

Pricing and Resource Allocation in Caching Services with Multiple Levels of Quality of Service

Kartik Hosanagar

The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, kartikh@wharton.upenn.edu

Ramayya Krishnan

Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, rk2x@cmu.edu

John Chuang

University of California, Berkeley, California, chuang@berkeley.edu

Vidyanand Choudhary

University of California, Irvine, California, veecee@uci.edu

Network caches are the storage centers in the supply chain for content delivery—the digital equivalent of warehouses. Operated by access networks and other operators, they provide benefits to content publishers in the forms of bandwidth cost reduction, response time improvement, and handling of flash crowds. Yet, caching has not been fully embraced by publishers, because its use can interfere with site personalization strategies and/or collection of visitor information for business intelligence purposes. While recent work has focused on technological solutions to these issues, this paper provides the first study of the managerial issues related to the design and provisioning of incentive-compatible caching services. Starting with a single class of caching service, we find conditions under which the profit-maximizing cache operator should offer the service for free. This occurs when the access networks' bandwidth costs are high and a large fraction of content publishers value personalization and business intelligence. Some publishers will still opt out of the service, i.e., cache bust, as observed in practice. We next derive the conditions under which the profit-maximizing cache operator should provision two vertically differentiated service classes, namely, premium and best effort. Interestingly, caching service differentiation is different from traditional vertical differentiation models, in that the premium and best-effort market segments do not abut. Thus, optimal prices for the two service classes can be set independently and cannibalization does not occur. It is possible for the cache operator to continue to offer the best-effort service for free while charging for the premium service. Furthermore, consumers are better off because more content is cached and delivered faster to them. Finally, we find that declining bandwidth costs will put negative pressure on cache operator profits, unless consumer adoption of broadband connectivity and the availability of multimedia content provide the necessary increase in traffic volume for the caches.

Key words: Web caching; content delivery; pricing; capacity allocation; quality of service (QoS)

History: Accepted by Rajiv D. Banker, information systems; received April 24, 2003. This paper was with the authors 8 months for 2 revisions.

1. Introduction

The phenomenal growth of content and applications on the Internet has helped make e-business a large and growing part of overall commerce. Internet infrastructure is a key enabler of e-business. The infrastructure consists of the following players intermediating between end users and content publishers:

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

(2) Local area transport (LAT) service providers such as local phone companies and cable franchises that connect end users' premises to the IAPs' points of presence (POPs).

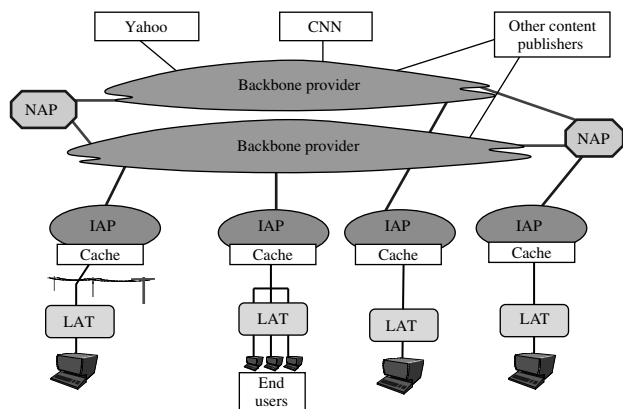
(3) Backbone networks such as AT&T and UUNET that operate long-haul data networks and

interconnect with one another through network access points (NAPs) and bilateral peering points to form the Internet backbone. IAPs connect to the backbone networks and pay them for bandwidth. Content publishers also pay the backbone networks for bandwidth consumed (either directly or indirectly through hosting services).

(4) Content delivery networks (CDNs) such as Akamai that deliver content on behalf of content publishers using proprietary networks of caching servers.

Together, these players make up the digital supply chain for delivery of information goods on the Internet (see Figure 1). The content publisher creates the content and networks help move the content. Caches—the digital equivalent of warehouses—store and deliver the content to the users. Our study

Figure 1 The Internet Industry Structure



focuses on caching because of the rapidly occurring transformation in caching services and the critical impact that these changes will have on the digital supply chain and therefore to e-business. In this paper, we will focus on caches operated by IAPs, although much of our analysis applies to CDNs as well.

Web caching refers to the temporary storage of content somewhere between the origin server and the end user to satisfy future requests for the same. The effectiveness of a cache is measured by the hit rate, which is the fraction of requests that are satisfied by the cache. Because IAP caches are located at points of aggregation, they are shared by a large number of users and hence demonstrate high hit rates. A recent survey of caching from a management science perspective is provided by Datta et al. (2003). By delivering content from the edge of the network, IAP caches reduce the latency experienced by the user in content delivery. Furthermore, because content served from the cache need not be fetched over the upstream backbone provider's network, this reduces the bandwidth payments the IAP makes to the backbone.¹ In addition, content publishers also derive the following significant benefits from caching:

(1) Handling flash crowds: *Flash crowd* is the term used to refer to sudden surges in demand for content that can bring servers down and render websites unreachable. Caches alleviate the problem by meeting a large fraction of the demand using locally stored content.

(2) Improving content delivery/response time: Response time is a key determinant of consumer switching behavior on the Web. According to Jupiter Research, nearly half of the Web users have stopped using a preferred website due to poor performance.

¹ It is estimated that the potential bandwidth savings to AOL from its caches was about \$430 million in 2002 (see the electronic companion pages at <http://mansci.pubs.informs.org/ecompanion.html>).

By improving response time, caches can increase customer retention rates for websites.

(3) Reducing bandwidth costs: Caching also reduces bandwidth costs for content publishers, as they do not have to deliver any data for requests satisfied from the cache.

(4) Reducing infrastructure costs and scaling content delivery globally.

Currently, content publishers do not pay for IAP caching services. Despite these seemingly large benefits, content publishers often choose to mark their content as noncacheable. Chuang et al. (2002) found that 26 out of MediaMetrix's list of 50 most popular websites prevent caching of their content. This practice, also known as cache busting, results in slower response times for end users and increased costs for IAPs and the publishers themselves. The primary reason for cache busting is that caching results in a loss of business intelligence regarding visitors to websites because the origin server is not informed whenever cached content is served. Accuracy of access reports and click-stream data is crucial to various firms for marketing and internal auditing. Furthermore, support for dynamic content caching in IAP caches has been minimal thus impeding personalization. A variety of e-marketing techniques rely on personalization of content. Jupiter Research reported that 40% of Fortune 500 companies had migrated to dynamic data-driven personalized content as early as 1998. 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. Caching can also create new security concerns (violation of confidentiality, integrity, or authentication) for publishers because content resides in a foreign location that is not under their immediate control.

These can be addressed by provisioning quality of service (QoS) in caching and ensuring comprehensive reporting, object consistency, security, etc. However, supporting these features is in general expensive, which deters the IAPs from provisioning them (Maggs 2002, Stargate 2002). Recent work has focused on technological solutions to these QoS issues, whereas this paper presents the first study of the managerial issues related to the design and provisioning of incentive-compatible caching services, appropriate pricing schemes, and resource allocation policies. This paper is organized as follows: We state the research question and summarize key results in §1.1. We review the related literature and present our integrated QoS framework in §2. In §3, we develop an analytical model for pricing and resource allocation. In §4, we relax a number of assumptions and test the robustness of our results. We conclude in §5 with a summary of our findings and directions for future research.

1.1. Research Question and Key Results

This paper analyzes the impact of recent developments such as increased demand for business intelligence, personalization of content, etc., on caching service adoption. We then proceed to examine how QoS-based caching services impact the market dynamics of content provision. Through analytical models, we hope to inform cache operators regarding service provisioning and optimal pricing policies. The models also provide valuable insight into the future of caching and content delivery.

The main contribution of this paper is in establishing the positive impact that QoS-based caching can have on publishers and IAPs. We have four primary findings. First, we find that today's best-effort services are free because the IAP derives significant bandwidth savings from caching, coupled with high publisher sensitivity for value-added features. However, free services cannot eliminate cache busting. Second, our research suggests that the design of QoS-based services can significantly benefit publishers, IAPs, and end users. Interestingly, we find that QoS provisioning differs from traditional vertical differentiation models in that the two services do not abut. Each publisher receives positive surplus from at most one of the two services. Thus, there are no cannibalization effects between the two services. Third, we find that resources will be increasingly directed toward premium services in the future. Fourth, declining bandwidth costs will have a significant negative impact on profits from caching services. This can be mitigated in the short term by a shift toward broadband access by end users coupled with provision of rich content by publishers.

2. QoS Framework and Literature Review

QoS caching is publisher centric and focuses on providing verifiable QoS and value-added services to publishers (Myers et al. 2001, Kelly et al. 1999, Chuang and Sirbu 2000). In contrast to the traditional caching policies that are traffic driven (also known as best-effort caching), QoS caching allows an IAP to adopt caching schemes based on publishers' willingness to pay.

Various techniques can be used to implement QoS caching. Object placement policy refers to the policy used to determine when data objects move into a cache. Traditional placement policies entail an object being moved into the cache only when a request is made for it. Push caching and prefetching allow a publisher's objects to be moved into the cache in anticipation of future requests.

Because caches have a finite size, an object may be purged from the cache when a new object is

moved in. The decision about which data object to replace is governed by the replacement policy. Replacement policies such as least recently used (LRU) and least frequently used (LFU) are commonly employed. LRU replaces objects that were least recently requested, assuming that they are also least likely to be requested again (Mookerjee and Tan 2002), and LFU replaces objects that were accessed the least frequently. Object replacement policies may also be modified to include differential treatment to data objects through differential caching techniques such as cache reservation and priority caches. Cache reservation involves reservation of a predetermined space in the cache for objects from a specific publisher. Priority caches allow the assignment of priorities to different classes of content and provide high priority data objects with higher hit rates. Chan et al. (1999) propose a market mechanism for a replacement policy in which content publishers bid for space in a cache.

