TECHNICAL RESEARCH REPORT

Optimal Cache Allocation Policies in Competitive Content Distribution Networks

by Ozgur Ercetin, Leandros Tassiulas

CSHCN TR 2001-3
(ISR TR 2001-4)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Optimal Cache Allocation Policies in Competitive Content Distribution Networks

Özgür Erçetin, Leandros Tassiulas
Institute for Systems Research, Electrical and Computer Engineering Department,
University of Maryland, College Park, MD 20740.

Abstract

Exponential expansion in network dimensionality and user traffic has created substantial traffic congestion in Internet. This congestion causes increased delays perceived by the users while downloading web pages. Users have considerably short patience, and when they do not start receiving information in a short while, they stop browsing the requested web page. As the commercial value of Internet has become prevalent, the importance of keeping users at a site started to have direct translation into business value. Proxy caching can alleviate problems caused by the increased user traffic. In this paper, we consider the effects of real-world non-cooperative behavior of the network agents (servers and proxy caches) in overall network performance. Specifically, we consider a system where the proxy caches sell their caching space to the servers, and servers invest in these caches to provide lower latency to their users to keep them browsing their web pages and in turn to increase their revenues. We determine optimal strategies of the agents that maximize their benefits. We show that such a system has an equilibrium point when no agent can increase it's benefit by unilaterally updating it's strategy. We show that under certain conditions this equilibrium leads to optimal cache allocation. We also show that an algorithm derived from this analysis is superior to the currently implemented caching algorithms.

1 Introduction

The Internet has expanded exponentially in network and user community size with the introduction of the World Wide Web (WWW). With this expansion average user delay has increased due to the increased traffic over the Internet. The traditional Internet model of browsers connecting to content servers (web sites) over the Internet is the root of the problems in Internet. Although this simple model of users reaching distant servers for information has brought the success of WWW, it has also become a limitation for it's development.

WWW pages have become complicated with larger size embedded objects and the users have come to expect interactive applications. This has put substantial burden on the underlying communications infrastructure which could not be improved as fast as the rate of increase in traffic demand. The business over Internet has thrived with many large and small companies making their presence on the Internet. The business of companies became global and continuous with potential customers from all over the world with transactions 24 hours a day. The performance of web sites in terms of average user latency has started to translate directly into business value.

In order to alleviate the problems of traffic congestion, i.e. to reduce high user latencies, user and proxy caches are deployed. The user and proxy caches store frequently requested information, so that the user requests do not always have to be served by the original server, which is in many occasions already overloaded and located far from the users. User caches are implemented at the clients and store the most recently accessed data by the users. Most modern browsers employ this type of caches. The proxy caches (or simply proxies) are located at the network edges. Proxies can be implemented by enterprises or Internet service providers to reduce the user latency and/or to reduce the total traffic flowing in/out of the network. Original web sites may also choose to implement proxies by storing their information partially or as a whole at the proxy sites, so that the users can access their information more rapidly.

The serious congestion encountered in Internet has resulted in the emergence of a new type of business of *content delivery/distribution*. The content delivery companies provide caching and repli-

cation services to the web sites. Among a few of those content delivery companies are Akamai [5], InfoLibria [9], Mirror Image [6] and Teleglobe [8]. The promise of the content delivery/distribution companies is to allow the web sites the ability to reach users with much less congestion and latency without large investment in mirror sites. This is possible by many geographically dispersed servers installed by the content delivery companies all over the globe. These servers selectively cache documents from the host web sites that are used to serve user requests locally.

An additional benefit of this type of system is the control of the intellectual property rights. The servers may set their content as un-cacheable for the rest of the Internet, while explicitly pushing their content to these trusted content delivery companies.

In the literature, the metric that is used for measuring the performance of the caching systems is the average latency observed by the users [1],[2]. This metric serves well if the requested object has no alternatives in other servers and thus users have no incentive of ceasing to download the object from one server and switching to another. Obviously this is not true in Internet. In the 1960s, IBM performed a study to determine how long a computer user would wait for an application screen to refresh before becoming impatient. The answer was about two seconds. Recent studies [3] have shown that as the latency increases users stop browsing the requested page (*bail-out*) with increasing probability (Figure 1). Since the Internet becomes more and more commercially oriented, the bail-out rate started to have a direct economic impact for the web sites. It is of interest of web sites to have a fast rate of delivery of the objects, so they do not lose customers.

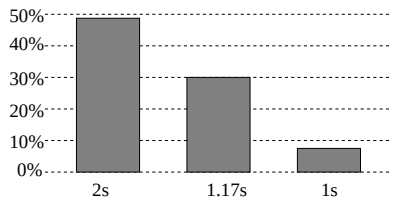

Figure 1: Bail-out rates for various download times [3].

The web sites (servers) can be considered as content providers. They make their revenue either by selling some information content (such as news, maps, etc.) or by selling tangible products. For

the first type of servers, the longer the user remains on the site higher the revenue will be. For the second type of servers, it is shown by marketing studies that when the users stay at a web-site longer, they are more likely to make a purchase.

