NAMED DATA NETWORKING IN SCIENTIFIC APPLICATIONS

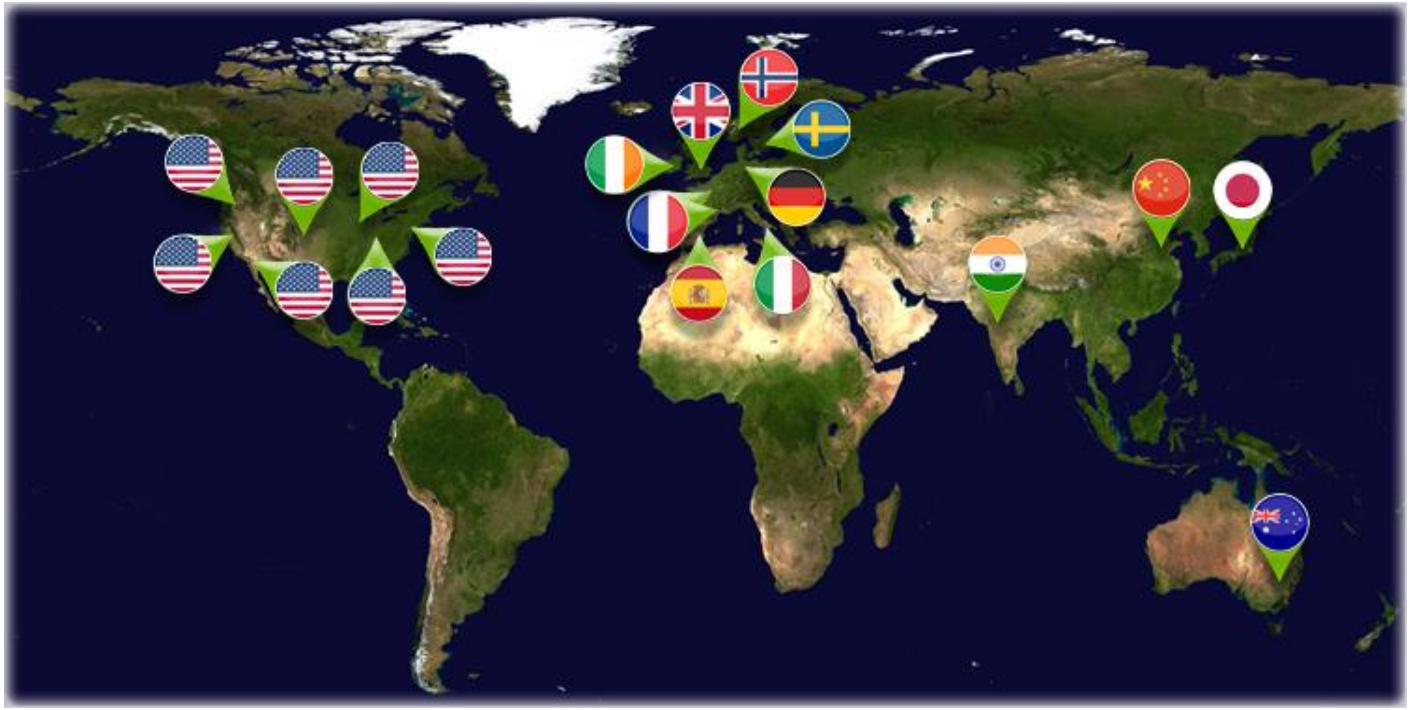
Susmit Shannigrahi, Chengyu Fan and Christos Papadopoulos

