Analyzing the IT Subsystem Failure Impact on Availability of Cloud Services

Guto Leoni Santos*, Patricia Takako Endo†, Glauco Gonçalves‡, Daniel Rosendo*, Demis Gomes*,
Judith Kelner*, Djamel Sadok* and Mozghan Mahloo§
*Federal University of Pernambuco, Recife, Brazil
Networking and Telecommunication Research Group (GPRT)
E-mail: {guto.leoni, daniel.rosendo, jk, jamel}@gprt.ufpe.br
†University of Pernambuco, Caruaru, Brazil
E-mail: patricia.endo@upe.br
‡Federal Rural University of Pernambuco, Recife, Brazil
E-mail: glauco.goncalves@ufrpe.br
§Ericsson Research, Stockholm, Sweden
E-mail: mozhgan.mahloo@ericsson.com

***Abstract*—Cloud computing has gained popularity in recent years due to its pay-as-you-go business model, high availability of services, and scalability. Service unavailability does not affect just user experience but is also translated into direct costs to cloud providers and companies. Part of the costs is due to SLA breaches, once interruptions time greater than those signed in the contract generate financial penalties. Thus, cloud providers have tried to identify failure points and have estimated the availability of their services. This paper proposes models to assess the availability of services running in a cloud data center infrastructure. The models follow the TIA-942 standard. We propose Tier I and IV models using the Reliability Block Diagram (RBD) to allow modeling of different types of applications, and Stochastic Petri Net (SPN) to represent the failure behavior of information technology (IT) components. We perform stationary analysis to measure the service availability, and sensitivity analysis to understand which metrics have major impacts on data center availability.**

I. Introduction

Cloud providers have gained popularity because they changed the traditional business models, replacing a huge initial investment with a pay-as-you-go model, in which users can deploy their applications with guarantees of high availability, scalability, and security. Currently, one of the biggest challenges for cloud providers is SLA (Service Level Agreement) violation. This is tightly connected to the failure management of data centers (across the whole stack, from hardware to software). Unplanned data center failures are expensive (for both sides, providers and users) and require special attention. According to [1], the average cost of a data center outage has steadily increased from $505,502 in 2010 to $740,357 in 2016. Beyond the direct financial costs, these failures also result in business disruption, lost revenue, diminished end-user productivity, and a blow to business reputation.

Therefore, cloud providers have become increasingly interested in understanding the operation of their data center infrastructure, in order to identify failure points and estimate the availability of their services. Several techniques can be used for building estimation models, such as Petri Nets, Markov Chains, and Fault Trees, among others ([2], [3], [4], [5] and [6]).

At a high level of granularity, a data center can be divided into three major subsystems: information technology (IT) infrastructure, power system and cooling system [7]. These subsystems are interdependent and an interruption of one subsystem impacts on the availability of others. The IT infrastructure basically comprehends processing, storage, and networking hardware and software [7] and can be considered as the main subsystem of a data center. According to [1], the IT downtime costs approximately $5,600 per minute.

This paper proposes models to assess the availability of a service running in a cloud data center infrastructure. We use Reliability Block Diagram (RBD) to model different types of applications, whereas we use Stochastic Petri Net (SPN) to represent failure behavior of IT components. Despite some existing works, our contribution is focused on scalable models based on existing standards. We also provide a sensitivity analysis to understand which components are involved and how they impact on data center availability.

The rest of this paper is organized as follows: Section II presents some basic concepts needed to understand our proposal, such as RBD, SPN and cloud data center infrastructure; Section III presents some related work; Section IV describes our models regarding Tier I and Tier IV; in Section V, we present availability and sensitivity analysis of our models, and a discussion about results; and finally, in Section VI, we conclude the work and delineate some future works.

II. Background

This section describes briefly some basic concepts needed to understand our proposal.

A. Reliability Block Diagram (RBD)

RBD is a graphical representation of a system's success logic using block structures [8]. RBD can be evaluated using analytical methods to obtain system reliability and availability.

