

Demand-Aware Content Distribution on the Internet

Srinivas Shakkottai, *Member, IEEE*, and Ramesh Johari, *Member, IEEE*

Abstract—The rapid growth of media content distribution on the Internet in the past few years has brought with it commensurate increases in the costs of distributing that content. Can the content distributor defray these costs through a more innovative approach to distribution? In this paper, we evaluate the benefits of a hybrid system that combines peer-to-peer and a centralized client-server approach against each method acting alone. A key element of our approach is to explicitly model the temporal evolution of demand. In particular, we employ a *word-of-mouth* demand evolution model due to Bass [2] to represent the evolution of interest in a piece of content. Our analysis is carried out in an order scaling depending on the total potential mass of customers N in the market. Using this approach, we study the relative performance of peer-to-peer and centralized client-server schemes, as well as a hybrid of the two—both from the point of view of consumers as well as the content distributor. We show how awareness of demand can be used to attain a given average delay target with lowest possible utilization of the central server by using the hybrid scheme. We also show how such awareness can be used to take provisioning decisions. Our insights are obtained in a fluid model and supported by stochastic simulations.

Index Terms—Bass diffusion, content distribution, delay guarantees, peer-to-peer (P2P).

I. INTRODUCTION

MEDIA content delivery over the Internet has been rapidly growing over the past few years. Content that is available spans a wide range, including software packages, music and video files for purchase, streamed music and video, and streamed real-time events. Different types of content require different kinds of quality-of-service (QoS) guarantees, and in turn, performance is dictated by investments into transit bandwidth and server capacity on the part of the content distributor. For example, YouTube, whose content consists of streaming stored video, acts as its own content distributor

(through Google); prior to acquisition by Google, it was estimated that YouTube spent over \$20 million per month for transit bandwidth [1].

Can the content distributor defray this cost through a more innovative approach to distribution? In particular, one possible solution is to leverage peer-to-peer (P2P) technology, which means that each user implicitly provides a fraction of the transit capacity required, hence reducing the cost to the content distributor. At the same time, P2P distribution can perform poorly if the number of peers available is insufficient to meet demand; thus, maintaining some central server capacity alongside P2P distribution seems desirable.

In this paper, we study hybrid schemes that combine P2P and centralized client-server mechanisms for file distribution. In contrast to most prior work on P2P mechanisms, we explicitly consider a *word-of-mouth* demand model for the evolution of interest in a piece of content. In such a model, interest in content grows as interested users contact others and make them interested. This is in contrast to the usual assumption that all demand for a piece of content arrives at once.

The use of this demand model allows us to analyze a delicate tradeoff in the design of hybrid content distribution systems. In the early phases of interest in a piece of content, few individuals possess the content; thus, P2P distribution would perform poorly, so centralized client-server distribution is preferred. On the other hand, as the number of individuals possessing the file increases, the peer cloud grows—and P2P dissemination becomes the preferred mode of operation. This suggests that by optimally combining the two schemes, one can reap the best of both worlds.

Our main goal is to provide qualitative performance analysis that can help guide server provisioning decisions in such hybrid systems. We use a fluid model to study the benefits of combining P2P and client-server distribution in the case of files that are popular and whose demand follows a word-of-mouth evolution. We offer several main insights. First, we characterize the relative performance of P2P, client-server, and a hybrid scheme for dissemination of a single file in a scaling regime where the target population size grows large. Second, we characterize the capacity provisioning necessary to achieve a given average delay target when multiple files are served by the same provider. Finally, we use numerical experiments to illustrate the impact of departures and the efficiency of P2P dissemination on our results. These insights are described fully in Section I-A.

We emphasize that we focus on stored content, i.e., content that is only used after downloading the entire file. Examples are those of software packages and of music or video files sold at online stores such as iTunes. The files do not have a hard delay constraint, and all that is needed is that, on average, the user does not experience a large waiting time. (We also show in Appendix B that streaming of stored content can be viewed as

Manuscript received November 20, 2008; approved by IEEE/TRANSACTIONS ON NETWORKING Editor J. Walrand. First published November 24, 2009; current version published April 16, 2010. This work was supported in part by NSF grants CNS-0644114, CNS-0904609, and CNS-0904520, DTRA grant HDTRA1-09-1-0051, the Office Naval Research, and the Cisco University Research Program. An earlier version [10] of this paper appeared in the Proceedings of the Allerton Conference on Communication, Control and Computing, Monticello, IL, September 2007.

S. Shakkottai was with Stanford University, Stanford, CA 94305 USA. He is now with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: sshakkot@tamu.edu).

R. Johari is with the Departments of Management Science and Engineering, Electrical Engineering, and Computer Science, Stanford University, Stanford, CA 94305 USA (e-mail: ramesh.johari@stanford.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNET.2009.2035047

a possible extension of this model in which each file is broken into chunks, with each chunk having a deadline before which it must be received.)

File distribution should be contrasted with streaming of real-time events, where all interested users are watching (almost) the same content at the same time and, hence, require some version of Internet multicast. In real-time streaming, the chunks of the file are generated in real-time, and there would be a hard delay constraint on the delay between the generation and reception of each chunk. *We do not study hybrid schemes for real-time streaming in this paper using Internet multicast*; that remains an interesting open direction for future work.

A. Main Results

In this paper, we employ a *fluid model* to study the performance of various file distribution schemes. We begin by first modeling demand, or *interest*, in the file using a word-of-mouth propagation model drawn from the seminal work of Bass [2]. In the Bass model, $I(0)$ users of a total population N are initially interested in the file. Uninterested users become interested through random “contact” with an interested user. Thus, interest initially grows approximately exponentially in time, and then tapers off once a level $\Theta(N)$ of interest is reached. The paradigm is very good for modeling the dynamics of interest evolution of popular goods and services. We describe the Bass model in detail in Section II.

We use the Bass diffusion model of demand together with different file delivery service models to estimate the amount of resources required to attain a certain average delay target. To begin, we assume that there are no departures of interested users and subsume the cost of their waiting into the delay, i.e., an infinite wait implies an infinite delay. As described above, we focus on a model of file distribution (rather than real-time streaming).

In Section III, we begin with the case of a *single file* with a potential interested population of N users. We study three types of distribution:

- *Centralized Distribution (CD)*: Here, the content distributor invests in resources (servers and transit bandwidth) such that the maximum rate at which the file can be served is C users per unit time. We show that the average delay experienced by users in this regime is $\Theta(N/C)$.
- *Peer-to-Peer Distribution (P2P)*: Here, the content distributor does not invest in significant central resources, but uses a P2P system to distribute the file. We show that the average delay experienced by users in this regime is $\Theta(\ln(N))$.
- *Hybrid CD-P2P Distribution*: Here, the content distributor uses the available capacity C until the demand exceeds C , and at this point, switches to P2P (and does not use the central server any longer). The average delay experienced by users in this regime is $O(\ln(N/C))$.

We observe that the hybrid regime requires a capacity of $C = N/\ln N$ to attain an average delay of $\ln \ln N$, which is essentially constant. *Thus, the hybrid scheme offers a $\ln N$ reduction in capacity required over centralized distribution for an essentially constant per-user average delay.* The section concludes with some stochastic simulations that show that the fluid model

TABLE I
COMPARISON OF PER-USER AVERAGE DELAY

Capacity	CD Delay	P2P Delay	Hybrid Delay
C	$\Theta(N/C)$	$\Theta(\ln N)$	$O(\ln(N/C))$
$C = N/\ln N$	$\Theta(\ln N)$	$\Theta(\ln N)$	$O(\ln \ln N)$

closely approximates the underlying discrete stochastic system. We summarize the findings in Table I.

In Section IV, we assume instead that a content provider must use common resources to serve *multiple files* simultaneously. We assume that files arrive at rate λ according to a Poisson process. The number of users interested in each file is $\Theta(N)$, i.e., all of them have a large potential user base.

In this setting, our observations are built closely on the single file analysis. We first note that the minimum required capacity for overall stability in the centralized distribution scheme is λN , and the per-user average delay in the P2P scheme is still $\Theta(\ln N)$. However, since the installed capacity is used only for a short time in the hybrid distribution scheme, it may be amortized over different files. We show that if a per-user average delay target of d_N is desired over all files, and $\lim_{N \rightarrow \infty} \sqrt{N} d_N e^{-d_N} = \infty$, then an installed capacity of $C_N \approx \lambda N e^{-d_N}$ meets the delay target. For an almost constant target delay $d_N = \ln \ln N$, the required capacity would be $\lambda N / \ln N$. The result indicates that a tiny sacrifice in per-user delay ($\ln \ln N$, rather than constant) reduces the required capacity in an order sense. In particular, this matches the insight obtained in our single file analysis. As we note in the Appendix, such an analysis can be used to study streaming dissemination of stored content as well.

In Section V, we extend the analytical model of hybrid file distribution to include peer departures and more efficient P2P methods. We study these extensions by numerical experiments. We first validate the scaling results for the analytical studies, and then show that the system can support departure rates of served users near 40% per unit time, without significant performance losses. We suggest that these losses can be mitigated by a dual hybrid scheme in which the CD scheme is used for a short time again *after* the P2P distribution phase. We conclude the section with an investigation of the effects of improving the efficiency of P2P dissemination; we find this does not alter the qualitative insights of our results. We conclude in Section VI.

B. Related Work

There has been considerable research into understanding the impact of P2P technologies on content distribution, with models for the capacity of P2P systems [3]–[7]. In all these models, the arrival rate of demand is either constant or all arrivals are assumed to occur instantaneously. This is in contrast to our work, where we assume a word-of-mouth evolution for interest in content.

