# Analysis of Methods for Providing Availability and Accessibility of Cloud Services

Max Yanovsky[1], Olga Yanovskaya[1], Vyacheslav Kharchenko[1,2]

[1]National Aerospace University named after N.E. Zhukovsky "KhAI"
17, Chkalova St., 61070 Kharkiv, Ukraine
[2]Centre for Safety Infrastructure-Oriented Research and Analysis
17, Chkalova St., 61070 Kharkiv, Ukraine

(M.Yanovsky, O.Yanovskaya}@csn.khai.edu, V.Kharchenko@khai.edu

**Abstract.** The article describes methods for dealing with reliability and fault tolerance issues of cloud datacenters. These methods are mainly focused on the elimination of single point of failure within any component of the cloud infrastructure, including the availability of infrastructure and accessibility of cloud services. The methods for providing the availability of hardware, software and network components are also presented. The analysis of the actual accessibility of the cloud services and the matching of cloud datacenter infrastructure with the level of reliability according to the Tier Classification System is described. Non-compliance of the actual accessibility with the level of High Availability for cloud web services was found.

**Keywords.** Availability, Accessibility, Cloud Datacenter, Service Reliability

**Key Terms**. ICTInfrastructure, ICTComponent, WebService, FormalMethod

## 1 Introduction

High availability is a critical issue for the cloud datacenter. Thus, estimating the financial loss due to the failure of datacenter components, which would result in unavailability of services, is a major part of an economic development plan for any cloud datacenter customer. The complexity of the architecture, meaning the large number of components and diversity approaches in designing the structure of the network infrastructure, causes issues in obtaining accurate evaluation of reliability, availability and accessibility of such systems.

In order to ensure end-user quality of service, the required cloud system should have the appropriate characteristics of reliability and performance. The term reliability refers to the property of an object or system to maintain, over time and within the prescribed limits, the ability to perform the required functions in set modes and conditions of use, maintenance, repair, storage and transportation according to ISO 2382-14:1978 [1]. A major property is failure-free operation, a property of an object that refers to permanent operability during some period of time. Time to failure – time to

ICTERI 2016, Kyiv, Ukraine, June 21-24, 2016

the first failure: This property is characterized by the probability of failure-free operation – likelihood of absence of failure within a given operating time.

The main internal property of reliability is availability. Availability reflects the system's ability to perform its functions continuously. The availability coefficient is defined as the probability that at any given time *t* the object is in working state s, except for maintenance periods during which there is no intended use of the system. However, according to the concept of cloud-based architecture, the concept of availability as the main internal property of reliability refers to the entire infrastructure: the value of the availability factor will be determined based on how efficient the functional state of the system is considered and under what conditions the state of the system can be considered workable. The article considers methods for providing the availability of cloud infrastructure and accessibility of cloud services, as well as analyzes the effectiveness of their application based on studies for actual availability of cloud providers' services.

## 2 State of the Art

Studies' analysis [2, 3] leads to the conclusion that currently there is ambiguous interpretation of cloud datacenter operability conditions as it depends on the number of available and unavailable services in relation to the total amount of services, at the current time. Seeing as the end user interacts with a specific service of the cloud datacenter, the term of cloud service accessibility is suggested to be used.

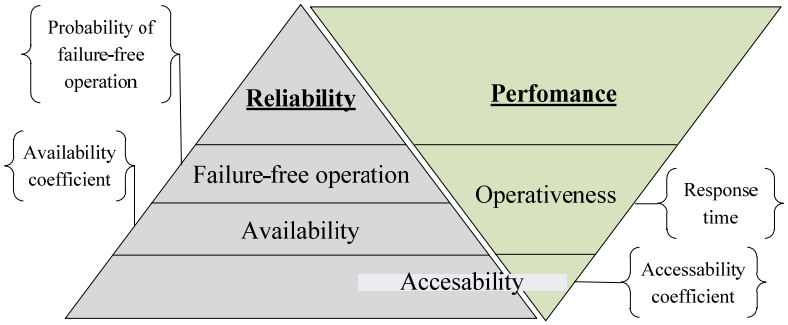


**Fig. 1.** Correlation between availability, reliability and accessibility indicators

Analysis of the sources [4, 5] shows that the most common availability indicator is determined by the following formula:

$$Ka = MTTF/(MTTF + MTTR), \quad (1)$$

where *Ka* – availability coefficient,
*MTTF* – mean time to failure,
*MTTR* – mean time to recover.

