UDC 621.391

DEVELOPMENT OF THE CONCEPT FOR THE COMPUTATIONAL RESOURCES MANAGEMENT SOFTWARE IN THE CUSTOMER SERVICE SYSTEMS

Mariia A. Skulysh [1], José Luis Pastrana Brincones[2], Dmytro O. Parhomenko[1]

[1]Institute of Telecommunication Systems
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
[2]Department of Computer Science and Programming Languages of the University of Malaga,
Malaga, Spain

Background. To date, there is no customer service system that does not involve information and computer systems. One of the most important issues in ensuring the reliability and reliability of such systems is the task of dynamic scaling and providing the required amount of computing resources at any time. This study was focused on the planning and deployment of computing infrastructure that is able to respond to significantly increased volumes of request flows, changes in the dynamics of load intensity, strict requirements for the quality of their service, etc.

Objective. The purpose of the paper is to create a concept of virtual computing space to meet the needs of distributed customer service system, which takes into account the peculiarities of service, the computing node load nature, service quality requirements, and provides energy efficient. Developed models, methods will control the performance of distributed computing infrastructure and flow maintenance processes, reduce downtime of computing resources and provide services to end users at a given level of quality.

Methods. Analysis of the operation of the node load assessment mechanism, which consists in a dynamic change in the intensity of control of the state of function nodes, showed the effectiveness of planning for a group of computing nodes..

Results. The proposed approach to managing a heterogeneous computation environment to improve the efficiency of the service maintenance process in new generation systems is a unified solution for highly loaded distributed systems. The developed concept made it possible to avoid a decrease in the quality of service during surges of congestion and to maintain the indicators of the quality of service at a given level, provided that the resource utilization ratio is kept within the given limits

Conclusions. In summary, a mathematical model of the problem of determining the maximum allowable load volume with a QoS level guarantee for a service node in a heterogeneous telecommunications environment was proposed.

Keywords: *information and communication network; dynamic resource allocation; ontology; model; computer system infrastructure; analysis; resources management; QoS.*

Introduction

A system that is represented by autonomous PCs connected using middleware is called a distributed system [1], [2]. It helps share resources and infrastructure to provide customers with a single, integrated system. A distributed system has some key features, such as sharing resources as well as software from other systems connected to the network, in other words, the components in the system are synchronous [3]. In the distributed model, the fault tolerance is much higher than in other network models, so the performance/price ratio is much better [4]. In a distributed system, the key goals are transparency, reliability, performance and scalability, as well as ensuring QOS.

Transparency is a model for presenting structure without hiding details from the user. The reliability of a distributed system consists in high masking of errors, security and consistency [5]. Productivity is measured by the ability to deliver the expected result. Distributed systems should be scalable in terms of topography [6]. When using a public network, fault tolerance is one of the main problems in a distributed model when it is built on unreliable basic resources [7]. Without proper protocols or policies, resource coordination and sharing is a major challenge in a distributed environment[8].

When using cloud computing, software is provided to users as an Internet service. The user has access to his own data, but cannot control and should not choose the infrastructure, operating system and software with which he works. The underlying hardware is regularly updated in huge data centers that use sophisticated virtualization techniques to provide high levels of scalability, availability, and flexibility [9]. However, the distribution of cloud resources has

become the subject of research, which has led to significant development of algorithms and methods. Vitally important concepts for continuing the execution of ready-made programs through the structure are the allocation of resources for calculations [10].

A fundamental system for taking advantage of spatio-temporal repetition in remote channels is dynamic resource allocation by handing over valuable resources such as range and power to expand or limit relevant measures of system performance[11]. In conventional static resource allocation procedures, subchannels are transmitted in a predetermined manner; that is, each client is assigned fixed frequency bands regardless of channel status.

The goal of the research was to create a concept of a virtual computing space to meet the needs of a distributed customer service system that takes into account the specifics of service, the nature of the load on computing nodes, requirements for the quality of service and ensures energy efficiency.

