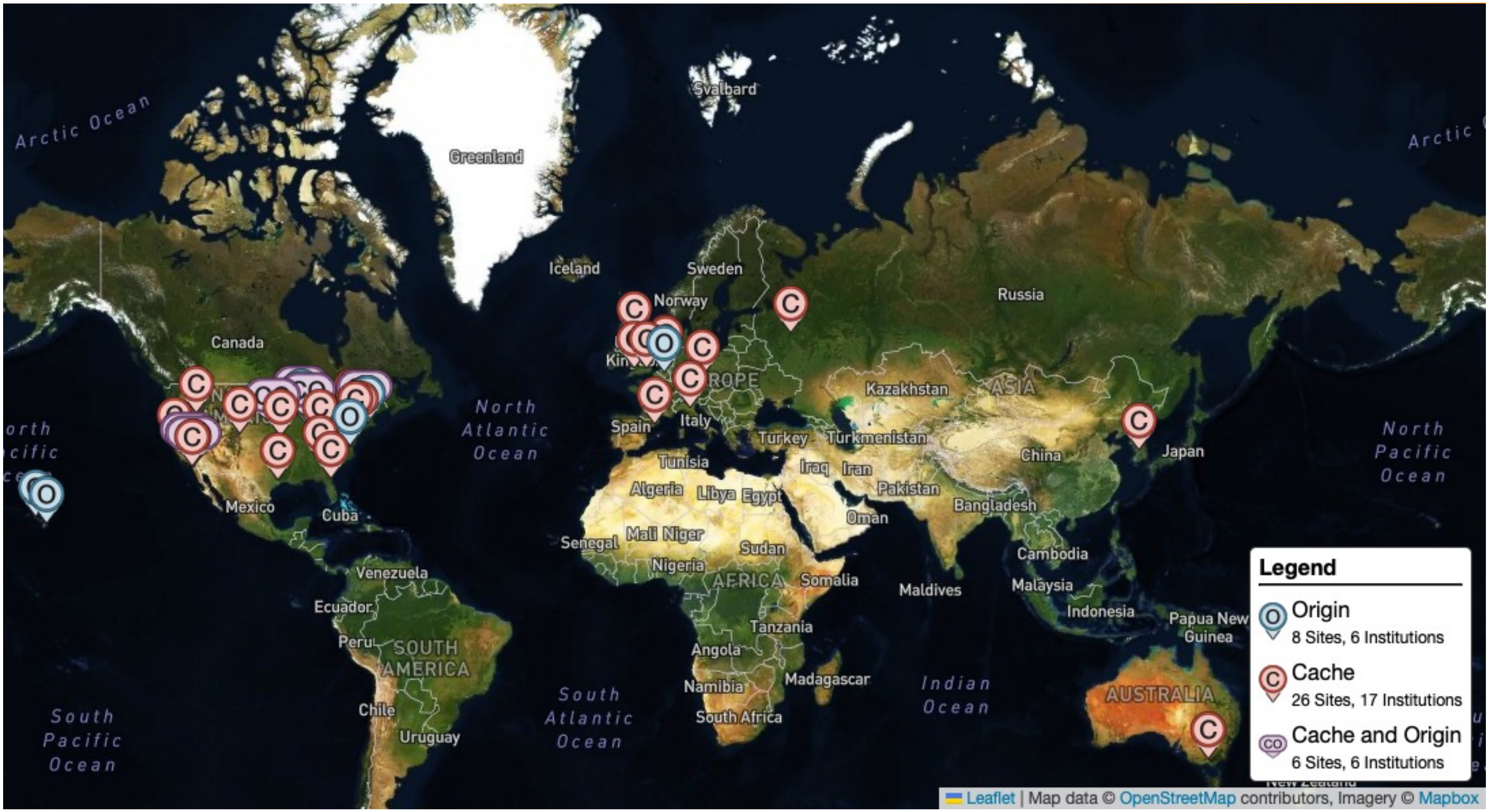
Where are all the AI/ML science use cases?

OSG Distributed Data viewed as Maps

Three Concepts

- **Data Origin** = storage server that hosts data accessible via the Open Science Data Federation (OSDF)
- **Data Cache** = cache server via which data in the OSDF is accessed.
  - **Access any data, anytime, from anywhere**
- Namespace = federated means to address objects/files registered in OSDF.
  - implemented as a “lazy crawler” across namespaces exported from data origins.

