JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CLOUD COMPUTING ENVIRONMENTS

Dr. Vishal Deshmukh
Associate Professor
Bharati Vidyapeeth Deemed to be University, Pune
Yashwantrao Mohite Institute of Management, Karad
Mr. Pravin Natha Jadhav
MBA Student
Bharati Vidyapeeth Deemed to be University, Pune
Yashwantrao Mohite Institute of Management, Karad

Abstract

Cloud computing providers have setup several data centres at different geographical locations over the Internet to optimally serve needs of their customers around the world. However, existing systems do not support mechanisms and policies for dynamically coordinating load distribution among different Cloud-based data centres to determine optimal location for hosting application services to achieve reasonable QoS levels. Further, the Cloud computing providers are unable to predict geographic distribution of users consuming their services, hence the load coordination must happen automatically, and distribution of services must change in response to changes in the load. To counter this problem, we advocate creation of federated Cloud computing environment (Inter Cloud) that facilitates just-in-time, opportunistic, and scalable provisioning of application services, consistently achieving QoS targets under variable workload, resource and network conditions. The overall goal is to create a computing environment that supports dynamic expansion or contraction of capabilities (VMs, services, storage, and database) for handling sudden variations in service demands.

Key words: Cloud environment, geographical, hosting application etc.

Introduction

Cloud computing is the on-demand availability of computer system resources, especially data storage (cloud storage) and computing power, without direct active management by the user. Large clouds often have functions distributed over multiple locations, each location being a data centre. Cloud computing relies on sharing of resources to achieve coherence and economies of scale. Cloud computing delivers infrastructure, platform, and software (application) as services, which are made available as subscription-based services in a pay-as- you-go model to consumers. These services in industry are respectively referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). A Berkeley Report in

Feb 2009 states “Cloud computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service.” Clouds aim to power the next generation data centers by architecting them as a network of virtual services (hardware, database, user-interface, application logic) so that users can access and deploy applications from anywhere in the world on demand at competitive costs depending on users QoS (Quality of Ser- vice) requirements. Developers with innovative ideas for new Internet services no longer require large capital outlays in hardware to deploy their service or hu- man expense to operate it . It offers significant benefit to IT companies by free- in them from the low-level task of setting up basic hardware (servers) and soft- ware infrastructures and thus enabling more focus on innovation and creating business value for their services. The business potential of Cloud computing is recognised by several market re- search firms including IDC, which reports that worldwide spending on Cloud ser- vices will grow from $16 billion by 2008 to $42 billion in 2012. Furthermore, many applications making use of these utility-oriented computing systems such as clouds emerge simply as catalysts or market makers that bring buyers and sellers together.

Application Scaling and Cloud Infrastructure:

Providers such as Amazon, Google, Sales force, IBM, Microsoft, and Sun Microsystems have begun to establish new data centers for hosting Cloud computing application services such as social networking and gaming portals, business applications (e.g., Sales Force.com), media content delivery, and scientific workflows. Actual usage patterns of many real-world application services vary with time, most of the time in unpredictable ways. To illustrate this, let us consider an “elastic” application in the business/social networking domain that needs to scale up and down over the course of its deployment.

Social Networking Web Applications

Social networks such as Face book and MySpace are popular Web 2.0 based applications. They serve dynamic content to millions of users, whose access and interaction patterns are hard to predict. In addition, their features are very dynamic in the sense that new plug-ins can be created by independent developers, added to the main system and used by other users. In several situations load spikes can take place, for instance, whenever new system features become popular or a new plug- in application is deployed. As these social networks are organized in communities of highly interacting users distributed all over the world, load spikes can take place at different locations at any time. In order to handle unpredictable seasonal and geographical changes in system workload, an automatic scaling scheme is pa- ramount to keep QoS and resource consumption at suitable levels. Social networking websites are built using multi-tiered web technologies, which consist of application servers such as IBM Web Sphere and persistency layers such as the MySQL relational database. Usually, each component runs in a separate virtual machine, which can be hosted in data centers that are owned by different cloud computing providers. Additionally, each plug-in developer has the freedom to choose which Cloud computing provider offers the services that are more suitable to run his/her plug-in. As a consequence, a typical social networking web application is formed by hundreds of different services, which may be hosted by dozens of Cloud data centers around the world. Whenever there is a variation in temporal and spatial locality of workload, each application component must dynamically scale to offer good quality of experience to users.