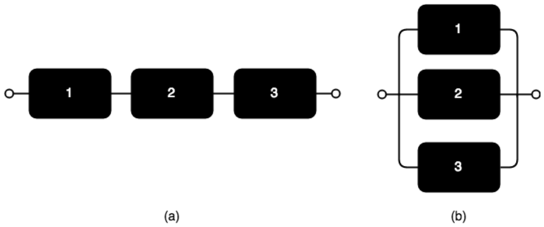

Fig. 1. RBD configurations

When all components of the system are strictly required for its operation, a failure of one of them causes the overall system to fail. In this case, the components are arranged in series as shows Figure 1.a. On the other hand, if the isolated failure of one component does not interrupt or shut down the system (i.e., the component is redundant), the blocks are disposed in parallel, as shown in Figure 1.b.

The availability can be defined as service uptime over total service time, where total time is described as the sum of service uptime and service downtime. These concepts can be associated with the average behavior of the system for the purpose of availability calculation.

Let's consider the series configuration with N components (Figure 1.a). The availability of each component, A_x, is calculated by division of the MTTF (Mean Time To Failure) and the MTBF (Mean Time Between Failures) of each component (Eq. 1). The MTBF also is defined as the sum of MTTF and MTTR (Mean Time to Repair), indicating the time between the detection of a failure and the detection of the next failure.

$$A_x = \frac{MTTF_x}{MTBF_x} = \frac{MTTF_x}{MTTF_x + MTTR_x} \tag{1}$$

In this way, the availability of the overall system, A_s, can be calculated as shown in Eq. 2.

$$A_s = \prod_{x=0}^{N} A_x \tag{2}$$

Considering that reliability, $R(t)$, is defined as Eq. 3, the reliability of overall system, $R_s(t)$ can be calculated as shown in Eq. 4.

$$R_x(t) = \mathrm{e}^{-\lambda_x t} \tag{3}$$

Where, λ means the failure rate.

$$R_s(t) = \prod_{x=0}^{N} R_x = \prod_{x=1}^{N} \mathrm{e}^{-\lambda_x t} \tag{4}$$

Then, knowing the λ_x values, we can calculate the $MTTF_s$ of the overall system following the Eq. 5.

$$MTTF_s = \frac{1}{\lambda_s} = \frac{1}{\sum\limits_{x=1}^{N} \lambda_x} \tag{5}$$

And, finally, we can calculate the MTTR of the system following the Eq. 1, considering the availability and MTTF of the overall system, instead an isolated component.

RBDs are commonly used due to their simplicity, but they are not suitable to model behavioral aspects of a system [2]. In this way, Petri Nets can be used with RBDs in order to have a set of comprehensive models addressing aspects of availability in cloud data centers.

B. Petri Nets

Graphically, a Petri Net is composed of circles (white circles represent places, and black circles represent tokens), rectangles (transitions), and arcs (see Figure 2). Places describe passive components (Figure 2.a), while transitions are active ones. There are two types of transitions: timed (Figure 2.b) and immediate (Figure 2.c). The timed transition is activated through a time parameter that follows a distribution, generally exponential. Immediate transitions are activated instantly. Case two immediate transitions be able at same time, priority can be used: the first transition to be triggered has a priority two, while the last has priority one. A Petri Net composed only of timed stochastic transitions is called a Stochastic Petri Net (SPN).

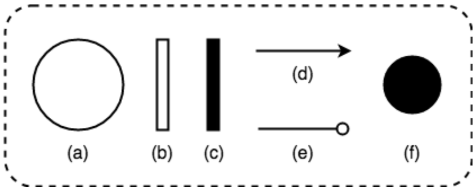

Fig. 2. Petri Net components

Regarding transitions' concurrency, it can be classified as single server or infinite server. When a transition has single server policy, the transition fires one token at a time, while infinity server policy can be understood as having an individual transition for each set of input tokens, all running in parallel [9].

C. Data Center Standardization

Data center standards define fundamental aspects, best practices, and recommendations regarding data center design and infrastructure. According to them, a generic data center system is basically composed of three subsystems: i) power infrastructure, ii) cooling infrastructure, and iii) IT infrastructure, whose dependency relations are shown in Figure 3.

These standards also define a classification that allows comparing data centers according to their availability. Such classification may be based on Tiers (ITU and TIA-942 standards) or Classes of availability (BICSI-002 and EN-50600 standards). In this paper, we focus on the IT infrastructure and our models are based on the TIA-942 Tier classification.

