NSF Cloud Workshop 2011

Title: CloudStor: Data Intensive Computing in the Cloud
Name: Sriram Krishnan, Chaitanya Baru
Affiliation: San Diego Supercomputer Center
Email: sriram@sdsc.edu

Topics of interest: Cloud Architectures & Systems, Programming Models for the Cloud, Cloud Test-beds

Research Summary

The CloudStor group at the San Diego Supercomputer Center (http://acid.sdsc.edu/projects/cloud) is exploring new strategies and technologies for data-intensive cloud computing. We are investigating application profiles that benefit from this paradigm; and developing corresponding applications. We are interested in evaluating and comparing the performance and price/performance of alternative, dynamic strategies for provisioning data intensive applications using cloud-computing paradigms such as MapReduce. We are currently investigating applications involving remote-sensed LiDAR data, in conjunction with the NSF-funded OpenTopography facility. These applications allows users to (i) subset remote sensing data (stored as "point cloud" data sets), (ii) process the data subsets in multiple steps, using various algorithms, and (iii) visualize the output. Cloud platforms with thousands of processors and access to hundreds of terabytes of storage provide a natural environment for implementing OpenTopography processing routines, which are highly data-parallel in nature. Our studies will contribute towards the understanding of performance tradeoffs and feasibility in dynamic provisioning strategies for serving large scientific data sets. Results from our experiments were recently published at the 2nd IEEE International Conference on Cloud Computing Technology and Science. In addition to this work, in collaboration with the Moores UCSD Cancer Center, we are investigating tools and techniques for large-scale next-generation sequencing. In particular, we are comparing the price/performance of shared-memory approaches, versus shared-nothing approaches provided by frameworks such as Apache Hadoop.

Our work is funded by the National Science Foundation's Cluster Exploratory (CluE) award #0844530, and an SDSC Triton Resource Opportunities (TRO) award. For our experiments, we have access to the Google/IBM CluE cluster, Triton and Gordon/Dash resources at SDSC, the FutureGrid test bed, and the Amazon Web Services (AWS) cloud infrastructure.

Abstract

Cloud computing provides a model for convenient on-demand access to a shared pool of resources, that can be rapidly and elastically provisioned with minimal management effort. Cloud computing has proven to be very successful in the industry, and increasingly popular in academic and scientific research as well. Researchers in the fields of bioinformatics, geosciences, healthcare, etc. are looking at cloud computing as a viable solution for their specific needs. We propose the notion of "*seeded clouds*" as an effective way by which to facilitate production of new results and generation of new collaborations in data intensive computing. Currently, there is a gap between "domain" scientists with large datasets who are faced with the data deluge and have the need for effective solutions, on the one hand, and computer scientists who are keen to work with and apply their techniques on "real world" data at unprecedented scale to uncover and study, on the other. The creation of seeded clouds as a well-defined strategy would greatly enhance innovation in this area. Seeded clouds will serve to close the gap between the community of researchers on one hand who have to deal with large and/or complex data sets and need innovative strategies for managing, processing, integrating, and mining these data, and the community of researchers who are interested in working with "real" datasets at scale, to try various processing strategies, data mining algorithms, and data integration techniques. There is a need for standardized access to real-world data at scale to foster computer science innovation with such data. Seeded clouds would essentially "provision" a large data set with the necessary computational and storage capability to allow in situ experiments to be performed with the entire dataset or subsets thereof. Just as supercomputer centers provide access to hardware (OS) platforms, seeded cloud centers would provide access to a data intensive platform that is pre-loaded with data, and expand the reach of such data to a broad research community, thereby accelerating discovery with these data and development of new algorithms for data mining and integration. Our group recently sponsored an NCI/NSF-funded workshop on 'Health Cyberinfrastructure: A Seeded Cloud Approach' (http://www.healthcyberinfrastructure.org/), where there was group consensus on this approach.

Another challenge is programming scientific applications on the cloud. The MapReduce programming model is currently the *de facto* standard for programming applications on cloud resources. It is a real problem to port existing codes *as-is* to the cloud, and observe acceptable performance. Hence, many groups in various domains are re-implementing scientific codes to conform to the MapReduce model. Although many real-world algorithms do map well to the MapReduce model, there is a large class of applications that do not conform. New programming models and infrastructures are being developed to address this issue – e.g. Google's Pregel infrastructure is optimized to mine relationships from graphs. Microsoft's Azure cloud provides a queue and role-based model for programming cloud applications. We propose to investigate the use of various programming models for a class of scientific applications (with our collaborators) in the area of the geosciences (for the NSF-funded OpenTopography facility) and bioinformatics (next generation sequencing for the Moores UCSD Cancer Center). We also propose to use the Software as a Service (SaaS) deployment model to accelerate the development of scientific pipelines on cloud computing resources.