Cloud Monitoring: definitions, issues and future directions

Giuseppe Aceto, Alessio Botta, Walter de Donato, Antonio Pescapè
University of Napoli Federico II (Italy), {giuseppe.aceto,a.botta, walter.dedonato, pescape}@unina.it

Abstract[*] - **Despite its importance for operating Cloud systems, Cloud monitoring has received limited attention from the research community. In this position paper, we provide an analysis of Cloud monitoring. More precisely, we discuss the main motivations, basic concepts and definitions, and point out open research issues and future directions for Cloud monitoring.**

Keywords: Cloud Monitoring and Measurements, SLA Monitoring, Cloud Resource Monitoring, Cloud Monitoring Metrics.

1. INTRODUCTION

Accurate and fine-grained monitoring activities are required to efficiently operate Cloud Computing [1] platforms and to manage their increasing complexity and security requirements. In literature, there is a large number of works proposing surveys and taxonomies of Cloud Computing in general [2, 3, 4], of Virtualization technologies [5, 6], and of Cloud Security [7, 8, 1]. To the best of our knowledge, however, there are no specific analysis on definitions, issues and future directions for *Cloud monitoring*. In this paper, we provide an analysis of Cloud monitoring, with specific focus on: (i) relevant motivations at the base of Cloud monitoring (Sec. 2); (ii) basic concepts, definitions, properties and related issues, and platforms for Cloud monitoring (Sec. 3). We further discuss open issues, main challenges and future research directions for Cloud monitoring (Sec. 4). We close the paper with conclusion remarks (Sec. 5).

2. MOTIVATIONS FOR CLOUD MONITORING

Monitoring of Cloud is a task of paramount importance for both Cloud Service Providers (called Providers in the following) and Cloud Service Consumers (called Consumers in the following). On the one side, it is a key tool for controlling and managing hardware and software infrastructures; on the other side, it provides information and Key Performance Indicators (KPI) for both platforms and applications. The continuous monitoring of the Cloud and of its Service Level Agreements (SLAs), - for example, in terms of availability, delay, etc. - supplies both the Providers and the Consumers with information such as the workload generated by the latter or the performance and Quality of Service (QoS) offered through the Cloud, also allowing to implement mechanisms to prevent or recover violations, for both the Provider and Consumers. Cloud Computing involves many activities for which monitoring is an essential task. The most important ones are:

- **Capacity and Resource Planning.** One of the most challenging tasks for application and service developers, before the large scale adoption of Cloud Computing, has always been resource and capacity planning (e.g. web services [9]).
- **Capacity and Resource Management.** The first step to manage a complex system like a Cloud consists in having a monitoring system able to accurately capture its state [10].
- **Data Center Management**. Cloud services are provided through large scale data centers, whose management is a very important activity. Data center management includes two fundamental tasks: (i) *monitoring*, that keeps track of desired hardware and software metrics; (ii) *data analysis*, that processes such metrics to infer system or application states for resource provisioning, troubleshooting, or other management actions [11].
- **SLA Management.** The unprecedented flexibility in terms of resource management provided by Cloud Computing calls for new programming models in which Cloud applications can take advantage of such new feature [12], whose underlying premise is monitoring.
- **Billing.** In order to offer "measured services" [1] allowing the Consumer to pay proportionally to a metered parameter, monitoring is fundamental, not only -trivially- for the Provider (or the Auditor), but also for the Consumer, in order to verify its effective usage of the Cloud services, and also to compare pricing over different providers (a non-trivial monitoring task [13]).
- **Troubleshooting.** The complex infrastructure of a Cloud represents a big challenge for troubleshooting (e.g. root cause analysis), as the cause of the problem has to be searched in several possible components (e.g. network, host, etc.), each of them made of several layers. Monitoring is therefore needed for Providers to understand where to locate the problem inside their complex infrastructure and for Consumers to understand if any occurring performance issue or failure is caused by the Provider or by other causes.
- **Performance Management.** Being the hardware infrastructure maintenance delegated to the Providers, the

[*]**Acknowledgements:** This work has been partially funded by LINCE project of the FARO programme jointly financed by the Compagnia di San Paolo and by the Polo delle Scienze e delle Tecnologie of the University of Napoli Federico II and by PLATINO, a MIUR project in the framework of the PON action.

