Adaptive Resource Management and Provisioning in the Cloud Computing: A Survey of Definitions, Standards and Research Roadmaps

Amin Keshavarzi[1], Abolfazl Toroghi Haghighat[1] and Mahdi Bohlouli[2]

[1] Department of Computer & Information Technology Engineering, Qazvin branch, Islamic Azad University, Qazvin, Iran
[e-mail: keshavarzi@miau.ac.ir]
[e-mail: haghighat@qiau.ac.ir]

[2] Institute for Web Science and Technologies (WeST) University of Koblenz-Landau Universitaetstr. 1, 56070 Koblenz, Germany
[e-mail: mahdi. bohlouli@uni-koblenz.de]

*Corresponding author: Abolfazl Toroghi Haghighat

Received November 20, 2016; revised March 28, 2017; accepted April 18, 2017; published September 30, 2017

Abstract

The fact that cloud computing services have been proposed in recent years, organizations and individuals face with various challenges and problems such as how to migrate applications and software platforms into cloud or how to ensure security of migrated applications. This study reviews the current challenges and open issues in cloud computing, with the focus on autonomic resource management especially in federated clouds. In addition, this study provides recommendations and research roadmaps for scientific activities, as well as potential improvements in federated cloud computing. This survey study covers results achieved through 190 literatures including books, journal and conference papers, industrial reports, forums, and project reports. A solution is proposed for autonomic resource management in the federated clouds, using machine learning and statistical analysis in order to provide better and efficient resource management.

Keywords: Cloud Computing, Federated Clouds Resource Provisioning, Research Challenges in the Cloud, Service Level Agreement

A preliminary version of this paper appeared as an oral paper at the 7th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications 12-14 September 2013, Berlin, Germany. The published paper in this conference was 6 pages and consist of a very short version of this submission. The current submission is 21 pages.

https://doi.org/10.3837/tiis.2017.09.006 **ISSN : 1976-7277**

1. Introduction

Cloud computing is one the most important developments in IT in recent times. The main idea of utility computing in a similar direction to cloud computing was proposed in the 1960s. At that time, John McCarthy [1] said that computing will become a utility, like the telephone is today. The practical implementation of the idea was not feasible due to the lack of appropriate infrastructure. In 1999, Salesforce[1] launched its first cloud computing product. Afterwards, Amazon corporation offered Amazon Web Service (AWS) in 2002 [2] and took an important step in order to implement the idea of cloud computing. Google, Microsoft and others, then have offered cloud based products and services to the market. Currently, many companies provide cloud based services and many others such as Japan Ministry of Economy[2] are using cloud based services to support their customers.

Nowadays, there are many definitions of the term cloud computing [3], [4], [5]. In the following, we present one of the most scholarly used definitions. The National Institute of Standards and Technology (NIST) focused on the configurability, on demand accessibility and rapid re-source provisioning in the clouds. The NIST definition for cloud computing "is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models"[3].

Public, private, hybrid and community clouds are four deployment models of cloud computing. Also essential characteristics of cloud computing are: On-demand Self-Service, A Broad network access, Rapid Elasticity, Resource pooling and Measured service [3].

Inter cloud or cloud of clouds defined a couple of years after advent of cloud computing. It has been defined by Global Inter-Cloud Technology Forum [6] as "a cloud model that, for the purpose of guaranteeing service quality, such as the performance and availability of each service, allows on-demand reassignment of resources and transfer of workload through an interworking of cloud systems of different cloud providers based on coordination of each consumers requirements for service quality with each providers Service Level Agreement (SLA) and use of standard interfaces."

SLA is a contract between cloud service user and provider consisting Quality of Services (QoS) criteria. The provider has to obey SLA guaranteed criteria and pay a penalty to the customer in case of any failure from the agreed SLA.

Inter cloud has several benefits for both customers and providers. Diverse geographical locations, better application resilience and avoidance of vendor lock-in are customers' benefits and moving workload to another provider which provides better SLAs to customers and leasing resources from other providers when workload increases are listed as providers' benefits [7]. Federations and multi clouds are two type of inter clouds [7]. In federation, a set of providers voluntarily interconnect their infrastructure and sharing resources between each other. In multi cloud a customer or an application broker use multiple independent clouds.

[1] What is Cloud Computing? - Salesforce UK. http://www.salesforce.com/uk/cloudcomputing/. Accessed 14 Oct 2016

[2] Case Studies in Cloud Computing. David Cearley and Gene Phifer. http://docplayer.net/12763747-Case-studies-in-cloud-computing-david-cearley-and-gene-phifer.html. Accessed 14 Oct 2016

Hybrid cloud, composition of different cloud infrastructure for example a public cloud and a private cloud, is a type of multi cloud [7].

There are several federate clouds definitions in the literature such as "a federated cloud (also called cloud federation) is the deployment and management of multiple external and internal cloud computing services to match business needs. A federation is the union of several smaller parts that perform a common action”[3]. Garcia et. al. reviewed open challenges in cloud federation as lack of a unified cloud interface, porting one VM to another provider with different hypervisor and authentication and authorization, via a federated identity management system.Discovering resources and capabilities that each of providers that exist in federation offers, lack of a unified accounting and billing method in federation are another reviewed challenges [8].

The rest of the paper is sketched as follows. Section 2 is about autonomic resource management. Key challenges and roadmaps are described in the section 3. Research challenges have been categorized into 5 directions with detailed literature review, related projects and research directions for each category. Section 4 is a bout a suggested solution approach to autonomic resource management in Federated Clouds. Finally, Section 5 concludes the paper.

2. Autonomic Resource Management in the Cloud

Auto scaling of resources is a vital requirement in cloud computing for both consumers and providers. There are two types of resource scaling: horizontal scaling and vertical scaling. In horizontal scalability more machines are being added when demand increases. Vertical scalability means the amount of physical resources required for computation is added to virtual nodes on-demand.

Also Resource Management (RM) is an important component of cloud computing. The RM of cloud environments aims to establish cloud essential characteristics, like those discussed in section 1. Due to the nature of clouds, RM should be done without involving users (provider and customer) intervention, thus, Autonomic Resource Management (ARM) systems are primarily used. The goal of ARM is to predict and prevent Service Level Agreement (SLA) violation, fault-tolerance, performance improvement, cost reduction and energy efficiency, especially in federated clouds. In order to attain these goals different Knowledge Management (KM) techniques such as case base reasoning, multi agent systems and fuzzy control are used [9]. In this section some of these techniques are studied with suggesting further challenges and detailed literature review.

Maurer et al. [10] used KM techniques in ARM to allocate physical resources to VMs for detecting SLA violations. They used Case Base Reasoning (CBR) for DM in a Monitoring, Analyzing, Planning and Executing (MAPE) cycle to automate SLA management. Low level metrics of the infrastructure’s resources such as free_disk or packets_sent are first monitored and then mapped (through the ‘MAP’ phase) to high level SLA parameters such as bandwidth, storage and response time. A following ‘Analysis’ phase is responsible for detecting SLA violations. In order to detect SLA violations, CBR was used. Each state of system is stored as a case, and each new case is compared with all cases in the Knowledge Base (KB). The most similar cases are retrieved and their corresponding knowledge is used for selecting best actions. As a result, achieved knowledge through new experiment is also being stored in the KB. The

[3] What is federated cloud (cloud federation)? - Definition from WhatIs.com. http://whatis.techtarget.com/definition/federated-cloud-cloud-federation. Accessed 14 Oct 2016

‘Planning’ phase maps Knowledge Based (KB) recommendations to Physical Machines (PM), which prevents oscillations and schedules execution of tasks through the “Execution” phase. Simultaneously, Emeakaroha et al. [11] mapped low level metrics such as downtime, uptime and available storage to the agreed SLA quantities using mapping rules that exist in the KB. In addition, the KB helps the ARM system to detect SLA violations before happening. In order to detect SLA violations, a predefined Threat Threshold (TT) is used. The calculated SLA values are compared to TT in order to react to violations before they happen. The system consists of an application deployment component, which manages the execution of user applications and an automatic VM deployment, to allocate required resources to requested services and arrange their deployment in VMs.

