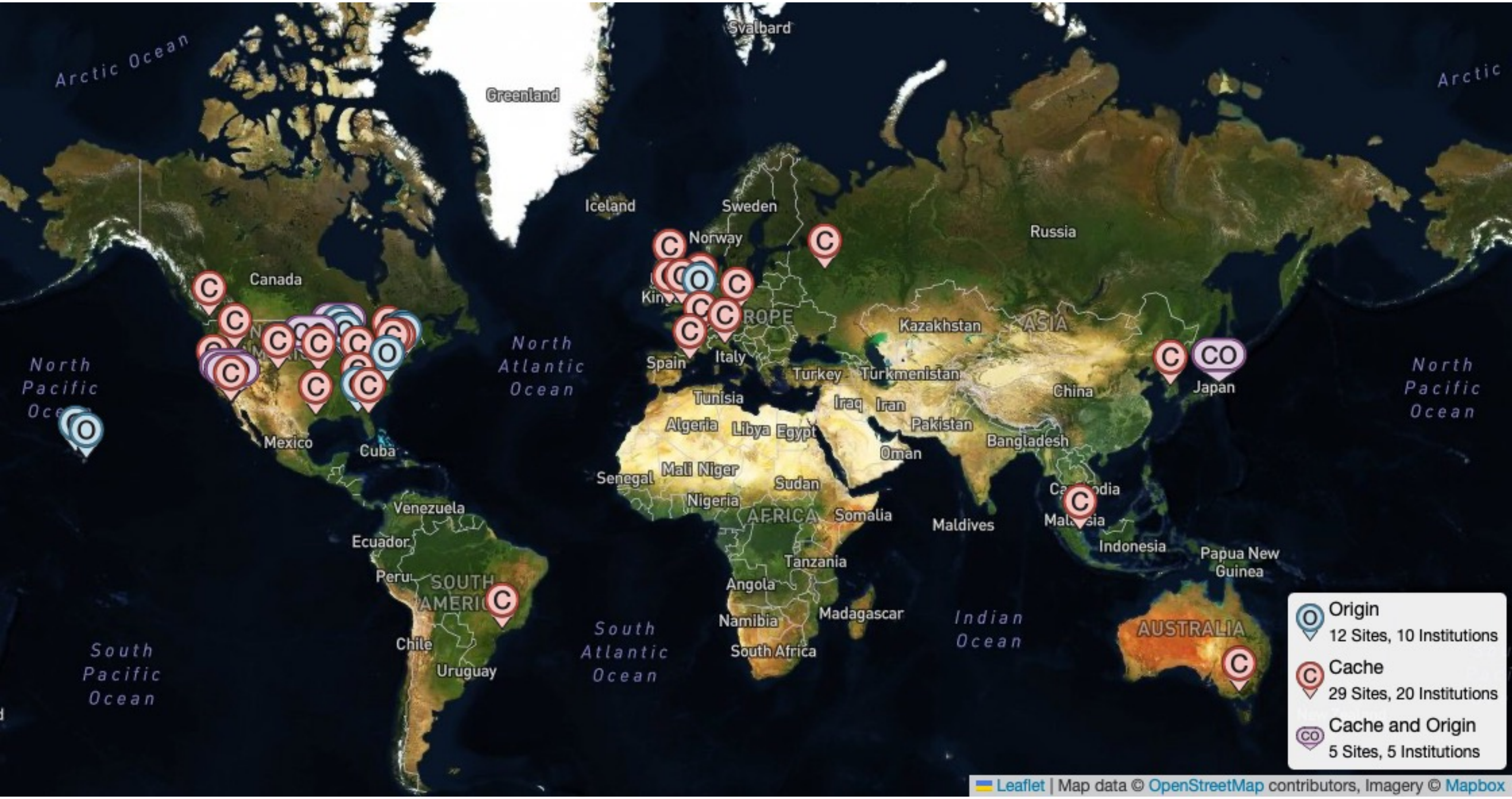
OSDF Deployment & Use

Frank Würthwein
OSG Executive Director
