Unfolding the Mutual Relation Between Timeliness and Scalability in Cloud Monitoring

Guilherme da Cunha Rodrigues[1], Rodrigo N. Calheiros[2], Glederson Lessa dos Santos[3],
Vinicius Tavares Guimarães[1], Lisandro Zambenedetti Granville[3], Liane Tarouco[3], Rajkumar Buyya[4]

[1]Federal Institute of Education, Science and Technology Sul Rio-Grandense, Charqueadas, Brazil
Email: {grodrigues,vicoguim}@charqueadas.ifsul.edu.br

[2]School of Computing, Engineering and Mathematics,
Western Sydney University, Sydney, Australia
Email: r.calheiros@westernsydney.edu.au

[3]Computer Networks Group, Institute of Informatics
Federal University of Rio Grande do Sul, Porto Alegre, Brazil
Email: {gledersons,granville,liane}@inf.ufrgs.br

[4]**Clou**d Computing and **D**istributed **S**ystems (CLOUDS) Laboratory
School of Computing and Information Systems
The University of Melbourne, Melbourne, Australia
Email: rbuyya@unimelb.edu.au

Abstract— **Cloud computing is a suitable solution for professionals, companies, and institutions that need to have access to computational resources on demand. Clouds rely on proper management to provide such computational resources with adequate quality of service, which is established by Service Level Agreements (SLAs), to customers. In this context, cloud monitoring is a critical function to achieve such proper management. Cloud monitoring systems have to accomplish requirements to perform its functions properly, and currently, there are plenty of requirements which includes: timeliness, adaptability, comprehensiveness, and scalability. However, such requirements usually have mutual influence, which is positive or negative, among themselves, and it has prevented the development of complete cloud monitoring solutions. This paper presents a mathematical model to predict the mutual influence between timeliness and scalability, which is a step forward in cloud monitoring because it paves the way for the development of complete monitoring solutions. It complements our previous work that identified the monitoring parameters (e.g., frequency sampling, amount of monitoring data) that influence timeliness and scalability. Evaluations present the effectiveness of the mathematical model based on a comparison of the results provided by the mathematical model and the results obtained via simulation.**

I. Introduction

Cloud computing is a suitable solution for professionals, companies, and institutions that require access to computational resources on demand along with advantages such as availability, flexibility, and reduced costs [1] [2]. Cloud computing is suitable because it delivers high-quality services based on its five essential characteristics, namely, on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service [3] [4].

The focus of this paper is the *measured service* characteristic of clouds. Measured service provides support to cloud users (e.g., service providers (SPs), infrastructure providers (InPs), and customers) based on resource management in clouds, commonly defined as cloud management [5] [6]. Cloud management is vital to delivering quality services to measured services, and cloud monitoring is essential to providing support for cloud management [7] [8] [9].

Cloud monitoring is a critical management function for cloud operators (e.g., SPs, InPs) that focus on delivering high-quality cloud services [10] [11]. This is because cloud services are based on contracts between cloud operators and customers. Such contracts are usually defined as Service Level Agreements (SLAs) [12]. Service level agreements have established the expectation of customers related to the quality of service provided by the cloud operators as well as the capacity that a specific cloud has to offer such services [13] [14].

Aiming to achieve fair and competitive SLAs, cloud operators need to have means (e.g., information, notifications, analysis) to define metrics adequately based on the cloud capacity to offer services to customers [15]. For instance, there is the case of how can a specific cloud operator ensures performance to a customer based on an amount of resources that it has to offer on demand. Thus, clouds have relied on cloud monitoring to provide means for the delivery of high-quality services [16] [17]. In this context, cloud monitoring has requirements that are essential to cloud monitoring systems

978-1-5386-6950-1/18/$31.00 ©2018 IEEE

that have the intention to offer complete monitoring solutions.

Cloud monitoring requirements are essential to cloud monitoring systems that want to perform its functions properly [18] [19]. According to the literature, there are several cloud monitoring requirements such as scalability, elasticity, accuracy, and timeliness. However, cloud monitoring systems usually accomplish one or more requirements, and there is not any cloud monitoring system that meets all the requirements. It happens because requirements have mutual influence among themselves. This mutual influence must be unveiled to pave the way for the development of complete cloud monitoring solutions. Efforts in this direction started with evaluations between specific cloud monitoring requirements [18] [20] [21]. In this context, in our early research [22] we paved the foundation for investigation of the mutual influence between timeliness and scalability, finding out the monitoring parameters (e.g., frequency sampling, amount of monitoring data) that mutually influence both. This paper is built on top of the findings of such previous work [22].

Timeliness and scalability are two significant cloud monitoring requirements. Timeliness is the competence that a cloud monitoring system has to detect events in time to support users to get information at the moment in which they need to use such information [22]. It is significant to cloud monitoring because cloud services are regulated by SLAs. In other words, if monitoring data is not timely, an action correcting violation in the SLA cannot be performed in time, resulting in penalties (i.e., costs) to cloud operators. On the other side, scalability is the competence to increase the amount of probes in a monitoring system to cope with resources in the cloud [22]. It is important to cloud monitoring systems because the cloud business model provides resources on demand, and it usually happens in quickly.

