ESnet
ENERGY SCIENCES NETWORK

ESnet In-Network Caching Pilot

Chin Guok
Chief Technology Officer
Energy Sciences Network
Lawrence Berkeley National Laboratory

LHCOPN-LHCONE Meeting #50
Prague, Czech Republic
April 19, 2023

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

Observations (from a data movement POV)

- Large data volume from scientific experiments and simulations
  - Challenging for geographically distributed collaborations
    - E.g., Large Hadron Collider (LHC) from High-Energy Physics (HEP) community
  - Data stored at a few locations
    - Requiring significant networking resources for replication and sharing
    - Long latency due to the distance
      - ATLAS Tier-1 site at Brookhaven National Laboratory, USA
      - CMS Tier-1 site at Fermi National Accelerator Laboratory, USA
    - Network traffic primarily carried by Energy Sciences Network (ESnet)

- Significant portion of the popular dataset is used by many researchers

- Storage cache allows data sharing among users in the same region
    - Reduce the redundant data transfers over the wide-area network
    - Decrease data access latency
    - Increase data access throughput
    - Improve overall application performance

ESnet

What is the objective (from a network POV)?

- Reduction of network bandwidth utilization
  - Science is a collaborative endeavor, implying common data sets being shared with different organizations.
  - Scientific data sets are growing exponentially, resulting in larger data movement requirements.
  - Scientific collaborations are borderless, requiring wider geographic footprints with corresponding network connectivity needs.
- “Dictating” the usage of the network
  - Understanding how data sets are shared, provides insight on network designed and traffic engineering.
  - Sharing network feedback to the data movement to schedule transfer
    - E.g., delaying a transfer to during peak congestion periods.
  - Integrating data movement requirements to (dynamically) provision the network to accommodate transfers
    - E.g., provisioning guaranteed bandwidth temporary circuits to bypass congestion points for large data transfers.

