Manuscript No.: POM-May-15-OA-0352.R2

Revenue Management of Reusable Resources with Advanced Reservations

Yiwei Chen[§], Retsef Levi[†], Cong Shi[‡*]

[§] Engineering Systems and Design, Singapore University of Technology and Design, Singapore 487372, yiwei_chen@sutd.edu.sg

[†] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, retsef@mit.edu

[‡] Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109, shicong@umich.edu

Abstract

We consider a revenue management problem wherein the seller is endowed with a single type resource with a finite capacity and the resource can be repeatedly used to serve customers. There are multiple classes of customers arriving according to a multi-class Poisson process. Each customer, upon arrival, submits a service request that specifies his service start time and end time. Our model allows customer advanced reservation times and services times in each class to be arbitrarily distributed and correlated. Upon arrival of each customer, the seller must instantaneously decide whether to accept this customer's service request. A customer whose request is denied leaves the system. A customer whose request is accepted is allocated with a specific item of the resource at his service start time. The resource unit occupied by a customer becomes available to other customers after serving this customer. The seller aims to design an admission control policy that maximizes her expected long-run average revenue.

We propose a policy called the ϵ-*perturbation class selection policy* (ϵ-CSP), based on the optimal solution in the fluid setting wherein customers are infinitesimal and customer arrival processes are deterministic, under the restriction that the seller can utilize at most $(1-\epsilon)$ of her capacity for any $\epsilon \in (0,1)$. We prove that the ϵ-CSP is near-optimal. More precisely, we develop an upper bound of the performance loss of the ϵ-CSP relative to the seller's optimal revenue, and show that it converges to zero with a square-root convergence rate in the asymptotic regime wherein the arrival rates and the capacity grow up proportionally and the capacity buffer level ϵ decays to zero.

Key words: algorithms; revenue management; loss network; advanced reservation; blocking probability; reusable resources.

History: Received May 2015; revisions received April 2016, September 2016; accepted November 2016 by Qi Annabelle Feng after two revisions.

*Corresponding author: shicong@umich.edu

1 Introduction

In the past several decades, revenue management has received extensive attention and has achieved huge success in a wide range of domains, such as hotel management, cloud computing, workforce management, call center service, and car rental management. As one example, in the hospitality industry, a hotel uses a finite number of rooms to dynamically accommodate customers who plan to stay. A customer who plans to stay in the hotel may book in advance by specifying the check-in and check-out dates. A room occupied by a customer becomes available after this customer checks out. The hotel dynamically makes the admission decision for each customer who submits the booking request. As another example, in the cloud computing industry, a cloud computing firm, such as Google, uses her limited computing capability to serve customer computing requests that dynamically arrive to the firm. A customer who plans to receive computational service may submit his request in advance by specifying the time periods during which he wants his job to be processed. The computing resource used in processing a customer's job becomes available after the job is finished. The firm dynamically determines whether to accept a customer's computing request. As the third example, in the workforce management, a firm, such as IBM, uses her finite workforce to work on client projects that dynamically arrive to the firm. A client who submits a project specifies when the firm can start to access to the project and when the firm has to complete the project. The workforce resource used in working on one project is released to service other clients after this project is completed. The firm dynamically determines whether to accept a client's project.

The commonalities of examples above motivate us to study a more general revenue management problem wherein the seller is endowed with a limited number of resource that are reusable over time and a customer may reserve his service in advance. To be specific, we consider the following model: The monopolist seller is endowed with a homogeneous pool of a single type of reusable resource with a known fixed capacity before the start of the selling season. The seller uses her resource to deliver service to customers over an infinite horizon. The resource occupied by a customer becomes available to other customers after finishing serving this customer. Customers are segmented into multiple classes. In each class, customers dynamically arrive to the system according to a Poisson process. Each arriving customer, at his time of arrival, submits a request for the service that specifies his service start time and end time. A customer is allowed to make an advanced reservation that the service may start after he submits the request. In each class, customers have the same per unit of time value for the service, but their advanced reservation times and service times may be potentially heterogeneous and arbitrarily correlated. For each arriving customer, the seller can observe which class this customer falls into and instantaneously decides whether to accept this customer's service request. The seller may accept a customer's service request only if serving this customer will not drive the seller to run out of resource capacity during the periods in which this customer requests to reserve. A customer whose request is denied permanently leaves the system. A customer whose request is accepted is allocated with a specific item of the resource at his service start time, rather than his arrival time. The seller aims to devise an admission control policy that maximizes her expected long-run average revenue.

The seller faces the following challenges when computing the optimal policy:

1. *Limited information.* Deriving the optimal solution requires the seller to have the perfect

information about the distributions and correlations of customer advanced reservation times and services times. However, the seller may lack such information.

2. *Curse of dimensionality.* Suppose the seller perfectly knows the information above. Then the seller needs to solve a dynamic optimization problem to compute the optimal policy. However, even in special cases (e.g., no advanced reservation allowed and with exponentially distributed service times), the resulting dynamic program seems computationally intractable because the corresponding state space grows up very fast.

We, therefore, seek a heuristic policy that can be readily implemented and also yield provably-good performance.

1.1 Main Contributions

We propose an *ϵ-perturbation class selection policy* (ϵ-CSP). This policy, defined later in optimization problem (1), is the optimal solution in the fluid setting wherein customers are infinitesimal and customer arrival processes are deterministic, under the restriction that the seller can use at most $(1-\epsilon)\times 100\%$ percent of her capacity. This policy has the following appealing features:

1. *Simple implementation.* Under the ϵ-CSP, the seller's admission decision is only based on a customer's class and whether the resource during the periods that a customer requests for is available. The seller always admits a customer as long as this customer is in some certain classes and the seller has available resource to serve this customer during his requested service periods. Otherwise, a customer's request is always rejected.

2. *Robustness.* Under the ϵ-CSP, while deciding whether to admit a customer into the system, the seller does not need to have any information about the distributions or the correlations of customer advanced reservation times and services times. Therefore, the seller avoids taking the risk of misspecifying customer demand models.

3. *Buffer for uncertainty.* The reason that we require the seller to reserve $\epsilon\times 100\%$ percent of capacity for not selling to customers in the fluid model is as follows. Recall from queuing theory that in a service system with demand uncertainty, such as customer arrival uncertainty and service time uncertainty, if the system's average utilization ratio is 1, then either customers wait in the system for extremely long times or a large fraction of customers are not admitted into the system due to non-availability of the resource. However, such bad situations can be significantly improved if the system has even only a small capacity buffer, $\epsilon\times 100\%$ percent of capacity not used in the fluid model, to deal with various uncertainties (e.g., uncertainties in customer arrival times, advanced reservation times, and service times) in our stochastic model.

Our main results about the performance of the ϵ-CSP are as follows:

1. *Finite upper bound.* The expected long-run average revenue loss of the ϵ-CSP has a finite upper bound. (This result is formally stated later in Theorem 1.)

2. *Asymptotic optimality.* In the asymptotic regime wherein the customer arrival rates and the seller's capacity proportionally grow large (multiplied by a sufficiently large number n) and the capacity buffer level ϵ goes down to zero with an appropriate speed (scaled by n as well), the ϵ-CSP is optimal. In addition, the performance loss of the ϵ-CSP relative to the seller's optimal revenue decays to zero in the speed that is almost as fast as $1/\sqrt{n}$. This relative performance loss in the asymptotic regime only depends on the capacity buffer level ϵ, rather than any other information, such as the seller's capacity level or any information about customer arrivals, advanced reservation times, or service times. (This result is formally stated later in Theorem 2.)

3. *Numerical optimality.* Our numerical experiments show that the ϵ-CSP performs within a few percentages of the optimality for a large set of parameters (even in the non-asymptotic regime).

We develop a novel approach to analyze the performance of the ϵ-CSP. We first show that the ϵ-CSP induces a well-structured stochastic process called a loss network system with advanced reservations (specifically, a $M/G/C/C$ loss system with advanced reservations). Loss network systems are concerned with the setting in which customers stochastically arrive to the system and are being served as long as there is available capacity. Customers who find a fully utilized system are lost (see the survey by Kelly (1991)). We are able to derive explicit upper bounds on the steady-state *blocking probability* (i.e., the probability that a random customer at the steady state will find a fully utilized system), and analyze them asymptotically in the above regimes. As seen from our literature review below, there have been relatively few successful attempts to characterize the blocking probabilities for loss network models with advanced reservations. Models with advanced reservations are significantly harder to analyze than those without advanced reservations. One of the major difficulties in models with advanced reservations is the fact that a randomly arriving customer effectively observes a nonhomogeneous Poisson process that is induced by the already reserved service intervals. Moreover, analyzing the blocking probability of an arriving customer requires considering the entire requested service interval instead of the instantaneous load of the system. Analyzing the load over an interval immediately introduces correlation that is hard to analyze. The upper bound on the blocking probability is obtained by considering an identical system with infinite capacity, where all customers are admitted (a $M/G/\infty$ system with advanced reservations). The probability of having more than C customers reserved in the infinite capacity system provides an upper bound on the blocking probability in the original system; we call this the *virtual blocking probability*. We obtain an exact analytical expression for this virtual blocking probability and then analyze it asymptotically. The analysis of the virtual blocking probability is tight and constitutes a contribution to the analysis of $M/G/\infty$ systems with advanced reservations. The analysis approaches significantly depart from the existing literature, and we believe that they can be effective in analyzing other core models in operations management.

The remainder of the paper is organized as follows. §1.2 provides a literature review. §2 presents our model. §3 establishes an upper bound of the seller's optimal revenue and formally defines the ϵ-CSP. §4 analyzes the performance of the ϵ-CSP. §5 extends our model to a pricing model. §6 presents the dynamic programming formulation and demonstrates the empirical effectiveness of our policy. §7 concludes the paper with some future research directions. The proofs of technical lemmas and propositions are provided in the Appendix.

1.2 Relevant Literature

Our work is closely related to the following two streams of literature: loss network systems and revenue management.

Loss Network Systems. Loss network systems without advanced reservation are well-known, and have been studied extensively, primarily in the context of communication networks (e.g., the survey by Kelly (1991)) and recently other application domains. Two of the major issues in the literature on loss networks have been the *study and design of heuristics* for admission control (e.g., Miller (1969), Ross and Tsang (1989), Key (1990), Kelly (1991), Hunt and Laws (1997), Puhalskii and Reiman (1998), Fan-Orzechowski and Feinberg (2006)), and the development of *approximations and bounds* as well as *sensitivity analysis* of blocking probabilities with respect to input parameters and resource capacities (e.g., Erlang (1917), Sevastyanov (1957), Kaufman (1981), Burman et al. (1984), Whitt (1985), Kelly (1991), Ross and Yao (1990), Zachary (1991), Louth et al. (1994), Kumar et al. (1998) and Adelman (2006)).

However, there have been relatively few successful attempts to analyze loss network systems with advanced reservation. Luss (1977) and Virtamo and Aalto (1991) analyze models for accepting and rejecting customer reservations for a discrete-time model. In their models, all customers arrive before the start of a finite service horizon and request a reservation for a start time that is uniformly distributed over the horizon. Closer to our setting, Greenberg et al. (1999) and Wischik and Greenberg (1998) consider a loss network system with advanced reservations in the context of large-bandwidth resource sharing in telecommunication. They consider instantaneous request calls as well as book-ahead calls. Their admission control policy is based on determining, under the assumption that the new call is admitted, whether the probability that a call in progress will eventually need to be interrupted exceeds a threshold value. Srikant and Whitt (2001) subsequently extend the model and results to cover multiple classes of instantaneous request calls. The interrupt and blocking probabilities are calculated by a normal approximation based on the central limit theorem. There are two major points of departure between the aforementioned works and ours. First, their instantaneous request calls have an unspecified (random) holding time, and could be interrupted during the service. In contrast, the service or holding times of our instantaneous requests are determined (drawn from a given distribution) upon arrival, and the service cannot be interrupted once admitted into the system. We remark that the service interruptions are common in large communication systems, but they are typically not allowed in many other service systems, such as hotel room management or car rental management. Second, their algorithms are based on approximating the interrupt and blocking probabilities, which work very well in simulation but do not admit any tractable analytical bounds. In contrast, we develop an explicit upper bound on the blocking probabilities and study its asymptotic behavior with a provable convergence rate.

Closer to our work, Coffman-Jr et al. (1999) derive explicit formulae for the limiting blocking probabilities for several *special cases*, for instance, in a setting where the reservation distribution is uniform and all requested intervals have unit length. They extend the result to more general reservation distributions by relating the problem to an online interval packing problem. Lu and Radovanovic (2007) study the asymptotic blocking probabilities when the capacity of the system approaches infinity with sub-exponential resource requirements. van de Vrugt et al. (2014) characterize the blocking probabilities for a single-server queue with deterministic short notice (time between arrival and starting service time) as well as discrete notice times.

Maillardet and Taylor (2016) bound the blocking probability via calculating the transient and stationary distributions for several performance measures for the infinite-server queue. From a technical viewpoint, our work is also related to the stream of literature approximating blocking probabilities for the $M_t/G/\infty$ queue as well as $M_t/G/C/C$ loss systems where the arrival process is nonhomogeneous Poisson (e.g., Eick et al. (1993b,a), Massey (1985)), and the stream of literature analyzing queues with future information (e.g., Spencer et al. (2014), Xu (2015)).

The counterpart systems with a deterministic sequence of arrivals and advanced reservation have been extensively considered in the appointment scheduling literature (see the survey by Gupta and Denton (2008)). Their objective is typically minimizing the expected costs of waiting and idle times (see, e.g., Kaandorp and Koole (2007), Begen and Queyranne (2011), Ge et al. (2013), Begen et al. (2012), Kong et al. (2013), Mak et al. (2014)). In contrast to this stream of literature, our model incorporates stochastic arrivals over time, and does not model waiting lines (as we consider loss queues). They focus on developing methods in stochastic programming, which is quite different from our work.

The theoretical results in the loss network systems with advanced reservation find interesting applications in a class of revenue management problems. The most relevant prior work in these applications is Levi and Radovanovic (2010) which use a simple knapsack-type linear program (LP) to devise a conceptually simple admission control policy called *class selection policy* (CSP) for the models without advanced reservation (i.e., all customers wish to start service upon arrival). The optimal solution obtained by solving the LP guides the policy to select the more profitable classes of customers. The LP provides an upper bound on the optimal expected long-run average revenue and can be used to analyze the performance of the CSP. The analysis is based on the fact that the CSP induces a stochastic process that can be reduced to a classical loss network model without advanced reservation. They develop explicit expressions for the resulting blocking probabilities induced by the CSP, and then show that the CSP is guaranteed to achieve at least half of the optimal long-run revenue. Also, the CSP is shown to be asymptotically optimal when the capacity goes to infinity. The knapsack-type LP considered by Levi and Radovanovic (2010) has been previously discussed by several other researchers (see, e.g., Key (1990), Hunt and Laws (1997)). In fact, a variant of the CSP has been discussed by Key (1990) and Kelly (1991), who analyze the randomized thinning policy. Moreover, Key (1990) has shown that the variant of the CSP for the single resource case without advanced reservation is asymptotically optimal in the critically loaded regime. Iyengar and Sigman (2004) have also used an identical LP to devise exponential penalty function control policies to approximately maximize the expected reward rate in a loss network. All of these works have considered models without advanced reservation.

Revenue Management. Revenue management is today a robust area of study with applications ranging from traditional domains such as hospitality to more modern ones, such as cloud computing. The text by Talluri and van Ryzin (2005), Özer and Phillips (2012), den Boer (2015) provides excellent overviews of this area.

A number of revenue management literature studies settings in which the seller uses a finite number of reusable resources to serve customers. Maglaras (2006) studies a setting wherein the seller is endowed with a single unit of resource that can be repeatedly used to serve customers. There are multiple classes of customers. Customers in each class arrive to the system according to a Poisson process, with the arrival rate modulated by the seller's dynamically posted prices

for the service. Customer service times are exponentially distributed. A customer waits in the system till the resource is idle and the seller decides to serve this customer. The seller dynamically determines prices for each class of customers and the serving priority sequence of all classes. The seller earns revenue from serving customers and incurs customer waiting costs and the server operating cost. The seller aims to find a joint pricing and sequencing policy that maximizes her long-run expected profit. Maglaras (2006) proposes a policy that is the optimal solution of the corresponding fluid model, and shows that this heuristic policy performs well. Based on this paper, Besbes and Maglaras (2009) study a setting wherein the seller cannot observe the customer arrival rate. Lei and Jasin (2016) propose two heuristic pricing policies. The first heuristic pricing policy is static that is optimal of an associated deterministic fluid model. The second heuristic pricing policy is constructed based on the static pricing policy by frequently making adjustments on previous demand realizations. They show that both heuristic pricing policies are asymptotically optimal. Nadarajah et al. (2015) study a setting wherein a customer is allowed to make a reservation in advance and request to use the resource for multiple periods. They show that pricing multiple-period usage as a bundle yields a higher revenue than pricing resource on each period separately and charge the sum of these prices over a multiple-period usage. The paper by Borgs et al. (2014) considers a setting where a firm with time-varying capacity sets prices over time to maximize revenues in the face of forward-looking customers. Inventory cannot be carried over from one epoch to the next. This paper assumes that customer arrival times, deadlines and valuations are assumed known by the firm. The focus of the paper is thus on the dynamic optimization problem that arises in this setting and the authors contribute a surprising dynamic programming formulation. Chen and Shi (2016) consider a setting with forward-looking customers. They allow customer arrival times, serve start times, service end times, service valuations, and waiting disutilities to be customer private information and heterogeneous among different customers. They adopt a mechanism design approach to show that a pricing policy that simply posts a static price for each period performs asymptotically optimal.

