THE DEVELOPMENT AND STUDY OF THE METHODS AND ALGORITHMS FOR THE CLASSIFICATION OF DATA FLOWS OF CLOUD APPLICATIONS IN THE NETWORK OF THE VIRTUAL DATA CENTER

Irina Bolodurina[1] and Denis Parfenov[2]

[1]Department of Applied Mathematics Orenburg State University Orenburg, Russia
[2]Faculty of Distance Learning Technologies Orenburg State University Orenburg, Russia

ABSTRACT

This paper represents the results of the research, which have allowed us to develop a hybrid approach to the processing, classification, and control of traffic routes. The approach enables to identify traffic flows in the virtual data center in real-time systems. Our solution is based on the methods of data mining and machine learning, which enable to classify traffic more accurately according to more criteria and parameters. As a practical result, the paper represents the algorithmic solution of the classification of the traffic flows of cloud applications and services embodied in a module for the controller of the software-defined network. This solution enables to increase the efficiency of handling user requests to cloud applications and reduce the response time, which has a positive effect on the quality of service in the network of the virtual data center.

KEYWORDS

Cloud applications; software-defined network; traffic flows; virtual data center; data mining; machine learning

1. INTRODUCTION

Today, telecommunication networks are the basis for deploying various types of applications and services in data centers. The growing use of the cloud computing concept to provide access to applications and services increases the amount of converged network traffic every year. The physical network infrastructure of data centers cannot always adapt timely and scale the existing solutions for users' current tasks. The main problem for data centers is a dynamically changeable structure of the circulating flow of network traffic [1-3]. To improve the efficiency of using and scaling the existing architectures in the data centers, they use solutions based on the virtualization of resources. One of the complex solutions is the creation of the virtual data centers by using a software-defined infrastructure. This solution is based on a software-defined network. However, the existing approaches based on a software defined-network don't provide enough flexible solutions able to adapt to the changes in traffic flows in real time [9-11]. During peak periods, this leads to an overabundance of traffic to specific physical network nodes which are not prepared to handle a large data flow [4].

DOI: 10.5121/ijcnc.2018.10202

We should monitor and analyze circulating traffic for the efficient use of physical network resources. This problem can be solved by using software and hardware solutions for monitoring objects and network resources of the data center. The existing systems for traffic flow analysis can be divided into two main classes: online and offline [5].

The first class includes the systems that analyze traffic in real time. The most significant drawback of these systems is a small amount of data for analysis (about 10-20% of the total amount of data transmitted) [6-7]. Another significant drawback is the limited time for data analysis. To handle user requests efficiently, the system should analyze and classify traffic for a limited period of time, which is set in the network control system in accordance with the requirements of the cloud applications and services. Data for analysis are selected randomly from the general flow of traffic, and the analysis itself is concise.

The second class includes the systems that perform a retrospective analysis of data collected from the network of the data center in the previous time intervals. The second class systems have some problems. The main problems are the relevance of the analyzed data for the current configuration of the network and the storage of the collected data.

In our investigation, we propose a hybrid approach based on a compromise solution, which unites real-time data processing and a retrospective data analysis. We have developed an approach, which aggregates the advantages and eliminates the disadvantages of the traffic analysis systems mentioned above.

The existing solutions are used for analyzing a small list of characteristics: the length of a network packet, the content of the headers of a network packet, a source and destination IP address, and others [8]. However, these approaches do not include characteristics, which describe the state of the physical objects in networks, communication schemes between data sources, physical and virtual routes, and the required parameters of the quality of service (QoS) for selected types of cloud applications and services. Typical solutions are based on a software-defined network using the network controller. In compliance with a standard scheme, the control rules are introduced automatically in the network controller by using the data set by the traffic routing policy. However, this approach does not allow us to have an effective control of the traffic flow routes. In our investigation, we have developed a solution based on the methods of data mining and machine learning. This allows the controller to change dynamically the rules of the traffic routing for cloud applications and services in a virtual data center.

The paper is organized as follows. Section 2 describes the main methods and approaches used in this research. In Section 3, a model of network traffic flows control of cloud applications and services is constructed. Section 4 describes a new algorithm of adaptive routing, which based on the classification of traffic flows of cloud application and services. Results of computational experiments on a comparison of the developed algorithm with other existing solutions are presented in Section 5.