Prior work also does not typically consider the question of delays leading to dissatisfied users in a scaling regime as we have done. In [8], an exponentially decaying arrival process is

considered in order to model a flash crowd, and client-server and P2P methods are compared. They find that the P2P method performs better, but is more unreliable with respect to corrupted files. On the other hand, the authors of [9] define a flash crowd to be a sudden spike in arrival rate. They show that under these circumstances, P2P would perform well in multihop wireless networks.

It has been increasingly clear that P2P systems suffer performance issues during the initial stage of file distribution and some help might be necessary. Contemporaneously with the initial presentation of the work presented here in [10], there have been several recent ideas of *server-assisted* P2P content distribution networks [6], [11]–[13] in which a central server is used to boost the performance of P2P systems to improve delay performance. In [6], the system is considered with some additional helpers during the startup phase, which improves performance. However, they do not analytically quantify the gains. Hybrid centralized-P2P schemes are considered in [11]–[13]. However, their focus is on streaming using a P2P multicast structure, which is very different from our focus on file distribution.

II. THE BASS DIFFUSION MODEL

Traditional models of demand growth in new innovations date back to the seminal work of Griliches [14] and Bass [2]. According to the widely employed *Bass model*, the demand for a new product evolves as a function of both direct communication by the manufacturer and by word-of-mouth spreading by interested persons. The resulting evolution of interest shows a characteristic logistic shape. The model has been verified in several different case studies and is used to model the growth of interest over time in a diverse range of products such as microwave ovens, fax machines, and music. Indeed, the model is the de facto standard for understanding the adoption of innovations [15].

We observe similar behavior in user interest in popular videos on CoralCDN [16], a content distribution network hosted on university infrastructure. A typical example is presented in Fig. 1(a), which shows the cumulative demand for one particular video of the Asian Tsunami seen over a month in December 2005, the one-year anniversary of the disaster. Its popularity follows the pattern that would be seen in a typical viral model, a logistic curve. Similar behavior has been observed in user interest in videos on YouTube [17] (although those observations are somewhat coarse as they are taken only once a day).

In our setting, we consider the following three key quantities for a fixed piece of content:

- $N(t)$, the total population of users at time t ;
- $I(t)$, the number of users interested in the content at time t ;
- $P(t)$, the number of users that possess the content at time t .

Quite often, the major driver in the adoption of a product is word-of-mouth interest propagation, which is modeled for a fixed population N using the following differential equation:

$$\frac{dI(t)}{dt} = \left(\frac{N - I(t)}{N} \right) I(t). \quad (1)$$

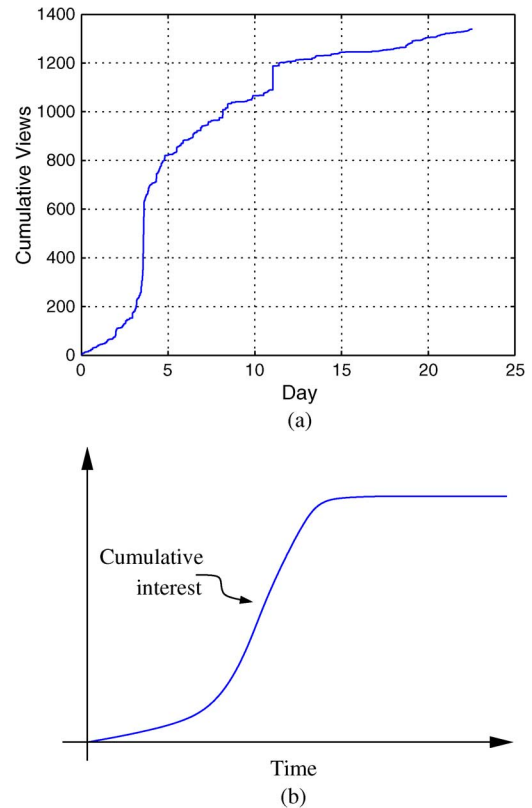


Fig. 1. (a) Single-file cumulative demand for a file over one month (December 2005–January 2006). The demand looks much like (b) the cumulative demand seen in a Bass diffusion of interest with world-of-mouth spreading

The preceding equation can be considered a “fluid limit” of a system where each interested member of the system attempts to cause a randomly selected member to become interested. The number of users that could potentially be interested is $N - I(t)$. Thus, the probability of finding such a user is $(N - I(t))/N$, and the fluid model follows. We assume that an interested user can interest other users at rate 1 per unit time without loss of generality. The above equation is easily solved and yields the so-called *logistic function* as its solution

$$I(t) = \frac{I(0)e^t}{1 - (I(0)/N)(1 - e^t)}. \quad (2)$$

The logistic function is illustrated in Fig. 1(a) and has a characteristic S-shape. (Such behavior is also seen in many different epidemic contexts [18].)

Note that in the model described in the previous paragraph, we implicitly assume that every user can randomly contact any other user. In fact, interested users are only likely to contact others to whom they are connected in a social network. Of course, this graph of users is not very likely to be complete. We can thus consider a modified version of the Bass model where users only contact others to whom they are connected in a given graph structure. As we show in Appendix A, in the case of a d -regular random graph, such a model still leads to the basic model described in (1), but with a different constant gain. More generally, for graphs with arbitrary degree distributions, we find that the logistic form described above is an upper bound on interest evolution with time.

It is instructive to study the arrival rate of interested users for logistic demand as a function of time. The time derivative of (2) is

$$\frac{NI(0)e^t(N - I(0))}{(N - I(0) + I(0)e^t)^2}. \quad (3)$$

It is clear from (3) that for $t \ll \ln N$, the arrival rate is exponentially increasing. For $t = \ln N$, we have from (2) that the number of interested users is approximately N , and for $t \gg \ln N$, from (3) the arrival rate is exponentially decreasing. Thus, *all crowds have two phases*—one where the arrival rate is *exponentially increasing* and one where the arrival rate is *exponentially decreasing*.

Given a fixed number of servers, the Bass diffusion model suggests that as the number of interested users increases, the delay in service perceived by users should first increase, and then gradually decrease. Informally, the servers would be overwhelmed for a period of time before being able to cope with the “crowd”—a characteristic of the so called “flash crowd” effect. It is important to note that even if we reduce the time-scale of events by a constant factor (so that interest grows more steeply), the order of magnitude of the delays (relative to N) would be unchanged. Exact performance will depend on the exogenous constants parametrizing the system, such as installed server capacity.

III. AVERAGE DELAY FOR DIFFERENT DISTRIBUTION METHODS

The total delay experienced by all users for any work conserving service regime is the area between the cumulative demand and service curves [19]. In this section, we will obtain analytical expressions that quantify this area for three different distribution methods: 1) Centralized Distribution (CD); 2) Peer-to-Peer (P2P) Distribution; and 3) Hybrid CD-P2P Distribution, in which the CD servers are used for a short time initially to boost the performance of P2P distribution.

Our focus in this section is on the dynamics of interest arousal and satisfaction of exactly one file, e.g., a piece of software. We do not consider departure of users in this section, as departure of an unsatisfied user entails an opportunity cost that has to be considered separately; user departures are considered in Section V. Here, we subsume the opportunity cost into the user delay since the longer one waits, the larger the area between the interest and service curves.

Our main insight at the end of this section will be that sacrificing delay performance by a small amount buys us a significant decrease in the required server capacity if a hybrid of CD and P2P is used. For example, we will see that if we are willing to allow the average per-user delay to scale with N as $\ln \ln N$, i.e., almost constant for any reasonable N , using a hybrid approach could achieve a decrease in required capacity of the CD by a factor of $\ln N$. Also, this capacity is used only for a short time (when interest in the file first builds). In other words, small sacrifices in the average delay desired have a very large impact on the capacity required to meet the target delay if we employ a hybrid distribution scheme.

In the next three subsections, we explicitly find the total delay experienced by all users under the CD, P2P, and hybrid schemes.

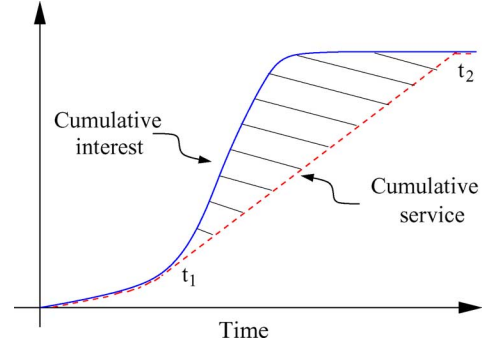


Fig. 2. The total delay using centralized distribution is the area between cumulative interest and service.

Since we analyze each of these models using the same interest curve (2), we begin by first finding the area under this curve, which is easily determined by integration of (2) as

$$\begin{aligned} I_A(t) &\triangleq \int_0^t I(t)dt = \int_0^t \frac{I(0)e^t}{1 - (I(0)/N)(1 - e^t)} dt \\ &= N \ln \left(1 - \frac{I(0)}{N}(1 - e^t) \right). \end{aligned} \quad (4)$$

Our analysis proceeds by calculating the area under the service curve for each of the three schemes we consider and subtracting the area under the interest curve.

A. Centralized File Distribution

We begin in this subsection by finding the average user delay when centralized distribution is used. In the centralized model, the content distributor purchases a certain amount of transit bandwidth C and sets up a central server-bank. Interested users have to download the file from the server-bank, and service rate is limited by the bandwidth C . We assume an *access-core* network topology, wherein users are only limited by their access bandwidth and the core network is uncongested. Furthermore, we assume that users only have an upload constraint. We make the following reasonable assumptions.

