University of Pennsylvania
ScholarlyCommons

Operations, Information and Decisions Papers | Wharton Faculty Research

12-2005

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

kartik Hosanagar
University of Pennsylvania

Ramayya Krishnan

John Chuang

Vidyanand Choudhary

Follow this and additional works at: https://repository.upenn.edu/oid_papers

 Part of the Marketing Commons, and the Operations and Supply Chain Management Commons

Recommended Citation

Hosanagar, k., Krishnan, R., Chuang, J., & Choudhary, V. (2005). Pricing and Resource Allocation in Caching Services With Multiple Levels of Quality of Service. *Management Science, 51* (12), 1844-1859. http://dx.doi.org/10.1287/mnsc.1050.0420

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/oid_papers/154
For more information, please contact repository@pobox.upenn.edu.

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

Abstract

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.

Keywords

web caching, content delivery, pricing, capacity allocation, quality of service (QoS)

Disciplines

Marketing | Operations and Supply Chain Management

This journal article is available at ScholarlyCommons: https://repository.upenn.edu/oid_papers/154

Pricing and Resource Allocation in Caching Services with Multiple Levels of QoS

Kartik Hosanagar[+], Ramayya Krishnan[+], John Chuang[++] and Vidyanand Choudhary[+++]

Abstract

Internet infrastructure consists of backbone networks, access networks, Content Delivery Networks and other cache operators. Caches are the storage centers in the supply chain for content delivery - the digital equivalent of warehouses. The benefits of caching to content publishers, namely scalable content delivery, reduction in bandwidth costs and improvements in response times, are well recognized. Yet, caching has not been fully embraced by content publishers since its use can interfere with site personalization strategies or result in loss of information about visitors to the site. Recent work on web caching has focused on the technological advances required to address these deficiencies. However, there has been no work on the managerial issues related to the design of incentive compatible caching services, appropriate pricing schemes and associated resource allocation issues. The primary question this paper investigates is – *Why have adoption rates for caching been rapidly declining and what can cache operators do to respond to these changes and prevent market failure?* We propose a framework to enable an access network provider to provision vertically differentiated caching services that we refer to as "QoS Caching". We demonstrate analytically how provisioning of such services can increase profits of cache operators as well as overall market welfare. The analytic models also investigate the optimal pricing and capacity allocation policies of cache operators. The primary contribution of the paper is to propose an efficient mechanism to align the incentives of cache operators and content publishers and prevent failure of markets for caching services.

Keywords: Web caching, Content Delivery, Pricing, Capacity allocation, Quality of Service (QoS).

+ Carnegie Mellon University
++ University of California, Berkeley
+++ University of California, Irvine

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

1 Introduction

e-business is a large and growing part of overall commerce conducted today[1]. The growth of content and applications on the Internet has been phenomenal and is likely to continue. These observations highlight the importance of Internet infrastructure as a key enabler of e-business. The infrastructure consists of the following players intermediating between the end user and the content publisher:

1. Internet Access Providers (IAPs) such as AOL and Earthlink that provide retail-level Internet access to 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 provide wholesale–level Internet access to IAPs. Backbone networks interconnect with one another through Network Access Points (NAPs) and bi-lateral peering points to form the Internet backbone.
4. Content Delivery Networks (CDNs) such as Akamai and Digital Island (now part of Cable & Wireless) that deliver content on behalf of content publishers using proprietary networks of caching servers.

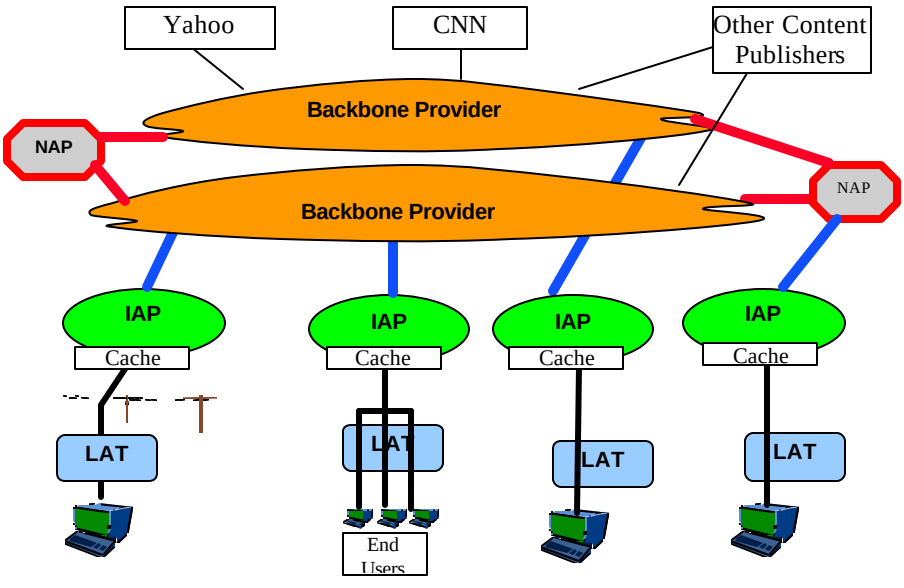

Figure 1. The Internet Industry Structure

[1] See "How big is e-commerce" (http://www.newsfactor.com/perl/story/18403.html) for some projections.

Figure 1 illustrates the interaction between the players on the Internet. The end users obtain access to the content through the access networks (IAPs). The IAPs in turn connect to the backbone networks and pay them for bandwidth. Content publishers such as Yahoo and CNN typically connect to the backbone providers or host their content at data centers of the backbone providers. Thus, publishers also pay the backbone networks for bandwidth consumed (either directly or pay their hosting providers who in turn pay backbones). Due to traffic and congestion on the network, latency and bandwidth costs can be high. Therefore, IAPs and CDNs often store local copies of content on caches located at the edge of the network.

Each of these players 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 publishers, 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 publisher 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 will have on the digital supply chain and therefore to e-business. This is underscored in IDC's projection that the caching market would be worth $4.5 billion in 2004.

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. If there is another request for the cached content (a web page), the local copy may be returned instead of requesting the origin server for it again. Cache performance is measured by its hit-rate, which denotes the fraction of requests satisfied by the cache. Caching makes use of the locality in web request patterns (i.e., a recently requested data object is likely to be requested again in the near future). By moving content to the edge of the network, caches help reduce latency for the end user and bandwidth costs for the IAP. A recent survey of caching from a management science perspective is provided by Datta et al. (2002a).

Caching can be implemented at various locations in the network. For example, browser caches permit re-use on the client desktop alone. Proxy caches permit caching of data objects in gateways of large organizations or IAPs. Because proxy caches are located at points of aggregation, they are shared by a large number of users and hence demonstrate higher hit rates.

Thus, they are very effective in reducing latency and traffic. Vendors such as Inktomi and Cisco develop solutions that IAPs may use to deploy proxy caches on their servers.

In addition, CDNs provide solutions that allow content publishers to store data objects at geographically dispersed servers and direct a client to an appropriate server in order to reduce latency and bandwidth costs and evenly distribute server load. These servers are located at the edge of the Internet (typically collocated with partnering IAPs around the world) and thus provide similar benefits as proxy caches at IAP locations. While many of the arguments presented in this paper apply to CDNs as well, the focus of this paper is on caching at IAP locations. A brief discussion on the extension to the CDN setting is provided in the concluding section.

We can quantify, in dollar terms, the magnitude of bandwidth savings that can be realized by an IAP from caching. AOL serves 13.5 billion URLs to its users every day (AOL 2002). Vendors specify proxy caches to be capable of delivering hit rates between 35-75% (Web-Caching.com 2002). Assuming a hit rate of 40% and a conservative size of 100 KB for the response to an average web request, this amounts to an average of 16.2 petabytes of data every month that AOL serves from its cache and hence need not fetch over the network. However this demand for content would not be spread evenly in a 24-hour period. Assuming that the bulk of the data is demanded in a 6-hour time frame, i.e., a conservative peak-to-average ratio of 4:1, AOL saves 64.8 petabytes per month or a little over 200 Gbps of bandwidth. This can be serviced by eighty OC 48 connections (a single OC 48 connection can handle up to 2.488 Gbps). Even with today's reduced bandwidth costs, this amounts to savings of about $430 million per annum. These savings are AOL savings alone and do not account for consumer surplus from faster downloads due to caching or content publishers' bandwidth and infrastructure savings from not having to serve the content from their servers.

Proxy caches at IAP locations help the network operators (IAPs) realize bandwidth savings for themselves and latency reduction for the end users. They are therefore beneficiaries of these services. However, content publishers also derive significant benefits. Caches are crucial to content publishers in:

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 web sites unreachable. Caches help in

alleviating 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 web site due to poor performance. By improving response time, caches can thus increase customer retention rates for web sites.
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 to content publishers, they often choose to mark their content as non-cacheable (Chuang et al. 2002). 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 why publishers cache-bust is that caching results in a loss of business intelligence regarding visitors to websites. Accuracy of access reports and click-stream data is crucial to various firms for marketing and internal auditing. Caching can result in a loss of accuracy of access statistics because the origin server may not be informed whenever cached content is served. Furthermore, a variety of e-marketing techniques rely on personalization through cookies. Jupiter Research reported that 40% 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. Support for dynamic content caching in IAP caches has been minimal. Furthermore, 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. In addition, 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 need to devise ways to reap the benefits of caching without incurring the costs of today's "Best Effort" caching that precipitates cache-busting. These can be addressed by provisioning premium services that ensure comprehensive reporting, object consistency, security,

etc. In addition, technologies such as differential caching, push caching, dynamic content caching, etc. further enable publishers to derive tangible benefits from caching services. However, provisioning these features is in general expensive for an IAP (Maggs 2002; Stargate 2002). These services would therefore need to be priced to provide the IAP with the right incentives to provision them. Much of the work done to date on Quality of Service (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. The paper is organized as follows – we state the research question and summarize key results in section 1.1. We review the related literature and present our integrated QoS framework in section 2. Section 3 presents the results of our empirical analysis of trace data. In section 4, we develop an analytical model for pricing and resource allocation. In section 5, we relax a number of assumptions and test the robustness of our results. We conclude the paper in section 6 with a summary of our findings and directions for future research.

1.1 Research Question and Key Results

As noted above, adoption rates for caching has been declining because content publishers have generated new requirements over the last few years. These include support for personalization of content, business intelligence, etc. The paper analyzes *whether these new developments might result in market failure*. We then proceed to analyze *whether QoS-based services can play a role in increasing profits of cache operators and overall market welfare*. Through analytical models, we hope to inform caching service design and provide insights into optimal pricing and capacity allocation policies.

