December 2002

# Optimal Pricing and Capacity Allocation in Vertically Differentiated Web Caching Services

Kartik Hosanager
*Carnegie Mellon University*

Ramayya Krishnan
*Carnegie Mellon University*

John Chuang
*University of California, Berkeley*

Viyanand Choudhary
*Carnegie Mellon University*

# OPTIMAL PRICING AND CAPACITY ALLOCATION IN VERTICALLY DIFFERENTIATED WEB CACHING SERVICES

**Kartik Hosanagar**
Carnegie Mellon University
Pittsburgh, PA USA
kartikh@andrew.cmu.edu

**Ramayya Krishnan**
Carnegie Mellon University
Pittsburgh, PA USA
rk2x@andrew.cmu.edu

**John Chuang**
University of California, Berkeley
Berkeley, CA USA
chuang@sims.berkeley.edu

**Vidyanand Choudhary**
University of California, Irvine
Irvine, CA USA
veecee@uci.edu

## Abstract

*Internet infrastructure is a key enabler of e-business. The infrastructure consists of backbone networks (such as UUNET and AT&T), access networks (such as AOL and Earthlink), content delivery networks (CDNs, such as Akamai) and other caching service providers. Together, all of the players make up the digital supply chain for information goods. Caches provisioned by CDNs and other entities are the storage centers, the digital equivalent of warehouses. These caches store and deliver information from the edge of the network and serve to stabilize and add efficiency to content delivery. While the benefits of caching to content providers with regard to scaling content delivery globally, reducing bandwidth costs and response times are well recognized, caching has not become pervasive. This is largely due to misaligned incentives in the delivery chain. Much of the work done to date on Web caching has focused on the technology to provision quality of service and has not dealt with issues of fundamental importance to the business of provisioning caching services, specifically, the design of incentive compatible services, appropriate pricing schemes, and associated resource allocation issues that arise in operating a caching service. We discuss the design of incentive compatible caching services that we refer to as quality of service caching. Pricing plays an important role in aligning the incentives. We develop an analytic model to study the IAP's optimal pricing and capacity allocation policies.*

## 1 INTRODUCTION

According to the Gartner Group, e-business will be a $4.4 trillion industry by 2003. Internet infrastructure is a key enabler of e-business. The infrastructure consists of the following players intermediating between the end user and the content provider:

(1) Internet access providers (IAP), such as AOL and Earthlink, that provide retail-level Internet access to end users.

(2) Local area transport (LAT) service providers: connecting the end user's premises to the IAP's point of presence (POP). These include dialup, DSL and cable modem service providers.

(3) Backbone networks such as AT&T and UUNET that own the Internet backbone and provide wholesale–level Internet access to IAPs. Backbones connect to each other through network access points (NAPs).