Timeliness and scalability are two requirements that have direct relation. It is confirmed by two factors that reinforce the mutual influence between both. First, to provide information in the time that cloud users need to access it (timeliness), a monitoring system has to be capable of growing in the number of probes to handle with all resources in the cloud (scalability). Second, the number of probes impairs the capacity of the system to be timely because it has an adverse influence on functions such as synchronization and data collection. Also, timeliness and scalability have direct influence in other requirements (*e.g.,* elasticity, accuracy). Thus, the understanding of the mutual influence between timeliness and scalability will assist in further studies about cloud monitoring requirements.

Besides, to cloud operators, unveiling the mutual influence between timeliness and scalability is useful to enhance their SLAs. In this way, cloud operators can propose SLAs to customers based on the relation between timeliness and scalability to avoid breaches in SLAs, and as a consequence increase their profits. According to our previous work [22], this mutual influence can be predicted based on monitoring parameters such as the amount of monitoring data, and frequency sampling. However, it lacks on a mathematical proof.

This paper proposes a step forward in cloud monitoring because it presents a mathematical model to predict the mutual influence between timeliness and scalability. This model is based on monitoring parameters that influence timeliness and scalability. The main contributions of this paper are:

- It discusses the interplay between timeliness and scalability along with monitoring parameters that influence the relation between both.
- It proposes a mathematical model to predict the impact of scalability over timeliness and conversely.
- It compares results provided by the mathematical model with results of a simulation to demonstrate the usefulness of the proposed model.

II. Predicting the Interplay Between Timeliness and Scalability

Previous research presented a wide discussion about the interplay among cloud monitoring requirements such as adaptability, accuracy, timeliness, and scalability [18] [21] [23].

In our previous work [22], we demonstrated that the mutual influence between timeliness and scalability is suitable to be mathematically represented based on monitoring parameters, namely, monitoring topology, amount of monitoring data and frequency sampling. Moreover, network bandwidth must be considered along with response time as an output metric. In this section, we analyze all of these factors (i.e., monitoring parameters, network bandwidth, response time) to mathematically model the interplay between timeliness and scalability based on traditional network architectures.

Monitoring topology is vital because it contributes to control the communication delay in a cloud monitoring system. It occurs as a result of the distance from agents to managers, which increases the time spent in the process of communication in a network depending on both, the placement of such agents and managers, and the number of communication links.

The amount of monitoring data (*a*) is critical because it causes delays between event occurrence and warning. It happens as a consequence of the growing of the cloud monitoring system that has more monitoring data to be gathered and managed. Then, timeliness is impaired when the cloud monitoring system escalates to cope with the cloud infrastructure.

Frequency sampling (*fs*) is an essential monitoring parameter because, when sampling in higher frequency, the amount of monitoring data is increased in a network, causing communication delay. Thus, timeliness is impaired when a cloud monitoring system is gathering samples in tiny intervals of time. It usually happens to resources such as CPU that has to be continuously monitored in narrow periods of time. In this scenario, to perform frequency sampling is important to unmask how the interval between data collection and response time impairs timeliness in accordance with the scalability of a cloud monitoring system.

Network bandwidth (*b*) must be considered to calculate the mutual influence between timeliness and scalability because it consists of physical links among a plenty of resources provided by the cloud infrastructure. It represents the capacity that a cloud has in terms of data transmission among its peers.

Response time (*rt*) must be used as the output metric because it means the time spent on the data collection and

978-1-5386-6950-1/18/$31.00 ©2018 IEEE

notification. Therefore, it represents the quantification of the influence of scalability over timeliness and vice-versa.

Conventional data center networks for clouds are based on either two or three level trees of host's (h) [24] [25] [26] [27] [28]. Equation 1 was developed based on all parameters discussed in this section and considering conventional data center networks.

$$rt = \frac{a * fs(1 + h + h * h_{l2})}{b} \qquad [ms] \tag{1}$$

Fat-tree is a special type of Clos Topology [25] [26] [28] [29] [30], and it is a topology based on trees. A usual Fat-tree topology is a three level tree with more communication channels. Thus, to adjust the Equation 1 to work in a Fat-tree topology we have to take into consideration the amount of communication channels provided which is represented by ac (Amount of Channels) in Equation 2.

$$rt = \frac{a * fs(1 + \frac{h}{ac} + \frac{h * h_{l2}}{ac})}{b} \qquad [ms] \tag{2}$$

The mathematical model presented in Equation 1 is useful to predict the mutual influence between timeliness and scalability and vice-versa in a conventional data center network. If necessary, it can be easily adjusted to estimate such mutual influence in Fat-Tree topologies as we demonstrated in Equation 2. Thus, a mathematical model is useful to predict the mutual influence between timeliness and scalability and vice-versa. Aiming to assess it in the next section, we compared the outcomes obtained by the mathematical model with the results obtained via simulations.

