MATHEMATICAL MODEL OF REQUEST ROUTING IN CONTENT DELIVERY NETWORKS WITH GUARANTEED QUALITY OF SERVICE

Yevsyeyeva O.Yu., Khader M.B.
Kharkov National University of Radioelectronics
14, Lenin Str., Kharkiv, 61166, Ukraine
Ph.: (+38 057) 7021320, e-mail: evseeva.o.yu@gmail.com

Abstract — The request routing in content delivery networks (CDNs) is traditionally related to choosing the appropriate mirrored server only. At the same time, the need of providing guaranteed quality of service necessitates conversion to solving of request routing on CDN together integrated with traffic routing in transport telecommunication network. The mathematical model that allows obtaining integral optimal solution of both tasks has been offered.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ МАРШРУТИЗАЦИИ ЗАПРОСОВ В СЕТЯХ ДОСТАВКИ КОНТЕНТА С ГАРАНТИРОВАННЫМ КАЧЕСТВОМ ОБСЛУЖИВАНИЯ

Евсеева О. Ю., Кадер М. Б.
Харьковский национальный университет радиоэлектроники
просп. Ленина, 14, Харьков, 61166, Украина
тел.: (+38 057) 7021320, e-mail: evseeva.o.yu@gmail.com

Аннотация — Маршрутизация запросов пользователей в сетях доставки контента (CDN) традиционно связывается лишь с выбором зеркалирующего сервера. В то же время необходимость гарантированного качества обслуживания требует рассмотрения задачи маршрутизации запросов в CDN совместно с маршрутизацией трафика в транспортной телекоммуникационной сети. В работе предложена математическая модель, которая обеспечивает комплексное оптимальное решение этих задач.

I. Introduction

Under rapid growth of number of users accessing different web-based content role of Content Delivery Networks (CDNs) is growing too. Concept of CDN was developed as effective way to perform reliable and timely delivery of content to end users by replicating content from original server over several mirrored servers and redirecting a request to the closest replica [1].

The efficiency of the CDN is essentially determined by efficiency of resource management and allocation. In turn this task includes two fundamental problems: request routing and content placement. Both of them arise in the process of functioning and are mutually dependent. First problem is related to selecting a server (original or replica) which holds required content according some objective function (cost function, for example). Content placement problem is caused by limited storage capacity of mirrored servers (replicas). As a result a replica can hold only subset of content.

Within the bounds of article we'll concentrate on request routing problem where mathematical modeling is a powerful effective tool that can be used to obtain optimal solution. Traditionally within CDN request routing problem is associated with choosing the "best" server only. But CDN is spanning network which works over Internet and uses its transport infrastructure to deliver content to end user. In this case quality of content delivering depends on two factors: (1) choosing the minimal loaded server which holds required content (traditional request routing), (2) choosing route (or set of routes) between user's router and selected server along which numerical values of parameters of quality will be satisfy the content requirements (QoS routing problem).

Thus within CDN in order to achieve reliable and timely delivery with guaranteed quality of service request routing problem must be solved jointly with QoS routing problem where required values of QoS-parameters will be taken into account when choosing server and routes.

II. Mathematical Model of Request Routing

In view of the aforesaid mathematical model of routing in CDN must use two different types of variables:

y_g^{kl} — binary variable of choosing p^{th} server, $p \in S_l$, as source of l^{th} type content for k^{th} pair source-destination $\{s_k, t_k\}$, $s_k \in S_l$,

$$y_g^{kl} = \begin{cases} 0, \ if \ \ p \neq s_k, \\ 1, \ if \ \ p = s_k, \end{cases} \quad (1)$$

where 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;

x_{ij}^{kl} — routing variable that contains portion of traffic between source and destination $\{s_k, t_k\}$, which will be transmitted along link (i, j), $(i, j) \in E$. Here indexes l and k are related to type of requested content and pair source-destination respectively, $l \in L$, $k \in K$, L is set of contents that are held at servers of given CDN, K is set of pairs source-destination, E is set of links in transport telecommunication network (TTN).

Variable x_{ij}^{kl} can be binary in single route case and not-binary ($0 \le x_{ij}^{kl} \le 1$) in multipath delivering case.

In accordance with the meaning defined variables must comply with conservation law:

$$\sum_{j \in N} x_{ij}^{kl} - \sum_{j \in N} x_{ji}^{kl} - \sum_{p \in S_l} y_p^{kl} x_{pi}^{kl} = \begin{cases} 0, \ if \ \ i \neq s_k, t_k; \\ -1, \ if \ \ i = t_k; \end{cases} \quad (2)$$

No nodes of TTN can be source of content within given CDN, so condition (2) allows two possible values: 0 for transit nodes and (-1) for node-destination (node requested some content). For servers which are pro-

 ISBN: 978-966-335-395-1. IEEE Catalog Number: CFP13788

spective sources of traffic (content) in CDN conservation constraint has form:

$$\sum_{j\in N} x_{pj}^{kl} = \begin{cases} 0, \ if \ \ p \neq s_k; \\ 1, \ if \ \ p = s_k; \end{cases} \quad p \in S_l . \qquad (3)$$

Due to limited network and server resources

$$\sum_{k\in K}\sum_{l\in L} r^{kl} x_{ij}^{kl} \le c_{ij} , \qquad (4)$$

$$\sum_{k\in K} y_p^{kl} \le R_p^l , \qquad (5)$$

where r^{kl} — intensity (rate) of traffic generated by server-source s_k (within k th pair) when it transmits l th type of content; c_{ij} — capacity of the link (i, j); R_p^l — maximum number of l th type session which p th server is able to maintain simultaneously (productivity of server).

