Data Access Trends in Southern California Petabyte Scale Cache

Alex Sim, Jack Ruize Han, John Wu at LBNL
Frank Würthwein, Diego Davila, Fábio Andrijauskas at UCSD
Inder Monga, Chin Guok at ESnet
Justas Balcas, Harvey Newman at Caltech

May 25, 2022

Introduction

- Data generation in scientific experiments and simulations
  - Data volume is large, challenging for geographically distributed large collaborations
    - E.g., Large Hadron Collider (LHC) from High-Energy Physics (HEP) community
  - Data stored at a few locations, requiring significant networking resources
    - ATLAS Tier-1 site at Brookhaven National Laboratory
    - CMS Tier-1 site at Fermi National Accelerator Laboratory
    - Network traffic primarily carried by Energy Sciences Network (ESnet)
- Observation
  - Significant portion of the popular dataset is transferred multiple times
  - Storage cache allows data sharing between users in same region
    - Reduce the redundant data transfers over the wide-area network
      - Save network traffic volume
      - Lower data access latency
      - Improve overall application performance

Introduction (cont'd)

- Use case
  - CMS:
    - Southern California Petabyte Scale Cache (SoCal Repo)
    - New deployments coming soon:
      - CHICAGO: ESnet/T2_US_Wisconsin 200TB/100TB
      - BOSTON: ESnet-only 200TB to be used by T2_US_MIT
  - Open Science Data Federation (OSDF)

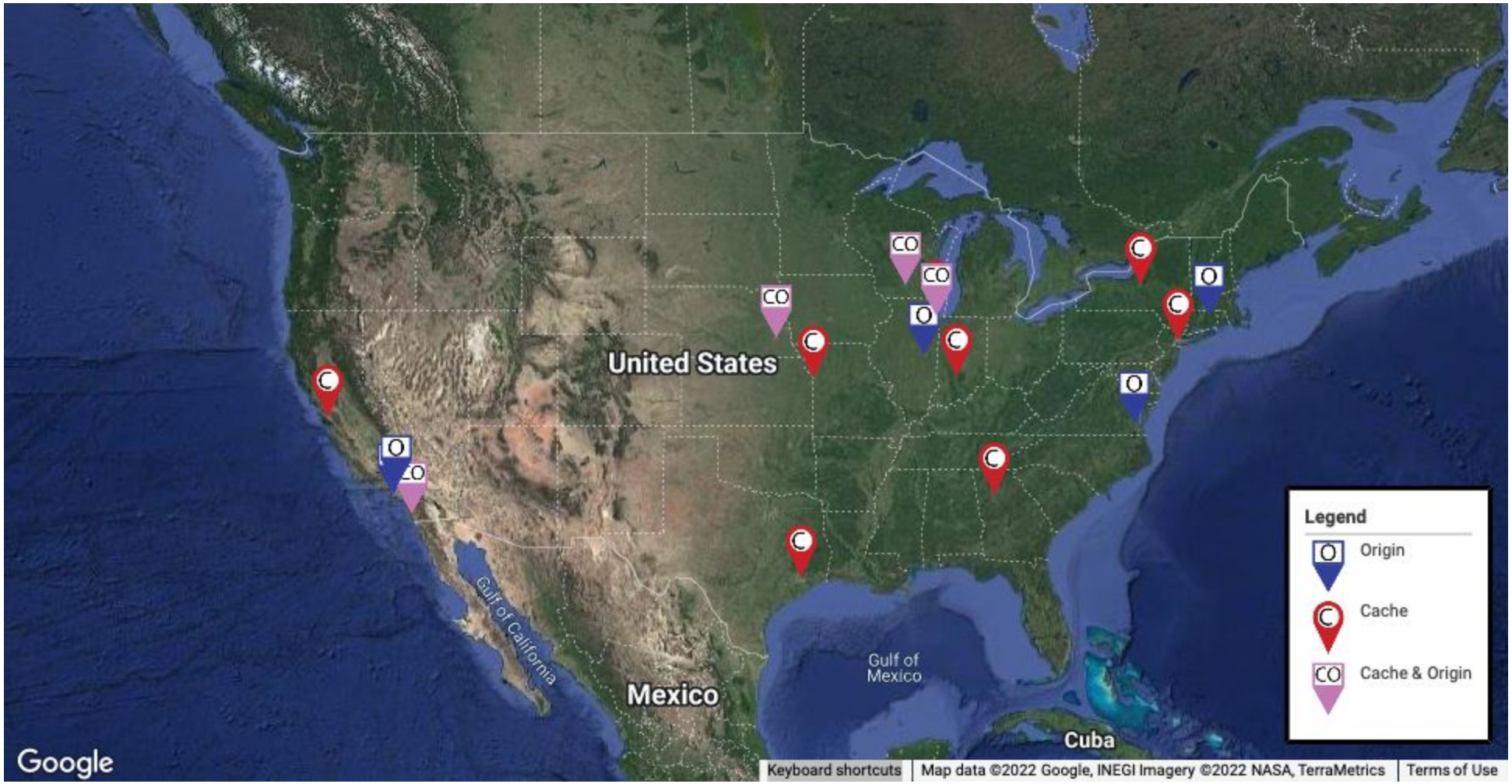
Open Science Data Federation

Providing data access and transfer services for Open Science

BYTES READ
29,570 TB
↑938 MB/s Last 1 Year

FILES READ
1,188,355,803
↑38/s Last 1 Year

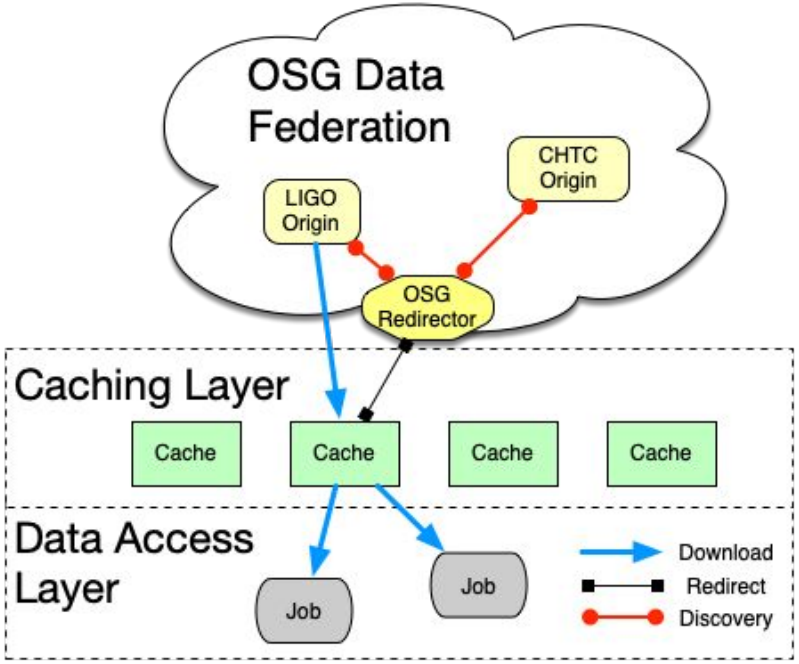
https://osdf.osg-htc.org



https://opensciencegrid.org/about/osdf/

Open Science Data Federation

- Origin: Storage of the data from multiple Virtual Organizations (VO);
- Redirector: Process the data request to direct the request to the appropriate origin;
- Cache: Storage to provide data geographically close to the execution points and access points;
- Data access: It is possible to use different tools. Using OSDF commands is possible to fetch the files from the closest cache (GeoIP);

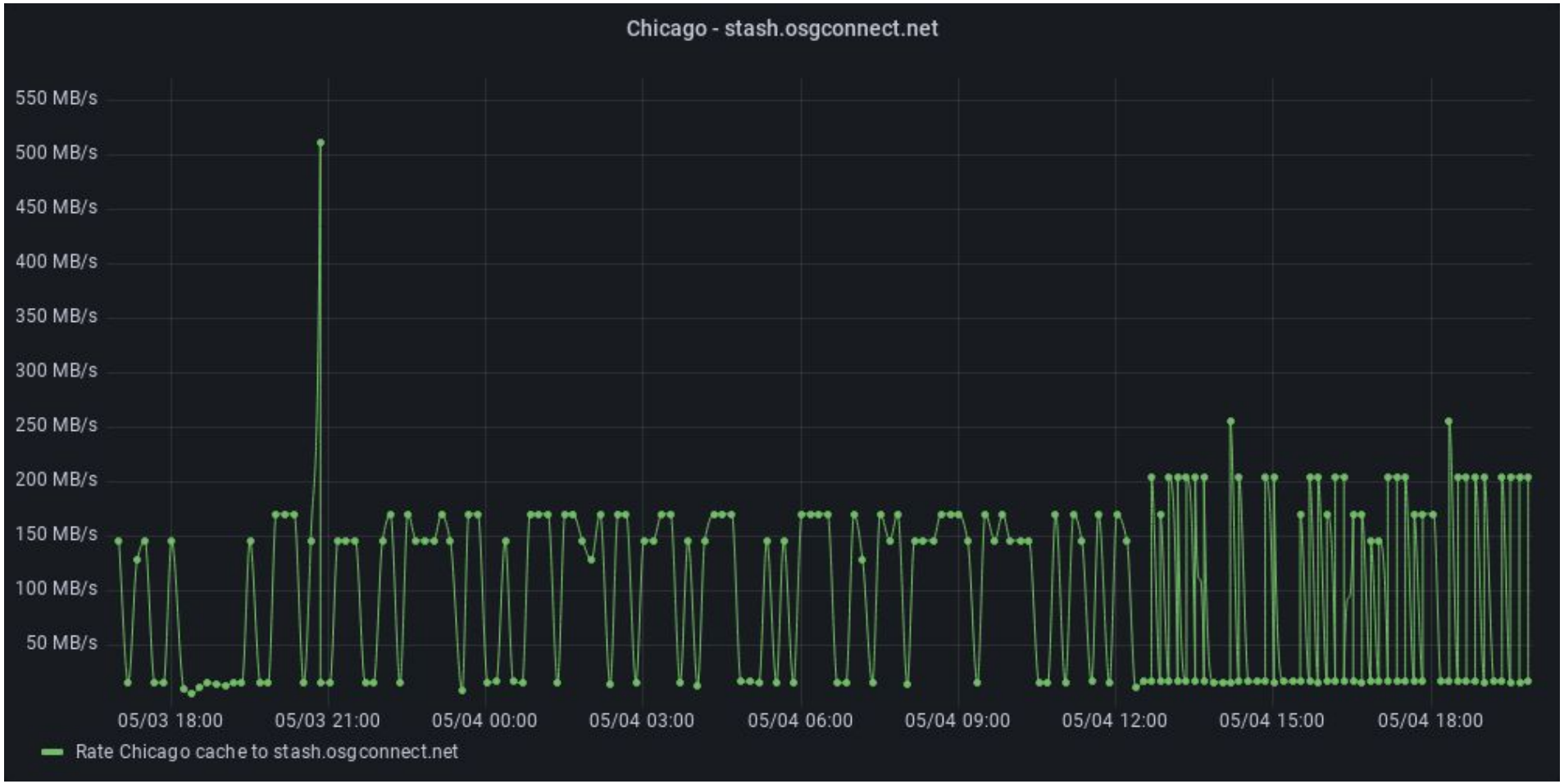


Cache, origins, and redirector based on the XRootD technology: https://xrootd.slac.stanford.edu

