New computing model could lead to quicker advancements in medical research

November 4 2013

In April of 2010, the National Science Foundation teamed with Microsoft on a collaborative cloud computing agreement. One year later they decided to fund 13 research projects. Wu Feng was selected to lead one of these teams. His target was to develop an on-demand, cloud-computing model to keep up with the data deluge in the DNA sequencing space. Credit: Virginia Tech

With the promise of personalized and customized medicine, one

extremely important tool for its success is the knowledge of a person's unique genetic profile.

This personalized knowledge of one's genetic profile has been facilitated by the advent of next-generation sequencing (NGS), where sequencing a genome, like the human genome, has gone from costing $95,000,000 to a mere $5,700. So, now the research problem is no longer how to collect this information, but how to compute and analyze it.

"Overall, DNA sequencers in the life sciences are able to generate a terabyte—or one trillion bytes—of data a minute. This accumulation means the size of DNA sequence databases will increase 10-fold every 18 months," said Wu Feng of the Department of Computer Science in the College of Engineering at Virginia Tech.

"In contrast, Moore's Law (named after Intel co-founder Gordon E. Moore) implies that a processor's capability to compute on such 'BIG DATA' increases by only two-fold every 24 months. Clearly, the rate at which data is being generated is far outstripping a processor's capability to compute on it. Hence the need exists for accessible large-scale computing with multiple processors … though the rate at which the number of processors needs to increase is doing so at an exponential rate," Feng added.

For the past two years, Feng has led a research team that has now created a new generation of efficient data management and analysis software for large-scale, data-intensive scientific applications in the cloud. Cloud computing is a term coined by computing geeks that in general describes a large number of connected computers located all over the world that can simultaneously run a program at a large scale. Feng announced his work in October at the O'Reilly Strata Conference + Hadoop World in New York City.

By background to Feng's announcement, one needs to go back more than three years. In April of 2010, the National Science Foundation teamed with Microsoft on a collaborative cloud computing agreement. One year later, they decided to fund 13 research projects to help researchers quickly integrate cloud technology into their research.

Feng was selected to lead one of these teams. His target was to develop an on-demand, cloud-computing model, using the Microsoft Azure cloud. It then evolved naturally to make use of the Microsoft's Hadoop-based Azure HDInsight Service. "Our goal was to keep up with the data deluge in the DNA sequencing space. Our result is that we are now analyzing data faster, and we are also analyzing it more intelligently," Feng said.

With this analysis, and the ability of researchers from all over the globe to see the same sets of data, collaborative work is facilitated on a 24/7 global perspective. "This cooperative cloud computing solution allows life scientists and their institutions easy sharing of public data sets and helps facilitate large-scale collaborative research," Feng added.

Think of the advantages of oncologists from Sloan Kettering to the German Cancer Research Center would have by maintaining simultaneous and instantaneous access to each other's data.

Specifically, Feng and his team, Nabeel Mohamed, a master's student from Chennai, Tamilnadu, India and Heshan Lin, a research scientist in Virginia Tech's Department of Computer Science, developed two software-based research artifacts: SeqInCloud and CloudFlow. They are members of the Synergy Lab , directed by Feng.

The first, an abbreviation for the words "sequencing in the clouds", combined with the Microsoft cloud computing platform and infrastructure, provides a portable cloud solution for next-generation

sequence analysis. This resource optimizes data management, such as data partitioning and data transfer, to deliver better performance and resource use of cloud resources.

The second artifact, CloudFlow, is his team's scaffolding for managing workflows, such as SeqInCloud. A researcher can install this software to "allow the construction of pipelines that simultaneously use the client and the cloud resources for running the pipeline and automating data transfers," Feng said.

"If this DNA data and associated resources are not shared, then life scientists and their institutions need to find the millions of dollars to establish and/or maintain their own supercomputing centers," Feng added.

Feng knows about high-performance computing. In 2011, he was the main architect of a supercomputer called HokieSpeed.

That year, HokieSpeed settled in at No. 96 on the Top500 List, the industry-standard ranking of the world's 500 fastest supercomputers. Its fame, however, came because of the machine's energy efficiency, recorded as the highest-ranked commodity supercomputer in the United States in 2011 on the Green500 List, a compilation of supercomputers that excel at using less energy to do more.

Economics was also key in Feng's supercomputing success. HokieSpeed was built for $1.4 million, a small fraction—one-tenth of a percent of the cost—of the Top500's No. 1 supercomputer at the time, the K Computer from Japan. The majority of funding for HokieSpeed came from a $2 million National Science Foundation Major Research Instrumentation grant.

Feng has also been working in the biotechnology arena for quite some

time. One of his key awards was the NVIDIA Foundation's first worldwide research award for computing the cure for cancer. This grant, also awarded in 2011, enabled Feng, the principal investigator, and his colleagues to create a client-based framework for faster genome analysis to make it easier for genomics researchers to identify mutations that are relevant to cancer. Likewise, the more general SeqInCloud and CloudFlow artifacts seek to achieve the same type of advances and more, but via a cloud-based framework.

More recently, he is a member of a team that secured a $2 million grant from the National Science Foundation and the National Institutes of Health to develop core techniques that would enable researchers to innovatively leverage high-performance computing to analyze the data deluge of high-throughput DNA sequencing, also known as next-generation sequencing.

Provided by Virginia Tech