УДК 621.3

М. І. Кирик
Національний університет “Львівська політехніка”

ДОСЛІДЖЕННЯ МЕХАНІЗМІВ УПРАВЛІННЯ КОНТЕНТОМ У МЕРЕЖАХ CDN

© Кирик М. І., 2016

Розглянуто концепції мережі доставки контенту (CDN) та балансування навантаження. Основне завдання мережі (CDN) – забезпечення якісної доставки інформації кінцевому користувачу. Запропоновано архітектуру для вибору оптимального маршруту і сервера для оброблення запитів користувачів на основі цільової функції. Основними критеріями цільової функції вибрано час затримки, навантаження на сервер роздавання контенту і ймовірність втрати пакетів.

У роботі проведено експериментальне дослідження відеоконтенту в мережі CDN. Представлено графічні залежності інтенсивності трафіку, який надходить на сервери досліджуваної мережі. Побудовано залежності часу затримки та джитера від інтенсивності надходження даних на кешуючі сервери за різних значень інтенсивності обслуговування. Розглянуто основні методи балансування навантаження, яке дасть змогу раціонально розподілити навантаження між сереверами мережі.

Ключові слова: мережа доставки контенту (CDN), балансування навантаження, якість обслуговування, час затримки, джитер, цільова функція.

M. Kyryk
Lviv Polytechnic National University

STUDY OF THE LOAD MANAGEMENT MECHANISMS IN CONTENT DELIVERY NETWORKS

© Kyryk M., 2016

Concept content delivery network (CDN) system and load balancing mechanisms were considered in this paper. The main task of the CND network - providing the qualitative information delivery to the end user. Content Delivery Network (CDN) – is a geographically distributed network, that contains a number of content servers and routers. As a rule, it consists of a main node (Origin), and caching nodes (Edges) – points of presence, which can be located in various parts of the world. All content are stored and updated on the Origin server. The main advantages of the CDN are the next:

- **increasing the content delivery speed;**
- **reducing the load on the Origin server;**
- **reducing the load on the primary server.**

We propose architecture, for select the optimum route and server for a user request, based on the objective functions. The main criterions of the objective function were taken delay time, server load and probability of packet lost. All user requests are sent to the nearest or lowest-load server to quicken the response.

As part of this paper experimental research video content in CDN network was conducted. The main attention was focused on the primary server to broadcast video content and one of Edge servers. The intensity of the packets flow to Origin server and Edge server were presented. With the increasing load on the Edge server, server resources increase is needed. The load balancing technology can ensure that the user request points to the nearest

edge server with minimum load in the network, so that network content is efficiently distributed. Load balancing system performs the following tasks: monitoring of servers and network equipment; choosing the server that will respond to the client's request; traffic control between client and server.

Dependencies delay time of the intensity receipt of user requests on a Edge server was presented. As we can see from the presented dependencies, delay time and jitter increases with the intensity of the input stream. Load balancing mechanism will allow more efficiently redirect user requests to the Edge servers. This makes it possible to provide services of better quality, ie less latency and jitter, which is critical for real-time services.

Key words: content delivery network, load balancing, QoS, delay time, jitter, objective function.

Вступ

Висока інтенсивність зростання трафіку в мережі Інтернет призводить до збільшення кількості запитів до популярних ресурсів, що доволі часто створює трафік, який далеко виходить за межі можливостей одного сервера. У зв'язку з цим можливі тривалі затримки у доступі користувачів до мережевих додатків. Зміна інфраструктури сайта чи будь якого іншого серверного додатка на локальний кластер не забезпечує повного вирішення цієї проблеми, оскільки канал між кластером і глобальною мережею може стати вузьким місцем цієї інфраструктури. Ефективнішим рішенням є розподіл серверів географічно так, щоб вони розташовувалися в окремих мережах. За використання такої архітектури роль балансування навантаження істотно зростає. Це зумовлено тим, що об'єкт, який розподіляє запити, може здійснювати перенаправлення як на основі завантаження мережі та серверів, так і на основі відстані між клієнтом і серверами.

Робота спрямована на поліпшення якості роботи алгоритмів балансування навантаження з метою покращення якості обслуговування в мережах розподілу контенту CDN [1].

