Journal of Network and Computer Applications 41 (2014) 424–440

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey

Sunilkumar S. Manvi [a], Gopal Krishna Shyam [b,*]

[a] Department of Electronics and Communication Engineering, Reva Institute of Technology & Management, Bangalore 560 064, India
[b] Department of Computer Science and Engineering, Reva Institute of Technology and Management, Bangalore 560 064, India

ARTICLE INFO

Article history:
Received 26 January 2013
Received in revised form
8 August 2013
Accepted 14 October 2013
Available online 25 October 2013

Keywords:
Cloud computing
Resource management

ABSTRACT

The cloud phenomenon is quickly becoming an important service in Internet computing. Infrastructure as a Service (IaaS) in cloud computing is one of the most significant and fastest growing field. In this service model, cloud providers offer resources to users/machines that include computers as virtual machines, raw (block) storage, firewalls, load balancers, and network devices. One of the most pressing issues in cloud computing for IaaS is the resource management. Resource management problems include allocation, provisioning, requirement mapping, adaptation, discovery, brokering, estimation, and modeling. Resource management for IaaS in cloud computing offers following benefits: scalability, quality of service, optimal utility, reduced overheads, improved throughput, reduced latency, specialized environment, cost effectiveness and simplified interface. This paper focuses on some of the important resource management techniques such as resource provisioning, resource allocation, resource mapping and resource adaptation. It brings out an exhaustive survey of such techniques for IaaS in cloud computing, and also put forth the open challenges for further research.

Contents

* Corresponding author. Tel.: +91 8065687564.
E-mail addresses: sunil.manvi@revainstitution.org (S.S. Manvi), gopalkrishna@revainstitution.org, gopalshyambabu@gmail.com (G. Krishna Shyam).

1084-8045/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2013.10.004

1. Introduction

A cloud is defined as a place over network infrastructure where information technology (IT) and computing resources such as computer hardware, operating systems, networks, storage, databases, and even entire software applications are available instantly, on-demand as given in Buyya and Ranjan (2011). Cloud computing is the use of cloud resources (hardware and software) that are delivered as a service over a network (typically the Internet). While cloud computing may not involve a lot of new technologies, it certainly represents a new way of managing IT. In many cases, this will not only change the workflow within the IT organization, it will often result in a complete reorganization of the IT department. Cost savings and scalability can be highly achieved from cloud computing.

Cloud computing is often compared with Service Oriented Architectures (SOA), Grid, Utility and Cluster computing as in http://cloudcomputing.sys-con.com. Cloud computing and SOA can be pursued independently or concurrently where cloud computing's platform and storage service offerings can provide a value added underpinning for SOA's efforts as in Dai and Rubin (2012). Cloud computing does not replace SOA or the use of distributed software's components as an integration technology. With grid computing, we can provision computing resources as a utility that can be turned on or off. Cloud computing goes one step further with on-demand resource provisioning. This eliminates over-provisioning when used with utility pricing. It also removes the need to over-provision in order to meet the demands of millions of users. Utility computing is paying for what we use on shared servers like we pay for a public utility (such as electricity, gas) as in http://www.ibm.com.

Clustering is the use of multiple computers, typically PCs or UNIX workstations, multiple storage devices, and redundant interconnections, to form what appears to users as a single highly available system. Cluster computing is a low-cost form of parallel processing for scientific and other applications that lend themselves to parallel operations. The summary of the features of each of the computing techniques is listed in Table 1.

Clouds can be broadly classified as follows:

- Infrastructures as a Service (IaaS).
- Platforms as a Service (PaaS).
- Software as a Service (SaaS).

IaaS refers to a combination of hosting, hardware provisioning and basic services needed to run a cloud. PaaS refers to the provision of a computing platform and the provision and deployment of the associated set of software applications (called a solution stack) to an enterprise by a cloud provider. Software as a Service (SaaS) is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network.

The uses of IaaS are as follows. (1) Provides access to shared resources on need basis, without revealing details like location and hardware to clients, (2) provides details like server images on demand, storage, queuing, and information about other resources, among others, and (3) offers full control of server infrastructure, not limited specifically to applications, instances and containers.

Table 1
Computing techniques.

Computing techniques	*Features*
Cloud computing	Cost efficient, almost unlimited storage, backup and recovery, easy deployment
SOA	Loose coupling, distributed processing, asset creation
Grid computing	Efficient use of idle resources, modular, parallelism can be achieved, handles complexity
Cluster computing	Reduced cost, processing power, improved network technology, scalability, availability

The major issues that are commonly associated with IaaS in cloud systems are resource management, network infrastructure management, virtualization and multi-tenancy, data management, application programming interfaces (APIs), interoperability, etc. These issues are briefly discussed in Section 2. This paper focuses on resource management due to the fact that resource management for IaaS in cloud computing offers following benefits: scalability, Quality of Service (QoS), specialized environment, reduction in overheads and latency, improved throughput, cost effectiveness and simplified interface. This paper focuses on survey of some of the important resource management schemes such as resource mapping, resource provisioning, resource allocation, and resource adaptation. It brings out an exhaustive survey of such schemes for IaaS in cloud computing, and also put forth the open challenges for further research.

Our contributions in the paper are as follows. (1) Classification of resource management schemes into resource provisioning, resource allocation, resource adaptation, resource mapping, resource modeling, resource estimation and resource brokering. (2) Bringing out exhaustive works in resource mapping, resource provisioning, resource allocation, and resource adaptation. Scheme's operation and drawbacks are presented with comparative analysis in terms of performance metrics such as reliability, deployment, QoS, delay and overheads. (3) Providing open challenges in each considered class of resource management, and (4) to facilitate novice researchers to work on the research problems.

The rest of the paper is organized as follows. In Section 2, we discuss different issues in IaaS. Resource management problems in IaaS are presented in Section 3. Tools and technologies for resource management in cloud computing are discussed in Section 4. Some solutions for resource management are given in Section 5. Finally, concluding remarks are presented in Section 6.

2. Issues in IaaS

The major issues that are commonly associated with IaaS in cloud systems are virtualization and multi-tenancy, resource management, network infrastructure management, data management, APIs, interoperability etc. We will briefly introduce the issues.

(i) *Virtualization and multi-tenancy*: Virtulization is an essential technological characteristic of clouds, hides the technological complexity from the user and enables enhanced flexibility (through aggregation, routing and translation). In a multi-tenancy environment, multiple customers share the same application, running on the same operating system, on the same hardware, with the same data-storage mechanism. The distinction between the customers is achieved during application design, thus customers do not share or see each other's data. In case of virtualization, components are abstracted enabling each customer application to appear to run on a separate physical machine. Lombardi and Pietro (2011) show how virtualization can increase the security of cloud computing by protecting both the integrity of guest virtual machines and the cloud infrastructure components. The effect of virtualization on new generation programming models and environments like Hadoop has been explored in Ibrahim and Shi (2009).

One of the most important issues in network virtualization is an efficient utilization of subtrate network (SN) resources. It will help to improve the resource utilization as well as avoiding congestion in the SN. Haider and Potter (2009) focus on issues

related to the problem of resource allocation in VNs. It provides a concise overview of various existing techniques for resource allocation in VNs. These can be helpful for developing detailed designs, specifications and performance evaluation techniques for VNs. However, this paper does not attempt to provide an exhaustive survey on the resource allocation/management techniques in VNs. Multi-tenancy is a highly essential issue in cloud systems, where the location of code and/or data is principally unknown and the same resource may be assigned to multiple users. This affects infrastructure resources as well as data/applications/services that are hosted on shared resources but need to be made available in multiple isolated instances. Multi-tenancy implies a lot of potential issues, ranging from data protection to legislator issues.

While hardware based virtualization has many benefits, it lacks from a high level of scalability required to offer cost effective cloud computing for masses. Multi-tenant virtualization remedies this bottleneck by focusing on software based virtualization. Siddhisena et al. (2011) present an approach to application virtualization using multi-tenant concept.

(ii) *Resource management*: At any time instant, resources are to be allocated to effectively handle workload fluctuations, while providing QoS guarantees to the end users. The computing and network resources are limited and have to be efficiently shared among the users in virtual manner. In order to perform effective resource management, we need to consider the issues such as resource mapping, resource provisioning, resource allocation and resource adaptation. The lack of mature virtualization tools and powerful processor's have prevented growth of cloud computing. Although relatively new, a fair amount of work by Urgaonkar et al. (2010a) and Vaquero et al. (2009) has been done to examine current and future challenges for both users and providers of cloud computing. However, little has been done to understand the range of operational challenge faced by users as they attempt to run applications in the cloud. Chase et al. (2010) have considered the problem of energy-efficient resource management of homogeneous resources in Internet hosting centres. The main challenge is to determine the resource demand of each application at its current request load level and to allocate resources in most efficient way.

Metering of any kind of resource and service consumption is essential in order to offer elastic pricing, charging and billing. It is therefore a pre-condition for the elasticity of clouds. The issue here is to see that the users are charged only for the services that they use for the specific period of time. Cloud computing alone will not help an organization to determine who will pay for what resource, but it can help provide a platform for an infrastructure design that establishes a charge-back model for metering and billing.

(iii) *Network infrastructure management*: Managing millions of network components (hubs, bridges, switches etc.) leads to unsustainable administrator costs, requiring automated methods for typical system management tasks. These automated methods needs to deal with increased monitoring data size of several orders of magnitudes higher than current systems. Gupta and Singh (2009) suggest putting network interfaces, links, switches and routers into sleep modes when they are idle in order to save the energy consumed by the Internet backbone and consumers. Chiaraviglio and Matta (2010a) have proposed cooperation between ISPs (Internet Service Providers) and content providers that allow the achievement of an efficient simultaneous allocation of compute resources and network paths that minimize energy consumption.

(iv) Security, privacy and compliance are obviously essential in all systems dealing with potentially sensitive data and code. To ensure adequate security in cloud computing, various security issues, such as authentication, data confidentiality, integrity, and non-repudiation need to be considered.

(v) Data management is an essential aspect in particular, for storage clouds, where data is flexibly distributed across multiple resources. Implicitly, data consistency needs to be maintained over a wide distribution of replicated data sources. At the same time, the system always needs to be aware of the data location (when replicating across data centres) taking latencies and workload into consideration.

(vi) APIs and/or programming enhancements are essential to exploit the cloud features. Common programming models require that the developer takes care of the scalability and autonomic capabilities, whilst a cloud environment provides the features in a fashion that allows the user to leave such management to the system.

(vii) Tools are generally necessary to support development, adaptation and usage of cloud services. Getting a final product in cloud can be a significant challenge. Cloud computing reduces some of these problems by introducing tools and processes that provide a complete server and storage environment without the need to interact with technical specialists.

Today, resources are pooled on storage and virtualization platforms, a degree of elasticity is present, and services are more “on demand” than ever before. Behind the scenes, there is some level of automation present, with many of the tools and processes in place to build a self-service portal. There are tools that monitor, tools that provision, and tools that cross the divide between both. There are a few vendors that offer pervasive approaches in handling provisioning and managing metrics in hybrid environments as in http://searchcloudcomputing.techtarget.com: RightScale, Kaavo etc.

RightScale's management environment provides interface to users for managing the resources. It is designed to walk a user through the initial process of migrating to the cloud using their templates and library. The core of Kaavo's product is called IMOD. IMOD handles configuration, provisioning and changes to the cloud environment across multiple vendors in a hybrid model. Like all major CIM (common information model) players, Kaavo's IMOD sits at the “top” of the stack, managing the infrastructure and application layers. One great feature in IMOD is its multi-cloud, single system tool. For instance, one can create a database backend in Rackspace while putting presentation servers on Amazon.

Both Kaavo and RightScale offer scheduled dynamic allocation based on demand and monitoring tools to ensure that information and internal metrics (like SLAs) are transparently available. The dynamic allocation even helps meet the demands of those SLAs. Both offer the ability to keep templates as well to ease the deployment of multi-tier systems. There are several tools that are available to do a specific task. The issue is to see that we develop tools which are effective and gives accurate results.

The non-technological issues also play a major role in realizing these technological aspects and in ensuring viability of the infrastructures in the first instance. Non-technological issues are as follows. (1) Economic aspects which cover knowledge about when, why, how to use which cloud system, how this impacts on the original infrastructure (providers long-term experience is lacking in all these areas). (2) Legalistic issues which come as a consequence from the dynamic handling of the clouds, their scalability and the partially unclear legislative issues in the Internet.

3. Resource management problems in IaaS

There are several problems to be considered while managing resources, such as, type of resource required (physical/logical), allocation, brokering, provisioning, mapping, adaptation and

Table 2
Physical and logical resources.

