# Emerging Paradigms and Practices in Cloud Resource Management


**Durga Prasad Sharma, Bhupesh Kumar Singh, Amin Tuni Gure, and Tanupriya Choudhury**

## 1 Introduction

Originally cloud computing is a computational model that facilitates on-demand resources of computing systems and services, especially virtual machines, cloud storage, computing, and communication software, with the least involvement of user's inactive modes. These computing resources are provided on-demand as metered services like a public utility. In the primary phase, the main architecture of cloud computing was designed based on characteristics of utility computing, and later the cloud resources were pooled at distributed places in centralized modes and characterized as data canters [1].

To realize the Journey of cloud computing technology, we need to go back in history. In the 1950s, scientist Herb Grosch (a well-known author of Grosch's law) hypothesized that the entire world would operate on dumb terminals powered by about 15 large data centers, i.e., perceived as modern cloud data canters [2]. Later, the term "Cloud" was used as a symbolic representation or metaphor for the Internet and an abstraction of the underlying network infrastructure. Cloud computing [3, 4] was introduced by John McCarthy in 1960 with a concept of the illusion of an infinite supply of resources. The actual term "Cloud" was borrowed from telephony in that telecom companies, who until the 1990s offered primarily dedicated point-to-point data circuits, began offering Virtual Private Network (VPN) services with comparable quality of service but at a much lower cost [5].

D. P. Sharma · B. K. Singh (✉) · A. T. Gure
Computing and Software Engineering, Arba Minch University, Arba Minch, Ethiopia
e-mail: sharma.dp@amu.edu.et; dr.bhupeshkumarsingh@amu.edu.et; amin.tuni@amu.edu.et

T. Choudhury
Department of Informatics, School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand, India



T. Choudhury et al. (eds.), *Autonomic Computing in Cloud Resource Management in Industry 4.0*, EAI/Springer Innovations in Communication and Computing,
https://doi.org/10.1007/978-3-030-71756-8_2

Next to the dot-com bubble, Amazon played a key role in the rapid development of cloud computing [6, 7] by modernizing their own data centers. It was an initial major step toward the computing paradigm and revolutionizing the cloud technology. As an innovator, Amazon opened the door for access to cloud computing resources to external consumers. In 2006, Amazon launched Amazon Web Service (AWS) on a utility computing platform for world consumers and became the pioneer of cloud computing in real sense.

In early 2008, Eucalyptus also entered into the cloud market and became the first open-source cloud service provider, AWS API-compatible platform for deploying and facilitating cloud computing resources privately [8]. Also, at the same time in early 2008, Open Nebula, was declared as the first and the foremost open-source software for deploying the private, hybrid, and other federated clouds [9]. In mid-2008, Gartner observed a scope for cloud computing (i.e., to shape the relationship among consumers of information technology services and information technology service providers). Later, this was revolutionized as "switching from company-owned hardware and software assets to per-use service-based models" [10].

Cloud computing is a general term for anything that involves the delivery of hosted services over the network, i.e., the Internet. Cloud computing can be viewed as access to resources from a set of pooled computing resources needed to perform functions with dynamically changing needs. Cloud can be perceived as a techno-business paradigm in which hosted resources are delivered over the Internet to perform certain tasks with dynamically changing needs of resources. In fact, the cloud is a convergence model for enabling convenient, on-demand access to a shared pool of computing resources such as computing servers, storage grids, networks, application software, computing tools, and services in a convenient and ubiquitous environment.

These infrastructure products and services are nothing but "resources" that can be provided as service with quick provision or reprovision at anytime, anywhere over any device, and released with nominal admin efforts or user intervention. The cloud models presented in Fig. 1 are composed of five essential characteristics (e.g., on-demand self-service, broad network access, resource pooling, location independence, rapid elasticity, and measured service), three service models (IaaS, PaaS, and SaaS), and four deployment models (Public, Private, Hybrid, and Community) [11–14].

The use of information technology (ITs) by multidimensional applications has been changing with respect to time dynamics. These dynamic changes create a new horizon of the vibrant echo system in computing, communication, and collaboration environment. The individual users of the big enterprises need on-demand computing, communication, and collaboration resources. This scenario reflects the fact that contemporary IT needs have been dynamically evolving, and motivation is shifting from owned infrastructure, i.e., capital expenditure to operational expenditure. This implies that the popularity of rent-based infrastructure is rapidly increasing than own infrastructure. The acceptance of cloud computing can be visualized by incremental growth and investment migration over the cloud rather than the purchase of new infrastructure. The enhanced business efficiency

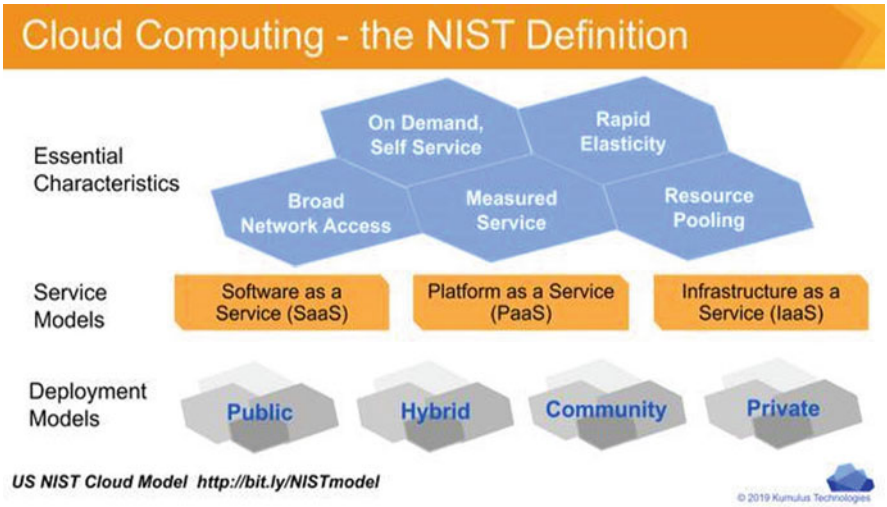


**Fig. 1** The cloud models

through the growing usage of IT services offered by cloud computing will further boost the growth in migration of IT services toward cloud, especially in small- and medium-scale enterprises (SMEs) [15].

In general, cloud computing [16–18] enables users to migrate their data storage and computational needs to a remotely available infrastructure with minimal user intervention and impact on computing system performance. Usually, this offers a variety of benefits that could not be experienced when computing over traditional infrastructure like, for instance, on-demand high scalability, both vertical and horizontal; freedom of provision and reprovision of computing resources with a variety of options; and high uptime, i.e., 99.99%. All such variety of IT tools, products, and services could be accessed in an easy and user-centric manner over cloud. These cloud resources can be anything like hardware or software (network, storage, applications, developing tools, high-performance system, etc.).

In SMEs and Big Enterprises, workflows have emerged as a technique to formalize and structure data analytics, perform computations over distributed cloud resources, gather the output of the processed data, and then repeat the analysis if required for desired results. As a matter of fact, the SMEs cannot afford rapid changing high-end modern IT needs for exploring the full potential of structured or unstructured data analytics collected from the salient distributed sources to manage and alleviate the competitive needs [19].

