NDN-SCI for Managing Large Scale Genomics Data

Susmit Shannigrahi
Colorado State University
susmit@cs.colostate.edu

Chengyu Fan
Colorado State University
chengyu.fan@colostate.edu

Christos Papadopoulos
Colorado State University
christos@colostate.edu

Alex Feltus
Clemson University
ffeltus@clemson.edu

ABSTRACT

Genomics datasets are currently managed by iRODS, the Integrated Rule-Oriented Data System, which is an open source data management software. iRODS provides several services, including indexing, publishing, integrity, storage, and provenance. In this work, we investigate how NDN can seamlessly integrate into iRODS to provide simplified and improved functionality such as name-based data discovery, replication, retrieval, and computation at the edge. We show how to integrate these NDN based mechanisms as well as caching into iRODs. Once completely functional, we aim to study NDN's benefits in a live, production system with real users.

CCS CONCEPTS

• **Networks** → **Network design principles**; **Naming and addressing**; *Network services*; • **Information systems** → *Information retrieval*;

KEYWORDS

Named Data Networking, NDN, Information Centric Networking, Scientific Data

ACM Reference Format:
Susmit Shannigrahi, Chengyu Fan, Christos Papadopoulos, and Alex Feltus. 2018. NDN-SCI for Managing Large Scale Genomics Data. In *5th ACM Conference on Information-Centric Networking (ICN '18), September 21–23, 2018, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3267955.3269022

1 INTRODUCTION

Just like Climate, and High-energy Particle Physics (HEP), the Genomics community needs to store and distribute large amounts of data. However, data can be distributed around the world, named arbitrarily, and may not always be easy to discover, retrieve and use.

In this work, we closely examine a Genomics workflow and investigate how NDN can facilitate such workflow by augmenting mechanisms such as discovering genomic data repositories around the world, and searching and retrieving data efficiently. For example, the National Center for Biotechnology Information (NCBI) in Maryland, USA, contains 14.4 petabytes of high-throughput DNA sequence data with varying degrees of associated metadata quality. There are several other similar repositories, for example, in Clemson University and Washington State University, that hold and distribute such data. Much of this data is quite useful for potentially thousands of genomics researchers. For example, imagine having the ability of easily searching open source human genome and open access gene expression files distributed around the world.

In this work, we investigate a framework based on Named Data Networking (NDN)[1] that stores and distributes a significant amount of genomics data layered on top of a distributed data grid (iRODS)[2]. Genomics data are already named in an evolution-based, hierarchical manner, which fortunately is easily mappable to an NDN framework. We translate the names and show how we can map them in NDN.

Modern genomic DNA data comes in the form of (A) "static" reference genomes with coordinate-based annotation files, and (B) "dynamic" measurements of genome out (e.g., RNAseq data files that contain RNA molecule snapshot strings in the tens of millions of records). We investigate how NDN helps with locating and retrieving static datasets as well pushing computations to the edge for generating dynamic datasets on-the-fly.

2 IRODS BACKGROUND

iRODS is an intermediate layer between clients and existing infrastructure strewn around various universities and data centers. Users can ask, through command line or an web application, for a set of data for analysis. For example, an user might be interested in comparing two genomes where one dataset is hosted at Clemson and another at Washington State. iRODS tries to provide the uniform access to distributed datasets without having to locate various data repositories, create accounts on them, and installing various software.

iRODS also provides the ability to replicate data across multiple organizations, add metadata to existing datasets, and publish data to a distributed parallel HPC file system. To provide faster data transfers, iRODS also provides the ability to create and utilize "chunked" datasets that are then transferred over the network in parallel. This mechanism not only speeds up data delivery but also provides some safeguard against failures since the user no longer need to download the full dataset but only the failed part(s).

3 PILOT INTEGRATING NDN WITH IRODS

NDN can bring several benefits to Genomics workflows utilizing iRODS as well as the iRODS system itself.