Cloud Computing model is attractive for most Consumers (primarily medium sized enterprises and research groups). However, despite the attention paid by Providers, some Cloud nodes may attain performance orders of magnitude worse than other nodes [14].

- **Security Management.** Cloud security is very important for several reasons. It is one of the most significant obstacles to the spread of Cloud Computing, especially considering certain kinds of applications (e.g. business-critical ones) and Consumers (e.g. governments) [15].

3. Definitions, Properties and Issues, and Platforms

Layers. According to the work of the Cloud Security Alliance, a Cloud can be modeled in seven layers [16, 17, 18], controlled by either a Provider or Consumer:

- *Facility* - at this layer we consider the physical infrastructure comprising the data centers that host the computing and networking equipment.
- *Network* - here we consider the network links and paths both in the Cloud and between the Cloud and the user.
- *Hardware* - here we consider the physical components of the computing and networking equipment.
- *Operating System* (OS) - at this layer we consider the software components forming the operating system of both the host (the OS running on the physical machine) and the guest (the OS running on the virtual machine).
- *Middleware* - at this layer we consider the software layer between the OS and the user application, typically present only in Cloud with SaaS and PaaS models.
- *Application* - here we consider the application run by the user of the Cloud system.
- *User* - at this layer we consider the user of the Cloud system and the applications that run outside the Cloud (e.g. a web browser running at the user's premise).

In the context of Cloud monitoring, these layers can be seen as where to put the probes of the monitoring system. In fact, the layer at which the probes are located has direct consequences on the phenomena that can be monitored and observed. Besides, due to the very high complexity of Cloud systems, it not possible to be sure that certain phenomena are actually observed or not. For example, if we put a probe into an application that runs into the Cloud, to collect information on the rate at which it exchanges information with other applications running in the same Cloud, we do not necessarily know if this rate comprises also the transfer rate of the network. It depends on if the two applications run on the same physical host or not, and this information is not always exposed by the Provider. Therefore, for Cloud monitoring, is it is important to have all the information in order to properly perform all the tasks described in the previous section.

Abstraction levels. In Cloud Computing, we can have both high- and low-level monitoring, and both are required [19]. High-level monitoring is related to information on the status of the virtual platform. This information is collected at the middleware, application and user layers by Providers or Consumers through platforms and services operated by themselves or by third parties. High-level monitoring information is generally of more interest for the Consumer than for the Provider (being closely related to the QoS experienced by the former). On the other hand, low-level monitoring is related to information collected by the Provider and usually not exposed to the Consumer, and it is more concerned with the status of the physical infrastructure of the whole Cloud (e.g. servers and storage areas, etc.). More precisely [16], for low-level monitoring, specific utilities collect information at the hardware layer (e.g., in terms of CPU, memory, temperature, voltage, workload, etc.), at the operating system layer, at middleware layer (e.g., bugs and software vulnerabilities), at the network layer (e.g., on the security of the entire infrastructure through firewall, IDS and IPS), and at the facility layer (e.g. on the physical security of involved facilities through monitoring of data center rooms using video surveillance and authentication systems). In the following the most common metrics and tests are defined.

Tests and Metrics. Monitoring tests can be divided in two main categories: *Computation-based* and *Network-based* [20]. Computation-based tests are operated by the provider or sometimes demanded to third parties. For example, in the case of EC2 and Google App Engine, Hyperic Inc publishes results of these test on CloudStatus [21]. *Network-based* tests are related to the monitoring of network-layer metrics. This set includes *round-trip time* (RTT), *jitter*, *throughput*, *packet/data loss*, *available bandwidth*, *capacity*, *traffic volume*, etc. [22, 23, 24, 25]. Several experimental studies compared traditional and Cloud-based hosting using these metrics [26].

Properties and Related Issues. In order to operate properly, a distributed monitoring system is required to have several properties that, when considered in the Cloud Computing scenario, introduce new issues, as discussed in the following.

