International Journal of Computers and Applications

ISSN: 1206-212X (Print) 1925-7074 (Online) Journal homepage: http://www.tandfonline.com/loi/tjca20

Efficient resource management techniques in cloud computing environment: a review and discussion

Frederic Nzanywayingoma & Yang Yang

To cite this article: Frederic Nzanywayingoma & Yang Yang (2018): Efficient resource management techniques in cloud computing environment: a review and discussion, International Journal of Computers and Applications, DOI: 10.1080/1206212X.2017.1416558

To link to this article: https://doi.org/10.1080/1206212X.2017.1416558

INTERNATIONAL JOURNAL OF COMPUTERS AND APPLICATIONS, 2018
https://doi.org/10.1080/1206212X.2017.1416558

Taylor & Francis
Taylor & Francis Group

Efficient resource management techniques in cloud computing environment: a review and discussion

Frederic Nzanywayingoma and Yang Yang

School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

ABSTRACT

The increasing number of cloud computing infrastructure and the users' demands for services has made the cloud resource management an impossible task to be manually performed by human operators. In this paper, we surveyed the state of the art of cloud resource management for infrastructure as a service. We provided an overview of the recent research findings and technologies while focusing on the resource management techniques such as resource provisioning, resource discovery, resource monitoring, resource mapping, resource allocation, resource consolidation, resource modeling, resource scheduling. This survey triggered innovative methods to handle the existing problems of resource management in cloud computing and we hope that it can be used as a source of interested readers to understand the existing methodologies in this research area for future enhancements.

ARTICLE HISTORY
Received 5 June 2017
Accepted 10 December 2017

KEYWORDS
Cloud computing; Dynamic resource management; Resource provisioning; Resource monitoring; Resource allocation; Resource discovery

1. Introduction

The National Institute of Standards and Technology (NIST) defines cloud computing as a model for providing services to computing resources such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [1]. The IaaS (Infrastructure as a Service) is a form of cloud computing that provides virtualized computing resources such as virtual machines, virtual storage, virtual infrastructure, other hardware assets as resources that can be easily scaled up and down with demand and only pay per use. The (PaaS) Platform as a Service, a cloud providers delivers a hardware and software tools such as physical and virtual machines, operating systems, applications, services, development framework, transactions, control structures. The (SaaS) Software as a Services also referred to as on-demand software, the cloud provider delivers a complete operating environment with applications, management, and the user interface (i.e. GoogleApps, SalesForce.com, Windows Azure Platform, Office 365). Cloud deployment models are private cloud where the infrastructure are deployed by an organization; community cloud where the cloud infrastructure is shared through multiple entities; public cloud where the infrastructure are possessed by organization which sells cloud services; and hybrid cloud which is a mixture of private and public cloud. The cloud computing features are: on-demand self-service, broadband network access, rapid elasticity where the abilities are provided fast and released rapidly by cloud users. In this work, we introduce some resource management techniques. Figure 1 shows the framework of this study.

Firstly, we outline different metrics or parameters studied so far by some researchers such as throughput, migration time, associated overhead, scalability, response time, resource utilization, performance.

The rest of the paper is organized as follows. In Section 2, we introduced the resource management in cloud computing. In Section 3, we presented the resource management techniques. In Section 4, we discussed the benefits to migration techniques. In Section 5, we discussed on cloud resource management challenges, strategies, and opportunities. In Section 6, we discussed on solutions to open problems. Section 7, showed the related work and literature surveys. Finally, we concluded in Section 8.

2. Introduction to resource management in cloud computing

The cloud infrastructure resources often sit in geographically dispersed data centers and the customers obtain services through network. A data center represents a physical area in which racks containing IT equipment such as disk

CONTACT Frederic Nzanywayingoma nzanywafre@yahoo.fr

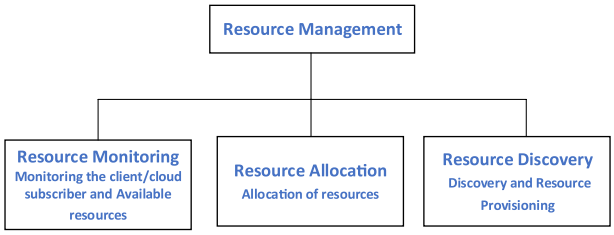

Figure 1. Elements of resource management.

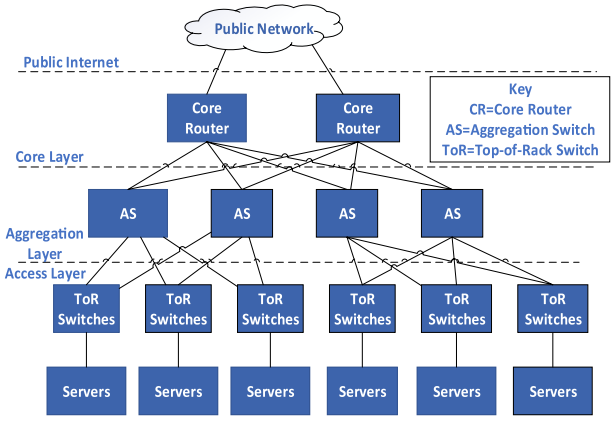

Figure 2. Representation of data center network topology.

enclosures, controller enclosure, servers, routers, switches, load balancers, firewalls, power delivery devices, cooling machines, unmanaged devices, other security devices. The IT equipment are enclosed in racks and they could be monitored and managed by monitoring system. The system monitors the equipment usage status and it can generate the alarms when it detects an abnormal usage which is greater than or equal to the specified alarm threshold. The alarms are manually or automatically cleared when the equipment return in normal status. The figure below shows the network topology of a data center [2] (Figure 2).

The rack contains a certain number of servers stacked one above the other. The rack server brings the advantage for easy cabling and saving floor space in data center. ToR (Top of Rack) is linked to AS (Aggregation Switch) for redundancy and provides connectivity to the servers mounted in a one rack.

The aggregation switch in the aggregation layer forwards traffic from multiple ToR switches to the core layer. A core router in core layer is a router designed to operate in the Internet backbone and must support multiple communication interfaces of the fastest speed and must be able to forward IP packets at full speed on all of them. It monitors communication channels between internal networks and external networks to prevent risks on external networks from affecting internal networks. Load balancers receive data traffic and increase effective network bandwidth by distributing network connection requests to multiple servers using a distribution algorithm. A server is a computer program running to serve the requests of other programs, the "clients" such as sharing data, information or hardware and software resources (i.e. file server, mail server, print server, web server, application server, so on.)

2.1.1. Physical and virtual resource management

The physical resources (PR) consist of the processors, memory, disk drive, Network Interface Controller (NIC), peripheral devices (floppy, keyboard), network resources (networking products), storage media, and other physical components. The physical resources are dispatched across multiple compute requests through virtualization and provisioning while the virtual resources (VR) are dynamically assigned and reassigned according to consumer demand. The physical resources and virtual resources are managed with resource management system (RMS). When a user orders resources; the RMS checks actual resource status to decide if there are sufficient resources to satisfy user's request. The RMS is responsible to monitor continuously the user's allocated resources while optimizing the system [3]. Virtual and Physical servers in data centers host many applications types such as webservers, databases, customer business apps. The physical servers host the virtual servers called virtual machines and those VMs run different OS/or applications [4]. Computing resources (such as CPUs and memory), GPUs (graphics processing units), CD/DVD-ROM drives or ISO images, USB devices, and network connections are provided by hosts; virtual disk storage space is provided physical storage devices. A resource pool provides computing resources on hosts for VMs. You can manage hosts, divide clusters, and configure scheduling policies for clusters on the Computing Pool page.

The researches have been done to develop infrastructure in cloud data centers. Server consolidation is one of the approaches used to consolidate the physical servers and save space (server location) in data center. To consolidate the VMs on one physical server reduces the capital, operating expenses, and cuts all unnecessary costs to maximize return on investment (ROI) in data center while optimizing the efficiency use of resources. A server consolidation technique is divided into three groups: physical consolidation, logical consolidation, and workloads consolidation. The physical consolidation means collocating servers at fewer locations. The logical consolidation means no physical relocation of servers done and the implementation procedures are done across the server application. Logically, the servers are consolidated through a set of management tools and processes. Workload consolidation means implementing multiple applications on fewer, more powerful platforms through workload management and partitioning also called rationalized consolidation.

Author [5] has categorized the type of workload in cloud computing. The web server and SaaS workload based such as communication, file storage, processing, online shops, interactive, DBMS are among the most workload used. Scientific applications based such as Big Data, Workflows, learning algorithms should be considered. Benchmark applications such as micro-benchmarks, system benchmarks, application benchmarks were mentioned. Some of the advantages of workload consolidation are simplified management, lower costs such as staff costs, hardware costs, energy conservation, software, facilities costs and also improving the data center services.

2.1.2. Logical resources management

Logical resources are system abstractions which have temporary control over physical resources (i.e. operating system, bandwidth, network throughput, energy load balancing mechanisms).

2.2. Requirements of resource management

According to the NIST [6], the definition of cloud computing, cloud services exhibit five essential resource management characteristics and their requirements: The table below summarized those requirements (Table 1).

