An Investigation on Applications of Cloud Computing in Scientific Computing

Huiying Chen[a], Feng Wang[b,*] and Hui Deng[c]

Yunnan Computer Technology Application Key Lab
Kunming University of Science and Technology
Kunming, China, 650500
{[a]chenhuiying, [b]wf, [c]dh}@ cnlab.net

Keywords: Cloud computing, scientific computing, computational science, virtualization

Abstract. In recent years, "cloud computing" is one of the most popular terminologies in computer society. Many IT providers and enterprises are eager to move closer to "cloud" in order to solve the current bottlenecks encountered in various fields such as scientific computing. In common, scientific computing needs a huge number of computers available to perform large-scale experiments. These should not only require enormous money to construct, but also need a lot of follow-up time and manpower to maintain and operate. Cloud computing provides a new computing pattern for scientific computing, which could make scientists dynamically access to computing infrastructure on demand, such as computing, storage resources and applications. The emergence of this new computing pattern has brought new opportunities and challenges for computational science. This thesis starts with the introduction of cloud computing, and then probes into the application prospect and results of cloud computing in scientific computing.

Introduction

Scientific computing is a kind of numerical calculation, and it is mathematical calculation in computer processing of scientific research and engineering[1]. Scientific computing involves the construction of mathematical models and numerical solution techniques to solve scientific, social scientific and engineering problems. These models often require a huge number of computing resources to perform large scale experiments and meet with a great quantity of complex mathematical calculations. Along with the rapid development of computer hardware technology, these needs have been initially addressed with dedicated high-performance computing (HPC) infrastructures such as clusters or supercomputers. However, the cost of construction and maintenance is tremendous. With the advent of Grid computing, new opportunities became available to scientists[2]. The use of computing Grids in scientific computing has become so successful that many international projects led to the establishment of world-wide infrastructures available for computational science.

Even though the widespread use of Grid technologies in scientific computing, the procedure cannot always carry through as you wish. There exist some problems, especially technical hurdles. In most cases scientific Grids feature a pre-packaged environment in which applications will be executed, sometimes specific tools and APIs have to be used and there could be limitations on the hosting operating systems or on the services offered by the runtime environment. In practice a limited set of options are available for scientists, and they could not be elastic enough to cover their needs. This problem could constitute a fundamental obstacle for scientific computing.

* Corresponding Author: Feng Wang, wf@cnlab.net

Cloud computing can solve these problems. By means of virtualization technologies, cloud computing offers to end users a variety of services covering the entire computing stack, from the hardware to the application level, by charging them on a pay per use basis. By using cloud based technologies scientists can have easy access to large distributed infrastructures and completely customize their execution environment, thus providing the perfect setup for their experiments. There are enough options available to scientists to cover any specific need for their research. In the meantime, leasing infrastructure from the third party can largely reduce costs for scientific research organizations.

The rest of the paper is organized as follows: first, we provide an overview of cloud computing by summarizing the key technologies applied to cloud computing. Then, we will introduce several solutions that are pertinent to scientific applications and provide a rough discussion of their features by highlighting how it can support computational science. Final thoughts and key observations about the future directions of cloud computing, as a valid support for scientific computing, are discussed at the end.

The key technologies of cloud computing

Cloud computing is a new computing mode with data-centric data-intensive supercomputing. So inevitable it has its own unique technology in data storage, data management, programming mode, concurrency control and so on.

Data storage technology. To ensure the high reliability, high availability and economy, cloud computing uses distributed data storage and redundant storage mechanism to store information, i.e. same multiple copies. In addition, cloud systems often need to satisfy multiple users' request and deal with multiple parallel applications, so data storage technology must have high throughput rate and high transfer rate. At present, GFS (Google File System)[3] and HDFS (Hadoop Distributed File System) are the widely used date storage technology. Many IT providers, including Yahoo and Intel, use HDFS in their cloud schemes[4-5].

Virtualization technology. Virtualization[6] is a technology to realize the partial or complete machine simulation and time sharing by cutting the rock-bottom computing resources into multiple or merging them into one running environment. After using virtualization technology, the movement way of system is logical, so virtualization technology could shield the complexity of physical movement, and then the system shows to users is the simple logic state of movement. As shown in Figure 1.

Fig.1 The operating of virtualization technology

Data management technology. The management of cloud data is read optimization. And in the database we usually adopt column storage methods to improve reading efficiency, so we can store cloud data after dividing tables by column. At present, there are some relatively mature cloud data management systems, such as BigTable (Google)[7], HBase (Hadoop), Sector/Sphere, etc.

