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
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.
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

- Statistics about US CMS data analysis with MINIAOD/NANOAOD
  - Analysis Object Data (AOD):
    - 384 PB of RAW
    - 240 PB of AOD
    - 30 PB of MINIAOD
    - 2.4 PB of NANOAOD
    - Mostly on Tape: accessed a few times per year
    - Mostly on disk: heavily re-used by many researchers
  - More than 90% of analyses work with either MiniAOD or NanoAOD
Data Access Summary*
(Jul 2021 - Jun 2022 study)

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

- Consisting of 8.7 million file requests between July 2021 and June 2022
- 5.9M (67.6%) file requests (out of 8.7M) were satisfied by the cache
- 4.5PB (35.4%) of requested bytes (out of 12.7PB) were served from the cache

*NB: Data used for the analysis is from 12 months of SoCal Repo’s operational logs from July 2021 to June 2022 (~8,433 log files, ~3GB)
Daily average file requests - 23,808 files
(Jul 2021 - Jun 2022 study)

- On average, 16,098 file requests per day were served from the storage cache nodes (i.e., cache hits), while 7,710 requests were cache misses.
- Daily file requests peaked at ~100K.
67.6% (average) of daily files requested were cache hits (Jul 2021 - Jun 2022 study)
Average daily bytes requests - 34.72TB
(Jul 2021 - Jun 2022 study)

• On average, 12.29TB per day were served from the storage cache nodes (i.e., cache hits), while 22.43TB were cache misses
• Daily byte requests peaked at 200TB
35.4% (average) of daily bytes requested were cache hits
(Jul 2021 - Jun 2022 study)

between Oct. 21 and Feb. 22, there were more requests for larger files
Cache usage involving large files - digging deeper (Jul 2021 - Jun 2022 study)

On Jan 13, 2022, there were ~60K file cache misses requiring ~200TB of data to be fetched (vs ~20K file cache hits with ~15TB of data reuse)

- On average, each of these files were about 3.3GB
- These files were requested by a small number of processing jobs
- On further analysis it was determined that files could be grossly divided into:
  - Preprocessing jobs - large files, single use
  - Analysis jobs - small files, multiple uses

**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
### Data Access Summary*
(Jul 2022 - Mar 2023 study)

<table>
<thead>
<tr>
<th></th>
<th># of accesses</th>
<th>Data transfer size (TB)</th>
<th>Shared data size (TB)</th>
<th># of cache misses</th>
<th># of cache hits</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>3,615,578</td>
<td>560.96</td>
<td>5,208.91</td>
<td>663,994</td>
<td>2,951,584</td>
</tr>
<tr>
<td><strong>Daily average</strong></td>
<td>13,147</td>
<td>2.04</td>
<td>18.94</td>
<td>2,414</td>
<td>10,733</td>
</tr>
</tbody>
</table>

*NB: Data used for the analysis is from 9 months of SoCal Repo’s operational logs from July 2022 to March 2023 (~5,838 log files, ~3GB)*

- Consisting of 3.6 million file requests between July 2022 and March 2023
- 3.0M (81.6%) file requests (out of 3.6M) were satisfied by the cache
- 5.2PB (90.2%) of requested bytes (out of 5.8PB) were served from the cache
Daily average file requests - 13,147 files
(July 2022 - Mar 2023 study)

- On average, 10,733 file requests per day were served from the storage cache nodes (i.e., cache hits), while 2,414 requests were cache misses
- Daily file requests peaked at ~55K
81.6% (average) of daily files requested were cache hits
(July 2022 - Mar 2023 study)
Average daily bytes requests - 20.98TB
(July 2022 - Mar 2023 study)

- On average, 18.94TB per day were served from the storage cache nodes (i.e., cache hits), while 2.04TB were cache misses
- Daily byte requests peaked at 140TB
90.2% (average) of daily bytes requested were cache hits
(July 2022 - Mar 2023 study)
### Summary observations

#### July 2021 - Jun 2022 study

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

#### July 2022 - Mar 2023 study

<table>
<thead>
<tr>
<th></th>
<th># of accesses</th>
<th>Data transfer size (TB)</th>
<th>Shared data size (TB)</th>
<th># of cache misses</th>
<th># of cache hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>3,615,578</td>
<td>560.96</td>
<td>5,208.91</td>
<td>663,994</td>
<td>2,951,584</td>
</tr>
<tr>
<td>Daily average</td>
<td>13,147</td>
<td>2.04</td>
<td>18.94</td>
<td>2,414</td>
<td>10,733</td>
</tr>
</tbody>
</table>

**Solution 1:** Separated the accesses to the cache nodes based on file types, which effectively prevents cache pollution

- July 2021 - Jun 2022 study identified large file single use patterns
- Investigation uncovered unique data requirements between Preprocessing and Analysis jobs
- Solution was to separate Preprocessing and Analysis data caching on distinct nodes
- July 2022 - Mar 2023 study shows significant cache hit improvements over July 2021 - Jun 2022 study
  - File cache hits: **81.6%** (Jun 2022 - Mar 2023 study), **67.6%** (Jul 2022 - Jun 2022 study)
  - Bytes cache hits: **90.2%** (Jun 2022 - Mar 2023 study), **35.4%** (Jul 2022 - Jun 2022 study)
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.
  – Boston DTNaaS to support LHCb use case.
  – Amsterdam DTNaaS to support Protein Data Bank (PDB) use case.
Publications and Presentations


Acknowledgements

- Alex Sim, K. John Wu (Lawrence Berkeley National Lab)
- Chin Guok, Damian Hazen, Ezra Kissel, Inder Monga, Engineering team, Infrastructure team (ESnet)
- Diego Davila, Dmitry Mishin, Fabio Adnrijauskas, Frank Wuerthwein (Univ. of California at San Diego)
- Justas Balcas, Harvey Newman (Caltech)
- Caitlin Sim, Jack Ruize Han (Univ. of California at Berkeley)
- Ellie Copps (Middlebury College)

- National Science Foundation through the grants OAC-2030508, OAC-1836650, MPS-1148698, PHY-2121686 and OAC-1541349
- US Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program
- Used resources of the National Energy Research Scientific Computing Center (NERSC)
Questions…

Chin Guok <chin@es.net>