UDC 004.043

PLANNING THE LOADING OF DATA CENTERS' RESOURCES BASED ON DOWNLOAD STATISTICS[1]

Globa Larysa prof, Dr.; Skulysh Mariia, PhD
National Technical University of Ukraine "Kiev Polytechnic Institute"
Kiev, Ukraine, mskulysh@gmail.com

ПЛАНУВАННЯ ЗАВАНТАЖЕННЯ РЕСУРСІВ ЦЕНТРУ ОБРОБКИ ДАНИХ НА ОСНОВІ СТАТИСТИЧНИХ ДАНИХ

Глоба Л.С., д.т.н., професор; Скулиш М.А., к.т.н.
Національний технічний університет України «Київський політехнічний інститут», м. Київ, Україна

Introduction

Modern infrastructure of the communication provider implies data center availability that allows simultaneous maintenance of a large number of subscribers with applications of different kind. However, the assistance as well as modernization of the infrastructure is rather cost-intensive for communication provider. Meanwhile the cloud computing resources can be flexibly enabled or disabled at the specific time, thus paying only for the utilized resource. Today the communication providers encounter two main realities that affect the processes of effective operation. The first is related to the effective usage of infrastructure, for instance, by applying the cloud technologies that allow increasing capabilities of own data centers with the use of leased ones. The second reality includes the significant increase in range and variety of provided services, the maintenance of which enlarges the load on the data centers of the operator. Thus, the operator should optimize the resources of the infrastructure in order to ensure the required quality of service (QoS) of subscribers' maintenance.

Currently, all the more and more widespread became hybrid Cloud infrastructure. A hybrid Cloud [1, 2, 3] is the integration and utilization of services from both public and private Clouds. The hybrid Cloud platform will help scientists and businesses to leverage the scalability and cost effectiveness of the public Cloud by paying only for IT resources consumed (server, connectivity, storage) while delivering the levels of performance and control available in private Cloud environments without changing their underlying IT setup. As a result, hybrid Cloud computing has received increasing attention recently. While using the hybrid clouds, the different solutions of stream management in the datacenter can be used. For this purpose, the part of streams is maintained by using the leased resources provided according to the IaaS technology (Infrastructure as

[1] http://radap.kpi.ua/radiotechnique/article/view/1198

a Service). Consequently, by using the hybrid infrastructure the amount of system resources can be regulated according to the load. Depending on the load the infrastructure can use various resource sets. Whereas there is some constant volume of own resources, the amount of leased resources can change on demand or in the prescribed time the required infrastructure can be added based on the IaaS technology.

The IaaS technologies have been examined in [1, 5], where the approaches of optimally chosen leased resources' have been described according to the criteria of minimal expanses of effective fulfillment as well as scalability (scheduling approach seems to perform very well both in terms of cost minimization, feasibility and scalability), the range of resource management optimization problems have been solved.

Work [4] proposes capacity allocation techniques able to minimize the cost of the provided Cloud resources at multiple providers, while guaranteeing Quality of Service (QoS) constraints. The rationale is to provide the distribution of workload over multiple IaaS providers and then to implement capacity allocation of multiple class of requests at each provider on a long-term (1 hour) time scale.

In spite of numerous newly proposed platforms for Cloud federation with dissimilar motivations in addition to incentives for parties to connect it [6], a lot of primary problems and questions regarding federation remain unanswered. One of these problems is deciding at what time providers ought to outsource their local requests to additional participants of the federation or how many and at what charge they ought to offer resources to the federation. The outsourcing difficulty is not measured only in the framework of federated clouds; it was also investigated as a means of rising capability or scalability of applications in hybrid Clouds [7], grid environment [8], and clusters [9].

In paper [10] the researchers present a profit-driven strategy for decisions correlated to outsourcing or selling idling resources. According to the authors, providers have the choice of shutting down idle nodes of the data center to save power. Though, they did not catch into account diverse types of virtual machines (e.g. on-demand and spot) in addition to probable actions like terminating low priority virtual machines. A consumer satisfaction-oriented scheduling algorithm for serving requests was developed in [11]. Such an algorithm tries to exploit Cloud providers' revenue by accepting as many service requests as it can, as long as QoS is reserved at a certain level. In this view, contracting with additional service providers was taken into explanation as a technique to avoid rejection of consumer requests.

