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 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, as expected in reality, volume discounts may be suboptimal and may even be replaced by volume taxes. Finally, a pricing policy that accounts for both the mean and variance in traffic such as percentile-based pricing is more profitable than pure volume-based pricing when there is heterogeneity in burstiness across content providers. This finding is in contrast to the current practices of many CDN firms that use pure volume-based pricing.

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 and intelligently deliver it to end 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 help Content Providers (CPs) to scale 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 are an important part 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 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 CDNs include Cable & Wireless, SyncCast and Mirror Image.

CDNs have traditionally offered services that enabled CPs to deliver part of their content (typically bandwidth-intensive content) through CDNs and the remainder on their own. In recent years, CDNs

---

1 Flash crowds refer to sudden surges in demand for content that often bring down web servers.
have introduced services that enable CPs to deliver entire websites from the edge servers, with Akamai’s Edgesuite being the best-known example. However, CDN executives (e.g., Maggs 2002) face challenges in determining how they should price these services, what factors influence service adoption by content providers, and how traffic patterns impact service adoption and pricing. Because optimal prices may be nonlinear in traffic serviced and CPs may be heterogeneous in mean demand, cost of outsourcing content delivery, and demand burstiness, the pricing problem is non-trivial. We discuss the unique aspects of the CDN pricing problem in the following section.

We use analytic models and numerical simulations to study this pricing problem, developing two central results. First, we characterize the optimal pricing policy when CP demand is Poisson distributed (i.e., traffic is Markovian), the CP has the option of self-provisioning, and there are some outsourcing costs to using a CDN (for example, sharing confidential data with a third party). We find that optimal prices for these services should provide volume discounts. In addition, the optimal pricing policy places a lower emphasis on value-based pricing and a greater emphasis on cost-based pricing as the relative density of CPs with high outsourcing costs increases. Second, we numerically determine how the pricing policy changes when CP demand is bursty. We find that volume discounts may continue to be offered when the CPs have similar levels of traffic burstiness, but that volume discounts may be suboptimal (and may even be replaced by volume taxes) if the CPs exhibit varying degrees of traffic burstiness. 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 when there is heterogeneity in burstiness across CPs. This finding is in contrast with current practices of many CDN firms that use pure volume-based pricing.

In addition, our paper makes an important methodological contribution to the literature. Since customers (CPs) are heterogeneous with respect to both the value they get from the CDN service, and from fulfilling their own content delivery needs, the pricing problem faced by the CDN is non-trivial. This is further complicated by the search for nonlinear prices and the uncertainty in volume of traffic experienced by CPs. To address this complexity, we first apply a dynamic optimization technique from the
calculus of variations to determine the optimal expected price charged to CPs with a given arrival rate. Based on that, we then determine an optimal usage-based price function. We discuss our approach in greater detail in Section 3.

The remainder of the paper is organized as follows. In Section 2, we review the literature on pricing information systems with congestion effects and pricing of information goods and services. We present our model in Section 3. In Section 3.1.1, we characterize the CP’s infrastructure decision and realized surplus when traffic is Poisson distributed and the CP is self-provisioning. Based on the CP’s surplus from self-provisioning, we present an analytical model to determine the CDN’s optimal price in Section 3.1.2. We present comparative statics for the pricing model in Section 3.1.3. In Section 3.2, we use numerical simulations to test the robustness of our key results under bursty traffic. We discuss our results in Section 4.

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. Other work has also analyzed CDN workloads (Saroiu et al. 2002), the effectiveness of CDNs in the presence of conventional web proxy caching (Gadde et al. 2000), and dynamic placement of replicas in a CDN (Chen et al. 2002). CDNs can also be constructed in a peer-to-peer fashion, for example CoralCDN (Freedman et al. 2004), and organized in the form of a market economy (Adler et al. 2004). 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. Management Science research can make significant contributions in this regard, as highlighted by Datta et al. (2003).

Pricing has been actively studied in the economics, marketing and operations management literatures. The two streams of work most closely related to our paper are the work on (a) Pricing Information Systems (IS) with congestion effects and (b) Pricing information goods and services.
With respect to pricing IS with congestion effects, Mendelson (1985) presents a model for pricing computer services in the presence of delay costs. The model has been extended to study incentive compatible pricing of priority computer services under heterogeneous job classes (Mendelson and Whang 1990) and pricing under nonlinear delay costs (Dewan and Mendelson 1990). Other relevant work includes the study of externalities in a single server queuing discipline (Haviv and Ritov 1998), optimal IS pricing under network externalities (Westland 1992), and pricing of congestible resources on the Internet (Gibbens and Kelley 1999; Mackie-Mason and Varian 1994). Gupta et al. (1997), Cocchi et al. (1993), and Zhang et al. (2004) have studied QoS pricing in the transmission domain (prioritized transmission of data packets based on QoS schemes such as DiffServ and IntServ). Keon and Anandalingam (2003) study pricing of multiple service classes in telecom networks in the presence of QoS guarantees and subsequently extend this work (Keon and Anandalingam 2005) to study the use of price discounts as a congestion avoidance scheme.

The literature has also studied pricing under bursty traffic. Prices under bursty traffic are often based on effective bandwidths (Kelly 1997). These are not predictable to subscribers as the computation of effective bandwidths is dependent on the characteristics of the multiplexed traffic and the link resources, which are beyond the control or knowledge of individual subscribers. Kelly (1997) proposes an approximation that includes a charge per unit time, a charge per unit volume of traffic carried and a charge per connection for real-time bursty connections in settings such as ATM networks. Marbach (2001) shows that it is not possible to provide absolute QoS guarantees in DiffServ networks with bursty traffic.

Our study, while related to this stream on IS pricing with congestion, differs from this literature in three important ways. First, the aforementioned papers focus on the interaction between pricing and QoS by studying the trade-off between congestion cost and capacity cost. When QoS requirements are stringent, pricing plays a key role in achieving desired QoS. In contrast, QoS requirements are not as stringent in the CDN setting and pricing is not necessarily driven by the goal of reducing congestion. Second, the
subscriber in network transmission both originates the traffic and pays for the service. Thus, the subscriber can shape its traffic in response to price signals. The network QoS pricing literature typically assumes that the service provider and the subscriber set up some traffic profile, and the service provider uses the information to make reservations and/or scheduling decisions. Payments can then be computed based on this traffic profile. In contrast, in the CDN setting it does not make sense for the CP to perform traffic shaping on end user requests as CPs do not want to drop or delay client requests because some traffic profile limit has been reached. For instance, flash crowds are not predictable in advance, and CPs subscribe to CDN services precisely to manage the traffic spikes. Since QoS implementation and traffic shaping are not the goals of pricing, the CDN’s objective function and subscriber response functions are fundamentally different in CDN pricing than in network transmission. Further, pricing schemes used by CDNs are often simpler than in network transmission. Prices could be based on actual usage in every period or percentile usage over several periods. Current CDN practice is to simply measure actual traffic levels, and apply pricing according to the measurements \textit{a posteriori}. Third, unlike in network transmission, CDN subscribers have the option of self-provisioning (i.e. delivering content from their own servers) and the benefit from this outside option is heterogeneous across CPs. Further, CPs incur a cost of outsourcing content delivery to a CDN which is also heterogeneous across CPs. This includes the cost of sharing confidential information and the cost of modifying content to facilitate delivery by a CDN. These "make-versus-buy" considerations are irrelevant to network pricing, but strongly influence CDN pricing.