The necessary (but of course not sufficient) condition to keep the users browsing the web site is the timely delivery of the content. However, one can easily see from Figure 1 that the user bail-out rate have nonlinear relationship with the retrieval time. Thus, from the content providers' view the important metric is the minimization of the lost revenue rather than the retrieval time.

In this paper, we consider a realistic model for the relationship between the content providers and the content distribution networks. The servers disseminate some of their content to the content distribution networks for improved user latency. Meanwhile, content distribution networks charge the servers for the amount of space they allocate in their proxy cache servers. However, there are multiple content distribution networks competing with each other. We investigate the effect of this competition on the system. Specifically, we show that such a price competition leads to an equilibrium, which under certain conditions, leads to the optimal cache allocation strategy for the servers. This approach provides a dynamical and distributed algorithm for determining the optimal (or near-optimal) cache content in the network.

The paper is organized as follows. In the next section, we provide the details of the model of our system. In the third section, we determine the optimal server information dissemination strategy, and in the following section we discuss the optimal proxy pricing strategy. Since the proxies compete among each other for the business of the servers, they update their pricies according to their competitors' strategies. In section 5, we investigate the outcome of this competition. We define the server-proxy game and show that this game has a Nash equilibrium solution. We determine certain conditions under which this equilibrium leads to the globally optimal cache allocation. In section 6, we describe a distributed pricing algorithm based on this analysis and show that the performance of this algorithm is better than the currently implemented caching algorithms.

2 System Model

Figure 2 illustrates the network set-up that we are interested in this paper. Consider a network where the users are always two hops away from the servers. Proxy caches are located between the users and the servers. Each user belongs to a network that is served by an unique proxy cache. User request first arrives to the proxy cache of the user's network. If the requested object is available at the proxy, the request is immediately served. Otherwise the request is forwarded to the main server that the object originally resides in.

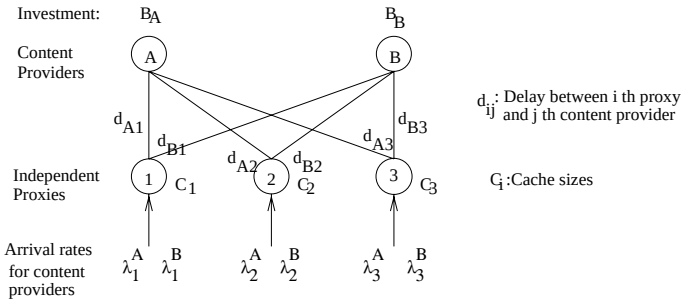

Figure 2: Content delivery system.

In order to reduce the expected user latency and thus to reduce the number of users bail-out and in turn to increase the revenue, the content providers may invest in the content delivery companies for the replication of some of their content to the sites that are closer to the users. We assume that the content delivery companies build their proxies at the network edges, and each proxy serve the users of that particular network. The proxies have limited sized caches which are shared among several servers. They charge servers according to the space that is allocated to each of the server.

The request arrival rates from each network to the content providers are independent. Proxies compete with each other for the business of the content providers. They announce their price of unit caching space to the content providers. The objective of the proxy is to maximize it's revenue by selling it's caching space at the maximum possible price. Servers have fixed initial investment that they spend completely for the purchase of the caching space from the proxies. The objective of the content provider is to maximize the *expected* revenue generated from a user request by determining

the appropriate cache allocation for each proxy. The server's decision depends on their relative proximity of the proxies, the user request arrival rates from each network and the prices of the unit cache spaces.

Assume that there are I different servers and J different proxy caches present in the network. Let λ_i^j denote the total request arrival rate at proxy j for the content in the ith content provider. The requests are distributed according to Zipf distribution [10]. That is, given that a request has arrived, the probability that the request is for object h is $q(h) = \frac{c}{h^{\alpha_i}}$. c is the normalization constant, and $0 < \alpha_i < 1$ is the distribution characteristic of the content provider i. The characterization of user request distribution as Zipf distribution is discussed in previous studies [11], [12] and is accepted as a good approximation to the actual web traffic behavior.

The average delay (propagation and transmission) between the content provider i and the proxy j is d_{ij}. The impact of bail-out phenomenon due to the observed user latency on the server revenues can be effectively modeled by a system where content providers charge their users according to the retrieval time of the objects. In this model, as the retrieval time of an object increases, the charge for the content the corresponding user pays decreases. The content provider i charges a user of proxy j $w(0)$, if the user's request can be satisfied at the proxy, and charges $w(d_{ij})$, otherwise. The pricing function $w(d)$ is determined by the server so that the expected revenue generated when the users bail-out with respect to some probability distribution when the observed delay is d, is equivalent to the expected revenue generated when the users are charged according to the retrieval time. We acknowledge that such a delay-dependent pricing strategy is not realistic. We consider this pricing strategy only as a simple mathematical reduction to the real pricing model. Even though these strategies may give different results on per request basis, on the average the total revenue generated in either case will be equal, if the pricing function $w(d)$ is determined as stated above. Notice that $w(d_{ij}) < w(0)$. From now on we assume that $w(d)$ is available from the bail-out rate, which is in turn determined from the marketing studies.

