Network-Aware Data Movement Advisor

Patrick Brown and Mengxia (Michelle) Zhu
Department of Computer Science
Southern Illinois University
Carbondale, IL 62901, USA
{patiek, mzhu}@cs.siu.edu

Qishi Wu and Xukang Lu
Department of Computer Science
The University of Memphis
Memphis, TN 38152, USA
{qishiwu, xlv}@memphis.edu

November 14th, 2011

The Network-Aware Data Management (NDM) Workshop
in conjunction with
The Supercomputing Conference (SC'11)
Washington State Convention Center, Seattle, WA

Sponsored by
Advanced Scientific Computing Research
Office of Science
Department of Energy
Outline

- Introduction
- NADMA
  - Architecture
  - Functional Components
- System Implementation
- Case Studies
  - Regular Data Transfer in Internet2
  - ION-based Data Transfer over Dedicated Path in Internet2
  - OSCARS-based Data Transfer over Dedicated Path in ESnet
  - Cross-domain Data Transfer over Dedicated Path in Internet2 and ESnet
- Conclusion
Introduction

- Large-scale scientific applications

- Climate research
- Nanoscience
- Astronomy
- Neutron sciences
- Computational biology
- Fusion simulation
- High energy physics
- Computational materials
• **Massive amount of scientific data**
  – Simulation
    • Fusion, combustion, astrophysics, etc.
  – Experimental
    • Spallation Neutron Source, Large Hadron Collider, etc.
  – Observational
    • Sensor networks, astronomy imaging, etc.

• **Data needs to be moved and managed**
  – Remote storage
  – Collaborative analysis
  – Distributed processing

• **Resources needed for data transfer**
  – Advanced Networking services
  – Storage and I/O services
Current status of resource provisioning and data transfer

- Several networking and storage research projects
  - Underway to support bulk data transfer, such as ANI, BeStMan, TeraPaths, ESCPS, etc.
  - Have a very limited user scope
- Application users are primarily domain experts
  - Manually configure and execute their routine data-centric tasks
  - Use software tools based on their own empirical studies

Challenges of utilizing existing services

- Their deployment requires a certain level of network/host reconfigurations
- Most science users are even not aware of their existence inside their own networks
- Involve an often steep learning curve towards their adoption
- Result in unsatisfactory performance

Discovering data and available networking and storage technologies with estimated performance is a critical step towards their wide adoption!
What can be done to assist in data transfer?

- Generally, a default Internet path is used for data transfer
- However, dedicated paths might be also available
  - Provisioned by hardware/software (such as OSCARS) on shared connections
- Transport protocols themselves lack abilities
  - To discover these dedicated paths
  - To choose efficient routes for data transmission
  - To predict the performance in terms of measured bandwidth and delay

Our approach – Network-aware Data Movement Advisor (NADMA)

- Acts as a route planner in a typical navigation system
- Aims to help users select the best possible route to transfer data among all possible routes
- Supports automatic resource discovery and performance estimation

Goal: make NADMA a user-friendly and lightweight toolkit that has great potential to impact the science community by augmenting the traditional way of data transfer
User Interface

Network-Aware Data Movement Advisor

- Source Endpoint: 131.230.142.82
- Destination Endpoint: alpha.ep-dev.com

Transfer Preferences
- Enable Transfer Preferences
- Transfer Data Size: 120 GB
- Maximum Duration: 1 Hour
- Earliest Start Time: 8/30/11 10:00 AM
- Latest Finish Time: 8/30/11 11:00 PM

Protocol Preferences

Usage Instructions:
Network-Aware Data Movement Advisor is a tool that can assist in data movement. It provides instructional recommendations for improved data movement and resource utilization.

To begin using this tool, please provide a source and destination endpoint. If you do not have a source or destination endpoint, you can use the Find Dataset button to search for available datasets.

Transfer and Protocol preferences can be specified if desired. The preferences are used when analyzing the network and providing data movement advice.

After providing the information above, you can click the Begin Advise button below to begin the advice process.