Also, the scientific workflows in scientific collaborations enable the sharing of data analytics results, and therefore the scientific workflows have been viewed as an emerging paradigm, where engineers and data scientists can handle complex scientific processes easily and conveniently to share worldwide for rapid result disseminations in scientific discoveries and research [20].

This cloud computing, business, and scientific workflows convergence is enabled by resource management, which includes resource provisioning and reprovisioning, scheduling and rescheduling, and allocation and reallocation [21, 22].

In a Cloud computing [23, 24] environment, smart and portable systems such as Mobiles, Tablets, and Fablets and services are highly anticipated, as provisioning of inefficient resources may result in the failure of timeliness of task processing [25] [26]. To avoid such issues and challenges, provisioning the most feasible computing resource, most fit storage space, and suitable application can significantly reduce the unpredictable monetary losses. Such critical cost savings with no substantial impact on application performance can be considered as a good sign toward efficient management of cloud resources.

## 2 Literature Review

### *2.1 Cloud Resource Management*

In the cloud resource management, the significant challenges are efficient allocation of resources to the workload based on the specifications, energy efficiency, uptime, on-demand horizontal and vertical scalability, consumer satisfaction, trust, transparency, and QoS. Resources are hardware or software entities used for computing and communications [27]. Resource management is the process or method of allocating appropriate computing, communication, storage, and other resources to run the applications as per the needs of cloud consumers. The cloud SLA specifications are kept in the center while allocating the resources. Cloud resource management is a dynamic process that deals with locating and releasing resources in an environment, where the dynamics of the needs and specifications frequently change. The efficient and effective utilization of the resources in any computing model like the cloud is highly anticipated. Today greenness or energy efficiency of the resources has been declared as one of the most important QoS. The other issues in resource management are violations of SLA and efficient load balancing with high service availability, i.e., uptime 99% [28].

It is easy to procure the resources but difficult to deploy, deliver, and manage the customer workloads in cloud environments, where worldwide customers and their resource dynamics fluctuate within a very small quantum of time. The arrival of the CSPs such as Amazon, Microsoft, and Google, which are ranging from scientific applications to the business, commerce, industry, academia, and personal use, creates the need for ultra-advanced resource management solutions with complex systems design and management strategies. The SLAs specify the need specification of the cloud resources, and it is the sole responsibility of the CSPs to fulfill the resource requirement of the customers to maintain the trust and transparency for customer satisfaction. In the cloud environment, heterogeneity of the resources is the major challenge as they need well-designed and well-tested robust solutions

for complex system management. The convergence of performance data analytics and automatic resource management carries new challenges and opportunities. It becomes difficult for the system integration and management designers to transform the theoretical models and conceptions into practically implementable solutions. In order to achieve this, resource management in the cloud requires well-structured and agreed policies and efficient decisions for multi-objective optimization of resources. These policies can be categorized into five classes or processes: (1) admission control, (2) capacity allocation, (3) load balancing, (4) energy optimization, and (5) quality of service guarantees [29, 30]. This chapter covers the general concepts of cloud resource management and investigates the trust, transparency, and QoS in service-level agreements (SLAs) as a case-based experimental analysis.

## *2.2 Essential Concepts and Definitions*

Since the beginning, the cloud models were designed by their built-in techno-business characteristics. Usually, the cloud infrastructures from its foundation are provided as utility computing resources and availed in cost-effective scale to utilize resource offering in a pay-per-use model [31, 32]. Several authors have defined the necessary components and their functionalities. According to [33], resource management comprises nine major processes:

- Resource provisioning: Assigning the desired resources to a workload based on need specifications.
- Resource discovery: Identification or discovery of a list of cloud resources that are available for workload handling or execution.
- Resource modeling: A standard framework that helps in predicting the resource specifications required by a workload based on attributes such as states, transitions, inputs, and outputs within a given cloud environment.
- Resource scheduling: Cloud resource scheduling can be defined as the mapping, allocation, and execution of workloads based on the cloud resources shortlisted (provisioned) in the provisioning phase. It can also be defined as a timetable of activities and resources, with start and end points along with the duration of the workloads. The quality attributes of services such as cost-effectiveness, timeliness, energy efficiency, etc. (i.e., as promised under service-level agreement (SLA)) are also aligned.
- Resource allocation: Balanced distribution of resources among competing workloads with minimum conflicts.
- Resource mapping: Negotiations between resources required by the workload and resources provided by cloud service providers.
- Resource estimation: Prediction of the actual resources required for executing a workload efficiently.

- Resource brokering: Arbitration or negotiation of cloud resources through a mediator entity (agent) to guarantee their availability at the right time to execute or handle the workload.
- Resource adaptation: Ability to dynamically adjust (elasticity) the desired resources to fulfill the dynamic requirements of the workload efficiently.

Complexity, the variety, and the nature of the data are dynamically changing. The scientific and business organizations nowadays rely on the analytics of the complex, varied, and voluminous data sets, and the processing must be done over on-demand scalable and auto-configurable computing resources. The substantial performance enhancement and overhead reduction in virtualization boosted its adoption as a key feature in cloud computing technology [34, 35].

There are salient underlying techniques, technologies, and their configurations that transform computing over the cloud in reality. Among these technologies, the most significant technologies are the virtualization of data center resources for access to enormous processing capabilities and scalable resources to handle complex data with unpredictable computing needs.

## 3 Functions of Cloud Resource Management

The main essence of resource management is to recognize the suitable resources for a specific workload to handle in the most proficient manner. The quality-of-service specifications are determined by the consumers. This process is known as provisioning of the most suitable computing resources [36].

As mentioned in Fig. 2, the cloud resource management consists of three main functions—provisioning, scheduling, and monitoring.


**Fig. 2** Resource management in cloud computing environment [redesign]

### *3.1 Cloud Recourse Provisioning*

In the resource provisioning, the first step is consumer authentication. Afterward, the consumer interacts with the cloud servers via a cloud portal and submits the resource requirements of the workload along with quality specifications. In this process, the Resource Information Centre (RIC) maintains the status of the pooled resources and provides this state information to the customer about the availability of the requested resources for handling or executing the workload. The Resource Provisioning Agent (RPA) is a responsible component that checks the availability of the requested resources by the customer, i.e., what is required and the states of availability. When the resource provisioning is over, the customer workloads are submitted to the next component, i.e., scheduler. Finally, the resource states information is submitted to the Workload Resource Manager (WRM) which forwards it to the Resource Provisioning Agent, and the final results are forwarded to the cloud customer.

### *3.2 Cloud Resource Scheduling*

Cloud resource scheduling consists of the three processes, i.e., mapping, allocation, and execution of workloads, based on the cloud resources shortlisted (provisioned) in the aforementioned provisioning phase. This is usually performed aligning with quality attributes of services such as cost, timeliness, energy efficiency, etc. as promised under service-level agreement (SLA) [37].

The whole process consists of the three activities that are: (1) Mapping—selection of the suitable resources based on the quality-of-service specifications (i.e., mentioned in SLA) of the customer, (2) detection—identification or discovery of the list of available cloud resources, and (3) Selection—choosing the most feasible resource from the list produced by detection based on SLA.

