



Analyzing scientific data sharing patterns for in-network data caching

Elizabeth Copps, Katherine Zhang, Alex Sim, John Wu at LBNL
Inder Monga, Chin Guok at ESnet
Frank Wuerthwein, Diego Davila, Edgar Hernandez at UCSD

Introduction

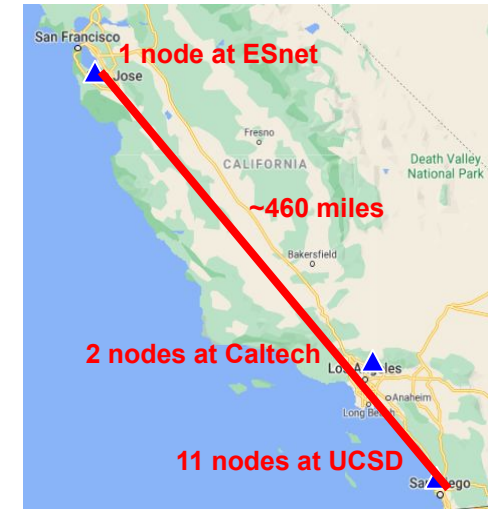
- **Data generation in newer scientific experiments and simulations**
 - Exponential data volume increase, particularly for geographically distributed large collaborations
 - E.g. Large Hadron Collider (LHC) and Large Synoptic Survey Telescope (LSST)
 - Network bandwidth requirement increase
- **Observation**
 - Significant portion of the popular dataset is transferred multiple times to different users as well as to the same user
 - Data sharing
 - Reduce the redundant data transfers
 - Save network traffic volume, consequently.
 - Lower data access latency
 - Overall application performance is expected to be improved

In-network data caching

- **In-network caching allows data sharing between users in same region**
 - **Reduces redundant transfers (costly, inefficient)**
 - **Reduce demand on transatlantic links**
 - **ESnet**
 - **HEP community**
- **Goals in this study**
 - **Analyze cache utilization and network utilization**
 - **Number of accesses, cache misses, cache hits, data transfer sizes, and shared data sizes**
 - **Identify and study the impact of node downtimes**
 - **Predict future resource loads and utilization to increase data availability**
 - **If users double, how will it affect accesses? cache hits? etc.**
 - **How many XCache nodes will maintain resource efficiency?**

Application use case with CMS

- Working with US CMS data analysis using MINIAOD/NANOAOD
- High-Luminosity Large Hadron Collider
 - HL-LHC aims to increase performance after 2025
 - Increase luminosity by factor of 10
 - Luminosity proportional to the number of collisions in a given time
- Southern California regional cache consists of 14 federated storage caches for US CMS
 - 11 at UCSD: each w/ 24 TB, 10 Gbps network connection
 - 2 at Caltech: each w/ 180 TB, 40 Gbps network connection
 - 1 at ESnet: 40 TB, 40 Gbps network connection
 - Spans 500 miles (socio-politically relevant distance)
- Measurement data
 - Regional data collected from XCache nodes from June - Aug 2020
 - ESnet node data collected from May - Dec 2020
 - Analysis on Cori at NERSC



Sunnyvale–San Diego is the relevant distance scale





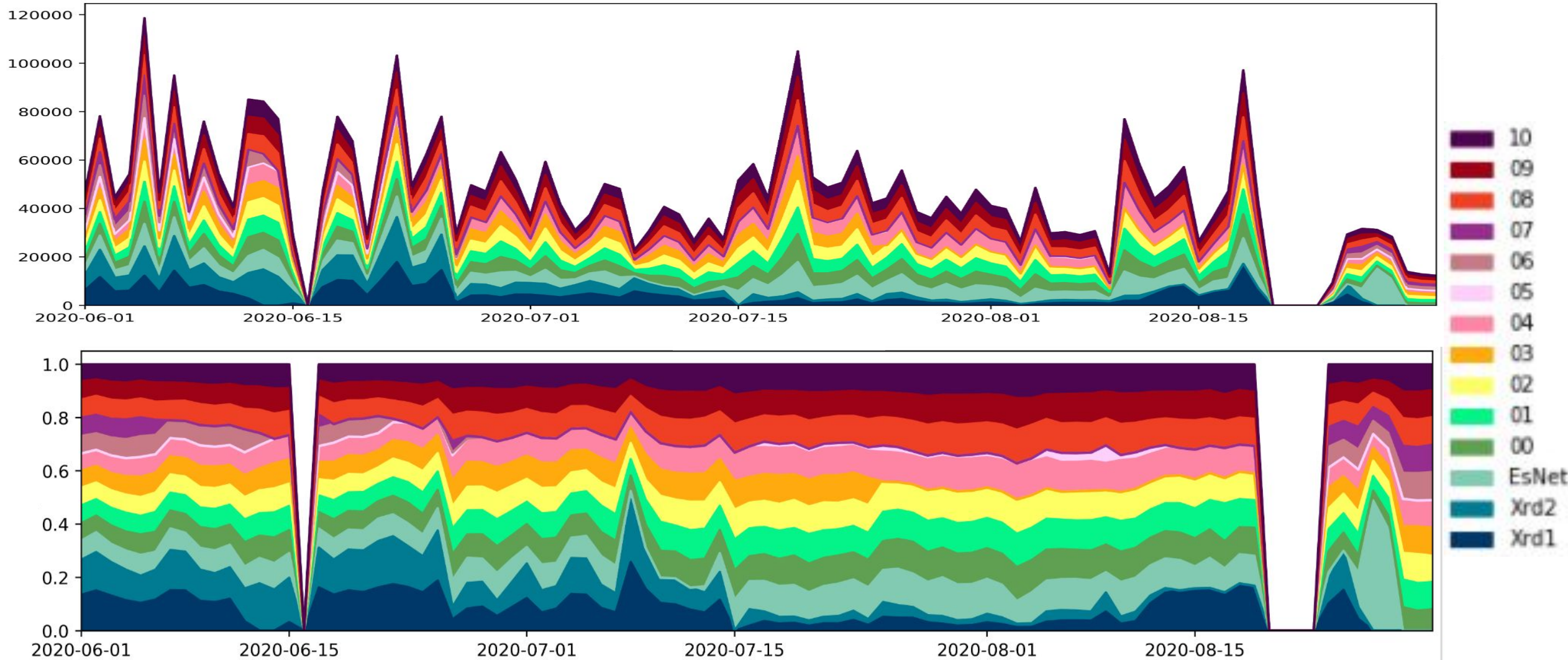
Cache utilization

Summary statistics for regional cache repo

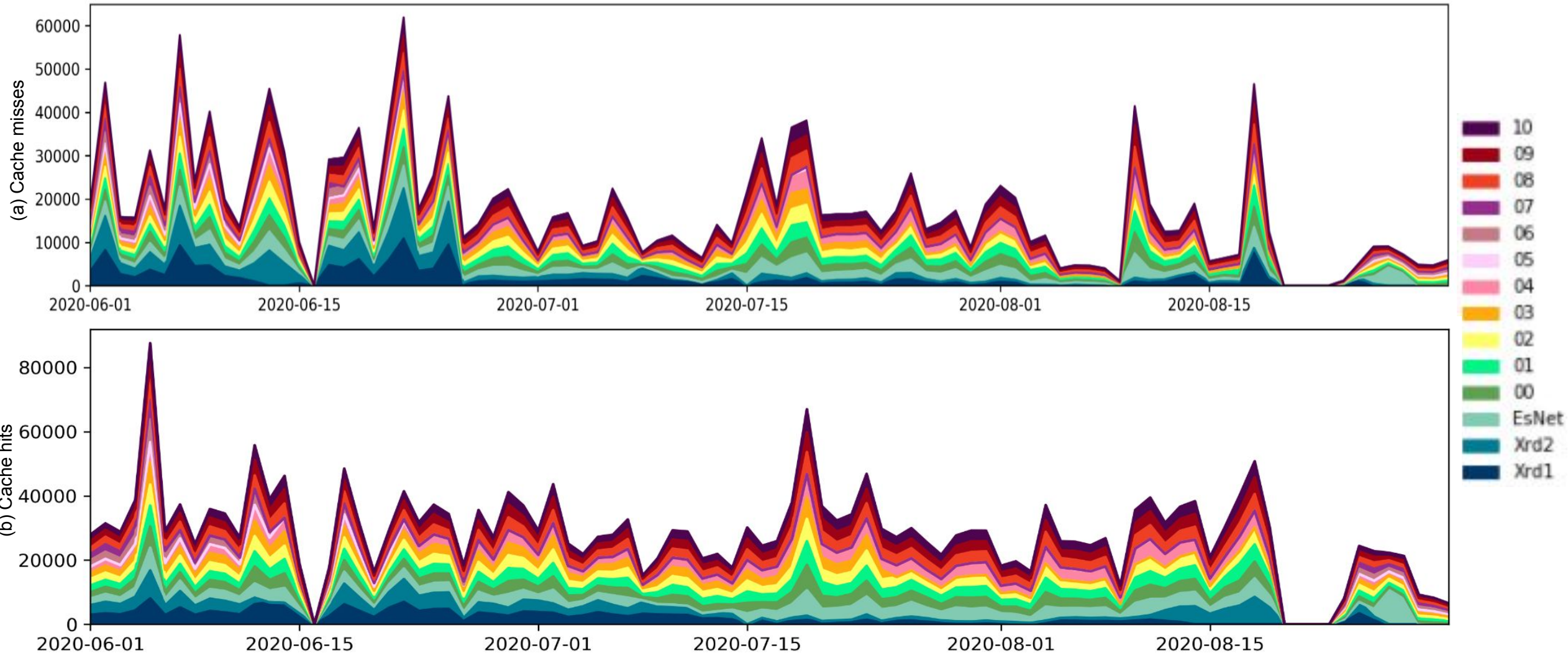
	Number of accesses	Data transfer size (TB)	Shared data size (TB)	Percentage of shared data size
June 2020	1,804,697	532.04	818.96	60.62%
July 2020	1,426,585	354.45	764.35	68.32%
Aug 2020	995,324	249.58	586.19	70.14%
Total	4,226,606	1,136.07	2,169.49	65.63%
Daily average	48,029.61	12.91	24.65	

- Data transfer size (= first time data access size, cache misses): From remote sites to the local node cache
- Shared data access size (= repeated data accesses, cache hits, network bandwidth savings): From the local node cache to the application, excluding the first time accesses (data transfers)