3. Resource management techniques

The objective of resource management is to reduce the service costs and time related to that service. Here, we will consider different resource management techniques. Those techniques based on the priorities and on parameters like service cost, time needed to access resources, task type, number of processors needed to run tasks [7]. Other techniques such as bin-packing algorithm and gradient search techniques are also mentioned. In the bin-packing technique [8,9], the heuristic-based VM migration scenario is partitioned as follows: (1) Divides machines into two pools – core nodes and accelerator nodes. (2) Adjusts the size of each pool to reduce cost or to increase resource utilization. This technique manages the resource nodes allocated and decides when to add/remove them from the resource pool. It also monitors the storage system to estimate the incoming data capacity. The author in [10] focused on maximizing the efficiency of the scheduling algorithm. Here, the Round Robin scheduling technique is used. It utilizes the turnaround time utility efficiently by differentiating it into a gain function and a loss function for a single task and also used to increase the efficiency gain. An overall improvement in the resource utilization and reduction in the processing cost is also shown.

The research in [11] considered a stochastic model based on load balancing and scheduling in cloud computing clusters, where tasks arrive according to a stochastic process and request resources like memory, CPU, and storage space. It takes the performance of Join-the-shortest-queue (JSQ) routing and two-choice routing algorithms with Max-Weight scheduling policy. Authors in [14] based on game theory models, proposed a market-based resource management policy that let participants to trade their services by means of a double-sided combinational auction.

A theory based on Stackelberg game was proposed and a Nash equilibrium solution was found. In [12], Zuling et al. proposed a new cloud resource management algorithm called CRAA/FA (Cloud Resource Allocating Algorithm via Fitness-enabled Auction), that creates a market for cloud resources and makes the resource agents and service agents bargain in that market. Authors in [13] suggested cloud management framework which allocates infrastructure resources to spot markets to best match customer demand in terms of supply and price in order to maximize the providers revenue and customer satisfactions.

A Minimum Cost Maximum Flow (MCMF) algorithm was proposed in [14] for optimal dynamic placement of virtual resources in cloud infrastructure to serve multiple users. In [15], authors formulate the resource management problem in a VM-multiplexing resource management scheme to manage decentralized resources to achieve increased resource utilization using the proportional share model (PSM), and also delivers adaptively optimal execution efficiency. The work suggested a novel scheme named DOPS (Dynamic Optimal Proportional Share) for virtual resource allocation on a Self-organizing cloud (SOC), and with three key contributions: Optimization of task's resource allocation under user's budget, increased resource utilization based on PSM and Lightweight resource query protocol with low contention.

An allocation of consumer resource to a proper data center using adaptive resource allocation model was proposed in [16]. It is based on geographical location of consumer and the workload of data center.

In [17], Feng, et al. presented a joint allocation and scheduling of network resource for Multiple Control Applications in SDN. The authors have evaluated the learning algorithms using the price paid by the user for both bandwidth and flow table capacities in order to guarantee fair allocation of the network resources. The bandwidth allocation was achieved by maximizing the sum of each control application in a logarithmic rate, where the control application rate is proportional to the bandwidth price for that particular application while flow table allocation favors flows with higher hit-rate to optimize OpenFlow

switches' throughput. This technique was tested against data-sets provided by Chinese organizations and compared with a random scheduling model.

Hierarchical network-aware placement technique of service-oriented applications in clouds is a technique which based on Integer Linear Programming (ILP) for the Cloud Application Placement Problem (CAPP) [18]. CAPP is used to determine how applications and services are allocated within the cloud. For instance, it determines in which machine a service should be allocated in order to satisfy multiple constraints such as CPU, memory, bandwidth, management policies. A comparison between the optimal ILP algorithm with two hierarchical algorithms based on Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) that find near-optimal solutions for the CAPP problem has been conducted. In [19], another resource management technique were proposed. The proposed technique uses the Self-Organizing Clouds (SOC) to achieve the maximum of the resource utilization and it also delivers optimal execution time. SOC connects a large number of desktop computers on the internet by P2P network. Each participating computer acts as a resource provider and resource consumer. SOC having two main issues: (a) Locating a qualified node to satisfy a user task's resource demand with bounded delay; (b) Optimizes a task's execution time by determining the optimal shares of the multi-attribute resources to allocate to the tasks with various QoS constraints, such as the expected execution time. In this work, the Dynamic optimal proportional (DOPS) share and multi range query protocols were proposed. DOPS is an algorithm used to redistribute available resources among running tasks dynamically, such that these tasks could use up the maximum capacity of each resource in a node while each task's execution time can be further minimized.

3.1. Virtualization techniques [20]

Virtualization technique has become popular in IT industry as a software-based solution to build shared hardware infrastructures [21]. Virtualization technique is a framework to divide the resources of a computer into execution environments to make the physical machine to be productive for the resource utilization and resource management. The main importance of virtualization technique is that it supports migration of VMs from existing host machine to other physical machines (in hosts migration or between hosts migration) [22]. Virtualization technology brought technical basis in cloud computing. There exists different kind of virtualization technology used in cloud computing such as storage virtualization, server virtualization, network virtualization, client virtualization, operation framework virtualization [23]. The network virtualization is based on virtualized network switch technology commonly known as vSwitch. Client Virtualization is a client virtualization technology based on creating a client desktop as a VM called VDI (virtual desktop infrastructure). A VDI uses rack-based servers distributed across the data center (DC) with a top of rack (ToR) switches at the network edge.

A virtual processor also called virtual CPU (vCPU) is a technique to share physical CPU with different VMs. By default, each VM is allocated its own vCPU. Storage Area Network (SAN) is a storage architecture that connects the storage system to the application server over the network. SAN storage devices provide data storage space for VMs in the system using shared software to provide block-level data sharing service.

Cloud storage is a solution which permits cloud users to save data or use storage resources on a unified cloud platform. Such a cloud storage system consolidates all types of storage devices on the network into a unified storage platform using various functions, such as cluster applications and distributed file system.

3.2. Virtual machine migration techniques

VM migration technique is the process of moving an operating system instance to another physical node without interrupting an operation. The virtualization separates hardware from software and has benefits of server consolidation and live migration. VM migration improves the system reliability and availability. Migration facilitates load balancing, fault-tolerant management, low level system maintenance, and energy efficiency in data center [24]. For VM migration, the overloaded VMs can migrate from light loaded to over loaded machines. To improve the reliability in data center, the migration is performed as a service restoration in response to hardware faults that prevent the system within the VMs to comply with the functional and/ or real-time specifications. In case of failure physical machines, if the hardware failure still allows for saving and transfer of the state of the VMs, its operation can be continued on another physical machine.

3.2.1. Live VM migration

The live migration is a technology which help to balance load and optimize VM deployment over physical nodes in cloud computing. It is used to increase the resource availability and decrease energy consumption in cloud computing.VM migration helps to migrate the data from the failed physical nodes to the healthy nodes. There are many techniques which attempt to reduce the downtime and to provide better performance in low bandwidth environment. We outline two techniques: pre-copy technique and post-copy technique. The pre-copy technique allows

the sources servers to transfer the memory to the target VMs over a series of iterations.

The memory and vCPUs are reserved on the destination host. For the post-copy technique, the information of device state on the target machines are transferred at first and also uses the following metrics to measure the performance: preparation, service downtime, resume, total migration time, CPU, and network consumption during the migration. The benefits of the live migration are: energy efficiency, load balancing, and convenient maintenance, etc.

3.2.2. The offline migration

During the offline migration, the service is stopped before the migration and resumed after the migration while for the live migration the process can execute at the user unaware state. The disadvantage of offline migration is the larger downtime.

3.2.3. Cold migration

Cold Migration also called Regular Migration is the migration of a power-off virtual machine. With cold migration, you have the option of moving the associated disks from one datastore to another [25,26].

3.2.4. Hybrid virtual machine migration

Hybrid migration is a combination of both pre-copy and post-copy virtual machine migration. It presents five phases: preparation phase; bounded pre-copy rounds phase; virtual machine state transfer phase; virtual machine resume phase; and on-demand paging phase.

3.3. Storage migration techniques

The storage migration can be performed while the VM still running. Whether a VM is reading or writing from virtual hard disk file for Microsoft Windows servers or for Oracle VMs or VMS. There exist many VM storage data migration solutions. In this paper, we only cite the following storage data migration solutions:

3.3.1. LVM mirroring-based data migration solution

Logical Volume Manager (LVM) migration solution is based on Linux LVM mirroring function. The disk space used by applications comes from logical volumes (LVs) created based on physical volumes (PVs) of the source storage. The storage model must be compatible with the Linux system. The host must be directly connected to a storage device (DAS networking mode) or it must be connected to the storage device using switch (SAN networking mode) and the number of idle ports on the switch must be greater than or equal to 2xn where n represents the number of controllers on the target storage. It supports both online and offline data migrations. Before performing LVM mirroring migration, it is recommended to collect the live network information and assess the risks and check the compatibility of the source and target storage systems.

3.3.2. LDM mirroring-based data migration solution

The LDM (Logical Disk Manager) is a migration technique to migrate data between different storage systems using the logical disk manager function on the Windows servers. After mirrors are created, mirror copies relevant to the source storage are removed from the target storage.

3.3.3. SmartVirtualization and SmartMigration-based data migration solution

These migration solutions describe how to migrate data from the source storage to target storage using SmartVirtualization and SmartMigration features. SmartVirtualization supports both DAS and SAN networking mode and is used while migrating data on peer vendors' products. It provides the benefits such as Online LUN migration and ensuring service continuity. Data on the source LUN is completely copied to the target LUN, ensuring data consistency. Services can be migrated between third-party storage systems. SmartMigration technique replicates all data from the source LUN to the target LUN and lets the target LUN take over services from the source LUN to complete the service migration process. SmartMigration enables you to migrate data within a storage system or between heterogeneous storage systems.