Property of accessibility determines the probability that at any time a certain cloud service will be available to the end user with a satisfactory response time. The main

factor is the accessibility coefficient, which includes not only the availability, but also the functional properties of the system.

With increasing demands on the quality of services in the IT-infrastructure, any kind of failures in the network are unacceptable. Even a relatively small packet loss can have a negative impact on the end-users' quality of service, especially for critical and business-critical processes, so the failure of the main switching node, link or interface may have serious consequences for the provider. The design of the cloud datacenter should help minimize network failures and the severity of the consequences of potential accidents.

Advances in technology and the pace of construction of virtual data centers and cloud infrastructures have caused the development of requirements for the distribution functions of management control across multiple geographically dispersed nodes, the division of responsibility between the teams of technical personnel, the extension of monitoring and diagnostics functions support high availability and disaster recovery. According to [6], the datacenter design should include redundant components and distributed platforms, so that the physical connection and access to resources remain constant, regardless of the location and value of the current availability and performance indicators. Furthermore, to protect the competitiveness of enterprises and organizations that are customers of cloud providers, critical business applications need to be available 24/7. In case of environmental or technological disasters, the data must be restored with minimal disruption, calling for an emergency backup and recovery of business applications and the virtual machine in a different availability zone will ensure that user data is protected and accessible from anywhere.

Typically, network architects predict a 4 or 5 "nines" system availability [6]. However, each additional digit = "9" can significantly increase the cost of deployment. To achieve near-zero downtime per year of the cloud data center, one must consider not only the reliability of the hardware and network infrastructure, but also part of software.

## 3 Classification of Cloud Data Centers Based on Tier Standard

Cloud datacenter reliability levels have been identified in the documents "Data Center Site Infrastructure Tier Standard: Topology" [7] and "Data Center Site Infrastructure Tier Standard: Operational Sustainability" [8] of the world organization Uptime Institute [9], which are engaged in the development and verification of detailed requirements for a fault-tolerant datacenter infrastructure, certification and issuance of recommendations and expert advice on data center infrastructure according to the level of reliability.

Uptime Institute Certification on levels of reliability meets the standard ANSI/TIA-942-A [6]. Classification of data center infrastructure by levels of reliability is carried out on the basis of these basic criteria, the degree of redundancy of equipment and communication channels, the meeting of performance characteristics, functionality, efficiency and expected availability level. The requirements and recommendations apply to the following systems and components:

– architecture and topology;

– power supply system;
– cooling system;
– security;
– fire alarm system;
– structured cabling system;
– maintenance.

There are 4 standard levels of reliability:

1. TIER I: Basic Site Infrastructure;
2. TIER II: Redundant Site Infrastructure Capacity Components;
3. TIER III: Concurrently Maintainable Site Infrastructure;
4. TIER IV: Fault Tolerant Site Infrastructure.

Datacenter infrastructure and operating expenses increase in accordance with the reliability level, which gives grounds for datacenter owners to choose the class of reliability at the designing stage and in accordance with their business needs. The results of the comparative analysis of the datacenter infrastructure reliability levels are shown in the table below.

**Table 1.** Results of the comparative analysis of the datacenter's infrastructure reliability levels

| **Properties** | **TIER I** | **TIER II** | **TIER III** | **TIER IV** |
|---|---|---|---|---|
| Level of active equipment redundancy | N | N+1 | N+1 | 2(N+1) |
| Redundant channels | 1 | 1 | 1 active + 1 passive | 2 active |
| Possibility of maintenance without downtime | No | No | Yes | Yes |
| Fault-free operation | No | No | No | Yes |
| Continuous cooling system | No | No | No | Yes |
| Availability coefficient, % | 99,671 | 99,749 | 99,982 | 99,995 |
| Downtime per year, h | 28,8 | 22 | 1,6 | 0,4 |

Based on the results of this analysis, the following conclusions are made:

1. The existing Tier Classification System for the reliability assessment of the datacenter infrastructure in terms of business requirements for system performance does not consider the reliability of software components.