### *3.3 Cloud Resource Monitoring*

The monitoring and surveillance processes of cloud resources are also required to be autonomic. Cloud resource monitoring supports in achieving the desired performance as promised (i.e., SLA specifications). As per the standard agreements of SLA, both the parties (cloud service provider and cloud service consumer) must specify and agree on the possible deviations or violations in service terms and conditions so as to manage promised quality attributes in SLA and avoid the conflicts. Subsequently, this phase also controls the rescheduling of activities in the cloud environment.

This phenomenon state is necessary for the optimization of the trust and transparency in metering and monitoring of cloud resources consumptions. It is

envisioned that the violations or deviation must be less than the defined thresholds for successful execution of a workload in the cloud environment. Logically, resource monitoring is also one of the important quality attributes that should be taken care of seriously when trust and transparency are categorically mentioned as the essential QoS specifications like availability of services, uptime, and performance specifications, and security [7]. In the monitoring process, the existing workload states are compared to the number of required cloud resources. In the case of less, more resources are demanded by the resource scheduler so as to maintain the SLA provisions and promises. If the resources are sufficiently available in the pool, the resources can also be released and made available for allocations.

### *3.4 Resource Management Techniques/Methods*

The cloud computing resource management has a variety of solutions and techniques that are accumulated and classified from the literature survey.

Effective and efficient resource utilization is confined to the optimization and assured by algorithms running in the cloud environment. The researchers [38] classified the cloud resource management into nine classes/categories. In this scheduling [39] solution, cost, time, success rate, scalability, make span, speed, resource utilization, reliability, and availability were considered. Usually, the reliability and availability have lots of similarities, but they were typically ignored; however, time, speed, and make span are sufficiently described as interconnected properties. The 12 properties were defined in the research [40] such as (1) *Time-based:* its deadline based by the blending of deadline and budget, (2) *Cost-based:* It is multi-QoS, application, virtualization, and scalability based; (3) *Compromised Cost: It is time* based either on workflows or workloads; (4) *QoS-based:* Created on several QoS aspects, such as resource utilization and security; (5) *SLA-based: Created on the baseline* SLA types, such as autonomic feature and workload; (6) *Energy-based: It connects the* deadlines and SLAs; (7) *Optimization-based*: It optimizes permutation and combinations of parameters; (8) *Nature and Bio-Inspired*: It includes the genetic algorithms and ant colony approaches; (9) *Dynamic*: It includes the dynamic aspects of resource management with salient permutation and combinations; (10) *Rule-based*: It considers the special cases for failures and hybrid clouds, (11) *Adaptive-based*: Prediction-based and Bin-Packing strategies; and (12) *Bargaining-based*: It is organized in market based, auction, and negotiations.

Another study of [41] classified the resource management solutions in relations to scalability, interoperability, flexibility, heterogeneity, localized autonomy, load balancing, information exposure, past scheduling records, unpredictability management, real-time data, geographical distribution, SLA compatibility, rescheduling, and intercloud compatibility. In this study, several properties are overlapping or correlated, such as rescheduling, scalability, and managing unpredictable phenomenon.

A research study of [42, 43] proposed nine categories to classify their references such as (1) *Best effort: Single* objective optimization by ignoring other factors;

(2) *Deadline constrained*: When the deadline is set, it schedules based on the execution time and monetary cost; (3) *Budget constrained:* finishing within budget, (4) *Multicriteria:* combining many objectives together, (5) *Workflow as a service:* A moment when the resource manager receives many workflow instances to perform; (6) *Robust scheduling*: Capability of handling uncertainties such as performance fluctuation and failure together; (7) *Hybrid environment*: ability of handling hybrid cloud requirements; (8) *Data intensive*: scheduling with data-aware workflows; and (9) *Energy aware*: ability of greenness while optimizing execution.

After rigorous review and analysis of resource management techniques, it is clear that this field still lags behind in terms of trust, transparency, and QoS in resource management and needs vigorous improvements and extensions in the existing techniques and methods. The claim becomes significantly more important when we explore cloud computing applications in a wider spectrum of multi-cloud and industry 4.0. The cyber-physical production system that combines ICTs, cyberspace, and intelligent systems is expanding the pathways of Industry 4.0 in salient dimensions toward multidimensional revolutions like traditional manufacturing to intelligent system-supported manufacturing.

Also, the convergence of the Internet of Things into Industry (i.e., Industrial Internet of Things (IIoTs)) has created new ways of computing, communication, collaboration, and control toward a new era of automation. In order to comprehend these transformations into reality, a wide range of resource management optimization and dynamics of connected resources will need reengineering. The existing process of computing and communication in automated environments such as cloud, fog, and due computing needs serious attention and collaborative research on salient tiers of research such as management and effective monitoring and control over the Service-Level Agreement (SLAs) for efficient and effective communication and interaction among the mobile system components and devices in autonomic manners [44, 45].

## *3.5 Service-Level Agreements (SLAs) Gaps in Cloud Computing*

In general, a service-level agreement (SLA) is the bond for performance arbitration between the CSPs and the customer. Most of the SLAs are standardized. The SLAs are also categorized at different levels: (1) Client-side SLA, (2) Service-level SLA, and (3) Multilevel SLA. Most of these contracts are more along the lines of operating-level agreements (OLAs) and may not be restricted by the court of law. On ample occasions, these SLAs are violated, and therefore they need to have an attorney to review before agreeing to the CSPs. The SLA contracts usually specify some measuring/metering parameters such as availability of the Service outage (uptime), the Response time (latency), QoS (greenness), Service Configuration, Service components reliability, and Warranties. If a CSP

fails to meet the stated/warranted requirements of minimums, then the CSP has to pay the compensation/penalty to the consumer. Microsoft publishes the Service-Level Agreements linked with the Windows Azure Platform components, which is demonstrative of industry practice for cloud service vendors. Each individual component has its own Service-Level Agreements such as Windows Azure SLA and SQL Azure SLA [46].

Effective cloud resource management needs robust implementation techniques to manage the resources of cloud data centers. However, the Service Level-Objective (SLO) is a judicious range in order to achieve optimum performance in business service operations.

The energy efficiency and ineffective resource metering, monitoring, and utilization can build better trust and transparency among CSPs and Cloud Service Consumers (CCC); however, it can lead to an increase in the operational costs. Also, an increase in the cloud resource utilization needs energy efficiency, as quality-of-service (QoS) parameters are nowadays towards green computing initiatives. However, combining cloud resources such as virtual machines can cause a serious violation of SLAs [47, 48].

The cloud service providers (CSPs) are responsible for metering and monitoring of the consumed cloud resources. The cloud resources for computing, communication, storage, and other purposes need efficient and effective frameworks for metering and monitoring mechanisms. The metering and monitoring systems must utilize the trusted and transparent scales so that the scheduling of globalized or localized resource allocation and utilization can be optimized.

It was envisioned that this rapid growth in the technology sectors will warrant a substantial techno democratic environment in terms of trust, transparency, and empowerment of loyal consumers. As a matter of fact, most of the computing and communication system services have been metered and billed in a monopolized method by the service provider companies/enterprises. Most often, the consumers have to believe in the metered service measurements and pay the bills accordingly. What if the service-level agreements are violated in terms of promises between the service provider and the consumer, and metering system reading outputs are manipulated? In most of the service-level agreements, a cross-verification or metered data tally at the client side is still at the embryonic stage toward judicious empowerment of the consumer rights. Also, the weak steps of statutory compliance, settlement, and decree make violations very thoughtful in the consumer market. A system for preserving such trust and transparencies in functional and nonfunctional attributes is highly anticipated in Telecom, ICT, and Clouds service sectors.