The primary contribution of the paper is to provide prescriptions to cache operators for design of incentive compatible caching services and help prevent the cache-busting that is rampant today. Our research suggests that the design of such incentive compatible QoS-based services can significantly improve market welfare. However, only large IAPs would be able to provide such services due to the role of transaction costs and economies of scale and scope. Smaller IAPs will need to partner with an intermediary to provide such services. CDNs can play an important role in facilitating such markets, particularly for small IAPs. The models also provide valuable insights on the future of the caching and content delivery market. For example,

we find that resources will be increasingly directed towards premium services in the future. Further, declining bandwidth costs will have a significant negative impact on profits from caching services.

2. QoS Framework and Literature Review

Publisher-centric caching has generated significant interest recently (Myers et al. 2001; Kelly et al. 1999; Chuang and Sirbu 2000). This permits introduction of verifiable QoS and focuses on providing value-added services to content publishers. While caching policies were previously traffic driven, an IAP may adopt caching schemes based on publishers' willingness to pay. Value added services such as differential caching, push caching, log reporting, maintaining object consistency, etc. enable preferential treatment to premium objects in a cache. The cache operator can thus price discriminate based on the desired level of QoS. Hereafter, we refer to caching with multiple levels of QoS as "QoS caching".

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 pre-fetching allow a publisher's objects to be moved into the cache even before a request is made for it. Thus, a publisher can enter into agreements with a cache operator that enables her objects to be moved into the cache in anticipation of future requests.

Since caches have a finite size, an object may have to be purged from the cache when a new object is moved in. The decision of which data object to replace is governed by the replacement policy. Replacement policies such as Least Recently Used (LRU) or Least Frequently Used (LFU) are commonly used to replace objects. The primary goal of replacement policies is to maximize the hit rate of the cache. LRU replaces objects that were least recently requested (assuming that they are also least likely to be requested again) and LFU replaces objects that were accessed 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. Alternatively, priority caches may be employed to increase the lifetime of the premium objects. Priority caches allow the assignment of priorities to different classes of content and provide high priority data objects with higher hit rates. Thus, the manner in which the IAP

chooses to allocate the available cache space between the different service classes determines the QoS experienced, with regard to 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). The cache operator can additionally 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) allows applications to specify the caching policies for the dynamic content generated by them. Datta et al. (2002b) also propose an approach to cache dynamic content at a proxy.

Kelly et al. (1999), Lu et al. (2001) and Feldman et al. (2002) discuss different hit rates for different service classes through biased placement or replacement policies. Zhu and Yang (2001) allow different invalidation schemes to be applied to different classes of dynamic objects. Barnes and Pandey (1999) provide language support for content publishers to specify the cache management policy for their content. Chuang and Sirbu (2000) and Pierre et al. (2001) discuss QoS for object replication. Myers et al. (2001) and Cao et al. (1998) address the security concerns that arise when caches are involved in content generation. While the technical solutions proposed by computer scientists address individual problems, an integration of these solutions provides us with a framework to provision QoS. Figure 2 summarizes the dimensions along which QoS may be varied in order to offer a wide array of vertically differentiated services.

	QoS Dimension	Best Effort caching	QoS Caching
1	Object Placement	Pull (traffic-Driven)	Push, Pre-fetching
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	No	Yes

Figure 2 Best Effort Caching Vs. QoS Caching

Provisioning QoS based services and pricing them is critical in aligning the incentives of the IAP and publishers. Despite the fact that publishers clearly receive benefits from caching today, an appropriate payment scheme does not currently exist. This has partly been due to the costs imposed by 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 developed a mechanism to correct both of these deficiencies. Provisioning of premium services ensures that publishers who value security, business intelligence, etc will have the incentives to cache. Pricing these services ensures that IAPs have incentives to provide premium 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; Rao and Peterson 1998), the domain of web caching poses unique challenges. Firstly, 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 service provider – the IAP – also derives a positive utility from the caching service (bandwidth cost reduction). That provides the IAP with incentives to provide discounts. Thus, there is a strong interaction between the service provider's surplus and the subscriber's (content publisher) surplus. Furthermore, pricing and resource allocation are strongly coupled in the domain of web caching. The price that the IAP can charge depends partially on how high a hit rate it can provision for the service classes, 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.

The analytical model in this paper is related to the models in (Mussa and Rosen 1978), (Bhargava et al 2000) and (Maskin and Riley 1984). 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. Maskin and Riley (1984) study optimal quantity discounts in a monopolized market with asymmetric information. Sundararajan (2002) 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). In the transmission domain, as in caching, pricing and resource allocation are strongly coupled. However, the resource allocation problems are quite different. In transmission, the router's queue management and scheduling operations provide the performance differentiation for data packets arriving in real time. Queue management controls the length of the packet queues and hence determines which data packets may be dropped when buffer overflow occurs. The scheduling policy determines which packet to send next and hence controls bandwidth allocation. Hence, a constrained buffer and bandwidth are allocated 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 in real time. 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 soon. Hence more elaborate QoS mechanisms, such as those specified in figure 2, can be justified. Additionally, this also allows for object level pricing. In contrast, even per-flow QoS (intserv) is deemed non-scalable 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 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 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. Pricing provides a means to align the incentives of IAPs and publishers and thus achieve the QoS goals. Finally, in contrast to network

QoS pricing, the IAP would choose prices that maximize its profit rather than social welfare. All these aspects make cache QoS pricing and resource allocation a unique and challenging problem.

3. Trace Analysis

Before we proceed to the analytical model for pricing and capacity allocation, we start with an empirical analysis of web traces in order to better understand the problem and calibrate some of the parameters for the model. In order to price the service, the IAP should be aware of the distribution of requests for content. For example, can we assume that requests are uniformly distributed or normally distributed across data objects? This is important because a publisher's valuation of caching an object depends on the number of requests for that object. If the object is requested 10,000 times, it may be valuable to cache this object as opposed to one that is requested 3 times. The distribution of requests for the population of data objects thus determines the price that the IAP can charge. In addition, we also need to determine whether and how differential allocation of cache space alters the benefit from caching. That is, how does the hit rate (fraction of requests served from the cache) increase with cache size? For example, is it acceptable to assume that hit rate increases linearly with cache size? Furthermore, is the increase in hit rate different for popular content versus relatively unpopular data objects? The answers to these questions will play an important role in determining the IAP's pricing strategy.

To answer these questions, we study two publicly available web proxy traces from the World Wide Web Consortium's web characterization repository – Boeing and DEC (Web Characterization Repository 2002). Trace data typically records requests for web pages made by end users and hence reflects the demand for content. Typical information available in a trace includes URL requested, time of request, object type, etc. Summary information about the two traces is provided below.

Trace	Date, Duration	Total No. of Requests	Number of Unique Objects
Boeing	Mar 99, 1 day	4,292,154	1,668,434
DEC	Sep 96, 7 days	7,866,111	2,047,000

Table 1. Summary Information of Traces

3.1 Pdf of Requests, *R*

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*. Although *R* is a discrete variable, we shall use a continuous approximation in the derivation below and the rest of the paper. The use of integrals

can be replaced by summations to derive the discrete version of the equations. Previous studies - for example, Breslau, et al. (1999) - have shown that requests-rank distribution follows a zipf-like distribution. That is, number of requests for the i^{th} most popular data object is $R = \Omega / i^d$, where Ω and d are constants. However, the distribution of requests has not been studied. Based on the zipf relationship, we can however analytically derive an approximate functional form for *f(R)*.

If *f(R)* denotes the pdf of *R*, then the rank of an object that is requested *R* times can be approximated as $Rank(R) = N \int_{R}^{R_{UB}} f(x)dx$, where *N* is the total number of objects and R_{UB} is the upper bound of the number of requests for an object (may be infinity). The expression assumes that all objects have unique number of requests and is hence an approximation. We also know from the zipf-like relationship that $Rank(R) = (\Omega / R)^{1/d}$. Thus, $N \int_{R}^{R_{UB}} f(x)dx = (\Omega / R)^{1/d}$. Taking the derivative with respect to *R*, we get $f(R) = \frac{bc^b}{R^{1+b}}$, where $b = 1/d$ and $c = \frac{\Omega}{N^d}$. This is the density function of a pareto distribution.

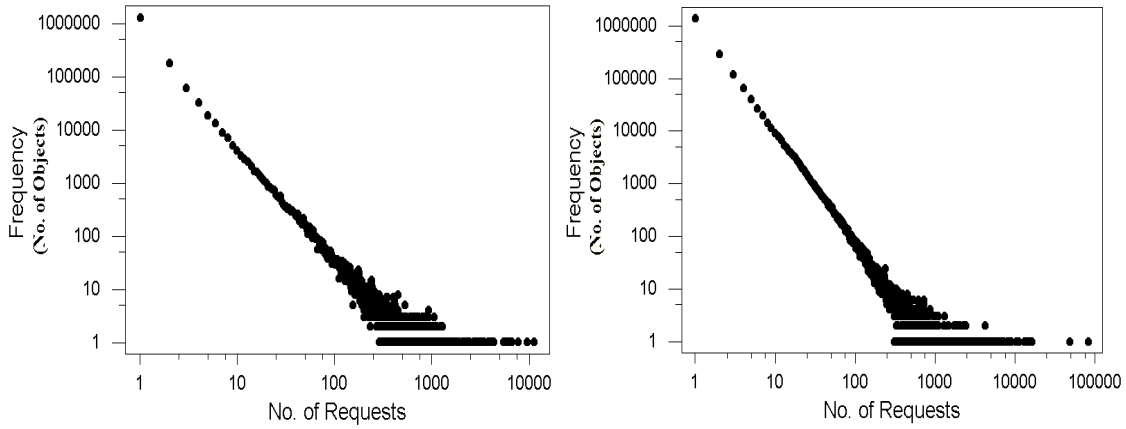

Figure 3. Histogram of R on a log-log scale for Boeing and DEC respectively

Figure 3 plots the histogram of *R* on a log-log scale for the two traces. *f(R)* may intuitively be considered as a measure of the number of objects with *R* requests. The distribution is modeled as $f(R) = \frac{bc^b}{R^{1+b}}$ where *R* lies in [c_1, c_2]. The parametric estimates for this distribution obtained from Maximum Likelihood Estimation (MLE) are in Table 2. The fit for both the traces is good.

For the purposes of the analytical model to follow, we need only note that requests follow a power law distribution. We shall later examine the implications of this distribution on our results.

