San Diego Supercomputer Center begins cloud computing research using the Google-IBM CluE cluster

February 18 2009

Researchers from the San Diego Supercomputer Center (SDSC) at the University of California, San Diego, have been awarded a two-year, $450,000 grant from the National Science Foundation to explore new ways for academic researchers to manage extremely large data sets hosted on massive, Internet-based commercial computer clusters, or what have become known as computing "clouds."

The NSF award focuses on the Cluster Exploratory (CluE), a distributed, large-scale computing resource formed in late 2007 between Google and IBM. The NSF joined the Google-IBM partnership early last year, hailing the CluE initiative as a partnership between private enterprise and the federal science agency to expand access to this research infrastructure to academic institutions across the nation. Last April, the NSF issued a solicitation for research projects aimed at developing software to make CluE a researcher-friendly resource to analyze and manage extremely large amounts of data.

Specifically, SDSC researchers will explore the use of compute clouds to dynamically provision and manage large-scale scientific datasets. This is in contrast to the current approach using a traditional parallel relational database management system (RDBMS) architecture, which is more structured but also more static. The SDSC team will investigate the feasibility of the cloud computing approach versus known conventional approaches, while evaluating the trade-offs, advantages, and

disadvantages.

"The CluE system provides access to a cloud computing environment characterized by relatively vast amounts of computational and storage resources," said SDSC Distinguished Scientist Chaitan Baru, who is heading the SDSC research project, called 'Performance Evaluation of On-Demand Provisioning of Data-Intensive Applications.' "This creates opportunities to rethink some of our strategies and ask ourselves some key questions: Could we use more dynamic strategies for resource allocation? Can this result in better overall performance for the user?"

Cloud computing - defined by the ACM Computer Communication Review as a large pool of easily usable and accessible virtualized resources that can be dynamically reconfigured to adjust to a variable load and operated on a pay-per-use model - has been generating considerable attention throughout the high-performance computing community, in both the commercial and academic sectors. This new model is seen as a possible way for researchers to move from processing and managing their own data sets locally, to relying on large, off-site, commercially managed data clusters.

Amazon.com, for example, although primarily an e-commerce retailer, has made pay-as-you-go, on-demand computing and storage available via its "Elastic Compute Cloud" or EC2 platform. Introduced in mid-2006, EC2 is now being used by both startup companies and established businesses as a 'virtual' computing resource.

Like many other supercomputer scientists, Baru is concerned that the ever-increasing volume of scientific data is beginning to overwhelm current approaches to data management.

"The broader impact of this research will be to reassess how scientific data archives are implemented, and how data sets are hosted and served

to the scientific community at large, using on-demand and dynamic approaches for provisioning data sets as opposed to the current static approach," said Baru. "This project has the potential to offer scientific researchers compute clouds as a complement to conventional supercomputing architectures used today, while creating new tools and techniques for commercial cloud computing."

SDSC's research will focus on using its already widely accepted GEON LiDAR Workflow (GLW) application, which is part of the Center's GEON Project, an open, collaborative project funded by the NSF's Information Technology Research (ITR) and Geoinformatics programs to develop cyberinfrastructure for the integration of three- and four-dimensional earth science data. LiDAR data have broad applicability in the areas of earth sciences, hydrology, ecology, environmental sciences, and hazards. The GLW application allows users to subset remote sensing data stored as "point cloud" data sets, process it using different algorithms, and visualize the output.

SDSC researchers will use the Google-IBM CluE cluster in combination with the Apache Hadoop programming environment, an open-source volunteer project developed under the Apache Software Foundation, a not-for-profit enterprise dedicated to open, collaborative software development projects by supplying hardware, communication, and business infrastructure. During the next 24 months, researchers will evaluate hybrid approaches that dynamically deploy a database instance for some subsets of the data, based on user directives or workload analysis, while the rest of the data are served using Hadoop.

"The resource-rich CluE environment represents the type of systems that are becoming available to real-world applications, as opposed to the resource-constrained environments that are currently common in most production systems for large scientific data management, especially in academia," said Baru. "The Google-IBM CluE cluster is a natural

environment for implementing GLW, which is highly data-parallel in nature. Our plan is to evaluate a series of implementations using the database system, as well as the distributed file system and MapReduce implementations provided by Hadoop."

In addition, SDSC researchers believe that the CluE- based or similar cluster-based implementations can be exposed to user communities via a Services Oriented Architecture (SOA) for access via other scientific workflow environments, visualization tools, and portals, such as the GEON Portal and GEON's new Open Topography portal, which now contains GEON's LiDAR data access and processing capabilities.

"This in turn will bring the benefits of massively-scaled computing resources to a large community of users," noted Baru.

Source: University of California - San Diego