Resource Sharing for Book-Ahead and Instantaneous-Request Calls using a CLT Approximation

R. Srikant *

Coordinated Science Laboratory and the Department of General Engineering, University of Illinois, 1308 W. Main Street, Urbana, IL 61801; rsrikant@uiuc.edu

Ward Whitt

AT&T Laboratories. Address: AT&T Labs, Room A117, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971; wow@research.att.com

This paper extends the admission control algorithm for book-ahead and instantaneous-request calls proposed by Greenberg, Srikant and Whitt (1997) to cover multiple classes of instantaneous-request calls, each with their own traffic characteristics and their own performance requirements. As before, book-ahead calls specify their starting and finishing times, and are assumed to book far ahead relative to the holding times of the instantaneous-request calls. The book-ahead calls may be constrained by an upper-limit on the capacity that can be reserved for them. Instantaneous-request calls are admitted if the probability of interruption (or some other form of service degradation in response to the conflict) for that call is below a threshold, but now this threshold can be class-dependent, and now the interrupt probability is calculated by a normal approximation based on the central limit theorem. Simulation experiments show that the normal approximation performs as well as the previous detailed calculation in single-class examples, and that the normal approximation can be applied to multi-class examples.

Keywords: book-ahead calls, advance reservation, admission control, integrated-services networks, multimedia, video teleconferencing, loss networks, grades of service.

AMS Subject classification: Primary 65B15, 65B99; Secondary 65C05

1. Introduction and summary

This paper is a sequel to Greenberg, Srikant and Whitt [13] in which we studied resource sharing in a telecommunications system when some customers are allowed to book ahead their service requests, i.e., make advanced reservations. In particular, in [13] we developed an admission control algorithm for *book-*

* Research supported by an NSF CAREER Award NCR 9701525

ahead (BA) *calls* that specify their starting and finishing times in advance and *immediate-request* (IR) *calls* that start to receive service immediately, if admitted, and have unspecified holding times governed by a probability distribution. We think of the IR calls as ordinary voice calls and the BA calls as large-bandwidth calls, such as video conference calls, possibly using multiple resources as with multicast, that might well need advance reservation in order to get access (avoid very high blocking probabilities). After allowing the BA calls to book ahead (usually subject to some constraints, such as a minimum book-ahead time and an upper-limit on the total capacity that can be reserved), we want to provide as good service as possible to the IR calls. There are also other potential applications, e.g., related to the Internet resource reservation protocol (RSVP); see [13] for additional references. See Coffman, Jelenkovic and Poonen [6] for a recent theoretical study of booking ahead for a resource with capacity for a single customer.

Our main idea in [13] was to achieve more efficient use of limited resources than can be achieved by strict partitioning by allowing a small probability of conflict between an admitted IR call of uncertain duration and BA calls with scheduled starting times. We think of the BA calls tending to book relatively far ahead (in the time scale of IR call holding times) and the most recently arriving IR calls being interrupted if necessary, although the actual resolution of the conflict could be different. In many applications, it will not actually be necessary to interrupt calls. Instead, the bandwidth or quality of service may be temporarily reduced, e.g., by bit dropping or coarser encoding in video. The admission control algorithm based on interruptions can also be applied with other forms of service degradation. Then the interruption probability should be interpreted as the probability of conflict.

A main conclusion of [13] was that allowing occasional service interruptions or service degradation can yield significantly greater resource utilization and revenue than admission control schemes that do not allow them. It was also observed that a relatively simple nearly-decomposable Markov-chain (ND-MC) algorithm provides a useful approximation for long-run average performance when BA calls book for ahead and have relatively long holding times, and provides an upper bound on revenue more generally. However, most of [13] was devoted to the design and performance of the admission control algorithm. That will be our focus here as well.

Thinking of the BA calls as large-bandwidth calls, such as video conference calls, and the IR calls as ordinary voice calls, in [13] we let the BA calls have very general bandwidth requirements, book-ahead times (time until starting to receive service) and holding times (service durations), but we considered only a single class of IR calls with unit bandwidth requirements and a common holding-time distribution. We let BA calls be admitted subject to an upper limit on the reserved bandwidth for BA calls and possibly a minimum book-ahead time. We let each successive IR call be admitted if the probability that it will be interrupted

is below a specified threshold. In [13] our admission control algorithm strongly exploited the fact that there was only a single IR class with common bandwidth requirements.

The main purpose of this paper is to extend our previous admission control algorithm to cover multiple classes of IR calls, with possibly different bandwidth requirements, holding-time distributions and interrupt-probability thresholds. To achieve this goal, we replace our previous interrupt probability calculation by a normal approximation based on the central limit theorem (CLT). We show that the normal approximation is effective, first, by showing that it does as well as our previous algorithm in the previous single-IR-class simulation experiments and, second, by showing that it also applies to treat multiple-IR-class examples.

The new CLT framework also allows us to exploit information about the IR calls in progress. In particular, for non-exponential holding-time distributions, we can exploit the ages (elapsed holding times) of IR calls in progress to compute the conditional distribution of the remaining holding time. In [13] we partially exploited the holding-time distribution by using the equilibrium-excess distribution for all IR calls in progress, but we did not exploit the ages.

The CLT-based algorithm for multiple IR classes proposed here is an alternative to an effective-bandwidth (EB) or large-deviations (LD) approach developed by Wischik and Greenberg [18]. The EB approach is an interesting extension of previous EB analysis because it includes a new time dimension. The EB approach also applies to multiple IR classes, allowing the exploitation of ages, and may have useful application, but preliminary studies indicated that it was less effective than the detailed interrupt probability calculation for scenarios as considered in [13]. We do not make direct comparison with the EB approach here, but since the CLT approach here closely matches the previous approximation in [13], we conclude that the CLT approach shares the advantages of the detailed calculation in [13] for the kinds of scenarios considered in [13].

We also think that CLT asymptotics tend to be more appropriate than LD asymptotics in this setting, because we are likely to be computing larger probabilities. This is so because we are basing our admission decision on an interrupt probability threshold that applies to an individual IR call. A key point is that the long-run interrupt probability, i.e., the long-run proportion of admitted IR calls that are interrupted, will tend to be much smaller than this threshold. For example, the interrupt probability threshold might be 10^{-2}, while the long-run interrupt probability might be 10^{-4}; then it is the larger probability 10^{-2} that we want to compute accurately. It is well known that the CLT tends to be more appropriate for larger tail probabilities. We make a specific numerical comparison here in Section 7.

The point just made also has implications for applying the admission control procedure. In applications, we directly set the interrupt probability thresholds, but our ultimate goal may be to achieve a certain long-run interrupt probability. For that purpose, we need to apply simulations or system measurements to de-

termine the interrupt probability threshold that meets the goal. As in [13], in our experiments we observe a nearly linear relation between the interrupt probability threshold and the long-run interrupt probability.

It is important to note that performance of the admission control algorithm for IR calls depends on both the long-run blocking probability and the long-run interrupt probability for IR calls. As we decrease the interrupt probability threshold, the long-run interrupt probability will decrease and the long-run blocking probability will increase, so that we obtain a curve of interrupt probability and blocking probability pairs, as shown in figures later in this paper. In [13] we demonstrated that some approximate interrupt probability calculations are better than others because the whole curve is lower (closer to the origin in the positive quadrant). Here we show that the CLT approximation is as good as the previous detailed approximation by showing that the curves tend to fall on top of each other.

The comparison of different algorithms raises the question of optimality. As noted in [13], the admission control problem can be formulated with appropriate costs and rewards as a Markov decision problem (MDP). However, the states are complicated, involving the BA reservation profiles. As a consequence, admission control based solely on an interrupt probability calculation is typically not optimal, because it fails to account for the impact of admission upon future calls. (We elaborate in Section 6.) Nevertheless, the proposed admission control algorithm seems to be effective.