OSDF - Monitoring

Monitoring end-to-end

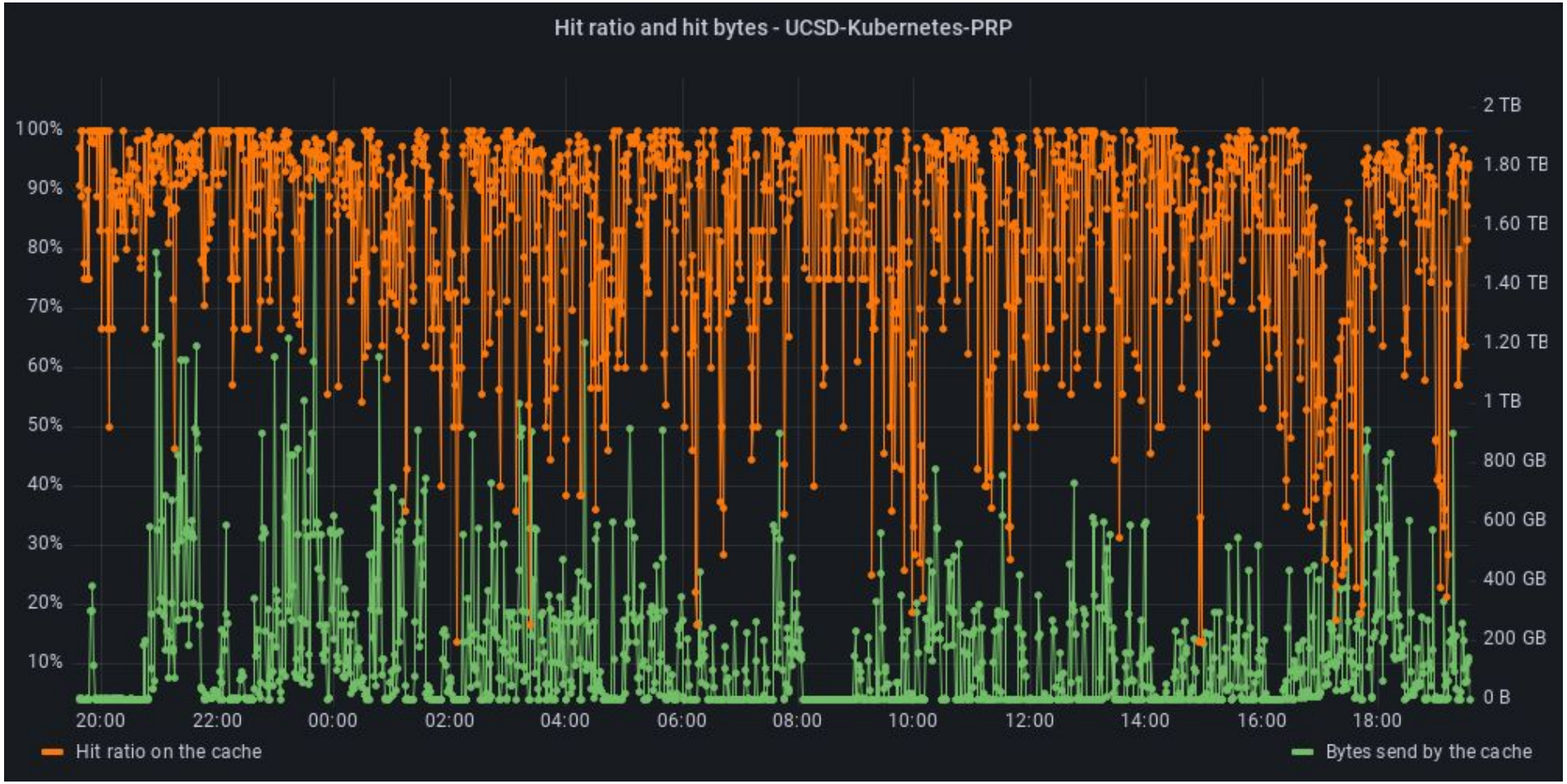
OSG Origin to Chicago Cache



OSDF - Monitoring

Hit ratio on the caches

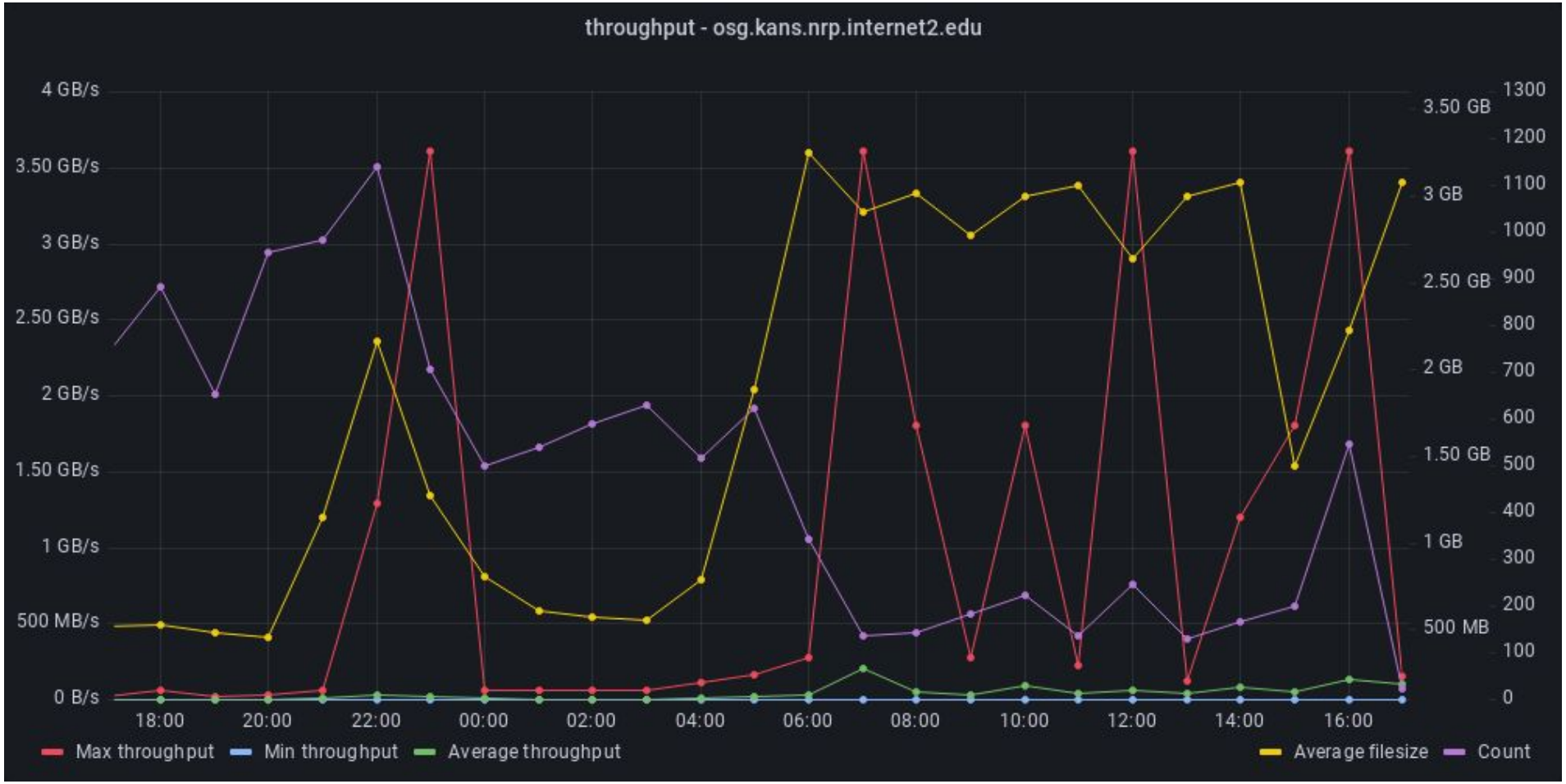
UCSD cache



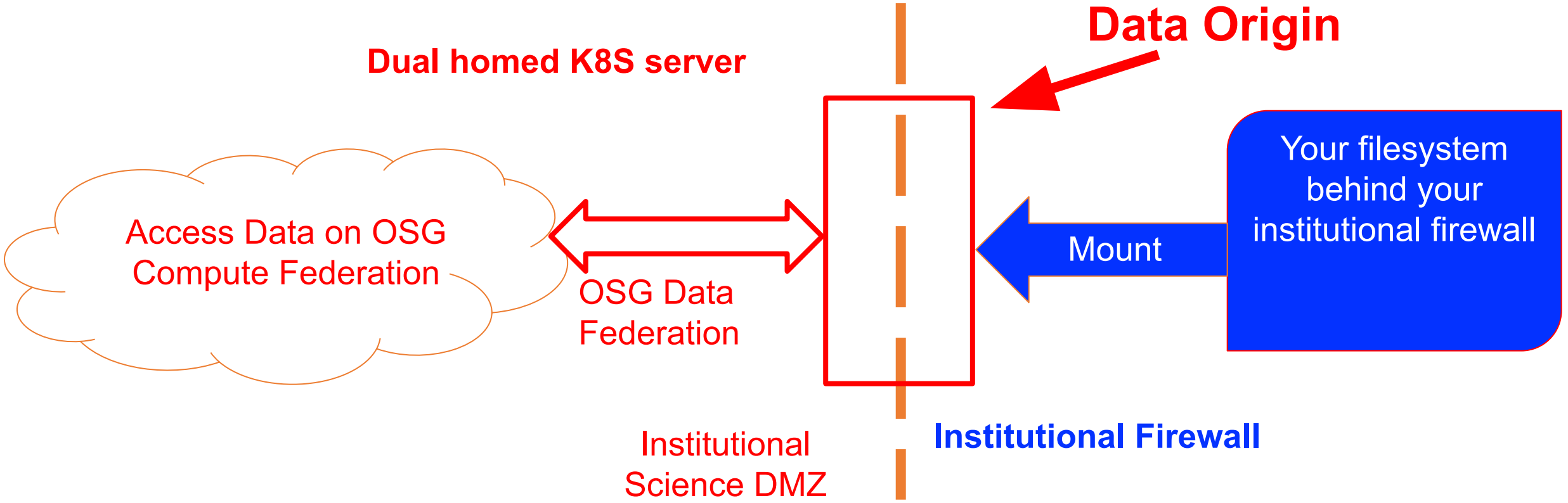
OSDF - Monitoring

Client cache monitoring

Kansas cache



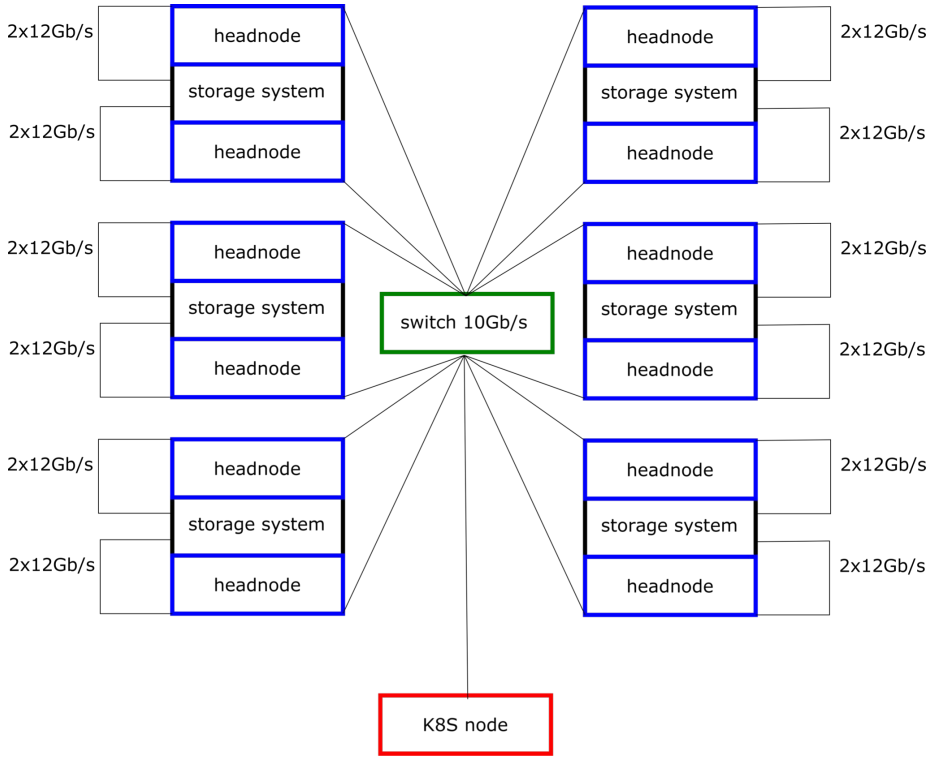
Campus Cyberinfrastructure (CC*)



The devOps model and the actual containers used to make NSF 22-582 data origins possible were developed by IRIS-HEP for use by the LHC.

Campus Cyberinfrastructure (CC*)

- CC* awards made via this solicitation will be supported in two program areas:
- **Data Storage** awards will be supported at up to $500,000 total for up to 2 years
- We created Storage Architecture suggestion to help the researcher to reach this award
- A update quote indicates is possible to reach 15PT using $500,000



Storage Architecture suggestion

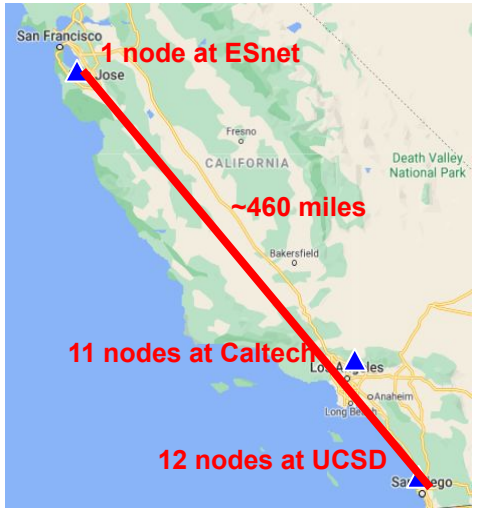
Southern California Petabyte Scale Cache

Southern California Petabyte Scale Cache (SoCal Repo)

- High-Luminosity Large Hadron Collider
  - HL-LHC aims to increase performance after 2025
- SoCal Repo consists of 24 federated storage caches for US CMS
  - 12 nodes at UCSD: each with 24TB and 10 Gbps network connection
  - 11 nodes at Caltech: each with storage sizes ranging from 96TB to 388TB and 40 Gbps network connections
    - 7 new nodes (xrd 3-8, 11) around Aug. 26, 2021
    - 2 new nodes (xrd 9-10) around Sep. 30, 2021
  - 1 node at ESnet: 44TB storage and 40 Gbps network connection
  - Approximately 2.5PB of total storage capacity
  - <200km between UCSD and Caltech nodes, round trip time (RTT) < 3 ms
  - ~700km between ESnet and UCSD nodes, RTT ~10 ms
- Working with US CMS data analysis using MINIAOD/NANOAOD
  - One Caltech node is for NANOAOD, and the rest are for MINIAOD
  - Analysis Object Data (AOD):
    - 384 PB of RAW
    - 240 PB of AOD

    } Mostly on Tape => accessed a few times per year
    - 30 PB of MINIAOD
    - 2.4 PB of NANOAOD

    } Mostly on disk => heavily re-used by many researchers



Sunnyvale–San Diego is the relevant distance scale

SoCal Repo

- Goals in this study: Explore measurements from Southern California Petabyte Scale Cache (SoCal Repo) to understand
  - Trends in cache utilization
  - Trends in network utilization
  - Effectiveness of the SoCal Repo in reducing network traffic volume
  - Predictability of traffic patterns
- Measurement data
  - Logs from SoCal Repo nodes from July 2021 - Jan 2022
  - Analysis on Cori at NERSC
- Key observations
  - On cache utilization and network utilization
    - Reduce the traffic by 2.35X during normal uses
  - On predictability of resource loads and utilization
    - Achieve 88.4% accuracy despite the dramatic changes in usage patterns
- Caching could supplement the existing local data storage and benefit wider user community

Data Access Trend

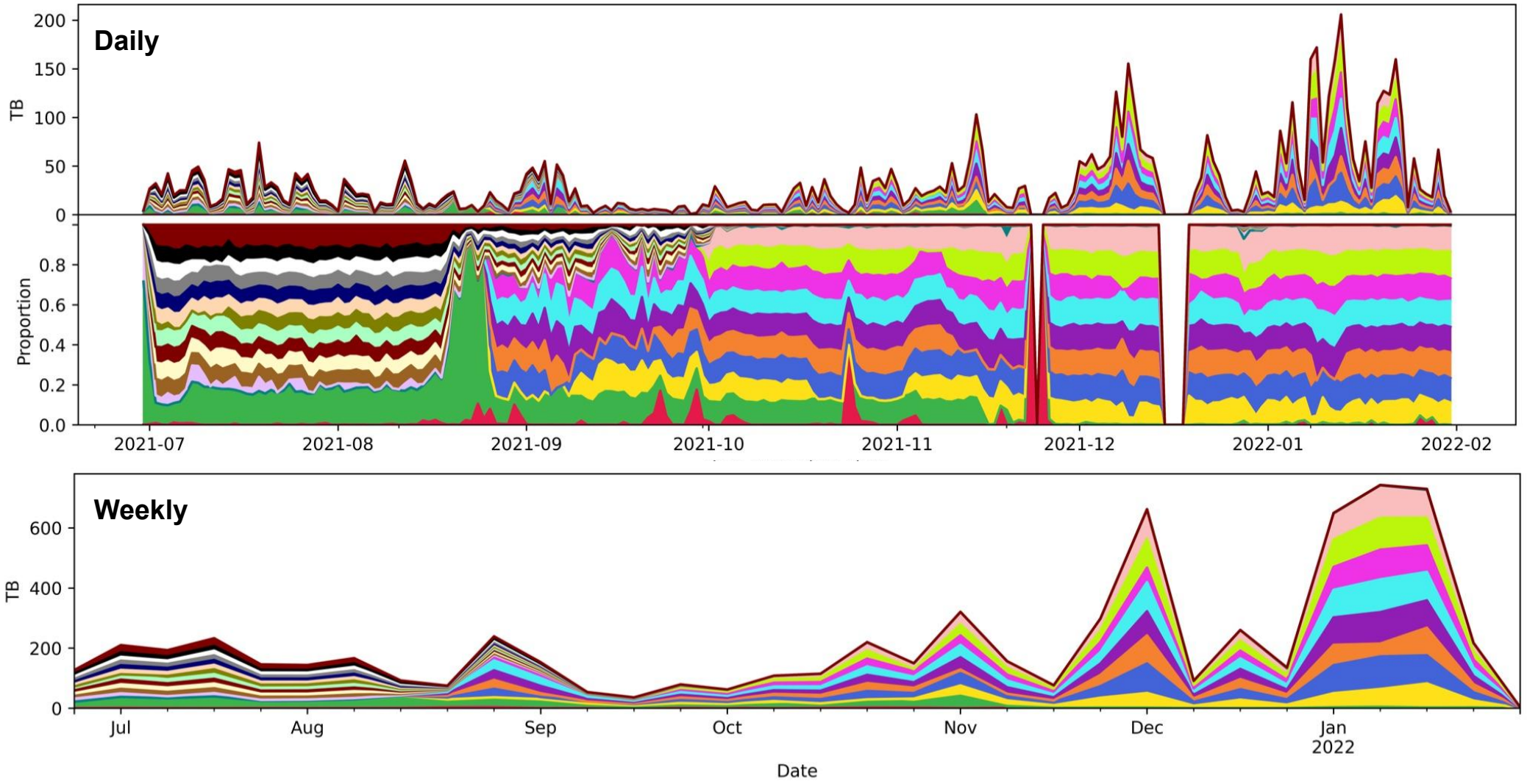
Summary data accesses July 2021- Jan 2022

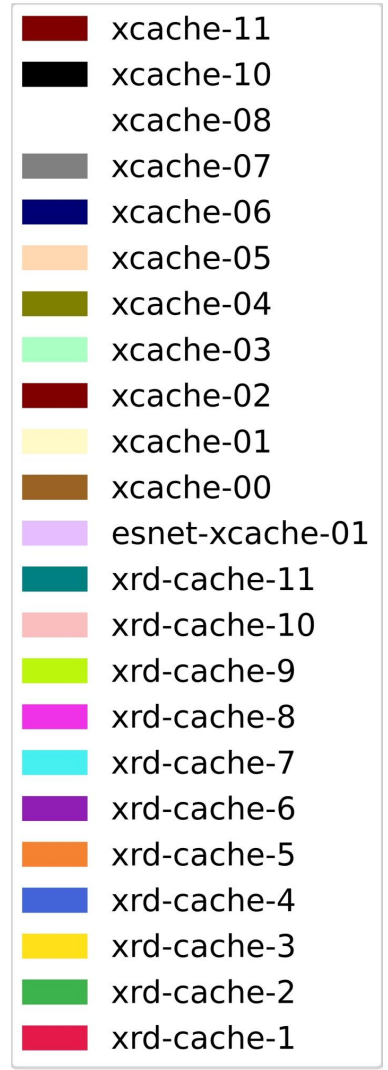
	Number of accesses	Transferred data size (TB)	Shared data size (TB)	Net Traffic Reduction
July 2021	1,182,717	385.78	519.25	2.35
Aug 2021	1,078,340	206.94	313.46	2.51
Sept 2021	1,089,292	206.96	257.18	2.24
Oct 2021	1,058,071	412.18	141.91	1.34
Nov 2021	878,703	649.30	82.67	1.13
Dec 2021	983,723	1,257.89	130.03	1.10
Jan 2022	1,207,332	2,238.59	148.26	1.07
Total	7,478,178	5,357.66	1,592.78	1.30
Daily average	35,441.60	25.51	7.55	

- Transferred data size
  - First time data access size, i.e., cache misses
  - From remote sites to the local node cache, then to application
- Shared data size
  - Repeated data accesses, i.e., cache hits, no need to transfer from remote sites
  - From the local node cache to the application
- Net (wide-area) Traffic Reduction
  - (Transferred data size + Shared data size) / Transferred Data Size

