End-to-End Data Movement Using MPI-IO Over Routed Terabits Infrastructures

Geoffroy Vallee, Scott Atchley, Youngjae Kim, Galen Shipman
Motivation

• Platform: HPC leadership computing centers linked via a dedicated WAN infrastructure

• Environment: Users run MPI applications on each site and want to exchange a large amount of data

• Specifics
  – DOE HPC platforms
  – DOE ESNet infrastructure

Can we transfer data using the platform and environment that are already in place?
Challenges

1. End-to-end communication over heterogeneous networks across both local-area and wide-area networks
   - From HPC system to HPC system via the WAN
   - Need a high-performance communication substrate for the WAN

   Common Communication Infrastructure Routing

2. Is it possible to assume that users can use MPI for data transfer?
   - Ease the integration with applications

Experiments using CCI with MPI on ESNet
Common Communication Infrastructure Routing

• Assumptions
  – Based on the Common Communication Infrastructure (CCI)
    • Support various technologies: verb, gni, ethernet (Linux kernel), TCP, UDP
    • Support reliable/unreliable and ordered/unordered connections
    • Support Remote Memory Access (RMA)
  – Heterogeneous networks
    1. First facility’s high-performance interconnect within the leader class system
    2. First facility’s local network
    3. Wide area network
    4. Second facility’s local network
    5. Second facility’s high-performance interconnect within the HPC system
  – Take advantage of the highest performing networking stack on each network
Example

AS: Autonomous system
SN: Single subnet
R: Router
Route Map

• Clients never have the route map
• Clients only have a static list of reachable routers
• Each router within AS has the same route map
  – Multiple route selection
    • Convert to Gb/s and divide by the larger target bandwidth
    • Lower value selected
    • Ex: 10 Gb/s link within a 1Tb/s network gives 100; a 100Gb/s gives 10
  – Routing table is typically NxN where N is the number of subnets
    • Left column is array of originators
    • Top row is array of destination subnets
    • Intersection row N and column M is the ordered list of subnets IDs from subnet N to subnet M
Example

<table>
<thead>
<tr>
<th></th>
<th>SN1</th>
<th>SN2</th>
<th>SN3</th>
<th>W*</th>
<th>W4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN1</td>
<td>-</td>
<td>1,2</td>
<td>1,3</td>
<td>1,W*</td>
<td>1,W4</td>
</tr>
<tr>
<td>SN2</td>
<td>2,1</td>
<td>-</td>
<td>2,3</td>
<td>2,1,W*</td>
<td>2,1,W4</td>
</tr>
<tr>
<td>SN3</td>
<td>3,1</td>
<td>3,2</td>
<td>-</td>
<td>3,1,W*</td>
<td>3,1,W4</td>
</tr>
<tr>
<td>W*</td>
<td>W*,1</td>
<td>W*,1,2</td>
<td>W*,1,3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>W4</td>
<td>W4,1</td>
<td>W4,1,2</td>
<td>W4,1,3</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
CCI Experiment with MPI on ESNet

• Preliminary results using TCP using nodes on WAN’s edge
  – Between ANL and NESCR
  – Using a 10Gb/s NIC

• Using the CCI BTL (developed by UTK)

• Read file from Site A; transfer the data to Site B; write the data into a file on Site B

• Throughput calculated with
  – \( t_0 \) = start to read the file on Site A
  – \( t_{\text{final}} \) = site A receives the acknowledgment the file is received and written to a file on Site B
Description of the Experiment

- Rank i: MPI_Send/MPI_Recv → Rank n
- Rank i+1: MPI_Send/MPI_Recv → Rank n+1
- Rank i+3: MPI_Send/MPI_Recv → Rank n+2
- Rank i+4: MPI_Send/MPI_Recv → Rank n+3
- Rank j: MPI_Send/MPI_Recv → Rank m

Operations:
- Site A: MPI or POSIX_File_read operations
- Site B: MPI or POSIX_File_write operations

Sites:
- Site A
- Site B
CCI TCP Performance

- RMA write operation
Results

Throughput (MB/s) vs. Number of Ranks

- CCI + POSIX
- CCI + MPI-IO
- TCP + POSIX
- TCP + MPI-IO
Results (2)

• Overlap file access and communications
• Split the file among the ranks (“chunks”); each rank read the assigned chunk by block of pre-defined size

• All ranks on Site A
  – Read a block and do a non-blocking send

• All ranks on Site B
  – Post all the non-blocking receives
  – When a receive completes, write to file using the POSIX API

<table>
<thead>
<tr>
<th>Block Size</th>
<th>1 GB file (throughput in MB/s)</th>
<th>10 GB file (throughput in MB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1MB</td>
<td>488.17</td>
<td>1012.42</td>
</tr>
<tr>
<td>10MB</td>
<td>481.71</td>
<td>963.52</td>
</tr>
</tbody>
</table>
Conclusion

• CCI+MPI is a solution for end-to-end data movement over routed terabit infrastructure

• MPI can be used to implement data transfer over WAN
  – MPI-IO implementation provides poor performance
  – File access with the POSIX API + overlap with MPI communications enable more than 1000MB/s

• Future work
  – Develop of RoCE CCI transport (or any future technology for the WAN)
  – Ongoing effort focusing on I/O optimization with layout-awareness on end-system hosts or bulk data movement
Acknowledgment

• This research is sponsored by the Office of Advanced Scientific Computing Research; U.S. Department of Energy and performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 and used resources of the Center for Computational Sciences at Oak Ridge National Laboratory.

• This research used resources of the ESnet Testbed, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH11231.