Buyya et al. [12] has proposed the following methods for decreasing energy consumption in cloud computing:

- Design new architectures for data center with respect to energy consumption.
- Develop efficient resource allocation mechanisms towards optimum energy consumption.
- Design and development of the software components for energy management in the clouds.

Energy efficiency is a criteria that can be consider in ARM. When a SLA violation is predicted, then the amount of resources should be increased or decreased relative to the requirements. Violation prevention, resource performance and cost should be considered together in resources increasing or decreasing. When the amount of resources is increased or decreased, some resources may be idle, effecting power consumption. In order to avoid this situation, VM migration can be used and PM with low load could be powered off.

To decrease the number of message exchanges for achieving an agreement on public criteria in SLA, Dual Agreement Protocol of Cloud-Computing (DAPCC) are used [13], in which the maximum allowed number of error components are tolerated within bounds. A Low Latency Fault Tolerance (LLFT) middleware [14] that uses leader/follower replication approach, makes distributed applications fault tolerant in a cloud environment. LLFT consists of Low Latency Messaging Protocol (LLMP), a Leader-Determined Membership Protocol (LDMP), and a Virtual Determiner Framework (VDF). LLMP is a reliable message delivery service; LDMP is a fast reconfiguration and recovery service for faulty replicas as well as joins or leaves; VDF ensures that the order of information is same in main version and backed up version. Higher replica consistency, application transparency and low end-to-end latency can be achieved in LLFT [14].

Islam et al. [15] used a feed-forward neural network with back-propagation learning algorithms for resource usage prediction. The current CPU usage is given to the network and it predicts the required amount of CPU resources required 12 minutes later. This duration of 12 minutes has been selected due to the required time of installing a VM in a cloud computing environment which is between 5 and 15 minutes. The number of hidden layers in the neural network is 1 and the number of neurons in the hidden layer is 7, chosen by empirical results. The effect of using a sliding window in prediction precision has been investigated in this work. At the time of using sliding window, input of the network is a vector of values instead of one value. Linear regression is another technique that has been used to make predictions. This algorithm has been used to predict future value of resource usage. In this technique, if input variables are as $x = [x_1, x_2, \dots, x_n]$, the output y is calculated as equation 1 and the value of β is calculated during training process. This technique also has been investigated in two modes with and without sliding window.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i \, x_i \qquad (1)$$

Another ARM method was proposed by Rolim et al. [16] using Multi Adaptive Neuro-Fuzzy Inference Systems (MANFIS) in order to predict and provision required resources. Additionally, the use of Multi-Layer Perceptron (MLP) was investigated and compared with MANFIS. In the MLP, determining characteristics such as the number of layers, the number of neurons in each layer, training algorithm, activation function of each layer, learning rate and a number of learning steps is a tedious task. Combining fuzzy system and neural network and constructing a hybrid network, is a solution for this problem. A few hybrid networks could be used, but given the required accuracy needed by ARM, the ANFIS model was selected. ANFIS is a network with one output, so the MANFIS network is an extension of this. The inputs to the two networks are as follow [16]:

- The amount of requests at the amount of simultaneous requests that the environment receives
- Number of seconds for each request to be completed
- The amount of RAM consumed by the requests in a certain snapshot of time the processor
- Number of Megabytes consumption from the requests

The number of cores necessary to allocate a new virtual instance or update one single running instance, amount of memory necessary to carry out the current workload from the environment, amount of storage needed for the environment to maintain the current workload are the predicted outputs. Also, a binary variable as an output predicts SLA violation [16].

In the MLP, using a 10-fold cross validation and constructing a confusion matrix, the best topology of a network included: 6 input nodes, 2 hidden layers with 7 and 12 nodes respectively, hyperbolic tangent activation function, and 4 output nodes. In the MANFIS network, configuration is done automatically by internal mathematic functions and based on input data, but the network designer selects the best manner of input data partitioning using subtractive clustering. This method does not produce many rules that improves its simplicity [16].

The EU funded research project, CELAR [17], focuses on autonomic and elastic resource allocation to cloud applications in order to increase performance and resource utilization and reduce administrative costs. In this project, resource allocation is done through intelligent decision making by performance metrics and cost evaluation through the scalable monitoring system and elastic modeling. An on-line gaming and scientific computing scenarios have been defined in the project to validate the usability and the significance of the project outcomes.

The Adaptable Management Architecture [18] project provides an adaptable service for changes in RM, in order to support elastic services, e.g. the services that run on cloud resources. The project aims to predict changes in deployed services due to the lead time needed for Virtual Machines (VM) addition/removal. The target platform of this project is an agent based architecture where agents receive an event and do further processing. The processing output can be a new event on the outgoing stream or modification of the state of the managed system. The system consists of the worker component that contains the service application and the delegator component that acts as a load balancer for receiving incoming requests and distributing across workers. This system follows the logic of Monitor, Analyze, Plan and Execute (MAPE) loop. The monitoring and analyzing are done by sensors and the planning phase by actors. The effecter agents execute actions. Agents are organized as a hierarchy and construct a management goal graph. At a higher level, a graph called elastic management graph is constructed and used for rapid elasticity of cloud environment according user defined policy. For each manageability capability, an effective agent is created.

The Cloud and Autonomic Computing Center (CAC) at the University of Mississippi State started a project called “a model-based framework for autonomic performance management of cloud computing systems” [19]. The main goal of this project is to develop a model based self-management system that meets the overall performance of the system that is mentioned in SLA consisting of QoS, reliability and availability. The project consist of: (1) advanced SLA definition between cloud provider and application provider, (2) developing a monitoring tool that consumes minimum resource and has minimum latency while simultaneously monitoring different aspects such as application performance and system health, (3) developing reliability metrics and models, and (4) solving the dynamic management problem in cloud computing.

According to the NIST’s definition of cloud computing, the rapid elasticity is a cloud essential characteristic that means provisioning unlimited resources for users with limited resources on the provider’s side. To establish full elasticity, complicated cloud infrastructure such as inter cloud is required. Resource management is one of the important requirement for complicated environments such as inter clouds [20]. Cloud broker is a solution to address this complexity. In [21], a control based autonomic approach uses an interaction balance based algorithm for performance management of a distributed cloud broker. In the proposed algorithm, computational resources are allocated to all service providers by interaction balance algorithm. Following this, the service level controller is responsible for utilizing allocated resources for each service and maintaining respective service SLAs. The global cost function is minimized by independently optimizing local cost function of each service in a cooperation manner.

Carella et al. [22] developed the Elasticity Engine of a cloud broker, FOKUS, that deployed and evaluated within BonFIRE, European large-scale, multi-site cloud experimental facility. BonFIRE supports three approaches for elasticity: manual, programmed and managed. In manual method, users can create or delete resources via a web based portal. Programmed elasticity is supported via RM monitoring and Open Cloud Computing Interface (OCCI) APIs. The last approach called Elasticity as a Service (EaaS), done by BonFIRE Elasticity Engine (EE). The main task of EE is automatically increasing or decreasing the computing resources in a running experiment. The dynamic allocation of resources is done automatically based on the rules expressed by the experimenter. Javadi et al. [23] proposed a failure aware provisioning algorithm for hybrid cloud platforms. Implementing the workload model and failure correlations, user’s request are redirected to appropriate cloud provider. In Lucas-Simarro et. al [24], a broker architecture for deployment of services across multi-cloud providers based on different optimization criteria, different user constraints and different environmental conditions was developed.

