A Multidimensional Model for Monitoring Cloud Services

Nuno Palhares[1], Solange Rito Lima[1] and Paulo Carvalho[1]

[1] Centro Algoritmi, Departamento de Informática, Universidade do Minho,
Campus de Gualtar, 4710-057 Braga, Portugal.

Abstract. The complexity of monitoring cloud environments and the lack of standards so far urge for a careful analysis, systematizing and understanding of key points involved when assessing the services provided. In this context, this paper proposes a layered model for Cloud Services monitoring, identifying the multiple dimensions of monitoring, while combining the perspectives of service providers and customers. This process involves the identification of relevant parameters and metrics for each monitoring dimension, focusing on monitoring of resources, quality of service, security and service contracts. Taking a stratified view of the problem, this study contributes to achieve a clearer and more efficient approach to cloud services monitoring.

Keywords: Cloud Computing, Cloud Services, Monitoring, Management.

1 Introduction

The provision of services based on Cloud Computing is becoming a trend and reality. This is mainly due to the decrease in capital and operational costs associated with this technology, combined with the potential advantages cloud computing brings.

In short, *Clouds* can be viewed as a large pool of virtualized resources (e.g. hardware, development platforms and services) easily usable and widely accessible. These resources can be dynamically reconfigured to be adjustable to variable loads, allowing enhanced and transparent resources utilization. The set of resources is typically available based on a *pay-per-use* business model, in which the infrastructure provider offers guarantees through customized SLAs (Service Level Agreements) [1].

The deployment of cloud services may follow distinct service models, commonly defined as IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service), and implementation models, either in the private or public sector, demanding different approaches for monitoring. In *Private Clouds*, resources and relevant data are typically under control and maintained within the organization premises. *Public Clouds* impose additional monitoring requirements mainly due to the wide geographical coverage and the large set of resources involved, requiring extra flexibility, scalability and security concerns. In particular, security issues can affect monitoring between cloud service providers, limiting interoperability.

A relevant topic to be taken into account in cloud monitoring is energy issues. The challenge in Green Cloud Computing involves minimizing the use of resources and

still meeting the required quality and robustness of the service, contributing to a reduction in operational costs and environmental impact [2], [3].

From a customer/provider perspective, offering cloud services also rises economic and contractual issues according to the negotiated SLAs, and corresponding QoS and QoE compromises. SLA compliance is the first step to a profitable interaction between customers and service providers, being a raw mechanism for mutual control. The role of monitoring is therefore reinforced to ground mutual SLA auditing.

In this context, articulating the various aspects cloud monitoring rises, this paper proposes a stratified approach to cloud services monitoring. For each layer, the model identifies the main parameters and metrics to consider, thereby creating an integrated approach for monitoring the different dimensions and perspectives of participating entities. Our aim and main contribution is therefore gathering, clarifying and systematizing major issues involved in cloud monitoring in order to ground and foster the development of comprehensive and flexible monitoring services.

This article is organized as follows: related work is discussed in Section 2; the proposed stratified monitoring model for cloud services is presented in Section 3; and the main conclusions and future work are included in Section 4.

2 Related Work

One of the constant concerns of service providers is related to monitoring and management of cloud services. Table 1 includes monitoring tools currently available to sustain these tasks, being here classified according to the technique and paradigm followed. As shown, the monitoring location can either be local or remote. In the latter, monitoring tools are distributed and scalable systems supporting high performance computing, such as cluster or grid platforms. Monitoring is usually complemented using web-based management platforms.

Table 1. Monitoring tools.

Type	Examples
Local	Sysstat (Isag, Ksars), Dstat.
Remote	Nagios, Ganglia, GroundWork, Cacti, MonALISA, GridICE.
Web Management Platforms	RightScale, Landscape, Amazon CloudWatch, Gomez, Hyperic/Cloud Status, 3Tera, Zenos, Logic Monitor, Nimsoft, Monitis, Kaavo, Tap in systems, CloudKick, Enstratus, YLastic, TechOut, ScienceLogic, Keynote, NewRelic.