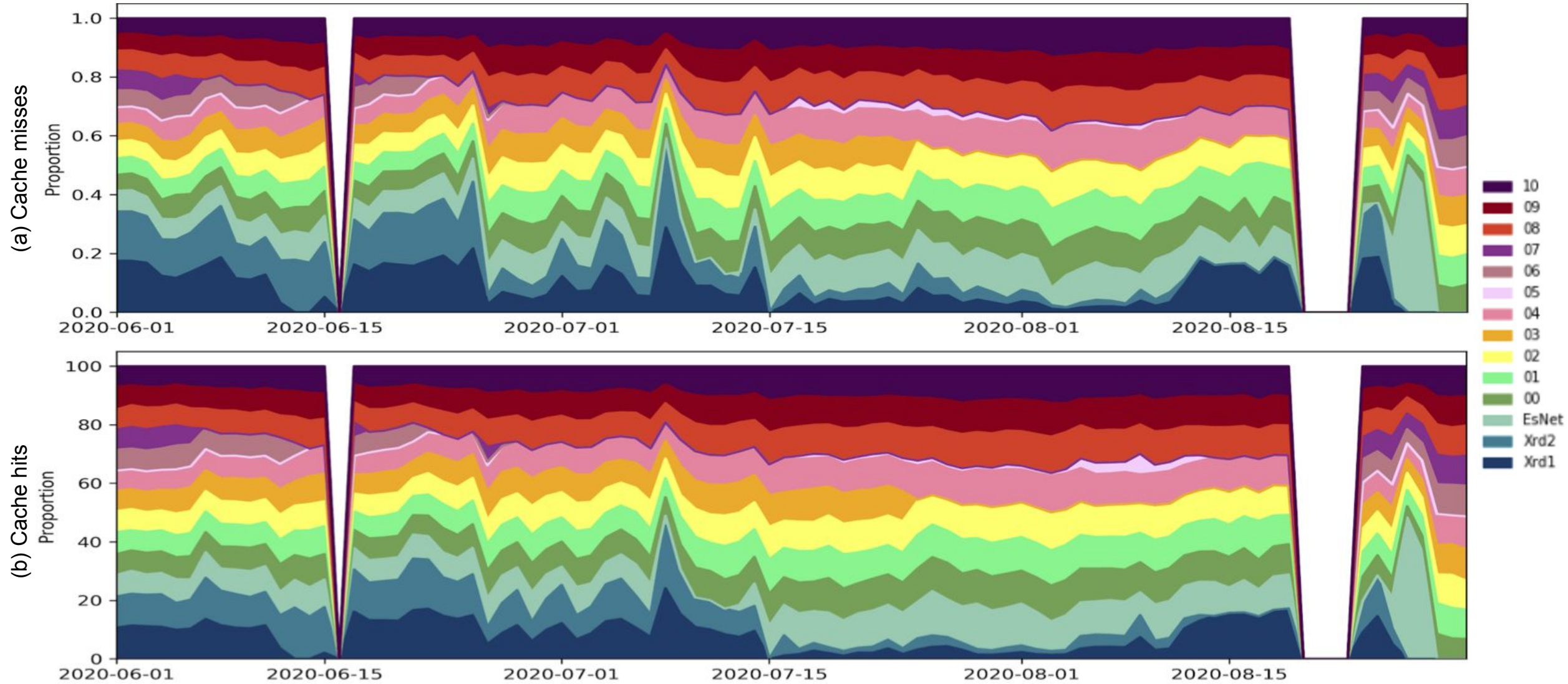
Daily total number of data accesses



Daily number of cache misses and hits

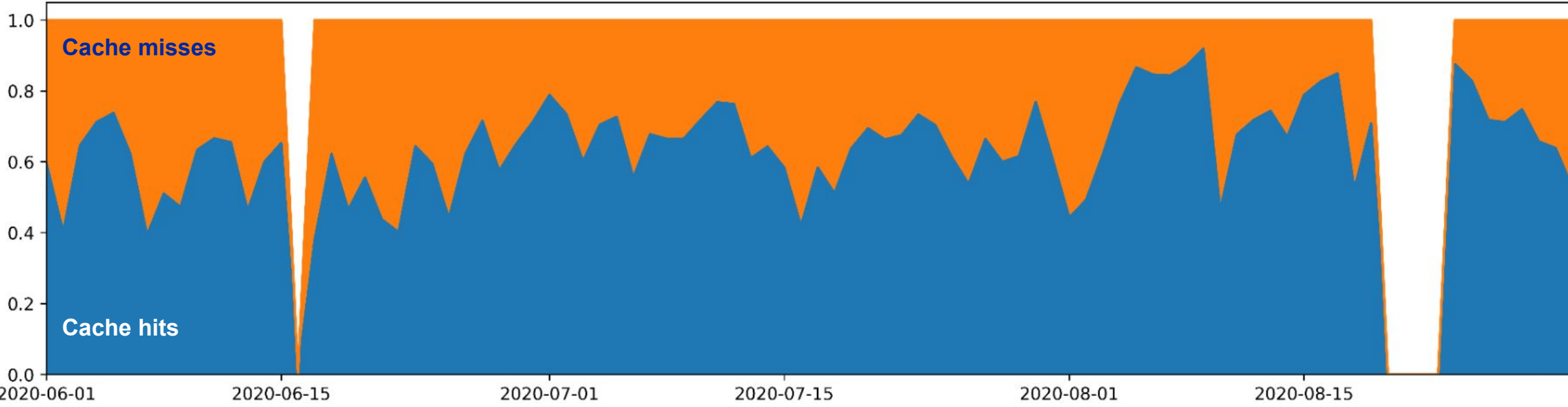


Number of cache misses and cache hits





Daily proportion of number of cache misses and cache hits

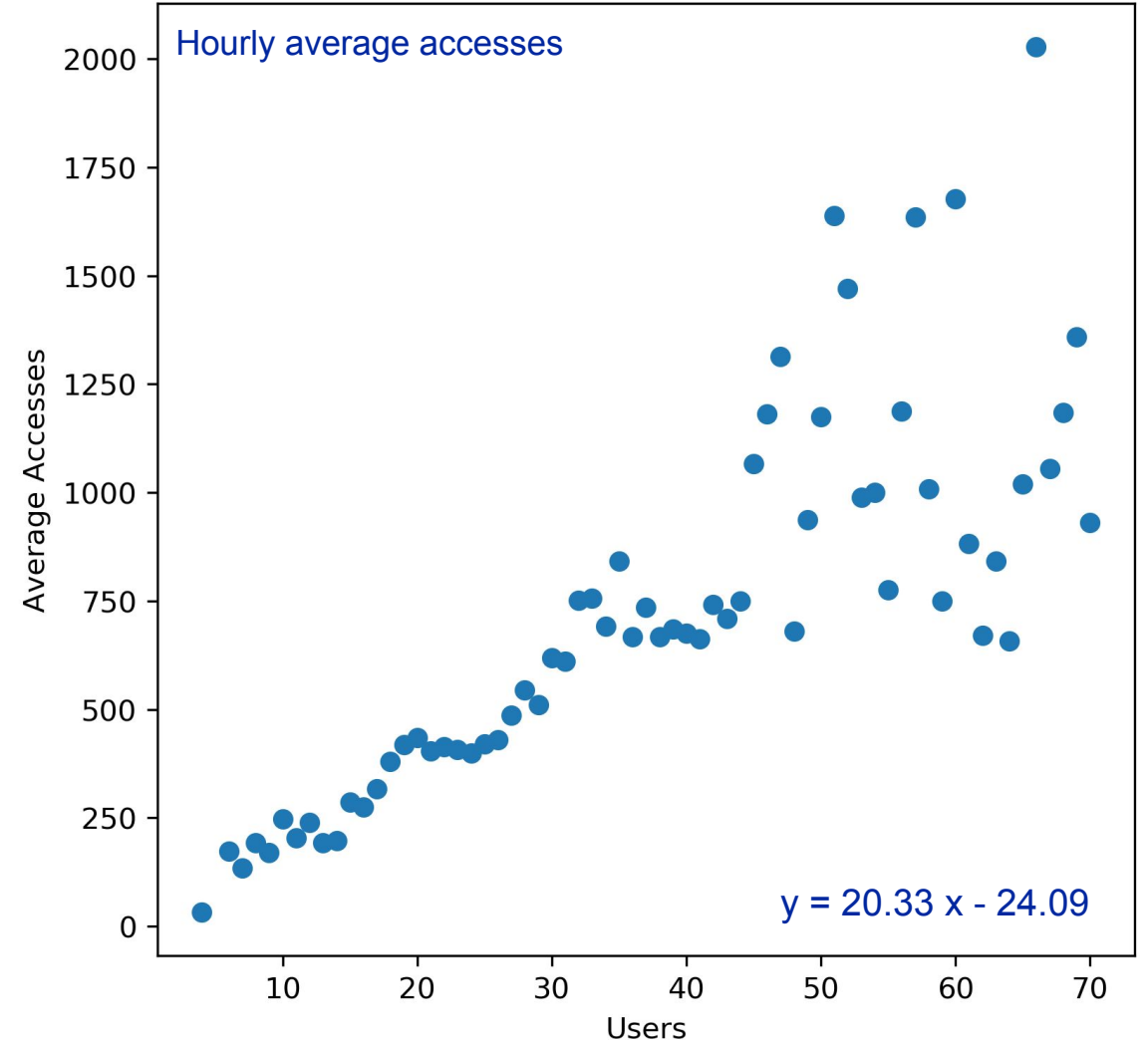
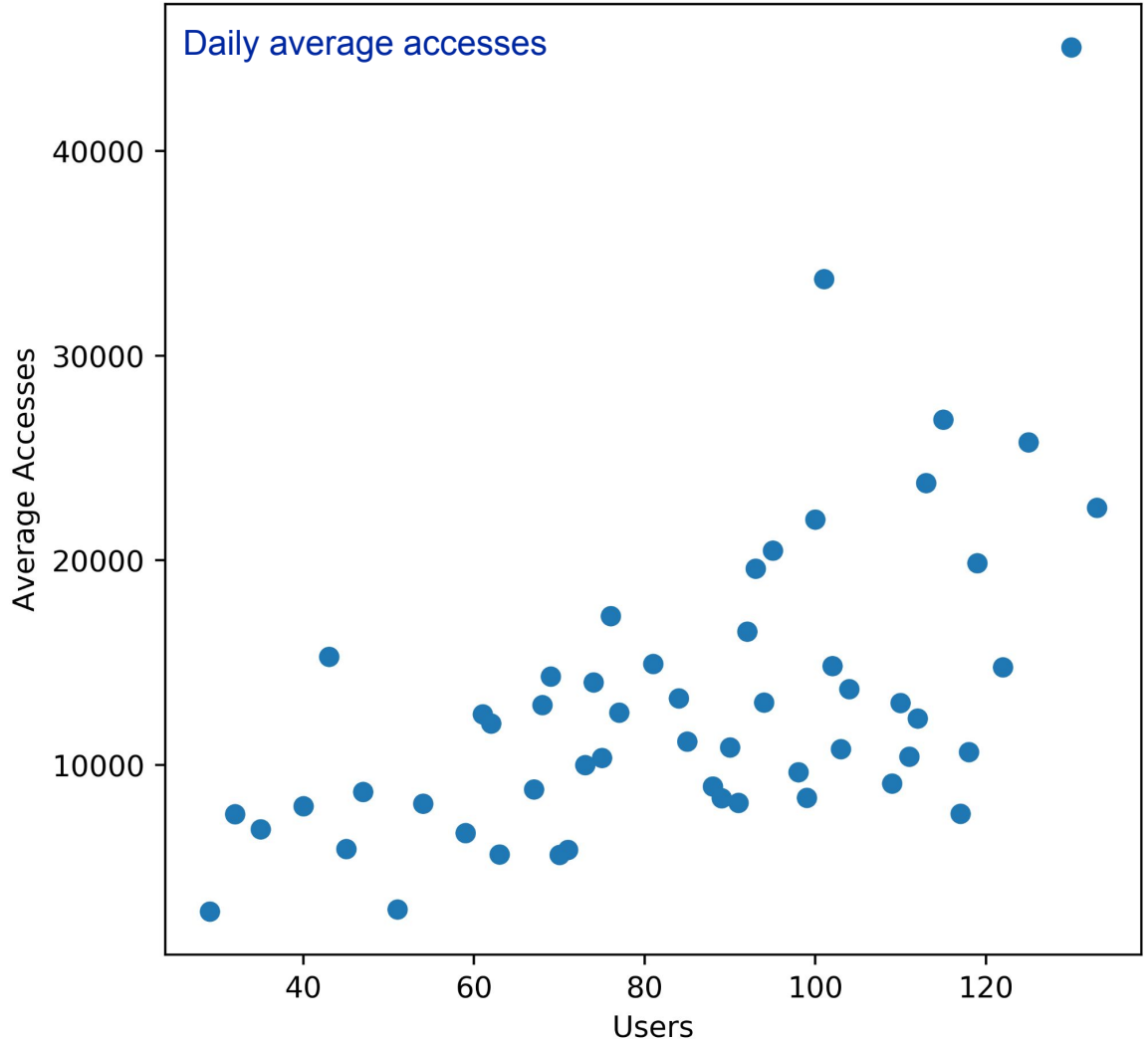