However, by using such solutions the amount of the required resources as well as point of time and the resource lease time should be clearly estimated since the load upon data center isn't homogeneous during the day. Additionally the balance between the amount of leased resources as well as QoS should be

calculated to minimize the permissible amount of service denials. In the current article the approach of resources utilization is under consideration that comprises two management methods for involved serving nodes. The one is utilized for current control of the resource sufficiency, permits evaluating of the dynamics of the input load based on the short-term statistics as well as current state of the technical means. The second approach is based on the long-term statistics that allows planning the additional resources involvement during the load peaks.

Problem statement. The distinguished feature of data-centers of communication operator is that the maintenance of subscribers applications includes the common procedures set, changing depending on the kind of service. The quantity of different applications is inherently countable and finite that allows calculating of the average number of applications, that can be maintained simultaneously during the maximum admissible loads of system resources.

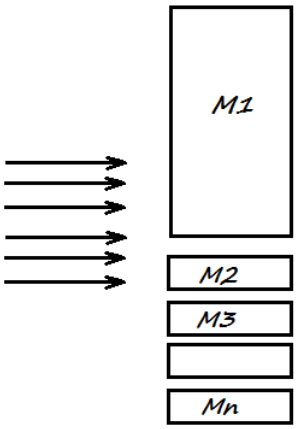

Fig. 1. Service model in the data center with hybrid infrastructure

The applications' maintenance can be considered as a classical queuing system. However, the distinguishing feature of the hybrid infrastructure is scalable service facility, i.e. the service parameter depends on the queue length. In works [12, 13] the operational methods of such systems have been presented.

Let the input flow goes into service in the operator's data center. The point of time when the additional leased resources are involved should be determined. In fig.1 the system model is schematically presented, where the main data center is considered to process M1 applications. Meanwhile the additional resources can be implicated to serve *M2, M3,..Mn* applications correspondingly.

System model

The component-based diagram in Fig. 2 illustrates the software architecture employed within the Hybrid Cloud Construction and Management (HICCAM) project. In this section, there is a brief overview of this architecture presented in [1] paper.

In the current article the proposed architecture of the hybrid infrastructure of the service provider with the improved functionality Optimization Engine has been used. The Optimization Engine includes two parts:

B1 – traffic distribution block the same as in article [1];

B2 – the additional decision-making block for inserting or removing the resources.

The challenges of workflow management of the billing system on the quantity of technical facilities have been described below. Let's consider the example solution of changing facilities' management problem based on the example of billing system that according to [14] became a "bottleneck" in the data center

operation of both communication and Internet access providers.

