HIGH-PERFORMANCE CLOUD COMPUTING AND DATA MINING: A SCIENTIFIC APPROACH

Priyanka Gautam
Associate Professor of Computer and Information Science, CPJCHS, NARELA

Abstract: Data mining techniques and applications are very much needed in the cloud computing paradigm. Cloud computing is a rapidly gaining popularity computing paradigm, prompted by much research efforts. However, not much work is done in the area of joining cloud computing with high performance computing in an efficient way, e.g., for scientific simulation purposes. Moreover, there is even less research effort in making cloud computing for scientific simulations more efficient and suitable for specific simulation code. Scientific computing often requires the availability of a massive number of computers for performing large scale experiments. Traditionally, these needs have been addressed by using high-performance computing solutions and installed facilities such as clusters and super computers, which are difficult to setup, maintain, and operate. Cloud computing provides scientists with a completely new model of utilizing the computing infrastructure. Compute resources, storage resources, as well as applications, can be dynamically provisioned on a pay per use basis. As examples of scientific computing in the Cloud, we present a preliminary case study on using Aneka for the classification of gene expression data and the execution of fMRI brain imaging workflow. This paper presents an ongoing "SimPaaS" project – our research efforts in building a cloud based platform for scientific simulations. It deals with some challenging features in cloud computing, such as performance. The concepts and methods proposed in this paper allow customizing and optimizing cloud infrastructure to increase its performance and to meet certain requirements with the help of Data Mining.

Keywords: Scientific computing, Aneka Cloud Computing, computational science, Cloud computing, high-performance computing, Data Mining

I. INTRODUCTION

"Cloud computing" is the next natural step in the evolution of on-demand information technology services and products. To a large extent, cloud computing will be based on virtualized resources. The term became "popular" sometime in October 2007 when IBM and Google announced collaboration in a domain. This was followed by IBM's announcement of the "Blue Cloud". Since then, everyone is talking about "Cloud Computing. 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 or to cut down the computational complexity into a reasonable time frame. These needs have been initially addressed with dedicated high-performance computing (HPC) infrastructures such as clusters or with a pool of networked machines in the same department, managed by some "CPU cycle scavenger "software such as Condor. With the advent of Grid computing new opportunities became available to scientists: in a complete analogy with the power Grid the computing Grid could offer on demand the horse power required to perform large experiments, by relying on a network of machines, potentially extended all over the world. Computing Grids introduced new capabilities such as dynamic discovery of services, the ability of relying on a larger number of resources belonging to different administrative domains and of finding the best set of machines meeting the requirements of applications. The use of Grids for scientific computing has become so successful that many international projects led to the establishment of world-wide infrastructures available for computational science.

The Internet is becoming an increasingly vital tool in our everyday life, both professional and personal, as its users are becoming more numerous. It is not surprising that business is increasingly conducted over the Internet. Cloud computing introduce a reference model for Cloud computing, cyber infrastructure, Aneka and identify the key services that this new technology offers.

II. OBJECTIVES OF STUDY

- To find out the different challenges and Issues of data mining techniques and High performance Cloud Computing.
- To predict the unknown values i.e. analysis of output of different techniques by Aneka Model.
- To Comparing different techniques on different factors e.g. input and time taken to train and test.
- To Analyze the Cloud Computing Reference Model to integrate the services available for g Aneka with public clouds.

Cloud Definition: Although, the term Cloud computing is too broad to be captured into a single definition it is possible to identify some key elements that characterize this trend. "Cloud computing refers to both the applications delivered as services over the Internet and the hardware and system software in the datacenters that provide those services". They then identify the Cloud with both the hardware and the software components of a datacenter. A more structured definition is who define a Cloud as a "type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more

unified computing resources based on service-level agreement". A key differentiating element of a successful information technology (IT) is its ability to become a true, valuable, and economical contributor to cyber infrastructure [4]. "Cloud" computing embraces cyber infrastructure, and builds upon decades of research in virtualization, distributed computing, "grid computing", utility computing, and, more recently, networking, web and software services. It implies a service oriented architecture, reduced information technology overhead for the end-user, greater flexibility, reduced total cost of ownership, on demand services and many other things. One of the key features characterizing Cloud computing is the ability of delivering both infrastructure and software as services. More precisely, it is a technology aiming to deliver on demand IT resources on a pay per use basis. Previous trends were limited to a specific class of users, or specific kinds of IT resources. Cloud computing aims to be global: it provides the a forementioned services to the mass, ranging from the end user that hosts its personal documents on the Internet, to enterprises outsourcing their entire IT infrastructure to external datacenters.