2. RESEARCH METHODS

One of the most important tasks for increasing the efficiency of the network is to analyze the performance of all the components, which support the work of the data center. A significant difference of the virtual data center is the use of virtual networks for routing data flows of cloud applications and services over the existing physical infrastructure. Virtualization at the network level allows us to run simultaneously many different types of applications with different performance requirements for compute nodes, network objects, and QoS. Therefore, we should be aware of the flows circulating in the network at present time to provide the information about the quality of service at the network of the virtual data center. Most modern methods of traffic classification define the flow as a group of packets, which have the same destination IP address and use the same transport protocol and port numbers. This definition allows considering bi-directional flows. Therefore, the packets that transfer requests from the user and answers from compute nodes are a part of the same flow in a network of the virtual data center. The approach to classification based on selected characteristics is a fairly simple task; however, the accuracy of such solutions for modern cloud applications and services are not quite high. In terms of architecture, modern cloud applications are comprehensive objects consisting of many clustered and distributed services. In turn, each cloud service used by the cloud application has its own set of requirements for QoS and dependencies. In the virtual data center, the procedure of user request processing can be represented as a multi-phase queuing system, where data flows are described by different laws of distribution. Thus, the problem of analysis and classification of the traffic flows of cloud applications and services in the virtual data center becomes non-trivial. Another feature of using a software-defined network is the use of dynamic port numbers for routing flows for the same type of cloud application. This feature makes it difficult to classify the traffic of applications by this attribute. The existing solutions based on deep packet inspection (DPI) work slowly and require a lot of processing power.

Another problem of data flows classification is traffic encryption. This feature does not allow us to use this method for analyzing the headers of the packets transmitted in the network. The analysis of research has shown that the methods and algorithms based on machine learning such as K-Means, Bayes filter, and C4.5 have much wider coverage for solving this task. These methods can be used anywhere in the network and can provide very fast detection of traffic flows of the cloud application and services. However, you should have prior training before using them, which leads to lower efficiency at the initial stage of work.

In this investigation, we propose an approach for accelerating the learning process and improving the accuracy for traffic flow classification of cloud applications and services in a network environment of the virtual data center at the initial stage of data analysis.

3. Model Of Network Traffic Flows Control Of Cloud Applications And Services

A classification model for traffic flows of cloud applications and services located in the virtual data center can be divided into the following elements: classification, clustering and identification of association rules. Let's consider these elements of the model separately.

For the effective classification of traffic flows in the virtual data center, we need to determine the set of all applications. Suppose that we have set of cloud applications and services defined as $X = \{x_1, \ldots, x_n\}$. Each cloud application and service is characterized by a set of attributes $X_j = \{a_1, \ldots, a_m, y\}$, where a_i is the observed attributes, whose values represent characteristics of the traffic of cloud application or service; y is the target attribute that identifies the class of a cloud application or service. Each attribute a_i takes a value from some set $A_i = \{a_{i,1}, a_{i,2}, \ldots\}$, which describes valid values of characteristics of the attributes in the subject area under study. In the framework of solving the tasks of traffic classification, suppose a limited number of classes applications $y \in C$ circulate in the network of the virtual data center, where $C = \{c_1, \ldots, c_k\}$. With regard to our task, we will explore the traffic flows of the cloud applications and services according to the communication scheme of their interaction with the network objects within the virtual data center.

To optimize application distribution in the cloud environment of a virtual data center, it is necessary to determine the traffic distribution laws for each application type and distribute the traffic into access objects (virtual servers, containers, and storage systems). For this purpose, it is necessary to set a certain route and make the control law for it within the time interval $T=[t_1,t_2]$.

The dynamic of traffic in cloud applications and services of the software-defined infrastructure of a virtual data processing center can be described by the following discrete system:

$$\begin{aligned} z_{i,j}(t+\Delta t) &= z_{i,j}(t) - \sum_{k=1}^{K}\sum_{l=1}^{N} s_{i,j}(t)u_{i,l}^{j,k}(t) + \\ &+ \sum_{m=1}^{N} s_{m,j}(t)u_{m,l}^{j}(t) + y_{i,j}(t) \end{aligned} \qquad (1)$$

