Resource Management based on Agent technology In Cloud Computing

Taha Chaabouni  
MIRACL Lab, FSEG,  
University of Sfax,  
Sfax, Tunisia  
chaabounitaha@yahoo.fr

Maher Khemakhem  
Faculty of Computing and Information Technology,  
University of King Abdulaziz, Jeddah, Saudi Arabia  
maherkhemakhem@yahoo.com

Abstract—Cloud computing is a new concept which combines different technologies like: service oriented architecture, distributed computing, virtualization... It represents a model used to deliver IT capabilities. In cloud computing, everything is provided as a service. We distinguish three main services: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The management of cloud computing resources in IaaS level is a very difficult task to achieve for the service provider. For example, in a simple cloud computing system which is composed of one data center, we easily find thousands of physical resources that are virtualized in millions of virtual resources. So, the manager needs an automatic solution to supervise the whole system in order to avoid or to minimize the human intervention or implication. Consequently, we propose in this paper a solution to this problem based on Agent technology which can offer an automatic management of the cloud system by doing a quick adaptation once a new situation appears. This work is in fact composed of two parts: the description of the agent system and the implementation.

Keywords- cloud computing, Agent technology, resource allocation, optimization management, automatic management.

I. INTRODUCTION

Cloud computing is a rising technology. It combines different existing technologies like: service oriented Architecture, distributed computing, virtualization... Cloud computing represents a business model in which IT capabilities are offered to client through a network (mostly Internet). In cloud computing, everything (any given resource) is provided as a service. Consequently, the service provider reacts according to the user (client or customer) request by providing him all resources that satisfy his requirement. We distinguish between three main services: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In SaaS, a full ready application is available to the customer. In PaaS, the user has the development tools to create his own application. Finally in IaaS, the infrastructure provider prepares the hardware part as a service. This computing infrastructure is offered to the client as Virtual Machine (VM). VM is a software implemented machine which runs on physical machine. VM works like a real machine and can run different operating systems and applications since it has the same behavior. One physical machine can execute in the same time several different VM. Pay as you go is the most important advantage of cloud computing. It has a direct effect in the cloud development. In the actual economic situation (world crisis), many companies attempt to reduce their financial expenses by using cloud services. So, why to buy a computing infrastructure when it can be rent. Different situations have influenced the development and the adoption of cloud computing.

The U.S. National Institute of Standards and Technology (NIST) proposes [1] the following definition: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models”. This definition seems to have captured the commonly agreed aspects of cloud computing, was adopted by the most researchers. It is important to agree on one definition to limit the scope of research and emphasize the potential business benefits.

Cloud computing is still in developing stage. Many challenges are not resolved yet especially optimization and scheduling problems. In fact, a cloud provider deals with thousands of physical machines and up to millions of VM. In addition, some devices such as hard disks are the most vulnerable parts in such infrastructure. The majority of hardware failure/replacement is caused by to hard disk. So, the provider has to manage a huge number of resources and to predict hardware failure to keep the cloud computing system running. The challenge is to satisfy the request of the client with a quality of service. It is important to create an efficient strategy to manage this entire infrastructure. In this work we present a management strategy (model) of cloud computing based on agent technology which attempts to ease the control of cloud computing. This work is composed of two sections. In the first one, we summarize some researches related to this work. The second section is dedicated to present our proposed management system for cloud computing.

II. RELATED WORK

A. Resource Management

Optimization in cloud computing represents one of the most active research areas. The motivation was the huge
number of the resources used in the system of cloud. Generally in cloud, there are two principal actors: the service provider and the service consumer. Depending on the point of view, there are different strategies to use cloud resource: the service provider wants to reduce the cost of running the system and the consumer wishes to reduce the cost of using the cloud infrastructure with the minimum execution time. The service provider must consider client's constraint and in the same time reduce his running cost. Any failure of this challenge can lead to serious losses.

Data center [2], operating under the cloud computing model, can host different types of running applications starting from those that work for few seconds (web applications like e-commerce) to those that need a longer periods of time (simulation or large data processing). The management of this various types of applications in one data center creates a challenge between on demand resource provisioning and allocation in response to time varying workloads.

Monitoring of cloud is one of the important tasks in the cloud management. It supplies the service provider with critical information like the workload, the performance and the Quality of Service (QoS). This information is important to keep a track of the hardware state to prevent a future failure that can lead to a stop of service. The most important activities, in cloud computing, which need monitoring, are [3]: resource planning and management, data center management, troubleshooting, performance and security management. So, a monitoring and data analysis system must be implemented to have a good management strategy.

B. Agent Technology in Cloud Computing

Agent technology is used to benefit of the intelligent behaviors, communications and negotiation capabilities. In cloud computing we have two actors: the client and the service provider. Different works have proposed and presented some ideas to create a negotiation system between these two actors to gain the best cost for the client. We have identified different specifications for the agent: like finding out the best offer or discovering a new cloud computing provider...

The work [4] presents an agent-based testbed system to bolster cloud resource discovery and service level agreement negotiation. There are different jobs for the agents in this system. Actually, every action is achieved by a specific agent. There is an agent responsible for finding out all cloud computing systems that can satisfy the client request. Another agent is dedicated to negotiate a given phase or step. Depending on the client requirements, this agent will choose the best offer presented by cloud providers. It can even split the client work to different parts on different cloud systems to obtain a better result.

The work [5] proposes a complete system based on agent technology for provisioning and management. An application was developed to help the client in the different phases starting by the cloud provisioning (selecting the cloud resource), then the cloud configuration and finally cloud monitoring.