Some aspects regarding Cloud Computing
Cloud computing represents both the software and the hardware delivered as services over the Internet. Cloud Computing is a new concept that defines the use of computing as a utility, that has recently attracted significant attention.

The computing paradigm shift on the last half century through six distinct phases:

- Phase 1: people used terminals to connect to powerful mainframes shared by many users.
- Phase 2: stand-alone personal computers became powerful enough to satisfy users' daily work.
- Phase 3: computer networks allowed multiple computers to connect to each other.
- Phase 4: local networks could connect to other local networks to establish a more global network.
- Phase 5: the electronic grid facilitated shared computing power and storage resources.
- Phase 6: Cloud Computing allows the exploitation of all available resources on the Internet in a scalable and simple way.

Cyber infrastructure: Cyberinfrastructure makes applications dramatically easier to develop and deploy, thus expanding the feasible scope of applications possible within budget and organizational constraints, and shifting the scientist's and engineer's effort away from information technology development and concentrating it on scientific and engineering research. Cyberinfrastructure 236 Cloud Computing – Issues, Research and Implementations also increases efficiency, quality, and reliability by capturing commonalities among application needs, and facilitates the efficient sharing of equipment and services." Today, almost any business or major activity uses, or relies in some form, on IT and IT services. These services need to be enabling and appliance-like, and there must be an economy of- scale for the total-cost-of-ownership to be better than it would be without cyberinfrastructure. Technology needs to improve enduser productivity and reduce technology-driven overhead. For example, unless IT is the primary business of an organization, less than 20% of its efforts not directly connected to its primary business should have to do with IT overhead, even though 80% of its business might be conducted using electronic means.

Cloud Computing Reference Model: This figure gives an overview of the scenario envisioned by Cloud computing. It provides a layered view of the IT infrastructure, services, and applications that compose the Cloud computing stack. It is possible to distinguish four different layers that progressively shift the point of view from the system to the end user. The lowest level of the stack is characterized by the physical resources on top of which the infrastructure is deployed. These resources can be of different nature: clusters, datacenters, and spare desktop machines. Infrastructures supporting commercial Cloud deployments are more likely to be constituted by datacenters hosting hundreds or thousands of machines, while private Clouds can provide a more heterogeneous environment, in which even the idle CPU cycles of spare desktop machines are used to leverage the compute workload. This level provides the "horse power" of the Cloud.
Cloud resources

Figure 1. Cloud computing layered architecture.

The physical infrastructure is managed by the core middleware layer whose objectives are to provide an appropriate runtime environment for applications and to exploit the physical resources at best. In order to provide advanced services, such as application isolation, quality of service, and sandboxing, the core middleware can rely on virtualization technologies. Among the different solutions for virtualization, hardware level virtualization and programming language level virtualization are the most popular. Hardware level virtualization guarantees complete isolation of applications and a fine partitioning of the physical resources, such as memory and CPU, by means of virtual machines. Programming level virtualization provides sandboxing and managed execution for applications

developed with a specific technology or programming language (i.e. Java, .NET, and Python). On top of this, the core middleware provides a wide set of services that assist service providers in delivering a professional and commercial service to end users. These services include: negotiation of the quality of service, admission control, execution management and monitoring, accounting, and billing. Together with the physical infrastructure, the core middleware represents the platform on top of which the applications are deployed in the Cloud. It is very rare to have direct user level access to this layer. More commonly, the services delivered by the core middleware are accessed through a user level middleware.

Cloud Computing Services Offering: The wide variety of services exposed by the Cloud Computing stack can be classified and organized into three major offerings that are available to end users, scientific institutions, and enterprises. These are: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

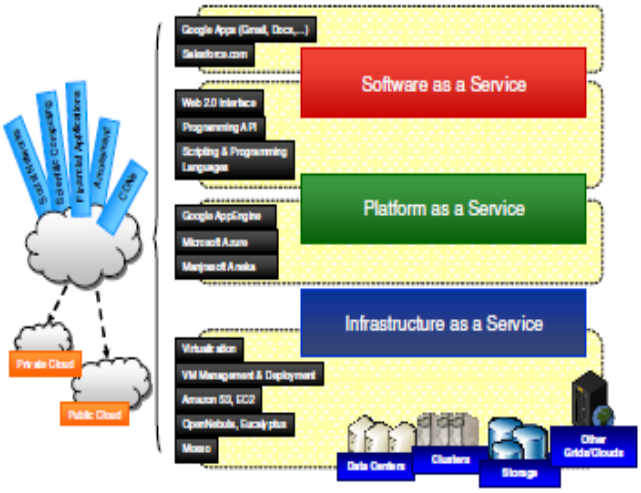

Figure 2. Cloud computing offerings by services.

