COMPUTATIONAL PROBLEMS OF ELECTRICAL ENGINEERING

Vol. 3, No. 1, 2014

SERVICE PROVISIONING BY USING A STRUCTURE STABILITY ALGORITHM IN A VIRTUALIZED DATA CENTER BASED ON CLOUD TECHNOLOGY

Bogdan Strykhalyuk, Olga Shpur, Andriy Masiuk
Lviv Polytechnic National University, Ukraine
bogdan_str@ukr.net, o.shpur@mail.ru

Abstract. This paper is devoted to the problem of reducing the time for processing of the requests that arrive at a data center to be serviced, taking into account a topological structure of the center. However, to take into account the structure of the data center is not enough, an important factor while servicing the requests is survivability of such a structure, because when the structure is stable, the system performs faster and redirects the required service to a user. In this paper, there is a method proposed for assessing stability of the structure of virtualized data centers taking into account intensity of the requests that arrive at VM. This will allow based on the structure stability at a particular moment the recalculation of an optimal transmission route to be skipped that will reduce service delays.

Key words: service provisioning, virtualized data center, structure stability, cloud technology, service provisioning in cloud networks.

1. Introduction

Distributed computing is a new branch in IT. It embodies a set of access technologies and systems that must provide end users with necessary services. The main component of cloud computing is a cloud system, which using the service component replication and large amounts of computing resources can ensure that any request of service will be processed with the lowest delay. The diversity of such requests and their processing would not be possible without virtualization. In cloud systems, the virtualization is a technology that allows abstracting from hardware to the point where software stacks can be deployed and put into action without binding to a single physical server. The virtualization contributes to creating a dynamic data center where servers are arranged into a pool of resources that can be used in case of need, and where applications for computing, storage and networking resources will change dynamically to fit needs. A user has access to his/her data but cannot control or does not have to take care of infrastructure, operating system, and his/her own software to deal with. However, there are very important aspects of providing cloud services: service rate, availability of free channels and necessary throughput to satisfy the users' needs. The rapid development of info-communication networks sets higher and higher requirements to the service providers [1]. That is why the accelaration of service provisioning is a very urgent question.

2. Statement of the problem

The advantage of cloud systems when performing distributed computing is the possibility of a dynamic deployment of virtual machines on servers depending on the users' requests that arrive at a data center and the possibility of dynamic migration of virtual machines from one server to another. A data center is the whole complex of engineering and IT systems that is an integral part of a great number of telecommunication structures. It must ensure a single information resource with guaranteed levels of authenticity, availability and data security. However, the virtual topological structure of such data centers is very unstable and can change dynamically, especially due to the migration of virtual machines form one server to another or even to another data center. Migration is a process enabling to shut down a virtual machine on one physical server, perform all necessary preparations in case they have not been done in advance to transfer a set of data belonging to this particular virtual machine to another server, and then start up the machine on this server, i.e initialize the virtual machine assigning a new IP address [2]. Redirection of the requests to another logical or physical channel will impact on the total service time. That is why such an "unstable" structure of the data center will cause greater delay in request servicing. Thus, one of the research tasks is to decrease the time assigned to service the requests arriving at a data center considering its logical topology. However, taking into consideration the structure of a data center is not sufficient. The important factor that must be considered is vitality of such a structure because the more stable the structure is, the faster the system performs and redirects the requested service to a user. Many scientists have analyzed this problem. For instance, in [3], the authors have evaluated

survivability and performed an analysis of "typical" structures (star, ring), in which the emergence of new units does not significantly impact on service delivery to an end user. The model proposed in [4] is intended for evaluating the efficiency of a highly virtualized cloud center with the Poisson distribution of the tasks, and the normal distribution of the task size. The model is based on a two-step approximation technique where the basic Markov process is first modeled as a built-in semi Markov process, which is then modeled as a Markov approximated process, but only when super tasks are received. However, this model does not involve changing the position of the virtual machines and the resulting impact on the time of service provided to an end user

That is why in this paper, we propose a method for evaluation of data center's structure stability considering the incoming requests intensity. This will make it possible based on the data concerning structure stability at particular moments of time to avoid recalculating an optimal transmission route that in turn will result in a decrease in service delays.