To the best of our knowledge there is no previous work which has used agent technology to optimize the resources management for the service provider.

III. PROPOSED WORK

A. Problem description

The management of cloud computing in IaaS level is a difficult task for the service provider. In fact it can’t be achieved manually. In a simple cloud computing system which is composed of one data center, we can easily find out thousands of physical resources that are virtualized in millions of virtual resources. So, the manager needs an automatic solution to supervise the system in order to minimize the human intervention. Moreover, this system must allow some leeway to the manager so he can set his own rules (customization) to supervise the system or adapt them to a new situation. Multi-agent system is a method that can be adopted as a solution to this problem. In fact, this solution offers an automatic management of the cloud system with a quick adaptation to the new situation. Multi-agent system can make a distributed and concurrent resolution of the problem. Different agents collaborate together to find a quicker and better solution. So, we attempt to deploy a multi-agent system to manage the cloud computing infrastructure. These agents will be charged to monitor the different resources. This control is deployed locally in the same data center or globally in the cloud computing system. To control the agent, a management interface is important to simplify the task of the cloud computing manager. Our work is composed of three essential phases. In the first part we define and describe the agent system (interaction, communication, type of agent...), then we create and define the agent and finally we create the management interface and the list of rules to be taken into account by the agent who will be charged in the resources management. We look to obtain a fully automatic system in which the client can connect to a website, inputs his requirements and loads his work. At this moment, the system agents will satisfy this request in an optimal manner according to the provider rules already customized.

B. Jade: Agent development framework

Jade (Java Agent DEvelopment Framework) is a software framework to develop agent applications in compliance with the FIPA (Foundation for Intelligence Physical Agents) specification for interoperable intelligent Multi-agent system [6] [7].

Agent communication in jade is based on messages exchange. Every agent in jade framework has task (code) to execute which is defined as behavior. i.e.: the behavior represents an action that an agent is able to perform.

C. CloudSim: Cloud computing Simulation Framework

CloudSim [8] [9] provides a support for modeling and simulation of virtualized cloud computing based data center environments. It provides a generalized and extensible simulation framework that enable modeling, simulation and experimentation of cloud computing infrastructure. CloudSim is used to develop and test scheduling policies, management strategy...

D. Agent system description

The Multi-agent system is composed of two types of agents: Supervisor Agent and Data Center Agent. We have one
The task of the supervisor agent is to receive client’s work and to deal with them to propose an offer based on rules defined by the manager. This agent is the main agent in the system. He coordinates between data center agent and makes all decisions. He is the most active and intelligent part in the system.

The data center agent controls resources in the data center. He plans the maintenance operation based on the state of resources (for example after 30 000 hours of work (~ 3 years) a hard disk is replaced to prevent his failure). He detects hardware failure and notified the human operator. When a request of resources comes from the supervisor agent, the data center agent analyses this request and looks in his schedule to determine the time when he can satisfy this request and send the response to the supervisor. He is also responsible of starting client’s work, finishing it and sending the result to the client.

The system is initiated whenever a resources request is arrived. The client inputs the list of resources he needs and the way and timeslots of the corresponding utilizations throw some wizards and interfaces. The supervisor agent receives the client’s demands and treats them one by one according to the rule “the first comes is the first served”. So, the client who comes first has a better chance to get served. To deal with the client request, we have chosen the following set of actions: when a command arrives, the supervisor agent sends a message to all data center agents to see who is active and who is not. Two cases can be found.

The first case: the data center agent does not return any response after a predefined timeslot. The problem can be caused by network latency or no agent in this data center is active. Therefore, the supervisor initiates a new data center agent. This data center agent has also a predefined timeslot during it, he has to indicate whether he is active and has the control of the data center or not. If this time is elapsed with no generated response, the supervisor considers that this data center has a network problem and it will try to repeat this scenario with the next received client’s request. After three attempts with no response, the human operator is notified (with email and/or SMS).

The second case: the supervisor agent receives a response from data center agent which means that this agent is active. In the next step, the supervisor estimates the time needed to complete the client’s work (task). Then he communicates this time and the resource requested to the client. Based on this information, the data center agent looks for available resources, its reservation planning and if there is a maintenance operation that is drafted between the starting of the client’s work and the estimated finish time. Then this agent returns the time when he can accept the client’s work and lock the requested resource until its confirmation or cancellation.

The supervisor agent chooses the minimum time from data center agent response and proposes it to the client. If the client accepts the proposal, the supervisor agent confirms this reservation on the selected data center and sends a message to the other agents to ignore their reservations.

If there is more than one data center that has the minimum time and in the case that the client accepts the proposed time, the supervisor agent computes the nearest data to the client based on network communication (a control message is sent from each data center to client. The message that arrives first, his corresponding data center is selected). This is the basic of our agent system. We are until now in the development stage. We are using Jade framework to create and to define the behavior of our agent system and cloudSim to simulate a cloud system and test it.

IV. Conclusion

In this work, an under development system for cloud management based on agent technology was presented. This system focuses on the service provider management resources. It attempts to control automatically the cloud computing resources to avoid (or to minimize) human intervention. The design phase and the definition of the system features were presented in this paper. As a future work, we will complete the implementation of the system and test it using jade framework and cloudSim. Then we will improve it and try to integrate it in an open source cloud computing system such as OpenNebula or Eucalyptus. This management system aims to offer the tools needed to ease the development of Holly Quran applications such as providing user several “tafsir” at the same time for a given verse. Cloud computing provide the infrastructure to serve a higher number of customer. So, it’s important to explore the potential capacity of this system to create a better Holly Quran services.
REFERENCES