Number of Accesses Fluctuates Around 31000 Per Day

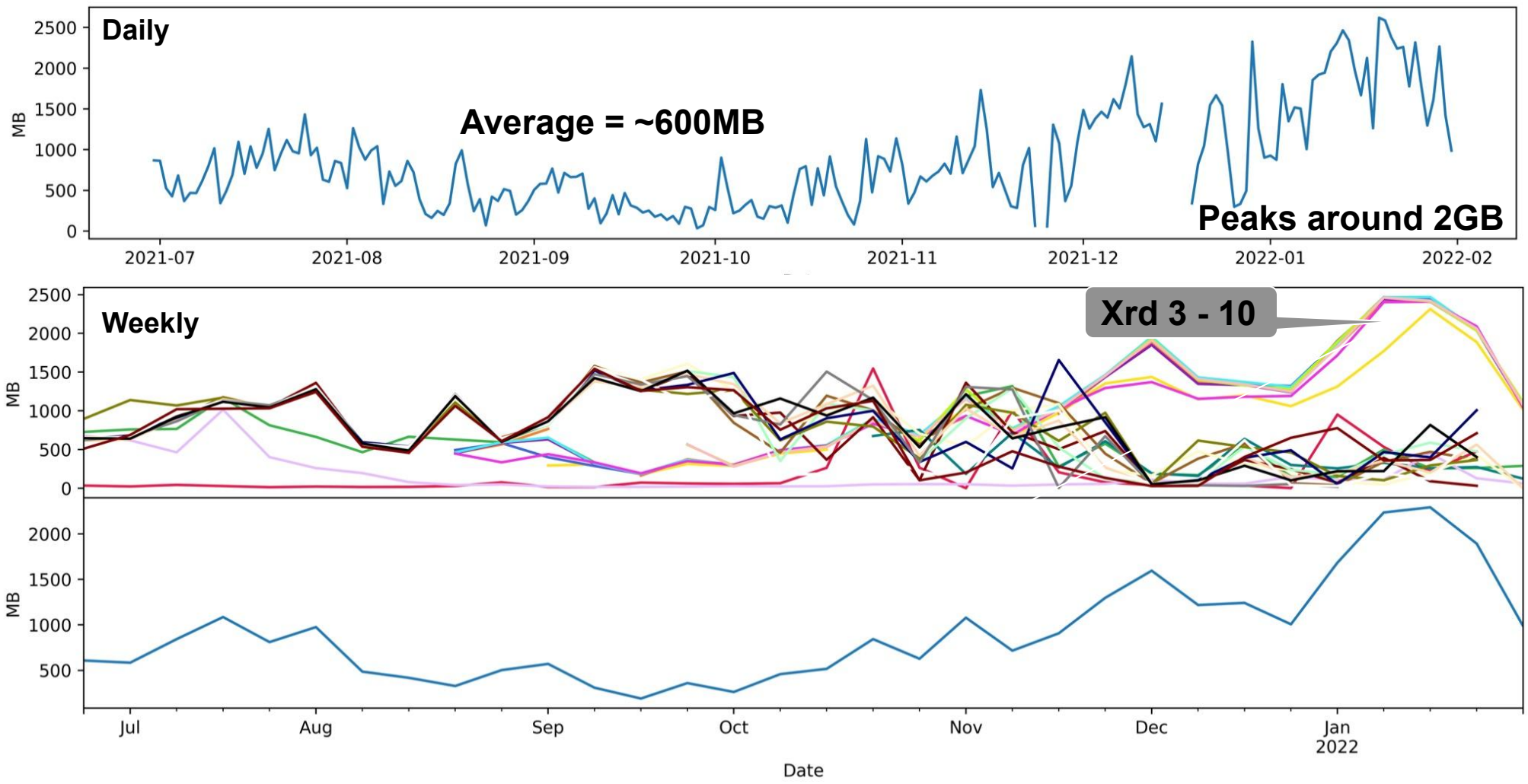
Volume of Data Accesses Hovers Around 21TB/day

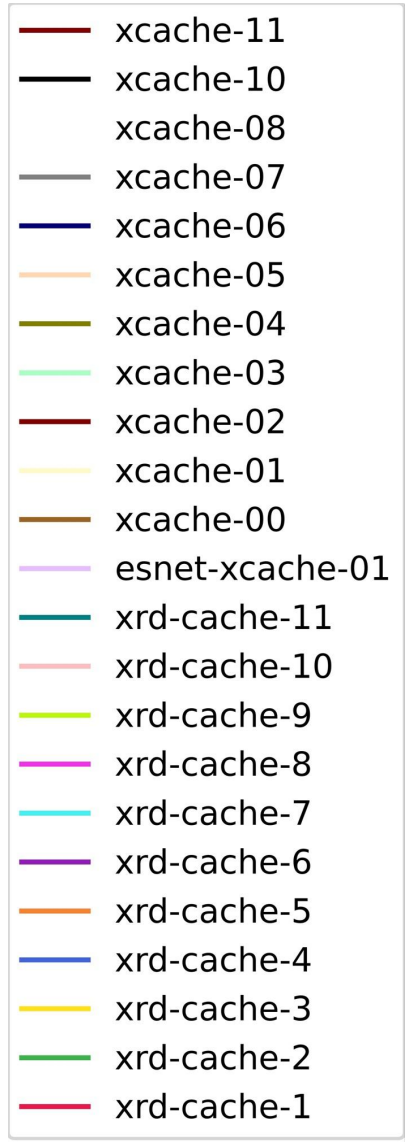




Large amounts of new file accesses in Dec and Jan

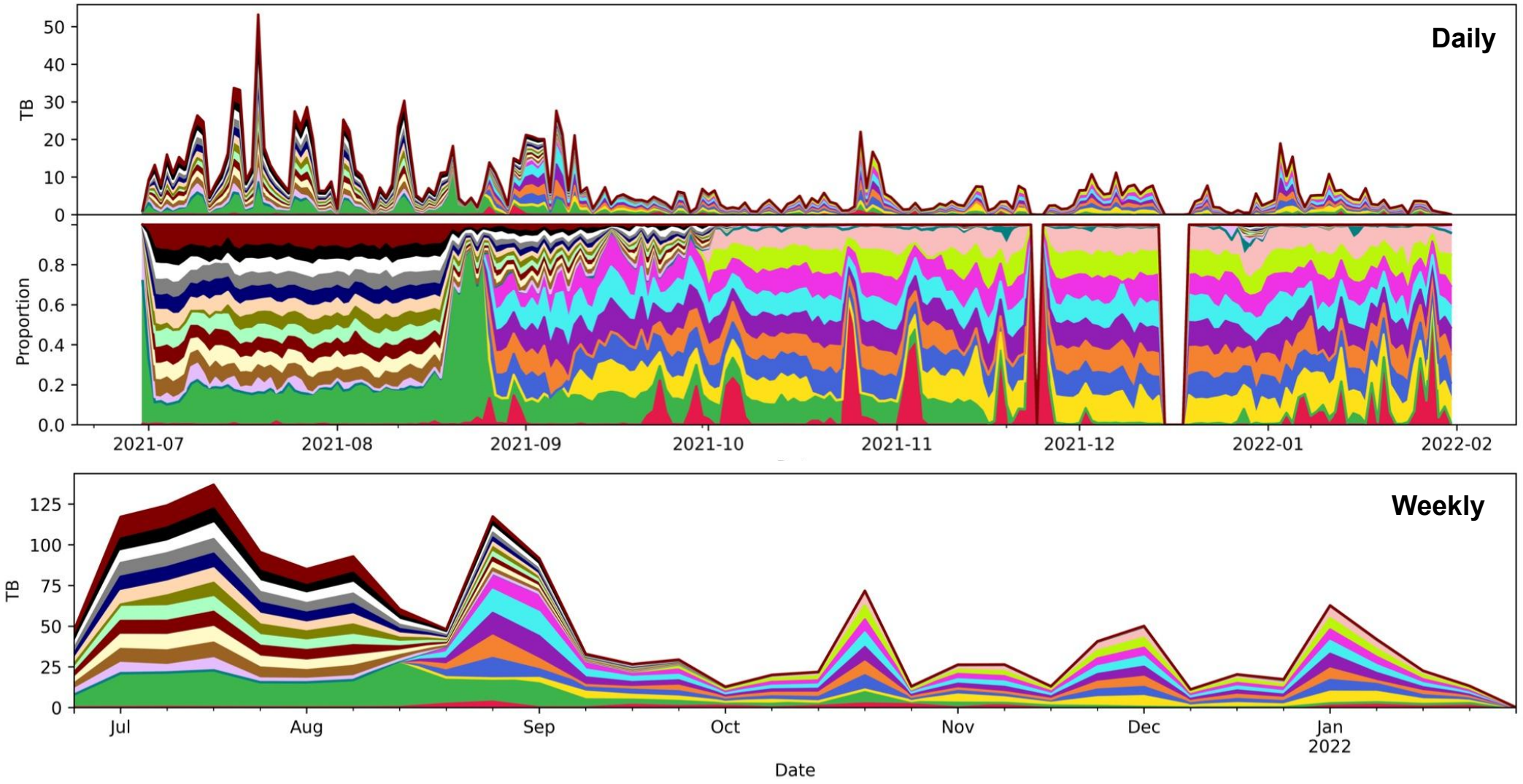
Average Data Size per Access

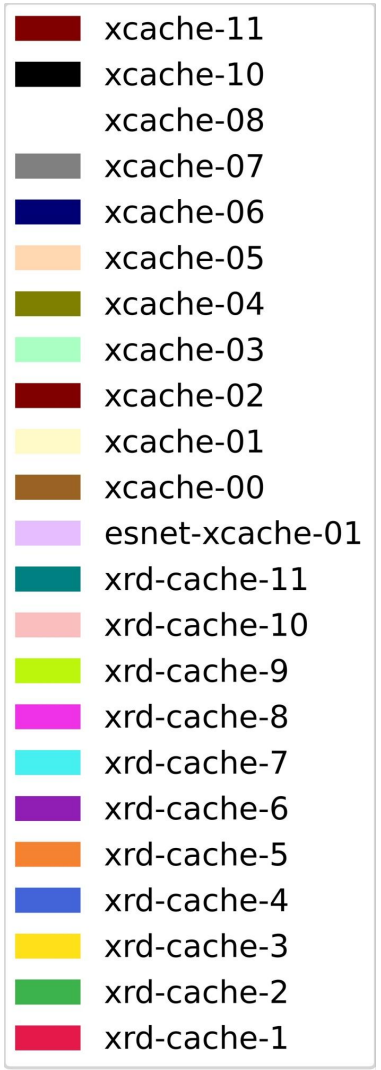




Average Size per access = (*Data Access Size*) / (*Data Access Counts*)

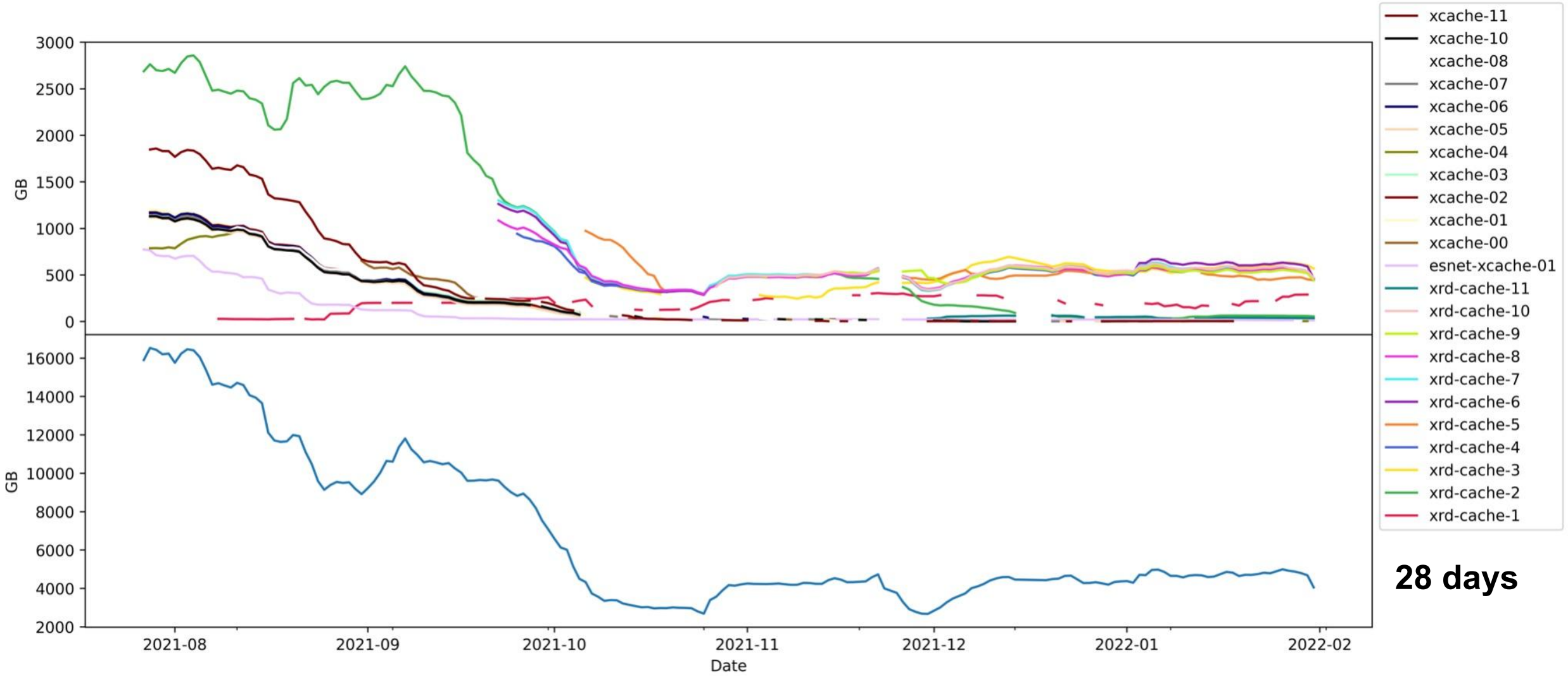
Shared Data (Cache Hits) Size





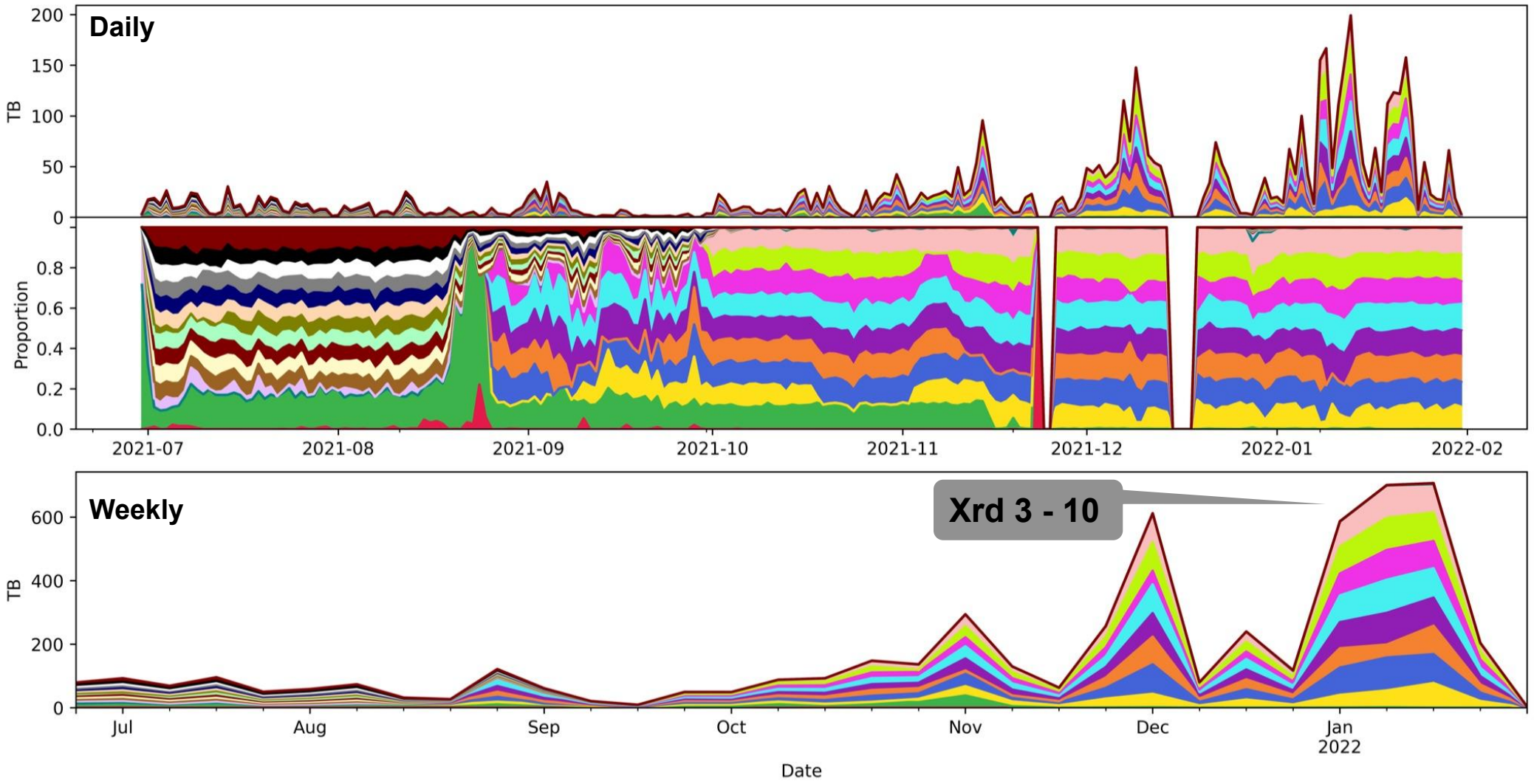
Significantly reduced shared data size since mid-Sep