Assumption 1: We assume that $P(0) = I(0) \in \Theta(1)$ for simplicity. We also assume that $C = C(N)$, i.e., C scales with N ; for example, if $C = \kappa \ln N$, we say that $\Theta(C) \equiv \Theta(\ln N)$. We assume that there exists a constant \underline{k} such that $\underline{k} \leq C(N) \leq N$ for all N .

Under the assumption, the service curve follows $I(t)$ until some time t_1 when $dI(t)/dt > C$, after which the service occurs at a constant rate. At time $t_2 \geq t_1$ the curves meet again. Fig. 2 illustrates the centralized distribution model, with the area between the curves (which is the total delay) being shaded. The following proposition characterizes three “phases” of the service curve.

Proposition 1: Under Assumption 1 with $C \in o(N)$, there exist times $t_1 \in \Theta(\ln(C/I(0)))$ and t_2 such that $N/2C + t_1 - P(t_1)/C < t_2 \leq N/C + t_1 - P(t_1)/C$ that divide the service curve $P(t)$ into three phases as follows.

- 1) Phase 1: For $t < t_1$, we have $P(t) = I(t)$ and $P(t_1) = I(t_1) = (N/2)(1 - \sqrt{1 - 4C/N})$.

- 2) Phase 2: For $t_1 \leq t \leq t_2$, we have linear growth at rate C , i.e., $P(t) = C \times (t - t_1) + P(t_1)$.
 3) Phase 3: For $t > t_2$, we again have $P(t) = I(t)$.

Proof: By definition, t_1 is the first time at which the slope of the interest curve is equal to C . Thus, we have

$$\left. \frac{dI(t)}{dt} \right|_{t=t_1} = \frac{NI(0)e^{t_1}(N - I(0))}{(N - I(0) + I(0)e^{t_1})^2} = C. \quad (5)$$

For $t = \ln(C/I(0))$, since $C \in o(N)$, we have

$$\left. \frac{dI(t)}{dt} \right|_{t=\ln(C/I(0))} = \frac{NC(N - I(0))}{N^2(1 - I(0)/N + C/N)^2} \rightarrow C, \quad \text{as } N \rightarrow \infty.$$

Also, rearranging (5), we have

$$\begin{aligned} \frac{I(t_1)(N - I(t_1))}{N} &= C \\ \Rightarrow I(t_1) &= P(t_1) = \frac{N - \sqrt{N^2 - 4NC}}{2} \\ &= \frac{N}{2} \left(1 - \sqrt{1 - \frac{4C}{N}} \right) \end{aligned}$$

which completes the first part of the proof.

For the second part, consider $\hat{t} \triangleq \ln((N - I(0))/I(0))$. It is clear that: 1) $dI(t)/dt|_{\hat{t}} = N/4$; and 2) $I(\hat{t}) = N/2$. By assumption, since $C \in o(N)$, the above observations imply that: 1) $t_2 > \hat{t}$; and 2) $I(t_2) > N/2$. Since by definition, $P(t_2) = I(t_2)$, this in turn yields $N/2C + t_1 - P(t_1)/C \leq t_2 \leq N/C + t_1 - P(t_1)/C$. ■

We can now determine the average per-user delay by considering the area between interest and service curves. Informally, if all N users were to become interested instantaneously at time 0, then the time taken to serve all N users using a server with capacity C would be N/C , thus the average delay per user is $\Theta(N/C)$. Of course, this analysis is very coarse, but we show below that this intuition is correct in the order sense.

Theorem 2: In a system with Bass diffusion of interest without departures and centralized distribution with a capacity C , the average delay per user is $\Theta(N/C)$.

Proof: From (4), the area under the interest curve in the interval $[0, t_2]$ is $I_A(t_2)$. Call this area A_1 . Similarly, define $A_2 \triangleq I_A(t_1)$. Now, we can find the area under the service curve in the interval $[t_1, t_2]$ by integration of $P(t)$ in Phase 2. This yields

$$A_3 \triangleq \frac{C \times (t_2^2 - t_1^2)}{2} - (Ct_1 - P(t_1))(t_2 - t_1)$$

Then, since $P(t) = I(t)$ in $[0, t_1]$ and $[t_2, N]$, the area between the curves is $A \triangleq A_1 - A_2 - A_3$, and hence the area between the interest and service curves is

$$\begin{aligned} A &= N \ln \left(\frac{N - I(0) + I(0)e^{t_2}}{N - I(0) + I(0)e^{t_1}} \right) - \frac{C \times (t_2^2 - t_1^2)}{2} \\ &\quad + (Ct_1 - P(t_1))(t_2 - t_1) \end{aligned} \quad (6)$$

which we can rewrite as

$$\begin{aligned} A &= \frac{N^2}{C} \left(\ln \left(\frac{N - I(0) + I(0)e^{t_2}}{N - I(0) + I(0)e^{t_1}} \right)^{C/N} - \frac{C^2}{2N^2} (t_2^2 - t_1^2) \right. \\ &\quad \left. + \left(\frac{C}{N^2} \right) (Ct_1 - P(t_1))(t_2 - t_1) \right). \end{aligned}$$

Substituting t_1, t_2 , and $P(t_1)$ from Proposition 1 and dividing by N^2/C gives

$$\lim_{N \rightarrow \infty} \frac{AC}{N^2} = \text{constant}. \quad (7)$$

Dividing the above result by N (the number of users served) yields the proof. ■

Not surprisingly, the per-user delay is completely determined by the service capacity C that the content distributor is able to purchase.

B. Peer-to-Peer File Distribution

We now consider a fluid model of a P2P system for file distribution. We first obtain an analytical expression for the cumulative number of served users as a function of time. We consider the following differential equation describing the evolution of served users:

$$\frac{dP(t)}{dt} = \eta(I(t) - P(t)) \left(\frac{P(t)}{N} \right). \quad (8)$$

The term on the right describes the rate at which the $I(t) - P(t)$ interested users who do not possess the file obtain pieces (or the whole file) from the $P(t)$ users who possess it, with η being a constant that depends on upload capacity of users as well as efficiency of dissemination [4]. Our equation reflects the following model. Suppose each interested user who does not possess the file randomly contacts another user. If that user possesses the file, then the file is exchanged with efficiency coefficient η . We assume that peers do not depart from the system, although we relax this assumption in Section V.

Clearly, our model is an inefficient P2P scheme since the peer selection is random. However, this means that our analysis gives a bound on the *worst-case performance of P2P*. Using this model for the P2P part of the hybrid scheme that we will discuss in the next subsection implies that a real hybrid system can only give better delay performance. Even with such a conservative model of P2P performance, we can show analytically that significant gains are possible by using a hybrid system. We provide some insights into the effect of using a more efficient P2P scheme in Section V, where users search for content only among other interested users. Our qualitative insights are robust to this change.

We make the following technical assumptions.

Assumption 2: We assume that $P(0)$ can be any function of N such that $P(0) \in o(N)$. We also assume that $\eta \in \Theta(1)$. The

assumption that $\eta \in \Theta(1)$ is consistent with an assumption that every peer has finite upload capacity.

We solve (8) jointly with the evolution of interest given in (2) to obtain the following proposition.

Proposition 3: The cumulative number of users that possess the file in the P2P model is given by

$$P(t) = N (N - I(0) + I(0)e^t)^\eta \\ \div \left(\eta \left((N - I(0))^\eta t + \sum_{i \neq 0} \binom{\eta}{i} (N - I(0))^{\eta-i} \frac{I(0)^i e^{it}}{i} \right) + Q \right)$$

where

$$Q = \frac{N^{\eta+1}}{P(0)} - \eta \sum_{i \neq 0} \binom{\eta}{i} (N - I(0))^{\eta-i} \frac{I(0)^i}{i}$$

and $\binom{\eta}{i}$ is defined in the usual sense for $\eta \in \mathbb{R}$ and $i \in \mathbb{N}$.

Proof: The cumulative number of users that possess the file is given by (8)

$$\frac{dP(t)}{dt} = \eta \left(\frac{I(t) - P(t)}{N} \right) P(t) \\ = \frac{\eta I(t)}{N} P(t) - \frac{\eta P^2(t)}{N}$$

which is a second-order Bernoulli differential equation. Substituting $V(t) = 1/P(t)$, we obtain

$$\frac{dV(t)}{dt} + \frac{\eta I(t)}{N} V(t) = \frac{\eta}{N}.$$

The above equation has a closed-form solution given by

$$V(t) = \frac{\eta \int J(t)/N dt + Q}{J(t)} \quad (9)$$

where Q is a constant and

$$J(t) = \exp \left(\int \left(\frac{\eta I(t)}{N} \right) dt \right) \\ = \exp \left(\int \frac{\eta I(0)e^t}{N - I(0)(1 - e^t)} dt \right) \\ = \exp \left(\eta \ln (N - I(0) + I(0)e^t) \right) \\ = (N - I(0) + I(0)e^t)^\eta. \quad (10)$$

Here, we have used the expression for $I(t)$ from (2). Now, in order to obtain the closed-form solution, we also require

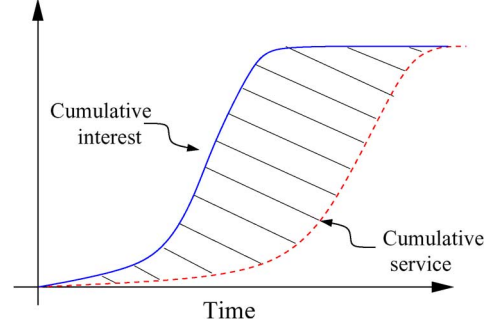


Fig. 3. The total delay using P2P delivery is the area between cumulative interest and service.