In addition to the tools mentioned above, examples of relevant ongoing projects are *Lattice* [4] and *PCMONS* [5]. Lattice is a framework designed primarily to monitor resources and services in virtual environments. Lattice uses a probe-based monitoring system to collect data for the management system. This framework was developed and implemented in conjunction with the RESERVOIR project. RESERVOIR is a cloud service that distinguishes service providers and infrastructure providers and aims to increase the efficiency of computing, enabling the development of complex services. Both geographical, quality and security issues are covered.

PCMONS assumes that monitoring can take advantage of concepts and tools already present in the management of distributed computing. Its main objective is to implement a monitoring system for Private Clouds and IaaS model, using open source software (e.g. Nagios). The architecture of the monitoring system comprises three layers and equates to a centralized model based on client/server connections. The base layer includes infrastructure components, the middle layer (Integration Layer) is responsible for abstracting the details of the infrastructure, allowing the system to be adaptable and extensible (plug-ins) to other scenarios/tools and the top layer provides an interface to assess the compliance with established policies and SLAs.

From the literature review it became evident that there is no consensus in defining monitoring solutions that satisfy all requirements of complex cloud environments. Through a layered view of the problem, we expect to contribute toward an efficient management and optimization of cloud services deployment.

3 Stratified Monitoring of Cloud Services

The proposed model for cloud services monitoring is stratified into four main layers, which are then divided into categories. As shown in Figure 1, the main layers are: Infrastructure; Network; Service/Application; and Customer/Provider. The Infrastructure layer covers monitoring of both physical and virtual resources involved in the cloud computing environment. Apart from the need to monitor distinct components that compose an entire infrastructure, there are other components that should be monitored at this level, namely energy and security. Aspects related to the IP service, throughput, performance and reliability are covered at Network layer. The Layer Service/Application is focused on assessing the availability, efficiency, reliability and safety of a service. Finally the relationship Customer/Service Provider is considered, targeting SLA auditing and accounting (usage and cost) of a specific service. The sections below discuss the four layers in more detail.

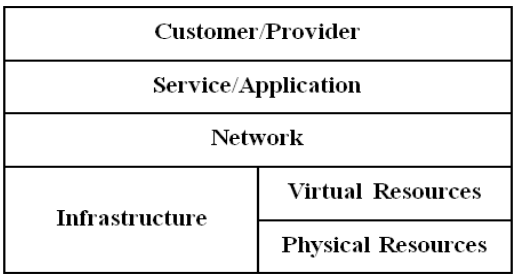

Fig. 1. Proposed Stratified Model for Monitoring Cloud Services.

3.1 Infrastructure

As the foundation for the cloud computing architecture, the physical infrastructure is the major focus of monitoring. All physical components, from processing and storage

devices to network equipment, should be monitored. Most studies in the area agree in considering the percentage of CPU utilization, RAM, storage memory, and statistics of physical network interfaces as the most relevant metrics [2], [3], [5], [6], [7]. As mentioned, network devices must also be monitored, as problems in switches, routers or communication links may affect the cloud topological connectivity. An unstable topology may cause problems which influence a whole range of aspects, such as traffic engineering, throughput, service availability, SLA fulfillment, economic issues, among others. Table 2 summarizes the metrics to consider at Infrastructure level.

Regarding energy consumption, it is known that high temperatures reduce the lifetime of the devices, influencing the reliability and availability of the system. In turn, energy management procedures can affect the system performance in a complex way, since the overall rate computation results from the speed and coordination of multiple elements in a system [8]. According to [8], where ecological and performance issues of resource management system are taken into account (metrics, techniques, models, policies, algorithms), the power consumption is considered an adequate metric to address (see also [3]). In [2], metrics for temperature control and backup power systems (generators, UPS) are also proposed. The purpose is mainly to assess and optimize the use of energy and reduce the emission of carbon monoxide.