Current best-effort caches maintain consistency by the use of expiration headers that specify the expiration time or time to live (TTL) of the document. However, data sources may be modified before the TTL expires, resulting in stale content being served. These problems can be circumvented by the use of invalidation schemes wherein the server sends invalidation messages (Yu et al. 1999) to the cache whenever content changes, or by using leases (Yin et al. 1998). Additionally, the cache operator can provide reports on access patterns to publishers (Mogul and Leach 1997). Furthermore, caches can add support to dynamically generated data and streaming data. The dynamic content caching protocol (DCCP) (Smith et al. 1999) and dynamic proxy cache (DPC) (Datta et al. 2002) allow origin servers to specify the caching policies for dynamic content. Table 1 summarizes the dimensions along which QoS may be varied.

Provisioning QoS-based services and pricing them is critical to aligning the incentives of the IAP and publishers. However, as highlighted in §1, current services do not address these issues. One generic explanation for the inability of markets to implement all the gains that can be achieved is that transaction costs to do so may be high (Zerbe and McCurdy 2000).

Table 1 Best-Effort Caching vs. QoS Caching

QoS dimension	Best-effort caching	QoS caching
1 Object placement	Pull (traffic driven)	Push, prefetching
2 Object replacement	LRU, LFU, and variants	Priority, reservation
3 Object consistency	TTL (time to live), if modified since (weak)	Invalidation, leases (strong)
4 Object types	Static	Dynamic, streaming
5 Accounting	Logging	Reporting
6 Security	None	Confidentiality, integrity, authentication

Generally, we expect that improved connectivity and communication across organizational boundaries has brought about significant reduction in barriers such as contracting costs (in our context, these are costs of metering and billing). However, for the sake of completeness, we will explore the impact of contracting costs on caching service provisioning and also identify alternative explanations for the nonexistence of contracts and pricing in current caching services.

2.1. Cache Pricing and Resource Allocation—Unique Challenges

While there are previous studies on pricing of priority services (Marchand 1974, Mendelson and Whang 1990), the domain of Web caching poses unique challenges. First, the subscribers of the caching service (content publishers) are not the generators of demand. The publishers subscribe to the service but the end users, who do not directly participate in the subscription, generate the demand (number of requests and, hence, objects served from the cache). In addition, the IAP also derives a positive benefit from the caching service (bandwidth cost reduction). Thus, there is also a strong interaction between the service provider's surplus and the subscriber's surplus.

The analytical model in this paper is related to the models in Mussa and Rosen (1978) and Bhargava et al. (2000). Mussa and Rosen (1978) consider the pricing of a product line by a monopoly, with buyers purchasing one good. Bhargava et al. (2000) study pricing strategies for intermediaries in electronic markets. Sundararajan (2004) studies optimal pricing of information goods when both fixed-fee and usage-based pricing are feasible. Dewan et al. (2000) study the relationship between proprietary content providers and IAPs in distribution channels for information goods on the Internet. As mentioned above, cache QoS pricing is a problem different in structure and scope.

QoS pricing has also been addressed in detail in the transmission domain (Gupta et al. 1997, Cocchi et al. 1993). However, the pricing and resource allocation problems are quite different. In transmission, the router's queue management and scheduling operations allocate a constrained buffer and bandwidth and provide the performance differentiation for data packets arriving in real time. The real-time nature also implies that the pricing has to be coarser than packet-level pricing as it would be too costly to implement. On the other hand, the resource allocation problem in caching relates to the allocation of the available cache space between the service classes. The IAP has to account for the fact that its allocation decision also impacts its own bandwidth savings (the allocation decision may lower the overall hit rate and hence increase the IAP's bandwidth costs). Data objects stay in a cache for at least a few hours, even for "one-timer" objects that get purged quickly. Hence, more

elaborate QoS mechanisms, such as those specified in Table 1, can be justified. Additionally, this also allows for object-level pricing. In contrast, even per-flow QoS (intserv) is deemed nonscalable in transmission and the focus has more recently been on per-class QoS (diffserv).

The performance objectives of QoS in caching and transmission are dissimilar as well. In transmission, the goal is to reduce delay, jitter, and/or packet loss for performance-sensitive applications. To achieve end-to-end QoS, it is necessary to provide network operators with incentives to ensure 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 the collective benefits of the system rather than the network operator's profits. In caching, the QoS goal is to provide higher hit rates for objects of publishers that value caching more, provide security, etc. Resources need not be allocated along the path as in transmission. Allocation at the caching node alone suffices and this makes QoS caching easier to realize. Further, the IAP can also choose prices that maximize its profit rather than social welfare. Pricing provides a means to align the incentives of IAPs and publishers and thus achieve the QoS goals.

3. Analytical Model

The unit of analysis in the analytical model presented in this section is an object. That is, we assume that the content publisher makes the caching decision by data object. This reflects reality in that publishers' decisions regarding the content they mark as cacheable vary from one data object to another. For example, a publisher may care about security for one data object that contains confidential data but may not insist on security for another.

We consider a monopoly pricing model in this paper. This is because users typically subscribe to particular IAPs and cannot switch IAPs instantaneously. Therefore, the IAP has monopolistic power over publisher's access to users. This arises from it being the only conduit to any particular set of end users. A different IAP can only provide access to a different set of users and hence cannot be treated as a perfect substitute. In addition, large IAPs such as AOL have considerable market share that enables them to provide significant value to the publishers that is hard to replace.

In the resource allocation section, we treat the cache size, S , as a fixed exogenous parameter and do not consider determination of the cache size. There are two reasons for this. First, the issue of optimal cache sizing has been considered in detail by Kelly and Reeves (2000). Second, we focus on the problem of

an IAP, with a cache in place, making the decision of provisioning a premium service (QoS cache). Therefore, in our setting, the IAP needs to determine how to allocate the available space to the different services.

We begin this section with an analysis of IAP caches as they are provisioned today (best effort). In §3.2, we analyze the equilibrium when a premium service is introduced by the IAP. The notation used in this section is summarized in Table 2. Note that in the table we calibrate three key parameters using trace-driven cache simulations. We simulated a cache with an LRU policy and request arrival at the cache was replicated from two benchmarked traces—Boeing and DEC (Web Characterization Repository 2002). Due to space constraints, we only summarize the results below. Interested readers are directed to the online appendix (available on the *Management Science* website at <http://mansci.pubs.informs.org/ecompanion.html>), which describes the empirical analysis in considerable detail.

(1) Distribution of requests for content: The number of requests for an object is denoted by R . We study traces to determine $f(R)$, the probability density function (pdf) of R . The pdf is given by $f(R) = \beta c^\beta / R^{1+\beta}$, where $R \in [c_1, c_2]$.

(2) Hit rate for cache: Hit rate denotes the fraction of requests answered by the cache. The hit rate of a cache varies as the logarithm of its size, S , when an LRU replacement policy is used, i.e., $H(S) = k_s \cdot \ln(S)$,

where k_s is a constant. The result is consistent with Breslau et al. (1999).

(3) Object specific hit rate: The object specific hit rate, $H(S, R)$, is the hit rate of an object with R requests in a cache of size S , i.e., the fraction of requests for that object that were satisfied by the cache. The object specific hit rate is given by $H(S, R) = k \cdot \ln(S) \cdot \ln(R)$, where k is a constant.

3.1. Single Best-Effort Service

A publisher derives some benefit from best-effort caches (bandwidth savings and faster delivery of content) but may incur a cost associated with compromising on security, business intelligence, and other value-added features. We use a piecewise separable benefit function

$$U = \theta \cdot (-q_L) + (Thc) \cdot (\eta + B). \quad (1)$$

The cost to the publisher of compromising on security and other value-added features is denoted by $\theta(-q_L)$. θ is a type parameter that denotes the weight the publisher attaches to value-added features for the specific object. This weight varies from publisher to publisher and even from object to object for a given publisher. For example, a data object containing confidential information may be associated with a high θ , whereas another not requiring any value-added features may correspond to an extremely low θ . We assume that θ is uniformly distributed in $[0, 1]$ across objects. We discuss the implications of this assumption at the end of this section. $(-q_L)$ denotes the level of quality of the value-added features, with the negative sign indicating that costs are being incurred. We will use the subscript L to denote the best-effort service and in the next section, we will use the subscript H for the premium service.

The publisher derives a benefit from being able to deliver her content faster to the end users. This benefit, represented by η and measured per hit, captures increased customer retention rates from faster delivery of an object. B denotes the average bandwidth savings realized by the publisher from a cache-based response to an object request. If R denotes the number of requests for an object and $H(S, R)$ denotes the object-specific hit rate, then the total hit count for the object is $Thc = RH(S, R)$. The benefit function captures the trade-off faced by the publisher—caching provides certain benefits related to bandwidth savings and faster content delivery but imposes costs, too. The specification of U assumes that the cost $\theta(-q_L)$ is independent of the number of requests or the hit rate of the object. In §4, we will explore correlations between these two components and demonstrate that this simple specification captures the primary features of the pricing problem.