2 Model

We consider a monopolist seller who uses a single type resource to serve customers over an infinite horizon. At time zero, the seller is endowed with the capacity of C units of the resource. The capacity does not change over the course of the season. The resource is reusable that the resource occupied by a customer becomes available to other customers after finishing serving this customer.

There are M different classes of customers. Class $k \in \{1, \ldots, M\}$ customers arrive to the system according to an independent Poisson process with a class-dependent rate λ_k. Each class-k customer who arrives at time t_k, at his time of arrival, requests to reserve one unit of the capacity for a specified service that starts at time $t_k + d_k$ and finishes at time $t_k + d_k + s_k$, where the advanced reservation time $d_k \in \{0, 1, \cdots, u\}$ is discrete with mean ν_k and variance $\sigma^2_{d,k}$, and the service time $s_k \in \{0, 1, \cdots, v\}$ is discrete with mean μ_k and variance $\sigma^2_{s,k}$. We use capital letters D_k and S_k to denote the distributions of class-k customer advanced reservation times and service times, respectively (see Figure 1). We assume that for customers in each class k, D_k and S_k are independent of customer arrival process and between customers; however, per

customer, D_k and S_k could be arbitrarily correlated.

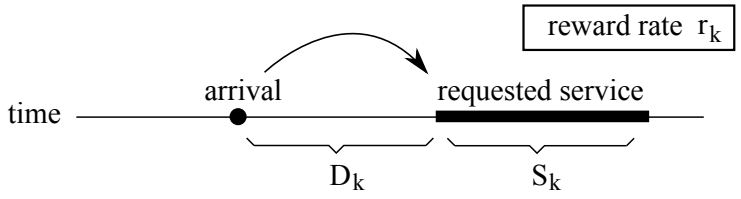


Figure 1: Reservation distributions and service distributions

Upon each customer's request, the seller has to instantaneously decide whether to accept his request. For a customer who arrives at time t and requests to get the service during time interval $[t+d, t+d+s)$, the seller accepts his request only if upon his arrival at time t, there is at least one unit of capacity that is available (i.e., not reserved before time t) throughout the entire requested interval $[t+d, t+d+s)$, i.e., this request can only be satisfied if the maximum number of already reserved resource units over $[t+d, t+d+s)$ is strictly smaller than the capacity C. However, the seller may also reject a customer's request if capacity is always available throughout the service time interval that this customer requests for. Doing so might possibly enable the seller to serve more profitable customers with the limited resource.

A customer whose request is not accepted upon his arrival permanently leaves the system without receiving any service throughout the season. Each accepted customer request cannot be canceled or changed during the season. For each class-k customer whose request is accepted, the seller collects revenue from this customer with r_k per unit of the service time.

For each customer whose request is already accepted, at his *service start time*, rather than his service request time, a specific unit among C units of the resource that is idle at that point of time is allocated to serve him. If more than one customer whose requests are accepted have the same service start time, at their common service start time, idle units among C units of the resource are allocated to these customers in a random order. This is usually the case in practice. For instance, a hotel typically assigns a room to a reserved customer only when he checks in the hotel. Also, a car rental company typically assigns a car to a reserved customer only when he arrives at the rental car pick-up location (e.g., the airport).

Denote Π to be the collection of all feasible admission policies with the resource allocation rule specified above. Every policy $\pi \in \Pi$ guarantees that an accepted customer can keep on using the same unit of the resource during his entire requested service interval. This avoids the undesirable situation in which an accepted customer is forced to switch to another resource unit during his requested service interval. We provide a formal proof of this result below.

Proposition 1. *Under every policy $\pi \in \Pi$, for a customer who arrives at time t and requests to get the service during time interval $[t+d, t+d+s)$, if his request is accepted, then he is guaranteed to keep on using the same unit of the resource during $[t+d, t+d+s)$.*

Denote $\mathcal{R}_\pi(T)$ to be the revenue that the seller achieves over the interval $[0, T]$ under policy $\pi \in \Pi$. Denote the expected long-run average revenue under policy $\pi \in \Pi$ as

$$\mathcal{R}(\pi) \triangleq \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}[\mathcal{R}_\pi(T)].$$

The seller aims to find a policy that maximizes her expected long-run average revenue (with the optimal revenue denoted by $\mathcal{R}(OPT)$):

$$\pi^* \in \arg\max_{\pi\in\Pi} \mathcal{R}(\pi).$$

It is typically quite challenging to solve this stochastic optimization problem optimally. Reasons are as follows:

1. *Limited information.* Deriving the optimal solution requires the seller to have the perfect information of the distributions and correlations of customer advanced reservation times and service times. However, our model does not require the seller to have such knowledge.

2. *Curse of dimensionality.* Suppose the information above is available to the seller, then the seller needs to solve a dynamic program. However, even in special cases (e.g., no advanced reservation allowed and with exponentially distributed service times), the resulting dynamic program seems computationally intractable because the corresponding state space grows up very fast.

Therefore, in the rest of this paper, rather than deriving the optimal solution to the above revenue maximization problem, we seek a heuristic policy that can be readily implemented and guarantee a good performance.

3 ϵ-Perturbation Class Selection Policy

We describe a simple deterministic linear program (LP) that provides an upper bound on $(1-\epsilon)$ times the achievable expected long-run average revenue for every small positive ϵ. The LP conceptually resembles to the one used by Levi and Radovanovic (2010), Key (1990) and Iyengar and Sigman (2004) who study models without advanced reservation. It is also similar in spirit to the one used by Adelman (2007) in the queueing network framework with unit resource requirements again without advanced reservation. We shall show how to use the optimal solution of the LP to construct a simple admission control policy that is called *ϵ-perturbation class selection policy* (ϵ-CSP). The original class selection policy is first analyzed by Levi and Radovanovic (2010) in models without advanced reservation. In the end of this section, we present our main results of the paper about the performance of the ϵ-CSP in the stochastic setting that we introduce in §2. We establish an upper bound of the revenue loss under the ϵ-CSP relative to the optimal revenue. We show that the relative revenue loss under the ϵ-CSP diminishes to zero in the asymptotic (high volume) regime wherein customer arrival rates $\{\lambda_k : k = 1, \cdots, M\}$ and the seller's capacity C grow up proportionally and ϵ is scaled down to converge to zero with an appropriate speed. We also characterize the speed that the relative revenue loss converges to zero in the asymptotic regime.

At any point of time t, the state of the system is specified by the entire booking profile consisting of the class, reservation and service information of each customer in the booking system as well as the customers currently served. Without loss of generality, we restrict our attention to *state-dependent policies*. Note that each state-dependent policy induces a Markov process over the state-space. Moreover, by following similar arguments as in Lu and Radovanovic (2007) and Sevastyanov (1957), one can show that the induced Markov process has a unique stationary distribution which is ergodic. (The detailed proof of ergodicity can be found in Levi and Shi (2015)). Since any state-dependent policy induces a Markov process on the state-space of the system that is ergodic, for a given state-dependent policy π, there exists a long-run stationary probability α^{π}_{dsk} for accepting a class-k customer who wishes to start service in d units of time for s units of time, which is equal to the long-run proportion of accepted customers of this type while running the policy π. In other words, any state-dependent policy π is associated with the stationary probabilities α^{π}_{dsk} for all possible reservation time d, service time s and class k. The mean arrival rate of accepted class-k customers with reservation time d and service time s is $\alpha^{\pi}_{dsk}\lambda_{dsk}$, where $\lambda_{dsk} \triangleq \lambda_k \mathbb{P}\left(D_k = d, S_k = s\right)$ is the arrival rate of a subset of class-k customers who wishes to start service in d units of time for s units of time. By applying Little's Law and the PASTA property (see Karlin and Taylor (1981)), the expected number of class-k customers with reservation time d and service time s being served in the system under state-dependent policy π is $\alpha^{\pi}_{dsk}\lambda_{dsk}s$. It follows that under policy π the expected long-run average number of resource units being used to serve customers can be expressed as $\sum_{k=1}^{M}\sum_{d,s}\alpha^{\pi}_{dsk}\lambda_{dsk}s$. Fix a small $\epsilon \in (0,1)$, this gives rise to the *knapsack* LP below:

$$\begin{aligned} \max_{\{\alpha^{\pi}_{dsk}\}} \quad & \sum_{k=1}^{M}\sum_{d,s} r_k \alpha^{\pi}_{dsk}\lambda_{dsk}s, \\ \text{s.t.} \quad & \sum_{k=1}^{M}\sum_{d,s} \alpha^{\pi}_{dsk}\lambda_{dsk}s \le (1-\epsilon)C, \\ & 0 \le \alpha_{dsk} \le 1, \ \forall d,s,k. \end{aligned} \tag{1}$$

Note that for each feasible state-dependent policy π, the vector $\alpha^{\pi} = \{\alpha^{\pi}_{dsk}\}$ is a feasible solution for the LP with objective value equal to the expected long-run average revenue of policy π.

Lemma 1. *Let $\{\alpha^{*}_{dsk}\}$ be the optimal solution of (1). Then the optimal objective value of (1) is at least $(1-\epsilon)$ times the optimal expected revenue, i.e.,*

$$\sum_{k=1}^{M}\sum_{d,s} r_k \alpha^{*}_{dsk}\lambda_{dsk}s \ge (1-\epsilon)\mathcal{R}(OPT).$$

The LP defined in (1) can be solved greedily. Without loss of generality, assume that classes are re-numbered such that $r_1 \ge r_2 \ge \ldots \ge r_M$. An optimal solution to (1) is as follows: there exists a class $M' \le M$ such that

$$\alpha^{*}_{k} \triangleq \alpha^{*}_{dsk} = \begin{cases} 1 & \text{if } k \le M'-1, \\ \min\left\{\frac{\left((1-\epsilon)C - \sum_{k=1}^{M'-1}\lambda_k\mu_k\right)}{\lambda_{M'}\mu_{M'}}, 1\right\} & \text{if } k = M', \\ 0 & \text{if } k \ge M'+1. \end{cases}$$

This solution from solving the LP gives rise to the ϵ-CSP:

(1) For each $k = 1, \ldots, M' - 1$, accept the customer upon arrival if there is sufficient unreserved capacity throughout the requested service interval.

(2) If $k = M'$, accept with probability $\alpha^*_{M'} \in (0, 1]$ if there is sufficient unreserved capacity throughout the requested service interval.

(3) For $k = M' + 1, \ldots, M$, reject.

Without loss of generality, we assume that there is no fractional variable in the optimal solution α^*, i.e., for each $k = 1, \ldots, M'$, $\alpha^*_k = 1$. (If $\alpha^*_{M'}$ is fractional, we think of class M' as having an arrival rate $\lambda'_{M'} = \alpha^*_{M'}\lambda_{M'}$ and then eliminate the fractional variable from α^*.)

We denote by

$$\rho \triangleq \sum_{k=1}^{M} \alpha^*_k \lambda_k \mu_k$$

the average number customers that are in service per unit of time. Therefore,

$$\rho = \min \left\{ (1 - \epsilon)C, \sum_{k=1}^{M} \lambda_k \mu_k \right\}. \tag{2}$$

The ϵ-CSP has the following salient features.

1. *Simple implementation.* The ϵ-CSP can be easily implemented by simply checking whether a customer's class $k \leq M'$ and whether the resource during the periods that a customer requests for is available. The seller always admits a customer as long as this customer is in some certain classes and the seller has available resource to serve this customer during his requested service period. Otherwise, a customer's request is always rejected.

2. *Robustness.* While deciding which classes of customers to admit into the system, the seller does not need to have any knowledge about the distributions or correlations of customer advanced reservation times and services times. Therefore, the ϵ-CSP avoids the seller to take the risk of misspecifying customer demand models.

3. *Buffer for uncertainty.* The reason that we do not allow the seller to fully utilize the capacity in the fluid model is as follows. Recall from queuing theory that in a service system with demand uncertainty, such as customer arrival uncertainty and service time uncertainty, if the system's average utilization ratio is 1, then either customers wait in the system for extremely long times or a large fraction of customers are not admitted into the system due to non-availability of resources. However, such bad situations can be significantly improved if the system has even only a small capacity buffer, $\epsilon \times 100\%$ percent of capacity not used in the fluid model, to deal with various uncertainties (e.g., uncertainties in customer arrival times, advanced reservation times, and service times) in our stochastic model.

We can assume that without loss of generality $\alpha^*_{M'} = 1$. (If it is fractional, we can simply consider the $\alpha^*_{M'}$-thinned Poisson arrival process.) If the M'-class arrival processes are merged,

the merged arrival process has an aggregate rate $\lambda = \sum_{k=1}^{M'} \lambda_k$, and a customer upon arrival has probability of λ_k/λ to be a class-k customer. Define $v = \max_k v_k$ and $u = \max_k u_k$. Let S (with finite discrete support $[1, v]$ and mean μ) and D (with finite discrete support $[0, u]$) be the *merged* service and reservation distributions. The joint probability mass function of S and D is $f_{D,S}(d, s) \triangleq \mathbb{P}(D = d, S = s) = \sum_{k=1}^{M'} \lambda_k/\lambda \cdot \mathbb{P}(D_k = d, S_k = s)$, for $d \in [0, u]$ and $s \in [1, v]$. Similarly, the marginal probability mass functions of S and D are $f_S(s) \triangleq \mathbb{P}(S = s) = \sum_{k=1}^{M'} \lambda_k/\lambda \cdot \mathbb{P}(S_k = s)$ and $f_D(d) \triangleq \mathbb{P}(D = d) = \sum_{k=1}^{M'} \lambda_k/\lambda \cdot \mathbb{P}(D_k = d)$, respectively, for $s \in [0, u]$ and $d \in [1, v]$. It is sufficient for our analysis to use only the marginal probability mass functions. This allows for arbitrary correlation between the reservation and service distributions of a given customer.

Now, we present our main results of this paper about the performance of the ϵ-CSP. We introduce the following notation to state our main results. Define

$$\delta \triangleq \left(\frac{\epsilon}{1-\epsilon} - \frac{\log(1+\rho)}{\rho \log \theta^{-1}} \right)^+, \tag{3}$$

where the traffic intensity ρ is given in (2), and $1 - \theta$ is the minimal non-zero reservation probability, i.e.,

$$\theta \triangleq 1 - \min_{\substack{M' \in \{1, \cdots, M\} \\ s \in [1,v],\, d \in [0,u]}} \left\{ \mathbb{P}\left(D \le d \mid k \le M', S = s\right) \;\middle|\; \mathbb{P}\left(D \le d \mid k \le M', S = s\right) > 0 \right\}. \tag{4}$$

The following theorem establishes a finite upper bound of the revenue loss of the ϵ-CSP.

Theorem 1. *Consider the revenue management model with reusable resources and advanced reservations. The expected long-run average revenue loss of the ϵ-CSP relative to the optimal expected long-run average revenue has the following finite upper bound:*

$$\frac{\mathcal{R}(\epsilon\text{-}CSP)}{\mathcal{R}(OPT)} \ge \left(1 - \frac{1}{\rho} - \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\rho} \right) (1 - \epsilon), \tag{5}$$

where ρ is given in (2) and δ is given in (3).

This finite bound spells out explicitly the dependence on the traffic intensity ρ and the buffer size ϵ. More importantly, this finite bound enables us to characterize the *convergence rate* to optimality when we scale both the arrival rate and the capacity simultaneously. The next theorem shows that the ϵ-CSP is asymptotically optimal with a provable convergence rate.

Theorem 2. *Consider the revenue management model with reusable resources and advanced reservations. Consider a sequence of problems where in the nth problem, $\lambda_k^{(n)} = n\lambda_k$ for all $k \in \{1, \cdots, M\}$, $C^{(n)} = nC$, and $\epsilon^{(n)} = \epsilon/\sqrt{n^{1-\alpha}}$ with $\alpha \in (0, 1)$. We have*

$$\frac{\mathcal{R}^{(n)}(\epsilon^{(n)}\text{-}CSP)}{\mathcal{R}^{(n)}(OPT)} \ge 1 - \frac{\epsilon}{\sqrt{n^{1-\alpha}}} + o\left(\frac{1}{\sqrt{n^{1-\alpha}}} \right). \tag{6}$$

In the asymptotic regime, for the sequence of problems as defined in the above theorem, the ϵ-CSP is optimal. In addition, the relative performance loss (6) decays to zero in the

speed that is arbitrarily close to $1/\sqrt{n}$ as α is sufficiently small. The relative performance loss in the asymptotic regime only depends on the capacity buffer level ϵ, rather than any other information, such as the seller's capacity level C or any information about customer arrivals, advanced reservation times, or service times.

In the next section, we are devoted to proving this theorem about the performance of the ϵ-CSP in the stochastic setting introduced in §2. The key idea is as follows. Since the ϵ-CSP accepts the profitable classes 1 to M' and rejects the nonprofitable classes $M'+1$ to M, it induces a well-structured stochastic process called loss network systems with advanced reservation. Each class $k = 1, \ldots, M$ induces a Poisson arrival stream with respective rate $\alpha_k^* \lambda_k$. Thus, for each class k with $\alpha_k^* = 1$, the arrival process is identical to the original process, each class k with $\alpha_k^* = 0$ can be ignored. The key aspect of the performance analysis of the ϵ-CSP boils down to finding an upper bound on the blocking probabilities in these loss network models.

4 Analysis of Blocking Probabilities

We analyze the performance of the ϵ-CSP in the stochastic setting introduced in §2. We will prove the performance loss (5) and its asymptotic behavior (6).

