Efficient Resource Monitoring and Prediction Techniques in an IaaS Level of Cloud Computing: Survey

Vivek Kumar Prasad(✉) and Madhuri Bhavsar

Nirma University, Ahmedabad 382481, Gujarat, India
{vivek.prasad,madhuri.bhavsar}@nirmauni.ac.in
http://www.nirmauni.ac.in/

Abstract. In this paper, we have discussed about the various techniques through which the cloud computing monitoring and prediction can be achieved, This paper provides the survey of the techniques related to monitoring and prediction for the efficient usages of the resources available at the IaaS level of cloud. As cloud provides the services, which are elastic, scalable or highly dynamic in nature, which binds us to make the correct usages of the resources, but in real situations the (Cloud Service Provider)CSP's has to face the situation of under provisioning and over provisioning, where the resources are not fully utilized and being wasted, though this is the survey paper, it ends up with the proposed model where both the concepts of the Monitoring and Prediction will be combined together to give a better vision of the future resource demand in IaaS layer of Cloud Computing.

Keywords: Cloud computing · Monitoring · Prediction
Under provisioning · Over provisioning · IaaS

1 Introduction

Cloud computing [24] is a techniques which allows suitable, on demand network access to the pool of various computing resources such as network, servers, storage application and other services, that can be quickly given back to the end user and released with minimal management effort. The cloud computing services [30] can be classified as Software as a Service Platform as a Service and Infrastructure as a Service along with different deployment models [8]. Essential characteristics of Cloud Computing [14,27] are On-Demand Self-Service, Broad Network Access, Resource pooling, Rapid elasticity, Measured service and metering and billing.

The resource management has to be efficiently used in the IaaS level of cloud computing, because the resources has to be allocated in a right amount [36] for an application. The interconnected resource management areas for efficient resource management are [5], resource discovery, Resource modeling, resource

Z. Patel and S. Gupta (Eds.): ICFITT 2017, LNICST 220, pp. 47–55, 2018.
https://doi.org/10.1007/978-3-319-73712-6_5

Table 1. The Literature survey on Monitoring in to the cloud computing for efficient resource utilization

Sr.No	Authors	Objectives and methodologies	Conclusion and future directives
1	Whiteaker et al. [35]	To identified the delay measurements of the virtual machines (VMs) that consume CPU, memory, I/O, hard disk, and network bandwidth. As Heavy network usage of these competing VMs can introduce high round-trip times	On the spot decision to select the appropriate scheduling algorithms based upon the current scenario has to be identified. Dynamic Scheduling algorithms has to be used to deal with these kinds of issues
2	Wang [34]	A system integrating monitoring with analytics, termed as "Monalytics"? has been discussed which can capture, aggregate, and incrementally analyse data on demand. The properties of the Monalytics are as follow:- 1. Zooming in to 'interesting' locations at regular periods of time. 2. Reducing 'Time to Insight' i.e. capturing the total delay between when 'interesting' events occur and by the time they are recognized (i.e., after analysis is complete)	Identifying patterns usage and. finding ways to reduce Datacentre energy use. Fault Patterns or cost /effectiveness needs
3	Clayman et al. [11]	The distributed model has been used which consists of Virtualisation Plane, the Management Plane, the Knowledge Plane, the Service Enablers Plane, and the Orchestration Plane. Working together these distributed systems form a software-driven virtual network control infrastructure that runs on top of all current network and service infrastructures	Monitoring should not affect the performance and account in case of elasticity, scalability, federation and adaptability without violating the performance instances
4	Hasselmeyer and d'Heureuse [16]	The architecture with a data stream management system has been discussed with the event propagation, filtering, and aggregation component. To developed adaptation which makes it easy to interact with the monitoring system	Still the dynamicity is not reached, and enhancements are going on. New architecture are still in demand that can handle the dynamicity. Prediction mechanism has to be analysed to deal with such situation
5	Mian et al. [26]	The cost model has been discussed, which balances resource costs and penalties from SLAs if the SLA's are violated	Usage of static provisioning to provide an initial configuration and then moving to the concepts of dynamic enhancement is yet to be analysed
6	Ayad and Dippel [4]	Continuously check the availability of the virtual machines and automatically intervene in the case of VM failure If agent report that the machine is no longer healthy (corrupt or intrusion) to run, the monitor will destroy those machine, rolling back to the nearest healthy backup available and restart again. Destroying, recovering and restarting the VM should not take more time	Agents has to be made intelligence using other Machine Learning Techniques to get better results Needs to add more functionality in the agents and monitoring systems