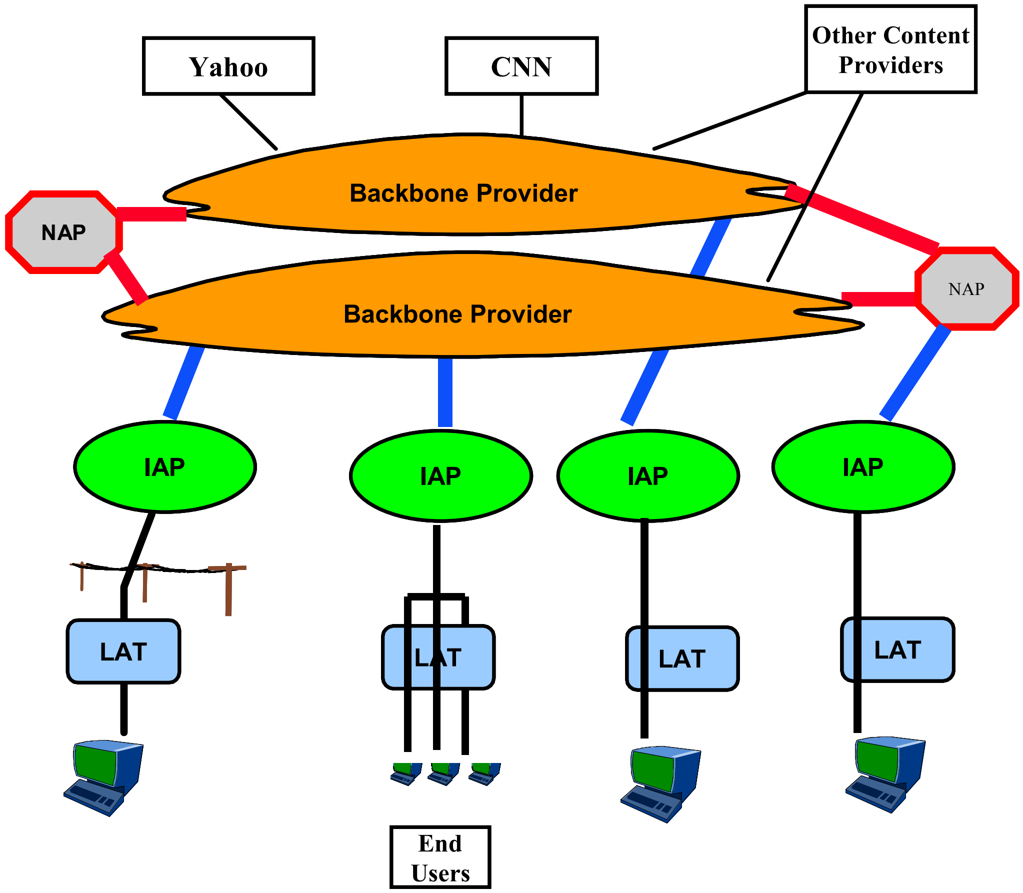


**Figure 1. The Internet Industry Structure**

Figure 1 illustrates the interaction between the players in the Internet value chain. Content publishers such as Yahoo and CNN typically connect to the backbone providers or host their content at data centers of the backbone providers. The end users obtain access to the content through the access networks (IAPs). Due to the growing traffic and congestion on the network, latency (delay in accessing content) and bandwidth costs tend to be high. Therefore, content delivery networks (CDNs, such as Akamai) and other caching service providers (IAPs such as AOL) often store local copies of content at the edge of the network (typically IAP servers) to minimize latency for the end users and bandwidth costs for IAPs and content providers.

Each player in the Internet value chain plays a distinct role. The networks are involved in providing transport services. They help move the content, created by content providers, over the Internet. Caches are the storage centers, the digital equivalent of warehouses. In this context, the Internet infrastructure makes up the *digital supply chain* for information goods. The content provider creates the content, networks help move the content, and the caches store and deliver it to the users. Our study focuses on caching because of the rapidly occurring transformation of these content storage and distribution centers and the critical impact that these changes might have on the digital supply chain and, therefore, to e-business. This is underscored in IDC's projection that the caching server appliance market would be worth $13 billion in 2004. In this paper, we focus on caching at IAP locations.

## 2 RESEARCH OBJECTIVES AND QUESTIONS

A Web cache can be conceptualized as an intermediary that stores local copies of Web content between the origin server and the client in order to satisfy future requests for the same (depicted in Figure 2). If data (a Webpage) is requested, the local copy is returned instead of requesting it from the origin server. Cache performance is measured by its hit-rate, which denotes the fraction of requests satisfied by the cache.

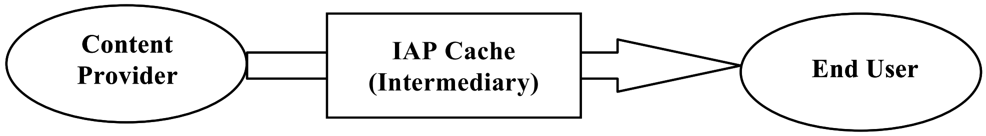


**Figure 2. Caching Service as Intermediation**

End users benefit from the latency reduction facilitated by caching. Caches also help the IAPs realize bandwidth savings as content served from the cache need not be transported over the backbone. In addition, content publishers also derive significant benefits. It reduces their bandwidth costs, as they do not have to deliver any data for requests satisfied from the cache. It also reduces the load on their servers and thus their infrastructure costs. Caching helps reduce response time in content delivery. According to Forrester Research, response time is a key determinant of consumer switching behavior on the Web and current expectations are around four seconds for downloading Web pages (Nelson 2000). Furthermore, by meeting a large fraction of the demand using locally stored content, it helps in handling flash crowds that are typical on the Internet. Flash crowd is the term used to refer to sudden surges in demand for content and is often responsible for server crashes and downtime due to the overwhelming activity levels. Content publishers do not currently pay for these caching services.

While the benefits of caching to content providers are well recognized, there are several reasons why caching is not pervasive. Principal among them is that caching results in loss of business intelligence regarding visits. Accuracy of access reports and click-stream data is crucial to various firms for marketing and internal audits. Hence, cache busting (marking content as non-cacheable) prevails among several content providers. Furthermore, a variety of e-marketing techniques rely on personalization through cookies. Jupiter Research reported that 40 percent of Fortune 500 companies had migrated to dynamic data driven personalized content as early as 1998. Traditional caching approaches have impeded personalization due to the possibility of content created for one user being displayed to another. In addition, some publishers cache-bust due to concerns relating to stale content being served to the end user. This occurs whenever changes in content at the origin server are not reflected at the cache. Finally, caching may potentially create new security concerns (violation of confidentiality, integrity, or authentication). For example, several publishers with confidential data such as medical records cache-bust due to security concerns arising from confidential data residing in a foreign location that is not under their immediate control.