Concerning the infrastructure security in terms of physical resources, several restrictions and audits to cloud security are recommended in [9], [10]. Based on work carried out by Cloud Security Alliance (CSA), the cloud is modeled in seven layers, namely: *Facility, Network, Hardware, OS, Middleware, Application and User*. From this model, the first three levels (*Facility, Network* and *Hardware*) need to be considered at physical resources level. The resulting metrics are shown in Table 2.

Table 2. Sample metrics for Physical Resources layer.

Layer	Category	Sample Metrics
Infrastructure Physical Resources	Components	CPU (usage, number of cores), RAM (usage), memory storage (usage, speed of reading and writing), network interface statistics, topology connectivity.
	Energy	Energy consumption, temperature, generator state.
	Security	Fire alarms/sensors, surveillance, access control, IDS and IPS monitoring, firewalls, authentication systems.

Regarding *Facility*, security is mainly handled at physical level, involving the implementation of access control through video surveillance, authentication systems, alarm systems and sensors, among others. The main objective is to prevent malicious intrusion and data manipulation, ensuring the integrity of facilities and components. At *Hardware* level, security metrics are in line with those adopted in the premises where security protocols should be followed up. Regarding *Network* level, which can be described as the boundary between customer data and the customers, mechanisms such as firewalls, Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS) can be adopted.

Within the Infrastructure layer, virtual resources assume a crucial role, increasing transparency, dynamics and scalability, therefore. Virtualization processes involve operations such as suspend/resume/migration and start/stop of Virtual Machines

(VMs). Common metrics at this level are mainly related to the percentage of CPU usage [8], RAM and memory storage of VMs (see Table 3). Statistics on the network interfaces of VMs are equally relevant. Operations related to creation and migration of VMs or number of active instances are also useful information [2], [3], [5], [6], [7].

The security of virtual infrastructure resources can be associated with *OS* and *Middleware* layers [9], [10]. In this case, the metrics should be extracted from monitoring OS-level events and system calls between VMs and hardware. The purpose is mainly to prevent copy and data violations. The *Middleware* layer is considered a potential weak point [10], due to its location between OS and Application layers, involving many components according to the service and architecture. At this layer, the metrics should then be related to monitoring of virtualization and safety systems in heterogeneous cloud architectures.

Table 3. Sample metrics for Virtual Resources layer.

Layer	Category	Sample Metrics
Infrastructure Virtual Resources	Components	CPU (usage, number of cores), RAM (usage), memory storage (usage, speed of reading and writing), statistics of VM interfaces, VM migration, number of active instances.
	Security	Monitoring events and OS at call system level between VMs and hardware.

3.2 Network

In this layer, the relevant metrics are mainly at IP service level. As illustrated in Table 4 these metric types are classified as: Throughput, Performance, Availability and Reliability/Efficiency. The metrics involved derive from telecommunications and computer networking areas, resulting mainly from standardization efforts within ITU-T and IETF IP Performance Metrics (IPPM) workgroup.

Throughput is considered an essential parameter in cloud monitoring [11], [12]. Apart from its relevance for traffic engineering decisions, the verification of SLA fulfillment involves the assessment of throughput related metrics [13]. When analyzing traffic volumes per time unit, monitoring at service class level can bring benefits, particularly for the optimization of network utilization, identification of configuration problems within service classes, etc. Bandwidth quantifies the volume of data that a link or path is able to transfer per time unit. The available bandwidth, thus represents a metric variable in time, where the available capacity is identified, taking into consideration the current load. The capacity, representing the upper bound on available bandwidth, is also a metric that fits in this category.

In [5], [6] statistics related to network traffic are also identified as important sources of monitoring data. This information can also be useful at network layer in addition to physical and virtual infrastructure layer, as mentioned previously.

Performance metrics related to the network level include traditional QoS metrics such as packet duplication, packet loss (OWPL - One-way packet loss, OWLP - One-way loss pattern, IPLR - IP packet loss ratio), delay (OWD - one way delay, RTT - round-

trip time, IPTD - IP packet transfer delay, IPDV - IP packet delay variation), IP packet error ratio (IPER) and Spurious IP packet ratio (SPR) [7].

