Chapter 34
Queuing Analysis of Cloud Load Balancing Algorithms

Santosh Kumar Majhi
VSS University of Technology, Department of Computer Science and Engineering, Burla, India

Shweta Bhuyan
VSS University of Technology, Department of Computer Science and Engineering, Burla, India

Shankho Subhra Pal
VSS University of Technology, Department of Computer Science and Engineering, Burla, India

Sunil Kumar Dhal
Sri Sri University, Faculty of Management Studies, Cuttack, India

ABSTRACT

The emergence of cloud-computing and the apparent shift to this new paradigm has led to the creation of data centres that consist of hundreds of thousands of servers. The Cloud is a distributed system that helps share data and provides resources to the users. The data and the distributed resources are stored in the open environment. This paper presents a model of cloud load balancing using queuing and probability theory. A queuing cloud model is discussed with load balancing perspective. We present analysis for two servers and then extended it to n server. In addition, an optimal strategy is modelled for cloud load balancing. The analytical results are verified through numeric simulation.

INTRODUCTION

The emergence of cloud - computing and the apparent shift to this new paradigm has led to the creation of data centres that consist of hundreds of thousands of servers. The Cloud is a distributed system that helps share data and provides resources to the users. The data and the distributed resources are stored in the open environment. The amount of data storage increases rapidly in this environment. Hence several data centres are installed at different geographical locations that are connected over the Internet in order to optimally serve a request by the client. But since the maintenance and the deployment cost of a data centre is extremely high, achieving high utilisation is essential. Load balancing addresses this issue and provides high resource utilisation and better response time.

DOI: 10.4018/978-1-7998-5339-8.ch034

The main aim of load balancing algorithm includes higher availability of resources, increase in the number of services with minimal resource utilization, increasing ease of use for user, reduction in the service time and waiting time of different service requests, performance improvement, system stabilization, accommodation facility for system upgradation and building of fault tolerant systems (Zhao et al., 2016; Aslam & Shah, 2015; Xu et al., 2013; Panda & Jena, 2015).

The data centre, the backbone of the cloud environment privileges the different services by utilizing the load balancing algorithms. Moreover, a data centre consists of multiple racks. The rack contains multiple slots which house the resources (servers) stacked one upon the other (Fang et al., 2016). A request to be served is first kept at a queue before being sent to a data centre. A load balancing algorithm (LB1) decides which data centre is to be chosen. After assignment to the data centre, the request resides in a local queue maintained at the data centre. This queue is accessed by the Task Management Server (TMS) that distributes requests from the queue optimally among the racks. Again, a load balancing algorithm (say LB2) is at play for efficient distribution. Each rack similarly maintains a queue before sending the request to a particular slot for servicing with the application of a load balancing algorithm (LB3) to ensure higher resource utilisation. Thus, each rack and the TMS both maintain a queuing system to handle tasks according to the FCFS principle. A similar structure is hence observed as the request moves from the client to the server servicing the request.

In this paper, a mathematical model is proposed that dynamically coordinates load distribution among the different racks in a data centre by using of an optimal load balancing algorithm (M / M / k queues). It is assumed that each rack behaves similarly i.e. are identical. In addition, each service requires same processing time and the same amount of resources. The model assumes request arrivals occur at rate λ_i by Poisson process and service time has an exponential distribution with the mean service rate of each rack being μ_i (where 'i' indicates the rack number). It also considers that there is no delay due to scheduling or transfer of requests.

LOAD BALANCING ALGORITHMS

Cloud computing is the recent paradigm of large scale distributed computing. Optimal resource utilisation is the necessary prerequisite of parallel and distributed systems. Load balancing is the method that distributes the workload among servers, data centres, hard - drives or other computing resources which are regarded as 'nodes' in the given environment such that a node is neither over - whelmed nor under - utilised. The unpredictability in the cloud environment is caused due to the dynamic behaviour of the cloud which makes load balancing a crucial concern in cloud computing. Load Balancing is nothing but an optimisation problem and an efficient load balancing algorithm is the one that adapts itself to the stochastic environment and the diverse tasks. However, allocating the request uniformly across the nodes is considered to be an NP - complete problem (Baca, 1989).

Since scheduling is NP – Hard optimisation problem, so (Babu & Krishna, 2013) have used a technique inspired by Honey Bee Behaviour, to schedule the load among different VM (Virtual Machines).

(Zhang & Zhang, 2010) has used complex network theory and Ant Colony inspired optimisation for load balancing in cloud – computing.

Response time and waiting time are the performance metrics of load balancing algorithms. In cloud computing environment, they are classified either as immediate mode scheduling or batch mode scheduling (Gopinath & Vasudevan, 2015). Immediate mode organises on basis of arrival of the requests while