UCSD/SDSC

July 8th 2024

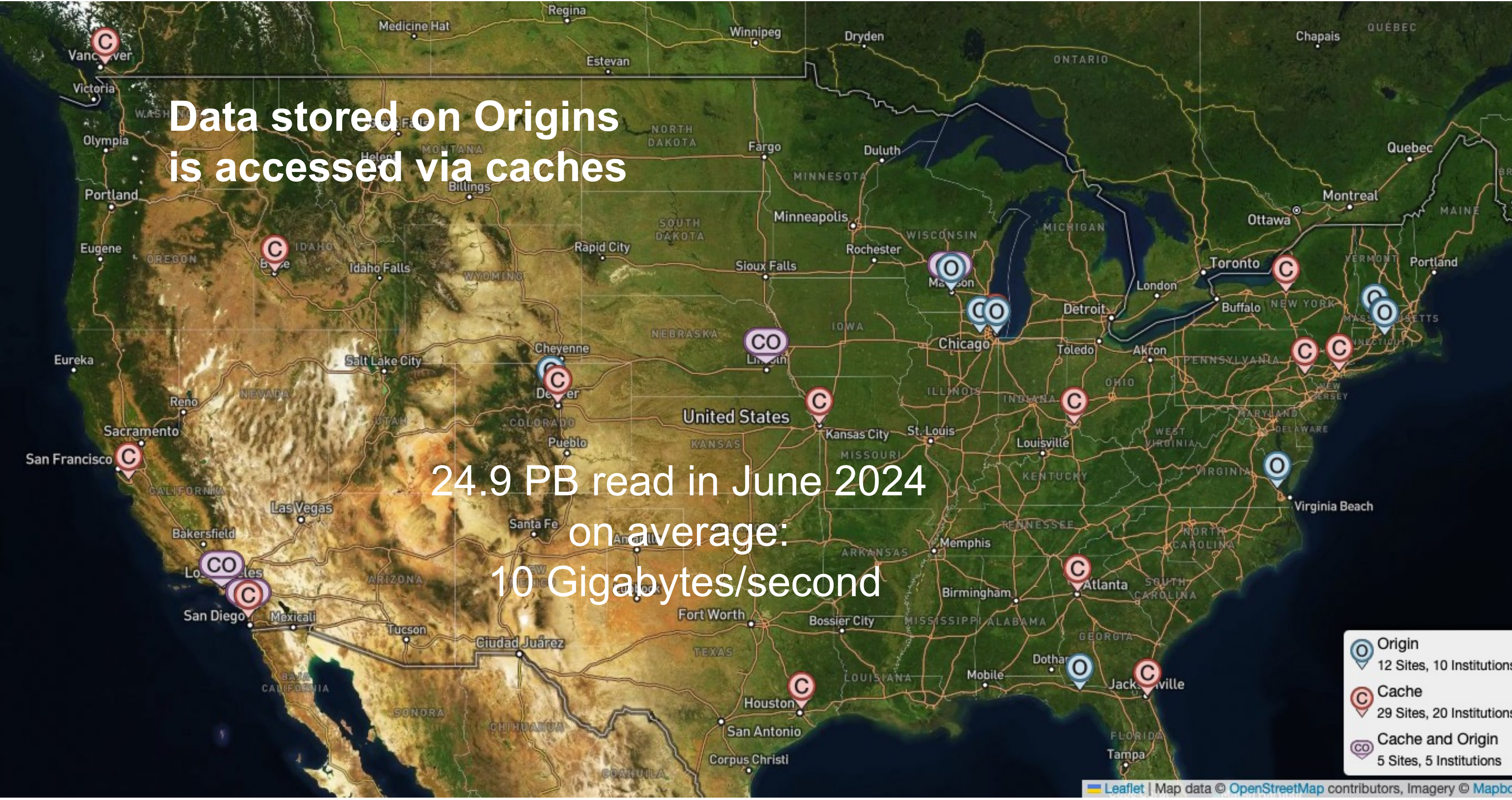
35 Institutions Contribute to OSDF Today