In addition to near optimality, we also require that the computation be fast, because it must be performed in real time, at each successive IR call arrival. Moreover, the BA call profile must be updated as BA calls arrive and depart. The simulation experiments demonstrate that the algorithm is sufficiently fast. In practice, it may be convenient to turn off the algorithm when the risk of conflict is negligible, e.g., when there is no chance of interruption, and then turn it on again when the interrupt probability calculation is needed. This can be done by simply monitoring the total bandwidth being used and reserved, assuming a fixed bandwidth allocation for each call.

In [13] we considered two cases for the IR holding-time distribution: (1) an exponential distribution and (2) a general distribution. The exponential case is easier because the remaining holding times of all calls in progress also have the same exponential distributions, whether or not we condition upon the ages, by the lack-of-memory property. However, even for the exponential IR holding-time distribution, the interrupt probability is difficult to compute, because the profile of reserved BA calls may be quite complicated, involving several different potential interruption times. For our interrupt probability calculation, we assume that the most recent arrival will be interrupted. Assuming the BA calls book relatively far ahead, this means that the interrupt probability calculation depends only on the BA calls in service or already reserved and the IR calls in service; i.e., at each IR arrival epoch it is not necessary to consider future service requests

of any kind. The first potential interruption time, if any, is the time that an interruption would occur if the candidate IR call is admitted and none of the IR calls in progress (including the newly admitted call) complete service. Subsequent times at which the BA call profile (reservation level) increase constitute additional potential interruption times for the admitted IR call.

The exact interrupt probability given a set of potential interruption times $\{T_1, \ldots, T_k\}$ was displayed in (5.2) of [13], but it is quite complicated because it involves the probability distribution of the vector giving the number of IR service completions in each of the intervals $(T_{i-1}, T_i]$, $1 \leq i \leq k$. Hence, in [13] we considered approximations for the interrupt probability, the most promising being the sum of the probabilities of an interruption at a single time over successive potential interruption times. The idea is that one term is likely to dominate the sum, but it might not be the first. It is clear that the sum of the interrupt probabilities is an upper bound on the exact interrupt probability, but the sum appears to be a good approximation for the exact interrupt probability. To see why this should be so, think of a fixed BA call profile and successive IR arrivals. The IR calls will be admitted until there is an initial conflict. If the IR calls require less bandwidth than BA calls, then the impact of admitting one IR call should be relatively small, so that this initial conflict, when there is one, is likely to occur at only one potential interruption time.

To substantiate the sum approximation, we now also consider the maximum interrupt probability over all potential interruption times, which is a lower bound on the exact interrupt probability. We report results showing that the maximum and sum bounds perform similarly, showing that both are good approximations for the exact interrupt probability. We use the sum because it is conservative.

A word of caution is appropriate, however, because it is possible to construct scenarios in which the max and sum bounds differ greatly. For (an unrealistic) example, if the BA calls require very small bandwidths and have very short holding times, then there can be an enormous number of potential interruption times. Thus, in applications it may be desirable, at least occasionally, to apply both the max and sum calculations to verify that they are good approximations for the exact interrupt probability.

At this point, it is appropriate to point out that an attractive feature of the EB approach in [18] is that the bandwidth requirements of each call over time are summarized by a deterministic function. In that framework, a new call is not admitted if the total (deterministic) required (effective) bandwidth in the future ever exceeds capacity. Thus, the interrupt probability calculation is avoided. We think that an explicit calculation of the interrupt probability should be preferable, provided that the computation can indeed be done sufficiently, accurately (allowing for approximation) and quickly. Otherwise, the EB approach is a viable alternative.

Here is how the rest of this paper is organized. In Section 2 we specify the new CLT-based interrupt probability calculation. In Section 3 we discuss simula-

tion experiments. In Section 4 we discuss additional ways to provide appropriate grades of service to different IR classes, to use in addition to the book-ahead feature. In Section 5 we discuss how appropriate minimum book-ahead times for BA calls might be determined. In Section 6 we indicate that even though our admission-control policy based on interrupt probabilities seems to be effective, it is not optimal in a Markov decision process framework. In Section 7 we provide insight into CLT and LD approximations by numerically comparing their performance for binomial distributions.

2. The CLT-Based Algorithm

Suppose that there are n IR calls in progress at an arrival epoch of a new IR call. Let the IR calls in progress be represented by triples (a_i, b_i, F_i), $1 \leq i \leq n$, where a_i is the age (elapsed holding time), b_i is the constant required bandwidth and F_i is the original cumulative distribution function (cdf) of the holding time, say Z_i, for IR call i, i.e., $F_i(t) \equiv P(Z_i \leq t)$, $t \geq 0$. Let F_i^c be the associated complementary cdf (ccdf), i.e., $F_i^c(t) \equiv 1 - F_i(t)$. Let $F_i(t|a_i)$ denote the conditional cdf and $F_i^c(t|a_i)$ the associated conditional ccdf of the remaining holding time given the age, i.e.,

$$F_i^c(t|a_i) \equiv P(Z_i > t + a_i | Z_i > a_i) = F_i^c(t + a_i)/F_i^c(a_i) \ . \tag{2.1}$$

Without loss of generality, suppose that 0 is the arrival epoch of the new IR call. Let it have constant required bandwidth b_I and holding-time cdf F_I. If the new call is admitted, then the bandwidth required for IR calls at time 0 is $b \equiv b_I + \sum_{i=1}^{n} b_i$. Let C be the total available capacity. Clearly, if $b > C$, then the new IR cannot be admitted.

If $b \leq C$, then we consider the probability that the new IR call will be interrupted. As in [13], the BA calls are assumed to book relatively far ahead, announcing their holding times, so that their service starting and finishing times are known. Hence we can construct a BA call profile giving the total reserved bandwidth as a function of time in the future. The first potential interruption time T_1 is the first time that b plus the reserved BA bandwidth exceeds C. Of course, interruptions may not actually occur at time T_1 because by that time some of the IR calls in progress may have departed.

Each new successive maximum of the reserved bandwidth for BA calls, beyond its value at T_1, represents a new potential interruption time. Let these times be denoted by T_j. We may elect to consider only the first L potential interruption times for some specified L. When the possible BA bandwidth requirements are all integer multiples of a minimum bandwidth, then it is easy to compute an upper bound on the number of potential interruption times, using the upper limit for BA calls or the total capacity if there is no upper limit. In our examples the maximum number of potential interruption times is usually less than 10.

For the (exact) overall interrupt probability, denoted by p_i, we need to consider all the potential interruption times $T_1, \ldots, T_L$. Let $p_i(T_j)$ denote the interrupt probability considering only only the single time point T_j. The *sum approximation* is

$$p_i \approx p_i(\text{sum}) \equiv \sum_{j=1}^{L} p_i(T_j), \tag{2.2}$$

while the *max approximation* is

$$p_i \approx p_i(\text{max}) \equiv \max_{1 \le j \le L} p_i(T_j) \ . \tag{2.3}$$

As indicated earlier, these two approximations are bounds, i.e.,

$$p_i(\text{max}) \le p_i \le p_i(\text{sum}) \ , \tag{2.4}$$

provided that no potential interruption times have been omitted in the calculation of $p_i(\text{sum})$. The admission control algorithm admits the new call if $p_i \le p_T$, where p_T is the interrupt probability threshold (for the customer class of the newly arriving IR call).

