Minimization of delivery cost in content delivery networks with multilevel hierarchical architecture

Article *in* Automatic Control and Computer Sciences · January 2006

CITATION
1

READS
17

1 author:

Mahammad Sharifov
Khazar University
6 PUBLICATIONS **1** CITATION

SEE PROFILE

МИНИМИЗАЦИЯ СТОИМОСТИ ДОСТАВКИ В CONTENT DELIVERY NETWORKS МНОГОУРОВНЕВОЙ ИЕРАРХИЧЕСКОЙ АРХИТЕКТУРОЙ

М.Г.Шарифов, *аспирант*
Институт Информационных Технологий Национальной Академии Наук Азербайджана
ул. Ф.Агаева 9, AZ-1141 Баку, Азербайджан
E-mail: azeridesigner@gmail.com

В статье предлагается модель многоуровневого кэширования в Content Delivery Networks, обеспечивающая минимизацию стоимости доставки Web-контентов конечным потребителям. Минимизация стоимости доставки достигается интегрированным подходом к решению задач: оптимальное распределение запросов, определение оптимальных точек расположения серверов и оптимальное распределение контент-реплик по серверам. Предложенная модель сводится к задаче дискретного программирования.

Ключевые слова: Content Delivery Network, многоуровневое кэширование, минимизация стоимости доставки, распределение запросов, дискретное программирование

---

1. ВВЕДЕНИЕ

В последние годы, среди поставщиков, так и потребителей информации наблюдается огромный рост популярности Internet. Такой рост популярности, а также экспоненциальное увеличение трафика в WWW неизбежно должны были породить серьезные проблемы, среди которых можно отметить таких – повышение нагрузок на опорной сети Internet и на Web-серверы, увеличение быстродействия отклика, потеря пакетов и т.д. Как известно, причиной появления таких проблем является принцип централизованной доставки информации конечным потребителям. При централизованной модели обслуживания, как только пользователь устанавливает соединение с сетью Internet Service Provider (ISP), он автоматически получает доступ к любому выбранному хосту вне зависимости от его удаленности. В этом случае, в центре сферы размещен контент-провайдер (источник первичной информации), осуществляющий информационное наполнение Web-сайта. Такой источник обычно размещается в одном географически удаленном сервере, куда со всего мира стекаются запросы на обслуживание, т.е. соединение организуются через опорную сеть Internet, которое любые узкие места (bottlenecks) могут замедлять доставку информации.

Для оптимизации нагрузки на опорной сети Internet применяется множество решений как экстренного, так и интенсивного характера. Экстенсивный путь решения требует больших материальных затрат и следовательно, с экономической точки зрения является не выгодным. Поэтому исследователям пришлось искать интенсивные пути решения. Усилие исследователей в поисках интенсивных путей решений привело к разработке новых технологий: Web-кэширование (Web caching) [1,2] и сеть доставки контента (Content Delivery Network – CDN) [3,4]. Появление таких технологий помогло перейти от централизованной модели обслуживания к децентрализованной, с множеством сервером, стратегически размещенных в распределенных географических точках. Web-кэширование и Content Delivery Network главным образом сосредотачиваются на уменьшение времени ответа, сохраняя при этом пропускную способность опорной сети Internet. Технология Web-кэширование полагается на тот факт, что информации дешевле и эффективнее хранить, чем передавать. Принцип работы Content Delivery Network то же опирается на технологию

кэширования, которая вносит элементы интеллектуальности в концепцию хранения и пересылки информации, определяя при этом какие информации будут скопированы из ядра на периферию сети, и как часто они будут обновляться [5]. В отличие от Web-кэширования в Content Delivery Network могут задействовать кластеры или иерархии серверов, позволяющие снизить нагрузку на опорной сети Internet и решить проблему масштабируемости серверов. По сути Content Delivery Network является наложенной (overlay) сетью, построенной поверх инфраструктуры Internet с целью доставкой Web-контентов на узлы, расположенные поблизости от клиентов. С использованием технологии Content Delivery Network решается и проблема внезапного всплеска числа запросов, приводящих к перегрузке Web-сервера. Такой поток запросов могут теперь обслужить многочисленные серверы вместо одного централизованного сервера.

С разработкой технологии Content Delivery Network, почти одновременно появились компании, оказывающие услуги Content Delivery Network, среди которых можно отметить Akamai [6], Digital Island [7] и т.д. Провайдеры, оказывающие услуги Content Delivery Network, позволяют контент-провайдерам существенно снизить затраты на распространение информации, при котором стремятся получить максимальную прибыль из своей распределенной инфраструктуры. Для получения максимальную прибыль провайдерам необходимо минимизировать стоимости доставки Web-контентов конечным потребителям [8,9], которая достигается путем решений следующих задач:

- нахождение стратегических точек расположения серверов в узлах сети;
- оптимальное распределение контент-реплик по серверам;
- переадресация запросов на серверы, способные наиболее эффективно их обслужить.