Leaflet | Map data © OpenStreetMap contributors, Imagery © Mapbox

17 Origins and 34 caches across 5 continents

Zoomed in on Continental USA



Data stored on Origins is accessed via caches

24.9 PB read in June 2024
on average:
10 Gigabytes/second

Origin
12 Sites, 10 Institution

Cache
29 Sites, 20 Institution

Cache and Origin
5 Sites, 5 Institutions

Leaflet | Map data © OpenStreetMap contributors, Imagery © Mapbo

80 Gigabit per second … that’s 80% of a 100G pipe

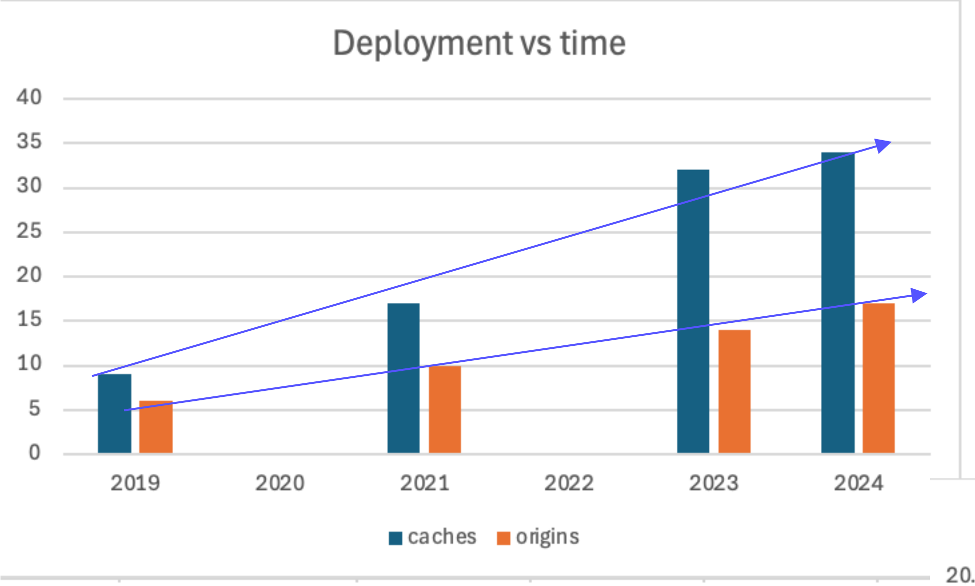
Observe <3% cache misses => OSDF caches save >75Gbps in network traffic

OSDF by Numbers

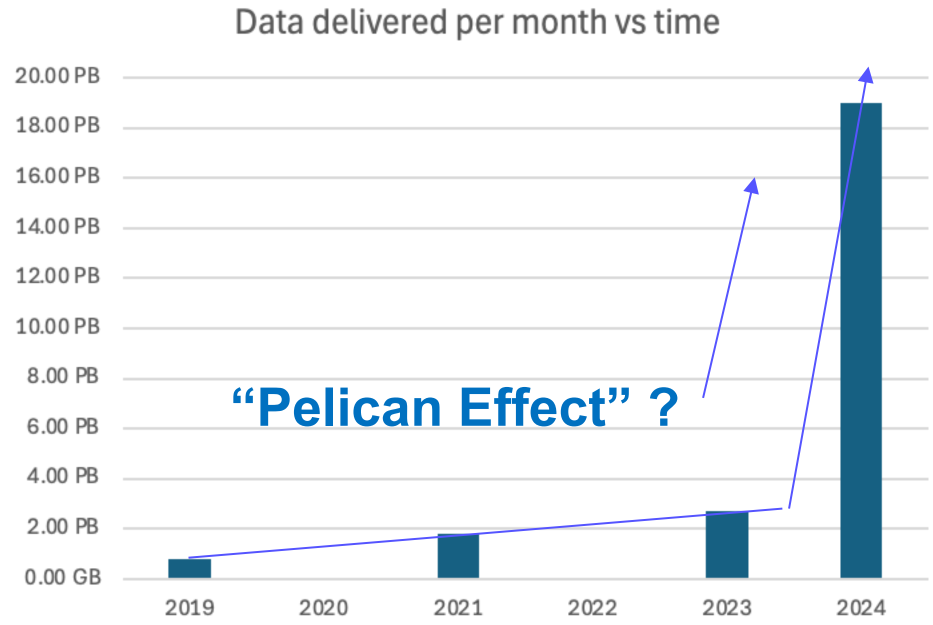
Realtime visualization at:

https://osdf.osg-htc.org

Historic Perspective



~5 caches added per year
~2 origins added per year



"Pelican Effect" ?

Data volume delivered per month
went from ~40% growth per year
between 2019 – 2023 to
7x growth in the last year

Fun Facts for the Month of June

24.9 PB read total

10% of this is accounted for by the OSPool

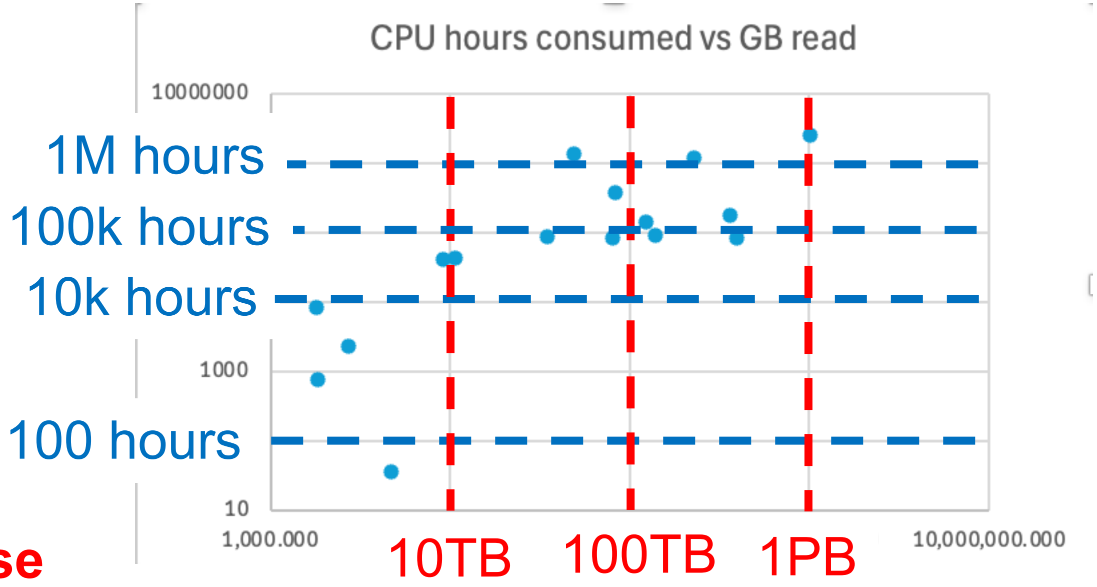
June 2024 OSPool Numbers

- 61 out of 172 users used OSDF
- 31 out of 98 projects used OSDF

~ 1/3 of OSPool uses OSDF !!!