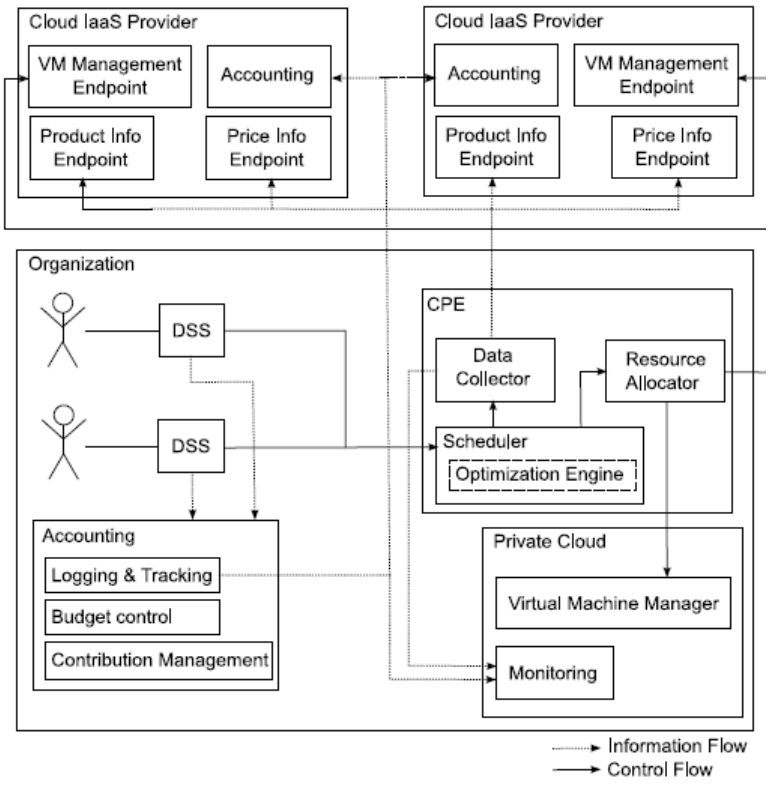

Fig. 2 Schematic component view of the HICCAM model

One of the obstructions to the external resources mass use in online billing systems is the absence of the corresponding methods and algorithms of well- organized workflow management of server group that ensure the operation of the whole online billing system.

Let's consider two algorithms that provide failure-free and economic operation of billing system. The cost effectiveness lies in the involvement of required quantity of servers ensuring the workflow of billing systems and subsystems according to the statistical data on input traffic.

The method of time point designation of additional resources involvement.

Communication operator should create a plan of servers inclusion that would meet all requirements of subscribers in the current time point accounting for the condition of lease payment's cost cutting of servers maintaining for used resources, located in the clouds or power resources for ensuring workflow of own servers.

The method of time point designation of additional resource (server) involvement concludes that within the specified time point the current statistics is evaluated according to which the linear approximation of the maintained input applications' number is built. Then the probability of exceeding of permissible limit of applications' number is forecasted. In this case the permissible limit is the acceptable number of applications that can be maintained by already used resource.

As for the time, for which the estimation is provided, let's choose the occupancy time point of one more resource that consequently permits involving the required number of resources.

The method is applied during the system monitoring in order to reveal the time points when the additional resources should be engaged.

The current problem can be solved for both the entire system and for subsys-

tems. In the first case the amount of billing applications' statistics is evaluated, that arrive in the online billing system. The second case occurs when the amount of application submissions to the subsystems is estimated. The choice of the place where the method should be used depends on the structure of the billing system, i.e. the number of the involved servers or other hardware resources either physical or virtual.

Hence, the probability of not exceeding the permissible limit of applications' number during the specified time period can be determined based on the current traffic statistics, created according to the subscribers' billing applications as well as estimation of the upper limit of the applications' number that can be simultaneously maintained by utilizing the capacities of available servers.

a) The resource sufficiency control algorithm for the application processing system

Input data:

- time interval *T1* – the period of time, during which the statistics analysis is handled;
- time interval *dt* – sampling time (small time interval);
- the data of monitoring system regarding the number of applications, submitted during small time intervals *(x(t0-T1), x(t0-T1+dt), x(t0-T1+2dt)…);*
- time period, during which servers can be occupied – *T*;
- permissible bit error rate;
- *M* – maximum allowed value of applications that can be processed while using the current amount of involved servers (resources).

The involvement of server is required in case if the probability of resource scarcity exceeds the permissible value (see fig.3).

The algorithm includes three main steps:

Step 1. Statistical data analysis during the time period (t0-T1), where t0 – is the current point of time, for which the calculation is accomplished. Based on the statistical data of monitoring system for the couple of values (t, x) the estimated coefficient $\hat{a}$ is calculated for the line (1) using the least square method:

$$x = \hat{a}t + b \tag{1}$$

Step 3. Bit error rate estimation during the specified period of time is calculated according to (2)

$$P_T \le 2 * P(N(0,1) \ge y) \tag{2}$$

where P_T – the probability that during the time period T the number of applications won't exceed the tolerable threshold *M*;

P(N(0,1)≥y) – the probability that stochastic process characterized by normal distribution $N(Ex, \sigma^2)$ will exceed the value of y and can be found from Laplace's function table*:*

$$y = ((M - x0) - \hat{a} * T) / \sqrt{T}$$

x0 – current traffic load on the system,

M – the tolerable traffic load on the system.

