Chapter 36

Scheduling Large-Scale DNA Sequencing Applications

Sudha Gunturu
Oklahoma State University, USA

Xiaolin Li
Oklahoma State University, USA

Laurence Tianruo Yang
St. Francis Xavier University, Canada

ABSTRACT

This chapter studies a load scheduling strategy with near-optimal processing time that is designed to explore the computational characteristics of DNA sequence alignment algorithms, specifically, the Needleman-Wunsch Algorithm. Following the divisible load scheduling theory, an efficient load scheduling strategy is designed in large-scale networks so that the overall processing time of the sequencing tasks is minimized. In this study, the load distribution depends on the length of the sequence and number of processors in the network and, the total processing time is also affected by communication link speed. Several cases have been considered in the study by varying the sequences, communication and computation speeds, and number of processors. Through simulation and numerical analysis, this study demonstrates that for a constant sequence length as the numbers of processors increase in the network the processing time for the job decreases and minimum overall processing time is achieved.

INTRODUCTION

Large-scale network-based computing has attracted tremendous efforts from both academia and industry because it is scalable, flexible, extendable, and economic with wide-spread applications across many disciplines in science and engineering. To address scalability issues for an important class of applications, researchers proposed a divisible load scheduling theory (DLT). These applications are structured as large numbers of independent tasks with low granularity (Bharadwaj, V., Ghose, D., & Thomas Robertazzi, G., 2003). They are thus amenable to embarrassingly parallelism, typically in master-slave fashion.
Such applications are called divisible load because a scheduler may divide the computation time among worker processes arbitrarily, both in terms of task and task sizes. Scheduling the tasks of a parallel application on the resources of a distributed computing platform efficiently is critical for achieving optimal performance (Bharadwaj, V., Ghose, D., & Mani, V., 1995).

The load distribution problem in distributed computing networks, consisting of a number of processors interconnected through communication links, has attracted a great deal of attention (Sameer Bataineh, Te-Yu Hsiung, & Thomas Robertazzi, 1994). Divisible Load Theory (DLT) is a methodology that is involved in the linear and continuous modeling of partitioning the computation and communication loads for parallel processing (Robertazzi, T.G., 2003). DLT is primarily used for handling large scale processing on network based systems.

The DLT paradigm has demonstrated numerous applications such as edge detection in image processing, file compression, joining operations in relational databases, graph coloring and genetic searches (Wong Han Min., & Bharadwaj Veeravalli, 2005). Some more examples of real divisible applications include searching for pattern in text, audio, graphic files, database and measurement processing, data retrieval systems, some linear algebra algorithms, and simulations (Maciej Drozdowski., & Marcin Lawenda., 2005).

Over the past few decades research in the field of molecular biology has made advancement that is coupled with advances in genomic technologies. This has led to an explosive growth in the biological information generated, in turn, led to the requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.

In this chapter a parallel strategy is designed to explore the computational characteristics of the Needleman-Wunsch algorithm that are used for biological sequence comparisons proposed in the literature. In designing the strategy the load is partitioned among the processors of the network using the DLT paradigm (Bharadwaj, V., Ghose, D., & Mani, V., 1995).

Two commonly used algorithms for sequence alignment are the Needleman-Wunsch Algorithm and Smith-Waterman Algorithm where the former is employed for Global Alignment and the latter is used for Local Alignment. The complexity of the Needleman-Wunsch Algorithm and Smith-Waterman Algorithm to align sequence of length x is given by $O(x^2)$ (Wong Han Min., & Bharadwaj Veeravalli, 2005).

The algorithm used in this study is the Needleman-Wunsch Algorithm. The way that has been adopted in this study to for parallelizing the Needleman-Wunsch Algorithm is by computing the matrix elements in diagonal fashion by using a Multiple Instruction Multiple Data Systems.

Divisible Load Theory is employed for handling the sequence alignment. The objective is to minimize the total processing time for sequence alignment. The partition of the load depends primarily on the matrix that is generated by the Needleman-Wunsch Algorithm. The network has been studied for variable link speed and constant link speed.

**RELATED WORK**

The merging of the two rapid advancing technologies of molecular biology and computer science resulted in a new informatics science, namely bio informatics (Wong Han Min., & Bharadwaj Veeravalli, 2005). Over the past few years, the interest and research in the area of biotechnology has increased drastically. This area of study deals primarily with the methodologies of operating on molecular biological information. The present days of molecular biology is characterized by collection of large volumes of data.
Related Content

Grid Computing for Ontology Matching
www.igi-global.com/chapter/grid-computing-ontology-matching/64565?camid=4v1

Mechanism for Privacy Preservation in VANETS
www.igi-global.com/article/mechanism-privacy-preservation-vanets/43881?camid=4v1

Scalability of Mobile Ad Hoc Networks
www.igi-global.com/chapter/scalability-mobile-hoc-networks/36430?camid=4v1

A QSQL-Based Service Collaboration Method for Automatic Service Composition, and Optimized Execution
Kaijun Ren, Jinjun Chen, Nong Xiao and Weimin Zhang (2009). Quantitative Quality of Service for Grid Computing: Applications for Heterogeneity, Large-Scale Distribution, and Dynamic Environments (pp. 249-270).
www.igi-global.com/chapter/qsqld-based-service-collaboration-method/28281?camid=4v1