With respect to pricing information goods and services, Varian (1997), and Shapiro and Varian (1998) study the implications of the unique characteristics of information goods (low or zero marginal costs, high fixed costs typically sunk before market) and discuss ways sellers can avoid ruinous price competition. These strategies include limit pricing, versioning, and price discrimination. Bakos and Brynjolfsson (1999) and Bakos et al. (1999) extend this work to discuss optimal pricing of bundled information goods, showing that sellers can extract nearly the full consumer surplus in large bundles of information goods. Sundararajan (2004a) studies optimal pricing of information goods when both fixed fee and

Our paper extends this stream of the literature by studying service adoption and pricing of CDNs. The analytical model employed here has a unique setup that arises from the specific features of the problem context. As previously noted, CPs are heterogeneous in terms of their mean traffic rates and outsourcing costs. Because the outsourcing cost is not observable to the CDN, the CDN has to set prices that induce CPs to self-select into two categories: self-provisioning content delivery or outsourcing to a CDN. On the other hand, the mean traffic rate is observable to the CDN, and CPs with different mean traffic rates have different responses to the price. Because of this, the CDN is faced with the challenge of announcing a single price function for all CPs while accounting for both unobservable outsourcing costs and observable mean traffic rates. We discuss this setup in greater detail in Section 3.1.

Given the importance of CDN services to the Internet economy, and relative lack of knowledge about optimal pricing and adoption drivers for these services, our paper helps answer key managerial questions relevant to both the providers and consumers of CDN services.

3. Model

We consider a monopoly CDN that provides content delivery services to a large number of CPs. CPs have the option of delivering content on their own (from their own servers or those of a hosting service) or outsourcing content delivery to a CDN. CPs are heterogeneous in terms of mean traffic rates and outsourcing costs and indexed by the corresponding type parameters: $\lambda$ and $C_o$. The mean arrival rate, $\lambda$,
is a measure of the volume of traffic handled by the CP. Following prior empirical studies (see Hosanagar et al. 2005), we assume that the number of CPs with arrival rate $\lambda$ is given by $g(\lambda) = \frac{\beta}{\lambda^\delta}$, where $\delta \in [1,2]$ and $\beta$ are constants. $C_o$ is the cost of outsourcing content delivery, such as the cost of modifying content to facilitate delivery by the CDN, or the cost of sharing confidential data with a third party (i.e., the CDN). We allow $C_o$ to vary across CPs to capture the fact that a CP with highly confidential content has higher outsourcing costs than a CP with more generic content. We assume that the cdf and pdf of $C_o$ are given by $H(C_o) = C_o^W$; $h(C_o) = WC_o^{W-1}$, where $W$ is a positive constant and $C_o \in [0,1]$. This distributional assumption is for convenience, but the parameter $W$ allows us to capture skews in the distribution. When $W=1$, $C_o$ is Uniformly distributed. $W>1$ captures negative skews and $W<1$ captures positive skews in the distribution. Note also that the upper bound of $C_o$ can be arbitrarily high but is normalized to 1 without loss of generality. We assume that $\lambda$ and $C_o$ are independent as we do not expect a significant correlation between demand for content and confidentiality requirements. We also assume that the CDN knows $\lambda$ for all CPs as the traffic can be directly observed by the CDN, but does not know the outsourcing cost $C_o$. However, we assume the CDN does know the distribution of outsourcing costs across CPs ($H(C_o)$).

We now discuss the CP’s surplus under self-provisioning and provisioning through a CDN.

**Self-Provisioning by Content provider**

Consider a CP of type $\lambda_i$ delivering content to users. Let $X_i$ be a random variable denoting the realized number of requests to CP $i$ in a given period. In any period, the distribution of $X$ is known *a priori*, but not its realized value. The CP can choose to serve this content directly by investing in infrastructure to process a mean $I$ requests per unit time. If it does so, its surplus from serving content is

$$U_{self}(X) = V(X) - C(I) - c \cdot L(I, X)$$  \hspace{1cm} (1)
where $V()$ is the CP’s benefit from responding to all $X$ requests, $C()$ is the cost of maintaining the infrastructure (servers, bandwidth, software, etc), which is concave in $I$ because of economies of scale. $L()$ is the number of lost requests and $c$ is the cost of each lost request.

The value of serving content, $V()$, includes all sources of revenue from the CP’s Internet operations (e.g., selling products on the Internet, indirect surplus from disseminating information). We model the CP’s infrastructure cost as: $C(I) = a_1 \cdot I - a_2 \cdot I^2$, $(I \leq a_1 / 2a_2 )$, which captures the concavity between $I$ and cost. The constraint $I \leq a_1 / 2a_2$ ensures that the infrastructure cost is always non-decreasing in infrastructure (note that $C'(I) < 0$ for $I > a_1 / 2a_2$). In this formulation, a large value for $a_1$ indicates high infrastructure costs and a large value for $a_2$ indicates significant economies of scale. $L()$, the number of lost requests, increases with $X$ but decreases with $I$.

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 choice of infrastructure level $I$ will ultimately determine the CP’s net surplus from delivering content on its own.

**Provisioning through a CDN**

The CP can also choose to deliver content through a CDN. The CP’s surplus from delivering content through the CDN is given by

$$U_{CDN}(X) = V(X) + \tau(N) \cdot X - C_o - P(X) \tag{2}$$

where $V()$, and $X$ are defined as above, $\tau()$ is the benefit per request from faster content delivery through a geographically distributed set of $N$ CDN servers, $C_o$ is cost of outsourcing content delivery, and $P()$ is the usage-based price the CDN charges the CP.

Empirical studies show that CDNs reduce the average latency in content delivery (Krishnamurthy et al. 2001) and that the latency benefit is primarily driven by the proximity of the CDN server to the end user (Huang and Abdelzaher 2004). Generally, the larger the number of replicas, the closer the replicas
are to the end users. In fact, Cronin et al. (2002) demonstrate that the benefit as measured by improvement in mean latency increases with the number of replicas in a concave fashion. In detailed simulations of a distributed CDN, we also find that the improvement in average latency increases with number of CDN servers in a concave fashion. The simulations were performed using CDNSim (Pallis et al. 2005), a trace-driven discrete event CDN simulator that simulates a network with clients, CDN servers and origin servers interconnected through several intermediate routers. CDNSim simulates TCP/IP request-response process and various object placement/replacement policies. The simulations are described in detail in the online Appendix. Based on the empirical and simulation results, we assume that \( \tau \) is increasing in \( N \).

The outsourcing costs incurred by CPs may be in the form of content modification costs or the cost of sharing confidential data with the CDN. The former is the cost associated with modifying content in order to facilitate delivery by the CDN. This cost is driven primarily by the technology choice of the CDN and is likely to be similar across CPs. The cost of sharing confidential data arises because the CP may be sharing sensitive information such as customer records, credit card information or patient medical history with the CDN. The cost may be in the form of perceived risk or may be due to additional steps needed to ensure security. The cost of sharing confidential data is expected to vary across CPs because of inherent differences in the type of content handled by CPs. One way to visualize the outsourcing cost is that the two costs are additive, i.e., \( C_o = (\text{data confidentiality cost}) + (\text{content modification cost}) \). In this case, the first term determines the shape of the distribution of \( C_o \), whereas the second term only impacts the boundary of the distribution not the shape. An alternative formulation is that the two costs have a multiplicative impact on the outsourcing cost, i.e., \( C_o = (\text{data confidentiality cost}) \times (\text{content modification cost}) \). Even here, the content modification cost impacts only the boundary of the distribution of \( C_o \) but not the shape of the distribution. We discuss the managerial implications of this interpretation of outsourcing costs in the context of the results of the analytical model in Section 3.1.