Step 4. If the probability PT exceeds the permissible threshold, the additional server is engaged.

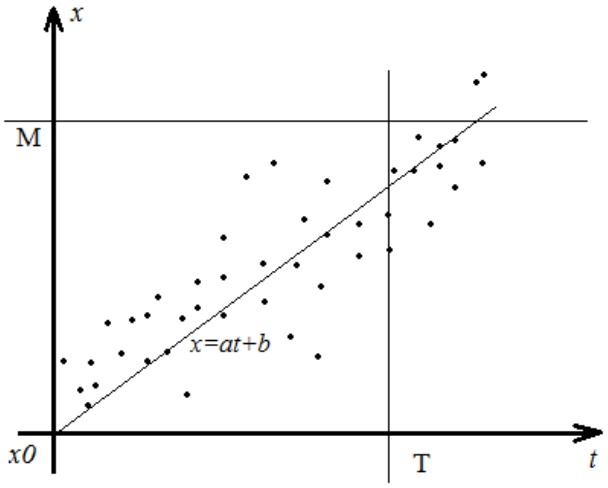

Fig. 3 The traffic dynamics analysis on the online billing server

By using the proposed algorithm the dynamics of traffic increase on the online billing server can be controlled as well as the time point of the additional resources involvement can be determined. Time interval T is the required time that lasts from the operation start to the full operational capability of the additional server.

As can be observed from fig.3 the line $x=at+b$ goes under the point with coordinates (T, M) since the current method accounts for the structure of the stochastic process as well as the mean square deviation that characterizes the stochastic process of billing application submission.

The abovementioned function is the part of monitoring system and starts with the specified periodicity to ensure reliable and unfailing system operation, exerts control over the resource sufficiency.

Scheduling method of servers' involvement

In order to ensure the unfailing billing server operation the set of means should be developed that account for both the current situation as well as the long-term statistics. Hence, the amount of resources can be planned as well as the traffic load balancing methods can be elaborated etc.

The the pocess of schedule table development include two stage. The first stage is to split into time lines with same dynamics of change in a random process. The second stage is to determine the number of servers required to service requests

a) Algorithm of the total time period splitting into time lines

Step 1. Long-term statistics analysis should be conducted. It is necessary to calculate the average load $\bar{x}$, for example, the number of applications that are received an average 15-20 Mondays from 8:00:00a.m. to 8:00:01a.m. The data are summarized as it is shown in table 1.

Table1

T	00:00:00	00:00:01	00:00:02	00:00:03	…
$\bar{x}$	$\bar{x}$(t)	$\bar{x}$(t)	$\bar{x}$(t)		

Step 2. The server loadtime T should be specified as well as arbitrary small quantity $\varepsilon_1>0$;

Step 3. The set of admissible values t should be split into subsets t_i so that $t_{i+1}-t_i=T$. At the beginning of algorithm execution the total number of matrices makes up $n=24\cdot3600/T$, where $(i+1)$-th matrix is given in the form as shown in table 2:

Table 2

T	t_i +00:00:01	t_i +00:00:02	…	t_{i+1}-00:00:01	t_{i+1}
$\bar{x}$	$\bar{x}$ (t)	$\bar{x}$ (t)		$\bar{x}$ (t)	$\bar{x}$ (t)

Step 4. By applying the least square method to each i-th matrix the coefficient $\hat{a}$ estimation of the approximated line (1) is calculated for the couple of values ($\bar{x}$, t).

Step 5. For all $i=1,..n$ the estimation of $\hat{a}_i$ is conducted. If $|\hat{a}_i-\hat{a}_{i+1}|<\varepsilon_1$, then the sets i та $(i+1)$ are assembled. The obtained sets are reindexed and the amount of new sets are denoted as n_{new}.

Afterwards, there is a shift to Step 4. Otherwise, in case if the following condition $|\hat{a}_i-\hat{a}_{i+1}|>\varepsilon_1$ is fulfilled for all $i=1,..n_{new}$, define n_{new} as n_{last}. The set partitioning has been found.