Trace	c	b	c_1	c_2
Boeing	1.00	1.37	1	11479
DEC	1.00	1.29	1	84286

Table 2. Estimates of parameters for the distribution of *R*

3.2 Cache Hit rate

Hit Rate denotes the fraction of requests answered by the cache. For instance, if the cache satisfies three out of every six requests, it is said to have a hit rate of 50%. Clearly, a larger cache has a higher hit rate because it can store a larger number of data objects. Therefore H(*S*), the hit rate for a cache of size *S*, increases monotonically with *S*. In order to estimate the impact of cache size, we simulated a cache using the WisWeb cache simulator (Cao and Irani, 1997). The simulator simulates requests using a trace as an input and provides information on the hit rates for different cache sizes. We simulate an LRU caching policy, the most popularly used cache replacement policy. We consider cache sizes of 50% (a cache size big enough to store all objects in the sample is 100%), 20%, 10%, 5% and 0.5%. We test the following model for the two traces: $H(S) = k_s.\ln(S)$. We find that the logarithmic specification fits well for both traces. The results are also consistent with earlier literature such as Breslau et al. (1999).

Trace	Parameter Estimate (k_s)	Standard Error	t Value	Pr > |t|	R-Square
Boeing	0.03644	0.00065	56.03	<.0001	0.9984
DEC	0.04784	0.00111	43.07	<.0001	0.9973

Table 3. Estimates from regression of *H(S)* on *ln(S)*

3.3. 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*. This denotes the fraction of requests for that object that were satisfied by the cache. The previous sub-section estimated how the overall hit rate may vary with cache allocation. While the IAP cares about the overall hit rate, the content publisher is only concerned about the object specific hit rates for her objects. Clearly, *H(S,R)* increases with cache size *S* but also varies with the object popularity *R*. The more popular an object, greater the likelihood of it being requested before it is replaced from the cache. Therefore, it enjoys a higher hit rate too. Thus

H(S,R) also increases monotonically with *R*. We model the object specific hit rate as follows: $H(S,R) = k \cdot \ln(S) \cdot \ln(R)$. In addition to noting the hit rates for the cache as a whole, we also recorded the hit rates for each data object in the trace-driven cache simulations. The parameter estimates are presented in table 4. Other specifications such as $H(S) = k_s.S^b$ and $H(S,R) = k.S^{b_1}.R^{b_2}$ also perform well. However, we use the chosen specification because of the common use of the logarithmic specification in modeling hit rates in the caching literature (Breslau et al. 1999). We conclude our empirical analysis with a summary of the findings. We found that distribution of requests follows a power law, specifically a pareto distribution. We found that increasing the cache size for any service class increases the hit rate in a concave manner, with a logarithmic relationship. The hit rate experienced by different objects varies across objects, with popular objects enjoying a higher hit rate. We use these observations in the analytical model that follows.

Trace	Parameter Estimate (k)	Standard Error	t Value	Pr > |t|	R-Square
Boeing	0.02296	0.00002463	932.09	<.0001	0.8281
DEC	0.02271	0.00002672	850.17	<.0001	0.7395

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

We summarize here the key results from the empirical analysis:

1. Distribution of Requests for content: The pdf of the number of requests for data objects is given by $f(R) = \frac{bc^b}{R^{1+b}}$ (a power law relationship)

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

3. Object Specific Hit rate: The hit rate experienced by an individual data object varies as the logarithm of the cache size and as the logarithm of the number of requests for the object. Thus, popular objects are intrinsically more likely to be served from the cache (for an LRU policy).

4 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 the real world situation very well. Publishers' decisions regarding the content they mark as

cacheable vary from one data object to another. Further, the publisher may care about security for a data object that contains confidential data but may not place a similar weight on quality (security, consistency, etc) on another object. Thus, it makes sense to model caching decisions by object, as valuations for different data objects may be different, even for the same publisher.

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 assume that a cache is already in place and the cache size, *S*, is therefore a fixed exogenous parameter. That is, we do not consider determination of the cache size in our model. There are two reasons for this. Firstly, the issue of optimal cache sizing has been considered in detail by Kelly and Reeves (2000). Secondly, 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.

Symbol	Explanation
q	Quality level of value-added features such as reporting, consistency.
q	The weight a publisher places on the quality for an object (type parameter)
S	Total size of cache
a	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 $S = k.\ln(S).\ln(R)$
Thc	Total Hit count or total number of times an object with R requests is served from cache = $R.H(S,R)$
B	Average Bandwidth cost for processing one object request = Bandwidth

	cost per byte*Average size (in bytes) of object
B_{IAP}	IAP's average bandwidth cost for processing one object request
h	Publisher's benefit from faster delivery of an object to an end user. This may come in the form of increased sales, 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.
p	IAP's expected profit

Table 5 Glossary of terms

We begin this section with an analysis of IAP caching services as they are provisioned today. As explained in section 1, IAPs maintain a best effort cache without any support for security, business intelligence and other value-added features. We explore why IAPs do not price the service, why cache-busting occurs, and study the potential loss in social welfare. In section 4.2, we analyze the equilibrium when a premium service is introduced by the IAP.

4.1 Single Best-effort Service

The content publisher has content (data objects) that is requested by an end user and the IAP is the 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 (the IAP derives some benefits too, namely bandwidth savings). We separate the value derived by the publisher from the caching service into two components – benefits derived from caching and costs incurred due to caching. The former represents benefits directly associated with caching (such as bandwidth savings and benefit from faster delivery of content). These benefits are derived every time an object is served from the cache (called a "hit"). The latter represents costs incurred due to the compromise of value added features such as personalization of content, security, reporting, etc.

The cost to the publisher of compromising on security and other value-added features is denoted by $q(-q_L)$. q is a type parameter that denotes the weight that the publisher attaches to value-added features for the specific object. This weight would clearly vary 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 q whereas another not requiring security, reporting, etc. may correspond to an extremely low q. We assume that q is

uniformly distributed in [0,1] across objects. We shall soon discuss the implications of this assumption. ($-q_L$) denotes the level of quality of the value-added features, with the negative sign indicating the "lack of quality" or the fact that costs are being incurred due to loss of security, personalization, etc.

The publisher derives a benefit from being able to deliver his content faster to the end users (consumers of content). This benefit captures increased customer retention rates from faster delivery of content. For every object delivered from the cache, h represents the aforementioned benefit to the publisher. Another important component of the benefit from caching is the bandwidth savings realized by the publisher. B denotes the average bandwidth cost of processing one object request. A piecewise separable benefit function is used, $U = q \cdot (-q_L) + (Thc) \cdot (h + B)$, where Thc or *"Total hit count"* is the count of the number of times the object is served from the cache. If R denotes the number of requests for an object and $H(S,R)$ denotes the object specific hit rate, then $Thc = RH(S,R)$. The benefit function captures the tradeoff faced by the publisher – caching provides certain benefits related to bandwidth savings and faster content delivery but imposes costs too.

We assume that the IAP charges a price P for every object served from this best-effort cache. Thus, the net surplus to the publisher derived from caching an object of type q and demand R is $U_L = q \cdot (-q_L) + R \cdot H(S,R) \cdot (h + B) - P \cdot R \cdot H(S,R)$. The publisher's decision problem is to choose whether to cache or opt out (cache bust). 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 q_i who is indifferent between caching and not caching (gets zero surplus from the caching service). By setting $U_L = 0$, we get $q_i = \frac{R \cdot \ln R \cdot k \ln S(h + B - P_L)}{q_L}$. Note that q_i varies with R as shown in figure 4. Any publisher with an object with $q > q_i(R)$ is more sensitive to value-added features and hence will not cache. Those with lower qs will cache because they incur lower costs from caching but derive the same benefits as the indifferent publisher. This is also illustrated in the figure.

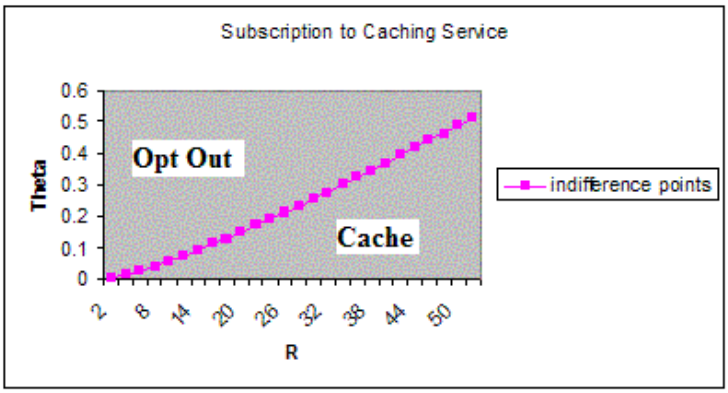

Figure 4. Sample Indifference points (for assumed price and quality levels)

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 $q \le q_i$ (see figure 4). 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:

$$p = R_L \cdot H(S) \cdot (P + B_{IAP} - T)$$

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). We sum up all requests for objects with $q \le q_i$ and plug it into the profit function to get:

$$p = \left[\int_{c_1}^{c_2} \int_{0}^{q_i(R)} f(q) dq f(R) R dR \right] k_s \ln(S) \cdot (P + B_{IAP} - T) = \left[\frac{k \ln S (h + B - P) k_1}{q_L} \right] \cdot k_s \ln(S) \cdot (P + B_{IAP} - T)$$

where $k_1 = \int_{c_1}^{c_2} R \ln R f(R) dR$ is a constant of integration. The IAP's decision problem is $\max_P \{p(P)\}$. Based on the first order condition, the optimal price that the IAP should charge is $P^* = \frac{h + B + T - B_{IAP}}{2}$. The global concavity of the profit function can be easily verified. Note that the IAP passes a part of its transaction cost of billing and metering to the publisher and thus the price charged increases with the transaction cost, T. The indifference point associated with

the optimal price is $q_i(P^*) = \frac{R \cdot \ln R \cdot k \ln S(h + B + B_{IAP} - T)}{2q_L}$. Thus, the number of subscribers to the caching service decreases as the transaction cost, *T*, increases. This is illustrated below.

