
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2020 SCPE. Volume 21, Issues 3, pp. 349–357, DOI 10.12694:/scpe.v21i3.1697

MONITORING AND PREDICTION OF SLA FOR IOT BASED CLOUD

VIVEK KUMAR PRASAD AND MADHURI D BHAVSAR*

Abstract. Internet of Things (IoT) and cloud computing are the expertise captivating the technology. The most astonishing thing is their interdependence. IoT deals with the production of an additional amount of information that requires transmission of data, storage, and huge infrastructural processing power, posing a solemn delinquent. This is where cloud computing fits into the scenario. Cloud computing can be treated as the utility factor nowadays and can be used by pay as you go manner. As a cloud is a multi-tenant approach, and the resources will be used by multiple users. The cloud resources are required to be monitored, maintained, and configured and set-up as per the need of the end-users. This paper describes the mechanisms for monitoring by using the concept of reinforcement learning and prediction of the cloud resources, which forms the critical parts of cloud expertise in support of controlling and evolution of the IT resources and has been implemented using LSTM. The resource management system coordinates the IT resources among the cloud provider and the end users; accordingly, multiple instances can be created and managed as per the demand and availability of the support in terms of resources. The proper utilization of the resources will generate revenues to the provider and also increases the trust factor of the provider of cloud services. For experimental analysis, four parameters have been used i.e. CPU utilization, disk read/write throughput and memory utilization. The scope of this research paper is to manage the Cloud Computing resources during the peak time and avoid the conditions of the over and under-provisioning proactively.

Key words: Internet of Thing, Cloud Computing; Monitoring; Reinforcement Learning; Prediction; LSTM; Resource Management; Trust

AMS subject classifications. 68M14

1. Introduction.

When Cloud computing combines with the heights of IoT, which continuously processes the information/ data stream, this combination proves to be a boon to the business applications/industries and allowing the services to work from anywhere, anytime, and from any devices. The abilities of the Cloud range from assigning infinite pool of resources through virtualization, handle on Demand services and rapid elasticity. The cloud refers to an IT environs which is designed for remote provisioning of the resources in measured and scalable ways [1]. Much of the internet is devoted to the access of content based IT resources published via WWW (World Wide Web). The cloud can be grounded on any protocols that permit for the remote access of the infrastructure/ resources. This provides mainly three services and is known as SaaS(Software as a service), PaaS (Platform as a service) and IaaS (Infrastructure as a service) [2]. In this paper, the monitoring and prediction approach of the cloud is described. The cloud resources management supports to harmonize the IT resources and its management required by both the end users and the cloud provider. The core part for the same will is performed by Virtual Infrastructure Manager (VIM) which coordinated to the server hardware to create the numbers of Virtual Machines(VMs) or instances as per the user demand [3].

Following are the automated tasks that should be implemented for resource management:

- Creating server images by managing Virtual IT resources.
- Releasing and allocating the virtual IT resources into the available physical infrastructure with responses, such as IT resource instances termination, resuming, starting and pausing.
- Coordinating resources such as replication of the resources fault tolerance and load balancing.
- Applying security policies and usages of cloud service instances
- Monitoring operative circumstances of the IT resources of the cloud.

The resources should be always available when the user needs this, for the same the SLA (service level agreement) should be maintained [4], A SLA is a form of the document that acts as a contract in between the

*Nirma University, India (vivek.prasad@nirmauni.ac.in, madhuri.bhavsar@nirmauni.ac.in)

service provider and the end user. The SLA defines the standard service [5] that the cloud service provider is obligated for.

The prediction [6] for the advance violation for the SLAs can be fruitful to maintain the SLAs in proactive ways. This is an encouraging task because this allows the cloud service provider not only to study from the past disappointment but actually avoid them in the first occurrence, this is achieved by monitoring the cloud resource utilization.

SLA Monitor: The SLA monitoring mechanism [7] is castoff to witness the runtime performance of the cloud facilities to ensure that they are satisfying the contractual QoS needs to which are listed in the SLAs. The data collected by the SLA monitoring is aggregated into the SLA reporting metrics to recognize the numerous influences which might create downtime in the cloud. The system can proactively failover or repair the cloud facilities when incomparable conditions occur (identified because of the operational policies by using metrics), like as when the SLA monitoring intelligence found the cloud facility as "down". Or in other words, the SLA monitoring construction key concern is to proactively monitor the SLA in the mandate to predict the probable violations before they happen.