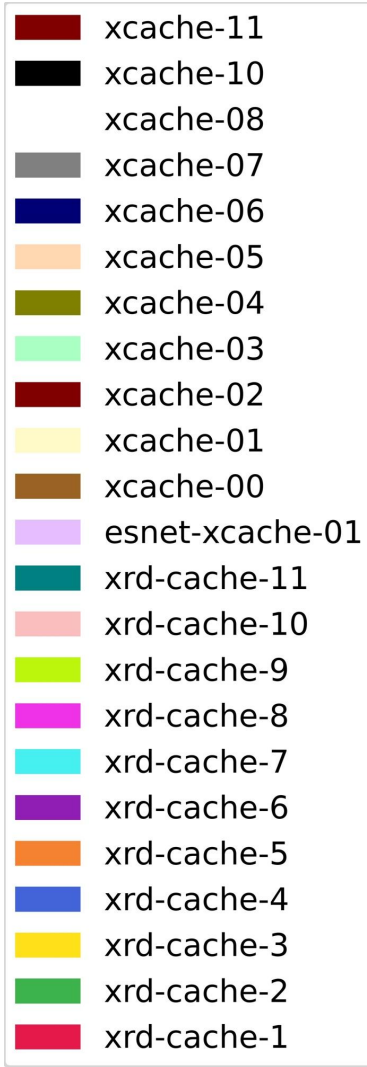
Shared data (cache hits) size with moving average



28 days

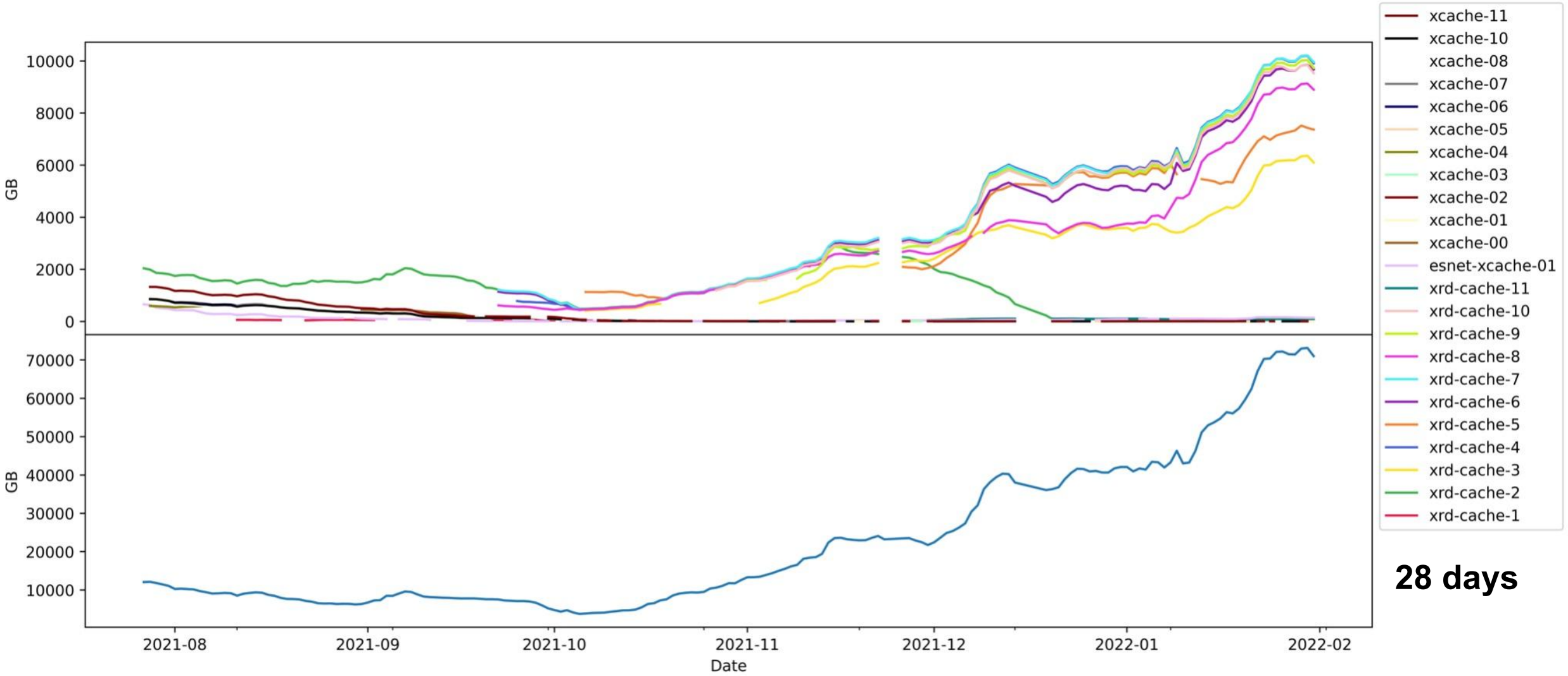
Transferred data size (Cache misses)





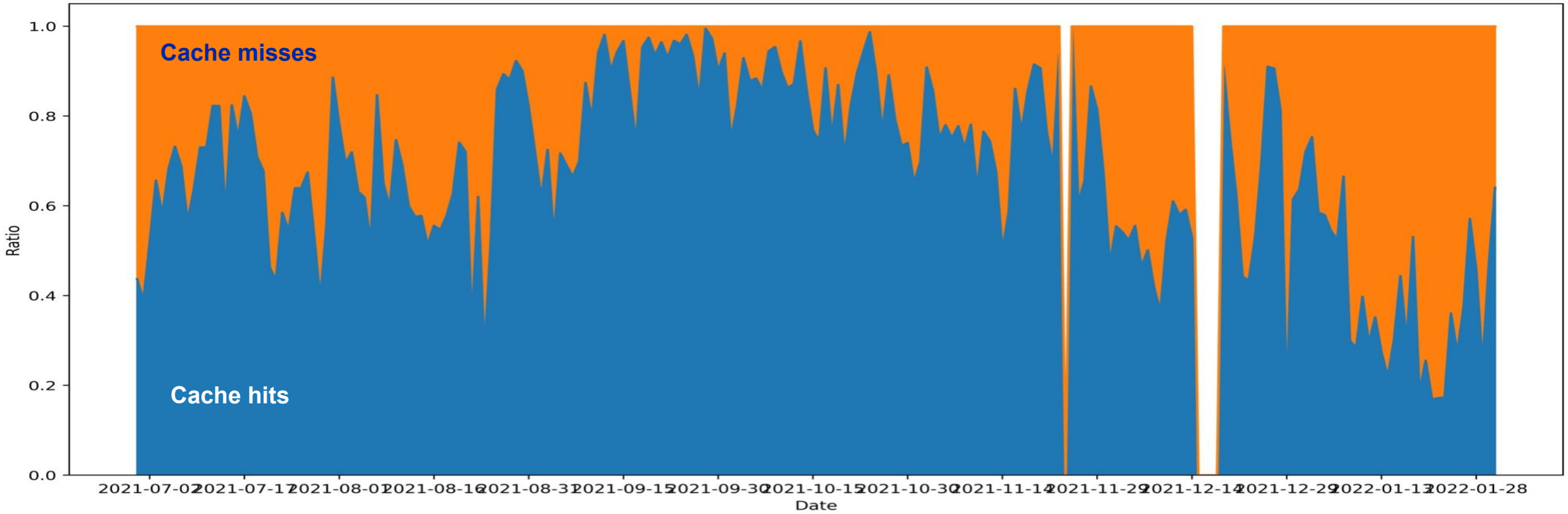
Much more data transferred in Dec and Jan

Data transfer (cache miss) size with moving average

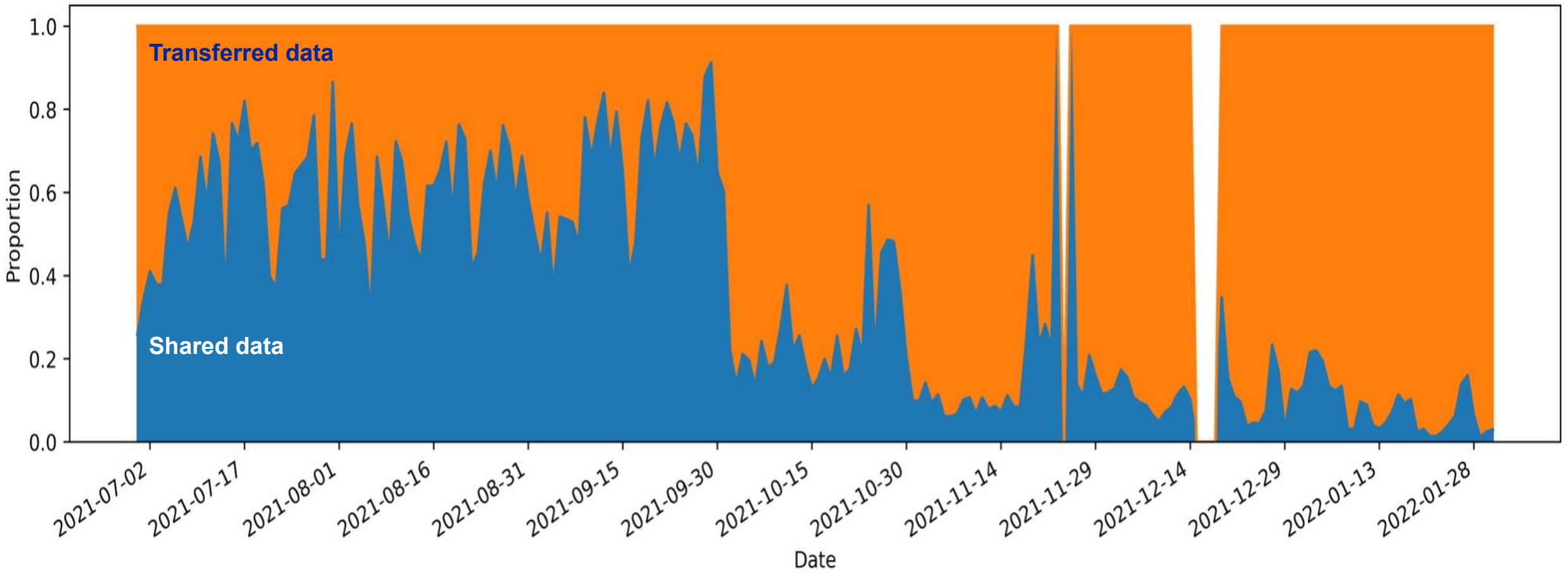


28 days

Daily proportion of number of cache misses and cache hits

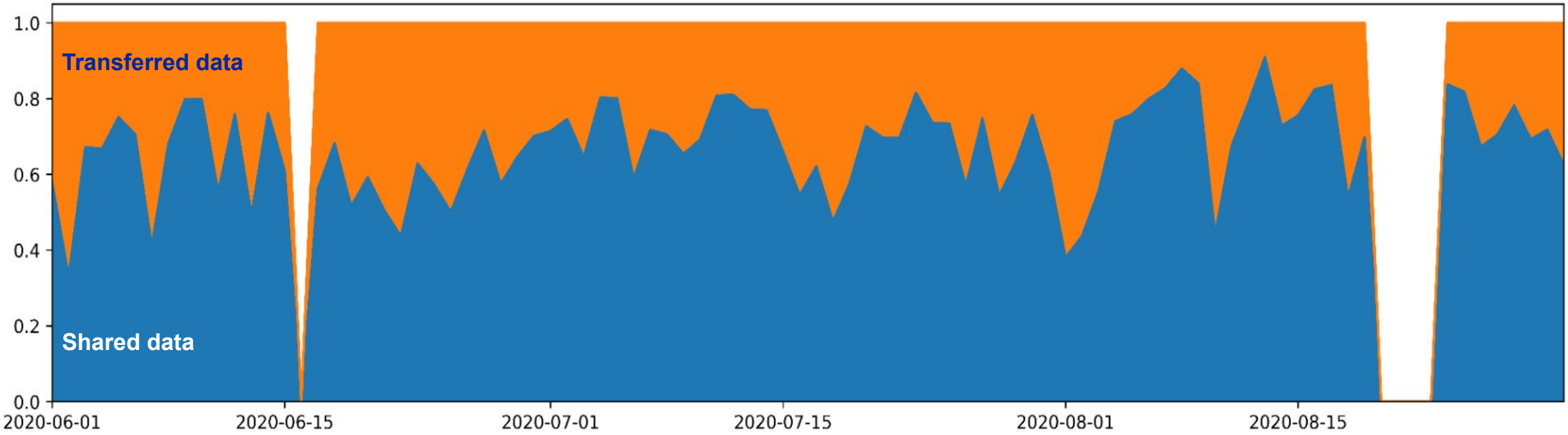


Daily Proportion of Data Transfer Size vs Shared Size



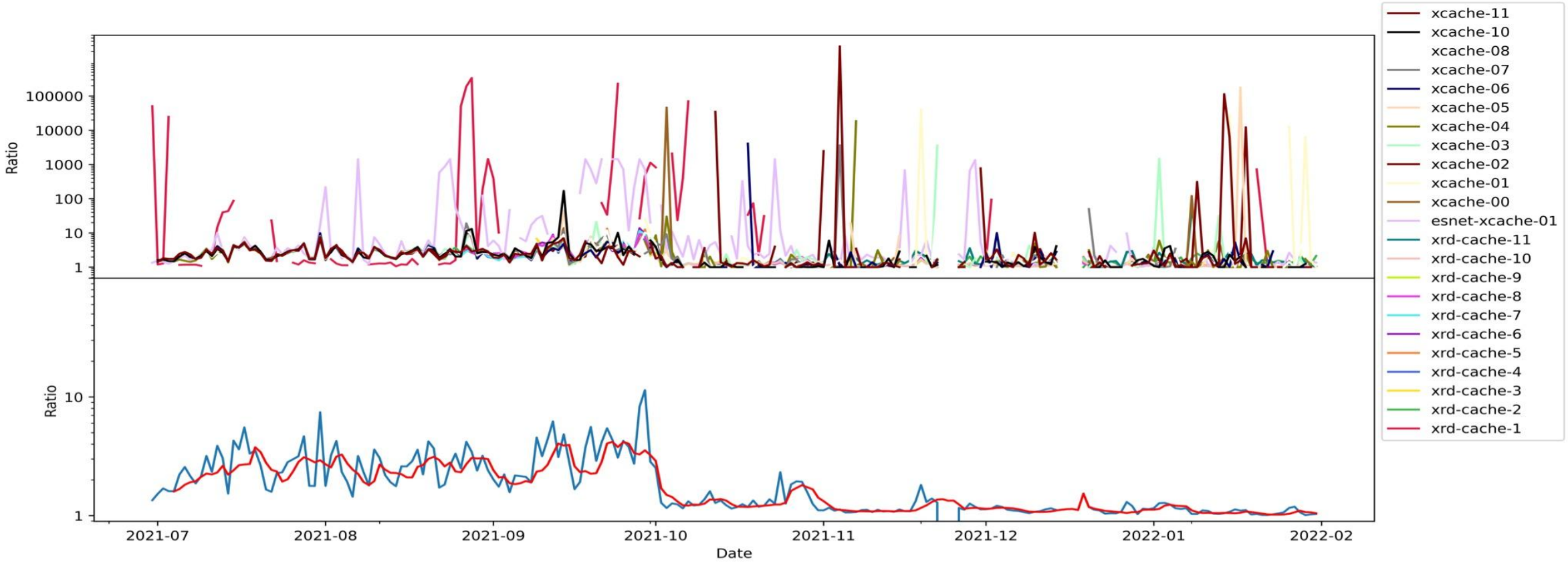
The sudden drop of proportion of data transfer size since Oct 2021 is due to several users streaming large amount of new files.