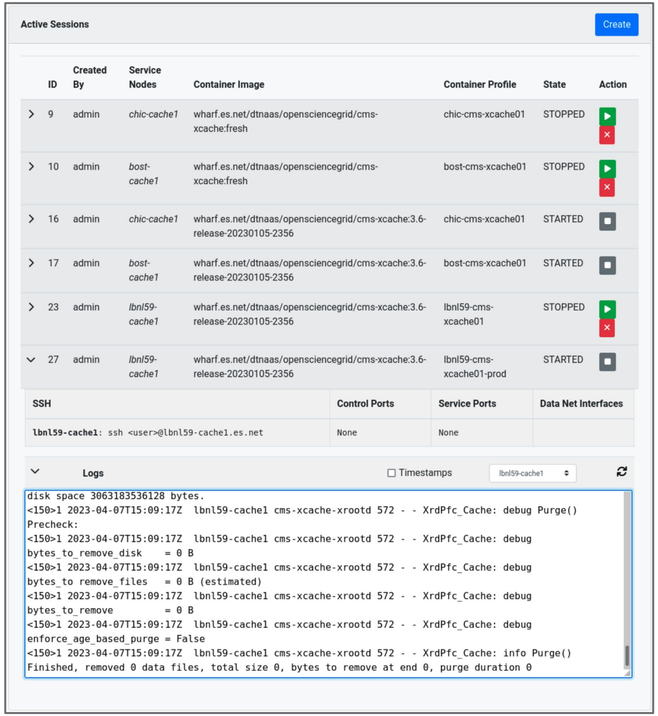
ESnet

Goals of the caching pilot

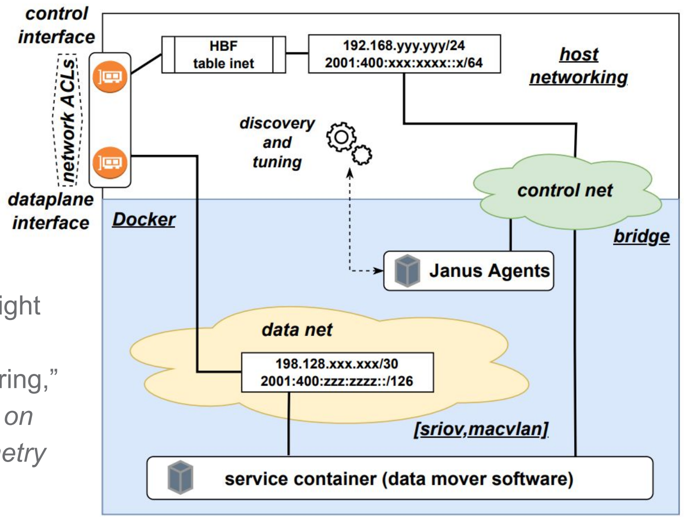
- Understand the networking characteristics
  - Explore measurements from Southern California Petabyte Scale Cache (SoCal Repo)
  - Characterise the trends of network and cache utilization
  - Study the effectiveness of in-network caching in reducing network traffic
- Explore the predictability of the network utilization
  - Help guide additional deployments of caches in the science network infrastructure
- Overall, study the effectiveness of the cache system for scientific applications

DTNaaS - Containerized DTN deployment model

- Janus is used to deploy DTNaaS for the ESnet In-Network caching pilot



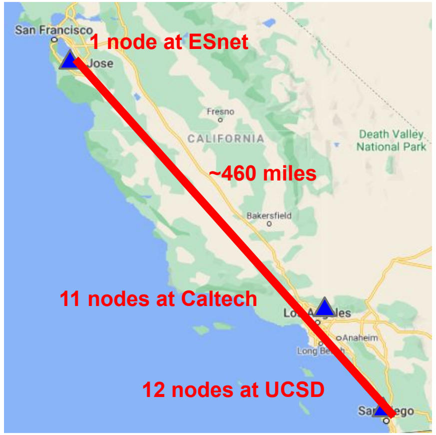
- Janus software provides:
  - Live profile updates and schema validation
  - A web-based user interface called Janus Web
  - Packaging of the Janus controller and open source availability on PyPI
  - Ansible-based deployment automation



Kissel, Ezra "Janus: Lightweight Container Orchestration for High-Performance Data Sharing," *Fifth International Workshop on Systems and Network Telemetry and Analytics*, June 2022

Southern California Petabyte Scale Cache (SoCal Repo)

- SoCal Repo consists of 24 federated storage nodes for US CMS
  - 12 nodes at UCSD: each with 24 TB, 10 Gbps network connection
  - 11 nodes at Caltech: each with storage sizes ranging from 96TB to 388TB, 40 Gbps network connections
  - 1 node at LBNL (by ESnet): 44 TB storage, 40 Gbps network connection
  - Approximately 2.5PB of total storage capacity
  - ~100 miles between UCSD and Caltech nodes, round trip time (RTT) < 3 ms
  - ~460 miles between LBNL and UCSD nodes, RTT ~10 ms



Sunnyvale–San Diego is the relevant distance scale

- Statistics about US CMS data analysis with MINIAOD/NANOAOD
  - 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**
  - More than 90% of analyses work with either MiniAOD or NanoAOD