The repeated violation of the SLAs experienced by the end user's will damage the image of the cloud service provider. Hence it's a duty of the provider to reduce the violations happened because of SLA degradation. A good monitoring and prediction technique will solve these types of delinquent.

Motivation: Dynamic provisioning via resource scheduling is a difficult task and requires proper management of cloud resources. As Cloud contains an infinite pool of resources and managing these resources, optimally is an open research challenge. Hence motivated by these, a self-learning based framework has been implemented through cloud resource monitoring using Reinforcement learning and prediction of the cloud resources using Long Short Term Memory has been analyzed on IoT based Cloud. This framework reduces the conditions of the over and underprovisioning. It ensures the timely management of cloud resources and increased revenue for the Cloud Service Provider.

Contributions

- The paper deals with the proactive technique to maintain the SLAs with the viewpoint of the cloud service provider.
- The prediction technique will make use of the historical and present status of the cloud, which will be fulfilled because of the monitoring mechanism.
- The cloud service provider can create trust and reputation for their services by making use of these features.
- These techniques will be useful for admission control and capacity management.
- The conditions like over provisioning and under-provisioning can be solved out.

The structure of this paper is systematized as follows. Section 2 designates related studies. Section 3 deliberates system architecture and the proposed approach. Section 4 deliberates implementation and evaluation. Section 5 concludes the paper with future work.

2. Related studies. Mian et al [8] discussed the cost model that balanced the SLAs violation penalties and the cost of the resources. Singh et al.[9] describes the various categories of resources and their associated cloud application requirements. In another research done by Ayad et al. [10], an action used to trigger when the availability of virtual machines found to be unsatisfactory and the intelligent will automatically move to the adjoining healthy virtual machine.

Kousiouris, G. et al.[11], proposed a solution for prediction was it predicts the user's behavior patterns through time series analysis. Rimal et al.[12] Compared the cloud mechanisms in terms of its architecture, load balancing terminology, security, the framework of the programming, storage and virtualization. In another approach Buyya et al.[13] The resource manager acts as an admission controller and reallocates the resources as and when the resources deviate.

3. Proposed Approach And Its Architecture. Reinforcement learning (RL) is one of the prominent areas of machine learning [14]. It is about captivating the appropriate actions to maximize the reward in a specific condition. It has been deployed by various machines and software to identify the finest likely behavior it should implement in the specific scenario. There are numerous research has been done to implement RL and is well suited to cloud environs as they do not need a priori information of the application performance.

It learns the environment as the task/job runs. To work with RL the policies are required from which the positive and negative rewards will be generated, which tends to change over time and stops when the goal will be received. RL works with the fundamentals of the Bellman equation, as described below

$$Q(s,a) = r(s,a) + \gamma \max_a Q(s',a) \tag{3.1}$$

$Q(s,a)$ is the target, $r(s,a)$ is the reward of taking that action at that corresponding state and $\gamma \max_a Q(s',a)$ is the discounted max Q value among all possible actions from next state.

LSTM: Long short term memory (LSTM) is an extension of the recurrent neural network, which essentially extends their memory [15]. These extended memories are used to store the imperative pieces of knowledge/ experiences that have very elongated time pauses in between. The LSTM's can delete, read and write data from its memory. It has three gates, which are named as output, forget and input gate for its memory. These gates regulate whether or not to let fresh input in, let it influence the output at the present time step and delete the data as this might not useful for the time being.

Architecture and Methodologies used:

- **Monitoring Agent and Reinforcement Learning:**Due to the vibrant or changing nature of resource loads and the complexity of the cloud ecosystem, it is difficult to set up the mathematical model for the energy-efficient resource provision policy. The lack of efficient resource provisioning due to real-time dynamic management of resources is handled with the aid of model-free Reinforcement Learning methodology. The RL intellect with the environmental circumstances of the cloud ecosystem and accomplish the best suitable management policies without precise domain knowledge. Hence this grantee the best effort to the resource allocation problem in the cloud. As the users' requests are dynamic, and the providers are very complex to handle these requests in real-time. RL seems to be a suitable candidate to handle such situations.
- **SLA Analyzer and Match Making Algorithm:** Every cloud users want to be sure about the quality of their services, which is issued by the cloud provider. In the cloud ecosystem, this quality guarantee includes assurances on services enactment and performance. The Service Level Agreement is used to manage the performance of the services. The SLA analyzer manages these agreed qualities of services by using metrics and various threshold values. The matchmaking algorithm is used for the resource allocation mechanism by comparing the number of incoming requests from the cloud users and various resources available in the Cloud ecosystem. SLA usages the matchmaking algorithm to define its metrics and calculate various performance parameters to maintain the quality of the service.
- **Cloud Manager:**The cloud manager takes care of the cloud resources and makes sure that they are working optimally and is integrating with the cloud users and their services.
- **Prediction Manager and LSTM:**LSTM analyzes or observes the future scope (prediction) of the cloud resources and identifies the behavior of the resource demand based on the present input sequences.

As shown in Fig. 3.1 the entire architecture and its tasks can be carried out in the following sequences.

Step 1: Monitoring Agent (Reinforcement Learning) will monitor the resources and will identify the time stamp or episodes by which the particular workload will be finished by making use of the policy to reach the final goal. The agent automatically identifies the standard conduct within the precise content and will exploit its performance.

Step 2: SLA Analyzing service will investigate the various parameters that need to be adjusted to maintain the SLA and form the policy by making use of metrics.

Step 3: Matchmaking algorithms will act as a capacity management process and will identify the resources available as per the current demand. Also identifies the tasks details and its associated infrastructure requirement, whether it's available or not, steps to bring the resources as per the current need.

Step 4: The resource manager will contain the database to carry out the resource management tasks and will receive the input from the matchmaking.

Step 5: Based on the data collected by the resource manager, the prediction approach using LSTM will be implemented to identify future resource demand.

The architecture discussed here will improve the availability and trust factor of the cloud and results in the increased revenue to the cloud service provider.

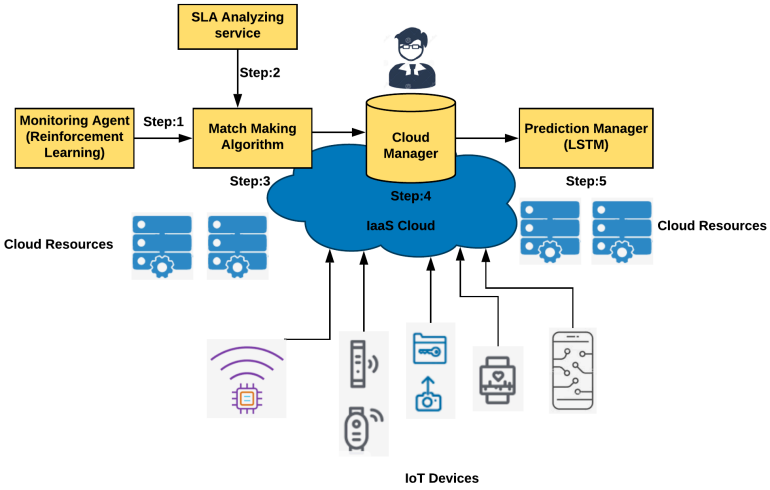

FIG. 3.1. *Architecture of the Cloud for SLA management*

4. Implementation and Evaluation. As discussed in section 3 the reinforcement learning experimentation have been carrying out into the data set of 1750 VMs of distributed data center and is used by various services that usages IoT plate form for the data processing. For the analysis of the experiments, the Openstack based private cloud have been used, with INTEL XEON SILVER CPU with 8 CPU cores and 16 number of threads and 128 GB of RAM. The parameter used is described in Table 4.1. The results are shown in Fig. 4.1 and Fig. 4.2 which make use of the CPU utilization. Here the workload has been classified as less than and greater than the threshold defined. The RL intellect with the environmental circumstances of the cloud ecosystem and accomplish the best suitable management policies without precise domain knowledge. Hence this grantee the best effort to the resource allocation problem in the cloud. The threshold is defined based upon the elastic nature of the cloud resources. As in the case of Figure 4.3 the CPU utilization is up to 35 percent.In an another case like in Figure 4.4 the variable peak rises up to 50 percent of the utilization of the CPU.

