Autonomic Fault Tolerant Scheduling for Multiple Workflows in Cloud Environment

Kumkum Sharma

Under guidance

of

Ms. Sapna Aggarwal

(HOD, JIET)

Department of Computer Science and Engineering

JIND INSTITUTE OF ENGINEERING AND TECHNOLOGY, JIND

Kurukshetra University, Kurukshetra, Haryana, INDIA.

Abstract

Cloud Computing is becoming an increasingly admired paradigm that owns the characteristics of existing paradigms through strong support for virtualization along with various additional features such as on demand resource provisioning, reduced cost, computing flexibility etc. As, the scientific workflows need a suitable paradigm for deployment and execution in conjunction with high availability of Cloud services. Thus, Cloud is a current benchmark for effective facilitation of the execution of scientific workflows through flexibility of accessible services such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) without allusion to the infrastructure on which these applications are hosted.

For the successful execution of the scientific workflows on Clouds, Cloud platform should be able to manage the faults through autonomic fault tolerant approaches during the scheduling of workflow tasks on Cloud resources. Cloud providers also entail efficient scheduling algorithms to schedule these workflows along with autonomic fault tolerant approaches. Although, Cloud Computing technology has evolved but still some of the key challenges like autonomic fault tolerance and workflow scheduling need to be achieved. Furthermore, a thorough study of failure prediction approaches, fault tolerant techniques and fault tolerant scheduling has been performed. Based on the literature survey, it is evident that the key challenge in scheduling workflow applications on Cloud that needs to be addressed is fault tolerant scheduling in autonomic way.

To achieve the set of challenges for the fault tolerant workflow scheduling, a comprehensive study of workflow scheduling algorithms along with the required set of Quality of Service (QoS) parameters is carried out. To address the assorted challenges and autonomic fault tolerant scheduling for multiple workflows is the main focus of this research work.

Introduction

Distributed computing interconnects the geographically distributed resources such as storage devices, data sources and compute resources utilized by users around the world as single amalgamated resource. Nowadays, a number of new concepts and terms related to distributed computing have surfaced that are promising to deliver IT as a service evolving from Cluster to Grid to Utility and now Cloud Computing. Cloud Computing owns the characteristics of existing paradigms through strong support for virtualization along with various additional features such as on demand resource provisioning, reduced cost, computing flexibility etc.

Cloud Computing is an internet-based computing, whereby shared resources, software and information are provided to computers and other devices on-demand like a public utility.

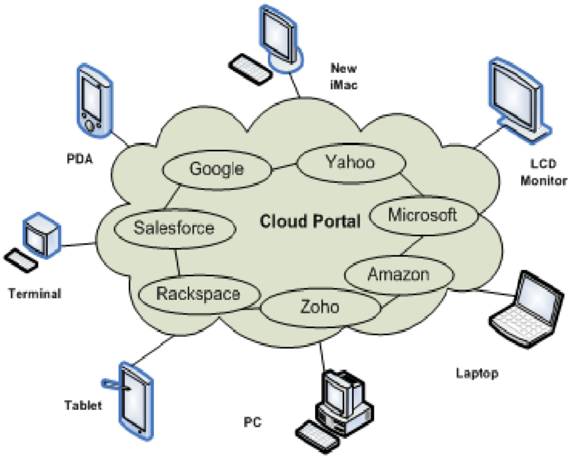
Cloud Computing services are offered by many of the Cloud providers such as Amazon, Yahoo, Google, Microsoft etc. as shown in below figure.

- Public Cloud
- Private Cloud
- Community Cloud
- Hybrid Cloud

Features of Cloud Computing includes following. These are just names and more information can be found from different sources on internet or books dedicated for the studies in the said field.

- Open Access
- Capacity for on-demand infrastructure and computational power
- Improved resource utilization
- Reduced information technology (IT) infrastructure needs
- Resource pooling
- Computing Flexibility
- Mobility

- Reliability



Cloud Computing services can be deployed using following Cloud models:
After a concise introduction to Cloud Computing, the succeeding section confers related technologies such as Autonomic Computing and Workflows which have been utilized for this research work are:

- Automatic Computing
- Automatic Fault tolerance

There are several challenges related to the growth of Cloud Computing such as Reliability and Availability, Integration and Interoperability, Scalable Monitoring of System Components, Efficient Scheduling Heuristics and Resource Management etc. This paper contributes in following fields as per result of experiments conducted in Automatic fault tolerance and Workflow Scheduling algorithm. These only are the focus in this paper.

