ESnet
ENERGY SCIENCES NETWORK

Experiences in deploying in-network data caches

Ezra Kissel, Chin Guok (ESnet)
Alex Sim (LBNL)

CHEP 2023 - Track 7
Norfolk, VA
May 9th, 2023

U.S. DEPARTMENT OF ENERGY
Office of Science

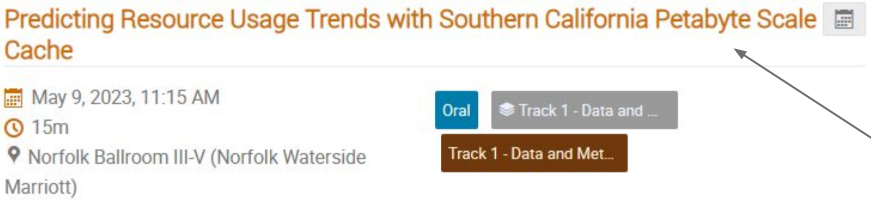
BERKELEY LAB

Motivating in-network caching

- Data volumes continue to grow at a dramatic rate
  - Scientific instruments, simulations, IoT and sensor networks, etc.
- A significant portion of popular datasets are re-used during analysis
- Storage caching allows data sharing among users in the same region
  - Reduce the repeated data transfers over the wide-area network
  - Decrease data access latency
  - Increase data access throughput
  - Improve overall application performance
- In-network caching presents opportunities to better dictate usage

ESnet data caching pilot

- Support geographically distributed collaborations
  - Large Hadron Collider (LHC) from High-Energy Physics (HEP) community
- Deploy regional caching nodes and understand their impact
- Use case: Southern California Petabyte Scale Cache (SoCal Repo)



Predicting Resource Usage Trends with Southern California Petabyte Scale Cache

May 9, 2023, 11:15 AM

15m

Norfolk Ballroom III-V (Norfolk Waterside Marriott)

analysis talk
earlier today

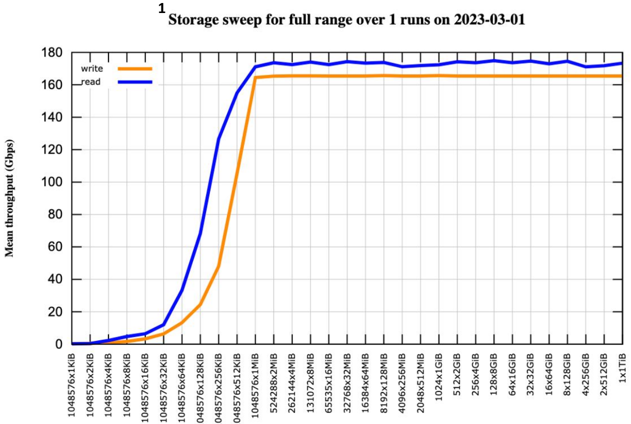
- Goals:
  - Characterise the trends of network and cache utilization
  - Study the effectiveness of in-network caching in reducing network traffic
  - Study the effectiveness of the cache system for scientific applications
  - Explore the logistics of hosting data movement services within an international science network such as ESnet

Deployment status

- 3 active nodes, 2 being provisioned
  - Boston, Chicago, Sunnyvale
  - London, Amsterdam coming soon
- System specs (latest)

SuperMicro SYS-2029U-TN24R4T
2x Xeon 5220S (18 cores, 2.7Ghz)
20 Samsung NVMe 15TB Gen 4
2 PEX9765 PCIe Gen 3 switches x16 to the CPU w/ 8 & 12 drives on each
100Gbps network interfaces

- Running cms-xcache containers[2]



System Block Diagram

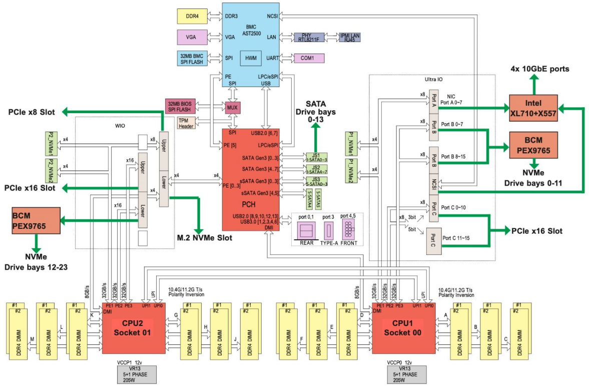


Figure 1-6. System Block Diagram, SYS-2029U-TN24R4T

[1] elbencho StorageSweep
https://github.com/breuner/elbencho/tree/master/contrib/storage_sweep
https://hub.docker.com/r/dtnaas/elbencho
[2] OpenScienceGrid CMS XCache
https://hub.docker.com/r/opensciencegrid/cms-xcache

