UDC 004.2 **doi: 10.20998/2522-9052.2018.1.09**

M. Skulysh, S. Sulima

The National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

HYBRID RESOURCE MANAGEMENT SYSTEM FOR TELECOMMUNICATION NETWORK

The implementation of modern telecommunication technologies (SDN, NFV, SDR, CloudRAN) will lead to the full dependence of the telecommunication network performance on the information and computing environment efficiency. **The purpose of the article** is to develop the method for managing telecommunication network resources by choosing the optimal interval for redistributing system resources. The proposed model of load management system allows presenting the method of load forecasting, taking into account long-term accumulated data of statistics, and the latest tendencies, which are observed in the network. So it becomes possible to achieve a rational ratio of management costs and the final value of service quality. A new method of monitoring and measurement, using which it is possible to avoid overloading of network functions was developed. **Conclusion.** Experiments to research the proposed methods were conducted in the Mathcad system. The results showed that proposed methods could rationally distribute the system resources, especially under transient conditions of overload.

Keywords: network resource management; load forecasting; dynamic allocation of resources; virtualization of network resources.

Introduction

Today the mobile operator telecommunication network is an organized system that includes the special equipment being serviced, observed and guided by operating data centers, which have installed computing servers and corresponding software that serves numerical information and service streams. Modern technologies SDN, NFV, SDR, CloudRAN and other are fast-growing. Their full-scale implementation will lead to the full dependence of the telecommunication network performance on the information and computing environment efficiency.

Integration of telecommunication systems (TCS) and distributed computing environment (CE) is observed today. As a result, a single heterogeneous service environment for telecommunications services, in which it is possible to control the process of information flow maintenance at each stage and to ensure compliance with high quality standards is created. At the same time, there is still no single concept, models and methods for organizing interactions that take into account the peculiarities of the CE functioning when servicing a large number of queries generated during the provision of telecommunication services. This leads to inefficient use of resources that provide a heterogeneous environment for telecommunication services.

Due to the lack of a methodological basis for organizing the work of a heterogeneous telecommunications system, system resources are chaotic, optimization problems are solved partly or locally, which leads to deterioration of control and QoS parameters for end users.

In this article the method of managing the resources of a telecommunication network by choosing the optimal interval for redistribution of system resources is proposed.

This method allows reducing the amount of redundant service information transmitted over the network and unloading the network nodes by dynamic adjustment of the system.

Formulation of the problem

Formulate the problem of optimal resource distribution when servicing the virtual network functions of NFV. The purpose of the method of providing resources is to allocate their sufficient number for network functions, so that their SLA can be met even in the periods of peak load. The algorithm involves solving two main tasks: finding the required amount of resources and the right time to provide them.

To solve the question of necessary amount of resources for each network function, an analytical model is constructed. The presented model adopts the input data intensity of incoming requests and the requirements of serving a separate request, and calculates the amount of resources required by each network functional block to meet requirements.

The decision on when to provide resources depends on the dynamics of loads. Telecommunication loads undergo long-term changes, such as daylight hours or seasonal effects, as well as short-term fluctuations such as crowd outbreaks. While long-term fluctuations can be predicted in advance, observing changes in the past, short-term fluctuations are less predictable, and in some cases even unpredictable. The proposed method uses two different approaches for working in conditions of changes that are observed at different time scales. Proactive resource management is used to assess the load and appropriate management, as well as reactive resource management to correct errors in long-term forecasts or to respond to unexpected outbreaks of the crowd.

Consider a network with several network functions. It is assumed that each such network function indicates the desired quality of service (QoS) requirement; in this case, we assume that the QoS requirements are defined in terms of the target response time. The purpose of the system is to ensure that the average response time (or some response time percentile) observed by the network function requests does not exceed the desired target response time. In general, several hardware and software resources on the server, such as CPU, NIC, disk, etc, service each incoming request. Assume that the given target response time is divided into several response time values for specific resources one for each such resource. Thus, if each request on each resource does not spend more time than the target value, then the overall target response time for the server will be satisfied. For ease of presentation, we assume that there is only one type of resource in the system.

Formally, d_i denotes the target time of the network function response; i and T_i – the observed average response time, then the network function needs to allocate such amount of resources to satisfy the condition $T_i \leq d_i$.

We use this formulation of the problem to obtain a mechanism for the dynamic allocation of resources, which is described below.

Literature review

Although the NFV promises substantial cost savings, flexibility and ease of deployment, potential problems in implementing virtualized network elements that can meet the demands of real-world performance are still open issues, and the NFV is still in its early stages of implementation.

