Managing of incoming stream applications in online charging system

Globa Larysa Prof, Dr., Skulysh Mariia, PhD,
Podgurskaj Tatiana
National Technical University of Ukraine "Kiev Polytechnic Institute" Kiev, Ukraine
mb_s@ukr.net, lgloba@its.kpi.ua

Andrii Reverchuk
NVision Czech Republic, Prague, CZ

Abstract— **Qualitative online billing system performance is essential for cost-effective service providers. This paper presents the method of technical resources' distribution of the billing server based on the resource requirements of different types of service that takes into account the daily load statistics of the different types of services and increases the revenue by providing an individual approach to services with different costs. The method of incoming billing requests' flow management is proposed according to the requirements for technical resources. The developed method can reduce the loss of tariff requirements caused by the exceedence of maintenance time. The results of incoming billing requests' flow management simulation have been obtained that prove the effectiveness of the proposed method.**

Keywords—3G mobile communication, multimedia communication, packet switching, quality of service, core function, online charging system, operators billing, packet switched networks, rating module

I. INTRODUCTION

The 3rd Generation Partnership Project (3GPP) Release 8 determines the specifications with low latency, high data speed of all-IP core network (CN) enabling the real-time packet service support for several access technologies, including novel Long-Term Evolution (LTE) access network. The industry of mobile communication is being developing. In order to stay competitive many mobile operators are being focused on the decrease in both capital and operating expenditures (CAPEX and OPEX) as well as concentrate on the outsourcing of several activities and/or by exchange of some network infrastructures.

The objective of this article is to provide an overview of potential evolution of 3GPP in terms of Policy and Charging Control (PCC) as well as quality of service (QoS) of architecture in order to ensure more effective support of Fixed-mobile convergence (FMC) and more flexible interchange of CN decisions. For this purpose more definite business role distribution (for example, access and network providers) is proposed as well as the Network Policy function's (NPF) introduction in order to control CN.

II. DESCRIPTION OF THE RESEARCH

The existing trend of service provisioning on IP-protocol gives rise to the wireless straightforward high bitrate Internet access for 3G/4G mobile customers, where mobile terminals serve as internet-connected IP-hosts [1, 10, 11].

Simplifying of the mobile terminals-to-Internet interfaces provides the steadily increasing range of services for the subscribers.

Services fees' evaluation require progressively more computation to determine their value, that consequently leads to the load increase on a tariff system, namely:

- low efficiency of call service;
- deterioration of both flexibility and efficiency of tariff approaches;
- inability of QoS guaranteeing;

The created 3G Partnership Project (3GPP) and ETSI TISPAN were intented to solve above mentioned problems. Within the framework of the project the Working Parties have been organized for developing IP Multimedia Subsystem (IMS) and dealing with:

- support of real-time personal multimedia information exchange between different terminals, i.e.: voice, video-conference, games etc.
- calls' differentiation in terms of services and Subscriber Data;
- service and application support in a single session connection;
- deterioration of both flexibility and efficiency of tariff approaches;
- flexibility of tariff approaches' enhancement;
- service monitoring on a network, session, application layers.

A significant amount of studies and approaches are dedicated to the subscriber service quality surveillance. Methods of QoS provision during the data transmission are

being developed at different stages of maintenance (i.e. network access, organization of transport information flow) [2, 3, 13]. Some part of publications are concerned with billing efficiency maintenance and analyze the development of PCC policy that defines the QoS policy for each subscriber (depending on the kind of service and tariff plan) [1, 4-6, 12, 13]. In publications [7, 8, 9] the optimization of billing systems are examined in terms of discipline of service's differentiation of the input flow as well as new rules of subscribers' billing establishment. However, the network operators face rather acute problem of overtime support request notably on the billing server. This occurs due to the unefficient principle of resource allocation of technical means that is especially critical in terms of QoS during the maximum load periods.

The solution may lie in the gradual increase in the capacity of computational resources of servers, which is quite costly, thus, unacceptable on the practice.The other possible approach is to prevent the occurrence of peak loads by managing incoming requests for billing.

The exceeding of the permissible duration of service results in call rejection and consequently leads to the economic losses as well as deterioration of the company's reputation.

