Semi-Markov Availability Model for Infrastructure as a Service Cloud Considering Hidden Failures of Physical Machines

Oleg Ivanchenko[1], Vyacheslav Kharchenko[2], Yurij Ponochovny[3], Ivan Blindyuk[1], Oksana Smoktii[4]

[1] University of Customs and Finance, Dnipro, Ukraine
vmsu12@gmail.com, ivanblindyuk@mail.ru
[2] National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine
v_s_kharchenko@ukr.net
[3] Poltava National Technical University named after Yurij Kondratyuk, Poltava, Ukraine
pnch1@rambler.ru
[4] Vasyl' Stus Donetsk National University, Vinniza, Ukraine
oksana.smokty@gmail.com

Abstract. Results of researches in different areas of science and technique confirm that an effective route in order to solve important tasks is the possibility of migrating data resource within the cloud. Therefore Cloud Computing services including Infrastructure as a Service Cloud (IaaS Cloud) should have high availability level. It is serious problem, because even large cloud providers face with sudden failures of IaaS Cloud. It comes no surprise, that scientists consider taxonomy of this system on base of using overall underlying components, such as physical machines (PMs) and virtual machines (VMs). However cloud providers should always remember that hidden failures of PMs are one of the main causes of damage for their cloud assets. In this paper we propose approach on base of using Semi-Markov model in order to determine availability level for the IaaS Cloud with Technical State Control System.

Keywords: Infrastructure as a Service Cloud, sudden and hidden failures of physical machines, Semi-Markov availability model.

Key terms: Infrastructure, Mathematical Model, Development, Characteristic.

1 Introduction

Cloud Infrastructure is one of the most widely used model of Cloud Computing. Therefore modern large cloud providers such as Amazon, Microsoft, Google, Rackspace need approaches and models for the quantification of reliability level. In particular, Infrastructure as a Service (IaaS) Cloud provider's data centers try to ensure quality of service (QoS) by using different approaches for determining of their availability level. Significance of issue ensuring of availability of the IaaS Cloud can hardly be exaggerated.

Moreover additional incentive for cloud providers is transformation of cyber assets for several important Critical Infrastructures into cloud. According to researches by Cornell University and Washington State University, group of scientists have made efforts to develop software platform for Grid Smart energy infrastructure [1]. Amazon EC2 was used by them in order to perform the cloud computing needs for the Critical Energy Infrastructure.

At the same time nowadays we can describe situation, when number of physical machines for IaaS Cloud data centers are climbing fast and different scientists try to help providers control their availability level [2], [3]. Most of the scientists usually prefer to use Markov models in order to solve different tasks for concrete computer systems, including Cloud Infrastructures. In fact Continuous Time Markov Chains are main toolkit and predominate among the different mathematical models of availability and reliability for IaaS Cloud [4]. Stochastic Petri Nets [5] also featured heavily in the list with Reliability Block Diagrams [6], Fault Trees and Reliability Graphs all among the best techniques, which researches of cloud computing systems prefer to use.

However researchers have to take into consideration that these types of models need to describe different events for IaaS Cloud, including sudden and hidden failures, repairs and monitoring services. We can't afford to ignore issues, that to relate to the monitoring of technical and information states of different components for cloud infrastructures. Our researches show, that due to combination of deterministic and stochastic durations into working cycle of Infrastructure as a Service Cloud, this availability model can be presented by us as Semi-Markov model. We will also try to consider solutions for IaaS Cloud on base of benefits added by description of Semi-Markov process with special states. It gives us a way to conduct quite deep analysis of IaaS Cloud behavior in different negative situations, involving accidents of data centers, failures of physical and virtual machines or even DoS and DDoS attacks. At the same time we won't theorize approaches and techniques that related to the availability of cloud infrastructure. We will only consider concrete situation for the IaaS Cloud with definite number of PMs.

In this paper, we consider how to build Semi-Markov availability models for the IaaS Cloud with three pools of physical machines (PMs) and Technical State Control System. Note that, IaaS Cloud provider can substitute one machine for another in case definite PM is failed. PMs are grouped into three pools such as: hot, warm and cold pools [7], [8].