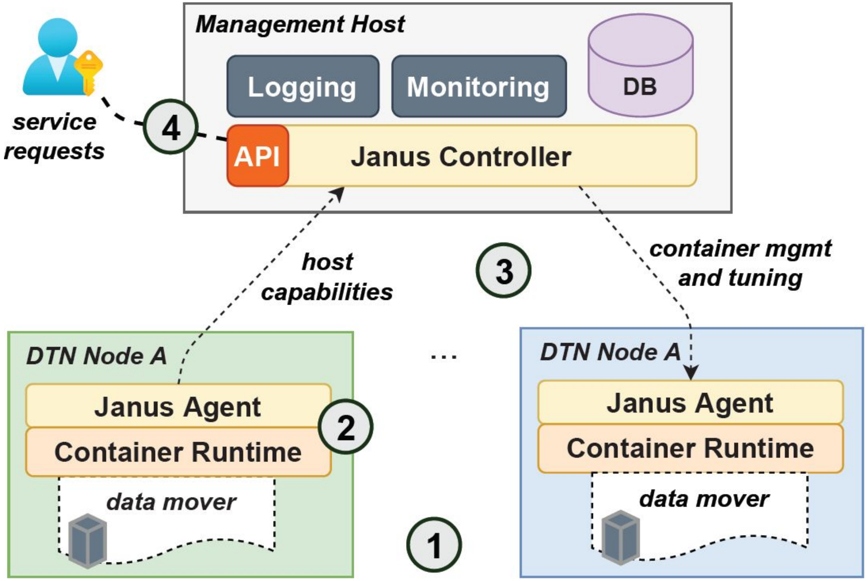
Janus container orchestration

- Develop a managed data movement service capability
  - Support a pool of transfer software images that “just work”
  - Reduce reliance on varying levels of network/system expertise for deployments
  - Enable automation on Data Transfer Nodes (DTNs): *DTN-as-a-Service*
- Make use of containerization supported by lightweight orchestration
- Target high-speed data transfer deployments with dual-stack and multi-homed networking requirements
- Evaluate container networking with data transfer tools used in R&E nets

ESnet

Janus concept

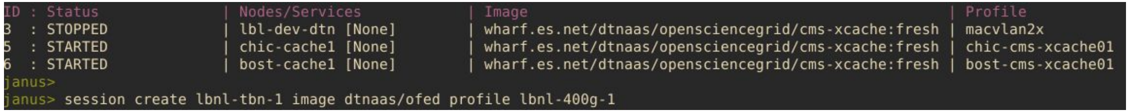
1. Data mover software in containers
2. Network and storage performance optimization
3. Configuration and tuning flexibility
4. Lightweight service orchestration



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

Extensible profiles

- Provide common configuration sets for service containers
- Helpful for consistency and re-use for larger deployments
- Specify capabilities once, then apply often



```
ID : Status              | Nodes/Services            | Image                                                  | Profile
3  : STOPPED             | lbl-dev-dtn [None]        | wharf.es.net/dtnaas/opensciencegrid/cms-xcache:fresh | macvlan2x
5  : STARTED             | chic-cache1 [None]        | wharf.es.net/dtnaas/opensciencegrid/cms-xcache:fresh | chic-cms-xcache01
6  : STARTED             | bost-cache1 [None]        | wharf.es.net/dtnaas/opensciencegrid/cms-xcache:fresh | bost-cms-xcache01
janus>
janus> session create lbnl-tbn-1 image dtnaas/ofed profile lbnl-400g-1
```