$\int J(t)/(N - t)dt$, so we proceed to integrate the above expression. We have

$$\int J(t)dt = \int (N - I(0) + I(0)e^t)^\eta dt \\ = \int \sum_i \binom{\eta}{i} I(0)^i e^{it} (N - I(0))^{\eta-i} dt \\ = (N - I(0))^\eta t + \sum_{i \neq 0} \binom{\eta}{i} (N - I(0))^{\eta-i} \frac{I(0)^i e^{it}}{i}.$$

Substituting into (9) yields the result. \blacksquare

We assume that the system is work-conserving given available capacity at any time. Under this assumption, in order to find the total delay in using P2P file distribution, we need to find the area between the interest and service curves. An example is shown in Fig. 3.

We can develop an informal characterization of the area between the curves in the following manner. Assume that all N users become interested instantaneously at time 0. Then, since the Bass diffusion is roughly exponential in the early stages, we expect that a P2P service discipline would require approximately $\ln(N/P(0))$ time to satisfy this interest, by which time $\Theta(N)$ individuals would have been served. Thus, we expect the total delay to be $\Theta(N \ln(N/P(0)))$, and the per-user average delay to be $\Theta(\ln(N/P(0)))$. Again, this analysis is quite coarse. However, the following theorem formally justifies our intuition.

Theorem 4: In a system with Bass diffusion of interest and a P2P service discipline, the average per-user delay is $\Theta(\ln(N/P(0)))$.

Proof: We first need to find the area under the service curve, which is $\int_0^t P(t)dt$. This can be obtained by straightforward integration to yield

$$\int_0^t P(t)dt = \frac{N}{\eta} \ln \left(\eta \left((N - I(0))^\eta t + \sum_{i \neq 0} \binom{\eta}{i} \right. \right. \\ \left. \left. \times (N - I(0))^{\eta-i} \frac{I(0)^i e^{it}}{i} \right) + Q \right) \\ - \frac{N}{\eta} \ln \left(\frac{N^{\eta+1}}{P(0)} \right). \quad (11)$$

We now find the total area between the curves by substituting $t = N$ in (4) and (11) and taking the difference. The difference of the two areas can be written as

$$\begin{aligned} & N \ln \left(\frac{e^N}{N} (Ne^{-N} - I(0)e^{-N} + I(0)) \right) \\ & - N \ln \left(\frac{e^N}{N} \left(\frac{\eta P(0)e^{-\eta N}}{N} (N - I(0))^\eta + \frac{\eta P(0)e^{-\eta N}}{N} \right. \right. \\ & \quad \times \sum_{i \neq 0} \binom{\eta}{i} (N - I(0))^{\eta-i} \frac{I(0)^i e^{iN}}{i} \\ & \quad \left. \left. + \frac{P(0)e^{-\eta N} Q}{N} \right)^{1/\eta} \right) \end{aligned}$$

which, for large N , converges to

$$\frac{N}{\eta} \ln \left(\frac{N}{\eta P(0)} \right). \quad (12)$$

Dividing by the number of users served, N , yields the proof. ■

The theorem essentially reveals that the performance of the P2P system can only be improved by increasing the number of initial hosts who possess the file $P(0)$. It also demonstrates that the delay per user is insensitive to the initial number of interested hosts $I(0)$. There is a simple intuition for this observation: Regardless of the value of $I(0)$, in a pure P2P system, roughly $\ln N/P(0)$ time units are required to ensure a large number of interested hosts also possess the file and are thus available to service all the remaining demand.

C. Hybrid File Distribution

We showed above how the average per-user delay in the P2P case depends strongly on the initial number of hosts with the file $P(0)$. If we first use a centralized distribution scheme, and then switch to P2P dissemination at a later time, the initial number of hosts $P(0)$ can be boosted to significantly improve performance. In this section, we consider precisely such a hybrid scheme.

We use a centralized distribution scheme with capacity C until the time t_1 in Proposition 1, the time at which demand outstrips C ; we switch to P2P at this point. Note that, in principle, we could envision a scheme where the server is always available to “boost” the P2P system. As we show via simulation in Section III-E, such a modification to the scheme does not significantly change our predictions.

We illustrate a typical case of the hybrid scheme in Fig. 4. It is straightforward to characterize the delay performance of the above scheme. Using the framework developed above, we have the following result.

Theorem 5: In a system with Bass diffusion of interest and a hybrid service discipline that switches from CD with capacity C to P2P when $dI(t)/dt > C$, the average per-user delay is $O(\ln(N/C))$.

Proof: Since the switch happens exactly when $dI(t)/dt > C$, there is no delay in the CD phase. Call the time at which

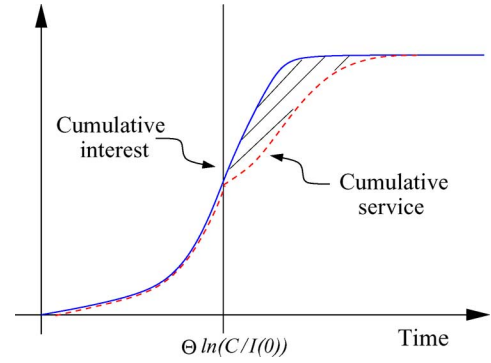


Fig. 4. The total delay in the hybrid CD-P2P scheme. CD is used in the initial phase to boost the performance of the P2P phase.

we switch as t_1 . To evaluate the total delay, we first have from Theorem 4 that for large N the area between the curves is

$$A \approx \frac{N}{\eta} \ln \left(\frac{N}{\eta P(t_1)} \right).$$

Substituting the value of $P(t_1)$ from Proposition 1, we obtain

$$\begin{aligned} A & \approx \frac{N}{\eta} \ln \left(\frac{N}{\eta(N/2)(1 - \sqrt{1 - 4C/N})} \right) \\ & = \frac{N}{\eta} \ln \left(\frac{2}{\eta(1 - \sqrt{1 - 4C/N})} \right). \end{aligned}$$

Now, since $C \in o(N)$, $4C/N < 1$ for N large enough. Thus

$$\begin{aligned} A & \approx \frac{N}{\eta} \ln \left(\frac{2}{\eta(1 - 1 + 2C/N + 2C^2/N^2 + \dots)} \right) \\ & \leq \frac{N}{\eta} \ln \left(\frac{N}{\eta C} \right) \end{aligned}$$

and we divide by N to obtain the result. ■

The intuition behind this result is as follows. Since there is no delay in the CD phase (we switch before delays can occur), delays only occur in the P2P phase. The P2P phase starts with an initial condition of $P(t_1) \approx C$ users who possess the file, and the result follows from (12).

D. Example

We illustrate our results by means of an example. Suppose a server capacity of $C = N/\ln N$ is provisioned. What is the per-user average delay performance of the three schemes? From Theorem 2, the per-user delay in the CD case is $\Theta(\ln(N))$. From Theorem 4, the per-user delay in the P2P case is $\Theta(\ln(N))$ (independent of C). Finally, from Theorem 5, the per-user delay in the hybrid case is $O(\ln \ln N)$, i.e., it is *practically constant*.

E. Simulations

We simulate our example as a stochastic system using Matlab. The purpose of simulation is twofold. First, we would like to confirm that the results we derived capture the correct order sense scalings in the system. Second, we would like to confirm

that the fluid approximation that we use does not impact the validity of our results. In the simulations, we assume that time is discrete, and there are $N = 100\,000$ users. At each time slot, each interested user chooses a Poisson distributed number of other users (with mean 1) to spread interest. We compare four possible service regimes.

First, we consider centralized distribution using a FIFO policy with a service rate of $C = 8000 \approx N/\ln N$. Our predicted value of average user delay from (6) is about 2.5 units, and the simulation yields a value of 2.2 units. Evolution of demand and service are shown in Fig. 5(a). In the plot, the left curve is the cumulative interest, while the right curve is the cumulative supply.

Second, we consider the P2P system with efficiency $\eta = 2$. The P2P dynamic is similar to the diffusion of interest. Each interested user without the file chooses a Poisson number of other users (with mean 2) from whom to obtain the file, and if at least one of the contacted users has it, the file is downloaded. From (12), the average user delay should be about 6 units, and our simulation yields a value of 8.5 units. Evolution is seen in Fig. 5(b).

Third, we consider the hybrid scheme, which we simulate by causing the system to switch from CD to P2P when the queue size at the server is $P(t_1)$ given in Proposition 1 (≈ 8700 users). From (12) the expected user delay should be about 1.2 units, and our simulation yields a value of 1.15 units. The state evolution can be seen in Fig. 5(c). We note that, as expected, the stochastic system matches our fluid model quite well. This lends further credence to our analysis.

Fourth, we simulate the system in which both the server and P2P are used simultaneously. In this simulation, users attempt to obtain the file using P2P. If they are unsuccessful, they may contact the server. The server serves the first 8700 users that contact it if capacity is available (as in the previous case), after which it serves only users who have experienced a delay greater than 2 units. This is in keeping with the idea of attempting to use the P2P system if at all possible and using the server to “boost” if needed. The average delay per user in this case is about 0.8 units, which is similar to the hybrid system with explicit server and P2P phases. Fig. 5(d) illustrates this case. We observe that in the initial phase, almost all service is due to the server, i.e., P2P has small effect here. In the latter part, server usage is quite small, i.e., P2P has sufficient capacity here.

We simulated the system with other values of population size and capacity and summarize all results in Table II. The performance of the hybrid system is superior to the other two for different scalings of population size and capacity.