Data Access Summary*

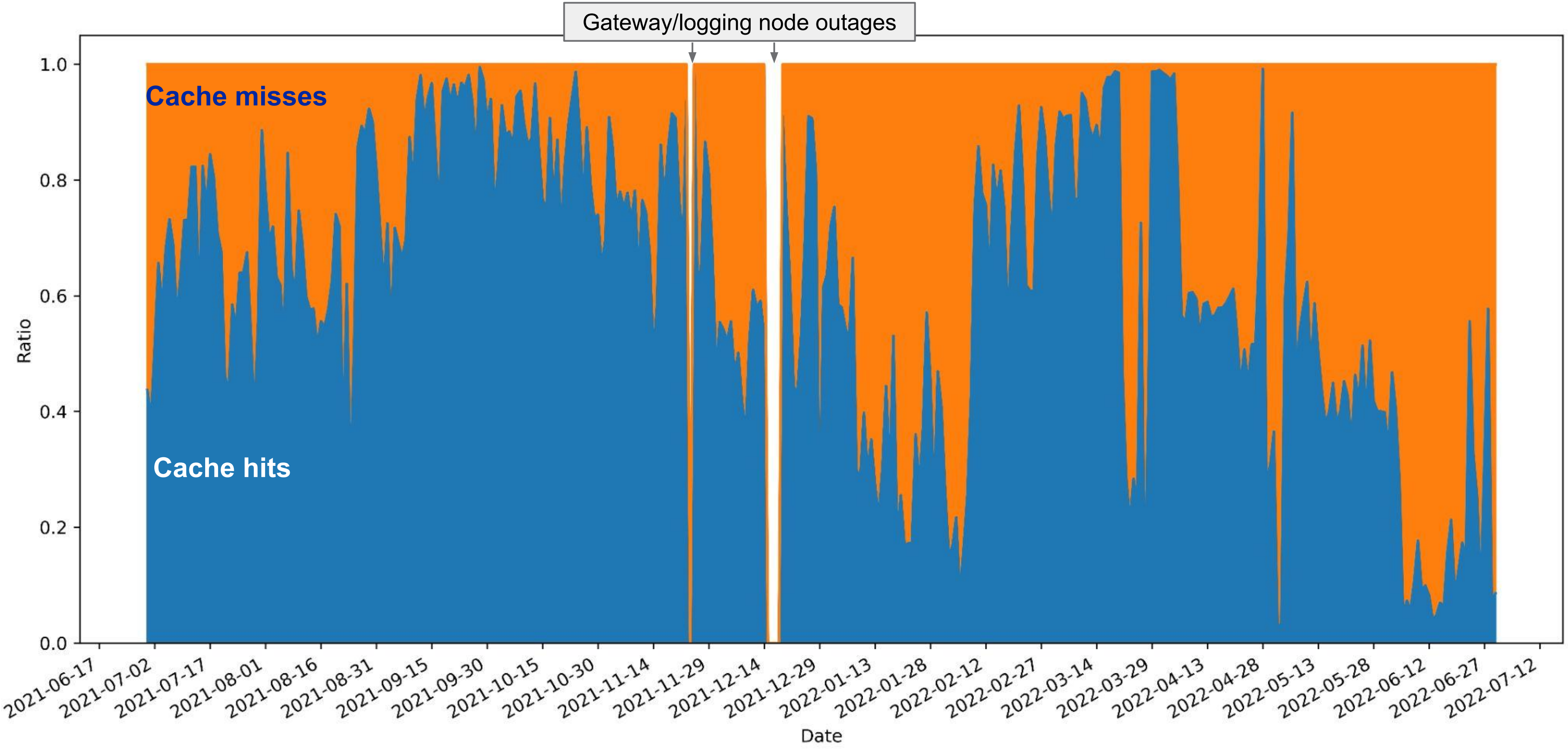
	# of accesses	Data transfer size (TB)	Shared data size (TB)	# of cache misses	# of cache hits
Total	8,713,894	8,210.78	4,499.44	2,822,014	5,891,880
Daily average	23,808	22.43	12.29	7,710	16,098

- Consisting of 8.7 million file requests between July 2021 and June 2022
- 4.5PB (35.4%) of requested bytes (out of 12.7PB) could be served from the cache
- 67.6% of file requests are satisfied by the cache