- **Scalability.** A monitoring system is *scalable* if it can cope with a large number of probes [27]. Such property is very important in Cloud Computing scenarios due to the large number of parameters to be monitored about a huge number of resources. This importance is amplified by the adoption of virtualization technologies, which allow to allocate many virtual resources on top of a single physical resource. In literature such issue has been mainly addressed by proposing architectures in which monitoring data and events are propagated to the control application after applying

aggregation and filtering operations, in order to reduce the volume of monitoring data.

- **Elasticity.** A monitoring system is *elastic* if it can cope with dynamic changes of monitored entities, so that virtual resources created and destroyed by expanding and contracting networks are monitored correctly [27]. The main challenge in providing elasticity is related with the fact that it is a new fundamental property introduced by Cloud monitoring and not previously considered as a requirement for monitoring generic distributed systems.
- **Adaptability.** A monitoring system is *adaptable* if it can adapt to varying computational and network loads in order not to be *invasive* (i.e. impeding for other activities) [27]. In fact, the workload generated by active measurements, as well as the collection, processing, transmission and storage of monitoring data and the management of the monitoring subsystem, require computing and communication resources and therefore constitute a cost for the Cloud infrastructure. Thus, the ability to tune the monitoring activities according to suitable policies is of significant importance to meet Cloud management goals [27].
- **Timeliness.** A monitoring system is *timely* if detected events are available on time for their intended use [11]. Timeliness is interdependent with other properties of the monitoring system, such as elasticity, autonomicity and adaptability. Thus, granting it implies the same challenges or trade-offs between opposing requirements.
- **Autonomicity.** An *autonomic* monitoring system is able to self-manage its distributed resources by automatically reacting to unpredictable changes, while hiding intrinsic complexity to Providers and Consumers [28]. Supporting autonomicity in a Cloud monitoring system is not trivial, since it requires to implement a control loop that receives inputs from a huge number of sensors (i.e. the monitoring data) and propagates control actions to a large number of distributed actuators.
- **Comprehensiveness, Extensibility and Intrusiveness.** A monitoring system is *comprehensive* if it supports different types of resources, several kinds of monitoring data, and multiple tenants [29]; it is *extensible* if such support can easily be extended; it is *intrusive* if its adoption requires significant modification to the Cloud [30]. Most non-Cloud-specific monitoring systems were already designed to provide extensibility and low intrusiveness. As for comprehensiveness, an holistic monitoring system has to supports different underlying architectures, technologies, and resources, while preserving isolation among different tenants, and a comprehensive monitoring system allows to better perform troubleshooting activities, which is an issue because of the dynamicity of Cloud environments and of the large number and heterogeneity of resources and parameters considered at different layers (e.g., through plug-ins or functional modules); it is i*ntrusive* if its adoption requires significant modification to the Cloud [30].
- **Resilience, Reliability, and Availability.** A monitoring system is *resilient* when the persistence of service delivery can justifiably be trusted when facing changes [31], that basically means to withstand a number of component failures while continuing to operate normally; it is *reliable* when it can perform a required function under stated conditions for a specified period of time; it is *available* if it provides services according to the system design whenever users request them [32]. The necessity to provide such properties for Cloud monitoring poses several issues, such as tracking and managing heterogeneous monitored and monitoring resources (also migrating from a physical computer to another), characterizing possible faults of the monitoring system itself and protecting against them.
- **Accuracy.** We consider a monitoring system to be *accurate* when the measures it provides are accurate, i.e. they are as close as possible to the real value to be measured. There are two main issues related to the accuracy of Cloud monitoring systems: the workload used to perform the measurements, and the impact of virtualization systems that add additional layers between applications and physical resources.

Platforms. The majority of monitoring approaches and platforms proposed for the Grid scenario have been customized for Cloud systems. Zanikolas et al. [33] surveyed the Grid monitoring research field by introducing the involved concepts, requirements, phases, and related standardization activities (e.g. Global Grid Forum's Grid Monitoring Architecture). According to the definitions reported in the previous sections, commercial and open source platforms implement both high- and low-level monitoring. The following are the main monitoring platforms:

- *Commercial*: CloudWatch, AzureWatch, CloudKick, CloudStatus, Nimsoft, Monitis, LogicMonitor, Aneka, GroundWork.
- *Open Source*: Hyperic-HQ, OpenNebula, CloudStack ZenPack, Nimbus, PCMONS, DARGOS, Sensu.