Elastic Applications -

In order to support a large number of application service consumers from around the world, Cloud infrastructure providers (i.e., IaaS providers) have established data centers in multiple geographical locations to provide redundancy and ensure reliability in case of site failures. For example, Amazon has data centers in the US (e.g., one in the East Coast and another in the West Coast) and Europe. However, currently they (1) expect their Cloud customers (i.e., SaaS providers) to express a preference about the location where they want their application services to be hosted and (2) don't provide seamless/automatic mechanisms for scaling their hosted services across multiple, geographically distributed data centers. This approach has many shortcomings, which include it is difficult for Cloud customers to determine in advance the best location for hosting their services as they may not know origin of consumers of their services and (3) Cloud SaaS providers may not be able to meet QoS expectations of their service consumers originating from multiple geographical locations. This necessitates building mechanisms for seam- less federation of data centers of a Cloud provider or providers supporting dynamic scaling of applications across multiple domains in order to meet QoS targets of Cloud customers. In addition, no single Cloud infrastructure provider will be able to establish their data centers at all possible locations throughout the world. As a result Cloud application service (SaaS) providers will have difficulty in meeting QoS expectations for all their consumers. Hence, they would like to make use of services of multiple Cloud infrastructure service providers who can provide better support for their specific consumer needs. This kind of requirements often arises in enterprises with global operations and applications such as Internet service, media hosting, and Web 2.0 applications. This necessitates building mechanisms for federation of Cloud infrastructure service providers for seamless provisioning of services across different Cloud providers. There are many challenges involved in creating such Cloud interconnections through federation. To meet these requirements, next generation Cloud service providers should be able to: (i) dynamically expand or resize their provisioning capability based on sudden spikes in workload demands by leasing available computational and storage capabilities from other Cloud service providers; (ii) operate as parts of a market driven resource leasing federation, where application service providers such as Salesforce.com host their services based on negotiated Service Level Agreement (SLA) contracts driven by competitive market prices; and (iii) deliver on demand, reliable, cost-effective, and QoS aware services based on virtualization technologies while ensuring high QoS standards and minimizing service costs. They need to be able to utilize market-based utility models as the basis for provisioning of virtualized software services and federated hardware infrastructure among users with heterogeneous applications and QoS targets.

Research Issues

The diversity and flexibility of the functionalities (dynamically shrinking and growing computing systems) envisioned by federated Cloud computing model, combined with the magnitudes and uncertainties of its components (workload, compute servers, services, workload), pose difficult problems in effective provisioning and delivery of application services. Provisioning means “high-level management of computing, network, and storage resources that allow them to effectively provide and deliver services to customers". Finding efficient solutions for following challenges is critical to exploiting the potential of federated Cloud infrastructures.

Application Service Behaviour Prediction

It is critical that the system is able to predict the demands and behaviours of the hosted services, so that it intelligently undertake decisions related to dynamic scaling or de-scaling of services over federated Cloud infrastructures. Concrete prediction or forecasting models must be built before the behaviour of a service, in terms of computing, storage, and network bandwidth requirements, can be predicted accurately. The real challenge in devising such models is accurately learning and fitting statistical functions to the observed distributions of service behaviours such as request arrival pattern, service time distributions, I/O system behaviours, and network usage. This challenge is further aggravated by the existence of statistical correlation (such as stationary, short- and long-range dependence, and pseudo- periodicity) between different behaviours of a service.

(i) Flexible Mapping of Services to Resources: With increased operating costs and energy requirements of composite systems, it becomes critical to maximize their efficiency, cost-effectiveness, and utilization The process of mapping services to resources is a complex undertaking, as it requires the system to compute the best software and hardware configuration (system size and mix of re- sources) to ensure that QoS targets of services are achieved, while maximizing system efficiency and utilization. This process is further complicated by the uncertain behaviour of resources and services. Consequently, there is an immediate need to devise performance modelling and market-based service mapping tech- niques that ensure efficient system utilization without having an unacceptable impact on QoS targets.

Further, there is huge amount of sensitive data in an enterprise, which is unlikely to migrate to the cloud due to privacy and security issues. As a result, there is a need to investigate issues related to integration and interoperability between the software on premises and the services in the cloud. Identity management: authentication and authorization of service users; provisioning user access; federated security model.

(ii) Data Management: Not all data will be stored in a relational database in the cloud, eventual consistency (BASE) is taking over from the traditional ACID transaction guarantees, to ensure sharable data structures that achieve high scalability.