Implementation of the main ideas of the study.

In modern systems for processing extremely large data flows, the problem of creating an environment for their processing is being deformed. There is an evolution from traditional processing on own resources with limited technical capabilities to the use of cloud technologies. In the study, this idea was developed, appropriate models and methods were proposed, which allow calculating a sufficient number of resources, which are necessary for the maintenance of extremely large flows, taking into account the peculiarities of the structure of computing resources.

The idea of organizing a heterogeneous computing environment was embodied in the development of a unified system for managing communication and computing resources of a heterogeneous environment, which made it possible to optimize the maintenance of extremely large heterogeneous arrays of information resources in accordance with the requirements described in the ontological data model.

The hypothesis regarding the efficiency of distribution of the load for the maintenance of information resources between the company's own resources and leased cloud resources was confirmed. It was found that when planning the use of computing resources, the enterprise becomes 30% more efficient, due to the avoidance of downtime and overloading, with the timely involvement of leased resources, which allows for an even distribution of the load. Also, the proposed computer environment management system allows you to ensure the required level of protection, fault tolerance and reliability of flow maintenance in the customer service systems.

New methods have been developed

1. A method has been developed for determining the location and volume of reserved computing resources for functions that are services and serve the flow of requests. The method takes into account the state of computing and network resources, requirements for service quality, network heterogeneity, and unlike the existing ones due to the dynamic provision of resources, it allows reducing the amount of used resources.

2. A mathematical model of the problem of determining the maximum allowable volume of load with a guarantee of the QoS level for a service node in a heterogeneous telecommunication environment is proposed, which takes into account the ergodic distribution of the number of requests, the limit delay, minimizes the loss of requests in the system due to a lack of computing service resources, and allows you to calculate the upper limit of the allowable load when planning the intensity of inbound traffic for service at a node.

3. A management model of the virtualization infrastructure of electronic information resources service resources is proposed, which takes into account the peculiarities of the location of service data centers, of the flows of hybrid telecommunication services that enter data centers for service, which allows for flexible management of the information and telecommunications system organized with using cloud computing.

An analysis of communication and computing resources required for work in the customer service systems was carried out. Possibilities of attracting cloud computing resources and the management system of a hybrid information and communication environment were investigated to ensure indicators of the quality of service of the system as a whole and of end users in particular.

Main part

A set of architectural solutions will be developed to organize the infrastructure of a computer system that meets the requirements for quality of service based on an intelligent platform.

Main idea of the infrastructure organization is to perform a dynamic allocation of resources each server will need to use three components:

A monitoring module that measures the load and performance of each function (such as the intensity of

receipt of requests, the average response time, etc.);

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

2) A resource allocation module that uses these estimates of load to determine the amount of resources that should be allocated by function s. Fig. 1 shows these three components.

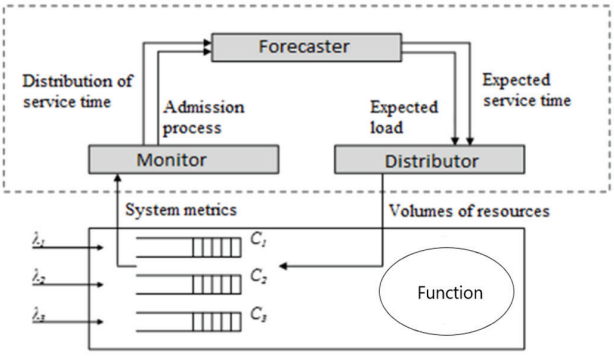


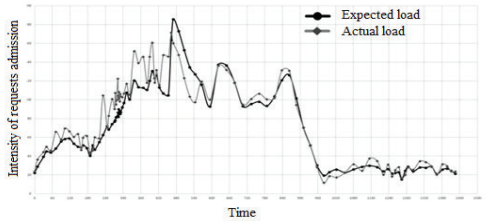
Fig. 1 Dynamic resource allocation system