Архітектура CDN мереж

Мережа доставки (розподілу) контенту являє собою географічно розподілену мережу передавання даних, яка містить велику кількість серверів обробки та трансляції контенту і мережевих маршрутів. Основне завдання такої мережі – забезпечення якісної доставки інформації до кінцевого користувача. Як правило, мережа складається із головного сервера, на якому міститься контент, та кешуючих серверів, які розташовані в різних географічних точках. Коли користувач відправляє запит до головного сервера, цей запит буде перенаправлений до найближчого кешуючого сервера. Завдяки цьому маршрут між кінцевим користувачем та сервером трансляції контенту істотно скорочується і клієнт має змогу отримати ресурс з більшою надійністю.

Основні переваги використання CDN:

- *зростання швидкості доставки контенту*. Користувачі із усього світу мають змогу отримати контент по оптимальному мережевому маршруту із мінімальною затримкою;
- *зниження навантаження на основний сервер*. Всі статичні дані будуть зберігатись на кешуючих серверах і всі запити будуть опрацьовувати, як правило, саме вони. На основному сервері залишатиметься тільки динамічний контент. Для прикладу, якщо мережа CDN використовується для трансляції відеоконтенту, то контент, який іде в реальному часі, буде транслюватись із основного сервера, а весь інший, наперед записаний контент, віддаватимуть кешуючі сервери;
- *використання файлових обмінників для передавання даних*. Багато користувачів використовують сховища даних (storage) для обміну даними. Як правило, там зберігаються великі файли відео, аудіо та фото. Завдяки використанню CDN ці файли можуть бути завантажені на високій швидкості із будь-якої точки світу.

Базова архітектура мережі CDN наведена на рис. 1.

Рис. 1. Архітектура мережі CND

Всі запити, що надходять від користувачів, направляються до найближчого або ж до найменш завантаженого сервера для отримання швидкої відповіді. Як правило, механізм роботи CDN полягає у виборі сервера віддачі контенту для кожного користувача. Найпоширеніші методи вибору такі:

– вибір сервера, найближчого до користувача, на основі служби доменних імен DNS;
– визначення завантаженості кожного сервера та вибір найменш завантаженого;
– вибір сервера, який забезпечує мінімальне значення затримки під час передавання даних на шляху до кінцевого користувача [2].

Всі ці методи необхідні для забезпечення ефективного використання ресурсів мережі та вузлів доставки контенту [3]. Не можна забувати і про якість обслуговування QoS у такій мережі. Для забезпечення задовільної якості обслуговування у нашій роботі введено поняття цільової функції. Основна її умова – це забезпечення заданої якості сервісу в CDN мережі. Цільову функцію подамо у такому вигляді:

$$F = \min(t_i, l_i, p_i), \tag{1}$$

де t_i – час затримки передавання даних від i-го кешуючого (Edge) сервера до кінцевого користувача; l_i – завантаженість i-го кешуючого (Edge) сервера; p_i – імовірність того, що пакет буде втрачений на шляху передавання від i-го кешуючого (Edge) сервера до кінцевого користувача.

Значення часу затримки, навантаження на сервер та імовірність втрати пакетів на шляху передавання повинні відповідати певним вимогам, а саме:

$$t_i < T_i, l_i < L_i, p_i < P_i, \tag{2}$$

де T_i – максимально дозволене значення часу затримки передавання даних; L_i – максимально дозволене значення завантаженості сервера; P_i – максимально допустиме значення імовірності втрати пакетів під час передавання даних.

Отже, використання цільової функції дасть можливість вибирати оптимальний кешуючий сервер, який опрацьовуватиме запити користувачів.

У межах цієї роботи проведено експериментальні дослідження CDN мережі доставки відеоконтенту. Увагу зосереджено передусім на основному сервері трансляції відеоконтенту (Origin) та одному із кешуючих (Edge) серверів. Інтенсивність надходження трафіку на основний сервер та від нього відображено на рис. 2.