Table 2 Glossary of Terms

Symbol	Explanation
q	Quality level of value-added features such as reporting and consistency
θ	Weight a publisher places on the quality for an object (type parameter)
S	Total size of cache
α	Fraction of cache space allocated to low-quality service
N	Total number of distinct objects
R	Number of requests for an object
$H(S)$	Hit rate for a cache of size $S = k_s \ln(S)$
$H(S, R)$	Hit rate for object with R requests in a cache of size $S = k \cdot \ln(S) \cdot \ln(R)$
Thc	Total hit count or total number of times an object with R requests is served from cache = $R \cdot H(S, R)$
B	Publisher's average bandwidth cost savings from delivering an object from cache = bandwidth cost per byte * average size (in bytes) of object
B_{IAP}	IAP's average bandwidth cost savings from delivering object from cache
η	Publisher's benefit from faster delivery of an object to an end user. This may come in the form of increased sales and advertising revenue
P	Price charged by the IAP for delivery of an object from cache
T	Marginal cost to the IAP of billing and metering
U	Publisher's surplus from caching the object
π	IAP's expected profit

We assume that the IAP charges a price P for every object served from this best-effort cache.² Thus, the net surplus to the publisher derived from caching an object of type θ and demand R is $U_L = \theta \cdot (-q_L) + R \cdot H(S, R) \cdot (\eta + B) - P \cdot R \cdot H(S, R)$. The publisher decides to cache if $U_L \geq 0$ and will cache bust otherwise. To determine the number of subscribers to the service, we consider a publisher with an object of type θ_i who is indifferent between caching and not caching (gets zero surplus from the caching service). By setting $U_L = 0$, we get

$$\theta_i = \frac{R \cdot \ln R \cdot k \ln S(\eta + B - P)}{q_L}.$$

Note that θ_i varies with R as shown in Figure 2. Any publisher with an object with $\theta > \theta_i(R)$ is more sensitive to value-added features and hence will not cache. Those with lower θ s will cache because they incur lower costs from caching but derive the same benefits as the indifferent publisher.

To compute the IAP's expected profit, we first determine the expected number of requests for objects that are cached by summing up the number of requests for all objects with $\theta \leq \theta_i$. This is denoted by R_L . Out of these R_L requests for objects, $R_L H(S)$ end up as hits (delivered from cache). The IAP's expected profit is thus given by

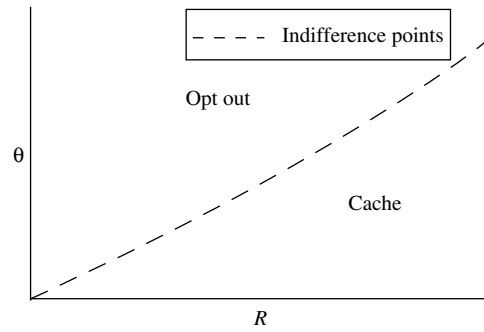
$$\pi = R_L \cdot H(S) \cdot (P + B_{\text{IAP}} - T). \quad (2)$$

For each object served from the cache, the IAP charges a price P . In addition, the IAP realizes bandwidth savings, B_{IAP} , from avoiding a request to the upstream backbone provider. Finally, the IAP incurs a marginal cost of metering and billing denoted by T . This represents costs associated with monitoring its caches, collection, customer-support costs, and additional accounting. For simplicity, we assume that these costs are linear in usage. If the IAP does not price the service, it incurs no such cost (i.e., $T = 0$ whenever $P = 0$). Summing all requests for objects with $\theta \leq \theta_i$,

$$\begin{aligned} \pi &= \left[N \int_{c_1}^{c_2} \int_0^{\theta_i(R)} Rf(\theta) d\theta f(R) dR \right] k_s \ln(S) \\ &\quad \cdot (P + B_{\text{IAP}} - T) \\ &= \left[\frac{Nk \ln S(\eta + B - P)k_1}{q_L} \right] \cdot k_s \ln(S) \\ &\quad \cdot (P + B_{\text{IAP}} - T), \end{aligned} \quad (3)$$

² Note that we use a per-object pricing scheme in this paper because of the prevalent pricing structure in the content delivery domain. The reader is referred to MacKie-Mason and Varian (1995) for a discussion on the merits of usage-based pricing for capacitated resources on the Internet.

Figure 2 Sample Indifference Points



where $k_1 = \int_{c_1}^{c_2} R^2 \ln R f(R) dR$ is a constant of integration. The IAP's decision problem is $\max_P \{\pi(P)\}$. Based on the first-order condition, the optimal price that the IAP should charge is

$$P^* = \frac{\eta + B + T - B_{\text{IAP}}}{2}. \quad (4)$$

The global concavity of the profit function can be easily verified. Note that high IAP bandwidth costs result in lower prices for the caching service. However, the IAP bandwidth cost will generally be lower than the publisher bandwidth cost due to volume discounts, i.e., $B_{\text{IAP}} < B$, and thus the IAP bandwidth cost alone does not account for the zero price for today's best-effort services. The indifference point associated with the optimal price is

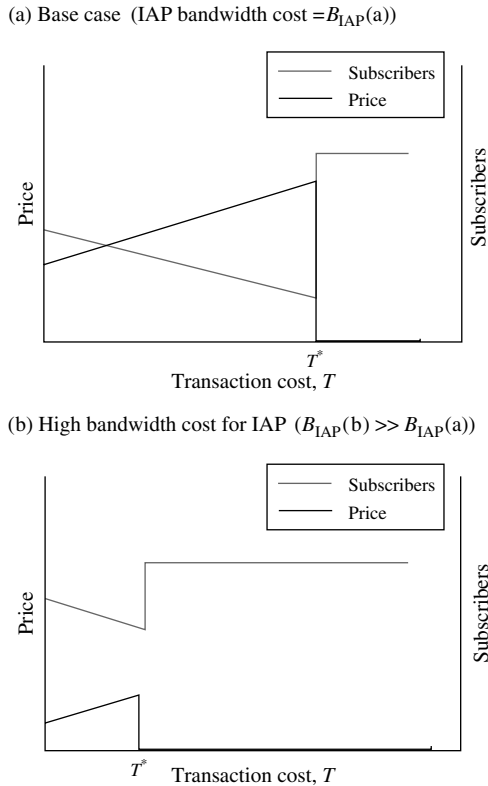
$$\theta_i(P^*) = \frac{R \cdot \ln R \cdot k \ln S(\eta + B + B_{\text{IAP}} - T)}{2q_L}.$$

This solution is valid for the case $\theta_i \leq 1 \forall R$. The boundary solution can be obtained along the same lines.

It can be inferred that the price increases and number of publishers that cache decreases as the transaction cost, T , increases. If $T = (\eta + B + B_{\text{IAP}})$, then $\theta_i = 0$, implying that there will be no publisher willing to pay for the caching services. At some transaction cost well below this level, the IAP would be better off setting the price to zero (T would also be zero) and maximizing its bandwidth savings instead. This occurs when $\pi(P=0) \geq \pi(P^*)$. Simplifying, if $T \geq (\eta + B + B_{\text{IAP}}) - 2\sqrt{(\eta + B)B_{\text{IAP}}} = T^*$, the optimal price for the IAP is $P^* = 0$ (see Figure 3a). Thus, one possible explanation for the free provisioning of today's best-effort service is that the transaction cost of metering and billing is greater than this threshold. The threshold may also be rewritten as $T^* = (\sqrt{\eta + B} - \sqrt{B_{\text{IAP}}})^2$. If IAP bandwidth costs are high, this threshold will be relatively low (Figure 3b). In other words, high IAP bandwidth costs will imply that the price can be zero for low (but nonzero) transaction costs as well.

To consider the impact of distributional assumptions, we solved the same model but introduced a

Figure 3 Impact of Transaction Cost and IAP Bandwidth Cost on Prices and Subscriptions



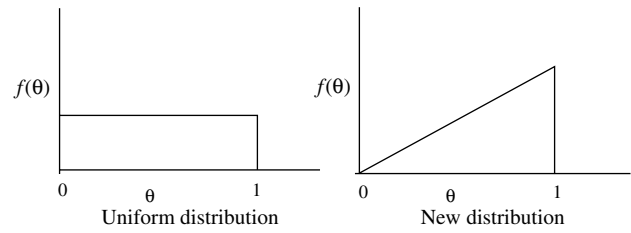
skew in the distribution of object types: $F(\theta) = \theta^2$ and $f(\theta) = 2\theta$. Relative to a uniform distribution, this distribution assumes that a larger fraction of publishers care about value-added features (see Figure 4). The new solution is

$$P^* = \frac{\eta + B + 2T - 2B_{IAP}}{3}$$

If $T \geq \eta + B + B_{IAP} - \sqrt[3]{(27/4)(\eta + B)^2 B_{IAP}}$, then $P^* = 0$. It can be verified that for lower values of T , the new price is lower than the optimal price in case of uniformly distributed valuations. Generally, the net impact of the negative skew in valuations is that the price charged reduces if it is not already zero.³ We also find that the price decreases with increasing IAP bandwidth costs. The higher the negative skew, the greater is this decrease. If $B_{IAP} = ((\eta + B)/2) + T$, then $P^* = 0$. In this case, the optimal price may be zero even with zero transaction costs. Thus, another explanation for the zero prices observed in reality is that

³ If $B_{IAP} > (729/1,024)(\eta + B)$, there is a small range of values with $T \in [\eta + B + B_{IAP} - 2\sqrt{(\eta + B)B_{IAP}}, \eta + B + B_{IAP} - \sqrt[3]{(27/4)(\eta + B)^2 B_{IAP}}]$, where the price in the case of the uniform distribution is zero but nonzero for the skewed distribution. However, the IAP bandwidth cost is so high under these settings that it turns out that the price for the skewed distribution is negative (the IAP pays publishers to induce them to cache). Thus, the impact of the skew is to reduce the price regardless.