This figure provides such categorization, Infrastructure as a Service or Hardware as a Service (HaaS) are terms that refer to the practice of delivering IT infrastructure based on virtual or physical resources as a commodity to customers. These resources meet the end user requirements in terms of memory, CPU type and power,storage, and, in most of the cases, operating system. Users are billed on a pay per use basis and have to set up their system on top of these resources that are hosted and managed in datacenters owned by the vendor. Amazon is one of the major players in providing IaaS solutions. Amazon Elastic Compute Cloud (EC2) provides a large computing infrastructure and a service based on hardware virtualization. By using Amazon Web Services, users can create Amazon Machine Images (AMIs) and save them as templates from which multiple instances can be run. It is possible to run either Windows or Linux virtual machines and the user is charged per hour for each of the instances running. Amazon also provides storage services with the Amazon Simple Storage Service (S3), users can use Amazon S3 to host large amount of data accessible from anywhere.Platform as a Service solutions provide an application or development platform in which users can create their own application that will run on the Cloud. PaaS implementations provide users with an application framework and a set of API that can be used by developers to program or compose applications for the Cloud. In some cases, PaaS solutions are generally delivered as an integrated system offering both a development platform and an IT infrastructure on top of which applications will be executed. The two major players adopting this strategy are Google and Microsoft. Google AppEngine is a platform for developing scalable web applications that will be run on top of server infrastructure of Google. It provides a set of APIs and an application model that allow developers to take advantage of additional services provided by Google such as Mail, Datastore, Memcache, and others. By following the provided application model, developers can create applications in Java, Python, and JRuby.

ANEKA: Aneka is a software platform and a framework for developing distributed applications on the Cloud. It harnesses the computing resources of a heterogeneous network of desktop PCs and servers or datacenters on demand. Aneka provides developers with a rich set of APIs for transparently exploiting such resources and expressing the logic of applications by using a variety of programming abstractions. System administrators can leverage a collection of tools to monitor and control the deployed infrastructure. This can be a public cloud available to anyone through the Internet, or a private cloud constituted by a set of nodes with restricted access within an enterprise. The flexible and service-oriented design of Aneka and its fully customizable architecture make Aneka Clouds able to support different scenarios. Aneka Clouds can provide the pure compute power required by legacy financial applications, can be a reference model for teaching distributed computing, or can constitute a more complex network of components able to support the needs of large scale scientific experiments. This is also accomplished by the variety of application programming patterns supported through an extensible set of programming models. These define the logic and the abstractions available to developers for expressing their distributed applications. As an example, in order to run scientific experiments it is possible to rely on a classic bag of tasks model, or to implement the application as a collection of interacting threads or MPI processes, a set of interrelated tasks defining a workflow .If the available options do not meet the requirements, it is possible to seamlessly extend the system with new programming abstractions. Aneka Clouds can be built on top of different physical infrastructures and integrated with other Cloud computing solutions such as Amazon EC2 in order to extend on demand their capabilities. In this particular scenario, Aneka acts as a middleman mitigating the access to public clouds from user applications. It operates as an application service provider that, by using fine and sophisticated pricing policies, maximizes the utilization of the rented virtual resources and shares the costs among users. Of a particular importance arethen, the accounting and pricing services and how they operate when Aneka integrates public clouds.

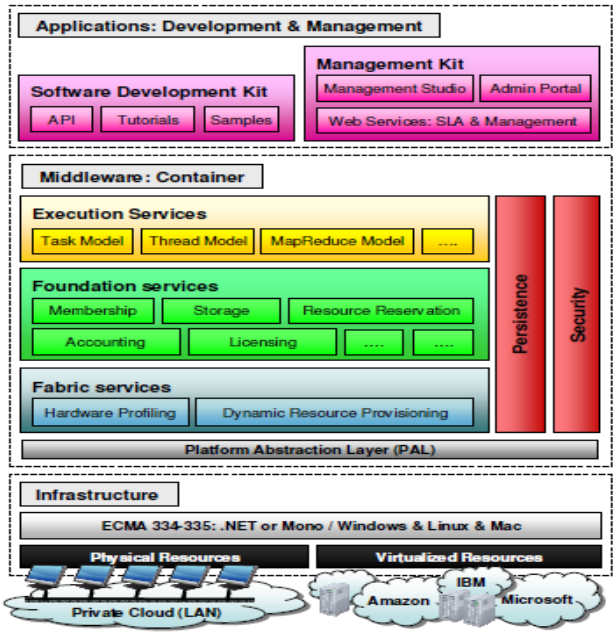

Figure 3. Aneka architecture.