We also assume that the CDN serves the CP’s entire site, as is the case with Akamai’s EdgeSuite service and that the CDN maintains sufficient capacity to nearly eliminate lost requests. Thus, the cost of
the minimal infrastructure needed and the cost of lost requests, $C()$ and $L()$ respectively, are both approximated to zero in the case of CDN delivery. Finally, we consider the pricing problem of a CDN with a network in place and treat the CDN’s network size, $N$, as an exogenous parameter. This is consistent with the observation that CDNs like Akamai have already rolled out significant infrastructure but have only recently rolled out Edgesuite-like services for edge assembly and delivery of entire websites. Thus, the relevant problem is the optimal pricing of these services given existing network size.

We apply our model by first analyzing the CP’s surplus when provisioning content directly. In order to determine this surplus, we need to solve the CP’s optimal infrastructure sizing problem. Based on the optimal infrastructure level, $I^*$, we can then use equation (1) to compute the CP’s surplus from self-provisioning. We then analyze the CP’s surplus when outsourcing to a CDN and determine the drivers of a CP’s subscription decision. Based on that, we then determine the CDN’s optimal pricing decision. We use this framework with two different traffic distributions. In Section 3.1, we analytically study the case in which CPs’ traffic is Poisson distributed. We then extend the results numerically in Section 3.2 to the case when traffic is bursty.

3.1. Pricing with Poisson Demand

In this section, we assume that the demand for content at any given CP site is Poisson distributed. We begin by characterizing CPs’ surplus under self provisioning. To do this, we first analyze the optimal infrastructure sizing decision in Section 3.1.1.

3.1.1. Optimal Infrastructure Sizing for Content Providers

We now analyze the CP’s optimal infrastructure sizing decision and compute the associated surplus. The expected surplus from delivering content is obtained from equation (1) as follows:

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

(3)
where \( L(I) = E[L(I, X)] \) and \( V = E[V(X)] \). We assume that the CPs and CDN are risk neutral. Risk neutrality implies that the CPs (CDN) care only about the expected surplus (profit) and not about the variance. The CP’s decision problem, given risk neutrality, is \( \max \{ U_{self}(I) \} \). We denote the optimal infrastructure level for a CP of type \( \lambda \) as \( I^*(\lambda) \) and the associated expected surplus as \( U_{self}(I^*(\lambda)) \).

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 \( \lambda \). In Section 3.2, we relax this assumption to consider a bursty, as opposed to Poisson, arrival process for requests. The service time distribution is arbitrary. The queuing model treats the delivery system as a single server.\(^3\) The queue length is a finite exogenous parameter \( K \), which is consistent with the observation that most commercial servers have similar \texttt{somaxconn} settings and most vendors recommend setting the queue size to \texttt{somaxconn} (see Stevens (1990) and Banga and Druschel (1997) for more on HTTP connection establishment). Multithreading in the server is modeled by a processor sharing queuing discipline.

For an \( M/G/1/K*PS \) queuing system, the expected number of lost requests is given by

\[
L(I) = \frac{\lambda \left( 1 - \frac{\lambda}{I} \right)^K}{1 - \frac{\lambda}{I}}. 
\]

The region of interest for our model is \( I \in (\lambda, \frac{a_1}{2a_2}] \).\(^4\) 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 is straightforward to verify that \( \frac{\partial L(I)}{\partial I} < 0 \), and thus the CP has to trade off the benefits and costs of added infrastructure. Under this model, the CP’s decision problem is

\(^3\) In the online appendix, we also consider a multiple server system.

\(^4\) The lower bound on \( I \) ensures that requests do not overwhelm the infrastructure. The upper bound is driven by the fact that our infrastructure cost function is defined for \( I \leq (a_1 / 2a_2) \).
\[
\max_I U_{\text{Self}}(I) = \max \left\{ V - (a \cdot I - b \cdot I^2) - c \cdot \left( 1 - \frac{\lambda}{I} \right) \left( \frac{\lambda}{I} \right)^K \right\}
\]

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

\[
-a + 2b \cdot I - \frac{c \cdot \lambda^{K+1}}{I^{K+1} - \lambda^{K+1}} + \frac{c \cdot (K + 1) \cdot \lambda^{K+1} (I - \lambda) I^K}{(I^{K+1} - \lambda^{K+1})^2} = 0.
\]

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

**Lemma 1:** If the cost of infrastructure increases, \( I^*(\lambda) \) decreases.

**Lemma 2:** If there are significant economies of scale in content delivery, \( I^*(\lambda) \) increases.

**Lemma 3:** If a content provider’s cost of losing requests is high, \( I^*(\lambda) \) is correspondingly higher.

**Lemma 4:** If the arrival rate of requests \( \lambda \) increases, \( I^*(\lambda) \) increases.

These lemmas conform to intuition. One result of particular relevance to our CDN pricing problem is how the CP’s optimal infrastructure, and thus surplus, varies with CP type parameter \( \lambda \). This relationship determines the surplus under self-provisioning for a CP of type \( \lambda \), which in turn influences the CP’s decision to subscribe to a CDN. Since there is no closed form solution to the analytical model, we employ numerical tests using parameter values conforming to typical bandwidth and hosting costs to determine an approximate relationship between \( \lambda \) and \( I^*(\lambda) \). The results in Appendix A for reasonable parameters suggest that the optimal infrastructure is approximately linear in the mean arrival rate. We find that the approximately linear relationship holds for a wide range of parameters.\(^5\) In subsequent analysis,

\(^5\) In the online appendix, we also show that this approximate relation holds for multi-server systems as well.
we assume that a CP’s capacity choice is given by \( I^*(\lambda) = M \cdot \lambda \), where \( M \) is a constant. With this assumption, the CP’s expected surplus from self-provisioning is

\[
U_{self}(I^*(\lambda)) = V - C(M\lambda) - c \cdot L(M\lambda)
\]  

(6)

We now analyze the CP’s surplus from using the CDN.

3.1.2 CDN Pricing Problem

Consider a CP that outsources content delivery to a CDN. The CP does not know how many requests \((X)\) will be made for its content in any period, but can compute the expected surplus from outsourcing to a CDN:

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

(7)

Given any price function \( P(X) \), the CP can compute expected surplus. The CP chooses the CDN if

\[
U_{CDN} \geq U_{self}(I^*(\lambda))
\]

Substituting (6) and (7) into this condition, a CP of type \{ \lambda, C_o \} subscribes to the CDN if

\[
C_o \leq \tau(N) \cdot \lambda + C(M\lambda) + c \cdot L(M\lambda) - E[P(X)]
\]  

(8)

Recall that \( C_o \in [0,1] \). If the right hand side of equation (8) is greater than 1, it implies that all CPs subscribe to the CDN. We assume here that the maximum outsourcing cost is high enough that there is at least one CP that does not subscribe to the CDN even when the price is zero. This is quite reasonable and is expected to hold for CPs with low mean traffic volume (\( \lambda \)). Similarly, if the right hand side of (8) is less than zero, then there are no subscribers of the CDN. This places an upper bound on the price that can be charged. We will revisit this bound later in this section, and focus here on the interior solution of (8).