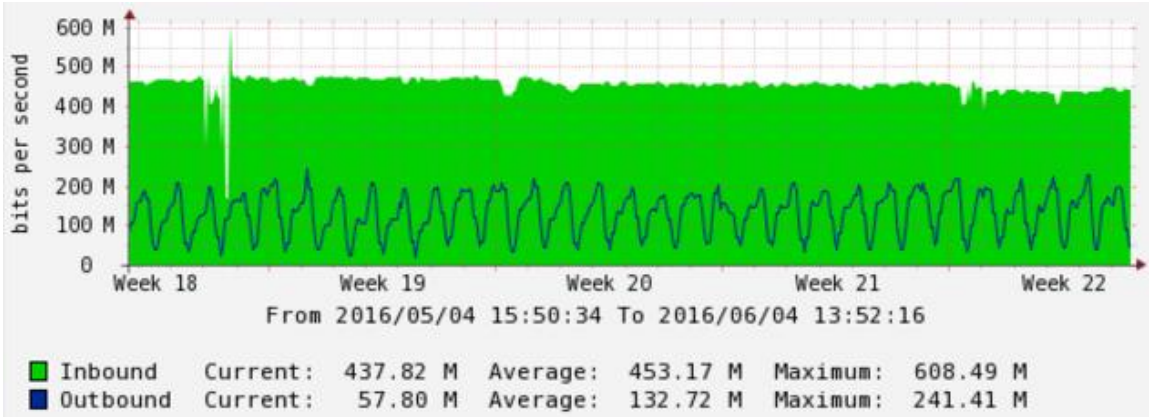


Рис. 2. Інтенсивність трафіку на основному сервері CDN мережі

Як бачимо із рис. 2, на основний сервер надходить потік даних доволі високої інтенсивності (Inbound), що являє собою відеоконтент для кінцевих користувачів. Вихідний потік із основного сервера – це дані, які передаються кінцевим користувачам через кешуючі сервери (Outbund). Як бачимо із графіка, середня інтенсивність вхідного потоку становить 453 Мбіт/с, середня інтенсивність вихідного – 132 Мбіт/с. На рис. 3 подано трафік, який опрацьовує кешуючий сервер.

На рис. 3 бачимо, що на кешуючий сервер потрапляє динамічний трафік із основного сервера, а контент є закешованим і потрапляє до користувачів сервісу прямо від нього. Вхідний трафік на кешуючому сервері є частиною вихідного трафіку із основного сервера (рис. 2. Outbound traffic), який віддає контент також й іншим кешуючим серверам. Вихідний трафік переважає вхідний і в середньому його інтенсивність становить 150–200 Мбіт/с. Весь цей трафік генерують запити на контент від кінцевих користувачів сервісу, який надає ця CDN мережа. Проаналізувавши результати експериментальних досліджень, можна сказати, що використання кешуючих серверів дає змогу істотно зменшити завантаженість основного сервера, знизити затримку, яка виникає у кінцевого користувача під час отримання контенту, а також знижує імовірність втрати даних на шляху передавання. Всі ці переваги відповідають критеріям цільової функції та дають змогу покращувати якість сервісу в мережах передавання даних.

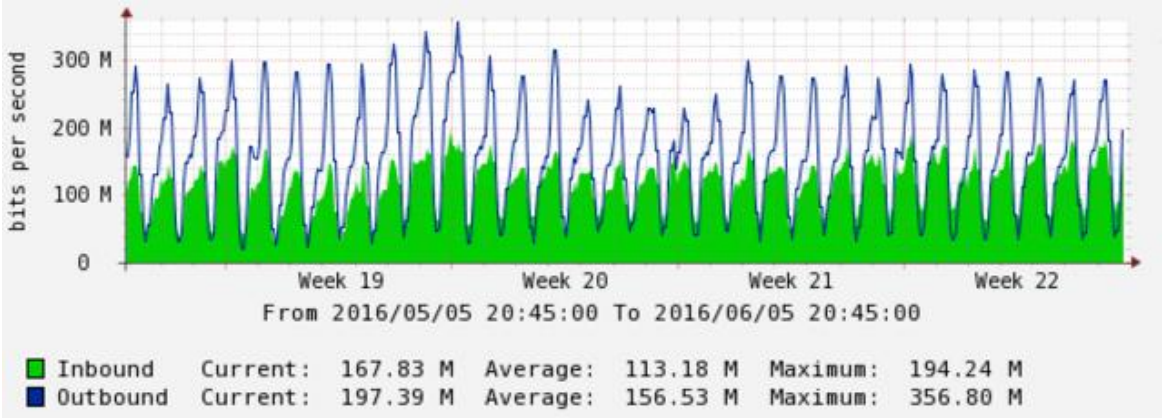


Рис. 3. Інтенсивність трафіку на кешуючому сервері CDN мережі

Механізми балансування навантаження у CDN мережах

Коли зростає навантаження на кешуючі сервери, виникає потреба у підвищенні їхньої продуктивності або ж збільшенні їх кількості. У разі використання другого принципу навантаження між серверами можна рівномірно розподіляти, використовуючи технологію балансування навантаження. Технологія балансування навантаження є одним з основних елементів у CDN [4]. Використання цієї технології може гарантувати, що запити користувачів будуть спрямовані до найближчого кешуючого сервера з мінімальним навантаженням у мережі, розподіляючи цим ефективне використання ресурсів CND мережі. На рис. 4 подано схему балансування навантаження у мережі CDN, яка містить n серверів.

