TRIUMF

Canada's national laboratory
for particle and nuclear physics
and accelerator-based science

TRIUMF status update, Storage experience

HSF WLCG Virtual Workshop, 20 November 2020

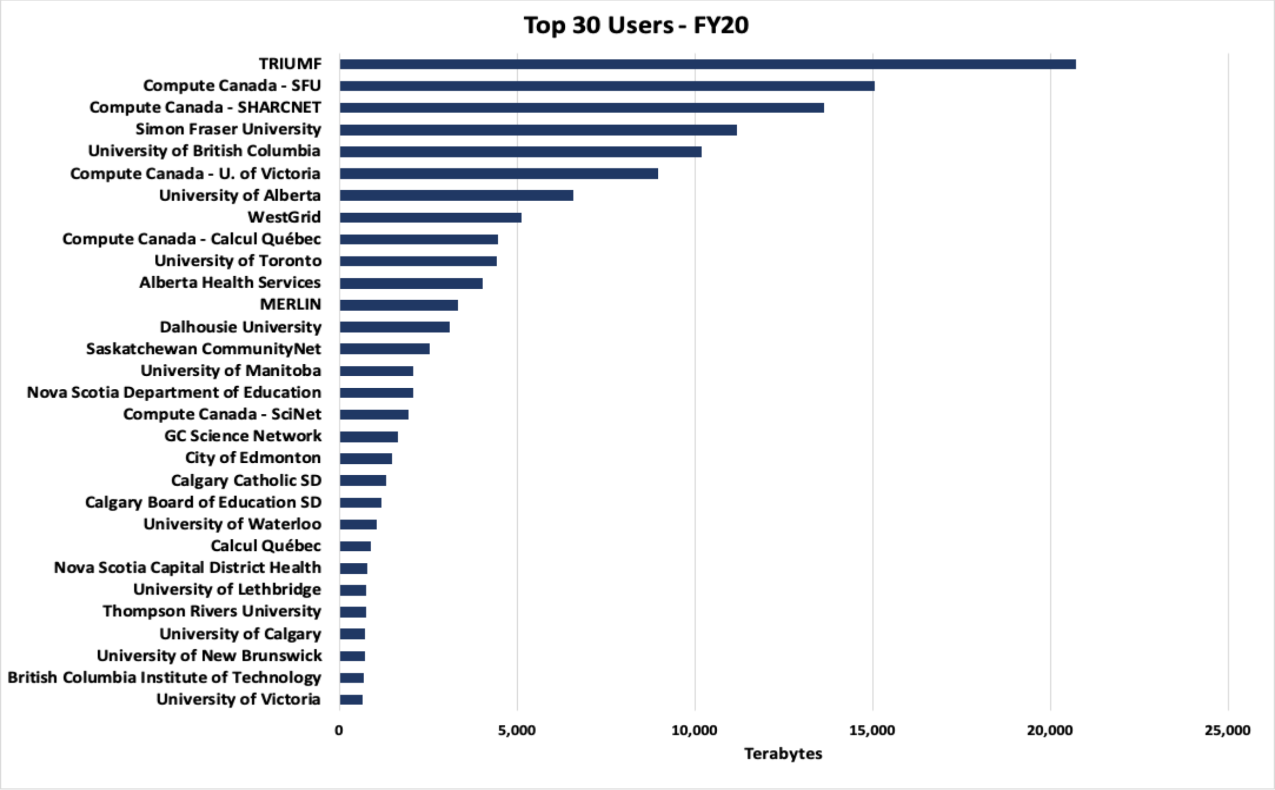
Xinli (Simon) Liu

compute | calcul
canada | canada

ATLAS T1 Status

Data and computing resources are located at Compute Canada SFU site

- Dedicated to ATLAS, also collaborating with CA T2s and Compute Canada
- 10% of ATLAS Data and Computing resources
- High Density and throughput compute cluster with 12k usable cores, 238.9 HEPSPEC06 computing resources
- Disk Storage 11 PB usable, 80GB/sec throughput
- Tape Storage 30 PB usable
- Ansible with a Git back-end as our configuration management
- Top Data transfer volume user over research network in CA