We can identify two types of optimization problems in this model: Servers' revenue maxi-

mization and the proxies' revenue maximization. We first determine the server's optimal caching strategy under a certain proxy pricing scheme.

3 Optimal Server Information Dissemination Strategy

Let B_i^j be the investment of the ith content provider in the jth proxy. Let $\mathcal{B}_i = \sum_j B_i^j$ be the total investment of the ith content provider. It is assumed that the information stored in the servers is continuous and can be replicated continuously to a proxy. The total information available at the server i is T_i. The server replicates its most popular part of the content to the proxies so that the average hit rate increases. Assuming that C_i units of cache space is allocated to the server, the probability that an incoming request is satisfied at the proxy is given by $\left(\frac{C_i}{T_i}\right)^{1-\alpha_i}$. Notice that in arriving this result we assumed that $q(0) = 0$.

Let p_j denote the price of the unit cache space in proxy j. Let the pricing policy, $\mathbf{p} = (p_1, p_2, \ldots, p_J)$, denote the set of unit cache space prices of all the proxies in the network. Let x_i^j be the cache space allocated to server i in proxy j. If i th server's investment in the j th proxy is B_i^j, then the total cache space allocated to server i in proxy j is $x_i^j = \frac{B_i^j}{p_j}$. The average revenue that server i generates by B_i^j investment in proxy j is $\lambda_i^j[w(0) - w(d_{ij})]\left(\frac{B_i^j}{p_j T_i}\right)^{1-\alpha_i}$. Define $\beta_i^j = \lambda_i^j[w(0) - w(d_{ij})]/T_i^{1-\alpha_i}$ as the gain factor for server i from proxy j.

The utility function, i.e. the total average revenue, $U_i(x_i)$, of the i th server is $U_i(x_i) = \sum_{j=1}^{J} \beta_i^j \left(x_i^j\right)^{1-\alpha_i}$. For a given pricing policy $\mathbf{p}$ the server optimization problem (S) can be given as:

$$\text{(S)} \quad \max_{\{x_i^j\}_{j=1}^J} \quad U_i(x_i)$$

$$\text{subject to} \quad \sum_{j=1}^{J} x_i^j p_j \leq \mathcal{B}_i.$$

Since $U_i(x_i)$ is a concave function, and the constraint set is compact, there exists a unique solution to (S).

Lemma 1 $x_i^{j^*} = \dfrac{\left(\frac{\beta_i^j}{p_j}\right)^{1/\alpha_i} \mathcal{B}_i}{\sum_{k=1}^{J} p_k \left(\frac{\beta_i^k}{p_k}\right)^{1/\alpha_i}}$ *is the unique optimal solution to the optimization problem (S).*

Proof Consider the Lagrangian function $L(\mathbf{x})$,

$$L(\mathbf{x}) = \sum_{j=1}^{J} \beta_i^j \left(x_i^j\right)^{1-\alpha_i} - \gamma \left(\sum_{j=1}^{J} x_i^j p_j\right),$$

where γ is the Lagrangian constant. From Karush-Kuhn-Tucker Theorem we know that the optimal solution is given by $\partial L(\mathbf{x})/\partial x_i = 0$ for $\gamma \geq 0$.

$$\partial L(\mathbf{x})/\partial x_i = \beta_i^j (1-\alpha_i) \left(x_i^j\right)^{-\alpha_i} - \gamma p_j = 0,$$

$$x_i^j = \left(\frac{\beta_i^j (1-\alpha_i)}{\gamma p_j}\right)^{1/\alpha_i}.$$

Using this result in the constraint equation, we can determine γ as

$$\gamma^{-1/\alpha_i} = \frac{\mathcal{B}_i}{\sum_{k=1}^{J} p_j \left(\frac{\beta_i^j (1-\alpha_i)}{p_j}\right)^{1/\alpha_i}}.$$

Now, optimal x_i^j can easily be determined. ■

This result has been analyzed for a two servers, three proxies system, where the total investment of each server is 20 cost units, and the sizes of the caches of each proxy is the same at 10 storage units. Figure 3 depicts the investment of the server in a proxy when a proxy's unit cache space price is varied, while the prices of the remaining two proxies' are kept the same. The arrival rates to each proxy and the gain factors of each server-proxy pair are the same. The analysis does not take into account the limited cache capacities of the proxies. The investment in the proxy decreases

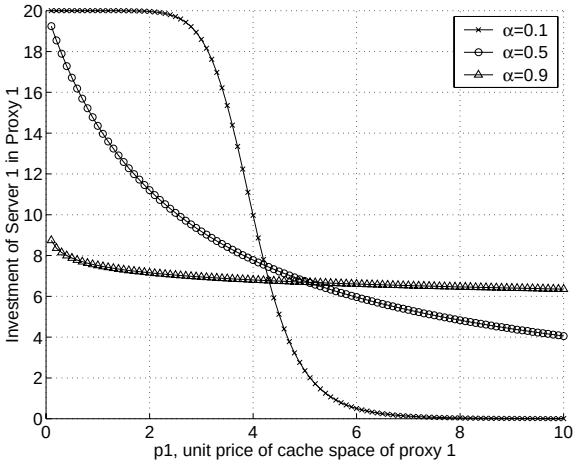

Figure 3: Elasticity of server investment when the cache size is infinite. $\alpha_2 = 0.5$. $\beta_i^j = 1, \forall i, j$. $p_2 = 4$, and $p_3 = 7$.

with the increasing price. However, more importantly the investment is quite dependent on the distribution of requests for the server's content. In fact $\alpha = 1$ represents a special case, where the server's investment in a proxy is the same regardless of the price of the proxy.

