Analyzing Scientific Data Sharing Patterns for In-Network Data Caching

Elizabeth Copps

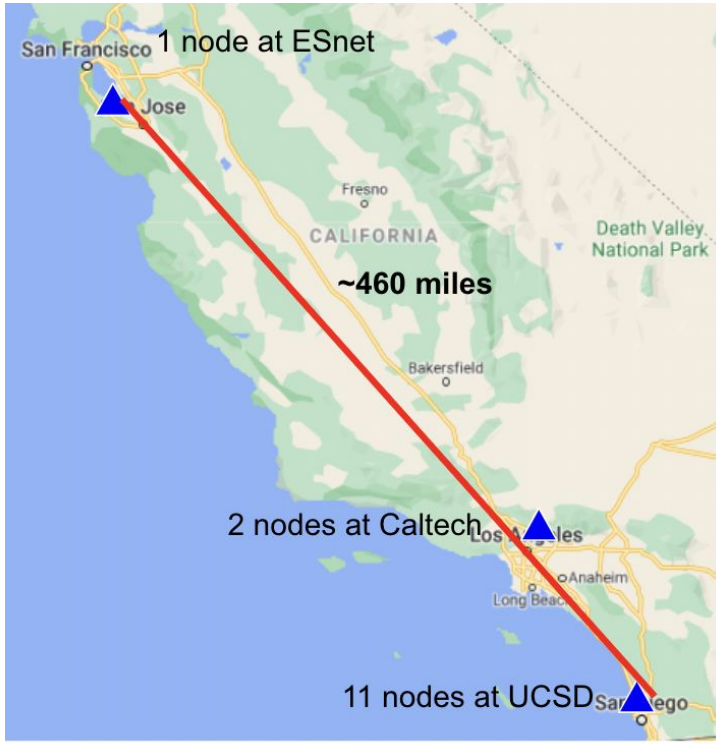
Middlebury College

Introduction

- Data volume from new scientific projects and simulations exponentially increasing
  - Drives up network traffic, data latency, and total data transfer costs
- Problem: Many files accessed multiple times by the same user or users in the same region
- Solution: In-network regional data caches
  - Reduce costly and redundant transfers by sharing data among regional users
  - Previously shown to reduce network demand by factor of ~3

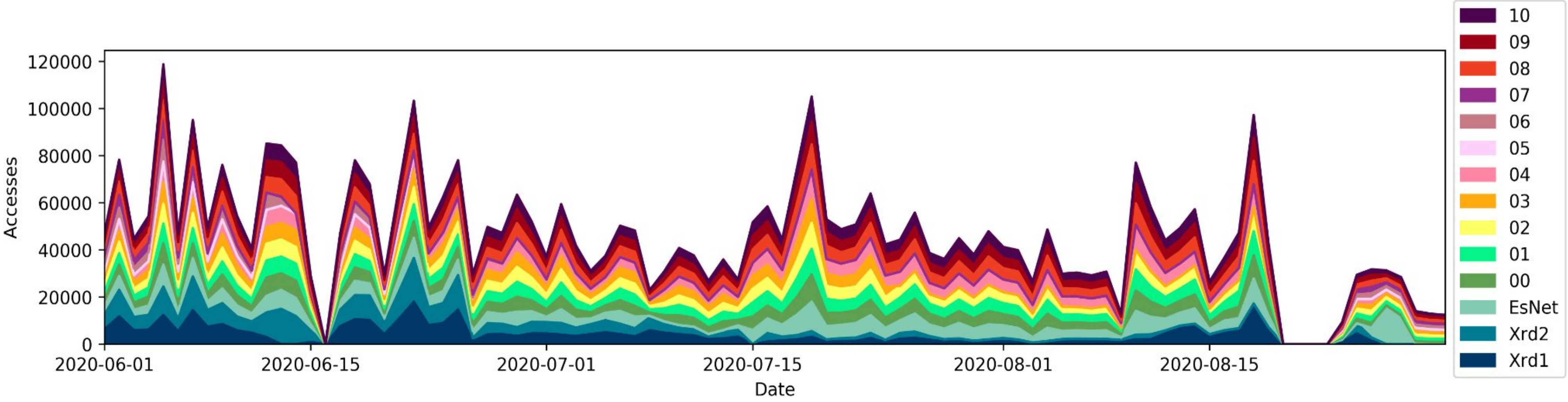
Southern California Regional Cache

- 14 XCache installations deployed in Southern California
  - 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
  - XCache fetches data from source to store locally
- Cache relevant from research standpoint
  - Spans almost 500 miles
  - Data from Large Hadron Collider



