9-2008

Service Adoption and Pricing of Content Delivery Network (CDN) Services

kartik Hosanagar
University of Pennsylvania

John Chuang

Ramayya Krishnan

Michael D. Smith

Recommended Citation
Hosanagar, k., Chuang, J., Krishnan, R., & Smith, M. D. (2008). Service Adoption and Pricing of Content Delivery Network (CDN) Services. *Management Science, 54* (9), 1579-1593. http://dx.doi.org/10.1287/mnsc.1080.0875

Service Adoption and Pricing of Content Delivery Network (CDN) Services

Abstract
Content delivery networks (CDNs) are a vital component of the Internet's content delivery value chain, servicing nearly a third of the Internet's most popular content sites. However, in spite of their strategic importance, little is known about the optimal pricing policies or adoption drivers of CDNs. We address these questions using analytic models of CDN pricing and adoption under Markovian traffic and extend the results to bursty traffic using numerical simulations.

When traffic is Markovian, we find that CDNs should provide volume discounts to content providers. In addition, the optimal pricing policy entails lower emphasis on value-based pricing and greater emphasis on cost-based pricing as the relative density of content providers with high outsourcing costs increases. However, when traffic is bursty and content providers have varying levels of traffic burstiness, volume discounts may be suboptimal and may even be replaced by volume taxes. Finally, when there is heterogeneity in burstiness across content providers, a pricing policy that accounts for both the mean and variance in traffic such as percentile-based pricing is more profitable than traditional volume-based pricing (metering bytes delivered in a given time window). This finding is in contrast to the current practices of many CDN firms that use traditional volume-based pricing.

Keywords
content delivery, content delivery networks, CDN, pricing, media delivery, hosting, bursty traffic

Disciplines
E-Commerce | Operations and Supply Chain Management | Other Business

Service Adoption and Pricing of Content Delivery Network (CDN) Services

Kartik Hosanagar[*], John Chuang[**], Ramayya Krishnan[***], Michael D. Smith[***]

Abstract

Content Delivery Networks (CDNs) are a vital component of the Internet's content delivery value chain, servicing nearly a third of the Internet's most popular content sites. However, in spite of their strategic importance little is known about the optimal pricing policies or adoption drivers of CDNs. We address these questions using analytic models of the market structure for Internet content delivery.

We find that, consistent with industry practices, CDNs should provide volume discounts to content providers when traffic burstiness is similar across content providers. However, when different content providers have varying traffic burstiness, as expected in reality, CDNs should provide relatively lower volume discounts, even leading to convex price functions in some cases. Surprisingly, we also find that content providers with bursty traffic provision less infrastructure compared to those with lower burstiness, that CDNs are able to charge more in the presence of bursty traffic, and that content providers with bursty traffic realize lower surplus. Similarly, we find that a pricing policy that accounts for both the mean and variance in traffic such as percentile-based pricing does better than pure volume based pricing. Finally, we show that larger CDN networks can charge higher prices in equilibrium, strengthening any technology-based economies of scale.

Keywords: Content Delivery, Content Delivery Networks, CDNs, Pricing, Hosting, Infrastructure Sizing, Bursty Traffic.

Acknowledgements: We thank participants at the 37th Hawaii International Conference on Information System Sciences (HICSS-37) and the 2003 Workshop on Information Technology and Systems (WITS) for their comments and suggestions.

[*] The Wharton School of the University of Pennsylvania; email: kartikh@wharton.upenn.edu

[**] School of Information Management and Systems, University of California and Berkeley, Berkeley, CA; email: chuang@sims.berkeley.edu

[***] H. John Heinz III School of Public Policy and Management, Carnegie Mellon University; email: {rk2x,mds}@cmu.edu

1. Introduction

A Content Delivery Network (CDN) is a network of servers that cache or store web content (i.e., web pages and embedded objects) and intelligently deliver it to users based on their geographic location. CDN servers are typically collocated with Internet Service Providers (ISPs) with which the CDN has alliances. When users request content, the request is redirected to the nearest CDN server, where *nearness* is based on expected latency, which is in turn determined by geographical proximity, server load, and network conditions. By delivering content from the edge of the Internet, CDNs speed content delivery, circumvent bottlenecks and provide protection from sudden traffic surges that can bring down servers, rendering web sites unreachable.

CDNs are an important element of the digital supply chain for the delivery of information goods. The supply chain consists of Content Providers (CPs) that create the content; backbone and access networks that help transport the content, and CDNs that store and deliver the content to the end users. CDNs thus function as content storage and distribution centers performing similar functions to those performed by distributors/retailer warehouses in traditional supply chains. In 2000, CDN services were used by 31% of the 127 most popular Internet websites (Krishnamurthy et al 2001). Akamai dominates the industry, with an 80% market share. Other prominent CDNs include Cable & Wireless, Speedera and Mirror Image.

Due to increasing traffic on the Internet and a shift towards high bandwidth multimedia content, CPs must periodically resize and upgrade their server farms and bandwidth capacities. In addition, the high variability in document request patterns creates challenging capacity allocation problems for CPs. If they allocate capacity based on peak traffic, the capacity will sit idle most of the time. If they under-provision, then the performance and uptimes of their web sites

decrease, resulting in customer dissatisfaction and reduced revenue. For example, flash crowds[1] on Sep 11 overwhelmed media sites such as CNN and MSNBC, reducing site availability to close to 0%, and increasing response times to nearly 40 seconds when the sites finally were available[2]. Because of these trade-offs, CPs have had to choose intermediate capacity levels and accept occasional down times as a necessary evil.

CDNs provide CPs with a viable alternative to scale content delivery. CDNs improve the scalability of content delivery in three primary ways. First, CDNs achieve economies of scale in infrastructure costs by aggregating traffic across multiple customer sites. Second, aggregation reduces the impact of variability in demand for content, reducing infrastructure needs per site and improving content availability. Third, since there are several nodes from which the content can be served, no single point will be a bottleneck. Replication of content across delivery locations improves the availability of content, especially during flash crowds or Denial of Service (DoS) attacks.

CDNs have traditionally offered services that enabled CPs to deliver part of their content (typically rich content) through CDNs and the remainder on their own. However, partial site delivery implied that CPs still needed to maintain significant infrastructure to deliver content (and thus were unable to fully realize infrastructure cost reduction). Further, this resulted in high costs for coordinating partial content delivery and for integrating business intelligence regarding end users. In the recent years, several CDNs have introduced services that enable CPs to deliver entire websites from the edge servers. A well-known example of such a service is Akamai's Edgesuite. Conversations with CDN executives (Maggs 2002) reveal that they face challenges in determining how they should price these services, what factors influence service adoption by con-

[1] *Flash crowds* refer to sudden surges in demand for content that often bring down web servers
[2] See http://news.com.com/2100-1023-272873.html. Retrieved March 2003.

tent providers, and how factors such as traffic patterns impact service adoption and pricing. We use analytic models and simulations to address these questions in this paper. We find that optimal prices for these services provide volume discounts when content providers have similar traffic burstiness profiles. However, these volume discounts may no longer be optimal if content providers exhibit varying degrees of traffic burstiness. The most likely purchasers of these services are high volume web sites with low security requirements for their content. Larger CDN networks can charge higher prices in equilibrium thus enhancing any technology-based economies of scale. Finally, a pricing policy that accounts for both the mean and variance in traffic such as percentile-based pricing does better than pure volume based pricing

2. Literature Review

CDNs have been widely studied in the computer science literature. Nottingham (2000) discusses the development of a framework to formally define the role of surrogate origin servers such as CDNs. Dilley et al. (2002) provide an overview of Akamai's network infrastructure and the technical challenges involved in operating a CDN. Saroiu et al. (2002) compare properties of CDN workloads with workloads from other content delivery architectures. Gadde et al. (2000) explore the effectiveness of CDNs in the presence of conventional web proxy caching. Chen et al. (2002) propose a protocol for dynamic placement of replicas in a CDN.

A popular theme for research has focused on the redirection schemes used by CDNs. Client requests are redirected to CDN servers using either URL rewriting or DNS-based redirection (Krishnamurthy et al. 2001). With URL rewriting, the origin server rewrites URL links with CDN server addresses so that any click-throughs are directed to the CDN server. With DNS redirection, the CDN controls the nameserver of the CP and resolves the name to the IP address of a CDN server. The Time-To-Live (TTL) of these DNS mappings are typically kept small so that

the CDN can map any given URL to different servers based on network conditions. Krishnamurthy et al. (2001) verify that CDNs reduce average download times but find that DNS redirection adds additional overheads. Johnson et al. (2000) also find that CDNs provide improvements in latency but find that they do not always choose the optimal server from which to serve the content. Kangasharju et al. (2000) find that it is best to retrieve different data objects of a single web page from the same CDN server.