Compared to daily proportion of data transfer sizes and shared data sizes (June 2020 - Aug 2020)



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**

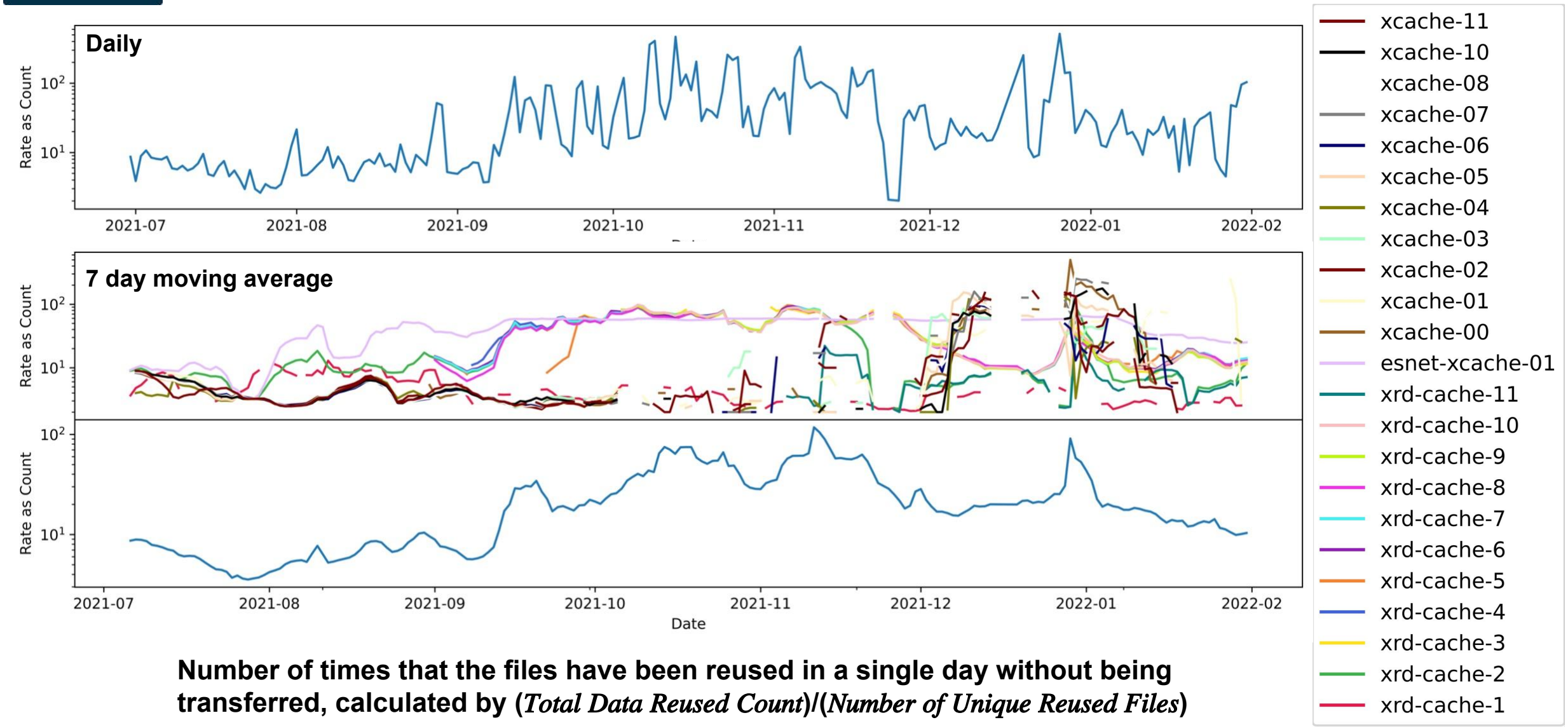
Network Traffic Volume Reduction Over Time



network traffic demand reduction rate = (total cache hit size + total cache miss size) / (total cache miss size)

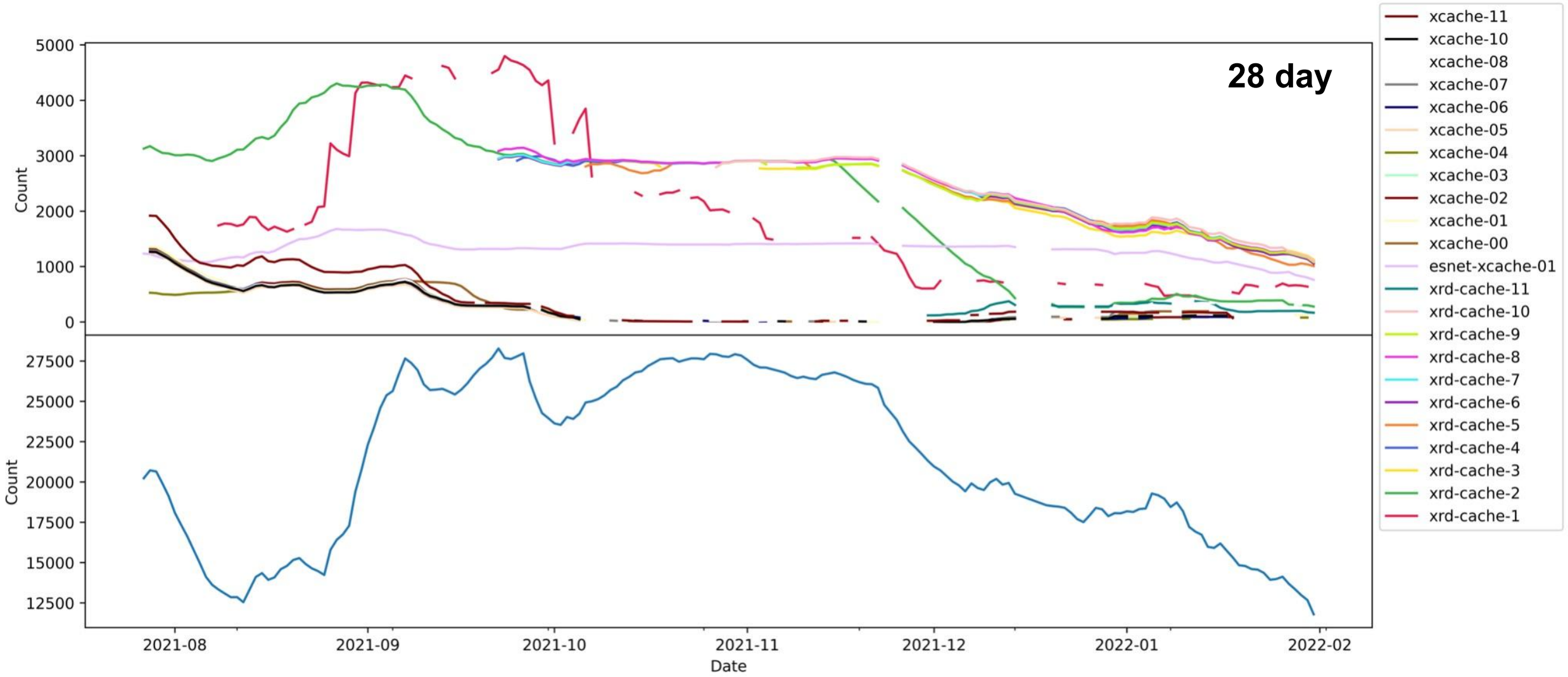
Avg traffic volume reduction for the whole period: 1.30
Avg traffic volume reduction from July 2021 - Sep 2021: 2.35
Avg traffic volume reduction from Oct 2021 - Jan 2021: 1.11

Number of Data Re-used Per Day (in log cale)

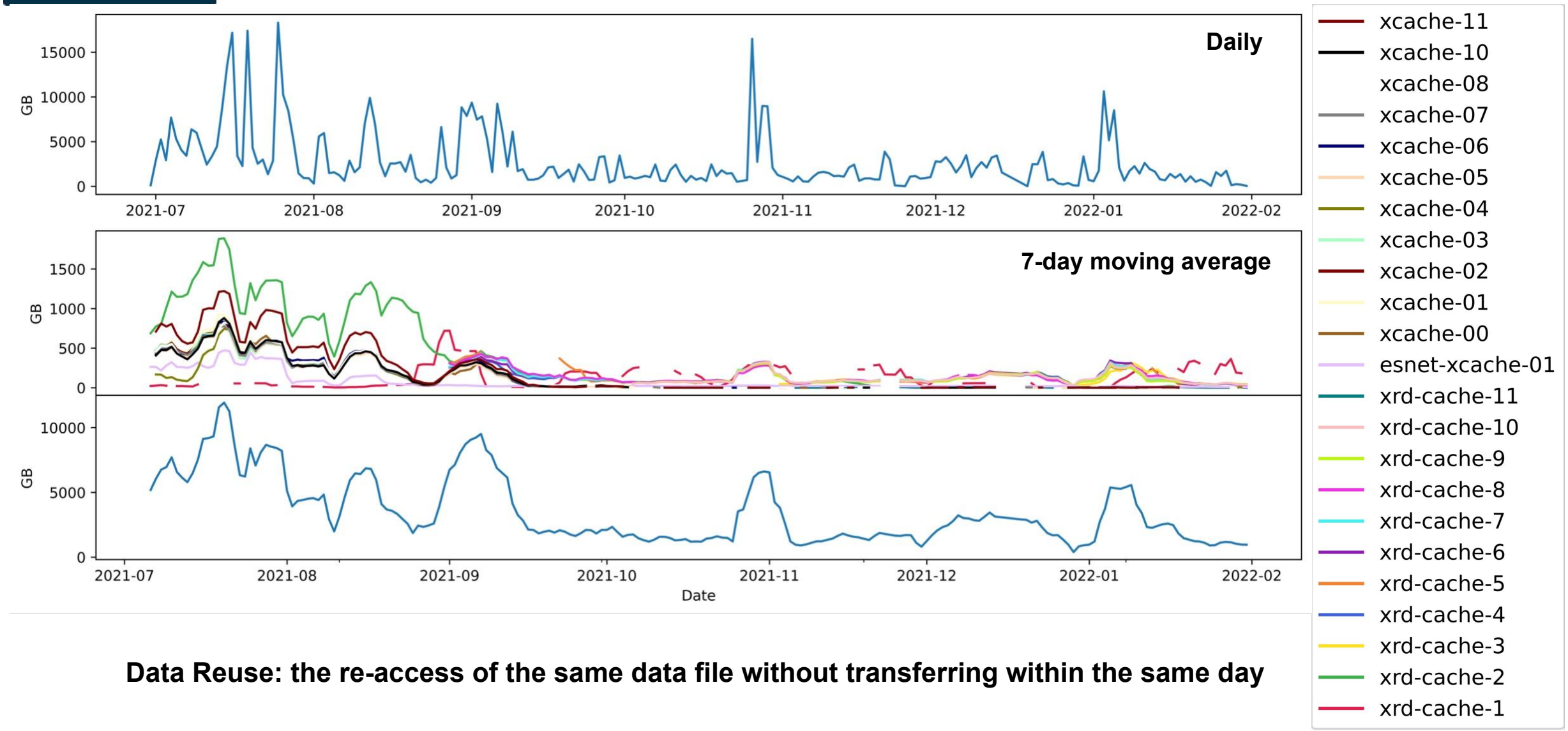


Number of times that the files have been reused in a single day without being transferred, calculated by (*Total Data Reused Count*)/(*Number of Unique Reused Files*)

Daily number of data re-use with moving average



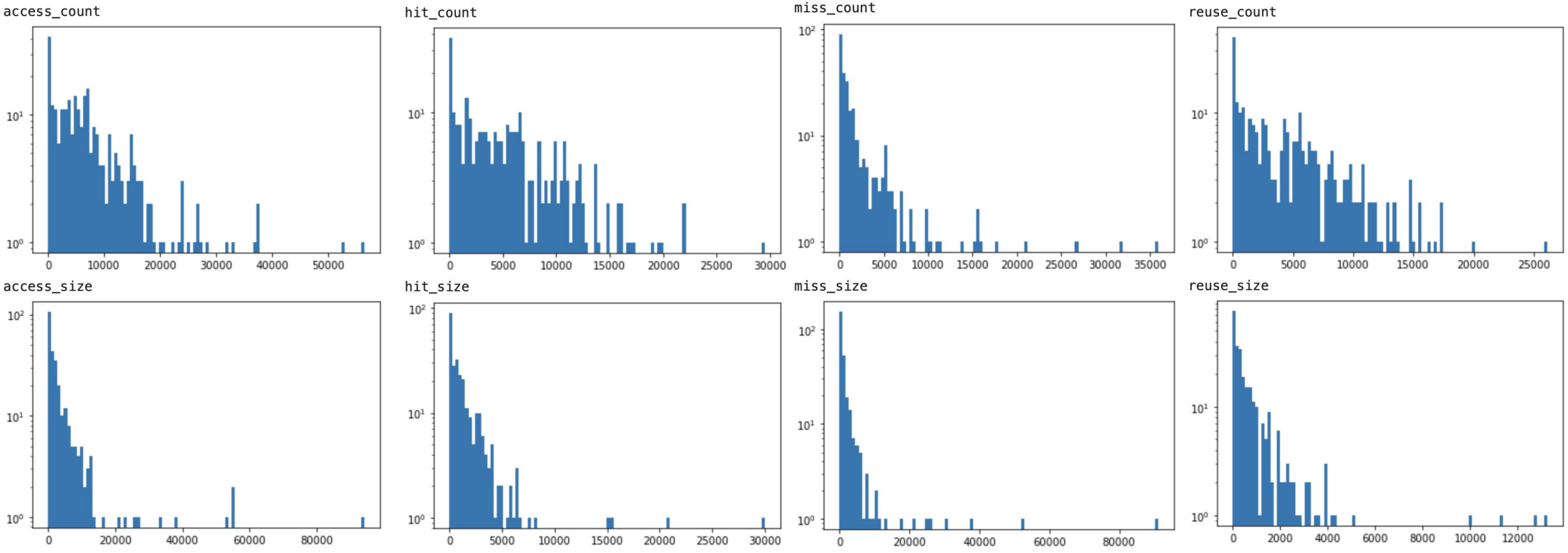
Total Volume of Data Re-use Is Somewhat Consistent Over Time



Data Reuse: the re-access of the same data file without transferring within the same day

MODELING AND PREDICTING CACHE UTILIZATION

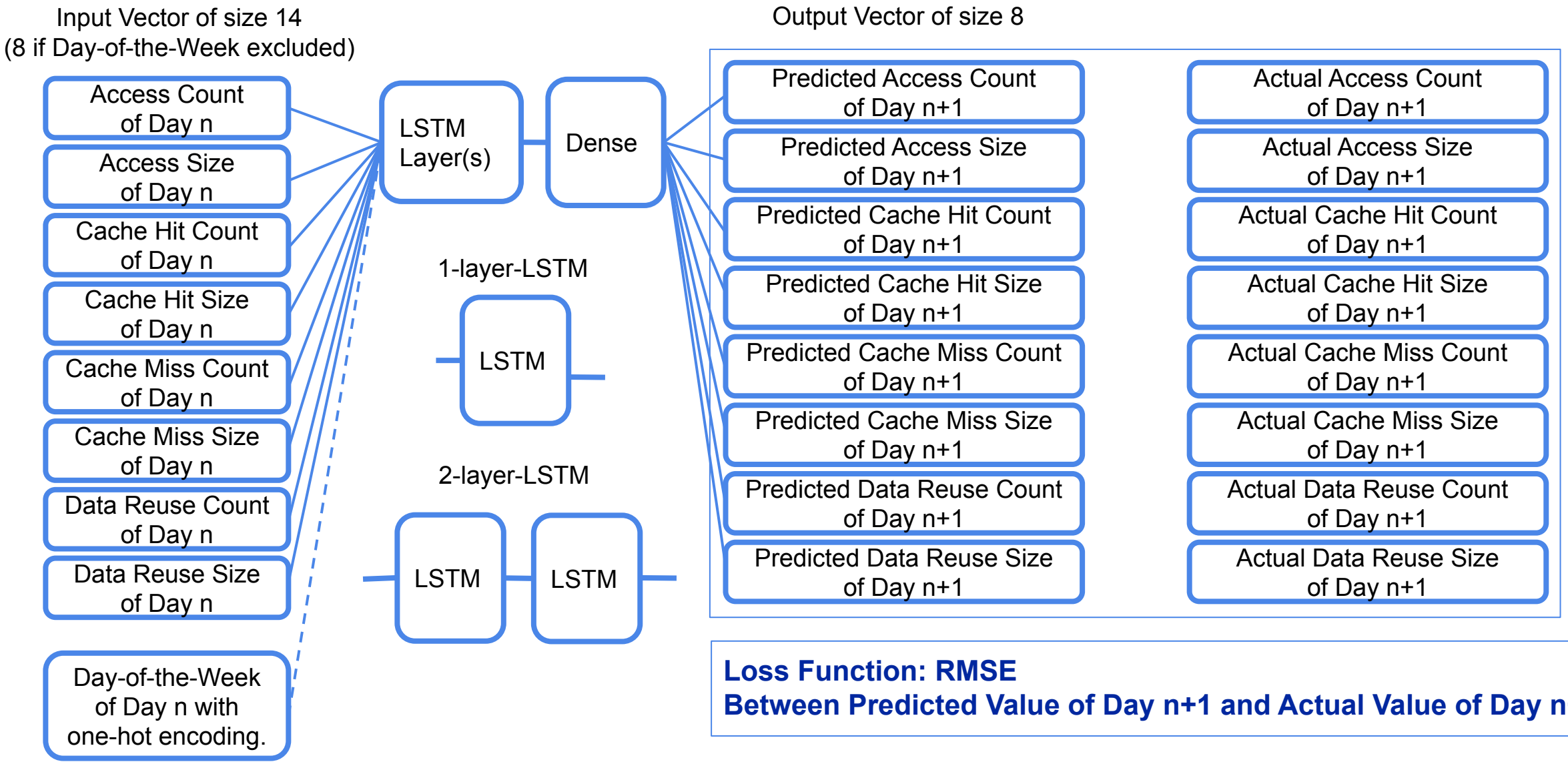
Distribution of Daily Data



Most features have highly skewed distributions, with small number of very large values