(*continued*)

Table 1. (*continued*)

Sr.No	Authors	Objectives and methodologies	Conclusion and future directives
7	Li [23]	The Systematic Literature Review (SLR) method was employed to collect applicable suggestions to investigate the Cloud services evaluation turn by turn The time to time collection of evidences are used to make updates to the knowledge to focus upon new research areas	The metrics can be made based upon the following points The data from SLR will be stored into structured database in support of the services of cloud evaluation methodology to develop superior evaluation metrics
8	Da Cunha Rodrigues [12]	Survey paper which describes various monitoring techniques	Monitoring has to be achieved without compromising application performance and SLA's Integrating different cloud monitoring techniques together When specific requirement will come, it is either negatively or positively affected by other requirements, thus balancing among cloud monitor requirements is a challenging and important trend
9	Hill and Humphrey [17]	Create clusters of machines on-demand and use them for small to medium scale computational problems.	Root cause analysis: techniques able to derive the causes of the observed phenomena, spotting the right thread in the complex fabric of the Cloud infrastructure. Root cause analysis here indicates the primary factor which results into the failure of the system
10	Aceto [1]	Survey Paper which highlights that monitoring is required at both CSP and as well as the client side too	Cross Layer Monitoring: Consumers and Providers make their decisions based on a limited horizon. Both of them has to be considered With Cloud monitoring requirements also focus on minimizing the related resource/ energy consumption and monitoring of Federated Clouds is also a challenge
11	Botta et al. [7]	The workload modelling and generation has been discussed. Different contributions have been provided in terms of studies of real and synthetic workloads.	An important challenge is about the workload generators specifically designed for Cloud scenarios (with adaptability) and it should give the correct value if used for analysis of results

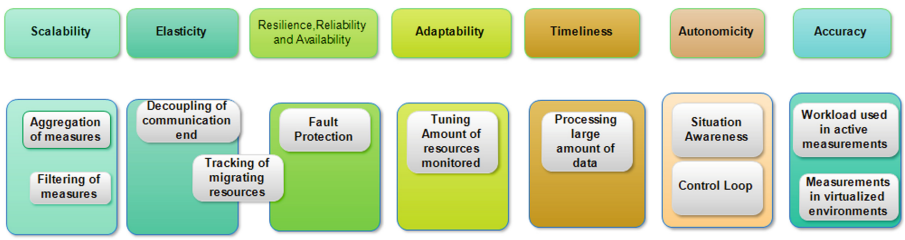

Fig. 1. Cloud monitoring necessity

Table 2. The Literature survey on prediction in to the cloud computing for efficient resource utilisation

Sr.no	Title authors	Objectives and methodology	Conclusion and future directions
1	Kousiouris [21]	It predicts the anticipated user behaviour (Behavioural level). Patterns are identified through a time series analysis	The potential usage of Support Vector Machines has to be analysed As it can perform better than ANN in various cases
2	Menascé and Almeida [25]	Analyse the customer behavioural pattern for website's workload characterization. Also make use of Customer Behaviour Model Graphs for calculating different metrics in order to find workload	To find the accuracy of these metrics
3	Almeida [2]	Different steps involved in capacity planning are discussed. This paper provides base for different activity of workload prediction	Different Tools like Matlab etc. can be used for forecasting, planning, analysis of work load depending on the applications type and its usages
4	Rimal et al. [28]	The author compare different cloud system on the basis of architecture, virtualization, storage, load balancing, interoperability, programming framework, security etc.	Which scheduling algorithms are suitable to which kind of environment? Is still an open research to be discussed among the researchers
5	Huang et al. [18]	To capture the relationship between the workload and the performance metrics. It is possible to ensure that performance of applications is above a minimum threshold. so the SLA violations can be avoided. To include the domain knowledge to model the application behaviour	Some controllers such as fuzzy controllers are based on the rule based approaches. The rules extraction is not easy for the resource management. The ability of the controllers depends on the defined rules and the rule based approaches do not have the learning capability. High availability is required
6	Buyya et al. [9]	The future demand of applications should be predicted accurately in a way that the resources manager is able to reallocate resources before the workload changes occurs	They cannot extract all useful patterns whose length is less/more than the fixed length. Choosing the length of the pattern (the length of the sliding window) for different regions of workloads is one of the most important challenges in these methods
7	Singh and Chana [29]	Different types of resources which include physical resources such as compute, memory, storage, servers, processors and networking are allocated to cloud applications were discussed	Most of the existing methods focus on one or two resources and ignore the correlation between resources. Researchers could investigate the correlation between these resources and provide more reasonable results for the resource manager
8	Urgaonkar [32]	Researchers could develop the new prediction approaches based on both of the reactive and the proactive methods. The proactive prediction methods should be able to extract all access patterns correctly	The reactive provisioning methods react to the surge of fluctuations or the deviation from the expected behaviour. They allocate the additional resources according to the workload increase to prevent SLA violation. Timeliness is the issue here
9	Buyya et al. [8]	Historical executions details (Statistical data) will be used for prediction of resources selection for the workload assigned	The market oriented principles for supply and demand of the resources should also be considered