Hence, there is an urgent scientific and technical problem focused on the improvement of requests' input flow control that would take into account the following: the need for technical resources for billing system and system load's evaluation as well as the corresponding mechanisms, methods, models, algorythms, protocols, interfaces and tools for call processing and billing control, which would overcome the above-mentioned disadvantages.

A. The approach of call processing on a cellular's company server

The main Network Host built according to the 3GPP standard is PCC (Policy and Charging Control) that provides the monitoring of QoS and billing. The monitoring is ensured by introducing the PCC rules' according to which the PCC solutions are made. The decision node in 3GPP networks is PCRF (Policy Control and Charging Rules Function). It is the functional element that makes decisions based on the service policy of subscribers by enabling / disabling the use of services as well as setting the QoS parameters. It also establishes the tariff regulations according to various conditions. The scheme of the PCRF in a communication network LTE, built on standards 3GPP, is shown in Fig. 1.

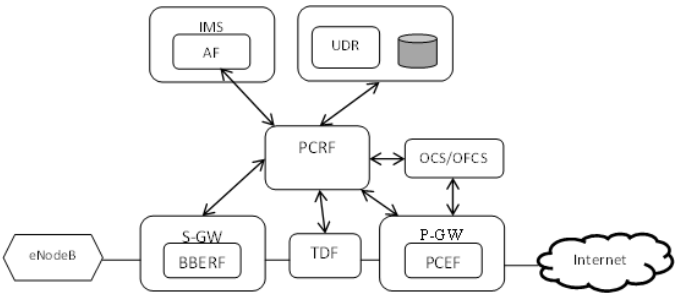

Fig. 1. Relation between the PCRF-server and 3GPP network elements

In order to provide call handling the PCRF interacts with various subsystems that ensure the call service. PCRF's functionality depends on the availability and functionality of the above mentioned nodes. The functional elements and interfaces, interacting with PCRF include the following:

- PCEF (Policy and Charging Enforcement Function) –
- applies the PCC rules from the PCRF to the flowing traffic as well as ensures the traffic billing in the OCS / OFCS.
- BBERF (Bearer Binding and Event Reporting Function) – notifies the PCRF of session setup and sends both an identifier and optional parameters of the subscriber in order to determine the corresponding QoS rules.
- TDF (Traffic Detection Function) –
- determines the traffic and notifies PCRF. Depending on the regulations the traffic can be transmitted to the subscriber, forwarded or limited in speed.
- UDR (User Data Repository) – stores the user data.
- AF (Application Function) – describes the service data flow and provides information about the required resources.
- OCS (Online Charging System) – real time credit control server. Provides the following functions: the service rating, the control of payment balance, the processing of information about the funds flow on subscriber's account, discounts and calculation of the consumed services' amount.

B. Online Charging System operation

The customer service on the billing server is an important part of the telecommunication services' provision. Since each service is paid, depending on the type of service plan the network operator provides billing of the request either real-time before the delivery of service or offline. In both cases the availability of funds on the subscriber's account should be checked as well as the conversion and the range of standard operations should be performed.

Since operation execution time is limited, the exceeding of it leads to the loss of the request. Eventually, the user is notified of the inability to obtain the service. Consequently, the operator incurs losses that spoil the reputation in the case of systemic failures. Therefore, by effective choosing of server resources, the possibility of service time exceeding of user's requests' can be reduced.

The complexity of the amount of resources' calculation required for different types of services' charging can be explained by four major prerequisites:

- Charging process involves sequential operations that require different amounts of resources.
- Each type of service, despite the uniformity of operations performed during the billing operations, requires different amount of resources.

- A large number of billing requests come simultaneously for various types of resources.
- Intensity of requests' arrival depends on day of the week and time of the day.

Scheme of requests maintenance on the billing server is shown in Fig. 2.

The modules EDP and CPA decode the flows coming according to the protocols, namely, CAP and Diameter.

The decoded sequence is sent to the business logic server SEE (Service Execution Environment) through the routing module BUS. The business logic server SEE is the core of the billing system and provides an environment to perform a sequence of operations involved in requests maintenance.