Рис. 4. Схема балансування навантаження у CDN мережі

Вважатимемо, що $l_i, i = 1...n$ – це інтенсивність надходження пакетів на вузол i у момент часу t. У такому випадку сумарна інтенсивність у мережі визначатиметься як:

$$l = \sum_{i=1}^{n} l_i \tag{3}$$

Варто зазначити, що окрім власних запитів, кожен кешуючий сервер може також отримувати запити від інших серверів. Очевидно, що процент завантаженості серверних ресурсів, які виділяються на опрацювання запитів від інших серверів, залежатиме від багатьох факторів. Отже, інтенсивність запитів, які надходять від інших серверів, буде визначатись за такою формулою:

$$a_i = \sum_{j=1}^{n} l_i \cdot w_{ji}, \tag{4}$$

де w_{ji} – частина потоку l_i, яка перенаправляється від сервера j до сервера i. В результаті середню інтенсивність вхідного навантаження, що надходить на вузол i у конкретний період часу, можна визначити за формулою:

$$a_i = l_i - \sum_{j=1}^{n} l_i \cdot w_{ij} + \sum_{j=1}^{n} l_j \cdot w_{ji} \tag{5}$$

Перший доданок визначає частину навантаження, що вузол i переадресовує до інших вузлів, наступний – частину навантаження, яке отримав вузол i від інших серверів. Інтенсивність обробки пакетів m_i у вузлі i має бути більшою, ніж інтенсивність приймання пакетів l_i у режимі нормальної роботи.

Коли ідеться про дослідження якості обслуговування в мережах доставки контенту чи про методи оптимізації та розподілу навантаження між вузлами мережі, важливе значення має середній час затримки передавання даних до кінцевого користувача. Будемо вважати, що потік заявок, які надходять на вузол обслуговування, є пуассонівським. Кожен вузол розглянемо як систему масового обслуговування типу M/M/1 [5–8]. В такому випадку середній час затримки буде визначатись як

$$T = \frac{1}{l} \cdot \sum_{i=1}^{n} a_i \cdot T_i = \frac{1}{l} \cdot \sum_{i=1}^{n} \left(\frac{a_i}{m_i - a_i}\right). \tag{6}$$

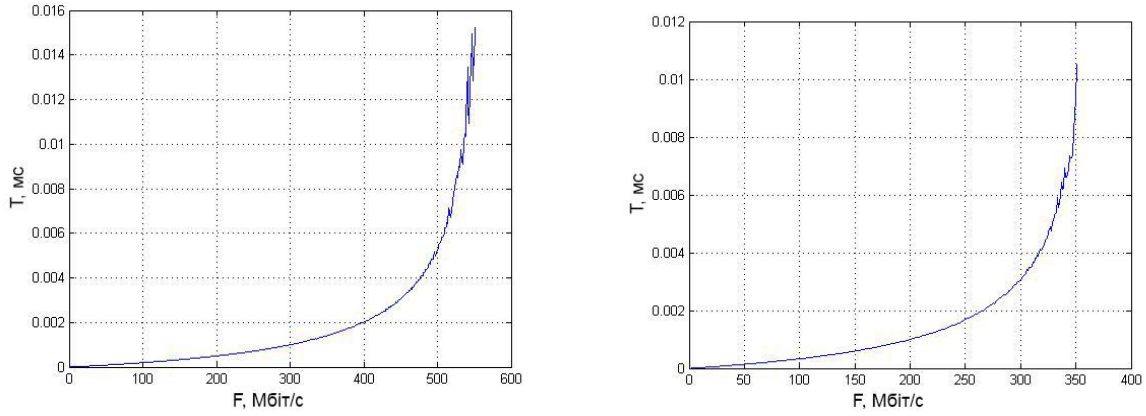
Враховуючи затримку під час передавання даних між обслуговуючими пристроями, інтенсивність обробки пакетів та інтенсивність надходження до вузла, середню затримку можна визначити так:

$$T = \frac{1}{l} \cdot \sum_{i=1}^{n} \left(\frac{F_i}{C_i - F_i} + T_{Fi}\right), \tag{7}$$

де T_{Fi} – затримка передавання між кешуючими серверами; C_i – інтенсивність обслуговування вузла оброблення контенту, яка в цьому випадку визначатиметься пропускною здатністю; F_i –