where *N* is the number of virtual nodes within the network; *K* is the number of application types within the network; $s_{i,j}(t)$ is the capacity of the channels between *i*-th computing node and *j*-th storage system $i \neq j$; $y_{i,j}(t)=\lambda_{i,j}(t)\Delta t$ is the traffic volume (the number of user requests) at the moment *t* on the virtual node *i*-th and intended for transferring to the storage system *j*-th; $\lambda_{i,j}(t)$ is the intensity of incoming load, which is defined as the total intensity of the user request flow connected to the virtual node *i*-th and used the storage system *j*-th; $u_{i,l}^{j,k}(t)$ is the part of the channel transmission capacity in a certain segment of the software-defined network*(i,l)* at the moment *t* for the user request flow to the application of type *k*, working with the data storage system *j*-th.

At the stage of clustering, imagine that many classified cloud applications and services X can be divided into a finite number of groups $G=\{g_1,\ldots,g_m\}$. Each group of cloud-based applications and services is characterized by a set of variables $G_j=\{h_1,\ldots,h_k\}$, which describe the network routes in the telecommunication environment of the virtual data center. Each variable h_i takes the value from some set $H_i=\{h_{i,1},h_{i,2},\ldots\}$. The clustering problem is to construct a set $F=\{f_1,\ldots,f_k\}$, where f_i is the cluster with similar objects from multiple cloud-based applications and services X relative to the introduced proximity measures $d(x_j,x_p)$ called the distance, i.e. $cm=\{x_j,x_p \mid x_j \in X, x_p \in X \,\&\, d(x_j,x_p)<\sigma\}$, where σ is a value defining the maximum distance between the elements of the same cluster. With regard to our task, we will explore the route of the traffic generated in the virtual data center to provide cloud applications and services.

For dynamic rearrangement of the routing rules of traffic flows of the cloud applications and services in a software-defined infrastructure of the virtual data center, we need to create a solution and apply a search of association rules. Assume that we have a set of source elements $I=\{i_1,\ldots,i_n\}$, which are predetermined in the controller of a software-defined network. In addition, we have a set of objects $D=\{d_1,\ldots,d_m\}$ describing the network devices, which are involved in traffic routing in the virtual data center. Thus, the group of network devices that are involved in the creation of a selected route for cloud applications is a subset of $I(d_i \subseteq I)$. In accordance with database terms, d_i is called a transaction, and D is the database. A rule is an implication such as $X \Rightarrow Y$, where $X,Y \subseteq D$ and $X \cap Y = \varnothing$. To identify the most plausible rules that reflect common dependencies between transactions in a database, we defined two metrics. The support of the X set denoted as $supp(X)$is the set of X with respect to the entire set of D. The support of the rule $supp(X \Rightarrow Y) = supp(X \cup Y)$. The confidence in the rule is determined by the formula

$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$. The more are the values of support and confidence, the more accurately the rule reflects the dependencies.

Thus, the problem of the efficient control over the traffic of the cloud applications and services with support of the developed model can be solved by constructing the decisive function that puts the class number I, which relates to the application and to the generated data flows in the network of the virtual data center, in correspondence for a particular vector of characteristics (attributes) X.

The model of traffic control can also be represented as a graph. The obtained graph illustrates all possible multi-level and multi-directional network routes of data flow of cloud applications and services according to the communication scheme of their interaction. In this graph, each node represents either precondition or sequence of actions to change the route of traffic in the specified initial conditions. The constructed graph or some of its elements can be used for creating new routing rules in the network of the virtual data center.

Since the graph contains full information about all known routes in the virtual data center and information about the communication schemes of interaction between cloud applications and services, we get a complete map of current traffic flows, which will enable us to predict the network performance by matching events.

The analysis enables to establish whether a particular event is critical for the stable work of the network. If the event is critical, we need to apply the rules of traffic redistribution and other measures of flows balancing and rerouting to mitigate its impact on the network in the virtual data center as a whole and on selected cloud applications and services in particular. The graph can be represented as follows:

$$SAG = (V, E) \tag{2}$$

where $V = Nc \cup Nd \cup Nr$ denotes the set of vertices that include three types, namely, Nc is the set of vertices (network objects across virtual data centers) involved in the current routing of traffic to cloud applications and services, Nd is the set of network objects involved in traffic rerouting and Nr is the result of traffic flows rerouting; $E = Epre \cup Epost$ denotes the set of directed edges. The edge e ($e \in Epre = Nd \times Nc$) suggests that Nd satisfies the reachability of Nc. The edge e, $e \in Epost = Nd \times Nc$ means that the sequence of Nd can be obtained if Nc is performed.