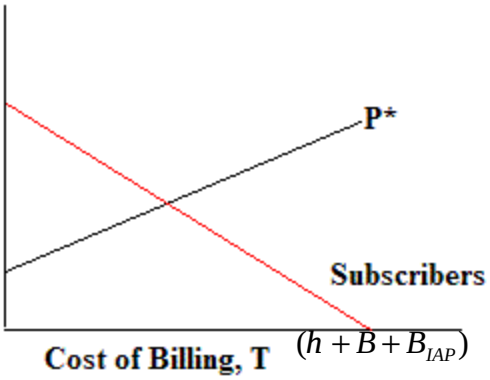

Figure 5. Impact of transaction costs on prices and subscriptions

When $T = (h + B + B_{IAP})$, 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 (transaction cost of billing, T would also be 0) and maximizing its bandwidth savings instead. The IAP's profit when $P^*=0$ is $p(P=0) = \frac{kk_1 \ln S(h+B)}{q_L} k_s \ln S \cdot B_{IAP}$. It would be optimal for the IAP to set the price to zero if $p(P=0) \geq p(P^*)$. Simplifying, if $T \geq (h + B + B_{IAP}) - 2\sqrt{(h+B)B_{IAP}}$, the optimal price for the IAP is P^*=0. One possible explanation for the fact that today's best effort service is free is that T is close to the above threshold. That is, the transaction cost of metering and billing may be high compared to the benefits that each player derives. Furthermore, the price is low if the IAP bandwidth costs are high.

To consider the impact of distributional assumptions, we solved the same model but introduced a skew in the distribution of object types: $F(q) = q^2$; $f(q) = 2q$. Relative to a flat distribution, this distribution assumes that there are a relatively higher number of publishers who care about value-added features than those who do not (see figure 6). The new solution is $P^* = \frac{h + B + 2T - 2B_{IAP}}{3}$. If $T \geq h + B + B_{IAP} - \sqrt[3]{\frac{27}{4}(h+B)^2 B_{IAP}}$, 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[2]. We also find that the price decreases with increasing IAP bandwidth costs. Higher the negative skew, greater is this decrease (the negative weight on B_{IAP} is higher with the skewed distribution than the uniform distribution). If $B_{IAP} = \left(\frac{h+B}{2}\right) + T$, then $P^* = 0$. Hence, another possible explanation for the zero prices observed in reality is that IAP bandwidth costs are reasonably high and publisher preferences are negatively skewed with a large number of publishers being sensitive to value-added features. Even when prices are zero, several publishers will cache-bust because the cost of caching ($-qq_L$) dominates the benefits for these publishers.

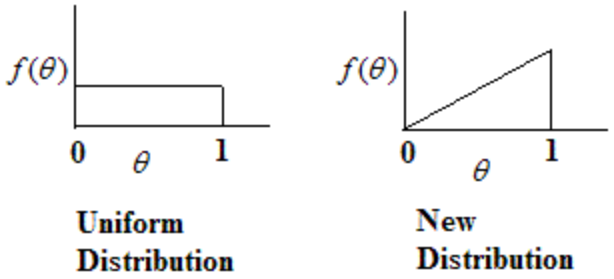

Figure 6. Negative skew in distribution of preferences

The model presented here highlights a number of reasons why prices may be zero today (high transaction costs, high bandwidth costs, skews in publisher preferences or a combination of these factors). The primary 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). Thus, market welfare decreases in general. Recent trends suggest that publishers are become increasingly sensitive to business intelligence and personalization as online business models have started to evolve. Thus, adoption

[2] If $B_{IAP} > \frac{729}{1024}(h+B)$, there is a small range of values with $T \in [h + B + B_{IAP} - 2\sqrt{(h+B)B_{IAP}}$, $h + B + B_{IAP} - \sqrt[3]{\frac{27}{4}(h+B)^2 B_{IAP}}]$ where the price in the case of the uniform distribution is zero but non zero 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.

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.

4.2 Provisioning 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 to the publishers. This higher quality is achieved 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 due to 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 et al. (2002) propose a multi-level replacement policy based on a number of interconnected LRU-based queues. All these authors use different techniques to achieve the same goal – providing different *effective* cache sizes to different content classes. Feldman et al. (2002) indicate how the effective cache sizes are related to the sizes of each queue in their two-level LRU cache. 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.

4.2.1 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[3]. The IAP offers a service at a higher quality level and charges a per-object price P_H. 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

[3] 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.

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:

$$U_L = q \cdot (-q_L) + R \cdot H(S_L, R) \cdot (h + B) - P_L \cdot R \cdot H(S_L, R) \quad (1)$$

$$U_H = q \cdot q_H + R \cdot H(S_H, R) \cdot (h + B) - P_H \cdot R \cdot H(S_H, R)$$

where $S_L = aS$ and $S_H = (1-a)S$.

As discussed previously, any priority scheme for differential caching can be expressed using different *effective* cache sizes for the two services. The cache space, *S*, is divided into 2 levels, with a denoting the fraction of cache space allocated to the best effort service. That is, aS is the size of the cache for best effort subscribers and the remainder of size (1-a)*S* is for the premium subscribers. To determine the number of subscribers to the service, we consider a publisher with object of type q_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 $q > q_L$ is more sensitive to value-added features and hence will not choose the best effort service to cache the object. Those objects associated with lower q will cache in the best effort service. Similarly, we consider a publisher with an object of type q_H who is indifferent between the premium service and not caching (gets zero surplus from the premium services). Objects associated with $q > q_H$ gain more benefits from the premium service and will be cached. Publishers with objects of type $q < q_H$ do not weigh the value-added features enough to be willing to pay the price P_H. By setting $U_L = 0$, we get q_L and by setting $U_H = 0$, we get q_H.

$$q_L(R) = \frac{R \ln R \cdot k \ln(S_L)(h + B - P_L)}{q_L} \text{ and } q_H(R) = \frac{R \ln R \cdot k \ln(S_H)(P_H - h - B)}{q_H} \quad (2)$$

Both q_L and q_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 7. The interesting

[4] "Quality 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. Since these points vary with *R*, we refer to them as QICs.

region is the intermediate one where $q \in [q_L, q_H]$. Publishers in this region care enough about security, business intelligence, etc. that they will not choose the best-effort service but not so much that they are willing to pay the premium price. Note also that in figure 7, publishers subscribe to the service which provides them with the maximum surplus, an Incentive Compatibility (IC) constraint. Additionally, the publishers will cache a data object only if the surplus from doing so is positive, which is an Individual Rationality (IR) constraint.

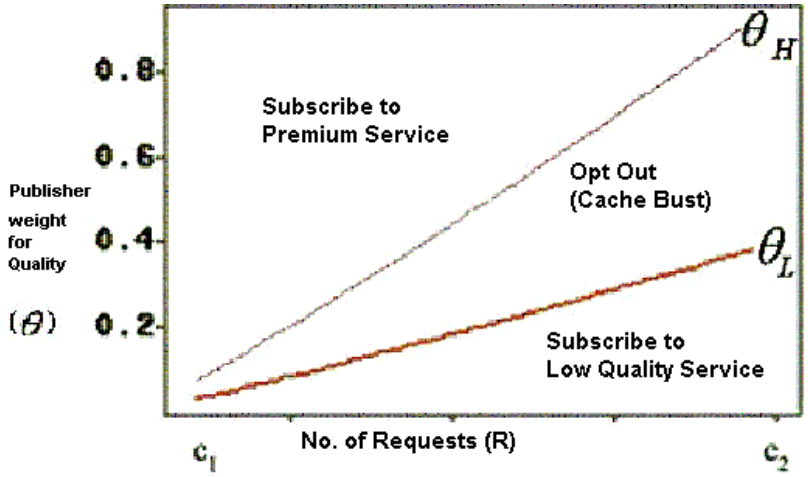

Figure 7. Sample Indifference points (for assumed price and quality levels)

4.2.2 Properties of the Quality Indifference Curves (QICs)

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 Quality Indifference Curves never cross each other.

Lemma 2: If the IAP prices optimally, then $q_H(R) \geq q_L(R)$.

The proofs for both lemmas are provided in the appendix. We will use these lemmas in lemma 3 and in computing IAP profit functions in the next section.

Segmentation Conditions

The market is segmented if there are customers for both the services. There exist subscribers to the best effort service only if $q_L > 0$ for some R. Similarly, there exist subscribers to the premium service only if $q_H < 1$ for some R (see figure 7). These two conditions are thus necessary conditions for segmentation.

i. $q_H < 1$, for some $R \in (c_1, c_2)$: If $\{q_H > 1, \forall R\}$, then there is no "feasible" subscriber[5] who prefers the premium service. Thus, this condition is necessary if the IAP has customers for the premium service. $q_H < 1$ results in the following inequality:

$$P_H < h + B + \frac{q_H}{k \cdot \ln S_H c_1 \ln c_1} \qquad (4)$$

Equation (4) indicates that if the price is above the specified threshold, no publisher will choose the premium service.

ii. $q_L > 0$, for some $R \in (c_1, c_2)$: If $\{q_L < 0, \forall R\}$, then there will be no subscriber for the best effort service. This condition can be rewritten as $P_L < h + B$. 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 $P_H < h + B + \frac{q_H}{k \cdot \ln S_H c_1 \ln c_1}$ and $P_L < h + B$, then the market will be segmented.

The proof is provided in the appendix. 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 optimal to segment the market depends on whether the optimal prices satisfy the conditions of lemma 3.

4.2.3 IAP's Decision Problem

Illustrative Example: Consider a discrete example where the population consists of 4 objects $\{o_1, o_2, o_3, o_4\}$. The IAP receives 10 requests per unit time and this consists of 5 requests for o_1, 3 for o_2, 1 for o_3 and 1 for o_4. Thus, the demand vector is given by $< R >= \{5,3,1,1\}$. Let us assume that, based on the IAP's pricing decision, o_1 subscribes to the premium service, o_2 and o_3 subscribe to the best-effort service and o_4 opts out. Thus, the subscription to the services is given by the 2-tuple (1,2) indicating that 1 object subscribed to the premium service and 2 objects to the best-effort service. However, the IAP's profit depends not on the subscription per se, but on how these objects contribute to the incoming request. The incoming request vector is given by the 2-tuple (5,4), indicating that 5 requests per unit time are for objects subscribed to the premium cache and 4 for objects subscribed to the best-effort cache. The IAPs profit is given

[5]By "feasible" subscriber, we mean a subscriber in the feasible region with $q \in [0,1]$. By definition, there exists no subscriber outside the feasible region.

by $5H(S_H)[P_H + B_{IAP} - T] + 4H(S_L)[P_L + B_{IAP} - T]$. The IAP's decision problem is thus to choose the prices P_H, P_L and the allocation S_H, S_L so as to maximize its profit. Note that the request vector, i.e. the 2-tuple (5,4), is itself determined by these decisions since the price charged influences the subscription to the two services. ◁

The IAP has a cache of size *S*. It allocates the space to the two services through its choice of a (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 find it desirable to choose prices that sets the q_H QIC below the q_L QIC. Hence, we only discuss the case $q_H(R) \geq q_L(R)$. 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 $q > q_H$ (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 $q \leq q_L$. The IAP's expected profit is given by:

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

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 incurs a marginal cost T of metering and billing. There is a fixed cost for the IAP to provision the value added services. This is the cost of the infrastructure the IAP needs in order to provide security, business intelligence, etc. The cost is $C(q_H)$, which is monotonically increasing in q_H.