Physical resources	*Logical resources*
CPU	Operating system
Memory	Energy
Storage	Network throughput/ bandwidth
Workstations	Information security, Protocols
Network elements	APIs
Sensors/actuators	Network loads, delays

estimation. In this section, we present significance of resource types and issues in resource management.

3.1. Type of resources

A resource is any physical or virtual component of limited availability within a computer system. Every device connected to a computer system is a resource. Every internal system component is a resource. We have listed out various physical and logical resources in Table 2. Now we shall discuss the impact of the above mentioned resources on the performance of the clouds.

3.1.1. Physical resources

Physical resources typically include processors, memory, and peripheral devices. Physical resources vary fairly dramatically from computer to computer. For example, a typical PC system might have 640 K of memory, one 20 Mbyte Winchester disk, one floppy disk drive, a single keyboard, and a single video display. A typical mainframe system has several parallel processors, hundreds of disks, tens of millions of bytes of memory, hundreds of terminals, tapes, and other special purpose peripherals, and is connected to a global network with thousands of other similar computers. Cloud providers can offer resource provisioning plans to consumers, namely short-term on-demand and long-term reservation plans. Outsourcing techniques can be used to take advantage of cloud computing infrastructure for providing scalability and high availability capabilities to the web applications deployed on it; this would definitely increase the number of cloud consumers and hence increase the resource utilization of clouds. Now let us discuss some of the important physical resources and its impact on cloud computing.

(i) *CPU (central processing unit)*: It performs most of the processing inside a computer. The issue in cloud computing is the CPU utilization. CPU utilization refers to a computer's usage of processing resources, or the amount of work handled by a CPU. Actual CPU utilization in cloud varies depending on the amount and type of managed computing tasks. Certain tasks require heavy CPU time, while others require less because of non-CPU resource requirements. Proper CPU usage makes it easy to consume massive amounts of compute power for batch processing, data analysis, and high performance computing needs.

(ii) *Memory*: The cloud computer architecture asks for a clustered structure of the memory resources in the form of virtual entities. Gone are the days when memory management was done using the static methods. As cloud environment is dynamic and volatile, there is a strong need to inculcate the dynamic memory allocation trends in the cloud based systems. The increased number of cores in cloud servers combined with the rapid adoption of virtualization technologies also creates huge demand for memory.

(iii) *Storage*: It refers to saving data to an off-site storage system maintained by a third party. Instead of storing information to computer's hard drive or other local storage device, we save it to a remote database. The Internet provides the connection between the computer and the database. Cloud storage systems generally rely on hundreds of data servers. Because computers occasionally require maintenance or repair, it is important to store the same information on multiple machines. This is called redundancy. Without redundancy, a cloud storage system could not ensure clients that they could access their information at any given time. Most systems store the same data on servers that use different power supplies. That way, clients can access their data even if one power supply fails. The two biggest concerns about cloud storage are reliability and security. Clients are not likely to entrust their data to another company without a guarantee that they will be able to access their information whenever they want and no one else will be able to get it. Through this, clouds provide storage as a service.

(iv) *Workstations*: IT managers are seeing a trend where more powerful PCs could be classified as workstations. Bigger CPUs, faster graphics, and upwards of 20 Gb of RAM–these are machines that are designed to do a lot of local processing. The challenge is to take advantage of cloud technology and economics and use workstations to bring high-performance computing (HPC) capabilities to the corporate user or any user connected to the Internet with good connectivity. To accomplish this, cloud vendors are working with applications that live and do some work on the local workstation, but then shunt much of the workload out into the cloud and back again. In order to see that the users get the best services at all times, proper functioning of workstations are very much desired. The systems used as workstations should be of higher configuration so that they withstand the overloaded network. The issue here is to see that workstations operate on its own, without human interference. Through this, clouds provide monitoring as a service.

(v) *Network elements*: Managing millions of network components (hubs, bridges, switches etc.) lead to unsustainable administrator costs, requiring automated methods for typical system management tasks. The automated methods need to deal with increased monitoring network sizes of several orders of magnitudes higher than current systems. Through this, clouds provide communication as a service.

(vi) *Sensors/actuators*: Proliferation of applications involving Internet-connected objects, has recently given rise to the notion of clouds of Internet-connected objects (i.e. sensors, actuators, devices), which are promoted as large-scale networks of spatially distributed entities with scalable processing and storage capabilities.

3.1.2. Logical resources

Logical resources are system abstractions which have temporary control over physical resources. They can support in development of applications and efficient communication protocols. The significance of logical resources in cloud computing is as follows.

(i) *Operating system*: It provides users with “logical” well-behaved environment to manage physical (hardware) resources as well as offers mechanisms and policies for the control of object/resources. The operating system performs file management, device management, performance management, security and fault tolerance management, thereby facilitating efficient utilization of the available resources.

(ii) *Energy*: The main technique applied to minimize energy consumption is concentrating the workload to the minimum of physical nodes and switching idle nodes off. This approach requires dealing with the power/performance trade-off, as performance of applications can be degraded due to the workload consolidation.

(iii) *Network throughput/bandwidth*: In cloud computing, we are concerned about measuring the maximum data throughput in bits

per second of a communications link or network access. A typical method of performing a measurement is to transfer a 'large' file from one system to another system and measure the time required to complete the transfer or copy of the file. Higher throughput is desired to make network work efficiently. Bandwidth management protocols are used to prevent congestion, essentially by accepting or refusing a new-arrival cell. The bandwidth allocation problem which is the most critical one is concerned with successful integration of link capacities through the different types of services.

(iv) *Load balancing mechanisms*: It is assumed that the physical facilities for providing cloud computing services are distributed over multiple centers in order to make it easy to increase the number of the facilities when demand increases, to allow load balancing, and to enhance reliability.

(v) *Information security*: For end-users to feel comfortable with a cloud solution that holds their software, data and processes; there should exist considerable assurance that services are highly reliable and available as well as secure and safe, and that privacy is protected. Hence we need to consider various security issues, such as authentication, data confidentiality, integrity, and non-repudiation.

(vi) *Delays*: Second or even millisecond can make a significant difference, when we talk about the quality of delay-sensitive traffic, the end user experience with cloud-based services, or the ability to trade fairly. A cloud service provider should be able to make accurate decisions for scaling up or down its data-centers while taking into account several utility criteria, e.g., the delay of virtual resources setup, the migration of existing processes, the resource utilization, etc.

(vii) *APIs/(Applications Programming Interfaces)*: It is a protocol intended to be used as an interface by software components to communicate with each other. An API may include specifications for routines, data structures, object classes, and variables. An API specification can take many forms, including international standard such as POSIX, vendor documentation such as the Microsoft Windows API, the libraries of a programming language, e.g. standard template library in C++ or Java API.

(viii) *Protocols*: Network protocols include mechanisms for devices to identify and make connections with each other, as well as formatting rules that specify how data is packaged into messages sent and received. Hundreds of different computer network protocols can be developed for specific purposes in cloud environments. Some of the examples of protocols are as follows.

Eludiora et al. (2011) propose user identity management protocol (U-IDM protocol) for cloud computing customers and cloud service providers. This protocol will authenticate and authorize customers/providers in order to achieve global security networks. The protocol will be developed to achieve the set of global security objectives in cloud computing environments.

A cloud application provider, a cloud storage provider and a network provider could implement different policies. The unpredictable interactions between load-balancing and other reactive mechanisms could lead to dynamic instabilities. The unintended coupling of independent controllers that manage the load, power consumption, and elements of the infrastructure could lead to undesirable feedback and instability similar to the ones experienced by the policy-based routing in the Internet Border Gateway Protocol (BGP) as in http://technet.microsoft.com

Wuhib et al. (2010) propose a protocol that can be used to meet design goals for resource management like fairness of resource allocation with respect to sites, efficient adaptation to load changes and scalability of the middleware layer in terms of both the number of machines in the cloud as well as the number of hosted sites.

(ix) *Network loads*: Cloud applications can present varying workloads. It is therefore essential to carry out a study of cloud services and their workloads in order to identify common behaviors, patterns, and explore load forecasting approaches that can potentially lead to more efficient resource provisioning and consequently improved energy efficiency.

3.2. Issues in resource management

The important issues identified in resource management are, resource provisioning, resource allocation, resource adaptation, resource mapping, resource modelling, resource estimation, resource discovery and selection, resource brokering, and resource scheduling. Their definitions are highlighted in Table 3. However, this paper limits discussion to only few important issues resource provisioning, resource allocation, resource adaptation, and resource mapping since they represent important entities in resource management software's, and we would like to have focused survey.

4. Tools and technologies for resource management in cloud computing

As the computing industry shifts toward providing Infrastructure as a Service (IaaS) for consumers and enterprises to access on demand resources regardless of time and location, there will be an increase in the number of available cloud platforms as in Ruth et al. (2010). Recently, several academic and industrial organizations have started investigating and developing technologies and infrastructure for cloud computing for IaaS. Keahey et al. (2011), http://www.opennebula.org, http://www.reservoir-fp7.eu provide

Table 3
Issues in resource management.

Issue	*Definition*
Resource provisioning	It is the allocation of a service provider's resources to a customer
Resource allocation	It is the distribution of resources economically among competing groups of people or programs
Resource adaptation	It is the ability or capacity of that system to adjust the resources dynamically to fulfill the requirements of the user
Resource mapping	It is a correspondence between resources required by the users and resources available with the provider
Resource modeling	Resource modeling is based on detailed information of transmission network elements, resources and entities participating in the network. It is a framework that illustrates the most important attributes of resource management: states, transitions, inputs and outputs within a given environment. Resource modeling helps to predict the resource requirements in subsequent time intervals
Resource estimation	It is a close guess of the actual resources required for an application, usually with some thought or calculation involved
Resource discovery and selection	It is the identification of list of authenticated resources that are available for job submission and to choose the best among them
Resource brokering	It is the negotiation of the resources through an agent to ensure that the necessary resources are available at the right time to complete the objectives
Resource scheduling	A resource schedule is a timetable of events and resources. Shared resources are available at certain times and events are planned during these times. In other words, It is determining when an activity should start or end, depending on its (1) duration, (2) predecessor activities, (3) predecessor relationships, and (4) resources allocated

academic efforts that include virtual workspaces, OpenNebula and Reservoir.

Vouk (2010) described cloud computing from a SOA perspective and talked about the Virtual Computing Laboratory (VCL) as an implementation of a cloud. VCL is an "open source implementation of a secure production-level on-demand utility and service oriented technology for wide-area access to solutions based on virtualized resources, including computational, storage and software resources". In this respect, VCL could be categorised as an IaaS layer service.

There are three groups currently working on standards for cloud computing: The Cloud Computing Interoperability Forum9, the Open Cloud Consortium10, and the DMTF Open Cloud Standards Incubator11. There is also a document called the Open Cloud manifesto12, in which various stakeholders express why open standards will benefit cloud computing. Kehey and Tsugua (2010) looked into the difficulties of developing standards and summarised the main goals of achieving interoperability between different IaaS providers as being machine-image compatibility, contextualization compatibility and API-level compatibility.

Eucalyptus presented in Nurmi and Wolski (2010) and Baun and Kunze (2009) is an open-source software package that can be used to build IaaS clouds from computer clusters. Eucalyptus emulates the proprietary Amazon EC2 SOAP and query interface, and thus an IaaS infrastructure set up using Eucalyptus can be controlled with the same tools and software that is used for EC2. The open source nature of Eucalyptus gives the community a useful research tool to experiment with IaaS provisioning as given in Nimbus and Nurmi et al. (2009).

Harmer and Wright (2009) present a cloud resource interface that hides the details of individual APIs to allow provider agnostic resource usage. They present the interface to create a new instance at Amazon EC2, at Flexiscale15, and at a provider of on-demand non-virtualized servers called NewServers16, and implemented an abstraction layer for these APIs.

Sotomayor et al. (2010) present two tools for managing cloud infrastructures: OpenNebula, a virtual infrastructure manager, and Haizea, a resource lease manager. To manage the virtual infrastructure, OpenNebula provides a unified view of virtual resources regardless of the underlying virtualization platform, manages the full lifecycle of the VMs, and supports configurable resource allocation policies including policies for times when the demand exceeds the available resources. Haizea can act as a scheduling backend for OpenNebula, and together they advance other virtual infrastructure managers by giving the functionality to scale out to external clouds, and providing support for scheduling groups of VMs. Another tool extends IBM data centre management software to be able to deal with cloud-scale data centre, by using a hierarchical set up of management servers instead of a central one. For resilience, the management CloudSim, modeling and simulation toolkit has been proposed in Sriram (2010). CloudSim goal is to provide a generalized and extensible simulation framework that enables modeling, simulation, and experimentation of emerging cloud computing infrastructures and application services, allowing its users to focus on specific system design issues that they want to investigate, without getting concerned about the low level details related to cloud-based infrastructures and services.

