Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2014; 2(5A):685-692 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

Flow-Based Model for Joint Routing and Request Routing in Hybrid Content Delivery Networks

Yevsyeyeva O.*, Khader Mohammed B.
Telecommunication Systems Department, Kharkiv National University of Radioelectronics, 14, Lenina str., Kharkiv, 61164, Ukraine

***Corresponding author**
Yevsyeyeva O
Email: evseeva.o.yu@gmail.com

Abstract: Growing number of users, increasing popularity of large-scale multimedia content and high requirements for quality of delivery the content lead to integration of Content Delivery Networks (CDN) and Peer-to-Peer (P2P) networks in whole hybrid (HCDN) architecture. Hybrid content delivery networks allow to combine low cost of services with high their quality, but at same time they toughen up requirements to management practices, including request routing. In order to maximize the productivity of content delivery system in whole, to provide guaranteed quality of service and to use resources in effective way a management of requests in HCDN should be based on a rigorous mathematical justification. In this paper we proposed a flow mathematical model that implies (a) downloading from several CDN-servers at same time that allows to use computing and network resources in a balanced manner and to improve the result quality of service, (b) solving of request routing on HCDN and traffic routing in transport telecommunication network jointly. A distinctive feature of the proposed approach is taking into account state of the underlying transport infrastructure to select CDN-servers. Numerical results show that proposed model provides multipath routing without differential delay and allow reducing downloading delay.
Keywords: content delivery network; hybrid content delivery network; flow model; quality of service; request routing; load balancing

INTRODUCTION

The info communication networks, including the Internet, are aimed at timely and effective content delivery. In order to deliver large-scale content of different types to growing number of users two architectures can be implemented. There are Content Delivery Networks (CDN) and Peer-to-Peer (P2P) networks. Whereas CDN concept is based on several replica (surrogate) servers which help origin server to serve numerous user's requests but within P2P-network every peer can work as server well as client. The CDN is able to ensure high quality, security, stability of servicing but has high cost of deployment and maintenance. On the other hand P2P has high scalability and low cost with instability of quality. As a result integrating of P2P and CDN within hybrid (HCDN) architecture promises to combine their advantages and achieve maximal effectiveness of content distribution with guaranteed quality of service and relatively low cost [1, 2].

In the hybrid network, user can download content from the CDN server, from P2P hosts or to use both types of source. Logical relationships between CDN and P2P components within hybrid content delivery network can be CDN-aided or P2P-aided [1]. In the paper we'll focus on CDN-aided P2P-network. In such type of HCDN P2P-network plays main role. P2P-network is complemented by CDN-servers and new request should be directed to CDN if the requested content can't be granted by P2P-network [1]. It means that servers of P2P-network must work as index-server and as gateway to CDN at same time. And from viewpoint of CDN the servers play the role of sources of aggregated requests.

Inevitably hybrid network architecture leads to increasing complexity of its control, which among other things includes the problem of request routing. In HCDN control decision should be based on the following factors:

Hierarchical control

HCDN has a well-defined hierarchical architecture, and therefore request management should realize the principles of hierarchical control.

Quality of Service

In order to provide guaranteed Quality of Service (QoS) all control decisions in HCDN, including

the request routing and content delivering, should be based on the required type and quality of service (QoS-demands).

Telecommunication infrastructure

Content delivery network is an overlay network. It includes two components, system of content storage which controls storage capacity servers, accessing, caching, billing, and so on, and underlying telecommunications network (or Internet) which provide transport functions for content delivery from CDN-server to client. Since each of the components has material effect on the resulting QoS, the state of each of them must be taken into account when control decisions are making. In other words, management of computing resources of CDN and P2P servers should be coordinated with the allocation link and buffer resources of the telecommunications network.

Load balancing

In order to achieve effective use of resources load between CDN and P2P-networks must be balanced, as well as the use of resources within each of them should be balanced to. The matter is to balance resources of CDN-servers and link's resources in telecommunications network at same time.

As analysis shows often the methods of request routing implemented in practice and proposed in the literature do not meet the requirements in full [1-3]. Conventionally problem of request routing in CDN is problem of closest replica to which the request will be redirected. In this formulation the problem is solved separately from the problems of traffic routing in transport telecommunications network. In addition, for solving the request routing problem a heuristic approaches dominate, where the selecting of the "best" replica is a result of passive or active monitoring of the content server's state.