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

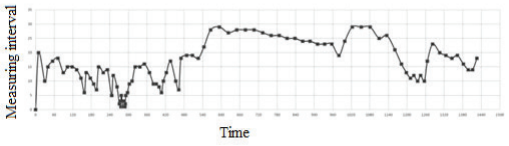
$$W(t) = I_{base} - K \cdot \sum_{j=t-h}^{t-1} \frac{\max(0; \lambda_{obs}(j) - \lambda_{pred}(j))}{h} I_{base}$$

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, $\lambda_{pred}(t)$ - 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 node; (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 function s 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:

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

As the resource is modeled as a Functions server, the intensity of the function request service is $\mu_i = C_i/s_i$, where C_i is the number of resources of the 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: (1)

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

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

$$T_i = \frac{q_i + 1}{\mu_i},$$

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 function s need to allocate the number of resources, so that $T_i \le d_i$, then the amount of resources allocated by the function C_i must satisfy the condition of equation:

$$C_i \geq s_i \frac{q_i + 1}{d_i}, \quad (5)$$

A modified load factor predictor based on the method 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 :

$$\lambda(t) = \lambda_{pred}(j) + \sum_{j=t-h}^{t-1} \frac{\lambda_{obs}(j) - \lambda_{pred}(j)}{h}.$$

3. Recommendations for its use in customer service information systems will also be developed. We propose the procedure for ensuring the quality of service provided.

The principle of dynamic quality control is as follows: the value of the delay in maintaining the application for connection (disconnect, recovery) is compared with the quality of service of the subscriber. If the metric does not match, then the metrics of quality in virtual nodes and virtual local area networks are compared with the limit values of the corresponding policies stored in the PCRF subsystem. This principle analyzes the following quantitative indicators of the effective operation of the system, such as: the time delay of the service flow request in the virtual node and the probability of loss of queries in the service node. Service node - a virtual machine that performs functions of managing the network node.

After determining the cause of the problem with the performance indicators, appropriate measures are taken. If there is a problem in the time of transmission between service nodes, then it is recommended to reconfigure the system, namely to change the location of virtual nodes in the physical nodes of the heterogeneous structure of the data center. If the problem is detected in a single service node, then it is recommended to increase the number of service resources. If there is a decline in service quality indicators in a group of related interface nodes, for example, which form the computation network, then it is recommended to limit the flow of applications sent to service the corresponding node. For this purpose it is recommended to calculate the intensity of the load on the group of nodes. Algorithm of procedure of indications on Fig. 3.

The proposed approach to managing a heterogeneous computation environment to improve the efficiency of the service maintenance process in new generation systems is a unified solution for highly loaded distributed systems. The developed concept made it possible to avoid a decrease in the quality of service during surges of congestion and to maintain the indicators of the quality of service at a given level, provided that the resource utilization ratio is kept within the given limits.

