400Gbps Benchmark of XRootD-HTTP third-party-copy Transfers

Aashay Arora, Jonathan Guiang, Diego Davila, Frank Würthwein, Justas Balcas, Harvey Newman
Introduction / Motivation

• The High-Luminosity LHC era will bring huge data challenges. We predict, ATLAS and CMS combined will accumulate on the order of an exabyte of raw data every year [1].

  • ESNet quotes (for T2 sites), “[Throughput] projections for the HL-LHC, with a planned start in 2027, are a 100 Gbps average over the year, with 400 Gbps bursts lasting hours” [1]

• Therefore, in addition to hardware upgrades, we need to verify the robustness of our software stack to make sure it can support this high throughput.

• HTTP third-party-copy (HTTP-TPC) is now the default for data transfers between LHC sites.

• In the US, all Tier 2 centers support XRootD for data access.

• What configuration of HTTP-TPC + XRootD give us the throughput we need? What are the minimum hardware requirements to run this configuration?
Previous Studies

• We studied the overhead of using XRootD-HTTPS over the globus protocol for data transfers
  • Transferred data over several 100Gbps links using the two protocols with varying RTTs between endpoints to test and compare the throughputs
  • Concluded that XRootD-HTTPS performs slightly better on average over high throughput links. [2]

• We benchmarked the performance on empty-links on low latency networks (Microsoft Azure)
  • We were able to get ~ 1 Tb/s within the same region
  • Want to test over higher RTT
Current Hardware Setup

- We have 13 data-transfer-nodes (DTNs) at Caltech (Full Specs in backup)
- 1 BIG DTN at UCSD
  - 2 x AMD EPYC 7763 64-Core Processor (with SMT)
  - 2.0 Ti Memory
  - 3 x 200 G Links (ConnectX-6)
- All hosts managed using Kubernetes
- Running 3 pods on UCSD host (each running its own XRootD service and separate interface), and 1 pod on each Caltech Host
- Dedicated network paths provisioned using SENSE-Autogole
XRootD Configuration

- XRootD configured in a cluster using a non-shared filesystem (each origin has its own set of files)
  - Multiple data origins subscribed to a single redirector.
  - Both origins and redirectors configured with the HTTP(S) directive.
  - Authentication using X509
  - Macaroons for delegation and authorization.
Results

• We can reach 400* Gbps and sustain it for hours!

• Well, 345 Gbps over a network path capable of doing 350 Gbps.

• Using 40 streams of 1 GB files for each of the 13 servers with Caltech as sink, i.e. 520 streams coming out of UCSD
Interesting Findings

1. How does number of streams affect the throughput?

   • Not surprising, Drastically! Over the same 200 Gbps capable links

   1 100GB file in-flight gives us 5 Gbps

   200 1GB files in-flight gives us 200 Gbps
In fact, how is throughput parametrized by number of cpu cores and streams?
2. What is the overhead of adding a redirector?

• Almost None! We get the same performance transferring data between clusters as we do transferring directly between origins.

cluster-to-cluster (using redirectors): 500 streams

origin-to-origin (manual load balancing): 40 streams per caltech host
3. How does the transfer tool affect throughput?

- We get slightly different performance when we transfer gfal-copy and curl (write-then-delete-then-repeat)

Using gfal-copy we get **200 Gbps**

Using curl, we get **170 Gbps**
Conclusion + Summary

• XRootD-HTTP is capable of supporting the high throughputs required for the HL-LHC era.
  • Systematically running transfers can enable us to parameterize by number of CPU cores, number of streams, etc.
  • Need at least $\mathcal{O}(10)$ streams per XRootD instance for ideal throughput.
  • Use of redirectors does not affect performance.
  • Choice of transfer tool does affect throughput.
Acknowledgement

The authors would like to thank the different funding agencies for this work, in particular the National Science Foundation through the following grants: OAC-1836650, OAC-2030508, OAC-1836650, MPS-1148698 and OAC-1541349
References