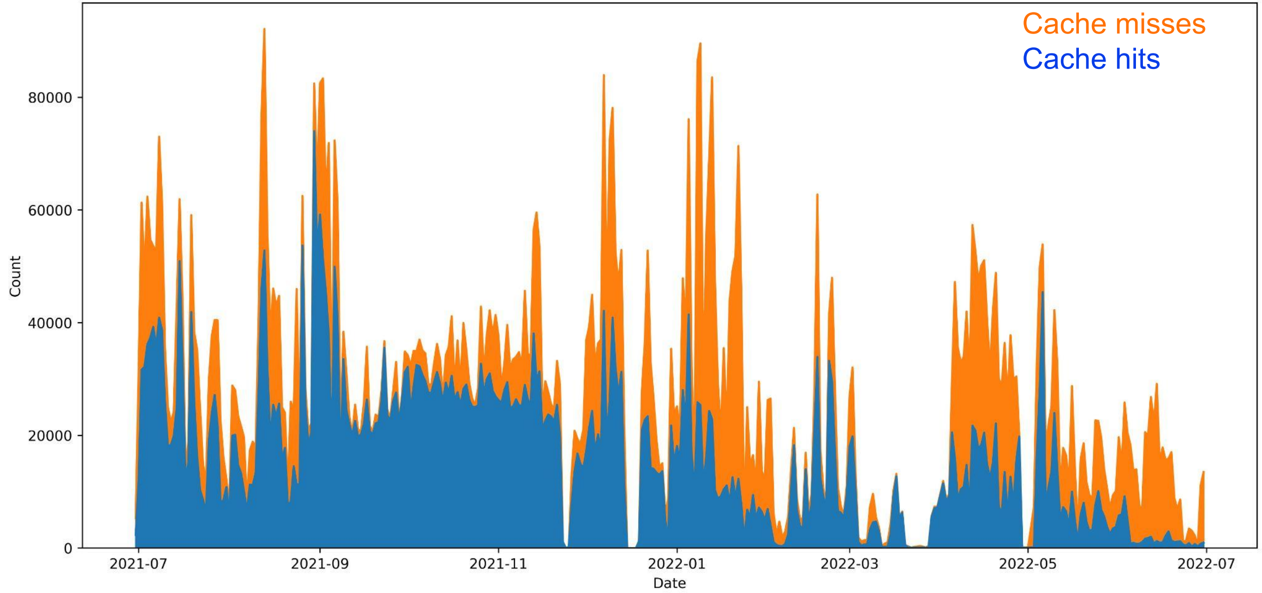
**NB: Data used for the analysis is from 1-year of SoCal Repo's operational logs from July 2021 to June 2022 (~8,433 log files, ~3TB)*