Before delving into details, we first give an overview of our analysis. Analyzing the original capacitated system as introduced in §2 (i.e., a $M/G/C/C$ system with advanced reservation) seems rather difficult. Instead, we consider the counterpart system with infinite capacity (i.e., a $M/G/\infty$ system with advanced reservation) while keeping all other problem parameters fixed. In this counterpart system, all customers are admitted since there are an infinite number of resources. It is not hard to see that, for each sample path and each time t, the admitted customers reserved to get service in the original capacitated system are a subset of those reserved in the infinite capacity counterpart system. Consider now a customer arriving at some random time t in the counterpart system with infinite capacity requesting service interval $[t+d, t+d+s]$. Define the *virtual blocking probability* to be the probability that the maximum reserved capacity over the requested service interval $[t+d, t+d+s]$ just prior to time t is larger than C. Since the set of served customers in the infinite capacity system is always a superset of that served in the original capacitated system, it follows that the virtual blocking probability is in fact an upper bound on the blocking probability in the original capacitated system. It makes sense to analyze the upper bounds of the virtual blocking probabilities, which, in turn, provides us upper bounds on the blocking probabilities in the original capacitated system. Let us start with the simplest non-trivial case (which gives us some insights into how to analyze such complex models), and gradually develop our main results for the general case.

4.1 The Simplest Non-Trivial Case: Two-Point Distribution

We will start the analysis with the simplest non-trivial case, and then extend it gradually to the more general case. Suppose that S takes only one value $s = 1$ deterministically. Then the traffic intensity is $\rho = \lambda\mu = \lambda$. In addition, assume that D follows a two-point distribution,

$$D = \begin{cases} 0 & \text{w.p. } \gamma, \\ 1 & \text{w.p. } 1-\gamma, \end{cases}$$

i.e., $f_D(0) = \gamma$ and $f_D(1) = 1-\gamma$. That is, an arriving customer either wants to start the service immediately or in one unit of time. Consider the counterpart system with an infinite number of servers in the steady state (note that the steady state exists due to the induced semi-Markov process). Upon a customer arrival to the system at some time t, all the starting service times of the customers who had arrived prior to t are already known. For ease of exposition, we call these starting service times *pre-arrivals*. Similarly, we call all the starting service times of the customers, who will arrive after t *post-arrivals*. Note that the pre-arrivals and post-arrivals are always defined with respect to the current time t. It is important to observe that the virtual blocking probability at time t (as well as the blocking probability in the original capacitated system) is independent of *post-arrivals*. Without loss of generality, we assume that $t = 0$ and the system reaches equilibrium.

Lemma 2 below characterizes the pre-arrival processes (i.e., the booking profile) observed by a customer arriving at time 0 in the steady state. Let $\lceil r \rceil$ be the smallest integer not less than r.

Lemma 2. *Consider the counterpart system with an infinite number of servers. Then a customer arriving at the system at time* 0 *in the steady state, observes that the pre-arrivals follow a non-homogeneous Poisson process with piecewise rate* $\eta(r)$ *at time* r*, where*

$$\eta(r) = \begin{cases} \rho, & \text{if} \quad r \leq 0, \\ (1-\gamma)\rho, & \text{if} \quad r \in (0,1], \\ 0, & \text{if} \quad r > 1. \end{cases}$$

The proof of Lemma 2 is simple by using Poisson splitting arguments. In order to figure out how likely this customer (arriving in equilibrium) gets blocked, it is important to know the entire booking profile (consisting of committed services not yet started) at the moment of his arrival. Lemma 2 gives a compact description of this pre-arrival process as seen from $t = 0$.

Let $N_d(r)$ where $r \in [0,1]$ be the Poisson counting process of the number of pre-arrivals over the interval $[d-1, d]$ *as seen from time 0*. The corresponding rates of this Poisson counting process are given by Lemma 2. Next, we introduce the notion of *mirror image* of a Poisson counting process. The mirror image of a Poisson counting process $N_d(r)$, denoted by $\tilde{N}_d(r)$, is a backward counting process of $N_d(r)$. More formally, $\tilde{N}_d(r) = N_d(1) - N_d(1-r)$ for each $r \in [0,1]$. It is evident that $\tilde{N}_d$ is also a Poisson process with the same rate as N_d. We will use $\tilde{N}_d(r)$ to model the departure process over the interval $[d, d+1]$ in reverse time.

Now let B be the event that a customer arriving at time 0 in the steady state is virtually blocked. The conditional long-run virtual blocking probability $P_d \triangleq \mathbb{P}(B \mid D = d)$, for each $d = 0, 1$. Lemma 3 below characterizes P_0 and P_1 based on the counting processes introduced above. Figure 2 gives a schematic representation of the two processes as observed a random customer arriving at $t = 0$. For clarity, we also use $N_d(\cdot; \rho)$ to denote a Poisson counting process with a given rate ρ.

Lemma 3. *Consider the counterpart system with an infinite number of servers. If a customer arrives at time* 0 *in the steady state and requests service* $S = 1$ *deterministically to commence in* D *units of time (*$D = 0$ *or* 1 *with probabilities* γ *and* $1-\gamma$*, respectively), the conditional virtual*

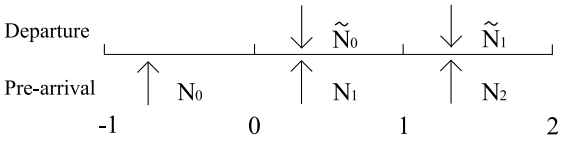


Figure 2: One-class departure and pre-arrival processes

blocking probabilities are given by

$$P_0 \triangleq \mathbb{P}(B \mid D=0) = \mathbb{P}\left(\max_{r\in[0,1]}\left\{\tilde{N}_0(1-r;\rho) + N_1(r;(1-\gamma)\rho)\right\} \geq C\right),$$
$$P_1 \triangleq \mathbb{P}(B \mid D=1) = \mathbb{P}\left(\tilde{N}_1(1;(1-\gamma)\rho) \geq C\right),$$

where N_d $(d=0,1,2)$ is a Poisson counting process with rate ρ_d and $\tilde{N}_d$ $(d=0,1)$ is the mirror image of N_d with the same rate ρ_d. Moreover, $\rho_0=\rho$, $\rho_1=(1-\gamma)\rho$ and $\rho_2=0$.

Lemma 3 essentially tells us that the virtual blocking probability can be expressed in terms of the maximum of the sum of these two Poisson counting processes running towards each other (see Figure 3), one of which representing the pre-arrival process (the committed services not yet started) and the other one representing the departure process. The random process (inside the max operator) is hard to analyze, which is very different than merging two Poisson counting processes running in the same direction.

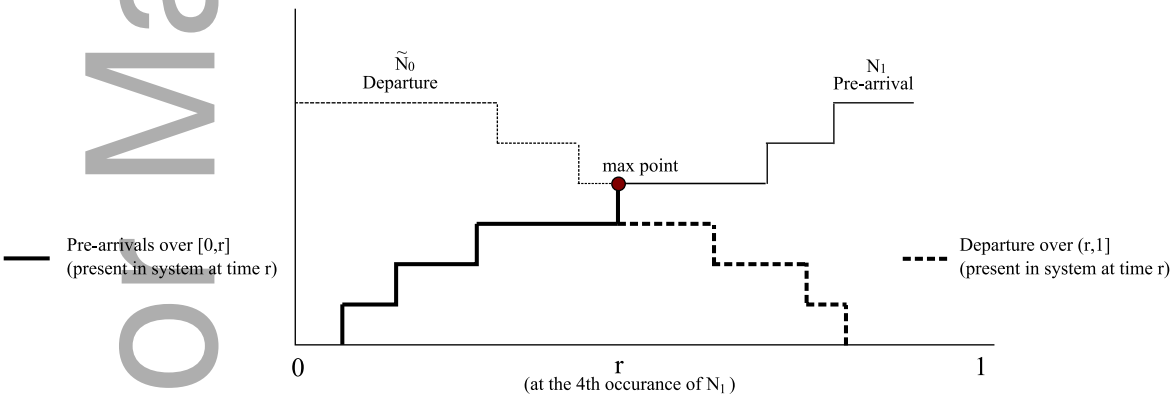


Figure 3: Two Poisson counting processes running towards each other

To analyze the above blocking probabilities, we shall prove a more general statement that will be the key building block in our analysis of the general case.

Proposition 2. *Let N_0, N_1 and N_2 be Poisson counting processes (mutually independent) with rates ρ, $\theta_1\rho$ and $\theta_2\rho$, respectively, where $1>\theta_1\geq\theta_2\geq 0$ are fixed constants. Let $\tilde{N}_0$ and $\tilde{N}_1$ be the mirror images of N_0 and N_1, respectively. Define two random variables X and Y as follows,*

$$X \triangleq \max_{r\in[0,1]}\left\{\tilde{N}_0(1-r;\rho) + N_1(r;\theta_1\rho)\right\}, \qquad Y \triangleq \max_{r\in[0,1]}\left\{\tilde{N}_1(1-r;\theta_1\rho) + N_2(r;\theta_2\rho)\right\}. \tag{7}$$

Then,

$$\max\{\mathbb{P}(X \geq C), \mathbb{P}(Y \geq C)\} \leq \frac{1}{\rho} + \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\rho},$$

where

$$\delta = \left(\frac{\epsilon}{1-\epsilon} - \frac{\log \rho}{\rho \log \theta_1^{-1}}\right)^{+}.$$

Note that P_0 and P_1 (in Lemma 3) can be obtained by simply setting $\theta_1 = (1-\gamma)$ and $\theta_2 = 0$ in X and Y above. To prove Proposition 2, we provide an alternative characterization of X and Y above based on a downward-drifting asymmetric random walk process that takes a down-step, for each departure, and an up-step, for each pre-arrival. We would like to show that asymptotically the maximum level of the random walk stays relatively close to its starting position by showing that the rate of the random walk going up is sublinear in $\sqrt{\rho}$.

Consider the merged process induced on $[0,1]$ by the two Poisson counting processes $\tilde{N}_0$ and N_1. Let $\mathcal{N} = \tilde{N}_0(1;\rho) + N_1(1;\theta_1\rho)$ denote the total number of occurrences (pre-arrivals and departures) over $[0,1]$ of the two independent Poisson counting processes of $\tilde{N}_0$ and N_1. Note that since $\tilde{N}_0$ and N_1 are independent of each other, $\mathcal{N}$ is a Poisson random variable with rate $(1+\theta_1)\rho$. Conditioning on $\mathcal{N} = n$, the induced merged process has n points uniformly distributed over the interval $[0,1]$. By the splitting argument applied to the merged process, each of these n points has independent probability $p = \frac{\theta_1\rho}{(1+\theta_1)\rho} = \frac{\theta_1}{1+\theta_1} < \frac{1}{2}$ to be from the process N_1 and probability $q = 1-p$ from the process $\tilde{N}_0$. If we associate $+1$ with each point from N_1, and -1 with each point from $\tilde{N}_0$, then each configuration of these n points induces a downward-drifting asymmetric random walk of length n. The random walk starts at the origin 0, with up probability p and down probability q.

Lemma 4. *Let $\mathcal{R}_n$ denote the corresponding random walk of length n described above, and M_n denote the maximum level attained by $\mathcal{R}_n$, and G_n denote the overall number of down-steps taken by $\mathcal{R}_n$. Also let $X_n \triangleq (X \mid \mathcal{N} = n)$ with X defined in (7). Then, $X_n = G_n + M_n$ almost surely.*

However, it should be noted that M_n and G_n are correlated. To address the correlation between M_n and G_n, we will replace M_n by M_∞. However, first we would like to obtain an expression for the hitting probability of a downward-drifting asymmetric random walk. This is done in Lemma 5 given below. (Lawler (2006) provides a proof in Chapter 2, Section 2.2; for completeness, we present a shorter proof.)

Lemma 5. *Consider a random walk defined by a sequence of independent random variables $\{E_i\}$ where E_i takes value 1 with probability p and 0 with probability $q = 1-p$. Let $S_n = \sum_{i=1}^{n} E_i$. Define $M_\infty \in [0,\infty) \bigcup \{\infty\}$ to be the maximum level attained by the random walk (i.e., $M_\infty = \max_n S_n$). Given that $0 \leq p < q \leq 1$ (downward drifting), then the probability that the random walk ever hits above level b is $\mathbb{P}(M_\infty \geq b) = (p/q)^b$.*

Proposition 2 can be proved by using Lemma 5. The complete proof of Proposition 2 can be found in the appendix.

4.2 The General Case

Next we extend the simple model to allow for an arbitrary finite discrete reservation distribution D with marginal probability mass function $f_D(d)$. We still assume that the service distribution remains fixed at $S = 1$, deterministically. Now let $\gamma_d \triangleq f_D(d)$ for $d \in [0, u]$. Thus, $\gamma_d \in [0, 1]$ and $\sum_{d=0}^{u} \gamma_d = 1$. Then the traffic intensity is $\rho = \lambda \left(\sum_{d=0}^{u} \gamma_d\right) \mu = \lambda$. Lemma 6 below is a generalization of Lemma 2. The proofs of all lemmas and propositions in this subsection can be found in the Appendix.

Lemma 6. *Consider the counterpart system with an infinite number of servers. Then a customer arriving at the system at time* 0 *in the steady state, observes that the pre-arrivals follow a non-homogeneous Poisson input process with piecewise rate* $\eta(r)$ *at time* r*, where*

$$\eta(r) = \begin{cases} \rho, & \text{if} \quad r \le 0, \\ \rho \left(1 - \sum_{d=0}^{\lceil r \rceil - 1} \gamma_d\right), & \text{if} \quad r > 0. \end{cases}$$

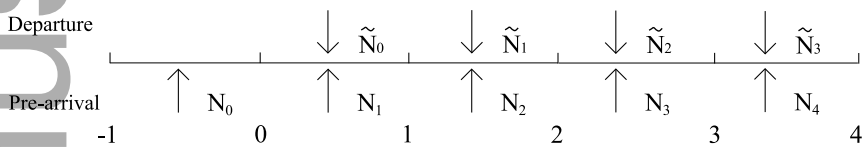


Figure 4: One-class departure and pre-arrival processes with general reservation distribution

Define N_d (for $d \in [0, u]$) to be the process of pre-arrivals prior to t over $(d-1, d]$. This process induces a departure process over the interval $(d, d+1]$, and let $\tilde{N}_d$ denote its mirror image. Figure 4 shows the pre-arrival and departure processes with general reservation distribution. The conditional virtual blocking probabilities are given in Lemma 7 below, which is a generalization of Lemma 3.

Lemma 7. *Consider the counterpart system with an infinite number of servers. If a customer comes at time* 0 *in the steady state and requests service (*$S = 1$*) deterministically to commence in* D *units of time (*$D \in [0, u]$*), then the conditional virtual blocking probability is given by*

$$P_d \triangleq \mathbb{P}(B \mid D = d) = \mathbb{P}\left(\max_{r \in [0,1]} \left\{\tilde{N}_d(1 - r; \rho_d) + N_{d+1}(r; \rho_{d+1})\right\} \ge C\right), \qquad \forall d \in [0, u],$$

where N_d *is a Poisson counting process with rate* $\rho_d = \rho\left(1 - \sum_{i=0}^{d-1} \gamma_i\right)$*, and* $\tilde{N}_d$ *is a mirror image of* N_d *with the same rate* ρ_d*.*

Proposition 3 is a generalization of Proposition 2 with general distribution functions of customer advanced reservation times.

Proposition 3. *Let the service distribution* $S = 1$ *deterministically, and the reservation distribution* D *be discrete with marginal probability mass function* $f_D(d) = \gamma_d$ *and bounded support* $[0, u]$*. Then for all* $d \in [0, u]$*,*

$$P_d \le \frac{1}{\rho} + \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\rho},$$

where δ *is defined by* (3)*.*

Next we extend the model further to allow for an arbitrary finite discrete service distribution. The total admitted customer arrival rate is λ, and the reservation distribution D is defined on $[0, u]$ defined as above. Now assume that the service time S is a general finite discrete distribution on $[1, v]$. More specifically, let $f_S(\cdot)$ be the marginal probability mass function with $f_S(s) = \mathbb{P}(S = s) \triangleq \kappa_s$. Thus, for each $s \in [1, v]$, $\kappa_s \in [0, 1]$ and $\sum_{s=1}^{v} \kappa_s = 1$.

We partition the arriving customers according to their requested service times, i.e., customers are partitioned into v disjoint sets numbered $1, \ldots, v$ according to their requested service times. For each $s \in [1, v]$, the arrival process of customers in set s follows a thinned Poisson process with rate $\kappa_s\lambda$. Moreover, these processes are independent of each other. Now, for each set $s \in [1, v]$, let the conditional reservation probability mass function be $\gamma_d^s \triangleq \mathbb{P}(D = d \mid S = s)$ for $d \in [0, u]$. Note that $\sum_{d=0}^{u} \gamma_d^s = 1$, for each $s \in [1, v]$.

Consider the counterpart system with an infinite number of servers. If a customer of set s ($s \in [1, v]$) arrives at time 0 in the steady state and requests s units of service time to commence after d units of time ($d \in [0, u]$), the conditional virtual blocking probability is defined as $P_d^s \triangleq \mathbb{P}(B \mid D = d, S = s)$. In addition, the traffic intensity is $\rho = \sum_{s=1}^{v} s\kappa_s\lambda = \mu\lambda$, where $\mu = \sum_{s=1}^{v} s\kappa_s$ is the mean service time.