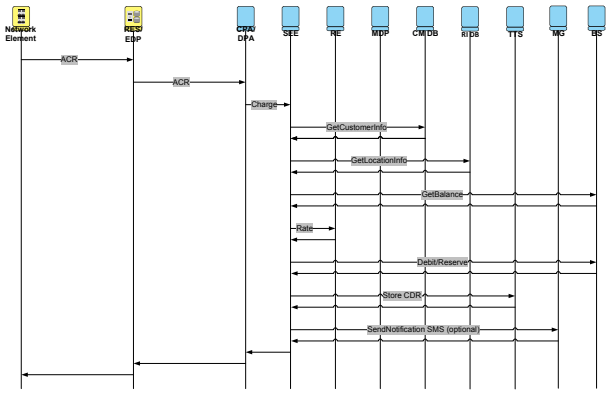

Fig. 2. The overall scheme of online tariffication on billing server

The sequence of operations for successful requests maintenance includes the following:

- Retrieving information about the subscriber. For this module SEE addresses the Customer Management data base (CM DB).
- Retrieving information about the location of the Customer. For this module SEE refers to the Resource Inventory data base (RE DB) that stores the network structure.
- Retrieving information about the status of the subscriber's account. For this module SEE refers to the subscriber's database of Balance Storage (BS).
- Calculation of the service fee based on tariff book and current service plan for the subscriber. For this module SEE addresses the Rating Engine module (RE).
- The fee is charged from subscriber's account. Depending on the type of service the payment reservation can be simultaneously performed only for such services, as: Session Charging with Unit Reservation (SCUR) and Event Charging with Unit Reservation (ECUR), in case of undefined duration of the service as well as unknown final amount of payment. For debit transactions and funds reservation module SEE refers to the subscriber's BS database.
- In order to report about the provided services, the call data record is made (SDR). For this purpose the SEE module refers to the Toll Ticket Server (TTS).
- If required, the subscriber can be sent the message on the outcome of the provided services. For this purpose the SEE module addresses the message GateWay (MG) module, by sending a message to the subscriber.

Apparently, the billing process is multistage. Thereby, the sequentially executed operations in the business logic core of SEE involve various subsystems and generally are diverse. Thus, they require different amount of Random Access Memory (RAM), CPU time and disk space. Let the stages of billing server maintenance name as functional blocks. Among the most significant parameters, that have impact on QoS is the rate of requests maintenance, that depends on the amount of processing technical means. Therefore, it is necessary to focus on the way the resources are used. Both the total number of resources that maintain the server as well as methods of resources' separation should be precisely considered. On the one hand, the last ones ensure the efficient service provision at each processing stage. However, the resources' separation is also a constrain since the subsystem uses only the allocated resources and doesn't have access to the other ones.

C. Management level

The operation of billing server subsystem can be represented as a multi-level queuing system, where the requests' flow control is accomplished at two levels.

The first logic level is an application level. Here, the submitted to the system requests differ in the type of service they represent. Additionally, the queue maintenance involves sequential operations with SEE module that are listed above. The management process involves queues' incoming requests management, namely, queues forming depending on the type of service, the usage of WRAD group's methods etc.. Thus, the first level is the queuing system that processes various types of arriving requests, let call them first-level requests. The serving devices in this system are the circuits of functional units, where each type of service is sequentially served in a separate circuit.

The second level is the level of technical computing. The scheme of requests maintenance by the type of the service involves the sequential operations that do require a given amount of hardware resources. Each operation can be represented as a request for service. Thus, at the second level the handling devices are the hardware resources. While the second level requests are organized in queues to appropriate resources. The policies of resources utilization are determined by the methods of computing system's resource management. The architecture of resource allocation as well as organization of second-level requests maintenance, significantly affect the speed of service. However, such system architecture of second -level requests' processing is constant. Its functionality can be assessed by statistical data delays while maintaining the first-level requests.

The incoming flow of second -level requests is uniquely determined by the number of first-level requests, served by the system. Therefore, the control system of the first-level incoming requests, which is built on the basis of statistics on second-level resources' load factor will allow reducing the requests' loss in terms of delays associated with lack of resources. According to the calculation scheme of resources' division in periods of high load the resources are allocated based on economical effectivity and the number of different types of arrived requests.

The question is in what way the incoming requests management should be organized so that the second-level requests' flow was the most uniform.

By providing maintenance of the first-level requests in functional blocks the second-level requests' flow is generated, that uses the given amount of server resources. So, let some functional block simultaneously processes a large number of the first-level requests. At the same time, the second-level requests, produced by the corresponding functional unit, require a significant amount of resources. Thus, the lack of server resources problem may arise, leading to the delays in first-level requests servicing, and, as a result, permissible time exceeding of service, loss of requests, reduce in QoS.

