RESEARCH SUPPORT CASE STUDY

Accelerating Genomic Research with Advanced Networking Collaborations

Transformational Science Data Mobility with 16.6X Big Data Transfer Gains from Innovative Network Architectures and Collaborative Optimization

THE PROJECT: OPTIMIZING THE TRANSFER OF GENOMICS DATA

Researchers and technologists at Clemson University (CU) in South Carolina and the National Center for Biotechnology Information (NCBI) in Maryland teamed up to optimize the transfer of genomics data between two endpoints connected by the new Internet2 Advanced Layer 2 Service (AL2S).

The work was supported in part by an NSF Campus Cyberinfrastructure–Network Infrastructure and Engineering (CC-NIE) Program award to CU. Kevin Thompson, program director in NSF's Division of Advanced Cyberinfrastructure, says, *"High performance end-to-end networking is essential for enabling distributed research collaborations such as this one in genomics and across the sciences. CC-NIE addresses the network infrastructure and innovation needs at the campus level that, in combination with Internet2's high-speed network, provide network paths capable of routinely supporting multigigabit data flows, and accelerating scientific discovery."*

In this case, the specific problem was to investigate how to optimize the transfer of a moderate quantity (12 terabytes) of production DNA sequence data between NCBI and CU over a dedicated AL2S connection. Transferring the data as fast as possible into DNA analysis workflows through optimized network facilities and parallel transfer mechanisms will enable researchers to mine data and perform experiments at an unprecedented scale. The ability to create and use dedicated high-speed connections as needed between data sources and analytical computing power is a powerful tool for increasing the throughput of genomics analyses.

> **THIS COLLABORATION AMONG CLEMSON AND NCBI RESEARCHERS AND TECHNOLOGISTS ILLUSTRATES HOW NEW NETWORK SERVICES CAN BE INTEGRATED INTO EXISTING WORKFLOWS TO IMPROVE RESEARCH PRODUCTIVITY AND REDUCE TIME TO DISCOVERY.**

THE IMPORTANCE OF COMMUNITY COLLABORATION AND OPEN NETWORKING

Although advanced technology such as high performance computing (HPC) and high-speed networks played an essential role in the success of this project, the human factor was a key component.

SOLUTION SUMMARY

Genomics research is rapidly becoming one of the leading generators of Big Data for science, with the potential to equal if not surpass the data output of the high-energy physics community. Like physicists, university-based life-science researchers must collaborate with counterparts and access data repositories across the nation and around the globe.

The National Center for Biotechnology Information (NCBI) in Maryland at the National Library of Medicine (NLM) is the largest repository providing access to massive genomic datasets. NCBI hosts almost 25 petabytes of research data and makes it available to a global community of scientists, including researchers who are collaborating to find better ways to diagnose, treat and prevent cancer, as well those working to accelerate agricultural discovery.

NCBI provides genomic sequence data to life scientists like the genomics researchers at Clemson University (CU) in South Carolina. Using their advanced Internet2 network connections and Internet2 services specialized to support research collaborations, NCBI and Clemson have teamed up to transform research workflow in the transfer of genomic Big Data.

COLLABORATORS

- Clemson University (CU)
- University of Utah
- National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM)

SOLUTIONS

- Internet2 Advanced Layer 2 Service (AL2S)
- perfSONAR

FUNDING SOURCES

- National Science Foundation (NSF) Campus Cyberinfrastructure Network Infrastructure and Engineering (CC-NIE) Program
- Broadband Technologies Opportunity Program (BTOP)
- Campus and Regional Network Investments

COMMUNITY RESOURCES

- Internet2 Research Support Center
- Focused Technical Workshops

CU and NCBI have long been leading Internet2 members, and both organizations sent representatives to an Internet2 and ESnet sponsored Focused Technical Workshop, "Network Issues for Life Sciences Research," held in July, 2013. At this workshop, an encounter between a CU genomics researcher and the head of systems at NCBI led to a plan to collaboratively test new network capabilities provided by Internet2 at CU and NCBI. The CU genomics researcher worked directly with CU engineers and faculty to move large datasets. By connecting with NCBI at the workshop, the right people from both organizations–technologists and scientists–were linked, and the project was launched. In addition to its convening role, Internet2 provided the network infrastructure and innovative architecture required to power this significant scientific collaboration.