As a result we have an actual task to develop a mathematical models and methods for request routing in a hybrid content delivery network which must take into account the state of the telecommunication component and be aimed at content delivery with guaranteed QoS and balanced use of server and link resources.

MATHEMATICAL MODEL OF HYBRID CONTENT DELIVERY NETWORK WITH GUARANTEED QUALITY OF SERVICE

In order to describe structural features of HCDN we'll use graph $G = \{N, E\}$ in which set of vertices N represents servers of P2P-networks (N^{P2P}), replica servers of CDN (N^{CDN}) and routers of underlying telecommunications network (N^{TTN}), i.e. $N = N^{P2P} \bigcup N^{TTN} \bigcup N^{CDN}$. Set of edges E in graph G represents link between routers of telecommunications network and between routers and servers. In accordance with the architecture of HCDN sources of content and the respective traffic flows are replica servers of CDN. But servers of P2P-networks are sources of aggregated requests and receivers for content from CDN replica servers. Thus source s_k^l and sink t_k^l in graph G are restricted by sets N^{CDN} and N^{P2P} respectively, i.e. $s_k^l \in N^{CDN}$, $t_k^l \in N^{P2P}$, where the index k is index of request within the set K, $k \in K$, l is requested type of content (requested object), $l \in L$, L is set of available in given HCDN objects.

In accordance to previously formulated requirements for control decision in HCDN request routing problem includes two sub problems which should be solved jointly:

(a) to choose appropriate set of servers as sources of requested content and to balance load between them and

(b) to define paths (routes) to deliver the content through telecommunications network from chosen servers to P2P-server initiated the request.

In order to solve the subproblems let us define two types of variables [4]. Variable y_g^{kl} describes selecting g th server, $g \in N^{CDN}$,as source for l th object within k th request,

$$y_g^{kl} = \begin{cases} 1, \text{ if request } k \text{ for object } l \text{ is directed to replica server } g, \\ 0, \text{ otherwise.} \end{cases}$$

----------(1)

In HCDN are quest to CDN servers is initiated by servers of P2P-networks, so it isn't single client request but aggregated request that units some number single ones. There appears to be sufficient reason for downloading from multiple CDN servers simultaneously. Then y_g^{kl} becomes part of l th object that will be downloaded from g th CDN server, i.e.

$$0 \le y_g^{kl} \le 1. \quad \text{---------- (2)}$$

Necessary condition for $y_g^{kl} > 0$ is availability of l th object on server g (replica or origin). In order to indicate the availability let us introduce coefficients z_g^l,

$$z_g^l = \begin{cases} 1, \text{ if object } l \text{ is available on server } g, \\ 0, \text{ otherwise.} \end{cases}$$

----------(3)

To solve problem of routing in telecommunication network let us define routing variables x_{ij}^{kl} which contain portion of traffic from sources (replica servers) s_k^l to destination t_k^l transmitted along link $(i, j), (i, j) \in E$. It's known that multipath routing allows to improve effectiveness of transport telecommunication network. The multipath routing for content delivery can be realized in the following way

$$0 \le x_{ij}^{kl} \le 1. \text{ ----------(4)}$$

By using described variables conservation law for routers of transport telecommunication network and P2P-servers can be written as [4]

$$\sum_{j \in N^{P2P} \cup N^{TTN}} x_{ij}^{kl} - \sum_{j \in N^{TTN}} x_{ji}^{kl} - \sum_{g \in N^{CDN}} z_g^l x_{gi}^{kl} = \begin{cases} 0, \text{ если } i \ne s_k^l, t_k^l, \\ -1, \text{ если } i = t_k^l. \end{cases}$$

$i \in N^{P2P} \bigcup N^{TTN}$.---------- (5)

The conservation law for CDN servers has form

$$\sum_{j \in N^{P2P} \bigcup N^{TTN}} z_g^l x_{gj}^{kl} = y_g^{kl}, \; g \in N^{CDN}. \text{ -----(6)}$$