With \( H(C_o) = C_o^W \), the probability that a CP with mean arrival rate \( \lambda \) subscribes to a CDN is

\[
\text{Prob}(\text{Subscription} \mid \lambda) = \left( \tau(N) \cdot \lambda + C(M\lambda) + c \cdot L(M\lambda) - E[P(X)] \right)^W
\]  

(9)
If $g(\lambda)$ denotes the number of CPs with mean arrival rate $\lambda$, the expected number of these CPs subscribing to the CDN is given by

$$Subs(\lambda) = g(\lambda)(\tau(N) \cdot \lambda + C(M\lambda) + c \cdot L(M\lambda) - E[P(X)])$$  \hspace{1cm} (10)

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 $P(X)$. The CDN’s expected profit is given by

$$\pi = \left\{ \int_{\lambda} Subs(\lambda) \left( \int_X \Pr ob(X | \lambda) \cdot P(X) dX \right) d\lambda \right\} - \left\{ b_1 \left( \int_{\lambda} \lambda \cdot Subs(\lambda) d\lambda \right) - b_2 \left( \int_{\lambda} \lambda \cdot Subs(\lambda) d\lambda \right)^2 \right\}$$ \hspace{1cm} (11)

In the expression above, the first term represents the CDN’s expected revenues obtained by summing the revenues over $\lambda$. The second term represents the CDN’s cost, which in the long term is assumed to be quadratic over the mean volume of traffic handled by the CDN. This cost includes the cost of keeping content consistent across replicas, an accounting mechanism that collects and tracks information related to request routing and delivery (Vakali and Pallis 2003), and the cost associated with content delivery. All these activities are expected to involve economies of scale and thus the costs are expected to be concave in volume. Note that the CP and CDN cost parameters are different (i.e., $a_i \neq b_1, a_2 \neq b_2$) because the CDN cost includes other factors, such as accounting cost and cost of maintaining consistency, in addition to the content delivery cost. Substituting equation (10) and $g(\lambda) = \beta / \lambda^\beta$ into (11) gives:

$$\pi = \int_{\lambda} \frac{\beta}{\lambda^\beta} \left( \tau(N) \cdot \lambda + C(I^*(\lambda)) + c \cdot L(I^*(\lambda)) - E[P(X)] \right)^\gamma \left( E[P(X)] \right) d\lambda$$

$$- \left\{ b_1 \left( \int_{\lambda} \frac{\beta}{\lambda^\beta} \left( \tau(N) \cdot \lambda + C(I^*(\lambda)) + c \cdot L(I^*(\lambda)) - E[P(X)] \right)^\gamma d\lambda \right) - b_2 \left( \int_{\lambda} \frac{\beta}{\lambda^\beta} \left( \tau(N) \cdot \lambda + C(I^*(\lambda)) + c \cdot L(I^*(\lambda)) - E[P(X)] \right)^\gamma d\lambda \right)^2 \right\}$$ \hspace{1cm} (12)
Note that CPs make their subscription decision based on \( E[P(X)] \) (see equation 8) and the CDN computes its expected profit by evaluating \( E[P(X)] \) for each subscribing CP (equation 12). Thus, for risk-neutral agents, the selected usage-based price function is relevant to the CPs and the CDN only through its expectation, i.e., the \( E[P(X)] \) function. Denote \( E[P(X)] \) by \( P(\lambda) \). \( P(\lambda) \) is a function that specifies the value of \( E[P(X)] \) for different values of \( \lambda \). With this notation, we rewrite (12) as follows:

\[
\pi = \int_{\lambda} \frac{\beta}{\lambda^\delta} \left( r(N) \cdot \lambda + C(I^*(\lambda)) + c \cdot L(I^*(\lambda)) - P(\lambda) \right) P(\lambda) d\lambda - \\
\left\{ b_1 \left( \int_{\lambda} \frac{\beta}{\lambda^\delta} \left( r(N) \cdot \lambda + C(I^*(\lambda)) + c \cdot L(I^*(\lambda)) - P(\lambda) \right) P(\lambda) d\lambda \right\} - \\
\left\{ b_2 \left( \int_{\lambda} \frac{\beta}{\lambda^\delta} \left( r(N) \cdot \lambda + C(I^*(\lambda)) + c \cdot L(I^*(\lambda)) - P(\lambda) \right) P(\lambda) d\lambda \right)^2 \right\}
\]

(13)

It is worth noting the setup used to obtain equation 13. First, the consumer type parameter \( C_o \) is unobservable to the CDN. Thus, given a price function \( P(X) \), consumers self-select between the option of self-provisioning or outsourcing to the CDN (equation 8). The subscription decision is similar to that in pricing models with unobservable consumer types. This decision results in a demand curve for each \( \lambda \) that specifies the number of CPs of type \( \lambda \) that subscribe to the CDN when the price charged is \( P(X) \). Now, if each \( \lambda \) were charged a unique expected price \( P(\lambda) \), the problem of determining the optimal \( P(\lambda) \) is more like a third-degree price discrimination problem once the demand function for each \( \lambda \) has been determined. However, the CDN announces a single price function, \( P(X) \), to all CPs. We can determine the optimal usage-based price function, \( P^*(X) \), by first solving for the \( P^*(\lambda) \) that maximizes equation (13) and then determining a \( P^*(X) \) such that \( E[P^*(X)] = P^*(\lambda) \). Determining the optimal \( P^*(\lambda) \) from equation (13) is itself complicated by the fact that the objective function is a function of a definite integral involving the price path \( P(\lambda) \). We apply the binomial theorem and a dynamic optimization technique from the calculus of variations, to derive the following optimal \( P^*(\lambda) \) (proof in appendix B):

**Proposition 1:** The expectation of the optimal usage-based price function is:
\( P^* (\lambda) = \frac{1}{1+W} \left[ \tau(N) + a_1 M + \frac{c(M-1)}{M^{K+1} - 1} + W \cdot (b_1 - 2b_2 \cdot \psi) \right] \lambda = \frac{a_2 \cdot M^2}{1+W} \lambda^2 \) (14)

where \( \psi \) is the total volume of traffic handled by the CDN at the optimal price. \( \psi \) can be solved self-consistently from the following equation:

\[
\psi = \int_0^{\lambda_M^*} \frac{1}{\lambda^\theta} \left( \frac{\tau(N) + a_1 M + \frac{c(M-1)}{M^{K+1} - 1} - a_2 \cdot M^2 \lambda^2 - P^* (\lambda)}{\beta \lambda} \right) \lambda^\theta d\lambda \] (15)

To determine the optimal price function, we need to find a \( P^* (X) \) such that \( E[P^*(X)] = P^* (\lambda) \).

**Corollary:** A usage-based price function, \( P(X) \) for which the \( E[P(X)] \) trajectory is given by equation (14), and hence an optimal usage-based price function, is:

\[
P^*(X) = \frac{1}{1+W} \left[ \tau(N) + a_1 M + a_2 \cdot M^2 + \frac{c(M-1)}{M^{K+1} - 1} + W \cdot (b_1 - 2b_2 \cdot \psi) \right] X + \frac{a_2 \cdot M^2}{1+W} X^2 \] (16)

It can be easily shown that \( E[P^*(X)] = P^* (\lambda) \) when \( X \) follows a Poisson distribution. We verify that equation (16) represents an optimal usage-based price function by contradiction: assume that there is a different price function, \( P^i(X) \) that performs better than \( P^* (X) \), i.e., yields a higher expected profit than \( P^* (X) \). But notice that any usage-based price function impacts the CDN’s expected profit only through its expectation \( E[P(X)] \) (see equation 12). In other words, our assumption implies that the expectation of the alternative price function, denoted by \( E[P^i(X)] \), should yield higher expected profit than \( E[P^*(X)] \) (i.e., \( P^* (\lambda) \)). However, this cannot be true since equation (14) is optimal for the problem of maximizing the expected profit in equation 13. Thus, equation (16) represents an optimal usage-based price function.

