Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 47 (2015) 30 – 36

Procedia Computer Science

Dynamic Resource Allocation Scheme in Cloud Computing

Saraswathi AT a, Kalaashri.Y.RA b, Dr.S.Padmavathi c[1]

a UG Student, Thiagarajar College of Engineering,Madurai,625015, India.
b UG Student, Thiagarajar College of Engineering,Madurai,625015, India.
c Associate Professor, Thiagarajar College of Engineering,Madurai,625015, India.

Abstract

Cloud Computing environment provisions the supply of computing resources on the basis of demand, as and when needed. It builds upon advances of virtualisation and distributed computing to support cost efficient usage of computing resources, emphasizing on resource scalability and on-demand services. It allows business outcomes to scale up and down their resources based on needs. Managing the customer demand creates the challenges of on-demand resource allocation. Virtual Machine (VM) technology has been employed for resource provisioning. It is expected that using virtualized environment will reduce the average job response time as well as executes the task according to the availability of resources. Hence VMs are allocated to the user based on characteristics of the job. Effective and dynamic utilization of the resources in cloud can help to balance the load and avoid situations like slow run of systems. This paper mainly focuses on allocation of VM to the user, based on analyzing the characteristics of the job. Main principle of this work is that low priority jobs (deadline of the job is high) should not delay the execution of high priority jobs (deadline of the job is low) and to dynamically allocate VM resources for a user job within deadline

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the Graph Algorithms, High Performance Implementations and Applications (ICGHIA2014)

Keywords: Cloud computing; Resource allocation; Virtualization; Lease type; Priority; Pre-emption.

1

*Corresponding author Tel.: +91-948-682-3139

E-mail address: spmcse@tce.edu

1877-0509 © 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the Graph Algorithms, High Performance Implementations and Applications (ICGHIA2014)
doi:10.1016/j.procs.2015.03.180

1. Introduction

A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource based on Service Level Agreements (SLA) established through negotiation between the service provider and consumers. Cloud computing is an internet-based computing in which large groups of remote servers are networked to allow sharing of data-processing tasks, centralized data storage, and an online access to computer services or resources. It relies on sharing of resources to achieve coherence and economies of scale, similar to a utility (like the electricity grid) over a network. Cloud computing also focuses on maximizing the effectiveness of the shared resources. Cloud resources are not only shared by multiple users but are also dynamically re-allocated on demand. The main enabling technology is virtualization. Virtualization software allows a physical computing device to be electronically separated into one or more "virtual" devices, each of which can be easily used and managed to compute tasks. Virtualization provides the agility required to speed up IT operations, and reduces cost by increasing infrastructure utilization.

Scheduling is an important of any operating system. CPU scheduling deals with problem of deciding which of the processes in the ready queue is to be allocated CPU time. When a job is submitted to a resource manager, the job waits in a queue until it is scheduled and executed. The time spent in the queue, or wait time, depends on several factors including job priority, load on the system, and availability of requested resources. Turnaround time represents the elapsed time between when the job is submitted and when the job is completed. It includes the wait time as well as the jobs actual execution time. Response time represents how fast a user receives a response from the system after the job is submitted. Resource utilization during the lifetime of the job represents the actual useful work that has been performed. System throughput is defined as the number of jobs completed per unit time. Mean response time is an important performance metric for users, who expect minimal response time.

In a typical production environment, many different jobs are submitted to cloud. So, the job scheduler software must have interfaces to define workflows and/or job dependencies, execute the submitted jobs automatically. The cloud broker has pre-configured and stored in the cloud all the necessary VM images to run users' jobs. All the incoming jobs are enqueued into a queue. A system-level scheduler, running on a dedicated system, manages all the jobs and a pool of machines, and decides whether to provision new VM from clouds and/or to allocate jobs to VMs. The scheduler is executed periodically. At each moment, the scheduler performs five tasks: (1) Predicting future incoming work-loads; (2) Provisioning necessary VMs in advance, from clouds; (3) Allocating jobs to VM; (4) Releasing idle VMs if its Billing Time Unit (BTU) is close to increase; (5) If the time of un-allocated jobs is high, starting the necessary number of VMs.

Cloud computing builds upon advances of virtualisation and distributed computing to support cost efficient usage of computing resources, emphasizing on resource scalability and on-demand services. Cloud computing allows business outcomes to scale up and down their resources based on needs. Managing the needs of the customer creates the challenges of on-demand resource allocation. Virtual machine technology has been employed for resource provisioning. Hence VM are allocated to the user based on characteristics of the job. Low priority jobs should not delay the execution of high priority job. This scenario leads to resource contention between low and high priority jobs to access resources. The outcome of the paper is priority-based preemption policy that improves resource utilisation in a virtualised environment.

The remainder of this paper has been organised as follows. Section 2 gives a brief review of related works regarding resource allocation in cloud environment. Section 3 presents a proposed algorithm for resource allocation and an overview of experimental environment. Section 4 shows the performance analysis of the proposed approach and finally Section 5 concludes the paper.

Download English Version:

https://daneshyari.com/en/article/490028

Download Persian Version:

https://daneshyari.com/article/490028

Daneshyari.com