Under using multiple servers simultaneously the integrity (wholeness) of downloaded content can be ensured by next expression

$$\sum_{g \in N^{CDN}} z_g^l y_g^{kl} = 1 \text{----(4)}$$

Limited available network and computing resource require conditions

$$\sum_{k \in K} \sum_{l \in L} r^{kl} x_{ij}^{kl} \le c_{ij}, \text{---------- (7)}$$

and $$\sum_{k \in K} z_g^l y_g^{kl} \le S_g^l, \text{ ----------(8)}$$

where r^{kl} is total rate at which k^{th} destination (P2P server t_k^l) downloads l^{th}object; c_{ij} is capacity of the link $(i, j) \in E$; S_g^l is productivity of g^{th} server measured as maximal number of l^{th} type sessions, which g^{th} server is capable to serve.

One from previously formulated requirements for control in HCDN is related to quality of services. In order to guarantee quality of content delivery the mathematical model must be appended by additional constraints. Use of constraints developed within tensor approach [5] allows taking into account flow nature of network traffic, nonlinear depending result quality of servicing on intensity of traffic, multipath fashion of transmitting and downloading from several sources at the same time. The constraints to ensure downloading with acceptable (required) rate and delay have form (because we'll have same constraints for every pair of indexes l and k, in order to simplify notations the indexes will be omitted)

$$\Lambda_\eta^{\langle \vartheta \rangle} \le \left(G_{\pi\eta}^{\langle 4,1 \rangle} - G_{\pi\eta}^{\langle 4,2 \rangle} \left[G_{\pi\eta}^{\langle 4,4 \rangle} \right]^{-1} G_{\pi\eta}^{\langle 4,3 \rangle} \right) T_\eta^{\langle \vartheta \rangle},$$

----------(9)

$$\sum_{w=1}^{\vartheta} \lambda_{(\eta)}^w = r \ge r^{\langle req \rangle}, \text{ ---------- (10)}$$

where $\Lambda_\eta^{\langle \vartheta \rangle}$ is vector with elements $\lambda_{(\eta)}^w = y_g r$, w is index of node pair between P2P server t_k^l and g^{th} server, $w = \overline{1, \vartheta}$; $\lambda_{(\eta)}^w$ is packet intensity of traffic flow downloaded from g^{th} server (within w^{th} node pair); ϑ is number of servers used as sources simultaneously; $T_\eta^{\langle \vartheta \rangle}$ is $\vartheta \times 1$ vector with same elements $\tau_{\langle req \rangle}$; $r^{\langle req \rangle}$ and $\tau_{\langle req \rangle}$ are numerical values of rate and delay, respectively, required for acceptable quality of playback of requested content; $G_{\pi\eta}^{\langle 4,1 \rangle}$ is the first element of the matrix $G_{\pi\eta}^{\langle 4 \rangle}$,

$$\begin{Vmatrix} G_{\pi\eta}^{\langle 4,1 \rangle} & | & G_{\pi\eta}^{\langle 4,2 \rangle} \\ --- & + & --- \\ G_{\pi\eta}^{\langle 4,3 \rangle} & | & G_{\pi\eta}^{\langle 4,4 \rangle} \end{Vmatrix} = G_{\pi\eta}^{\langle 4 \rangle}; \quad G_{\pi\eta}^{\langle 4 \rangle} \text{ is square } \phi \times \phi \text{ submatrix of matrix } \begin{Vmatrix} G_{\pi\eta}^{\langle 1 \rangle} & | & G_{\pi\eta}^{\langle 2 \rangle} \\ --- & + & --- \\ G_{\pi\eta}^{\langle 3 \rangle} & | & G_{\pi\eta}^{\langle 4 \rangle} \end{Vmatrix} = G_{\pi\eta}; \phi = m - 1,$$

m is the number of nodes in the network; $G_{\pi\eta}$ is $n \times n$ matrix calculated according to $G_{\pi\eta} = A^t G_v A$; n is the number of links in the telecommunication network; A and C are $n \times n$ matrices of co- and contravariant transformation of coordinates (they connect set of basic

circuits and node pairs in structure of telecommunication network with set of links in the structure); $G_v = \left\| g_v^{ij} \right\|$ is diagonal $n \times n$ matrix where i^{th}, $i = \overline{1,n}$, element connects rate of traffic through the i^{th}link with delay along the link. If assume queuing model $M/M/1/N$ as model of given link, then i th element of G_v is calculated according to