The model of traffic flows control includes a classification of cloud applications in a software-defined infrastructure of the virtual data centers and a clustering of route schemes to balance the flows depending on the types of transmitted data. The developed approach allows considering various schemes for data access including a database level (SQL – cleared data and NoSQL for raw or semi-structured data). To improve the efficiency, we have used a proactive approach to the traffic analysis in a software-defined infrastructure of the virtual data center based on the identification of moments of changing volumes and directions of traffic, as well as on the methods of package mining. To reduce the time of boosting, we have used the maps of the location of cloud applications in the network based on the data received from the controller of a software-defined infrastructure of the virtual data center.

The developed model is used to create the adaptive routing algorithm for the traffic flows of cloud applications and services in the virtual data center.

4. ALGORITHM OF ADAPTIVE ROUTING

From an algorithmic point of view, we can represent classification or clustering of a traffic flow as the function $f : X \to C$, which puts the label $c_j \in C$ in correspondence with each object $x_i \in X$. The set of C is defined in advance. In the task of clustering, neither the set of C nor its dimensionalities are determined. In this research, we have used the classification and clustering of traffic flows of the cloud applications and services to improve the efficiency of routing inside the virtual data center.

The generalized algorithm for classification of the traffic flows of cloud applications and services can be represented as the following sequence of steps:

Step 1. To identify the routes of traffic flows, we use the data received from the controller of a software-defined network, which controls the placement of cloud applications and services in the virtual data center.

Step 2. Basing on the obtained data, we formed a graph in compliance with the communication schemes of interaction between cloud applications and services.

Step 3. The obtained schemes are overlaid on the current topology of the software-defined network to evaluate the network bandwidth, the usage of virtual channels and the analysis of primary transmitted thereon data basing on moments of the packets reception time distribution.

Step 4. The data obtained are ranked in descending order by the load of the communication channel and by the priority of the traffic flows of cloud applications. To adjust the routes, 20% of the most loaded channels are selected.

Step 5. For the selected virtual channels, the traffic flows of cloud applications are classified more thoroughly to identify the degree of channel usage by particular types of applications.

Step 6. For the applications identified in the previous step, we analyze the traffic route, get the parameters of state from physical network devices and identify the most loaded objects.

Step 7. High loaded devices are excluded from the current route and the traffic is redistributed between less loaded nodes by using association rules for routing. These changes are also updated in the communication schemes of interaction between applications and services.

Step 8. The results of the analysis are used to apply QoS policies on the controller of a software-defined network in the virtual data center as well as in a retrospective analysis of data for error correction.

Thus, the main goal of the developed algorithm is to find the optimal solution and to maximize the performance of the physical network taking into account the existing flows of applications and services and their demands for delays in the work of the virtual data center.

5. EXPERIMENTAL RESULTS

To assess the effectiveness of the developed algorithm for optimizing the adaptive routing of data flow balancing in the applications and services in the virtual data center, we have conducted a pilot study. We have chosen the Open stack cloud as the basic platform. For comparison, we have applied the algorithms used in the Open Flow version 1.4, for route control of the software-defined network in the experiment. For the experimental research, a prototype has been created, including basic nodes, as well as software modules for the developed algorithms that redistribute data flows and applications. To verify the developed algorithm of optimal routing and traffic balancing in case of dynamic changes channels in a software-defined network of the virtual data center, several experimental networks consisting of 25, 50, 100, 200, 300, and 400 objects have been deployed. All generated requests were played consequently on two pilot sites: the traditional routing technology (Platform 1, NW) and the technology of the software-defined networks (Platform 2, SDN). This restriction is caused by the need to compare the results to a traditional network infrastructure, which is not capable of dynamic reconfiguration. Two tests were carried out on site 2. In the first case, the model OpenFlow version 1.4 routing algorithms were in use, in the second case (Platform 3, NEW SDN), the developed routing optimization algorithm was applied. The experiment time was one hour, which corresponds to the most prolonged period of peak demand, recorded in a real traffic network of a heterogeneous cloud platform. We have chosen response time of applications and services that work in a cloud platform as a basic metrics to assess the efficiency of the proposed solutions. The results of the experiment are provided in Fig. 1.