3.3.4. LUN copy based data migration solution

The LUN (Logical Unit Number) copy-based migration is the migration technique based on LUN copy feature. LUN copy feature are used to copy data from remote LUNs on storage devices to the target storage using fiber channel links. The LUN Copy migration technique provides quick data distribution, and centralized, and generates multiple data duplicates to ensure data security.

3.3.5. ASM-based data migration solution

The basic principles of ASM data migration solution are as follows: the extents of files are moved among disks using Oracle ASM Rebalance to realize I/O balance on the ASM disk group. LUNs of the target storage are added to the ASM disk group of the Oracle database, and LUNs of the source storage are deleted from the ASM disk group. To migrate data using Oracle ASM Rebalance automatically rebalances data and migrate data from the LUNs of the source storage to those of the target storage. During the rebalance process, I/O performance (mainly throughput and response time) may be affected, depending on the capacity of the storage and the degree of rebalance. This

migration solution supports DAS, and SAN networking model and support online and offline data migration. Oracle ASM Rebalance can save space during data migration. The number of LUNs on the target storage can be smaller than that of LUNs on the source storage, but the total capacity of LUNs on the target storage must be greater than that of LUNs on the source storage.

3.3.6. Oracle RMAN-based data migration solution

The Oracle RMAN migration technique is used to migrate data files from different storage systems using Oracle RMAN for the Oracle DB and SUSE platform. The oracle data files are backed up to the ASM disk groups. Oracle RMAN migrates data files from the source storage to the target storage and restores the database. In this way, data of the source LUNs are migrated to the target storage.

3.3.7. SVM Mirroring-based data migration solution

The Solaris Volume Manager(SVM) mirroring is the online migration technique used to migrate data between different storage systems using the SVM mirroring function.

3.3.8. VIS-based data migration solution

VIS-based migration technique is used to migrate data from the source storage to the target storage. This technique consists of taking over the source storage and migrate data from source storage to the target storage.

3.3.9. VxVM Mirroring-based data migration solution

The VxVM (Veritas Volume Manager) Mirroring is a migration technique to migrate data between storage systems to target systems. After the mirroring is completed, the mirror copy of the source storage should be deleted.

3.3.10. FastCopy-based data migration solution

The FastCopy-based data migration is the migration technique used to migrate data between windows server OS. The source and target file systems are mounted on the migration server. The FastCopy is used to synchronize files between the source and target systems, achieving data migration between the source and target storage.

There are other functions to be considered in storage system such as SmartThin, smartTier, SmartQoS, SmartMotion, SmartPartition. The SmartThin function automatically expands capacity, and help on improving disk utilization. The SmartTier function intelligently migrates data among different storage tiers based on the data access frequencies. It dynamically matches hotspot data and storage medium, improving system performance and reducing the TCO (total cost of ownership).

SmartQoS function intelligently schedules storage resources based on the priority of services, optimizing system resource allocation. It dynamically allocates resources of storage systems, meeting specific performance requirements on the IOPS, bandwidth, or latency. SmartMotion function dynamically relocates data based on the service changes to balance loads of storage systems. SmartPartition function sets cache partition requirements for critical services, dynamically allocates cache resources to services based on the requirements, and isolates the cache resources between services, preventing unnecessary cache competition and ensuring performance of critical services. Resource Performance Tuning: View and manage resource optimization features including SmartTier, SmartQoS, SmartPartition, SmartMigration, and SmartCache.

4. Benefits of virtual machine migration techniques

4.1. Server migration optimizations

The optimization for server migration helps on improving the live migration and optimizing the performance metrics such as total migration time, and downtime while providing uninterruptible services to applications running in VMs, some techniques have been proposed: (1) Memory Page Compression: Live migration performance is improved by minimizing the amount of data transferred to the destination using the technique called memory page compression which compress the source PM memory pages and decompress the memory pages in target PM. (2) Delta page transfer: This technique reduces the network bandwidth consumption by maintaining a cache of already transferred memory pages. It improves the live migration process and reduces service interruption risks. (3) Data de-duplication: the data de-duplication is a compression technique which is specialized to find the duplicates data inside the memory and disk of a single VM and removes them during the live migration process. It improves the storage data utilization and in network data transfer. (4) Post-copy technique: The post-copy migration technique means the transfer of a VM's memory contents until after its processor state has been sent to the target host and resumed there [27]. The performance of this technique depends on the way and which the VM's memory contents are fetched from source machines during the live migration. (5) Hybrid pre and post copy: This technique performs single round of pre-copying which precedes the virtual CPU state transfer. This technique improves live migration process. (6) Server consolidation: To reduce server sprawl in data centers, server consolidation algorithms should be the requirements. Among them are: stochastic bin packing [28,29], multi-capacity bin packing [30], VM packing heuristics, and so on. Consolidating servers reduces the power consumption and the costs of data center administrators. The figure below illustrates the

server consolidation modeling based on memory buddies technique (Figure 3).

Memory Buddies Algorithm [31] is also a type of server consolidation technique which determines which candidate servers to be shutting down and attempts to migrate virtual machines to hosts with high sharing opportunities. The Memory Buddies' algorithm comprises three steps: The first step is to identify server to consolidate by examining memory utilization statistics of every hosts. The second step is to identify the target hosts. After determining the server candidates, the algorithm has to determine the new physical server to house the VM but one of the existing challenges reside on how to find the proper PM in order to minimize the cost function. This can happen especially when the VM has a large CPU size, network or especially, when the existing servers are heavily utilized. The third step concerned with the migration of the VMs to their target hosts. Once the new destinations have been discovered, the algorithm can perform VM migration process with minimal service downtime and with the minimum resource consumption.

The live migration ensures system transparency, and near zero downtimes of the applications running inside the migrated VMs [31]. The migration are done concurrently, and once the migration is finished, the original servers are then shut down and moved to the shutdown pool so that they can be reinitialized later if memory requirements increase.

4.2. Resource location optimization

The cloud infrastructure resources often sit in data centers dispersed in different location and the customers obtain services through network, therefore, it is an important decision to select an appropriate cloud data center to host an application given a fixed budget and QoS constraints. Furthermore, in the dynamic and open cloud environment, the user groups are dynamic, and hence, the resource location optimization is also a dynamic process. The VM live migration techniques are important mechanisms used to strategically place application VMs across disparate geographic locations to optimize user-perceived latency and cost in real time, in which the selection of resource location is a multi-objective optimization problem [32].

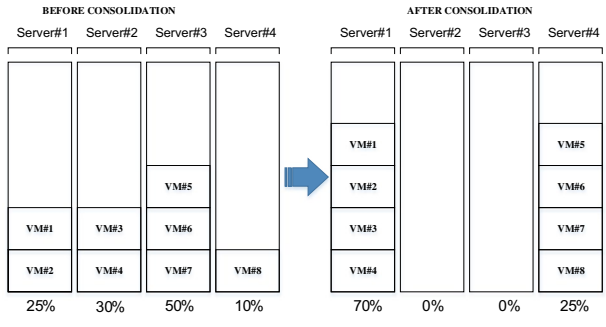

Figure 3. Server consolidation modeling.

4.3. Load balancing optimization

The increased number of acting VMs results in maximizing the number of server migration and the maximizing of the energy efficiency. The more VMs can be placed on the same host, the greater the energy efficiency can be obtained. However, a large number of VMs at the same host will increase the VMs migration and degrade the system performance. Therefore, the hotspot mitigation algorithm [33] came up to help to determine which heavy-loaded VMs to migrate first.

Load balancing algorithm such as round-robin, minimal queue depth, or minimal tasks are invoked. The advantage of load balancing algorithm is to enhance the performance in data migration, load balancing to minimize the makespan time, and the resource utilization rate [34].

5. Resource management issues

There exist several problems to be considered while allocating virtual resources, such as type of the resource provisioning (logical, physical resources), resource discovery, resource scheduling, resource brokering, resource modeling.

In this section we outlined the resource types and problems, in resource management; we investigated data center resource management problems, and opportunities, and strategies. We considered the problems and the impact on the cloud data center's performance. In data center there are many problems such as the ones based on physical resources, and logical resources. The ones based on physical resources are CPU, memory, storage, workstations, processors, networking elements, sensors, actuators, and so on while the ones based on the logical resources are operating systems, energy, network throughput, bandwidth, protocols APIs, network loads, delays, and so on [35].

5.1. Challenges, strategies, and opportunities of resource management in cloud computing

5.1.1. Resource management challenges in cloud computing

Efficient challenges for resource management should be based on the following metrics [36–38]: Energy-efficient, Bandwidth cost minimization, Performance optimization, etc. According to the work in [39], stated that the

utilization rate of server resources in the data centers is 20% and the rest 80% represent the idle servers and those idle servers need 60% of the total consumed power. The rest of power coupled with cooling equipment, security system, and so on. Consider the CPU as the main components to consume energy. Turning unused hosts off improves the energy efficiency.

According to [39], Dynamic Component Deactivation (DCD), Dynamic Performance Scaling (DPS), and Dynamic Voltage and Frequency Scaling (DVFS) are ones of the dynamic power management techniques suggested. DPS technique is for the automatic adjustment of the performance proportional to the power consumption. DVFS and DCD are applied to different computer component and at OS level such as CPU, Memory, disk, network interface, other power-aware OS such as KVM, VMware solution, Xen Hypervisor. The application of DVFS technique decreases power consumption of a computing resource significantly. This technique was firstly used in portable and laptop systems to conserve battery power, and now it has been implemented on the server chipsets. Lowering the CPU frequency may lead to power savings and potential energy savings but may also impact application performance. Therefore, to maximize the energy efficiency while meeting the SLA constraints, scheduling algorithms are involved to determine a good operating frequency of the CPU to meet the application deadlines. The scheduling algorithms have to consider both cost and energy factors on the decision-making [40].