ESnet

About 2/3rd of daily file requests satisfied by SoCal Repo



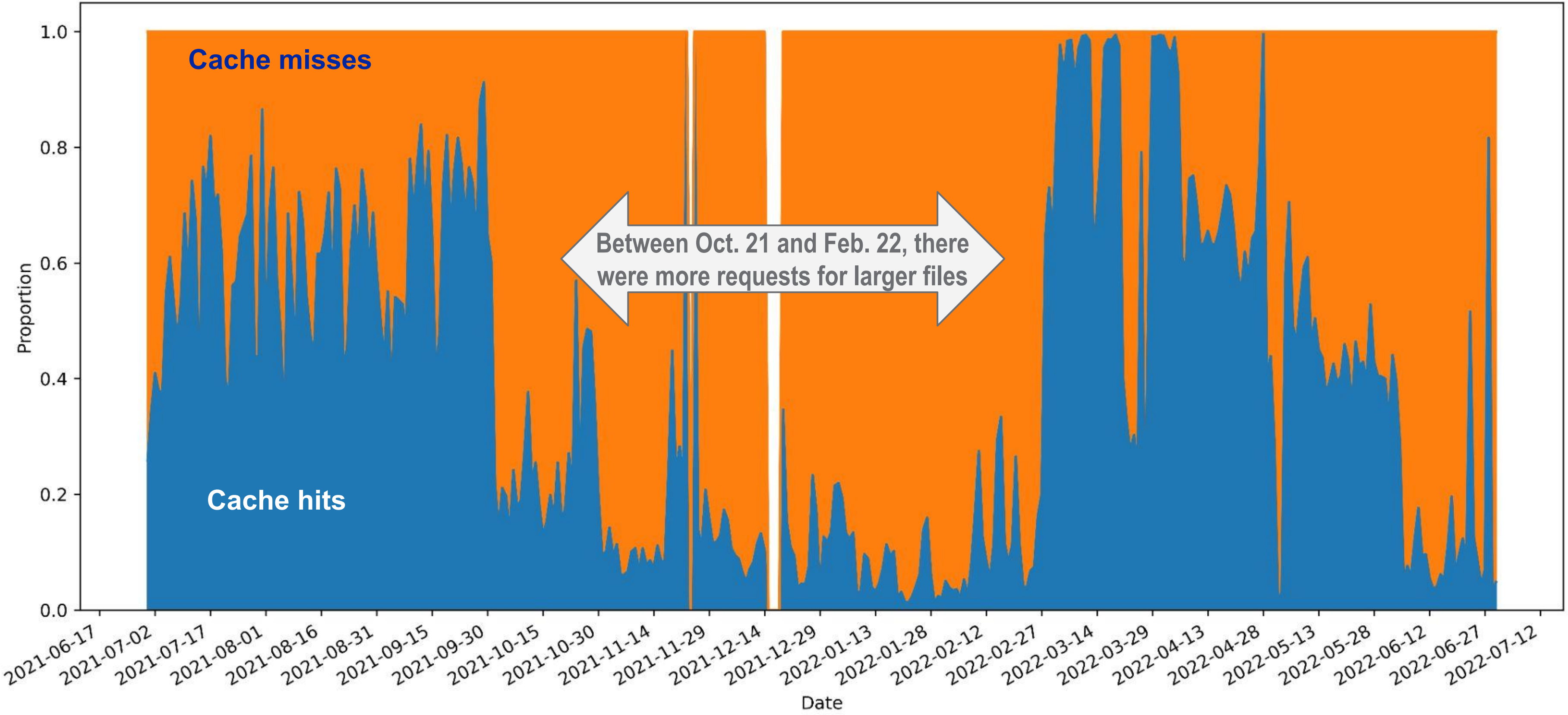
File requests per day, some days peaking to nearly 100,000 requests



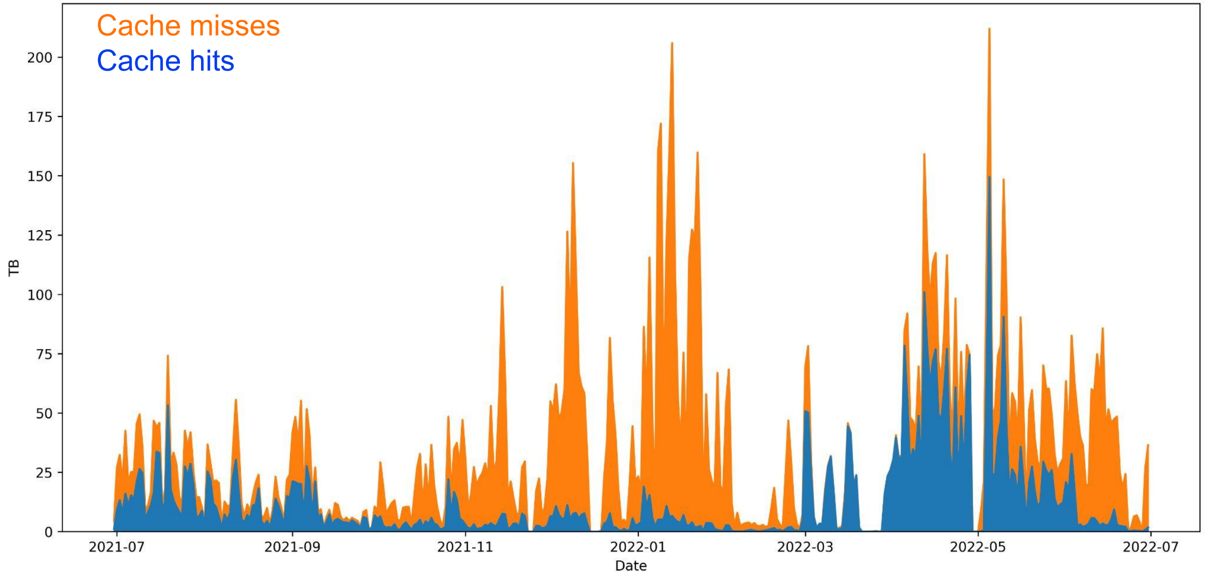
- On average, ~16,000 file requests per day are served from the storage cache nodes (i.e., cache hits), while 8,000 requests are cache misses
- Only file requests that miss the cache trigger remote file transfers