Network transfer savings = ~2.6 million transfers

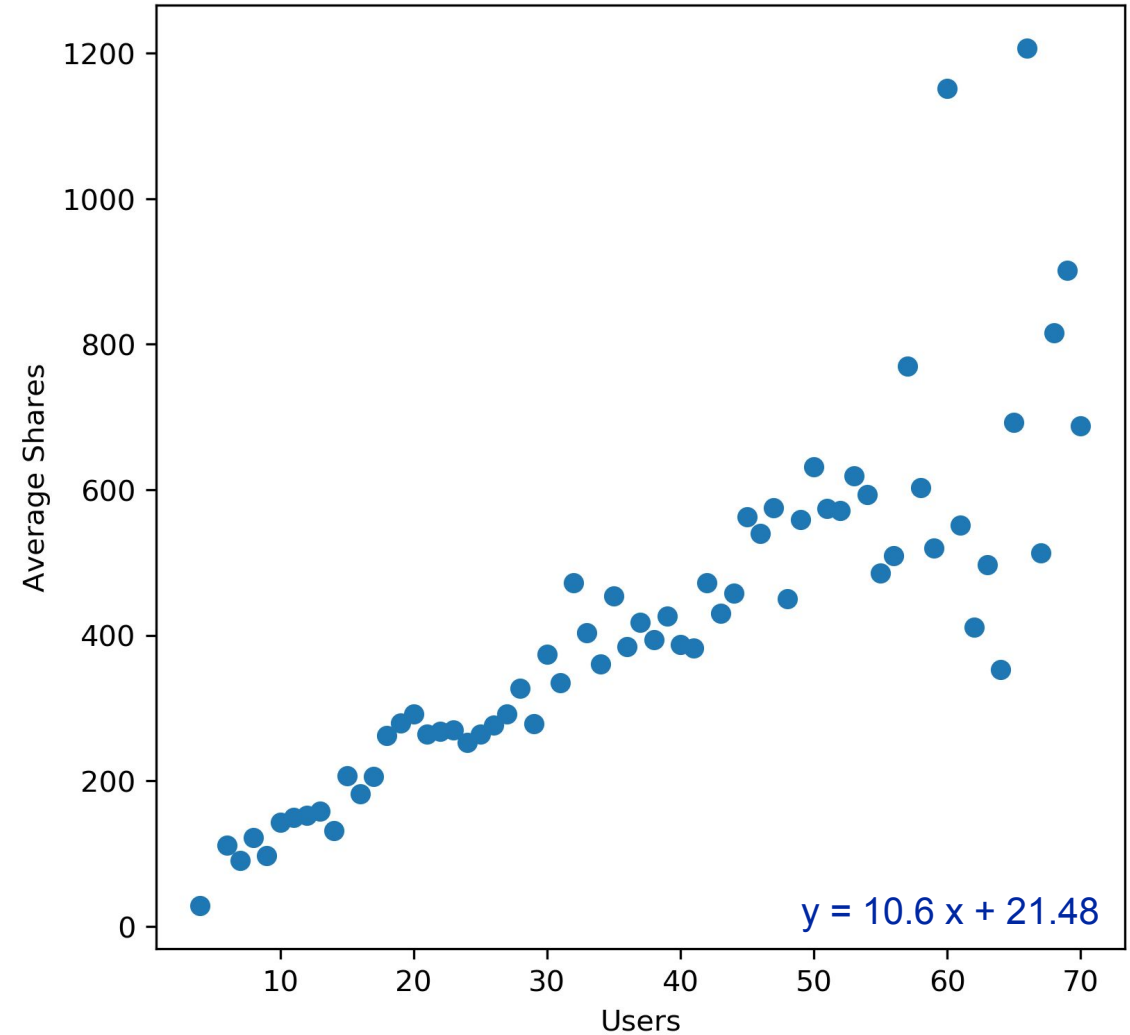
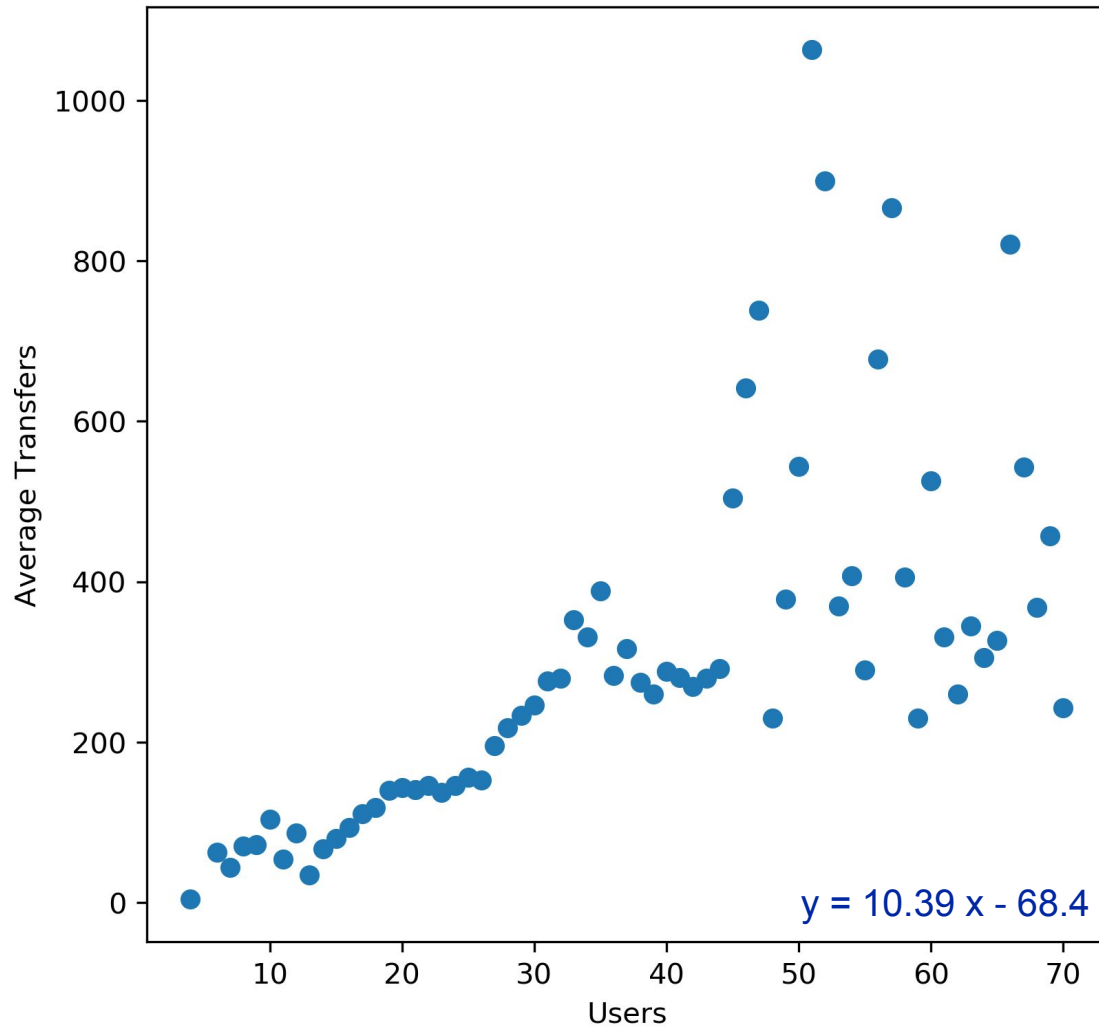
Network demand frequency reduction rate = (Cache misses + Cache hits) / Cache misses = **2.617**



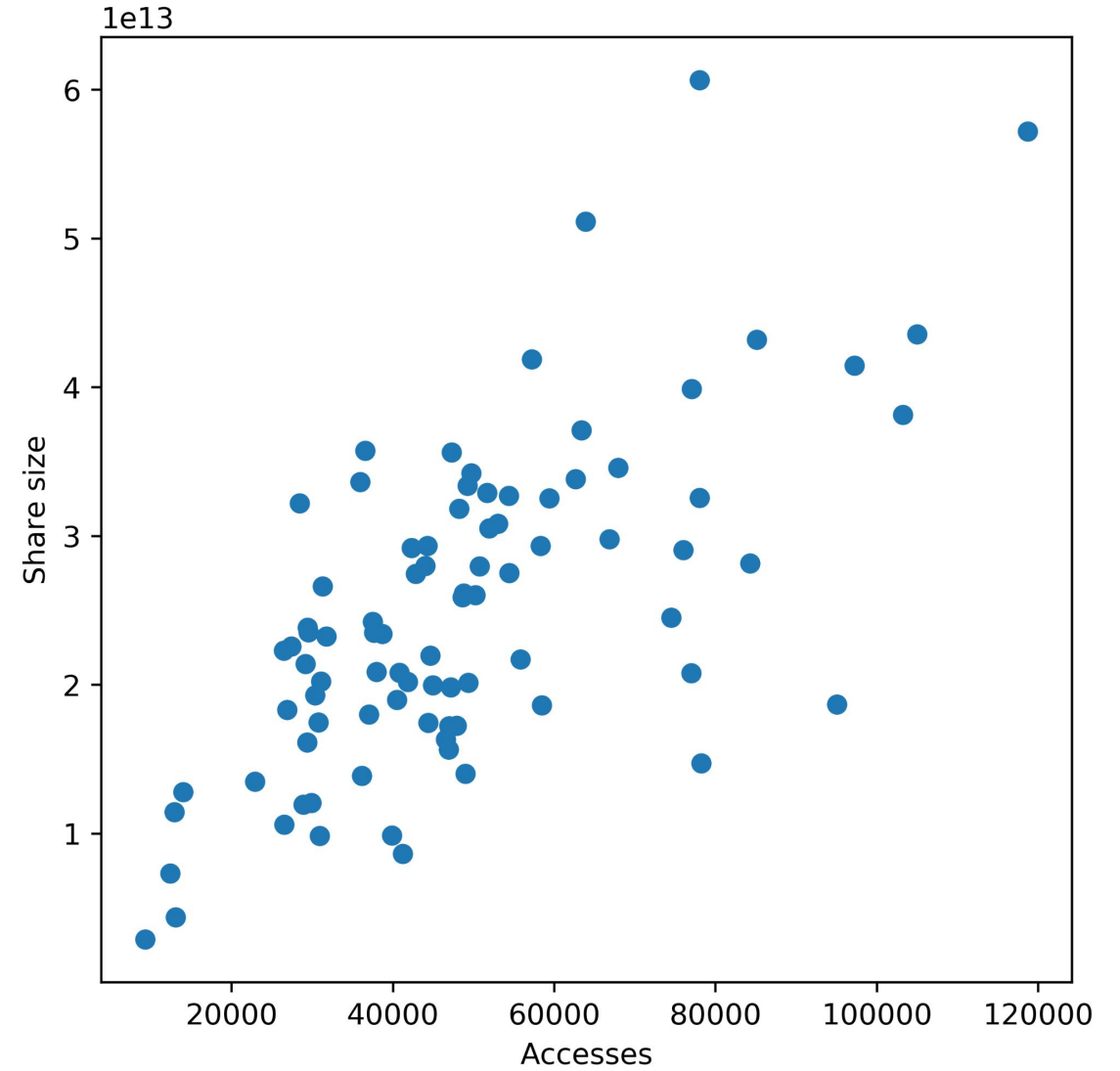
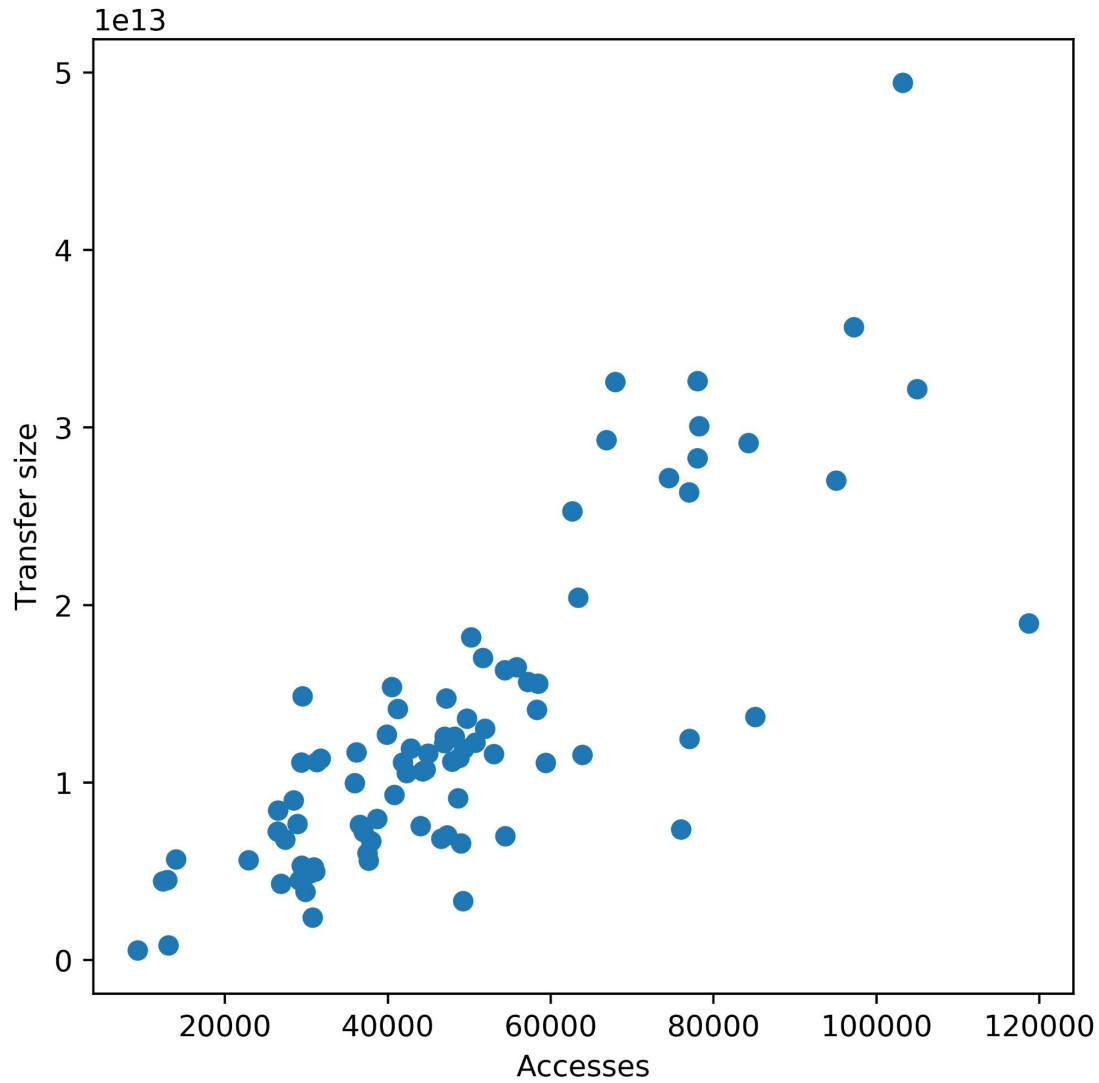
Distribution of number of users and average data accesses