Basically, a tier is different from another tier due to the number of redundant components (i.e., N means no redundancy; and $N+1$ means component redundancy) and distribution paths (i.e. single or multiple paths that may be

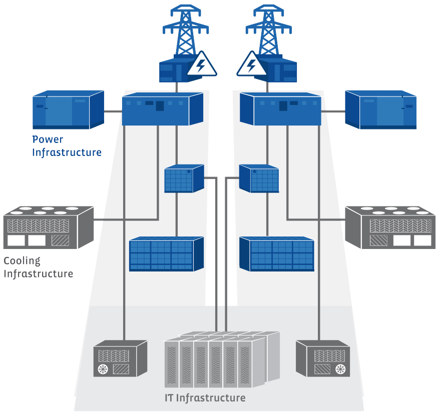

Fig. 3. Cloud Data Center Subsystems

active or passive). Tier classification goes from I to IV and higher tiers inherit requirements of lower tiers and are less susceptible to system disruptions (Tier II), may avoid system disruptions (Tier III), or are fault tolerant (Tier IV). Higher tiers provide greater availability, which results in higher costs and operational complexities. Therefore, the tier selection depends on the business requirements, such as minimum service availability, employment costs, and downtime financial consequences.

A data center IT subsystem is basically composed of servers, storage, and network components. The storage is illustrated as Network Attached Storage (NAS) Disk Array. Network is represented by Edge, Core, Aggregation routers, and Access switch. The Storage Area Network (SAN) is a network component used in Tier IV to connect array disks to servers.

Figures 4 and Figure 5, depict a Tier I and Tier IV data center IT subsystem, respectively. As one may note, the IT subsystem Tier I does not present any component redundancy, being susceptible to system disruptions from planned and unplanned activities. On the other hand, Tier IV is a fully redundant architecture ($2(N+1)$).

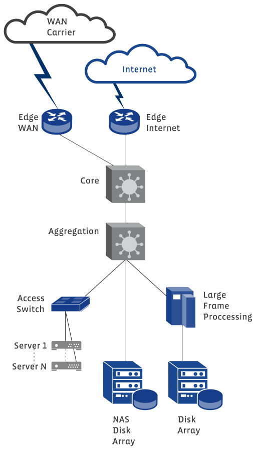

Fig. 4. IT architecture - Tier I

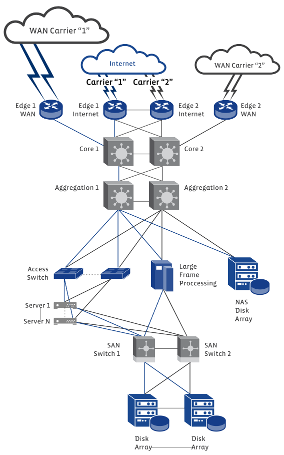

Fig. 5. IT architecture - Tier IV

D. Sensitivity Analysis

The sensitivity analysis is defined by [10] as *"the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input"*.

tAccording to [11], the technique of sensitivity analysis by means of percentage difference consists of changing one parameter over a list of values, while holding the other parameters fixed, and calculating the percentage difference (SI) in the output metric considered. One performs this step for each parameter in the list, and sorts them from the highest difference (D_{max}) to the lowest (D_{min}). The formula for obtaining the percentage difference is shown in Equation 6:

$$SI = \frac{D_{max} - D_{min}}{D_{max}} \tag{6}$$

III. Related Work

A digital library service inside the data center infrastructure is modeled in [12]. The service architecture is composed of a front-end and a node. The front-end has cloud management components, and the node runs the digital library application. A RBD was used to represent the dependability between components of the application architecture, and an SPN was used to model the components redundancy. Results show that the cold-standby redundancy achieves 2.65 number of nines of availability and the hot-standby 2.66. When cold-standby and hot-standby redundancy were considered for the front-end, the system reached 4.1 and 5 number of nines of availability, respectively.

In [13], authors propose an availability model of a Eucalyptus cloud environment that runs a video stream application. To estimate availability, authors use RBD and Markov Chains. The RBD models the components of Eucalyptus architecture, while the Markov Chain models the behavior of the stream service. The results of sensitivity analysis show that the repair rate of the front-end module is the most important parameter with respect to the availability.

A model for evaluating availability of private clouds is proposed by [14] and they use RBD and Markov Chains. The architecture is based on Eucalyptus, and employs warm-standby in the main components. RBD is used to model the dependency between components, while Markov Chains model the redundant behavior of cloud components. Authors evaluate three architectures with one, two and three clusters, achieving 99.9938749%, 99.9969376% and 99.9969377% of availability, respectively.