While the focus of this literature has generally been on the design of efficient CDN architectures, pricing and service adoption aspects of CDN services have generally been ignored, and Management Science research can make significant contributions in this regard. For example, Datta et al. (2003) motivate the importance of research on pricing of CDNs. Furthermore, managers in CDN firms face challenges in accounting for various technological factors and Internet traffic patterns while determining their pricing strategies.

While pricing of traditional Telecommunications services has been studied in the past, pricing of content delivery services is a relatively new and unexplored area. Mendelson and Whang (1990) have studied the pricing of priority computer services. Gupta et al. (1997) and Cocchi et al. (1993) have studied QoS pricing in the transmission domain (prioritized transmission of data packets based on QoS schemes such as Diffserv and Intserv). Hosanagar et al. (2002) have studied the optimal pricing of priority-based web proxy caching services. Our paper extends this stream of research on telecom pricing by studying service adoption and pricing of CDNs.

3. Model

Consider a CP indexed by i delivering content to users. Let X_i be a random variable denoting the number of requests to CP i in any given period.[3] In any period, the distribution of X is known a priori, but the realized value of X is unknown. The publisher can choose to serve this content directly by investing in infrastructure to process a mean of I requests per unit time. If it does so, its surplus from serving content is $U_{self}(X) = V(X) - C(I) - c \cdot L(I, X)$, where $V()$ is the CP's benefit from responding to X requests, $C()$ is the cost for maintaining the infrastructure (servers, bandwidth, software, etc) which is concave in I because of economies of scale, $L()$ is the number of lost requests which increases with X but decreases with I, and c is the cost of each lost request.

$V()$ includes all sources of revenue to the CP from its Internet operations (e.g., revenue from selling products on the Internet, indirect surplus from disseminating information). The CP faces a trade-off in determining the optimal infrastructure capacity I. The CP can choose a low capacity but will incur a high cost of lost requests, or it can reduce the number of lost requests by incurring high infrastructure costs. The net expected surplus from delivering content is

$$U_{self} = E[U_{self}(X)] = V - C(I) - c \cdot L(I) \tag{1}$$

where $L(I) = E[L(I, X)]$ and $V = E[V(X)]$. In this section, we assume that all agents (content providers and CDN) are risk neutral. The risk neutrality implies that the CPs (CDN) care only about the expected surplus (profit) and not about the variance. We discuss implications of this assumption in Section 4. The CP's decision problem, given risk neutrality, is $\max_I \{U_{self}(I)\}$. We denote the optimal infrastructure level as I^* and associated expected surplus as $U_{self}(I^*)$.

[3] In the subsequent model development we drop *the* subscript i for simplicity.

The CP can choose to deliver content from its own servers or through a CDN. The CDN is assumed to be a monopoly. The CP's surplus from delivering content through the CDN is given by $U_{CDN}(X) = V(X) + \tau(N) \cdot X - C_O - P(X)$. *V()* is defined as above; $\tau()$ is the per-request benefit from faster content delivery through a geographically distributed set of *N* CDN servers. We assume that τ is concave in *N*, implying diminishing returns in improvements in response time from a larger network size. C_o is cost of outsourcing content delivery (e.g., cost of sharing confidential data or cost of modifying content to facilitate delivery by CDN). This cost is assumed to vary across CPs. *P()* is the usage-based price the CP is charged by the CDN. Note that the CDN serves the CP's entire site, as is the case in Akamai's popular EdgeSuite product and that the CDN maintains sufficient capacity to nearly eliminate lost requests. Thus, the cost of the minimal infrastructure needed and the cost of few lost requests, *C()* and *L()* respectively, are both approximated to zero. Since a CP cannot precisely predict *X* in any period, it can compute the expected surplus $U_{CDN} = E[U_{CDN}(X)]$. The CP will choose the CDN if $U_{CDN} \geq U_{self}(I^*)$. Based on these subscription decisions, one can evaluate the optimal price function *P(X)* for the CDN.

We apply this model by first analyzing the CP's optimal infrastructure decision when provisioning content directly, and then by analyzing the CDN's optimal pricing decision.

3.1. Optimal Infrastructure Sizing

We begin this section with a brief discussion of the infrastructure resources required to service HTTP requests. In the HTTP protocol, exchange of data between a server and a client occurs after a TCP connection has been established. When a client attempts to establish a TCP connection, it begins by sending a SYN message to the server. The server acknowledges the SYN message by sending a SYN-ACK message to the client. In addition, the server creates a

socket for the incoming connection and places it in the SYN-RCVD queue. Subsequently, the client responds with an ACK message. Upon receiving the ACK message, the server moves the corresponding socket to the accept queue. The connection between the client and the server is then open. Whenever a web server process is ready to respond to a connection request, it executes an accept() system call and receives a socket number from the accept queue in return. In other words, requests are queued and wait for their turn to be processed. The sum of the SYN-RCVD and accept queues is also referred to as the backlog queue (requests waiting to be processed). The maximum value of the backlog queue is determined by the operating system kernel variable `somaxconn`. For further information on HTTP connection establishment, the reader is referred to Stevens (1990) and Banga and Druschel (1997).

Following previous literature (for example, Cao et al. 2003), we model a web server as an *M/G/1/K* Processor Sharing (PS) queuing system. That is, we assume that requests follow a Poisson process with mean arrival rate λ. The service time distribution is arbitrary. The queuing model treats the delivery system as a single server and the queue length as a finite exogenous parameter *K*, which is consistent with the observation that most commercial servers have similar `somaxconn` settings and most vendors recommend setting the queue size to `somaxconn`. Multithreading in the server is modeled by a processor sharing queuing discipline. Later in the paper, we will relax our assumptions to include multiple servers and a bursty, as opposed to Poisson, arrival process for requests.

We model the CP's infrastructure cost as: $C(I) = a \cdot I - b \cdot I^2$, $(I \le a/2b)$, which captures the concavity between *I* and cost. In this formulation, a large value for *a* would indicate high infrastructure costs and a large value for *b* would indicate significant economies of scale. For an *M/G/1/K*PS* queuing system, the expected number of lost requests is given by

$L(I) = \dfrac{\lambda\left(1-\dfrac{\lambda}{I}\right)\left(\dfrac{\lambda}{I}\right)^{K}}{1-\dfrac{\lambda}{I}^{K+1}}$. A well-known result in queuing theory indicates that *I* should be greater than λ, else the system "blows up".[4] Thus, the region of interest for our model is $I \in (\lambda, \dfrac{a}{2b}]$. The CP can choose a high infrastructure level and reduce the expected number of lost requests *L(I)* but will incur high infrastructure cost, *C(I)*. It can be verified that $\dfrac{\partial L(I)}{\partial I} < 0$ and thus the CP has to trade off the benefits and costs of added infrastructure. Under this model, the CP's decision problem is

$$\max_{I} U_{Self}(I) = \max\left\{ V - (a \cdot I - b \cdot I^{2}) - c \cdot \frac{\lambda\left(1-\dfrac{\lambda}{I}\right)\left(\dfrac{\lambda}{I}\right)^{K}}{1-\dfrac{\lambda}{I}^{K+1}} \right\} \tag{2}$$

and the associated first-order necessary condition is given by:

$$-\mathrm{a} + 2\mathrm{b} \cdot \mathrm{I} - \frac{\mathrm{c} \cdot \lambda^{\mathrm{K}+1}}{I^{\mathrm{K}+1} - \lambda^{\mathrm{K}+1}} + \frac{c \cdot (K+1) \cdot \lambda^{\mathrm{K}+1} (I-\lambda) I^{K}}{(I^{\mathrm{K}+1} - \lambda^{\mathrm{K}+1})^{2}} = 0 . \tag{3}$$

While this polynomial lacks a closed form solution, we can use the conjugate pairs theorem from calculus (Currier 2000) to analyze the properties of I^{*} (the optimal infrastructure level). The theorem states that for the maximization problem $\max_{x} F(x,a)$, the derivative $\dfrac{\partial x^{*}}{\partial a}$ and the cross partial F_{xa} have the same sign. The following results follow:

i) If the cost of infrastructure increases, I^{*} decreases.