Figure 4 Negative Skew in Distribution of Preferences



IAP bandwidth costs are high and a large number of publishers are sensitive to value-added features (have high θ). Even when prices are zero, several publishers will cache bust because the cost of caching ($-\theta q_L$) dominates the benefits for these publishers.

The model presented here highlights a number of reasons why prices may be zero today (high bandwidth costs for IAP, publisher sensitivity toward value-added features, nonzero transaction costs, or a combination of these factors). The implication of the model is that publishers have to trade off the benefits from caching with the loss of business intelligence and security, even in the absence of pricing. This trade-off results in cache busting by a large number of publishers. This further leads to loss of surplus for end users (slower delivery of content), IAPs (higher bandwidth costs), and publishers (unable to reap the benefits from caching). Recent trends suggest that publishers are becoming increasingly sensitive to business intelligence and personalization as online business models have started to evolve. Thus, adoption rates for caching will likely continue to decline. In the next section, we explore how the equilibrium changes when an additional premium caching service is also provided by the IAP.

3.2. Provision of Premium Service

In this section, we assume that the IAP provisions a premium service in addition to the best-effort service. The premium service offers a higher quality level by supporting dynamic content (personalization), object consistency, security, business intelligence, etc., and by providing premium objects with higher hit rates. Support for value-added features eliminates the cost incurred by the publisher from caching.

By using an appropriate priority scheme, the IAP can provide premium content with a higher hit rate. For example, Kelly et al. (1999) propose a scheme where different objects are assigned different weights and a server-weighted replacement policy is used to provide higher hit rates to objects with higher weights. Lu et al. (2001) achieve performance differentiation by dividing the cache space differentially between the content classes. Feldman and Chuang (2002) propose a multilevel replacement policy based on a number of interconnected LRU-based queues.

These authors use different techniques to achieve the same goal—providing different *effective* cache sizes to different content classes. For the purposes of our model, it does not matter which scheme is used to achieve differential hit rates. Our model prescribes the optimal *effective* cache sizes (or equivalently, optimal hit rates for the two content classes). Any of these schemes may be used by the IAP to achieve the differential hit rates. In the rest of the paper, the terms cache size or cache space will refer to the *effective* cache size/space without any reference to the underlying priority scheme.

3.2.1. Content Publisher’s Decision Problem. We assume that the IAP offers two services—a best-effort service and a premium service. The IAP charges a price, P_L , for every object served from its best-effort cache and charges a per-object price, P_H , for the premium service. We denote the level of value-added features for the premium service by q_H . The positive sign on the quality indicates a benefit to the publisher. This is because the IAP can provide the publisher with superior business intelligence than the publisher can gather on her own. For example, the IAP can provide valuable information about end users (type of Internet connection, user profile, etc.) or aggregate information across publishers. The publisher’s benefit function has the same form as in the previous section. Thus, the publisher’s net surplus from the two services is given by

$$\begin{aligned} U_L &= \theta \cdot (-q_L) + R \cdot H(S_L, R) \\ &\quad \cdot (\eta + B) - P_L \cdot R \cdot H(S_L, R), \\ U_H &= \theta \cdot q_H + R \cdot H(S_H, R) \cdot (\eta + B) \\ &\quad - P_H \cdot R \cdot H(S_H, R), \end{aligned} \quad (5)$$

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

The cache space, S , is divided into two levels, with α denoting the fraction of cache space allocated to the best-effort service. That is, αS is the size of the cache for best-effort subscribers and the remainder of size $(1 - \alpha)S$ is for the premium subscribers. To determine the number of subscribers to the service, we consider a publisher with object of type θ_L who is indifferent between the best-effort service and not subscribing to the service at all (gets zero surplus from the best-effort service). Any publisher with $\theta > \theta_L$ is more sensitive to value-added features and hence will not choose the best-effort service to cache the object. Objects with lower θ will cache in the best-effort service. Similarly, we consider a publisher with an object of type θ_H who is indifferent between the premium service and not caching. Objects associated with $\theta > \theta_H$ gain more benefits from the premium service and will be cached. Publishers with objects of type $\theta < \theta_H$ do not weigh

the value-added features enough to be willing to pay the price P_H . By setting $U_L = 0$, we get θ_L and by setting $U_H = 0$, we get θ_H :

$$\begin{aligned} \theta_L(R) &= \frac{R \ln R \cdot k \ln(S_L) \cdot (\eta + B - P_L)}{q_L}, \\ \theta_H(R) &= \frac{R \ln R \cdot k \ln(S_H) \cdot (P_H - \eta - B)}{q_H}. \end{aligned} \quad (6)$$

Both θ_L and θ_H vary with demand R and are thus curves that represent indifferent content publishers. We call these the quality indifference curves (QICs) for the publishers.⁴ A sample QIC, based on assumed prices and quality levels, is shown in Figure 5. Publishers in the region $\theta \in [\theta_L, \theta_H]$ care enough about the value-added features that they will not choose the best-effort service but not so much that they are willing to pay the premium price. Note that in Figure 5, the upper region representing premium service subscribers does not abut the lower region representing subscribers to the best-effort service. Lemma 2 in §3.2.2 formalizes this result.

3.2.2. Properties of the Quality Indifference Curves. This section presents some properties and observations regarding the QICs.

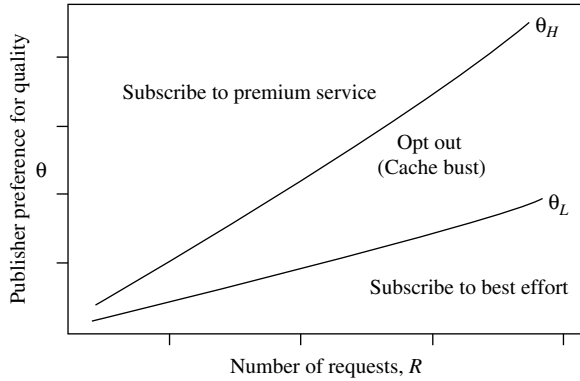
Well-behaved curves. One important property that we desire from the QICs is that they never cross each other. If they do, then interpreting the QICs becomes difficult.

LEMMA 1. *The QICs never cross each other.*

LEMMA 2. *If the IAP prices optimally, then $\theta_H(R) \geq \theta_L(R)$.*

The proofs for both lemmas are provided in the appendix. Lemma 2 implies that in equilibrium, no publisher derives positive surplus from both services. This is a unique feature of our setting. In traditional models, one has to verify that buyers choose the service which provides them with the maximum surplus, an incentive compatibility (IC) constraint. Additionally, the buyer buys only if the surplus from doing so is positive, which is an individual rationality (IR) constraint. However, in our setting, the participation (IR) constraint alone suffices because the prices are set in a manner such that only one service can potentially satisfy a publisher’s participation constraints.

⁴Quality indifference curves are not the same as indifference curves in economics. Traditional indifference curves denote how consumers trade off one good for another (thus indicating the marginal rate of substitution). QICs denote the points in space where publishers are indifferent between services. They are usually called indifference points in microeconomic models. Because these points vary with R , we refer to them as QICs.

Figure 5 Sample Indifference Points for Assumed Price and Quality Levels

Segmentation Conditions. The market is segmented if there are customers for both services. There exist subscribers to the best-effort service only if $\theta_L > 0$ for some R . Similarly, there exist subscribers to the premium service only if $\theta_H < 1$ for some R (see Figure 5). These two conditions are thus necessary conditions for segmentation.

(i) $\theta_H < 1$ for some $R \in (c_1, c_2)$ results in the following inequality:

$$P_H < \eta + B + \frac{q_H}{k \cdot \ln S_H c_1 \ln c_1}. \quad (7)$$

Equation (7) indicates that if the price is above the specified threshold, no publisher will choose the premium service.

(ii) $\theta_L > 0$ for some $R \in (c_1, c_2)$: If $\{\theta_L < 0 \forall R\}$, then there will be no subscriber for the best-effort service. This condition can be rewritten as $P_L < \eta + B$. This condition sets a simple upper bound for the price that the IAP can charge for the best-effort service.

LEMMA 3. *If the IAP chooses prices and allocation policy so that*

$$P_H < \eta + B + \frac{q_H}{k \cdot \ln S_H c_1 \ln c_1} \quad \text{and} \quad P_L < \eta + B,$$

then the market will be segmented.

It is clear from the discussion above that these two conditions are necessary. Their sufficiency is established in the appendix. Note that the optimal prices need not satisfy these conditions. Thus, whether it is desirable for the IAP to segment the market depends on whether the optimal prices satisfy Lemma 3.