III. Evaluating The Mathematical Model

In this section, we demonstrate that the mathematical model is useful to predict the mutual influence between timeliness and scalability. Firstly, it evaluates the behaviour of the mathematical model in monitoring topologies based on conventional data center networks. After, it evaluates such behaviour in a monitoring topology based on Fat-tree.

A. Evaluating Monitoring Topologies Based on Conventional Data Center Network

In our previous work [22], we evaluated the interplay between timeliness and scalability. This evaluation was a simulation that provided results to such interplay based on response time (RT). The evaluation unmasked the mutual influence between timeliness and scalability to topologies based on trees like conventional data center networks with either two or three level trees. It provides results that are significant to the development of the mathematical model that aims to predict the mutual influence between such requirements. In this scenario, if the mathematical model presents similar results comparing to the simulation, it shows to be useful and suitable to conventional data center network topologies.

In this section, we compare the results provided by simulation with results obtained by the mathematical model. To organize such comparison we define the results as follow:

Table I. Predicted and Average response time (RT) for 120 bytes of monitoring data with interval sampling of 1 second.

Topology	Hosts/Aggregators	Predicted RT (ms)	Average RT (ms)
Topology 1	64 / 10	7.62	7.72
Topology 1	256 / 18	30.12	30.07
Topology 1	576 / 26	67.62	67.77
Topology 1	1296 / 38	151.99	151.54
Topology 2	64 / 21	127.62	127.93
Topology 2	216 / 43	936.68	936.33
Topology 2	512 / 73	15210.46	15210.46
Topology 2	1331 / 133	362127.78	362127.72

Table II. Predicted and Average response time (RT) for 120 bytes of monitoring data with interval sampling of 10 seconds.

Topology	Hosts/Aggregators	Predicted RT (ms)	Average RT (ms)
Topology 1	64 / 10	7.62	7.67
Topology 1	256 / 18	30.12	30.19
Topology 1	576 / 26	67.62	67.54
Topology 1	1296 / 38	151.99	151.81
Topology 2	64 / 21	127.62	128.02
Topology 2	216 / 43	936.68	936.32
Topology 2	512 / 73	3900.12	3900.22
Topology 2	1331 / 133	36155.59	36155.62

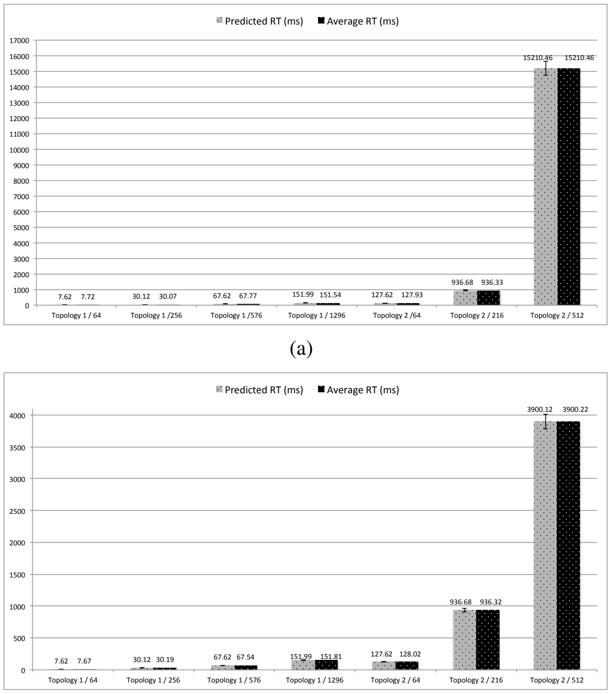
- Predicted RT: It is the result for response time (RT) obtained by the mathematical model.
- Average RT: It is the result for average response time (RT) obtained by the simulation.

Table I, Table II, Table III and Table IV present the results to predicted RT and average RT for the amount of monitoring data for 120 bytes and 150 bytes with frequency sampling based on an interval of 1 and 10 seconds.

Table I and Table II compare results to predicted RT and average RT for 120 bytes of monitoring data with interval sampling of 1 and 10 seconds. Table I results show that predicted RT and average RT are equal to topology 2 with 512 hosts. Moreover, we highlight that the most relevant difference among results is to topology 1 with 64 hosts that are 1.29% apart. Results demonstrate that the behaviour of the mathematical model is compatible with the simulation as observed in Figure 1 (a) which bars to predict RT considers 3% as a margin of error. It demonstrates that the mathematical model is useful to 120 bytes of monitoring data with interval sampling of 1 second, considering 1.29% as a maximum margin of error.