Storage Systems

DISK, current usable capacity 11PB

2 X DDN sfa14kx, and few IBM dcs3700 storage systems

20 Lenovo servers, Each 16 cores,196GB mem, 40GbE

dCache 5.2.35

Provides SRM, gridftp, https, xroot protocols. nfs4.1, dcap also supported, not being used.

https and xroot TPC have been tested by smoke test, and function tests. Production soon

TAPE current usable capacity 30PB

One library, 20 LTO8 drives, 12 LTO7 drives

HSM

one hsmhead node,10 hsm pool nodes, Disk buffer 820 TB

Tape system, Tapeguy/smallhsm, using ENDIT HSM interface plugin, provides high throughput for ATLAS computing

DISKLESS (Xcache)

Four xcache server cluster, 500 TB GlusterFS. Together with ARC6, created pre-caching feature.

Will move on to use ARC6 caching and index services, GlusterFS as cache storage.

DISKLESS Setup

3 * IBM dcs3700 storage systems, 4 servers, all out of warranty

GlusterFS as backend cache storage(500 TB)

Initially we deployed Ceph cluster, later dropped if off, turned to a GlusterFS.

Xcache

4 xrootd instances, no CMSD, no redirector, a simple DNS-RR, 4 xrootd services use the same GlusterFS cache backend.

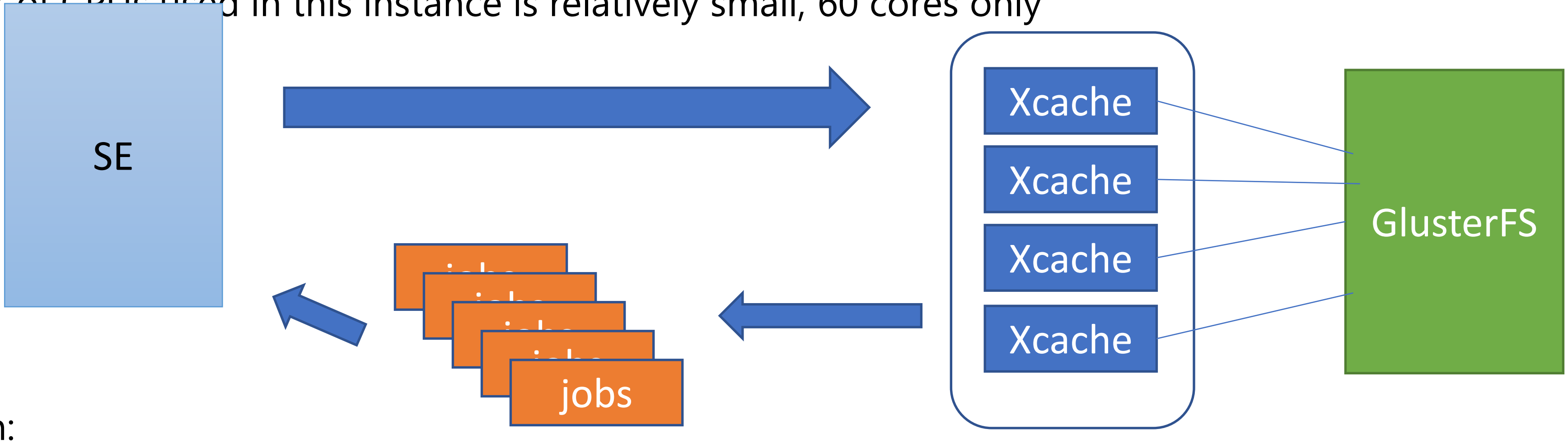
Pre-cache script runs on each xcache node, fetch queued jobs required data

Initially planned to test index service, didn’t make it

We also have a script to convert xcache log into billing db, so we can analysis data access

CPUs

Number of CPUs used in this instance is relatively small, 60 cores only



Conclusion:

This model works of course, it cached 367k files(250TB) since March, half of them are cached files

Not all functions are integrated, putting things together and make it run reliably takes some effort.

Will, evaluate ARC6 + GlusterFS cache model in coming weeks.