3. Implementation of route search algorithm with minimal transmission delay criteria

A cloud data center logically consists of five main levels: aggregation, access, applications, storage, optical channels. Form the perspective of logical topology, outer servers of the main distributing subsystem are logically separated from the servers that host HAD subsystem application, which in turn are separated from the servers that host EDA subsystem. The traffic is first transmitted from a client to an outer server, then from the outer server to the application server and, finally, from the application server to the database servers. The logical separation implies that each level is a special functional zone and has its own logical channels. A request of service is transmitted through the physical channels to the aggregation level when the management system searches for and provides necessary service resources. Here, the resources are the availability of free physical servers with necessary software installed. At the access level, based on free physical and logical channels, the resource allocation system finds necessary resources, and executes the request processing in order to access to a respective service. Between the levels of access and application, allocation of physical channels, as well as startup of an algorithm for searching logical channels to get access to virtual machines take place, making use of the routing algorithm. It should be mentioned that one physical machine can host scores of virtual machines that can execute only one type of a software complex.

Access to and search for both an optimal physical and logical route of transmitting requests to the software on the virtual machines for a service to be provided are performed using an algorithm of minimal spanning tree. [5] However, such an algorithm automaticly blocks the redundant at this particular moment connections for full coherency of the ports, and as a result cannot ensure sufficient quality of service. The metric of this algorithm takes into consideration a current workload of the channels only towards the “central” physical machines (those that receive the greatest number of requests) and does not analyze the structure of other connections, i.e. does not have the ability every time to analyze the topology between the virtual machines. Such an analysis becomes necessary especially when parts of the virtualized data center migrate from one server to another.

Let us assume that a user sends a request of service to a cloud data center. The data center initializes the assignment of the physical server that hosts the necessary type of service. When a few requests arrive at the same physical machine, the performance of such services drops, for the service access to the resources of the physical machine is scheduled according to the time-division method. As a result, when the number of services on one physical machine increases, and when the intensity of incoming requests to the given PM increases, and when there is a rise in the intensity of incoming requests of all services on this PM, then the request processing time by each service increases. In case the service performance goes down, the management system transfers this service to another physical machine. In this case, the total time of request processing including the time to transfer a request from a user to a data center and backwards can be calculated as:

$$t_{пер} = \sum_{1}^{n} t_{комут.} + \sum_{1}^{n-1} t_{п.к.з.} + t_{обр.}, \qquad (1)$$

where n is the number of requests; $t_{комут.}$ represents the time of passing a request through a switching system; $t_{п.к.з.}$ denotes the time of searching request transmission channels; $t_{обр.}$ is the request processing time which is the total time of request processing by the service and consists of k- atomic services:

$$t_{обр.} = \sum_{1}^{k} t_{а.с.}, \qquad (2)$$

The optimal transmission route changing, the time of transmission channels search increases - $t_{п.к.з}$, that in turn will cause an increase in the transmission time. In order to decrease this time, we propose to use an algorithm of route search based on the minimal transmission delay criteria [6]. This algorithm calculates an optimal transmission route based on dissemination of information and changes in network topology.

The metric of this algorithm is determined by the formula below:

$$M = \left(K_1 * \min C + \left(\frac{K_2 * \min C}{256 - L} \right) + K_3 * D \right) * \left(\frac{K_5}{R + K_4} \right) * 256, \quad (3)$$

where K_i is the coefficient set by a network administrator for configuration of the composite metric; *min C* stands for the minimal value of the throughput along the route used for data transfer; *L* represents the load of each section of the network; *D* denotes the total delay on interfaces; *R* is the route reliability. In fact, the problem of decreasing the time necessary for servicing the requests that arrive at a data center, considering the topology of this data center, comes to a decrease in the average summary delay *D*. However, such a delay must satisfy the following conditions:

$$D = \frac{1}{P_{cm}} \sum_{Deg.n} \frac{w_T(n)}{C - w_T(n)} \le D_{зад}, \quad (4)$$

where P_{cm} represents the probability of system structure stability, $W_T(n)$ is the load of the vertices during the interval *T*, *C* denotes the throughput of a communication channel, *Deg n* is the number of edges connected to the vertex *n*. For the algorithm of optimal route search based on the minimal transmission delay criteria, the delay equals 100 microseconds when the speed is 100 Mbit/s.

The delay can be reduced only in case the structure of the data center is stable during the period *T*. For the structure stability (vitality) to be evaluated, we propose a method involving the assessment of network structure (network model), the load of each virtual machine at the moment of time *T*, and also the coherency of nodes.

4. Algorithm of structure vitality evaluation

The structure of a cloud computing system can be presented as a graph. However, taking into consideration failures or migration of the virtual machines, such a graph will be non-determined, and mathematically it can be described only by using the theory of random graphs (models of Balobashi-Albert or Erdesh-Renni [7]).