[4] The intuition is that if $I \le \lambda$, the mean arrival rate is greater than the mean service rate and the server keeps lagging further and further behind.

$U_{Ia} = -1$. This implies that $\frac{\partial I^*}{\partial a} < 0$. As expected, if the infrastructure costs (cost of processing and bandwidth) decrease, the optimal level of investment in infrastructure increases.

ii) If there are significant economies in scale in content delivery, I^* increases.

$U_{Ib} = 2I > 0$. Thus $\frac{\partial I^*}{\partial b} > 0$. That is, if server or bandwidth sellers provide high volume discounts, infrastructure levels of CPs will increase.

iii) If a content provider's cost of losing requests is high, I^* is correspondingly higher.

Proof: $U_{Ic} = \frac{(K+1)\cdot\lambda^{K+1}(I-\lambda)I^K}{(I^{K+1}-\lambda^{K+1})^2} - \frac{\lambda^{K+1}}{I^{K+1}-\lambda^{K+1}}$. We know from the first order condition that $-\{a-2b\cdot I\} - \frac{c\cdot\lambda^{K+1}}{I^{K+1}-\lambda^{K+1}} + \frac{c\cdot(K+1)\cdot\lambda^{K+1}(I-\lambda)I^K}{(I^{K+1}-\lambda^{K+1})^2} = 0$. Since $-\{a-2b\cdot I\} < 0$ (the rate at which infrastructure costs increase with infrastructure level *I*), it follows that $-\frac{c\cdot\lambda^{K+1}}{I^{K+1}-\lambda^{K+1}} + \frac{c\cdot(K+1)\cdot\lambda^{K+1}(I-\lambda)I^K}{(I^{K+1}-\lambda^{K+1})^2} = cU_{Ic} > 0$. Since $c > 0$, it follows immediately that $U_{Ic} > 0$. From the conjugate pairs theorem, $\frac{\partial I^*}{\partial c} > 0$. In other words, if the cost (c) of losing a request increases, the optimal infrastructure level also increases.

iv) If the arrival rate of requests λ increases, I^* increases.

Proof: This statement follows from conjugate pairs theorem if $U_{I\lambda} > 0$ is true. Computing the cross partial with respect to I, λ and simplifying, $U_{I\lambda} = \frac{c(K+1)\cdot\lambda^K I^K\left\{K(I-\lambda)(I^{K+1}+\lambda^{K+1}) - 2I\lambda(I^K-\lambda^K)\right\}}{(I^{K+1}-\lambda^{K+1})^3}$. Thus, $U_{I\lambda} > 0$ if and only if $K(I-\lambda)(I^{K+1}+\lambda^{K+1}) - 2I\lambda(I^K-\lambda^K) > 0$. This can be restated as

$U_{I\lambda} > 0$ iff $Kp^{K+2} - (K+2)p^{K+1} + (K+2)p - K > 0$ (A)

where $p = I/\lambda$. For $I >> \lambda$, it follows that $p >> 1$. Thus

$(K+2)p > K$ (B)

Also, for large K (queue size), we know the following is true: $p > 1 + \frac{2}{K}$. This can be restated as:

$Kp^{K+2} > (K+2)p^{K+1}$ (C)

Adding (B) and (C) yields $Kp^{K+2} + (K+2)p > (K+2)p^{K+1} + K$. Combining this result with (A), it follows that $U_{I\lambda} > 0$. That is, $\frac{\partial I^*}{\partial \lambda} > 0$. QED.

Supplemental numerical tests using parameter values conforming to typical bandwidth and hosting costs were conducted to determine the relationship between λ and I^*. For the infrastructure cost function, $C(I) = a \cdot I - b \cdot I^2$, we assume that *a=3.56* and *b=0.000043*. These parameter values roughly correspond to current infrastructure costs. For example, under these parameter values the cost of serving 233 requests/min is $804 per month. If we assume that the average size of the response to a request is 50 Kbytes, this implies that the cost of serving data at 1.55 Mbps is $804 per month. This is reasonable given the cost of a T1 connection (approximately $400 per month) and maintaining a workstation. Likewise, the cost of serving 6,975 requests per minute is $22,042, which is approximately the cost of a T3 connection and the associated cost of maintaining a server. Finally, the cost of serving 23,255 requests per minute is $57,208 per month, roughly equivalent to the cost of an OC3 connection. These costs are also comparable to managed hosting costs at the time of this study.

We assume that the cost of a lost request, *c*, is $10. This is based on an assumption that 10% of visitors purchase products/services, the average purchase is $100, and a customer leaves

a website if a request does not go through. Finally, we assume that the queue size, *K*, (for requests waiting to be processed) is 8 requests. Our settings for *c* and *K* are biased towards incurring high cost of lost requests in order to eliminate boundary solutions where I^* is set to the lower bound λ. This is because we are interested in the nature of the relationship for interior solutions. Figure 1 shows the optimal infrastructure level (in requests/min) for different arrival rates ranging from 5,000 to 20,000 requests per minute. The relationship is approximately linear.

To test for robustness, we repeated the numerical analysis for a variety of other settings for buffer size *K* and cost of lost requests *c*, and found that the relationship is approximately linear in all cases. For example, Figure 2 shows the relationship for the case where $\{a = 3.46; b = 0.000043; K = 4; c = 20\}$. Note that the special case where $I^* = \lambda$ (boundary solution) is also linear

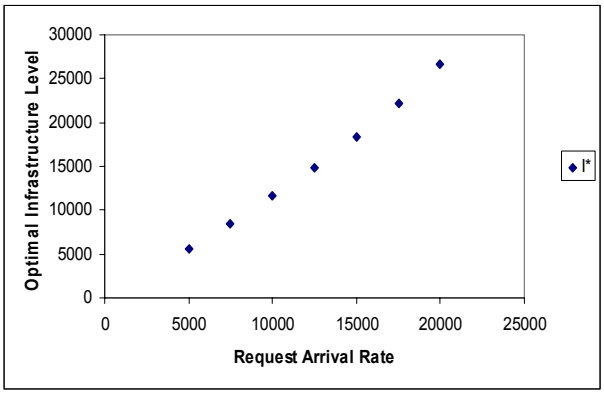

Figure 1: Optimal Infrastructure Level versus Arrival Rate (Case 1)

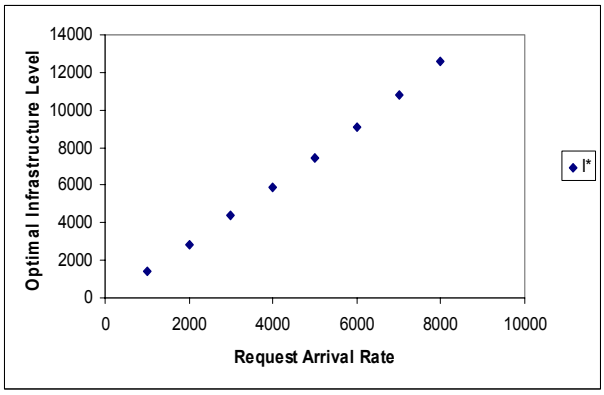

Figure 2. Optimal Infrastructure Level versus Arrival Rate (Case 2)

3.1.1 Multiple Servers

In this subsection, we relax the assumption of a single server system and numerically evaluate the characteristics of a multiple server system. For illustration purposes, we test a three server system. The queue size is assumed to be 5, the cost of a lost request is assumed to be $10, and the remaining settings are as before, i.e., $\{a = 3.46; b = 0.000043; K = 5; c = 10\}$. The optimal infrastructure level for different arrival rates is plotted in Figure 3. The relationship continues to

be linear. As is intuitive, the optimal infrastructure level for each server (I^*) can now be lower than the mean arrival rate, λ, as three servers are sharing the load.

