Hierarchical Control Method for Hybrid Content Delivery Network

Oksana Yevsyeyeva, Mohammed B. Khader

Abstract - **A hierarchical control method for hybrid content delivery network that integrates advantages of P2P and CDN architectures is offered. The method implies two layers where lower layer is associated to control traffic within P2P-network but upper layer is related to CDN. The method is based on a flow mathematical model which allows solving traffic control problem in optimal way. The model includes conditions to ensure required rate and delay and allows balancing of network and computing resources using by multipath routing and redistribution requests between content servers.**

Keywords – **Hierarchical control method, Hybrid content delivery network, Optimal request routing.**

I. Introduction

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; both of them were development to improve effectiveness of content distribution over Internet. Whereas CDN concept is based on usage of 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].

II. Mathematical model of CDN

The hybrid content delivery networks give rise to a set of problems, one of which is management and control problem. The control system for the hybrid content delivery network should combine centralized, based on the global state information approach used in the CDN, and the principles of local, decentralized control realized in P2P. Whereas user's requirements for the quality of service are increasing, request management, including the selection of the server or the set of servers and content delivering, should be based on the type of service and the requested quality. The additional control subproblem is related to load balancing between CDN and P2P-networks, as well as the sustainable use of resources within each of them. Moreover, in order to improve the effectiveness of HCDN the management of different resources should be highly consistent, i.e. management of computing server's resources should be coordinated with management of telecommunications resources. Here under management of telecommunications resources we will understand problem of channel and buffer distribution within telecommunications network that is providing transport functions for content delivering.

Some of the requirements can be satisfied by using proposed in [2] the mathematical model of CDN. The model allows to solve two problems jointly:

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

(2) to define paths (routes) to deliver the content through telecommunications network from chosen servers to user.

The model is given by the following:

$$\sum_{j\in N} x_{ij}^{kl} - \sum_{j\in N} x_{ji}^{kl} - \sum_{g\in G_l} x_{gi}^{kl} = \begin{cases} 0, \text{ if } i \neq s_k, t_k, \\ -1, \text{ if } i = t_k, \end{cases} \quad (1)$$

$$\sum_{j\in N} x_{gj}^{kl} = \begin{cases} 0, \text{ if } g \neq s_k, \\ y_g^{kl}, \text{ if } g = s_k, \end{cases} \quad (2)$$

$$0 \le x_{ij}^{kl} \le 1, \quad (3)$$

$$\begin{cases} y_g^{kl} = 0, \text{ if } g \neq s_k, \\ 0 < y_g^{kl} \le 1, \text{ if } g = s_k, \end{cases} \quad \sum_{g\in S_l} y_g^{kl} = 1, \quad (4)$$

$$\sum_{k\in K}\sum_{l\in L} r^{kl} x_{ij}^{kl} \le c_{ij}, \quad \sum_{k\in K} y_g^{kl} \le P_g^l, \quad (5)$$

where x_{ij}^{kl} is routing variable that contains portion of traffic from sources (replica servers) s_k to destination t_k, which will be transmitted along link (i, j), $(i, j) \in E$; y_g^{kl} is variable of choosing g^{th} server, $g \in S_l$, as source of l^{th} type content for k^{th} destination; S_l is set of servers (replicas and origin servers) that hold l^{th} type content, $S_l \in S$, S – set of all servers in CDN; $l \in L$, $k \in K$, L is set of contents that are held at servers of given CDN, K is set destinations (users which are requested some content),

Oksana Yevsyeyeva, Mohamed B. Khader – Kharkiv National University of Radioelectronics, 14, Lenina str., Kharkiv, 61164, UKRAINE,
E-mail: evseeva.o.yu@gmail.com,
mohamedbasil.84@gmail.com

E is set of links in telecommunication network; c_{ij} is capacity of the link (i, j); P_g^l is productivity of g^{th} server; r^{kl} is total rate at which k^{th} destination downloads l^{th} type of content.

Eqs. (1) and (2) represent the conservation law for the i^{th} router of transport telecommunication network (TTN), the Eq. (3) is related to the multipath style of routing, the Eq. (4) ensures the integrity (wholeness) of content downloaded from multiple servers, Eq. (5) is caused by limited available network and computing resources.