Let N_d^s (for $s \in [1, v]$ and $d \in [0, u]$) denote the pre-arrival process of set-s customers over (i.e., customers requesting s units of service time) the interval $(d - s, d - s + 1]$. This induces a departure process over the interval $(d, d + 1]$, and let $\tilde{N}_d^s$ denote its mirror image. The rate of N_d^s and $\tilde{N}_d^s$ are given in Lemma 8 below.

Lemma 8. *Let N_d^s and $\tilde{N}_d^s$ be defined as above. Then, for each $s \in [1, v]$ and each $d \in [0, u]$, N_d^s and $\tilde{N}_d^s$ are Poisson processes with the same rate*

$$\rho_d^s = \rho_0^s \left(1 - \sum_{i=0}^{d-s} \gamma_i^s\right) = \kappa_s \rho \left(1 - \sum_{i=0}^{d-s} \gamma_i^s\right). \tag{8}$$

Moreover, N_d^s is independent of $N_{d'}^{s'}$ for $d \neq d'$ or $s \neq s'$.

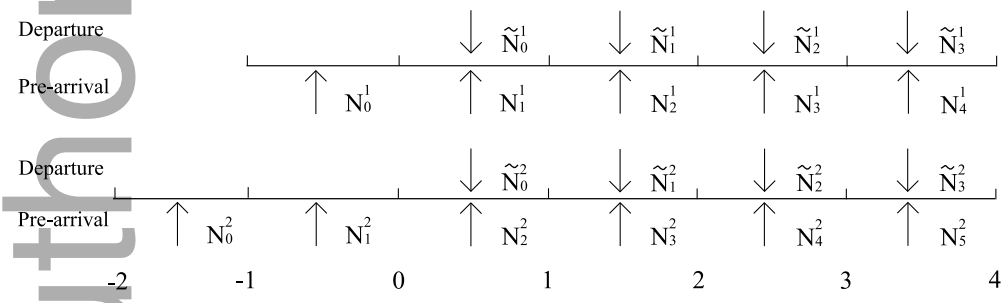


Figure 5: Two-service-set departure and pre-arrival processes

First assume that $\exists s \in [1, v]$ such that $\gamma_0^s > 0$, i.e., the probability of an arriving customer requesting to start the service immediately upon arrival is strictly positive. This assumption can be dropped later. Let A_d be the maximum number of customers in the system over the interval $(d, d+1]$ for $d \in [0, u]$. In fact, one can derive an exact mathematical expression of each

A_d for $d \in [0,u]$,

$$A_d = \sum_{s=2}^{v} \sum_{i=d+1}^{d+s-1} N_i^s(1;\rho_i^s) + \max_{r\in[0,1]} \left\{ \sum_{s=1}^{v} \tilde{N}_d^s(1-r;\rho_d^s) + \sum_{s=1}^{v} N_{d+s}^s(r;\rho_{d+s}^s) \right\}. \tag{9}$$

For $r \in [0,1]$, the term $\sum_{s=1}^{v} \tilde{N}_d^s(1-r;\rho_d^s)$ captures all the departures over $(d+r, d+1]$, the term $\sum_{s=1}^{v} N_{d+s}^s(r;\rho_{d+s}^s)$ captures all the pre-arrivals over $(d, d+r]$, and the term $\sum_{s=2}^{v}\sum_{i=d+1}^{d+s-1} N_i^s(1;\rho_i^s)$ captures all the customers being served over $(d, d+1]$. The sum captures exactly all the customers being served at time $d+r$. It is important to note that since $\tilde{N}_i^s$ and N_i^s do not simultaneously appear in A_d, for each $i \in [0,u]$ and $s \in [1,v]$, all the Poisson counting processes in the expression of A_d are independent of each other (see Lemma 8).

We shall further explain (9) by providing the following example when $v = 2$ (refer to Figure 5),

$$\begin{aligned} A_0 &= N_1^2(1;\rho_1^2) + \max_{r\in[0,1]} \left\{ \tilde{N}_0^1(1-r;\rho_0^1) + \tilde{N}_0^2(1-r;\rho_1^2) + N_1^1(r;\rho_1^1) + N_2^2(r;\rho_2^2) \right\}, \\ A_1 &= N_2^2(1;\rho_2^2) + \max_{r\in[0,1]} \left\{ \tilde{N}_1^1(1-r;\rho_1^1) + \tilde{N}_1^2(1-r;\rho_1^2) + N_2^1(r;\rho_2^1) + N_3^2(r;\rho_3^2) \right\}, \text{ and so on.} \end{aligned}$$

More specifically, A_0 represents the maximum customers in the system over the interval $(0,1]$. At time $r \in (0,1]$, the number of departures over $(r,1]$ is equal to $\tilde{N}_0^1(1-r;\rho_0^1) + \tilde{N}_0^2(1-r;\rho_0^2)$, capturing customers in both sets starting before 0 and still in the system at time r. (Note that the service time is at least 1.) In addition, the number of pre-arrivals over $(0,r]$ is equal to $N_1^1(r;\rho_1^1) + N_2^2(r;\rho_2^2)$, capturing pre-arrivals of customers with service time 1 and 2, respectively, starting service over $(0,r]$. Finally, N_1^2 captures set-2 customers with service time 2 who started service within $(-1,0]$. These customers will continue service over the entire interval $(0,1]$. Therefore $N_1^2(1;\rho_1^2)$ appears in the expression A_0 outside the max. The same reasoning applies to A_i for each $i \in [1,u]$.

Now for each $d \in [0,u]$ and $s \in [1,v]$, we have $P_d^s = \mathbb{P}(\max(A_d, \ldots, A_{d+s-1}) \geq C)$. It should be noted that A_d and $A_{d'}$ can be correlated. To analyze the upper bound of P_i^j, we first analyze the upper bound of $\mathbb{P}(A_d \geq C)$, for each $d \in [0,u]$.

Lemma 9. *Assume that there exists $s \in [1,v]$ such that $\gamma_0^s > 0$. Then,*

$$\mathbb{P}(A_0 \geq C) \leq \frac{1}{\rho} + \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\rho,$$

where δ is defined by (3).

Proposition 4. *Let the service distribution S be discrete with marginal probability mass function $f_S(s) = \kappa_s$ and bounded support $[1,v]$, and the reservation distribution D be discrete with marginal probability mass function $f_D(d) = \gamma_d$ and bounded support $[0,u]$. The traffic intensity is given by $\rho = \sum_{s=1}^{v} s\kappa_s\lambda = \sum_{s=1}^{v} s\lambda_0^s = \mu\lambda$. Then for all $d \in [0,u]$ and $s \in [1,v]$,*

$$P_d^s \leq \frac{1}{\rho} + \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\rho,$$

where δ is given by (3).

Now, we prove the main results of this paper, Theorems 1 and 2.

Proof of Theorem 1. The expected long-run average revenue of the ϵ-CSP is $\sum_{k=1}^{M'}\sum_{d,s} r_k\lambda_{dsk}s(1-Q_{dsk})$, where Q_{dsk} is the stationary probability of blocking a class-k customers with reservation time d and service time s. However, $\sum_{k=1}^{M'}\sum_{d,s} r_k\lambda_{dsk}s$ is the optimal value of the LP defined in (1), which is an upper bound on $(1-\epsilon)\mathcal{R}(OPT)$ for small positive ϵ by Lemma 1. Thus, a key aspect of the performance analysis of the ϵ-CSP is to obtain an upper bound on the probabilities $\{Q_{dsk}\}$. Specifically, if $Q_{dsk} \le \xi$, for each d, s, and k, then

$$\mathcal{R}(\epsilon\text{-}CSP) = \sum_{k=1}^{M'}\sum_{d,s} r_k\lambda_{dsk}s(1-Q_{dsk}) \ge \sum_{k=1}^{M'}\sum_{d,s} r_k\lambda_{dsk}s(1-\xi) \ge (1-\xi)(1-\epsilon)\mathcal{R}(OPT).$$

By Proposition 4, we have

$$\xi \le \frac{1}{\rho} + \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\rho}.$$

Therefore,

$$\frac{\mathcal{R}(\epsilon\text{-}CSP)}{\mathcal{R}(OPT)} \ge \left(1 - \frac{1}{\rho} - \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\rho}\right)(1-\epsilon).$$

□

Proof of Theorem 2. Consider a sequence of problems where in the nth problem, $\lambda_k^{(n)} = n\lambda_k$ for all $k \in \{1,\cdots,M\}$, $C^{(n)} = nC$, and $\epsilon^{(n)} = \epsilon/\sqrt{n^{1-\alpha}}$ with $\alpha \in (0,1)$.

Following Equation (2), we have

$$\begin{aligned}\rho^{(n)} &= \min\left\{\left(1-\epsilon^{(n)}\right)C^{(n)}, \sum_{k=1}^{M}\lambda_k^{(n)}\mu_k\right\} \ge \min\left\{(1-\epsilon)\,C^{(n)}, \sum_{k=1}^{M}\lambda_k^{(n)}\mu_k\right\}\\ &= n\min\left\{(1-\epsilon)\,C, \sum_{k=1}^{M}\lambda_k\mu_k\right\} = n\rho,\end{aligned}$$

where the first inequality follows the property that $\epsilon^{(n)} \le \epsilon$.

Thus,

$$\delta^{(n)} = \left(\frac{\epsilon^{(n)}}{1-\epsilon^{(n)}} - \frac{\log(1+\rho^{(n)})}{\rho^{(n)}\log\theta^{-1}}\right)^{+} = \epsilon^{(n)} + o\left(\epsilon^{(n)}\right) = \frac{\epsilon}{\sqrt{n^{1-\alpha}}} + o\left(\frac{1}{\sqrt{n^{1-\alpha}}}\right).$$

We have

$$\begin{aligned}
\left(\frac{e^{\delta^{(n)}}}{(1+\delta^{(n)})^{1+\delta^{(n)}}}\right)^{\rho^{(n)}} &= \exp\left(\log\left(\left(\frac{e^{\delta^{(n)}}}{(1+\delta^{(n)})^{1+\delta^{(n)}}}\right)^{\rho^{(n)}}\right)\right) \\
&= \exp\left(\rho^{(n)}\left(\delta^{(n)}-\left(1+\delta^{(n)}\right)\log\left(1+\delta^{(n)}\right)\right)\right) \\
&= \exp\left(\rho^{(n)}\left(-\left(\delta^{(n)}\right)^2+o\left(\left(\delta^{(n)}\right)^2\right)\right)\right) \\
&\leq \exp\left(n\rho\left(-\left(\delta^{(n)}\right)^2+o\left(\left(\delta^{(n)}\right)^2\right)\right)\right) \\
&= \exp\left(-n^{\alpha}\rho\epsilon^2\right)+o\left(\exp\left(-n^{\alpha}\right)\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
1-\frac{1}{\rho^{(n)}}-\left(\frac{e^{\delta^{(n)}}}{(1+\delta^{(n)})^{1+\delta^{(n)}}}\right)^{\rho^{(n)}} &\geq 1-\frac{1}{n\rho}-\exp\left(-n^{\alpha}\rho\epsilon^2\right)+o\left(\exp\left(-n^{\alpha}\right)\right) \\
&= 1-\frac{1}{n\rho}+o\left(\frac{1}{n}\right).
\end{aligned}$$

The equality follows the property that for $\alpha > 0$,

$$\lim_{n\to\infty}\frac{\exp\left(-n^{\alpha}\rho\epsilon^2\right)}{1/n}=0.$$

Therefore,

$$\begin{aligned}
\frac{\mathcal{R}^{(n)}(\epsilon^{(n)}\text{-}CSP)}{\mathcal{R}^{(n)}(OPT)} &\geq \left(1-\frac{1}{\rho^{(n)}}-\left(\frac{e^{\delta^{(n)}}}{(1+\delta^{(n)})^{1+\delta^{(n)}}}\right)^{\rho^{(n)}}\right)\left(1-\epsilon^{(n)}\right) \\
&\geq \left(1-\frac{1}{n\rho}+o\left(\frac{1}{n}\right)\right)\left(1-\frac{\epsilon}{\sqrt{n^{1-\alpha}}}\right) \\
&= 1-\frac{\epsilon}{\sqrt{n^{1-\alpha}}}+o\left(\frac{1}{\sqrt{n^{1-\alpha}}}\right).
\end{aligned}$$

The equality follows the property that for $\alpha \in (0,1)$,

$$\lim_{n\to\infty}\frac{1/n}{1/\sqrt{n^{1-\alpha}}}=0.$$

□

We note that the above performance analysis (on blocking probabilities) for the model with customer advanced reservation is completely different from the one used in Levi and Radovanovic (2010) for models without advanced reservation. In the following, we also discuss the key distinction of our results from two related models.

Key Distinction from the Models without Advanced Reservation. In contrast to models without advanced reservation (e.g., Levi and Radovanovic (2010)), the major challenge in analyzing the blocking probabilities in loss network systems with advanced reservation is that the blocking event depends on the maximum reserved capacity over a particular requested service interval. This requires the characterization of the booking profile (i.e., the pre-reserved arrival and departure processes) over the service interval. As a simple example with capacity $C = 2$ shown in Figure 6, in the models without advanced reservation, it suffices to check the instantaneous load of the system upon arrival of a customer. However, in the models with advanced reservation, we cannot guarantee one's request by merely checking the instantaneous load of the system at his starting service time upon his arrival, because his request may be potentially blocked by reserved slots of those customers who booked prior to him but will start services after him. (In this simple example, the system has only 1 customer in service when his service begins; however, during his requested service interval, there is a point in time that the system has 3 customers (who were reserved before him). Thus, he has to be rejected by the system.) This introduces much difficulty in handling this correlation issue between the incoming requests and the booking profiles.

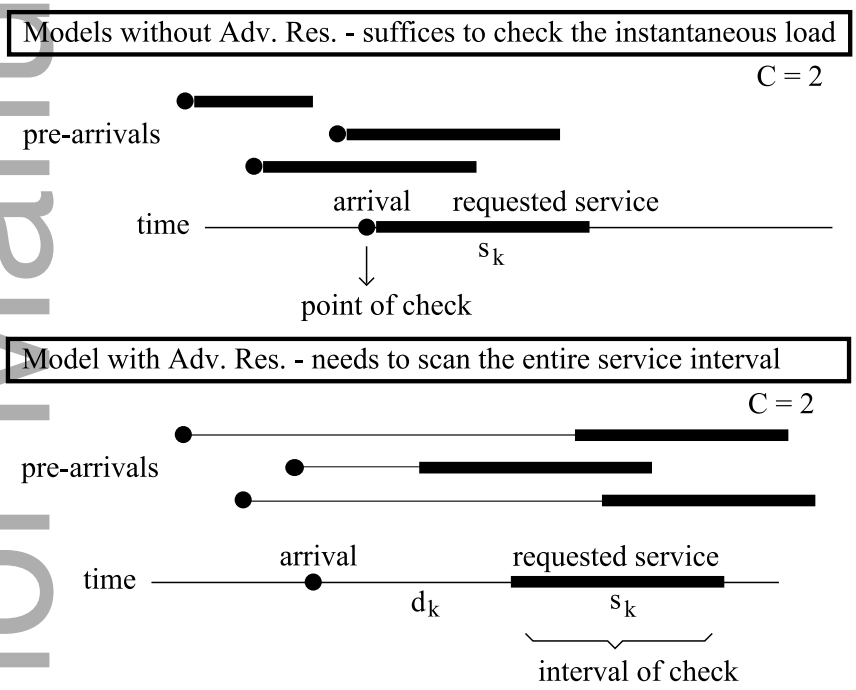


Figure 6: Challenges in analyzing the blocking probabilities in loss queues with adv. reservation

Key Distinction from the Tandem Queues of Two Stations. Tandem queueing model is closely related to our model since one may regard the original system with advanced reservation as a tandem queueing model of two stations, where the advance reservation is spent in the first station and the service is spent in the second station. The first station has infinite capacity and the second station has finite capacity, customers first enter the system from the first station, but if the second station is full when customers arrive, they will be rejected or lost. Note that

the decision whether a customer is blocked is made only after the service in the first station is over. However, in our model, we have to make the decision as soon as the customer enters the first station. The critical difference is that in the tandem queueing model it suffices to check the instantaneous load of the second station to determine if the customer is blocked, while in our model we have to check the maximum occupancy over the entire requested service interval as seen from the moment the customer arrives. When we relax the finite capacity assumption on the second station, the blocking probability can seemingly be approximated by the probability that the number of customers in the second station is bigger than C (see, e.g., Boxma (1984), Schmidt (1987)). However, this approximation cannot be used to upper bound the blocking probabilities in our system due to the difference in system dynamics. It may serve as an approximation of the blocking probabilities but we are unsure how good the approximation is, since the stationary distribution can no longer be expressed as a product-form.

5 Pricing Extensions

We present an interesting pricing extension, in which the arrival rates of the different classes of customers are affected by prices. Specifically, consider a two-stage decision. At the first stage, we set the respective prices $r_1, \ldots, r_M$ for each class. This determines the respective arrival rates $\lambda_1(r_1), \ldots, \lambda_M(r_M)$. (The rate of class-$k$ customers is affected only by price r_k.) Then, given the arrival rates, we wish to find the optimal admission policy that maximizes the expected long-run revenue rate. In particular, we assume that $\lambda_k(r_k)$ is nonnegative, differentiable, and decreasing in r_k for each $k \in \{1, \cdots, M\}$. In addition, we assume that all prices are nonnegative real numbers and that there exists a price r_∞ such that, for each $i = k, \ldots, M$, we have $\lambda_k(r_\infty) = 0$. (The latter condition is required to guarantee that the problem has an optimal solution.)