Parallel programming model. Programming model must be simple for unprofessional users in cloud computing system. Actually the complex data calculation and task scheduling in the background of system are executed in parallel, only transparent to users and programmers. So far Google's Map Reduce[8] is the main programming model for cloud computing systems.

The application prospect of cloud computing in scientific computing

Since Google has proposed the concept of cloud computing[9], the major business organizations have invested in constructing cloud platform. Now, let us explore the specific application of the cloud computing for scientific computing, and you'll have a preliminary understanding from table 1.

Table 1. Cloud Computing Application Fields

Application Fields	Application Scenarios	Application Fields	Application Scenarios
Scientific Research	Meteorological Data Processing	Graphic Image Processing	Animation Material Storage and Analysis
	Seismic Survey		
	Ocean Information Monitoring		High Simulation Animation
	Astronomical Information Processing		Mass Image Retrieval
Medical Science	DNA Information Analysis	Internet	E-mail Services
	Mass Patient Cases Storage and Analysis		Online Real-time Translation
	Medical Image Processing		Network Information Retrieval
Internet Security	Virus Database Storage and Matching	Social Security	Citizen Information File Storage
			Criminal Record Storage
	Spam Shielding		Suspect Fingerprint/DNA Matching

All scenarios listed in the table 1 have features like large data storage or finding information in large data sets, and actually they are the instances of scientific computing. In addition, cloud computing can store and manage large data sets, and also achieve massive data search. So cloud computing can be used in scientific computation like grid computing, or even better.

At present, the use of cloud computing in computational science is still limited, but the first steps towards this goal have been already done. In 2009, the Department of Energy (DOE) National Laboratories started exploring the use of cloud services for scientific computing. On April 2009, Yahoo Inc. announced that it has extended its partnership with the major top universities in United States of America to advance cloud computing research and applications to computational science and engineering[10]. Next we'll describe the results of cloud computing so far achieved in the scientific computing.

Science Clouds. Science Clouds[11-12], one of the first cloud-based infrastructures for computational science, was initiated by the University of Chicago (UC) and the University of Florida (UFL) with two objectives: Make it easy for scientific and educational projects to experiment with EC2-style cloud computing, and Better understand the potential and challenges that cloud computing poses for these communities and what can be done to overcome them. The first cloud, at the University of Chicago, became available on March 3, 2008, and was named "nimbus"[13]. The University of Florida cloud[14], made available on May 13, 2008.

The "nimbus cloud model" has proved popular among resource providers. In fact, the GridFTP and container scalability tests at UC proved so popular that two new private clouds were configured on newly purchased infrastructure to support this mode of usage for internal UC projects.

AzureBlast. AzureBlast[15] is a parallel BLAST engine running on the Windows Azure that can marshal the compute power of thousands of Azure instances. BLAST(Basic Local Alignment Search Tool)[16] is one of the most widely used bioinformatics algorithms in life science applications. It is not only relevant to a large number of research communities; it represents a large-number of science applications. These applications are usually computation intensive, data intensive and can be parallelized by a simple coarse-grained data-parallel computational pattern. While high performance is often considered desirable, scalability and reliability are usually more important for this class of applications. The experience presented in literature [15] demonstrates that Windows Azure can support the BLAST and associated class of applications very well due to its scalable and fault-tolerant computation and storage services. Moreover the pay-as-you-go model, together with elasticity scalability of cloud computing greatly facilitates the democratization of research. Research services in the cloud such as AzureBlast can make any research group competitive with the best funded research organizations in the world.

SciCloud. The Scientific Computing Cloud (SciCloud)[17-18] is a project established at the University of Tartu. The main goal of this project is to study the scope of establishing private clouds at universities. With these clouds, students and researchers can efficiently use the already existing resources of university computer networks, in solving computationally intensive scientific, mathematical, and academic problems. Traditionally such computationally intensive problems were targeted by batch-oriented models of the Grid computing domain, where as current project tries to achieve this with the more interactive and service oriented models of cloud computing that fits a larger class of applications. The established interoperable private clouds also provide better platforms for collaboration among interesting groups of universities and in testing internal pilots, innovations and social networks. The project mainly targets the development of a framework, including models and methods for establishment, proper selection, state management (managing running state and data) and interoperability of the private clouds. Once such clouds are feasible, the networks can also be leased to commercial enterprises or governmental institutions for such diverse applications as drug discovery, seismic analysis, and back-office data processing in support of e-commerce and Web services[19].