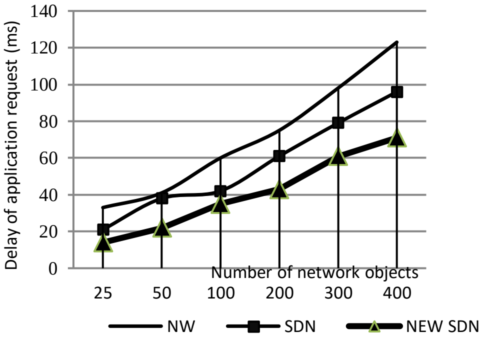

Fig 1.Schedule of dependence of applications and services in a cloud platform response time from the quantities of network objects in the data center.

6. CONCLUSION

The study proposed a model adaptive control of network traffic based on the statistical properties of the flow and defined a systematic approach to the selection of the optimal set of attributes of the traffic flow. The results show that the classification of the traffic flows of cloud applications enable to improve the quality of service by 20-25% by reducing the response time and load on physical network devices. It became possible due to the identification of applications at the initial stage by using the data of the application placement in the network of the virtual data center. We are going to assess our approach with a larger number of flows of applications since it will allow us to assess the accuracy of the proposed solution. However, a software package developed for the simulation of the routing algorithms enables to identify applications and services in the virtual data center and create the optimal routes for data transmission.

ACKNOWLEDGMENT

The research work was funded by Russian Foundation for Basic Research, according to the research projects No. 16-37-60086 mol_a_dk and 16-07-01004, and the President of the Russian Federation within the grant for state support of young Russian scientists (MK- 1624.2017.9).

REFERENCES

[1] Bein D., Bein W., Venigella S. “Cloud Storage and Online Bin Packing” Proc. of the 5th Intern. Symp. on Intelligent Distributed Computing, 2011, Delft: IDC, P. 63-68.

[2] Nagendram S., Lakshmi J.V., Rao D.V., et al “Efficient Resource Scheduling in Data Centers using MRIS” Indian J. of Computer Science and Engineering, 2011, V. 2. Issue 5, P. 764-769.

[3] Arzuaga E., Kaeli D.R. “Quantifying load imbalance on virtualized enterprise servers” Proc. of the first joint WOSP/SIPEW international conference on Performance engineering, 2010, San Josa, CA: ACM, P. 235-242.

[4] Mishra M., Sahoo A. “On theory of VM placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach” Cloud Computing (CLOUD), IEEE International Conference, 2011, Washington: IEEE Press, P.275-282.

[5] Korupolu M., Singh A., Bamba B. “Coupled placement in modern Data Centers” IEEE Intern. Symp. on Parallel & Distributed Processing. N. Y.: IPDPS, 2009. P. 1-12.

[6] Singh A., Korupolu M., Mohapatra D. “Server-storage virtualization: integration and load balancing in Data Centers” Proc. of the 2008 ACM/IEEE Conf. on Supercomputing. Austin: IEEE Press, 2008. P.1- 12.

[7] Plakunov A., Kostenko V. “Data center resource mapping algorithm based on the ant colony optimization” Proc. of Science and Technology Conference (Modern Networking Technologies) (MoNeTeC), Moscow: IEEE Press, 2014. P.1- 6.

[8] Darabseh, A., Al-Ayyoub, M., Jararweh, Y., Benkhelifa, E., Vouk, M., Rindos, A. “SDStorage: A Software Defined Storage Experimental Framework” Proc. of Cloud Engineering (IC2E), Tempe: IEEE Press, 2015. p.341- 346.

[9] Parfenov D., Bolodurina I. “Approaches to the effective use of limited computing resources in multimedia applications in the educational institutions” WCSE 2015-IPCE, 2015.

[10] Parfenov D., Bolodurina I. “Development and research of models of organization storages based on the software-defined infrastructure” 39th International Conference on Telecommunication and signal processing : materials of conference 27-29 June 2016, Vienna, Austria. 2016. - . - P. 1-6.

[11] Parfenov D., Bolodurina I. “Development and Research of Models of Organization Distributed Cloud Computing Based on the Software-defined Infrastructure” Procedia Computer Science, Volume 103, 2017. P. 569-576.