Rest parts of this paper are organized as follows. In Section 2 a special approach for availability analysis of IaaS Cloud with three multiple pools is described by us, considering sudden and hidden failures of PMs. Final Section 3 introduces the conclusions statement of our researches.

2 Statement of the Researches Results

2.1 Approach for Availability Analysis of IaaS Cloud

We will not try describing concrete architecture for IaaS Cloud, but we will consider that user has access to IaaS Cloud. At the same time we assume that IaaS Cloud consists of three pools of PMs. According to the research, three pools of PMs allow to reduce infrastructure cost, cooling cost and power consumption of IaaS Cloud [9].

The approach uses the means of failure detection that two main components could employ. First component is Resource Provisioning Decision Engine (RPDE) by which the pools implement capture the resource provisioning decision process [10]. As second component that are functioned in system, namely, Technical State Control System (TSCS), which is working in monitoring and diagnostic modes [8]. Perhaps inspired by necessity of through high availability level maintenance, cloud providers will get possibilities for effective repair and migration of physical resources (PMs). Figure 1 shows the taxonomy model for availability analysis of IaaS Cloud.

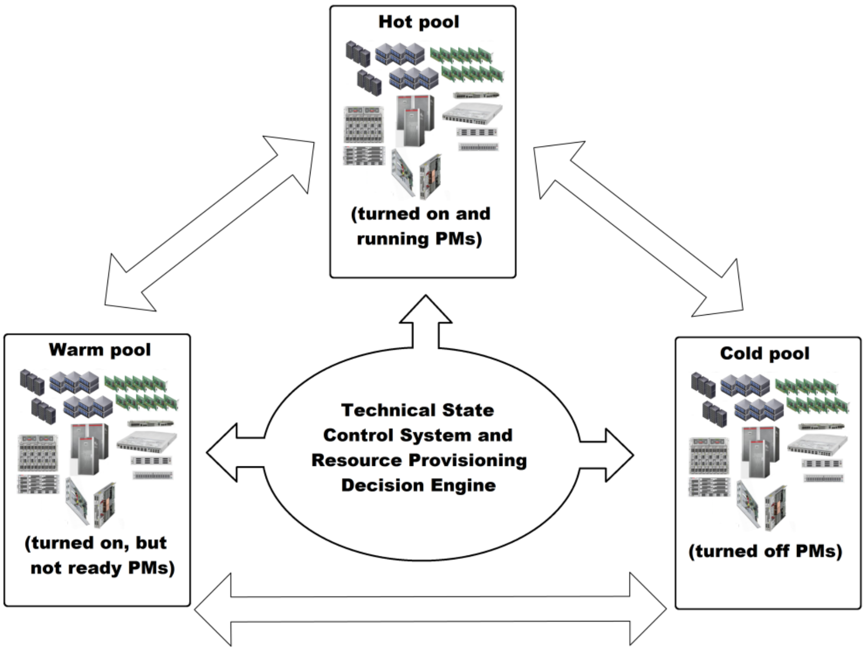

Hot pool
(turned on and running PMs)
Warm pool
(turned on, but not ready PMs)
Cold pool
(turned off PMs)
Technical State Control System and Resource Provisioning Decision Engine

Fig. 1. Taxonomy model for availability analysis of IaaS Cloud

As this taxonomy (Fig. 1) shows, that provider has to consider different regimes and operational time of TSCS in order to ensure high availability level of the IaaS Cloud. Furthermore, researchers must build mathematical availability models for IaaS Cloud considering different types of physical machine's failures. In [11], the authors presented data about several typical cloud services' downtime. In particular, they calculated that in 2011 year Amazon cloud services downtime equals about 8,5 days and the corresponding availability was about 97,67%. This information teaches us, how to analyze, what is causing the downtime of cloud services. Results of analysis have shown that hidden failures of PMs is one the most important causes of the IaaS Cloud downtime.