Since the cloud resources are deployed on the virtual infrastructure, therefore, consumers lack or have limited privileges of metering and monitoring consumed services and resources. CSPs have domination, and therefore there are ample chances of violations or deviations on dynamical alterations in the prices charged for leasing the infrastructure, while cloud users can alter the costs by changing application parameters and usage levels only. However, the cloud consumers have limited privilege for resource management, being embarrassed to generate workload requests and control when and where the workloads are to be found.

The client-side monitoring and tally system is required to be implemented by regulatory authorities, and both the client and the service providers must abide by the monitoring and tally system regulatory framework and mechanisms. This system can enhance the trust and transparency in metering and billing systems for the betterment of consumer rights.

### *3.6 The CSPs SLA Monitoring Mechanism (Table 1)*

### *3.7 Client-Centric SLA Framework for Enhancing the Trust, Transparency, and QoS*

Figure 3 presents a proposed Client-Centric SLA Framework. The framework labels the metering and monitoring of consumed cloud resource services in an enhanced democratic manner to empower consumer rights. The client-centric SLA systems can help support in cross-verifying the warranties of the quality-of-service (QoS) attributes promised in SLAs such as energy efficiency and standard certifications. The proposed framework is the vigorously unique conceptualization of cloud consumer empowerment through incremental growth in managing the trust and transparency in metering and monitoring of consumed cloud resource services.

### *3.8 The Client-Centric SLA Framework*

The prime aim of the framework is to empower the client rights in a trusted and transparent manner with QoS. This framework is an effort toward the betterment of two-party business relations, i.e., between cloud service consumers (CSCs) and cloud service providers (CSPs). Cloud Service Providers (CSPs): CSPs as mentioned in Fig. 3 are cloud service provider organizations that need to be regulated for consumer rights. The CSPs deliver cloud services to the consumers per their business workload specifications and requirements as guaranteed in SLAs. In general, the CSPs collect the consumer feedbacks and store them in the Service Management Databases (SMDB). This feedback is analyzed for consistent improvements in the cloud service delivery mechanisms.

***Cloud Service User (Customers/Consumers):*** The cloud service consumers or customers are the end users of CSP services and may be an individual or organizational entity that maintains a business relationship with CSPs to consume cloud services. Generally, the existing state-of-the-art metering and monitoring systems at CSPs are single sided, monopolized with the lack of consumer rights, and empowered in service provisioning and monitoring. In today's democratic system governance and management environment, consumerism is an essential stratum, and consumer trust and transparency should be maintained for the vibrant ecosystem

**Table 1** Describes the cloud service provider and monitoring of SLA [49]

| CSP | Types of service | Step 1 dis-covering the service provider | Step 2 defining SLA | Step 3 making agreement | Step 4 monitoring the violations of SLA | Step 5 terminate SLA | Step 6 penalty for SLA violation |
|---|---|---|---|---|---|---|---|
| Amazon Ec2 | Computing (IaaS) | Discovering manually via website | Predefined SLA terms and QoS specifications | Predefined SLA document of CSPs. | Third-party monitoring system (e.g., cloud watch) can be used | Monitoring by third regulatory party/using program by CSP or manually by both the parties | Additional service crediting by CSP or refunds or legal actions |
| Amazon S3 | Storage (IaaS) | Discovering manually | Predefined SLA terms and QoS specifications | Predefined SLA document of CSPs. | Third-party monitoring system (e.g., cloud status) can be used | Monitoring by third regulatory party/using program by CSP or manually by both the parties | Additional service crediting by CSP or refunds or legal actions |
| MS azure compute | PaaS | Discovering manually (e.g., website) | Predefined SLA terms and QoS specifications | Predefined SLA document of CSPs. | Third-party monitoring system (e.g., monitis) can be used | Monitoring by third regulatory party/using program by CSP or manually by both the parties | Additional service crediting by CSP or refunds or legal actions |
| MS azure storage | PaaS | Discovering manually | Pre-defined SLA terms and QoS specifications | Predefined SLA document of CSPs. | Third-party monitoring system (e.g., monitis) can be used | Monitoring by third regulatory party/using program by CSP or manually by both the parties | Additional service crediting by CSP or refunds or legal actions |
| MS azure software | SaaS | Discovering manually | Pre-defined SLA terms and QoS specifications | Predefined SLA document of CSPs. | Third-party monitoring system (e.g., monitis) can be used | Monitoring by third regulatory party/using program by CSP or manually by both the parties | Additional service crediting by CSP or refunds or legal actions |

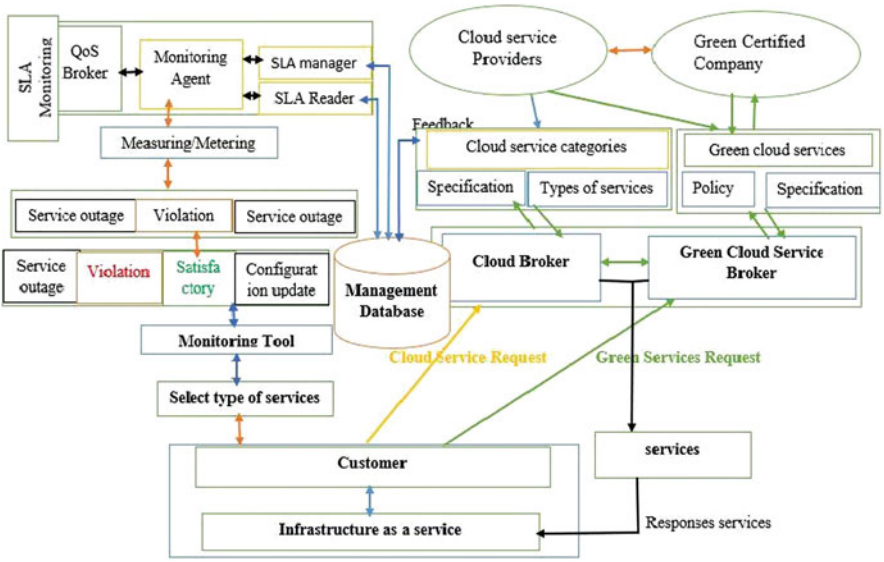


**Fig. 3** Client-centric SLA framework

in the CSP industry. It can minimize the conflict between consumerism and professionalism. The proposed framework (i.e., in Fig. 3) proposes a new way for client-side metering and monitoring of consumed cloud services. This framework can be implemented as a metering and monitoring tool on cloud consumer devices. The central cloud or IT regulatory agencies may authorize/certify as a tally or auditing tool for better relationships among consumers and CSPs. This framework "Client-Side SLA" will empower the consumers judiciously.

