NAMED DATA NETWORKING IN
SCIENTIFIC APPLICATIONS

Susmit Shannigrahi, Chengyu Fan and Christos Papadopoulos
Colorado State University

March 23, 2017
Work supported by NSF #1345236 and #13410999
3 Years of CMIP5 Data Access

- CMIP5 is a 3.3PB archive of climate data, made available to the community through ESGF (~25 nodes) (CMIP6 estimated into the exabytes)
- We look at one server log collected at the LLNL ESGF node
- Approximately 3 years of requests (2013 to 2016)
- 18.5 million total requests (many duplicate)
- 1.5M Unique datasets requested
  - Total size Requests (with dupes) = 1,844TB
Client Locations
ASN Map

- Done using reverse traceroute
- Little path overlap, but view from only one ESGF node
## User/Clients Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Users</td>
<td>5692</td>
</tr>
<tr>
<td>Unique Clients (IP addresses)</td>
<td>9266</td>
</tr>
<tr>
<td>Unique ASNs</td>
<td>911</td>
</tr>
</tbody>
</table>
User Distribution per ASN

![User Distribution per ASN Chart]

- **Colorado State University**

---

**Graph Description**

- The left chart illustrates the number of users distributed across different Autonomous System Numbers (ASNs) ranging from 0 to 910.
- The right chart shows the Cumulative Distribution Function (CDF) of users with respect to the number of ASNs, indicating how the number of users is distributed across a range of ASNs.
Dataset Size Distribution

95% percentile: 1.34GB
98% of the datasets was requested by 10 users or less.)
Successful vs Failed Requests

- Successful Requests
- Failed Requests

Number of Requests at Producer

Time (10 minute bins)
## Summary: Data Statistics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMIP5 Archive Size</td>
<td>3.3PB</td>
</tr>
<tr>
<td>Total Data Requested</td>
<td>Equivalent of 1.8PB (18.5M requests)</td>
</tr>
<tr>
<td>Total Data Successfully Retrieved</td>
<td>234 TB (1.9M requests)</td>
</tr>
<tr>
<td>Total Data Successfully Retrieved (Excluding Duplicates)</td>
<td>113 TB (415K requests)</td>
</tr>
<tr>
<td>Number of unique datasets requested</td>
<td>1.5 million</td>
</tr>
</tbody>
</table>
A Closer Look at Failures

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of requests</td>
<td>18.5 million</td>
</tr>
<tr>
<td>Successful Requests</td>
<td>1,935,256</td>
</tr>
<tr>
<td>Failed Requests</td>
<td>16,673,815</td>
</tr>
</tbody>
</table>
Client Request Failures

CDF of Failure

Number of Clients
Duplicate Requests by Failure Group

- Failure rate < 10%
- Failure rate 10-90%
- Failure rate > 90%
Failure Heatmap
CMIP5 Data Retrieval Today

- HTTP://someESGFnode:/CMIP5/output/MOHC/HadCM3/decadal1990/day/atmos/tas/r3i2p1/tas_Amon_HADCM3_historical_r1i1p1_185001-200512.nc
CMIP5 Retrieval with NDN

HTTP://someESGFnode./CMIP5/output/MOHC/HadCM3/decadal1990/day/atmos/tas/r3i2p1/tas_Amon_HADCM3_historical_r1i1p1_185001-200512.nc
Why make the change?

☐ Does it improve performance?
☐ Does it improve publishing?
☐ Does it improve discovery?
☐ Does it improve resilience/availability?
☐ Does it improve security/integrity?

We begin to answer these questions by analyzing a real CMIP5 log
NDN Catalog and Retrieval

[Diagram showing NDN with nodes and data storage.]
(1) Publish Dataset names

Publisher

Catalog node 1

Data storage

Catalog node 2

NDN

Consumer

Catalog node 3

Data storage
NDN Catalog and Retrieval
NDN Catalog and Retrieval

Catalog node 1

Data storage

Catalog node 2

(2) Sync changes

Catalog node 3

Data storage

Publisher

NDN

Consumer
NDN Catalog and Retrieval

Catalog node 1

Data storage

Catalog node 2

Publisher

NDN

Catalog node 3

Consumer

Data storage
NDN Catalog and Retrieval

(3) Query for Dataset names
NDN Catalog and Retrieval

Publisher

Catalog node 1

Catalog node 2

Catalog node 3

Data storage

Data storage

Consumer

NDN

Colorado State University
NDN Catalog and Retrieval

Publisher

Catalog node 1

Catalog node 2

NDN

Data storage

Catalog node 3

Consumer
NDN Catalog and Retrieval

Publisher

Catalog node 1

(1) Publish Dataset names

Catalog node 2

NDN

(2) Sync changes

Catalog node 3

Consumer

(3) Query for Dataset names

Data storage
NDN Catalog and Retrieval

(1) Publish Dataset names

(2) Sync changes

(3) Query for Dataset names

(4) Retrieve data
NDN Catalog and Retrieval

(1) Publish Dataset names

(2) Sync changes

(3) Query for Dataset names

(4) Retrieve data
NDN Catalog and Retrieval

(1) Publish Dataset names
(2) Sync changes
(3) Query for Dataset names
(4) Retrieve data
Improvements with NDN

- **Performance** – seamless retrieval from the best performing locations
- **Publishing** – authenticated, only owner can publish
- **Discovery** – distributed catalog, anycast-style discovery
- **Resilience/availability** - seamless retrieval from multiple locations
- **Security/integrity** – enabled by signed data
Science NDN Testbed

- NSF CC-NIE campus infrastructure award
  - 10G testbed (courtesy of ESnet, UCAR, and CSU Research LAN)
- Currently ~50TB of CMIP5, ~20TB of HEP data
Vision: Integration with OS and FS

With Alex Afanasyev and Lixia Zhang

Colorado State University
Conclusions

- NDN encourages common **data** access methods where IP encourages common **host** access methods
  - NDN encourages interoperability at the content level
- NDN unifies scientific data access methods
  - Eliminates repetition of functionality
  - Adds significant security leverage
  - Rewards structured naming
For More Info

christos@colostate.edu
susmit.shannigrahi@gmail.com

http://named-data.net
http://github.com/named-data