D. Technical means' distribution for requests' maintenance for various types of services' tariffication

The task of resources' allocation between different types of services is generally based on economic effectiveness and involves the optimal portion of resources' determination that are required to provide requests' maintenance for various types of services. By solving this problem the economic effectiveness of requests maintenance will be maximized as well as the statistics of workloads for different types of services will be determined.

The input data have been listed below.

- The volume of resources of each type to ensure the maintenance of various requests in a functional unit.
- Profit from every type of request.
- Maximal allowable amount of distributed resources for billing server.

The distribution of resources should be determined for the particular requests' maintenance.

Suppose it is necessary to allocate the two types of resources, namely: Random Access Memory (RAM) and Read-only memory (ROM) on the disk.

Let k_i is the search value of i-type requests in the billing system;

v_{ij}^{Rg} – amount of g-th resource, required for single i-type request maintenance in j-th functional unit;

S_i – profit from every type of i-type request.

Thus, the amount of g-th resource, required for i-type request maintenance in all functional units constitutes $\sum_i v_{ij}^{Rg}$.

The total amount of income received from all i-type requests maintenance makes up:

$$S=\sum_i k_i S_i$$

The total number of resources occupied by all i-type requests, which are simultaneously located in the system and distributed between all functional blocks equal:

$$v^{Rg}=\sum_i k_i \left(\sum_i v_{ij}^{Rg}\right). \qquad (1)$$

Since it is necessary to maximize profit, the objective function is as follows:

$$\sum_i k_i S_i \rightarrow max. \qquad (2)$$

The amount of utilized resources can't exceed the available one. The additional restrictions include the values of the average number of requests' ratio for various types of services, based on daily statistics' analysis:

$$\begin{cases} \sum_i k_i \left(\sum_i v_{ij}^{Rg}\right) \le V_{Rg},\ g=1,..G & (3) \\ a_{ij} \le k_i/k_j \le b_{ij},\ \forall\, i,j=1,..m & (4) \end{cases}$$

where V_{Rg} – the amount of g-th resource, provided by billing server;

k_i and k_j – the search values of tariffication's requests, obtained from i-th and j-th types of service, respectively;

a_{ij} and b_{ij} – numerical limits calculated on the basis of statistical data on the hourly number of billing requests coming to the system;

m – number of services for which the billing is accomplished.

By solving the given optimization problem on constrained extremum, the obtained result is a sequence $\{k_i\}$ $(i=1,..m)$, i.e. the number of requests of each type of service that can be simultaneously processed in the billing system. Then according to the formula (1) we can calculate the required amount of resources allocated to service.

As a result, the total amount of technical resources distribution is defined for billing requests of different types of services during the call maintenance in OCS, which gives a maximal profit.

E. Method of incoming billing requests' flow management

The main feature of this method is to control the permissible load of processing requests by delaying the rest ones. It allows avoiding overloading of resources and prevention of inefficient resource loading by the requests' processing during more than assigned time.

The input data of the requests' flow control problem, which are submitted to the mobile operator's server, are:

The information about the amount of resources required for the operations' provision that are assigned by functional block corresponding to the request maintenance of specified service.

The information about the duration of information resources' use while processing the request of specified service in every functional block.

Statistical information on the duration of requests maintenance of specified service in every functional block.

The resources capacity allocated to the specified service's maintenance.

The server settings that are characterized as system resources for requests maintenance are usually calculated for the average values of the parameters of the input flow. However, the system also receives the peak amount of simultaneous submitted requests'.

Under the peak load of the input stream we consider the exceeding amount of simultaneous arriving requests comparing to the calculated above acceptable value (see Section 5).

In order to provide the efficient requests' processing as well as to prevent the resource deficiency in management system the following strategy can be proposed:

- Two and more bursts of input flow's load shouldn't be simultaneously processed in the functional units that require a significant amount of resources;
- For this purpose a delay of requests is introduced, the arrival of which is coincided with the load-peak. The requests don't come in the system until the last burst is successfully processed in functional unit that determines the delay time.