In summary, ARM in the cloud computing is a multi-objective problem. Prediction and prevention of SLA violation, increasing energy efficiency, cost reduction, performance improvement, fault tolerance, and guaranteed security, especially in federated clouds, are major challenges in this area. Developing a fuzzy control system and formulating human knowledge can better the performance of the ARM system. Integration of rule base systems, mapping low level metrics to high level parameters in SLA, considering multiple resources concurrently, are further open issues in this regard.

3. Research Challenges and Road Maps in further Areas associated with Cloud Computing

This section summarizes open research issues and challenges in areas connected to cloud computing, which are organized in the four different topics. Section 3.1 covers security and trustworthy in the cloud, Section 3.2 covers cloud-based development and benchmarking of the systems, Section 3.3 covers big data and cloud computing, and section 3.4 covers social and mobile-clouds.

3.1 Security and Trustworthy in the Cloud

In cloud computing, users outsource their computation and data [25] , relinquishing control with no knowledge of where the services and data reside. As such, data security is paramount. There are some techniques in order to address security and privacy concerns, such as access control, cryptography, and integrity checking. However, these techniques have some difficulties in novel cloud infrastructures and should be adopted to be feasible for cloud environments.

Pearson et al. [26] defined privacy as “for organizations, privacy entails the application of laws, policies, standards and processes by which Personally Identifiable Information (PII) of individuals is managed”. Privacy is a main challenge for cloud computing due to the shared environment, remote access and data processing, as well as combined service and information flow across provider boundaries [27]. Identity and access management is one of the mechanisms that is used to preserve privacy. In access management, access to various resources and services are controlled through mechanisms such as authentication and authorization. In authentication, identity of an applicant is verified and his/her access level is controlled in authorization step. It requires an access control framework that integrates access policies. Security Assertion Markup Language (SAML), Extensible Access Control Markup Language (XACML) and web service standards are various frameworks for cross-domain access specification and verification [28]. Identity federation is the way that organizations and public cloud providers trust each other and share digital identity and attributes and support single sign-on [29]. SAML and OpenID standards can be used to accomplish identity federation [28]. XACML uses an XML-base language to define policies and decision making tasks. Identity Management (idM) method has been used for access control to cloud services [30].

Traditional integrity checking techniques, such as hashing, cannot be applied to data and computational integrity in cloud computing, because hashing of such huge volumes of data through the internet is not feasible. Provable Data Possession (PDP) can support integrity checks in cloud computing. In order to preserve confidentiality, traditional techniques cannot be applied to cloud computing because of existing threats inside the systems. In the confidentiality method, privacy of users is protected from others and authorized users can only access to the environment. Virtualization helps to separate the shared environment of users, helping address confidentiality issues. Subashini and Kavith [30] used protocols such as Secure Sockets Layer (SSL) or Internet Protocol Security (IPSec) in encryption phase to prevent security threats in clouds. There are two main types of threats for cloud service availability [31]: (1) flooding attack via bandwidth starvation and (2) Fraudulent Resource Consumption (FRC) attack. Zissis and Lekkas [32] used Trusted Third Party (TTP) as an entity, which both parties trust on a third protocol, to provide a secure interaction. In this model, a set of TTP create a Public Key Infrastructure (PKI) and are responsible for ensuring the security of the cloud environment. PKI with a directory of certificates are used in order to provide an access control. PKI is also used with Single-Sign-On (SSO), which the user does not need to sign in repetitive actions.

MASSIF [33] is the European funded project in the field of SIEM (Security Information and

Event Management), which researches management of security information and events in service infrastructure. SIEM consists of the log management with Security Information Management (SIM) and real-time event management with Security Event Management (SEM). In this project, a SIEM framework is developed that supports intelligent, scalable, resilient, and multi-level/multi domain security event processing and predictive security monitoring.

The Cryptographic Cloud Storage Service (C2S) [34] is a Microsoft research project for building a secure cloud based storage service on top of public clouds in order to eliminate cloud customers' concerns, whenever they do not trust on the provider. Confidentiality and integrity are two impediments to customers adopting cloud computing specialty when they outsource personally identifiable information. the Cryptographic Cloud Storage Service has two different applications for customer and enterprise sides. Consumer side client consists of 4 core components: Data Processor (DP), Data Verifier (DV), Token Generator (TG) and Credential Generator (CG). In first-time execution of the application, it generates a cryptographic key which is called master key and is stored locally in consumers' machine. The DP attaches some metadata to data and encrypts before uploading to the cloud. The DV checks the integrity of the data using a master key. Whenever a consumer wants to retrieve his/her data, the TG produces a token that is sent to cloud provider in order to retrieve encrypted files. Whenever a consumer wants to share the data with other users, his/her application generates a token and a credential to be sent to target users. The token should be sent to the provider in order to share the cloud hosted and encrypted data with third parties. Generally, the enterprise side application is similar to the consumer side. The main difference is that employees of an enterprise should receive a credential from CG and use it for their transactions. For security purpose these components can be implemented as open source.

In summary, users outsource data into cloud hosted resources, which can be manipulated by internal/external people, thus robust information integrity checking and confidentiality mechanisms is required in clouds [35]. In addition, rapid elasticity of the clouds needs to distinguish denial of service attacks from the further resource demands [36]. Strategy makers and scientists should also focus on the definition of proper security standards in order to support interoperability and construct between multiple public, private, community as well as federated clouds.

3.2 Cloud Development and Benchmarking of the Systems

With increased use of cloud computing, the need of cloud benchmarks is also increased. Cloud benchmarks help in assessing cloud performance and comparing different cloud services and providers. In addition to the cloud adoption concerns, variations in definitions of cloud standards and tools is inevitable.

Yahoo! Cloud Serving Benchmark (YCSB) [37] is the evaluation and comparing PNUTS system, a parallel and distributed database system for Yahoo!'s web applications, with other cloud based Database (DB) systems. This tool is used in order to compare performance of Cassandra, HBase, Yahoo!'s PNUTS, and shared MySQL. This tool helps to understand which workload is suited for which system. Performance, elasticity, availability and replication are some of the indicators used in this tool.

The tool is limited to the Cloud DB systems. Schad et al. [38] used CPU, I/O and network variants as indicators to evaluate the efficiency of Amazon EC2 and compare it with local cluster systems. They used micro-benchmarks to measure performance variance in CPU, I/O, network resources, and used a multi-node MapReduce application to quantify the impact on real data intensive applications. They show that variance of EC2 is high.

Kossmann et al. [39] investigate how cloud computing platform meet cloud promises. The focus was online transaction processing workloads (OLTP) in public clouds. They compared performance of different PaaS provider in 2010. Cost and elasticity are two measurements in this comparison. Cost can be several orders between different providers and also in one provider depending on number of concurrent users. Also a fully elastic service should never hit its limits. All experiments in this study were done with the TPC-W benchmark [38].

Another benchmarking suite is CloudGauge, a dynamic and experimental cloud benchmarking suite developed by El-Refaey et al. [40]. Using specified metrics and workload the main focus of the tool is on performance evaluation of virtual systems that can be used for performance models of virtual systems and clouds. Due to a wide variety of cloud services and providers, an integration suite for cloud hosted tools and package is fundamental basic in cloud development models. Some of the key capabilities of this framework include dynamically injecting workload, configuring and customizing virtualization layer, measuring the performance and communicating measurements, statistic's report and intelligent load balancing [39].