IV. A DYNAMIC MODEL: SERVING MULTIPLE FILES

The results of the previous section applied to the case of dissemination of a single file. However, in reality, files arrive at random time instants as they are generated, and interest in each file evolves from that point on. Thus, each such file will have a different interested user population associated with it. In this section, we study the performance of a system with files arriving over time and with common resources employed to serve the content. Hence, we now consider the system with a Poisson arrival rate of λ files per unit time. We ignore files that are un-

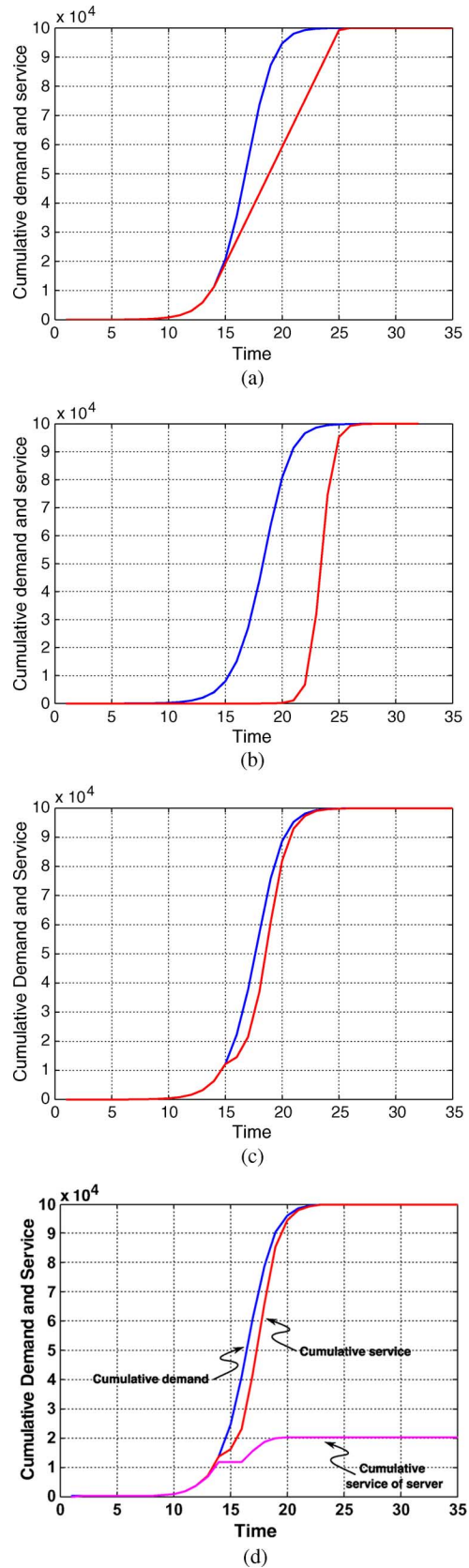


Fig. 5. (a) Centralized distribution and (b) P2P distribution. The hybrid scheme (c) switches between them, while (d) uses both simultaneously.

popular (population $\ll N$), and assume instead for simplicity that all files have the same population of interested users up to

TABLE II
COMPARISON OF PER-USER AVERAGE DELAY

Population	Capacity	CD Delay	P2P Delay	Hybrid Delay
1000	100	1.16	6.0	0.98
10000	1000	1.2	7.6	0.9
100000	4000	2.2	8.5	1.15

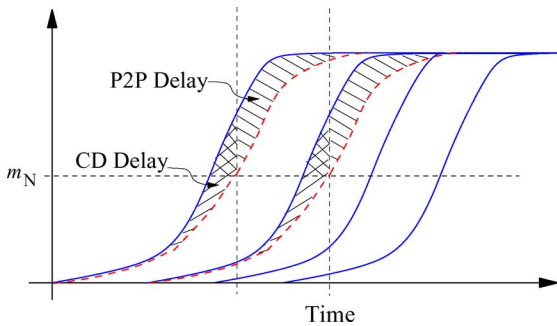


Fig. 6. The cumulative demand (solid line) and supply (dashed) curves for multiple files being served using the hybrid scheme. The centralized distribution method is used to serve m_N users for each file. Delays are now experienced in both CD and P2P phases.

a constant factor. Since the constant makes no difference to the order results, we take the user base of all files as N .

We investigate the following question: Given that for each file we choose to serve m_N customers using the centralized method and then switch to P2P, what is the minimum transit capacity C_N required in order to ensure an average delay d_N per user? (In answering the question, we will make some observations on the pure CD and P2P cases as well.) We have changed the notation slightly from the previous section to indicate the dependence of capacity on N explicitly. Also, note that we are allowing per-user delay to scale with N , i.e., we are allowing non-constant user delays. Our results complement those obtained in the single file analysis of Section III. In particular, we find that if we wish to provision for a per-user average delay per file of $d_N = O(\ln \ln N)$, then capacity $C_N = \lambda N / \ln N$ suffices to meet the delay target.

Fig. 6 illustrates the scenario we consider. Files arrive at some time, and each file is associated with a Bass diffusion of interest. Each file is distributed using the hybrid scheme. There are three delay terms that must be considered for each file:

- 1) the delay incurred in serving m_N users in the CD phase;
- 2) the delay incurred by users who arrive after the m_N th user, but have to wait until the P2P phase begins;
- 3) the delay incurred by the $N - m_N$ users who are served using the P2P method.

In the next three subsections, we analytically characterize the delay incurred in each of these phases. The total delay incurred in each of these phases is the relevant shaded area between the

interest and service curves in Fig. 6. We note that we could potentially reduce the delay 2) by simultaneously using the centralized and P2P methods. However, we will see that the other two delays dominate, and this would not change our results.

A. Delay Incurred in the P2P Phase

The average delay incurred in the P2P phase for any user can be estimated using the framework that we have developed in the previous sections. Recall that our target is to achieve a per-user average delay of d_N . By Little's Law, since the average arrival rate of files into the system is λ , the average number of files in the system should be λd_N . We need to ensure that there is sufficient capacity in the P2P system to serve these files. Let the per-user capacity be denoted by ξ_N . In the single-file case, we had $\xi_N \in \Theta(1)$. In the current setting, we continue to assume that each peer would utilize only $\Theta(1)$ capacity for each file. Suppose that we provision the server such that m_N users are served before the beginning of the P2P phase; then, the following result follows from Theorem 4.

Proposition 6: Consider using the hybrid dissemination method, where for each file $m_N \in o(N)$ users are served during the CD phase. Assume that the capacity of each user in the system is $\xi_N \in \Omega(\lambda \ln(N/m_N))$, but that each peer only utilizes $\Theta(1)$ capacity for each file. Then, the average delay incurred by each user served during the P2P phase is $O(\ln(N/m_N))$. In other words, the number of users that must be served in the CD phase to achieve an average delay d_N per user in the P2P phase is $m_N \in \Omega(Ne^{-d_N})$.

The above proposition yields two main conclusions. First, sufficient capacity will be present in the P2P phase if the per-user capacity scales as $\Omega(\lambda \ln(N/m_N))$. If the desired delay is $\Theta(\ln \ln N)$, then m_N is $N / \ln N$, and a sufficient per-user capacity is $\Omega(\lambda \ln \ln N)$, which is practically constant. Second, the above proposition gives an estimate of the number of users m_N to be served in the CD phase in order to meet a per-user average delay target of d_N in the P2P phase¹. We subsume the proportionality constant in $m_N \in \Omega(Ne^{-d_N})$ into d_N and say that we require $m_N = Ne^{-d_N}$. We assume that $d_N \geq 1$. Indeed, to realize this value of delay, we would need $\Theta(N)$ capacity. We need to provision sufficient capacity C_N so as to serve these m_N users (for each file) with a per-user delay that does not exceed d_N . We determine such a choice of C_N in the next subsection.

B. Delay Incurred in the CD Phase

We first find the smallest necessary value of C_N such that the number of files in the system does not grow unboundedly large. We assume that the client-server system serves users in the order that they arrive (first-come, first-served). Let $n(t)$ be the number of files being served in the client-server phase at time t , $x(t)$ be the rate at which all these files are being served, and T be the service time per file after which the system switches to P2P distribution. A simple fluid approximation then suggests

$$\dot{n}(t) = \lambda - \frac{n(t)x(t)}{m_N}.$$

¹Note that from Theorem 4, pure P2P distribution yields a per-user delay of $\Theta(\ln N)$ as before.

Taking $C_N = \lambda m_N$, we have

$$x(t) = \min\left(\frac{m_N}{T}, \frac{\lambda m_N}{n(t)}\right).$$

Thus

$$\dot{n}(t) = \lambda - \min\left(\frac{n(t)}{T}, \lambda\right).$$

Hence:

- if $\lambda > n(t)/T$ (i.e., $n(t) < \lambda T$), then $\dot{n}(t) > 0$.
- if $\lambda \leq n(t)/T$, then $\dot{n}(t) = 0$.

If the system starts at an initial condition $n(0) \leq \lambda T$, then the system will converge to the steady-state value. Hence, for $C_N > \lambda m_N$, the number of files in the system does not grow unbounded. (Note that using the same argument, if pure CD were used, the minimum required capacity for stability is λN .)

Although the number of files is bounded with this choice of capacity, such a choice does not ensure that we meet a per-user delay target since the total arrival rate of users could exceed the installed capacity at some time instants. From Proposition 6, we know that we have to serve $m_N = Ne^{-d_N}$ users in the CD phase in order to achieve the delay target d_N in the P2P phase. We first study the case where we let $C_N \rightarrow \lambda m_N$ (i.e., heavy traffic) and characterize the per-user delay in this case to see how it compares to d_N . If it is smaller than d_N , then $C_N \approx \lambda m_N$ is sufficient to meet our per-user delay target per file. Otherwise, we have to provision $C_N \gg \lambda m_N$ in order to meet the delay target.

