The Open Science Data Federation

Frank Wuerthwein
Director, San Diego Supercomputer Center
Executive Director, OSG
PI, National Research Platform

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Open Science Data Federation (OSDF)

20 Caches … 6 of which are in R&E network backbone

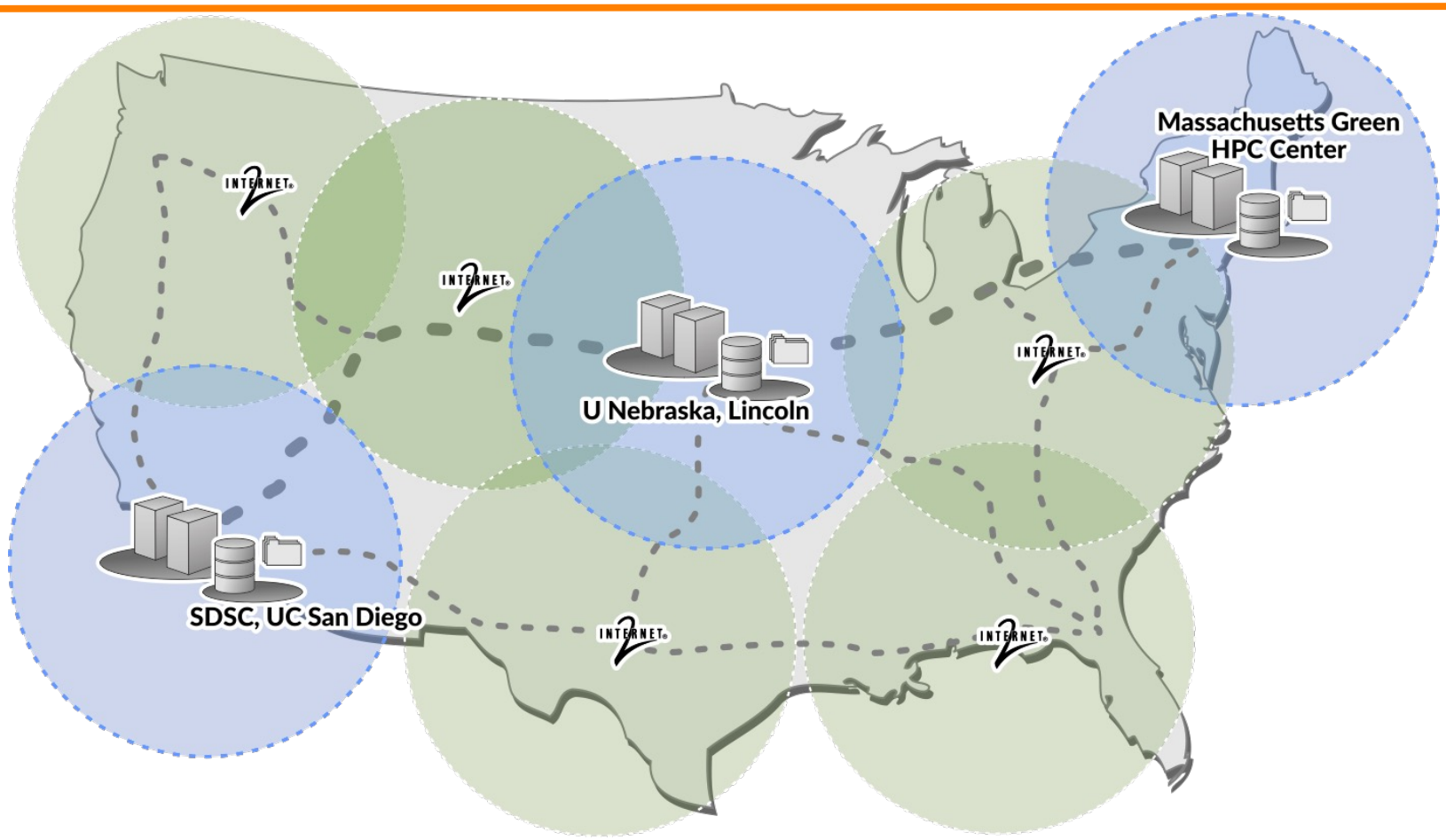
10 Data Origins … incl. one in Europe

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Origins and Caches

- Origins are places that own data.
- Caches are places via which data is accessed.
- By having a network of caches that spans the entire USA, we can provide access to any data that lives in any origin from anywhere.
  - **Any Data, Anytime, Anywhere**

