Embedded Self Organizing Systems (Vol 9. No 3. 2022) (pp.39-41)

Embedded Selforganizing Systems

Special Issue Topic: "International Symposium on Computer Science and Educational Technology "

Development of an Overload Management Model in IMS

Nurmatova Sevara
EMU University
Uzbekistan, Tashkent
sevara-mmm@inbox.ru

***Abstract*— The concept of IP Multimedia Subsystem (IMS) describes a new network architecture, the main element of which is a packet transport network that supports all access technologies and provides implementation of a large number of info communication services. In this paper, formalization of the task of balancing the load of IMS virtual subsystems.**

***Index Terms*—quality of service, a network of mass service, imitative model, probability of timely delivery of packets, coefficient of a variation of a time delay of packets.**

I. INTRODUCTION

Currently, the second stage of the IMS-based communication networks concept is being implemented in Uzbekistan, and operators, providers and other market participants in the industry have ample opportunities to use the advantages of IMS. It is shown that today the main trends in the global telecommunications market are reduced to building networks based on IMS, on the platform of which almost all modern products and services are created, including for the telecommunications market of Uzbekistan.

In accordance with recommendations and technical specifications, SIP (Session Initiation Protocol) [1] was chosen as the main signaling protocol of IMS technology. This choice is explained by such principles of the protocol as extensibility, which consists in the possibility of adding new functions to the protocol by adding new headers and messages, which allows you to add new functionality to the network without changing the protocol; independence from the underlying transport layer; scalability; the ability to interact with other signaling protocols (ISUP, BICC).

II. TASK SETTING

In the Republic of Uzbekistan, the state standard of the Republic of Uzbekistan "Switching and signaling systems, the protocol for initializing communication sessions SIP. Basic requirements" [2]. This standard establishes the basic requirements for the structure and format of the SIP protocol, connection management procedures, as well as procedures for ensuring security on the telecommunications networks of Uzbekistan when using the SIP protocol, and is intended for suppliers (manufacturers) of equipment using the SIP protocol and telecommunications operators operating the corresponding equipment.

The requirements of this standard apply to next generation networks and are mandatory when using equipment using the SIP protocol on the territory of the Republic of Uzbekistan.[3]

The main elements of the IMS subsystem are (Figure 1.1):

- CSCF (Call Session Control Function), which contains the main call processing logic, and consists of 3 functional blocks:
- P-CSCF (Proxy CSCF), which acts as an intermediary when interacting with subscriber terminals (UE - User Equipment);
- I-CSCF (Interrogating CSCF), which acts as an intermediary for interaction with external networks;
- S-CSCF (Serving CSCF), which is the central functional block of the IMS network, it processes all SIP messages transmitted by end devices;
- HSS (Home Subscriber Server), which is a database with subscriber profiles.

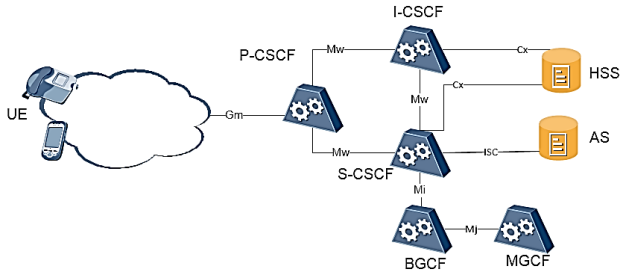

Fig.1.1. Simplified diagram of the IMS subsystem

To implement IMS services, AS (Application Server) application servers are used, which are connected to the IMS core through a standardized ISC interface.

The performance of IMS servers is determined by the number of calls that the server can handle per unit of time. When the volume of calls exceeds the capacity of the server, it goes into overload mode, the waiting time for service on the server increases. The client waits for a response to the request and, without waiting, sends a repeated request to establish a connection. This leads to an even greater increase in the input load on the server and exacerbates the overload [4].

Congestion in the IMS network can be caused by various reasons. This may be the result of user activity trying to establish connections around the same time. Another common cause of overload is the failure of one of the elements in a cluster of SIP servers, which reduces overall performance and requires distribution of the incoming load among the remaining servers in the cluster.