Distribution of number of users and average transfer and shared data accesses



Distribution of the data transfer and shared data sizes

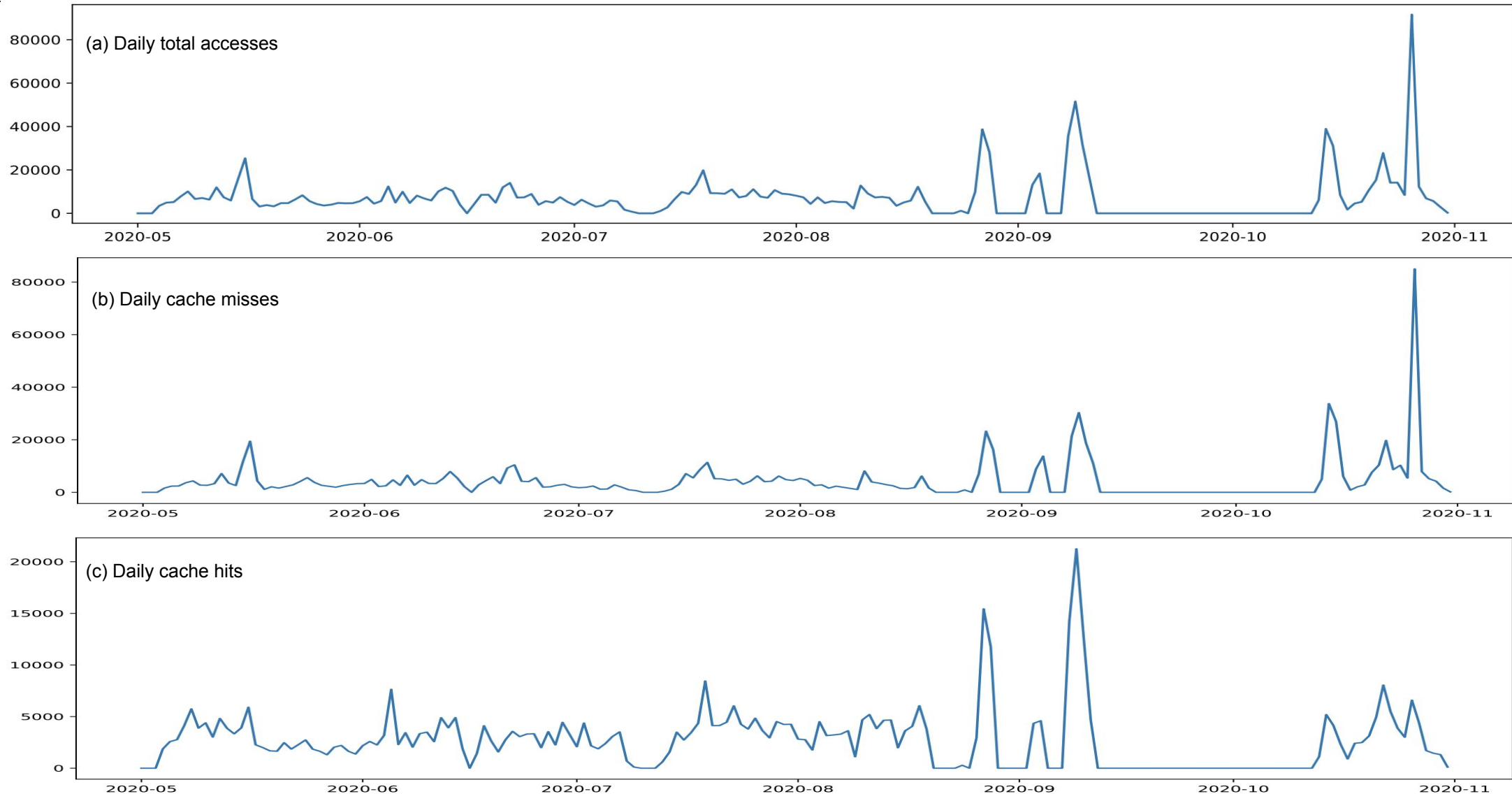


Summary statistics for ESnet node

	Number of accesses	Data transfer size (GB)	Shared data access size (GB)
May 4-31, 2020	189,984	30,150.50	47,986.56
June 2020	215,452	40,835.23	55,929.47
July 2020	205,478	33,399.81	66,457.35
Aug 2020	203,806	30,819.80	68,723.19
Sep 2020	165,910	10,153.97	38,036.19
Oct 2020	306,118	22,723.93	45,614.91
Nov 2020	276	3.33	47
Dec 2020	8514	1236.81	4523
Total (May-Oct)	1,286,748	168,083.27	322,747.67
Daily average	9,674.79	1,263.8	2,426.67

- Total number of active days until the end of Oct 2020: 133
- Monitoring issues from later Sep to Dec that measurements were not collected properly

Number of data accesses over time on ESnet node

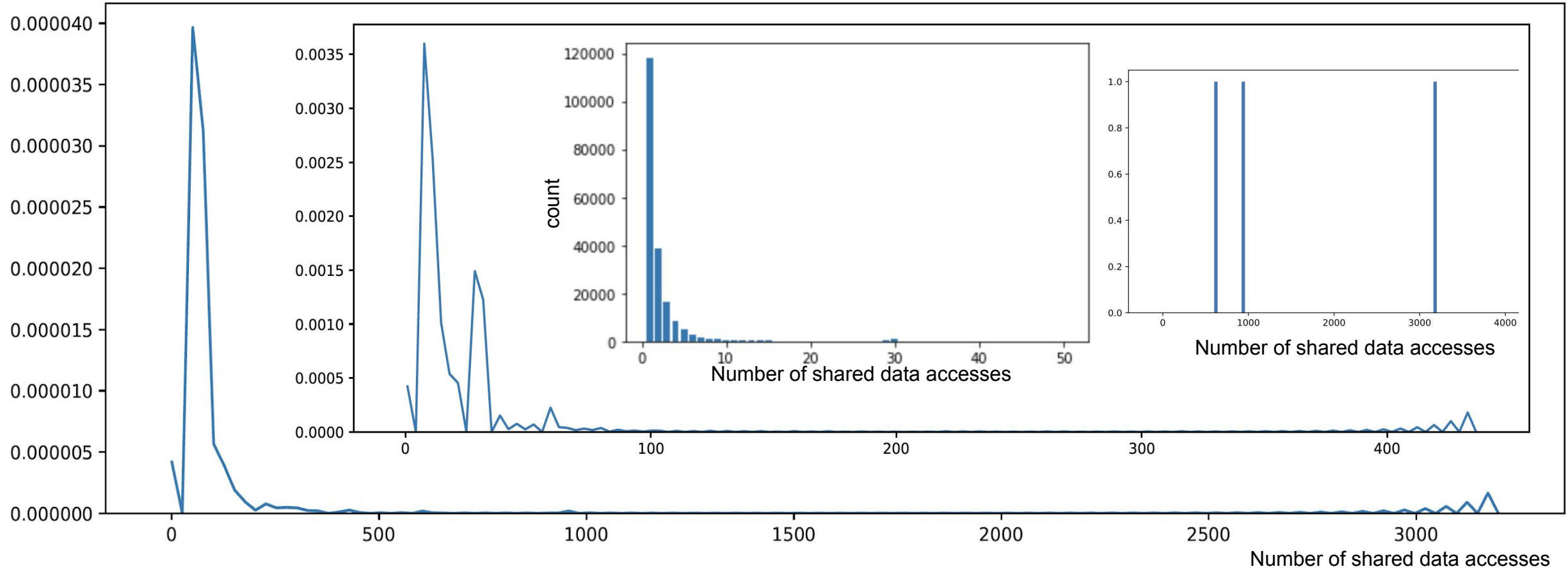


Notes on data transfer spike on ESnet node in 10/2020

	data_user	data_transfer	data_zero_operation_time	metadata_id	data_file_lfn		data_read_bytes		data_file_size		data_operation_time
					count	nunique	count	max	sum	max	sum
0	User 1	False	False	1	1	1	775856549	775856549	775856549	775856549	325.000000
1	User 1	True	False	1	1	1	0	0	729145901	729145901	621.000000
2	User 2	True	False	8	8	8	0	0	10921350	20772110	83.500000
3	User 2	True	True	1	1	1	0	0	0	0	0.000000
4	User 3	True	False	14	14	14	0	0	459441739	726622941	270.285714
5	User 3	True	True	2	2	2	0	0	0	0	0.000000
6	User 4	False	False	45	45	45	13723775	433735867	13723775	433735867	12.444444
7	User 4	True	False	11763	6547	11763	0	0	14537889	19644221225	6.160928
8	User 4	True	True	3295	2757	3295	0	0	0	0	0.000000
9	User 5	True	False	1	1	1	0	0	0	0	1152.000000

- 12pm-1pm, October 26, 2020, 15058 transfer operations for one user
- 3295 records has 0 file transfer size and 0 transfer time.
- For the other 11,763 transfer records, 6547 unique files are requested.

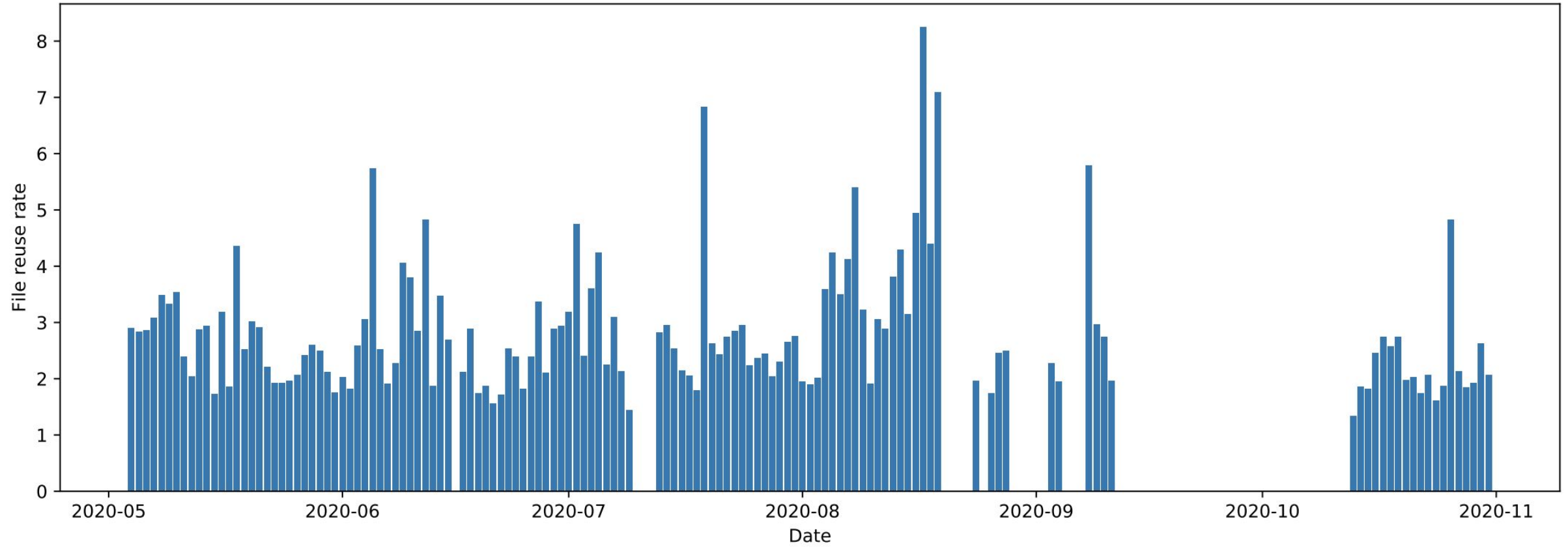
Distribution of the shared data accesses on ESnet node



- Distribution of the shared data access count during 05/2020-10/2020
total shared access count=490,944, unique file count=198,940
- Distribution of the shared data access count (≤ 500), total shared access count=486,182, unique file count=198,937
- Density plot of shared data access count (> 500), total shared access count=4,762, unique file count=3



Daily data file reuse rates for ESnet node



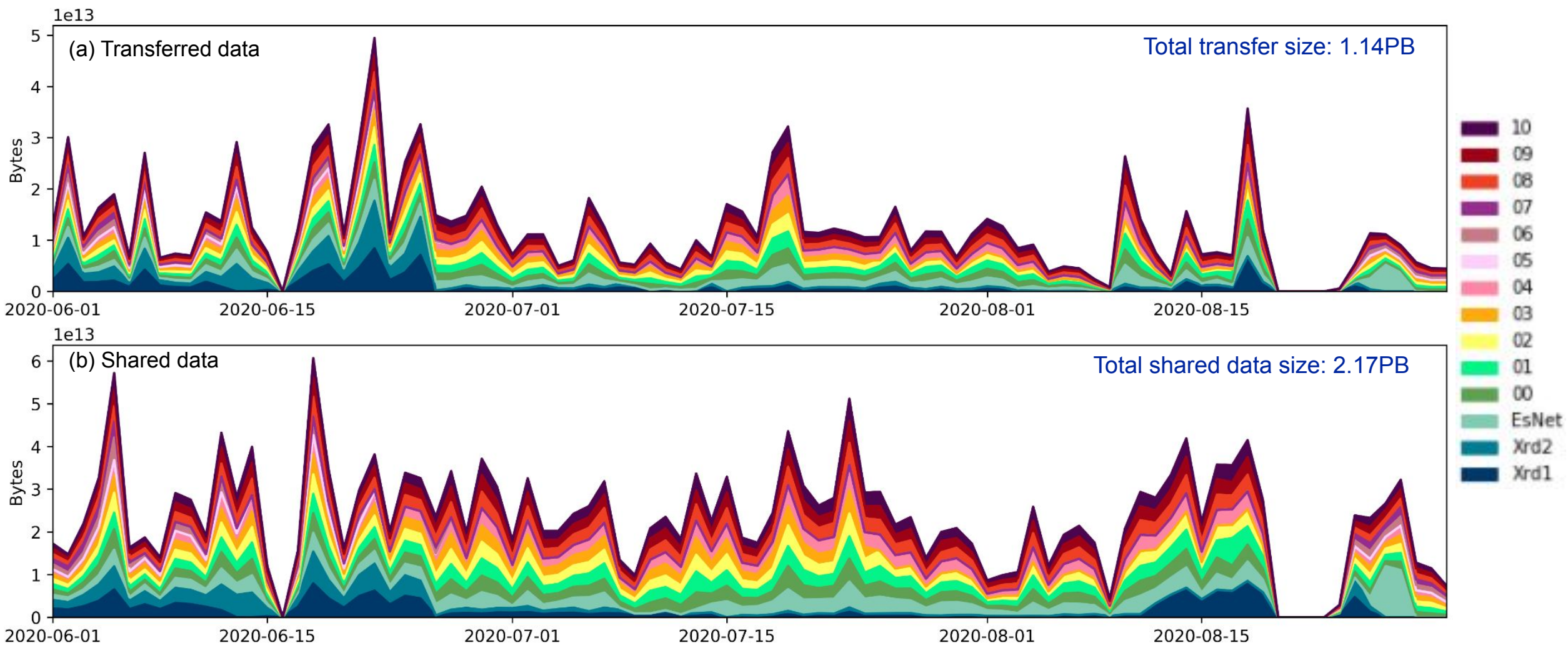
- Data file reuse rate = (sum of accesses) / (total number of unique files)
 - Sum of accesses = all accesses for the day
 - Total number of unique files = number of unique files for the day