Cloud service brokerage is an intermediate between providers and consumers [41] which helps users in the allocation and management of service consumptions. Grozev and Buyya [7] defined inter cloud requirements as: data location awareness, geo-location awareness, pricing awareness, legislation/policy and local resources. Data location awareness is important for job-based and data-centric applications, geo-location for compute-intensive and data-centric applications, pricing for job based, compute intensive and data centric, legislation/policy for job based and data centric and local resources for job based, compute intensive and data centric [7].

Nair et al. [42] proposed a secure cloud broker architecture that can be used to implement brokering of multiple providers and provide an SLA-based pricing model to users. This broker consists of components such as data confidentiality, scaling resources, identity and access management, risk analysis, cloud bursting, secure consumers' data transferring, SLA management.

A cloud marketplace is market oriented cloud computing that consists of a wide variety of cloud resources and services from multiple vendors and platforms. The Market Maker as an interactive front-end is responsible for providing the best services according to the users' budget, constraints and desired QoS. The Cloudbus Toolkit [43] is an option that consists of a collection of components and can be used as a comprehensive simulation environment for cloud Marketplace.

The Cloudbus provides a service brokering infrastructure that users can deploy their applications in the cloud. The Cloudbus middleware consists of components such as Aneka, Workflow Engine, Broker, Market Maker/Meta-broker, InterGrid, MetaCDN, Energy Efficient Computing, and ClouSim [44]. The Aneka is a Platform as a Service for developing and deploying applications in the cloud. The Broker component is used to access physical and virtual resources. The workflow engine supports cloud users in representing their applications in the form of workflows. The Market Maker is an interface between consumers and providers. The responsibility of InterCloud is providing an interconnection between the islands of clouds. The MetaCDN exploits different storage resources from different IaaS providers and creates an overlay network that provides a Content Delivery Network (CDN). In Energy efficient Computing component various algorithms and techniques are deployed in order to decrease energy consumption.

The Contrail project [45] aims at the integration of heterogeneous cloud resources with vertical and horizontal platforms. The vertical integration provides a unified platform for different kinds of resources and horizontal integration covers different cloud providers. The interface layer provides mechanisms for users and other Contrail components to interact with federation. It contains both HTTP and CLI interfaces. The core layer is responsible for functional requirements such as application life cycle management and non-functional requirements such as security. The adapter layer is to operate information retrieval and operate on different cloud providers, which also copes with heterogeneity of providers. In addition, the user identity module is responsible for binding identities between providers and users. The local identities are stored in state module. Authentication and authorizing mechanisms are used for isolation and data integrity guarantees. The Federate Runtime manager (FRM) maps submitted applications with federation resources by using a set of heuristics that consider different economical and performance aspects. The FRM collects information from the status module and manages application life cycle. The Image Manager (IM) module decides when image to be packed into Open Virtual Format (OVF) archive and when referenced within the OVF files by means of the URI. The Adapter layer sends monitored information to the Provider Watcher component.

The SLA organizer component works at the federation level and composed of SLA coordination, SLA negotiation, and SLA template repository. The SLA coordination checks user defined SLA and the status of currently running applications. Upon the SLA violation detection, it logs the event and investigates the status of all applications and resources for providing consistent action, in order to prevent SLA violation. The SLA negotiation module compares protocols to match user needs with providers. The last component of SLA organizer is an SLA template repository. It contains SLA templates of various providers. All information that is needed during Contrail work are aggregated and stored in the State module.

Another benchmarking project titled RESERVOIR [46] (Resources and Services Virtualization without Barriers) aims to introduce a next generation federated infrastructure of various infrastructure providers across different geographies. It reduces investment and operational costs and increases energy efficiency, elasticity, ensures QoS, and guarantees security. In RESERVOIR, two different virtualization technologies of virtual machines (VMs) and Virtual Java Service Containers (VJSCs) are used in order to develop an abstract layer that is not tied to any environments and help to establish federated cloud [47]. The RESERVOIR project has a layered architecture, where for each layer, software and specifications are available.

The highest abstract layer is service manager (SM). The next layer is the Virtual Execution Environment Manager (VEEM) that is responsible for optimal placement of VEEs into VEE hosts regarding to constraints determined by SM. The last abstract layer in RESIVOIR is Virtual Execution Environment Host (VEEH). The responsibility of VEEH is basic control and monitoring of VEEs and their resources such as creating a VEE, monitoring VEE, migrating a VEE, and allocating additional resource [46].

Cloud interoperability and data and VM exchange between different cloud vendors' scientific efforts in order to improve the proficiency and develop the cost algorithms and measurements. In this regard, scientific and technical people should define new standards and tools. Cloud bursting as well as cloud broker strategies need further efforts and improvements by scientists.

3.3 Big Data and the Cloud

The world of information is doubling every two years [48]. International Data Corporation (IDC) says that digital universe will be 40,000 Exabyte in 2020 [49]. There are three different dimensions for big data, namely “Volume”, “Velocity” and “Variety” [50].

Chaudhuri et al. [51] defined six challenges for big data in the cloud: data privacy; data accuracy; data exploration to enable deep analytics; enterprise data enrichment with web; social media; query optimization. Developing a novel cloud based platform for big data management systems is at the forefront of topics being researched. Having a decision support system (DSS) [52] and applying machine learning for large scale, disparate data sources, that can be delivered as a service to consumers is another main research challenges in this area. For this purpose, DSS should be developed in a component based approach that can be characterized by reusability, substitutability, extensibility, scalability, customizability, reliability, low cost of ownership, and economy of scale [53].

Web applications have different storage requirements that traditional Relational Database Management Systems (RDBMS) do not fulfill. Strong consistency and integrity are not vital for web applications instead low latency, distribution and scalability are important. According to CAP theorem [54], a distributed system can satisfy only two of the following three requirements: Consistency, Availability and Partition tolerance. The NoSQL (Not Only SQL or Not relational) data stores have any characteristics such as schema free, easy replication support, simple API, eventually consistent/BASE (not ACID), a huge amount of data. BASE means (basically available, soft state, eventually consistent) and ACID means (Atomicity, Consistency, Isolation, Durability). Cattell et al. [55] compared some NoSQL data stores. His focus is on scalability of NoSQL data store versus traditional database systems. He compares new database systems on their data model, consistency mechanisms, storage mechanisms, durability guarantees, availability, query support, and other dimensions.

Cloud storage for Big data has some challenges such as security, control, performance, support, configurability and vendor lock-in.

A framework has been offered to enable the execution of large scale data mining applications on top of cloud computing services [56]. This framework has been developed using Windows Azure and consists of 5 components: a set of binary and text data containers, a task queue, a task status table, a pool of workers, and a website. The data containers are used to store data to be mined and the results of data mining tasks. The task queue contains the data mining tasks to be executed, while the next component keeps information about the status of all tasks. Workers are responsible of executing the data mining tasks submitted by the users and web site is a dashboard for managing and monitoring tasks.

Demirkan et al. [53] proposed a conceptual model for evaluation of Service-Oriented Decision Support Systems (SODSS). The SODSS consists of 4 major components: information technology as enabler, process as beneficiary, people as user and organization as facilitator. Also, operational systems, data warehouses, online analytic processing and end-user’s components can be delivered as a service to users. In Data as a Service (DaaS) [57], data can reside anywhere and business processes have 24 hour access. Information as a service (IaaS) provides a fast access to information across a business for users and processes. IaaS offers an integrated platform of information that provides a set of interfaces and standards to access data easily. In Analytics as a Service (AaaS), cloud computing is used for analytic work which is also called Agile Analytics, delivering scalability and cost reduction benefits. Encrypted data, analyzing and using data mining algorithms for big data are challenges of Cloud Analytics [50].