In order to guarantee quality of content delivery the mathematical model must be appended by additional constraints. Use of constraints developed within tensor approach [2] 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)

$$\Lambda_\eta^{\langle\vartheta\rangle} \le \left(G_{\pi\eta}^{\langle 4,1\rangle} - G_{\pi\eta}^{\langle 4,2\rangle}\left[G_{\pi\eta}^{\langle 4,4\rangle}\right]^{-1} G_{\pi\eta}^{\langle 4,3\rangle}\right) T_\eta^{\langle\vartheta\rangle}, \quad (6)$$

$$\sum_{z=1}^{\vartheta} \lambda_{(\eta)}^z = r \ge r^{\langle req\rangle}, \quad (7)$$

where $\Lambda_\eta^{\langle\vartheta\rangle}$ is vector with elements $\lambda_{(\eta)}^z = y_g r$, z is index of node pair between k^{th} user and g^{th} server (Fig.1), $z = \overline{1, \vartheta}$; $\lambda_{(\eta)}^z$ is packet intensity of traffic flow downloaded by k^{th} user from g^{th} server (within z^{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 playback quality of requested content; $G_{\pi\eta}^{\langle 4,1\rangle}$ is the first element of the matrix $G_{\pi\eta}^{\langle 4\rangle}$, $\left\| \begin{array}{ccc} G_{\pi\eta}^{\langle 4,1\rangle} & | & G_{\pi\eta}^{\langle 4,2\rangle} \\ --- & + & --- \\ G_{\pi\eta}^{\langle 4,3\rangle} & | & G_{\pi\eta}^{\langle 4,4\rangle} \end{array} \right\| = G_{\pi\eta}^{\langle 4\rangle}$; $G_{\pi\eta}^{\langle 4\rangle}$ is square $\phi \times \phi$ submatrix of matrix $\left\| \begin{array}{ccc} G_{\pi\eta}^{\langle 1\rangle} & | & G_{\pi\eta}^{\langle 2\rangle} \\ --- & + & --- \\ G_{\pi\eta}^{\langle 3\rangle} & | & G_{\pi\eta}^{\langle 4\rangle} \end{array} \right\| = G_{\pi\eta}$; $\phi = m - 1$, 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 TTN; 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 TTN 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

$$g_v^{ij} = \frac{\rho_i^v - (\rho_i^v)^{N+2} - (N_i^v + 1)(\rho_i^v)^{N+1}(1 - \rho_i^v)}{(1 - (\rho_i^v)^{N+1})(1 - \rho_i^v)\lambda_i^v}, \quad (8)$$

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).

Eqs. (6) and (7) guarantee that under sufficient amount of available resources result delay will be less than $\tau_{\langle req\rangle}$ and download rate will be more than $r^{\langle req\rangle}$. Here unknown variables are λ_i^v which are related to routing variables as $\lambda_i^v = r x_{mj}$, where indexes i and (m, j) define same link in the TTN. Then traffic control problem in CDN can be formulated as optimization problem:

$$\textit{Minimize } Q_x \vec{x} + Q_y \vec{y} \quad (9)$$

Subject to Eqs. (1) – (7),

where $\vec{x}$, $\vec{y}$ are vectors of variables x_{ij}^{kl} and y_g^{kl} respectively; Q_x, Q_y are weight matrices of using the network and computing resources respectively.

III. Hierarchical Control Method for HCDN

In accordance with the hierarchical organization of a HCDN a hierarchical method of traffic control can be proposed. The method assumes two levels, where the lower level is associated with P2P-networks, and the upper level is answerable for request distributing to CDN servers. The method assumes a centralized control scheme within the P2P network, where central P2P-server works as index-server and performs monitoring functions. When P2P-server receives a new request (step 1 in Fig.2) it tries to handle the request without recourse to CDN servers (step 2). This decision is based on current state of the P2P-network (primarily on the number of active peers) and analysis of incoming requests.