Для увеличения своих прибылей провайдеры Content Delivery Network попытаются привлечь значительного количества клиентов. А это реализуемо лишь увеличением масштабируемости Content Delivery Network. Когда платформа имеет сложную, например, многоуровневую архитектуру, то задача масштабируемости Content Delivery Network становится еще сложнее. При проектировании Content Delivery Network на многоуровневой архитектуре одной из сложных задач является переадресация запросов по серверам. Например, в работах [10,11] для удовлетворения запросов пользователей предлагается такая стратегия: на каждом уровне сначала производится кластеризация серверов по контенту, потом на каждом кластере с целью переадресации запросов ближайшему кластеру определяется представительный сервер. При этом удовлетворение запроса пользователя осуществляется с такой последовательностью:

Шаг 1. Запрос сначала адресуется к локальному серверу;

Шаг 2. Если на шаге 1 запрос не удовлетворен, то он переадресуется серверу, в котором расположен в одном кластере с локальным сервером;

Шаг 3. Если на шаге 2 запрос не удовлетворен, то он переадресуется серверу, расположенному в ближайшем кластере;

Шаг 4. Если на шаге 3 запрос не удовлетворен, то он переадресуется к оригинальному серверу.

Как видно, в этих работах речь идет только о переадресации запросов, при этом подразумевается, что узлы расположения серверов Content Delivery Network заранее известны и Web-контенты реплицированы в этих серверах. Если заранее не известны узлы расположения серверов Content Delivery Network, то задача переадресации запросов становится более трудной. В этом случае требуется интегрированный подход к решению данной проблемы. При интегрированном подходе, который и данная работа посвящена, серверы Content Delivery Network следуют размещать в тех стратегических узлах каждого уровня таким образом, чтобы при репликации Web-контентов в них и перераспределении запросов по серверам общая стоимость доставки Web-контентов конечным потребителям была минимальной. В данной работе для решения этой проблемы предлагается модель, математическое описание, которое опирается на задачу дискретного программирования.

2. ФОРМУЛИРОВКА ЗАДАЧИ И ОБОЗНАЧЕНИЯ

Рассматривается сеть многоуровневой иерархической структурой. Отметим, что такую архитектуру обладает и глобальная компьютерная сеть Internet. Предполагается, что Web-контенты доставляются из узлов верхнего уровня к узлам низких уровней и на каждом уровне между узлами (автономными системами) заключены прямые (одноранговые) соглашения об обмене трафиком. Наша цель состоит в минимизации стоимости доставки Web-контентов конечным потребителям, которая достигается: 1) определением оптимальных точек расположения серверов на каждом уровне; 2) оптимальным размещением контент-реплик по серверам; 3) и оптимальным распределением запросов по серверам. Здесь и далее под сервером понимается сервер Content Delivery Network.

Введем некоторые обозначения:

L – число уровней;

I_k – число узлов в k-м уровне, $k=\overline{1,L}$;

J_k – число потенциальных узлов в k-м уровне для которых должны быть размещены серверы, $k=\overline{1,L}$;

wc_m – часто запрашиваемые Web-контенты узлами всех уровней, $m=\overline{1,M}$;

$N_{i_k m}$ – общее число обращений узла i_k k-го уровня Web-контенту wc_m, $m=\overline{1,M}$, $i_k=\overline{1,I_k}$, $k=\overline{1,L}$;

$X_{i_k j_p m}$ – число обращений узла i_k Web-контенту wc_m, реплицированному в сервере, который расположен в потенциальном узле j_p, $m=\overline{1,M}$, $i_k=\overline{1,I_k}$, $k=\overline{1,L}$, $j_p=\overline{1,J_p}$, $p=\overline{k,L}$;

$N^{+}_{i_k j_p m}$ – верхняя граница числа обращений узла i_k Web-контенту wc_m, реплицированному в сервере, который расположен в потенциальном узле j_p, $m=\overline{1,M}$, $i_k=\overline{1,I_k}$, $k=\overline{1,L}$, $j_p=\overline{1,J_p}$, $p=\overline{k,L}$;

$V_{i_k j_p}$ – объем передаваемой информации из узла j_p в узел i_k, $i_k=\overline{1,I_k}$, $k=\overline{1,L}$, $j_p=\overline{1,J_p}$, $p=\overline{k,L}$;

$C_{i_k j_p}$ – стоимость передачи единичной информации из узла j_p в узел i_k, $i_k=\overline{1,I_k}$, $k=\overline{1,L}$, $j_p=\overline{1,J_p}$, $p=\overline{k,L}$;

S – число серверов, в которых должны быть расположены в потенциальных узлах многоуровневой глобальной сети;

Y_{mj_k} – булево переменное, равное 1, если Web-контент wc_m размещен в сервере, расположенном в потенциальном узле j_k, и равное 0, в противном случае, $m=\overline{1,M}$, $j_k=\overline{1,J_k}$, $k=\overline{1,L}$;

Z_{j_k} – булево переменное, равное 1, если в потенциальном узле j_k расположен сервер, и равное 0, в противном случае, $j_k=\overline{1,J_k}$, $k=\overline{1,L}$;

Cap_{j_k} – способность (capacity) сервера, расположенного в потенциальном узле j_k, $j_k=\overline{1,J_k}$, $k=\overline{1,L}$.