Thus, there is a growing need to devise ways of gaining the benefits of caching without incurring the costs that precipitate cache busting. Recent innovations in caching technology provide the means to address the aforementioned problems. For example, dynamic content caching protocol (DCCP) (Smith et al. 1999) enables caching of dynamic content and thus facilitates personalization. Servers can also send caches invalidation messages (Yu et al. 1999) whenever content changes to circumvent problems associated with delivery of stale content. In addition, value added services such as differential caching, push caching, prefetching, etc further enable publishers to derive tangible benefits. Differential caching entails differential treatment to different data objects by reserving cache space or allocating a larger fraction of the cache space to high priority content (Kelly et al. 1999; Myers et al. 2001). Push caching and prefetching allow a publisher's content to be moved into the cache even before a request is made (in anticipation of future requests). Building on these developments, Chuang and Sirbu (1999) propose the *stor-serv* framework and discuss the technical requirements and design issues associated with service specification and provisioning. The bundling of content delivery, content targeting, and business intelligence with these value-added technologies will provide assured quality of service (QoS) in Web caching. We refer to the bundle of these new technologies (represented in Table 1) as *QoS caching*.

The use of these technologies to deploy value-added services is expensive for IAPs (Maggs 2002; Stargate 2002) and hence needs to be done in an incentive compatible fashion. Much of the work done to date on QoS in caching has focused on the technology and has not dealt with issues of fundamental importance to the business of provisioning caching services, specifically, the design of incentive compatible services, appropriate pricing schemes, and associated resource allocation issues that arise in operating a caching service. This is the focus of the paper.

**Table 1. Best Effort Caching Versus Quality of Service Caching**

| | QoS Dimension | Best Effort Caching | QoS Caching |
|---|---|---|---|
| 1 | Object Placement | Pull (Traffic-Driven) | Push, Prefetching |
| 2 | Object Replacement | Least Recently Used (LRU), LFU | Priority, Reservation |
| 3 | Object Consistency | Expiration Headers (weak) | Invalidation, Leases (strong) |
| 4 | Object Types | Static | Dynamic, Streaming |
| 5 | Business Intelligence | Logging | Reporting |
| 6 | Security | No | Yes |

# 3 RELEVANT LITERATURE

The analytic model used in this paper is related to the models in Mussa and Rosen (1978) and Bhargava et al. (2000). Mussa and Rosen consider the pricing of a product line by a monopoly, with buyers purchasing one good. Bhargava et al. study pricing strategies for intermediaries in electronic markets. Pricing of priority services has also been studied by others (Marchand 1974; Mendelson and Whang 1990). MacKie-Mason and Varian (1995) provide an overview of pricing issues on the Internet. The domain of Web caching poses unique challenges. First, the IAP also derives a positive utility from the caching service (bandwidth cost reduction). Thus, there is a strong positive interaction between the IAPs surplus and the content provider's surplus. This unique feature raises new questions and leads to fresh insights. In addition, content providers subscribe to the service but the end users, who do not participate in the subscription, generate the demand. In contrast, the aforementioned studies address pricing when the subscriber controls the demand (typically unit demand). Furthermore, pricing and resource allocation are strongly coupled in Web caching. The price that the IAP can charge depends partially on the hit rate it can provision, which is determined by the allocation decision. The optimal allocation decision depends on the traffic profile in the various service classes, which is in turn determined by the pricing.

QoS pricing has been addressed in detail in the network transmission domain. The reader is referred to Kaufmann and Walden (2001) for an overview of the literature on network pricing. The performance objectives in transmission and caching are dissimilar. In transmission, the goal is to reduce delay, jitter, or packet loss for applications that are sensitive to the same. It is also necessary to provide network operators with incentives to provide appropriate service levels to users from different subscriber bases. Thus, Gupta et al. (1997) and Cocchi et al. (1993) consider a pricing policy that maximizes social welfare rather than the network operator's profits. In caching, the QoS goal is to provide higher hit rates for objects that value caching more, providing security, consistency, etc. Resources need not be allocated along the path from the content publisher to the end user as in transmission. Allocation at the caching node alone suffices and this makes QoS caching easier to realize. Pricing provides the means to align the incentives of IAPs and content provider and thus achieve the QoS goals. The IAP would choose prices that maximize its profit rather than social welfare. Additionally, data objects stay in a cache for at least a few hours, even for "one-timer" objects that get purged soon. Hence more elaborate QoS mechanisms, such as those specified in Table 1, can be justified. Additionally, this also allows for object level pricing. Finally, the resource allocation issues are quite different too. In transmission, the router's queue management and scheduling operations provide the performance differentiation as opposed to space allocation in caching. All of these aspects make cache QoS pricing and resource allocation a unique and challenging problem.

# 4 TRACE ANALYSIS

Before we proceed to the analytic model, we start with an empirical analysis of Web traces in order to calibrate some of the parameters for the model. Trace data typically records requests for Web pages made by end users and hence reflects the demand for content. The Boeing trace (2002) consists of 4,292,154 requests over one day. The DEC trace (2002) has 7,866,111 requests.

*Distribution of Demand:* We denote the fractional demand for an object by $R$. That is, if four out of every 10 requests are for object $o_1$, then it has a fractional demand, $R = 0.4$. Note that R varies with time for a given object and hence is determined by a stochastic process. We assume that such a process is stable in that it has pdf associated with it. We study traces to determine $f(R)$, the probability density function (pdf) of $R$. $f(R)$ may intuitively be considered as a measure of the number of objects with fractional demand equal to $R$. The distribution is modeled as $f(R) = \frac{c}{R^{\beta}}$ where $R$ lies in [$c_1$, $c_2$]. The Maximum likelihood estimates of the parameters are in Table 2. Goodness of fit tests indicate that the fit for both of the traces is good.