In Figure 4, the variation of total revenue generated by a proxy server for varying proxy prices is depicted. In this case, as the proxy lowers its price, it receives higher investment from the servers. However, lowering the price more than a certain price reduces the revenue, because the proxy has a limited cache space, and the request for more space from the servers can not be satisfied.

4 Optimal Proxy Pricing Strategy

We now consider the optimal pricing strategies for the proxies that maximize the proxies' revenues. Let $\mathbf{p}^{-j} = (p_1, p_2, \ldots, p_{j-1}, p_{j+1}, \ldots, p_J)$ be the set of unit cache space prices of all the proxies in the network except the jth one. We assume that there is no collaboration among the proxies, and each proxy tries to maximize their revenue non-cooperatively.

Lemma 2 *Proxy j's best pricing strategy under a given fixed pricing policy $\mathbf{p}^{-j}$ is to set a price p_j that satisfies $\sum_{i=1}^{I} x_i^j(p_j) = C_j$, i.e. when the proxy cache space is completely allocated.*

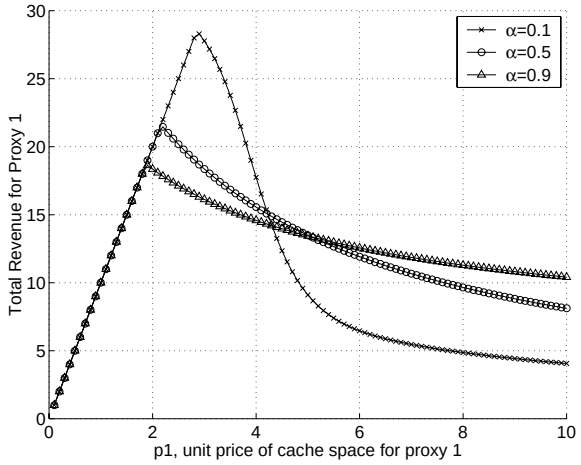

Figure 4: Total revenue of proxy for varying prices. $\alpha_2 = 0.5$. $\beta_i^j = 1, \forall i, j$. $p_2 = 4$, $p_3 = 7$.

Proof As illustrated in Figure 4, the proxy's revenue decreases when the price is either increased or decreased beyond a certain price. Let $r_j(p_j)$ denote the revenue of proxy j.

$$r_j(p_j) = \begin{cases} \sum_i x_i^j(p_j) p_j & \text{for } \sum_i x_i^j(p_j) \leq C_j \\ C_j p_j & \text{for } \sum_i x_i^j(p_j) > C_j \end{cases} .$$

The optimal point lies either at the boundaries or at the irregularity. If $\sum_i x_i^j(p_j) \leq C_j$,

$$\begin{aligned} \partial r_j / \partial p_j &= \mathcal{B}_i \frac{(\beta_i^j)^{1/\alpha_i}(-1/\alpha_i) p_j^{-1-1/\alpha_i} \sum_k p_k^{1-1/\alpha_i} (\beta_i^k)^{1/\alpha_i} - (\beta_i^j)^{1/\alpha_i} p_j^{-1/\alpha_i} (1 - 1/\alpha_i)(\beta_i^j)^{1/\alpha_i} p_j^{-1/\alpha_i}}{\left(\sum_k p_k^{1-1/\alpha_i} (\beta_i^k)^{1/\alpha_i} \right)^2} \\ &= \mathcal{B}_i \frac{-(\beta_i^j)^{2/\alpha_i} p_j^{-2/\alpha_i} - 1/\alpha_i (\beta_i^j)^{1/\alpha_i} p_j^{-1-1/\alpha_i} \sum_{k \neq j} p_k^{1-1/\alpha_i} (\beta_i^k)^{1/\alpha_i}}{\left(\sum_k p_k^{1-1/\alpha_i} (\beta_i^k)^{1/\alpha_i} \right)^2} \\ &< 0. \end{aligned}$$

Thus, $r_j(p_j)$ is monotonically decreasing for p_j such that $\sum_i x_i^j(p_j) \leq C_j$.

If $\sum_i x_i^j(p_j) > C_j$, then it is clear that $r_j(p_j)$ is maximized at the boundary when $\sum_i x_i^j(p_j) = C_j$. This concludes the proof. ■

Lemma 2 suggests that assuming that the pricing policies of other proxies are given and fixed, the best pricing policy of the proxy is to set a price at which the demand for caching space equals

to the supply.

5 Server-Proxy Game

Until now, we discussed the optimal strategies of the servers and the proxies given that system is at a steady state. However, we have not discussed whether such a steady state exists. Notice that when a proxy re-evaluates it's pricing policy according to the pricing policies of the rival proxies, the remaining proxies will do the same. At each different pricing policy the servers' optimal investments will be different as well.