Buyya and Ranjan (2009) seek to optimise change management strategies, which are necessary for updates and maintenance for low energy consumption of a cloud data centre. One of the key aspects of cloud computing is elasticity, which will make it difficult to estimate the load from the service-level-agreements (SLAs) in place.

Recently, green clouds concept is gaining more importance. Green cloud refers to the potential environmental benefits that information technology (IT) services delivered over the Internet can offer to the society as in Lakshmi et al. (2012). Green cloud bills itself as a truly green cloud computing service, which according to the company means a public cloud infrastructure powered entirely by renewable energy (rather than using carbon offset credits), and providing customers with the means to live monitor their energy metrics and carbon savings.

Younge et al. (2011), present a novel Green Cloud framework for improving system efficiency in a data center. To demonstrate the potential of their framework, authors have presented new energy efficient scheduling, VM system image, and image management components that explore new ways to conserve power. Though the research presented in the paper, new ways to save vast amounts of energy has been seen while minimally impacting performance.

Server consolidation is an approach to the efficient usage of computer server resources in order to reduce the total number of servers or server locations that an organization requires. Server consolidation describes a variety of ways of reducing capital and operating expenses associated with running servers. Gartner research divides consolidation projects into three categories as in Leostream: logical consolidation means implementing common processes and management across a range of server applications; physical consolidation means collocating servers in fewer locations; rationalized consolidation means implementing multiple applications on fewer, more powerful platforms. The main reasons why companies undertake server consolidation are to simplify management by reducing complexity and eliminating server sprawl; reduce costs, particularly staff costs but also hardware, software and facilities costs; and to improve service. Data on the ROI (return on investment) of server consolidation projects is hard to come by but anecdotal evidence from big companies indicates that typical savings run into millions of dollars. A server consolidation project may also provide the opportunity to improve scalability and resilience (including disaster recovery) and consolidate storage.

Energy conservation can be achieved through server consolidation, moving VM instances away from lightly loaded computing nodes so that they become empty and can be switched to low-power mode. Marzolla et al. (2011) present VMAN, a fully decentralized algorithm for consolidating VMs in large cloud datacenters. VMAN can operate on any arbitrary initial allocation of VMs on the Cloud, iteratively producing new allocations that quickly converge towards the one maximizing the number of idle hosts. VMAN uses a simple gossip protocol to achieve efficiency, scalability and robustness to failures.

5. Solutions to resource management problems

In this section, we present significant research carried out in resource provisioning, resource allocation, resource adaptation and resource mapping for IaaS in cloud computing, and bring out the open challenges. The performance metrics are used to compare different works under resource management techniques. For each of the metric, we have assigned value as either high or medium or low. We arrive at the value by literature reading, analysis of results, relative comparisons of the results in different research papers, mathematical complexity involved, and complexity of the scheme. The metrics considered are reliability, deployment ease, Quality of Service, delay and control overhead.

Reliability is defined as the ability of machine, or system to consistently perform its intended or required function or mission, on demand without degradation or failure. In our observation of network reliability in different papers, we have considered factors such as availability of end to end functionality for customers and

ability to experience failures or systematic attacks, without impacting customers or operations. We analyzed whether the proposed system (in different research papers) scales itself with increase in number of users. If the system does not complicate itself with increase in number of users, it is said to be highly reliable. If the system works with too much constraints, then it is not reliable.

Ease of deployment refers to the easiness in implementing the system model. The value for ease of deployment has been assigned as high if the infrastructures are easily available for deployment. Quality of Service (QoS) refers to a broad collection of networking technologies and techniques. The goal of QoS is to provide guarantees on the ability of a network to deliver predictable results. The elements of network performance in our work which decides QoS includes availability (uptime), bandwidth (throughput), latency (delay), and error rate. The higher availability, higher bandwidth, lower latency, and lower error rate offers higher QoS. Delay is the time taken from point-to-point in a network. Higher delay degrades performance of the system and vice versa.

Control overhead refers to the extra consideration required by a system that is not directly related to data. In this paper, we made an analysis of whether any resources are consumed or lost in completing a process that does not contribute directly to the end-product. If the resources are lost or extra resources required during transmission of the data, it means that the control overhead is high.

5.1. Resource provisioning

The development of efficient service provisioning policies is among the major issues in cloud research. The issue here is to provide better quality of service in IaaS by provisioning the resources to the users or applications via load balancing mechanism, high availability mechanism, etc. In this context, game theoretic methods in Teng and Magoules (2010) allows us to gain an in depth analytical understanding of the service provisioning problem. Earlier game theory has been successfully applied to diverse problems such as Internet pricing, congestion control, routing, and networking. Resource provisioning can encompass three dimensions as per Sotomayor et al. (2009): hardware resources, the software available on those resources, and the time during which those resources must be guaranteed to be available. A complete resource provisioning model must allow resource consumers to specify requirements across these three dimensions, and the resource provider to efficiently satisfy those requirements.

The service provisioning procedure according to Hill and Varaiya (2009) is based on a solution of the problem of allocating bandwidth and buffers to meet several types of service requests, differentiated by bounds on the average rate and burstiness of the message and on the end-to-end delay. Here, the users decide the resources they need and the network coordinates their choices via resource pricing in order to optimize an overall measure of network performance.

Among the existing works that we have at present for resource provisioning in cloud computing, we observed that only few researchers have addressed the problem in multi-tier applications. The work given in Urgaonkar et al. (2010b) presents a model which can be best described as a analytical model using queuing networks in which the behaviour of each tier has been captured. This analytical model is able to predict parameters such as the think time, service time and visit ratio. The most recent work in this area given in Singh et al. (2010) which presents a technique to model dynamic workloads for multi-tier Web applications using k-means clustering. The method uses queuing theory to model the system reaction to the workload and to identify the number of instances required for an Amazon EC2 cloud to perform well under a given workload. Although this work does model system behavior on a per-tier basis, it does not perform multi-tier dynamic resource provisioning. In particular, database tier scaling is not considered. Also, the authors do not follow any approach toward dynamic resource management on clouds.

Inefficiency of resource provisioning leads to either overprovisioning or underprovisioning problem. Vijayakumar et al. (2010a) propose a robust cloud resource provisioning (RCRP) algorithm to minimize the total resource provisioning cost (i.e., overprovisioning and underprovisioning costs). Various types of uncertainty are considered in the algorithm. Dailey et al. (2011) propose a method for identifying and retracting overprovisioned resources in multi-tier cloud-hosted Web applications. They demonstrate the feasibility of approach in an experimental evaluation with a testbed EUCALYPTUS based cloud and a synthetic workload. But the problem is that they only address scaling of the Web server tier and a read-only database tier. In particular, they do not address software configuration management.

Buyya et al. (2011) point out many challenges in addressing the problem of enabling SLA-oriented resource allocation in data centers to satisfy competing applications demand for computing services. In particular, the user applications are becoming more complex and need multiple services to execute instead of a single service. In particular, work on cloud management in Cunningham and Holmes (2011) and Armbrust and Fox (2010) has focused on the provisioning and scaling of services within infrastructure clouds. Among the problems that are faced by the users, performance and virtualization problems are the most persistent and prevalent problems owing in part due to the fact that users have no visibility into the cloud and are thus forced to consult the cloud operators for help. The surveys indicate that to offer more effective support, clouds should develop tools to automate operators task.

Vijayakumar et al. (2010b) have considered the problem of cost-sensitive resource provisioning for adaptive data streaming applications in virtualized or cloud environments. This framework dynamically achieves the user-specified accuracy level by adapting a adaptive parameter at runtime. Chaisiri et al. (2012) have proposed an optimal cloud resource provisioning (OCRP) algorithm to provision resources offered by multiple cloud providers. The optimal solution obtained from OCRP is found by formulating and solving stochastic integer programming. The OCRP algorithm can be used as a resource provisioning tool for the emerging cloud computing market in which the tool can effectively save the total cost.

Warneke and Kao (2011) have discussed the challenges and opportunities for efficient parallel data processing in cloud environments and presented Nephele, the first data processing framework to exploit the dynamic resource provisioning offered by today IaaS clouds. They have described Nephele basic architecture and presented a performance comparison to the well-established data processing framework Hadoop. The performance evaluation gives a first impression on the ability to assign specific virtual machine types to specific tasks of a processing job, as well as the possibility to automatically allocate/deallocate virtual machines in the course of a job execution. This can help in improving overall resource utilization. and consequently reduce the processing cost.

Juve and Deelman (2012) have discussed several techniques based on resource provisioning that may be used to reduce network overheads. These techniques include: advance reservations, multi-level scheduling, and Infrastructure as a Service (IaaS). They have discussed the advantages and disadvantages of these techniques in terms of cost, performance and usability.

In Huang et al. (2011), an architectural design of on-demand service for grid computing is proposed. A profile-based approach to capture expert knowledge of scaling applications was proposed in which extra demanded resources can be more efficiently provisioned as in Jie et al. (2011). In Kee and Kesselman (2011),

Table 4
Resource provisioning schemes.

Name of the scheme	*Functioning*
Nash equilibrium approach using Game theory (Teng and Magoules, 2010)	Run time management and allocation of IaaS resources considering several criteria such as the heterogeneous distribution of resources, rational exchange behaviors of cloud users, incomplete common information and dynamic successive allocation
OpenNebula (infra-structure manager) and and Haizea (resource lease manager) (Sotomayor et al., 2009)	Allows resource consumers to specify requirements across these three dimensions-hardware resources, the software available on those resources, and the time during which those resources must be guaranteed to be available for the resource provider to efficiently satisfy those requirements
Resource pricing (Hill and Varaiya, 2009)	The provisioning procedure consists of two algorithms, one executed by the network and the other by individual users The network offers resources freely to meet their desired quality based on their own traffic parameters and delay requirements The network periodically adjusts resource prices based on user requests
Network queuing model (Urgaonkar et al., 2010b)	Presents a model based on a network of queues, where the queues represent different tiers of the application.The model sufficiently captures the behavior of tiers with significantly different performance characteristics and application idiosyncrasies such as session-based workloads, concurrency limits, and caching at intermediate tiers
Prototype provisioning (Singh et al., 2010)	Employs the k-means clustering algorithm to automatically determine the workload mix and a queuing model to predict the server capacity for a given workload mix. A prototype provisioning system evaluate its efficiency on a laboratory Linux data center running the TPC-W web benchmark
Resource provisioning (Vijayakumar et al., 2010a)	Uses virtual machines (VMs) that run on top of the Xen hypervisor. The system provides a Simple Earliest Deadline First (SEDF) scheduler that implements weighted fair sharing of the CPU capacity among all the VMs The share of CPU cycles for a particular VM can be changed at runtime
Adaptive resource provisioning (Dailey et al., 2011)	Automatic bottleneck detection and resolution under dynamic resource management which has the potential to enable cloud infrastructure providers to provide SLAs for web applications that guarantee specific response time requirements while minimizing resource utilization. Demonstrates the feasibility of the approach with a testbed EUCALYPTUS-based cloud and a synthetic workload
SLA oriented methods (Buyya et al., 2011)	Handling the process of dynamic provisioning to meet user SLAs in autonomic manner through Aneka platform. Additional resources are provisioned for applications when required and are removed when they are not necessary
Dynamic and automated framework (Armbrust and Fox, 2010)	Presents a dynamic and automated framework which can adapt the adaptive parameters to meet the specific accuracy goal, and then dynamically converge to near-optimal resource allocation to handle unexpected changes in the data distribution characteristics and/or rates
Optimal cloud resource provisioning (OCRP) (Chaisiri et al., 2012)	The demand and price uncertainty is considered using optimal cloud resource provisioning (OCRP) including deterministic equivalent formulation, sample-average approximation, and Benders decomposition

Table 5
Performance metrics for resource provisioning schemes.

Schemes	**Metrics**				
	Reliability	*Ease of deployment*	*QoS*	*Delay*	*Control overhead*
Nash equilibrium using Game theory (Teng and Magoules, 2010)	High	Medium	High	Medium	High
OpenNebula (infrastructure manager) and Haizea (resource lease manager) (Sotomayor et al., 2009)	High	High	High	High	High
Resource pricing (Hill and Varaiya, 2009)	Med	Medium	High	Medium	Medium
Network queuing model (Urgaonkar et al., 2010b)	Medium	Low	Medium	Medium	Medium
Prototype provisioning (Singh et al., 2010)	Medium	Medium	Medium	High	High
Resource provisioning (Vijayakumar et al., 2010a)	High	High	High	Medium	High
Adaptive resource provisioning (Dailey et al., 2011)	High	Medium	High	Medium	Medium
SLA oriented methods (Buyya et al., 2011)	Medium	Medium	Medium	Medium	Medium
Dynamic and automated framework (Armbrust and Fox, 2010)	High	Medium	High	Medium	Medium
Optimal cloud resource provisioning (OCRP) (Chaisiri et al., 2012)	High	Medium	High	Medium	Low

the concept of resource slot was proposed whose objective was to address uncertainty of resources availability. A binary integer program to maximize revenues and utilization of resource providers was formulated in Filali et al. (2009). However, some of the works did not consider uncertainty of future consumer demands. In Kusic and Kandasamy (2010), an optimization framework for resource provisioning was developed. This framework considered multiple client QoS classes under uncertainty of workloads (e.g., demands of computing resources). The arrival pattern of workloads is estimated by using online forecasting techniques. Miyashita et al. (2011) consider heuristic method for service reservation where prediction of demand was performed to define reservation prices.