Aneka. Aneka[20] is a platform and a framework for developing distributed applications on the cloud. It is based on the .NET framework and this is what makes it unique from a technology point of view as opposed to the widely available Java based solutions. Aneka, which is an interesting solution for different types of applications in educational, academic, and commercial environments, has been used to provide support for distributed execution of evolutionary optimizers and learning classifiers. In these cases a significant speed up has been obtained compared to the execution on a single local machine. The preliminary results presented in literature [21] have shown that the use of Aneka has contributed to reduce the execution time of the learning process to the twenty percent of the execution on a single machine.

Conclusions and Future Work

This paper presented the background, key technologies and scientific application prospect of cloud computing, a new on-demand and service-oriented computing model. Cloud computing can completely overturn the original calculation mode of scientific fields, and provide scientists completely customizable and flexible services. It adopts virtualization technology to share resources worldwide transparently, so as to achieve maximum utilization of resources naturally. The adoption of cloud computing as a technology and a paradigm for the new era of computing has definitely become popular and appealing within the enterprise and service providers. It also has widely spread among end users, which more and more host their personal data to the cloud. Nevertheless, this trend for scientific computing is still at an early stage. What could make cloud computing attractive for scientific institutions is the possibility of having a fully customizable runtime environment for their experiments. The active interest of government bodies such as the Department of Energy (DOE) in cloud computing will probably open pathways to the establishment of more science clouds.

According to the survey results, the platforms based on cloud computing can provide a very good operating environment for scientific experiments, increase the performance of applications and reduce the execution time of scientific experiments. Nevertheless, the practical application is still limited, so our next job is to do more research to expand the applications of computing clouds in computing science areas, and also to improve the restrictive conditions of clouds such as the necessity of lasting of high-speed network connection and the insecurity of cloud data.

Acknowledgment

This work is primarily supported by Yunnan Computer Technology Application Key Lab in Kunming University of Science and Technology. We appreciate the support from Natural Science Foundation of China (10878009), and also gratefully acknowledge the helpful comments and suggestions of the anonymous reviewers.

References

[1] "Scientific computing : http://baike.baidu.com/view/605662.htm.."

[2] J. Chin, *et al.*, "Scientific grid computing: the first generation," *Computing in Science & Engineering*, vol. 7, pp. 24-32, 2005.

[3] S. Ghemawat, *et al.*, "The Google file system," *ACM SIGOPS Operating Systems Review*, vol. 37, pp. 29-43, 2003.

[4] N. A. L. Gillam, "Cloud Computing Principles, Systems and Applications," 2010.

[5] C. Vecchiola, *et al.*, "High-performance cloud computing: A view of scientific applications," 2009, pp. 4-16.

[6] M. Armbrust, *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, pp. 50-58, 2010.

[7] F. Chang, *et al.*, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, pp. 1-26, 2008.

[8] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, pp. 107-113, 2008.

[9] "Cloud Computing : http://baike.baidu.com/view/1316082.htm."

[10] ""HP, Intel and Yahoo! Create Global Cloud Computing Research Test Bed http://www.businesswire.com/portal/site/google/?ndmViewId=news_view&newsId=20080729005585&newsLang=en."."

[11] K. Keahey, *et al.*, "Science clouds: Early experiences in cloud computing for scientific applications," *Cloud Computing and Applications,* vol. 2008, 2008.

[12] "Science Clouds : http://scienceclouds.org/."

[13] "The Nimbus Cloud : http://www.nimbusproject.org/nimbus_cloud."

[14] B. Ludascher, *et al.*, "Scientific workflow management and the Kepler system," *Concurrency and Computation,* vol. 18, p. 1039, 2006.

[15] W. Lu, *et al.*, "AzureBlast: a case study of developing science applications on the cloud," presented at the Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, Chicago, Illinois, 2010.

[16] S. F. Altschul, *et al.*, "Basic local alignment search tool," *Journal of molecular biology,* vol. 215, pp. 403-410, 1990.

[17] "Scientific Computing on the Cloud (SciCloud) : http://ds.cs.ut.ee/research/scicloud."

[18] "SciCloud : http://dougdevel.org/scicloud/trac/wiki/SciCloud."

[19] S. Srirama, *et al.*, "SciCloud: Scientific Computing on the Cloud," presented at the Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010.

[20] C. Vecchiola, *et al.*, "Aneka: a software platform for .NET-based Cloud computing," *High Performance & Large Scale Computing, Advances in Parallel Computing. IOS Press, Amsterdam,* 2009.

[21] X. Chu, *et al.*, "Aneka: Next-generation enterprise grid platform for e-science and e-business applications," 2008, pp. 151-159.