2. The classification system of datacenter reliability on Tier levels does not explicitly consider the characteristics of the equipment, such as mean time to failure, which is not correct in assessing the availability of the system.

3. Deploying cloud and business-critical applications requires the highest level of availability TIER IV datacenter infrastructure.

4. During operation of the cloud datacenter and appending servers and equipment within a constant engineering infrastructure needs may change in the required resources, which may lead to a change in the datacenter reliability. Thus, it is necessary

to review and confirm the level of reliability of the datacenter, as well as to ensure the effective operation by highly trained personnel and administration.

In order to meet the levels of reliability and maintenance of the set level of availability and accessibility of cloud infrastructure services, a variety of methods are used to ensure fault tolerance in the core, aggregation and access layers. The main purpose of the application of methods is to ensure availability, which means to eliminate points of single failure of any component of the cloud infrastructure (hardware, software, network) at any layer (core, aggregation, access). Since hardware failure and software faults may appear in components at any layer, there are methods of fault tolerance for each of them. The methods are used to ensure the availability of cloud services at different levels of the architecture will be considered.

## 4 Methods for Providing Availability of Hardware Components

The objective of this group of methods is to maintain the availability of cloud services and applications in case a particular server becomes unavailable. The method can operate at multiple levels within the datacenter infrastructure. Hardware component accessibility methods are used at the physical layer ISO/OSI model. These include the following.

- Grouping of network adapters and communication channels.

In order to eliminate single points of failure at the level of communication with network, access layer servers have multiple (two or more) network interfaces (Fig. 1). This method is named NIC-Teaming and it involves the grouping of multiple physical connections into one logical channel - LAG (Link Aggregation). The logical connection may be in active-active mode that combines multiple channels into a single logical load sharing or active-passive mode, wherein the second interface is idle as long as the first interface operates as usual.

- Using hot-swappable interfaces.

This method requires the ability to install or remove the interface card on the router or switch without having to power off the device. The controller dynamically recognizes the new interface and begins the data exchange. As a result, new components can be inserted and removed without interrupting the system’s operation.

- Use of highly reliable server access layer.

Hardware components of highly reliable servers have the highest values of MTTF.

## 5 Methods for Providing Availability of Software Components

An analysis [10, 11] found that the main methods of resilience at the application level are the use of pooling resources, protected applications and resources of critical applications as well as and the migration of virtual machines and the use of Unified In-Service Software Upgrades.

- Using application pools of resources.

This method is based on the fact that multiple instances of applications are combined to the pool of resources that are distributed throughout the network (Fig. 2).

According to [11], the use of resource pools is an effective solution for resiliency, but the main disadvantage of this approach is the problem of synchronization. This solution requires more effective methods of planning, synchronization and load balancing on the coordination sites.

