CHAPTER 8

CONCLUSIONS AND FUTURE SCOPE

In this Chapter,

8.1 Major Contributions
8.2 Limitations and Future Scope
8.3 Concluding Remark

8.1 MAJOR CONTRIBUTIONS

Distributed Computing Environment (DCE) provides collaborative infrastructural support to satisfy the large computational needs of scientific and business applications. Various participating hardware and software resources in distributed computing environment are heterogeneous in nature and make resource management a challenging task. Dynamic load balancing is a distributed scheduling technique used to improve the performance of the distributed computing elements. It ensures the effective utilization of the processing capabilities of the nodes by even work load distribution.

The vision of this research work is to utilize the resource in heterogeneous environment effectively with the belief that optimum resource utilization improves the overall performance of the systems under consideration. Throughout the research work, attempts have been made to develop dynamic load balancing algorithms, exploiting heterogeneous hardware resources to their maximum potential. Although the proposed dynamic load balancing models can be applied to the systems wherever workload distribution among the heterogeneous resources are required, the proposed solutions are highly effective especially in the area of clusters, routers and multi-core nodes. The Major
contributions arising out of the research are being summarized in following subsections:

(a) Content Aware Load Balancing in Heterogeneous Web servers: This research proposes a dynamic load balancing algorithm for scalable heterogeneous web server cluster using content awareness. The algorithm uses server's processing capability, queue length, utilization ratio as load indices. As the cluster supports multiple services at the primary level, we have used content aware forwarding algorithm and at the secondary level, waited round robin algorithm has been used. To implement content-aware load balancing, web requests are classified into static requests, transaction requests, secure transaction requests and multimedia requests. Further the classification is on the basis of I/O bound and CPU bound jobs or the combination of both. In the web server cluster, the servers are also categorized on the request basis i.e. static servers, transaction servers, secure transaction servers and multimedia servers and respond the requests of that category only. To distribute the workload evenly among the nodes of the web server cluster, heterogeneity factor of the nodes has been considered. Servers of the WSC respond to client's requests directly using one-way mechanism to reduce the workload of the scheduler.

The performance of proposed algorithm has been compared with weighted round robin and Lin's algorithm. For simulation set of 400, 800 ....4000 requests in each category are sent to the WSC and their mean response time parameter is used for performance comparison. The results show that proposed algorithm gives better result as the load on the WSC increases. For lightly loaded server clusters, WRR algorithm can be used which is less complex than the other two algorithms and hence reduces the complexity of the scheduler. However, for large workload, the proposed algorithm can be used effectively [Tiwari, 2010b].

(b) Enhanced Dynamic Load Balancing in Heterogeneous Cluster using Linear Programming: The research addresses a load balancing policy in heterogeneous multi server environment. All the servers of the WSC are of
different service rates and each server can serve all types of client requests. Requests are categorized into classes with varying service requirements and each category request is assured a reserved service rate. As optimal resource utilization is one of the objectives of the Linear Programming (LP) model, load balancer uses deterministic LP model to identify the resource of the WSC for optimal allocation of client requests. In the second stage, master node of each group acts like a load balancer and uses sender initiated technique to balance the load among the nodes of the WSC. The throughput of the WSC is used to compare the proposed model with three popular scheduling techniques random, round robin and least loaded.

The proposed model is simulated for three sets of different category requests on cluster of heterogeneous groups comprising of nine nodes. The throughput of the WSC is used to compare the proposed model with three popular scheduling techniques random, round robin and least loaded. The WSC throughput is compared using same set of input for different scheduling technique. The simulation results show that the proposed model performs better for all categories of requests and therefore justify the use of linear programming for scheduling client requests at load balancer level. The proposed model is not a solution to a specific problem but can be applied the variety of situations where scheduling is required [Tiwari, 2013c].

(c) An Open Source S/Ws based DLB Model for Distributed Computing Environment: The research highlights the significance of open source software for distributed computing environment. Available open source software are identified for distributed computing environment and proposed a need of incorporation of dynamic load balancing feature in the commonly used Open source software. The work also proposed an adaptive load balancing model using open source software where different load balancing strategies are chosen for lightly loaded, moderately loaded and heavily loaded systems.
The simulation was carried out on real network on Red Hat Cluster Suite with Apache tomcat and My-SQL open source software being used by the WSC. The proposed model consists of Internal Proxy, Decision Maker (DM), Load Data Base (LDB) and Load Controller (LC) modules and these modules are developed using multi threaded approach of Java programming. The proposed model is compared with random, round robin and weighted round robin algorithms. The same set of input requests is simulated for random, round robin, weighted round robin and proposed algorithm and mean response time and mean processor utilization parameters are calculated for all the nodes of the WSC. Simulation results show improved performance with the proposed model. The model can be useful for incorporation of load balancing feature in existing and new open source software [Tiwari, 2009b; Tiwari, 2012].

(d) A Linear Programming Based DLB Model for Router: The research highlights recent advances and design issues of state-of-art routers and routing algorithms. A Linear Programming (LP) based dynamic load balancing model has also been proposed for routers.