(iii) Business process orchestration: How does integration at a business process level happen across the software on premises and service in the Cloud boundary? Where do we store business rules that govern the business process orchestration?

(iv) Scalable Monitoring of System Components: Although the components that contribute to a federated system may be distributed, existing techniques usually employ centralized approaches to overall system monitoring and man- agreement. We claim that centralized approaches are not an appropriate solution for this purpose, due to concerns of scalability, performance, and reliability arising from the management of multiple service queues and the expected large volume of service requests. Monitoring of system components is required for effecting on-line control through a collection of system performance characteristics. Therefore, we advocate architecting service monitoring and management services based on decentralized messaging and indexing models

Objectives

To meet requirements of auto-scaling Cloud applications, future efforts should focus on design, development, and implementation of software systems and policies for federation of Clouds across network and administrative boundaries. The key elements for enabling federation of Clouds and auto-scaling application services are Cloud Coordinators, Brokers, and an Exchange. The re- source provisioning within these federated clouds will be driven by market- oriented principles for efficient resource allocation depending on user QoS targets and workload demand patterns. To reduce power consumption cost and improve service localization while complying with the Service Level Agreement (SLA) contracts, new on-line algorithms for energy-aware placement and live migration of virtual machines between Clouds would need to be developed. The approach for realisation of this research vision consists of investigation, design, and development of the following:

- Architectural framework and principles for the development of utility- oriented clouds and their federation.
- A Cloud Coordinator for exporting Cloud services and their management driven by market-based trading and negotiation protocols for optimal QoS delivery at minimal cost and energy.
- A Cloud Broker responsible for mediating between service consumers and Cloud coordinators.
- A Cloud Exchange acts as a market maker enabling capability sharing across multiple Cloud domains through its match making services.
- A software platform implementing Cloud Coordinator, Broker, and Ex- change for federation.

The rest of this paper is organized as follows: First, a concise survey on the existing state-of-the-art in Cloud provisioning is presented. Next, the comprehensive description related to overall system architecture and its elements that forms the basis for constructing federated Cloud infrastructures is given. This is followed by some initial experiments and results, which quantifies the performance gains delivered by the proposed approach. Finally, the paper ends with brief conclusive remarks and discussion on perspective future research directions.

State-of-the-art in Cloud Provisioning

The key Cloud platforms in Cloud computing domain including Amazon Web Services Microsoft Azure Google AppEngine Eucalyptus, and GoGrid offer a variety of pre-packaged services for monitoring, managing, and provisioning resources and application services. However, the techniques implemented in each of these Cloud platforms vary (refer to Table 1. The three Amazon Web Services (AWS), Elastic Load Balancer, Auto Scaling and CloudWatch together expose functionalities which are required for undertaking provisioning of application services on Amazon EC

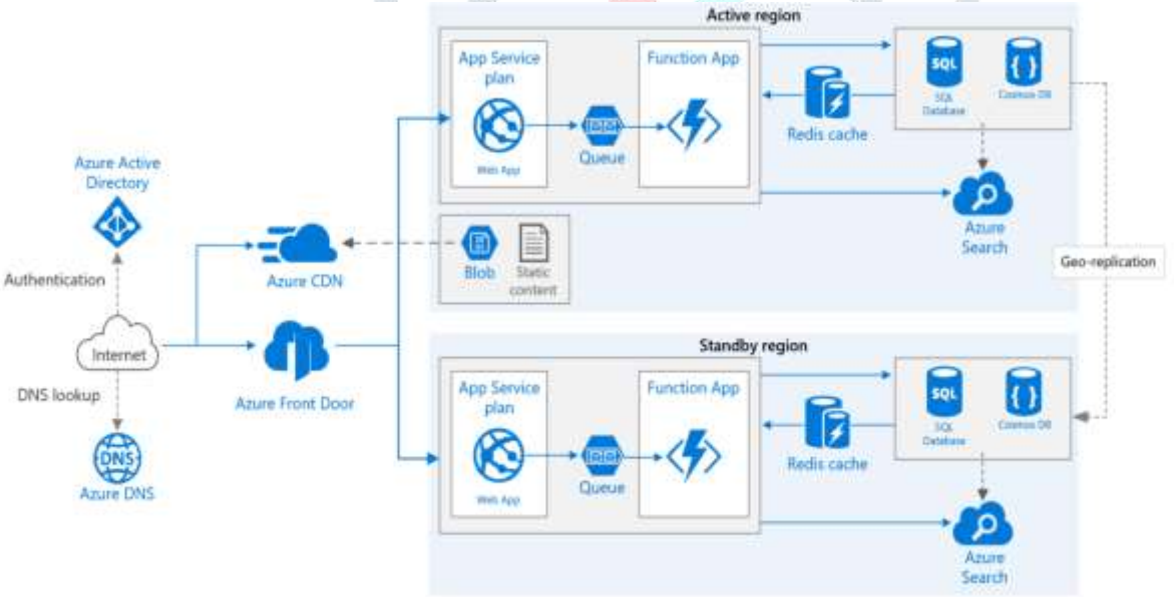
2. Elastic Load Balancer service automatically provisions incoming application workload across available Amazon EC2 instances. Auto-Scaling service can be used for dynamically scaling-in or scaling-out the number of Amazon EC2 instances for handling changes in service demand patterns. And finally, the CloudWatch service can be integrated with above services for strategic decision making based on real-time aggregated resource and service performance information.