- OSPool users transfer small files with HTCondor and large files with OSDF:
  - 43% of all bytes transferred by OSDF
    - But only 2.3% of all files

About a dozen projects read 10TB to 1PB consuming 10k to 1M CPU-h during the month of June



Data use is only very loosely correlated with CPU use

Top OSPool Data Users in June 2024

PI	Institution	Science	Description	TB's Read	CPU-h
Chun Shen	Wayne State	Nuclear Physics	Dynamical Modelling of relativistic heavy Ions	1,020	2.5M
Paul Vaska	Stonybrook	Biology	MC for developing better image reconstruction	398	85k
Jeffrey D. Jensen	ASU	Biology	Population genetics to study evolutionary processes	363	178k
J. Pixley	Rutgers	Physics	Condensed Matter Theory incl. quantum phase transitions of many-body systems	230	1.2M
D. Katz	CSU Northridge	Math	Searchers for binary sequences with identical autocorrelation spectra	140	91k
O. Isayev	CMU	Chem	QC & ML insights into supra-molecular organization of molecular Xtals	123	140k
H. Fricker	UCSD	Geo & Earth Sciences	Use satellite remote sensing data to study processes that affect mass loss of Antarctic Ice Sheet	81	83k

The top data users span a wide range of sciences … and institutions … and locations

Fun Facts for the last year

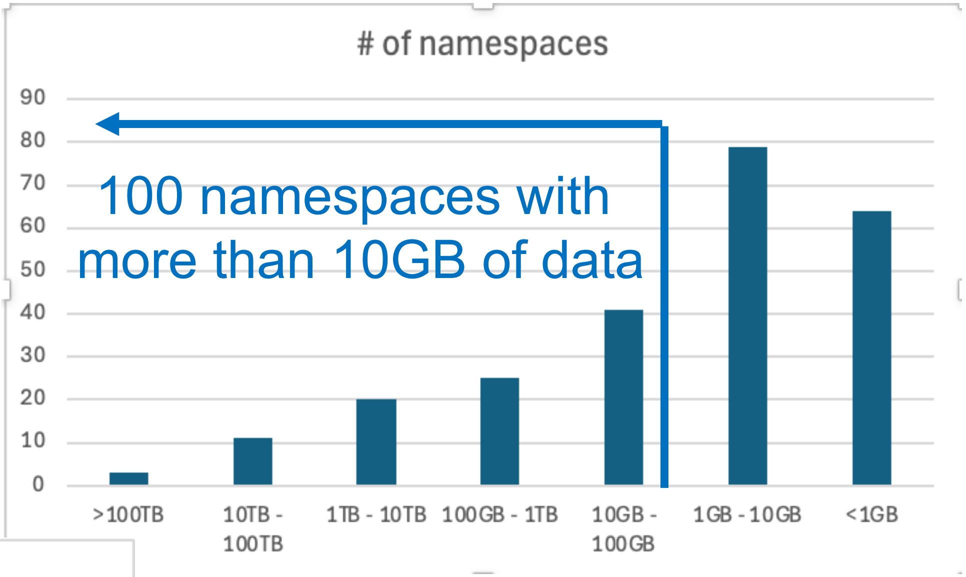
Working set size = volume of unique data read last year