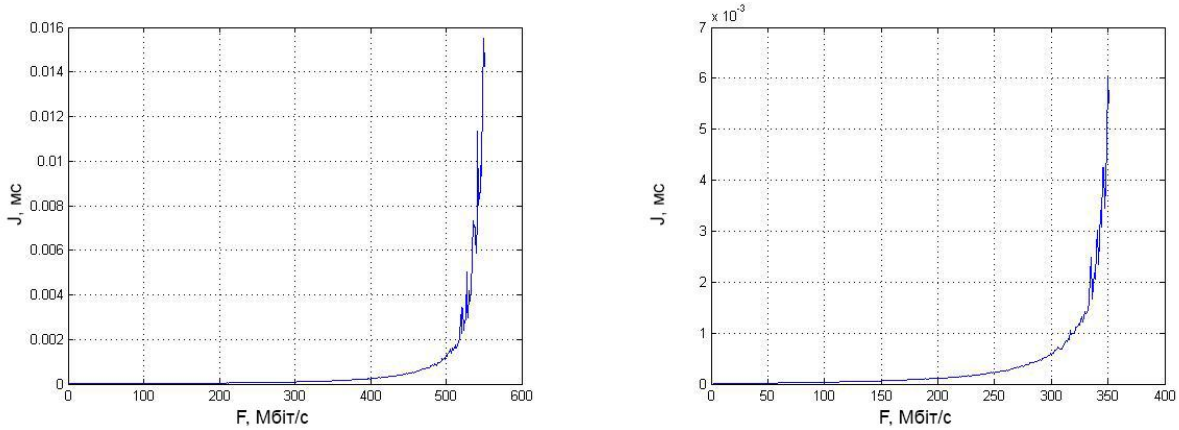
інтенсивність надходження пакетів на вузол обслуговування. Із формули (6) очевидно, що затримка зростатиме, коли вузол перевантажений обробкою запитів. Однією з проблем розподілу ресурсів у системі доставки контенту є вибір оптимальної структури мережі вузлів [9, 10]. Для мінімізації затримки потрібно зменшити навантаження на кожен із серверів та прокласти оптимальні маршрути між серверами обробки контенту. Наведемо графічні залежності часу затримки від кількості кешуючих серверів та інтенсивності потоку, який вони опрацьовують.

Як бачимо із рис. 5, 6, затримка та джитер збільшуються зі зростанням інтенсивності вхідного потоку. Для того, щоб мінімізувати це значення, потрібно збільшити продуктивність обслуговуючих пристроїв або ж мінімізувати затримки між кешуючими серверами. Оскільки не завжди доцільно нарощувати ресурси обслуговуючих пристроїв, то один з варіантів зменшення завантаженості серверів обробки запитів кінцевих користувачів – використання балансування навантаження між серверами CDN мережі. Механізм балансування навантаження дасть змогу раціональніше перенаправляти запити до серверів, які зможуть надати сервіс із кращою якістю, а саме з меншою затримкою та джитером, що критично для послуг реального часу.



а *б*

Рис. 5. Залежність часу затримки обслуговування на кешуючому сервері від інтенсивності надходження запитів користувачів за різних значень інтенсивності обслуговування: а – C_i =600 Мбіт/с; б – C_i =400 Мбіт/с



а *б*

Рис. 6. Залежність джитера від інтенсивності надходження запитів користувачів за різних значень інтенсивності обслуговування: а – C_i =600 Мбіт/с; б – C_i =400 Мбіт/с

Висновки

У роботі розглянуто концепцію мережі доставки контенту CDN та основні принципи балансування навантаження між серверами. Запропоновано архітектуру для вибору оптимального

маршруту і сервера для обробки запитів користувачів на основі цільової функції. Основними критеріями цільової функції взято час затримки, навантаження на сервер і ймовірність втрати пакетів. Експериментально досліджено роботу кешуючих серверів мережі доставки відеоконтенту. Представлено інтенсивності трафіку на основному сервері мережі та одному із кешуючих Edge серверів. Проаналізувавши результати експериментальних досліджень, можна стверджувати, що використання кешуючих серверів дає змогу істотно зменшити завантаженість основного сервера, знизити затримку, яка виникає у кінцевого користувача під час отримання контенту, а також знижує імовірність втрати даних на шляху передавання. Всі ці переваги відповідають критеріям цільової функції та дають змогу покращувати якість сервісу в мережах передавання даних. Побудовано залежності часу затримки та джитера від інтенсивності надходження даних на кешуючі сервери за різних значень інтенсивності обслуговування. З графіків стає зрозуміло, що зі зростанням навантаження збільшується і затримка обслуговування. З цією метою пропонуються методи балансування навантаження між серверами роздавання контенту. Розглянуто основні методи балансування навантаження, які дадуть змогу раціонально розподілити навантаження між серверами мережі й покращити якість обслуговування QoS.