Network utilization

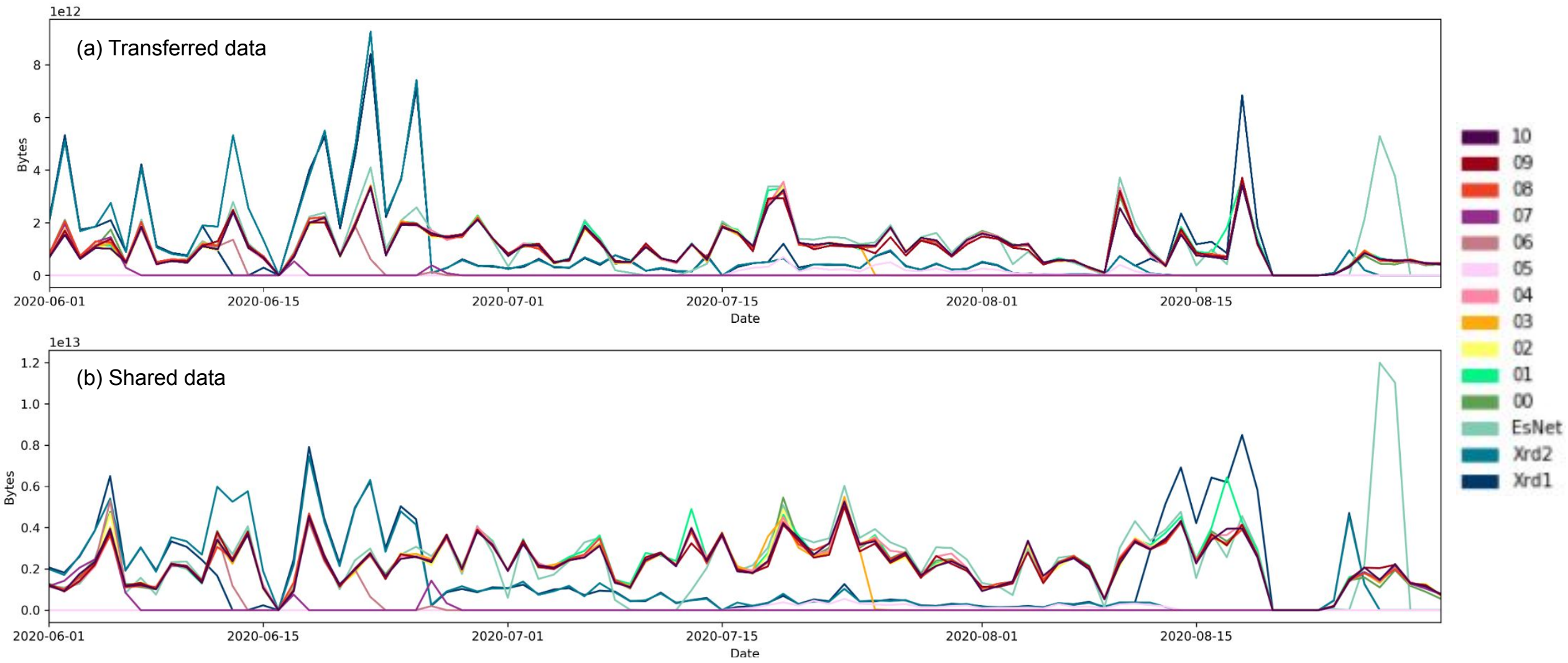


Daily data access sizes

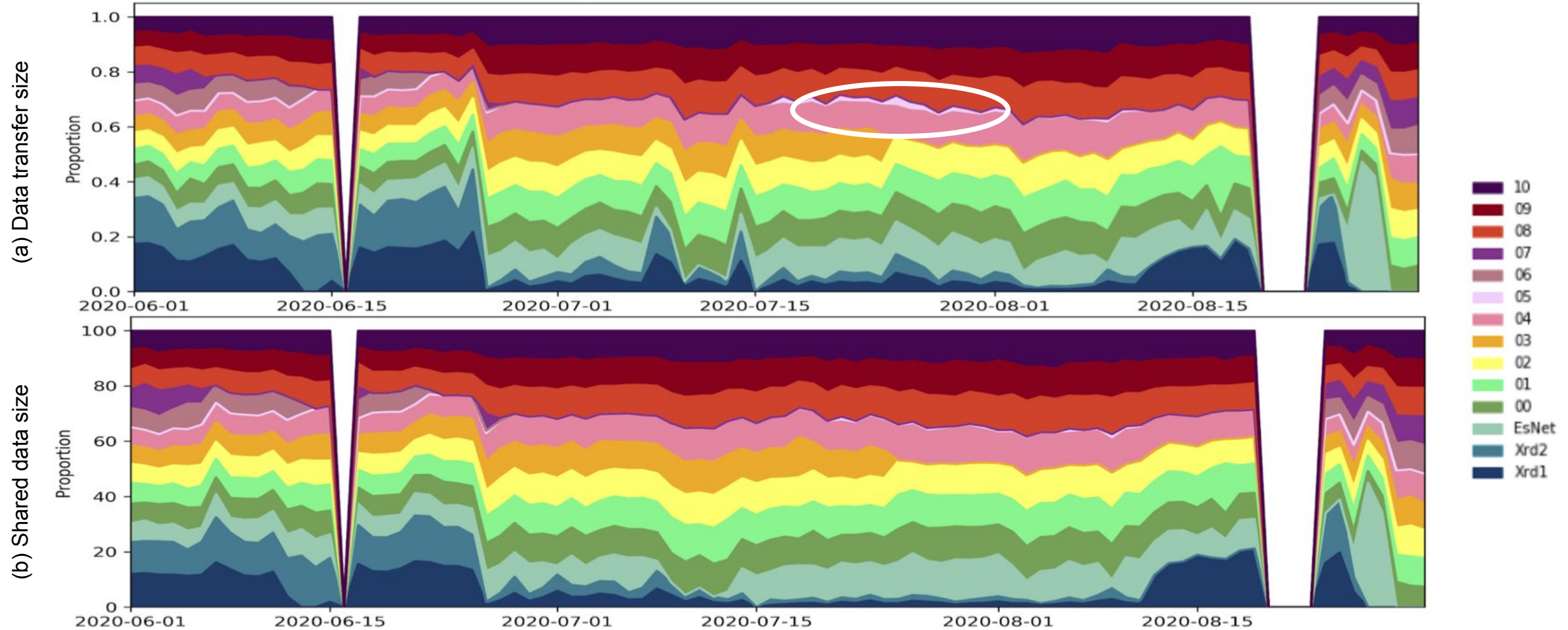




Daily data access sizes



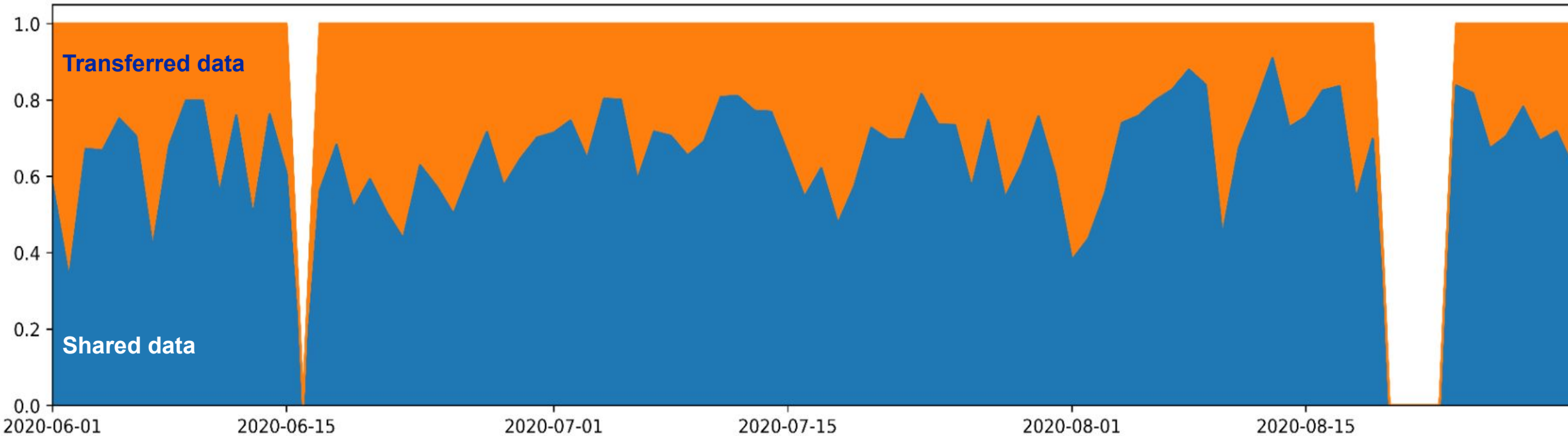
Data transfer size and shared data size



Two larger nodes have higher transfer proportion than share proportion, may be due to redirector policy