The problem of SIP server overloads arises not only from user behavior during peak hours, but also from the provision of certain multimedia services that significantly change the nature of the signaling traffic. For example, a presence service involves sending notification messages simultaneously to a large number of users [5]. Another congestion cause is when end terminals try to register at the same time after a network failure.

The SIP protocol has flaws in the congestion control mechanism, according to which, in the event of an overload of the proxy server, a 503 Service Unavailable message is sent. In particular, the following issues have not been resolved:

- the problem of aggravation of overload, which consists in the tendency of a significant increase in load during the period of overload;
- the problem of incomplete use of the server cluster.

In conditions of overload of SIP servers, it is necessary to apply load control mechanisms: threshold and priority control.

Distinguish between local, inter-node and end-to-end load management. One of the simplest tools to prevent congestion in SIP servers is a mechanism for local control of incoming traffic based on queue length thresholds [5].

Work on new mechanisms for managing congestion was entrusted to the working group SOC (SIP Overload Control) of the IETF committee [6]. The group's current work focuses on two inter-node congestion control schemes:

- scheme with resetting messages on the sender side (LBOC-Loss-Based Overload Control);
- scheme with limiting the rate of the flow of signaling messages (RBOC-Rate-Based Overload Control).

The main idea of the LBOC scheme is that the sender, at the request of the recipient, reduces the number of messages sent by a percentage of the total number of messages specified by the recipient. The principle of operation of the RBOC scheme is based on indicating to the sender the maximum number of messages that the recipient would like to receive from the sender during the specified time interval. The LBOC scheme is intended to replace the existing basic congestion control mechanism and should eventually be integrated into the existing version of the SIP protocol.

Cloud-based or virtualized IMS architecture (vIMS) is gaining popularity these days. In "cloud" SIP servers, it becomes possible to manage congestion not only by LBOC and RBOC mechanisms, but also by changing the number of virtual machines. Based on this, the work develops a model for managing congestion in "cloud" SIP servers based on limiting the intensity of the incoming flow of requests (LBOC) and (or) increasing the number of virtual machines serving applications.

III. Methodological recommendations for assessing the probabilistic and temporal characteristics of the entity of the SIP server object

The process of servicing a connection request (INVITE) in the IMS subsystem is multi-stage, involving several IMS servers. Each server in the multi-stage process of servicing the request for the upstream server is a client, i.e. the i-1 server is a client of the i-th server, and the i-th server is a client of the i+1 server.

Let's consider the functioning of the i-th server as a managed CFR. The controlled parameters are the intensity of receipt of requests (requests, messages) λ and the number of enabled virtual machines v serving requests with intensity μ. The number of virtual machines v can vary from 1 to the maximum value V depending on the number of requests in the system n(t) at time t>0, $0 \leqslant n(t) \leqslant N$, N=L+V, where L is the maximum queue length.

IV. Calculation of probabilistic and temporal characteristics of the virtual SIP server under study

At the initial moment, the server contains one virtual machine, the remaining virtual machines can be turned on when the number of requests in the system reaches the threshold value $n(t) \leq M_1$. The distance between thresholds is

$$m = \left\lfloor \frac{N}{V} \right\rfloor, \quad (1)$$

where $\lfloor\ \rfloor$ is the sign of the operation of rounding a number down.

If the number of requests in the system $n(t) \leq M_1$, then the control parameters do not change

$$\lambda_1 = \lambda\,,\ \ \mu_1 = \mu \quad (2)$$

If n(t) increases and passes the threshold M_1, then the intensity of service increases by increasing the number of virtual machines by one, provided that the i+1th server is not overloaded

$$\mu_2 = 2\mu(1-P) + P\mu, \quad (3)$$

If the i+1-server is overloaded, then the intensity of requests from the i-1-th server decreases

$$\lambda_2 = P\lambda(1-q_1) + (1-P)\lambda, \quad (4)$$

where q_1 is the proportion of the decrease in the intensity of the incoming flow, when

$M_1 < n(t) \leq M_2, 0 < q_1 \leq 1.$

Thus, when $n(t) > M_i (i = \overline{2,V})$ the intensities of incoming and servicing requests change according to the formulas

$$\mu_i = i\mu(1-P) + P\mu, \quad (5)$$
$$\lambda_i = P\lambda(1-q_i) + (1-P)\lambda, \quad (6)$$

where $q_1 < q_2 < q_3 \ldots . < q_v$.