Turning to building of models, we will look at specific type of models that is Semi-Markov models with special states. Let's look now at overall methods of solution tasks to assess the IaaS Cloud availability level. Research has shown that in order to solve different tasks we propose to enhance the State Space Modeling Taxonomy [12], [13] with new type of Semi-Markov models. Semi-Markov regenerative process is described by this type of models. In fact we can characterize these models, that relate to the Semi-Markov birth-death processes.

So far, we have tried to use state-space models for monolithic cloud computing system, but not for separate components of IaaS Cloud. It was serious problem, because this model couldn't give accurate assessments for whole system. Now we can obtain the availability allocation for all IaaS Cloud and for concrete PM using Semi-Markov modeling approach, by which cloud provider determines possibilities of their own resources in working environment. As an illustrative example, sudden and hidden failures for IaaS Cloud with three pools of PMs and TSCS will be considered by us.

2.2 Analytical and Stochastic Availability Model for an IaaS Cloud Considering Hidden Failures of PMs

The case study chosen is a model of IaaS Cloud. Figure 2 shows finite graph for Semi-Markov model of the IaaS Cloud with three PMs.

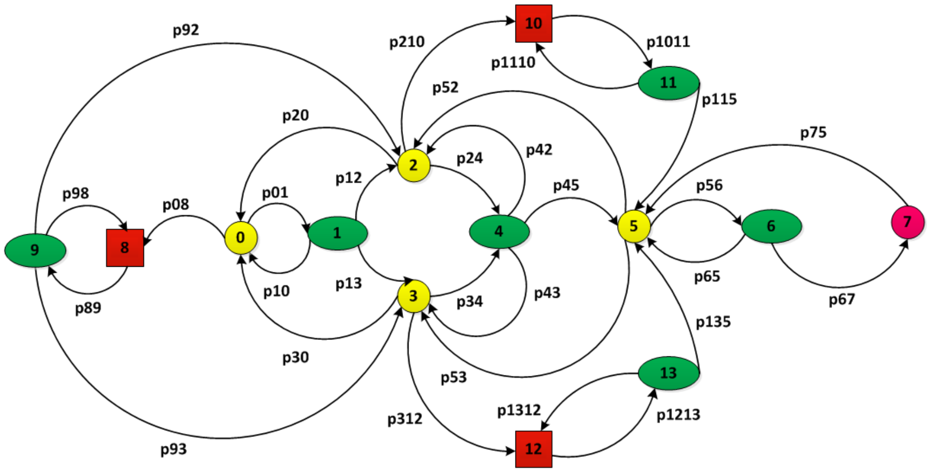


Fig. 2. Semi-Markov availability model of the IaaS Cloud with three PMs

In Fig. 2, if three PMs fail, the cloud computing system becomes unavailable. State S_7 is unavailable state of IaaS Cloud. Obviously the IaaS Cloud becomes available, when the model enters states S_0,S_2,S_3,S_5. In state S_0 three PMs are operational. At the same time states S_2 and S_3 are states with two operational PMs. Then state S_5 is state with one operational PM. From now available states are yellow, whereas unavailable states are red and states of TSCS are green.

Suppose our IaaS Cloud works throughout a particular time with operated duration $t \in [0,T]$. We use the following assumptions and limitations for our modeling process.