HSM and Tape system at TRIUMF

TAPE current usable capacity 30PB

One TS4500 library, 20 LTO8 drives, 12 LTO7 drives

Mixed medias, 3319 lto6 ,1000 lto7 ,1400 lto8 tape cartridges

Disk buffer 820 TB, one hsmhead node,10 HSM pool nodes

Adopted ENDIT HSM interface plugin, we did minor code change for our environment, as well as performance improvement.

Tape system, Tapeguy/smallhsm, fully developed at TRIUMF for ATLAS.

Features:

- Files grouping on write
- Reorder requests on read
- Minimize ape mounts and Maximize reads per mounts
- Open format
- No extra server/disk buffer for tape operation

Study - How files distribution on tape impact reading

Test data

125TB, 69k files from production

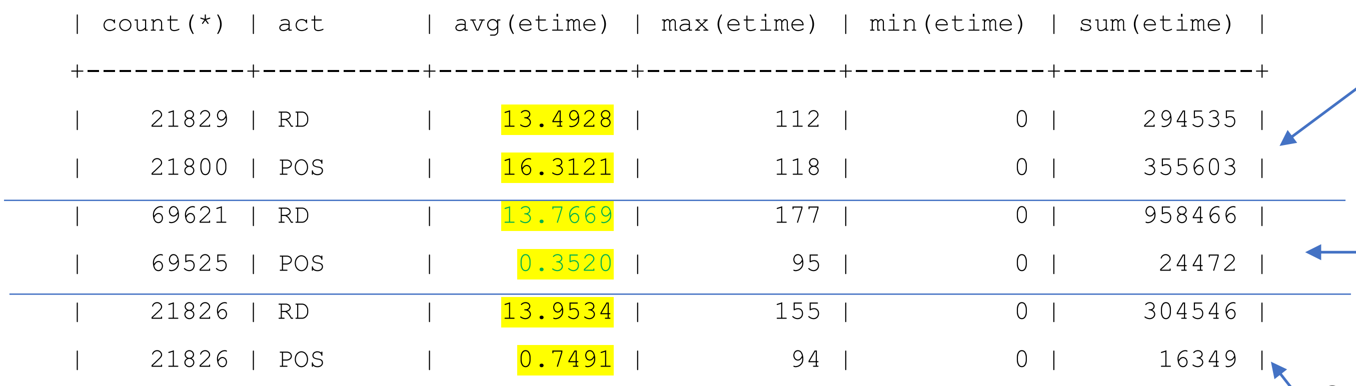
Min fsize, 451MB, avg fsize 1.8GB, max fsize 13GB