Proposition 7: Suppose that the distribution of multiple files is undertaken using the hybrid CD-P2P scheme. Let the potentially interested population for each file be N , and assume that interest in each file evolves according to the Bass model. Let files arrive according to a Poisson process at rate λ , which is constant independent of N .

Suppose the required delay target is a per-user delay of $d_N \geq 1$, and we serve $m_N = Ne^{-d_N}$ interested users per file using an installed transit capacity C_N , where $C_N, m_N \in o(N)$. In a heavy traffic regime with $\lambda m_N/C_N = 1 - 1/\sqrt{N}$, we have that for large N , the per-user delay D in the CD phase of any file satisfies

$$\mathbb{P}(D > d_N) \leq \exp\left(-\frac{2\lambda N\sqrt{N}d_N e^{-d_N}}{(\sqrt{N}-1)^2}\right).$$

Thus, if the target average delay d_N is such that $\lim_{N \rightarrow \infty} \sqrt{N}d_N e^{-d_N} = \infty$, then choosing $C_N \approx \lambda m_N$ yields a per-user delay in the CD phase of d_N with high probability.

Proof: We first approximate the arrival rate of users in the Bass model for the CD phase by Poisson arrivals of files, each with m_N users. Such an arrival process takes the form of an M/D/1 process. Furthermore, we assume that service for any file begins only after these m_N users have requested it. Thus, we first have a delay associated with waiting for m_N users to arrive, followed by a delay in serving these users. This yields an upper bound on the delay.

Step 1: We first show that under our assumptions, the error in calculating the per-user average service delay due to the M/D/1 approximation is $\Theta(1)$. Solving the differential (1) for t , and using $I(t) = m_N = Ne^{-d_N}$, we obtain

$$t = \ln\left(\frac{N}{I(0)}\right) - d_N - \ln\left(\frac{N(1 - e^{-d_N})}{N - I(0)}\right).$$

The per-user average delay experienced by users up till this time can be obtained from (4) as

$$\ln\left(1 - \frac{I(0)}{N} + \frac{e^{-d_N}\left(1 - \frac{I(0)}{N}\right)}{1 - e^{-d_N}}\right). \quad (13)$$

Now, for $d_N \geq 1$ and $\lim_{N \rightarrow \infty} \sqrt{N}d_N e^{-d_N} = \infty$, we have

$$\frac{e^{-d_N}}{1 - e^{-d_N}} \leq \frac{d_N e^{-d_N}}{1 - e^{-d_N}} < \frac{1}{\sqrt{N}(1 - e^{-1})}. \quad (14)$$

The result now follows from (13) and (14).

Step 2: We study the heavy traffic regime for the M/D/1 process in a manner similar to [20]. Assume that time is discrete and indexed by k . Let $a(k)$ be the number of users who arrive at time k . The total arrivals up to time $k - 1$ is

$$A(k) \triangleq \sum_{i=0}^{k-1} a(i).$$

It can be shown that the queue length at time k

$$q(k) = V(k) - \min_{1 \leq i \leq k} V(i)$$

where $V(k) \triangleq A(k) - kC_N$. Then

$$\frac{q(\lfloor Nt \rfloor)}{\sqrt{N}} = \frac{V(\lfloor Nt \rfloor)}{\sqrt{N}} - \min_{s=1/N, 2/N, \dots, \lfloor Nt \rfloor/N} \frac{V(\lfloor Ns \rfloor)}{\sqrt{N}}.$$

Assume that $X(t) \triangleq \lim_{N \rightarrow \infty} q(\lfloor Nt \rfloor)/\sqrt{N}$ exists. Also, it can be shown that

$$\lim_{N \rightarrow \infty} \frac{V(\lfloor Nt \rfloor)}{\sqrt{N}} = W(t) \sim \mathcal{N}(-C_N, \lambda m_N).$$

Furthermore, $W(t)$ is a Brownian motion. Hence

$$X(t) = W(t) - \inf_{0 \leq s \leq t} W(s)$$

and we can show that

$$\lim_{t \rightarrow \infty} \mathbb{P}(X(t) > x) \approx \exp\left(-\frac{2xC_N}{\lambda m_N}\right).$$

Thus, for large N

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbb{P}(q(\lfloor Nt \rfloor) > B) &= \lim_{t \rightarrow \infty} \mathbb{P}\left(\frac{q(\lfloor Nt \rfloor)}{\sqrt{N}} > \frac{B}{\sqrt{N}}\right) \\ &\approx \exp\left(-\frac{2BC_N}{\sqrt{N}\lambda m_N}\right) \\ &= \exp\left(-\frac{2B}{\sqrt{N}-1}\right) \end{aligned}$$

where the last equality arises from the heavy traffic relation $\lambda m_N/C_N = 1 - 1/\sqrt{N}$. Finally, using the relations $d_N = B/C_N$ and $m_N = Ne^{-d_N}$, we have the proof. ■

In the case where $\lim_{N \rightarrow \infty} \sqrt{N} d_N e^{-d_N} < \infty$, the delay target is growing large as N grows. While this is a relatively less interesting regime, for completeness we analyze this setting using a general result from [21] for $M/D/1$ systems.

Proposition 8: Let $\rho = \lambda m_N / C_N$. Then, for $d_N \rightarrow \infty$

$$\mathbb{P}(D > d_N) \approx \alpha e^{-\gamma d_N}$$

where γ is the real solution of

$$\frac{\rho(\lambda - \gamma) + \gamma - \gamma e^{(\lambda - \gamma)(m_N / C_N)}}{(m_N / C_N)(\lambda - \gamma)(\lambda - \gamma e^{(\lambda - \gamma)(m_N / C_N)})} = \frac{1}{\rho}$$

and

$$\alpha = \frac{(1 - \rho)(\lambda - \gamma)}{2\lambda(1 - \rho) - \gamma\rho(2 - \rho)}.$$

Note that for $d_N \rightarrow \infty$, we automatically have from (13) that the $M/D/1$ approximation error in per-user average delay is $\Theta(1)$. We define \bar{C}_N as the value of C_N that satisfies Proposition 8 with $\lim_{N \rightarrow \infty} \mathbb{P}(D > d) = 0$. The proposition says that if the desired delay d_N is large, we must choose $\bar{C}_N \geq \lambda m_N$ in order to satisfy the delay constraint. Note that $m_N = N e^{-d_N}$, so if d_L and d_S are large and small target delays, respectively, with $d_L \gg d_S$, and C_L and C_S are the capacities required to attain these delay targets, then $\lambda N e^{-d_S} \approx C_S < \lambda N e^{-d_L} \leq C_L$.

C. Delay Incurred Between CD and P2P

As mentioned above, users who arrive after the m_N^{th} user but before he has been served, all have to wait until he has been served in order to proceed to the P2P phase. The total amount of time spent in this phase is indicated by the cross-hatched region in Fig. 6. This wait is artificial in the sense that one would not use such a binary service rule in a practical system.

However, we can easily show that this waiting time is small on average by the following argument. Since we know that the m_N^{th} interested user waits at most for time d_N with high probability, the time spent by the system (for each file) in the phase between CD and P2P is at most d_N as well. The arrival rate of users during this interval is at most $m_N e^t$, which means that the cumulative number of users varies as $m_N e^t$. When integrated between time 0 to d_N , this yields a total delay of $m_N e^{d_N} - m_N$, which is the area of the cross-hatched region in Fig. 6. However, the number of users who arrive in this interval is $m_N e^{d_N}$, and thus the delay incurred per user is $\Theta(1)$. Thus, the technical artifice of splitting the file dissemination into two distinct phases—CD and P2P—does not significantly affect the average delay.

D. Provisioning for an Average Delay Target

We may combine Propositions 6–8, along with the observation of the preceding subsection, to yield the following theorem.

Theorem 9: For large N , the installed capacity C_N required to satisfy an average per-user delay target d_N while using the hybrid CD-P2P scheme is

$$C_N = \begin{cases} \lambda N e^{-d_N}, & \text{if } \lim_{N \rightarrow \infty} \sqrt{N} d_N e^{-d_N} = \infty \\ \bar{C}_N, & \text{otherwise.} \end{cases}$$

The theorem determines the provisioning required for a fixed per-user average delay target, given an exogenously specified arrival rate of files.

We illustrate the main insights of this section using a simple example. Suppose that we would like to serve a file arrival rate of λ files per unit time, with a per-user average delay d_N per file of $O(\ln \ln N)$. From Proposition 6, each user needs to contribute capacity $\xi_N \in \Omega(\lambda \ln \ln N)$. Also, since $\lim_{N \rightarrow \infty} \sqrt{N} \ln(\ln(N)) e^{-\ln(\ln(N))} = \infty$, Theorem 9 implies that $C_N = \lambda N / \ln N$ is a sufficient transit capacity required to attain the average delay target.

V. SIMULATIONS: PEER DEPARTURES AND P2P EFFICIENCY

Our scaling results provided an analytical characterization of the amount by which a hybrid system can outperform either P2P or CD dissemination alone. In this sections, we employ simulations to investigate two modifications to our basic model: first, the inclusion of peer departures; and second, the effects of a more efficient P2P system design.

We start by considering the effects of peer departures. Our objective is to show that the impact on average delay and server utilization is limited and relatively straightforward to account for. In our experiments, we consider a single file, and a total population of $N = 10000$ users. The capacity that we provision is $C = N / \ln N = 1085$ users per unit time. Also, as in Section III-E, we allow both the server and P2P to be used simultaneously, with the server being used to boost in the latter part if the delay is over 2 units.