Open Science Data Federation



12 Institutions provide Data Origins … 23 Institutions provide data caches

Usage of OSDF in 2022

- Total Data available in aggregate:
  - 15 Science Collaborations
    - 385 TB of data
  - ~120 OSPool users
    - 125 TB of data, out of which 4.5TB is private
  - Top users … those with more than 1TB of data
    - 6 Science Collaborations
    - 7 OSPool users
- Total Data Read in last year:
  - 11 PB across the Science Collaborations
  - 21 PB across the OSPool users

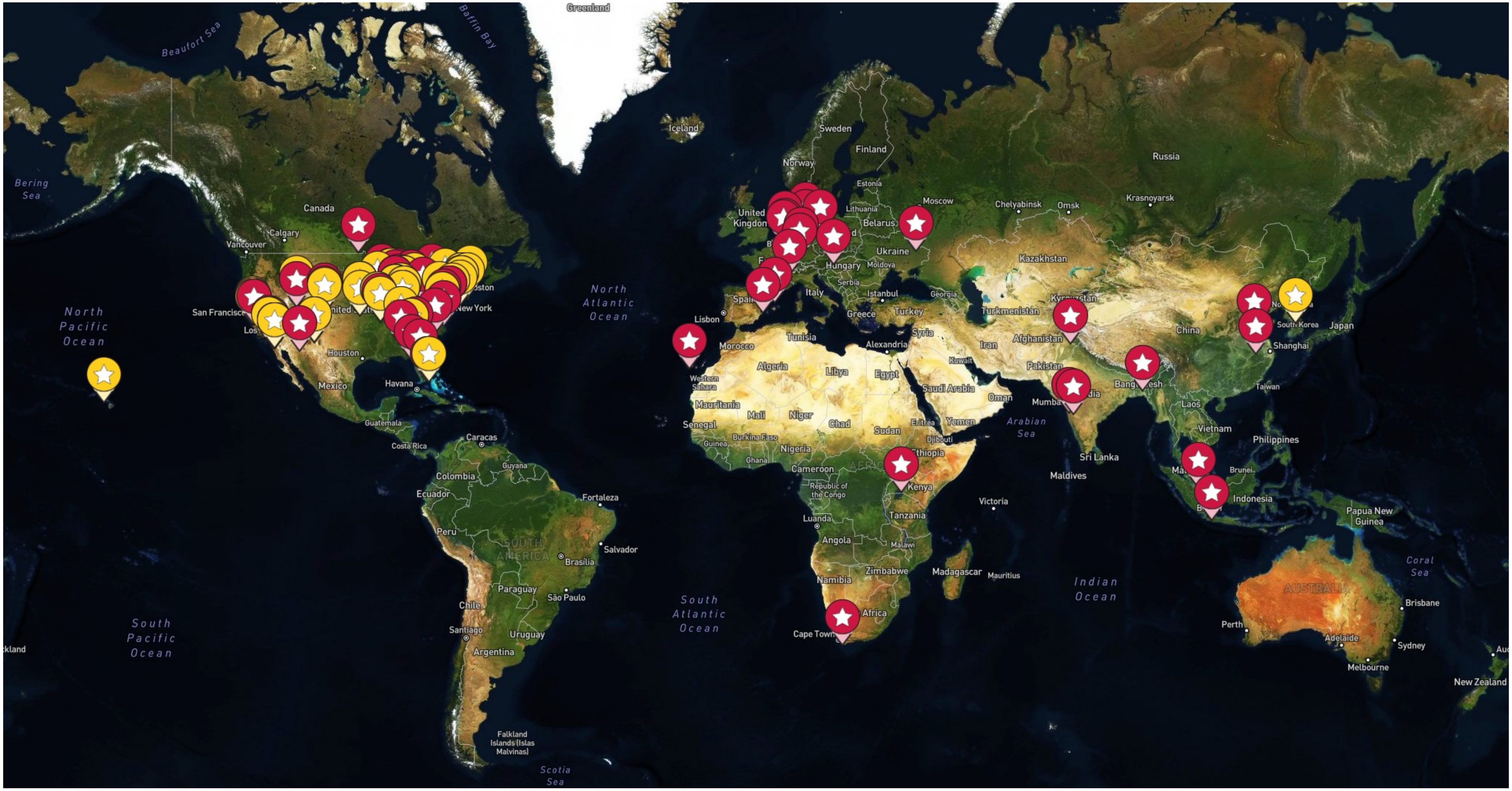
We support public & private user data

1 Billion files accessed per year

Majority of data users from OSPool
Majority of data volume from Collaborations

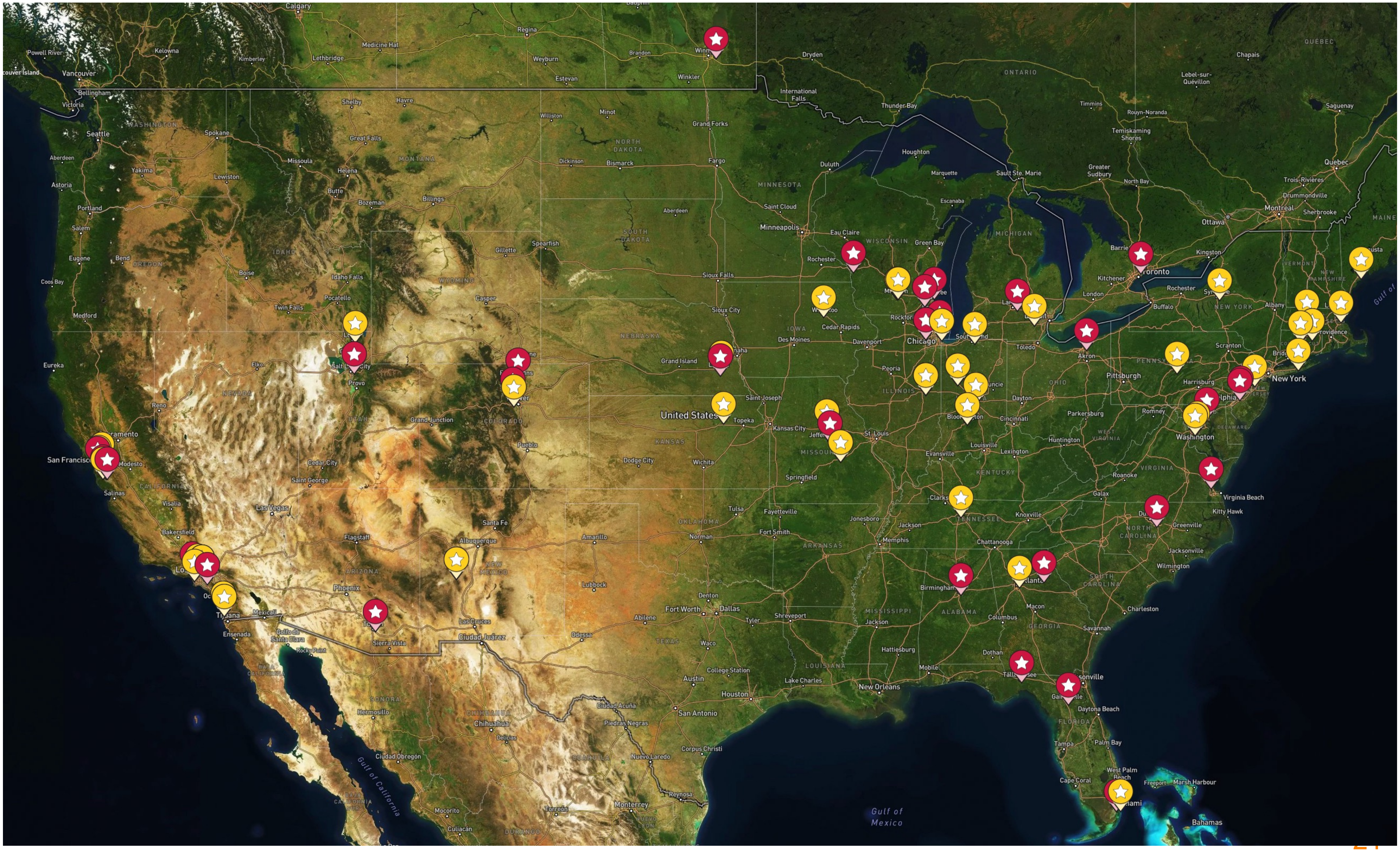
Walk Through the Week’s Program

Registration for HTC23



Yellow = in person … Red = remote registration

Zoom in on continental USA



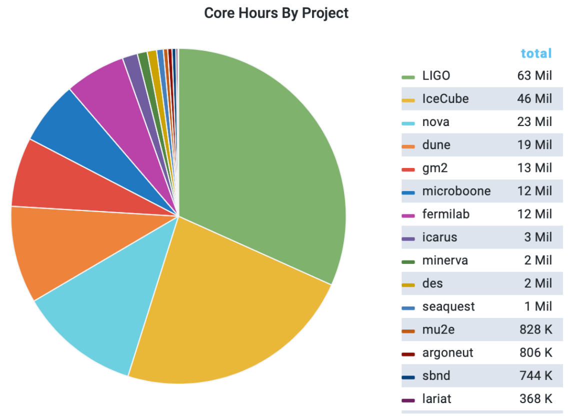
Rough Block Schedule (I)

- Monday Morning
  - Big picture overview of the state of OSG
- Monday Afternoon
  - Science on OSG … incl. **David Swanson Award**
- Tuesday Morning
  - **HTC services for Campuses**
    - Incl. CC* Program & Kevin Thompson Keynote
- Tuesday Afternoon
  - Technical talks on OSG software stack
  - **Ending the day with Campus Q&A session**

Monday & Tuesday is all plenary

Rough Block Schedule (II)

- Wednesday Parallel Sessions
  - HTCondor training & Tutorials
  - US-LHC parallel & joint sessions
  - HTC for global Science Collaborations
    - 200M core hours in last year
  - End the day with HTCondor Discussion Panel



- Thursday & Friday all Plenary
  - HTCondor week program of talks
  - Some highlights:
    - **Science Keynote on Gravitational Waves by Laura Cadonati**
    - **Strategic Directions for HTCSS by Miron Livny**
    - **What’s new, what’s coming by Todd Tannenbaum**

OSG

Summary & Conclusion

UCSD

- OSG continues to **advance all of open science via the practice of distributed HTC, and the advancement of its state of the art.**
  - Lot’s of “Big Data” across many science domains
- Open Science Pool as strategy to **democratize access to HTC**

SDSC SAN DIEGO SUPERCOMPUTER CENTER

PATh

Acknowledgements

- This work was partially supported by the NSF grants OAC-2030508, OAC-1841530, OAC-1836650, and MPS-1148698