In order to understand the behavior of the proxies, we model the two-stage proxy-server system as a non-cooperative game [13]. In this *server-proxy game*, , (J, S, P), the players, J, are the proxies, the strategy set S_j for a proxy j is given by the proxy's unit cache space price and the payoff function $P_j(s)$ of each proxy j is given by the profit of the jth proxy. This system is similar to the Cournot oligopoly discussed in the economics literature. Assume that each proxy has a fixed cost for it's cache, but has no control over the size of the cache, i.e. the size of the cache is determined before the system implementation.

We first show that this game has a Nash equilibrium solution, where no proxy has incentive to change unilaterally it's strategy, since each proxy maximizes it's own individual payoff given the strategies of others.

Theorem 1 *The non-cooperative proxy-server game , (J, S, P) has at least one Nash Equilibrium solution.*

Proof We first show that the strategy sets are convex and compact. The profit for proxy j is $r_j(\mathbf{p}) - c_j$, where c_j is the cost of the proxy j's cache. We will assume that there exists some price $\hat{p}_j$ at which demand for the cache space of proxy j is zero regardless of the prices of other proxies. Considering the revenue curve $r_j(\mathbf{p})$, this is not a restricting assumption. In Figure 4, the revenue of proxy j increases until a certain price p_j^* beyond which it starts to decrease again. Then, we

may limit the strategy set S_j to the interval $[0, \hat{p}_j]$, and still be able to cover the complete range of payoff function. Thus, the strategy set S_j is convex and compact.

The profit of each proxy is bounded from below by zero and since the total investment of all servers is limited, the profit can never exceed $\sum_i \mathcal{B}_i - c_j$. We assume that proxy takes its' rivals actions as given, supposes they will remain constant, and chooses its own best course of action accordingly. This assumption is called *Cournot behavioral assumption*. The payoff function under this assumption is given by $r_j(\mathbf{p})$. In Lemma 2, we have shown that there is an unique *best reply function*, $R_j(\mathbf{p}) = \text{argmax}_{p_j}\{r_j(\mathbf{p})\}$ for proxy j, which is also continuous. Define a mapping $\mathbf{R}(\mathbf{s}) = (R_1(\mathbf{s}), \ldots, R_J(\mathbf{s}))$. By Brouwer's Theorem [13] $\mathbf{R}$ must have at least one fixed point $\mathbf{s}^* \in S$, where $\mathbf{s}^* = \mathbf{R}(\mathbf{s}^*)$. The definition of the best reply function $R_j(\mathbf{s})$ and Brouwer's Theorem tell us that $P_j(\mathbf{s}^*) \geq P_j(\mathbf{s}^*/t_j)$ for all $t_j \in S_j$ and $j = 1, \ldots, J$, where $\mathbf{s}^*/t_j$ is the strategy set when the jth proxy's strategy is changed to t_j in the complete strategy set $\mathbf{s}^*$. This result is the definition of Nash equilibrium. ■

We have shown that there exists a set of equilibrium prices for such a system. The question that remains to be addressed is what physical interpretation such an equilibrium has.

Consider the server optimization problem (S) discussed in the previous section. In our system, every server tries to maximize their own revenue regardless of others, subject to the availability of funds and caching space. The optimization problem of each server is related to each other with constraint $\sum_i x_i^j \leq C_j$ for all proxies, i.e. servers compete for the available cache resources. Thus, we can reiterate the optimization problem for individual servers as:

$$(S_i) \quad \max_{\{x_i^j\}_{j=1}^J} \quad U_i(x_i)$$

$$\text{subject to} \qquad (1) \qquad \sum_{j=1}^{J} x_i^j p_j \leq \mathcal{B}_i$$

$$(2) \qquad \sum_i x_i^j \leq C_j, \;\; j = 1, \ldots, J.$$

Theorem 2 *When there is an unique equilibrium to the server-proxy game, , (J, S, P), the equilibrium prices solve the optimization problem S_i for all servers $i = 1, \ldots, I$.*

Proof Assume that each proxy uses the best reply function $R_j(\mathbf{p})$ to update it's price. The best price for proxy j given the pricing policy $\mathbf{p}^{-j}$ is calculated from $\sum_i x_i^j = C_j$. At the equilibrium this condition is satisfied as well. Furthermore, the servers calculate x_i^j as given by Lemma 1, which guarantees local optimality of the solution and the feasibility of the first condition in S_i. Uniqueness of the equilibrium guarantees that the feasible solution is also the optimal solution. Under these conditions, the outcome of server-proxy game is the solution of $S_i, \forall i = 1, \ldots, I$. ■

Theorem 2 states that the outcome of the non-cooperative game is the optimal solution to the revenue maximization problem of the individual servers. This result is very important, since it shows that a distributed resource allocation algorithm relying on this game leads to optimal solution. Unfortunately, often there are multiple Nash equilibria and depending on the initial prices as well as price update strategies, we may not always get the optimal solution to server optimization problem as an outcome to our game. In the following, we discuss a special case of the proxy cache allocation problem, where the delay between each server-proxy pair is the same and user request arrival rates to each proxy and Zipf distributions for each server are identical. For this case, we determine the condition for which unique equilibrium exists.