National Research Platform



By fall 2022 we will have deployed 8 additional caches via NSF 2112167
"Any" location within the continental USA is within 500 miles of a cache.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Functionality of a Data Origin

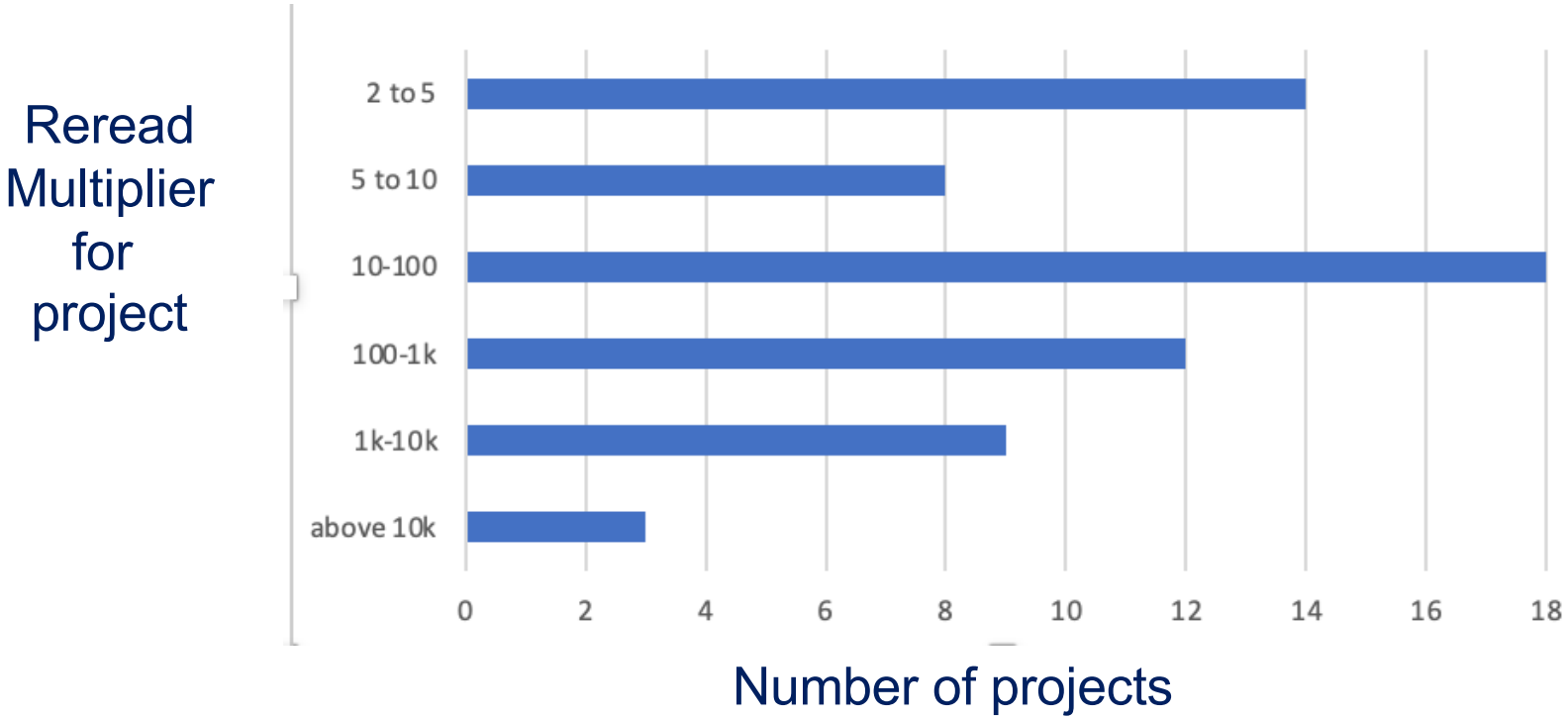
- Export your data read-only into the Data Federation
  - You choose what part of your filesystems namespace you want to export.
  - You can change this dynamically any time you want.
  - **Data can be public or private**
  - Origin uses HTTPS as protocol => works as general webserver in addition to OSG Data Federation.
- Store output data produced on OSG
  - Put via HTTPS as part of HTCondor workflows
    - authorized only to those people you want to support.
  - Read-only access possible to data stored this way.
    - What is put into an origin may be read via the federation if desired.