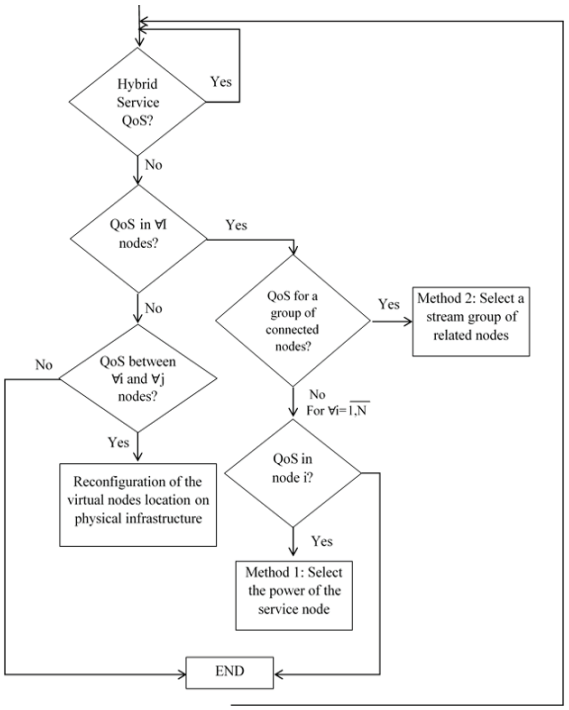


Fig. 3. Procedure for guaranteeing the preset quality of service

Most modern solutions offer a static resource allocation scheme. Where the redistribution of resources does not occur during the operation of a virtual computing network. There are a limited number of decentralized and dynamic solutions. And the approaches that offer solutions for the dynamic embedding of a virtual network allocate a fixed amount of resources for virtual nodes and channels for the entire period of existence. As the nature of the load changes over time, this can lead to inefficient use of shared computing and network resources, especially if the physical network rejects requests to embed new virtual network functions, while reserving resources for

virtual network functions that are under low load conditions.

In summary, the difference between the approach proposed in this study and the one mentioned above is that the resources reserved for use by virtual functions do not remain constant throughout the lifetime of the virtual network. Virtual nodes are monitored and based on their actual resource needs, resources are reallocated, in which case unused resources are returned to the physical network for use by other virtual networks.

Simulation

A number of simulation models were presented that implemented the models and methods used in the concept on the basis of which the estimation of the quality of service of the tasks was calculated in accordance with the proposed concept. Service quality assessment was performed on a set of statistical data that was received from the company of the communication operator. The evaluation process proceeded as follows, used modeling the set of kernel nodes of the communication operator in the GPSS system.

Input data of the simulation model:

dt = 0,1 ms – interval for time sampling;

T_{simul}=864000 intervals – total simulation time (discrete);

N_{Σ}^{i} – the number of requests that came in the system for time i-th simulation, $i = \overline{1,100}$;

$$M = \begin{cases} 28, when\ using\ additional\ resources \\ 14,\ else \end{cases};$$

$Vj = \{V_{R1j}, V_{R2j}\}$ – available resources of the node j.

For each simulation, the statistics of servicing quality indicators for each j-th node, as well as the system as a whole, were recorded:

t_j^i – average delay time at node j ($j = \overline{1,M}$), during the i-th simulation $i = \overline{1,100}$;

z_j^i – number of lost queries in node j ($j = \overline{1,M}$), during the i-th simulation $i = \overline{1,100}$;

t_{all}^i – the average delay time in the system during the i-th simulation $i = \overline{1,100}$;

z_{all}^i – the number of lost queries in the system during the i-th simulation $i = \overline{1,100}$.

Also, the monitoring system keep statistic data on the number of resources in each small interval of time:

${R_{1j}}^i = \{R_{1j}^1, \dots R_{1j}^k, \dots R_{1j}^{T_{мод}}\}$ – a set data monitoring of resource R1 in node j ($j = \overline{1,M}$), during the i-th simulation $i = \overline{1,100}$;

${R_{2j}}^i = \{R_{2j}^1, \dots R_{2j}^k, \dots R_{2j}^{T_{мод}}\}$ - – monitoring of resource R2 in node j ($j = \overline{1,M}$), during the i-th simulation $i = \overline{1,100}$.

To assess the performance of hybrid services, the calculation of the likelihood of violating the requirements of standards and specifications regarding the service time and the probability of timely service of the service was performed, the corresponding formulas for calculating probabilities were given in Table 1. The simulation results are summarized in Table 2.