Cloud Platforms	Load Balancing	Provisioning	Auto Scaling
Amazon Elastic Compute Cloud	√	√	√
Microsoft Windows Azure	√	√	√
Google App Engine	√	√	√
Eucalyptus	√	√	×

Table1: Summary of provision in capabilities exposed by public Cloud platforms

System Architecture -

This figure shows the high-level components of the service-oriented architectural framework consisting of client's brokering and coordinator services that support utility-driven federation of clouds: application scheduling, resource allocation and migration of workloads. The architecture cohesively couples the administratively and topologically distributed storage and computes capabilities of Clouds as parts of single resource leasing abstraction. The system will ease the cross-domain capabilities integration for on demand, flexible, energy-efficient, and reliable access to the infrastructure based on emerging virtualization technologies



Conclusions and Future Directions

Development of fundamental techniques and software systems that integrate distributed clouds in a federated fashion is critical to enabling composition and deployment of elastic application services. We believe that outcomes of this research vision will make significant scientific advancement in understanding the theoretical and practical problems of engineering services for federated environments. The resulting framework facilitates the federated management of system components and protects customers with guaranteed quality of services in large, federated and highly dynamic environments. The different components of the proposed framework offer powerful capabilities to address both services and resources management, but their end-to-end combination aims to dramatically improve the effective usage, management, and administration of Cloud systems. This will provide enhanced degrees of scalability, flexibility, and simplicity for management and delivery of services in federation of clouds.

In our future work, we will focus on developing comprehensive model driven approach to provisioning and delivering services in federated environments. These models will be important because they allow adaptive system management by establishing useful relationships between high-level performance targets (specified by operators) and low-level control parameters and observables that system com- ponents can control or monitor. We will model the behaviour and performance of different types of services and resources to adaptively transform service requests. We will use a broad range of analytical models and statistical curve-fitting tech- niques such as multi-class queuing models and linear regression time series. These models will drive and possibly transform the input to a service provisioner, which improves the efficiency of the system. Such improvements will better ensure the achievement of performance targets, while reducing costs due to improved utilization of resources. It will be a major advancement in the field to develop a robust and scalable system monitoring infrastructure to collect real-time data and re- adjust these models dynamically with a minimum of data and training time. We believe that these models and techniques are critical for the design of stochastic provisioning algorithms across large, federated Cloud systems where resource availability is uncertain. Lowering the energy usage of data centers is a challenging and complex issue because computing applications and data are growing so quickly that increasingly larger servers and disks are needed to process them fast enough within the required time. Green Cloud computing is envisioned to achieve not only efficient processing and utilization of computing infrastructure, but also minimization of energy consumption. This is essential for ensuring that the future growth of Cloud Computing is sustainable. Otherwise, Cloud computing with increasingly pervasive front-end client devices interacting with back-end data centers will cause an enormous escalation of energy usage. To address this problem, data centre resources need to be managed in an energy-efficient manner to drive Green Cloud computing. In particular, Cloud resources need to be allocated not only to satisfy QoS targets specified by users via Service Level Agreements (SLAs), but also to reduce energy usage. This can be achieved by applying market-based utility models to accept requests that can be fulfilled out of the many competing requests so that revenue can be optimized along with energy- efficient utilization of Cloud infrastructure.

References-

- http://www.amazon.com
- http://appengine.google.com
- http://www.microsoft.com
- http://www.springframework.net
- www.microsoft.com