OSDF's Global Namespace

- Global Namespace is separate from the origins that hold the data
  - **You can move data between origins via HTTPS without changing how the data is accessed via the OSDF.**
    - Literally, nobody will notice !!!
- This allows federation of namespace that is separate from federation of server hardware that serves the namespace.
  - Lot's of interesting ways of using the power this provides you with.

Use of the OSG Data Federation

In 2021,
92 research groups, 9 collaborations, 1 campus
read 32PB of data out of a working set of 420TB
for an average re-read factor of 75.



SDSC SAN DIEGO SUPERCOMPUTER CENTER

"Working Set"

- What we really mean is the sum of sizes of all unique files accessed via the data federation within a given time period.
- This is not the same as the total disk space used across origins.
  - Somebody could access foo then delete it, and put foo2 in its place, then access foo2.
  - The working set would include the sum of the filesizes of foo and foo2 even though these two files never existed at the same time.

Top Users of Data Federation

Project	Data Read	Working Set	Reread multiplier
LIGO (Private)	10PB	38TB	264
Minerva	5.6PB	3.1TB	1,789
NOVA	2.6PB	1.9TB	1,348
LIGO (Public)	2.4PB	38TB	67
Tufts_Hempstead	2.0PB	380GB	5,321
DUNE	1.6PB	185GB	8,658
Steward	1.0PB	11TB	92
REDTOP	874TB	95TB	9.2
Molcryst	570TB	5GB	115,650
BiomedInfo	530TB	66GB	8,090

17 projects have >TB working sets

11 of these are individual researchers and small groups

Tufts Computer Architecture Lab

Steward Observatory Data Analytics

R&D towards future particle physics experiment

Quantum chemical and machine learning insights into supra-molecular organization of molecular crystals

Development and application of software tools for performing large-scale biomedical informatics on microbial genome sequence data.