Several research papers have focused on developing adaptive systems that can respond to changes in the load in the context of storage systems, common operating systems, network services, web servers, and Internet datacenters.

This article discusses the abstract model of the server resource and presents the methods for dynamically allocating resources. The proposed model and resource allocation methods are applicable to many scenarios where the system or resource can be abstracted using a GPS server.

One of the key issues in the virtualization area of the network is the allocation of physical resources to the virtual network functions. Virtual Network Embedding (VNE) is a well-studied task. Nevertheless, most modern solutions offer a static resource allocation scheme, in which, when a virtual network is displayed, the redistribution of resources does not occur throughout its life cycle. There is a limited number of decentralized and dynamic VNE solutions (like [1] or [2]). Moreover, even the approaches that offer solutions for dynamic virtual network embedding still allocate a fixed amount of resources for virtual nodes and channels for the entire period of existence. As network traffic is not static, this can lead to inefficient use of shared network resources, especially if the physical network rejects the requests for embedding new virtual network functions, while reserving resources for virtual network functions that are in low load conditions [3].

Most of the existing work on dynamic resource management is based on three approaches: management theory, modeling of work dynamics and load forecasting [3]. Among adaptive systems using a method based on the theory of control [4]). Among works based on the dynamics of work - [5]. The authors [6] use prediction of the load.

Summing up, the difference between the approach proposed in this article and the described above is that the resources reserved for use with virtual network functions do not remain unchanged throughout the life of the virtual network. The monitoring of virtual nodes is carried out, and, resources are redistributed, based on real resource needs. In this case, unused resources are returned to the physical network for use by other virtual networks.

The next problem that arises is how information is obtained about the current situation on the network. In this aspect, resource management in NFV networks is similar to managing application programs in data centers and clouds. Existing solutions of server resource management can be classified as predictive and reactive solutions. The prognostic allocation of resources involves the presence of a predictable and stable template in requirements and distributes volumes, usually on a time scale of several hours or days based on a template. However, large, unpredictable bursts of requirements can cause serious SLA violations. On the other hand, the reactive allocation of resources allocates resources at short intervals (for example, every few minutes) in response to load changes. Reactive policies can quickly respond to changes in load, but problems such as unpredictability, instability and high management costs limit their application in practice [7].

Therefore, predicting peak application load and allocating resources based on the worst case scenario is extremely difficult [8]. Given the difficulty in predicting peak loads, an application program should use a combination of predictive and reactive control. While prognostic methods work well for online prediction in large time intervals from a few minutes to several hours, reactive methods can predict load for short time intervals up to several minutes and respond quickly to non-stationary overloads [9].

Several approaches combine prognostic and reactive control [7, 8]. Although these approaches have the common features with the hybrid approach proposed in this article, they differ in several aspects. The offered approach is directed on productivity optimization, energy consumption and cost allocation of resources simultaneously. Based on the hybrid system of resource management, a dynamic monitoring method is developed to effectively manage network resources and reduce the amount of service information - resource management intervals have variable lengths, while in other management approaches, simple fixed intervals are used. [7] similarly uses a variable length of intervals, but in contrast to this approach, the approach proposed in the article defines the length of the intervals dynamically, depending on the actual situation on the network.

The main objective is to develop a system that can respond to transient load changes, while a theoretical mass-service approach tries to plan queries based on stationary load. The resource management method based on the model that performs resource allocation based on resource modeling to correlate QoS metrics and resource particles allocated to the application was proposed in [9]. This work is similar to the proposed approach, but in it the model of mass service in the time domain is adapted to the task of virtualization of network functions, and also supplemented by the mechanism of measurement and prediction.

Description of the method

To perform a dynamic allocation of resources each server will need to use three components:

1) a monitoring module that measures the load and performance of each network function (such as the intensity of receipt of requests, the average response time, etc.);

2) a forecasting module that uses measurements from the monitoring module to assess the load characteristics in the near future;

3) a resource allocation module that uses these estimates of load to determine the amount of resources that should be allocated by network functions.

Fig. 1 shows these three components [9].