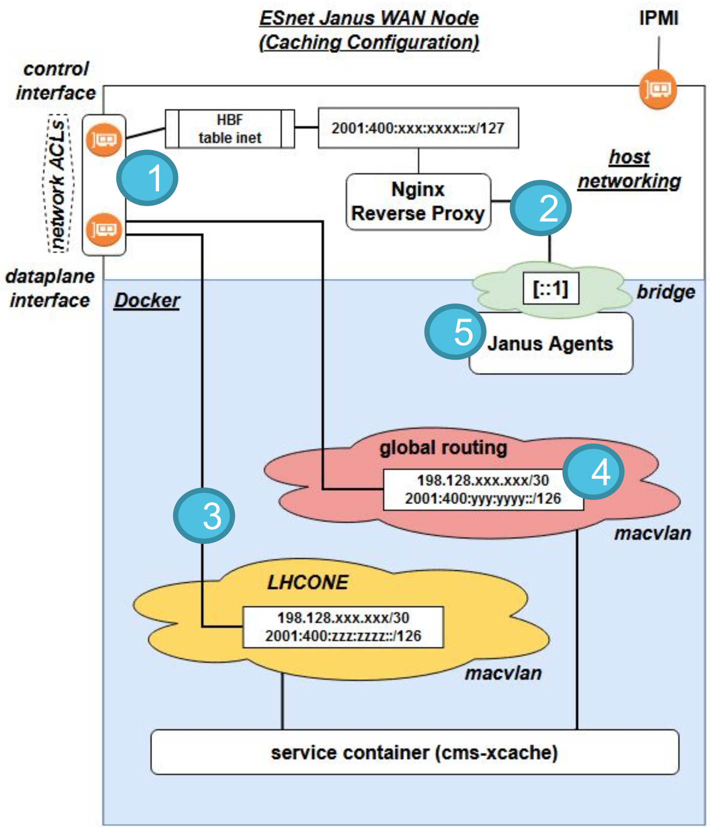
```
features:
  rdmacap:
    devices:
      - devprefix: "/dev/infiniband"
        names:
          - rdma_cm
          - uverbs
    caps:
      - IPC_LOCK
      - NET_ADMIN
    limits:
      - Name: memlock
        Soft: -1
        Hard: -1

profiles:
  lbnl-400g-1:
    cpu: 4
    affinity: network
    mgmt_net: bridge
    data_net:
      name: net3001_eth200
      ipv4_addr:
        - 10.33.1.20
      ipv6_addr:
        - 2001:400:2202:2191::3
    features:
      - rdmacap
    privileged: false

  lbnl-400g-2:
    cpu: 8
    affinity: network
    mgmt_net: bridge
    data_net:
      name: net3002_eth200
      ipv4_addr:
        - 10.33.2.20
        - 10.33.2.21
    features:
      - rdmacap
    privileged: false
    volumes:
      - data
```

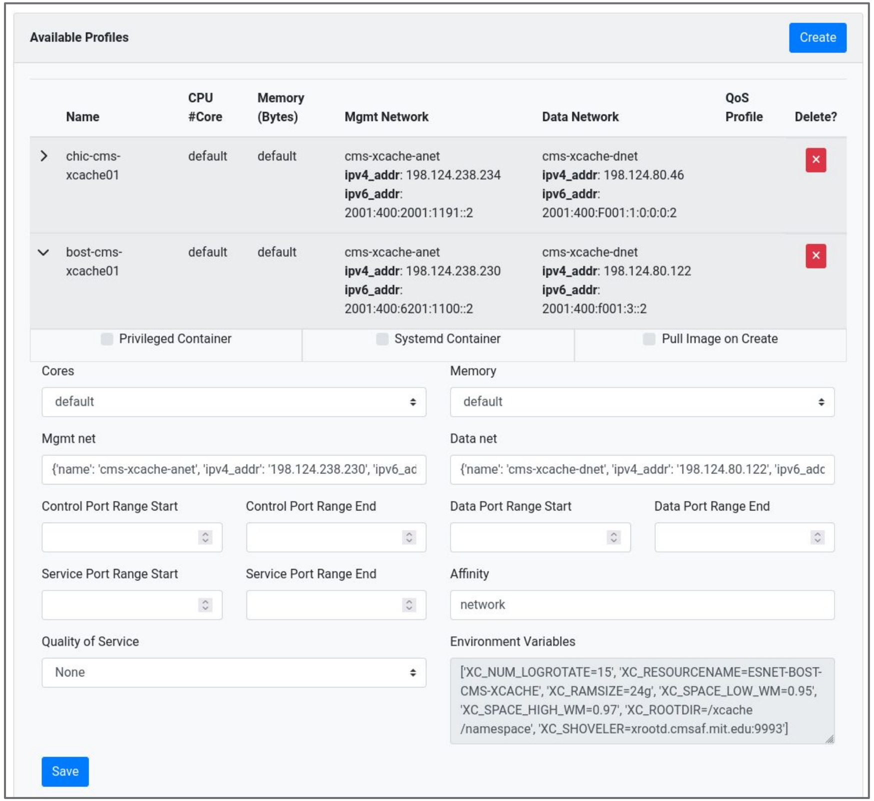
Caching node: network features

1. Multi-homed physical nodes
2. Slow path control
3. Fast path data plane
4. Dual-stack IP networking
5. Local agents for resource discovery and customized tuning