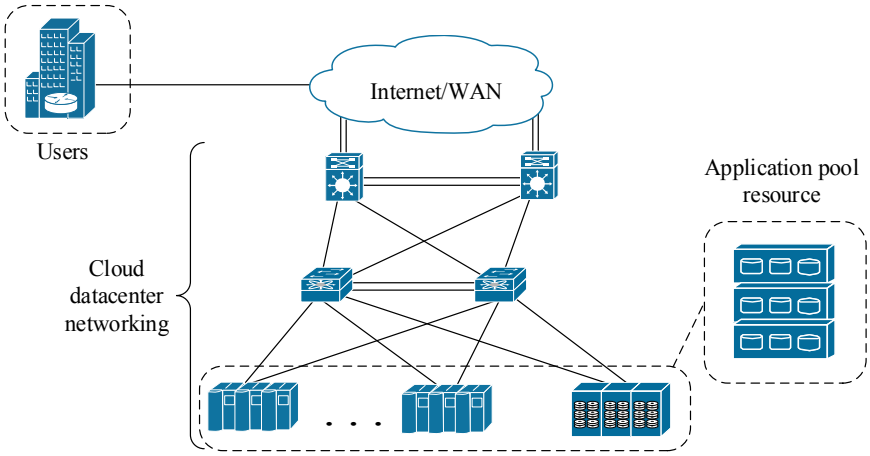


**Fig. 2.** Use of application pools of resources

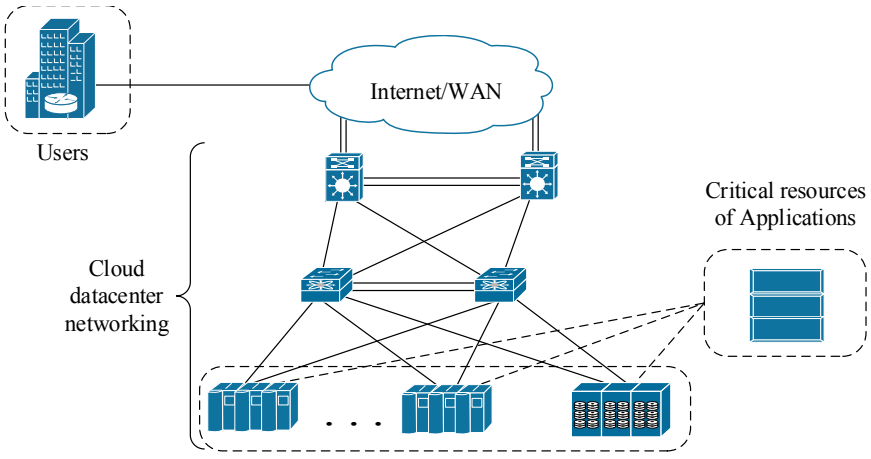


**Fig. 3.** Transfer of critical resources of applications

- Transfer of critical application resources.

Some applications have critical resources, that for various reasons are not possible or desirable to replicate. They can either work on the basis of high-performance servers, that makes replication too expensive, or they can include critical resources that makes replication not possible due to security threats or exploitation reasons.

Under these conditions, one application server is a single point of failure. In order to minimize the risk of failure for critical resources of applications, the execution of these applications are made on several powerful servers, failover and high availability provided by the active / standby configuration mode for disaster recovery. Connection problems are solved by multisession network connections between the server and clients, and multiple network routes (Fig. 3).

Conditions of effective application of this method are the presence of redundant network links and backup systems as well as continuous monitoring of the status of servers and data replication, in order to maintain synchronization of the active and standby systems.

- Migration of virtual machines.

Virtual machine migration is an effective method for providing fault-tolerance and for maintaining service availability in the event of a failure of the physical server on which it is running. This method assumes that the virtual machine has its own running copy on a server located in another rack or in another datacenter. In this case, services that are deployed on the initial virtual machine are replicated on another virtual machine.

- Using a single integrated service system updates.

The ability to provide unified system ISSU (Unified In-Service Software Upgrades) updates of an operating system without shutting down network devices that are scheduled for preliminary verification of compatibility, is supported by some versions of operating systems within a number of network equipment manufacturers [11], thus avoiding the risks associated with downtime and failed updates of network operating systems.

## 6 Methods for Providing Availability of Network Components

- Redundant network devices

The analysis of the examined standards and guidelines for the design of the datacenter allows us to determine that the redundancy of network devices as a method of fault tolerance, involves the duplication of the core level routers, access layer and distribution switches.

Additional mechanisms for balancing the load between them increase network performance and reduce latency.

Apart from that, in order to minimize the effects of a single point of failure, the network device may also be used in methods such as hot-swappable interface, Unified In-Service Software Upgrades, redundant switching and routing mechanisms.

- Redundant switching and routing mechanisms

The main purpose of this method is to create redundant switching for network devices.

