Cloud Computing In Bioinformatics: Current Status and Future Research

Ritushree Narayan
Department of Computer science
R.L.S.Y.College, Ranchi-834001

Abstract— The problems faced by bioinformatics researchers in order to carry out their research in an economic and fast manner can be solved easily with the help of Cloud computing concepts. Thus, cloud computing is a boon for bioinformatics research. In this paper we have discussed how the cloud computing will be helpful for bioinformatics researchers. Cloud Computing offers large scalable computing and storage, data sharing, on-demand anytime and anywhere access to resources and applications, and it supports easy but powerful distributed computing models, for facing those issues. In fact, in the recent years it has been adopted for the deployment of several applications in healthcare and bioinformatics both in academia and in the industry. However, cloud computing presents several issues regarding the security and privacy of data.

Key words: cloud computing, bioinformatics, HPC architecture

I. Introduction

Cloud computing is a newly emerging technology for the future with it roots based on the rapidly increasing demands on data centers that needs to be catered to. Cloud computing is defined as the use of computing resources to access data over the internet. It is a means or a mechanism to enhance the existing capabilities of Information technology by many folds. The terminology cloud comes from the fact that the data is not stored on your desktop or your device but is located far away similar to a cloud in literal terms, but despite of it being away its within your reach, you can access it irrespective of your geographical location using a computing device via an internet. Cloud computing is a technology for the future and will change the entire scenario of the IT industry, being a cost efficient approach, with reduced exigency of buying the software or the hardware resources. It is an on demand form of utility computing for those who have access to cloud. There has been a gradual shift in interest of researchers towards cloud computing in the past few years. Several researchers have started the use of cloud computing in bioinformatics research. According to researchers who are working in the field of bioinformatics are confronted with analysis of ultra large-scale data sets right now. They have also come up with a problem that growth of data will boost at a shocking velocity in upcoming years and current developments in real world open source software, the Hadoop project and associated software will provide an establishment for scaling to petabyte size data warehouses on Linux clusters. Hadoop also provides fault tolerant parallelized investigations on such data by means of a programming approach called Map Reduce[6].

II. Cloud Computing

Cloud computing comes in four different forms of a cloud: public, private, hybrid and community.

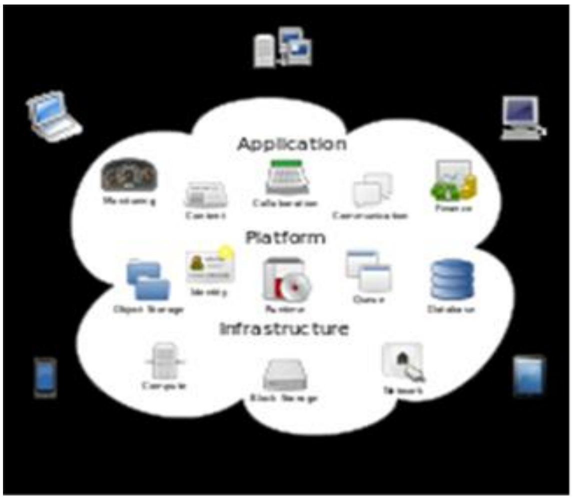

Fig. 1: Cloud Computing

A. *Private Cloud*

Private cloud is cloud infrastructure which is managed for a single organization; a private cloud can be owned and managed either internally or externally by an organization. For Example Intel, Hewlett Packard (HP) and Microsoft have their own internal private cloud.

B. *Public Cloud*

Public cloud applications, resources, storage, and other services are made available to the general public by a service provider. These cloud service providers are organizations engaged in selling cloud services like Amazon AWS, Microsoft and Google. These services can be offered either free of cost or may be made available through pay per use. The quality of services offered by the service providers is mentioned in a Service level agreement i.e. it is an agreement between a consumer and a cloud service provider. It include services that will be offered by a service provider in terms of privacy, security, backup procedures Some of the examples of public cloud are Amazon Web Services (AWS) and Microsoft Azure .

C. *Community Cloud*

In community cloud the cloud infrastructure is shared by the organizations that have common resource requirements (security, jurisdiction and policy), whether managed internally or by a third-party and hosted internally or externally. They take slight advantage of cloud computing in terms of sharing the costs, as the cost is not alone born by a single organization but rather shared by different organizations For Example Google Gov (google apps for government).