The VM resource scheduling algorithm based on auction mechanism was proposed in [41,42], which consider the multiple factors such as network bandwidth cost minimization, auction deadline, profits of service providers and lowering VMs resource utilization rate. Kim et al. [43] studied different adaptive DVFS schemes to maximize energy savings and user profits.

Their proposed algorithm selects the least expensive VM and placement that meets the required application throughput (MIPS rate) to minimize user's cost. The work in [44] proposed an energy-aware algorithm that supports DVFS. Their algorithm schedules deadline-constrained virtual machines to the cores of the PMs (physical machines) to provide more computation resource within a certain power budget. The physical machines are prioritized depending to their performance-power ratio.

Thermal Management: At any time, a CPU chip consumes power depending on its workloads. The large amount of energy is consumed by cooling systems. The thermal hotspots may be occurred when workload concentration is on its peak and the energy consumption by cooling systems is increased. The challenge is to distribute the workload between the available hosts while avoiding the thermal hotspots.

The hotspot is mitigated with virtual machine migration techniques from overloaded to less loaded hosts.

Chen et al. [45] suggested an approach where the temperature is considered to determine VM placement to avoid hotspot and achieve thermal balance.

Paper [46] studied a Multi-agent-Based resource management for Energy Minimization. The authors addressed the allocation of VMs to Physical Machines (PMs) in order to minimize the energy consumption. Traditional energy-aware VMs resource allocations either allocate VMs to Physical Machines (PMs) in a centralized manner or implement VMs migrations for energy reduction without considering the migration cost. Authors proposed Multi Agent which dispatches a cooperative agent to each Physical Machines (PMs) to ensure that they can manage the VMs resources. Considering the power-aware scheduling techniques [47], the same variables such as resource management [39,48], live migration [20,49], and a minimal virtual machine design would be greatly with minimal performance overhead. This allows for the creation of an efficient scheduling system which decreases power consumption of a system while attempting to maximize performance [41].

Edwardo and Fabio [50] have presented a DPA (dynamic programming approach) based on mathematical formulation to optimize maintenance budget and regarding the system reliability constraints. The DPA approach was combined with successive approximations, branch and bound, and binary programming to get a good reliability index. Other methods have been suggested along years to compare the cost and the system reliability. The dynamic programming approach (DPA) was first developed to solve optimization problem. The author adopted this approach to compare DPA with hybrid Genetic Algorithm (HGA) developed previously.

5.1.2. Resource management strategies in cloud computing

Resource management strategies are: Resource monitoring, Resource allocation, Resource consolidation (such as Virtualization Consolidation, storage consolidation, Data center Consolidation), Resource scheduling, QoS (Quality of Service) control, and overload avoidance for live migration of less utilized resources.

Authors in [51] reviewed various strategies developed so far for resource management based on linear scheduling strategies named TARA (Topology-Aware Resource Allocation) and DRA (Dynamic Resource Allocation) for Parallel Data Processing. The author also enumerated different advantages such as: (1) Topology-Aware Resource Allocation. (2) Linear Scheduling strategy for resource management designed to increase the profit of cloud resource utilization, and minimize the response time. The

linear scheduling considers performing task and resource simultaneously. This saves the waiting time occurred when task and resources scheduled separately. Network-Aware Resource Management technique was proposed to solve [52] efficient network resource sharing issue, and network congestion issues in multi-tenant data centers. Their study is based on SDN (Software-Defined Network) resource management strategy to minimize the overall cost while managing network resources.

This is the multi-objective optimization problem due to that the management strategy of the network resource should decrease the cost for tenant, maintain the performance of tenants' applications, and increase the profit of the cloud service provider. Paper [53] suggested a "skewness algorithm" to manage and reclaim the dynamic resources using virtualization concept. According to paper [54], T.R Gopalkrishnan Nair, et al., presented a model named RBRAM (Ruled Based Resource Allocation Management) to solve two critical resource management issues: resource arbitration and resource allocation. Authors discussed three main points: resource allocation strategies (RAS), complexity of the system allocation, and transition panorama of resource allocation. Authors revealed that resource allocation rate should be greater than resource request rate. In [55], Chenn-Jung Huang, et al., outlined a resource management strategy based on a combination of GA (Genetic Algorithm) and (SVR) Support Vector Regression.

Authors designed application service prediction module with Support Vector Regression (SVR) to maximize the number of resource utilization according to the Service Level Agreement (SLA) of each process. Authors mainly focused on application level resource allocation instead of focusing on the mapping of physical resources to virtual resources. Based on parameters like completion time and bandwidth, cloud consumers' tasks had been classified. According to the characteristics and preferences of tasks, resources were assigned to the cloud consumers.

In [56], authors proposed a threshold based dynamic resource allocation scheme for cloud computing.

Authors mainly focused on application level resource allocation instead of mapping between physical resources and virtual resources for better utilization of resources. A threshold is used to optimize the decision of resource reallocation. The proposed algorithm consists of two procedures: Data center resides at the data centers central computer and Broker-runs on user's machine with the application. Both procedures interact with each other for dynamic resource management.

In [57], authors proposed job scheduling algorithm based on Berger Model with dual fairness constraints. Authors had mainly concentrated on fairness of resource management and cloud consumers' satisfaction to the provided services. Based on parameters like completion time and bandwidth, cloud consumers' tasks had been classified. According to the characteristics and preferences of tasks, resources were assigned to the cloud consumers. Authors implemented their algorithm on CloudSim toolkit and compared with optimal completion time algorithm. Results show that algorithm based on Berger Model is better.

5.1.3. Virtual resource management in cloud computing: opportunities [58]

Clouds are designed to deliver as much as computing capacity as any users want. Cloud provider rent or lease resources to users in self-service manner with additional services such as resource scalability which provide greater flexibility for users since user is only paying for what needed. Therefore, cloud computing has many benefits such as Scalability, QoS (Quality of Service), Cost effectiveness, simplified interfaces.

5.2. Resource provisioning

The increase on the resource demand leads to overprovisioning or underprovisioning of resources. In [59], authors proposed a Fault-Tolerant Scheduling for Real-Time Scientific Workflows with Elastic Resource Provisioning in Virtualized Clouds based on PB (primary backup) scheduling technique.

This technique uses the vertical scaling-down scheme to avoid unnecessary and ineffective resource changes due to fluctuated workflow requests to make full use of the idle resources. To solve pricing issues linked to under provisioning; an Online Auction Framework for Dynamic Resource Provisioning was studied in [60]. It is based on efficient system optimization and heterogeneous VMs provisioning.

A randomized auction mechanism that efficiently allocates resources according to users' bids was presented. The auction mechanism in each round manages resources uses the one-round resource allocation method and decides the payments from the winning bidders.

The objective of this mechanism is to minimize users' budget spending and the latencies for VMs which are the major concerns in IaaS cloud provisioning. To solve various challenges in workflow applications, author in [61] combined resource provisioning and scheduling algorithm such as Particle Swarm Optimization (PSO) to solve the overall workflow execution cost and deadline constraints.

There is a concern to evaluate the performance of the resource provisioning algorithms. Paper in [62], suggested an approach for modeling and analyzing the performance of the resource provisioning. The suggested model uses a

Stochastic Process Algebra (SPA) to simulate the components involved in the resource provisioning process along with the interactions among them.

To build the system model and evaluate the performance of the resource provisioning process, the SPA framework was used. SPA is built based on continuous-time Markov chains. The overview of SPA can be found in [63–66]. In SPA context, a system is composed of subsystems which are connected by algebraic operators and perform some private. The figure is the example of a storage resource provisioning structure in cloud computing (Figure 4).

Consider a storage system with enough disks capacity to satisfy the users' requests with the physical disks logically grouped into disk domains. The disk domains contain a set of disks of the same type or different types and are isolated from each other to carry different services, and they do not interfere with one another. The storage pools are the container created in the disk domains to provide storage space for hosts. The LUNs create logical disks in a storage pool. Hosts can access LUNs that have been added to a LUN groups and mapped. The mapping views view the access permissions and mappings among LUN groups, port groups, and host groups. Add initiators to hosts and add the hosts to host groups to establish a logical connection between application servers and the storage system. Also we can create file systems to allocate storage resources of the storage system as file directories. After file systems are created, you can use the NFS, CIFS, and FTP file-sharing services to enable clients to access storage resources.

5.3. Resource discovery

Resource discovery and resource allocation are critical issue in designing an enhanced and practical distributed cloud. Paper [67] introduces resource discovery by suggesting a Distributed Cloud Architecture to make use of independent resources provided by the users. Authors proposed multi-valued distributed hash tables for efficient resource discovery.

Authors then proposed a new auction method, using a reserve bid formulated rationally by each user for the optimal allocation of discovered resources. The paper [68] considers a fully decentralized resource discovery service based on an unstructured overlay. The major challenge is to locate desired resources without the global knowledge about sharing resource information.

The nodes which involved into resource discovery scheme should acquire higher network overhead. Therefore, the direction-aware resource discovery scheme comes to improve the overall performance.

The unstructured network constitutes the membership. Then, the direction-aware strategy helps to discover the desired resources according to the queries. Paper [69] proposed an approach of managing dynamic and autonomous resources in data centers. This technique adopts adjustments of the parameters in order to continuously manage the resources in a data center with less human intervention. The figure below represents the proposed resource discovery and resource allocation framework (Figure 5).