In the work of the communication operator, an important indicator of the functioning of the system as a whole is the utilization rate of resources. The practice of the telecommunication company has shown that the resource utilization rate should vary from 30% to 80%. Since, if the resource utilization rate exceeds 80%, start unforeseen crashes , if the resource utilization rate is less than 30%, then not used efectively hardware and arises surplus of their maintenance costs are recorded. Therefore, in the simulation process, the probability that the system resources are used less than a given threshold value a, as well as the probability that the system resources are used more than a given threshold value b

$${R_{2j}}^i \supset {R_{2j\_a}}^i = \{R_{2j}^k | R_{2j}^k < a * V_{R2j}; k = \overline{1, T_{simul}}\}$$
$$\left|{R_{2j\_a}}^i\right| = A^i \quad (20)$$

$${R_{2j}}^i \supset {R_{2j\_b}}^i = \{R_{2j}^k | R_{2j}^k > b * V_{R2j}; k = \overline{1, T_{simul}}\}$$
$$\left|{R_{2j\_b}}^i\right| = B^i \quad (21)$$

Table 1. Quality scores and appropriate quality scores

Quality score	Threshold value	Score	Values of evaluation
t_d- delay time	P_{tj} =0.8 ms	p_{1j} $\forall j = \overline{1,M}$	$p_{1j} = 1 - (\sum_{i=1}^{100} k_{ij})/100$ $k_{ij} = \begin{cases} 1 & t_j^i > P_{tj} \\ 0 & else \end{cases}$
	P_{tall}=8 ms	p_{1all}	$p_{1all} = 1 - (\sum_{i=1}^{100} K_{iall})/100$ $K_{iall} = \begin{cases} 1 & t_{all}^i > P_{tall} \\ 0 & else \end{cases}$
P – probability of successful service	P_{zi}=0,98	p_{2j} $\forall j = \overline{1,M}$	$p_{2j} = 1 - (\sum_{i=1}^{100} K_{iz})/100$ $K_{iz} = \begin{cases} 1 & \frac{N_{\Sigma}^i - z_j^i}{N_{\Sigma}^i} > P_{zi} \\ 0 & else \end{cases}$
	P_{zall}=0,98	p_{2all}	$p_{2all} = 1 - (\sum_{i=1}^{100} K_{iall})/100$ $K_{i2all} = \begin{cases} 1 & \frac{N_{\Sigma}^i - z_{all}^i}{N_{\Sigma}^i} > P_{zall} \\ 0 & else \end{cases}$

$\boldsymbol{\alpha}$ – resource utilization rate HTE	a=0,3	$P_{3R_{2j}}$ $\forall j = \overline{1,M}$	$P_{3R_{2j}} = (\sum_{i=1}^{100} \frac{A_i}{T_{simul}})/100$
	b=0,8	$P_{4R_{2j}}$ $\forall j = \overline{1,M}$	$P_{4R_{2j}} = (\sum_{i=1}^{100} \frac{B_i}{simul})/100$

Table 2 shows that such indicators of quality of service as the average delay in servicing the service at the node and in the system as a whole, the probability of successful servicing of the service in the node and the system as a whole were kept within the limits of acceptable values and acquired a slight improvement. However, the utilization rate of resources for servicing according to the proposed models and methods is kept within the specified limits with an average probability of 32%.

Table 2. Simulation results

	Standard service	Management based on a proposed method	Standard service	Management based on a proposed method
	Average delay in servicing the service at the node ($\overline{t_j} = \sum_{i=1}^{100} t_j{}^i/100$)		Assessment of timely service in the node ($p_1 = (\sum_{j=1}^{M} p_{1j})/M$)	
Max_j	7	9	0,8	0,805
Min_j	1	1		
Average	3,8	4,4		
	Average service servicing delay over the system ($\overline{t_{all}} = \sum_i t_{all}{}^i/100$)		Assessment of timely service ($p_{2=}p_{1all}$)	
Max	80	90	0,82	0,84
Min	20	23		
Average	55	62		
	Probability of successful servicing in the node ($\overline{z_j} = \sum_i \frac{N_\Sigma^i - z_j^i}{N_\Sigma^i}/100$)		Estimation of the probability of successful service in the node ($p_3 = (\sum_{j=1}^{M} p_{2j})/M$)	
Max_j	1	1	0,95	0.99
Min_j	0,96	0.99		
Average	0,98	0.999		
	Probability of successful service in the system ($\overline{z_{all}} = \sum_i z_{all}{}^i/100$)		Estimation of the probability of successful service in the system ($p_4 = p_{2all}$)	
Max_j	1	1	0,94	0.99
Min_j	0,95	0.99		
Average	0,975	0.999		
	Resource utilization rate HTE $\boldsymbol{\alpha}$		Probability of using the resource R1 and R2 with a coefficient of utilization of computing resources less than a given threshold value ($\overline{P_{3R_{1j}}}$ i $\overline{P_{3R_{2j}}}$)	
$(Max_j\overline{R_{1j}})/V_{R2j}$	0.95	0.9	0.4 i 0,35	0,15 i 0,25
$(Min_j\overline{R_{1j}})/V_{R2j}$	0.15	0.25.	Probability of using the resource R1 i R2 with the coefficient of computing resources more than a given threshold value ($\overline{P_{4R_{1j}}}$, $\overline{P_{4R_{2j}}}$)	