Some of these platforms have the properties discussed above. However, most of the properties and the related issues have still to be properly studied and implemented, as discussed in the following section.

4. Open Issues and Future Directions

Effectiveness. Main open issues reside in the possibility to have a clear view of the Cloud and to pinpoint the original causes of the observed phenomena. To achieve this, improvements are needed in terms of: (i) custom algorithms

and techniques that provide effective summaries, filtering and correlating information coming from different probes; (ii) 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; (iii) very importantly, accurate measures in an environment dominated by virtualized resources. We believe that the Cloud complexity requires more effort in each of these three research areas (e.g. see [34] for similar issues on 3G network monitoring). As the monitoring system has become a strategic subsystem for Cloud environments, its resilience is to be considered a fundamental property. On this topic, the analysis of the literature highlighted important contributions focused on resilience to faults and to VM migration and reconfiguration (e.g. [35, 36]). Building on this, we believe that more effort is required for currently available Cloud monitoring systems in order to be also reliable. Even if implicitly addressed in Scalability and Adaptability issues, Timeliness in itself is explicitly considered and evaluated only in [11]. This is a fundamental property that can be effectively used to quantitatively evaluate a Cloud monitoring system and objectively compare it with alternatives (e.g. by defining a specific kind of monitored event and measuring the time needed for the information to reach the managing application). Future proposals and comparisons of Cloud monitoring systems should include the use of the related metric, *Time to Insight*, and further research is needed in this field to model the relations among the parameters involved in Timeliness. Similar considerations can be made about the property of Availability of a monitoring system: though it is closely related with Scalability and Reliability, at the best of the knowledge of the authors there are no evaluations in terms of percentage of missed events, unanswered queries and similar failures in employing the monitoring subsystem and no explicit design constraints in ensuring a given level of availability – possibly 100% as monitoring is a critical functionality. The implications in terms of costs of obtaining less than 100% availability should be considered and assessed as well.

Efficiency. Main improvements in terms of efficiency are expected for data management. In particular, algorithms and techniques more and more efficient are needed to manage, quickly and continuously, the large volume of monitoring data necessary to have a comprehensive view of the Cloud, without putting too much burden on the Cloud and monitoring infrastructures both in terms of computing and communication resources. The monitoring system should be therefore able to do several operations on data (collect, filter, aggregate, correlate, dissect, store, etc.) respecting strict requirements in terms of time, computational power, and communication overhead. These requirements become more and more strict with the increasing spread of Cloud Computing and therefore, the increasing number of users and resources.

Besides the improvements reported above, in the next future we foresee different possible research directions for Cloud monitoring, as detailed in the following:

- **New monitoring techniques and tools**. Effective monitoring techniques should be able to provide, on the one hand, very fine grained measures, and, on the other hand, a synthetic outlook of the Cloud, involving all the variables affecting the QoS and other requirements. At the same time, the techniques should not add performance burden to the system. Finally, they should be integrated with a control methodology that manages the performance of the enterprise system. For all these reasons, new monitoring techniques and tools specifically designed for Cloud Computing are needed.
- **Cross-layer monitoring.** The complex structure of Cloud is made of several layers to allow for functional separation, modularity and thus manageability. However, such strong layering poses several limits on the monitoring system, in terms of kinds of analysis and consequent actions that can be performed. These limits include the inability for Consumers to access lower-layer metrics and for Providers to access upper-layer ones. As a consequence, Consumers and Providers make their decisions based on a limited horizon. Overcoming this limitation is very challenging, technology-, privacy- and administration-wise.
- **Monitoring of Federated Clouds.** The standardized collaboration across multiple cloud infrastructures is referred to as resource federation. However, such standardization process is still at an early stage [37]. The high heterogeneity among different Cloud monitoring infrastructures challenges the possibility to obtain a comprehensive monitoring solution for federated Clouds, and this has not been properly addressed in literature yet.
- **Workload generators for Cloud scenarios**. While different contributions have been provided in terms of studies of real and synthetic workloads, an important remaining challenge is that of workload generators specifically designed for Cloud scenarios (e.g. see [38] for emerging networking scenarios).
- **Energy and cost efficient monitoring.** Monitoring activities can be highly demanding in terms of computing and communication resources, and therefore in terms of energy and cost. Another important challenge for next generation Cloud monitoring systems is that of performing monitoring activities satisfying their basic requirements (accuracy, completeness, reliability, etc.), but minimizing the related energy consumption and cost.