Literature Review

For fault tolerance: AWS administration the definitive guide and Implementing cloud design patterns for AWS gave a very lucid picture of what is the current state of system and what all it needs in future with complete work done by different organizations and people in this area. The comparative analysis has been performed that analyses these tools for implementing the fault tolerance techniques.

• HAProxy: It stands for High Availability Proxy and is used by companies such as right scale for load balancing and server fail over in the Cloud. It is also being utilized to handle the fault tolerance through various technique such as replication, job migration and proactive fault tolerance in virtual machine environments.

• Assure: introduces rescue points which are locations in existing application codes for handling programmer anticipated failures. It can be used to bypass the path which induces software failures and recover from software failures by using the error virtualization technique to force an error return using an observed value in a function. ASSURE uses a production system and a triage system to implement rescue points and error virtualization.

• Shelp: SHelp is a lightweight runtime system that can survive software failures in the framework of virtual machines.

Hadoop: Hadoop is an open-source java-based software platform developed by the Apache Software Foundation. Hadoop implements Googles Map-Reduce programming model on top of a distributed file system called the Hadoop Distributed File System (HDFS) which is designed to reliably store very large files across machines in a large cluster. These blocks of a file are replicated for fault tolerance as data replication technique.

• Amazon Elastic Compute Cloud (EC2): It provides a virtual computing environment that enables a user to run Linux-based applications. Amazon Web Services (AWS) provides a platform that is ideally suited for building fault-tolerant software systems. The AWS platform is unique because it enables to build fault-tolerant systems that operate with a minimal amount of human interaction and with minimum cost. It can be used to implement replication, SGuard technique for Cloud.

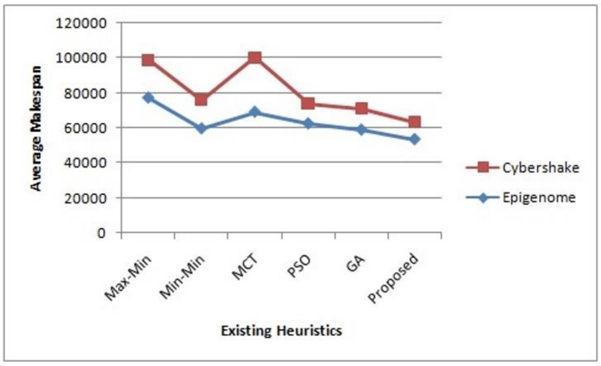
Latest implementation and kind of services and infrastructure provided was found from AWS.

Results and Discussion

The experimental results have been verified and categorized into the following sections.

Autonomic Fault Tolerant Workflow Scheduling Algorithm:

For the comparative analysis, very few approaches are found which have utilized the concept of fault tolerant scheduling in Cloud; none of them has incorporated the implementation of workflow scheduling approaches along with VM migration. As PSO has used to optimize the cost for workflows and GA for optimizing the execution time, but both heuristics do not consider the fault tolerance aspect. Therefore, the comparative analysis of proposed approach has been done after implementing the existing scheduling heuristics with VM migration approach in Cloud environment. The experimental results in Figure 5.21 evidently confirm that the proposed scheduling approach is efficient for large-scale workflows also, such as Cybershake and Epigenome with 1000 jobs. These results have been validated using the existing heuristics such as PSO, GA , Min-Min, Max-Min , MCT and proposed heuristics. It is apparent from the experimental results for Epigenome workflow that the PSO performs better than Max-Min and MCT, whereas GA also enhances the performance over Max-Min, Min-Min, MCT, and PSO. Similarly, for Cybershake workflow, PSO performs better than Max-Min, Min-Min, MCT, and GA.



Furthermore, the hybrid heuristic combines the features of all other heuristics along with VM migration approaches, thus the proposed heuristic outperforms all the existing heuristics by reducing the average makespan for large-scale scientific workflows such as Cybershake and Epigenome.

This section thus verified the maximum accuracy of Naive Bayes by implementing proposed model on Cloud infrastructure. Then, Naive Bayes model has been utilized to predict the task failures. An autonomic fault tolerant technique migrates the faulty virtual machines on other working hosts autonomically by appreciably reducing mean execution time, standard deviation of mean execution time and SLAV. Further, the experimental results for proposed scheduling algorithm for multiple workflows along with autonomic fault tolerant scheduling have been validated on CloudSim and WorkflowSim toolkits by reducing average makespan.

References

- J. Blythe,S. Jain, E. Deelman, Y. Gil, K. Vahi, A. Mandal, and K. Kennedy. Task Scheduling Strategies for Workflow-Based Applications in Grids. CCGrid 2005, 2005.
- Hadoop: The Definitive Guide, 4th Edition by Tom White
- https://aws.amazon.com/
- https://machinelearningmastery.com/