We now approximate the probability of interruption $p_i(T_j)$ at T_j. by a normal approximation. This normal approximation is justified by the CLT for independent, non-identically distributed random variables, where the sum is not dominated by a few summands; see p. 262 of Feller [10].Given a normal approximation for $p_i(T_j)$ for each j, we obtain the CLTsum and CLTmax approximations for p_i by inserting the normal approximation for $p_i(T_j)$ into (2.3)-(2.4).

Let m_j and σ_j^2 denote the mean and variance of the total bandwidth required by all IR calls at time T_j. Then, since the presence of each IR call at time T_j is a Bernoulli random variable,

$$m_j = b_I F_I^c(T_j) + \sum_{k=1}^{n} b_k F_k^c(T_j|a_k) \tag{2.5}$$

and

$$\sigma_j^2 = b_I^2 F_I^c(T_j) F_I(T_j) + \sum_{k=1}^{n} b_k^2 F_k^c(T_j|a_k) F_k(T_j|a_k) \ . \tag{2.6}$$

Let B_j be the reserved BA bandwidth at time T_j. To express the approximate probability of interruption at time T_j alone, let $N(m, \sigma^2)$ be a random variable having the normal distribution with mean m and variance σ^2, let Φ be the cdf of $N(0,1)$, i.e., $\Phi(t) \equiv P(N(0,1) \le t)$, and let Φ^c be the associated ccdf. The approximate interrupt probability at time T_j, denoted by $\hat{p}_i(T_j)$, is then

$$\hat{p}_i(T_j) \equiv P(N(m_j, \sigma_j^2) > C - B_j) = \Phi^c\left(\frac{C - B_j - m_j}{\sigma_j}\right) \ . \tag{2.7}$$

Remark 2.1. If there are at most L potential interruption times for some finite L, and if interruptions are indeed rare events, then $p_i(\max)$ and $p_i(\text{sum})$ should be approximately equivalent. The rare event scenario naturally occurs in large systems with a large number n of small IR sources and a small number of large-bandwidth (order n) BA calls reserved at any time. We might then have for large n

$$\log p_i(T_j) \approx e^{-n\delta_j}, \quad 1 \le j \le L \ , \tag{2.8}$$

or, more precisely,

$$\frac{1}{n} \log p_i(T_j) \to -\delta_j \quad \text{as} \quad n \to \infty \ . \tag{2.9}$$

However, a standard large deviations consequence of (2.9) is that

$$\frac{1}{n} \log p_i(\max) \to - \min_{1 \le j \le L} \delta_j \quad \text{as} \quad n \to \infty \tag{2.10}$$

and

$$\frac{1}{n} \log p_i(\text{sum}) \to - \min_{1 \le j \le L} \delta_j \quad \text{as} \quad n \to \infty \ , \tag{2.11}$$

so that $p_i(\max)$ and $p_i(\text{sum})$ are asymptotically equivalent in the sense of (2.10) and (2.11). (Note that $e^{-n\delta}$ and $Le^{-n\delta}$ are equivalent in this scaling.) We give a supporting large deviations result in the appendix.

Remark 2.2. So far, we have assumed that BA calls are of known duration. However, we may not always want to require that BA calls specify their holding times. It is significant that the CLT algorithm extends if only a holding-time cdf is available for each BA call, along with the start time and bandwidth.

Given this revised information, we can act as if each BA call has no termination time, but instead have a random required bandwidth at each time, just like the IR calls in progress. We can consider all BA arrival times as potential interruption times. Instead of the known required BA bandwidth B_j at time T_j, we can compute instead the mean and variance, just as in (2.5) and (2.6) without the ages. Instead of (2.7), we now have

$$\begin{aligned} \tilde{p}_i(T_j) &\approx P(N(m_j + EB_j, \sigma_j^2 + Var\, B_j) > C) \\ &= \Phi^c \left((C - m_j - EB_j)/\sqrt{\sigma_j^2 + Var\, B_j} \right) \ . \end{aligned} \tag{2.12}$$

Remark 2.3. In this section we have assumed that all calls use constant fixed bandwidth while they are in progress. That is clearly appropriate when constant bandwidths are assigned to each call, as in circuit-switched networks. However, it is also possible to consider advanced reservation with uncertain variable bandwidth requirements for each source. It is again significant that the CLT approach can be extended to serve as an approximation in that setting too.

First, however, with variable bandwidth requirements, conflict is more likely to mean service degradation (e.g., packet loss) rather than total interruption. With that in mind, it is natural to change the control for an arriving call. We would admit the IR call if the probability of conflict at time t is less than a specified conflict threshold for all t.

Thus it is directly appropriate to use the algorithm CLTmax, except that we must consider more time points. It is natural to use a finite collection of points including BA call arrival epochs. With variable bandwidth processes, we can still use the CLT, but the formulas for the means and variances in (2.5) and (2.6) are no longer valid. For IR calls, we could use a semi-Markov bandwidth process as in Duffield and Whitt [8], which includes the familiar on-off model with general on-time and off-time distributions as a special case. Algorithms for computing the means and variances for the generalizations of (2.5) and (2.6) are given in [8].

If the bandwidth requirements of BA calls are also uncertain, then the BA bandwidth requirements can be included in the approximation; i.e., instead of (2.7), we would have (2.12).

Due to randomness in the BA call profile, either due to the uncertainty in the holding times or due to variable bandwidth requirements, it is possible for an admitted BA call to be denied service when its service is about to commence. Thus, in these cases, we also have to consider an "interrupt probability" for BA calls. Further investigation is required to fully understand the role of the various admission controls for these scenarios.

3. Simulation Experiments

We start by considering previous examples in [13] in order to see how well the CLT approximation compares to the detailed calculation, the previous independent-peaks approximation (IPA), in the single-class case. We also want to see how the maximum lower bound for the interrupt probability compares to the sum upper bound, both for IPA and the new CLT approximation.

We first reconsider the simulation experiment in Example 3 from [13]. The available capacity (number of servers) is $s = 100$. The arrival rates, required bandwidths and mean holding times for IR and BA calls, respectively, are $\lambda_I = 60$, $\lambda_B = 2$, $b_I = 1$, $b_B = 10$, $\mu_i^{-1} = \mu_B^{-1} = 1$. We assume that the arrival processes are independent Poisson processes and that the service times are independent exponential random variables. We assume that the BA bookahead times (times before starting) are all 20, so that the BA calls indeed book relatively far ahead. (Other cases were also considered; e.g., see Section 8 of [13].)

As in [13], we consider 10 different values of the IR interrupt probability threshold p_T. For each value of p_T, the simulation run length was 100,000 time units, after deleting 25 times units to eliminate transients. Thus, we simulate roughly 6 million IR arrivals and 2 million BA arrivals for each threshold value.

Note that the offered load and capacity are 80 and 100, so that the loading is normal. Without the booking ahead feature, the blocking probabilities for the BA and IR calls can be determined by the Kaufman [14] - Roberts [16] recursion or by numerical inversion [4]. Without booking ahead, the blocking probabilities for the BA and IR calls are 0.151 and 0.0112, respectively, but with booking ahead, the BA blocking probability goes way down to 0.000038 (without using any upper limit on BA calls). Clearly, the book-ahead feature greatly helps the BA calls. On the other hand, the IR calls suffer as a consequence. If we allow no possibility of interruption, then the IR blocking probability increases to 0.264.

We might well decide that the change from (0.151, 0.011) for the BA and IR blocking probabilities to (0.000038, 0.264) is giving too much advantage to BA calls at the expense of IR calls. Indeed, with our admission control scheme, the overall utilization is maximized (yielding 76.3 out of 80) by first placing an upper limit of 50 on BA calls. This upper limit causes the BA blocking probability to increase to 0.037, which is still a big improvement over the 0.151 value without booking ahead. With the upper limit of 50 imposed on BA calls, the IR blocking probability allowing no interruptions decreases to 0.19.