Once the results of the monitoring will be identified based on the epochs and rewards, the prediction approach using LSTM will determine the future demand of the resources for a particular type of workload in a proactive way as displayed in Fig.4.3, and Fig.4.4 for different ranges of CPU utilization.As such this can be analyzed that in Fig. 4.3 the range of the CPU utilization is upto 35 % and as per figure 4.4 the utilization has been reached till 50 %. These percentage of the work can be treated as different workload criteria. The parameter used here is depicted in Table 4.2. A total of five LSTM layers has been used and the every layer the neurons has been increased with percentage of 100. The grid search technique has been implemented to carry out the experimental work and to identify the hyper parameters.

TABLE 4.1
Parameters for reinforcement learning

Parameters	Values
Discount( )	0.8
Tau	0.01
Batch size	32
Layers	(50,50)
Learning rate	0.001
Epsilon decay fraction	0.4
Memory fraction	0.80
Memory type	Deque
Process_observation	Standardizer
Process_target	Normalizer

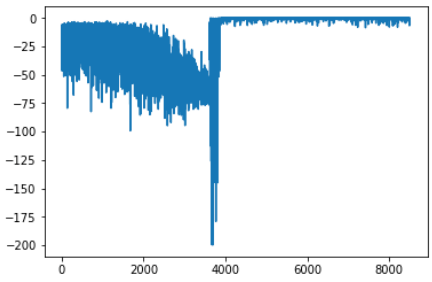

FIG. 4.1. *Graph for rewards generation for the cloud environment policy is to manage the resources in less than 70% threshold. X-Axis: Episodes and Y –Axis: Rewards*

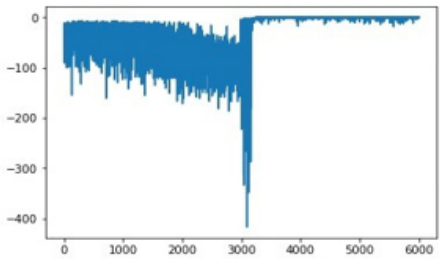

FIG. 4.2. *The CPU utilization is very high and more than the threshold defined i.e 70%. X-Axis: Episodes and Y –Axis: Rewards*

TABLE 4.2
Parameters used for LSTM

Parameters	Values
Batch size	64
Epochs	120
Time steps	10
Input layer	10 nodes
Output layer	10 nodes
Parameters for input layer	4 * LSTM output size * (weights of LSTM output size + 1 Bias variable)
Parameters for output layer	4 * LSTM output size * (weights of LSTM output size + 1 Bias variable)
Optimizer	Adam

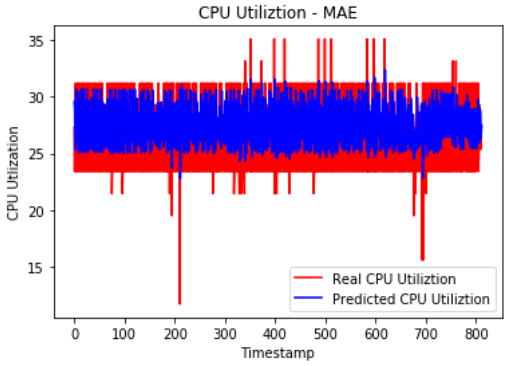

FIG. 4.3. *Predicted values using LSTM for CPU utilization where the usages of the CPU is from 0 to 35 percent*

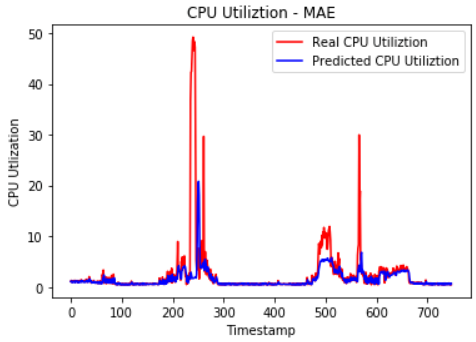

FIG. 4.4. *Predicted values using LSTM for CPU utilization with a range of 0 to 50 percent of CPU usage*