Colorado State University

March 23, 2017

Work supported by NSF #1345236 and #13410999

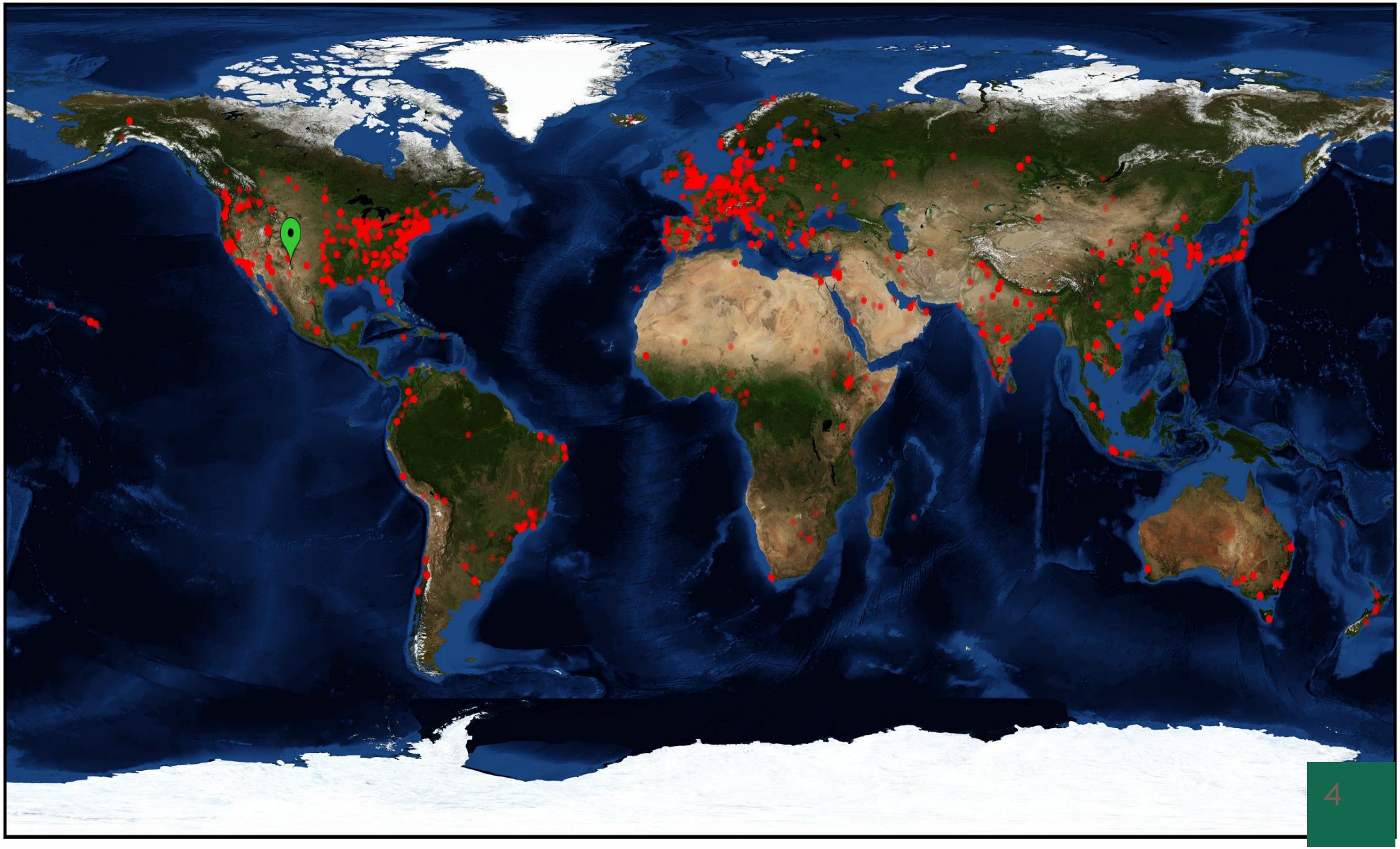
CMIP5 Servers



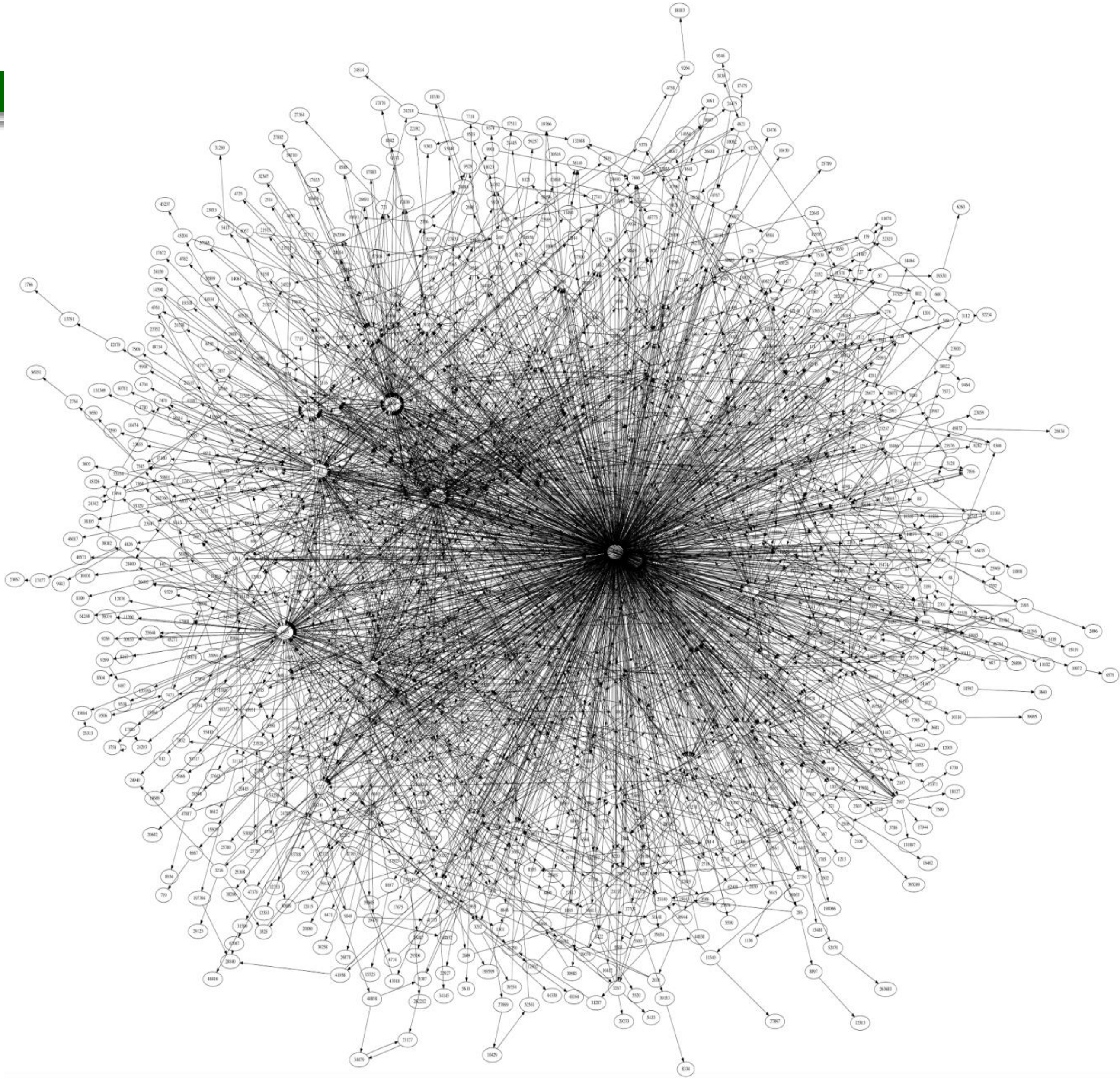
3 Years of CMIP5 Data Access

- 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 dups) = 1,844TB

Client Locations



ASN Map

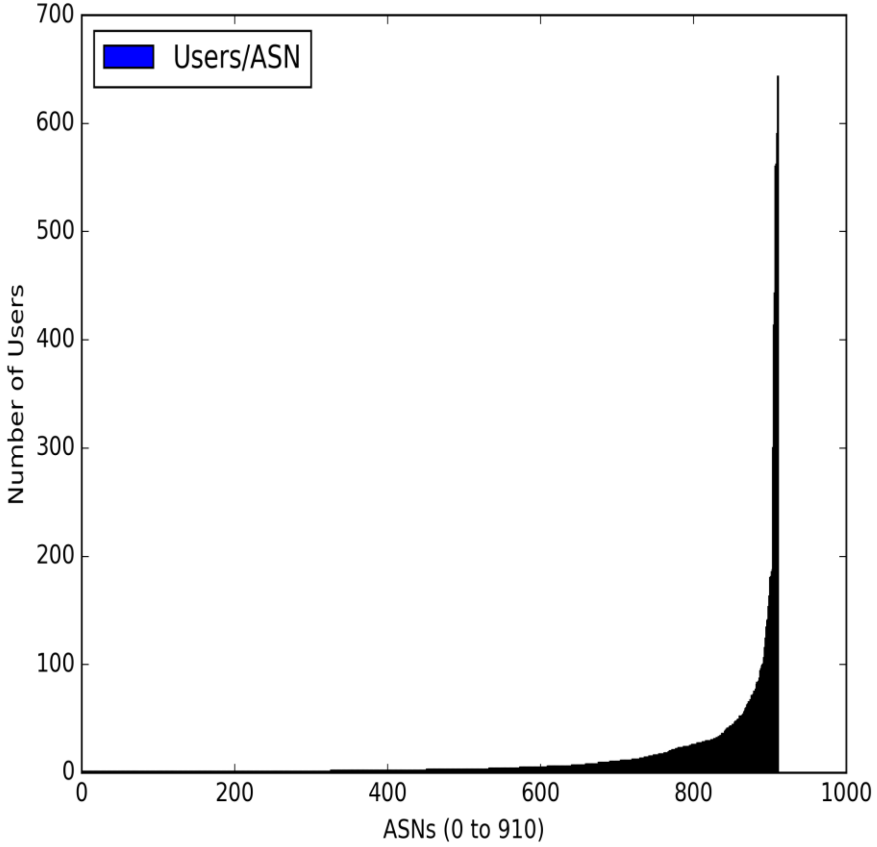


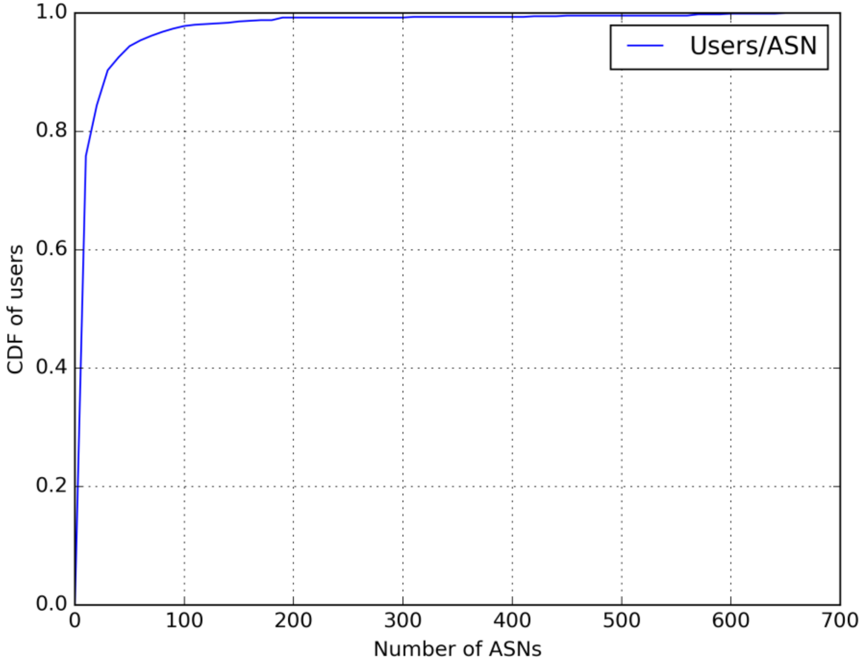
- Done using reverse traceroute
- Little path overlap, but view from only one ESGF node

User/Clients Statistics

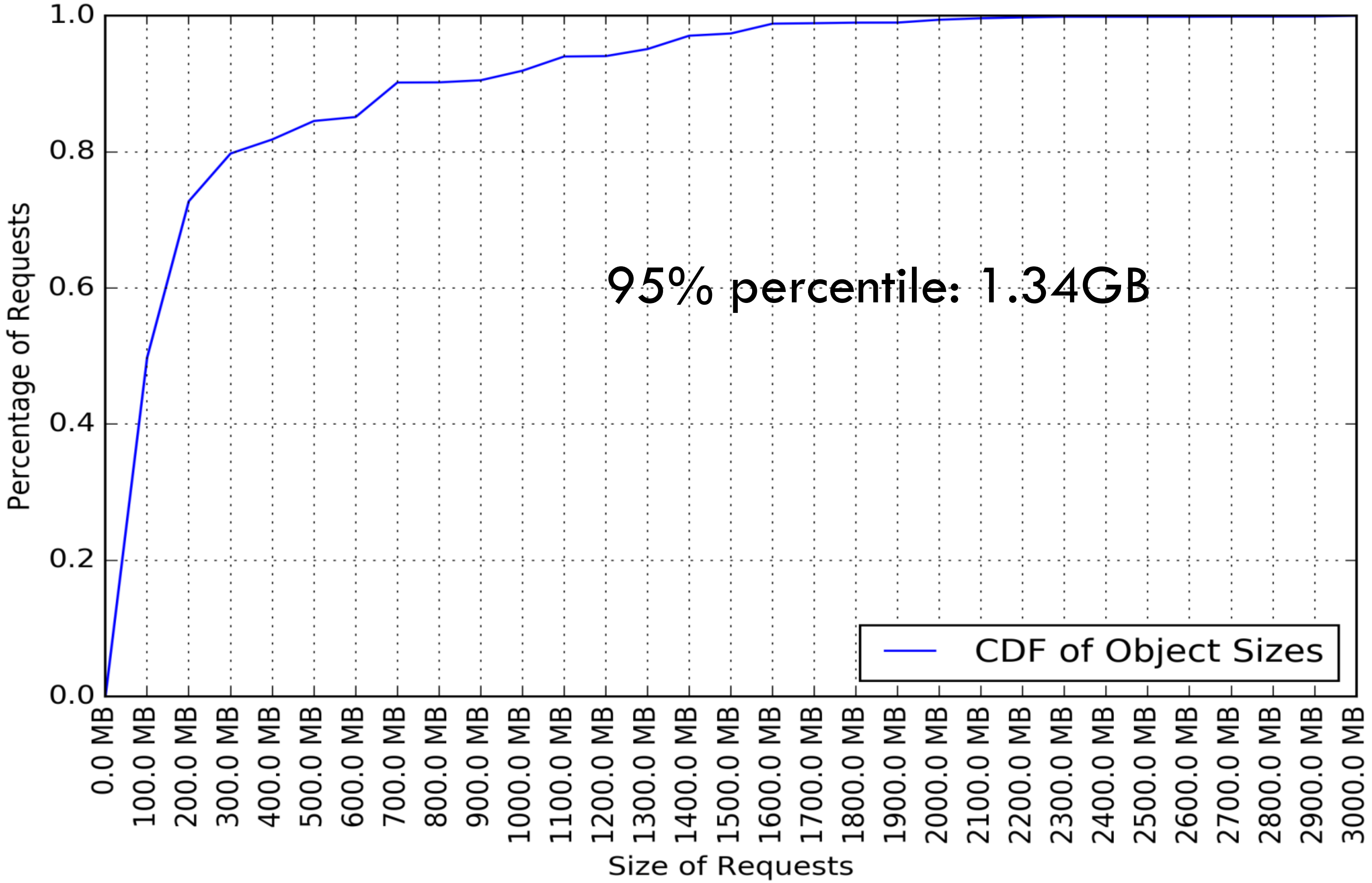
Unique Users	5692
Unique Clients (IP addresses)	9266
Unique ASNs	911