- Hot, warm, and cold PMs are identical PMs [4]. Provider can do replacement of failed hot PM by available warm or cold PM, respectively.
- Solution for this model can be obtained when replacement process of PMs is instantaneous. It means that we consider immediate transitions.
- IaaS Cloud provider can perform technical state control (CTS) of hot PM. The duration of this interval is τ_c.
- Overall effect several types of possible failures in PMs with an aggregated mean time to failure (MTTF) is considered here [14]. We also use assumptions that all times to failures for all PMs are exponentially distributed. Despite the fact that hot, warm and cold PMs have different operating time in order to simplify modeling process, we will suppose that MTTF $1/\lambda_s$ can be represented as equal values. Apparently, it is reasonable to consider case for three hot, warm and cold PMs ($n_h = n_w = n_c = 1$), when MTTF $1/\lambda_s = 1/\lambda_{sh} = 1/\lambda_{sw}$, where sudden failure rates λ_{sh} for hot PM and sudden failure rates λ_{sw} for warm PM.
- IaaS Cloud provider haven't enough time in order to perform repair operations for failed PMs. Therefore we need to take into consideration that all times to repair are not exponentially distributed. We use Erlang-k distribution in preference to exponential distribution, where $k = 2$ [15]. We also assume that mean time to repair (MTTR) of warm PM $1/\mu_w$ is higher than MTTR of hot PM $1/\mu_h$ by a factor of two.
- Specific feature of architecture for IaaS Cloud is implementation of migration process of PMs. We consider the migration operations of physical machines as operations to restore the working capacity of warm and hot PMs with repair rates μ_w and μ_h, respectively.
- We assume that hot and warm PMs can fail due to the occurrence of hidden failures with rates $\lambda_h = \lambda_{hh} = \lambda_{hw}$. In this case MTTF equals $1/\lambda_h = 1/\lambda_{hh} = 1/\lambda_{hw}$ for hot and warm PMs, respectively. Hidden unavailable hot or warm PM will repair after next CTS with rate λ_h^*.
- IaaS Cloud becomes unavailable when the SM model enters the state S_7.

According to the Fig. 2, our IaaS Cloud is processing workload into the states S_0 (at the initial moment $t = 0$), S_2, S_3 and S_5, that is available sub-set space S_A^1 for IaaS Cloud was created by states $S_A^1 = \{S_0,S_2,S_3,S_5\}$. Other states for IaaS Cloud can be described as: a) CTS sub-set space $S_{CTS}^{1}=\{S_1,S_4,S_6,S_9,S_{11},S_{13}\}$; b) unavailable sub-set space $S_{UA}^{1}=\{S_7,S_8,S_{10},S_{12}\}$. In order to solve this task we will employ analytical and stochastic method on base of using embedded Markov Chains [8], [15]. Then steady-state probability vector $\pi=\{\pi_0,\pi_2,\pi_3,\pi_5\}$ is solution of this task.

Turning to solution, we will interpret Semi-Markov process as follows. As our research shows CTS performs deterministic period of time T, therefore transitions from states S_i to states S_j are given by:

$$Q_{01}(t)=Q_{24}(t)=Q_{34}(t)=Q_{56}(t)=Q_{89}(t)=Q_{1011}(t)=Q_{1213}(t)=\begin{cases}0,t<T,\\1,t\geq T.\end{cases}$$

At the same time transitions for TSCS from states S_j to states S_i can be written as:

$$Q_{10}(t)=Q_{42}(t)=Q_{43}(t)=Q_{65}(t)=Q_{98}(t)=Q_{1110}(t)=Q_{1312}(t)=\begin{cases}0,t<\tau_c,\\1,t\geq \tau_c.\end{cases}$$

Let we turn now to sudden failures for hot, warm and cold PMs. In this case distribution function for transitions from state S_1 to state S_2, from state S_1 to state S_3, from state S_4 to state S_5 and from state S_6 to state S_7 are given by:

$$Q_{12}(t)=Q_{13}(t)=Q_{45}(t)=Q_{67}(t)=1-e^{-\lambda_s t}.$$

Now we will move on to hidden failures for hot and warm PMs. Distribution functions for transitions from state S_0 to state S_8, from state S_2 to state S_{10}, from state S_3 to state S_{12} can be written as:

$$Q_{08}(t)=Q_{210}(t)=Q_{312}(t)=\begin{cases}1-e^{-\lambda_h t},t<T,\\0,t\geq T.\end{cases}$$

Next distribution functions for hidden unavailable hot or warm PM after next CTS are given by:

$$Q_{92}(t)=Q_{93}(t)=Q_{115}(t)=Q_{135}(t)=1-e^{-\lambda_h^* t}.$$

Distribution functions of transitions from state S_2 to state S_0, from state S_3 to state S_0 and from state S_7 to state S_5 can be written as:

$$Q_{20}(t)=Q_{30}(t)=Q_{75}(t)=1-(1+\mu_h t)e^{-\mu_h t},$$

Simultaneously, distribution functions of transitions from state S_5 to state S_2, from state S_5 to state S_3 are given by:

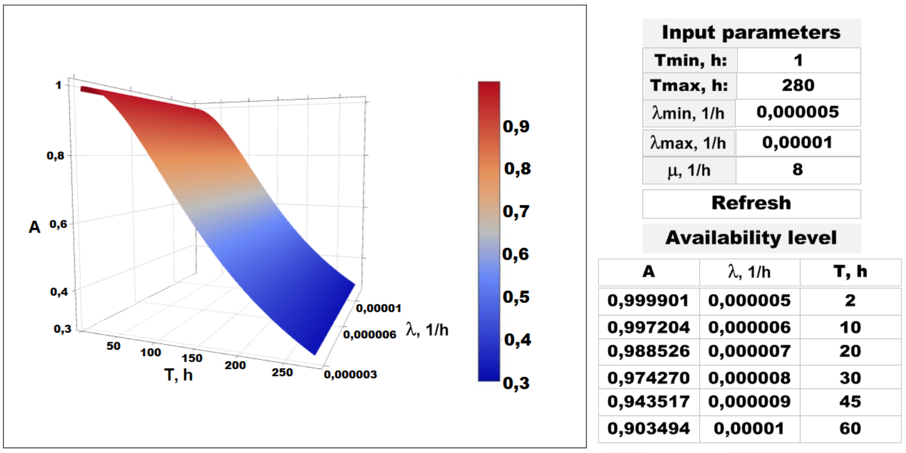
$$Q_{52}(t)=Q_{53}(t)=1-(1+\mu_w t)e^{-\mu_w t}.$$

Taking the total probability relation $\sum_{i=0}^{13}\pi_i=1$ and taking the steady-state probability vector, we can compute the required result as

$$A=\pi_0+\pi_2+\pi_3+\pi_5,$$

where $\pi_0, \pi_2, \pi_3, \pi_5$ are steady-state probabilities for states S_0, S_2, S_3, S_5.

In other words, states S_0, S_2, S_3, S_5 really are states, when IaaS Cloud can perform operational required functions. But we also consider others states, when IaaS Cloud can't perform a certain amount of useful work. Overall results of IaaS Cloud modeling based on embedded Markov Chains are shown in Fig. 3 and Fig. 4.



Input parameters

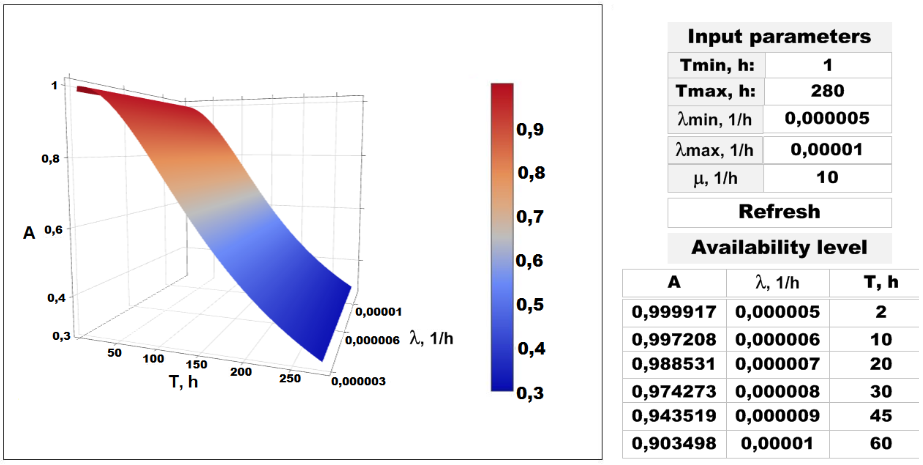
Tmin, h:	**1**
Tmax, h:	**280**
λmin, 1/h	**0,000005**
λmax, 1/h	**0,00001**
μ, 1/h	**8**

Refresh

Availability level

A	λ, 1/h	T, h
0,999901	0,000005	2
0,997204	0,000006	10
0,988526	0,000007	20
0,974270	0,000008	30
0,943517	0,000009	45
0,903494	0,00001	60

Fig. 3. Depending of steady-state availability $A(\lambda_h, T)$ for $T = 250$ h, $\mu = 8$ 1/h