The optimal price function can be rewritten as

\[
P(X) = \frac{1}{1+W} \left[ \tau(N) + a_1 M + a_2 \cdot M^2 + \frac{c(M-1)}{M^{K+1} - 1} - a_2 \cdot M^2 X \right] X + \frac{W}{1+W} \cdot (b_1 - 2b_2 \cdot \psi) X \] (17)
where the first term reflects some of the value to the CP from the CDN and the second term reflects some of the CDN cost (in fact, $b_1 - 2b_2 \cdot \psi$ is the CDN’s marginal cost at the optimal price). Hence as expected, the optimal price is a combination of value based pricing and cost based pricing.

Recall that the right hand side of equation (8) must be greater than zero for some $\lambda$, else there are no subscribers to the CDN service. Substituting (16) into (8), we get

$$(b_1 - 2b_2 \cdot \psi) \leq (\tau(N) \cdot \lambda + C(M\lambda) + c \cdot L(M\lambda)).$$

Thus, the CDN’s marginal cost must be sufficiently low relative to the value the CPs derive from the CDN in order for the interior solution in this section to be valid.

### 3.1.3. Comparative Statics and Discussion

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

- **a)** When the CDN cost function does not have any economies of scale (i.e., $b_2 = 0$), the price charged by the CDN 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 $\lambda$, and thus does not care about the distribution of CP mean arrival rates. However, when volume discounts exist ($b_2 \neq 0$), the CDN is interested in how many CPs have a given mean arrival rate $\lambda_1$. Thus, the CDN may be willing to discount the price for CPs with a certain arrival rate $\lambda_1$ in return for cost savings, when there are relatively large number of CPs of type $\lambda_1$.

- **b)** *Volume discounts*: It is straightforward to see that the optimal price is quadratic in the traffic level, $X$, and that $\frac{\partial P(X)}{\partial X} > 0$ and $\frac{\partial^2 P(X)}{\partial X^2} < 0$ (proof in appendix B). Thus,

$\textit{Proposition 2:}$ The optimal pricing policy when content providers have Markovian traffic 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 (16) indicates that the volume discounts follow from the economies of scale in the CPs’ content delivery costs \((a_2>0)\). In other words, if bandwidth sellers reduce their volume discounts, so should the CDN. Interestingly, the volume discount is not driven by the concavity in the CDN’s own costs (i.e., the volume discount is independent of \(b_2\)).

c) Impact of Network Size: Taking the derivative of (17) with respect to \(N\):

\[
\frac{\partial P}{\partial N} = \frac{X}{1 + W} \left( \frac{\partial \tau(N)}{N} + W \frac{\partial (b_1 - 2b_2 \psi)}{\partial N} \right) = \frac{X}{1 + W} \left( \frac{\partial \tau(N)}{\partial N} - 2b_2 W \frac{\partial \psi}{\partial N} \right)
\]  

(18)

The direct impact of a larger CDN network is that replicas are likely to be closer to eventual users thus driving up CP surplus (i.e., \(\frac{\partial \tau(N)}{\partial N}\) is positive). This results in a direct increase in the price (the first term in equation 18). However, the CDN does not extract all the increase in CP surplus. Hence, CPs are also better off and a number of CPs who previously preferred self-provisioning will now subscribe to the CDN. This increases mean traffic volume at the CDN and reduces the CDN’s marginal cost of servicing the traffic. This decrease in the CDN’s marginal cost drives down prices and acts as a counteracting force that tends to limit the extent of the price increase (term 2 in equation 18). Thus, the impact of a larger network size is composed of price-increasing component that results from greater subscriber value and a price-decreasing component that results from higher traffic (i.e., lower marginal costs).

d) Impact of Outsourcing Cost: Substituting the optimal price function into the subscription condition in equation (8), a CP with mean arrival rate \(\lambda\) subscribes to the CDN if:

\[
C_a \leq \frac{W}{1 + W} \left[ \tau(N) + a_1 M + \frac{c(M-1)}{M^{\kappa+1} - 1} - b_1 + 2b_2 \psi \right] \lambda - \frac{a_2 M^2 W}{1 + W} \lambda^2
\]  

(19)
As seen in Figure 1, CPs likely to subscribe to a CDN are those with high volume of traffic and low outsourcing costs (e.g., low data confidentiality requirements). What happens when the market has relatively more CPs with data confidentiality requirements? The impact can be determined by assuming a higher value of $W$ in the distribution of outsourcing cost, $H(C_o) = C_o^W$. For example, Figure 2 plots a Uniform distribution ($W=1$) and a new distribution with $W=2$. Relative to the uniform distribution, the new distribution assumes there are more CPs with high outsourcing cost. As $W$ increases, the first term in equation 17 decreases and the second term increases. Thus, this proposition follows:

**Proposition 3:** If the relative density of CPs with high outsourcing costs increases, the CDN’s optimal pricing strategy entails lower emphasis on value-based pricing relative to cost-based pricing.

Recall that the outsourcing costs incurred by CPs may be in the form of content modification costs or cost of sharing confidential data with the CDN. The content modification cost impacts only the boundary of the distribution of $C_o$ but not the shape of the distribution, whereas confidentiality cost influences the shape of the distribution. Hence, we can interpret the previous results regarding skews in the distribution of $C_o$ as relating to the cost of sharing confidential data. We now discuss the impact of the content modification cost.

Content modification is closely tied to the request redirection technology 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 name server of the CP and resolves the name to the IP address of a CDN server, keeping the Time-To-Live (TTL) of these DNS mappings small so that the same URL can be mapped to different servers based on network conditions. Krishnamurthy et al. (2001) find that DNS redirection adds additional overheads and thus URL rewriting is more efficient in terms of users’ realized response times.

However, URL rewriting would entail higher outsourcing costs for CPs because of the significant cost incurred in modifying their entire content. For example, 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. Even when content modification can be significantly automated, costs are incurred in dynamic content modification (Vakali and Pallis 2003). This will result in lower prices as the distribution of $C_o$ shifts to the right. Thus, the CDN will need to trade-off efficiency-based benefits of any redirection technology with the outsourcing costs imposed on the CPs.

### 3.2. Pricing with Bursty Demand

The model in Section 3.1 assumes 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 this section, we examine the impact of traffic burstiness on CDN service adoption and pricing. Because of the analytical intractability in studying bursty traffic, we use Monte Carlo simulations to determine optimal pricing below. The objective of the simulations in this section is to verify which of the results from Section 3.1 are robust under bursty demand.

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. When the Markov chain is in state $j$, arrivals follow a Poisson process with arrival rate $\lambda_j$.