Deploying in-network caches in support of distributed scientific data sharing.
Alex Sim, Ezra Kissel, Chin Guok
https://arxiv.org/abs/2203.06843

Caching node: Janus WebUI



Available Profiles

	Name	CPU #Core	Memory (Bytes)	Mgmt Network	Data Network	QoS Profile	Delete?
›	chic-cms-xcache01	default	default	cms-xcache-anet **ipv4_addr**: 198.124.238.234 **ipv6_addr**: 2001:400:2001:1191::2	cms-xcache-dnet **ipv4_addr**: 198.124.80.46 **ipv6_addr**: 2001:400:F001:1:0:0:0:2		✕
⌄	bost-cms-xcache01	default	default	cms-xcache-anet **ipv4_addr**: 198.124.238.230 **ipv6_addr**: 2001:400:6201:1100::2	cms-xcache-dnet **ipv4_addr**: 198.124.80.122 **ipv6_addr**: 2001:400:f001:3::2		✕

□ Privileged Container □ Systemd Container □ Pull Image on Create

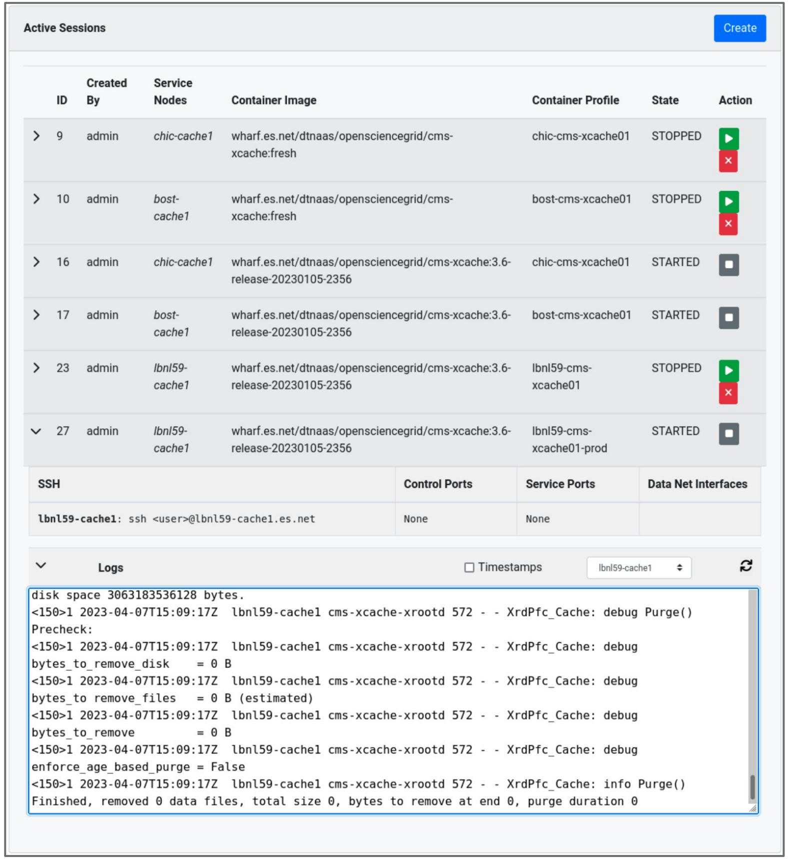
Cores: default	Memory: default
Mgmt net: {'name': 'cms-xcache-anet', 'ipv4_addr': '198.124.238.230', 'ipv6_ad	Data net: {'name': 'cms-xcache-dnet', 'ipv4_addr': '198.124.80.122', 'ipv6_adc

Control Port Range Start | Control Port Range End | Data Port Range Start | Data Port Range End

