A Dynamic Load Balancing Strategy with Adaptive Thresholds (DLBAT) for Parallel Computing System

Taj Alam, Jawaharlal Nehru University, New Delhi, National Capital Territory of Delhi, India

Zahid Raza, Jawaharlal Nehru University, New Delhi, National Capital Territory of Delhi, India

ABSTRACT

The primary objective of scheduling is to minimize the job execution time and maximize the resource utilization. Scheduling of 'm' jobs to 'n' resources with the objective to optimize the QoS parameters has been proven to be NP-hard problem. Two broad approaches that are defined for dealing with NP-hard problems are approximate and heuristic approach. In this paper, a centralized dynamic load balancing strategy using adaptive thresholds has been proposed for a multiprocessors system. The scheduler continuously monitors the load on the system and takes corrective measures as the load changes. The threshold values considered are adaptive in nature and are readjusted to suite the changing load on the system according to the mean of the available load. Effectively, the load is leveraged towards the mean, transferring only the appropriate number of jobs from heavily loaded nodes to lightly loaded nodes. In addition, the threshold values are designed in such a way that the scheduler avoids excessive load balancing. Therefore, the scheduler always ensures a uniform distribution of the load on the processing elements with dynamic load environment. Simulation study reveals the effectiveness of the model under various conditions.

Keywords: *Centralized Scheduling, Load Balancing, Parallel and Distributed System, Threshold, Turnaround Time*

INTRODUCTION

Parallelism in the computing systems can be viewed at two levels viz. hardware and software. At hardware level, it can be realized in the form of multiplicity of the processing elements and/or functional units whereas for the software level it can be seen as the multiple modules of the job demanding execution that can run in parallel. The efficiency of a parallel system is governed by the degree of matching between the hardware and the underlying software parallelism. More is this match better is the efficiency (Barney, 2012; Foster, 1995; Steen, 2012; Brucker, 2007).

DOI: 10.4018/ijdst.2014010104

The job should be scheduled in such a way that no resources are underutilized and that the turnaround time is minimized. This can be ensured by exploiting the inherent parallelism in the job by distributing the entire workload on the available computational job to run simultaneously. As per (Brucker, 2007) scheduling is defined as the method by which threads, processes or data flows are given access to system resources e.g. processor time, communications bandwidth. Scheduling has been broadly classified into local and global scheduling schemes with global scheduling further divided into static and dynamic schemes. The division further extends hierarchically downward as given by Casavant and Kuhl which gives the proper picture of hierarchical division of scheduling approaches (Casavant & Kuhl, 1988). Further refinement of dynamic scheduling approaches can be seen in Singhal and Shivaratri with flat classification (Shivaratri, Krueger & Singhal, 1992). However, based on above two classifications many approaches have emerged. The core of all these approaches is to achieve the objective functions of designing an optimum schedule. Scheduling of jobs should be done in such a way that each computing node has its proper share of work so that eventually the job turnaround time can be minimized. Dynamic Scheduling is a methodology to distribute workload across multiple computers, network links, central processing units, disk drives, or other resources, to achieve optimal resource utilization, maximize throughput, minimize response time while avoiding overload (Shivaratri, Krueger & Singhal, 1992). Dynamic scheduling approaches are concerned with load balancing resulting in effective utilization of the resources. An effective dynamic scheduling is very important for a system to balance the system effectively and achieve the target quality of services while addressing issues like synchronization, communication overhead, data locality and scalability. In order to achieve the above goals dynamic scheduling must exhibit the following features (Shivaratri, Krueger & Singhal, 1992):

1. Effective information policy for decimations of load information among the computing nodes;
2. Effective transfer policy which will decide the number of jobs to transfer with minimization of communication delays;
3. Effective placement policy as to where to allocate the process in the best possible way.

This paper presents a centralized dynamic scheduling policy based on adaptive threshold values for the job reallocations on a parallel computing system consisting of multiprocessors. As the load on the system is bound to change dynamically, defining thresholds for the workload on the system with adaptive nature is a must. The system reacts by redistributing the load as soon as these threshold values get crossed, thereby ensuring a uniform distribution of the workload. The core of the policy is its approach to bring the load around the mean of available workload, with least number of job transfers in the minimum possible time.

The remainder of the paper is organized as follows. In second section, related work pertaining to the dynamic scheduling is discussed with emphasis on DLB approach suggested by Lan Youran. Section three presents the proposed scheduler and the working of the model. Section four presents the illustrative example for better understanding the model. Section five discusses the simulation study and the implementation of the model on Sun Fire x4470 server and its comparison with the similar approaches. Finally, the paper ends in section six with the concluding remarks.

RELATED WORK

The issue of load balancing has gained attention of many researchers. A number of load balancing strategies using various approaches have been reported in the literature, with discussion further extending towards the application of load balancing in various emerging fields of parallel