Map-Reduce [58] is a parallel programming model that can process large scale data volumes. In this model users define map and reduce functions and then the underlying runtime system dose computation automatically and in parallel across large scale clusters of machines [59]. Specifically, the model has two functions: map and reduce. The map function is responsible for processing a key/value pair and produces a set of intermediate key/vale pairs. The reduce function is used for merging all intermediate values with the same key. This model also facilitates programming on multicore chips [57]. Hadoop [60] is a framework for distributed large data set computation across a cluster of machines by using map-reduce programming model. This project includes four modules. The first module is the Hadoop common, which supports other Hadoop modules. The second module is the Hadoop Distributed File System (HDFS) that is a distributed file system for data accessing. The Hadoop YARN is the third module that is responsible for resource management. The last module is Hadoop MapReduce that is used for parallel large data set processing.

Apache Mahout [61] is also an open source tool designed to build scalable machine learning algorithms. The initial version of the Mahout was supporting only few machine learning algorithms such as clustering, categorization, collaborative filtering (CF), and evolutionary programming. It has been extended with supporting new algorithms. **Table 1** provides a list of current supported algorithms in the Apache Mahout.

Table 1. Selected supported algorithms in the Apache Mahout [4]

Algorithm	Brief description	Use case
Logistic Regression, solved by Stochastic Gradient Descent (SGD)	Blazing fast, simple, sequential classifier capable of online learning in demanding environments	Recommend ads to users, classify text into categories
Hidden Markov Models (HMM)	Sequential and parallel implementations of the classic classification algorithm designed to model real-world processes when the underlying generation process is unknown	Part-of-speech tagging of text; speech recognition
Singular Value Decomposition (SVD)	Designed to reduce noise in large matrices, thereby making them smaller and easier to work on	As a precursor to clustering, recommenders, and classification to do feature selection automatically
Dirichlet Clustering	Model-based approach to clustering that determines membership based on whether the data fits into the underlying model	Useful when the data has overlap or hierarchy
Spectral Clustering	Family of similar approaches that use a graph-based approach to determine cluster membership	Like all clustering lgorithms, useful for exploring large, unseen data sets
Minhash Clustering	Uses a hashing strategy to group similar items together, thereby producing clusters	Same as other clustering approaches
Numerous recommender improvements	Distributed co-occurrence, SVD, Alternating Least-Squares	Dating sites, e-commerce, movie or book recommendations
Collocations	Map-Reduce enabled collocation implementation	Finding statistically interesting phrases in text

In summary, new mechanisms and algorithms are required in this area in order to fully support cloud and big data integration. These include, but not limited to:

- Scalable decision making algorithms for big data analytics in through the cloud.
- Scalable data management and machine learning techniques in the clouds.
- Adapt traditional machine learning, data mining, decision making and artificial intelligence methods and algorithms with cloud hosted scalable algorithms and methodologies.

[4] Apache Mahout: Scalable machine learning and data mining. https://mahout.apache.org/users/basics/algorithms.html. Accessed 14 Oct 2016

3.4 Social and Mobile Clouds

Mobile Cloud Computing (MCC) [62] [63] integrates cloud and mobile computing in order to use cloud properties, to overcome mobile computing performance, environment and security obstacles [64]. Integrating social networks with cloud computing raises two issues. The first is to overcome privacy concerns of cloud computing, and the second is scalability concerns and requirements of social networks. Li et al. [65] have demonstrated a social service computing ecosystem. The ecosystem has five basic elements including: service providers and consumers, services, local services, physical things, and cloud computing platforms. In this ecosystem, there are four types of networks , social networks between service providers and service consumers, service networks, cloud computing networks and physical thing networks. In this ecosystem, computer systems and social individuals are connected together through social networks. The main task of social service computing is service classification, clustering, migration, recommendation, composition, as well as services discovery and publishing in social context [65].

Chard et al. [66] proposed a social cloud system that users can limit the resources shared with different groups of friends based on type according to social network friend category. In a social cloud, resource sharing is done based on users' online relationships in social networks. In a social network each group can establish a social cloud with specific policies and market metaphors.

Mobile devices such as smart phones, PDAs and tablets are good candidates for thin clients. Dinh et al. [64] investigated mobile cloud computing architecture, applications and approaches. Due to the limitations of mobile devices, migration of the processing and storage of mobile devices in the cloud can be beneficial for end users and providers. In this architecture, mobile devices connect to the network operator and use services as AAA methodology (Authentication, Authorization and Accounting) based on subscribers' data stored in DBs. Satria et. al. [67] proposed two different recovery schemes for overloaded Mobile Edge Computing (MEC). One scheme is where an overloaded MEC offloads its work to available neighboring MECs and another is for situations when there is no available neighboring MEC within transfer range.

Mobile Cloud Networking (MCN)[5] is a project that is supported by EU FP7 framework. This project as aims to deliver an atomic service that is a combination of storage, computing and mobile network. Extended cloud services to mobile end users, offering a mobile architecture that supports cloud computing, define a new business actor that is called mobile cloud provider, use the concept of an end to end mobile cloud for novel applications are other characteristics of this service.

In order to establish mobile cloud network two scenarios have been regarded. The first one uses cloud computing as an infrastructure for future mobile network and decreases capital expenditure or operating expenditures. The second is more visionary; that the cloud provider offers commercial services to the users. In this framework two types of data centers have been regarded: macro and micro data center. Macro datacenters are large scale computing farms in selected locations. The micro data centers are medium to small scale server clusters that place in a certain geographical area. By the invention of MCN providers, organization can sign a contract with a MCN provider instead of one or more mobile operator, and use its services such as mobile communication, computation and storage. The MCN provider itself sign contracts with micro and macro datacenters and cloud ready mobile networks in order to provide services by using Radio Access Networks (RAN) virtualization, mobile Evolved

[5] MCN. http://www.mobile-cloud-networking.eu/site/. Accessed 14 Oct 2016

Packet Core as a Service (EPCaaS) that is on demand development of mobile Evolved Packet Core (EPC) instances on top of IaaS on micro and macro datacenters, IP Multimedia Subsystem as a service (IMSaaS) that is on demand development of IP Multimedia Subsystem (IMS) instances on top of IaaS on micro and macro datacenters that is used for voice/video services, on demand development of content / storage / application distribution services, End-to-End MCN Service Orchestration and XaaS in MCN.

The social cloud project [68] is a project at Karlsruhe Service Research Institute which aims to share underutilized resources and services in a social network between a social network members.

As a summary, we argue following points as open research issues in this area:

- A need of proper market protocols for social clouds.
- Algorithms for mobile device resource sharing based on trust between connections in the social networks
- Optimization of unreliable mobile networks in mobile clouds

4. A Suggested Solution Approach to ARM in Federated Clouds

An IaaS required resource for running a workload should be provisioned autonomously. In this direction two goals should be regarded: (1) resources should be provided in **a way,** such that SLA is not being violated and provider shouldn't pay penalties, and (2) resource provisioning algorithms shouldn't result in wasting resources. An adaptive resource management (ARM) in the cloud aims at promoting cost reduction and energy efficiency. The ARM systems are free from administration and maintenance. In ARM, customers get resources when needed and release them when resources are not required anymore without human intervention. In this way, cost decreases for both customers and cloud providers, since customers pay costs based on usage. Self-management is vital for autonomic system. Self-configuration, self-optimization, self-healing and self-protection are four aspects of self-management [69] that need to be taken into account. An autonomic system should change its operation with temporal and spatial changes in context environment. According to NIST, rapid elasticity is an essential characteristic of cloud computing which should be supported through ARM. With respect to rapid elasticity. In an autonomic system, both reactive and proactive methods can be used. Using the reactive method, SLA violation happens and then system reacts based on the violation and provider should pay penalties. In the proactive method, a system predicts resource usage before any violation happens and prevents SLA violation. However, the prediction process is not easy and straightforward due to sudden changes in cloud workloads [11]. In addition, it deals with the quality issue and performance concerns.