Files distributed on 99 LTO5 tapes, 1 to 3195 files on each tape. 81 full tapes

Full bulk staging, 0 repeat mounts

Staging requests submitted in one big bulk, files had optimized read

Random, Grouped, strip grouped staging read/position time cost comparison



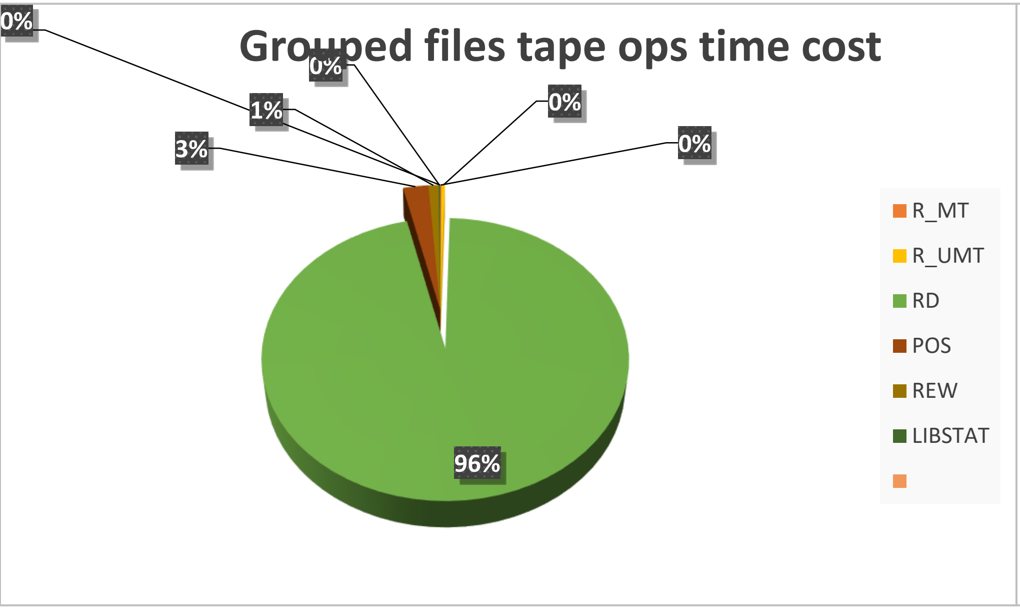
```
| count(*) | act      | avg(etime) | max(etime) | min(etime) | sum(etime) |
+----------+----------+------------+------------+------------+------------+
|    21829 | RD       |    13.4928 |        112 |          0 |     294535 |
|    21800 | POS      |    16.3121 |        118 |          0 |     355603 |
|    69621 | RD       |    13.7669 |        177 |          0 |     958466 |
|    69525 | POS      |     0.3520 |         95 |          0 |      24472 |
|    21826 | RD       |    13.9534 |        155 |          0 |     304546 |
|    21826 | POS      |     0.7491 |         94 |          0 |      16349 |
```