In Chen et al. (2011), K-nearest-neighbors algorithm was applied to predict the demand of resources. Montero et al. (2011) have presented an elastic architecture for clusters that allow a flexible management of these computing platforms by: (i) supporting the execution of heterogeneous application domains; (ii) dynamically partitioning the cluster capacity, adapting it to variable demands; and (iii) efficiently isolating the cluster workloads. Moreover, this architecture is able to transparently grow the cluster capacity using an external cloud provider. Kong

et al. (2011) propose an efficient dynamic task scheduling scheme for virtualized data centers. Considering the availability and responsiveness performance, the general model of the task scheduling for virtual data centers is built and formulated as a two-objective optimization. A graceful fuzzy prediction method is given to model the uncertain workload and the vague availability of virtualized server nodes, by using fuzzy logic systems.

There are a few published papers on cloud computing performance prediction model. For instance, Vianna (2012) proposed an analytical model to predict performance for a Hadoop online prototype using intra-job pipeline parallelism with no reference to power consumption.

Xie (2010) focuses on the optimization of the MapReduce performance in heterogeneous Hadoop clusters. The work shows performance improvements for placing data across multiple nodes so that each node has a balanced data processing performance. But, it does not provide a prediction model to verify and estimate performance variations for different disks and processor architectures. The work does not analyze disk, I/O latency variation for different patterns, nor does it show any improvement in the power consumption associated with the proposed optimized data placing method.

A summary of some of the resource provisioning schemes is given in Table 4. Table 5 lists out the performance metrics of the resource provisioning schemes.

5.1.1. Open challenges in resource provisioning

The challenges in resource provisioning are as follows.

- How to make the applications hosted on the cloud to be elastic so that we can achieve economy of scale while preserving the application-specific Service Level Agreements (SLAs) such as, response time, throughput?
- How do we develop resource prediction models for facilitating proactive scaling in the cloud so that hosted applications are able to withstand the variation in workload with least drop in performance and availability?
- How resources may be provisioned to an application mix such that the SLAs of all applications are met?
- How to design resource provisioning algorithm that correctly converges to the optimal CPU allocation based on the data arrival rate and computational needs ?
- How to design a system to support n-tier clustered applications hosted on a cloud ?
- How to extend the prediction model, which is currently only used to retract over-provisioned resources, to also perform bottleneck prediction in advance, in order to overcome the virtual machine boot-up latency problem ?

5.2. Resource allocation

Resource allocation has a significant impact in cloud computing, especially in pay-per-use deployments where the number of resources are charged to application providers. The issue here is to allocate proper resources to perform the computation with minimal time and infrastructure cost. Proper resources are to be selected for specific applications in IaaS. Once the required types of resources are determined, instances of these resources are allocated to execute the task. Resource determination and allocation for each atomic task is managed by task modules.

IaaS cloud allocates resources to competing requests based on pre-defined resource allocation policies. Presently, most of the cloud providers rely on simple resource allocation policies like immediate and best effort as in http://aws.amazon.com. Amazon EC2 given in Bhowmik et al. (2010) is a public cloud which provides computing resources to general public on pay-per-use model. Zhang et al. (2011) showed that by strategically co-locating network I/O applications together, considerable performance gain could be obtained. However, they did not show how to utilize this strategy to help decision making in the cloud.

Kim et al. (2011) have presented a vision for the creation of global cloud exchange for trading services. Chabarek et al. (2010) describe mechanisms that automatically allocates service resources suitable for mobile devices in cloud computing environment supporting social media services. The model is able to recommend efficient virtualization by analyzing user context and the state of system. In addition, this model analyzes social media service resource in real time, learning user context for virtualization.

In virtualized data centers, VMs often communicate with each other by establishing virtual network topologies. However, due to VM migrations or a non-optimized allocation, the communicating VMs may end up hosted on logically distant physical nodes providing costly data transfers between each other. If the communicating VMs are allocated to the hosts in different racks or enclosures, the network communication may involve additional network switches and links, which consume significant amount of energy as per Chiaraviglio and Matta (2011).

There have been recent research efforts on the optimization in allocation of communicating applications to minimize the network data transfer overhead as in Chiaraviglio and Matta (2010b). However these works have not directly addressed the problem of energy consumption by the network infrastructure. Moreover, the proposed approaches do not optimize the placement of VMs at run-time depending on the current network load, which is effective for variable communication patterns and should be applied to virtualized data centers.

In tenant-based resource allocation model, Batini et al. (2011) recommend some work to be done to improve and continue validating the infrastructure. It is recommendable to deploy a different platform over the cloud infrastructure, such as High-Performance Computing (HPC) or scenarios such as online transactional applications. In Upton (2010), the resource allocation is entirely done by an online algorithm that is based on profiling active and idle time periods of desktop activity. On the other hand, in Mei et al. (2010), the resource allocation on a virtual desktop is entirely done by an offline algorithm that is based on resource predictions from profiling user workloads in traditional desktops.

In Morikawa and Ikebe (2011), authors propose a dynamic resource allocation method based on the load of VMs on IaaS, abbreviated as DAIaS. This method enables users to dynamically add and/or delete one or more instances on the basis of the load and the conditions specified by the user. They have implemented a prototype to evaluate the effectiveness and efficiency of DAIaS. Furthermore, they have performed an experiment to extract the prototype on a real cloud service, namely, Amazon EC2.

In He et al. (2011) , authors propose an efficient resource management solution specially designed for helping small and medium sized IaaS cloud providers to better utilise their hardware resources with minimum operational cost. Such an optimised resource utilization is achieved by a well-designed underlying hardware infrastructure, an efficient resource scheduling algorithm and a set of migrating operations of VMs. Ishakian and Sweha (2010) consider the case of a single cloud provider and address the question of how to best match customer demand in terms of both supply and price in order to maximize the providers revenue and customer satisfactions while minimizing energy cost.

In conventional congestion control, even when only a specific resource type is congested, use of all resource types is restricted. This brings down the efficiency in the use of other resource types,

and consequently the serviceability. To solve this problem, Tomita and Kuribayashi (2011) proposed a congestion control method that attempts to reduce the resource size allocated to the request that requires a large size of the congested resource type as in Hatakeyama et al. (2009). The proposed method is designed for a cloud computing environment in which both processing ability and bandwidth are allocated simultaneously and leased on a per-hour basis. The authors also considered another congestion control method which delays the allocation of resources while keeping the allocated resource size unchanged. It was found that the first method is more advantageous than second method in cases where there are many requests for services that require at least a minimum resource size to be allocated at the time when a request is generated, or for services that does not allow delaying the allocation of resources as in Yoshino et al. (2010).

Mao and Humphrey (2012) present an approach whereby the basic computing elements are virtual machines (VMs) of various sizes/costs. They dynamically allocate/deallocate VMs and schedule tasks on the most cost-efficient instances. Alvarez and Humphrey (2012) have presented an approach to data allocation for resource management in cloud computing. But the drawback is that they do not take into account factors such as the hourly billing of cloud providers, the VM startup time and the shape of the computation (single-threaded, workflow, etc.).

Pawar and Wagh (2012) present dynamic resource allocation mechanism for preemptable jobs in cloud. They propose priority based algorithm, in which they consider multiple SLA objectives of job for dynamic resource allocation. The recent trend shows that dynamic resource allocation is growing need of cloud providers for more number of users and with the less response time. Hence the on-demand resource allocation based SLA as per defined task priority helps to satisfy the efficient provisioning of cloud resources to multiple cloud users.

The dynamic resource allocation based on distributed multiple criteria decisions in computing cloud is explained in Ruiz-Alvarez and Humphrey (2011). In it, author contributes in two ways. First distributed architecture is adopted, in which resource management is divided into independent tasks, each of which is performed by Autonomous Node Agents (NA) in a cycle of three activities: (1) VMPlacement, in it suitable physical machine (PM) is found which is capable of running a given VM and then assigning VM to that PM, (2) Monitoring, in which total resources use by hosted VM are monitored by NA, (3) In VMSelection, if local accommodation is not possible, a VM need to migrate at another PM and process loops back to into placement. And second, using PROMETHEE method, NA carry out configuration in parallel through multiple criteria decision analysis. This approach is potentially more feasible in large data centers than centralized approaches.

The problem of resource allocation is considered in Yazir et al. (2010), to optimize the total profit gained from the multi-dimensional SLA contracts for multi-tier application. In it the upper bound of total profit is provided with the help of force-directed resource assignment (FRA) heuristic algorithm, in which initial solution is based on provided solution for profit upper bound problem. Next, distribution rates are fixed and local optimization step is used for improving resource sharing. Finally, a resource consolidation technique is applied to consolidate resources to determine the active (ON) servers and further optimize the resource assignment.

In Alvarez and Humphrey (2011), an automated approach to the selection of cloud storage services that can meet the user requirements is described. In Hill and Humphrey (2011), the authors goal with CSAL(Cloud Storage Abstraction Layer) is to leverage application portability to explore multi-cloud application deployments and management as well as dynamic resource allocation optimization for cost and performance metrics.

Lai et al. (2005) describe Tycoon, a market based distributed resource allocation system based on proportional share. The key advantages of Tycoon are that it allows users to differentiate the values of their jobs. Its resource acquisition latency is limited only by communication delays, and it imposes no manual bidding overhead on users. Buyya et al. (2009) have proposed architecture for market-oriented allocation of resources within clouds. They have discussed some representative platforms for cloud computing covering the state-of-the-art.

Kuribayashi (2011) have proposed an optimal joint multiple resource allocation method, assuming that both processing ability and bandwidth are allocated simultaneously for each request and rented out on an hourly basis. The allocated resources are dedicated to each service request. Venugopal et al. (2009) present a bilateral protocol for SLA negotiation using the alternate offers mechanism wherein a party is able to respond to an offer by modifying some of its terms to generate a counter offer. The authors apply this protocol to the negotiation between a resource broker and a provider for advance reservation of compute nodes, and implement and evaluate it on a real grid system.

To reduce communication overhead between consumer and provider of cloud and increase resource utilization on cloud provider side, negotiation is necessary. The algorithm in Tyagi and Pathak (2011) generates counter offers considering constraint's flexibilities to maximize the chances of acceptance. Using ranking algorithm, consumers will get suitable offers sorted according to their needs. It will reduce consumer's efforts to go through all the provided counter offers and choose best suitable one.

Apostol and Cristea (2011) focus on adding new features to the cloud resource allocation mechanism that enhances on demand elasticity. Most of the resource managers that are now on the market use static allocation. The authors propose a novel solution that uses dynamic allocation based on well defined policies. Moreover, the solution offers authentication and accountability for the actions of users which is very important for commercial aspect of public clouds.

Soundararajan et al. (2011) give an effective multi-resource allocation technique based on a unified resource-to-performance model incorporating (i) pre-existing generic knowledge about the system and inter-dependencies between system resources e.g., due to cache replacement policies and (ii) application access tracking and baseline system metrics captured on-line.

In Bobro et al. (2010), a dynamic server migration and consolidation algorithm is introduced. The algorithm provides substantial improvement over static server consolidation in reducing the amount of required capacity and the rate of service level agreement violations. Benefits accrue for workloads that are variable and can be forecast over intervals shorter than the time scale of demand variability. The management algorithm reduces the amount of physical capacity required to support a specified rate of SLA violations for a given workload by as much as 50 percent as compared to static consolidation approach.

Tai et al. (2011) present a smart load balancer, which leverages the knowledge of burstiness to predict the changes in user demands and on-the-fly shifts between the schemes that are greedy (i.e., always select the best site) and random (i.e., randomly select one) based on the predicted information. The result shows that this new load balancer can adapt quickly to the changes in user demands and thus improve performance by making a smart site selection for cloud users under both bursty and non-bursty workloads.

Vendor lock-in is one of the major issues in cloud based services. Migration from one cloud environment to another would be much more challenging than migrating within one's premise software. Since cloud computing is still relatively new, standards are still being developed. Many cloud platforms and services are proprietary, i.e., they are built on the specific standards, tools, and protocols are developed by a particular vendor for its particular

Table 6
Resource allocation schemes.