3.2.3. IAP's Decision Problem. The IAP has a cache of size S . It allocates the space to the two services through its choice of α (fraction of space allocated to the best-effort service). In addition, the IAP also chooses the prices for the two services. As indicated in Lemma 2, the IAP will not choose prices that

set the θ_H QIC below the θ_L QIC. Hence, we only discuss the case $\theta_H(R) \geq \theta_L(R)$. 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 requests for all objects with $\theta > \theta_H$ (see Figure 5). 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 $\theta \leq \theta_L$. The IAP's expected profit is given by

$$\begin{aligned} \pi = & R_H H(S_H) [P_H + B_{\text{IAP}} - T_L] \\ & + R_L H(S_L) [P_L + B_{\text{IAP}} - T_H] - C(q_H). \end{aligned} \quad (8)$$

Out of the R_H requests for premium objects, $R_H H(S_H)$ end up as hits (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 realizes bandwidth savings, B_{IAP} , and charges a fee from the publisher. The IAP's marginal costs of metering and billing for the two services are denoted by T_L and T_H . We consider the general case where these marginal costs are not equal. There is a fixed cost, $C(q_H)$, of the infrastructure the IAP needs to provision the value-added services. Substituting for R_L and R_H ,

$$\begin{aligned} \pi = & N \left[E(R) - \frac{k k_1 (P_H - \eta - B) \ln S_H}{q_H} \right] \cdot k_s \ln S_H \\ & \cdot [P_H + B_{\text{IAP}} - T_H] + \frac{N k k_1 (\eta + B - P_L) \ln S_L}{q_L} k_s \ln S_L \\ & \cdot [P_L + B_{\text{IAP}} - T_L] - C(q_H), \end{aligned} \quad (9)$$

where k_1 is a constant of integration (see Appendix 2, for full derivation of expected profit). The IAP's decision problem is $\max_{P_H, P_L, \alpha} \pi(P_H, P_L, \alpha)$. Following the traditional approach (see, for example, Neven and Thisse 1990) in pricing, we model the IAP's problem as a two-stage process. In the first stage, we look at the pricing problem assuming that the allocation problem has been solved. That is, we address the pricing problem assuming that α is an exogenous parameter that has already been determined and hence the IAP is only interested in pricing. In the second stage, we analyze the allocation problem (determining optimal α). Note that this procedure maps well to reality where pricing follows production or service design.

3.2.4. Pricing Problem. The first-order conditions associated with the pricing problem are as follows:

$$\begin{aligned} \frac{\partial \pi}{\partial P_H} = & N \left[E(R) - \frac{k k_1 \ln S_H (P_H - \eta - B)}{q_H} \right] k_s \ln S_H \\ & - \frac{N k \cdot k_1 \ln S_H}{q_H} k_s \ln S_H [P_H + B_{\text{IAP}} - T_H] = 0. \end{aligned}$$

The first term in the derivative indicates the increase in revenue from being able to charge an

infinitesimal amount more for the premium service. The second term captures the decrease in revenue from subscribers opting out of the premium service due to this small increase in price.

$$\frac{\partial \pi}{\partial P_L} = \frac{Nkk_1 \ln S_L (\eta + B - P_L)}{q_L} k_s \ln S_L - \frac{Nkk_1 \ln S_L}{q_L} k_s \ln S_L [P_L + B_{IAP} - T_L] = 0.$$

The first term reflects the increased revenue from being able to charge more from the best-effort customers. However, some subscribers opt out of the service due to the increased price (second term). Solving the two first-order conditions, we have

LEMMA 4. When $T_L < (\eta + B + B_{IAP}) - 2\sqrt{(\eta + B)B_{IAP}}$, the optimal prices are

$$P_L = \left(\frac{\eta + B + T_L - B_{IAP}}{2} \right), \quad (10)$$

$$P_H = \frac{E(R)q_H}{2kk_1 \ln S_H} + \left(\frac{\eta + B + T_H - B_{IAP}}{2} \right).$$

When $T_L \geq (\eta + B + B_{IAP}) - 2\sqrt{(\eta + B)B_{IAP}}$, the optimal price for the best-effort service is $P_L^* = 0$ and P_H remains the same.

It can be verified that the Hessian is negative definite implying that the prices in Lemma 4 represent a maxima. Substituting the optimal prices from Lemma 4 into the segmentation conditions of Lemma 3, the following proposition is immediate.

PROPOSITION 1. If

$$T_H > (\eta + B + B_{IAP}) + \frac{q_H}{k \ln S_H} \left[\frac{2}{R_{LB} \cdot \ln R_{LB}} - \frac{E(R)}{k_1} \right],$$

then it is not optimal for the IAP to segment the market.

If transaction costs of metering and billing are exceedingly high, then it is obvious that a market for premium services may not exist. However, information technology (IT) has played an important role in reducing the transaction cost of metering and billing and can thus help facilitate such markets for premium Web-caching services. In addition, if the premium service provides a very high level of support for value-added features (q_H is high), a market for premium service will exist even at reasonably high transaction costs. Improvements in IT can also help facilitate an increase in q_H as has been witnessed in the last several years in caching technologies.

The QICs associated with the optimal prices of Equation (10) are

$$\theta_L^* = R \ln(R) \frac{k \ln S_L}{q_L} \left(\frac{\eta + B + B_{IAP} - T_L}{2} \right), \quad (11)$$

$$\theta_H^* = R \ln(R) \left\{ \frac{E(R)}{2k_1} - \left(\frac{\eta + B + B_{IAP} - T_H}{2} \right) \frac{k \ln S_H}{q_H} \right\}.$$

Special Case. Note that the first-order conditions in this section display no cannibalization effects. Hence, the optimization of P_L and P_H is done independently. Thus, this process may artificially push the θ_H QIC below the θ_L QIC (and may incorrectly count the subscribers in the intermediate region twice—once as subscribers of the best-effort service and once of the premium service). When this occurs, the optimal prices are at a boundary condition where $\theta_H(R) = \theta_L(R)$. The corresponding solution is given by

$$P_L = \eta + B - \frac{E(R)q_H q_L}{2k \cdot k_1 (q_L + q_H) \ln S_L} + \frac{\eta + B + B_{IAP} - T}{2} \cdot \left(\frac{\ln S_H - \ln S_L}{\ln S_L} \right) \left(\frac{q_L}{q_L + q_H} \right) \text{ and}$$

$$P_H = \eta + B + \frac{E(R)q_H^2}{2k \cdot k_1 (q_L + q_H) \ln S_H} - \frac{\eta + B + B_{IAP} - T}{2} \cdot \left(\frac{\ln S_H - \ln S_L}{\ln S_H} \right) \left(\frac{q_H}{q_L + q_H} \right).$$

Note that this special case represents the case where the market is completely captured (all publishers cache). This special case is not very realistic and thus we focus only on the interior solution in the rest of this paper.

3.2.5. Comparative Statics. Based on the QICs in Equation (11), we can conduct sensitivity analysis to determine the impact of improvements in the quality of value-added features on subscriptions. Improvements in quality will increase publisher surplus. The IAP can extract the additional value by increasing prices. However, the net impact on subscriptions depends on whether the price increases outpace increase in surplus for the indifferent publishers.

PROPOSITION 2. If the quality of the best-effort service is increased and the IAP reacts optimally with respect to prices, subscription to the best-effort service increases and to the premium service is unchanged.

PROOF. From Equation (11), it follows that an decrease in q_L (note that an increase in quality corresponds to a decrease in q_L) causes θ_L to increase and θ_H is not affected. This implies that the number of subscribers to the best-effort service increases and that to the premium service is unchanged. When $T_L \geq (\eta + B + B_{IAP}) - 2\sqrt{(\eta + B)B_{IAP}}$, the best-effort service is free ($P_L^* = 0$). Substituting this into Equation (6), we again find that θ_L increases with quality. Thus, subscription to the best-effort service increases with quality, irrespective of the transaction costs.

PROPOSITION 3. If the quality of the premium service is increased and the IAP reacts optimally, subscription to the best-effort service is not affected but to the premium service decreases, is unchanged, or increases depending on whether T_H is less than, equal to, or greater than $\eta + B + B_{IAP}$.

PROOF. From Equation (11), it follows that an increase in q_H has no effect on θ_L . Thus, the number of subscribers to the best-effort service is not affected. The impact on the premium service is readily verified for three cases.

There are two aspects worth noting in the propositions. First, there are no cannibalization effects. That is, changes in quality of one service do not impact the other, unlike traditional vertical differentiation models. This is because of the negative quality of the best-effort service that results in the no-subscription region being sandwiched between the premium and best-effort subscribers (see Figure 5). Thus, changes in parameters of any service affects that service but not the other. In contrast, there are direct effects to other services from changing any service parameter in classical segmentation models. In addition, we observe that the direction of the impact of increasing quality is different for the services. For the low-quality service, increase in quality consistently results in an increase in subscribers. This is because the IAP is unable to increase its price as the benefit from quality is still negative (i.e., $-\theta \cdot q_L < 0$). Thus, publisher surplus increases, resulting in an increase in the subscription base. On the other hand, price increases with quality for the premium service. Whether the price increase outpaces the benefit from quality increase for the indifferent publisher at θ_H depends on the relative magnitude of the hit-based benefits and transaction costs. The IAP may lose subscribers but earn higher margins per subscriber when benefits from caching are high and transaction costs are low.

PROPOSITION 4. *As bandwidth costs decrease, subscriptions to both the services decrease.*

PROOF. As bandwidth costs drop, both B and B_{IAP} decrease in Equation (11). This results in an increase in θ_H and a decrease in θ_L . Thus, subscriptions drop for both the services.

Equation (10) indicates that the IAP charges the content publisher a part of her surplus from bandwidth reduction and faster content delivery ($\eta + B$). However, the IAP gives back to the publisher a part of its own surplus from bandwidth reduction (B_{IAP}). If the IAP bandwidth costs are high and there are a large number of publishers sensitive to value-added features (negative skew in θ) or transaction costs are nonzero, the best-effort service will be free. The price for the best-effort service is the same as in the single-service case. Hence, we expect best-effort services to remain free even when a premium service is introduced. Further, the best-effort QIC is the same as in the single-service case (i.e., there is no cannibalization), so the number of subscribers to the best-effort service remains the same.

