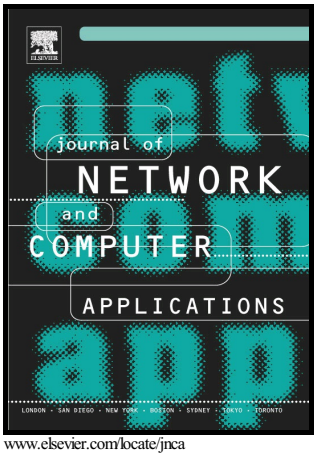
A TAXONOMIC SURVEY ON LOAD BALANCING IN CLOUD

Avnish Thakur, Major Singh Goraya



www.elsevier.com/locate/jnca

PII: S1084-8045(17)30282-5
DOI: http://dx.doi.org/10.1016/j.jnca.2017.08.020
Reference: YJNCA1966

To appear in: *Journal of Network and Computer Applications*

Received date: 17 April 2017
Revised date: 5 August 2017
Accepted date: 25 August 2017

Cite this article as: Avnish Thakur and Major Singh Goraya, A TAXONOMIC SURVEY ON LOAD BALANCING IN CLOUD, *Journal of Network and Computer Applications,* http://dx.doi.org/10.1016/j.jnca.2017.08.020

A TAXONOMIC SURVEY ON LOAD BALANCING IN CLOUD

Avnish Thakur[a], Major Singh Goraya[a]
[a]Sant Longowal Institute of Engineering and Technology, Longowal, India

Abstract: Cloud computing aims to provide seamless computing services to the millions of consumers across the world. Datacenter, the engine of cloud computing, hosts large scale computing resources (hardware and software) at the backend of cloud. In the recent years, the rising demand for cloud computing services is increasing the load on datacenters. Moreover, cloud computing environment is highly dynamic in terms of workload patterns and system behaviour, which often leads to load imbalance across the datacenter resources. Due to load imbalance, some of the datacenter resources may get overloaded/underloaded, which leads to performance degradation and resource wastage. Load balancing helps to maximize resource utilization and achieve desired Quality of Service (QoS) in cloud by employing optimal resource allocation and workload distribution approaches both at schedule time and runtime. Recognizing its relevance and importance, a great deal of research has been attributed to the issue of load balancing in cloud and many load balancing algorithms has been proposed in literature. This paper presents a systematic survey of the state-of-the-art load balancing algorithms proposed for cloud computing environment. The survey is systemized by proposing a novel taxonomy of load balancing algorithms in cloud. For the ready reference of the readers, a brief summary and a comparative analysis of various promising load balancing algorithms is also included in the survey. The survey also discusses various issues and challenges along with trail to their possible solutions.

Keywords: Cloud computing, Load balancing, Resource utilization, Energy efficiency, Quality of Service

1 Introduction

Cloud computing [1-7] is an Internet based service delivery model for providing on-demand access to the shared pool of computing resources (software/hardware). In the computing world, cloud computing evolves to provide everything-as-a-service [4]. In the present time, various cloud based computing services are being offered to the consumers as utility services like electricity and telephony. On-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service are some of the essential characteristics of cloud computing [5]. Cloud service providers follow "pay-as-you-use" pricing model where consumers need to pay only for the resources/services they use. It leverages the consumers to focus on their primary objectives rather than worrying about their computing infrastructure needs. The computing infrastructure needs of the consumers are fulfilled by the cloud service providers.

Cloud computing from its very origin has evolved to provide strategic and economic benefits to both the service providers and service consumers. Cloud services are implemented through datacenters at the back-end. Datacenters host large scale computing resources to provide seamless computing services to the consumers. Due to continuous increase in the demand for cloud services, the QoS [8,9] and efficient utilization of the datacenter resources [10-13] has become a matter of great concern. Continuous hike in the number of cloud service consumers may lead to the scarcity of datacenter resources which will affect the quality of cloud services. In the context of increased service requests, the desired QoS can be maintained by efficiently utilizing the datacenter resources. The efficient utilization of the resources is further an important economic consideration in the datacenter deployment. Resource allocation and balancing of workload in the datacenter are important aspects of