Name of the scheme	*Functioning*
Novel, non-intrusive method (Bhowmik et al., 2010)	Proposes a novel, non-intrusive method for application and remoting protocol agnostic desktop responsiveness monitoring. Moreover, desktop workload usage which enables to discover and leverage workload patterns that can lead to increased efficiency both in terms of desktop responsiveness and resource usage, is also highlighted
Market-oriented resource allocation (Zhang et al., 2011)	Considers the case of a single cloud provider and address the question how to best match customer demand in terms of both supply and price in order to maximize the providers revenue and customer satisfactions while minimizing energy cost.In particular, it models the problem as a constrained discrete-time optimal control problem and uses Model Predictive Control(MPC) to find its solution
Intelligent multi-agent model (Kim et al., 2011)	Proposes an intelligent multi-agent model based on virtualization rules for resource virtualization (IMAV) to automatically allocate service resources suitable for mobile devices. It infers user demand by analyzing and learning user context information. In addition, it allocates service resources according to use types so that users are able to utilize reliable service resources
Mixed integer optimization techniques (Chabarek et al., 2010)	Applies a generic model for router power consumption model in a set of target network configurations and uses mixed integer optimization techniques to investigate power consumption, performance and robustness in static network design and in dynamic routing
Energy-Aware Resource allocation (Chiaraviglio and Matta, 2010b)	Resource allocation is carried out by mimicking the behavior of ants, that the ants are likely to choose the path identified as a shortest path, which is indicated by a relatively higher density of pheromone left on the path compared to other possible paths
Measurement based analysis on performance (Mei et al., 2010)	Focuses on measurement based analysis on performance impact of co-locating applications in a virtualized cloud in terms of throughput and resource sharing effectiveness, including the impact of idle instances on applications that are running concurrently on the same physical host
Dynamic resource allocation method (Morikawa and Ikebe, 2011)	Proposes a dynamic resource allocation method based on the load of VMs on IaaS, which enables users to dynamically add and/or delete one or more instances on the basis of the load and the conditions specified by the user
Real time resource allocation mechanism (He et al., 2011)	Proposes an efficient resource management solution specially designed for helpingsmall and medium sized IaaS cloud providers to better utilize their hardware resources with minimum operational cost by a well-designed underlying hardware infrastructure, an efficient resource scheduling algorithm and a set of migrating operations of VMs
A dynamic scheduling and consolidation mechanism (Ishakian and Sweha, 2010)	Presents the architecture and algorithmic blueprints of a framework for workload co-location, which provides customers with the ability to formally express workload scheduling flexibilities using Directed Acyclic Graphs (DAGs), and optimizes the use of cloud resources to collocate client's workloads
Congestion control method (Tomita and Kuribayashi, 2011)	Proposes a definition of fairness in congested situation, assuming that multiple resource types are allocated simultaneously to each service request. Also, identifies a measure for evaluating fair resource allocation

Table 7
Performance metrics for resource allocation schemes.

Schemes	**Metrics**				
	Reliability	*Ease of deployment*	*QoS*	*Delay*	*Control overhead*
Novel, non-intrusive method (Bhowmik et al., 2010)	Medium	Medium	Medium	Medium	High
Market-oriented resource allocation (Zhang et al., 2011)	Medium	High	Medium	High	High
Intelligent multi-agent model (Kim et al., 2011)	Medium	High	Medium	Low	Medium
Mixed integer optimization techniques (Chabarek et al., 2010)	Medium	Medium	High	High	Medium
Energy aware resource allocation (Chiaraviglio and Matta, 2010b)	High	Medium	High	Medium	Medium
Measurement based analysis on performance (Mei et al., 2010)	Medium	High	Medium	Medium	Medium
Dynamic resource allocation method (Morikawa and Ikebe, 2011)	High	Medium	High	Low	Medium
Real time resource allocation mechanism (He et al., 2011)	High	Low	High	High	Medium
A dynamic scheduling and consolidation mechanism (Ishakian and Sweha, 2010)	Medium	High	Medium	Low	High
Congestion control method (Tomita and Kuribayashi, 2011)	High	Low	High	Low	High

cloud offering. This can make migrating off a proprietary cloud platform prohibitively complicated and expensive as given in http://www.forbes.com.

Three types of vendor lock-in can occur with cloud computing as presented in http://community.zenoss.org. (i) Platform lock-in: cloud services tend to be built on one of several possible virtualization platforms, for example VMWare or Xen. Migrating from a cloud provider using one platform to a cloud provider using a different platform could be very complicated. (ii) Data lock-in: since the cloud is still new, standards of ownership, i.e., who actually owns the data once it lives on a cloud platform, are not yet developed, which could make it complicated if cloud computing users ever decide to move data off of a cloud vendor's platform. (iii) Tools lock-in: if tools built to manage a cloud environment are not compatible with different kinds of both virtual and physical infrastructure, those tools will only be able to manage data or applications that live in the vendor's particular cloud environment.

Heterogeneous cloud computing prevents vendor lock-in, and aligns with enterprise data centers that are operating hybrid cloud models. The absence of vendor lock-in lets cloud administrators select his or her choice of hypervisors for specific tasks, or to deploy virtualized infrastructures to other enterprises without the need to consider the flavor of hypervisor in the other enterprise as in Vada and Eirik (2011).

A heterogeneous cloud is considered one that includes on-premise private clouds, public clouds and software-as-a-service clouds. Heterogeneous clouds can work with environments that are not virtualized, such as traditional data centers as discussed in Geada and Dave (2011). Heterogeneous clouds also allow for the use of piece parts, such as hypervisors, servers, and storage, from multiple vendors as given in http://www.neovise.com.

Table 6 summarizes some of the resource allocation schemes. Table 7 lists out the performance metrics of the resource allocation schemes.

Table 8
Resource mapping schemes.

Name of the scheme	*Functioning*
Mapping logical plane to underlying physical plane (Hou et al., 2009)	Presented a novel set of feasibility checks for node assignments based on graph cuts
Symmetric mapping pattern (Grehant and Demeure, 2011)	Presents the symmetric mapping pattern, an architectural pattern for the design of resource supply systems. It divides resource supply in three functions: (1) users and providers match and engage in resource supply agreements, (2) users place tasks on subscribed resource containers, and (3) providers place supplied resource containers on physical resources
Load-aware mapping (Chen et al., 2009)	Explores how to simplify VM image management and reduce image preparation overhead by the multicast file transferring and image caching/reusing. Additionally, the Load-Aware Mapping, a novel resource mapping strategy, is proposed in order to further reduce deploying overhead and make efficient use of resources
Minimum congestion mapping (Bansal et al., 2011)	Proposes a general framework for solving a natural graph mapping problem arising in cloud computing. And applying this framework to obtain offline and online approximation algorithms for workloads given by depth-d trees and complete graphs
Iterated local search based request partitioning (Leivadeas et al., 2011)	A novel request partitioning approach based on iterated local search is introduced that facilitates the cost-efficient and on-line splitting of user requests among eligible Cloud Service Providers (CSPs) within a networked cloud environment
SOA API (Xabriel et al., 2012)	The solution is designed to accept different resource usage prediction models and map QoS constraints to resources from various IaaS providers
Impatient task mapping (Mehdi et al., 2011)	Proposes batch mapping via genetic algorithms with throughput as a fitness function that can be used to map jobs to cloud resources
Distributed ensembles of virtual appliances (DEVAs) (Villegas and Sadjadi, 2011)	Requirements are inferred by observing the behavior of the system under different conditions and creating a model that can be later used to obtain approximate parameters to provide the resources. These models are usually measured by treating the application as a black-box (i.e., without employing any knowledge of the internal implementation or design)
Opportunistic resource	Adopts a simple greedy heuristic to all virtual nodes to sort in a decreasing order of their CPU constraints and places them in a queue
Sharing and topology-aware node ranking (ORSTA) (Zhang et al., 2012)	Then, maps each virtual node in the sorted queue to the unused substrate node with the highest rank Hence, minimizes the length of the substrate paths that virtual links are mapped to
Mapping a virtual network onto a substrate network (Lu and Turner (2006))	Developed an effective method (using backbone mapping) for computing high quality mappings of virtual networks onto substrate networks. The computed virtual networks are constructed to have sufficient capacity to accommodate any traffic pattern allowed by user-specified traffic constraints

5.2.1. *Open challenges in resource allocation*

The challenges in resource allocation are as follows.

- How to design a resource allocation scheme that spans several clusters and data centers?
- How to devise a mechanism that allows controlling the trade-off between the cost of reconfiguration and maximizing the cloud utility?
- How to develop a tree-based protocol for resource management in cloud environments and how such a protocol compares with a gossip-based protocol with similar functionality?
- How to bring out the techniques for allocation of services to applications depending on energy efficiency and expenditure of service providers?
- How and when to reallocate VMs to minimize the power drawn by the cooling system, while preserving a safe temperature of the resources and minimizing the migration overhead and performance degradation?
- How to design SLA-oriented resource allocation strategies that encompass customer-driven service management, computational risk management, and autonomic management of clouds in order to improve the system efficiency, minimize violation of SLAs, and improve profitability of service providers.
- How to move from one cloud to another cloud considering vendor lock-in issues? What if a good part of our application infrastructure resides with a single cloud provider?

5.3. *Resource mapping*

Mapping of virtual resources to physical resources has an impact on cloud clients. Resource mapping is a system-building process that enables a community to identify existing resources and match those resources to a specific purpose. The issue here is to maximize cloud utilization in IaaS by calculating the capacity of application requirements so that minimal cloud computing infrastructure devices shall be procured and maintained. This can be achieved by using cognitive architecture that automatically builds a model of the machine behavior based on prior training data.

In a cloud computing environment, a logical network (i.e a set of virtual machines) must be deployed on to physical network (servers). This requires mapping of VMs to physical resources. The mapping problem is dealt in Hou et al. (2009) which translates virtual machines assignment onto physical servers and assigns flows in the physical network with bandwidth allocation so that requirements of logical communication can be met.

An allocation which is directed by a decision system under user control can result in high resource supply costs and an allocation directed by a decision system under provider's control can result in low user-perceived resource value. Instead of compromising with them, symmetric mapping referred in Grehant and Demeure (2011) builds on these differences from the system design. It relies on the idea that a system benefits from the involvement of different participants if it induces them to adopt predictable behaviors and uses these behaviors as part of its mechanism.

Chen et al. (2009) have worked towards an efficient resource management system for on-line virtual cluster provision. In particular, they focus on two crucial problems namely efficient VM image management and intelligent resource mapping. Additionally, they have proposed an intelligent resource mapping strategy, named load-aware mapping, in order to reduce deploying overhead and balance resource utilization.

In cloud computing, the underlying resource is a physical network (also called the substrate) consisting of servers that are inter-connected via communication links. The allocation of a workload to the substrate can be viewed as mapping one graph into another. This consists of two aspects: (a) node-mapping, the assignment of processes to servers, and (b) path-mapping, the

Table 9
Performance metrics for resource mapping schemes.

Schemes	Metrics				
	Reliability	*Ease of deployment*	*QoS*	*Delay*	*Control overhead*
Mapping logical plane to underlying physical plane (Hou et al., 2009)	Medium	Med	High	Medium	High
Symmetric mapping pattern (Grehant and Demeure, 2011)	Medium	Medium	Medium	Medium	Medium
Load-aware mapping (Chen et al., 2009)	Medium	Medium	Medium	Medium	Medium
Minimum congestion mapping (Bansal et al., 2011)	Medium	High	Medium	Low	Medium
Iterated local search based request partitioning (Leivadeas et al., 2011)	High	Medium	Medium	High	High
SOA API (Xabriel et al., 2012)	Medium	Medium	Medium	Medium	High
Impatient task mapping (Mehdi et al., 2011)	Medium	High	Medium	High	High
Distributed ensembles of virtual appliances (DEVAs) (Villegas and Sadjadi, 2011)	Medium	Medium	Medium	High	Medium
Opportunistic resource sharing and topology-aware node ranking (Zhang et al., 2012)	High	Medium	High	Medium	Medium
Mapping a virtual network onto a substrate network (Lu and Turner, 2006)	High	Medium	Medium	Medium	High

assignment of each communication request (i.e. edge between two processes) to a path in the substrate between the respective servers as in Bansal et al. (2011) .

In Leivadeas et al. (2011), the cloud resource mapping problem over a networked cloud computing environment is studied by providing high performance algorithms in terms of embedding effectiveness and run time complexity. Within the proposed framework, an Iterated Local Search(ILS)-based heuristic is employed to provide a cost efficient resource allocation realizing the partitioning of the user request.

Xabriel et al. (2012) propose a SOA API, in which users provide a cloud application model and get back possible resource allocations in an IaaS provider. The solution emphasizes the assurance of quality of service (QoS) metrics embedded in the application model. An initial mapping is done based on heuristics, and then the application performance is monitored to provide scaling suggestions. Underneath the API, the solution is designed to accept different resource usage prediction models and can map QoS constraints to resources from various IaaS providers.