As discussed previously, R is randomly distributed across objects and has a pdf given by the pareto distribution. The hit rate is given by $H(S) = k_s \ln(S)$. R_H is obtained by summing all the requests for objects in the topmost region of figure 7. R_L is obtained by summing up all requests in the lowermost region. That is, $R_L = N\int_{c1}^{c2} q_L(R) Rf(R) dR$ and

$R_H = N\int_{c1}^{c2} (1 - q_H(R)) Rf(R) dR$. This gives us:

$$p = N\left[E(R) - \frac{kk_1(P_H - h - B)\ell nS_H}{q_H}\right] \cdot k_s \ell nS_H \cdot [P_H + B_{IAP} - T]$$
$$+ \frac{Nkk_1(h + B - P_L)\ell nS_L}{q_L} k_s \ell nS_L \cdot [P_L + B_{IAP} - T] - C(q_H) \quad (5)$$

where k_1 is a constant of integration (see Appendix A2 for full derivation of expected profit). The IAP's decision problem is $\max_{P_H, P_L, a} p(P_H, P_L, a)$. Following the traditional approach (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 a 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 a). Note that this procedure maps well to reality where pricing follows production or service design.

4.2.4 Pricing Problem

As discussed above, we assume that a is an exogenous parameter and focus on the pricing problem in this section. The first order conditions associated with the problem are as follows:

1) $$\frac{\partial p}{\partial P_H} = N\left[E(R) - \frac{kk_1 \ln S_H (P_H - h - B)}{q_H}\right] k_s \ln S_H - \frac{Nk \cdot k_1 \ln S_H}{q_H} k_s \ln S_H [P_H + B_{IAP} - T] = 0$$

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.

2) $$\frac{\partial p}{\partial P_L} = \frac{Nkk_1 \ln S_L (h + B - P_L)}{q_L} k_s \ln S_L - \frac{Nkk_1 \ln S_L}{q_L} k_s \ell nS_L [P_L + B_{IAP} - T] = 0$$

The interpretation is similar. 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). The interpretation of the first-order conditions and trade-offs associated with pricing is summarized in Table 6.

	Impact on Premium Service		Impact on Best-Effort Service	
Marginal	Subscription	Revenue from	Subscription	Revenue from

Increase in:		Service		Service
P_H	Decrease	+/-	No Impact	No Impact
P_L	No Impact	No Impact	Decrease	+/-

Table 6. Impact of decision variables on profit

Solving the two first-order conditions gives us:

Lemma 4: When $T < (h + B + B_{IAP}) - 2\sqrt{(h + B)B_{IAP}}$, the optimal prices are:

$$P_L = \left(\frac{h + B + T - B_{IAP}}{2} \right)$$
$$P_H = \frac{E(R)q_H}{2kk_1 \ln S_H} + \left(\frac{h + B + T - B_{IAP}}{2} \right) \quad (6)$$

When $T \geq (h + B + B_{IAP}) - 2\sqrt{(h + B)B_{IAP}}$, 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 equation 6 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 $T > (h + B + B_{IAP}) + \frac{q_H}{k \ln S_H} \left[\frac{2}{R_{LB} \cdot \ln R_{LB}} - \frac{E(R)}{k_1} \right]$, then it is not optimal for the IAP to segment the market.

If transaction costs of metering and billing are exceedingly high, then a market for premium services may also fail. However, 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. Improvements in Information technology (IT) can help facilitate increase in q_H as has been witnessed in the last several years in caching technologies. In addition, IT can play an important role in reducing the transaction cost of metering and billing and help facilitate such markets for premium web caching services.

The Quality Indifference Curves associated with the optimal prices of equation (6) are:

$$q_L^* = R\ln(R) \frac{k \ln S_L}{q_L} \left(\frac{h + B + B_{IAP} - T}{2} \right)$$
$$q_H^* = R\ln(R) \left\{ \frac{E(R)}{2k_1} - \left(\frac{h + B + B_{IAP} - T}{2} \right) \frac{k \ln S_H}{q_H} \right\} \quad (7)$$

Special Case

Note that the first order conditions in section 4.2.4 display no cannibalization effects. Hence, the optimization of P_L and P_H is done independently. Thus, this process may artificially push the q_H QIC below the q_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 $q_H(R)=q_L(R)$. The corresponding solution is given by

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

$$P_H = h + B + \frac{E(R)q_H^2}{2k \cdot k_1(q_L + q_H)\ln S_H} - \frac{h + B + B_{IAP} - T}{2}\left(\frac{\ln S_H - \ln S_L}{\ln S_H}\right)\left(\frac{q_H}{q_L + q_H}\right)$$. Note that this special case represents the case where the market is completely captured (all publishers cache either in best effort or premium service). This special case is not very realistic and thus we focus only on the interior solution in the rest of this paper.

4.2.5 Comparative Statics

Based on the QICs in equation 7, we can conduct sensitivity analysis to determine how various parameters impact the subscriptions to the services. For example, how would subscriptions to a service change if technology facilitates improvements in the quality of value-added features. 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 (7), it follows that an decrease in q_L (note that an increase in quality corresponds to a decrease in q_L) causes q_L to increase and q_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 $T \geq (h + B + B_{IAP}) - 2\sqrt{(h + B)B_{IAP}}$, the best effort service is free ($P_L^*=0$). Substituting this into equation 2, we again find that q_L increases with quality. Thus, subscription to 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 is less than, equal to or greater than $h + B + B_{IAP}$.

Proof: From equation (7), it follows that an increase in q_H has no effect on q_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 exist 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 low quality service that results in the no-subscription region being sandwiched between the premium and best-effort subscribers (see figure 7). Thus, changes in parameters of any service affects that service but the other service is shielded from experiencing any impact. 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. $-q \cdot q_L < 0$). 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 q_H depends on the relative magnitude of transaction costs compared to the hit-based benefits. The IAP may lose subscribers but earn higher margins per subscriber with low transaction costs. Note that proposition 3 is consistent with proposition 1. If $T > (h + B + B_{IAP})$, the quality of the premium service will have to be considerably high for the service to exist .

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 (7). This causes q_H to increase and therefore subscription to the premium service drops. Simultaneously, q_L decreases implying that subscription to best effort service decreases.

Table 7 lays out the impact of changes in quality and bandwidth costs on the subscription to the two services.

Increase in	q_L	q_H	Subscription to Best Effort	Subscription to Premium Service
$(-q_L)$	Increases	-	Increases	-
q_H	-	Varies	-	Varies
B,B_{IAP}	Increases	Decreases	Increases	Increases

Table 7. Impact of quality and bandwidth on Subscription

Equation (6) provides valuable insight into the IAP's pricing decision. The IAP charges the content publisher a part of her surplus from bandwidth reduction and faster content delivery $(h + B)$. The IAP gives back to the publisher a part of its own surplus from bandwidth reduction (B_{IAP}) but also passes on a part of the transaction cost of billing. If the transaction costs or IAP bandwidth costs are high or there are a large number of publishers sensitive to value-added features (negative skew in q), 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 the best effort services to remain free even when a premium service is introduced. 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. Assuming a nonlinear surplus function results in non-linearity in pricing. For example, assuming the following surplus function: $U_H = q \cdot q_H - Cq_H^2 + Thc(S_H, R) \cdot (h + B) - P_H \cdot Thc(S_H, R)$, where C is a normalization constant, results in the following nonlinear optimal price: $P_H = \frac{E(R)q_H(1-Cq_H)}{2kk_1 \ln S_H} + \left(\frac{h + B + T - B_{IAP}}{2} \right)$. Space constraints prevent us from providing a detailed derivation. Interested readers may contact the authors. There also exists some non-linearity with regard to impact of cache sizes. Equation (6) 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.

4.3 Space Allocation

In the previous section, we looked at the pricing problem assuming that the cache sizes $S_L = aS$ and $S_H = (1-a)S$ 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 are never over-provisioned. This is due to diminishing returns from increasing 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. As stated earlier, our model prescribes the optimal *effective* cache sizes for the two services. An appropriate priority scheme may be used to implement the same.

We obtain the first order condition of the expected profit with respect to a, conditional on the prices. The equation is of the form $a = g(a)$, where g() is continuous over a in [e,1-e] (see Appendix A3). The equation does not yield a closed form solution, although the existence can be guaranteed by Brouwer's Fixed Point Theorem (Brouwer 1910). The proof is based on the observation that $g(e) \geq e$ and $g(1-e) \leq (1-e)$. Therefore, $g(e) - e \geq 0$ and $g(1-e) - (1-e) \leq 0$. Since g() is continuous, there exists an a^* in [e,1-e] such that $g(a^*) - a^* = 0$. Thus, the optimal allocation of cache space is given by the solution to $a^* = g(a^*)$. The detailed proof is presented in the appendix. To illustrate how the IAP may solve the allocation problem, we consider a numerical example below.

4.3.1 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 publisher with 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. Thus, the IAP bandwidth cost is approximately 36% that of the typical publisher. We assume that the publisher's benefit from faster delivery of content is of the same order of magnitude as bandwidth savings, i.e., $h = 0.03$c. While appropriate values for q_L and q_H can be best calibrated from surveys of content publishers, one can calibrate these parameters by intuitively considering $q \cdot q$ 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 non-hit rate attributes such as security, consistency, reporting, etc ($q = 1$) 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 of compromising on security, business intelligence, etc is 0.4 c per object. This choice of q_L sets the cost of compromising on value-added features high enough that it dominates benefits from bandwidth savings and faster content delivery for a large number of data objects. Finally, the transaction cost of billing to the IAP is set as $T = h + B + B_{IAP} = 0.071c$ per object. This sets the cost high enough that the best effort service will be free (as observed in reality today). The cache size is assumed to be 6 GB. All the remaining parameters are empirically derived from the Boeing trace. A cache at an IAP location such as AOL would be of the order of a few Terabytes in size. However since we use the parameters derived from the Boeing trace, which has a lower number of requests as well as distinct objects, we use a proportionately smaller cache size. Figure 8 illustrates that the IAP would find it optimal to allocate 17.32% of the cache space (or 1.04 GB) to the lower level cache for the simulated setting (i.e., $a^* = 0.1732$).