Our work differs from the literature because we propose a more detailed and scalable IT subsystem model (other works are focused only on the software level), and we are also considering data center standards to guide our models.

IV. IT Subsystem Model

In this work, we are considering an application running on an IT infrastructure of a cloud data center. For our analysis, we are taking into consideration the Tier I and IV to evaluate the service availability.

The IT infrastructure is composed of network and storage components, and servers. The network and storage components are modeled using SPN, in order to represent the failure and repair behavior. To model servers, we use RBD, in order to represent the dependency relationship between the server components. These models are described next.

A. RBD model of server

The Figure 6 shows the RBD model that represents the server components. This system is composed of four serial components: hardware (HW), operating system (OS), virtual machine (VM) and the application (APP) instance that is running on this server.

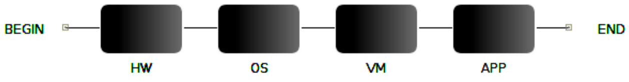

Fig. 6. RBD model of server components, based from [12]

The availability level, and MTTF and MTTR values are calculated as described in subsection II-A.

B. SPN Model of Tier I

Figure 7 shows our SPN model of Tier I. We disregard the large frame processing and disk array components, because they are used as backup, and so their failure does not impact on service availability. Edge router WAN was also not taken into account, because we are considering the Internet access.

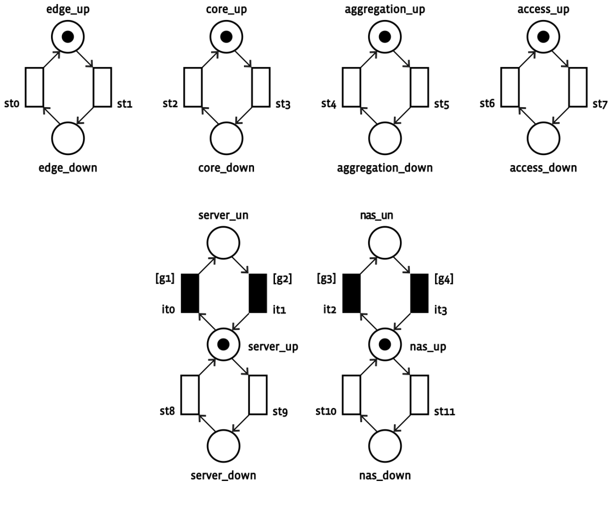

Fig. 7. SPN model of IT infrastructure - Tier I

Each component of the data center network is modeled as a building block, that is, a set of two places (up and down) and two transitions (failure and repair). For instance, the edge router building block has two places, named $edge\_up$ and $edge\_down$, that represent when the edge router is running and when it has failed, respectively. The transition $st0$ represents the MTTF of the edge router, and $st1$ represents its MTTR. Other network components are similar, each one with their respective building blocks, with MTTF and MTTR values.

Storage and server are composed of one more place and two immediate transitions. For instance, considering servers, the place $server\_up$ represents a server running, while the place $server\_down$ indicates that a server is down. Stochastic transitions model components failure and repair, similarly to the network components. When one or more network component fails, the immediate transition $it0$ fires, the token present in $server\_up$ is consumed, and a token is produced in the $server\_un$ place, indicating the server unavailability. When the failed network component is repaired, the immediate transition $it1$ fires, and a token returns to the $server\_up$ place, indicating that the server is available again. The storage has a similar behavior.

Immediate transitions (from $it0$ to $it3$) have guard functions to assure the behavior described above. These guard functions are presented in Table I. As one can observe in Figure 4, the components connected to the server are edge router, core router, aggregation router, and access switch. So if one of these components fails, the server will be unavailable. This behavior is guaranteed in guard function $g1$. However, if all components are up, the server is available, and it is modeled by the guard function $g2$. The transitions $it2$ and $it3$ are similar, but only edge, core and aggregation routers are connected to the storage component. So if one of these components fails, the storage will be unavailable. This behavior is modeled in guard function $g3$. The guard function $g4$ models these components' repair, allowing storage to become available.

