Mitigating the dilemma of huge data flow at the University of North Carolina, Greensboro

Jacob Fosso Tande, Ph.D
Research Computing Administrator
University of North Carolina at Greensboro

February 12, 2020

Abstract

Researchers at the University of North Carolina in Greensboro, just as researchers at other research institutions, are faced with the dilemma of huge data management. To mitigate this dilemma, we use a multi-prong approach: by designing and building a dedicated, low-latency and isolated network; provisioning a resilient, persistent and voluminous storage infrastructure; establishing a data management plan and transfer workflow. All of which are aimed at reducing bottlenecks in the operation of experimental instrument and at the same time, optimizing the use of our compute resources.

1 Introduction

The proximity of the University of North Carolina at Greensboro (UNCG) and the North Carolina A&T State University (NC A&T) has enabled very close collaboration leading to the development of research activities and infrastructure. The Joint School of Nanoscience and Nanoengineering (JSNN) is an example of academic collaboration between the two institutions. The JSNN has a vast catalogue of multi-user imaging instruments [1] capable of generating very large amount of data. Also, researchers work collaboratively across disciplines and often have to move data around or to access computational resources remotely. Collaborative and data intensive research activities require the development of efficient workflows, data management plans and the setting up of efficient, effective and reliable network infrastructure.

The Information Technology Services (ITS) at UNCG and at NC A&T are aware of the challenges researchers face as they go about their research activities. Our two institutions are working to upgrade and improve cyberinfrastructures while building efficient workflows to manage data generated from research activities. Above all, we want to create awareness and train our researchers to efficiently use the new infrastructure and services.

2 Method

To mitigate the challenge posed by huge data, we use a multi-prong approach: First we design and build a dedicated, low-latency and isolated network; provision a resilient, persistent and voluminous storage infrastructure; establish data management plans and transfer workflows.

2.1 Science DMZ network

The Gate City Research Network (gcrNet) [2] is being set up to foster collaboration between researchers at UNCG and at NC A&T. The gcrNet is designed for low-latency and is dedicated to moving research data. Once, completed, gcrNet will connect UNCG and NC A&T to the North Carolina Research and Education Network [3] through a dedicated link and the national Internet2 [4] grid. Funded by the national science foundation (NSF) [5], the gcrNet is committed to the open access design [6]. The sustainability and governance documentation, use and performance data, testing and operations protocols will be implemented to fulfill the NSF statutory mission and vision.

2.2 Data management plan

Researchers at our institutions generally generate data that could be put into four categories: observational, experimental, simulation and derived or compiled. Each of these data categories will have specific data management plans that reflects the project requirements.

Given, the number of experiments and the shared number of investigators, it is important for us to set a general minimum standard with documented guidelines on how to meet the standards. We encourage our users to implement best practices as data is generated such as the use of descriptive and informative file names, use of file folders, choosing file formats that will ensure long-term access, tracking different versions of their documents, creation of metadata for every experiment or analysis run and finding helpful tools for analyzing their data. More so, having a plan to transfer knowledge, when and if a member leaves or to fulfill project sponsor requirement for knowledge transfer.

2.3 Data transfer workflow

We design our data transfer processes and develop workflows that reduce bottlenecks in the operation of experimental instrument and at the same time, optimizes the use of our compute resources.

In shared-resource environment, instruments are in high demand and are continually being used. Thus, tying up an acquisition system computer for data processing or analysis is foolhardy, as it will inevitably limit instrument access and data acquisition time. Also, computers dedicated to running basic imaging systems with slow image-acquisition speeds are generally not that powerful; high-end image processing,

particularly for 3D-over-time experiments, demands a system dedicated to image analysis. The ideal solution is the co-existence of two dedicated platforms: one for imaging, and the other for analysis.

A good percentage of our researchers work off-site and while they may have access to data processing computers, interaction with the compute unit may not be evident. A central file server drastically increases the ease of data writing and access for shared-facility users. Data is collected on acquisition system, and immediately transferred to the facility server, which keeps the instrument free for another user to collect data. With such an offline system, users can access their collected data using any number of dedicated and networked analysis systems.

Our cyberinfrastructure planning is designed to meet changing needs. A CyberInfrastructure innovation lab is being established by ITS to enable researchers explore, evaluate, get trainning and professional expertise on emerging technologies. We have adopted a network storage service for on-premise data, only using cloud storage for adapted needs. A data transfer node will be set up to move data from off-campus through the science DMZ network to on-campus servers. Hence, giving our users the impression of locally accessed data.

3 Conclusion

By using a multi-prong approach and collaborating with NC A&T we are able to establish a template for extending our collaboration range. Knowledge acquired will be useful for upscaling as the demand for huge data grows.

References

[1] https://jsnn.ncat.uncg.edu/facilities/equipment/

[2] https://gcrnet.org/

[3] https://www.mcnc.org/about/ncren-footprint

[4] https://www.internet2.edu/products-services/advanced-networking/

[5] https://app.dimensions.ai/details/grant/grant.8463859

[6] https://muninetworks.org/content/open-access