For big data researchers, network and compute capabilities are lynchpin to success

April 13 2015

For many researchers in the life sciences, Big Data is not just a buzz word—it's the daily reality for carrying out their work in areas like genomics, which is expected to equal if not surpass the data output of the particle physics community. For many scientists, in order to keep pace with the data deluge, the often less glamorous side of big data research—the network, computing and cloud architecture required to support their work—must be at the forefront of their minds. At the Internet2 Global Summit meeting taking place April 26-30 in Washington, D.C., researchers like Genetics and Biochemistry Associate Professor Alex Feltus of Clemson University will come together with network engineers, chief information officers, and other technology leaders in the research and education community to discuss ways they can collaborate to advance research capabilities in IT infrastructure and applications.

Reducing the bottleneck in data transfer

At Clemson, Feltus uses genomics research to develop new agricultural crop varieties that address population pressure, bioenergy, food security and climate change. He will be at the Global Summit meeting to present how his team and collaborators at The National Center for Biotechnology Information (NCBI) in Maryland at the National Library of Medicine (NLM) are leveraging the advanced Internet2 infrastructure, including its Advanced Layer 2 Service high-speed connections and perfSONAR network monitoring, to substantially accelerate genomic big

data transfers and transform researcher collaboration.

As DNA data sets get bigger and bigger, Feltus sees a need to change the way data is stored and transferred. "I hope that as DNA sequencing becomes cheaper it will make more sense to regenerate the data than to store it long term," he says. "Of course we need bigger boxes, but we also need faster ways to put stuff into them. There is a serious data transfer bottleneck at the network-hard-drive interface. Thus, we need faster, reasonably-priced storage that can keep up with the advanced networks such as the Internet2 Network."

Feltus' supercomputing capabilities at Clemson are best-in-class thanks to the university's high-performance computing (HPC) resource, Palmetto. But, Feltus says, when it comes to collaborating with other research teams across the country, one's compute power is only as good as the connection it's hooked up to—which is where Internet2 comes in. "With the Internet2 Network, I can quickly download more data to Palmetto from public repositories like the National Center for Biotechnology Information and scale up my crop genomics experiments," he says. "Besides scale up, low latency networks like Internet2 have opened up new possibilities for my research. In collaboration with Melissa Smith's group at Clemson for example, we were able to run GPU-enabled visualization algorithms on the Palmetto cluster at Clemson and beam results to our Supercomputing Conference booth in New Orleans for near real-time visualization of gene interaction networks. You can process data on the fastest nodes in the world, but it's pointless for real-time applications if the supercomputer is hooked up to a slow pipe."

These advanced technology capabilities allow Feltus and his team to focus on their work, rather than attempting to build their own network and computing infrastructure to enable that work.

Enhancing the model for big data computing

Across the country, Arizona State University (ASU) is also looking to do just that for its researchers in the College of Life Sciences' Adaptive Complex Systems Science program, which studies highly interactive and dynamic systems that change over individual and evolutionary time scales, such as epidemics, obesity and cancer.

ASU's response to the Big Data challenge has been to develop what they call the Next Generation Cyber Capability (NGCC)—a "First Generation Data Science Research Instrument" that they liken to instruments such as the Hubble Space Telescope or the Large Hadron Collider. NGCC is already proving useful in the growing area of personalized medicine—specifically tailored disease treatment that takes into account an individual patient's own molecular constitution and that of the disease.

Also called precision medicine, this field of study comes with the challenge of managing and analyzing big data related to both genomic information and associated imaging data for each individual. ASU's Director of Operations, Research Computing and Senior HPC Architect Jay Etchings, who will also present at Internet2's Global Summit meeting this month, says the NGCC is poised to address precision medicine challenges and beyond.

"The potential for patient-focused, precision medicine care roadmaps crystalizes with personalized medicine if we can simply sort out the data," says Etchings. "Additionally, genomic data is only one of the 'varieties' of large volume Big Data. Diverse clinical observations and patient-reported outcomes also must be integrated and interpreted."

Etchings says enabling this integration of multidimensional molecular and clinical data is where NGCC's new model of computing becomes

essential.

"The traditional model of accessing data at the node still has its place in the greater datacenter/cloud infrastructure arena," says Etchings. "However, discounting many core, software-defined and virtual instances would be in error."

The NGCC's new model marries several essential capabilities for big data research. The first is physical capacity, which means being connected to the ultrahigh bandwidth Internet2 Network, having large-scale storage—on the order of 2 Petabytes or more—and integrating multiple types of computation, including utility computing, traditional HPC and new big data computing. The second element to NGCC is advanced logical capabilities such as software-defined storage and networking, metadata processing and semantics. The final element is the human factor—transdisciplinary teams of researchers, network engineers and computing professionals working together on the system as a whole.

It's this last capability that both Etchings and Feltus note is essential to the success of their work in Big Data.

"A key aspect is the side of cyberinfrastructure that can't be coded: personal relationships," says Feltus. "Recently, my collaboration with network and storage researchers and engineers has opened my eyes to innovative possibilities that will impact my research via the human network."

That "human network" will be made possible at the Global Summit meeting, which will provide yet another opportunity for researchers like Feltus and architects like Etchings to check in, collaborate and potentially develop even more meaningful interactions that just may lead to the next big breakthrough in Big Data science.

More information: Presentations:

- "Innovations in High-Volume Life Sciences Research," Jay Etchings, Monday, April 27, 3 p.m., Room: Mount Vernon A ,Renaissance Washington DC Downtown Hotel
- "From CC-NIE/IIE/DNI to Building a Cohesive Platform for Collaboration Over Advanced Cyberinfrastructure," Alex Feltus et al., Tuesday, April 28, 8:45 a.m., Room: Mount Vernon A, Renaissance Washington DC Downtown Hotel

Provided by Internet2