TABLE I
GUARD FUNCTIONS OF IMMEDIATE TRANSITION OF SPN - TIER I

Identification	Guard Function
$g1$	$((\#edge\_up = 0)OR(\#core\_up = 0)$ $OR(\#aggregation\_up = 0)OR(\#access\_up = 0))$
$g2$	$((\#edge\_up > 0)AND(\#core\_up > 0)$ $AND(\#aggregation\_up > 0)AND(\#access\_up > 0))$
$g3$	$((\#edge\_up = 0)OR(\#core\_up = 0)$ $OR(\#aggregation\_up = 0))$
$g4$	$((\#edge\_up > 0)AND(\#core\_up > 0)$ $AND(\#aggregation\_up > 0))$

The system is considered available if there are one or more tokens in $server\_up$ and nas_up, i.e., if there are one or more servers and storage running. Then, the availability formula of Tier I, A_{tierI}, is shown in Eq. 7, and means the probability of having more than zero tokens in these two places ($server\_up$ and nas_up).

$$A_{tierI} = P\{(\#server\_up > 0)AND(\#nas\_up > 0)\} \quad (7)$$

C. SPN Model of Tier IV

In this model, large frame processing was disregarded, however different from Tier I, array disks are used to storage. The edge router WAN was disregarded too. As can be seen in Figure 5, all network and storage components are replicated in order to keep the data center available in case of unexpected failures. In addition, there are two switches that are connected to disk arrays, which are not in Tier I. The SPN model regarding Tier IV is presented in Figure 8. The model is similar to Tier I; however to model that redundant component we are using two tokens with infinite server transitions.

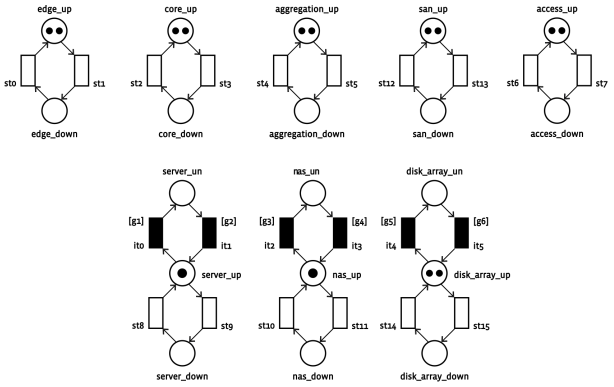

Fig. 8. SPN model of IT infrastructure - Tier IV

There are also components that represent the SAN switches and the new array disk. The SAN switch building block is composed of places san_up and san_down, while the array disks building block is composed of $disk\_array\_un$, $disk\_array\_up$ and $disk\_array\_down$.

The formula to calculate the availability is different to Tier I, because now it takes into consideration the array disk component. The availability of Tier IV, A_{tierIV} is defined in Eq. 8 and it means the probability of having more than zero tokens in these three places ($server\_up$, nas_up and $disk\_array\_up$).

$$A_{tierIV} = P\{(\#server\_up>0) AND ((\#nas\_up>0) \\ OR (\#disk\_array\_up>0))\} \quad (8)$$

Our Tier IV model uses all the transitions of the Tier I model, shown in Table IV, and additionally four stochastic transitions ($st12$, $st13$, $st14$, and $st15$), and two immediate transitions ($it4$ and $it5$). Transition $st12$ represents a failure of the SAN switch, while transition $st13$ represents the repair of this component. Transitions $st14$ and $st15$ are similar to other transitions that represent failure and repair behavior.

Furthermore, since places that represent components' availability have two tokens, we had to use an infinity server policy in stochastic transitions, in order to model independent behavior of redundant components.

The immediate transition $it4$ models the behavior when SAN switches are down, and then the array disks are also unavailable. It fires when there are no tokens in place san_up (guard function $g5$), that means that SAN switches are unavailable. The transition $it5$ fires when there is at least one token in place san_up (guard function $g6$). The guard functions of these transitions are presented in Table II. Guard functions $g1$ to $g4$ are the same as the Tier I model, presented in Table I.

TABLE II
GUARD FUNCTIONS OF IMMEDIATE TRANSITION OF SPN - TIER IV

Identification	Guard Function
$g5$	$(\#san\_up = 0)$
$g6$	$(\#san\_up > 0)$

V. Availability and Sensitivity Analysis

We performed stationary analysis using the Mercury tool[1] with the purpose of calculating the availability of models previously described.

Regarding the RBD model, the MTTF and MTTR values are presented in Table III. We are considering a digital library service (APP), but we can easily use other kind of applications here. From this RBD model, we calculate the MTTR and MTTF of the server, and these values are used in our SPN (specifically in transitions $st8$ and $st9$, respectively), in order to estimate the overall data center availability.