We obtain substantially lower IR blocking probabilities by allowing a small probability of IR call interruption. Figure 1 shows the pairs of (long-run) blocking probabilities and interrupt blocking probabilities achievable with different interrupt probability thresholds (when an upper limit of 50 has been placed on the BA calls). Figure 1 shows that, by allowing a small IR interrupt probability of 0.0002 (which is achieved by choosing an interrupt threshold of about 0.05), the IR blocking probability is reduced from 0.19 to about 0.05.

Table 1 shows IR blocking probability and interrupt probability pairs as a function of the interrupt probability threshold for the two approximations CLTmax and CLTsum. In Table 1 the blocking probability (interrupt probability) is consistently slightly smaller (larger), showing that the performance of the two approximations is very similar. (There is one exception for P_I in line 2 of the table, which we attribute to estimation error.)

Figure 1 shows four curves, one each for IPA, IPAmax, CLTsum and CLTmax. The four curves are very close. In particular, there is no clear separation between the curves comparable to the curves for the three different approximation procedures in Figure 2 of [13]. Since these curves here nearly fall on top of each other, we conclude that the four procedures perform approximately the same. Since the long-run interrupt probability is quite small in this example (being between 10^{-4} and 10^{-3}), there remains considerable variability in this example. Hence we will also support the conclusion with other examples having larger IR blocking and interrupt probabilities.

Results for such an example, corresponding to Example 2 of [13] are displayed in Figure 2. The four curves for IPA, IPAmax, CLTsum and CLTmax even more clearly fall on top of each other in this example. Moreover, this example is potentially more difficult because the reserved BA calls have a more

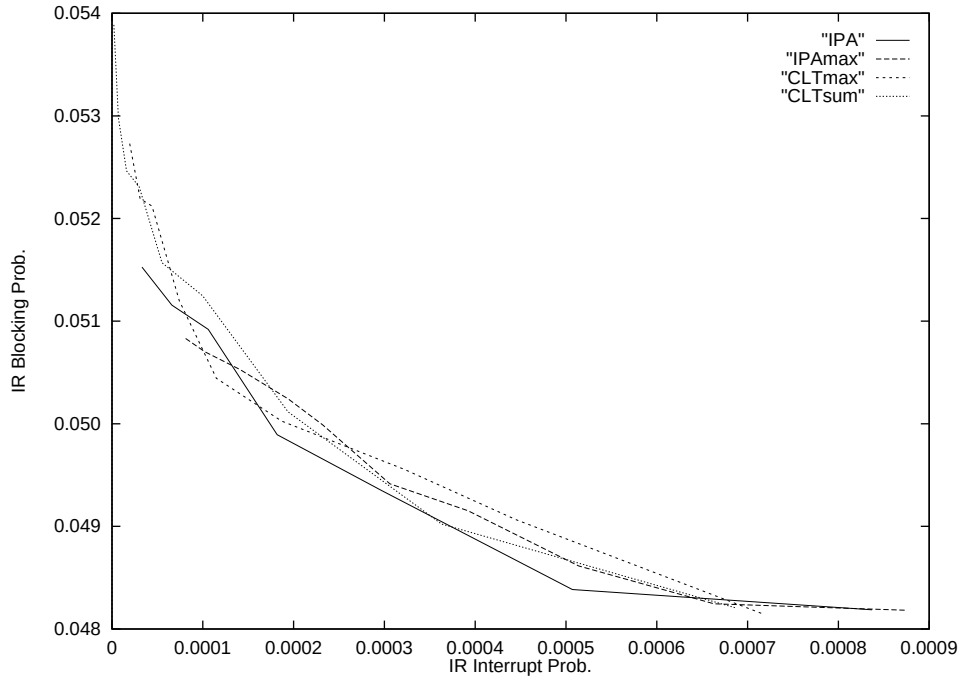

Figure 1. CLT and IPA approximations for $s = 100$, $\lambda_I = 60$, $\lambda_B = 2$, $b_I = 1$, $b_B = 10$, $\mu_I = \mu_B = 1$, UL on BA calls is 50

P_T	p_I under CLTmax	p_I under CLTsum	P_I under CLTmax	P_I under CLTsum
0.0774	0.000715	0.000686	0.048153	0.048208
0.0599	0.000445	0.000363	0.049065	0.049023
0.0464	0.000322	0.000194	0.049555	0.050115
0.0359	0.000187	0.000100	0.050026	0.051246
0.0278	0.000115	0.000055	0.050443	0.051569
0.0215	0.000074	0.000030	0.051201	0.052308
0.0167	0.000044	0.000016	0.052121	0.052464
0.0129	0.000031	0.000007	0.052193	0.052982
0.0100	0.000019	0.000002	0.052755	0.053889

Table 1
The realized interrupt probability p_I and blocking probability P_I as a function of the interrupt probability threshold P_T under CLTmax and CLTsum

variable profile, because of higher arrival and service rates and smaller bandwidths, in particular, $\lambda_B = 16$, $\mu_B = 4$ and $b = 5$ now, instead of $\lambda_B = 2$, $\mu_B = 1$ and $b = 10$ before. By allowing an IR interrupt probability of 0.1, the IR blocking probability is reduced from more than 0.9 to about 0.6.

This second example is interesting because it demonstrates the large detrimental effect booking ahead can have on IR calls. Even though the total offered

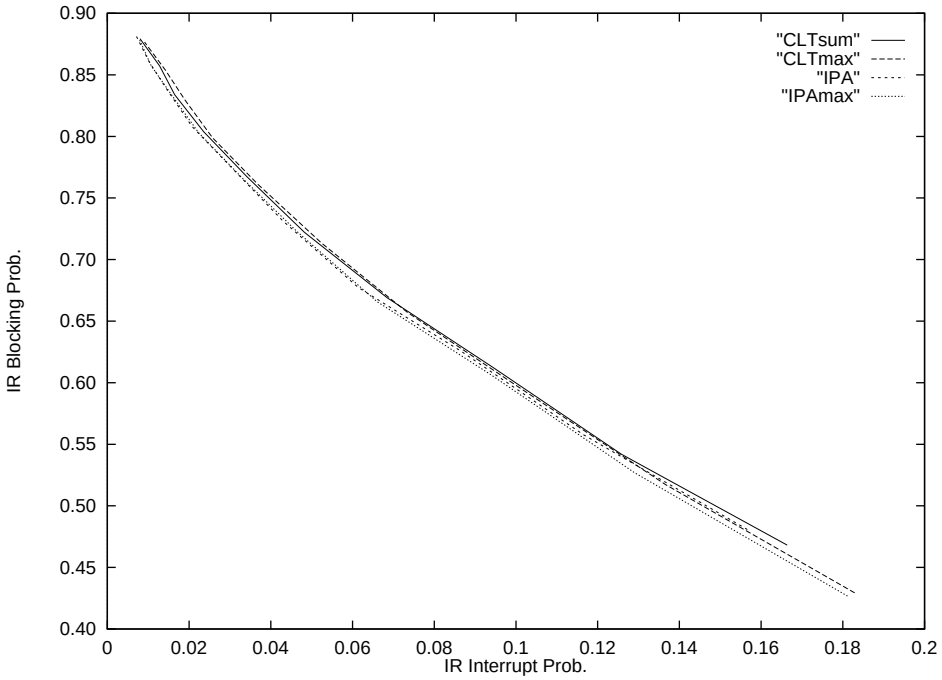

Figure 2. CLT and IPA approximations for $s = 40$, $\lambda_I = 24$, $\lambda_B = 16$, $b_I = 1$, $b_B = 5$, $\mu_I = 1$, $\mu_B = 4$, no UL on BA calls

load 34 is less than the capacity 40, the IR blocking probability can reach 0.90. This shows the importance of being able to impose upper limit on the BA calls, which has not yet been done here. Figure 3 shows the CLTsum curve when an upper limit is introduced on the BA calls. This improves both the blocking and interrupt performance of IR calls. However, the BA blocking probability increases from 0.03 with no BA upper limit, to 0.12 with the upper limit of 30. Thus, the choice of the upper limit is a trade-off between IR and BA call performance. For this example with an upper limit of 30 on BA calls, if we let the IR interrupt probability threshold be zero, the blocking probability for IR calls is 0.61, whereas from Figure 3, with an IR interrupt probability of 0.01, the IR blocking probability drops to about 0.5.