В принятых обозначениях уровень L считается верхним уровнем, т.е. Web-контенты доставляются из узлов верхнего уровня L к узлам нижележащих уровней.

3. *МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ЗАДАЧИ*

Пусть Web-контент wc_m реплицирован в узле j_p k-го уровня. Тогда согласно архитектуры рассматриваемой сети к этому Web-контенту могут обращаться узлы того же уровня и узлы, расположенных в нижележащих уровнях. Предполагается, что один и тот же Web-контент может реплицироваться на каждом уровне. Здесь во избежание от дополнительных расходов, вызванных расположением дополнительных серверов, предполагается, что на каждом уровне Web-контенты могут реплицироваться максимум один раз. Тогда соблюдением минимальности стоимости доставки, задачу распределения запросов узла i_k по серверам, в котором реплицируется Web-контент wc_m, можно сформулировать таким образом:

$$\sum_{p=k}^{L}\sum_{j_p=1}^{J_p}\sum_{m=1}^{M} X_{i_k j_p m} V_{i_k j_p} C_{i_k j_p} Y_{m j_p} \to \min, \tag{1}$$

где $k=\overline{1,L}$, $i_k=\overline{1,I_k}$.

В формуле (1) если произвести суммирование по k и i_k, то получим математическую модель поставленной задачи:

$$\sum_{k=1}^{L}\sum_{i_k=1}^{I_k}\sum_{p=k}^{L}\sum_{j_p=1}^{J_p}\sum_{m=1}^{M} X_{i_k j_p m} V_{i_k j_p} C_{i_k j_p} Y_{m j_p} \to \min. \tag{2}$$

В рамках сделанных предположений и обозначений задача (2) должна решаться при соблюдении некоторых ограничений. Во-первых, должно обеспечиваться то условие, что суммарное число запросов узла i_k адресованных к Web-контенту wc_m, реплицированному в узлах разных уровней, не должно превышать общее число обращений $N_{i_k m}$ узла i_k к Web-контенту wc_m:

$$\sum_{p=k}^{L}\sum_{j_p=1}^{J_p} X_{i_k j_p m} Y_{m j_p} \le N_{i_k m}, \quad \forall i_k, m. \tag{3}$$

Для предотвращения всплеска запросов в серверах налагается ограничение на число запросов узла i_k адресованных к Web-контенту wc_m, реплицированному в сервере, расположенном в потенциальном узле j_p:

$$0 \le X_{i_k j_p m} \le N^{+}_{i_k j_p m}, \quad \forall i_k, j_p, m. \tag{4}$$

Согласно предположению, Web-контенты на каждом уровне могут реплицироваться максимум один раз, т.е. должно удовлетворяться условие:

$$\sum_{j_k=1}^{J_k} Y_{m j_k} \le 1, \quad \forall k. \tag{5}$$

Поскольку способность серверов ограничена, то адресованных к ним число запросов не должно превышать их способностей:

$$\sum_{m=1}^{M}\sum_{q=1}^{p}\sum_{i_q=1}^{I_q} N_{i_q j_p m} Y_{m j_p} \le Z_{j_p} Cap_{j_p}, \quad \forall j_p. \tag{6}$$

Легко видеть, что условие (6) не только налагает ограничение на число запросов, оно также предотвращает адресацию запросов к узлам, в которых не размещены серверы.

Поскольку количества серверов равняется S, то выполнение следующего условия является обязательным:

$$\sum_{k=1}^{L}\sum_{j_k=1}^{J_k} Z_{j_k} = S, \tag{7}$$

Очевидно, что если в узле не расположен сервер, то в нем ни возможно реплицировать Web-контент, а также ясно, что Web-контент не реплицируется ни на каждом сервере, то отсюда следует, что между переменными Y_{mj_k} и Z_{j_k} справедливо соотношение такого вида:

$$Y_{mj_k} \le Z_{j_k}, \quad \forall m, j_k . \tag{8}$$

Наконец, согласно определению переменных Y_{mj_k} и Z_{j_k}:

$$Y_{mj_k} \in \{0,1\}, \quad \forall m, j_k . \tag{9}$$

$$Z_{j_k} \in \{0,1\}, \quad \forall j_k . \tag{10}$$

Итак, мы пришли к задаче дискретного программирования, цель, которая состоит в минимизации суммы (2) при ограничениях (3)-(10).