Thank You!
Backup
• We see interesting behavior when overwriting files,

With gfal-copy, we get 170 Gbps initially, then it drops down to 80 Gbps

Similar behavior with curl,
<table>
<thead>
<tr>
<th>Server</th>
<th>CPU</th>
<th>Cores</th>
<th>Mem</th>
<th>NIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>sandie-1</td>
<td>2x E5-2667 v3 @ 3.20GHz</td>
<td>32 (HT ON)</td>
<td>16 x Kingston DDR4 16GB 2400</td>
<td>MT27700 ConnectX-4 (100G)</td>
</tr>
<tr>
<td>sandie-3</td>
<td>2x Silver 4110 CPU @ 2.10GHz</td>
<td>32 (HT ON)</td>
<td>12 x Micron DDR4 8GB 2666</td>
<td>MT27500 ConnectX-3 (40G)</td>
</tr>
<tr>
<td>sandie-5</td>
<td>1x AMD 7551P @ 2Ghz</td>
<td>64 (SMT ON)</td>
<td>8 x Kingston DDR4 32GB 2666</td>
<td>MT28800 ConnectX-5 (100G)</td>
</tr>
<tr>
<td>sandie-6</td>
<td>1x AMD 7551P @ 2Ghz</td>
<td>64 (SMT ON)</td>
<td>8 x Kingston DDR4 32GB 2666</td>
<td>MT28800 ConnectX-5 (100G)</td>
</tr>
<tr>
<td>sdn-dtn-1-7</td>
<td>2x E5-2687W v3 @ 3.10GHz</td>
<td>20 (HT OFF)</td>
<td>8 x Micron DDR4 16GB 2133</td>
<td>MT27700 ConnectX-4 (100G)</td>
</tr>
<tr>
<td>sdn-dtn-2-09</td>
<td>2x E5-2690 v2 @ 3.00GHz</td>
<td>40 (HT ON)</td>
<td>16 x Samsung DDR3 8GB 1600</td>
<td>MT27500 ConnectX-3 (40G)</td>
</tr>
<tr>
<td>sdn-dtn-2-11</td>
<td>2x E5-2670 v3 @ 2.30GHz</td>
<td>48 (HT ON)</td>
<td>16 x Micron DDR4 8GB 2133</td>
<td>MT28800 ConnectX-5 (100G)</td>
</tr>
<tr>
<td>neu-sc-01</td>
<td>2x E5-2667 v4 @ 3.20GHz</td>
<td>32 (HT ON)</td>
<td>8 x Hynix DDR4 16GB 2133</td>
<td>MT28908 ConnectX-6 (100G)</td>
</tr>
<tr>
<td>sdn-sc-03</td>
<td>2x E5-2667 v4 @ 3.20GHz</td>
<td>32 (HT ON)</td>
<td>8 x Hynix DDR4 16GB 2133</td>
<td>MT28908 ConnectX-6 (100G)</td>
</tr>
<tr>
<td>sdn-sc-04</td>
<td>2x E5-2667 v4 @ 3.20GHz</td>
<td>32 (HT ON)</td>
<td>8 x Hynix DDR4 16GB 2133</td>
<td>MT28908 ConnectX-6 (100G)</td>
</tr>
<tr>
<td>sdn-sc-05</td>
<td>2x E5-2667 v4 @ 3.20GHz</td>
<td>32 (HT ON)</td>
<td>8 x Hynix DDR4 16GB 2133</td>
<td>MT28908 ConnectX-6 (100G)</td>
</tr>
<tr>
<td>sdn-sc-06</td>
<td>2x E5-2667 v4 @ 3.20GHz</td>
<td>32 (HT ON)</td>
<td>8 x Hynix DDR4 16GB 2133</td>
<td>MT28908 ConnectX-6 (100G)</td>
</tr>
<tr>
<td>sandie-9</td>
<td>2x E5-2667 v3 @ 3.20GHz</td>
<td>32 (HT ON)</td>
<td>8 x Hynix DDR4 16GB 2133</td>
<td>MT28800 ConnectX-5 (100G)</td>
</tr>
</tbody>
</table>
Plot Showing 300 Gbps sustained
Plot showing throughput vs. latency