Probing Endpoint Networks
1. Resolving Endpoint Hostnames
2. Searching for Reserved Network Paths
3. Searching for Common Network Protocols
4. Calculating Available Network Paths
5. Building Network Analysis
Transfer Protocol Discovery

- NADMA scans for open ports of popular transfer protocols such as GSIFTP, HTTP, HTTPS, BBFTP, SCP, SFTP, and FTP
- Users can choose protocols of interest to scan for
- An open protocol port detected by the port scanner is considered available only if it is supported in the transfer scenario
  - Local host to remote host
  - Remote host to local host
  - Remote host to remote host
Data Discovery

- NADMA allows users to specify dataset characteristics
  - Users can also discover datasets of interest using the built-in data discovery component
- The data discovery component interacts with the remote metadata services and replica location services (RLS) to identify the physical locations of these data items
- The dataset properties including source location and dataset size are automatically retrieved
  - Earth System Grid (ESG) data discovery web service is used as an example
### ESG Data Discovery

#### Data discovery interface for Earth System Grid

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Filename</th>
<th>Filesize</th>
<th>Parent Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>ucars.gd.pcm.B06.97.atm.proc.monthly_ave</td>
<td>BEVAP.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>48.50 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>BTRAN.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>48.50 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CGS.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>920.94 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CH4VMR.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>48.50 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CLDGHG.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>48.50 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CLDLOW.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>48.50 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CLDMED.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>48.50 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CLDTOT.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>48.50 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CLOUD.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>872.47 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CLOUDP.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>872.47 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CMFDQ.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>872.47 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CMFDT.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>872.47 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
<tr>
<td></td>
<td>CMFMG.B06.97.atm.1870-01_cat_1999-12.nc</td>
<td>920.94 MB</td>
<td>ucars.gd.pcm.B06.97...</td>
</tr>
</tbody>
</table>
Network Profile Management Using Web Services

- The network profile needs to be updated and maintained in a timely manner
  - Network topology, availability and locations of provisioning services, data management and movement protocols
- The database is accessible by
  - Regular users for information lookup
  - System administrators for information updates via web service calls
- The user interface interacts with the remote centralized database to retrieve network profile
  - Add, remove and edit entries
Network Profile Management Using Web Services Interface
Network Quality of Service Discovery

- NADMA tells users if there exists any dedicated path between two end hosts
- NADMA discovers existing SRM and GridFTP endpoints in the corresponding domains of interests
- NADMA displays the intermediate routers along the path for both Internet-based and dedicated paths in Google map
- NADMA dynamically measures the bandwidth and delay between two end hosts using Bing or available transfer protocols such as FTP and SCP
Bandwidth and Delay Estimation and Measurement

A linear regression line to estimate the path bandwidth and latency between a local SIUC node and a remote node in Connecticut: the x-axis represents the file sizes in KB, and the y-axis represents the round-trip time in seconds.
Implementation Details

- **NADMA Frontend**
  - Java client with native library socket support
- **NADMA Backend**
  - Java Hessian web service built on Apache Tomcat
  - MySQL database
    - Advanced network resource information
    - IP Geolocation
- **Transfer Strategy Generation**
  - Endpoint Specification
  - Network Discovery
  - Route Generation and Performance Estimation
Case Studies

1. Regular Data Transfer in Internet2
   • siu.edu to nyu.edu
2. ION-based Data Transfer in Internet2
   • stanford.edu to nyu.edu
3. OSCARS-based Data Transfer in ESnet
   • lbl.gov to bnl.gov
4. Cross-domain Data Transfer in Internet2 and ESnet
   • lbl.gov to nyu.edu
Case Study 1: Regular Data Transfer

While your source and destination endpoints appear to have access to an advanced network, we could not immediately determine a dedicated reserved path from the source endpoint to the destination endpoint. However, we found a non-dedicated path using an advanced network.
Case Study 2: ION-based Data Transfer
Case Study 3: OSCARS-based Data Transfer

Network-Aware Data Movement Advisor

Source
IP Address: 128.3.30.109
Hostname: sim.lbl.gov
Location: Berkeley, CA
Organization: LBNL

Destination
IP Address: 130.199.185.72
Hostname: grid15.racf.bnl.gov
Location: Upton, NY
Organization: BNL

Reserved Path
Source (A): lbl.gov
Destination (B): bnl.gov
Network: OSCARS
Maximum Backbone Capacity: 10000 Mbps

Link Capacities:
A → B: 20000 Mbps (2*MAN 10G RING)
B → C: 10000 Mbps
C → D: 20000 Mbps
D → E: 20000 Mbps
E → F: 20000 Mbps
F → G: 20000 Mbps
G → H: 60000 Mbps

Path:
A: lbl.gov (LBNL) - LBNL
B: (SUNY) - SUNY
C: (DENV) - DENV
D: (KANS) - KANS
E: (STAR) - STAR
F: (CLEV) - CLEV
G: (AOFA) - AOFA

Advisement Summary
Your source and destination endpoints appear to have access to a reserved bandwidth network. We have found a possible path using that uses this reserved bandwidth to offer significantly faster transfer and network performance over standard internet channels.