$$g_v^{ij} = \frac{\rho_i^v - (\rho_i^v)^{N+2} - (N_i^v + 1)(\rho_i^v)^{N+1}(1-\rho_i^v)}{(1-(\rho_i^v)^{N+1})(1-\rho_i^v)\lambda_i^v} \quad (11)$$

where $\rho_i^v = \dfrac{\lambda_i^v}{c_i^v}$, λ_i^v is packet intensity of traffic transmitted through the i^{th}link, c_i^v is capacity of the i^{th}link (number of packets per second).

Figure 1 shows example of network where, $n = 12$, $m = 8$, $\phi = 7$. In the network three servers are used as sources simultaneously, i.e. $\vartheta = 3$. Every w^{th}node pair has own quality parameters, downloading rate $\lambda_{(\eta)}^w$ within it and delay $\tau_w^{(\eta)}$.

Eqs. (9) and (10) guarantee that under sufficient amount of available resources result delay $\max\left\{\tau_w^{(\eta)}\right\}$ will be less than $\tau_{\langle req \rangle}$ and total download rate $\sum \lambda_{(\eta)}^w$ will be more than $r^{\langle req \rangle}$. Here unknown variables are λ_i^v which are related to routing variables as $\lambda_i^v = rx_{mj}$ where indexes i and (m, j) define same link in the telecommunication network.

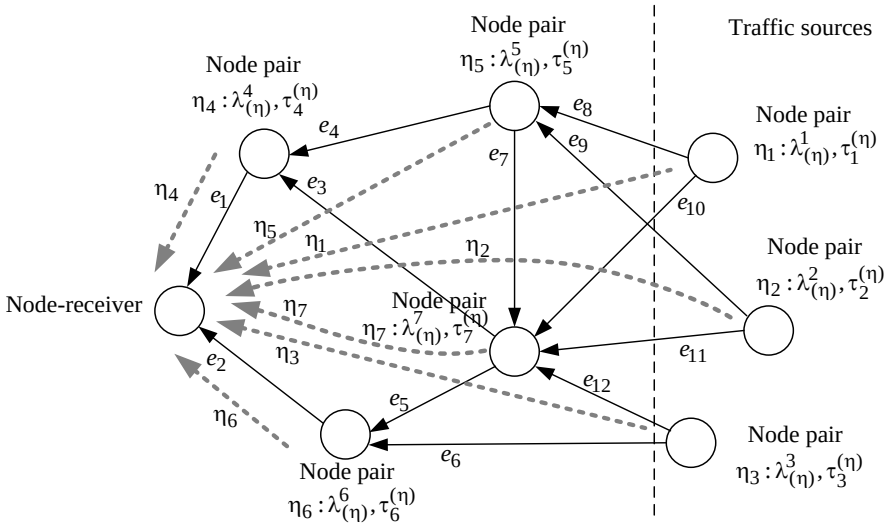


Fig. 1: Example of set of node pairs

In order to balance resources of CDN-servers and link's resources in telecommunication network let us define [6]

$$\sum_{k \in K} \sum_{l \in L} x_{ij}^{kl} \le \alpha \le 1, \text{ --------- (12)}$$

$$\frac{\sum_{k \in K} y_g^{kl}}{S_g^l} \le \beta \le 1 \text{ ---------- (13)}$$

where α is controlled threshold of use of link resources in transport telecommunication network, $0 \le \alpha \le 1$; β is controlled threshold of use of server resources in CDN, $0 \le \beta \le 1$.

Then request routing problem in HCDN can be formulated as optimization problem

$$Minimize\, Q_\alpha \alpha + Q_\beta \beta \quad (14)$$

Subject to Eqs. (2) – (13)

where $\vec{x}$, $\vec{y}$ are vectors of unknown variables x_{ij}^{kl} and y_g^{kl} respectively; Q_α, Q_β are weight matrices.

Thus formulated optimization problem (14) provides

- Optimal selecting of CDN-servers for content downloading;
- Multipath routing content from several selected servers to P2P-server initiated request:
- Guaranteed quality of service (downloading rate and delay);
- Balanced use of link resources in transport telecommunication network and server resources of CDN.