It is easy to see that if the best-reply mapping $\mathbf{R}(\mathbf{p})$ is a *Contraction mapping*, then the equilibrium is unique [14]. A mapping $T(p)$ is called contraction mapping, if $|T(p) - T(q)| \leq \lambda |p - q|$ for $\lambda < 1$ or if the mapping is differentiable $\partial T(p)/\partial p < 1$.

Lemma 3 *When $\beta_i^j = \beta$, $\forall i, j$ and $\alpha_i = \alpha$, $\forall i$, then the best reply function $R_j(\mathbf{p})$ is a contraction mapping if the price vector $\mathbf{p}$ is limited to the region given by*

$$\frac{(1-\alpha)\sum_i \mathcal{B}_i / C_j p_l^{-1/\alpha}}{\left(\sum_{k \neq j} p_k^{1-1/\alpha}\right)^{1+\alpha}} < 1$$

Proof

$$\sum_i \frac{{p_j}^{-1/\alpha}}{\sum_k p_k^{1-1/\alpha}} \mathcal{B}_i = C_j$$

$${p_j}^{-1/\alpha}\left(1 - \frac{C_j}{\sum_i \mathcal{B}_i} p_j\right) = \frac{C_j}{\sum_i \mathcal{B}_i} \sum_{k \neq j} p_k^{1-1/\alpha} \tag{1}$$

Let $\gamma_j = \sum_i \mathcal{B}_i / C_j$. Taking the derivative of eq. (1) with respect to p_l, we determine $\partial p_j / \partial p_l$.

$$-\frac{1}{\alpha p_j} p_j^{-1/\alpha} \frac{\partial p_j}{\partial p_l}(1 - p_j/\gamma_j) - \frac{1}{\gamma_j} p_j^{-1/\alpha} \frac{\partial p_j}{\partial p_l} = \frac{1}{\gamma_j}(1 - 1/\alpha) p_l^{-1/\alpha}$$

$$\frac{\partial p_j}{\partial p_l} = \frac{{p_l}^{-1/\alpha}}{p_j^{-1/\alpha}\left(\frac{\gamma_j}{(1-\alpha)p_j} - 1\right)} = \frac{(1-\alpha)\frac{p_j}{\gamma_j} {p_l}^{-1/\alpha}}{p_j^{-1/\alpha}\left(1 - (1-\alpha)\frac{p_j}{\gamma_j}\right)} \tag{2}$$

Notice that the denominator in eq. (2) is similar to the left hand side of eq. (1). Remember that $0 < \alpha < 1$. Hence, $p_j^{-1/\alpha}\left(1 - (1-\alpha)\frac{p_j}{\gamma_j}\right) > p_j^{-1/\alpha}\left(1 - \frac{p_j}{\gamma_j}\right)$. Then,

$$\frac{\partial p_j}{\partial p_l} < \frac{(1-\alpha)\frac{p_j}{\gamma_j} {p_l}^{-1/\alpha}}{p_j^{-1/\alpha}\left(1 - (1-\alpha)\frac{p_j}{\gamma_j}\right)}$$

$$= \frac{(1-\alpha) p_j {p_l}^{-1/\alpha}}{\sum_{k \neq j} p_k^{1-1/\alpha}}$$

Also notice that $p_j < \gamma_j$. Then $0 < 1 - p_j/\gamma_j < 1$. From eq. (1)

$$p_j^{1/\alpha} = \frac{1 - p_j/\gamma_j}{\frac{1}{\gamma_j}\sum_{k \neq j} p_k^{1-1/\alpha}}$$

$$< \frac{1}{\frac{1}{\gamma_j}\sum_{k \neq j} p_k^{1-1/\alpha}}$$

$$p_j < \frac{\gamma_j^\alpha}{\left(\sum_{k \neq j} p_k^{1-1/\alpha}\right)^\alpha}$$

Thus, the best response function $R_j(\mathbf{p})$ is a contraction mapping if

$$\frac{\partial p_j}{\partial p_l} < \frac{(1-\alpha)\gamma_j^{\alpha} p_l^{-1/\alpha}}{\left(\sum_{k \neq j} p_k^{1-1/\alpha}\right)^{1+\alpha}} < 1$$

■

Notice that the condition given in Lemma 3 is not a necessary but a sufficient condition which is probably more restrictive than the necessary condition. Let $\mathcal{R}_j$ be the region given by the above Lemma. Following Corollary gives the condition for optimality of the outcome of the game for the identical case.

Corollary 1 *If the range of the price vector,* $\mathbf{p}$*, is in the region* $\cap_{j=1}^{J} \mathcal{R}_j$*, the server-proxy game will have an unique equilibrium.*

6 Numerical Analysis

The results given in the previous sections suggest that we may use a price-directed market-based distributed algorithm for solving the two-stage server-proxy cache resource allocation problem. We consider the following algorithm for this purpose:

1. Proxy caches announce a set of initial prices $\mathbf{p}^{(0)} = (p_1^{(0)}, p_2^{(0)}, \ldots, p_J^{(0)})$.

2. At iteration k, each server i calculates it's optimal cache demand for proxy j, $x_i^{j\,(k)}$ as given in Lemma 1. Forward these demands to the proxies.