For simulation, three sets of inputs, each consisting different category requests with reserved forwarding rate ratio have been considered. The proposed model is compared with popular load balancing techniques viz. random and round robin. Simulation uses artificial load comprising of various categories of requests. To begin with, the performance of the proposed model is tested on a single router. Instantaneous throughput $Z$ have been calculated using random scheduling, round robin scheduling and proposed LP model for the three given sets of inputs on a single router. Thereafter, the proposed model is tested on a Java network simulator. The routers of the network simulator use one scheduling technique out of random, round robin and proposed LP model for load balancing. All the links used in the simulator are homogeneous and shortest path routing metric is used by the switching fabric of the routers to route the outgoing packets. For experimental purpose, simulator uses three inline cards, three outline cards and three virtual input queues in each of the routers and backbone links used by the simulator are
27. The observation of the simulation shows that, as the number of routers in the network increase, the proposed model gives better results [Tiwari, 2009a; Tiwari, 2013b].

(e) Performance Enhancement of Symmetric Multicore Processors using DLB: The research proposes an adaptive load balancing model for even workload distribution among the cores of symmetric multi-core servers by using two state-of-art framework, one for I/O and another for parallel programming.

To compare the performance of proposed work, various experiments are executed using proposed DLB algorithm without external load balancing. The experiments include Matrix Multiplication, Merge Sort and Fibonacci series. For matrix multiplication, iterative method was used whereas merge sort uses divide-and-conquer algorithm to divide the list into two equal sub-lists until the sub list is reduced to two elements. For each experiment, number of subtasks executed on each core and the total execution time with and without DLB is collected. Experimental results show that the proposed model is efficient for large tasks and all the processes finish almost at the same time which improves the overall performance of the multi-core processors.

8.2 LIMITATIONS AND FUTURE SCOPE

The thesis presents solutions to several problems related to resource management in clusters, routers and multi-core processors. Although most of the objectives have been achieved, yet few issues are still under study and several new issues related to load balancing in distributed computing environment were identified during the research work. The limitation and future scope of the research work is as follows:

(a) Adaptive Dynamic Load Balancing Algorithms for Resource Management: Most of the load balancers use the same complex DLB algorithm for all load situations. It has been observed that the performance of these algorithms is better for moderate and heavily loaded system but for lightly loaded
system the performance of these proposed algorithms is poor than even random scheduling on certain occasions. The issue can be considered as future research work.

(b) Queuing Theory and Game Theory based DLB Models for DCEs: The thesis proposes Linear Programming based DLB algorithm for load balancer where load balancer uses deterministic LP model to identify the resources of the WSC for optimal allocation of client requests. Queuing Theory and Game Theory models can also be applied and tested as the assumption of these models satisfy the criteria for DLB.

(c) Priority and Link Level based DLB for Routers: A Linear Programming based dynamic load balancing model has also been proposed for routers in this research work. The model proposes a virtual input queue and balances the incoming traffic of input queues among the virtual input queues. Second level load balancing can be applied at link level as the links are also heterogeneous i.e. the packet carrying cost of different links are different. Thus the combination of internal and external load balancing can be more effective.

(d) Performance Enhancement of Asymmetric Multicore Servers: The research work proposed a DLB model for symmetric multi-core processors. The situation is more challenging when the cores in multi-core processors are heterogeneous i.e. all or some cores are of different processing capability. Special DLB algorithms can be designed to distribute the workload among the cores of such processors.

(e) DLB Strategies for Clouds and Grids: Clouds and girds are becoming increasingly popular alternative to custom built parallel computers. Clouds popularity is based on the ability to offer cost effective environment for running all category of requests. Various DLB strategies can be proposed for Cloud and grid environment.
Management of Information Overload in Distributed Computing Environment: The rapid growth and improved accessibility to Internet has presented users with huge amount and variety of information. The increasing Information through various sources causes information overload problem in distributed computing environment. The nature and causes of information overload can be identified and Information measurement tools can be developed which measures the Information in quantitative way rather than qualitative.

8.3 CONCLUDING REMARK

The results obtained by scheduling and resource management algorithms proposed in the thesis prove that the dynamic load balancing is a flexible, cost effective and reliable strategy to support distributed scheduling without modifying the systems kernel or underlying hardware without deploying expensive servers and workstations. The work presented in the thesis has a wide scope in the field of distributed computing system and includes cluster, grids and clouds. The thesis contributes in three major areas: Clusters, Routers and Multicore Processors. In addition to the theoretical interest, algorithmic methodologies suggested in the proposed work have potential applications in practical fields in distributed operating systems, which are under development stage and can be incorporated in existing open source software like Linux, Virtual Linux Server, and Red Hat Cluster Suit etc. The algorithms can be useful in increasing the resource utilization of Internet infrastructure and to minimize the response time of client requests without adding extra hardware or software cost as our main emphasis was to utilize the existing infrastructure rather than its upgradation.