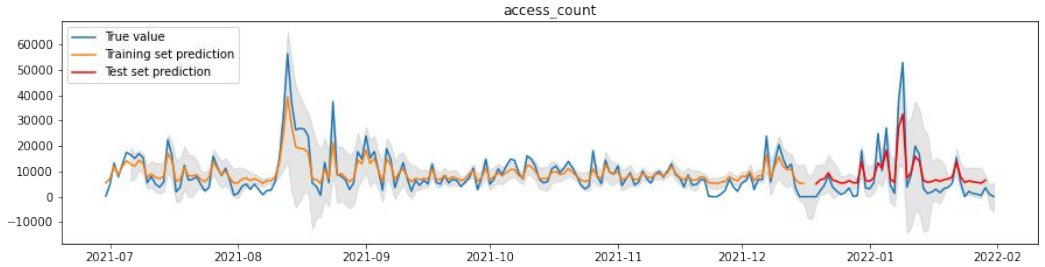
80% Training July 1, 2021 to Dec. 16, 2021, 20% Testing Dec. 19, 2021 to Jan. 29, 2022

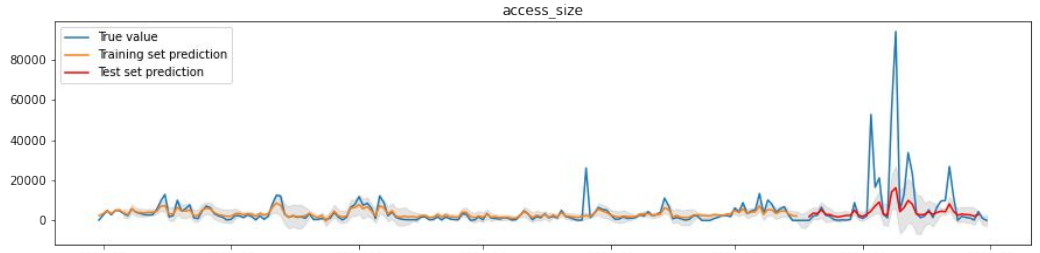
Model Construction

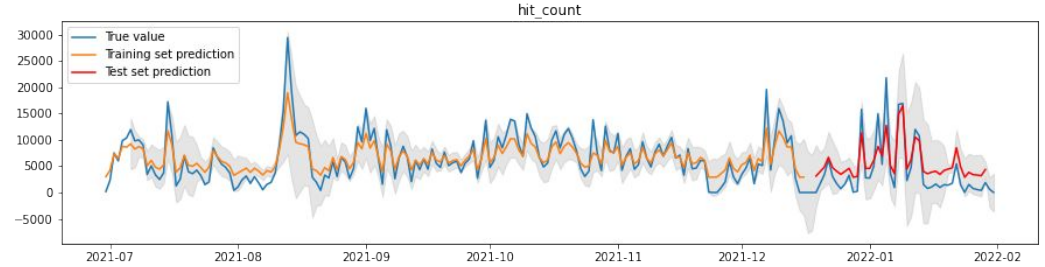


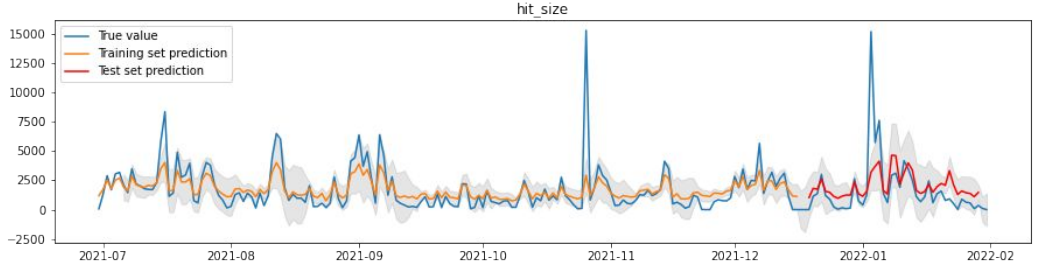
Loss Function: RMSE
Between Predicted Value of Day n+1 and Actual Value of Day n+1

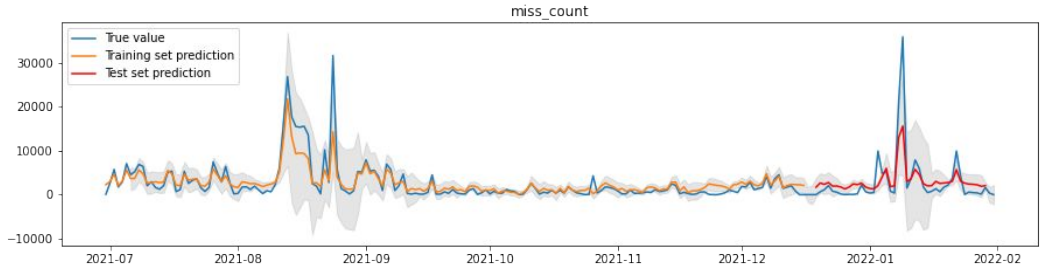
Daily LSTM model results

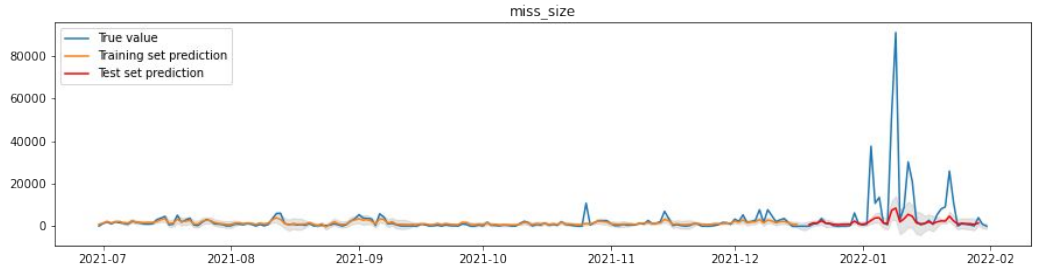


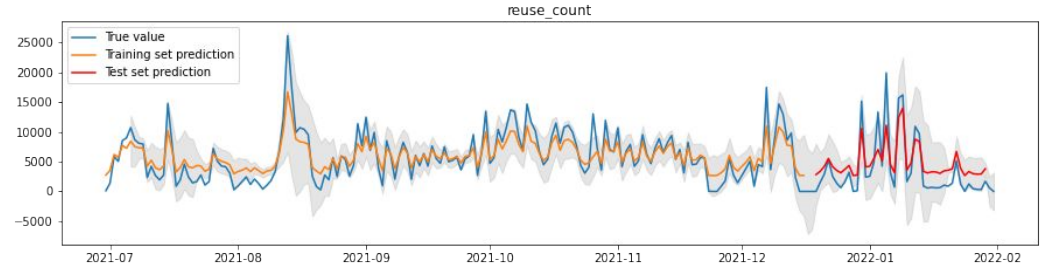


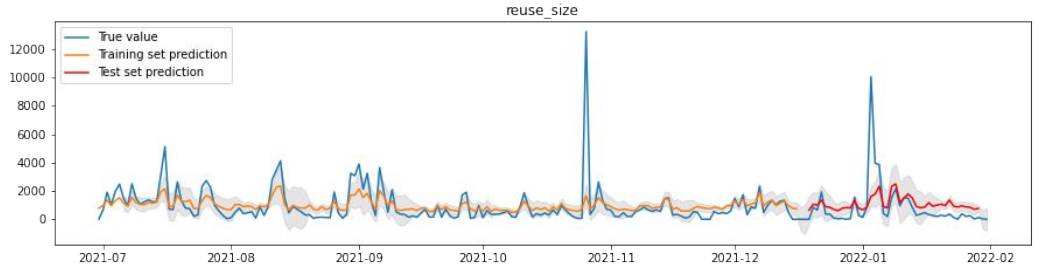




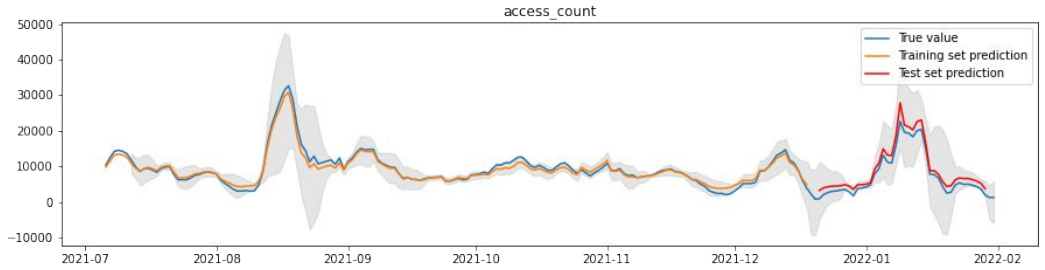


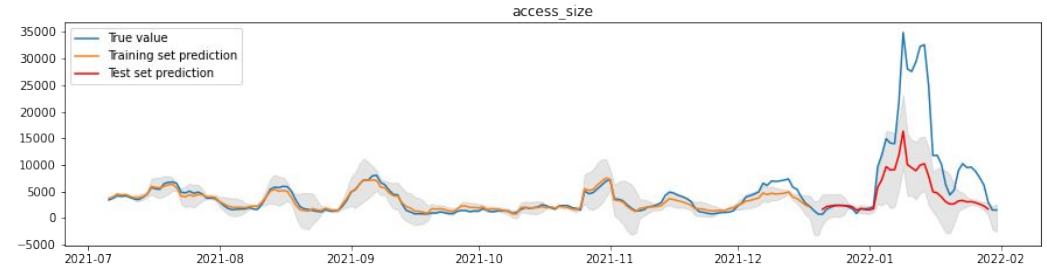


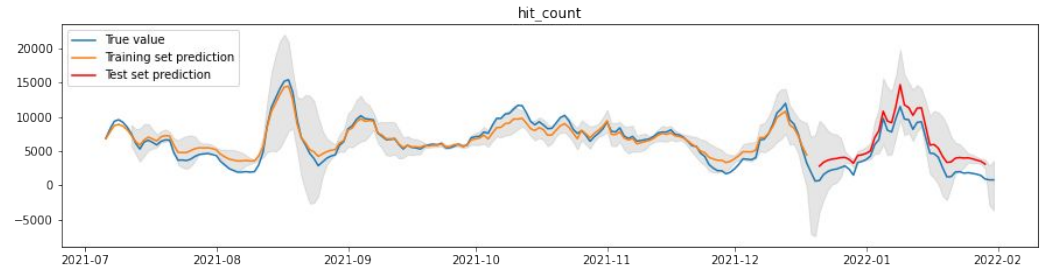


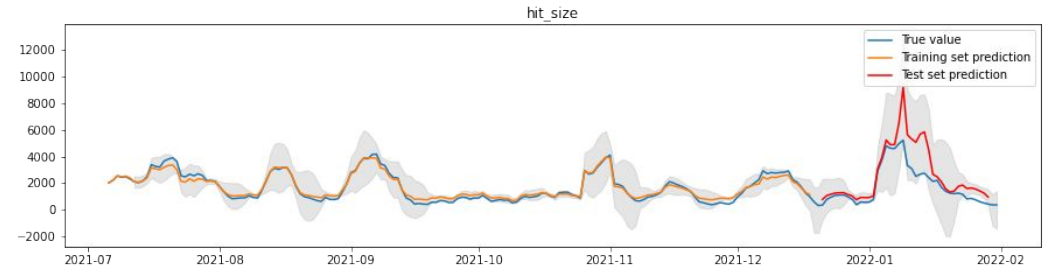


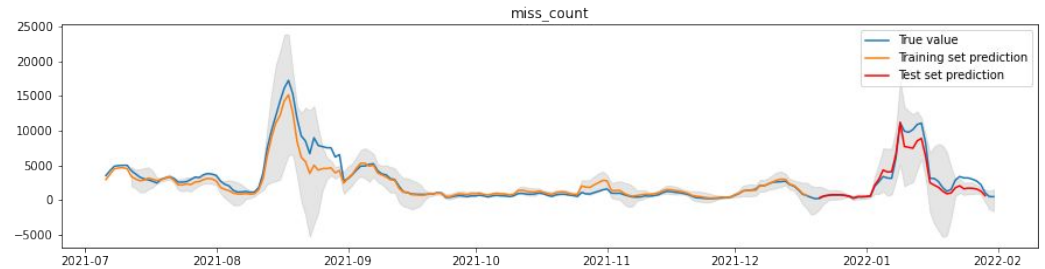
Daily LSTM model results with 7 day moving average

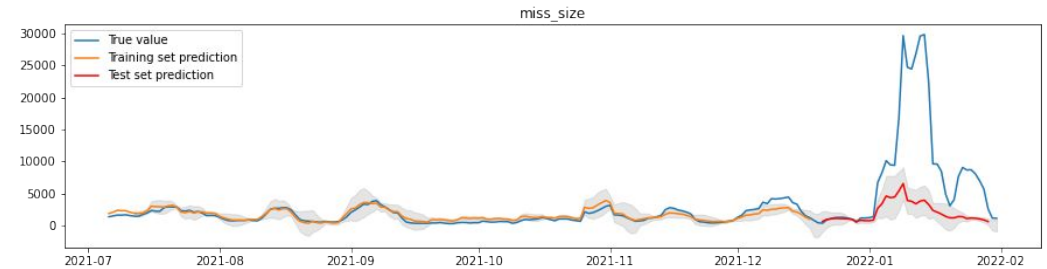


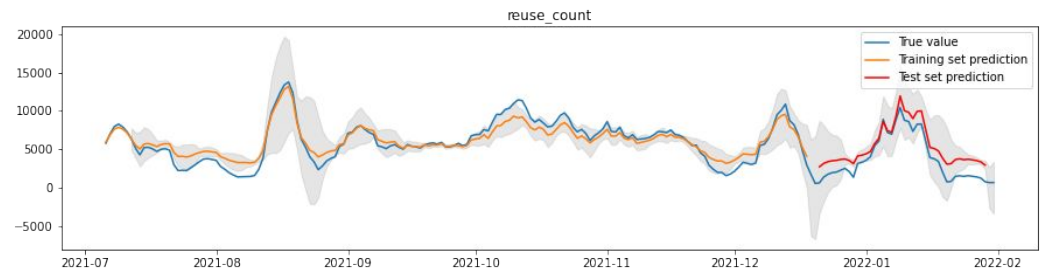


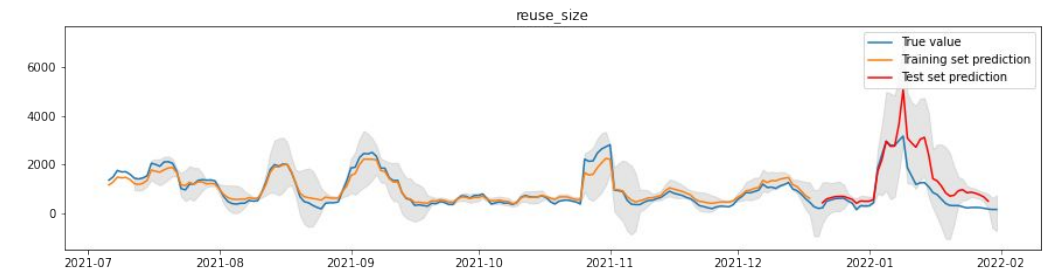












Accuracy of Daily LSTM model

	Daily LSTM model	Daily LSTM model with 7 day moving average
Access Count	0.93	0.93
Access Size	0.85	0.83
Cache Hit Count	0.95	0.88
Cache Hit Size	0.85	0.91
Cache Miss Count	0.91	0.86
Cache Miss Size	0.90	0.77
Data Reuse Count	0.93	0.87
Data Reuse Size	0.73	0.92

Summary

- Southern California Petabyte Scale Cache
  - Reduced the redundant data transfers, saved network traffic volume
    - Network traffic savings 2.35X during normal operations
- Daily Cache Utilization can be predicted by LSTM
  - LSTM can achieve about 88.4% accuracy
  - LSTM works better with moving average data, as there are less extreme values
- Further studies
  - Shared data rate drops since Oct 2021
  - LSTM models with more data
  - Longer term traffic modeling
  - Impact of resource utilization on data access performance

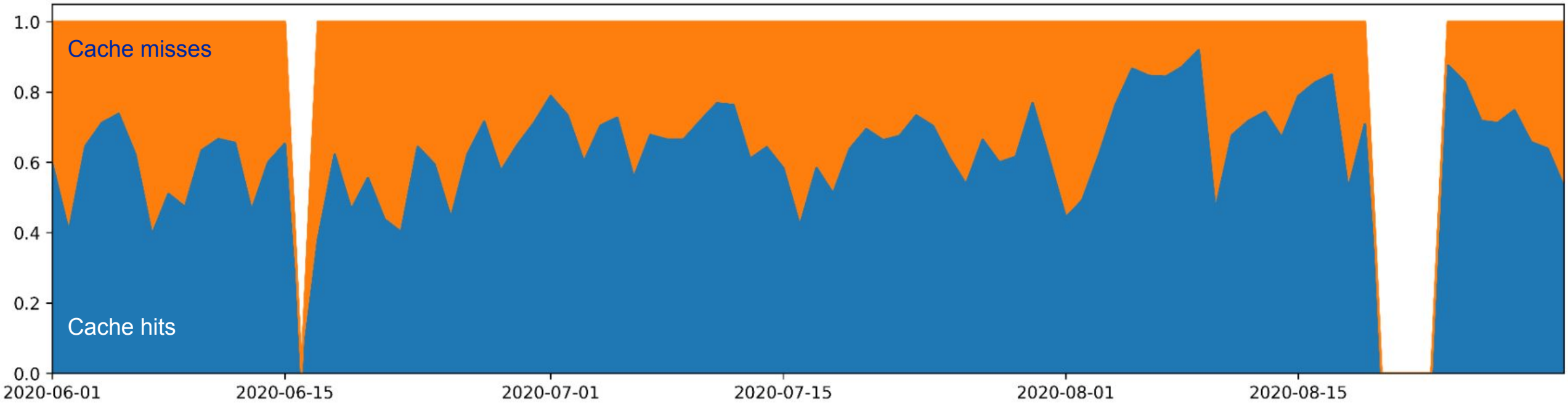
Backup

SoCal Repo file request steps

- "Trivial File Catalogue" (TFC) handles user requests
  - "Local redirector" knows all caches
  - If cache hit, local redirector routes to the node
  - if cache miss, an XRootD client to fetch the file from the national XRootD data federation