In Table II the results demonstrate that predicted RT and average RT are close in topology 2 with 216, 512 and 1331 hosts. It provides indications that to deep topologies the mathematical model reaches results with small margins of error, in this example 0.04%, 0.01% and 0.01%, respectively. Additionally, the most significant difference among results is to topology 1 with 64 hosts (0.65%). It evidences that the mathematical model results are consistent with the simulation as demonstrated in Figure 1 (b) which bars to predict RT considers 3% as a margin of error. Therefore, the mathematical model is useful to 120 bytes of monitoring data with interval sampling of 10 seconds, considering 0.65% as a maximum margin of error.

Table III and Table IV assess results for predicted RT and average RT for 150 bytes of monitoring data with interval sampling of 1 and 10 seconds. Table III shows that predicted RT and average RT are closer in deep topologies



(a)

(b)

Figure 1. Comparison between predicted and average response time to 120 bytes of monitoring data with interval sampling to 1 and 10 seconds. (a) 1 Second . (b) 10 Seconds.

when compared to results from topology 1 and topology 2. It strengthens the finding that the mathematical model is more accurate to deep topologies and reiterates that this model is more efficient in larger environments if compared to smaller ones. Moreover, we highlight that the most relevant difference among results is to topology 1 with 64 hosts (0.85%), which is the smaller environment evaluated. The mathematical model results (predicted RT) are coherent with the simulation results (average RT) as showed in Figure 2 (a) which bars to predict RT considers 3% as a margin of error. It attests that the solution proposed is useful to 150 bytes of monitoring data with interval sampling of 1 second, considering 0.85% as a maximum margin of error.

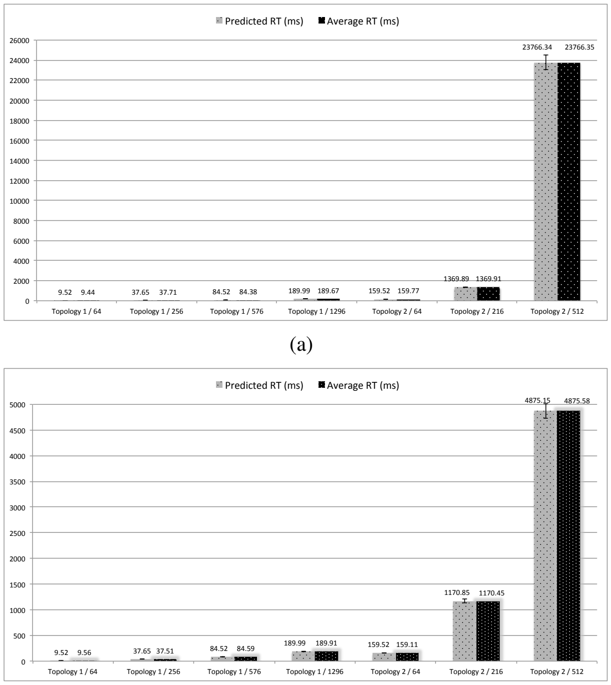
Table IV confirms that predicted RT and average RT are closer to larger topologies such as topology 2 with 216 hosts (0.03%), 512 hosts (0.01%) and 1331 hosts (equal). It demonstrates that the mathematical model is more accurate when applied to larger topologies. On the other hand, if applied to smaller topologies the mathematical model presents results with small margin of error such as in topology 1 with 64 hosts (0.42%). Figure 2 (b) shows the behaviour of the solution compared to the simulation considering 3% as a margin of error to predicted RT. Finally, the solution proves to be useful to 150 bytes of monitoring data with interval sampling of 10 seconds, considering 0.42% as a maximum margin of error.

Table III. PREDICTED AND AVERAGE RESPONSE TIME (RT) FOR 150 BYTES OF MONITORING DATA WITH INTERVAL SAMPLING OF 1 SECOND.

Topology	Hosts/Aggregators	Predicted RT (ms)	Average RT (ms)
Topology 1	64 / 10	9.52	9.44
Topology 1	256 / 18	37.65	37.71
Topology 1	576 / 26	84.52	84.38
Topology 1	1296 / 38	189.99	189.67
Topology 2	64 / 21	159.52	159.77
Topology 2	216 / 43	1369.89	1369.91
Topology 2	512 / 73	23766.34	23766.35
Topology 2	1331 / 133	565787.41	565787.41

Table IV. PREDICTED AND AVERAGE RESPONSE TIME (RT) FOR 150 BYTES OF MONITORING DATA WITH INTERVAL SAMPLING OF 10 SECONDS.

Topology	Hosts/Aggregators	Predicted RT (ms)	Average RT (ms)
Topology 1	64 / 10	9.52	9.56
Topology 1	256 / 18	37.65	37.51
Topology 1	576 / 26	84.52	84.59
Topology 1	1296 / 38	189.99	189.91
Topology 2	64 / 21	159.52	159.11
Topology 2	216 / 43	1170.85	1170.45
Topology 2	512 / 73	4875.15	4875.58
Topology 2	1331 / 133	56374.18	56374.18



(a)

(b)

Figure 2. Comparison between predicted and average response time to 150 bytes of monitoring data with interval sampling to 1 and 10 seconds. (a) 1 Second . (b) 10 Seconds.