Collaborators report the importance of face-to-face interactions between researchers from the two organizations. According to Alex Feltus, Associate Professor, Genetics and Biochemistry, Clemson University, "Without the correct technical contacts in both organizations, the lessons learned in the project would have been severely limited. Given the connections made at the Internet2 workshop, we were able to rapidly communicate technical details and interact with hardware for experimentation."

> **THE RIGHT PEOPLE FROM BOTH ORGANIZATIONS–TECHNOLOGISTS AND SCIENTISTS–WERE LINKED, AND THE PROJECT WAS LAUNCHED.**

THE SOLUTION: NETWORK INNOVATION

The Internet2 AL2S is an effective and efficient wide area 100 gigabit Ethernet technology that played a key role in making this project successful. This service allows members to build dedicated Layer 2 circuits (VLANS) between endpoints on the Internet2 Network and beyond.

With the high-speed connection in place, the researchers and technologists explored several methodologies to find the optimal method to transfer 7,535 SRA database files–ranging in size from 20MB to 60GB–from NCBI's FTP servers to the crop genome researcher at CU.

CU and NCBI deployed perfSONAR servers near their respective endpoints to identify and isolate performance issues, and to set a baseline of what speed should be possible across the network itself. perfSONAR is a suite of tools, including tests for bandwidth and latency, used to assure high-performance network paths. A strong community of perfSONAR users and developers work together through an online international forum to fix bugs and develop new features.

THE RESULTS

Using a dedicated 10 Gbps AL2S circuit between CU and the NCBI, and by varying file transfer protocols, storage system characteristics and network software tuning, the experimenters were able to achieve network transfer rates of 7.5 Gbps using the Aspera transfer client. This compares to a rate of 0.5 Gbps seen before the optimizations were applied. After optimization, the team was able to transfer the datasets totaling 12 terabytes across the Internet2 AL2S circuit in 11.6 hours instead of eight days. The overall improvement in throughput reflects the combination of higher speeds with the new AL2S paths, plus the tuning and application of higher performance storage subsystems and file transfer software.

> **THE TEAM WAS ABLE TO TRANSFER THE DATASETS TOTALING 12 TERABYTES ACROSS THE INTERNET2 AL2S CIRCUIT IN 11.6 HOURS INSTEAD OF EIGHT DAYS—16.6X FASTER**

Not only did the genomics scientists obtain their data far more quickly, but this high-speed transfer now makes it reasonable for a genomic researcher to download datasets, process them, re-design the experiment–and repeat.

A SUMMARY OF KEY LESSONS LEARNED

> **ONE OF THE BIGGEST LESSONS LEARNED WAS THE POWER OF COMMUNITY TO REMOVE SIGNIFICANT ROADBLOCKS TO SCIENTIFIC RESEARCH—WHAT THE COLLABORATORS CALLED "THE SOCIAL LEVEL."**

This included the interactions between CU and NCBI team members at the Internet2 workshop and then, having established face-to-face relationships, ongoing work and collaboration.

"It should be emphasized that even with the clear objective and skilled expertise on both sides, hundreds of exchanges occurred during the collaboration," said Alex Feltus, Associate Professor, Genetics and Biochemistry at CU. *"This project was invaluable in learning how to link two organizations via Internet2's AL2S, and we intend to take these lessons to heart as we optimize our research computing connections with other institutions."*

And, in fact, this is happening. Subsequently, the NCBI announced that it is opening up its high-speed storage servers for a second round of CU experiments and a parallel experiment with the University of Utah, also a user of the Internet2 Advanced Layer 2 Service.

The Internet2 community continues to collaborate on advances that enable science, accelerate discovery, and jump-start technologies that add up to huge impacts around the world.