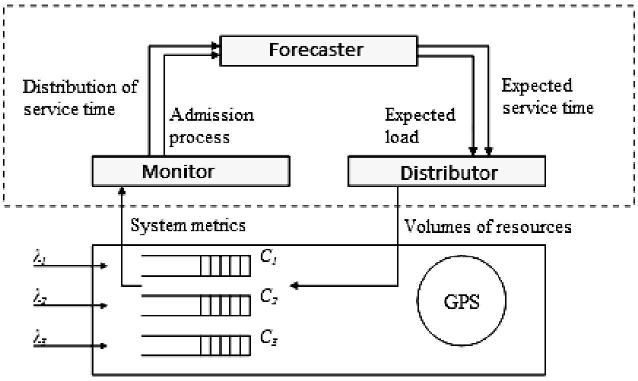


Fig. 1. Dynamic resource allocation system

Using traditional methods of monitoring the state of network resources, excessive service information increases significantly, which can negatively affect the overall performance of the network due to the capacity of the channels. Therefore, it is proposed to apply a mechanism, the essence of which is to dynamically change the intensity of the control of the network element state, depending on the difference between the predicted value of load and the actual. This equation describes the principle of changing the frequency of monitoring:

$$W(t) = I_{base} - \\ -K \cdot \sum_{j=t-h}^{t-1} \frac{\max(0;\lambda_{obs}(j)-\lambda_{pred}(j))}{h} I_{base}, \qquad (1)$$

where W – the control interval,

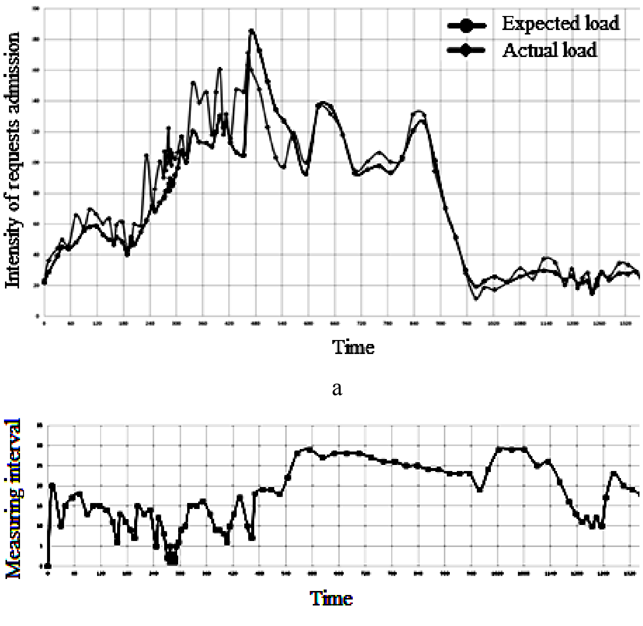
I_{base} – the base value of the interval,

K – the normalization constant,

$\lambda_{obs}(t)$ – the real intensity of the load flow during the interval t,

$\lambda_{pred}(t)$ - the provided intensity of the load receipt at the interval t.

Fig. 2 shows the adaptation of the frequency of the network element status control to the rejection of the real load from the predicted on the network element.



a

b

Fig. 2. Dynamic change of the adaptation frequency (a) Load of the network element; (b) Change the intensity of adaptation in accordance with the deviation of the actual load from the expected

The resource allocation module is called periodically (each window of adaptation or when threshold is reached) to dynamically divide the resource volume between different network functions that work on common servers on the network.

As already mentioned, the algorithm of adaptation is started every W time units. Let q^0_i is the length of the queue at the beginning of the adaptation window; λ_i is an estimate of the rate of receipt of applications, and μ_i denotes an assessment of the intensity of service in the next window of adaptation (that is, the next W timelines).

Then assuming that the values λ_i and μ_i are constant, the queue length at any time t within the next adaptation window is given by equation:

$$q_i(t) = \max(0; q^0{}_i + (\lambda_i - \mu_i)t). \qquad (2)$$

As the resource is modeled as a GPS server, the intensity of the network function request service is

$$\mu_i = C_i/s_{i,},$$

where C_i is the number of resources of the network function and s_i is the average service time of the request by one resource unit.

The average length of a queue in the window of adaptation is determined by equation:

$$q_i = \frac{1}{W}\int_0^W q_i(t)dt. \qquad (3)$$

The average response time T_i at the same time interval is estimated by equation:

$$T_i = \frac{q_i + 1}{\mu_i}. \qquad (4)$$

Parameters of such a model depend on its current characteristics, so this model is applicable in the online scenario for responding to dynamic changes in the load.

The network functions need to allocate the number of resources, so that

$$T_i \le d_i,$$

then the amount of resources allocated by the network function C_i must satisfy the condition of equation:

$$C_i \ge s_i \frac{q_i + 1}{d_i}. \qquad (5)$$

A modified load factor predictor based on the method proposed in [8] uses past load monitoring to predict peak demand that will occur over time W.

Assume that $\lambda_{pred}(t)$ – the predicted intensity of receipt during a certain interval t, that is obtained from the analysis of historical data for the past days. Let $\lambda_{pred}(t)$ is the real intensity of the flow during this interval.