Service Port Range Start | Service Port Range End | Affinity: network

Quality of Service: None

Environment Variables: ['XC_NUM_LOGROTATE=15', 'XC_RESOURCENAME=ESNET-BOST-CMS-XCACHE', 'XC_RAMSIZE=24g', 'XC_SPACE_LOW_WM=0.95', 'XC_SPACE_HIGH_WM=0.97', 'XC_ROOTDIR=/xcache /namespace', 'XC_SHOVELER=xrootd.cmsaf.mit.edu:9993']



Active Sessions

	ID	Created By	Service Nodes	Container Image	Container Profile	State	Action
›	9	admin	*chic-cache1*	wharf.es.net/dtnaas/opensciencegrid/cms-xcache:fresh	chic-cms-xcache01	STOPPED	▶ ✕
›	10	admin	*bost-cache1*	wharf.es.net/dtnaas/opensciencegrid/cms-xcache:fresh	bost-cms-xcache01	STOPPED	▶ ✕
›	16	admin	*chic-cache1*	wharf.es.net/dtnaas/opensciencegrid/cms-xcache:3.6-release-20230105-2356	chic-cms-xcache01	STARTED	■
›	17	admin	*bost-cache1*	wharf.es.net/dtnaas/opensciencegrid/cms-xcache:3.6-release-20230105-2356	bost-cms-xcache01	STARTED	■
›	23	admin	*lbnl59-cache1*	wharf.es.net/dtnaas/opensciencegrid/cms-xcache:3.6-release-20230105-2356	lbnl59-cms-xcache01	STOPPED	▶ ✕
⌄	27	admin	*lbnl59-cache1*	wharf.es.net/dtnaas/opensciencegrid/cms-xcache:3.6-release-20230105-2356	lbnl59-cms-xcache01-prod	STARTED	■

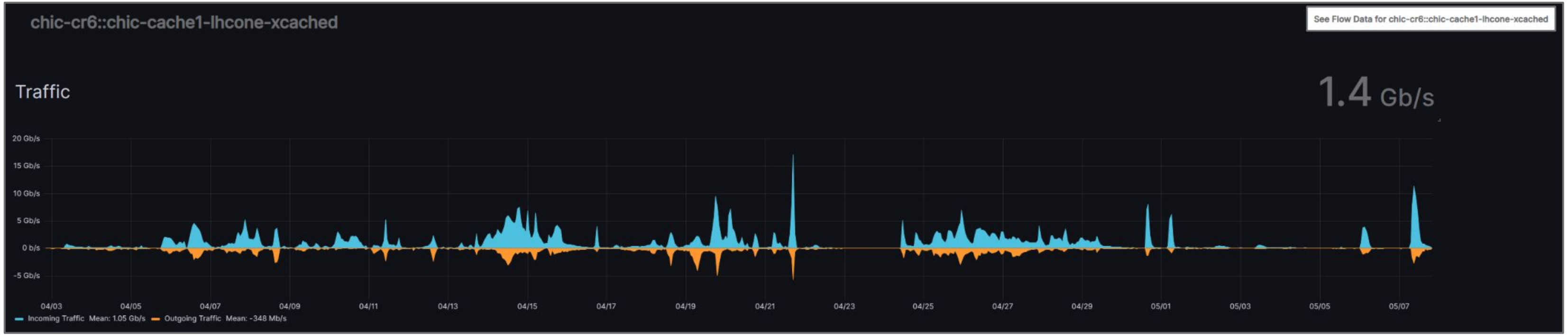
SSH	Control Ports	Service Ports	Data Net Interfaces
lbnl59-cache1: ssh <user>@lbnl59-cache1.es.net	None	None	