(*continued*)

Table 2. (*continued*)

Sr.no	Title authors	Objectives and methodology	Conclusion and future directions
10	Ullrich and Lassig [31]	Make use of the pattern extracted from previous executions. Three different categories of load balancing: black box, grey box and white box were discussed	Predict the necessary resource adaption in real time if not even in advance. Resource Consumption can be applied based on the type of application
11	da Silva Dias [13]	Made use of monitoring agents for self-configuration	Self-Adaptive Capacity Management: Monitor and will respond to certain conditions that is overload or underutilisation of the resources at run time has to be analysed
12	Amiri and Mohammad-Khanli [3]	Survey Paper	The new approaches should be able to extract all the behavioural patterns of workloads independent of the fixed pattern length Capabilities of online learning has to be analyzed, able to identify the interesting trends or patterns of the workload variations. Researchers could develop the new prediction approaches based on both of the reactive and the proactive methods

prediction and resource monitoring [22]. The Literature survey on Monitoring in to the cloud computing for efficient resource utilization are as follow (Tables 1 and 2).

Fig. 1 shows [19] the relationship between the cloud properties to the below mentioned key points as, scalability depends upon the aggregation of measures and filtering of measure etc.

2 Prediction

There are various case studies [20,25] that indicates that the workload prediction plays very important role for any company. Basic Steps Required for Workload Prediction [2] and understanding the environment where the workload has to be executed, characterize the workload based upon its availability of resources or capacity planning, behaviour pattern etc., are key features for the prediction mechanism. The next step is to identify the parameter of the workload modelling, which depends upon the type of applications [2]. Lets now analyze the literature review of various prediction techniques.

3 Combining Monitoring and Prediction Techniques [10]

The above Fig. 2 indicates the connectivity between the monitoring and prediction mechanism to the cloud scenarios, their relationship has been identified [37] Prioritization Engine. As we can have more task aligned in the message queue in the cloud computing for different clients asking for the same resources in such scenario Business policies defined by the MSP (Managed Service Provider) helps

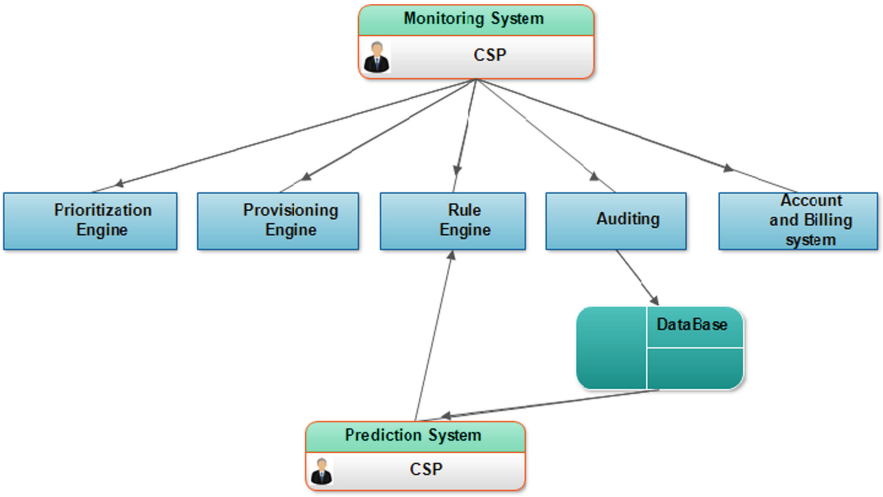

Fig. 2. The combined architecture

to identifying the requests whose execution should be prioritized with respects to the services that they want, in case of resource contentions [15]. The rules engine evaluates the data captured by the monitoring system [33]. Rules engine and the operational policy is the key to guaranteeing SLA agreements.