For the model with price-driven demand we use the following nonlinear program (NLP1):

$$\begin{aligned} \max_{\{\alpha_{dsk}, r_k\}} \quad & \sum_{k=1}^{M} \sum_{d,s} r_k \alpha_{dsk} \lambda_{dsk}(r_k) s, && (10) \\ \text{s.t.} \quad & \sum_{k=1}^{M} \sum_{d,s} \alpha_{dsk} \lambda_{dsk}(r_k) s \le (1-\epsilon) C, \\ & 0 \le \alpha_{dsk} \le 1, \qquad \forall d, s, k, \\ & 0 \le r_k \le 1, \qquad \forall k. \end{aligned}$$

In particular, it can be verified that any optimal solution of (NLP1) has only nonnegative prices. Also, observe that for any fixed prices $r_1, \ldots, r_M$, the corresponding solution of $\{\alpha_{dsk}\}$ has the same knapsack structure defined in §2 above. Let $(r^*, \alpha^*) = \{r_k, \alpha_{dsk}\}$ be the corresponding optimal solution. Note that if one can solve (NLP1) and obtain the solution (r^*, α^*) then one can construct a similar ϵ-CSP that will be amenable to the same performance analysis discussed in §4 above. However, solving (NLP1) directly may be computationally hard. Next, we show that one can reduce (NLP1) to an equivalent nonlinear program that is more tractable; we denote it by (NLP2). (By equivalent we mean that they have the same set of optimal solutions.) Consider the following nonlinear program (NLP2):

$$\begin{aligned} \max_{\{r_k\}} \quad & \sum_{k=1}^{M} \sum_{i,j} r_k \lambda_{dsk}(r_k) j, \\ \text{s.t.} \quad & \sum_{k=1}^{M} \sum_{d,s} \lambda_{dsk}(r_k) s \le (1-\epsilon) C, \\ & 0 \le r_k \le 1, \qquad \forall k. \end{aligned} \tag{11}$$

It can be readily verified that as long as $\lambda_{dsk}(r_k)$ is nonnegative (and decreasing) it is always optimal to have nonnegative prices, so the nonnegativity constraints can be dropped.

Theorem 3. *The programs (NLP1) and (NLP2) are equivalent.*

Theorem 3 implies that we can solve (NLP2) instead of solving (NLP1). However, (NLP2) is computationally more tractable and can be solved relatively easily in many scenarios. Specifically, Lagrangify (dualize) the constraint in (NLP2) with some Lagrange multiplier Θ and consider the unconstraint problem $\max_{r_k \in [\Theta, r_\infty)} \sum_{1 \le k \le M} \sum_{d,s} (r_k - \Theta) \lambda_{dsk}(r_k) s$, which is separable in $r_1, \ldots, r_{M'}$. In fact, one aims to find the minimal Θ for which the resulting solution satisfies the constraint in (NLP2). This can be done by applying bi-section search on the interval $[0, p_\infty]$. The complexity of this procedure depends on the complexity of maximizing $\sum_{1 \le k \le M} \sum_{d,s} (r_k - \Theta) \lambda_{dsk}(r_k) s$ for each $k \in \{1, \cdots, M\}$. It is not hard to check that there are at least two tractable cases: (i) $\lambda_{dsk}(r_k)$ is a concave function on $[0, r_\infty)$, for each $k \in \{1, \cdots, M\}$; (ii) $\lambda_{dsk}(r_k)$ is convex, but $r_k \lambda_{dsk}(r_k)$ is concave function on $[0, r_\infty)$, for each $k \in \{1, \cdots, M\}$.

6 Numerical Experiments

We examine the empirical performance of the proposed ϵ-CSP to the optimal solution provided by the benchmark linear program (1). Solving the ϵ-CSP solutions is extremely efficient and our extensive simulation results show that the ϵ-CSP performs near-optimal in the heavy traffic regime (with average optimality gap less than 4%), and is also very robust with respect to different input distributions and parameters. This is consistent with our theoretical development of the proposed policy. Additionally, the numerical results suggest that the ϵ-CSP performs quite well even in the light or medium traffic regimes (with average optimality gap less than 10%). All the simulations were implemented in an Intel Xeon 3.50GHz PC.

6.1 Design of Experiments

We first describe a baseline model, and then vary the distributions and parameters of the baseline model to comprehensively study the empirical performance of the ϵ-CSP. The baseline model (e.g., in a hotel room management setting) is constructed as follows: the number classes of customers $M = 8$, the total capacity $C = 40$ units, and the buffer size $\epsilon = 0.001$. In addition, the class-dependent arrival rates, revenue rates, advanced reservation and service distributions for the baseline model are specified in Table 1 below. Note that $\mathcal{N}(a, b)$ denotes the truncated normal distribution with mean a and variance b.

Class No.	Arrival Rate	Adv. Reservation	Service Time	Revenue Rate ($)
1	2	Short $\sim \mathcal{N}\left(3, 1^2\right)$	Short $\sim \mathcal{N}\left(3, 1^2\right)$	150
2	3	Short $\sim \mathcal{N}\left(3, 1^2\right)$	Long $\sim \mathcal{N}\left(10, 2^2\right)$	140
3	2	Long $\sim \mathcal{N}\left(30, 10^2\right)$	Short $\sim \mathcal{N}\left(3, 1^2\right)$	130
4	2	Long $\sim \mathcal{N}\left(30, 10^2\right)$	Long $\sim \mathcal{N}\left(10, 2^2\right)$	120
5	1	Short $\sim \mathcal{N}\left(3, 1^2\right)$	Short $\sim \mathcal{N}\left(3, 1^2\right)$	110
6	2	Short $\sim \mathcal{N}\left(3, 1^2\right)$	Long $\sim \mathcal{N}\left(10, 2^2\right)$	100
7	3	Long $\sim \mathcal{N}\left(30, 10^2\right)$	Short $\sim \mathcal{N}\left(3, 1^2\right)$	90
8	1	Long $\sim \mathcal{N}\left(30, 10^2\right)$	Long $\sim \mathcal{N}\left(10, 2^2\right)$	80

Table 1: Parameters for the baseline model with $n = 1$

In our numerical experiments, we use the scaling factor $n = \{1, 2, \ldots, 20\}$ to scale the arrival rate and the capacity simultaneously. More specifically, for a given scaling factor n, the arrival rate is $\lambda^{(n)} = n\lambda$ (where λ is given in Table 1), and the capacity is $C^{(n)} = nC = 40n$. For convenience, we refer to the settings with $n = 1, 2, 3$ as light traffic, $n = 3, 4, 5$ as medium traffic, and $n \geq 6$ as heavy traffic. In particular, the baseline model corresponds to the light traffic setting with $n = 1$.

Besides using different scaling factors, we also conduct our simulations by experimenting with different values for the load factors, the mean and variance of the advanced reservation distribution, the mean and variance of the service distribution, as well as the correlation coefficient between these two distributions. We refer readers to §6.3 for the detailed numerical results and discussions.

6.2 Performance Measure

We compare the long-run average revenue of the ϵ-CSP with the optimal solution provided by the benchmark linear program (1). The performance ratio is defined as follows.

$$\text{Performance Ratio} \triangleq \frac{\mathcal{R}(\epsilon\text{-}CSP)}{\text{The optimal objective value of (1)}} \leq \frac{\mathcal{R}(\epsilon\text{-}CSP)}{\mathcal{R}(OPT)}.$$

Note that the above-defined performance ratio of the ϵ-CSP is conservative in the sense that it understates the relative revenue ratio of the ϵ-CSP to the optimal policy. The actual performance of the ϵ-CSP is better than the performance ratio.

6.3 Numerical Results

Figure 7 shows that the performance ratio of the ϵ-CSP increases very quickly as the scaling factor n increases. When $n = 1$ (corresponding to $C = 40$), the optimality gap is around 12%. When $n = 10$ (corresponding to $C = 400$), the optimality gap is reduced significantly to less than 3%.

Table 2 shows the performance ratios of the ϵ-CSP with varying load factors, where the load factor is defined as the ratio of $\lambda\mathbb{E}[S]$ to C. (In our numerical experiments, we change the load

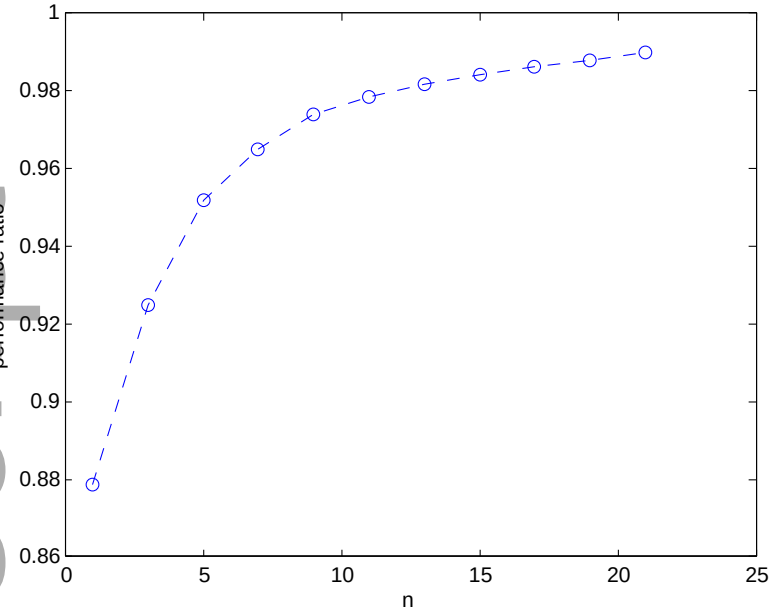


Figure 7: Performance ratios as the scaling factor n increases

factors by fixing λ and varying C.) Our results show that as the resource capacity decreases while keeping all other parameters fixed, the performance ratio drops steadily in the low traffic regime but drops sluggishly in the high traffic regime.

Tables 3–7 show the performance ratios of the ϵ-CSP by varying parameters of the advance reservation and the service time distributions (such as their means, variances and correlation coefficients). In summary, our results demonstrate that the ϵ-CSP performs near-optimally in the heavy traffic regime (with average optimality gap less than 4%), which is consistent with our theoretical development. Moreover, the ϵ-CSP performs reasonably well in the light and medium traffic regimes as well (with average optimality gap less than 10%). It is also worth noting that the performance of the ϵ-CSP is robust with respect to input distributions and parameters, which could be widely adopted in many practical scenarios.

Load Factor \ n	1	2	3	4	5	6	7	8	9	10
1.15	91.9%	94.3%	95.6%	96.4%	97.0%	97.5%	98.0%	98.4%	98.6%	99.0%
1.50	90.7%	93.6%	95.1%	95.9%	96.5%	97.0%	97.7%	98.0%	98.3%	98.6%
1.85	89.3%	92.6%	94.1%	95.0%	95.8%	96.4%	96.9%	97.4%	97.8%	98.1%
2.20	88.3%	91.7%	93.5%	94.3%	95.2%	95.8%	96.4%	97.0%	97.2%	97.8%
2.55	87.0%	90.8%	92.6%	93.8%	94.6%	95.2%	95.8%	96.4%	97.0%	97.4%
2.90	87.1%	91.0%	92.7%	93.9%	94.7%	95.4%	96.1%	96.7%	97.0%	97.5%
3.25	86.9%	90.6%	92.2%	93.6%	94.5%	95.2%	95.8%	96.3%	96.7%	97.4%
3.60	85.6%	89.8%	91.8%	93.2%	93.9%	94.8%	95.4%	96.1%	96.4%	96.9%
3.95	85.2%	89.5%	91.5%	92.9%	93.8%	94.4%	95.2%	95.7%	96.2%	96.7%
4.30	84.6%	88.9%	91.0%	92.4%	93.4%	94.1%	94.8%	95.6%	96.0%	96.6%
4.65	83.8%	88.3%	90.6%	92.0%	92.9%	93.6%	94.6%	95.0%	95.4%	96.0%
5.00	83.6%	88.1%	90.4%	91.7%	92.8%	93.4%	94.1%	94.9%	95.2%	96.1%

Table 2: Performance ratios for varying load factors

$\zeta_1 \setminus n$	1	2	3	4	5	6	7	8	9	10
1.0	87.9%	91.5%	93.3%	94.3%	95.2%	95.8%	96.3%	96.7%	97.3%	97.5%
1.1	88.0%	91.6%	93.3%	94.3%	95.1%	95.8%	96.3%	96.7%	97.2%	97.4%
1.2	88.0%	91.5%	93.2%	94.3%	95.1%	95.8%	96.4%	96.8%	97.3%	97.2%
1.3	87.8%	91.6%	93.2%	94.4%	95.0%	95.8%	96.3%	96.7%	97.3%	97.1%
1.4	88.0%	91.7%	93.2%	94.4%	95.2%	95.9%	96.4%	96.7%	97.4%	96.9%
1.5	88.0%	91.5%	93.2%	94.3%	95.0%	95.7%	96.2%	96.8%	97.2%	96.8%
1.6	88.2%	91.7%	93.3%	94.3%	95.1%	95.8%	96.3%	96.7%	97.3%	96.7%
1.7	87.9%	91.6%	93.2%	94.3%	95.3%	95.9%	96.4%	96.7%	97.1%	96.6%
1.8	88.0%	91.7%	93.1%	94.4%	95.1%	95.7%	96.4%	96.7%	97.1%	96.4%
1.9	87.7%	91.5%	93.3%	94.4%	95.1%	95.8%	96.2%	96.8%	97.0%	96.3%
2.0	87.8%	91.7%	93.3%	94.4%	95.2%	95.8%	96.3%	96.8%	96.9%	96.3%

Table 3: Performance ratios for varying adv. res. means ($\zeta_1 \times$baseline mean)

$\zeta_2 \setminus n$	1	2	3	4	5	6	7	8	9	10
1	88.1%	91.6%	93.2%	94.4%	95.1%	95.9%	96.4%	96.8%	97.3%	97.4%
0.87	87.7%	91.3%	93.2%	94.4%	95.0%	95.7%	96.3%	96.7%	97.3%	97.5%
0.88	87.6%	91.6%	93.2%	94.3%	95.1%	95.9%	96.3%	96.6%	97.3%	97.5%
0.89	87.6%	91.4%	93.2%	94.3%	95.0%	95.7%	96.2%	96.7%	97.3%	97.5%
0.90	87.8%	91.5%	93.2%	94.2%	95.1%	95.7%	96.3%	96.6%	97.3%	97.4%
0.92	87.6%	91.3%	93.1%	94.2%	95.1%	95.8%	96.3%	96.7%	97.3%	97.4%
0.95	87.6%	91.4%	93.0%	94.3%	95.0%	95.7%	96.3%	96.7%	97.4%	97.3%
0.99	87.3%	91.3%	93.1%	94.2%	95.0%	95.7%	96.2%	96.7%	97.2%	97.4%
1.04	87.5%	91.4%	93.1%	94.3%	95.0%	95.7%	96.3%	96.7%	97.3%	97.4%
1.07	87.5%	91.3%	93.0%	94.1%	95.0%	95.7%	96.3%	96.6%	97.3%	97.4%

Table 4: Performance ratios for varying adv. res. variances ($\zeta_2 \times$baseline variance)

$\zeta_3 \setminus n$	1	2	3	4	5	6	7	8	9	10
1.0	87.8%	91.5%	93.4%	94.4%	95.2%	95.8%	96.3%	96.7%	97.4%	97.5%
1.1	87.9%	91.5%	93.2%	94.4%	95.2%	95.8%	96.3%	96.8%	97.3%	97.6%
1.2	88.0%	91.5%	93.2%	94.4%	95.1%	95.8%	96.4%	96.8%	97.4%	97.6%
1.3	87.8%	91.6%	93.4%	94.4%	95.2%	95.9%	96.3%	96.9%	97.5%	97.8%
1.4	88.3%	91.8%	93.5%	94.6%	95.3%	96.0%	96.7%	97.0%	97.6%	97.9%
1.5	88.5%	92.0%	93.6%	94.6%	95.4%	95.9%	96.6%	97.0%	97.7%	98.0%
1.6	88.8%	92.1%	93.8%	94.9%	95.6%	96.3%	96.7%	97.2%	97.8%	98.1%
1.7	88.7%	92.0%	93.9%	94.8%	95.5%	96.1%	96.9%	97.2%	97.9%	98.2%
1.8	88.9%	92.0%	93.7%	94.8%	95.6%	96.3%	96.8%	97.1%	97.9%	98.2%
1.9	89.0%	92.2%	93.8%	95.0%	95.7%	96.1%	96.8%	97.2%	98.0%	98.2%
2.0	88.9%	92.4%	93.8%	94.9%	95.6%	96.4%	96.8%	97.4%	97.9%	98.2%

Table 5: Performance ratios for varying service time means ($\zeta_3 \times$baseline mean)

$\zeta_4 \setminus n$	1	2	3	4	5	6	7	8	9	10
0.86	88.3%	91.7%	93.2%	94.4%	95.2%	95.8%	96.2%	96.7%	97.2%	97.0%
0.87	88.1%	91.6%	93.2%	94.5%	95.1%	95.8%	96.3%	96.6%	97.2%	97.0%
0.88	88.3%	91.8%	93.4%	94.3%	95.2%	95.8%	96.3%	96.7%	97.2%	97.1%
0.89	88.1%	91.6%	93.3%	94.4%	95.1%	95.9%	96.4%	96.6%	97.2%	97.1%
0.90	88.2%	91.6%	93.3%	94.4%	95.1%	95.7%	96.3%	96.7%	97.3%	97.3%
0.92	88.1%	91.6%	93.1%	94.4%	95.1%	95.7%	96.2%	96.7%	97.3%	97.4%
0.95	87.9%	91.6%	93.5%	94.2%	95.2%	95.8%	96.4%	96.7%	97.4%	97.4%
0.99	87.8%	91.5%	93.2%	94.4%	95.1%	95.7%	96.4%	96.7%	97.3%	97.4%
1.04	87.9%	91.7%	93.3%	94.3%	95.1%	95.8%	96.4%	96.8%	97.4%	97.5%
1.07	88.0%	91.5%	93.3%	94.4%	95.2%	95.9%	96.3%	96.8%	97.3%	97.7%

Table 6: Performance ratios for varying service time variances ($\zeta_4 \times$baseline variance)

$\zeta_5 \setminus n$	1	2	3	4	5	6	7	8	9	10
-0.9	88.3%	91.8%	93.4%	94.5%	95.3%	96.0%	96.5%	97.0%	97.4%	97.9%
-0.6	87.9%	91.7%	93.3%	94.4%	95.2%	95.8%	96.4%	96.7%	97.2%	97.7%
-0.3	88.0%	91.6%	93.1%	94.1%	95.0%	95.8%	96.4%	96.7%	97.2%	97.6%
0.3	87.8%	91.5%	93.2%	94.4%	95.2%	95.8%	96.3%	96.8%	97.2%	97.7%
0.6	88.1%	91.6%	93.2%	94.2%	95.2%	95.7%	96.5%	96.7%	97.3%	97.8%
0.9	88.2%	91.9%	93.5%	94.6%	95.3%	95.9%	96.5%	96.9%	97.5%	97.9%

Table 7: Performance ratios for varying correlations between adv. res. and service times (ζ_5)

7 Conclusion

In this paper, we have studied an important class of revenue management problems with reusable resources and advanced reservations. We have devised an effective and efficient admission control policy termed the ϵ-CSP that is proven to be asymptotically optimal when both the demand rate and the capacity grow large. We have also explicitly characterized its convergence rate. This class of revenue management problems finds many applications in real life, such as hotel management, cloud computing, car rental, workforce management, and call centers. In addition, the methods developed also contribute to the existing queueing literature.