Along with a redundant configuration, switching fabric with two switch modules is used to increase the capacity and performance the switch. The third module, if present, provides an additional precision (2 + 1) for switching functions, so that if one of the two functional modules becomes inoperable, a third module can take over the function of the failed module. Redundant routing mechanisms provide simultaneous operation of multiple routing protocols as well as protection from the routing and switching loops based on the following protocols, technology and standards:

- L3 dynamic routing protocols in the core level (OSPF, RIP, or static routing);
- Multiple Spanning Tree Protocol (MSTP);
- MPLS in the core level;
- 802.3ad LAG;
- 802.1q Virtual LANs;
- RTG (Redundant trunk groups);
- VRRP;
- MPLS in the aggregation level.

Based on the analysis, the presented methods highlight a number of common disadvantages in their application: complexity of the architecture, the processes of its maintenance and operation due to demand for high priced resources, additional overhead costs on excess equipment and permanent high-quality maintenance. In order to determine the effectiveness of the methods and architecture considered, an analysis of the accessibility of services from known cloud providers should be conducted.

## 7 Analysis of Actual Services Accessibility of Cloud Providers

Figures 4 - 5 are bar graphs that illustrate the results of statistical data processing regarding the services accessibility of cloud providers.

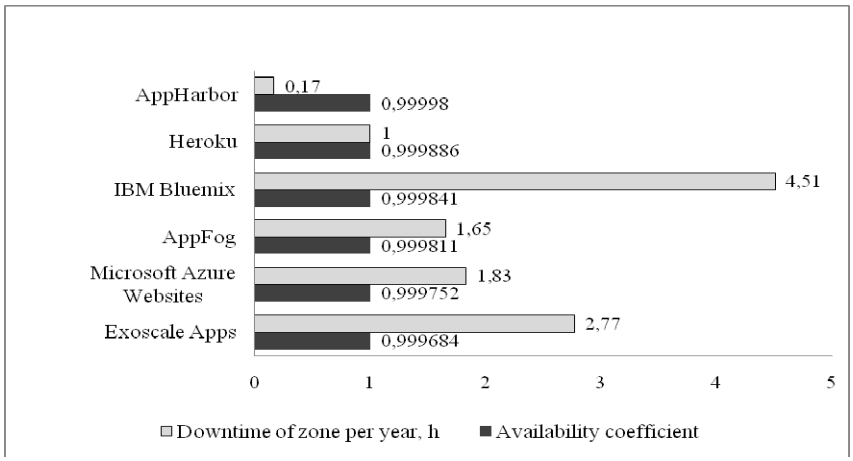


**Fig. 4.** Actual services accessibility of cloud providers: PaaS service model

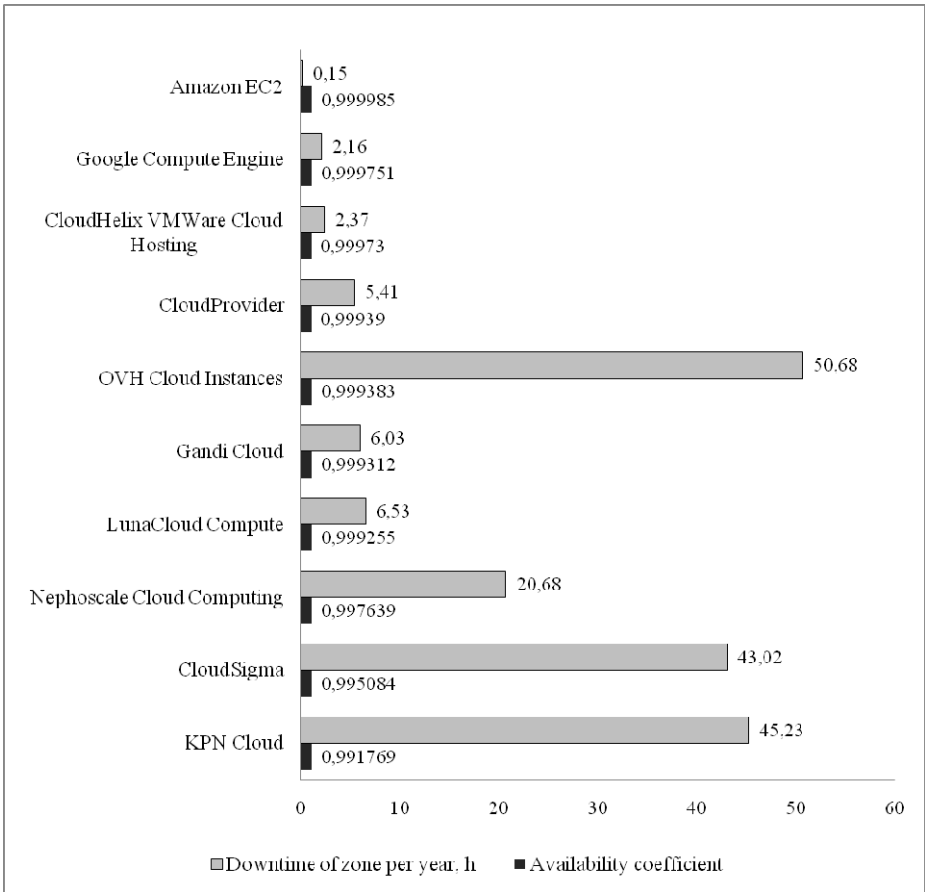


**Fig. 5.** Actual services accessibility of cloud providers: IaaS service model

The values shown in the above histograms are obtained by analyzing the actual values of downtime for a time period of 1 year, as published by Cloud Harmony [12] for the service models of PaaS and IaaS [13].

According to the analysis of the actual accessibility of cloud providers' services, and as shown by the values represented in the histograms, we can conclude that the average accessibility of a datacenter's cloud services corresponds to a value of 0.999.

In order to determine whether the claimed quality of the service is in compliance with the actual quality of service, an analysis of the service-level agreement (SLA) of known cloud providers was performed. The results of the analysis are summarized in the table 2.