***Cloud Brokers:*** There are two types of brokers, i.e., (1) Cloud Service Broker and (2) Green Cloud Broker. This component of the framework (i.e., in Fig. 3) proposes to offer the most suitable services with better prices, efficiency, and QoS based on needs or service specifications of the client/consumer. This component has two major responsibilities as a typical cloud service broker with an additional QoS feature, i.e., energy efficiency/ green certification of cloud resources or services. The green cloud service broker verifies the suitability of the consumer's energy efficiency specifications/green certifications and recommends the services based on their preference and specifications included in SLA. Further, the Green Cloud Service Broker (GCSB) verifies greenness of services declared by CSPs with the certification issued by competent authorities and generates a validation certification to filter the false claims [50–52].

***SLA Metering (Measuring) and Monitoring Agent:*** The framework as presented in Fig. 3 consists of three different agents, i.e., (1) SLA Readers, (2) Monitoring Agent, and (3) QoS Broker. The functionality of different agents may vary from one CSPs to another. SLA reader reads the signed SLA from the CSPs and Consumer both which are stored in the database having the exact value of

parameters with full transparency and trust via the Internet. It feeds the signed SLA to the monitoring agent. The QoS Broker is responsible for monitoring the nonfunctional requirements (i.e., greenness of services and others) and collects data from the customer and disseminates the information to the CSPs.

***Layers of SLA Monitoring Framework:*** The functionality of the layers may vary as they work based on the assigned values. As presented in Fig. 3, the following layers are included in the framework:

***Application Layer:*** The logic tier is pulled out from the presentation tier and, as its own layer; controls an application's functionality by performing detailed processing. The application layer receives the results from the lower layer. The Metering and the Monitoring agent provide notifications about SLA state (violated or not?). So, the application layer forwards the results of the monitoring/ agent to the consumers which are displayed by the presentation layer. Actually, presentation is the topmost level of the application. The presentation layer displays the results of browsing merchandise, purchasing, and shopping cart contents.

***Service Management Layer:*** The monitoring layer provides the results to the upper layer (i.e., the presentation layer). The monitoring layer includes different types of components such as SLA Reader, QoS Broker, and Monitoring Agent. The QoS Broker gathers the data from the client and disseminates to the CSPs to set judicious compensation for the violated services.

***Database Layer:*** This layer stores information of the SLA with the exact parameter specified in the SLA. The database contains SLAs that are specified by the CSPs and Cloud Service Consumer. Thus, client-side measuring and monitoring SLA can improve the trust and transparency between the involved parties. The following diagram presented in Fig. 4 presents the metering and monitoring layers of clients-side SLA.

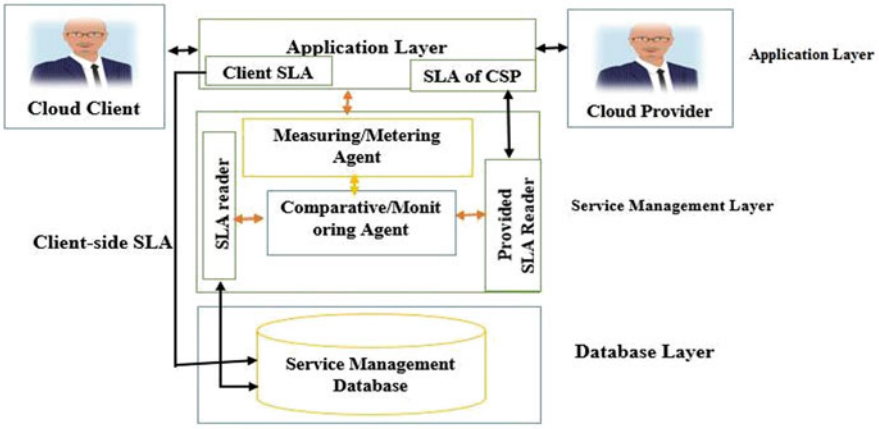


**Fig. 4** Metering and monitoring layers of client-side SLA

## 4 Experimental Analysis and Discussions

For the testing of the conceptual framework, the client-side SLA was implemented over cloud-based AppNeta. The Alternative to AppNeta, the tools that were studied, analyzed, and compared with parametric analysis are (1) Microsoft System Center, (2) Datadog, (3) LogicMonitor, (4) ThousandEyes, (5) NinjaRMM, (6) Zabbix, and (7) Wireshark.

Finally, AppNeta was selected as a cloud tool to monitor and manage applications and network performance. The first experimental setup test results of AppNeta for AWS and Azure Cloud Data Centre are presented in Fig. 5. The test results were recorded for a week from October 25, 2019, to October 31, 2019. In this test, some of the selected services were ordered by a customer to Amazon AWS. In this experimental test, as presented in Fig. 5, the (1) service outage—0.192%, (2) SLA violation—2.154%, (3) satisfactory services—93.531%, and (4) configuration update—4.122% over cloud data centers were recorded. Also, in an ideal state, 99.99% (uptime), i.e., the service availability was promised in cloud data center SLAs, but in actual it was recorded as 95.686%. The experiments confirmed the

