HPC wire

October 21, 2010

Supercomputing Meets Social Media

In supercomputing these days, it's usually the big science applications (astrophysics, climate simulations, earthquake predictions and so on) that seem to garner the most attention. But a new area is quickly emerging onto the HPC scene under the general category of informatics or data-intensive computing. To be sure, informatics is not new at all, but its significance to the HPC realm is growing, mainly due to emerging application areas like cybersecurity, bioinformatics, and social networking.

The rise of social media, in particular, is injecting enormous amounts of data into the global information stream. Making sense of it with conventional computers and software is nearly impossible. With that in mind, a **story** in MIT Technology Review about using a supercomputer to analyze Twitter data caught my attention. In this case, the supercomputer was a Cray XMT machine operated by the DOE at Pacific Northwest National Lab (PNNL) as part of their **CASS-MT** infrastructure.

The application software used to drive this analysis was GraphCT, developed by researchers at Georgia Tech in collaboration with the PNNL folks. GraphCT is short for Graph Characterization Toolkit, and is designed to analyze really massive graph structures, like for example, the type of data that makes up social networks such as Twitter.

For those of you who have been hiding under a rock for the last few years, Twitter is a social media site for exchanging 140-character microblogs, aka tweets. As of April 2010, there were over 105 million registered users, generating an average of 55 million tweets a day. The purpose of Twitter is, of course... well, nobody knows for sure. But it does represent an amazing snapshot of what is capturing the attention of Web-connected humans on any given day. If only one could make sense of it.

Counting tweets or even searching them is a pretty simple task for a computer, but sifting out the Twitter leaders from the followers and figuring out the access patterns is a lot trickier. That's where GraphCT and Cray supercomputing comes in.

GraphCT is able to map the Twitter network data to a graph, and make use of certain metrics to assign importance to the user interactions. It measures something called "betweenness centrality," to rank the significance of tweeters.

Because of the size of the Twitter data and the highly multithreaded nature of the GraphCT software, the researchers couldn't rely on the vanilla Web servers that make up the Internet itself, or even conventional HPC computing gear. Fine-grained parallelism plus sparse memory access patterns necessitated a large-scale, global address space machine, built to tolerate high memory latency.

The Cray XMT, a proprietary SMP-type supercomputer is such a machine, and is in fact specifically designed for this application profile. I suspect the reason you don't hear more about the XMT is because most of them are probably deployed at those top secret three-letter government agencies, where data mining and analysis are job one.

The XMT at PNNL is a 128-processor system with 1 terabyte of memory. The distinguishing characteristic of this architecture is that each custom "Threadstorm" processor is capable of managing up to 128 threads simultaneously. Tolerance for high memory latencies is supported by efficient management of thread context at the hardware level.

The system's 1 TB of global RAM is enough to hold more than 4 billion vertices and 34 billion edges of a graph. To put that in perspective, one of the Twitter datasets from September 2009 was encapsulated in 735 thousand vertices and 1 million edges, requiring only about 30 MB of memory. Applying the GraphCT analysis, the data required less than 10 seconds to process. The researchers estimated that a much larger Twitter dataset of 61.6 million vertices and 1.47 billion edges would require only 105 minutes.

When the Georgia Tech and PNNL researchers ran the numbers, they found that relatively few Twitter accounts were responsible for a disproportionate amount of the traffic, at least on the particular datasets they analyzed. The largest dataset was made up of all public tweets from September 20th to 25th in 2009, containing the hashtag #atlflood (to capture tweets about the Atlanta flood event). In this case, at least, the most influential tweets originated with a few major media and government outlets.

We're likely to be hearing more about the graph applications in HPC in the near future. Data sets and data streams are outpacing the capabilities of conventional computers, and demand for digesting all these random bytes is building rapidly. Since the optimal architectures for this scale of data-intensive processing is apt to be quite different than that of conventional HPC platforms (which tend to be optimized for compute-intensive science codes), this could spur a lot more diversity in supercomputer designs.

To that end, a new group called the **Graph 500** has developed a benchmark aimed at this category of applications, and intends to maintain a list of the top 500 most performant graph-capable systems. The first Graph 500 list is scheduled to be released at the upcoming Supercomputing Conference (SC10) in New Orleans next month.

In the meantime, if you're interested in giving GraphCT a whirl, a pre-1.0 release of the software can be downloaded for free from the **Georgia Tech website**. You'll just need a spare Cray XMT or POSIX-compliant machine to run it on.

Find this article at:
http://www.hpcwire.com/blogs/Supercomputing-Meets-Social-Media-105493293.html

□ Check the box to include the list of links referenced in the article.