ESnet

Fraction of daily requested bytes varies significantly during different time periods

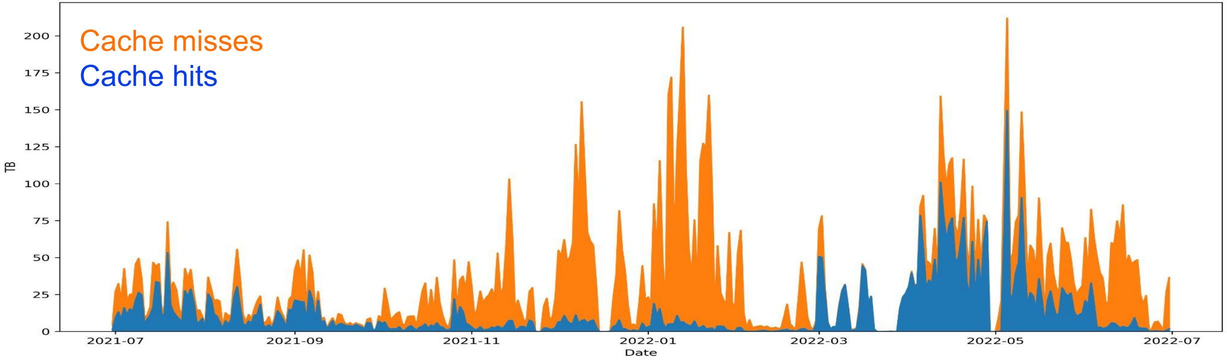


Bytes requested can peak to 200 TB daily



- On average, 12.3 TB per day are served out of the cache during the whole year
- Between Jul 2021 and Sep 2021, the network traffic is reduced by ~13 TB per day
- Between Mar 2022 and May 2022, the network traffic is reduced by ~29TB per day

Cache usage involving large files



- On Jan 13, 2022, there were ~60K cache misses with ~200TB of network traffic (vs ~20K cache hits with ~15TB)
  - On average, each of these files were about 3.3GB
  - These files were requested by a small number of data analyses jobs involving larger files
  - Challenge: This particular usage pattern has the potential of evicting the smaller files (that are used more frequently) and reducing the overall effectiveness of the cache system
  - **Solution 1**: Separated the accesses to the cache nodes based on file types, which effectively prevents cache pollution
  - **Solution 2**: In cases where the cache usages couldn't be differentiated based on simple known characteristics, an alternative strategy could be to have those requests bypass the cache system

ESnet

Summary observations

- SoCal Repo could serve on average about **67.6% of files** from its disk cache, while on average only **35.4% of bytes** requested could be served from the cache
  - Because the large files are less likely to be reused
  - To avoid cache pollution from this particular usage pattern with large files, the operators have separated the two different types of files requests with different storage nodes.
- Over the whole period of observation, there is a five-month period where the large file requests are noticeably high, resulting in an average reduction of wide-area network traffic of about **12.3TB per day**
- During the period where fewer large files were requested (3/2022 – 5/2022), the network traffic was reduced by about **29TB per day**

ESnet

What's next?