D. *Hybrid Cloud*

Hybrid cloud is a composition of two or more clouds i.e. it can be a composition of either a private and public or a private and community cloud etc. By utilizing hybrid cloud architecture, organizations and individuals are able to obtain degrees of fault tolerance combined with locally immediate usability without dependency on internet connectivity. Hybrid clouds have few limitations such as the lack of flexibility, security and certainty of in-house client applications. Hybrid cloud ensures the provision and flexibility of in house applications with the fault tolerance and scalability of cloud based services.[2]

1) Cloud computing service delivery models

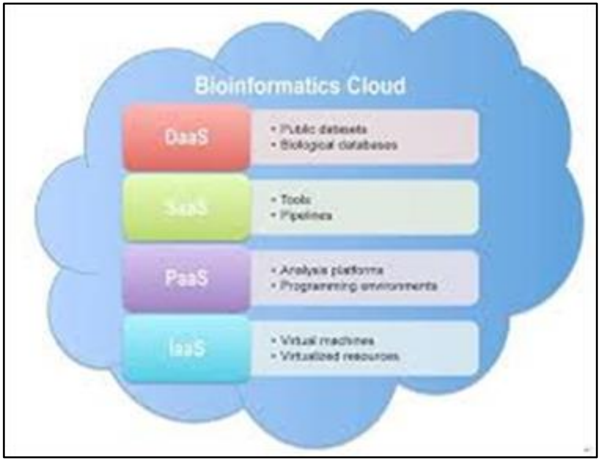

Fig. 2: Bioinformatics Cloud

Cloud computing providers offer three fundamental service models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS):

– *Data as a service (DaaS)* :Bioinformatics clouds are dependent on data for downstream analyses. "It is reported that annual worldwide sequencing capacity is beyond 13 Pbp and on an increase by a factor of five every year" [17]. Due to this unrevealed explosion of data, Data as a Service (DaaS) delivery via Internet has gained importance. It provides dynamic data access on demand, along with up-to-date data access to a wide range of devices, connected over the Web. Amazon Web Services (AWS) provide a centralized cloud of public data sets (e.g. archives of GenBank, Ensembl databases, 1000 Genomes, Model Organism Encyclopedia, Unigene, etc.) of biology, chemistry, economics, etc. as services

– *Infrastructure as a service (IaaS):* SaaS delivers a large variety of software services online for different types of data analysis facilitating remote access of various heavy bioinformatics softwares. Thus, it eliminates the need for local installation, thereby easing software maintenance. Up-to-date cloudbased services for bioinformatic data analysis has made life easy for the users. Efforts have been made to develop cloud-scale and cloud-based sequence mapping , multiple sequence alignment , expression analysis, identification of epistatic interactions of SNPs (single nucleotide polymorphisms), and NGS (Next-Generation Sequencing).

– *Platform as a service (PaaS):* PaaS allow users to develop, test and use cloud applications in an environment where computer resources scale to match application demand automatically and dynamically. This scalability factor helps in developing applications for biological data.

  Two PaaS platforms:

  1) Eoulsan, cloud-based-for high-throughput sequencing analyses;
  2) Galaxy Cloud, cloud-scale-for large-scale data analyses.

– *Software as a Service (SaaS):* IaaS delivers all kinds of resources (virtualized) including CPU (hardwares), OS (softwares) etc. summing up a full computer infrastructure, reaching to the full potential of computer resources via Internet. Virtualized resources can be accessed as a public utility by users and thereby paying for the cloud resources that they utilize. Flexibility and customization give freedom to different users to access different cloud resources, as per their requirement, thus meeting the customized needs of different users.

  Examples:

  1) Cloud BioLinux is a virtual machine that is publicly accessible for high-performance bioinformatics computing.
  2) CloVR is a portable virtual machine that incorporates several pipelines for automated sequence analysis. [5]