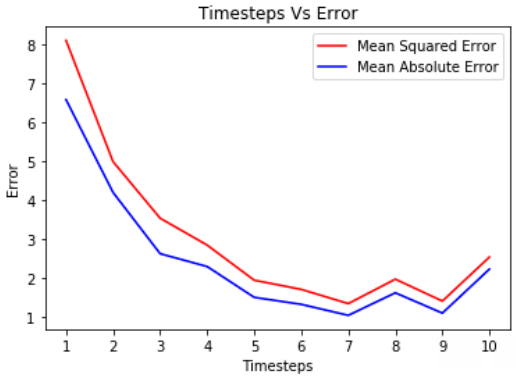

FIG. 4.5. *Variations in error with changing timestamps*

Results and Discussion

The results discussed can be well suited for capacity management of the cloud resources and is based on real-time monitoring and prediction features which usages the reinforcement learning and LSTM to support this framework. The monitoring results will identify the possible scenarios in which the type of tasks (workloads) can be completed in ideal situations with maximized enactment. The prediction results indicate the RMSE (root mean square error) rates as 1.09 for Fig. 4.3 and 1.28 for Fig. 4.4. The MAE (mean absolute error) for Fig. 4.3 and Fig. 4.4 are 0.89 and 0.85.

The changes in the timestamp or setting the window size for the prediction tend to deviate the values for the errors too and have been shown in Fig. 4.5 and the same can be identified by grid search too. Likewise, the parameters such as Disk read throughput[KB/s] and Disk write throughput [KB/s] were analyzed, and the concept of LSTM were implemented.The MAE and RMSE for Disk Read throughput has been calculated as 1.79 and 2.04 and is its graph is shown in Fig.4.6. similarly for Disk write throughout the values for MAE and RMSE are 1.36 and 2.15 and its graph is depicted in Fig.4.7.

The another parameter which has been taken into the consideration is memory utilization in percentage and the experimental value for MAE and RMSE values is calculated as 0.98 and 1.52.The same is depicted in the Fig. 4.8. By understanding the future usage demand from the current and past usage patterns of the resources, the cloud service provider can manage their resources. The prediction of the resources such as CPU Utilization, Disk Read Throughput, Disk write Throughput, and Memory utilization is of great importance for handling dynamic scaling of the resources support. It will achieve enhanced efficiency in terms of energy and cost consumption, and this also maintains the QoS.

Table 5.1 shows the various characteristics of Cloud Computing and its mapping with the proposed approach. It signifies that the proposed approach can handle Scalability [16] , Elasticity[17], Adaptability [18],

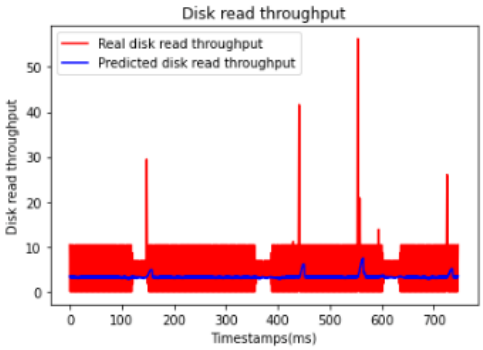

FIG. 4.6. *Disk read throughput[KBs] prediction using LSTM*

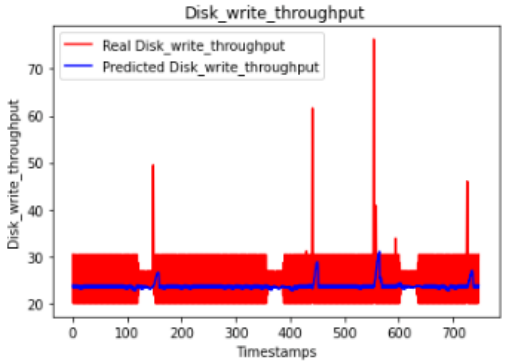

FIG. 4.7. *Disk write throughput[KBs] prediction using LSTM*

Autonomicity [19], Comprehensiveness [20], and Availability [21]. The Extensibility [22], Intrusiveness [23], Resilience [24], and Reliability [25] still needs to be identified and tested for the proposed scheme.

For the fulfillment of the cloud services with proper QoS requirements, a required amount of resources are provisioned by the Cloud Service Provider. Hence based upon the QoS, the SLA will be designed and defined for the smooth conduction of the services [26]. Even the SLA violations are detected regularly to impose the penalty among the parties [27]. So there is a requirement to provide an adequate amount of resources dynamically by the service provider and always avoid the SLA violations proactively. Our proposed approach will proactively avoid the conditions of the SLA violations and thus will be useful for managing the resources as well.