- Follow on usage analysis of ESnet's Chicago and Boston caching nodes.
  - Chicago DTNaaS will support CMS use case in collaboration with University of Wisconsin (Madison), Notre Dame, and Purdue.
  - Boston DTNaaS will support CMS use case in collaboration with MIT.
- Deployment of additional caching nodes in Amsterdam and London.
  - Both Amsterdam and London DTNaaS will support DUNE/LIGO use cases mainly in collaboration with Open Science Data Federation (OSDF).
- Deployment of multiple DTNaaS instances of on a physical caching node.
  - Chicago DTNaaS to support LHCb use case.
  - Amsterdam DTNaaS to support Protein Data Bank (PDB) use case.

ESnet

Publications and Presentations

1. C. Sim, K. Wu, A. Sim, I. Monga, C. Guok, F. Wurthwein, D. Davila, H. Newman, J. Balcas, "Effectiveness and predictability of in-network storage cache for Scientific Workflows", International Conference on Computing, Networking and Communication (ICNC 2023), 2/2023. https://sdm.lbl.gov/oapapers/icnc23-xcache-sim.pdf

2. C. Sim, C. Guok, A. Sim, K. Wu, "Data Throughput Performance Trends of Regional Scientific Data Cache", ACM/IEEE The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'22), ACM Student Research Competition (SRC), 11/2022. https://sdm.lbl.gov/oapapers/sc22-src-poster-sim.pdf

3. R. Han, A. Sim, K. Wu, I. Monga, C. Guok, F. Würthwein, D. Davila, J. Balcas, H. Newman, "Access Trends of In-network Cache for Scientific Data", 5th ACM International Workshop on System and Network Telemetry and Analysis (SNTA), in conjunction with The 31st ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC), 6/2022, doi:10.1145/3526064.3534110

4. A. Sim, "Data Access Trends in Southern California Petabyte Scale Cache", WLCG Data Organization, Management and Access (DOMA) general meeting, CERN, 5/2022.

5. A. Sim, E. Kissel, C. Guok, "Deploying in-network caches in support of distributed scientific data sharing", the US Community Study on the Future of Particle Physics (Snowmass 2021), 3/2022. doi:/10.48550/arXiv.2203.06843. https://arxiv.org/abs/2203.06843.

6. E. Copps, A. Sim, K. Wu, "Analyzing scientific data sharing patterns with in-network data caching", ACM Richard Tapia Celebration of Diversity in Computing (TAPIA 2021), ACM Student Research Competition (SRC), 9/2021. https://sdm.lbl.gov/oapapers/tapia21-copps-poster.pdf

7. E. Copps, H. Zhang, A. Sim, K. Wu, I. Monga, C. Guok, F. Würthwein, D. Davila, E. Fajardo, "Analyzing scientific data sharing patterns with in-network data caching", 4th ACM International Workshop on System and Network Telemetry and Analysis (SNTA 2021), 6/2021, doi:10.1145/3452411.3464441

8. A. Sim, "Exploring in-network data caching, ESnet-US CMS collaboration study", LHC GDB meeting, CERN, 2/2021.

ESnet

Acknowledgements

Chin Guok, Damian Hazen, Ezra Kissel, Inder Monga
(ESnet)

Alex Sim, K. John Wu
(Lawrence Berkeley National Lab)

Caitlin Sim, Jack Ruize Han
(Univ. of California at Berkeley)

Ellie Copps
(Middlebury College)

Diego Davila, Frank Wuerthwein
(Univ. of California at San Diego)

Justas Balcas, Harvey Newman
(Caltech)

Questions…

Chin Guok <chin@es.net>

ESnet