In this section, we consider a 2-state MMPP. The arrival matrix, transition matrix, and limiting state probabilities of the phase process are given by

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad R = \begin{bmatrix} -\sigma_{12} & \sigma_{12} \\ \sigma_{21} & -\sigma_{21} \end{bmatrix}$$

and $q = (q_1, q_2)$ respectively, where $\lambda_j$ denotes the mean arrival rate in state $j$, $\sigma_{jk}$ is the probability of a transition from state $j$ to state $k$, and $q_j$ is the steady state probability of being in state $j$. The mean and variance of the number of requests in a unit time period are denoted $\bar{\Lambda}$, and $\Psi$ respectively. 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 $\lambda_2$ 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 our MMPP parameters in this section. In other words, the mean arrival rate during bursts is ten times the regular mean arrival rate; and the system, on average, bursts 10% of the time. Different values of $\bar{\Lambda}$ are simulated by varying $\lambda_1$. Further, when the mean arrival rate $\bar{\Lambda}$ is increased, we also adjust $\sigma_{12}$ in order to maintain constant burstiness (constant value for $\sqrt{\Psi/\bar{\Lambda}}$).

As in Section 3.1, we begin by modeling a CP’s infrastructure sizing problem in order to compute the CP’s surplus under self-provisioning and then proceed to the CDN pricing problem.

3.2.1. Optimal Infrastructure Sizing for Content Providers

The CP’s expected surplus under self-provisioning is given by equation (3). The expected number of requests lost, $L(I)$, under MMPP arrivals can be computed numerically as highlighted by Baiocchi and Blefari-Melazzi (1993). The optimal infrastructure level is obtained by trading off the infrastructure cost with the cost of lost requests. In Appendix C (Figure C1), we present numerical results on optimal infrastructure sizing. Interestingly, we find that the CP’s optimal infrastructure level with bursty traffic is
lower than that with Poisson traffic, even though MMPP traffic has higher variance than Poisson traffic for the same mean arrival rate. This is driven by the fact that most of the lost requests are due to bursts that occur when the system is in state 2. In order to see a marked reduction in lost requests, the infrastructure level has to be raised above $\lambda_2$. However, this raises the infrastructure cost substantially and is suboptimal. Thus, the high disparity in arrival rates between the two states implies that the CP is often willing to accept server downtime during bursts, and thus the optimal infrastructure level is driven by $\lambda_1$ and not $\overline{\lambda}$. Since $\lambda_1 < \overline{\lambda}$, the optimal infrastructure level is also lower than that with Poisson traffic. As expected, the CP loses a large number of requests with highly bursty traffic and the CP’s surplus, conditional on optimal infrastructure sizing, with MMPP traffic is lower than with Poisson traffic (Figure C2 in Appendix C). From the perspective of our model in Section 3.1, the primary assumption that the optimal infrastructure level is approximately linear in mean arrival rate continues to hold with MMPP traffic as well (Figure C1).

3.2.2. CDN’s Optimal Pricing Policy

When traffic burstiness is heterogeneous across CPs, there is no closed form analytic solution to the optimal price function. We simulate web traffic and numerically compute the optimal usage-based price function for 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 3.2.1. 3) Mixed traffic: 500 CPs have Poisson traffic and 500 CPs have MMPP traffic.

For each case, we first compute the CPs’ expected utility from self-provisioning. The mean arrival rates for the CPs are drawn from a Pareto distribution in [1000, 8000]. For the infrastructure cost function, $C(I) = a_1 \cdot I - a_2 \cdot I^2$, we assume that $a_1=3.56$ and $a_2=0.000043$. We assume that the cost of a lost request, $c$, is $10$ and the queue size, $K$ is 1024 requests. These values are motivated by real-world parameters and justified in detail in Appendix A. The expected number of lost requests can be computed.

---

6 If the CP’s cost is convex in the number of lost requests, then the CP may be more willing to incur higher infrastructure costs in return for lower cost of lost requests. The resulting infrastructure level can be much higher than suggested in Figure C1. Regardless, bursts will continue to negatively impact the CP’s surplus.
numerically for MMPP traffic (Baiocchi and Blefari-Melazzi (1993)) and from equation 4 for Poisson traffic. Under these settings, we compute optimal infrastructure and associated expected utility under self-provisioning for each CP from equation 3.

Next, we compute the CP’s expected utility from CDN-provisioning for a given CDN price function. The benefit per request from faster delivery is set at $\tau = 1$ and CP outsourcing cost is drawn from a $U[0,30000]$ distribution. For the CDN price function, we restrict attention to quadratic functions specified by $P(X) = p_0 \cdot X \pm p_1 \cdot X^2$, and perform a grid search for optimal values of $p_0$ and $p_1$. For each CP, we draw 1000 values of $X$ from the corresponding arrival distribution (Poisson or MMPP). Given $p_0$ and $p_1$, we can compute the price $P(X)$ corresponding to each value of $X$ and also the expected price for the CP by averaging over the 1000 values of $X$. Given these parameters, the expected utility from using the CDN is computed from equation 7. The CP subscribes to the CDN if the expected utility is higher than under self-provisioning. The CDN’s expected profit is obtained by summing expected revenues from each subscribing CP and subtracting the CDN’s cost as in equation 11 ($b_1=3$ and $b_2=0.000047$). We compute the optimal price in 50 replications of the simulation. The optimal price functions, from one simulation run, are specified in Table 1 and plotted in Fig. 5. The variance in the computed values of $p_0$ and $p_1$ across the replications was small.

<table>
<thead>
<tr>
<th></th>
<th>Poisson (analytic computation from 3.1)</th>
<th>Poisson (numeric computation)</th>
<th>MMPP</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(X)</td>
<td>4.3X - 4.65e - 05X^2</td>
<td>Error! Objects cannot be created from editing field codes.</td>
<td>8.5X - 4.9e - 05X^2</td>
<td>Error! Objects cannot be created from editing field codes.</td>
</tr>
</tbody>
</table>

**Table 1: Optimal Price Functions for the Three Cases**

It can be seen that the CDN is able to charge higher prices as traffic burstiness increases. That is, $\text{Price(MMPP)} > \text{Price(Mixed)} > \text{Price(Poisson)}$. This is because the CDN’s value proposition to CPs in terms of avoiding lost requests is enhanced in the presence of bursty traffic. Interestingly, the extent of volume discounts provided to CPs is much lower with mixed traffic than when the level of burstiness is constant across CPs (Poisson or MMPP). In fact, for the settings above, the price function is convex for mixed traffic, corresponding to a volume tax rather than a volume discount. This suggests that proposition
2, which indicates that volume discounts are always optimal with Poisson traffic, need not hold when there is heterogeneity in burstiness across CPs.

We now explore the convexity further. Consider the pricing scheme with volume discounts shown in Figure 6. CP$_1$ has a mean arrival rate given by $\lambda$. Without loss of generality, assume that CP$_1$ has a deterministic arrival process. Every period, CP$_1$ receives $\lambda$ requests (point A in figure) and pays an expected price $P_1$ to the CDN. CP$_2$ on the other hand has the same mean $\lambda$ as CP$_1$ but has higher variance. With some high probability, CP$_2$ receives requests shown by point B; but for the remainder of the time it receives a high number of requests shown by point C. 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 under self-provisioning exerts a force on the price function that tends to make it concave (see equation 16). Note also that such convexity arises only when the traffic burstiness profile is mixed and not 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 $2\lambda$. Compared to a CP with fixed arrivals of $\lambda$, the CP pays a high tax for using the CDN. In contrast, this CP gets volume discounts when self-provisioning 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.
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. The policy will either penalize CPs with low burstiness or CPs with high volume, depending on whether a concave or convex price function is used. We thus consider an alternative policy, which entails pricing based on a certain high percentile of usage. Specifically, the CDN monitors the request rate, $X$, over a period of time (e.g., a month) and computes the $95^{th}$ percentile of the request rate for each CP. The price to the CP is then based on the $95^{th}$ percentile of his/her usage rate. Let $Z$ be the $95^{th}$ percentile of request rate, $X$. As before, we restrict attention to quadratic price functions ($P(Z) = p_0 \cdot Z \pm p_1 \cdot Z^2$) to simplify computation. Using the same grid search approach used to obtain the results in Table 1, we numerically computed the optimal percentile-based price when the traffic profile is mixed. In Figure 7, we plot the CDN’s optimal expected profit under percentile-based pricing and traditional usage-based pricing. When the traffic profile is mixed, the CDN’s profit with a percentile-based pricing strategy is higher than with a traditional 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,\(^7\) 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 traffic.