In order to verify compliance of the datacenter infrastructure cloud providers with levels of reliability according to the Tier Classification System, made analysis of the data with characteristics of the datacenter provided by cloud providers. The results of the analysis are presented in the table 3.

**Table 2.** Results of the comparative analysis of the datacenter infrastructure reliability levels

| Cloud provider / Service type | Claimed levels of service accessibility (replication services provided at least two availability zones) |
|---|---|
| Microsoft Azure / Microsoft Online Services [14] | 99.95 % |
| Microsoft Azure / Virtual Machines [15] | 99.95 % |
| Microsoft Azure / Cloud Services [16] | 99.95 % |
| Google Cloud Platform / Google Compute Engine [17] | 99.95 % |
| Google Cloud Platform / Google App Engine [18] | 99.95 % |
| Amazon EC2 [19] | 99.95 % |
| Rackspace Cloud Servers [20] | 99.9 % |

**Table 3.** Reliability levels of cloud providers' datacenters

| Cloud provider | Tier Reliability Level |
|---|---|
| Amazon [21] | IV |
| Microsoft Azure [22] | IV |
| Rackspace Cloud [23] | IV |

Analysis of the results leads to the conclusion: despite the fact that the cloud datacenter infrastructure matches the fourth level of reliability with an availability coefficient of 0,99995, the actual average access-infrastructure of cloud service providers, on average, corresponds to a value of 0.999. Therefore, it is necessary to improve the models and methods of assessing the availability and accessibility for services of client-server cloud infrastructure to obtain more accurate estimates of the reliability indices.

## 8 Case study

This section provides the results of a case study on accessibility, based on the simulation model of the cloud server that is running 3 virtual machines.

The results of the statistical analysis of time characteristics of the WEB-applications servers' performance [24] confirm that, based on a mathematical model of computing systems, it can be assumed that the random variables: time of the requests towards the server has an exponential distribution, and the input query is a Poisson series ( QS M/M/1). Based on this assumption, a model was found. Requests arrive to the physical server network card (Physical NIC), then are distributed among the virtual network interfaces and are processed there. In this model, another element to the input stream applications was added: that of lost requests, due to loss of server performance (hardware failure, operating system or hypervisor refusal). It can be

assumed that hardware failures occur on average once every 300,000 hours, with mean time between OS failures is 1440 hours, and the hypervisor is 2880 hours.