III. CLOUD COMPUTING IN BIOINFORMATICS

Traditional bioinformatics research usually involves downloading huge data sets from publicly available data repositories and then analyzing these data sets through inhouse infrastructure. The major problem faced by these scientists who are working in the field of bioinformatics is that they need machines having high computation power requirements but such high power computation machines require a great deal of economies of scale posing as a hurdle to the researchers. It is here that cloud computing comes into picture by providing rescue to researchers facing economic crises. Through cloud computing many high power machines can be taken on rent basis depending upon the user requirements and usage. Software's and data required for carrying out research can also be placed in cloud and then accessed as a service as per the usage and requirements. Apart from meeting the infrastructure needs of researchers it also makes the entire process of installing software's now a redundant task. The biologist need no longer to develop any expertise for installing software's operating systems etc. They can just log on to the respective cloud and use the software required by them without any initial setups. Thus, saving a lot of money and time required for performing initial laboratory setups. Cloud computing also has its application in Healthcare and Biomedicine domain. In the authors have discussed the main cloud-based healthcare and biomedicine applications along with a special focus on bioinformatics solutions and its underlying important issues and problems related to the use of cloud computing environment for the storage and analysis. They have used data of patients in this study. They have also discussed recent studies which show that cloud computing can improve healthcare services and benefit biomedical research through proposing innovative solutions and applications. Reduced costs are essential drivers that have led to recognition of cloud as a technology in the healthcare domain. The cost of basic healthcare conveyance has increased to such a huge amount that government is facing severe funding issues and several patients are on the verge of remaining unattended to basic medical amenities. The recognition of cloud computing as a technology can enhance patient's care and overall well-being of an individual along with reduced costs which means that government can move the usually slow healthcare business to a quicker step of acceptance. In the researchers have discussed, Azures pricing model, the price of their cloud-based telemedicine service is based on the amount of CPU utilization. The cost of resource usage is about USD $ 0.1 per hour and usage of its database price is approximately USD $9.99per GB for one month.. Hadoop works in bioinformatics field because it deals with bringing computation to data rather than moving the entire data to computation thus reducing performance bottleneck characterized with poor network latency and slow computational devices. MapReduce which is a programming framework supported by Hadoop can be used in cloud environment for performing computations in cloud environment in a parallel fashion. Data in Hadoop environment can be stored in Hadoop distributed file system (HDFS) which can be used for storage of data of any volume. HDFS works by storing data in form of chunks on multiple data nodes. Thus, HDFS can be another characteristic of cloud computing which supports bioinformatics application which usually involves capturing of data, storage of huge amount of data and also analysis of large datasets. Bioinformatics applications use different services offered by cloud computing such as data as a service which involves dynamic on demand data access and also makes available the latest data available. Amazons AWS makes available datasets such as GenBank, 1000 Genomes, Influenza virus and Unigene [. These data sets can be integrated transparently with the existing cloud applications. Bioinformatics application tool can also be accessed through software as a service in cloud. Orthology detection and peak caller for CHIP-seq data are some of the bioinformatics tools available as a service. Apart from software and data as a service platforms such as Galaxy Cloud for analysis of huge amount of data is also offered through cloud techniques. Though cloud computing can be an effective technology for meeting the computation needs in bioinformatics research there are certain issues that can affect the performance of the available cloud platforms. One such hindrance is the speed of data transfer. Usually the speed of data transfer is slow and at present there are not many solutions available for moving the existing physical hard drives to cloud. Therefore, we need cloud solutions for integrating cloud computing with efficient data transfer technologies.[4]

IV. HPC ARCHITECTURES

High performance computing has been employed for addressing bioinformatics problems that would otherwise be impossible to solve. A first example was the preliminary assembly of the human genome, a huge effort in which the Human Genome Project was challenged by Celera Genomics with a different approach, consisting in a bioinformatics post analysis of whole sets of shotgun sequencing runs instead of the long-standing vector cloning technique, arriving very close to an unpredictable victory.

A. *GPU Computing*

Driven by the demand of the game industry, graphics processing units (GPUs) have completed a steady transition from mainframes to workstations to PC cards, where they emerge now a days like a solid and compelling alternative to traditional computing platforms. GPUs deliver extremely high floating point performance and massively parallelism at a very low cost, thus promoting a new concept of the high performance computing (HPC) market; i.e. heterogeneous computing where processors with different characteristics work together to enhance the application performance taking care of the power budget. This fact has attracted many researchers and encouraged the use of GPUs in a broader range of applications, particularly in the field of Bioinformatics, where developers are required to leverage this new landscape of computation with new programming models which ease the developers task of writing programs to run efficiently on such platforms altogether.

B. *Supercomputers*

Many of the supercomputers in the TOP500 list are heavily involved in computational biology research. the SuperMUC cluster, installed at the Leibniz Supercomputer Centre in Monaco, is often employed in bioinformatics, for example in analysis of linkage disequilibrium in genotyping and Piz Daint, installed at the CSCS/Swiss Bioinformatics Institute in Lugano, has been successfully employed for a challenge of evolutionary genomics, for calculating selection events in genes many times more quickly.

C. *Grid and Cloud computing*

In current bioinformatics, the cost of buying and maintaining an in-house cluster is very important and this explains why the grid computing paradigm gained a great success in the mid-1990sThe problem is that even if the low-level details of the grid infrastructures are hidden via middlewares, often the application of bioinformatics methods on HPC facilities requires specialized knowledge. A suitable solution to offer easy-to-use and intuitive access to applications are science gateways, which offer specific services tailored to the users needs. Nonetheless, ten years later, cloud computing was presented as a more flexible solution with respect to grid [13]. Cloud computing overcomes the idea of volunteer computing for resource sharing by proposing an on-demand paradigm in which users pay for what they use[1].