Daily proportion of data transfer sizes and shared data sizes

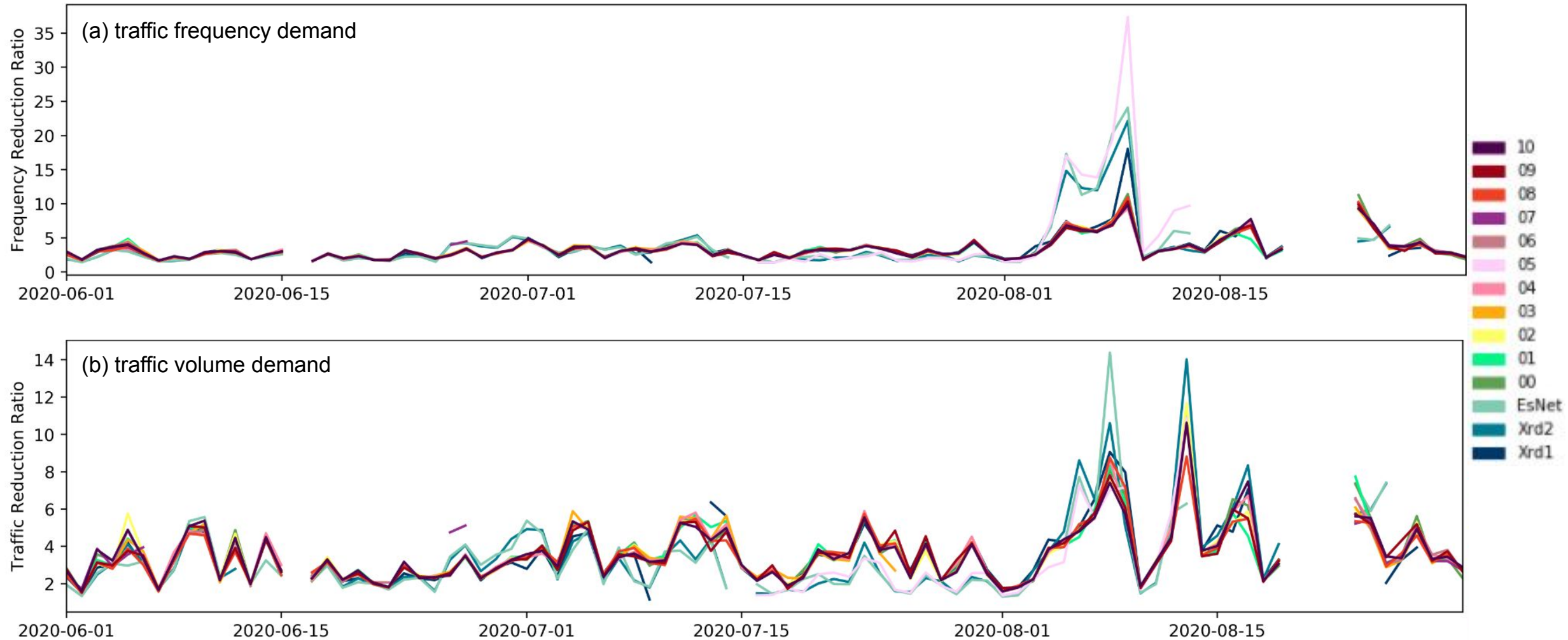


Network traffic volume savings during the study period = 2.17 PB

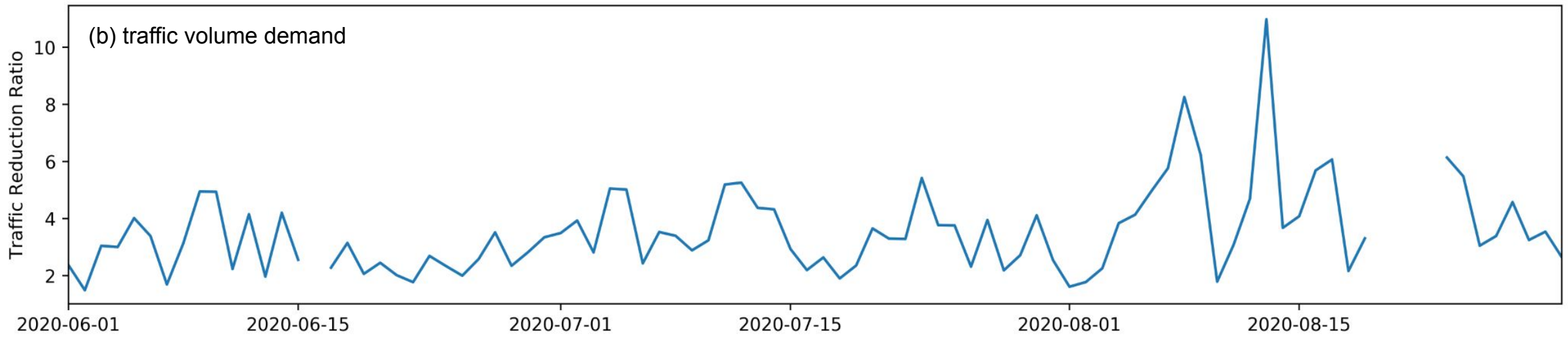
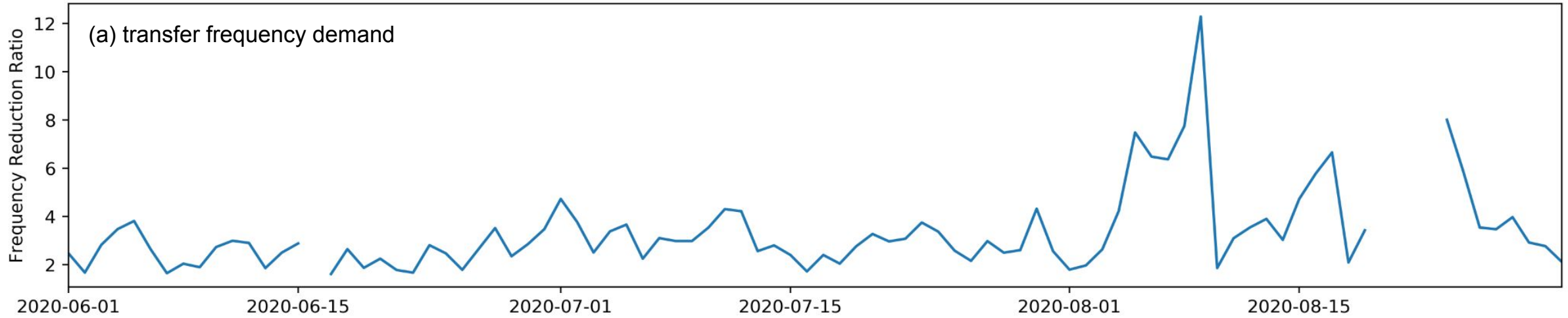
Network demand traffic reduction rate = (sum of total shared data + sum of total transfer data) / (sum of total transfer data)

Network demand traffic reduction rate = **2.9096**

Daily network demand reduction rate (each node)



Daily network demand reduction rate (all nodes)



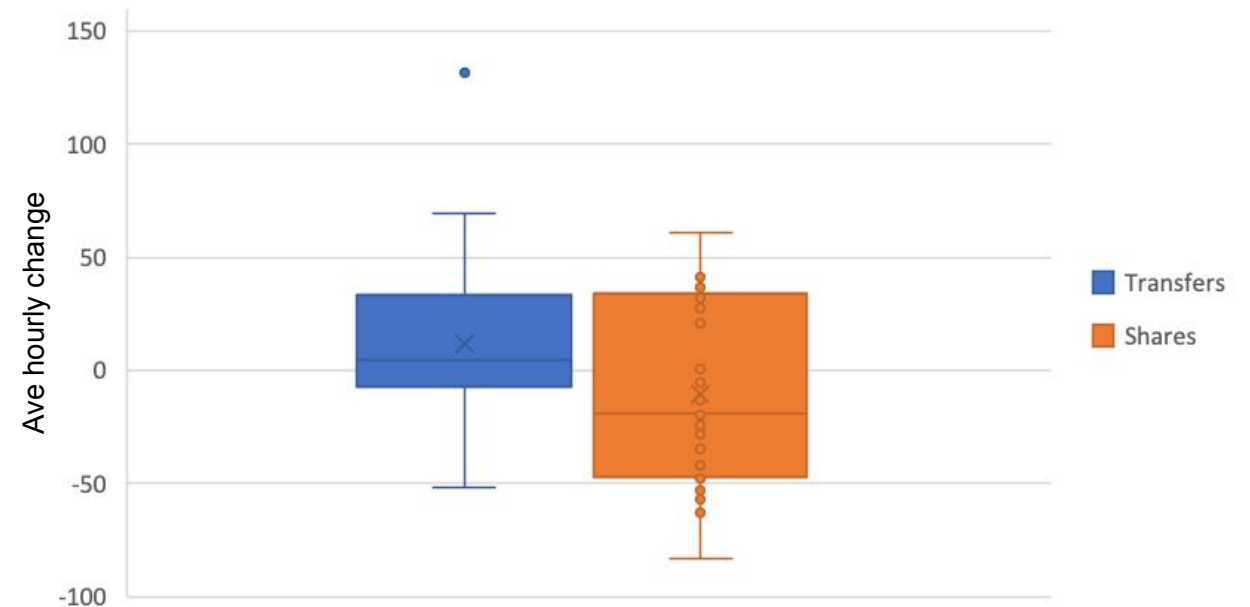
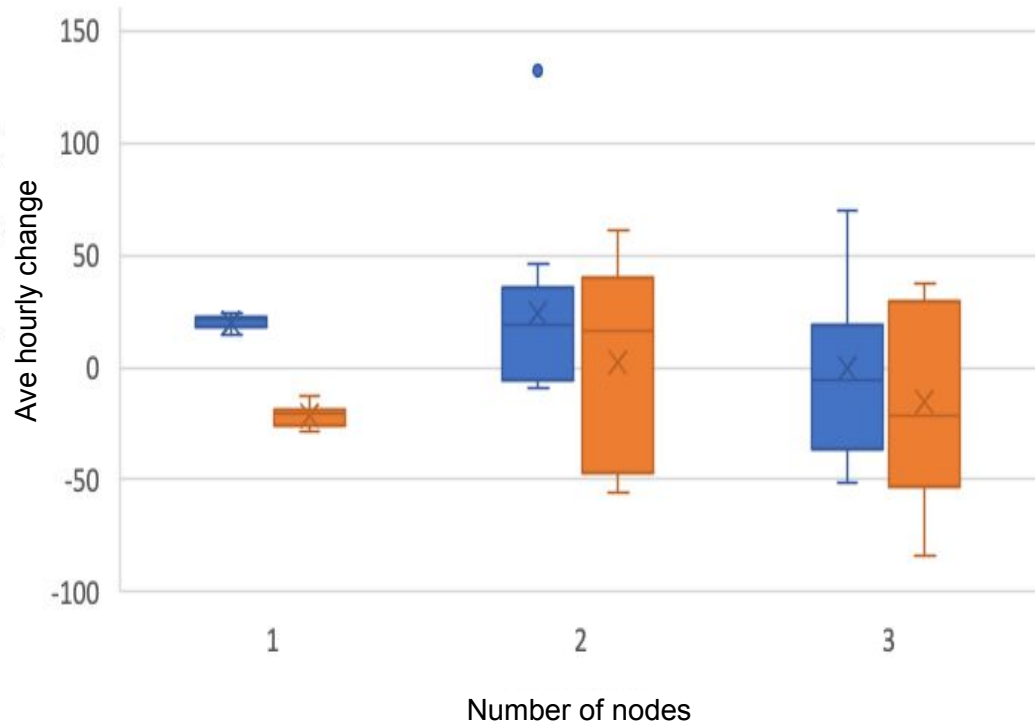


Impact of a cache node

Downtime of cache nodes

- **XCache nodes went down occasionally throughout the summer**
 - Identified all times that a node had 0 accesses over a full hour or longer
 - Grouped downtimes based on how many nodes were already down
- **Downtime effectively reduces number of nodes**
 - Allows to study the effects of adding/subtracting nodes from the network
 - How proportional data accesses changed
 - Number of cache misses and cache hits
 - Data transfer sizes and shared data sizes
- **Observations**
 - **Proportional loads**
 - Remaining nodes tend to split proportional load of the downed node
 - Particularly for access proportions
 - ESnet node proportion increases more
 - **Larger nodes**
 - Downtimes of larger nodes (Xrd1 and Xrd2) did not act as expected compared to other downtimes

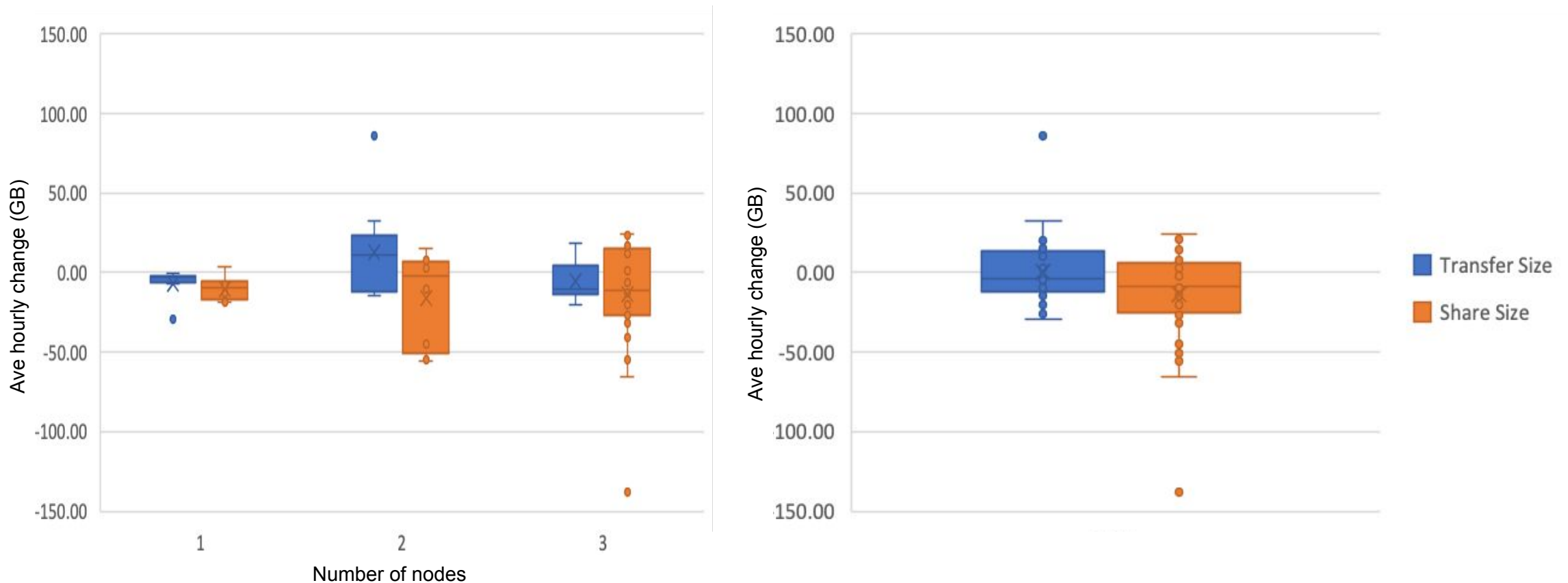
Change of effects of node down in data accesses



Avg hourly change over **all** downtimes studied: Shares: -58.64, Transfers: +105.67

Avg hourly change **without** data from two larger nodes: Shares: -264.59, Transfers: +67.81

Change of effects of node down in data access sizes



Avg hourly change over all downtimes studied: Share size: +21.4GB, Transfer size: +77.3GB

Avg hourly change without downtimes of two larger nodes: Shares: Share size: -149GB, Transfer size: +6.76GB

Load estimation

- **What would happen when the number of users double to the regional cache?**
 - Number of data accesses would double
 - Number of data transfers & data shares would roughly double
 - Number of data shares likely to increase more as more files are cached
 - Data transfer size slightly less than doubles
 - Shared data size slightly more than doubles
- **Effects of adding a node**
 - Evenly takes loads off of other nodes
 - Evidence that further research will indicate a higher proportion of data being shared as the number of accesses/nodes increase
 - Will reduce time needed to access data files on average
 - Larger nodes do not necessarily take proportionally larger loads
 - Able to hold data files in cache for longer
 - Consequently increasing the network demand reduction rate as time goes on

Summary

- **Shared data caching mechanism**
 - **Reduced the redundant data transfers, saved network traffic volume**
 - **Network transfer savings: ~2.6 million transfers**
 - **Network traffic volume savings during the study period: 2.17 PB**
 - **Network traffic demand reduced by a factor of ~3**
 - **Increases as time goes on and more files are cached**
 - **Data transfer frequency reduction rate: ~2.62**
 - **Proportion of shared data increases as more files are cached**
 - **Decreases average applications wait time**
 - **Decreases network loads**
- **Additional nodes allow more data files to be cached**
 - **Increase cache hits**
 - **Reduces data access latency and increases overall application performance**
- **Further studies**
 - **How are the repeated cache misses affecting the application performance?**
 - **Anomaly detection and prevention**
 - **How many XCache installations are needed for certain data access loads?**
 - **What size of each disk cache would be appropriate?**

Acknowledgements

- **Acknowledgements**

- Adam Slagell, Anne White, Eli Dart, Eric Pouyoul, George Robb, Goran Pejovic, Kate Robinson, Yatish Kumar at ESnet
- Dima Mishin, Michael Sinatra at UCSD
- Justas Balcas at Caltech
- ESnet Infrastructure team: System build, config, and management
- ESnet Engineering team: Network connectivity
- **LBNL work is supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Contract No. DE-AC02-05CH11231, and also used resources of the National Energy Research Scientific Computing Center (NERSC).**



- **UCSD work is supported by the National Science Foundation through the grants OAC-2030508, OAC-1836650, MPS- 1148698 and OAC-1541349.**