The study carried out in Mehdi et al. (2011) proposes an algorithm that can find a fast mapping using genetic algorithm and ensures all task deadlines. Mapping time and makespan are the performance metrics that are used to evaluate the proposed system.

In Villegas and Sadjadi (2011), the design and implementation of an Infrastructure as a Service cloud manager is discussed, such that non-functional requirements determined during the requirements analysis phase can be mapped to properties for a group of virtual appliances running the application. Zhang et al. (2012) re-examine the virtual network mapping problem through two novel aspects, opportunistic resource sharing and topology-aware node ranking, and proposed a novel framework called opportunistic resource sharing topology-aware node ranking (ORSTA), which provides efficient physical resource utilization and deployment.

Lu and Turner (2006) have developed an effective method for computing high quality mappings of virtual networks onto substrate networks. The computed virtual networks are constructed to have sufficient capacity to accommodate any traffic pattern allowed by user-specified traffic constraints. The computational method produces high quality results that are close to a lower bound and is fast enough to handle networks of practical size.

Table 8 summarizes some of the resource mapping schemes. Table 9 lists out performance metrics of the resource mapping schemes.

5.3.1. Open challenges in resource mapping

The challenges in resource mapping are as follows.

- Will all applications run in the cloud? Should we attempt to port all of our existing applications to the cloud?
- Mapping the logical nodes on to the physical nodes and finding physical resource allocation to meet the logical network demands, subject to physical network constraints.
- Designing algorithm that can find a fast mapping using genetic algorithms to speed up the mapping process and ensures the respecting of all task deadlines.
- Minimizing the cost of mapping the request into the substrate (embedding cost).
- Mapping the application's attributes to cloud attributes to validate whether cloud services are suitable for the application, and identifying which types of services to use.
- Evaluating cloud service providers as possible candidates for hosting the applications, identifying which types of services are available from the chosen provider(s), and then determining specific implementation attributes of the services offered.
- Developing models that are able to predict applications performance considering different parameters such as processor, memory, network and disk usage.
- Load balancing on substrate networks and partial reconfiguration of virtual networks.

5.4. Resource adaptation

The primary reason for adapting cloud computing from a user perspective is to move from the model of capital expenditure (CAPEX) to operational expenditure(OPEX). Instead of buying IT resources like machines, storage devices etc. and employing personnel for operating, maintaining etc., a company pays another company (the provider) for the actual resources used (pay-as-you-go). An important aspect of this is that a company no longer needs to overprovision its IT resources. It is typical today, when a company invests in its own resources, that the amount of resources invested in corresponds to the maximum amount of resources needed at peak times with the result that much of these resources are not needed at all during regular periods.

The key conceptual component of framework discussed in Zhu and Agrawal (2010) is a dynamic resource adaptation algorithm, which is based on control theory. A reinforcement learning guided control policy is applied to adjust the adaptive parameters so that application benefit is maximized within the time constraint using modest overhead. Such a control model can be trained fast and accurately. Furthermore, a resource model is proposed to map any given combination of values of adaptive parameters to resource requirements in order to guarantee that the resource cost stays under the budget.

Duong et al. (2009) have proposed an extensible framework for dynamic resource provisioning and adaptation in IaaS clouds. The core of this framework is a set of resource adaptation algorithms which utilize workload and resource information to make

informed provisioning decisions in light of dynamically changing users demands.

Jung et al. (2010) present Mistral, a holistic optimization system that balances power consumption, application performance, and transient power/performance costs due to adaptation actions and decision making in a single unified framework. By doing so, it can dynamically choose from a variety of actions with differing effects in a multiple application, and dynamic workload environment.

Calyam et al. (2011) use OnTimeMeasure-enabled performance intelligence to compare utility-driven resource allocation schemes in virtual desktop clouds. The results from the global environment for network innovations(GENI) infrastructure experiments carried out by the authors demonstrated how performance intelligence enables autonomic nature of FI (Future Internet) applications to mitigate the costly resource overprovisioning and user QoE (Quality of Experience) guesswork, which are common in the current Internet.

Senna et al. (2011) present an architecture for management and adaptation of virtual networks on clouds. Their infrastructure allows the creation of virtual networks on demand, associated with the execution of workflows, isolating and protecting the user environment. The virtual networks used in workflow execution has its performance monitored by the manager which acts preemptively in the case of performance dropping below stated requirements.

Flexibility enables the adaptation of cloud solutions to all users to ensure that they get exactly what they want and need. By that, cloud computing not only introduces a new way of how to perform computations over the Internet, but some observers also observed that it holds the potential to solve a range of ICT (information and communications technology) problems identified within disparate areas such as education, healthcare, climate change, terrorism, economics etc. as per Schubert (2010).

Resource adaptation of the virtual hosts should dynamically scale to the updated demands (cloud computing) as well as co-locate applications to save on energy consumption (green computing) as per Sclater (2011). Most importantly, resource transitions during workload surges should occur while minimizing the expected loss due to mismatches of the resource predictions and actual workload demands.

A system that can automatically scale its share of infrastructure resources is designed in Charalambous (2010). The adaptation manager monitors and autonomically allocates resources to users in a dynamic way. However, this centralized approach cannot fit in the future multiprovider cloud environment, since different providers may not want to be controlled by such a centralized manager.

There have been great advances towards automatically managing collections of inter-related and context-dependent VMs (i.e. a service) in a holistic manner by using policies and rules. The degree of resource management, the bonding to the underlying API and coordinating resources spread across several clouds in a seamless manner while maintaining the performance objectives are major concerns that deserve further study. Also, dynamically scaling LBS (location based services) and its effects on whole application scalability are reported in Vaquero et al. (2011) and Amazon auto scaling service.

The goal of authors in Baldine et al. (2009) is to manage the network substrate as a first-class resource that can be co-scheduled and co-allocated along with compute and storage resources, to instantiate a complete built-to-order network slice hosting a guest application, service, network experiment, or software environment. The networked cloud hosting substrate can incorporate network resources from multiple transit providers and server hosting or other resources from multiple edge sites (a multi-domain substrate).

Jung et al. (2008) propose a novel hybrid approach for enabling autonomic behavior that uses queuing theoretic models along with optimization techniques to predict system behavior and automatically generate optimal system configurations. Marshall et al. (2010) have implemented a resource manager, built on the Nimbus toolkit to dynamically and securely extend existing physical clusters into the cloud. The elastic site manager interfaces directly with local resource managers, such as Torque.

Raghavan et al. (2009) present the design and implementation of distributed rate limiters, which work together to enforce a global rate limit across traffic aggregates at multiple sites, enabling the coordinated policing of a cloud-based service network traffic. This abstraction not only enforces a global limit, but also ensures that congestion-responsive transport-layer flows behave as if they traversed a single, shared limiter.

Table 10 summarizes some of the resource adaptation schemes. Table 11 lists out the performance metrics of the resource adaptation schemes.

5.4.1. Open challenges in resource adaptation

- How is the demand for using the cloud services provided by the vendor? Is it mostly constant or widely varying?
- What is the frequency of usage of cloud resources? Is it highly frequent? Very frequent usage in fact makes less economic sense to go for cloud based pay-as-you-go model.
- Do we need highly customized services/API (application programming interfaces) to be exposed by the vendor? Cloud vendors would not find it economically attractive to provide highly customized services and hence price for enterprise (users of cloud) might also be not very attractive.
- Is the application mission critical? A mission critical application would need very stringent SLAs, which cloud vendors could not be able to satisfy as yet. An industry or application with highly stringent compliance requirements might still not find it suitable to consume key services from a vendor due to inherent risks involved.
- Can a problem occurence in our slice environment that impacts our QoE be identified and notified to our application to adapt and heal?
- Can problem information also be shared with us and our application service provider if our application cannot automatically heal itself?
- Can we monitor all the detailed active (e.g., Ping, traceroute, iperf) and passive (e.g., TCP dump, netow, router-interface statistics) measurements at end-to-end hop, link, path and slice levels across multiple federated ISP domains?
- Can we analyze all the measurements to offline provision adequate resources to deliver satisfactory user QoE, and online i.e., real-time identify anomalous events impacting user QoE?

6. Conclusions

After so many years, cloud computing today is the beginning of network based computing over Internet in force. It is the technology of the decade and is the enabling element for new computing models. Traditional monitoring and management systems are typically centralized. These approaches will not scale to potentially millions of management objects in cloud systems. Approaches that are more distributed and have scalability properties that allow easy scale-up and scale-down of the monitoring and management systems to elastically meet cloud requirements are needed. Infrastructure-as-a-Service (IaaS) cloud computing provides the

Table 10
Resource adaptation schemes.

Name of the scheme	*Functioning*
Reinforcement learning guided control policy (Zhu and Agrawal, 2010)	Proposes a framework that is a multi-input multi-output feedback control model-based dynamic resource provisioning algorithm which adopts reinforcement learning to adjust adaptive parameters to guarantee the optimal application benefit within the time constraint
Web-service based prototype (Duong et al., 2009)	Developed a fully functional web-service based prototype framework, and used it for performance evaluation of various resource adaptation algorithms under different realistic settings, e.g. when input data such as job's wall times are inaccurate
Mistral framework (Jung et al., 2010)	A framework that optimizes power consumption, performance benefits, and the transient costs incurred by various adaptations and the controller itself to maximize overall utility in multiple distributed applications and large-scale infrastructures through a multi-level adaptation hierarchy and scalable optimization algorithm
OnTimeMeasure service (Calyam et al., 2011)	Presents an application – adaptation case study that uses OnTimeMeasure-enabled performance intelligence in the context of dynamic resource allocation within thin-client based virtual desktop clouds to increase cloud scalability, while simultaneously delivering satisfactory user quality-of-experience
Virtual networks (Senna et al., 2011)	Proposes virtual networks architecture as a mechanism in cloud computing that can aggregate traffic isolation, improving security and facilitating pricing, also allowing customers to act in cases where the performance is not in accordance with the contract for services between the customer and the provider of the cloud
DNS-based Load Balancing (Vaquero et al., 2011)	Proposes a system that contain the appropriate elements so that applications can be scaled by replicating VMs (or application containers), by reconfiguring them on the fly, and by adding load balancers in front of these replicas that can scale by themselves(even relying on DNS to do so)
On-demand creation of virtual network re-sources (Amazon auto scaling service)	Proposes a mechanism that allows to scale the capacity up or down automatically according to conditions that customers define and also ensuring that the number of Amazon EC2 instances that customer uses increases seamlessly during demand spikes to maintain performance, and decreases automatically during demand lulls to minimize costs
A control framework for a Multi-level cloud Network (Baldine et al., 2009)	Explains advancement in cloud resource control to cloud networks with multiple substrate providers, including networktransit providers and enables cloud applications to request virtual servers at multiple points in the network, together with bandwidth-provisioned network pipes and other network resources to interconnect them
Hybrid approach (Jung et al., 2008)	Proposes a mechanism for providing dynamic management in virtualized consolidated server environments that host multiple multi-tier applications using layered queuing models for Xen-based virtual machine environments, which is a novel optimization technique that uses a combination of bin packing and gradient search

Table 11
Performance metrics for resource adaptation schemes.

Schemes	**Metrics**				
	Reliability	*Ease of deployment*	*QoS*	*Delay*	*Control overhead*
Reinforcement learning guided control policy (Zhu and Agrawal, 2010)	High	Medium	High	Medium	High
Web-service based prototype (Duong et al., 2009)	Medium	High	Medium	Medium	High
Mistral framework (Jung et al., 2010)	Medium	Low	Medium	High	High
OnTimeMeasure service (Calyam et al., 2011)	Medium	Medium	High	Medium	Low
Virtual networks (Senna et al., 2011)	Medium	Medium	Medium	High	High
DNS-based load balancing (Vaquero et al., 2011)	High	Medium	High	Medium	Medium
On-demand creation of virtual network resources (Amazon auto scaling service)	High	Medium	High	Medium	High
A control framework for a multi-level cloud network (Baldine et al. (2009))	Medium	Medium	Medium	Medium	Medium
Hybrid approach (Jung et al., 2008)	High	Medium	High	Medium	Medium

ability to dynamically acquire extra or release existing computing resources on-demand to adapt to dynamic application workloads.

The success of companies offering Infrastructure-as-a-Service (IaaS) based on cloud computing is a strong indication that cloud computing will become increasingly more important over time. Meanwhile, as the debate over the exact definition of cloud computing continues in academic circles and technical chat rooms, the reality of cloud computing is giving companies cost efficiencies and flexibility that have never before been possible.