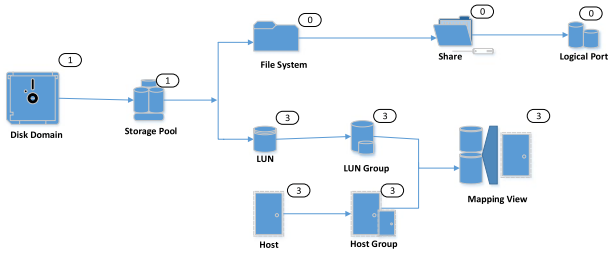

Figure 4. Storage resource provisioning structure.

6. Solution to resource management issues

6.1. Resource provisioning schemes

See Table 2.

6.2. Resource monitoring and performance management

This section explains how to manage resources' performance of a cloud resource in real-time environment. The measurement units based are total IOPS, read IOPS, write IOPS, total bandwidth, write bandwidth, total response time, read response time or write response time.

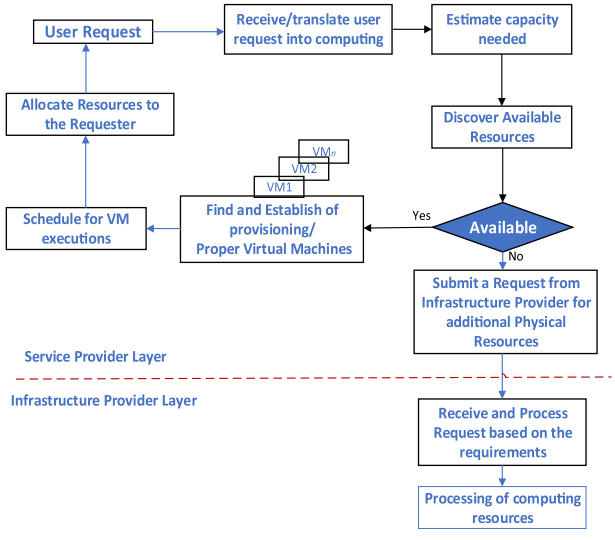

Figure 5. The proposed resource discovery and resource allocation framework.

Table 1. Summary for the requirements of cloud computing dynamic resource management.

	Characteristics	Requirements	Objectives
Cloud computing	On-demand self-service	Intelligent and business-related resource management	Quality of service (QoS) guarantee and Cost optimization
	Broad network access	End-to-end resource chain management resource location optimization	
	Resource pooling	Dynamic deployment management	
	Rapid elasticity	Dynamic adaptive resource management	
	Measured service	Monitoring and reporting of resource usage	
Green Cloud computing		Energy-efficient resource management	

Table 2. Performance metrics for resource provisioning schemes.

Name of the schemes	Functioning
Fault-tolerant scheduling [59]	Based on dynamic fault-tolerant scheduling algorithm to provide high resource utilization and high schedulability
Online Auction framework [60]	Based on auction mechanisms to optimize system efficiency
Meta-heuristic optimization technique [61]	Based on particle swarm optimization (PSO) to minimize the overall workflow execution cost
Agent based Automated Service Composition (A2SC) algorithm [70]	Based on A2SC algorithm to reduce the cost of VMs services.
Failure-aware Resource provisioning algorithm [71]	Based on Failure-Aware Resource provisioning algorithm to redirect users' requests to the appropriate cloud providers
On-demand provisioning, resource availability, [72]	Based on ASFLA (Augmented Shuffled Frog Leaping Algorithm) technique to minimize execution cost/ time
High bandwidth provisioning, low latency, low bit error rate and energy [73]	Based on polynomial-time energy-aware routing algorithm, Integer Linear Programming (ILP) and heuristic to design energy-aware paths and allocate servers and switches to accommodate the traffic requests
Fault tolerance, fault detection and fault recovery [74]	Surveyed different techniques such as self-healing, self-detection, preemptive migration, checkpoint, restart, replication, system node recovery, and job migration techniques
Hybrid approach [75]	Combined the concept of the autonomic computing and the (Reinforcement Learning) RL-based agent to predict the future demands of the cloud services.

The Monitoring information includes: CPU monitoring, to monitor the available/total resources, reserved capacity/reservation rate.

The Memory monitoring, to monitor the available/total capacity, reserved capacity/Reservation rate. VMs Monitoring: Name, ID, status (running/stooped), type, CPU Cores, CPU Usage, Memory Size, Memory Usage, used Memory, IP Address, Cluster, Host, tools status, operation, and more.

The Host Monitoring, to monitor the host status, maintenance mode, host CPU usage, host memory usage, cluster name, host IP address, and host operations.

The Task Tracing to trace the task information (task name, object name, start time, task status, task end time, and operator).

The Alarm Monitoring, to monitor the alarm information such as, current alarms, alarm Thresholds, alarm Masking, alarm Statistics, alarm dump. Event monitoring system shows you the past events.

The system alarm checks all the historical actions of the whole system. For example, the events name: virtual NIC abnormality, IP address of the VM changed, failed to get the VM ip, etc.

6.3. Energy-aware monitoring

Author [76], considered energy as a QoS metric. Author presented service framework which allow to monitor the power consumption of a cloud infrastructure, calculated its energy efficiency to put in place an effective virtual machine management. Author calculated the energy consumed by virtual and physical hosts by applying power models over the utilization of the resource and use external sensors and devices for measuring the power consumption.

Author [77] studied the optimization of the significant energy consumed in data center during the data storage and data processing. The author [77] has proposed the energy conservation framework and has enumerated three technologies to take into consideration: energy saving technologies in high-performance computing, energy conservation technologies for computer rooms, and renewable energy applications during the construction and operation of data centers. The application of the above mentioned technologies lead to the cost reduction, to the minimization of the environmental impact.

6.4. Resource mapping

The current resource management methodologies [5,11] are having the capability to map virtual system resources with physical systems dynamically depends on workload. To map VMs onto PMs is a key problem for cloud providers since PM utilization impacts revenue significantly. VM initial placement: refer to the process of mapping VMs onto PMs with optimization objectives. The challenge is

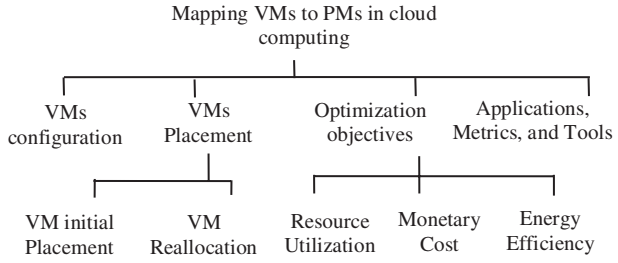

Figure 6. Mapping VMs onto PMs in cloud computing.

to know which VMs to allocate to which PMs to achieve the targeted goals. This action is formulated as an integer programming problem. Various application and system constraints may also be considered when making placement decisions. The applications with short deadlines are given a higher priority to prevent the SLA violations or minimize the delay if the deadline cannot be met or when execution can finish in time [78] (Figure 6).

Open challenge issues in resource mapping:

- Designing the algorithm that can find a fast mapping using genetic algorithms to speed up the mapping process and ensures the respecting of all task deadlines
- Minimize the cost of mapping the request
- Evaluating the services providers as possible candidates for hosting the applications,
- Efficient load balancing on substrate networks and partial reconfiguration of virtual networks.

6.5. Resource adaptation

6.5.1. Optimization objectives

The optimization objective, attempts to optimize three key categories: Resource utilization, cost, and energy consumption. The resource utilization is one of the main purposes of cloud providers. The resource wastage would lead to increased energy consumption and costs due to underloaded hosts.

Therefore, optimizing resources is a performance-related problem, it may be done with fairness. Host consolidation is the process of allocating multiple virtual machines onto physical machines to share the hardware resources and it can be done to increase the resource utilization.

Workload balancing techniques: Since the data storage increases rapidly in open environment, workload balancing distributes the dynamic load across multiple nodes to avoid node overload.

It helps in resource utilization and helps to increase the performance of the applications running to the host. There are different resources like memory, CPU, and network resources. The workload balancing is the process of finding heavy loaded resources and then transferring the extra load to the under loaded machines. The workload policies include round-robin scheduling to distribute evenly the requests among the available hosts [79,80]. Join the shortest Queue scheduling to balance the number of waiting VMs at the queues by assigning the request to the shortest waiting queue. Workload balancing techniques helps network and resource to achieve good network throughput with minimum response time by dividing the traffic between available hosts.

Opportunistic load balancing (OLB) [81] assigns tasks in free order to present node of useful which leads to the poor completion time. Min-Min algorithm [82] assigns tasks with minimum completion time which leads to the load starvation. Honeybee foraging behavior [83] is population based technique which achieves a global load balancing.

7. Open problems and suggestions

In this work, we have theoretically analyzed issues in resource management such as allocation, provisioning, adaptation, mapping, monitoring and surveyed models, techniques, the approaches work and discussed the advantages such as scalability, quality of service, optimal utility, reduced latency, reduced overheads, specialized environment, communication cost, computational time, energy efficiency, simplified interface. Many resource management allocation models have been investigated by different researchers under different virtualized environments. The performance of those approaches has been evaluated in referenced papers from academia and adopted at industry level.

7.1. Performance issues

Performance guarantee should be an important issue between the cloud service provider and cloud users. As mentioned in the previous paragraphs, there are many challenges in terms of performance of the cloud resources and the applications. Suggestions were given on how to improve the performance of resources and applications running in cloud environment. There are possible ways to improve the performance of cloud to model the system more accurately by identifying the parameters and the metrics which cause the system to be less productive.