Total read = volume of data read last year

Re-use multiplier = total read / working set size

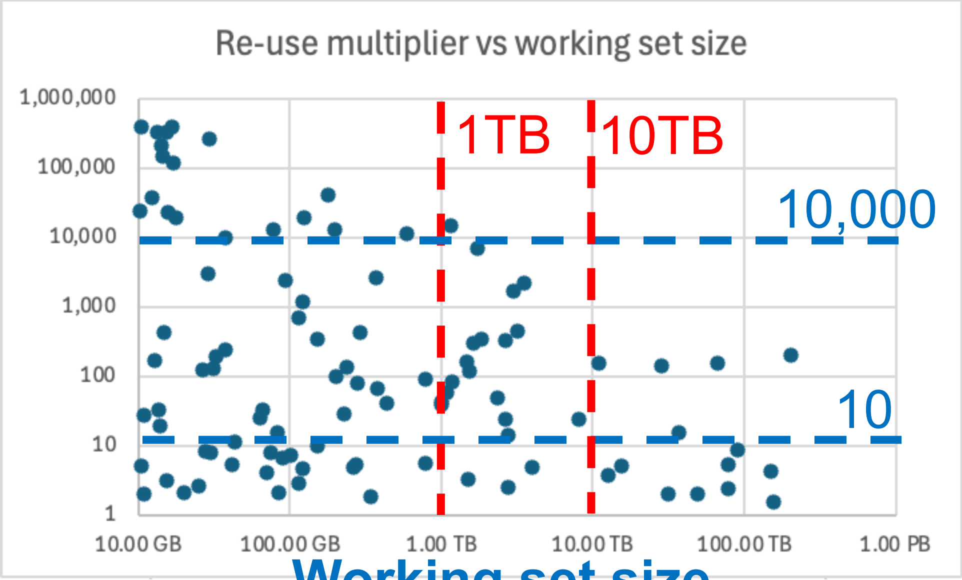
OSDF Usage Accounting by namespace

Looking at the 100 namespaces with >10GB of working set size



100 namespaces with more than 10GB of data

Working set size

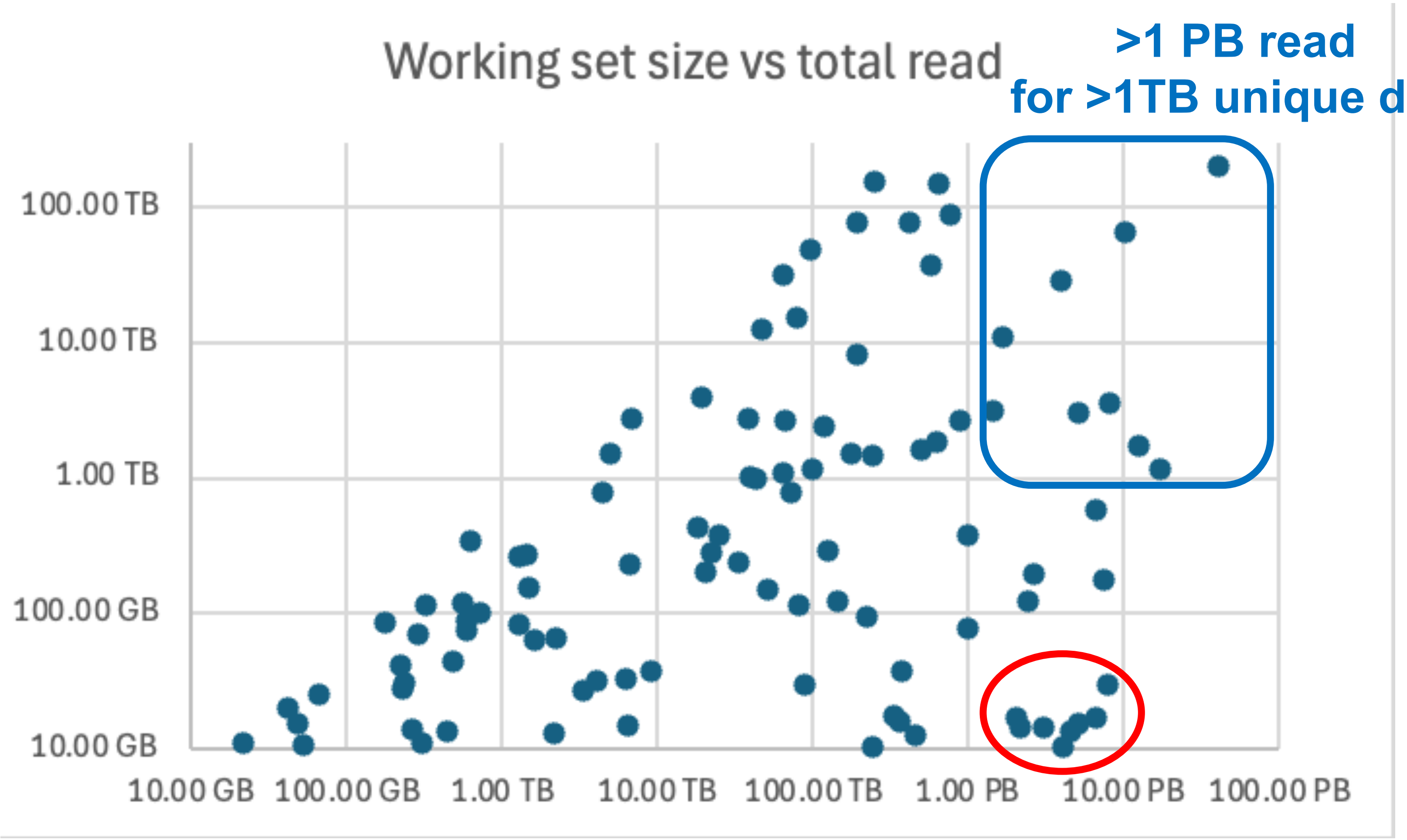


Working set size

TB datasets were read between a few to 10,000 times

Little correlation between size of a namespace & how often it's read

Let's look at two patterns



>1 PB read
for >1TB unique data

>1 PB read
for <50 GB unique data

Each of these patterns comprise ~1/3 of the namespaces with >1 PB read

>PB read of a >TB dataset

- There are 9 namespaces like this, and all 9 belong to international collaborations

=> See Panel Discussion Tuesday Afternoon

name	Read	Unique data
LIGO IGWN	40 PB	203 TB
IceCube	10 PB	66 TB
LIGO users	4 PB	28 TB
IGWN shared	1.7 PB	11 TB
KOTO	8 PB	3.5 TB

name	Read	Unique data
Einstein Telescope	1.5 PB	3.2 TB
Nova	5 PB	3 TB
MicroBoone	12 PB	1.7 TB
IGWN CIT	17 PB	1.2 TB

Gravitational Wave Observatories Community dominates unique data

Next come neutrino physics experiments (IceCube, Nova, MicroBoone)

>PB read of a 10-50GB dataset

- There are 7 namespaces like this, and all 7 belong to OSPool users

name	Read	Unique data
J. Pixley 1	7.8 PB	29 GB
G. Thomson	2 PB	17 GB
Chin Shen 1	5.2 PB	15 GB

name	Read	Unique data
Chin Shen 2	3 PB	14 TB