Input parameters

Tmin, h:	**1**
Tmax, h:	**280**
λmin, 1/h	**0,000005**
λmax, 1/h	**0,00001**
μ, 1/h	**10**

Refresh

Availability level

A	λ, 1/h	T, h
0,999917	0,000005	2
0,997208	0,000006	10
0,988531	0,000007	20
0,974273	0,000008	30
0,943519	0,000009	45
0,903498	0,00001	60

Fig. 4. Depending of steady-state availability $A(\lambda_h, T)$ for $T = 250$ h, $\mu = 10$ 1/h

Analysis of modeling results confirmed that value of steady-state availability A is increased by means of increasing of repair rate μ of hot PMs and reduction of hidden failure rate λ_h of hot PMs, as it results by comparing Fig. 3 with Fig. 4.

Next we will illustrate how to use our stochastic approach in order to describe the behavior of IaaS Cloud with three pools of PMs. Figure 5 shows finite graph for second type of Semi-Markov model of the IaaS Cloud with nine PMs. Note that the model was solved for only one type of hidden failures. It is serious lack, because in real situation we can observe different types of hidden failures. For example, hidden failures of hot and warm physical machines can be different.

In considering the second model, we consider that two different types of hidden failures of PMs are made possible. Previous study have shown that we can interpret two branches of hidden failures for PMs using the following probability transitions: 1) p_{03}, p_{711}, p_{1325}, p_{819}, p_{1429}, p_{2750}, p_{2239}, p_{3564}, p_{5371}, p_{4261}, p_{5775}, p_{7385}, p_{2447}, p_{4668}, p_{6782}, p_{8189}, p_{8893}, p_{9297} (first branch for PMs of hot pool); 2) p_{04}, p_{816}, p_{2243}, p_{717}, p_{1437}, p_{3558}, p_{1332}, p_{2754}, p_{5378} (second branch for PMs of warm pool).

According to the Fig. 5, available sub-set space S_A^2 for second model of IaaS Cloud was created by states $S_A^2 = \{S_0, S_7, S_8, S_{13}, S_{14}, S_{22}, S_{24}, S_{27}, S_{35}, S_{42}, S_{46}, S_{53}, S_{57}, S_{67}, S_{73}, S_{81}, S_{88}, S_{92}\}$. Other states for second Semi-Markov availability model can be described as unavailable. Then steady-state probability vector $\pi = \{\pi_0, \pi_7, \pi_8, \pi_{13}, \pi_{14}, \pi_{22}, \pi_{24}, \pi_{27}, \pi_{35}, \pi_{42}, \pi_{46}, \pi_{53}, \pi_{57}, \pi_{67}, \pi_{73}, \pi_{81}, \pi_{88}, \pi_{92}\}$ is solution of this task.

Next we will also move on to hidden failures for hot and warm PMs. Distribution functions for first branch of transitions can be written as:

$$Q_{ij}(t) = \begin{cases} 1 - e^{-\gamma_h t}, t < T, \\ 0, t \geq T. \end{cases}$$

At the same time distribution functions for second branch of transitions is given by:

$$Q_{ij}(t) = \begin{cases} 1 - e^{-\gamma_w t}, t < T, \\ 0, t \geq T, \end{cases}$$

where hidden failure rates of warm PMs γ_w is lower than hidden failure rates of hot PMs γ_h by a factor of two to four [7].

Distribution functions of transitions for repair time of the hot and warm PMs are given by:

$$Q_{ji}(t) = 1 - (1 + \delta_h t)e^{-\delta_h t},$$

$$Q_{ji}(t) == 1 - (1 + \delta_w t)e^{-\delta_w t},$$

where repair rate of hot PMs δ_h is higher than repair rate of warm PMs δ_w.