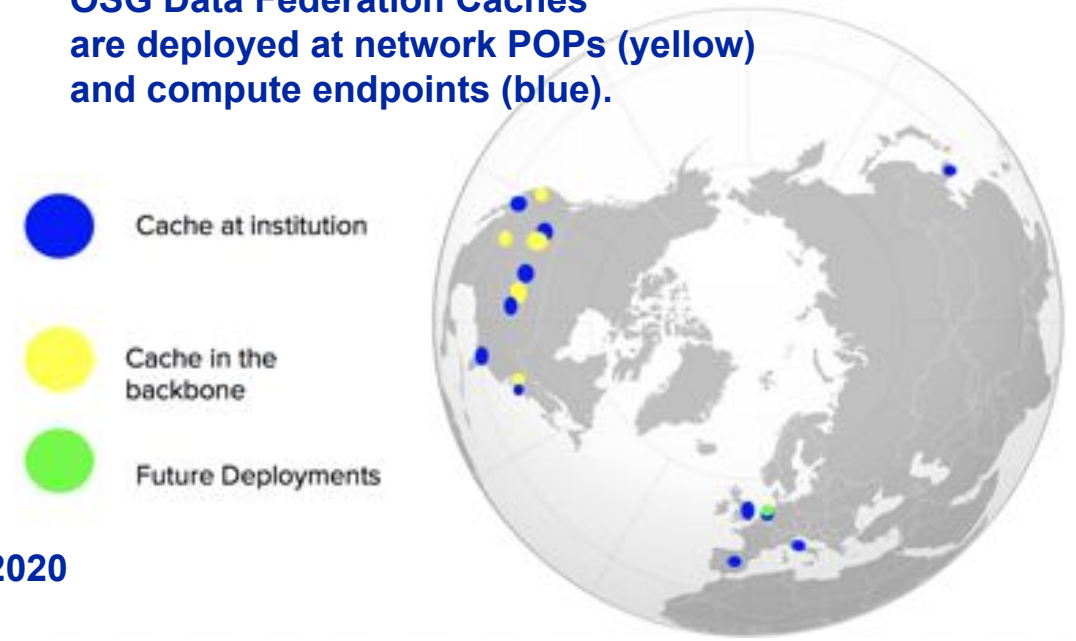
Backup

Data caching background

- **Diverse science relevant to DOE HEP & NP**
 - Regional storage repo and data caching are one of the hot topics at HSF/WLCG meetings
 - At present, there are caches in production for ATLAS, CMS, and OSG, all based on XRootD
 - OSG cache use dominated by Dune, LIGO, Virgo, MINERVA, DES, NOVA, and a liquid XENON detector R&D for future dark matter and neutrinoless double beta decay experiments. Electron Ion Collider R&D is in planning

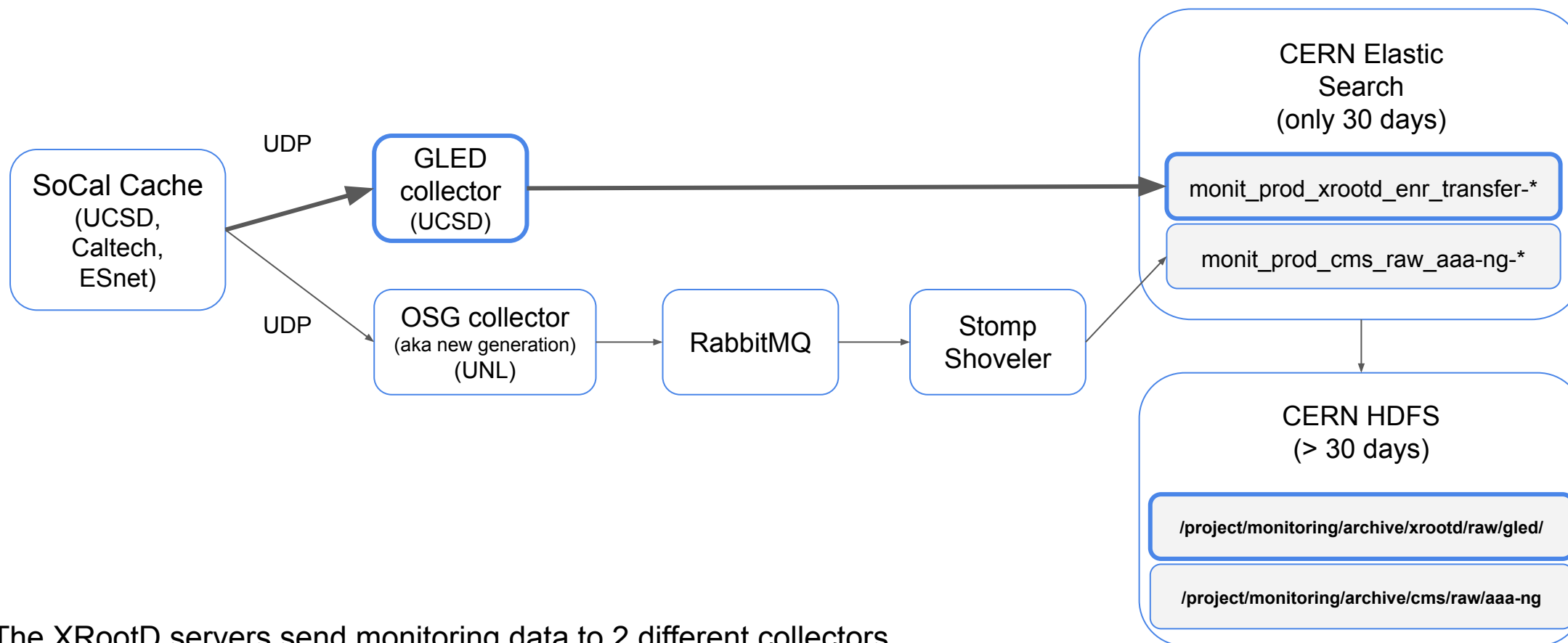
Collaboration	Working Set	Data Read	Reread Multiplier
DUNE	25 GB	131 TB	5400
LIGO (private)	41.4 TB	3.8 PB	95
LIGO (public)	4.3 TB	1.5 PB	318
MINERVA	351 GB	116 TB	340
DES	268 GB	17 TB	66
NOVA	268 GB	308 TB	1200
RPI_Brown	67 GB	541 TB	8300

OSG Data Federation Caches are deployed at network POPs (yellow) and compute endpoints (blue).



Data read from OSG data federation caches in 6 month period 3/2020-8/2020

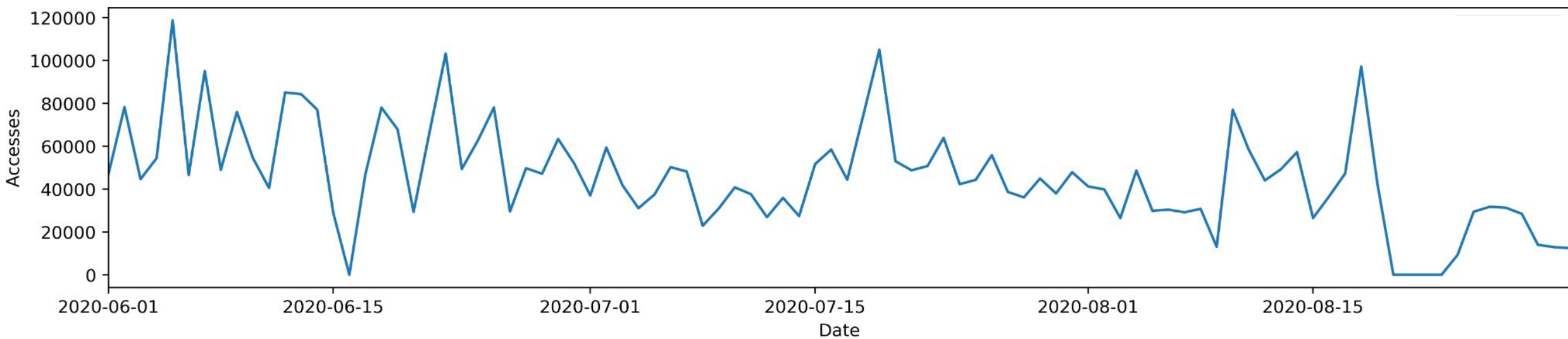
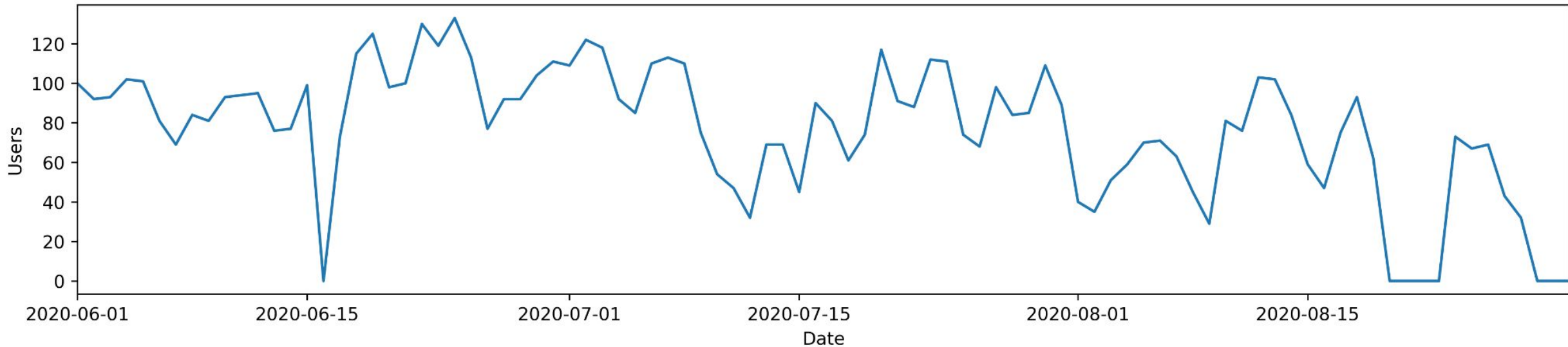
Current monitoring data path