1. Awduche D., Chiu A., Elwalid A., Widjaja I. Overview and principles of internet traffic engineering. IETF. RFC3272. 2002. 2. Sivasubramanian S., Szymaniak M., Pierre G. Replication for web hosting systems // ACM Computing Surveys. 2004. Vol. 36, no. 3. P. 291–334. 3. Data Buffering Multilevel Model at a Multiservice Traffic Service Node / Mykhailo Klymash, Maryan Kyryk, Nazar Pleskanka, Volodymyr Yanyshyn // Smart Computing Review. Korea. Vol. 4. No. 4. August 31. 2014. P. 294–306. 4. OLNN.CN.CDN. The Content distributed network technology, server, May 28, 2007. URL: http://olnn.cn/html/server/6/20070528/3660.html. 5. Дмитриев Г. А., Марголис Б. И., Музанна М. М. Решение задачи оптимальной маршрутизации по критерию загруженности сети // Программные продукты и системы. 2013. № 4. С. 17–19. 6. Парфенов В. И., Золотарев С. В. Об одном алгоритме решения задачи оптимальной маршрутизации по критерию средней задержки // Вестник ВГУ: сер. Физика. Математика. 2007. № 2. С. 28–32. 7. Шварц М. Сети связи: протоколы, моделирование и анализ: пер. с англ.: в 2 ч. Москва: Наука, 1992. Ч. 1. 336 с. 8. Клейнрок Л. Теория массового обслуживания / пер. с англ. И. И. Грушко; ред. В. И. Нейман. М.: Машиностроение 1979. С. 292–320. 9. Replication based on Objects Load under a Content Distribution Network / Pallis George, Konstantinos Stamos, Athena Vakali "and other" // Proceedings of the 22nd (In conjunction with ICDE'06). Atlanta, 2006. 53 p. 10. Load balancing through efficient distributed content placement / Tim Wauters, Jan Coppens, Bart Dhoedt, Piet Demeester // In proceeding of: Next Generation Internet Networks. NY. 2005. P. 99–105.

References

1. D. Awduche, A. Chiu. A Elwalid, I. Widjaja Overview and principles of internet traffic engineering. IETF, RFC3272, 2002. 2. Sivasubramanian S., Szymaniak M., Pierre G., "Replication for web hosting systems," ACM Computing Surveys, vol. 36, no. 3, no. 3, pp. 291–334, 2004. 3. Mykhailo Klymash, Maryan Kyryk, Nazar Pleskanka, Volodymyr Yanyshyn Data Buffering Multilevel Model at a Multiservice Traffic Service Node // Smart Computing Review. Korea – Vol. 4. No. 4. August 31, 2014, p. 294–306. 4. OLNN.CN.CDN, The Content distributed network technology, server, May 28, 2007, http://olnn.cn/html/server/6/20070528/3660.html. 5. G. Dmitriev, B. Margolis, M. Muzanna. Solution of the optimal routing problem on the criterion of network congestion // Software products and systems. – 2013. – No. 4. – P. 17–19. 6. V. Parfenov, S. Zolotarev. An algorithm for solving the optimal routing problem by the criterion of average delay // Vestnik VSU: Ser. Physics. Mathematics. – 2007. – No. 2. – P. 28–32. 7. M. Schwartz. Communication Networks: Protocols, Modeling and Analysis: Trans. from English .: 2 h. – M .: Nauka, 1992. – Part 1 – 336, 2. – P. 28–32. 8. L. Kleinrock Queueing theory. Trans. with Eng. / Ed. I. Grushko; Ed. V. Neiman. – M .: Engineering, 1979. – P. 292–320. 9. Pallis George, Konstantinos Stamos, Athena Vakali "and other". Replication based on Objects Load under a Content Distribution Network // Proceedings of the 22nd (In conjunction with ICDE'06). – Atlanta, 2006. – 53 p. 10. Tim Wauters, Jan Coppens, Bart Dhoedt, Piet Demeester. Load balancing through efficient distributed content placement // In proceeding of: Next Generation Internet Networks. – NY, 2005. – P. 99–105.