Chapter 1

Introduction
This chapter introduces the context of the research. It starts with our motivation toward the present work followed by a description of general area of load balancing in parallel and distributed computing and objectives of the thesis. Finally an overview of the entire thesis has been given.

1.1 Motivation

What would be our first reaction if we join a long queue in railway reservation ticket counter or in a bank for transactions? Obviously we try to find a small queue or try to join the queue which is being served quickly.

The major cause behind load balancing is that there is lot of distributed processing power remains unused in the distributed networks. So instead of overloading a few nodes with huge numbers of tasks, it is always better to allocate some of these tasks under loaded nodes to finish the tasks as early as possible without affecting the other users on other machines in a network.

Other motivations for load balancing include providing high availability, signifying reconfiguration and utilizing special machine capabilities. Higher availability can be achieved by migrating tasks from the machines that are more prone to failure to the more reliable machines. Thus the process will continue to run even after the failure of machines or disconnection from network. The availability is indispensible during the process reconfiguration as some processes are needed to be continually operational (e.g. servers). At last, some machines have high processing capability in a distributed network and these machines should always be targeted for processing purpose although the processes can be started on regular machines.

In short, there are many motivations for load balancing but the goal of the load balancing in distributed computing is to improve performance by distributing tasks among the processors to utilize idle processing power.
1.2 Introduction

As the domain of applications is continuously growing day by day, the present and future era need more and more powerful computing systems. Some applications like large-scale parallel and distributed systems for database, real time and large-scale scientific and commercial applications need huge computational power to finish the computation within the calculated time. So to accomplish these types of applications, the concept of parallel and distributed computing has been emerged. In parallel distributed computing systems, a number of processors are connected by a computer network which consists of cheap and easily available autonomous systems like workstations or PCs [1, 2]. Distributed computing system provides high performance environment that are able to provide huge processing power [3]. One of the most challenging issues is to maintain the load or tasks equally to each node in a network. In a network generally the tasks are not equally distributed over the nodes causing some nodes overloaded, some nodes under loaded and some others even in idle state [1, 2, 4]. It is also noticed in this field that even in pick time of computation slower nodes are over loaded and faster nodes are lightly loaded or even in idle state. So the performance of the network is highly affected due to the uneven distribution of tasks among the nodes present in that underlying network. Thus, it is highly necessary to distribute the tasks equally among the nodes present in a network to get the efficient performance of the system [4, 5]. The distribution of loads to the processing elements is simply called the load balancing problem. Load balancing algorithms guarantee the near equal distribution of tasks among the computing nodes. Load balancing algorithms minimizes execution time, communication delays, response time, over heads and maximizing resource utilization. So load balancing algorithms play most important role to enable efficient use of multi-processor systems by nearly equalizing load among the processors. Equalization of load among the computing nodes does not keep any node in idle state. Even a simple load balancing algorithm may significantly increase the performance of a network than a network having no load balancing algorithm. Thus, without load balancing algorithm the basic objective of a computer network for sharing resources will not be fulfilled effectively.
Chapter I: Introduction

The applications can be partitioned into manageable tasks by applying suitable task partitioning algorithms and the independent tasks may be assigned to any node. There are mainly two load balancing algorithms: static load balancing (SLB) and dynamic load balancing algorithm (DLB). In static load balancing algorithms the assignment of tasks to the processors are done before runtime [6]. After the assignment of jobs no change is possible that means redistribution of tasks is not possible. In dynamic load balancing algorithms tasks are dynamically distributed across the nodes and the redistribution of tasks is always possible [6]. In SLB priori knowledge about the tasks and the system are known, so the task are distributed according to the performance of the nodes and the other factors like communication speed, input output buffer size etc. remains unchanged. Once assignment of tasks is done the reassignment is frizzed, so the communication over-heads is negligible here. But in DLB jobs are reassigned at the runtime; the load will be transferred from heavily loaded nodes to the lightly loaded or idle nodes. So considerable communications over-heads occur and become more when number of processors increases.

Quality of a load balancing algorithm is dependent on two factors. Firstly number of steps that are needed to get a balanced state. Secondly, the extent of load that moves over the link to which nodes are connected [4].