User Distribution per ASN





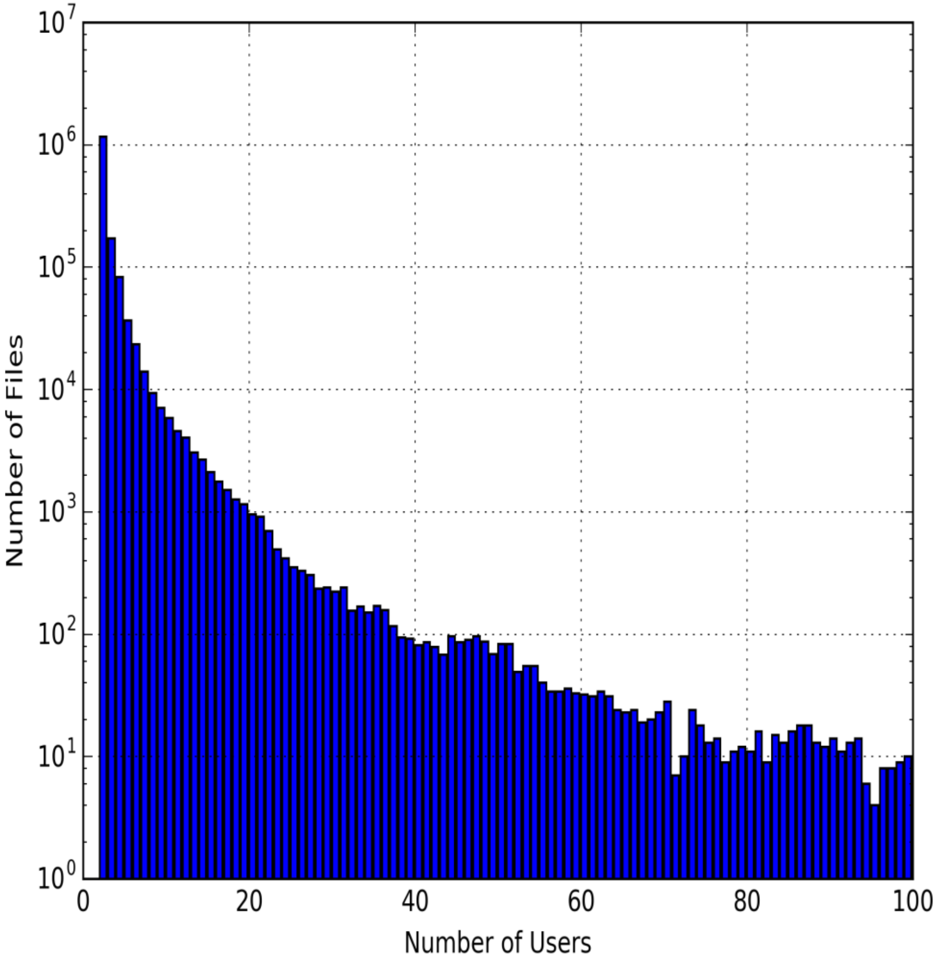
Dataset Size Distribution



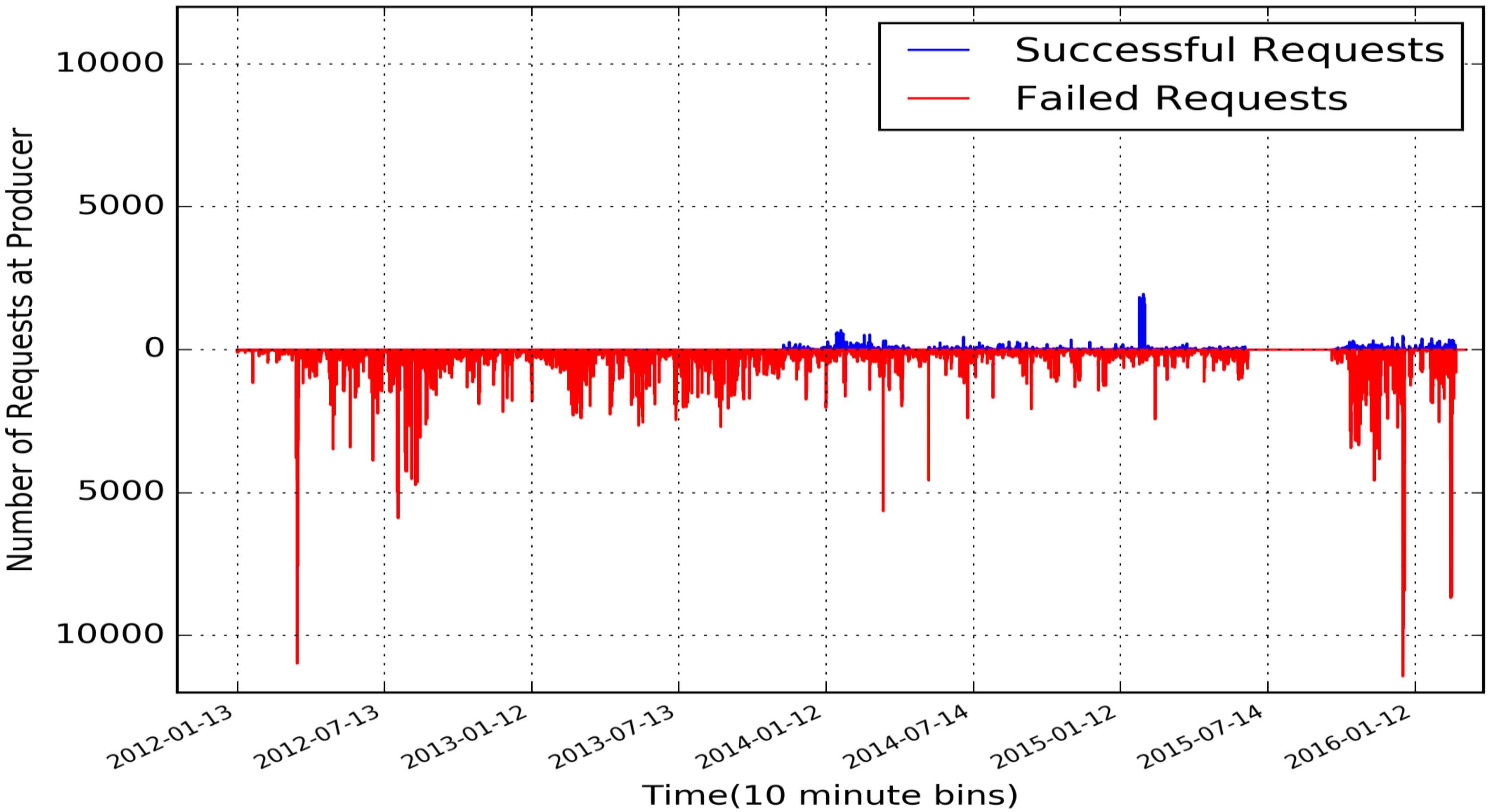
95% percentile: 1.34GB

Data Popularity

98% of the datasets
was requested
by 10 users or less)



Successful vs Failed Requests



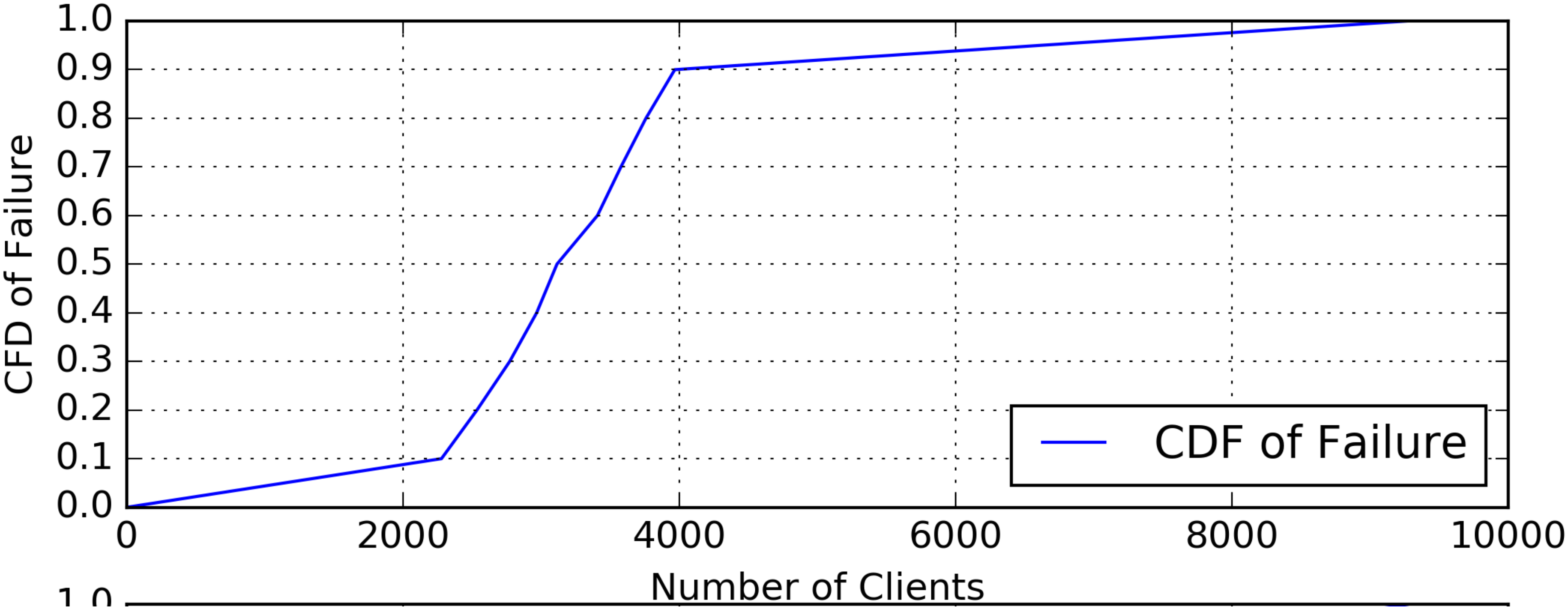
Summary: Data Statistics

CMIP5 Archive Size	3.3PB
Total Data Requested	Equivalent of 1.8PB (18.5M requests)
Total Data Successfully Retrieved	234 TB (1.9M requests)
Total Data Successfully Retrieved (Excluding Duplicates)	113 TB (415K requests)
Number of unique datasets requested	1.5 million

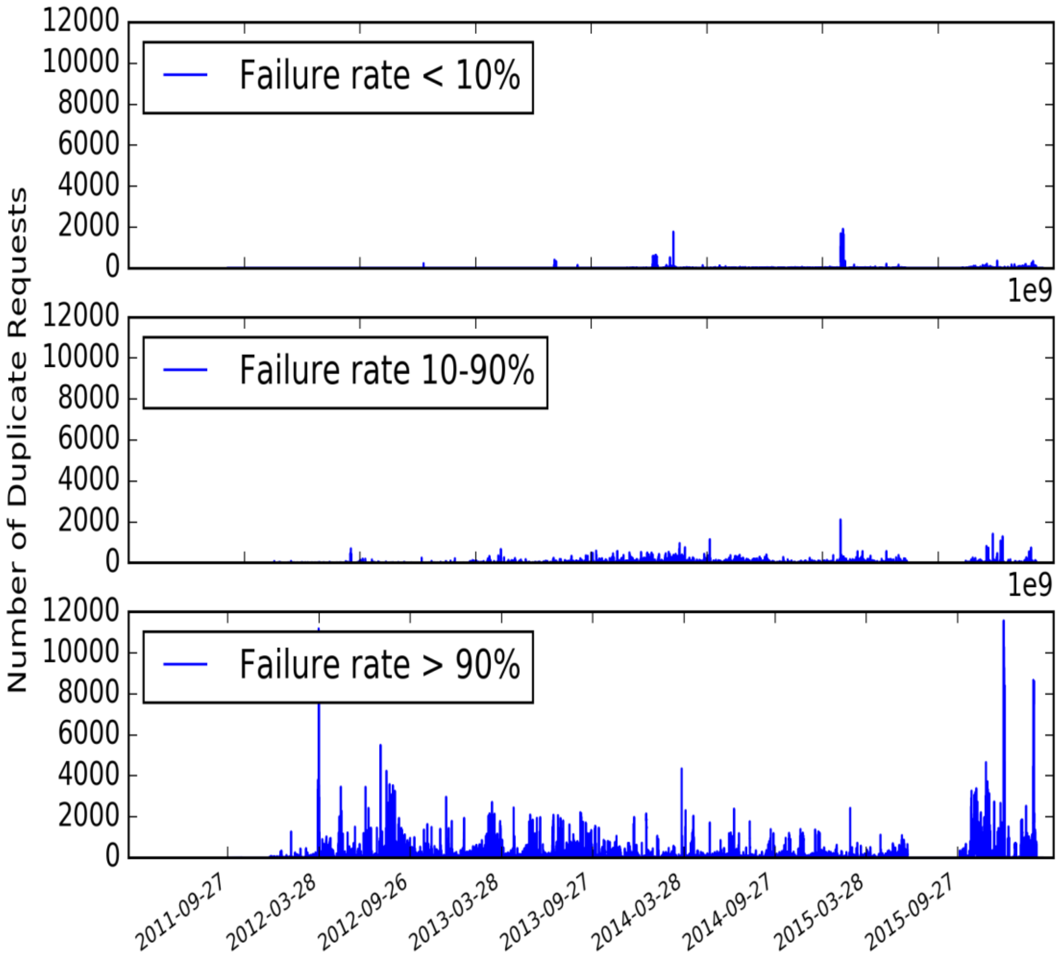
A Closer Look at Failures

Number of requests	18.5 million
Successful Requests	1,935,256
Failed Requests	16,673,815