To close this paper, we would like to point out three potential research directions. (a) One may consider a dynamic pricing model and derive similar policies. Assume that there is a single-class time-homogenous Poisson arrival process with rate λ. Each customer's reservation and service-time are drawn from D and S, respectively. The system offers a price from a fixed price menu $\{r_1, \ldots, r_n\}$ to an arriving customer with d and s, depending on the current state. The state is characterized by the booking profile, d, and s. Moreover, one can introduce a reservation price distribution denoted by R. The customer only accepts the offer if the price offered falls below the reservation price. (b) Gallego and Hu (2014) consider a competition network model of perishable resources. One could also study a counterpart model of reusable resources. Each firm has a fixed capacity of reusable resources and competes in setting prices to sell them. Assuming deterministic customer arrival rates, one can potentially show that any equilibrium strategy has a simple structure, and then show that there exists a similar asymptotic equilibrium strategy in a stochastic version where the arrival rates and the capacity are scaled together to infinity. (c) To capture seasonality of demands, one could also consider the model studied in this paper with non-homogeneous Poisson arrivals. However, this may require new methodologies to be developed.

Acknowledgment. The authors thank the department editor Qi Annabelle Feng, the anonymous senior editor, and the anonymous referees for their constructive comments, which helped significantly improve both the content and the exposition of this paper. The authors also thank Ana Radovanovic, Yuan Zhong, Armando Bernal for their valuable discussions. The author's names are alphabetically ordered, and the third author is the corresponding author. The second author's research is partially supported by NSF grants DMS-0732175 and CMMI-0846554 (CAREER Award), an AFOSR award FA9550-08-1-0369, an SMA grant and the Buschbaum Research Fund of MIT. The third author's research is partially supported by NSF grants CMMI-1362619 and CMMI-1451078 and CMMI-1634505.

Appendix

Omitted Proofs of Technical Lemmas and Propositions

Proof of Proposition 1. Consider any policy $\pi \in \Pi$. For the customer who is the nth of being allocated with a specific item of the resource since time 0, we denote by $t^{(n)}$ his service request time, $d^{(n)}$ his advanced reservation time, and $s^{(n)}$ his service time. Therefore, $t^{(n)} + d^{(n)}$ is non-decreasing in $n \in \mathbb{N}$.

Now, we prove this proposition by induction. First, the statement of this proposition is trivially true for the customer with index 1, since all items of the resource have not been allocated to any customer before this customer is allocated with a specific item.

Second, suppose the statement of this proposition is true for customers with indexes $1, \cdots, n$. We now prove that the statement of this proposition is still true for the customer with index $n+1$. For any $t \in \left[t^{(n+1)} + d^{(n+1)}, t^{(n+1)} + d^{(n+1)} + s^{(n+1)}\right)$, we have

$$\begin{aligned}
& \sum_{n'=1}^{n} \mathbf{1}\left\{t \in \left[t^{(n')} + d^{(n')}, t^{(n')} + d^{(n')} + s^{(n')}\right)\right\} \\
\leq & \sum_{n'=1}^{n} \mathbf{1}\left\{t^{(n+1)} + d^{(n+1)} \in \left[t^{(n')} + d^{(n')}, t^{(n')} + d^{(n')} + s^{(n')}\right)\right\} \\
= & \sum_{n'=1}^{n+1} \mathbf{1}\left\{t^{(n+1)} + d^{(n+1)} \in \left[t^{(n')} + d^{(n')}, t^{(n')} + d^{(n')} + s^{(n')}\right)\right\} - 1 \\
\leq & \sum_{n'=1}^{\infty} \mathbf{1}\left\{t^{(n+1)} + d^{(n+1)} \in \left[t^{(n')} + d^{(n')}, t^{(n')} + d^{(n')} + s^{(n')}\right)\right\} - 1 \\
\leq & C - 1.
\end{aligned}$$

The first inequality follows from the property that $t \geq t^{(n+1)} + d^{(n+1)} \geq t^{(n)} + d^{(n)}$. The third inequality follows from the feasibility property that under any policy $\pi \in \Pi$, given any time t, the total number of admitted customers who use the service at time t is no more than capacity C. Therefore, for the customer with index $n+1$, at each point of time during his service interval $\left[t^{(n+1)} + d^{(n+1)}, t^{(n+1)} + d^{(n+1)} + s^{(n+1)}\right)$, there is always at least one unit of resource that is available to serve him.

In addition, we notice that since every customer with index in $\{1, \cdots, n\}$ keeps on using the same item of the resource during his service interval, at time $t^{(n+1)} + d^{(n+1)}$, an item of the resource not occupied by any customer with index in $\{1, \cdots, n\}$ will not be occupied by any of these customers after time $t^{(n+1)} + d^{(n+1)}$. We also notice that no customer with index $n' > n+1$ is allocated with a specific item of the resource before the customer with index $n+1$ is allocated. Therefore, any item that is not occupied by any customer with index in $\{1, \cdots, n\}$ at time $t^{(n+1)} + d^{(n+1)}$ can be allocated to the customer with index $n+1$ at his service start time $t^{(n+1)} + d^{(n+1)}$, with the guarantee that this item is able to serve him during his entire service interval $\left[t^{(n+1)} + d^{(n+1)}, t^{(n+1)} + d^{(n+1)} + s^{(n+1)}\right)$. □

Proof of Lemma 1. Consider the following LP

$$\max_{\{\alpha^{\pi}_{dsk}\}} \sum_{k=1}^{M} \sum_{d,s} r_k \alpha^{\pi}_{dsk} \lambda_{dsk} s, \qquad \text{s.t.} \qquad \sum_{k=1}^{M} \sum_{d,s} \alpha^{\pi}_{dsk} \lambda_{dsk} s \le C, \ 0 \le \alpha_{dsk} \le 1, \ \forall d, s, k. \quad (12)$$

Note the LP defined in (1) differs from the LP defined in (12) by changing the right hand side of the capacity constraint to $(1-\epsilon)C$. Suppose the optimal solution of the LP defined in (12) is $\{\hat{\alpha}_{dsk}\}$. Now consider $\{\tilde{\alpha}_{dsk}\} = \{(1-\epsilon)\hat{\alpha}_{dsk}\}$. Since

$$\sum_{k=1}^{M} \sum_{d,s} \tilde{\alpha}_{dsk} \lambda_{dsk} s = (1-\epsilon) \sum_{k=1}^{M} \sum_{d,s} \hat{\alpha}_{dsk} \lambda_{dsk} s \le (1-\epsilon) C, \quad (13)$$

$\{\tilde{\alpha}_{dsk}\}$ is a feasible solution to (1). Then we have

$$\sum_{k=1}^{M} \sum_{d,s} r_k \alpha^{*}_{dsk} \lambda_{dsk} s \ge \sum_{k=1}^{M} \sum_{d,s} r_k \tilde{\alpha}_{dsk} \lambda_{dsk} s = (1-\epsilon) \sum_{k=1}^{M} \sum_{d,s} r_k \hat{\alpha}_{dsk} \lambda_{dsk} s \ge (1-\epsilon) \mathcal{R}(OPT). \quad (14)$$

The last inequality holds since the optimal objective value in (12) provides an upper bound on the optimal expected revenue rate, since the capacity constraint is enforced on expectation whereas in the original problem this capacity constraint has to hold, for each sample path. This completes the proof. □

Proof of Lemma 2. If $r \le 0$, we focus on the interval $(\lceil r \rceil - 1, \lceil r \rceil]$ and its preceding interval $(\lceil r \rceil - 2, \lceil r \rceil - 1]$. The arrival process in $(\lceil r \rceil - 2, \lceil r \rceil - 1]$ follows a Poisson process with rate ρ. Each arrival has γ probability of starting services immediately in $(\lceil r \rceil - 2, \lceil r \rceil - 1]$, and $1-\gamma$ probability of starting services in 1 unit of time in $(\lceil r \rceil - 1, \lceil r \rceil]$. By the Poisson splitting argument, the pre-arrivals in $(\lceil r \rceil - 1, \lceil r \rceil]$ follow a Poisson process with rate $(1-\gamma)\rho$. Using a similar argument, we conclude that the pre-arrival process in $(\lceil r \rceil - 1, \lceil r \rceil]$ induced by customers arriving to the system in $(\lceil r \rceil - 1, \lceil r \rceil]$ follows a Poisson process with rate $\gamma\rho$. Note that these two processes are independent of each other since they are generated by customers arriving in disjoint intervals. Now merge these two pre-arrival processes, and the resulting pre-arrival process in $(\lceil r \rceil - 1, \lceil r \rceil]$ follows a Poisson process with rate $(1-\gamma)\rho + \gamma\rho = \rho$.

If $r \in (0, 1]$, focus on the interval $(0, 1]$ and its preceding interval $(-1, 0]$. By an argument similar to the above, there is a Poisson process of pre-arrivals with rate $(1-\gamma)\rho$ induced by customers arriving in $(-1, 0]$. There is also a Poisson process with rate $\gamma\rho$ induced by customers arriving in $(0, 1]$. However, the latter process consists of post-arrivals. Thus, the resulting pre-arrivals at time 0 over $(0, 1]$ follow a Poisson process with rate $(1-\gamma)\rho$.

Finally, since the maximum reservation time is 1, it is impossible for customers arriving prior to 0 to start service at any time greater than 1. Thus, the rate of pre-arrivals from 1 onwards is 0. This completes the proof. □

Proof of Lemma 3. Suppose that a customer arrives at time 0 in the steady state and requests the service to commence immediately ($D = 0$), i.e., requesting the service interval $(0, 1]$. Focus solely on the pre-arrivals as seen from 0. By Lemma 2, the pre-arrivals over the time interval $(-1, 0]$ follow a Poisson process with rate ρ, denoted by N_0. However, this implies that, over the

time interval $(0,1]$, the customers depart the system following a Poisson process with rate ρ (a shift of N_0 by 1 unit of time). Let $\tilde{N}_0$ be the mirror image of the departure process induced by N_0 over $(0,1]$. By Lemma 2, we also know that the pre-arrivals over $(0,1]$ follow a Poisson process with rate $(1-\gamma)\rho$. We denote this pre-arrival process by N_1. Figure 2 shows the pre-arrival and departure processes.

Consider now the number of customers in the system at some time r. These fall exactly into one of the two types; customers that start service over $(0,r]$ and customers that start service over $(r-1,0]$ and will depart over $(r,1]$. It follows that the number of customers in service at time $r \in (0,1]$ can be expressed as $\tilde{N}_0(1-r) + N_1(r)$. Specifically, in time r the number of departures over $(r,1]$ (equal to $\tilde{N}_0(1-r)$) captures customers starting service before 0, and still in the system at time r. In addition, the number of pre-arrivals over $(0,r]$ (equal to $N_1(r)$) captures customers arriving before 0, starting service over $(0,r]$ and still being served in time r. The sum of the two is exactly equal to the total number of customers in the system at time r. Note that by the Poisson splitting argument, it follows that $\tilde{N}_0$ and N_1 are independent of each other. The virtual blocking probability is expressed in terms of the maximum of the sum of these two Poisson counting processes running towards each other (see Figure 3), i.e.,

$$P_0 \triangleq \mathbb{P}(B \mid D=0) = \mathbb{P}\left(\max_{r\in[0,1]}\left\{\tilde{N}_0(1-r;\rho) + N_1(r;(1-\gamma)\rho)\right\} \geq C\right).$$

Consider now the case that the arriving customer requests the service to commence in $D=1$ unit of time, i.e., the service will cover the interval $(1,2]$. The departure process in $(1,2]$ is a shift of the pre-arrival process N_1 in $(0,1]$ by 1 unit of time, and its mirror image is denoted by $\tilde{N}_1$. Moreover, by Lemma 2, the pre-arrival process N_2 in $(1,2]$ has rate 0. Thus, we have

$$P_1 \triangleq \mathbb{P}(B \mid D=1) = \mathbb{P}\left(\max_{r\in[0,1]}\left\{\tilde{N}_1(1-r;(1-\gamma)\rho) + N_2(r;0)\right\} \geq C\right) = \mathbb{P}\left(\tilde{N}_1(1;(1-\gamma)\rho) \geq C\right).$$

This completes the proof. □

Proof of Lemma 4. Note again that for each $r \in [0,1]$, $\tilde{N}_0(1-r;\theta_1\rho) + N_1(r;\rho)$ is equal to the sum of the number of occurrences of N_0 over $(1-r,1]$ and the number of occurrences of N_1 over $[0,r)$. Also observe that the value of X is obtained either at time 0 or upon on occurrence of N_1. Now conditioning on $\mathcal{R}_n = \omega_n$ (a specific realization of the random walk $\mathcal{R}_n$), and consider the l^{th} occurrence of N_1 ($l \in \{0,\ldots,n\}$), at time, say r. Then we have (see Figure 3),

$$\begin{aligned}
&\tilde{N}_0(1-r;\rho) + N_1(r;\theta_1\rho) = (\#\text{ up-steps before and including } l + \#\text{ down-steps after } l)\\
=\ &(\#\text{ up-steps before and including } l - \#\text{ down-steps before and including } l)\\
&+(\#\text{ down-steps before and including } l + \#\text{ down-steps after } l).
\end{aligned}$$

The first term is exactly the location of the random walk after l steps and the second expression is exactly G_n. Since X is the maximum of the above sum over all arrivals $l = 0,1,\ldots,n$, it follows that indeed $X_n \mid (\mathcal{R}_n = \omega_n) = (G_n + M_n) \mid (\mathcal{R}_n = \omega_n)$, from which the result follows. □

Proof of Lemma 5. Define the stopping time τ as follows:

$$\tau \triangleq \inf\{t \geq 1 : S_t \leq -a \text{ or } S_t \geq b\}.$$

It is straightforward to check the following two conditions,

$$\mathbb{E}(\tau) \leq \infty, \qquad \mathbb{E}(|E_{t+1} - E_t| \mid \mathcal{F}_t) \leq 2, \qquad \forall t \in \tau. \tag{15}$$

The Wald's identity (see Karlin and Taylor (1981))

$$G_n(\theta) \triangleq \frac{e^{\theta S_n}}{[\phi(\theta)]^n} \tag{16}$$

is a martingale where the moment generating function $\phi(\theta) \triangleq \mathbb{E}(e^{\theta Y}) \geq 1$. First we compute $\hat{\theta}$ that solves the equation $\mathbb{E}(e^{\hat{\theta} Y}) = 1$, i.e.,

$$\mathbb{E}(e^{\hat{\theta} Y}) = pe^{\hat{\theta}} + qe^{\hat{\theta}} = 1 \qquad \Rightarrow \qquad e^{\hat{\theta}} = \frac{q}{p}. \tag{17}$$

By Optional Sampling Theorem (see Karlin and Taylor (1981)),

$$\mathbb{E}\left[\frac{e^{\hat{\theta} S_\tau}}{[\phi(\hat{\theta})]^\tau}\right] = \mathbb{E}\left[e^{\hat{\theta} S_\tau}\right] = \mathbb{E}\left[e^{\hat{\theta} S_0}\right] = 1. \tag{18}$$

This leads to

$$\mathbb{P}(S_\tau \geq b) \underbrace{\mathbb{E}(e^{\hat{\theta} S_\tau} \mid S_\tau \geq b)}_{E_b} + (1 - \mathbb{P}(S_\tau \geq b)) \underbrace{\mathbb{E}(e^{\hat{\theta} S_\tau} \mid S_\tau \leq -a)}_{E_a} = 1. \tag{19}$$