This paper presented a survey of resource management in IaaS based cloud computing by considering schemes such as resource provisioning, resource allocation, resource mapping and resource adaptation. It is observed that there are many issues to be addressed in cloud resource management with respect to flexibility, scalability, adaptability, customization and reusability. Also, performance metrics such as delay, bandwidth overhead, computation overhead, reliability, security and Quality of Experience have to be taken into consideration while designing a resource management scheme. The intelligent computational and cognitive software agents may provide flexible, adaptable and customized services. Human reasoning can be embedded in agents by using cognitive models and may provide better performance metric values compared to traditional classical approaches.

References

Alvarez AR, Marty Humphrey. An automated approach to cloud storage service selection. In: Proceedings of the 2nd international workshop on scientific cloud computing, vol. 8, no. 3; 2011. p. 39–48.
Alvarez AR, Humphrey M. A model and decision procedure for data storage in cloud computing. In: Proceedings of the IEEE/ACM international symposium on cluster, cloud, and grid computing , vol. 12, no. 4; 2012. p. 50–2.
Amazon auto scaling service. Available online at: http://aws.amazon.com/autoscaling/ (accessed 16.08.12).
Apostol E, Valentin Cristea. Policy based resource allocation in cloud systems. In: Proceedings of the 2011 international conference on P2P, parallel, grid, cloud and internet computing, vol. 7, no. 1; 2011. p. 165–72.
Armbrust, Armando Fox. A view of cloud computing. In: Proceedings of the communications of the ACM, vol. 53, no. 4; 2010. p. 50–8.
Available online at: http://aws.amazon.com/ec2/ (accessed on 09.07.2012).
Available online at: http://cloudcomputing.sys-con.com/node/612375 (accessed on 12.07.2012).

Available online at: http://community.zenoss.org/blogs/zenossblog/2010/06/09/three-cloud-lock-in-considerations (accessed on 02.08.2013).
Available online at: http://searchcloudcomputing.techtarget.com/report/Cloud-management-tools-guide-for-beginners (accessed on 01.08.2013).
Available online at: http://technet.microsoft.com/en-us/magazine/dn271884.aspx (accessed on 02.08.2013).
Available online at: http://www.forbes.com/sites/joemckendrick/2011/11/20/cloud-computings-vendor-lock-in-problem-why-the-industry-is-taking-a-step-backwards/ (accessed on 02.08.2013).
Available online at: http://www.ibm.com/developerworks/web/library/wa-cloudgrid (accessed on 10.07.2012).
Available online at: http://www.neovise.com/cloud-computing-in-2012-what-is-already-happening (accessed on 02.08.2013).
Baldine I, Xin Y, Evans D, Heerman C, Chase J, Marupadi V, Yumerefendi A. The missing link: putting the network in networked cloud computing. In: Proceedings of the international conference on virtual computing initiative, vol. 8, no. 5; 2009. p. 212–30.
Bansal N, Kang-Won Lee, Viswanath Nagarajan, Murtaza Zafer. Minimum congestion mapping in a cloud. In: Proceedings of the 2011 IEEE conference on PODC, vol. 6, no. 5; 2011. p. 267–76.
Batini C, Simone Grega, Andrea Maurino. Optimal enterprise data architecture. In: Proceedings of the 19th ACM international symposium on high performance distributed computing, vol. 8, no. 4; 2011. p. 541–47.
Baun C, Kunze M. Building a private cloud with eucalyptus. In: Proceedings of the international conference on E-science workshops, vol. 4, no. 4; 2009. p. 33–8.
Bhowmik R, Kochut A, Beaty K. Managing responsiveness of virtual desktops using passive monitoring. In: Proceedings of the IEEE integrated network management symposium, vol. 28, no. 4; 2010. p. 45–51.
Bobro N, Andrzej Kochut, Kirk Beaty. Dynamic placement of virtual machines for managing SLA violations. In: Proceedings of the 8th USENIX conference on file and storage technologies, vol. 10, no. 3; 2010. p. 20–21.
Buyya, Ranjan R, Modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: challenges and opportunities. In: Proceedings of the international conference on high performance computing and simulation, vol. 8, no. 4; 2009. p. 29–35.
Buyya Rajkumar, Ranjan Rajiv. Federated resource management in grid and cloud computing systems. J Future Gener Comput Syst 2011;26(5):1189–91.
Buyya Rajkumar, Shin Yeo Chee, Venugopal Srikumar. Market-oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities. J Future Gener Comput Syst Arch 2009;25(6):599–616.
Buyya R, Saurabh Kumar Garg, Rodrigo N Calheiros. SLA-oriented resource provisioning for cloud computing: challenges, architecture, and solutions. In: Proceedings of the international conference on cloud and service computing, vol. 17, no. 5; 2011. p. 71–9.
Calyam Prasad, Sridharan Munkundan, Xu Yingxiao, Zhu Kunpeng, Berryman Alex, Patal Rohit. Enabling performance intelligence for application adaptation in the future internet. J Commun Networks 2011;13(6):67–78.
Chabarek J, Sommers J, Barford P, Estan C, Wright DTS. Power awareness in network design and routing. In: Proceedings of the 27th IEEE conference on computer communications, vol. 20, no. 4; 2010. p. 457–65.
Chaisiri S, Bu-Sung Lee, Dusit Niyato. Optimization of resource provisioning cost in cloud computing. In: IEEE transactions on service computing, vol.5, no. 2; 2012. p. 67–78.
Charalambous T. Decision and control. In: Proceedings of the 49th IEEE conference on CDC, vol. 28, no. 11; 2010. p. 3778–83.
Chase JS, Darrell C Anderson, Prachi N Thakar, Amin M Vahdat. Managing energy and server resources in hosting centers. In: Proceedings of 11th IEEE/ACM international conference on grid computing (GRID), vol. 12, no. 4; 2010. p. 50–2.
Chen Y, Tianyu Wo, Jianxin Li. An efficient resource management system for on-line virtual cluster provision. In: Proceedings of the 2009 IEEE international conference on cloud computing, vol. 12, no. 3; 2009. p. 72–9.
Chen J, Soundararajan G, Amza C. Autonomic provisioning of backend databases in dynamic content web servers. In: Proceedings of the IEEE international conference on autonomic computing, vol. 7, no. 3; 2011. p. 231–42.
Chiaraviglio L, Matta I, GreenCoop. Co-operative green routing with energy efficient servers. In: Proceedings of 1st ACM international conference on energy-efficient computing and networking, vol. 15, no. 8; 2010a. p. 191–4.
Chiaraviglio L, Matta I. Resource allocation using energy-efficient servers. In: Proceedings of the 1st ACM international conference on energy-efficient computing and networking, vol. 6, no. 4; 2010b. p. 54–67.
Chiaraviglio L, Matta I. GreenCoop: co-operative green routing with energy efficient servers. In: Proceedings of the 1st ACM international conference on energy-efficient computing and networking, vol. 3, no. 2; 2011. p. 191–94.
Cunningham S, Holmes G. Developing innovative applications of machine learning. In: Proceedings of the Southeast Asia regional computer confederation conference, vol. 6, no. 4; 2011. p. 67–76.
Dai W, Rubin. Service-oriented knowledge management platform. In: Proceedings of the 13th IEEE international conference on information reuse and integration, vol. 15. no. 4; 2012. p. 255–9.
Dailey MN, Carrera David, Janecek Paul. Adaptive resource provisioning for read intensive multi-tier applications in the cloud. J Future Gener Comput Syst 2011;27(3):871–9.
Duong TNB, Xiaorong Li, Rick Siow Mong Goh. A framework for dynamic resource provisioning and adaptation in IaaS clouds. In: Proceedings of the 2011 IEEE third international conference on cloud computing technology and science, vol. 5, no. 4; 2009. p. 312–9.
Eludiora Safiriyu, Abiona Olatunde, Oluwatope Ayodeji, Oluwaranti Adeniran, Onime Clement, Kehinde Lawrence. A user identity management protocol for cloud computing paradigm. Int J Commun Network Syst Sci 2011;1(4):152–63.
Filali A, Hafid AS, Gendreau M. Adaptive resources provisioning for grid applications and services. In: Proceedings of the IEEE international conference on network communications, vol. 2, no. 14; 2009. p. 607–14.
Geada G, Dave D. The case for the heterogeneous cloud. Cloud Comput J 2011;11(3):521–5.
Grehant X, Isabelle Demeure. Symmetric mapping: an architectural pattern for resource supply in Grids and clouds. In: Proceedings of the 2011 IEEE conference on IPDPS , vol. 11, no. 4; 2011. p. 1–8.
Gupta M, Singh S. Greening of the internet. In: Proceedings of the ACM conference on applications, technologies, architectures, and protocols for computer communication, vol. 15, no. 5; 2009. p. 19–26.
Haider A, Richard Potter, Akihiro Nakao. Challenges in resource allocation in network virtualization. In: Proceedings of the 20th ITC specialist seminar on network virtualization, vol. 38, no. 4; 2009. p. 34–40.
Harmer T, Wright P. Provider-independent use of the cloud. In: Proceedings of the international workshop on cloud computing, vol. 5, no. 4; 2009. p. 296–305.
Hatakeyama K, Osana Y, Tanabe M, Kuribayashi S. Proposed congestion control method reducing the size of required resource for all-IP. In: Proceedings of the IEEE pacific rim conference on communications, computers and signal processing, vol. 11, no. 3; 2009. p. 124–35.
He S, Li Guo, Yike Guo. Real time elastic cloud management for limited resources. In: Proceedings of the 4th IEEE international conference on cloud computing, vol. 3, no. 6; 2011. p. 622–29.
Hill Z, Humphrey M. CSAL: a cloud storage abstraction layer to enable portable cloud applications. In: Proceedings of the 2nd IEEE international conference on cloud computing technology and science, vol. 12, no. 5; 2011. p. 67–72.
Hill M, Varaiya P. An algorithm for optimal service provisioning using resource pricing. In: Proceedings of the 13th IEEE international conference on networking for global communications, vol. 1, no. 2; 2009. p. 368–73.
Hou Y, Murtaza Zafer, Kang-won Lee, Dinesh Verma, Kin K Leung. On the mapping between logical and physical topologies. In: Proceedings of the IEEE conference on COMSNETS, vol. 6, no. 4; 2009. p. 67–75.
Huang Z, He C, Wu J. Architecture design, and implementation. In: Proceedings of the 11th international conference on parallel and distributed systems, vol. 10, no. 7; 2011. p. 65–75.
Ibrahim S, Shi X. Evaluating mapreduce on virtual machines: the hadoop case. In: Proceedings of the 1st international conference on cloud computing, vol. 5, no. 4; 2009. p. 345–50.
Ishakian V, Raymond Sweha. Dynamic pricing for efficient workload colocation. In: Proceedings of the 4th IEEE international conference on utility and cloud computing, vol. 7, no. 6; 2010. p. 117–25.
Jie Y, Jie Q, Ying L. A profile-based approach to just-in-time scalability for cloud applications. In: Proceedings of the IEEE international conference on cloud computing, vol. 3, no. 2; 2011. p. 101–18.
Jung G, Kaustubh R Joshi, Matti A Hiltunen. Generating adaptation policies for multi-tier applications in consolidated server environments. In: Proceedings of the 2008 international conference on autonomic computing, vol. 11, no. 4; 2008. p. 23–32.
Jung, Matti A Hiltunen, Kaustubh R Joshi, Richard D Schlichting, Calton Pu. Mistral: dynamically managing power, performance, and adaptation cost in cloud infrastructures. In: Proceedings of the 2010 IEEE 30th international conference on distributed computing systems, vol. 11, no. 5; 2010. p. 18–31.
Juve G, Deelman E. Resource provisioning options for large-scale scientific workflows. In: Proceedings of the 4th IEEE international conference on E-science, vol. 13, no. 7; 2012. p. 608–13.
Keahey K, Foster I, Freeman T, Zhang X. Achieving quality of service and quality of life in the grid, scientific programming. J Future Gener Comput Syst 2011;13(4):265–75.
Kee Y, Kesselman C. Grid resource abstraction, virtualization, and provisioning for time-target applications. In: Proceedings of the IEEE international symposium on cluster computing and the grid, vol. 11, no. 3; 2011. p. 199–203.
Kehey K, Tsugua. Sky computing. In: Proceedings of IEEE international conference on internet computing, vol. 13, no. 5; 2010. p. 43–51.
Kim M, Hanku Lee, Hyogun Yoon, Jee-In Kim, HyungSeok Kim. IMAV: an intelligent multi-agent model based on cloud computing for resource virtualization. In: Proceedings of the international conference on information and electronics engineering, vol. 6, no. 5; 2011. p. 30–5.
Kong Xiangzhen, Lin Chuang, Jiang Yixin, Yan Wei, Chu Xiaowen. Efficient dynamic task scheduling in virtualized data centers with fuzzy prediction. J Network Comput Appl 2011;34(4):1068–77.
Kuribayashi Shinichi. Optimal joint multiple resource allocation method for cloud computing environments. J Res Rev Comput Sci 2011;2(1):155–68.
Kusic D, Kandasamy N. Risk-aware limited lookahead control for dynamic resource provisioning in enterprise computing systems. In: Proceedings of the IEEE international conference on autonomic computing, vol. 10, no. 3; 2010. p. 337–50.
Lai Kevin, Rasmusson Lars, Adar Eytan, Zhang Li, Bernardo A. Tycoon: an implementation of a distributed, market-based resource allocation system. J Multiagent Grid Syst 2005;1(3):169–82.
Lakshmi SVSS, Sarwani Lalita, Nalini Tuveera M. A study on green computing: the future computing and eco-friendly technology. J Eng Res Appl 2012;2(4):1282–5.
Leivadeas A, Papagianni C, Papavassiliou S. Efficient resource mapping framework over networked clouds via iterated local search based request partitioning. IEEE Trans Parallel Distributed Syst Spec Sec Cloud Comput 2011;11(3):521–5.