The IAP also charges the publisher for the support provided to value-added features (denoted by q_H). The price charged varies linearly with the quality level. This linearity is largely driven by the fact that our model assumes that the publisher's surplus varies linearly with q_H . Nonlinear prices result if we assume a nonlinear surplus function. For example, assuming the following surplus function:

$$U_H = \theta \cdot q_H - Cq_H^2 + Thc(S_H, R) \cdot (\eta + B) - P_H \cdot Thc(S_H, R),$$

where C is a normalization constant, results in the following nonlinear optimal price:

$$P_H = \frac{E(R)q_H(1 - Cq_H)}{2kk_1 \ln S_H} + \left(\frac{\eta + B + T_H - B_{IAP}}{2} \right).$$

Space constraints prevent us from providing a detailed derivation. Interested readers may contact the authors. There also exists some nonlinearity with regard to impact of cache sizes. Equation (10) indicates that the per-object price for the premium service decreases with increasing cache size. This is analogous to quantity discounts in conventional pricing theory. The total price charged to a publisher for caching an object is $P_H \cdot R \cdot k \ln R \ln S_H$, which is increasing in cache size.

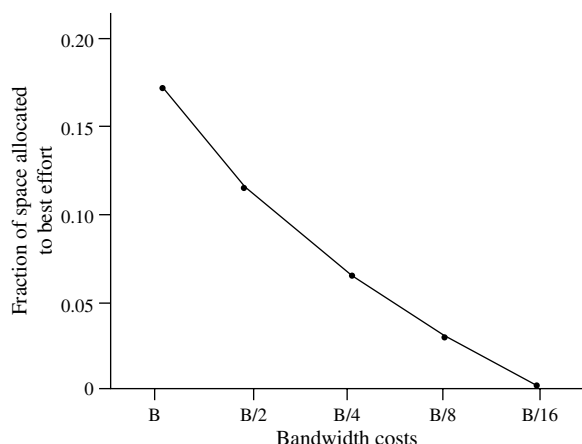
3.3. Space Allocation

Caches sizes are typically optimized based on traffic profiles and are seldom over-provisioned. This is due to diminishing returns from larger cache sizes and costs dominating 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 space allocation is an important consideration.

The first-order condition with respect to α does not yield a closed-form solution, although the existence can be guaranteed. The proof is presented in the online appendix. To illustrate how the IAP may solve the allocation problem, we consider a numerical example below.

3.3.1. Illustrative Example. We simulate values for the various parameters in the model. We consider an average publisher with bandwidth cost of 0.03c per object. This corresponds to a T1 connection priced at \$750 per month, an average object of size 50 KB, and a peak to average bandwidth ratio of 4:1. The IAP handles more traffic and hence would have lower bandwidth costs. Assuming that the IAP uses an OC-48 connection, this gives us IAP bandwidth cost of 0.011c per object, approximately 36% of that of the typical publisher. We assume that the publisher's benefit from faster delivery of content is

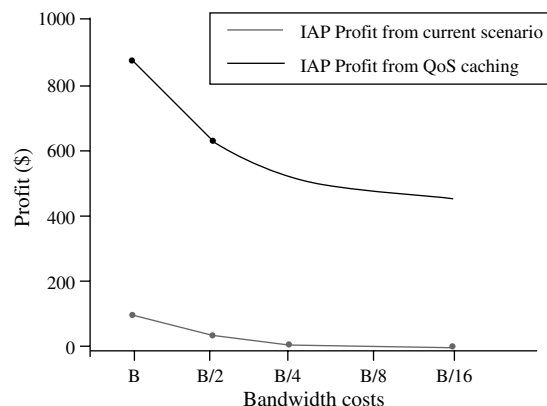
Figure 6 Impact of Decreasing Bandwidth Costs on Allocation



of the same order of magnitude as bandwidth savings, i.e., $\eta = 0.03c$. We calibrate values for q_L and q_H by considering $\theta \cdot q$ as the dollar cost/benefit of value-added features (note that $R \cdot H(S, R) \cdot B$ is the dollar value of bandwidth savings). The publisher who is most sensitive to value-added features ($\theta = 1$) is assumed to value these features an order of magnitude more than the bandwidth savings ($q_H = 0.3$). q_L is assumed to be 0.4 with the implication that the cost to the most sensitive publisher is $0.4c$ per object. The transaction cost of billing for both services is set as $T = \eta + B + B_{IAP} = 0.071c$ per object. This sets the cost high enough so that the best-effort service will be free (note that we could also have set higher IAP bandwidth cost and a skewed distribution for θ to achieve the same result). The cache size is assumed to be 6 GB. All the remaining parameters were empirically derived from the Boeing trace (Web Characterization Repository 2002). For these settings, the optimal solution is to allocate 17.32% of the cache space to the best-effort service (i.e., $\alpha^* = 0.1732$). When publisher and IAP bandwidth costs are reduced to half their previous values ($B = 0.015c$ and $B_{IAP} = 0.0055c$), the IAP allocates 10.9% of the cache to the best-effort service. Figure 6 lays out the impact of lowering bandwidth costs on α^* . In each successive simulation, we halve the bandwidth costs from the previous simulation (both B and B_{IAP} are halved). It is optimal for the IAP to reduce the size of the best-effort cache and increase that of the premium cache as bandwidth costs decline.

In addition, an IAP's profit from caching also decreases when bandwidth costs fall.⁵ In Figure 7, the upper curve indicates the IAP's profits from QoS caching for different bandwidth prices. The lower curve plots the IAP's profits if it does not pursue

Figure 7 Impact of Decreasing Bandwidth Costs on IAP Profit (in Dollars/Day)

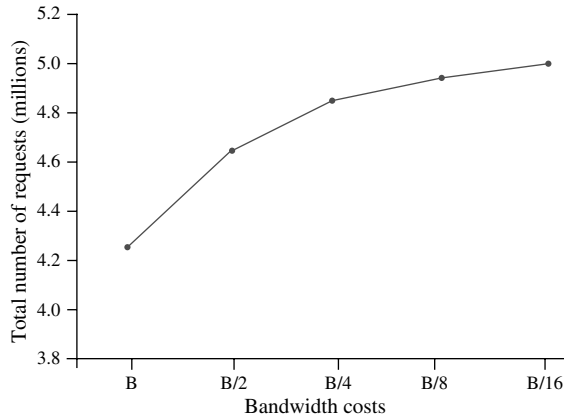


any QoS caching policies (IAP's surplus from caching consists only of its bandwidth savings). The figure does not account for other possible trends such as reduction in end user access charges facilitated by lower bandwidth costs that might ultimately result in higher demand for content by end users (both the type of content requested and number of requests may change). While it is difficult to ascertain the exact nature or magnitude of these changes, we can, however, determine the rate at which traffic will need to increase to maintain IAP profits from caching at the same level. As shown in Figure 8, a near-linear traffic growth is needed to sustain IAP profits from QoS caching at the same level (note that the x axis in Figure 8 is on a log scale for consistency with prior graphs). Such an increase may in fact be feasible in the short term with greater adoption of broadband services and a shift towards rich content such as multimedia. In Figure 9, we have plotted the bandwidth costs and traffic handled by backbone providers in January of 2000, 2001, and 2002. For example, in 2000, bandwidth cost for an OC3 leased line from Chicago to Los Angeles was \$400,000 per annum and backbone traffic was 8 petabytes/month (100%).⁶ In 2001, these figures were \$200K and 23 petabytes (287.5%). Because the bandwidth costs for the base case in our numerical analysis are based on bandwidth prices in 2002, we have overlaid traffic growth and bandwidth costs from Figure 8 on Figure 9 as well (dotted line). As is clear, the required traffic growth rates are lower than historic growth between 2000–2002. Thus, the shift towards rich content and greater broadband penetration that has fueled the growth in traffic between 2000–2002 may help sustain returns from caching in the near future.

⁵ This effect can also be identified by applying the envelope theorem to Equation (5).

⁶ Bandwidth costs were obtained from www.telegeography.com and Internet traffic statistics from <http://www.cibcwm.com/conferences/hour/foundingfather/roberts.pdf>.

Figure 8 Increase in Traffic Needed to Maintain IAP Profits with Declining Bandwidth Costs

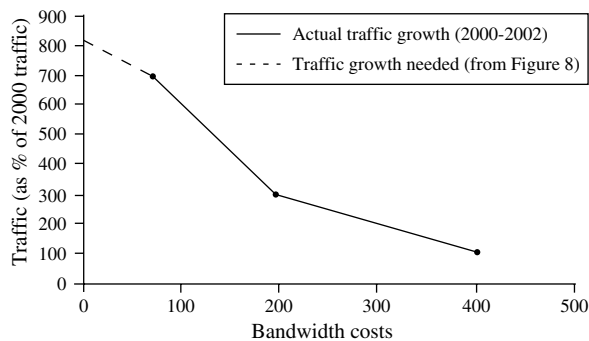


4. Robustness of the Model

We conclude our discussion of the model with a review of our modeling assumptions and the robustness of our analytical insights. In the model, we account for heterogeneity among objects in terms of demand but ignore size related differences (which would bring in heterogeneity in terms of bandwidth costs and speedup benefits as well). While heterogeneity in object sizes impact the efficiency of caching policies, they did not seem to critically affect the pricing strategies because the price is usage based. Accounting for bandwidth costs and speedup benefits for an average data object enables us to keep the model tractable and derive valuable broader insights.