TABLE III
RBD PARAMETERS OBTAINED FROM [12]

Components	MTTF (in hours)	MTTR (in hours)
HW	8760	1.667
OS	1440	1
VM	1880	0.167
APP	6865.3	0.167

Table IV presents all MTTF and MTTR values of the stochastic transitions we used in our SPN models, including MTTR and MTTR obtained from our RBD model.

TABLE IV
PARAMETERS OF STOCHASTIC TRANSITIONS OBTAINED FROM [15], [16], AND [17]

Transition	Meaning	Value (in hours)
$st0$	Edge Router MTTF	796
$st1$	Edge Router MTTR	1
$st2$	Core Router MTTF	16243
$st3$	Core Router MTTR	0.78
$st4$	Aggregation Router MTTF	8247
$st5$	Aggregation Router MTTR	0.63
$st6$	Access Switch MTTF	13043.48
$st7$	Access Switch MTTR	0.35
$st8$	Server MTTF	768.35
$st9$	Server MTTR	0.7445
$st10$	NAS MTTF	1200000
$st11$	NAS MTTR	12
$st12$	SAN MTTF	255358
$st13$	SAN MTTR	7.66
$st14$	Disk Array MTTF	1200000
$st15$	Disk Array MTTR	12

Table V shows the availability level and the downtime of both Tier I and IV. Tier IV presents approximately 99.90% of availability, that is about 8.49 hours of downtime in a year, while the Tier I is only 99.76%, meaning 20.86 hour of downtime.

[1] http://www.modcs.org/?page_id=1397

TABLE V
AVAILABILITY EVALUATION

Tier	Availability (in %)	Downtime (in hours/year)
I	99.7618235	20.8642
IV	99.9030398	8.4937

The formula used to calculate the downtime, D, in hours/year, is shown in Eq. 9.

$$D = (1 - A) * 8760 \tag{9}$$

We also performed sensitivity analysis (through simulations) to verify which parameters affect more the overall data center availability. The setup simulation is presented in Table VI.

TABLE VI
SETUP SIMULATIONS

Parameter	Value
Confidence Level	95
Maximum relative error	10
Warm-up period	50
Batch size	50000

A. Sensitivity Analysis - Tier I

Table VII shows the sensitivity result of our Tier I SPN containing the three higher and lower indices. Parameters with values equal to zero are not considered.

TABLE VII
SENSITIVITY RANKING OF TIER I

Parameter	Sensitivity Index
$edge\_mttr$	2.50 x 10^{-3}
$server\_mttr$	3.31 x 10^{-4}
$edge\_mttf$	2.55 x 10^{-4}
$aggregation\_mttf$	6.38 x 10^{-6}
$core\_mttf$	4.68 x 10^{-6}
$access\_mttf$	1.83 x 10^{-6}

Results indicate that the edge router MTTR ($edge\_mttr$) has the greatest impact in the data center availability considering the Tier I. In other words, a variation in this value impacts more significantly on the availability. The second and third values that have more impact are the server MTTR ($server\_mttr$) and edge router MTTF ($edge\_mttf$), respectively. The metric with the smallest index was Access Switch MTTF ($access\_mttf$). When the value of this metric was changed from 13043.48h to 13943.48h, it was registered as a small availability variation (99.763308% to 99.762586%).

Figure 9 shows the Tier I data center availability when we varied the $edge\_mttr$ value. A variation of 2h in the $edge\_mttr$ results in an availability drop from 99.76% to 99.56%, that means a considerable downtime increase from 21.024h to 38.544h. Figure 10 shows that the availability decreases from 99.76% to 99.73%, when $server\_mttr$ increases from 0.7445h to 0.9945h. And Figure 11 shows the impact of the $edge\_mttf$ parameter; it is a MTTF value, its impact is positive and increases the overall availability.