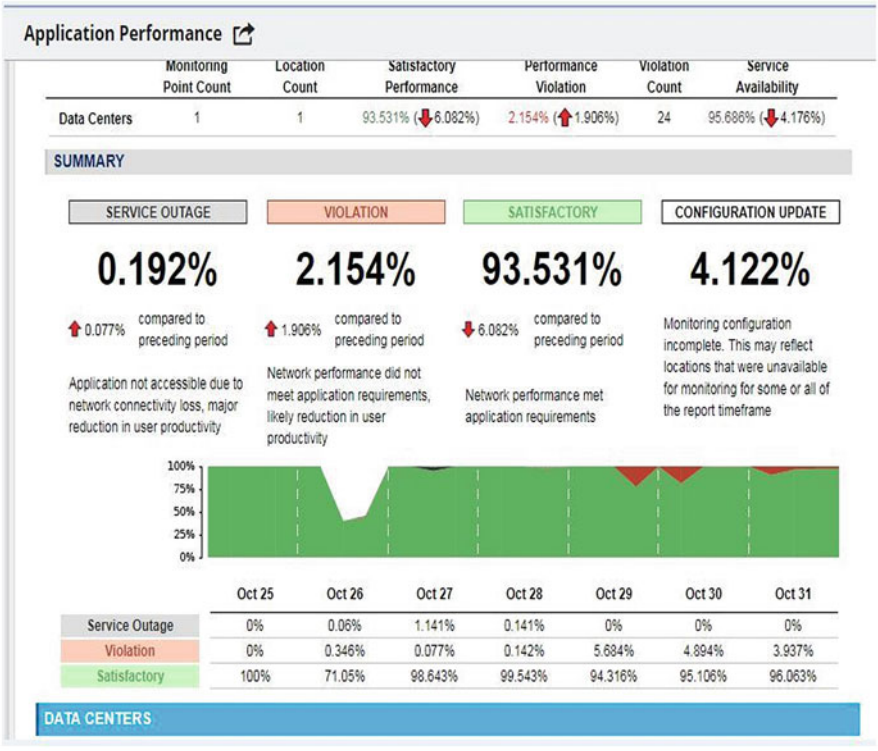


**Fig. 5** Experiment 1- AWS data center measured services by AppNeta on the client-side machine

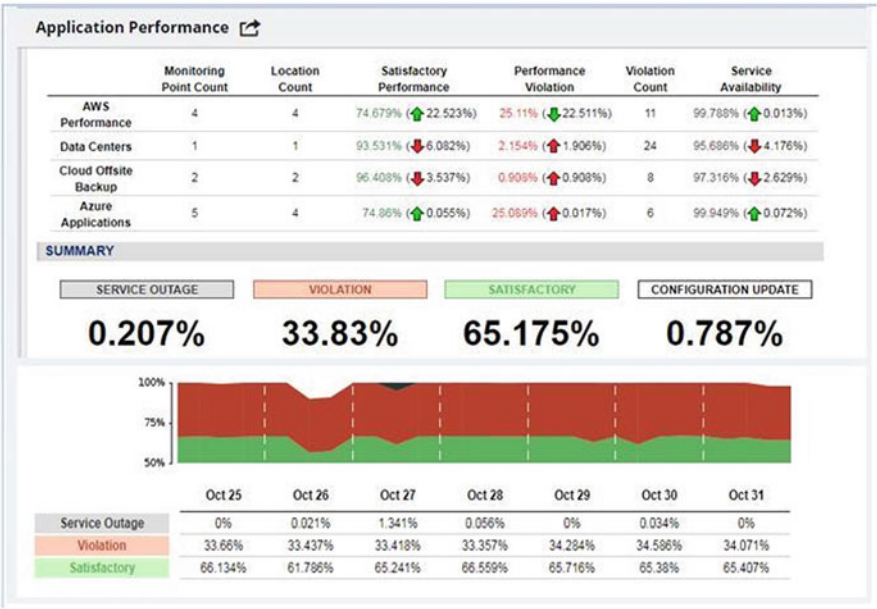


**Fig. 6** The applications performance on Amazon AWS and MS Azure run on the client-side machine

SLA violation, and therefore using the measured/metered results on the client-side machine, the customer can ask the compensation for the deviation of the terms and standard promised in SLA. Or otherwise, case customers can terminate the contract agreements.

In the second experimental test (Fig. 6), another CSP, i.e., MS Azure and Amazon AWS, was considered. The test results were recorded for a week from October 25 to 31 October 2019. In this test, some of the selected services were ordered by a customer to Amazon AWS & MS Azure. This CSPs promised 99.99% (uptime) service availability. In this experimental setup, the test results at Azure and AWS cloud data center services by the AppNeta were recorded as: (1) service outage—0.207%, (2) SLA violation—33.83%, (3) satisfactory services—65.175%, and (4) configuration update—0.787%. This clearly implies that if cloud consumers have their own metering and monitoring tools for SLA, they can negotiate with CSPs either in terms of compensation for the service violation or terminate the agreement. The consumers can ask for compliance settlement through regulatory authorities or negotiations and hence the trust and transparency will be enhanced. The performance of applications was also observed and recorded in this experimental setup. As presented in Fig. 6; the fluctuations (ups and downs) and violations can be clearly observed in AWS performance, data center services, cloud off-site backups, and Azure applications.

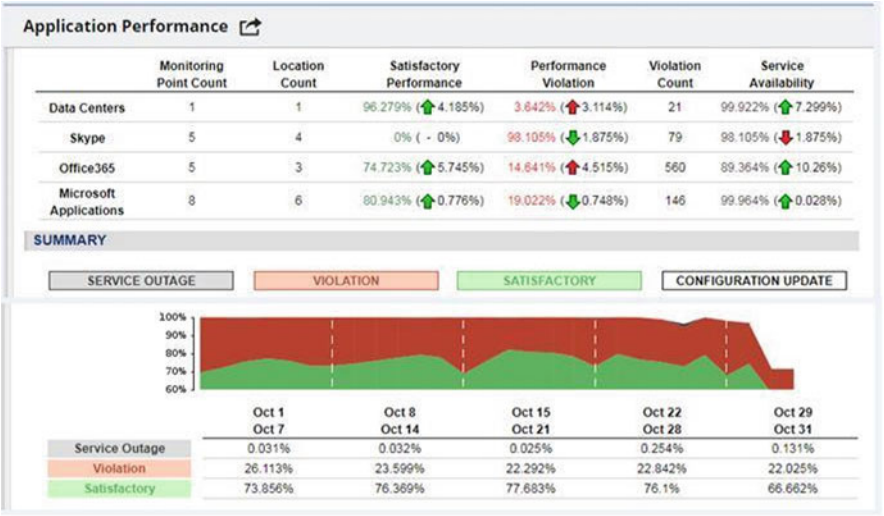


**Fig. 7** Amazon application performance metering on the client-side machine (month report)


**Fig. 8** Performance of the Amazon AWS and MS Azure cloud service providers

Another experimental setup presents the metered result for the application performance of the services provided by Amazon AWS. The result presented in Fig. 7 clearly indicates that SLA violation is at second position and satisfaction of services is in the first position. Here service outage is negligible. The customer can also view the performance of the cloud services provided by CSPs.

The experimental results in Fig. 8 present the performance of both Amazon AWS and MS Azure. In addition, the customer can also measure and monitor the service performance to check which performs better before they place an order and sign SLA.

The following experimental results in Fig. 9 present the performance of the Amazon AWS measured for one week of the Service Outage, violated and Satisfactory

The experimental result presents the performance of MS Azure products and services. Using the following result, the customer can decide which service provider is the best or the most suitable based on consumer requirement specifications. Based on the one-week reports presented in Fig. 10, the SLA promised 99.99%

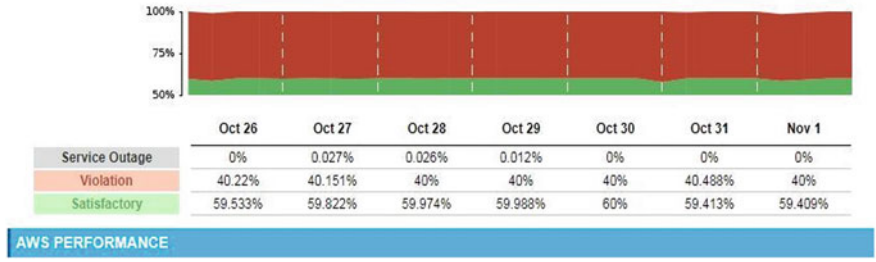


**Fig. 9** The Performance testing of the Amazon AWS for one week on the client-side machine

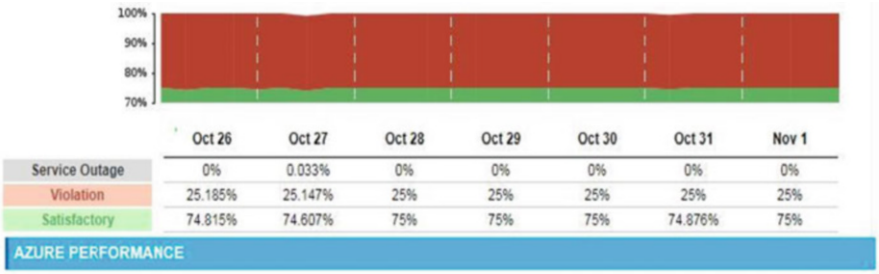


**Fig. 10** The Performance measurement of MS Azure for a one-week report on the client-side machine

(uptime) but measured and found 25.1% (uptime) which is a typical violation. The satisfactory result of the client side is 74.899 % out of the promised result of 99.999% which is written in the Service-Level Agreement.