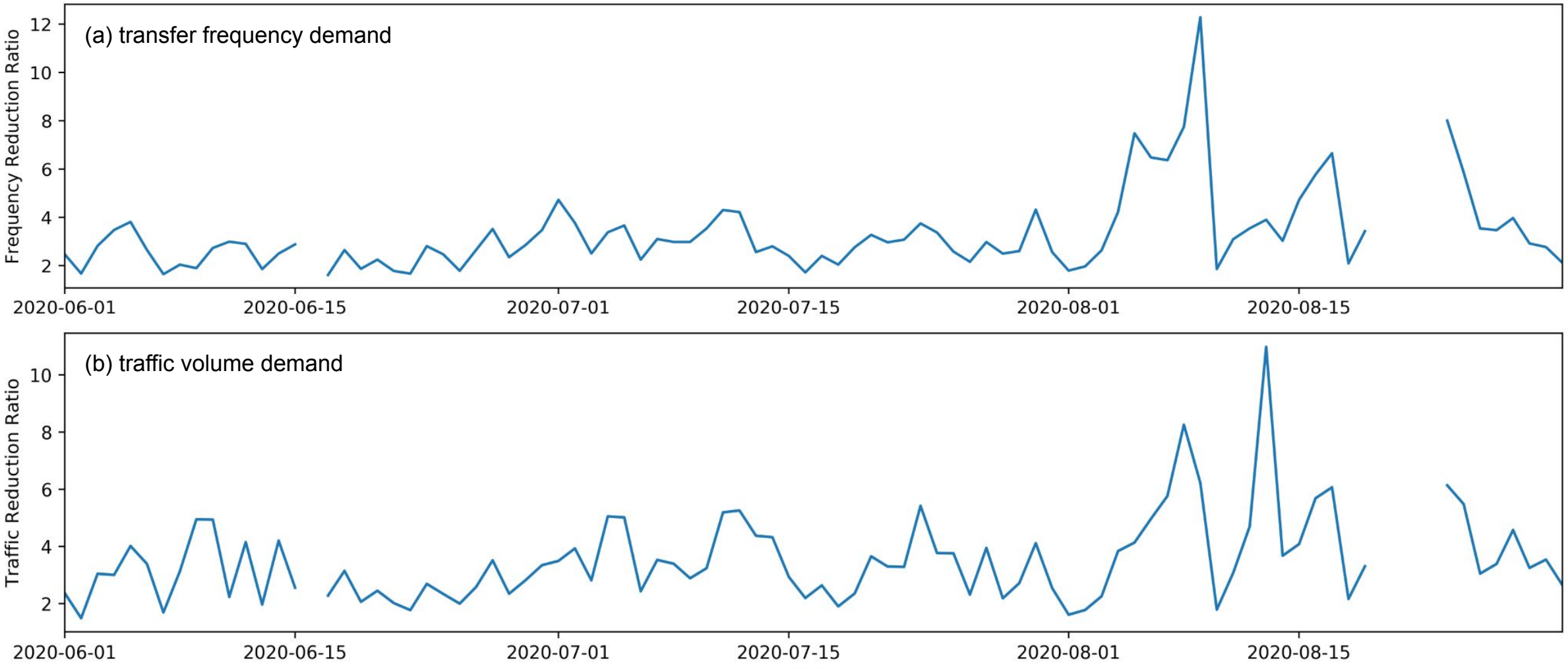
- Handling cache miss
  - Cache miss file goes to nodes with empty space first
    - nodes with empty spaces receive most of data accesses
      - cache miss goes to empty space
      - cache hits mostly on newly transferred data, which are previous cache misses
    - when new nodes added
      - new Caltech nodes (Xrd 3 - 8, 11) added around Aug 26th, 2021
      - new Caltech nodes (Xrd 9 - 10) added around Sept 30th, 2021
  - Delete old data if space are full
    - After the system running for sufficient time without adding new nodes
    - All cache nodes divide the data access more evenly

Compared to daily proportion of number of cache misses and cache hits (June 2020 - Aug 2020)



Network transfer savings = ~2.6 million transfers
Network demand frequency reduction rate = (Cache misses + Cache hits) / Cache misses = **2.617**

Compared to daily network demand reduction rates (June 2020 - Aug 2020)



Summary statistics for regional cache repo

	Number of accesses	Data transfer size (GB)	Shared data size (GB)	Percentage of shared data size
June 2020	1,804,697	532,037.7	818,956.9	60.62%
July 2020	1,426,585	354,452.8	764,351.3	68.32%
Aug 2020	995,324	249,583.5	586,188.8	70.14%
Total	4,226,606	1,136,074.0	2,169,497.0	65.63%
Daily average	48,029.61	12,909.93	24,653.37	

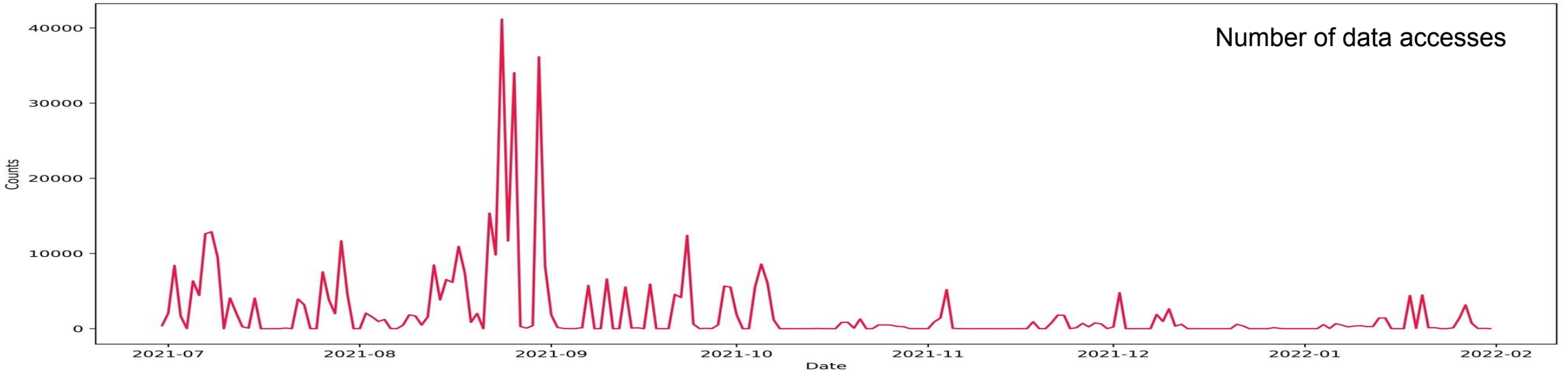
- Data transfer size (= first time data access size, cache misses): From remote sites to the local node cache
- Shared data access size (= repeated data accesses, cache hits, network bandwidth savings): From the local node cache to the application, excluding the first time accesses (data transfers)

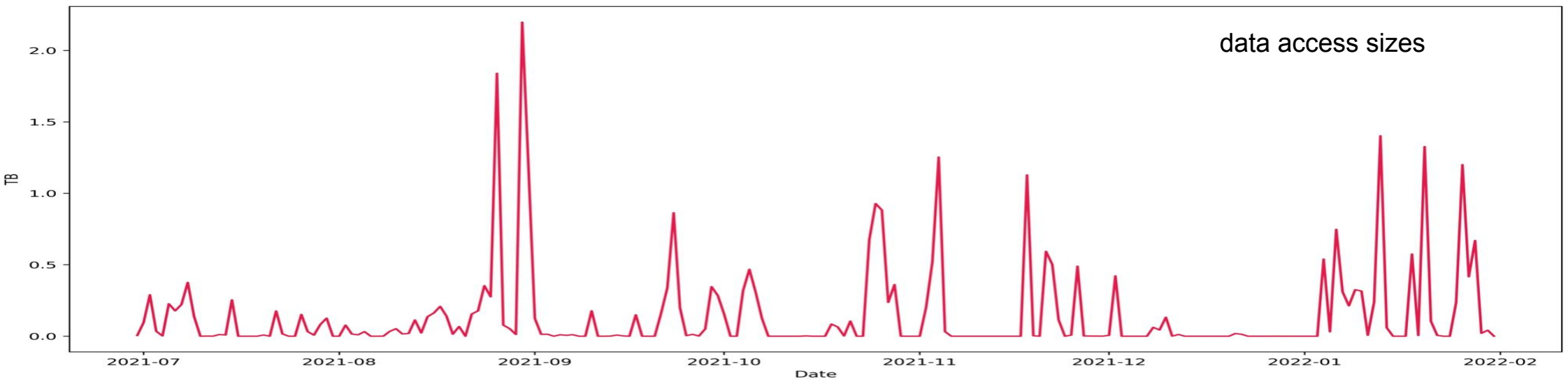
Summary data accesses Jan-May 2021

	Number of accesses	Data transfer size (TiB)	Shared data size (TiB)	Percentage of shared data size
Jan 2021	1,402,696	269.62	269.62	51.42%
Feb 2021	1,078,545	279.33	173.69	38.34%
Mar 2021	1,166,506	319.77	226.57	41.47%
Apr 2021	365,068	81.85	72.04	46.81%
May 2021	757,555	216.50	186.29	46.24%
Total	4,770,370	1152.14	928.21	44.62%
Daily average	32,451.50	7.84	6.31	

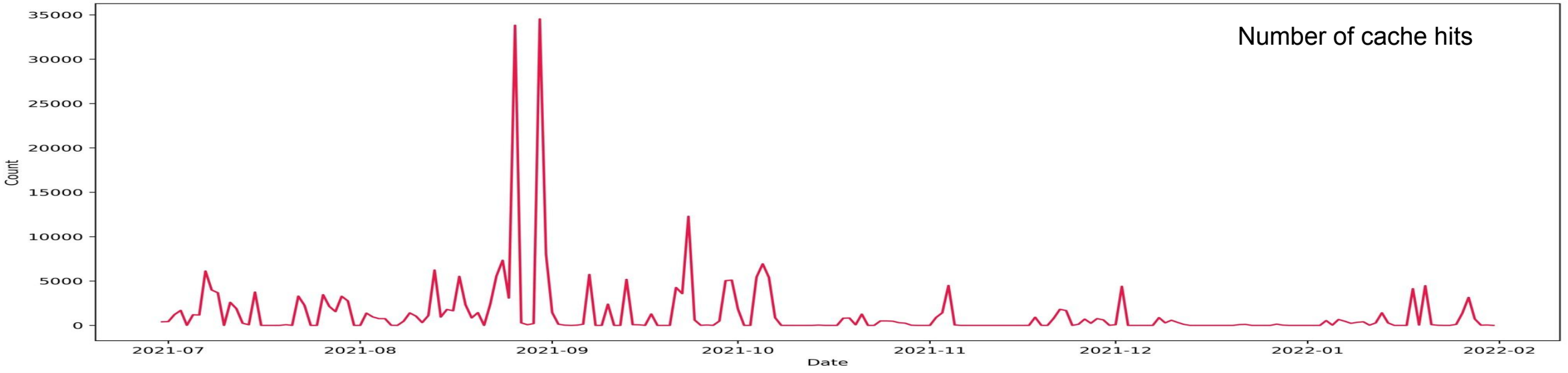
- Data transfer size (= first time data access size, cache misses): From remote sites to the local node cache
- Shared data access size (= repeated data accesses, cache hits, network bandwidth savings): From the local node cache to the application, excluding the first time accesses (data transfers)

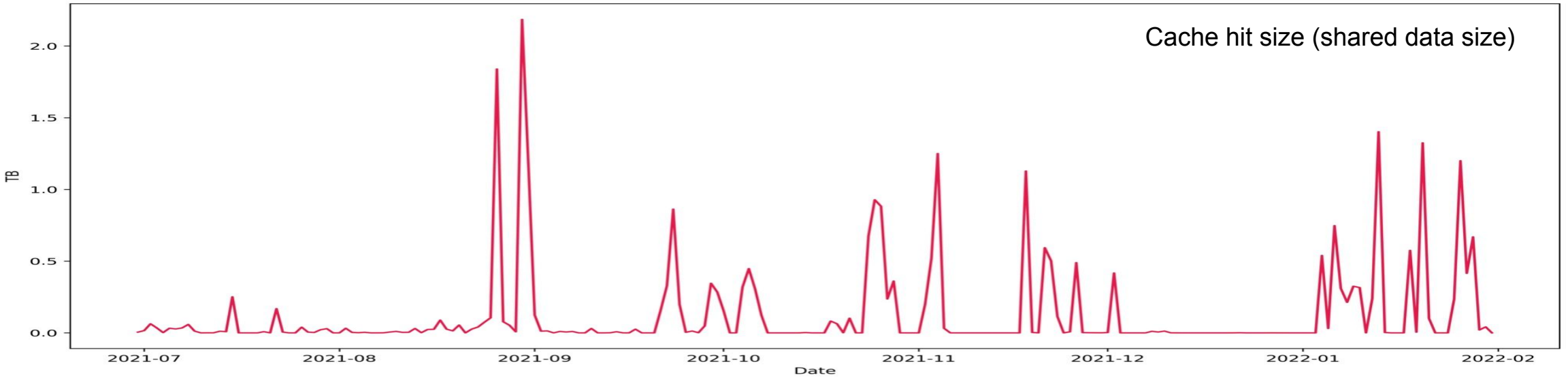
Daily data accesses for xrd-cache-1 (NanoAOD)



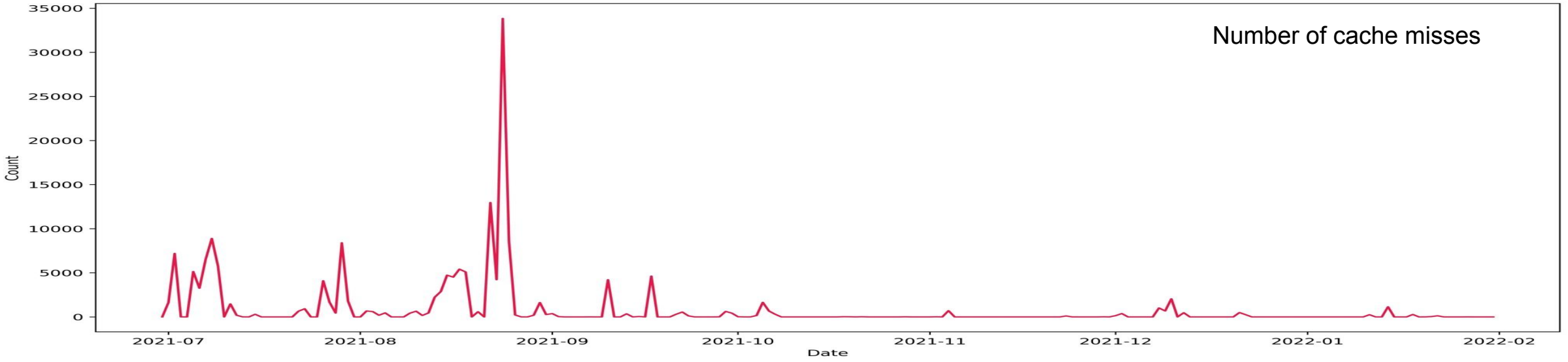


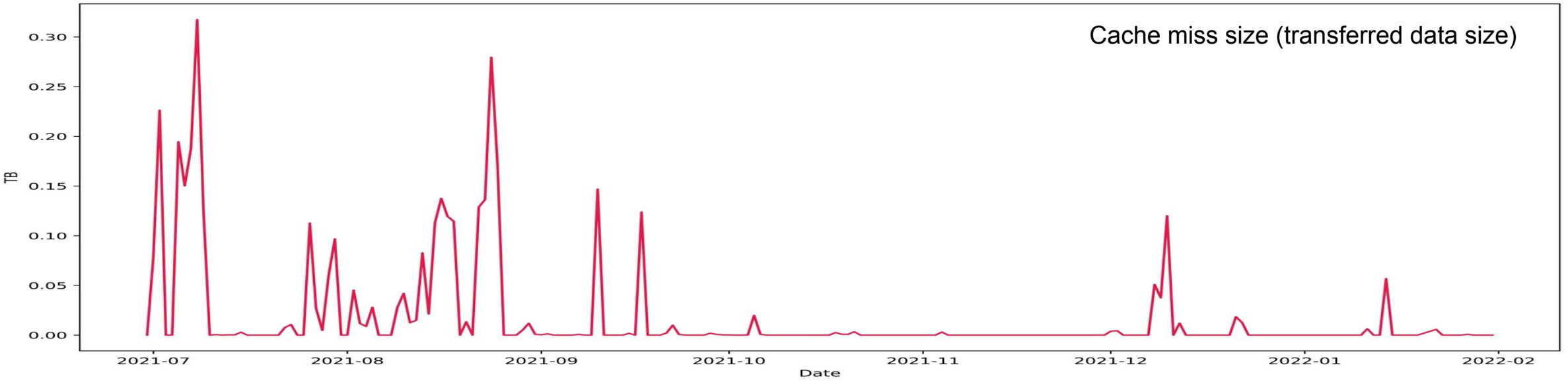
Daily cache hits for xrd-cache-1 (NanoAOD)