- **Standard and common testbeds and practices.** In literature, it is very difficult to find standards on procedures, formats, and metrics for Cloud monitoring. It is authors' opinion that an effort should be made in this direction. For example, Open Cirrus [39] is an open Cloud Computing research testbed to support research into the design, provisioning, and management of services at a global, multi-datacenter scale. The open nature of the testbed is designed to encourage research into all aspects of service and datacenter management. The collaborative use of research facilities provides ways to share tools, lessons learned and best practices, and ways to benchmark and compare alternative approaches for Cloud monitoring. To foster the progress of the state of the art, open platforms for fair comparison and experimentations of Cloud monitoring tools and techniques are needed.

5. CONCLUSION

In this paper we have discussed the main activities in Cloud environment that have strong benefit from or actual need of monitoring, the main properties that Cloud monitoring systems should have, and the issues arising from these properties. To contextualize and study Cloud monitoring, we have provided background and definitions for key concepts. We also have listed some of the main platforms for Cloud monitoring and we found how the current platforms, still being useful, sometimes fail to fulfill all the requirements expressed in the paper. Finally, we have reviewed open issues and future research directions in the field of Cloud monitoring.

REFERENCES

[1] P. Mell, T. Grance, "The NIST Definition of Cloud Computing", NIST Special Publication 800-145, September 2011.

[2] Chunye Gong; Jie Liu; Qiang Zhang; Haitao Chen; Zhenghu Gong; "The Characteristics of Cloud Computing", ICPPW'10, 13-16 Sept. 2010.

[3] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared", GCE'08 Workshop, 1–10, 2008.

[4] L. Atzori, F. Granelli, A. Pescapè, "A Network-Oriented Survey and Open Issues in Cloud Computing: methodology, system, and applications", CRC, Taylor & Francis group, 2011.

[5] NM Chowdhury, and R. Boutaba, "A survey of network virtualization", Computer Networks 54(5), Elsevier, 862–876, 2010.

[6] SN Chiueh, "A survey on virtualization technologies", RPE Report, 2005.

[7] D. Zissis, D. Lekkas, "Addressing cloud computing security issues", Future Generation Computer Systems, 28 (3), March 2012, Pages 583-592.

[8] Md.T. Khorshed, A.B.M.S. Ali, S.A. Wasimi, "A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing", Future Generation Computer Systems, 28(6), June 2012.

[9] D.A. Menascé and V.A.F. Almeida, "Capacity planning for Web services: metrics, models, and methods", Prentice Hall, 2002.

[10] A. Viratanapanu, A. K. A. Hamid, Y. Kawahara, T. Asami, "On demand fine grain resource monitoring system for server consolidation". Kaleidoscope: Beyond the Internet? – IFNS 2010 ITU-T. IEEE, pp. 1-8.

[11] C. Wang *et al.,* "A flexible architecture integrating monitoring and analytics for managing large-scale data centers", Proceedings of ICAC, 2011.

[12] M. Rak, S. Venticinque, T. Mahr, G. Echevarria, G. Esnal, "Cloud application monitoring: The mOSAIC approach". IEEE CloudCom 2011.

[13] A. Li, X. Yang, S. Kandula, and M. Zhang, "CloudCmp: comparing public cloud providers", IMC '10, pp. 1-14.

[14] M. Armbrust *et al.,* "Above the Clouds: A Berkeley View of Cloud Computing. EECS Department, UCB, Tech. Rep. UCB/EECS-2009-28, 2009.

[15] Y. Chen, V. Paxson, R. H. Katz, "What's New About Cloud Computing Security?", Technical Report No. UCB/EECS-2010-5, 2010.