According to the result presented in Fig. 11, the MS Azure almost failed to fulfill the promised performance mentioned in SLA. According to 1-week report, the service outage found was 6.868%, violated result was 55.568%, and the satisfying result was 36.524%. According to the energy efficiency standards, benchmark, and measurements, the power usage Effectiveness (PUE) can be calculated by cloud-based metering greenness/energy efficiency tools such as 42U. The measurement results can be compared to the energy efficiency declared by CSPs of Data Centre resources. The customers have the right to terminate the cloud service agreement or to ask for the compensation credits for the greenness violations or service outage results.

## 5 Challenges in Data Centre Resource Management

Resource management in a cloud environment is a typical challenge due to the following issues in the modern data centers [27, 53].

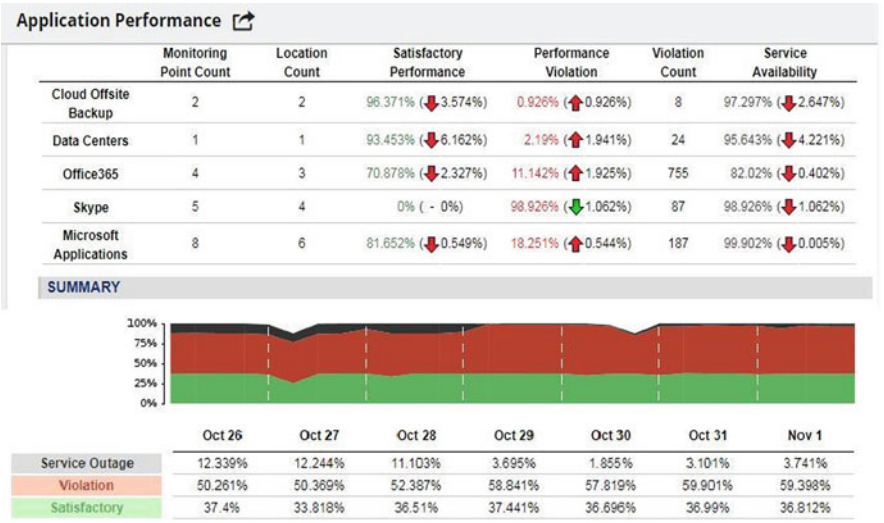


**Fig. 11** The application performance on the MS Azure platform for the 1-week report

- The fault tolerance of the resources
- The interoperability and the interdependence of the resources
- The high level of scalability of the resources
- The heterogeneity of the resource in multicloud
- The variability and unpredictability of the resource load
- The salient players and multi-objectivity in a cloud ecosystem
- The data-intensive workflows
- The energy-aware resource scheduling
- The reliability to performance fluctuations
- Communication and transfer costs of resources
- The dominance of the execution time and cost
- The data placement strategies design for decision-making while resource provisioning
- The performance fluctuations in multitenant resource sharing
- The workflow scheduling in the management of workflow execution in cloud environments

## 6 Conclusion

This chapter provides an exhaustive investigation and analysis of concepts, definitions of cloud resource management, and the review of existing techniques of management, SLA, and violations. The chapter started from evolution of the concepts, defining the terms and references on the subject area, covering the

basics of the foundation published in salient publications from the research and academia. Among the salient common tasks in resource management, each phase of the resource life cycle, such as resource discovery, allocation, scheduling, and monitoring, is also covered. Moreover, the critical objective in all cases is to enable task execution while optimizing infrastructural efficiency. These most important issues related to cloud resource management are also covered. A rigorous review of literature is incorporated for the characterization and selected solutions of the pinpointed issues, challenges, and gaps in resource management. Finally, the chapter concluded that trust, transparency, and QoS (greenness) in cloud resource metering and monitoring are necessary. The experimental analysis of the proposed framework for the client-side metering and monitoring of SLA proved that there are violations in the metering of cloud services along with the QoS claimed and the QoS provided. The serious violations are observed at CSP sides, but neither clients nor CSPs have abundant attention to resolve such issues seriously for the betterment of the trust and transparency between consumers and CSPs. The essential solutions must be designed, developed, and deployed with future research recommendations in high dynamic and scalable environment of cloud resource management.

## 7 Unanswered Questions as Recommendations for the Future Research Efforts

- How to discourse the particularities of data-intensive workflows and address the particularities of large-scale cloud setups with more complex environments in terms of resource heterogeneity and distribution, such as hybrid and multicloud?
- How to handle the fluctuations in workflow progress due to performance variation and reliability and to maintain reliability based on actual and measurable metrics.

## References

1. Marinescu, D. C. (2013, July 8). *Cloud computing: Theory and practice*. Cambridge, MA: Elsevier.
2. Kumbhar, P. (2019, July 6). Summary of computing laws amdahl, dennard, gustafson, little more and more.
3. Dinote, A., Sharma, D. P., Gure, A. T., Singh, B. K., & Choudhury, T. (2020). Medication processes automation using unified green computing and communication model. *Journal of Green Engineering, 10*(9).
4. Singh, B. K., Sharma, D. P., Alemu, M., & Adane, A. (2020). Cloud-based outsourcing framework for efficient IT project management practices. *(IJACSA) International Journal of Advanced Computer Science and Applications, 11*(9), 114–152.
5. Chana, I., & Kaur, T. (2013, May). Delivering IT as a utility – A systematic review. *IJFCST, 3*(3), 11–30.
6. Tomar, R., Khanna, A., Bansal, A., & Fore, V. (2018). An architectural view towards autonomic cloud computing – data engineering and intelligent. *Computing*.