In this section, the comparison between predicted RT and average RT demonstrates that the proposed mathematical model is useful to predict the mutual influence between timeliness and scalability to conventional data center network topologies. This model presents accurate results to larger and deep topologies when comparing to smaller and non-deep ones which is desirable to massive environments such as clouds. On the other side, to smaller environments the mathematical model shows to be useful because it reaches closer results with small margin of error.

Nevertheless, the mathematical model must be evaluated based on other common cloud network topology such as Fat-tree. In the next section, we perform a set of tests to assess the behaviour of the mathematical model in a Fat-tree topology to fulfil this gap.

B. Evaluating Monitoring Topologies Based on Fat-Tree

In the previous section, we compared the results provided by simulation with results obtained by the mathematical model to conventional data center networks. Aiming to expand the evaluation in this section we present experiments based on Fat-tree topologies and compare the results obtained with the mathematical model.

A third monitoring topology is built based on a Fat-tree topology consisting of three levels of switches [25]. To each switch is added one aggregator as deployed in Topology 1 and Topology 2. This monitoring topology is depicted in Figure 3, and it is defined as Topology 3. Topology 3 is evaluated to timeliness based on the addition of pods, and it was extended to 4, 6, 8, 10, and 12 pods (*e.g.,* K= 4). Moreover, Topology 3 was evaluated to both 120 and 150 bytes of monitoring data, and frequency sampling of 1 second as performed in the previous topologies to maintain consistency.

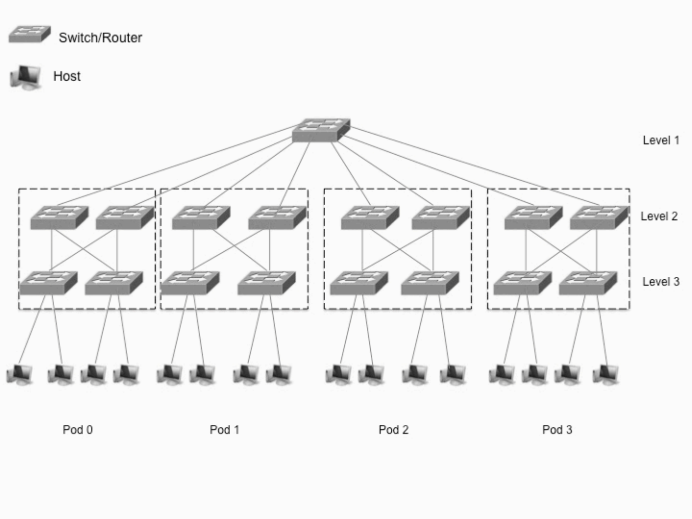


Figure 3. Topology 3, a monitoring topology based on a Fat-tree topology (k=4)

To organize the comparison between simulation and the mathematical model to Fat-tree topologies we define the results as follow:

- Predicted FAT-RT: It is the result for response time (RT) obtained by the mathematical model to Fat-tree.
- Average FAT-RT: It is the result for average response time (RT) obtained by the simulation to Fat-tree.

Table V and Table VI present the results to predicted FAT-RT and average FAT-RT for the amount of monitoring data for 120 bytes and 150 bytes with frequency sampling based on an interval of 1 second. Otherwise to conventional data center networks experiments, the experiments based on Fat-tree topologies does not exceed the time of 1 second to response time, so there is no need to extend the experiments to this topology to 10 seconds as we have done for topologies 1 to 4.

Table V compares results between predicted FAT-RT and average FAT-RT for 120 bytes of monitoring data with interval sampling of 1 second. In Table V, predicted FAT-RT shows results that are useful to estimate the mutual influence between timeliness and scalability when considering - 6.72% as a maximum margin of error. Such margin of error is the result to the topology that has 4 pods (K= 4), which is the topology with fewer pods. On the other side, predicted FAT-RT has a minimum margin of error when considering 12 pods (K= 12) which is the topology with more pods.

Table V. PREDICTED FAT-RT AND AVERAGE FAT-RT FOR 120 BYTES OF MONITORING DATA WITH INTERVAL SAMPLING OF 1 SECOND.

Topology/pods	Predicted fat-rt (ms)	Average fat-rt (ms)	Deviation(%)
Topology 3/k=4	34.92	32.57	- 6.72%
Topology 3/k=6	74.76	70.17	- 6.13%
Topology 3/k=8	125.80	118.72	- 5.63%
Topology 3/k=10	199.45	189.60	- 4.94%
Topology 3/k=12	284.30	272.27	- 4.23%

Table VI. PREDICTED FAT-RT AND AVERAGE FAT-RT FOR 150 BYTES OF MONITORING DATA WITH INTERVAL SAMPLING OF 1 SECOND.