The functional blocks that require the greatest amount of resources are assumed to be resource-consuming. The maximal number of requests for given type of service is calculated on the basis of the sequence $\{k_i\}$, derived from the solution of the problem (2) and according to the formula (5):

$$k_{i\,max}=a*max_g(k_i v_{iw}^{Rg}/\sum_j v_{ij}^{Rg})+b, \qquad (5)$$

where w – the index number of resource-consuming functional unit,

a, b =const numerical limits, obtained from simulation.

By providing the information about the amount of resources required to process operations for specified service maintenance, the resource-consuming functional units should be defined.

The algorithm of the management method is shown in Figure 3.

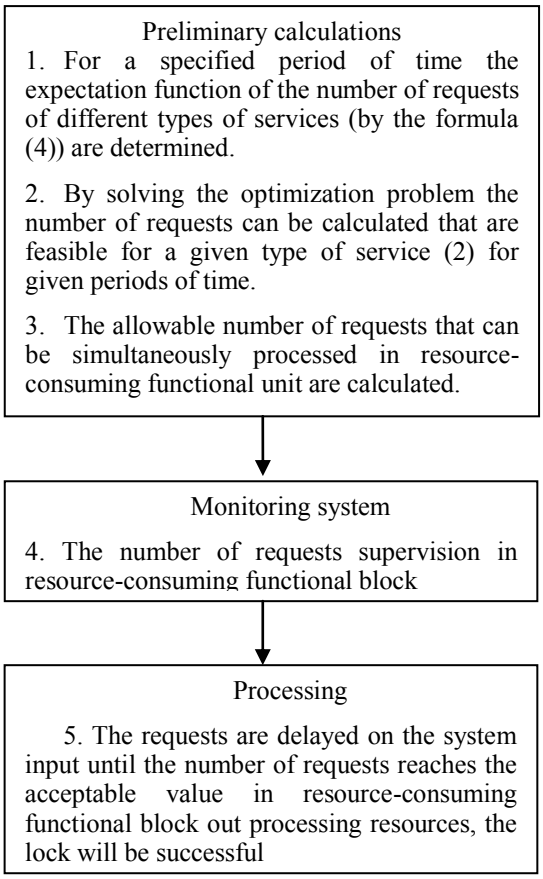

Fig. 3. The algorithm of the method of incoming requests' flow management in order to prevent deficiency of server resources

The number of acceptable requests for a given type of service is calculated by redistribution of technical means between requests of different types of services described above, taking into consideration the maintenance effectiveness for all types of services with the availability of system resources.

III. SIMULATION RESULTS

The simulation of the method of incoming billing requests' flow management has been conducted using GPSS package.

During the simulation process the model with two resources and the services' flow has been studied. The resources taken into account during the simulation include RAM and Permanent storage.

The processing of an application includes four functional units that imitated such operations as: retrieving information about the subscriber from data base, calculation of service cost, the formation of the subscriber notification as well as the conduction of final calculations and deduct costs.

For the simulation the amount of allocated resources has been intended for simultaneous processing of 50 thousand of billing requests with uniform distribution between the functional units. The request maintenance was followed by corresponding resources' blocking until the served request moved to the next functional unit.

For this purpose a delay of requests has been introduced, the arrival of which is coincided with the load-peak. The requests don't enter the system until the last burst is successfully processed in functional unit. The queuing time corresponds to the delay time. If the arrival of request is coincided with the lack of processing resources, it is delayed until the release of the required amount of resources. At each stage the residence time of request in the system is computed and compared with allowed servicing time. The chosen values correspond to the real systems' parameters.

The input stream was modeled by the Poisson law. Based on the real systems' analysis, the most resources are spent during the notification formation for the subscriber. Therefore, in the model the number of requests has been monitored served at the current moment in the third functional unit as well as the messages have been delayed until the number of requests would not be less than the maximal number, calculated by the formula (5).

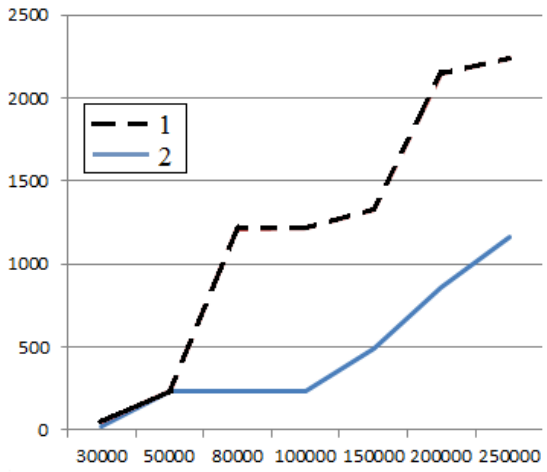

Fig. 4. The amount of lost requests

The dependence of lost requests on the arrival rate is shown in Figure 4. The proposed method can significantly decrease the amount of lost requests (line 2 in fig.4) by deploying time delays compared to the case without managing the input stream (line 1).