The proposed algorithm results in the $\{t_i\}$ (i=1,.. n_{last}), that represents the time splitting into the periods with constant of variation. The time points t_i (i=1,.. n_{last}) are the timeperiods of dynamics' variation of input (fig.4).

b) In order to design the necessity of additional resource insertion (or removement) for expected traffic load processing, the following points are required:

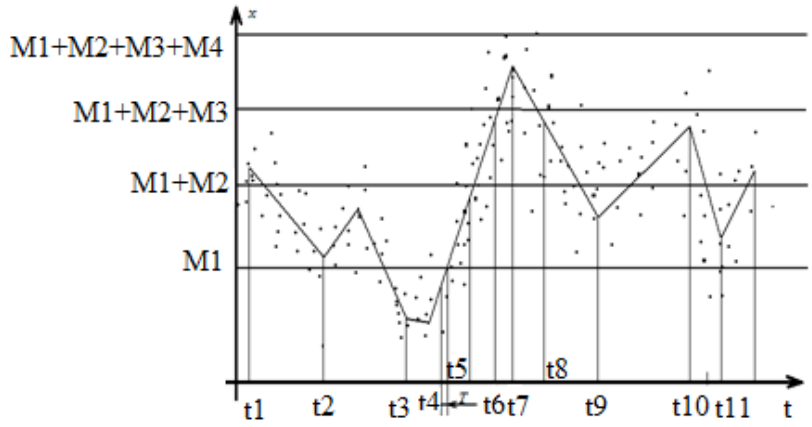

Fig.4 The dynamic variation of the applications' amount during the day

- the amount of resources for load maintenance of $\bar{x}(t_i)$ should be determined;
- if the amount of operating resources for $\bar{x}(t_i)$ and $\bar{x}(t_{i+1})$ load maintenance differs by more than one, then the time interval (t_i,t_{i+1}) should be split so that the amount of resources would be sufficient for each time interval. Otherwise, the amount of involved resources may stay unchangeable for the point of time t_{i+1}.

The second stage results in matrix that shows the switching time as well as the amount of additional resources (servers) required to process the applications

flow and determines the switch on/switch off operation.

The execution of the algorithm is only the component part of the technical support of online billing system since except for the static schedule of servers' involvement the load balance problem between servers requires considerations.

Performance evaluation

The variety of problems that can be solved by applying the proposed method and algorithms is rather wide. The main feature of systems for which the proposed method can be applied is the execution of a great number of procedures, where the initiators can be people, software or services. The procedures' completion is accomplished by using the software of the server while utilizing the technical resources of the system.

Today the overloading problem can be solved by debarring the system from redundant billing applications. Hence, in case if the server is overloaded, the specific signal comes to the control device resulting that the billing applications are temporally rejected. As a rule, the communication provider operates under the condition that in a steady state 20-30% resources are not used. The mentioned resources are reserved for usage in case of overload.

In order to confirm the effectiveness of the proposed method was used to simulate the billing system using three servers that served the flow of applications for billing. The aim was to, tracking download server memory, the ability to quickly identify overload and take measures to attract additional servers. The simulation was performed using GPSS packet.

In order to prove the proposed algorithm's effectiveness the simulation has been conducted in *two modes*:

- The limited technical resource has been allocated for maintenance, i.e. in case of overloading the applications were discarded (the mode in which the proposed methods aren't used).
- The applications' processing has been hold while using between one and three servers with identical technical resources. Based on the statistical sampling of the communication provider the schedule of switching on/off of the servers has been built according to the proposed algorithm. On the ground of the obtained schedule the billing system has been simulated with the input flow that is maximally brought into the proximity with the reality. The servers' inclusion was conducted according to the schedule. According to the proposed method of additional technical equipment's inclusion the sufficiency of the resources has been checked every three minutes in order to maintain the input flow.

From the results of simulation the following conclusions have been made:

1) Infrastructure maintenance expenditures have been shortened by 60% (fig.5).
2) the amount of the lost applications has decreased fivefold, namely from five per cents to one (fig.6).