Monitoring System. Monitoring system collects the defined metrics in SLA. These metrics are used for monitoring resource failures, evaluating operational policies as well as for auditing and billing work. Monitoring system have to interact [6] with the other systems to optimise the its objective if careful usages of the resources available at IaaS of cloud.

Auditing. The adherence to the predefined SLA needs to be monitored and recorded. It is essential to monitor the values of SLA, as any noncompliance leads to strict penalties.

Prediction System: From auditing, the prediction model can be derived, which can be used to predict the resource consumption into the cloud and prediction can be merged with the Machine learning capabilities to increase its effectiveness.

Accounting/Billing System: Based on the payment model and metering The outcome of this model will predict the accurate resources usage for the specific type of work load and the problems that arises because of the under provisioning and over provisioning can be avoided.

4 Conclusion

In this survey paper we have highlighted the concepts of monitoring and prediction, which are an essentials for cloud computing environment, as the IaaS

gives us a vision of infinite pool resources and managing such a huge resources while serving at local level as well as at remote level site is a tedious task and can be handled efficiently if the mechanism of monitoring and prediction concepts should be mapped to the cloud environment. The algorithms related to these two can be merged with the techniques of mathematical modelling, artificial intelligence and machine learning for better accuracy of results and analysis. The model which has been discussed at the concluding portion in this paper will allow the researchers to impose the techniques to implement monitoring and prediction at the correct position with respect to its associated attributes of the cloud computing environment.

References

1. Aceto, G., Botta, A., De Donato, W., Pescapè, A.: Cloud monitoring: a survey. Comput. Netw. **57**(9), 2093–2115 (2013)
2. Almeida, V.A.F.: Capacity planning for web services techniques and methodology. In: Calzarossa, M.C., Tucci, S. (eds.) Performance 2002. LNCS, vol. 2459, pp. 142–157. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45798-4_7
3. Amiri, M., Mohammad-Khanli, L.: Survey on prediction models of applications for resources provisioning in cloud. J. Netw. Comput. Appl. (2017)
4. 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)
5. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener. Comput. Syst. **28**(5), 755–768 (2012)
6. Bose, S.K., Sundarrajan, S.: Optimizing migration of virtual machines across datacenters. In: International Conference on Parallel Processing Workshops, ICPPW 2009, pp. 306–313. IEEE (2009)
7. Botta, A., Dainotti, A., Pescapé, A.: A tool for the generation of realistic network workload for emerging networking scenarios. Comput. Netw. **56**(15), 3531–3547 (2012)
8. Buyya, R., Broberg, J., Goscinski, A.M.: Cloud Computing: Principles and Paradigms, vol. 87. Wiley, New York (2010)
9. Buyya, R., Calheiros, R.N., Li, X.: Autonomic cloud computing: open challenges and architectural elements. In: 2012 Third International Conference on Emerging Applications of Information Technology (EAIT), pp. 3–10. IEEE (2012)
10. Chen, H., Fu, X., Tang, Z., Zhu, X.: Resource monitoring and prediction in cloud computing environments. In: 2015 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence (ACIT-CSI), pp. 288–292. IEEE (2015)
11. Clayman, S., Galis, A., Mamatas, L.: Monitoring virtual networks with lattice. In: 2010 IEEE/IFIP Network Operations and Management Symposium Workshops (NOMS Wksps), pp. 239–246. IEEE (2010)
12. Da Cunha Rodrigues, G., Calheiros, R.N., Guimaraes, V.T., dos Santos, G.L., de Carvalho, M.B., Granville, L.Z., Tarouco, L.M.R., Buyya, R.: Monitoring of cloud computing environments: concepts, solutions, trends, and future directions. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 378–383. ACM (2016)