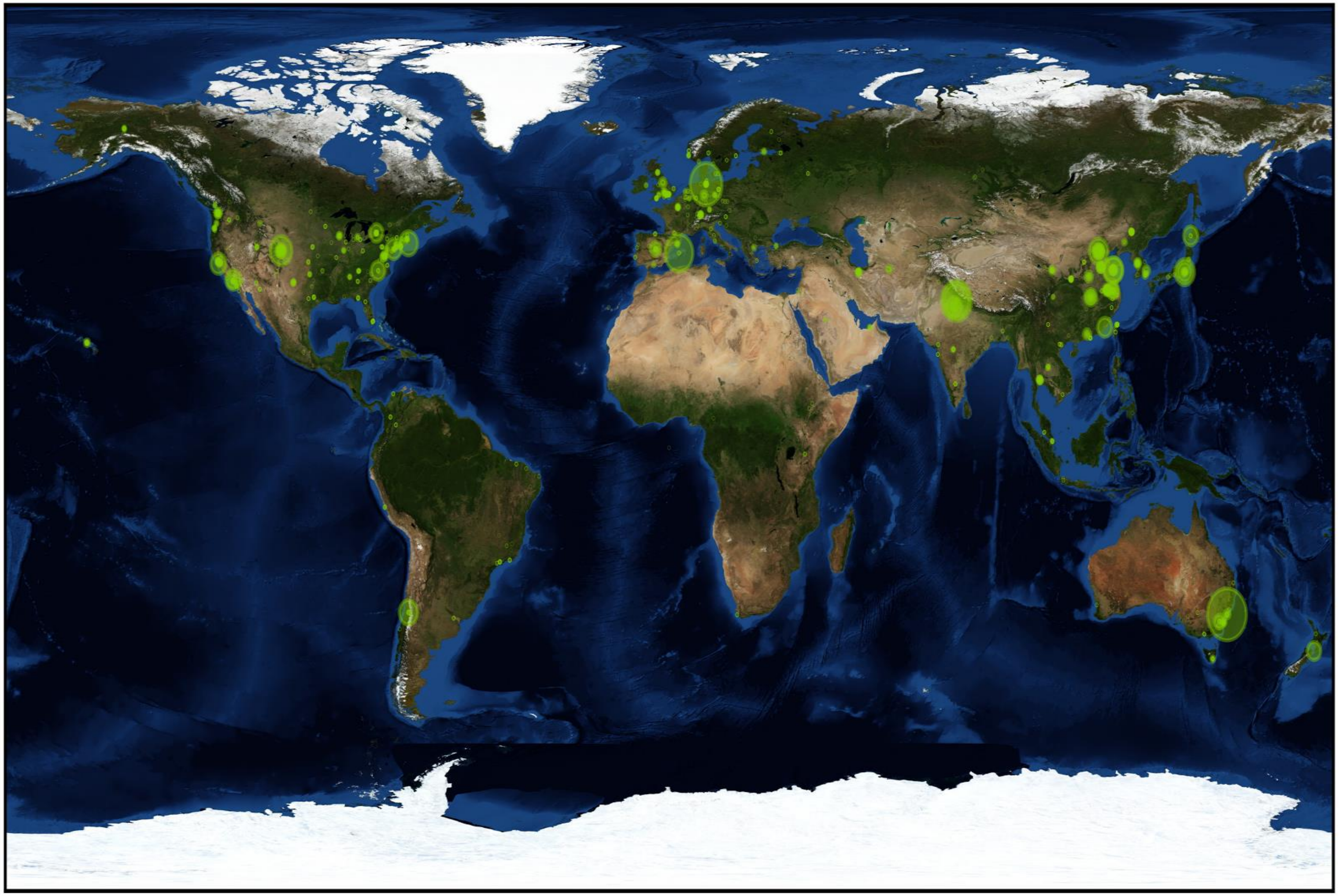
Client Request Failures



Duplicate Requests by Failure Group



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

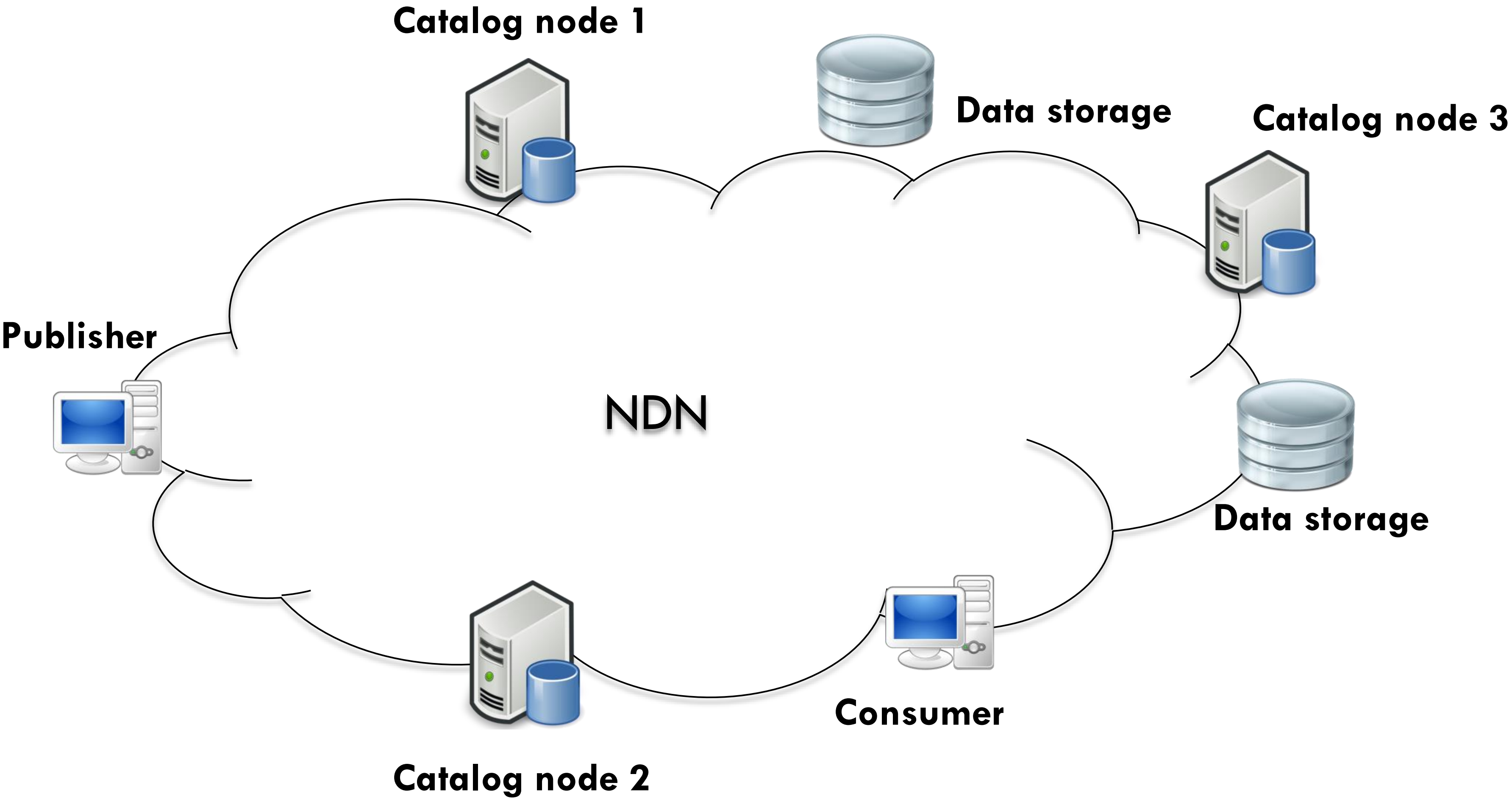
CMIP5 Retrieval with NDN

□ ~~HTTP://someESGFnode:~~/CMIP5/output/MOHC/HadCM3/decadal1990/day/atmos/tas/r3i2p1/tas_Amon_HADCM3_historical_r1i1p1_185001-200512.nc

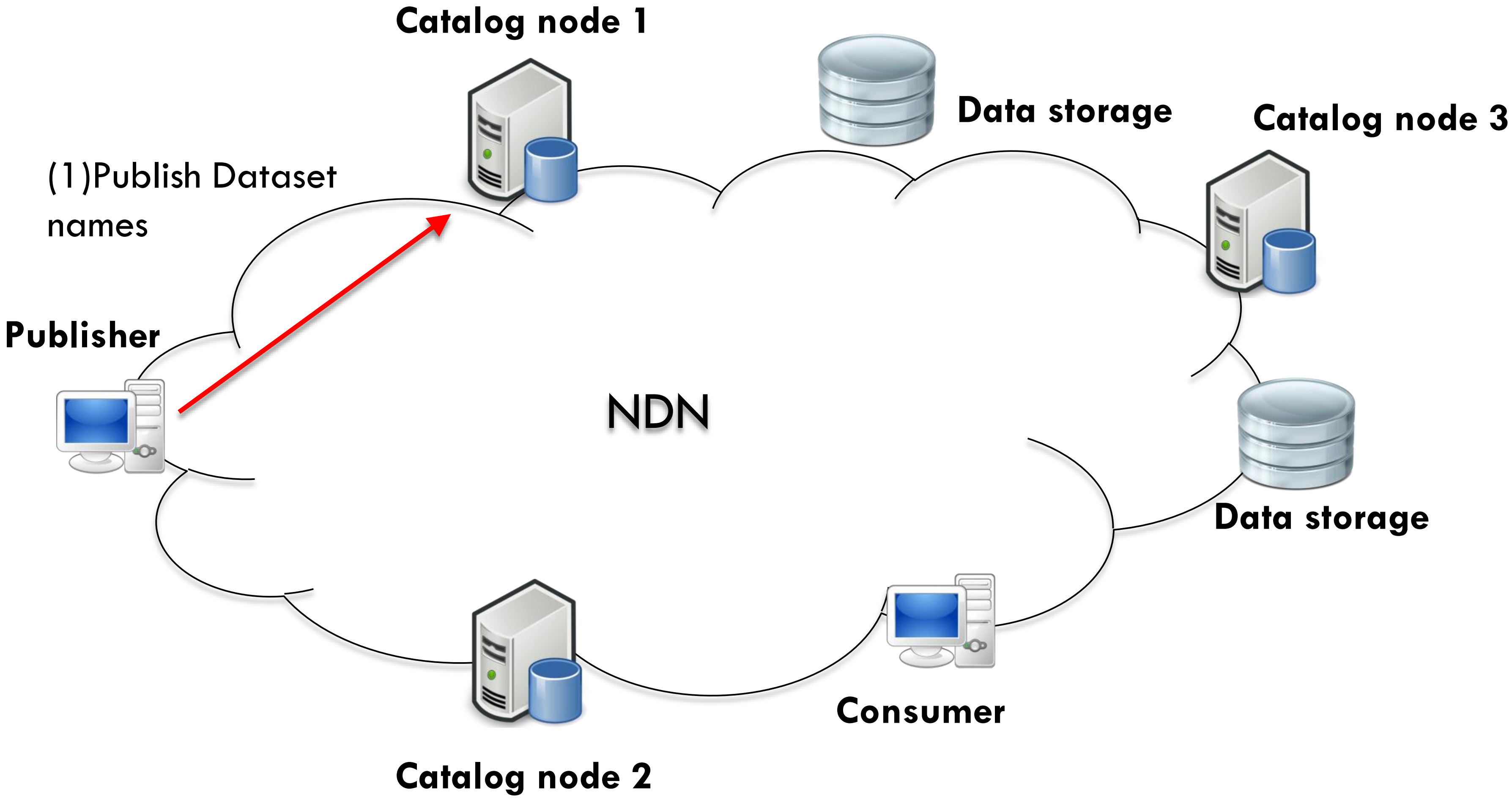
Why make the change?