V. APPLICATIONS IN BIOINFORMATICS

Bioinformatics research is generally synonyms with high computations, huge data sets and expensive computational equipments. These computational requirements involve high end laboratories to meet the computational needs of the research personnel's. Thus, this need for an expensive setup is becoming a bottleneck in biological discovery at the computational level. Therefore, cloud computing comes as a handy solution to researchers in field of bioinformatics. Cloud computing helps in easing down the burden of bioinformatics researchers by reducing the computation time and optimizing costs involved. With the help of cloud computing researcher can get result easily and quickly without wasting time and money involved in laboratory setups. In the authors have discussed how cloud computing offers vibrant set of resources to small and medium-sized laboratories to quickly adjust their computational capacity depending on their requirements. In this work the authors have benchmarked two established cloud computing services, Amazon Web Services Elastic Map Reduce and Google Compute Engine (GCE), using widely available genomic datasets and a standard bioinformatics channel based on Hadoop platform.[2] V Marx has discussed in his paper about big data challenges that Biologists are facing. According to them biologists are joining the big-data club. According to them with the advent of high-throughput genomics, scientists engaged in life sciences have started grappling with massive data sets. It is also stated that with every transitory year, scientist rotate extra frequently to big data to investigate the whole thing from the regulation of genes and the progression of genomes to why coastal algae bloom, what microbes dwell where in human body cavities and how the genetic composition of dissimilar cancers influences the cancer patients. [7]

Platform Description Cloud Technology Cloud BLAST Bioinformatics application using a combination of virtualization technology and Map Reduce Hadoop, Map Reduce, Blast Myrna RNA-sequencing through differential expression analysis Bio- Linux, Amazon EC2, Galaxy Cloud Cloud Burst Performs sensitive read mapping and SNP calling through Hadoop Map Reduce Amazon EC2, Hadoop Map Reduce Peak Ranger Peak caller for Chip-seq data on cloud Hadoop Map Reduce Rainbow Whole genome sequence data analysis through cloud computing Map Reduce, SOAPsnp, Cross Bow, perl, picard BioPig Analytical toolkit for large scale data analysis using hadoops pig Apache Pig, Hadoop SeqPig Analytical toolkit for large sequencing data set analysis using hadoops pig Apache Pig, Hadoop Galaxy Scientific workflow system available online for genomic research Python, SQL database, web server Galaxy CloudMan Delivers clusters using cloud computing through Amazon EC2 for bioinformatics application Amazon EC2,galaxy, Biolinux .[3]

VI. Conclusions

Applications and services in bioinformatics and biomedicine pose quite demanding requirements. The fulfillment of these requirements could results in an improvement in the provision of services to patients, as well as an increase in knowledge in the biomedical field. Although supercomputing or Grid Computing can provide the computational power and the storage required in biomedical applications such as medical imaging or electronic medical record, the elasticity in providing such resources, a clear definition of Service Level Agreement, and the possibility for the customer to use a pay-per-use model, make the Cloud more suitable to support those applications. Naturally, the adoption of this technology with its benefits will determine a reduction of costs and the possibility of also providing new services. Different comparative analysis demonstrate the potentials of cloud technology in reducing cost of IT organization: in the case of cloud adoption, institutions and laboratories are free from the expense and having to install and maintain applications locally. Specifically, they list all direct costs and estimate all indirect costs and, with the help of a simulation study they estimate the costs for typical research projects in bioinformatics. large amount of bioinformatics data that are nowadays produced. However, it is important to emphasize that the use of cloud in these fields is featured still by a number of open issues, such as the security and privacy, that require a rapid and efficient solution.

This manuscript discusses about the role of cloud computing in the field of bioinformatics for management of huge data sets and for carrying out complex computations which eventually led to its application in Big data analytics as well. We have also focused on how to manage big data using cloud computing. It also discusses about the current state of art of cloud computing in big data analytics along with the scope of future research based on cloud in both bioinformatics as well as big data analytics.

References

[1] Horacio Per´ez-S´anchez, Jos´e M. Cecilia, and Ivan Merelli”The role of High Performance Computing in Bioinformatics” Proceedings IWBBIO 2014.

[2] Leo, S., Santoni, F., Zanetti, G.: Biodoop: bioinformatics on Hadoop. In Parallel Processing Workshops, 2009. ICPPW'09. International Conference on (pp. 415-422). IEEE, (2009).

[3] CloudStore file system. [Online]. Available: http://kosmosfs.sourceforge.net,(2015)

[4] Nidal M. Turab, Anas Abu Taleb Shadi R. Masadeh “CLOUD COMPUTING CHALLENGES AND SOLUTIONS” International Journal of Computer Networks & Communications (IJCNC) Vol.5, No.5, September 2013

[5] Mansaf Alam a and Kashish Ara Shakil ”Recent Developments in Cloud Based Systems: State of Art” “IT-3_Cloud_Computing” A news Letter for IT professionals Issue 3 2012.

[6] Prachi Singh ,“Big Genomic Data in Bioinformatics Cloud” Appli Microbio Open Access ISSN:2471-9315 AMOA, an Open Access Journal 2016.