We may obtain an idea of the dynamics by studying the modifications to the differential equations describing the P2P system to account for departures. We assume that some fraction α/N of users who possess the file depart per unit time. Thus, we have a new variable $R(t)$, which is the cumulative number of users who possess the file but have departed. $P(t)$ now refers to users who possess the file and are still in the system. We replace the expression (8) with the following:

$$\frac{dP(t)}{dt} = \eta \frac{I(t) - R(t) - P(t)}{N - R(t)} P(t) - \frac{dR(t)}{dt} \quad (15)$$

$$\frac{dR(t)}{dt} = \frac{\alpha P(t)}{N}. \quad (16)$$

Notice that in the modified system, peers search for content only among users currently in the system, which translates into the term $N - R(t)$ in the denominator of the expression above. (In the second part of this section, we consider a more efficient P2P search model.)

We simulate the stochastic system using the same method described in Section III-E (i.e., with demand following a stochastic Bass diffusion, random peer selection, and a fixed capacity server) with $\eta = 1$, and vary α from 0 to $0.9N$: i.e., in the extreme case we allow 90% of the users who possess the file to leave per unit time. These users are picked randomly from those who possess the file. Note that since the server is always present, all users would eventually be served in spite of departures.

Fig. 7(a) shows the average per-user delay as a function of α . For $\alpha = 0$, the average delay is about 0.7 as expected. As α increases, the average delay rises approximately linearly. As α

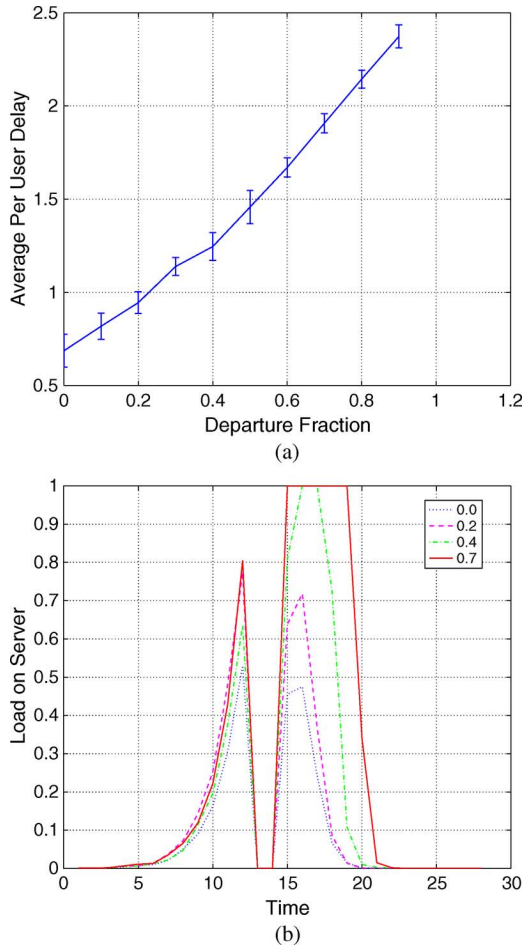


Fig. 7. (a) Average delay as a function of the fraction of departing users. The delay performance is close to optimal even for $\alpha = 0.3N$. (b) Server load and the utilization of the server over time, for different fractions of departing users (as depicted in the legend). Note that even for 20%–40% departure rates, the server utilization remains similar to the profile achieved with no departures at all.

gets close to 1, the system behaves like a pure CD system, with appropriate delays.

However, the increase in average delay can be somewhat misleading: Even though the fraction of users departing increases, it may not be the case that a significant additional assistance is needed from the server. The intuition is that if a large number of users are departing, then it must mean that a large number of users have been served as well, in which case P2P dissemination must have been reasonably efficient (for moderate departure rates).

Indeed, we can see this effect in the server utilization trajectories (as a function of time) shown in Fig. 7(b). The initial spike is the server boost common to all trajectories, and the second is the extra boost needed to account for departures. Note that only when the departure rate reaches 40% is the server capacity beginning to approach saturation. For departure rates below this point, the server capacity we provisioned based on a model with no departures is adequate—the system is instead constrained by performance of the P2P phase in this region.

Our second set of simulations account for the fact that the P2P system we analyzed is inefficient due to random peer selec-

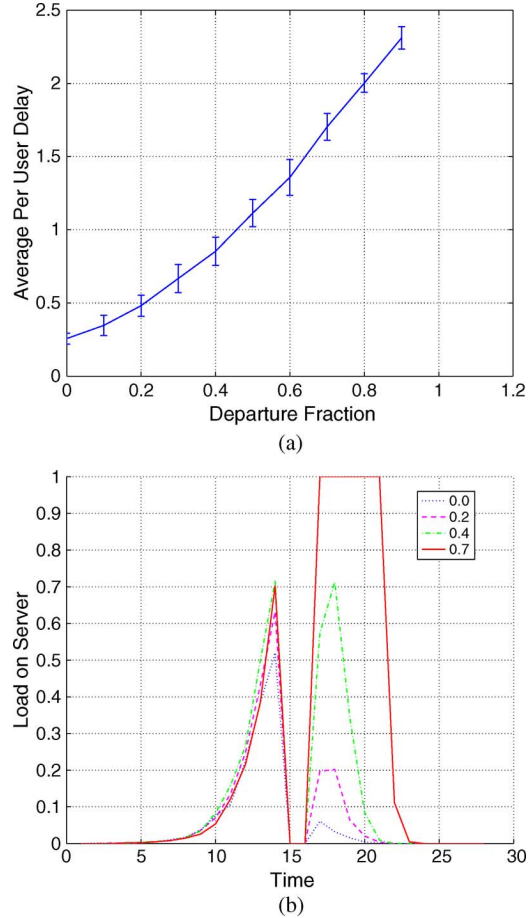


Fig. 8. Simulation of a more efficient P2P system. (a) Average delay as a function of the fraction of departing user. (b) Trajectories of the server load for different departure rates.

tion. Even with such a worst-case model of P2P dissemination, we have found that significant gains are possible using a hybrid scheme. However, one way to improve performance and to make the model more realistic is to limit the peer choice only to interested users. Thus, when the user connects to the P2P system, he is given a list of interested users (in the system) that is constantly updated. He may contact any of these users to try to obtain the file. The (fluid) dynamics of P2P would now be

$$\frac{dP(t)}{dt} = \eta \frac{I(t) - R(t) - P(t)}{I(t) - R(t)} P(t) - \frac{dR(t)}{dt} \quad (17)$$

$$\frac{dR(t)}{dt} = \frac{\alpha P(t)}{N} \quad (18)$$

where the change is that $I(t) - R(t)$ appears in the denominator in (17).

As before, we simulate the stochastic system with $\eta = 1$ and vary α from 0 to $0.9N$. Thus, users who possess the file leave randomly with fraction of such leaving users at each time being α/N . Fig. 8(a) shows the average per-user delay as a function of α . Although the average delay is less than in the inefficient P2P system shown in Fig. 7(a), as α increases, the performance is close to identical.

The server utilization trajectories are shown in Fig. 8(b). Here, we see that as compared to Fig. 7(b), the server utilization

in the initial phase is almost identical. This observation implies that even with higher efficiency, a P2P system alone is not sufficient; server boosting is essential. The amount of boosting required as peers depart is less with higher efficiency (as it should be), but again, as the departure fraction α/N increases, the performance is the same.

Our observations suggest that: 1) a small boost from the CD (server) system is sufficient to account for substantial rates of peer departures; and 2) server boosting in the initial phase is essential even with high-efficiency P2P systems. Regarding departures, another possibility is to incentivize users to stay long enough that the rate of departures is small. Such incentives may be provided by trading a local currency in exchange for files, as in the system analyzed in [22].

VI. CONCLUSION

This paper has studied the delay performance of a variety of mechanisms for distribution of stored content. We modeled demand for a file as a function of time using the Bass diffusion model, representing the temporal evolution of popular files, that the model represents popular files, and unpopular ones would never do well in a P2P scenario—they must be handled almost entirely by the central server. We studied three distribution methods—centralized, peer-to-peer, and hybrid distribution. We calculated the average per-user delay in each setting and explicitly characterized the extent to which the hybrid approach reduces the capacity required to achieve a target delay, in both a single-file model and a model with files arriving over time.

A significant open question concerns the design of algorithms that could be used in real CD-P2P hybrid content distribution systems. Our analysis is deliberately from a performance analytic view: Our results provide guidance to understand how server provisioning affects performance of hybrid content distribution systems. However, in practice, online algorithms are required to determine allocation of scarce server capacity across multiple pieces of content, without knowing the eventual total realized demand that can be achieved. The eventual goal is the practical design of such schemes, to achieve the performance outlined in this paper.

APPENDIX A

INTEREST EVOLUTION ON GENERAL SOCIAL GRAPHS

The model we consider in (1) assumes that the social graph of users is fully connected, which might not be accurate in reality. However, a similar evolution of interested users is observed even when the social graph is not fully connected. In this case, consider the following quantities:

- N_k , the population of users of degree k ;
- $I_k(t)$, the number of users of degree k interested at time t ;
- γ , the rate at which an interested user contacts his neighbors.