Such a graph G will have the random number of nodes (a node implies VM) and edges (physical channels). Delays directly depend on the system structure (on the graph forming model) and its stability. Graph reliability can be determined as one minus the sum of probability of node coherency that cannot exceed a maximal possible load of the system:

$$P_{cm} = 1 - \sum_{i=1}^{n} w_i(t) + P_{зв_i}, \quad (5)$$

$w_i(t)$ is the load of the nodes at the moment of time t; $P_{зв_i}$ is the probability of nodes coherency.

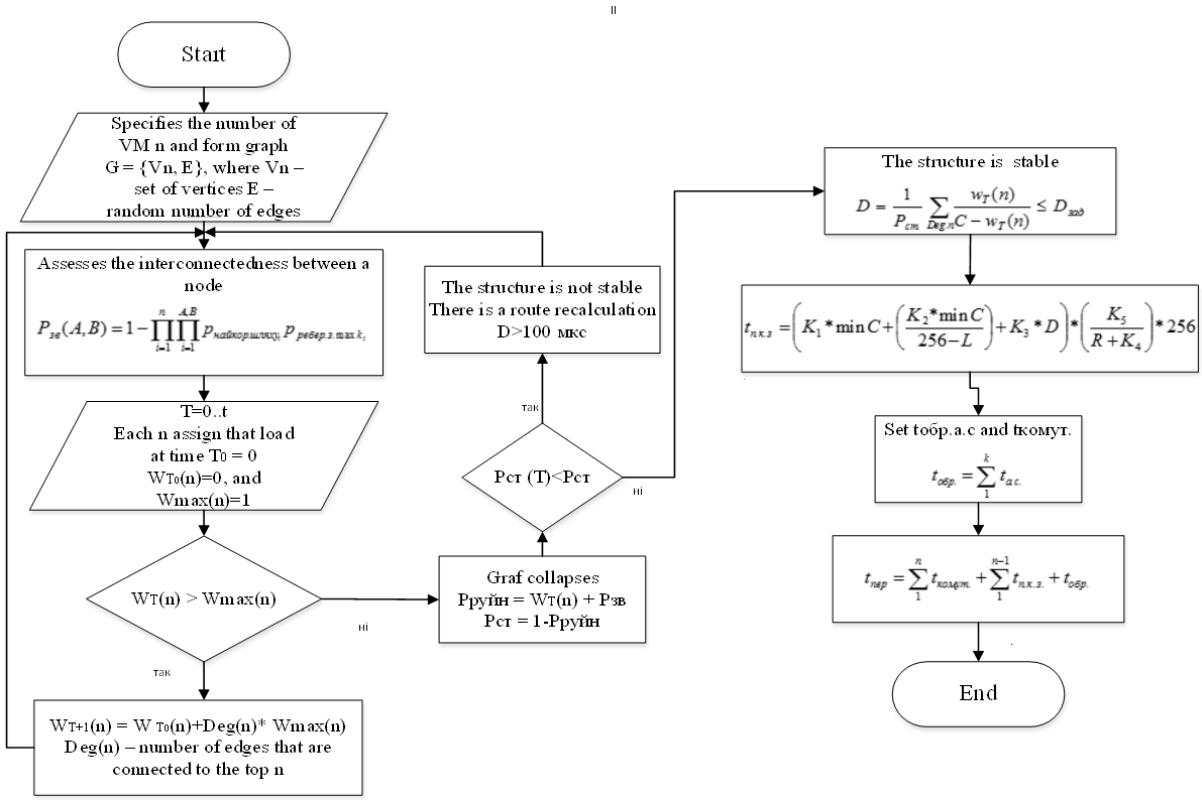

Fig. 1. Algorithm evaluation of structure vitality.

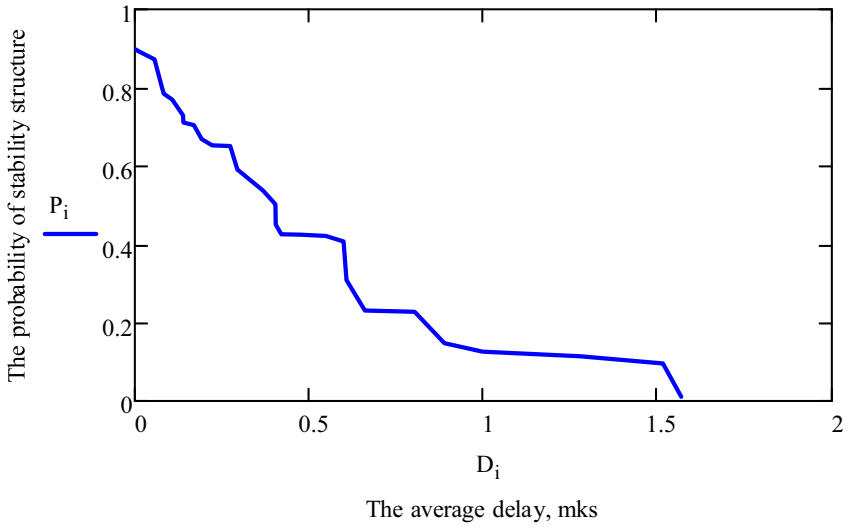

Fig. 2. Dependence of the probability of structure stability on the average total delay.