[16] J. Spring, "Monitoring Cloud Computing by Layer, Part 1", Security & Privacy, IEEE 9(2), IEEE, 66–68, 2011.

[17] J. Spring, "Monitoring Cloud Computing by Layer, Part 2", Security & Privacy, IEEE 9(3), IEEE, 52–55, 2011.

[18] "Security Guidance for Critical Areas of Focus in Cloud Computing v2.1", Cloud Security Alliance, Dec. 2009.

[19] F. Desprez, E. Caron, L. Rodero-Merino Adrian Muresan, "Auto-scaling, load balancing and monitoring in commercial and open-source clouds", Chapter in "Cloud computing: methodology, system and applications", 2011.

[20] Y. Mei, L. Liu, X. Pu, S. Sivathanu, "Performance measurements and analysis of network I/O applications in virtualized cloud", CLOUD 2010

[21] CloudStatus. http://www.hyperic.com/products/cloud-status-monitoring

[22] R. P. Karrer, I. Matyasovszki, A. Botta, A. Pescapè, "MagNets - experiences from deploying a joint research-operational next-generation wireless access network testbed", IEEE TRIDENTCOM 2007.

[23] M. Bernaschi, F. Cacace, A. Pescapè, S. Za, "Analysis and Experimentation over Heterogeneous Wireless Networks", IEEE TRIDENTCOM'05.

[24] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, A. Pescapè, "Broadband Internet Performance: A View From the Gateway", ACM SIGCOMM 2011, Toronto, ON, Canada, August 15-19.

[25] A. Botta, A. Pescapè, G. Ventre, "Quality of Service Statistics over Heterogeneous Networks: Analysis and Applications", European Journal of Operation Research, Vol. 191, Issue 3, 16 Dec. 2008, Pages 1075-1088.

[26] Vinod Venkataraman, Ankit Shah, Yin Zhang, "Network-based Measurements on Cloud Computing Services", Citeseer, 2008

[27] S. Clayman, A. Galis, and L. Mamatas, "Monitoring virtual networks with lattice", IEEE/IFIP NOMS, 2010.

[28] Rizwan Mian, Patrick Martin, Jose Luis Vazquez-Poletti, "Provisioning data analytic workloads in a cloud", FGCS Journal, February 2012.

[29] P. Hasselmeyer and N. d'Heureuse, "Towards holistic multi-tenant monitoring for virtual data centers", NOMS 2010 IEEE/IFIP Network Operations and Management Symposium, Apr. 2010

[30] G. Katsaros, R. Küandbert, and G. Gallizo, "Building a Service-Oriented monitoring framework with REST and nagios", SCC 2011, pp. 426-431.

[31] Jean-Claude Laprie "From Dependability to Resilience", IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2008.

[32] Robert W. Shirey, "Internet Security Glossary", Internet Engineering Task Force RFC 4949, informational, 2007.

[33] S. Zanikolas, and R. Sakellariou, "A taxonomy of grid monitoring systems", Future Generation Computer Systems 21(1), Elsevier, 2005.

[34] A. Botta, A. Pescapè, C. Guerrini, M. Mangri, "A customer service assurance platform for mobile broadband networks", IEEE Communications Magazine, vol.49, no.10, pp.101-109, Oct. 2011.

[35] P. Massonet, et al, "A monitoring and audit logging architecture for data location compliance in federated cloud infrastructures", IEEE International Symposium on Parallel and Distributed Processing, May 2011.

[36] G. Xiang, H. Jin, D. Zou, X. Zhang, S. Wen, and F. Zhao, "VMDriver: A Driver-based Monitoring Mechanism for Virtualization", Reliable Distributed Systems, 2010 29th IEEE Symposium on, 72–81, 2010.

[37] CompatibleOne. http://www.compatibleone.org/

[38] A. Dainotti, A. Botta, A. Pescapè, "A tool for the generation of realistic network workload for emerging networking scenarios", Computer Networks, Vol. 56, Issue 15, pp 3531–3547, October 2012.

[39] Open Cirrus. https://opencirrus.org/