Bandwidth is currently priced using either a usage-based model or capacity-based model. Examples of the former include ATM and frame relay-based services. Leased line services (T1, T3, and OC3) employ the capacity-based model in that a fixed monthly charge is paid for guaranteed bandwidth. It is clear that a single model cannot capture both usage-based pricing and pricing for capacity. Hence, we consider average bandwidth cost per object served from the cache. The average bandwidth cost is clearly nonzero

Figure 9 Actual Traffic Growth vs. Required Growth to Sustain QoS Cache Profits



and represents publisher’s bandwidth considerations rather well. Furthermore, recent trends in bandwidth pricing, such as burstable packages for leased lines, have focused on usage-based metering.

The model presented in §3 incorporates transaction costs of contracting for the IAP but not for the publisher. The model incorporating publisher transaction costs is relatively tedious but the insights are straightforward. We summarize the main points here. Contracting costs reduce publishers’ surplus from the service and thus they only have incentives to contract with large IAPs, whereas benefits from the service can outweigh the contracting costs. Thus, we expect that such QoS services can be rolled out successfully by large IAPs only. For smaller IAPs, an aggregator can play an important role in reducing the transaction costs for publishers. Content delivery networks (CDNs) can play an important role in facilitating such markets. Publishers will only need to contract with one CDN, which in turn can contract with a large number of IAPs. Thus, whether an IAP markets QoS caching services independently or through an aggregator may depend on contracting costs.

The model in §3 assumed that the QICs ($\theta_i, \theta_L, \theta_H$) are always less than or equal to 1 for all R . However, there may be situations where these QICs reach 1 for some $R \in [c_1, c_2]$. When that occurs, the solutions in §3 are no longer valid. However, we have analyzed this boundary condition and found that the qualitative nature of our results does not change. The primary difference in the results is that the optimal price of the best-effort service now depends on the cache size (S) and the quality of the best-effort service (q_L) at the boundary, whereas it is independent of these parameters in the interior. A detailed analysis of the boundary condition is provided in the online appendix.

The publisher’s benefit function in §3 assumed that the benefit/cost from value-added features is independent of the number of requests for an object. However, a publisher may value security or business intelligence more for an object that is in relatively higher demand. Hence, we consider two variants of the publisher surplus function. The first is: $U = \theta Rq' + R \cdot \ln R(\eta + B) - P \cdot R \cdot \ln R$. This surplus function assumes that each object has an intrinsic requirement (or lack thereof) for value-added features, denoted by θ . However, this function also assumes that given two objects with the same θ , the publisher values the premium service more for the relatively popular object. The impact of the change is that the slopes of both the QICs in Figure 7 decrease. The per-object price charged for the best-effort service is not affected and that for the premium service decreases. Broader insights from the model continue to hold.

Next, we let the benefit/cost from the value-added features increase at a faster rate with R than the hit-rate benefits by assuming that

$$U_L = \theta \cdot R^2 \cdot (-q_L'') + R \cdot k \ln S_L \cdot \ln R \cdot (\eta + B) - P_L \cdot R \cdot k \ln S_L \cdot \ln R,$$

$$U_H = \theta \cdot R^2 q_H'' + R \cdot k \ln S_H \cdot \ln R \cdot (\eta + B) - P_H \cdot R \cdot k \ln S_H \cdot \ln R.$$

q_L'' and q_H'' are rescaled quality levels such that $\theta \cdot R^2 \cdot q''$ represents the benefit or cost derived from quality level q'' . This transformation ensures consistency across the two models. In this new specification as well, the slopes of the two QICs decrease. However, the impact is strong enough to cause the QICs to slope downwards as illustrated in Figure 10. Downward sloping QICs imply that more popular objects prefer the premium service. The optimal prices are

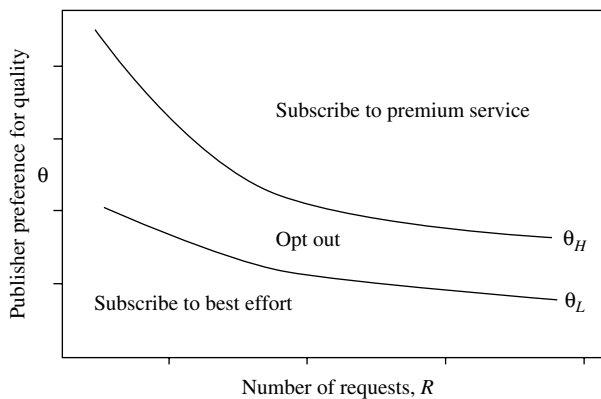
$$P_L = \left(\frac{\eta + B + T_L - B_{IAP}}{2} \right) \text{ and}$$

$$P_H = \frac{E(R)q_H''}{2kk_1 \ln S_H} + \left(\frac{\eta + B + T_H - B_{IAP}}{2} \right).$$

It can be shown that the price of the premium service decreases (note that the price functions have the same form as in §3) from that in §3, whereas that of best-effort service remains the same.

We can expect reality to lie between the cases of complete independence (Figure 5) and very strong correlation (Figure 10). In fact, when the benefit from value-added features varies linearly with the number of hits, the QICs are horizontal lines and the distribution of R becomes irrelevant. The net impact of assuming a correlation between number of requests and benefit from value-added features is that slope of the QICs and the magnitude of price for premium service may change but the broader insights from the model in §4 (such as Propositions 1–3, impact on profit and allocation, etc.) are not altered.

Figure 10 Sample QICs for the New Model (Assumed Price and Quality Levels)



5. Conclusions

Quality of service (QoS) is the leading performance consideration in e-business today. We introduce a framework to analyze the QoS issues in Web caching. If designed prudently, QoS caching can 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 great interest to e-marketing—to the edge of the network as well. Thus, content publishers 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.

Best-effort caching worked well when the Web consisted primarily of static, nonpersonalized content. As e-markets have matured over the last few years, publishers have developed new requirements which current caching services do not meet resulting in significant cache busting. Our paper discusses a QoS framework to determine optimal pricing and capacity allocation policies for an IAP provisioning best-effort and premium-caching services to respond to the new needs of publishers.

Starting with a single class of caching service, we found cost conditions under which the profit-maximizing price is zero. Specifically, IAPs offer the service for free because their bandwidth costs are high and a large number of publishers are sensitive to value-added features. Yet, some publishers will still opt out of the service, i.e., cache bust, as observed in practice. We next derived the conditions under which the profit-maximizing cache operator should offer two vertically differentiated service classes, namely, premium and best effort. Interestingly, caching service differentiation is different from the traditional vertical differentiation model, in that the premium and best-effort market segments do not abut. Thus, optimal prices for the two service classes can be set independently, and cannibalization does not occur. Indeed, it is possible for the cache operator to continue to offer the best-effort service for free while charging for the premium service. In this new scheme, publishers are better off and cache more content, as are IAPs who reap higher bandwidth savings and are able to generate revenues from the premium service. Furthermore, consumers benefit from the reduced cache busting by publishers.

We find that subscriptions to caching services will drop with falling bandwidth costs. This effect will be mitigated by an increase in the traffic on the network. Increased broadband penetration and a shift towards multimedia content can help facilitate the same. Based on traffic growth in the recent past, the increase in

traffic needed to maintain profits from caches at the same level may be feasible.

Our analysis has also shown that changes in publisher preferences will diminish the role of best-effort caching services. Declining bandwidth costs further reduce their relevance. Thus, managers are better off directing their resources towards provisioning value-added services. This finding is also corroborated by recent articles in the business press (Mears 2002). Services like Akamai’s Edgestuite that enable delivery of entire sites from the edge caches, bundled with business intelligence and content targeting, may well become the norm. This is an indication of the impending metamorphosis of the content delivery value chain.

An electronic companion to this paper is available on the *Management Science* website (<http://mansci.pubs.informs.org/ecompanion.html>).

Acknowledgments

This research was funded in part by NSF CISE/IIS/KDI 9873005 and ITR 0085879. The authors thank seminar participants at the University of Washington, Purdue University, University of Rochester, New York University, Penn State, Tulane University, and The Wharton School for their feedback and comments. They acknowledge the research assistance provided by Kim Norlen.

Appendix 1

PROOF OF LEMMA 1. Equating θ_L and θ_H leads to the following equation:

$$\frac{(P_H - \eta - B) \ln(S_H)}{q_H} \cdot (R \ln(R)) = \frac{(\eta + B - P_L) \ln(S_L)}{q_L} (R \ln(R)).$$

Note that if $\theta_L = \theta_H$ for any R in $[c_1, c_2]$, then either $\ln(R) = 0$ or

$$\frac{(P_H - \eta - B) \ln(S_H)}{q_H} = \frac{(\eta + B - P_L) \ln(S_L)}{q_L}.$$

The former is the simple and uninteresting case where $R = 1$ (the object has only one request). The latter implies that $\theta_L = \theta_H$ for all R in $[c_1, c_2]$. That is, if θ_L and θ_H ever meet, they are always equal (market is exactly captured), thus ensuring that the two QICs never cross each other. The quality indifference curves are therefore always “well behaved.”

PROOF OF LEMMA 2. Let us assume that the converse is true. That is, $\theta_H(R) < \theta_L(R)$ for the optimal prices. In this case, the entire market is captured for the prices chosen (i.e., everyone derives positive utility from at least one service). In fact, the region $\theta \in (\theta_H, \theta_L)$ represents subscribers that derive a positive surplus from both services. We define θ_{LH} as the object whose publisher is indifferent between caching it in the premium service or the best-effort service. By setting $U_L = U_H$, we get

$$\theta_{LH}(R) = \frac{R \ln R \{ \ln S_L (\eta + B - P_L) + \ln S_H (P_H - \eta - B) \}}{q_L + q_H}.$$

This can be rewritten as $\theta_{LH} = (\theta_L q_L + \theta_H q_H) / (q_L + q_H)$. Thus, θ_{LH} is a weighted average of θ_L and θ_H , implying that the θ_{LH} QIC lies between the other two QICs. Publishers

with objects of type $\theta > \theta_{LH}$ will join the premium service because they weigh the value-added features more; those with $\theta < \theta_{LH}$ will choose the best-effort service.