There are some drawbacks of percentile-based billing, including complicated billing relative to traditional usage-based billing and the lack of standardization (e.g., choice of sampling times can affect the bill). This has resulted in some debate in the content delivery industry regarding the optimal billing policy. For example, consider this statement from a newsletter from Servicelevel.net “Peak demand billing does not happen regularly in other utility markets...That customers are paying for it in the Internet space may simply be a sign of immaturity in this sector...It just doesn’t connect to user value.” Similarly, several CDNs such as SyncCast use traditional usage-based billing because of its simplicity. However, our results suggest that when different CPs have different levels of burstiness, as expected in reality, percentile-based pricing performs much better than volume-based pricing. When burstiness levels of subscribers are somewhat similar, CDNs can choose volume-based pricing schemes to simplify billing.

\(^7\) For Poisson, the variance is equal to the mean. For our MMPP process, the variance is equal to the square of the product of burstiness (a constant) and the mean.

Figure 7: 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 variable 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.

In this paper, we first develop analytic models to determine optimal pricing when content providers’ traffic is Poisson distributed. In this scenario, our model shows that CDNs should provide volume discounts to content providers and that the most likely subscribers to CDN services are those content providers with high traffic volume and low data confidentiality requirements. 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.

Next, we numerically determine how the optimal pricing policy changes when traffic is bursty. We find that volume discounts may continue to be offered when CPs have similar traffic burstiness but are no longer optimal, and may even replaced by volume taxes, if CPs exhibit varying degrees of traffic burstiness. Further, we find that pure usage-based pricing policies used by a number of CDNs may be suboptimal in such cases as well. A pricing policy that accounts for both the mean and variance in traffic, such as percentile-based pricing, is more profitable than pure volume-based pricing when there is heterogeneity in burstiness across CPs. A percentile-based pricing policy allows for volume discounts for content providers with high mean traffic rates while simultaneously charging content providers with highly bursty traffic.
In addition to presenting these results, our paper makes a methodological contribution. Since customers (CPs) are heterogeneous with respect to both the value they get from CDNs, and from fulfilling their own content delivery, the pricing problem faced by the CDN is non-trivial. The problem is further complicated by the search for a nonlinear price schedule and the uncertainty in the traffic experienced by CPs. To address this complexity, we first apply a dynamic optimization technique from the calculus of variations to determine the optimal expected price charged to CPs with a given arrival rate. Based on that, we then determine an optimal usage-based price function.

Several interesting directions for future research emerge from our study. While we focus on risk neutral agents in our paper, it is worth investigating service adoption and pricing with risk-averse CPs. For example, CPs may be averse to website downtime due to the negative attention that may result. Furthermore, the tradeoffs between percentile-based pricing and traditional usage-based billing are worth further investigation given the significant interest in industry and limited extant research on the subject.

References


Appendix A: Numerical Results on Optimal Infrastructure Sizing (Poisson Traffic)

We employ numerical tests using parameter values conforming to typical bandwidth and hosting costs to determine the approximate relationship between $\lambda$ and $I^*$. The CP’s optimal infrastructure is determined by trading off the cost of infrastructure against the cost of lost requests. We now fix the parameters associated with these two costs.

For the infrastructure cost function, $C(I) = a_1 \cdot I - a_2 \cdot I^2$, we assume that $a_1 = 3.56$ and $a_2 = 0.000043$. These 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 also 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 with an average purchase of $100$, 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 1024 requests. Figure A1 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 A2 shows the relationship for the case where $\{a_1 = 3.46; a_2 = 0.000043; K = 256; c = 20\}$. While the plotted solutions are all interior solutions, note that the special case where $I^* = \lambda$ (boundary
solution) is also linear. We find that the approximately linear relationship holds for multi-server systems as well (see online appendix for a case with 3 servers). Thus, we assume that a CP’s capacity choice is given by \( I^* = M \cdot \lambda \), where \( M \) is a constant.

[Graphs showing Optimal Infrastructure Level versus Arrival Rate (Case 1) and Case 2)]

**Appendix B. Determining the Optimal Price Function**

Substituting \( I^* = M \cdot \lambda \) and the expressions for \( C(I) \) and \( L(I) \) into equation 13, we get:

\[
\pi = \int_{\lambda} \frac{B}{\lambda^\delta} \left( \tau(N) + a_1M + \frac{c \cdot (M-1)}{M^{K+1} - 1} \right)^W \left( \lambda - a_2M^2 \lambda^2 - P(\lambda) \right)^P d\lambda
\]

\[
\left[ b_1 \int_{\lambda} \frac{\lambda}{\lambda^\delta} \left( \tau(N) + a_1M + \frac{c \cdot (M-1)}{M^{K+1} - 1} \right)^W \left( \lambda - a_2M^2 \lambda^2 - P(\lambda) \right)^P d\lambda \right]
\]

\[
- \left[ b_2 \int_{\lambda} \frac{\lambda}{\lambda^\delta} \left( \tau(N) + a_1M + \frac{c \cdot (M-1)}{M^{K+1} - 1} \right)^W \left( \lambda - a_2M^2 \lambda^2 - P(\lambda) \right)^P d\lambda \right]^2
\]

(B1)

In the integrals above, the limits of \( \lambda \) are from 0 to \( \lambda_{\text{MAX}} \), where \( \lambda_{\text{MAX}} \) is the mean arrival rate of the most popular CP. Letting \( R = \left( \tau(N) + a_1M + \frac{c \cdot (M-1)}{M^{K+1} - 1} \right)^W \left( \lambda - a_2M^2 \lambda^2 \right) \), and denoting \( P(\lambda) \) by \( P \), equation B1 may be rewritten as follows:
\[\pi = \int_\lambda \lambda^{-\delta} \beta (R - P)^W Pd\lambda \left\{ b_1 \left( \int_\lambda \lambda^{-\delta} \beta (R - P)^W d\lambda \right) \right\} - \int_\lambda \lambda^{-\delta} \beta (R - P)^W (P + \Delta P)d\lambda \left\{ b_2 \left( \int_\lambda \lambda^{-\delta} \beta (R - P)^W d\lambda \right) \right\}^{2} \]  
\quad \text{(B2)}

Note that both \(P\) and \(R\) are functions of \(\lambda\). The change in profit from a small perturbation in the price, \(\pi(P + \Delta P) - \pi(P)\), can be obtained by summing the changes in each of the three terms in (B2). We start with the change in the first term,

\[\int_\lambda \lambda^{-\delta} \beta (R - P - \Delta P)^W (P + \Delta P)d\lambda - \int_\lambda \lambda^{-\delta} \beta (R - P)^W Pd\lambda \left\{ b_1 \left( \int_\lambda \lambda^{-\delta} \beta (R - P)^W d\lambda \right) \right\} \]
\[= \int_\lambda \lambda^{-\delta} \beta \cdot P \left\{ (R - P - \Delta P)^W - (R - P)^W \right\} + \lambda^{-\delta} \beta \cdot \Delta P (R - P - \Delta P)^W d\lambda \]  
\quad \text{(B3)}