Leostream A. White paper on server consolidation technologies. Available at: http://www.slashdocs.com/mztqnz/server-consolidation-technologies-a-leostream-white-paper.html (accessed on 02.08.2013).
Lombardi Flavio, Di Pietro Roberto. Secure virtualization for cloud computing. J Network Comput Appl 2011;41(1):45–52.
Lu J, Turner J. Efficient mapping of virtual networks onto a shared substrate. Technical report on Washington University WUCSE-2006, vol. 1, no. 1; 2006. p. 34–49.
Mao M, Marty Humphrey. Auto-scaling to minimize cost and meet application deadlines in cloud workflows. In: Proceedings of the international conference on high performance computing, networking, storage and analysis, vol. 37, no. 4; 2012. p. 337–48.
Marshall P, Kate Keahey, Tim Freeman. Elastic site: using clouds to elastically extend site resources. In: Proceedings of the 10th IEEE/ACM international conference on cluster, cloud and grid computing, vol. 13, no. 5; 2010. p. 43–52.
Marzolla M, Ozalp Babaoglu, Fabio Panzieri. Server consolidation in clouds through gossiping. In: Proceedings of the 2011 IEEE international symposium on a world of wireless, mobile and multimedia network (WoWMoM), vol. 1, no. 1; 2011. p. 12–9.
Mehdi Nawfal A, Mamat Ali, Ibrahim Hamidah, Subramaniam Shamala K. Impatient task mapping in elastic cloud using genetic algorithm. J Comput Sci 2011;9(4):877–83.
Mei Y, Ling Liu Xing Pu, Sankaran Sivathanu. Cloud computing performance measurements and analysis of network I/O applications in virtualized cloud. In: Proceedings of the 3rd IEEE international conference on cloud computing, vol. 5, no. 6; 2010. p. 59–66.
Miyashita K, Masuda K, Higashitani F. Coordinating service allocation through flexible reservation. IEEE Trans Serv Comput 2011;5(2):65–9.
Montero Ruben S, Moreno-Vozmediano Rafael, Llorente Ignacio M. An elasticity model for high throughput computing clusters. J Parallel Distributed Comput Arch 2011;71(6):750–7.
Morikawa T, Ikebe M. Proposal and evaluation of a dynamic resource allocation method based on the load of VMs on IaaS. In: Proceedings of the 4th IFIP international conference on new technologies, mobility and security, vol. 5, no. 6; 2011. p. 1–6.
Nimbus. Available online at: http://www.nimbusproject.org (accessed on 09.06.2012).
Nurmi D, Wolski R. The eucalyptus open-source cloud-computing system. In: Proceedings of the international conference cloud computing and its applications, vol. 10, no. 6; 2010. p. 234–6.
Nurmi D, Wolski R, Grzegorczyk C, Obertelli G, Soman S, Youseff L, et al. The eucalyptus open-source cloud-computing system in cloud computing and applications. In: Proceedings of the CCA09, vol. 8, no. 4; 2009. p. 10–2.
OpenNebula project available at: http://www.opennebula.org (accessed on 09.05.2012).
Pawar CS, Wagh RB. A review of resource allocation policies in cloud computing. In: Proceedings of the national conference on emerging trends in information technology (NCETIT-2012), world journal of science and technology, vol. 2, no. 3; 2012. p. 165–67.
Raghavan B, Vishwanath K, Ramabhadran S, Yocum K, Snoeren AC. Cloud control with distributed rate limiting. In: Proceedings of the 2007 conference on applications, technologies, architectures, and protocols for computer communications, vol. 17, no. 4; 2009. p. 337–48.
Reservoir project available at: http://www.reservoir-fp7.eu (accessed on 09.05.2012).
Ruiz-Alvarez A, Humphrey M. An automated approach to cloud storage service selection. In: Proceedings of the 2nd workshop on scientific cloud computing, vol. 10, no. 2; 2011. p. 44–56.
Ruth P, McGachey P, Xu D. Viocluster: virtualization for dynamic computational domains. In: Proceedings of the IEEE international conference on cluster computing, vol. 4, no. 3; 2010. p. 110–4.
Schubert. The future of computing opportunities for European cloud computing beyond 2010. Available from http://bit.ly/b7faxz (accessed on 26.07.12).
Sclater. Cloud computing in education. Policy Brief Unesco Inst Inf Technol Educ 2011;11(5):78–81.
Senna CR, Milton A Soares Jr, Luiz F Bittencourt, Edmundo RM Madeira. An architecture for adaptation of virtual networks on clouds. In: Proceedings of the network operations and management symposium, vol. 13, no. 7; 2011. p. 1–8.
Siddhisena B, Lakmal Wruasawithana, Mithila Mendis. Next generation mutitenant virtualization cloud computing platform. In: Proceedings of 13th international conference on advanced communication technology (ICACT), vol. 12, no. 3; 2011. p. 405–10.
Singh R, Sharma U, Cecchet E, Shenoy P. Autonomic mix-aware provisioning for non-stationary data center workloads. In: Proceedings of the 7th IEEE international conference on autonomic computing and communication, vol. 8, no. 4; 2010. p. 24–31.
Sotomayor B, Rubn S Montero, Ignacio M Llorente, Ian Foster. An open source solution for virtual infrastructure management in private and hybrid clouds. In: Proceedings of the IEEE international conference on internet computing, vol. 10, no. 6; 2009. p. 78–89.
Sotomayor, Llorente A, Foster I. Virtual infrastructure management in private and hybrid clouds. In: Proceedings of the IEEE international workshop on internet computing, vol. 13, no. 5; 2010. p. 14–22.
Soundararajan G, Daniel Lupei, Saeed Ghanbar. Dynamic resource allocation for database servers running on virtual storage. In: Proceedings of the 7th USENIX conference on file and storage technologies, vol. 6, no. 2; 2011. p. 71–84.
Sriram A. Simulation tool exploring cloud-scale data centres. In: Proceedings of the 1st international conference on cloud computing, vol. 9, no. 4; 2010. p. 381–92.
Tai J, Juemin Zhang, Jun Li, Waleed Meleis, Ningfang Mi. Adaptive resource allocation for cloud computing environments under bursty workloads. In: Proceedings of the international conference on communications (ICC'11), vol. 2, no. 2; 2011. p. 55–68.
Teng F, Magoules F. A new game theoretical resource allocation algorithm for cloud computing. In: Proceedings of the 1st international conference on advances in grid and pervasive computing, vol. 6, no. 4; 2010. p. 321–30.
Tomita T, Shin-ichi Kuribayashi. Congestion control method with fair resource allocation for cloud computing environments. In: Proceedings of the IEEE pacific rim conference on communications, computers and signal processing, vol. 10, no. 3; 2011. p. 67–75.
Tyagi Neha, Pathak Rajesh. Negotiation for resource allocation for multiple processes infrastructure as a service cloud. J Adv Eng Res 2011;10(2):65–71.
Upton D. Enabling efficient online proling of homogeneous and heterogeneous multicore systems. In: Proceedings of the symposium on microarchitecture, vol. 3, no. 4; 2010. p. 22–31.
Urgaonkar B, Shenoy P, Chandra A, Goyal P, Wood T. Agile dynamic provisioning of multi-tier Internet applications. ACM Trans Auton Adaptive Syst 2010a;5(5):139–48.
Urgaonkar B, Pacifici G, Shenoy P, Spreitzer M, Tantawi, A. An analytical model for multi-tier internet services and its applications. In: Proceedings of the ACM SIGMETRICS international conference on measurement and modeling of computer systems, vol. 33, no. 5; 2010b. p. 291–302.
Vada, Eirik T. Creating flexible heterogeneous cloud environments. Text Book Heterogeneous Cloud Environ 2011;1(1):23–45.
Vaquero LM, Merino LR, Caceres J, Lindner M. A break in the clouds: towards a cloud definition. ACM SIGCOMM Comput Commun Rev 2009;39(1):67–72.
Vaquero LM, Luis Rodero-Merino, Rajkumar Buyya. Dynamically scaling applications in the cloud. In: Proceedings of the ACM SIGCOMM computer communication review, vol. 41, no. 1; 2011. p. 45–52.
Venugopal Srikumar, Chu Xingchen, Buyya Rajkumar. A negotiation mechanism for advance resource reservation using the alternate offers protocol. J Future Gener Comput Syst 2009;25(6):599–616.
Vianna E. Modeling performance of the hadoop online prototype. Int Symp Comput Archit 2012;53(1):72–7.
Vijayakumar S, Qian Zhu, Agrawal G. Dynamic resource provisioning for data streaming applications in a cloud environment. In: Proceedings of the 2nd IEEE international conference on cloud computing technology and science, vol. 5, no. 6; 2010a. p. 1023–39.
Vijayakumar S, Qian Zhu, Agrawal G. Automated and dynamic application accuracy management and resource provisioning in a cloud environment. In: Proceedings of the 11th IEEE/ACM international conference on grid computing, vol. 5, no. 6; 2010b. p. 33–40.
Villegas D, Sadjadi SM. Mapping non-functional requirements to cloud applications. In: Proceedings of the 2011 IEEE conference on SEKE, vol. 8, no. 4; 2011. p. 527–32.
Vouk MA. Issues, research and implementations in information technology interfaces. In: Proceedings of 30th international conference on ITI, vol. 4, no. 5; 2010. p. 31–40.
Warneke D, Odej Kao. Exploiting dynamic resource allocation for efficient parallel data processing in the cloud. In: IEEE transactions on parallel and distributed systems, vol. 6, no. 4; 2011. p. 34–48.
Wuhib F, Rolf Stadler, Mike Spreitzer. Gossip-based resource management for cloud environments. In: Proceedings of 6th international conference on network and service management, vol. 5, no. 4; 2010. p. 65–71.
Xabriel J, Collazo-Mojica Jorge Ejarque, Masoud Sadjadi S, Rosa M Badia. Cloud application resource mapping and scaling based on monitoring of QoS constraints. In: Proceedings of the 2012 international conference on software engineering and knowledge engineering, vol. 7, no. 4; 2012. p. 88–93.
Xie J. Improving map reduce performance through data placement in heterogenous Hadoop clusters. IEEE Int Symp Parallel Distributed Process 2010;66(10):1322–37.
Yazir YO, Matthews C, Farahbod R, Neville S, Guitouni A, Ganti S, Coady Y. Dynamic resource allocation based on distributed multiple criteria decisions in computing cloud. In: Proceedings of the 3rd international conference on cloud computing, vol. 28, no. 1; 2010. p. 91–8.
Yoshino T, Osana Y, Kuribayashi S. Evaluation of congestion control methods for joint multiple resource allocation. In: Proceeding of the 13th international conference on network-based information systems, vol. 10, no. 3; 2010. p. 16–8.
Younge AJ, von Laszewski G, Wang L, Lopez-Alarcon S, Carithers W. Efficient resource management for cloud computing environments. In: Proceedings of the international conference on green computing, vol. 50, no. 12; 2011. p. 60–71.
Zhang Q, Quanyan Zhu, Raouf Boutaba. Dynamic resource allocation for spot markets in cloud computing environment. In: Proceedings of the 4th IEEE international conference on utility and cloud computing, vol. 10, no. 6; 2011. p. 177–85.
Zhang S, Zhuzhong Qian, Jie Wu. An opportunistic resource sharing and topology-aware mapping framework for virtual networks. In: Proceedings of the IEEE INFOCOM 2012, vol. 13, no. 4; 2012. p. 2408–16.
Zhu Q, Gagan Agrawal. Resource provisioning with budget constraints for adaptive applications in cloud environments. In: Proceedings of the HPDC 2010, vol. 8, no. 3; 2010. p. 304–07.