We next describe a two-IR-class example. As before, we consider independent Poisson arrival processes and exponential service times. We let the IR and BA arrival rates, mean service times and constant required bandwidths be: $\lambda_{I1} = 30$, $\lambda_{I2} = 60$, $\lambda_B = 8$, $\mu_{I1}^{-1} = 1$, $\mu_{I2}^{-1} = 1/2$, $\mu_B^{-1} = 1$, $b_{I1} = b_{I2} = 1$ and $b_B = 5$. The offered load is thus 100. We also let the capacity be 100 and assume that there is no UL on BA calls. The blocking probability of BA calls is 0.000159.

The IR class-1 and class-2 blocking probabilities and interrupt probabilities for ten different interrupt probability thresholds are displayed in Figures 4 and 5, while the combined performance is displayed in Figure 6. If there were no

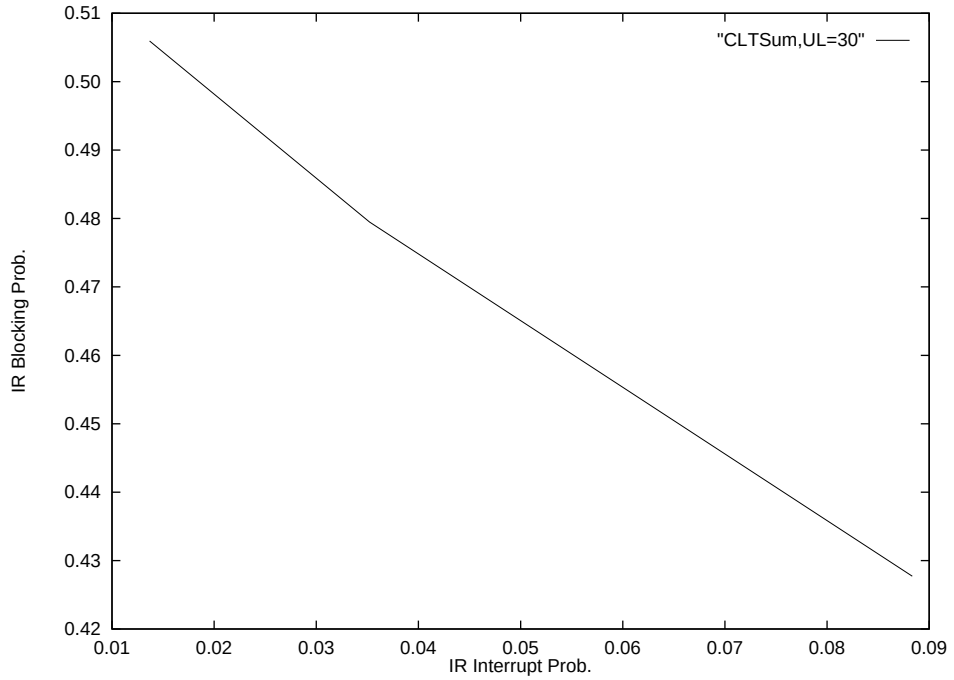

Figure 3. CLTsum algorithm for $s = 40$, $\lambda_I = 24$, $\lambda_B = 16$, $b_I = 1$, $b_B = 5$, $\mu_I = 1$, $\mu_B = 4$, UL on BA calls is 30

bookahead feature, then the two IR classes would have the same blocking probabilities because they require common (unit) bandwidths and have Poisson arrival processes. However, because they have different holding-time means, they face different interruption probabilities, and thus different performance overall. However, Figures 4 and 5 show that the differences between the two IR classes this example is not great. Class 1 does have slightly higher interrupt probabilities for common thresholds, though.

If we had allowed no interrupt probability, then the blocking probabilities of IR Classes 1 and 2 are 0.44 and 0.42, respectively, and the overall IR blocking probability is 0.42. Thus, again we see that allowing small interrupt probabilities greatly enhances the performance of the system. As in the single-IR-class examples, Figures 4-6 show that the CLTmax and CLTsum algorithms perform very similarly, indicating that both approximate the interrupt probability sufficiently well.

4. Different Grades of Service for IR Calls

In this paper we have allowed for multiple IR customer classes with their own traffic characteristics and their own interrupt probability thresholds. In this

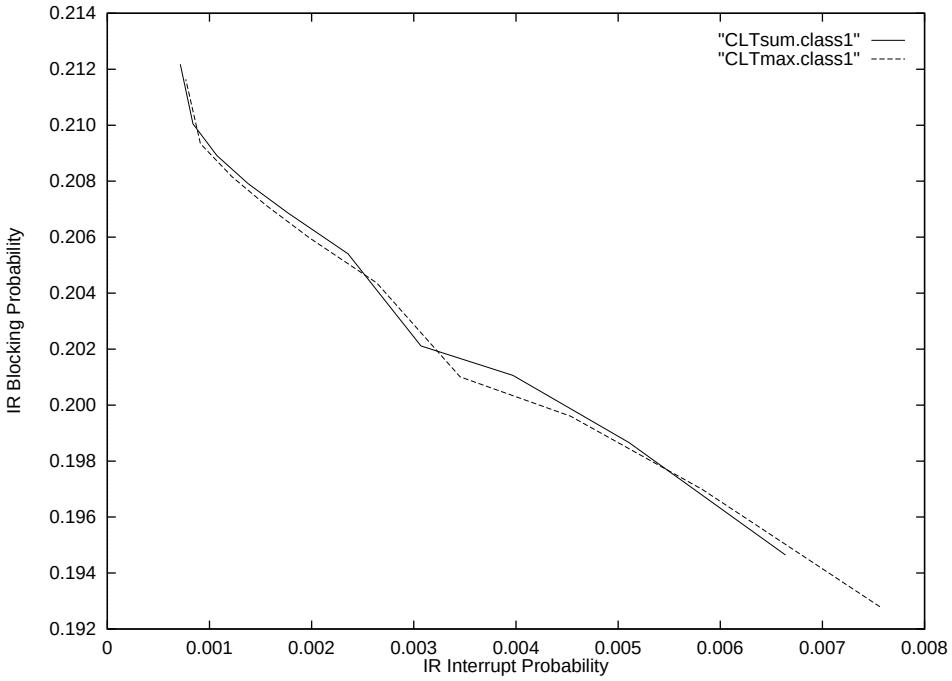

Figure 4. Plots for IR Class1 with two IR classes: $s = 100$, $\lambda_{I1} = 30$, $\lambda_{I2} = 60$, $\lambda_B = 8$, $\mu_{I1} = 1$, $\mu_{I2} = 2$, $\mu_B = 1$, $b_{I1} = b_{I2} = 1$, $b_B = 5$, no UL on BA calls

setting, we may also want to take other measures to provide the IR classes with appropriate grades of service.