3. At iteration k, each proxy j updates it's price according to the server demands. If the total demand $\sum_i x_i^{j\,(k)}$ is greater than the cache capacity C_j, then the new price $p_j^{(k)}$ is increased by Δ from $p_j^{(k-1)}$, otherwise it is decreased by Δ. Announce the new prices $\mathbf{p}^{(k)}$ to the servers.

4. If total demand $\sum_i x_i^{j\,(k)} = C_j, \forall j$, then stop, otherwise repeat from Step 2.

In this model, the system operates as follows: An initial set of prices is announced to the servers. The servers determine their resource (cache) demands according to these prices as well as the request rates, and the observed delays from the proxies. The servers request these resources from the proxies. Prices are then iteratively changed to accommodate the demands for resources until the total demand equals to the total amount of resource available.

In this algorithm we use a very simple price update policy, where we change the price by a fixed factor, Δ, at each iteration. We acknowledge that more sophisticated price update strategies can be developed, which may converge to the equilibrium faster. However, we do not pursue the design of such more sophisticated price update policies, since our objective in this paper is to demonstrate the existence of such algorithms, as well as the benefits of using them.

We compare the outcome of our algorithm with current caching systems that store the most popular data in their cache. We model the current system for our purposes as follows: Proxy j allocates $\frac{\frac{\lambda_i^j}{\sum_k \lambda_i^k}}{\sum_k \frac{\lambda_k^j}{\sum_l \lambda_k^l}}$ portion of the caching space to server i information. Notice that, in fact this algorithm is better than the current implementation, since it considers the importance of a particular proxy for a server. That is, if the requests of server i are mainly arriving from the network serviced by the proxy j, proxy j gives more caching space to server i than the rest of servers.

We compare the performance of the game-theoretical and conventional caching algorithms according to the total server revenues. We again consider the two server and three proxy system illustrated in Figure 2. We compare the performances of two methods when the skewness of the system is increased. For this purpose, we set several parameters fixed while varying the others. Specifically, we consider the case when one of the servers receives more benefit from one of the proxies while the other server receives more benefit from another proxy. We expect each of the methods to find the appropriate allocation that maximizes the server benefits.

In this analysis we assume that the total investment of each server and the cache sizes of all proxies are the same. Let $\zeta_i^j = [w(0) - w(d_{ij}]/T_i^{1-\alpha_i}$. Notice that $\beta_i^j = \lambda_i^j \zeta_i^j$.

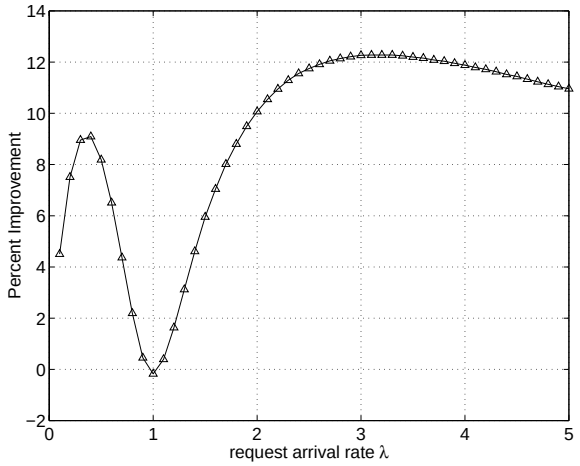

Figure 5: The percentage of improvement of using the game-theoretic algorithm over the conventional caching methods. $\alpha_1 = \alpha_2 = 0.1$. $\zeta_i^j = 1, \forall i, j$. $\lambda_1^1 = \lambda_2^3 = \lambda$. $\lambda_i^j = 1, \forall (i,j) \neq (1,1), (2,3)$.

Figure 5 depicts the improvement of game-theoretic algorithm over the conventional caching solution. In this figure we compare the two algorithms for varying request arrival rates. When the request arrival rates are equal to 1, then the solution given by the game-theoretical algorithm and the conventional algorithm is the same resulting in no improvement. However, as the arrival rates become less or more than 1, we observe that game-theoretical algorithm gives better performance.

In Figures 6 and 7 we consider the performance improvement when the arrival rates are fixed, but ζ is varied. From the definition of ζ_i^j one can notice that by varying ζ_i^j, basically we change the delay between proxy j and server i. As illustrated in Figure 6, as the skewness of the system increases the performance of game-theoretic algorithm gets better compared to the conventional algorithm. Figure 7 also depicts the performance when the server request characteristic is varied. That is, for the ith server the α_i parameter in the Zipf distribution is changed. We observe that for larger values of α_i the improvement of the game-theoretic algorithm is smaller. This is reasonable considering that when $\alpha_i = 1$, the investment of a server is independent of the prices. In that case, the game-theoretic algorithm's solution reduces to the conventional algorithm's solution.