RESULTS AND DISCUSSION

The novelty of the proposed model (2) - (14) is related to tensor constraints (9) – (10) which are implemented for solving the request routing problem for the first time. Their satisfactions mean providing two QoS-requirements, the required downloading rate and delay. Moreover the tensor constraints ensure a very important gain. It's known multipath routing suffers from grave shortcomings that are related to difference between delays along different paths (so called differential delay). But conditions (9) - (10) allow eliminating the drawback. If the conditions (9) - (10) are satisfied, the delays along different paths should besame. In other words, the proposed model (2) - (14) provides such load balancing between CDN-servers and paths in transport telecommunication network that differential delay becomes zero.

In order to demonstrate the gain let us discuss network shown in Fig. 2 that contains two P2P-networks, and correspondingly two P2P-servers which aggregate requests to CDN within own peering network. Assume all requests to CDN-servers will be initiated by the first P2P-server and same content is requested. But intensity of requests will be varied. The structure of given network allows the use of two CDN-servers (simultaneously or separately), and three paths for content delivery from the servers to node 1 which representsserver of first P2P-network. Capacities of network's links are shown in Fig.3. Total network capacity is bounded by link's capacities and equal to 2000 Mbps. A simulation results that were obtained in accordance with (2) – (14) are shown in Fig.3.

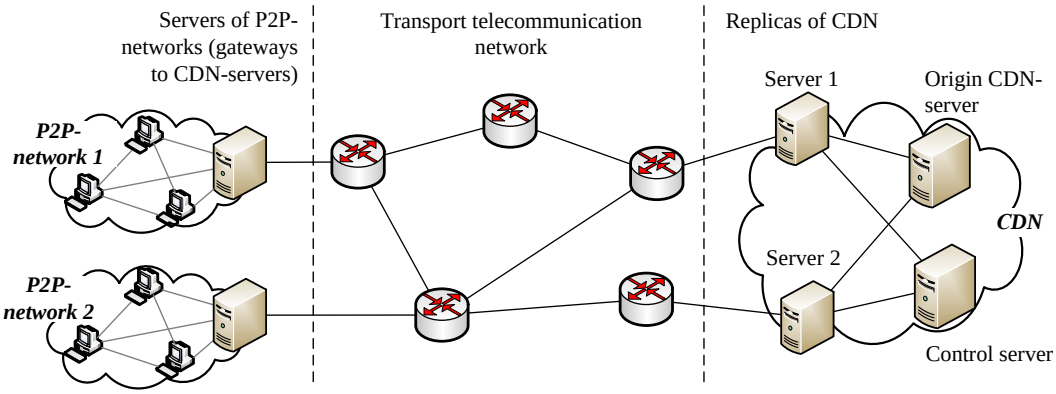
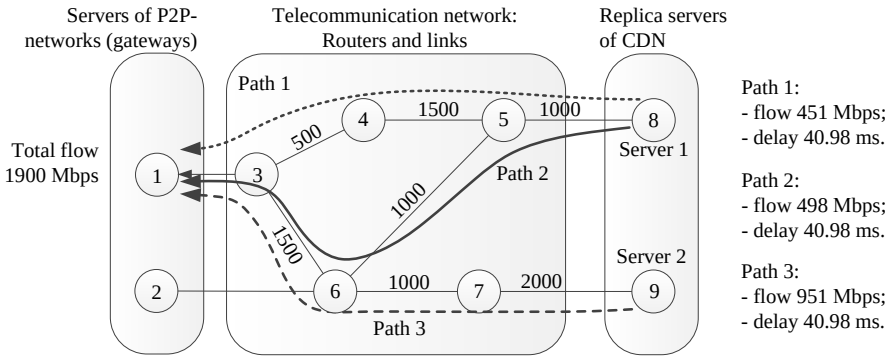


Fig. 2: Example of HCDN

As numerical results show under high request's intensity two CDN-servers of given network (nodes 8 and 9) are used simultaneously and load is divided between them approximately in equal proportion (Fig. 3a). Under medium load CDN-server 2 (node 9) is preferred because path from 9 to 1 is shortest and it gives less delay (Fig. 3b). Under low load resources of one server and one path are enough (Fig. 3c). But every time when two or more paths are used for content delivery delays along each of them are the same. Here delay was estimated according to queuing model M/M/1.