As in Choudhury, Leung and Whitt [4], [5], for this purpose we propose using upper-limit (UL) and guaranteed-minimum (GM) bounds for the individual IR classes or subsets of the IR classes. Similar measures could also be taken for subclasses of BA calls, but we will not discuss that possibility.

For an IR class, a UL bound is an upper limit on the total bandwidth that can be used by calls from that class, while a GM bound is a guaranteed minimum amount of capacity that is reserved for that class. When there are only two classes, UL and GM bounds are equivalent, because a UL bound for one class is equivalent to a GM bound for the other class, but more generally that is not so.

We can use common GM and UL bounds for one class to give one class a high grade of service. Then that class has a guaranteed amount of bandwidths allocated to it, with zero probability of interruption. Thus, for this one class, the UL/GM bound is equivalent to strict partitioning. However, we could instead have a higher UL bound. Then the class could use more than its guaranteed minimum, with the understanding that calls exceeding the guaranteed minimum would be subject to interruption. Those calls exceeding the upper limit could be notified upon arrival about the possibility of interruption.

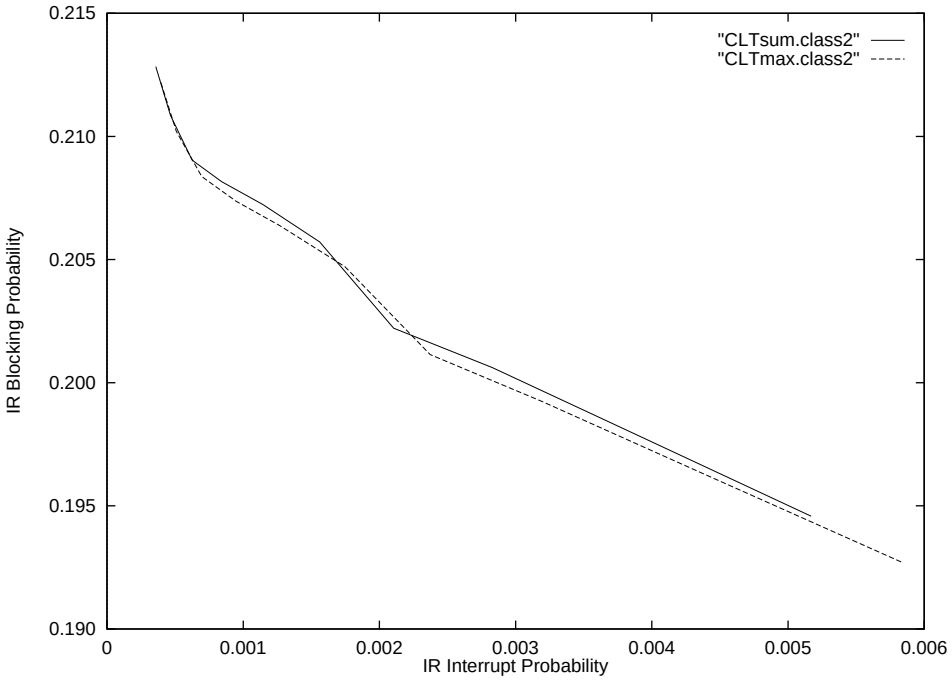

Figure 5. Plots for IR Class2 with two IR classes: $s = 100$, $\lambda_{I1} = 30$, $\lambda_{I2} = 60$, $\lambda_B = 8$, $\mu_{I1} = 1$, $\mu_{I2} = 2$, $\mu_B = 1$, $b_{I1} = b_{I2} = 1$, $b_B = 5$, no UL on BA calls

The UL and GM bounds can also extended to networks. However, the size of emerging networks necessitates the use of analytic approximation algorithms to be able to effectively evaluate the impact of these bounds on blocking probability. For this purpose, one can use the algorithm presented in [12] to compute the blocking probability given UL/GM parameters for each link and each call class in the network.

An alternative to UL and GM bounds is to use trunk reservation for each class. However, with trunk reservation, there is no computationally-efficient analytical technique to calculate the blocking probability and simulation is too slow to be useful. When the bandwidth of each class is small compared to the link capacity, various analytical approximations are available [11,1,2] to compute the blocking performance effectively. The numerical results in [12] suggest that the algorithm in [1] is the most effective of these.

It is not difficult to incorporate the UL and GM bounds or trunk reservation into the admission control algorithm, because they can be considered before calculating the interrupt probability. Upon an arrival of a new IR call, we first identify its class and then see if it can be admitted from the perspective of the UL and GM bounds or trunk reservation in effect. If it can be admitted from that perspective, we then compute the interrupt probability and see if it is below the threshold for that class. If the calculated interrupt probability is below the

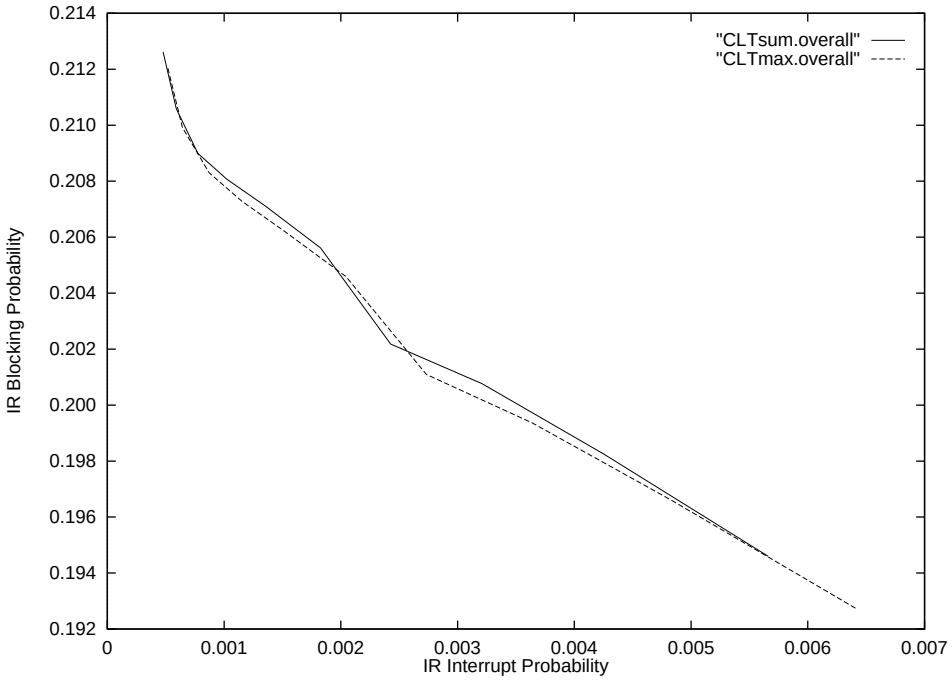

Figure 6. Plots for all IR Calls with two IR classes: $s = 100$, $\lambda_{I1} = 30$, $\lambda_{I2} = 60$, $\lambda_B = 8$, $\mu_{I1} = 1$, $\mu_{I2} = 2$, $\mu_B = 1$, $b_{I1} = b_{I2} = 1$, $b_B = 5$, no UL on BA calls

threshold, then we admit the call. In the case of UL and GM bounds, we also need to update and maintain state information upon arrival and departure of IR calls, which can be done efficiently, as indicated in [5].

5. A Minimum Book-ahead Time

We have assumed that BA calls book relatively far ahead in the time scale of IR holding times. In this section we consider more carefully what might be an appropriate minimum book-ahead time.

What we seek is the minimum book-ahead time such that BA arrivals do not significantly alter and invalidate IR interrupt probability determinations. We want the interrupt probability for each IR arrival, up to a close approximation, to depend only on the BA calls that have made reservations up to that time, and not upon later BA service requests.