**Table 2. Estimates of Parameters for the Distribution of *R***

| Trace | $c$ | $\beta$ | $c_1$ | $c_2$ |
|---|---|---|---|---|
| Boeing | 2.08996E-04 | 1.51146 | 2.33E-07 | 0.002674 |
| DEC | 1.27382E-03 | 1.35583 | 1.27E-07 | 0.010715 |

For the purposes of the analytic model, we use the simpler form: $f(R) = \frac{c}{R}$. This merely changes the constant of integration in our results.

*Cache Hit Rates*: Hit Rate, H(S), denotes the fraction of requests answered by a cache of size S. Clearly, a larger cache has a higher hit rate because it can store a larger number of data objects. We used a cache simulator to simulate cache performance for requests in the two traces. We find that the logarithmic specification, $H(S) = k_s.\ln(S)$, fits well for both traces (see Table 3). The results are consistent with earlier literature such as (Breslau et al. 1999).

**Table 3. Estimates for $H(S) = k_s.\ln(S)$**

| Trace | Estimate ($k_s$) | Standard Error | Pr > \|t\| | R-Square |
|---|---|---|---|---|
| Boeing | 0.05004 | 0.00003063 | <.0001 | 0.9367 |
| DEC | 0.04523 | 0.00002860 | <.0001 | 0.9076 |

*Object Specific Hit Rate:* We also seek the functional form of the object specific hit rate, H(S,*R*), the hit rate of an object with fractional demand *R* in a cache of size S. This denotes the fraction of requests for a specific object that were satisfied by the cache. Intuitively, the hit rate denotes the probability of a hit for a random request whereas the object specific hit rate denotes the probability of a hit given the object that is requested. While the IAP cares about the overall hit rate, the content provider is only concerned about the object specific hit rates for her objects. We model the object specific hit rate as follows: $H(S,R) = k.\ln(S).\ln(\lambda R)$, where $\lambda$ denotes the total number of requests. We find that this specification also fits well with the data (see Table 4).

**Table 4. Estimates for $H(S,R) = k.\ln(S).\ln(\lambda R)$**

| Trace | Estimate (k) | Standard Error | Pr > \|t\| | R-Square |
|---|---|---|---|---|
| Boeing | 0.02296 | 0.00002463 | <.0001 | 0.828 |
| DEC | 0.02271 | 0.00002672 | <.0001 | 0.740 |

## 5 OPTIMAL PRICING

We assume that the content provider/publisher makes the caching decision by data object. This reflects the real world situation very well. For example, publishers such as CNN decide which objects are stored and delivered from Akamai's server and which are delivered from their own servers. Further, valuations for different data objects may be different, even for the same publisher. We also assume that the IAP uses a usage based pricing policy as that is the norm in the caching industry today.

## 5.1 Content Provider's Decision Problem

The content provider has content (data objects) that is requested by an end user. The IAP is a conduit through which the content is delivered. Due to its unique position in content delivery, the IAP can provide the publisher with additional value through caching. We separate the value derived by the publisher from the caching service into two components: hit rate based benefit and non-hit rate based benefit. The former represents benefits directly associated with caching (such as bandwidth savings and faster content delivery). These benefits are derived every time an object is served from the cache (called a hit). The latter represents benefit from value-added features such as support for caching dynamic content, security, etc.

The quality of the value-added features is denoted by $q \in \Re$. We denote the benefit to the publisher from the non-hit rate based quality for an object by $\theta.q$. $\theta$ is a type parameter that denotes the weight that the publisher attaches to non-hit rate quality for the specific object. We shall also refer to an object associated with weight $\theta$ as an object of type $\theta$. This weight would clearly vary from object to object, even for a given publisher. For example, a data object containing confidential information may be associated with a high $\theta$ whereas another not requiring security, reporting, support for dynamic data caching, etc. may correspond to an extremely low $\theta$. We assume that $\theta$ is uniformly distributed in [0,1]. We find that the general nature of the results do not change for several other well-known distributions.

The publisher additionally derives a benefit from being able to deliver his content faster to the consumers of his content. This benefit captures reduced churn from faster delivery of content. For every object delivered from the cache, $\eta$ represents the aforementioned benefit to the publisher. Another important component of the benefit from the service is the bandwidth savings realized by the publisher. B denotes the average bandwidth cost for processing one object request (and hence the benefit to the publisher from caching an object). A piecewise separable benefit function is used, $U = \theta.q + (Thc).\eta + (Thc).B$, where $Thc$ is the count of the number of objects served from the cache. If $\lambda$ is the arrival rate of requests and $R$ the fractional demand for an object, then $\lambda R$ is the number of requests for that object in a unit time period. If $H(S, R)$ denotes the object specific hit rate, then $Thc = \lambda RH(S, R)$.