The goal of this research was to focus on proactive methods. We have utilized Neural Networks to predict SLA violation. The neural network is a powerful method in order to deal with SLA violation prediction in autonomic systems [16]. The major part of neural network utilization is to provide learning to algorithms from our training data. The structure of network and quality of training data plays a key role in the performance of predictions [70].

In the proposed framework of this research, we use I-MAPE (Initial-MAPE provider), which is an extension of MAPE. As stated earlier, MAPE is Monitor, Analyze, Plan and Execute. In the MAPE cycle, the behavior of autonomic system is first monitored and parameters values are collected from the operating environment. These values are analyzed for prediction of the future state of system. In the plan phase, and an associated policy rules are designed for system based on earlier predictions. Finally, the devised plan is executed in an execution phase. Pre-processing in the initialization phase prepares historical data that exists in the KB. This is used for training and solving the neural network problems such as assessing the quality of training data. A few preprocessing techniques such as outlier detection, data association detection and sampling, have been investigated in this work to evaluate the impact of them in neural network performance. In order to use neural networks, it is important to define: input nodes, number of hidden layers, number of nodes in each hidden layer, output nodes, type of the network, training algorithm, number of epochs (i.e. each time the network is presented with a new input pattern), activation function of each layer, training data, test data and validation data.

Multi-layer feed-forward model and recurrent model are two neural network models that frequently have been used in prediction problems [70]. The architecture of neural network is simple recurrent neural network or Elman [71]. The recurrent neural networks have the ability to process temporal data and are being used in time series modeling. The Elman network is one of the simplest versions of recurrent neural network. The neural network considered in this work has one input layer, one hidden layer and one output layer. In all neural networks, one input layer and one output layer is used. Although, the number of hidden layers and the number of nodes in each hidden layer is an empirical task, but in most prediction problems only one hidden layer is used [70].

The number of neuron in hidden layer is specified using trial and error. The input is selected from the data set and output is predicted values. The inputs of network are response time values from WS-Dream data set. The last k values of response time series is input and the k+1 value is output. In order to eliminate the effect of oscillation in our prediction we use multi-step-ahead prediction. In multi-step-ahead prediction a network with a few output nodes is used. The number of output nodes is n and is determined experimentally. The back propagation algorithm is frequently used as a training algorithm in prediction problems [70]. Back propagation through time [72] is our selected training algorithm. This algorithm is fast, but the local optima problem is more significant than backpropagation algorithm [73].

The sigmoid activation function is used in both the hidden and output layer. This function is differentiable and is suitable to be used in BackPropagation Through Time (BPTT) training algorithm. The sigmoid activation function relation has been described in the Equation 2. The output of this function is between 0 and 1.

$$f(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

We use WS-DREAM [74] dataset in our approach . This dataset contains response times (s) and throughputs (kbps) of 142 users for 4532 web services on 64 different time slots. Because the activation function is sigmoid, the data set values first should be normalized. Equation (3) indicates the processing phase in this approach. The goal is to predict QoS values. Thus, denormalization is necessary in the end.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \qquad (3)$$

x_i is current value of x, $\min(x)$ is minimum value of x, $\max(x)$ is maximum value of x and z_i is normalized value. We calculate mean absolute error (MAE) and root mean squared error (RMSE) to evaluate the prediction quality of the proposed approach. These metrics are defined as follows:

$$MAE = \frac{\sum_i |\widehat{y_n} - y_n|}{N} \qquad (4)$$

$$RMSE = \sqrt{\frac{\sum_i(\widehat{y_n} - y_n)^2}{N}} \qquad (5)$$

where $\widehat{y_n}$ is the predicted QoS value of a QoS sequence y for the current time slot t_n, y_n denotes the actual QoS value of y for t_n, and N is the number of predicted QoS values.

Table 2 shows MAE and RMSE of proposed approach with different matrix density. In our experiments, matrix desnsity id defined as the density of training data set. For example if matrix density be 70% means 70 percent of QoS data set is used for training neural network and the remaining 30 percent is used for test. The result of our experiments to predict the response time of WS-DREAM dataset has been depicted in Table 2. As shown in this table our approach has a good prediction accuracy. The best result of MAE and RMSE are achieved using 60% and 70% respectively.

Table 2. The experiments results of proposed approach

Matrix Density / Metrics	10%	20%	30%	40%	50%	60%	70%	80%	90%
MAE	0.515	0.57	0.615	0.683	0.564	0.493	0.514	0.683	1.01
RMSE	0.254	0.26	0.266	0.282	0.203	0.157	0.147	0.16	0.177

5. Conclusion

The basic idea of cloud computing has been proposed in 1966 by McCarthy. Salesforce.com has announced its first cloud hosted service in 1999. Public, Private, Hybrid and Community clouds are deployment models of cloud computing. Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) are service models of cloud computing. There are over 65 definitions for cloud computing published in the literature. After reviewing most of the definitions, the authors of the paper defined cloud computing as transparent, scalable and easily accessible computing system with various service levels in public and/or private forms which provides on demand access to a virtualized pool of resources, and targets cost reductions in computing and improvements in software deployment process and IT solutions.

Research challenges and directions have been categorized into 4 groups, namely security and trustworthy, cloud development and benchmarking, big data technologies and cloud computing as well as social and mobile clouds. The Autonomic Resource Management (ARM) in the clouds is separately reviewed in this paper. This is the most important issue in this regard, especially towards virtualization technology, fault tolerance, SLA violation and feder-ated clouds. Security is the most common challenge of cloud computing.

An improved solution approach to ARM in federated clouds is given in this paper which uses machine learning and statistical analysis in order to provide better performance and prediction quality. This paper represents the outcome of literature review of 190 resources including books, journal and conference papers, project reports and deliverables, European Commission roadmaps and calls for proposals, online weblogs, white papers, business reviews, company websites, profiles and offered services by cloud providers. As a result, it represents today’s state-of-the-art from a wide perspective.

References

[1] H. Abelson, *Architects of the information society: Thirty-five years of the laboratory for computer science at MIT*. MIT Press, 1999.

[2] M. Siegel and F. Gibbons, "Amazon enters the Cloud computing business," *CasePublisher Com May*, vol. 20, 2008.

[3] P. Mell and T. Grance, "The NIST definition of cloud computing," 2011.

[4] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Gener. Comput. Syst.*, vol. 25, no. 6, pp. 599–616, 2009. Article (CrossRef Link)

[5] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *Proc. of Grid Computing Environments Workshop, 2008. GCE'08*, pp. 1–10, 2008. Article (CrossRef Link)

[6] Global Inter-Cloud Technology Forum, "Use Cases and Functional Requirements for Inter-Cloud Computing. Technical Report."