Overall equation for steady-state availability of IaaS Cloud can be written as:

$$A = \pi_0 + \pi_7 + \pi_8 + \pi_{13} + \pi_{14} + \pi_{22} + \pi_{24} + \pi_{27} + \pi_{35} + \pi_{42} + \pi_{46} + \pi_{53} + \pi_{57} + \pi_{67} + \pi_{73} + \\ + \pi_{81} + \pi_{88} + \pi_{92}.$$

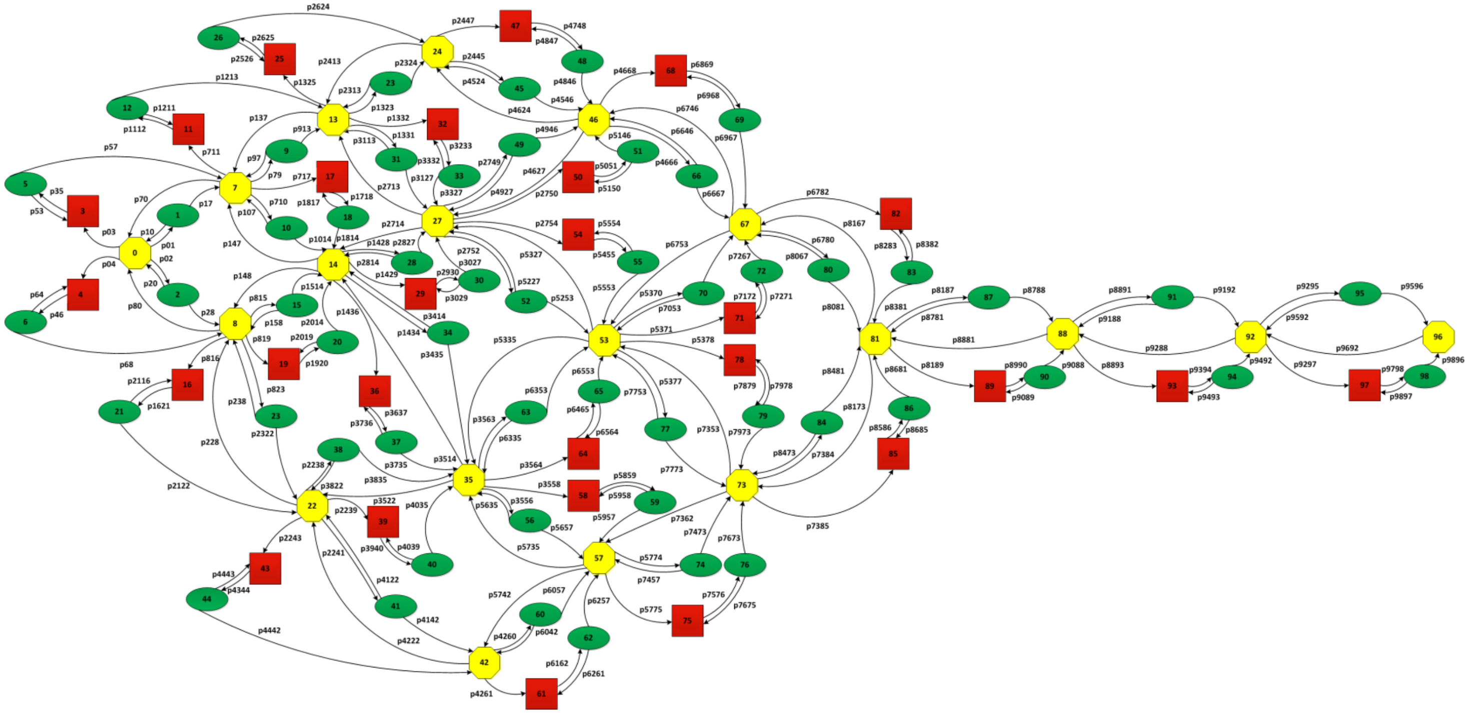


Fig. 5. Semi-Markov availability model of the IaaS Cloud with ten PMs

Finally, in spite of the fact that second Semi-Markov model (Fig. 5) is more complex than first model (Fig. 2), nevertheless we can use similar approach in order to get solution based on embidded Markov Chains. Nowadays we can allege that it is one of the most notable advantages of Markovian modeling all among the famous mathematical methods and stochastic approaches, on which scientific researches for Cloud Computing area is based.