What we want, then, is a time such that future IR departures are very likely to compensate for the new BA arrival. A conservative (upper bound) is the maximum of the maximum number of prevailing (independent) IR holding times (when they fill the entire system). This bound can often be well approximated by extreme value limits. For example, if the capacity is n and IR calls have unit

bandwidth requirements and exponential distributions with mean μ^{-1}, then the time T_n for all n IR calls to clear satisfies

$$P(\mu T_n - \log n \leq x) \;\rightarrow\; \exp(-e^{-x}) \quad \text{as} \quad n \rightarrow \infty \;, \tag{5.1}$$

so that $\mu T_n / \log n \rightarrow 1$ in probability as $n \rightarrow \infty$; see p. 19 of Leadbetter, Lindgren and Rootzén [15]. Hence, we might approximate T_n by $\mu^{-1} \log n$. Note that, for large n, T_n clusters quite closely around the deterministic value $\mu^{-1} \log n$.

The previous estimate may be needlessly conservative, however, because it may not be necessary to clear all IR calls. Instead it suffices to clear only enough to compensate for future BA service starts. Suppose that BA calls have an upper limit of capacity L in a total capacity C. At some instant, the system might be filled entirely with IR calls. Immediately afterwards, we may have one or more BA service requests that simultaneously make BA reservations taking the reserved BA bandwidth up to its upper limits L after a minimum book-ahead time T. In this worst case scenario, we need sufficient IR calls to depart to lower the IR bandwidth being used from C to $C - L$ by time T.

With that in mind, it is natural to let the minimum book-ahead time be the time required for the bandwidth used by the initial IR calls (not counting new arrivals) to decrease from C to $C - L$. Following Duffield and Whitt [7], we suggest approximating that random recovery time by the deterministic time for the IR mean bandwidth to decrease from C to $C - L$. More generally, we can use a CLT approximation as in Section 2. For example, then we can define T to be the time such that the probability of recovery is greater than $1 - \epsilon$ for suitable small ϵ. These approximations are convenient, because we are already performing such computations for our interrupt probability calculations. Duffield and Whitt [7] show that the recovery time for the mean is asymptotically correct as an approximation for the random recovery time in large systems. That analysis also shows that the recovery time is asymptotically deterministic as the number of calls increases.

A major point of [7] is that, if the IR holding-time distribution is non-exponential and ages are used, then the recovery time can be significantly altered by conditioning when there is an exceptional initial age distribution. In particular, very slow recovery can occur in such circumstances with long-tail distributions. We suggest caution in that case.

In this section we have indicated how the minimum book-ahead time can be analyzed. We have also predicted that the required book-ahead time, should be nearly deterministic in large systems, which relates to (helps explain) the notion of "critical bookahead time" in Section 4 of Wischik and Greenberg [18].

6. Optimality

In this section we point out that, even though the proposed admission control algorithm should be effective, it should not be considered optimal. To discuss optimality, we assign costs and rewards, as in Section 3 of [13], and obtain a Markov decision problem (MDP).

The reason the admission control algorithm is not optimal is because it fails to consider differences in the BA call profile (which must be in the MDP state) and the impact of the current IR call being considered for admission upon future IR calls.

To construct a specific counterexample to optimality, let the system capacity be 2 and consider two different BA call profiles. The first candidate BA call profile is 0 up until time T_1 and then at a constant level 1 thereafter. The second candidate BA call profile is also 0 up until time T_1, but is at the same level 1 only during a short interval $[T_1, T_1 + \epsilon)$, and is thereafter 0.

Let IR calls arrive according to a Poisson process at rate 1 and let IR holding times be i.i.d. exponential random variables. Suppose that an IR call arrives at time 0 and that there is already one IR call in service. Then T_1 is the only potential interruption time.

Suppose that T_1 is chosen so that the optimal policy with the first BA call profile is indifferent between admitting or rejecting the new IR call. Then the optimal policy should strictly favor admission, and not be indifferent, for the second BA call profile, because future IR calls are less likely to be blocked if the current arrival is admitted.

7. CLT vs LD

In this section we consider a simple numerical example to provide insight into the performance of CLT and LD approximations for the interrupt probability calculations (and more generally). It is well known that the LD principles are intended for smaller tail probabilities. Consider a partial sum S_n of n i.i.d. random variables each with mean m and variance σ^2. The CLT is intended to describe the likelihood of values differing from the mean nm by the order of the standard deviation $\sqrt{n\sigma^2}$, while the LD principle describes the likelihood of values differing from the mean nm by the order n, where n is suitable large.

Since the random quantity under consideration in Section 2 is the sum of independent $\{0, b\}$-valued random variables, where b may vary, it seems reasonable to consider the special case of a sum of i.i.d. $\{0, b\}$-valued random variables, with one fixed b, which is equivalent (except for a scale parameter) to considering the binomial probability distribution.

Hence we compare the CLT and LD approximations for the tail probabilities of a binomial distribution. Let $S_n = X_1 + \cdots + X_n$ where X_i are i.i.d. Bernoulli random variables with $P(X_i = 1) = p$; i.e., S_n has a binomial distribution with

parameters n and p. We are interested in calculating the tail probability $P(S_n > \lfloor x \rfloor)$, where $x > ES_n = np$ and $\lfloor x \rfloor$ is the largest integer less than or equal to x. The exact tail probability is

$$P(S_n > \lfloor x \rfloor) = \sum_{k=\lceil x \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k} \ . \tag{7.1}$$

The standard CLT approximation is

$$P(S_n > x) \approx P\left(N(0,1) > \frac{x - ES_n}{\sqrt{Var\, S_n}}\right) = \Phi^c\left(\frac{x-np}{\sqrt{np(1-p)}}\right) \ . \tag{7.2}$$

The CLT approximation for the binomial distribution is discussed in detail in Section VII of Feller [9]. As in (3.16) on p. 185 there, one could use a refined approximation to account for the discreteness, but we do not.

As on p. 22 of Shwartz and Weiss [17], the LD approximation is

$$P(S_n > x) \approx e^{-nI(x/n)} \ , \tag{7.3}$$

where the rate function I is

$$I(x) = x \log \frac{x}{p} + (x-1) \log\left(\frac{1-p}{1-x}\right) \ . \tag{7.4}$$

For x such that $P(S_n = x) \neq 0$, the LD approximation can be refined using the Bahadur-Rao theorem as

$$P(S_n > x) \approx \frac{1}{\sqrt{2\pi np(1-p)}(1-e^{-\theta_x})} e^{-nI(x/n)} \ , \tag{7.5}$$

where

$$\theta_x = \log\left(\frac{(1-p)x}{(n-x)p}\right);$$

e.g., see [3, page 121].

For our numerical calculations, we let x exceed the mean by a specified number, α, of standard deviations for various values of α. We then compute α', the number of standard deviations by which $\lfloor x \rfloor$ exceeds the mean. Thus, we let $\sigma^2 = p(1-p)$ and $x = np + \alpha/\sqrt{n\sigma^2}$, which leaves the parameter triple (n, p, α). In Tables 2 and 3 we display numerical results for $(n, p) = (100, 0.4)$ and $(100, 0.2)$ for a range of α values. In addition to the exact values, we display the ratios approx./exact for the three candidate approximations. As expected, the CLT is much more accurate for larger tail probabilities, e.g., of order 10^{-2}.

References

[1] N. Bean, R. J. Gibbens, and S. Zachary. The performance of single resource loss systems in multiservice networks. In *Proceedings of the 14th International Teletraffic Congress*, 1994.