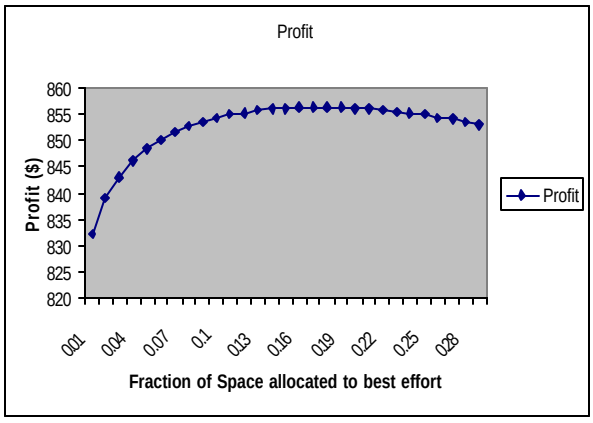

Figure 8. Optimal allocation for the IAP

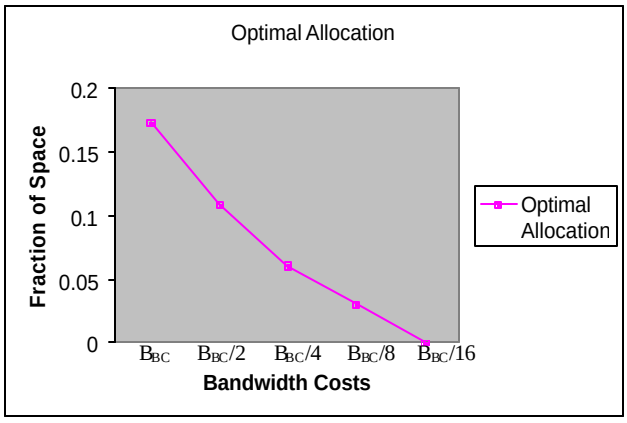

Figure 9. Impact of decreasing bandwidth costs on allocation

One interesting question that arises from recent trends is how do decreasing bandwidth costs impact the allocation decision and profit. We repeat the simulations but halve the bandwidth costs for both the IAP and the content publisher (B = 0.015c and B_{IAP} = 0.0055c). Under the new settings, it is optimal for the IAP to allocate 10.9% of the cache to the lower level. Figure 9 lays out the impact of lowering bandwidth costs on a^*. B_{BC} represents the base case of B = 0.03c and B_{IAP} = 0.011c. In each successive simulation, we halve the bandwidth costs from the previous simulation (both B and B_{IAP} are halved). 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.

We find that the IAP's profit also decreases when bandwidth costs fall[6]. In figure 10, the upper curve shows the change in IAP's profits (the data is calibrated to Boeing trace) as bandwidth costs decrease. The lower curve plots the IAP's profits if it does not pursue any QoS caching policies. This is the current scenario wherein an IAP's surplus from caching consists only of its bandwidth savings. It must be noted that these simulations do not account for other possible trends that might accompany reduction in bandwidth costs. For example, the reduction in costs might reduce access charges for end users causing a greater demand for content (both the type of content requested and number of requests may change). Also, there may be accompanying changes in response times causing changes in h (publisher's benefit from faster delivery of content). It is difficult to ascertain the exact nature or magnitude of these changes. However, it is possible to determine the rate at which traffic will need to go up in order to maintain IAP profits at the same level. Figure 11 indicates how the total volume of traffic will need to rise for our trace (Boeing) in order for the IAP profits from QoS caching to remain constant despite declining bandwidth costs. The sub-linear relationship (note also that the x axis is on a log scale) indicates that such an increase may be feasible.

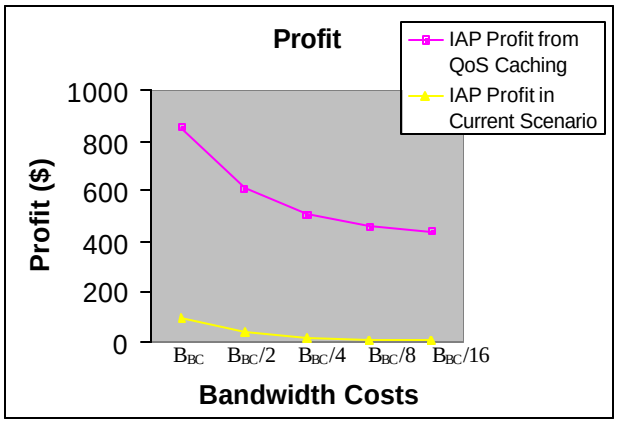

Figure 10. Impact of decreasing bandwidth costs on IAP Profit (in cents/day)

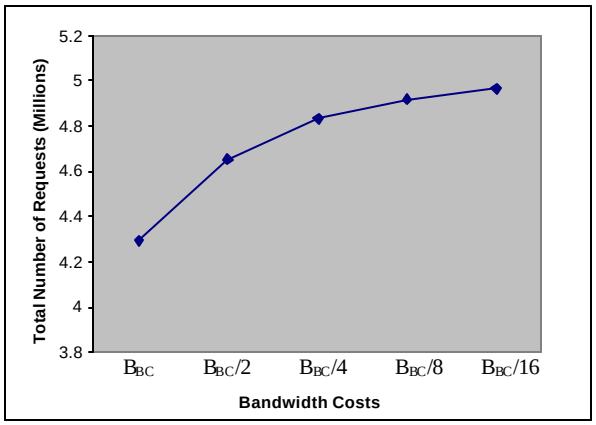

Figure 11. Increase in traffic needed to maintain IAP profits with declining bandwidth costs

[6] This effect can also be identified by applying the envelope theorem to equation 5.

5. Robustness of 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. That is, we assumed that all objects have the same size. However, accounting for size related heterogeneity adds an additional dimension to the QICs without providing greater insight. While heterogeneity in object sizes impact the efficiency of caching algorithms and policies, they did not seem to critically affect the pricing strategies (especially since the pricing is usage based). By ignoring size differences, the model thus ignores heterogeneity in bandwidth cost B and benefit from faster delivery of content h. 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 -- irrespective of whether the capacity is put to use or not. Finally, there are burstable T1, T3 and OC3 services which enable organizations to use up to the maximum capacity of the service (e.g., 1.5 Mbps in the case of T1) to handle peak loads but pay based on their usage pattern organized into tiers. It is clear that a single model cannot capture both usage-based pricing and pricing for capacity irrespective of usage. Hence, we consider average bandwidth cost per object served from the cache (or equivalently, bandwidth cost per byte). The average bandwidth cost is clearly non-zero and represents publisher's bandwidth considerations rather well. Furthermore, recent trends in bandwidth pricing (for example, burstable packages for leased lines) have focused on usage-based metering.

The model assumed that q is uniformly distributed across data objects. The broader insights of the model did not change for other well-known distributions. The net impact of a negative skew was to reduce the price that the IAP can charge. The model also assumed that the publisher surplus from value-added features, q, was linear in q. Introducing non-linearity in preferences also introduces non-linearity in pricing, as indicated in section 4.2. In section 5.1, we revisit our model and consider correlation between benefit from value-added features and number of requests for an object. In section 5.2, we discuss the sensitivity of the results to trace parameters and finally discuss the implications of publishers' contracting costs in 5.3.

5.1 Correlation between benefits from value-added features and demand, *R*

The publisher's benefit function in section 4 assumed that the benefit/cost from value-added features is independent of the number of requests for an object. The two may in fact be correlated. That is, the 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: $U = qRq' + R \cdot \ln R(h + B) - P \cdot R \cdot \ln R$. This surplus function is similar to the one in section 4. It assumes that each object has an intrinsic requirement (or lack thereof) for value-added features. This is denoted by q. For example, a publisher with an object with confidential data always values security more for this object than another without sensitive information (all else held constant). However, this function also assumes that the importance of value-added features increases with popularity of an object. Given two objects with the same q, the publisher values the premium service higher for the relatively popular object. The impact of the change is that the slopes of both the QICs in figure 5 decrease. The optimal prices are given by: $P_L = \left(\frac{h + B + T - B_{IAP}}{2} \right)$ and $P_H = \frac{E(R)q'_H}{2kk'_1 \ln S_H} + \left(\frac{h + B + T - B_{IAP}}{2} \right)$, where $k'_1 = \int_{c_1}^{c_2} \ln R \cdot f(R) \cdot dR < k_1$. The per-object price charged to the best effort service is not affected and that to the premium service changes (we will soon show that it 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:

$$U_L = q \cdot R^2 \cdot (-q''_L) + R \cdot k \ln S_L \cdot \ln R \cdot (h + B) - P_L \cdot R \cdot k \ln S_L \cdot \ln R \qquad (8)$$

$$U_H = q \cdot R^2 q''_H + R \cdot k \ln S_H \cdot \ln R \cdot (h + B) - P_H \cdot R \cdot k \ln S_H \cdot \ln R$$

Note that this model is also similar to the model in section 4 except that it assumes that the benefit/cost from value-added features increases as the square of *R*. We had previously defined q_H as the dollar benefit from the value-added features to the most sensitive publisher. From that definition, it follows that $q''_H = q_H / R_{UB}{}^2$. Similarly, it follows that $q''_L = q_L / R_{UB}{}^2$. These transformations ensure consistency across the two models so that we can continue to infer each term in the surplus function as a dollar cost or benefit. In this new specification, as in the preceding case, the slopes of the 2 QICs decrease. However, the impact is strong enough to cause

the QICs to slope downwards as illustrated in the figure below. Downward sloping QICs imply that more popular objects prefer the premium service.

$$q_L(R) = \frac{\ln R \cdot k \ln S_L(h + B - P_L)}{q_L'' R} \text{ and } q_H(R) = \frac{\ln R \cdot k \ln S_H(P_H - h - B)}{q_H'' R} \qquad (9)$$

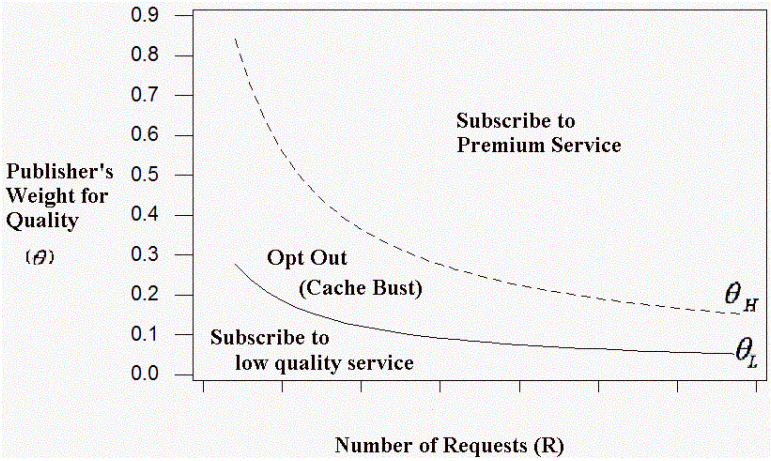

Figure 12. Sample QICs for new model (assumed price and quality levels)

The functional form of the optimal prices do not change, with $P_L = \left(\frac{h + B + T - B_{IAP}}{2}\right)$ and $P_H = \frac{E(R)q_H''}{2kk_1'' \ln S_H} + \left(\frac{h + B + T - B_{IAP}}{2}\right)$, where $k_1'' = \int_{R_{LB}}^{R_{UB}} (\ln R / R) \cdot f(R) \cdot dR$. Now, $\frac{q_H''}{k_1''} = \frac{q_H}{R_{UB}^2 k''} = \frac{q_H}{\int_{R_{LB}}^{R_{UB}} R_{UB}^2 (\ln R / R) \cdot f(R) \cdot dR} < \frac{q_H}{\int_{R_{LB}}^{R_{UB}} R^2 (\ln R / R) \cdot f(R) \cdot dR} = \frac{q_H}{k_1}$. Thus, the price of the premium service decreases from that in section 4. The broader insights from the model in section 4 (such as propositions 1-3, impact on profit and allocation, etc) are not altered. We can expect reality to lie between the cases of complete independence (figure 5) and strong correlation (figure 12). Thus, 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. Both the variants considered in this sub-section suggest that the broader insights from the model in section 4 do not change with correlation assumptions.