3 Conclusions Statement of the Researches Results

In this paper we performed Semi-Markov modeling for the IaaS Cloud with TSCS based on embedded Markov Chains. The contributions of this paper are the following.

The purpose of the model, through the tool implementing it, is studied and used IaaS Cloud availability measures. It really is the significant assessments for provider. Indeed sudden and hidden failures of PMs are serious problem for IaaS Cloud provider. We illustrated these results in an availability Semi-Markov model with fourteen states.

Our model can be used in order to make profound analysis of different architectures for IaaS Cloud, in particular during accidents, disasters and other negative events, such as DoS and DDoS attack. Several optimization problems for IaaS Cloud regarding resource availability can be formulated using our analytical and stochastic model described in this paper.

In conclusion, increasing scalability and flexibility of this type of models is future of IaaS Clouds development.

References

1. Gamage, T., et al. Mission-Critical Cloud Computing for Critical Infrastructures. Smart Grids: Clouds, Communications, Open Source, and Automation. CRC Press, pp. 1–16 (2014).
2. Birke, R., Chen, L. Y., Smirni, E. Data centers in the cloud: A large scale performance study. Cloud Computing (CLOUD), IEEE 5th International Conference on, pp. 336–343 (2012).
3. Patel, C., Shah, A. Cost model for planning, development and operation of a data center. HP Laboratories Palo Alto, Tech. Rep. (2005).
4. Ghosh, R., Longo, F., Frattini, F., Russo, S., Trivedi, K.: Scalable analytics for IaaS cloud availability. In IEEE Transactions on Cloud Computing, 2(1), pp. 57–70 (2014).
5. Silva, B.: A framework for availability, performance and survivability evaluation of disaster tolerant cloud computing systems. Diss. Federal University of Pernambuco (2016).
6. Matos, R., Araujo, J., Oliveira, D., Maciel, P., Trivedi, K.S.: Sensitivity analysis of a hierarchical model of mobile cloud computing. Simulation Modelling Practice and Theory, no. 50, pp. 151–164 (2015).

7. Ghosh, R, Longo, F., Xia, R., Naik, V. and Trivedi, K.S.: Stochastic model driven capacity planning for an infrastructure-as-a-service cloud. In IEEE Transactions on Services Computing, vol. 7, no. 4, pp. 667–680 (2014).
8. Ivanchenko, O, Kharchenko, V. Semi-markov availability models for an Infrastructure as a Service Cloud with multiple pools. In Proc. International Conference on ICT in Education, Research, and Industrial Applications, pp. 349–360 (2016).
9. Longo, F, Ghosh, R, Naik,V.K, Trivedi, K.S. A scalable availability model for Infrastructure-as-a-Service Cloud. In Proc. The 41st IEEE/IFIP International Conference on Dependable Systems and Networks, pp. 335–346 (2011).
10. Ghosh, R.: Scalable stochastic models for cloud services. Diss. Duke of University (2012).
11. Li, Zheng, Liang, Mingfei, O'Brien, Liam, Zhang, He. The cloud's cloudy moment: A systematic survey of public cloud service outage. arXiv preprint arXiv:1312.6485 (2013).
12. Trivedi, K.S., and Sahner, R.: SHARPE at the age of twenty two. ACM Sigmetrics Performance Evaluation Review, vol. 36, no. 4, pp. 52–57 (2009).
13. Cai, B.L., Zhang, R.Q., Zhou, X.B., Zhao, L.P., Li, K.Q.: Experience Availability: Tail-Latency Oriented Availability in Software-Defined Cloud Computing. In Journal of Computer Science and Technology, vol. 32, no. 2, pp. 250–257 (2017).
14. Lanus, M., Yin, L., and Trivedi, K.S.: Hierarchical Composition and Aggregation of State-Based Availability and Performability Models. In IEEE Trans. Reliability, vol. 52, no. 1, pp. 44–52 (2003).
15. Ivanchenko, O., Lovyagin, V., Maschenko, E., Skatkov, A., Shevchenko, V.: Distributed critical systems and infrastructures. National Aerospace University named after N. Zhukovsky "KhAI", Kharkiv (2013).