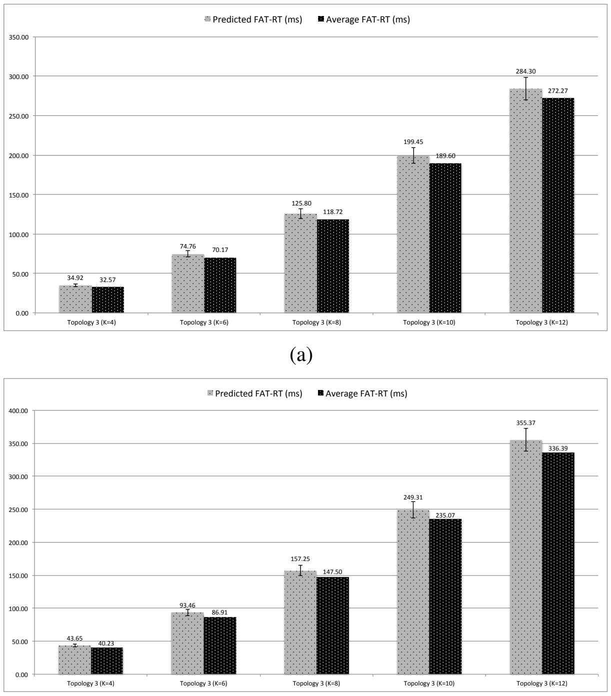
Topology/pods	Predicted fat-rt (ms)	Average fat-rt (ms)	Deviation(%)
Topology 3/k=4	43.65	40.23	- 7.84%
Topology 3/k=6	93.46	86.91	- 7.01%
Topology 3/k=8	157.25	147.50	- 6.20%
Topology 3/k=10	249.31	235.07	- 5.71%
Topology 3/k=12	355.37	336.39	- 5.34%

Moreover, we highlight two significant issues from the Table V. First, the margin of error was reduced, when more pods were added. In other words, when the monitoring topology escalates the margin of error decreases, which is significant to clouds that are environments usually composed of a plethora of probes and managers. Second, the margin of error is negative to Fat-Tree topologies which mean that the mathematical model can be useful to support the development of SLA's based on timeliness and scalability. It happens because the estimative provided by the mathematical model (Predicted FAT-RT) is always a bigger value when comparing to Average FAT-RT; thereby it is useful to avoid SLA breaches since it provides a big margin of tolerance.

Results demonstrate that the behaviour of the mathematical model is compatible with the simulation as observed in Figure 4 (a) which bars to predict FAT-RT considers 6% as a margin of error. Such results demonstrate that the mathematical model is useful to 120 bytes of monitoring data when considering 8 pods or more because of the margin of error assumed. On the other hand, monitoring topologies with fewer pods present values that extrapolate the margin of error. However, the absolute values are closed owing to the small size of the structure; thereby the results are useful as a reliable reference.

Table VI compares results between predicted FAT-RT and average FAT-RT for 150 bytes of monitoring data with interval sampling of 1 second. Table VI shows that the results to predicted FAT-RT are useful to estimate the mutual influence between timeliness and scalability when considering - 7.84% as a maximum margin of error. This margin of error is the result to the topology that has 4 pods (K= 4), which is the topology with fewer pods. On the other side, predicted FAT-RT has a minimum margin of error when considering 12 pods (K= 12) which is the topology with more pods.

Moreover, we highlight two significant findings from the Table VI. First, the margin of error was reduced, when more pods were added. In other words, when the monitoring topology escalates the margin of error decreases. Second, the margin



(a)

(b)

Figure 4. Comparison between predicted and average response time to 120 and 150 bytes of monitoring data with interval sampling to 1 second to Fat-Tree Topologies. (a) 120 bytes . (b) 150 bytes.

of error is negative to Fat-Tree topologies. The behaviour is practically the same if comparing to experiments performed to 120 bytes (Table V). In this sense, we highlight that the amount of monitoring data accentuates the difference as observed when comparing Table V and Table VI. For example, in Topology 3 with 12 pods (K= 12), the difference is - 4.23% and - 5.34% respectively to 120 bytes and 150 bytes of monitoring data.

Results demonstrate that the behaviour of the mathematical model is compatible with the simulation as observed in Figure 4 (b) which bars to predict FAT-RT considers 6% as a margin of error. Such results demonstrate that the mathematical model is useful to 150 bytes of monitoring data when considering 10 pods (K= 10) or more because of the margin of error assumed. On the other hand, monitoring topologies with fewer pods present values that extrapolate the margin of error. However, the absolute values are closed owing to the small size of the structure; thereby the results are useful as a reliable reference.

IV. Applying the Mathematical Model

The model to predict the mutual influence between timeliness and scalability introduced in this paper is helpful not only for allowing a better understanding of how this two metrics affect each other but also for serving as basis for future research in this field. It can be used as a support tool towards enhancing SLAs and billing, saving energy, raising profits to cloud operators, and reducing costs to customers.