This figure gives an architectural overview of Aneka. In order to develop cloud computing applications developers are provided with a framework that is composed by a software development kit for programming applications, a management kit for monitoring and managing Aneka Clouds and a configurable service based container that constitute the building blocks of Aneka Clouds. In this section we will mostly focus on three key features: the architecture of Aneka clouds, the application model, and the services available for integrating Aneka with public clouds.

Aneka Clouds: The Aneka cloud is a collection of software daemons –called containers – that can be hosted on either physical or virtual resources and that are connected through a network such as the Internet or a private intranet. The Aneka container is the building block of the entire system and exposes a collection of services that customize the runtime environment available for applications. It provides the basic management features for a single node and leverages the hosted services to perform all the other operations. We can identify fabric and foundation services. Fabric services directly interact with the node through the Platform Abstraction Layer (PAL) and perform hardware profiling and dynamic resource provisioning. Foundation services identify the core system of the Aneka infrastructure; they provide a set of basic features on top of which each of the Aneka containers can be specialized to perform a specific set of tasks. One of the key features of Aneka is the ability to provide multiple ways of expressing distributed applications by offering different programming models; execution services are mostly concerned with providing the middleware with an implementation for these models. Additional services such as persistence and security are transversal to the entire stack of services that are hosted by the container. The network of containers can be the result of different deployment scenarios: it can represent a private cloud completely composed by physical machines (desktop PCs and clusters) within the same administrative domain such as an enterprise or a university department. On the other hand, a totally virtual infrastructure is possible and the entire Aneka Cloud can be hosted on a public cloud such as Amazon EC2 or a private datacenter.

Data mining in Cloud Computing

Data mining techniques and applications are very much needed in the cloud computing paradigm.
As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining.
"Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources.
The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users."
As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way.

The main effects of data mining tools being delivered by the Cloud are:

- The customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive.
- The customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.
- Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments.

Research Issues: The general cloud computing approach discussed so far, as well as the specific VCL implementation of a cloud continues a number of research directions, and opens some new ones For example, economy-of-scale and economics of image and service construction depends to a large extent on the ease of onstruction and mobility of these images, not only within a cloud, but also among different clouds. Of special interest is construction of complex environments of resources and complex control images for those resources, including workflow-oriented images. Temporal and spatial feedback large scale workflows may present is a valid research issue. Underlying that is a considerable amount of meta-data, some permanently attached to an image, some dynamically attached to an image, and some kept in the cloud management databases. Cloud provenance data, and in general metadata management, is an open issue.

The classification we use divides provenance information into followings:

• Cloud Process provenance – dynamics of control flows and their progression, execution information, code performance tracking, etc.
• Cloud Data provenance – dynamics of data and data flows, file locations, application input/output information, etc.
• Cloud Workflow provenance – structure, form, evolution, of the workflow itself
• System (or Environment) provenance – system information, O/S, compiler versions,Loaded libraries, environment variables, etc.

Open challenges include:

- How to collect provenance information in a standardized and seamless way and with minimal overhead –modularized design and integrated provenance recording;
- How to store this information in a permanent way so that one can come back to it at anytime, – standardized schema;
- How to present this information to the user in a logical manner – an intuitive user web interface:Dashboard .Some other image- and service-related practical issues involve finding optimal image and service composites and optimization of image and environment loading times.

There is also an issue of the image portability and by implication of the image format. For example, VCL currently uses standard image snapshots that may be an operating system, hypervisor and platform specific, and thus exchange of images requires relatively complex mapping and additional storage.