[7] N. Grozev and R. Buyya, "Inter-Cloud architectures and application brokering: taxonomy and survey," *Softw. Pract. Exp.*, vol. 44, no. 3, pp. 369–390, 2014. Article (CrossRef Link)

[8] Á. L. García, E. F. del Castillo, and P. O. Fernández, "Standards for enabling heterogeneous IaaS cloud federations," *Comput. Stand. Interfaces*, vol. 47, pp. 19–23, 2016. Article (CrossRef Link)

[9] A. Keshavarzi, A. T. Haghighat, and M. Bohlouli, "Research challenges and prospective business impacts of cloud computing: A survey," in *Proc. of Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2013 IEEE 7th International Conference on*, vol. 2, pp. 731–736, 2013. Article (CrossRef Link)

[10] M. Maurer, I. Brandic, and R. Sakellariou, "Simulating autonomic SLA enactment in clouds using case based reasoning," in *Proc. of Towards a Service-Based Internet*, Springer, pp. 25–36, 2010. Article (CrossRef Link)

[11] V. C. Emeakaroha, M. A. Netto, R. N. Calheiros, I. Brandic, R. Buyya, and C. A. De Rose, "Towards autonomic detection of SLA violations in Cloud infrastructures," *Future Gener. Comput. Syst.*, vol. 28, no. 7, pp. 1017–1029, 2012. Article (CrossRef Link)

[12] R. Buyya, A. Beloglazov, and J. Abawajy, "Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges," *ArXiv Prepr. ArXiv10060308*, 2010.

[13] S.-S. Wang, K.-Q. Yan, and S.-C. Wang, "Achieving efficient agreement within a dual-failure cloud-computing environment," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 906–915, 2011. Article (CrossRef Link)

[14] W. Zhao, P. M. Melliar-Smith, and L. E. Moser, "Fault tolerance middleware for cloud computing," in *Proc. of Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, pp. 67–74, 2010. Article (CrossRef Link)

[15] S. Islam, J. Keung, K. Lee, and A. Liu, "Empirical prediction models for adaptive resource provisioning in the cloud," *Future Gener. Comput. Syst.*, vol. 28, no. 1, pp. 155–162, 2012. Article (CrossRef Link)

[16] C. O. Rolim, F. Schubert, A. G. Rossetto, V. R. Leithardt, C. F. Geyer, and C. Westphall, "Comparison of a Multi output Adaptative Neuro-Fuzzy Inference System (MANFIS) and Multi Layer Perceptron (MLP) in Cloud Computing Provisioning," in *Proc. of 29th Brazilian Symposium on Computer Networks and Distributed Systems, Paris*, 2012.

[17] G. Copil *et al.*, "On controlling elasticity of cloud applications in celar," *Emerg. Res. Cloud Distrib. Comput. Syst.*, p. 222, 2015. Article (CrossRef Link)

[18] P. Martin, A. Brown, W. Powley, and J. L. Vazquez-Poletti, "Autonomic management of elastic services in the cloud," in *Proc. of Computers and Communications (ISCC), 2011 IEEE Symposium on*, pp. 135–140, 2011. Article (CrossRef Link)

[19] S. Srivastava, R. Mehrotra, I. Banicescu, and S. Abdelwahed, "A Model-Based Framework for Autonomic Performance Management of Cloud Computing Systems."

[20] O. Rogers and D. Cliff, “A financial brokerage model for cloud computing,” *J. Cloud Comput.*, vol. 1, no. 1, pp. 1–12, 2012. Article (CrossRef Link)

[21] R. Mehrotra, S. Srivastava, I. Banicescu, and S. Abdelwahed, “Towards an autonomic performance management approach for a cloud broker environment using a decomposition–coordination based methodology,” in *Proc. of Future Gener. Comput. Syst.*, vol. 54, pp. 195–205, 2016. Article (CrossRef Link)

[22] G. Carella, T. Magedanz, K. Campowsky, and F. Schreiner, “Elasticity as a service for federated cloud testbeds,” in *Proc. of Communications Workshops (ICC), 2013 IEEE International Conference on*,pp. 256–260, 2013. Article (CrossRef Link)

[23] B. Javadi, J. Abawajy, and R. Buyya, “Failure-aware resource provisioning for hybrid Cloud infrastructure,” *J. Parallel Distrib. Comput.*, vol. 72, no. 10, pp. 1318–1331, 2012. Article (CrossRef Link)

[24] J. L. Lucas-Simarro, R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, “Scheduling strategies for optimal service deployment across multiple clouds,” in *Proc. of Future Gener. Comput. Syst.*, vol. 29, no. 6, pp. 1431–1441, 2013. Article (CrossRef Link)

[25] “A Fast and Secure Scheme for Data Outsourcing in the Cloud,” in *Proc. of KSII Trans. Internet Inf. Syst.*, vol. 8, no. 8, Aug. 2014.

[26] S. Pearson and A. Benameur, “Privacy, security and trust issues arising from cloud computing,” in *Proc. of Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, pp. 693–702, 2010. Article (CrossRef Link)

[27] C. N. Höfer and G. Karagiannis, “Cloud computing services: taxonomy and comparison,” *J. Internet Serv. Appl.*, vol. 2, no. 2, pp. 81–94, 2011. Article (CrossRef Link)

[28] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, “Not so different after all: A cross-discipline view of trust,” *Acad. Manage. Rev.*, vol. 23, no. 3, pp. 393–404, 1998. Article (CrossRef Link)

[29] S. Bradshaw, C. Millard, and I. Walden, “Contracts for clouds: comparison and analysis of the Terms and Conditions of cloud computing services,” *Int. J. Law Inf. Technol.*, vol. 19, no. 3, pp. 187–223, 2011. Article (CrossRef Link)

[30] S. Subashini and V. Kavitha,“A survey on security issues in service delivery models of cloud computing,” *J. Netw.Comput.Appl.*,vol.34, no.1,pp. 1–11, 2011. Article (CrossRef Link)

[31] C. S. Alliance, *Top threats to cloud computing*. March, 2010.

[32] D. Zissis and D. Lekkas, “Addressing cloud computing security issues,” *Future Gener. Comput. Syst.*, vol. 28, no. 3, pp. 583–592, 2012. Article (CrossRef Link)

[33] E. Prieto, R. Diaz, L. Romano, R. Rieke, and M. Achemlal, “MASSIF: A promising solution to enhance olympic games IT security,” in *Proc. of Global Security, Safety and Sustainability & e-Democracy*, Springer, pp. 139–147, 2012. Article (CrossRef Link)

[34] S. Kamara and K. Lauter, “Cryptographic cloud storage,” in *Proc. of Financial Cryptography and Data Security*, Springer, pp. 136–149, 2010. Article (CrossRef Link)

[35] C. Rong, S. T. Nguyen, and M. G. Jaatun, “Beyond lightning: A survey on security challenges in cloud computing,” *Comput. Electr. Eng.*, vol. 39, no. 1, pp. 47–54, Jan. 2013. Article (CrossRef Link)

[36] R. Buyya, R. N. Calheiros, and X. Li, “Autonomic cloud computing: Open challenges and architectural elements,” in *Proc. of Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*, pp. 3–10, 2012. Article (CrossRef Link)

[37] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, “Benchmarking cloud serving systems with YCSB,” in *Proc. of the 1st ACM symposium on Cloud computing*, pp. 143–154, 2010. Article (CrossRef Link)

[38] J. Schad, J. Dittrich, and J.-A. Quiané-Ruiz, “Runtime measurements in the cloud: observing, analyzing, and reducing variance,” in *Proc. of VLDB Endow.*, vol. 3, no. 1–2, pp. 460–471, 2010. Article (CrossRef Link)

[39] D. Kossmann and T. Kraska, “Data management in the cloud: promises, state-of-the-art, and open questions,” *Datenbank-Spektrum*, vol. 10, no. 3, pp. 121–129, 2010. Article (CrossRef Link)

[40] M. A. El-Refaey and M. A. Rizkaa, “CloudGauge: a dynamic cloud and virtualization benchmarking suite,” in *Proc. of Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on*, pp. 66–75, 2010. Article (CrossRef Link)

[41] B. J. Lheureux, D. C. Plummer, T. Bova, M. Cantara, E. Knipp, P. Malinverno, “Who’s Who in Cloud Services Brokerage,” Gartner, 2011.