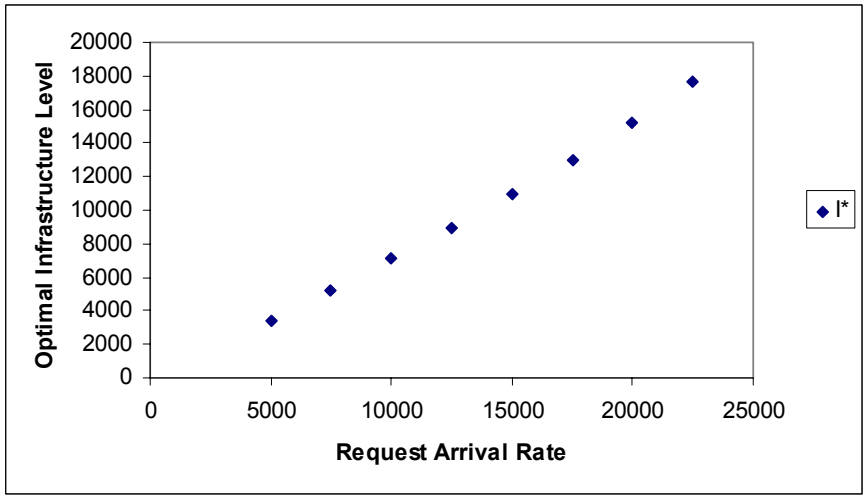

Figure 3. Optimal Infrastructure Level Vs. Arrival Rate (with Three Servers)

3.2. CDN Pricing Problem

As stated earlier, the CP's surplus from choosing a CDN is given by

$$U_{CDN}(X) = V(X) + \tau(N) \cdot X - C_O - P(X) \tag{4}$$

The CP does not know exactly how many requests (X) will be made for its content in any period, but can compute the expected surplus given by

$$U_{CDN} = E[U_{CDN}(X)] = V + \tau(N) \cdot \lambda - C_O - E[P(X)] \tag{5}$$

Given any price function $P(X)$, the CP can compute its expected surplus. The CP chooses the CDN if $U_{CDN} \geq U_{self}(I^*)$. Substituting equations (1) and (5) into this condition, a CP with arrival rate λ subscribes to the CDN if

$$C_O \leq \tau(N) \cdot \lambda + C(I^*) + c \cdot L(I^*) - E[P(X)] \tag{6}$$

Since the outsourcing cost C_o varies across CPs, we denote *H()* as the cumulative distribution function of C_o and *h()* as the associated probability density function. The probability that a CP with mean arrival rate λ subscribes to a CDN is then given by $H\left(\tau(N)\cdot\lambda + C(I^*) + c\cdot L(I^*) - E[P(X)]\right)$. If $g(\lambda)$ denotes the number of CPs with mean arrival rate λ, then the expected number of these CPs subscribing to the CDN is given by $g(\lambda)H\left(\tau(N)\cdot\lambda + C(I^*) + c\cdot L(I^*) - E[P(X)]\right)$. Any subscribing CP pays *P(X)* for a realized level of requests *X*. Since *X* is not known a priori, the CDN does not know its realized profit in any period associated with a price function *P(X)*. Under the standard assumption of zero marginal costs, the CDN's expected profit is given by

$$\pi = \int_{AllSubscribers}\int_{X} \Pr ob(X \mid \lambda)\cdot P(X)$$

$$= \int_{\lambda} g(\lambda)H\left(\tau(N)\cdot\lambda + C(I^*) + c\cdot L(I^*) - E[P(X)]\right)\left(\int_{X} P(X)\cdot\left(\frac{e^{-\lambda}\lambda^{X}}{X!}\right)dX\right)d\lambda \qquad (7)$$

Risk neutrality: A risk neutral CDN chooses the price function *P(X)* in order to maximize its expected profit (π). Note that the implication of the CPs' risk neutrality is that they make their subscription decision based on *E[P(X)]* (as in equation 6). Similarly, the risk neutrality of the CDN implies that it computes its expected profit by evaluating *E[P(X)]* for each subscribing CP. Thus, the CDN can achieve the same subscription levels and expected profits by charging each CP a fixed amount equal to its expected price *E[P(X)]*. This is illustrated in Figure 4. Consider the optimal price function, denoted by $P^*(X)$ and a CP with mean arrival rate λ_1. In each period, the CP receives a stochastic number of requests, *X*, and pays *P(X)* in that period. Over a long period of time, the CP expects to receive a mean of λ_1 requests per period and expects to pay *E[P(X)]*. In fact, if the CDN offers an alternative pricing scheme, wherein it charged the CP a

fixed amount *E[P(X)]* per period, the CP would still make the same subscription decision. Similarly, the CDN could charge all CPs with arrival rate λ_2 the corresponding expected price of *E[P(X)]* per period as shown in the Figure (note *E[P(X)]* is different for CPs with mean λ_1 and λ_2). Thus, for any optimal price function $P^*(X)$, there exists a corresponding "mean-usage-based" price function P_λ, obtained by following the trajectory of *E[P(X)]* for different values of λ, that achieves the same results. We can use this observation to simplify the problem to that of determining the optimal P_λ and then determining a corresponding *P(X)*.

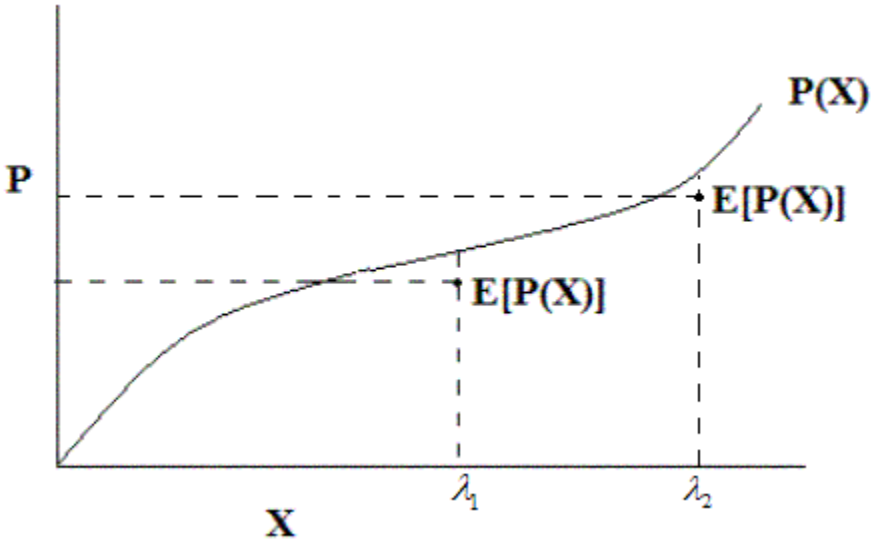

Figure 4. Pure usage based price and mean price

Now consider CPs with mean arrival rate λ_1. The CDN charges all such CPs a fixed price $P_{\lambda 1}$. The CDN's expected profit from these CPs is given by

$$\pi = g(\lambda_1)H\left(\tau(N)\cdot\lambda_1 + C(I^*) + c\cdot L(I^*) - P_{\lambda 1}\right)\left(P_{\lambda 1}\right) \tag{8}$$

The optimal price $P_{\lambda 1}$ obtained directly by applying the necessary first order condition is given by the solution to the following equality: $P_{\lambda 1} = \dfrac{H\left(\tau(N)\cdot\lambda_1 + C(I^*) + c\cdot L(I^*) - P_{\lambda 1}\right)}{h\left(\tau(N)\cdot\lambda_1 + C(I^*) + c\cdot L(I^*) - P_{\lambda 1}\right)}$.

The optimal "mean-usage-based" price function, P_λ, is thus given by

$$P_\lambda = \frac{H\left(\tau(N)\cdot\lambda + C(I^*) + c\cdot L(I^*) - P_\lambda\right)}{h\left(\tau(N)\cdot\lambda + C(I^*) + c\cdot L(I^*) - P_\lambda\right)} \tag{9}$$

A special case with uniform distributions: To illustrate a few properties of the optimal price function, we make the following two assumptions:

(1) The outsourcing cost, C_o, is uniformly distributed in [0,1]. That is, *H(x) =x* and *h(x) = 1*.

(2) The optimal infrastructure level I^* for a CP with mean arrival rate λ is given by $I^* = i_o\lambda$, where i_o is a constant. Assumption 2 is consistent with the numerical results in Section 3.1, Figures 1 and 2. Assumption 1 is for convenience and will be relaxed below to test its impact on our results. Under these assumptions, the optimal price function obtained by simplifying equation (9) is as follows:

$$P_\lambda = \frac{1}{2}\left[\tau(N) + ai_o + \frac{c(i_o - 1)}{i_o^{M+1} - 1}\right]\lambda - \frac{b\cdot i_o^2}{2}\lambda^2 \tag{10}$$

Note that the price does not depend on $g(\lambda)$, the distribution of mean arrival rate across CPs. This is because the CDN can observe the mean arrival rate and customize the price for each unique value of λ, and thus does not care about the distribution of CP mean arrival rates. A usage-based price function, $P^*(X)$ for which the *E[P(X)]* trajectory is given by equation (10) is:

$$P(X) = \frac{1}{2}\left[\tau(N) + ai_o + bi_o^{\,2} + \frac{c(i_o - 1)}{i_o^{M+1} - 1}\right]X - \frac{b\cdot i_o^2}{2}X^2 \tag{11}$$

To verify that equation (11) represents an optimal usage-based price function, assume that there is a different price function, $P^A(X)$ that performs better than *P(X)*, i.e., yields a higher expected profit than *P(X)*. In that case, the corresponding "mean-usage-based" price function represented

by $E[P^A(X)]$ should also provide higher expected profit than $E[P(X)]$. However, this cannot be true since equation (10) represents the optimal "mean-usage-based" price. This proof by contradiction shows that equation (11) represents an optimal usage-based price function.