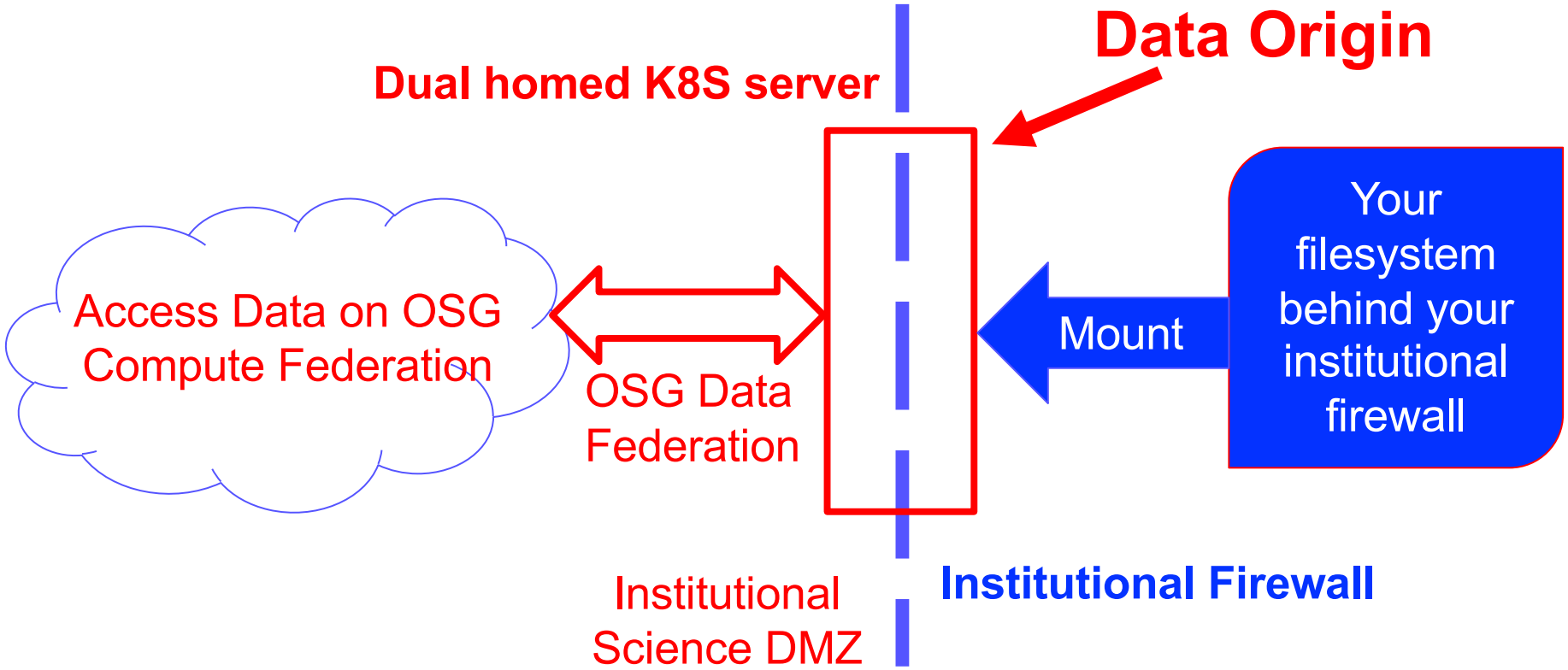
SDSC SAN DIEGO SUPERCOMPUTER CENTER

You Can Join your data in 2 ways

1) Transfer your data via HTTPS to the PATh supported Origin
2) **Federate the filesystem at your institution with the OSDF**

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Federating Data with the OSDF



Your fileserver with your data can be behind your institutional firewall.
A dual homed K8S server mounts only those filesystems you want to export.
We operate the OSDF API by deploying a “Data Origin” container into K8S.

CC* Solicitation (NSF 22-582)

OSG's Support for Campus Cyberinfrastructure Proposals and Awardees

Upcoming Deadline: June 27th, 2022

The National Science Foundation Campus Cyberinfrastructure (CC*) program (NSF 22-582) invests in coordinated campus and regional-level cyberinfrastructure improvements and innovation. The 2022 solicitation has two program areas, both of which explicitly mention and encourage the use of OSG services to meet requirements.

The NSF supports awards in 2 CC* program areas:

- Data Storage awards, which mention the OSG Open Science Data Federation, encouraging responses that would add data origins or caches at campuses
- Regional Computing awards, in which the NSF strongly encourages joining PATh, and using our services to contribute to the Open Science Pool

Let OSG Help with your CC* Proposal

The Partnership to Advance Throughput Computing (PATh), which develops technologies and operates services for the OSG, has significant experience working with CC* applicants and awardees, and offering letters of support and consulting for:

- Sharing data with authorized users via the Open Science Data Federation (OSDF)
- Bringing the power of high throughput computing via the OSPool to your researchers
- Gathering science drivers and planning local computing resources
- Meeting CC*-required resource sharing as specified in (NSF 22-582), and other options for integrating with the OSG Consortium
- Providing connections to help with data storage systems for shared inter-campus or intra-campus resources
- Building regional computing networks
- Developing science gateways to utilize high throughput computing via the OSPool

CC* applicants are encouraged to email OSG Support with questions or requests for letters of support regarding their CC* proposal.

Contents

Storage Proposals

- $500,000 over 2 years
- Up to 25% labor
  - $375k hardware
  - $125k labor
- From the solicitation:
  - At least 20% of the disk/storage space on the proposed storage system must be made available as part of the chosen federated data sharing fabric.

Fkw's Ideas for 20% Sharing

- Using some of the storage to share curated institutional data with the public.
  - Personally, I would find it exciting if this lead to OSDF being used for making curated data available, including FAIR principles.
- Using some of the storage as cache space for OSPool use of the institutional computing resources.
  - This makes a lot of sense for those institutions that make significant contributions to the OSPool.
- Using some of the storage as Origin space for the OSPool community.
- **Some combination of all of the above, or something different that you come up with.**

Summary & Conclusion

- The Open Science Data Federation (OSDF) is about to undergo significant growth
  - Via the addition of caches across the continental USA
  - Via the addition of origins in response to NSF 22-582
- The OSDF may be used in conjunction with the better known Open Science Compute Federation, or independent of it.
- If you have any questions regarding OSDF in the context of NSF 22-582 please send email to:
  - cc-star-proposals@opensciencegrid.org

Acknowledgements

- This work was partially supported by the NSF grants OAC-2112167, OAC-2030508, OAC-1841530, OAC-1836650

SDSC SAN DIEGO SUPERCOMPUTER CENTER