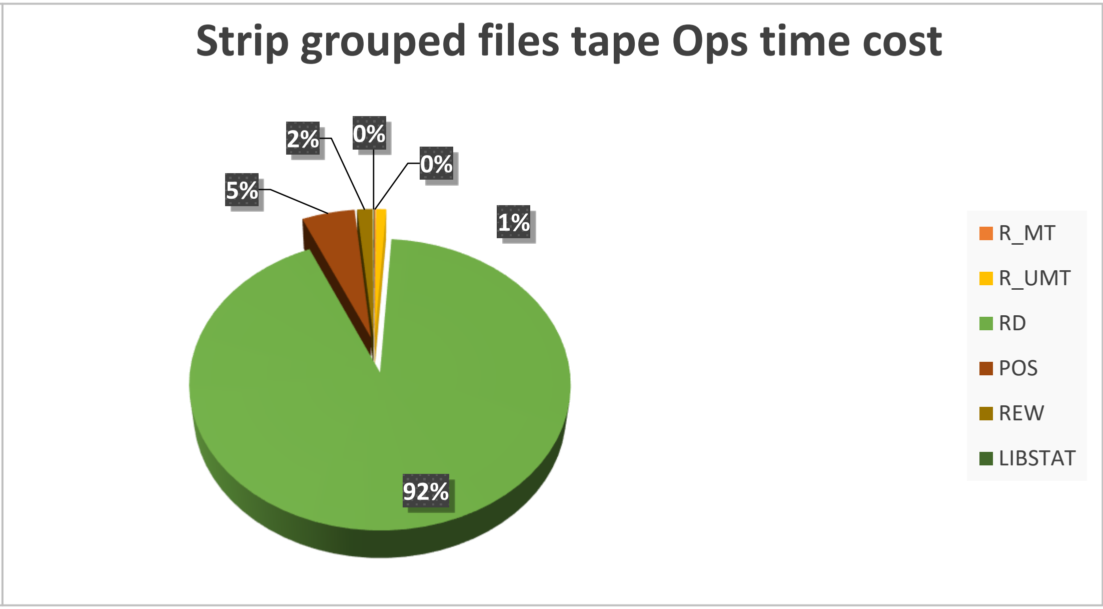
Random data read

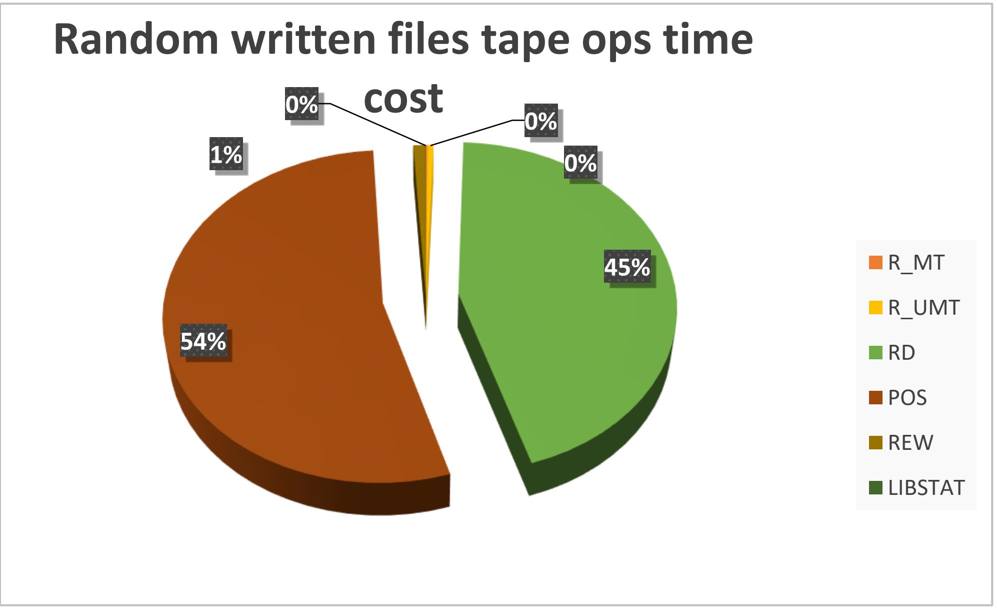
grouped data read

Small/strip group data read

Study - How files distribution on tape impact reading


Grouped files tape ops time cost


Strip grouped files tape Ops time cost


Random written files tape ops time cost

By grouping data, either big group, or small group, tape head seek time reduced to almost 1-2%, from 54%

Tape read rate doubles

Tape drive use efficiency increase by near 100%

Tape mounts, rewinding are time cost operations, however overall they don’t take big fraction of time