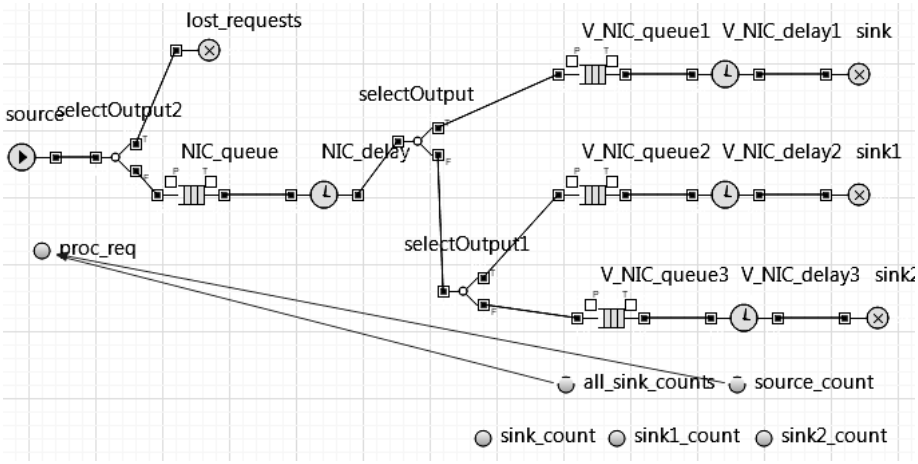


**Fig. 6.** Simulation model of the cloud server presented in AnyLogic environment

The resulting value, obtained as a percentage of processed requests on the time line, is as follows:

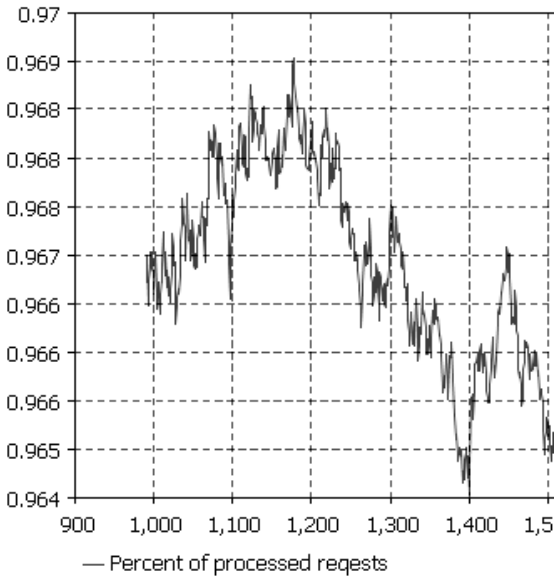


**Fig. 7.** Simulation results

According to the results shown in the graph, it is evident that during the day (24-hour model) each unit of the system's configuration, that has its test parameters on the average percentage of time, processes user requests of about 96%.

Almost all of the input parameters depend on the specific hardware and software implementation of the system. The exception is the MTTR of hardware failure, the operating system and the hypervisor as well as the corresponding recovery rate. The experiments s conducted showed the recovery time takes an average of 1 to 2 hours. The simulation results, when changing the data in this range, show a decrease in the recovery times of up to 1 hour while it is possible to get an increase in availability features in 4 digits.

It is possible to improve this figure in several ways:

1) increase the average time between failures of hardware and software server;

2) decrease the time of exchange between physical and virtual network adapters, which can be varied from a few milliseconds to tens of milliseconds, depending on the specific application platform and hypervisor which implements virtual server infrastructure.