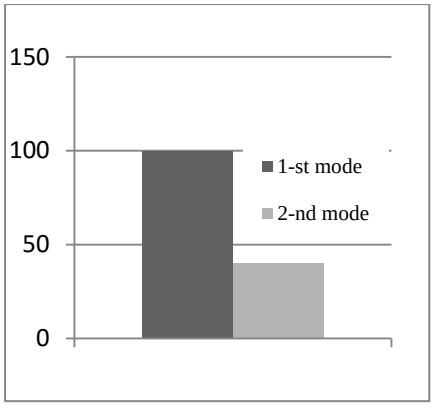

Fig. 5 The comparison of infrastructure maintenance expendintures (in percent)

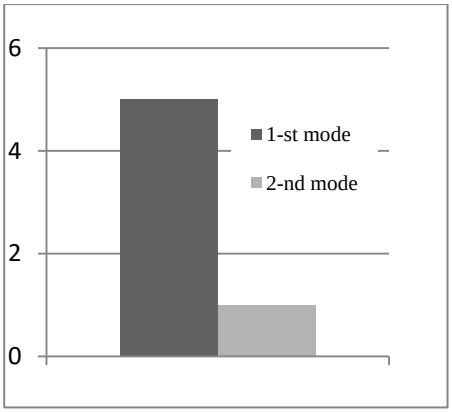

Fig. 6 The comparison of the amount of the lost billing applications (in percent)

Conclusions

In the article the challenges of workflow management in providing services have been considered for the heterogeneous environment with changing infrastructure as well as the mathematical simulation of the system's operation has been conducted based on the example of online billing system of communication provider.

According to the proposed approach of system operation control the system is considered as a network, in which flows' management is accomplished by using the balancing systems, where the amount of available units varies with the load (billing applications).

The method of time point determination when the additional resources are involved has been proposed that evaluates the dynamics of the input load as well as current state of the technical means and permits forehanded additional resources' switching on and prevents from the overloading of the existing resources.

The schedule development method for servers involvement ensures the long-term schedule of servers' switching on, based on long-term statistics data and allows planning the operation of technical means for long period of time.

The conducted simulation has shown that infrastructure maintenance expenditures have decreased by 60%, while the amount of the lost applications due to server's busyness have shortened fivefold, from five to one per cent.

References

1. Ruben Van den Bossche, Kurt Vanmechelen and Jan Broeckhove (2010) Cost-Optimal Scheduling in Hybrid IaaS Clouds for Deadline Constrained Workload. IEEE 3rd International Conference on Cloud Computing, pp. 228-235.

2. Javadi B., Abawajy J. and Buyya R. (2012) Failure-aware resource provisioning for hybrid Cloud infrastructure. Journal Parallel Distrib. Comput., No 72, pp. 1318–1331.

3. Vecchiola Ch., Calheiros R. N., Karunamoorthy D. and Buyya R. (2012) Deadline-driven provisioning of resources for scientific applications in hybrid clouds with Aneka. Future Generation Computer Systems, No 28, pp. 58–65.

4. Ardagna D. and Ciavotta M. (2014) Long-term Auto-Scaling Algorithm for Multi-Cloud IaaS Systems. Politecnico di Milano, Technical Report, No 13.

5. Kumar Das A., Adhikary T., Razzaque Md. A., Cho E. J. and Hong Ch. S. (2014) A QoS and Profit Aware Cloud Confederation Model for IaaS Service Providers. IMCOM(ICUIMC)'14. Siem Reap, Cambodia, pp. 9-11.

6. Buyya R., Ranjan R. and Calheiros R. N. (2010) InterCloud: Utility-oriented federation of Cloud computing environments for scaling of application services. 10th Int'l Conf. on Algorithms and Architectures for Parallel Processing (ICA3PP), pp 13-31.

7. den Bossche R. V., Vanmechelen K. and Broeckhove J. (2010) Cost-optimal scheduling in Hybrid IaaS Clouds for deadline constrained workloads / R. V den Bossche, // IEEE Int'l Conf. on Cloud Computing, pp. 228-235.

8. Hyunjoo Kim, Yaakoub el-Khamra, Shantenu Jha and Manish Parashar (2010) Exploring application and infrastructure adaptation on hybrid Grid-Cloud infrastructure / // Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, pp. 402-412.