Similarly, the change in the second term from a small perturbation in price is:

\[b_1 \int_\lambda \lambda^{-\delta} \beta (R - P - \Delta P)^W d\lambda - b_1 \int_\lambda \lambda^{-\delta} \beta (R - P)^W d\lambda = \int_\lambda \lambda^{-\delta} \beta \left\{ (R - P - \Delta P)^W - (R - P)^W \right\} d\lambda \]  
\quad \text{(B4)}

Lastly, the change in the third term of equation (B2) from a small perturbation in price is:

\[b_2 \left( \int_\lambda \lambda^{-\delta} \beta (R - P - \Delta P)^W d\lambda \right) - b_2 \left( \int_\lambda \lambda^{-\delta} \beta (R - P)^W d\lambda \right)^2 \]
\[= 2b_2 \psi \int_\lambda \lambda^{-\delta} \beta \left\{ (R - P - \Delta P)^W - (R - P)^W \right\} d\lambda \]  
\quad \text{(B5)}

where \(\psi\) is the expected volume of traffic handled by the CDN. \(\psi\) can be solved self-consistently from:

\[\psi = \int_\lambda \lambda^{-\delta} \beta \psi \left\{ (R - P)^W \right\} d\lambda = \int_0^{\lambda_{\text{max}}} \beta \lambda^{-\delta} \left\{ \left( \tau(N) + a_1 M + \frac{c(M - 1)}{M^{K+1} - 1} \right) - a_2 \cdot M^2 \lambda^2 - P(\lambda) \right\}^W d\lambda \]  
\quad \text{(B6)}

Adding (B3), (B4) and (B5), we get

\[\pi(P + \Delta P) - \pi(P) = \int_\lambda \lambda^{-\delta} \beta \cdot P \left\{ (R - P - \Delta P)^W - (R - P)^W \right\} + \lambda^{-\delta} \beta \cdot \Delta P (R - P - \Delta P)^W \]
\[b_1 \lambda^{-\delta} \beta \left\{ (R - P - \Delta P)^W - (R - P)^W \right\} + 2b_2 \cdot \lambda^{-\delta} \beta \psi \left\{ (R - P - \Delta P)^W - (R - P)^W \right\} d\lambda \]  
\quad \text{(B7)}
Substituting the following Taylor expansion

\[(R - P - \Delta P)^w = (R - P)^w + \left(\begin{array}{c} N \\ 1 \end{array}\right)(-\Delta P)(R - P)^{w-1} + \cdots + \left(\begin{array}{c} N \\ N \end{array}\right)(-\Delta P)^w\]  

(B8)

into equation B3 and setting the coefficient of \(\Delta P\) to zero, we get

\[P^*(\lambda) = \frac{1}{1 + W}\left[\tau(N) + a_1M + \frac{c(M - 1)}{M^{K+1} - 1} + W(b_1 - 2b_2 \cdot \psi)\right] - \frac{a_2 \cdot M^2}{1 + W} \lambda^2\]  

(B9)

A usage-based price function, \(P^*(X)\) for which the \(E[P(X)]\) trajectory is given by (B6) is:

\[P(X) = \frac{1}{1 + W}\left[\tau(N) + a_1M + a_2 \cdot M^2 + \frac{c(M - 1)}{M^{K+1} - 1} + W(b_1 - 2b_2 \cdot \psi)\right]X - \frac{a_2 \cdot M^2}{1 + W} X^2\]  

(B10)

**Proof of Proposition 2**

\[P^*(X) = \frac{1}{1 + W}\left[\tau(N) + a_1M + a_2 \cdot M^2 + \frac{c(M - 1)}{M^{K+1} - 1} + W(b_1 - 2b_2 \cdot \psi)\right]X - \frac{a_2 \cdot M^2}{1 + W} X^2\]  

(B11)

\[\frac{\partial P(X)}{\partial X} = \frac{1}{1 + W}\left[\tau(N) + a_1M + a_2 \cdot M^2 + \frac{c(M - 1)}{M^{K+1} - 1} + W(b_1 - 2b_2 \cdot \psi)\right] - \frac{a_2 \cdot M^2}{1 + W} (2X)\]  

(B12)

\[= \frac{1}{1 + W}\left[\tau(N) + a_2 \cdot M^2 + \frac{c(M - 1)}{M^{K+1} - 1}\right] + W\left[b_1 - 2b_2 \cdot \psi\right] + \left[a_1M - a_2 \cdot M^2 (2X)\right]\]  

(B13)

Note that \(\left[\tau(N) + a_2 \cdot M^2 + \frac{c(M - 1)}{M^{K+1} - 1}\right]\), the first term in (B13), is always positive. Similarly, \(b_1 - 2b_2 \cdot \psi\) is the marginal cost of the CDN at the optimal price and is always non-negative. Let us now consider the term \(a_1M - a_2 \cdot M^2 (2X)\). Recall that CP’s infrastructure cost is \(C(I) = a_1 \cdot I - a_2 \cdot I^2\).

---

\[^8\text{Note that the optimal price is obtained by setting } \lim_{\Delta P \to 0} \left\{\pi(P + \Delta P) - \pi(P)\right\}/\Delta P = 0. \text{ Thus, we can set the coefficient of } \Delta P \text{ in } \pi(P + \Delta P) - \pi(P) \text{ to zero as higher order terms involving } (\Delta P)^2, (\Delta P)^3, \text{ etc can be approximated to zero when computing } \left\{\pi(P + \Delta P) - \pi(P))/\Delta P\right\}.\]
Thus, for any value \( X \), the term \( a_1 M - a_2 \cdot M^2 (2X) \) represents the CP’s marginal cost of infrastructure when \( I = MX \). This marginal cost of infrastructure is also positive within the feasible region of \( I \). Thus, it follows that \( \frac{\partial P(X)}{\partial X} > 0 \). Also, \( \frac{\partial^2 P(X)}{\partial X^2} = -\frac{2a_2 M^2}{1+W} < 0 \). Thus, \( \frac{\partial P(X)}{\partial X} > 0 \) and \( \frac{\partial^2 P(X)}{\partial X^2} < 0 \).

Appendix C: Numerical Results on Optimal Infrastructure Sizing (MMPP Traffic)

We now study optimal infrastructure sizing under bursty traffic. All parameters except the traffic at the CP are the same as in Appendix A. This includes cost of infrastructure, cost of a lost request and buffer size. For the traffic at the CP server, we assume MMPP parameters as described in section 3.2.1 (\( \lambda_2 = 10\lambda_1 \); \( q_1 = 0.9, q_2 = 0.1 \)). In addition, we maintain constant burstiness (\( \sqrt{\Psi / \bar{X}} \)) across CPs.

The number of lost requests associated with any choice of \( I \) can be computed numerically and the optimal infrastructure level can be computed by trading off the infrastructure cost and cost of lost requests. In Figure C1, we present the optimal infrastructure level associated with a given mean arrival rate under the assumption of MMPP traffic. For the sake of comparison, we also present the infrastructure levels if the traffic were Poisson (from Figure A2). The optimal infrastructure under MMPP traffic is lower than that with Poisson traffic. In Figure C2, we plot the CPs’ net cost (infrastructure cost + cost of
lost requests) and observe that it is higher with MMPP traffic than under Poisson traffic. Thus, the CPs’ expected surplus is lower when the traffic is MMPP.