5.2 Sensitivity to Trace Parameters

The results from the numerical analysis in section 4.3.1 are based on parameters obtained from the trace analysis (specifically Boeing trace). In this section, we highlight the sensitivity of the results to the trace parameters. We first highlight the results of the trace analysis - The pdf of the number of requests for data objects is given by $f(R) = \frac{bc^b}{R^{1+b}}$ (a power law relationship); The hit rate of a cache when an LRU replacement policy is used is $H(S) = k_s.\ln(S)$; Hit rate experienced by an individual data object is $H(S,R) = k \cdot \ln S \cdot \ln R$.

Since various studies (Breslau et al. 1999) have shown that requests-rank relationship for requests follows a zipf distribution, our calibrated distribution (pareto) is likely to hold. What may vary from one trace to another is the parameter b and the upper bound on the number of requests, R_{UB} (c is a function of b and R_{UB}). Traces that are favorable for caching are those with low b and high R_{UB}. These traces have a relatively higher number of popular objects and popular objects tend to be extremely popular (see figure 13). We consider the impact of trace parameters on the results by considering two traces, which we shall call regular (high b, low R_{UB}) and favorable (low b, high R_{UB}). For the analytical model and numerical analysis presented in section 4, the results are as follows:

$P_L^*(fav) = P_L^*(reg)$; $P_H^*(fav) < P_H^*(reg)$; and $a^*(fav) > a^*(reg)$. For content publisher surplus functions of the form presented in the previous subsection (equation 8), the results are as follows - $P_L^*(fav) = P_L^*(reg)$; $P_H^*(fav) > P_H^*(reg)$; and $a^*(fav) < a^*(reg)$. The direction of the effects shown in figures 8-11 remain for all traces irrespective of the surplus functions.

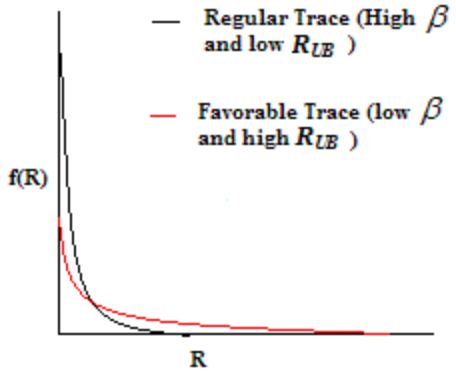

Figure 13. Two traces with different properties

When the relative density of popular objects increases, the IAP has incentives to provide more resources to publishers of such objects. Figure 7 indicates that for a given value of q, popular objects are more likely to be in the best effort service. The IAP thus allocates a larger fraction of the cache space to the best effort service to ensure that a large fraction of the data objects continue to be cached. This reduces the surplus for premium objects and thus it reduces P_H to compensate for the same.

In contrast, when the benefit function is as in equation 8, the indifference points are as shown in figure 12. Popular objects are more likely to be in the premium service because surplus from value-added features increases with R at a faster rate than hit rate benefits. When the relative density of popular objects increases, the IAP caters to their publishers by increasing the fraction of space allocated to the premium service. Since the surplus of the publishers increases, the IAP is able to charge higher per-object prices as well. Thus, the impact on the allocation and price depends on how popularity of data objects impacts the trade-off between hit rate benefits and value-added benefits. The bottom line is that sensitivity to value-added features will have to increase at a very high rate with R (faster than $R \cdot \ln R$) for prices and cache size of premium service to increase, else we expect both to decrease when relative density of popular objects increases.

In the cache simulations, we found that that hit rate varies as the logarithm of the cache size with an LRU replacement policy. This drives the result that the total price charged increases as the logarithm of the cache size for any service level (or per object price decreases as the inverse logarithm of cache size). The concavity of hit rate Vs. cache size is expected to hold for other replacement policies and thus total price is expected to increase non-linearly (concavely) with cache size irrespective of the replacement policy.

5.3 Contracting Costs for Content Publishers

The model presented in section 4 incorporates transaction costs of contracting for the IAP but not for the publisher. The model incorporating publisher transaction costs is relatively tedious and hence we only highlight the key insights from the model here. Including a transaction cost of contracting for publishers reduces their surplus from the service. The net result is that publishers only have incentives to contract with large IAPs. IAPs with large subscriber bases would handle a large number of requests for the publisher's content and thus benefits from the service can outweigh the contracting costs. On the other hand, smaller IAPs

will find it difficult to attract publishers because the contracting costs may outweigh the publisher's benefit from the service. 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 as the aggregators that facilitate 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 transaction costs of contracting and economies of scale and scope.

6. Conclusions

Quality of Service (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 great 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 for information goods, this suggests that 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.

There exists no previous research in the area of cache QoS pricing and capacity allocation. While IAPs currently deploy caches, they do not charge publishers for these services because of the best effort nature of these services. Best effort caching worked well as a caching paradigm a few years back. As electronic markets have matured over the last couple of years, publishers have developed new requirements (such as personalization, business intelligence) which current caching services do not meet. The best effort nature of caching is thus contributing to significant cache-busting by publishers. Our paper proposes a QoS framework wherein an IAP can provision value-added services to respond to the new needs of publishers and thus realize more efficient markets. QoS-based caching can help prevent a lot of the cache-busting that is prevalent today. In addition to increasing publisher welfare, appropriate pricing of these services can ensure that IAPs are also better off. Thus, this aligns the incentives of publishers and IAPs and increases overall market welfare. We have introduced a framework to determine optimal

pricing strategies for an IAP provisioning best effort and premium caching services. We also address the capacity allocation issues that arise from the provisioning of these services.

We find that value-added services would allow cache operators to price discriminate effectively. Additionally, we find that subscriptions to both the services would drop with falling bandwidth costs. This effect can be mitigated by either provisioning superior value-added services or through increased broadband penetration. The former allows the IAP to charge higher prices. Increased broadband penetration would likely increase Internet usage by end users as well as the volume of traffic on the Internet, thus reinforcing the value of caching and increasing adoption. For example, consider the fact that @Home (a broadband ISP now part of Comcast) had 0.3% of Internet subscribers but constituted 5% of Internet traffic.

Our analysis has shown that recent changes in publisher preferences 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). Resources may be diverted towards serving the maximum number of data objects from the premium cache. Services like Akamai's Edgesuite 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.

References

[1] America Online (AOL) Fast Facts, http://corp.aol.com/whoweare/fastfacts.html, Retrieved December 2002.

[2] J. Fritz Barnes and Raju Pandey, "CacheL: Language support for customizable caching policies," In Fourth International WWW Caching Workshop, San Diego, CA, USA, April 1999.

[3] H. K. Bhargava, V. Choudhary, and R. Krishnan, ``Pricing and Product Design: Intermediary Strategies in an Electronic Market," *International Journal of Electronic Commerce* Volume 5, Number 5 (Fall 2000), pp. 37-56.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," Proceedings of IEEE Infocom 1999, NY , March1999.

[5] L.E.J. Brouwer, "Über Abbildung von Mannigfaltigkeiten," *Mathematische Annalen* 71, 97-115. 1910.

[6] P. Cao and S. Irani, "Cost-aware WWW proxy caching algorithms," Proceedings of the 1997 USENIX Symposium on Internet Technology and Systems, 193-206.

[7] P. Cao, J. Zhang, and K. Beach, "Active cache: Caching dynamic contents (objects) on the web," Proceedings of the IFIP International Conference on Distributed Systems Platforms and Open Distributed Processing, The Lake District, England, September 1998.

[8] Y. M. Chan, J. Womer, J. K. MacKie-Mason and S. Jamin, "One Size Doesn't Fit All: Improving Network QoS Through Preference-driven Web Caching," Proceedings of the 27th Annual Telecommunications Policy Research Conference, Alexandria, VA, August 1999.

[9] J. Chuang and M. Sirbu, "Distributed Network Storage with Quality-of-Service Guarantees," *Journal of Network and Computer Applications* 23(3): 163-185, July 2000.

[10] J. Chuang, S. Kafka and K. Norlen, "Efficiency and Performance of Web Cache Reporting Strategies," Proceedings of IEEE International Workshop on Data Semantics in Web Information Systems, Singapore, December 2002.

[11] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, "Pricing in computer networks: Motivation, formulation and example," *IEEE/ACM Transactions on Networking*, vol. 1, December 1993.

[12] A. Datta, K. Datta, H. Thomas, D. VanderMeer, "WORLD WIDE WAIT: A Study of Internet Scalability and Cache-Based Approaches to Alleviate it," Georgia Institute of Technology Working Paper.

[13] A. Datta, K. Dutta, H. Thomas, D. VanderMeer, Suresha, K. Ramamritham, "Proxy-Based Acceleration of Dynamically Generated Content on the World Wide Web: An Approach and Implementation," Proceedings of ACM SIGMOD, June, 2002.

[14] R. Dewan, M. Freimer and A.Seidmann, "Organizing Distribution Channels for Information Goods on the Internet," *Management Science*, vol. 46, No. 4, April 2000.

[15] M. Feldman and J. Chuang, "Service Differentiation in Web Caching and Content Distribution," Proceedings of IASTED International Conference on Communications and Computer Networks, Cambridge MA, November 2002.