Evaluation metrics: the metrics used to evaluate the performance of different VM mapping approaches and they are divided into four different groups of metrics:

Application performance metrics: it measures user satisfaction trying to capture some QoS level of performance which may include response time or execution time, the ratio of the capacity provided to the application to the

maximum capacity at optimal allocation, the number of SLA violations and delay.

Host consolidation Metrics: The host consolidation uses resource utilization, and system throughput.

Energy Efficiency metrics: The good way to measure the energy improvements is to consider overall power and energy consumption. Some power saving methods were proposed such as DVFS [84–86] on/off [87–91], DVS [92,93], VM migration [94], VM placement [95], Hibernation [96].

Monetary metrics: This evaluation is based on the service provider's cost, and user cost.

7.2. Security and privacy issue

The security and privacy are the major concerns in cloud computing since the users have no control over their data. Through the service agreement, cloud computing provider is delegated the right to access the data at any time. The data can be deliberately deleted or altered by the hackers. In Cloud Security Alliance reports [97] and studies on cloud security issues [98] have mentioned several security threats such as data breaches, data loss, denial of service, and more other malicious which can occur in cloud system information management. The security solution approaches were proposed such as authentication and authorization, Identity and Access Management, Confidentiality, Integrity, Availability. The study on authentication solution was conducted based on MiLAMob [99] to handle the real-time authentication events on behalf of consumer devices with minimal HTTP traffic. Another approach on identification and authentication is Public Key Infrastructure (PKI) X.509 certificates [100]. The authors propose cryptographic approach such as IBE, IBS to enforce identity and access management policies. The other proposed approaches are CloudProof [101], trusted cloud computing platform (TCCP) [102,103], Fuzzy authorization (FA) for cloud storage [104] to enforce the confidentiality, integrity, and availability of cloud resources.

7.3. Resource scheduling issues

7.3.1. Processor resources scheduling

Efficient processor resource scheduling leads to the reduction of dynamic energy consumption. Currently, researches on single and multiprocessor scheduling were conducted to reveal techniques to reduce energy consumption such as power-aware task scheduling (PTS) based on dynamic voltage scaling (DVS) [47], real-time DVS algorithms RT-DVS [105], dynamic voltage scaling-adaptive body biasing (DVS-ABB) [106].

7.3.2. Server resource scheduling

The purpose for server resource scheduling is to find the idle machines to be turn off when no services running on them to save the energy. We categorized the energy-saving solution into two categories: the VMs-based servers and PMs-based servers. One PM can support multiple VMs. The VMs can dynamically start, hang, or turn off. The resource assigned to the VMs can dynamically be adjusted. MV can dynamically migrate. VM migration increases the energy consumption and the migration cost. Kim, K. et al. [107] studied a new strategy combining the resources allocated dynamically and DVS VM to achieve the energy-saving of server.

7.3.3. Workload scheduling

The majority of the service oriented applications like internet service providers, online gaming websites, and online shopping websites generate a huge unexpected workload especially on weekends and holidays. The resource management methods may fail under some circumstances. Therefore, the on-going study to develop new resource management techniques is a must.

8. Related work and literature survey

Author in [1] did a survey study to analyze dynamic resource allocation approaches in SDN and virtual network. In the context of SDN and virtual network, the objectives of allocating resource dynamically are based on the following metrics but not limited here: Cost reduction, improved resource utilization, management overhead reduction for service provider, etc. These dynamic resource approaches are as the followings: Joint allocation of scheduling of network resource; Hierarchical network-aware placement of service oriented applications; Design and Evaluation of learning algorithms for dynamic resource management in virtual network. These approaches are based on Integer Linear Programming model and help to determine how applications and services are allocated within the cloud. These approaches combine both dynamic VM with cloud application auto-scaling to reduce operational costs to cloud providers and clients.

In these approaches, clients themselves deploy, manage, and scale up applications in an on-demand model and pay-per-use manner using three operations: VM relocation, VM Consolidation, and VM placement.

Benefits of SDN in cloud computing are: Network resources, network communication security, focusing on network bandwidth management, network device management, and network traffic management.

Parikh, S. M. (2013), surveyed on dynamic resource allocation in SDN to provide a cost reduction while improving the resource usage [36].

In paper [38], Vakilinia, et al., has proposed performance modeling techniques with application of power management in mobile cloud computing environment. To model the system, authors adopted two processes. Poisson Process and birth–death processes. Firstly, authors consider modeling a system with homogeneous VMs, constant job sizes and simultaneous release times. Secondly, they consider modeling a system with heterogeneous VMs, constant job sizes and simultaneous release times. They considered single, multiple server, and multiple server pools cases. Authors group jobs into classes and assign them to a system according to joint probability distribution. Each class of jobs is allocated to VMs according to Poisson Process.

Consider two systems, one with finite resources, and system with infinite resources. For finite resources some jobs will be blocked if resources are not enough to serve them. But for infinite resources, there are always idle VMs ready to execute the service immediately. The dynamic resource allocation based on IEDA (Intelligent Economic Approach) is explained in [108]. In it, authors have proposed an IEDA uses the improved combinatorial double auction protocol to allow various kinds of resources traded between cloud consumers and cloud providers. Authors considered bidding in the following services: VMS (virtual machine service), (CPS) computation service, DBS (database service), and STS (storage service). Authors present the improved combinatorial double auction protocol, including tender description, price formation, and winner determination.

Auction mechanisms have attracted attention of man researchers as an efficient approach for pricing and allocating resources.

An Online Auction Framework for Dynamic Resource Provisioning was studied in [60]. It is based on optimized system efficiency model dynamic and provisioning of heterogeneous VMs types in practice. A randomized auction mechanism which efficiently allocates resources according to users' bids was presented.

The auction mechanism in each round allocates resources based on the one-round resource allocation problem and decides the payments from the winning bidders. [109], reviewed the state-of-the-art application prediction methods and the existing challenging issues in resource provisioning.

Author in [110] has suggested an efficient resource management approach by analyzing the resource allocation logs of Virtual servers, SLA agreements and follows the resource prediction algorithms to predict future resource requirements.

To address the above problem, in their paper they introduced PFRRF framework called Predicting Future Resource Requirement Framework to predict the future resource needs. PFRRF considers to analyze the log files of resource allocation system, and to estimate the future requirements. To achieve it, their framework uses the Resource Prediction Algorithm (RPA).

The semi-structured log files should be transformed to the structured log files, which on the other hand are having the periodic resource allocation charts (PRAC) for every PM (physical machine) running on cloud environment. Authors have proposed hour bounded and day-bounded resource prediction methodologies.

The author in [3] has provide the rich materials on self-organizing RMS and has highlighted the main open challenges related to this area such as security issue, QoS or SLA guarantees, energy efficiency. Some techniques such as bio-inspired (such as ACO for workload consolidation, Honey Bees for load balancing) computing, multi-agent systems, evolutionary techniques were also studied to make resources more robust and adaptable.

Author in [35], has surveyed the state of the art in resource provisioning, resource allocation, resource mapping, and resource adaptation. He also mentioned the resource modeling, resource estimation, resource discovery, resource brokering and resource scheduling. Paper [5] also has surveyed and categorized with short explanations about some of these areas in their work.

Author in [78] conducted a survey study on VM placement for cloud computing. Author focused on scheduling and mapping problems using different optimization objectives. In his survey, his first priority is to analyze how the VMs can be mapped onto PMs to optimize different objectives and he evaluated the efficiency of different solutions in different situations.

9. Conclusion

In our study, we have surveyed different resource management techniques which can be applied to manage cloud resources. Among those techniques are resource provisioning, resource discovery, resource monitoring, resource mapping, and so on. We also have mentioned the opening problems which can be solved to make the cloud system more manageable and efficient. Those open challenges are: resource performance issues, security and privacy issue, and server, processor, workload resources scheduling. The following metrics were applied: CPU utilization, VMs allocation SLA violation, total cost, and profit. In our next paper, we will focus on cloud resource scheduling based on heuristic techniques.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation of China [grant number 61202508], [grant number 61272432], [grant number 61370132], [grant number 61472033], [grant number 61370092]; Fundamental Research Funds for the Central Universities [grant number FRF-TP-14–045A2]; Rwanda Education Board.

Notes on contributors

Frederic Nzanywayingoma was born in Rwanda, September, 1981. He received his BS degree in Electronic and Communication Systems Engineering from National University of Rwanda, in 2010, and his MS degree in Information Communication Engineering from University of Science and Technology, Beijing, China, in 2013. Currently, he is a PhD candidate at the same university. His research interests include Machine to Machine Communications, Cloud Computing, Scheduling Algorithms, and Optimization Methods.

Yang Yang graduated from Beijing Iron and Steel Institute of Automation System in 1982. He received his PhD in Information Engineering, from University of Science and Technology in Lillie, France, in 1988. He has been a professor at University of Science and Technology, Beijing since 1988. He has published in journals and conferences both at home and abroad more than 200 papers, completed books. His research interests include Service Science and Cloud Computing, Intelligent Control, Image Processing and Pattern Recognition, Multimedia Communication, Grid Technology.