The predicted value for the next interval is corrected using the observed error in accordance with equation :

$$\lambda(t) = \lambda_{pred}(j) + \sum\nolimits_{j=t-h}^{t-1} \frac{\lambda_{obs}(j) - \lambda_{pred}(j)}{h} \qquad (6)$$

Conducting an experiment

Consider the problem in the Mathcad system. Let us consider the work of one block in one day (1440 minutes).

We assume that basic component of the predicted intensity of applications admission during each minute is λ_{pred}, and the average value of the service life of the application be respected by one unit resource s_i and it does not change, we also assume the availability of a resource of the same type .

Let the adaptation window dynamically determine based on its base value and the last four values from the monitoring history, where the base value of the adaptation window is 5 minutes.

We do not introduce restrictions on the queue size, but for calculations, we assume that at the end of the adaptation interval, the theoretical value of the queue size in the current conditions is 40 queries.

Fig. 3 shows a change in the value of the control interval and the change in the predicted value of the load intensity compared to the actual load received.

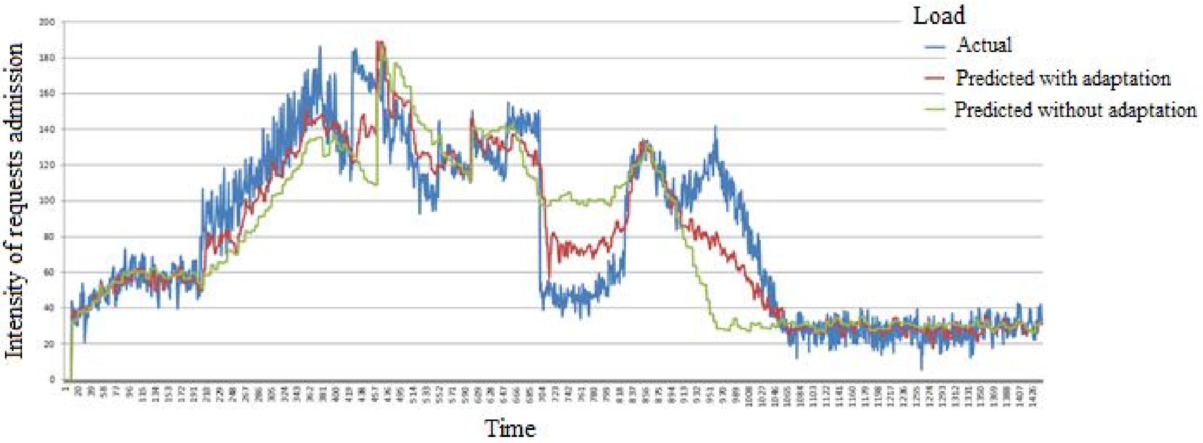


Fig. 3. Results of system modeling with dynamic adaptation of control interval and load forecasting and systems without them

The simulation results showed that the error in the predicted value compared to the real one could be 16%, while the "positive" error is 9% of the actual rate of receipt of applications.

If you do not apply the dynamic adjustment system for the size of the adaptation window and the system for incorporating historical data into forecasting, then the error will be 26%, and the "positive" error will be 15%, which can further be drawn from the conclusion that resources will be extremely ineffective.

Conclusions

The article proposes a method of dynamic allocation of resources of the telecommunication network for increasing the efficiency of its work. A system that combines online measurements with forecasting and resource allocation methods was presented. In particular, a new method of monitoring and measurement, using which it is possible to reduce the amount of service information circulating in the network and to avoid negative phenomena of overload of network functions has been developed. The estimation of methods using modeling in the system Mathcad was carried out.

The results showed that these methods could rationally distribute the system resources, especially under transient conditions of overload.

The system can be used to control the deployment of virtualized network functions on the underlying physical infrastructure to minimize operator costs and improve customer service quality.