Thus, we have

$$\mathbb{P}(S_\tau \geq b) = \frac{1 - E_a}{E_b - E_a} = \frac{1 - e^{-\hat{\theta} a}}{e^{\hat{\theta} b} - e^{-\hat{\theta} a}} = \frac{1 - \left(\frac{q}{p}\right)^{-a}}{\left(\frac{q}{p}\right)^{b} - \left(\frac{q}{p}\right)^{-a}}. \tag{20}$$

Let $S_\tau^a \triangleq S_\tau$ be the stopping time location of the process. Let B_a be the event that the random walk hits b before $-a$. Observe that $\mathbb{P}(B_a) = \mathbb{P}(S_\tau^a \geq b)$ and also note that $B_i \subset B_{i+1}$ for all i. Define $B = \bigcup_{i=1}^{\infty} B_i$, i.e., there exists an i that the random walk hits b before $-i$. Therefore $\mathbb{P}(M_\infty \geq b) = \mathbb{P}(B)$. By properties of probability measures, we have

$$\mathbb{P}(M_\infty \geq b) \quad = \quad \mathbb{P}(\bigcup_{i=1}^{\infty} B_i) = \lim_{a \to \infty} \mathbb{P}(B_a) = \lim_{a \to \infty} \left(\frac{1 - \left(\frac{q}{p}\right)^{-a}}{\left(\frac{q}{p}\right)^{b} - \left(\frac{q}{p}\right)^{-a}} \right) = \left(\frac{p}{q}\right)^{b}. \tag{21}$$

This completes the proof. □

Proof of Proposition 2. First we establish an upper bound for $\mathbb{P}(X \geq C)$. Let M_∞ be the maximum level attained by the infinite-step random walk defined above. Since the random walk has a negative drift, it follows from Lemma 5 above that $\mathbb{P}(M_\infty \geq -\log \rho / \log \theta_1) \leq 1/\rho$. (Note

that $\theta_1 < 1$, so $-\log\rho/\log\theta_1 > 0$.) Now, we have

$$\begin{aligned}
\mathbb{P}(X_n \geq C) &= \mathbb{P}\left(G_n + M_n \geq C\right) \qquad (22)\\
&= \mathbb{P}\left(G_n + M_n \geq C \bigcap M_n \geq -\frac{\log\rho}{\log\theta_1}\right) + \mathbb{P}\left(G_n + M_n \geq C \bigcap M_n < -\frac{\log\rho}{\log\theta_1}\right)\\
&\leq \mathbb{P}\left(M_n \geq -\frac{\log\rho}{\log\theta_1}\right) + \mathbb{P}\left(G_n \geq C + \frac{\log\rho}{\log\theta_1}\right)\\
&\leq \mathbb{P}\left(M_\infty \geq -\frac{\log\rho}{\log\theta_1}\right) + \mathbb{P}\left(G_n \geq C + \frac{\log\rho}{\log\theta_1}\right)\\
&\leq \frac{1}{\rho} + \mathbb{P}\left(G_n \geq C + \frac{\log\rho}{\log\theta_1}\right).
\end{aligned}$$

The first equality follows from Lemma 4. The first inequality follows from the fact that $M_\infty \geq M_n$ almost surely. The second inequality follows from Lemma 5 above. Since G_n is distributed as $(\tilde{N}_0(1;\rho) \mid \mathcal{N} = n)$, we get from (22) that,

$$\begin{aligned}
\mathbb{P}(X \geq C) &= \sum_{n=1}^{\infty} \mathbb{P}(X_n \geq C)\mathbb{P}(\mathcal{N} = n)\\
&\leq \frac{1}{\rho} + \sum_{n=1}^{\infty} \mathbb{P}\left(G_n \geq C + \frac{\log\rho}{\log\theta_1}\right)\mathbb{P}(\mathcal{N} = n)\\
&= \frac{1}{\rho} + \mathbb{P}\left(\tilde{N}_0(1;\rho) \geq C + \frac{\log\rho}{\log\theta_1}\right)\\
&= \frac{1}{\rho} + \mathbb{P}\left(\text{Poisson}(\rho) \geq C + \frac{\log\rho}{\log\theta_1}\right)\\
&= \frac{1}{\rho} + \mathbb{P}\left(\text{Poisson}(\rho) \geq C - \frac{\log\rho}{\log\theta_1^{-1}}\right)\\
&\leq \frac{1}{\rho} + \mathbb{P}\left(\text{Poisson}(\rho) \geq \frac{\rho}{1-\epsilon} - \frac{\log\rho}{\log\theta_1^{-1}}\right)\\
&\leq \frac{1}{\rho} + \mathbb{P}\left(\text{Poisson}(\rho) \geq \frac{\rho}{1-\epsilon} - \frac{\log(1+\rho)}{\log\theta_1^{-1}}\right).
\end{aligned}$$

The second inequality follows Equation (2) that $\rho \leq (1-\epsilon)C$. The third inequality follows the property that $\rho \leq 1 + \rho$ and the property that $\theta_1 < 1$ implies $\log\theta_1^{-1} > 0$.

Next, we establish an upper bound for $\mathbb{P}(Y \geq C)$. Note that $\theta_1 < 1$. We define

$$\bar{Y} = \max_{r\in[0,1]} \left\{\bar{N}_1(1-r;\rho) + N_2(r;\theta_2\rho)\right\}.$$

Thus, $\bar{Y}$ stochastically dominates Y. Following the same argument that we make above for establishing an upper bound for $\mathbb{P}(X \geq C)$, we have

$$\begin{aligned}
\mathbb{P}(Y \geq C) &\leq \frac{1}{\rho} + \mathbb{P}\left(\text{Poisson}(\rho) \geq \frac{\rho}{1-\epsilon} - \frac{\log(1+\rho)}{\log\theta_2^{-1}}\right)\\
&\leq \frac{1}{\rho} + \mathbb{P}\left(\text{Poisson}(\rho) \geq \frac{\rho}{1-\epsilon} - \frac{\log(1+\rho)}{\log\theta_1^{-1}}\right).
\end{aligned}$$

The second inequality follows the property that $\theta_2 \le \theta_1$ and the property that $\log(1+\rho) \ge 0$.

Now, we establish an upper bound for $\mathbb{P}\left(\text{Poisson}(\rho) \ge \frac{\rho}{1-\epsilon} - \frac{\log(1+\rho)}{\log \theta_1^{-1}}\right)$. Using the formula of Chernoff bound of the upper tail for a Poisson random variable, we have

$$\begin{aligned}\mathbb{P}\left(\text{Poisson}(\rho) \ge \frac{\rho}{1-\epsilon} - \frac{\log(1+\rho)}{\log \theta_1^{-1}}\right) &= \mathbb{P}\left(\frac{1}{\rho}\left(\text{Poisson}(\rho) - \rho\right) \ge \frac{\epsilon}{1-\epsilon} - \frac{\log(1+\rho)}{\rho \log \theta_1^{-1}}\right) \\ &\le \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\rho},\end{aligned}$$

where

$$\delta = \left(\frac{\epsilon}{1-\epsilon} - \frac{\log(1+\rho)}{\rho \log \theta_1^{-1}}\right)^{+}.$$

Therefore,

$$\max\left\{\mathbb{P}(X \ge C), \mathbb{P}(Y \ge C)\right\} \le \frac{1}{\rho} + \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\rho}.$$

□

Proof of Lemma 6. Lemma 6 is a generalized version of Lemma 2. For $r \le 0$, consider the time interval $(\lceil r \rceil - 1, \lceil r \rceil]$. By arguments similar to those used in Lemma 2, for each $l \in [0, u]$, the interval $(\lceil r \rceil - 1 - l, \lceil r \rceil - l]$ generates a stream of pre-arrivals over $(\lceil r \rceil - 1, \lceil r \rceil]$ that follow a Poisson process of rate $\gamma_l \rho$. These processes are independent of each other and the overall merged process has rate $\rho = \gamma_0 \rho + \gamma_1 \rho + \ldots + \gamma_u \rho$. For $\lceil r \rceil = d$ for $d \in [1, u]$, then the pre-arrivals prior to t over $(\lceil r \rceil - 1, \lceil r \rceil]$ are induced by arriving customers over the intervals $(\lceil r \rceil - l - 1, \lceil r \rceil - l]$, for $l \in [d, u]$, and the total rate is $\gamma_d \rho + \gamma_{d+1}\rho + \ldots + \gamma_u \rho$. Note again that the rate $\gamma_i \rho$ is induced from the Poisson arrival stream of customers over $(\lceil r \rceil - l - 1, \lceil r \rceil - l]$ who wish to start in l units of time. Since we only consider pre-arrivals prior to t, the terms $\gamma_{d-1}\rho, \gamma_{d-2}\rho, \ldots, \gamma_0 \rho$ are missing. □

Proof of Lemma 7. By Lemma 6, for each $d \in [0, u]$, the pre-arrival process N_d over the interval $(d-1, d]$ follows a Poisson process with rate $\rho_d = \rho\left(1 - \sum_{i=0}^{d-1} \gamma_i\right)$. This implies that over the interval $(d, d+1]$, the customers depart the system following a Poisson process with rate ρ_d (a shift of N_d by 1 unit of time). Let $\tilde{N}_d$ be the mirror image of the departure process induced by N_d over $(d, d+1]$, and therefore $\tilde{N}_d$ has the same rate ρ_d. The rest of arguments are identical to that of Lemma 3. □

Proof of Proposition 3. First we assume that $\gamma_0 > 0$. By Lemma 7, we have that $\rho_0 > \rho_1$ and $\rho_d \ge \rho_{d+1}$ for each $d \in [1, u]$. By Proposition 2, it follows that, for all $d \in [0, u]$,

$$P_d \le \frac{1}{\rho} + \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\rho},$$

where δ is defined by (3).

In fact, we can relax the assumption of $\gamma_0 > 0$. If $\gamma_0 = 0$, it implies that over the interval $(0, 1]$ (recall that the customer arrives at time 0 in the steady state), the departure rate is equal

to the pre-arrival rate, i.e., $\rho_0 = \rho_1 = \rho$. Proposition 2 cannot be applied under this case. However, the fact that $\gamma_0 = 0$ implies that no arriving customers will start the service right away. Therefore, we do not have to consider the probability P_0 in the expression of P. Let the index $i = \min\{d : \gamma_d > 0\}$. Then we have $\gamma_d = 0$ for $d \in [0, i-1]$, by the same argument, we can ignore the probabilities $P_0, \ldots, P_{i-1}$. Therefore, the above results that we prove in this proposition still hold. This completes the proof. □

Proof of Lemma 8. For each set $s \in [1, v]$, and $d \in [0, u]$, the pre-arrivals prior to t over $(d-s, d-s+1]$ (i.e., N_d^s) are induced by arriving customers over the intervals $(d-s-i, d-s-i+1]$ for $i = (d-s+1)^+, \ldots, u$, and the total rate ρ_d^s is therefore

$$\gamma^s_{(d-s+1)^+}\rho_0^s + \ldots + \gamma_u^s\rho_0^s = \rho_0^s\left(1 - \sum_{i=0}^{d-s}\gamma_i^s\right) = \kappa_s\rho\left(1 - \sum_{i=0}^{d-s}\gamma_i^s\right). \tag{23}$$

Note again that the rate γ_i^s is induced from the Poisson arrival stream of customers over $(d-s-i, d-s-i+1]$ who wish to start in i units of time. It follows from the Poisson splitting argument that N_d^s and $N_{d'}^{s'}$ are independent if $(d, s) \neq (d', s')$. This completes the proof. □

Proof of Lemma 9. The assumption $\gamma_0^s > 0$ for some $s \in [1, v]$ implies that in the interval $(0, 1]$, the total departure rate is strictly greater than the total pre-arrival rate, i.e., $\sum_{s=1}^v \rho_0^s > \sum_{s=1}^v \rho_s^s$. For subsequent intervals $(d, d+1]$ for $d \geq 1$, we have $\sum_{s=1}^v \rho_d^s \geq \sum_{s=1}^v \rho_{d+s}^s$. Therefore the conditions of Proposition 2 are satisfied. Proposition 2 implies that

$$\mathbb{P}(A_0 \geq C) \leq \frac{1}{\rho} + \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\rho,$$

where δ is defined by (3). This completes the proof. □

Proof of Proposition 4. First we assume that $\gamma_0^s > 0$ for some $s \in [1, v]$. For each $s \in [1, v]$, by applying union bound and Lemma 9, we have for all $d \in [0, u]$ and $s \in [1, v]$,

$$P_d^s \leq \frac{1}{\rho} + \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\rho,$$

where δ is defined by (3).

We then drop the assumption that $\gamma_0^s > 0$ for some $s \in [1, v]$. Suppose now $\gamma_0^s = 0$ for all $s \in [1, v]$. This implies that no arriving customers at time 0 will start the service over $(0, 1]$, and hence we can ignore the blocking probability over this interval. Therefore, the above result continues to hold by following the same arguments. □

Proof of Theorem 3. First, we show that for each solution $\{r_k\}$ of (NLP2), we can construct a solution of (NLP1) with the same objective value. Specifically, consider a solution $\{r'_k, \alpha'_{dsk}\}$ such that $r'_k = r_k$ and $\alpha'_{dsk} = 1$ if and only if $\sum_{d,s} \lambda_{dsk}(r_k)s > 0$. It can be verified that the resulting solution is feasible for (NLP1) and has the same objective value.

Next, we show how to map optimal solution $\{r_k^*, \alpha_{dsk}^*\}$ of (NLP1) to a feasible solution of (NLP2) with the same objective function. For each $k = 1, \ldots, M' - 1$, set $r_k = r_k^*$, and

for each $k = M' + 1, \ldots, M$, set $r_k = r_\infty$. It is clear that, for each $k \neq M'$, the resulting contributions to the objective value and constraint in (NLP2) are the same as in (NLP1). Consider now the possible fractional value $\alpha^*_{M'}$ for class M'. The respective contribution of class M' to the objective value is $\sum_{d,s} r^*_{M'} \alpha^*_{dsM'} \lambda_{ijM'}(r^*_{M'})j$. Similarly, the contribution to constraint in (1) is $\sum_{d,s} \alpha^*_{dsM'} \lambda_{dsM'}(r^*_{M'})s$. Thus, it is sufficient to show that there exists a price $r_{M'}$ such that $\sum_{d,s} r_{M'} \lambda_{dsM'}(r_{M'})s \geq \sum_{d,s} r^*_{M'} \alpha^*_{dsM'} \lambda_{dsM'}(r^*_{M'})s$ and $\sum_{d,s} \lambda_{dsM'}(r_{M'})s \leq \sum_{d,s} \alpha^*_{dsM'} \lambda_{dsM'}(r^*_{M'})s$.

Since $\sum_{d,s} r^*_{M'} \lambda_{dsM'}(r^*_{M'})s \geq \sum_{d,s} r^*_{M'} \alpha^*_{dsM'} \lambda_{dsM'}(r^*_{M'})s$, by the properties of $\lambda_{dsM'}(r_{M'})$, we know that there exists $\bar{r} \in [r_{M'}, r_\infty)$ such that $\sum_{d,s} \bar{r} \lambda_{dsM'}(\bar{r})s = \sum_{d,s} r^*_{M'} \alpha^*_{dsM'} \lambda_{dsM'}(r^*_{M'})j$. Note that $\bar{r} \geq r^*_{M'}$, and thus we obtain $\sum_{d,s} r^*_{M'} \lambda_{dsM'}(\bar{r})s \leq \sum_{d,s} \bar{r} \lambda_{dsM'}(\bar{r})s = \sum_{d,s} r^*_{M'} \alpha^*_{dsM'} \lambda_{dsM'}(r^*_{M'})s$. We conclude that $\sum_{d,s} \lambda_{dsM'}(\bar{r})s \leq \sum_{d,s} \alpha^*_{dsM'} \lambda_{dsM'}(r^*_{M'})s$, which completes the proof. □

References

Adelman, D. 2006. A simple algebraic approximation to the Erlang loss system. *Operations Research Letters* **36**(4) 484–491.

Adelman, D. 2007. Price-directed control of a closed logistics queuing network. *Operations Research* **55**(6) 1022–1038.

Begen, M. A., R. Levi, M. Queyranne. 2012. A sampling-based approach to appointment scheduling. *Operations research* **60**(3) 675–681.

Begen, M. A., M. Queyranne. 2011. Appointment scheduling with discrete random durations. *Mathematics of Operations Research* **36**(2) 240–257.

Besbes, O., C. Maglaras. 2009. Revenue optimization for a make-to-order queue in an uncertain market environment. *Operations Research* **57**(6) 1438–1450.

Borgs, C., O. Candogan, J. Chayes, I. Lobel, H. Nazerzadeh. 2014. Optimal multiperiod pricing with service guarantees. *Management Science* **60**(7) 1792–1811.

Boxma, O. J. 1984. $M/G/\infty$ tandem queues. *Stochastic Processes and their Applications* **18** 153–164.

Burman, D. Y., J. P. Lehoczky, Y. Lim. 1984. Insensitivity of blocking probabilities in a circuit-switching network. *J. Appl. Probab.* **21**(4) 850–859.

Chen, Y., C. Shi. 2016. Optimal pricing policy for service systems with reusable resources and forward-looking customers. Working paper, University of Michigan.