a

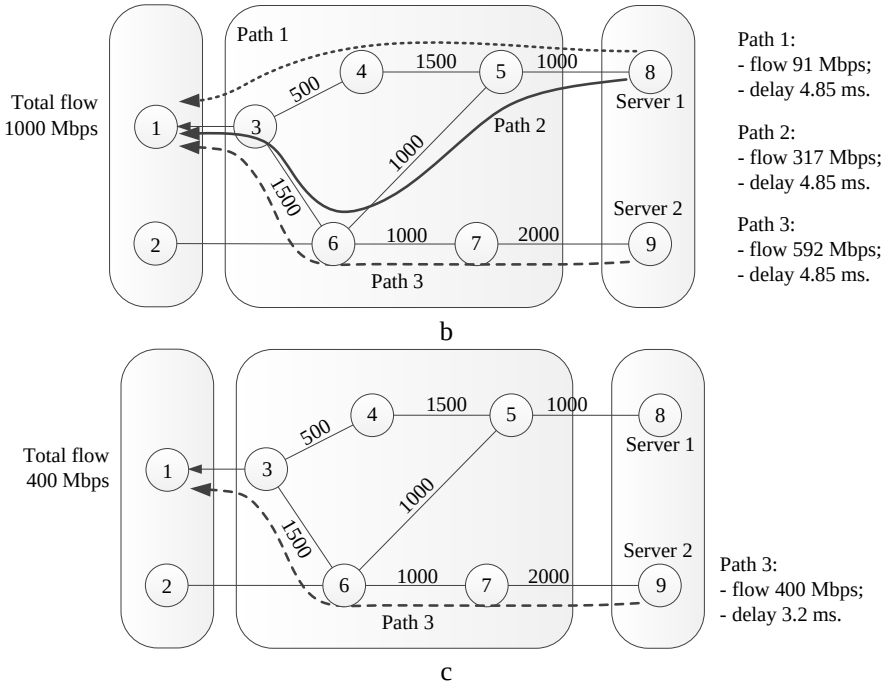


Fig. 3: Routing and request routing solution under high (a), medium (b) and low (c) request's intensity

Fig. 4 and 5 show how server's resources are allocated under different intensity of requests. Here to estimate quality of load balancing expression $\left(\sum_k \sum_l y_g^{kl}\right) / \left(\sum_g \sum_k \sum_l y_g^{kl}\right)$ was implemented.

For comparison, Fig. 6 shows the results of the model without tensor conditions (9) - (10). In this case we have even load allocation between servers. On the one hand, it provides stable operation of servers. But it rises to the differences in delays along different paths. For a given network, these delays are shown in Fig. 7. Moreover conditions (9) - (10) allow to reduce result delay as it shown in Fig. 8.