REFERENCES

1. Cai, Z., Liu, F., Xiao, N., Liu, Q. and Wang Z. (2010), “Virtual network embedding for evolving networks”, *Global Telecommunications (GLOBECOM 2010): IEEE conference*, Miami, Florida, pp. 1–5.
2. Sun, G., Yu, H., Anand, V. and Li, L. (2013), “A cost efficient framework and algorithm for embedding dynamic virtual network requests”, *Future Generation Computer Systems*, Vol. 29, No. 5, pp. 1265–1277.
3. Mijumbi, R., Gorricho, J.-L., Serrat, J., Claeysy, M., Turcky, F. D. and Latr S. (2014), “Design and Evaluation of Learning Algorithms for Dynamic Resource Management in Virtual Networks”, *Network Operations and Management Symposium (NOMS)*, *IEEE,* Krakow, pp. 1–9.
4. Patikirikorala, T., Colman, A., Han, J. and Wang L. (2011), “Multi-model framework to implement self-managing control systems for QoS management”, *Software Engineering for Adaptive and Self-Managing Systems: symposium*, Waikiki, Honolulu, pp. 218–227.
5. Lai, W., Chiang, M., Lee, S. and Lee, T. (2013), “Game Theoretic Distributed Dynamic Resource Allocation with Interference Avoidance in Cognitive Femtocell Networks” *Wireless Communications and Networking: IEEE conference*, Shanghai, pp. 3364–3369.
6. Jokhio, F., Ashraf, A., Lafond, S., Porres, I. and Lilius, J. (2013), “Prediction-Based Dynamic Resource Allocation for Video Transcoding in Cloud Computing”, *Parallel, Distributed, and Network-Based Processing: 21st Euromicro International Conference*, Belfast, pp. 254–261.
7. Gandhi, A., Chen, Y., Gmach, D., Arlitt, M. and Marwah, M. (2011), “Minimizing Data Center SLA Violations and Power Consumption via Hybrid Resource Provisioning” *Green Computing: International Conference and Workshops*, Orlando, FL, pp. 1–8.
8. Urgaonkar, B., Shenoy, P., Chandra, A., Goyal, P. and Wood, T. (2008), “Agile dynamic provisioning of multi-tier Internet applications” *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, Vol. 3, No. 1, pp. 1–39.
9. Chandra, A., Gong, W. and Shenoy, P. (2003), “Dynamic resource allocation for shared data centers using online measurements”, *Quality of service : 11th International Workshop*, Berkeley, pp. 381–398.
10. Koval, A., Globa, L. and Novogrudska. R. (2016), “The approach to web services composition” *Hard and Soft Computing for Artificial Intelligence, Multimedia and Security*, Vol. 534 of the series Advances in Intelligent Systems and Computing. – Springer international publication, pp. 293-304.

Надійшла (received) 27.02.2018
Прийнята до друку (accepted for publication) 18.04.2018

Гібридна система управління ресурсами телекомунікаційної мережі

М. А. Скулиш, С. В. Сулима

Впровадження сучасних телекомунікаційних технологій (SDN, NFV, SDR, CloudRAN) призводить до повної залежності продуктивності телекомунікаційної мережі від ефективності інформаційного та обчислювальної середовища. **Метою статті** є розробка способу керування ресурсами телекомунікаційної мережі шляхом вибору оптимального інтервалу для перерозподілу системних ресурсів. Запропонована модель системи управління навантаженням дозволяє представити метод прогнозування навантаження з урахуванням довгострокових даних статистики і останніх тенденцій, які спостерігаються в мережі. Це дозволяє досягти оптимального співвідношення витрат на управління і якості обслуговування. Був розроблений новий метод моніторингу та вимірювання, за допомогою якого можна уникнути перевантаження мережевих функцій. **Висновок.** Експериментальні дослідження запропонованих методів були проведені в системі Mathcad. Результати показали, що запропоновані методи дозволяють раціонально розподіляти системні ресурси, особливо при перехідних умовах перевантаження.

Ключові слова: керування ресурсами мережі; прогнозування навантаження; динамічний розподіл ресурсів; віртуалізація мережевих ресурсів.

Гибридная система управления ресурсами телекоммуникационной сети

М. А. Скулиш, С. В. Сулима

Внедрение современных телекоммуникационных технологий (SDN, NFV, SDR, CloudRAN) приводит к полной зависимости производительности телекоммуникационной сети от эффективности информационной и вычислительной среды. **Целью статьи** является разработка способа управления ресурсами телекоммуникационной сети путем выбора оптимального интервала для перераспределения системных ресурсов. Предложенная модель системы управления нагрузкой позволяет представить метод прогнозирования нагрузки с учетом долгосрочных данных статистики и последних тенденций, которые наблюдаются в сети. Это позволяет достичь рационального соотношения затрат на управление и качества обслуживания. Был разработан новый метод мониторинга и измерения, с помощью которого можно избежать перегрузки сетевых функций. **Вывод.** Экспериментальные исследования предложенных методов были проведены в системе Mathcad. Результаты показали, что предложенные методы позволяют рационально распределять системные ресурсы, особенно при переходных условиях перегрузки.

Ключевые слова: управление ресурсами сети; прогнозирование нагрузки; динамическое распределение ресурсов; виртуализация сетевых ресурсов.