Chin Shen 3	2.2 PB	14 GB
Paul Vaska	4.5 PB	13 GB
J. Pixley 2	4.1 PB	10 GB

J. Pixley: Condensed Matter Theory
G. Thomson: Telescope Array (TA) is the largest cosmic ray detector in the Northern hemisphere, which is located in Millard county, Utah.
Ch. Shen: Nuclear Physics Theory
P. Vaska: MC simulation for better image reconstruction for biological sciences

OSDF use from outside of OSG

10 namespaces in OSDF that belong to NRP communities

1.5 PB was read from 311 TB of unique data for these.

We assume that at least some of this reading was done via native NRP access mechanisms, i.e. from outside OSG.

Summary & Conclusion

- **The Open Science Data Federation has seen a 7x increase of use within the last year.**
  - At this point our caching saves >75% of a 100G transnational network pipe.
- **Roughly 1/3 of all OSPool users now use OSDF**
  - OSPool users account for roughly 10% of the total reads.
  - The top projects by OSDF reads span Biology, Physics, Math, Chemistry, and Geological & Earth Sciences
- The usage pattern we observe from international collaborations and OSPool users are quire different
- **We start to see usage of OSDF from outside OSG**

Acknowledgements

- This work was partially supported by the NSF grants OAC-2112167, OAC-2030508, OAC-1841530, OAC-1836650, the CC* program, and in kind contributions by many institutions including ESnet, Internet2, and the Great Plains Network.