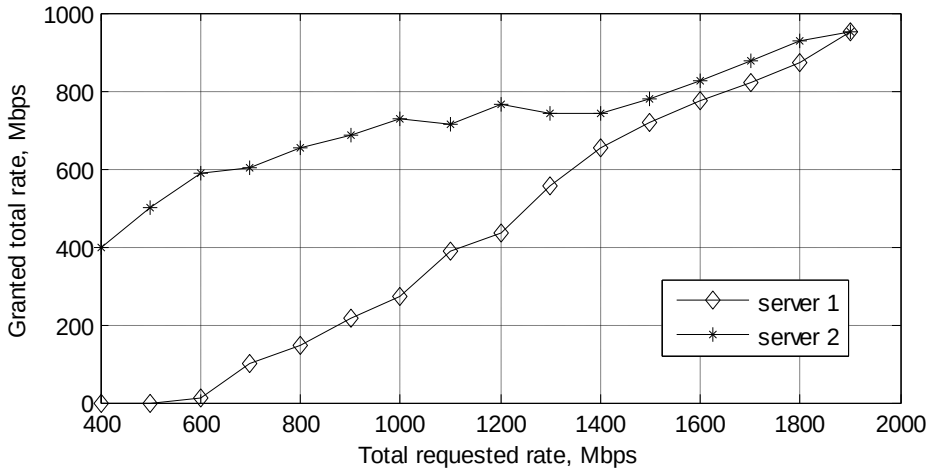


Fig. 4: CDN-server's load under different request's intensity

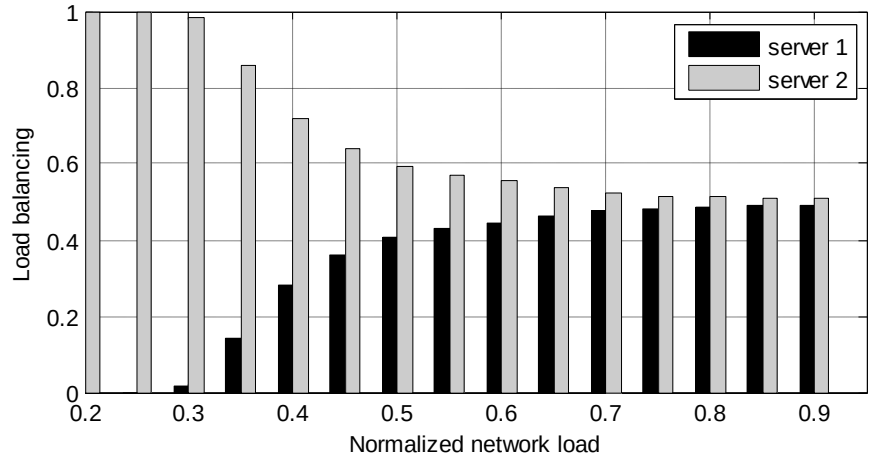


Fig. 5: Load balancing between servers according to proposed model

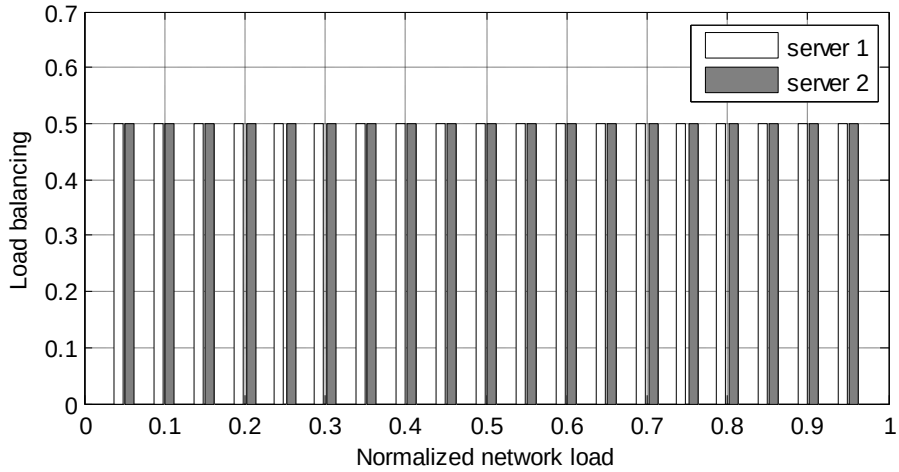


Fig. 6: Load balancing between servers without conditions (9) - (10)

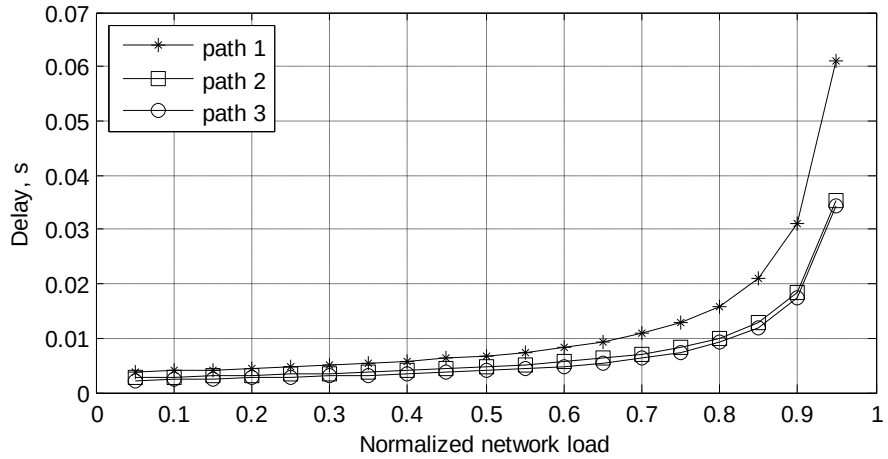


Fig. 7: Delay along different paths without tensor conditions (9) - (10)