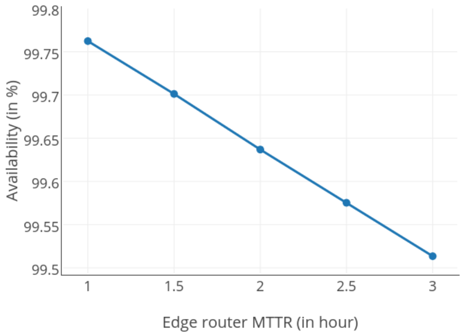

Fig. 9. Edge Router MTTR impact on Tier I availability

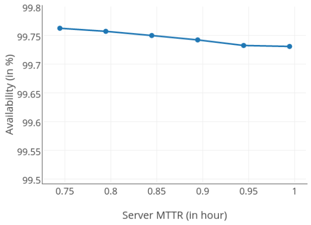

Fig. 10. Server MTTR impact on Tier I availability

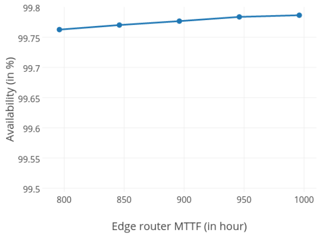

Fig. 11. Edge Router MTTF impact on Tier I availability

B. Sensitivity Analysis - Tier IV

Table VIII shows results of sensitivity analysis regarding Tier IV architecture with the three higher and lower indices. The result is different from Tier 1 because all architecture is duplicated; that influences which components are most critical for availability. The server MTTR ($server\_mttr$) has the greatest impact on data center availability; and the second and third values that have the most impact are server MTTF ($server\_mttf$) and edge router MTTR ($edge\_mttr$), respectively. The metric with the smaller sensitivity analysis index was SAN MTTR (san_mttr). When SAN MTTR value varied from 4.5h to 7.66h, there was a negligible variation of availability (99.903023% to 99.903038%).

TABLE VIII
SENSITIVITY RANKING OF TIER IV

Parameter	Sensitivity Index
$server\_mttr$	3.32 x 10^{-4}
$server\_mttf$	3.82 x 10^{-5}
$edge\_mttr$	1.25 x 10^{-5}
$core\_mttf$	2.47 x 10^{-10}
$access\_mttf$	1.90 x 10^{-10}
san_mttr	1.00 x 10^{-10}

As presented in Figure 12, a variation of 2 hours in $server\_mttr$ resulted in an availability decrease from

99.904% to 99.87%. On the other hand, an increase of 30 hours in $server\_mttf$ resulted in an availability increase from 99.903% to 99.907% (Figure 13). Figure 14 shows the impact of the Edge Router MTFF ($edge\_mttr$). The impact is negative, and decreases the availability of data center.

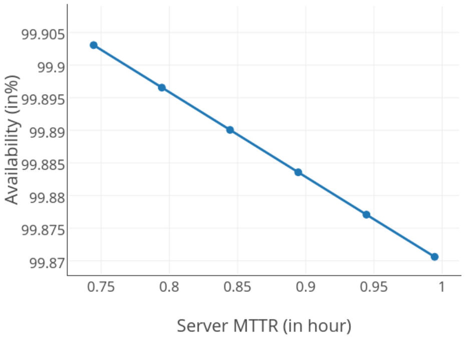

Fig. 12. Server MTTR impact on Tier IV availability

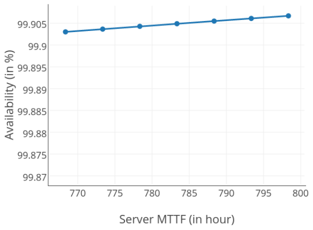

Fig. 13. Server MTTF impact on Tier IV availability

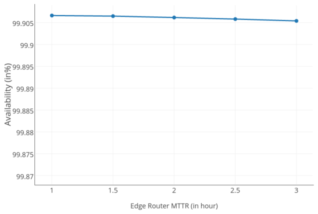

Fig. 14. Edge Router MTTR impact on Tier IV availability

C. Discussion

For the stationary analysis results, we noted that availability increased considerably, from 99.761% (Tier I) to 99.903% (Tier IV). This addition represents a downtime decrease from 20.86h to 8.49h per year, which can greatly impact on the application performance. So, to applications that need a minor unavailable time (e.g. critical applications), it is necessary an architecture with redundant components.

About the sensitivity analysis performed in architectures, the component that has more impact on availability was the edge router for Tier I, and the server for the Tier IV. It happen due to the redundancy of the components, which changes the critical fault points. In order to improve the architectures' availability, an investment can be made in these components, either in redundancy or in new equipment with greater reliability.

VI. Conclusion and Future Works

SLA violations and unavailability of services can be translated into direct costs to companies and cloud providers. For this reason, estimating the availability of the infrastructure of a data center can help cloud providers to minimize their costs.