- 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

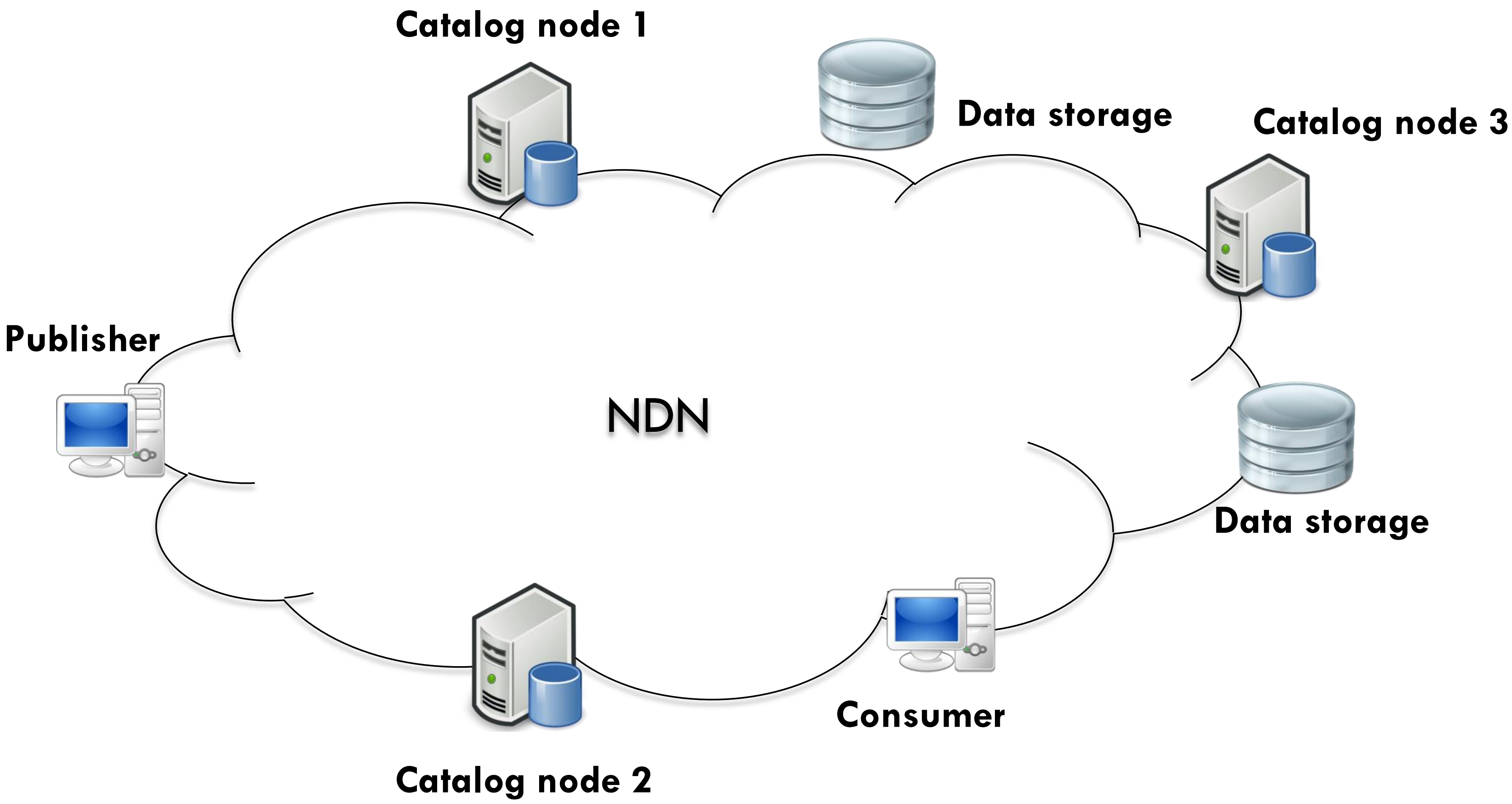
NDN Catalog and Retrieval



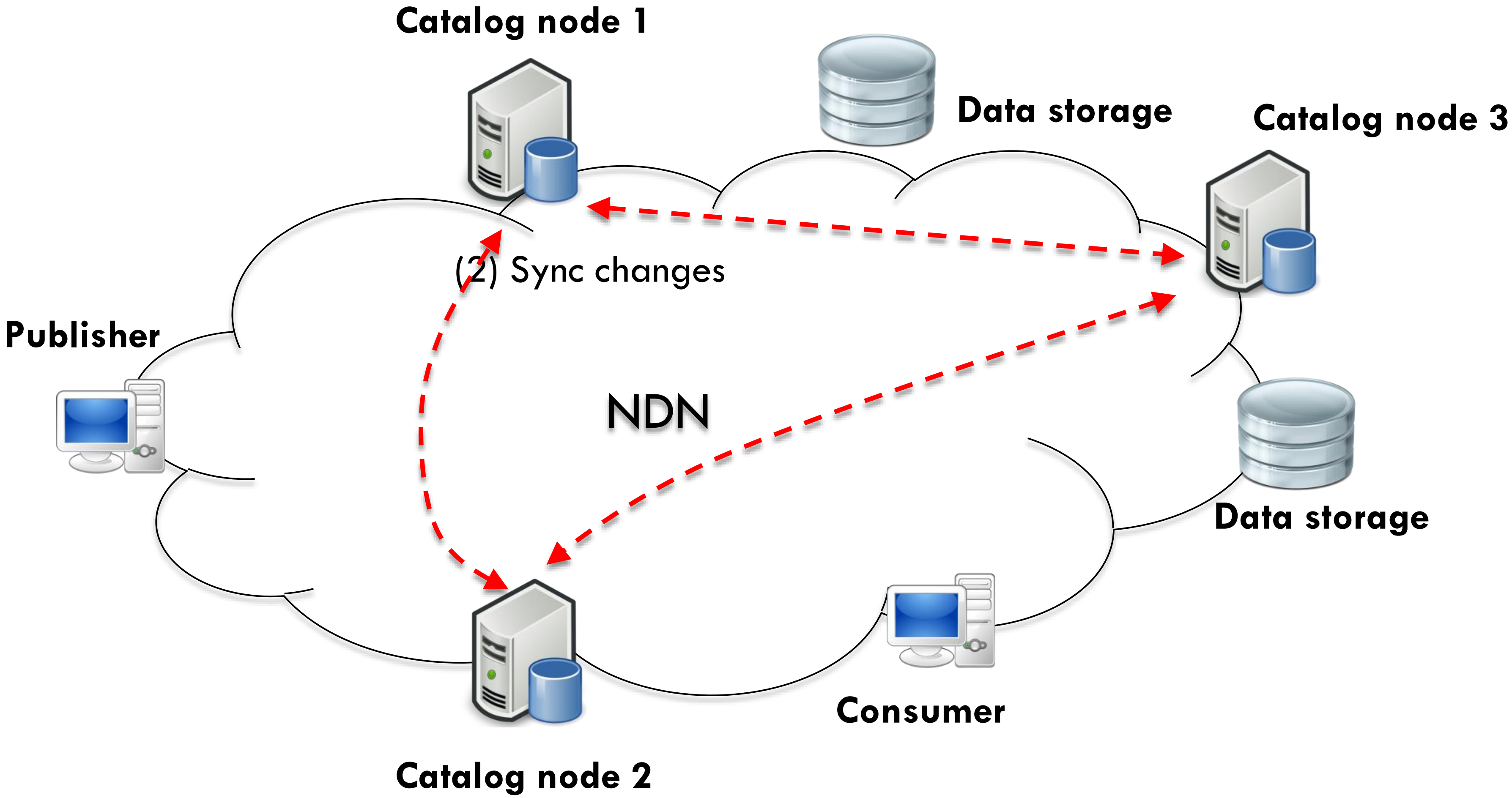
NDN Catalog and Retrieval



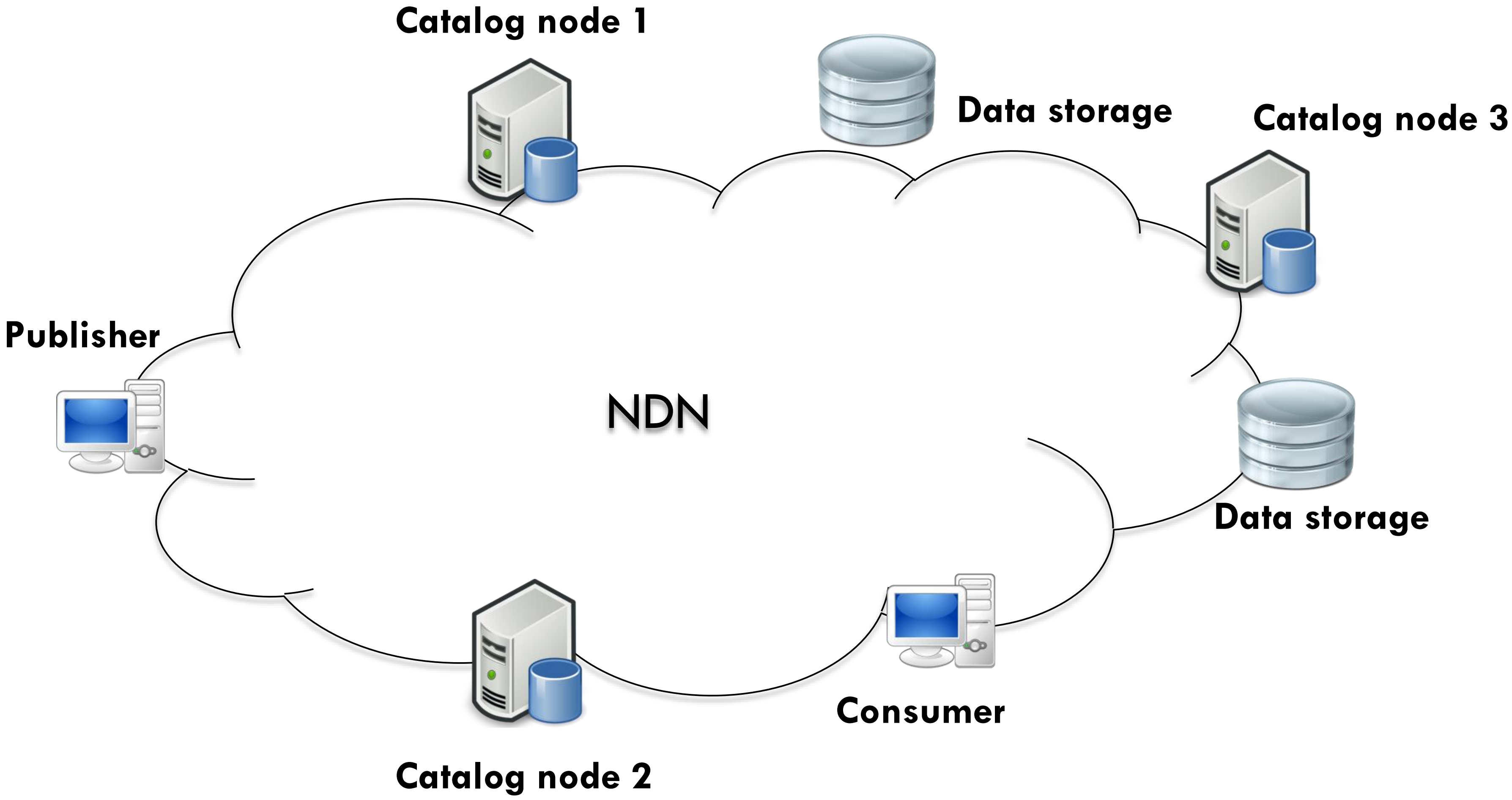
NDN Catalog and Retrieval



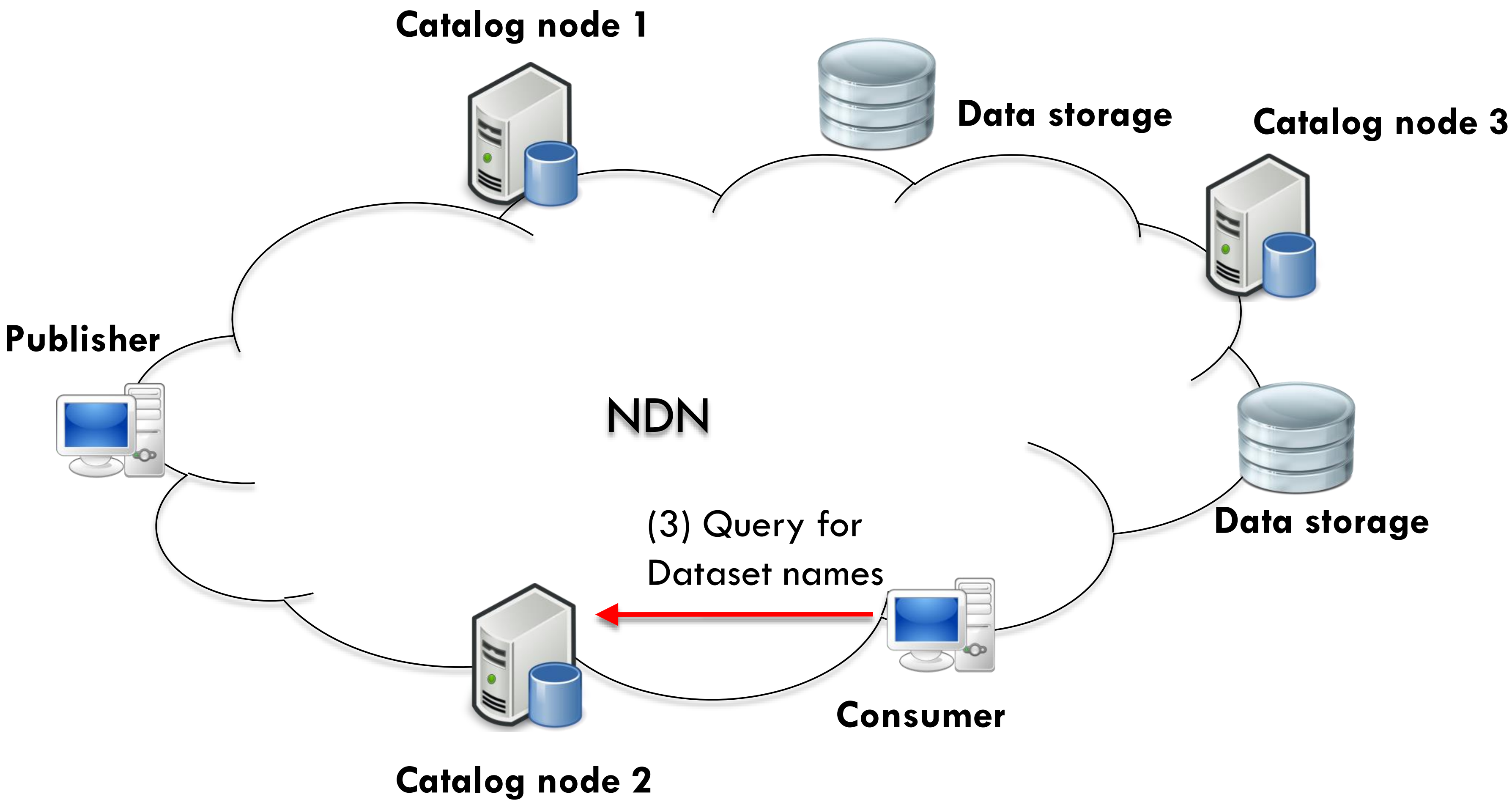
NDN Catalog and Retrieval



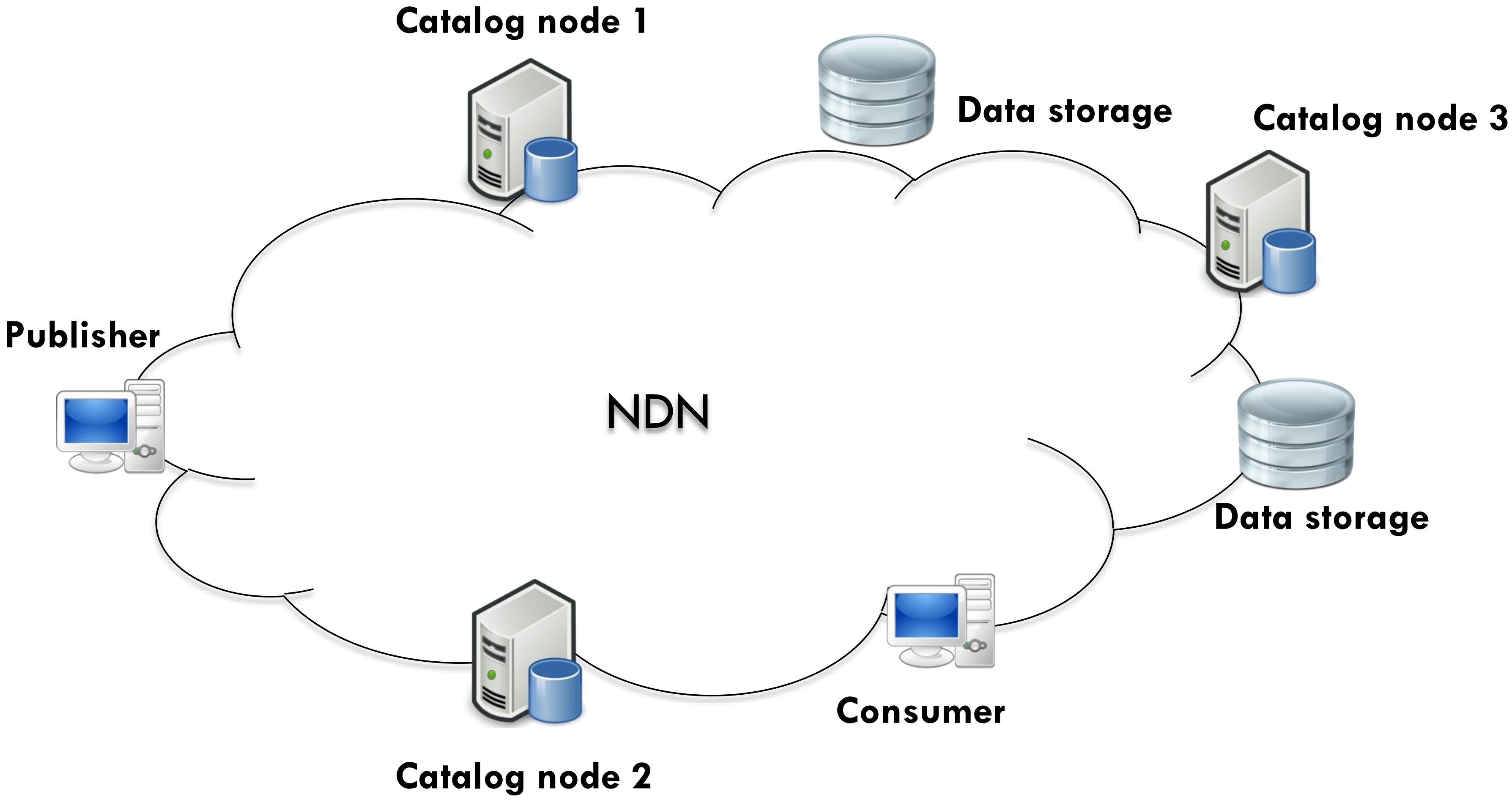
NDN Catalog and Retrieval



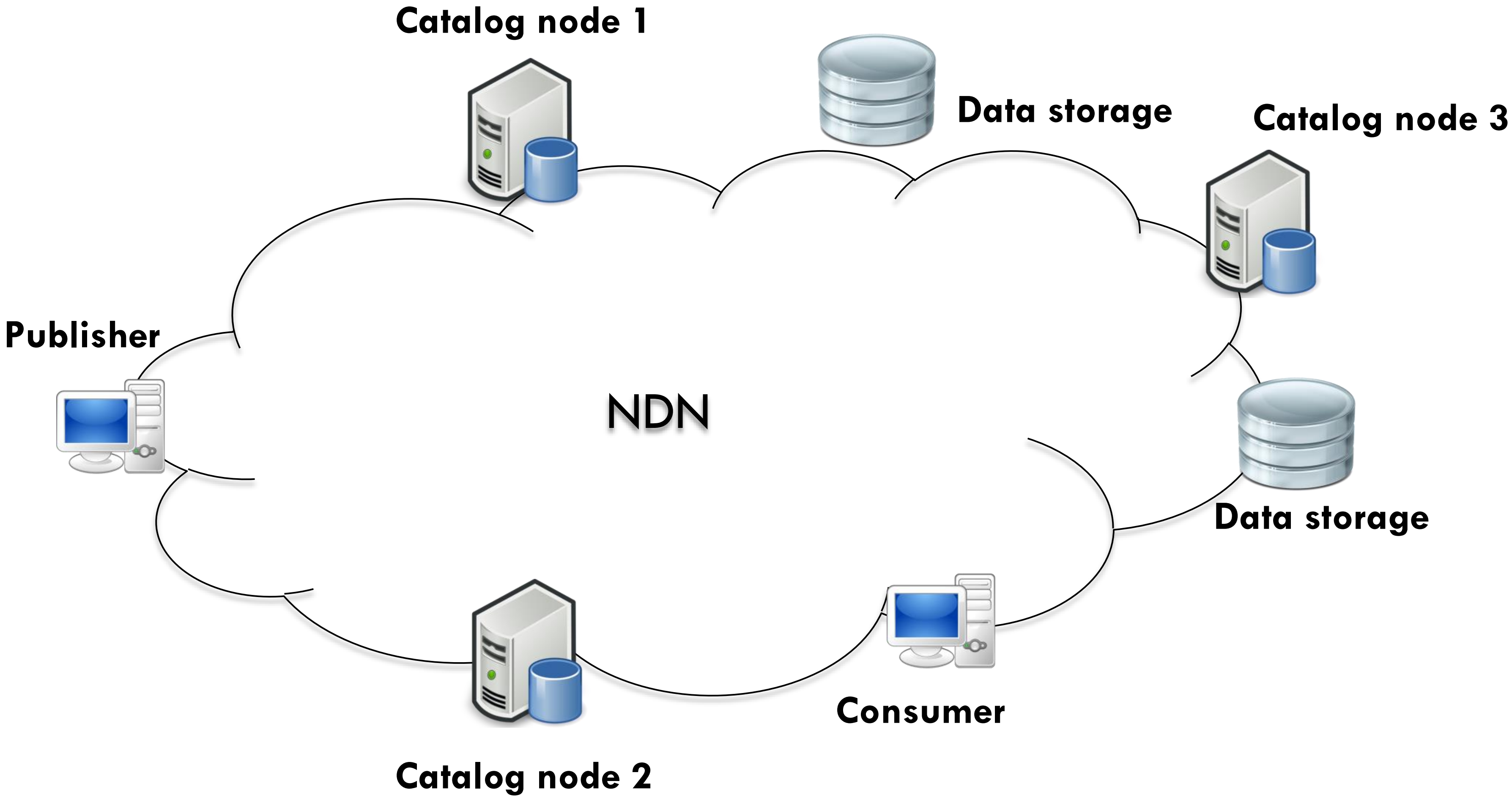
NDN Catalog and Retrieval



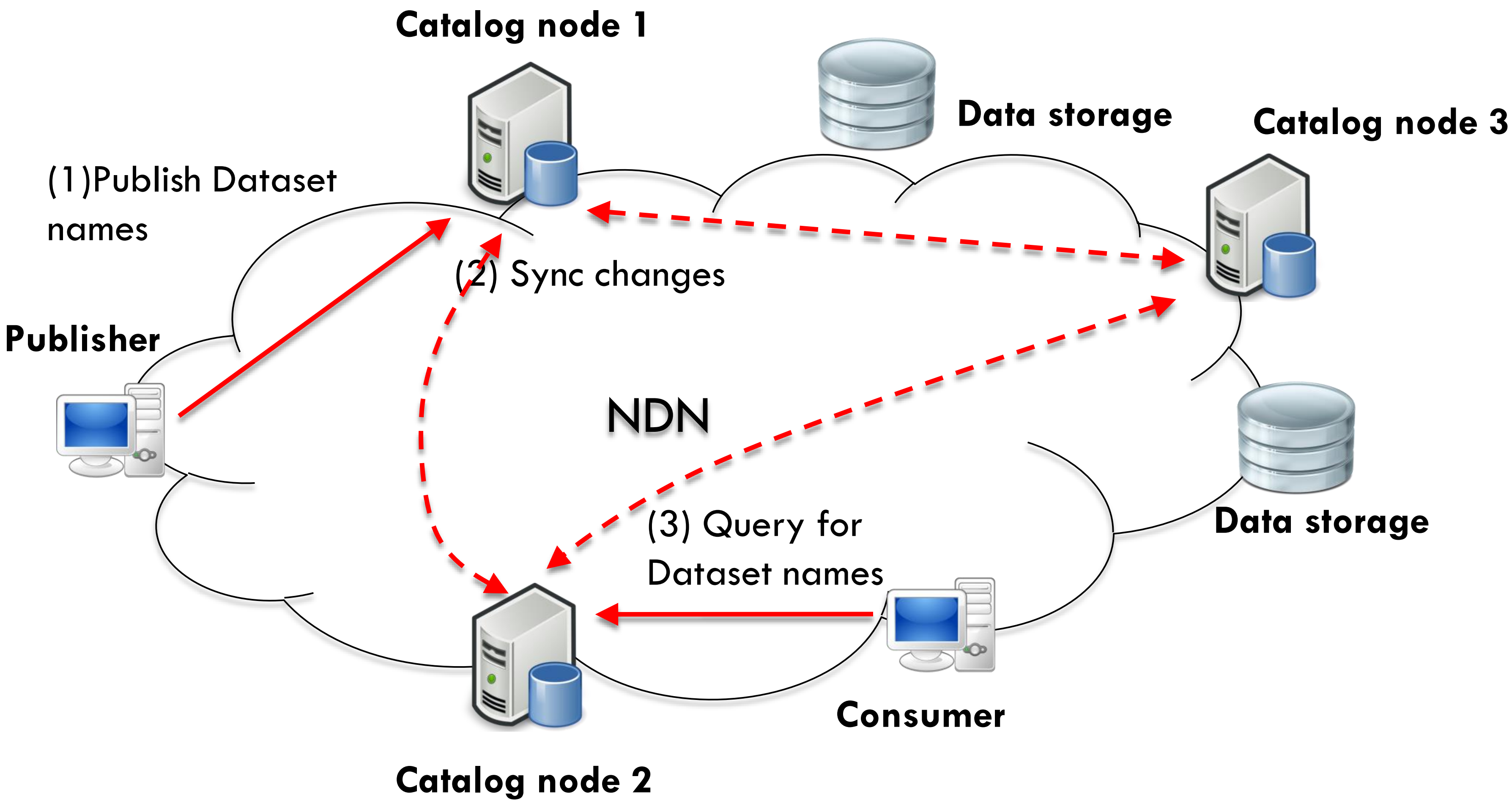