ICN '18, September 21–23, 2018, Boston, MA, USA

ACM ISBN 978-1-4503-5959-7/18/09.
https://doi.org/10.1145/3267955.3269022

```
Genome/
[genus]_[species]{_[infraspecific name]}/[assembly_name]
[genus]_[species]{_[infraspecific name]}-[assembly_name].fa
[genus]_[species]{_[infraspecific name]}-[assembly_name].other_extensions
```

Figure 1: Naming convention for DNA sequence datasets

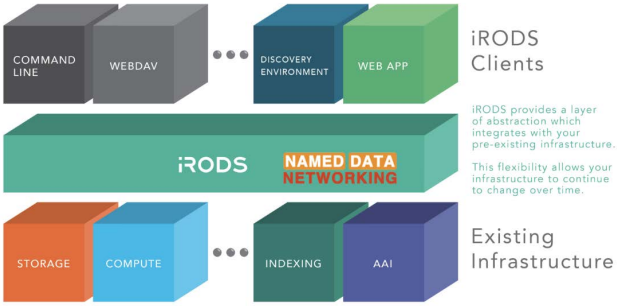

Figure 2: Integrating NDN and iRODS.

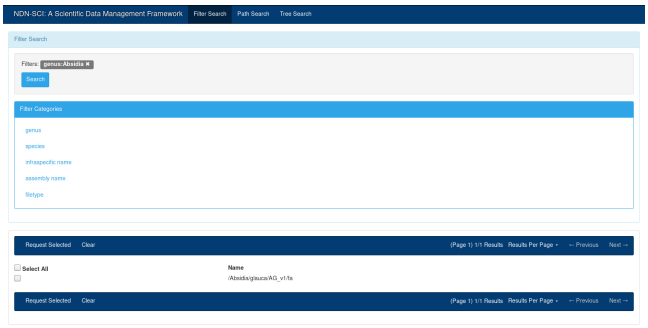

Figure 3: The UI for the NDN-based genomics catalog system.

First, NDN can help with the proper naming of datasets. We found it is possible to annotate DNA sequence datasets using the hierarchical naming convention above (Figure 1). These sequences then can be imported into the NDN-SCI framework along with the actual datasets and metadata. It would be then possible to discover and pull these datasets into genomics workflows that require HPC resources. A user trying to run analysis can provide the names of these data sets that are then brought into computing nodes by the network in a manner that is entirely transparent to the user.

One such actual name may look like: /Absidia/glauca/AG_v1/fa. We could find that some of these names may or may not contain certain components, for example, infraspecific name in the above example. These names are already translated and imported into our NDN-SCI data management framework.

In addition to name discovery, our system is also able to provide additional benefits. NDN makes it easier to federate existing repositories. Repositories are no longer needed to be tracked by their IP addresses and what datasets they host, a replicated and distributed catalog of the names should be sufficient for discovering datasets as well as retrieving them. This makes publishing new datasets and retrieving them much easier. The NDN-SCI prototype already provides this functionality.

An NDN based system can also make parallel data retrieval much easier. For example, separate threads can express Interests for various chunks of the same file, thereby retrieving different parts of a file in parallel. This mechanism not only speeds up the retrieval of the data but in case of failure, allows the user to fetch only the failed chunks.

Finally, NDN can provide specialized functionality such as pushing computations to the edge by integrating computations in the name as a name component. Once the computation reaches the data producer, the producer generates and returns the resulting dataset. These NDN based capabilities can be beneficial for simplifying the iRODS system and adding new functionality to it.

4 CONCLUSIONS

In this work, we show how we can utilize NDN to simplify an existing data management system for Genomic data. We have taken the first step of taking the existing genomics naming convention and translated it into NDN names. We have also created a catalog that publishes the data and allows users to search for it through various ways.

We will next move towards simplifying other components of the iRODS system, for example, NDN-based protocols for discovery, computation, and data retrieval. Many of these aspects are implemented naturally in NDN; for example, retrieving parts of files from various sources and retrieving files in parallel are inherently simple in NDN. We plan to integrate these mechanisms as well as caching into iRODs using NDN and study NDN's benefit in a live, production grade system with a set of invited users.

REFERENCES

[1] Fan, C., Shannigrahi, S., DiBenedetto, S., Olschanowsky, C., Papadopoulos, C., and Newman, H. Managing scientific data with named data networking. In Proceedings of the Fifth International Workshop on Network-Aware Data Management (2015), ACM, p. 1.

[2] Rajasekar, A., Moore, R., Hou, C.-y., Lee, C. A., Marciano, R., de Torcy, A., Wan, M., Schroeder, W., Chen, S.-Y., Gilbert, L., et al. irods primer: integrated rule-oriented data system. Synthesis Lectures on Information Concepts, Retrieval, and Services 2, 1 (2010), 1–143.