We presented models based on SPN and RBD in order to estimate the availability of service running on IT infrastructure. Tiers I and IV of IT infrastructure was modeled. Experiments showed that availability increases from Tier I to Tier IV. With sensitivity analysis were evaluated the components that most impact the availability, and how a variation in parameters of components impact the availability of data center.

As future work we plan to integrate the IT subsystem with power and cooling subsystems and model other applications, with different architectures (such as multi-tier services).

Acknowledgment

This work was supported by the RLAM Innovation Center, Ericsson Telecomunicações S.A., Brazil. We would like to thank the design team of the Networking and Telecommunication Research Group (GPRT) for the support.

References

[1] P. Institute, "Cost of data center outages: Data center performance benchmark series." http://www.emersonnetworkpower.com/en-US/Resources/Market/Data-Center/Latest-Thinking/Ponemon/Documents/2016-Cost-of-Data-Center-Outages-FINAL-2.pdf/, 2016. Accessed: 2016-10-22.

[2] J. Dantas, R. Matos, J. Araujo, and P. Maciel, "Eucalyptus-based private clouds: availability modeling and comparison to the cost of a public cloud," *Computing*, vol. 97, no. 11, pp. 1121–1140, 2015.

[3] M. Miglierina, G. P. Gibilisco, D. Ardagna, and E. Di Nitto, "Model based control for multi-cloud applications," in *Modeling in Software Engineering (MiSE), International Workshop on*, pp. 37–43, IEEE, 2013.

[4] H. Khazaei, J. Mišić, V. B. Mišić, and N. B. Mohammadi, "Availability analysis of cloud computing centers," in *Global Communications Conference (GLOBECOM), IEEE*, pp. 1957–1962, IEEE, 2012.

[5] A. J. Gonzalez and B. E. Helvik, "Hybrid cloud management to comply efficiently with sla availability guarantees," in *Network Computing and Applications (NCA), 12th IEEE International Symposium on*, pp. 127–134, IEEE, 2013.

[6] M. Jammal, A. Kanso, P. Heidari, and A. Shami, "A formal model for the availability analysis of cloud deployed multi-tiered applications," in *Cloud Engineering Workshop (IC2EW), IEEE International Conference on*, pp. 82–87, IEEE, 2016.

[7] L. A. Barroso, J. Clidaras, and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 8, no. 3, pp. 1–154, 2013.

[8] A. K. Verma, A. Srividya, and D. R. Karanki, *Reliability and safety engineering*, vol. 43. Springer, 2010.

[9] G. Callou, D. Tutsch, J. Ferreira, J. Araújo, P. Maciel, and R. Souza, *A Petri net-based approach to the quantification of data center dependability*. INTECH Open Access Publisher, 2012.

[10] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons, 2004.

[11] D. Oliveira, "The mercury scripting language cookbook,"

[12] J. Araujo, P. Maciel, M. Torquato, G. Callou, and E. Andrade, "Availability evaluation of digital library cloud services," in *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on*, pp. 666–671, IEEE, 2014.

[13] R. Melo, M. C. Bezerra, J. Dantas, R. Matos, I. Melo, and P. Maciel, "Video on demand hosted in private cloud: Availability modeling and sensitivity analysis," in *Dependable Systems and Networks Workshops (DSN-W), IEEE International Conference on*, pp. 12–18, IEEE, 2015.

[14] R. Matos, J. Dantas, J. Araujo, K. S. Trivedi, and P. Maciel, "Redundant eucalyptus private clouds: Availability modeling and sensitivity analysis," *Journal of Grid Computing*, pp. 1–22, 2016.

[15] A. P. Guimaraes, P. Maciel, and R. MATLAs JR, "Design of it infrastructures of data centers: An approach based on business and technical metrics," *Quantitative Assessments of Distributed Systems: Methodologies and Techniques*, p. 265, 2015.

[16] Y. Yue, B. He, L. Tian, H. Jiang, F. Wang, and D. Feng, "Rotated logging storage architectures for data centers: Models and optimizations," *IEEE Transactions on Computers*, vol. 65, no. 1, pp. 203–215, 2016.

[17] B. Schroeder and G. A. Gibson, "Disk failures in the real world: What does an mttf of 1, 000, 000 hours mean to you?," in *FAST*, vol. 7, pp. 1–16, 2007.