The reliability and average delay time of data frames largely depend on the characteristics of the communication channel. [6]

Knowing the stationary state probabilities, we can estimate the probabilistic-temporal characteristics of the studied virtual SIP server, the formulas for calculating which are given below.

Average number of requests in the SIP server:

$$\bar{N} = \sum_{k=0}^{N} kP_k \quad (7)$$

Probability of losing requests (system blocking) [7]:

$$P_{loss} = P_N \quad (8)$$

Average delay (stay) time of requests in the server:

$$\bar{T} = \frac{\bar{N}}{\lambda(1-P_{loss})} \quad (9)$$

The calculation of the probabilistic-temporal characteristics of the virtual server was carried out with the following initial data:

$N = 50, m = 10, V = 5, \mu = 1\ мс^{-1}, P = 0.5,\ q_1 = 0.2,\ q_2 = 0.4,\quad q_3 = 0.6,\quad q_4$=0.8.

The problems of effective functioning of the IMS-based network involve various aspects of the construction and implementation of technologies that reflect the various capabilities of the IMS platform. The presence of multivariance in the processing of information and signal flows, the heterogeneity of communication systems and technologies, quality of service requirements and a number of other factors determine the need to manage a large number of devices and process a variety of information flows with a given quality of service for each type of flow, and therefore, solving the problems of ensuring efficient use of network resources becomes much more complicated.

CONCLUSION

In order to study the issues of efficient use of the resources of the reference network for the transmission and processing of information flows of various types, this article formulates an optimization problem related to the distribution of flows in the IMS network, the complexity of which lies in the presence of many scenarios for the interaction of functional modules in the network architecture and heterogeneity of traffic.

The tasks of effective functioning of IMS networks are formulated as optimization tasks related to the search for the extremum of the cost functional in the presence of certain restrictions (for example, on probabilistic-temporal characteristics).

Calculations show that the advantage of the proposed method of congestion control is observed in areas of high load ρ>0.8. In general, the proposed method reduces the average delay time of requests from 1.2 times to 1.7 times in areas of high load compared to the LBOC (Loss-based Overload Control) mechanism [8].

REFERENCES

[1] Goldstein B.S. Handbook of Telecommunication Protocols: SIP Protocol / B.S. Goldstein, A.A. Zarubin, V.V. Samorezov.- St. Petersburg: BHV - Petersburg, 2007

[2] O‘zDSt 2873:2014 Telecommunication networks. SIP session initialization protocol. Primary requirements".

[3] Nurmatova S.B., Amirsaidov U.B. Assessment of the quality of services in next generation networks. "Problems of information technology and telecommunications" Republican scientific and technical conference of young scientists, applicants, undergraduates and students, collection of articles 3-part, Tashkent, March 14-15, 2013, 71-73 pages.

[4] E.S. Sopin. Models of servers of the IMS subsystem with group receipt of requests. XII All-Russian Conference on Management Problems, Moscow, 2014, pp. 8735-8742.

[5] Samuylov K.E., Zaripova Z.R. Model of the local mechanism for SIP server congestion control. T-Comm, No. 7, 2012, pp. 185-187.

[6] Nurmatova S.B., Amirsaidov U.B. Method for calculating the probability of packet loss in a QS with self-similar traffic Journal “Vestnik TUIT”, Tashkent 2014, No. 4 (32) pp. 29-36

[7] Samuilov K.E., Abaev P.O., Gaidamaka Yu.V. and others. Analytical and simulation models for assessing the performance of SIP servers under congestion T!Comm No. 8, 2014, pp. 83-88.

[8] Nurmatova S.B. Development of a model for calculating the probability of timely delivery of packages in IMS ICISCT 2020