Clouds have relied on SLAs to regulate the commercial relation between cloud operators and customers. Cloud monitoring is essential to certify that the SLA accomplishment is fair to both cloud operators and customers. In this scenario, the acknowledgement of the mutual influence between timeliness and scalability provided by the model supports significant issues such as assisting cloud operators to satisfy SLAs based on response time (timeliness), assisting infrastructure providers to assess its structures, assisting cloud operators to fulfil SLA without being invasive, and avoiding SLA breaches to reduce penalties. Timeliness is closely related to the accomplishment of SLAs based on response time. Predicting the influence of scalability over timeliness is important to avoid the development of weak SLAs that are vulnerable to the effects of scalability over response time and, as a consequence, may not be suitable to clouds.

Cloud resource (particularly virtual machine) migration is a meaningful cloud monitoring requirement to save energy, and it is directly affected by timeliness. VM migration is significant because it changes the location of computational resources according to the goals of a specific application or system. Migration depends on timeliness to work properly because the monitoring data is useful to migration only when it is timely. It allows for cloud operators to perform actions in time to either correct or adjust the deployment of resource to save energy. In this scenario, predictions based on timeliness and scalability are key to provide support towards definition of placement of resources to save energy because such predictions can estimate what size and architecture of network topology that is more adequate to migrate resources.

V. Conclusion and Future Works

This paper proposed a mathematical model to estimate the mutual influence between timeliness and scalability in cloud monitoring systems. This mathematical model is a step forward in cloud monitoring and, as a consequence, in cloud management because it provides means to cloud operators (e.g., CPs, InPs) to enhance their services. In this context, cloud operators may use the estimation provided by the mathematical model to raise its profits, improve its quality of service and use it to develop fair SLAs based on timeliness, for example.

Aiming to demonstrate the mathematical model effectiveness to conventional data center networks, a comparison was performed comparing results provided by the mathematical model (predicted RT) and the results obtained by a simulation (average RT). Such results revealed some important issues such as the solution is useful, and it presents accurate results to larger and deep topologies, and in smaller topologies the mathematical model provides meaningful outcomes.

In a second moment, a comparison was performed to monitoring topologies based on Fat-tree. It compares results provided by the mathematical model (predict FAT-RT) and the results obtained by a simulation (average FAT-RT). Even that the results have not had the precision demonstrated for conventional data center networks, they confirmed that the mathematical model based on monitoring metrics is useful to another kind of monitoring topology commonly used in clouds.

As future works, we aim to follow two research directions. First, we plan to study the economic impact of the mathematical model in clouds based on the increase of the profits to cloud operators (e.g., cloud providers, cloud infrastructures). Second, we will investigate the mutual influence between others cloud monitoring requirements such as scalability and comprehensiveness, timeliness and adaptability, timeliness and accuracy. It will be helpful to the development of comprehensive cloud monitoring systems based on cloud monitoring requirements.

References

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Gener. Comput. Syst.*, vol. 25, no. 6, pp. 599–616, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.future.2008.12.001

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, apr 2010.

[3] P. M. Mell and T. Grance, "Sp 800-145. the nist definition of cloud computing," Gaithersburg, MD, United States, Tech. Rep., 2011.

[4] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.

[5] C. Xu, J. Yang, X. Ling, Y. Wang, and L. Li, "Architecture design for management as a service cloud," in *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*, May 2013, pp. 860–863.

[6] R. Mijumbi, J. Serrat, and J.-L. Gorricho, "Self-managed resources in network virtualisation environments," in *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, May 2015, pp. 1099–1106.

[7] G. Da Cunha Rodrigues, R. N. Calheiros, V. T. Guimaraes, G. L. d. Santos, M. B. de Carvalho, L. Z. Granville, L. M. R. Tarouco, and R. Buyya, "Monitoring of cloud computing environments: Concepts, solutions, trends, and future directions," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, ser. SAC '16. New York, NY, USA: ACM, 2016, pp. 378–383. [Online]. Available: http://doi.acm.org/10.1145/2851613.2851619

[8] S. Meng and L. Liu, "Enhanced monitoring-as-a-service for effective cloud management," *IEEE Transactions on Computers*, vol. 62, no. 9, pp. 1705–1720, 2013. [Online]. Available: http://dx.doi.org/10.1109/TC.2012.165

[9] G. Rodrigues, V. Guimaraes, G. Santos, L. Tarouco, and L. Granville, "Network and services monitoring: A survey in cloud computing environments," in *Proceedings of the 11th International Conference on Networks*, ser. ICN '12, 2012, pp. 7–13.

[10] M. Barbosa de Carvalho, R. Pereira Esteves, G. da Cunha Rodrigues, C. Cassales Marquezan, L. Zambenedetti Granville, and L. Rockenbach Tarouco, "Efficient configuration of monitoring slices for cloud platform administrators," in *Computers and Communication (ISCC), 2014 IEEE Symposium on*, June 2014, pp. 1–7.

[11] S. De Chaves, R. Uriarte, and C. Westphall, "Toward an architecture for monitoring private clouds," *Communications Magazine, IEEE*, vol. 49, no. 12, pp. 130 –137, december 2011.