## 9 Conclusion

The analysis methods for fault tolerance and availability of client-server cloud infrastructure services were presented. These methods are more focused on the points of single failure elimination for each component of the cloud infrastructure (hardware, software, network) in every level (core, distribution, access), as well as models and methods of estimating the availability and accessibility of cloud-based architectures. The results of the analysis of the actual datacenter services accessibility of cloud service provider for service models PaaS and IaaS, as well as compliance of datacenter infrastructure of cloud providers with levels of reliability according to the Tier classification were presented. Despite the use of different methods of fault tolerance in cloud infrastructures, there is the problem of inconsistency of the actual system availability level of "High Availability" for critical and business-critical web applications. Thus, the direction of future research would be towards the improvement of the models and methods so as to ensure accessibility of cloud services. Furthermore, according to analysis results, it is important for the direction of future research results to find an effective combination of the discussed methods for providing the required level of availability and accessibility.

## References

1. ISO/IEC 2382-14: Information technology. Vocabulary. Part 14: Reliability, maintenance and availability (1997)
2. R. S. Couto, M. E. M. Campista and L. H. M. K. Costa: A reliability analysis of datacenter topologies, Global Communications Conference (GLOBECOM), IEEE, Anaheim, CA, pp. 1890--1895. doi: 10.1109/GLOCOM.2012.6503391 (2012)
3. K. V. Vishwanath, N. Nagappan: Characterizing cloud computing hardware reliability, in Proceedings of the 1st ACM symposium on Cloud computing (SoCC '10). ACM, New York, NY, USA, pp. 193--204. doi: 10.1145/1807128.1807161 (2010)

4. Fernandes, S., Tavares, E., Santos, M., Lira, V., & Maciel, P.: Dependability assessment of virtualized networks. IEEE International Conference on Communications (ICC), pp. 2711--2716, Ottawa (2012)
5. Downtime Statistics of Current Cloud Solution,
6. http://iwgcr.org/wp-content/uploads/2014/03/downtime-statistics-current-1.3.pdf
7. TIA/EIA-942: Telecommunications Infrastructure Standard for Data Centers (2005)
8. Data Center Site Infrastructure Tier Standart: Topology,
9. https://uptimeinstitute.com/publications/asset/tier-standard-topology
10. Data Center Site Infrastructure Tier Standart: Topology,
11. https://uptimeinstitute.com/publications/asset/tier-standard-operational-sustainability
12. About Uptime Institute, https://uptimeinstitute.com/about-ui
13. Cisco Data Center Infrastructure 2.5 Design Guide,
14. http://www.scn.rain.com/~neighorn/PDF/Cisco_Data_Center_Infrastructure_Design_Guide.pdf
15. Juniper Networks, "Cloud Ready Data Center Network Design Guide",
16. http://www.juniper.net/us/en/local/pdf/design-guides/8020014-en.pdf
17. Research and compare cloud providers and services, https://cloudharmony.com/status
18. Cloud Computing Vendors Taxonomy, http://cloudtaxonomy.opencrowd.com/taxonomy
19. SLA for App Service, https://azure.microsoft.com/en-us/support/legal/sla/app-service
20. SLA for Virtual Machines, https://azure.microsoft.com/en-us/support/legal/sla/virtual-machines
21. SLA for Cloud Services, https://azure.microsoft.com/en-us/support/legal/sla/cloud-services
22. Google Compute Engine Service Level Agreement, https://cloud.google.com/compute/sla
23. Google App Engine Service Level Agreement, https://cloud.google.com/appengine/sla
24. Amazon EC2 Service Level Agreement, https://aws.amazon.com/ru/ec2/sla
25. Cloud Service Level Agreement, https://www.rackspace.com/information/legal/cloud/sla
26. Amazon called out over cloud security,
27. http://www.techworld.com.au/article/326287/amazon_called_over_cloud_security_secrecy
28. Microsoft Cloud Services,
29. https://assets.digitalmarketplace.service.gov.uk/documents/93064/4504230568132608-pricing-document.pdf
30. About Rackspace. Global Infrastructure and Uptime Guarantee,
31. http://www.rackspace.com/about/datacenters
32. Gorbenko, A., Romanovsky, A.: Time-Outing Internet Services. Security & Privacy, IEEE, 11(2), pp. 68--71. doi: 10.1109/MSP.2013.43 (2013)