ORCID

Frederic Nzanywayingoma http://orcid.org/0000-0001-5883-4946

References

[1] de Oliveira-GRR20112021 FAN, Risso-GRR20120726 JVT. Dynamic resource allocation in software defined and virtual networks: a comparative analysis.
[2] Bari MF, Boutaba R, Esteves Ret al. Data center network virtualization: a survey. IEEE Commun Surv Tutorials. 2013;15(2):909–928.
[3] Endo PT, Batista MS, Goncalves GE, et al. Self-organizing strategies for resource management in Cloud Computing: State-of-the-art and challenges. 2nd IEEE Latin American Conference on Cloud Computing and Communications (LatinCloud); Maceio, Brazil; 2013.
[4] Nguyen Van H, Dang Tran F, Menaud J-M. Autonomic virtual resource management for service hosting platforms. Proceedings of the ICSE Workshop on Software Engineering Challenges of Cloud Computing; IEEE Computer Society; Vancouver, Canada; 2009.
[5] Ullrich M, Lässig J, Gaedke M. Towards efficient resource management in cloud computing: a survey. IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud); Vienna, Austria; 2016.
[6] Mell P, Grance T. The NIST definition of cloud computing; 2011.
[7] Lee G, Katz RH. Heterogeneity-aware resource allocation and scheduling in the cloud. HotCloud; 2011.
[8] Ngenzi A, Nair SR. Dynamic resource management in Cloud datacenters for Server consolidation. arXiv preprint arXiv:1505.00577; 2015.
[9] Coffman EG, Garey MR, Johnson DS. Dynamic bin packing. SIAM J Comput. 2006;12(2):227–258.
[10] Koneru S, Uddandi VR, Kavuri S. Resource allocation method using scheduling methods for parallel data processing in cloud. Int J Comput Sci Inf Technol [IJCSIT]. 2012;3(4):4625–4628.
[11] Maguluri ST, Srikant R, Ying L. Stochastic models of load balancing and scheduling in cloud computing clusters. International Conference on Computer Communications; Orlando, FL; 2012.
[12] Kang Z, Wang H. A novel approach to allocate cloud resource with different performance traits. IEEE International Conference on Services Computing (SCC); Santa Clara, CA; 2013.
[13] Zhang Q, Cheng L, Boutaba R. Cloud computing: state-of-the-art and research challenges. J Internet Services Appl. 2010;1(1):7–18.
[14] Hadji M, Zeghlache D. Minimum cost maximum flow algorithm for dynamic resource allocation in clouds. IEEE 5th International Conference on Cloud Computing (CLOUD); Honolulu, HI; 2012. p. 876–882.
[15] Yang HC, Dasdan A, Hsiao RL, et al. Map-reduce-merge: simplified relational data processing on large clusters. Proceedings of the ACM SIGMOD international conference on Management of data; Beijing, China: ACM; 2007.
[16] Jung G, Sim KM. Agent-based adaptive resource allocation on the cloud computing environment. 40th International Conference on Parallel Processing Workshops (ICPPW); Taipei City, Taiwan; 2011.
[17] Feng T, Bi J, Wang K. Joint allocation and scheduling of network resource for multiple control applications in SDN. IEEE Network Operations and Management Symposium (NOMS); Krakow, Poland; 2014.
[18] Moens H, Hanssens B, Dhoedt B, et al. Hierarchical network-aware placement of service oriented applications in clouds. IEEE Network Operations and Management Symposium (NOMS); Krakow, Poland; 2014.
[19] Di S, Wang C-L. Dynamic optimization of multiattribute resource allocation in self-organizing clouds. IEEE Trans Parallel Distrib Syst. 2013;24(3):464–478.
[20] Leelipushpam PGJ, Sharmila J. Live VM migration techniques in cloud environment–a survey. IEEE Conference on Information & Communication Technologies (ICT); Thuckalay, Tamil Nadu, India; 2013.
[21] Rabbani M. Resource management in virtualized data center; 2014.
[22] Sonkar S, Kharat M. A review on resource allocation and VM scheduling techniques and a model for efficient resource management in cloud computing environment. International Conference on ICT in Business Industry & Government (ICTBIG); Anaheim, California; 2016.
[23] Singh G, Behal S, Taneja M. Advanced Memory Reusing Mechanism for Virtual Machines in Cloud Computing. Procedia Comput Sci. 2015;57:91–103.

[24] Kale RMCO. Virtual machine migration techniques in cloud environment: a survey.
[25] Vyas N, Chauhan A. A survey on virtual machine migration techniques in cloud computing.
[26] Yang Y, et al. Disk failure prediction model for storage systems based on disk SMART technology. Int J Comput Appl. 2015;37(3–4):111–119.
[27] Hines M, Deshpande, U, Gopalan K. Post-copy based live virtual machine migration using adaptive pre-paging and dynamic self-ballooning. OSR. 2009;43(3): 14–26.
[28] Jin H, Pan D, Xu J, et al. Efficient VM placement with multiple deterministic and stochastic resources in data centers. Global Communications Conference; Anaheim, California; 2012.
[29] Zhang L, et al. Moving Big Data to the cloud: an online cost-minimizing approach. IEEE Journal on Selected Areas in Communications. 2013;31(12):2710–2721.
[30] Hwang I, Pedram M. Hierarchical Virtual Machine Consolidation in a Cloud Computing System. International Conference on Cloud Computing; Santa Clara, California; 2013.
[31] Wood T, et al. Memory buddies: exploiting page sharing for server consolidation in virtualized data centers. Citeseer; 2007. (Technical Report).
[32] Vernekar A, Anandalingam G, Dorny C. Optimization of resource location in hierarchical computer networks. Comput Oper Res. 1990;17(4):375–388.
[33] Wood T, Shenoy PJ, Venkataramani A, et al. Black-box and gray-box strategies for virtual machine migration. NSDI'07 Proceedings of the 4th USENIX Conference on Networked Systems Design & Implementation; Cambridge, MA; 2007. p. 17.
[34] Kumar M, Sharma SC. Dynamic load balancing algorithm to minimize the makespan time and utilize the resources effectively in cloud environment. Int J Comput Appl. 2017;1–10.
[35] Manvi SS, Shyam GK. Resource management for Infrastructure as a Service (IaaS) in cloud computing: a survey. J Network Comput Appl. 2014;41:424–440.
[36] Parikh SM. A survey on cloud computing resource allocation techniques. Engineering (NUiCONE). Nirma University International Conference; Ahmedabad, Gujarat India; 2013.
[37] Madni SHH, et al. Resource scheduling for infrastructure as a service (IaaS) in cloud computing: Challenges and opportunities. J Network Comput Appl. 2016;68:173–200.
[38] Vakilinia S, Ali MM, Qiu D. Modeling of the resource allocation in cloud computing centers. Comput Network. 2015;91:453–470.
[39] Lee L-T, et al. A dynamic resource management with energy saving mechanism for supporting cloud computing. Int J Grid Distrib Comput. 2013;6(1):67–76.
[40] Zhang Y. Classified scheduling algorithm of big data under cloud computing. Int J Comput Appl. 2017;1–6.
[41] Younge AJ, Von Laszewski G, Wang L, et al. Efficient resource management for cloud computing environments. International Green Computing Conference; Chicago, IL; 2010.
[42] Kong W, Lei Y, Ma J. Virtual machine resource scheduling algorithm for cloud computing based on auction mechanism. Optik. 2016;127(12):5099–5104.
[43] Kim N, Cho J, Seo E. Energy-credit scheduler: an energy-aware virtual machine scheduler for cloud systems. Future Gener Comput Syst. 2014;32:128–137.
[44] Ding Y, et al. Energy efficient scheduling of virtual machines in cloud with deadline constraint. Future Gener Comput Syst. 2015;50:62–74.
[45] Chen LY, Ansaloni D, Smirni E, et al. Achieving application-centric performance targets via consolidation on multicores: myth or reality? Proceedings of the 21st international symposium on High-Performance Parallel and Distributed Computing; ACM; 2012.
[46] Wang W, Jiang Y, Wu W. Multiagent-based resource allocation for energy minimization in cloud computing systems. IEEE Trans Syst Man Cybern Syst; 2017;47:205–220.
[47] Kim KH, Beloglazov A, Buyya R. Power-aware provisioning of cloud resources for real-time services. Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science; 2009.
[48] Mishra M, Das A, Kulkarni P. Dynamic resource management using virtual machine migrations. IEEE Commun Mag. 2012;50(9):34–40.
[49] Clark C, Fraser K, Hand S, et al. Live migration of virtual machines. Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation. Vol. 2; Cairo, Egypt: USENIX Association; 2005. p. 273–286.
[50] Bacalhau ET, Usberti FL, Lyra C. A dynamic programming approach for optimal allocation of maintenance resources on power distribution networks. IEEE Power and Energy Society General Meeting (PES); 2013.
[51] Awasare V, Deshmukh S. Survey and comparative study on resource allocation strategies in cloud computing environment. IOSR J Comput Eng. 2014;16(2):94–101.
[52] Abdelaal MA, Ebrahim GA, Anis WR. Network-aware resource management strategy in cloud computing environments. 11th International Conference on Computer Engineering & Systems (ICCES); Cairo, Egypt; 2016. p. 26–31.
[53] Verma M, et al. Dynamic resource demand prediction and allocation in multi-tenant service clouds. Concurrency and Computation: Practice and Experience; 2016.
[54] Nair TG, Vaidehi M. Efficient resource arbitration and allocation strategies in cloud computing through virtualization. IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS); Beijing, China; 2011. p. 397–401.
[55] Huang C-J, et al. An adaptive resource management scheme in cloud computing. Eng Appl Artif Intell. 2013;26(1):382–389.
[56] Lin W, et al. A threshold-based dynamic resource allocation scheme for cloud computing. Procedia Eng. 2011;23:695–703.
[57] Xu B, et al. Job scheduling algorithm based on Berger model in cloud environment. Adv Eng Software. 2011;42(7):419–425.
[58] Singh AN, Prakash S. Challenges and opportunities of resource allocation in cloud computing: a survey. 2nd International Conference on Computing for Sustainable Global Development (INDIACom); New Delhi, India; 2015. p. 2047–2051.