$(\mathrm{Max}_j\overline{R_{2j}})/V_{R2j}$	1	0.85	0,2 i 0,15	0,05 i 0,1
$(\mathrm{Min}_j\overline{R_{2j}})/V_{R2j}$	0.1	0.2	Estimation of the coefficient of utilization of computing resources used within the norm $p_5 = 1 - \frac{\sum_{g=1}^{2} w_g p_{3Rgj}}{2} - \frac{\sum_{g=1}^{2} w_g p_{4Rgj}}{2}$	
Average	0.4	0.2	0.43	0.75

Conclusion

In this paper, a method was presented for determining the place and amount of reserved computing resources for functions that are services and serve the flow of requests. The method takes into account the state of computing and network resources, quality of service requirements, network heterogeneity, and unlike the existing ones due to the dynamic provision of resources, it allows reducing the amount of resources used. A mathematical model of the problem of determining the maximum allowable load volume with a QoS level guarantee for a service node in a heterogeneous telecommunications environment was proposed. A model for managing the infrastructure of virtualization of resources for servicing electronic information resources is also proposed, taking into account the peculiarities of the location of service data processing centers, as well as the of the flows of hybrid telecommunications services entering data centers. The basic idea behind organizing the infrastructure is to dynamically allocate the resources that each server must use to make use of the three components.

The HTE management technology is proposed, where the maintenance of hybrid telecommunication services is carried out using software in many cloud data centers, the technology made it possible to avoid a decrease in the quality of service during surges of congestion and to maintain the indicators of the quality of service at a given level, provided that the resource utilization ratio is kept within the given limits.

In the work of a communication operator, an important indicator of the functioning of the system as a whole is the resource utilization ratio. The practice of the telecommunications company showed that the resource utilization ratio should range from 30% to 80%. As a result of the work, it can be seen that such indicators of service quality as the average service delay in the node and in the system as a whole, the probability of successful service in the node and in the system as a whole were kept within acceptable values and slightly improved. However, the ratio of resource utilization during maintenance according to the proposed models and methods is kept within the given limits with an average probability of 32% higher.

References

1. Z. N. Rashid, S. R. Zeebaree, and A. Shengul, “Design and Analysis of Proposed Remote Controlling Distributed Parallel Computing System Over the Cloud,” pp. 118–123.
2. L. M. Haji, R. R. Zebari, S. R. M. Zeebaree, M. A. WAFAA, H. M. Shukur, and O. Alzakholi, “GPUs Impact on Parallel Shared Memory Systems Performance – International Journal of Psychosocial Rehabilitation,” May 22, 2020. https://www.psychosocial.com/article/PR280814/24791/ (accessed May 22, 2020).
3. Z. Xiao, W. Song, and Q. Chen, “Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment,” IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1107–1117, Jun. 2013, doi: 10.1109/TPDS.2012.283.
4. N. Harki, A. Ahmed, and L. Haji, “CPU Scheduling Techniques: A Review on Novel Approaches Strategy and Performance Assessment,” J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 48–55, 2020.
5. Z. N. Rashid, S. R. Zebari, K. H. Sharif, and K. Jacksi, “Distributed Cloud Computing and Distributed Parallel Computing: A Review,” pp. 167–172.