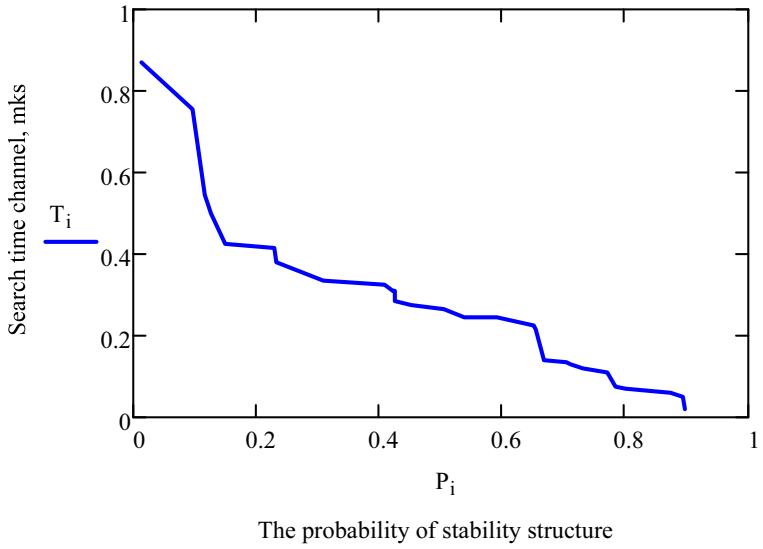

Fig. 3. Dependence of channel search on probability of structure stability

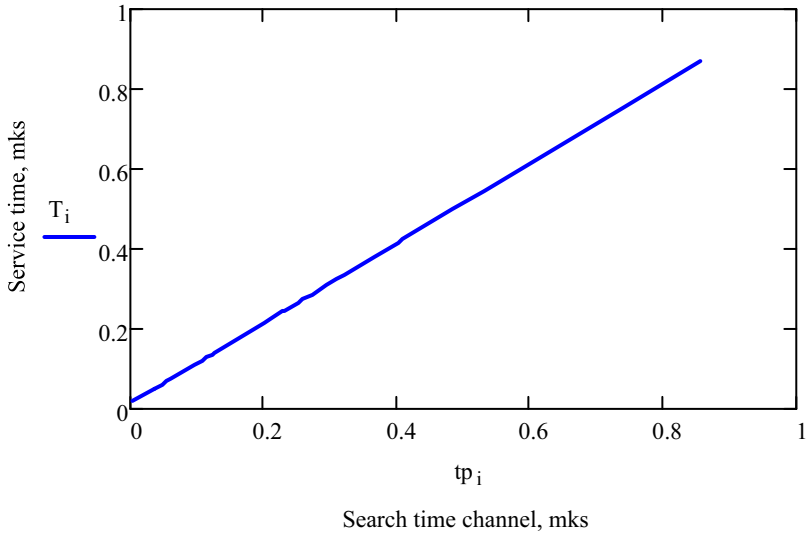

Fig. 4. Dependence of total service transmission time on transmission channels searching time.

The probability of coherency between any two of nodes can be assessed as:

$$P_{зв}(A,B) = 1 - \prod_{i=1}^{n} \prod_{i=1}^{A,B} p_{найкор.шляху_i} p_{ребер.з.\max k_i}, \quad (6)$$

For the method efficiency to be evaluated, it is necessary to form connections at random between the nodes (in accordance with the network model) and assess the probability of coherence of the arbitrary pairs of the nodes first, and then the total coherence of the network that is determined as the sum of coherency of all the node pairs.

The evaluation of the load of each virtual machine during the interval T is done using Norros method considering that each VM in accordance with the theory of queues is a G\G\1system. To simplify the model, we assume that the intensity of incoming requests is described by exponential distribution, and the coefficient of interval variation between the requests and duration of their processing will be constant values. A maximal load must be determined as the maximal number of service requests.