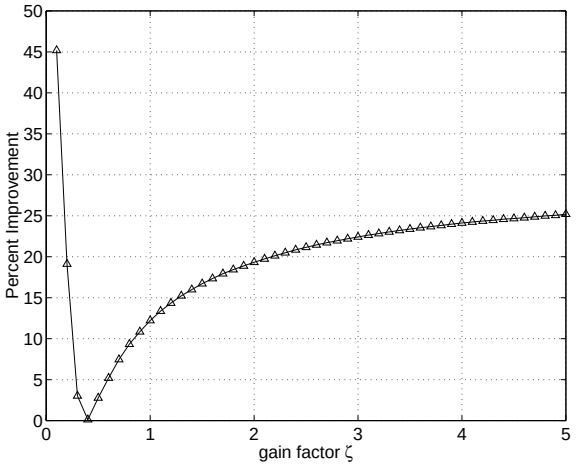

Figure 6: The percentage of improvement of using the game-theoretic algorithm over the conventional caching methods. $\alpha_1 = \alpha_2 = 0.1$. $\zeta_1^1 = \zeta_2^3 = \zeta$. $\zeta_i^j = 1, \forall (i,j) \neq (1,1), (2,3)$. $\lambda_1^1 = \lambda_2^3 = 3$. $\lambda_i^j = 1, \forall (i,j) \neq (1,1), (2,3)$.

We also noticed that game-theoretic algorithm is worse than the conventional method for some values of ζ. We believe this is due to the imprecision of the simulation algorithm that we have to accept for achieving reasonable simulation running times.

7 Conclusions and Future Work

In this paper we analyzed a two-stage server-proxy resource allocation system by a market-based approach. In this analysis, we have shown that the server-proxy game that models the system leads to an equilibrium. And under certain conditions we have shown that this equilibrium is the optimal solution for the non-cooperative resource allocation problem. The importance of our model is that it closely resembles the real-world situation, where the servers and users will not collaborate to achieve the system optimal solution. Instead every agent in the system will try to maximize their benefits without consideration of others. We have also shown that the competition among proxy caches leads to a solution that is better than the solution provided conventional caching methods.

We are going to consider the extension of this model for the multi-stage case. We will also

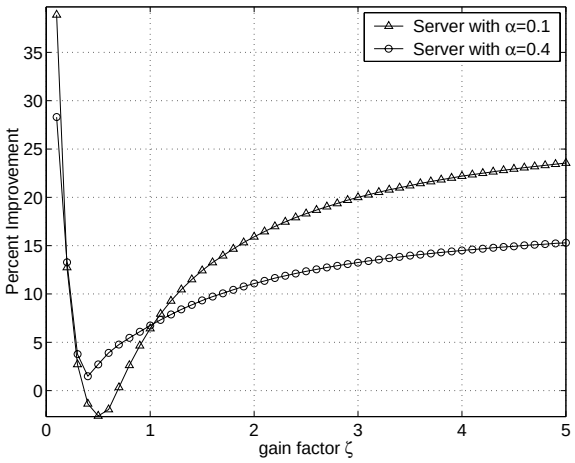

Figure 7: The percentage of improvement of using the game-theoretic algorithm over the conventional caching methods. $\alpha_1 = 0.1$, $\alpha_2 = 0.4$. $\zeta_1^1 = \zeta_2^3 = \zeta$. $\zeta_i^j = 1, \forall (i,j) \neq (1,1), (2,3)$. $\lambda_1^1 = \lambda_2^3 = 3$. $\lambda_i^j = 1, \forall (i,j) \neq (1,1), (2,3)$.

examine the implications of replaceable objects. That is, when there is an equally substitutable object in another server. In that case, when the user bails-out, it will direct its request to another server.

References

[1] B. Li, X. Deng, M. Golin, and K. Sohraby, "On the optimal placement of web proxies in the Internet", *INFOCOM'99*, pp 1282-1290

[2] J. F. Kurose and R. Simha, "A Microeconomic Approach to Optimal Resource Allocation in Distributed Computer Systems," *IEEE Trans. on Computers*, 38(5):705-717, 1989.

[3] Zona Research, "Economic Impact of Unacceptable Web Site Download Speeds," http://www.zonaresearch.com/deliverables/white_papers/wp17/index.htm

[4] P. B. Danzig, R. S. Hall, and M. F. Schwartz, "A Case for caching file objects inside internetworks." *Proceedings of SIGCOMM'93*, pp 239-248, 1993.

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

[6] http://www.mirror-image.com

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

[8] http://www.teleglobe.com

[9] http://www.infolibria.com

[10] G. K. Zipf *Human Behavior and the Principle of Least Effort*, Addison-Wesley Press, Cambridge MA, 1949.

[11] Lee Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," *Proceedings of INFOCOM'99.*

[12] T. P. Kelly, S. Jarmin and J. K. MacKie-Mason, "Variable QoS from Shared Web Caches: User-Centered Design and Value Sensitive Replacement," *MIT workshop on Internet Service Quality Economics*, Dec'1999.

[13] G. Alexander Jehle, *Advanced Microeconomic Theory*, Prentice-Hall, 1991.

[14] N. Van Long and A. Soubeyran, "Existence and uniqueness of Cournot Equilibrium: A contraction mapping approach," *Economics Letters*, vol 67 (2000), pp 345-348, Elsevier Publishing.