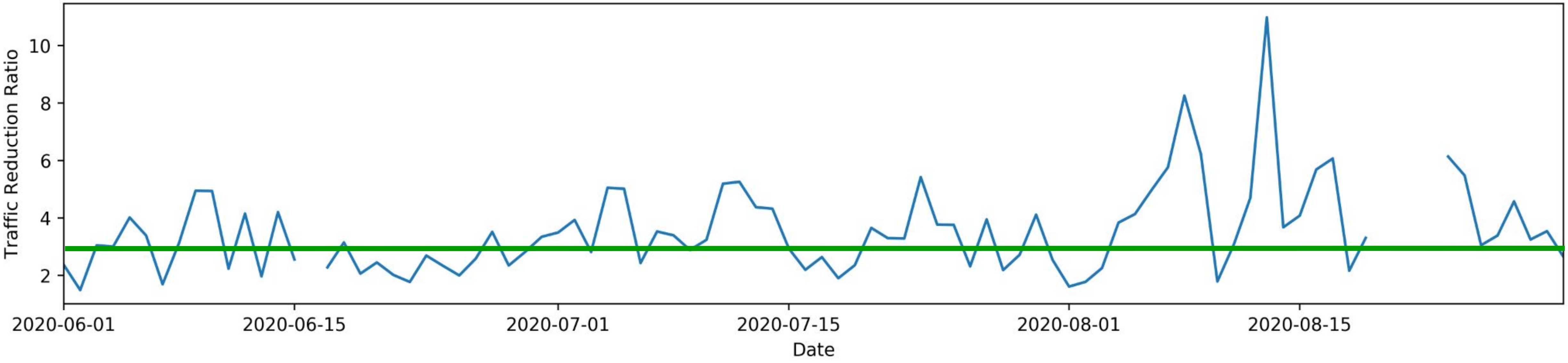
Summary Statistics & Cache Utilization

	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	2,169.50	65.63%
Daily average	48,029.61	12.91	24.65	



Traffic Reduction Ratio

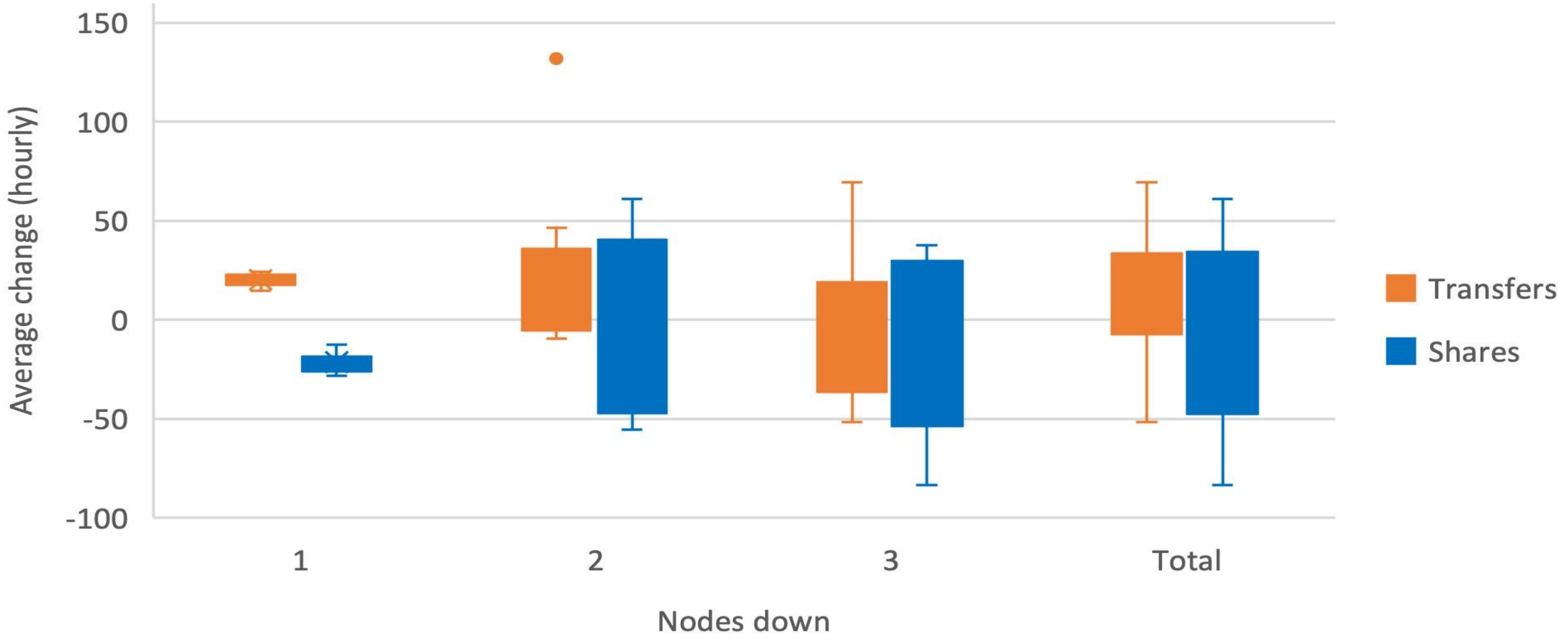
- **Ratio measures the sizes of the data that the cache shares rather than transfers**
  - **Traffic Reduction Ratio = Total access size / Total transfer size**
  - **Total access size = Share size + Transfer size**
- **Average ratio: 2.91**



Impact of a single node

- **Studied node downtimes to analyze impact of a single node**
  - Downtimes reduce number of nodes in cache
  - Predict what happens when we add nodes

- **Downtime changes**
  - Hourly shares and share sizes decrease
  - Hourly transfers and transfer sizes increase
  - Remaining nodes evenly split load



Conclusions

- Proportion of data volume being shared increases over time
    - More files in cache
    - Increases traffic reduction rate
    - Increase stops once cache is full
- Single node affects rest of cache proportionally
    - More transfers after downtime because less data stored in cache
- Future
    - Adding nodes expected to proportionally reduce demands on other nodes
        - Larger nodes store more data but do not take proportionally more accesses
        - More total disk space = higher traffic reduction rate
    - As users increase, accesses increase proportionally