5. Simulation results

With the implementation of this algorithm in Matlab, there has been obtained dependences indicating the impact of the network structure stability assessment on the delay in servicing requests and on the time necessary to search appropriate transmission channels. To carry out the simulation, each atomic service was assumed to be served by a virtual machine at $t_{a.c.}$ = 0.005 ms. Fig. 2 demonstrates the dependence showing that the more stable network structure is, the less the delay is. This speeds up the delivery of services to the end user.

As a result of the simulation, it has been found that with the stability of structure increasing, the time of transmission channels searching reduces that results in decreasing the total time of transmitting services to DC and backwards (Fig. 3 and Fig. 4)

6. Conclusion

Cloud computing is one of the most advanced technologies of distributed computing, which allows users to get all necessary resources. In this paper, we suggest reducing the time of servicing the requests coming to a data center by using the method of assessement of stability of virtualized data centers structure. Moreover, taking into account the intensity of the requests received by VM will give a possibility based on the data concerning the structure stability at a particular time point not to recalculate an optimal transmission route that will lead to reducing service delays. The average delay was established to directly depend on the stability of the structure that can reduce the time of service. In further studies, much attention will be focused on the assessement of the virtual machines workload, and optimization of their resource usage that may result in a more profound effect on request processing in data centers.

References

[1] B. Yang, F. Tan, Y. Dai, and S. Guo, "Performance evaluation of cloud service considering fault recovery," in *First Int'l Conference on Cloud Computing,* pp. 571–576, Dec., 2009.

[2] D. Katsaros, P. Mehra, G. Pallis, and A. Vakali, "Cloud computing: distributed internet computing for IT and scientific research", *IEEE Internet Computing*, vol. 13, no. 5, pp. 10–13, Sept.–Oct. 2009.

[3] G. A. Ptitsyin, "The models of probabilistic collapse of dynamic networks", *Electrical and information complexes and systems,* vol. 2, no. 4, pp. 54–58, 2006. (Russian)

[4] H. Khazaei, J. Mišić, and V. B. Mišić, "Performance of cloud centers with high degree of virtualization under batch task arrivals", *IEEE Transactions on Parallel and Distributed Systems*, vol. 10, no. 5, pp. 1, 2012.

[5] Ye Wu Bang and Kun-Mao Chao, *Spanning trees and optimization problems*, 1th ed. CRC Press, 2004.

[6] R. G. Gallager, "A minimum delay routing algorithm using distributed computation", *IEEE Trans. on communications*, vol. 25, no. 1, pp.73–85, 1975.

[7] A. A. Kochkarov, M. B. Salpagarov, and L. M. Elkanova, "The discrete model of the structural failure of complex systems", *Problems and Management*, vol. 1, no. 5, pp. 21–26, 2007. (Russian)

НАДАННЯ СЕРВІСУ З ВИКОРИСТАННЯМ АЛГОРИТМУ СТІЙКОСТІ СТРУКТУРИ У ВІРТУАЛІЗОВАНІЙ ЧАСТИНІ ЦОД НА ОСНОВІ ХМАРИНКОВИХ ТЕХНОЛОГІЙ

Богдан Стрихалюк, Ольга Шпур, Андрій Масюк

Висвітлено проблему зменшення часу обслуговування (обробки) запитів, які надходять на обслуговування до центру обробки даних, з урахуванням топологічної структури такого центру. Проте врахування структури ЦОД недостатньо, важливим фактором живучість такої структури, адже чим стійкіша структура, що система швидше виконує та перенаправляє користувачеві необхідний сервіс. Для цього у статті запропоновано метод оцінки стійкості структури віртуалізованого ЦОД з урахуванням інтенсивності запитів, що надходять до VM, що дасть змогу на підставі даних про стійкість структури в конкретні моменти часу не здійснювати повторний перерахунок оптимального шляху передачі, що призведе до зменшення затримки в разі надання сервісу.

Bogdan Strykhalyuk – Ph. D, Associate Professor of the Department of Telecom-munications at Lviv Polyte-chnic National University, Ukraine. His research interests include theoretical foundations of construction and operation of next generation networks and cloud-NGN technology.

Olga Shpur – postgraduate student of the Department of Telecom-munications at Lviv Polytechnic National University, Ukraine. Her research interests include design features and operation of networks based on service-oriented architecture, mesh- and cloud-technology.

Andriy Masiuk – postgraduate student of the Department of Telecom-munications at Lviv Polytechnic National University, Ukraine. His research interests include design features and operation of networks based on service-oriented architecture, mesh- and cloud-technology.