III. CONCLUSION

"Cloud" computing builds on decades of research in virtualization, distributed computing, utility computing, and, more recently, networking, web and software services. It implies a service-oriented architecture, reduced information technology overhead for the end-user, great flexibility, reduced total cost of ownership, on demand services and many other things. This paper discusses the concept of "cloud" computing, the issues it tries to address, related research topics, and "Aneka cloud" terminology. We have discussed the potential opportunities and the current state-of-the-art of high-performance scientific computing on public clouds. 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. Science computing Grids such as Open Science Grid and EGEE already provide a large scale infrastructure, a set of well established methods and tools, and huge community of users. What could make interesting the use of computing. At present, some preliminary works have investigated the cost of doing science in the Cloud, by taking the Amazon EC2 and S3 infrastructure. From an operational point of view, the first science computing cloud has been has been already deployed as a result of the joint efforts of a consortium of universities. The active interest of government bodies such as the Department of Energy in Cloud computing, will probably open pathways to the establishment of more science Clouds. A stronger adoption of Cloud computing for computational science will also contribute to advance research in other functional aspects such as security and jurisdiction. Many scientific projects are funded by the government bodies that sometimes impose significant restrictions on the use of data. This paper provides an overview of the necessity and utility of data mining in cloud computing. As the need for data mining tools is growing every day, the ability of integrating them in cloud computing becomes more and more stringent.

REFERENCES

[1] Amazon Elastic Compute Cloud (EC2):http://www.amazon.com/gp/browse.html node=201590011, accessed Dec 2008.

[2] ILKAY ALTINTAS, BERTRAM LUDAESCHER, SCOTT KLASKY, MLADEN A. VOUK. "Introduction to scientific workflowmanagement and the Kepler system",Tutorial, in Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, Tampa, Florida, TUTORIAL SESSION, Article No. 205, 2006,ISBN:0-7695-2700-0, also given at Supercomputing 2007 by Altintas, Vouk, Klasky, Podhorszki.

[3] ILKAY ALTINTAS, GEORGE CHIN, DANIEL CRAWL,TERENCE CRITCHLOW, DAVID KOOP, JEFF LIGON,BERTRAM LUDAESCHER, PIERRE MOUALLEM1,MEIYAPPAN NAGAPPAN, NORBERT PODHORSZKI, CLAUDIO SILVA, MLADEN VOUK, "Provenance in Kepler-based ScientificWorkflow Systems", Poster # 41, at Microsoft eScience Workshop Friday Center,University of North Carolina, Chapell Hill.

[4] D.E. ATKINS ET AL., "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of theNational Science FoundationBlue-ribbonAdvisoryPanel on Cyberinfrastructure", NSF, Report of theNational Science FoundationBlue-ribbonAdvisory Panel on Cyberinfrastructure, January 2003,http://www.nsf.gov/od/oci/reports/atkins.pdf

[5] Ditto, Appendix A (http://www.nsf.gov/od/oci/reports/APXA.pdf)

[6] SAM AVERITT, MICHALE BUGAEV, AARON PEELER,HENRY SCHAFFER, ERIC SILLS, SARAH STEIN, JOSH THOMPSON, MLADEN VOUK, "The Virtual Computing Laboratory", Proceedings of the International Conference on Virtual Computing Initiative, May 7-8, 2007, IBM Corp.

[7] ROSELYNE BARRETO, TERENCE CRITCHLOW, AYLA KHAN, SCOTT KLASKY, LEENA KORA, JEFFREY LIGON, PIERRE MOUALLEM, MEIYAPPAN NAGAPPAN, NORBERT PODHORSZKI, MLADEN VOUK, "Managing and Monitoring Scientific Workflows

through Dashboards", Poster # 93, at Microsoft eScience Workshop Friday Center, University of North Carolina, Chapell Hill, NC, October 13 – 15,2007.

[8] D.A. BATCHELOR, M. BECK, A. BECOULET, R.V. BUDNY, C. S. CHANG, P. H.DIAMOND, J. Q.DONG, G. Y. FU, A. FUKUYAMA, T. S. HAHM, D. E.KEYES,Y. KISHIMOTO, S. KLASKY, L. L. LAO1, K. LI1, Z. LIN1, B. LUDAESCHER, J. MANICKAM, N. NAKAJIMA1,T. OZEKI1, N. PODHORSZKI, W. M. TANG.

[9] MICHAEL BELL, "Introduction to Service-oriented Modeling", Service-oriented Modeling: Service Analysis, Design, and Architecture. Wiley & Sons, 3. ISBN 978-0-470-14111-3, 2008.

[10] W. M. BULKELEY, "IBM, Google, Universities Combine 'Cloud' Foces", Wall Street Journal,October,2007,http://online.wsj.com/public/articleprint/SB119180611310551864.html.