Coffman-Jr, E. G., P. Jelenkovic, B. Poonen. 1999. Reservation probabilities. *Adv. Perf. Anal.* **2** 129–158.

den Boer, A. V. 2015. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in operations research and management science* **20**(1) 1–18.

Eick, S. G., W. A. Massey, W. Whitt. 1993a. Infinite-server approximations for multi-server loss models with time-dependent arrival rates. Working paper, Columbia University.

Eick, S. G., W. A. Massey, W. Whitt. 1993b. The physics of the $M_t/G/\infty$ queue. *Operations Research* **41**(4) 731–742.

Erlang, A. K. 1917. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren* **13** 5–13.

Fan-Orzechowski, X., E. A. Feinberg. 2006. Optimality of randomized trunk reservation for a problem with a single constraint. *Adv. Appl. Probab.* **38**(1) 199–220.

Gallego, G., M. Hu. 2014. Dynamic pricing of perishable assets under competition. *Management Science* **60**(5) 1241–1259.

Ge, D., G. Wan, Z. Wang, J. Zhang. 2013. A note on appointment scheduling with piecewise linear cost functions. *Mathematics of Operations Research* **39**(4) 1244–1251.

Greenberg, A. G., R. Srikant, W. Whitt. 1999. Resource sharing for book-ahead and instantaneous-request calls. *IEEE/ACM Transactions on Networking* **7**(1) 10–22.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions* **40**(9) 800–819.

Hunt, P. J., C. N. Laws. 1997. Optimization via trunk reservation in single resource loss systems under heavy traffic. *Ann. Appl. Probab.* **7**(4) 1058–1079.

Iyengar, G., K. Sigman. 2004. Exponential penalty function control of loss networks. *Ann. Appl. Probab.* **14**(4) 1698–1740.

Kaandorp, G. C., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Science* **10**(3) 217–229.

Karlin, S., H. E. Taylor. 1981. *A second course in stochastic processes*. Elsevier, New York.

Kaufman, J. S. 1981. Blocking in a shared resources environment. *IEEE Trans. Comm.* **29** 1474–1481.

Kelly, F. P. 1991. Effective bandwidths at multi-class queues. *Queueing Systems* **9**(1-2) 5–16.

Key, P. 1990. Optimal control and trunk reservation in loss networks. *Probab. Engrg. Informs. Sci.* **4** 203–242.

Kong, Q., C.-Y. Lee, C.-P. Teo, Z. Zheng. 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations research* **61**(3) 711–726.

Kumar, S., R. Srikant, P. R. Kumar. 1998. Bounding blocking probabilities and throughput in queueing networks with buffer capacity constraints. *Queueing Systems* **28**(1-3) 55–77.

Lawler, G. F. 2006. *Introduction to Stochastic Processes*. Chapman & Hall.

Lei, Y., S. Jasin. 2016. Real-time dynamic pricing for revenue management with reusable resources and deterministic service time requirements. Working paper, University of Michigan, available at https://ssrn.com/abstract=2816718.

Levi, R., A. Radovanovic. 2010. Provably near-optimal LP-based policies for revenue management in systems with reusable resources. *Operations Research* **58**(2) 503–507.

Levi, R., C. Shi. 2015. Dynamic allocation problems in loss network systems with advanced reservation. Working paper, University of Michigan, available at arXiv:1505.03774.

Louth, G., M. Mitzenmacher, F. Kelly. 1994. Bounding blocking probabilities and throughput in queueing networks with buffer capacity constraints. *Theoret. Comput. Sci.* **125** 45–59.

Lu, Y., A. Radovanovic. 2007. Asymptotic blocking probabilities in loss networks with subexponential demands. *J. Appl. Probab.* **44**(4) 1088–1102.

Luss, H. 1977. A model for advanced reservations for intercity visual conferencing services. *Journal of the Operational Research Society* **28**(2) 275–284.

Maglaras, C. 2006. Revenue management for a multiclass single-server queue via a fluid model analysis. *Operations Research* **54**(5) 914–932.

Maillardet, R. J., P. G. Taylor. 2016. Queues with advanced reservations: an infinite-server proxy for the bookings diary. *Advances in Applied Probability* **48**(1) 13–31.

Mak, H.-Y., Y. Rong, J. Zhang. 2014. Appointment scheduling with limited distributional information. *Management Science* **61**(2) 316–334.

Massey, W. A. 1985. Asymptotic analysis of the time dependent $M/M/1$ queue. *Mathematics of Operations Research* **10**(2) 305–327.

Miller, B. 1969. A queueing reward system with several customer classes. *Management Science* **16**(3) 234–245.

Nadarajah, S., Y. F. Lim, Q. Ding. 2015. Dynamic pricing for hotel rooms when customers request multiple-day stays. Working paper, University of Illinois at Chicago, available at https://ssrn.com/abstract=2639188.

Özer, Ö., R. Phillips. 2012. *The Oxford handbook of pricing management*. Oxford University Press.

Puhalskii, A. A, M. I. Reiman. 1998. A critically loaded multirate link with trunk reservation. *Queueing Systems* **28**(1-3) 157–190.

Ross, K., D. Tsang. 1989. The stochastic knapsack problem. *Management Science* **37**(7) 740–747.

Ross, K., D. Yao. 1990. Monotonicity properties for the stochastic knapsack. *IEEE Trans. Inform. Theory* **36**(5) 1173–1179.

Schmidt, V. 1987. On joint queue-length characteristics in infinite-server tandem queues with heavy traffic. *Adv. Appl. Prob.* **19** 474–486.

Sevastyanov, B. A. 1957. An ergodic theorem for markov processes and its application to telephone systems with refusals. *Theory of Probability and its Applications* **2**(1) 104–112.

Spencer, J., M. Sudan, K. Xu. 2014. Queue with future information. *Annuals of Applied Probability* **24**(5) 2091–2142.

Srikant, R., W. Whitt. 2001. Resource sharing for book-ahead and instantaneous-request calls using a CLT approximation. *Telecommunication Systems* **16**(3-4) 233–253.

Talluri, K. T., G. J van Ryzin. 2005. *The theory and practice of revenue management*. Springer, New York.

van de Vrugt, M., N. Litvak, R. J. Boucherie. 2014. Blocking probabilities in Erlang loss queues with advance reservation. *Stochastic models* **30**(2) 187–196.

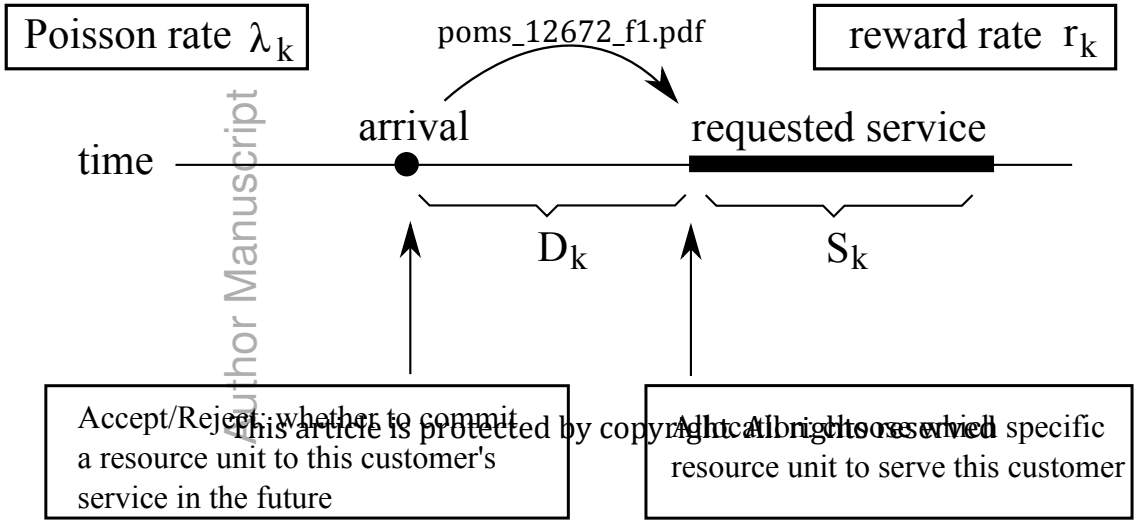
Virtamo, J. T., S. Aalto. 1991. Stochastic optimization of reservation systems. *European Journal of Operational Research* **51**(3) 327–337.

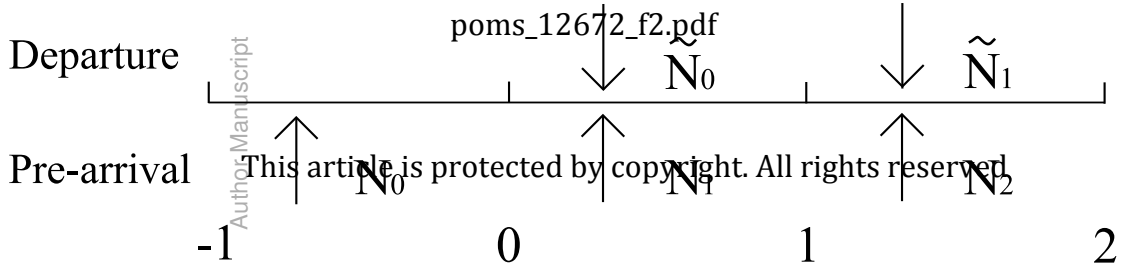
Whitt, W. 1985. Blocking when service is required from several facilities simultaneously. *AT&T Tech. J.* **64** 1807–1856.

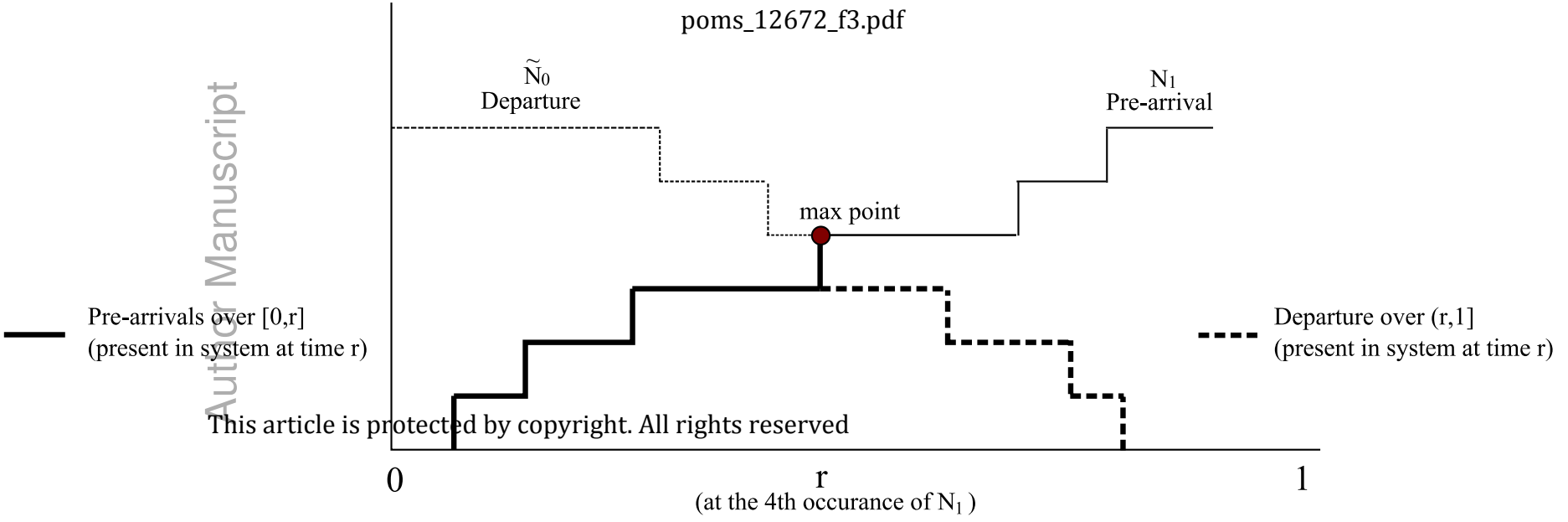
Wischik, D., A. Greenberg. 1998. Admission control for booking ahead shared resources. *INFOCOM'98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2. IEEE, 873–882.

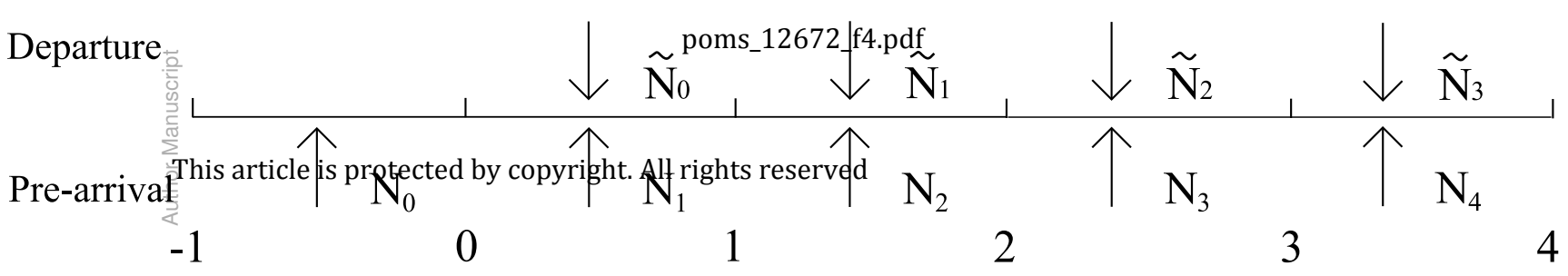
Xu, K. 2015. Necessity of future information in admission control. *Operations Research* **63**(5) 1213–1226.

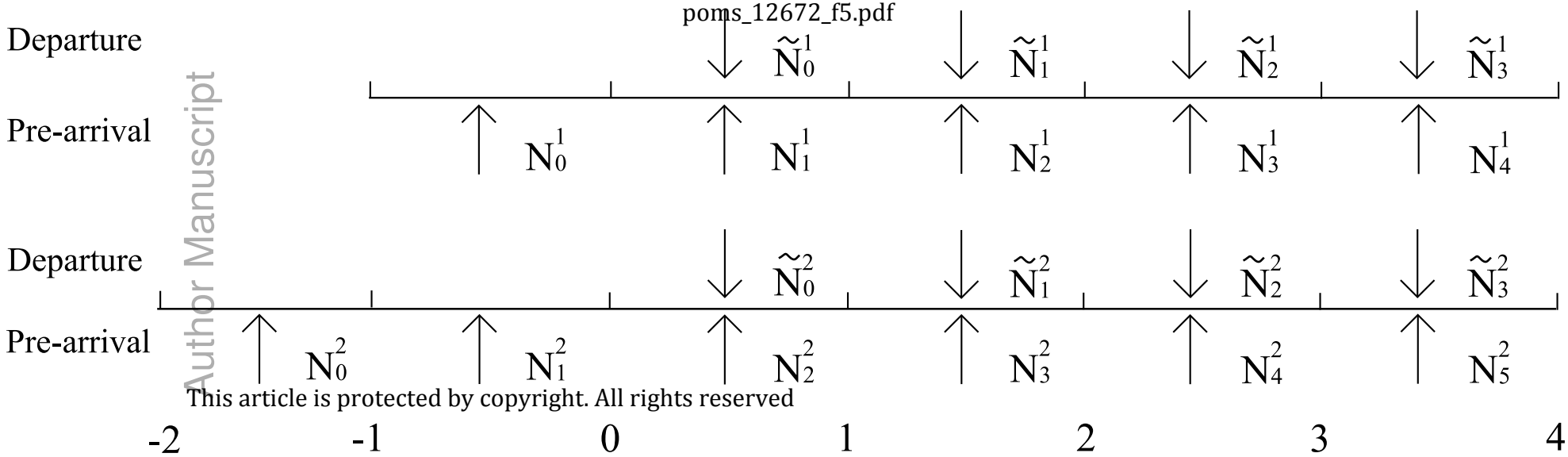
Zachary, S. 1991. On blocking in loss networks. *Adv. Appl. Probab.* **23**(2) 355–372.



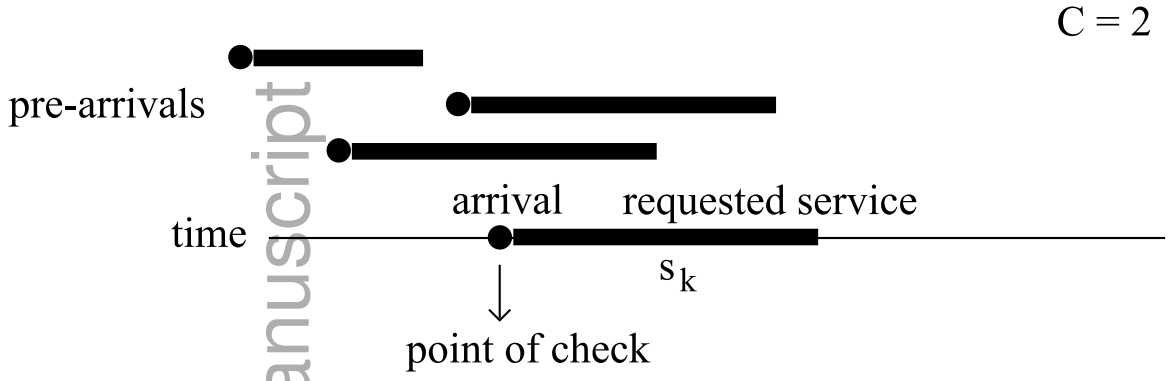








Models without Adv. Res. - suffices to check the instantaneous load



Model with Adv. Res. - needs to scan the entire service interval