Note that demand is driven by user requests and, therefore, the publisher has no direct control over it. Further, users typically subscribe to particular IAPs and cannot switch IAPs instantaneously. Therefore, the IAP has monopolistic power over the publisher's access to users. This arises from it being the only conduit to any particular end user. A different IAP can only provide access to a different set of users. Furthermore, IAPs such as AOL have considerable market share which enables them to provide significant value to publishers that is hard to replace. We therefore consider a monopoly pricing model.

We assume that the IAP offers two services: a best effort service and a premium service. The best effort service provides lower quality service and is ideally suited for publishers with lower valuation for value-added features like business intelligence, security, etc. The IAP charges price $P_L$ for every object served from its best effort cache. In addition, it offers a service at a higher quality level and charges a per object price $P_H$ for the same. The higher quality is achieved by supporting dynamic content, object consistency, reporting, etc. and provisioning a larger fraction of the cache for premium objects. The publisher chooses the service that provides her with the maximum surplus.

We denote the higher quality and lower non-hit qualities by $q_H$, $q_L$ respectively. Thus, the publisher's net surplus from the two services is given by:

$$U_L = \theta.q_L + \lambda.R.H(S_L, R)(\eta + B) - P_L \lambda.R.H(S_L, R) \tag{1}$$

$$U_H = \theta.q_H + \lambda.R.H(S_H, R)(\eta + B) - P_H.\lambda.R.H(S_H, R)$$

where $S_L$ and $S_H = (1 - \alpha)S$.

The cache space, S, is divided into two levels: the lower one of size $\alpha S$ for the best effort subscribers and the higher one of size (1- $\alpha$)S for the premium publishers. To determine the number of subscribers to the service, we consider a publisher of an object with type $\theta_L$ who is indifferent between paying for the low quality service and not subscribing to the service at all (gets zero surplus from the best effort service) and the publisher of an object with type $\theta_H$ who is indifferent between the two services (gets equal surplus from the two services). Any publisher with $\theta_L < \theta < \theta_H$ will choose the best-effort service and a publisher with $\theta > \theta_H$ will subscribe to the premium service. Objects of type $\theta < \theta_L$ do not subscribe to either service. By equating the surplus from the lower quality service to 0, we get $\theta_L$ and by equating the surplus from the two services, we get $\theta_H$.

$$\theta_L = \frac{(P_L - \eta - B).\ell n(S_L)k\lambda.R.\ell n(\lambda R)}{q_L}$$

$$\theta_H == \frac{\{(P_H - \eta - B).\ell n(S_H) - (P_L - \eta - B).\ell n(S_L)\}k\lambda R.\ell n(\lambda R)}{q_H - q_L} \tag{2}$$

Both $\theta_L$ and $\theta_H$ vary with demand $R$ and are thus curves that represent indifferent content publishers. We call these the *quality indifference curves* (QIC) for the publishers. These are the indifference points for consumers used in microeconomic models except that in our case, they are a function of the fractional demand $R$. A sample QIC, based on assumed prices and quality levels is shown in Figure 3.

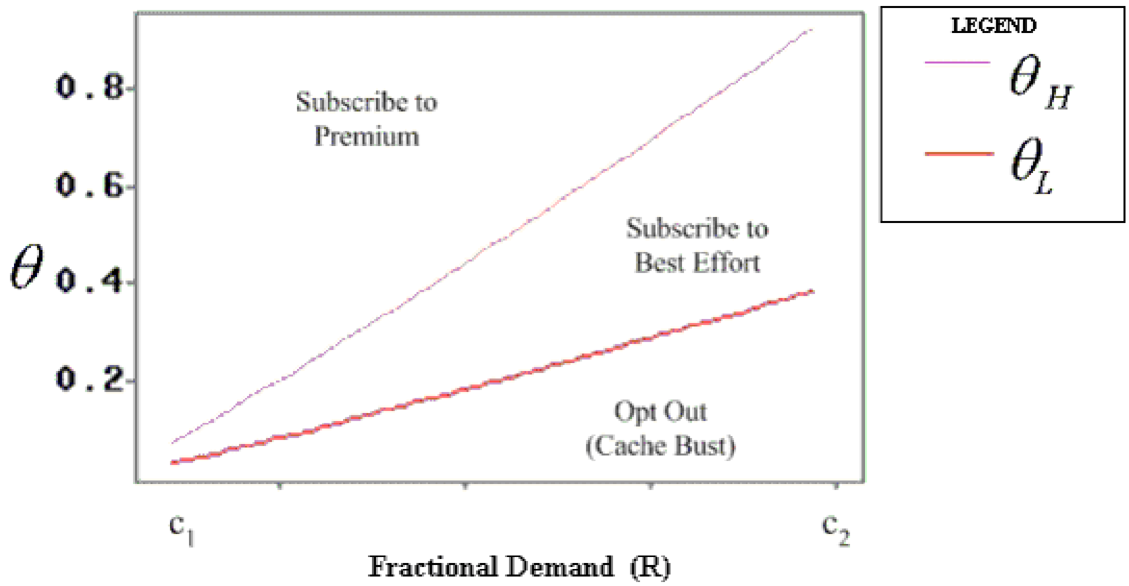


**Figure 3. Sample Indifference Curves (for Assumed Price and Quality Levels)**

## 5.2 IAP's Decision Problem

To compute the IAP's profit function, we first determine the expected number of requests for objects in the premium service by summing up the number of requests for all objects with $\theta > \theta_H$ (see Figure 3). This is denoted by $R_H$. Similarly, the expected number of requests for best-effort objects, $R_L$, is obtained by summing up requests for objects of type $\theta \in [\theta_L, \theta_H]$. The IAP's expected profit is given by:

$$\pi = R_H H(S_H)[P_H + B_{IAP}] + R_L H(S_L)[P_L + B_{IAP}] - C(q_H)$$

Out of the $R_H$ requests for premium objects, a fraction $R_H H(S_H)$ are delivered from cache and, similarly, $R_L H(S_L)$ are the number of objects delivered from cache for the best-effort service. For each object served from the cache, the IAP has bandwidth savings, $B_{IAP}$. The IAP charges the publisher a per-object price $P_L$ or $P_H$ depending on the service that the object is subscribed to. It is expensive for the IAP to provision value added services. The cost to the IAP is $C(q_H)$. Substituting all of the variables into the profit function and integrating, we get the following expression for the IAP's profit function:

$$\pi = \lambda\left[1 - \frac{\lambda N k k_1 c\{\overleftrightarrow{(P_H - \eta - B)\ell n S_H}^{X} - (P_L - \eta - B)\ell n S_L\}}{(q_H - q_L)}\right] \overleftrightarrow{k_s \ell n S_H}^{Y} \overleftrightarrow{[P_H + B_{IAP}]}^{Z}$$

$$+ \frac{\lambda^2 N k k_1 c\{q_L \overleftrightarrow{(P_H - \eta - B)\ell n S_H}^{D} - q_H (P_L - \eta - B)\ell n S_L\}}{q_L(q_H - q_L)} . \overleftrightarrow{k_s \ell n S_L}^{E} \overleftrightarrow{[P_L + B_{IAP}]}^{F} - C(q_H)$$

The term denoted X is the number of requests for objects in the premium service. Y denotes the hit rate for the same and Z the sum of price charged and bandwidth savings per object (profit per object). Similarly, D is the number of requests for objects in the best-effort service, E the associated hit rate, and F the profit per object. The IAP's decision problem is: $\max_{P_H, P_L} \pi(P_H, P_L)$. The optimal prices are computed from the first order conditions.

**Lemma 1:** The optimal prices that the IAP should charge are

$$P_L = \frac{q_L}{(2\lambda N k k_1 c).\ln S_L} + \left(\frac{\eta + B - B_{IAP}}{2}\right)$$

$$P_H = \frac{q_H}{(2\lambda N k k_1 c).\ln S_H} + \left(\frac{\eta + B - B_{IAP}}{2}\right)$$

The optimal quality indifference curves associated with the optimal prices are

$$\theta_L = R\ln(\lambda R)\left\{\frac{1}{2Nk_1 c} - \left(\frac{\eta + B + B_{IAP}}{2}\right)\frac{k\lambda \ln S_L}{q_L}\right\}$$

$$\theta_H = R\ln(\lambda R)\left\{\frac{1}{2Nk_1 c} - \left(\frac{\eta + B + B_{IAP}}{2}\right)\frac{k\lambda(\ln S_H - \ln S_L)}{q_H - q_L}\right\}$$

## 5.3 Analysis

The prices charged vary linearly with the quality offered. The IAP charges the content publisher a part of her surplus from bandwidth reduction and faster content delivery ($\eta + B$) and gives back to the publisher a part of its own surplus from bandwidth cost reduction ($B_{IAP}$). If the IAP's bandwidth costs start to dominate, then the discount can be substantial. The IAP may even find it optimal to provision the best effort service for free if the IAP's bandwidth costs are high enough and the quality of the best effort service is low enough. This makes intuitive sense considering that the best effort service is currently provisioned at zero cost to the publishers. The following propositions follow directly from the expressions for $\theta_L$ and $\theta_H$

**Proposition 1:** If the quality of the best effort service is increased, subscription to the best effort service decreases and to the premium service increases.

**Proposition 2:** If the quality of the premium service is increased, subscription to the best effort service increases and to the premium service decreases.

Propositions 1 and 2 seem counterintuitive at first glance. One would expect the impact of modifying either $q_L$ or $q_H$ to be in the opposite direction. However, an increase in $(q_H - q_L)$ differentiates the services sufficiently to enable the IAP to charge even more for the premium service. Objects with high $\theta$ would still continue with the premium service (but pay more) but those objects located close to the indifference points would rather just switch to the best effort service than pay the steep price for the premium service. In contrast, a decrease in $(q_H - q_L)$ prevents the IAP from being able to charge a premium on the high quality service.

Since the price is low, a larger fraction of the publishers subscribe to the premium service. This causes subscriptions to change as indicated in the propositions.