The following observations can also be made regarding the optimal pricing policy:

a) *Volume discounts*: It is straightforward to show that $\frac{\partial P(X)}{\partial X} > 0$ and $\frac{\partial^2 P(X)}{\partial X^2} < 0$ for the relevant range of X. Thus, the optimal pricing policy entails volume discounts to CPs. This is consistent with Akamai's pricing statement:

"…Customers commit to pay for a minimum usage level over a fixed contract term and pay additional fees when usage exceeds this commitment. Monthly prices currently begin at $1,995 per megabit per second, with discounts available for volume usage."

Equation (11) indicates that the volume discounts essentially follow from the economies of scale in content delivery costs ($b>0$). In other words, if bandwidth sellers reduce their volume discounts, so can the CDN.

b) *Market power*: Since $\frac{\partial P}{\partial N} > 0$, larger CDNs are able to charge higher prices in equilibrium.

c) *Returns to scale*: Since $\frac{\partial^2 P}{\partial N \partial X} > 0$, a larger CDN can extract a higher increase in price than a smaller CDN for the same increase in volume of traffic. That is, given that the amount of traffic handled by CPs is on the rise, a larger CDN is able to leverage this trend more effectively.

d) *Subscription decision*: A CP with mean arrival rate λ subscribes to the CDN if $C_o \le \frac{1}{2}\left[\tau(N) + Ai_o + \frac{c(i_o - 1)}{i_o^{M+1} - 1}\right]\lambda - \frac{B \cdot i_o^2}{2}\lambda^2$. This is obtained by substituting the optimal price function into the subscription condition in equation (6). As seen in Figure 5, CPs likely to sub-

scribe to a CDN are those with high volume of traffic and low content delivery outsourcing cost (for example, content with minimal data confidentiality requirements).

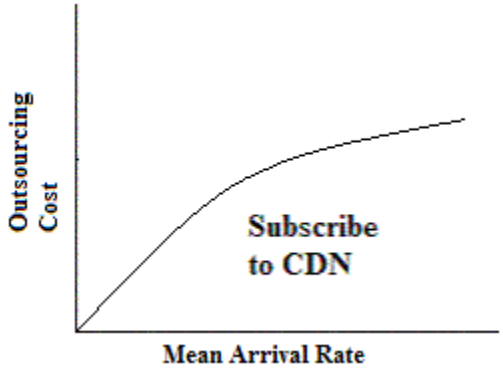

Figure 5: CP subscription decision

e) *Impact of technology choice*: We solve the same model as above but introduce a skew in the distribution of outsourcing cost across CPs by setting $H(C_o) = C_o^{\,2}; h(C_o) = 2C_o$. Relative to the uniform distribution, this distribution assumes that there are more CPs with high outsourcing cost (see Figure 6). The new solution is given by $P(X) = \frac{1}{3}\left[\tau(N) + Ai_o + \frac{c(i_o - 1)}{i_o^{M+1} - 1}\right]X - \frac{B \cdot i_o^2}{3}X^2$. This price is lower than in equation (11), suggesting that the price decreases as the relative number of CPs with high outsourcing costs increases.

Outsourcing cost, C_o, may include cost of sharing confidential information with a third party, or the transactions cost of interfacing with a third party or modifying content in order to enable delivery by the third party. For example, in the context of modifying content to facilitate delivery by CDN, switching to Akamai would require a CP to make its content ESI (Edge Side Includes – a technology developed by Akamai to enable edge delivery) compatible. As mentioned in Section 2, the CDN may choose either URL rewriting or DNS redirection as the technology for directing requests to CDN servers. Krishnamurthy et al (2001) found that DNS redirection adds additional overhead and URL rewriting is thus more efficient in terms of users' real-

ized response times. However, URL rewriting would entail higher outsourcing costs for CPs because of the significant cost incurred in modifying their entire content. This can result in lower prices. Thus, the CDN will need to trade-off efficiency-based benefits of any technology with the outsourcing costs imposed on the CPs.

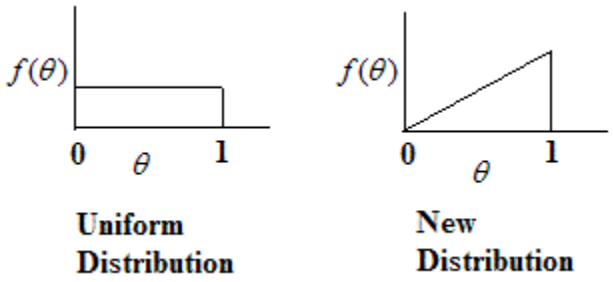

Figure 6: Negative skew in distribution of outsourcing cost

f) *Impact of bandwidth cost*: As bandwidth, memory and processor costs decline, the price that the CDN can charge will also decrease.

3.3. Modeling Bursty Traffic

The model presented in Sections 3.1 and 3.2 assume that requests for content at a web server follow a Poisson arrival process. However, some web traffic engineering studies suggest that web traffic exhibits bursts that cannot be captured by a Poisson arrival process (Crovella and Bestavros 1996). Furthermore, a feature of a Poisson process is that the burstiness reduces with increasing mean arrival rates. For example, one measure of burstiness — standard deviation/mean = $1/\sqrt{\lambda}$ — clearly decreases as arrival rate increases. In real-world traffic, burstiness tends to remain the same at high arrival rates too.

In order to model traffic burstiness, we assume request arrivals follow a Markov Modulated Poisson Process (MMPP). MMPP is commonly used to model bursty traffic to communications systems such as web servers (Scott et al. 2003, Anderson et al. 2003). MMPP is a doubly stochastic Poisson process in which the arrival rate is given by an *m*-state Markov process. At

any given instant, the system can be in any one of the *m* Markovian states. When the Markov chain is in state *i*, arrivals follow a Poisson process with arrival rate λ_i. Bursts can be captured by modeling a system transition to a state with very high arrival rate. The system is specified by the following matrices:

1) $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_m)$, where λ_i denotes the mean arrival rate in state *i*.

2) $R = \begin{bmatrix} -\sigma_1 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & -\sigma_2 & \dots & \sigma_{2m} \\ . & . & . & . \\ \sigma_{m1} & \sigma_{m2} & & -\sigma_m \end{bmatrix}$, where R is the $m \times m$ transition rate matrix of the phase process underlying the MMPP. In the matrix, σ_{jk} denotes the probability of a transition from state *j* to state *k*. In addition, we define the following matrices in order to compute the loss probability:

3) $q = (q_1, q_2, ..., q_m)$: m-dimensional vector containing the limiting state probabilities of the phase process.

4) $\pi(0)$: m-dimensional vector whose *j*th element is the probability, at the imbedded epochs, of having 0 users in the system and being in state *j*. $\pi(0)$ can be numerically computed as demonstrated in Baiocchi and Blefari-Melazzi (1993).

5) *e*: m-dimensional unit vector whose elements are all equal to 1.

$\overline{\lambda}$, the mean number of requests in a unit time period is given by $\overline{\lambda} = q_1\lambda_1 + q_2\lambda_2 + ... + q_m\lambda_m = q\Lambda e$.

The loss probability for an MMPP system can be computed as follows (Baiocchi and Blefari-Melazzi 1993):

$\Pr(Loss) = 1 - \dfrac{(1+\pi(0)(\Lambda - R)^{-1} eI)^{-1}}{q\Lambda e / I}$, where $q\Lambda e / I$ is a measure of the offered traffic and $(1+\pi(0)(\Lambda - R)^{-1} eI)^{-1}$ is a measure of the carried traffic. The expected number of lost requests is given by $L(I) = \bar{\lambda} \Pr(Loss) = \bar{\lambda} - \dfrac{I}{(1+\pi(0)(\Lambda - R)^{-1} eI)}$. It is straightforward to show that increasing I reduces $L(I)$.