7. Yadav, A. K., Tomar, R., Kumar, D., & Gupta, H. (2012). Security and privacy concerns in cloud computing, *Computer Science and Software Engineering, 2*(5).
8. Wikipedia. (2013). Cloud computing. The free encyclopaedia.
9. Ward, J. S., & Barker, A. (2014). Observing the clouds: A survey and taxonomy of cloud monitoring. *Journal of Cloud Computing, 24*(3).
10. unctad. (2019). Digital economy report. United Nation.
11. Douglas, D. D., & Barry, K. (2013). Cloud computing model. *Science Direct*.
12. Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. Gaithersburg, MD: National Institute of Standards & Technology.
13. Sharma, D. P., Sharma, R. K., & Ayodele, A. (2008). Convergence of intranetware in project management for effective enterprise management. *Journal of Global Information Technology (JGIT)-USA, 4*(2), 65–85.
14. Amin Tuni, D. P. S. (2019). Assessment of knowledge sharing practices in higher learning institutions: A new exploratory framework–AT-DP KSPF. *The IUP Journal of Knowledge Management, 17*(4), 7–20.
15. Intelligence, D. M.. (2020, June 12). Cloud migration services market, size, share, opportunities and forecast. *DDIC131*.
16. Dewangan, B. K., Agarwal, A., Choudhury, T., & Pasricha, A. (2020). Cloud resource optimization system based on time and cost. *International Journal of Mathematical, Engineering and Management Sciences, 5*(4). https://doi.org/10.33889/IJMEMS.2020.5.4.060
17. Wadhwa, M., Goel, A., Choudhury, T., & Mishra, V. P. (2019). Green cloud computing-A greener approach to IT. 2019 international conference on computational Intelligence and knowledge economy (ICCIKE) (pp. 760–764).
18. Kaur, A., Raj, G., Yadav, S., & Choudhury, T. (2018). Performance evaluation of AWS and IBM cloud platforms for security mechanism. 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 516–520).
19. Deelman, E., Gannon, D. B., Shield, M., & Taylor, I. J. (2014). *Workflows for E-science: Scientific workflows for grids*. London: Springer.
20. Li, Y., Raicu, I., Lu, S., Tian, W., Liu, H., & Zhao, Y. (2015). Enabling scalable scientific workflow management in the cloud. *Future Generation Computer System, 46*, 3–16.
21. Bubendorfer, K., & Arabnejad, V. (2015). Cost effective and deadline constrained scientific workflow scheduling for commercial clouds. In: Network Computing and Applications (NCA). *IEEE, 14th International Symposium On* (Vol. 33, pp. 106–113).
22. Shyam, G. K., & Manvi, S. S. (2014). Resource management for infrastructure as a service (IaaS) in cloud computing: A survey. *Journal of Networking and Computer Application, 141*, 424–440.
23. Choudhury, T., Gupta, A., Pradhan, S., Kumar, P., & Rathore, Y. S. (2018). Privacy and security of cloud-based internet of things (IoT). In Proceedings – 2017 international conference on computational intelligence and networks, CINE 2017. https://doi.org/10.1109/CINE.2017.28
24. Bansal, S., Gulati, K., Kumar, P., & Choudhury, T. (2018). An analytical review of PaaS-cloud layer for application design. In Proceedings of the 2017 international conference on smart technology for smart nation, SmartTechCon 2017. https://doi.org/10.1109/SmartTechCon.2017.8358374
25. Chard, K., Bubendorfer, K., Lacinski, L., Madduri, R., Foster, I., & Chard R. (2015). Cost-aware elastic cloud provisioning for scientific workloads. In IEEE 8th international conference on cloud computing (Vol. 130, pp. 971–974).
26. Yi, S., Andrzejak, A., & Kondo, D. (2012). Monetary cost-aware checkpointing and migration on amazon cloud spot instances. *IEEE Transactions on Services Computing, 15*(4), 512–524.
27. Jennings, B., & Stadler, R. (2015). Resource management in clouds: Survey and research challenges. *Journal of Network and System Management, 23*(3), 567–619.
28. Mustafa, S., Nazir, B., Hayat, A., Khan, A. R., & Madani, S. A. (2015). Hayat resource management in cloud computing: Taxonomy, prospects, and challenges. *Computer and Electrical Engineering, 47*, 186–203.

29. Marinescu, D. C. (2013). *Cloud computing: Theory and practice*. Waltham: Morgan Kauffman.
30. Asres, K., Gure, A. T., & Sharma, D. P. (2019). Automatic surveillance and control system framework-DPS-KA-AT for alleviating disruptions of social Media in Higher Learning Institutions. *Journal of Computer and Communications, 8*(1), 1–15.
31. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., et al. (2010). A view of cloud computing. Communications of ACM. New York.
32. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., et al. (2010). A view of cloud computing and communication. *ACM, 53*(4), 50–58.
33. Manvi, S. S., & Shyam, G. K. (2014). Resource management for infrastructure as a service (IaaS) in cloud computing: A survey. *Journal of Network and Computer Application, 141*, 424–440.
34. Chieu, T. C., Mohindra, A., Karve, A. A., & Segal, A. (2009). Dynamic scaling of web applications in a virtualized cloud computing environment. Washington, DC.
35. Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., & Zagorodnov, D.. (2009). The eucalyptus open-source cloud-computing system. Shanghai.
36. Chana, I., & Singh, S. (2014). Quality of service and service level agreements for cloud environments: Issues and challenges. In Challenges, limitations and R&D solutions, Switzerland.
37. Chana, I., & Singh, S. (2016). A survey on resource scheduling in cloud computing: Issues and challenges. *Journal of Grid Computing, 14*(2), 217–264.
38. Bala, A., & Chana, I. (2011). A survey of various workflow scheduling algorithms in cloud environment In Nagpur.
39. Dewangan, B. K., Jain, A., & Choudhury, T. (2020). GAP: Hybrid task scheduling algorithm for cloud. *Revue d'Intelligence Artificielle, 34*(4), 479–485. https://doi.org/10.18280/ria.340413.
40. Singh, S., & Chana, I. (2016). Cloud resource provisioning: Survey, status and future research directions. *Knowledge and Information System, 49*(3), 1005–1069.
41. Sotiriadis, S., Bessis, N., & Antonopoulos, N. (2011). Towards inter-cloud schedulers: A survey of meta-scheduling approaches. In 2011 international conference on, Barcelona.
42. Wu, F., Wu, Q., & Tan, Y. (2015). Workflow scheduling in cloud: A survey. *The Journal of Supercomputing, 71*(9), 3373–3418.
43. Shekhawat, H. S., & Sharma, D. P. (2012). Hybrid cloud computing in E-governance: Related security risks and solutions. *Research Journal of Information Technology, 4*(1), 1–6.
44. Wan, J., Chen, B., Imran, M., Tao, F., Li, D., Liu, C., & Ahmad, S. (2018). Toward dynamic resources management for IoT-based manufacturing. In IEEE Communications Magazine (Vol. 56(2), pp. 52–59). IEEE.
45. Raptis, T. P., Passarella, A., & Conti, M. (2019). Data management in industry 4.0: State of the art and open challenges. In Access (Vol. 7, pp. 97052–97093). IEEE.
46. Ibrahim, A. A. Z. A., Kliazovich, D., & Bouvry, P. (2016). Service level agreement assurance between cloud services providers and cloud customers. Cartagena.
47. Mandal, R., Mondal, M. K., Banerjee, S., & Biswas, U. (2020). An approach toward design and development of an energy-aware VM selection policy with improved SLA violation in the domain of green cloud computing. *Journal of Super Computer, 76*, 7374–7393.
48. Daraghmeh, M., Melhem, S. B., Agarwal, A., Goel, N., & Zaman, M. (2018). Linear and logistic regression based monitoring for resource Management in Cloud Networks. Barcelona.
49. Wu, L., & Buyya, R. (2012). *Service level agreement (SLA) in utility cloud systems* (p. 25). IGI Global: Melbourne.
50. Gure, A. T., & Sharma, D. P. (2019). Assessment of knowledge sharing practices in higher learning institutions: A new exploratory framework–AT-DP KSPF. *The IUP Journal of Knowledge Management, 17*(4), 7–20.
51. Muda, J., Tumsa, S., Tuni, A., & Sharma, D. P. (2020). Cloud-enabled E-governance framework for citizen centric services. *Journal of Computer and Communications, 8*(7), 63–78.

52. Chakraborty, P. (2017). Simulation of reducing broadcasting protocol in ad hoc wireless networks. *International Journal of Scientific & Engineering Research, 8*(7), 295–301.
53. Wu, F., Wu, Q., & Tan, Y. (2015). Workflow scheduling in cloud: A survey. *Journal of Super Computer, 71*(9), 3373–3418.