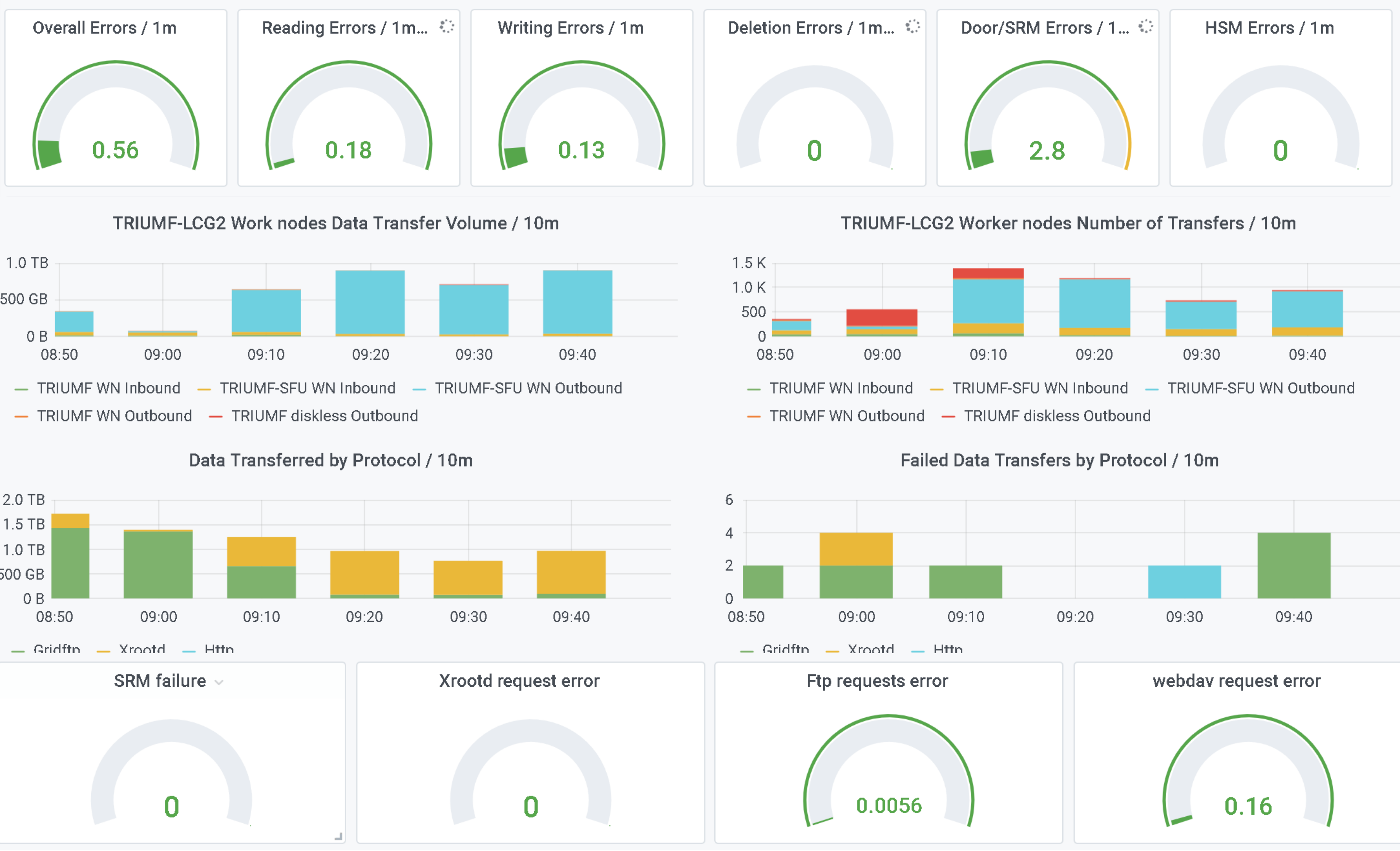
Data Analytic for storage

USE Elasticsearch, logstash, kibana, and Grafana as the platform

Injected some log data into ES

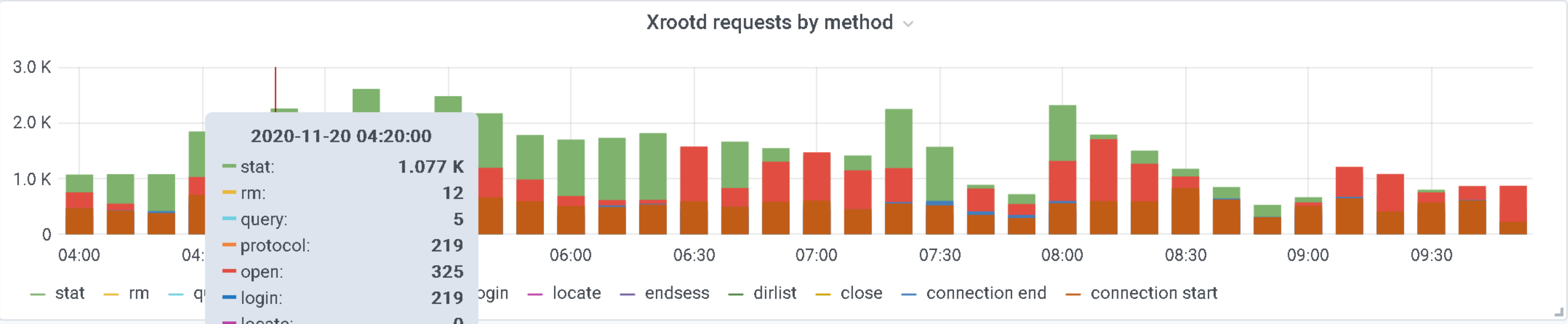
- Billing data
- Srm,ftp,http,xroot door access log
- Use packetbeat to track particular ports activities , 8443,1094,2880
- Also, we have other data poured into ES, router logs, database status metrics, security etc.. More to do
- Not meant to repeat DDM dashboard, but local monitoring and log analysis.
- Still at early stage