$\dfrac{\partial L(I)}{\partial I} = -\left\{ \dfrac{1 - I^2 (\Lambda - R)^{-1} e \dfrac{\partial \pi(0)}{\partial I}}{(1+\pi(0)(\Lambda - R)^{-1} eI)^2} \right\}$. At small values of I, there is a significant reduction in number of lost requests from increasing I. However, for large values of I, the gains are much smaller (i.e., decreasing marginal returns).

3.3.1 Optimal Infrastructure Sizing

In this section, we consider a 2-state MMPP. The arrival matrix and the transition matrix are given by $\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ and $R = \begin{bmatrix} -\sigma_{12} & \sigma_{12} \\ \sigma_{21} & -\sigma_{21} \end{bmatrix}$ respectively. The mean $\bar{\lambda}$ and variance Ψ of the number of requests in a unit time period are given by: $\bar{\lambda} = q\Lambda e = \dfrac{(\lambda_1 \sigma_{21} + \lambda_2 \sigma_{12})}{(\sigma_{21} + \sigma_{12})}$, and

$$\Psi = \bar{\lambda}\left(1 + \frac{2(\lambda_1 - \lambda_2)^2 \sigma_{12}\sigma_{21}}{(\sigma_{12} + \sigma_{21})^2 (\lambda_1 \sigma_{21} + \lambda_2 \sigma_{12})} - \frac{2(\lambda_1 - \lambda_2)^2 \sigma_{12}\sigma_{21}}{(\sigma_{12} + \sigma_{21})^3 (\lambda_1 \sigma_{21} + \lambda_2 \sigma_{12})} (1 - e^{-(\sigma_{12} + \sigma_{21})}) \right).$$

Poisson traffic is a special case of MMPP with $\lambda_1 = \lambda_2$. On the other hand, a burst in traffic is modeled by assuming a very large value of λ_2 along with a non-zero probability of transitioning to state 2. We set $\lambda_2 = 10\lambda_1$ and $(q_1 = 0.9, q_2 = 0.1)$ as the MMPP parameters. In other words, the mean arrival rate during bursts is ten times the regular mean arrival rate and the system bursts 10% of the time. Different values of $\bar{\lambda}$ are simulated by varying λ_1. Further, when the mean arrival rate

$\bar{\lambda}$ is increased, we also change σ_{12} in order to maintain constant burstiness (constant value for $\sqrt{\Psi}/\bar{\lambda}$). This addresses the issue of decreasing burstiness with increasing arrival rates associated with Poisson traffic modeling. The loss probability and the optimal infrastructure level can be numerically computed for any given set of MMPP parameters.

Figure 7 presents the optimal infrastructure level with Poisson traffic and MMPP traffic for given mean arrival rates. Counter intuitively, we find that the CP's optimal infrastructure level with bursty traffic is lower than that with Poisson traffic (note that for a given mean arrival rate, MMPP has much higher variance than the Poisson traffic). Furthermore, the difference between the Poisson optimal infrastructure level and the MMPP infrastructure level increases as mean arrival rate increases. This also seems counterintuitive because as arrival rates increase, Poisson traffic is far less bursty than the MMPP traffic.

To help explain the result, consider a specific point on the graph. For the case where the mean arrival rate is given by 1,000, the corresponding values of (λ_1, λ_2) are given by (526.31, 5,263.15). That is, the CP faces a mean arrival rate of 526.31 requests per period approximately 90% of the time and faces 5,263.15 requests per period 10% of the time. The computed optimal infrastructure level of 941 requests per minute is sufficient to handle the state associated with low arrivals but is insufficient to handle bursts. However, small increases in infrastructure levels do not have much impact in reducing the number of lost requests, but only increase the infrastructure cost. This is because most of the lost requests are associated with state 2, which cannot be reduced unless service rate increases substantially. In order to see a marked reduction in lost requests, the infrastructure level has to be raised above 5,263 so that state 2 does not completely overwhelm the service center. However, this also raises the cost substantially. Thus, the high disparity between arrival rates in the two states (which follows the definition of a burst) implies

that the CP has to accept downtime during the high bursts. Because of this, the optimal infrastructure level is driven by λ_1 and not $\overline{\lambda}$. Since $\lambda_1 < \overline{\lambda}$, the optimal infrastructure level is also lower than that with Poisson traffic. Even with MMPP traffic, the optimal infrastructure level continues to be nearly linear with the mean arrival rate. As expected, the CP loses a large number of requests with highly bursty traffic and the CP's surplus with MMPP traffic is lower than with Poisson traffic (conditional on optimal infrastructure sizing in both situations). In Figure 8, we plot the "net cost" (the sum of infrastructure cost and cost of lost requests) to a CP. The net cost is higher with MMPP traffic despite lower infrastructure level because of the significantly higher loss of requests.

Our model assumes that the CP's utility depends on the expected number of lost requests and that this cost is linear in the number of lost requests. However, if the CP's cost is convex in the number of lost requests or if the CP's utility depends on *E[maximum number of lost requests]*, then the infrastructure level with MMPP will be higher than indicated in Figure 7. However, high bursts will continue to negatively impact the CP's surplus in either case.

We also found that the optimal infrastructure gradually increases as we increase the limiting state probability of being in state 2 (from 0.1 converging to 1) while decreasing the probability of being in state 1. The mean arrival rate was kept at 1,000 requests per minute. The infrastructure level approaches the optimal infrastructure with Poisson traffic and exceeds it as the probability of being in state 2 increases. For example, the optimal infrastructure level with $\{q_1 = 0.1, q_2 = 0.9, \lambda_1 = 109.89, \lambda_2 = 1098.9, \overline{\lambda} = 1000\}$ was 1,502 requests per minute. Note however that the case where the MMPP system is in a high arrival state with high probability and in a low arrival state with low probability is the reverse of bursty traffic patterns and does not model reality.

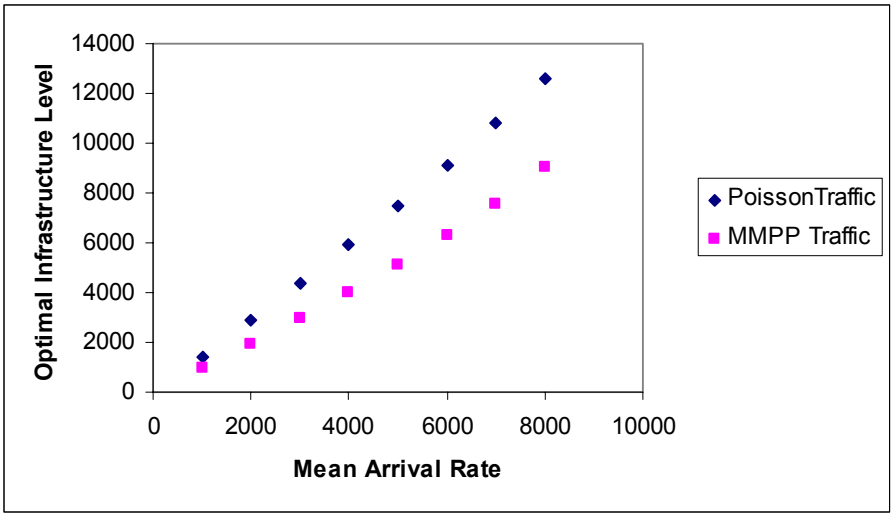

Figure 7: Optimal Infrastructure Level for Poisson and MMPP traffic (constant burstiness)