IV. CONCLUSION

The paper presents the method of billing server's resources' allocation. It accounts for the required resources for different types of services' requests maintenance as well as daily load statistics of different types of services and provides the significant enhancement of economic efficiency of the service. The improved method of incoming billing requests' flow management has been proposed that considers technical resources' requirements at each servicing stage as well as reduces the amount of lost requests by introducing delays that did not significantly affect the overall service time of billing requests, but at the same time limits the amount of arriving requests that require a lot of resources. By providing the simulation in GPSS package the amount of unprocessed billing requests due to excess of servicing time has decreased approximately three-fold by applying the method of incoming billing requests' flow management.

The future work is to improve the billing system by ensuring the efficient operation of the operator through the use of leased resources in order to increase the billing systems' capacity during peak loads.

REFERENCES

[1] T. V. K. Chaitanya and E. G. Larsson, "Improving 3GPP-LTE uplink control signaling performance using complex-field coding," IEEE Transactions on Vehicular Technology, vol. 62, no. 1, pp. 161–171, 2013.

[2] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," IEEE Wireless Communications, vol. 17, no. 3, pp. 10–22, 2010.

[3] X.Li, W.Bigos, D. Dulas, Y. Chen, U. Toseef, C. Goerg, A. Timm-Giel, A. Klug, "Dimensioning of the LTE Access Network for the Transport Network Delay QoS", Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd , Page(s): 1 - 7

[4] J. Costa-Requena, "SDN integration in LTE mobile backhaul networks" Information Networking (ICOIN), 2014 International Conference 10-12 Feb. 2014, p 264 – 269

[5] Wei-Ching Ho ; Li-Ping Tung ; Tain-Sao Chang ; Kai-Ten Feng "Enhanced component carrier selection and power allocation in LTE-advanced downlink systems" Wireless Communications and Networking Conference (WCNC), 2013 IEEE Page(s): 574 – 579

[6] Nuaymi, L. ; Sato, I. ; Bouabdallah, A. "Improving Radio Resource Usage with Suitable Policy and Charging Control in LTE" Next Generation Mobile Applications, Services and Technologies (NGMAST), 2012 6th International Conference Page(s): 158 - 163

[7] Ouellette, S. ; Marchand, L. ; Pierre, Samuel. "A potential evolution of the policy and charging control/QoS architecture for the 3GPP IETF-based evolved packet core" Communications Magazine, IEEE, 2011 , Page(s): 231 - 239

[8] Sok-Ian Sou ; Jeu-Yih Jeng ; Yinman Lee, "Signaling overhead of Policy and online Charging Control for bearer sessions in LTE network" Consumer Electronics, 2009. ISCE '09. IEEE 13th International Symposium, 2009 , Page(s): 593 – 597

[9] Francesco Malandrino, Claudio Casetti, Carla-Fabiana Chiasserini, "LTE offloading: When 3GPP policies are just enough" , Wireless On-demand Network Systems and Services (WONS), 2014 11th Annual Conference, Page(s): 1 – 8

[10] Antonio Cuevas, Jose Ignacio Moreno, Hans Einsiedler. IMS Service Platform: A Solution for Next-Generation Network Operators to Be More than Bit Pipes.// IEEE Communication Magazine .2006. Aug. P.75-81.

[11] Syed A. Ahson, Mohammed Ilyas. IP Multimedia subsystem (IMS) handbook, CRC Press, 2009. P. 250

[12] Globa L., Dyadenko A., Reverchuk A. The charging problems in mobile service deployment.// EUROCON 2009 The IEE Region 8 Conference devoted to 150 Anniversary of Alexander Popov, May 19-23, 2009. Saint Petersburg, Russia .

[13] Larisa Globa, Mariia Slukysh. Nodal routing with traffic classification// Polish association for knowlage management Series: Studies&Proceedings №42, 2011, pp 37-46