Regarding Availability, a network can present downtime periods caused by problems in network components, routing configurations, among other aspects. Thus, it is important to monitor the (un)availability of a network, as well as connectivity.

For Reliability/Efficiency assessment, the response time to a network configuration can be a relevant indicator. Upon the occurrence of a network failure, the mean time between failures or the average time to recover are common reliability indicators.

Table 4. Sample metrics for Network layer.

Layer	Category	Sample Metrics
Network	Throughput	Traffic volume per time unit, used and available bandwidth, capacity.
	Performance	Packet duplication, packet loss (OWPL, OWLP, IPLR), delay (OWD, RTT, IPTD, IPDV), IPER, SPR.
	Availability	Uptime, (un)availability of the network connectivity (one or two-way).
	Reliability/ Efficiency	Response time (average/ maximum), mean time to repair upon failure, mean time between failures.

3.3 Service/Application

In the Service/Application layer, the nature of the monitored parameters and how they should be collected depends essentially on the software being monitored and not on the cloud infrastructure per se. One of the main concerns to be taken into account is the availability of a Service/Application. Measuring availability includes registering the periods of time during which a service is running and when it is unavailable. This topic also involves economic issues because in case of unavailability of a Service/Application, SLA violations and subsequent penalties at supplier side may occur. Apart from Availability, relevant metrics for this layer are classified in Reliability/Efficiency and Security (see Table 5). Regarding Reliability/Efficiency, the response time of a given service is a common indicator of efficiency. For instance, in [13], the average and maximum response time are defined as metrics for an online games scenario in Cloud Computing. In case of service failures (due to service unavailability or to QoS degradation), the time to repair should be provided to customers or third-parties responsible for monitoring. The time interval between occurrence of failures is also a measure of efficiency.

The insecure nature of the environment where services and applications are offered turns security into a fundamental aspect to control. As indicated in Table 5 the number of security vulnerabilities is a relevant metric, since it is necessary to monitor behavior to detect possible violations. Other aspects to be monitored and safeguarded are mostly digital certificates, private keys, etc.

The user behavior may also be considered at this layer, including relevant metrics such as login processes, access patterns and associated IPs. Monitoring should also focus on managing passwords, controlling the format of passwords and how often they should be renewed [11]. Specific metrics for each Service/Application type

should also be considered. Furthermore, it is useful to maintain a history, which may contain the IP addresses accessing the service and the login times for each client.

Table 5. Sample metrics for Service/Application layer.

Layer	Category	Sample Metrics
Service/ Application	Availability	Uptime, service (un)availability.
	Reliability/ Efficiency	Response time (average/ maximum), mean time to repair upon failure, mean time between failures.
	Security	Number of security vulnerabilities, access patterns, login processes, password management.
	Others	Login times and IP access records (historic), specific metrics depending on service/application.

3.4 Customer/Provider

In technical terms, the Customer/Provider relationship relies on SLA negotiation. Formally, an SLA is a service contract specifying administrative and technological issues for the type of services provided, and a complete description of each service regarding QoS, uptime, security, privacy, backup procedures, responsibilities and compensation of both parties, among others [14]. A further reference may exist to issues related to geographic location of resources (e.g. datacenters) according to national and international laws. This is an important decision criterion for companies planning to invest in cloud-based solutions [15].

Regarding the management of services, an SLA acts as a valuable auditing instrument both for clients and service providers. The verification of SLA compliance is a cross-layer task spanning all the layers described above. For example, in [5], a metric for average SLA violations is obtained based on the average of CPU usage that was not allocated to an application when requested. Monitoring cloud services usage is also relevant due to the elastic nature of cloud environments, associated with the business model "*pay-as-you-go*", therefore, measuring use and cost become vital aspects [7], [13]. The accounting of services and corresponding revenue allows service providers to adapt pricing and business strategies according to market needs. The sample metrics for the Customer/Provider layer are summarized in Table 6.

Table 6. Sample metrics for Customer/Provider layer.