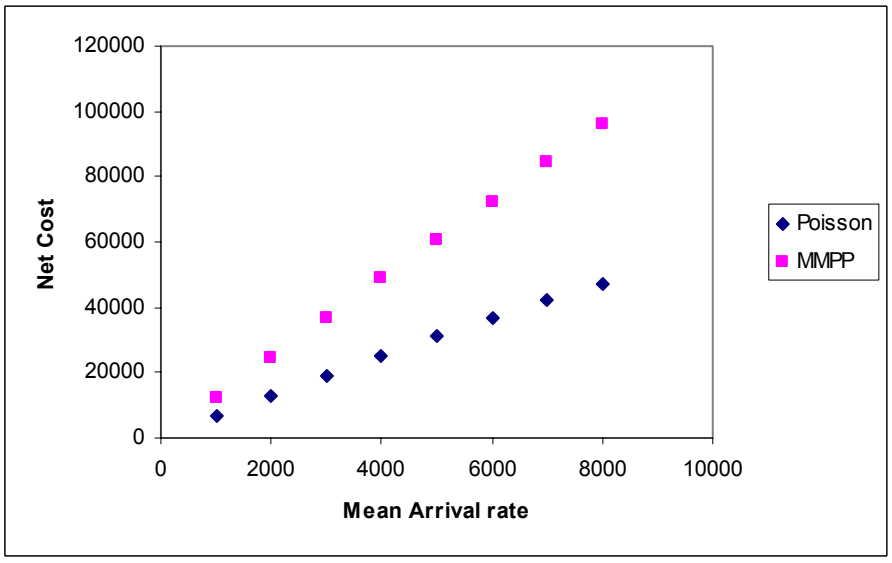

Figure 8: Net cost incurred by a CP

3.3.2 CDN's Optimal Pricing Policy

Using the arguments presented in deriving equation (9), we can derive an analytical expression for the optimal price function if all CPs have MMPP traffic with the same burstiness but different means (i.e., they effectively have infrastructure given by $I^* = i_0\lambda$, where i_o is the same constant for all CPs). However, when CPs have different burstiness levels (hence the infrastructure scaling constant i_o varies across CPs), it is difficult to analytically derive the optimal price

function. We numerically computed the optimal usage-based price function with a population of 1000 CPs for three cases: 1) All 1000 CPs have Poisson distributed traffic. 2) All 1000 CPs have MMPP traffic with parameters as specified in Section 3.3.1. 3) Mixed traffic: 500 CPs have Poisson traffic and 500 CPs have MMPP traffic.

The mean arrival rates for the CPs are drawn from a Uniform distribution in [1000, 8000]. All other parameters such as cost of lost requests, infrastructure cost, etc. are the same as those used in Figues 2, 7 and 8. To simplify computation, we restricted attention to quadratic price functions specified by $P(X) = p_0 \cdot X \pm p_1 \cdot X^2$, and performed a grid search for optimal values of p_0 and p_1. The optimal price functions for the three cases are specified in Table 1 and plotted in Fig. 9.

	Poisson (analytic computation)	***Poisson (numeric computation)***	***MMPP***	***Mixed***
P(X)	$3.39X - 4.65e-05X^2$	$3.2X - 6.6e-05X^2$	$8.6X - 3.4e-05X^2$	$4.4X + 7.6e-05X^2$

Table 1: Optimal Price Functions for the Three Cases

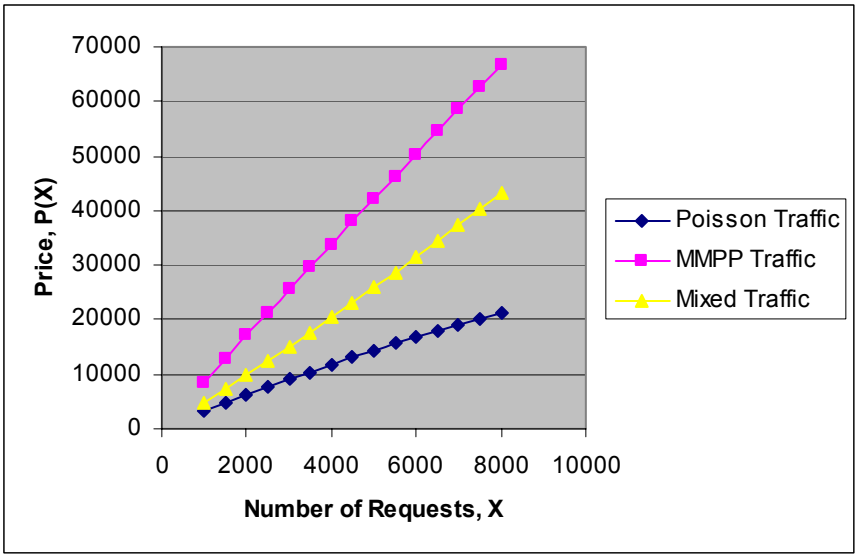

Figure 9: Optimal Price functions for the Three Cases

It can immediately be seen that the CDN is able to charge higher prices when traffic burstiness increases. That is, *Price(MMPP) > Price(Mixed) > Price(Poisson)*. This is because

the CDN's value proposition to CPs in terms of avoiding lost requests is enhanced by bursty traffic. However, the CDN will also incur a higher fixed cost of infrastructure because of bursty traffic, which does not factor in our price or profit computation. Interestingly, the extent of volume discounts provided to CPs is much lower with mixed traffic than with traffic with one fixed level of burstiness (Poisson or MMPP). In fact, the price function is convex with mixed traffic, corresponding to a volume tax rather than a volume discount.

To illustrate the reasoning behind this, consider the pricing scheme with volume discounts shown in Figure 10. CP '1' has a mean arrival rate given by $\bar{\lambda}$. CP '2' has the same arrival rate, but a higher variance. Without loss of generality, assume that '1' has a deterministic arrival process. Every period, '1' receives $\bar{\lambda}$ requests (point A in figure) and hence pays an expected price P_1 to the CDN. CP '2' on the other hand has some variability. With some high probability, '2' receives requests shown by point B and the remainder of the time, the CP receives a high number of requests shown by point C. CP '2' has the same mean $\bar{\lambda}$ as CP '1' but has higher variance. The expected price, P_2 paid by CP '2' is shown in the Figure and is clearly lower than P_1. This is an artifact of the concave price function. However, this is not desirable as the CP with higher variance derives greater surplus from the CDN and hence the CDN should ideally charge CP '2' a higher expected price. For this reason, the CDN may choose a convex price function even though the concavity in infrastructure costs exerts a force on the price function that tends to make it concave. Note also that such convexity arises only when the traffic burstiness profile is mixed and it does not arise when all CPs with the same mean arrival rate also have the same variance (pure Poisson or MMPP with same burstiness across CPs).

If the CDN chooses a convex price function, CPs with high mean arrival rates are penalized. Consider a CP with a fixed deterministic arrival rate of 2λ. Compared to a CP with fixed

arrivals of λ, the CP pays a high tax for using the CDN. In contrast, this CP gets volume discounts for infrastructure costs and may thus be tempted to deliver content on its own. Thus, a convex price function dissuades CPs with high volume and low variability traffic from subscribing to the CDN. Thus, whether the optimal price function is concave, convex, or linear in the mixed traffic case depends on the distribution of traffic burstiness across CPs and the amount of volume discounts in CP's own infrastructure costs.

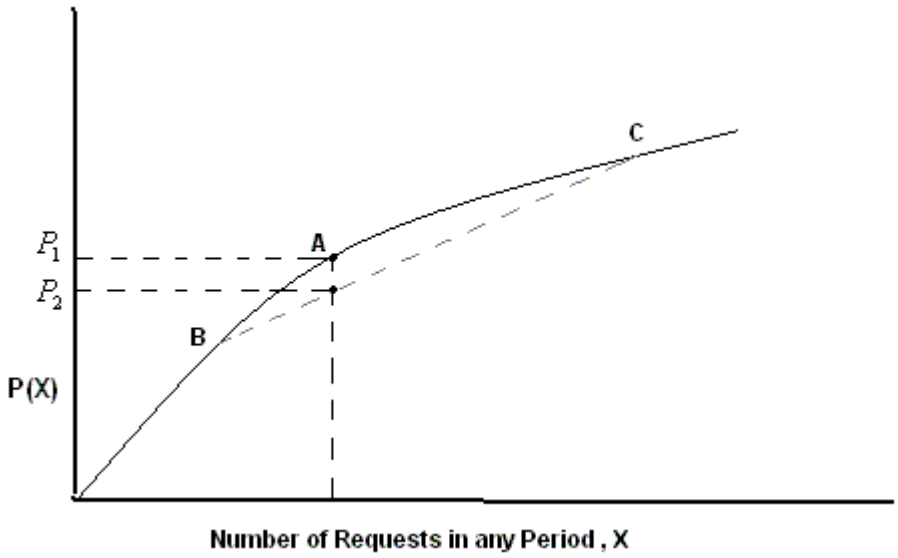

Figure 10: Expected prices for a concave price function

The analysis above indirectly suggests the inefficiency of a pure usage-based pricing policy when the traffic profile is mixed. Such a policy does not permit a CDN to provide volume discounts to CPs and simultaneously charge a higher price to CPs with greater traffic burstiness. We thus consider an alternative policy, which entails pricing based on a certain high percentile of usage. Specicially, charging a price based on the 95th percentile of usage. In such a policy, a CDN monitors the request rate, *X*, over a period of time (say a month) and computes the 95th percentile of the request rate. The price to the CP is then based on the 95th percentile rather than the observed usage rates. We computed numerically the optimal price charged at the 95th percen-

tile of usage when the traffic profile is mixed as $P(Z) = 1.6Z - 4e - 06Z^2$, where Z is the 95th percentile of request rate, X. As shown in Figure 11, when the traffic profile is mixed, the CDN's profit with a percentile-based pricing strategy is higher than with a usage-based pricing policy. At the same time, there is no noticeable difference in profit from usage-based and percentile-based pricing policies for pure Poisson and MMPP traffic. This is not surprising because once the mean request rate is fixed, the variance is also determined in both these cases[5] and hence a mean based pricing policy can be converted to a percentile based policy or vice versa. With mixed traffic, a usage based pricing scheme cannot simultaneously account for both the mean and variance in the request rate.