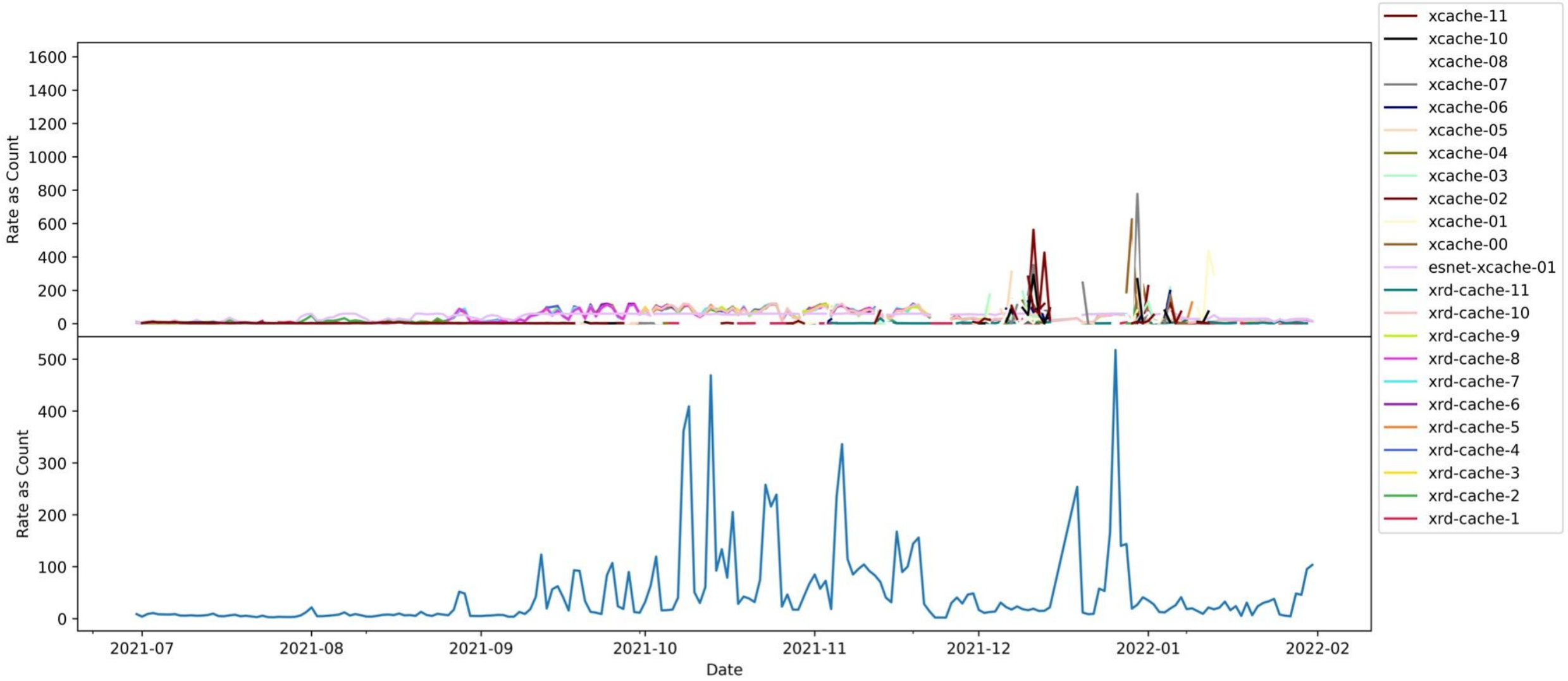


Daily cache misses of xrd-cache-1 (NanoAOD)

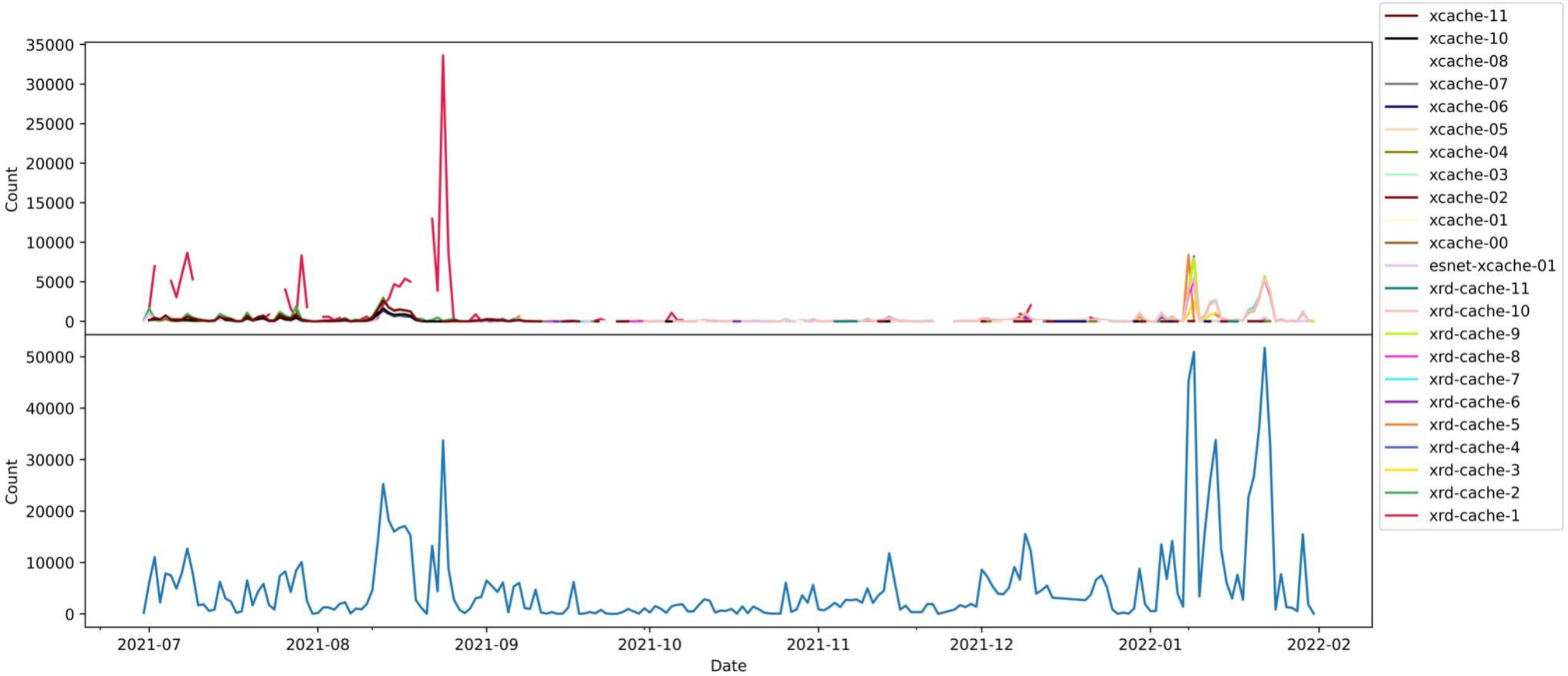




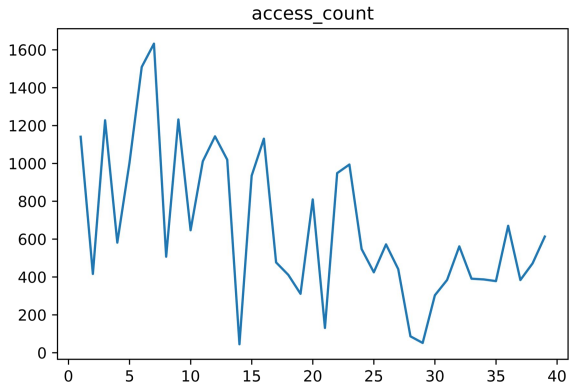
Daily data re-use rates (without log cale)

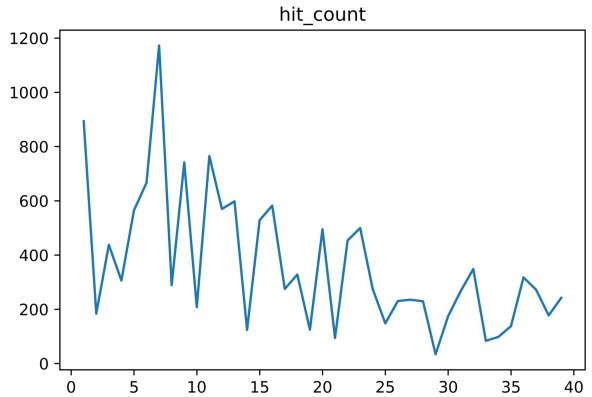


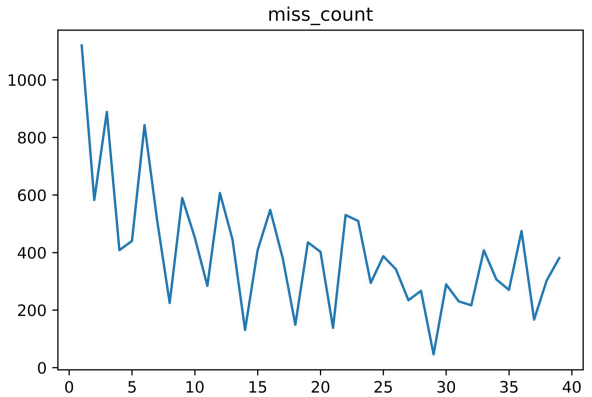
Daily number of file re-access (without log scale)

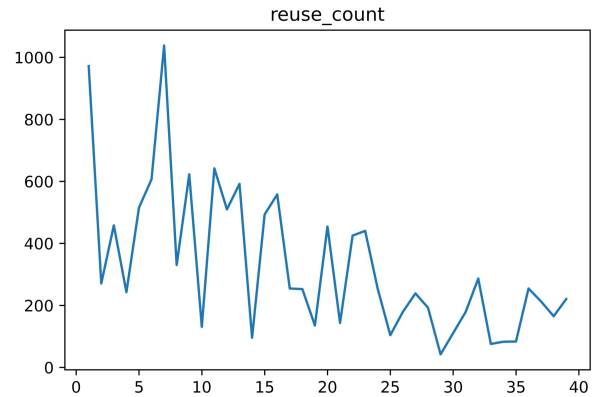


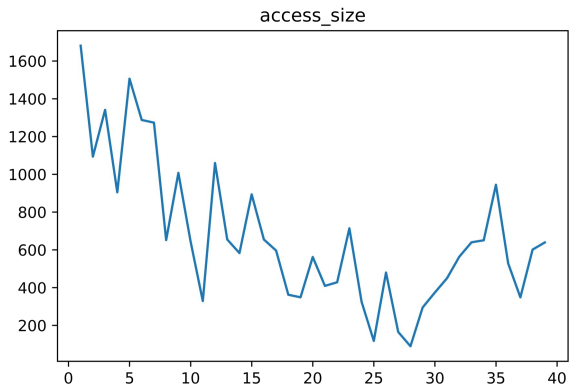
Periodogram of Daily data

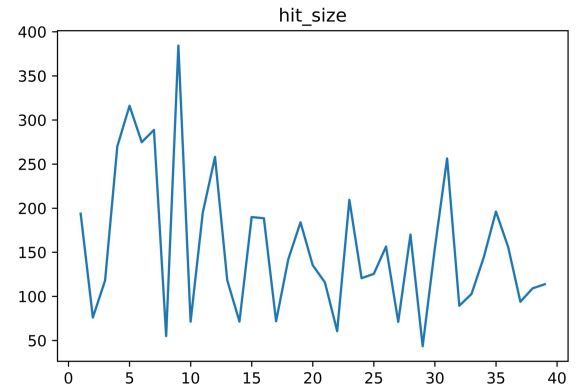


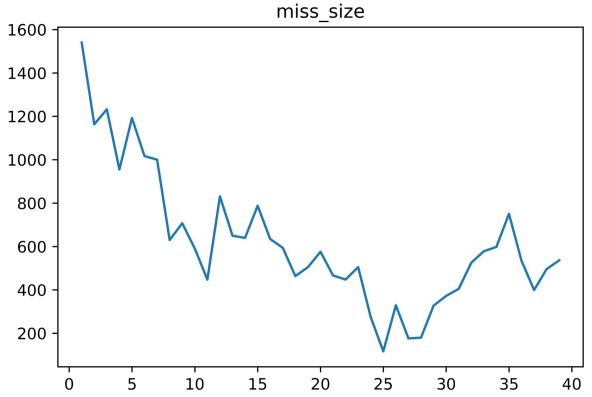


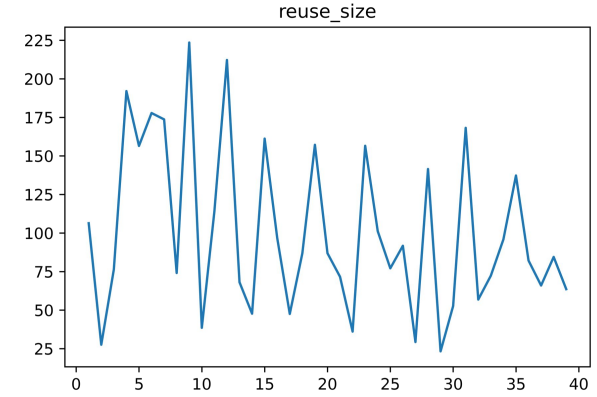












Hyperparameters for daily LSTM model

Explored Parameter:

Parameter	values
# of first layer LSTM unit	16, 32, 64, 128, 256
# of second layer LSTM unit	0, 16, 32, 64, 128, 256
first layer activation function	tanh, relu
second layer activation function	tanh, relu
dropout rate	0, 0.04, 0.1, 0.15
# of epchs	5, 10, 15, 25, 50, 75, 100

Model Parameter:

	# of LSTM units	activation function	dropout rate	# of training epoch
values	128	tanh	0.04	50

RMSE of Daily LSTM model

	Without day-of-week		With day-of-week		
	Training RMSE	Test RMSE	Train RMSE	Test RMSE	Accuracy
Access Count	3,861.14	4,944.34	3,492.61	4,220.19	0.93
Access Size	2,480.61	16,621.57	2,612.90	16,571.21	0.85
Cache Hit Count	2,459.72	3,158.99	2,179.03.99	2,917.99	0.95
Cache Hit Size	1,425.66	2,144.92	1,375.42	2,154.87	0.85
Cache Miss Count	2,261.62	2,954.13	2,302.29	2,970.10	0.91
Cache Miss Size	1,265.84	17,324.68	1,298.15	16,426.95	0.90
Data Reuse Count	2,224.82	3,066.91	2,063.65	2,646.69	0.93
Data Reuse Size	1,135.80	1,482.21	1,099.14	1,466.38	0.73

Hyperparameters for daily LSTM model with 7 day moving average

Explored Parameters:

Parameter	values
# of first layer LSTM unit	16, 32, 64, 128, 256
# of second layer LSTM unit	0, 16, 32, 64, 128, 256
first layer activation function	tanh, relu
second layer activation function	tanh, relu
dropout rate	0, 0.04, 0.1, 0.15
# of epchs	5, 10, 15, 25, 50, 75, 100

Model Parameters:

	# of LSTM units	activation function	dropout rate	# of training epoch
values	128	tanh	0.00	100

Daily LSTM model results with 7 day moving average

	Train RMSE of MA LSTM model	Test RMSE of MA LSTM model (reduction compare to Daily LSTM Model)	Accuracy	Test RMSE of Daily LSTM model
Access Count	1,122.15	2,169.72 (48.6%)	0.93	4,220.19
Access Size	744.56	7,729.04 (53.6%)	0.83	16,571.21
Cache Hit Count	829.23	2025.21 (30.6%)	0.88	2,917.99
Cache Hit Size	223.00	1,573.72 (27.1%)	0.91	2,154.87
Cache Miss Count	1,127.30	781.83 (73.0%)	0.86	2,970.10
Cache Miss Size	612.94	9616.83 (58.5%)	0.77	16,426.95
Data Reuse Count	808.80	1,228.71 (53.6%)	0.87	2,646.69
Data Reuse Size	208.27	812.33 (44.6%)	0.92	1,466.38