[12] P. Wieder, J. M. Butler, W. Theilmann, and R. Yahyapour, *Service Level Agreements for Cloud Computing*. Springer Publishing Company, Incorporated, 2011.

[13] V. C. Emeakaroha, M. A. S. Netto, R. N. Calheiros, I. Brandic, R. Buyya, and C. A. F. De Rose, "Towards autonomic detection of sla violations in cloud infrastructures," *Future Gener. Comput. Syst.*, vol. 28, no. 7, pp. 1017–1029, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.future.2011.08.018

[14] J. Simao and L. Veiga, "Flexible slas in the cloud with a partial utility-driven scheduling architecture," in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, vol. 1, Dec 2013, pp. 274–281.

[15] A. Antonescu and T. Braun, "Improving management of distributed services using correlations and predictions in sla-driven cloud computing systems," in *2014 IEEE Network Operations and Management Symposium, NOMS 2014, Krakow, Poland, May 5-9, 2014*, 2014, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/NOMS.2014.6838320

[16] M. Ribas, C. Furtado, G. Barroso, A. Lima, N. Souza, and A. Moura, "Modeling the use of spot instances for cost reduction in cloud computing adoption using a petri net framework," in *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, May 2015, pp. 1428–1433.

[17] D. Ardagna, G. Casale, M. Ciavotta, J. F. Pérez, and W. Wang, "Quality-of-service in cloud computing: modeling techniques and their applications," *J. Internet Services and Applications*, vol. 5, no. 1, pp. 11:1–11:17, 2014. [Online]. Available: http://dx.doi.org/10.1186/s13174-014-0011-3

[18] G. Aceto, A. Botta, W. De Donato, and A. Pescapè, "Survey cloud monitoring: A survey," *Comput. Netw.*, vol. 57, no. 9, pp. 2093–2115, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2013.04.001

[19] S. Clayman, A. Galis, C. Chapman, G. Toffetti, L. Rodero-Merino, L. M. Vaquero, K. Nagin, and B. Rochwerger, "Monitoring service clouds in the future internet." in *Future Internet Assembly*, G. Tselentis, A. Galis, A. Gavras, S. Krco, V. Lotz, E. P. B. Simperl, B. Stiller, and T. Zahariadis, Eds. IOS Press, 2010, pp. 115–126. [Online]. Available: http://dblp.uni-trier.de/db/conf/fia/fia2010.html#ClaymanGCTRVNR10

[20] S. Clayman, A. Galis, and L. Mamatas, "Monitoring virtual networks with lattice," in *Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP*, april 2010, pp. 239 –246.

[21] J. Montes, A. Sánchez, B. Memishi, M. S. Pérez, and G. Antoniu, "Gmone: A complete approach to cloud monitoring," *Future Gener. Comput. Syst.*, vol. 29, no. 8, pp. 2026–2040, Oct. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.02.011

[22] G. da Cunha Rodrigues, R. N. Calheiros, M. Barbosa de Carvalho, C. Raniery Paula dos Santos, L. Zambenedetti Granville, L. Rockenbach Tarouco, and R. Buyya, "The interplay between timeliness and scalability in cloud monitoring systems," in *IEEE Symposium on Computers and Communications(ISCC), 2015, International Conference on*, Jul 2015, pp. 776–781.

[23] G. da Cunha Rodrigues, G. Lessa dos Santos, V. Tavares Guimaraes, L. Zambenedetti Granville, and L. Rockenbach Tarouco, "An architecture to evaluate scalability, adaptability and accuracy in cloud monitoring systems," in *Information Networking (ICOIN), 2014 International Conference on*, Feb 2014, pp. 46–51.

[24] M. Fiorani, S. Aleksic, P. Monti, J. Chen, M. Casoni, and L. Wosinska, "Energy efficiency of an integrated intra-data-center and core network with edge caching," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 6, no. 4, pp. 421–432, April 2014.

[25] M. Bari, R. Boutaba, R. Esteves, L. Granville, M. Podlesny, M. Rabbani, Q. Zhang, and M. Zhani, "Data center network virtualization: A survey," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–20.

[26] Y. Sun, M. Chen, Q. Liu, and J. Cheng, "A high performance network architecture for large-scale cloud media data centers," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 1760–1766.

[27] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: A scalable and flexible data center network," in *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication*, ser. SIGCOMM '09. New York, NY, USA: ACM, 2009, pp. 51–62. [Online]. Available: http://doi.acm.org/10.1145/1592568.1592576

[28] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, Aug. 2008. [Online]. Available: http://doi.acm.org/10.1145/1402946.1402967

[29] Z. Guo and Y. Yang, "On nonblocking multicast fat-tree data center networks with server redundancy," *Computers, IEEE Transactions on*, vol. 64, no. 4, pp. 1058–1073, April 2015.

[30] D. Li, C. Guo, H. Wu, K. Tan, Y. Zhang, and S. Lu, "Ficonn: Using backup port for server interconnection in data centers," in *INFOCOM 2009, IEEE*, April 2009, pp. 2276–2285.