4. ЗАКЛЮЧЕНИЕ

Несмотря на то, что технология Content Delivery Network существует недавно, его главная цель уже утвердилась – это часто запрашиваемые Web-контенты необходимо переместить на периферию сети, ближе к конечному потребителю, с целью уменьшения загрузки опорной сети, экономии пропускной способности, снижения времени отклика и минимизация стоимости доставки. При этом Web-контенты реплицируются во множество регионов, т.е. рассеивается географически, что существенно уменьшает нагрузку на Web-серверы и, следовательно, повышается устойчивость их работы. С первого взгляда идея Content Delivery Network выглядит очень простой, однако, использование этой технологии для доставки Web-контентов конечным потребителям порождает ряд проблем, среди которых как было отмечено выше, доминирующую роль играют три задачи. Несмотря на то, что в последние годы к решению каждой из этих задач посвящены многочисленные работы, пока некоторые вопросы остаются нерешенными. Данная работа посвящена к решению одной из сложных вопросов, а именно к проблеме распределения запросов по серверам Content Delivery Network многоуровневой иерархической архитектурой, с целью минимизации стоимости доставки конечным потребителям. При иерархической схемы репликации, репликация производится на уровне пользователей, так и на уровне организаций, регионов и стран, т.е. при распределении запросов приходится согласовывать интересы значительного количества узлов каждого уровня. В данной работе в отличие от работ, посвященные к этой проблеме, предлагается интегрированный подход, который распределение запросов согласуется с нахождением оптимальных точек расположения серверов и оптимальности распределения контент-реплик по серверам. Эта проблема сводится к задаче дискретного программирования.

СПИСОК ЛИТЕРАТУРЫ

[1] *Aggarwal C., Wolf J.L., and Yu P.S.* Caching on the World Wide Web //IEEE Transactions on Knowledge and Data Engineering. 1999. Vol. 11. № 1. PP. 94-107.

[2] *Gadde S., Chase J., and Rabinovich M.* Web caching and content distribution: a view from interior //*Computer Communications*. 2001. Vol. 24. № 2. PP. 222-231

[3] *Tang X., Xu J., Chanson S. T.* Web content delivery //Springer-Verlag. 2005. 394 p.

[4] *Plagemann T., Goebel V., Mauthe A., Mathy L., Turletti T. and Urvoy-Keller G.* From content distribution networks to content networks – issues and challenges //Computer Communications. 2006. Vol. 29. № 5. PP. 551-562.

[5] *Bakiras S., Loukopoulos T.* Combining replica placement and caching techniques in content distribution networks //Computer Communications. 2005. Vol. 28. № 9. PP. 1062-1073.

[6] http://www.akamai.com

[7] http://www.digitalisland.com

[8] *Almeida J.M., Eager D.L., Vernon M.K., Wright, S.J.* Minimizing delivery cost in scalable streaming content distribution systems //IEEE Transaction on Multimedia. 2004. Vol. 6. №2. PP.356-365.

[9] *Wang B., Sen S., Adler M., Towsley D.* Optimal proxy cache allocation for efficient streaming media distribution //IEEE Transaction on Multimedia. 2004. Vol. 6. №2. PP. 366-374.

[10] *Ni J., Tsang D.H.K.* Large-scale cooperative caching and application-level multicast in multimedia content delivery networks //IEEE Communications. 2005. Vol. 43. №5. PP. 98-105.

[11] *N. Jian, D. Tsang, I. Yeung, and H. Xiaojun.* Hierarchical content routing in large-scale multimedia content delivery network //Proceedings of the IEEE International Conference on Communications (ICC 2003). Anchorage. Alaska. USA. May 11-15. 2003. Vol. 2. PP. 854-859.

MINIMIZATION OF DELIVERY COST IN CONTENT DELIVERY NETWORKS WITH MULTILEVEL HIERARCHICAL ARCHITECTURE

M. H. Sharifov, PhD candidate
Institute of Information Technology Azerbaijan National Academy of Sciences
9, F.Agayev str., Baku, Azerbaijan, Az1141

Abstract. *In paper the multilevel caching model in Content Delivery Networks, providing minimization of delivery cost of Web-contents to end users is offered. Minimization of delivery cost is reached by integrated approach to the decision of problems: optimum distribution of demands, determination of optimum placement points of servers and optimum distribution of content-replications on servers. The proposed model is reduced to discrete programming problem.*

Keywords: *Content Delivery Network, multilevel caching, minimizing delivery cost, demands distribution, discrete programming.*