**Proposition 3:** As bandwidth costs decrease, subscription to both the services decreases.

# 6 SPACE ALLOCATION

In the previous section, we looked at the pricing problem assuming that the cache sizes $S_L = \alpha S$ and $S_H = (1 - \alpha)S$ were exogenously decided. Space allocation is, however, a real-world problem faced by cache operators. Typically caches sizes are optimized based on traffic profiles and never over-provisioned. This is because the gains from increasing cache sizes increase at a decreasing rate and costs dominate beyond the optimal cache size (Kelly and Reeves 2000). The cost of incremental upgrades at various caching nodes tends to be high, hence they are rarely resized unless the traffic profile changes substantially (Maggs 2002). Thus, caches are a capacitated resource and allocation of the space is an important consideration. The first order conditions with respect to $\alpha$ do not yield a closed form solution, hence we consider a numerical example to illustrate how the IAP may solve its allocation problem.

We simulate values for the parameters in the model. We consider an average publisher with bandwidth cost of 0.03c per object per month. This corresponds to a publisher with a T1 connection priced at $750 per month and an average object of size 50 KB. The IAP's bandwidth costs would be lower and is assumed to be 0.02c per object per month. The publisher's benefit from faster delivery of content to the end consumer is assumed to be $\eta = 0.03$c. The publisher who is most sensitive to non-hit rate attributes such as security, consistency, reporting, etc. ($\theta = 1$) values these features an order of magnitude more than the bandwidth savings ($q_H = 0.3$). Finally, $q_L$ is set artifically low at 0.001. The cache size is assumed to be 5 GB. All of the remaining parameters are derived from the Boeing trace. For these settings, we found that the IAP would find it optimal to allocate 16.8 percent ($\alpha^* = 0.168$) of the cache space to the lower level cache.

An interesting question relates to the impact of recent trends in decreasing bandwidth costs on the allocation decision. We repeat the simulations but halve the bandwidth costs for both the IAP and the content publisher ($B = 0.15$ and $B_{IAP} = 0.1$). Under the new settings, it is optimal for the IAP to allocate 11.4 percent of the cache to the lower level. Figure 4 lays out the impact of lowering bandwidth costs on $\alpha^*$. $B_{BC}$ represents the base case of $B = 0.3$ and $B_{IAP} = 0.2$. In each successive simulation, we halve the bandwidth costs from the previous simulation. We observe that the IAP finds it optimal to reduce the size of the best effort cache and increase the size of the premium cache as bandwidth costs decrease.

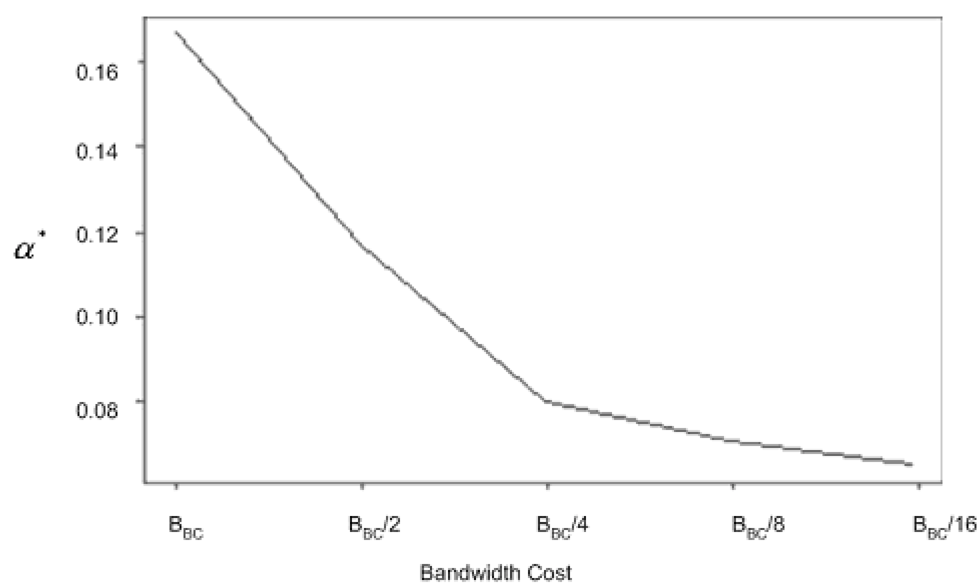


**Figure 4. Impact of Decreasing Bandwith Costs on Optimal Allocation**

## 7 CONCLUSIONS

QoS is the leading performance consideration in e-business today. We introduce a framework to structure and analyze the QoS issues in Web caching in an integrated manner. If designed prudently, QoS caching would move content delivery almost entirely to the edge. This could change the structure of the digital supply chain and have significant impact on e-business infrastructure. For example, it could move "intelligent" processing of collateral information, of interest to e-marketing, to the edge of the network as well. When combined with our conceptual view of content delivery as a digital supply chain, this suggests that content providers would gradually become manufacturers of content and caches would handle the storage and retailing of content. This is a significant reinvention of content delivery as it exists today.