Data Analytic for storage



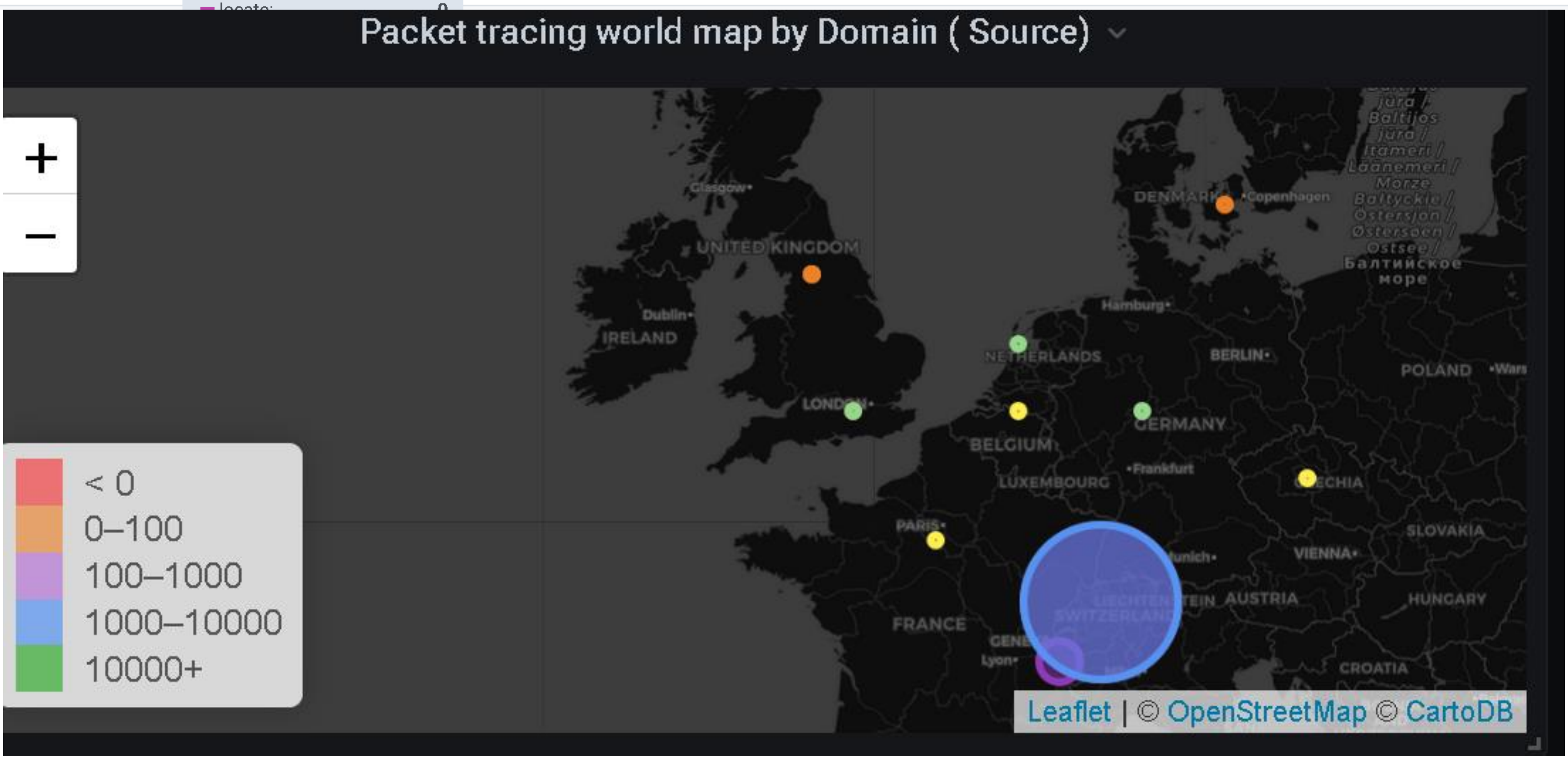
TRIUMF

Data Analytic for storage



Xrootd requests by method

2020-11-20 04:20:00
- stat: 1.077 K
- rm: 12
- query: 5
- protocol: 219
- open: 325
- login: 219
- locate: 0



Packet tracing world map by Domain (Source)

INNOVATION.CA

CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

THANK YOU !

ATLAS

canarie

BCNET

Shared Services for Higher Education & Research