If the "bottleneck" is the network interface, the restriction (6) related to the performance of the server can be represented as:

$$\sum_{k\in K}\sum_{l\in L} r^{kl} x_{pj}^{kl} \le c'_{pj} , \ \ p \in S_l , \qquad (6)$$

where c'_{pj} — capacity of the interface, which connects p th server of CDN to j th router of TTN.

In order to achieve required quality of service we need include relevant constraints in the model. As analysis shows known request routing models use simple QoS constraints in general form. But in order to take into account flow nature of network traffic, nonlinear depending result quality of servicing on intensity of traffic, multipath fashion of transmitting we must use more difficult mathematical expressions. One of such QoS constraints was developed within tensor approach [2] and has next form

$$\lambda^{\langle req\rangle} \le \left(G_{\pi\eta}^{\langle 4,1\rangle} - G_{\pi\eta}^{\langle 4,2\rangle} \left[G_{\pi\eta}^{\langle 4,4\rangle} \right]^{-1} G_{\pi\eta}^{\langle 4,3\rangle} \right) \tau_{\langle req\rangle} , \qquad (7)$$

where $\lambda^{\langle req\rangle}$ and $\tau_{\langle req\rangle}$ are numerical values of rate and delay, respectively, required for acceptable quality of playback of requested content; $G_{\pi\eta}^{\langle 4,1\rangle}$ – the first element of the matrix $G_{\pi\eta}^{\langle 4\rangle}$, $\left\| \begin{matrix} G_{\pi\eta}^{\langle 4,1\rangle} & | & G_{\pi\eta}^{\langle 4,2\rangle} \\ --- & + & --- \\ G_{\pi\eta}^{\langle 4,3\rangle} & | & G_{\pi\eta}^{\langle 4,4\rangle} \end{matrix} \right\| = G_{\pi\eta}^{\langle 4\rangle}$; $G_{\pi\eta}^{\langle 4\rangle}$ — square $\phi\times\phi$ submatrix of matrix $\left\| \begin{matrix} G_{\pi\eta}^{\langle 1\rangle} & | & G_{\pi\eta}^{\langle 2\rangle} \\ --- & + & --- \\ G_{\pi\eta}^{\langle 3\rangle} & | & G_{\pi\eta}^{\langle 4\rangle} \end{matrix} \right\| = G_{\pi\eta}$;

$\phi = m-1$, m — the number of nodes in the network; $G_{\pi\eta}$ — $n\times n$ matrix calculated according to $G_{\pi\eta} = A^t G_v A$; n — the number of links in the TTN; A and C — $n\times n$ 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); $G_v = \left\| g_v^{ij} \right\|$ — 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 $M/M/1/N$ [3] as model of given link, then i th element is calculated according to

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

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

Condition (7) guarantees that under the stipulation that amount of availability of resources is sufficient result delay will be less than $\tau_{\langle req\rangle}$ and result intensity (rate) will be more $\lambda^{\langle req\rangle}$ by calculating variables λ_i^v which were defined before as routing variables.

Thus, within the framework of the model (1) – (8) the process of request routing in CDN is associated with calculation variables y_p^{kl}, and the presence of unknown variables x_{ij}^{kl} allows solve this problem together with the traditional problem of multipath QoS-routing in TTN. So routing problem in CDN is formulated as the optimization problem where a cost function can be used as the objective function

$$W = Q_x \vec{x} + Q_y \vec{y} \to \mathbf{min} , \qquad (9)$$

where $\vec{x}$ and $\vec{y}$ — vectors of variables x_{ij}^{kl} and y_g^{kl}, respectively; Q_x, Q_y — vectors of weighting coefficient, that determine the cost of using the network and server resources, respectively.

III. Conclusion

Thus, the offered mathematical model (1) — (9) formalizes the request routing problem in the CDN as a constrained optimization problem. Its main advantages are, firstly, the possibility of joint (and therefore highly coordinated) solutions of request routing problem and multipath routing in the TTN, secondary, guaranteed quality of service. In order to achieve reliable and timely content delivery the offered model takes into account two metrics (rate and delay) and by reallocating traffic between links in transport telecommunication network the model can satisfy these two requirements simultaneously (under the stipulation that amount of availability of resources is sufficient).

IV. References

[1] Rajkumar Buyya, Mukaddim Pathan, Athena Vakali (Eds.) *Content Delivery Networks* (*Series: Lecture Notes in Electrical Engineering)*. Springer, 2008. 418 p.

[2] Lemeshko A.V., Yevseyeva O.Yu. Tensor model of multipath routing based on multiple QoS metrics. *Problems of telecommunication*, 2012, No 4 (9), pp. 16-31. Available at: http://pt.journal.kh.ua/2012/4/1/124_lemeshko_tensor.pdf. (accessed 30 July 2013).

[3] Kleinrock L. *Queueing Systems: Volume I – Theory.* New York: Wiley Interscience, 1975. 417 p.

ISBN: 978-966-335-395-1. IEEE Catalog Number: CFP13788