Chapter 11
Big Data in Massive Parallel Processing:
A Multi-Core Processors Perspective

Vijayalakshmi Saravanan
The State University of New York at Buffalo, USA

Anpalagan Alagan
Ryerson University, Canada

Isaac Woungang
Ryerson University, Canada

ABSTRACT

With the advent of novel wireless technologies and Cloud Computing, large volumes of data are being produced from various heterogeneous devices such as mobile phones, credit cards, and computers. Managing this data has become the de-facto challenge in the current Information Systems. According to Moore's law, processor speeds are no longer doubling, the processing power also continuing to grow rapidly which leads to a new scientific data intensive problem in every field, especially Big Data domain. The revolution of Big Data lies in the improved statistical analysis and computational power depend on its processing speed. Hence, the need to put massively multi-core systems on the job is vital in order to overcome the physical limits of complexity and speed. It also arises with many challenges such as difficulties in capturing massive applications, data storage, and analysis. This chapter discusses some of the Big Data architectural challenges in the perspective of multi-core processors.

INTRODUCTION

Over the recent years, Chip Multi-Processors (CMPs) have become one of the most prominent ways to build high performance microprocessors capable of handling large volumes of data. An increasing trend has been observed in using multi-core and multiprocessor for high performance computing applications such as weather forecasting, astronomy, energy applications and supercomputers applications. In the

DOI: 10.4018/978-1-5225-3142-5.ch011

"Big Data" Distributed Computing research area, some of the main challenges have been to address the questions: (1) How Big Data can be processed in a timely fashion and with faster computational efficiency in a multi-core environment using the current state-of-the-art technology? (2) What is the impact of computations on multi-core and multi-processors shared memory systems on the performance of Massive Parallel Processing on Big Data environment?

In recent years, multi-core processors have been considered as the standard for all computing systems ranging from handheld devices to high-end server farms. On the other hand, parallel programming techniques that could exploit more than one processor have been advocated as promising solutions to handle large volumes of data in distributed systems. However, utilizing the underlying hardware in an effective way and exploiting the parallelism from multiple cores are still open challenges.

The existing parallel programming paradigm that uses traditional programming techniques such as Message Passing Interface (MPI) and shared memory implementations for large scale computing are far from being widely adopted and the processing of large volumes of data often requires the use of specialized hardware and software tools as well as distinct programming and analytical skills. The Big Data MapReduce and Hadoop techniques are meant to provide high throughput, fault tolerance and efficient processing and utilization of large volumes of data. As such, they have significant advantages over traditional parallel processing. This chapter presents some of the Big Data architectural challenges in the perspective of multi-core processors.

The remainder of the Chapter is organized as follows. The Big Data concept is introduced followed by our proposed chip multiprocessors architecture model for Big Data. Next, a case study of a Big Data scenario using Massively Parallel Processing (MPP) is presented along with some concluding remarks.

WHY BIG DATA MATTERS IN THE PERSPECTIVE OF MASSIVELY PARALLEL PROCESSING

In the ever changing landscape of chip industries, the data collected from different computing systems need to be managed and the process information from multiple sources and various formats of data such as "Structured, Semi, Quasi-structured and Unstructured data" also needs to be examined. Examples of data types constituting Big Data include: transaction of data, OLAP, spreadsheet and relational databases, wireless sensory data to name a few. An unstructured data type has no inherent data structure and it is usually stored in different formats including PDF, textual files, videos and images to name a few. On the other hand, semi-structured data types are textual data file patterns that enable parsing. Examples of such data types include XML data files and HTML tagged texts. Quasi-structured data types can be described as textual data with erratic data formats that require more formatting efforts, tools and time. Web click streams are types of data that may contain some inconsistencies in their values and content.

Big Data is the Buzz word everywhere. It does not only mean large volumes of data but it also represents a variety of data that needs to be stored, managed and processed beyond the state-of-the-art technologies in multi-core CPUs. Massively Parallel Processing is a technology that deals with processing large volumes of datasets with fast processing speeds using multiple processors in which current technology may not necessarily be able to process. When handling Big Data, the data collected for analysis should be linked, matched against some benchmark data, cleansed and transformed across the systems so that possible relationships between data features, hierarchies and multiple data linkages are correlated. Therefore, data management is a complex process especially when large volumes of data are