Layer	Category	Sample Metrics
Customer/ Provider	Auditing	Monitoring SLA violations, penalties.
	Accounting	Monitoring of usage and cost, revenue.

4 Conclusions

Cloud monitoring is a recent and active research area where the lack of related standards is evident. This fact is particularly important and complex when trying to

perform monitoring of cloud services across multiple clouds, involving geographical, quality and legal issues. Contributing to the efforts toward modeling and standardization, this paper has proposed a stratified approach identifying and suggesting parameters, metrics and best practices for efficient monitoring of cloud services and environments. Future work includes validating and tuning the proposed model resorting to an experimental scenario and forthcoming activities in the area.

Acknowledgments. This work is funded by FEDER through the Competitiveness Factors Operational Programme - COMPETE and Portuguese National Funds through FCT - Foundation for Science and Technology under the Project: FCOMP-01-FEDER-0124-022674.

References

1. Vaquero, L., Merino, L., Caceres, J. and Lindner, M.: A Break in the Clouds: Towards a Cloud Definition. SIGCOMM CCR, Vol.39(1), pp. 50-55, Jan. 2009.
2. Werner, J., Geronimo, G., Westphall, C., Koch, F. and Freitas. R.: Simulator Improvements to Validate the Green Cloud Computing Approach. In 7th Latin American Network Operations and Management Symposium (LANOMS), pp. 1-8, Oct. 2011.
3. Beloglazov, A., Abawajy, J. and Buyya, R.: Energy Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing. ELSEVIER, Future Generation Computer Systems 28 (2012), pp. 755-768, May 2012.
4. Chaves, S., Uriarte, R. and Westphall, C.: Toward an Architecture for Monitoring Private Clouds. IEEE Communications Magazine, pp. 130-137, Dec. 2011.
5. Clayman, S., Galis, A., Chapman, C., Toffetti, G., Merino, L., Vaquero, L., Nagin, K. and Rochwerger. B.: Monitoring Service Clouds in the Future Internet. IOS Press, pp. 115-126, Apr. 2010.
6. Peng, Y. and Chen, Y.: SNMP-Based Monitoring of Heterogeneous Virtual Infrastructure in Clouds. In 13th APNOMS, pp. 1-6, Sept. 2011.
7. Choi, T., Kodirov, N., Lee, T., Kim, D. and Lee, J.: Autonomic Management Framework for Cloud-based Virtual Networks. In 13th APNOMS, pp. 1-7, Sept. 2011.
8. Sheikhalishahi, M. and Grandinetti, L.: Revising Resource Management and Scheduling Systems. In 2nd Int. Conf. on Cloud Computing and Services Science, pp. 121-126, 2012.
9. Spring, J.: Monitoring Cloud Computing by Layer, Part 1. IEEE Security & Privacy Magazine, Vol.9(2), pp. 66-68, Mar./Apr. 2011.
10. Spring, J.: Monitoring Cloud Computing by Layer, Part 2. IEEE Security & Privacy Magazine, Vol.9(3), pp. 52-55, May/June 2011.
11. Chaves, S., Westphall, C. and Lamin, F.: SLA Perspective in Security Management for Cloud Computing. In 6th Int. Conf. on Networking and Services, IEEE Computer Society, pp. 212-217, Mar. 2010.
12. Sousa, F., Moreira, L., Santos, G. and Machado, J.: Quality of Service for Database in the Cloud. In 2nd Int. Conf. on Cloud Computing and Services Science, pp. 595-601, 2012.
13. Patel, P., Ranabahu, A. and Sheth, A.: Service Level Agreement in Cloud Computing. Technical Report, Sept. 2009.
14. Cloud Computing Use Case Discussion Group: Cloud Computing Use Cases, white paper v4.0. Technical Report, July 2010.
15. Stamou, K., Morin, J., Gateau, B. and Aubert, J.: Service Level Agreements as a Service - Towards Security Risks Aware SLA Management. In 2nd International Conference on Cloud Computing and Services Science, pp. 663-669, 2012.