[42] S. K. Nair *et al.*, “Towards secure cloud bursting, brokerage and aggregation,” in *Proc. of Web services (ecows), 2010 ieee 8th european conference on*, pp. 189–196, 2010. Article (CrossRef Link)

[43] R. Buyya, S. Pandey, and C. Vecchiola, “Cloudbus toolkit for market-oriented cloud computing,” in *Proc. of Cloud Computing*, Springer, pp. 24–44, 2009. Article (CrossRef Link)

[44] R. Buyya, S. Pandey, and C. Vecchiola, “Market-oriented cloud computing and the cloudbus toolkit,” *ArXiv Prepr. ArXiv12035196*, 2012.

[45] E. Carlini, M. Coppola, P. Dazzi, L. Ricci, and G. Righetti, “Cloud federations in contrail,” in *Proc. of Euro-Par 2011: Parallel Processing Workshops*, pp. 159–168, 2011. Article (CrossRef Link)

[46] B. Rochwerger *et al.*, “The reservoir model and architecture for open federated cloud computing,” *IBM J. Res. Dev.*, vol. 53, no. 4, p. 4: 1-4: 11, 2009. Article (CrossRef Link)

[47] J. Tordsson, R. S. Montero, R. Moreno-Vozmediano, and I. M. Llorente, “Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers,” *Future Gener. Comput. Syst.*, vol. 28, no. 2, pp. 358–367, 2012. Article (CrossRef Link)

[48] “J. Catone, How Much Data Will Humans Create & Store This Year? [INFOGRAPHIC], Mashable Social Media, June 2011.” .

[49] J. Gantz and D. Reinsel, “The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,” *IDC IView IDC Anal. Future*, vol. 2007, pp. 1–16, 2012.

[50] P. Russom, “Big data analytics,” *TDWI Best Pract. Rep. Fourth Quart.*, pp. 1–35, 2011.

[51] S. Chaudhuri, “What next?: a half-dozen data management research goals for big data and the cloud,” in *Proc. of the 31st symposium on Principles of Database Systems*, pp. 1–4, 2012. Article (CrossRef Link)

[52] M. Bohlouli *et al.*, “Towards an integrated platform for big data analysis,” in *Proc. of Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*, Springer, pp. 47–56, 2013. Article (CrossRef Link)

[53] H. Demirkan and D. Delen, “Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud,” *Decis. Support Syst.*, vol. 55, no. 1, pp. 412–421, 2013. Article (CrossRef Link)

[54] S. Gilbert and N. Lynch, “Brewer’s conjecture and the feasibility of consistent, available, partition-tolerant web services,” *ACM SIGACT News*, vol. 33, no. 2, pp. 51–59, 2002. Article (CrossRef Link)

[55] R. Cattell, “Scalable SQL and NoSQL data stores,” *ACM SIGMOD Rec.*, vol. 39, no. 4, pp. 12–27, 2011. Article (CrossRef Link)

[56] “ Marozzo F, Talia D, Trunfio P. Large-Scale Data Analysis on Cloud Systems. ERCIM News. 2012.” .

[57] M. Bohlouli, J. Dalter, M. Dornhöfer, J. Zenkert, and M. Fathi, “Knowledge discovery from social media using big data-provided sentiment analysis (SoMABiT),” *J. Inf. Sci.*, vol. 41, no. 6, pp. 779–798, 2015. Article (CrossRef Link)

[58] C. Chu *et al.*, “Map-reduce for machine learning on multicore,” *Adv. Neural Inf. Process. Syst.*, vol. 19, p. 281, 2007.

[59] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008. Article (CrossRef Link)

[60] T. White, *Hadoop: The definitive guide*. O’Reilly Media, Inc., 2012.

[61] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in action*. Manning Shelter Island, 2011.

[62] "Adaptive Cloud Offloading of Augmented Reality Applications on Smart Devices for Minimum Energy Consumption," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 8, pp. 3090–3102, Aug. 2015. Article (CrossRef Link)

[63] "A Classification-Based Virtual Machine Placement Algorithm in Mobile Cloud Computing," *KSII Trans. Internet Inf. Syst.*, vol. 10, no. 5, May 2016. Article (CrossRef Link)

[64] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wirel. Commun. Mob. Comput.*, vol. 13, no. 18, pp. 1587–1611, 2013. Article (CrossRef Link)

[65] S. Li and Z. Chen, "Social services computing: Concepts, research challenges, and directions," in *Proc. of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*, pp. 840–845, 2010. Article (CrossRef Link)

[66] K. Chard, K. Bubendorfer, S. Caton, and O. F. Rana, "Social cloud computing: A vision for socially motivated resource sharing," *Serv. Comput. IEEE Trans. On*, vol. 5, no. 4, pp. 551–563, 2012. Article (CrossRef Link)

[67] D. Satria, D. Park, and M. Jo, "Recovery for overloaded mobile edge computing," *Future Gener. Comput. Syst.*, 2016. Article (CrossRef Link)

[68] S. Caton, C. Haas, K. Chard, K. Bubendorfer, and O. F. Rana, "A social compute cloud: allocating and sharing infrastructure resources via social networks," *Serv. Comput. IEEE Trans. On*, vol. 7, no. 3, pp. 359–372, 2014. Article (CrossRef Link)

[69] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003. Article (CrossRef Link)

[70] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," *Int. J. Forecast.*, vol. 14, no. 1, pp. 35–62, 1998. Article (CrossRef Link)

[71] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990. Article (CrossRef Link)

[72] O. De Jeses and M. T. Hagan, "Backpropagation through time for a general class of recurrent network," in *Proc. of Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, vol. 4, pp. 2638–2643, 2001. Article (CrossRef Link)

[73] M. P. Cuéllar, M. Delgado, and M. C. Pegalajar, "An application of non-linear programming to train recurrent neural networks in time series prediction problems," in *Proc. of Enterprise Information Systems VII*, Springer, pp. 95–102, 2007. Article (CrossRef Link)

[74] Z. Zheng and M. R. Lyu, "Collaborative reliability prediction of service-oriented systems," in *Proc. of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*, pp. 35–44, 2010. Article (CrossRef Link)

Amin Keshavarzi is a Ph.D. candidate in software systems at Islamic Azad university Qazvin branch. He received his M.Sc. in Software Engineering from the Islamic Azad University science and research branch, Iran, in 2008. He is an Instructor at the Faculty of Computer Engineering of the Islamic Azad university Marvdasht branch in Iran. His research involves Cloud Computing, sensor networks, Data mining and semantic web.

Abolfazl Toroghi Haghighat was born in Mashhad, Iran, on November 14, 1969. He is Assistant Professor, Ph.D. in Computer Engineering department at Qazvin Islamic Azad University, Qazvin, Iran. His research interests include distributed systems, distributed operating system, computational intelligence, wireless and mobile networks. Haghighat holds a Ph.D. in Computer Engineering from Amirkabir University (Iran).

Mahdi Bohlouli is postdoctoral researcher at the University of Koblenz, Germany and holds Ph.D. in computer science from the university of Siegen, Germany. He has been involved in numerous EU projects as well as Program Committee (PC) member of over 30 high qualified conferences and reviewer in over 15 scholarly journals. His research interests are in big data, scalable machine learning and misinformation management. Further information can be found in: www.bohlouli.com