Our study provides insights into design of incentive-compatible caching services in general, and the pricing of these services and capacity allocation in particular. Aligning the incentives of the players is a key requirement. Despite the fact that publishers clearly receive benefits from caching, currently an appropriate payment scheme does not exist. This has partly been due to the best-effort nature of Web caching. In addition, publishers have requirements that are currently not fulfilled by the IAP such as provisioning business intelligence, content personalization, etc. This in turn has been due to the lack of appropriate payment schemes. We have outlined a mechanism to correct both these deficiencies. The framework also has natural extensions to the CDN market, which is being explored. No previous research exists in the area of cache QoS pricing and capacity allocation.

We find that value-added services would allow cache operators to charge significantly higher prices and thus price discriminate effectively. If the IAP increases the quality, it can charge more but the subscription to the premium service decreases. Furthermore, it also has to consider the cost of provisioning these value-added services and be strategic about its choice of the level of the premium service quality. Additionally, we find that the market share in both of the services would drop with falling bandwidth costs. This further underscores the need to provision value-added services. Our analysis has shown that best effort services will play a diminishing role with decreasing bandwidth costs. Resources may be directed toward serving the maximum number of data objects from the premium cache (in the limit, this suggests that edge delivery of entire sites would be the norm). This is an indication of the impending metamorphosis of the content delivery value chain.

We have assumed that all objects have the same size but find that the general nature of the results do not change even if size is accounted for (although the quality indifference curves become three dimensional). In addition, an object with high demand may also have a higher valuation for quality. It is an empirical exercise to determine the correlation between quality valuation and demand. Future work includes studying the impact of alternative pricing schemes (usage-based versus fixed fee) and extending the framework to CDNs.

## 8 REFERENCES

Bhargava, H., Choudhary, V., and Krishnan, R. "Pricing and Product Design: Intermediary Strategies in an Electronic Market," *International Journal of Electronic Commerce* (5:5), 2002, pp. 37-56.

Boeing. "Boeing Proxy Logs," 2002 (available online at ftp://researchsmp2.cc.vt.edu/pub/boeing/).

Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S. "Web Caching and Zipf-Like Distributions: Evidence and Implications," paper presented at the IEEE Infocom Conference, New York, 1999.

Chuang, J., and Sirbu, M. "Stor-Serv: Adding Quality-of-Service to Network Storage," paper presented at the MIT Workshop on Internet Service Quality Economics, Cambridge MA, December 1999.

Cocchi, R., Shenker, S., Estrin, D., and Zhang, L. "Pricing in Computer Networks: Motivation, Formulation and Example," *IEEE/ACM Transactions on Networking* (1), December 1993.

Digital Equipment Corporation. "Digital's Web Proxy Traces," 2002 (available online at ftp://ftp.digital.com/pub/DEC/traces/proxy/webtraces.html).

Gupta, A., Stahl, D., and Whinston, A. "Priority Pricing of Integrated Services Networks," in L. McKnight and J. Bailey (eds.), *Internet Economics*. Boston: MIT Press, 1997.

Kauffman, R., and Walden, E. "Economics and Electronic Commerce: Survey and Directions for Research," *International Journal of Electronic Commerce* (5:4), 2001.

Kelly, T., Jamin, S., and MacKie-Mason, J. "Variable QoS from Shared Web Caches: User-Centered Design and Value-Sensitive Replacement," paper presented at the MIT Workshop on Internet Service Quality Economics, Cambridge, MA, December, 1999.

Kelley, T., and Reeves, D. "Optimal Web Cache Sizing: Scalable Methods for Exact Solution," in A. Tsalgatidou (ed.), *Proceedings of the Fifth International Conference on Web Caching and Content Delivery*, Lisbon, Portugal, May 2000.

MacKie-Mason, J., and Varian, H. "Pricing the Internet," in B. Kahin and J. Keller (eds.), *Public Access to the Internet*. Cambridge, MA: MIT Press, 1995, pp. 269-314.

Maggs, B., Vice President, Akamai. Personal Communication, 2002.

Marchand, M. "Priority Pricing," *Management Science* (20), 1974.

Mendelson, H., and Whang, S. "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research* (38), 1990, pp. 870-883.

Mussa, M., and Rosen, S. "Monopoly and Product Quality," *Journal of Economic Theory* (18), 1978.

Myers, A., Chuang, J., Hengartner, U., Xie, Y., Zhuang, W., and Zhang, H. "A Secure, Publisher-Centric Web Caching Infrastructure," *IEEE Infocom*, April 2001.

Nelson, M. "Fast Is Nno Longer Fast Enough," *InformationWeek Online*, June 6, 2000 (available online at www.informationweek.com/789/web.htm, accessed December 9, 2000).

Smith, B., Acharya, A., Yang, T., and Zhu, H. "Exploiting Result Equivalence in Caching Dynamic Web Content," in *Proceedings of 1999 USENIX Symposium on Internet Technologies and Systems*, Boulder, CO, October 1999.

Stargate, Inc. Personal Communication, 2002.

Yu, H., Breslau, L., and Shenker, S. "A Scalable Web Cache Consistency Architecture," in *Proceedings of ACM SIGCOMM '99,* Cambridge, MA, August 1999.