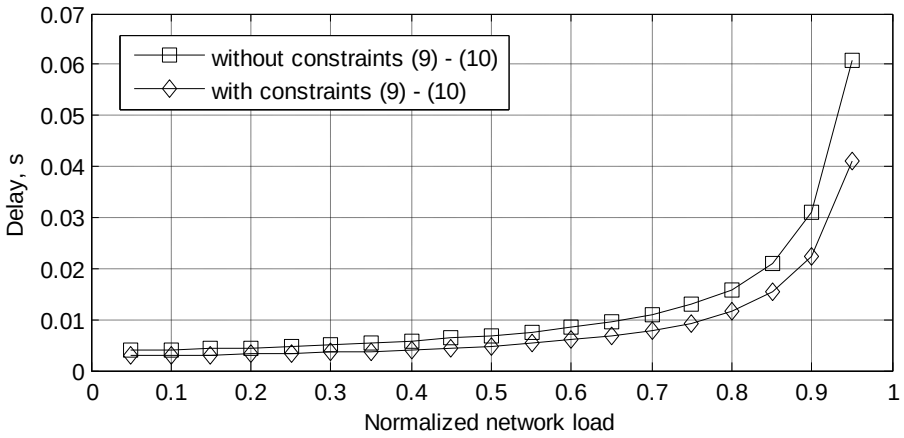


Fig. 8: Delay's gain affected by proposed model

Thus, numerical results demonstrate main logic of proposed model:

- under low request's intensity it's irrationally to use all servers;
- load balancing among servers is rational only under medium and high load, when network load is 30% and above;
- load balancing among servers is based on paths from the servers to clients. Length of paths, current residual capacity and delay along them affect selecting of server and its load. Server and path with higher quality is more preferable.

CONCLUSION

Hybrid content delivery networks allow to combine low cost of services with high their quality, but at same time they toughen up requirements to management practices, including management of requests.

In order to maximize the effectiveness of content delivery system in whole, to provide guaranteed quality of service and to use resources in effective way a management of requests in HCDN should be based on a rigorous mathematical justification.

In this paper we have proposed a flow mathematical model that allows using a different number of CDN-servers as content sources.

A distinctive feature of the proposed approach is taking into account state of the underlying transport infrastructure for selecting CDN-servers to redirect request. So, problems of request routing in HCDN and traffic routing in transport telecommunication network are solved jointly.

Tensor QoS-constraints (9) – (10) make possible to provide guarantees on the quality of delivery, and due to chosen criterion of optimality all resources in HCDN (computational resources of CDN-servers and link resources in transport network) are used in a balanced manner.

REFERENCES

1. Lu ZH, Wang Y, Yang YR; An Analysis and Comparison of CDN-P2P-hybrid Content Delivery System and Model. Journal of Communications, 2012; 7(3): 232–245.
2. Jiang H, Li J, Li Z, Bai X; Efficient large-scale content distribution with combination of CDN and P2P networks. International Journal of Hybrid Informational technology,2009; 2(2):13–24.
3. Buyya R, Pathan M, Vakali A; Content Delivery Networks. Springer, 2008: 418.
4. Yevsyeyeva O, Khader MB; Hierarchical control method for hybrid content delivery network. Modern Problems of Radio Engineering, Telecommunications and Computer Science. Proceedings of the international Conference TCSET'2014,Lviv-Slavske, Ukraine, February 25 - March 1, 2014: 554–556.
5. Yevsyeyeva O; Tensor model of multipolar telecommunications network. Radiotekhika, 2013; 175: 154–159. (Russian)
6. Lemeshko AV, Vavenko TV; Improvement of flow model for multipath routing based on load balancing. Problemitelekomunìkacìj, 2012; 6: 12–29. (Russian)