5. Conclusion and future work. To satisfy the end users, SLA violations should be avoided by the cloud service provider. Most of the research proposed the solutions or explanations of violations after they have occurred, the research paper solves this by making use of a proactive approach using the mechanism of monitoring and prediction. Reinforcement learning and LSTM has been used to implement the same. The proposed solution takes the input from the monitoring data and accordingly does the prediction about the resources and manages the capacity of the resources as per the demand.The proposed technique will help us to solve many real-time problems in the cloud environment, such as can be used for matchmaking algorithms, SLA management, capacity planning, and admission control.

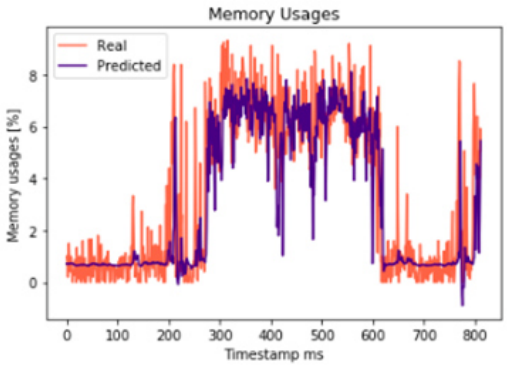

FIG. 4.8. *Predicted values using LSTM for Memory utilization with a range of 0 to 10 percent of Memory usage*

TABLE 4.3
Cloud Characteristics and its mapping with our proposed approach

Cloud Characteristics	**Proposed Approach**
Scalability	Yes
Elasticity	Yes
Adaptability	Yes
Automaticity	Yes
Comprehensiveness	Yes
Availability	Yes
Extensibility	No
Intrusiveness	No
Resilience	No
Reliability	No

REFERENCES

[1] JOSEP, A. D., KATZ, R., KONWINSKI, A., GUNHO, L. E. E., PATTERSON, D., AND RABKIN, A.. A view of cloud computing. Communications of the ACM, 53 (4) (2010).
[2] IOSUP, A., OSTERMANN, S., YIGITBASI, M. N., PRODAN, R., FAHRINGER, T., AND EPEMA, D. Performance analysis of cloud computing services for many-tasks scientific computing. IEEE Transactions on Parallel and Distributed Systems, 22(6), 931-945 (2011).
[3] YEUNG, M., EL AJALTOUNI, E., PHILLIPS, A., AND ANDERSEN, P. U.S. Patent Application No. 15/655,607 (2019).
[4] GUPTA, S., GUPTA, S. C., MAJUMDAR, R., AND RATHORE, Y. S.. Measuring Cloud Security from risks perspective. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 214-220). IEEE (2016).
[5] CASOLA, V., DE BENEDICTIS, A., ERAÇU, M., MODIC, J., AND RAK, M.. Automatically enforcing security slas in the cloud. IEEE Transactions on Services Computing, 10(5), 741-755 (2017).
[6] LEITNER, P., FERNER, J., HUMMER, W., AND DUSTDAR, S.. Data-driven and automated prediction of service level agreement violations in service compositions. Distributed and Parallel Databases, 31(3), 447-470 (2013).
[7] EMEAKAROHA, V. C., NETTO, M. A., BRANDIC, I., AND DE ROSE, C. A.. Application-level monitoring and SLA violation detection for multi-tenant cloud services. In Emerging Research in Cloud Distributed Computing Systems (pp. 157-186). IGI Global (2015).
[8] PRASAD, V. K., AND BHAVSAR, M. Efficient Resource Monitoring and Prediction Techniques in an IaaS Level of Cloud Computing: Survey. In International Conference on Future Internet Technologies and Trends (pp. 47-55). Springer, Cham (2017).
[9] SINGH, S., AND CHANA, I. QoS-aware autonomic resource management in cloud computing: a systematic review. ACM Computing Surveys (CSUR), 48(3), 42 (2016).
[10] AYAD, A., DIPPEL, U.: AGENT-BASED MONITORING OF VIRTUAL MACHINES. In: 2010 International Symposium in Information Technology (ITSim), vol. 1, pp. 1–6. IEEE (2010)
[11] KOUSIOURIS, G., MENYCHTAS, A., KYRIAZIS, D., GOGOUVITIS, S., VARVARIGOU, T., Dynamic, behavioral-based estimation of resource provisioning based on highlevel application terms in cloud platforms. Future Gener. Comput. Syst. 32, 27–40 (2014)
[12] RIMAL, B. P., CHOI, E., AND LUMB, I.. A taxonomy and survey of cloud computing systems. In 2009 Fifth International Joint Conference on INC, IMS and IDC (pp. 44-51). IEEE (2009).
[13] BUYYA, R., BROBERG, J., AND GOSCINSKI, A.. Cloud computing. Principles and Paradigms. Wiley (2011).