13. da Silva Dias, A., Nakamura, L.H.V., Estrella, J.C., Santana, R.H.C., Santana, M.J.: Providing IaaS resources automatically through prediction and monitoring approaches. In: 2014 IEEE Symposium on Computers and Communication (ISCC), pp. 1–7. IEEE (2014)
14. Dillon, T., Wu, C., Chang, E.: Cloud computing: issues and challenges. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 27–33. IEEE (2010)
15. Gor, K., Ra, D., Ali, S., Alves, L., Arurkar, N., Gupta, I., Chakrabarti, A., Sharma, A., Sengupta, S.: Scalable enterprise level workflow and infrastructure management in a grid computing environment. In: IEEE International Symposium on Cluster Computing and the Grid, CCGrid 2005, vol. 2, pp. 661–667. IEEE (2005)
16. Hasselmeyer, P., d'Heureuse, N.: Towards holistic multi-tenant monitoring for virtual data centers. In: 2010 IEEE/IFIP Network Operations and Management Symposium Workshops (NOMS Wksps), pp. 350–356. IEEE (2010)
17. Hill, Z., Humphrey, M.: A quantitative analysis of high performance computing with Amazon's EC2 infrastructure: the death of the local cluster? In: 2009 10th IEEE/ACM International Conference on Grid Computing, pp. 26–33. IEEE (2009)
18. Huang, D., He, B., Miao, C.: A survey of resource management in multi-tier web applications. IEEE Commun. Surv. Tutor. **16**(3), 1574–1590 (2014)
19. KaurSahi, S., Dhaka, V.S.: A review on workload prediction of cloud services. Int. J. Comput. Appl. **109**(9), 1–4 (2015)
20. Kohavi, R., Longbotham, R.: Online experiments: lessons learned. Computer **40**(9), 103–105 (2007)
21. Kousiouris, G., Menychtas, A., Kyriazis, D., Gogouvitis, S., Varvarigou, T.: Dynamic, behavioral-based estimation of resource provisioning based on high-level application terms in cloud platforms. Future Gener. Comput. Syst. **32**, 27–40 (2014)
22. Li, A., Yang, X., Kandula, S., Zhang, M.: CloudCmp: comparing public cloud providers. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, pp. 1–14. ACM (2010)
23. Li, Z., Zhang, H., O'Brien, L., Cai, R., Flint, S.: On evaluating commercial cloud services: a systematic review. J. Syst. Softw. **86**(9), 2371–2393 (2013)
24. Mell, P., Grance, T., et al.: The NIST definition of cloud computing (2011)
25. Menascé, D.A., Almeida, V.A.F.: Challenges in scaling e-business sites. In: Interantional CMG Conference, pp. 329–336 (2000)
26. Mian, R., Martin, P., Vazquez-Poletti, J.L.: Provisioning data analytic workloads in a cloud. Future Gener. Comput. Syst. **29**(6), 1452–1458 (2013)
27. Nida, P., Dhiman, H., Hussain, S.: A survey on identity and access management in cloud computing. Int. J. Eng. Res. Technol. **3**(4) (2014)
28. Rimal, B.P., Choi, E., Lumb, I.: A taxonomy and survey of cloud computing systems. In: INC, IMS and IDC, pp. 44–51 (2009)
29. Singh, S., Chana, I.: QoS-aware autonomic resource management in cloud computing: a systematic review. ACM Comput. Surv. (CSUR) **48**(3), 42 (2016)
30. Turab, N.M., Taleb, A.A., Masadeh, S.R.: Cloud computing challenges and solutions. Int. J. Comput. Netw. Commun. **5**(5), 209 (2013)
31. Ullrich, M., Lassig, J.: Current challenges and approaches for resource demand estimation in the cloud. In: 2013 International Conference on Cloud Computing and Big Data (CloudCom-Asia), pp. 387–394. IEEE (2013)
32. Urgaonkar, B., Shenoy, P., Chandra, A., Goyal, P., Wood, T.: Agile dynamic provisioning of multi-tier internet applications. ACM Trans. Auton. Adapt. Syst. (TAAS) **3**(1), 1 (2008)

33. Von Halle, B.: Business Rules Applied: Building Better Systems using the Business Rules Approach. Wiley Publishing, New York (2001)
34. Wang, C., Schwan, K., Talwar, V., Eisenhauer, G., Hu, L., Wolf, M.: A flexible architecture integrating monitoring and analytics for managing large-scale data centers. In: Proceedings of the 8th ACM International Conference on Autonomic Computing, pp. 141–150. ACM (2011)
35. Whiteaker, J., Schneider, F., Teixeira, R.: Explaining packet delays under virtualization. ACM SIGCOMM Comput. Commun. Rev. **41**(1), 38–44 (2011)
36. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)
37. Zhang, W., Song, Y., Ruan, L., Zhu, M.-F., Xiao, L.-M.: Resource management in internet-oriented data centers. Ruanjian Xuebao/J. Softw. **23**(2), 179–199 (2012)