Detectable Network Protocols

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Port</th>
<th>Bandwidth(kbps)</th>
<th>Latency(ms)</th>
<th>Test Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>GridFTP</td>
<td>2811</td>
<td>unknown</td>
<td>unknown</td>
<td>Test Bandwidth</td>
</tr>
</tbody>
</table>

<< Configure Endpoints

View Detailed Advisement >>

Illinois University
Carbondale
Case Study 4: Cross-domain Data Transfer

Network-Aware Data Movement Advisor

Source:
- IP Address: 128.3.41.146
- Hostname:.lbl.gov
- Location: Albany, CA
- Organization: LBNL

Destination:
- IP Address: 128.122.119.209
- Hostname: nyu.edu
- Location: New York, NY
- Organization: New York University

Reserved Path

Source (A): lbl.gov
Destination (G): nyu.edu
Network: OSCARS / Internet2 CION
Maximum Backbone Capacity: 1000 Mbps

Link Capacities:
- A → B: 20000 Mbps (2*MAN 10G RING)
- B → C: 10000 Mbps (SUNN / Internet2 CENC)
- C → D: 20000 Mbps (CENC)
- D → E: 10000 Mbps (I2-CHIC-KANS-10GE-05567)
- E → F: 10000 Mbps (NYSERNet)
- F → G: 1000 Mbps

Path:
- A: lbl.gov (LBNL) - LBNL
- B: SUNN - SUNN
- C: cenic.org - CENC
- D: rtr.kans.net.internetc2.edu
- E: rtr.chic.net.internet2.edu
- F: nysernet.org - NYSERNet, Inc.
- G: nyu.edu - New York University

Detectable Network Protocols

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Port</th>
<th>Bandwidth(kbps)</th>
<th>Latency(ms)</th>
</tr>
</thead>
</table>

Advisement Summary

Your source and destination endpoints appear to have access to a reserved bandwidth network. We have found a possible path using that uses this reserved bandwidth to offer significantly faster transfer and network performance over standard internet channels.

View Detailed Advisement >>
Select a real source and destination endpoint to receive a detailed step-by-step advisement and performance estimation. You can choose different combinations of source and destination endpoints to determine a transfer strategy that is well suited to your needs.

**Summary**

**Requirements and Notes**
- Local installation of Globus Toolkit or SSH access to a machine to use Globus Toolkit remotely.
- If you do not have access to Globus Toolkit, consider using Globus Online (https://www.globusonline.org/) to facilitate the transfer. Globus Online provides many functions of the Globus Toolkit through its website as well as GUI-enabled servers.
- Grid Credentials / Authorized access to both sim.lbl.gov and grid15.racf.bnl.gov. You can inquire at your institution about acquiring grid-enabled certificates.

**Path Reservation (Assisted)**
Your source and destination endpoints support path reservation over OSCARS. We can assist you in creating this reservation directly from this program. Click the Prepare Transfer button at the bottom of this window to begin.

**Path Reservation (Manual)**
Your source and destination endpoints support path reservation over OSCARS. Creating dedicated circuits varies depending on your institution but can usually be accomplished through a client tool or web interface. You should inquire about your ability to create dedicated circuits with your local administrator.

**Transfer Instructions (Manual)**
To begin the transfer, you must SSH into a machine that has Globus Toolkit installed or open a local terminal if your local machine has access to Globus Toolkit.

1. grid-proxy-init: This Globus Toolkit establishes a proxy certificate that can be used by applications to authenticate themselves on your behalf.
2. globus-url-copy -r gsiftp://sim.lbl.gov:2811 gsiftp://grid15.racf.bnl.gov:2811/path/to/destfolder/: Recursively copy files from sim.lbl.gov to grid15.racf.bnl.gov using GridFTP. The file or folder that you specify will be stored into the path that you specify.
Current & Future Works

- Integration of Globus Online API
  - Facilitate transfers using Globus Online from NADMA
- Intelligent interaction with OSCARS 0.6
  - Expose new multiple-plan functionality of OSCARS 0.6 (shortest duration versus earliest finish time)
- Improved Support for DICE InterDomain Controller Protocol
  - Dynamically query topology of networks supporting IDCP
  - Allow easy subscription to various IDCP networks
Thanks! ☺️

Questions?