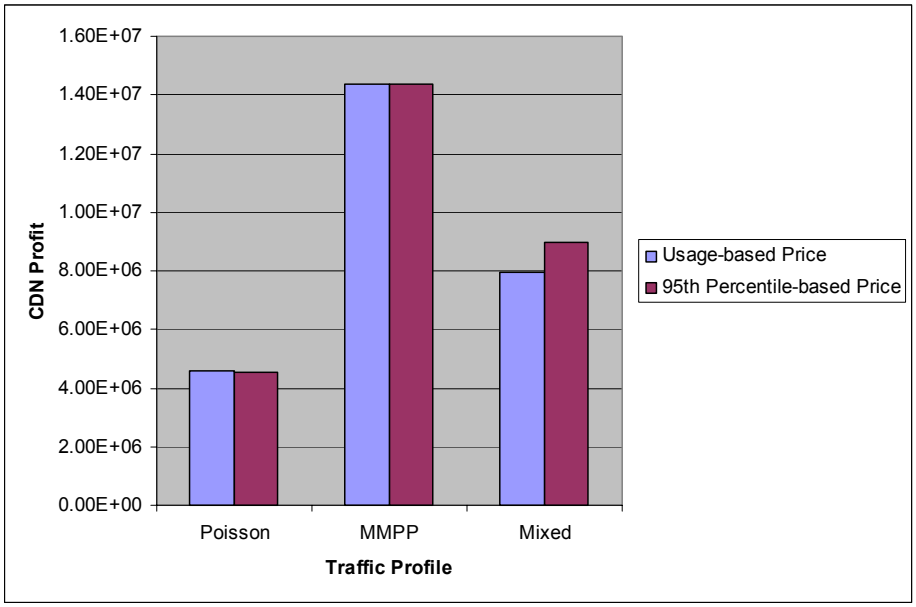

Figure 11: CDN Profit with Different Pricing Policies and Traffic Profiles

4. Conclusions

Content Delivery Networks have become an important component of the Internet content delivery value chain. These services bring content closer to consumers, and by aggregating vari-

[5] For Poisson, the variance is equal to the mean and for our chosen MMPP process, the variance is equal to the square of the burstiness (a constant) times the mean.

able traffic across a variety of sources, they minimize a content provider's risk of facing bursty traffic when using a stand-alone content delivery system. Because of the importance of timely and reliable delivery of content, nearly one-third of the most popular content sites on the Internet use CDN services.

However, despite their strategic importance for the delivery of content, there has been little academic work that has examined the pricing and adoption of these services. In particular, it is important for CDN managers and industry participants to understand the optimal pricing strategies for CDN services under different traffic patterns, the adoption drivers of CDN services, and the drivers of profitability within CDN services.

We develop analytic models to answer these questions. Our model shows that CDN pricing functions should provide volume discounts to content providers when all content providers have similar levels of traffic burstiness. It also shows that the most likely subscribers to CDN services are those content providers with high traffic volumes and low security requirements. Larger CDN networks can charge higher prices in equilibrium, which should strengthen any technology-based economies of scale. Traffic patterns play a major role in determining the infrastructure sizing decisions of content publishers as well as the optimal pricing strategies for CDNs. Surprisingly, we find that the optimal infrastructure level for highly bursty traffic is lower than for Poisson traffic. This is because small increases in infrastructure level do not suffice in handling peak traffic. Furthermore, volume discounts should be reduced, if not replaced by volume taxes, when the population consists of content providers with varying levels of traffic burstiness. Further, the pure usage based pricing policy that is used by a number of CDNs is suboptimal in such cases as well. A percentile-based pricing policy allows for volume discounts

for content providers with high mean traffic and also additional charges for content providers with highly bursty traffic, which cannot be achieved by usage-based pricing policies.

References

M. Andersson, J. Cao, M. Kihl, and C. Nyberg, "Performance Modeling of an Apache Web Server with Bursty Arrival Traffic", International Conference on Internet Computing (IC), June 2003.

A. Baiocchi and N. Blefari-Melazzi, "Steady state analysis of the MMPP/G/I/K queue," IEEE Transactions on Communications, vol. 41, no. 4, Apr. 1993.

P. Barford and M. E. Crovella. Generating representative web workloads for network and server performance evaluation. Proceedings of ACM SIGMETRICS, July 1998.

G. Banga and P. Druschel. Measuring the capacity of a Web server under realistic loads. World Wide Web Journal (Special Issue on World Wide Web Characterization and Performance Evaluation), 2(1), May 1999.

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.

J. Cao, M. Andersson, C. Nyberg, and M. Kihl, "Web Server Performance Modeling using an M/G/1/K*PS Queue,", International Conference on Telecommunication (ICT), Feb 2003.

Y. Chen, R. Katz, and J. Kubiatowicz. Dynamic Replica Placement for Scalable Content Delivery. In Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS) 2002.

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.

M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic evidence and possible causes. In Proceedings of the ACM SIGMETRICS 96, pages 160--169, Philadelphia, PA, May 1996.

K. Currier. Comparative Statics Analysis in Economics. World Scientific Publishing Company. August 2000.

A. Datta, Kaushik Datta, Helen Thomas, Debra VanderMeer. WORLD WIDE WAIT: A Study of Internet Scalability and Cache-Based Approaches to Alleviate it. Georgia Institute of Technology Working Paper.

J. Dilley, Bruce Maggs, Jay Parikh, Harald Prokop, Ramesh Sitaraman, and Bill Weihl. Globally Distributed Content Delivery. IEEE Internet Computing. Sep-Oct 2002.

S. Gadde, Jeff Chase, and Michael Rabinovich. Web caching and content distribution: A view from the interior. In Proceedings of the Fifth International Web Caching and Content Delivery Workshop, Lisbon, Portugal, May 2000.

M. Graff. Sun Microsystems Security Bulletin, October 1996, http://www.networkcomputing.com/unixworld/security/004/004.add.html.

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.

K. Hosanagar, R. Krishnan, J. Chuang, and V. Choudhary. Pricing Vertically Differentiated Web Caching Services.Proceedings of the International Conference on Information Systems (ICIS), Barcelona, December 2002.

K. Johnson, John Carr, Mark Day, and M. Frans Kaashoek. The measured performance of content distribution networks. In Proceedings of the Fifth International Web Caching and Content Delivery Workshop, Lisbon, Portugal, May 2000.

J. Kangasharju, Keith W. Ross, and James W. Roberts. Performance evaluation of redirection schemes in content distribution networks. In Proceedings of the Fifth International Web Caching and Content Delivery Workshop, Lisbon, Portugal, May 2000.

B. Krishnamurthy, C. Wills, and Y. Zhang. On the use and performance of content distribution networks. ACM SIGCOMM Internet Measurement Workshop, 2001.

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

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

National Laboratory for Applied Network Research (NLANR). Ircache project. www.ircache.net/, 2002.

M. Nottingham. On defining a role for demand-driven surrogate origin servers. In Proceedings of the Fifth International Web Caching and Content Delivery Workshop, Lisbon, Portugal, May 2000.

S. Saroiu, Krishna P. Gummadi, Richard J. Dunn, Steven D. Gribble, and Henry M. Levy. An Analysis of Internet Content Delivery Systems. Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI), Boston, MA, December 2002.

S. L. Scott, P. Smyth, “The Markov Modulated Poisson Process and Markov Poisson Cascade with Applications to Web Traffic Modelling”, Bayesian Statistics, Oxford University Press, 2003.

R. Stevens. UNIX Network Programming, Prentice Hall, 1990.