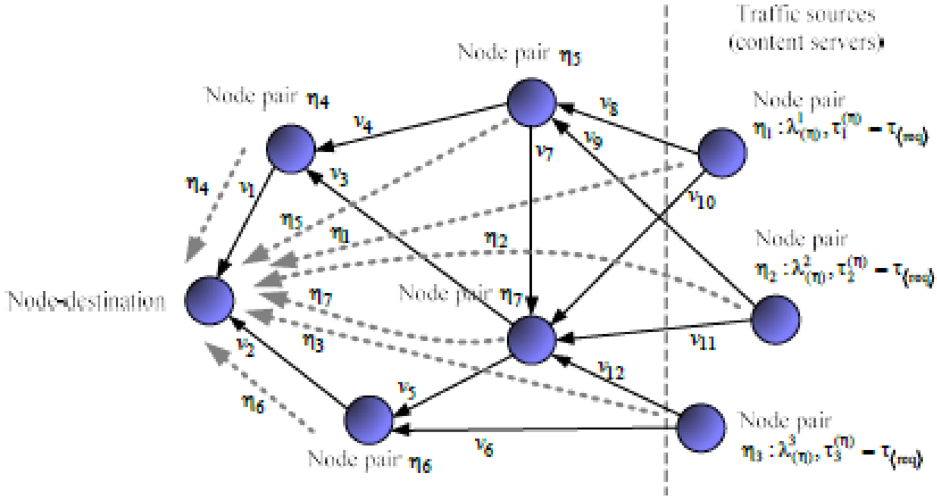

Fig. 1. Example of network with several sources.

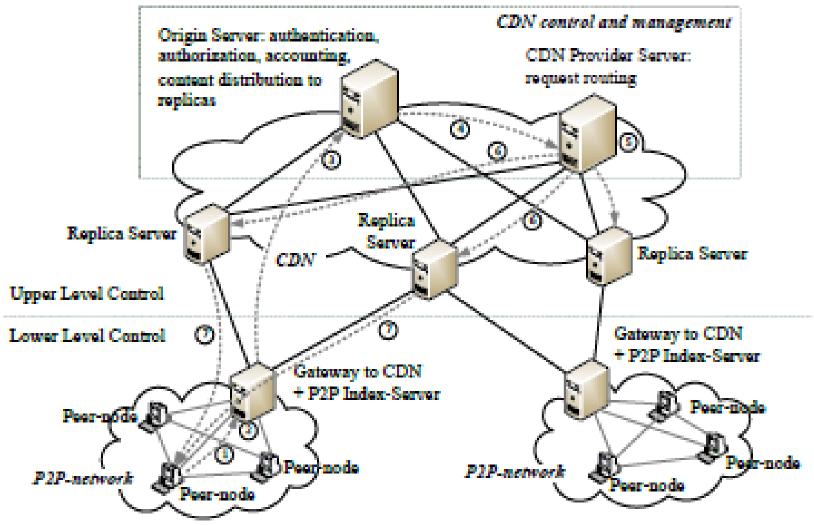

Fig.2. Hierarchical control in hybrid content delivery network.

If the number of active peers that are hold the requested content is not enough to download it at desired quality, the P2P server generates a request to the CDN origin server (the upper level control center) (step 3 in Fig.2). After the authentication and authorization procedures origin server sends a request to the CDN provider server (step 4). CDN provider server performs request routing to one or more of mirror servers (steps 5 and 6). This solution is formed by solving the optimization problem Eq. (9) based on the current workload of CDN servers, usage of transmission links of telecommunication network and the requested quality of service. Loading from several CDN-servers (step 7) allows to use computing and network resources in a balanced manner and to increase the result download speed.

IV. Conclusion

The HCDN allows to achieve maximal effectiveness of content distribution with guaranteed quality of service and relatively low cost. It becomes possible by proposed two layers control method. The method implies (1) 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, (2) solving of request routing on HCDN and traffic routing in transport telecommunication network jointly.

References

[1] Zhi Hui Lu, Ye Wang, Yang Richard Yang, "An Analysis and Comparison of CDN-P2P-hybrid Content Delivery System and Model" *Journal of Communications*, vol. 7(3), 2012. pp. 232-245.

[2] O. Yu. Yevsyeyeva, M.B. Khader, "Mathematical model for resource management in hybrid content delivery network with guaranteed quality of service" *Radiotekhika. All-Urk. Sci. Interdep. Mag.*, № 175, 2013. pp. 26-33 [in Russian].