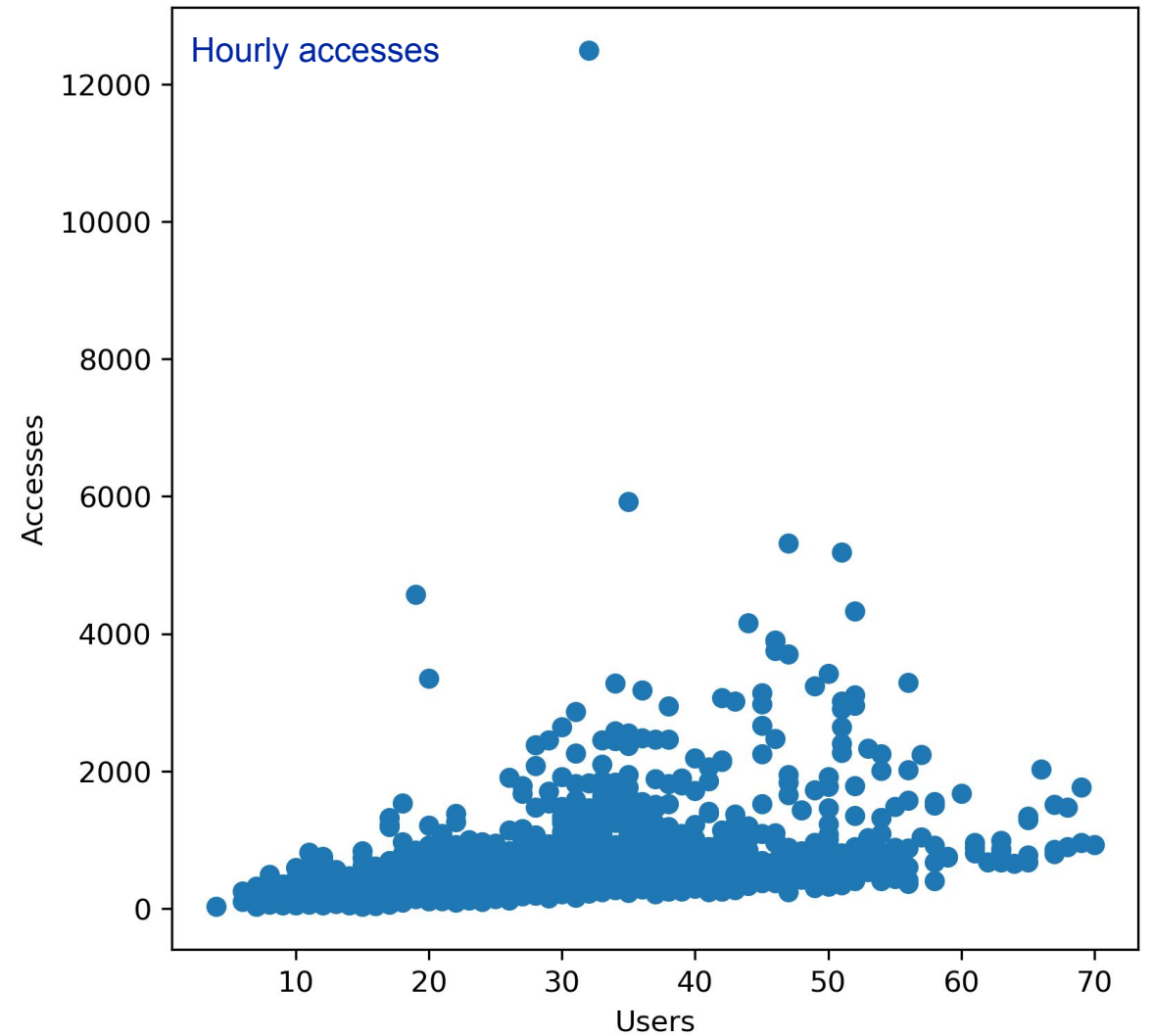
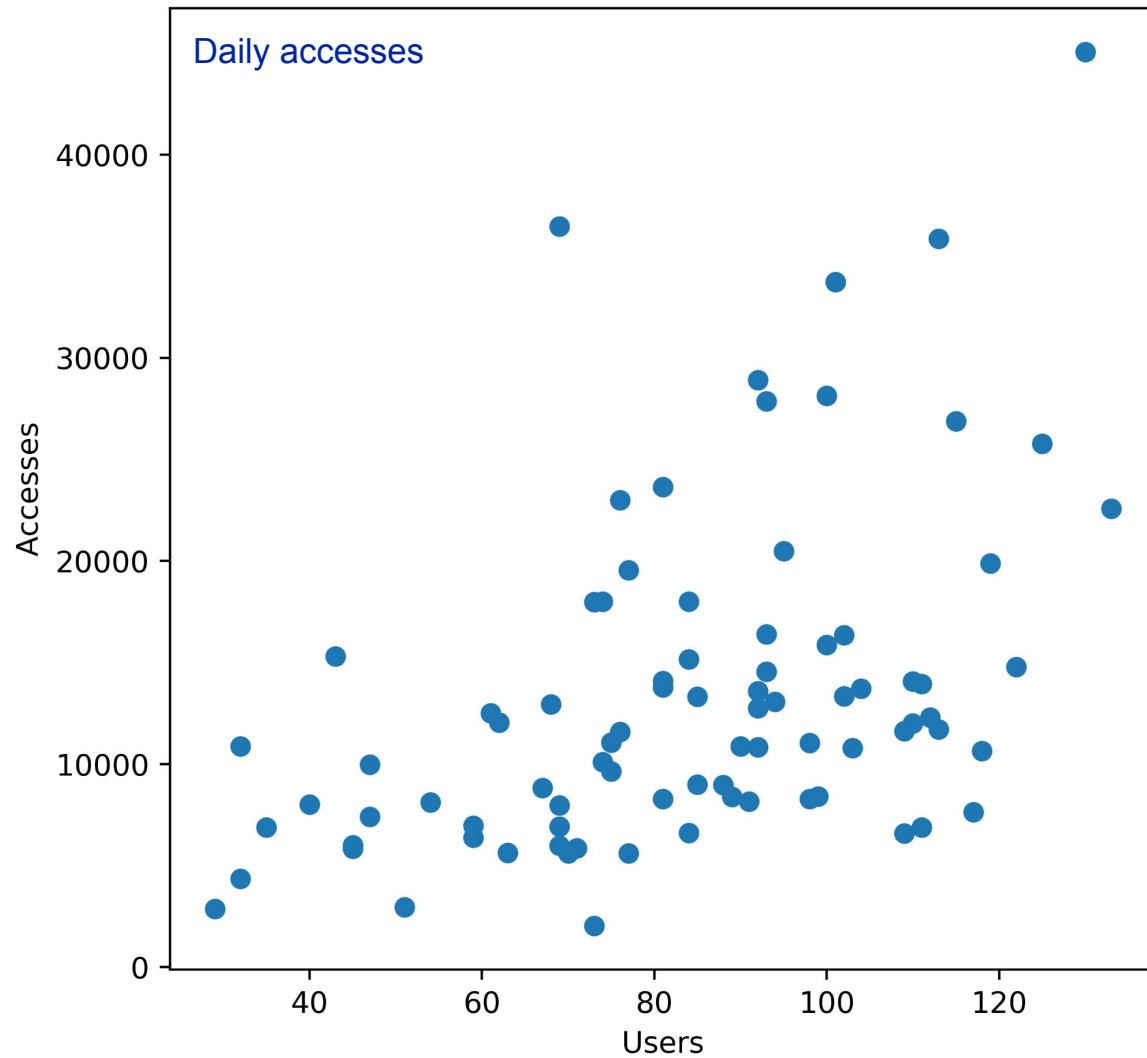
- The XRootD servers send monitoring data to 2 different collectors
- The data ends up in 2 different indexes at Elastic Search and kept there for a maximum of 30 days
- Periodically, the Elastic Search data gets stored in HDFS for long-term storage
- Given the proximity of the GLED collector to the SoCal cache, the chances for UDP packets being dropped are less on this path



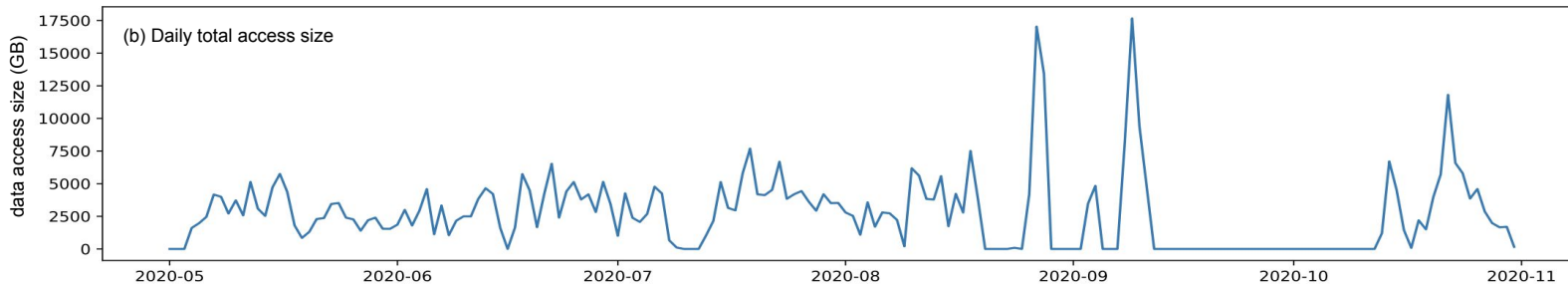
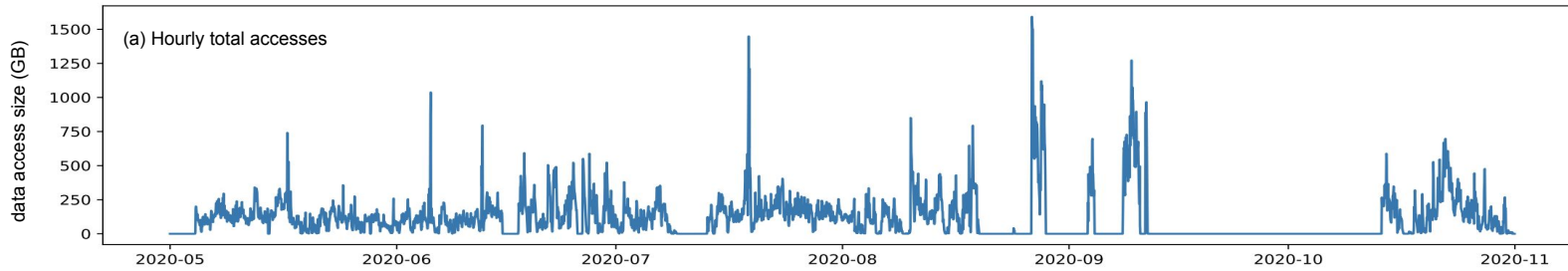
Daily number of users and data accesses



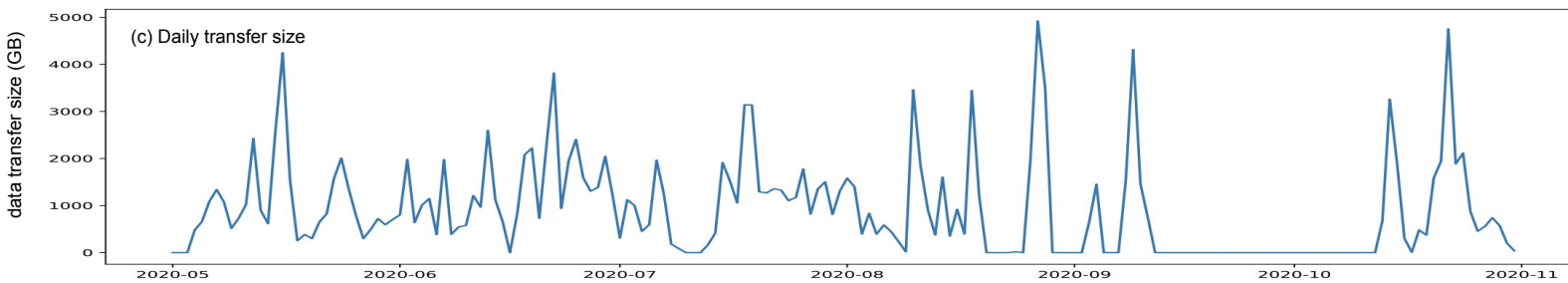
Distribution of number of users and data accesses



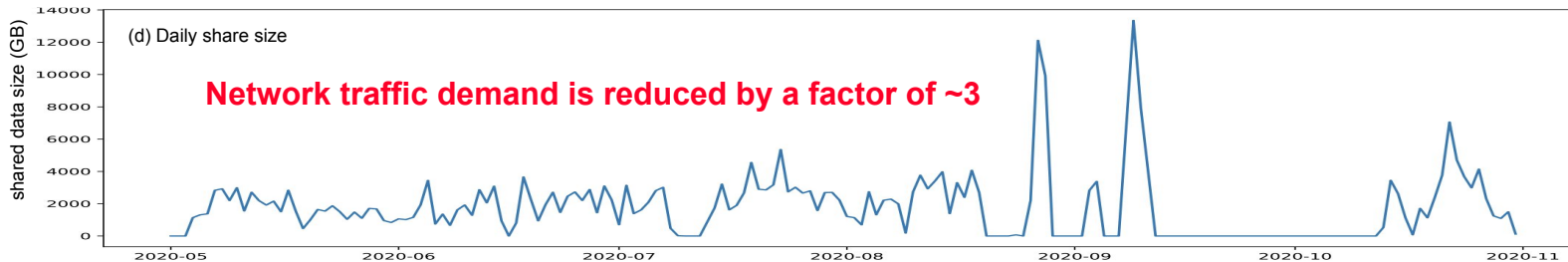
Total data access sizes over time on ESnet node



Total size=490.831TB

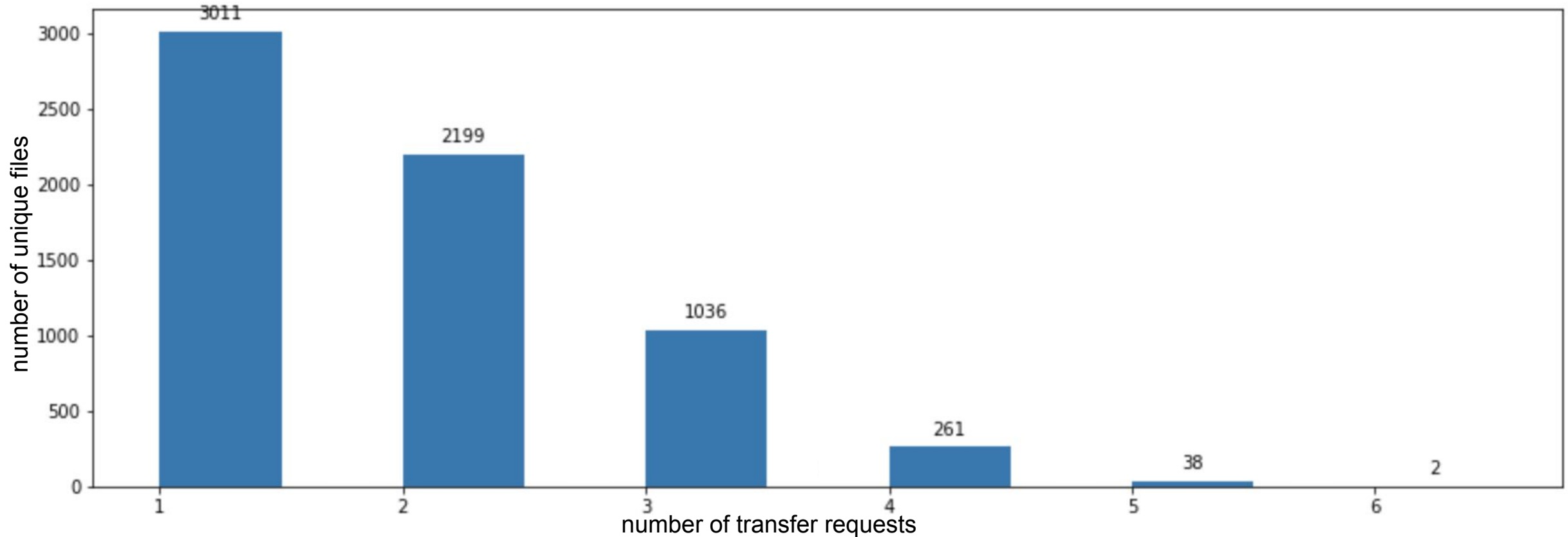


Cache misses
Total size=168.083TB



Cache hits
Total size=322.748TB

Notes on data transfer spike on ESnet node in 10/2020



- 15058 file transfer operations, 12pm-1pm, October 26, 2020 for one user
- 3295 records has 0 file transfer size and 0 transfer time.
- From 11,763 transfer records, 6547 unique files.
- Total transferred file size: 19.64 GB