Chapter 129
Virtual Supercomputer Using Volunteer Computing

Rajashree Shettar
R. V. College of Engineering, India

Vidya Niranjan
R. V. College of Engineering, India

V. Uday Kumar Reddy
CA Technologies, India

ABSTRACT

Invention of new computing techniques like cloud and grid computing has reduced the cost of computations by resource sharing. Yet, many applications have not moved completely into these new technologies mainly because of the unwillingness of the scientists to share the data over internet for security reasons. Applications such as Next Generation Sequencing (NGS) require high processing power to process and analyze genomic data of the order of petabytes. Cloud computing techniques to process this large datasets could be used which involves moving data to third party distributed system to reduce computing cost, but this might lead to security concerns. These issues are resolved by using a new distributed architecture for De novo assembly using volunteer computing paradigm. The cost of computation is reduced by around 90% by using volunteer computing and resource utilization is increased from 80% to 90%, it is secure as computation can be done locally within the organization and is scalable.

1. INTRODUCTION TO NEXT GENERATION SEQUENCING

Modern quantitative biology has changed the perspective of data rich genomic sequencing technology. Large scale genomic data analysis requires the need for a new computational framework supported by High Performance Computing. One such application is the Next Generation Sequencing (NGS), which deals with terabytes or petabytes of genome data requiring high computational power.

DOI: 10.4018/978-1-7998-5339-8.ch129

Next Generation Sequencing (NGS) (Wilson et al., 2002; Narzisi et al., 2011) is a technique of sequencing the exact order of nucleotides which form the basic building blocks of Deoxyribonucleic Acid (DNA). NGS with a market size of over 2.7 billion dollars has diverse uses in fields of biological sciences ranging from identification of diseases in human beings to invention of sequence for novel species. Traditionally sequencing was done by treating DNA chemically and identifying nucleotides using color codes, but this technique of sequencing is not suitable for organisms with just thousands of nucleotides. Earlier, the cost of producing base pair information stored as ‘reads’ was limited to wet laboratory techniques and was very expensive. Hence the rate of production of data was very slow, but new sequencing technologies combined with wet lab techniques and information technology started producing millions to billions of short ‘reads’ quickly. The traditional assembly tools used earlier was incapable to handle this huge data.

To overcome these problems a number of assembly technologies have been invented that uses computations performed by computer, also known as the *In silico* approach. These assemblers started with small datasets and were effective. As the size of ‘reads’ increased, the assemblers required either a single computer with very large amounts of memory and computing resources or the data to be sent to third party for execution such as cloud computing which might lead to security concerns. These constraints make the analysis of huge amount of genomic data a tedious task.

An alternate solution to Cloud and Hadoop is to use volunteer computing which is proposed and explained in this chapter. In particular emphasis is on recommending a solution to Next Generation Sequencing (NGS) which uses an open source grid middleware namely Berkeley Open Infrastructure for Network Computing (BOINC) designed to handle various applications that require high computational power, data storage or both. This will be a great enabler for bioinformatics scientists to create applications that use public computing resources.

1.1 Importance of Big Data and Cloud in Sequencing

Bioinformatics domain has brought in lot of challenges with respect to management of enormous amount of genomic data that is growing exponentially. Modern Biology has moved from wet lab computing techniques to big data analytics, cloud computing and open source techniques for analysis and inference of bioinformatics data.

From global perspective, Big Data is a recent trend which has attracted many researchers and scientist to work upon (Sharma et al., 2015). MapReduce with Hadoop, a programming model is one of the techniques to analyze big data. It provides a platform for data processing in parallel fashion (Seema et al., 2015). Big data analytics enables analyzing large data of size varying from Petabyte (PB) to Exabyte (EB) to extract hidden pattern and useful information from large datasets (Fisher et al., 2012).

Conventional database services suffer from performance issues when fed with large amount of data. Many machine learning techniques and data mining algorithms can be developed and integrated with MapReduce, a general purpose parallel programming model to improve efficiency of systems with large data. Hence, normally a large data is divided into smaller datasets and are assigned to different mapper nodes to work in parallel. The intermediate results from all these nodes are collected. At reduce nodes, computations of algorithms will be carried out to form the final result. Hadoop is an open source platform used for computation of large data developed by Apache foundation (White et al., 2009). Hadoop adopts Hadoop Distributed Files System (HDFS) for storage and MapReduce. Hadoop requires com-