Publishers in the region (θ_H, θ_{LH}) all subscribe to the best-effort service and yet derive positive surplus from joining the premium service. Similarly, publishers in the region (θ_{LH}, θ_L) all subscribe to the premium service and yet derive positive surplus from joining the best-effort service. Under this scenario, the IAP can increase the prices of the two services by the same amount, which causes θ_H to move up and θ_L to move down without impacting θ_{LH} . The IAP can charge more without impacting its subscriptions and thus increase its profits. Hence, the original prices cannot be optimal.

PROOF OF LEMMA 3. Lemma 2 establishes that $\theta_H(R) \geq \theta_L(R)$. Now, if $\theta_L(R) > 0$, we know that all publishers located between $[0, \theta_L]$ will subscribe to the best-effort service because they derive a positive surplus from it but negative surplus from the premium service (see Figure 7). Thus, $\theta_L(R) > 0$ for some value of R guarantees the existence of subscribers for the best-effort service. The condition is sufficient. Similarly, publishers with $\theta > \theta_H$ will derive a positive surplus from the premium service. Because $\theta_H \geq \theta_L$, all these publishers will also derive a negative surplus from the best-effort service. Thus, publishers located between $[\theta_H, 1]$ will all subscribe to the premium service. If $\theta_H < 1$ for some R , it is guaranteed that the premium service also has some subscribers. It follows that the market is segmented.

Appendix 2. IAP Profit Function

The IAP’s expected profit consists of revenues from charging for the two services and bandwidth savings from the cache. As explained in §3.2.3, this is given by $\pi = R_H H(S_H) [P_H + B_{IAP} - T_H] + R_L H(S_L) [P_L + B_{IAP} - T_L] - C(q_H)$, where R_H is the expected number of requests for objects in the premium service and R_L is the expected number of requests for objects in the best-effort service. R_H and R_L are obtained by summing up requests for objects of type $\theta > \theta_H$ and $\theta \leq \theta_L$, respectively. The number of objects requested R times is given by $Nf(R)$. Thus, these objects constitute a total of $NRf(R)$ requests. The fraction of these requests that are for content in the best-effort service is $\int_0^{\theta_L(R)} f(\theta) d\theta = \theta_L(R)$. Thus, the total number of requests for content in the best-effort service is given by summing $N\theta_L(R)Rf(R)$ for all values of R : $R_L = N \int_{c_1}^{c_2} \theta_L(R)Rf(R) dR$. Substituting the expression for $\theta_L(R)$ gives us $R_L = (Nkk_1(\eta + B - P_L) \ln S_L) / q_L$, where $k_1 = \int_{c_1}^{c_2} R^2 \cdot \ln R \cdot f(R) dR$ is a constant of integration. Similarly,

$$R_H = N \int_{c_1}^{c_2} (1 - \theta_H(R))Rf(R) dR = N \left[E(R) - \frac{kk_1(P_H - \eta - B) \ln S_H}{q_H} \right],$$

where $E[R]$ is the expected value of R . Substituting these expressions for R_H and R_L into the profit function, we get

$$\pi = N \left[E(R) - \frac{kk_1(P_H - \eta - B) \ln S_H}{q_H} \right] \cdot k_s \ln S_H \cdot [P_H + B_{IAP} - T_H] + \frac{Nkk_1(\eta + B - P_L) \ln S_L}{q_L} k_s \ln S_L \cdot [P_L + B_{IAP} - T_L] - C(q_H).$$

References

- Bhargava, H. K., V. Choudhary, R. Krishnan. 2000. Pricing and product design: Intermediary strategies in an electronic market. *Internat. J. Electronic Commerce* 5(5) 37–56.
- Breslau, L., P. Cao, L. Fan, G. Phillips, S. Shenker. 1999. Web caching and Zipf-like distributions: Evidence and implications. *Proc. IEEE Infocom*, IEEE, New York, 126–134.
- Brouwer, L. E. J. 1910. Über Abbildung von Mannigfaltigkeiten. *Math. Ann.* 71 97–115.
- Chan, Y. M., J. Womer, J. K. MacKie-Mason, S. Jamin. 1999. One size doesn't fit all: Improving network QoS through preference-driven web caching. *Proc. 27th Annual Telecomm. Policy Res. Conf.*, Alexandria, VA.
- Chuang, J., M. Sirbu. 2000. Distributed network storage with quality-of-service guarantees. *J. Network Comput. Appl.* 23(3) 163–185.
- Chuang, J., S. Kafka, K. Norlen. 2002. Efficiency and performance of web cache reporting strategies. *Proc. IEEE Internat. Workshop on Data Semantics in Web Inform. Systems*, Singapore, 120–129.
- Cocchi, R., S. Shenker, D. Estrin, L. Zhang. 1993. Pricing in computer networks: Motivation, formulation and example. *IEEE/ACM Trans. Networking* 1 614–627.
- Datta, A., K. Datta, H. Thomas, D. VanderMeer. 2003. WORLD WIDE WAIT: A study of Internet scalability and cache-based approaches to alleviate it. *Management Sci.* 49(10) 1425–1444.
- Datta, A., K. Dutta, H. Thomas, D. VanderMeer, Suresha, K. Ramamritham. 2002. Proxy-based acceleration of dynamically generated content on the World Wide Web: An approach and implementation. *Proc. ACM SIGMOD*, ACM Press, Madison, WI, 97–108.
- Dewan, R., M. Freimer, A. Seidmann. 2000. Organizing distribution channels for information goods on the Internet. *Management Sci.* 46(4) 483–495.
- Feldman, M., J. Chuang. 2002. Service differentiation in web caching and content distribution. *Proc. IASTED Internat. Conf. Comm. Comput. Networks*, Cambridge, MA.
- Gupta, A., D. O. Stahl, A. B. Whinston. 1997. Priority pricing of integrated services networks. Mcknight, Bailey, eds. *Internet Economics*. MIT Press, Cambridge, MA, 323–352.
- Kelly, T., D. Reeves. 2000. Optimal web cache sizing: Scalable methods for exact solution. Presented at 5th Internat. Conf. Web Caching Content Delivery, Lisbon, Portugal, 163–173.
- Kelly, T., S. Jamin, J. K. MacKie-Mason. 1999. Variable QoS from shared web caches: User-centered design and value-sensitive replacement. MIT Workshop on Internet Service Quality Economics, Cambridge, MA.
- Lu, Y., A. Saxena, T. F. Abdelzaher. 2001. Differentiated caching services; A control-theoretical approach. Presented at Internat. Conf. Distributed Comput. Systems, IEEE, Phoenix, AZ, 615–624.
- MacKie-Mason, J., H. R. Varian. 1995. Pricing congestible network resources. *IEEE J. Selected Areas Comm.* 13(7) 1141–1149.
- Maggs, B. 2002. Vice President, Akamai. Personal communication.
- Marchand, M. 1974. Priority pricing. *Management Sci.* 20(7) 1131–1140.
- Mears, J. 2002. CDNs are not just for content anymore. *Network World* (January 14), www.networkworld.com/news/2002/0114specialfocus.html.
- Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* 38 870–883.
- Mogul, J., P. Leach. 1997. Simple hitmetering and usage-limiting for HTTP. RFC 2227, www.faqs.org/rfcs/rfc.2227.html.
- Mookerjee, V., Y. Tan. 2002. Analysis of a least recently used cache management policy for web browsers. *Oper. Res.* 50(2) 345–357.
- Mussa, M., S. Rosen. 1978. Monopoly and product quality. *J. Econom. Theory* 18 301–317.
- Myers, A., J. Chuang, U. Hengartner, Y. Xie, W. Zhuang, H. Zhang. 2001. A secure, publisher-centric web caching infrastructure. *Proc. IEEE INFOCOM 2001*, Anchorage, AK, 1235–1243.
- Neven, D., J. F. Thisse. 1990. Quality and variety competition. J. Gabszewicz, J. Richard, L. Wolsey, eds. *Economic Decision Making: Games, Econometrics and Optimisation*. North-Holland, Amsterdam, The Netherlands, 175–199.
- Smith, B., A. Acharya, T. Yang, H. Zhu. 1999. Exploiting result equivalence in caching dynamic web content. *Proc. USENIX Sympos. Internet Tech. Systems*, USENIX Association, Bolder, CO.
- Stargate Inc. Now part of Earthlink Inc. 2002. Personal communication, Pittsburgh, PA.
- Sundararajan, A. 2004. Non-linear pricing of information goods. *Management Sci.* 50(12) 1660–1673.
- Web Characterization Repository. 2002. World Wide Web consortium. Retrieved October 2000 from http://repository.cs.vt.edu/.
- Yin, J., L. Alvisi, M. Dahlin, C. Lin. 1998. Using leases to support server-driven consistency in large-scale systems. *Proc. 18th Internat. Conf. Distributed Systems*, Amsterdam, The Netherlands, IEEE, 285–294.
- Yu, H., L. Breslau, S. Shenker. 1999. A scalable web cache consistency architecture. *Proc. ACM SIGCOMM*, ACM Press, Cambridge, MA, 163–174.
- Zerbe, R., H. McCurdy. 2000. The end of market failure. *Regulation* 23(2) 10–14.