6. Q. Zhang, Q. Zhu, and R. Boutaba, "Dynamic Resource Allocation for Spot Markets in Cloud Computing Environments," 2011, pp. 178–185.
7. S. R. Zeebaree, K. Jacksi, and R. R. Zebari, "Impact analysis of SYN flood DDoS attack on HAProxy and NLB cluster-based web servers," Indones. J. Electr. Eng. Comput. Sci., vol. 19, no. 1, pp. 510–517, 202
8. J. W. J. Xue, A. P. Chester, L. He, and S. a. Jarvis, "Dynamic Resource Allocation in Enterprise Systems," 2008 14th IEEE Int. Conf. Parallel Distrib. Syst., pp. 203–212, 2008.
9. S. R. Zeebaree, R. R. Zebari, and K. Jacksi, "Performance analysis of IIS10. 0 and Apache2 Cluster-based Web Servers under SYN DDoS Attack," 2020.
10. R. R. Zebari, S. R. Zeebaree, K. Jacksi, and H. M. Shukur, "E-Business Requirements For Flexibility And Implementation Enterprise System: A Review."
11. O. Alzakholi, L. Haji, H. Shukur, R. Zebari, S. Abas, and M. Sadeeq, "Comparison Among Cloud Technologies and Cloud Performance," J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 40–47, Apr. 2020, doi: 10.38094/jastt

Скулиш М.А., Хосе Луїс Пастрана Брінконес, Пархоменко Д.О.

Розробка концепції програмного забезпечення управління обчислювальними ресурсами в системах обслуговування клієнтів

Проблематика. На сьогоднішній день не існує жодної системи обслуговування клієнтів, яка б не включала інформаційно-обчислювальні системи. Одним із найважливіших питань забезпечення надійності таких систем є завдання динамічного масштабування та забезпечення необхідного обсягу обчислювальних ресурсів у будь-який момент часу. Дане дослідження було зосереджено на плануванні та розгортанні обчислювальної інфраструктури, здатної реагувати на значно збільшені обсяги потоків запитів, зміни динаміки інтенсивності навантаження, жорсткі вимоги до якості їх обслуговування тощо.

Мета досліджень. Створення концепції віртуального обчислювального простору для задоволення потреб розподіленої системи обслуговування споживачів, яка враховує особливості обслуговування, характер навантаження на обчислювальні вузли, вимоги до якості обслуговування та забезпечує енергоефективність. Розроблені моделі, методи дозволять контролювати продуктивність розподіленої обчислювальної інфраструктури та процеси обслуговування потоків, скорочувати час простою обчислювальних ресурсів і надавати послуги кінцевим користувачам на заданому рівні якості.

Методика реалізації. Аналіз роботи механізму оцінки навантаження вузла, який полягає в динамічній зміні інтенсивності контролю стану функціональних вузлів, показав ефективність планування для групи обчислювальних вузлів.

Результати досліджень. Запропонований підхід до управління гетерогенним обчислювальним середовищем для підвищення ефективності процесу обслуговування сервісів у системах нового покоління є уніфікованим рішенням для високонавантажених розподілених систем. Розроблена концепція дозволила уникнути зниження якості обслуговування під час сплесків заторів та підтримувати показники якості обслуговування на заданому рівні за умови збереження коефіцієнта використання ресурсів у заданих межах.

Висновки. У підсумку запропоновано математичну модель задачі визначення максимально допустимого обсягу навантаження із забезпеченням рівня QoS для вузла обслуговування в неоднорідному телекомунікаційному середовищі.

Ключові слова: *інформаційно-комунікаційна мережа; динамічний розподіл ресурсів; онтологія; модель; інфраструктура комп'ютерної системи; аналіз; управління ресурсами; QoS.*