1.3 Thesis Objectives

The objective of this thesis is to develop efficient algorithms for dynamic load balancing in distributed computing systems. Direction of this research leads to the development of some fast and efficient dynamic load balancing algorithms. In this thesis we propose four load balancing algorithms which are fully distributed in nature. In chapter 6, we propose An Efficient Diffusion Load Balancing Algorithm in Distributed System which only considers the interested nodes for load balancing to participate in load balancing. In chapter 7, we propose An Improved Local Hierarchical Load Balancing Algorithm (ILHLBA) in Distributed Computing which removes the unnecessary overheads caused in previously existing algorithm. In chapter 8, we propose An Efficient Local Hierarchical Load Balancing Algorithm (ELHLBA) in Distributed Computing which
Chapter 1: Introduction

uses the parents of leaf nodes as front-end nodes and improves the performance. In chapter 9, we propose A New Distributed Load Balancing Algorithm which maintains the balanced load state among nodes and targets the more powerful nodes for processing.

1.4 Organization of the Thesis

This thesis consists of nine chapters and the organization of the thesis is as follows:

Chapter 1 presents a brief introduction of the research problems to be addressed in this thesis. It starts with an introduction to the general area of load balancing in parallel and distributed computing, motivation and objectives of the thesis. Finally an overview of the entire thesis has been given.

Chapter 2 gives an overview on parallel computing that depicts all the aspects of parallel computing viz. parallel computing hardware, parallel computing model, parallel programming design, and implementation, software support for parallel programming, parallel performance measurement and parallel benchmarks used in HPCC. The parallel computing is based on the special hardware and software. The hardware and the software are specially designed for parallel algorithm and programming. This paper explores all the aspects of parallel computing and its usefulness.

Chapter 3 gives overview distributed computing. This chapter presents the difference between parallel and distributed computing, terminologies used in distributed computing, task allocation in distributed computing and performance parameters in distributed computing system, parallel distributed algorithm models, advantages of distributed computing and scope of distributed computing.

Chapter 4 presents a survey on load balancing algorithms in distributed computing. A study on the load balancing strategies has been done lucidly in detail. A comparison has been made between SLB and DLB introducing some new parameters. Some important dynamic load balancing algorithms has been analyzed and their comparison has been evaluated to focus their importance in different situations.
Chapter 5 gives a review on wide range of parallel programming simulation tools. This chapter also compares the features of different tools with a particular focus on Parallel Virtual Machine (PVM) and Message Passing Interface MPI which is mostly used in today’s parallel and distributed computing system.

Chapter 6 presents a dynamic load balancing algorithm titled “An Efficient Diffusion Load Balancing Algorithm in Distributed System”. In this chapter it has been shown that the number of communication overheads depends on the number of overloaded nodes present in the domain of an under loaded nodes and vice-versa. It has been also shown that communication overhead for load balancing is always fairly less than $KN$ but in worst case this algorithm’s complexity becomes equal to $KN$.

Chapter 7 presents another algorithm named “An Improved Local Hierarchical Load Balancing Algorithm (ILHLBA) in Distributed Computing”. ILHLBA is compared with another existing algorithm LHLBA and the simulation results show that ILHLBA produces better result than LHLBA in respect of both response time and throughput against system utilization.

Chapter 8 represents another algorithm titled “An Efficient Local Hierarchical Load Balancing Algorithm (ELHLBA) in Distributed Computing” which is the extended version of ILHLBA. ELHLBA is compared with ILHLBA and LHLBA. The simulation results show that ELHLBA produces better result than ELHLBA and LHLBA in respect of response time and throughput against system utilization.

Chapter 9 presents a load balancing algorithm in distributed computing system titled “A New Distributed Load Balancing Algorithm”. It has been assumed that each node will maintain a local load table to hold the load status of immediate neighbors. The aim of this algorithm is to achieve balanced load among the processors according to their speed of computation and also to reduce communication over heads. This algorithm also targets most powerful nodes for load transfer in the system. Performance of this algorithm shows better performance over previously existing algorithm.
Finally conclusions, references and list of publications have been provided at the end of this thesis.