Progress Report: CCC's Support for Data-Intensive Computing

Randal E. Bryant, Carnegie Mellon University and Thomas T. Kwan, Yahoo! Research

Nov. 3, 2008

The CCC and Yahoo! cosponsored two "community building" events for a *Big Data Computing Study Group*, an organization created to explore and enable opportunities for the research and application of high-performance computing over very large data sets. The CCC funding supported two meetings: the first ever Hadoop Summit on March 25 in Santa Clara, followed by the first Data-Intensive Computing Symposium on March 26 at Yahoo!'s Sunnyvale headquarters.

A report of these meetings was published in the May 2008 edition of *Computing Research News*, Vol. 20/No. 3. In addition, both slides and videos of the presentations are available at http://research.yahoo.com/node/2104. Here are some highlights from the actual meetings:

The Hadoop Summit brought together leaders from the Hadoop developer and user community for the first time. (Apache Hadoop, an open source distributed computing project of the Apache Software Foundation, is a distributed file system and parallel execution environment that enables its users to process massive amounts of data.) Originally planned for an audience of 100, the venue was changed to accommodate the enthusiastic response from the open source community. Close to 350 people attended the summit to listen to the talks. Participants included members of the Hadoop development community, researchers at well-established companies such as IBM and Microsoft, and developers from a number of startup companies.

About 100 researchers from academia, industry, and government laboratories and agencies attended the Data-Intensive Computing Symposium at Yahoo!'s Sunnyvale headquarters. The symposium brought together experts in system design, programming, parallel algorithms, data management, scientific applications, and information-based applications to better understand existing capabilities in the development and application of large-scale computing systems, and to explore future opportunities.

Follow-On Activities

Since that meeting, a number of important activities and initiatives in data-intensive computing have arisen. Some can be traced directly to these two workshops, while others were helped along by these meetings. Since the meeting, the term "cloud computing" has become widely used for describing large-scale cluster computing facilities, as can be seen by these follow-on projects.

- In July 2008, the National Science Foundation (NSF) initiated a cross-cutting program for data-intensive computing (NSF 08-578), with up to 30 awards and $10 million in funding for data-intensive computing.

- The NSF and Google, along with the University of Washington and Haverford College sponsored the 2008 NSF Data-Intensive Scalable Computing in Education Workshop, held at the University of Washington July 16—18, 2008. This workshop brought together educators from a number of US institutions to get training in the programming tools (mainly Hadoop) available on the Google/IBM facility.

- Based on conversations between people from Yahoo!, Hewlett-Packard (HP), and Intel that started at the Data-Intensive Computing Symposium, these companies have formed a partnership to create an international cloud computing testbed. As announced in July, 2008, they will team up with the University of Illinois at Urbana-Champaign (which is receiving NSF funding), the Infocomm Development Authority of Singapore, and the Karlsruhe Institute of Technology. These organizations are setting up six data centers and will explore the issues behind supporting large-scale, geographically distributed data-intensive computing.

- Engineers and researchers at Intel and Yahoo! have initiated the open source project Tashi, a cluster management system for cloud computing on big data. Tashi will serve a critical need for managing the processing resources in a large-scale facility operating with multiple users.

- Ian Foster and Robert Grossman organized a workshop on cloud computing and applications, CCA-08, held in Chicago in October, 2008, with an organizing committee that included many of the participants in the Data Intensive Computing Symposium.

Concluding Thoughts

The timing of the March workshops was ideal. Large-scale, data-intensive computing has been underway at large companies, especially Google, Yahoo!, and Microsoft, for a number of years. Programs by Google/IBM, and HP/Intel/Yahoo! to provide access for academics had just begun. The National Science Foundation could see the importance of data-intensive computing both as a research area and for its potential impact on all areas of scientific research. The March workshops provided a chance for academics, researchers, engineers, and government representatives to share their ideas and visions. It fostered connections between individuals that have led to additional linkages and activities. We anticipate many more activities in data-intensive computing will arise as a result of CCC-sponsored activities.