9. Marcos Dias de Assunção, Alexandre di Costanzo and Rajkumar Buyya (2010) A cost-benefit analysis of using cloud computing to extend the capacity of clusters. Cluster Computing, Vol. 13., No 3, pp. 335-347.

10. Goiri I., Guitart J. and Torres J. (2010) Characterizing Cloud Federation for Enhancing Providers' Profit. 2010 IEEE 3rd International Conference on Cloud Computing. pp. 123-130.

11. Lee Y., Wang C., Taheri J., Zomaya A. and Zhou B. (2010) On the effect of using third-party Clouds for maximizing profit. Algorithms and Architectures for Parallel Processing, Springer Berlin Heidelberg, Vol. 6081, pp. 381-390.

12. Hamilton J. Cost of power in large-scale data centers. Available at: http://perspectives.mvdirona.com/.

13. Barroso L. A. and Hölzle U. (2007) The case for energy-proportional computing. Computer, vol. 40, No 12, pp. 33–37.

14. Zhernovyi K. Y. and Zhernovyi Y. V. (2012) An M θ/G/1/m system with two-threshold hysteresis strategy of service intensity switching. Journal of Communications Technology and Electronics,Vol. 57, No 12, pp. 1340-1349.

Глоба Л. С., Скулиш М. А. ***Планування завантаження ресурсів центру обробки даних на основі статистичних даних.*** *Якість обслуговування клієнтів залежить від процедури підтримки прикладних програм в центрі обробки даних постачальника зв'язку. У статті розглядається підхід контролю динамічного використання ресурсів для забезпечення обслуговування вхідного потоку, який враховує випадковий характер надходження заявок і використовує короткострокові і довгострокові статистичні навантаження. Запропонований підхід складається з двох методів, які керують кількістю обслуговуючих вузлів. Результати моделювання управління технічними ресурсами були представлені для інфраструктури ЦОД провайдера зв'язку, що доводить ефективність запропонованих методів.*

Ключові слова: *телекомунікаційна система, тарифікація послуг, якість обслуговування, гібридна хмара, хмарні обчислення, центри обробки даних, utility-комп'ютинг, віртуалізація.*

Глоба Л. С., Скулиш М. А. ***Планирование загрузки ресурсов центра обработки данных на основе статистических данных.*** *Качество обслуживания клиентов зависит от процедуры поддержки приложений в центре обработки данных поставщика*

связи. В статье рассматривается подход контроля динамического использования ресурсов для обеспечения обслуживания входящего потока, который принимает во внимание случайный характер поступления заявок и использует краткосрочные и долгосрочные статистические нагрузки. Предложенный подход состоит из двух методов, которые управляют количеством обслуживающих узлов. Результаты моделирования управления техническими ресурсами были представлены для инфраструктуры ЦОД провайдера связи, что доказывает эффективность предложенных методов.

Ключевые слова: *телекоммуникационная система, тарификация услуг, качество обслуживания, гибридное облако, облачные вычисления, центры обработки данных, utility-компютинг, виртуализация.*

*Globa Larysa, Skulysh Mariia****. Planning the loading of data centers' resources based on download statistics.*** *The customer service quality depends on the procedure of the application maintenance in data center of the communication provider. In the article the control approach of dynamic resource involvement has been suggested in order to ensure the input flow maintenance that takes into account the random nature of applications' inflow and utilizes both short-term and long-term load statistics. The proposed approach consists of two methods that manage the number of the implicated serving nodes. The first one verifies the resource amount adequacy, provides the evaluation of input load's dynamics based on the short-term statistics as well as the current state of the technical facilities. The second one accounts for the long-term statistics according to which the implication of additional resources can be scheduled during the load peaks. The simulation results of technical resources management have been presented for the data center infrastructure of the communication provider, that prove the effectiveness of the proposed methods.*

Keywords: *Telecommunication system, Tariffing of services, Quality of services, Hybrid cloud, Cloud computing, Data Centers, Utility computing, Virtualization, Market-oriented resource allocation*