Logs □ Timestamps lbnl59-cache1

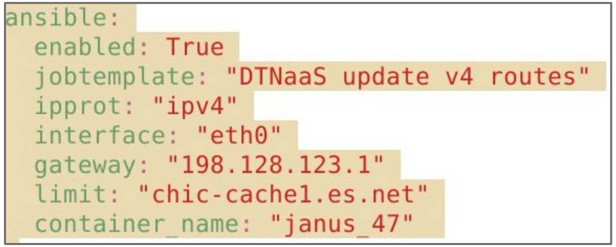
```
disk space 3063183536128 bytes.
<150>1 2023-04-07T15:09:17Z  lbnl59-cache1 cms-xcache-xrootd 572 - - XrdPfc_Cache: debug Purge()
Precheck:
<150>1 2023-04-07T15:09:17Z  lbnl59-cache1 cms-xcache-xrootd 572 - - XrdPfc_Cache: debug
bytes_to_remove_disk    = 0 B
<150>1 2023-04-07T15:09:17Z  lbnl59-cache1 cms-xcache-xrootd 572 - - XrdPfc_Cache: debug
bytes_to remove_files   = 0 B (estimated)
<150>1 2023-04-07T15:09:17Z  lbnl59-cache1 cms-xcache-xrootd 572 - - XrdPfc_Cache: debug
bytes_to_remove         = 0 B
<150>1 2023-04-07T15:09:17Z  lbnl59-cache1 cms-xcache-xrootd 572 - - XrdPfc_Cache: debug
enforce_age_based_purge = False
<150>1 2023-04-07T15:09:17Z  lbnl59-cache1 cms-xcache-xrootd 572 - - XrdPfc_Cache: info Purge()
Finished, removed 0 data files, total size 0, bytes to remove at end 0, purge duration 0
```

Supporting infrastructure

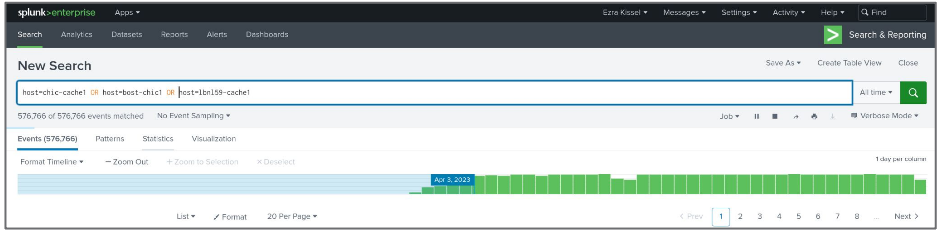
Stardust monitoring



Ansible integration



```
ansible:
  enabled: True
  jobtemplate: "DTNaaS update v4 routes"
  ipprot: "ipv4"
  interface: "eth0"
  gateway: "198.128.123.1"
  limit: "chic-cachel.es.net"
  container_name: "janus_47"
```



New Search

```
host=chic-cache1 OR host=bost-chic1 OR host=lbnl59-cache1
```

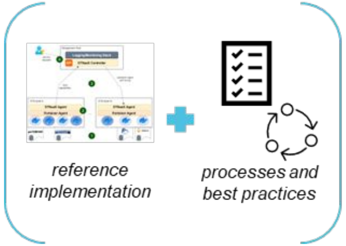
576,766 of 576,766 events matched

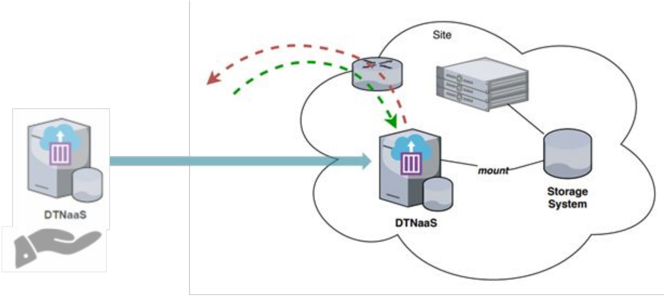
Log collection

ESnet

Observations and lessons learned

- Pilot efforts are challenging but rewarding
  - Effectively socializing new ideas is often half the battle
- Networking and application service concerns may be worlds apart
- The variety in OSG deployments and configurations can be daunting
  - Having a supportive technical contact is invaluable
- At the end of the day, the desire is for a solution that works





Summary

- Successful deployment of in-network caches on ESnet
  - Homegrown Janus/DTNaaS approach has shown promise for this application
  - Data caching pilots are expanding
  - Learning curve was overcome with help from OSG community and collaborators
- Characterization study including new nodes is ongoing
  - Existing SoCal Repo caching use has been effective (18.9TB cache hits per day)

- Future work:
  - London and Amsterdam nodes targeting LIGO and DUNE for TA traffic from US to EU
  - Enhancing XRootD Monitoring Shoveler placement and deployment
  - Multi-tenancy of shared physical caching nodes
  - Janus integration with additional container technologies and frameworks

Acknowledgements

Chris Cummings, Chin Guok, Damian Hazen, Ezra Kissel, Charles Shiflett, Goran Pejovic, Alex Sim, Fatema Bannat Wala

Diego Davila, Frank Würthwein