α	x	α'	exact	approximations CLT	LD	BR
1.0	44	0.82	0.179e-0	1.16	4.0	2.2
1.5	47	1.43	0.638e-1	1.20	5.7	1.9
2.0	49	1.84	0.271e-1	1.22	7.0	1.9
2.5	52	2.45	0.576e-2	1.24	9.3	2.0
3.0	54	2.86	0.171e-2	1.24	10.9	2.1
3.5	57	3.47	0.209e-3	1.24	13.6	2.2
4.0	59	3.88	0.425e-4	1.24	15.6	2.4
4.5	62	4.49	0.289e-5	1.23	18.9	2.6
5.0	64	4.90	0.392e-6	1.23	21.4	2.8
5.5	66	5.31	0.449e-7	1.24	24.0	3.0
6.0	69	5.92	0.125e-8	1.29	28.5	3.3
6.5	71	6.33	0.914e-10	1.36	31.8	3.6
7.0	74	6.94	0.125e-11	1.56	37.4	4.0

Table 2
A comparison of approximation with exact binomial tail probabilities. The ratio approx./exact is shown for the case $n = 100$ and $p = 0.4$. The final number of standard deviations x exceeds the mean is α'.

α=α'	x	exact	approximations CLT	LD	BR
1.0	24	0.131e-0	1.21	4.7	2.3
1.5	26	0.558e-1	1.20	6.3	2.2
2.0	28	0.200e-1	1.14	8.0	2.2
2.5	30	0.606e-2	1.02	9.9	2.4
3.0	32	0.155e-2	0.87	11.9	2.5
3.5	34	0.336e-3	0.69	14.2	2.8
4.0	36	0.619e-4	0.51	16.7	3.0
4.5	38	0.968e-5	0.35	19.2	3.2
5.0	40	0.129e-5	0.22	22.0	3.5
5.5	42	0.147e-6	0.13	25.1	3.8
6.0	44	0.143e-7	0.07	28.3	4.1
6.5	46	0.119e-8	0.034	31.8	4.5
7.0	48	0.850e-10	0.015	35.4	4.8
7.5	50	0.518e-11	0.0062	39.4	5.2
8.0	52	0.270e-12	0.0023	43.7	5.6

Table 3
A comparison of approximation with exact binomial tail probabilities. The ratio approx./exact is shown for the case $n = 100$ and $p = 0.2$.

[2] N. Bean, R. J. Gibbens, and S. Zachary. Asymptotic analysis of single resource loss systems in heavy traffic, with applications to integrated networks *Adv. Appl. Prob.,* 27 (1995), 273-292.

[3] J. A. Bucklew. *Large Deviation Techniques In Decision, Simulation, And Estimation,* Wiley, New York, 1990.

[4] G. L. Choudhury, K. K. Leung and W. Whitt, An inversion algorithm to compute blocking probabilities in loss networks with state-dependent rates, *IEEE/ACM Trans. Networking*

3 (1995), 585–601.

[5] G. L. Choudhury, K. K. Leung and W. Whitt, Efficiently providing multiple grades of service with protection against overloads in shared resources, *AT&T Tech. J.* **74** (1995), 50–63.

[6] E. G. Coffman, Jr., P. Jelenkovic and B. Poonen, Reservation probabilities, Bell Laboratories, Murray Hill, NJ, 1998.

[7] N. G. Duffield and W. Whitt, Control and recovery from rare congestion events in a large multi-server system, *Queueing Systems* **26** (1997), 69–104.

[8] N. G. Duffield and W. Whitt, A source traffic model and its transient analysis for network control, *Stochastic Models*, **14**, pp. 51–78, 1998.

[9] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I, third edition, Wiley, New York, 1968.

[10] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, second edition, Wiley, New York, 1971.

[11] A. Gersht and K.J. Lee. A bandwidth management strategy in ATM networks, Preprint, GTE Laboratories, 1990.

[12] A. G. Greenberg and R. Srikant, Computational techniques for accurate performance evaluation of multirate, multihop communication networks. *IEEE/ACM Transactions on Networking*, 5 (1997), pp. 266–277.

[13] A. G. Greenberg, R. Srikant and W. Whitt, Resource sharing for book-ahead and instantaneous-request calls, *Teletraffic Contributions for the Information Age, Proceedings of ITC 15*, V. Ramaswami and P. E. Wirth (eds.), Elsevier, Amsterdam, 1997, 539–548. Longer version to appear in *IEEE/ACM Transactions on Networking*, 1999.

[14] J. S. Kaufman, Blocking in a shared resource environment, *IEEE Trans. Commun.* **COM-29** (1981), 1474–1481.

[15] M. R. Leadbetter, G. Lindgren and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York, 1983.

[16] J. W. Roberts, A service system with heterogeneous user requirements, in *Performance of Data Communication Systems and Their Applications*, G. Pujolle (ed.), North-Holland, Amsterdam, 1981, 423–431.

[17] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*, Chapman and Hall, London, 1995.

[18] D. Wischik and A. Greenberg, Admission control for booking ahead shared resources, Proceedings of *IEEE Infocom '98*, San Francisco, CA, April 1998.

Appendix

Suppose that there are K classes of customers, with independent exponentially distributed service times. Let the number of customers of class k at time 0 be $n_k n$, where n is the parameter we will let grow. A class k customer requests b_k units of bandwidth. Let the peaks occur at times $T_1, T_2, \ldots, T_L$, and let nx_l be the available bandwidth at peak l. We are interested in the exact interrupt probability given by

$$p_I(n) = P(N_1 > nx_1 \text{ or } N_2 > nx_2 \text{ or } \ldots \text{ or } N_L > nx_L),$$

where N_l is the required capacity at T_l. We have the following large deviations result:

Theorem 7.1. In the K-class model specified above,

$$\lim_{n\to\infty} \frac{1}{n} \log p_I(n) = -\min_l I_l(x_l),$$

where

$$I_l(x) = \sup_{s>0}\{sx - \sum_{k=1}^{K} n_k \log M(sb_k)\},$$

$$M(s) = E(e^{sX_{kl}}),$$

X_{kl} is a Bernoulli random variable with parameter p_{kl}, and p_{kl} is the probability that a customer of class k has not departed the system at time T_l. For the special case $K = 1$, $I_l(x)$ is the rate function of a Bernoulli random variable with parameter p_{kl}, i.e.,

$$I_l(x) = x \log \frac{x}{p_{1l}} + (1-x) \log \frac{1-x}{1-p_{1l}}.$$

Proof. Let $l^* = \arg\max_l I_l(x_l)$. It is obvious that

$$p_I(n) \geq P(N_{l^*} > nx_{l^*}).$$

Chernoff's theorem directly gives us

$$\lim_{n\to\infty} \frac{1}{n} \log p_I(n) \geq -I_{l^*}(x_{l^*}).$$

Consider the upper bound

$$p_I(n) \leq \sum_{l=1}^{L} P(N_l > nx_l) \leq \sum_{l=1}^{L} e^{-nI_l(x_l)},$$

where the first inequality is the sum bound and the second one is the Chernoff bound. It readily follows that

$$\frac{1}{n}\log p_I(n) \leq -I_{l^*}(x_{l^*}) + \frac{1}{n}\log(1 + \sum_{l=1, l\neq l^*}^{L} e^{-n(I_l(x_l) - I_{l^*}(x_{l^*}))}).$$

Taking the limit as $n \to \infty$ gives us the desired result. $\diamond$

A more general version of the above result has been proved by Duffield using sample-path large-deviations techniques (personal communication). Duffield allows for non-exponential holding times and uses the ages of the various calls in progress to compute the interrupt probability.