An NDN Testbed for Large-scale Scientific Data

Huhnkuk Lim
Korea Institute of Science & Technology Information (KISTI)

NDNComm 2015
Sep. 28, 2015
Motivations on NDN for Large-scale Scientific Application

- As the data volumes and complexity increase, data-intensive science cannot rely on extension in the storage infrastructure.

- It needs to investigate new methods of intelligent processing and data distribution over networks.

- Use of caching technique changes traffic pattern in the network and improves corrupted data rate.

- NDN based large-scale scientific application
  - Climate modeling application as an initial focus
  - Extension of NDN architecture to various data-intensive science application such as HEP and astronomy with hierarchical naming strategies

- Innovative data management lead to traffic pattern change
Backgrounds on NDN for Climate Modeling Application

Why climate data transfer using NDN Architecture

- Current CMIP5 data transfer using ESGF, long time latency and corrupted data occur
- To provide innovative transfer, management, and security function for scientific big data using the NDN architecture
- Movement of traffic pattern in data-intensive science and reduction of data explosion on it

R&D on NDN based data-intensive science application

- NDN testbed for climate modeling application (CSU univ.)
- NDN architecture design, development, and deployment for LHC big data transfer (Fermi Lab)
- ESnet for research networks in US

Data-intensive science applications

1. Climate Modeling
2. HEP (LHC, CMS)
3. Astronomy

Climate modeling NDN testbed in US
NDN Testbed for Climate Modeling Application

**Graphic User Interface (Web browser)**

- **NDN Consumer**
- **NDN Consumer SW**
  - NDNJ5; SimpleHTTPServer; firefox addon
  - Ndn-cxx
  - Tables: CS, BIT, FIB
  - Forwarding Engine: name-based routing
  - Faces: Local, Remote
  - TCP, UDP, IP...
  - Ethernet
- NDN router
- **NDN Producer**
- NDN Name Translator
- **Climate data Repository**
- NDN Producer SW
- Ndn-cxx
  - Tables: CS, BIT, FIB
  - Forwarding Engine: name-based routing
  - Faces: Local, Remote
  - TCP, UDP, IP...
  - Ethernet

**Functions of front-end system in consumer**
- To provide GUI for climate modeling application based on NDN architecture
- CMIP5 data search using controlled vocabulary
- NDN name based CMIP5 data downloading

**Kisti-ndn-atmos package**

**Functions of back-end system in producer**
- To translate .nc file names to NDN names
- NDN based repository establishment for CMIP5 data management
- NDN name database establishment, in order to search a CMIP5 data of interest in producer
Key Components in the NDN Testbed

- Name lists sorting
- To show metadata corresponding to each searched CMIP5 data
- Search results is changed to CMIP5 file name following DRS syntax

Works to support NDN based Climate Modeling Application

- NDN Name Translator for climate modeling application
- To translate CMIP5 data files stored in NDN repository to NDN names and to store them in DB
- NDN name translation following DRS structure

NDN network for climate modeling in Korea

- GUI to support NDN based climate modeling application

Key Components in the NDN Testbed
- Forwarding and caching of interest/data packets
- Synchronised FIB table management in the NDN testbed
- NDN platform (ver 0.3.4)
  - NDN-cxx, NFD
  - NDN-js (one of NDN-ccl)
  - NDNfs-port
Features of GUI (1)

- Reflection of the ESGF system workflow
- CMIP5 climate data searching following climate DRS structure
  - To show original CMIP5 nc file names changed from NDN names, together with metadata sets corresponding to .nc file names
  - Key word based CMIP5 data search and user-friendly sorting for search results
Features of GUI (2)

- **CMIP5 data downloading in metadata window**
  - Download button have the address corresponding to an NDN name of interest in producer side
    - Address: NDN name based URI
    - “ndn:/catalog/myUniqueName/<CMOR filename.nc>”
      - ex) ndn:/catalog/myUniqueName/psl_amip_MIROC5_historical_r1i1p1_1950010100-xx.nc

<Downloading of CMIP5 climate data>
Features of Name Translator

- To translate all nc file names stored in repository to NDN names
  - Parsing of each name component
  - To check time variable in an nc file has the same value in metadata
    - Sometimes, time in metadata is slightly different from one in real data.
    - For allowable error range, name translation for an nc file name.
    - If they are outside from it, no translation for that one.

6 nc files in NDN file system (repository)

6 CMIP5 NDN names translated in Mysql DB repository

Database schema => http://redmine.named-data.net/projects/ndn-atmos/wiki/Schema
### Summary of kisti-ndn-atmos SW package

<table>
<thead>
<tr>
<th>Key function</th>
<th>kisti-ndn-atmos</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User Interface</strong></td>
<td></td>
</tr>
<tr>
<td>Data search</td>
<td>To show .nc file name lists following DRS structure</td>
</tr>
<tr>
<td>Metadata</td>
<td>Supported</td>
</tr>
<tr>
<td>File downloading</td>
<td>Supported</td>
</tr>
<tr>
<td>User-friendly functions</td>
<td>Sorting and key word based searching</td>
</tr>
<tr>
<td>Name translator</td>
<td>NDN name translation for valid climate data</td>
</tr>
<tr>
<td>Repository for NDN</td>
<td>To provide a repository using ndnfs-port</td>
</tr>
</tbody>
</table>

There have been significant code sharing between KISTI and CSU project, in order to develop each ndn-atmos SW package for climate application.
Climate Data Transfer by Federated NDN Testbed in Korea and US

- Transfer by the Earth System Grid Federation (ESGF) infrastructure
  - ESGF: Distributed CMIP5 data management protocol in current IP based networks
  - Data explosion for duplicate big data requests results in BW waste

- Transfer by federated NDN testbeds
  - Smart transfer for duplicate big data requests
  - Change of traffic pattern results in traffic reduction in networks
  - Prevention of data explosion in networks

ESGF architecture based CMIP5 delivery

- Current works on federated NDN Testbed in Korea and US
  - Interoperability for front and back-end systems in each domain
  - To create synchronized FIB tables to search for all CMIP5 data sets at each producer using NLSR
  - Caching scheme for large scale scientific data

NDN based CMIP5 delivery
Summary and Future Works

- Current climate data transfer by ESGF results in long time latency and high corrupted data rate.

- To provide large-scale scientific data with innovative transfer and management.

- To change traffic pattern in data-intensive science and to prevent data explosion in networks.

- **NDN testbed with kisti-ndn-atmos package for climate application**
  - Front-end system in consumer and back-end system in producer
  - To show original climate .nc file names following DRS and corresponding metadata sets
  - Key word based climate data search and downloading
  - To translate all .nc file names stored in the NDN repository to NDN names
  - Forwarding and caching of interest/data packets on climate modeling application

- **Future works**
  - Federated NDN testbed in Korea and US for climate modeling application
  - Performance analysis for ESGF and NDN based transfer
  - Caching and mobility to consider characteristics of large-scale scientific data