An approximate expression for the evolution of interest of users of degree k is then given by

$$\frac{dI_k(t)}{dt} = k(N_k - I_k(t)) \sum_j q_j \frac{I_j(t)}{N_j} \frac{\gamma}{j}. \quad (19)$$

Equation (19) can be derived via a “mean-field” analysis as follows. The number of uninterested users of degree k at time t is $N_k - I_k(t)$, and each such user can potentially be interested by any of his k neighbors. The probability that a given link points to a user of degree j is $q_j = j p_j / \sum_\ell \ell p_\ell$, where $(p_\ell, \ell \geq 1)$ is the degree distribution of the graph. The probability that such a user of degree j is interested is $I_j(t)/N_j$. Also, such a user of degree j tries to interest his neighbors uniformly and at random with rate γ , resulting in any given neighbor being selected with probability $1/j$. Define

$$\beta_j = \frac{N_j}{N}; \quad \zeta_j(t) = \frac{I_j(t)}{I(t)}; \quad \delta_j(t) = \frac{N_j - I_j(t)}{N - I(t)}.$$

Summing over all k , we have

$$\frac{dI(t)}{dt} = K(t) (N - I(t)) \frac{I(t)}{N} \quad (20)$$

where

$$K(t) = \sum_k k \delta_k(t) \left(\sum_j \frac{\lambda q_j \zeta_j(t)}{j \beta_j} \right). \quad (21)$$

The differential (20) is similar to (1) and can be solved in a like manner, yielding the solution

$$I(t) = \frac{I(0) e^{\int K(t) dt}}{K_0 - (I(0)/N) (K_0 - e^{\int K(t) dt})} \quad (22)$$

where $K_0 = e^{\int K(t) dt}|_{t=0}$. The above expression holds for any degree distribution on the graph. For example, in the case of a d -regular graph, $K(t)$ would be a constant, and (22), upon normalizing time, would look identical to (2). In the case of other graphs, $K(t)$ might be decreasing with time resulting in a “time stretched” version of the logistic function. In this paper, we restrict ourselves to demand that evolves according to the logistic function, which acts as an upper bound on all other viral demand that could arise on different graphs.

APPENDIX B

A NOTE ON STREAMING OF STORED CONTENT

Consider a single video or music file that is broken up into chunks. Let the user interest in the file as a whole follow a Bass diffusion. It is clear that since all users start with the first chunk, the demand for the first chunk is a Bass diffusion with N potentially interested users. The second chunk arrives deterministically after the first. Assuming that, in an order sense, the number of interested users remains the same over all the chunks, the demand looks much like Fig. 6, only instead of files, we now have chunks and they all arrive *deterministically* at rate λ .

Suppose we use the same hybrid approach to serving these users, and we have a per-user average delay target of d_N . If the number of users served per chunk in the CD phase of each chunk is m_N , then from Proposition 6, we choose $m_N = N e^{-d_N}$. We now check whether choosing a capacity of λm_N for the CD phase is sufficient to guarantee an average delay d_N . An upper

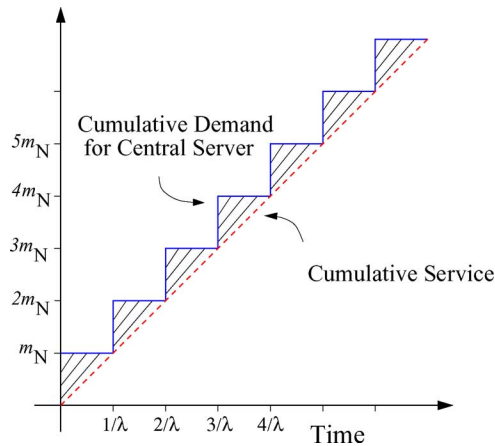


Fig. 9. An upper bound on the delay experienced in the CD phase with deterministic arrivals of chunks. The cumulative arrivals appear as a staircase as each chunk is assumed to bring m_N users instantaneously upon arrival.

bound on the cumulative demand and service curves (for the CD phase of each file) appear as illustrated in Fig. 9.

Consider a time interval N/λ . Then, the area under the demand curve can be calculated as

$$\sum_{i=0}^{N-1} m_N \frac{N-i}{\lambda} = \frac{m_N N(N+1)}{2\lambda}$$

whereas the area under the service curve is simply $(C_N N^2)/(2\lambda^2)$. With $C_N = m_N \lambda$, this is $(m_N N^2)/(2\lambda)$. Subtracting the two and dividing by the total number of users Nm_N who arrive in the time interval N/λ , we obtain the average delay per user to be $1/(2\lambda)$. Noting that the average delay requirement has to be $d_N \geq 1/\lambda$ (since the chunks arrive at spacing of $1/\lambda$), we see that provisioning for stability automatically ensures that the P2P delay is dominant. Thus, a capacity of $\lambda N e^{-d_N}$ would suffice to provide an average delay of $d_N \geq 1/\lambda$.

REFERENCES

- [1] W. B. Norton, "Internet video: The next wave of massive disruption to the U.S. peering ecosystem," Equinix, White Paper, 2007 [Online]. Available: <http://www.equinix.com>
- [2] F. M. Bass, "A new product growth model for consumer durables," *Manage. Sci.*, vol. 15, pp. 215–227, 1969.
- [3] G. Veciana and X. Yang, "Fairness, incentives and performance in peer-to-peer networks," presented at the 41st Annu. Allerton Conf. Control, Commun., Monticello, IL, Oct. 2003.
- [4] D. Qiu and R. Srikant, "Modeling and performance analysis of BitTorrent-like peer-to-peer networks," in *Proc. ACM SIGCOMM*, Portland, OR, Aug. 2004, pp. 367–378.
- [5] M. Vojnovic and L. Massoulié, "Coupon replication systems," *IEEE/ACM Trans. Netw.*, vol. 16, no. 3, pp. 603–616, Jun. 2008.
- [6] J. Wang, C. Yeo, V. Prabhakaran, and K. Ramchandran, "On the role of helpers in peer-to-peer file download systems: Design, analysis and simulation," presented at the IPTPS, Bellevue, WA, Feb. 2007.
- [7] S. Sanghavi, B. Hajek, and L. Massoulié, "Gossiping with multiple messages," in *Proc. IEEE INFOCOM*, May 2007, pp. 2135–2143.

- [8] K. Leibnitz, T. Hoßfeld, N. Wakamiya, and M. Murata, "Peer-to-peer vs. client/server: Reliability and efficiency of a content distribution service," in *Proc. 20th Int. Teletraffic Congress*, Ottawa, ON, Canada, Jun. 2007, pp. 1161–1172.
- [9] V. N. Padmanabhan, H. J. Wang, P. A. Chou, and K. Sripanidkulchai, "Distributing streaming media content using cooperative networking," in *Proc. NOSSDAV*, Miami, FL, May 2002, pp. 177–186.
- [10] S. Shakkottai and R. Johari, "Content distribution on the internet: Peer to peer vs. client-server," presented at the Allerton Conf. Control, Commun. Comput., 2007.
- [11] E. Setton and J. Apostolopoulos, "Towards quality of service for peer-to-peer video multicast," in *Proc. IEEE ICIP*, San Antonio, TX, Sep. 2007, vol. 5, pp. 81–84.
- [12] S. Liu, R. Zhang-Shen, W. Jiang, J. Rexford, and M. Chiang, "Performance bounds for peer-assisted live streaming," in *Proc. ACM SIGMETRICS*, Jun. 2008, pp. 313–324.
- [13] M. Chen, M. Ponec, S. Sengupta, J. Li, and P. A. Chou, "Utility maximization in peer-to-peer systems," in *Proc. ACM SIGMETRICS*, Jun. 2008, pp. 169–180.
- [14] A. Griliches, "Hybrid corn and the economics of innovation," *Science*, vol. 132, pp. 275–280, 1960.
- [15] G. Moore, *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers*, Rev. ed. New York: Harper-Business, 1999.
- [16] M. J. Freedman, E. Freudenthal, and D. Mazières, "Democratizing content publication with Coral," in *Proc. NSDI*, Mar. 2004, vol. 1, p. 18.
- [17] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM Internet Meas. Conf.*, San Diego, CA, Oct. 2007, p. 1–14.
- [18] D. J. Daley and J. Gani, *Epidemic Modelling: An Introduction*. Cambridge, UK: Cambridge Univ. Press, 1999.
- [19] D. Bertsekas and R. Gallager, *Data Networks*. Boston, MA: Longman Higher Education, 1987.
- [20] M. Harrison, *Brownian Motion and Stochastic Flow Systems*. New York: Wiley, 1985.
- [21] R. Egorova, B. Zwart, and O. Boxma, "Sojourn time tails in the M/D/1 processor sharing queue," *Probab. Eng. Inf. Sci.*, vol. 20, no. 3, pp. 429–446, 2006.
- [22] C. Aperjis and R. Johari, "A peer-to-peer system as an exchange economy," in *Proc. Workshop Game Theory Netw.*, Pisa, Italy, Oct. 2006, Article no. 10.



Srinivas Shakkottai (S'00–M'08) received the Bachelor of Engineering degree in electronics and communication engineering from Bangalore University, Bangalore, India, in 2001, and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 2003 and 2007, respectively.

He was a Post-Doctoral Scholar at Stanford University, Stanford, CA, until December 2007, and is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station. His research interests include peer-to-peer systems, pricing approaches to resource allocation, game theory, congestion control, and the measurement and analysis of Internet data.



Ramesh Johari (M '05) received the A.B. degree in mathematics from Harvard University, Cambridge, MA, in 1998, the Certificate of Advanced Study in Mathematics from the University of Cambridge, Cambridge, U.K., in 1999, and the Ph.D. degree in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, in 2004.

He is currently an Assistant Professor of management science and engineering, and by courtesy, electrical engineering, at Stanford University, Stanford, CA. His research interests include game theory, optimization, and competition and cooperation in networked systems.