[14] Rao, J., Bu, X., Xu, C. Z., Wang, L., and Yin, G.. VCONF: a reinforcement learning approach to virtual machines auto-configuration. In Proceedings of the 6th international conference on Autonomic computing (pp. 137-146). ACM (2009)

[15] Lai, C. F., Chien, W. C., Yang, L. T., and Qiang, W. (2019). LSTM and Edge Computing for Big Data Feature Recognition of Industrial Electrical Equipment. IEEE Transactions on Industrial Informatics.

[16] Yang, J., Qiu, J., and Li, Y. A profile-based approach to just-in-time scalability for cloud applications. In 2009 IEEE International Conference on Cloud Computing (pp. 9-16). IEEE (2009).

[17] Coutinho, Emanuel Ferreira, Flávio Rubens de Carvalho Sousa, Paulo Antonio Leal Rego, Danielo Gonçalves Gomes, and José Neuman de Souza. Elasticity in cloud computing: a survey. annals of telecommunications-annales des télécommunications 70, no. 7-8 (2015): 289-309.

[18] Khan, Suleman, Muhammad Shiraz, Ainuddin Wahid Abdul Wahab, Abdullah Gani, Qi Han, and Zulkanain Bin Abdul Rahman. A comprehensive review on adaptability of network forensics frameworks for mobile cloud computing. The Scientific World Journal (2014).

[19] Caromel, D. ProActive Parallel Suite: Multi-cores to Clouds to autonomicity. In 2009 IEEE 5th International Conference on Intelligent Computer Communication and Processing (pp. xi-xii). IEEE (2009).

[20] Durao, Frederico, Jose Fernando S. Carvalho, Anderson Fonseka, and Vinicius Cardoso Garcia. A systematic review on cloud computing. The Journal of Supercomputing 68, no. 3 : 1321-1346 (2014).

[21] Lei, L., Dagang, L., Lianwen, J., and Lihong, M.. Constructing a high availiable private cloud storage platform based on OpenStack Swift. Experimental Technology and Management, (5), 37 (2015).

[22] Copil, Georgiana, Daniel Moldovan, Hong-Linh Truong, and Schahram Dustdar. Sybl: An extensible language for controlling elasticity in cloud applications. In 2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, pp. 112-119. IEEE (2013).

[23] Bolte, Matthias, Michael Sievers, Georg Birkenheuer, Oliver Niehörster, and André Brinkmann. Non-intrusive virtualization management using libvirt. In 2010 Design, Automation and Test in Europe Conference and Exhibition (DATE 2010), pp. 574-579. IEEE (2010).

[24] Suciu, George, Alexandru Vulpe, Simona Halunga, Octavian Fratu, Gyorgy Todoran, and Victor Suciu. Smart cities built on resilient cloud computing and secure internet of things. In 2013 19th international conference on control systems and computer science, pp. 513-518. IEEE (2013).

[25] Garg, Ritu, and Mamta Mittal. Reliability and energy efficient workflow scheduling in cloud environment. Cluster Computing 22, no. 4 (2019): 1283-1297.

[26] Vivek Kumar Prasad and Madhuri Bhavsar, Preserving SLA Parameters for Trusted IaaS Cloud: An Intelligent Monitoring Approach", Recent Patents on Engineering (2019) 13: 1.

[27] Prasad, Vivek Kumar, and Madhuri D. Bhavsar. Monitoring IaaS Cloud for Healthcare Systems: Healthcare Information Management and Cloud Resources Utilization. International Journal of E-Health and Medical Communications (IJEHMC) 11.3 (2020): 54-70.

Edited by: Anand Nayyar
Received: Feb 10, 2020
Accepted: May 31, 2020