[16] A. Gupta, D. O. Stahl, and A. B. Whinston, "Priority Pricing of Integrated Services Networks," In Mcknight and Bailey, Eds., *Internet Economics*, MIT Press, 1997.

[17] T. Kelly, S. Jamin, and J. K. MacKie-Mason, "Variable QoS from Shared Web Caches: User-Centered Design and Value-Sensitive Replacement," in MIT Workshop on Internet Service Quality Economics, Cambridge, Massachusetts, December, 1999.

[18] T. Kelly, D. Reeves, "Optimal Web Cache Sizing: Scalable Methods for Exact Solution," 5th International Conference on Web Caching and Content Delivery, Lisbon, Portugal, May 2000.

[19] Y. Lu, A. Saxena, and T. F. Abdelzaher, ``Differentiated Caching Services; A Control-Theoretical Approach," International Conference on Distributed Computing Systems, Phoenix, Arizona, April 2001.

[20] MacKie-Mason, J., and Varian, H.R, "Pricing Congestible Network Resources," *IEEE Journal of Selected Areas in Communications*, 13, 7 (September 1995), 1141-49.

[21] B. Maggs, Vice President, Akamai. Personal Communication. 2002.

[22] Marchand, M. "Priority Pricing", *Management Science*. 20. 1974.

[23] Maskin, E., and Riley, J. (1984), "Monopoly with Incomplete Information", *RAND Journal of Economics*,15, 171-196.

[24] J. Mears, "CDNs are not just for content anymore," *Network World*, 01/14/02.

[25] Mendelson, H., and S. Whang "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research*, 38, 870-83, 1990.

[26] J. Mogul and P. Leach, "Simple hitmetering and usage-limiting for HTTP," RFC 2227, 1997.

[27] M. Mussa and S. Rosen, "Monopoly and Product Quality," *Journal of Economic Theory*, vol. 18, 1978.

[28] A. Myers, J. Chuang, U. Hengartner, Y. Xie, W. Zhuang, H. Zhang, A Secure, Publisher-Centric Web Caching Infrastructure. Proceedings of IEEE INFOCOM 2001, Anchorage AL, April 2001.

[29] D. Neven and J. F. Thisse, "Quality and Variety Competition," in *Economic Decision Making: Games, Econometrics and Optimisation*, J. Gabszewicz, J. Richard and L. Wolsey (eds), North-Holland, 1990.

[30] G. Pierre, I. Kuz, M. van Steen, and A. Tanenbaum. "Differentiated Strategies for Replicating Web Documents." *Computer Communications*, 24(2):232--240, Feb. 2001.

[31] S. Rao and E. R. Petersen, "Optimal Pricing Of Priority Services", *Operations Research*, Vol. 46, No. 1, January-February 1998.

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

[33] Stargate, Personal Communication. 2002.

[34] Sundararajan, A. "Non-linear pricing of information goods". Working paper, New York University, 2002.

[35] Web-Caching.com, Proxy Cache Comparison, http://www.web-caching.com/proxy-comparison.html, Retrieved March 2002.

[36] Web Characterization Repository, World Wide Web Consortium, http://repository.cs.vt.edu/, 2002.

[37] J. Yin, L. Alvisi, M. Dahlin, and C. Lin, "Using leases to support server-driven consistency in large-scale systems," *Proceedings of the 18th International Conference on Distributed Systems.* IEEE, May 1998.

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

[39] H. Zhu, T. Yang, "Class-based Cache Management for Dynamic Web Content," *IEEE INFOCOM*, 2001.

Appendix A1

Proof of Lemma 1

Equating q_L and q_H leads to the following equation:

$$\frac{(P_H - h - B)\ln(S_H)}{q_H}.(R\ln(R)) = \frac{(h + B - P_L)\ln(S_L)}{q_L}(R\ln(R)). \quad (3)$$

Note that if $q_L = q_H$ for any R in $[c_1, c_2]$, then either $\ln(R) = 0$ or $\frac{(P_H - h - B)\ln(S_H)}{q_H} = \frac{(h + B - P_L)\ln(S_L)}{q_L}$. The former is the simple and uninteresting case where $R = 1$ (the object has only 1 request). The latter implies that $q_L = q_H$ for all R in $[c_1, c_2]$. That is, if q_L and q_H ever meet then 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, $q_H(R) < q_L(R)$ 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 1 service). In fact, the region $q \in (q_H, q_L)$ represents subscribers that derive a positive surplus from both the services. We define q_{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: $q_{LH}(R) = \frac{R\ln R\{\ln S_L(h + B - P_L) + \ln S_H(P_H - h - B)\}}{q_L + q_H}$. This can be rewritten as $q_{LH} = \frac{q_L q_L + q_H q_H}{q_L + q_H}$. Thus, q_{LH} is a weighted average of q_L and q_H implying that the q_{LH} QIC lies in between the other two QICs. Publishers with objects of type $q > q_{LH}$ will join the premium service because they weigh the value-added features more. Those with $q < q_{LH}$ will choose the best effort service.

Publishers in the region (q_H, q_{LH}) all subscribe to the best effort service and yet derive positive surplus from joining the premium service. Similarly, publishers in the region (q_{LH}, q_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 q_H to move up and q_L to move down, without impacting q_{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 $q_H(R) \geq q_L(R)$. Now, if $q_L(R) > 0$, we know that all publishers located between $[0, q_L]$ 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, $q_L(R) > 0$ for some value of R guarantees the existence of subscribers for the best effort service. Thus, the condition is sufficient. Similarly, publishers with $q > q_H$ will derive a positive surplus from the premium service. Since $q_H \geq q_L$, all these publishers will also derive a negative surplus from the best effort service. Thus, publishers located between $[q_H, 1]$ will all subscribe to the premium service. Thus, if $q_H < 1$ for some *R*, then it is guaranteed that the premium service also has some subscriber. It follows that the market is segmented.

Appendix A2 - IAP Profit Function

The IAP's expected profit consists of revenues from charging for the 2 services and bandwidth savings from the cache. As explained in section 4.2.3, this is given by:

$p = R_H H(S_H)[P_H + B_{IAP}] + R_L H(S_L)[P_L + B_{IAP}] - C(q_H)$, 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 $q > q_H$ and $q \leq q_L$ 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 $\int_0^{q_L(R)} f(q)dq = q_L(R)$. Thus, the total number of requests for content in the best effort service is given by summing $Nq_L(R)Rf(R)$ for all values of R. Thus, $R_L = N\int_{c1}^{c2} q_L(R)Rf(R)dR$. Substituting the expression for $q_L(R)$, this gives us $R_L = \frac{Nkk_1(h + B - P_L)\ell nS_L}{q_L}$, where $k_1 = \int_{c_1}^{c_2} R^2 \cdot \ln R \cdot f(R)dR$ is a constant of integration.

Similarly, $R_H = N\int_{c1}^{c2}(1-q_H(R))Rf(R)dR = N\left[E(R) - \frac{kk_1(P_H - h - B)\ell nS_H}{q_H}\right]$, where $E[R]$ is the expected value of R. Substituting these expressions for R_H and R_L into the profit function, we get:

$$p = N\left[E(R) - \frac{kk_1(P_H - h - B)\ell nS_H}{q_H}\right]\cdot k_s \ell nS_H \cdot [P_H + B_{IAP} - T]$$

$$+ \frac{Nkk_1(h + B - P_L)\ell nS_L}{q_L} k_s \ell nS_L \cdot [P_L + B_{IAP} - T] - C(q_H)$$

Appendix A3 – Space Allocation

We look at the allocation decision contingent on the equilibrium prices derived in section 4.2. The IAP profit evaluated at (P_L^*, P_H^*) is given by

$$p = \frac{Nkk_1k_s}{q_L}\left(\frac{h + B + B_{IAP} - T}{2}\right)^2 (\ln(aS))^2$$

$$+ Nk_s \ln\{(1-a)S\}\left[\frac{E(R)}{2} + \frac{kk_1 \ln\{(1-a)S\}}{q_H}\left(\frac{h + B + B_{IAP} - T}{2}\right)\right]\left[\frac{E(R)q_H}{2kk_1 \ln\{(1-a)S} + \frac{h + B + B_{IAP} - T}{2}\right]$$

The FOC with respect to a:

$$\frac{\partial p}{\partial a} = \frac{Nkk_1k_s}{q_L}\left(\frac{h + B + B_{IAP} - T}{2}\right)^2 \frac{2\ln(aS)}{a}$$

$$- \frac{Nk_s}{(1-a)}\left[\frac{E(R)}{2} + \frac{kk_1 \ln\{(1-a)S\}}{q_H}\left(\frac{h + B + B_{IAP} - T}{2}\right)\right]\left[\frac{E(R)q_H}{2kk_1 \ln\{(1-a)S} + \frac{h + B + B_{IAP} - T}{2}\right]$$

$$- Nk_s \ln\{(1-a)S\}\left[\frac{kk_1}{q_H(1-a)}\left(\frac{h + B + B_{IAP} - T}{2}\right)\right]\left[\frac{E(R)q_H}{2kk_1 \ln\{(1-a)S} + \frac{h + B + B_{IAP} - T}{2}\right]$$

$$+ Nk_s \ln\{(1-a)S\}\left[\frac{E(R)}{2} + \frac{kk_1 \ln\{(1-a)S\}}{q_H}\left(\frac{h + B + B_{IAP} - T}{2}\right)\right]\left[\frac{E(R)q_H}{2kk_1 \ln\{(1-a)S\}^2(1-a)}\right] = 0$$

The FOC indicates that if the IAP increases a infinitesimally, both the hit rate and subscription to the best effort service goes up (first term), hit rate of the premium service decreases (second term), subscriptions to the premium service decreases (third term) but the per-object price for the

premium service is increased by the IAP (fourth term). Although, the per-object price increases, it can be shown that the total price paid by any publisher decreases with a, as expected.

Let us denote the 4 terms in the derivative as A, B, C and D respectively. Note that all these terms are functions of a. Then, the first order condition may be rewritten as:
A + B + C + D = 0; That is, $a = A + B + C + D + a = g(a)$.

$\frac{\partial g(a)}{\partial a}$ is defined for all a in [0,1] except a =0 and a =1. Thus, we restrict our attention to the interval $[e, 1-e]$. e is arbitrarily small, hence $a^* = e$ is interpreted as a non existent lower-level cache (i.e., $a^* = 0$) and similarly $a^* = (1-e)$ is interpreted as a non-existent premium cache. $g(a)$ is continuous in $[e, 1-e]$ and the interval $[e, 1-e]$ is compact and convex. Therefore, the existence of a solution is guaranteed by Brouwer's Fixed Point Theorem (Brouwer 1910). □