NDN Catalog and Retrieval



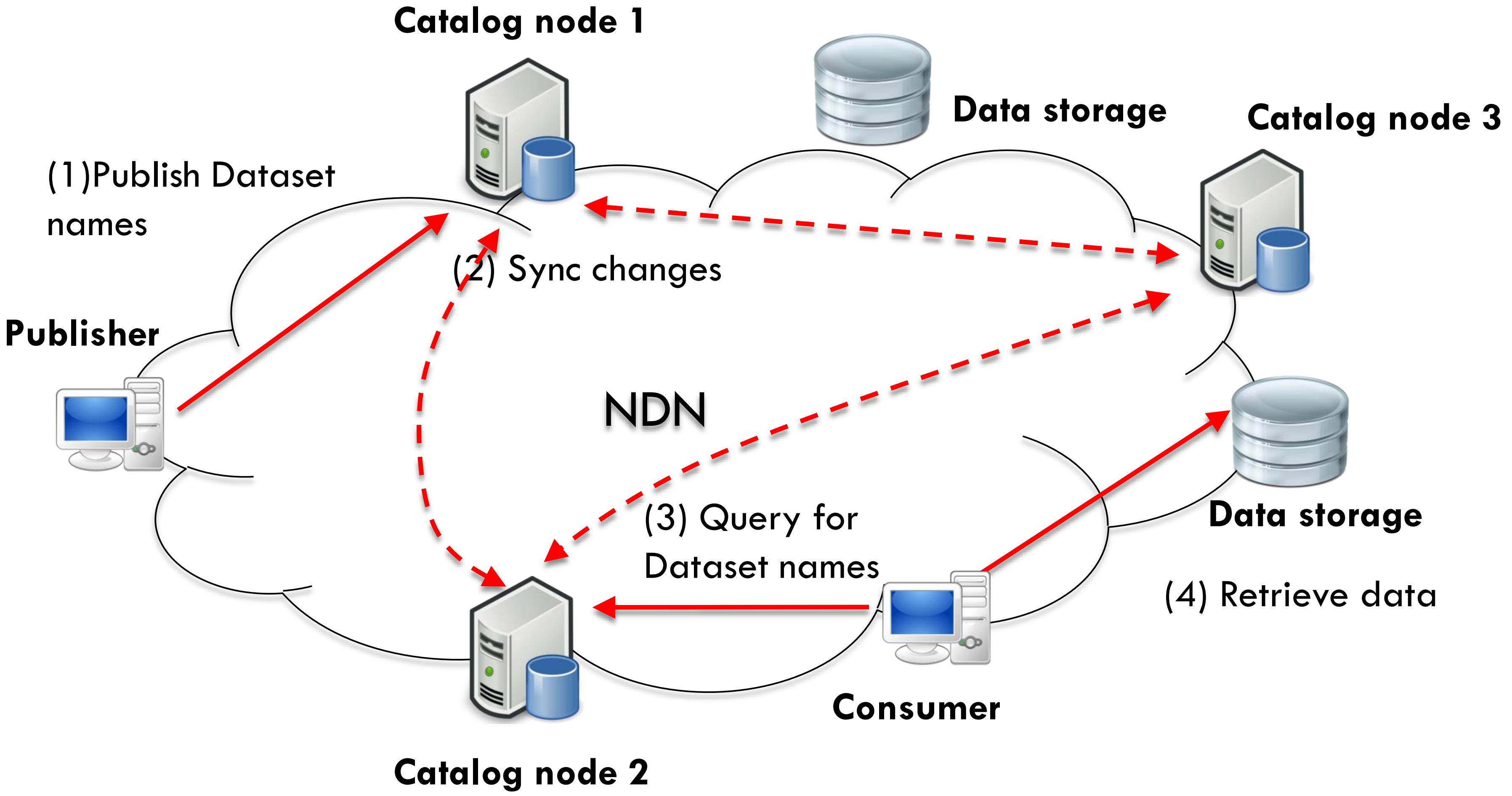
NDN Catalog and Retrieval



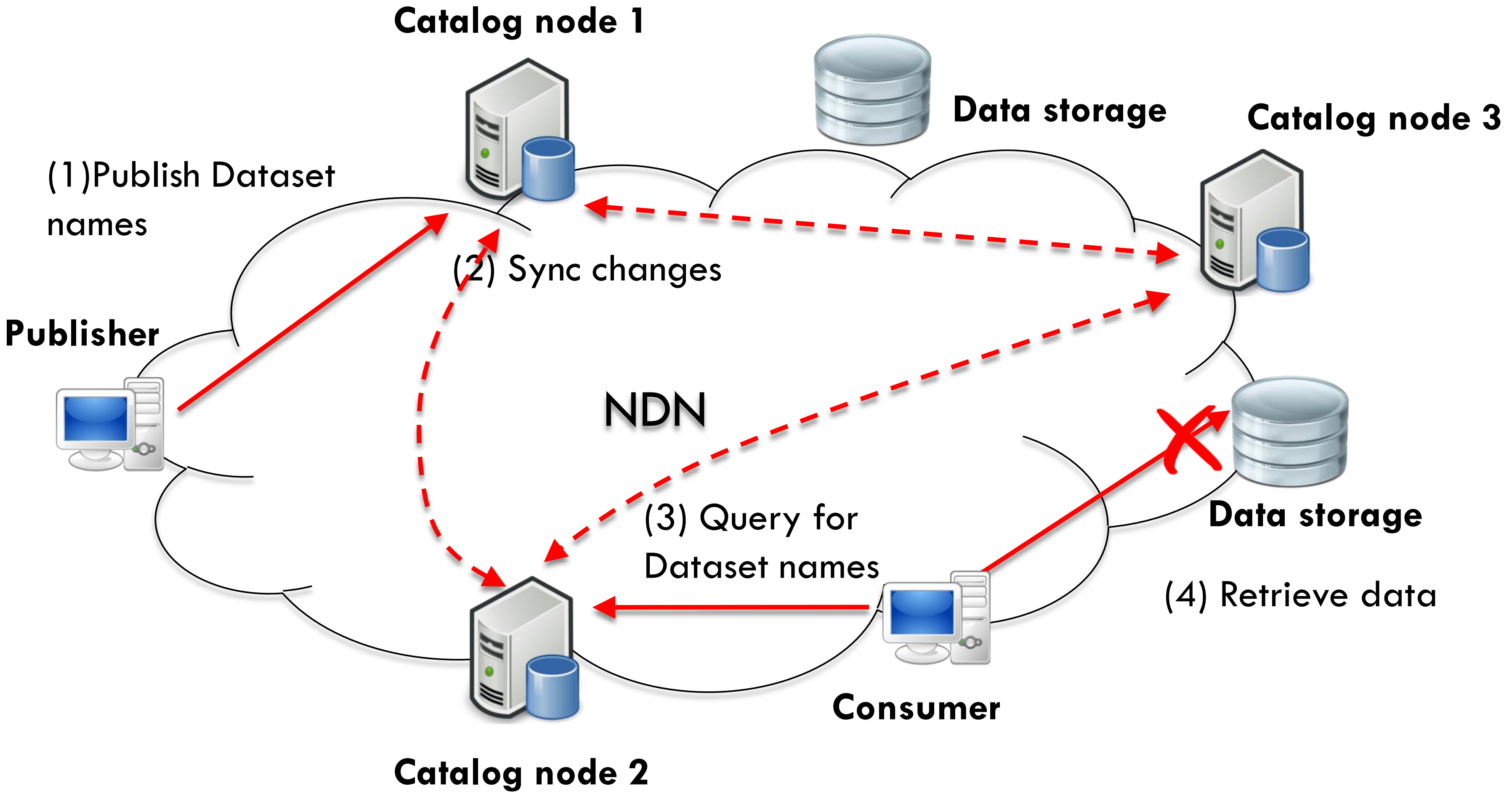
NDN Catalog and Retrieval



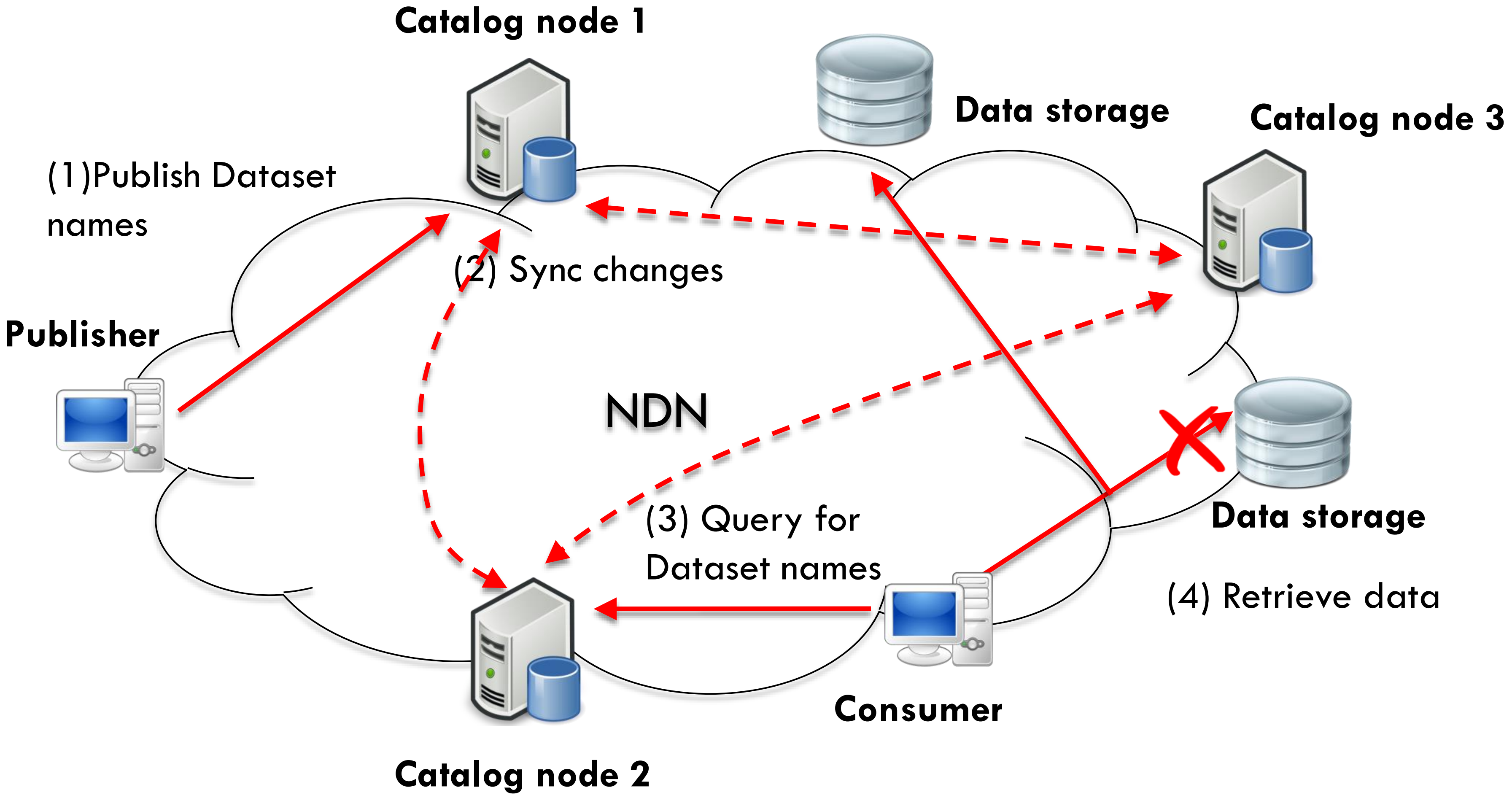
NDN Catalog and Retrieval



NDN Catalog and Retrieval



NDN Catalog and Retrieval



Catalog node 1

Data storage

Catalog node 3

(1)Publish Dataset names

(2) Sync changes

Publisher

NDN

(3) Query for Dataset names

Data storage

(4) Retrieve data

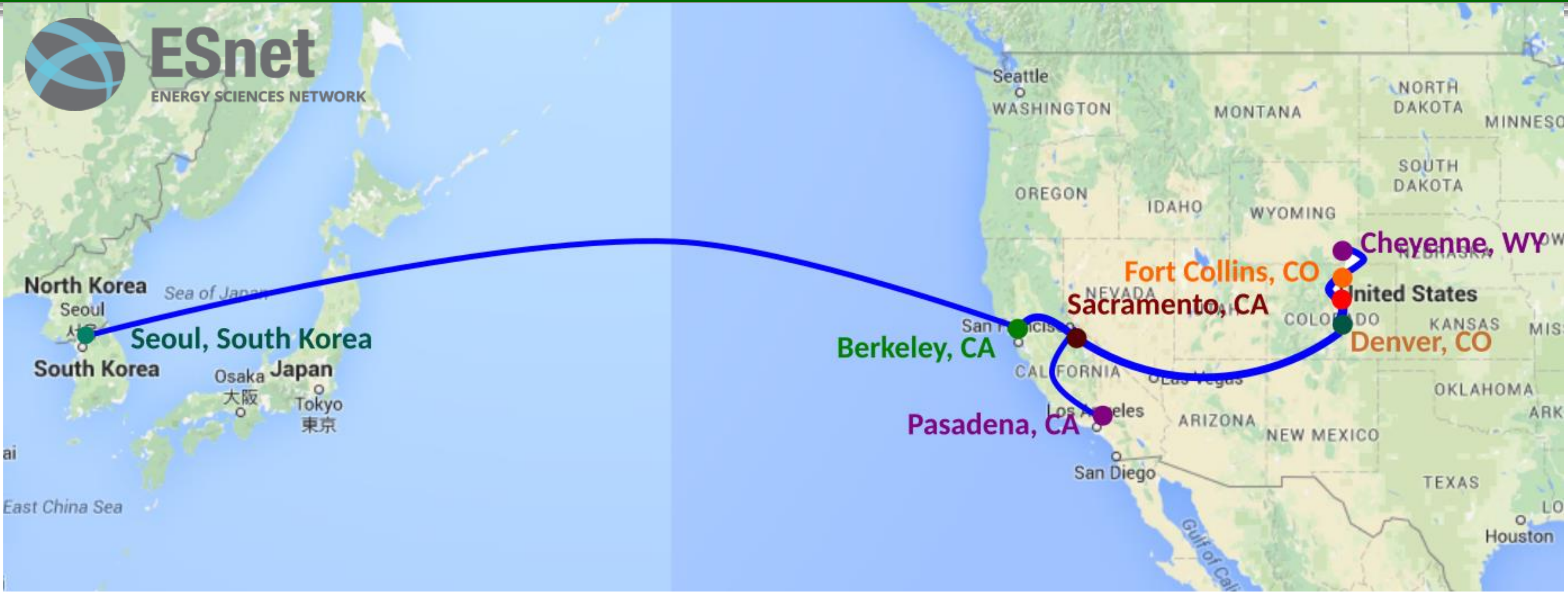
Consumer