[59] Zhu X, Wang J, Guo H, et al. Fault-tolerant scheduling for real-time scientific workflows with elastic resource provisioning in virtualized clouds. IEEE Trans Parallel Distrib Syst. 2016;27(12):3501–3517.

[60] Shi W, et al. An online auction framework for dynamic resource provisioning in cloud computing. IEEE/ACM Trans Network. 2016;24(4):2060–2073.

[61] Rodriguez, MA, Buyya R. Deadline based resource provisioningand scheduling algorithm for scientific workflows on clouds. Cloud Comput. 2014;2(2):222–235.

[62] Shawky DM. Performance evaluation of dynamic resource allocation in cloud computing platforms using Stochastic Process Algebra. 8th International Conference on Computer Engineering & Systems (ICCES); Cairo, Egypt; 2013. p. 39–44.

[63] Bernardo M, Donatiello L, Ciancarini P. Stochastic process algebra: from an algebraic formalism to an architectural description language. IFIP International Symposium on Computer Performance Modeling, Measurement and Evaluation; Rome, Italy: Springer; 2002. p. 236–260.

[64] Baeten JC, Weijland WP. Process algebra, volume 18 of Cambridge tracts in theoretical computer science; Cambridge: University Press Cambridge; 1990.

[65] Hermanns H, Herzog U, Mertsiotakis V. Stochastic process algebras as a tool for performance and dependability modelling. Proceedings of the International Computer Performance and Dependability Symposium; Erlangen, Germany; 1995. p. 102–111.

[66] Herzog U. Formal description, time and performance analysis a framework. Entwurf und Betrieb verteilter Systeme; Springer; 1990, p. 172–190.

[67] Khethavath P, Thomas J, Chan-Tin E, et al. Introducing a distributed cloud architecture with efficient resource discovery and optimal resource allocation. IEEE Ninth World Congress on Services (SERVICES); 2013.

[68] Chung WC, Hsu CJ, Lai KC, et al. Direction-aware resource discovery service in large-scale grid and cloud computing. IEEE International Conference on Service-Oriented Computing and Applications (SOCA); Irvine, CA; 2011. p. 1–8.

[69] Yazir YO, Matthews C, Farahbod R, et al. Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis. IEEE 3rd International Conference on Cloud Computing (CLOUD); 2010.

[70] Singh A, Juneja D, Malhotra M. A novel agent based autonomous and service composition framework for cost optimization of resource provisioning in cloud computing. J King Saud University Comput Inf Sci; 2017;29(1):19–28.

[71] Javadi B, Abawajy J, Buyya R. Failure-aware resource provisioning for hybrid Cloud infrastructure. J Parallel Distrib Comput. 2012;72(10):1318–1331.

[72] Kaur P, Mehta S. Resource provisioning and work flow scheduling in clouds using augmented Shuffled Frog Leaping Algorithm. J Parallel Distrib Comput. 2017;101:41–50.

[73] Yang S, et al. Energy-aware provisioning in optical cloud networks. Comput Networks. 2017;118:78–95.

[74] Cheraghlou MN, Khadem-Zadeh A, Haghparast M. A survey of fault tolerance architecture in cloud computing. J Network Comput Appl. 2016;61:81–92.

[75] Ghobaei-Arani M, Jabbehdari S, Pourmina MA. An autonomic resource provisioning approach for service-based cloud applications: a hybrid approach. Future Gener Comput Syst. 2017.

[76] Katsaros G, et al. A service framework for energy-aware monitoring and VM management in clouds. Future Gener Comput Syst. 2013;29(8):2077–2091.

[77] Rong H, et al. Optimizing energy consumption for data centers. Renew Sust Energy Rev. 2016;58:674–691.

[78] Pietri I, Sakellariou R. Mapping virtual machines onto physical machines in cloud computing: a survey. ACM Comput Surv (CSUR). 2016;49(3):49.

[79] Chaudhari A, Kapadia A. Load Balancing Algorithm for Azure Virtualization with Specialized VM. algorithms. 2013;1:2.

[80] Sran N, Kaur N. Comparative analysis of existing load balancing techniques in cloud computing. Int J Eng Sci Invent. 2013;2(1):60–63.

[81] Katyal M, Mishra A. A comparative study of load balancing algorithms in cloud computing environment. arXiv preprint arXiv:1403.6918; 2014.

[82] Pinto P. Introducing the Min-Max Algorithm. Submited to the AI Depot article contest. 2002;1–10.

[83] Ld DB, Krishna PV. Honey bee behavior inspired load balancing of tasks in cloud computing environments. Appl Soft Comput. 2013;13(5):2292–2303.

[84] Kaur S, Kaur A, Gobindgarh M. Energy aware resources allocation heuristic for efficient management of data centers for cloud computing; 2016.

[85] Beloglazov A, Abawajy J, Buyya R. Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. Future Gener Comput Syst. 2012;28(5):755–768.

[86] Garg SK, et al. Environment-conscious scheduling of HPC applications on distributed Cloud-oriented data centers. J Parallel Distrib Comput. 2011;71(6):732–749.

[87] Burge J, Ranganathan P, Wiener JL. Cost-aware scheduling for heterogeneous enterprise machines (CASH'EM). IEEE International Conference on Cluster Computing; Austin, TX; 2007. p. 481–487.

[88] Bradley DJ, Harper RE, Hunter SW. Workload-based power management for parallel computer systems. IBM J Res Dev. 2003;47(5.6):703–718.

[89] Lefèvre L, Orgerie A-C. Designing and evaluating an energy efficient Cloud. J Supercomput. 2010;51(3):352–373.

[90] Salfner F, Lenk M, Malek M. A survey of online failure prediction methods. ACM Computing Surveys (CSUR). 2010;42(3):10.

[91] Wang, C.-F, Hung W-Y, Yang C-S. A prediction based energy conserving resources allocation scheme for cloud computing. IEEE International Conference on Granular Computing (GrC); Noboribetsu, Hokkaido, Japan; 2014. p. 320–324.

[92] Chen Y, Das A, Qin W, et al. Managing server energy and operational costs in hosting centers. ACM SIGMETRICS performance evaluation review; 2005.

[93] Mezmaz M, Melab N, Kessaci Y, et al. A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems. J Parallel Distrib Comput. 2011;71(11):1497–1508.

[94] Tesfatsion SK, Wadbro E, Tordsson J. A combined frequency scaling and application elasticity approach for energy-efficient cloud computing. Sustainable Comput Inf Syst. 2014;4(4):205–214.
[95] Khosravi A, Garg SK, Buyya R. Energy and carbon-efficient placement of virtual machines in distributed cloud data centers. European Conference on Parallel Processing; Aachen, Germany: Springer; 2013. p. 317–328.
[96] Subrata R, Zomaya AY, Landfeldt B. Cooperative power-aware scheduling in grid computing environments. J Parallel Distrib Comput. 2010;70(2):84–91.
[97] Islam T, Manivannan D, Zeadally S. A classification and characterization of security threats in cloud computing. Int J Next-Gener Comput. 2016;7(1).
[98] Gholami A, Laure E. Security and privacy of sensitive data in cloud computing: a survey of recent developments. arXiv preprint arXiv:1601.01498; 2016.
[99] Lomotey RK, Deters R. Saas authentication middleware for mobile consumers of iaas cloud. IEEE Ninth World Congress on Services (SERVICES); 2013.
[100] Kim H, Timm SC. X. 509 authentication and authorization in fermi cloud. Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing; London, England: IEEE Computer Society; 2014. p. 732–737.
[101] Popa RA, Lorch JR, Molnar D, et al.. Enabling security in cloud storage SLAs with CloudProof. USENIX Annual Technical Conference; 2011.
[102] Garfinkel T, Pfaff B., Chow J, et al. Terra: a virtual machine-based platform for trusted computing. ACM SIGOPS Operating Systems Review; 2003.
[103] Santos N, Gummadi KP, Rodrigues R. Towards Trusted Cloud Computing. Proceedings of the 2009 Conference on Hot Topics in Cloud Computing; San Diego, CA; 2009. p. 1–5.
[104] Zhu S, Gong G. Fuzzy authorization for cloud storage. IEEE Trans Cloud Comput. 2014;2(4):422–435.
[105] Pillai P, Shin KG. Real-time dynamic voltage scaling for low-power embedded operating systems. ACM SIGOPS Operating Systems Review; 2001.
[106] Martin SM, Flautner K, Mudge T, et al. Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads. Proceedings of the 2002 IEEE/ACM International Conference on Computer-Aided Design; 2002.
[107] Kim KH, Beloglazov A, Buyya R. Power-aware provisioning of virtual machines for real-time Cloud services. Concurr Comput Pract Exper. 2011;23(13):1491–1505.
[108] Wang X, et al. An intelligent economic approach for dynamic resource allocation in cloud services. IEEE Trans Cloud Comput. 2015;3(3):275–289.
[109] Amiri M, Mohammad-Khanli L. Survey on prediction models of applications for resources provisioning in cloud. J Network Comput Appl. 2017;82:93–113.
[110] Prasad B, Angel S. Predicting future resource requirement for efficient resource management in cloud.