Catalog node 2

Improvements with NDN

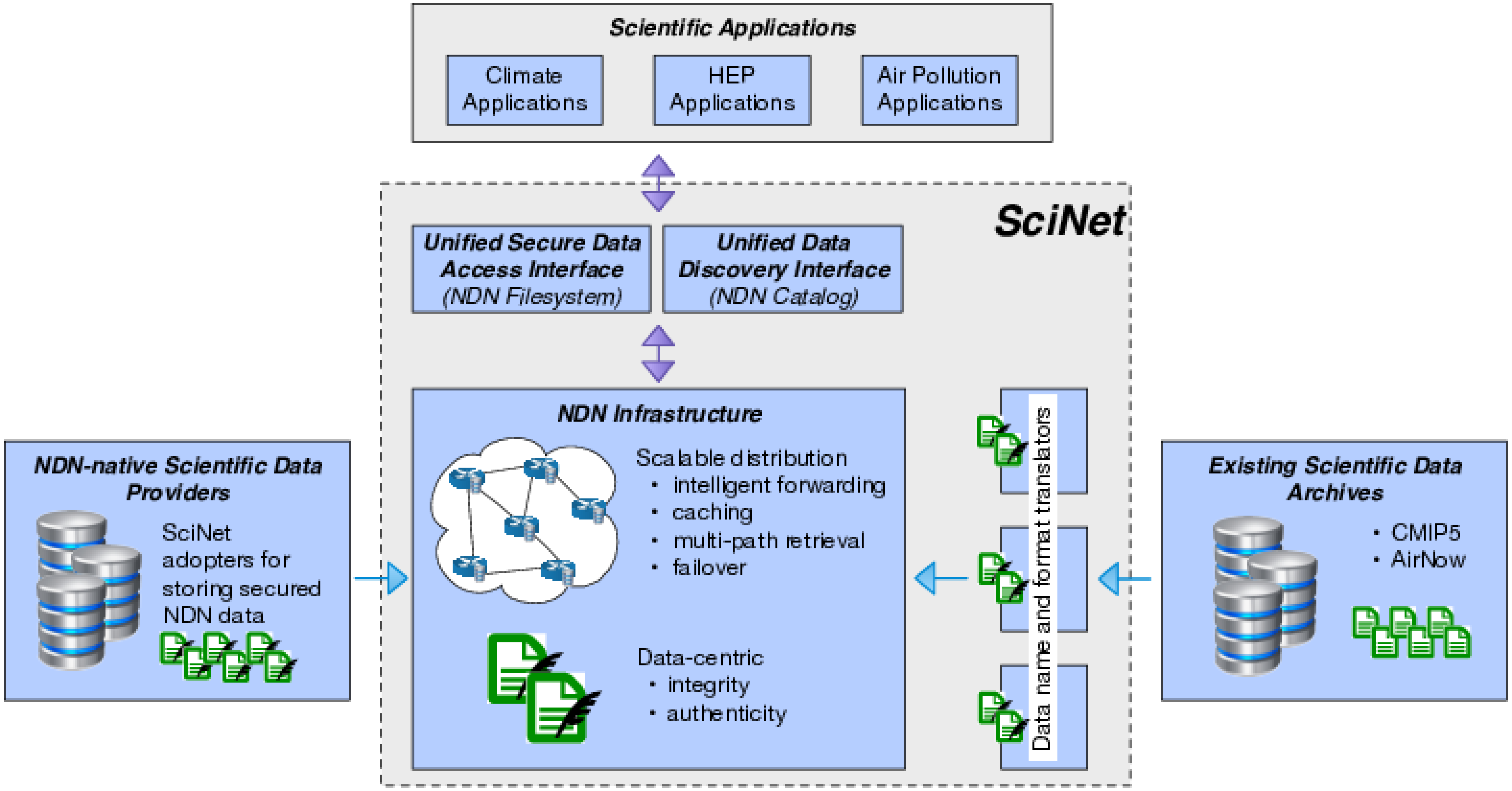
- **Performance** – seamless retrieval from the best performing locations
- **Publishing** – authenticated, only owner can publish
- **Discovery** – distributed catalog, anycast-style discovery
- **Resilience/availability** - seamless retrieval from multiple locations
- **Security/integrity** – enabled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


Scientific Applications
Climate Applications
HEP Applications
Air Pollution Applications
SciNet
Unified Secure Data Access Interface (NDN Filesystem)
Unified Data Discovery Interface (NDN Catalog)
NDN Infrastructure
Scalable distribution
• intelligent forwarding
• caching
• multi-path retrieval
• failover
Data-centric
• integrity
• authenticity
Data name and format translators
NDN-native Scientific Data Providers
SciNet adopters for storing secured NDN data
Existing Scientific Data Archives
• CMIP5
• AirNow

With Alex Afanasyev and Lixia Zhang

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